THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Bioinformatique

École doctorale CBS2

Unité de recherche 9002

Design and implementation of bioinformatic tools for RNA sequencing data analysis

Présentée par Claudio LORENZI
Le 20/10/2021

Sous la direction de William RITCHIE
et Alban MANCHERON

Devant le jury composé de

Hervé SEITZ, Directeur de Recherche, CNRS – Institut de Génétique Humaine de Montpellier	Président du jury
Daniel GAUTHERET, Directeur de Recherche, CNRS – Institute for Integrative Biology of the Cell	Rapporteur
Eduardo EYRAS, Full Professor, ANU - John Curtin School of Medical Research	Rapporteur
Camille MARCHET, Chargée de Recherche, CNRS – Centre de Recherche en Informatique, Signal et Automatique de Lille	Examinatrice
Alban MANCHERON, Maître de Conférences, Université de Montpellier, Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier	Co-Directeur de thése
William RITCHIE, Directeur de Recherche, CNRS – Institut de Génétique Humaine de Montpellier	Directeur de thése

UNIVERSITÉ DE MONTPELLIER

Table of contents

Table of contents 1

Abstract 2

Résumé 3

Publications 12

Preamble 14

Introduction 15

RNA: a versatile macromolecule with a key role in the cellular system 15

RNA quantification: a technological breakthrough 23

RNA-seq data analysis 33

Identification of IR events 45

IRFinder-S: a comprehensive suite to discover and explore intron retention 46

Alternative approaches for the RNA-seq data analysis 63

GECKO is a genetic algorithm to classify and explore high throughput sequencing data 63

iMOKA: κ-mer based software to analyze large collections of sequencing data 72

Conclusions 93

References 95

Annexes 118

A cell-to-patient machine learning transfer approach uncovers novel basal-like breast cancer prognostic markers amongst alternative splice variants 118

PickPocket: Pocket binding prediction for specific ligand families using neural networks. 138

NF90 modulates processing of a subset of human pri-miRNAs 164

Translesion DNA synthesis-driven mutagenesis in very early embryogenesis of fast cleaving embryos 180

Acknowledgements 198

Abstract

A large portion of the information contained in next-generation sequencing data is potentially lost through classical bioinformatics analysis. Both the mapping of sequencing reads to a genome or transcriptome and filtering results to focus on known gene regions eliminate useful information. This is especially true in cancer studies where patient transcriptomes or genomes may vary from their references. We created a novel approach that makes use of recent advances in genetic algorithms, neural networks and feature selection to comprehensively explore massive volumes of sequencing data to classify samples without these biases. Our approach, called GECKO for GEnetic Classification using κ-mer Optimisation maximizes the sequencing information used when trying to explain the difference between 2 or more samples. Our algorithm has been effective at classifying data from large-scale cancer studies using mRNA-seq, circulating DNA or whole-genome resequencing.

iMOKA (interactive multi-objective *κ-mer* analysis) is a software that enables the comprehensive analysis of sequencing data from large cohorts to generate robust classification models or explore specific genetic elements associated with disease aetiology. iMOKA uses a fast and accurate feature reduction step that combines a Naïve Bayes classifier augmented by an adaptive entropy filter and a graph-based filter to rapidly reduce the search space. By using a flexible file format and distributed indexing, iMOKA can easily integrate data from multiple experiments and also reduces disk space requirements and identifies changes in transcript levels and single nucleotide variants.
Our software could be run on a desktop computer and enable scientists and clinicians to discover novel informative sequences in their own NGS data.

Accurate quantification and detection of intron retention levels require specialized software. Building on our previous software, we have created a suite of tools: IRFinder-S, to analyse and explore intron retention events in multiple samples. Specifically, IRFinder-S allows a better identification of true intron retention events using a convolutional neural network, allows the sharing of intron retention results between labs, integrates a dynamic database to explore and contrast available samples and provides a tested method to detect differential levels of intron retention.

Résumé

Introduction
Dans ce résumé en français sont inclus les principaux concepts de chacun des paragraphes de cette thèse.

L'ARN : une macromolécule polyvalente avec un rôle clé dans le système cellulaire

Épissage d'ARNm

Dans les génomes eucaryotes, l'information pour produire une protéine spécifique n'est pas continue mais divisée en segments, appelés exons, divisés par des régions non codantes ou ne codant pas pour cette protéine, les introns.
La rétention d'intron (IR) se produit lorsque des séquences qui sont habituellement épissées sont maintenues dans le transcrit mature.
L'IR n'est pas simplement le résultat d'un mauvais épissage, mais il a été rapporté qu'il est omniprésent et susceptible d'affecter plus de 80 % de tous les gènes codant pour les protéines, contribuant à la régulation du transcriptome et jouant un rôle non seulement dans les maladies, mais aussi dans les processus physiologiques. .
Habituellement, les isoformes IR (IRI) contiennent des codons de terminaison prématurés qui déclenchent leur dégradation rapide par la voie NMD. Dans certains cas, au cours de la spermatogenèse, par exemple, le transcrit IRI peut être retenu dans le noyau ou le cytoplasme et être soumis à un épissage supplémentaire en réponse à des stimuli, montrant une demi-vie plus longue que les transcrits correctement épissés. Enfin, l'IRI peut également échapper à la NMD et subir une traduction, produisant des isoformes protéiques alternatives, généralement tronquées et nocives pour la cellule.

Réseaux de régulation des gènes
Chaque cellule d'un même individu contient une copie du même génome, appelé génotype, mais elles peuvent se différencier en plusieurs types cellulaires aux formes, dimensions, fonctions et propriétés très différentes, appelés phénotypes.
Ce qui détermine le phénotype de chaque cellule est non seulement son génotype, constant dans chaque cellule, mais aussi le milieu environnant et, surtout, son interaction entre les produits du génome et l'environnement.
Cette interaction affecte la façon dont le génome est utilisé dans chaque cellule, quels transcrits d'ARN sont exprimés, quand et combien, générant différents modèles dans un réseau de régulation génique complexe
En raison de le contrôle de la qualité de la transcription et des étapes de régulation, la quantité de transcrits ne correspond pas toujours à la quantité de la protéine correspondante, mais les informations recueillies à partir des données d'expression de l'ARN-seq sont parfois suffisantes pour déduire des modèles informatiques des

parties des réseaux de régulation génique sous-jacents pour reproduire son comportement dans des environnements contrôlés.
Plutôt que de tenter de décrire les interactions entre les éléments connus du réseau, au cours des dernières années, les approches d'apprentissage automatique et d'apprentissage profond se sont massivement développées: des modèles de boîte noire sont entraînés pour prédire des phénotypes spécifiques à l'aide de données de grande dimension. Ces méthodes peuvent utiliser différents types de caractéristiques d'entrée, telles que l'expression des gènes, la méthylation de l'ADN, les modifications des histones et le génotype, en les considérant individuellement ou en combinaison à partir de grandes cohortes de patients.

Quantification de l'ARN : une rupture technologique

Séquençage de première génération

Également connue sous le nom de séquençage de Sanger, la méthode qui permettait de déterminer la séquence de longs fragments de toute molécule d'ADN a été publiée en 1975 et affinée au cours des années suivantes.
L'application de cette technologie s'étend de l'assemblage *de novo* du génome, comme le projet du génome humain (HGP) dont la première ébauche a été publiée en 2001, à la biologie évolutive, pour déterminer la phylogénie des organismes ou l'évolution des gènes, en passant par les applications cliniques, comme la détection de pathogènes ou les tests de mutations génomiques dans les pathologies congénitales, ou encore à l'identification médico-légale et aux tests de paternité, grâce aux empreintes génétiques.

Séquençage de deuxième génération

La réduction des coûts et l'augmentation de l'accessibilité ont permis d'appliquer les NGS dans un large éventail de domaines : le reséquençage du génome, c'est-à-dire la lecture de séquences cartographiques sur un génome de référence pour identifier des variantes génétiques ; les tests prénataux non invasifs, la classification moléculaire du cancer et le diagnostic des maladies mendéliennes ne sont que quelques exemples des nombreuses applications cliniques qui sont devenues des routines réalisables dans les hôpitaux.
Plusieurs méthodes ont été dérivées de l'ADN-seq standard pour quantifier différentes molécules et événements. Un exemple est le séquençage de l'ARN qui, en utilisant la transcriptase inverse et des protocoles dédiés, a presque complètement remplacé la technologie des puces à ADN pour la quantification de l'expression génique.

Séquençage de troisième génération

Deux décennies de travail et d'avancées technologiques ont été nécessaires pour un premier prototype fonctionnel de nanopore et à la fondation en 2005 de la société Oxford Nanopore Technology (ONT).
L'ONT utilise une différence de tension appliquée aux bains d'électrolytes de chaque côté d'une membrane isolée pour produire un courant ionique.

La précision de l'ONT était inférieure à 60 % lors de sa première introduction, mais les améliorations du *base calling* au cours des dernières années ont permis des valeurs de 85 % en 2018 et jusqu'à 98,3 % en 2021 et promettant 99 % avec la version chimique Q20+.
Si l'entreprise britannique atteint cet objectif impensable d'ici quelques années, la technologie des nanopores aura toutes les caractéristiques pour remplacer l'Illumina dominant sur le marché mondial et plus encore : l'absence d'étape d'imagerie permet la production d'appareils moins chers et plus petits, avec l'appareil MinION étant aussi gros qu'un smartphone et coûtant 1000 dollars ; la longueur de lecture peut aller de lecture courte à ultra-longue (plus de 2Mb d'ADN et plus de 20Kb d'ARN) ; il permet une analyse en temps réel et la préparation de la bibliothèque est rapide, ne nécessitant que dix minutes, et standardisée, grâce à un dispositif automatisé qui augmente la reproductibilité des expériences.

Nanostring nCounter: quantification directe d'ARN
Une technologie émergente qui permet la quantification directe de molécules d'ARN à l'aide d'un protocole simple et rapide est Nanostring Technologies nCounter. La société Nanostring, fondée en 2003 et installée à Seattle, propose une technique efficace en termes de coût et de temps pour quantifier des ensembles de séquences spécifiques.
Cette plate-forme automatisée hybride les marqueurs moléculaires fluorescents directement à des séquences d'acides nucléiques spécifiques, permettant la mesure non amplifiée de jusqu'à 800 cibles dans un échantillon et de multiplexer jusqu'à 96 échantillons dans le même cycle.

Analyse des données RNA-seq
Conception expérimentale
À l'instar d'autres expériences scientifiques, le RNA-seq nécessite une préparation minutieuse des données qui doivent être générées ou collectées. Une étude peut être exploratoire, avec l'objectif de découvrir de futures tâches de recherche, ou formelle, avec une hypothèse à tester.
L'application standard des données RNA-seq est l'étude de l'expression différentielle (DE) des gènes et, moins fréquemment, des transcrits.
Différentes applications nécessitent différentes dimensions d'échantillon : si nous voulons, par exemple, associer un SNP à un phénotype particulier, nous devons appliquer des tailles d'échantillon d'étude d'association pangénomique (GWAS), avec un minimum de 100 échantillons jusqu'à plus de 2000.
Pour ce qui concerne la profondeur, puisque plus de 80 % des lectures sont attribuées aux 10 % de gènes les plus exprimés et qu'augmenter le nombre de lectures n'augmente que marginalement la couverture des gènes faiblement exprimés, surtout au-delà des 10 millions de lectures, il vaut mieux utiliser le budget pour avoir plus de réplicats plutôt que peu d'échantillons avec un séquençage profond.

Enfin, lorsque l'expérience doit être exécutée en plusieurs lots, il est important de répartir équitablement les conditions entre les lots. Le traitement de groupes d'échantillons à des jours différents, à l'aide de différentes machines et par différents opérateurs peut refléter de faibles écarts entre les lots qui peuvent être interprétés à tort comme des signaux biologiques.

Alignement de lecture ARN-seq

Pour quantifier l'abondance de la molécule d'ARN au niveau du transcrit, en considérant chaque isoforme comme une entité indépendante, ou au niveau du gène, où l'expression d'un gène est la somme de l'expression de ses isoformes, il est nécessaire d'aligner les lectures à un génome ou transcriptome de référence. La cartographie des lectures sur un génome de référence présente le principal défi pour aligner correctement une lecture qui comprend une jonction d'épissage (SJ).

Quantification au niveau des gènes et des transcrits

Indispensable pour la plupart des analyses en aval, l'évaluation de l'abondance des gènes et des transcrits se caractérise également par une longue liste d'outils qui atteignent le même objectif en utilisant différentes stratégies et avec des performances différentes.

HTSeq, featureCounts, l'option intégrée à l'outil STAR et d'autres outils comptent directement les fragments chevauchant les caractéristiques du gène après l'étape d'alignement, différant les uns des autres par la façon dont ils gèrent certaines situations, comme les alignements multiples fragmentés, les fragments qui correspondent à plusieurs caractéristiques et des fragments s'alignant partiellement sur une caractéristique. Cette approche est limitée par des changements dans la composition des exons qui n'ont pas d'impact direct sur le nombre de lectures au niveau des gènes, tels que la capacité d'un même gène à produire différentes isoformes.

Pour surmonter ces obstacles, la quantification au niveau du transcrit est de plus en plus utilisée, même pour estimer l'expression au niveau du gène avec de meilleures performances sur l'analyse en aval. Il convient de mentionner que, contrairement aux transcrits, le gène n'est pas une entité physique mais une abstraction utile n'ayant pas de cible claire pour la quantification.

Des approches récentes utilisent des pseudo-alignements de κ-mers pour accélérer le processus, contourner l'étape d'alignement et produire une estimation précise.

Enfin, les outils classiques de quantification des gènes et des transcrits ne prennent pas en compte les éléments répétitifs et transposables. Des logiciels dédiés, comme TEtranscripts, télescope et SalmonTE, abordent ce problème, en appliquant des approches similaires à celles utilisées pour les gènes classiques aux familles d'éléments transposables.

Signatures d'épissage alternatif

L'abondance des transcrits et des gènes ne sont pas les seules caractéristiques quantifiables qui peuvent être déduites du séquençage de l'ARN : le pourcentage de

l'épissage (PSI) est utilisé dans les études d'épissage pour quantifier la fréquence d'inclusion d'exons spécifiques.
Parmi les événements d'épissage alternatifs possibles, la rétention d'intron (IR) nécessite des ajustements supplémentaires afin d'être correctement quantifiée. Sans une approche appropriée, des sites d'épissage donneurs ou accepteurs alternatifs non annotés et des transcrits qui se chevauchent pourraient conduire à des événements mal classés. De plus, les introns enrichis en séquences de faible complexité et répétitives peuvent restreindre la cartographie unique des données de séquençage.

Analyse différentielle
La plupart des modèles expérimentaux visent à identifier les différences d'expression entre deux ou plusieurs conditions, l'une utilisée comme contrôle et l'autre comme cible. Avec cet objectif, l'analyse d'expression différentielle (DE) formule et teste une hypothèse statistique pour chaque caractéristique dans les échantillons.
Habituellement, seul un nombre limité de réplicats est disponible (3 à 5 réplicats par condition) et, combiné au grand nombre de fonctionnalités testées simultanément, la puissance statistique réalisable serait très faible sans stratégies dédiées mises en œuvre et affinées au cours des années par la communauté statistique.
La plupart de ces approches, telles que le limma-voom largement utilisé, ont été initialement développées pour les données de microarrays et dans un second temps adaptées au séquençage d'ARN.

Approches basées sur les κ-mers
Quantifier l'abondance de transcrits connus ou d'événements d'épissage n'est pas le seul moyen d'obtenir des caractéristiques significatives : compter les occurrences de sous-chaînes de longueur k, appelées κ-mers, dans les données brutes de séquençage est une autre approche largement utilisée dans différents domaines, tels que la métagénomique, l' assemblages *de novo* et la phylogénie.
Ce type de représentation a l'avantage d'être sans référence, puisque le dénombrement des occurrences de κ-mers est indépendant de tout génome, transcriptome ou annotation de référence.
L'inconvénient est qu'il est très redondant et avec une grande dimensionnalité.
La procédure de comptage, bien que simple, présente des défis de calcul pour ce qui concerne les exigences de temps et d'espace.
Une fois les comptes κ-mers obtenus, une approche courante consiste à créer des graphes de Bruijn (dB), un graphe direct représentant les κ-mers en tant que sommets et le chevauchement entre eux en tant qu'arêtes.
L'application de la théorie des graphes aux graphes κ-mer de de Bruijn est l'une des clés du succès de cette méthodologie : cette représentation est gérée efficacement par la machine et il existe un grand nombre d'algorithmes pour rechercher, parcourir, trouver des chemins et représenter ses propriétés.

Identification des événements IR

Huit ans après la publication de la première version d'IRFinder, avec plus de 400 citations cumulées, le logiciel est une référence pour l'analyse IR.

Les raisons de son succès résident non seulement dans la qualité de l'analyse mais aussi dans l'implémentation de bout en bout qui prend en charge tous les aspects de l'analyse des données brutes, y compris la génération de référence du du logiciel d'alignement STAR, le découpage de l'adaptateur et les procédures d'analyse différentielle.

Les aspects de la convivialité du logiciel ont été améliorés au cours de ces années, également grâce aux commentaires des utilisateurs qui ont aidé à résoudre différents bugs. Néanmoins, il reste encore quelques aspects qui nécessitent un effort supplémentaire: le séquençage à lecture longue prend de plus en plus d'importance, en particulier dans les études impliquant la structure des transcrits. Le pipeline est calibré autour d'un séquençage à lecture courte, non seulement pour ce qui concerne le type d'aligneur mais aussi pour les hypothèses qui sont posées pour le calcul de l'IRratio.

Malgré les stratégies utilisées pour masquer les régions chevauchant des régions difficiles à aligner et des caractéristiques connues, telles que des exons supplémentaires et des ARN non codants, il existe une proportion considérable d'événements IR faussement positifs qui peuvent être discriminés par inspection visuelle sur un navigateur génomique.

La base de données IR, IRbase, construite en 2017 à partir de 2000 échantillons humains est obsolète et ne permet pas à l'utilisateur de visualiser et de comparer facilement ses propres données avec celles incluses dans la base de données.

L'approche IR différentielle n'a pas été validée dans les travaux antérieurs et nécessite la connaissance du logiciel R.

Au cours de ma dernière année de doctorat, j'ai travaillé avec mon collègue Sylvain Barrier pour améliorer IRFinder, en me concentrant non seulement sur les quatre points décrits précédemment, mais également en améliorant l'aspect de la convivialité et de la vitesse qui ont conduit à son succès.

Approches alternatives pour l'analyse des données RNA-seq

Des méthodes telles que DE-kupl, KOVER et HAWK ont démontré qu'il n'est pas nécessaire d'intégrer les informations dans un format compréhensible et interprétable par l'homme, tel que des gènes ou des transcriptions, pour comparer les informations contenues dans les données de séquençage. Les κ-mers nous permettent de comparer des groupes d'échantillons de manière agnostique, sans biais induit par aucune séquence de référence ou annotation, ce qui conduit à des résultats hautement reproductibles : les décomptes de κ-mers ne changeront pas, tandis que notre connaissance de la composition du génome de référence s'améliore chaque année. De plus, les κ-mers permettent la comparaison de petites fractions de la molécule d'ARN, évitant la perte d'informations dérivées de l'agrégation de

plusieurs lectures sous une seule caractéristique, c'est-à-dire un gène, un transcrit ou une jonction d'épissage.
Enfin, en utilisant un nombre suffisamment important d'échantillons, il serait possible d'associer des variations, telles que le SNP ou les indels, à une population spécifique, de la même manière que les études d'association pangénomique (GWAS).

GECKO est un algorithme génétique pour classer et explorer les données de séquençage à haut débit

GEnetic Classification using κ-mer Optimization, GECKO, est la première méthode permettant d'identifier des groupes de κ-mers capables de classer deux ou plusieurs groupes d'échantillons dans l'étude de grandes cohortes.
La méthode, décrite en détail dans l'article présenté ultérieurement dans le manuscrit, montre qu'il est possible d'identifier des groupes de κ-mers qui, seuls ou en synergie, sont capables de classer différents groupes de patients, avec une meilleure performance en ce qui concerne le nombre de gènes. L'approche a été testée sur différents types de données de séquençage, tels que les données de séquençage de miARN, d'ARNm et de bisulfite.
Brièvement, GECKO prend des séquences brutes en entrée et utilise Jellyfish2 pour compter l'abondance des κ-mers dans chaque échantillon. Il assemble ensuite une matrice de κ-mers, où chaque ligne est un κ-mer et chaque colonne est un échantillon.
La dernière étape du prétraitement consiste à filtrer des κ-mers considérés comme non informatifs, bruités et redondants.
Enfin, GECKO implémente un algorithme génétique adaptatif, un algorithme d'optimisation métaheuristique efficace, pour sélectionner des sous-ensembles de κ-mers qui maximisent la précision de la classification des groupes d'échantillons à l'aide d'un classificateur de vecteur de support linéaire (LinSVC).

iMOKA : logiciel basé sur κ-mer pour analyser de grandes collections de données de séquençage

iMOKA, Interactive Multi Objective κ-mer Analysis, a d'abord été pensé comme un filtre pour sélectionner les κ-mers informatifs : la plupart des κ-mers sélectionnés par GECKO ont pu individuellement classer avec une assez bonne précision les échantillons dans les groupes respectifs, même à l'aide d'une procédure de validation croisée. À l'instar de GECKO, les détails de l'algorithme sont décrits dans l'article présenté ultérieurement dans le manuscrit, y compris une référence sur quatre ensembles de données dans lesquels les κ-mers extraits par iMOKA sont comparés aux valeurs PSI, à l'expression des gènes et des transcriptions en tant que caractéristiques de classification selon un modèle de forêt aléatoire.
Brièvement, le logiciel peut prendre en entrée à la fois des fichiers de séquençage, tels que fastq ou bam, ou des identifiants de lien externe, http, ftp ou SRR, en téléchargeant les données requises avant le début de l'analyse.

À l'aide de KMC3, iMOKA extrait le décompte des κ-mers triés de chaque échantillon et les convertit en fichiers binaires. Un fichier JSON contient les métadonnées des échantillons appartenant à l'analyse, comprenant pour chaque échantillon : le nom, l'étiquette du groupe, l'emplacement du fichier binaire et la somme totale des comptages des *κ-mers*, utilisés pour normaliser les données .
La première étape de réduction considère un *κ-mer* à la fois et, à l'aide d'un classificateur bayésien, estime la précision de la caractéristique permettant de classer les échantillons dans les groupes respectifs. Cette étape est par défaut couplée à un filtre d'entropie adaptatif qui accélère le processus en supprimant très peu d'éléments réellement informatifs.
Enfin, une procédure d'agrégation regroupe les *κ-mer*s en fonction de leur séquence, en construisant des graphes de de Bruijn et de leur pertinence biologique, en cartographiant les séquences générées à partir des graphes sur un génome de référence et en utilisant une annotation de référence pour attribuer des « événements » aux *κ-mer*s les plus informatifs de dans chaque groupe.
Surtout, le logiciel est couplé à une interface utilisateur graphique (GUI) qui permet d'exécuter en local ou sur un cluster distant toutes les étapes de l'algorithme. L'utilisateur peut également explorer le résultat final de l'étape d'agrégation sous forme de tableau interactif, visualiser l'alignement des *κ-mer*s sur un génome de référence avec une version javascript du navigateur de génome IGV, générer des cartes auto-organisatrices et des classificateurs basés sur des forêts aléatoires.

Conclusion

Les trois dernières décennies ont été marquées par des avancées technologiques incroyables, tant du point de vue biotechnologique que informatique.
Pour les suivre, nous avons adapté IRFinder pour prendre en charge les séquences de troisième génération et utiliser de nouvelles méthodologies, le réseau de neurones convolutifs, pour affiner et améliorer ses résultats. De plus, nous avons proposé IRBase, une plateforme où les utilisateurs peuvent non seulement visualiser leurs données mais aussi les comparer avec celles partagées par d'autres utilisateurs.
La possibilité de séquencer à faible coût et haute fidélité de larges cohortes de personnes donne l'opportunité d'approfondir nos connaissances sur les mécanismes sous-jacents aux pathologies et de générer des modèles pour prédire les réponses aux médicaments, aux traitements et aux modifications environnementales.
En introduction, nous avons vu comment les approches classiques, basées sur la cartographie à un génome de référence et utilisant des annotations de référence, présentent de nombreux niveaux de variabilité induits par les différentes versions des références et des logiciels utilisés. De plus, une grande partie des informations sont généralement rejetées car elles ne correspondent pas aux caractéristiques considérées dans l'étude. Nous avons montré comment les approches basées sur les *κ-mers* peuvent être une représentation optimale et agnostique des données de séquençage, utiles pour identifier des biomarqueurs pouvant être appliqués à des fins cliniques et de recherche. Dans cette optique, nous avons mis en place iMOKA,

un logiciel capable de sélectionner efficacement un groupe de *κ-mer*s avec une faible redondance d'informations et une grande capacité de discrimination des phénotypes en analyse au sein d'une cohorte de très grands échantillons.

Publications

The results of the work done during the realization of this thesis are included in the main text of this thesis.
Three articles had been published in two different peer-reviewed scientific journals[1–3].

1. Thomas, A., Barriere, S., Broseus, L. *et al.* GECKO is a genetic algorithm to classify and explore high throughput sequencing data. *Commun Biol* **2,** 222 (2019). https://doi-org.insb.bib.cnrs.fr/10.1038/s42003-019-0456-9
2. Lorenzi, C., Barriere, S., Villemin, JP. *et al.* iMOKA: *κ-mer* based software to analyze large collections of sequencing data. *Genome Biol* **21,** 261 (2020). https://doi-org.insb.bib.cnrs.fr/10.1186/s13059-020-02165-2
3. Lorenzi, C., Barriere, S., Arnold, K. *et al.* IRFinder-S: a comprehensive suite to discover and explore intron retention. *Genome Biol* **22,** 307 (2021). https://doi.org/10.1186/s13059-021-02515-8

As a result of the collaboration with three laboratories in our institute, two other publications and two preprints are included in the Annexes:

1. Villemin, JP., Lorenzi, C., Cabrillac, MS. *et al.* A cell-to-patient machine learning transfer approach uncovers novel basal-like breast cancer prognostic markers amongst alternative splice variants. *BMC Biol* **19,** 70 (2021). https://doi-org.insb.bib.cnrs.fr/10.1186/s12915-021-01002-7
2. PickPocket: Pocket binding prediction for specific ligands family using neural networks. Benjamin Thomas VIART, Claudio Lorenzi, María Moriel-Carretero, Sofia Kossida. bioRxiv 2020.04.15.042655; https://doi.org/10.1101/2020.04.15.042655
3. Giuseppa Grasso, Takuma Higuchi, Victor Mac, Jérôme Barbier, Marion Helsmoortel, Claudio Lorenzi, Gabriel Sanchez, Maxime Bello, William Ritchie, Shuji Sakamoto, Rosemary Kiernan, NF90 modulates processing of a subset of human pri-miRNAs, *Nucleic Acids Research*, Volume 48, Issue 12, 09 July 2020, Pages 6874–6888, https://doi-org.insb.bib.cnrs.fr/10.1093/nar/gkaa386
4. Translesion DNA synthesis-driven mutagenesis in very early embryogenesis of fast cleaving embryos. Elena Lo Furno, Isabelle Busseau, Claudio Lorenzi, Cima Saghira, Stephan Zuchner, Domenico Maiorano. bioRxiv 2020.11.28.401471; doi: https://doi.org/10.1101/2020.11.28.401471 (Currently under revision for Nucleic Acid Research)

The softwares developed during the thesis are available at the following GitHub repositories:

1. https://github.com/RitchieLabIGH/GECKO (partial partecipation)
2. https://github.com/RitchieLabIGH/iMOKA
3. https://github.com/LucoLab/Villemin_2020 (shared project)
4. https://github.com/RitchieLabIGH/IRFinder

Preamble

With this manuscript, I wish not only to give a general overview of my past three years of passionate work but also to convey a progressive view of why researchers all over the world cooperate every day to advance the knowledge about ourselves and the world around us. The acceptance of risks and failures in everyday challenges, the constant curiosity and the awareness that every piece of certainties that we have can be questioned thanks to technological advances are the keys for success in this field, together with a bit of luck.
We'll explore biological events essential for life, shaped by evolution in hundreds of thousands of years. We'll analyse machinery built to quantify those events that would have been considered sci-fi products by our grandfathers. Finally, I will introduce my work that aims in part to use the data generated by those instruments using well-corroborated methods to identify fine regulatory elements in complex systems and in part to change perspective on how we use this huge amount of information.

Introduction

RNA: a versatile macromolecule with a key role in the cellular system

Ribonucleic acid (RNA) is one of the two main classes of nucleic acids together with deoxyribonucleic acid (DNA), two polynucleotide chains that carry all the information required to orchestrate the organization of the cell.
RNA is synthesized by the RNA polymerases in complexes that use a DNA segment as a template and involves a wide network of regulators.
Although the existence of such molecules had been known since 1869, more than a century was required to reveal their chemical composition[4,5], their role [6,7], their structures[8], and only in 1977[9] we were able to read the information carried by those molecules using techniques that we'll describe in detail in the following chapter.
Furthermore, despite the first draft of full human genome assembly being available for 20 years[10,11], our knowledge about the complex mechanism underlying the generation of a multicellular organism from a single omnipotent cell and the effect of small genomic variations on such organisms is still limited.
The analysis of RNA and protein behaviour in response to genomic alteration can be the key for further understanding since those molecules are the effectors that use the information to act in the cellular environment.

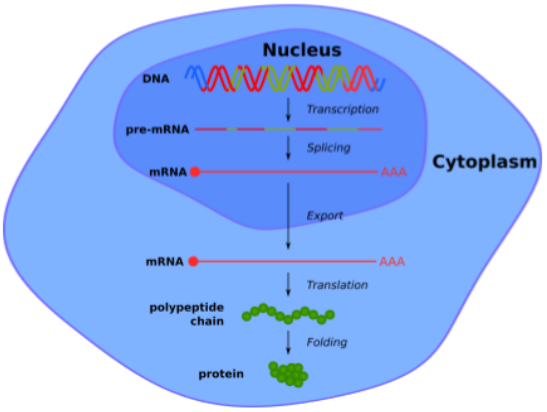


Figure 1: Schematic representation of the central dogma of molecular biology

According to the central dogma of molecular biology formulated by Crick in 1957[12,13], the information to produce a protein is encoded as a four-letter alphabet sequence in the nucleus, it's transcribed into RNA molecules (messenger RNA, mRNA) that have the role to export the information from the nucleus compartment to the cytosol, the liquid matrix surrounding the organelles, where it is translated into a twenty letter alphabet amino acid sequence.

Figure 1 shows a schematic version of this process, from the transcription of the precursor mRNA molecule, pre-mRNA, its maturation through splicing, 5' capping,

polyadenylation and formation of the messenger ribonucleoprotein particle (mRNP), the export in the cytoplasm and its final translation in a polypeptide chain, that will fold in a functional protein. In sixty years from the formulation of the first version of this dogma, we discover that this process is part of a complex network made of effectors and regulators that interact to ensure the survival and the reproduction not only of the single cell but of the whole organism of which the cell is part of.
In this system, the role of RNA molecules goes far beyond the mere carrier of information from DNA to protein: they can have catalytic, structural and regulatory functions[14].
Within the following paragraphs, we'll focus our attention on the RNA regulatory strategies that take place in eukaryotic cells focusing our attention on the underlying informational flow.

mRNA splicing

In eukaryotic genomes the information to produce a specific protein is not continuous but split into segments, called exons, divided by non-coding regions or not coding for that protein, the introns.
The spliceosome is a ribonucleoprotein complex in which five small nuclear RNAs (snRNAs), approximately 300 proteins and magnesium ions cooperate to remove the introns from the pre-mRNA molecule in a two-step transesterification reaction[15].
Donor, acceptor and branch sites are *cis*-acting elements necessary for the recognition of the splice boundaries by the spliceosome and are located respectively at the 5', 3' and 20-50 nucleotides upstream the 3' end of the intron[16].

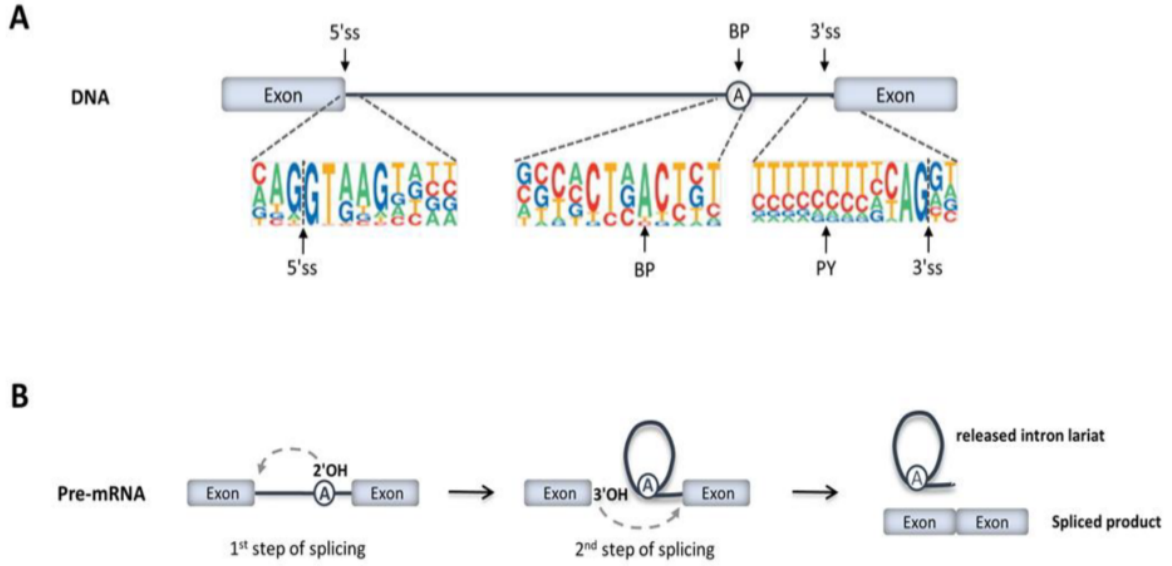


Figure 2: Precursor messenger RNA (pre-mRNA) splicing. **A)** Exons are represented by boxes and introns by lines. The most conserved nucleotides at the 5' splice site (5'ss), branch point (BP), polypyrimidine tract (PY), and 3' splice site (3'ss) are indicated. **B)** The two transesterification reactions that result in the excision of introns from pre-mRNA are represented.
From Int. J. Mol. Sci. 2020, 21(4), 1329; https://doi.org/10.3390/ijms21041329

The process is highly dynamic and not deterministic: the sites are not always recognised by the spliceosome with the same efficiency. This flexibility leads to the inclusion and exclusion of different portions in different mature mRNA isoforms, and, as consequence, the possible formation of a variety of different proteins from the same gene, increasing the genetic diversity. This phenomenon, called alternative splicing (AS), gives plasticity to the transcriptome playing a key role during cell development and differentiation[17]. AS is finely regulated by *cis*-acting elements, *trans*-acting factors, transcription and chromatin structure, whose combinatorial effect determines the final outcome [18,19] .

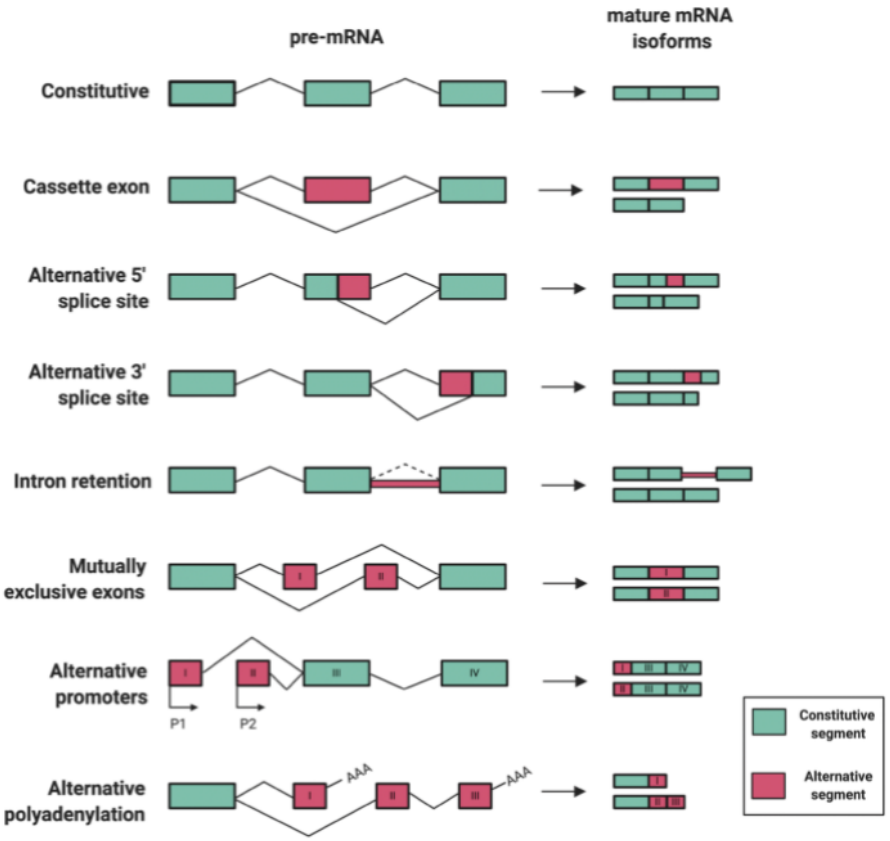


Figure 3: Representation of seven alternative splicing configurations. Boxes represent exons and lines represent introns.
From Bhadra, M., Howell, P., Dutta, S. *et al.* Alternative splicing in aging and longevity. *Hum Genet* **139,** 357–369 (2020). https://doi-org.insb.bib.cnrs.fr/10.1007/s00439-019-02094-6

AS can occur in different locations (Figure 3), but not all of their combinations result in a functional protein. The mature transcripts undergo degradation if specialized surveillance systems detect abnormalities in the mRNA sequence, such as the nonsense-mediated decay (NMD) and the non-stop decay (NSD)[20,21].
The NMD is mediated by proteins associated with the exon-exon junction (EJ Complex, EJC) and the ribosome.
During the first round of translation, the ribosome removes the EJC but, in case of a premature stop codon, the ribosome is released before reaching the last EJC. The translation termination recruits additional factors that, in the case of residual EJC on the mRNA, trigger the degradation of the mRNA by the exosome complex. Similarly, the NSD occurs when the ribosome stalls at the poly-A, discharging the ribosome and redirecting the mRNA to degradation.
Those abnormalities are more frequent in transcripts containing introns.
Intron retention (IR) occurs when sequences that are usually spliced out are maintained in the mature transcript.
IR is not simply the outcome of mis-splicing but has been reported to be ubiquitous and likely to affect over 80% of all protein-coding genes [22,23], contributing to the transcriptome regulation [24] and having a role not only in diseases [25–27], but also in physiological processes [28,29].

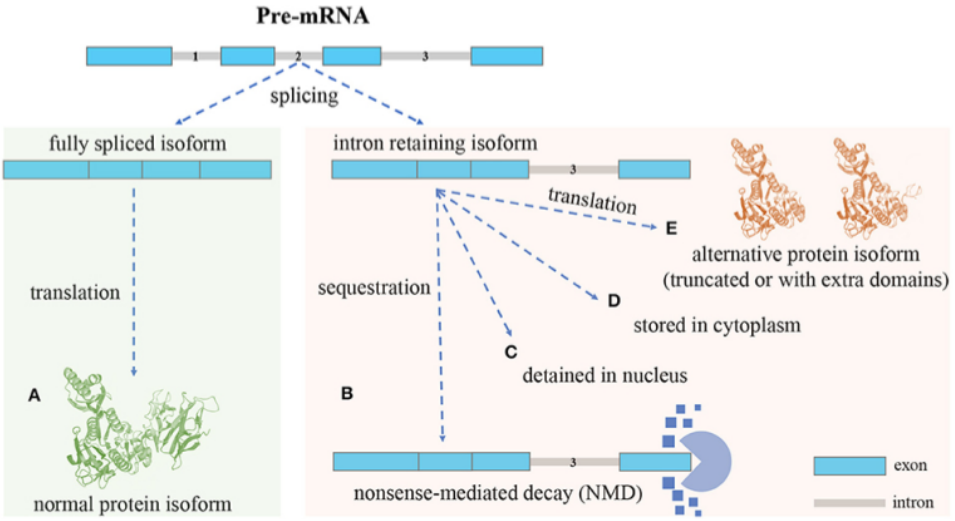


Figure 4: An overview of the IR mechanism. **A)** Fully spliced isoforms are sent out of the nucleus for translation. **B)** the IRIs are degraded by the NMD pathway. **C)** the IRIs are detained in the nucleus, and in response to stimuli these IRIs can undergo further splicing to remove the retained intron, before being exported out of nucleus for translation **D)** In the case of cytoplasmic splicing, IRIs are shuttled to the cytoplasm for preservation and may be subject to further splicing
E) IRIs escape from the NMD pathway and are translated into protein isoforms, which are often truncated and may lose domains; the alternative protein isoforms may include extra domains formed by the amino acid sequences translated from retained introns.
From: Zheng J-T, Lin C-X, Fang Z-Y and Li H-D (2020) Intron Retention as a Mode for RNA-Seq Data Analysis. *Front. Genet.* 11:586. doi: 10.3389/fgene.2020.00586

Usually, IR isoforms (IRI) contain premature termination codons that trigger their rapid degradation by the NMD pathway. In some cases, during spermatogenesis, for example, IRI transcript can be retained in the nucleus or cytoplasm and be subject to further splicing in response to stimuli, showing a longer half-life than properly spliced transcripts[30]. Finally, IRI can also escape NMD and undergo translation, producing alternative protein isoforms, usually truncated and harmful to the cell[31–33].
Although this type of gene regulation requires the formation of the mature RNA and its degradation, therefore inefficient under the energetical point of view respect the downregulation at the transcription level, it's more specific compared to transcription factors, whose action covers a wide panel of genes. Since the energetic cost at the transcription level is much lower than the one at the protein level[34] and the speed of translation is much higher than the transcription one[35], the generation of a reservoir of IRI transcripts allows to have an energetically efficient and fast way to produce proteins in response to external stimuli.
In the next paragraph, we'll focus our attention on the gene regulatory network, which is the complex system where proteins, DNA and RNA molecules interact to ensure the survival of the living organism.

Gene regulatory networks

Cooperation is strength and complex multicellular organisms are the perfect incarnation of this concept. Unicellular organisms are self-sufficient cells able to provide all the functions needed for the survival and reproduction of their species. In multicellular organisms each cell depends on the activity of each other, generating complex systems having emergent properties.
Each cell of the same individual contains a copy of the same genome, called genotype, but they can differentiate in several cell types with very different shapes, dimensions, functions and properties, called phenotypes.
What determines the phenotype of each cell cannot be only its genotype, constant in each cell, but also the surrounding environment and, more importantly, by the interaction between the genome products and the environment.
This interaction affects how the genome is used in each cell, which RNA transcripts are expressed, when and how much, generating different patterns in a complex gene regulatory network[36].
In such a network the abundances of each transcript and protein have to be finely tuned by pathways characterized by regulatory intercommunicating loops.
Traditionally, the transcriptional pattern is modulated at two interconnected levels: a first level having transcription factors (TF) that bind enhancer elements and recruit cofactors and RNA polymerase II to target genes [37], and a second at the epigenetic level that involves chromatin, its regulators and the DNA methylation[38]. As we saw in the previous paragraph, however, there are additional control levels that influence the network: RNA-binding proteins and non-coding RNAs, such as miRNA[39] and siRNA[40], regulate the mRNA processing[41–43], transport and degradation [44,45].
Furthermore, protein translation and degradation are finely regulated, the first at the levels of initiation, elongation, localization and ribosome composition[46–48], the second with the ubiquitin-proteasome system[49]. Finally, the phenotype arises from the protein's activity, their composition, influenced also by post-translational modifications and their interaction with other proteins and biomolecules[50,51].

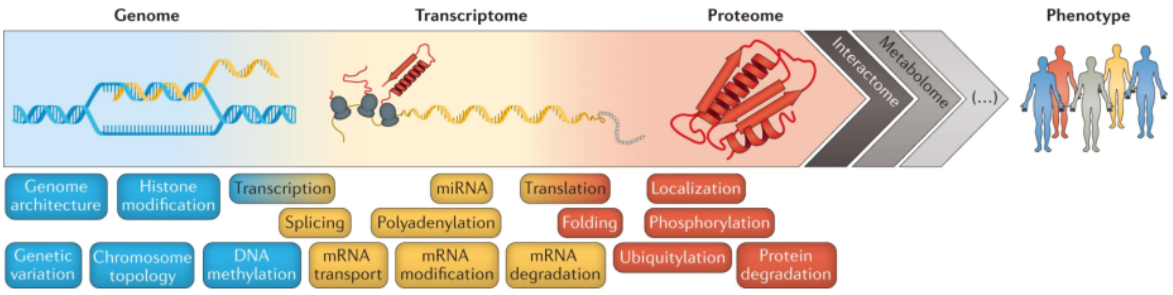


Figure 5: From genotype to phenotype: the processes at different steps in the gene regulatory pathway that confer regulatory control are indicated at the bottom.
From Buccitelli, C., Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet 21, 630–644 (2020).
https://doi-org.insb.bib.cnrs.fr/10.1038/s41576-020-0258-4

Mammals contain thousands of cell types, each with a specific transcriptome and proteome pattern where the alteration of a single key component can cause diseases. For example, the oncogenic transcription factor TAL-1, overexpressed in almost half of T cell lymphoblastic leukaemia cases, forms an interconnected autoregulatory loop with several key TF partners[52]. Alteration of the information content is a source of intraspecies variability, but if a mutation disrupts the balance of the regulatory network it can cause developmental deficiencies, like missense mutations in the RNA polymerase II Mediator subunit MED12 that cause intellectual disability and multiple congenital anomalies[53].

Because of transcript control quality and regulation steps, the quantifiable amount of a transcript doesn't always correspond to the amount of the corresponding protein[35,54], but the information gathered from RNA-seq expression data is sometimes sufficient to infer computational models of portions of the underneath gene regulatory networks to reproduce its behaviour in controlled environments. The reaction kinetics in those models can be described using mathematical models, such as sets of coupled ordinary or stochastic differential equations [55,56] , boolean and bayesian networks. The strengths and the weakness of those methods are accurately described in the review of M. Banf [57], where the author highlights the importance of those methods in prescreening *in silico* the potential interactions, limiting the extent of experimentation needed. However, the high complexity of the gene regulatory network, its interaction with other cellular pathways and the difficulties to correctly quantify all its components at the same time are the main obstacles for the creation of a complete descriptive computational model, especially when the models are based uniquely on expression data and not integrated with consistent, large-scale multiple data types.
Rather than attempt to describe the interactions between the known elements of the network, in the last few years machine learning and deep learning approaches are flourishing: black-box models are trained to predict specific phenotypes using high-dimensional data[58,59]. Those methods can use different types of input features, such as gene expression, DNA methylation, histone modifications and genotype, either considering these individually or in combination from large cohorts of patients [60–63].
The main challenges of this approach are to gather data correctly annotated and having a dimension and composition such that it can be a representative sample of the population in analysis. Projects like the human phenotype ontology[64] aim to standardize the medical annotation of the biological data to facilitate the integration of data from different sources. For what concerns the data availability, large projects like The Cancer Genome Atlas (TCGA)[65], the Personal Genome Project[66] and the Human Protein Atlas[67] gives access to large collections of standardized omic data, but still small compared to the huge amount of data generated by hospitals and research centers every year. Ethical and legal issues are intrinsically linked with patient data: is it safe to share patient data? To which extent an individual is aware of

the risk and benefit of sharing his medical record and biological data? Is it possible to efficiently anonymize those types of data without a drastic loss of information?
An international effort of the bureaucratic bodies is required to face those questions, together with the instauration of clear and efficient communication between the scientific community and the general population to raise the interest about the possible benefits and problems that this type of data sharing could bring in everyday life.

RNA quantification: a technological breakthrough

Technological advances allow novel definitions of basic concepts, such as the one of life: “Life is an organized matter that provides genetic information metabolism” [68] given by Tetz in 2019. The author defines genetic information metabolism as “functioning, reproduction, and creation of genes and their distribution among the living and non-living carriers of genetic information”.

Many definitions of life have been given throughout history, some of which focus the attention more on the physical properties, as in Schrodinger’s book “What is Life?”[69], and some on the biological properties, like the notorious seven pillars of life[70].
In 1944, before the discovery of the DNA as the carrier of information[71], Schrodinger defined life as a partially closed environment that, thanks to the genetic information stored in an "aperiodic crystal" under the form of covalent chemical bonds, can maintain a low internal entropy increasing the environmental one.
Fifty-eight years later, Koshland proposed seven principles that define any living system. The genetic information, called the program, is the first pillar and is defined as the organized way to handle the system components and their interactions.
Most of those evergreen definitions emerged from the enthusiasm led by the possibility to study, analyse and quantify different biological properties, but the content of information is a characteristic present in all of them, even when its physical carrier was still unknown.
In the case of Tetz’s definition of life, the trigger is the high throughput sequencing technology that allows to easily read the genetic information, opening the gates for its decryption.
In the following paragraphs, we’ll go through three generations of sequencing technologies that had an impact not only in the research field but also in everyday life. Despite several detailed reviews written by the main characters that contribute to this fascinating journey[72,73], it’s important to remember the milestones that drove us where we are.

First-generation sequencing

The story of sequencing flows in the opposite direction with respect to the flow of information: the first biological sequence decipher was the amino acid sequence of the insulin protein in 1951[74], followed in 1965 by the first RNA sequence (alanine tRNA), which required five people working three years with one gram of pure material isolated from 140 kg of yeast to determine 76 nucleotides[75].
The processes to sequence those two classes of molecules were similar: fragmentation of the polymer followed by separation by chromatography and electrophoresis, then deciphering of the individual fragments by sequential exonuclease digestion and finally the sequence was deduced from overlaps.
The first successful sequencing of a DNA molecule was published in 1968 by We and Kaiser: they measured the incorporation of radiolabeled nucleotides by *Escherichia Coli* Polymerase in reactions that extended the 3' end to fill in the complementary cohesive end sequences of a phage lambda DNA of only 12 nucleotides[76–78].
The cohesive portion was necessary for the polymerase to start the synthesis of the complementary strand.
Copying the lactose-repressor binding site of *E.Coli* into RNA allowed its sequencing by Gilbert and Maxam: 24 bases in two years[79].
Thanks to the discovery of type II restriction enzymes by H.Smith[80,81], it was possible to generate short fragments from large molecules of DNA having ends that could function as primers, starting points for the polymerase reaction.

Also known as Sanger sequencing, the method that allowed the determination of the sequence of long fragments of any DNA molecule was published in 1975 and fine-tuned in the following years[9,82–85].

Sanger's method involves four extensions of a labelled primer by DNA polymerase, each with trace amounts of one chain-terminating dideoxynucleotides (dNTPs), to produce fragments of different lengths. The sizes of fragments present in each base-specific reaction were measured by electrophoresis on polyacrylamide slab gels, which enabled the separation of the DNA fragments by size with single-base resolution. The gels, with one lane per base, were put onto X-ray film, producing a ladder image from which the sequence could be read off immediately, going up the four lanes by size to infer the order of bases.
Notably, Maxam and Gilbert developed during the same period a similar method that, instead of dNTPs, took a terminally labelled DNA restriction fragment and, in four reactions, used chemicals to create base-specific partial cleavages[86].
The application of Sanger sequencing was dominant and it was enhanced when Messing and collaborators published a method for cloning into the single-stranded phage M13[85], the shotgun sequencing: any fragment of DNA can be inserted into a specific location in the phage genome (bacterial artificial chromosomes, BAC), allowing primers designed on the known vector sequence to amplify the insert.

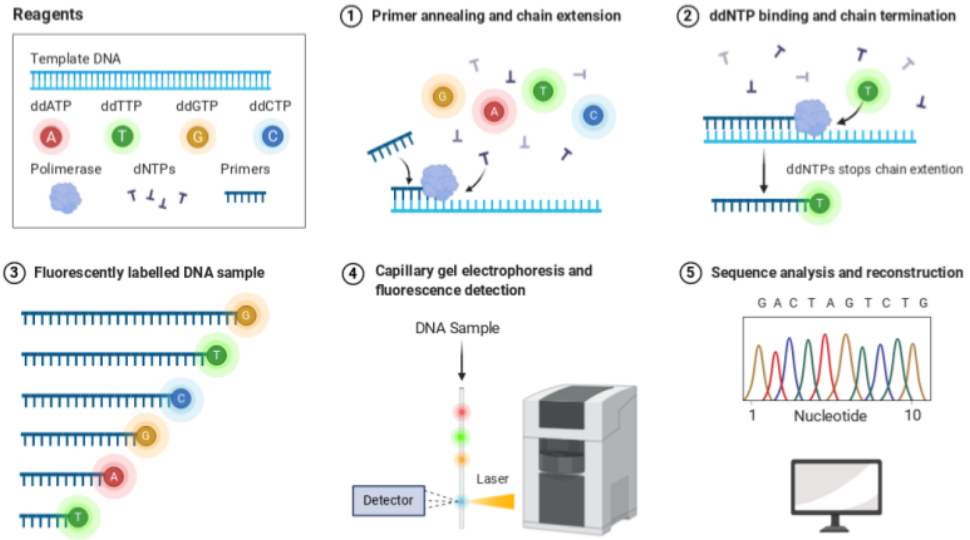


Figure 6: representation of Sanger sequencing. The reaction uses normal deoxynucleoside triphosphates (dNTPs) and modified dideoxynucleoside triphosphates (ddNTPs) for strand elongation. The ddNTPs are chemically altered with a fluorescent label and with a chemical group that inhibits phosphodiester bond formation, causing DNA polymerase to stop DNA extension whenever a ddNTP is incorporated. The resulting DNA fragments are subjected to capillary electrophoresis, where the fragments flow through a gel-like matrix at different speeds according to their size. Each of the four modified ddNTPs carries a distinct fluorescent label. The emitted fluorescence signal from each excited fluorescent dye determines the identity of the nucleotide in the original DNA template.

By 1987, the company Applied Biosystems developed automated fluorescence-based Sanger sequencing machines, shown in figure 6, able to generate around 1,000 bases per day[87], a number that reached 10 million bases per day by 2001 in a small number of academic genome centres thanks to additional technical improvements.

The application of this technology span from *de novo* genome assembly, such as the human genome project (HGP) of which the first draft was published in 2001[10,11], evolutionary biology, to determine organism phylogenies or the evolution of genes[88–91], clinical, as the detection of pathogens or testing for genomic mutations in congenital pathologies[92–94], to forensic identification and paternity testing, thanks to DNA fingerprinting[95–97].

Second-generation sequencing

By 2004, Sanger automated instruments allowed to sequence 600-700 bp at cost of 1 dollar per read, but the technology reached a plateau in its evolution since additional improvements had little or marginal effects[73]. Luckily, several groups throughout the 80s and 90s explored alternative sequencing methods and, after the HGP, those efforts paid off: next-generation sequencing (NGS) methods were becoming more and more competitive and were destined to take over the Sanger sequencing.

Also known as high throughput sequencing, the common denominator between those novel methods are the multiplexing and the sequence by synthesis (SBS) strategies.

Multiplexing replaced the one tube per reaction approach: a complex library of DNA templates is densely immobilized onto a chemically treated surface, with all templates accessible to a single reagent volume, allowing large numbers of libraries, that could be created also from different samples, to be pooled and sequenced simultaneously during a single sequencing run.

This step could be coupled by *in vitro* amplification: the most famous is the bridge amplification, a process that amplifies a complex template library with primers immobilized on a surface, such that copies of each template remain tightly clustered[98–100]. Other techniques that allow to amplify *in vitro* the input DNA are clonal PCR in emulsion, such that copies of each template are immobilized on beads[101,102], and rolling circle amplification in solution to generate clonal 'nanoballs'[103], followed in both cases by arraying on a surface for sequencing.

Finally, the SBS evolved in three main strategies:

1. The first system available was the pyrosequencing, used by the Roche 454 instruments, which consists in the detection of the light generated by a firefly luciferase, that use as substrate the pyrophosphate released by the incorporation of each dNTP, in a discrete step-wise manner[104]. This technology is no longer being maintained since 2013. A similar approach detects the incorporation of hydrogen ions released during the polymerization of DNA, used by Ion Torrent.
2. A second approach uses the specificity of DNA ligases to attach fluorescent oligonucleotides to templates in a sequence-dependent manner, used by SOLiD[8]. This approach generates reads shorter than the competitor's and has issues with palindromic regions[106].
3. The approach that became dominant since 2015 is Solexa that consists in a stepwise, polymerase-mediated incorporation of fluorescently labelled dNTPs. The development of engineered polymerase, reversibly terminating and reversibly fluorescent dNTPs are the keys that allow the incorporation of a single nucleotide in each cycle. After that the fluorescent colours are detected by imaging, the blocking and fluorescent groups are removed to set up the next extension[107,108].

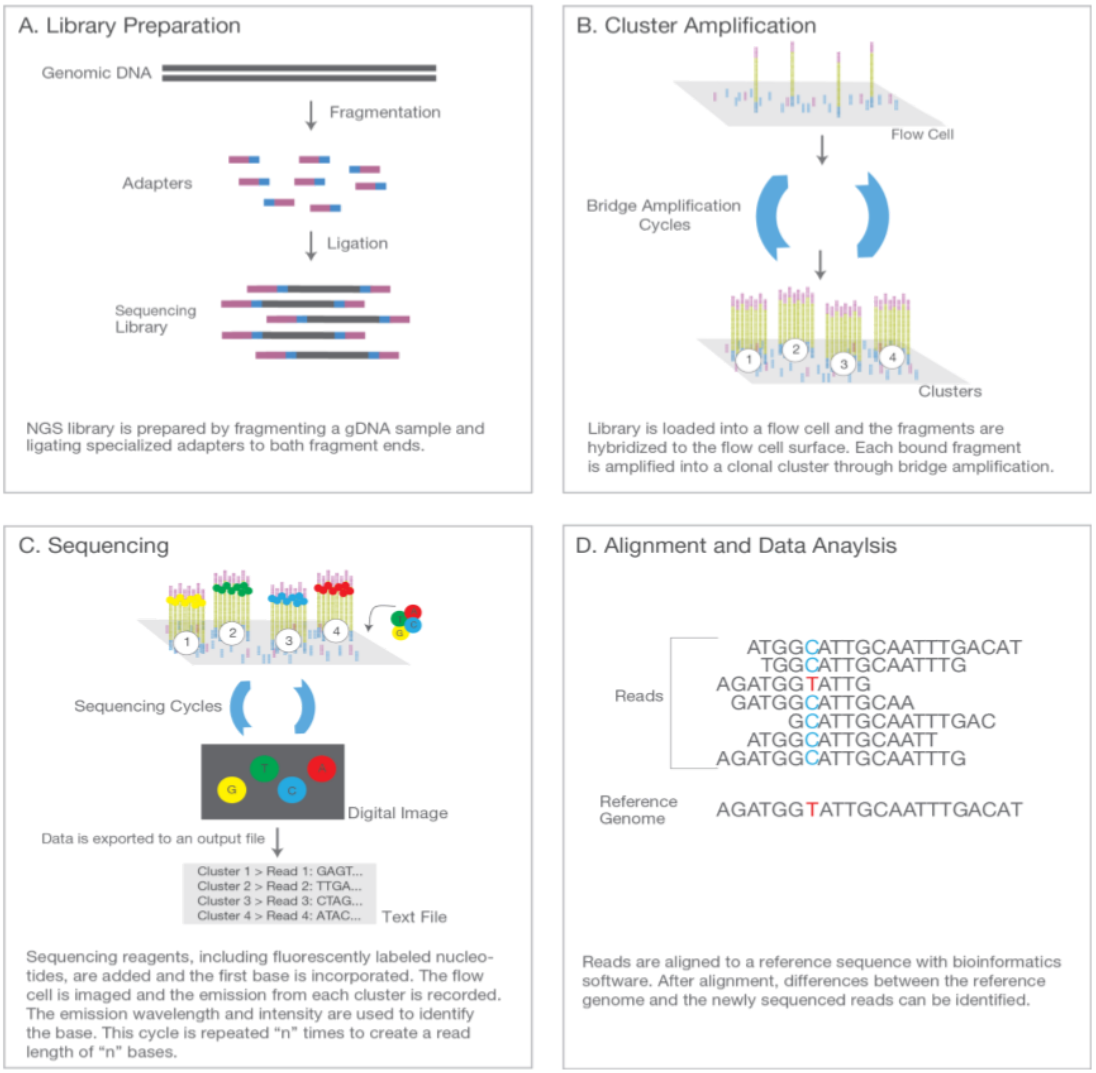


Figure 7: Illumina sequencing workflow.
From https://www.illumina.com/

Unlike Sanger sequencing, where Applied Biosystems had the monopoly, NGS technologies gave birth to several companies, competing in terms of cost, accuracy and read length. Few examples are the 454 and Solexa technologies, acquired respectively by Roche and Illumina; Agencourt (Applied Biosystems); SOLiD (ABI); Helicos (Quake), Complete Genomics (Drmanac) and Ion Torrent (Rothberg).
Those companies invested large capitals in several different approaches, allowing a fast growth of the field and a democratization of the sequencing capacity: individual laboratories could instantly access a vast catalogue of new methods, results, genomes and services.
Between 2007 and 2012, the raw per-base cost decreased by four orders of magnitude[109], keeping an accuracy of over 99.9%, though the length of each read is still shorter than Sanger sequencing.
In comparison with the first sentence of the paragraph, a single graduate student can generate over a billion independent reads, roughly a terabase of sequence, on one instrument for a few thousand dollars in a couple of days.

Reduction in costs and increased accessibility allowed NGS to be applied in a wide spectrum of fields: genome resequencing, i.e. mapping sequence reads to a reference genome to identify genetic variants; non-invasive prenatal testing, cancer molecular classification and Mendelian disease diagnosis are just a few examples of the many clinical applications that became feasible routines in hospitals.
Furthermore, *de novo* assemblies increased vastly, thanks to new assembly algorithms based on de Bruijn graphs that partially overcome the length issue[110,111]: with NGS many short reads generated from repetitive elements have only a single or no base difference, leading to ambiguous connections in the assembly. Instead of finding overlaps between reads, the EULER assembler [110] was the first to use a different representation of the data: the de Bruijn graphs. The method is organized around words of k nucleotides, the κ-mers, and the reads are mapped as paths through the graph. This data structure naturally handles the high redundancy without affecting the number of nodes: each repeat is present only once in the graph with explicit links to the different start and end points.

Several methods have been derived from the standard DNA-seq to quantify different molecules and events. Some examples are the RNA sequencing that, making use of the reverse transcriptase and dedicated protocols, replaced almost completely the microarray technology for gene expression quantification and allowed researchers to unveil the RNA world that we took into consideration in the previous chapter[112]; the ChIP-Seq, a method used to quantify the protein-DNA interactions[113]; the Bisulfite sequencing, that used to determine the DNA methylation patterns[114]; and the since cell RNA-seq, one of the many adaptations of the RNA-seq technique that allows to sequence the sparse transcriptome of individual cells.
An important approach is the paired-end sequencing that allows to sequence both ends of a single biological fragment, generating more accurate read alignment and the ability to detect insertion-deletion (indel) variants[115].

Third-generation sequencing

The second-generation sequencing has two important limitations: the short length of the reads, reaching nowadays a maximum of 300 bp[116], and the PCR amplification step. The first issue has repercussions on *de novo* assemblies of repetitive regions and on the determination of the single-molecule RNA isoforms, the second add time and complexity in the library preparation, loss of information, such as the lack of information about eventual nucleotide modifications, and the introduction of copying errors and sequence-dependent biases.
Due to those limitations, only recently the telomere to telomere (T-2-T) consortium was able to complete the assembly of the full human genome, including the constitutive heterochromatin regions, thanks to the combination of Illumina

sequencing and a new generation of sequencing technology: third-generation sequencing[117,118].

A parallel research field, started back in the 1980s, aimed to sequence single molecules in real-time (SMRT) and gave birth to two promising approaches: PacBio and Nanopore sequencing.

Initiated by Webb and Craighead and developed by Korlach, Turner and Pacific Bioscience, PacBio is the first approach capturing sequence information during the replication process of a single DNA molecule and was released in 2011[119,120].

The template is a closed, single-stranded circular DNA that is loaded into a chip divided into a sequencing unit, the zero-mode waveguide (ZMW), a hole less than half the wavelength of light that provides the smallest available volume for light detection. In each ZMW, a single engineered polymerase is immobilized at the bottom and can bind to the circular DNA to start the replication. Four different fluorescent-labelled nucleotides are incorporated by the polymerase, generating distinct emission spectrums that are recorded in a temporal sequence, the continuous long read (CLR). A base-calling software analyses the CLR and estimates the sequence based on the light-pulse spectrum.

Each strand can be sequenced multiple times, allowing the generation of multiple subreads, whose consensus increases the accuracy of the technique, going from a median error of 11% for a single pass to 1% with four passes and 0.1% with nine[121,122]. The errors consist of more indels than mismatch and are distributed randomly, a factor that allows reducing efficiently the error rate increasing the CLR depth.

Base-calling can also detect nucleotide modifications, such as N^6-methyladenine (m^6A) and n^4-methylcytosine (m^4C), analyzing the kinetic variation from the light-pulse of the temporal sequence[123].

PacBio's read length is limited by the longevity of the polymerase: with chemistry v3 released in 2018, the average RL is 30 kbp, spanning from 250bp to 50 kbp.

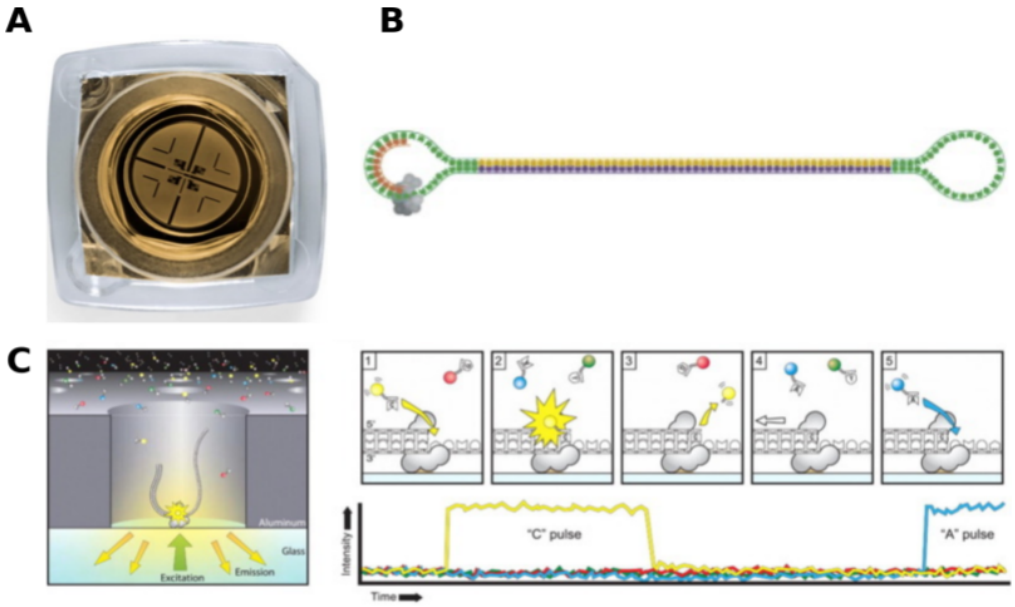


Figure 8: A) PacBio's SMRT cell, a chip containing 150'000 sequencing unit ZMW; **B)** the SMRT bell, a single-stranded circular DNA created by ligating hairpin adaptors to both ends of a target dsDNA; **C)** ZMW cell where the four fluorescent-labelled nucleotides are incorporated by the polymerase, generating distinct emission spectrums.
Adapted from: Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015 Oct;13(5):278-89. doi: 10.1016/j.gpb.2015.08.002. Epub 2015 Nov 2. PMID: 26542840; PMCID: PMC4678779.

A simple but revolutionary idea, hypothesized in the 1980s, is at the basis of the second methodology: considering a hole through which water is streaming, the passage of a body, let's say a tennis ball, modifies the flux of water in a different way a bowling ball would do. Detecting and decrypting the changes of the flux can tell the dimensions of the object that obstructed the channel, being able to discriminate between a bowling or a tennis ball. Similarly, but in a much smaller dimension, detecting the patterns in the flow of ions generated when an ssDNA passes through a narrow channel can be deciphered into the sequence of nucleotides that compose the polymer[124]. Two decades of work and technological advancement were required to move from this idea to the first successful nanopore prototype and the foundation in 2005 of the company Oxford Nanopore Technology (ONT).
ONT uses a voltage difference applied across electrolyte baths on either side of an insulated membrane to produce an ion current.
The current streams through a single channel protein, in the first chemistry versions *Mycobacterium smegmatis* porin A (MspA), pulling the DNA through the nanopore in a linear, head-to-tail fashion by electrophoresis. The passage would be too fast to be detectable, that's why another protein, called enzyme motor, acts as a molecular stop, preventing the DNA from travelling any further through the nanopore[124,125].

The signal detected is then analysed using bonito[126], a base caller that uses methods widely used in speech recognition problems (in particular a recurrent neural network (RNN) model trained using connectionist temporal classification (CTC) and conditional random field (CRF)) to decode the electric signal into a sequence of nucleotides.
ONT accuracy was less than 60%[127,128] when first introduced, but the base caller improvements over recent years allowed values of 85% in 2018 [129] and up to 98.3% in 2021 and promising a 99% with the chemistry version Q20+ [130,131].
If the British company accomplishes this goal, unthinkable a couple of years ago, nanopore technology would have all the characteristics to replace the dominant Illumina in the global market and more: the lack of an imaging step allows the production of cheaper and smaller devices, with the MinION device being as big as a smartphone and costing 1000 dollars; the read length can go from short to ultra-long read (more than 2Mb DNA and more than 20Kb RNA[132,133]); it allows real-time analysis[134,135] and the library preparation is quick, requiring only ten minutes, and standardized, thanks to an automated device that increase the reproducibility of the experiments; RNA molecules can be directly sequenced without needs of any cDNA intermediates, reducing the time, costs and introductions of errors.

Nanostring nCounter: direct RNA quantification

Sequence technologies are fundamental in research, but for clinical application most of the time it's sufficient to know the abundance of specific subsets of sequences, representing features like genes, specific isoforms, splicing junctions, chimeric transcripts and SNP.
An emerging technology that allows the direct quantification of RNA molecules using a simple and fast protocol is Nanostring Technologies nCounter[136]. The company Nanostring, founded in 2003 and settled in Seattle, offers a cost and time efficient technique to quantify specific sets of sequences[137].
This automated platform hybridizes fluorescent barcodes directly to specific nucleic acid sequences, allowing for the non amplified measurement of up to 800 targets within one sample and to multiplex up to 96 samples in the same run[136].
Nanostring's nCounter has been used within different clinical and research applications, such as assays to predict recurrence for gastric cancer after surgery[138], subtype molecular classification of diffuse large B-cell lymphoma[139], the identification of known oncogenic fusion genes in lung cancer[140] and many others[141–144].
The robustness, sensibility and high reproducibility of this technology outdo microarrays, the most similar technology available, which are often expensive and lack flexibility and reproducibility when evaluating low-quality RNA samples, such as those from formalin-fixed paraffin embedded[145].

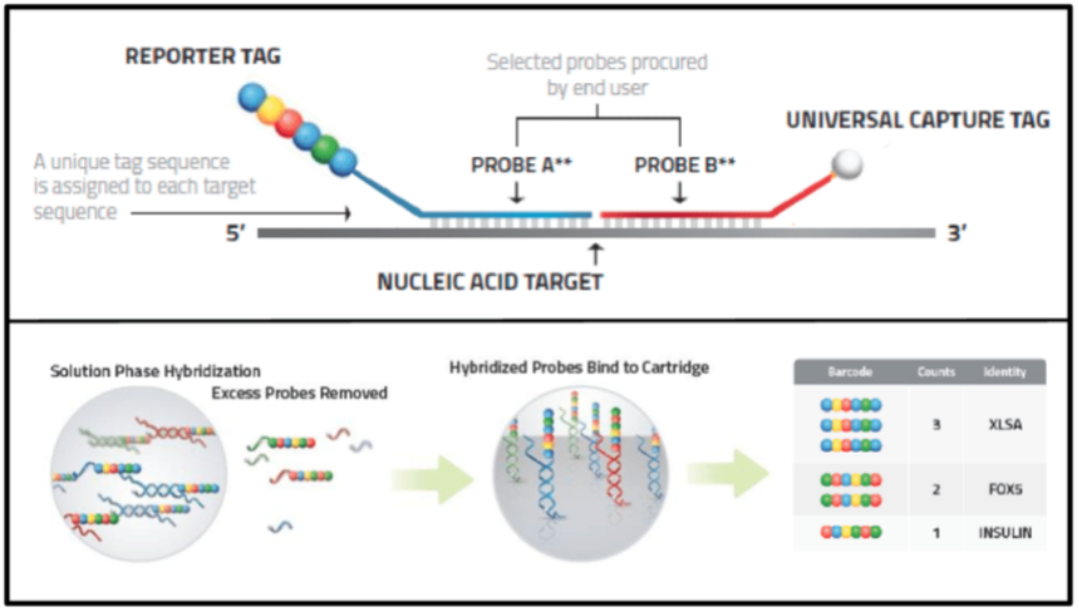


Figure 9: Nanostring nCounter principles: two probes, designed to target a specific nucleic acid, are ligated respectively to a unique reporter tag and an universal capture tag. A single run can contain up to 800 different probes analysing up to 96 samples.
From: Bobée, Victor. (2017). Détermination moléculaire des sous-types de lymphomes B diffus à grandes cellules par un classifieur de type retrotranscription multiplex ligation-dependent probe amplification.

RNA-seq data analysis

In the previous chapters, we considered the biological aspects of the information represented in the RNA-seq data and the evolution of the technologies that allow its extraction into a series of four-letter strings. Now, we'll focus on the common approaches used to exploit this information, starting from the experimental design, moving to read mapping, transcript quantification and concluding with differential gene analysis. Finally, we'll overview existing reference-free approaches that extrapolate and analyse the information using *κ-mer*s.

Experimental design

Similarly to other scientific experiments, RNA-seq requires a careful design of the data that has to be generated or collected. A study may be exploratory, with the objective of discovering future research tasks, or formal, with a hypothesis to test. An important factor to consider is whether the data comes from experiments, where the researcher has control of the variables in the study, or *ex post facto*, where the investigator cannot manipulate the variables, such as clinical data.
The sampling design must consider heterogeneous samples, representative of the population in analysis, and balance between case and control, randomizing the experimental units to treatment in order to reduce confounding factors.
Budget is one of the most limiting factors and it's determined by the number of samples processed and the number of reads generated for each of them, also called sequencing depth.
Tools like "Scotty"[146], "RNAseqPS"[147], "PROPER"[148] and "ssizeRNA"[149] estimates the optimal sample size required to achieve the desired statistical power and, although most of them diverge significantly in the results[150], can help the scientist in this crucial step.

The standard application for RNA-seq data is differential expression (DE) study of genes and, less frequently, of transcripts. In any experimental design, selecting the appropriate number of biological replicates is a trade-off between cost and precision. A misconception is that three replicates are enough in a DE study: Schurch et al. show that to identify differentially expressed genes having a low fold change it's necessary to have at least six replicates per condition and that using only three replicates per condition most of the DE analysis tools found only 20-40% of the significant DE genes [151].
Different applications require different sample dimensions: if we want, for example, to associate SNP to a particular phenotype, we need to apply genome-wide association study (GWAS) sample sizes, with a minimum of 100 samples up to more than 2000[152].
For what concerns the depth, since more than 80% of the reads are attributed to the 10% most expressed genes and increasing the number of reads only marginally increases the coverage of lowly expressed genes, especially over the 10 million

reads[153], it's better to use the budget to have more replicates rather than few samples with deep sequencing.
Finally, when the experiment has to be run in multiple batches, it's important to equally distribute the conditions between the batches. Processing groups of samples on different days, using different machines and by different operators can reflect on small variances between the batches that can be misinterpreted as biological signals.

RNA-seq read alignment

The final output of any sequencer is generally a FASTQ file, in which a read is represented by four parts[154]: the first is the header, starting with a '@' character and including a unique ID attributed to the read, useful especially in paired-end sequencing to identify the two mates; the second part contains the raw sequence, usually encoded using the standard IUPAC single letter codes for DNA and RNA; the third part, starting with a '+' character, can contain additional description but is usually empty. Finally, the last part encodes the quality values for each nucleotide. Since v1.8 Illumina sequencers use the same quality score as the Sanger and PacBio sequencer: the Phred quality score (Q_{phread}), that is the $-\log_{10}$ of the probability that the corresponding base call is incorrect.
Millions to billions of short cDNA reads contain information about what RNA molecules are in the original sample, their abundance and sequences. This information is randomly scattered across the reads: subsequential reads in the FASTQ file can represent completely different RNA molecules.
To quantify the abundance of the RNA-molecule at the transcript level, considering each isoform as an independent entity, or at the gene level, where the expression of a gene is the sum of the expression of its isoforms, it's necessary to align the reads to a reference genome or transcriptome.
In organisms for which only a *de novo* transcriptome is available, or it's much better characterized than the reference genome, unspliced alignment is a feasible solution. Mapping on a reference transcriptome, however, induces a high degree of multi-mapping since different isoforms can share the same intron and isn't flexible enough to deal with novel splicing or expression patterns. Pseudo-alignment and fast mapping to transcriptome is part of the strategy used by recent transcript abundance estimators and we'll focus on this subject in the next chapter.

Mapping the reads to a reference genome presents the main challenge to correctly align the read that includes a splice junction (SJ). Bowtie[155], STAR[156], HISTAT[157] and GMAP[158] are the most famous of a long list of splice-aware aligners that use known or empirically deduced SJ sites to guide the alignment. Each software uses a different approach, resulting not only in different performances in terms of time but also in terms of the final result, adding a layer of variability to the experiment.

Once mapped, the reads are stored in dedicated files, such as BAM, the binary and compressed version of the SAM format (Sequence Alignment Map), and CRAM, a reference-based storage format promoted by EBI from 40 to 50% smaller than the BAM one[159].
The problem of the lack of an international standard is being addressed by the moving picture expert group (MPEG), mostly known for the audio and video coding, who released the first version of the MPEG-G in 2019 proposing a new file format: "The standard will offer high levels of compression, approximately 100 times compared to raw data, i.e. more than one order of magnitude than possible with currently used formats. Furthermore, the MPEG-G standard will provide new functionalities such as native support for selective access, data protection mechanisms, flexible storage and streaming capabilities. This will enable various new applications scenarios, such as real-time streaming of data from a sequencing machine to remote analysis centres during the sequencing and alignment processes."[160]

Gene and transcript level quantification

Essential for most of the downstream analysis, assessing the gene and transcript level abundances is also characterized by a long list of tools that achieve the same goal using different strategies and with different performances.
HTSeq, featureCounts, the built-in STAR option and other tools count directly the fragment overlapping the gene features after the mapping step, differing one from the other by the way they handle certain conditions, like multi mapping fragments, fragments that map to multiple features and fragments mapping partially in the feature. This approach is limited by changes in the composition of the exons that do not directly impact the gene-level read count, such as isoform switching.
To overcome those obstacles, transcript-level quantification is getting more and more used, even to estimate the gene-level expression with better performances on the downstream analysis[161]. It's worth mentioning that, in contrast with the transcripts, the gene is not a physical entity but it's a useful abstraction having no clear target for quantification.
Methods like RSEM[162] and Cufflinks[163] define a generative model of RNA-seq reads and use such a model to infer the transcript abundance, assigning in a probabilistic way the ambiguous fragments to the different isoforms. Recent approaches make use of pseudo-alignments of *κ-mer*s to speed up the process, bypass the alignment step and produce an accurate estimation [164,165].
The pseudo alignment procedure uses the reference transcriptome in the form of the de Bruijn graph to assign a read to a set of transcripts without alignment.

In particular, Kallisto uses the transcriptome in the form of de Bruijn graph to assign through expectation-maximization (EM) algorithm the read to the transcript from which most likely was generated[166].
Another widely used *κ-mer* based tool is Salmon[167]: it first builds a sample-specific bias model to correct effects like fragment GC-content bias; after that, it uses a lightweight mapping procedure called quasi-mapping, similar to pseudo alignment in the use of the transcriptome and *κ-mer*s.
This strategy, proposed by Srivastava A. et al with RapMap[168] was first applied to Sailfish[169], has the same outcomes as the transcriptome pseudo alignment using a different data structure[168].
Easy to use, fast, with low computational requirements and high performances in terms of speed and disk usage, the mapping-free *κ-mer* based approaches have become popular for assessing transcript and gene-level abundance, gaining the top tiers of the most recent benchmark studies [170–172].
Finally, traditional gene and transcript quantification tools don't consider repetitive and transposable elements. Dedicated softwares, like TEtranscripts[173], telescope[174] and SalmonTE, address this problem, applying similar approaches like the ones used for classical genes to transposable element families.

Alternative splicing signatures

Transcript and gene abundances are not the only quantifiable features that can be inferred from RNA-sequencing: the percentage of the splice in (PSI) is used in splicing studies to quantify the frequency of inclusion of specific exons.
Tools like MISO[175], rMATS[176] and Whippet[177] compute one PSI value for each exon using the following formula:

$$PSI = 100 \times \frac{a+b}{a+b+2c}$$

Where a and b represent the reads overlapping the splice junctions that support the inclusion of the alternative exon to downstream and upstream constitutive exons and c represents the ones that support the exclusion.

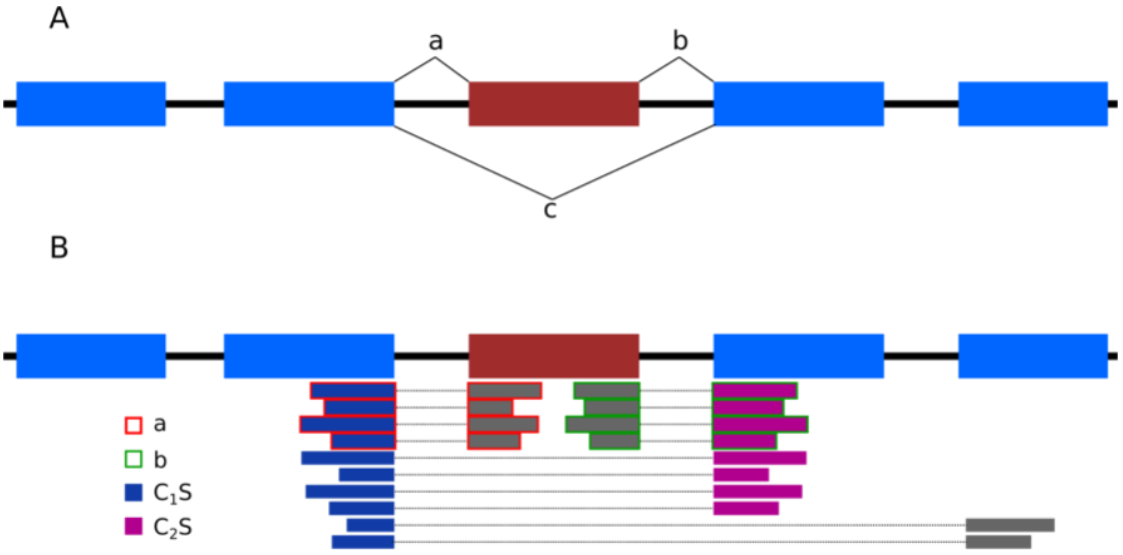


Figure 10: Representation of a cassette exon alternatively spliced, in red, and the constitutive exons, in blue. **A)** splices junctions used for the PSI evaluation in rMATS, MISO and Whippet. **B)** splices junctions used for the PSI evaluation in PSI-sigma.

PSI-Sigma uses a different PSI value:

$$PSI_{sigma} = 100 \times \frac{a+b}{\sum_{i=1}^{n} C_1 S_i + \sum_{j=1}^{m} C_2 S_j}$$

Where S_i and S_j are the splice-junction reads of all isoforms in the region between two constitutive exons C_1 and C_2, generating multiple PSI in case an exon is used by different isoforms[178].

MAJIQ[179] quantifies the PSI values for each isoform as well, using a combination of read rate modelling, Bayesian PSI modelling and bootstrapping.

Finally, SUPPA2[180] uses the transcript abundances to infer the PSI and delta PSI (ΔPSI, difference in PSI between two conditions). This approach, though fast, produces suboptimal results[178].

All the aforementioned methods can be applied to second and third generation sequencing since they take as input FASTA, FASTQ or BAM file formats, but only PSI-Sigma was tested using long reads[178] showing a more complete and precise transcriptome profile.

Among the possible alternative splicing events, intron retention (IR) requires additional adjustments in order to be correctly quantified: MAJIQ, for example, filters the events having consecutive windows across the intron lower than a user definable threshold; PSI-Sigma estimates the abundance of the IR isoform counting the number of intronic reads crossing the first, 25th, 50th, 75th and 99th percentile positions of an intron. Without a proper approach, unannotated alternative donor or

acceptor splicing sites and overlapping transcripts could lead to misclassified events. Furthermore, introns that are enriched in low-complexity and repetitive sequences may restrict the unique mapping of sequencing data[181].
IRFinder, the first software dedicated to IR analysis, addressed this problem proposing a new metric, the IRratio, defined as:

$$IRratio = \frac{I_a}{I_a + E_a}$$

Where: I_a is the intronic abundance, estimated as the median depth of the intron excluding low mappability regions and regions overlapping with other features; E_a is the abundance of the flanking exon, estimated as the maximum number of reads that map the 5' or the 3' flanking exon splice site[22,182].
In 2015, Bai et al developed IRcall, a ranking strategy, and IRClassifier, a random forest classifier, to detect IR events[183]. The first generates a joint score for IR events, based on intron read counts, flanking exon read counts and splice junctions. The latter uses 21 features extracted from other IR detection methods (IRFinder, MATS and ExpressionPlot) to build a Random Forest classifier to predict IR events.
Finally, iREAD[184] uses the Shannon entropy[185] to quantify the uniformity of the distribution of reads across the intron and considers only introns that don't overlap with any other feature. Due to the lack of experimentally validated intron retention events dataset availability, iREAD authors compared their tool with IRFinder using simulated reads. In this benchmark, IRFinder achieved a precision of 0.73, iREAD of 0.99 and similar time performances. However, few considerations are due to contextualize those results:

1. iREAD excludes all the introns overlapping known IR events or any other features, considering so far a much smaller set of events, meanwhile IRFinder includes the known IR events and masks the intronic annotations, such as miRNA and antisense transcripts. In the evaluation of the metric, the authors considered the same number of events to evaluate the performances of the two software, using the minimum number of hits found in the two methods. In the iREAD manuscripts, it's not specified which method outputs the limiting number of events, but for the aforementioned reasons, lots of IRFinder positive results have not been taken into consideration.
2. The results are further biased by the criteria chosen to generate IR events in the simulated RNA-seq data, which are the same criteria iREAD uses to identify IR events. For example, an intron is considered retained if it has at least 10 reads and one junction read that spans the exon-intron junction, regardless of the length of the intron and the number of exon-exon splice junctions, factors considered in IRFinder. Furthermore, IR events in isoforms having known intron retention are not considered as IR by their gold standard sets and iREAD algorithm but are generated anyway, increasing the number of events considered false positive in IRFinder's results.
3. The simulated data doesn't take into account the possible presence of intronic reads originated by unannotated intronic transcripts, which might affect the global performances of the softwares.

4. The speed comparison is biased: the authors use a machine with 20 cores and don't specify that IRFinder is a single-core process, while iREAD can multithread the process, leading to unfair comparison. Finally, IRFinder is optimized to have a low RAM footprint, an aspect that has not been considered in the benchmark.

Differential analysis

Most of the experimental designs aim to identify differences in expression between two or more conditions, one used as control and the other as a target. With this objective, differential expression (DE) analysis formulates and tests a statistical hypothesis for each feature across the samples.
Usually, only a limited number of replicates are available (3-5 replicates per condition) and, combined with the large number of features that are tested simultaneously, the achievable statistical power would be very small without dedicated strategies implemented and refined during the years by the statistical community. Most of those approaches, such as the widely used limma-voom[186], were initially developed for microarray data and in a second place adapted to RNA-sequencing.

The general workflow includes:

1. **Preprocessing**: encompasses the filtering of uninformative features, such as lowly expressed genes. Bourgon et al. showed that filtering independent of the test statistic achieves a higher detection power[187]. To facilitate across-sample comparison, the counts have to be directly normalized or, in software like DESeq2[188], scaling factors have to be computed to accompany the analysis. In addition, few highly expressed genes can drive the sampling of fragments, leading to inaccurate scaling measures. Calculating sample-wise size factor can in part overcome this issue: this procedure consists in generating a pseudo-reference sample, derived from the averages of each gene across all the samples in the study; for each sample, compute the ratio between the sample gene count and the pseudo-reference one and use the median value as size factor, by which the raw count is divided to obtain the normalized values. It can be considered a robust global fold change between the current sample and an ideal reference sample, derived from all the samples[189,190].
2. **Statistical model specification and estimation of its parameters**: due to the small sample size with respect to the number of features, DE tools mainly implement parametric methods. The variability in gene expression across technical replicates follows a Poisson distribution[189], for which the variance is equal to the mean:

$E(Y_{fi}) = \mu_{fi} = Var(Y_{fi})$

Where Y_{fi} is the observed count for class *i* and feature *f* and μ_{fi}its mean.

The biological replication introduces additional variability between the samples, approximately following an extension of the Poisson distribution: the gamma-Poisson (or negative binomial NB) distribution, that presents an additional dispersion parameter and a quadratic mean-variance relationship:

$Y_{fi} \sim NB(\mu_{fi}, \varphi_f)$

$Var(Y_{fi}) = \mu_{fi} + \varphi_f \mu^2_{fi}$

Where $\varphi_f$is the dispersion associated with the feature *f*. The limited number of samples is again a problem for a reliable estimation of φ_f. Different approaches have been developed to solve this issue, whose details go beyond the scope of this introduction[191]. Finally, the generalized linear model (GLM) framework, an extension of classical linear models to non-Gaussian responses, allows the inclusion of multiple treatments or covariates to the study[192]. The NB GLM model can be formulated as:

$log(\mu_{fi}) = \eta_{fi} = X_i \beta_f + log(s_i)$

Where $\eta_{fi}$is the linear predictor, $X_i$is the design matrix, β_f represents the regression parameters and $s_i$is the normalization scaling factor. β_f can be fitted using standard GLMs algorithms and the estimated dispersion values φ_f .

3. **Statistical inference**: for each feature, fitted the GLM, it's now the time to test the null hypothesis H_0 that there is no DE between conditions, generally that the log-fold-change (LFC) is zero, against the alternative hypothesis H_1 that the $LFC \neq 0$. The LFC can be represented as *L*, a single regression parameter (vector) or a linear combination of parameters (matrix) in the GLM framework as:

   $H_0: LFC = L\beta_f = 0$

   There are several tests available for GLM, such as the likelihood ratio tests (LRTs), implemented in edgeR[193], that compare the likelihood of a full model with the likelihood of a reduced model, where one or some of the parameters are constrained according to H_0. DESeq2, besides LRTs, implements also the Wald test, a faster approach that achieves approximately the same results as the LRT[188] , assuming a symmetric likelihood distribution and asserting the significance of the relation between the independent variable and the outcome within the logistic model.
4. **Adjustment for multiple testing:** to avoid excess false positives, the p-values obtained from the statistical inference must be corrected for multiple testing. Family wise error rate corrections, such as the Bonferroni correction, are usually too stringent for DE analysis, where a small proportion of false

positive (FP) can be tolerated to obtain a large number of true positive (TP). The false discovery rate (FDR) controlling procedure is widely used to control the expected fraction of false positives in the detected set of features. One example is the Benjamini–Hochberg (BH) procedure[194,195], which has become a common practice in high-dimensional data analysis thanks to its simplicity and solid theoretical justification, accepted from both frequentist and Bayesian perspectives[196]. The BH adjusted p-value, p^{adj}, is computed ranking in ascending order the p-values and applying the following formula:

$$p^{adj}_{j} = \frac{p_j \times m}{j}$$

where p_j is the p-value of the jth test and *m* is the total number of tests.

Though this pipeline is optimized for gene DE analysis, it can also support transcript level DE analysis to detect differential transcript expression (DTE).
Another type of analysis considers the change in the relative abundance of the isoform for a specific gene, called differential transcript usage (DTU), and of the individual exons, called differential exon usage (DEU).
Tools like DEXSeq[197], DRIMSeq[198] and BayesDRIMSeq[199] are specialized in this type of analysis, adopting different strategies whose description goes beyond the scope of this introduction.
To discover alternative splicing events between conditions, the difference of the PSIs between is used.
rMATS[176] uses likelihood ratio tests (LRTs), the same used in DGE analysis, while SUPPA2[180] test is based on comparing the observed difference in PSIs across conditions to the empirical cumulative density function of the within-replicates differences of PSIs of splice junctions from similarly expressed transcripts.
Finally, the differential IR analysis in IRFinder is performed using an Audic and Claverie Test[200], in case of a single replicate for each condition, or a GLM model, using a wrapper of DESeq2, fitted with the intron and exon abundances of each sample.

κ-mer based approaches

Quantifying the abundance of known transcripts or splicing events is not the only way to obtain meaningful features: counting the *κ-mer*s occurrences in the raw sequencing data is another approach widely used in different fields, such as metagenomics, *de novo* assemblies and phylogeny.
This kind of representation has the advantages of being reference-free since to count the *κ-mer*s occurrences is independent of any reference genome, transcriptome or annotation.

The drawback is that it's highly redundant and with high dimensionality: each transcript of length *L* will generate *L*-*k*+1 κ-mers and, globally, there are 4^k possible combinations of the four nucleotides in a string of length *k*.
The length of the *κ-mer*s is chosen according to the dimension and complexity of the genome of interest: the bigger and more complex the reference is, the longer the *κ-mer* needs to be in order to have a sufficiently high fraction of uniquely mapping *κ-mer*s.
Computationally, values close to a multiple of 8 (the number of bits in a byte) are efficient values to be represented in binary form, where for example A can be represented as 00, C as 01, G as 10 and T as 11.
Odd numbers are preferred to avoid reverse palindromic sequences: the central nucleotide won't ever be complementary of itself. Additionally, some tools use the final bit to represent the original strand of the *κ-mer*. Altogether, a common formula to select the *κ-mer* size is:

$$k = (8 \times d) - 1$$

Where *d* is an integer arbitrarily chosen to have a good tradeoff between the *κ-mer* precision, representing the proportion of *κ-mer*s mapping uniquely on a reference genome, and the tractable number of possible combinations. In human studies, for example, *d* is set to 4, resulting in a *k* equal to 31.
To compare organisms with smaller genomes, dedicated tools like KITSUNE can be used to determine the optimal *k*[201,202].
For genomic applications, "canonical" *κ-mer* representation is usually used to reduce the total number of *κ-mer*s and have a unique representation of the DNA sequence. The term canonical indicates the aggregation of the counts of a *κ-mer* and its reverse complementary to one of the two comings first using a relation order, generally the lexicographic one[203].
The counting procedure, though simple, presents computational challenges for what concerns the time and space requirements. A recent benchmark of S.C. Manekar [204] compared ten famous *κ-mer* counters, where KMC3, DSK and Gerbil showed the best performances. Among the three, DSK[205] is optimal in case of low RAM availability, thanks to its algorithm design that subdivide efficiently hash tables into multiple files on the hard disk; Gerbil[206] is optimal in case a GPU is available, being the only one supporting this type of processor able to massively parallelize procedures; finally, KMC3[207] presents the best tradeoff between time and resources, it's stable and offers a convenient C++ library.
Once obtained the *κ-mer* counts, a common approach is to create de Bruijn graphs (dBG), a direct graph representing the *κ-mer*s as vertices and the overlap of length *κ*-1 between them as edges. A compressed representation of the dBG, the cDBG, is obtained by merging two adjacent simple nodes, which means nodes linked to at most two other nodes[208].
The application of graph theory to de Bruijn *κ-mer* graphs is one of the keys to the success of this methodology: this representation is efficiently handled by the machine and there are a large number of algorithms for searching, traversing, finding paths and representing its properties. It's important to mention that most of the *κ-mer*

dBGs are not complete but just subgraphs since not all the 4^{κ} possible vertices are represented and therefore not all the dBG properties and algorithms can be applied. Due to the high dimensionality of these graphs, *κ-mer* representation is mostly used for small genome assemblies and comparisons[209–213].
For example, kover[214,215] implements a rule-based machine learning approach to identify *κ-mer*s of bacterial genomes that can be used as biomarkers for antibiotic resistance. CLARK[216] and KrakenUniq[217] are two tools able to classify metagenomes using unique *κ-mer*s found in different taxa.

In RNA sequencing experiments, *κ-mer*s are used not only to estimate the transcript abundances, such as with the already mentioned kallisto[166], but also to perform specific tasks, such as the HLA (Human Leukocyte Antigen) alleles profile[218], detect virus RNA in plants sequencing data[219], detect targeted and *de novo* variants[208,220–223], motif identification[224], identify fusion, noncoding and novel transcripts[225,226] and *de novo* transcriptome assembly [227].

For what concerns the differential analysis, there are few methods available that use *κ-mer*s to identify biological markers: KISSPLICE[208], HAWK[228,229] and DE-kupl[230]. KISSPLICE is a software initially designed to find alternative splicing events from RNA-seq data, but which also outputs indels and SNPs. Those events correspond to recognisable patterns, called bubbles, in a de Bruijn graph.

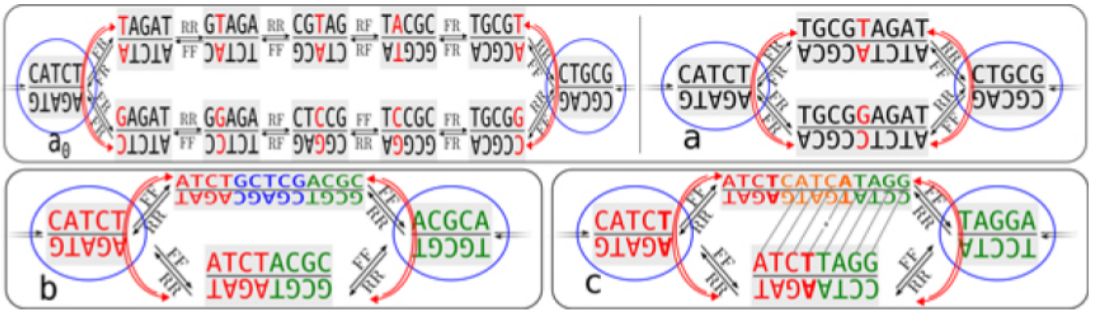


Figure 11: Part of non-compressed **(a_0)** and compressed **(a, b, c)** de Bruijn graphs (k = 5). Each node contains a word (upper text of each node) and its reverse complement (lower text of each node). In the uncompressed graph, the word is a k-mer. Encircled nodes are switching with respect to red paths (pointed out by red arrows). **(a_0, a)** Bubble due to a substitution (red letter). Starting from the forward strand in the leftmost (switching) node would generate the sequences CATCT A CGCAG (upper path) and CATCT C CGCAG (lower path). **(b)** Bubble due to the skipped exon GCTCG (blue sequence). This bubble is generated by the sequences CATCT ACGCA and CATCT GCTCG ACGCA. **(c)** Bubble due to an inexact tandem repeat. This bubble is generated by the sequences CATCT TAGGA and CATCT CATCA TAGGA, where CATC**T** CATC**A** is an inexact tandem repeat.
From: Sacomoto GA, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot MF, Peterlongo P, Lacroix V. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. BMC Bioinformatics. 2012 Apr 19;13 Suppl 6(Suppl 6):S5. doi: 10.1186/1471-2105-13-S6-S5. PMID: 22537044; PMCID: PMC3358658.

KissDE[231] performs a likelihood ratio test on the abundance of the alleles found using KissSplice and mapped to a reference genome using BLAT to identify condition-specific SNP.
Hitting association with *κ-mer*s, HAWK, is a method that aims to identify *κ-mer*s with counts that are statistically significant between two phenotypes in whole-genome sequencing reads, applying GWAS techniques such as the correction for population

stratification and other confounders. The approach consists in counting the *κ-mer*s in each sample using Jellyfish[203], test the differential expression using a Poisson distribution based likelihood ratio test, correcting for confounders and finally merging the *κ-mer* using ABySS, a notorious assembler for short reads. In contrast with other genomic classification tools, HAWK uses *κ-mer* counts and differential expression analysis, therefore it could be applied to RNA-seq data with the appropriate modifications in the assembly of the *κ-mer*s.
Finally, DE-kupl is the first tool to compare *κ-mer* abundances across two groups of human replicates, removing *κ-mer*s represented in the reference transcriptome and the ones considered noise due to low expression to identify differentially expressed events that are not represented in existing transcript catalogues. Each *κ-mer* is then tested using either a t-test or DESeq2, reducing the set of *κ-mer*s to only the ones considered differentially expressed between the two groups of samples. Finally, overlapping *κ-mer*s are merged in sequences that can be mapped on a reference genome to identify its biological meaning, such as differential splicing, polyadenylation, lincRNA, allele-specific expression, repeats and IR.
Importantly, in DE-kupl publication it is shown that in RNA-seq the sequence diversity from the reference genome and transcriptome is much bigger than in WGS, suggesting the existence of a significant amount of biological information n RNA-seq that cannot be accessed using reference-based approaches.

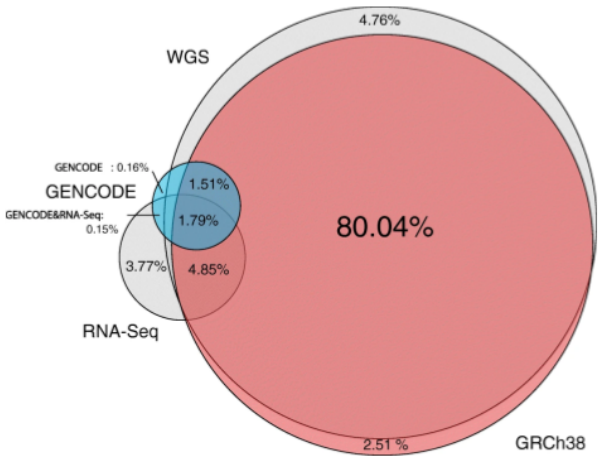


Figure 12: The diversity of non-reference *k*-mers is greater for RNA-seq than for WGS. Intersection of *k*-mers between GENCODE transcripts, the human genome (GRCh38), RNA-seq, and WGS data. RNA-seq and WGS data originate from the same lymphoblastoid cell line (HCC1395).
From Audoux J, Philippe N, Chikhi R, Salson M, Gallopin M, Gabriel M, Le Coz J, Drouineau E, Commes T, Gautheret D. DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. Genome Biol. 2017 Dec 28;18(1):243. doi: 10.1186/s13059-017-1372-2. PMID: 29284518; PMCID: PMC5747171.

Concluding, *κ-mer*s have a large potential as biomarkers: they are agnostic since their extraction from the raw data is independent of any reference genome or annotation; they are interpretable, since they can be mapped to a reference genome to derive the underlying biological meaning, and they can be aggregated by overlapping their sequence, reducing the big issue of redundancy. Furthermore, the specificity of the sequence allows the application of *κ-mer*s as biomarkers for clinical

applications, using counting sequencing such as Nanostring nCounter described in the previous chapter.

Identification of IR events

Eight years after the publication of the first version of IRFinder[182], with more than 400 cumulative citations, the software is a reference for IR analysis.
The reasons for its success reside not only in the quality of the analysis but also in the end-to-end implementation that takes care of all the aspects of the analysis of raw data, including the STAR mapper reference generation, adapter trimming and differential analysis procedures.
The aspects of usability of the software had been improved during those years, also thanks to users feedback that helped to solve different bugs. Nevertheless, there are still a few aspects that require additional effort:

1. Long read sequencing is gaining more and more importance, especially in studies involving transcript structure. The pipeline is calibrated around short-read sequencing, not only for what concerns the type of aligner but also for the assumptions that are postulated computing the IRratio.
2. Despite the strategies used to mask regions overlapping low mappability regions and known features, such as additional exons and non-coding RNAs, there is a considerable portion of false-positive IR events that can be discriminated by visual inspection on a genome browser.
3. The IR database, IRbase, built in 2017 from 2000 human samples is outdated and doesn't allow the user to easily visualize and compare his own data with the ones included in the database.
4. The differential IR approach was not validated in previous works and requires knowledge of the software R.

During my last year of PhD, I worked with my colleague Sylvain Barrier to improve IRFinder, focusing not only on the four points described before but also enhancing the aspect of usability and speed that lead to its success.
The result of our work is IRFinder-S^3, a suite of tools including a second version of IRFinder and a completely revised version of IRBase, described in the paper below. My contribution to this work comprehends: the design of each new component, enriched by frequent and useful discussions with S.B. and W.R.; the implementation of the new component with the exclusion of the CNN model, trained, tested and optimized by S.B.

The CNN Model

Convolutional Neural Networks (CNNs) are a special case of Artificial Neural Networks (ANNs) in which the connections have been arranged in a way that produces a convolution operation, hence their name. A detailed explanation of this important field can be found in the book Deep Learning[232].

Convolutional neural networks have a special type of layer, the convolutional layer, where the convolution is produced. Intuitively, a convolution consists of matching a pattern present in the kernel across all possible positions in the image. In this sense, matching is an element-wise multiplication between the kernel and each possible position in the image. The element-wise product of each position is then summed to generate an output value, as shown in Figure 13.

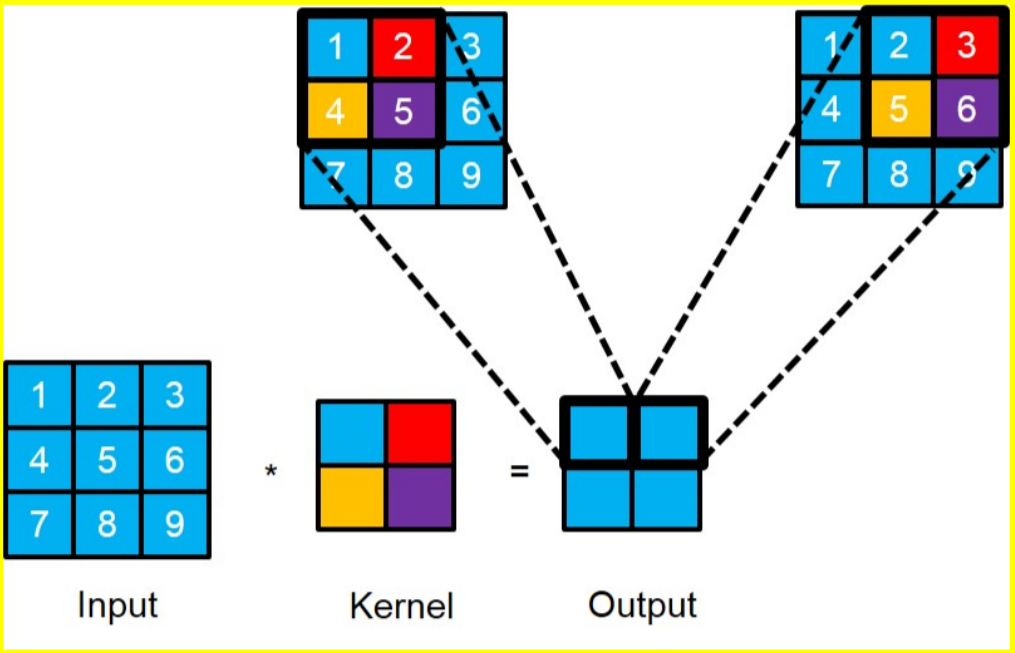


Fig. 13: Illustration of a discrete convolution between an 2D input image and a 2D convolution kernel. From: http://www.theses.fr/2020GRALM043

The output of the convolution of the image with a kernel is called a feature map, and each value of this matrix is obtained by taking the image values within a window, having the same shape as the kernel, and multiplying them element-wise with the kernel. These are then summed to obtain a single value. The window of image values is then moved by a certain amount, the kernel stride, and the element-wise multiplication and summation are repeated. Each of the values in the output feature maps represents the absence or the presence of the filter's pattern inside the image. The whole point with CNNs is to find "features" allowing them to represent objects by learning them directly from data, instead of hand-crafting or manually selecting them. This is done by updating the weights (kernels in the image context) in an iterative manner such that these updates help minimize an error measuring function. The adopted solutions to this problem are two well-known algorithms: gradient descent[233] and error backpropagation[234].

Usually, a pooling layer is added after the convolutional one: the aim is to replace the output at a certain location with a summary statistic, usually the maximum, of nearby outputs. This makes the representation invariant to small translations of the input

and therefore allows the detection of the presence of a feature more than its precise location. Replacing a region with a summary, this layer also improves the computational and memory efficiency of the model, reducing the number of inputs in the next layer.
In IRFinder-S we trained a CNN model using image-like vectors generated during the main process of IRFinder where the BAM file is processed to estimate the IRratio. Those vectors contain the information of the potential retained introns, including 15 nucleotides of the flanking exons, in a one-dimensional array with two channels. The only dimension represents the genomic position, the first channel represents the number of reads that cover the related position and the second the number of reads that are spliced.
Considering only introns having an IR ratio higher than 0.05, therefore presenting a considerable level of intron retention, the goal of the model is to classify introns that are truly retained from the ones that aren't.
To determine the ground truth, if an intron is truly retained or not, we use long reads and we filter the introns whose coverage isn't sufficient to have a good degree of confidence about their retention state.
The evaluation of the model performances requires a cross-validation procedure where the dataset is divided into *n* equal partitions. *n* models are trained using the data from *n*-1 partitions and tested on the remaining one.
This process allows us to estimate the performances of the model on unseen data.
Finally, to evaluate if the model could be generalized on different biological sources, we tested the model trained on a whole dataset using two external cohorts, one generated using the same cell line in a different differentiation state and a second one generated using a different cell line.

IRFinder-S: a comprehensive suite to discover and explore intron retention

Lorenzi *et al. Genome Biology* (2021) 22:307
https://doi.org/10.1186/s13059-021-02515-8

Genome Biology

SOFTWARE Open Access

IRFinder-S: a comprehensive suite to discover and explore intron retention

Claudio Lorenzi[†], Sylvain Barriere[†], Katharina Arnold, Reini F. Luco, Andrew J. Oldfield and William Ritchie[*]

* Correspondence: william.ritchie@igh.cnrs.fr
IRFinder-S is freely available at: https://github.com/RitchieLabIGH/IRFinder
[†]Claudio Lorenzi and Sylvain Barriere contributed equally to this work.
Institut de Génétique Humaine, Centre National de la Recherche Scientifique (CNRS), Université de Montpellier, Montpellier, France

Abstract

Accurate quantification and detection of intron retention levels require specialized software. Building on our previous software, we create a suite of tools called IRFinder-S, to analyze and explore intron retention events in multiple samples. Specifically, IRFinder-S allows a better identification of true intron retention events using a convolutional neural network, allows the sharing of intron retention results between labs, integrates a dynamic database to explore and contrast available samples, and provides a tested method to detect differential levels of intron retention.

Keywords: Intron retention, Splicing efficiency, RNA sequencing

Background

Intron retention (IR) occurs when an intron is transcribed into pre-mRNA and remains in the final mRNA. It is a type of alternative splicing that is gaining increased interest in human health and disease research. Originally described in plants and viruses, IR has now been shown to be a common form of alternative splicing in mammalian systems with a major impact on normal biology and disease [1–7]. However, detecting IR events poses several specific difficulties. Introns are highly heterogeneous genomic regions, both in length and sequence features. In mammals, IR levels are generally low and thereby subject to incomplete coverage and higher count overdispersion. As a result, software that is not specifically tuned for IR detection generally performs poorly and databases that provide transcript isoform sequences fail to list many IR events [4, 8].

We previously published a method called IRFinder, an algorithm for detecting and quantifying IR events, that is frequently used as a benchmark for IR detection and quantification [8–12]. This software and its associated database have been critical in the detection and interpretation of IR events in numerous studies [13–19]. However, building on 4 years of user feedback, it is apparent that IRFinder is lacking features that would enable bench scientists to more reliably identify actionable IR events, share IR data, and dynamically analyze changes in IR levels between multiple samples. We have implemented a suite of features in a new version of our software called IRFinder-

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

S. Specifically, we have (1) created a dynamic database that allows users to perform a meta-analysis, contrast IR from multiple samples, and view IR in an internal browser; (2) created an infrastructure allowing users to share IR detection results from their own samples; (3) implemented a convolutional neural network that analyzes genomic coordinates, as a genome browser would display, and pinpoints IR events that are most likely candidates for further wet-lab analysis; (4) implemented IR detection from third-generation long sequencing technologies; and (5) implemented and tested differential analysis of IR levels between samples.

Results and discussion

IRBase enables the visualization and contrast of IR events as well as data sharing

It is essential to visualize and contrast specific intron retention events detected by computational approaches before spending resources on their experimental validation. This allows users to understand the transcriptional context of a predicted IR event but also to assess whether the event is common to other cell types or specific to their experiment of interest. We therefore created a web application that allows users to upload their own data, decide whether to keep them private, or share them with other users and visualize the results in a javascript version of the IGV genome browser. We propose two types of tracks to visualize the IR events: a bar mode, showing the ratio values like a BedGraph and an IRFinder track to visualize the abundances of the flanking regions, the number of reads spliced and intron read depth (Fig. 1A and Additional file 1: Fig. S1). These views can integrate results from publicly available datasets and shared data from other users (Fig. 1B). Currently, IRbase accepts results from hg38 and ENSEMBL annotation and contains 935 cell lines (downloaded from https://portals.broadinstitute.org/ccle). This database is fully integrated within the IRFinder detection tool; users who have predicted IR events using our software are prompted to upload and share their results. By facilitating the upload process and allowing easy integration using flexible labelling of experiments using user-defined tags, we ensure that the database can grow steadily. The database is accessible for meta-analyses across tissue types and conditions and allows users to contrast multiple experiments in one interface.

IRFinder-S integrates long read detection of IR

Third-generation sequencing technologies, especially direct RNA sequencing, represent a unique opportunity for the detection, characterization, and validation of IR. Because these technologies are capable of sequencing individual RNA molecules from start to end, they can elucidate the full structure of transcripts with retained introns. As a consequence, long reads can be considered as a means of validating IR predictions obtained from SR data. The increased availability of long reads facilitates the study of splicing structure, including a more reliable identification of IR events. IRFinder-S proposes a dedicated version of the algorithm for long-read sequencing (Fig. 2 and Materials and Methods). In this long-read mode, we make multiple adjustments to the algorithm to account for the specificities of long-read data but also to account for the fact that these reads will often serve as the validation of IR and thus the default parameters are more stringent. Firstly, the mapping algorithm STAR is replaced by Minimap2 [20], a

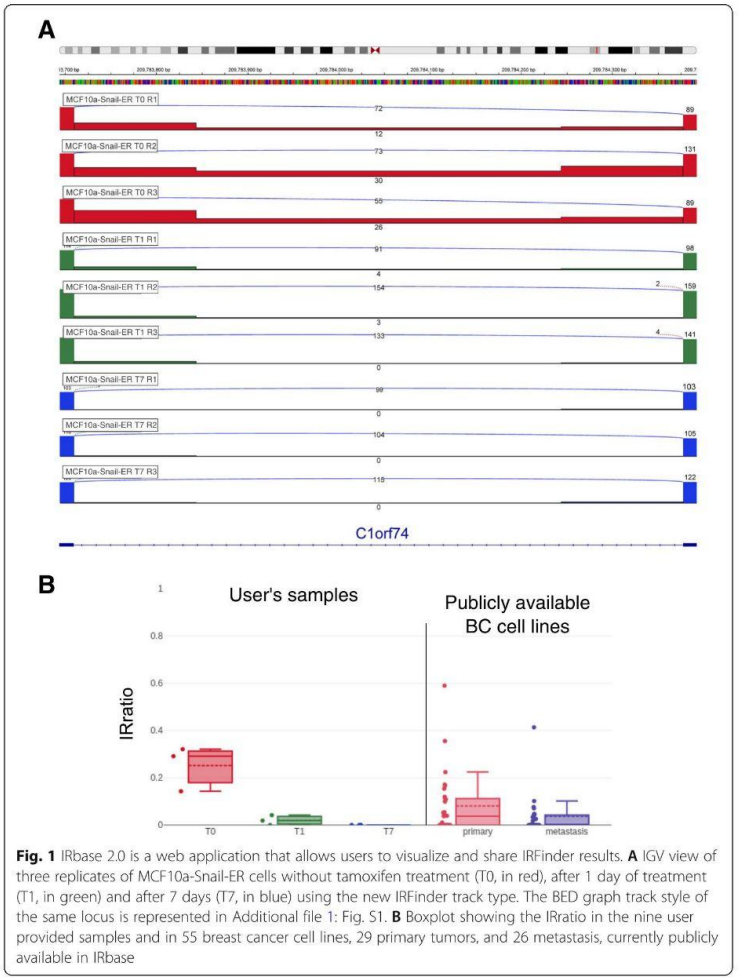


Fig. 1 IRbase 2.0 is a web application that allows users to visualize and share IRFinder results. **A** IGV view of three replicates of MCF10a-Snail-ER cells without tamoxifen treatment (T0, in red), after 1 day of treatment (T1, in green) and after 7 days (T7, in blue) using the new IRFinder track type. The BED graph track style of the same locus is represented in Additional file 1: Fig. S1. **B** Boxplot showing the IRratio in the nine user provided samples and in 55 breast cancer cell lines, 29 primary tumors, and 26 metastasis, currently publicly available in IRbase

specialized aligner for long reads providing competitive alignment accuracy and low computational requirements. Secondly, because long-reads have a higher error rate that often leads to slight imprecision in the definition of exonic boundaries (Fig. 2B), we allow by default up to three nucleotide jitter in exonic boundaries when calculating correctly spliced introns (parameter -j). Thirdly, we only consider the minimum read depth rather than the median when considering retained intron abundance. These modifications allow us to use more long reads when measuring IR levels and also filter out reads for which IR calls would be uncertain (Fig. 2C).

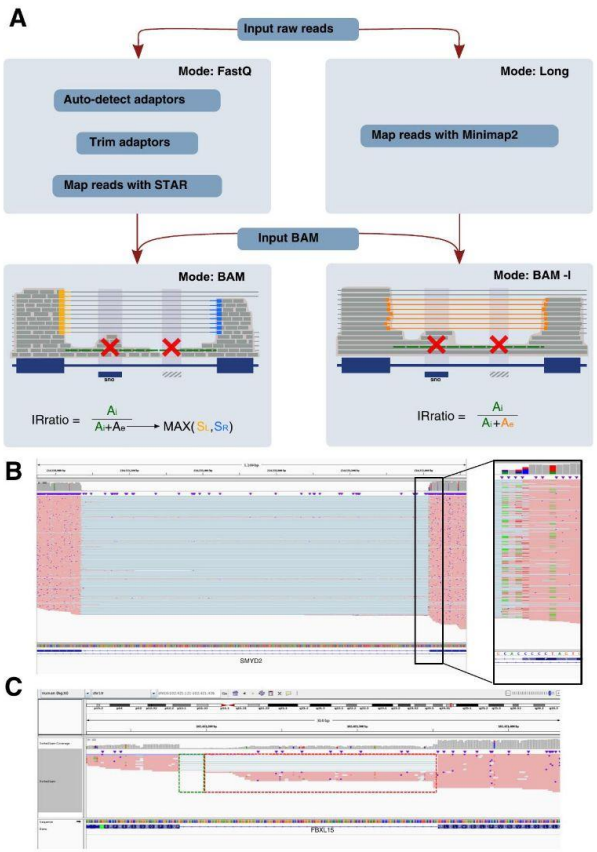


Fig. 2 IRratio estimation for long reads. **A** Main changes in the pipeline between short reads (left) and long reads (right). A_i = intron abundance and A_e = exon abundance. In short reads, A_i is estimated as the median value of the intron depth; in long reads, we use the minimum intron depth. The exon abundance in short reads is estimated as the maximum value between the number of reads spliced at the 5′ or 3′ of the intron; in long reads, it is the number of reads that are spliced in 5′ and 3′. Red crosses indicate regions that will be excluded from further analysis either due to other overlapping transcripts (left cross) or low mappability regions (right cross). Green lines indicate intronic abundance (A_i). **B** Example of poor alignment to exonic borders due to sequencing errors that creates a jitter effect on the mapping of long reads around the splice junctions. Without jitter (option -j 0), the software identifies only 110 SpliceLeft reads, 0 SpliceRight, and 0 ExactSplice. With the jitter option (-j 3), it identifies 183 SpliceLeft reads, 210 SpliceRight reads, and 69 ExactSplice. Having an intron abundance (*Ai*) of 4 and using the ExactSplice as Ae, the IRratio in the first case is 1 and it raise the LowCover warning; in the second case, the IRratio is 0.055, raising the MinorIsoform warning due to the imbalance between the ExactSplice and the max(SpliceLeft, SpliceRight)". **C** Example of alternative 3′ end that is considered as intron retention by the standard method (IRratio=0.40, due to the reads in the red box that do not extend fully between the two exons) and not by the long read mode (IRratio=0, due to the absence of reads covering the green and red boxes)

Convolutional neural networks enable users to pinpoint actionable IR events

Feedback from the users of our first version of IRFinder confirmed that visual inspection of IR events was a crucial step in selecting candidates. Specific patterns that an

expert could detect in a genome browser increased the likelihood of selecting good candidates. Features such as the regularity of intronic coverage, the presence of well-defined exons, and other features contributed to the review of IRFinder candidates. However, this process is time-consuming and variable from user to user. Thus, we tried to reproduce this expert viewing by using a deep-learning approach that would detect these patterns from a dataset of high-quality IR events. To this end, we trained a convolutional neural network (CNN) using high confidence retained introns confirmed by long reads as ground truth. This CNN filter is directly integrated into IRFinder, and it works by transforming coverage data into visual arrays that are submitted to the CNN (Fig. 3A). To test this approach, we used an inducible cell reprogramming system based on human MCF10A cells that recapitulates the epithelial-mesenchymal transition (EMT, Materials and Methods) for which we had access to both short- and long-read RNA-seq data (Fig. 3B). In this system, MCF10a cells stably express the EMT-inducing transcription factor Snail fused to the estrogen receptor. Upon treatment with tamoxifen, the first changes in alternative splicing can be observed as soon as 24h, while a complete cell reprogramming is reached upon 7 days of treatment. We thus used as a training set three biological replicates of untreated epithelial cells and three replicates treated for 1 day with tamoxifen, which corresponds to the first day of the EMT transition. As a first external validation set, we used three biological replicates of cells treated for 7 days with tamoxifen, corresponding to the fully induced mesenchymal-like state. This division aims to validate the model on new IR events that are likely to emerge in the mesenchymal-like state and therefore never seen by the model in the training dataset. As a second external validation set, we used long-read data of GM12878 B-Lymphocite cell lines, provided by the nanopore consortium [21]. Because there was no short read (SR) dataset provided with this experiment, we used the GM12878 Illumina data from an earlier ENCODE study, processing the data as described in our previous study [22]. We considered IR events detected in both short reads and long-reads as bonafide IR events to measure true positives (Material and Methods). We trained the model to recognize the true positive introns from the false positive ones in a 10-fold cross-validation procedure. We then evaluated our model on a biologically distinct dataset where the cells had fully transitioned to their mesenchymal-like state. On this independent test set, it achieved a sensitivity of 0.90 and a specificity of 0.88, with a balanced accuracy of 0.89 (Fig. 3B, right). We then evaluated our model on a different cell line, GM12878, where the model achieved a sensitivity of 0.81, specificity of 0.83, and a balanced accuracy of 0.82.

We then benchmarked IRFinder-S against iREAD [12], a recent software dedicated to the analysis of intron retention, MAJIQ [23], a software designed for the analysis of alternative splicing events that adjust the PSI value of retained intron, and Whippet [24], another software that uses fastq files to compute PSI values. These software were selected based on their popularity but also on whether they could output a measure of retained versus spliced out introns. The results are shown in Table 1 and Additional file 1: Fig. S2. It is worth noting that Whippet excludes a high number of introns prior to their quantification steps since it builds its reference based only on known retained introns and would thus be unable to detect rare or unannotated IR events. iREAD excludes all the introns overlapping with other features. For IRFinder, we excluded the introns reporting any warnings. IRFinder-S achieves the best overall

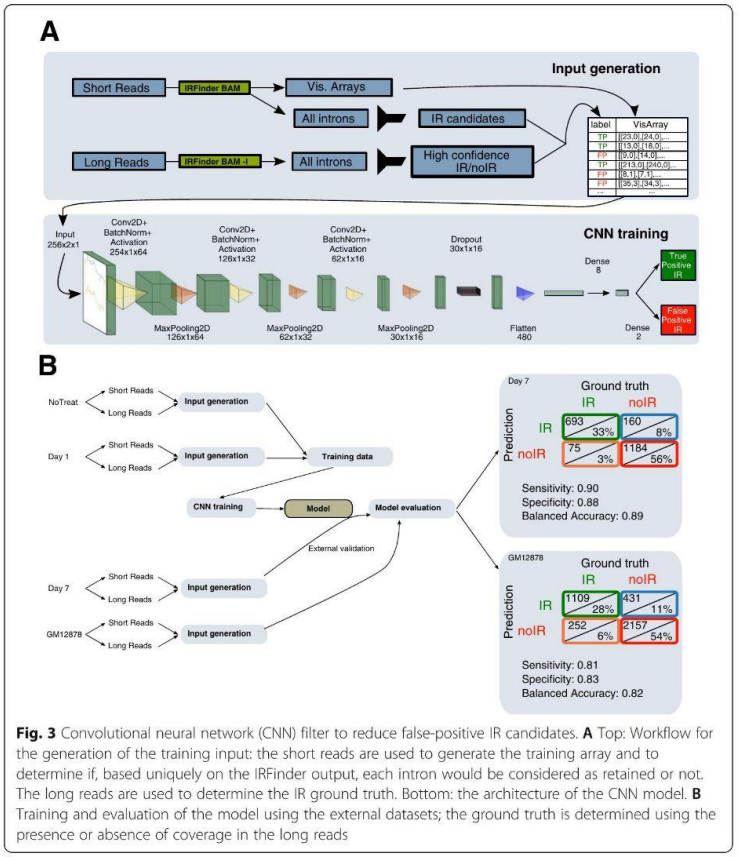


Fig. 3 Convolutional neural network (CNN) filter to reduce false-positive IR candidates. **A** Top: Workflow for the generation of the training input: the short reads are used to generate the training array and to determine if, based uniquely on the IRFinder output, each intron would be considered as retained or not. The long reads are used to determine the IR ground truth. Bottom: the architecture of the CNN model. **B** Training and evaluation of the model using the external datasets; the ground truth is determined using the presence or absence of coverage in the long reads

performance, excludes the least introns before analysis, and thanks to the CNN it does not require the user to set a threshold on IR ratio. To benchmark execution time, we ran a single sample (the third replicate of the EM test sample) on a single core. IRFinder-S processed a single BAM in 20 min, MAJIQ 31 min, and iREAD 50. Whippet took 194 min to process a sample; however, Whippet starts from FASTQ files instead of already aligned BAM files, for which the alignment takes 120 min using STAR on a single core. Interestingly, when we add the CNN on top of the other benchmarked algorithms, it reduces the number of false positive introns, at the expense of a small number of true positives (Additional file 1: Fig. S3) making the CNN a valuable approach for our algorithm but also for other approaches. An example of an intron correctly filtered out by the CNN is presented in Additional file 1: Fig. S4.

Inspection of examples where the CNN was mistaken reveal that the same mistakes would probably have been made by visual examination by an expert; the false positives generally present a homogenous coverage across the intron (Additional file 1: Fig. S5A top right) and false negatives seem to present unevenly covered intronic regions

Table 1 Table representing the results of the benchmark on the EMT test dataset (**A**) and on the GM12878 test dataset (**B**) using a threshold for the PSI values and IR ratios of 0.10

Method	Excl.	TP	TN	FP	FN	TPR	TNR	PPV	Acc.	FDR
A. EMT test results										
IRFinder	**25989**	695	95198	428	138	**0.83**	0.99	0.62	0.98	0.38
IRFinder-S	**25989**	673	35515	111	160	0.81	**1.00**	**0.86**	**0.99**	**0.14**
iREAD	28221	18	33994	86	129	0.12	**1.00**	0.17	**0.99**	0.83
Whippet	59822	443	1978	87	118	0.79	0.96	0.84	0.92	0.16
MAJIQ	30179	388	29572	1951	358	0.52	0.94	0.17	0.93	0.83
B. GM12878 test results										
IRFinder	**30943**	1228	50720	729	185	0.87	0.99	0.63	0.98	0.37
IRFinder-S	**30943**	1077	51123	326	336	0.76	0.99	**0.77**	**0.99**	**0.23**
iREAD	37905	71	45501	179	149	0.32	**1.00**	0.28	**0.99**	0.72
Whippet	80125	772	2459	347	102	**0.88**	0.88	0.69	0.88	0.31
MAJIQ	50932	826	30626	917	504	0.62	0.97	0.47	0.96	0.53

Excl intron excluded, *TP* true positive, *TN* true negative, *FP* false positive, *FN* false negative, *TPR* true positive rate (sensitivity), *TNR* true negative rate (specificity), *PPV* positive predicted value (precision), *Acc.* accuracy, *FDR* false discovery rate

(Additional file 1: Fig. S5A bottom left). Finally, the performance of our CNN may be underestimated because the misclassified IR events are generally borderline with IRratios close to the threshold of 0.1, and mislabeled introns, due to incongruences between long- and short-read resolution (Additional file 1: Fig. S5B).

Implementation and validation of differential IR analysis

In our first version of IRFinder, we suggested methods to analyze differential IR (DIR) using either standalone scripts written in a different coding language or a procedure requiring the user to have extensive knowledge of data transformation and statistical languages such as R. In IRFinder-S, we include IRFinder Diff, an integrated method that allows end to end analysis using either the density-based approach, DESeq2 [25], or the PSI-based approach, SUPPA2 [26] adapted for IR ratios (Material and methods). The output can be used in SUPPA2 downstream analysis for clustering analysis for example. Our choice of algorithms was based on the popularity of these two approaches for the analysis of transcriptomic data. We now wanted to test if they were suitable for the detection of differential IR.

In order to corroborate and compare DESeq2 and SUPPA2 as methods to identify differentially retained introns, we used the aforementioned EMT system (Materials and methods). We compared three replicates of EMT-induced MCF10a cells (mesenchymal-like state) and three untreated control replicates (epithelial state) to detect differentially retained introns between the mesenchymal and epithelial states (Fig. 4A). Using standard settings for both algorithms (BH adjusted p value < 0.05 for both, absolute FC > 1.5 for DESeq2 and delta ratio ≥ 0.1 for SUPPA2), we found that DESeq2 identified 148 differentially retained introns and SUPPA2 found 46 (Additional file 2: Table S1 and Additional file 3: Table S2). 31 differential IR events were common between the two. In both cases, introns were considered if at least one sample had IRratio > 0.05.

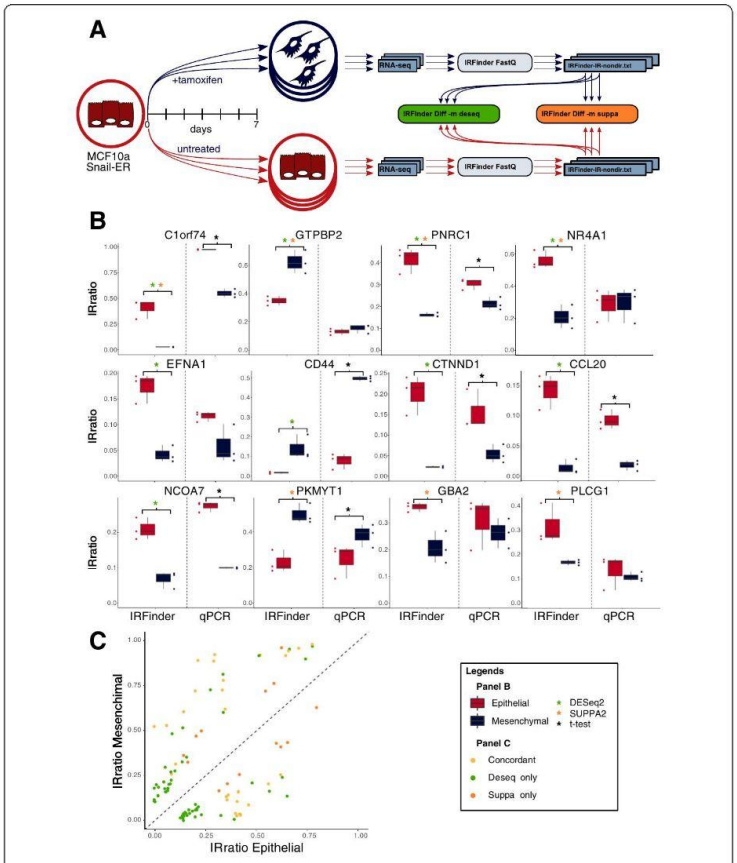


Fig. 4 Differential Intron retention validation between stages of EMT differentiation. **A** EMT is induced in MCF10a Snail-ER system. RNA-seq data is analyzed with IRFinder, and the replicates are compared using DESeq2 and SUPPA2 wrappers. **B** qPCR validation of 12 introns called as differentially retained by DESeq2 and/or SUPPA2. The complete list of primers and characteristics of each intron is described in supplementary file 1. **C** Scatter plot showing the IRratios of the introns called as differentially retained by DESeq2, in green, SUPPA2, in orange, and by both in yellow

We selected 12 introns called as differentially retained and that were suitable for clean primer design in that they did not overlap with other exons or have any known alternative donor or acceptor sites. The selected introns were the following: four introns, in the genes C1ORF74, GTPBP2, PNRC1, and NR4A1 called by both methods; six introns, in the genes EFNA1, CD44, CTNND1, CLL20, and NCOA7, called only by DEseq2; and 3 introns, in the genes PKMYT1, GBA2, and PLCG1, called only by SUPPA2. Figure 4B shows the delta IRratios between epithelial and mesenchymal replicates as computed by IRFinder and the ones obtained by qPCR validations. Of the 12 tested introns, 7 were confirmed using RT-qPCR. The comparison between IRFinder-S and RT-qPCR results showed that both approaches display comparable changes

between epithelial and mesenchymal IR ratios. However, we observed that DESeq2 identifies more DIR events in samples with an average lower IRratio (Fig. 4C). This may be explained by the fact that events with low intronic coverage produce highly variable IR ratio values. As a consequence the ratio values may be highly variable within replicates and methods such as SUPPA2 which make use of replicate variability to determine uncertainty may not produce statistically significant scores. As such, DESeq2 is chosen as the default with SUPPA2 available if required.

Conclusion

Until recently, IR detection ran parallel with the analysis of other splicing events without taking into account inherent difficulties in measuring intronic expression. As a result, IR has been systematically underestimated. Despite the recent development of specialized software for detecting IR, the measurement of IR levels has been problematic. Here, we introduce IRFinder-S to overcome major obstacles in IR detection and exploration. These include a database to explore IR in numerous tissue types and share IRFinder results, the addition of a CNN filter to drastically reduce the false-positive rate of IR detection, the inclusion of an experimentally validated approach to detect differential IR, and the ability to analyze long-read sequencing data. In addition, IRFinder-S overcomes many issues unveiled in the last 4 years thanks to community feedback, such as the possibility to give pre-computed low mappability areas, whose creation step takes most of the time during the reference creation, the possibility to link pre-existing STAR reference folders, and a detailed help divided by run modes. Finally, Docker and Singularity images including all the dependencies required to run IRFinder on any Linux distribution are available in dockerhub (cloxd/irfinder:2.0) and in GitHub (https://github.com/RitchieLabIGH/IRFinder).

Methods

IRbase 2.0

The new version of IRbase consists in a frontend, implemented with Angular 10, a mySQL database containing the basic information about each sample submitted and the introns having IRratio higher than 0.05, warning different than “LowCover” and a tag-based aggregation system that allows fast queries to obtain statistics on large number of samples.

The backend is implemented in node express version 4.17.1. We generated two novel tracks to show IRFinder results (IRFinder-IR-[non]dir.txt files) directly on igv.js, one displaying the IRratio as bedgraph and one that combines the additional information included in the file allowing the representation in detail of the flanking exons, the spliced reads, and the intron depths, as shown in Fig. 1.

The user authentication is managed by Google’s service firebase and is necessary in order to upload new samples. Currently, IRbase requires results from hg38 with ENSEMBL reference.

Measuring intron retention in long reads

In order to adapt the IRratio computation in long read, we adapted the estimation of intron and exon abundance keeping unchanged the formula:

$$IRratio = \frac{Intronic\ abundance}{(\ Intronic\ abundance + exonic\ abundance\ )}$$

A visual representation of the main changes is shown in Fig. 2. The intron abundance in long reads is evaluated as the minimum coverage in the intron instead of the median, offering a more stringent but reliable IRratio. The exon abundance in long reads is estimated as the exact number of reads spliced between the acceptor and the donor site, rather than the highest number of reads spliced between donor and acceptor sites. Finally, in order to take into account the long reads' higher error rate, the count of the splits is considered not only for the exact split nucleotide annotated but also the three flanking positions.

This alternative version is used by default in IRFinder long mode and is triggerable by the "-l" flag argument using IRFinder BAM.

Convolutional neural networks

The network was trained on the epithelial datasets labeled T0 and T1 (days 0 and 1 of treatment) and validated on the mesenchymal dataset T7, described in our previous work [22] and having biological samples sequenced with both unstranded short and stranded long-read technologies. We use IRFinder to analyze the raw data, and for each pair of data belonging to a sample, we selected the introns with IRratio above 0.05 and no warnings in short reads, as putative IR candidates. We then used the long reads as ground truth of the corresponding intron: we labeled as true positive IR, the introns with no warning, depth (intron abundance + exon abundance) of 25, and IRratio above 0.1 and as false positive IR, the introns with 50 depth and IRratio of 0. Our rationale is that it is easier to assert the existence of IR events than to assert their absence; thus, we pushed the required depth for negative events to 50 to increase their likelihood of being true negatives.

To allow the model to use directional and non-directional libraries and to reduce mislabeled events, we considered only the introns having a congruent label between the directional and non-directional long reads IRFinder results. Due to the scarcity of FP, we included in the training set also true negative introns having IRratio higher than 0.01 in the SR to ensure a balanced dataset.

Benchmark

To compare IRFinder's results with the output of iREAD [12], Whippet [24] (v1.6.1) and MAJIQ [23] (Build v2.1-c3da3ce), we used the reference genome hg38 and ENSEMBL v100 annotation, generating the required reference files for each software. We paired the results of each method with the introns of the ground truth determined from the long reads in the test datasets as described in the previous chapter.

We used two arbitrary thresholds, 0.05 and 0.10, for the PSI values of Whippet and IRFinder's IRratio to classify the introns in IR and non-IR. For what concerns MAJIQ, we considered as no IR the introns without a PSI value adjusted for intron retention and the introns having an adjusted PSI value lower than the two arbitrary thresholds.

Differential intron retention

The DESeq2 constructor is used to fit a GLM based on the intronic abundance (intron depth column) and the exonic abundance (the maximum between LeftSplice and RightSplice) to test the fold change of IR between two conditions.

The SUPPA2 wrapper uses IRratio values instead of percent splice in (PSI) values, both spanning from 0 to 1, and the exon abundance instead of transcript per million (TPM) values, considering so far the expression of the exons surrounding each intron rather than the average transcript expression.

In both cases, the user can decide to remove introns with warnings (by default, introns with LowCover in at least one sample are removed) and to set a threshold on the minimum IRratio that at least one sample has to meet (by default 0.05).

The command line interface offers a simple tool to use DESeq2 or SUPPA2 on two or more sets of samples, requiring only the location of the IRFinder result files IRFinder-IR-[non]dir.txt. In case of more than two sets, all the pairwise comparisons are reported in the output folder.

Cell line culture

Non-transformed human female breast epithelial cells (MCF10a cells) were cultured at 37°C and 5% CO_2 in DMEM/F12 (Sigma) supplemented with 5% horse serum (ThermoFisher), 10 ng/ml EGF (Sigma), 10 μg/ml insulin (Sigma), 0.1 μg/ml cholera toxin (Sigma), 0.5 μg/ml hydrocortisone (Sigma), 1% L-glutamine (Sigma), and 1% penicillin/streptomycin (Sigma; culture medium). Cells were kept in high confluency (approx. 70%) in order to maintain their epithelial character and passed every 2–3 days by trypsinization (0.25% Trypsin (Sigma) for 15–20 min).

Epithelial-mesenchymal transition (EMT)

MCF10a-Snail-ER cells were used as cellular model for EMT. In this model, EMT is induced by addition of exogenous 4-hydroxy-tamoxifen to the cells, which changes Snail-ER conformation and can thus be translocated to the nucleus for silencing of key epithelial markers and expression of mesenchymal genes within 24 h. Prior to induction, 850,000 cells were seeded in 15-cm culture plates and grown in 17-ml culture medium for approximately 24 h. Twelve hours before tamoxifen treatment, the cells were synchronized by exchanging the medium to serum free medium (culture medium without horse serum). Cells were incubated for 6 days in culture medium with 100 nM 4-hydroxy-tamoxifen (Sigma). Controls were performed by adding equivalent volumes of methanol.

Primer design

We selected IR events by visual inspection, selecting introns without neither antisense transcript nor known exon in each sample and without excessive noise in the intron body. Two sets of primers were designed for each intron, one pair overlapping the exon-exon junction and one covering the intron-exon junction.

RT-qPCR

RT-qPCRs were performed in biological triplicates. RNA was extracted from cells using QIAshredder (Qiagen, 79656) and GeneJET RNA purification kit (Thermo Scientific, #K0732) following the manufacturers' instructions. 500 ng of the total RNA was DNase treated (Promega, M6101) and reverse-transcribed using oligo(dT) primers (Transcriptor First Strand cDNA Synthesis kit, Roche 04897030001).

For each biological replicate, qPCRs were performed in technical duplicates using Bio-Rad CFX-96 Real-Time PCR System and iTaq Universal SYBR green Super-mix (Bio-Rad #1725121). For each intron of interest, two primer pairs were designed that includes the exon-exon (of the flanking exons) and an intron-exon junction, respectively.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02515-8.

Additional file 1. Supplemental Figures.
Additional file 2. Supplemental Table S1.
Additional file 3. Supplemental Table S2.
Additional file 4. Review history.

Peer review information
Anahita Bishop was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 4.

Authors' contributions
C.L., W.R., and S.B. designed the algorithm; C.L. and S.B. coded the software; C.L., S.B. W.R. designed the experiments; K.A, A.O., and R.L. designed and performed the qPCR validation experiments. W.R., S.B., and C.L. wrote the article. The authors read and approved the final manuscript.

Funding
We wish to acknowledge the Agence Nationale de la Recherche (ANRJCJC - WIRED), the Labex EpiGenMed, and the MUSE initiative for their financial support.

Availability of data and materials
Cell line data used to help populate the database was taken from: https://portal.gdc.cancer.gov.
Direct RNA Nanopore and Illumina RNA-seq MCF10A samples have been deposited on GEO under accession number GSE126638 [27].
GM12878 cell line, the long read data, was available from the Nanopore consortium at https://github.com/nanopore-wgs-consortium/NA12878. We made use of the *Run1* (MinION ONT direct-RNA, kit SQK-RNA001, pore R9.4) generated by the UCSC laboratory. These long reads were corrected using short read data from the same cell line sequenced by a separate consortium. These data were available from the GEO website (https://www.ncbi.nlm.nih.gov/sra/SRX159827). After the quality control using FastQC, we kept and pooled together runs SRR521447, SRR521448, SRR521453, SRR521454, and SRR521455.
An OceanCode capsule is available at https://codeocean.com/capsule/0822057/tree [28] that reproduces the main functionalities of IRFinder-S.
IRFinder-S is available at https://github.com/RitchieLabIGH/IRFinder [29] under the MIT license.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 3 June 2021 Accepted: 12 October 2021
Published online: 08 November 2021

References

1. Braunschweig U, et al. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. 2014; 24:1774–86.
2. Wong JJ-L, et al. Orchestrated intron retention regulates normal granulocyte differentiation. Cell. 2013;154:583–95.
3. Middleton R, et al. IRFinder: assessing the impact of intron retention on mammalian gene expression. Genome Biol. 2017;18:51.
4. Broseus L, Ritchie W. Challenges in detecting and quantifying intron retention from next generation sequencing data. Comput Struct Biotechnol J. 2020;18:501–8.
5. Grabski DF, et al. Intron retention and its impact on gene expression and protein diversity: a review and a practical guide. Wiley Interdiscip Rev RNA. 2020:e1631. https://doi.org/10.1002/wrna.1631.
6. Jacob AG, Smith CWJ. Intron retention as a component of regulated gene expression programs. Hum Genet. 2017;136: 1043–57.
7. Smart AC, et al. Intron retention is a source of neoepitopes in cancer. Nat Biotechnol. 2018;36:1056–8.
8. Vanichkina DP, Schmitz U, Wong JJ-L, Rasko JEJ. Challenges in defining the role of intron retention in normal biology and disease. Semin Cell Dev Biol. 2018;75:40–9.
9. Broseus L, et al. TALC: transcription aware long read correction. bioRxiv. 2020:2020.01.10.901728. https://doi.org/10.1101/2020.01.10.901728.
10. de la Fuente L, et al. tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. Genome Biol. 2020;21:119.
11. Lee S, et al. Covering all your bases: incorporating intron signal from RNA-seq data. NAR Genomics Bioinforma. 2020;2: lqaa073.
12. Li H-D, Funk CC, Price ND. iREAD: a tool for intron retention detection from RNA-seq data. BMC Genomics. 2020;21:128.
13. Sachamitr P, et al. PRMT5 inhibition disrupts splicing and stemness in glioblastoma. Nat Commun. 2021;12:979.
14. Tan DJ, Mitra M, Chiu AM, Coller HA. Intron retention is a robust marker of intertumoral heterogeneity in pancreatic ductal adenocarcinoma. NPJ Genomic Med. 2020;5:55.
15. Zhang D, et al. Intron retention is a hallmark and spliceosome represents a therapeutic vulnerability in aggressive prostate cancer. Nat Commun. 2020;11:2089.
16. Ashraf U, et al. Influenza virus infection induces widespread alterations of host cell splicing. NAR Genomics Bioinforma. 2020;2:lqaa095.
17. Green ID, et al. Macrophage development and activation involve coordinated intron retention in key inflammatory regulators. Nucleic Acids Res. 2020;48:6513–29.
18. Ullrich S, Guigó R. Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development. Nucleic Acids Res. 2020;48:1327–40.
19. Burke EE, et al. Dissecting transcriptomic signatures of neuronal differentiation and maturation using iPSCs. Nat Commun. 2020;11:462.
20. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
21. Workman RE, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods. 2019;16:1297–305.
22. Broseus L, et al. TALC: transcript-level aware long read correction. Bioinformatics. 2020. https://doi.org/10.1093/bioinformatics/btaa634.
23. Green CJ, Gazzara MR, Barash Y. MAJIQ-SPEL: web-tool to interrogate classical and complex splicing variations from RNA-Seq data. Bioinformatics. 2018;34:300–2.
24. Sterne-Weiler T, Weatheritt RJ, Best AJ, Ha KCH, Blencowe BJ. Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. Mol Cell. 2018;72:187–200.e6.
25. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
26. Trincado JL, et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 2018;19:40.
27. Broseus L, Severac D, Oldfield AJ, Dubois E, Ritchie W. Short and long read sequencing of human mammary epithelial MCF10a-Snail-ER cells after epithelial-to-mesenchymal transition initiation. Datasets. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126638.
28. Lorenzi C, Barriere S, et al. OceanCode. IRFinder-S: a comprehensive suite to discover and explore intron retention; 2019. https://doi.org/10.24433/CO.5556419.v1. https://codeocean.com/capsule/0822057/tree/v1
29. Ritchie W. github. https://github.com/RitchieLabIGH/IRFinder

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Alternative approaches for the RNA-seq data analysis

Methods like DE-kupl, KOVER and HAWK demonstrated that embedding the information in a human-understandable and interpretable format such as genes or transcripts is not necessary to compare the information contained in sequencing data. κ-mers allow us to compare groups of samples in an agnostic way, unbiased by any reference sequence or annotation, leading to highly reproducible results: the κ-mer counts won't change in the feature but our knowledge about the composition of the reference genome improves every year. Furthermore, κ-mers allow the comparison of small fractions of the RNA molecule, avoiding the loss of information derived from aggregating multiple reads under a single feature, that is gene, transcript or splice junction.
Finally, using a large enough number of samples, it would be possible to associate variations, such as SNP or indels, to a specific population, similarly to genome-wide association studies (GWAS).
The two following paragraphs present the work of my team on two different algorithms designed to identify κ-mers able to classify two or more distinct groups of samples in large cohorts of samples.

GECKO is a genetic algorithm to classify and explore high throughput sequencing data

GEnetic Classification using κ-mer Optimization, GECKO, is the first method that aims to identify groups of κ-mers able to classify two or more groups of samples in large cohort studies.
The method, described in detail in the paper included below, shows that It's possible to identify groups of κ-mers that, alone or in synergy, can classify different groups of patients, with a better performance with respect to gene counts. The approach has been tested on different sequencing data types, such as miRNA, mRNA and bisulfite sequencing data.

In brief, GECKO takes in input raw sequences and uses Jellyfish2 to count the κ-mer abundances in each sample. It then assembles a κ-mer matrix, where each row is a κ-mer and each column is a sample.
The last step of the preprocessing consists of the filtering of the κ-mers considered uninformative, noisy and redundant.

Finally, GECKO implements an adaptive genetic algorithm, an efficient metaheuristic optimization algorithm, to select subsets of κ-mers that maximize the accuracy in classifying the sample groups using a linear support vector classifier (LinSVC).

I joined the lab when GECKO was almost finished and I contributed by implementing the step to reduce the redundancy, the optional step to filter the κ-mers based on the ANOVA f-test and by fixing some bugs.
Some crucial issues of working with κ-mers on large datasets emerged:
1 - The unfiltered κ-mer matrix is sparse and can easily occupy one terabyte of space on a disk in a study with one thousand samples, even in binary form. Furthermore, its fixed structure requires the user to allocate one matrix for each experiment.
2 - Despite the redundancy reduction step, the final output presents several κ-mers mapping to the same biological entity. Though the information content is similar, it might be different enough to escape the symmetric uncertainty based filter.
3 - The process is nondeterministic: running several times the genetic algorithm, different subsets of κ-mers are selected and there is not a clear procedure to select a robust group of κ-mers.

For what concerns the implementation, the use of Nextflow to coordinate different scripts written in different languages (C++, Perl and Python) makes not only the maintenance of the software challenging, but also requires an advanced user for the installation and usage.

COMMUNICATIONS BIOLOGY

ARTICLE

https://doi.org/10.1038/s42003-019-0456-9 OPEN

GECKO is a genetic algorithm to classify and explore high throughput sequencing data

Aubin Thomas[1,6], Sylvain Barriere[1,6], Lucile Broseus[1], Julie Brooke[1], Claudio Lorenzi[1], Jean-Philippe Villemin[1], Gregory Beurier[2], Robert Sabatier[3], Christelle Reynes[3], Alban Mancheron[4,5] & William Ritchie[1]

Comparative analysis of high throughput sequencing data between multiple conditions often involves mapping of sequencing reads to a reference and downstream bioinformatics analyses. Both of these steps may introduce heavy bias and potential data loss. This is especially true in studies where patient transcriptomes or genomes may vary from their references, such as in cancer. Here we describe a novel approach and associated software that makes use of advances in genetic algorithms and feature selection to comprehensively explore massive volumes of sequencing data to classify and discover new sequences of interest without a mapping step and without intensive use of specialized bioinformatics pipelines. We demonstrate that our approach called GECKO for GEnetic Classification using *k-mer* Optimization is effective at classifying and extracting meaningful sequences from multiple types of sequencing approaches including mRNA, microRNA, and DNA methylome data.

[1] Institute of Human Genetics, CNRS UPR1142, Machine learning and gene regulation, University of Montpellier, Montpellier, France. [2] AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France. [3] IGF, Centre National de la Recherche Scientifique, INSERM U1191, University of Montpellier, Montpellier, France. [4] LIRMM, Université de Montpellier, CNRS, UMR5506, Montpellier, France. [5] Institut Biologie Computationnelle, Montpellier, France. [6] These authors contributed equally: Aubin Thomas, Sylvain Barriere. Correspondence and requests for materials should be addressed to W.R. (email: william.ritchie@igh.cnrs.fr)

Studies of variation in gene expression, initially through probe-based technology and more recently high throughput sequencing (HTS), have considerably advanced knowledge of disease etiology and classification[1–3]. The recent promotion of HTS across a wide spectrum of diseases has generated a wealth of data that measure gene expression and transcript diversity but also explore its putative genetic and epigenetic regulators. Still, despite more than a decade of development, computational analysis and integration of these data presents a major challenge. Each type of HTS experiment is compartmentalized to a set of computational pipelines and statistical approaches that often require a full-time bioinformatics specialist. In addition, most of these pipelines rely on a reference genome or transcriptome and thus cannot inherently account for the diversity in non-reference transcripts or individual variations[4]. To remove the requirement of a reference, recent methodologies use *k-mer* representation; they directly compare the counts of nucleotide sequences of length k between samples[5]. These approaches have been successful at detecting novel transcripts but only on a very small subset of RNA sequencing data[4] and would be impossible to implement for the classification of large patient cohorts using the entire transcriptome. In the field of metagenomics, numerous algorithms have been developed to discover unique *k-mers* or *k-mer* signatures to classify organisms[6,7]. However, these were developed for organisms with smaller genomes that do not have billions of different *k-mers*. In addition, they were designed for inter-species studies where unique *k-mers* can be attributed to the genomes of different taxonomic identities.

Exploring a large set of *k-mers* to classify samples can be framed as a global optimization problem for which many recent approaches have been published and compared[8]. Amongst these is a class of nature-inspired algorithms termed Genetic Algorithm which are based on the processes of mutation, crossing over and natural selection. These have appealing properties that could apply to the exploration of a large set of *k-mers*. They have low memory requirements because they explore only part of the data at each stage and they can produce multiple solutions that fit well with biological interpretation of data. However, despite these properties, genetic algorithms are rarely used to optimize problems with relatively small sample sizes and such a large number of parameters, in this case billions of *k-mers*.

We have created a novel approach and associated software called GECKO for genetic classification using *k-mer* optimization that is especially designed for HTS data. GECKO is based on *k-mer* decomposition coupled with an adaptive genetic algorithm that explores HTS data from two or more input conditions. This algorithm searches for groups of *k-mers* that, combined together are highly informative; they are able to classify the input categories with high accuracy. Because GECKO uses *k-mer* counts, it can theoretically be applied to any type of HTS experiment and does not rely on a reference genome or transcriptome. Here, we successfully apply GECKO to a variety of biological problems and sequencing data. These include microRNA (miRNA) sequencing to classify normal blood cells, mRNA sequencing to classify subtypes of breast cancer and to predict response to chemotherapy, and bisulfite sequencing (BS-seq) on normal versus chronic lymphocytic leukemia (CLL) samples. Regardless of the type of data, GECKO finds small, accurate signatures that classify these samples and could thus be used as diagnostic and prognostic markers. In addition, by visualizing how the genetic algorithm evolves to find solutions, GECKO can be used to explore novel sequences or groups of functionally related sequences associated with normal biology and disease.

Results

GECKO is designed around two main steps; these are a *k-mer* matrix preparation step and an adaptive genetic algorithm (Fig. 1).

The *k-mer* matrix preparation, uses an input sequencing file (.bam or .fastq) to create a matrix of *k-mer* counts; that is the number of times a sequence of length k appears in each sample ($k = 30$ by default). This matrix is filtered for *k-mers* with low counts and non-informative or redundant *k-mers* (see the section "Methods"). Then, during the second step an adaptive genetic algorithm will explore the matrix to discover combinations of *k-mers* that can accurately classify input samples. The adaptive genetic algorithm starts by creating thousands of digital individuals; these are groups of randomly selected *k-mers*. The set of individuals is called a population. This population will then go through phases of mutation, where individuals replace one of their *k-mers* with another randomly selected *k-mer*; a phase of crossing-over where individuals exchange a portion of their *k-mers* with each other and selection, where individuals that do not classify the input samples well enough will be removed from the population and replaced. Mutation allows GECKO to explore local solutions similar to the individual to be mutated; crossing-over, allows GECKO to explore a broader set of solutions and reduces the chances of getting stuck in a local minimum (see the section "Methods"). Each cycle of mutation, crossing-over, and selection is called a generation. By default, GECKO will iterate through 20,000 generations or stop when the number of new solutions discovered throughout generations slows down (see stopping criteria in the section "Methods"). This algorithm is called adaptive because the mutation and crossing-over rates depend on how well individuals in the population perform. Individuals that perform well have lower rates to prevent them from changing drastically and thus enabling them to converge

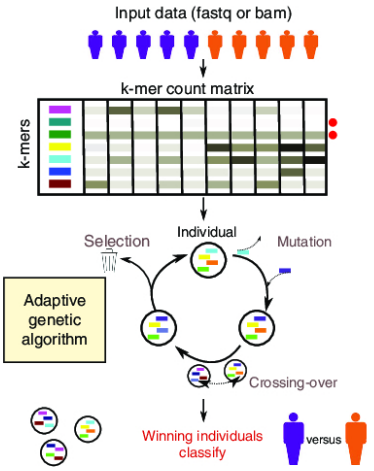


Fig. 1 Overview of the GECKO algorithm. Input fastq or bam files from two or more conditions are transformed into a matrix of *k-mer* counts across all samples. The *k-mers* for which the counts are below a noise threshold or that do not vary across samples are removed (red dots on the right of the *k-mer* matrix). The adaptive genetic algorithm randomly selects groups of *k-mers* from the *k-mer* matrix to form individuals. These individuals will go through rounds of mutation, crossing-over and selection to discover individuals capable of classifying the input samples with high accuracy

faster to a solution; individuals that do not perform well will have higher rates to enable wider exploration of solutions.

In the analyses presented in this study and by default in the software, GECKO's performance is systematically tested on 1/6th of the data that is randomly selected and set aside before running the algorithm (see the section "Methods"). This test set allows us to evaluate the accuracy and overfitting for each run; it measures whether the algorithm fits too closely to the training set and thus will not correctly predict future input samples. GECKO is thus run on the remaining 5/6th of the data with cross-validation at each generation of the algorithm.

Classifying miRNA sequencing data of blood cells. We first tested GECKO's performance on a miRNA expression data of seven types of blood cells sorted from 43 healthy patients for a total of 413 samples[9]. We ran GECKO on this dataset using 20-mers (*k-mer* size of 20; miRNAs generally vary in size from 20 to 23) to find a set of *k-mers* that could correctly classify the seven blood-cell types.

After 6000 generations (15 h on 15 cores; see Supplementary Table 1 for parameters and Supplementary Fig. 1 for runtimes and memory usage) GECKO discovered an individual composed of only three *k-mers* (ACCCGTAGAACCGACCTTGC, CCCCA GGTGTGATTCTGATA, AGTGCATGACAGAACTTGGG) that could distinguish the groups with 0.96 accuracy (Fig. 2a, b and Supplementary Data 1 and 2).

In the initial study, the authors described a signature of 136 cell-type-specific miRNAs. These 136 miRNAs could classify the groups with 0.97 accuracy. Thus, we found a much smaller signature that could classify the seven blood-cell types with similar accuracy without the use of a miRNA-dedicated bioinformatics pipeline.

We then aligned the three *k-mers* discovered by GECKO to a database of known miRNAs[10]. Two of these mapped perfectly to miRNAs 152-3p and 99b-5p, which were annotated in the original study as specific to NK cells and T helper cells, respectively. The third mapped to miRNA 361-3p which was not found to be specific to any of the seven cell types and was thus ignored in the initial study. Separately, the first two *k-mers* could classify one cell-type each and the third would have been overlooked. Together these three *k-mers* classify all seven groups with high accuracy because of their contrasting expression between each cell types (Fig. 2c).

Classifying breast cancer subtypes using mRNA sequencing data. Breast cancer is a heterogeneous disease in regards to response to treatment and its transcriptional background. Defining the subtypes luminal A (LumA), luminal B (LumB), HER2-enriched (HER2) and basal-like are crucial for prognosis and predicting outcome of breast cancer. These subtypes were initially defined through unsupervised clustering of gene expression and are currently identified using a standard qPCR assay of 50 genes called the PAM50[11,12]. To assess whether GECKO could identify *k-mers* that classify breast cancer subtypes, we used a dataset of 1087 mRNA-Seq breast cancer samples from the Cancer Genome Atlas Pan-Gyn cohort[13] (patients per class: Basal 175, Her2 73, LumA 513, LumB 185). We ran GECKO for 20,000 generations (75 h on 15 cores; see Supplementary Table 1 for parameters and Supplementary Fig. 1 for runtimes and memory usage) and extracted the highest scoring individual at its term (Supplementary Table 2). We then tested how well these *k-mers* classified the four cancer subtypes compared to PAM50 expression values calculated as transcript per million (TPM). Both the *k-mer* counts and PAM50 TPMs were trained using a linear support vector machine (see the section "Methods") with identical training data and evaluated on the same test set. The 10 *k-mers* had higher accuracy rates compared to the PAM50 on all four classes (Fig. 3 and Table 1).

We then further inspected the 10 *k-mers* discovered by GECKO by mapping them to the human genome. We found

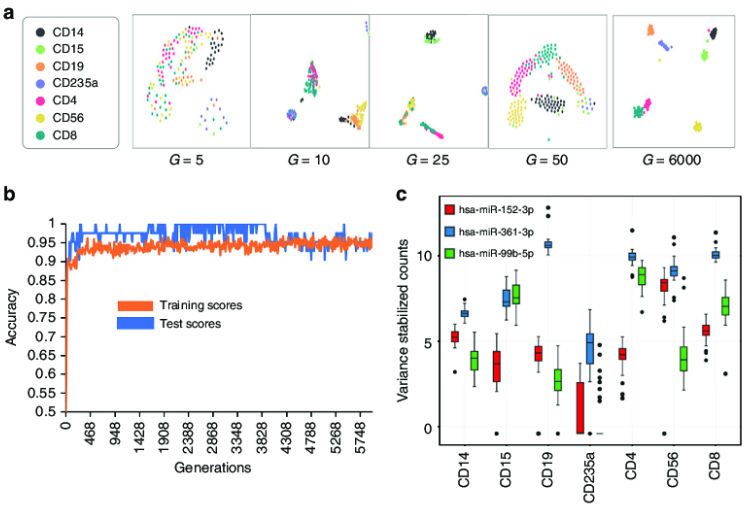


Fig. 2 GECKO can accurately classify miRNA data from seven types of blood cells using three *k-mers*. **a** GECKO output showing the separation of the seven blood-cell types at each generation (G) of GECKO analysis using t-SNE visualization applied to *k-mer* counts. **b** GECKO output showing the accuracy of separation for the training and test set across 6000 generations. **c** variance stabilized counts of the three miRNAs that correspond to the three *k-mers* discovered by GECKO across the seven blood-cell types ($n = 43$ biologically independent donors)

that four of the *k-mers* mapped to genes from the PAM50 list (FOXC1, ESR1, KRT14, KRT17). Three others mapped to genes NISCH, TPX2, and ATF3, the first of which is linked to breast cancer aggressiveness[13] and the two latter both affect cell viability in breast cancer cells[14,15]. The three last *k-mers* mapped to three genes KLHL6, KANSL2, and PHF10 shown to be involved in tumorigenesis but not in breast cancer[16–18]. Of the 10 *k-mers*, 3 map to coding regions and 7 map to 3′ untranslated regions for which multiple isoforms exist. *k-mer* counting can thus integrate alternative transcription to classify mRNA-Seq samples.

Classifying response to chemotherapy of triple negative breast cancer on small sample sizes of mRNA-Seq. We then tested GECKO on a dataset with more heterogeneous cell populations and smaller sample sizes. We used a cohort of triple-negative breast cancer patients, an aggressive, heterogeneous subtype of breast cancer with poor outcomes. This cohort taken from the Breast Cancer Genome Guided Therapy (BEAUTY) study[19,20]

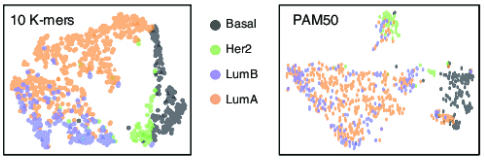


Fig. 3 GECKO discovers 10 *30-mers* that classify breast cancer subtypes. Comparison of breast cancer subtype classification using the frequency of *k-mers* discovered by GECKO and the transcript per million values of the PAM50 gene. Panels show the t-SNE separation of the four classes

was divided into 19 patients that had a complete response to chemotherapy and 20 patients that did not. In such cases of small sample size and high heterogeneity, we recommend using GECKO's voting mode (Fig. 4a).

This mode compensates for bias that may be introduced when splitting a small number of samples between training and test datasets and may thus accentuate batch effects. The voting mode will run 10 instances of the genetic algorithm for 10,000 generations. At their term, it will select *k-mers* from the top individuals across the 10 instances and run a final genetic algorithm on this subset of *k-mers* for another 10,000 generations. Running multiple genetic algorithms and aggregating their results prevents overfitting on a specific split of the data between the training and test set. In addition, the voting mode introduces Gaussian noise by default into the data to further prevent overfitting. This option is recommended for experiments with <30 samples per condition.

Using the voting mode (83 h using 15 cores; see Supplementary Table 1 for parameters and Supplementary Fig. 1 for runtimes and memory usage), we found an individual that was able to classify patients with 0.93 accuracy (Fig. 4b) with only five *k-mers* of length 30 (Supplementary Table 3). As expected three of these *k-mers* mapped to genes that had clear roles in resistance to chemotherapy; *JAK3* is involved in chemotherapy resistance in triple-negative breast cancer[8], *BOP1* reduces chemotherapy resistance[21] and *VTCN1* is associated with poor clinical outcomes in numerous cancers including breast cancer[22].

Classifying BS-seq data. We then wanted to see if GECKO could accurately classify samples using epigenetic sequencing data, such as BS-seq generated to investigate DNA methylation. BS-seq requires extensive bioinformatics processing to discover changes

Table 1 Confusion matrices of breast cancer subtype classification using the frequency of *k-mers* discovered by GECKO and the transcript per million values of the PAM50 gene set

		Classification with GECKO *k-mers*					Classification with PAM50 TPM values				
Predicted class	Basal	97.7	2.2	0	0	Predicted class	Basal	86	5.2	5.5	3.3
	Her2	2	87.5	6.2	4.2		Her2	15.3	60.6	3.6	20.6
	LumA	1.5	1.5	92.3	4.6		LumA	15.3	2.2	88.1	8.6
	LumB	0	3.4	18.8	77.8		LumB	5.9	15.4	36.5	42.2
		Basal	Her2	LumA	LumB			Basal	Her2	LumA	LumB
		True class						True class			

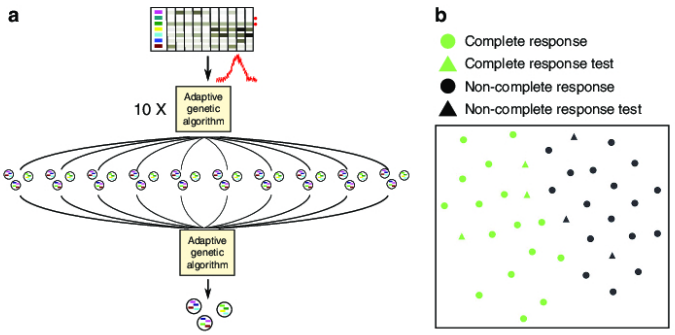


Fig. 4 GECKO voting mode for small sample sizes. **a** GECKO's voting mode will run 10 separate genetic algorithms with added Gaussian noise. The best solutions of these runs will be fed into a final genetic algorithm to produce a final solution. **b** GECKO output showing the t-SNE separation of patients with complete response to chemotherapy from those that did not using five *k-mers* from the winning individual. Triangles correspond to the test dataset that was excluded from GECKO training can thus be used to estimate overfitting

in methylation and thus, a method that could directly classify BS-seq samples could be of great interest. To test GECKO on BS-seq we downloaded raw sequencing files from a study on methylome diversity in 104 primary CLLs samples compared with 26 normal B cell samples[23]. Although global hypomethylation has been well described in cancer, these alterations are highly variable between CLL samples[23] and thus present a challenge for classification.

We ran GECKO for 20,000 generations (39 h; see Supplementary Table 1 for parameters and Supplementary Fig. 1 for runtimes and memory usage) and found a winning individual that was able to classify normal from CLL samples with an accuracy of 1 using 20 *k-mers* (Fig. 5a; Supplementary Table 4). In addition to this final classification, GECKO plots the evolution of winning organisms across the 20,000 generations (Fig. 5b). This graph can be used to identify individual *k-mers* that are essential for classification and thus worth investigating. Here we found three *k-mers* that were most frequently used by winning individuals for classification (Supplementary Table 5).

We verified the methylation status of the loci where these *k-mer* sequences were mapped using the Bismark software[24] and found that all three of them displayed dramatic changes in DNA methylation between normal and CLL samples (Fig. 5c). Interestingly the two *k-mers* that were finally selected after 20,000 generations, K107977 and K90528 overlapped binding sites for CTCF and GATA3, both of which are affected by DNA methylation status[25,26]. K107977 overlaps a CTCF-binding site for the ATP6V1G1 gene[27], which codes for a proton pump responsible for acidification of the cell, a hallmark of cancer promotion. K90528 overlaps a GATA3-binding site for the SULF2 gene that has already been identified as a diagnostic and prognostic marker in multiple cancers[28–30].

Discussion

HTS data analysis often requires extensive data transformations through tailored bioinformatics pipelines to organize the sequences in a manner that is coherent with our understanding of biology. Mapping to a reference, using ad hoc statistical thresholds and grouping sequences by functional elements, such as transcripts are common steps in most bioinformatics pipelines.

We designed GECKO with the aim of creating a classifier that could explore HTS data without a reference genome or transcriptome and without the need of bioinformatics pipelines dedicated to a specific library preparation or technology. The approach we describe here can in theory explore any type of sequencing data. Because GECKO considers groups of *k-mers* for classification, it can make use of co-dependencies between sequences to find smaller and more accurate classifiers. Thus, GECKO is capable of better classification than the commonly used approach that consists of selecting genes for which the expression is statistically significant between conditions to build a classifier (Supplementary Fig. 2). In the miRNA analysis of blood cells for example, one of the *k-mers* that participated in making an excellent classifier was not statistically significant by itself and would have been overlooked.

Using *k-mer* counts removes the requirement of a mapping step and makes GECKO applicable to numerous types of sequencing experiments. In addition, we found that using *k-mers* instead of other metrics, such as fragments per kilobase million (FPKM) or read counts resulted in higher predictive power even when run with the same genetic algorithm (Supplementary Fig. 3). This can be explained by the fact that *k-mers* can measure changes in transcription, isoform abundance, and sequence simultaneously. When applied to bisulfite converted data, each epigenetic change can potentially lead to the appearance of a novel *k-mer* in samples where the modification is present. These sample-specific *k-mers* allow GECKO to make very efficient classifications and to pinpoint the exact location of the modification.

Unlike regression analysis our approach provides multiple solutions (Supplementary Fig. 4). For research purposes this allows us to investigate why different groups of solutions work well together, explore co-dependencies between sequences and functional pathways that allow a good separation of input samples. In a clinical setting, providing multiple good solutions allows more flexibility for selecting diagnostic or prognostic targets. Importantly, the *k-mers* used for classification are not biased towards higher expressed genes (Supplementary Fig. 5) and mostly map to unique locations in the genome or transcriptome (Supplementary Fig. 6). Thus, GECKO can make use of unique transcriptional elements across a large spectrum of expression.

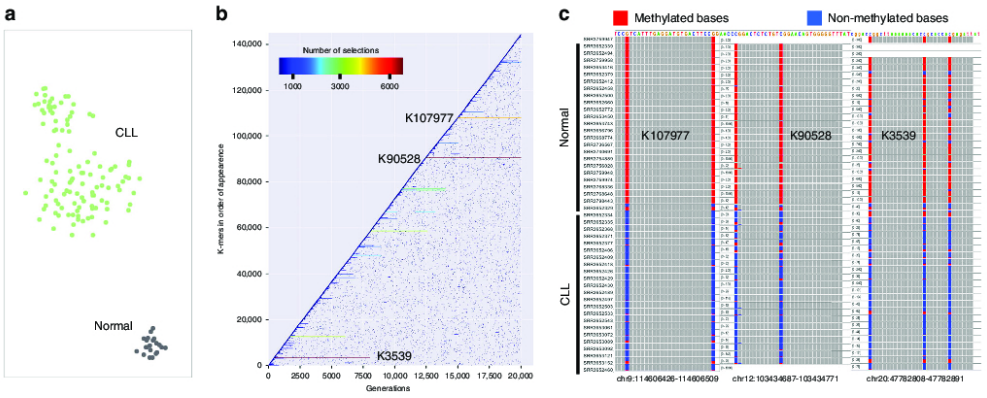


Fig. 5 GECKO can accurately classify normal and CLL patients using *k-mers* from bisulfite sequencing data. **a** GECKO output showing the t-SNE separation of CLL and normal samples using 20 *k-mers* from the winning individual. **b** GECKO output of *K-mer* exploration across 20,000 generations; *k-mers* that are frequently found in winning organisms are displayed as horizontal lines across generations; dots represent *k-mers* that were selected in one generation but eliminated in the following generation often due to a decrease in fitness of the model. **c** IGV screenshots showing the methylation status of normal and CLL samples of regions corresponding to three most frequently used *k-mers* in winning organisms determined by the Bismark software

GECKO's ability to work across multiple types of data without the need of dedicated bioinformatics tools could make it invaluable for cross-platform large-scale analyses but also for individual researchers and clinicians who would be able to compare HTS data between cohorts of patients with no bioinformatics training. It is worth noting that the longest and computationally intensive part of our procedure is obtaining the *k-mer* matrix. This step need be performed only once per dataset however and providing a *k-mer* matrix for online datasets along with sequencing files could result in widespread use of non-biased approaches such as GECKO. In addition, *k-mer*-based approaches, such as GECKO have the advantage of being portable; *k-mer* sequences will not change with new versions of the genome.

Methods

Data preparation. The *k-mer* decomposition into a matrix of *k-mer* counts is performed using Jellyfish 2[31]. This step can be preceded by a filtering of sequencing adaptors by Trim Galore (bioinformatics.babraham.ac.uk/projects/trim_galore/) if the user selects this option in GECKO. GECKO will then eliminate *k-mers* for which the count is below a noise threshold, *k-mers* that are uninformative for the given study and *k-mers* that are redundant (i.e. that share the same information as another *k-mer*).

The noise threshold is determined empirically from the input samples and is calculated for each separate run of GECKO. To do this, we count the number of times a *k-mer* count appears in one sample with null values in all other samples from the same group for the same *k-mer*. Starting at a *k-mer* count of 1, we search how many times the value 1 appears for a *k-mer* in one sample with 0 in every other sample for the same *k-mer*. We then iterate this process for *k-mer* counts 2, 3, etc. When this frequency drops dramatically as determined by the slope of frequency counts (determined by calculating the derivative at each point), we consider that we are above background and set the threshold as the *k-mer* count just before the greatest inflection of the slope (Supplementary Fig. 7).

To determine uninformative *k-mers*, that is *k-mers* that do not vary across input samples, we first discretize the *k-mer* counts using a chi-square statistic that determines the minimum number of discrete intervals with minimum loss of class attribute interdependence[32]. This algorithm is unsupervised and determines the existence and number of separate levels in continuous data. If there are no clear categories, the discretization will output a vector of 1's. Following this discretization, if there is not a minimum of 10% of samples with a different level, then this *k-mer* is considered uninformative. By default, this minimum number is set at 10% of the size of the input condition with the least replicates. For example, if the condition with the least replicates has 30 samples, then at least three samples must have a different discretized level to the other samples.

To eliminate redundant *k-mers* we use symmetric uncertainty (SU) between pairs of *k-mers*. Instead of comparing each *k-mer* to all other *k-mers*, we first split the *k-mers* into buckets of equal size and perform pairwise comparisons within a bucket. To determine which *k-mers* will be bucketed together, we calculate the sum of their counts across samples. *k-mers* with a similar sum across samples are put together; *k-mers* within a bucket have a higher chance of being redundant than if they were randomly bucketed. When all *k-mers* within buckets have been compared and redundant *k-mers* filtered, this process of bucketing by sum and filtering is repeated. This process of bucketing the *k-mers* by sum lead to 10 times faster filtering process on smaller samples and larger gains with larger matrices.

The SU between two *k-mers A* and *B* is given by the formula:

$$\mathrm{SU}(A, B) = 2\times((H(A) + H(B) - H(A, B)) \div (H(A) + H(B))]$$

where $H(A)$ and $H(B)$ are the entropies of the two *k- mers* along the samples and $H(A, B)$ is the entropy of the combined *k-mer* counts A and B along the samples.

The Entropy is given by the formula:

$$H(A) = -\sum_{i}^{G} Mi/N^{*}\log 2(Mi/N)$$

where G is the total number of *k-mer* frequencies given by the discretization step, Mi is the number of samples at the given discretization level N is the total number of samples. In our analysis, we empirically set the limit of SU at 0.7, above which two *k-mers* were considered as redundant.

GECKO keeps a record of all *k-mers* eliminated due to redundancy along with the ID of the *k-mer* that caused it to be eliminated. Thus, when the genetic algorithm finds a solution, GECKO can provide all the redundant *k-mers* that would have provided a similar solution.

All code for the data preparation was implemented in C++.

The adaptive genetic algorithm. The algorithm begins by splitting the input data into a training and test set. The test set is created by randomly selecting a number of samples from each input category. By default the number of samples selected is 1/6th of the category with the smallest amount of samples. The test set is used to establish a final test score that will have no impact on the genetic algorithm's evolution but allows us to estimate how well GECKO performs on a given dataset.

Training: At each generation of the AG, all individuals are scored based on their ability to classify the input samples using a machine learning algorithm. In this study, the algorithm used was a Linear Support Vector Classification (LinSVC). This method combines excellent results on smalls datasets and unbalanced groups with a good generalization potential, for a small computational resource cost. LinSVC is implemented in GECKO via the Scikit-learn package[33]. GECKO can also be used with a random forest model or neural networks, however these have higher computational costs and require dedicated hardware to be implemented within reasonable time-frames.

To calculate the fitness score of an individual at each generation we randomly split up the training set into two. 2/3 of the training set becomes the inner training set and the remaining 1/3 becomes the inner test set. We contrast the inner test set, which is used to score individuals at each generation of the adaptive genetic algorithm with the test set which is not used to train the adaptive genetic algorithm but instead is used to estimate the performance of our model. The inner split on the training data is random and is performed five times. The score of each individual is an average of these five iterations trained on the inner training sets and tested on the inner test sets. This rotation of the training data avoids sample batch effect biases at each generation.

Natural selection: After testing the fitness of each individual of our population we delete individuals with lower fitness scores. By default, this is 30% of the population. We call this process natural selection.

We sort the individuals by ascending rank and then apply the following probabilistic rule:

$$P - \mathrm{value} = \alpha X + \beta$$

where X is the individual rank and the following conditions are satisfied:

$$\sum_{n}^{N} P - \mathrm{value} = 1$$

$$\overset{N}{P - \mathrm{value}} = \frac{P - \mathrm{value}}{2}\overset{N/2}{P - \mathrm{value}}$$

where α, β are scalar values, N is the size of the population, and $\overset{N}{P - \mathrm{value}}$ and $\overset{N/2}{P - \mathrm{value}}$ are, respectively, the probability for the individual rank N and rank $N/2$ to be deleted.

Mutation and crossing over rates: GECKO makes use of three different types of Genetic Algorithm. These adapt the mutation and cross-over probabilities depending on the homogeneity and the performances of the population in order to converge faster and more accurately.

The three algorithms are:

A simple adaptative genetic algorithm[34]. This algorithm has a fixed factor for individuals for which the fitness is inferior to the average and a decreasing linear function for the better performing half of individuals.

Another improved adaptive genetic algorithm[35] that, similar to the simple adaptive genetic algorithm, has a crossover probability fixed above the average fitness, but uses exponential instead of the linear function for fitness values below the average.

An improved adaptive genetic algorithm[36] that models the probabilities with two linear functions, with a breakpoint for the individuals that have a fitness equal to the average fitness.

We recommend using the last model as it shows better exploration and higher convergence rates for the kind of data used for GECKO. This approach aims to maintain the population's diversity while protecting good individuals from modifications. The mutation and cross-over probabilities are decreased when the individual's fitness is high compared to the average and increased if it is low. Similarly, the probabilities are decreased when the population is heterogeneous and increased when the population is homogeneous to favor exploration of novel solutions. These probabilities are modeled by two linear functions depending on whether the individual is above the average fitness of the population or below it and is given by the formula below.

$$Pm = \begin{cases} \frac{k_1(f_{\mathrm{avg}} - f) + k_2(f - f_{\mathrm{min}})}{f_{\mathrm{avg}} - f_{\mathrm{min}}}, & f < f_{\mathrm{avg}} \\ \frac{k_2(f_{\mathrm{max}} - f) + k_3(f - f_{\mathrm{avg}})}{f_{\mathrm{max}} - f_{\mathrm{avg}}}, & f \geq f_{\mathrm{avg}} \end{cases}$$

Here f is the individual's fitness, f_{min} is the fitness of the population's worst individual, f_{avg} is the population's average fitness and f_{max} is the fitness of the population's best individual. $k1$ is the rate applied when $f = f_{\mathrm{min}}$, $k2$ when $f = f_{\mathrm{avg}}$ and $k3$ when $f = f_{\mathrm{max}}$.

Stopping criteria: By default, GECKO will run for an input number of generations. The user may however choose to make use of a stopping criteria that will stop the algorithm prematurely. The stopping criteria is checked after at least

5000 generations of the genetic algorithm. At this moment, the number of occurrences of each *k-mer* in the population is calculated across bins of 500 generations from the start of the algorithm to the current generation. The top 1% of most frequent *k-mers* in each bin are selected. We then estimate the difference in *k-mer* composition between the current bin and all previous ones using a Hamming distance. This distance measures the quantity of highest scoring *k-mers* that are changing across generations. When the slope of Hamming distance across generations drops below 1%, the stopping criteria is triggered.

Adding Gaussian noise: The user may add Gaussian noise to the model to prevent overfitting. The characteristics of this noise are determined for each *k-mer* separately. They are a mean of 0 and a standard deviation equal to the standard deviation of the *k-mer* in the training set. The user can modify the level of noise by changing noisefactor which multiplies the standard deviation by the input value. This noise is generated at each training of machine-learning model and for each individual.

tSNE visualization: t-SNE plots are generated using scikit-learn with the default parameters but initialization with PCA. This initialization option allows for better reproducibility of t-SNE graphs. Below is the corresponding command-line: manifold.TSNE (n_components = 2, init = 'pca', random_state = 0, perplexity = 30.0, early_exaggeration = 12.0, learning_rate = 200.0, n_iter = 1000, n_iter_without_progress = 300).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available from NCBI Gene Expression Omnibus under the accession numbers GSE100467 and GSE58889; the Cancer Genome Atlas under the Pan-Gyn cohort name; the database of Genotypes and Phenotypes under the accession numbers phs000435.v2.p1 and phs001050.v1.p1 but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available by submitting a request to these repositories.

Code availability

GECKO is available at https://github.com/RitchieLabIGH/GECKO under the CeCILL license.

Received: 18 December 2018 Accepted: 8 May 2019
Published online: 20 June 2019

References

1. Learn, C. A. et al. Resistance to tyrosine kinase inhibition by mutant epidermal growth factor receptor variant III contributes to the neoplastic phenotype of glioblastoma multiforme. *Clin. Cancer Res.* **10**, 3216–3224 (2004).
2. Zhang, Z.-M. et al. Pygo2 activates MDR1 expression and mediates chemoresistance in breast cancer via the Wnt/β-catenin pathway. *Oncogene* **35**, 4787–4797 (2016).
3. Martín-Martín, N. et al. Stratification and therapeutic potential of PML in metastatic breast cancer. *Nat. Commun.* **7**, 12595 (2016).
4. Audoux, J. et al. DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. *Genome Biol.* **18**, 243 (2017).
5. Kirk, J. M. et al. Functional classification of long non-coding RNAs by k-mer content. *Nat. Genet.* **1**, https://doi.org/10.1038/s41588-018-0207-8 (2018).
6. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mer0s. *BMC Genom.* **16**, 236 (2015).
7. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* **19**, 198 (2018).
8. Sergeyev, Y. D., Kvasov, D. E. & Mukhametzhanov, M. S. On the efficiency of nature-inspired metaheuristics in expensive global optimization with limited budget. *Sci. Rep.* **8**, 453 (2018).
9. Juzenas, S. et al. A comprehensive, cell specific microRNA catalogue of human peripheral blood. *Nucleic Acids Res.* **45**, 9290–9301 (2017).
10. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).
11. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
12. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304. e6 (2018).
13. Maziveyi, M. & Alahari, S. K. Breast cancer tumor suppressors: a special emphasis on novel protein nischarin. *Cancer Res.* **75**, 4252–4259 (2015).
14. Hasim, M. S., Nessim, C., Villeneuve, P. J., Vanderhyden, B. C. & Dimitroulakos, J. Activating transcription factor 3 as a novel regulator of chemotherapy response in breast cancer. *Transl. Oncol.* **11**, 988–998 (2018).
15. Gijn, S. E. van et al. TPX2/Aurora kinase A signaling as a potential therapeutic target in genomically unstable cancer cells. *Oncogene* **1**, https://doi.org/10.1038/s41388-018-0470-2 (2018).
16. Choi, J. et al. Loss of KLHL6 promotes diffuse large B-cell lymphoma growth and survival by stabilizing the mRNA decay factor roquin2. *Nat. Cell Biol.* **20**, 586–596 (2018).
17. Solari, N. E. F. et al. The NSL chromatin-modifying complex subunit KANSL2 regulates cancer stem-like properties in glioblastoma that contribute to tumorigenesis. *Cancer Res.* **76**, 5383–5394 (2016).
18. Tatarskiy, V. V. et al. Stability of the PHF10 subunit of PBAF signature module is regulated by phosphorylation: role of β-TrCP. *Sci. Rep.* **7**, 5645 (2017).
19. Goetz, M. P. et al. Tumor sequencing and patient-derived xenografts in the neoadjuvant treatment of breast cancer. *J. Natl. Cancer Inst.* **109**, 7 (2017).
20. Thomas, S. J., Snowden, J. A., Zeidler, M. P. & Danson, S. J. The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. *Br. J. Cancer* **113**, 365–371 (2015).
21. Sapio, R. T. et al. Inhibition of post-transcriptional steps in ribosome biogenesis confers cytoprotection against chemotherapeutic agents in a p53-dependent manner. *Sci. Rep.* **7**, 9041 (2017).
22. Podojil, J. R. & Miller, S. D. Potential targeting of B7-H4 for the treatment of cancer. *Immunol. Rev.* **276**, 40–51 (2017).
23. Landau, D. A. et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26**, 813–825 (2014).
24. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
25. Wang, H. et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22**, 1680–1688 (2012).
26. Fleischer, T. et al. DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat. Commun.* **8**, 1379 (2017).
27. Lesurf, R. et al. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res.* **44**, D126–D132 (2016).
28. Alhasan, S. F. et al. Sulfatase-2: a prognostic biomarker and candidate therapeutic target in patients with pancreatic ductal adenocarcinoma. *Br. J. Cancer* **115**, 797–804 (2016).
29. Rosen, S. D. & Lemjabbar-Alaoui, H. Sulf-2: an extracellular modulator of cell signaling and a cancer target candidate. *Expert Opin. Ther. Targets* **14**, 935–949 (2010).
30. Lui, N. S. et al. SULF2 expression is a potential diagnostic and prognostic marker in lung cancer. *PLoS ONE* **11**, e0148911 (2016).
31. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
32. Gonzalez-Abril, L., Cuberos, F. J., Velasco, F. & Ortega, J. A. Ameva: an autonomous discretization algorithm. *Expert Syst. Appl.* **36**, 5327–5332 (2009).
33. Pedregosa, F. et al. Scikit-learn: machine learning in python. ArXiv12010490 Cs (2012).
34. Zhang, J., Chung, H. S. H. & Hu, B. J. Adaptive probabilities of crossover and mutation in genetic algorithms based on clustering technique. In *Proc. 2004 Congress on Evolutionary Computation* (ed Greenwood, G. W.) (*IEEE Cat. No. 04TH8753*), Vol. 2, 2280–2287 (IEEE Portland, OR, USA, USA, 2004).
35. Ravindran, S., Jambek, A. B., Muthusamy, H. & Neoh, S.-C. A novel clinical decision support system using improved adaptive genetic algorithm for the assessment of fetal well-being. *Comput. Math. Methods Med.* **2015**, 283532 (2015). https://doi.org/10.1155/2015/283532.
36. Yan, M. et al. Improved adaptive genetic algorithm with sparsity constraint applied to thermal neutron CT reconstruction of two-phase flow. *Meas. Sci. Technol.* **29**, 55404 (2018).

Acknowledgements

We wish to acknowledge the Genotoul platform (genotoul.fr) for providing us with calculation time on their servers. We thank Jerome Audoux, Jean-Philippe Villemin and Giacomo Cavalli for their advice. We wish to acknowledge the Agence Nationale de la Recherche (ANRJCJC-WIRED), the Labex EpiGenMed and the MUSE initiative for their financial support.

iMOKA: κ-mer based software to analyze large collections of sequencing data

iMOKA, interactive Multi Objective κ-mer Analysis, was initially thought as a filter to select the informative κ-mers: most of the κ-mers selected by GECKO were able individually to classify with relatively good accuracy the samples in the respective groups, even using a cross-validation procedure. As for GECKO, the details of the algorithm are described in the paper below, including a benchmark on four datasets in which the κ-mers extracted by iMOKA are compared to PSI values, gene and transcript expression as classifying features in a Random Forest classifier model.

In brief, the software can take as input both sequencing files, such as fastq or bam, or external link, HTTP, FTP or SRR ids, downloading the required data before the beginning of the analysis.
Using KMC3, iMOKA extracts the sorted κ-mer counts from each sample and converts them into binary files. A JSON file contains the metadata of the samples belonging to the analysis, including for each sample: the name, the label of the group, the location of the binary file and the total sum of the κ-mer counts, used to normalize the data.
The first step of reduction considers one κ-mer at the time and, using a Bayes Classifier, estimates the accuracy of the feature to classify the samples in the respective groups. This step is by default coupled to an adaptive entropy filter that speeds up the process discarding few truly informative features.
Finally, an aggregation procedure groups the κ-mers based on their sequence, building de Bruijn graphs, and their biological relevance, mapping the sequences generated from the graphs on a reference genome and using a reference annotation to assign “events” to the most informative κ-mers in each group.

Importantly, the software is coupled with a graphical user interface (GUI) that allows running in local or on a remote cluster all the steps of the algorithm. The user can also explore the final output of the aggregation step as an interactive table, visualize the κ-mers mapping on a reference genome with a javascript version of IGV genome browser, generate self-organizing maps and Random forest classifiers.

The key novelties of iMOKA are represented by:

1. The scalability: the κ-mer matrix is generated on the flight by combining the κ-mer counts of each sample, stored independently in sorted binary files. No matter how many samples there are in input, iMOKA adapts to the user-defined RAM limits and is going to keep in memory only a small buffer for each column, allowing it to run the first step of the algorithm with few resources. This representation is compact since the “zero” values are

represented by the absence of a determined κ-mer in a sample, solving the issue of sparsity. Furthermore, thanks to its flexible structure, the same sample can be used in different studies. Currently, the aggregation step doesn't consider the available memory and it could require a large amount of RAM in case of numerous κ-mers in input.

2. The reduction step is based on a machine learning procedure and not on a statistical test. This method, though slower with respect to the differential expression analysis, used for example in DE-kupl, is robust to outliers and scales efficiently with the number of samples.
3. The aggregation step reduces the redundancy based not only on the sequence but also on its biological meaning. Furthermore, the software assigns different types of events, such as mutations, indel, splice, alternative splice and DE based on the information obtained by the alignment and the gene annotations.
4. Finally, the GUI is an uncommon feature for bioinformatics tools and it's useful to interactively explore the results, visualizing not only the individual κ-mers but also the context in which it resides.

A dedicated k-mer structure

The selection of a performant and compacted data structure to store and access the *k*-mer sequences and abundances must be aware of the application and the context required by the software.
The literature offers detailed reviews about techniques to store and query a set of *k*-mers[235] and large collections of sequencing data sets[236].
Designing iMOKA we were looking for a data structure able to dynamically generate a *k*-mer matrix, to store efficiently the sample's *k*-mer counts and to load only small portions of the files in order to handle large datasets on virtually any architecture.
We implemented a prefix-suffix structure similar to the one used as database format in the first version of the k-mer counter software KMC[237]. Rather than using two files (*.kmc_pre* and *.kmc_suf*), we store both the prefix and suffix information in the same binary file. The prefix data contains, for each prefix:

- The binary encoded DNA symbols (A=00, C = 01, G = 10, T=11) of *p* stored as *char* values.
- The position in the suffix array that corresponds to the first suffix associated with the *p*, stored as *uint64*. To know the range of the suffixes of *p* is therefore sufficient to retrieve the position of the first suffix of *p+1*.

Similarly, the suffix data contains the binary encoded DNA symbols of the suffixes and the related counts, stored as *uint32*.
The length of the prefix is chosen according to the total number of *k*-mers present in the dataset, following a formula adapted by A. Mancheron in [238] and described in the article.
This type of data structure allows loading small buffers of suffix and prefix data at the time.

Furthermore, prefixes and suffixes are sorted lexicographically, which allows generating the k-mer matrix dynamically using an n-way merge algorithm:

1. Each sample's database loads a buffer of prefixes and suffixes.
2. A pointer is associated with the first k-mer in each database.
3. A copy of the k-mer associated with the pointers is stored in a sorted set S.
4. A pop operation retrieves the lexicographically smallest *k*-mer from S, k_n, which will correspond to the current row of the k-mer matrix.
5. For each sample m, if the pointer corresponds to k_n, the count of k_n is assigned to m in the current row and the pointer moves one position forward, updating S, otherwise the count of k_n is 0.
6. Repeat from 4 until all the databases are empty, refilling the buffers when needed.

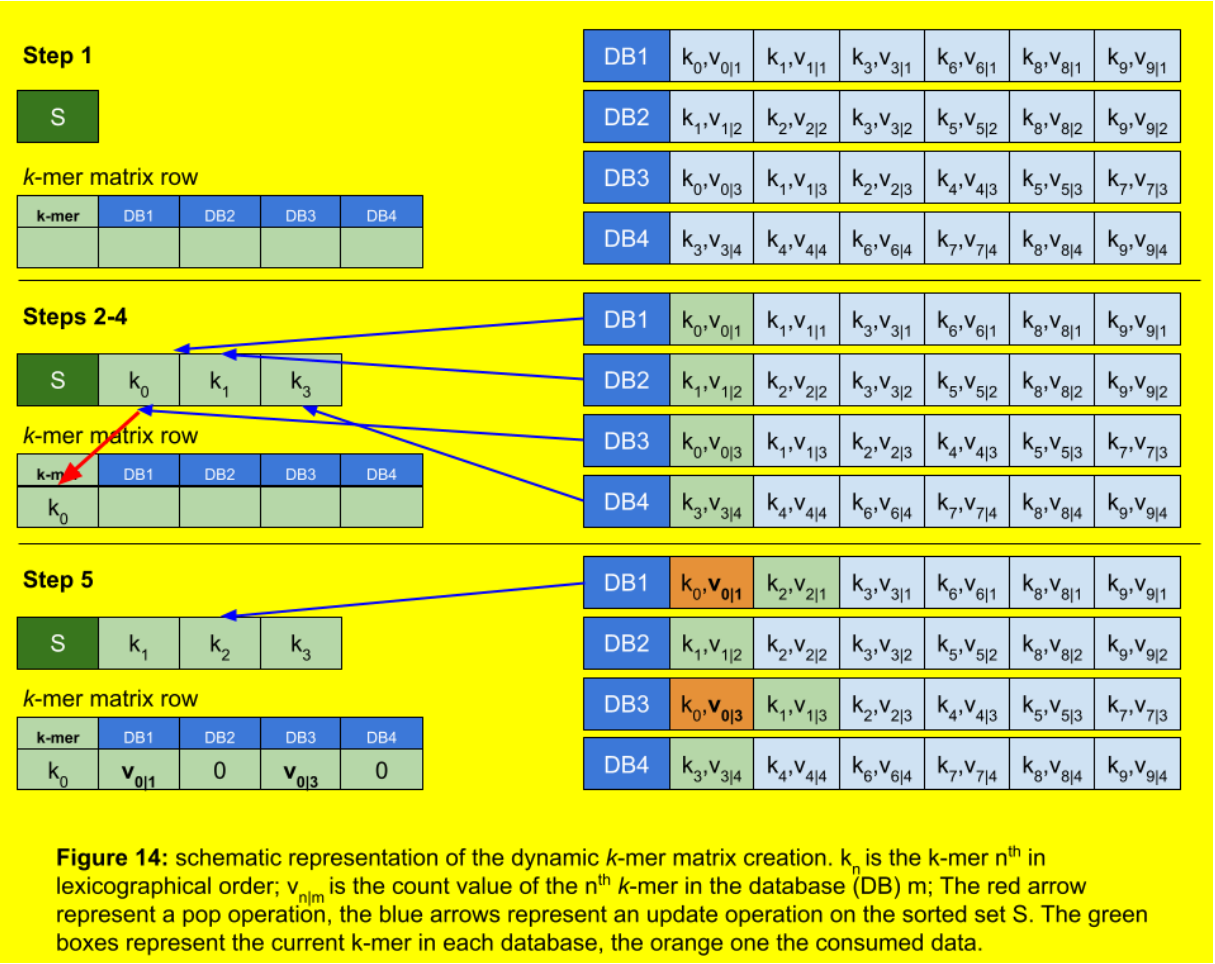


Figure 14: schematic representation of the dynamic *k*-mer matrix creation. k_n is the k-mer n^{th} in lexicographical order; $v_{n|m}$ is the count value of the nth *k*-mer in the database (DB) m; The red arrow represent a pop operation, the blue arrows represent an update operation on the sorted set S. The green boxes represent the current k-mer in each database, the orange one the consumed data.

Finally, to allow a multithreading generation of the matrix it's possible to generate it starting from any *k*-mer k_n: using a binary search, each database can be initialized to k_n or, if absent, to the closest *k*-mer following k_n.
Graphs based data structures, such as de Bruijn graphs, offer great querying performances and allow to quickly query for *k*-mers in the neighbour nodes but we didn't consider them since the reduction step of iMOKA doesn't require a

navigational data structure since each feature is considered independently from the others.
Most of the recent k-mer counting tools, such as Jellyfish[203] or KMC3[207], use Hash Tables (HT) or Bloom Filter (BF) to store the k-mer counts.
Though optimal for querying and modification operations, those data structures don't store the *k*-mers in sorted order and require loading the full index in memory or performing frequent disk reading operations compared to the prefix-suffix structure aforementioned to produce a *k*-mer matrix.

Lorenzi et al. Genome Biology (2020) 21:261
https://doi.org/10.1186/s13059-020-02165-2

Genome Biology

METHOD Open Access

iMOKA: *k*-mer based software to analyze large collections of sequencing data

Claudio Lorenzi[1], Sylvain Barriere[1], Jean-Philippe Villemin[1], Laureline Dejardin Bretones[1], Alban Mancheron[2] and William Ritchie[1*]

* Correspondence: william.ritchie@igh.cnrs.fr
[1]IGH, Centre National de la Recherche Scientifique, University of Montpellier, Montpellier, France
Full list of author information is available at the end of the article

Abstract

iMOKA (interactive multi-objective *k*-mer analysis) is a software that enables comprehensive analysis of sequencing data from large cohorts to generate robust classification models or explore specific genetic elements associated with disease etiology. iMOKA uses a fast and accurate feature reduction step that combines a Naïve Bayes classifier augmented by an adaptive entropy filter and a graph-based filter to rapidly reduce the search space. By using a flexible file format and distributed indexing, iMOKA can easily integrate data from multiple experiments and also reduces disk space requirements and identifies changes in transcript levels and single nucleotide variants. iMOKA is available at https://github.com/RitchieLabIGH/iMOKA and Zenodo https://doi.org/10.5281/zenodo.4008947.

Keywords: *k*-mer, NGS analysis, Personalized medicine, Bioinformatics software, Data reduction, Machine learning

Background

Studies of variation in gene expression have considerably advanced knowledge of disease etiology and classification [1–3]. To capitalize on genomic data generated from numerous clinical studies, recent initiatives have aggregated high-throughput sequencing (HTS) experiments from multiple cohorts that measure gene expression, RNA isoform usage, and genome variation. For example, the Genomic Data Commons program controls access to over 84,000 cases [4]. Still, despite these efforts to aggregate and provide data from multiple studies, their computational analysis and integration presents a major challenge; each type of HTS data requires specific bioinformatics pipelines that need to be implemented by a bioinformatics specialist. In addition, most of these approaches require reference genomes or transcriptomes and thus cannot inherently account for the diversity in non-reference transcripts or individual variations [5]. To alleviate the requirement of a reference, recent methodologies use *k-mer* representation; they directly compare the counts of nucleotide sequences of length *k* between samples [6]. These *k*-mer based approaches have been core to the field of metagenomics, where they are used to discover unique *k*-mers or *k*-mer signatures to classify organisms [7, 8]. However, when translated to mammalian genomes, *k*-mer

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

representation results in a *k*-mer count matrix with as many columns as there are samples and as many rows as there are *k*-mers, generally billions. Exploring such large matrices to find biologically relevant *k*-mers is intractable unless the analysis focuses only on a very small subset of the sequencing data [5] or by using metaheuristics that provide partial solutions [9].

Here we present iMOKA (interactive multi-objective *k*-mer analysis), a novel approach and software that allows non-specialists to make use of *k*-mers to explore large amounts of mammalian sequencing data. This approach is agnostic of the type of sequencing data used, is not biased towards annotated genetic elements, and can analyze transcript levels and single nucleotide variations in one pass. Importantly, iMOKA is interactive; it allows the user to import and merge samples from different studies and tailor their exploration of *k*-mers to specific genomic elements of interest such as splicing events, mutations, or global gene expression. We tested iMOKA on four clinical datasets: the classification of breast cancer subtypes and response to chemotherapy of breast, ovarian cancer, and diffuse large B cell lymphoma (DLBCL). We find that iMOKA found features that are more accurate than classical bioinformatics approaches, takes up less space, uses less memory, has faster runtimes, and can be run on a computer cluster or on a laptop.

Results

IMOKA design

iMOKA imports sequencing files in FASTQ, FASTA, BAM format, or SRR identifiers via its user interface. It then counts the occurrences of all sequences of given length *k* (default 31) [9] using the KMC3 software [10] in each sample (Fig. 1). It then extracts labels from the sequencing metadata so that the user can define groups they wish to compare. Importantly, each sample is stored as a sorted vector of *k*-mer counts in a dedicated binary file using a custom prefix-suffix structure that drastically reduces the disk space requirements ("Methods" section). For each sample, a JSON file is created that contains metadata and a rescaling factor for *k*-mer count normalization that allows the user to remove or add samples without having to recalculate an entire *k*-mer matrix. It then uses our feature reduction step that combines a Bayes classifier augmented by an adaptive entropy filter to rapidly remove non-relevant *k*-mers (Fig. S1). The aim of this filter is to evaluate each *k*-mer individually by combining the accuracy of the Bayes classifier with the speed of calculating Shannon's entropy. This evaluation is performed using a Monte Carlo cross validation with a high number of iterations and an early break ("Methods" section) that efficiently reduces overfitting and generates predictions that overcome batch effects. In order to reduce the number of features evaluated, the entropy filter works simultaneously and, learning from the entropies of the *k*-mers that successfully passed the accuracy filter, discards *k*-mers with low entropy. Following this filtering, *k*-mers for which the sequences overlap are assembled into graph structures. These are used to aggregate the *k*-mers that are likely to have been generated from the same biological sequence and are used to eliminate false positive *k*-mers that are mainly singletons (1 *k*-mer) or very short branches in the graph structure. Bifurcations or bubbles in these graphs generally arise from the existence of multiple sequence isoforms that differ by point mutations or alternative splicing events [11]. By

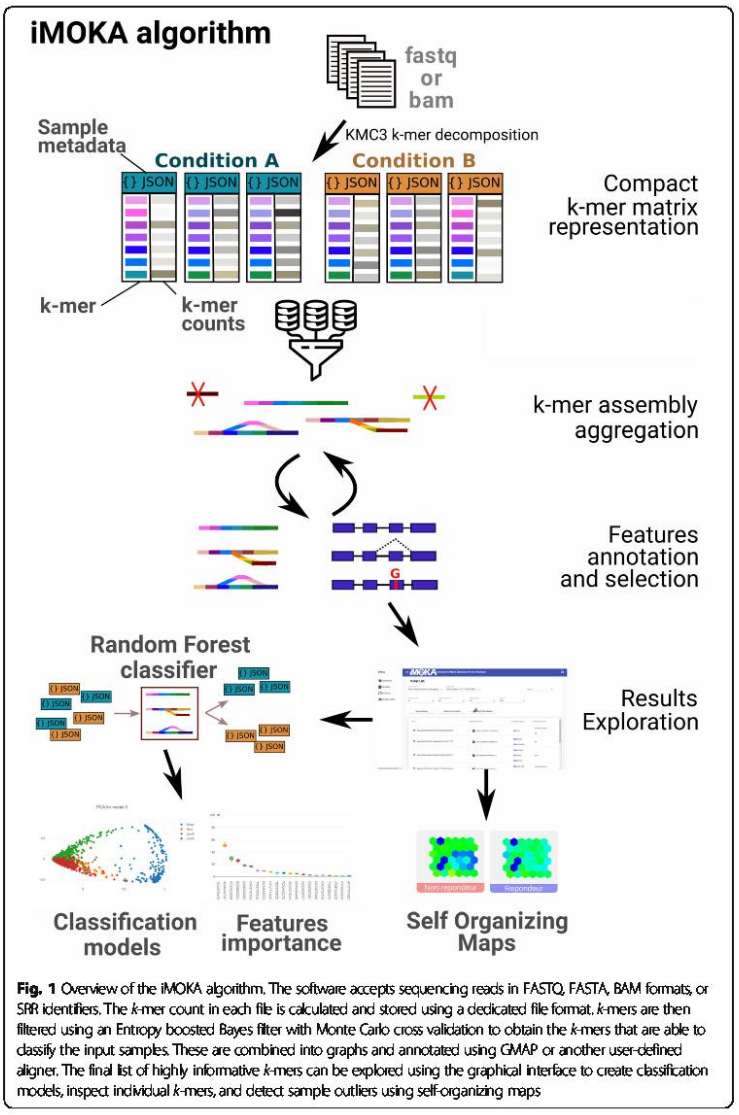


Fig. 1 Overview of the iMOKA algorithm. The software accepts sequencing reads in FASTQ, FASTA, BAM formats, or SRR identifiers. The *k*-mer count in each file is calculated and stored using a dedicated file format. *k*-mers are then filtered using an Entropy boosted Bayes filter with Monte Carlo cross validation to obtain the *k*-mers that are able to classify the input samples. These are combined into graphs and annotated using GMAP or another user-defined aligner. The final list of highly informative *k*-mers can be explored using the graphical interface to create classification models, inspect individual *k*-mers, and detect sample outliers using self-organizing maps

combining this graph assembly with the relatively permissive Bayesian filter, we are able to generate a list of informative *k*-mers in a manner that is fast and accurate.

iMOKA allows the user to align the *k*-mer graphs to a reference genome to annotate them with known genomic features such as known RNA transcripts, point mutations, or mRNA splicing events. iMOKA provides a random forest classifier that uses filtered *k*-mer graphs as features (Supplementary methods) and provides the user with a

classification model and a sorted list of *k*-mer graphs that were most used in the tree models and that are thus of higher interest (Fig. 1). The user may even build classification models based solely on specific genomic features such as point mutations or gene expression for example. Finally, iMOKA uses self-organizing map clustering on the *k*-mer graphs to enable users to identify subgroups or outliers amongst their input samples.

Benchmarking datasets and algorithms

iMOKA uses a *k*-mer based analysis to detect sequence features and create classification models from large cohorts of mammalian RNA sequencing data. To test its performance, we selected four studies that were distinct in their data structures, classification objectives, and sizes. The first was a non-binary classification of 1038 patients aiming to define 4 subtypes of breast cancer which were luminal A (LumA), luminal B (LumB), HER2-enriched (HER2), and basal-like. The second was a cohort of 240 ovarian cancer patients where the objective was to predict response to chemotherapy. The third was a smaller cohort of 118 breast cancer patients where the objective was also to predict response to chemotherapy. The last was an even smaller cohort of 17 DLBCL patients divided according to their responsiveness to the chemotherapy.

In our benchmark, we included methods based on four different types of features which were *k*-mer counts, percentage-spliced-in (PSI), transcripts per kilobase million (TPM), and sequencing counts. The two latter were measured and tested across annotated genes and transcripts separately. The algorithms we benchmarked were DESeq2 [12], edgeR [13], and limmaVoom [14] for TPM and sequencing counts; iMOKA for *k*-mer counts; and Whippet [15] for alternative splice site usage. We excluded four other *k*-mer based methods HAWK [16], KOVER [17], Kissplice [11], and GECKO [9] because they were respectively impossible to run on such big datasets due to segmentation fault errors, were unable to find *k*-mers that could classify the input samples or, for the last two methods, were killed after 2 weeks of runtime on our computer cluster.

In our benchmark, we compared the list of features output by each algorithm by using them in a random forest classifier and determining their out of bag scores (OOB score). The out of bag score tests how well each classifier performs without having to set aside a portion of the data specifically as a test set. It is as reliable as using a test set [18, 19] without having to set aside part of the data. We chose the random forest classifier because it is a non-parametric approach and because the importance of each input feature is easy to evaluate.

Finally, for the largest dataset, the molecular classification of breast cancer, we performed a 5-fold cross validation of the entire iMOKA procedure and all other benchmarked algorithms, using 4/5 of the dataset for data reduction and creation of a random forest model and 1/5 of the dataset as the test set.

Classification of breast cancer subtypes

Breast cancer is a transcriptionally heterogeneous disease with multiple subtypes that determine prognosis, treatment, and patient outcome. Although breast cancer classification is constantly being updated, a broadly accepted stratification defines four groups

which are luminal A (LumA), luminal B (LumB), HER2-enriched (HER2), and basal-like [20]. We benchmarked iMOKA on a dataset of 1038 mRNA-Seq breast cancer samples from the Cancer Genome Atlas (TCGA) Pan-Gyn cohort [21] (patients per class: basal 190, Her2 82, LumA 559, LumB 207) and tested how well the outputs of each approach could accurately predict the four classes. We found that the list of *k*-mers output by iMOKA (Additional file 1, Fig. S5) was above all other methods in their ability to classify the four types of breast cancer (Fig. 2a). The worst performing features were the splice site usage statistics given by Whippet. This could be expected because the breast cancer stratifications were originally created using gene expression profiles, not splicing events.

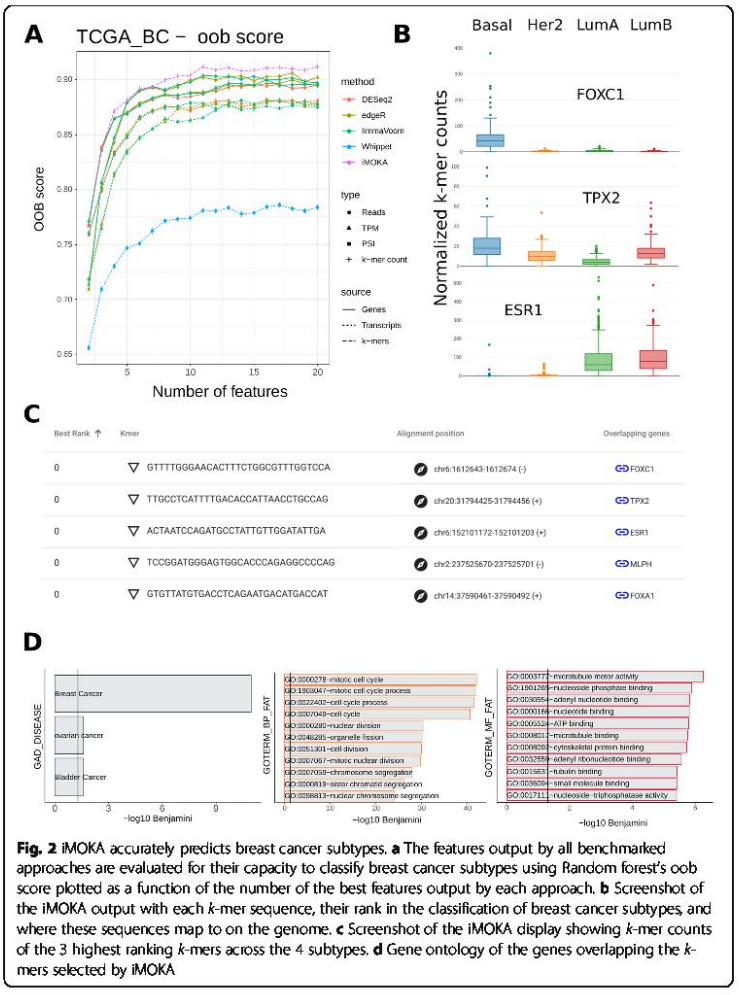


Fig. 2 iMOKA accurately predicts breast cancer subtypes. **a** The features output by all benchmarked approaches are evaluated for their capacity to classify breast cancer subtypes using Random forest's oob score plotted as a function of the number of the best features output by each approach. **b** Screenshot of the iMOKA output with each *k*-mer sequence, their rank in the classification of breast cancer subtypes, and where these sequences map to on the genome. **c** Screenshot of the iMOKA display showing *k*-mer counts of the 3 highest ranking *k*-mers across the 4 subtypes. **d** Gene ontology of the genes overlapping the *k*-mers selected by iMOKA

We additionally performed a 5-fold cross validation of the entire iMOKA procedure and all other benchmarked algorithms including feature reduction and model generation. The accuracies of the final models (Fig. S2) show a consistent behavior to the oob scores in Fig. 2a.

iMOKA identified 3002 *k*-mers overlapping different types of events (Table S1 and Additional file 1). Using iMOKA's interface, we were able to explore the genes to which these *k*-mers mapped (Fig. 2b). As expected, within the best ranking *k*-mers, iMOKA found overlaps with genes that have been extensively linked to breast cancer subtypes and are already used in the clinic such as estrogen receptor 1 (ESR1) [22], Forkhead Box A1 (FOXA1) [23], Forkhead Box C1 (FOXC1) [24], xenopus kinesin-like protein 2 (TPX2) [25], and Melanophilin (MLPH) [26]. By clicking on the *k*-mer sequence in the iMOKA interface, we can visualize the representation of each *k*-mer in the 4 classes (Fig. 2c). The top three *k*-mers, whose gene expression is shown in Fig. S3, have representation profiles that clearly explain iMOKA's high classification accuracy with a small number of *k*-mers.

It is worth noting that iMOKA picked up 120 potential alternative splicing events. Amongst these were 4 extensively studied splicing isoforms (MYO6, TPD52, IQCG, and ACOX2) [27] identified to be amongst the 5 most important isoforms differentially expressed between ER+HER2− and ER-HER2 primary breast tumors (Fig. S4).

Finally, we used DAVID [28] to perform a functional annotation of the genes overlapping the *k*-mer selected by iMOKA. The gene list is strongly enriched for breast cancer-associated genes and of genes associated with the function commonly dysregulated in cancer cells, such as cell cycle, cell division, and motility (Fig. 2d and Additional file 4).

iMOKA identifies events associated with the response to treatment in ovarian cancer patients

Our second benchmark was performed on a dataset of high-grade serous ovarian cancers taken from the TCGA_OV cohort [29]. We included patients having an annotated [30] response to a first-line treatment to the combination platinum and taxane chemotherapy (patients per class: 174 responsive, 66 non-responsive). iMOKA identified 138 *k*-mers with individual accuracy between 65 and 75% (Table S1 and Additional file 2). Again, the *k*-mers found by iMOKA gave the most accurate oob scores for response to chemotherapy (Fig. 3a). The gain compared to other methods is much higher than for the previous breast cancer classification. This can be explained by the fact that most of the methods we benchmark against only make use of gene or transcript expression or splicing sites. Breast cancer stratification is mainly based on gene expression, and therefore, these methods compare well with iMOKA. However, in the case of response to chemotherapy in ovarian cancer, iMOKA is able to also make use of single nucleotide variants (SNVs) and splice site usage to make its predictions (Fig. 3b). Via the iMOKA interface, we can visualize the SNVs with the highest feature importance. Thus, we can observe that iMOKA detected a known nonsense mutation (SNP id: rs10794537) in the alpha-L-iduronidase (IDUA) gene. IDUA is responsible for the degradation of the mucopolysaccharides, heparan sulfate, and dermatan sulfate

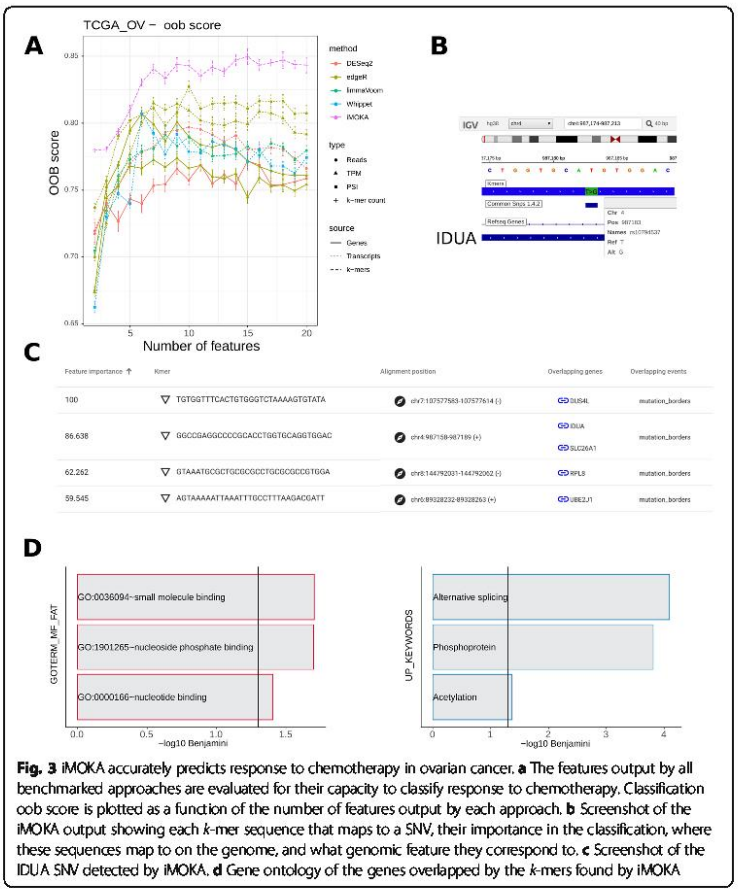


Fig. 3 iMOKA accurately predicts response to chemotherapy in ovarian cancer. **a** The features output by all benchmarked approaches are evaluated for their capacity to classify response to chemotherapy. Classification oob score is plotted as a function of the number of features output by each approach. **b** Screenshot of the iMOKA output showing each *k*-mer sequence that maps to a SNV, their importance in the classification, where these sequences map to on the genome, and what genomic feature they correspond to. **c** Screenshot of the IDUA SNV detected by iMOKA. **d** Gene ontology of the genes overlapped by the *k*-mers found by iMOKA

which modulate angiogenesis, cell invasion, metastasis, and inflammation [26] and importantly are ligand receptors for polynuclear platinum anticancer agents [27]. In agreement with this, the gene ontology (Fig. 3d) analysis shows a functional enrichment of small molecule binding proteins.

iMOKA identifies events associated with the response to neoadjuvant chemotherapy in breast cancer patients

The third test dataset was taken from the Breast Cancer Genome Guided Therapy (BEAUTY) study [31] and consisted of patients with all 4 types of breast cancer for which we tested the response to neoadjuvant chemotherapy with paclitaxel and anthracycline. This allowed us to test the binary classification of more heterogeneous cell populations on smaller sample sizes: 36 patients that had a complete response to chemotherapy and 82 that did not. It is worth noting that this dataset presented a

significant batch effect, detected using the R package DASC [32], associated with the load date of the samples (Fig. S5). Despite this, iMOKA identified 1248 *k*-mers with an individual accuracy between 70 and 83.8% (Table S1 and Additional file 3). Again, the *k*-mers discovered by iMOKA give the highest oob scores for the response to chemotherapy (Fig. 4a).

Our method can identify multiple events on the same gene that are useful for classification. For example, as shown in Fig. 4b for the highest scored *k*-mers overlapping the gene TBC1D9, iMOKA discovers that the gene as a whole is differentially expressed between conditions but also discovers alternatively expressed introns (Fig. 4c) that were confirmed as being a retained intron using a dedicated algorithm, IRFinder [33].

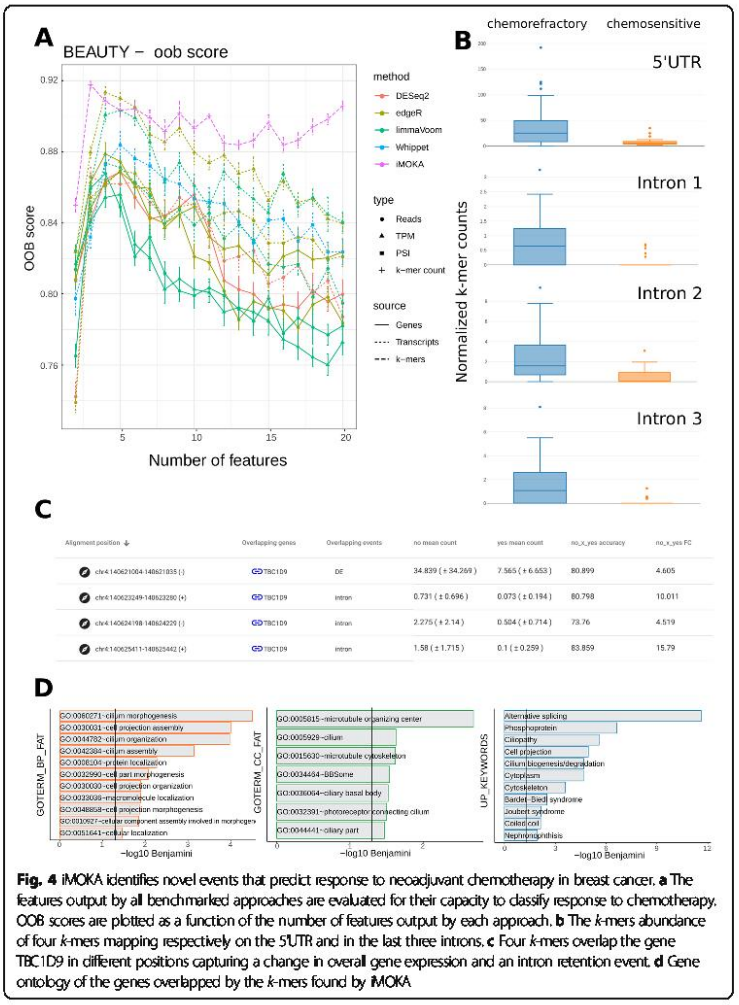


Fig. 4 iMOKA identifies novel events that predict response to neoadjuvant chemotherapy in breast cancer. **a** The features output by all benchmarked approaches are evaluated for their capacity to classify response to chemotherapy. OOB scores are plotted as a function of the number of features output by each approach. **b** The *k*-mers abundance of four *k*-mers mapping respectively on the 5'UTR and in the last three introns. **c** Four *k*-mers overlap the gene TBC1D9 in different positions capturing a change in overall gene expression and an intron retention event. **d** Gene ontology of the genes overlapped by the *k*-mers found by iMOKA

The gene ontology analysis of the genes overlapping the *k*-mers selected by iMOKA reveals a strong relationship with microtubules and cilia, components influenced by paclitaxel [34, 35], an anti-microtubule agent of the taxane family used as part of the therapy on all the patients in the study. Although the study included heterogeneous cancer types and an unbalanced dataset, iMOKA was able to detect features useful for classification.

iMOKA identify DE genes associated with DLBCL chemoresistance

In the last dataset, we tested iMOKA in a frequent scenario where differential representation of transcripts is assessed in a very small cohort. To this end, we considered 17 DLBCL patients [36], 10 responsive to an anthracycline-based regimen R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone) and 7 non-responsive. The RNA-seq used for this dataset is targeted, making it impossible to evaluate the PSI values, so only the abundance of the genes and transcripts were considered in the benchmark (Fig. 5 and Fig. S7). iMOKA identified 1928 *k*-mers having an individual accuracy over 80% and five with 100% accuracy. They corresponded to the genes AKT1, BTBD9, ZBTB45, ZBTB17, and BHLHE40. Amongst those, AKT1 is known to play a role in DLBCL chemosensitivity [37] but was not detected as differentially expressed in the original publication [36].

This study highlights another advantage of using *k*-mers; they are agnostic to transcript annotation. For example, the *k*-mer overlapping ZBTB17, a gene involved in B cell development and differentiation [38], is located on the splicing site at position chr1:15,947,123-15,948,295 and is part of Refseq transcript NM_001242884. However, this transcript was not annotated in the GENCODE annotation (Fig. 5b) and thus not detected by salmon.

iMOKA runtimes and disk space

iMOKA was designed to be scalable; the user can control the number of threads used and the dedicated RAM, allowing the software to run not only on HPC clusters, but also on a laptop. In Fig. 6, we report the times to analyze three experiments described in the previous sections on a computer with 8-cores and 32 GiB of RAM. Importantly, the higher the number of samples in the cohort, the bigger iMOKA's gains are.

iMOKA's most intensive task is the generation of informative *k*-mers, where a large amount of data is filtered and aggregated, while the other benchmarked approaches handle data that are already filtered (reads are already mapped to annotated regions). Finally, most methods that calculate differential expression are designed for relatively small cohorts and do not scale well in memory with large cohorts: DESeq2 and edgeR for example required additional RAM in order to analyze the differential expressed transcripts in the TCGA BRCA (TCGA_BC) analysis (61 GiB and 46 GiB, respectively) (Fig. 6).

Discussion

Recent efforts to aggregate and annotate patient HTS data should facilitate our understanding of health trajectories through multiple molecular mechanisms. In theory,

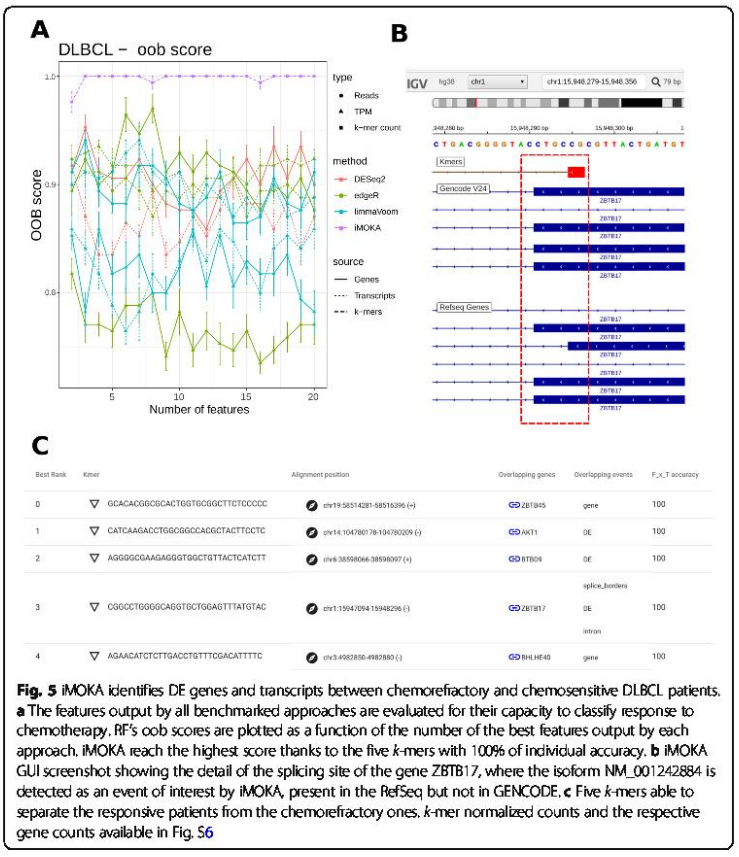


Fig. 5 iMOKA identifies DE genes and transcripts between chemorefractory and chemosensitive DLBCL patients. **a** The features output by all benchmarked approaches are evaluated for their capacity to classify response to chemotherapy. RF's oob scores are plotted as a function of the number of the best features output by each approach. iMOKA reach the highest score thanks to the five *k*-mers with 100% of individual accuracy. **b** iMOKA GUI screenshot showing the detail of the splicing site of the gene ZBTB17, where the isoform NM_001242884 is detected as an event of interest by iMOKA, present in the RefSeq but not in GENCODE. **c** Five *k*-mers able to separate the responsive patients from the chemorefractory ones. *k*-mer normalized counts and the respective gene counts available in Fig. S6

combining gene expression, isoform usage and single nucleotide variation should allow for more nuanced stratification and prediction of disease etiology. However, HTS data analysis often requires extensive data transformations that are often performed with little transverse coherence; each type of analysis produces lists of features that pass a

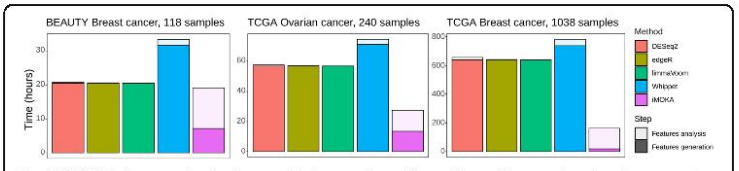


Fig. 6 iMOKA is faster and scales better with large cohorts. Comparison of the running times between the benchmarked methods. Solid bars represent the time dedicated to the generation of the features (transcript abundance, PSI evaluation, and *k*-mer count), and the lighter bars represent the time dedicated to the analysis of the features (differential expression, differential splicing, and the machine learning-based filters)

given test and these are then analyzed separately. Mapping to a reference, using ad hoc statistical thresholds for each type of analysis, and grouping sequences by functional elements are common steps in bioinformatics pipelines that may not reflect the complex interaction between each of the processes that make up an individual's transcriptome.

We designed iMOKA with the aim of analyzing HTS data in the reverse manner; we wished to first discover all sequences that were informative, group them according to how well they could classify the input samples, and then break them down into the different components of gene expression, isoform representation, and SNV presence. In doing so, we created a classifier that could explore HTS data without a reference genome or transcriptome and without the need of dedicated bioinformatics pipelines for each type of transcriptional event.

Using *k*-mer counts removes the requirement of a mapping step and allows iMOKA to explore and combine multiple transcriptional events to make more accurate predictions and to explore all these events simultaneously without having to apply multiple pipelines. *k*-mers can measure changes in transcription, isoform abundance, and sequence simultaneously and were thus able to create better predictive models than other metrics such as transcripts per million (TPM), read counts, or splice site usage.

By creating a reliable, cross-platform user interface, iMOKA allows non-specialists to leverage the predictive power of our approach in a manner that is fast and accurate. In addition, iMOKA uses a flexible data structure that allows the easy integration of new samples and uses only a fraction of the disk space required for stocking compressed sequencing files. In addition, *k-mer* based approaches such as iMOKA have the advantage of being portable; *k-mer* sequences will not change with new versions of the genome. This is crucial for the integration of omics data with other clinical data such as imaging or patient file records.

Methods

Preprocessing

The input data can be given as SRR identifier, BAM, FASTA, or FASTQ files. In the first and second cases, the corresponding FASTQ files are automatically generated using sra-tools' fastq-dump [39] and SAMtools [40], respectively. If the data is stranded paired end sequencing, the user can reverse complement one or both the files using SeqKit [41]. In order to assert the quality of the FASTQ files, the user can use FAST QC [42] by adding the flag "-q".

For each sample, KMC3 [9] is used to count the *k*-mers of the length chosen by the user (default $k = 31$). Its output is converted into a sorted binary file optimized for the following steps of iMOKA and a JSON file containing the metadata information.

The binary file is divided into two parts: a suffix portion, containing the nucleotidic sequence and the relative count, and a prefix portion, which contains the prefixes and the positions of the respective suffixes.

The length of the prefix is defined using the following formula, an adaptation from [43]:

$$p = 0.5 \times \log_2(t) - 0.5 \times \log_2(\log_2(t))$$

where p is the prefix size and t is the total number of different k-mers for the current sample.

Matrix generation

The input to the feature reduction step is a JSON file containing the name, group, and localization of the sorted binary k-mer count file of each sample in the analysis. The JSON file also stores the sum of all the k-mer counts that will be used as a normalization factor:

$$N_{ij} = C_{ij} \times \frac{\mathrm{RF}}{T_j}$$

where

N_{ij} is the normalized count of the ith k-mer of the sample j

C_{ij} is the raw count of the ith k-mer of the sample j

T_j is the sum of the counts of all the k-mers of the sample j

RF is a rescaling factor, used to increase the value of all the normalized values and avoid computational problems related to precision. By default, RF = 1e9

Each thread starts the creation of the matrix and the reduction step in parallel, using an OpenMP [44] implementation, at a different point of the matrix according to the number of threads available using the following formula:

$$K_t = \frac{4^k - 1}{T} \times t$$

where

T is the total number of threads available

K_t is the first k-mer analyzed by the thread t (from 0 to T excluded) considering all the possible ordered combination from 0 to 4^k

k is the length of the k-mers (default 31)

The last k-mer analyzed by each thread is $K_{t+1} - 1$. For example, with 2 threads ($T = 2$) and $k = 31$, the first k-mers for each threads will be:

$$K_0 = \frac{4^{31} - 1}{2} \times 0 = 0 = AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA$$

$$K_1 = \frac{4^{31} - 1}{2} \times 1 = 2305843009213693952 = GAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA$$

Finally, the buffer size reserved for each sample is dependent on the number of parallel processes, the number of total samples, and the available memory reserved:

$$buff = \frac{\mathrm{RAM}_{\mathrm{avail}}}{\alpha \times N \times T}$$

where

Buff is the length of the buffer

$\mathrm{RAM}_{\mathrm{avail}}$ is the available RAM in GiB, defined by the user using the environmental variable "IMOKA_MAX_MEM_GB"

N is the number of samples in the matrix

T is the total number of threads available

α is a factor representing the GiB occupied by 1000 k-mers, approximated to 0.011

Bayesian classifier k-mer accuracy assessment

The accuracy of each k-mer is calculated using the NaiveBayesClassifier method implemented in the library mlpack [45]. For each k-mer, the samples are randomly divided into test and training sets, with an equal number of samples for each group scaled to the smallest one:

$$n_{\text{test}} = \text{round}(n_{\text{min}} * p_{\text{test}})$$

$$n_{\text{train}} = n_{\text{min}} - n_{\text{test}}$$

where:

n_{min} is the dimension of the smallest group

n_{test} and n_{train} are respectively the dimension of the test and training sets

p_{test} is the test fraction, 0.25 by default

Using one feature (k-mer count) x_k at a time, the NaiveBayesClassifier class computes for each label y_j:

$$P\left(X = x_k \vee Y = y_j\right)$$

$$P(Y = y_i)$$

Given that we use a pairwise comparison with a constant number of training samples amongst the labels, all the N_{labels} have the same probability

$$P(Y = y_i) = P\left(Y = y_{j+1}\right) = \frac{1}{N_{\text{labels}}}$$

The label prediction of a sample i based on the k-mer count x_k is then given by:

$$y_i = argmax(P(Y = y))$$

The accuracy of the k-mer k is computed considering only the samples part of the test set:

$$acc_k = \frac{T}{n_{\text{test}}} \times 100$$

where

acc_k is the accuracy of the k-mer k

T is the number of correct labels assigned in the test set

Because the accuracies depend on the random division of the training and test sets, we use a Monte Carlo cross validation [46] with a given number of iterations (-c argument, default 100). This cross validation can be ended by a conditional break that is triggered when the standard error across iterations drops beneath a given threshold (-s argument, default 0.5).

The k-mers that achieve an accuracy higher than the accuracy threshold (-a argument, default 65) in at least one of the pairwise comparisons are saved in a text file, along with the accuracy values.

Entropy filter booster

In order to speed up the process of accuracy estimation, we introduced an additional filter based on the Shannon entropy [47] of the counts of each k-mer that runs in parallel to the Bayesian filter (BF).

For a given k-mer k and its counts in the different samples $C_k = (c_{k0}, c_{k1}, \ldots c_{kn})$, we compute its entropy value H_k as follows:

$$H_k = -\sum_{i=0}^{n} f_{ki} \times \log_2(f_{ki})$$

$$f_{ki} = \frac{c_{ki}}{\sum_{j=0}^{n} c_{kj}}$$

The filter uses an adaptive threshold, H_{thr}, tuned according to the lowest entropy detected in the previous batch of k-mers that passed the accuracy filter (H_{min}).

Initially $H_{thr} = 0$, so all the k-mers in the first batch are evaluated by the BF and the lowest entropy is saved as H_{min}. During the analysis, H_{thr} is updated when more than E_{up} (initially equal to 30) passes the BF. The first assignment is always:

$$H_{thr} = H_{min} - (H_{min} \times a_1 \times 2)$$

Subsequently:

$$\text{IF}(H_{thr} > H_{min} - (H_{min} \times a_1)):$$

$$H_{thr} = H_{min} - (H_{min} \times a_1)$$

ELSE:

$$H_{thr} = H_{min} + (H_{min} \times a_2)$$

The adjustment parameters a_1» a_2 ensure that the new threshold is not set too close to the minimum H_{min}.

The number of k-mers required to update the threshold (E_{up}) increases by 30 at each update in order to reduce the number of computations and reduce the fluctuations of the threshold. Figure S1 shows the entropy in function of the BF estimated accuracy of a sample of k-mers from the previously defined datasets showing that the number of k-mer would have been rejected by the entropy filter but would have had an accuracy higher than 60% are rare and that the adaptive threshold is able to find a mild cutoff that can save more than 50% of the computation, like in TCGA BC, or can let the BF evaluate most of the k-mers in case of difficult datasets, like in BEAUTY.

k-mer graph generation

The k-mers that successfully passed the reduction are used as nodes in a graph. A link between two nodes is created if they overlap by a minimum number of nucleotides defined by parameter w (default = 1). This parameter can be increased if the user notices multiple small sequences in the final result, caused usually by k-mers with accuracy close to the given threshold arguments -T and -t, respectively the minimum accuracy required to consider a k-mer in the graph construction and the minimum accuracy required to generate a sequence from a graph.

iMOKA then prunes short bifurcations in the graph where there is only one node following the bifurcation. If there are multiple sequential bifurcations, then the branch with the lowest accuracy is removed.

The accuracy values are then rescaled from 0 to 100 for each pairwise comparison in order to normalize the accuracy values and favor the features that are able to classify pairs of classes that are more difficult to separate.

Since each bifurcation could correspond to a biological event such as a point mutation or splicing isoform, each separate path that results from a bifurcation will be kept as a separate sequence for downstream analysis using a depth-first graph traversal approach. When the traversal meets a bifurcation, the branch having the most similar accuracies values to the bifurcating node is kept in the current sequence and others will generate new sequences. Furthermore, to maintain the context of the bifurcations, three *k*-mers preceding the bifurcation are added to each of those new sequences.

Graph mapping and annotation

The sequences generated from the graphs can be aligned to a reference genome. Currently, iMOKA supports any aligner that provides an output in SAM or pslx format and uses the information given in the JSON configuration file "mapper-config" (-m argument) to align and to retrieve the annotation file, in GTF format. In this manuscript, we used gmap v. 2019-05-12 with the human genome GRCh38 and the GENCODE annotation v29, excluding from the file the entries with the transcript type "retained_intron".

Once the *k*-mer graphs are aligned, iMOKA identifies the following "alignment derived features" (ADF):

- Mutations, insertions, deletions, and clipping are identified by the letters "M", "I", "D" and "S," respectively, in the alignment's CIGAR string.
- Alternative splice sites are identified when a *k*-mer graph is split across exons.
- Differential expression (DE) is identified if 50% (set by parameter d) of an annotated transcript is covered by the *k*-mer graphs. Since regions with sequence variations not associated with the classes generate holes in the graphs reducing the portion of the transcripts that generate useful *k*-mers, a higher threshold might result in classifying DE event as general "gene" event, that is, the best *k*-mer in a gene.
- Alternative intronic events are identified if 50% (set by parameter d) of an annotated intron is covered by the *k*-mer graphs.
- Intergenic events are identified if the *k*-mer graph maps to the genome but not to any annotated transcript.
- Unmapped or multimapped events are created for those *k*-mer graphs that have no mapping or map to multiple sites.

iMOKA will preserve one *k*-mer per event, the one with the highest accuracy score. Table S2 contains the list of events with a detailed description.

iMOKA Implementation

The feature reduction component of iMOKA is implemented in C++ using the following libraries: MLpack [45], armadillo [48], cephes [49], cxxopts [50], and nlohmann/json [51]. The self-organizing map and the random forest are implemented in python 3 using the following libraries: numpy [52], pandas [53], sklearn [54], and SimpSOM [55]. The whole software is included in a ready-to-use Docker and Singularity [56] image and is released under the Open Source CeCILL license.

Benchmark

Transcript abundance was computed using Salmon [57] version 1.1.0 using the index built on the reference transcriptome GENCODE v29 (hg38). The PSI values were computed using Whippet [15] version v0.10.4. We processed the samples in parallel in 4 processes allowing 2 threads and a maximum of 8 GiB of RAM each. The differential expression analysis was performed between each pair of classes in R v3.6.3 using the parameters and functions described in a recent benchmark [58] for the methods DESeq2 [12], edgeR [13], and limmaVoom [14]. Significantly different PSI values between two subsets were detected using whippet-delta.jl, included in the Whippet package.

Random Forest classifier feature selection and oob score comparison

In order to compare the same number of features extracted by each pipeline, we used the sklearn method SelectFromModel to select 20 features using a decision tree classifier (DTC) trained with all the samples and all the features in order to identify twenty features that, in combination, can be good classifiers. Using an increasing number of features, from 2 to 20, we trained multiple RandomForestClassifier to retrieve the out of the box scores.

We also performed a 5-fold cross validation of the largest and better characterized dataset, TCGA BRCA, to evaluate the accuracy of a model on unseen data. For each fold, we performed the feature reduction using only the training in each method. The final list of features is reduced similarly as for the oob score determination and the balanced accuracy score is estimated for the test set.

Supplementary information

Supplementary Information accompanies this paper at https://doi.org/10.1186/s13059-020-02165-2.

Additional file 1. TCGA_BC_aggregated.json - iMOKA results for the dataset TCGA_BC.
Additional file 2. TCGA_OV_aggregated.json - iMOKA results for the dataset TCGA_OV.
Additional file 3. BEAUTY_aggregated.json - iMOKA results for the dataset BEAUTY.
Additional file 4. DLBCL_aggregated.json - iMOKA results for the dataset DLBCL.
Additional file 5. GO - folder containing the DAVID gene ontology result for each dataset.
Additional file 6. iMOKA_supplementary.docx - Supplementary materials.
Additional file 7. Supplementary Figures S1-S7.
Additional file 8. Review history.

Acknowledgements

We wish to acknowledge the Genotoul platform (genotoul.fr) for providing us with calculation time on their servers. The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Review history

The review history is available as Additional file 8.

Peer review information
Yixin Yao was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
C.L., W.R., and A.M. designed the algorithm; C.L. coded the software; S.B. designed and coded the SOM; C.L., W.R., A.M., S.B., and J.P.V. designed the experiments.; L.D.B. contributed to the binary data structure optimization during her internship; W.R. and C.L. wrote the article. The authors read and approved the final manuscript.

Funding
We wish to acknowledge the Agence Nationale de la Recherche (ANR.IC.IC - WIRED), the Labex EpiGenMed, and the MUSE initiative for their financial support.

Availability of data and materials
The data used in this manuscript are available from the Cancer Genome Atlas under the project ID TCGA-BRCA [21] and TCGA-OV [29] with dbGaP study accession identifier phs000178.v11.p8 [59]; the BEAUTY dataset [31] is available under the dbGaP study accession identifier phs001050.v1.p1 [59]. Restrictions apply to the availability of these data, which were used under license for those studies, and so are not publicly available. Data are however available by submitting a request to the respective repositories.
The DLBCL targeted RNA-seq data [36] are publicly available in the EMBL-EBI ArrayExpress with the accession number E-MTAB-6597 [60].
iMOKA is available at https://github.com/RitchieLabIGH/iMOKA [61] under the Open Source CeCILL license. The copy of the scripts used for the benchmark is available under the subfolder https://github.com/RitchieLabIGH/iMOKA/paper_codes.
The DOI for the source version used in this article is https://doi.org/10.5281/zenodo.4008947 [62].

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors have no competing interests to declare.

Author details
[1]IGH, Centre National de la Recherche Scientifique, University of Montpellier, Montpellier, France. [2]LIRMM, Université de Montpellier, CNRS, Montpellier, France.

Received: 6 May 2020 Accepted: 10 September 2020
Published online: 13 October 2020

References
1. Learn CA, et al. Resistance to tyrosine kinase inhibition by mutant epidermal growth factor receptor variant III contributes to the neoplastic phenotype of glioblastoma multiforme. Clin. Cancer Res. 2004;10:3216–24.
2. Zhang Z-M, et al. Pygo2 activates MDR1 expression and mediates chemoresistance in breast cancer via the Wnt/β-catenin pathway. Oncogene. 2016;35:4787–97.
3. Martín-Martín N, et al. Stratification and therapeutic potential of PML in metastatic breast cancer. Nat Commun. 2016;7:12595.
4. Grossman RL, et al. Toward a shared vision for cancer genomic data. N. Engl. J. Med. 2016;375:1109–12.
5. Audoux J, et al. DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. Genome Biol. 2017;18:243.
6. Kirk, J. M. et al. Functional classification of long non-coding RNAs by k-mer content. Nat. Genet. 50, 1474–1482 (2018).
7. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers. BMC Genomics. 2015;16:236.
8. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol. 2018;19:198.
9. Thomas A, et al. GECKO is a genetic algorithm to classify and explore high throughput sequencing data. Commun. Biol. 2019;2:222.
10. Kokot M, Dlugosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. Bioinforma. Oxf. Engl. 2017;33:2759–61.
11. Sacomoto GAT, et al. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. BMC Bioinformatics. 2012;13(Suppl 6):S5.
12. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
13. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
14. Ritchie ME, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
15. Sterne-Weiler T, Weatheritt RJ, Best AJ, Ha KCH, Blencowe BJ. Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop. Mol. Cell. 2018;72:187–200.e6.
16. Rahman A, Hallgrímsdóttir I, Eisen M, Pachter L. Association mapping from sequencing reads using k-mers. eLife 2018;7:e32920.

17. Drouin A, et al. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genomics. 2016;17:754.
18. Hastie T, The T. Elements of statistical learning second edition. Math Intell. 2017;27:83–5.
19. Breiman, L. Out-of-bag estimation. In (1996).
20. Bastien RRL, et al. PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. BMC Med Genomics. 2012;5:44.
21. Hoadley KA, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173:291–304.e6.
22. Jeannot E, et al. A single droplet digital PCR for ESR1 activating mutations detection in plasma. Oncogene. 2020; 39:2987–95.
23. Ciriello G, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell. 2015;163:506–19.
24. Han B, et al. FOXC1: an emerging marker and therapeutic target for cancer. Oncogene. 2017;36:3957–63.
25. Yang Y, et al. TPX2 promotes migration and invasion of human breast cancer cells. Asian Pac J. Trop. Med. 2015;8:1064–70.
26. Thakkar A, et al. High expression of three-gene signature improves prediction of relapse-free survival in estrogen receptor-positive and node-positive breast tumors. Biomark. Insights. 2015;10:103–12.
27. Bjørklund SS, et al. Widespread alternative exon usage in clinically distinct subtypes of invasive ductal carcinoma. Sci. Rep. 2017;7:5568.
28. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 2009;4:44–57.
29. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474:609–15.
30. Villalobos VM, Wang YC, Sikic BI. Reannotation and analysis of clinical and chemotherapy outcomes in the ovarian data set from the Cancer Genome Atlas. JCO Clin. Cancer Inform. 2018;2:1–16.
31. Goetz, M. P. et al. Tumor sequencing and patient-derived xenografts in the neoadjuvant treatment of breast cancer. J Natl Cancer Inst. 2017;109(7):djw306. https://doi.org/10.1093/jnci/djw306.
32. Yi H, Raman AT, Zhang H, Allen GI, Liu Z. Detecting hidden batch factors through data-adaptive adjustment for biological effects. Bioinforma. Oxf. Engl. 2018;34:1141–7.
33. Middleton R, et al. IRFinder: assessing the impact of intron retention on mammalian gene expression. Genome Biol. 2017;18:51.
34. Shi X, Sun X. Regulation of paclitaxel activity by microtubule-associated proteins in cancer chemotherapy. Cancer Chemother. Pharmacol. 2017;80:909–17.
35. Buljan VA, et al. Calcium-axonemal microtubuli interactions underlie mechanism(s) of primary cilia morphological changes. J. Biol. Phys. 2018;44:53–80.
36. Fornecker L-M, et al. Multi-omics dataset to decipher the complexity of drug resistance in diffuse large B-cell lymphoma. Sci. Rep. 2019;9.
37. Agarwal NK, et al. Transcriptional regulation of serine/threonine protein kinase (AKT) genes by glioma-associated oncogene homolog 1. J. Biol. Chem. 2013;288:15390–401.
38. Zhu C, Chen G, Zhao Y, Gao X-M, Wang J. Regulation of the development and function of B cells by ZBTB transcription factors. Front. Immunol. 2018;9.
39. *ncbi/sra-tools*. (NCBI - National Center for Biotechnology Information/NLM/NIH, 2020) https://github.com/ncbi/sra-tools.
40. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16): 2078–9. https://doi.org/10.1093/bioinformatics/btp352.
41. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One. 2016; 11(10):e0163962. Published 2016 Oct 5. https://doi.org/10.1371/journal.pone.0163962.
42. FastQC: a quality control tool for high throughput sequence data – https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
43. Park G, Hwang H-K, Nicodème P, Szpankowski W. Profiles of tries. SIAM J. Comput. 2009;38:1821–80.
44. L. Dagum and R. Menon, "OpenMP: an industry standard API for shared-memory programming," in IEEE Computational Science and Engineering, 1998;5(1):46–55. https://doi.org/10.1109/99.660313.
45. Curtin R, et al. mlpack 3: a fast, flexible machine learning library. J. Open Source Softw. 2018;3:726.
46. Dubitzky, W., Granzow, M. & Berrar, D. P. Fundamentals of data mining in genomics and proteomics. (Springer Science & Business Media, 2007).
47. Shannon, C. E. The mathematical theory of communication. 1963. MD Comput. Comput. Med. Pract. 14, 306–317 (1997).
48. Sanderson C, Curtin R. Armadillo: a template-based C++ library for linear algebra. J. Open Source Softw. 2016;1:26.
49. CEPHES Mathematical function library. http://www.netlib.org/cephes/.
50. Lightweight C++ command line option parser. jarro2783/cxxopts. 2020. https://github.com/jarro2783/cxxopts.
51. JSON for Modern C++, N. nlohmann/json. 2020. https://github.com/nlohmann/json.
52. van der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. 2011. https://doi.org/10.1109/MCSE.2011.37.
53. McKinney, W. Data structures for statistical computing in Python. Proc. 9th Python Sci. Conf. (2010).
54. Pedregosa F, et al. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 2011;12:2825–30.
55. Federico Comitani. fcomitani/SimpSOM: v1.3.4. (Zenodo, 2019). https://doi.org/10.5281/zenodo.2621560.
56. Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. PLOS ONE. 2017;12: e0177459.
57. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods. 2017;14:417–9.
58. Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. BMC Bioinformatics. 2017;18:38.
59. dbGaP/database of genotypes and phenotypes/ National Center for Biotechnology Information, National Library of Medicine (NCBI/NLM) https://www.ncbi.nlm.nih.gov/gap.
60. Athar A. et al., 2019. ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res, https://doi.org/10.1093/nar/gky964, Pubmed ID 30357387.

61. Lorenzi, C. et al. iMOKA: k-mer based software to analyze large collections of sequencing data. (GitHub, 2020). https://github.com/RitchieLabIGH/iMOKA.
62. Lorenzi, C. et al. iMOKA: k-mer based software to analyze large collections of sequencing data. (Zenodo, 2020). https://doi.org/10.5281/zenodo.4008947.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

Conclusions

The last three decades were marked by incredible technological advancements, both from the biotechnological and computational points of view.
To keep up with them, we adapted IRFinder to support the third-generation sequences and use new methodologies, the convolutional neural network, to refine and improve its results. Furthermore, we proposed IRBase, a platform where users can not only visualize their data but also compare them with the ones shared by other users.
The possibility to sequence at low cost and high fidelity large cohorts of people gives the opportunity to increase our knowledge about the mechanisms underlying pathologies and generate models to predict the responses to drugs, treatments and environmental modification.
In the introduction, we saw how classical approaches, based on mapping to a reference genome and using reference annotations, present lots of levels of variability caused by different versions of the references and softwares used. Additionally, a large portion of the information is usually discarded because it doesn't fit with the features considered in the study.
We showed how κ-mer based approaches can be an optimal and agnostic representation of sequencing data, useful to identify biomarkers that can be applied for clinical and research purposes.
In this optic, we implemented iMOKA, a software that can efficiently select a group of κ-mers with a low redundancy of information and high capacity in discrimination of the phenotypes in analysis within a large cohort of samples.

Bioinformatics is a young field and its identity is still shaping, trying to find its place in the middle between statistics, informatics and biology.
The technological advancements we saw taking place in the last few decades are causing a revolutionary shift from hypothesis-driven to data-driven science, requiring wet-lab researchers to spend more time in front of a computer to analyse and understand the data they produced.
Bioinformatics classes are given in most of the university biological science courses, forming the next generation of researchers in the usage of the basic tools and resources currently available.
Developing user-friendly, maintainable and powerful platforms is therefore the direction that not only lots of private companies are taking, such as Geneious and QIAGEN CLC Genomics Workbench, but also the open-source community, of which Galaxy is the most successful example.
Additionally, more and more pure bioinformatics laboratories are rising in the research centres that use publicly available data to perform novel analyses and

develop new algorithms, supported by classical wet labs only for the validation of the findings.
Unfortunately, this shift is not affecting the way data is stored and distributed.
For companies such as Google, Amazon and Microsoft, it's enough to accept the general conditions with a single click to have access to any user data: e-mails, browsing history, what colour was the t-shirt we bought three years ago and our exact location every minute we keep our phone in our pocket.
They use that information to feed you with "the best advertisement for you", to influence your opinion on social media and to direct your next vote[239], in the most efficient way possible. They can store and share within their platforms this huge amount of information, legally and in the name of profit.
On the other side, we have fragmented national health systems that don't take care of how or where the clinical data are stored, leaving the burden to the individual hospitals.
In my opinion, creating an international platform for data storage is the next big but necessary challenge the scientific community has to face to fully exploit the potential of the new sequencing technologies.
Such a platform should use light and standard data format, a complete and flexible ontology and ensure the privacy of the information, allowing certified laboratories to access in agreement with detailed rules of conduct.
An interesting report from NIH[240] predicts that sequencing and analysing the whole human genome of patients will become a routine procedure for any research lab by 2030, that transcriptome and epigenetic analysis will be routinely incorporated into predictive models and that "an individual's complete genome sequence along with informative annotations will, if desired, be securely and readily accessible on their smartphone."
Those forecasts need a large and international effort to generate new tools able to generate, store and analyse such data using fast, efficient, robust and privacy-aware procedures.
κ-mer based algorithms have all the prerequisites to not only offer a compressed representation of the data but also to analyse large cohorts of samples to identify biomarkers useful for personalized medicine.
Future works should focus on the generation of κ-mer based algorithms for the efficient and lossless compression of raw sequencing data, their anonymization and application to new biological questions using different types of data.
For example, a recent study[241] used compressed k-mer groups, a set of k-mers having similar counts across the samples, to cluster single cells in scRNA-seq data, a task usually performed using gene counts.
Another interesting field of application for k-mers using human whole-genome sequencing analysis are genome-wide association studies, identification of mutational events in cancer, copy number variations analysis and the identification of translocation events.
Improving the interpretability of the k-mers results would allow more and more researchers to accept k-mer based softwares as part of standard analysis.

A file format based on k-mers able to store both the abundance and the order of the k-mers in a compressed, lossless and efficient way would take over the current standard BAM and CRAM files, especially if paired to a genome browser able to quickly reproduce a visualization of this information on a reference genome.
Finally, IRFinder-S is just one of the several examples of how artificial intelligence methodologies and large cohorts of data can help in solving biological related problems where classical approaches struggle to face.
The methodology used to train the CNN model of IRFinder can be, with the proper adaptations, extended to identify other types of alternative splicing events, new transcripts and other transcriptomics related elements, such as promoter upstream transcripts, in a reference-free manner.
To achieve this goal, a finely annotated training set is necessary in order to correctly train the model, together with a fast and efficient implementation of a genome-wise features generator.
Third generation sequencing data would facilitate the task because they are more likely to give the information concerning the full structure of the transcripts isoforms.

References

1. Lorenzi, C. *et al.* iMOKA: k -mer based software to analyze large collections of sequencing data. *Genome Biol.* **21**, 1–19 (2020).
2. Thomas, A. *et al.* GECKO is a genetic algorithm to classify and explore high throughput sequencing data. *Commun. Biol.* **2**, 222 (2019).
3. Lorenzi, C. *et al.* IRFinder-S: a comprehensive suite to discover and explore intron retention. *Genome Biol.* **22**, 307 (2021).
4. Dahm, R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum. Genet.* **122**, 565–581 (2008).
5. His, W. F. Miescher. *Histochem. Physiol. Arb. Von Friedrich Miescher* **1**, 5–32 (1897).
6. Lederberg, J. The transformation of genetics by DNA: an anniversary celebration of Avery, MacLeod and McCarty (1944). *Genetics* **136**, 423–426 (1994).
7. Miescher, F. Uberdie Chemische Zusammen-setzung der Eiterzellen. *Hoppe-Seylers Med Chem Untersuchgn Berl.* 441 (1870).
8. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
9. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).
10. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).
11. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
12. Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).

13. Cobb, M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol.* **15**, (2017).
14. Serganov, A. & Patel, D. J. Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat. Rev. Genet.* **8**, 776–790 (2007).
15. Nilsen, T. W. The spliceosome: the most complex macromolecular machine in the cell? *BioEssays* **25**, 1147–1149 (2003).
16. House, A. E. & Lynch, K. W. Regulation of Alternative Splicing: More than Just the ABCs *. *J. Biol. Chem.* **283**, 1217–1221 (2008).
17. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**, 437–451 (2017).
18. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010).
19. Wang, Y. *et al.* Mechanism of alternative splicing and its regulation (Review). *Biomed. Rep.* **3**, 152–158 (2015).
20. Chang, Y.-F., Imam, J. S. & Wilkinson, M. F. The Nonsense-Mediated Decay RNA Surveillance Pathway. *Annu. Rev. Biochem.* **76**, 51–74 (2007).
21. Venkataraman, K., Guja, K. E., Garcia-Diaz, M. & Karzai, A. W. Non-stop mRNA decay: a special attribute of trans-translation mediated ribosome rescue. *Front. Microbiol.* **5**, (2014).
22. Middleton, R. *et al.* IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol.* **18**, 51 (2017).
23. Wong, J. J.-L., Au, A. Y. M., Ritchie, W. & Rasko, J. E. J. Intron retention in mRNA: No longer nonsense. *BioEssays* **38**, 41–49 (2016).
24. Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* gr.177790.114 (2014)

doi:10.1101/gr.177790.114.

25. Inoue, D. *et al.* Minor intron retention drives clonal hematopoietic disorders and diverse cancer predisposition. *Nat. Genet.* (2021) doi:10.1038/s41588-021-00828-9.
26. Zhang, Y. *et al.* Comprehensive characterization of alternative splicing in renal cell carcinoma. *Brief. Bioinform.* (2021) doi:10.1093/bib/bbab084.
27. Tyzack, G. E. *et al.* Aberrant cytoplasmic intron retention is a blueprint for RNA binding protein mislocalization in amyotrophic lateral sclerosis. *Brain J. Neurol.* (2021) doi:10.1093/brain/awab078.
28. Ullrich, S. & Guigó, R. Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development. *Nucleic Acids Res.* **48**, 1327–1340 (2020).
29. Yeom, K.-H. *et al.* Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring. *Genome Res.* (2021) doi:10.1101/gr.273904.120.
30. Naro, C. *et al.* An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation. *Dev. Cell* **41**, 82-93.e4 (2017).
31. Brady, L. K. *et al.* Transcriptome analysis of hypoxic cancer cells uncovers intron retention in EIF2B5 as a mechanism to inhibit translation. *PLoS Biol.* **15**, e2002623 (2017).
32. Kanagasabai, R. *et al.* Alternative RNA Processing of Topoisomerase IIα in Etoposide-Resistant Human Leukemia K562 Cells: Intron Retention Results in a Novel C-Terminal Truncated 90-kDa Isoform. *J. Pharmacol. Exp. Ther.* **360**, 152–163 (2017).

33. Uzor, S. *et al.* Autoregulation of the human splice factor kinase CLK1 through exon skipping and intron retention. *Gene* **670**, 46–54 (2018).
34. Lynch, M. & Marinov, G. K. The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci.* **112**, 15690–15695 (2015).
35. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **21**, 630–644 (2020).
36. Davidson, E. & Levin, M. Gene regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 4935 (2005).
37. Lelli, K. M., Slattery, M. & Mann, R. S. Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.* **46**, 43–68 (2012).
38. Alajem, A. *et al.* DNA methylation patterns expose variations in enhancer-chromatin modifications during embryonic stem cell differentiation. *PLoS Genet.* **17**, e1009498 (2021).
39. Cannell, I. G., Kong, Y. W. & Bushell, M. How do microRNAs regulate gene expression? *Biochem. Soc. Trans.* **36**, 1224–1231 (2008).
40. Carthew, R. W. & Sontheimer, E. J. Origins and Mechanisms of miRNAs and siRNAs. *Cell* **136**, 642–655 (2009).
41. Meyer, K. D. & Jaffrey, S. R. The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nat. Rev. Mol. Cell Biol.* **15**, 313–326 (2014).
42. Geula, S. *et al.* Stem cells. m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* **347**, 1002–1006 (2015).
43. Su, A. a. H. & Randau, L. A-to-I and C-to-U editing within transfer RNAs. *Biochem. Biokhimiia* **76**, 932–937 (2011).
44. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of

RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* **19**, 327–341 (2018).

45. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* **15**, 163–175 (2014).
46. Tahmasebi, S., Khoutorsky, A., Mathews, M. B. & Sonenberg, N. Translation deregulation in human disease. *Nat. Rev. Mol. Cell Biol.* **19**, 791–807 (2018).
47. Teixeira, F. K. & Lehmann, R. Translational Control during Developmental Transitions. *Cold Spring Harb. Perspect. Biol.* **11**, a032987 (2019).
48. Emmott, E., Jovanovic, M. & Slavov, N. Ribosome Stoichiometry: From Form to Function. *Trends Biochem. Sci.* **44**, 95–109 (2019).
49. Schwartz, A. L. & Ciechanover, A. Targeting proteins for destruction by the ubiquitin system: implications for human pathobiology. *Annu. Rev. Pharmacol. Toxicol.* **49**, 73–96 (2009).
50. Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **21**, 255–261 (2003).
51. Ryan, C. J. *et al.* High-resolution network biology: connecting sequence with function. *Nat. Rev. Genet.* **14**, 865–879 (2013).
52. Sanda, T. *et al.* Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell* **22**, 209–221 (2012).
53. Zhou, H. *et al.* MED12 mutations link intellectual disability syndromes with dysregulated GLI3-dependent Sonic Hedgehog signaling. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19763–19768 (2012).
54. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013).
55. Zhu, R., Ribeiro, A. S., Salahub, D. & Kauffman, S. A. Studying genetic

regulatory networks at the molecular level: Delayed reaction stochastic models. *J. Theor. Biol.* **246**, 725–745 (2007).

56. Batt, G. *et al.* Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in Escherichia coli. *Bioinformatics* **21**, i19–i28 (2005).
57. Banf, M. & Rhee, S. Y. Computational inference of gene regulatory networks: Approaches, limitations and opportunities. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1860**, 41–52 (2017).
58. Milanez-Almeida, P., Martins, A. J., Germain, R. N. & Tsang, J. S. Cancer prognosis with shallow tumor RNA sequencing. *Nat. Med.* **26**, 188–192 (2020).
59. Alimadadi, A. *et al.* Machine learning-based classification and diagnosis of clinical cardiomyopathies. *Physiol. Genomics* **52**, 391–400 (2020).
60. Zhang, G., Xue, Z., Yan, C., Wang, J. & Luo, H. A Novel Biomarker Identification Approach for Gastric Cancer Using Gene Expression and DNA Methylation Dataset. *Front. Genet.* **12**, 644378 (2021).
61. Chen, J. *et al.* Sparse deep neural networks on imaging genetics for schizophrenia case-control classification. *Hum. Brain Mapp.* (2021) doi:10.1002/hbm.25387.
62. López-García, G., Jerez, J. M., Franco, L. & Veredas, F. J. Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *PloS One* **15**, e0230536 (2020).
63. Woo, G. *et al.* DeepCOP: deep learning-based approach to predict gene regulating effects of small molecules. *Bioinforma. Oxf. Engl.* **36**, 813–818 (2020).
64. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Res.*

49, D1207–D1217 (2021).

65. The Cancer Genome Atlas Program - National Cancer Institute. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (2018).
66. The Harvard Personal Genome Project (PGP) – enabling participant-driven science. https://pgp.med.harvard.edu/.
67. A pathology atlas of the human cancer transcriptome | Science. https://science-sciencemag-org.insb.bib.cnrs.fr/content/357/6352/eaan2507.
68. Tetz, V. V. & Tetz, G. V. A new biological definition of life. 6.
69. Schrödinger, E. *What is Life? The Physical Aspect of the Living Cell*. (Cambridge University Press, 1944).
70. Koshland, D. E. The Seven Pillars of Life. *Science* **295**, 2215–2216 (2002).
71. Hershey, A. D. & Chase, M. INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE. *J. Gen. Physiol.* **36**, 39–56 (1952).
72. Hutchison, C. A., III. DNA sequencing: bench to bedside and beyond †. *Nucleic Acids Res.* **35**, 6227–6237 (2007).
73. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).
74. Sanger, F. & Tuppy, H. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochem. J.* **49**, 481–490 (1951).
75. Holley, R. W. *et al.* STRUCTURE OF A RIBONUCLEIC ACID. *Science* **147**, 1462–1465 (1965).
76. Kaiser, A. D. & Wu, R. Structure and function of DNA cohesive ends. *Cold*

Spring Harb. Symp. Quant. Biol. **33**, 729–734 (1968).

77. Wu, R. & Kaiser, A. D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.* **35**, 523–537 (1968).
78. Jou, W. M., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. *Nature* **237**, 82–88 (1972).
79. Gilbert, W. & Maxam, A. The Nucleotide Sequence of the lac Operator. *Proc. Natl. Acad. Sci.* **70**, 3581–3584 (1973).
80. Kelly, T. J. & Smith, H. O. A restriction enzyme from Hemophilus influenzae. II. *J. Mol. Biol.* **51**, 393–409 (1970).
81. Middleton, J. H., Edgell, M. H. & Hutchison, C. A. Specific fragments of phi X174 deoxyribonucleic acid produced by a restriction enzyme from Haemophilus aegyptius, endonuclease Z. *J. Virol.* **10**, 42–50 (1972).
82. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–448 (1975).
83. Sanger, F. The Croonian Lecture, 1975. Nucleotide sequences in DNA. *Proc. R. Soc. Lond. B Biol. Sci.* **191**, 317–333 (1975).
84. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687–695 (1977).
85. Heidecker, G., Messing, J. & Gronenborn, B. A versatile primer for DNA sequencing in the M13mp2 cloning system. *Gene* **10**, 69–73 (1980).
86. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560–564 (1977).
87. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).

88. Albariño, C. G., Posik, D. M., Ghiringhelli, P. D., Lozano, M. E. & Romanowski, V. Arenavirus phylogeny: a new insight. *Virus Genes* **16**, 39–46 (1998).

89. Ibrahim, A., Goebel, B. M., Liesack, W., Griffiths, M. & Stackebrandt, E. The phylogeny of the genus Yersinia based on 16S rDNA sequences. *FEMS Microbiol. Lett.* **114**, 173–177 (1993).

90. Pollock, D. D., Eisen, J. A., Doggett, N. A. & Cummings, M. P. A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol. Biol. Evol.* **17**, 1776–1788 (2000).

91. Doyle, J. J. & Gaut, B. S. Evolution of genes and taxa: a primer. *Plant Mol. Biol.* **42**, 1–23 (2000).

92. Wildin, R. S. & Cogdell, D. E. Clinical utility of direct mutation testing for congenital nephrogenic diabetes insipidus in families. *Pediatrics* **103**, 632–639 (1999).

93. Kolbert, C. P. & Persing, D. H. Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. *Curr. Opin. Microbiol.* **2**, 299–305 (1999).

94. Rossetti, S. *et al.* Mutation analysis of the entire PKD1 gene: genetic and diagnostic implications. *Am. J. Hum. Genet.* **68**, 46–63 (2001).

95. van Daal, A. DNA profiling in forensic science and parentage testing. *Australas. Biotechnol.* **4**, 92–96 (1994).

96. Boonsaeng, V. DNA fingerprinting and forensic medicine. *Southeast Asian J. Trop. Med. Public Health* **26 Suppl 1**, 296–300 (1995).

97. Debenham, P. G. DNA fingerprinting. *J. Pathol.* **164**, 101–106 (1991).

98. Adams, C. P. & Kron, S. J. Method for performing amplification of nucleic acid with two primers bound to a single solid support. (1997).

99. Chetverina, H. V. & Chetverin, A. B. Cloning of RNA molecules in vitro. *Nucleic*

Acids Res. **21**, 2349–2353 (1993).

100. Adessi, C. *et al.* Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* **28**, e87 (2000).
101. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
102. Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci.* **100**, 8817–8822 (2003).
103. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
104. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84–89 (1996).
105. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
106. Huang, Y.-F., Chen, S.-C., Chiang, Y.-S., Chen, T.-H. & Chiu, K.-P. Palindromic sequence impedes sequencing-by-ligation mechanism. *BMC Syst. Biol.* **6**, S10 (2012).
107. Seo, T. S. *et al.* Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 5926–5931 (2005).
108. Ruparel, H. *et al.* Design and synthesis of a 3’-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 5932–5937 (2005).

109. DNA Sequencing Costs: Data. *Genome.gov* https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

110. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 9748–9753 (2001).

111. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).

112. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).

113. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).

114. Li, Y. & Tollefsbol, T. O. DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods Mol. Biol. Clifton NJ* **791**, 11–21 (2011).

115. Paired-End vs. Single-Read Sequencing Technology. https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html.

116. Sequencing Read Length | How to calculate NGS read length. https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html.

117. Nurk, S. *et al.* The complete sequence of a human genome. *bioRxiv* 2021.05.26.445798 (2021) doi:10.1101/2021.05.26.445798.

118. Reardon, S. A complete human genome sequence is close: how scientists filled in the gaps. *Nature* **594**, 158–159 (2021).

119. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).

120. Real-Time DNA Sequencing from Single Polymerase Molecules | Science. https://science-sciencemag-org.insb.bib.cnrs.fr/content/323/5910/133.
121. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).
122. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
123. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).
124. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).
125. Stein, D. Nanopore Sequencing: Forcing Improved Resolution. *Biophys. J.* **109**, 2001–2002 (2015).
126. *nanoporetech/bonito*. (Oxford Nanopore Technologies, 2021).
127. Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* **25**, 1750–1756 (2015).
128. Laver, T. *et al.* Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **3**, 1–8 (2015).
129. Rang, F. J., Kloosterman, W. P. & Ridder, J. de. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 1–11 (2018).
130. Accuracy. *Oxford Nanopore Technologies* http://nanoporetech.com/accuracy.
131. Oxford Nanopore announces multiple releases, for high-accuracy, content-rich, high-throughput whole-genome sequencing, and dynamic targeted sequencing. *Oxford Nanopore Technologies*

http://nanoporetech.com/about-us/news/oxford-nanopore-announces-multiple-releases-high-accuracy-content-rich-high.

132. Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**, 2193–2198 (2019).
133. Viehweger, A. *et al.* Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Res.* **29**, 1545–1554 (2019).
134. Magi, A. *et al.* Nano-GLADIATOR: real-time detection of copy number alterations from nanopore sequencing data. *Bioinformatics* **35**, 4213–4221 (2019).
135. Sanderson, N. D. *et al.* Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices. *BMC Genomics* **19**, 714 (2018).
136. Technology for Research and Translational Medicine. *NanoString* https://www.nanostring.com/.
137. Goytain, A. & Ng, T. NanoString nCounter Technology: High-Throughput RNA Validation. *Methods Mol. Biol. Clifton NJ* **2079**, 125–139 (2020).
138. Lee, J. *et al.* Nanostring-based multigene assay to predict recurrence for gastric cancer patients after surgery. *PloS One* **9**, e90133 (2014).
139. Veldman-Jones, M. H. *et al.* Reproducible, Quantitative, and Flexible Molecular Subtyping of Clinical DLBCL Samples Using the NanoString nCounter System. *Clin. Cancer Res.* **21**, 2367–2378 (2015).
140. Lira, M. E. *et al.* A single-tube multiplexed assay for detecting ALK, ROS1, and RET fusions in lung cancer. *J. Mol. Diagn. JMD* **16**, 229–243 (2014).
141. Martin, J. W. *et al.* Digital expression profiling identifies RUNX2, CDC5L, MDM2, RECQL4, and CDK4 as potential predictive biomarkers for neo-adjuvant

chemotherapy response in paediatric osteosarcoma. *PloS One* **9**, e95843 (2014).

142. Sivendran, S. *et al.* Dissection of immune gene networks in primary melanoma tumors critical for antitumor surveillance of patients with stage II-III resectable disease. *J. Invest. Dermatol.* **134**, 2202–2211 (2014).
143. M'Boutchou, M.-N. & van Kempen, L. C. Analysis of the Tumor Microenvironment Transcriptome via NanoString mRNA and miRNA Expression Profiling. *Methods Mol. Biol. Clifton NJ* **1458**, 291–310 (2016).
144. Huang, B. *et al.* Diagnosis and typing of influenza using fluorescent barcoded probes. *Sci. Rep.* **7**, 18092 (2017).
145. Veldman-Jones, M. H. *et al.* Evaluating Robustness and Sensitivity of the NanoString Technologies nCounter Platform to Enable Multiplexed Gene Expression Analysis of Clinical Samples. *Cancer Res.* **75**, 2587–2593 (2015).
146. Busby, M. A., Stewart, C., Miller, C. A., Grzeda, K. R. & Marth, G. T. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* **29**, 656–657 (2013).
147. Guo, Y., Zhao, S., Li, C.-I., Sheng, Q. & Shyr, Y. RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment. *Cancer Inform.* **13**, 1–5 (2014).
148. Wu, H., Wang, C. & Wu, Z. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinforma. Oxf. Engl.* **31**, 233–241 (2015).
149. Bi, R. & Liu, P. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics* **17**, 146 (2016).

150. Poplawski, A. & Binder, H. Feasibility of sample size calculation for RNA-seq studies. *Brief. Bioinform.* **19**, 713–720 (2018).

151. Schurch, N. J. *et al.* How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22**, 839–851 (2016).

152. Hong, E. P. & Park, J. W. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics Inform.* **10**, 117–122 (2012).

153. Liu, Y., Zhou, J. & White, K. P. RNA-seq differential expression studies: more sequence or more replication? *Bioinforma. Oxf. Engl.* **30**, 301–304 (2014).

154. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2010).

155. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, 1–10 (2009).

156. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

157. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

158. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).

159. Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G. & Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* **21**, 734–740 (2011).

160. MPEG-G Genomic Information Representation and Transport. *MPEG-G Genomic Information Representation* https://mpeg-g.org/.
161. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, (2016).
162. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
163. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
164. Kanitz, A. *et al.* Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16**, 1–26 (2015).
165. Zhang, C., Zhang, B., Lin, L.-L. & Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18**, 583 (2017).
166. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
167. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
168. Srivastava, A., Sarkar, H., Gupta, N. & Patro, R. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics* **32**, i192–i200 (2016).
169. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform

quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32**, 462–464 (2014).

170. Germain, P.-L. *et al.* RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Res.* **44**, 5054–5067 (2016).
171. Everaert, C. *et al.* Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Sci. Rep.* **7**, 1559 (2017).
172. Zheng, H., Brennan, K., Hernaez, M. & Gevaert, O. Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. *GigaScience* **8**, (2019).
173. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).
174. Bendall, M. L. *et al.* Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLOS Comput. Biol.* **15**, e1006453 (2019).
175. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
176. Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E5593-5601 (2014).
177. Sterne-Weiler, T., Weatheritt, R. J., Best, A. J., Ha, K. C. H. & Blencowe, B. J. Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. *Mol. Cell* **72**, 187-200.e6 (2018).

178. Lin, K.-T. & Krainer, A. R. PSI-Sigma: a comprehensive splicing-detection method for short-read and long-read RNA-seq analysis. *Bioinformatics* **35**, 5048–5054 (2019).

179. Green, C. J., Gazzara, M. R. & Barash, Y. MAJIQ-SPEL: web-tool to interrogate classical and complex splicing variations from RNA-Seq data. *Bioinformatics* **34**, 300–302 (2018).

180. Trincado, J. L. *et al.* SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 40 (2018).

181. Broseus, L. & Ritchie, W. Challenges in detecting and quantifying intron retention from next generation sequencing data. *Comput. Struct. Biotechnol. J.* **18**, 501–508 (2020).

182. Wong, J. J.-L. *et al.* Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**, 583–595 (2013).

183. Bai, Y., Ji, S. & Wang, Y. IRcall and IRclassifier: two methods for flexible detection of intron retention events from RNA-Seq data. *BMC Genomics* **16**, 1–9 (2015).

184. Li, H.-D., Funk, C. C. & Price, N. D. iREAD: a tool for intron retention detection from RNA-seq data. *BMC Genomics* **21**, 128 (2020).

185. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

186. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, 1–17 (2014).

187. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci.* **107**,

9546–9551 (2010).

188. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
189. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, 1–9 (2010).
190. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
191. Van den Berge, K. *et al.* RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. *Annu. Rev. Biomed. Data Sci.* **2**, 139–173 (2019).
192. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
193. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
194. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
195. Benjamini, Y. & Yekutieli, D. The Control of the False Discovery Rate in Multiple Testing under Dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
196. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035 (2003).
197. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).

198. Nowicka, M. & Robinson, M. D. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research* **5**, 1356 (2016).
199. Papastamoulis, P. & Rattray, M. Bayesian estimation of differential transcript usage from RNA-seq data. *Stat. Appl. Genet. Mol. Biol.* **16**, 367–386 (2017).
200. Audic, S. & Claverie, J. M. The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995 (1997).
201. Pornputtapong, N. *et al.* KITSUNE: A Tool for Identifying Empirically Optimal K-mer Length for Alignment-Free Phylogenomic Analysis. *Front. Bioeng. Biotechnol.* **8**, 556413 (2020).
202. Bai, X., Tang, K., Ren, J., Waterman, M. & Sun, F. Optimal choice of word length when comparing two Markov sequences using a χ 2-statistic. *BMC Genomics* **18**, 732 (2017).
203. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinforma. Oxf. Engl.* **27**, 764–770 (2011).
204. Manekar, S. C. & Sathe, S. R. A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience* **7**, (2018).
205. Rizk, G., Lavenier, D. & Chikhi, R. DSK: k-mer counting with very low memory usage. *Bioinforma. Oxf. Engl.* **29**, 652–653 (2013).
206. Erbert, M., Rechner, S. & Müller-Hannemann, M. Gerbil: a fast and memory-efficient k-mer counter with GPU-support. *Algorithms Mol. Biol. AMB* **12**, 9 (2017).
207. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinforma. Oxf. Engl.* **33**, 2759–2761 (2017).
208. Sacomoto, G. A. T. *et al.* KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics* **13 Suppl 6**, S5 (2012).

209. Jaillard, M., Palmieri, M., van Belkum, A. & Mahé, P. Interpreting k-mer–based signatures for antibiotic resistance prediction. *GigaScience* **9**, (2020).
210. Souvorov, A., Agarwala, R. & Lipman, D. J. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol.* **19**, 153 (2018).
211. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
212. Jaillard, M. *et al.* A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet.* **14**, e1007758 (2018).
213. Bernard, G., Greenfield, P., Ragan, M. A. & Chan, C. X. k-mer Similarity, Networks of Microbial Genomes, and Taxonomic Rank. *mSystems* **3**, (2018).
214. Drouin, A. *et al.* Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics* **17**, 754 (2016).
215. Drouin, A. *et al.* Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci. Rep.* **9**, 4071 (2019).
216. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16**, 236 (2015).
217. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* **19**, 198 (2018).
218. Buchkovich, M. L. *et al.* HLAProfiler utilizes k-mer profiles to improve HLA calling accuracy for rare and common alleles in RNA-seq data. *Genome Med.* **9**, 86 (2017).
219. Baizan-Edge, A. *et al.* Kodoja: A workflow for virus detection in plants using

k-mer analysis of RNA-sequencing data. *J. Gen. Virol.* **100**, 533–542 (2019).

220. Audemard, E. O. *et al.* Targeted variant detection using unaligned RNA-Seq reads. *Life Sci. Alliance* **2**, (2019).
221. Nordström, K. J. V. *et al.* Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k -mers. *Nat. Biotechnol.* **31**, 325–330 (2013).
222. Shajii, A., Yorukoglu, D., William Yu, Y. & Berger, B. Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics* **32**, i538–i544 (2016).
223. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 1–14 (2016).
224. Menzel, M., Hurka, S., Glasenhardt, S. & Gogol-Döring, A. NoPeak: k-mer-based motif discovery in ChIP-Seq data without peak calling. *Bioinforma. Oxf. Engl.* **37**, 596–602 (2021).
225. Vellichirammal, N. N., Albahrani, A., Li, Y. & Guda, C. Identification of Fusion Transcripts from Unaligned RNA-Seq Reads Using ChimeRScope. *Methods Mol. Biol. Clifton NJ* **2079**, 13–25 (2020).
226. Li, A., Zhang, J. & Zhou, Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics* **15**, 311 (2014).
227. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
228. Rahman, A., Hallgrímsdóttir, I., Eisen, M. & Pachter, L. Association mapping from sequencing reads using k-mers. *eLife* **7**, (2018).
229. Mehrab, Z., Mobin, J., Tahmid, I. A. & Rahman, A. Efficient association mapping

from k-mers—An application in finding sex-specific sequences. *PLOS ONE* **16**, e0245058 (2021).

230. Audoux, J. *et al.* DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. *Genome Biol.* **18**, 243 (2017).
231. Lopez-Maestre, H. *et al.* SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Res.* **44**, e148–e148 (2016).
232. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. (MIT press, 2016).
233. Ruder, S. An overview of gradient descent optimization algorithms. *ArXiv Prepr. ArXiv160904747* (2016).
234. Kelley, H. J. Gradient theory of optimal flight paths. *Ars J.* **30**, 947–954 (1960).
235. Chikhi, R., Holub, J. & Medvedev, P. Data Structures to Represent a Set of *k*-long DNA Sequences. *ACM Comput. Surv.* **54**, 17:1-17:22 (2021).
236. Marchet, C. *et al.* Data structures based on k-mers for querying large collections of sequencing data sets. *Genome Res.* **31**, 1–12 (2021).
237. Deorowicz, S., Debudaj-Grabysz, A. & Grabowski, S. Disk-based k-mer counting on a PC. *BMC Bioinformatics* **14**, 160 (2013).
238. Park, G., Hwang, H.-K., Nicodème, P. & Szpankowski, W. Profiles of Tries. *SIAM J. Comput.* **38**, 1821–1880 (2009).
239. Social Media's Impact on the 2020 Presidential Election: The Good, the Bad, and the Ugly. http://research.umd.edu/news/news_story.php?id=13541.
240. Green, E. D. *et al.* Strategic vision for improving human health at The Forefront of Genomics. *Nature* **586**, 683–692 (2020).
241. Sun, Q., Peng, Y. & Liu, J. A reference-free approach for cell type classification with scRNA-seq. *iScience* **24**, 102855 (2021).

Annexes

A cell-to-patient machine learning transfer approach uncovers novel basal-like breast cancer prognostic markers amongst alternative splice variants

During my last year of PhD, I contributed to the revision of the following paper. The main contributions before the revisions were the discussions about an effective strategy to apply to effectively transfer the information from cell line data to patients. During the revisions, my main contributions were the cleaning of the code and the implementation of the mixed feature model.

Villemin et al. BMC Biology (2021) 19:70
https://doi.org/10.1186/s12915-021-01002-7

BMC Biology

RESEARCH ARTICLE Open Access

A cell-to-patient machine learning transfer approach uncovers novel basal-like breast cancer prognostic markers amongst alternative splice variants

Jean-Philippe Villemin, Claudio Lorenzi, Marie-Sarah Cabrillac, Andrew Oldfield, William Ritchie* and Reini F. Luco*

Abstract

Background: Breast cancer is amongst the 10 first causes of death in women worldwide. Around 20% of patients are misdiagnosed leading to early metastasis, resistance to treatment and relapse. Many clinical and gene expression profiles have been successfully used to classify breast tumours into 5 major types with different prognosis and sensitivity to specific treatments. Unfortunately, these profiles have failed to subclassify breast tumours into more subtypes to improve diagnostics and survival rate. Alternative splicing is emerging as a new source of highly specific biomarkers to classify tumours in different grades. Taking advantage of extensive public transcriptomics datasets in breast cancer cell lines (CCLE) and breast cancer tumours (TCGA), we have addressed the capacity of alternative splice variants to subclassify highly aggressive breast cancers.

Results: Transcriptomics analysis of alternative splicing events between luminal, basal A and basal B breast cancer cell lines identified a unique splicing signature for a subtype of tumours, the basal B, whose classification is not in use in the clinic yet. Basal B cell lines, in contrast with luminal and basal A, are highly metastatic and express epithelial-to-mesenchymal (EMT) markers, which are hallmarks of cell invasion and resistance to drugs. By developing a semi-supervised machine learning approach, we transferred the molecular knowledge gained from these cell lines into patients to subclassify basal-like triple negative tumours into basal A- and basal B-like categories. Changes in splicing of 25 alternative exons, intimately related to EMT and cell invasion such as ENAH, CD44 and CTNND1, were sufficient to identify the basal-like patients with the worst prognosis. Moreover, patients expressing this basal B-specific splicing signature also expressed newly identified biomarkers of metastasis-initiating cells, like CD36, supporting a more invasive phenotype for this basal B-like breast cancer subtype.

(Continued on next page)

* Correspondence: william.ritchie@igh.cnrs.fr; reini.luco@igh.cnrs.fr
Institut de Génétique Humaine (IGH-UMR9002), Centre National de la Recherche Scientifique (CNRS), University of Montpellier, Montpellier, France

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(Continued from previous page)
Conclusions: Using a novel machine learning approach, we have identified an EMT-related splicing signature capable of subclassifying the most aggressive type of breast cancer, which are basal-like triple negative tumours. This proof-of-concept demonstrates that the biological knowledge acquired from cell lines can be transferred to patients data for further clinical investigation. More studies, particularly in 3D culture and organoids, will increase the accuracy of this transfer of knowledge, which will open new perspectives into the development of novel therapeutic strategies and the further identification of specific biomarkers for drug resistance and cancer relapse.

Keywords: Alternative splicing, Breast Cancer, Survival, Basal-like, Epithelial-to-mesenchymal transition, Machine learning classification

Background

Breast cancer is a heterogenous disease with multiple molecular drivers and disrupted regulatory pathways [1, 2]. The development of large-scale genomics and transcriptomics methods has increased the capacity to identify clinically-relevant tumour subtypes with distinct molecular signatures. These can be used for a better choice of treatment and/or prediction of potential metastasis which can improve survival outcome [3, 4]. However, patients are still facing a high percentage of misdiagnosis in which undetected early metastasis and/or inappropriate choice of treatment can lead to deadly complications with the use of unnecessary severe chemotherapies or the apparition of drug resistance and subsequent tumour relapse [5]. Currently, breast cancer is classified into five major categories (normal-like, luminal A, luminal B, Her2-positive and basal-like) based on expression of three receptors: oestrogen and progesterone hormonal receptors (ER and PR) and the epidermal growth factor receptor ERBB2 (Her2). Basal-like are the most aggressive, and difficult to treat, type of breast cancer tumour. They are usually negative for the three receptors, and thus called triple negative breast cancer (TNBC), which represents 10–20% of all breast cancers. These tumours are usually found in younger patients with a larger size and higher probability of lymph node infiltration and metastasis [2, 6]. Furthermore, the absence of all three receptors reduces the number of targeted therapeutic strategies to be used, leaving nonspecific chemotherapy as the standard treatment of choice, which soon leads to dose-limiting side-effects, resistance to treatment and finally clinical relapse in less than 5 years [6]. A better understanding of the molecular differences in between these tumour categories will improve the choice of treatment and detection of early metastasis, which will significantly impact patient's outcome. There have been many attempts to identify novel therapeutic targets and/or prognostic biomarkers to better subclassify breast cancer tumours [7]. Over 170 independent breast cancer susceptibility genomic variants have been identified. Many of which have been associated with a specific tumour category, such as ER positiveness or Her2 amplification. However, no clear subcategories exist despite tumour heterogeneity and differences in clinical response to treatment and tumour relapse within the same category [8–10]. Interestingly, alternative splicing is an emerging source of new biomarkers and therapeutic targets in cancer [11–15].

The alternative processing of mRNA precursors enables one gene to produce multiple protein isoforms with different functions, increasing protein diversity and the capacity of a cell to adapt to new environments. An increasing number of splice variants, and their respective splicing regulators, have been shown to confer a selective advantage to tumour cells. For instance, the splicing regulators RBM5, 6 and 10 favour tumour cell proliferation and colony formation by regulating the alternative splicing of the membrane-bound protein NUMB [16]. Post-translational activation of the splicing factor SRSF1 (also known as ASF/SF2) confers resistance to apoptosis by inducing inclusion of the anti-apoptotic splice variant in a network of functionally related genes, such as *Bcl-X* and *Mcl1* [17]. Regulation of VEGF splicing is detrimental for stimulation of angiogenesis [18]. A change in the alternative splicing of the pyruvate kinase pre-mRNA can switch tumour cells metabolism to adapt to the increased proliferation [19, 20]. Finally, a list of well-known alternatively spliced variants related to cell adhesion (CTNND1, CD44) and cytoskeleton organisation (ENAH, FLNB) is responsible for the acquisition of migratory and invasive phenotypes necessary for distal metastasis [13, 21–24]. The existence of functionally relevant cancer-specific isoforms is therefore a promising new source of highly specific and less toxic therapeutic targets for the development of isoform-specific antibodies and/or splice-switching antisense oligonucleotides [25, 26].

By taking advantage of an extensive transcriptomics and anti-tumour compound screening information publicly available in cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) [27], we identified a splicing signature that can stratify basal breast cancer cell lines into two well-known subtypes, basal A and basal B. In contrast to basal-like breast cancer patients, basal breast cancer cell lines are divided into two subgroups, basal A and basal B, depending on the expression profile of a subset of basal

(cytokeratins, integrins), stem cell (CD44, CD24) and mesenchymal markers (Vimentin, fibronectin, MSN, TGFBR2, collagens, proteases) [28–30]. Basal B cell lines are mostly triple negative breast cancer cells that express classical mesenchymal and stem cell markers characteristic of the epithelial-to-mesenchymal transition (EMT), a biological process in which epithelial cells acquire mesenchymal features that are advantageous for the cancer cell, such as increased cell motility to invade distal organs in metastasis, resistance to apoptosis, refractory responses to chemotherapy and immunotherapy, and acquisition of stem cell-like properties like in cancer stem cells [31, 32]. In concordance, basal B cells are morphologically less differentiated, with a mesenchymal-like shape, and a more invasive phenotype in culture assays than basal A and luminal cells [28, 33, 34]. We aimed to transfer this basal A/basal B splicing classification into the clinic by using a semi-supervised machine learning approach. We successfully classified 40% of basal-like breast cancer patients (75/188) from the Cancer Genome Atlas (TCGA) [35] as basal B-like based on a unique 25 spliced gene signature characteristic of cells undergoing EMT. In this signature, we found well-known markers of malignancy, such as ENAH EMT splice variant that promotes lung metastasis [36] or CSF1 variant which promotes macrophage infiltration and distal metastasis [37], together with new promising splicing candidates of tumour progression and invasiveness (PLOD2, CTNND1, SPAG9). Finally, expression of this basal B signature was sufficient to identify triple negative breast cancer tumours with poor survival, highlighting the prognostic value of the newly identified splicing biomarkers to subclassify one of the most heterogenous and difficult to treat type of breast cancer. More studies in cell lines, particularly regarding resistance to treatment and cell invasion will be essential to refine this splicing signature in view of orienting treatment or predicting metastasis sites.

In conclusion, by adapting a machine learning approach, we were able to transfer the molecular knowledge obtained in experimental cell lines to identify novel biomarkers of poor prognosis and metastasis amongst triple negative breast cancers in patients. Furthermore, the study of the regulatory pathway involved in this specific splicing signature pointed to RBM47 as one of the splicing regulators responsible for the basal B-specific splicing signature, and for which differential expression levels also correlate with distinct prognostic values, turning this splicing factor a promising novel therapeutic target. Further clinical and functional validation of the 25 splicing events proposed in our basal B-specific splicing signature will open new perspectives in the understanding of triple negative breast cancers and the improvement of currently available therapeutic strategies and survival outcome.

Results

A distinctive basal B-like breast cancer splicing signature

Data mining of large-scale genomics and transcriptomics datasets in breast cancer cell lines are a promising source of novel biomarker and therapeutic targets [23, 38, 39]. We sought to leverage the wealth of transcriptomics and functional data available in cancer cell lines to better understand different profiles of breast cancer. Hierarchical clustering of changes in alternative splicing of cassette exons and gene expression profile of 80 breast cancer cell lines from two extensive and complementary projects (Additional file 2: Table S1) revealed basal B cell lines as a distinctive group of cells with an expression and splicing profile significantly different from basal A and luminal cancer cells (Additional file 1: Fig. S1). To identify the transcriptional signature characteristic of basal B cells, we repeated the hierarchical clustering in just basal A and basal B cell lines to merge all the differentially expressed and spliced transcripts responsible for the segregation of basal B cell lines (Fig. 1). We found 635 genes and 217 spliced isoforms with significantly different levels between basal A and basal B cells (Fig. 1a, b). In line with published tissue-specific and EMT transcriptomics analyses [40–42], most of the genes differentially spliced were not affected at the expression level, suggesting that two different subsets of genes, and thus regulatory layers, are responsible for the basal B phenotype (Fig. 1c). Gene set enrichment analysis (GSEA) [43] between basal B and basal A cells confirmed the EMT and stem cell-like phenotype characteristic of basal B cell lines (Fig. 2a, b), which was supported with a higher CD44+/CD24− stem cell score (Fig. 2e) [28–30]. DAVID gene ontology analysis of differentially expressed and spliced genes also underlined biological terms that are hallmarks of EMT and cell invasiveness, such as cell-cell junction (Fig. 2d) [44]. However differentially expressed genes were also enriched in their own unique terms, related to extracellular vesicles/plasma membrane organisation. Whilst differentially spliced genes were specifically enriched in terms related to GTPase activity, cytoskeletal protein and cadherin binding, which reinforces the existence of two complementary regulatory pathways (Fig. 2d). Finally, another malignant characteristic acquired by cancer cells undergoing EMT is resistance to chemotherapy, which often leads to clinical relapse. Gene set enrichment analysis found upregulation of genes resistant to the Epidermal Growth Factor Receptor (EGFR) inhibitor Gefitinib (Fig. 2c), which is an alternative to hormonal therapy in Her2+ breast cancer tumours, but is not efficient in triple negative tumours [45]. Available drug assays from the Genome Drug Sensitivity in Cancer portal (GDSC) [46] confirmed the need of a higher concentration (IC50) of Gefitinib, and other EGFR inhibitors (Erlotinib,

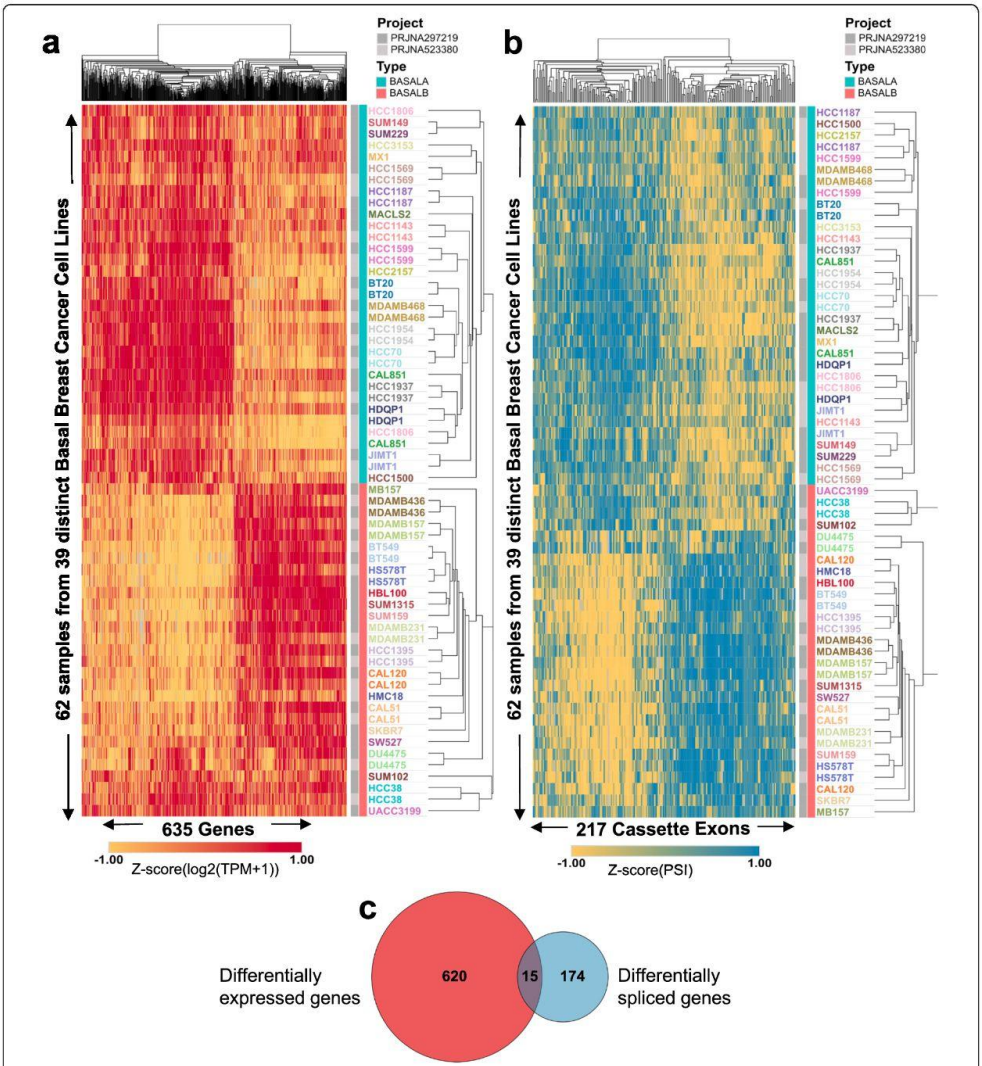


Fig. 1 Basal cell lines are divided in two subgroups based on gene expression and splicing patterns. **a** Heatmap of the transcripts per million (TPM) values of the 635 genes which differential expression can cluster breast cancer cell lines into basal A and basal B (P value $< 10^{-3}$ by Kruskal-Wallis test). **b** Heatmap of the percentage spliced-in (PSI) values of the 217 exons which differential splicing can cluster breast cancer cell lines into basal A and basal B (P value $< 10^{-3}$ by Kruskal-Wallis test). **c** Venn Diagram of the genes differentially expressed and/or spliced between basal A and basal B cancer cell lines. The overlap is not higher than expected by Fisher's exact test, two tail ($p = 0.098$)

Sapitinib), to have the same deleterious effect on basal B compared to basal A cancer cells (Fig. 2f). Basal B cell lines also showed a significant resistance to well-known inhibitors of the cell cycle (irinotecan, taselisib, 5-fluorouracil), drug inducers of cell death (AZD5582, AZD5991) and other receptor tyrosine kinase inhibitors, such as savolitinib which inhibits c-MET to reduce tumour persistence and metastasis [47].

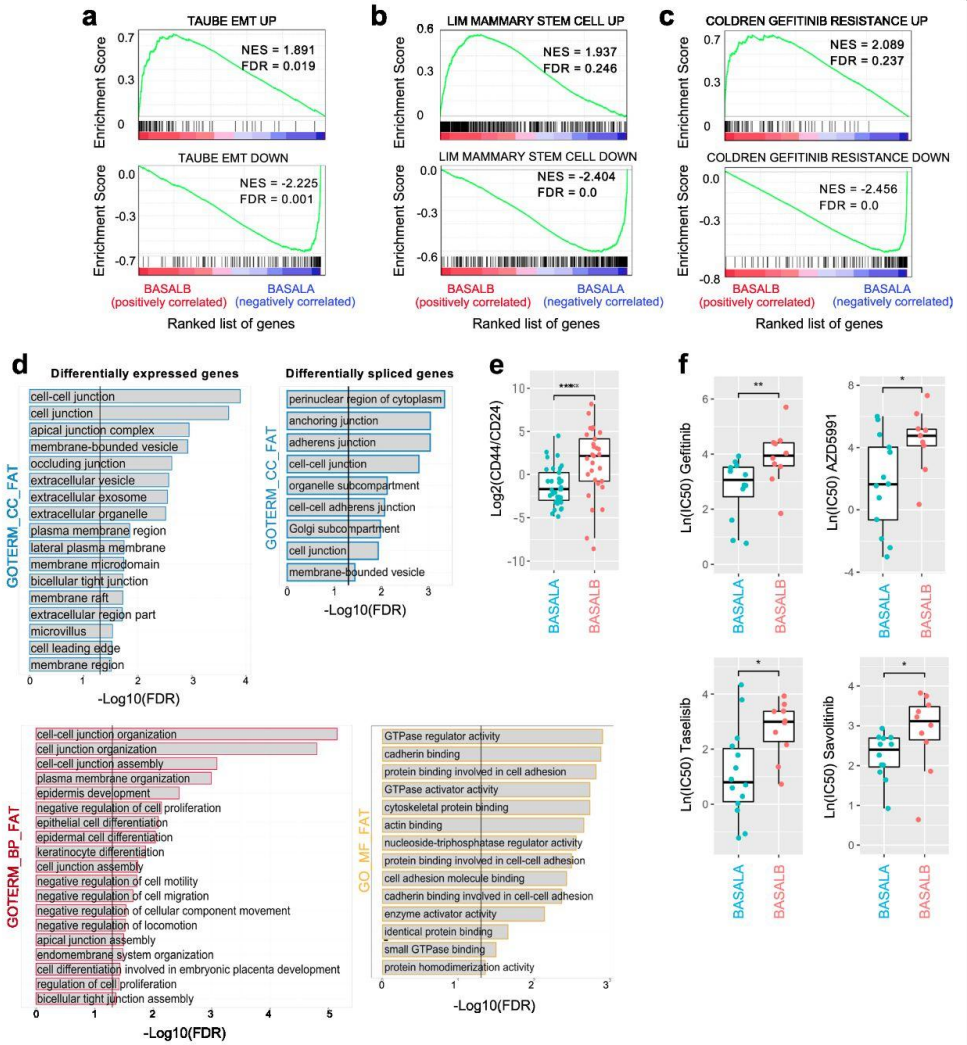


Fig. 2 Basal B cell lines show mesenchymal, stem-like and resistance to treatment characteristics. **a**–**c** Gene Set Enrichment Analysis (GSEA) of differentially expressed genes between basal A and B cell lines for three different signatures: Mammary Stem Cell, EMT and Resistance to Gefitinib. Up-regulated genes in all signatures are enriched in basal B cell lines (FDR < 0.25). **d** Gene ontology analysis bar graphs for differentially expressed (left) and differentially spliced (right) genes between basal A and B cell lines. Gene ontology terms related to Cellular Component (GO_CC_FAT), Molecular Function (GO_MF_FAT) and Biological Process (GO_BP_FAT) are shown in the y axis in blue, yellow and red, respectively. Benjamini false discovery rate (FDR, −log10) is shown on the *x*-axis. Vertical lines mark an FDR threshold of FDR = 0.05 (−log10(0.05) = 1.3) for differentially expressed and spliced genes, respectively. **e.** Box plots of the median and 25th percentile of the CD44/CD24 log2 expression ratio for basal A and B cell lines. *P* value is calculated using the Wilcoxon rank-sum test. **f** Boxplots comparing IC50 values in basal A and B cell lines upon treatment with different drugs from the Genomics of Drug Sensitivity in Cancer 2 (GDS2) dataset. *P* values are calculated using the Wilcoxon rank-sum test

In summary, we have identified two distinct transcriptional and splicing signatures, specific of basal B cell lines, that underline an EMT phenotype with molecular characteristics related to cell invasion, stemness and resistance to chemotherapy. We next sought to investigate whether this basal B-specific splicing signature could also be used to subclassify basal-like/triple negative breast cancer patients.

A semi-supervised machine learning approach to subclassify basal-like breast cancer patients

As a first and simple approach, we performed a hierarchical clustering followed by a k-means clustering (k = 2 for "A-like" and "B-like") of the 188 patients, annotated as basal-like in The Cancer Genome Atlas Program (TCGA), using the 635 differentially expressed or 217 differentially spliced cassette exons characteristic of basal B cell lines (Additional file 1: Fig. S2a,b). Using such method, patients were forced to classify in one of the two groups based on differences in gene expression or splicing patterns. Since basal B cell lines show more invasive, cancer stem cell-like phenotypes, we assessed whether these aggressive characteristics were translated to the "B-like" patient group through differences in disease specific survival (DSS) rates. Kaplan-Meier analysis of DSS did not show significant differences between the two subgroups of basal-like patients (Additional file 1: Fig. S2c,d). However, we did observe a tendency for "B-like" patients to have a poor survival compared to "A-like" when just looking at differences in splicing, contrary to expression levels (p value = 0.09 vs 0.57, respectively—Additional file 1: Fig. S2c,d).

In fact, it was not surprising that the transcript-level and splicing signatures did not translate directly from simplistic cell culture models to much more complex tumour patients with specific cell micro-environments and differences in cell heterogeneity. However, because the patients showed clear "A-like" and "B-like" signatures, we sought to develop a machine learning approach that would allow us to transfer part of the molecular and phenotypic observations found in cell-lines to patient data. Transfer learning is a recent research methodology that focuses on storing the knowledge gained when solving a problem, to apply it to a different, but related, one. Because we wanted to ensure that the newly developed cell-to-patient transfer learning algorithm could create interpretable models, we used a decision tree-based approach called Random Forest. In this cell-to-patient random forest classification method, we started by classifying basal A or basal B cell-lines based on their splicing and/or expression profile (Fig. 3a and Additional file 1: Figs. S3-S4). Then, once the model was trained on cell-lines, we would start integrating patient data gradually into the model. This was done iteratively by integrating at each round of classification the patients best predicted to be basal A-like and basal B-like, so their added informative value could be used back to train the system and improve the next round of classification (Fig. 3a). With this semi-supervised approach, the probability of assigning a patient to a specific subgroup evolves and improves at each round based on the updated information obtained from the best predicted patients, reaching at the end a stable population with the labels 'basal A-like', 'basal B-like' or 'unclassified' determined by the algorithm after 10–12 rounds (Fig. 3b,c and Additional file 1: Figs. S3b,c-S4b,c). Thanks to the gradual addition of patients at each round of training, there is a progressive increase, or decrease, in the feature importance of the splicing variants used to classify patients (Fig. 3d and Additional file 1: Figs. S3d-S4d). Out of the 188 basal-like patients, 75 were classified as basal B-like, 88 as basal A-like and 25 could not be classified based on their splicing signature. Using only expression levels, there was a slight biased towards the basal A-like phenotype, with 56 patients classified as basal B-like, 122 as basal A-like and 10 unclassified (Additional file 1: Fig. S3b-c). Combining differentially spliced and expressed features seemed to be the most performant classifier with 84 patients as basal B-like, 100 as basal A-like and just 4 unclassified (Additional file 1: Fig. S4b-c). Taken together, depending on the features used (splicing patterns, expression levels or both), patients were differently classified in basal A-like or basal B-like.

An EMT-related basal B-specific splicing signature that marks poor prognosis

To address which classifier translates the best to patients the invasive, EMT-like and drug-resistant basal B phenotype found in cancer cells, we calculated the 5-year survival rate for each group of basal A-like and basal B-like issued from the three types of classification. Only basal B-like patients classified based on splicing levels had a poor prognosis compared to basal A-like patients (log-rank test p = 0.0067, HR = 4.87; 95% IC: [1.37–17.28] in Kaplan-Meier analysis and univariate Cox regression) (Fig. 3e). Basal B-like patients subclassified based on gene expression levels, or gene expression and splicing features, did not show significant differences in disease survival rate (Additional file 1: Fig.S3e-4e), suggesting that splicing biomarkers might be more informative to further subclassify basal-like patients based on prognosis. We thus decided to focus on the role of alternative splicing in identify triple negative basal-like breast cancer with poor prognosis.

To extract the most informative splicing features from the cell-to-patient transfer learning classifier, we used the Boruta feature selection method [48]. This allowed us to select the key splicing events responsible for the

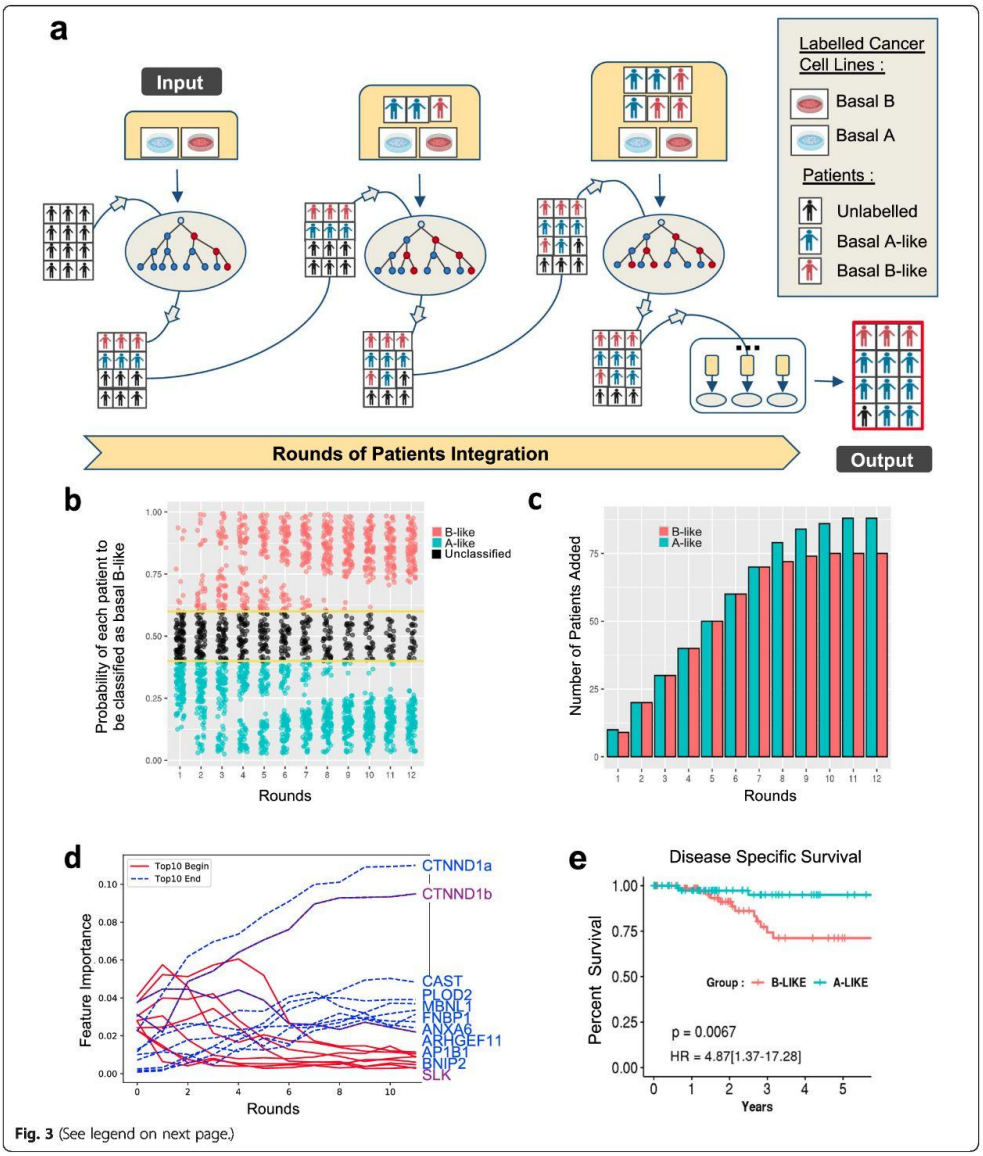


Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 A Random Forest Classifier using knowledge transfer from cell lines to patients. **a.** Workflow scheme: a random forest (RF) model is built using cell lines labelled as basal B (red) or basal A (blue). It is then run iteratively, integrating at each round patients whose probability to be classified in one group or the other is amongst the ten highest. The classifier stops when no more patients can be classified. **b** Probability of a basal-like patient to be classified as basal B-like, basal A-like or unclassified over each round. Yellow lines indicate thresholds used to classify a patient as basal B-like (> 0.6) or basal A-like (< 0.4). **c** Bar plot of the number of patients added at each round. Patients with the highest probability to be classified are sequentially incorporated to the input cell lines in order to create a new classifier for the next round of integration. **d** Evolution of the feature importance at each round of iterative training. In red are the 10 splicing variants (features) most informative at the beginning of the transfer learning process. In blue are the 10 splicing variants most informative at the end. Only two exons remained informative from the beginning to the end (in blue and red). The name of the top 10 final most informative spliced genes are written in blue and in sequential order. **e** Kaplan-Meier plots of disease specific survival in basal A-like (blue) and basal B-like patients (red). Hazard ratio (HR) and logrank *p* value (*P*) discriminating the two groups are shown

basal A/B classification without the need to predefine arbitrary thresholds (Fig. 4a). Out of the 217 differentially spliced exons between basal A/B cell lines, just 25 were needed to subclassify breast cancer patients in basal A or basal B-like tumours (Fig. 4a and Additional file 3: Table S2). Sashimi plots representing the splicing patterns of some of these basal B-specific splicing events, such as the well-known splicing biomarker of cancer metastasis ENAH [26] and the newly identified splicing biomarkers PLOD2, SPAG9 and KIF13a, validated the observed changes in splicing between basal A and basal B-like patients (Fig. 4b-c and Additional file 1: Fig. S5a-b). Moreover, the changes in percentage of spliced-in (PSI) of the 25 basal B-specific splicing events between the two subtypes of basal-like patients correlated with the observed splicing changes between basal A/B cell lines (Additional file 1: Fig. S5c-d), further supporting the transfer of knowledge from the laboratory to the clinic. Finally, in the absence of publicly available RNA-seq data on a second cohort of basal-like breast cancer patients, we took advantage of three independent sequencing projects on breast cancer cell lines, different from the ones used for the training of the semi-supervised classifier (Additional file 2: Table S1). Distribution of 52 independent breast cancer cell lines showed a 93% accuracy in the spatial segregation (t-SNE) of basal A from basal B cells based on the splicing pattern of the 25 newly identified splicing events (Fig. 4d). Just three cell lines were misclassified as basal A (HCC38, SUM102 and MDA-MB-157). It is worth noting that one of these, HCC38, was also labelled as basal A in the DepMap portal (www.depmap.org), which validated our methodology and the specificity of the splicing signature towards a basal B-like phenotype.

Consistent with basal B cell lines being more mesenchymal, differences in the alternative splicing of these 25 basal B-specific splicing events in four different cellular models of EMT, coming from different cell types and methods of EMT induction [49–52], successfully clustered epithelial cells from mesenchymal with a pattern of splicing equivalent to basal A and basal B-like patients, respectively (Fig. 4e). Of note, another 25 gene-based EMT-like splicing signature characteristic of luminal breast cancer tumours has also been identified capable of subclassifying mesenchymal-like breast cancer tumours with poor prognosis [38]. Consistent with a more luminal-specific signature, despite both marking EMT phenotypes, not more than six splicing events were found in common between the two splicing signatures (ATP5C1, CTNND1, KIF13a, PLOD2, SEC31a and SPAG9), which further supports the specificity of our newly identified splicing signature for basal-like triple negative breast cancer. Finally, using one of the first established molecular subtypes of triple negative breast cancer tumours based on gene expression, which is the Lehman classification [53], we found that basal B-like patients are mostly found in the categories associated with mesenchymal stem-like (MSL) and immunomodulatory (IM) subtypes (Fig. 5a), which goes in line with a gene set enrichment of terms related to inflammatory responses and hallmark of EMT (Fig. 5b).

When looking at the expression of well-known basal and EMT biomarkers in the two subpopulations of basal A/B-like patients, we found that basal A-like patients express classical basal/epithelial markers, such as E-cadherin, EPCAM and cytokeratin KRT5/KRT6/KRT14, together with ERBB3 and TOB1 which are markers of more differentiated, non-invasive cells [2]. On the other hand, basal B-like patients express classical EMT/mesenchymal markers such as Fibronectin, the EMT inducers Twist and Slug, and the Zinc-finger transcriptional regulators Zeb1 and Zeb2 which have recently been shown to confer stemness properties that can increase the plasticity and invasive capacity of the tumour cells [54] (Fig. 5c, d). In line with a more aggressive, invasive phenotype, basal B-like patients express cytoskeletal (MSN, FN1) and extracellular matrix signalling proteins (TGFB1, TGFBR2, FBN1, AXL), collagens (COL3A1, COL6A3) and proteases (MMP2, TIMP1, CTSC, PLAU, SERPINE1/2, PLAT), which are necessary for cell's migration and dissemination to distal organs during metastasis [2]. Finally, basal B-like patients overexpress a recently identified new marker of metastasis-initiating cells, the fatty acid receptor CD36 [20]. Clinically, the presence of CD36-positive cells has

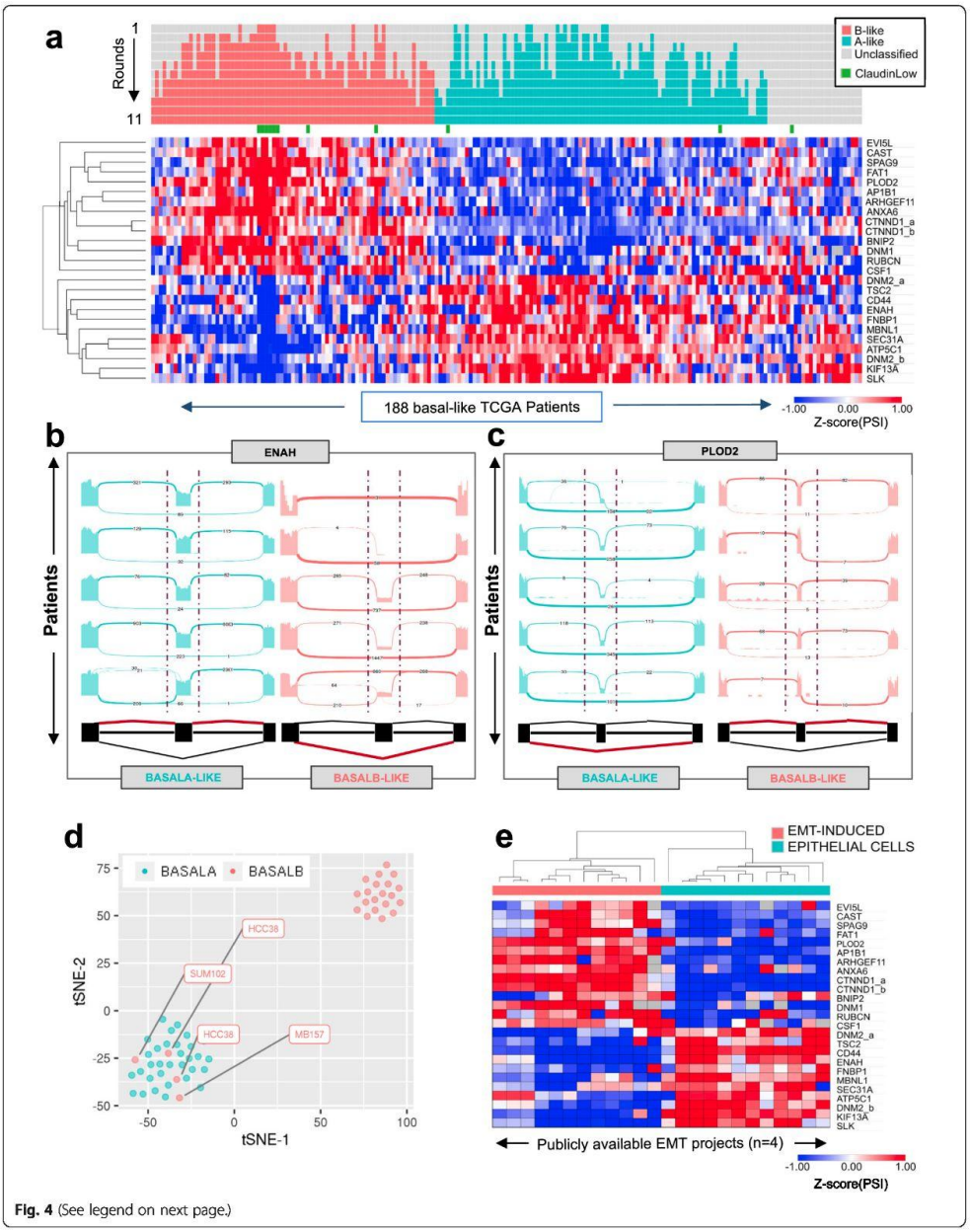


Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 The basal B-specific splicing signature is associated to EMT features. **a** Heatmap of the Percentage Spliced-In (PSI) values of the 25 cassette exons most informative to classify TCGA basal-like patients into basal B-like (red) or basal A-like (blue). Claudin low tumours are highlighted in green. **b**, **c** Sashimi plots displaying ENAH and PLOD2 splicing patterns in randomly selected patients classified as basal A-like and basal B-like. **d** Changes in alternative splicing of these 25 basal B-specific splicing events is sufficient to properly cluster 55 basal breast cancer cell lines from 3 unrelated sequencing projects into basal B and basal A using t-SNE. Of note, three basal B cell lines, HCC38, MDA-MB-157 and SUM102 were misclassified as Basal A cell lines (red dots). Although HCC38 has also been classified as Basal A in the DepMap portal (www.depmap.org). **e** Heatmap of the PSI values of the 25 basal B-specific splicing signature in public RNA-seq datasets from four different EMT projects. Basal B-like events have the same splicing patterns as EMT-induced cells

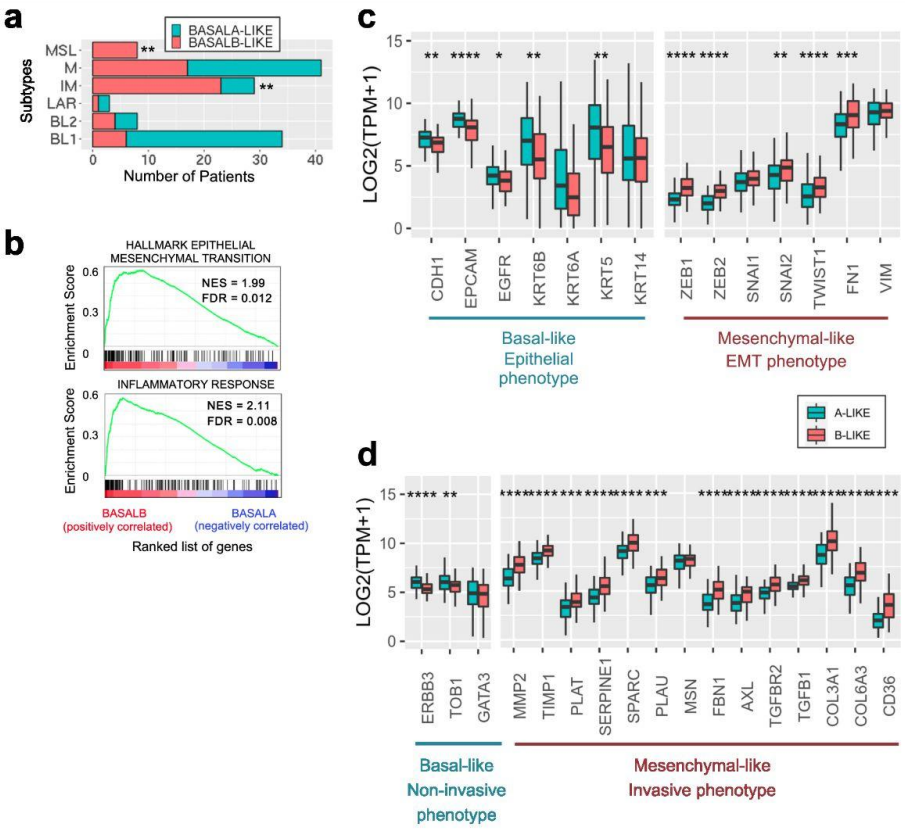


Fig. 5 Basal B-like patients express hallmarks of EMT and metastasis that leads to a poor prognosis. **a** Lehman classification for basal A- and B-like patients. $^{**}p < 0.01$ in Fisher's exact test, two tail, comparing basal B to basal A. **b** Gene Set Enrichment Analysis (GSEA) of the genes differentially expressed between basal A- and B-like patients. Hallmark EMT and inflammatory response signatures are enriched in basal B-like patients. **c.** Box plots of the median and 25th percentile of the expression levels (in TPM) of major epithelial and mesenchymal-like EMT markers in basal A-like (blue) and basal B-like (red) patients. **d** Box plot of the mean and 25th percentile of the expression levels (in TPM) of Basal-like non-invasive and mesenchymal-like invasive markers in basal A-like (blue) and basal B-like (red) patients. ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$ in Wilcoxon rank-sum test comparing basal A-like to basal B-like

been correlated with a lower survival rate in many carcinomas, including breast cancer, and inhibition of CD36 impairs metastasis in breast cancer-derived tumours, turning this receptor into an important biomarker of tumour cell dissemination and a potential new target to reduce cell invasion. The fact that basal B-like tumour cells co-express this metastasis-initiating marker further strengthens the aggressive nature of this tumour subclass and the clinical relevance of the basal B-specific splicing signature in tumour progression and relapse.

Overall, we have identified a novel splicing signature, specific of triple negative breast cancer tumours, that marks patients with the poorest prognosis. This basal B-like splicing signature is responsible of a stem-like, EMT phenotype that favours tumour growth, invasion of distal organs and increased drug resistance, which eventually leads to tumour relapse and metastasis. Interestingly, some of the genes differentially expressed in these basal B-like patients are well-known markers of metastasis-initiating cells, such as the alternatively spliced CTNN D1 and PLOD2 genes or the fatty acid receptor CD36, turning these biomarkers into promising new targets for innovative therapies, such as the use of splicing specific antibodies [6, 26].

A metastasis-related common regulatory pathway for the basal B-specific splicing signature

Hierarchical clustering of basal A and B cell lines based on the differential expression of RNA-binding proteins highlighted six RNA regulators, ESRP1, ESRP2, RBM47, TMEM63A, KRR1 and RBMS3 (Fig. 6a) (Kruskal-Wallis $p < 10^{-9}$). Interestingly, ESRP1/2 and RBM47 are significantly less expressed in basal B-like than basal A-like patients (Fig. 6b), consistently with the known inhibitory effect of these three splicing regulators in EMT progression and metastasis [52, 55, 56]. Available transcriptomics data in ESRP1/2 and RBM47 lung carcinoma NCI-H358-depleted cells [52] and RBM47 overexpressing breast cancer metastatic MDA-MB-231 cells [57] showed that 19 of the 25 splicing events responsible for the newly identified basal B-specific splicing signature could potentially be regulated by ESRP1/2 and/or RBM47 in breast cancer cells (Fig. 6c, d). Importantly, in the cell types analysed, ESRP1/2 and RBM47 induce the epithelial, basal A-like splicing phenotype, suggesting a potential tumour suppressor effect for these splicing regulators (Figs. 6e–g, 4e and Additional file 1: S5c-d). Consistently with this observation, low expression of RBM47 in basal-like breast cancer patients was associated with poor overall survival (log-rank test $p = 0.031$, HR = 3.36, 95% IC:[1.05–10.79] Fig. 6h, i), which supports previous experimental evidence of a role for RBM47 in supressing breast cancer metastasis and progression [56]. In fact, RBM47-dependent basal B-specific splicing events were found to be functionally interconnected by physical and/or genetic interactions, which points to the existence of a common basal B-specific regulatory network associated with tumour malignancy (Additional file 1: Fig. S6a). In support, most of RBM47-dependent basal B-specific splicing events play well-known roles in cell-cell adhesion (CTNND1) [58], cytoskeleton organisation (ENAH, SLK, FNBP1) [59, 60], endocytosis (KIF13A, DNM2) [61] and association with the extracellular matrix (PLOD2) [62], which are all key processes for gaining the cell motility and invasiveness necessary in tumour metastasis (54–58). Of note, expression of just one of these basal B-specific splice variants, which are CTNND1, ENAH and PLOD2, is sufficient to lower the disease-specific survival rate of basal B-like breast cancer patients compared to basal A-like (Additional file 1: Fig. S6b-g). These splicing events could turn into promising new therapeutic strategies aiming at specific key regulatory genes instead of a pleiotropic splicing regulator that could have unsuspected secondary effects.

In summary, by taking advantage of extensive large-scale transcriptomics data from breast cancer cell lines and patients, we identified the first splicing signature capable of subclassifying basal-like tumours based on their aggressiveness and drug resistance. Importantly, novel splicing biomarkers of poor prognosis were identified that should be further studied in more functional assays to test their capacity to inhibit tumour invasion and metastasis. Results from these assays will open new perspectives in the development of improved target therapies and more accurate diagnostic profiles to identify the basal-like triple negative breast cancer patients with a higher chance of relapse.

Discussion

Cancer-specific dysregulation of alternative splicing is a promising source of cancer biomarkers and therapeutic targets to improve diagnostics and thus overall survival rate [63]. An increasing number of mutations at core spliceosome components, such as S3FB1 and U2AF1, or upregulation of specific splicing factors, such as SRSF1 and other members of the SR protein family, which are now considered oncogenes, have been intimately linked to tumour progression and malignancy [64]. Furthermore, an increasing number of alternatively spliced events, like CD44, ENAH, CTNND1 and FLNB, have been shown to impact cell invasion and metastasis on their own, making them promising new targets for more specific therapeutic strategies compared to the inhibition of splicing regulators [22, 23, 65, 66]. Effectively, splicing regulators are not only responsible for the regulation of splicing of a subset of genes, but they are also responsible for other RNA related functions such as translation, mRNA export and nonsense-mediated mRNA decay [56,

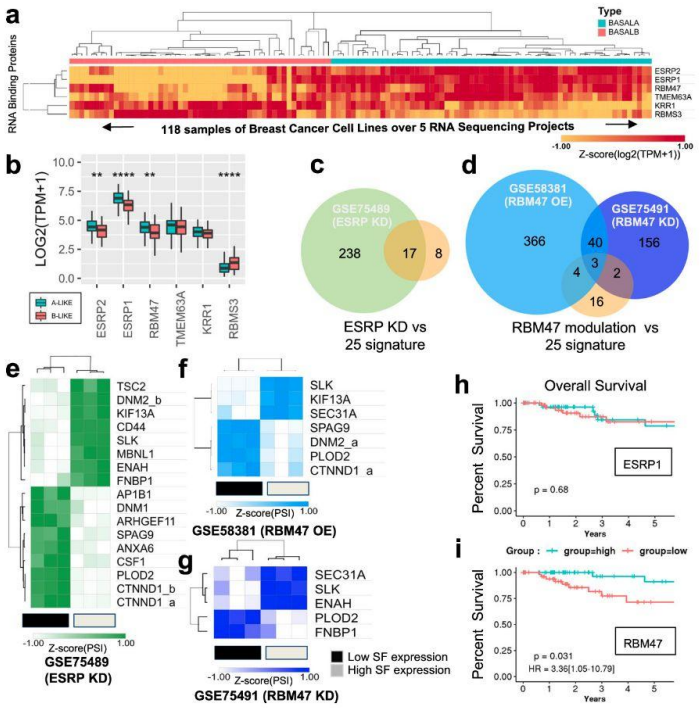


Fig. 6 The basal B-specific splicing signature is co-regulated by ESRP1 and RBM47. **a** Heatmap of transcripts per million values for RNA binding proteins (RBP) differentially expressed in basal A and basal B cell lines (*P* value < 10^{-9} by Kruskal-Wallis test). **b** Box plots of the mean and 25th percentile of the expression levels (in TPM) of the same RBP as in a, but in basal A-like and basal B-like patients. **c**, **d** Venn diagrams of the number of splicing events from the basal B-specific splicing signature dependent on the splicing factors (SF) ESRP1/2 and RBM47 using a cutoff of |DeltaPsi| > 0.1 and a higher probability > = 0.95. **e**–**g** Heatmaps of the PSI values of the ESRP and RBM47-dependent exons from c and d in ESRP1/2 knock downed H358 cells, RBM47 overexpressed MDA-MB-231 cells and RBM47 knock downed H358 cells. **h**, **i** Kaplan-Meier plots for overall survival in basal-like TCGA patients expressing the highest tercile (blue) or the lowest tercile (red) of ESRP1 and RBM47 expression levels. HR (hazard ratio) and logrank *p* values (P) discriminating between groups are shown

64], which can have numerous downstream deleterious effects when inhibited in a targeted therapy. By specifically targeting a key downstream splicing event, as in splicing-specific immunotherapy, a more cancer-specific and direct impact on the cell phenotype might be achieved (134, 135).

Large scale public molecular data sets on genomics (copy number and mutation), epigenomics, transcriptomics, proteomics, in vitro and in vivo cell invasiveness and response to anti-tumour compounds in a large number of patients (11,000 patients across 33 different tumour types from the Genome Cancer Atlas) and human-derived cell lines (1000 cancer cell lines across 36 tumour types from the Broad Institute's Cancer Cell Line Encyclopedia) has become an extraordinary toolbox to identify novel prognostic markers of early metastasis and/or resistance to specific drugs, which are the two major reasons for clinical relapse and low survival rate [67–69]. Unfortunately, the translatability of these preclinical findings is often limited since culture cells are not representative of the variety of individuals nor the biological reality of the tumour's multicellular environment. Yet, culture procedures are improving with the creation of organoids, and machine learning approaches combined with large-scale data mining are bypassing some of these important caveats. This is the case of our cell-to-patient random forest classifier approach, in which the addition at each round of selection of novel informative features, based on the patients classified in previous rounds, allows an algorithm to make use of the information learned from cell lines. Thanks to this approach, we were able to identify the first splicing

signature, composed of 25 alternatively spliced exons, capable of subclassifying basal-like breast cancer patients into two subtypes with different prognoses: basal A- and basal B-like.

Actually, this newly identified basal B-like splicing signature underlined a stem cell-like EMT signature, with hallmarks of cell invasiveness and drug resistance. Five of these 25 alternatively spliced genes are well-known to play a role in cancer (ARHGEF11, CD44, CTNND1, ENAH, MBNL1) [70–72]. Six have been indirectly linked to tumour malignancy and are thus new splicing targets to study (CAST, CSF1, PLOD2, SLK, SPAG9, TSC2) [60, 62, 73–76]. The rest are completely unknown for their splicing role in cancer, even though changes in expression of some of them have been shown to play a role in tumour progression, chemosensitivity and metastasis without specifically addressing which splice variant (ATP5C1, BNIP2, FAT1, FNBP1, SEC31A, ANXA6, DNM1, DNM2) [61, 77]. Of special interest are ARHGEF11 and CTNND1 splice variants. Both proteins are involved in cell-cell adhesion and the basal B-specific splice variants promote cell migration and invasiveness in several cancer types, such as breast cancer (13,54,74, 67). Moreover, depletion of ARHGEF11 in basal breast cancer cells is sufficient to alter cell morphology, which suppresses the cancer cell growth and survival in vitro and in vivo [71]. On the other hand, the existence of an isoform-specific antibody for CTNND1 pro-invasive splice variants turns this splicing candidate as a valuable new target to reduce tumour metastasis [78]. ENAH and CD44 are amongst the most studied splicing events impacting cancer and are well-known biomarkers of poor prognosis. ENAH's inhibition decreases metastasis by slowing down tumour progression and reducing cell invasion and intravasation [79–81]. Whilst the change to basal B splicing signature of CD44, a transmembrane protein that maintains tissue structure, is sufficient to drive an EMT and to increase cell invasion and plasticity by promoting stem cell characteristics [22, 82]. Interestingly, MBNL1 splicing regulation has also been involved in pluripotent stem cell differentiation [83] and cell viability via inhibition of DNA damage response [84]. Promising new splice variants with a potential link with cancer are CSF1, PLOD2, SLK, SPAG9 and TSC2. CSF1 is a macrophage marker which splice variant could correlate with infiltration of tumour-promoting macrophages [73, 85]. Changes in the alternative splicing of the procollagen-lysine PLOD2, which catalyses the deposition and cross-link of collagens in the extracellular matrix, have been intimately linked to EMT progression and cervical, breast, lung, colon and rectal cancer prognosis [40, 86]. Its inhibition reduced proliferation, migration and invasion of cancer cells, while its overexpression promoted cancer stem cell properties and resistance to drugs [62, 87]. SLK was identified as a prognostic biomarker in several cancers and is necessary for the induction of cell migration and invasion during EMT [60, 72, 88]. SPAG9 is a scaffold protein that organises mitogen-activated protein kinases and has been associated with invasion in several types of tumours and prognosis [75, 89, 90]. Finally, TSC2 basal B-specific splicing isoform cannot be phosphorylated by AKT, which leads to a continuously activated mTOR pathway and oncogenic autophagy [74]. More functional studies on the impact of each of these cassette exons splice variants in cancer will increase our knowledge on tumour progression and metastasis with the long term goal of improving diagnostics and treatment. Of note, other types of splicing events, different from the studied cassette exons, have also been shown to play important roles in tumorigenesis, such as alternative splice sites and intron retention [91–93]. It is necessary to extend this type of approaches to all types of splicing events and validate them using independent cohorts of patients. The increase of accessible sequencing data in primary tumours will thus be essential to continue with this type of approaches.

Finally, it is interesting to note that these 25 alternatively spliced exons are basically dependent on three well-known splicing regulators, ESRP1/2 and RBM47, which are intimately linked to EMT and metastasis. ESRP1 is the major regulator of a newly identified epithelial-specific splicing signature [52]. Its expression in cancer cells promotes tumour growth and a mesenchymal-to-epithelial transition which are essential for the formation of new tumours at distal organs during metastasis [94, 95]. RBM47 is a newly identified splicing regulator of EMT that has also been associated with metastasis [56, 96, 97]. Through integrative analysis of clinical breast cancer gene expression datasets, cell line models and mutation data from cancer genome resequencing studies, RBM47 was identified as a suppressor of breast cancer progression and metastasis. It was found mutated in patients with brain metastasis and its expression was necessary to inhibit brain and lung metastatic progression in vivo [56]. Interestingly, despite regulating just 9/25 splicing events of the basal B-specific splicing signature, low expression of RBM47, and not ESRP1, correlated with a poor prognosis and lower survival rate in basal-like breast cancer patients, which increases the interest to design new therapies targeting this splicing regulator.

In fact, this basal B-specific splicing signature has highlighted a subpopulation of basal-like triple negative breast cancer patients differentially expressing several hallmarks of invasive, EMT-like aggressive cancer, such as the newly identified biomarker of metastasis CD36 [20]. CD36 is a fatty receptor expressed in metastasis-

initiating cells. Neutralising antibodies that block CD36 completely inhibited the formation of metastasis in orthotopic mouse models of human oral cancer, and CD36 inhibition impaired metastasis in human melanoma and breast cancer-derived tumours. Interestingly, the fatty acid-binding protein 7 (FABP7) correlates with a higher incidence of brain metastasis and lower survival rate in breast cancer patients, which all together points to a potential connection between fatty acid metabolism and metastasis in our subclass of basal-like breast cancer patients [98]. Furthermore, cells expressing our newly identified basal B-specific splicing signature also showed resistance to several EGFR inhibiting drugs. Therapies targeting EGFR have variable and unpredictable responses in breast cancer [99]. By better subclassifying sensitive from resistant tumour cells, diagnoses could be improved, which will impact the choice of treatment and thus the chances of tumour relapse. Extensive drug screening of cells derived from basal B-like patients combined with machine learning strategies to transfer the splicing knowledge obtained will certainly improve the identification of much more suitable treatments for triple-negative breast cancer cells and reduce tumour relapse, thus improving the survival rate.

Conclusion

Taking advantage of extensive available experimental data in breast cancer cell lines, we performed a knowledge transfer to clinical data to identify the first splicing signature capable of subcategorizing the most aggressive and difficult to treat type of breast cancer, which is basal-like triple negative breast cancer. Based on the pattern of splicing of 25 splicing biomarkers, we could identify two new subclasses of clinically relevant basal-like tumours, basal A and basal B-like, with different sensitivity to drugs and capacity to invade distal organs, which has a direct impact on prognosis. We propose that by testing all basal-like patients with this novel signature, patients with increased chances of creating early metastasis or tumour relapse could be closely monitored to improve their chances of survival. Similarly, by correlating alternative splicing patterns with drug resistance in cancer cell lines, or even cancer cells isolated from patients, more specific splicing biomarkers could be identified for the most adequate and personalised choice of treatment, which is one of the major challenges in triple negative breast cancer. Finally, the newly identified basal B-specific splice variants underline a stem cell-like, highly invasive EMT phenotype, with increased drug resistance, that could be used as novel therapeutic targets to reduce cancer metastasis and relapse, opening new perspectives into the development of improved and more specific treatments for triple negative breast cancer tumours.

Methods

RNA-seq transcriptomics analysis: gene expression and alternative splicing

RNA-seq reads were aligned to the human genome (GRCh38, primary assembly) using STAR [100] version 2.5.2b with standard parameters. Gencode v25 (derivated from Ensembl v85) was used for all analysis requiring annotation.

TPMCalculator [101] (v0.0.1) was used to compute transcripts per million (TPM) values and obtain read counts. Q parameter was set to 255 to keep only unique mapped reads and ExonTPM value was used to consider only reads mapped to exons.

Whippet-quant from Whippet software (v10.4) was used to compute Percentage Spliced-In (PSI) values for splicing analysis. Conjointly to Kruskal-Wallis testing, the output from Whippet-quant was further filtered to include only events for which the sum of inclusion counts (IC) and skipping counts (SC) was greater or equal to 10 for both sets of samples. Whippet-delta was used to compute differential splicing (deltaPsi) and probability that there is some change in splicing between conditions. Two heuristic filters were applied on splicing events as advised in whippet documentation; |deltaPsi| > 0.1 and P(|deltaPsi| > 0.0) > = 95% were considered reliable parameters to filter biologically relevant AS events.

When necessary, Biobambam2 [102] (v 2.0.87) was used to transform bam files into fastq in order to be processed by Whippet.

Gene ontology (GO) analysis was done using the DAVID (v 6.8) [103] functional annotation tool (https://david.ncifcrf.gov/home.jsp) using Benjamini-Hochberg adjusted *P* value cutoff of 0.05 to define a term as enriched. Go terms enrichment was restricted to GOTERM BP-FAT, GOTERM MF-FAT, and GOTERM CC-FAT, KEGG_PATHWAY and REACTOME_PATHWAY.

Gene Set Enrichment Analysis (GSEA v20.0.5) was carried out on the GenePattern [104] web platform using phenotype for permutation type and 1000 for the number of permutations to execute. FDR cutoff of 25% for potential true positive finding was used as documented in the GSEA user guide. Read counts were previously normalised using DESeq2 [105] (v 1.10.1) on the same Platform.

R version 3.6.2 was used all along this study excepted for GSEA.

All heatmaps were done online using Morpheus https://software.broadinstitute.org/morpheus/. Values were adjusted by *Z*-score. (subtract mean and divide by standard deviation). Hierarchical clustering was done in Morpheus. We selected "Metric One minus Pearson correlation" as a measure of distance between pairs of

observation and "Average" as the linkage method. The clusters were done using rows and columns together. Columns were grouped by cancer subtypes.

Sashimi plots to look cassette exons events were done using ggsashimi tool [106].

Machine learning and feature selection

First, we construct a classifier to distinguish basal B/A cell lines using a Random Forest with 1000 trees. After, we applied this model to the TCGA patients. Based on Gini impurity, we computed the class probability to predict patient labelled as B-like or A-like. Then, mixing initial cell lines with a subset of patients classified with the more reliability (the ones picked up with higher class probability not passing below a threshold of $P = 0.6$), we create a new model. Each addition of patients is called a round, during which a new model is created, giving new predictions (probabilities) for the remaining patients. By limiting the number of new patients added at each round (10 x n_current_round) (Fig. 3c and Additional file 1: Figs. S3c-4c), the model can gradually learn from the patient data and avoid overfitting. With such conditions, we can observe a gradual shifting in feature importance from the ones informative to classify cell lines to the ones informative to classify patients and cell lines (Fig. 3d and Additional file 1: Figs. S3d-4d). The algorithm stops when it can no longer incorporate the patients into one or the other group given the cutoff of $P = 0.6$. ML analyse was done with Python 3.7.3 based on scikit-learn version 0.21.2.

To select the more efficient features that were able to separate B-like from A-like patients, we used Boruta package (0.3) implemented in python. We ran it 10 times with different random states, on the 217 features related to splicing and kept the ones that were present at least 7 times on 10. We ended with 25 AS features. Considering only these 25 AS features, we applied TSNE function from manifold package (with perplexity = 20) to 3 other datasets of basal cell lines ($n = 56$) to check the features were sufficient to distinguish spatially these cell lines according to their labels.

For the classification using only differentially expressed genes (Additional file 1: Fig. S3) or a mix of differentially spliced and expressed features (Additional file 1: Fig. S4), we applied the same strategy using the information from the 635 differentially expressed genes and the 217 differentially spliced exons scaling independently the values from the cell lines and patients with sklearn's StandardScaler. We also had to reduce the probability threshold to 0.55 in the mixed model.

Breast cancer annotation

Basal B and A cells were labelled according to literature: Neve et al. [28], Kao et al. [33], Marcotte et al. [107], Dai et al. [108]. PAM50 intrinsic subtype was retrieved from https://www.cell.com/cancer-cell/fulltext/S1535-6108(18)30119-3 [109].

Claudin Low status was defined with script downloaded from https://github.com/clfougner/ClaudinLow/blob/master/Code/TCGA.r [110] using dataset from http://download.cbioportal.org/brca_tcga_pan_can_atlas_2018.tar.gz [111, 112].

Survival analysis

Log-rank tests were performed using the functions surv and survfit from R package (survival v3.1.8). A different survival was considered significative if log rank test p value was < 0.05. Coxph function was also used for univariate Cox regression analysis in order to compute Hazard Ratio and 95% Interval of confidence. Kaplan-Meier curve was plotted using function ggsurvplot from R package survminer (0.4.6). Plots were truncated at 5 years, but the analyses were conducted using all of the data. All endpoints used for survival analysis in this study were retrieved from this study [113].

Statistics

Wilcoxon rank-sum test was used to assess statistical significance within boxplots.

They were noted. $P < 0.05$ (*), $P < 0.01$ (**), and $P < 0.001$ (***), $P < 0.0001$ (****).

Kruskal-Wallis test was used to keep differential features for expression (TPM values) or splicing (PSI values) when Luminal, basal A and B cell lines were compared and displayed in heatmap figures. A threshold of p value $< 10–5$ was used to filter out potential false positive and reduce the number of features in order to apply hierarchical clustering. This threshold was adapted depending on the number of samples in the comparison. For RNA binding proteins, a higher cut off of $p < 10–9$ was used because 5 projects were pulled together.

Abbreviations

AS: Alternative splicing; CE: Cassette exons; EMT: Epithelial-to-mesenchymal transition; CSC: Cancer stem cells; CTC: Circulating tumour cells; PSI: Percentage spliced-in; TPM: Transcripts per million; DSS: Disease-specific survival; TCGA: The Cancer Genome Atlas; RBPs: RNA binding proteins

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-021-01002-7.

Additional file 1: Fig. S1. Allele-specific alternative splicing and its functional genetic variants in human tissues. **Fig. S2.** Hierarchical clustering and k-means of patients based on differential gene expression and splicing. **Fig. S3.** Semi-supervised Random Forest Classifier to transfer cell lines knowledge to patients using expression levels. **Fig. S4.** Semi-supervised Random Forest Classifier to transfer cell lines knowledge to patients using alternative splicing and expression levels. **Fig. S5.** In silico validation of basal B splicing signature. **Fig. S6.** Prognostic value of individual alternatively spliced genes from the basal B-specific signature.

Additional file 2: Table S1. GEO accession numbers for all the datasets analysed.

Additional file 3: Table S2. Name, coordinates (Hg38) and PSI mean value and standard error for the 25 exons of the basal B-specific signature in Basal A and Basal B cancer cells and patients. The difference in splicing levels between basal B and basal A is shown as deltaPSI.

Acknowledgements
We would like to thank Yaïza Nuñez-Alvarez and Sylvain Barrière for discussions.

Code
Code and annotation files are available here.
https://github.com/LucoLab/Villemin_2020.

Authors' contributions
JPV performed all the analyses. CL helped with the development of the semi-supervised classifier. MSC and AO helped with the discussion and writing of the manuscript. JPV, RL and WR designed the study and wrote the manuscript. All authors read and approved the final manuscript.

Funding
Luco team is supported by the Agence Nationale de la Recherche [ANRJCJC - 2016 - EpiSplicing] and the Labex EpiGenMed [ANR-10-LABX-12-01]. Ritchie team is supported by the Agence Nationale de la Recherche [ANRJCJC - WIRED], the Labex EpiGenMed [ANR-10-LABX-12-01] and the MUSE initiative [GECKO].

Availability of data and materials
All datasets are available in the Gene Expression Omnibus (GEO): GSE75489, GSE58381, GSE75491, GSE61220, PRJEB25042, GSE74881, GSE75492, PRJNA523380, PRJNA297219, PRJNA210428, PRJNA251383, PRJEB30617 (detailed in Additional file 2: Table S1) and The Cancer Genome Atlas (TCGA) repositories upon request (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v11.p8).

Declarations

Ethics approval and consent to participate
Patients data was obtained from The Cancer Genome Atlas upon agreement of TCGA ethics and policies (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/policies)

Consent for publication
All patients gave consent for publication of their personal information.

Competing interests
The authors declare no competing interests.

Received: 13 November 2020 Accepted: 9 March 2021
Published online: 12 April 2021

References

1. Sims AH, Howell A, Howell SJ, Clarke RB. Origins of breast cancer subtypes and therapeutic implications. Nat Clin Pract Oncol. 2007;4(9):516–25.
2. Toft DJ, Cryns VL. Minireview: basal-lik e breast cancer: from molecular profiles to targeted therapies. Mol Endocrinol. 2011;25(2):199–211. https://doi.org/10.1210/me.2010-0164.
3. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lonning PE, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001;98(19):10869–74. https://doi.org/10.1073/pnas.191367098.
4. Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res. 2015;5(10):2929–43.
5. Cardoso F, Van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. N Engl J Med. 2016;375(8):717–29. https://doi.org/10.1056/NEJMoa1602253.
6. Jiang Y-Z, Ma D, Suo C, Shi J, Xue M, Hu X, et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. Cancer Cell. 2019;35(3):428–40.e5.
7. Marcotte R, Sayad A, Brown KR, Sanchez-Garcia F, Reimand J, Haider M, et al. Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. Cell. 2016;164(1-2):293–309.
8. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92–4. https://doi.org/10.1038/nature24284.
9. Milne RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, Lindström S, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet. 2017;49(12):1767–78.
10. Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, Brook MN, et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat Genet. 2013;45(4):392–8, 398e1–2.
11. Karni R, De Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat Struct Mol Biol. 2007;14(3):185–93. https://doi.org/10.1038/nsmb1209.
12. Climente-González H, Porta-Pardo E, Godzik A, Eyras E. The functional impact of alternative splicing in cancer. Cell Rep. 2017;20(9):2215–26. https://doi.org/10.1016/j.celrep.2017.08.012.
13. Sebestyén E, Zawisza M, Eyras E. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. Nucleic Acids Res. 2015;43(3):1345–56. https://doi.org/10.1093/nar/gku1392.
14. Kahles A, Lehmann K-V, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, et al. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. Cancer Cell. 2018;34(2):211–224.e6
15. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev. 2010;24(21):2343–64. https://doi.org/10.1101/gad.1973010.
16. Bechara EG, Sebestyén E, Bernardis I, Eyras E, Valcárcel J. RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. Mol Cell. 2013;52(5):720–33. https://doi.org/10.1016/j.molcel.2013.11.010.
17. Moore MJ, Wang Q, Kennedy CJ, Silver PA. An alternative splicing network links cell-cycle control to apoptosis. Cell. 2010;142(4):625–36. https://doi.org/10.1016/j.cell.2010.07.019.
18. Amin EM, Oltean S, Hua J, Gammons MVR, Hamdollah-Zadeh M, Welsh GI, Cheung MK, Ni L, Kase S, Rennel ES, Symonds KE, Nowak DG, Royer-Pokora B, Saleem MA, Hagiwara M, Schumacher VA, Harper SJ, Hinton DR, Bates DO, Ladomery MR. WT1 mutants reveal SRPK1 to be a downstream angiogenesis target by altering VEGF splicing. Cancer Cell. 2011;20(6):768–80. https://doi.org/10.1016/j.ccr.2011.10.016.
19. Chen M, Zhang J, Manley JL. Turning on a fuel switch of cancer: hnRNP proteins regulate alternative splicing of pyruvate kinase mRNA. Cancer Res. 2010;70(22):8977–80. https://doi.org/10.1158/0008-5472.CAN-10-2513.
20. Pascual G, Avgustinova A, Mejetta S, Martín M, Castellanos A, Attolini CSO, Berenguer A, Prats N, Toll A, Hueto JA, Bescós C, di Croce L, Benitah SA. Targeting metastasis-initiating cells through the fatty acid receptor CD36. Nature. 2017;541(7635):41–5. https://doi.org/10.1038/nature20791.
21. Xu Y, Gao XD, Lee JH, Huang H, Tan H, Ahn J, Reinke LM, Peter ME, Feng Y, Gius D, Siziopikou KP, Peng J, Xiao X, Cheng C. Cell type-restricted activity of hnRNPM promotes breast cancer metastasis via regulating alternative splicing. Genes Dev. 2014;28(11):1191–203. https://doi.org/10.1101/gad.241968.114.
22. Brown RL, Reinke LM, Damerow MS, Perez D, Chodosh LA, Yang J, Cheng C. CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. J Clin Investig. 2011;121(3):1064–74. https://doi.org/10.1172/JCI44540.
23. Li J, Choi PS, Chaffer CL, Labella K, Hwang JH, Giacomelli AO, et al. An alternative splicing switch in FLNB promotes the mesenchymal cell state in human breast cancer. eLife. 2018;7:1–28.
24. Ranieri D, Rosato B, Nanni M, Magenta A, Belleudi F, Torrisi MR. Expression of the FGFR2 mesenchymal splicing variant in epithelial cells drives epithelial-mesenchymal transition. Oncotarget. 2016;7(5):5440–60. https://doi.org/10.18632/oncotarget.6706.
25. Lee SCW, Abdel-Wahab O. Therapeutic targeting of splicing in cancer. Nat Med. 2016;22(9):976–86. https://doi.org/10.1038/nm.4165.
26. Bonomi S, Gallo S, Catillo M, Pignataro D, Biamonti G, Ghigna C. Oncogenic alternative splicing switches: role in cancer progression and prospects for therapy. Int J Cell Biol. 2013;2013:1–17. https://doi.org/10.1155/2013/962038.

27. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483(7391):603–7. https://doi.org/10.1038/nature11003.
28. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe JP, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo WL, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A, Gray JW. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell. 2006;10(6):515–27. https://doi.org/10.1016/j.ccr.2006.10.008.
29. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, Zhou AY, Brooks M, Reinhard F, Zhang CC, Shipitsin M, Campbell LL, Polyak K, Brisken C, Yang J, Weinberg RA. The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell. 2008;133(4):704–15. https://doi.org/10.1016/j.cell.2008.03.027.
30. Hennessy BT, Gonzalez-Angulo A-M, Stemke-Hale K, Gilcrease MZ, Krishnamurthy S, Lee J-S, Fridlyand J, Sahin A, Agarwal R, Joy C, Liu W, Stivers D, Baggerly K, Carey M, Lluch A, Monteagudo C, He X, Weigman V, Fan C, Palazzo J, Hortobagyi GN, Nolden LK, Wang NJ, Valero V, Gray JW, Perou CM, Mills GB. Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. Cancer Res. 2009;69(10): 4116–24. https://doi.org/10.1158/0008-5472.CAN-08-3441.
31. Thiery JP, Acloque H, Huang RYJ, Nieto MA. Epithelial-mesenchymal transitions in development and disease. Cell. 2009;139(5):871–90. https://doi.org/10.1016/j.cell.2009.11.007.
32. Ye X, Tam WL, Shibue T, Kaygusuz Y, Reinhardt F. Distinct EMT programs control normal mammary stem cells and tumour-initiating cells. Natue. 2016;525(7568):256–60. https://doi.org/10.1038/nature14897.
33. Kao J, Salari K, Bocanegra M, La Choi Y, Girard L, Gandhi J, et al. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. Plos One. 2009;4(7):e6146. https://doi.org/10.1371/journal.pone.0006146.
34. Charafe-Jauffret E, Ginestier C, Monville F, Finetti P, Adélaïde J, Cervera N, Fekairi S, Xerri L, Jacquemier J, Birnbaum D, Bertucci F. Gene expression profiling of breast cell lines identifies potential new basal markers. Oncogene. 2006;25(15):2273–84. https://doi.org/10.1038/sj.onc.1209254.
35. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70. https://doi.org/10.1038/nature11412.
36. Yae T, Tsuchihashi K, Ishimoto T, Motohara T, Yoshikawa M, Yoshida GJ, et al. Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell. Nat Commun. 2012;3:883. https://doi.org/10.1038/ncomms1892.
37. De Faria Poloni J, Bonatto D. Influence of transcriptional variants on metastasis. RNA Biol. 2018;15(8):1006–1024. https://doi.org/10.1080/15476286.2018.1493328.
38. Qiu Y, Lyu J, Dunlap M, Harvey SE, Cheng C. A combinatorially regulated RNA splicing signature predicts breast cancer EMT states and patient survival. RNA. 2020;26(9):1257–67. https://doi.org/10.1261/rna.074187.119.
39. Sebestyén E, Singh B, Miñana B, Pagès A, Mateo F, Pujana MA, et al. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res. 2016;26:732–44.
40. Shapiro IM, Cheng AW, Flytzanis NC, Balsamo M, Condeelis JS, Oktay MH, Burge CB, Gertler FB. An EMT–driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. PLoS Genet. 2011; 7(8):e1002218. https://doi.org/10.1371/journal.pgen.1002218.
41. Warzecha CC, Jiang P, Amirikian K, Dittmar KA, Lu H, Shen S, Guo W, Xing Y, Carstens RP. An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. EMBO J. 2010;29(19):3286–300. https://doi.org/10.1038/emboj.2010.195.
42. Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD, Frey BJ, Blencowe BJ. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell. 2004;16(6):929–41. https://doi.org/10.1016/j.molcel.2004.12.004.
43. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102.
44. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4(5):P3.
45. Dragowska WH, Weppler SA, Qadir MA, Wong LY, Franssen Y, Baker JHE, Kapanen AI, Kierkels GJJ, Masin D, Minchinton AI, Gelmon KA, Bally MB. The combination of gefitinib and RAD001 inhibits growth of HER2 overexpressing breast cancer cells and tumors irrespective of trastuzumab sensitivity. BMC Cancer. 2011;11(1). https://doi.org/10.1186/1471-2407-11-420.
46. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41:D955–61. https://doi.org/10.1093/nar/gks1111.
47. Ho-Yen CM, Jones JL, Kermorgant S. The clinical and functional significance of c-Met in breast cancer: a review. Breast Cancer Res. 2015;17(1):52. https://doi.org/10.1186/s13058-015-0547-6.
48. Kursa MB, Rudnicki WR. Feature selection with the boruta package. J Stat Softw. 2010;36:1–13.
49. Tian B, Li X, Kalita M, Widen SG, Yang J, Bhavnani SK, et al. Analysis of the TGFβ-induced program in primary airway epithelial cells shows essential role of NF-KB/RelA signaling network in type II epithelial mesenchymal transition. BMC Genomics. 2015;16(1):529. https://doi.org/10.1186/s12864-015-1707-x.
50. Pillman KA, Phillips CA, Roslan S, Toubia J, Dredge BK, Bert AG, et al. miR-200/375 control epithelial plasticity-associated alternative splicing by repressing the RNA -binding protein Quaking. EMBO J. EMBO J. 2018;37(13): e99016. https://doi.org/10.15252/embj.201899016.
51. Pattabiraman DR, Bierie B, Kober KI, Thiru P, Krall JA, Zill C, et al. Activation of PKA leads to mesenchymal-to-epithelial transition and loss of tumor-initiating ability. Science. 2016;351(6277):aad3680. https://doi.org/10.1126/science.aad3680.
52. Yang Y, Park JW, Bebee TW, Warzecha CC, Guo Y, Shang X, Xing Y, Carstens RP. Determination of a comprehensive alternative splicing regulatory network and combinatorial regulation by key factors during the epithelial-to-mesenchymal transition. Mol Cell Biol. 2016;36(11):1704–19. https://doi.org/10.1128/MCB.00019-16.
53. Lehmann BD, Shyr Y, Pietenpol JA, Lehmann BD, Bauer JA, Chen X, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011; 121(7):2750–67. https://doi.org/10.1172/JCI45014.
54. Caramel J, Ligier M, Puisieux A. Pleiotropic Roles for ZEB1 in Cancer. Cancer Res. 2018;78(1):30–5.
55. Bebee TW, Park JW, Sheridan KI, Warzecha CC, Cieply BW, Rohacek AM, et al. The splicing regulators Esrp1 and Esrp2 direct an epithelial splicing program essential for mammalian development. eLife. 2015;4:1–27.
56. Vanharanta S, Marney CB, Shu W, Valiente M, Zou Y, Mele A, et al. Loss of the multifunctional RNA-binding protein RBM47 as a source of selectable metastatic traits in breast cancer. eLife. 2014;2014:1–24.
57. Park SH, Brugiolo M, Akerman M, Das S, Urbanski L, Geier A, et al. Differential functions of splicing factors in mammary transformation and breast cancer metastasis. Cell Rep. 2019;29:2672–2688.e7.
58. \Hendley AM, Wang YJ, Polireddy K, Alsina J, Ahmed I, Lafaro KJ, Zhang H, Roy N, Savidge SG, Cao Y, Hebrok M, Maitra A, Reynolds AB, Goggins M, Younes M, Iacobuzio-Donahue CA, Leach SD, Bailey JM. p120 catenin suppresses basal epithelial cell extrusion in invasive pancreatic neoplasia. Cancer Res. 2016; 76(11):3351–63. https://doi.org/10.1158/0008-5472.CAN-15-2268.
59. Braeutigam C, Rago L, Rolke A, Waldmeier L, Christofori G, Winter J. The RNA-binding protein Rbfox2: an essential regulator of EMT-driven alternative splicing and a mediator of cellular invasion. Oncogene. 2014;33(9):1082–92. https://doi.org/10.1038/onc.2013.50.
60. Roovers K, Wagner S, Storbeck CJ, O'Reilly P, Lo V, Northey JJ, et al. The Ste20-like kinase SLK is required for ErbB2-driven breast cancer cell motility. Oncogene. 2009;28(31):2839–48. https://doi.org/10.1038/onc.2009.146.
61. Meng J. Distinct functions of dynamin isoforms in tumorigenesis and their potential as therapeutic targets in cancer. Oncotarget. 2017;8(25):41701–16. https://doi.org/10.18632/oncotarget.16678.
62. Song Y, Zheng S, Wang J, Long H, Fang L, Wang G, et al. Hypoxia-induced PLOD2 promotes proliferation, migration and invasion via PI3K/Akt signaling

in glioma. Oncotarget. 2017;8(26):41947–62. https://doi.org/10.18632/oncotarget.16710.
63. Urbanski LM, Leclair N, Anczuków O. Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. Wiley Interdisciplinary Reviews: RNA. 2018;9(4):e1476. https://doi.org/10.1002/wrna.1476.
64. Anczukow O, Krainer AR. Splicing-factor alterations in cancers. Rna. 2016; 22(9):1285–301. https://doi.org/10.1261/rna.057919.116.
65. Pagliarini V, Naro C, Sette C. Splicing regulation: a molecular device to enhance cancer cell adaptation. Biomed Res Int. 2015;2015:1–13. https://doi.org/10.1155/2015/543067.
66. Di Modugno F, Iapicca P, Boudreau A, Mottolese M, Terrenato I, Perracchio L, et al. Splicing program of human MENA produces a previously undescribed isoform associated with invasive, mesenchymal-like breast tumors. Proc Natl Acad Sci U S A. 2012;109(47):19280-5. https://doi.org/10.1073/pnas.1214394109.
67. Weinstein JN. Cell lines battle cancer. Nature. 2012;483(7391):544–5. https://doi.org/10.1038/483544a.
68. Jiang G, Zhang S, Yazdanparast A, Li M, Pawar AV, Liu Y, et al. Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics. 2016;17 Suppl 7(Suppl 7):525. https://doi.org/10.1186/s12864-016-2911-z.
69. Yu K, Chen B, Aran D, Charalel J, Yau C, Wolf DM, van 't Veer LJ, Butte AJ, Goldstein T, Sirota M. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. Nat Commun. 2019;10(1): 3574. https://doi.org/10.1038/s41467-019-11415-2.
70. Warzecha CC, Carstens RP. Complex changes in alternative pre-mRNA splicing play a central role in the epithelial-to-mesenchymal transition (EMT). Semin Cancer Biol. 2012;22(5-6):417–27. https://doi.org/10.1016/j.semcancer.2012.04.003.
71. Itoh M, Radisky DC, Hashiguchi M, Sugimoto H. The exon 38-containing ARHGEF11 splice isoform is differentially expressed and is required for migration and growth in invasive breast cancer cells. Oncotarget. 2017;8(54): 92157–70. https://doi.org/10.18632/oncotarget.20985.
72. Zhao N, Guo M, Wang K, Zhang C, Liu X. Identification of pan-cancer prognostic biomarkers through integration of multi-omics data. Front Bioeng Biotechnol. 2020;8:268. https://doi.org/10.3389/fbioe.2020.00268.
73. Wang H, Shao Q, Sun J, Ma C, Gao W, Wang Q, Zhao L, Qu X. Interactions between colon cancer cells and tumor-infiltrated macrophages depending on cancer cell-derived colony stimulating factor 1. Oncolmmunology. 2016; 5(4):e1122157. https://doi.org/10.1080/2162402X.2015.1122157.
74. Chen Y, Lu Y, Ren Y, Yuan J, Zhang N, Kimball H, et al. Starvation-induced suppression of DAZAP1 by miR-10b integrates splicing control into TSC2-regulated oncogenic autophagy in esophageal squamous cell carcinoma. Theranostics. 2020;10(11):4983–96. https://doi.org/10.7150/thno.43046.
75. Yan Q, Lou G, Qian Y, Qin B, Xu X, Wang Y, et al. SPAG9 is involved in hepatocarcinoma cell migration and invasion via modulation of ELK1 expression. OncoTargets Ther. 2016;9:1067–75. https://doi.org/10.2147/OTT.S98727.
76. Chen X, Zhao C, Guo B, Zhao Z, Wang H, Fang Z. Systematic profiling of alternative mRNA splicing signature for predicting glioblastoma prognosis. Front Oncol. 2019;9. https://doi.org/10.3389/fonc.2019.00928.
77. Zhang L, Liu X, Zhang X, Chen R. Identification of important long non-coding RNAs and highly recurrent aberrant alternative splicing events in hepatocellular carcinoma through integrative analysis of multiple RNA-Seq datasets. Mol Genet Genomics. 2016;291(3):1035–51. https://doi.org/10.1007/s00438-015-1163-y.
78. Venhuizen JH, Sommer S, Span PN, Friedl P, Zegers MM. Differential expression of p120-catenin 1 and 3 isoforms in epithelial tissues. Sci Rep. 2019;9(1):90. https://doi.org/10.1038/s41598-018-36889-w.
79. Roussos ET, Wang Y, Wyckoff JB, Sellers RS, Wang W, Li J, et al. Mena deficiency delays tumor progression and decreases metastasis in polyoma middle-T transgenic mouse mammary tumors. Breast Cancer Res. 2010;12(6): R101. https://doi.org/10.1186/bcr2784.
80. Philippar U, Roussos ET, Oser M, Yamaguchi H, Kim H Do, Giampieri S, et al. A mena invasion isoform potentiates EGF-induced carcinoma cell invasion and metastasis. Dev Cell. 2008;15(6):813–28. https://doi.org/10.1016/j.devcel.2008.09.003.
81. Li Q, Su YL, Zeng M, Shen WX. Enabled homolog shown to be a potential biomarker and prognostic indicator for breast cancer by bioinformatics analysis. Clin Invest Med. 2018;41(4):E186–E195. https://doi.org/10.25011/cim.v41i4.32221.
82. Zhang H, Brown RL, Wei Y, Zhao P, Liu S, Liu X, Deng Y, Hu X, Zhang J, Gao XD, Kang Y, Mercurio AM, Goel HL, Cheng C. CD44 splice isoform switching determines breast cancer stem cell state. Genes Dev. 2019;33(3-4):166–79. https://doi.org/10.1101/gad.319889.118.
83. Venables JP, Lapasset L, Gadea G, Fort P, Klinck R, Irimia M, et al. MBNL1 and RBFOX2 cooperate to establish a splicing programme involved in pluripotent stem cell differentiation. Nat Commun. 2013;4:2480. https://doi.org/10.1038/ncomms3480.
84. Tabaglio T, Low DHP, Teo WKL, Goy PA, Cywoniuk P, Wollmann H, Ho J, Tan D, Aw J, Pavesi A, Sobczak K, Wee DKB, Guccione E. MBNL1 alternative splicing isoforms play opposing roles in cancer. Life Sci Alliance. 2018;1(5): e201800157. https://doi.org/10.26508/lsa.201800157.
85. Soncin I, Sheng J, Chen Q, Foo S, Duan K, Lum J, et al. The tumour microenvironment creates a niche for the self-renewal of tumour-promoting macrophages in colon adenoma. Nat Commun. 2018;9(1):582. https://doi.org/10.1038/s41467-018-02834-8.
86. Markus MA, Yang YHJ, Morris BJ. Transcriptome-wide targets of alternative splicing by RBM4 and possible role in cancer. Genomics. 2016;107(4):138–44. https://doi.org/10.1016/j.ygeno.2016.02.003.
87. Sheng X, Li Y, Li Y, Liu W, Lu Z, Zhan J, Xu M, Chen L, Luo X, Cai G, Zhang S. PLOD2 contributes to drug resistance in laryngeal cancer by promoting cancer stem cell-like characteristics. BMC Cancer. 2019;19(1):840. https://doi.org/10.1186/s12885-019-6029-y.
88. Conway J, Al-Zahrani KN, Pryce BR, Abou-Hamad J, Sabourin LA. Transforming growth factor β-induced epithelial to mesenchymal transition requires the Ste20-like kinase SLK independently of its catalytic activity. Oncotarget. 2017;8(58):98745–56. https://doi.org/10.18632/oncotarget.21928.
89. de Miguel FJ, Pajares MJ, Martínez-Terroba E, Ajona D, Morales X, Sharma RD, et al. A large-scale analysis of alternative splicing reveals a key role of QKI in lung cancer. Mol Oncol. 2016;10(9):1437–49. https://doi.org/10.1016/j.molonc.2016.08.001.
90. Yang X, Zhou W, Liu S. SPAG9 controls the cell motility, invasion and angiogenesis of human osteosarcoma cells. Exp Ther Med. 2016;11(2):637–44. https://doi.org/10.3892/etm.2015.2932.
91. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. Genome Med. 2015;7(1):45. https://doi.org/10.1186/s13073-015-0168-9.
92. Jung H, Lee D, Lee J, Park D, Kim YJ, Park WY, et al. Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat Genet. 2015; 47(11):1242–8. https://doi.org/10.1038/ng.3414.
93. Chen J, Weiss WA. Alternative splicing in cancer: implications for biology and therapy. Oncogene. 2015;34(1):1–14. https://doi.org/10.1038/onc.2013.570.
94. Jeong HM, Han J, Lee SH, Park HJ, Lee HJ, Choi JS, et al. ESRP1 is overexpressed in ovarian cancer and promotes switching from mesenchymal to epithelial phenotype in ovarian cancer cells. Oncogenesis. 2017;6(10):e389. https://doi.org/10.1038/oncsis.2017.87.
95. Hayakawa A, Saitoh M, Miyazawa K. Dual roles for epithelial splicing regulatory proteins 1 (ESRP1) and 2 (ESRP2) in cancer progression. In: Advances in Experimental Medicine and Biology. 2017;925:33–40. https://doi.org/10.1007/5584_2016_50.
96. Sakurai T, Isogaya K, Sakai S, Morikawa M, Morishita Y, Ehata S, Miyazono K, Koinuma D. RNA-binding motif protein 47 inhibits Nrf2 activity to suppress tumor growth in lung adenocarcinoma. Oncogene. 2017;36(35):5083. https://doi.org/10.1038/onc.2017.191.
97. Rokavec M, Kaller M, Horst D, Hermeking H. Pan-cancer EMT-signature identifies RBM47 down-regulation during colorectal cancer progression. Sci Rep. 2017;7(1):4687. https://doi.org/10.1038/s41598-017-04234-2.
98. Cordero A, Kanojia D, Miska J, Panek WK, Xiao A, Han Y, Bonamici N, Zhou W, Xiao T, Wu M, Ahmed AU, Lesniak MS. FABP7 is a key metabolic regulator in HER2+ breast cancer brain metastasis. Oncogene. 2019;38(37): 6445–60. https://doi.org/10.1038/s41388-019-0893-4.
99. Savage P, Blanchet-Cohen A, Revil T, Badescu D, Saleh SMI, Wang YC, Zuo D, Liu L, Bertos NR, Munoz-Ramos V, Basik M, Petrecca K, Asselah J, Meterissian S, Guiot MC, Omeroglu A, Kleinman CL, Park M, Ragoussis J. A targetable EGFR-dependent tumor-initiating program in breast cancer. Cell Rep. 2017;21(5):1140–9. https://doi.org/10.1016/j.celrep.2017.10.015.
100. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
101. Alvarez RV, Pongor LS, Mariño-Ramírez L, Landsman D. TPMCalculator: One-step software to quantify mRNA abundance of genomic features. Bioinformatics. 2019; 35(11):1960–2. https://doi.org/10.1093/bioinformatics/bty896.

102. Tischler, G., Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol Med. 2014;9:13. https://doi.org/10.1186/1751-0473-9-13.
103. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.https://doi.org/10.1038/nprot.2008.211.
104. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P. GenePattern 2.0. Nat Genet. 2006;38(5):500–1. https://doi.org/10.1038/ng0506-500.
105. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
106. Garrido-Martín D, Palumbo E, Guigó R, Breschi A. ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. Plos Comput Biol. 2018;14(8):e1006360. https://doi.org/10.1371/journal.pcbi.1006360.
107. Mills GB, Sanchez-Garcia F, Virtanen C, Marcotte R, Pe'er D, Brown KR, et al. Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. Cell. 2016;164:293–309.
108. Dai X, Cheng H, Bai Z, Li J. Breast cancer cell line classification and its relevance with breast tumor subtyping. J Cancer. 2017;8(16):3131–41. https://doi.org/10.7150/jca.18457.
109. Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. Cancer Cell. 2018;33:690–705.e9.
110. Fougner C, Bergholtz H, Norum JH, Sørlie T. Re-definition of claudin-low as a breast cancer phenotype. Nat Commun. 2020;11:756411.
111. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4. https://doi.org/10.1158/2159-8290.CD-12-0095.
112. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1. https://doi.org/10.1126/scisignal.2004088.
113. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell. 2018;173:400–416.e11.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

During the last three years I helped Benjamin Viart in the implementation of PickPocket, focusing especially on the implementation in Python of the feature extraction process and the ML approaches.
Though promising, the method performances are not yet satisfying enough and the project needs more time and effort to be concluded.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Benjamin Viart[1*], Claudio Lorenzi[2], María Moriel-Carretero[3], Sofia Kossida[1]

1 IMGT®, IGH, CNRS - Univ Montpellier, Montpellier, France

2 Machine learning and gene regulation, IGH,CNRS - Univ Montpellier, Montpellier, France.

3 Cytoplasmic Control of Genome Stability, CRBM, CNRS - Univ Montpellier, Montpellier, France

* corresponding author

E-mail: benjamin.viart@igh.cnrs.fr

Pocket binding prediction for ligand families

Abstract

Most of the protein biological functions occur through contacts with other proteins or ligands. The residues that constitute the contact surface of a ligand-binding pocket are usually located far away within its sequence. Therefore, the identification of such motifs is more challenging than the linear protein domains. To discover new binding sites, we developed a tool called PickPocket that focuses on a small set of user-defined ligands and

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

uses neural networks to train a ligand-binding prediction model. We tested PickPocket on fatty acid-like ligands due to their structural similarities and their under-representation in the ligand-pocket binding literature.

Our results show that for fatty acid-like molecules, pocket descriptors and secondary structures are enough to obtain predictions with accuracy >90% using a dataset of 1740 manually curated ligand-binding pockets. The trained model could also successfully predict the ligand-binding pockets using unseen structural data of two recently reported fatty acid-binding proteins. We think that the PickPocket tool can help to discover new protein functions by investigating the binding sites of specific ligand families. The source code and all datasets contained in this work are freely available at https://github.com/benjaminviart/PickPocket .

Author Summary :

Most of the protein biological functions are defined by its interactions with other proteins or ligands. The cavity of the protein structure that receives a ligand, also called a pocket, is made of residues that are usually located far away within its sequence. Therefore understanding the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

complementarity of pocket and ligand is a real challenge. To discover new binding sites, we developed a tool called PickPocket that focuses on a small set of user-defined ligands to train a prediction model. Our results show that for fatty acid-like molecules, pocket descriptors (such as volume, shape, hydrophobicity…) and secondary structures are enough to obtain predictions with accuracy >90% using a dataset of 1740 manually curated ligand-binding pockets. The trained model could also successfully predict the ligand-binding pockets using unseen structural data of two recently reported fatty acid-binding proteins. We think that the PickPocket tool can help to discover new protein functions by investigating the binding sites of specific ligand families.

Introduction :

One of the main tasks of bioinformatics is to associate biological roles to proteins using the always increasing biological data (1,2). To predict the function of a protein based on its sequence, computational methods look for sequence patterns in biological databases of known and already annotated proteins. Homology search (3,4), motif search (5) and functional

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

domain search (6,7) are the most common methods, among the many available tools. Other strategies exploit different data types, such as gene expression (8) or even a combinatorial approach (9,10).

Most of the protein biological functions occur through interactions with other proteins or ligands (11). The residues making the protein contact surface and cavity shape are often located far away within the protein sequence. Therefore, the identification of such motifs is more difficult. Fortunately, the quantity of protein structures and models available in the Protein Data Bank (PDB) archive (12) has increased rapidly (13), providing abundant data for structural analyses. The study of ligands and cavities or pockets to which they bind is of particular interest, especially for drug discovery.

Different algorithms exist to compute the structure of ligand-binding pockets and to predict binding sites using geometric criteria, such as SURFNET (14), APROPOS (15), Q-SiteFinder (16), LIGSITEcsc (17), ConCavity (17,18), fpocket (19),DEPTH (20) or PDBinder (21) among others. Artificial intelligence also has been useful in this field with the development of algorithms for convolutional neural networks, such as

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FRSite (22), DeepSite (23) and DeepDrug3D (24) or of tools based on the random forest algorithm, such as P2Rank (25).

With the progressive increase in structure availability, the need to store and compare ligand-binding pocket data has led to the creation of dedicated databases, such as the Computed Atlas of Surface Topography of proteins (CASTp) (26,27) and the PoSSuM database (28). The collection of all pockets present in a single organism is called a pocketome and has its dedicated database: Pocketome (29). The extensive knowledge of all the pocket structures and their comparison are valuable for drug designers. For special needs, the tool PocketPipe can be used to analyze the pocketome of a single organism (30). Recently, comparative analyses of binding sites have gained momentum due to their capacity to reveal ligand-binding similarities among proteins, regardless of their evolution (31,32). As different proteins can evolve to bind to the same ligand type (33) the accurate classification of binding sites has become an important tool for designing drugs and predicting their possible side effects through unwanted binding (34,35).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

One limitation of the available tools is that they are tailored for drug design or for the analysis of large pockets, thus excluding the possibility of executing other tasks, such as determining the specific ligand-pocket binding complementarity. Yet, to discover new binding sites, the reverse approach needs to be possible: to focus on a small set of specific ligands and to take into account their molecular and structural specificity.

In this work, we developed a tool called PickPocket to generate a dataset of pockets that interact with a specific user-defined ligand family. The workflow consists of different R (36) and python (36,37) programs organized using bash scripts. From the pocket dataset, we computed a descriptive matrix of all pockets that contains structural information on the cavity and the residue secondary structure. Then, we trained a neural network multilayer perceptron (38) to predict whether a cavity is a 'true' pocket (i.e. can interact with ligands of the family under study) or a 'false' pocket.

To test PickPocket, we decided to use fatty acid-like ligands due to their structural similarities and their under-representation in the ligand-pocket binding literature (39). The source code and all datasets contained

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in this work are freely available at

https://github.com/benjaminviart/PickPocket.

Results :

The input data for the PickPocket tool consisted of 42 fatty acid-like ligands (Supplementary Information Table 1 : Detail of ligands used as input for PickPocket), 301 structures containing one of the selected ligands and 242 structures containing no ligand. For each pocket, a descriptive matrix is computed that gathers 21 features including structural information of the cavity as well as secondary structure (Table 1). In order to ensure the quality of the training data, the pockets labels were manually checked using pymol (40). A few mis-annotations were detected and corrected. The final training matrix contained 339 ligand-binding (true) pockets and 1401 empty ones (false).

Table 1 : Descriptors used to train the predictive model

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Pocket score	Charge score
Druggability score	Local hydrophobic density score
Number of vertices	Number of apolar alphaspheres
Mean radius of alpha sphere	Proportion of apolar alpha sphere
Mean alpha sphere solvent accessibility	Solvent accessible surface
Mean B factor	Number of alpha helices
Hydrophobicity score	Number of coils
Polarity score	Number of strands
Volume score	Number of turns
Real volume	Number of bridges
	Number of 310 helices

We obtained the best results using a neural network multilayer perceptron classifier with an architecture of (15, 10, 5). To avoid overfitting, we trained the model using a 5-fold cross-validation. Furthermore, in order to reduce problems associated with unbalanced classes, we downsampled the largest groups according to the smallest one. The model displayed an Area Under the Curve (AUC) of 97.2% (ROC curve in Figure 1). The model accuracy was 93.4%.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 1: ROC curve for pocket prediction after 5-fold cross-validation. The Area Under the Curve (AUC) of the best model was 97.2%.

To demonstrate that PickPocket can predict ligand binding on unseen data, we selected the two most recent (at the time of writing) human protein structures containing fatty acid(s) from the PDB archive: perdeuterated human myelin protein P2 (PDBID=6S2M), which contains two possible fatty acids in the same pocket (vaccenic acid and palmitic acid), and human angiopoietin-like 4 (PDBID=6U1U), which contains palmitic acid. Pockets with a score ≥0.5 were considered positive and were colored in red, the others were in random colors. For both structures, PickPocket correctly identified the fatty acid binding cavity. Careful analysis of the 6S2M structure showed that the cavity fatty acid occupied two pockets (red and blue in Figure 2A). The red pocket, which is the deep part inside the protein and contained the carboxyl part, had a score of 0.78. The blue pocket,

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

which is at the opening of the cavity and contained the fatty acid tail, had a score of 0.03. The fatty acid-binding cavities were large and, as illustrated in this case, fpocket tended to consider them as more than one pocket. As both pockets corresponded to the same cavity and the red pocket had a score well-above the threshold, we considered that PickPocket discovered the fatty acid-binding cavity of the structure. For 6U1U (Figure 2B), the red pocket, corresponding to the palmitic acid cavity, received a score of 0.69.

Fig 2. Prediction details for the recently published structures 6S2M (A) and 6U1U (B). Each set of colored balls represents a pocket. Only the relevant parts of the structure are shown here. Some parts of the protein were set transparent to facilitate the visualization, and some negative pockets distant from the fatty acid were hidden to help visibility. The red pockets are correctly predicted, while the blue pocket, which corresponds to the palmitic acid tail, was incorrectly categorized as false.

Table 2 : Prediction results

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Structure	Pocket color	Prediction Score
6S2M(A)	RED	0.78
6S2M(A)	BLUE	0.03
6S2M(A)	WHITE	0.18
6S2M(A)	YELLOW	0.2
6S2M(A)	PURPLE	0.04
6U1U(B)	RED	0.69
6U1U(B)	YELLOW	0.0
6U1U(B)	PURPLE	0.0
6U1U(B)	LIGHT GREEN	0.19
6U1U(B)	GREEN	0.0

Prediction scores for the pockets of structure 6S2M and 6U1U. Score >0.5 are considered positives and colored green. Color corresponds to Figure 2. Complete prediction results can be found in supplementary information (SI Table 2 : Full prediction results for 6S2M and 6U1U PDB structures).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Discussion and perspectives

Thanks to the increased availability of structural data and the improvement of protein 3D modeling, new pocket-based methodologies can now be developed to help protein function discovery. Our results show that combining pocket descriptors and the residue secondary structure is sufficient to train a model and to predict the pockets for specific ligand families with high accuracy, including when using unseen data. We believe that careful analysis of protein ligand families and their corresponding binding pockets coupled with the high prediction capacity of neural networks is the way forward to close the gap between protein sequences and their functions. PickPocket aims to simplify the procedure for building a ligand family dataset and train a model to recognize the corresponding cavity. The ligand selection and the corresponding structure still need to be done manually, but these steps are made easier by the PDB ligand research tool. Other databases, such as PoSSuM (28), may be used to create the dataset.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

PickPocket automatic labelling of 'true' and 'false' pockets is very fast, but still needs manual checking because sometimes it makes mistakes when pockets are very close to each other. The training data quality is extremely important for good predictions. Therefore, we strongly recommend users to manually check the generated dataset. Looking in detail at the results for the 6S2M structure, fpocket considered that it had two different pockets because the fatty acid binding cavity is very deep and narrow. This can be corrected by changing the alpha-sphere radius or the maximum distance between pockets. One of the challenges we faced was to tune the fpocket parameters in order to have a big enough pocket size without merging different cavities. PickPocket easy tuning of these parameters allows users to adapt the input to the ligand specificity.

In order to cluster ligand protein complexes, Deepdrug3D and FRsite use an atom-based voxelization. This step also allows generating a compatible input for convolutional neural networks. On the other hand, our methodology uses matrix properties that are faster to generate, but contain less information. We also chose to use the fpocket software, although DeepSite is more accurate against the sc-PDB database of binding sites

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(40). However, fpocket is fast, and pocket descriptors data is easily retrieved from output files.

PickPocket can help to discover new protein functions by investigating the binding sites of a specific ligand family. The results we obtained prove that for fatty acid-like molecules, pocket descriptors and secondary structure are enough to obtain predictions with >90% accuracy. Thanks to its high prediction accuracy, PickPocket can be used as a tool for *in silico* screens, and should boost novel research.

Material and Methods

PickPocket methodology can be divided into five steps (Figure 3). First a selection of ligands and structure, second the combination of fpocket and Stride to generate the descriptive matrice of the pockets, third the selection of 'true' pocket (pocket that binds the selected ligands), fourth training the desired model and last computing the prediction on the full PDB pocket dataset.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

To select a set of fatty acid-like ligands present in the PDB we used the ligand search tool. We selected ligands having this SMILES 'CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC(O)=O' as superstructure and a molecular weight superior to 100.0 g/mol.

This resulted in a list of mostly fatty acids and other molecules composed of aliphatic chains including some with alcohols.From the structures containing our ligands we only selected the one with human proteins, resolved using X-ray technique and excluding DNA and RNA molecules. A representative subset at 70% sequence identity was then used as the input. A set of random structures, using the same criteria (human, X-ray, redundancy), not containing any of the previously selected ligands was used as negative data. All pockets were computed using fpocket (19) and all secondary structure information with STRIDE (41).

The 'true' pockets containing ligands were identified using euclidean distance. For each pocket in each structure, if any voronoi vertices from a pocket and any ligand atom distance is inferior or equal to 1 angstrom, then the pockets are labelled as 'true'. All other pockets in the PDB file are by default categorized as 'false'.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

To train the predictive model PickPocket offers the possibility to use random forest, support vector machine algorithms or neural networks implemented in Python. The neural network is configured to test multiple architectures and automatically select the model with the best accuracy. The Python program included the following libraries: scikit-learn(41), numpy (42) and pandas (43).

Once the desired model it can be saved and used to make predictions on unseen data. A file containing all the Protein Data Bank pockets (all pockets from all structures) can be found in the data attached to the software.

Figure 3 : methodology workflow. Five steps of the PickPocket methodology for specific ligands binding prediction.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Funding

This work was supported by Merck Sharp and Dohme Avenir (GnoSTic) to S. Kossida and by the ATIP-Avenir program and La Ligue contre le Cancer (France) to M. Moriel-Carretero.

Bibliographie

1. Hawkins T, Kihara D. FUNCTION PREDICTION OF UNCHARACTERIZED PROTEINS [Internet]. Vol. 05, Journal of Bioinformatics and Computational Biology. 2007. p. 1–30. Available from: http://dx.doi.org/10.1142/s0219720007002503

2. Consortium TU, The UniProt Consortium. The Universal Protein Resource (UniProt) 2009 [Internet]. Vol. 37, Nucleic Acids Research. 2009. p. D169–74. Available from: http://dx.doi.org/10.1093/nar/gkn664

3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389–402.

5. Bairoch A. The PROSITE database, its status in 1995 [Internet]. Vol. 24, Nucleic Acids Research. 1996. p. 189–96. Available from: http://dx.doi.org/10.1093/nar/24.1.189

6. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro, progress and status in 2005. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D201–5.

7. Coggill P, Finn RD, Bateman A. Identifying protein domains with the Pfam database. Curr Protoc Bioinformatics. 2008 Sep;Chapter 2:Unit 2.5.

8. Troyanskaya OG. Integrated Analysis of Microarray Results [Internet]. Methods in Molecular Biology. 2007. p. 429–37. Available from: http://dx.doi.org/10.1007/978-1-59745-304-2_27

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

9. Rzhetsky A, Seringhaus M, Gerstein M. Seeking a new biology through text mining. Cell. 2008 Jul 11;134(1):9–13.

10. Si L, Yu D, Kihara D, Fang Y. Combining gene sequence similarity and textual information for gene function annotation in the literature [Internet]. Vol. 11, Information Retrieval. 2008. p. 389–404. Available from: http://dx.doi.org/10.1007/s10791-008-9053-0

11. Petrey D, Chen TS, Deng L, Garzon JI, Hwang H, Lasso G, et al. Template-based prediction of protein function. Curr Opin Struct Biol. 2015 Jun;32:33–8.

12. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, et al. The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res. 2013 Jan;41(Database issue):D475–82.

13. Brylinski M. Is the growth rate of Protein Data Bank sufficient to solve the protein structure prediction problem using template-based modeling? [Internet]. Vol. 11, Bio-Algorithms and Med-Systems. 2015. p. 1–7. Available from: http://dx.doi.org/10.1515/bams-2014-0024

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

14. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph. 1995 Oct;13(5):323–30, 307–8.

15. Peters KP, Fauck J, Frömmel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol. 1996 Feb 16;256(1):201–13.

16. Laurie ATR, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics. 2005 May 1;21(9):1908–16.

17. Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol. 2006 Sep 24;6:19.

18. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol. 2009 Dec;5(12):e1000585.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

19. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. 2009 Jun 2;10:168.

20. Tan KP, Varadarajan R, Madhusudhan MS. DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W242–8.

21. Bianchi V, Gherardini PF, Helmer-Citterich M, Ausiello G. Identification of binding pockets in protein structures using a knowledge-based potential derived from local structural similarities. BMC Bioinformatics. 2012 Mar 28;13 Suppl 4:S17.

22. Jiang M, Wei Z, Zhang S, Wang S, Wang X, Li Z. FRSite: Protein drug binding site prediction based on faster R–CNN [Internet]. Vol. 93, Journal of Molecular Graphics and Modelling. 2019. p. 107454. Available from: http://dx.doi.org/10.1016/j.jmgm.2019.107454

23. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017 Oct 1;33(19):3036–42.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

24. Pu L, Govindaraj RG, Lemoine JM, Wu H-C, Brylinski M. DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network. PLoS Comput Biol. 2019 Feb;15(2):e1006718.

25. Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminform. 2018 Aug 14;10(1):39.

26. Tian W, Chen C, Lei X, Zhao J, Liang J. CASTp 3.0: computed atlas of surface topography of proteins. Nucleic Acids Res. 2018 Jul 2;46(W1):W363–7.

27. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues [Internet]. Vol. 34, Nucleic Acids Research. 2006. p. W116–8. Available from: http://dx.doi.org/10.1093/nar/gkl282

28. Ito J-I, Ikeda K, Yamada K, Mizuguchi K, Tomii K. PoSSuM v.2.0: data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs [Internet]. Vol. 43, Nucleic Acids

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Research. 2015. p. D392–8. Available from: http://dx.doi.org/10.1093/nar/gku1144

29. Kufareva I, Ilatovskiy AV, Abagyan R. Pocketome: an encyclopedia of small-molecule binding sites in 4D. Nucleic Acids Res. 2012 Jan;40(Database issue):D535–40.

30. Ansar S, Centre for Bioinformatics, Kamalnayan Bajaj Institute for Research in Vision and Ophthalmology, Vision Research Foundation, SankaraNethralaya, - C, et al. PocketPipe: A computational pipeline for integrated Pocketome prediction and comparison [Internet]. Vol. 15, Bioinformation. 2019. p. 295–8. Available from: http://dx.doi.org/10.6026/97320630015295

31. Brylinski M. Local Alignment of Ligand Binding Sites in Proteins for Polypharmacology and Drug Repositioning [Internet]. Methods in Molecular Biology. 2017. p. 109–22. Available from: http://dx.doi.org/10.1007/978-1-4939-7015-5_9

32. Najmanovich RJ. Evolutionary studies of ligand binding sites in proteins. Curr Opin Struct Biol. 2017 Aug;45:85–90.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

33. Barelier S, Sterling T, O'Meara MJ, Shoichet BK. The Recognition of Identical Ligands by Unrelated Proteins [Internet]. Vol. 10, ACS Chemical Biology. 2015. p. 2772–84. Available from: http://dx.doi.org/10.1021/acschembio.5b00683

34. Ehrt C, Brinkjost T, Koch O. Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. J Med Chem. 2016 May 12;59(9):4121–51.

35. Naderi M, Lemoine JM, Govindaraj RG, Kana OZ, Feinstein WP, Brylinski M. Binding site matching in rational drug design: algorithms and applications. Brief Bioinform [Internet]. 2018 Aug 31; Available from: http://dx.doi.org/10.1093/bib/bby078

36. Tierney L. The R Statistical Computing Environment [Internet]. Lecture Notes in Statistics. 2012. p. 435–47. Available from: http://dx.doi.org/10.1007/978-1-4614-3520-4_41

37. Python Language Reference [Internet]. Python for Bioinformatics. 2009. p. 457–538. Available from: http://dx.doi.org/10.1201/9781584889304.axd

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

38. Schmidhuber J. Deep learning in neural networks: An overview [Internet]. Vol. 61, Neural Networks. 2015. p. 85–117. Available from: http://dx.doi.org/10.1016/j.neunet.2014.09.003

39. Bhagavat R, Sankar S, Srinivasan N, Chandra N. An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure. Structure. 2018 Mar 6;26(3):499–512.e2.

40. Desaphy J, Bret G, Rognan D, Kellenberger E. sc-PDB: a 3D-database of ligandable binding sites—10 years on [Internet]. Vol. 43, Nucleic Acids Research. 2015. p. D399–404. Available from: http://dx.doi.org/10.1093/nar/gku928

41. Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins [Internet]. Vol. 32, Nucleic Acids Research. 2004. p. W500–2. Available from: http://dx.doi.org/10.1093/nar/gkh429

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.15.042655; this version posted April 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

NF90 modulates processing of a subset of human pri-miRNAs

During my first year of PhD I helped G.G. to perform statistical analysis for her project.

6874–6888 *Nucleic Acids Research, 2020, Vol. 48, No. 12* *Published online 19 May 2020*
doi: 10.1093/nar/gkaa386

NF90 modulates processing of a subset of human pri-miRNAs

Giuseppa Grasso[1], Takuma Higuchi[2], Victor Mac[1], Jérôme Barbier[1], Marion Helsmoortel[1], Claudio Lorenzi[3], Gabriel Sanchez[1], Maxime Bello[1], William Ritchie[3], Shuji Sakamoto[2] and Rosemary Kiernan[1,*]

[1]UMR9002 CNRS-UM, Institut de Génétique Humaine-Université de Montpellier, Gene Regulation lab, Montpellier 34396, France, [2]Laboratory of Molecular Biology, Science Research Centre, Kochi Medical School, Kochi University, Kochi 783-8505, Japan and [3]UMR9002 CNRS-UM, Institut de Génétique Humaine-Université de Montpellier, Artificial Intelligence and Gene Regulation lab, Montpellier 34396, France

Received July 16, 2019; Revised April 24, 2020; Editorial Decision April 30, 2020; Accepted May 01, 2020

ABSTRACT

MicroRNAs (miRNAs) are predicted to regulate the expression of >60% of mammalian genes and play fundamental roles in most biological processes. Deregulation of miRNA expression is a hallmark of most cancers and further investigation of mechanisms controlling miRNA biogenesis is needed. The double stranded RNA-binding protein, NF90 has been shown to act as a competitor of Microprocessor for a limited number of primary miRNAs (pri-miRNAs). Here, we show that NF90 has a more widespread effect on pri-miRNA biogenesis than previously thought. Genome-wide approaches revealed that NF90 is associated with the stem region of 38 pri-miRNAs, in a manner that is largely exclusive of Microprocessor. Following loss of NF90, 22 NF90-bound pri-miRNAs showed increased abundance of mature miRNA products. NF90-targeted pri-miRNAs are highly stable, having a lower free energy and fewer mismatches compared to all pri-miRNAs. Mutations leading to less stable structures reduced NF90 binding while increasing pri-miRNA stability led to acquisition of NF90 association, as determined by RNA electrophoretic mobility shift assay (EMSA). NF90-bound and downregulated pri-miRNAs are embedded in introns of host genes and expression of several host genes is concomitantly reduced. These data suggest that NF90 controls the processing of a subset of highly stable, intronic miRNAs.

INTRODUCTION

MicroRNAs (miRNAs) are short non-coding RNAs that negatively regulate the expression of a large proportion of cellular mRNAs, thus affecting a multitude of cellular and developmental pathways (1,2). The canonical miRNA biogenesis pathway involves two sequential processing events catalysed by RNase III enzymes. In the nucleus, the microprocessor complex, comprising the RNase III enzyme Drosha, the double-stranded RNA-binding protein, DGCR8 and additional proteins carries out the first processing event, which results in the production of precursor miRNAs (pre-miRNAs) (3,4). These are exported to the cytoplasm, where a second processing event is carried out by another RNase III enzyme, DICER, leading to the production of miRNA duplexes. The duplexes are loaded into the RISC complex and the release of the 'passenger' strands leads to the formation of mature miRNAs and mature RISC complexes (5).

Due to the central role of miRNAs in the control of gene expression, their levels must be tightly controlled. Indeed, deregulation of miRNA expression is associated with aberrant gene expression and leads to human disease (6–9). Consequently, miRNA biogenesis is tightly regulated at multiple steps, both transcriptional and post-transcriptional. Increasing evidence suggests that RNA binding proteins (RBPs) act as post-transcriptional regulators of miRNA processing. Many RBPs modulate the processing efficiency of Microprocessor, either positively or negatively, by binding to regions of the pri-miRNA. A number of RBPs have been shown to bind the terminal loop, which can either facilitate or inhibit cropping by Microprocessor. For example, LIN28B binds the terminal loop of pri-let-7, which prevents its processing by Microprocessor (10). Binding of hnRNP A1 to the terminal loop has been shown to exert either positive or negative effects on Microprocessor activity, depending on the pri-miRNA target. It promotes cropping of pri-miR-18A while it inhibits processing of pri-let-7. KSRP is another terminal loop-binding RBP that facilitates Microprocessor cleavage of several pri-miRNA targets, including pri-let-7 where it acts as a competitor of hn-

*To whom correspondence should be addressed. Tel: +33 4 34359939; Fax: +33 4 34359901; Email: Rosemary.Kiernan@igh.cnrs.fr

© The Author(s) 2020. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

RNP A1 (11,12). Several other RBPs, including SMAD, TPD-43, SRSF1 and RBFOX, have been shown to bind pri-miRNA terminal loops to influence Microprocessor activity (see (13) for review). In most cases, they have been shown to bind specific pri-miRNAs, such as pri-let-7, or a limited subset of pri-miRNAs. To date, only NF90/NF45 heterodimer and ADAR1,2 have been shown to bind the double stranded stem region of pri-miRNAs (14–17). Both factors negatively affect Microprocessor activity. Indeed, NF90 has been shown to bind double stranded RNA in a mode similar to that of ADAR2 (18). Like terminal loop binding RBPs, binding of NF90/NF45 or ADAR1,2 has thus far been demonstrated for a very limited number of pri-miRNAs. NF90 has been shown to associate with pri-miR-7-1, pri-let-7A and pri-miR-3173 in human cells (14,15,19).

We have previously shown that NF90 associates with pri-miR-3173, which is located in the first intron of Dicer pre-mRNA (19). Binding of NF90 prevented cropping of pri-miR-3173 by Microprocessor and promoted splicing of the intron, thereby facilitating expression of DICER. By modulating DICER expression, NF90 was found to be an independent prognostic marker of ovarian carcinoma progression (19). Levels of NF90 are known to be elevated in hepatocellular carcinoma (HCC) and the effect of NF90 on processing of pri-miR-7-1 contributes to cellular proliferation in HCC models (14,20). Here, we have used genome-wide approaches to identify pri-miRNAs that are associated with and modulated by NF90 in HepG2 model of HCC. We identified 38 pri-miRNAs that are associated with NF90, in a manner that is for the most part exclusive of Microprocessor. Of these, 22 showed increased abundance of mature miRNAs products upon loss of NF90. NF90-targeted pri-miRNAs appear to be highly stable, having a lower free energy and fewer mismatches compared to all pri-miRNAs. Destabilization of the structures by mutation reduced NF90 association as determined by RNA EMSA. Of the 22 NF90-modulated pri-miRNAs, 20 are embedded exclusively in introns of host genes. Transcriptomic analysis revealed that the expression of the host gene is concomitantly downregulated for several, including an oncogene implicated in metastasis of hepatocellular carcinoma, TIAM2. These data suggest that NF90 controls the processing of a subset of intronic miRNAs, which in some cases affects the expression of the host gene.

MATERIALS AND METHODS

Cell culture

Human HepG2 cell line was grown in Dulbecco's modified Eagle's medium—high glucose (Sigma-Aldrich®, D6429) supplemented with 10% fetal bovine serum (PAN Biotech, 8500-P131704), 1% penicillin–streptomicin (v/v) (Sigma Aldrich®, P4333) and 1% L-glutamine (v/v) (Sigma Aldrich®, G7513). Human HEK-293T cells were grown in Dulbecco's modified Eagle's high glucose medium with HEPES (Sigma-Aldrich®, D6171) supplemented with 10% fetal bovine serum, 1% penicillin–streptomicin and 1% L-glutamine. Cells were cultured at 37°C in a humidified atmosphere containing 5% CO_2. To perform small RNA-seq and RNA-seq, HepG2 were seeded at 1.5×10^6 cells in six-well plates the day of siRNA transfection while HEK-293T were seeded at 6×10^5 cells in six-well plates.

To perform RNA Immunoprecipitation, HepG2 were seeded at 8×10^6 cells in 100 mM culture dishes the day of siRNA transfection.

Transfection of small interfering RNAs

Double-stranded RNA oligonucleotides used for RNAi were purchased from Eurofins MWG Operon or Integrated DNA Technologies. Sequences of small interfering RNAs (siRNAs) used in this study have been described previously (19) and are shown in Supplementary Table S1.

HepG2 or HEK-293T cells were transfected with siRNA (30 nM final concentration) using INTERFERin® siRNA transfection reagent (Polyplus Transfection) according to the manufacturer's instructions. To perform small RNA-seq and RNA-seq, two rounds of transfection were performed. The first transfection was carried out the day of seeding; on the fourth day cells were passaged and a second round of transfection was performed. Cells were collected for RNA extraction or protein purification ∼65 h after the second transfection. To perform RNA Immunoprecipitation, one round of siRNA transfection was carried out, as explained, the day of seeding. Cells were collected ∼65 h after siRNA transfection.

Immunoblot

HepG2 were lysed using RIPA buffer (50 mM Tris–HCl pH 7.5, 150 mM NaCl, 1% NP40, 0.5% Sodium Deoxycholate, 0.1% SDS, Halt™ Phosphatase Inhibitor Cocktail (Thermo Fisher Scientific)). Protein extracts (30 μg for NDUFS8, 50 μg for TIAM2 and 5 μg for all other proteins) were immunoblotted using the indicated primary antibodies (Supplementary Table S2) and anti-mouse, anti-rabbit or anti-rat IgG-linked HRP secondary antibodies (GE Healthcare) followed by ECL (Advansta).

Small RNA-seq and RNA-seq

Total RNA was extracted using TRIzol (Thermo Fisher Scientific) according to the manufacturer's instructions. Small RNA-seq (single end, 50 bp) was carried out by BGI Genomic Services (HepG2) or Fasteris (HEK-293T) in triplicate samples. Raw data were processed using the Subread package (version 1.6.0) as previously described (21) and the reference annotation was obtained from miRBase release 22.1 database (22). Statistical analysis was performed using DESeq2 (version 2.11.40.2). RNA-seq (paired-end, 125 bp) was carried out by BGI Genomic Services in triplicates. Raw data were processed using HISAT2 (version 2.1.0) and featureCounts (version 1.6.3), statistical analysis was performed using DESeq2. Reference annotation was obtained from ENSEMBL (GRCh38.96).

RT-qPCR, modified 5′ RLM RACE and RNA EMSA

Total RNA was extracted from HepG2 cells using TRIzol reagent (Thermo Fisher Scientific) and RNA was treated with DNAse I (Promega) according to the manufacturer's

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

instructions. RNA was used for RT-PCR and modified 5′ RLM-RACE as described previously (19).

For RT-qPCR, RT was performed using TaqMan™ Reverse Transcription Reagent or TaqMan™ Advanced miRNA cDNA Synthesis Kit (Thermo Fisher). qPCRs were performed using GoTaq® Probe qPCR Master Mix (Promega) or TaqMan® Fast Advanced Master Mix (Thermo Fisher).

Modified 5′ RLM RACE was performed according to the manufacturer's instructions (FirstChoice™ RLM-RACE kit, ThermoFisher Scientific). In order to detect premature miRNAs, the step using calf intestine alkaline phosphatase was omitted. Sequences of the primers used for PCR amplification are shown in Supplementary Table S3.

RNA EMSA was performed as described previously (15) using recombinant NF90 and recombinant DGCR8 dsRBDs (amino acids 484–773) in at least three replicates. The pri-miRNA probes were amplified by PCR using the primers shown in Supplementary Table S3. Sequences of mutant pri-miRNAs are shown in Supplementary Table S4.

RNA immunoprepicipation (RIP)

RIP was performed as previously described (23). HepG2 were seeded in 100 mm culture dishes and transfected with siRNAs the day of seeding as aforementioned. Cells were harvested ∼65 h after the treatment and lysed for 15 min in RIP buffer (20 mM HEPES, pH 7.5, 150 mM NaCl, 2.5 mM $MgCl_2 \bullet 6H_2O$, 250 mM sucrose, 0.05% (v/v) NP-40 and 0.5% (v/v) Triton X-100) containing 20 U ml^{-1} of RNasin (Promega), 1 mM DTT, 0.1 mM PMSF and EDTA-free protease and phosphatase inhibitor. After centrifugation, lysates were incubated for 4 h at 4°C with 2 μg of antibodies recognizing NF90, Drosha and IgG control and then incubated for 1 h at 4°C with Dynabeads™ Protein A (ThermoFisher Scientific). After incubation, beads were washed five times with RIP buffer for 5 min at 4°C and RNA was extracted as previously explained. RNA was treated with DNAse I (Promega) and RT was performed using SuperScript™ III Reverse Transcriptase (ThermoFisher Scientific) according to the manufacturer's instructions. cDNA was treated with RNAse H (ThermoFisher Scientific) and the samples were used to perform qPCRs using QuantiTect SYBR® Green PCR Kit (Qiagen) according to the manufacturer's instructions.

Splicing analysis

Splicing analyses were carried out as previously described (19). HepG2 were seeded in six-well plates and transfected with siRNAs, as aforementioned. Approximately 65 h after the second transfection, RNA was extracted using TRIzol reagent (ThermoFisher Scientific) and treated with DNAse I (Promega) according to the manufacturer's instructions. RT was performed using SuperScript™ III Reverse Transcriptase (ThermoFisher Scientific) and cDNA was treated with RNAse H (ThermoFisher Scientific). qPCRs were performed using QuantiTect SYBR® Green PCR Kit (Qiagen) using primers overlapping exon–intron boundaries to detect unspliced pre-mRNAs or primers amplifying exon-exon boundaries to detect the spliced mRNA.

Bioinformatic analyses

Enhanced UV crosslinking followed by immunoprecipitation (eCLIP) data for NF90, DGCR8 and DROSHA obtained in HepG2 cells by Nussbacher and Yeo (24) were retrieved from the NCBI database (NF90 eCLIP: ENCSR786TSC; DGCR8 eCLIP: ENCSR061SZV; DROSHA eCLIP: ENCSR834YLD). Peaks were filtered based on Fold Change (FC ≥ 1.5) and *P*-value (Bonferroni-Adj *P*-val ≤ 0.05). Distribution of eCLIP reads along the miRNAs was evaluated using deeptools software (version 3.1.3). Bigwig files from different replicates were merged using bigWigMerge v2. The base pair probability at each position of miRNA hairpins was calculated using RNAfold software (version 2.4.7).

Free energy analysis was performed using RNAfold software, version 2.4.7. Statistical analysis was performed using R (version 3.5.1).

Validated targets of the double positive miRNAs were extracted from MirTarBase database, release 7.0 (25). Gene ontology was performed on the expressed validated target using DAVID Functional Annotation Tool database version 6.8 (https://david.ncifcrf.gov) (26). Motif search was performed using MEME (version 5.0.5).

RESULTS

NF90 affects the abundance of a subset of human miRNAs

To determine the effect of NF90 on the abundance of miRNAs, we performed small RNA-seq of biological triplicate samples obtained from HepG2 cells that had been transfected with a non-targeting control siRNA (siScr) or an siRNA targeting NF90 (siNF90) (Figure 1A, top panel). Of 1917 miRNA precursors annotated in miRBase, 1105, which corresponds to 1661 mature 5p and 3p miRNA products, were found to be expressed in HepG2 cells. Following loss of NF90, differential expression analysis (fold change ≥ 1.5 or ≤ 0.667; Adj*P*-value ≤ 0.05) showed that 268 mature miRNAs, corresponding to 212 precursor miRNAs, were upregulated while 149, corresponding to 126 precursor miRNAs, were downregulated (Figure 1B). The number of upregulated and downregulated miRNAs in HepG2 cells after loss of NF90 is summarized in Figure 1C. MiRNAs that have previously been shown to be repressed by NF90, miR-7-1 (14) and miR3173 (19), were found to be upregulated in HepG2 cells following loss of NF90 (Figure 1B, red dots).

The effect of NF90 on the abundance of miRNAs observed by miRNA profiling were validated by RT-qPCR analysis of selected miRNAs, miR-3173-3p, miR-186-5p, miR-1273c and miR-3189-3p, from biological triplicate samples. The results obtained confirmed the effects observed by miRNA profiling (Figure 1B, D). In addition, RNA was extracted from cells transfected with an independent non-targeting siRNA (Scr#2) and an NF90-targeting siRNA (NF90#2) that has been described previously (19) (Figure 1A, lower panel). Quantification of miRNAs 3173-3p, -186-5p, -1273c and -3189-3p in biological triplicate samples (Figure 1D, lower panels) showed similar results to those obtained in Figure 1D upper panel, and also validated the results obtained by small RNA-seq. While we cannot

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

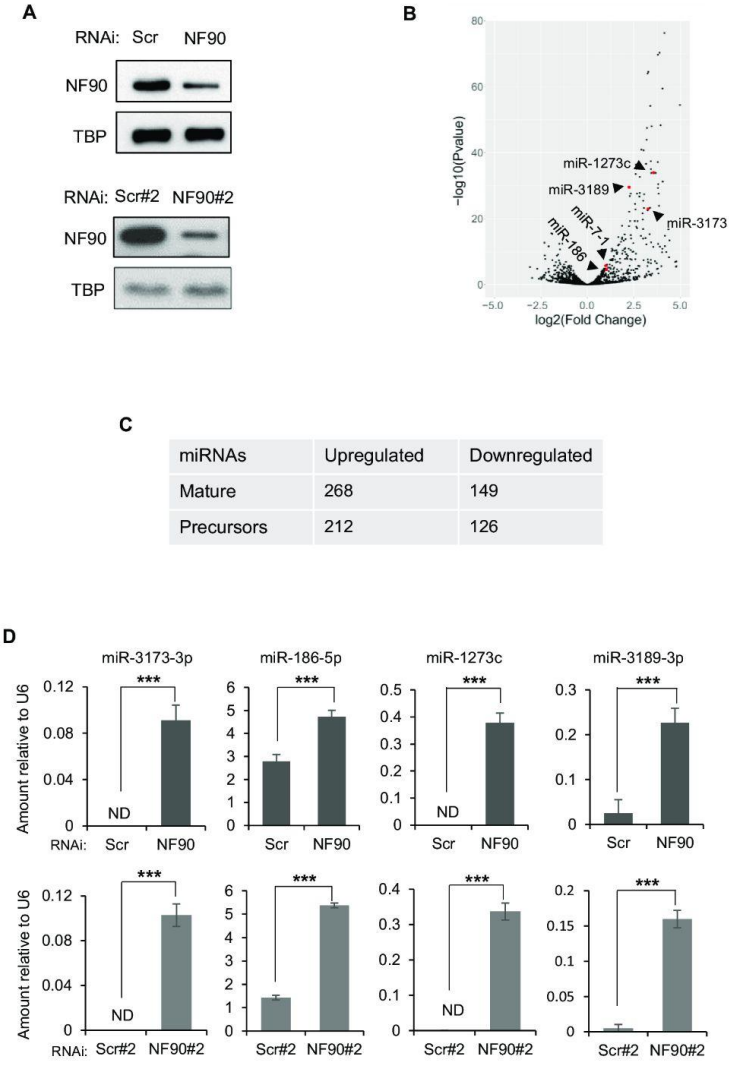


Figure 1. NF90 modulates the expression level of a subset of miRNAs in HepG2 cells. (**A**) Extracts of HepG2 cells transfected with non-targeting control siRNAs (Scr, Scr#2) or siRNA targeting NF90 (NF90, NF90#2) as indicated were analyzed by immunoblot using the antibodies indicated. (**B**) Total RNA extracted from cells transfected with siScr or siNF90 were analyzed by small RNA-seq. Results are shown as $\log_2$ fold change versus $-\log_{10}$ *P*-value. (**C**) Table summarizing the number of mature miRNAs and pri-miRNAs modulated in HepG2 cell line upon loss of NF90, according to small-RNA seq. (**D**) Total RNA extracted from cells described in (A) were analyzed by Taqman RT-qPCR as indicated. Results were normalized by those obtained for U6 abundance in the same samples. ND indicates 'not detected'. Data represent mean ± SEM obtained from three independent experiments (***$P < 0.001$, independent Student's *t* test).

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

exclude the possibility that a proportion of the small RNA-seq results could be due to off-target effects of the siRNAs, since only a single control and NF90-targeting siRNA were used, validation of a subset of the results using additional control and NF90-targeting siRNA suggests that the data are, to some extent, robust.

To evaluate whether the effect of NF90 on miRNA abundance might be cell type specific, we performed small RNA-seq in biological triplicate in HEK-293T cells transfected with control or NF90-targeting siRNA (Supplementary Figure S1A). Of 1917 annotated miRNA precursors, 1121, corresponding to 1647 mature miRNAs, were expressed in HEK-293T. Differential expression analysis (fold change $\geq$ 1.5 or $\leq$ 0.667; Adj*P*-value $\leq$ 0.05) revealed that 278 mature miRNAs, corresponding to 217 miRNA precursors, were upregulated following loss of NF90 while 84 mature miRNAs, corresponding to 77 precursors, were downregulated (Supplementary Figures S1B, C). Comparing upregulated miRNAs in the two cell types, we found 139 miRNAs that were upregulated in both cell lines after NF90 knockdown (Supplementary Figure S1D). This represents >65% of miRNAs upregulated in HepG2 and 64% of those upregulated in HEK-293T. Thus, NF90 appears to regulate a common subset of miRNAs.

NF90 associates with a subset of pri-miRNAs

To determine which of the miRNAs upregulated upon loss of NF90 (Figure 1B) are direct targets of NF90, that is, pri-miRNAs that are bound by NF90, we took advantage of enhanced UV crosslinking followed by immunoprecipitation (eCLIP) dataset obtained in HepG2 cells (24). Analysis of HepG2 eCLIP data revealed 38 pri-miRNAs for which eCLIP peaks overlapped annotated pri-miRNA localizations +/− 25 nt of flanking region (FC $\geq$ 1.5 and Bonferroni Adj*P* $\leq$ 0.05), as depicted in Figure 2A and Supplementary Table S5. Pri-miR-3173 and pri-miR-7-1 were among the 38 NF90-associated pri-miRNAs (Figure 2A, red dots).

We next analysed eCLIP read coverage across the pri-miRNA hairpin ±200 bp for the 38 NF90-associated miRNAs compared to all pri-miRNAs (Figure 2B). As expected, analysis of all pri-miRNAs did not show significant read coverage for NF90 association. In contrast, NF90-associated miRNAs showed highest read coverage over the region having the strongest base pair probability and therefore likely corresponding to the double stranded pri-miRNA stem (Figure 2B). The region corresponding to the terminal loop, which has a low base pair probability, was not significantly bound by NF90. Interestingly, NF90 also appeared to bind to the pri-miRNA flanking region. Browser shots showing NF90 association with pri-miR-7-1, pri-miR-186 and pri-miR-1273c by eCLIP are shown in Supplementary Figure S2A.

To validate NF90 association with pri-miRNAs identified by eCLIP analysis (Figure 2A), we performed RNA EMSA using pri-miR-186, pri-miR-3173, pri-miR-1273c and pri-miR-3189 as radiolabeled probes together with recombinant NF90 (Supplementary Figure S2B), as described previously for pri-miR-7-1 and pri-miR-3173 (14,19). RNA EMSA, performed in triplicate, confirmed NF90 association with pri-miR-186, pri-miR-3173, pri-miR-1273c and pri-miR-3189 (Figure 2C and Supplementary Figure S2C). Similarly, RNA EMSA confirmed that NF90 was not highly associated with pri-miR-200a, as indicated by eCLIP (Figure 2C). NF90 association with the pri-miRNAs identified by eCLIP analysis was also validated for several endogenous pri-miRNAs by performing RNA immunoprecipitation (RIP). RIP confirmed the association of NF90 with region proximal to the endogenous pri-miRNA (Figure 2D and Supplementary Figure S2D), while negative controls, pri-miR-200a and DALRD3, were not significantly associated with NF90. In contrast, pri-miR-200a was significantly bound by Drosha (Figure 2D and Supplementary Figure S2D). While not all NF90-bound pri-miRNAs identified by eCLIP have been tested, RIP analysis confirmed the association with NF90 *in vivo* for at least several.

Previous studies have indicated that NF90 may act as a competitor of Microprocessor for binding to pri-miRNAs (13–15,19). We therefore analysed eCLIP data for DGCR8 and Drosha performed in HepG2 cells (24). Association of DGCR8 was detected at 203 pri-miRNAs, while 147 pri-miRNAs were positive for Drosha binding (Figure 3A). Not surprisingly, there was a significant overlap between pri-miRNAs that were bound by both subunits of Microprocessor (Figure 3A). Indeed, 125 pri-miRNAs were associated with both factors, which represents approximately 60% and 85% of pri-miRNAs positive for DGCR8 and Drosha, respectively. Interestingly, only 10 pri-miRNAs bound by NF90 overlapped with those bound by either DGCR8 or Drosha, which represents approximately 24% overlap with DGCR8 and 13% overlap with Drosha (Figure 3A). This result indicates that NF90-associated pri-miRNAs are not highly associated with Microprocessor. Analysis of eCLIP reads showed association of DGCR8 with both apical and stem regions of pri-miRNAs (Figure 3B), as expected (27).

We further analysed eCLIP read coverage over the pri-miRNAs that were found to be associated with both NF90 and DGCR8. While the profile for DGCR8 was similar to that for all DGCR8 positive pri-miRNAs (compare Figure 3C, top panel to Figure 3B), the profile for NF90 read coverage was somewhat different to that for all NF90-positive pri-miRNAs (compare Figure 3C, lower panel, to Figure 2B). Interestingly, for pri-miRNAs that are bound by both DGCR8 and NF90, the profiles appear to be complementary (Figure 3C, compare top and lower panels). Plot profiles of DROSHA and DGCR8 eCLIP data suggest that pri-miRNAs common with NF90 (shown with red dots) are not among the most enriched for Microprocessor binding (Figure 3D and Supplementary Figure S3).

To further explore the competition between NF90 and the Microprocessor for the binding of pri-miRNAs, we performed RNA EMSA on pri-miR-3189 and pri-miR-1273c using recombinant NF90 and the dsRNA-binding domains of DGCR8 (Supplementary Figures S2B and S4A). Upon addition of rNF90, a shift corresponding to the formation of NF90-pri-miRNA complex and a reduction in the intensity of the band corresponding to DGCR8-pri-miRNA complex could be detected (Figure 4A). These results indicate that NF90 competes with Microprocessor for binding to certain pri-miRNAs, at least *in vitro*. Further analysis will be required to determine whether this competition also occurs *in vivo*.

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

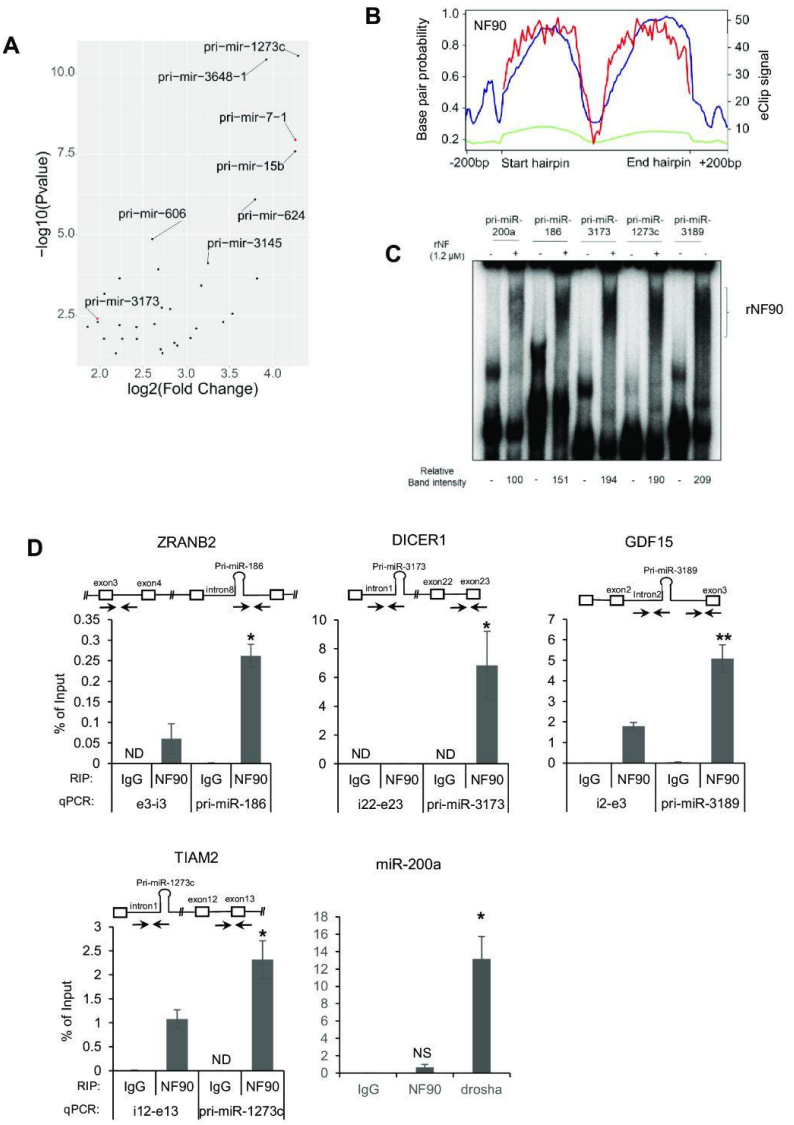


Figure 2. NF90 is associated with a subset of pri-miRNAs in HepG2 cells. (**A**) Dot plot representation of eCLIP data showing the 38 pri-miRNAs significantly associated with NF90 in HepG2 cells. Graph shows log2 fold change versus $-\log_{10}$ *P*-value. (**B**) Distribution of NF90 eCLIP reads along the region ±200 bp of NF90-associated pri-miRNAs (blue) or all miRNAs (green) and base pair probability of NF90-associated hairpins (red). (**C**) RNA EMSA performed using recombinant NF90 was probed with radiolabelled pri-miRNAs as indicated. rNF90-pri-miRNA complexes are indicated on the figure. (**D**) RIP analysis of HepG2 cells transfected with NF90 targeting siRNA or a non-targeting control (Scr), as indicated using anti-NF90, anti-Drosha or a control IgG antibody. Immunoprecipitates were analyzed by RT-qPCR amplifying a region proximal or distal to the miRNAs. ND indicates 'Not Detected'. NS indicates 'Not Significant'. Data represent mean ± SEM obtained from 3 independent experiments (*$P < 0.05$, **$P < 0.01$, independent Student's *t* test).

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

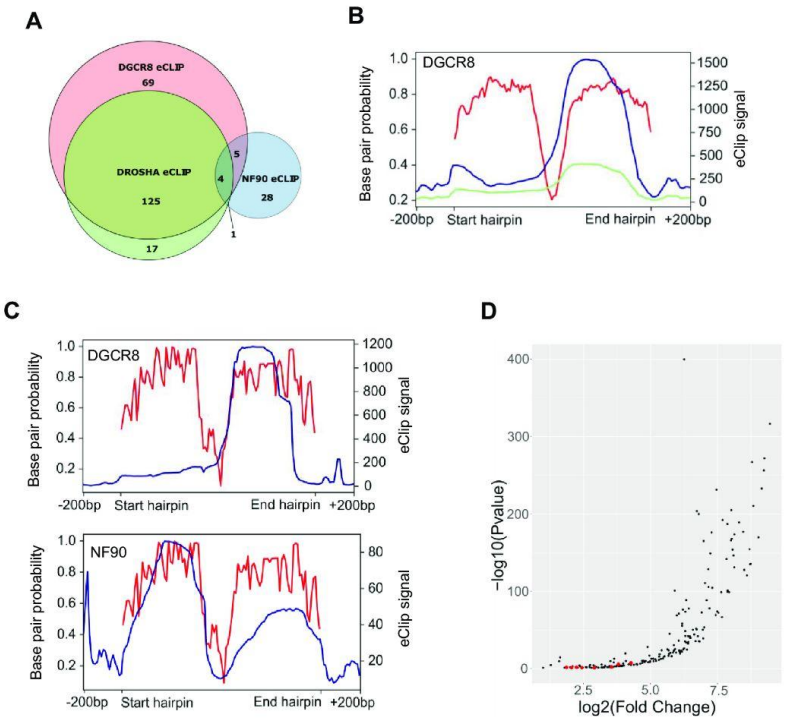


Figure 3. NF90-associated pri-miRNAs are poorly associated with Microprocessor. (**A**) Venn diagram showing the number of pri-miRNAs associated with DGCR8, Drosha or NF90 detected by eCLIP, as indicated. (**B**) Distribution of DGCR8 eCLIP reads along the region ±200 bp of DGCR8-associated pri-miRNAs (blue) or all miRNAs (green) and base pair probability of DGCR8-associated hairpins (red). (**C**) Distribution of eCLIP reads along the region ±200 bp of pri-miRNAs associated with both DGCR8 and NF90 (blue) and base pair probability of the hairpins (red). Left panel shows DGCR8 eCLIP reads, right panel shows NF90 eCLIP reads in blue. (**D**) Dot plot representation of eCLIP data showing 203 pri-miRNAs significantly associated with DGCR8 in HepG2 cells. Graph shows log2 fold change versus $-\log_{10}$ *P*-value. Red dots indicate pri-miRNAs that are also significantly associated with NF90.

We next tested whether loss of NF90/NF45 or Drosha/DGCR8 complexes could affect the binding of the complexes to endogenous pri-miRNAs *in vivo*. We performed RIP of NF90, Drosha or IgG control after downregulation of either NF90/NF45 or Drosha/DGCR8. Drosha association with the region surrounding the target pri-miRNAs was significantly enhanced after downregulation of NF90/NF45, while NF90 association was significantly enhanced after downregulation of Drosha/DGCR8 only for pri-miR-1273c (Figure 4B and Supplementary Figure S4B). This could be explained considering that these miRNAs are already poorly bound by the Microprocessor. To test this hypothesis, we analysed the association of NF90 to two pri-miRNAs poorly bound by NF90, pri-miR-200a and pri-miR-425. Notably, NF90 association with these miRNAs was significantly increased after loss of Drosha/DGCR8 complex (Supplementary Figure S4C). On the other hand, downregulation of NF90/NF45 complex did not significantly affect the association of pri-miR-200a and pri-miR-425 with Drosha (Supplementary Figure S4C), possibly because these miRNAs are poorly bound by NF90/NF45 under control conditions. Taken together, these results suggest that target pri-miRNAs may have binding preferences for either NF90/NF45 or Microprocessor under wild-type conditions, but that the relative abundance of these complexes can also influence the observed binding to specific pri-miRNAs.

Pri-miRNAs that are bound and downregulated by NF90 are highly stable

We next asked whether NF90 association with pri-miRNAs might affect their cropping by Microprocessor. If so, loss of NF90 would be predicted to increase the abundance of the mature miRNA products, as observed previously (14,15,19). MiRNA profiling revealed that of the 38 NF90-associated pri-miRNAs, 22 showed an increase in mature miRNA products, representing more than 57% of NF90-

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

associated pri-miRNAs, while only two were decreased (Supplementary Tables S6 and S7). Thus, we identified a subset of 22 pri-miRNAs that are bound by NF90 and whose abundance is increased following loss of NF90, which we named 'double-positive' pri-miRNAs. Both pri-miR-7-1 and pri-miR-3173 were identified within the double positive subset. Thus, NF90 downregulates the expression of most of its target pri-miRNAs.

Gene ontology of validated mRNA targets of double positive miRNAs revealed an implication particularly in cancer and infection by viruses, such as Epstein Barr Virus (EBV), hepatitis B virus (HBV), and human T lymphoma virus type 1 (HTLV1), as well as viral carcinogenesis (Supplementary Figure S5). This result is interesting given that NF90 translocates from the nucleus to the cytoplasm following viral infection of cells (28). Thus, viral infection could result in the coordinated processing of the NF90-modulated subset of pri-miRNAs, whose target mRNAs are implicated in viral replication. Interestingly, several miRNAs upregulated following loss of NF90 in this study have been shown to target RNAs expressed by influenza A virus subtypes. For instance, miR-3682 is involved in viral replication by targeting the NS gene of pH1N1 and H3N2 subtypes (29). Similarly, miR-4753 and miR-3145, which target PS and PB1 genes of H5N1 and H3N2 subtypes, are overexpressed in response to viral infection and inhibit viral transcription and replication (30).

We wondered whether pri-miRNAs that are associated with NF90 and downregulated upon its loss might share a common characteristic that would make them targets for NF90 binding. A MEME search did not reveal a simple binding motif common to the 22 pri-miRNA sequences. Compared to all human pri-miRNAs, the subset of 22 double-positive pri-miRNAs did not show any significant difference in their overall length (mean = 82.5 nt compared to 81.88 nt) or in the size of the terminal loop (mean = 7.87 nt compared to 7.92 nt) (Figure 5A). In contrast, however, the minimal stretch containing a mismatch $\leq$1 nt was significantly longer for double-positive pri-miRNAs compared to all pri-miRNAs, with a mean of 27.68 nt for double-positive pri-miRNAs compared to 21.11 nt for all pri-miRNAs (Figure 5A). This analysis suggests that double-positive pri-miRNAs might be more stable, having a longer duplex and less bulges compared to all human pri-miRNAs. To further investigate this possibility, we compared the free energy of the 22 double-positive pri-miRNAs compared to all pri-miRNAs. The 22 double-positive pri-miRNAs had a lower free energy (mean = −42.26) compared to all pri-miRNAs (mean = −38.19), as shown in Figure 5B. Taken together, these data suggest that double positive pri-miRNAs are more stable and have less mismatches than all pri-miRNAs. Predicted folding of double-positive pri-miRNA sequences also revealed highly stable structures with very few bulges, compared to pri-miR-200a, which is not highly associated with NF90 (Supplementary Figure S6A).

To test the idea that NF90 can bind to pri-miRNAs that have a stable structure with few bulges, we designed mutations within NF90-binding pri-miRNAs predicted to reduce stability and form bulge-like regions that might disrupt NF90 association. For each of the NF90-associated pri-miRNAs tested, we designed two mutant structures that

A

Feature (nt)	All pri-miRNAs	Double positive pri-miRNAs
Length	81.88	82.5
Terminal loop size	7.92	7.87
Longest duplex	21.11	27.68

B

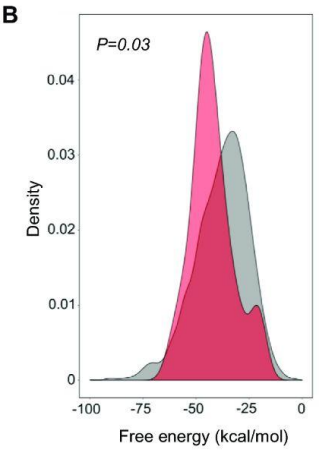


Figure 5. NF90 associates with a subset of highly stable pri-miRNAs. (**A**) Structural characteristics of all human pri-miRNAs and NF90 double positive pri-miRNAs. (**B**) Graph showing the free energy of all pri-miRNAs (grey) and NF90 double positive pri-miRNAs (red).

would be less stable than wild-type structures. (Figure 6A). WT and mutated pri-miRNAs were tested for NF90 association by RNA EMSA. As shown in Figure 6B and Supplementary Figure S6B, mutation of pri-miR-3173 or pri-miR-186 to less stable structures diminished NF90 binding. On the other hand, mutation of pri-miR-200a to a more stable structure enhanced NF90 binding. These data suggest that NF90 shows a preference for association with stable pri-miRNA hairpin structures having few bulge regions.

We then wondered whether pri-miRNAs whose mature products increased following loss of NF90, but were not considered eCLIP-positive using the applied cut-offs, might share the characteristics identified for double-positive pri-miRNAs. We therefore calculated the longest duplex length, allowing a mismatch of 1 nt, for the group of 181 upregulated but eCLIP negative pri-miRNAs, and 124 downregulated pri-miRNAs, as well as for those falling outside these groups (other) (Figure 7A). Interestingly, pri-miRNAs upregulated after loss of NF90 and eCLIP negative have a significantly longer duplex than all pri-miRNAs or other pri-miRNAs. Indeed, the duplex length is similar to that observed for the double positive group. In contrast, pri-

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

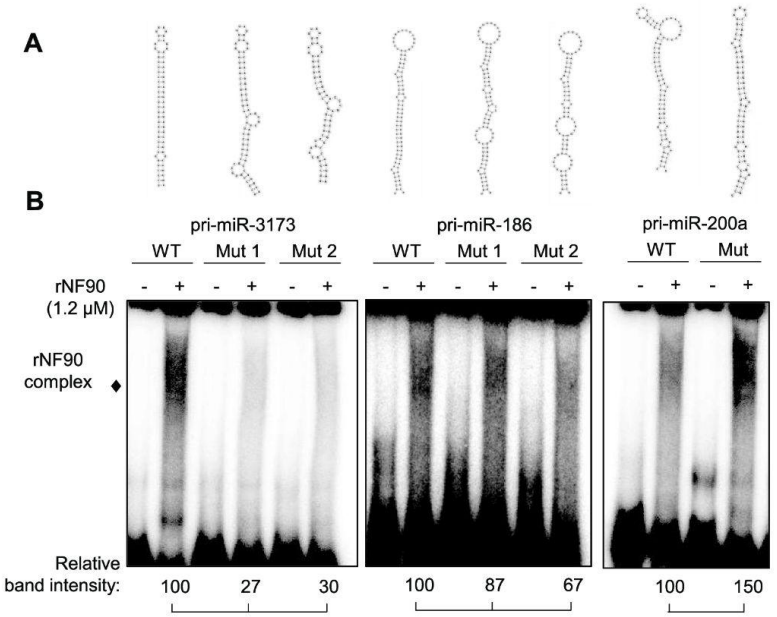


Figure 6. Modification of pri-miRNA structure alters NF90 binding. (**A**) Representations of wt or mutant pri-miRNAs sequences, as indicated. (**B**) RNA EMSA performed using recombinant NF90 and probed with radiolabelled pri-miRNAs as indicated. rNF90-pri-miRNA complexes are indicated on the figure. Relative band intensities (normalized to signal for wt) are shown below.

miRNAs downregulated upon loss of NF90 have a shorter duplex compared to all pri-miRNAs or other pri-miRNAs. We then calculated the mean free energy for the upregulated, eCLIP-negative group and the downregulated group of pri-miRNAs (Figure 7B). Similarly, when compared to all pri-miRNAs, the upregulated, eCLIP-negative group of pri-miRNAs had a significantly lower free energy. Free energy of the downregulated group was similar to that of all pri-miRNAs. In contrast, terminal loop size was comparable between the two groups; 7.86 nt (downregulated group) compared with 8.64 nt (upregulated eCLIP-negative group). Of note, total pri-miRNA length was higher for the upregulated eCLIP-negative group (87.01 nt) compared to the downregulated group (77.79 nt). These analyses suggest that upregulated, eCLIP-negative pri-miRNAs share some characteristics with double-positive pri-miRNAs. It is feasible that some NF90-associated pri-miRNAs were not detected by eCLIP analysis or did not pass the selection criteria used to identify eCLIP-positive pri-miRNAs. To test this idea, we selected two pri-miRNAs, pri-miR-4755 and pri-miR-4766, from the upregulated, eCLIP-negative group whose structure corresponds to the defined criteria for NF90 association, that is, having low free energy and few mismatches (Supplementary Figure S7A). NF90 binding to the pri-miRNAs was tested by RNA EMSA (Figure 7C and Supplementary Figure S7B). Indeed, both pri-miR-4755 and pri-miR-4766 were found to be significantly associated with NF90.

NF90 modulates the expression of a subset of genes hosting NF90-associated pri-miRNAs

Approximately 70% of human miRNAs are located in an intron of a host gene. Out of 22 double-positive pri-miRNAs, 20 are exclusively intronic. Two double-positive pri-miRNAs are found in either the 3′ UTR or an intron depending on transcript usage (Supplementary Table S6).

To determine whether loss of NF90 also affected the expression or splicing efficiency of the host genes, we performed RNA-seq in HepG2 cells transfected with control siRNA or siRNA targeting NF90. Loss of NF90 significantly diminished expression of three genes containing NF90-associated pri-miRNA; growth differentiation factor 15 (GDF15) hosting pri-miR-3189, 1-acylglycerol-3-phosphate *O*-acyltransferase 5 (AGPAT5) hosting pri-miR-4659a and zinc finger RAN-binding domain containing two (ZRANB2) hosting pri-miR-186 (Figure 8A). Furthermore, the splicing efficiency of introns containing pri-miRNAs downregulated by loss of NF90 was determined by RT-PCR for several targets (Figure 8B). Splicing efficiency was diminished for three pre-mRNAs containing NF90-associated pri-miRNAs: T-cell lymphoma invasion

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

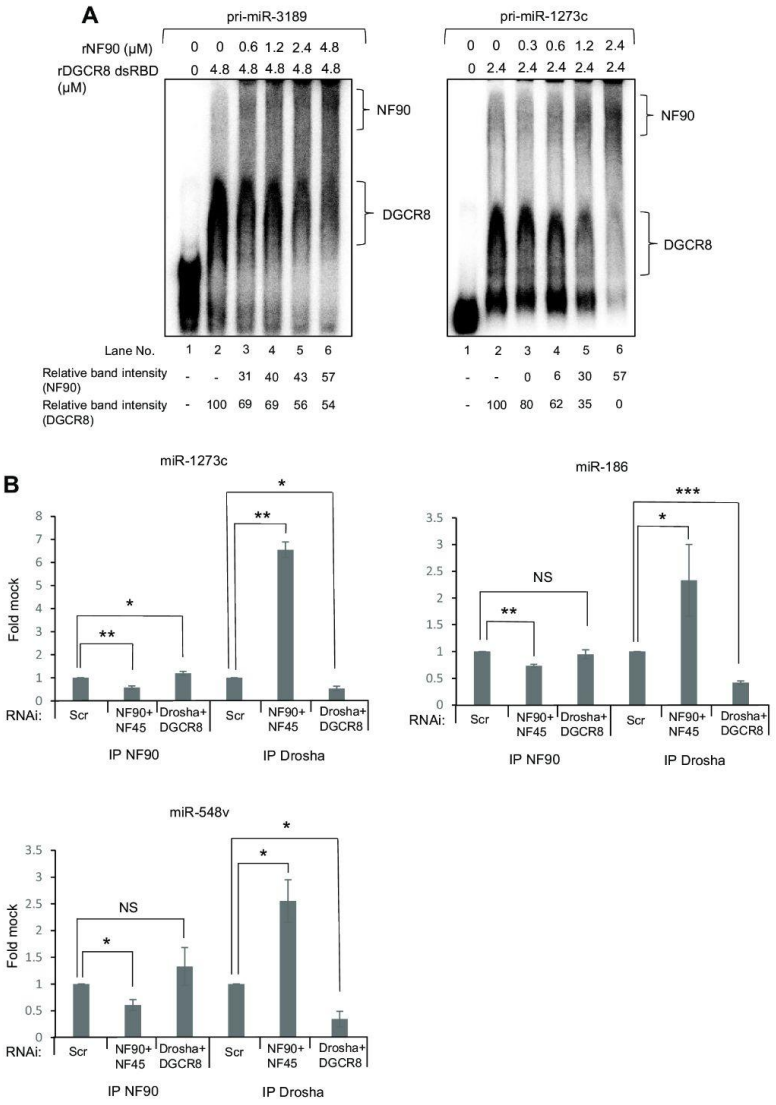


Figure 4. NF90 competes with the Microprocessor for the binding to pri-miRNAs. (**A**) RNA EMSA carried out using rDGCR8 dsRBD either alone or together with increasing amounts of rNF90 and probed with radio-labelled pri-miR-3189 or pri-miR-1273c. (**B**) Immunoprecipitates obtained using anti-NF90, anti-Drosha or a control antibody were analyzed by RT-qPCR amplifying a region proximal to the pri-miRNAs. The fold change relative to the control antibody sample was calculated and results are presented relative to the control sample (siScr), which was attributed a value of 1. Data represent mean ± SEM obtained from three independent experiments (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, NS 'Not Significant'. independent Student's t test).

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

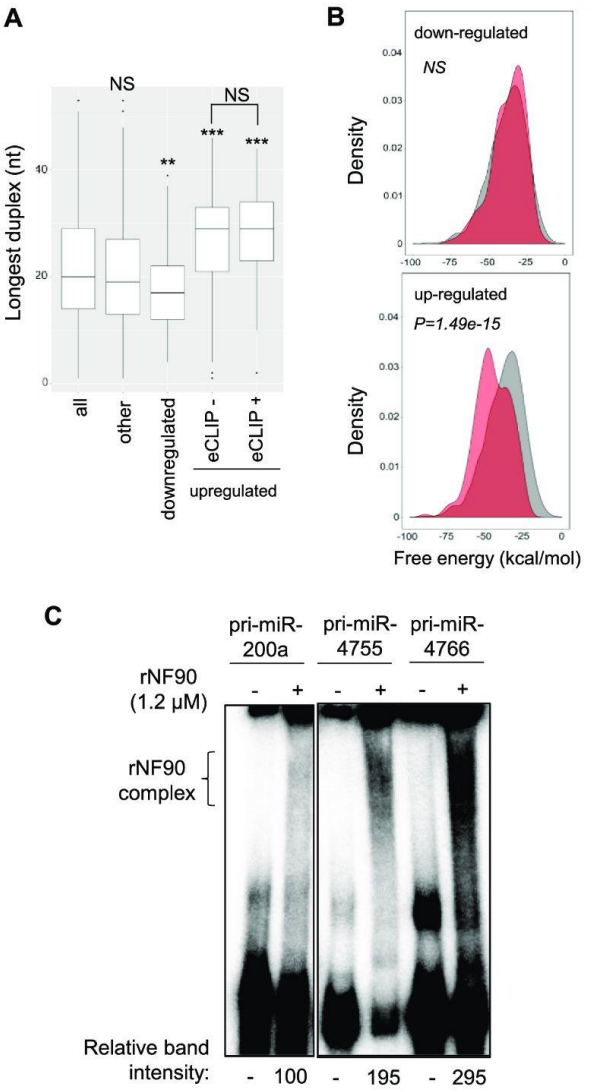


Figure 7. Pri-miRNAs whose mature products are upregulated following loss of NF90 share a similar structure. (**A**) Box plot representation of the longest duplex length of pri-miRNAs sorted into the indicated categories (* $P < 0.05$, *** $P < 0.001$, NS, not significant, Wilcoxon test). (**B**) Graphical representation of the free energy of pri-miRNAs whose mature products are downregulated or upregulated as indicated following loss of NF90 (red) compared to all pri-miRNAs (gray). (**C**) RNA EMSA performed using recombinant NF90 and probed with radiolabelled pri-miRNAs as indicated. rNF90-pri-miRNA complexes are indicated on the figure. Relative band intensities (normalized to pri-miR200a) are shown below.

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

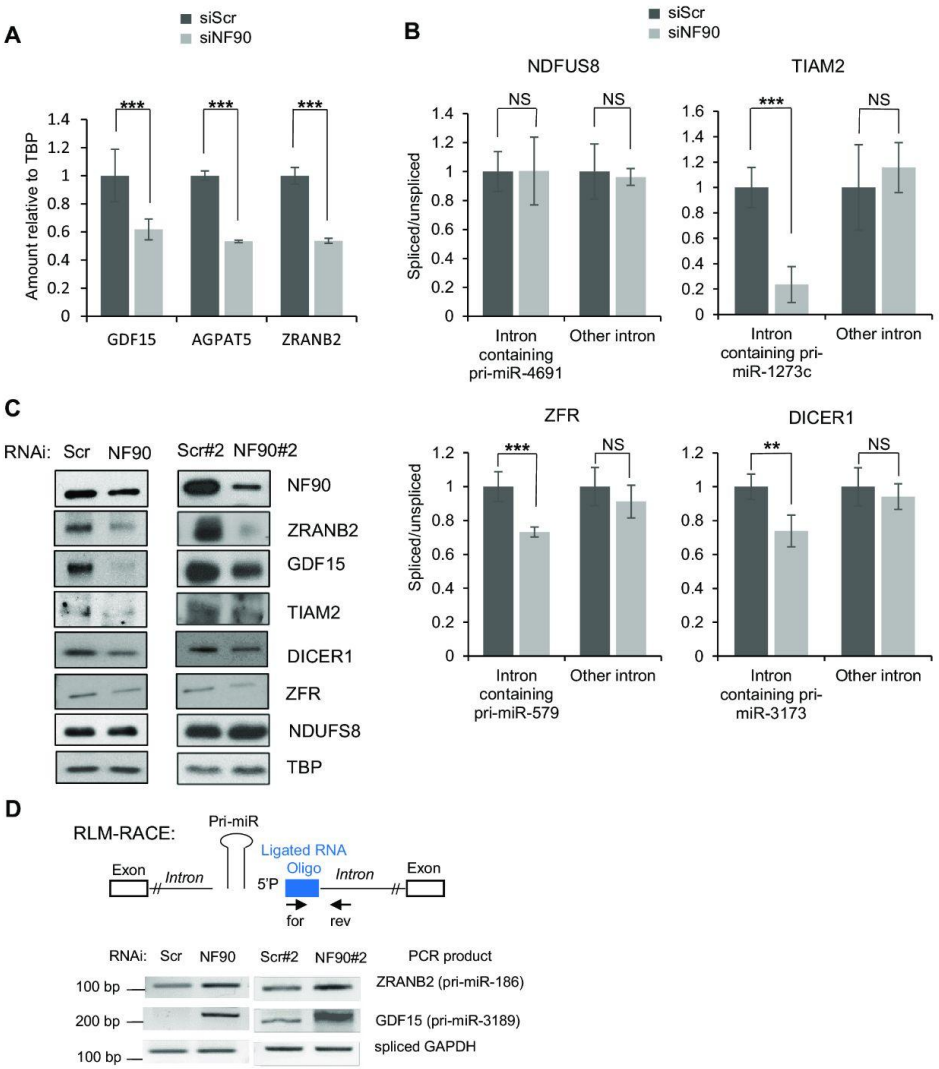


Figure 8. NF90 impacts expression of genes hosting pri-miRNAs. (**A**) Extracts of HepG2 cells transfected with siRNA targeting NF90 or a non-targeting control (Scr) as indicated were analyzed by RNA-seq and DESeq2. Data represent mean ± SEM obtained from three independent samples (***$P <$ 0.001, independent Student's *t* test). (**B**) The abundance of exon-intron junctions and exon-exon junctions in samples described in A was measured by RT-qPCR using PCR primers amplifying spliced or unspliced transcripts including introns containing pri-miRNAs or other introns. The splicing efficiency was calculated by the ratio of spliced to unspliced transcripts. Values obtained for the control sample (siScr) were attributed a value of 1. NS indicates 'Not Significant'. The graphs represent the mean ± SEM obtained from three or more independent experiments (**$P < 0.01$, ***$P < 0.001$, independent Student's *t* test). (**C**) Extracts of HepG2 cells transfected with siRNA targeting NF90 (NF90, NF90#2) or a non-targeting control (Scr, Scr#2) as indicated were analyzed by immunoblot using the antibodies indicated. (**D**) NF90 modulates transcript cleavage at the region containing miRNA. Extracts of HepG2 cells transfected with siRNAs targeting NF90 (NF90, NF90#2) or non-targeting controls (Scr, Scr#2) as indicated were analyzed by modified 5′ RLM-RACE. Forward and reverse primers used, and the predicted sizes of the PCR products are indicated.

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

and metastasis 2 (TIAM2), hosting pri-miR-1273c, Zinc Finger RNA binding protein (ZFR), hosting pri-miR-579, and DICER1, hosting pri-miR-3173 (Figure 8B). Interestingly, the splicing defect was detected for the intron containing the pri-miRNA but not for another intron within the same transcript (Figure 8B). In contrast, no significant effect was observed for NDUFS8, which hosts pri-miR-7113 and pri-miR-4691 that are not bound by NF90 and whose abundance are not affected by NF90 (Figure 8B).

The expression of these genes was analysed by western blot of extracts obtained from HepG2 cells transfected with control (Scr and Scr#2) and NF90-targeting (NF90 and NF90#2) siRNAs. All genes tested showed diminished expression upon loss of NF90, except NDUFS8 that showed no significant difference in expression (Figure 8C). Thus, NF90 modulates the expression of certain pri-miRNA host genes, including TIAM2, a known oncogene and metastasis factor in HCC (31,32).

Finally, to determine whether loss of gene expression correlated with increased pri-miRNA cropping following knock down of NF90, we performed modified RLM-5′ RACE as described previously (19), using extracts of cells transfected with control (Scr and Scr#2) and NF90-targeting (NF90 and NF90#2) siRNAs. Indeed, RLM RACE analysis showed enhanced cleavage of the intronic region of ZRANB2 hosting pri-miR-186 and GDF15 hosting pri-miR-3189 in extracts of NF90 knock down cells compared to controls (Figure 8D). This analysis indicates that loss of NF90 enhances transcript cleavage in the vicinity of the hosted pri-miRNA.

DISCUSSION

We and others have previously shown that NF90 can inhibit the processing of certain miRNA precursors (14,15,19). However, it was unclear how widespread the impact of NF90 might be on human miRNA biogenesis. Here, we have used genome-wide approaches to address the effect of NF90 on the miRNA pool in HepG2 HCC cells. Our data indicate that NF90 modulates the processing of a specific subset of miRNA precursors. NF90 is associated with at least 38 human pri-miRNAs, as indicated by analysis of eCLIP data obtained by Nussbacher and Yeo (24). Of these, 22 showed increased abundance of mature miRNA products following knock-down of NF90. Thus, association of NF90 with a pri-miRNA is likely to influence its fate. Most NF90-associated pri-miRNAs did not overlap with those bound by either DGCR8 or Drosha. Moreover, results obtained by RNA-EMSA support the idea that NF90 and Microprocessor may compete for the binding of the subset of pri-miRNAs, at least *in vitro*. Further analysis will be required to determine whether the competition also occurs *in vivo*. Of note, RIP analysis showed that loss of NF90/NF45 complex led to increased binding of Drosha at pri-miRNAs that were highly bound by NF90 in control conditions. Conversely, loss of Microprocessor increased binding by NF90 to pri-miRNAs that were not highly bound by NF90 in wild-type cells. Interestingly, for those pri-miRNAs that were bound by both NF90 and DGCR8, the binding profiles of the two factors were largely complementary. Furthermore, while the binding profile of DGCR8 was not noticeably different for this group compared to all pri-miRNAs bound by DGCR8, the binding profile of NF90 differed somewhat for this group compared to all pri-miRNAs bound by NF90. This could suggest that NF90 and DGCR8 might bind simultaneously to the pri-miRNA, and that the binding of DGCR8 may alter the binding mode of NF90 for such pri-miRNAs.

Since NF90 is a highly abundant and ubiquitously expressed protein, it might be expected that NF90-associated pri-miRNAs would be poorly processed in most cells. Indeed, the mature miRNA products of NF90 bound pri-miRNAs are very poorly expressed, or not expressed at all in control cells. They become readily detectable only upon loss of NF90. An exception is pri-miR-7-1, although interestingly, this miRNA shows tissue specific expression, being highly expressed only in brain and pancreas (33).

Our data suggests that pri-miRNAs upregulated after loss of NF90 share a common structure that might facilitate NF90 association with the stem region. This finding is consistent with a previous report showing structure-based recognition of adenovirus-expressed VA1 RNA by NF90 (34). Extensive mutational analysis of VA1 association with NF90 showed no specificity for nucleotide sequence but rather the requirement for a minihelix structure within the stem region. The pri-miRNAs identified in this study also exhibit a minihelix-like structure that appears to be necessary for NF90 binding. Indeed, RNA EMSA showed that NF90 association with pri-miR-3173 and pri-miR-186 could be diminished by introducing destabilizing mutations, while NF90 association could be acquired by increasing the stability of the stem region, as for pri-miR-200a.

Interestingly, our data predict that the subset of NF90-associated pri-miRNAs may extend beyond those detected by eCLIP analysis. Using the characteristics determined from the eCLIP-positive, upregulated pri-miRNA group, that is duplex length and free energy, we found that pri-miRNAs whose mature products were upregulated following loss of NF90 but were not positive by eCLIP analysis shared the same characteristics as the double positive group. The length of the duplex region and the free energy of the structure was comparable to that of double positive pri-miRNAs. RNA EMSA confirmed the predicted association with NF90 for two of these pri-miRNAs. Interestingly, both groups were significantly different to all pri-miRNAs or those that are unaffected by NF90 (other). Thus, it appears that the high specificity of eCLIP revealed a subset of pri-miRNAs that share a common structure. When this information was used to interrogate the group of pri-miRNAs who share the same biological response to loss of NF90, that is, upregulation of their mature products, we observed that both groups share the same characteristics. We predict that a certain number of the upregulated group likely do bind to NF90 but may escape detection by eCLIP. For example, as noted above, many of the pri-miRNAs are expressed at extremely low levels in control cells, which could make their association with NF90 difficult to detect.

Interestingly, pri-miR-7-1 processing has been shown to be influenced by another RBP, HuR, which recruits MSI2 to the terminal loop. Binding of HuR/MSI2 was found to

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

stabilize the stem region and led to diminished processing by microprocessor (35). It would be interesting to determine whether binding of HuR/MSI2 to pri-miR-7-1 might facilitate NF90 binding to the stem region, and compete with microprocessor. Similarly, it would be interesting to determine whether HuR/MSI2 can bind the terminal loop of other NF90-modulated pri-miRNAs in addition to pri-miR-7-1. NF90 may cooperate with other RBPs, such as HuR/MSI2 to control the processing of a subset of pri-miRNAs.

Another feature that the subset of NF90-modulated pri-miRNAs share is their restriction to human or primate lineages. Again, pri-miR-7-1 is an exception, being highly conserved throughout evolution. Thus, given that the subset of NF90-modulated pri-miRNAs are young and almost perfect hairpins, it is tempting to speculate that this group may have originated through recent insertion of repeat elements in the genome.

Interestingly, GO analysis of validated mRNA targets of the mature miRNAs showed significant enrichment for infection by viruses such as Epstein Barr Virus (EBV), hepatitis B virus (HBV) and human T lymphoma virus type 1 (HTLV1) and in viral carcinogenesis. Indeed, viral infection of cells induces translocation of NF90 from the nucleus to the cytoplasm (28). Thus, it is conceivable that pathological conditions such as viral infection could result in the coordinated processing of the NF90-modulated subset of pri-miRNAs, which target mRNAs important for viral replication.

Finally, transcriptomic analysis showed that association of NF90 with pri-miRNAs may diminish the expression of certain host genes, as described previously (19). Among the pri-miRNA-hosting transcripts that are downregulated after loss of NF90, two are noteworthy. The expression of TIAM2, hosting pri-miR-1273C, is down-regulated upon loss of NF90. TIAM2 is a known oncogene and metastasis factor in HCC (31,32). Levels of NF90 are elevated in HCC (14,20) and it would be interesting to determine whether NF90-dependent modulation of TIAM2 might contribute to pathogenesis. Loss of NF90 also diminished expression of growth differentiation factor 15 (GDF15), hosting pri-miR-3189. GDF15 is expressed and secreted by a limited number of tissues, including liver. When complexed with its receptor, GFRAL, in brain and CNS, GDF15 supresses appetite (see (36) for review). Cancer patients express high circulating levels of GDF15, which contributes to anorexia/cachexia. On the other hand, enhancement of GDF15 expression is a promising therapeutic strategy in the treatment of obesity. It would be interesting to determine whether high levels of NF90 in HCC may have a role in promoting expression of GDF15 from liver cells in cancer patients.

In summary, we have identified a subset of human pri-miRNAs that are bound by NF90. Analysis indicates that this subset shares a similar structure that appears to be favorable for NF90 binding. These data extend our knowledge of how processing of pri-miRNAs can be modulated by RBPs. This may be beneficial for understanding perturbations of miRNA levels in pathological conditions and could also open up novel treatment strategies using nanotherapeutics.

DATA AVAILABILITY

Small RNA-seq and RNA-seq data have been deposited at GEO (GSE132341).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We wish to thank Catherine Dargemont and Xavier Contreras for critical reading of the manuscript, and the Gene Regulation lab and Hervé Seitz for helpful discussions.

FUNDING

European Research Council [RNAmedTGS to R.K.]; MSD Avenir [HideInflame&Seq to R.K.]; Ministère de l'Enseignement Supérieur et de la Recherche et de l'Innovation scholarship (to G.G.); Japan Society for the Promotion of Science (Grant-in-aid for Young Scientists (B)) [17K15601, 19K16523 to T.H.]; Grant-in-aid for Scientific Research (C) [16K08590, 19K07370 to S.S.]. Funding for open access charge: [ERC, RNA MedTGS].
Conflict of interest statement. None declared.

REFERENCES

1. Ebert,M.S. and Sharp,P.A. (2012) Roles for microRNAs in conferring robustness to biological processes. *Cell*, **149**, 515–524.
2. Shenoy,A. and Blelloch,R.H. (2014) Regulation of microRNA function in somatic stem cell proliferation and differentiation. *Nat. Rev. Mol. Cell. Biol.*, **15**, 565–576.
3. Denli,A.M., Tops,B.B., Plasterk,R.H., Ketting,R.F. and Hannon,G.J. (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature*, **432**, 231–235.
4. Gregory,R.I., Yan,K.P., Amuthan,G., Chendrimada,T., Doratotaj,B., Cooch,N. and Shiekhattar,R. (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature*, **432**, 235–240.
5. Ha,M. and Kim,V.N. (2014) Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.*, **15**, 509–524.
6. Finnegan,E.F. and Pasquinelli,A.E. (2013) MicroRNA biogenesis: regulating the regulators. *Crit Rev. Biochem. Mol. Biol.*, **48**, 51–68.
7. Krol,J., Loedige,I. and Filipowicz,W. (2010) The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.*, **11**, 597–610.
8. Mendell,J.T. and Olson,E.N. (2012) MicroRNAs in stress signaling and human disease. *Cell*, **148**, 1172–1187.
9. Winter,J., Jung,S., Keller,S., Gregory,R.I. and Diederichs,S. (2009) Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat. Cell Biol.*, **11**, 228–234.
10. Viswanathan,S.R., Daley,G.Q. and Gregory,R.I. (2008) Selective blockade of microRNA processing by Lin28. *Science*, **320**, 97–100.
11. Guil,S. and Caceres,J.F. (2007) The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nat. Struct. Mol. Biol.*, **14**, 591–596.
12. Michlewski,G., Guil,S., Semple,C.A. and Caceres,J.F. (2008) Posttranscriptional regulation of miRNAs harboring conserved terminal loops. *Mol. Cell*, **32**, 383–393.
13. Michlewski,G. and Caceres,J.F. (2019) Post-transcriptional control of miRNA biogenesis. *RNA*, **25**, 1–16.
14. Higuchi,T., Todaka,H., Sugiyama,Y., Ono,M., Tamaki,N., Hatano,E., Takezaki,Y., Hanazaki,K., Miwa,T., Lai,S. *et al.* (2016) Suppression of MicroRNA-7 (miR-7) biogenesis by nuclear factor 90-Nuclear factor 45 complex (NF90-NF45) controls cell proliferation in hepatocellular carcinoma. *J. Biol. Chem.*, **291**, 21074–21084.

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

15. Sakamoto,S., Aoki,K., Higuchi,T., Todaka,H., Morisawa,K., Tamaki,N., Hatano,E., Fukushima,A., Taniguchi,T. and Agata,Y. (2009) The NF90-NF45 complex functions as a negative regulator in the microRNA processing pathway. *Mol. Cell. Biol.*, **29**, 3754–3769.
16. Heale,B.S., Keegan,L.P., McGurk,L., Michlewski,G., Brindle,J., Stanton,C.M., Caceres,J.F. and O'Connell,M.A. (2009) Editing independent effects of ADARs on the miRNA/siRNA pathways. *EMBO J.*, **28**, 3145–3156.
17. Yang,W., Chendrimada,T.P., Wang,Q., Higuchi,M., Seeburg,P.H., Shiekhattar,R. and Nishikura,K. (2006) Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat. Struct. Mol. Biol.*, **13**, 13–21.
18. Jayachandran,U., Grey,H. and Cook,A.G. (2016) Nuclear factor 90 uses an ADAR2-like binding mode to recognize specific bases in dsRNA. *Nucleic Acids Res.*, **44**, 1924–1936.
19. Barbier,J., Chen,X., Sanchez,G., Cai,M., Helsmoortel,M., Higuchi,T., Giraud,P., Contreras,X., Yuan,G., Feng,Z. *et al.* (2018) An NF90/NF110-mediated feedback amplification loop regulates dicer expression and controls ovarian carcinoma progression. *Cell Res.*, **28**, 556–571.
20. Jiang,W., Huang,H., Ding,L., Zhu,P., Saiyin,H., Ji,G., Zuo,J., Han,D., Pan,Y., Ding,D *et al.*2015) Regulation of cell cycle of hepatocellular carcinoma by NF90 through modulation of cyclin E1 mRNA stability. *Oncogene*, **34**, 4460–4470.
21. Seco-Cervera,M., Gonzalez-Rodriguez,D., Ibanez-Cabellos,J.S., Peiro-Chova,L., Pallardo,F.V. and Garcia-Gimenez,J.L. (2018) Small RNA-seq analysis of circulating miRNAs to identify phenotypic variability in Friedreich's ataxia patients. *Sci Data*, **5**, 180021.
22. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
23. Bennasser,Y., Chable-Bessia,C., Triboulet,R., Gibbings,D., Gwizdek,C., Dargemont,C., Kremer,E.J., Voinnet,O. and Benkirane,M. (2011) Competition for XPO5 binding between Dicer mRNA, pre-miRNA and viral RNA regulates human Dicer levels. *Nat. Struct. Mol. Biol.*, **18**, 323–327.
24. Nussbacher,J.K. and Yeo,G.W. (2018) Systematic discovery of RNA binding proteins that regulate MicroRNA Levels. *Mol. Cell*, **69**, 1005–1016.
25. Chou,C.H., Shrestha,S., Yang,C.D., Chang,N.W., Lin,Y.L., Liao,K.W., Huang,W.C., Sun,T.H., Tu,S.J., Lee,W.H *et al.*2018) miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **46**, D296–D302.
26. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
27. Nguyen,T.A., Jo,M.H., Choi,Y.G., Park,J., Kwon,S.C., Hohng,S., Kim,V.N. and Woo,J.S. (2015) Functional anatomy of the human Microprocessor. *Cell*, **161**, 1374–1387.
28. Li,X., Liu,C.X., Xue,W., Zhang,Y., Jiang,S., Yin,Q.F., Wei,J., Yao,R.W., Yang,L. and Chen,L.L. (2017) Coordinated circRNA biogenesis and function with NF90/NF110 in viral infection. *Mol. Cell*, **67**, 214–227.
29. Zheng,B., Zhou,J. and Wang,H. (2020) Host microRNAs and exosomes that modulate influenza virus infection. *Virus Res.*, **279**, 197885.
30. Khongnomnan,K., Makkoch,J., Poomipak,W., Poovorawan,Y. and Payungporn,S. (2015) Human miR-3145 inhibits influenza A viruses replication by targeting and silencing viral PB1 gene. *Exp. Biol. Med. (Maywood)*, **240**, 1630–1639.
31. Chen,J.S., Su,I.J., Leu,Y.W., Young,K.C. and Sun,H.S. (2012) Expression of T-cell lymphoma invasion and metastasis 2 (TIAM2) promotes proliferation and invasion of liver cancer. *Int. J. Cancer*, **130**, 1302–1313.
32. Yen,W.H., Ke,W.S., Hung,J.J., Chen,T.M., Chen,J.S. and Sun,H.S. (2016) Sp1-mediated ectopic expression of T-cell lymphoma invasion and metastasis 2 in hepatocellular carcinoma. *Cancer Med.*, **5**, 465–477.
33. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
34. Gwizdek,C., Ossareh-Nazari,B., Brownawell,A.M., Evers,S., Macara,I.G. and Dargemont,C. (2004) Minihelix-containing RNAs mediate exportin-5-dependent nuclear export of the double-stranded RNA-binding protein ILF3. *J. Biol. Chem.*, **279**, 884–891.
35. Choudhury,N.R., de Lima Alves,F., de Andres-Aguayo,L., Graf,T., Caceres,J.F., Rappsilber,J. and Michlewski,G. (2013) Tissue-specific control of brain-enriched miR-7 biogenesis. *Genes Dev.*, **27**, 24–38.
36. Tsai,V.W.W., Husaini,Y., Sainsbury,A., Brown,D.A. and Breit,S.N. (2018) The MIC-1/GDF15-GFRAL pathway in energy homeostasis: implications for obesity, cachexia, and other associated diseases. *Cell Metab.*, **28**, 353–368.

Downloaded from https://academic.oup.com/nar/article/48/12/6874/5840576 by INIST-CNRS BiblioVie user on 13 July 2021

Translesion DNA synthesis-driven mutagenesis in very early embryogenesis of fast cleaving embryos

During my second and third year of PhD I helped E.L.F. to analyse the WGS data for her project. The paper is currently under revision in Nucleic Acid Research.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

In early embryogenesis of fast cleaving embryos DNA synthesis is short and surveillance mechanisms preserving genome integrity are inefficient implying the possible generation of mutations. We have analyzed mutagenesis in *Xenopus laevis* and *Drosophila melanogaster* early embryos. We report the occurrence of a high mutation rate in *Xenopus* and show that it is dependent upon the translesion DNA synthesis (TLS) master regulator Rad18. Unexpectedly, we observed a homology-directed repair contribution of Rad18 in reducing the mutation load. Genetic invalidation of TLS in the pre-blastoderm *Drosophila* embryo resulted in reduction of both the hatching rate and Single Nucleotide Variations on specific chromosome regions in adult flies. Altogether, these findings indicate that during very early *Xenopus* and *Drosophila* embryos TLS strongly contributes to the high mutation rate. This may constitute a previously unforeseen source of genetic diversity contributing to the polymorphisms of each individual with implications for genome evolution and species adaptation.

Keywords: *Xenopus, Drosophila*, ubiquitin, nucleus, chromatin, PCNA

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Translesion DNA synthesis-driven mutagenesis in very early embryogenesis of fast cleaving embryos

Elena Lo Furno[1], Isabelle Busseau[2], Claudio Lorenzi[3], Cima Saghira[4], Matt C Danzi[4], Stephan Zuchner[4] and Domenico Maiorano[1*]

[1] Institut de Génétique Humaine, Université de Montpellier, Genome Surveillance and Stability laboratory. CNRS-UMR9002, 34000 Montpellier, France.

[2] Institut de Génétique Humaine, Université de Montpellier, Systemic impact des of small regulatory RNAs laboratory. CNRS-UMR9002, 34000 Montpellier, France.

[3] Institut de Génétique Humaine, Université de Montpellier, Machine learning and gene regulation laboratory.CNRS-UMR9002, 34000 Montpellier, France.

[4] Department of Human Genetics, Hussman Institute for Human Genomics, University of Miami (Florida, USA).

*Correspondance

Email: domenico.maiorano@igh.cnrs.fr

Tel.: +33(0)4 34 35 99 46

Fax: +33(0)4 34 35 99 01

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Materials and Methods

Experiments with *Xenopus* were performed in accordance with current institutional and national regulations approved by the Minister of Research under supervision of the Departmental Direction of Population Protection (DDPP). *Xenopus* embryos were prepared by *in vitro* fertilization as previously described (*10*). Two-cell stage embryos were microinjected in the animal pole using a Nanoject auto oocyte injector under a stereomicroscope (2 injections of 9 nL^{-1} in one blastomer). Each embryo was injected with 12 ng (pre-MBT) or 72 ng (post-MBT) of supercoiled plasmid, undamaged or irradiated with 200 J/m^2 of UV-C with a UV-Stratalinker, and/or 5 ng of RAD18 mRNAs. Embryos were collected at 16- or 32-cell stage, according to Nieuwkoop and Faber normal tables, snap frozen in liquid nitrogen and stored at -80 °C.

Drosophila melanogaster stocks were maintained and experiments were performed following standard procedures on standard cornmeal-yeast medium, inside a thermostatic room at 25 °C with alternating light and dark for an equal amount of hours per day. The following stocks were from Bloomington *Drosophila* Stock Center : *dPolηExc2.15* (#57341), OreRmE (#25211). The latter was homogeneized through serial individual backcrosses for nine generations. $dPol\eta^{12}$ was a kind gift of Benjamin Loppin (*23*). Balancer stocks were from our laboratory. Quantifications of hatching rates (eggs to larvae) were determined as previously described (*24*). Hatch rate is the ratio of hatched eggs to total eggs laid expressed as a percentage. Two hundred or more embryos were scored twice per genotype. In addition, hatching of adult flies was estimated by calculating the percentage of larvae (counted 2-2,5 days after fertilization) developed to mature flies (counted 10 days after fertilization).

Plasmid DNAs

lacZ-containing plasmid (pEL1) was obtained by subcloning the *lac* operon from pBluescript into the *Spe*I-*Kpn*I restriction sites of pRU1103 vector, which contains full-length *lacZ*. pEL1 was transformed and amplified in *E. coli* and purified using a standard protocol (QIAGEN) at a temperature lower or equal to 12 °C to obtain a near 100 % supercoiled DNA, as previously described (*25*). This procedure greatly minimizes DNA damage and background mutations. pCS2-MLH1 plasmid was obtained by subcloning human MLH1 cDNA from pCEP9MLH1 (*72*)), into the *BamH*I-*Xho*I restriction sites of pCS2 vector. Rad18 wild-type, C28F and C207F mutant plasmids were previously described (*10*). The $Rad18^{C28FC207}$ double mutant was generated by standard site-directed mutagenesis from the $Rad18^{C28F}$ mutant plasmid.

In vitro transcription

mRNA synthesis was performed with mMESSAGE mMACHINE kit SP6® (AM1340, Thermofisher). mRNAs were recovered by phenol-chloroform extraction and isopropanol precipitation. Following centrifugation and ethanol wash, mRNAs were dissolved in 20 μL^{-1} of RNase-free water. mRNAs quality was checked by formaldehyde gel electrophoresis.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Xenopus embryos and eggs protein extracts

An average of 20 embryos were lysed in Xb buffer (5 μL^{-1} of buffer per embryo; 100 mM KCl, 0.1 mM $CaCl_2$, 1 mM $MgCl_2$, 50 mM sucrose, 10 mM HEPES pH 7.7) supplemented with cytochalasin (10 $\mu g\ mL^{-1}$), phosphatases (PhosSTOP 1X,) and proteases inhibitors (5 $\mu g\ mL^{-1}$ Leupeptin, Pepstatin A and Aprotinin). After 10 min centrifugation at maximum speed in a benchtop centrifuge at 4 °C, cytoplasmic fraction was recovered, neutralized in an equal volume of Laemmli buffer 2X and boiled at 95 °C for 5 min. Embryos lysates were loaded on precast gradient gels (4-12 %, NuPAGE, Invitrogen). Gels were transferred to a nitrocellulose membrane for western blotting and incubated with the indicated antibodies. Interphasic *Xenopus* egg extracts were prepared and used as previously described (*10*).

Ribonucleotide incorporation assay

Upon thawing, *Xenopus* eggs extracts were supplemented with cycloheximide (250 $\mu g\ mL^{-1}$) and an energy regeneration system (1 mM ATP, 2 mM $MgCl_2$, 10 mM creatine kinase, 10 mM creatine phosphate). M13mp18 ssDNA was added as a template for DNA replication at the indicated concentrations in presence of α-^{32}PdCTP (3000Ci $mmol^{-1}$, Perkin Elmer). At the indicated time points half of the samples were neutralized in 10 mM EDTA, 0,5 % SDS, 200 $\mu g\ mL^{-1}$ Proteinase K and incubated at 52 ° C for 1 hour. Samples were treated with 0.3 M NaOH at 55°C for 2 hours to digest incorporated ribonucleotides in the plasmid and loaded on 5M urea 8% acrylamide gel TBE 0,5 X urea after formamide denaturation at 55 °C for 3 minutes. After migration, the gel was exposed to autoradiography.

Plasmid DNA isolation from embryos

Frozen embryos were crushed in STOP MIX supplemented with fresh Proteinase K (600 $\mu g\ \mu L^{-1}$). Embryos were homogenized with a tip in this solution while thawing. Immediately after, proteins digestion at 37 °C for 1 hour, total DNA was extracted as described above by phenol-chlorophorm extraction and ethanol precipitation. Recovered DNA was digested with *Dpn*I to destroy unreplicated plasmids and subsequently purified with QIAGEN gel extraction kit.

Somatic A6 cell culture

A6 epithelial cells were grown in modified Leibowitz L-15 medium containing 20 % sterile distilled water, 10 % foetal bovine serum and 100 U/ml penicillin/streptomycin at 25°C. A subcultivation ratio of 1:3 was employed. Cells were detached after a single wash with PBS by incubation with 0.25 % trypsin 0,03 % EDTA for 4 minutes at 37 °C. The day prior to transfection, 3 X 10^{-6} cells were seeded in 10 cm^2 dishes. One day later, cells were transfected with 60 µL Lipofectamine 2000 (Thermo Fisher Scientific) and 24 µg of plasmid following the manufacturer's recommendations.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Plasmid recovery from A6 cells

Cells were harvested by trypsinization and centrifugation 48 hours post transfection. After washing cell pellet with PBS, cells were crushed in lysis buffer (10 mM Tris pH 8,0; 100 mM NaCl; 10 mM EDTA pH 8,0; 0,5 % SDS) supplemented with fresh Proteinase K (600 µg μL^{-1}) by means of a tip (500 μL^{-1} per sample). Immediately after protein digestion at 37 °C, SDS was precipitated by adding half volume of saturated (6M) NaCl to each tube and centrifugation at 4°C for 10 minutes at 5000 rpm after 10 minutes incubation on ice. Total DNA was extracted from the supernatant as described above by phenol-chlorophorm extraction and ethanol precipitation. Recovered DNA was digested with *Dpn*I restriction enzyme (NEB) to destroy unreplicated plasmids and subsequently purified with QIAGEN gel extraction kit.

White/blue colonies selection and mutation frequency

DNA extracted from embryos was transformed in electrocompetent indicator bacteria (MBM7070 strain bearing an amber mutation in the *lacZ* gene) for white/blue screening and plated on selective petri dishes (40 µg mL^{-1} Xgal,; 200 µM IPTG). Over one thousand colonies were scored at least for each condition in each replicate. Plasmid DNA was isolated from mutant clones using a standard protocol (QIAGEN). After paired-end Sanger sequencing, polymorphisms were filtered for sequencing quality > 30 and analyzed on both strands using Geneious or Snapgene softwares. Mutation rates were estimated from the proportion of blue colonies observed (P_0). Before calculating the proportion of blue colonies observed (P_0), the basal percentage of white colonies prior to microinjection was subtracted from the percentage of white colonies in each experimental condition. The observed P_0 was substituted for P_0 to obtain the mutation rate (µ) using the following formula: $\mu=-\ln(P_0)$ and normalized to the number of cell cycles before embryo collection.

Antibodies

The following antibodies were used: Gapdh (ab9484, Abcam); $Pcna^{mUb}$ Lys 164 (13439, Cell Signaling Technology); PCNA (PC10, Sigma); XlRad18 (*10*); SMAUG (*27*); Tubulin (DM1A, Sigma), Mlh1 (ab14206 Abcam). The PC10 antibody cross-reacts with *Drosophila melanogaster* PCNA (*28*).

DAPI staining of Drosophila embryos

Embryos collection (0-2 hours unless otherwise indicated) was carried out using standard techniques (*29*). Embryos were dechorionated in 50% bleach and fixed by shaking in a mixture of PFA 4% in PBS and heptane (1:1) for 30 minutes and the aqueous layer containing formaldehyde was removed. Embryos were devitellinised upon washing in methanol-heptane mixture (1:1) and conserved in methanol at -20 °C overnight and for up to a week. Embryos

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

were rehydrated by sequential incubations of 10 minutes in Ethanol/PBS-T 7:3 (1X PBS + 0.1% Triton X-100), Ethanol/PBS-T 3:7 and PBS-T. Embryos were incubated for 30 minutes at room temperature in DAPI-PBS-T (1µg mL^{-1}) in the dark and rinsed three times in PBS-T. The third wash was performed overnight with mild shaking on a wheel in the dark at 4 °C. Samples were mounted in coverslips using Vectashield. Images were acquired with a Zeiss Axiovert Apotome microscope at 5X using Coolsnap HQ CDD camera (Photometrics) and processed using Omero 5.2.0 software. *P*-values were obtained using a two-tailed, unpaired Student's *t*-test.

Drosophila embryo protein lysates preparation

60 females and 10 males were incubated together inside embryo's collectors with embryo dishes for a certain number of hours according to the desired stage of embryos to be harvested. Collected embryos were gently rinsed off the medium with embryo collection buffer (Triton X-100 0,03 %; NaCl 68 mM). Embryos were removed from the medium using a brush and poured into a sieve (Falcon Cell Strainer 40 µM Nylon 352340). Harvested embryos were washed again and collected in a fresh tube (up to 50 μL^{-1} of embryos corresponding to 100 embryos). Laemmli 2X was added in ratio 1:1 in comparison to harvested volume and embryos were lysed by means of a pestle. After boiling embryo's mush at 95 °C for 5 min, chorion residues were removed by centrifugation with a benchtop centrifuge (max speed) at room temperature. Protein concentration was estimated by Amido Black staining using BSA of known concentration as a reference.

Genomic DNA extraction from single flies for Next Generation Illumina Sequencing.

Each fly was crushed with a pestle in a 1,5 mL^{-1} tube containing 170 µL of extraction buffer (Tris-HCl pH 8.0; 50mM; EDTA 50mM; SDS 1 %). Proteinase K (555 µg μL^{-1}) was added once the tissues had been completely grinded. After incubating 15 min at room temperature, cellular debris was removed twice by potassium acetate addition (0,83 M) and centrifugation with a bench-top centrifuge (max speed) at 4 °C for 10 min. DNA was isolated by double phenol/chloroform/isoamyl alcohol (25:24:1) extraction and ethanol precipitation overnight at -20 °C with glycogen (20 µg) followed by centrifugation at 4 °C. Precipitated DNA was washed with cold ethanol 70 %, dried at room temperature for 30 min and dissolved in water. The quality of extracted DNA (about 150 ng) was verified by agarose gel electrophoresis.

Illumina Next Generation Sequencing

Two genomic DNA samples per condition (extracted from single heterozygous *dpolη*$^{EXC215/+}$ males either dPolη maternally-depleted or maternally-provided, see text) were sequenced by Illumina NGS. After library construction and shotgun, whole *Drosophila* genomes were paired-end sequenced and assembled as previously described (*30*). Data were filtered for Genotype Quality > 35 and Depth > 10 before sequences alignment against the *Drosophila* reference genome (Flybase release 6). Variant calling was done using Freebayes software version 0.9.20.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Genotype ratio was not changed from recommended settings. Alignment was performed with BWA version 0.7.12-r1039. Variant calling was performed using Freebayes version 0.9.20 and then annotated with Ensembl VEP version 82. SNVs and Indels were then separated for downstream analyses. The threshold generally is above 33% to call an allele variant from the reference.

Statistics

Statistical analysis was performed using the Prism software (version 8). Means were compared using analysis of one-way ANOVA. Post-hoc tests were performed with a two-tailed unpaired Student's t test unless otherwise specified. Stars indicate significant differences * P< 0.05, ** P< 0.01, *** P <0.001, **** P <0.0001, "ns" denotes non-significant statistical test.

Results

High mutagenesis rate in the pre-MBT *Xenopus* embryo

We employed a classical *lacZ*-based reporter assay to measure mutagenesis in pre-MBT *Xenopus laevis* embryos. In this experimental procedure, a plasmid containing the whole 3 kb *lacZ* gene is microinjected in *in vitro* fertilized *Xenopus* embryos at the 2-cell stage (Figure 1A) and development is allowed to continue until before MBT (16-cell stage). Upon injection, plasmid DNAs form minichromosomes and replicate as episomes once per cell cycle with no sequence specificity (*31*, *32*). Total DNA is then extracted, purified and plasmid DNA is recovered in *E. coli* by transformation, since only plasmid DNA can transform bacteria (see Materials and Methods). Bacteria are plated on a chromogenic substrate (X-gal) to screen colonies for white or blue color. Wild-type *lacZ* produces active β-galactosidase which stains colonies in blue in the presence of X-gal and IPTG, while mutations generated in the *lacZ* gene that affect β-galactosidase activity will leave colonies colorless (white) or pale blue. A pre-MBT dose of supercoiled plasmid DNA (12 ng/embryo, Supplementary Figure S1A) was used in most of the experiments as previously described (*6*).

Recovery of the *lacZ*-containing plasmid DNA isolated from pre-MBT embryos into *E. coli,* generated white colonies with a frequency of 0,5 %, compared to the non-injected plasmid (pre-injection, Figure 1B) or to the same plasmid transfected into *Xenopus* somatic cells (Supplementary Figure S1E). Accordingly, mutation rate was calculated by normalization to the number of cell cycles (see Materials and Methods) and estimated to be in the order of 10^{-3} (Figure 1C, *lacZ*). Importantly, the mutation rate dropped to a background level when embryos were injected with a post-MBT amount of plasmid DNA, a situation that increases the N/C ratio and induces a cell cycle delay (*6*). Analysis of mutations by DNA sequencing revealed the presence of both single nucleotides variations (SNVs) and unexpectedly large deletions ranging from 100 bp to 1,5 kb (Figure 1D and Supplementary Figure S1B). Mutations inspection on the *lacZ* gene showed that they are generally widespread over the entire sequence with no hotspots (Supplementary Figure S1B). Analysis of the mutation spectrum shows that most SNVs detected were C>A and C>T changes (Figure 1D). Another frequent

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

signature was G>A transitions and T>A transversions, as well as nucleotides insertions and deletions. This mutation spectrum is close to that reported for TLS Pols on undamaged templates (*14*), in particular Polη and Polκ (*33*, *34*), although C>T transitions are also thought to be due to spontaneous deamination of 5-methyl cytosine to thymine.

The high frequency of base substitution and deletion prompted us to test the contribution of the mismatch repair system in the mutagenesis rate. For this, we overexpressed either wild-type or a catalytically inactive mutant (N38H) of Mlh1, a critical MMR component (*10*). We co-injected the *lacZ*-containing plasmid together with *in vitro*-transcribed MLH1 mRNAs to act as dominant negative by antagonizing the function of the endogenous protein (Figure 1E). While expression of $Mlh1^{WT}$ only slightly increased the mutation rate, this latter was increased 2-fold upon expression of the $Mlh1^{N38H}$ catalytically-inactive mutant (Figure 1E-G) suggesting that the MMR is functional and contributes to restrain mutagenesis. Altogether, these results show that the mutation spectrum observed in pre-MBT *Xenopus* embryos is similar to that expected for TLS Pols and that mutagenesis is restrained by the MMR system, suggesting that TLS Pols may actively contribute to mutagenesis in very early embryogenesis.

Rad18-dependent mutagenesis in the early *Xenopus* embryos

We have previously shown that in the pre-MBT *Xenopus* embryo TLS may be constitutively primed at replication forks in absence of external DNA damage (*10*). To determine the possible contribution of TLS to the mutagenesis in *Xenopus* embryos, we made use of a Rad18 TLS-deficient mutant in a dominant negative assay as done for Mlh1 (Figure 2A). Mutagenesis was analyzed as described in the previous paragraph. Expression of the TLS-deficient $Rad18^{C28F}$ mutant strongly reduced both the frequency of white colonies and the mutagenesis rate of about 100-fold compared to injection of either $Rad18^{WT}$ or *lacZ* alone (Figure 2B-C). In contrast, $Rad18^{WT}$ overexpression did not alter the mutagenesis rate compared to embryos injected with *lacZ* plasmid only, although it generated a different mutational spectrum, consisting of T>A transversions, C and G insertions, and remarkably no large deletions (Figure 2E and Supplementary Figure S2A). T>A transversions were reported to be significantly decreased in mice bearing the $pcna^{K164R}$ mutation that cannot support $PCNA^{mUb}$ (*35*), suggesting that this signature is Rad18 TLS activity-dependent. Compared to $Rad18^{WT}$, the residual mutagenesis observed in embryos injected with the $Rad18^{C28F}$ mutant showed a drastically reduced frequency of T>A transversion as well as C and T insertions, a TLS Polη and Polκ signature, suggesting that these mutations are $PCNA^{mUb}$-dependent, while the frequency of T>C transitions increased. These latter mutations are consistent with a Rev1 signature, a TLS Pol that can also be recruited independently of $PCNA^{mUb}$ (*36*, *37*).

We also tested the effect of expressing the homology-directed repair (HDR)-deficient, TLS-proficient, $Rad18^{C207F}$mutant, which is predicted to behave as $Rad18^{WT}$ (Figure 2A). Unexpectedly, however, expression of this mutant increased the number of white colonies of 2-fold compared to $Rad18^{WT}$ or *lacZ* alone, and the mutagenesis rate increased accordingly (Figure 2B-C) notwithstanding a similar expression level (Figure 2D). Compared to $Rad18^{WT}$,

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

expression of the Rad18^{C207F} mutant produced a reduction in both T>A transversions (Figure 2E) and generated large deletions (Supplementary Figure S2C and see below). The significant increase in single nucleotide substitutions generated by this mutant is consistent with the occurrence of unproofed mutations generated by its TLS activity. As expected, expression of the Rad18^{C28FC207F} double mutant produced a mutation burden similar to that of the TLS-deficient Rad18^{C28F} mutant, strongly suggesting that the mutagenesis restricted by the Rad18 HDR activity is TLS-dependent. Expression of this mutant increased G>A transitions and also produced large deletions (Figure 2E and Supplementary Figure S2D). In parallel, we analyzed mutagenesis when TLS is normally activated by UV irradiation and observed a very modest increase. The mutation spectrum was similar to that of the –UV condition and showed an increase in T insertions, as expected, which corresponds to TLS Polη and Polκ mutational signature (*33*)(Figure 2E and Supplementary Figure S1C-D), as well as disappearance of C>G transversions and reduction of C>T transitions. The modest increase in UV-induced mutagenesis is expected if TLS is constitutively activated, and is consistent with an error-free bypass of UV lesions by TLS Polη. Interestingly, no large deletions were detected (Supplementary Figure S1C and see discussion). Collectively, these results show that mutagenesis in the pre-MBT *Xenopus* embryo is Rad18-dependent and that, unexpectedly, the extent of TLS-dependent mutagenesis is alleviated by the error-free Rad18-dependent HDR activity.

Reduced hatching rate in *dpolη* maternally-deprived flies

In the aim to assess whether TLS-dependent mutagenesis is a general feature of fast cleaving embryos, and to obtain genetic evidence for this process, we turned to *Drosophila melanogaster*, a more genetically amenable system compared to allotetraploid *Xenopus*. First, we wished to establish whether developmental regulation of PCNA^{mUb} also occurs during *Drosophila* embryogenesis. Similar to *Xenopus*, *Drosophila* early development occurs through a rapid and synchronous series of embryonic cleavages before activation of zygotic transcription (MBT, Figure 3A)(*38*). Total protein extracts were prepared from *Drosophila* embryos before and after MBT and both total PCNA and PCNA^{mUb} levels were analyzed by western blot with specific antibodies (see Materials and Methods and Supplementary Figure S3A). Figure 3B-C shows that similar to what previously observed in *Xenopus* (*10*), PCNA^{mUb} is detectable in pre-MBT *Drosophila* embryos (0-2 hours) and declines at later stages (3-5 hours, post-MBT). The developmental stage where a decline in PCNA^{mUb} is observed coincided with that of Smaug, a mRNA polyadenylation factor destabilized just after MBT (*39*). We could not probe Rad18 expression since a *Drosophila* ortholog could not be found, neither by sequence homology, nor by structure-specific alignments (Busseau, Lo Furno, Bourbon, and Maiorano, unpublished). Altogether, these observations suggest that in pre-MBT *Drosophila* embryos TLS may be constitutively primed. In line with this conclusion, previous observations have shown that *Drosophila* Polη (dPolη) is highly expressed in pre-MBT embryos, localizes into interphase nuclei, similar to what previously observed in *Xenopus* (*10*), while *dpolη* mutant embryos are sensitive to UV-irradiation (*40*).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Very early embryogenesis of fast cleaving embryos is characterized by unusually contracted cell cycles, made of a periodic and synchronous succession of DNA synthesis (S-phase) and mitosis with virtually absent gap phases. S-phase length is dramatically short (15 minutes in *Xenopus* and only 4 minutes in *Drosophila*) and feedback mechanisms controlling genome integrity (checkpoints) are largely repressed, as there is no time to slow down the cell cycle (*1*) for review, and references therein). These include the ATR-dependent checkpoint that monitors replication fork progression (*2*)for review). This checkpoint is activated close to the midblastula transition (MBT) in concomitance with activation of zygotic transcription (*3–5*). Experiments in *Xenopus* have shown that checkpoint activation is sensitive to the DNA-to-cytoplasmic (N/C) ratio, since it can be triggered by artificially increasing the amount of DNA in the embryo over a threshold level, a situation that mimics the increase in DNA content reached close to the MBT (*6*). Previous observations in *Caenorhabditis elegans* (*7–9*) and more recently in *Xenopus laevis* (*10*) have implicated the translesion DNA synthesis (TLS) branch of DNA damage tolerance in silencing the DNA damage checkpoint. In *Xenopus* cleavage-stage embryos, constitutive recruitment of at least one Y-family TLS polymerase (Pol η) onto replication forks, driven by the TLS master regulator Rad18 (E3) ubiquitin ligase, minimizes replication fork stalling in front of UV lesions thereby limiting ssDNA production which is essential for replication checkpoint activation (*10–13*). This configuration is lost prior to MBT following a developmentally-regulated decline of Rad18 abundance (*10*).

TLS Pols have the unique capacity to replicate damaged DNA thanks to a catalytic site more open than that of replicative polymerases which can accommodate damaged bases. Because TLS Pols cannot discriminate the insertion of the correct nucleotide and lack proofreading activity, they can be highly mutagenic especially on undamaged templates (*14*)for review). Recruitment of Y-family TLS pols (ι, η, κ and Rev1) requires monoubiquitination of the replication fork-associated protein PCNA ($PCNA^{mUb}$) by Rad18 (E3) and Rad6 (E2) ubiquitin ligases complex (*15*, *16*). Aside from its TLS function, Rad18 is also implicated in error-free homology-directed DNA repair (HDR) in response to both double strand breaks (DSBs) and interstrand cross-links (*17–21*) . These functions are separable and lie in distinct domains of the Rad18 protein. The Rad18 TLS activity is confined to its ring finger domain (*22*), while the HDR activity mainly depends upon its zinc finger and ubiquitin binding domain (*17*, *18*). We have previously shown that in early *Xenopus* embryos PCNA is constitutively monoubiquitinated, irrespective of the presence of DNA damage (*10*). Whether TLS is active during the early embryonic cleavage stages is currently unclear. Previous work in *C. elegans* has shown that mutations in some TLS Pols do not influence global mutagenesis although a *polη* and *polκ* double mutant accumulate DNA deletions (*9*). In this work, we provide evidence for TLS-dependent mutagenesis in early *Xenopus* and *Drosophila* embryos and show that in *Xenopus*, both Rad18 HDR activity and the mismatch repair system (MMR) alleviate mutagenesis, thus reducing the mutation load.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

chromosomes revealed a significant SNV depletion on chromosome 3 of maternally-depleted flies compared to maternally provided flies (Supplementary Figure S4A-B and Figure 5C), in particular within a cluster of Responder (Rsp) satellite DNA repeat within the pericentromeric heterochromatin of chromosome 3L (Figure 5D). The SNVs difference in this region accounts for 100-fold decrease in the mutagenesis rate in the maternally-depleted flies compared to the maternally-provided flies, consistent with an error-prone activity of dPolη. A difference was also observed in the pericentromeric region of the right part of the same chromosome (3R) including a shift far from the centromere in the maternally-depleted flies (Figure 5D), while no gross variations were observed on other chromosomes, except some significant variations on chromosome 4, X and Y (Supplementary Figure S4D-F). Analysis of the mutation spectrum on chromosomes 3R and 3L (Figure 5C) revealed a predominant reduction of C>T, T>C transitions and T>A transversions. This is in line with the observation that unlike yeast and humans, dPolη misincorporates G opposite T template leading to T>C transitions (*45*).

Because approximately two third of the genes located on 3L pericentromeric heterochromatin are required for developmental viability and/or adult fertility (*46*) we evaluated the predicted effects of mutations in either maternally-depleted or maternally-provided adults by attributing variant effect predictor (VEP) score to each variation. VEP score determines the effect of variants (SNVs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. In both maternally-depleted and maternally-provided flies genomes, the majority of variants present a modifier score more enriched in the maternally-depleted mutant on the pericentromeric region of chromosome 3L and 3R (Supplementary Figure S5A). Most variants in this category affect intron splicing or noncoding regions (intergenic variants, Supplementary Figure S5A-B). However, once the VEP score modifier removed, the most recurrent SNVs presented a moderate score or a low score, (Figure 5C-D). Scoring the consequences of these variants shows that they lead to missense mutations in coding genes in the maternally-deprived mutant. This category of variants changes the genetic code, which may potentially alter the function of a protein.

Taken altogether, these data show that *dpolη* maternally-depleted adults are characterized by decreased mutations on specific chromosomes regions that may depend upon dPolη for efficient replication during the very fast cleavage stages and containing genes important from embryos viability (*46*).

Discussion

High mutagenesis in the very early *Xenopus* embryo

In this work, we have provided evidence for the occurrence of a surprisingly high mutation rate in the very early, pre-MBT *Xenopus* embryo. Mutation rate was estimated to be in the range of 10^{-3} and corresponds to 0.8 mutations per cell cycle, a value very close to that observed in the human germline (0.4-1.2) (*47*) but slightly lower than that estimated for pre-implantation human embryos (2.8) (*48*). This mutation rate, which is within the range

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

observed for Y-family TLS Pols, was greatly reduced upon expression of either the TLS-deficient Rad18^{C28F} or the Rad18^{C28FC207F} mutant. The corresponding mutation spectrum is also consistent with the mutagenesis spectrum of TLS Pols on undamaged DNA templates. Collectively, these findings indicate that in pre-MBT *Xenopus* embryos TLS strongly contributes to the observed mutagenesis.

The residual mutations observed in the Rad18^{C28F} mutant includes C>T transitions and C>A transversions. These mutations, that were recently reported to be also predominant in early human embryos (*48*), can be a consequence of either ribonucleotides incorporation or generated as a result of cytosine deamination into uracil, which is then turned into a thymine upon replication. The concentration of ribonucleotides exceeds of about 1000-fold that of deoxyribonucleotides, and we have observed a high level of ribonucleotides incorporation during DNA synthesis in *Xenopus* egg extracts that depends upon the DNA-to-cytoplasmic ratio (Supplementary Figure S2E). This may depend upon TLS activity, in line with evidence demonstrating ribonucleotides incorporation *in vitro* by human Polη (*49*, *50*). Because ribonucleotides slow down DNA replication, constitutive TLS activation facilitates their bypass, a strategy that the *Xenopus* embryos may have evolved to cope with a highly contracted cell cycle. Notwithstanding, it cannot be excluded that these mutations might be also a consequence of unproofed errors of replicative DNA Pols.

Unexpectedly, we have also observed large deletions in the *lacZ* gene recovered from pre-MBT embryos. These rearrangements are unlikely to be an artifact of the plasmid assay, since they were not detected neither in plasmids isolated from embryos co-injected with Rad18^{WT}, nor from UV-irradiated embryos. In this respect, genomic deletions have been observed in the *S* subgenome of adult *Xenopus laevis* (*51*) as well as in *Drosophila melanogaster* (*52*), suggesting that such genomic rearrangements might be generated naturally during evolution in these organisms. Although Y-family TLS Pols can generate deletions, their extent is rather small (1-3 bp), implying other mechanisms such as replication fork instability which is a common feature of DNA damage checkpoints inefficiency (*5*, *53*, *54*). Replication fork collapse can also happen when repriming and template switch are inefficient, however we have found no evidence for repriming nor for template switching in the context of *Xenopus* early embryogenesis (*10*) (and our unpublished observations). Replication fork collapse can be a consequence of suboptimal TLS activity due to limiting Rad18 levels (*10*). Consistent with this interpretation, *C. elegans* strains with reduced TLS function accumulate spontaneous genomic deletions as a result of double strand breaks at forks arrested by endogenous DNA lesions (*43*). Notwithstanding, it cannot be excluded that these rearrangements are the result of rare intermediates, Polη-dependent, that turned into deletion upon transformation into *E.coli*. Rad18^{WT} overexpression would reduce fork stalling by boosting both TLS and HDR, suppress NHEJ toxic effect at collapsed replication forks and therefore reduce deletions. This scenario is in line with evidence showing that Rad18 has a negative effect on NHEJ (*55*) and that NHEJ is predominant over HDR in the early *Xenopus* embryo (*56*). Externally applied DNA damage may stimulate repriming and/or template switch by so far unclear mechanisms, thus facilitating replication fork restart, and suppressing replication fork collapse.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Functional conservation of constitutive TLS in the early embryogenesis of fast cleaving embryos

Similar to what observed in *Xenopus* (*10*), we have provided evidence for both developmental regulation of $PCNA^{mUb}$ in *Drosophila* early embryogenesis, and TLS activity, suggesting that this process is also conserved in invertebrates. Because a *Drosophila* Rad18 ortholog in *Drosophila* could not be identified, it has not been possible to analyze mutagenesis in a complete Y-family TLS-free context.

Detailed genome-wide analysis of SNVs in maternally-depleted *dpolη* adults revealed a strong SNVs reduction in the pericentromeric region of chromosome 3, as well as SNVs depletion on the Y chromosome and the pericentromeric region of chromosome X. It is currently unclear why dPolη maternal depletion mainly affects SNVs abundance on chromosome 3. This is the largest *Drosophila* chromosome which includes the largest cluster of 120 bp *Responder* DNA repeats of α-satellite DNA within the pericentromeric heterochromatin (*57*). Such DNA sequences form secondary structures that constitute a challenge for a canonical replication fork, and Polη, and not Pol ι, has been previously shown to be important to replicate unusual DNA sequences in somatic cells (*58*, *59*). Because *Drosophila* lacks Polκ, Polη may be essential to assist the replisome in the replication of heterochromatin, in particular on chromosome 3 that contains the largest block of *Responder* DNA repeats. This interpretation is consistent with SNVs depletion observed on chromosome Y which is highly heterochromatic and with the mutation spectrum that corresponds to the reported incorporation errors of Polη on undamaged templates (*33*, *35*, *45*). Either replicative polymerases, or dRev1 may compensate for dPolη absence, although less efficiently. Due to inefficiency of the replication checkpoint in the *Drosophila* pre-blastoderm syncytium, embryos may accumulate chromosome abnormalities and undergo apoptosis at MBT (*5*), thus explaining the reduced hatching rate and chromosome abnormalities observed in maternally-depleted *dpolη* embryos. A caveat of this interpretation is that mapping to highly repetitive genomic regions is not very accurate. However, we do not have any indication that this contributed to the difference in SNVs identified on 3L and 3R chromosome arms between the maternally-depleted and maternally-provided dPol η flies.

The 3L chromosome region also contain a set of genes involved in development and viability. A great majority of SNVs in this region are predicted to generate mutations with low or moderate impact on genes functions. Hence, it cannot excluded that the phenotype observed in the *dpolη* homozygous flies may also be a consequence of mutations in essential genes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Consequences of a high mutagenic rate in early embryogenesis: good or bad?

The occurrence of a high mutation rate in early developing embryos of fast cleaving organisms is rather surprising but somehow not completely unexpected since these embryos are characterized by a highly contracted cell cycle that does not leave enough time to allow quality control (*1*). In this situation, the toll to pay is an increased risk of mutagenesis and genomic instability, as we have reported in this work. Several reports have highlighted the occurrence of genomic instability and mutations in early embryos (*1*) for review), which is apparently compatible with normal development (*63*). These observations suggest that active protection mechanisms must be operating to reduce the mutation load. Consistent with this possibility, mutagenesis dropped to background levels when *Xenopus* embryos were injected with a high dose of DNA, which mimics a pre-MBT stage (*6*). At MBT, cell cycle extension, activation of the DNA damage checkpoint and apoptosis would ensure repair of errors introduced during the cleavage stages, thereby limiting the propagation of cells having gross chromosomal alterations, and explaining both the low level of developmental defects and embryonic mortality (*5*, *60–62*). In this work, we have unveiled that in *Xenopus*, Rad18 has a protective function through an error-free HDR activity that reduces its TLS mutagenic activity. In addition, we have shown that the MMR pathway also contributes to reduce mutagenesis in the pre-MBT embryo. However, we have shown that in *Drosophila*, mutations generated in the pre-blastoderm embryo are inherited in the adult, suggesting that protection mechanisms against genomic instability are not very stringent.

Introduction of random mutations, generated by so far unclear mechanisms, has been also recently observed in human embryos (*48*). In addition, pre-implantation human embryos display genomic instability characterized by gross chromosomal rearrangements that can lead to cleavage arrest at the 2-4 cell stage (*63*). Introduction of random mutations may constitute an unexpected and novel source of genetic variation contributing to genome evolution that may be advantageous for the adaptation of the species, but at the same time might be dangerous for life. For example, an overall high mutation rate may be important for pseudogenization, a process that silences the expression of pseudogenes (*64*) and be also important to adaptation to a new environment. A recent study identified several genes located on the *Drosophila* 3L chromosome involved in adaptation (*65*). We have observed that Polη mutagenic activity may be important to maintain the stability of centromeric DNA sequences in the *Drosophila* pre-blastoderm embryo, thus being good for life. Genome-wide association studies have implicated hundreds of thousands of single-nucleotide polymorphisms (SNPs) in human diseases and traits (*66*). In the future, it will be important to explore which is the level of DNA damage inherited in the post-MBT embryo, and its contribution to the polymorphisms that characterize each individual.

Acknowledgments

We wish to thank Marcel Méchali for technical advises, J-S Hoffmann and SE Kearsey for critical reading of the manuscript, B Rondinelli, J Sale, A-M Martinez, A Goriely for useful discussions and H-M Bourbon for help with sequence alignments. This project was supported

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

by ANR (ANR-12-BSV2-0022) and MSD Avenir grants to D.M. E.L.F. was supported by a 3-year PhD fellowship from the "Ligue contre le Cancer" and 1-year fellowship from "Fondation ARC contre le cancer".

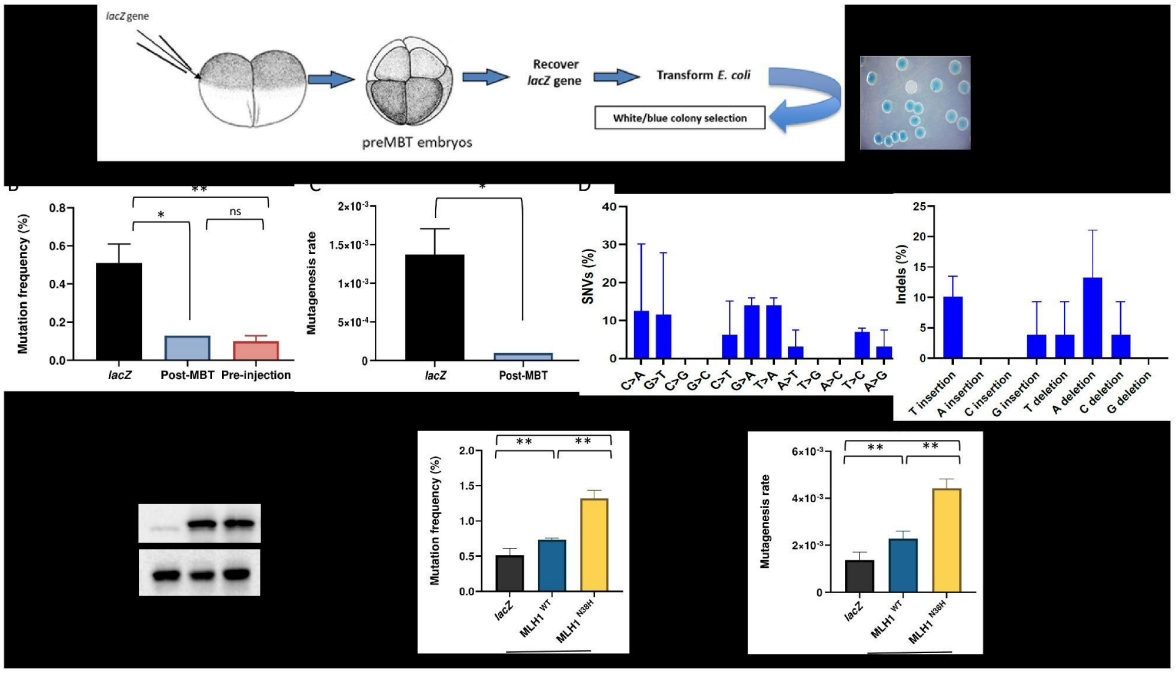
Competing Interests: The authors declare that they have no competing interests

Author contributions

Conceptualization, D.M., E.L.F.; Methodology, D.M., E.L.F., I.B.; Investigation, E.L.F., I.B.; Formal Analysis, C.S., S.Z., C.M.D., C.L.; Writing, E.L.F., D.M.; Funding Acquisition, D.M.; Supervision, D.M., I.B.; Visualization, D.M. and E.L.F.; Resources, D.M. and I.B.

Data repository

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. WGS data have been submitted to Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/; ID GSE161335).


lacZ gene
preMBT embryos
Recover lacZ gene
Transform E. coli
White/blue colony selection
Mutation frequency (%)
lacZ
Post-MBT
Pre-injection
Mutagenesis rate
SNVs (%)
C>A
G>T
C>G
G>C
C>T
G>A
T>A
A>T
T>G
A>C
T>C
A>G
Indels (%)
T insertion
A insertion
C insertion
G insertion
T deletion
A deletion
C deletion
G deletion
MLH1 WT
MLH1 N38H

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.28.401471; this version posted June 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1. Pre-MBT *Xenopus* embryos accumulate polymorphisms and deletions

(A) Drawing of the experimental strategy adopted to analyze mutagenesis in *Xenopus laevis* embryos. 2-cell stage embryos are injected with a supercoiled plasmid containing *lacZ*-reporter gene (pEL1) and allowed to replicate for further 3 divisions. After embryos collection, plasmid DNA is extracted and transformed in *lacZ*-deficient bacteria for white/blue screening.

(B) Mutation frequency expressed as percentage of white colonies in each condition. The mutation frequency of *lacZ* recovered from embryos injected with a post-MBT amount of plasmid DNA is also included as comparison. pre-MBT and post-MBT n=3, pre-injection: n=2.

(C) Mutagenesis rate in the indicated different experimental conditions expressed as mutations per base pair/locus per generation (see Materials and Methods), normalized to the pre-injection background values, n=3.

(D) Mutation spectra of the *lacZ* gene recovered from *Xenopus* pre-MBT embryos after Sanger sequencing, n=3.

(E) Western blot of total protein extracts obtained from *Xenopus* embryos subjected to the indicated experimental conditions, n=2.

(F) Mutation frequency and (G) mutagenesis rate of *lacZ* isolated from *Xenopus* embryos injected as indicated. *lacZ* n=3; Mlh1: n=2.

Data are presented as means ± SD. Means were compared using unpaired Student's t test.

Acknowledgements

Probabilmente non riuscirò mai ad usare questa barca,
ma costruendola potrò mettere in gioco me stesso
e insegnare ai miei figli che con dedizione e tenacia
si può raggiungere traguardi inattesi.
Mario Gadda

The work included in this manuscript has my name on it, but it's really the outcome of a large network of people that, more or less directly, contributed to its realization.
I'm not the type of person who likes long and melodramatic poems, but I'll have to make an exception since this will be probably the last occasion I'll have to write down a sincere acknowledgement. And I'll do it my way!

https://cloxd.github.io/Thesis_Acknowledgment/