UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Modèles, méthodes et algorithmes en biologie, santé et environnement**

Arrêté ministériel : 7 août 2006

Présentée par

Laure Sambourg

Thèse dirigée par **Nicolas Thierry-Mieg**

préparée au sein du **laboratoire TIMC-IMAG**
et de **l'école doctorale EDISCE**

Décrypter les données omiques : importance du contrôle qualité. Application au cancer de l'ovaire

Thèse soutenue publiquement le **18 décembre 2013**,
devant le jury composé de :

Mme Christine Brun
Directeur de recherche CNRS, TAGC, Marseille, Rapporteur

Mr Éric Rivals
Directeur de recherche CNRS, LIRMM, Montpellier, Rapporteur

Mr Olivier François
Professeur des universités Grenoble INP, TIMC-IMAG, Grenoble, Examinateur

Mr Daniel Kahn
Directeur de recherche INRA, LBBE, Lyon, Examinateur

Mr Nicolas Thierry-Mieg
Chargé de recherche CNRS, TIMC-IMAG, Grenoble, Directeur de thèse

Résumé

Contexte

Depuis le 14 avril 2003, date historique marquant la parution du séquençage du premier génome humain, d'immenses progrès technologiques ont été faits en matière de séquençage de l'ADN : un génome humain complet peut aujourd'hui être séquencé en une semaine pour un coût raisonnable. En parallèle, et souvent en s'appuyant sur les technologies de séquençage, de nombreuses autres techniques de biologie moléculaire ont également été développées pour des applications à très haut débit. Pour citer quelques exemples, le RNA-seq, le ChIP-seq, ainsi que des méthodes d'exploration des interactions entre protéines génèrent une quantité massive de données dites "omiques" (données génomiques et post-génomiques produites à haut débit).

L'analyse de ces données omiques promet de nombreuses découvertes en génétique et en médecine, mais nécessite de créer et d'adapter continuellement de nouvelles méthodes bioinformatiques afin d'être capable de stocker, d'intégrer et de comprendre cet immense afflux d'information (1, 2).

En particulier, une étape qui me semble primordiale et sur laquelle j'ai mis l'accent tout au long de ma thèse est le contrôle de la qualité des données. En effet, les données omiques sont parfois très incomplètes, et contiennent de nombreux biais et erreurs systématiques qu'il est facile de confondre avec de l'information biologiquement intéressante (3). Ces remarques s'appliquent à tous types de données, mais je me suis concentrée sur des données d'interaction protéine-protéine chez *S. cerevisiae* d'une part, et d'autre part sur des données de séquençage next-generation (NGS) de patientes atteintes d'un cancer séreux de l'ovaire.

Incomplétude de l'interactome de *S. cerevisiae*

Les interactions entre protéines régissent la plupart des mécanismes cellulaires, et les réseaux d'interactions protéine-protéine sont donc essentiels à l'étude des processus cellulaires. Par conséquence, plusieurs projets de cartographie à haut débit des interactomes ont été entrepris, et les interactions protéine-protéine sont rassemblées dans les bases de données par curation de la littérature. Cependant, les données d'interactions protéine-protéine sont encore loin d'être complètes, même chez l'organisme modèle *S. cerevisiae*. Estimer la taille de l'interactome est important pour évaluer la complétude des jeux de données, et pour mesurer les efforts qu'il reste à fournir.

Nous avons examiné l'interactome de la levure d'un point de vue original, en prenant en compte le degré d'étude des protéines par la communauté scientifique. Nous avons découvert que l'ensemble des interactions protéine-protéine de la littérature est qualitativement différent quand il est restreint aux protéines qui ont reçu une attention particulière de la communauté scientifique. En particulier, ces interactions s'appuient moins souvent sur la méthode double hybride, et plus souvent sur des expériences plus complexes, comme des analyses de l'activité biochimique. Notre analyse montre que les interactomes obtenus à haut débit et par curation de la littérature sont plus corrélés que ce qui est communément supposé, mais que cette corrélation s'estompe en se concentrant sur les protéines très étudiées. Nous proposons une méthode simple et fiable pour estimer la taille d'un interactome, en combinant les données concernant les protéines très étudiées par la communauté scientifique et les données haut débit. Notre méthode conduit à une estimation d'au moins 37 600 interactions physiques directes chez *S. cerevisiae*. Cette estimation est plus élevée et plus fiable que les estimations précédentes, car elle prend en compte les interactions difficiles à détecter avec les expériences communément utilisées. Cela montre que nous sommes encore plus loin de compléter l'interactome de la levure que ce à quoi on s'attendait.

Importance des variants germinaux dans le développement du cancer séreux de l'ovaire

The Cancer Genome Atlas (TCGA) (4) est un consortium international ayant pour but de caractériser les différentes tumeurs et de cartographier leurs modifications aux niveaux génomique, épigénétique et transcriptomique. Entre autres, ce consortium a récemment publié le séquençage de l'exome de 520 adénocarcinomes séreux ovariens et des tissus sains correspondants (5). Des mutations somatiques (*i.e.* présentes dans la tumeur et absentes du tissu sain) ont été identifiées comme potentiellement importantes dans la cancerogénèse, mais les SNVs germinaux (ou constitutionnels) n'ont pas été analysées (David Wheeler, Baylor College of Medicine, communication personnelle). Nous avons donc décidé d'étudier les exomes des tissus sains de ces patientes dans le but d'identifier des prédispositions génétiques jouant un rôle important dans le cancer. Ce travail a d'abord nécessité d'aligner et de génotyper les short-reads produits par TCGA, puis de prioritariser les variants obtenus afin d'identifier ceux impliqués dans le développement du cancer.

Appeler des génotypes fiables à partir des données NGS

La caractérisation d'un exome passe par plusieurs étapes expérimentales : la construction de librairies génomiques, la capture des parties supposées codantes, et enfin le séquençage à proprement parler, produisant des millions de short-reads. Ces short-reads doivent ensuite être alignés sur le génome de référence, afin d'identifier les variations entre le génome étudié et la référence (appel des génotypes). Pour ce faire, nous avons utilisé l'aligneur MAGIC (6) et développé une méthode d'appel de génotypes.

L'appel de génotypes consiste à classifier les loci du génome étudié en homozygotes références, hétérozygotes ou homozygotes variants, en se basant sur la proportion de reads variants s'alignant sur le locus. Ces proportions devraient théoriquement valoir 0 (homozygote référence), 0.5 (hétérozygote) ou 1 (homozygote variant), mais la présence de biais d'échantillonage et d'erreurs de séquençage ou d'alignement les fait dévier et peut rendre difficile la différenciation entre les vrais variants et les erreurs. Les méthodes pour répondre à ce problème sont souvent classées en deux catégories : les méthodes heuristiques, qui utilisent des seuils fixés, et les méthodes statistiques qui font l'hypothèse que la proportion de reads variants suit une loi binomiale (7).

Pour les données TCGA, nous montrons que les positions présentant entre 20 et 40% de read variants ne peuvent pas être appelés en se basant uniquement sur la proportion de reads variants. Pour palier à ce problème, nous avons donc développé une extension simple de la méthode des seuils, basée sur l'utilisation de zones tampons. De plus, notre méthode permet de filtrer les erreurs systématiques en analysant les reads alignés sur les deux brins. La comparaison des nombreux réplicats techniques produits par le consortium TCGA montre l'efficacité de cette méthode, avec une diminution importante du taux de faux positifs. Nous obtenons ainsi une liste de SNVs germinaux fiables pour les 520 patientes atteintes de cancer de l'ovaire.

Prédisposition au cancer de l'ovaire : prioritariser les SNVs candidats

Le principal défi de l'analyse de données NGS est d'identifier, parmi l'ensemble des variants germinaux d'une patiente, ceux susceptibles d'avoir un impact sur le développement de la maladie (variants "de susceptibilité").

Il est probable qu'une majorité des variants de susceptibilité aient un impact sur les protéines, voir les rendent non fonctionnelles, et nous avons choisi de nous concentrer sur ceux-ci. Nous avons donc estimé l'impact des SNVs sur les gènes connus, en nous basant sur les transcrits de la base de données Aceview (6). Ainsi, nous avons filtré les mutations synonymes et les mutations non codantes (à l'exception des pieds d'introns). Quant aux mutations faux sens, certaines peuvent complètement modifier la conformation de

la protéine, ou dénaturer les sites de liaisons, empêchant ainsi la protéine d'accomplir son rôle, alors que d'autres peuvent être sans grande conséquence. Pour discriminer ces mutations, nous avons utilisé les logiciels SIFT (8) et Polyphen-2 (9), qui prédisent l'impact d'une mutation en se basant sur la conservation des protéines ou les caractéristiques biochimiques des acides aminées. De surcroît, nous avons gardé uniquement les variants affectant les protéines exprimées dans l'ovaire.

Dans un deuxième temps, afin d'identifier les variants significativement associés au cancer de l'ovaire, nous avons comparé les SNVs avec deux cohortes contrôles : les données du 1000 Genomes Project et de ESP6500. Nous trouvons en moyenne de 44 SNVs par patiente, répartis sur 334 gènes dans l'ensemble de la population : ces gènes jouent probablement un rôle dans la cancerogénèse (gènes "candidats"). Parmi ces derniers, 42 ont été reportés comme impliqués de près ou de loin dans la cancerogénèse. Le rôle de certains de nos gènes candidats dans le cancer est bien établi, mais pour d'autres, des études fonctionnelles sont nécessaires afin de confirmer leur implication dans le développement de la maladie. Un intérêt particulier doit être porté à la protéine MAP3K8, dont le rôle de suppresseur de tumeur a été très récemment proposé dans d'autres cancers, et sur laquelle 106 de nos patientes portent une mutation délétère.

Table des matières

1 Introduction **11**

1.1 Du séquençage du génome à la compréhension du vivant . . . 11
1.2 Les défis de l'ère post-génomique . . . 11
1.3 Diversité des données expérimentales . . . 12
1.3.1 Le séquençage nouvelle génération . . . 12
1.3.2 Séquencer un exome . . . 13
1.3.3 Interactions protéine-protéine . . . 13
1.3.4 Épigénétique et transcriptomique . . . 14
1.4 Importance de la bioinformatique . . . 16
1.4.1 Séquencer l'ADN : le travail *in silico* . . . 16
1.4.2 Interactomique . . . 16
1.4.3 Stocker et partager . . . 18
1.4.4 Intégrer et visualiser . . . 18
1.5 Grands projets, grandes découvertes . . . 19
1.6 Problématique : biais, incomplétude et erreurs systématiques des données omiques . . . 20

2 Interactomique : de surprenantes découvertes sur les données d'interactions protéine-protéine conduisent à une estimation plus élevée de la taille de l'interactome de *S. cerevisiae*. **21**

2.1 Données haut débit et données de curation de la littérature . . . 21
2.2 Les données LC sont enrichies en interactions facilement détectables par double hybride . . . 22
2.3 Utiliser le niveau d'étude des protéines permet de réduire la corrélation entre les données LC et HT . . . 24
2.4 Prendre en compte le niveau d'étude des protéines améliore les estimations de la taille de l'interactome . . . 25
2.4.1 Estimations existantes . . . 25
2.4.2 Faux positifs . . . 25
2.4.3 Amélioration des estimations de la taille de l'interactome de *S. cerevisiae* . . . 26
2.5 Conclusions . . . 27
2.6 Document joint : New insights into protein-protein interaction data lead to increased estimates of the *S. cerevisiae* interactome size . . . 27

3 Génotyper les individus à partir de données next-generation et filtrer les erreurs systématiques **39**

3.1 Alignement et appel des génotypes, des méthodes en développement . . . 39
3.2 Choix d'un pipeline de séquençage : le logiciel MAGIC et le projet SEQC . . . 40
3.3 Obtenir des génotypes fiables à partir des données NGS . . . 41

3.3.1 Données utilisées et pré-analyse . . . 42
3.3.2 Une méthode efficace pour appeler les génotypes et filtrer les erreurs systématiques 44
3.4 Conclusions . . . 48
3.5 Document joint : Filtering systematic errors in next-generation data : stranding matters . . 48

4 Prioritarisation des candidats, application au cancer de l'ovaire 63
4.1 Le paysage génétique des cancers . . . 63
4.2 L'analyse TCGA du cancer de l'ovaire . . . 65
4.3 Variants germinaux et prédisposition au cancer de l'ovaire . . . 66
4.4 Séquençage de l'exome de 520 patientes atteintes du cancer de l'ovaire : analyse des variants germinaux . . . 67
4.5 Sélectionner les variants "loss-of-function" . . . 68
4.6 Étude d'association . . . 70
4.6.1 Comparaison à une population contrôle . . . 70
4.6.2 Équivalence des indels . . . 71
4.6.3 Sélection des variants significativement associés au cancer . . . 72
4.7 Comparaison des gènes mutés avec les gènes connus comme potentiellement impliqués dans le cancer . . . 72
4.8 Analyse d'enrichissement en terme GO . . . 75
4.9 Conclusions . . . 76

5 Conclusions et perspectives 77

Bibliographie 81

Annexes 98
Puissance et limitations du RNA-seq . . . 98
Gènes significativement mutés chez 520 patientes atteintes du cancer de l'ovaire . . . 142

Remerciements

Remerciements

Je tiens à remercier chaleureusement mon directeur de thèse, Nicolas Thierry-Mieg, pour sa patience, son enthousiasme et son implication sans faille dans nos travaux de recherche. Son soutien, ses conseils, sa bonne humeur et son humour ont rendu ces 4 années de doctorat très instructives et très agréables : il a réussi à m'enseigner le travail de recherche tout en me faisant rire.

Je veux également remercier Éric Rivals et Christine Brun pour avoir accepté de rapporter ma thèse. Leurs remarques et commentaires constructifs m'ont permis d'aboutir à la version définitive de ce manuscrit, et leurs questions lors de la soutenance ont ouvert de nouvelles pistes de recherche pertinentes. Je remercie sincèrement Daniel Kahn et Olivier François pour avoir accepté de faire partie de mon jury de thèse, pour leur intérêt, et pour leurs commentaires constructifs.

J'adresse également un grand merci à Gary Bader et à son équipe, qui m'ont gentiment accueillie pendant 6 mois à l'Université de Toronto et qui m'ont fait découvrir de nouveaux horizons. Par ailleurs, je remercie tout particulièrement Danielle et Jean Thierry-Mieg qui m'ont beaucoup aidée dans mes recherches. Ils m'ont accueillie à Washington, au NCBI et m'ont beaucoup appris.

De plus, je souhaite exprimer à quel point j'ai apprécié l'excellente ambiance du laboratoire TIMC-IMAG. Merci, entre autres, à Flora pour son soutien, à Arnaud pour le chocolat, à Agnès pour les blagues, aux "bébés" pour leur jeunesse éternelle, à Mickaël et Olivier pour leurs conseils en statistiques et en escalade, à Sean pour les insectes, à Éric D. pour son sarcasme, à toute l'équipe BCM, ainsi qu'à l'équipe de basket TIMC-IMAG. Je leur souhaite le meilleur pour les années à venir.

Enfin, j'adresse un grand merci à ma famille, à Guillaume, à mes amis et à la fanfare Pinkitblack pour m'avoir donné l'énergie pour mener à bien cette thèse, et pour me permettre de garder ma bonne humeur en (presque) toutes circonstances. Un grand merci également à l'ensemble de mes colocataires qui ont rendu ma vie très agréable durant ces 4 ans.

Je souhaite à toutes les personnes que j'ai rencontrées pendant ces 4 ans une excellente continuation, et j'espère que nos routes vont continuer à se croiser.

Chapitre 1

Introduction

1.1 Du séquençage du génome à la compréhension du vivant

Au début du XXI^{ième} siècle, la publication de la séquence du premier génome humain (appelé génome humain de référence, HRG) promet d'impressionnantes découvertes en génétique humaine et de grandes avancées dans notre compréhension des mécanismes biologiques et cellulaires. Dix ans plus tard, le déchiffrage de cette séquence a permis, entre autres, l'identification et l'annotation de nouveaux gènes. Cependant, ces résultats sont très loin d'expliquer le fonctionnement du vivant : un phénotype, ou un comportement cellulaire, n'est pas le simple résultat de l'expression d'un gène, mais des interactions complexes entre génétique, épigénétique et environnement. Comprendre le fonctionnement de la Vie implique donc l'identification et la description de l'ensemble de ses composants et de leurs interactions. Ainsi, le génome humain de référence a servi de base à de nombreuses études de génomique (étude de l'ADN), transcriptomique (étude de l'expression des gènes), métabolomique (étude des petites molécules présentes dans les cellules) et protéomique (étude des protéines)(10). Si ces études ont permis de répondre à certaines questions, elles ont surtout montré que les concepts de gène et de régulation de la transcription sont bien plus complexes que ce qu'on imaginait.

1.2 Les défis de l'ère post-génomique

Au cours des 10 dernières années, des progrès spectaculaires ont été faits en matière de séquençage de l'ADN, avec l'apparition des technologies NGS (next-generation sequencing) (11). Alors que le séquençage du premier génome humain a nécessité 10 ans de travail, nous sommes aujourd'hui capable de séquencer un génome complet en 1 semaine, pour un prix raisonnable (autour de 5000 dollars, Figure 1.1).

En parallèle, et souvent grâce à ces technologies NGS, des domaines aussi variés que l'épigénétique, la protéomique, la transcriptomique et l'interactomique ont également connu un essor énorme, avec le développement de méthodes haut débit permettant d'interroger ces différents aspects de la cellule. La baisse du coût de production, la diversification des méthodes haut débit et l'amélioration du rendement entraînent la génération d'une quantité massive de données dites 'omiques'.

Ce déluge de données promet des avancées importantes dans notre connaissance des mécanismes moléculaires et cellulaires, mais pose des défis considérables à la communauté bioinformatique. Cela nécessite d'une part, d'avoir une vue d'ensemble des technologies existantes, de leurs mécanismes et de leurs limitations, et d'autre part, de créer et d'adapter continuellement de nouvelles méthodes bioinformatiques

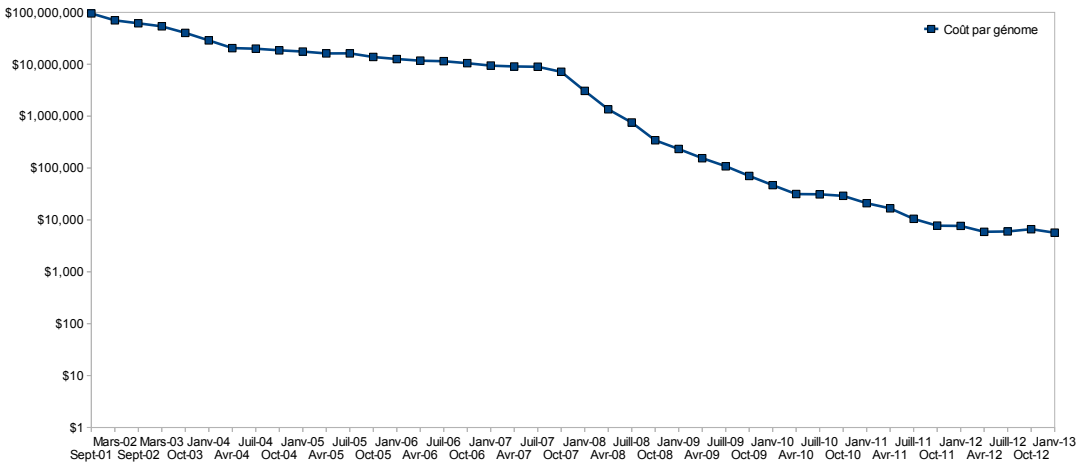


FIG. 1.1 – **Évolution du coût de séquençage d'un génome.** Le coût représenté ne prend pas en compte les aspects bioinformatiques et les analyses en aval. Adapté de (K. A. Wetterstrand, http ://www.genome.gov/sequencingcosts/)

afin d'être capable de stocker, analyser, intégrer, visualiser et comprendre cet immense afflux de données (1, 12, 2).

1.3 Diversité des données expérimentales

Au cours de ma thèse, je me suis principalement intéressée aux données de séquençage de l'ADN et aux données d'interaction protéine-protéine. Cette section décrit brièvement les technologies utilisées pour générer ces données, et donne un aperçu des autres techniques disponibles.

1.3.1 Le séquençage nouvelle génération

Il existe plusieurs plateformes de séquençage dites 'next-generation', produisant en parallèle des millions de courtes séquences d'ADN appelées 'short-reads'. Les plus utilisées actuellement sont les technologies Ion Torrent, SOLiD, Illumina, Roche 454 et Complete Genomics. Dans un premier temps, ces technologies nécessitent la préparation de templates : l'ADN génomique est fragmenté par des enzymes de restriction ou par sonication, puis les fragments sont liés à une séquence connue (adaptateur). L'adaptateur vient ensuite se fixer à une amorce universelle, permettant ainsi d'immobiliser les templates (11). Complete Genomics utilise un procédé différent : des enzymes de restriction sont utilisés pour créer une molécule d'ADN circulaire contenant 4 adaptateurs, séparés par des séquences inconnues de 13 et 26 paires de bases (13). Les technologies n'étant pas conçues pour séquencer des molécules uniques, ces templates doivent ensuite être amplifiés (par émulsion PCR pour les technologies Roche et SOLiD, par amplification en phase solide pour Illumina, par réplication circulaire pour Complete Genomics).

Les méthodes de séquençage et d'imagerie appliquées par la suite diffèrent selon les technologies. Illumina utilise des nucléotides terminaux 3' réversibles, marqués avec des fluorophores, qu'une ADN polymérase incorpore au template. Les nucléotides non fixés sont lavés, et les bases sont identifiées par imagerie. Le fluorophore est ensuite séparé et lavé, et le cycle recommence pour l'identification du nucléotide suivant. Les fragments d'ADN des templates illumina sont relativement longs (200-500 paires de bases (bp)). Les

75 bp du début du fragment vont d'abord être séquencés, suivi des 75 bp de la fin du fragment, générant des 'paired-end' reads. Ceux-ci facilitent l'alignement, et la détection de réarrangements structurels.

La première étape du séquençage par ligation utilisé par SOLiD consiste à construire des sondes constituées de bases dégénérées et de deux bases 'interrogatrices', et étiquetées par fluorescence. Ces sondes s'hybrident par complémentarité sur le brin à séquencer et sont jointes à l'amorce par ligation. Les sondes non liées sont lavées, et les bases interrogées sont identifiées par imagerie. Une partie de la sonde contenant le fluorophore est ensuite coupée, mais 3 bases dégénérées restent en place. Lors du prochain cycle de ligation, une sonde vient se lier à ces bases dégénérées : un ensemble de cycles séquence donc deux bases toutes les 5 bases. Pour séquencer les bases manquantes, ce processus est répété 5 fois, avec une amorce décalée d'une base à chaque fois. La particularité de SOLiD est l'utilisation du 'color-space' : les 16 dinucléotides des sondes sont encodés par 4 couleurs, et chaque base est donc interrogée deux fois. C'est cette redondance qui permet de reconstituer la séquence originale.

Lors du procédé utilisé par Complete Genomics, des nano-billes d'ADN sont attachées à des puces avant d'être séquencées, également par ligation. Contrairement à la méthode de SOLiD, les sondes ne contiennent ici qu'une seule base interrogatrice. De plus, après chaque imagerie, la sonde complète est enlevée et remplacée par une sonde contenant une base interrogatrice située un nucléotide plus loin, qui permet de lire la base suivante.

La méthode du pyroséquençage (Roche/454) repose sur la détection de la libération d'un pyrophosphate lors de l'ajout d'un nucléotide. Les quatre désoxynucléotides triphosphates (dNTP) sont ajoutés chacun à leur tour. Si le dNTP ajouté est complémentaire du brin à séquencer, il est incorporé par l'ADN polymérase, ce qui provoque la libération d'un pyrophosphate. S'ensuit une série de réactions enzymatiques convertissant cette libération en lumière, qui sert à détecter de l'ajout du nucléotide.

1.3.2 Séquencer un exome

Pour de nombreuses applications, le séquençage de génomes complets reste trop cher, et bien souvent, les études se restreignent aux parties supposées codantes du génome, les exons, diminuant ainsi grandement le coût de production et d'analyse des données. En effet, les 180000 exons représentent seulement 1% des 3 gigabases (3 milliards de bases) du génome humain. Le procédé de séquençage d'exome est le même que pour séquencer un génome complet, mais l'étape de séquençage est précédée d'une étape d'enrichissement, dans laquelle les exons sont capturés par hybridation à des sondes. Il est à noter que cette technique ne permet de séquencer que les gènes connus, ciblés par les sondes.

1.3.3 Interactions protéine-protéine

Les interactions protéine-protéine peuvent être de différentes natures. On distingue les interactions physiques binaires, possiblement transitoires, permettant par exemple la transmission d'un message dans une voie de signalisation, des interactions au sein d'un complexe ou encore des interactions génétiques. Plusieurs méthodes expérimentales ont été conçues pour détecter ces différents types d'interactions.

Les interactions génétiques, ou phénomènes épistatiques, décrivent la modification des effets d'un gène par un ou plusieurs autres gènes. Ces interactions peuvent être de natures différentes. À titre d'exemple, on peut citer les études de léthalité synthétique, qui testent si la combinaison de plusieurs gènes mutés conduit à la mort cellulaire, alors que ces mêmes mutations prises indépendemment conduisent à un phénotype viable.

Par ailleurs, il existe principalement 3 méthodes de détection d'interactions protéine-protéine physiques utilisées à haut débit (Figure 1.2). TAP-MS (Tandem Affinity Purification-Mass Spectrometry) est une mé-

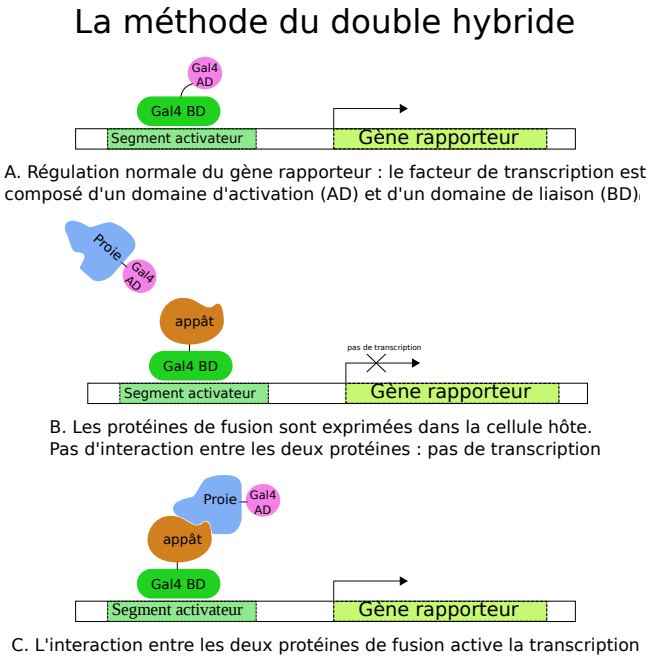
thode permettant d'identifier les composants d'un complexe. Il s'agit donc d'interactions stables, pouvant être indirectes. Elle implique la construction d'une séquence exprimant une protéine de fusion formée de la protéine d'intérêt (appât) et d'une étiquette (tag). Cette séquence est intégrée dans une cellule hôte, où la protéine, si elle est exprimée, peut alors interagir avec d'autres protéines (proies). L'appât est récupéré par colonne d'affinité, entraînant avec lui les éventuelles protéines du complexe. Finalement, le complexe est purifié grâce à une deuxième colonne d'affinité et ses composants sont identifiés par spectrométrie de masse.

Y2H (Yeast Two Hybrid) et PCA (Protein Complementation Assay) sont utilisées pour cribler un ensemble de paires de protéines (Figure 1.2), afin de déterminer celles qui interagissent physiquement, l'interaction pouvant être transitoire. La méthode du double hybride (14) repose sur une propriété de certains facteurs de transcription. En effet, ces derniers sont composés de deux domaines principaux : le domaine de liaison, qui se lie sur la molécule d'ADN en amont du gène régulé, et le domaine d'activation, qui active la transcription du gène. La transcription est activée si le domaine d'activation est à proximité du gène régulé, sans qu'il soit nécessaire que les deux domaines soient en contact direct. Ainsi, dans la méthode double hybride, les protéines étudiées sont attachées l'une au domaine de liaison (l'appât), l'autre au domaine d'activation (la proie). Les deux plasmides sont introduits simultanément dans l'organisme hôte (la levure, quand cette méthode est appliquée à haut débit). Si les deux protéines interagissent, le domaine d'activation est porté à proximité du gène rapporteur, et active sa transcription. Une interaction entre les deux protéines induit donc un phénotype différent et permet la sélection.

Lors d'une expérience PCA (15, 16), les deux protéines d'intérêt sont fusionnées à deux fragments complémentaires d'une protéine rapporteuse. Si les deux protéines interagissent, les deux fragments de la protéine rapporteuse sont alors à proximité, ce qui restaure l'activité rapporteuse. La méthode PCA a été appliquée avec différents types d'activité rapporteuse, notamment la fluorescence (luciférase) et la sélection par croissance cellulaire. L'avantage principal de cette méthode est la détection d'interactions dans le compartiment cellulaire natif.

1.3.4 Épigénétique et transcriptomique

De nombreuses techniques utilisent le séquençage NGS pour interroger d'autres aspects des cellules (12). Par exemple, l'expression des gènes et l'épissage alternatif peuvent être étudiés par RNA-seq, une méthode permettant de séquencer les ARNm après les avoir isolés et convertis en ADNc. La technique ChIP-seq permet d'identifier les sites de liaison ADN-protéine, en séquençant les portions d'ADN obtenues par immunoprécipitation de chromatine. En utilisant un procédé similaire, l'immunoprécipitation d'ADN méthylé suivi de séquençage (MeDIP-seq) identifie les parties méthylées du génome. Les techniques de DNase-seq (17) et FAIRE-seq permettent de séquencer les régions hypersensibles à la DNase 1, c'est-à-dire des régions d'ADN décondensé (principalement des régions régulatrices). Par ailleurs, les variations du nombre de copies d'un gène (CNV, copy number variation) peuvent être détectées par puce d'hybridation génomique comparative (CGH array). Cette technique est typiquement utilisée pour identifier les CNVs dans les tumeurs : l'ADN tumoral et l'ADN normal sont marqués différemment, puis l'hybridation préférentielle d'un ADN à une sonde est détectée par fluorescence. Des techniques pour identifier les CNVs à partir de données NGS ont également été proposées (18).



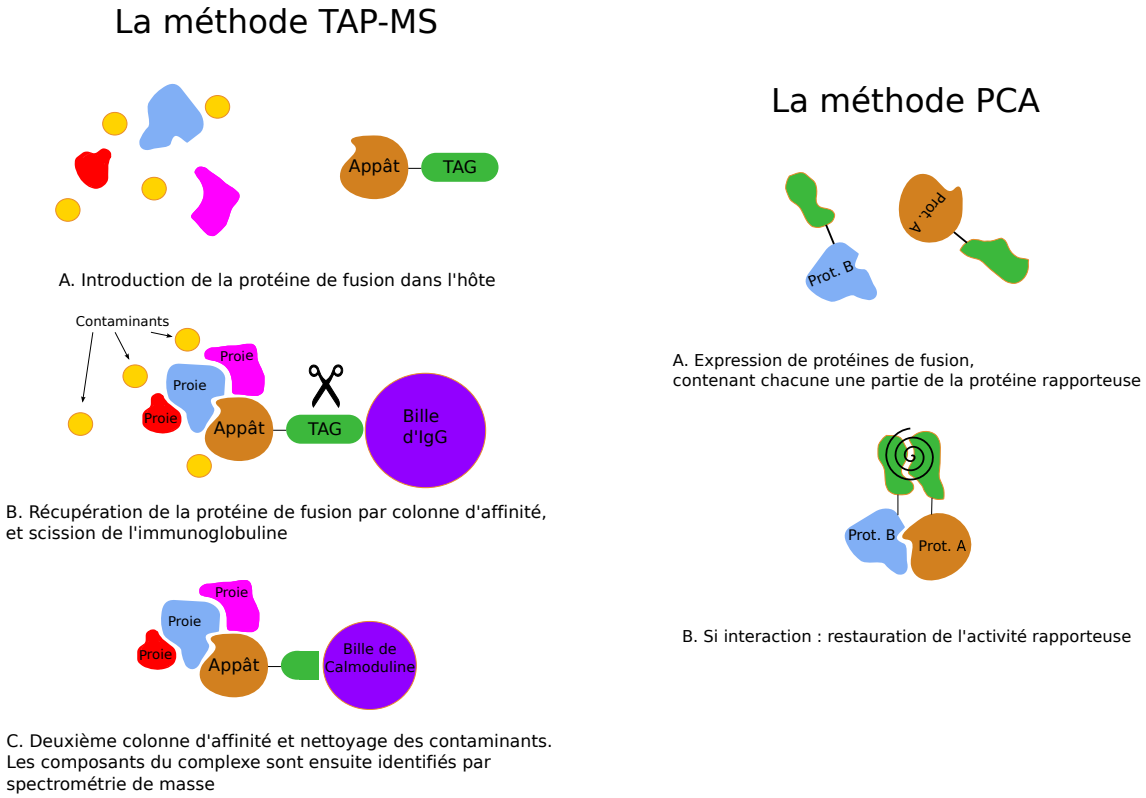


FIG. 1.2 – **Détection d'interaction protéine-protéine.** Schémas simplifiés de 3 méthodes de détection d'interactions protéine-protéine : la méthode double hybride (d'après Anna K., wikipédia (2007)), la méthode TAP-MS et la méthode PCA.

1.4 Importance de la bioinformatique

La bioinformatique est une thématique de recherche interdisciplinaire, qui combine les outils informatiques et mathématiques dans le but d'extraire de l'information des données biologiques. Cette filière a été énormément stimulée par l'émergence des technologies haut débit, qui font de la bioinformatique un passage obligé dans le traitement des données : seul un ordinateur offre la possibilité d'analyser des données à l'échelle du protéome ou du génome. Ses champs d'application sont très variés et je me suis principalement intéressée au séquençage de l'ADN et à l'interactomique. Cette section présente ces deux sous-disciplines et décrit brièvement d'autres aspects importants de la bioinformatique.

1.4.1 Séquencer l'ADN : le travail *in silico*

Avant d'être séquencé, l'ADN est fragmenté, puis les fragments sont amplifiés et séquencés en parallèle, produisant des millions de 'short-reads'. Ceux-ci doivent ensuite être ré-assemblés afin de reconstruire la séquence originale. Dans le cas où il n'existe pas de génome connu pour l'espèce étudiée, on parle d'assemblage *de novo*. Si un génome de référence existe, les short-reads peuvent être alignés sur ce génome. Le problème vient alors des différences entre le génome étudié et la référence (présence d'indels, de SNVs (Single Nucleotide Variation) et de réarrangements), de séquences répétées difficiles à identifier et de la taille des données (temps de calcul et coût en mémoire des algorithmes). La plupart des algorithmes d'alignement indexent le génome de référence ou les short-reads par des k-mers (19, 20, 21, 22, 23) et cherchent les correspondances exactes de ces k-mers. Ces algorithmes diffèrent entre autres par la façon d'indexer et de localiser les k-mers (table de hachages, transformée de Burrow-wheeler (23, 22, 21), Gk-array (22)), et par la taille des k-mers utilisés, paramètre important pour la spécificité et la sensibilité du mapping (24). Une fois l'alignement réalisé, il faut génotyper le génome étudié, c'est-à-dire établir ses différences avec le génome de référence (Figure 1.3). Dans le cas d'un génome diploïde, l'appel de génotypes consiste donc à classifier les loci du génome étudié en homozygotes références, hétérozygotes ou homozygotes variants, en se basant sur la proportion de reads variants s'alignant sur le locus. Ces proportions devraient théoriquement valoir 0 (homozygote référence), 0.5 (hétérozygote) ou 1 (homozygote variant), mais la présence d'erreurs de séquençage, d'échantillonnage (pour les hétérozygotes) ou d'alignement les fait dévier et peut rendre difficile la différenciation entre les vrais variants et les erreurs. Les méthodes pour répondre à ce problème sont souvent classées en deux catégories : les méthodes heuristiques, qui utilisent des seuils fixés, et les méthodes statistiques qui font l'hypothèse que la proportion de reads variants suit une loi binomiale (7). L'application de ces méthodes produit une liste de variants, caractérisant le génome séquencé.

1.4.2 Interactomique

Principaux acteurs de la cellule, les protéines interviennent dans tous les mécanismes cellulaires, à de nombreux niveaux : elles peuvent avoir un rôle structurel, catalytique, régulateur, messager, etc. Pour accomplir leur rôle, elles interagissent de façons très diverses, faisant ainsi fonctionner la machinerie cellulaire. Par exemple, d'une part, la formation de complexes leur permet d'accomplir des tâches qu'une protéine seule ne pourrait pas réaliser et d'autre part, des interactions plus transitoires sont essentielles dans les voies de transduction du signal. L'interactomique a pour but l'étude de ces mécanismes d'un point de vue global : les interactions entre les différentes protéines du système étudié sont collectées et représentées sous la forme d'un graphe, appelé interactome (voir (27) pour une revue intéressante). Cela permet entre autres d'étudier la structure globale de ces réseaux (études topologiques (28), comparaisons entre différentes espèces (29)), de prédire les fonctions de gènes (30, 16, 31) ou encore, à plus petite échelle, de

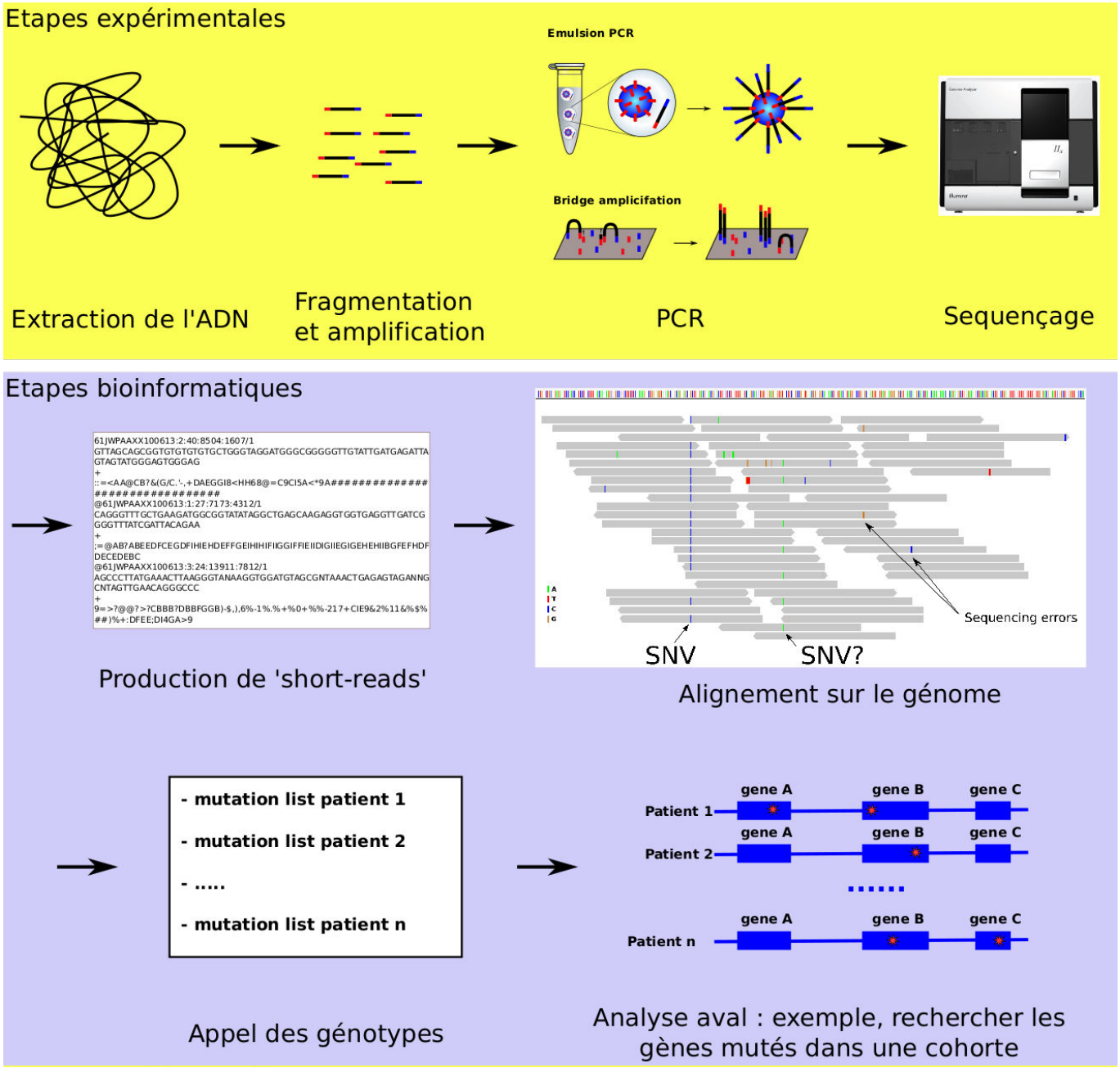


FIG. 1.3 – **Les différentes étapes du séquençage.** L'ADN est d'abord extrait, puis fragmenté. Les fragments sont liés à des adaptateurs et amplifiés, puis isolés et amplifiés une seconde fois avant d'être séquencés. (Le schéma émulsion PCR est adapté de (25)). Le séquençage produit des short-reads, représentés sur le schéma au format fastq (contenant les séquences ainsi que des codes qualité pour chaque base). Les short-reads sont alignés sur le génome de référence (figure réalisée avec IGV (26)) et les variants sont caractérisés, afin d'obtenir une liste de mutations pour chaque individu. Les gènes variants sont ensuite généralement identifiés et prioritarisés.

visualiser l'ensemble des interactions intervenant dans un mécanisme afin d'avoir une meilleure compréhension de son fonctionnement (outils de visualisation tels que Cytoscape (32)). Le terme interactome (33) peut aussi être utilisé dans un sens plus large, regroupant tous les types de graphes d'interactions biologiques - interactions pouvant être des liens physiques, biochimiques ou fonctionnels entre métabolites et/ou macromolécules (27). Je me concentre ici uniquement sur les réseaux d'interactions protéine-protéine, et le terme interactome sera employé dans ce sens dans la suite de mon manuscrit.

1.4.3 Stocker et partager

Afin de pouvoir paralléliser l'analyse des données omiques et comparer les résultats obtenus par différents laboratoires, il est nécessaire de pouvoir partager facilement les données. Cela passe premièrement par le développement de formats de données standards, facilitant le partage des outils bioinformatiques. D'importants efforts ont été faits dans ce sens, aussi bien dans le domaine de la protéomique que de la génétique. Le HUPO-PSI (Human Proteome Organisation - Proteomic Standard Initiative) (34) s'est chargé de tous les aspects protéomiques, proposant des directives, des recommandations, un vocabulaire contrôlé et des formats de données pour représenter les interactions protéine-protéine (PSI-MI), les données de spectrométrie de masse (PSI-MS) et les procédés de séparation (sepCV). Au niveau génomique, des formats standards existent pour représenter différentes étapes de l'analyse. Pour citer quelques-uns des plus utilisés, les formats FASTA ou FASTQ (35) permettent de représenter les séquences brutes, le format SAM (36), ou sa version compressée BAM, permet d'en représenter une version alignée sur le génome de référence et le format VCF (37) est utilisé pour lister les variants.

Dans un deuxième temps se pose la question du stockage et de l'accessibilité des données. En effet, les données omiques peuvent rapidement devenir massives : à titre d'exemple, les données de séquençage de l'exome de 520 tumeurs de l'ovaire et des tissus sains correspondants (5) représentent à elles seules environ 30T de données (au format BAM). Une partie du travail consiste à développer des algorithmes de compression efficaces (38). Par exemple, pour les données de séquence, une solution repose sur le stockage des différences entre la séquence et le génome de référence. D'autres algorithmes exploitent les répétitions présentes dans le génome avec des techniques classiques de compression de texte comme gzip ou la transformée de Burrows-Wheeler (très utilisée par les logiciels d'alignement). Par ailleurs, il est nécessaire de développer des dépôts performants, afin de permettre à l'ensemble de la communauté de stocker, de cataloguer et d'accéder à ces informations. De nombreuses plateformes de stockage et bases de données existent, souvent en partie redondantes. Il peut alors devenir chronophage de collecter l'ensemble des informations relatives à une question particulière et des outils d'intégration ont été développés pour permettre d'effectuer des requêtes simultanées dans plusieurs bases. Par exemple, PSICQUIC (39, 40) est un service web basé sur les formats PSI qui permet d'intégrer les bases de données d'interaction moléculaires. De plus, dans le cas de données humaines, se pose la question de la confidentialité : des protocoles sécurisés et des comités éthiques doivent être mis en place pour assurer le contrôle de l'accès aux données.

1.4.4 Intégrer et visualiser

Il est primordial de pouvoir visualiser ensemble des données de différentes sources afin de formuler de nouvelles hypothèses. Ainsi, de nombreux logiciels ont été développés. L'exemple typique en interactomique est le logiciel très utilisé Cytoscape (32) qui permet de visualiser facilement des réseaux d'interactions protéine-protéine, ainsi que des annotations, des données d'expression, etc. Il permet également de réaliser diverses analyses et a été étendu par le développement de nombreux plugins par la communauté scientifique (41). Au niveau génomique, on peut citer IGV (Integrative Genome Viewer (26)), une interface

permettant par exemple de visualiser les alignements des short-reads sur le génome, ou encore le génome browser d'UCSC (42, 43), permettant de juxtaposer un grand nombre d'annotations prédéfinies le long du génome, ainsi que d'ajouter ses propres données facilement.

1.5 Grands projets, grandes découvertes

Le potentiel de découverte des données omiques a stimulé la création de consortiums internationaux coordonnant les efforts de nombreux laboratoires. Par exemple, une suite naturelle de la publication du génome humain de référence est l'analyse de la diversité génétique de l'espèce humaine. Plusieurs grands projets internationaux d'annotation des polymorphismes ont été entrepris en ce sens. Le projet HapMap (44, 45), dont le but est de cartographier les haplotypes (groupes d'allèles transmis ensembles de génération en génération), a identifié les SNPs (Single Nucleotide Polymorphism) de 4 populations. Dans la même lignée, le 1000 Genomes Project (46) et ESP6500 (Exome Sequencing Project (47)) utilisent les technologies NGS dans le but de déterminer l'ensemble des mutations de nombreux individus. Cependant, alors que le projet HapMap utilise principalement des puces à SNPs, génotypant les variants communs (fréquence supérieure à 5% dans la population), le 1000 Genomes Project et ESP6500 espèrent identifier la majorité des variants rares. Pour ce faire, le 1000 Genomes Project a séquencé l'exome de 900 personnes, séparées en 27 populations d'origines variées. De son côté, l'Exome Sequencing Project (ESP) a séquencé l'exome de 6503 échantillons provenant de cohortes de patients de divers hôpitaux américains. Ces projets ont notamment permis d'identifier de nombreux variants liés à des maladies, grâce à des études d'association (*e.g.* (48, 49) pour les exemples les plus récents).

Initié en 2003, le projet ENCODE (Encyclopedia of DNA Elements (50)) avait pour but de cartographier tous les éléments fonctionnels du génome, en se concentrant sur les parties non codantes. Des techniques aussi diverses que le ChIP-seq, le RNA-seq, et des analyses de la méthylation et de l'hypersensibilité à la DNase I ont été mises en oeuvre pour produire plus de 1000 jeux de données. Un des résultats les plus surprenants serait la mort de la notion d'ADN poubelle (ADN non fonctionnel) : plus de 80% de notre génome serait fonctionnel. Même si ce point est controversé (51), l'importance de l'épigénome dans le fonctionnement des cellules n'est plus à démontrer et cette étude massive a permis d'identifier un grand nombre de ses composants.

Par ailleurs, les technologies NGS ont également permis de grandes avancées dans notre compréhension du développement des cancers : plusieurs consortiums internationaux (4, 52) s'appliquent à caractériser de façon systématique un grand nombre de tumeurs. Pour donner une idée de la taille de ces études, à ce jour, le consortium TCGA (The Cancer Genome Atlas) a analysé de façon exhaustive 27 types de cancer, avec la caractérisation systématique de nombreux échantillons (1007 pour l'étude la plus importante concernant le cancer du sein). De nouvelles cibles thérapeutiques potentielles, ainsi que de nombreux gènes impliqués dans le développement tumoral (gènes 'driver') ont été identifiés. Les gènes 'drivers' affectent principalement la différenciation cellulaire, la survie cellulaire et la réparation de l'ADN, et leurs mutations sont généralement au nombre de 2 à 8 par tumeur (53). Ces mutations s'accompagnent d'un plus ou moins grand nombre (selon le type de cancer) de mutations dites 'passengers', qui ne présentent pas d'avantage sélectif (54).

L'interactome (réseau d'interactions protéine-protéine) a lui aussi été très étudié à haut et moyen débit (27). Par exemple, chez l'homme, la méthode du double hybride a été appliquée sur des protéines de transduction du signal (55), sur les protéines se liant au splicéosome (56), ou pour étudier les interactions avec la méthyltransférase (technique du Y2H-seq, (57)). L'étude de ces réseaux a permis de mettre en évidence des propriétés globales du système, avec en particulier la présence de noeuds (protéines) très fortement

connectés, appelés 'hubs', dont le rôle est controversé (58, 28), ou la découverte de modules fonctionnels, régions très reliées de l'interactome portant sur un fonctionnement cellulaire particulier (27). Les interactomes peuvent aussi être utiles pour prédire la fonction de gènes inconnus, ou leur impact sur des maladies (59), en utilisant le principe de la culpabilité par association ('guilt-by-association') : une protéine interagissant avec des protéines impliquées dans une fonction cellulaire ou une maladie pourrait jouer un rôle similaire. Les maladies peuvent également être étudiées sous l'angle des réseaux : elles sont alors vues comme une perturbation du système global.

1.6 Problématique : biais, incomplétude et erreurs systématiques des données omiques

Bien que très informatives, les données omiques peuvent aussi être trompeuses si elles sont mal utilisées. En effet, ces données comportent de nombreux biais, sont incomplètes et contiennent de nombreux faux positifs. Par exemple, la corrélation entre centralité et létalité, observée pour les réseaux de curation de la littérature (58) ne semble pas être valable pour les interactomes haut débit (28). Les propriétés topologiques des interactomes sont aussi remises en question, car elles reposent sur des échantillons de l'interactome complet (3, 60). Il est également connu que les données NGS contiennent des erreurs systématiques, difficiles à différencier des vrais variants (61). De plus, la qualité des données contenues dans dbSNP, la principale base de données de variants, a été critiquée (62). Il est également utile d'avoir une idée de la taille totale du système étudié afin, d'une part, de pouvoir planifier les expériences en connaissance de cause et d'autre part, d'estimer la validité des conclusions que l'on en tire. Ainsi, de nombreuses études ont été conduites pour évaluer les biais, les taux d'erreur et la complétude des données omiques. Par exemple, le projet MAQC (MicroArray Quality Control) a, dans ses premières phases, estimé la précision et la comparabilité des puces à ADN (63, 64). Lors du passage aux technologies NGS, la phase 3 (appelée SEQC, Sequencing Quality Control), à laquelle j'ai eu la chance d'apporter ma modeste contribution, s'est appliquée à comparer les plateformes de séquençage RNA-seq et les pipelines bioinformatiques (article soumis à *Nature Biotechnology*, voir Annexe 1). Malgré les efforts fournis jusqu'à présent, des biais et faux positifs polluent encore l'information disponible et les pipelines bioinformatiques d'analyse de séquences ou d'interactomes peuvent encore être améliorées.

C'est dans ce cadre que se situe mon travail de thèse. Nous avons analysé plusieurs jeux de données omiques en nous concentrant sur le contrôle de la qualité des données. Nous avons estimé les biais et les taux d'erreur, et proposé des méthodes rigoureuses pour les réduire au maximum afin d'obtenir des données aussi propres que possible. D'une part, nous avons analysé les biais existants dans les données d'interaction protéine-protéine chez *S. cerevisiae* et développé une méthode pour les estomper, basée sur les protéines très étudiées par la communauté scientifique. Cela nous a permis d'estimer de façon fiable la taille de cet interactome. Les résultats de cette étude ont été publiés dans *BMC Bioinformatics*. D'autre part, nous avons étudié des données de séquençage NGS de patientes atteintes du cancer de l'ovaire. Nous avons présenté une méthode originale pour filtrer les erreurs systématiques de séquençage et pour caractériser les génotypes de ces patientes. Nous avons ensuite développé un pipeline pour identifier les SNVs et les gènes potentiellement impliqués dans le cancer de l'ovaire, en nous concentrant sur les variants affectant la fonction des protéines. Nous avons découvert 35 gènes connus pour être impliqués dans le développement du cancer contenant des variants significativement plus présents chez nos patientes que dans la population générale. Par ailleurs, d'autres gènes encore non caractérisés ont été identifiés comme conférant probablement un risque accru au cancer de l'ovaire. Ce travail a donné lieu à un article, qui sera soumis très prochainement (voir section 3.5).

Chapitre 2

Interactomique : de surprenantes découvertes sur les données d'interactions protéine-protéine conduisent à une estimation plus élevée de la taille de l'interactome de *S. cerevisiae*.

Les protéines, principaux acteurs de la cellule, interagissent de façon très diverses pour faire fonctionner la machinerie cellulaire. De ce fait, leurs interactions ont été très étudiées. Plusieurs projets de cartographie d'interactomes ont été entrepris, et une curation intensive de la littérature reporte bon nombre d'interactions dans les bases de données. Cependant, les scientifiques doivent rester vigilants dans l'analyse qu'ils font de ces données : elles sont incomplètes et contiennent de nombreux biais et faux positifs (3). Ces caractéristiques pouvant conduire à une mauvaise interprétation des résultats, surtout dans le cas d'analyse des propriétés globales du système, il est nécessaire de les étudier et de les prendre en compte. C'est dans cette optique que nous avons analysé les données d'interactions protéine-protéine de *S. cerevisiae*. Nous montrons que les jeux de données haut débit et de curation de la littérature sont beaucoup plus corrélés que ce qui est communément admis, mais que se restreindre aux protéines très étudiées par la communauté scientifique permet d'estomper cette corrélation. Nous proposons une estimation de la taille de l'interactome de *S. cerevisiae* basée sur ce sous-ensemble de l'interactome, plus fiable et plus élevée que les estimations existantes. Ces résultats ont été publiés dans *BMC Bioinformatics* en 2010 (section 2.6), et sont résumés ici.

2.1 Données haut débit et données de curation de la littérature

Les données d'interactions protéine-protéine peuvent être classées en deux catégories : les données haut débit (high throughput, HT), issues de cribles systématiques, et les données de curation de la littérature (literature curated, LC). De par les différences intrinsèques à leur génération, ces deux types de données

Jeu de données	Taille	Description
LowBP-LC	6272	Interactions binaires physiques extraites de BioGRID (65)
Ito-Core	834	Interactions HT-Y2H haute qualité de (66)
Ito-Full	4549	Toutes les interactions HT-Y2H de (66)
Uetz-Screen	674	Interactions HT-Y2H de (67)
CCSB-YI1	1809	Interactions HT-Y2H de (28)
Y2H-Union	2903	Union de *Ito-Core*, *Uetz-Screen* et *CCSB-YI1*
Tarassov	2761	Interactions PCA de (16)
HT-Union	5560	Union de *Ito-Core*, *Uetz-Screen*, *CCSB-YI1* et *Tarassov*
LowBP-LC-pre2000	1787	Interactions de *LowBP-LC* identifiées avant 2000

TAB. 2.1 – Description des jeux de données utilisés

sont, d'un certain point de vue, complémentaires. Premièrement, alors qu'un crible haut débit applique une unique méthode de détection à un protéome (ou à un sous-ensemble), une grande variété de méthodes est utilisée à bas débit. Chaque méthode de détection ayant ses propres limitations, des interactions peuvent être difficiles, voir impossibles, à détecter avec certaines méthodes et facilement identifiables avec d'autres. Un exemple typique est la méthode double hybride échouant à détecter les interactions impliquant des protéines membranaires. Les données LC et HT explorent donc deux sous-ensembles différents de l'interactome. Par ailleurs, les études bas débit sont souvent conduites pour tester une hypothèse particulière ("biais d'inspection"), contrairement aux méthodes haut débit qui visent à capturer de façon systématique l'ensemble des interactions. En conséquence, les interactomes HT et LC présentent des topologies très différentes, et résultent de l'échantillonnage de différents sous-espaces de l'interactome complet (28). Beaucoup d'études considèrent les interactions détectées à ce jour comme un sous-ensemble représentatif de l'interactome complet mais les biais susmentionnés font douter de la validité de ces suppositions. Par ailleurs, il est souvent admis que les données HT et LC sont indépendantes.

Afin de vérifier ces hypothèses, nous nous sommes intéressé à caractériser mieux les biais existants dans les données d'interactions protéine-protéine, grâce à l'analyse de 4 jeux de données HT (3 obtenus par Y2H et un par PCA), et des interactions bas débit physiques binaires reportées dans BioGRID (65) (voir Table 2.1).

2.2 Les données LC sont enrichies en interactions facilement détectables par double hybride

En examinant *LowBP-LC* (jeu de données de curation de la littérature, table 2.1), nous avons fait une découverte inattendue : la méthode Y2H, méthode la plus utilisée à haut débit, a également été très utilisée à bas débit. Parmi les interactions de *LowBP-LC*, 53% sont soutenues par Y2H. On s'attend donc à ce que les données LC soient enrichies en interactions détectables par Y2H. Ce résultat est confirmé par notre analyse du jeu de données Y2H produit par Ito et ses collègues (66) : les interactions facilement détectées par cette expérience sont sur-représentées dans les données LC (Figure 2.1).

L'habitude de considérer les jeux de données LC et HT comme des jeux de données orthogonaux et indépendants est donc à remettre en question.

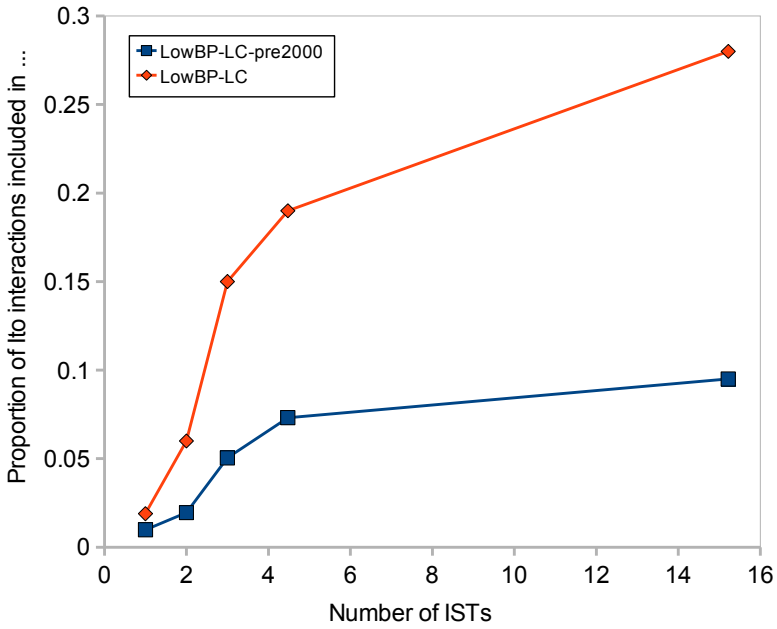


FIG. 2.1 – Lors d'une expérience double hybride reposant sur le criblage d'une banque, chaque interaction, identifiée par un IST (Interaction Sequence Tag), peut être détectée plusieurs fois. Ainsi, le nombre d'ISTs obtenus modélise la facilité d'une interaction à être identifiée par double hybride. Nous avons groupé les interactions de *Ito-Core* par paquets de 200, par nombre d'ISTs croissant (cette donnée n'est pas disponible pour les autres jeux de données). Chaque groupe est représenté par la moyenne du nombre d'ISTs, et par la proportion de ses interactions incluses dans *LowBP-LC*. On remarque que la proportion d'interactions incluses dans *LowBP-LC* est plus importante pour les interactions facilement détectables par double hybride. Comme il est possible que certaines études bas débit faites après la parution de *Ito-Core* aient été conduites pour confirmer ces interactions, nous vérifions que ce phénomène existe pour les interactions de *LowBP-LC* identifiées avant 2000 (année de parution de *Ito-Core*).

2.3 Utiliser le niveau d'étude des protéines permet de réduire la corrélation entre les données LC et HT

La base de données SGD (Saccharomyces Genome Database) met à disposition une table contenant les correspondances entre gènes et articles scientifiques. Cela nous a permis d'étudier les interactomes sous un nouvel angle, en comparant les caractéristiques des interactions impliquant les protéines très et peu étudiées par la communauté scientifique (le niveau d'étude d'une protéine étant modélisé par le nombre d'articles dans lesquels elle a été citée). À ma connaissance, c'est la première fois que le niveau d'étude des protéines a été pris en compte dans l'analyse de données interactomiques (ou dans tout autre type d'analyse). Cela nous a permis d'identifier des propriétés intéressantes des sous-réseaux impliquant les protéines très étudiées.

En premier lieu, nous avons examiné la corrélation entre le niveau d'étude d'une protéine et son degré (le nombre d'interaction dans lesquelles elle est impliquée). Comme on pouvait s'y attendre, le jeu de données LC contient en proportion plus d'interactions pour les protéines ayant été très étudiées ($R^2 = 0.59$, $P = 2 \cdot 10^{-4}$). De façon plus inattendue, ce résultat est vrai également pour les jeux Y2H, ce qui suggère que la densité de l'interactome est plus importante pour les protéines très étudiées. Cela pourrait être dû au fait que les protéines très connectées sont plus sujettes à jouer un rôle important dans des mécanismes cellulaires variés, et donc attirent plus l'attention de la communauté scientifique.

Nous nous sommes ensuite concentrés sur un sous réseau de *LowBP-LC*, appelé "*LowBP-LC* très étudié", contenant uniquement les interactions impliquant des protéines très étudiées (protéines citées dans plus de 100 articles). Nous espérions ainsi obtenir un sous-réseau complet de l'interactome. Pour vérifier cette hypothèse, nous avons estimé son taux de faux négatifs (interactions non annotées), en nous basant sur l'intersection de ce sous-réseau avec les jeux de données HT. Bien loin de confirmer cette supposition, nous nous sommes au contraire rendu compte que ce taux est de l'ordre de 60% : les données d'interactions protéine-protéine sont donc très incomplètes, même pour les protéines très étudiées. Cela est d'autant plus vrai que de par la corrélation entre les données LC et Y2H, ce taux est une sous-estimation du taux de faux négatifs réel.

Même si les données LC concernant les protéines très étudiées sont loin d'être complètes, elles nous ont permis de mettre en évidence un sous-espace de l'interactome détecté uniquement pour les protéines très étudiées. En effet, nous montrons que les données HT couvrent mieux *LowBP-LC* que "*LowBP-LC* très étudié", ce qui pourrait être dû au fait que les études approfondies de certaines protéines ont permis la détection d'interaction difficiles à déceler par les méthodes classiques (section 2.6). Pour vérifier cette hypothèse de façon plus approfondie, nous avons regardé si les méthodes expérimentales utilisées pour détecter les interactions de *LowBP-LC* très étudié sont significativement différentes de celles utilisées pour détecter les autres interactions. Nous montrons que les interactions du sous-réseau très étudié s'appuient moins souvent sur le Y2H (58,4% pour le réseau peu étudié contre 44,7% pour le réseau très étudié, soit une baisse de 13,9%, $P < 2{,}2 \cdot 10^{-16}$), et plus souvent sur des expériences nécessitant un travail de laboratoire intensif, par exemple la détection de phosphorylation ou d'ubiquitination (code BioGRID "biochemical activity", augmentation de 12%, $P < 2{,}2 \cdot 10^{-16}$) ou des expériences in-vitro de purification de protéines (code BioGRID "reconstituted complex", augmentation de 8,5%, $P = 5{,}5 \cdot 10^{-12}$). Par conséquence, ce sous-réseau est plus complet que le sous-réseau peu étudié, et plus représentatif de l'interactome.

2.4 Prendre en compte le niveau d'étude des protéines améliore les estimations de la taille de l'interactome

Malgré les efforts intenses entrepris pour cartographier les interactomes, les bases de données d'interactions protéine-protéine sont loin d'être exhaustives. Les conclusions tirées d'informations partielles peuvent être erronées, et estimer la complétude des données est donc important pour évaluer la pertinence des résultats des études interactomiques. Par exemple, la topologie des réseaux d'interactions actuels, qui suit une loi de puissance (*e.g.* (67, 66)), était considérée comme extrapolable à l'interactome complet. Han et ses collègues (3) ont montré que cette topologie peut provenir de l'échantillonage de réseaux présentant des topologies très diverses (aléatoire, exponentielle, loi de puissance ou normal tronquée), ce qui infirme cette hypothèse. De plus, la corrélation entre centralité et létalité, observée pour les réseaux de curation de la littérature (58) ne semble pas être valable pour les interactomes haut débit (28). Par ailleurs, estimer la taille de l'interactome permet d'évaluer la quantité de travail restant à fournir avant l'obtention d'une cartographie complète et aide à planifier les expériences à venir.

2.4.1 Estimations existantes

Plusieurs estimations de la taille de l'interactome de la levure ont été proposées. L'approche principalement utilisée pour ces estimations est une méthode statistique simple qui modélise la taille de l'intersection entre deux jeux de données par une loi hypergéométrique (68, 69, 70, 71). Ainsi, connaissant la taille de chacun des jeux de données et de leur intersection, on peut facilement déduire la taille de l'espace échantillonné, *i.e.* l'interactome complet. Cela suppose que les deux jeux de données sont échantillonnés indépendamment dans l'interactome. D'autres méthodes plus simples extrapolent le nombre d'interactions d'un sous réseau au réseau complet, et donc n'utilise qu'un seul jeu de données (72, 73). Enfin, une estimation basée sur des retests expérimentaux dans les conditions de production des données a également été proposée (28).

Une des principales limitations de ces estimations est qu'aucune ne prend en compte simultanément les données haut débit (HT) et les données de la littérature (LC) (ou alors un très petit échantillon utilisé comme gold-standard). Ces types de données étant complémentaires, cela constitue une perte importante d'information. De plus, beaucoup d'estimations sont basées uniquement sur deux jeux de données haut débit Y2H : ces estimations prennent donc en compte uniquement le sous-espace des interactions identifiables par Y2H.

En conséquence, les estimations actuelles sont en fait des estimations de la taille d'un sous-espace de l'interactome, et sont donc probablement des sous-estimations. Nous proposons une méthode simple pour estimer la taille d'un interactome, qui prend en compte toutes les données disponibles ainsi que les biais existant entre les données LC et HT grâce à l'information présente dans les sous-réseaux très étudiés. De plus, notre estimation tient compte des faux positifs présents dans les données.

2.4.2 Faux positifs

Au cours de ces dernières années, les études interactomiques se sont appliquées à améliorer la qualité des données produites, mais ces dernières contiennent encore beaucoup de faux positifs. Ces faux positifs doivent impérativement être pris en compte lors de l'estimation de la taille de l'interactome, et le problème est alors d'estimer les taux de faux positifs des jeux de données utilisés. Dans notre étude, les taux de faux positifs du jeu de données LC et d'un jeu de données HT "de référence" sont choisis d'après une revue de la littérature. Ceux des autres jeux HT sont calculés grâce à leur intersection avec les données LC et le jeu HT de référence.

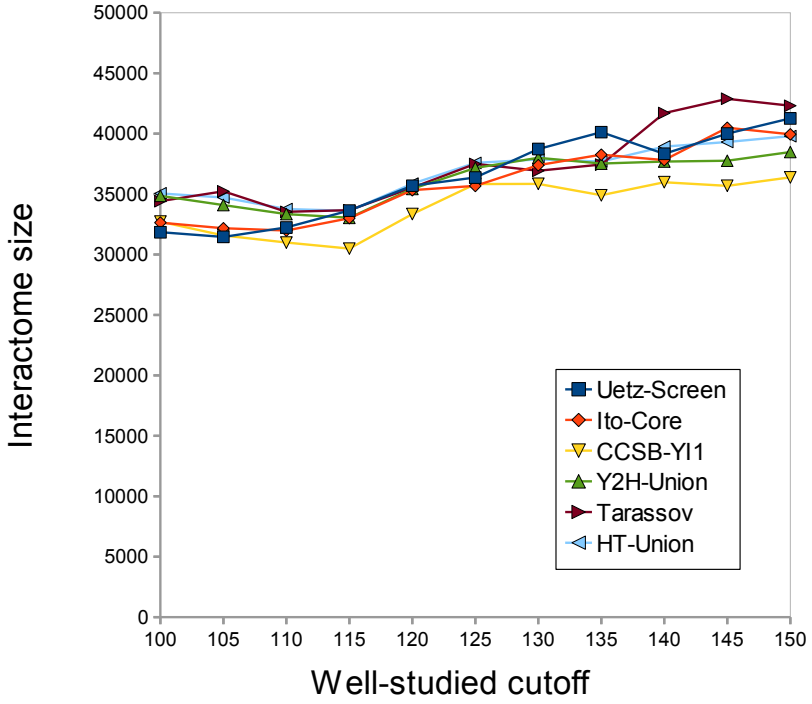


FIG. 2.2 – **Estimation de la taille de l'interactome de la levure.** Le nombre total d'interactions physiques binaires de la levure est estimé en fonction du seuil choisi pour qu'une protéine soit considérée très étudiée, en utilisant différents jeux de données HT.

2.4.3 Amélioration des estimations de la taille de l'interactome de *S. cerevisiae*

Pour estimer la taille de l'interactome de la levure, nous nous inspirons de la principale méthode existante : nous modélisons l'intersection de deux jeux de données par une loi hypergéométrique. Cependant, alors que les estimations actuelles reposent sur deux jeux de données haut débit, typiquement deux jeux Y2H, nous utilisons l'union de toutes les données haut débit d'interactions binaires physiques et un jeu de curation de la littérature. D'une part, cela permet de considérer l'ensemble des données disponibles. D'autre part, ce modèle repose sur l'hypothèse que les deux jeux de données utilisés sont échantillonnés indépendamment dans l'interactome, ce qui est faux dans le cas de données obtenues par la même méthode expérimentale. Des biais d'échantillonnage existent également entre les données HT et LC, mais ils sont moins importants qu'entre deux jeux générés par la même méthode, et l'estimation obtenue en combinant les données HT et LC est donc plus fiable.

Par ailleurs, nous avons montré que les biais d'échantillonnage entre les jeux HT et le sous-ensemble des interactions LC impliquant les protéines très étudiées ("*LowBP-LC* très étudié") sont moins importants qu'avec le jeu de données LC complet. Ainsi, utiliser "*LowBP-LC* très étudié" conduit à une estimation plus fiable de la taille de l'interactome. Cela permet de considérer le sous-espace des interactions difficiles à identifier par les méthodes classiques, puisque ces interactions sont détectées pour les protéines très étudiées. Malgré nos précautions, notre estimation est probablement toujours une sous-estimation, car il existe sans doute un sous-espace non détectable même par les études ciblées et approfondies qui n'est pas pris en compte.

Nous avons appliqué cette méthode à différents jeux de données et - contrairement à la plupart des

études précédentes - les résultats obtenus sont très robustes (Figure 2.2). Nous estimons la taille de l'interactome à au moins 37 600 interactions physiques binaires, confirmant l'idée que les estimations existantes - typiquement autour de 20 000 interactions - sont sous évaluées.

2.5 Conclusions

Nous avons montré que les données haut débit et les données de curation de la littérature sont plus corrélées que ce qui était communément présumé, en particulier car les données LC sont enrichies en interactions facilement détectables par double hybride. Les résultats tirés de ces données doivent donc impérativement être analysés en conséquence : par exemple, l'hypothèse couramment faites d'indépendance des jeux de données LC et HT est fausse. Par ailleurs, nous nous sommes intéressés aux caractéristiques des interactions impliquant les protéines très étudiées par la communauté scientifique. Celles-ci sont plus complètes et qualitativement différentes des interactions entre protéines peu étudiées, notamment car les études approfondies des protéines permettent de détecter des interactions difficile à identifier. Utiliser le sous-réseau impliquant les protéines très étudiées montre que l'interactome de la levure est encore plus incomplet que précédemment estimé. De plus, aucune méthode ne peut détecter toutes les interactions existantes, et les faux négatifs sont inévitables. Par conséquence, l'obtention d'une cartographie complète de l'interactome de la levure nécessite des efforts intensifs, et l'utilisation d'une grande variété de méthode de détection.

2.6 Document joint : New insights into protein-protein interaction data lead to increased estimates of the *S. cerevisiae* interactome size

Sambourg and Thierry-Mieg *BMC Bioinformatics* 2010, **11**:605
http://www.biomedcentral.com/1471-2105/11/605

BMC Bioinformatics

RESEARCH ARTICLE Open Access

New insights into protein-protein interaction data lead to increased estimates of the *S. cerevisiae* interactome size

Laure Sambourg, Nicolas Thierry-Mieg*

Abstract

Background: As protein interactions mediate most cellular mechanisms, protein-protein interaction networks are essential in the study of cellular processes. Consequently, several large-scale interactome mapping projects have been undertaken, and protein-protein interactions are being distilled into databases through literature curation; yet protein-protein interaction data are still far from comprehensive, even in the model organism *Saccharomyces cerevisiae*. Estimating the interactome size is important for evaluating the completeness of current datasets, in order to measure the remaining efforts that are required.

Results: We examined the yeast interactome from a new perspective, by taking into account how thoroughly proteins have been studied. We discovered that the set of literature-curated protein-protein interactions is qualitatively different when restricted to proteins that have received extensive attention from the scientific community. In particular, these interactions are less often supported by yeast two-hybrid, and more often by more complex experiments such as biochemical activity assays. Our analysis showed that high-throughput and literature-curated interactome datasets are more correlated than commonly assumed, but that this bias can be corrected for by focusing on well-studied proteins. We thus propose a simple and reliable method to estimate the size of an interactome, combining literature-curated data involving well-studied proteins with high-throughput data. It yields an estimate of at least 37, 600 direct physical protein-protein interactions in *S. cerevisiae*.

Conclusions: Our method leads to higher and more accurate estimates of the interactome size, as it accounts for interactions that are genuine yet difficult to detect with commonly-used experimental assays. This shows that we are even further from completing the yeast interactome map than previously expected.

Background

As the chief actors within the cell, proteins participate in every cellular process, from metabolism to mechanical structure, immune system or signaling pathways. To successfully fulfill their role, they stably or transiently interact with each other, forming a complex protein interaction network, or interactome. Thus, the comprehensive mapping and deciphering of theses interactomes is a prerequisite for the full understanding of any cellular system. Furthermore, interactomes can be used to infer the function and regulation of novel proteins (*e.g.* Tarassov *et al.* predict that the previously uncharacterized proteins YML018C, YMR221C and YDR119W are involved in autophagy [1]). However, when trying to extract information from protein interaction networks, one must be aware that they are far from comprehensive. Estimating the size of an interactome provides insight into the biological relevance of the conclusions drawn. For example, partial sampling from networks presenting a variety of degree distributions can result in apparent scale-free subnetworks, irrespective of the initial network's topology [2]: topology analyses based on incomplete data may not be valid. Moreover, the number of protein-protein interactions is an important parameter for evaluating the completeness of databases and current high-throughput experiments, in order to measure the remaining efforts and build a framework for future experiments [3,4]. We focus here on *S. cerevisiae*, one of the most studied eukaryotic model

* Correspondence: Nicolas.Thierry-Mieg@imag.fr
Laboratoire TIMC-IMAG, BCM, CNRS UMR5525, Faculté de médecine, 38706 La Tronche cedex, France

BioMed Central

© 2010 Sambourg and Thierry-Mieg; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

organisms and a widely-used test platform for new experimental techniques, in particular for protein-protein interaction (PPI) detection methods.

Available data

The available datasets of protein-protein interactions fall into two categories: literature-curated (LC) and high-throughput (HT). LC data reports manually curated interactions described in the literature, usually obtained by low-throughput experiments [5]. While high-throughput datasets are typically produced by testing all pairs of proteins within a subspace determined solely by the availability of reagents, low-throughput experiments are often hypothesis-driven, for example targeted at proteins involved in a disease or in a particular cellular function. Additionally, both LC and HT data can be of different nature: some assays identify proteins that belong to the same complexes, and find mainly stable but potentially indirect interactions (*e.g.* Affinity purification followed by mass spectrometry [6,7]), while others such as HT-Y2H (high-throughput yeast two-hybrid [8-10]) or PCA (protein complementation assay [1]) search essentially for direct binary interactions that may be transient [11]. Finally, synthetic lethality, genetic suppression and genetic enhancement are examples of genetic interactions, which occur at the phenotypic level and rarely correspond to physical interactions [12]. In this study, we focus on direct binary physical interactions.

Any dataset may contain errors, and particular attention must be paid to false positives (proteins erroneously annotated as interacting). Since interacting proteins in Y2H are not expressed in their natural cellular context, false positives are restricted here to 'technical' false positives that are due to stochastic or systematic detection method artifacts, and we ignore 'biological' false positives where an interaction is indeed physically possible but not biologically relevant (*e.g.* if the proteins are never expressed in the same cellular compartment).

Existing estimates

Since the publication of the first HT-Y2H datasets, several methods for estimating the size of the *S. cerevisiae* interactome have been proposed [5,10,13-18]; it is typically inferred to contain around 20,000 binary interactions, with extreme estimates ranging from 10,000 to 30,500. These methods are often based on analyses of the HT-Y2H genome-wide screens of the yeast interactome [8-10], and can be broadly divided into two categories. A first class involves the study of the overlap between two or more datasets [14-16,19], usually assumed to follow a hypergeometric distribution. Conceptually these methods differ mainly in their choice of datasets and estimations of error-rates. The second class of methods focuses on a single dataset. Two such methods [5,13] are based on an extrapolation of the number of interactions in an HT [13] or LC [5] subnetwork to the total number of yeast proteins. Another approach applied in the paper reporting the latest HT-Y2H dataset [10] relies on the estimation of their assay's characteristics within a sophisticated framework [3]. This provides detailed information but requires intimate knowledge of the dataset and/or performing additional experiments, hence it may be difficult to accomplish outside the laboratory that produced the data. Finally, Huang and coworkers [17,18] adapted capture-recapture theory and applied it using Interaction Sequence Tag (IST) counts. This is an interesting approach but is only applicable to library-screen-based HT datasets where the number of IST hits is available (a single dataset [8] among those considered in this study). Other estimates based on affinity purification-mass spectrometry data [19] have been proposed but these count indirect interactions and, as this work focuses on the binary interactome, are not directly relevant.

To date, most studies have not explicitly and comprehensively taken into account both LC and HT data. One recent method [10] did use a 'positive reference set' derived from LC data to assess the 'assay sensitivity' of their Y2H assay, but this dataset represents only a small sample of the available LC interactions and is focused on high confidence rather than wide coverage. However, recent results demonstrate the radically different view that these data offer. For example, the correlation between centrality and lethality, established in 2001 (Jeong et al. [20]) and considered as a given since then, was based on Uetz [9] and LC [21] data; this correlation does not exist [10] in the *Y2H-Union* dataset (the union of the 3 genome-wide HT-Y2H library screening results [8-10], see Methods, Datasets). One possible explanation lies in the intrinsically different strategies underlying low-throughput and high-throughput data collection (hypothesis-driven versus systematic). Additionally, only Y2H and PCA have been applied in a high-throughput setting whereas a wide variety of detection methods have been used at low-throughput. Thus high-throughput and low-throughput experiments may have explored different subspaces of the interactome: these two data sources appear complementary, and current estimates of the interactome size are questionable because usually based exclusively on one or the other. Finally, LC data includes highly focused and thorough studies of particular proteins, which may have allowed the identification of some interactions that are intrinsically difficult to detect. This has also never been considered.

We propose here a method for estimating the size of an interactome. It is based on dataset overlap, but takes into account both HT and LC data, as well as interactions that are hard to detect by taking advantage of the

extensive literature curation efforts undertaken at SGD (the Saccharomyces Genome Database [22]).

Results

Method overview

Our method is based on a comparison between low-throughput binary physical data curated from the literature (*LowBP-LC*, obtained from the BioGRID database after filtering), and a binary physical high-throughput dataset (*HT-Union*, the union of a PCA [1] and three HT-Y2H [8-10] datasets, see Methods). Assuming that HT interactions are randomly drawn within the interactome, and thus independently of their presence in *LowBP-LC*, allows to estimate the interactome size. Indeed, under this assumption, the number of true positive HT interactions included in *LowBP-LC* follows a hypergeometric distribution $\mathcal{H}(N, m, n)$, with N the total number of genuine interactions, m the number of true positive *LowBP-LC* interactions and n the number of true positive HT interactions. Thus, given an estimation of the false-discovery rate ($FDR = FP/(TP + FP)$ with FP and TP the numbers of false positives and true positives, respectively) of each dataset, one can compute the number of genuine interactions in the whole interactome. This is the basis for most methods relying on the overlap between datasets [14-16,19].

However, all assays have their biases and limitations: some interactions may be easy to detect with one assay and difficult or impossible with another. In addition, most HT datasets were obtained with Y2H, but this assay is also widely used in low-throughput studies - it provides support for 53% of *LowBP-LC* interactions according to BioGrid evidence codes. It follows that *LowBP-LC* is expected to be enriched in interactions that are readily detectable with Y2H. This hypothesis is supported by studying Ito and co-workers' data [8]. Indeed, we used the number of IST hits (interaction sequence tags) for each interaction as an indicator of the difficulty to detect it: interactions with more ISTs are easier to detect, at least in Ito and coworkers' version of the Y2H protocol. We observed that the number of IST hits is clearly correlated with over-representation in *LowBP-LC* (See Figure 1 and Methods). As this phenomenon exists with both *LowBP-LC* and *LowBP-LC-pre2000* (interactions reported before 2000), it is not due to the fact that low-throughput experiments could have been designed to confirm *Ito-Core* interactions (HT-Y2H interactions seen at least 3 times in Ito et al. [8], 2001). In addition, although the lower representation observed for interactions with 1 and 2 IST hits is likely partly due to higher FDRs among these interactions, reported as lower confidence in the original article [8], the coverage by *LowBP-LC* keeps increasing with the number of ISTs for interactions with 3 or more ISTs. These putative interactions - including any false positives among them - are well reproducible in this particular experimental system, hence the FDR is not expected to decrease when the number of ISTs increases. We conclude that the presence of an interaction in *LowBP-LC* is positively correlated with the ease of finding it by Y2H: *LowBP-LC* is indeed enriched in Y2H-strong interactions. Thus the assumption that HT and LC data are independent subsets of the complete interactome does not hold, and the simple dataset overlap method described above leads to underestimating the interactome size.

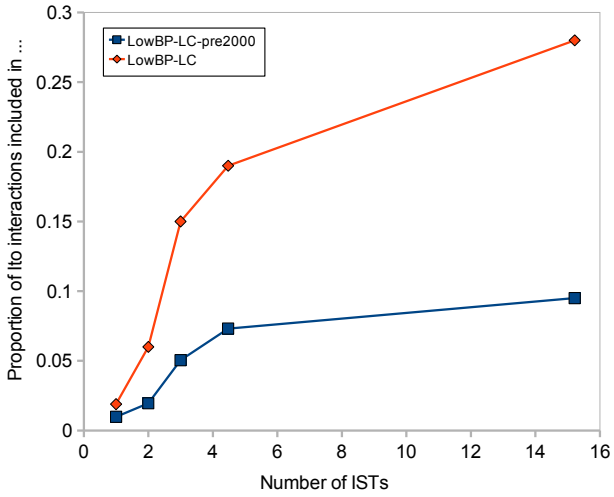


Figure 1 Increased coverage by literature-curated datasets of interactions that are easier to detect by Y2H. The proportion of Ito interactions present in *LowBP-LC* and in *LowBP-LC-pre2000* (literature-curated interactions reported before 2000) is plotted as a function of the number of IST hits. Each point represents at least 200 interactions, and the number of IST hits is the weighted mean for these interactions.

Our method can be summarized as follows. In order to alleviate this problem, we restrict the *LowBP-LC* dataset to interactions involving proteins that have been thoroughly studied: we show that these proteins have likely been subjected to a wider variety of assays, leading to a less biased view of the interactome. We then estimate the FDRs of *LowBP-LC* and of each HT dataset, using dataset overlap to relate the HT FDRs to one another. Finally, we model the number of HT true positives included in *LowBP-LC* restricted to well-studied proteins by a hypergeometric distribution $\mathcal{H}(N, m', n)$, with N and n as described above and m' the number of true positive *LowBP-LC* interactions involving well-studied proteins (equation (5)). This leads to an estimation of the interactome size N.

Taking into account how thoroughly proteins have been studied

We examined the relation between a protein's degree (*i.e.* the number of interactions it is involved in) and

how thoroughly it has been studied, modeled as the number of papers in which the protein has been cited (according to the Saccharomyces Genome Database [23], see Methods). This revealed a strong correlation between these two quantities for the *LowBP-LC* dataset (Figure 2a): as expected, literature curation has reported many more interactions for highly studied proteins than for poorly studied ones. More surprisingly, a small but significant correlation also exists for the *Y2H-Union* dataset (Figure 2b). We see no reason why a proteome-wide Y2H screen would identify a larger proportion of the interactions that can be established by well-studied proteins, therefore this observation suggests that the density of the complete interactome is higher for well-studied proteins than for poorly studied ones. The statistical test is inconclusive with the *Tarassov* data (Figure 2c). Another unexpected observation is that even for well-studied proteins, *LowBP-LC* data are far from comprehensive: based on the available HT data for these proteins, we estimate the false negative rate ($FNR = FN/(TP + FN)$ with TP and FN the numbers of true positives and false negatives) of *LowBP-LC* restricted to well-studied proteins at approximately 60% (see Methods and Tables 1 and 2).

Well-studied data comprise interactions that are difficult to detect

A closer look at the interaction data concerning well-studied proteins leads to another surprising discovery: HT data covers *LowBP-LC* much better than it does *LowBP-LC* restricted to interactions involving well-studied proteins (Figure 3). Note that this is not due to the fact that *LowBP-LC* has a better coverage of the complete interactome restricted to well-studied proteins: indeed, the completeness of *LowBP-LC* should not affect the proportion of its interactions that are present in an independent subset of the interactome. Thus, we see only two possible explanations.

First, this could be simply because the rate of false positives in *LowBP-LC* increases when restricting this dataset to well-studied proteins. Cusick et al. [24] recurated 100 literature-curated yeast interactions, which allows us to invalidate this hypothesis: for these interactions, we found that false positives are not over-represented among *LowBP-LC* interactions involving well-studied proteins (well-studied interactions represent 21.4% of the false positives and 22% of the true positives, see Methods).

As an alternative explanation, we propose that in-depth studies discover interactions that are difficult to detect by most widespread methods, hence are under-represented in HT datasets. To test this hypothesis, we examined whether the experimental methods used to demonstrate *LowBP-LC* well-studied interactions

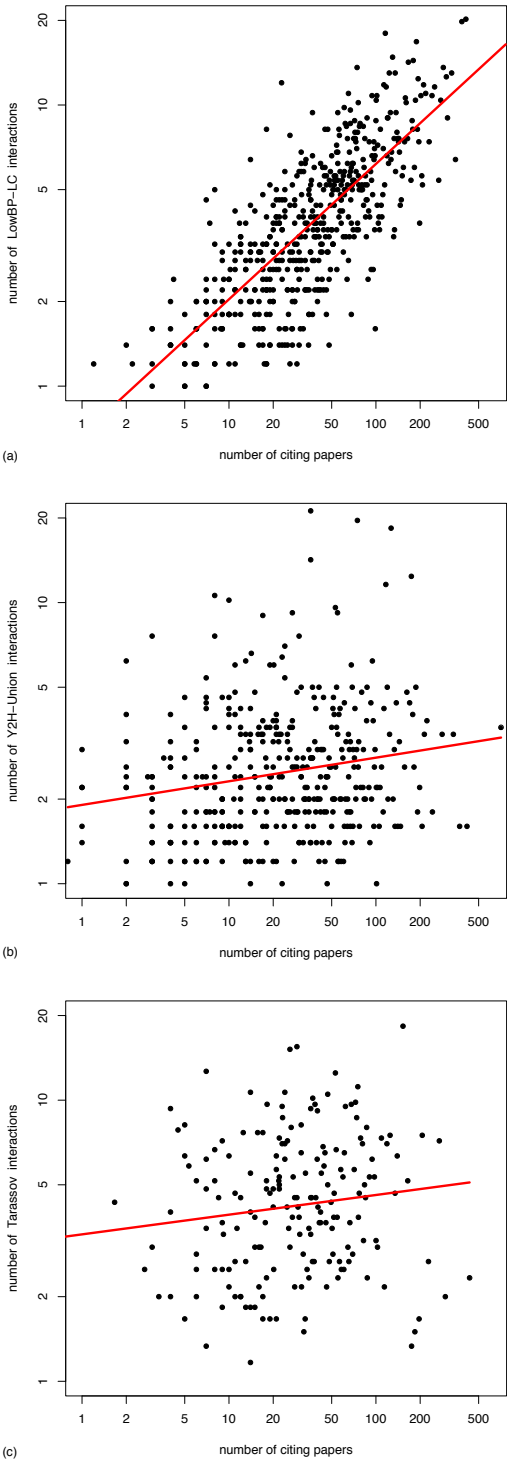


Figure 2 Relation between the level of study and the degree of proteins in various datasets. Log-log scale linear regression between the number of interactions (in the indicated dataset) involving a protein and the number of papers referencing that protein, using binned data (each point represents 5 proteins). (a) *LowBP-LC* interactions, $R^2 = 0.59$, $P = 2 \cdot 10^{-103}$, slope = 0.48. (b) *Y2H-Union* interactions, $R^2 = 0.04$, $P = 1.0 \cdot 10^{-4}$, slope = 0.08. (c) *Tarassov* interactions, $R^2 = 0.01$, $P = 0.07$, slope = 0.07.

Table 1 Estimated false negative rate of LowBP-LC restricted to interactions involving well-studied proteins.

Well-studied cutoff	*Uetz-Screen*	*Ito-Core*	*CCSB-YI1*	*Y2H-Union*	*Tarassov*	*HT-Union*
100	0.64	0.66	0.66	0.68	0.63	0.67
105	0.63	0.65	0.65	0.67	0.61	0.66
110	0.64	0.66	0.65	0.67	0.61	0.66
115	0.65	0.66	0.64	0.67	0.61	0.66
120	0.66	0.65	0.63	0.66	0.62	0.65
125	0.66	0.66	0.65	0.67	0.61	0.66
130	0.63	0.49	0.63	0.63	0.62	0.63
135	0.60	0.49	0.62	0.61	0.62	0.62
140	0.60	0.47	0.62	0.61	0.64	0.62
145	0.62	0.45	0.61	0.60	0.64	0.61
150	0.63	0.45	0.61	0.61	0.64	0.62

The false negative rate is computed separately with each high-throughput dataset, using a cutoff to consider proteins well-studied ranging from 100 to 150 and a reference FDR for *CCSB-YI1* set at 0.25.

differed significantly from those used to demonstrate other *LowBP-LC* interactions, using the BioGrid experimental evidence codes. We observed that interactions in the well-studied subset are less frequently supported by Y2H (down 13.9% from 58.6% to 44.7%, p-value < 2.2e-16), while they are significantly more frequently supported by biochemical activity assays such as those detecting phosphorylation or ubiquitination (Biochemical Activity, up 12.4% from 11.1% to 23.5%, p-value < 2.2e-16), as well as in vitro assays using purified proteins (Reconstituted Complex, up 8.5% from 33.5% to 42%, p-value = 5.5e-12). Thus well-studied proteins have more often been subjected to labor-intensive interaction detection methods, which may allow the detection of a wider variety of interactions. To sum up, this supports the hypothesis that literature-curated interaction data involving well-studied proteins comprise interactions that, although genuine, are difficult or impossible to detect using labor-efficient methods such as Y2H.

Taking into account the level of study of proteins may thus allow to account for these interactions, hence lead to more accurate estimates of the size of an interactome.

LowBP-LC false positives

Literature-curated data has been commonly assumed of excellent quality, but a recent study showed that curation errors may not be so infrequent. Cusick *et al.* [24] recurated 100 yeast interactions supported by a single paper, assigning a confidence score to each. They reported that 35% of these interactions were erroneous and that 40% could be not verified. For this study, we considered that among *LowBP-LC-Unique* (interactions from *LowBP-LC* supported by a single paper, and not found in the HT dataset), 35% were false positives. The initial report has been debated [25,26] and this may be an overestimate, which would result in our underestimating the interactome size. Interactions reported in more than one paper, or also detected by an HT experiment, were considered true positives.

Table 2 Influence of the CCSB-YI1 FDR on the LowBP-LC well-studied false negative rate.

CCSB-YI1 FDR	*Uetz-Screen*	*Ito-Core*	*CCSB-YI1*	*Y2H-Union*	*Tarassov*	*HT-Union*
0.15	0.67	0.66	0.65	0.68	0.57	0.66
0.25	0.63	0.62	0.61	0.64	0.53	0.62
0.35	0.59	0.57	0.57	0.58	0.48	0.56

The false negative rate of *LowBP-LC* restricted to interactions involving well-studied proteins is computed with the different datasets, when the *CCSB-YI1* FDR ranges from 0.15 to 0.35, using a well-studied cutoff set at 125.

HT false positives

The initial mistrust of HT-Y2H assays was largely based on an analysis [27] benchmarking HT datasets against a

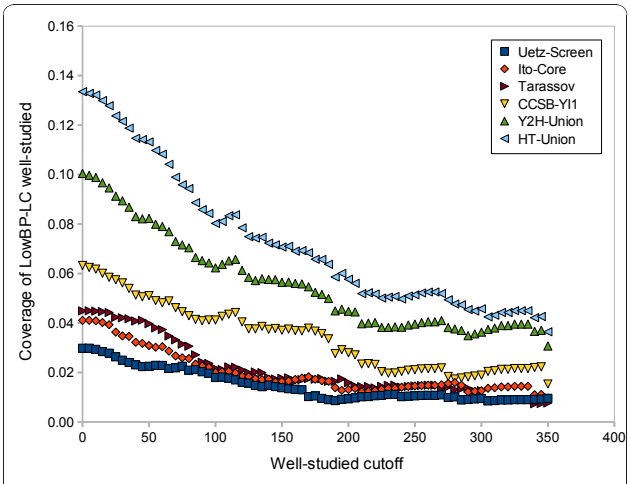


Figure 3 Coverage of *LowBP-LC* well-studied by each high-throughput dataset. The proportion of *LowBP-LC* interactions involving well-studied proteins that are covered by each HT dataset is plotted as a function of the 'well-studied cutoff', *i.e.* the minimum number of papers referencing a protein for it to be considered well-studied.

set of protein complexes expanded with the matrix model, and does not seem relevant anymore [10,18]. Indeed, after the publication of the first HT-Y2H datasets, several methods estimated their FDRs at ~ 50% (*e.g.* [14,16]). However, by retesting their own data with orthogonal assays, Yu *et al.* [10] have estimated the FDR of *CCSB-YI1*, their proteome-wide HT-Y2H dataset, at 0-6%, and showed that *Uetz-Screen* (the Uetz *et al.* HT-Y2H library screening result [9]) and *Ito-Core* are also of high quality. Based on the capture/recapture method, Huang *et al.* [18] have evaluated the FDR of *Ito-Full* to 26%. *Ito-Full* is comprised of all interactions from Ito *et al.* [8] including those reported as low confidence in the original publication, and is known to have the lowest quality (*e.g.* [10,14,28]). As there is no consensus on the order of magnitude of these FDRs, we decided to apply our method with different FDR values. The *CCSB-YI1* FDR is taken ranging from 15% to 35% and the other HT FDRs are computed as described below.

We developed a simple method for comparing the FDRs of high-throughput datasets, based on the hypothesis that the *LowBP-LC* coverage of HT true positives is the same for each HT dataset (see Methods). Under this assumption, we established a simple relation between the FDRs of HT datasets (Methods, equation (1)). However, if some low-throughput experiments were performed to verify interactions reported in high-throughput datasets, an important bias may favor older datasets, which will 'artificially' have more interactions in common with *LowBP-LC*. This problem can be addressed by restricting *LowBP-LC* to interactions reported before 2000 (the publication date of the oldest HT dataset), yielding another dataset called *LowBP-LC-pre2000*. In fact, *Ito-Core* and *Uetz-Screen* (published in 2001 and 2000) have a higher proportion of interactions in common with *LowBP-LC* than *CCSB-YI1* (published in 2008), whereas with *LowBP-LC-pre2000*, the proportions are similar (Table 3). We therefore used *LowBP-LC-pre2000* to estimate the HT FDRs. For example, assuming a *CCSB-YI1* FDR of 25%, FDRs of Y2H datasets range from 15% to 25% (Table 4).

Likewise, historical reasons may favor Y2H over PCA. Indeed, Y2H was proposed in 1989 [29], and has been widely used in low-throughput experiments, whereas PCA was first described in 2000 [30]. We cannot correct for this bias because restricting *LowBP-LC* to interactions reported before 1989 yields a very small dataset. As a consequence the FDR of 73% that can be computed for *Tarassov* (PPIs detected by high-throughput protein complementation assay [1]) may be largely overestimated and is only a rough upper bound.

Table 4 Estimated false discovery rate of each high-throughput dataset.

	CCSB-YI1	*Uetz-Screen*	*Ito-Core*	*Tarassov*	*Ito-Full*
FDR	0.25	0.15	0.21	0.73	0.76

The FDRs are computed with eq (3), setting the *CCSB-YI1* FDR at 0.25. As discussed, the FDR that can be computed for the *Tarassov* dataset is a rough upper bound.

Estimating the interactome size

Starting with the number of *LowBP-LC* interactions involving well-studied proteins (2572 interactions), we removed the expected number of false positives (35% of *LowBP-LC-Unique*). We then calculated on the one hand the number of interactions, all considered as genuine, in the intersection between the *LowBP-LC* well-studied subset and the HT dataset (144 interactions for *HT-Union*, see Table 5 for the other datasets), and on the other hand the estimated number of true positives in the whole HT dataset, taking into account HT false positives by using the HT FDRs estimated as described above and assuming an FDR of 25% for *CCSB-YI1* (~ 2814 true positives in *HT-Union*, see Table 5 for the other datasets). Taken together, this allows to estimate the size of the binary yeast interactome at ~ 37, 600 interactions (95% confidence interval: 32252-43472, constructed with the normal approximation method [31]). Details on the calculation are provided in Methods.

The *LowBP-LC* well-studied subset was defined with a cutoff (number of referencing papers for a protein to be considered well-studied) of 125 papers, which seems a good compromise between the number of proteins in the subset and how thoroughly they have been studied (Figure 4). The choice of this cutoff or even changes in the HT datasets have little influence on the estimate: it varies between 30,500 and 43,000 interactions, with a cutoff ranging from 100 to 150 and using all the different HT datasets, either singly or merged (Figure 5). Because of the *LowBP-LC* /HT correlation, which is likely still present even when using the well-studied subset of *LowBP-LC*, the results presented here may be underestimated. Obviously, increasing the estimated HT FDRs decreases the interactome size (Figure 6), and more precise results could be obtained with better estimates of these FDRs.

Table 3 Proportion of HT interactions included in LowBP-LC-pre2000 and LowBP-LC for the different datasets.

	Uetz-Screen	*Ito-Core*	*CCSB-YI1*	*Y2H-Union*	*Tarassov*	*Ito-Full*
LowBP-LC-pre2000	0.0831	0.0767	0.0734	0.0634	0.0264	0.0235
LowBP-LC	0.2017	0.2254	0.1617	0.1601	0.0746	0.0637

Table 5 Calculation steps leading to the interactome size. The well-studied cutoff is set at 125 papers and the CCSB-YI1 FDR at 0.25.

	Uetz-Screen	*Ito-Core*	*CCSB-YI1*	*Y2H-Union*	*Tarassov*	*HT-Union*
LowBP-LC well-studied size				2572		
LowBP-LC well-studied TPs	1905.95	1908.4	1911.55	1916.45	1909.45	1922.75
HT TPs	572.4	654.2	1349.3	2171.8	746.2	2814
HT∩*LowBP-LC* well-studied	30	35	72	112	38	144
Estimated size	36366	35670	35822	37163	37494	37574

By and large, our estimates are higher than previous ones, which is reasonable as we used all available datasets and took advantage of their complementarity, and we accounted for interactions that are difficult to detect.

Discussion

As mentioned in the introduction, several methods based on dataset overlap have been proposed for estimating the yeast interactome size [14-16]. The main differences between these methods lie in the error-rate estimations and in the datasets used. While Grigoriev and co-workers [15] consider that false positives and false negatives compensate each other, d'Haeleseer and Church [16] estimate false-discovery rates thanks to the overlap of two HT datasets with a reference LC dataset, and Sprinzak and co-workers' FDR estimation [14] is based on co-localization data. In our method, a reference FDR for one dataset was chosen following a review of the literature, and the overlap between high-throughput and literature-curated data is used to derive the FDRs of other HT datasets from the reference FDR, somewhat similarly to d'Haeleseer and Church. Another important factor for this class of methods lies in the choice of datasets, beyond the necessity of selecting appropriate data (e.g. genetic interactions or co-complex membership may not be directly relevant when studying binary physical interactions). While considering only HT datasets [15] restricts the estimation to interactions that can be detected with the HT method, using a gold standard reference set that is assumed error-free [14,16] is also problematic. In our method carefully selected LC and HT data are combined, taking into account error-rate estimations for each dataset.

The main advantages of our method are the following. First and foremost, by leveraging the available knowledge of how extensively proteins have been studied, our method accounts for interactions that are genuine yet difficult to detect with commonly-used experimental assays. This significantly increases the predicted interactome size, and has never been taken into account. Secondly, it is applicable to any dataset or union of datasets, and it allows to use most of the available data independently of the experimental detection methods.

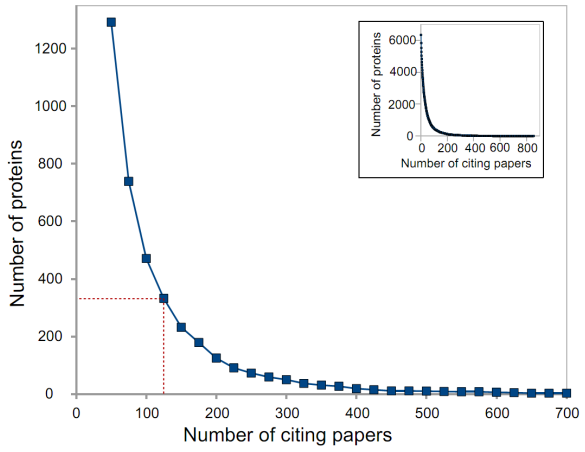


Figure 4 Number of well-studied proteins. The number of proteins in the well-studied subset is plotted as a function of the well-studied cutoff value. The main figure is restricted to proteins cited in at least 50 papers, while the inset shows the complete graph (starting at one paper). The well-studied cutoff value is the minimum number of papers referencing a protein, for this protein to be considered well-studied.

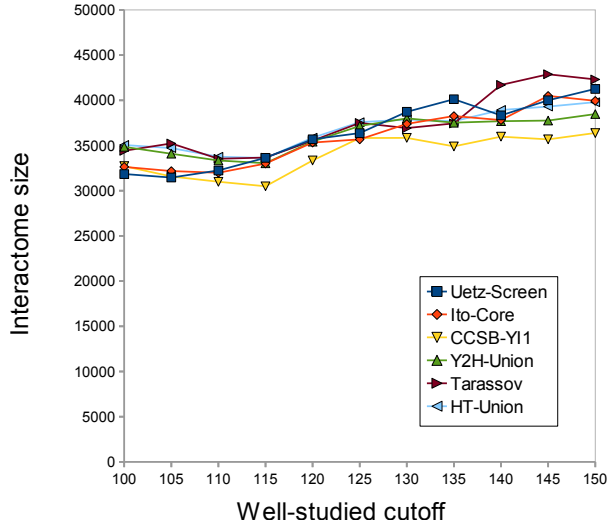


Figure 5 Estimated size of the yeast interactome. The predicted number of binary physical protein-protein interactions that can occur in *S. cerevisiae* is plotted as a function of the well-studied cutoff value, using each high-throughput dataset and a *CCSB-YI1* FDR of 0.25. The well-studied cutoff value is the minimum number of papers referencing a protein, for this protein to be considered well-studied.

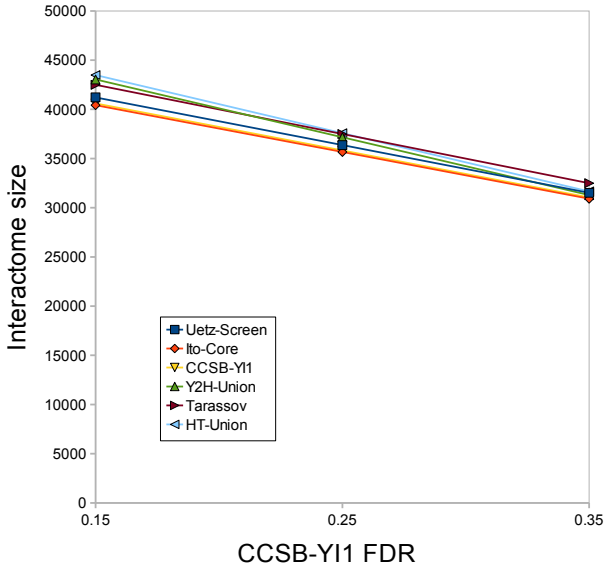


Figure 6 Influence of the *CCSB-YI1* FDR on the estimated interactome size. The predicted size of the *S. cerevisiae* interactome is plotted using each high-throughput dataset, when the *CCSB-YI1* FDR ranges from 0.15 to 0.35. The well-studied cutoff (number of papers for a protein to be considered well-studied) is set at 125 papers.

Thus, the estimates are easy to update when new datasets become available. Furthermore, our model does not directly rely on a gold standard (*i.e.* a subset assumed to contain only true positives), which can be difficult to construct and can introduce biases of its own. Likewise, as no dataset is error-free, it is important to consider error rates of both HT and LC datasets.

We have also shown that well-studied proteins appear capable of establishing more interactions than poorly studied ones (Figure 2b). This probably stems from the fact that well connected proteins are more likely to play important roles in diverse cellular functions, and therefore attract more attention from the community. Our method inherently takes into account this bias. In addition, our method is robust with respect to the choice of HT datasets. Contrary to other estimates [13,14], which increase by 90% and 66% when substituting datasets (respectively *Ito-Full* for *Uetz* and *Uetz* for *Ito-Core*), ours only changes by at most 15% when using different Y2H datasets (at any given well-studied cutoff). Even when comparing estimates based on data obtained by very different assays (Y2H and PCA), the variation remains low (20%). Lastly, the results presented here are for *S. cerevisiae*, but our method could be applied to other organisms, as long as a genome-wide screen as well as significant literature curation have been performed. A potential weakness of our method is that it relies on overlap between datasets that can be small, which may affect the robustness of the estimates.

Conclusion

In this work, we have analyzed HT and LC data while considering how thoroughly each protein has been studied. This has provided novel insight into existing interactome datasets: on the one hand, well-studied proteins seem capable of establishing more interactions than poorly studied ones, and on the other hand, in-depth studies of these well-studied proteins have allowed to identify interactions that are difficult to detect. Together with the combined use of LC and HT data, these observations allow to accurately estimate the interactome size. Our results show that the size of interactomes tend to be underestimated, as previous estimates are usually based on only one source of data and do not take into account interactions difficult to detect. No high-throughput technique can detect all interactions, and false negatives are unavoidable [32]. As a consequence, a variety of methods must be considered when working with interactome mapping, and new strategies such as prioritization and smart-pooling should be employed [4,33,34]. Extensive efforts will be required before an interactome map can be called 'complete', and until then biological conclusions based on the analysis of available data must be drawn with care.

Methods

Datasets

LowBP-LC contains 6, 272 low-throughput binary physical interactions gathered from BIOGRID-ORGANISM-Saccharomyces_cerevisiae-3.0.64.tab (downloaded from the BioGRID website) [35]. All papers referencing more than 100 interactions were considered as high-throughput, and their interactions were excluded. Among the remaining interactions, only binary physical data was kept, *i.e.* interactions whose detection method was by Reconstituted Complex, Two-hybrid, Far Western, Biochemical Activity, Co-crystal Structure, Protein-peptide, PCA or FRET (fluorescence resonance energy transfer).

Ito-Core [8], *Uetz-Screen* [9], *CCSB-YI1* and *Y2H-Union* [10] are HT-Y2H datasets: *Ito-Core* contains the interactions seen at least 3 times by Ito *et al.*, *Uetz-Screen* is the Uetz *et al.* genome-wide library screening result, and *Y2H-Union* is the union of these two datasets with *CCSB-YI1* [10]. All these Y2H datasets were downloaded from the Center for Cancer Systems Biology website [36]. *Ito-Full* contains all interactions from Ito *et al.* [8]. It was downloaded from the Ito Laboratory website [37]. *Tarassov* are the PPIs detected by high-throughput protein complementation assay [1] (provided as supplementary material in the original publication). *HT-Union* contains all interactions from all HT datasets.

The level of study of a protein is modeled by the number of papers in which it has been cited, computed from a table of associations between literature and genes

(downloaded from the Saccharomyces Genome Database [23] on 2010/05/03). Comparing HT FDRs requires to restrict the *LowBP-LC* dataset to interactions reported before 2000 in *LowBP-LC-pre2000*. *LowBP-LC-Unique* are interactions from *LowBP-LC* supported by a single paper, and not found in the considered HT dataset.

Additional file 1 presents the number of interactions and unique proteins in each dataset and intersection of datasets. All datasets are provided in Additional file 2.

The false positive rate of *LowBP-LC* does not depend on the level of study

Cusick *et al*. recurated 100 literature-curated yeast interactions, assigning confidence score for each one: 0 for no confidence, 1 for low confidence or unsubstantiated and 2 for substantiated or high confidence. We therefore considered interactions with a score of 0 to be false positives, and those with a score of 2 to be true positives. We then computed the proportion of these interactions that involve well-studied proteins for each category. Among the 35 false positive interactions and the 25 true positives, respectively 21.4% and 22% involve a well-studied protein.

LowBP-LC false negatives

Hypothesizing that HT well-studied and *LowBP-LC* well-studied are independent allows to estimate the expected number of genuine interactions involving well-studied proteins, and thus the *LowBP-LC* well-studied false negative rate:

$$FNR_{LowBP-LC_{WS}} = 1 - \frac{TP_{HT_{WS} \cap LowBP-LC}}{TP_{HT_{WS}}}$$

with $TP_{HT_{WS}}$ the estimated number of true positives in HT_{WS}, the HT dataset restricted to interactions involving well-studied proteins, and $TP_{HT_{WS} \cap LowBP-LC}$ the number of true positives within the intersection between HT_{WS} and *LowBP-LC*.

A relation between HT FDRs

To decrease the potential correlation between *LowBP-LC* and older HT-Y2H datasets due to recent studies that could have been designed to confirm HT interactions, the *LowBP-LC* dataset used for the FDR calculations contains only interactions reported in publications published before 2000 (publication date of the oldest HT dataset). Consider two HT datasets, denoted 1 and 2 (*e.g. Ito-Core* and *CCSB-YI1*), each partitioned into three subsets A, B and C, respectively the true positives included in *LowBP-LC-pre2000*, the true positives not included in *LowBP-LC-pre2000* and the false positives. We consider that HT interactions also present in *LowBP-LC-pre2000* are true positives (because detected by two independent methods). Therefore, *LowBP-LC-pre2000* and C are disjoint. Hypothesizing that the proportion of true positive HT interactions in *LowBP-LC-pre2000* is independent of the HT dataset yields:

$$\frac{A_1}{B_1} = \frac{A_2}{B_2}.$$

The proportion of HT interactions included in *LowBP-LC-pre2000* ($A/(A + B + C)$) can be easily computed, and denoting α as

$$\frac{A_1}{A_1 + B_1 + C_1} = \alpha \cdot \frac{A_2}{A_2 + B_2 + C_2},$$

we obtain a relation between the false-discovery rates of the two datasets, defined as $FDR = \frac{C}{A+B+C}$

$$FDR_1 = \alpha \cdot FDR_2 + 1 - \alpha. \quad (1)$$

In the rest of this work, we always use *CCSB-YI1* for set 2.

Computing the interactome size

Parameters

- HT : the *HT* dataset used.
- *Well-studied cutoff*: number of papers referencing a protein to consider it well-studied.
- $FDR_{Y\ I1}$: the *CCSB-YI1* FDR, required to compute the FDRs of other HT datasets.

Abbreviations and notations

- *WS*: well-studied.
- $TP_{Dataset}$: estimated number of true positives in *Dataset*.
- $|Dataset|$: size of Dataset.
- *Is*: Interactome size.

HT true positives

- The FDR of *Ito-Core, Uetz-Screen* and *Tarassov* is calculated from the FDR of *CCSB-YI1* as described in Methods, A relation between HT FDRs:

$$FDR_{HT} = \alpha \cdot FDR_{YI1} + 1 - \alpha$$

- The number of HT true positives is then computed as follows:

$$TP_{HT} = |HT| - \sum |HT_i| * FDR_{HT_i} \quad (2)$$

where HT_i iterates over the datasets making up HT for union datasets (*e.g.* for *Y2H-Union*: *Ito-Core*, *Uetz-Screen* and *CCSB-YI1*), or HT itself for individual datasets such as *Ito-Core*.

LowBP-LC true positives

$$TP_{LowBP\text{-}LC_{WS}} = \left|LowBP\text{-}LC_{WS}\right| - 35\% \cdot \left|LowBP\text{-}LC\text{-}Unique_{WS}\right| \quad (3)$$

Where $LowBP\text{-}LC\text{-}Unique_{WS}$ contains *LowBP-LC* interactions involving well-studied proteins, supported by a single paper and not in the HT dataset.

True positives in the intersection

All interactions in the intersection between HT and *LowBP-LC* are considered true positive, so:

$$TP_{HT \cap LowBP\text{-}LC_{WS}} = \left|HT \cap LowBP\text{-}LC_{WS}\right|. \quad (4)$$

Interactome size

The hypergeometric assumption discussed in Results, Method overview leads to:

$$Is = \frac{TP_{HT} \cdot TP_{LowBP-LC_{WS}}}{TP_{HT \cap LowBP-LC_{WS}}} \quad (5)$$

with TP_{HT}, $TP_{LowBP-LC_{WS}}$ and $TP_{HT \cap LowBP-LC_{WS}}$ computed as described above (equations (2), (4) and (4)).

This can be expanded to:

$$Is = \frac{TP_{CCSB\text{-}YI1} \cdot TP_{LowBP\text{-}LC_{WS}} \cdot \left|HT \cap LowBP\text{-}LC\text{-}pre2000\right|}{\left|CCSB\text{-}YI1 \cap LowBP\text{-}LC\text{-}pre2000\right| \cdot \left|HT \cap LowBP\text{-}LC_{WS}\right|}$$

This expanded form allows to study the influence of the various parameters. All relevant scripts are distributed under the GNU General Public License in Additional file 2.

Presence of 'Y2H-strong' interactions in *LowBP-LC*

To examine whether interactions that are more easily detected in Y2H are also overrepresented in *LowBP-LC*, we gathered *Ito-Full* hits and binned them by increasing number of ISTs, each bin containing at least 200 interactions. Each bin is represented by the weighted mean of the number of ISTs, and the proportion of interactions present in *LowBP-LC*. In order not to separate interactions with the same number of ISTs, some bins (particularly single hits) are larger than others. This analysis is performed both with the complete *LowBP-LC* and with *LowBP-LC-pre2000* (*LowBP-LC* interactions reported before 2000)(Figure 1).

Additional material

Additional file 1: Number of interactions and proteins in each dataset. Additional file 1 presents the number of interactions and unique proteins in each dataset and intersection of datasets.

Additional file 2: Datasets and scripts. Additional file 2 is an archive that includes all scripts, distributed under an open source license, as well as all datasets used in this study.

Abbreviations

PPI: protein-protein interaction; LC: literature-curated; HT: high-throughput; Y2H: yeast two-hybrid; PCA: protein complementation assay; FDR: false-discovery rate; IST: interaction sequence tag.

Acknowledgements

This work was supported by a grant from the Region Rhone-Alpes (to NTM).

Authors' contributions

NTM designed the study. LS implemented the method and performed the analyses. Both authors drafted and revised the manuscript. They have read and approved its final version.

Received: 11 June 2010 Accepted: 21 December 2010
Published: 21 December 2010

References

1. Tarassov K, Messier V, Landry CR, Radinovic S, Molina MMS, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW: **An in vivo map of the yeast protein interactome.** *Science* 2008, **320**:1465-1470.
2. Han JDJ, Dupuy D, Bertin N, Cusick ME, Vidal M: **Effect of sampling on topology predictions of protein-protein interaction networks.** *Nat Biotechnol* 2005, **23**:839-844.
3. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabasi AL, Vidal M: **An empirical framework for binary interactome mapping.** *Nat Methods* 2008, **6**:83-90.
4. Schwartz AS, Yu J, Gardenour KR, Finley RL Jr, Ideker T: **Cost-effective strategies for completing the interactome.** *Nat Methods* 2009, **6**:55-61.
5. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D, Andrews B, Boone C, Troyanskya OG, Ideker T, Dolinski K, Batada NN, Tyers M: **Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae.** *J Biol* 2006, **5**:11.
6. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier MA, Homan V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G: **Proteome survey reveals modularity of the yeast cell machinery.** *Nature* 2006, **440**:631-636.
7. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF: **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.** *Nature* 2006, **440**:637-643.
8. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, **98**:4569-4574.
9. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403**:623-627.

10. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet AS, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabási AL, Tavernier J, Hill DE, Vidal M: **High-quality binary protein interaction map of the yeast interactome network.** *Science* 2008, **322**:104-110.
11. Vinayagam A, Stelzl U, Wanker EE: **Repeated two-hybrid screening detects transient protein-protein interactions.** *Theor Chem Acc* 2010, **125**:613-619.
12. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, Prinz J, St Onge RP, VanderSluis B, Makhnevych T, Vizeacoumar FJ, Alizadeh S, Bahr S, Brost RL, Chen Y, Cokol M, Deshpande R, Li Z, Lin ZY, Liang W, Marback M, Paw J, San Luis BJ, Shuteriqi E, Tong AH, van Dyk N, Wallace IM, Whitney JA, Weirauch MT, Zhong G, Zhu H, Houry WA, Brudno M, Ragibizadeh S, Papp B, Pal C, Roth FP, Giaever G, Nislow C, Troyanskaya OG, Bussey H, Bader GD, Gingras AC, Morris QD, Kim PM, Kaiser CA, Myers CL, Andrews BJ, Boone C: **The Genetic Landscape of a Cell.** *Science* 2010, **327**:425.
13. Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C: **Estimating the size of the human interactome.** *Proc Natl Acad Sci USA* 2008, **105**:6959-6964.
14. Sprinzak E, Sattath S, Margalit H: **How Reliable are Experimental Protein-Protein Interaction Data?** *J Mol Biol* 2003, **327**:919-923.
15. Grigoriev A: **On the number of protein-protein interactions in the yeast proteome.** *Nucleic Acids Res* 2003, **31**:4157-4161.
16. D'haeseleer P, Church GM: **Estimating and improving protein interaction error rates.** *Proc IEEE Comput Syst Bioinform Conf* 2004, 216-23.
17. Huang H, Jedynak BM, Bader JS: **Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps.** *PLoS Comput Biol* 2007, **3**:e214.
18. Huang H, Bader JS: **Precision and recall estimates for two-hybrid screens.** *Bioinformatics* 2009, **25**:372-378.
19. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7**:120.
20. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411**:41-42.
21. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: **DIP: the database of interacting proteins.** *Nucleic Acids Res* 2000, **28**:289-291.
22. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe TY, Schroeder M, Weng S, Botstein D: **SGD: Saccharomyces genome database.** *Nucleic Acids Res* 1998, **26**:73.
23. **SGD project. "Saccharomyces Genome Database".** [http://downloads.yeastgenome.org/literature_curation/gene_literature.tab].
24. Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, Rual JF, Borick H, Braun P, Dreze M, Vandenhaute J, Galli M, Yazaki J, Hill DE, Ecker JR, Roth FP, Vidal M: **Literature-curated protein interaction datasets.** *Nat Methods* 2009, **6**:39-46.
25. Salwinski L, Licata L, Winter A, Thorneycroft D, Khadake J, Ceol A, Aryamontri AC, Oughtred R, Livstone M, Boucher L, Botstein D, Dolinski K, Berardini T, Huala E, Tyers M, Eisenberg D, Cesareni G, Hermjakob H: **Recurated protein interaction datasets.** *Nat Methods* 2009, **6**:860-861.
26. Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, Rual JF, Borick H, Braun P, Dreze M, Vandenhaute J, Galli M, Yazaki J, Hill DE, Ecker JR, Roth FP, Vidal M: **Addendum: Literature-curated protein interaction datasets.** *Nat Methods* 2009, **6**:934-935.
27. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417**:399-403.
28. Lee I, Date SV, Adai AT, Marcotte EM: **A probabilistic functional network of yeast genes.** *Science* 2004, **306**:1555-1558.
29. Fields S, Song O: **A novel genetic system to detect protein-protein interactions.** *Nature* 1989, **340**:245-246.
30. Michnick SW, Remy I, Campbell-Valois FX, Vallee-Belisle A, Pelletier JN: **Detection of protein-protein interactions by protein fragment complementation strategies.** *Methods Enzymol* 2000, **328**:208.
31. Sahai H, Khurshid A: **A note on confidence intervals for the hypergeometric parameter in analyzing biomedical data.** *Comput Biol Med* 1995, **25**:35-38.
32. Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, De Smet AS, Venkatesan K, Rual JF, Vandenhaute J, Cusick ME, Pawson T, Hill DE, Tavernier J, Wrana JL, Roth FP, Vidal M: **An experimentally derived confidence score for binary protein-protein interactions.** *Nat Methods* 2009, **6**:91-97.
33. Xin X, Rual JF, Hirozane-Kishikawa T, Hill DE, Vidal M, Boone C, Thierry-Mieg N: **Shifted Transversal Design smart-pooling for high coverage interactome mapping.** *Genome Res* 2009, **19**:1262.
34. Aryee MJA, Quackenbush J: **An Optimized Predictive Strategy for Interactome Mapping.** *J Proteome Res* 2008, **7**:4089-4094.
35. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Res* 2006, **34**:D535.
36. **Center for Cancer Systems Biology.** [http://interactome.dfci.harvard.edu/S_cerevisiae/index.php?page=download].
37. **Ito Laboratory.** [http://itolab.cb.k.u-tokyo.ac.jp/Y2H/full_data.txt].

doi:10.1186/1471-2105-11-605
Cite this article as: Sambourg and Thierry-Mieg: New insights into protein-protein interaction data lead to increased estimates of the *S. cerevisiae* interactome size. *BMC Bioinformatics* 2010 **11**:605.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Chapitre 3

Génotyper les individus à partir de données next-generation et filtrer les erreurs systématiques

Le consortium TCGA (The Cancer Genome Atlas) s'applique à caractériser un grand nombre de cancers au niveau génomique, transcriptomique et épigénétique. Nous nous sommes intéressés à l'étude du cancer ovarien de type séreux de haut grade, une analyse intégrative de séquençage d'exomes, de variations du nombre de copies des gènes, d'expression et de méthylation, réalisée sur 489 patientes (5). Entre autres, cette étude a permis d'identifier 9 gènes "driver" mutés de façon somatique (mutation présente dans le tissu cancéreux et absente du tissu sain) dans une proportion statistiquement significative de patientes : TP53, BRCA1, BRCA2, RB1, NF1, FAT3, CSMD3, GABRA6 et CDK12 (74, 75). Comme dans beaucoup d'études sur la génétique des cancers (*e.g.* (76, 77)), les mutations germinales (ou mutations constitutives) sont ignorées, les scientifiques préférant se concentrer sur les mutations somatiques, plus faciles à identifier. Les exomes des tissus sains ont été séquencés à haute couverture, mais ont été utilisés uniquement comme référence pour appeler les mutations somatiques, et les génotypes des tissus sains n'ont même pas été caractérisés (David Wheeler, Baylor College of Medicine, communication personnelle). Cette pratique est regrettable car les mutations germinales sont connues pour leur importance en oncologie, notamment pour prédire les prédispositions génétiques des patients (78), et pour leur impact sur la résistance aux thérapies ciblées (79). Ainsi, nous avons décidé d'analyser les mutations germinales des patientes atteintes du cancer de l'ovaire séquencées par TCGA. La première étape a naturellement été d'identifier les SNVs présents chez ces patientes, et nous avons donc développé un pipeline d'appel de génotypes fiable, qui cherche à identifier et écarter les erreurs systématiques.

3.1 Alignement et appel des génotypes, des méthodes en développement

En génomique, l'analyse bioinformatique des données est tout aussi importante que leur production. Les traitements à effectuer sont coûteux en temps de calcul et en mémoire, et la présence de nombreuses erreurs de séquençage nécessite d'être particulièrement vigilant dans le tri et le contrôle qualité. Comme expliqué dans l'introduction (Figure 1.3), la caractérisation d'un exome passe par plusieurs étapes expéri-

mentales produisant des millions de short-reads. Ces short-reads doivent ensuite être alignés sur le génome de référence, afin d'identifier les variations, en général les SNVs et les indels, entre le génome étudié et la référence (appel des génotypes). Nous nous sommes plus particulièrement intéressés au problème de l'appel des génotypes.

L'appel de génotypes consiste à classer les loci du génome étudié en homozygotes références, hétérozygotes ou homozygotes variants, en se basant sur la proportion de reads variants alignés sur le locus. Ces proportions devraient théoriquement valoir 0 (homozygote référence), $0,5$ (hétérozygote) ou 1 (homozygote variant), mais la présence d'erreurs de séquençage ou d'alignement, ainsi que l'échantillonnage dans le cas des hétérozygotes, les fait dévier et peut rendre difficile la différenciation entre les vrais variants et les erreurs. Les méthodes pour répondre à ce problème sont souvent classées en deux catégories (7) : les méthodes heuristiques (80), qui utilisent des seuils fixés, et les méthodes statistiques, qui font l'hypothèse que la proportion de reads variants suit une loi binomiale (81, 36, 82, 36).

Bien qu'aussi vieux que le développement des technologies NGS, ce problème n'a pas encore de solution consensuelle, et les solutions proposées nécessitent d'être améliorées. La preuve en est que les quelques études qui comparent les résultats des aligneurs et des méthodes d'appel de génotypes les plus utilisés trouvent peu de concordance dans les variants appelés. Une étude de O'Rawe et ses collègues (83) compare 5 pipelines d'appel de génotype sur 15 exomes provenant de 4 familles séquencées à haut débit (Illumina) et trouve 57% de SNVs concordants et $26,8$% d'indels concordants. Dans cette étude, les variants sont considérés concordants s'ils sont appelés de façon identique par les 5 pipelines. Altmann et ses collègues (84) ont exécuté plusieurs pipelines d'alignement suivis de SAMtools ou de GATK pour l'appel des génotypes sur l'exome d'un individu séquencé par le 1000 Genomes Project. Ils trouvent une concordance de 85% en comparant les méthodes de génotypes utilisées avec le même aligneur, et encore 85% de concordance entre 2 pipelines d'alignement différents suivis du même appel de génotype. Étudier des données contenant autant de faux positifs n'est pas prudent, surtout dans le cas de données aussi massive que celles auxquelles nous nous intéressons. Cela nous a conduit à étudier de plus près ces erreurs pour essayer de mieux les filtrer, notamment en comparant les reads alignés sur les deux brins de l'ADN. Nous avons également développé une variation de la méthode heuristique, plus flexible et mieux paramétrée que les méthodes existantes, grâce à des comparaisons des nombreux réplicats techniques réalisés par le consortium TCGA.

3.2 Choix d'un pipeline de séquençage : le logiciel MAGIC et le projet SEQC

La première étape de notre analyse consiste à aligner des short-reads sur le génome de référence (GRCH37/hg19). Pour réaliser cette étape, nous avons utilisé le logiciel MAGIC (6), sur lequel j'ai travaillé dans le cadre du projet SEQC, lors d'un séjour de 6 mois à Toronto (CCBR, Bader Lab) et à Washington (NIH, NCBI). Les principaux résultats de ces analyses sont décrits dans un article annexe 5 en révision à Nature Biotechnology, dont je suis co-auteur, et sont résumés brièvement ici.

Le but du projet SEQC est d'évaluer les performances des plateformes RNA-seq sur des jeux de données de référence (benchmark), en comparant les résultats obtenus par plusieurs centres de séquençage et pipelines bioinformatiques RNA-seq, par qPCR et par puces à ADN. Les jeux de données de référence, constitués de mélanges en proportions connues d'ARN humain bien caractérisés, ont été séquencés dans plusieurs centres par les technologies Illumina HiSEQ 2000 et SOLiD 5500. Les short-reads produits ont été alignés sur différents modèles de gènes : $85,9$% de ces derniers s'alignent sur RefSeq, 92% sur GENCODE et $97,1$% sur AceView. Dans le même ordre d'idée, le nombre de gènes et de jonctions exon-exon détectés étant plus important pour la base de données AceView, c'est celle-ci qui a été sélectionnée pour

la suite de l'analyse. Cette étude a ensuite évalué les performances de 3 pipelines dans la découverte de nouvelles jonctions exon-exon : MAGIC, r-make (85) et Subread (86). 2,9 millions de nouvelles jonctions ont été identifiées, mais seulement 23% d'entre elles systématiquement par les 3 méthodes. Cependant, 173 jonctions nouvellement détectées, parmi lesquelles 136 étaient découvertes par les 3 pipelines, ont été testées par qPCR : toutes celles trouvées par les 3 pipelines et 80% des autres ont été confirmées. Cela montre que malgré les annotations déjà très complètes existantes dans AceView, de nouvelles formes alternatives de gènes sont encore à découvrir. Par ailleurs, plusieurs mesures permettant d'effectuer un contrôle qualité ont été définies afin d'évaluer les performances des différents pipelines. Par exemple, le taux d'expression d'un gène dans chacun des échantillons doit théoriquement être cohérent avec les proportions des différents mélanges. Dans l'ensemble, on observe une excellente concordance dans les mesures d'expressions différentielles entre les différents centres de séquençage et les différentes technologies (Illumina et SOLiD), en particulier pour les gènes fortement exprimés. Par contre, mesurer l'expression absolue des gènes, même avec un benchmark précis, reste difficile. Enfin, par rapport aux pipelines bioinformatiques testés, MAGIC a montré d'excellente performances dans l'identification de gènes et de transcrits alternatifs. En résumé, cette étude constitue une étape importante dans le développement du RNA-seq : la quantité de données produites et la qualité des analyses contribueront à comprendre la puissance et les limites du RNA-seq.

J'ai contribué à cette étude en participant au développement de la plateforme MAGIC, un des principaux outils utilisés pour les analyses. J'ai en particulier réalisé de nombreux tests qui m'ont amenée à rechercher et corriger des bugs. J'ai également participé à l'amélioration de la documentation et de l'interface utilisateur ainsi qu'à la conception de nouvelles fonctionnalités du logiciel. Ce travail a été initié lors d'un séjour à Washington, au NCBI, en novembre 2011, et notre collaboration s'est poursuivie dans la suite de ma thèse.

Les short-reads du projet "ovaire" ont donc été alignés avec le logiciel MAGIC, qui implémente une version améliorée de l'algorithme "hash and extend". Les short-reads sont indexés (hashed) par des 16-mers, puis le génome est scanné et les short-reads sont ancrés au génome selon les correspondances exactes de ces 16-mers. D'autres cibles sont également scannées, comme par exemple le génome mitochondrial et un génome "imaginaire", construit par complémentarité des bases sans inverser l'ordre et utilisé pour mettre au point les filtres qualités. Enfin, les alignements sont étendus en autorisant les indels et les mésappariements (mismatches), et un score d'alignement est calculé en ajoutant 1 pour chaque correspondance et en enlevant 4 pour chaque différence. Les meilleurs alignements pour lesquels on aligne au moins 70% du read ou au moins 130 bp sont conservés. Dans le cas de séquençage "paired-end", les reads provenant des paires sont alignés indépendamment puis la compatibilité des paires est vérifiée. Les paires alignées de manière incompatible sont supprimées. Les reads qui passent ces filtres qualités sont considérés comme "correctement alignés". Une des caractéristiques originales de MAGIC est le calcul de score de qualité des bases a posteriori, plutôt que l'utilisation des scores de qualité fournis a priori par les séquenceurs, jugés peu pertinents. Pour chaque expérience, MAGIC calcule un score de qualité par position dans le read, basé sur le nombre de mésappariements à cette position sur l'ensemble de l'expérience. Les comptages reportés à chaque position génomique sont modulés par ce score (typiquement autour de 80% du nombre réel de reads). Ce pipeline est disponible sur le site AceView du NCBI, à l'onglet Downloads, Software (ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/Software/Magic).

3.3 Obtenir des génotypes fiables à partir des données NGS

Les résultats décrivant notre pipeline de génotypage et nos filtres d'erreurs systématiques sont détaillés dans un article qui sera soumis très prochainement (section 3.5). La section ci-dessous résume les principales contributions de ce travail et développe la partie pré-traitement des données qui n'a pas été présentée

BCM SOL	35		
BI ILL	3	0	
WUGSC	0	19	51
	BCM SOL	**BI ILL**	**WUGSC ILL**

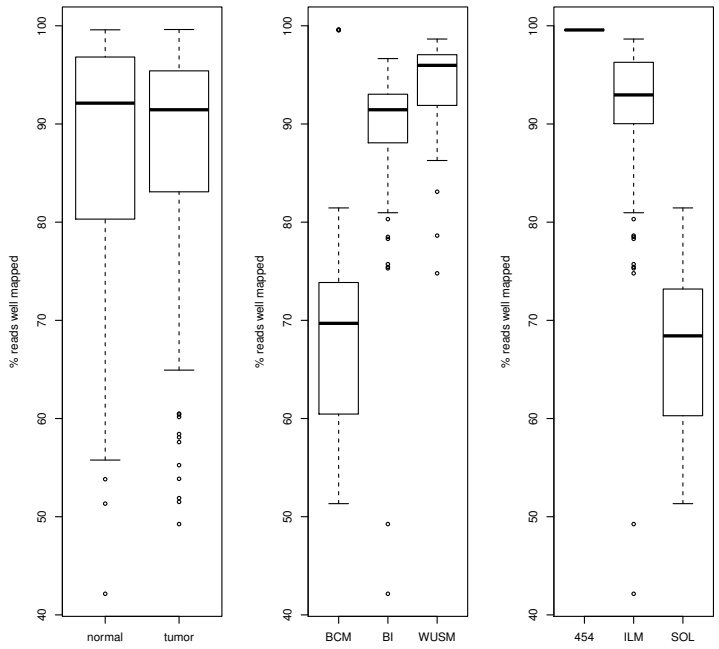
TAB. 3.1 – **Nombre de réplicats réalisés intra et inter laboratoires.** BCM : Baylor College of Medicine, BI : Broad Institute, WUGSC : Washington University Genome Sequencing Center, SOL : séquençage SOLiD, ILL : séquençage Illumina

dans l'article.

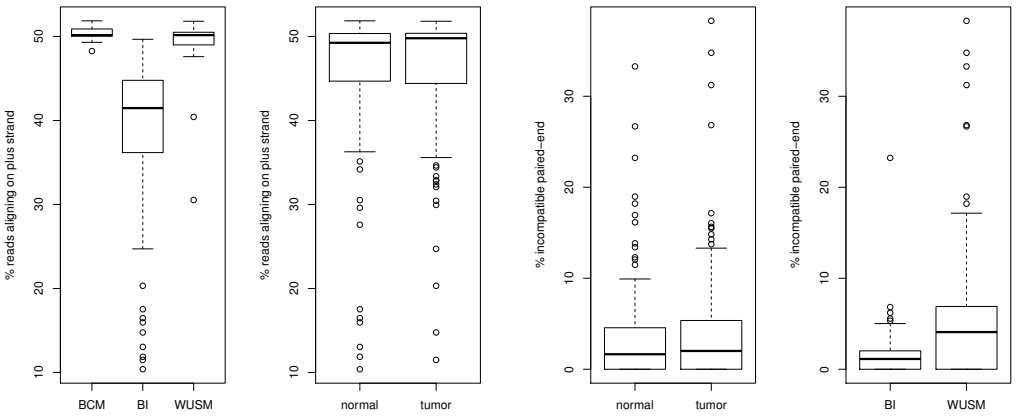
3.3.1 Données utilisées et pré-analyse

Les exomes de nombreuses patientes ont été séquencés par 3 laboratoires, Baylor College of Medicine, Broad Institute et Washington University Genome Center, et certains échantillons ont été séquencés dans plusieurs laboratoires, ou plusieurs fois dans le même laboratoire (réplicats). Afin de mettre au point notre méthode, nous avons étudié un sous-ensemble des données TCGA, contenant uniquement des réplicats. Cela permet d'avoir un contrôle qualité fiable : la reproductibilité du résultat. La table 3.1 liste le nombre de réplicats analysés dans cette étude pour chaque plateforme et laboratoire.

Les short-reads provenant de chaque "run" (expérience de séquençage) ont été aligné avec le pipeline MAGIC décrit ci-dessus. Pour avoir une idée de la qualité *a priori* des données et pour vérifier que les alignements se déroulent correctement, pour chaque "run", nous nous sommes intéressés à plusieurs statistiques : le pourcentage de reads correctement alignés, le pourcentage de reads alignés sur le brin plus, et le pourcentage de reads paired-end incompatibles. Le pourcentage de reads correctement alignés donne une idée de la qualité générale de l'expérience (du "run"), et les reads alignés sur chaque brin reflètent de potentiels biais (de capture, par exemple). Quant à l'incompatibilité des reads paired-end, elle peut indiquer des erreurs d'alignement ou de séquençage : les paires de short-reads devraient s'aligner à une distance raisonnable l'une de l'autre (de 20 à 500 bp), et sur des brins opposés. Une incompatibilité trahit un réarrangement chromosomique ou une erreur. Nous avons regroupé les différents runs en plusieurs catégories, par laboratoire, par plateforme (SOLiD ou Illumina), par type d'échantillon (tissu sain ou tissu cancéreux) et par individu, et nous avons réalisé des analyses de variance multiple (MANOVA) afin d'étudier l'influence de chacune des catégories sur les statistiques. Le laboratoire a des effets significatifs pour chacune des 3 statistiques et la plateforme influence le pourcentage de reads bien alignés. Les autres catégories n'ont pas d'impact significatif sur les alignements. La figure 3.1 montre les boîtes à moustaches réalisées pour chacun des effets observés. Notamment, on observe que la compatibilité des reads paired-end est meilleure pour les short-reads du Broad Institute que pour ceux de Washington University. Il est intéressant de remarquer que il n'y a pas de différence entre les échantillons tumoraux et sains, ce qui montre que les nombreux SNVs et réarrangements présents dans les tumeurs ne gênent pas l'alignement. Dans l'ensemble, la qualité du séquençage et de l'alignement est satisfaisante. Au total, 3137 Gb (giga paires de bases) ont été alignées, avec une couverture moyenne de 323x.



(a) Reads correctement alignés



(b) Reads alignés sur le brin plus

(c) Paires de reads incompatibles

FIG. 3.1 – **Analyse des résultats d'alignement**. Les boîtes à moustaches montrent l'influence des catégories sur chacune des statistiques. Chaque point représente les résultats d'un "run", c'est-à-dire une expérience de séquençage réalisée sur un échantillon. Les boîtes contiennent les 2^{e}et 3^{e}quartiles. Les moustaches s'étendent au point le plus extrême jusqu'à $1,5$ fois la taille de la boîte.

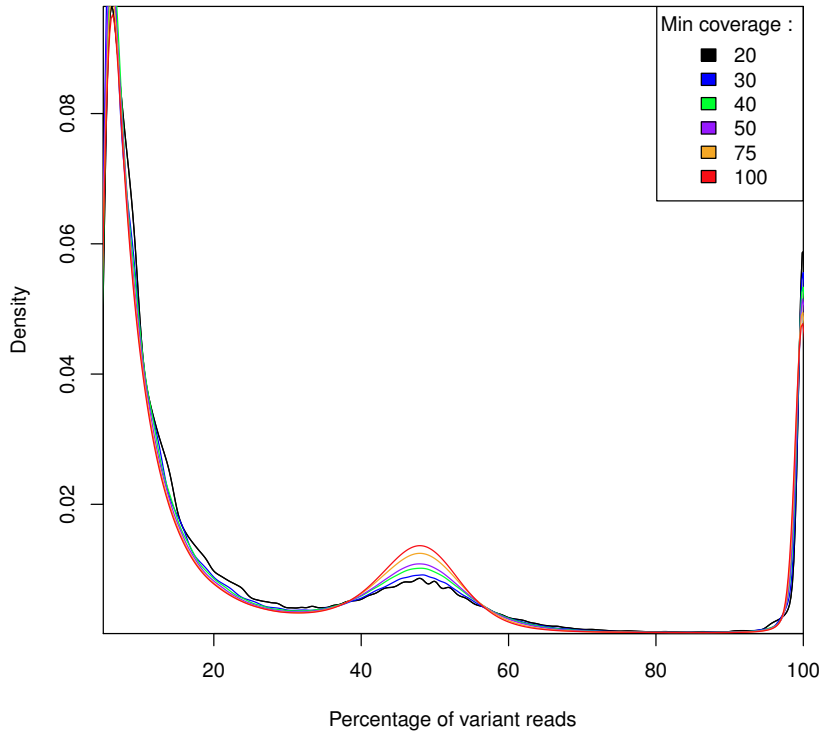


FIG. 3.2 – Densité des proportions de reads variants, pour différentes valeurs de couverture minimale. Comme les homozygotes références évidents ne nous intéressent pas, les positions représentées sont restreintes à celles contenant au moins 5% de reads variants, ce qui correspond à 83 millions de loci.

3.3.2 Une méthode efficace pour appeler les génotypes et filtrer les erreurs systématiques

Choisir une méthode d'appel de génotype

Une fois les alignements effectués, il faut identifier les positions contenant un SNV, et leur assigner un génotype. L'appel des génotypes reposant sur la proportion de reads variants à la position considérée, il est important de regarder la répartition globale de ces proportions avant de choisir une méthode d'appel (Figure 3.2). Comme attendu, on observe des pics à 0, 50 et 100% de reads variants, correspondant respectivement à des positions homozygotes références, hétérozygotes ou homozygotes variantes. La démarcation entre hétérozygotes et homozygotes variants est très claire, et les séparer ne pose pas de problème. Par contre, entre les homozygotes références et les hétérozygotes, on observe une zone allant de 20 et 40% de reads variants, dans laquelle il est impossible de faire un appel en se basant uniquement sur la proportion de reads variants. Nous proposons une méthode d'appel de génotype qui améliore la méthode des seuils afin de prendre en compte ce phénomène, et qui permet de filtrer de nombreuses erreurs systématiques, en comparant les reads alignés sur chaque brin.

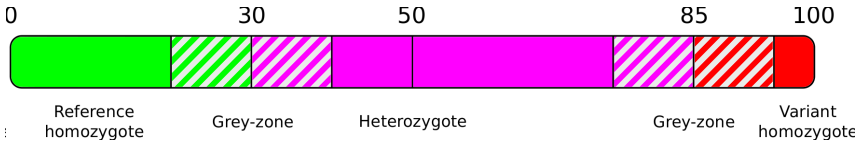


FIG. 3.3 – **Appel des génotypes.** Cette figure illustre l'appel des génotypes en fonction du pourcentage de reads variants. Les "zones grises" sont centrées autour de 30 et 85% de reads variants. Des appels indépendants sont réalisés sur chaque brin, le génotype du brin pouvant être homozygote, hétérozygote ou "zone grise". Le génotype général est appelé si les génotypes des deux brins sont les mêmes, ou si l'un des deux est dans une zone grise adjacente.

Appel des génotypes

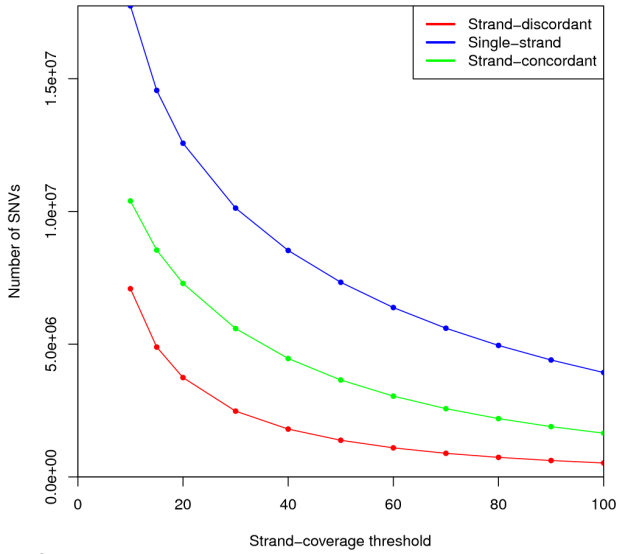
La méthode des seuils nécessite de fixer des limites dans les proportions de reads variants : habituellement, une position ayant entre 20 et 80% de reads variants est appelée hétérozygote. La figure 3.2 montre que ces seuils ne sont pas adaptés à nos données, et suggère d'utiliser 30 et 85%. Cependant, l'utilisation d'un seuil fixe ne permet pas de prendre en compte l'incertitude sur les positions contenant entre 20 et 40% de reads variants. Pour ce faire, nous avons défini une "zone grise" (Figure 3.3) autour des seuils. Des appels indépendants sont réalisés sur chaque brin, c'est-à-dire en considérant uniquement les reads alignés dans le sens du brin concerné, et le génotype du brin peut être homozygote, hétérozygote ou "zone grise". Ce génotype est appelé uniquement si la couverture sur le brin est supérieure au seuil de couverture minimale par brin. Finalement, le génotype général est appelé si les génotypes des deux brins sont les mêmes, ou si l'un des deux est dans une zone grise adjacente. La taille des zones grises et la couverture minimale par brin sont des paramètres de la méthode que nous choisissons en comparant les réplicats (voir plus bas).

La section suivante explique notre choix d'appeler les génotypes indépendamment sur chaque brin : nous montrons que ceux-ci sont trop souvent en désaccord, indiquant la présence d'erreurs systématiques. Faire des appels par brin permet de filtrer ces erreurs. Nous montrons également que les positions pour lesquelles la couverture est insuffisante sur un des deux brins sont très nombreuses, et que l'appel à ces positions n'est pas fiable. Nous conseillons par conséquence de les filtrer.

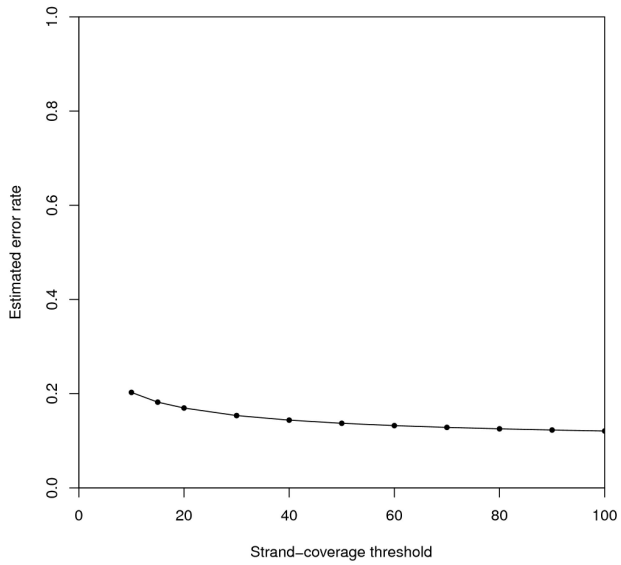
Une histoire de sens

Les erreurs systématiques peuvent être présentes à forte couverture, et avec une grande proportion de reads variants (61). Celles-ci ne peuvent donc pas être filtrées en se basant uniquement sur les proportions de reads variants. Heureusement, la plupart de ces erreurs ont la particularité de dépendre du contexte : par exemple, les séquenceurs Illumina (GAII) vont avoir tendance à souvent se tromper à la position T du motif GGT (61). Cette propriété peut être utilisée se débarrasser de ce type d'erreurs : elles vont *a priori* se produire uniquement dans le sens contenant le motif, et les reads alignés sur le brin complémentaire présenteront la séquence correcte. Ce filtre, appelé "strand bias", est mentionné dans l'article de revue de Nielsen et ses collègues (7) mais est pourtant très peu utilisé en pratique. Seul le GATK (81) en a implémenté une version, et aucun pipeline ne l'utilise pour faire l'appel des génotypes à proprement parler. Pour évaluer les effets du "strand bias", nous avons réalisé des appels indépendamment dans les 2 sens, et nous les avons comparés. Les positions sont classées en 3 catégories : "strand-concordant", si l'appel sur les deux brins est le même ou si l'un des deux est dans la zone grise adjacente, "single-strand", si la couverture est insuffisante sur un brin, et que l'appel est franc sur l'autre, ou "strand-discordant", si les génotypes appelés sont différents, ou si l'un d'eux est dans la zone grise éloignée.

La figure 3.4 a) montre la répartition des SNVs dans chacune des catégories, en fonction de la couverture minimale requise pour un appel. Étonnamment, les "strand-discordants" sont très nombreux. Dans ce cas,



a)



b)

FIG. 3.4 – a) Le nombre d'appels strand-discordants, strand-concordants et single-strand en fonction de la couverture minimale. b) La proportion d'appels strand-discordants parmi les positions appelées sur les deux brins, en fonction de la couverture minimale. C'est une estimation du taux de faux positifs des appels single-strand. Même à très forte couverture, ce taux reste élevé.

au moins un des deux appels est erroné. Faire un appel indépendamment sur chaque brin est donc crucial, car contrairement aux méthodes classiques qui risquent d'appeler un génotype faux sur ces positions, notre méthode permet de les filtrer facilement. De plus, les positions "single-strand" représentent la majorité des données, et au vu de la découverte précédente, il est légitime de se poser des questions sur leur fiabilité. En faisant l'hypothèse d'une part, que le taux de faux positifs est le même pour les positions vues sur les deux brins et pour celles vues sur un seul brin et d'autre part, que le mauvais génotype est appelé pour la moitié de positions strand-discordantes , nous avons estimé le taux de faux positifs des SNVs "single-strand" ainsi :

$$\frac{\#\textit{strand-discordants}}{2\cdot(\#\textit{strand-discordants}+\#\textit{strand-concordants})},$$

avec *#strand-discordants* le nombre total de positions strand-discordantes et *#strand-concordants* le nombre de positions strand-concordantes hétérozygotes ou homozygotes références. Cette estimation fait également l'hypothèse que les strand-concordants sont des vrais positifs, ce qui est faux car ceux-ci contiennent de nombreux faux positifs (voir section suivante). Cela conduit à sous-estimer le taux de faux positifs.

La figure 3.4 b) montre l'évolution de l'estimation du taux d'erreurs des appels "single-strand" en fonction de la couverture minimale. Ce taux est très élevé ($20,3\%$ à 10x de couverture minimale sur chaque brin), même avec une couverture minimale importante (12% à 100x). Nous avons donc décidé de filtrer l'ensemble des appels "single-strand", et nous conseillons aux développeurs des autres pipelines d'appel de génotypes de faire de même : cela permet de diminuer considérablement le taux de faux positifs.

Utiliser les réplicats TCGA pour choisir les paramètres d'appel et proposer de nouveaux filtres

Nous avons réalisé les appels des génotypes pour l'ensemble des données décrites dans la section 3.3.1, avec la méthode d'appel de génotype décrite dans la section 3.3.2 pour différentes valeurs de couverture minimale et en faisant varier la taille des "zones grises". Afin d'estimer le taux de faux positifs restant dans nos données, nous avons ensuite comparé les génotypes obtenus pour les réplicats. Pour un même échantillon, à une même position, deux génotypes sont appelés "run-discordants" s'ils sont différents, et "run-concordants" s'ils sont identiques. Ainsi, de la même manière que pour les erreurs single-strands, le taux de faux positifs se modélise par :

$$\frac{\#\textit{run-discordants}}{2\cdot(\#\textit{run-discordants}+\#\textit{run-concordants})}.$$

Estimer ce taux pour les différentes valeurs des paramètres d'appel de génotype nous aide à choisir une couverture minimale et une taille de zones grises réalisant un bon compromis entre spécificité et sensibilité. Nous ne pouvons pas calculer directement la sensibilité de la méthode car le génotype étudié est inconnu, mais le nombre de run-concordants nous permet de comparer la sensibilité de la méthode pour différents paramètres. Alors que la plupart des études utilisent une couverture minimale de 10x sur l'ensemble des deux brins, nous conseillons d'exiger une couverture minimale de 10x sur chaque brin (section 3.5). Par ailleurs, nous choisissons d'utiliser des zones grises de taille 20%. Ces valeurs conduisent à appeler $3,1\%$ de faux positifs pour les indels, et $1,2\%$ pour les SNVs. Pour les échantillons séquencés plusieurs fois, il est facile de supprimer les génotypes "run-discordants", et d'obtenir ainsi des données très propres ne contenant presque plus de faux positifs. Cependant, pour des raisons de coût, il est rare que des réplicats soient réalisés pour l'ensemble des échantillons analysés. Par exemple, parmi les 1043 échantillons séquencés dans le cadre de l'étude du cancer de l'ovaire, 807 n'ont été séquencés qu'une seule fois. Afin d'essayer de débarrasser les données séquencées une seule fois des faux positifs restants, nous avons regardé la répartition des positions "run-discordantes" sur le génome : de nombreuses positions semblent sujettes à erreurs pour

tous les échantillons séquencés par une même plateforme (section 3.5). Cela rejoint le fait que certaines erreurs systématiques dépendent du contexte génomique : une même technologie va avoir tendance à se tromper toujours aux mêmes endroits. Ainsi, les nombreux réplicats réalisés dans cette étude nous ont permis d'annoter bon nombre de loci comme "sujets à erreur", ce qui permet de les filtrer pour les échantillons séquencés une seule fois. Pour vérifier le bien-fondé de cette méthode, nous avons réalisé un bootstrapping simple : nous apprenons les positions sujettes à erreur sur la moitié des échantillons, et nous estimons les taux de faux positifs de l'autre moitié avec ou sans ces positions. Les résultats sont concluants, avec une baisse moyenne relative du taux de faux positifs de 24% (soit $0,3\%$ de faux positifs filtrés) sans perte de sensibilité.

3.4 Conclusions

Obtenir des génotypes fiables à partir de données NGS n'est pas encore fait de manière routinière et les méthodes utilisées demandent à être améliorées, en particulier car les génotypes appelés contiennent de nombreuses erreurs. Nous montrons qu'utiliser le sens d'alignement des reads sur le génome permet de grandement réduire le taux de faux positifs, et nous utilisons les réplicats techniques pour filtrer certaines des erreurs systématiques restantes grâce à l'annotation de loci "sujets à erreurs". Par ailleurs, notre estimation du taux de faux positifs basée sur la reproductibilité des résultats indique que les seuils de couverture minimale généralement utilisés dans la littérature sont trop faibles. Nous montrons également que les méthodes proposant un génotypage pour des positions contenant entre 20 et 40% de reads variants appellent probablement un grand nombre de faux positifs. Cette observation est d'autant plus importante pour les génotypes appelés dans des tissus tumoraux, qui sont souvent basés sur de telles proportions.

Notre méthode permet de filtrer un grand nombre d'erreurs systématiques, et utilise des valeurs de paramètres très stringentes. Malgré toutes ces précautions, nous estimons qu'il reste $0,96\%$ de faux positifs dans nos appels de génotypes, faux positifs pouvant provenir d'erreurs de séquençage, d'alignement ou d'appel de génotype. Appliquée aux données de séquençage des tissus sains des patientes de l'étude TCGA, cette méthode nous a permis d'annoter 5 833 985 positions variantes chez au moins une patiente pour un total de 16 706 986 SNVs variants.

Le défi majeur est ensuite d'identifier, parmi cette multitude de variations génétiques, celles prédisposant au développement du cancer. C'est l'objet du chapitre suivant : nous combinons des informations fonctionnelles et d'expression de gènes afin de sélectionner les SNVs et gènes probablement impliqués dans la cancerogénèse.

3.5 Document joint : Filtering systematic errors in next-generation data : stranding matters

Filtering systematic errors in next-generation genotype calls: stranding matters

Laure Sambourg[1], Thierry-Mieg Jean[2], Thierry-Mieg Danielle[2], and Thierry-Mieg Nicolas[1,*]

[1]UJF-Grenoble 1 / CNRS / TIMC-IMAG UMR 5525, Computational and Mathematical Biology (BCM), Grenoble, F-38041, France

[2]NIH/NCBI

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Background

Next generation sequencing technologies have enabled the production of massive datasets, opening up new avenues of research. In particular, exome capture and sequencing provides a cost-effective means of genotyping the coding portion of the genome for large cohorts of individuals. However calling genotypes from NGS data remains difficult, and high rates of discrepancies between called genotypes are observed when using different sequencing technologies on the same samples, or different software pipelines on the same raw sequencing data. Some of these difficulties may be due to systematic biases and errors arising at the sequencing, alignment or genotype-calling steps. In order to investigate these possibilities, we analyzed 108 deep germline exome-seq replicates produced by The Cancer Genome Atlas with ABI SOLiD and Illumina GAII platforms, searching for sources of systematic errors.

Results

We aligned the TCGA short-reads using the MAGIC pipeline, which allowed us to distinguish reads aligning on the forward and reverse strands of the genome, and developed a simple and effective algorithm for calling genotypes. This algorithm favors specificity over sensitivity: calls are only made when the data is unambiguous and sufficiently deep. Furthermore, we called genotypes independently on each strand and compared the resulting calls. Surprisingly, we discovered that they disagree in 20.3% of the positions where high-confidence calls can be made on both strands. This observation is not an artifact of the MAGIC aligner, as shown by reanalyzing a published dataset initially studied with GATK. These strand-discordant positions appear due to the sequence context, which is strand-specific and can lead to systematic sequencing errors on one strand. Furthermore, the TCGA replicates allowed us to identify systematic error-prone positions in the genome, some of which are specific to the ABI or Illumina sequencers and some of which are cross-platform. Filtering these positions significantly improves the genotyping quality.

Conclusions

Our results clearly show that strand-specific read counts should always be provided, and that a reliable genotype can only be called when the two strands are compatible and sufficiently covered. In addition, lists of error-prone positions are provided and should help to filter out systematic errors. Beyond genotype-calling, our findings have implications on the experimental design of exome-capture experiments: capture libraries should be short enough to allow a significant proportion of positions to be sequenced on both strands.

1 INTRODUCTION

The great progress of DNA-sequencing technologies has enabled the routine sequencing of human genomes, generating an unprecedented amount of data. For instance, international consortia are currently sequencing thousands of genomes of patients, aiming at understanding the molecular bases of genetic diseases and promising the development of personalized medicine.

This approach requires the characterization of individuals' genomes, ranging from small variants to CNVs or larger rearrangements. Briefly, it consists of several steps: library construction, sometimes followed by exome-capture, sequencing, short-reads alignment to a reference genome, and finally genotype calling or rearrangement identification. Our study focuses on the genotype calling problem, that is the classification of each locus in a given individual as reference homozygote, heterozygote or variant homozygote, depending mainly on the proportion of sequencing reads aligned at the considered position that carry the reference or variant nucleotide. The main challenge is the presence of sequencing and alignment errors, some of them systematic, that need to be distinguished from true variations. This is particularly important because between-sample reproducible errors could easily mislead biological conclusions, for example in association study.

© Oxford University Press 2005

Several genotype calling pipelines have been developed, often classified as heuristic or probabilistic methods (Nielsen *et al.*, 2011). Heuristic methods use arbitrary or empirically determined cutoffs: typically, positions with a proportion of non-reference alleles between 20% and 80% are called heterozygotes (Hedges *et al.*, 2009). Those methods are criticized because they do not allow to associate a likelihood to the genotype calls, and because they do not take into account read mapping quality. Probabilistic methods are usually based on the binomial assumption and estimate the most probable genotype accordingly. For instance, the popular GATK (Genome Analysis Toolkit, (McKenna *et al.*, 2010)) developed by the Broad Institute proposes a bayesian classifier: knowing the base call of each read aligning at a given locus, it computes the posterior probability of each genotype and assigns the greatest one to the locus, with prior probabilities depending on each base's Phred quality score. The SOAPsnp (R. Li *et al.*, 2009) caller uses the same kind of approach, except that the prior probability of each genotype is based on estimated SNV rates and transversion/transition ratios. Furthermore, to improve sensitivity for low-depth sequencing, a higher heterozygote prior probability is used for known SNVs (those present in dbSNP). SeqEM (Martin *et al.*, 2010) uses the same model, but it computes genotypes for several individuals jointly, taking advantage of the studied population to estimate the model parameters instead of relying on external sources. The SAMtools package (H. Li *et al.*, 2009) allows to compute the genotype maximum-likelihood, but also proposes several statistics to analyze short-read data without calling the variants. Primarily design for the analysis of pooled-NGS data, Wei and colleagues (Wei *et al.*, 2011) use hypothesis-testing to discriminate between rare and common variants within a population.

When looking at cancer genomes, the problem is even more challenging. Indeed, firstly, cancer cells are highly heterogeneous and secondly, extracted samples usually contain a mix of tumor and normal cells (estimated to 30% of contamination by normal DNA (Carter *et al.*, 2012)), making the classification of variants as somatic (present only in the tumor) or germline (in both samples) difficult. Proposed solutions are based on comparing normal and tumor proportions of variant reads. For instance, VarScan2 (Koboldt *et al.*, 2012) performs heuristic genotype call on normal cells and uses fisher's test to discriminate germline from somatic variants. SomaticSniper (Larson *et al.*, 2012) estimates the likelihood of the combined normal and tumor genotypes, based on estimated mutation rates. Along the same lines, MuTect (Cibulskis *et al.*, 2013) couples 2 bayesian classifiers to first call the variants in normal and cancer samples, and then classify each variant as germline or somatic.

The large number of software and algorithms available shows that genotype calling is a very active area of research, and the questions that naturally arise are how those pipelines perform, and how they compare to one another. A study from O'Rawe and colleagues (O'Rawe *et al.*, 2013) compares 5 genotype calling pipelines on 15 exomes from four families sequenced at high coverage on the Illumina HiSEQ2000 platform, and found as little as 57% of concordant SNVs and 26.8 % of concordant indels between the 5 pipelines. (Altmann *et al.*, 2012) ran several alignment pipelines, followed by the SAMtools or the GATK genotype caller, on an individual's exome sequenced by the 1000 genomes project (Illumina sequencing with a 10x minimum coverage). They found a mean concordance rate of 85% between the two genotype callers used with the same aligner, and again a mean concordance rate of 85% between 2 alignment pipelines used with the same genotype caller. Lam and colleagues (Lam *et al.*, 2012) compared Illumina (with BWA and GATK for alignment and genotype calling) and Complete Genomics sequencing platforms on whole-genome sequencing of an individual's blood and saliva samples. They found ~ 88% concordant SNVs between the two pipelines, imputed to a high false positive rate. One the other end, for indels, the 26.5% concordance was suggested to be attributed to low specificity rather than low sensitivity. Note that those values cannot be compared to the other studies, as here, heterozygotes and variant homozygotes are considered concordant.

This emphasizes the fact that DNA-seq bioinformatics pipelines need further development before a consensus is reached and applied routinely, and one of the main challenges is to reduce error rates. To this end, scientists are applying pre and post-alignment filtering steps. Reumers and his colleagues (Reumers *et al.*, 2012) listed the filters generally used by the community, and estimated the efficiency of each one of them on Complete Genomics and Illumina data. For Illumina's pipelines, coverage depth, variant score, allelic imbalance, clustered SNVs and consensus mapping and calling appear to be the most efficient ones. Another issue of genotype calling pipelines is the lack of assessment of the best calling parameters. For instance, the choice of the minimum coverage required for calling a SNV ranges from 4x to typically 10-11x, and is mostly chosen empirically. In their framework, Reumers and colleagues advise to use a 20x minimum coverage to better distinguish true variants. In addition, an important insight is the context-dependence of the errors. This phenomenon has been highlighted by (Meacham *et al.*, 2011): they performed Methyl-seq of an individual, and used overlapping mate-pairs to detect error prone loci from the Illumina pipeline. They found clear evidence of systematic errors, and particular sequence motifs surrounding systematic false-positives.

In this study, we took advantage of 108 deep exome-seq replicates produced by TCGA (The Cancer Genome Atlas) to propose and test original filtering steps. In particular, we showed that loci exhibiting between 20 to 40% variant reads cannot be classify only based on read variant proportion. In addition, we discovered that taking into account the concordance between reads aligning on the plus and the minus strand allows to greatly reduce the error rate. We also point the existence of error-prone positions, specific or not to sequencing platforms, that are highly likely false positives. To perform those analyses, we proposed an improved version of the cutoff method, based on buffer zones. Finally, we estimated the remaining error rate in the data, showing that even if careful filtering allows to greatly reduce the number of false-positives, drawing biological conclusions from DNA-seq experiments should be done with care.

2 RESULTS

2.1 Using TCGA replicates to finely tune pipelines

Consortia such as The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC, International Cancer Genome Consortium *et al.*, 2010) are producing massive amounts of exome-seq data, deep-sequencing thousands of tumor and matched normal exomes. These data have been extensively analyzed, giving insight into the cellular mechanisms driving cancers. However, a noteworthy sideline possibility has surprisingly been poorly (if at all) investigated: these data also allow for an unprecedented sequencing platform and bioinformatics pipeline comparison. Indeed, often a sample is sequenced several times independently by different labs of the consortium. This is a mine of information for comparing Illumina and/or SOLiD sequencing platforms, and for testing bioinformatics pipelines.

We used the ovarian TCGA exome-seq data (The Cancer Genome Atlas, 2011) to assess the variability between and within sequencing platforms, to finely tune our genotype caller, and to propose additional filtering steps for improving the quality of genotype calls in general. Table 1 shows the number of TCGA ovarian replicates used in our analysis. They consist of non-tumor samples (mainly blood) sequenced by the Baylor College of Medicine (SOLiD sequencing), the Broad Institute and/or Washington University (Illumina sequencing). Supplementary table 1 summarizes the data generation protocol used by each center, and a more detailed description can be found in the TCGA article supplementary material.

Baylor SOLiD	35		
Broad Illumina	3	0	
WU Illumina	0	19	51
	Baylor SOLiD	**Broad Illumina**	**WU Illumina**

Table 1. Number of replicates performed between and within labs. For instance, 51 samples have been sequenced twice or more at Washington University. Baylor: Baylor College of Medicine, Broad: Broad Institute, WU: Washington University Genome Sequencing Center, SOLiD: SOLiD sequencing, Illumina: Illumina sequencing

TCGA short-reads were pre-processed and aligned by the MAGIC aligner (see methods). Altogether, we aligned 3137 Gb, leading to an average 323 coverage per bp. Read counts were collected for positions with a 10x minimum coverage and carrying at least 20% of the variant allele in at least one of the samples. This results in a total of ~ 7,800,000 positions.

2.2 Choosing a variant calling method

In order to get the general picture, we first looked at the distribution of the proportion of variant reads (Fig. 1), for different minimum coverage cutoffs. Note that positions are restricted to those with more than 5% variant alleles: we are not interested in obvious reference homozygotes, that represent the vast majority of the data. As expected, we observe peaks around 0%, 50% and 100% of variant reads, corresponding respectively to reference homozygotes, heterozygotes and variant homozygotes. The clear demarcation between the variant homozygote and the heterozygote peaks shows that those categories can be differentiate easily. However, there is a zone, roughly from 20 to 40 % variant reads, which contains both heterozygotes and reference homozygotes SNVs: only based on variant reads proportion, it is impossible to distinguish those categories. Such proportion of variant reads could be due to sequencing errors, but also to duplicated genes not annotated in the reference genome : in that case, the aligner assigns reads coming from both copies to the one existing in the reference. Provided that only one copy carries the variant allele, 25% variant reads are expected. Note that this problem cannot be address by the use of a higher minimum coverage, as the phenomenon exits even with a 100x minimum coverage. As a consequence, when taking into account only NGS data, one cannot conclude. It may be resolved using external information, *e. g.* linkage disequilibrium for common SNPs. As we focus on NGS data, we decided to classify those positions as “no-call”, thanks to an improved version of the heuristic method based on a buffer zone.

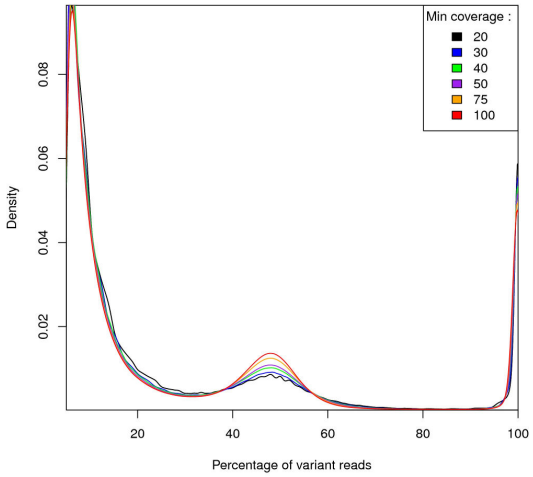


Fig. 1. Density of the proportion of variant reads, for different minimum coverage values. Positions are restricted to those containing at least 5% of variant reads, that is a total of 83 million observations. The plot was performed with the R density function, which generate a Gaussian density estimate (default parameters used).

2.3 Calling genotypes

Using the cutoff method, one needs to set variant read proportion thresholds for assigning a genotype to a locus. The density plot in Fig. 1 suggests to use cutoffs around 30% and 85%. However, as explained previously, calling the genotype for SNVs with 20 to 40 % of variant reads is not possible only based on variant read proportion. Thus, in order to limit wrong classification, we defined a “Grey-zone” (Fig. 2) around the thresholds, where we consider the information as ambiguous, and which will be treated accordingly depending on the situation. At each position in the genome, a “strand-genotype” is called independently on each strand, and can be homozygote, heterozygote or Grey-zone (see 2.4). In fact the strand-genotype is only called if the number of reads covering the position in the considered strand is higher the user-selected minimum strand-coverage. Then, the “general genotype” is called if both strands have the same call or if one of them is in the near-by Grey-zone. Otherwise, we assigned “no-call” to the SNV, which is preferred to a miscall.
The buffer zone might not be necessary for differentiating variant homozygotes from heterozygotes, as there is very little SNVs in the Grey-zone region. However, as a precaution, we decided to use it anyway: while not preventing true positives from being called, it may avoid few false positive calls.

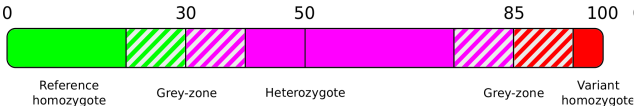


Fig. 2. This figure illustrates the genotype called depending on the percentage of variant reads. The Grey-zones are centered around 30 % and 85 % of variant reads. Independent calls are done for each strand: strand-genotype can be either homozygote, heterozygote or Grey-zone. The general genotype is called if strand-genotypes have the same call or if one of them fall in the adjacent Grey-zone. Otherwise, it is a “no-call”.

The next section explains why we choose to perform independent strand-genotype calls: we show that many loci have discordant calls on the plus and the minus strand. Those are systematic errors that must be filtered.

2.4 A strand story

Systematic sequencing errors, which can be present at high coverage and high variant proportion (*e.g.* (Meacham et al., 2011)), cannot be filtered only based on variant read proportion. Some of them are context-dependent and can be filtered by taking into account the "strand bias": for true variants, the proportion of variant reads aligned on the plus and the minus strand should be the same, and a deviation from this equilibrium can indicate a false positive, or, for exome sequencing, bias in the exome capture (Nielsen *et al.*, 2011). As a consequence, the GATK proposes a tool to filter strand bias posterior to the genotype call, based on a fisher test. However, most of the genotype callers do not use this information at all (Hedges *et al.*, 2009; R. Li *et al.*, 2009; Wei *et al.*, 2011; Martin *et al.*, 2010; Li *et al.*, 2008; Li, 2011). Furthermore, none of the genotype caller has directly compared the calls on each strand, or uses this information to perform the genotype call itself. Here, we perform a comparison of the genotype calls obtained for each strand, and we show that lots of them are discordant, and that positions where only one strand can be called are very low quality and should be excluded.

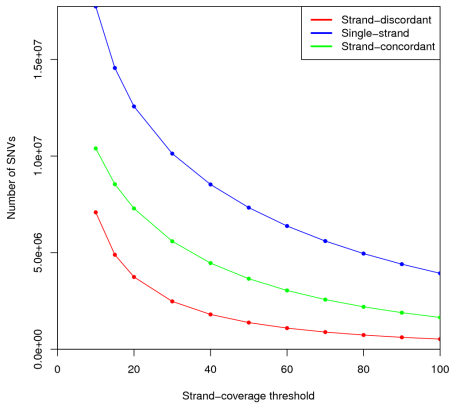
For each position on the genome, the MAGIC aligner returns 4 (weighted, see methods) reads counts: the number of reads carrying the variant allele aligning on the plus strand, the number of reads carrying the reference allele aligning on the plus strand, and the same counts on the minus strand. We make independent call for each strand and compare them. We name "strand-concordant" a position having the same genotype call for both strands or those with one called genotype and the other in an adjacent Grey-zone (Fig. 2). Otherwise the locus is considered "strand-discordant". If only one strand has enough coverage, it is named "single-strand". Otherwise it is ignored. Note that "strand-concordant" reference homozygotes are not taken into account here: their large number hide interesting patterns.

Fig. 3 a) shows the number of SNVs in each category, for a minimum strand-coverage varying from 10 to 100. Surprisingly, even at high coverage thresholds many calls are strand-discordant. Reads aligning at those loci clearly contain many errors: they do not agree amongst themselves. We do not know which genotype is correct, so those SNVs must be filtered. This is not the case in classic approaches that consider only the general genotype. They are likely making erroneous calls on those positions.
Furthermore, most of the calls are single-strand (*e.g.*, single-strand positions represent 50.3 % of the called positions at a 10x minimum strand-coverage). Single-strand positions are also questionable. How reliable are those calls? Should they be included in further analysis?

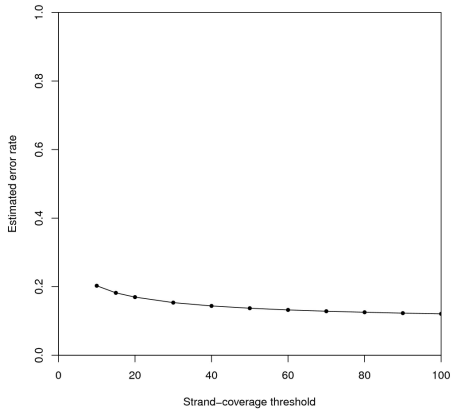
To address this question, we examined the proportion of strand-discordant calls among the positions that were called on both strands. Indeed, assuming on the one hand that the error rate is similar for positions called on both strands and for positions called on only one strand, and on the other hand that half of the strand-discordant position are correctly called, we can estimate the error rate for single-strand calls to be:

$$\frac{strandDiscordantCalls}{2(strandDiscordantCalls + StrandConcordantCalls)}$$

This assumes that all strand-concordant SNVs are true positives, which is unlikely. Accordingly, this is an underestimate of the error rate. Fig. 3 b) shows this estimated lower-bound error rate for different minimum strand coverage values. For a minimum strand-coverage of 10x, the error rate is surprisingly important (~ 20.3%), clearly stating that for single-strand SNVs, a higher threshold should be chosen. Even with high thresholds, the error rate remains important (12% for a 100x coverage), indicating that those SNVs should probably be filtered as well.
To sum up, making independent call on each strand, and keeping only the strand-concordant ones is an easy way to greatly improve the reliability of the calls. First, the number of strand-discordant calls is surprisingly high, and we advise to always filter them in genotype calling pipelines. Secondly, the majority of SNVs are single-strand, that have a high error rate (at 10x strand-coverage, it represents 50.3 % of the positions, of which 20.3% are false-positive). Obtaining high quality genotype thus require to exclude single strand positions.



a)



b)

Fig. 3. a) The number of strand-discordant (*d*), strand-concordant (*c*) and single-strand calls as a function of the minimum strand-coverage. b) Estimated error rate (*d/2(d+c)*) among positions called on both strands, as a function of the minimum strand-coverage.

2.5 Strand bias is not a MAGIC artifact

In order to assess the validity of our observations on other platforms and to confirm the correction of our alignment pipeline, we compared the MAGIC and GATK alignments on a very well studied sample : the NA12878 exome dataset, sequenced by the 1000 Genome Project. The GATK pipeline has been applied to this dataset by (DePristo *et al.*, 2011), and the vcf and bam output files are publicly available. The MAGIC pipeline was run on the extracted short-reads for the their bam file. To estimate the correspondences between MAGIC and GATK alignments, we compared the coverage obtained by both pipeline for each SNV : they should be equal. This is indeed what we observe: despite very little outliers, the coverages generally agree (R^2 = 0.948, see Fig. 4). In order to further verify that the small differences observed are not typical to strand discordant SNVs, we compared the coverage of strand discordant SNVs only. The obtained R^2 (0.902) is slightly smaller, but should not be compared to the R^2 for all SNVs : the number of SNVs and the coverage distribution greatly influence the results. Thus, we randomly sample SNVs from the non discordant ones, following the coverage distribution of the discordant ones. A hundred runs give around the same correlation than when restricted to discordant positions. This imply that the strand bias observed in our data is not a MAGIC artifact, but instead real systematic sequencing errors that should be taken into account by any short-read analysis pipeline.

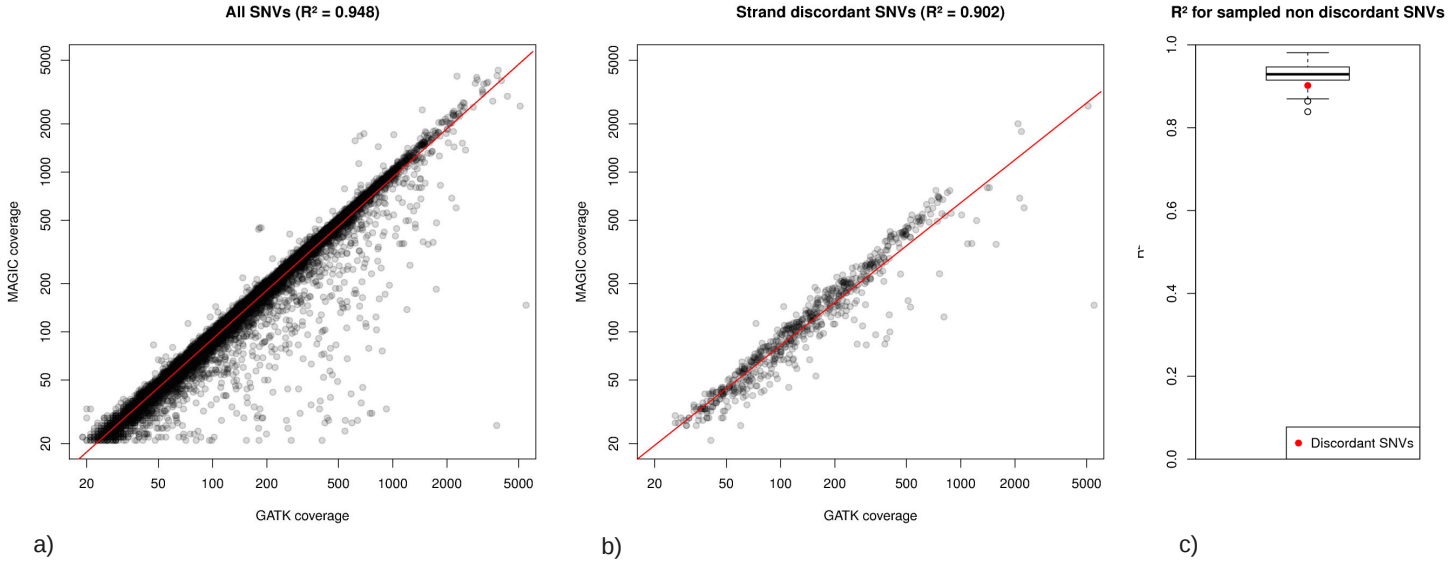


Fig. 4. a) Log-log scale linear correlation between the coverage obtained by MAGIC and GATK pipelines on the NA12878 exome dataset, for every SNVs. b) Same but restricted to SNVs annotated as strand discordant by the MAGIC pipeline. c) R^2 values for the same correlation, but for SNVs randomly sampled amongst the non strand discordant SNVs, sampled following the strand discordant SNVs coverage distribution. Red point is the R^2 for the correlation in b).

2.6 Getting double strand coverage : importance of experimental design

Considering the relevance of the strand coverage for calling the SNVs, it is crucial to maximize the size of the regions covered by both strands when designing an experiment. Looking at the strand-coverage as a function of the genomic position (Fig. 5, a)), we made an interesting discovery : the regions covered by both strands greatly depend on the size of the DNA fragments and the length of the short-reads. Indeed, during an exome-seq assay, single strand DNA fragments are captured from a genomic library. Then, both ends of the fragments are sequenced, always in the same direction (Fig. 5, b)), leading to paired-end short reads. If the size of the short-reads is small compared to the size of the fragments, ends of the fragments are covered only in one strand, and the regions covered in each strand might overlap very little.

2.7 Comparing replicates

Even SNVs seen in both strands with sufficient coverage might not be error-free. To assess the remaining errors, we checked the reproducibility of our genotype calls in replicates. Here, a SNV is named “run-discordant” if two different genotypes are called from the same sample in different experiments, and “run-concordant” otherwise.
Similarly to the stranding study, we estimate a lower-bound of the remaining error rate as:

$$\frac{runDiscordantPosition}{2(runDiscordantPosition + runConcordantPosition)}$$

This estimated error rate is computed for varying parameters thresholds (minimum coverage and Grey-zone size), and used to choose them at best.

Figure Fig. 6 a) and b) shows the evolution of the error rate and the number of concordant SNVs called as a function of the minimum coverage, computed with Grey-zone sizes of 20%. It is important to note that, as we are computing independent calls on both strand, we make a distinction between the strand-coverage, that is the coverage on one strand, and the total-coverage, that is the total number of reads aligning at the locus. For example, a 10x minimum strand-coverage correspond to a 20x minimum total-coverage for the position, but additionally requiring at least 10x coverage on each strand. Despite the stringency of the Grey-zone criteria, the genotype call error rate is still pretty high. Increasing the minimum strand-coverage clearly decreases the error rate, falling from 1.6 % at 5x, 1.2 % at 10x and 0.96% at 20x. However, this is at the price of sensitivity (Fig. 6 b)), and 10x strand-coverage seems a reasonable compromise between sensitivity and specificity. This is twice what is typically used as coverage threshold, and we think that the scientific community should revise upward the usual 10x total-coverage. This is also the point of view of the other study that estimated filters efficiency (Reumers *et al.*, 2012).

Similarly, increasing the Grey-zone size decreases the number of false positives. This was expected, as a higher Grey-zone conducts to ignore more SNVs as both strands fall in the Grey-zone. A Grey-zone size of 20 seems a good compromise between sensitivity and specificity (Fig. 6 c) and d)).

To sum up, we advise to choose a threshold of 10x strand-coverage (or 20x total-coverage if strand-coverage is not available), to remove strand discordant SNVs but to allow for 20% tolerance (Grey-zone). For those thresholds, we observe an overall concordance of 98.8%, attesting the validity of our pipeline. We also confirm that, as observed in many other studies, the error-rate of next-generation sequencing platform is higher for indels than for substitutions. For the chosen thresholds, we observe an error-rate of 3.1% for indels and 1.2% for substitutions. Choosing stringent thresholds is required to obtain reliable enough genotype calls (of course, at the price of sensitivity). However, this is not sufficient, as we still observe discordance between runs. When replicate sequencing is done, removing run-discordant calls is elementary. However, due to cost-constraints, it is very rarely the case, and further filtering steps are needed. As a consequence, we took advantage of the list of run-discordant SNVs we obtained from this massive study to find predictive methods for detecting them in a replicate-free study.

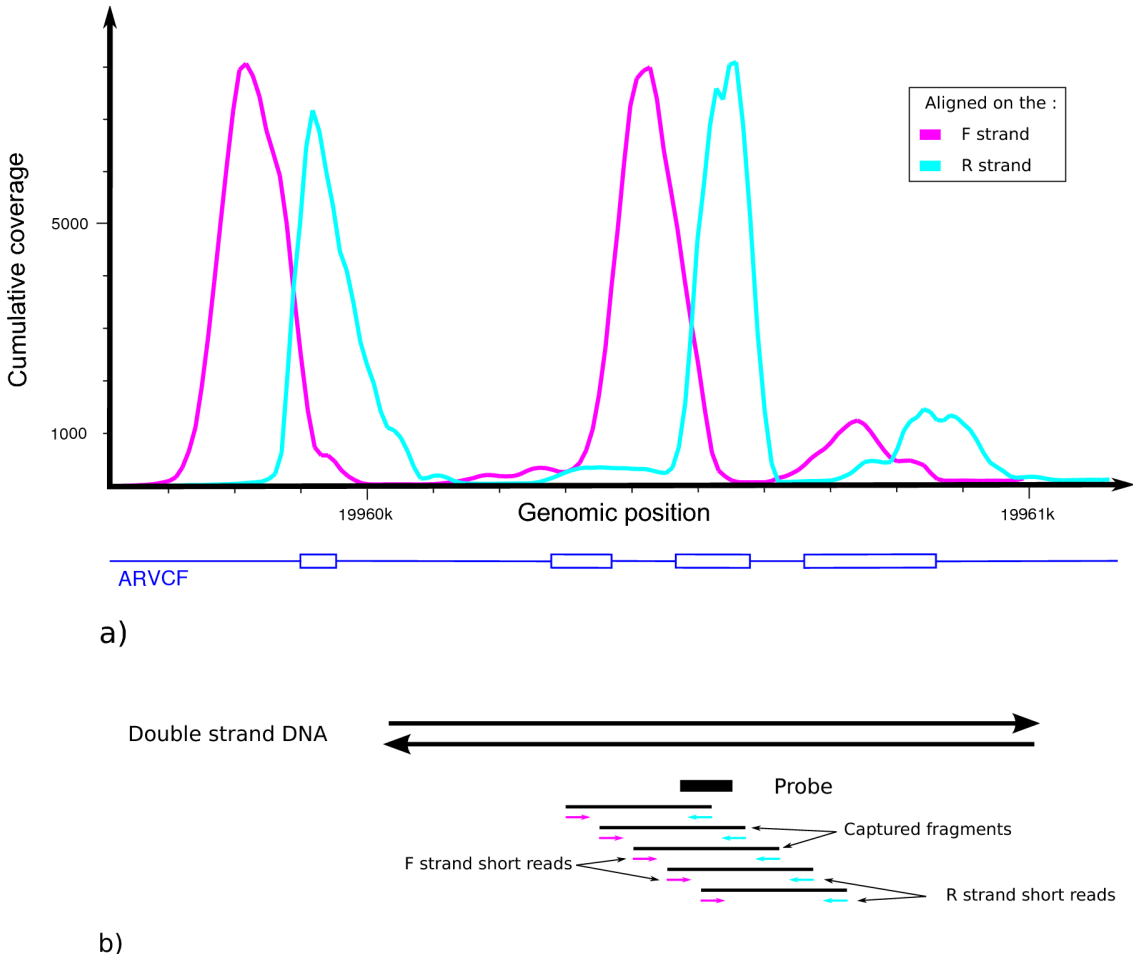


Fig. 5. **Importance of the experimental design on the double strand coverage.** Panel a) shows an example of the strand-coverage (cumulative on 120 runs) as a function of the genomic position. Here the regions covered by both strand are small. During an exome-seq experiment, both ends of the captured DNA fragments are sequenced, always in the same strand. If the fragments size is large compared to the short-read size, then the regions covered in each strand overlap very little (panel b), not scaled). The distance between the red and the blue peaks is $L - \lambda$ where L is the average fragment length and λ the read length. The same figure is obtained in paired-end and single-end sequencing. Note that the alignments used here are not the TCGA data, but come from another experiment (data not shown).

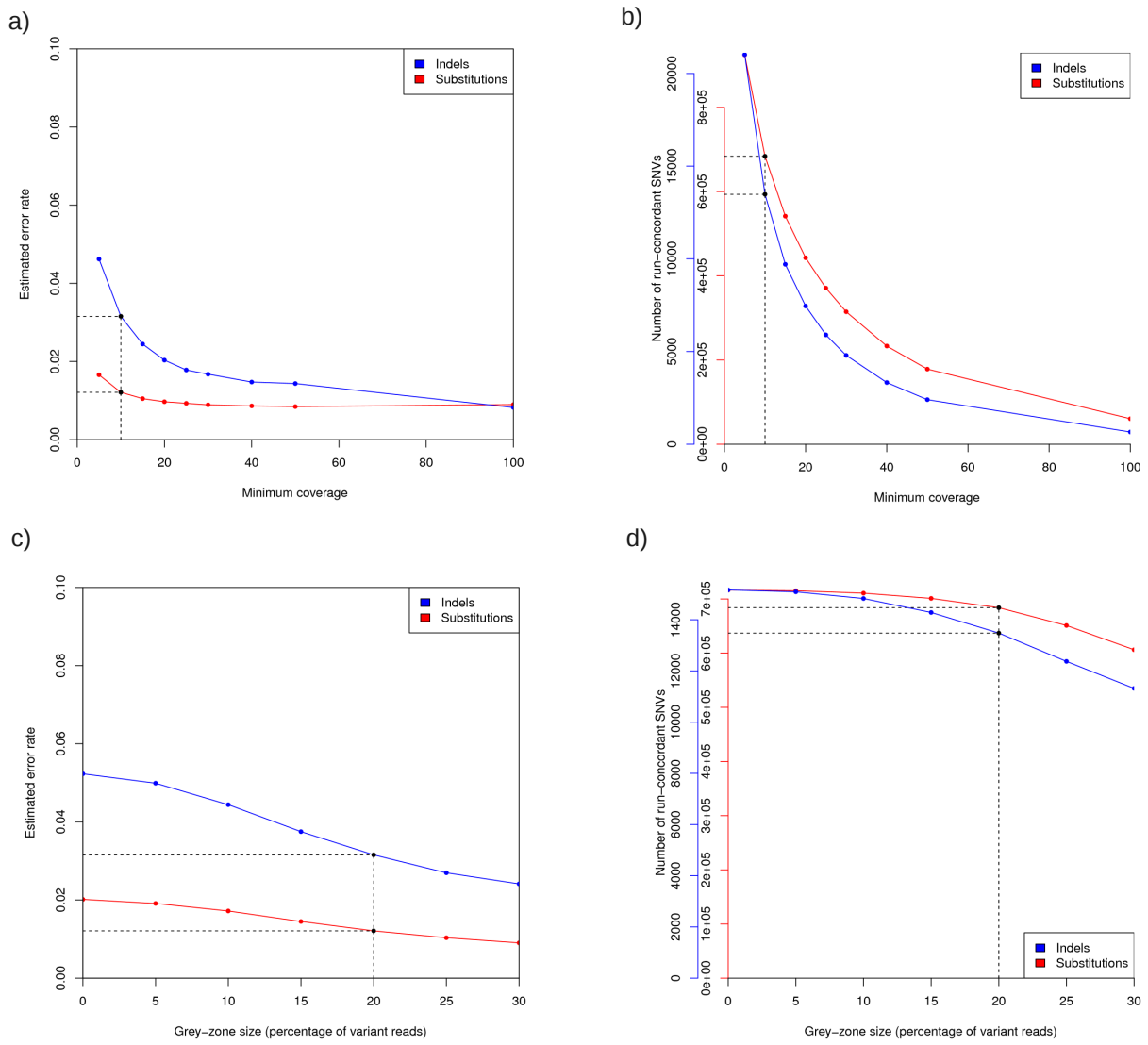


Fig. 6. Left panels show the estimated error rate (run-concordant/(run-concordant+run-discordant)) as a function of genotype call parameters (the minimum strand-coverage (a) and the Grey-zone size (c)). Right panels show the number of concordant calls for the same parameters. Panel c) and d): Grey-zone are centered around 30 and 85%. *E.g.*, a Grey-zone size of 20% correspond to Grey-zones between 10 and 40% and between 75 and 95%.

2.8 Run-discordant SNVs on the genome

To gain insight into the remaining systematic errors, we looked at the distribution of the run-discordant SNVs on the genome (Fig. 7). Errors seem to cluster at particular positions. While some positions seem to be error-prone independently of the sequencing-platform used, others are specific to Illumina or SOLiD.

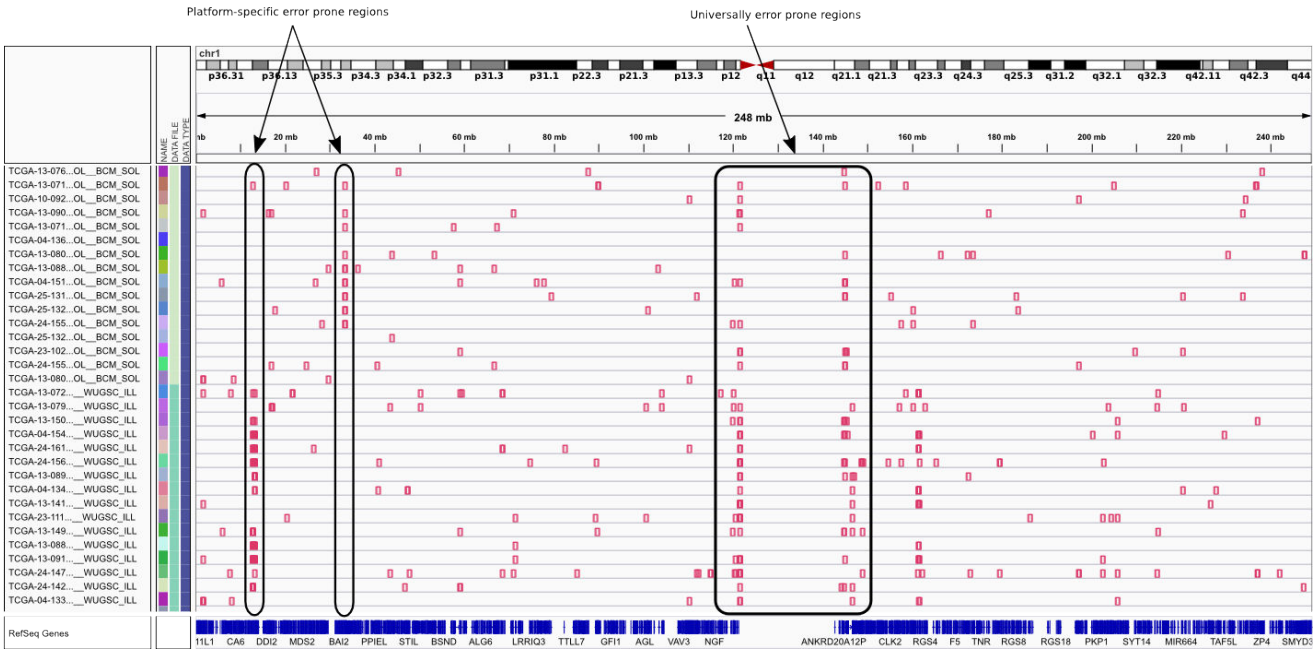


Fig. 7. Positions of run-discordant SNVs are shown for chromosome 1, for a subset of replicates. The first 16 lines correspond to SOLiD replicates, and the next ones to Illumina replicates. This figure was done with IGV (Integrative Genome Viewer (Thorvaldsdóttir *et al.*, 2013)).

Having such a number of replicates allows us to propose another filtering procedure. We define as “error-prone position” for a particular platform a position seen discordant in at least 3 patients. Then, for any sample sequenced by this platform, typically from a non-replicate experiment, filtering the SNV falling within those error-prone positions should reduce error rate.

To evaluate the predictive power of this method, we performed resampling. We learned the list of discordant positions from a training set of randomly chosen samples (half of the complete set), and then estimated the error rate of the remaining samples keeping or ignoring the error-prone positions. The results clearly show the efficiency of this filter, with an error rate relative drop of 24 % on average (that is 0.3% false positive filtered), and almost no concordant SNVs excluded (Fig. 8).

For future reference, we provide a list of 9673 error-prone positions for Illumina and SOLiD platforms (supplementary table 1), that can be used to efficiently filter systematic errors. If high sensitivity is required for a specific reason, we advise to at least annotate those positions as 'dubious', and treat them with particular care in downstream analysis. The minimum number of discordant sample to annotate a position as error-prone can also be chosen higher than 3.

Furthermore, regions near heterochromatin seem to contain more false-positives. We tried to remove SNVs near those regions (using the USCS gap track), but it resulted in very little improvement of the false-positive rate, and many true positives discarded. Thus, this filter was not used.

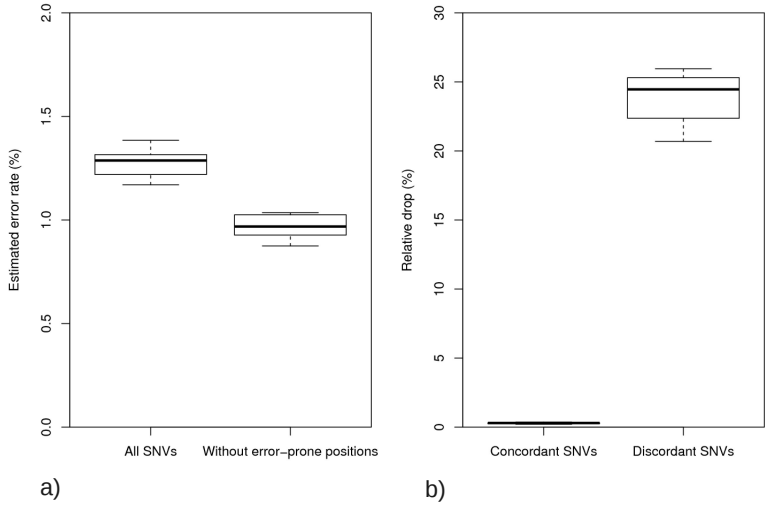


Fig. 8. a) Estimated error rate of test sets for the resampling runs, considering all SNVs or ignoring the error-prone positions found in the training set. The estimated error rate is the proportion of run-discordant SNVs amongst run-discordant and run-concordant SNVs. b) The relative drop of concordant and discordant SNVs between the test sets considering all SNVs and ignoring error-prone positions.

3 DISCUSSION AND CONCLUSIONS

We presented an efficient genotype calling method, finely tune thanks to estimated error rate from numerous replicate experiments. Performing independent call on each strand allows to filter an important number of systematic errors, and we advise to always take this parameter into account. In particular, we showed that single strand positions are very numerous, and that they contain many false positives. Note that the strand differences observed here are not an artifact of MAGIC, as our alignments are very similar to the GATK ones. Another important concern we raised is the general use of too low cutoff for coverage, resulting in high false positive rates. A 4x coverage should definitely be banned from genotype calling pipeline, and we advise to use, when possible, a 10x strand-coverage (or at least a 20x total-coverage if strand information is not available).

We emphasize the presence of lots of noise around 20-40 % variant reads. In normal sample, this problem can be address by the use of a Grey-zone allowing to ignore dubious SNVs. However, when analyzing cancer cells, somatic mutations are usual called with low percent of variants, typically ~ 20-30%. Indeed, cancer cells are highly heterogeneous, and tumor samples are usually contaminated with surrounding normal cells, so we do not expected 50 % of variant reads for heterozygotes. As a consequence, distinguish true somatic mutations from the noise we observed here remains challenging, and this question should be address by somatic variant callers.

Furthermore, we showed that even high coverage and strand concordance are not sufficient to filter all systematic errors, and confirm that errors are not uniformly spread along the genome. We discovered positions that are error-prone in many samples, depending or not of the sequencing platform. We listed them for both Illumina and SOLiD data, and advise to exclude those positions for downstream analysis, or at least to treat them with care. Excluding them allows to greatly diminish the number of systematic-errors, but even after all those filtering steps, a 0.96% false positive rate remains. The left errors are likely punctual, and should differ from a patient to another, meaning that the quality of the data is sufficient for GWAS-like study. However, when highly precise individual genome is required, for example for diagnostic, the problem is not solved and more effort should be done in this direction.

4 METHODS

4.1 Data

TCGA metadata for the ovarian study were downloaded from CGHub, and non-tumor exome-seq data with replicate sequencing were chosen randomly. Corresponding TCGA data were downloaded in BAM format from the CGHub (Cancer Genomics Hub) using GeneTorrent (October 2012). BAM files were

converted back to original short-reads using SAMtools (H. Li *et al.*, 2009) followed by our script for SOLiD data (supplementary information) or picard for Illumina data (picard.sourceforge.net). *A priori* base call quality scores are not used by the MAGIC aligner, which compute its own base call quality score according to alignments. Thus, to save disk space, fastq and csfasta files were converted to fastc, a format, internal to MAGIC, that contains only the base call information and maintains fragment pairs as single entities: attggctctg><ttgacccga.

Below are the command-line used for download and conversion:

```
##### download metadata
cgquery -a "analyte_code=ANALYTE_CODE&disease_abbr=OV"
# with analyte_code in
# D DNA
# G Whole Genome Amplification (WGA) produced using GenomePlex (Rubicon) DNA
# W Whole Genome Amplification (WGA) produced using Repli-G (Qiagen) DNA
# X Whole Genome Amplification (WGA) produced using Repli-G X (Qiagen) DNA (2nd Reaction)

##### download the chosen samples from CGHub
GeneTorrent -v -c <cghub_key_file> -d <sample_URI>

##### convert BAM to csfasta
samtools view -o <samfile> <bamfile> # convert BAM to SAM
convertSam2Csfasta.pl <samfile> <csfastafile> # convert SAM to csfasta

##### convert BAM to fastq
bam2fastq -o <fastqfile> <bamfile>

##### convert fastq and csfasta to fastc
MAGIC MAKEFASTC
```

4.2 Alignment strategy

Short-reads were aligned to the reference genome with the MAGIC aligner (Thierry-Mieg and Thierry-Mieg, 2006), which performs a tuned version of the hash and extend algorithm. In this pipeline a dense set of 16-mers present in the short reads is hashed. Then the genome is scanned and exact matches are extended to obtain an alignment, allowing indels and mismatches. Finally the best alignment is selected, preserving the compatibility of paired reads if available. Notably, as MAGIC was originally designed for RNA alignment, non-complete and non-consecutive read alignments are allowed, avoiding loss of information or miss-alignment due to rearrangements.

An original feature of the MAGIC aligner is the computation of an *a posteriori* base call quality score. For each sequencing run, MAGIC computes a quality score per position in the read, based on the number of mismatches at this position over the entire run. Specifically, the quality score at position p in a read is defined as z(p) = (q(p)-10)/20 bound to [0,1], where q(p) is the base call quality score of the position p after alignment (-log10(m), m being the mismatch rate at p). Therefore z(p)=1 for any position where q(p)>=30 and z(p)=0 if q(p)<=10. Each base call at position p in a read is then weighted by z(p).
Typically, ends of reads will have lower quality, but the error-rate can also be unusually high at specific positions in a given run, due to artifacts during the sequencing reaction. The “read counts” reported by MAGIC (and referred to as read counts in this manuscript) actually incorporate this quality score, *i.e.* each base contributes z(p) (rather than one) “counts”, where p is the position of the base in its read. Note that MAGIC weighted read counts can only be lower than exact read counts, since quality scores are in the interval [0,1]. Furthermore, the weighted read count is typically close to 80% of the exact read count, on average the correction remains moderate.

4.3 Comparing MAGIC and GATK alignments

The GATK read counts were downloaded from the GATK website (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20101201_cg_NA12878/NA12878.ga2.exome.vcf.gz), and the MAGIC pipeline was run on the NA12878 1000 Genomes short-reads. Then, log scale linear regression between both read counts was performed for all SNVs (Fig. 4) and restricted to strand discordant SNVs (according to MAGIC counts). Finally, SNVs were randomly sampled from the non discordant SNVs, following the same coverage distribution than the strand discordant ones. The boxplot summarizes the R^2 obtained by computing the same regression on each set of sampled SNVs (100 runs performed). The box represents the second and third quartile, and the whiskers extends to the most extreme data point which is no more than 1.5 times the interquartile range (R default boxplot).

ACKNOWLEDGMENTS

The results published here are in whole or part based upon data generated by The Cancer TCGA Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at

http://cancergenome.nih.gov. The dbGaP accession number of the dataset analyzed is phs000178.v5.p5.c1. This research was supported in part by the Intramural Research Program of the NIH, National Library of Medicine.

REFERENCES

Altmann,A. *et al.* (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum. Genet.*, **131**, 1541–1554.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413–421.

Cibulskis,K. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.

DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Hedges,D.J. *et al.* (2009) Exome sequencing of a multigenerational human pedigree. *PloS One*, **4**, e8232.

International Cancer Genome Consortium *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.

Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

Lam,H.Y.K. *et al.* (2012) Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.*, **30**, 78–82.

Larson,D.E. *et al.* (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinforma. Oxf. Engl.*, **28**, 311–317.

Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinforma. Oxf. Engl.*, **27**, 2987–2993.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,R. *et al.* (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, **19**, 1124–1132.

Martin,E.R. *et al.* (2010) SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinforma. Oxf. Engl.*, **26**, 2803–2810.

McKenna,A. *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

Meacham,F. *et al.* (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, **12**, 451.

Nielsen,R. *et al.* (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12**, 443–451.

O'Rawe,J. *et al.* (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.*, **5**, 28.

Reumers,J. *et al.* (2012) Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat. Biotechnol.*, **30**, 61–68.

The Cancer Genome Atlas (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.

Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.*, **7 Suppl 1**, S12.1–14.

Thorvaldsdóttir,H. *et al.* (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.

Wei,Z. *et al.* (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.*, **39**, e132.

Chapitre 4

Prioritarisation des candidats, application au cancer de l'ovaire

Nous avons développé une méthode fiable d'appel de génotypes et de filtrage d'erreurs systématiques, qui permet de caractériser les variants d'un génome à partir de données NGS. Cette méthode a été appliquée à des données de séquençage d'exome de 520 patientes atteintes d'un cancer de l'ovaire. Ayant ainsi obtenu une liste de variants pour chaque patiente, le problème est alors d'identifier parmi ces SNVs ceux jouant un rôle dans le développement du cancer. C'est l'objet de cette étude.

4.1 Le paysage génétique des cancers

Le cancer est une classe large et hétérogène de maladies génétiques. Ses causes sont variées et souvent impossibles à identifier. Elles peuvent être environnementales (tabac, obésité, infection, radiation) ou héréditaires (facteurs de risque génétiques). Le cancer se caractérise par une prolifération cellulaire anormalement élevée, conduisant à la formation de tumeurs et/ou à la dispersion de métastases dans l'organisme. Ces comportements sont dus à des modifications génétiques et épigénétiques de la cellule altérant certains mécanismes cellulaires.

On observe notamment des dérégulations des signaux de contrôle de l'homéostasie cellulaire (équilibre entre division et mort des cellules). Alors que dans les tissus sains, ces signaux sont transmis entre cellules voisines, certaines cellules cancéreuses acquièrent la capacité de s'auto-stimuler, par exemple en produisant elles-mêmes les facteurs de croissance, ou en devenant hypersensibles à ces ligands grâce à l'augmentation du nombre de récepteurs (54). D'autres cellules altèrent la transmission du signal plus en aval. Par exemple, des mutations affectant les oncogènes BRAF ou PIK3CA activent les signaux de croissance, ou encore des mutations des gènes RAS ou PTEN, qui normalement s'assurent que ces signaux soient transitoires, empêchent la rétroaction négative de se faire correctement. De plus, pour assurer sa prolifération, la cellule cancéreuse doit échapper aux suppresseurs de tumeurs, qui limitent la prolifération cellulaire ou induisent l'apoptose. En particulier, quand la cellule est en situation de stress, par exemple à cause de dommages de l'ADN, la protéine P53 bloque la progression du cycle cellulaire ou déclenche l'apoptose (mort cellulaire programmée). Cette protéine est mutée dans la majorité des tumeurs.

Des modifications des télomères peuvent également permettre aux cellules cancéreuses de se soustraire à l'apoptose. Les télomères jouent un rôle important dans le contrôle de la vie cellulaire : ils protègent les extrémités des chromosomes de détérioration, ou de fusion avec les chromosomes voisins. Lors des divi-

sions successives, la réplication de l'ADN nécessite de raccourcir les chromosomes, et donc les télomères. Ainsi, la taille des télomères indique le nombre de cycles de réplication que la cellule va pouvoir subir. Une enzyme, la télomérase, permet de reconstruire les télomères et donc d'allonger le temps de vie des cellules. Alors que cette enzyme est absente dans la plupart des cellules, elle est parfois anormalement exprimée dans les cellules cancéreuses, leur permettant d'échapper à la mort cellulaire. Combinée avec une inactivation de TP53, son action peut grandement favoriser l'apparition du cancer. Dans un premier temps, des mutations de TP53 permettent aux cellules cancéreuses de survivre à l'érosion des télomères, ce qui augmente leurs mutabilités. Le génome de ces cellules mutantes est ensuite stabilisé par activation de la télomérase (54).

De plus, les tumeurs, tout comme les tissus sains, ont besoin pour survivre d'être approvisionnées en nutriments et en oxygène et de pouvoir évacuer les déchets du métabolisme. Elles doivent donc être vascularisées. Pour cela, elles utilisent un processus appelé angiogénèse, qui permet de faire croître de nouveaux vaisseaux sanguins à partir des vaisseaux existants. Ce processus est contrôlé par de nombreux signaux pro-angiogéniques ou anti-angiogéniques se liant aux récepteurs des cellules endothéliales (VEGF-A ou TSP-1, par exemple (54)). Ces signaux sont anormalement activés dans les tissus tumoraux.

Acquérir les capacités mentionnées ci-dessus nécessite que la cellule subisse des mutations de son génome. Ce phénomène se retrouve dans l'ensemble des cellules de l'organisme : des mutations surviennent à chaque division cellulaire. Ces mutations sont appelées mutations somatiques, par opposition aux variations dites constitutionnelles ou germinales, qui sont présentes dans toutes les cellules de l'individu. Les mutations somatiques apparaissent de façon plus ou moins aléatoire le long du génome, mais dans une cellule cancéreuse, certaines ont conféré un avantage sélectif à la cellule induisant le développement du cancer. Il est donc important de distinguer les mutations dites "driver", ayant des conséquences sur la cancerogénèse, des mutations dites "passenger", n'ayant pas d'influence sur le développement du cancer.

Par ailleurs, cela implique que le cancer a plus de risques de se développer dans des lignées cellulaires subissant beaucoup de mutations. Ainsi, 90% des cancers (87) apparaissent dans les tissus épithéliaux, dont les cellules se renouvellent continuellement. De plus, le taux de mutation des cellules peut être augmenté par des mutagènes, par exemple certaines irradiations ou le benzopyrène contenu dans la fumée de cigarette. En outre, si des mutations apparaissent dans des gènes impliqués dans les mécanismes de réparation de l'ADN, cela augmente l'instabilité du génome de la cellule et facilite l'apparition du cancer.

Grâce au développement des techniques de séquençage, de nombreuses analyses haut débit de cellules cancéreuses ont été conduites afin de mieux caractériser leur génome. Ces études ont révélé la complexité du paysage génétique des cancers. Premièrement, les types de modifications génétiques affectant les cellules cancéreuses sont très variés : mutations ponctuelles, variations dans le nombre de copie des gènes, anomalies dans le nombre de copies des chromosomes, réarrangements, altérations de la méthylation, etc. De plus, bien que certains gènes soient connus pour être impliqués dans de nombreux cancers, la plupart des gènes cancéreux identifiés diffèrent fortement suivant les cancers, mais aussi au sein d'un même type de cancer. Par ailleurs, le nombre de mutations somatiques (dont la grande majorité sont passenger) détectées dans les cancers est très élevé, typiquement autour de 1 000 à 10 000 mutations (88), et très variable (de quelques mutations à 100 000 pour certains cancers du poumon). Une étude récente de Vogelstein et ses collègues (53) a évalué le nombre de mutations altérant la fonctionnalité des protéines à entre 33 et 66 par tumeur. De plus, sur 3 248 tumeurs analysées à haut débit, 125 des $\sim$ 20 000 gènes séquencés sont présumés driver du cancer.

Malgré les progrès importants réalisés ces dernières années dans notre compréhension du cancer et des modifications génomiques sous-jacentes, il reste encore beaucoup à découvrir. Par exemple, Vogelstein et ses collègues (53) estiment que la majorité des gènes fortement mutés (appelés "mountains") ont été détectés mais que les gènes plus difficiles à identifier car mutés dans un petit nombre de tumeurs ("hills")

	Broad	**WU**	**Baylor**
Séquenceur	Illumina GA II	Illumina GA II	SOLiD
Aligneur	BWA (21)	MAQ3 (89)	Bfast (90)
Appel des mutations somatiques	Somatic Sniper (91) VarScan2 (93)	MuTect (92)	Samtools (36)

TAB. 4.1 – **Pipelines utilisés par les différents centres.** Broad : Broad Institute, WU : Washington University Genome Sequencing Center, Baylor : Baylor College of Medicine

sont très peu caractérisés.

4.2 L'analyse TCGA du cancer de l'ovaire

The Cancer Genome Atlas a catalogué les aberrations moléculaires existant dans 489 adénocarcinomes séreux ovariens de haut grade (5). Des analyses d'expression d'ARNs messagers et de micro-ARNs, de méthylation des promoteurs de la transcription et de variabilité du nombre de copies de gènes (CNV) ont été réalisées pour ces tumeurs et pour des tissus sains provenant des mêmes patientes. De plus, une grande partie de l'exome ($\sim$ 180 000 exons provenant d'environ 18 500 gènes) de 316 paires d'échantillons cancer/sain a été séquencée. Nous nous sommes intéressé plus précisément aux données de séquençage d'exome. Le séquençage et l'alignement des short-reads ont été réalisés dans 3 grands centres : Broad Institute, Baylor College of Medicine et Washington University Sequencing Center. La table 4.1 liste les pipelines utilisés par chacun des centres pour séquencer, aligner et appeler les mutations somatiques.

De nombreuses mutations somatiques ont été identifiées, sur lesquelles deux algorithmes (MuSiC (75) et MutSigCV (74)) ont été appliqués afin de différencier les mutations driver des mutations passenger. Ces algorithmes calculent le taux de mutation moyen ("background mutation rate") des cellules cancéreuses, et sélectionnent les gènes significativement plus mutés que ce qui est attendu au hasard. Grâce aux 19 356 mutations somatiques annotées, plusieurs gènes driver ont été identifiés : le gène TP53, muté dans 96% des tumeurs (une partie des mutations a été détectée par re-séquençage à plus bas débit), et d'autres gènes à plus faible prévalence (NF1, BRCA1, BRCA2, RB1 et CDK12). Une autre méthode (CHASM (94)), basée sur un algorithme de classification de type forêt d'arbre décisionnel, a identifié 122 gènes potentiellement driver.

Les chiffres et les résultats présentés ci-dessus sont ceux de l'article paru en 2011 (5). Notre analyse, décrite dans les sections suivantes, utilise des données plus récentes. En particulier, les résultats de l'article TCGA se basent sur le séquençage de 316 tumeurs, alors qu'à la période où nous les avons téléchargées, les données de séquençage de 520 patientes étaient disponibles.

Comme beaucoup d'analyses haut débit de tumeurs, l'étude TCGA ne s'intéresse qu'aux variations somatiques, et ignore complètement les variations germinales. L'analyse des mutations somatiques est importante car elle peut faire progresser notre connaissance des mécanismes conduisant au cancer et aider à découvrir de nouveaux traitements anti-cancéreux. D'un autre côté, l'identification de facteurs de risque génétiques, présents dans les variations germinales, permet d'améliorer les outils de diagnostic (78). Cet aspect est fondamental pour les patients, puisqu'un diagnostic précoce influe énormément sur l'issue des traitements. Cela est particulièrement vrai dans le cas du cancer de l'ovaire : ce cancer a un pronostique assez sombre, car souvent diagnostiqué au grade 3 ou 4, quand l'efficacité des traitements s'effondre (taux de survie à 5 ans d'environ 40%) (95). Le retard du diagnostic est dû à la non spécificité des symptômes de la maladie (douleurs dorsales, abdominales, symptômes urinaires) et il est donc particulièrement important

de se concentrer sur les outils de diagnostic génétique.

Les données massives produites par le consortium TCGA peuvent aider à répondre à ces questions. Il s'agit de la plus grosse étude de séquençage jamais réalisée sur le cancer de l'ovaire : les exomes sains de 520 patientes ont été séquencés à très haute couverture (plus de 300x), ce qui permet une caractérisation précise des variants germinaux.

4.3 Variants germinaux et prédisposition au cancer de l'ovaire

Plusieurs études d'association gènes maladies ont permis d'identifier des facteurs de risque génétiques associés au cancer de l'ovaire. Cette section résume les résultats obtenus.

De nombreux gènes impliqués dans la réparation de l'ADN sont connus pour leur impact sur le développement du cancer de l'ovaire. En particulier, des mutations des gènes BRCA1 et BRCA2 ont été observées chez 5 à 15% des patientes ((96, 97, 98), par exemple), et le risque de développer la maladie est élevé pour les porteuses de ces mutations (risque cumulé à 70 ans de 59% contre 16,5%) (99). La protéine BARD1 (BRCA1-associated RING domain protein 1) interagit avec BRCA1 pour former un complexe nécessaire à la stabilité de BRCA1. Cette interaction pourrait bien être un aspect essentiel de la fonction de suppresseur de tumeur de BRCA1 et elle est souvent perturbée dans les cancers associés à BRCA1 (100). Le syndrome de Lynch, caractérisé par des mutations des gènes également impliqués dans les mécanismes de réparation de l'ADN (MSH2, MLH1 ou MSH6 ou plus rarement EPCAM), augmente le risque de développer un cancer de l'ovaire (101), mais de façon moins prononcée (observé dans 2% des cas (102)). Les 15 gènes "FA" impliqués dans l'anémie de Fanconi, une maladie génétique provoquée par une anomalie dans la réparation de l'ADN, sont également de potentiels candidats. Parmi eux, FANCD1 n'est autre que BRCA2 et BRIP1 (103), RAD51C (104, 105) et RAD51D (106) ont été identifiés comme potentiels facteurs de risque dans le cancer de l'ovaire. La liste des candidats a été étendue à d'autres gènes du même réseau de régulation : CHEK2, BARD1, MRE11A, NBN, RAD50 et PALB2. Ces gènes, ainsi que d'autres suppresseurs de tumeurs (BRCA1, MSH6, RAD51C et TP53) ont été analysés par une étude de Walsh et ses collègues (107), et présentent des mutations chez des patientes atteintes de cancer de l'ovaire. Cependant, cette étude, tout comme d'autres études des variants germinaux (108, 96), ne compare pas les fréquences des SNVs des patientes avec une population contrôle. Étant donné le grand nombre de mutations "de novo" détectées par chaque expérience haut débit, les conclusions de ces recherches sont difficilement interprétables, et des comparaisons cas/contrôle devraient être obligatoires dans toute publication de ce type. Récemment, une étude (109) a identifié PPM1D, une protéine activée par P53 et impliquée dans le déclenchement de l'apoptose, comme clairement associée aux cancers du sein et de l'ovaire.

D'autres variants associés au cancer de l'ovaire ont été identifiés par des études d'association (9p22.2 (110), 2q31, 8q24 (111), 19p13 (112)) basées sur des puces à SNPs, mais leur effet reste indéterminé, impliquant potentiellement les gènes HOXD1, MYC, TIPARP, SKAP et C19orf62 (101). Ces variants, ainsi que d'autres variants liés au cancer du sein et de la prostate, ont été rassemblés dans une puce à SNPs spécifique au cancer (appelée iCOGS) qui a été utilisée par le consortium COGS (Collaborative Oncological Gene-environment Study) sur de grandes cohortes de patients atteints de ces 3 types de cancers. Les résultats de l'étude COGS, publiés conjointement dans 13 articles, sont résumés par Sakoda et ses collègues (113), avec notamment 14 variants associés au cancer de l'ovaire (voir table 4.3).

Les données de séquençage d'exome produites par le consortium TCGA sont un excellent moyen de confirmer ou d'infirmer ces hypothèses, et d'en proposer de nouvelles.

Disease subtype	SNP	Chr.	Position	RAF	OR	Nearest gene	Ref	R. Hom.	Het.	V. Hom.
Invasive	rs711830	2	177037311	32 %	1.12	HOXD3	(114, 111)			
Invasive		3	157917010	5 %	1.44	TIPARP	(114, 111)			
BRCA1	rs4691139	4	165908721	48 %	1.2	TRIM61	(115)			1
Serous	rs10069690	5	1279790	26 %	1.09	TERT	(116)			
LMP	rs7705526	5	1285974	33 %	1.51	TERT	(116)			
Invasive	rs11782652	8	82653644	7 %	1.19	CHMP4C	(114)			
Invasive	rs7814937	8	129541475	87 %	1.18	PVT1	(114, 111)			
Invasive	rs3814113	9	16915021	68 %	1.21	BNC2	(114, 110)			
Invasive	rs7084454	10	21821274	33 %	1.1	MLLT10	(114)			
Serous	rs7405776	17	36093022	38 %	1.12	HNF1B	(114)		1	
Clear cell	rs11651755	17	36099840	51 %	1.29	HNF1B	(117)	4	1	1
Serous	rs2077606	17	43529293	17 %	1.15	PLEKHM1	(118)			
Invasive	rs7218345	17	46502917	27 %	1.12	SKAP1	(114, 111)			
Serous	rs8170	19	17389704	19 %	1.14	BABAM1	(114, 112)	103	29	2

TAB. 4.2 – **Variants germinaux identifiés comme associés au cancer de l'ovaire par le consortium COGS, et comparaison avec les données TCGA.** RAF : "Risk Allele Frequency", la fréquence de l'allèle dans la population générale, OR : odds ratio, R. Hom., Het. et V. Hom. : le nombre de patientes TCGA portant le génotype homozygote référence, hétérozygote ou homozygote variant, respectivement. La majorité des variants identifiés par COGS sont intergéniques, et leurs positions sont très peu couvertes par les short-reads TCGA. Seule rs8170, une mutation synonyme du gène BABAM1, a été détectée dans un nombre important de patientes.

4.4 Séquençage de l'exome de 520 patientes atteintes du cancer de l'ovaire : analyse des variants germinaux

Le consortium TCGA a séquencé à haute couverture l'exome des tissus sains et tumoraux de 520 patientes. Étonnement, les variants germinaux n'ont pas du tout été analysés par cette étude, ni même identifiés. Ainsi, la première étape a été de caractériser les génomes sains de ces patientes. TCGA met à disposition de la communauté scientifique des fichiers BAM correspondant aux alignements réalisés pour chaque expérience. Ces alignements ont été obtenus grâce à plusieurs pipelines, résumés table 4.1. Afin d'unifier les données, et de réaliser nos propres contrôles qualité, nous avons choisi d'analyser ces données à partir des short-reads et non des alignements. Ainsi, nous avons téléchargé sur CGHub (Cancer Genomics Hub) les fichiers d'alignements d'exome disponibles, desquels nous avons extrait les short-reads "d'origine", c'est-à-dire le résultat des séquenceurs. Nous avons ensuite aligné l'ensemble des short-reads avec l'aligneur MAGIC, et enfin nous avons appelé les génotypes et filtré les erreurs de séquençage grâce à la méthode décrite dans le chapitre 3.

Pour certains échantillons, plusieurs expériences de séquençage ("runs") ont été réalisées. Afin d'obtenir un seul jeu de variants par patiente, nous avons fusionné les génotypes appelés pour les runs provenant des mêmes échantillons. Pendant cette phase, seuls les SNVs ayant le même génotype pour tous les runs du même échantillon ont été gardés, ce qui permet de réduire le nombre de faux positifs. De surcroît, en plus des filtres décrits dans le chapitre 3, pour qu'un variant soit retenu, nous avons également requis qu'au moins une des patientes présente l'allèle de référence. En effet, une position pour laquelle l'ensemble des patientes portent l'allèle variant de façon homozygote est très probablement une erreur dans le génome de référence ou un artefact de capture, de séquençage ou d'alignement. En outre, afin de filtrer certaines des erreurs restantes, nous avons comparé les génotypes des tissus sains aux génotypes des tissus tumoraux. En effet, si une patiente porte un variant germinal homozygote, il est hautement improbable que l'allèle de référence réapparaisse par hasard dans la tumeur. Ainsi, les SNVs pour lesquels la patiente présente l'allèle homozygote variante dans le tissus sain et au moins une version de l'allèle de référence dans le tissu tumoral sont suspects, et nous les avons donc filtrés ($0,02\%$ des SNVs homozygotes filtrés). Nous estimons qu'une

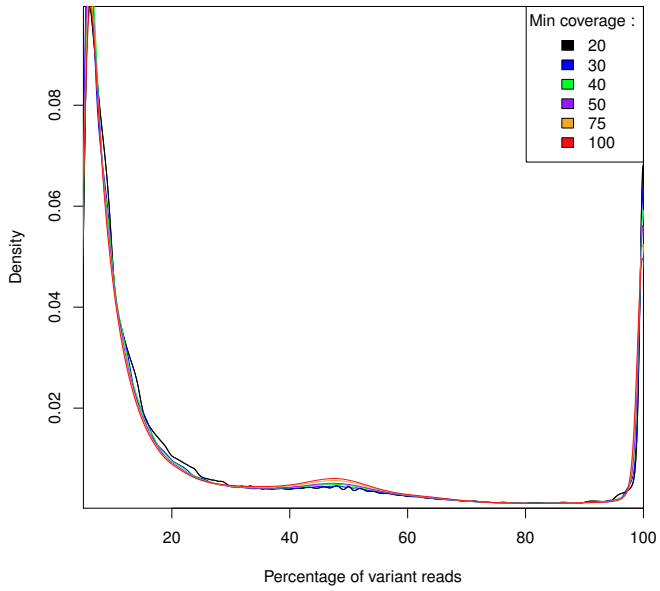


FIG. 4.1 – **Densité des proportions de reads variants, pour différentes valeurs de couverture minimale, pour les SNVs des échantillons tumoraux.** Comme les homozygotes références évidents ne nous intéressent pas, les positions représentées sont restreintes à celles contenant au minimum 5% de reads variants. Nous considérons que les SNVs présentant moins de 70% de read variants portent au moins une version de l'allèle de référence.

patiente porte l'allèle de référence dans le tissu tumoral si la position considérée présente moins de 70% de reads variants, pour une couverture totale d'au moins 20x. Ce choix a été réalisé grâce à l'observation de la densité des pourcentages de read variants dans les SNVs tumoraux (voir figure 4.1).

Au total, 5 649 Gb (giga paires de bases) ont été alignées avec une couverture moyenne de 324x, et 7 798 712 positions ont été examinées pour un potentiel appel de génotype. Les positions examinées sont les positions vues avec une couverture minimale totale de 10x et un taux de reads variants supérieur à 20% chez au moins une patiente. Finalement, 15 924 750 variants ont passé nos filtres qualité et ont été appelés sur 868 855 positions (voir table 4.3). Cela représente $\sim$ 30 625 variants en moyenne par échantillon, soit un peu plus que la moyenne de $\sim$ 24 000 variants observée par le 1000 Genomes Project(119). Étant donné la différence de couverture entre les deux expériences (324x pour nos données contre 80,3x pour les données du 1000 Genome), ce résultat n'est pas surprenant : *a priori*, plus la couverture est élevée, plus le taux de faux négatif est bas.

La plupart de ces variants n'ont pas de lien avec le cancer, et constituent simplement des différences génétiques normales entre les individus. Le défi est alors d'identifier parmi cette multitude de variants ceux constituant un facteur de risque du cancer.

4.5 Sélectionner les variants "loss-of-function"

Pour des raisons techniques, il est souvent plus simple de se restreindre aux SNVs inactivant la fonction de la protéine encodée (variants "loss-of-function", LoF) : ceux-ci sont beaucoup plus faciles à interpréter biologiquement, surtout quand l'activité de la protéine affectée est bien caractérisée. De plus, d'après

	Nb loci	Nb hétérozygotes	Nb homozygotes	moyenne/patient	Nb gènes
Non filtré	868 855	9 560 195	6 364 555	30 625	12 397
LoF	21 014	149 634	48 613	381	10 944
Non filtré significatif	10 434	372 256	337 052	1 364	3 214
LoF significatif	415	12 096	10 611	44	334

TAB. 4.3 – Cette table liste, pour différents jeux de variants, le nombre de loci (associés à un SNV, par exemple “délétion d’un A à la position 3 245 644 du chromosome 17”), le nombre de SNVs hétérozygotes et homozygotes (somme sur toutes les patientes), la moyenne du nombre de variants par patiente et le nombre de gènes contenant au moins une mutation chez au moins une patiente. “Non filtré” contient l’ensemble des SNVs ayant passé les filtres qualités, “LoF” ceux induisant une perte de fonction des protéines, pour celles très exprimés dans l’ovaire. “Non filtré significatif” et “LoF significatif” se restreignent aux variants ayant des fréquences alléliques significativement différentes chez les cas et chez les contrôles (comme détaillé section 4.6.3).

notre connaissance actuelle des mécanismes cellulaires, les mutations LoF ont plus probablement une incidence sur le fonctionnement des cellules, et donc sur le développement d’une maladie. C’est pourquoi nous avons décidé de regarder avec plus d’attention les variants LoF de nos patientes. Évidemment, faire ce choix implique potentiellement de perdre des informations importantes, car il est possible qu’une partie des mutations non-codantes aient des conséquences sur le développement de la maladie (120), mais celles-ci sont *a priori* peu nombreuses et difficiles, voir impossibles, à identifier.

Dans un premier temps, nous avons sélectionné les mutations affectant les parties supposées codantes du génome, à partir des coordonnées intron/exon de la base de données Aceview (6). Aceview aligne toutes les séquences publiques d’ARNm, d’ADNc, d’EST, et de RNA-seq (ARNm de RefSeq et GenBank, ADNc de dbEST et Trace, et les séquences NGS de SRA et GEO) sur le génome de référence, et les regroupe pour obtenir un nombre minimal de transcrits. Il en résulte 37 463 gènes encodant plus de 205 000 transcrits. Cette base cherche à être aussi complète que possible, et inclut des transcrits peu soutenus. Ainsi, parmi l’ensemble des transcrits Aceview, nous nous sommes restreints à ceux annotés comme codant pour une “bonne protéine”, soit 191 507 transcrits. La classification des transcrits Aceview repose entre autres sur la taille des protéines présumées et la présence d’introns et de domaines protéiques.

Ainsi, grâce aux coordonnées de ces transcrits, nous avons classé les mutations ponctuelles en plusieurs catégories : introniques ou intergéniques (hors pieds d’introns), synonymes, faux sens, non-sens (introduction d’un codon STOP), perte d’un codon d’initiation, perte d’un codon STOP et mutations affectant l’épissage (pieds d’introns à ± 2 bp). Quant aux insertions et aux délétions, nous les avons cataloguées en : perte d’un ou plusieurs acides aminés, décalage du cadre de lecture et mutations affectant l’épissage. L’effet des mutations intergéniques, introniques ou silencieuses est assez mal connu. Quelques-unes semblent altérer l’épissage alternatif ou la stabilité des ARNm (120), mais dans l’ensemble, estimer leur répercussion est très difficile. Ces variants ont donc été ignorés. Les mutations franches, comme les non-sens, les pertes de codons STOP et d’initiation ou les décalages du cadre de lecture, sont au contraire probablement LoF. L’incidence des mutations faux sens ou des pertes d’acides aminés est plus délicat à estimer. Par exemple, alors que l’introduction d’un acide aminé hydrophile dans un domaine hydrophobe peut complètement modifier la conformation de la protéine et la rendre non fonctionnelle, certains changements d’acide aminé n’ont que très peu de conséquences. Plusieurs logiciels (*e.g.* (8, 9)) permettent de prédire l’impact d’une mutation sur la protéine en se basant par exemple sur la conservation de la séquence entre espèces ou les caractéristiques biochimiques des acides aminés. Cependant, une étude a mis en évidence les désaccords (entre $40,6\%$ et $28,6\%$) existant entre les résultats de 4 logiciels différents (121). Ainsi, pour limiter les erreurs,

nous avons combiné les résultats de 2 logiciels, SIFT (8) et Polyphen-2 (9), qui utilisent des approches complémentaires, et nous avons conservé les mutations que les 2 logiciels prédisent LoF.

Néanmoins, la modification de la séquence ne suffit pas à déterminer l'effet d'une mutation. Par exemple, il est vraisemblable qu'une mutation impliquant une protéine non exprimée dans le tissu n'aura pas d'effet sur le fonctionnement de la cellule. Nous avons donc également pris en compte le taux d'expression des protéines. Pour ce faire, nous avons utilisé les données Aceview : le taux d'expression des transcrits a été estimé dans chacun des tissus grâce à l'ensemble des données d'expression alignées par Aceview (Thierry-Mieg Jean et Danielle, communication personnelle). Ces données nous ont permis de conserver uniquement les mutations altérant les gènes fortement exprimés dans l'ovaire. Les variants appelés LoF dans la suite du manuscrit sont les variants induisant une perte de fonction de la protéine, uniquement pour les protéines exprimées dans l'ovaire.

Nous obtenons ainsi pour l'ensemble des patientes respectivement 149 634 et 48 613 variants LoF hétérozygotes et homozygotes, répartis sur 21 014 positions. Cela représente en moyenne $\sim$ 380 variants LoF annotés par individu, soit un peu plus que les 250-300 estimés par le 1000 Genomes Project (46). Comme pour les variants non filtrés, cet écart s'explique par la différence de couverture entre les deux expériences. Sur l'ensemble des patientes, les variants LoF impliquent 10 944 des 13 625 gènes exprimés dans l'ovaire.

Nous disposons donc de deux listes de SNVs présents chez les patientes étudiées, l'une sur l'ensemble de l'expérience, appelée "SNVs non filtrés", et l'autre se restreignant aux variants LoF ("SNVs LoF"). Les gènes de ces listes sont parfois appelés "gènes LoF" ou "gènes non filtrés" dans la suite du manuscrit. Afin d'identifier parmi ces mutations celles potentiellement impliquées dans le développement du cancer, nous avons comparé les fréquences alléliques de nos patientes à celles observées dans des populations contrôles. Cela nous a permis de sélectionner les mutations significativement plus présentes dans le cancer de l'ovaire (appelées mutations significatives). Nous avons ensuite comparé l'ensemble des gènes présentant des mutations significatives avec les gènes connus pour être impliqués dans le cancer, puis nous avons réalisé une étude d'enrichissement en termes GO (Gene Ontology).

4.6 Étude d'association

4.6.1 Comparaison à une population contrôle

Les études d'associations "classiques" (GWAS, Genome Wide Association Studies) utilisent généralement des puces à SNPs pour génotyper un grand nombre d'individus malades (cas) et sains (contrôles) à des positions spécifiques, puis identifient les SNPs significativement associés à la maladie par des modèles statistiques. Par exemple, dans le cas de l'étude COGS (113), 22 252 loci ont été génotypés chez 26 105 cas et 35 350 contrôles. Le coût du séquençage de l'exome complet étant bien supérieur à celui d'un génotypage par puce, le nombre de cas analysés par séquençage est d'ordinaire 1 à 2 ordres de grandeur plus faible, et les études ne séquencent généralement pas de contrôle. C'est peut-être ce qui explique que bon nombre de ces études ne comparent pas les fréquences alléliques obtenues à celles d'une population contrôle. Pourtant, des données d'exomes de plusieurs populations sont disponibles publiquement, produites par les projets 1000 Genomes Project (46) et ESP6500 (Exome Sequencing Project) (47). La phase 1 du 1000 Genomes Project a publié le séquençage à haute couverture (80x en moyenne) de 1 092 individus échantillonnés dans 14 populations, et ESP6500 a séquencé l'exome de 6 503 individus atteints de maladies du coeur, du sang ou des poumons provenant de cohortes de divers hôpitaux américains. Nous avons utilisé ces données comme contrôle dans notre étude du cancer de l'ovaire.

Pour réaliser une étude d'association fiable, les individus malades doivent être comparés à des individus

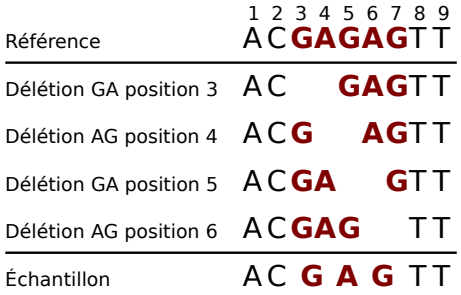


FIG. 4.2 – **Indels équivalents.** Sur cet exemple, des délétions du dinucléotide GA (ou AG) aux positions 3, 4, 5 ou 6 conduisent à la même séquence. Le problème est symétrique dans le cas d'insertions. L'appel et l'identification des indels ne sont pas unifiés dans les bases de données principales. Une manière simple de les harmoniser serait de choisir une convention, par exemple d'utiliser la position équivalente la plus à gauche, ce que nous avons fait dans notre analyse.

sains provenant de la même population. En effet, les études de génétique des populations montrent que les fréquences alléliques dépendent de la population échantillonnée : typiquement, un individu d'origine africaine n'aura pas les mêmes fréquences alléliques qu'un individu d'origine européenne. Ainsi, utiliser comme contrôle une population éloignée de la population malade peut conduire à de fausses associations, car les SNPs spécifiques de la population ne pourront pas être différenciés des SNPs associés à la maladie. Les patientes de l'étude TCGA sont américaines, pour la plupart d'origine européenne avec quelques patientes d'origine africaine, asiatique ou hispanique. Les données ESP6500 provenant d'hôpitaux américains, nous avons supposé que l'ensemble des individus de cette étude est assez représentatif de notre population de cas. Quant aux données du 1000 Genomes, nous avons gardé uniquement les génotypes des 302 individus appartenant à la population "EUR" (pour européens) d'origine américaine, italienne, finlandaise, espagnole ou anglaise.

4.6.2 Équivalence des indels

Avant de comparer les fréquences alléliques des populations, il est important de noter que, contrairement aux SNVs, les indels ne sont pas identifiés de manière unique par leur position et la description de la mutation. En effet, si la séquence génomique environnante présente des répétitions, les délétions (ou, de manière symétrique, les insertions) de différentes régions peuvent conduire à la même séquence (voir exemple figure 4.2). Krawitz et ses collègues (122) ont proposé un algorithme simple pour lister l'ensemble des indels équivalents d'une séquence, dont nous avons implémenté une version modifiée qui cherche la position la plus à gauche des positions équivalentes. Cet algorithme a été appliqué à nos données ainsi qu'aux données de ESP6500 (les données du 1000 Genomes ne contiennent pas d'indel), afin de pouvoir les comparer de façon correcte. Nous avons pu remarquer que ni les données d'ESP6500, ni les identificateurs "rs" (identificateurs de dbSNP) ne sont cohérents à ce sujet : il existe des redondances, c'est-à-dire des positions équivalentes correspondant à 2 entrées différentes et ne présentant pas la même fréquence allélique. Par ailleurs, une étude récente (123) a identifié plus d'un million (sur $\sim$ 6 millions) d'entrées redondantes dans la base de données Ensembl. Il est surprenant d'observer un si grand nombre d'erreurs alors qu'une solution très simple consiste à définir une convention, par exemple utiliser la position équivalente la plus à gauche.

4.6.3 Sélection des variants significativement associés au cancer

Une fois les positions équivalentes à gauche calculées pour les indels des données TCGA et ESP6500, le but est de sélectionner les SNVs et les indels associés au cancer en comparant les fréquences alléliques entre les cas et les contrôles. Un problème des données exome est que le génotypage n'est généralement fait qu'aux positions d'intérêt, c'est-à-dire les positions variantes chez au moins un individu de la population. Par conséquence, les fichiers de variants ne suffisent pas pour différencier les positions où le génotype n'a pas été appelé car la couverture n'était pas suffisante des loci où toutes les patientes étaient homozygote référence. Dans un cas, la fréquence allélique de la population est disponible et peut être comparée à la fréquence chez les malades, alors que dans l'autre elle est inconnue. Pour discriminer ces deux cas, nous avons utilisé les fichiers de couverture fournis par ESP6500 : pour chaque position, on dispose de la couverture moyenne et du nombre d'individus pour lesquels la position est couverte. Ainsi, nous estimons qu'une position absente de la liste des SNVs d'ESP6500 est homozygote référence chez le nombre d'individus couverts, dans le cas où la couverture moyenne est supérieure à 15x. Pour une position couverte à moins de 15x, nous considérons les données comme insuffisantes et nous ignorons la position. Nous avons ensuite comparé les fréquences alléliques de l'ensemble des patientes avec les 2 populations contrôles indépendamment grâce au test exact de Fisher. Les p-valeurs obtenues ont ensuite été triées et nous avons appliqué la procédure de Benjamini-Hochberg pour corriger pour la multiplicité des comparaisons, toujours indépendamment pour chaque contrôle. Finalement, les variants sont sélectionnés s'ils sont significativement plus présents ($\alpha = 5\%$) dans le cancer que dans la(es) population(s) contrôle(s) pour lesquelles une fréquence allélique existe (les positions absentes de 1000 Genomes et non couvertes par ESP6500 sont ignorées). De plus, nous conservons seulement les variants observés chez au moins 2 patientes, toujours dans une optique de limiter au maximum le nombre de faux positifs. À partir de la liste "SNVs non filtrés", nous obtenons ainsi 10 434 variants, appelés "variants significatifs", présentant des fréquences alléliques chez les patientes atteintes du cancer de l'ovaire significativement supérieures aux fréquences dans les populations contrôles. Ces variants sont localisés dans 3 214 gènes. Quant aux mutations LoF, nous sélectionnons parmi celles-ci 415 variants significatifs impliquant 334 gènes.

4.7 Comparaison des gènes mutés avec les gènes connus comme potentiellement impliqués dans le cancer

L'étape suivante consiste à comparer les résultats obtenus avec ceux des autres études sur la génétique des cancers. Comme expliqué dans l'introduction de ce chapitre, COGS a identifié 14 SNPs facteurs de risque dans le cancer de l'ovaire. Nous voulions comparer les fréquences alléliques observées dans l'ensemble de nos patientes avec celles identifiées par cette étude. Malheureusement, la plupart des SNVs identifiés par COGS se trouvent probablement hors de la zone capturée par TCGA, car très peu de loci ont assez de couverture pour qu'un génotype soit appelé (table 4.3). Seul le SNV rs8170, mutation synonyme du gène BABAM1 (ou MERIT40), est observé chez un nombre important de nos patientes, mais la taille de la population des cas TCGA ne permet pas de détecter d'association au cancer.

Par ailleurs, nous avons comparé les gènes identifiés dans notre étude avec les gènes connus pour être impliqués dans le cancer (appelés "gènes cancer"). Nous disposons de deux listes de gènes "candidats" présentant des SNVs significativement plus présents dans nos données : 3214 gènes "non filtrés significatifs", et 334 gènes "LoF significatifs" (voir table 4.3). Il est à noter que 40 gènes LoF significatifs sont absents de la liste non filtrée, car la correction pour tests multiples est plus sévère dans le cas des mutations non filtrées. Par ailleurs, nous avons choisi 5 listes de "gènes cancer". La première

	Gènes	Curation de la littérature	Vogelstein	Glad4U	ClinVar	OMIM
Nb gènes	334	40	124	2273	125	184
Intersection		5	7	37	5	7
P-valeur		$5.60 \cdot 10^{-5}$	$3.13 \cdot 10^{-4}$	$1.40 \cdot 10^{-3}$	$9.71 \cdot 10^{-3}$	$3.08 \cdot 10^{-3}$
	ATM	X	X	X	X	X
	BRCA1	X	X	X	X	X
	NF1	X	X	X		
	NQO1			X	X	X
	BARD1			X	X	X
	MAP3K8			X		X
	PARP1	X		X		
	CRTC1					X
	MLL3		X			
	NOTCH2		X			
	RAD51D	X		X	X	X
	SPOP		X			
	MLL2		X			
	EPOR			X		
	USP33			X		
	ERBB3			X		
	PTGER3			X		
	STC1			X		
	UTP14C			X		
	CD34			X		
	HSPG2			X		
	HSPA8			X		
	SRSF3			X		
	SMARCA5			X		
	SPAG9			X		
	RAD23B			X		
	TNFSF10			X		
	CD248			X		
	PLXND1			X		
	SPON1			X		
	EPHA2			X		
	KRT18			X		
	LMNA			X		
	HELLS			X		
	MLLT4			X		
	LTBP1			X		
	MUC16			X		
	NR4A1			X		
	BIRC6			X		
	CTBP2			X		
	ENTPD1			X		
	IRF7			X		

TAB. 4.4 – **Intersection entre les gènes LoF issus de notre analyse des données TCGA et les gènes connus pour être impliqués dans le cancer.** Nous avons cherché parmi les 235 gènes LoF ceux présents également dans d'autres listes de gènes impliqués dans le cancer. La colonne "Gènes" contient tous les gènes de notre liste de candidats apparaissant dans au moins une autre liste. La ligne "Intersection" donne le nombre de gènes présents dans notre liste et dans la liste concernée. La P-valeur est le résultat du test exact de Fisher comparant les proportions de gènes cancer (*i.e.* appartenant à la liste concernée) contenus dans notre liste de candidats et dans l'ensemble des gènes capturés par l'expérience. Pour une description des listes de gènes, voir section 4.7.

est une liste que j'ai construite par une revue manuelle de la littérature. Celle-ci, basée sur les articles (116, 124, 125, 126, 127, 128, 129, 130, 115, 101, 131, 122, 96, 97, 99, 107, 103, 98, 105, 106, 104, 109), contient 40 gènes et est décrite dans la section 4.3. Elle est restreinte aux gènes couverts par la capture TCGA. La liste de gènes couverts (ou liste "background") est en réalité le résultat d'une approximation réalisée grâce aux fichiers de "coverome" produits par MAGIC. Ces fichiers donnent la couverture moyenne par région génomique de taille variant de $\sim$ 50 à 400 bp. La liste "background" contient l'ensemble des gènes recouvrant, même partiellement, une région couverte pour au moins un échantillon et avec une couverture d'au moins 10x, soit 32 266 gènes. Les gènes associés au cancer provenant des bases des données OMIM (132) et ClinVar (133) constituent deux autres listes de 184 et 125 gènes respectivement, après suppression des redondances et des gènes non couverts par notre étude. De plus, une revue récente de Vogelstein et ses collègues (53) a identifié 125 gènes potentiellement driver du cancer (dont 124 sont couverts par nos données) parmi les 18 306 gènes affectés par des mutations de la base de données COSMIC (Catalogue Of Somatic Mutation In Cancer (134)). Finalement, nous avons effectué la requête "ovarian cancer" dans GLAD4U (135), une application qui cherche les publications et les gènes associés à des maladies, et obtenu 2 273 gènes.

Afin de faire correspondre les identifiants utilisés dans nos listes de gènes candidats (identifiants de la base de données Aceview) et ceux des listes de gènes cancer (divers identifiants dont principalement les "gene symbols"), nous avons utilisé les fichiers `ftp://ftp.ncbi.nih.gov/repository/acedb/ncbi_37_Aug10.human.genes/AceView.ncbi_37.mRNA2GeneID2NM.txt.gz` et `ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz`, fournis par le NCBI.

42 gènes LoF candidats sont présents dans au moins une liste de gènes cancer, et jouent très probablement un rôle dans le développement de la maladie chez nos patientes (table 4.4). La liste de candidats LoF est significativement enrichie en gènes soupçonnés d'être impliqués dans le cancer ($P = 9.71 \cdot 10^{-3}$ à $P = 5.60 \cdot 10^{-5}$), ce qui renforce la confiance que l'on peut accorder à notre méthode. De plus, cela permet de mieux prioritariser les gènes de cette liste : les candidats appartenant aux listes cancer confèrent probablement un risque accru de développer la maladie.

Parmi ces gènes, le rôle de BRAC1, NF1, ATM, BARD1 et RAD51D a été décrit dans l'introduction, et je présente ci-dessous quelques exemples de bons candidats à être impliqués dans le développement du cancer de l'ovaire (gène "de susceptibilité").

Par exemple, l'un des 4 récepteurs de la voie de signalisation Notch, NOTCH2, présente des variants LoF chez nos patientes. Cette voie est importante pour la communication entre cellules, et est notamment impliquée dans l'angiogénèse. NOTCH2 a été identifié comme différentiellement exprimé dans des tumeurs du sein et du côlon (136, 137, 138), et il a été suggéré que ce gène joue un rôle de suppresseur de tumeur dans des cancers du sein et des poumons (137, 139). NOTCH2 pourrait donc influencer la cancerogénèse chez les patientes présentant ces mutations.

La kinase MAP3K8, une enzyme capable de transférer un groupement phosphate d'une ATP vers une protéine, pourrait elle aussi être impliquée dans le développement du cancer de l'ovaire. Cette protéine active les réseaux de signalisation MAP kinase et JNK kinase, qui régulent entre autres la prolifération, l'expression des gènes, la différenciation, la mitose et l'apoptose. Le gène MAP3K8 a été identifié comme sur-exprimé dans plusieurs cancers (140, 141, 142, 143), et est un oncogène reconnu. Au contraire, une étude récente (144) démontre l'effet suppresseur de tumeur de ce gène, en identifiant des aberrations génétiques conduisant à une perte de la fonction de MAP3K8 dans des cancers du poumon. C'est ce que nous observons également chez 106 de nos patientes, qui présentent une mutation délétère sur ce gène.

Parmi les gènes pour lesquels moins de liens avec le cancer ont été établis, l'histone-lysine N-méthyltransférase MLL3 est une enzyme membre du complexe ASCOM, et est impliquée dans la méthylation des histones

et l'activation de la transcription. Quelques indices suggèrent que ce gène pourrait être associé à la leucémie (145, 146). Par ailleurs, 3 patientes présentent une mutation LoF homozygote sur le gène NQO1, sur-exprimé dans des adénocarcinomes du pancréas (147). L'association de ce gène à la leucémie aiguë myéloblastique, identifiée par Larson et ses collègues (148), n'a pas été retrouvée dans une étude plus récente (149). Quant au gène STC1, il code pour une glycoprotéine sécrétée qui pourrait avoir des fonctions autocrines et paracrines. Cette protéine semble jouer un rôle dans la prolifération et la migration cellulaire (150) et moduler la signalisation de l'ATP extracellulaire (151). Par contre, la mutation observée dans nos données implique une forme alternative non épissée du gène, dont le rôle n'est pas démontré. Meilleure candidate, la protéine SPOP inhibe l'expression de SRC-3, un oncogène reconnu, et semble donc jouer un rôle de suppresseur de tumeur (152, 153). Ces études ont identifié des pertes de fonction de SPOP dans des cancers du sein et de la prostate. Cette protéine présente une mutation LoF significative chez 3 de nos patientes.

En conclusion, nous avons identifié des mutations LoF significativement associées au cancer de l'ovaire altérant 42 gènes cancer. Nos connaissances sur l'implication de ces gènes dans le développement de la maladie sont très variables. Certains, comme ATM, BRCA1, BARD1, NF1, MAP3K8 ou PARP1 sont des gènes cancer connus, alors que d'autres sont très peu caractérisés. Nous confortons les hypothèses existantes et les étendons au cancer de l'ovaire en identifiant des mutations LoF altérant ces gènes chez nos patientes. En particulier, le gène MAP3K8 a récemment été indiqué comme suppresseur de tumeur (144), avec la découverte de mutations LoF affectant ce gène dans des cellules pulmonaires cancéreuses. C'est également ce que nous observons dans le cancer de l'ovaire : 106 de nos patientes présentent une mutation délétère sur ce gène. Quant aux 292 autres gènes affectés par des mutations LoF significatives chez nos patientes, ils sont d'excellents candidats à conférer un risque accru de développer le cancer de l'ovaire et des études fonctionnelles seraient nécessaires pour confirmer ces hypothèses.

Pour finir, 530 des 3 214 gènes candidats non filtrés sont présents dans au moins une liste de gènes cancer : notre liste de candidats non filtrée est donc significativement enrichie en gènes cancer ($P = 3,6 \cdot 10^{-4}$ à $P = 2,94 \cdot 10^{-81}$). Pour information, nous fournissons cette liste (table 5.1, en annexe). Les mutations affectant ces gènes sont souvent localisées dans les introns, les UTRs ou les régions non transcrites voisines. Certaines pourraient modifier la régulation de l'expression et contribuer ainsi à la cancérogénèse, alors que d'autres sont probablement sans conséquence. D'autres études seraient donc nécessaires pour discriminer, parmi les mutations non filtrées significatives, celles ayant un impact sur le phénotype.

4.8 Analyse d'enrichissement en terme GO

Afin d'identifier les mécanismes cellulaires altérés par les mutations présentes chez nos patientes, nous avons utilisé l'ontologie GO (Gene Ontology). Les termes GO décrivent les gènes selon 3 axes, les composants cellulaires, la fonction moléculaire et le processus biologique, et sont organisés hiérarchiquement, allant des descriptions très larges (par exemple, "pigmentation", ou "régulation des processus biologiques") à des définitions beaucoup plus précises, comme "régulation positive de la pigmentation de la cuticule". Cela permet à la communauté scientifique d'annoter les fonctions des gènes avec un vocabulaire contrôlé. Il est ensuite facile d'identifier les termes sur-représentés dans un ensemble de gènes, étude que nous avons réalisée grâce au logiciel DAVID (154, 155). DAVID identifie les termes GO significativement sur-représentés dans des gènes d'intérêt par rapport à une liste dite "background", qui contient l'ensemble des gènes auxquels les gènes d'intérêt doivent être comparés. La liste que nous avons utilisée comme "background" contient l'ensemble des gènes capturés par l'expérience (voir section 4.7 pour les détails du calcul de cette liste). Un test exact de Fisher modifié est réalisé pour chaque terme GO, suivi de la correction de

Benjamini-Hochberg (correction pour multiplicité des tests), pour identifier les termes GO significativement sur-représentés avec un seuil de 10%. DAVID propose 5 niveaux de spécificité des termes GO recherchés, et nous nous sommes intéressés à la catégorie la plus large des processus biologiques ("goterm_BP_1"), afin d'avoir une vue d'ensemble des données.

Le terme "metabolic process" est sur-représenté parmi les annotations des gènes LoF significatifs (P-valeur corrigée de $8,9\%$, 172 gènes). Ce terme s'applique à l'ensemble des gènes impliqués dans des réactions chimiques et des signaux de régulation qui transforment les substances chimiques. Cela inclue la transformation des petites molécules, mais aussi la réplication et la réparation de l'ADN et la synthèse et la dégradation des protéines. Quant aux gènes non filtrés, nous observons une sur-représentation des termes "cellular process" ($P = 2,12 \cdot 10^{-12}$), "biological adhesion" ($P = 4,56 \cdot 10^{-10}$), "metabolic process" ($P = 4,24 \cdot 10^{-7}$), "cellular component organization" ($P = 5,73 \cdot 10^{-6}$) et "cellular component biogenesis" ($P = 1,82 \cdot 10^{-3}$). Ces enrichissements sont cohérents avec une implication importante de nos candidats dans le développement du cancer et confirment là encore la validité de nos listes de gènes.

4.9 Conclusions

Notre étude constitue la plus grosse analyse des variations germinales de patientes atteintes du cancer de l'ovaire. Le séquençage à haute couverture de 520 patientes a permis de détecter en moyenne 30 632 variants par patiente. Afin d'identifier parmi ces derniers les variants responsables du phénotype, nous avons sélectionné d'une part les variants LoF, c'est-à-dire induisant une perte d'activité de la protéine encodée, et ce uniquement pour les protéines exprimées dans l'ovaire. D'autre part, nous conservons seulement les variants significativement plus présents chez nos patientes que dans des populations contrôles. Cela conduit à une moyenne de 44 SNVs par patiente, répartis sur 334 gènes dans l'ensemble de la cohorte : ces gènes pourraient jouer un rôle dans la cancerogénèse (gènes "candidats"). Afin d'identifier dans cette liste les gènes "de susceptibilité", nous avons comparé nos gènes candidats avec des gènes connus pour être associés au cancer, dont une partie provient d'une importante curation manuelle de la littérature : pour 42 d'entre eux, des études indiquent un lien avec la cancerogénèse. Un intérêt particulier doit être porté à la protéine MAP3K8, dont le rôle de suppresseur de tumeur a été très récemment proposé dans d'autre cancer, et sur laquelle 106 de nos patientes portent une mutation délétère.

Chapitre 5

Conclusions et perspectives

Au cours des dix dernières années, la taille et la complexité des données biologiques ont littéralement explosé, et une attention particulière doit être portée au contrôle qualité. En effet, certaines données omiques contiennent de nombreux faux positifs. À titre d'exemple, Musumeci et ses collègues (62) estiment qu'au moins 8% des SNPs de dbSNP sont des artéfacts dûs à des paralogues, et Cusick et ses collègues (156) évaluent à 35% le taux d'erreurs des données d'interaction protéine-protéine de curation de la littérature. De plus, plusieurs études montrent le peu de concordance existant entre les pipelines d'appel de génotypes (83, 84). D'autre part, il est important de garder à l'esprit qu'actuellement, les faux négatifs sont inhérents à tout type de données biologiques. Dans le cas des données interactomiques, nous montrons que la communauté scientifique est loin d'avoir cartographié l'ensemble des interactions existantes, même chez l'organisme modèle *S. cerevisiae*. Pour celui-ci, nous estimons le taux de faux négatifs des données de curation de la littérature à plus de 60%. De plus, la diversité génétique de la population humaine est très loin d'être caractérisée entièrement. Par exemple, le 1000 Genomes Project estime avoir découvert seulement $\sim$ 50% des SNPs ayant une fréquence de $\sim$ 1% dans la population, et évalue le taux de mutations "de novo" par génération par paire de base à $\sim 10^{-8}$, interdisant l'espoir d'obtenir un jour une cartographie complète des variations génétiques humaines. Avant toute étude, une première étape de tri est donc nécessaire afin de sélectionner les données les plus fiables, les plus complètes et répondant le mieux au problème posé.

À cela, on peut ajouter de nombreux biais, qui, s'ils ne sont pas identifiés et pris en compte, peuvent conduire à des erreurs d'analyse. Par exemple, les interactions protéine-protéine issues de curation de la littérature et les interactions identifiées à haut débit sont souvent considérés comme orthogonales, en particulier pour estimer la taille d'un interactome. Nous avons découvert que ces jeux de données sont beaucoup plus corrélés que ce qui est communément admis. La méthode du double hybride, la plus employée à haut débit, est également très utilisée à bas débit, et les interactions de curation de la littérature sont enrichies en interactions facilement détectables par double hybride. Cela remet en cause l'hypothèse d'indépendance de ces données et invalide donc les estimations actuelles de la taille des interactomes : elles sont sous-évaluées. Afin d'améliorer ces prédictions, nous avons examiné l'interactome de la levure d'un point de vue original, en prenant en compte le degré d'étude des protéines par la communauté scientifique. Nous avons découvert que l'ensemble des interactions protéine-protéine de la littérature est qualitativement différent quand

il est restreint aux protéines qui ont reçu une attention particulière de la communauté scientifique. En particulier, ces interactions s'appuient moins souvent sur la méthode double hybride, et plus souvent sur des expériences plus complexes, comme des analyses de l'activité biochimique. La corrélation existant entre les données haut débit et de curation de la littérature s'estompe en se restreignant aux protéines très étudiées. Nous proposons une méthode simple et fiable pour estimer la taille d'un interactome, en combinant les données concernant les protéines très étudiées par la communauté scientifique et les données haut débit. Notre méthode conduit à une estimation d'au moins 37 600 interactions physiques directes chez *S. cerevisiae*. Cette estimation est plus élevée et plus fiable que les estimations précédentes, car elle prend en compte les interactions difficiles à détecter avec les expériences communément utilisées. La présence de ces biais et notre réévaluation de l'estimation de la taille de l'interactome ont été pris en compte dans de nombreuses études (par exemple (157, 158, 159)). Nos travaux sur les protéines très étudiées ont été repris par Lewis et ses collègues (160, 161), qui s'intéressent à la prédiction d'interactions basée sur les protéines homologues (interologues). Ils ont notamment estimé le taux de conservation des interologues, et montré que, pour les protéines très étudiées par la communauté scientifique, des biais d'inspection (la connaissance d'une interaction va promouvoir la recherche de son interologue dans une autre espèce) induisent une sur-évaluation de ce taux. Une perspective de ce travail serait de prendre en compte le degré d'étude des protéines dans la prédiction des interologues et dans l'estimation de leur taux de conservation. Par exemple, on pourrait imaginer que la confiance dans la prédiction d'un interologue soit plus faible pour les protéines très étudiées : l'interaction a possiblement déjà été testée et non validée.

À travers l'exemple des données interactomiques, nous soulignons le fait qu'identifier les biais est critique pour une interprétation correcte et peut permettre de trouver des solutions pour les atténuer. Sur des données de séquençage de l'ADN, l'analyse des biais existant entre les reads alignés sur un brin ou sur l'autre nous a permis d'améliorer la qualité des génotypages. En effet, nous montrons qu'appeler les génotypes indépendamment sur chaque brin permet de filtrer un nombre important d'erreurs. De plus, nous constatons que les positions présentant des reads uniquement sur un brin sont très fréquentes, et qu'elles contiennent de nombreuses erreurs : elles doivent donc être filtrées. Par ailleurs, nous alertons la communauté scientifique sur l'usage général de seuils de couverture minimale trop faibles lors de l'appel des génotypes. Pour obtenir un génotype fiable, une couverture minimale de 10x sur chaque brin est nécessaire, et appeler des génotypes sur des positions présentant 4 reads doit être définitivement banni des pratiques. Nous soulevons un autre problème important concernant l'appel des génotypes : l'ensemble des positions contenant entre 20 et 40% de reads variants contient un mélange d'homozygotes références et d'hétérozygotes qui ne peuvent pas être différenciés en se basant uniquement sur le taux de read variants. Nous avons donc développé une extension simple de la méthode des seuils, basée sur l'utilisation de zones tampons et sur la comparaison des 2 brins, qui permet d'éliminer facilement ces positions. Nous avons testé cette méthode sur des exomes d'échantillons sains séquencés par le consortium TCGA, et la comparaison des nombreux réplicats techniques existants montre son efficacité, avec une diminution importante du taux d'erreur. Par contre, dans le cas d'échantillons tumoraux, le problème est plus délicat. En effet, les cellules cancéreuses sont très hétérogènes et les échantillons tumoraux sont souvent contaminés par les cellules saines voisines. De plus, le génome des tumeurs contient souvent des altérations à l'échelle chromosomique (variants structurels, CNVs, modifications de la ploïdie). Le pourcentage de reads variants attendu pour les hétérozygotes n'est donc pas 50%, et les mutations somatiques sont généralement appelées avec des pourcentages plus faibles, typiquement autour de $\sim$ 20-30%. Par conséquence, distinguer les vraies mutations somatiques

du bruit que nous observons ici reste difficile. De plus, nous montrons qu'une couverture élevée et une concordance entre les brins ne sont pas des conditions suffisantes pour éliminer toutes les erreurs, et nous confirmons que les erreurs ne sont pas réparties uniformément le long du génome. Nous avons découvert des positions "sujettes à erreurs" dans de nombreux échantillons, indépendemment ou non de la plateforme de séquençage. Exclure ces positions, que nous avons listées pour chacune des plateformes, permet de filtrer davantage d'erreurs. Après application de l'ensemble de ces filtres, il reste au moins $\sim 0.96\%$ d'erreurs dans nos appels de génotype. Ces erreurs sont a priori des erreurs ponctuelles, différentes d'une patiente à l'autre. Notre pipeline appelle donc les génotypes de manière fiable, et la qualité des données produites est suffisante pour étudier par exemple des associations entre variants et maladies.

Nous avons appliqué notre méthode sur des données de séquençage d'exome de cellules saines de 520 patientes atteintes du cancer de l'ovaire, produites par le consortium TCGA. Nous détectons en moyenne 30 632 variants par patiente. Le principal défi est alors d'identifier, parmi l'ensemble des variants germinaux d'une patiente, ceux conférant un risque accru de développer la maladie (variants "de susceptibilité"). Il est probable qu'une majorité des variants de susceptibilité affecte la fonctionnalité des protéines et nous choisissons de nous concentrer sur ceux-ci. Nous estimons donc l'effet des variants sur les gènes connus, en nous basant sur les transcrits de la base de données Aceview (6). Ainsi, nous ignorons les mutations synonymes et les mutations non codantes (à l'exception des pieds d'introns). Quant aux mutations faux sens, certaines peuvent complètement modifier la conformation de la protéine ou les sites de liaisons, empêchant ainsi la protéine d'accomplir son rôle, alors que d'autres peuvent être sans grande conséquence. Pour discriminer ces mutations, nous utilisons les logiciels SIFT (8) et Polyphen-2 (9), qui prédisent l'impact d'une mutation en se basant entre autres sur la conservation des protéines ou les caractéristiques biochimiques des acides aminées. De plus, nous sélectionnons uniquement les variants affectant les gènes bien exprimés dans l'ovaire. Dans un deuxième temps, nous comparons ces SNVs avec deux cohortes contrôles : les données du 1000 Genomes Project et de ESP6500, ce qui nous permet de sélectionner les variants significativement plus présents chez nos patientes que dans la population générale. Cela conduit à une moyenne de 44 SNVs par patiente, répartis sur 334 gènes dans l'ensemble de la cohorte : ces gènes pourraient jouer un rôle dans la cancerogénèse (gènes "candidats"). Afin de valider cette liste de candidats et d'identifier les plus prometteurs, nous les comparons avec des gènes connus pour être associés au cancer. D'une part, nous avons effectué une importante revue de la littérature, sélectionnant ainsi 40 gènes très probablement impliqués dans le cancer de l'ovaire, et d'autre part, nous avons utilisé des outils et bases de données existants pour obtenir des listes élargies de gènes cancer. 42 de nos candidats ont été reportés comme impliqués dans la cancerogénèse, notre liste est donc enrichie en gènes cancer ($P = 9.71 \cdot 10^{-3}$ à $P = 5.60 \cdot 10^{-5}$). Le rôle de certains d'entre eux dans le cancer est bien établi, mais pour d'autres, des études fonctionnelles sont nécessaires afin de confirmer leur implication dans le développement de la maladie. Un intérêt particulier doit être porté à la protéine MAP3K8, dont le rôle de suppresseur de tumeur a été très récemment proposé dans d'autres cancers, et sur laquelle 106 de nos patientes portent une mutation délétère.

De ces analyses, les parties les plus chronophages sont les étapes de récupération des données, de prétraitement, de tri, et de contrôle qualité. Ces étapes sont nécessaires et il est crucial de développer des méthodes expérimentales plus fiables et des outils bioinformatiques adaptés afin de faciliter ce processus.

Par manque de temps, je n'ai pas pu analyser les variants associés au cancer de l'ovaire aussi profondément que je l'aurais souhaité. Par exemple, il aurait été très intéressant de réaliser une étude interactomique de ces variants. Après cartographie des gènes mutés chez nos patientes sur l'interactome humain, des méthodes de clustering pourraient permettre d'identifier des modules fonctionnels impliqués dans le cancer de l'ovaire. De plus, cela améliorerait les filtres des variants germinaux sans effet sur la cancerogénèse (les variants isolés dans le graphe sont probablement sans effet) et supprimerait donc les potentiels faux positifs de nos listes de gènes candidats. En outre, nous avons laissé de côté de nombreux gènes mutés car nous n'avions que peu d'indices de leur rôle dans le développement de la maladie, mais certains d'entre eux sont sans doute des gènes à risque. Le principe de culpabilité par association pourrait être utilisé pour les identifier : ceux interagissant avec de nombreux gènes connus pour être impliqués dans le cancer sont de bons candidats. De surcroît, ce type d'analyse permettrait de prendre en compte la redondance dans les voies de signalisation, en identifiant des voies de signalisation significativement plus mutées chez nos patientes que dans la population générale. Analyser les interactions protéine-protéine serait donc utile pour identifier les variants associés au cancer, et pour comprendre leur effet sur la cancerogénèse.

Malgré les énormes progrès réalisés grâce aux données omiques, le fonctionnement de la Vie est encore loin d'être totalement compris. Par exemple, les interactomes modélisent le fonctionnement de la cellule de manière statique, alors qu'en réalité, les interactions protéine-protéine dépendent de l'environnement de la cellule et du type de tissu, ce qui est rarement considéré dans les analyses (162, 163). Par ailleurs, l'épissage alternatif, mécanisme fondamental permettant aux gènes de coder pour plusieurs protéines (formes alternatives), est très peu pris en compte, que ce soit dans les études interactomiques ou dans les études transcriptomiques. Cela est dû au fait que des informations sur les formes alternatives concernées sont généralement absentes des données expérimentales : pour les données RNA-seq, la taille des short-reads rend difficile l'identification de la forme alternative impliquée et les méthodes de détection d'interactions ne différencient pas les formes alternatives utilisées. La base de données Aceview, que nous avons utilisée pour localiser les variants cancer dans les transcrits, constitue une ressource très complète en ce qui concerne l'épissage. Une autre suite possible de notre travail serait donc d'essayer d'identifier quelle(s) forme(s) alternative(s) des gènes est(sont) impliquée(s) dans le développement du cancer. Une solution serait de construire des interactomes sur les différents transcrits des gènes. Dans les interactomes "classiques", un unique noeud représente l'ensemble des formes alternatives des protéines. Un interactome transcrit-spécifique représenterait chaque forme alternative par un noeud différent, et les interactions entre les formes alternatives pourraient être inférées par exemple en intégrant dans l'analyse des données d'expressions ou de domaines protéiques.

Par ailleurs, l'amélioration des technologies permet de caractériser un grand nombre de tumeurs. Cependant, la plupart des études se concentrent sur les mutations somatiques alors que les variants germinaux, très importants pour améliorer la prévention et le diagnostic, sont très peu analysés. Il s'agit là d'un choix étonnant au niveau santé publique, car ces aspects sont essentiels dans la lutte contre le cancer. Notre analyse a mis en évidence de nouveaux gènes qui pourraient conférer un risque accru de développer un cancer de l'ovaire. Bien que des analyses fonctionnelles soient nécessaires pour confirmer le rôle de ces derniers dans le développement de la maladie, ce sont d'excellents candidats pour une amélioration des diagnostics génétiques. Notre pipeline pourrait facilement être appliqué aux nombreux autres types de cancers séquen-

cés par le consortium TCGA, l'INCA ou l'ICGC. Complété par une analyse interactomique et des études fonctionnelles, cela permettrait d'identifier de nombreux variants germinaux prédisposant à ces cancers. Il serait ensuite intéressant de réaliser une analyse comparative des variants et des gènes identifiés. Par ailleurs, le futur du diagnostic génétique pourraient potentiellement se situer dans l'identification de gènes cancers et la prédiction de l'effet des variants sur ces gènes, plutôt que dans l'identification de tous les variants impliqués. Encore une fois, ce type d'analyse devra prendre en compte l'effet des variants sur les modules fonctionnels et les voies de signalisation. Pour cette tâche, les interactomes seront d'une grande aide.

Bibliographie

[1] McPherson JD : **Next-generation gap**. *Nature methods* 2009, **6**(11 Suppl) :S2–5. 3, 12

[2] Nekrutenko A, Taylor J : **Next-generation sequencing data interpretation : enhancing reproducibility and accessibility**. *Nature Reviews Genetics* 2012, **13**(9) :667–672. 3, 12

[3] Han JDJ, Dupuy D, Bertin N, Cusick ME, Vidal M : **Effect of sampling on topology predictions of protein-protein interaction networks**. *Nat Biotechnol* 2005, **23** :839–844. 3, 20, 21, 25

[4] Collins FS, Barker AD : **Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies**. *Scientific American* 2007, **296**(3) :50–57. 4, 19

[5] The Cancer Genome Atlas : **Integrated genomic analyses of ovarian carcinoma**. *Nature* 2011, **474**(7353) :609–615. 4, 18, 39, 65

[6] Thierry-Mieg D, Thierry-Mieg J : **AceView : a comprehensive cDNA-supported gene and transcripts annotation**. *Genome biology* 2006, **7 Suppl 1** :S12.1–14. 4, 69, 79

[7] Nielsen R, Paul JS, Albrechtsen A, Song YS : **Genotype and SNP calling from next-generation sequencing data**. *Nature reviews. Genetics* 2011, **12**(6) :443–451. 4, 16, 40, 45

[8] Kumar P, Henikoff S, Ng PC : **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm**. *Nature protocols* 2009, **4**(7) :1073–1081. 5, 69, 70, 79

[9] Adzhubei I, Jordan DM, Sunyaev SR : **Predicting functional effect of human missense mutations using PolyPhen-2**. *Current protocols in human genetics* 2013, **Chapter 7** :Unit7.20. 5, 69, 70, 79

[10] Mardis ER : **A decade/'s perspective on DNA sequencing technology**. *Nature* 2011, **470**(7333) :198–203. 11

[11] Metzker ML : **Sequencing technologies - the next generation**. *Nature reviews. Genetics* 2010, **11** :31–46. 11, 12

[12] Hawkins RD, Hon GC, Ren B : **Next-generation genomics : an integrative approach**. *Nature reviews. Genetics* 2010, **11**(7) :476–486. 12, 14

[13] Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F, Fernandez A, Staker B, Pant KP, Baccash J, Borcherding AP, Brownley A, Cedeno R, Chen L, Chernikoff D, Cheung A, Chirita R, Curson B, Ebert JC, Hacker CR, Hartlage R, Hauser B, Huang S, Jiang Y, Karpinchyk V, Koenig M, Kong C, Landers T, Le C, Liu J, McBride CE, Morenzoni M, Morey RE, Mutch K, Perazich H, Perry K, Peters BA, Peterson J, Pethiyagoda CL, Pothuraju K, Richter C, Rosenbaum AM, Roy S, Shafto J, Sharanhovich U, Shannon KW, Sheppy CG, Sun M, Thakuria JV, Tran A, Vu D, Zaranek AW, Wu X, Drmanac S, Oliphant AR, Banyai WC, Martin B, Ballinger DG, Church GM, Reid CA : **Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays**. *Science (New York, N.Y.)* 2010, **327**(5961) :78–81. 12

[14] Fields S, Song O : **A novel genetic system to detect protein-protein interactions**. *Nature* 1989, **340** :245–246. 14

[15] Michnick SW, Remy I, Campbell-Valois FX, Vallee-Belisle A, Pelletier JN : **Detection of protein-protein interactions by protein fragment complementation strategies.** *Methods Enzymol* 2000, **328** :208. 14

[16] Tarassov K, Messier V, Landry CR, Radinovic S, Molina MMS, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW : **An in vivo map of the yeast protein interactome**. *Science* 2008, **320** :1465–1470. 14, 16, 22

[17] Crawford GE, Holt IE, Mullikin JC, Tai D, Green ED, Wolfsberg TG, Collins FS : **Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(4) :992–997. 14

[18] Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kura D, Lam HYK, Leng J, Li R, Li Y, Lin CY, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stutz AM, Urban AE, Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Ye K, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll SA, Korbel JO : **Mapping copy number variation by population scale genome sequencing**. *Nature* 2011, **470**(7332) :59–65. 14

[19] Li H, Homer N : **A survey of sequence alignment algorithms for next-generation sequencing**. *Briefings in bioinformatics* 2010, **11**(5) :473–483. 16

[20] Philippe N, Salson M, Commes T, Rivals E : **CRAC : an integrated approach to the analysis of RNA-seq reads**. *Genome Biology* 2013, **14**(3) :R30. 16

[21] Li H, Durbin R : **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics (Oxford, England)* 2009, **25**(14) :1754–1760. 16, 65

[22] Langmead B, Trapnell C, Pop M, Salzberg SL : **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome biology* 2009, **10**(3) :R25. 16

[23] Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J : **SOAP2 : an improved ultrafast tool for short read alignment**. *Bioinformatics (Oxford, England)* 2009, **25**(15) :1966–1967. 16

[24] Philippe N, Boureux A, Bréhélin L, Tarhio J, Commes T, Rivals Ã : **Using reads to annotate the genome : influence of length, background distribution, and sequence errors on prediction capacity**. *Nucleic Acids Research* 2009, **37**(15) :e104–e104. 16

[25] Barillot E, Calzone L, Hupe P, Vert JP, Zinovyev A : *Computational systems biology of cancer*, *Volume 47*. CRC Press 2012. 17

[26] Thorvaldsdóttir H, Robinson JT, Mesirov JP : **Integrative Genomics Viewer (IGV) : high-performance genomics data visualization and exploration**. *Briefings in bioinformatics* 2013, **14**(2) :178–192. 17, 18

[27] Vidal M, Cusick ME, Barabási AL : **Interactome networks and human disease**. *Cell* 2011, **144**(6) :986–998. 16, 18, 19, 20

[28] Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet AS, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabasi AL, Tavernier J, Hill DE, Vidal M : **High-quality binary protein interaction map of the yeast interactome network**. *Science* 2008, **322** :104–110. 16, 20, 22, 25

[29] Gandhi TKB, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A : **Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets**. *Nature genetics* 2006, **38**(3) :285–293. 16

[30] Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, Prinz J, St Onge RP, VanderSluis B, Makhnevych T, Vizeacoumar FJ, Alizadeh S, Bahr S, Brost RL, Chen Y, Cokol M, Deshpande R, Li Z, Lin ZY, Liang W, Marback M, Paw J, San Luis BJ, Shuteriqi E, Tong AH, van Dyk N, Wallace IM, Whitney JA, Weirauch MT, Zhong G, Zhu H, Houry WA, Brudno M, Ragibizadeh S, Papp B, Pal C, Roth FP, Giaever G, Nislow C, Troyanskaya OG, Bussey H, Bader GD, Gingras AC, Morris QD, Kim PM, Kaiser CA, Myers CL, Andrews BJ, Boone C : **The Genetic Landscape of a Cell**. *Science* 2010, **327** :425. 16

[31] Brun C, Chevenet F, Martin D, Wojcik J, Guénoche A, Jacq B : **Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network**. *Genome biology* 2003, **5** :R6. 16

[32] Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD : **Integration of biological networks and gene expression data using Cytoscape**. *Nature protocols* 2007, **2**(10) :2366–2382. 18

[33] Sanchez C, Lachaize C, Janody F, Bellon B, Röder L, Euzenat J, Rechenmann F, Jacq B : **Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database**. *Nucleic acids research* 1999, **27** :89–94. 18

[34] Orchard S : **Data Standardization and Sharing-The work of the HUPO-PSI**. *Biochimica et biophysica acta* 2013. 18

[35] Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM : **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic Acids Research* 2010, **38**(6) :1767–1771. 18

[36] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R : **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16) :2078–2079. 18, 40, 65

[37] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R : **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**(15) :2156–2158. 18

[38] Berger B, Peng J, Singh M : **Computational solutions for omics data**. *Nature reviews. Genetics* 2013, **14**(5) :333–346. 18

[39] Aranda B, Blankenburg H, Kerrien S, Brinkman FSL, Ceol A, Chautard E, Dana JM, De Las Rivas J, Dumousseau M, Galeota E, Gaulton A, Goll J, Hancock REW, Isserlin R, Jimenez RC, Kerssemakers J, Khadake J, Lynn DJ, Michaut M, O'Kelly G, Ono K, Orchard S, Prieto C, Razick S, Rigina O, Salwinski L, Simonovic M, Velankar S, Winter A, Wu G, Bader GD, Cesareni G, Donaldson IM, Eisenberg D, Kleywegt GJ, Overington J, Ricard-Blum S, Tyers M, Albrecht M, Hermjakob H : **PSICQUIC and PSISCORE : accessing and scoring molecular interactions**. *Nature methods* 2011, **8**(7) :528–529. 18

[40] del Toro N, Dumousseau M, Orchard S, Jimenez RC, Galeota E, Launay G, Goll J, Breuer K, Ono K, Salwinski L, Hermjakob H : **A new reference implementation of the PSICQUIC web service**. *Nucleic Acids Research* 2013, **41**(W1) :W601–W606. 18

[41] Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T : **A travel guide to Cytoscape plugins**. *Nature methods* 2012, **9**(11) :1069–1076. 18

[42] Goldman M, Craft B, Swatloski T, Ellrott K, Cline M, Diekhans M, Ma S, Wilks C, Stuart J, Haussler D, Zhu J : **The UCSC Cancer Genomics Browser : update 2013**. *Nucleic Acids Research* 2012, **41**(D1) :D949–D954. 19

[43] Kuhn RM, Haussler D, Kent WJ : **The UCSC genome browser and associated tools**. *Briefings in Bioinformatics* 2013, **14**(2) :144–161. 19

[44] The International HapMap Consortium : **A second generation human haplotype map of over 3.1 million SNPs**. *Nature* 2007, **449**(7164) :851–861. 19

[45] The International HapMap 3 Consortium : **Integrating common and rare genetic variation in diverse human populations**. *Nature* 2010, **467**(7311) :52–58. 19

[46] The 1000 Genomes Project Consortium : **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319) :1061–1073. 19, 70

[47] **Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA** [http://evs.gs.washington.edu/EVS/]. 19, 70

[48] Lin WY, Brock IW, Connley D, Cramp H, Tucker R, Slate J, Reed MWR, Balasubramanian SP, Cannon-Albright LA, Camp NJ, Cox A : **Associations of ATR and CHEK1 Single Nucleotide Polymorphisms with Breast Cancer**. *PloS one* 2013, **8**(7) :e68578. 19

[49] Shah S, Kim Y, Ostrovnaya I, Murali R, Schrader KA, Lach FP, Sarrel K, Rau-Murthy R, Hansen N, Zhang L, Kirchhoff T, Stadler Z, Robson M, Vijai J, Offit K, Smogorzewska A : **Assessment of SLX4 Mutations in Hereditary Breast Cancers**. *PLoS ONE* 2013, **8**(6) :e66961. 19

[50] The ENCODE Project Consortium : **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414) :57–74. 19

[51] Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E : **On the immortality of television sets : "function" in the human genome according to the evolution-free gospel of ENCODE**. *Genome biology and evolution* 2013, **5**(3) :578–590. 19

[52] International Cancer Genome Consortium : **International network of cancer genome projects**. *Nature* 2010, **464**(7291) :993–998. 19

[53] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz J Luis A, Kinzler KW : **Cancer genome landscapes**. *Science (New York, N.Y.)* 2013, **339**(6127) :1546–1558. 19, 64, 74

[54] Hanahan D, Weinberg RA : **Hallmarks of cancer : the next generation**. *Cell* 2011, **144**(5) :646–674. 19, 63, 64

[55] Vinayagam A, Stelzl U, Foulle R, Plassmann S, Zenkner M, Timm J, Assmus HE, Andrade-Navarro MA, Wanker EE : **A directed protein interaction network for investigating intracellular signal transduction**. *Science signaling* 2011, **4**(189) :rs8. 19

[56] Hegele A, Kamburov A, Grossmann A, Sourlis C, Wowro S, Weimann M, Will C, Pena V, Lührmann R, Stelzl U : **Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome**. *Molecular Cell* 2012, **45**(4) :567–580. 19

[57] Weimann M, Grossmann A, Woodsmith J, ÃŰzkan Z, Birth P, Meierhofer D, Benlasfer N, Valovka T, Timmermann B, Wanker EE, Sauer S, Stelzl U : **A Y2H-seq approach defines the human protein methyltransferase interactome**. *Nature methods* 2013, **10**(4) :339–342. 19

[58] Jeong H, Mason SP, Barabasi AL, Oltvai ZN : **Lethality and centrality in protein networks**. *Nature* 2001, **411** :41–42. 20, 25

[59] Fraser HB, Plotkin JB : **Using protein complexes to predict phenotypic effects of gene mutation**. *Genome Biology* 2007, **8**(11) :R252. 20

[60] Annibale A, Coolen ACC : **What you see is not what you get : how sampling affects macroscopic features of biological networks**. *Interface focus* 2011, **1**(6) :836–856. 20

[61] Meacham F, Boffelli D, Dhahbi J, Martin DIK, Singer M, Pachter L : **Identification and correction of systematic error in high-throughput sequence data**. *BMC bioinformatics* 2011, **12** :451. 20, 45

[62] Musumeci L, Arthur JW, Cheung FS, Hoque A, Lippman S, Reichardt JK : **Single Nucleotide Differences (SNDs) in the dbSNP Database May Lead to Errors in Genotyping and Haplotyping Studies**. *Human mutation* 2010, **31** :67–73. 20, 77

[63] MAQC Consortium : **The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements**. *Nature biotechnology* 2006, **24**(9) :1151–1161. 20

[64] MAQC Consortium : **The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.** *Nature biotechnology* 2010, **28**(8) :827–838. 20

[65] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M : **BioGRID : a general repository for interaction datasets**. *Nucleic Acids Res* 2006, **34** :D535. 22

[66] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y : **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *Proc Natl Acad Sci U S A* 2001, **98** :4569–4574. 22, 25

[67] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields

S, Rothberg JM : **A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403** :623–627. 22, 25

[68] Sprinzak E, Sattath S, Margalit H : **How Reliable are Experimental Protein–Protein Interaction Data ?** *J Mol Biol* 2003, **327** :919–923. 25

[69] Grigoriev A : **On the number of protein-protein interactions in the yeast proteome**. *Nucleic Acids Res* 2003, **31** :4157–4161. 25

[70] D'haeseleer P, Church GM : **Estimating and improving protein interaction error rates**. In *Proc IEEE Comput Syst Bioinform Conf* 2004 :216–23. 25

[71] Hart GT, Ramani AK, Marcotte EM : **How complete are current yeast and human protein-interaction networks ?** *Genome Biol* 2006, **7** :120. 25

[72] Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C : **Estimating the size of the human interactome**. *Proc Natl Acad Sci U S A* 2008, **105** :6959–6964. 25

[73] Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D, Andrews B, Boone C, Troyanskya OG, Ideker T, Dolinski K, Batada NN, Tyers M : **Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae**. *J Biol* 2006, **5** :11. 25

[74] Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortés ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Lee RS, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CWM, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G : **Mutational heterogeneity in cancer and the search for new cancer-associated genes**. *Nature* 2013, **499**(7457) :214–218. 39, 65

[75] Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, Ding L : **MuSiC : identifying mutational significance in cancer genomes**. *Genome research* 2012, **22**(8) :1589–1598. 39, 65

[76] Hodis E, Watson I, Kryukov G, Arold S, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C, DiCara D, Ramos A, Lawrence M, Cibulskis K, Sivachenko A, Voet D, Saksena G, Stransky N, Onofrio R, Winckler W, Ardlie K, Wagle N, Wargo J, Chong K, Morton D, Stemke-Hale K, Chen G, Noble M, Meyerson M, Ladbury J, Davies M, Gershenwald J, Wagner S, Hoon D, Schadendorf D, Lander E, Gabriel S, Getz G, Garraway L, Chin L : **A Landscape of Driver Mutations in Melanoma**. *Cell* 2012, **150**(2) :251–263. 39

[77] Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, McLendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz LA, Hartigan J, Smith DR, Strausberg RL, Marie SKN, Shinjo SMO, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW : **An Integrated Genomic Analysis of Human Glioblastoma Multiforme**. *Science* 2008, **321**(5897) :1807–1812. 39

[78] Kilpivaara O, Aaltonen LA : **Diagnostic cancer genome sequencing and the contribution of germline variants**. *Science (New York, N.Y.)* 2013, **339**(6127) :1559–1562. 39, 65

[79] Cheng EH, Sawyers CL : **In cancer drug resistance, germline matters too**. *Nature medicine* 2012, **18**(4) :494–496. 39

[80] Hedges DJ, Hedges D, Burges D, Powell E, Almonte C, Huang J, Young S, Boese B, Schmidt M, Pericak-Vance MA, Martin E, Zhang X, Harkins TT, Züchner S : **Exome sequencing of a multigenerational human pedigree**. *PloS one* 2009, **4**(12) :e8232. 40

[81] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel

S, Daly M, DePristo MA : **The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Research* 2010, **20**(9) :1297–1303. 40, 45

[82] Martin ER, Kinnamon DD, Schmidt MA, Powell EH, Zuchner S, Morris RW : **SeqEM : an adaptive genotype-calling approach for next-generation sequencing studies**. *Bioinformatics (Oxford, England)* 2010, **26**(22) :2803–2810. 40

[83] O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, Wei Z, Wang K, Lyon GJ : **Low concordance of multiple variant-calling pipelines : practical implications for exome and genome sequencing**. *Genome Medicine* 2013, **5**(3) :28. 40, 77

[84] Altmann A, Weber P, Bader D, Preuss M, Binder EB, Müller-Myhsok B : **A beginners guide to SNP calling from high-throughput DNA-sequencing data**. *Human genetics* 2012, **131**(10) :1541–1554. 40, 77

[85] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR : **STAR : ultrafast universal RNA-seq aligner**. *Bioinformatics (Oxford, England)* 2013, **29** :15–21. 41

[86] Liao Y, Smyth GK, Shi W : **The Subread aligner : fast, accurate and scalable read mapping by seed-and-vote**. *Nucleic acids research* 2013, **41**(10) :e108. 41

[87] Frank SA : *Dynamics of cancer : incidence, inheritance, and evolution*. Princeton University Press 2007. 64

[88] Stratton MR : **Exploring the genomes of cancer cells : progress and promise**. *Science (New York, N.Y.)* 2011, **331**(6024) :1553–1558. 64

[89] Li H, Ruan J, Durbin R : **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome research* 2008, **18**(11) :1851–1858. 65

[90] Homer N, Merriman B, Nelson SF : **BFAST : An Alignment Tool for Large Scale Genome Resequencing**. *PLoS ONE* 2009, **4**(11) :e7767. 65

[91] Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L : **SomaticSniper : identification of somatic point mutations in whole genome sequencing data**. *Bioinformatics (Oxford, England)* 2012, **28**(3) :311–317. 65

[92] Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G : **Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples**. *Nature biotechnology* 2013, **31**(3) :213–219. 65

[93] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK : **VarScan 2 : somatic mutation and copy number alteration discovery in cancer by exome sequencing**. *Genome research* 2012, **22**(3) :568–576. 65

[94] Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R : **Cancer-specific high-throughput annotation of somatic mutations : computational prediction of driver missense mutations**. *Cancer research* 2009, **69**(16) :6660–6667. 65

[95] Azvolinsky A : **Diagnostic lens turns to difficult-to-detect ovarian cancer**. *Nature medicine* 2013, **19**(2) :117. 65

[96] Zhang S, Royer R, Li S, McLaughlin JR, Rosen B, Risch HA, Fan I, Bradley L, Shaw PA, Narod SA : **Frequencies of BRCA1 and BRCA2 mutations among 1,342 unselected patients with invasive ovarian cancer**. *Gynecologic oncology* 2011, **121**(2) :353–357. 66, 74

[97] Pal T, Permuth-Wey J, Betts JA, Krischer JP, Fiorica J, Arango H, LaPolla J, Hoffman M, Martino MA, Wakeley K, Wilbanks G, Nicosia S, Cantor A, Sutphen R : **BRCA1 and BRCA2 mutations account for a large proportion of ovarian carcinoma cases**. *Cancer* 2005, **104**(12) :2807–2816. 66, 74

[98] Ramus SJ, Gayther SA : **The contribution of BRCA1 and BRCA2 to ovarian cancer**. *Molecular oncology* 2009, **3**(2) :138–150. 66, 74

[99] Mavaddat N, Peock S, Frost D, Ellis S, Platte R, Fineberg E, Evans DG, Izatt L, Eeles RA, Adlard J, Davidson R, Eccles D, Cole T, Cook J, Brewer C, Tischkowitz M, Douglas F, Hodgson S, Walker L, Porteous ME, Morrison PJ, Side LE, Kennedy MJ, Houghton C, Donaldson A, Rogers MT, Dorkins H, Miedzybrodzka Z, Gregory H,

Eason J, Barwell J, McCann E, Murray A, Antoniou AC, Easton DF, EMBRACE : **Cancer risks for BRCA1 and BRCA2 mutation carriers : results from prospective analysis of EMBRACE**. *Journal of the National Cancer Institute* 2013, **105**(11) :812–822. 66, 74

[100] Thai TH, Du F, Tsan JT, Jin Y, Phung A, Spillman MA, Massa HF, Muller CY, Ashfaq R, Mathis JM, Miller DS, Trask BJ, Baer R, Bowcock AM : **Mutations in the BRCA1-associated RING domain (BARD1) gene in primary breast, ovarian and uterine cancers**. *Human molecular genetics* 1998, **7**(2) :195–202. 66

[101] Pennington KP, Swisher EM : **Hereditary ovarian cancer : beyond the usual suspects**. *Gynecologic oncology* 2012, **124**(2) :347–353. 66, 74

[102] Malander S, Rambech E, Kristoffersson U, Halvarsson B, Ridderheim M, Borg A, Nilbert M : **The contribution of the hereditary nonpolyposis colorectal cancer syndrome to the development of ovarian cancer**. *Gynecologic oncology* 2006, **101**(2) :238–243. 66

[103] Rafnar T, Gudbjartsson DF, Sulem P, Jonasdottir A, Sigurdsson A, Jonasdottir A, Besenbacher S, Lundin P, Stacey SN, Gudmundsson J, Magnusson OT, le Roux L, Orlygsdottir G, Helgadottir HT, Johannsdottir H, Gylfason A, Tryggvadottir L, Jonasson JG, de Juan A, Ortega E, Ramon-Cajal JM, García-Prats MD, Mayordomo C, Panadero A, Rivera F, Aben KKH, van Altena AM, Massuger LFAG, Aavikko M, Kujala PM, Staff S, Aaltonen LA, Olafsdottir K, Bjornsson J, Kong A, Salvarsdottir A, Saemundsson H, Olafsson K, Benediktsdottir KR, Gulcher J, Masson G, Kiemeney LA, Mayordomo JI, Thorsteinsdottir U, Stefansson K : **Mutations in BRIP1 confer high risk of ovarian cancer**. *Nature genetics* 2011, **43**(11) :1104–1107. 66, 74

[104] Loveday C, Turnbull C, Ramsay E, Hughes D, Ruark E, Frankum JR, Bowden G, Kalmyrzaev B, Warren-Perry M, Snape K, Adlard JW, Barwell J, Berg J, Brady AF, Brewer C, Brice G, Chapman C, Cook J, Davidson R, Donaldson A, Douglas F, Greenhalgh L, Henderson A, Izatt L, Kumar A, Lalloo F, Miedzybrodzka Z, Morrison PJ, Paterson J, Porteous M, Rogers MT, Shanley S, Walker L, Breast Cancer Susceptibility Collaboration (UK), Eccles D, Evans DG, Renwick A, Seal S, Lord CJ, Ashworth A, Reis-Filho JS, Antoniou AC, Rahman N : **Germline mutations in RAD51D confer susceptibility to ovarian cancer**. *Nature genetics* 2011, **43**(9) :879–882. 66, 74

[105] Meindl A, Hellebrand H, Wiek C, Erven V, Wappenschmidt B, Niederacher D, Freund M, Lichtner P, Hartmann L, Schaal H, Ramser J, Honisch E, Kubisch C, Wichmann HE, Kast K, Deissler H, Engel C, Müller-Myhsok B, Neveling K, Kiechle M, Mathew CG, Schindler D, Schmutzler RK, Hanenberg H : **Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene**. *Nature genetics* 2010, **42**(5) :410–414. 66, 74

[106] Loveday C, Turnbull C, Ruark E, Xicola RMM, Ramsay E, Hughes D, Warren-Perry M, Snape K, Breast Cancer Susceptibility Collaboration (UK), Eccles D, Evans DG, Gore M, Renwick A, Seal S, Antoniou AC, Rahman N : **Germline RAD51C mutations confer susceptibility to ovarian cancer**. *Nature genetics* 2012, **44**(5) :475–476 ; author reply 476. 66, 74

[107] Walsh T, Casadei S, Lee MK, Pennil CC, Nord AS, Thornton AM, Roeb W, Agnew KJ, Stray SM, Wickramanayake A, Norquist B, Pennington KP, Garcia RL, King MC, Swisher EM : **Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing**. *Proceedings of the National Academy of Sciences* 2011, **108**(44) :18032–18037. 66, 74

[108] Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, Carter SL, Cibulskis K, Hanna M, Kiezun A, Kim J, Lawrence MS, Lichenstein L, McKenna A, Pedamallu CS, Ramos AH, Shefler E, Sivachenko A, Sougnez C, Stewart C, Ally A, Birol I, Chiu R, Corbett RD, Hirst M, Jackman SD, Kamoh B, Khodabakshi AH, Krzywinski M, Lo A, Moore RA, Mungall KL, Qian J, Tam A, Thiessen N, Zhao Y, Cole KA, Diamond M, Diskin SJ, Mosse YP, Wood AC, Ji L, Sposto R, Badgett T, London WB, Moyer Y, Gastier-Foster JM, Smith MA, Auvil JMG, Gerhard DS, Hogarty MD, Jones SJM, Lander ES, Gabriel SB, Getz G, Seeger RC, Khan J, Marra MA, Meyerson M, Maris JM : **The genetic landscape of high-risk neuroblastoma**. *Nature genetics* 2013, **45**(3) :279–284. 66

[109] Ruark E, Snape K, Humburg P, Loveday C, Bajrami I, Brough R, Rodrigues DN, Renwick A, Seal S, Ramsay E, Duarte SDV, Rivas MA, Warren-Perry M, Zachariou A, Campion-Flora A, Hanks S, Murray A, Ansari Pour

N, Douglas J, Gregory L, Rimmer A, Walker NM, Yang TP, Adlard JW, Barwell J, Berg J, Brady AF, Brewer C, Brice G, Chapman C, Cook J, Davidson R, Donaldson A, Douglas F, Eccles D, Evans DG, Greenhalgh L, Henderson A, Izatt L, Kumar A, Lalloo F, Miedzybrodzka Z, Morrison PJ, Paterson J, Porteous M, Rogers MT, Shanley S, Walker L, Gore M, Houlston R, Brown MA, Caufield MJ, Deloukas P, McCarthy MI, Todd JA, Breast and Ovarian Cancer Susceptibility Collaboration, Wellcome Trust Case Control Consortium, Turnbull C, Reis-Filho JS, Ashworth A, Antoniou AC, Lord CJ, Donnelly P, Rahman N : **Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer**. *Nature* 2013, **493**(7432) :406–410. 66, 74

[110] Song H, Ramus SJ, Tyrer J, Bolton KL, Gentry-Maharaj A, Wozniak E, Anton-Culver H, Chang-Claude J, Cramer DW, DiCioccio R, Dörk T, Goode EL, Goodman MT, Schildkraut JM, Sellers T, Baglietto L, Beckmann MW, Beesley J, Blaakaer J, Carney ME, Chanock S, Chen Z, Cunningham JM, Dicks E, Doherty JA, Dürst M, Ekici AB, Fenstermacher D, Fridley BL, Giles G, Gore ME, De Vivo I, Hillemanns P, Hogdall C, Hogdall E, Iversen ES, Jacobs IJ, Jakubowska A, Li D, Lissowska J, Lubiński J, Lurie G, McGuire V, McLaughlin J, MÄŹdrek K, Moorman PG, Moysich K, Narod S, Phelan C, Pye C, Risch H, Runnebaum IB, Severi G, Southey M, Stram DO, Thiel FC, Terry KL, Tsai YY, Tworoger SS, Van Den Berg DJ, Vierkant RA, Wang-Gohrke S, Webb PM, Wilkens LR, Wu AH, Yang H, Brewster W, Ziogas A, Houlston R, Tomlinson I, Whittemore AS, Rossing MA, Ponder BAJ, Pearce CL, Ness RB, Menon U, Kjaer SK, Gronwald J, Garcia-Closas M, Fasching PA, Easton DF, Chenevix-Trench G, Berchuck A, Pharoah PDP, Gayther SA : **A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2**. *Nature Genetics* 2009, **41**(9) :996–1000. 66, 67

[111] Goode EL, Chenevix-Trench G, Song H, Ramus SJ, Notaridou M, Lawrenson K, Widschwendter M, Vierkant RA, Larson MC, Kjaer SK, Birrer MJ, Berchuck A, Schildkraut J, Tomlinson I, Kiemeney LA, Cook LS, Gronwald J, Garcia-Closas M, Gore ME, Campbell I, Whittemore AS, Sutphen R, Phelan C, Anton-Culver H, Pearce CL, Lambrechts D, Rossing MA, Chang-Claude J, Moysich KB, Goodman MT, Dörk T, Nevanlinna H, Ness RB, Rafnar T, Hogdall C, Hogdall E, Fridley BL, Cunningham JM, Sieh W, McGuire V, Godwin AK, Cramer DW, Hernandez D, Levine D, Lu K, Iversen ES, Palmieri RT, Houlston R, van Altena AM, Aben KKH, Massuger LFAG, Brooks-Wilson A, Kelemen LE, Le ND, Jakubowska A, Lubinski J, Medrek K, Stafford A, Easton DF, Tyrer J, Bolton KL, Harrington P, Eccles D, Chen A, Molina AN, Davila BN, Arango H, Tsai YY, Chen Z, Risch HA, McLaughlin J, Narod SA, Ziogas A, Brewster W, Gentry-Maharaj A, Menon U, Wu AH, Stram DO, Pike MC, The Wellcome Trust Case-Control Consortium, Beesley J, Webb PM, Cancer) TACSO, Group TAOCS, (ocac) tOCAC : **A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24**. *Nature Genetics* 2010, **42**(10) :874–879. 66, 67

[112] Bolton KL, Tyrer J, Song H, Ramus SJ, Notaridou M, Jones C, Sher T, Gentry-Maharaj A, Wozniak E, Tsai YY, Weidhaas J, Paik D, Van Den Berg DJ, Stram DO, Pearce CL, Wu AH, Brewster W, Anton-Culver H, Ziogas A, Narod SA, Levine DA, Kaye SB, Brown R, Paul J, Flanagan J, Sieh W, McGuire V, Whittemore AS, Campbell I, Gore ME, Lissowska J, Yang HP, Medrek K, Gronwald J, Lubinski J, Jakubowska A, Le ND, Cook LS, Kelemen LE, Brook-Wilson A, Massuger LFAG, Kiemeney LA, Aben KKH, van Altena AM, Houlston R, Tomlinson I, Palmieri RT, Moorman PG, Schildkraut J, Iversen ES, Phelan C, Vierkant RA, Cunningham JM, Goode EL, Fridley BL, Kruger-Kjaer S, Blaeker J, Hogdall E, Hogdall C, Gross J, Karlan BY, Ness RB, Edwards RP, Odunsi K, Moyisch KB, Baker JA, Modugno F, Heikkinenen T, Butzow R, Nevanlinna H, Leminen A, Bogdanova N, Antonenkova N, Doerk T, Hillemanns P, Dürst M, Runnebaum I, Thompson PJ, Carney ME, Goodman MT, Lurie G, Wang-Gohrke S, Hein R, Chang-Claude J, Rossing MA, Cushing-Haugen KL, Doherty J, Chen C, Rafnar T, Besenbacher S, Sulem P, Stefansson K, Birrer MJ, Terry KL, Hernandez D, Cramer DW, Vergote I, Amant F, Lambrechts D, Despierre E, Fasching PA, Beckmann MW, Thiel FC, Ekici AB, Chen X, Group tAOCS, Cancer) tACSO, Consortium obotOCA, Johnatty SE, Webb PM, Beesley J, Chanock S, Garcia-Closas M, Sellers T, Easton DF, Berchuck A, Chenevix-Trench G, Pharoah PDP, Gayther SA : **Common variants at 19p13 are associated with susceptibility to ovarian cancer**. *Nature Genetics* 2010, **42**(10) :880–884. 66, 67

[113] Sakoda LC, Jorgenson E, Witte JS : **Turning of COGS moves forward findings for hormonally mediated cancers**. *Nature Genetics* 2013, **45**(4) :345–348. 66, 70

[114] Pharoah PDP, Tsai YY, Ramus SJ, Phelan CM, Goode EL, Lawrenson K, Buckley M, Fridley BL, Tyrer JP, Shen

H, Weber R, Karevan R, Larson MC, Song H, Tessier DC, Bacot F, Vincent D, Cunningham JM, Dennis J, Dicks E, Study AC, Group AOCS, Aben KK, Anton-Culver H, Antonenkova N, Armasu SM, Baglietto L, Bandera EV, Beckmann MW, Birrer MJ, Bloom G, Bogdanova N, Brenton JD, Brinton LA, Brooks-Wilson A, Brown R, Butzow R, Campbell I, Carney ME, Carvalho RS, Jenny Chang-Claude, Chen YA, Zhihua Chen, Chow WH, Cicek MS, Coetzee G, Cook LS, Cramer DW, Cybulski C, Dansonka-Mieszkowska A, Despierre E, Doherty JA, Dörk T, du Bois A, Dürst M, Eccles D, Edwards R, Ekici AB, Fasching PA, Fenstermacher D, Flanagan J, Gao YT, Garcia-Closas M, Gentry-Maharaj A, Giles G, Gjyshi A, Gore M, Gronwald J, Guo Q, Halle MK, Harter P, Hein A, Heitz F, Hillemanns P, Hoatlin M, Hogdall E, Hogdall CK, Hosono S, Jakubowska A, Jensen A, Kalli KR, Karlan BY, Kelemen LE, Kiemeney LA, Kjaer SK, Konecny GE, Krakstad C, Kupryjanczyk J, Lambrechts D, Lambrechts S, Le ND, Lee N, Lee J, Leminen A, Lim BK, Lissowska J, Lubiński J, Lundvall L, Lurie G, Massuger LFAG, Matsuo K, McGuire V, McLaughlin JR, Menon U, Modugno F, Moysich KB, Nakanishi T, Narod SA, Ness RB, Nevanlinna H, Nickels S, Noushmehr H, Odunsi K, Olson S, Orlow I, Paul J, Pejovic T, Pelttari LM, Permuth-Wey J, Pike MC, Poole EM, Qu X, Risch HA, Rodriguez-Rodriguez L, Rossing MA, Rudolph A, Runnebaum I, Rzepecka IK, Salvesen HB, Schwaab I, Severi G, Shen H, Shridhar V, Shu XO, Sieh W, Southey MC, Spellman P, Tajima K, Teo SH, Terry KL, Thompson PJ, Timorek A, Tworoger SS, van Altena AM, van den Berg D, Vergote I, Vierkant RA, Vitonis AF, Wang-Gohrke S, Wentzensen N, Whittemore AS, Wik E, Winterhoff B, Woo YL, Wu AH, Yang HP, Zheng W, Ziogas A, Zulkifli F, Goodman MT, Hall P, Easton DF, Pearce CL, Berchuck A, Chenevix-Trench G, Iversen E, Monteiro ANA, Gayther SA, Schildkraut JM, Sellers TA : **GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer**. *Nature Genetics* 2013, **45**(4) :362–370. 67

[115] Couch FJ, Gaudet MM, Antoniou AC, Ramus SJ, Kuchenbaecker KB, Soucy P, Beesley J, Chen X, Wang X, Kirchhoff T, McGuffog L, Barrowdale D, Lee A, Healey S, Sinilnikova OM, Andrulis IL, OCGN, Ozcelik H, Mulligan AM, Thomassen M, Gerdes AM, Jensen UB, Skytte AB, Kruse TA, Caligo MA, von Wachenfeldt A, Barbany-Bustinza G, Loman N, Soller M, Ehrencrona H, Karlsson P, SWE-BRCA, Nathanson KL, Rebbeck TR, Domchek SM, Jakubowska A, Lubinski J, Jaworska K, Durda K, Zlowocka E, Huzarski T, Byrski T, Gronwald J, Cybulski C, Górski B, Osorio A, Durán M, Tejada MI, Benitez J, Hamann U, Hogervorst FBL, HEBON, van Os TA, van Leeuwen FE, Meijers-Heijboer HEJ, Wijnen J, Blok MJ, Kets M, Hooning MJ, Oldenburg RA, Ausems MGEM, Peock S, Frost D, Ellis SD, Platte R, Fineberg E, Evans DG, Jacobs C, Eeles RA, Adlard J, Davidson R, Eccles DM, Cole T, Cook J, Paterson J, Brewer C, Douglas F, Hodgson SV, Morrison PJ, Walker L, Porteous ME, Kennedy MJ, Side LE, EMBRACE, Bove B, Godwin AK, Stoppa-Lyonnet D, GEMO Study Collaborators, Fassy-Colcombet M, Castera L, Cornelis F, Mazoyer S, Léoné M, Boutry-Kryza N, Bressac-de Paillerets B, Caron O, Pujol P, Coupier I, Delnatte C, Akloul L, Lynch HT, Snyder CL, Buys SS, Daly MB, Terry M, Chung WK, John EM, Miron A, Southey MC, Hopper JL, Goldgar DE, Singer CF, Rappaport C, Tea MKM, Fink-Retter A, Hansen TVO, Nielsen FC, Arason A, Vijai J, Shah S, Sarrel K, Robson ME, Piedmonte M, Phillips K, Basil J, Rubinstein WS, Boggess J, Wakeley K, Ewart-Toland A, Montagna M, Agata S, Imyanitov EN, Isaacs C, Janavicius R, Lazaro C, Blanco I, Feliubadalo L, Brunet J, Gayther SA, Pharoah PPD, Odunsi KO, Karlan BY, Walsh CS, Olah E, Teo SH, Ganz PA, Beattie MS, van Rensburg EJ, Dorfling CM, Diez O, Kwong A, Schmutzler RK, Wappenschmidt B, Engel C, Meindl A, Ditsch N, Arnold N, Heidemann S, Niederacher D, Preisler-Adams S, Gadzicki D, Varon-Mateeva R, Deissler H, Gehrig A, Sutter C, Kast K, Fiebig B, Heinritz W, Caldes T, de la Hoya M, Muranen TA, Nevanlinna H, Tischkowitz MD, Spurdle AB, Neuhausen SL, Ding YC, Lindor NM, Fredericksen Z, Pankratz VS, Peterlongo P, Manoukian S, Peissel B, Zaffaroni D, Barile M, Bernard L, Viel A, Giannini G, Varesco L, Radice P, Greene MH, Mai PL, Easton DF, Chenevix-Trench G, kConFab investigators, Offit K, Simard J, Consortium of Investigators of Modifiers of BRCA1/2 : **Common variants at the 19p13.1 and ZNF365 loci are associated with ER subtypes of breast cancer and ovarian cancer risk in BRCA1 and BRCA2 mutation carriers**. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2012, **21**(4) :645–657. 67, 74

[116] Bojesen SE, Pooley KA, Johnatty SE, Beesley J, Michailidou K, Tyrer JP, Edwards SL, Pickett HA, Shen HC, Smart CE, Hillman KM, Mai PL, Lawrenson K, Stutz MD, Lu Y, Karevan R, Woods N, Johnston RL, French JD,

Chen X, Weischer M, Nielsen SF, Maranian MJ, Ghoussaini M, Ahmed S, Baynes C, Bolla MK, Wang Q, Dennis J, McGuffog L, Barrowdale D, Lee A, Healey S, Lush M, Tessier DC, Vincent D, Bacot F, Australian Cancer Study, Australian Ovarian Cancer Study, Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer (kConFab), Gene Environment Interaction and Breast Cancer (GENICA), Swedish Breast Cancer Study (SWE-BRCA), Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON), Epidemiological study of BRCA1 & BRCA2 Mutation Carriers (EMBRACE), Genetic Modifiers of Cancer Risk in BRCA1/2 Mutation Carriers (GEMO), Vergote I, Lambrechts S, Despierre E, Risch HA, González-Neira A, Rossing MA, Pita G, Doherty JA, Alvarez N, Larson MC, Fridley BL, Schoof N, Chang-Claude J, Cicek MS, Peto J, Kalli KR, Broeks A, Armasu SM, Schmidt MK, Braaf LM, Winterhoff B, Nevanlinna H, Konecny GE, Lambrechts D, Rogmann L, Guénel P, Teoman A, Milne RL, Garcia JJ, Cox A, Shridhar V, Burwinkel B, Marme F, Hein R, Sawyer EJ, Haiman CA, Wang-Gohrke S, Andrulis IL, Moysich KB, Hopper JL, Odunsi K, Lindblom A, Giles GG, Brenner H, Simard J, Lurie G, Fasching PA, Carney ME, Radice P, Wilkens LR, Swerdlow A, Goodman MT, Brauch H, Garcia-Closas M, Hillemanns P, Winqvist R, Dürst M, Devilee P, Runnebaum I, Jakubowska A, Lubinski J, Mannermaa A, Butzow R, Bogdanova NV, Dörk T, Pelttari LM, Zheng W, Leminen A, Anton-Culver H, Bunker CH, Kristensen V, Ness RB, Muir, et al. : **Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer**. *Nature genetics* 2013, **45**(4) :371–384, 384e1–2. 67, 74

[117] Shen H, Fridley BL, Song H, Lawrenson K, Cunningham JM, Ramus SJ, Cicek MS, Tyrer J, Stram D, Larson MC, Köbel M, Practical Consortium, Ziogas A, Zheng W, Yang HP, Wu AH, Wozniak EL, Ling Woo Y, Winterhoff B, Wik E, Whittemore AS, Wentzensen N, Palmieri Weber R, Vitonis AF, Vincent D, Vierkant RA, Vergote I, Van Den Berg D, Van Altena AM, Tworoger SS, Thompson PJ, Tessier DC, Terry KL, Teo SH, Templeman C, Stram DO, Southey MC, Sieh W, Siddiqui N, Shvetsov YB, Shu XO, Shridhar V, Wang-Gohrke S, Severi G, Schwaab I, Salvesen HB, Rzepecka IK, Runnebaum IB, Anne Rossing M, Rodriguez-Rodriguez L, Risch HA, Renner SP, Poole EM, Pike MC, Phelan CM, Pelttari LM, Pejovic T, Paul J, Orlow I, Zawiah Omar S, Olson SH, Odunsi K, Nickels S, Nevanlinna H, Ness RB, Narod SA, Nakanishi T, Moysich KB, Monteiro ANA, Moes-Sosnowska J, Modugno F, Menon U, McLaughlin JR, McGuire V, Matsuo K, Mat Adenan NA, Massuger LFAG, Lurie G, Lundvall L, Lubiński J, Lissowska J, Levine DA, Leminen A, Lee AW, Le ND, Lambrechts S, Lambrechts D, Kupryjanczyk J, Krakstad C, Konecny GE, Krüger Kjaer S, Kiemeney LA, Kelemen LE, Keeney GL, Karlan BY, Karevan R, Kalli KR, Kajiyama H, Ji BT, Jensen A, Jakubowska A, Iversen E, Hosono S, Hogdall CK, Hogdall E, Hoatlin M, Hillemanns P, Heitz F, Hein R, Harter P, Halle MK, Hall P, Gronwald J, Gore M, Goodman MT, Giles GG, Gentry-Maharaj A, Garcia-Closas M, Flanagan JM, Fasching PA, Ekici AB, Edwards R, Eccles D, Easton DF, Dürst M, du Bois A, Dörk T, Doherty JA, Despierre E, Dansonka-Mieszkowska A, Cybulski C, Cramer DW, Cook LS, Chen X, Charbonneau B, Chang-Claude J, Campbell I, Butzow R, Bunker CH, Brueggmann D, Brown R, Brooks-Wilson A, Brinton LA, Bogdanova N, Block MS, Benjamin E, Beesley J, Beckmann MW, Bandera EV, Baglietto L, Bacot F, Armasu SM, Antonenkova N, Anton-Culver H, Aben KK, Liang D, Wu X, Lu K, Hildebrandt MAT, Group AOCS, Study AC, Schildkraut JM, Sellers TA, Huntsman D, Berchuck A, Chenevix-Trench G, Gayther SA, Pharoah PDP, Laird PW, Goode EL, Pearce CL : **Epigenetic analysis leads to identification of HNF1B as a subtype-specific susceptibility gene for ovarian cancer**. *Nature Communications* 2013, **4** :1628. 67

[118] Permuth-Wey J, Lawrenson K, Shen HC, Velkova A, Tyrer JP, Chen Z, Lin HY, Ann Chen Y, Tsai YY, Qu X, Ramus SJ, Karevan R, Lee J, Lee N, Larson MC, Aben KK, Anton-Culver H, Antonenkova N, Antoniou AC, Armasu SM, Study AC, Study AOC, Bacot F, Baglietto L, Bandera EV, Barnholtz-Sloan J, Beckmann MW, Birrer MJ, Bloom G, Bogdanova N, Brinton LA, Brooks-Wilson A, Brown R, Butzow R, Cai Q, Campbell I, Chang-Claude J, Chanock S, Chenevix-Trench G, Cheng JQ, Cicek MS, Coetzee GA, Brca1/2 CoIoMo, Cook LS, Couch FJ, Cramer DW, Cunningham JM, Dansonka-Mieszkowska A, Despierre E, Doherty JA, Dörk T, Bois Ad, Dürst M, Easton DF, Eccles D, Edwards R, Ekici AB, Fasching PA, Fenstermacher DA, Flanagan JM, Garcia-Closas M, Gentry-Maharaj A, Giles GG, Glasspool RM, Gonzalez-Bosquet J, Goodman MT, Gore M, Górski B, Gronwald J, Hall P, Halle MK, Harter P, Heitz F, Hillemanns P, Hoatlin M, Hogdall CK, Hogdall E, Hosono S, Jakubowska A, Jensen A, Jim H, Kalli KR, Karlan BY, Kaye SB, Kelemen LE, Kiemeney LA,

Kikkawa F, Konecny GE, Krakstad C, Kjaer SK, Kupryjanczyk J, Lambrechts D, Lambrechts S, Lancaster JM, Le ND, Leminen A, Levine DA, Liang D, Lim BK, Lin J, Lissowska J, Lu KH, Lubiński J, Lurie G, Massuger LFAG, Matsuo K, McGuire V, McLaughlin JR, Menon U, Modugno F, Moysich KB, Nakanishi T, Narod SA, Nedergaard L, Ness RB, Nevanlinna H, Nickels S, Noushmehr H, Odunsi K, Olson SH, Orlow I, Paul J, Pearce CL, Pejovic T, Pelttari LM, Pike MC, Poole EM, Raska P, Renner SP, Risch HA, Rodriguez-Rodriguez L, Rossing MA, Rudolph A, Runnebaum IB, Rzepecka IK, Salvesen HB, Schwaab I, Severi G, Shridhar V, Shu XO, Shvetsov YB, Sieh W, Song H, Southey MC, Spiewankiewicz B, Stram D, Sutphen R, Teo SH, Terry KL, Tessier DC, Thompson PJ, Tworoger SS, Altena AMv, Vergote I, Vierkant RA, Vincent D, Vitonis AF, Wang-Gohrke S, Weber RP, Wentzensen N, Whittemore AS, Wik E, Wilkens LR, Winterhoff B, Woo YL, Wu AH, Xiang YB, Yang HP, Zheng W, Ziogas A, Zulkifli F, Phelan CM, Iversen E, Schildkraut JM, Berchuck A, Fridley BL, Goode EL, Pharoah PDP, Monteiro ANA, Sellers TA, Gayther SA : **Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31**. *Nature Communications* 2013, **4** :1627. 67

[119] The 1000 Genomes Project Consortium : **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012, **491**(7422) :56–65. 68

[120] Sauna ZE, Kimchi-Sarfaty C : **Understanding the contribution of synonymous mutations to human disease**. *Nature reviews. Genetics* 2011, **12**(10) :683–691. 69

[121] Jaffe A, Wojcik G, Chu A, Golozar A, Maroo A, Duggal P, Klein AP : **Identification of functional genetic variation in exome sequence analysis**. *BMC proceedings* 2011, **5 Suppl 9** :S13. 69

[122] Krawitz P, Rödelsperger C, Jäger M, Jostins L, Bauer S, Robinson PN : **Microindel detection in short-read sequence data**. *Bioinformatics (Oxford, England)* 2010, **26**(6) :722–729. 71, 74

[123] Assmus J, Kleffe J, Schmitt AO, Brockmann GA : **Equivalent Indels - Ambiguous Functional Classes and Redundancy in Databases**. *PLoS ONE* 2013, **8**(5) :e62803. 71

[124] Ingham SL, Warwick J, Buchan I, Sahin S, O'Hara C, Moran A, Howell A, Evans DG : **Ovarian cancer among 8,005 women from a breast cancer family history clinic : no increased risk of invasive ovarian cancer in families testing negative for BRCA1 and BRCA2**. *Journal of medical genetics* 2013, **50**(6) :368–372. 74

[125] Gan A, Green AR, Nolan CC, Martin S, Deen S : **Poly(adenosine diphosphate-ribose) polymerase expression in BRCA-proficient ovarian high-grade serous carcinoma ; association with patient survival**. *Human pathology* 2013, **44**(8) :1638–1647. 74

[126] Kobayashi H, Ohno S, Sasaki Y, Matsuura M : **Hereditary breast and ovarian cancer susceptibility genes (Review)**. *Oncology reports* 2013. 74

[127] Bjornslett M, Knappskog S, Lonning PE, Dorum A : **Effect of the MDM2 promoter polymorphisms SNP309T>G and SNP285G>C on the risk of ovarian cancer in BRCA1 mutation carriers**. *BMC cancer* 2012, **12** :454. 74

[128] Wickramanyake A, Bernier G, Pennil C, Casadei S, Agnew KJ, Stray SM, Mandell J, Garcia RL, Walsh T, King MC, Swisher EM : **Loss of function germline mutations in RAD51D in women with ovarian carcinoma**. *Gynecologic oncology* 2012, **127**(3) :552–555. 74

[129] Pennington KP, Walsh T, Lee M, Pennil C, Novetsky AP, Agnew KJ, Thornton A, Garcia R, Mutch D, King MC, Goodfellow P, Swisher EM : **BRCA1, TP53, and CHEK2 germline mutations in uterine serous carcinoma**. *Cancer* 2013, **119**(2) :332–338. 74

[130] Pilarski R, Patel DA, Weitzel J, McVeigh T, Dorairaj JJ, Heneghan HM, Miller N, Weidhaas JB, Kerin MJ, McKenna M, Wu X, Hildebrandt M, Zelterman D, Sand S, Shulman LP : **The KRAS-variant is associated with risk of developing double primary breast and ovarian cancer**. *PloS one* 2012, **7**(5) :e37891. 74

[131] Ramus SJ, Antoniou AC, Kuchenbaecker KB, Soucy P, Beesley J, Chen X, McGuffog L, Sinilnikova OM, Healey S, Barrowdale D, Lee A, Thomassen M, Gerdes AM, Kruse TA, Jensen UB, Skytte AB, Caligo MA, Liljegren A, Lindblom A, Olsson H, Kristoffersson U, Stenmark-Askmalm M, Melin B, SWE-BRCA, Domchek SM, Nathanson KL, Rebbeck TR, Jakubowska A, Lubinski J, Jaworska K, Durda K, Zlowocka E, Gronwald J,

Huzarski T, Byrski T, Cybulski C, Toloczko-Grabarek A, Osorio A, Benitez J, Duran M, Tejada MI, Hamann U, Rookus M, van Leeuwen FE, Aalfs CM, Meijers-Heijboer HEJ, van Asperen CJ, van Roozendaal KEP, Hoogerbrugge N, Collée JM, Kriege M, van der Luijt RB, HEBON, EMBRACE, Peock S, Frost D, Ellis SD, Platte R, Fineberg E, Evans DG, Lalloo F, Jacobs C, Eeles R, Adlard J, Davidson R, Eccles D, Cole T, Cook J, Paterson J, Douglas F, Brewer C, Hodgson S, Morrison PJ, Walker L, Porteous ME, Kennedy MJ, Pathak H, Godwin AK, Stoppa-Lyonnet D, Caux-Moncoutier V, de Pauw A, Gauthier-Villars M, Mazoyer S, Léoné M, Calender A, Lasset C, Bonadona V, Hardouin A, Berthet P, Bignon YJ, Uhrhammer N, Faivre L, Loustalot C, GEMO, Buys S, Daly M, Miron A, Terry MB, Chung WK, John EM, Southey M, Goldgar D, Singer CF, Tea MK, Pfeiler G, Fink-Retter A, Hansen TvO, Ejlertsen B, Johannsson OT, Offit K, Kirchhoff T, Gaudet MM, Vijai J, Robson M, Piedmonte M, Phillips KA, Van Le L, Hoffman JS, Ewart Toland A, Montagna M, Tognazzo S, Imyanitov E, Issacs C, Janavicius R, Lazaro C, Blanco I, Tornero E, Navarro M, Moysich KB, Karlan BY, Gross J, Olah E, Vaszko T, Teo SH, Ganz PA, Beattie MS, Dorfling CM, van Rensburg EJ, Diez O, Kwong A, Schmutzler RK, Wappenschmidt B, Engel C, Meindl A, Ditsch N, Arnold N, Heidemann S, Niederacher D, Preisler-Adams S, Gadzicki D, Varon-Mateeva R, Deissler H, Gehrig A, Sutter C, Kast K, Fiebig B, Schäfer D, Caldes T, de la Hoya M, Nevanlinna H, Aittomäki K, Plante M, Spurdle AB, kConFab, Neuhausen SL, Ding YC, Wang X, Lindor N, Fredericksen Z, Pankratz VS, Peterlongo P, Manoukian S, Peissel B, Zaffaroni D, Bonanni B, Bernard L, Dolcetti R, Papi L, Ottini L, Radice P, Greene MH, Mai PL, Andrulis IL, Glendon G, Ozcelik H, OCGN, Pharoah PDP, Gayther SA, Simard J, Easton DF, Couch FJ, Chenevix-Trench G, Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA) : **Ovarian cancer susceptibility alleles and risk of ovarian cancer in BRCA1 and BRCA2 mutation carriers**. *Human mutation* 2012, **33**(4) :690–702. 74

[132] **Online Mendelian Inheritance in Man, OMIMő. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD)** [http://omim.org/]. 74

[133] **Clin Var** [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/clinvar/]. 74

[134] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA : **COSMIC : mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer**. *Nucleic acids research* 2011, **39**(Database issue) :D945–950. 74

[135] Jourquin J, Duncan D, Shi Z, Zhang B : **GLAD4U : deriving and prioritizing gene lists from PubMed literature**. *BMC genomics* 2012, **13 Suppl 8** :S20. 74

[136] Fu YP, Edvardsen H, Kaushiva A, Arhancet JP, Howe TM, Kohaar I, Porter-Gill P, Shah A, Landmark-Hoyvik H, Fossa SD, Ambs S, Naume B, Borresen-Dale AL, Kristensen VN, Prokunina-Olsson L : **NOTCH2 in breast cancer : association of SNP rs11249433 with gene expression in ER-positive breast tumors without TP53 mutations**. *Molecular cancer* 2010, **9** :113. 74

[137] Parr C, Watkins G, Jiang WG : **The possible correlation of Notch-1 and Notch-2 with clinical outcome and tumour clinicopathological parameters in human breast cancer**. *International journal of molecular medicine* 2004, **14**(5) :779–786. 74

[138] Chu D, Zheng J, Wang W, Zhao Q, Li Y, Li J, Xie H, Zhang H, Dong G, Xu C, Li M, Chen D, Ji G : **Notch2 expression is decreased in colorectal cancer and related to tumor differentiation status**. *Annals of surgical oncology* 2009, **16**(12) :3259–3266. 74

[139] Sriuranpong V, Borges MW, Ravi RK, Arnold DR, Nelkin BD, Baylin SB, Ball DW : **Notch signaling induces cell cycle arrest in small cell lung cancer cells**. *Cancer research* 2001, **61**(7) :3200–3205. 74

[140] Christoforidou AV, Papadaki HA, Margioris AN, Eliopoulos GD, Tsatsanis C : **Expression of the Tpl2/Cot oncogene in human T-cell neoplasias**. *Molecular cancer* 2004, **3** :34. 74

[141] Chan AM, Chedid M, McGovern ES, Popescu NC, Miki T, Aaronson SA : **Expression cDNA cloning of a serine kinase transforming gene**. *Oncogene* 1993, **8**(5) :1329–1333. 74

[142] Aparecida Alves C, Silva IDCG, Villanova FE, Nicolau SM, Custódio MA, Bortoletto C, Gonçalves WJ : **Differential gene expression profile reveals overexpression of MAP3K8 in invasive endometrioid carcinoma**. *European journal of gynaecological oncology* 2006, **27**(6) :589–593. 74

[143] Clark AM, Reynolds SH, Anderson M, Wiest JS : **Mutational activation of the MAP3K8 protooncogene in lung cancer**. *Genes, chromosomes & cancer* 2004, **41**(2) :99–108. 74

[144] Gkirtzimanaki K, Gkouskou KK, Oleksiewicz U, Nikolaidis G, Vyrla D, Liontos M, Pelekanou V, Kanellis DC, Evangelou K, Stathopoulos EN, Field JK, Tsichlis PN, Gorgoulis V, Liloglou T, Eliopoulos AG : **TPL2 kinase is a suppressor of lung carcinogenesis**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(16) :E1470–1479. 74, 75

[145] Tan YC, Chow VT : **Novel human HALR (MLL3) gene encodes a protein homologous to ALR and to ALL-1 involved in leukemia, and maps to chromosome 7q36 associated with leukemia and developmental defects**. *Cancer detection and prevention* 2001, **25**(5) :454–469. 75

[146] Ruault M, Brun ME, Ventura M, Roizès G, De Sario A : **MLL3, a new human member of the TRX/MLL gene family, maps to 7q36, a chromosome region frequently deleted in myeloid leukaemia**. *Gene* 2002, **284**(1-2) :73–81. 75

[147] Lyn-Cook BD, Yan-Sanders Y, Moore S, Taylor S, Word B, Hammons GJ : **Increased levels of NAD(P)H : quinone oxidoreductase 1 (NQO1) in pancreatic tissues from smokers and pancreatic adenocarcinomas : A potential biomarker of early damage in the pancreas**. *Cell biology and toxicology* 2006, **22**(2) :73–80. 75

[148] Larson RA, Wang Y, Banerjee M, Wiemels J, Hartford C, Le Beau MM, Smith MT : **Prevalence of the inactivating 609C–>T polymorphism in the NAD(P)H :quinone oxidoreductase (NQO1) gene in patients with primary and therapy-related myeloid leukemia**. *Blood* 1999, **94**(2) :803–807. 75

[149] Malik E, Cohen SB, Sahar D, Dann EJ, Rund D : **The frequencies of NAD(P)H quinone oxidoreductase (NQO1) variant allele in Israeli ethnic groups and the relationship of NQO1*2 to adult acute myeloid leukemia in Israeli patients**. *Haematologica* 2006, **91**(7) :956–959. 75

[150] Liu G, Yang G, Chang B, Mercado-Uribe I, Huang M, Zheng J, Bast RC, Lin SH, Liu J : **Stanniocalcin 1 and ovarian tumorigenesis**. *Journal of the National Cancer Institute* 2010, **102**(11) :812–827. 75

[151] Block GJ, DiMattia GD, Prockop DJ : **Stanniocalcin-1 regulates extracellular ATP-induced calcium waves in human epithelial cancer cells by stimulating ATP release from bystander cells**. *PloS one* 2010, **5**(4) :e10237. 75

[152] Li C, Ao J, Fu J, Lee DF, Xu J, Lonard D, O'Malley BW : **Tumor-suppressor role for the SPOP ubiquitin ligase in signal-dependent proteolysis of the oncogenic co-activator SRC-3/AIB1**. *Oncogene* 2011, **30**(42) :4350–4364. 75

[153] Geng C, He B, Xu L, Barbieri CE, Eedunuri VK, Chew SA, Zimmermann M, Bond R, Shou J, Li C, Blattner M, Lonard DM, Demichelis F, Coarfa C, Rubin MA, Zhou P, O'Malley BW, Mitsiades N : **Prostate cancer-associated mutations in speckle-type POZ protein (SPOP) regulate steroid receptor coactivator 3 protein turnover**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(17) :6997–7002. 75

[154] Huang DW, Sherman BT, Lempicki RA : **Bioinformatics enrichment tools : paths toward the comprehensive functional analysis of large gene lists**. *Nucleic acids research* 2009, **37** :1–13. 75

[155] Huang DW, Sherman BT, Lempicki RA : **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2009, **4** :44–57. 75

[156] Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, Rual JF, Borick H, Braun P, Dreze M, Vandenhaute J, Galli M, Yazaki J, Hill DE, Ecker JR, Roth FP, Vidal M : **Literature-curated protein interaction datasets**. *Nat Methods* 2009, **6** :39–46. 77

[157] Pesch R, Zimmer R : **Complementing the Eukaryotic Protein Interactome**. *PloS one* 2013, **8**(6) :e66635. 78

[158] Garcia-Garcia J, Schleker S, Klein-Seetharaman J, Oliva B : **BIPS : BIANA Interolog Prediction Server. A tool for protein-protein interaction inference**. *Nucleic acids research* 2012, **40**(Web Server issue) :W147–151. 78

[159] Yu H, Tardivo L, Tam S, Weiner E, Gebreab F, Fan C, Svrzikapa N, Hirozane-Kishikawa T, Rietman E, Yang X, Sahalie J, Salehi-Ashtiani K, Hao T, Cusick ME, Hill DE, Roth FP, Braun P, Vidal M : **Next-generation sequencing to generate interactome datasets**. *Nature Methods* 2011, **8**(6) :478–480. 78

[160] Lewis ACF, Jones NS, Porter MA, Deane CM : **What Evidence Is There for the Homology of Protein-Protein Interactions ?** *PLoS Comput Biol* 2012, **8**(9) :e1002645. 78

[161] Lewis A : **Communities and Homology in Protein-protein Interactions**. *PhD thesis*, University of Oxford 2011. 78

[162] Ideker T, Krogan NJ : **Differential network biology**. *Molecular systems biology* 2012, **8** :565. 80

[163] Souiai O, Becker E, Prieto C, Benkahla A, De las Rivas J, Brun C : **Functional integrative levels in the human interactome recapitulate organ organization**. *PloS one* 2011, **6**(7) :e22051. 80

[164] Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabasi AL, Vidal M : **An empirical framework for binary interactome mapping**. *Nat Methods* 2008, **6** :83–90.

[165] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P : **Comparative assessment of large-scale data sets of protein–protein interactions**. *Nature* 2002, **417** :399–403.

[166] Schwartz AS, Yu J, Gardenour KR, Finley Jr RL, Ideker T : **Cost-effective strategies for completing the interactome**. *Nat Methods* 2009, **6** :55–61.

[167] Bader GD, Hogue CWV : **Analyzing yeast protein–protein interaction data obtained from different sources**. *Nat Biotechnol* 2002, **20** :991–997.

[168] Huang H, Jedynak BM, Bader JS : **Where have all the interactions gone ? Estimating the coverage of two-hybrid protein interaction maps**. *PLoS Comput Biol* 2007, **3** :e214.

[169] Huang H, Bader JS : **Precision and recall estimates for two-hybrid screens**. *Bioinformatics* 2009, **25** :372–378.

[170] Chiang T, Scholtens D, Sarkar D, Gentleman R, Huber W : **Coverage and error models of protein-protein interaction data by directed graph analysis**. *Genome Biol* 2007, **8** :R186.

[171] Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF : **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae**. *Nature* 2006, **440** :637–643.

[172] Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G : **Proteome survey reveals modularity of the yeast cell machinery**. *Nature* 2006, **440** :631–636.

[173] Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, De Smet AS, Venkatesan K, Rual JF, Vandenhaute J, Cusick ME, Pawson T, Hill DE, Tavernier J, Wrana JL, Roth FP, Vidal M : **An experimentally derived confidence score for binary protein-protein interactions**. *Nat Methods* 2009, **6** :91–97.

[174] Lee I, Date SV, Adai AT, Marcotte EM : **A probabilistic functional network of yeast genes**. *Science* 2004, **306** :1555–1558.

[175] Gentleman R, Huber W : **Making the most of high-throughput protein-interaction data**. *Genome Biol* 2007, **8** :112.

[176] Xin X, Rual JF, Hirozane-Kishikawa T, Hill DE, Vidal M, Boone C, Thierry-Mieg N : **Shifted Transversal Design smart-pooling for high coverage interactome mapping.** *Genome Res* 2009, **19** :1262.

[177] Aryee MJA, Quackenbush J : **An Optimized Predictive Strategy for Interactome Mapping**. *J Proteome Res* 2008, **7** :4089–4094.

[178] Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe TY, Schroeder M, Weng S, Botstein D : **SGD : Saccharomyces genome database**. *Nucleic Acids Res* 1998, **26** :73.

[179] Hakes L, Pinney JW, Robertson DL, Lovell SC : **Protein-protein interaction networks and biology–what's the connection ?** *Nat Biotechnol* 2008, **26** :69–72.

[180] Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D : **DIP : the database of interacting proteins**. *Nucleic Acids Res* 2000, **28** :289–291.

[181] Vinayagam A, Stelzl U, Wanker EE : **Repeated two-hybrid screening detects transient protein–protein interactions**. *Theor Chem Acc* 2010, **125** :613–619.

[182] Salwinski L, Licata L, Winter A, Thorneycroft D, Khadake J, Ceol A, Aryamontri AC, Oughtred R, Livstone M, Boucher L, Botstein D, Dolinski K, Berardini T, Huala E, Tyers M, Eisenberg D, Cesareni G, Hermjakob H : **Recurated protein interaction datasets**. *Nat Methods* 2009, **6** :860–861.

[183] Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, Rual JF, Borick H, Braun P, Dreze M, Vandenhaute J, Galli M, Yazaki J, Hill DE, Ecker JR, Roth FP, Vidal M : **Addendum : Literature-curated protein interaction datasets**. *Nat Methods* 2009, **6** :934–935.

[184] Sahai H, Khurshid A : **A note on confidence intervals for the hypergeometric parameter in analyzing biomedical data**. *Comput Biol Med* 1995, **25** :35–38.

[185] Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G : **Absolute quantification of somatic DNA alterations in human cancer**. *Nature Biotechnology* 2012, **30**(5) :413–421.

[186] Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ, Ji HP, Snyder M : **Performance comparison of whole-genome sequencing platforms**. *Nature biotechnology* 2012, **30** :78–82.

[187] Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J : **SNP detection for massively parallel whole-genome resequencing**. *Genome Research* 2009, **19**(6) :1124–1132.

[188] Reumers J, De Rijk P, Zhao H, Liekens A, Smeets D, Cleary J, Van Loo P, Van Den Bossche M, Catthoor K, Sabbe B, Despierre E, Vergote I, Hilbush B, Lambrechts D, Del-Favero J : **Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing**. *Nature Biotechnology* 2012, **30** :61–68.

[189] Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H : **SNVer : a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data**. *Nucleic Acids Research* 2011, **39**(19) :e132.

Annexes

Puissance et limitations du RNA-seq

Ci-joint la version de l'article SEQC soumise à Nature Biotechnology. Les commentaires des critiques et de l'éditeur sont positifs et encourageants, et des réponses à leur remarques sont en cours de rédaction.

Power and Limitations of RNA-Seq: findings from the SEQC (MAQC-III) consortium

MAQC Consortium*

In the US FDA-led SEQC (*i.e.*, MAQC-III) project, different sequencing platforms were tested across more than ten sites using well-established reference RNA samples with built-in truths in order to assess the discovery and expression-profiling performances of platforms and analysis pipelines. The results demonstrate that novel exon-exon junctions can still be discovered beyond existing comprehensive annotations and sequencing depth. Extensive investigations encompassing diverse performance metrics characterizing reproducibility, accuracy, and information content were combined with comparisons to qPCR and microarray platforms showing high levels of inter-site and cross-platform concordance for differentially expressed genes. However, performance is clearly platform and pipeline dependent, and transcript-level profiling shows larger variation. Together with applications of RNA-Seq to several preclinical and clinical problems, the entire SEQC data sets comprise over 100 billion reads (10 Tb) and provide a unique resource for testing future developments of RNA-Seq in clinical and regulatory settings.

*A full list of authors and their affiliations appears at the end of the paper.
Correspondence and requests for materials should be addressed to L.S. (leming.shi@gmail.com), C.E.M. (chm2042@med.cornell.edu), or D.P.K. (david.kreil@boku.ac.at / d.kreil@warwick.ac.uk).

Next-generation sequencing has revolutionized the scale and scope of genomic research. This technological advance has made deep whole-transcriptome sequencing ('RNA-Seq') feasible, which has since expanded our view of the transcriptome[1] and promises to permit quantitative profiling with large dynamic range.[2] Recent comparisons with established technologies for differential expression analysis have found good overall agreement between RNA-Seq, qPCR, and microarrays. In general, RNA-Seq has provided increased sensitivity and opened new avenues of research in transcriptome work, such as the study of gene fusions, allele-specific expression, and novel alternative transcripts, whereas the measurement noise of RNA-Seq was shown to be a direct consequence of the random sampling process.[3–6]

While new platforms and protocols for RNA-Seq have emerged in recent years, the comparability of results across platforms and laboratories has remained unclear. With the widespread adoption of RNA-Seq in large-scale research efforts such as ENCODE,[7] TCGA,[8] and ICGC,[9] a comprehensive, cross-site and cross-platform analysis of the performance of RNA-Seq is now essential. Reproducibility across laboratories, in particular, is a crucial requirement for any new experimental method in research and clinical applications, and can only be tested in an extensive multi-site and multi-platform comparison. Just as in the first phase of the MicroArray Quality Control (MAQC-I) project,[10] which tested multi-site and multi-platform agreement in gene-expression microarrays, the FDA has coordinated the work presented here, which represents a large-scale community effort called the Sequencing Quality Control (SEQC/MAQC-III) project to assess the performance of RNA-Seq, testing different sequencing platforms and analysis pipelines.

Without an independent 'gold standard', however, the objective assessment of a novel genome-scale measurement method remains non-trivial. A multitude of properties must be examined, including different metrics of reproducibility, accuracy, and information content. Such a multi-dimensional characterization is critical for the development of more powerful studies of the underlying biological mechanisms in complex data sets because often there is a trade-off between one desirable property and another (such as accuracy *vs* precision). We thus set out to achieve such a characterization efficiently in a controlled test setting, where truths built into the study design could be directly validated, and then related this to the performance observed in genome-scale research applications (**Fig. 1a**).

Specifically, we utilized the well-characterized reference RNA samples A (Universal Human Reference RNA) and B (Human Brain Reference RNA) from the MAQC consortium,[10] adding spike-ins of synthetic RNA from the External RNA Control Consortium (ERCC).[11] Samples C and D were then constructed by combining A and B in known mixing ratios, 3:1 and 1:3, respectively (**Suppl. Fig. S1**). All samples were distributed to several independent sites for RNA-Seq library construction and profiling by Illumina's HiSeq 2000 platform and Life Technologies' SOLiD 5500 platform. Also, vendors created their own cDNA libraries that were then distributed to each test site, in order to examine the degree of a 'site effect' that was independent of the library preparation process (**Fig. 1b**). To support an assessment of gene models, samples A and B were also sequenced at independent sites by the Roche 454 GS FLX platform, providing

longer reads. For comparison to other technologies, these data were also compared to the Affymetrix U133Plus2.0 microarrays used in MAQC-I, several current microarray platforms, and also assessed by 20,801 PrimePCR reactions. We here report an analysis of RNA-Seq measurement performance in this controlled setting, whereas analyses focusing on measurement quality metrics (S. Li *et al., submitted*), the ERCC spike-ins and limits of detection (S.A. Munro *et al., submitted*), and the effects of pipeline choice (J.H. Phan *et al., submitted*) are presented separately.

These studies are complemented by applications of RNA-Seq to a number of research applications (**Fig. 1a**), including the critical assessment of performance in neuroblastoma outcome prediction (W. Zhang *et al., submitted*) and a comparative investigation of toxicogenomic samples with a comprehensive study design testing chemicals with multiple modes of action (C. Wang *et al., submitted*). In total, over 100 billion reads (10 terabases) of RNA-Seq data were produced and studied, representing the most ambitious effort so far by the RNA-Seq community to generate and analyze large reference data sets. A rigorous dissection of sources of noise and signal indicates that, given appropriate data treatment and analysis, RNA-Seq is highly reproducible, particularly in differential gene-expression analysis.

RESULTS

Gene detection and junction discovery depend on read depth

Efficient quantitative expression profiling takes advantage of known gene models,[3] and the choice of a reference annotation can considerably affect results, including performance assessments. Specifically, our data showed that the number of reads mapped to known genes depends on the accuracy and completeness of the gene models. Among all 23.2 billion reads that could be mapped to genes other than mitochondrial or ribosomal RNAs, 85.9% were mapped to RefSeq,[12] whereas 92.9% mapped to GENCODE,[13] and 97.1% to NCBI AceView[14] (**Fig. 2a**). This is not a property of the samples examined (A, B, C, and D), as a similar trend is seen when adding all reads from the extensive SEQC neuroblastoma project (Supplement). The read fraction unique to AceView is genuinely due to the higher accuracy of its gene models (**Fig. 2a**).

Since the data constitute the deepest sequencing of any set of samples yet performed, comprising a total of 12 billion mapped HiSeq 2000 RNA-Seq fragments, we examined how well the known genes could be detected as a function of aggregate read depth, taking all replicate libraries, sites, and samples together (**Fig. 2c** for HiSeq 2000 and **Suppl. Fig. S2** for SOLiD). We report read depth as the number of read fragments because the mapping and counting of paired ends are highly correlated; single-ended reads can thus be used when the additional long-range information from read pairs is not required. At a sequencing depth of 10 million aligned fragments, about 35,000 of all the genes annotated in AceView[14] (55,674) were already found by at least one read. Some of these are due to background noise, for instance, from genomic DNA contamination. Using the background estimates of the NCBI Magic pipeline and requiring four reads above this intergenic noise still yielded about 26,000 genes (Thierry-Mieg, *pers. comm.*). For a comparison of alternative pipelines, annotations, and the effect of read depth, we next focused on genes with strong support (16 reads or more). At this stringency, we found that about 20,000 genes were detected at a sequencing depth of 10 million aligned fragments, which covered the majority of strongly expressed genes. Detection increased to over 30,000 at 100 million fragments, and finally to over 45,000 at about one billion fragments. While the number of additional known genes detected successively decreased for each doubling of read depth, additional genes were still being detected even at high read depths of >1 billion fragments, indicative of cellular rarity or low expression levels per cell.

Exon-exon junction detection was also found to increase with read depth with a similar RefSeq, GENCODE, and AceView annotation dependency (**Fig. 2d**). Generally, many additional known junctions were detected with each doubling of the read depth for the more comprehensive annotations, even at high read depths exceeding one billion. Since the samples A and B in the study are very different, we expected to see more transcriptional complexity when combining samples. Indeed, a detection boost was seen for the addition of the biologically distinct sample B (**Fig. 2d**), which indicates that different samples contribute more to exploring the complex transcriptome space than merely increasing read depth (W. Zhang *et al.,* Y. Yu *et al.; submitted*). The number of additional known junctions decreased fastest for RefSeq, which provides the least

complex annotation (**Suppl. Fig. S3**), with practically all annotated junctions observed at the highest read depth. Considering more complex annotations, the AceView database is the most comprehensive and has the highest number of junctions supported by reads from this study, reaching over 300,000 junctions at the maximum read depth, more than three times the number detected at 10 million reads. Although GENCODE and AceView have similar total numbers of genes and similar footprints on the genome, considerably fewer annotated genes and junctions in GENCODE were supported by the observed reads (**Fig. 2b** and **Suppl. Fig. S4**). Therefore, all subsequent analyses presented in this manuscript are based on AceView, unless stated otherwise.

We first examined the reproducibility of detecting genes and junctions across measurement sites, platforms, and analysis pipelines, since a key strength of RNA-Seq is its inherent ability to identify splice sites *de novo*. To test this ability of RNA-Seq to discover junctions, we first examined the HiSeq 2000 data because of the greater read length and depth. We considered three independent pipelines for *de novo* discovery of junctions independent of existing gene models: NCBI Magic,[14] r-make (which uses STAR)[15] and Subread.[16] All pipelines reported millions of junctions, with r-make predicting about 50% more than Subread and Magic, although almost all junctions found by Subread or Magic were also found by r-make. There was substantial but lower overlap with the junctions found by TopHat2,[17] whether it was run with or without gene model guided alignment. The five analyses together predicted 2.9 million new splice junctions, yet only 671,916 (23%) were consistently found by all the methods (**Suppl. Fig. S5**), illustrating the significant difficultly in reliable splice junction detection with current methods (more below).

Novel junctions were then examined for independent discovery in both HiSeq 2000 and SOLiD data, as junctions only found by a single platform or library preparation protocol could be due to technical artifacts (**Suppl. Fig. S6**). Reflecting lower read length and depth (*cf.* simulation, **Suppl. Table S1**), we discovered 87,117 junctions predicted to be novel in the SOLiD data, of which 74,561 (86%) were also independently discovered from HiSeq 2000 reads (Subread). The curves for the detection of these new junctions in the four samples (**Fig. 2e**) followed an expected order corresponding to sample complexity: B<A<D<C.

We then used the built-in truths of the data sets to examine the accuracy of the sample specific levels of support of novel junctions in terms of their ability to capture the expected A/B sample mixing ratio and yield a consistent titration order. The examples in **Fig. 1c** illustrate the concept of titration order consistency: if A > B then we expect A > C > D > B because C = ¾ A + ¼ B and D = ¼ A + ¾ B, and we expect the inverse order if B > A. This consistency test is affected both by systemic distortions reducing accuracy and random variations reducing reproducibility. Another complementary test assesses the A/B mixing ratio recovery in the samples C (3:1) and D (1:3), which can be examined in a plot of $\log_2$(C/D) *vs* $\log_2$(A/B). Deviations from the ideal line (**Fig. 1d**) are also affected by both systemic distortions reducing accuracy and random variations reducing reproducibility. Because both tests only reflect reproducibility when A~B, we also require a clear differential signal as assessed by the mutual information, a measure of information content. (See Methods for further details.) These filters reduced the

discoveries to only 8% 'well-behaving' junctions among the 74,561 novel junctions found. Such a reduction, however, is also observed for well-established junctions, where the same filters showed that only 38% of known junctions were 'well-behaving'. We observed that the greatest determinant of splice junction detection reliability was simply the expression level, reflecting the number of supporting fragments. Indeed, requiring a consistent titration order and the correct A/B mixing ratio clearly enriched for junctions with higher expression levels (**Fig. 2f**), which were the easiest to measure and quantify. A comparison with junctions detected in Roche 454 data confirmed that the more abundant junctions could reliably be detected across different RNA-Seq platforms (Suppl. Fig. S7).

For an examination of how well junctions newly discovered by RNA-Seq could independently be confirmed by a different technology, qPCR was performed with primer pairs designed to specifically validate 173 detected junctions: We randomly selected 136 well-supported junctions that had been discovered *de novo* by all three RNA-Seq analysis pipelines, in both HiSeq 2000 and SOLiD data. These junctions were chosen so that they log-uniformly covered a range of ~10–3,000 supporting reads, and so that half of them met all consistency tests. In addition, we also tested 13 known AceView junctions as positive controls and 24 novel junctions that had only been discovered by a single analysis pipeline in the HiSeq 2000 or SOLiD data, despite having support by many reads (~300–3,000). Only 5 assays were non-informative (see Methods). Interestingly, in the remaining assays, not just all of the 13 positive controls but also all of the 133 well-supported junctions could reliably be identified by qPCR, with the numbers of supporting RNA-Seq reads largely reflecting qPCR expression level estimates (slope 0.95, Pearson/Spearman correlation 0.74/0.77, $N = 146$), even for junctions with low read numbers or not meeting all consistency tests (Suppl. Fig. S8). Moreover, 18 of the 22 pipeline specific junctions (>80%) could be confirmed at least qualitatively. For the most comprehensive surveys of potential new junctions, one may therefore want to consider all discovered junctions, although it makes sense to prioritize well-supported junctions found consistently (Fisher $p < 4\times10^{-4}$).

In the complete data set comprising SEQC samples A, B, C, and D, we observed consistent detection of some 44,000 known genes (**Fig. 2g**) and 310,000 known exons across pairs of replicate sites (**Suppl. Fig. S9**), constituting about 79% and 47% of all known genes and exons, respectively. Nearly 200,000 splice junctions were seen consistently, making up about 50% of all known junctions. This corresponds to about 90%, 87%, and 83% of all detected known genes, exons, and junctions, respectively, reflecting that larger features aggregating more reads were easier to measure reproducibly. The fluctuations in the detection of sequence features stemmed largely from sequencing depth-dependent sampling noise, which were reflected in the very similar intra- and inter-site agreements in the detection of known genes (**Fig. 2g**), exons (**Suppl. Fig. S9**), and junctions (**Fig. 2h**). Considering the low technical variance besides the unavoidable sampling noise, biological replicates are advisable.

Filters improve the reliability of differential expression analysis

Studies on microarrays have shown that results of typical statistical differential expression tests thresholded by *p*-value need to be filtered and sorted by effect strength (fold-change) in order to attain robust comparisons across platforms and sites.[10] We thus sought to identify corresponding requirements for RNA-Seq, examining the reproducibility of rank ordered lists of differentially expressed genes (DEGs), as well as the False Discovery Rate reflecting the information content of the measurements. Since the exact same samples were profiled at all sites, the true number DEGs is zero when comparing the same sample between any two sites. Any DEGs found for self–self comparisons thus represent technical differences and can be considered 'false positives' (shown as dots in **Fig. 3a**). We then examined the number of inter-site A *vs* A 'false positives' relative to the number of DEGs in A *vs* B comparisons, giving an empirical estimate of the False Discovery Rate (eFDR). We tested several RNA-Seq data analysis pipelines for the six HiSeq 2000 sites, focusing on the set of 23,000 genes present on the Affymetrix U133Plus2.0 microarrays for comparison.

We found that unfiltered data for both RNA-Seq and microarrays show many DEGs, both for false positives (A *vs* A) and the likely 'real' DEGs in the A *vs* B comparison (**Fig. 3a**), with the ratio of false positives *vs* true positives unacceptably high for both platforms (**Fig. 3b**). Also, we observed that different analysis pipelines vary in these measures of performance (**Figs 3a-e**). Next, we applied the $|\log_2$ fold-change$|>1$ filter advocated in the MAQC study on microarrays,[10] and for microarrays observed a reduction in the eFDR to below 1.5%, save for one outlier site (**Figs 3c-d**). Importantly, we could then show that the application of pipeline dependent filters for *p*-value, fold-change, and expression-level (lowest third of all examined AceView genes**, Suppl. Tables S2** and **S3**) reduced the RNA-Seq eFDR across sites to close to zero false-positives without sacrificing sensitivity (**Fig. 3d**). We note that results for SOLiD were very similar (**Suppl. Figs S10** and **S11**), with expression level thresholds reflecting the lower read depth for that platform.

After the application of these filters, our data showed that most (but not all) RNA-Seq pipelines demonstrated high inter-site reproducibility of differential expression calls with up to 95% concordance in DEGs (**Fig. 3e** and **Suppl. Fig. S12**). This concordance between sites was highest for those genes with the greatest expression levels, just as in the similar relation for splice junction agreement and titration. Moreover, the filters resulted in a good agreement of differential expression calls across platforms (*e.g.*, A *vs* B on HiSeq 2000 compared to A *vs* B on SOLiD, **Suppl. Figs S13** and **S14**), suggesting that differential expression analyses from different platforms can be combined to extend existing studies with additional samples.

Relative but not absolute expression measures satisfy tests of built-in truths

After an examination of the reliability of differential expression analysis of genes, we next examined the quantification of RNA using four consistency tests exploiting ground truths built into the study design (**Fig. 4**). First, we considered titration order consistency as introduced in **Fig. 1c**, a metric affected both by systemic distortions reducing accuracy and random variations reducing reproducibility. The majority of genes (59%) correctly titrated (**Fig. 4a**), with little disagreement between platforms (**Suppl. Table S4**). Genes with large differential expression behaved best, with all genes showing consistent titration in several HiSeq 2000 and SOLiD sites, and no contradiction regarding the direction of change (blue curve). For the second built-in truth, we examined the A/B mixing ratio recovery (**Fig. 1d)** as another test reflecting accuracy and reproducibility. We observed the correct ratio for the majority of genes (**Fig. 4b**), with better agreement at higher expression levels (top 25%). Notably, the scatter of genes marked as titrating in this plot indicates that consistent titration does not guarantee a reliable recovery of the mixing ratio (and *vice versa*).

The third and fourth built-in truths leveraged the ERCC spike-ins.[11] These data complement work in the ERCC paper, which examined the measurement of fold-change recovery for these synthetic RNAs (S.A. Munro *et al.; submitted*). Across platforms, we observed that with sufficiently high expression levels ($\log_2$[conc]>3), the expected ratios of 1/2, 2/3, 1, and 4 were accurately recovered using about 90 million mapped fragments (**Fig. 4c**), with high precision indicating good reproducibility. Finally, we examined the ERCC absolute titration levels, since the ERCC RNAs were spiked into the samples A and B before the C and D samples were created (**Suppl. Fig. 1**). We observed, however, that the fraction of reads aligning to ERCC spike-ins for a given sample varied widely between libraries and platforms, with ranges of measured ERCCs between 1–2.5% for HiSeq 2000 and 2.5–4.7% for SOLiD, with a clear 'library effect' observed for all sites and platforms, affecting reproducibility. (**Fig. 4d** and **Suppl. Fig. S15**). Indeed, when using the vendor-prepared library as the cross-site control, we observed very consistent measurements of the percentages of reads mapping to ERCCs, which indicates a large degree of variation from the preparation of libraries at a single site.

In addition, when the 92 ERCC spike-in RNAs are compared to their nominal concentrations, some of them (*e.g.*, ERCC-116) are systematically measured up to 10 times below or above their expected concentrations, perturbing even the order of the ERCC scale. These discrepancies are highly reproducible, suggesting that the bias is sequence dependent. While marginal trends could be observed as functions of GC content and average sequence region unfolding probabilities (Suppl. Figs S16 and S17), these did not pass tests for significance. No consistent trend could be observed as function of mRNA length (Suppl. Fig. S18), and the majority of the deviations is not explained by any of these co-variates, indicating a need for further investigations of such distortions and their possible sources. We observed, however, that the effect is also protocol dependent and is reduced in the absence of poly-A selection (**Suppl. Fig. S19**). This is in line with results in other studies, further underscoring the impact of protocol choice on quantification (S.A. Munro *et al*; S. Li *et al*; submitted), where fragmentation time, poly-

A enrichment by columns or beads / ribo-depletion, hexamer or oligo-dT priming, library isolation by gel or beads, different ligation efficiency and RNA quality at the start of library preparation have all been shown to have an effect. Consequently, just as for microarrays, absolute measurements by RNA-Seq using a particular protocol are well reproducible but not very accurate. This observation implies that the use of external spike-in controls to accurately infer absolute expression levels of a gene of biological interest remains challenging.

RNA-Seq relative gene expression measurements compare well with alternative platforms

Since we observed good performance for RNA-Seq in consistency tests of relative expression levels, we then sought to compare alternative measurement platforms. We first examined the differential expression of 843 genes measured by TaqMan for samples A and B in the MAQC-I study. Although more strongly expressed than typical AceView genes, these genes nevertheless span a wide range of expression levels (**Suppl. Fig. S20**). We found good and comparable agreement among different platforms (**Fig. 5a**, where Pearson/Spearman correlation coefficients are given; also *cf.* **Suppl. Fig. S21**), with the HiSeq 2000 and SOLiD sequencing platforms showing the highest correlation to one another. This is in line with other comparisons of relative expression measures.[4,6,18,19]

For absolute expression levels, correlations to TaqMan were slightly better for RNA-Seq than for microarrays (0.83 *vs* 0.79, $p = 0.02$), and the average trend follows a more linear shape (**Suppl. Fig. S22b**, $R^2 = 0.68$ *vs* 0.62). While, on average, agreement could be found for absolute expression levels between different platforms, there were substantial deviations across the entire dynamic range for large numbers of individual genes. These deviations are systematic, *i.e.*, are not a question of reproducibility but rather affect the accuracy of absolute expression measures. In particular, comparing expression level estimates from HiSeq 2000 and SOLiD RNA-Seq (**Fig. 5b**), we observed hundreds of genes that were expressed according to one platform but not the other. This effect is only partly due to the non-stranded nature of the Illumina protocol used here, and the presence of 11,066 genes antisense to genes annotated on the opposite strand (*cf.* **Fig. 5c**).

As an independent additional test, we generated 20,801 PrimePCR measurements of the SEQC samples A, B, C, and D. We again observed that a substantial portion of genes was not considered expressed by one platform but showed strong expression in the other (**Suppl. Figs. S23 and S24**). While quantitative PCR based methods have traditionally been used as a reference 'gold' standard due to their high sensitivity and dynamic range, it is noteworthy in this context that specific primer selection and protocol calibration are challenging in their own right.[20] PCR is affected by GC bias[21] and considerable differences in expression level measurements from different PCR based assays can be observed (**Fig. 5d**).

Performance assessment is metric dependent

A major promise of RNA-Seq is the extension of expression profiling to the discovery and quantification of alternative transcripts. For transcript-specific profiling, however, no large-scale expression data from other technologies are available as an external reference point. The SEQC data represent a first opportunity for the multi-platform comparison of transcript-specific measurements.

To support a balanced performance study of gene-level and transcript-specific expression profiling, we combined multiple metrics for a robust characterization of platforms, sites, and data processing options: (1) average measurement precision,[3] directly assessing reproducibility, (2) titration order consistency[22] and (3) recovery of the expected A/B mixing ratio, providing two complementary assessments reflecting both measurement accuracy and reproducibility, as well as (4) differential expression and (5) the mutual information of sample titration, capturing different aspects of information content (see Methods). For a summary view, we first focused on genes with a clear directional signal, *i.e.*, those that allowed an ordered discrimination of the samples A to D, as indicated by mutual information (5). We then count how many of these genes also satisfy a second metric (1)-(4). Such an integration of tests through counting genes that fulfill multiple assay criteria allows a comprehensive consideration of all the genes instead of restricting comparisons to a common subset of genes always identified as expressed. This is necessary for a meaningful comparison of pipelines and platforms with varying degrees of sensitivity (**Suppl. Figs S25a** and **S25b**). The resulting four assays are complementary, *i.e.*, a gene satisfying one does not generally satisfy the others (**Suppl. Fig. S26**). The average of the four assays then provides a Consistency Score for robust characterization of measurement performance (**Suppl. Fig. S27**).

For gene-level profiling, on average, pipelines showed similar performances (**Fig. 6a**). Providing known gene models always considerably improved results (*cf.* TopHat2). Relatively lower scores for transcript-level profiling indicate that the discrimination of alternative transcripts is more difficult, which is also reflected in stronger effects of pipeline choice (**Fig. 6b**). RNA-Seq has sparked an interest in transcript-specific profiling and the development of advanced algorithms for estimating alternative transcript abundances. With known gene models, similar approaches can now also be applied for microarrays. We thus next focused on a test set of 782 genes with multiple alternative transcripts of varying complexity and specifically selected to represent the full subset of spliced genes in AceView (*cf.* Methods). Covering 5,691 alternative transcripts, this test set allows a first comparison of transcript-specific expression level estimates by RNA-Seq and a high-resolution transcript-level microarray. We found that efficient transcript-specific measurements with good precision on microarrays for quantitative expression profiling (**Fig. 6d**, **Suppl. Figs S27d** and **S28d**) could complement the power of RNA-Seq in the discovery and identification of new alternative transcripts (**Fig. 2**). In other words, the novel transcripts found by RNA-Seq can lead to efficient measurements with good precision on microarrays, which can in turn aid in the confirmation and functional study of new transcript variants.

Finally, each metric showed a different and platform specific response to signal strength (**Suppl. Figs S29** and **S30**), which for RNA-Seq increases with transcript expression level and read depth. The read depth at which average RNA-Seq performance meets or exceeds that of another platform thus directly depends on the chosen metric and the distribution of expression strength and differential signal in the samples measured. As a result, it also depends on the set of tested genes, over which the average performance is being computed. We show results here for the mutual information metric (**Suppl. Fig. S31**), which is of direct relevance for classifier performance. As expected, RNA-Seq performance improved with increasing numbers of mapped fragments (**Fig. 6e**). In particular, Life Technologies' SOLiD and Illumina's HiSeq 2000 performed similarly well for comparable effective read depths (**Suppl. Fig. S32a**). The choice of reference platform affects the number of RNA-Seq reads required for obtaining comparable mutual information per gene considerably (**Suppl. Fig. S33**). For some of the microarrays and data-processing methods tested, as little as 5 M mapped RNA-Seq fragments would more than suffice (U133plus2 with MAS5), while about 50 M mapped fragments were required for others (PrimeView with gcRMA/affyPLM). RNA-Seq pipeline choice also had an effect, with some tools requiring up to twice as many aligned fragments (*cf.* TopHat2+Cufflinks,[23] **Suppl. Fig. S34**).

DISCUSSION

In a multi-center study led by the US FDA, different sequencing platforms were tested using four well-characterized reference RNA sample mixtures with built-in truths to test accuracy, precision, reproducibility, sensitivity and specificity in a detailed analysis of over 30 billion reads on the reference samples alone. The data presented here provide the deepest molecular characterization of any RNA samples published to date.

Leveraging this extraordinarily deep data set and the known truths built into the study design, we have tested the reliability and power of RNA-Seq in exploring the complexity of the transcriptome, studying the detection of known splice junctions and the discovery of potential new junctions. We found robust inter-site agreement in *de novo* feature discovery, both at low and at high sequencing depths, even beyond 10 billion aligned fragments. This capacity for discovery constitutes a key strength of RNA-Seq, and is reflected in the expansive transcriptional landscapes observed from different cells and tissues in the transcriptome re-annotation projects for human and rat (Y. Yu *et al.,* P. Li *et al.; submitted*) and the rich profiles collected in clinical and toxicogenomic applications in which many terabases of additional RNA-Seq data were collected and analyzed (W. Zhang *et al.,* C. Wang *et al.; submitted*).

Several points can be made from the compiled data. First, the collected profiles underscore how crucial comprehensive gene model annotations are to accurate expression profiling.[3] The human genome now has more than 55,000 well-validated genes, and the majority of them are not protein coding.[7] Additional genes, splice junctions and transcripts are still being discovered, even beyond the already expansive gene annotation from ENCODE.[13] Interestingly, the NCBI AceView[14] database, which has over 50,000 genes annotated from cDNA evidence since 2004, still holds by far the largest and most-extensively validated set of splice junctions, with over 300,000 well supported by the RNA-Seq data reported in this study alone. Genes that may explain a particular phenotype may be missed by less extensive annotations, stressing that the most comprehensive annotation for expression profiling is vital to accurate clinical research (W. Zhang *et al.,* C. Zhao *et al.; submitted*). While analogous considerations will apply to research on mouse and rat models, and also other complex transcriptomes, simpler organisms such as *C. elegans* may be less affected.[24]

New splice junctions were found to be robustly supported by multiple platforms and pipelines, and their expression levels were the single largest predictor of reliability. Similar to observations by ENCODE at the gene level,[7] we observed three distinct classes of expression levels for splice junctions: highly expressed known junctions, known and novel junctions at medium level, and many novel junctions found only at low expression level. While it has been proposed that an abundance of weakly expressed transcripts may reflect biological noise,[25] the lower-expression levels of these junctions may alternatively be the reason that they have so far received less attention in traditional experiments. Future studies can be conducted to identify novel alternative transcripts through full gene models, which allow the filtering of spurious junctions that cannot be explained by expression levels of alternative transcripts consistent with exon mapping reads. With alternative transcript expression being cell type- and condition-dependent, deep RNA-Seq

will continue to play a key role in fully exploring the transcriptomic repertoire, including extensive maps of alternative transcripts (Y. Yu *et al.,* P. Li *et al.; submitted*) highlighting splicing variants as well as alternative start and polyadenylation sites.[7]

With substantial disagreements even between different types of qPCR based assays, we conclude that there is no single 'gold standard'. Tests of truths built into the SEQC study design as well as cross-platform comparisons demonstrated that, while common trends for absolute expression levels could be seen, drastic systemic differences remain. Reference data sets such as the compendium presented here, however, are invaluable for a systematic characterization of measurements, forming the basis of reliable conclusions from large-scale experiments. Specifically, a closer examination of the varying amount of detected ERCCs per sample indicates substantial differences and inconsistencies even across libraries prepared from the same sample at the same site and sequenced by the same machine, which implies inherent limitations for the read out of absolute expression level estimates and absolute quantification.[26] The observed variations likely arose in library construction, as the vendor-prepared libraries gave very uniform results across sites (**Fig. 4d**). This finding may partially be explained by the varying degrees of poly-adenylation in different samples,[26] and platform-specific differences in library preparation chemistry. RNA-Seq experiments ideally should therefore use multiple libraries *per* examined condition or sample. Notwithstanding these limitations, RNA-Seq across sites showed that the vast majority of genes still satisfied constraints based on the truths built into the study design.

Although no technology tested could provide reliable absolute quantification, relative expression measures agreed well across platforms, including RNA-Seq, qPCR, and microarrays, and the majority of genes satisfied constraints based on the truths built into the study design. Going beyond earlier platform comparisons that considered individual performance metrics,[3–6,18,19,27] we combined complementary metrics for a robust characterization of measurement performance that can be combined with further assays such as tests for strandedness, considerations based on the ERCC spike-in response, and tests for the exclusion of non-specific background. Notably, only the Magic pipeline attempted the important but difficult task of estimating and removing background noise (**Suppl. Fig. S35b**), typically improving accuracy at the expense of precision (**Suppl. Fig. S28**).

Almost all multi-exon genes exhibit alternative splicing, and spliced human genes have on average over nine alternative transcribed forms.[14] However, our data show that the area of greatest improvement needed in RNA-Seq is for the characterization and quantification of alternative transcripts. While expression profiling of alternative transcripts is feasible, reattributing measurements to a set of alternative transcripts requires knowledge of all the alternative transcript forms of a gene, and involves combining information across the transcripts. For genome-scale RNA-Seq, this is particularly difficult because of the sampling noise from low read counts for many transcripts, with recent work observing 300M read fragments to be required for the detection of a specific human alternative splicing event with 80% power.[28] This highlights the value of targeted RNA-Seq.[29]

The need for longer-range information indicates that certain complex gene models cannot be resolved by the local information provided by either microarray probes or individual short RNA-Seq reads alone. While longer reads and read pairs of size-controlled fragments can improve on this, they also limit the recovery of shorter transcripts. Full alternative transcript profiling will thus greatly benefit from longer RNA-Seq reads, which may eventually approach the full length of complex cDNAs. Combining deep RNA-Seq for alternative transcript discovery with modern high-resolution microarrays for genome-scale quantification may provide an efficient approach for systematic transcript-level expression profiling.[18]

Our analysis indicates that filters can improve robustness of differential expression calls and consistency across sites and platforms. For RNA-Seq, removing small fold-changes as well as excluding low-expression measurements successfully reduced the false discovery rate considerably, often close to zero, and in general gave an improvement over microarrays[10] at similar sensitivity. These filters also achieved good inter-site agreement of lists of differentially expressed genes, with several (but not all) RNA-Seq pipelines comparable to microarrays (**Fig. 3e**). Even though a direct general quantitative comparison of absolute expression levels from different platforms was not possible, the filters nonetheless yielded good agreement of differential expression calls between platforms (*e.g.*, A *vs* B on HiSeq 2000 compared to A *vs* B on SOLiD (**Suppl. Figs S13** and **S14**), suggesting that differential expression analyses from different platforms could be combined.

Importantly, the observed sensitivity of results to pipeline choice suggests that the need remains for significant improvement in short-read RNA-Seq analysis, particularly for transcript identification and quantification. The data we collected in this multi-center study can serve as a benchmark set for further advances. Some recent progress directly reflects the impact of more successful read mappers.[16] In addition, while systematic and sample-specific variations in GC bias,[30] sequence bias, and non-specific signal (**Suppl. Fig. S35**) can contribute to unwanted or missed differential expression calls, continued study of the confounding factors in RNA-Seq can be expected to improve signal quality,[3,19,30,31] just as methodological developments have improved microarray signal read out.[32–38] Conversely, with several microarray designs tested here probing less than half of all known AceView genes, new microarray designs can take advantage of updated gene annotation and its refinement by RNA-Seq,[18] as we have shown here with pilot microarrays.

Already today, RNA-Seq can be employed as a versatile tool for relative expression profiling, with comparable or superior performance to microarrays in many applications, given sufficient read depth and appropriate choice of analysis pipeline. An effective sequencing depth is clearly contingent on the experimental goals, with simple gene-level expression profiling only requiring 5–50 M single-ended reads for an appropriate analysis pipeline (*cf.* **Suppl. Figs S11** and **S33, Fig. 6e**), while a comprehensive characterization of alternative transcript expression benefits from the longer-range information of read pairs and requires considerably deeper sequencing. In our data set, at five million mapped fragments, over 15,000 AceView genes could already be detected with strong support (16 reads), including ~10,000 RefSeq genes (**Fig. 2c**).

Moreover, 10 million mapped fragments sufficed for differential expression analysis of the most strongly expressed genes in our study, reliably across sites (see **Suppl. Fig. S11** for adapted filter parameters). Other applications may require deeper sequencing, as is reflected by different metrics responding differently to an increase in reads for the samples and genes studied. Classifier performance, for instance, is directly related to the mutual information metric. While 5 M mapped fragments easily gave a mutual information per gene comparable to that of U133Plus2.0 microarrays, performance comparable to PrimeView microarrays required about 50 million mapped fragments in this study (**Fig. 6e** and **Suppl. Fig. S33**), or required considerably more, depending on the analysis pipeline (*cf.* TopHat2+Cufflinks, **Suppl. Fig. S34**). In addition, the required read depth is also dependent on the genome size, transcriptional complexity,[24] and cellular distribution of stored *vs* active RNAs, biological noise, as well as the panoply of all other factors of cell biology and RNA dynamics.

In summary, the study and data collection presented here forms an important milestone in the development and dissection of RNA-Seq as a method for transcriptome profiling. The results based on data sets of this unprecedented scope and an array of independent measures as provided by this study will contribute to a better understanding of the power and limitations of RNA-Seq. These results are complemented by SEQC companion studies analyzing the application of RNA-Seq to specific biological research and clinical questions (S. Li *et al.,* S.A. Munro *et al.,* J.H. Phan *et al.,* W. Zhang *et al.,* C. Wang *et al.,* Y. Yu *et al.,* P. Li, *et al.,* C. Zhao *et al.; submitted*). The cumulative SEQC data sets with over 100 billion reads (10 Tb) provide a unique resource for testing future developments of RNA-Seq in clinical and regulatory settings.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

ACCESSION CODES

All SEQC (MAQC-III) data sets are available through GEO (series accession number: GSE47792).

ACKNOWLEDGMENTS

All SEQC (MAQC-III) participants freely donated their time and reagents for the completion and analyses of the project. Many participants contributed to the sometimes-heated discussions on the topic of this paper during numerous e-mail exchanges, teleconferences, and face-to-face project meetings. The common conclusions and recommendations reported in this paper evolved from this extended discourse. The authors gratefully acknowledge support by the NCBI's Supercomputing Center, the FDA's Supercomputing Center, China's National Supercomputing Center of Tianjin, the Vienna Scientific Cluster HPCF (VSC), the Vienna Science and Technology Fund (WWTF), Baxter AG, the Austrian Institute of Technology, and the Austrian Centre of Biopharmaceutical Technology. This work was supported in part by the National Institutes of Health (NIH) grants R01CA163256, R01HG006798, R01NS076465, R44HG005297, U54CA119338, PO1HG00205, R24GM102656 and the Intramural Research Program of the NIH, National Library of Medicine, National Institute of Environmental Health Sciences (NIEHS) Z01 ES102345-04, Shriners Research Grant 85500, an Australia National Health and Medical Research Council (NH&MRC) Project Grant (1023454) and Victorian State Government Operational Infrastructure Support (Australia), the National 973 Key Basic Research Program of China (2010CB945401), the National Natural Science Foundation of China (31240038 and 31071162), and the Science and Technology Commission of Shanghai Municipality (11DZ2260300). We greatly appreciate SAS Institute, Inc. for kindly hosting several face-to-face meetings of the SEQC (MAQC-III) project.

DISCLAIMER

The views presented in this article are those of their authors and do not necessarily reflect current or future opinion or policy of the authors' institutions or agencies. Any mention of commercial products is for clarification and not intended as endorsement.

COMPETING FINANCIAL INTERESTS

Some authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

FIGURE LEGENDS

Figure 1 The SEQC (MAQC-III) project and experimental design. (**a**) SEQC project overview. We report on a group of studies assessing different sequencing platforms in real-world use cases, including transcriptome annotation and other research applications, as well as clinical settings. This paper focuses on the results of a complementary elaborate multi-center experiment with built-in ground truths. (**b**) SEQC main study design. Similar to the MAQC-I benchmarks, well-characterized RNA samples A and B were augmented by samples C and D comprised of A and B in known mixing ratios 3:1 and 1:3, respectively. These allow tests for titration consistency (**c**) and the correct recovery of the known mixing ratios (**d**). Synthetic RNAs from the External RNA Control Consortium (ERCC) were both pre-added to samples A and B before mixing and also sequenced separately to assess dynamic range (samples E and F). Samples were distributed to independent sites for RNA-Seq library construction and profiling by Illumina's HiSeq 2000 (3+3x) and Life Technologies' SOLiD 5500 (3+1x). Unless mentioned otherwise, data presented shows results from the three official sites (*italics*). In addition to the replicate libraries A1…D4 at each site, for each platform, one vendor-prepared library A5…D5 was being sequenced at all sites, giving a total of 240 libraries. At each site, each library has a unique bar-code sequence and all libraries were pooled before sequencing, so each lane was sequencing the same material, allowing a study of lane specific effects. To support a later assessment of gene models, samples A and B were also sequenced by Roche 454 (3x, no replicates, see Supplement). (**c**) Schema illustrating tests for titration order consistency. Four examples are shown. The dotted lines represent the ideal mixture of samples A and B (blue and red) expected for samples D and C (purple and green). (**d**) Schema illustrating a consistency test for recovering the expected sample mixing ratio. The green lines mark a 10% deviation from the expected response (red) for a perfect mixing ratio. Both tests (**c**) and (**d**) will reflect both systemic distortions (bias) and random variation (noise).

Figure 2 Gene detection and junction discovery. (**a**) The fraction of all reads aligned to gene models from different annotations, RefSeq, Encode, and NCBI AceView (green). Reads aligning only to specific annotations are shown in dark green. (**b**) Known genes (left) and exon junctions (right) supported by at least 16 HiSeq 2000 or SOLiD reads are in green; genes or junctions annotated but not observed at this threshold are shown in grey. (**c**)-(**e**) show sensitivity as a function of read depth. (**c**) Known genes detected. We show the number and percentage of all AceView annotated genes detected for three RNA-Seq analysis pipelines, Subread (red), r-make (blue) and Magic (green). The *x*-axis marks cumulative aligned fragments from all replicates and sites. Vertical lines indicate boundaries between samples A…D. (**d**) Known junctions detected. The numbers and percentages of all exon-exon junctions (supported by 8 or more reads) are shown for different gene model databases (line style). Horizontal lines show the respective total numbers of annotated junctions. (**e**) New junctions supported by multiple platforms and pipelines. Subsets of novel junctions have expression levels with correct titration orders and mixing ratios (*cf.* **Figs 1b-d** and **4a-b**). (**f**) Distribution of junction expression levels. New junctions, then new junctions supported by multiple platforms and pipelines, and known junctions show increasing expression levels (colors). Subsets expressed with mutual information about the samples and correct titration order and mixing ratio display a further shift towards higher expression levels (dashed lines). (**g**)-(**h**) Intra- (blue) and inter-site reproducibility (orange) of detected known genes (**g**) and junctions (**h**). Pairwise agreement is shown by bars (with whiskers) representing mean counts (with standard deviations). Lines mark percentages.

Figure 3 Sensitivity, specificity, and reproducibility of differential expression calls. Robust cross-site analyses depend on pipeline choice and appropriate filter rules. Results are shown for five MAQC-III RNA-Seq pipelines and the long established MAQC-I Affymetrix microarray platform (color). (**a** and **c**) Number of standardized differential expression calls. All possible pairwise inter-site A *vs* A comparisons (•) are shown next to all intra-site A *vs* B comparisons (+) as indicators for specificity and sensitivity. (**b** and **d**) Ratios of A *vs* A calls and A *vs* B calls give an estimate of the false discovery rate (eFDR). For all platforms and pipelines, differential expression calls identify thousands of differences in inter-site comparisons of identical samples. These can be controlled for microarray by additional filters for effect size. In addition, RNA-Seq also requires filters for expression strength due to the high sampling fluctuations at lower read counts. These were set to give similar numbers of A *vs* B expression calls (**b**), improving the eFDR to less than 1.5% but for a number of outliers. (**e**) Inter-site reproducibility of differential expression calls. Comparing the identities and the directions of change for differentially expressed genes (DEGs) across sites, agreement is plotted for lists including the top-ranked N genes as sorted by effect size (x-axis). The observed response curves depend on pipeline and filter choice, showing more variation for shorter lists. Several RNA-Seq pipelines are comparable to the MAQC-I microarrays, with a small advantage for microarrays when considering only the differentially expressed genes with the strongest fold change (left side of **e**).

Figure 4 Built-in truths for assessing RNA-Seq. (**a**) Titration order A/C/D/B. Log_2 fold-change is related to cross-platform titration consistency. At sufficiently strong log_2 fold-change, reliable titration is also found across platforms: The blue line represents 'unmissable' genes showing the correct titration order with no contradiction in at least 14 HiSeq 2000 and 6 SOLiD samples. Most genes with high differential expression are in this class. (**b**) Known A/B mixing ratios in samples C and D. The yellow solid line traces the expected values after mRNA/total-RNA shift correction. The 1%, 10%, and 25% most highly expressed genes are shown in red, cyan, and magenta. On average, the most strongly expressed genes recover the expected mixing ratio best. Genes with inconsistent titration (*cf.* **a**) are colored grey. Black and grey symbols intermixing indicates that consistent titration (black) does not guarantee reliable recovery of the mixing ratio (and *vice versa*). (**c**) ERCC spike-in ratios can be recovered increasingly well at higher expression levels. From the response curves, one can calculate signal thresholds for the detection of a change.[10] (**d**) Variation of the total amounts of detected ERCC spikes. The lack of reliable titration indicates that the considerable differences between libraries of a given site and protocol are random, implying limits for absolute expression level estimates, in general, and using spike-ins for the calibration of absolute quantification, in particular. The observed variations likely arise in library construction, as the vendor-prepared libraries (colored bars) gave constant results across different sites. For (**a**) and (**b**), all 55,674 AceView genes tested.

Figure 5 Cross-platform agreement of expression levels. (**a**) Comparison of $\log_2$ fold-change estimates for 843 selected genes. Good and comparable concordances were observed between relative expression measures from the MAQC-III HiSeq 2000 and SOLiD sequencing platforms, MAQC-I TaqMan, and the MAQC-III Affymetrix HuGene2 arrays (Pearson/Spearman correlation coefficients are shown; *cf.* **Suppl. Fig. S21**). (**b**) Comparison of absolute expression levels from HiSeq 2000 and SOLiD in a rank scatter density plot. Expression level ranks for sample A are shown on the *x*-axis for HiSeq 2000, and on the *y*-axis for SOLiD. Genes are represented by dots, and areas with several genes are shown in blue, with darker blue corresponding to a higher gene density in the area. Large cross-platform deviations are seen even for highly expressed genes and these variations are systematic. The genes in the vertical 'spur', for instance, are not detected by SOLiD RNA-Seq but show strong expression levels in HiSeq 2000 RNA-Seq, with an analog comparison to 19,604 qPCR measurements giving a similar picture (**Suppl. Fig. S24**. The ERCC spike-ins are shown as red symbols (+). ERCC spike-in signals are systematically lower in the HiSeq 2000 data, which may be explained by their shorter poly-A tails and differences in the library construction protocols. (**c**) The same plot as (**b**) but removing the 11,066 genes that can be affected by the non-stranded nature of the applied Illumina protocol. While the actual number of genes in the vertical spur that are not detected by SOLiD but show strong expression levels in the HiSeq 2000 is now smaller, it is still substantial. (**d**) Comparison of TaqMan and PrimePCR for 843 selected genes. Expression level estimates vary considerably for individual genes, with some genes showing high expression in one platform while not being detected at all by the other.

Figure 6 Multiple performance metrics for the quantification of genes and alternative transcripts. The *y*-axes show a Consistency Score, secondary *y*-axes mark the percentage of the maximal possible score. Panels show the three official HiSeq 2000 and SOLiD sites and compare a few analysis variants (red=TopHat2, Green=TopHat2 guided by known gene models, Black=Magic, Grey=BitSeq, and orange=Subread). Panels **a** and **b** consider all AceView annotated genes. Panels **c** and **d** focus on a subset of expressed complex genes with multiple alternative transcripts where comparison to a high-resolution test microarray (rightmost bar) can be conducted. Panel (**e**) compares RNA-Seq to four different microarrays and data-processing methods (red bars) by plotting the mutual information (*y*-axes) at different read depths (*x*-axes). For the microarrays, the number of probes used is shown. The numbers given for RNA-Seq state the number of fragments mapped to genes as well as the [total fragments]. SOLiD and HiSeq 2000 performed similarly well for comparable effective read depths (**Suppl. Fig. S32a**). HiSeq 2000 data is plotted here. Each bar shows the minima and maxima across the three official sites. The read depth for which average RNA-Seq performance met or exceeded that of the array is marked by a green bar. The corresponding read depths varied widely from 5 M (U133plus2 with MAS5) to about 50 M fragments (PrimeView with gcRMA/affyPLM), showing the strong effect of the reference gene set implied by the probes on the respective arrays and the employed microarray data-processing methods. Results are shown for the Subread pipeline. Alternative RNA-Seq data analysis pipelines, however, can require up to double the number of fragments (TopHat2+Cufflinks, **Suppl. Fig. S34**). See **Suppl. Figs S32** and **S33** for comparisons of other platforms and read depths.

METHODS

Study design and data

The SEQC (MAQC-III) main study design is based on the well-characterized MAQC-I RNA samples: the Universal Human Reference RNA (UHRR, from 10 pooled cancer cell lines, Agilent Technologies, Inc.) and the Human Brain Reference RNA (HBRR, from multiple brain regions of 23 donors, Life Technologies, Inc.).[10] To these, two different ERCC spike-in mixes were added[11] (50 µl of ERCC mix was spiked into 2500 µl of total RNA) to give: Sample A – UHRR with ERCC spike-in mix E, and Sample B – HBRR with ERCC spike-in mix F. These were then combined in ratios of 3:1 and 1:3, respectively, to generate samples C and D (**Fig. 1b** and **Suppl. Fig. S1**).

Illumina HiSeq 2000 data were provided by 6 sites (**Suppl. Tables S5** and **S6**): 1) Australian Genome Research Facility; 2) Beijing Genomics Institute*; 3) City of Hope; 4) Weill Cornell Medical College*; 5) Mayo Clinic*; and 6) Novartis, generating 100+100 nt read-pairs.

Life Technologies SOLiD 5500 data were provided by 4 sites (**Suppl. Tables S7** and **S8**): 1) Liverpool; 2) Northwestern University*; 3) Penn State University*; and 4) SeqWright Inc.*, generating 51+36 nt read-pairs, except for Liverpool which applied a protocol variant giving single 76 nt reads.

* Each platform vendor designated three 'official sites' before samples were distributed. Data produced by the official sites were used in all of the performed analyses. In addition, data produced by the non-official sites were incorporated in some analyses, e.g., the analysis of gene detection and junction discovery as a function of read depth (**Figs 2, S2-S4, S6, S7 and S36**) and the study of sensitivity, specificity, and reproducibility of differential expression calls (**Figs 3, S10-S14 and S37-S39**).

All official sites created 4 replicate measurements of each sample A to D, and also sequenced a vendor-prepared fifth replicate (**Fig. 1b**). The other HiSeq 2000 sites sequenced 4 replicate libraries of each sample A to D. In Liverpool, one site-prepared library and one vendor-provided library of each of the samples A to D were sequenced.

For comparisons of gene-level expression profiling, samples A to D were also hybridized to a variety of commercial microarray platforms: 1) Affymetrix HuGene2.0 (one site: Stanford); 2) Affymetrix PrimeView (one site: Stanford); 3) Agilent 60k (one site: Boku University Vienna); and 4) Illumina Bead arrays (two sites: City of Hope, and University of Texas Southwestern Medical Center). In addition, MAQC-I Affymetrix U133Plus2.0 data from six sites were reanalyzed. Providing another independent platform, 20,801 PrimePCR measurements were also performed, with at least 10 qPCR reactions per assay to assure good specificity, efficiency, linear dynamic range, and background from negative controls (see Supplementary Methods for more detail).

For comparisons of transcript-level profiling, exploring the potential of high-density microarrays for alternative transcript specific quantification, an Agilent 1M feature microarray was tested at Boku University Vienna. The microarray contained 1 million probes of length 60 nt, covering 782 AceView genes with 5,691 alternative transcripts, and including the ERCC spike-ins, averaging: 33 probes per exon (7x coverage, 9 nt spacing) and 55 probes per junction (about 1 nt spacing). The set of genes was selected to: 1) show expression in one of the samples in an SEQC RNA-Seq pilot study; 2) have a similar average expression distribution as the full set of AceView genes in the pilot study; 3) have a similar differential expression distribution as the full set of AceView genes in the pilot study; and 4) have a similar distribution of the number of transcripts *per* gene as the full set of genes annotated in AceView. These and similar microarrays can be ordered from Agilent, and the design of the test microarray is published together with this paper. Affymetrix also manufactures high-density transcriptome microarrays, which were released early 2013, not in time to be included in the SEQC study.

Roche 454 GS FLX data were provided by: 1) the Medical Genomes Project; 2) the New York University Medical Center; and 3) SeqWright Inc.. At each site, one replicate of samples A and B was sequenced (two runs).

Reads were mapped to a human reference and the ERCC spike-in sequences. Depending on the pipeline, genomic DNA (hg19) or transcript sequences were used as human reference. Unless otherwise stated, results for the gene model annotation of AceView 2010 are shown. Other annotations considered included RefSeq v104 and GENCODE v15.

The HiSeq 2000 sites produced on average 110 million read-pairs *per* replicate, for a total of 2,200 million *per* site (**Suppl. Tables S5** and **S6**). The official SOLiD sites produced on average 50 million read-pairs *per* replicate, for a total of 980 million *per* site (**Suppl. Table S7**). Liverpool generated 545 million single reads (**Suppl. Table S8**). The Roche 454 sites produced on average 1 million reads *per* replicate, for a total of about 2.1 million reads *per* site (**Suppl. Table S9**).

For the validation of junctions discovered by RNA-Seq, for a random selection of 173 junctions to test, qPCR measurements were performed with primers designed to specifically validate the particular junction, running 2 qPCR reactions per assay for all samples A…D (see Supplementary Methods for more detail). Specificity was confirmed by analyses of PCR product lengths. This allowed the identification of non-specific assays, identifying the target but also picking up cross-targets (giving qualitative validation but no meaningful quantitative readout) and of non-informative assays, failing to pick up the target but picking up cross-targets. We provide information on RNA-Seq read coverage flanking all 250 candidate junctions considered for validation in **Supplementary File 1**. **Supplementary File 2** provides the employed qPCR primer sequences, qPCR results and expression level estimates, as well as the corresponding RNA-Seq expression level estimates for the 173 performed assays.

Data processing – assessing expression estimates

A variety of tools/pipelines to process RNA-Seq data were compared:

TopHat2 std: TopHat v2.0.0[17] + CuffDiff v2.0.0[23].

TopHat2 G: TopHat v2.0.0 with -G parameter (providing the reference GTF file) + CuffDiff v2.0.0.

Magic: NCBI AceView MAGIC[14].

BitSeq: SHRiMP2 v2.2.2[39] + BitSeq v0.4.2[40].

Subread: Subread 1.3.0[16] The Subread pipeline uses the subjunc function to identify exon-exon junctions and the featureCounts function to obtain count summaries for each gene and spike-in transcript (see Supplementary Methods for more detail).

r-make: Cornell's r-make pipeline incorporating STAR[15] (http://physiology.med.cornell.edu/faculty/mason/lab/r-make/).

For LifeTech reads, all alignments were processed in color space.

Applying consistency tests based on truths built into the study design to expression levels of individual junctions, we consider the number of reads hitting a specific exon-exon junction as indicator of expression level.

Except for r-make, which provides raw read counts, each pipeline already has a built-in approach to normalization. In order to analyze Agilent 1M microarray data, a variance stabilizing normalization (vsn)[33] was used. Probe sequences-specific signals have been modeled using established methods, saturation effects detrended, and outlier probes downweighted.[35–37] Transcript variant expression levels have been estimated using a hierarchical Bayesian approach similar to modern methods applied for RNA-Seq data analysis (Stegle *et al.* in preparation) (see 'Transcript quantification for Agilent high-density microarrays' section).

CustomCDFs (v16, re-mapped to the latest AceView) were used for an analysis of the Affymetrix data (U133Plus2.0, PrimeView, and HuGene2.0),[41] respectively covering 24,623, 17,984, and 29,879 genes. PrimeView and HuGene2.0 data were analyzed using established methods (correction for probe sequence specific effects by gcRMA,[34] conservative normalization across arrays by vsn,[33] and robust probe set summarization by affyPLM), whereas for the U133Plus2.0 microarrays, a combination of more recent tools that appeared to be more efficient were used (correction for probe specific saturation effects by Hook,[36] conservative normalization across arrays by vsn,[33] and factor based probe set summarization by FARMS[35]).

For Illumina Bead microarrays and Agilent 60K microarrays variance stabilization normalization (vsn[33]) was applied.

Discovery of transcriptome complexity at high read depth

The detection and discovery of junctions was performed by Subread using data from all 6 HiSeq 2000 sites as well as all 4 SOLiD sites, and compared to results by r-make using data from all 6 HiSeq 2000 sites.

We applied consistency tests for the truths built into the study design to junction expression levels. The sample expression levels of almost 1/3 of all known AceView junctions (and 38% of all detected) follow the expected titration order while also correctly yielding the expected A/B mixing ratio and show a clear differential signal as assessed by the mutual information, a measure of information content (**Suppl. Fig. S36** and **Suppl. Table S10**). Of the well-supported new junctions, 5,189 passed these rigorous filters. Conservatively assuming a 1:3 ratio as in the AceView junctions, there may easily be three times as many new junctions, thereby adding over 15,000 likely new junctions to the already extensive AceView annotation. Furthermore, considering that we essentially required junctions to be independently detected in by SOLiD, where 98% of junctions were detected by a single site (LIV) and thus corresponding to just about 1/65 of the total HiSeq 2000 sequencing volume, and detection power being about 2x lower (**Suppl. Table S1**), there may well be up to 5,000×3×65×2 = 2 million new junctions to be discovered in samples A and B alone. Even more junctions than in samples A and B were discovered in the SEQC neuroblastoma study.[15]

Transcript quantification for Agilent high-density microarrays

Quantification of transcript expression from the probe level information was carried out using a linear mixed model independently for each gene and sample. Denoting the expression level of probe p as y_p, we model the probe expression as the sum of effects from transcripts with a probe-match:

$$y_p = \sum_{t=1}^{T} \delta_{t,p}\, x_t \,.$$

The Kronecker delta $\delta_{t,p}$ is one exactly if the probe p is matching the transcript t, and zero otherwise, whereas x_t denotes the unknown abundance for transcript t. Further, we assume Gaussian additive and multiplicative error variances. Probe-level noise tends to exhibit a strong spatial correlation structure, which we account for by using a latent Gaussian process function.[42] We employ a squared exponential covariance function where the probe distance in transcript space is used to parameterize the covariance.

Inference is performed by maximizing the joint marginal likelihood of all considered probes given with respect to the hidden transcript abundances (x_t) and the noise covariance parameters. To mitigate the computational complexity of Gaussian process models (cubical scaling in the number of probes), we randomly choose probes for each gene selecting a subset of at most 700 probes, including probes falling onto junctions.

Discrete nature of RNA-Seq data

With the discrete nature of RNA-Seq data and considering that most analysis tools work on a $\log_2$ scale, consistent ways need to be found for dealing with not expressed features, which are supported by zero reads. As the lowest positive expression is just a single read, a common approach is the addition of a pseudo-count (*e.g.*, 0.5 in voom[32]). An alternative well established for microarray data analysis is the application of asinh as a variance stabilizing transform assuming an additive-multiplicative error model. The transform is approximately linear for small values; for larger values it is well approximated by a logarithm. Another approach (natively applied *e.g.* in Magic[14]) is to use an artificial 'background' value. Here a threshold is set to the highest minimum read count of all measurement samples and this threshold is set as floor to all expression levels. We have applied this approach in our studies for each pipeline and platform (thus we were not adding the pseudo-count of 0.5 reads when using voom).

In order to identify genes, transcripts, and junctions with clear support of sequence reads, thresholds were applied: Support was considered sufficient when at least 16, 16, or 8 at reads were observed, respectively. For **Figure 2b**, expression above background level as determined by the Magic pipeline was additionally required for genes.

Sensitivity, specificity, and reproducibility of differential expression calls

In this part of the study the subset of 23,420 AceView genes (of the version frozen for the SEQC study) present on the MAQC-I Affymetrix U133Plus2.0 microarray was used.

As the array data were already processed and normalized with state-of-the-art methods (see 'Data processing' sections) no further processing was required. For RNA-Seq data, weighted trimmed mean of log fold-change normalization[31] improved results (data not shown), in agreement with a recent performance comparison of normalization methods.[43] For this normalization step, the TMM implementation provided in the Bioconductor R package edgeRM[44] was employed.

Several of the examined RNA-Seq pipelines exploit multi-mapping reads. This increases power (J.H. Phan *et al., submitted*) and, more generally, also allows the analysis of alternative transcripts. Those pipelines report expression-level estimates rather than read counts. For a uniform approach to differential expression analysis, RNA-Seq data were therefore analyzed using an established approach supporting such pipelines. Precision-based weights were attached to normalized expression estimates on the log-scale to account for higher variability at low expression levels using voom of the limma package.[32] The voom function has been developed to account for different variances as a function of signal intensity. For count data, such a variation is expected by theory, whereas for expression level estimates it is empirically justified. So we in general apply the voom model for expression level dependent variance to account for the different

platform specific noise characteristics as a function of the expression level (**Suppl. Fig. S37**).

Differential expression was then assessed for both microarray and sequencing platforms using the empirical Bayes moderated *t*-statistic of the limma package.[32] A *p*-value threshold of 0.01 unadjusted for multiple testing was used, as suggested in the MAQC-I study.[10] As the number of differentially expressed genes (DEGs) was similar for $p<0.01$ and the $q_{BY}<5\%$, where q_{BY} is the Benjamini-Yekutieli adjusted False Discovery Rate, downstream analysis is not qualitatively affected by this choice (**Suppl. Fig. S38** *vs* **Fig. S39**).

An estimate of the empirical False Discovery Rate (eFDR) was computed by comparing the number of DEGs for intra-site A *vs* B and inter-site A *vs* A comparisons. For each A *vs* A analysis two eFDRs were calculated (using the A *vs* B comparison of the two sites considered in the matching A *vs* A comparison).

Further filters were applied in order to control the eFDR, with parameters chosen to give similar numbers of A *vs* B differential expression calls:

AFX: $|\log_2(\text{fold-change})| > 1$

MAGIC: $|\log_2(\text{fold-change})| > 1.7$ and $\text{AveExp} > 32\%$

r-make: $|\log_2(\text{fold-change})| > 1.7$ and $\text{AveExp} > 33\%$

Subread: $|\log_2(\text{fold-change})| > 1.7$ and $\text{AveExp} > 32\%$

BitSeq: $|\log_2(\text{fold-change})| > 2$ and $\text{AveExp} > 19\%$

TopHat2 -G $|\log_2(\text{fold-change})| > 2$ and $\text{AveExp} > 23\%$

When filtering for average $\log_2$ expression level (AveExp), the stated fraction of weakly expressed genes to remove also included in that percentage the genes that were not observed at all. This is to allow comparisons across different pipelines observing varying numbers of genes.

Metrics for a robust characterization of platforms, sites, and data processing options

The complementary metrics examined react differently to rescaling, shifts, and other consequences of data-processing. As a result, individual pipelines can show varying performance in specific assays. For a robust performance characterization we combine the complementary metrics.

Different analysis pipelines and platforms, however, identify varying numbers of targets (Supplement). With this number not constant, performance needs to be assessed in terms of actual counts or fractions of all genes, rather than fractions of observed genes. An increased sensitivity of some pipelines and platforms can be demonstrated by not limiting analysis only to genes observed by all pipelines and platforms.

1) We count a gene as preserving titration monotonicity, when A > B and, as expected A ≤ C ≤ D ≤ B or, conversely, for A < B.

2) A gene is considered precise for a sample, if the standard error across technical replicates is below 10%.

3) A deviation less than 10% from the expected behavior of

$$\log\frac{C}{D} = \log\left(k_1\frac{A}{B} + (1-k_1)\right) - \log\left(k_2\frac{A}{B} + (1-k_2)\right)$$

where the correction z of the known mixing coefficients $k_1 = 3z/(3z+1)$ and $k_2 = z/(z+3)$ arising out of different ratios of mRNA *versus* total RNA in the samples A and B has been determined by a non-linear robust fit (nlrob) from an independent RNA-Seq library (library #5). The obtained value, 1.45±0.01 is very much in line with the experimental estimate[22] of 1.43±0.10. The plots and statistics shown in the paper give the same picture with either value. To ensure pipeline independence, the experimental value 1.43 is being used.

4) For the purpose of this metric, we call a gene differentially expressed if it is significant at a Benjamini-Yekutieli corrected FDR of 5% in an empirical Bayes moderated *t*-test across the expression level estimates of samples A and B (limma).

5) Finally, we calculate the mutual information of sample titration by extending the approach introduced in for two state measurements.[38] The mutual information between gene or transcript expression and titration requires modeling the probability of a measurement being from sample A, C, D, or B, under the constraint that these labels are ordered. To avoid a dependency of our assessment on choosing mutual information of sample titration as evaluation measure, we complemented this assay with three other alternative measures. For that purpose, we evaluated the mutual information for discriminating A *vs* B, and the mutual information for discriminating C *vs* D using an established approach.[38] In order to add a further measure which does not depend on modeling assumptions, for all genes and transcripts, we also calculated a non-parametric estimate of the probability that the respective measurement fulfils the order constraint which is implied by the titration experiment. All four measures are illustrated in **Suppl. Fig. S40** for the official HiSeq 2000 and SOLiD sites and a number of different quantification pipelines. Although the complementary measures suggest different numbers of 'good' transcripts and genes, they qualitatively agree with no exception on how they rank the different platforms and pipelines. This confirms that we can select the mutual information of sample titration to represent this class of information preserving measures to add an independent robust viewpoint for characterizing quantification performance in **Figure 6**.

References

1. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).
2. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).
3. Łabaj, P. P. *et al.* Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics* **27,** i383–i391 (2011).
4. Liu, S., Lin, L., Jiang, P., Wang, D. & Xing, Y. A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. *Nucleic Acids Res.* **39,** 578–588 (2011).
5. McIntyre, L. M. *et al.* RNA-seq: technical variability and sampling. *Bmc Genomics* **12,** 293 (2011).
6. Toung, J. M., Morley, M., Li, M. & Cheung, V. G. RNA-sequence analysis of human B-cells. *Genome Res.* **21,** 991–998 (2011).
7. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
8. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
9. (Chairperson), T. J. H. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
10. Shi, L. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24,** 1151–1161 (2006).
11. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. *Nat. Methods* **2,** 731–734 (2005).
12. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40,** D130–D135 (2011).
13. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
14. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts. *Genome Biol.* **7,** S12 (2006).
15. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* (2012). doi:10.1093/bioinformatics/bts635
16. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41,** e108–e108 (2013).
17. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
18. Xu, W. *et al.* Human transcriptome array for high-throughput clinical studies. *Proc. Natl. Acad. Sci.* **108,** 3707–3712 (2011).
19. 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).
20. VanGuilder, H., Vrana, K. & Freeman, W. Twenty-five years of quantitative PCR for gene expression analysis. *BioTechniques* **44 Supplement,** 619–626 (2008).

21. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12,** R18 (2011).
22. Shippy, R. *et al.* Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nat. Biotechnol.* **24,** 1123–1131 (2006).
23. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2013).
24. Agarwal, A. *et al.* Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *Bmc Genomics* **11,** 383 (2010).
25. Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *Plos Genet* **6,** e1001236 (2010).
26. Qing, T., Yu, Y., Du, T. & Shi, L. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. *Sci. China Life Sci.* **56,** 134–142 (2013).
27. Raghavachari, N. *et al.* A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *Bmc Med. Genomics* **5,** 28 (2012).
28. Liu, Y. *et al.* Evaluating the Impact of Sequencing Depth on Transcriptome Profiling in Human Adipose. *Plos One* **8,** e66883 (2013).
29. Levin, J. Z. *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol.* **10,** R115 (2009).
30. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40,** e72–e72 (2012).
31. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11,** R25 (2010).
32. Smyth, G. K. in *Bioinforma. Comput. Biol. Solutions Using R Bioconductor* (Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 397–420 (Springer New York, 2005).
33. Huber, W., Heydebreck, A. von, Sültmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18,** S96–S104 (2002).
34. Wu, Z., Irizarry, R., Gentleman, R., Murillo, F. M. & Spencer, F. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *Johns Hopkins Univ. Dept Biostat. Work. Pap.* (2004). at <http://biostats.bepress.com/jhubiostat/paper1>
35. Hochreiter, S., Clevert, D.-A. & Obermayer, K. A new summarization method for affymetrix probe level data. *Bioinformatics* **22,** 943–949 (2006).
36. Fasold, M., Stadler, P. F. & Binder, H. G-stack modulated probe intensities on expression arrays - sequence corrections and signal calibration. *BMC Bioinformatics* **11,** 207 (2010).
37. Mueckstein, U., Leparc, G. G., Posekany, A., Hofacker, I. & Kreil, D. P. Hybridization thermodynamics of NimbleGen Microarrays. *BMC Bioinformatics* **11,** 35 (2010).
38. Sykacek, P. *et al.* The impact of quantitative optimization of hybridization conditions on gene expression analysis. *BMC Bioinformatics* **12,** 73 (2011).
39. David, M., Dzamba, M., Lister, D., Ilie, L. & Brudno, M. SHRiMP2: Sensitive yet Practical Short Read Mapping. *Bioinformatics* **27,** 1011–1012 (2011).

40. Glaus, P., Honkela, A. & Rattray, M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* **28,** 1721–1728 (2012).
41. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33,** e175–e175 (2005).
42. Rasmussen, C. E. Gaussian processes for machine learning. in (MIT Press, 2006).
43. Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* (2012). doi:10.1093/bib/bbs046
44. 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).

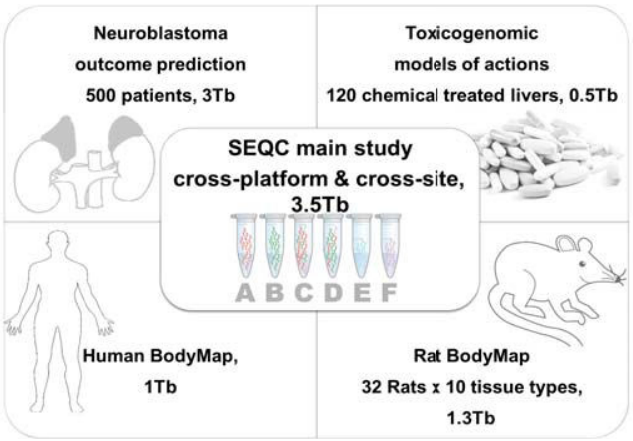
No.	Note	Full_Name	Organization	Address	E-mail	Last_Name	First_Name
1	**1st**	Zhenqiang Su	FDA/NCTR	3900 NCTR Road, Jefferson, AR	zhenqiang.su@fda.hh	Su	Zhenqiang
2	**1st**	Paweł P Łabaj	Boku University Vienna	Chair of Bioinformatics, Boku Uni	Pawel.Labaj@boku.a	Łabaj	Paweł P
3	**1st**	Sheng Li	Cornell University	Department of Physiology, Biophy	shl2018@med.cornell	Li	Sheng
4	1st if possible	Jean Thierry-Mieg	NIH/NCBI	Bldg 38A/Room 8S808, 8600 Roc	mieg@ncbi.nlm.nih.g	Thierry-Mieg	Jean
5	1st if possible	Danielle Thierry-Mie	NIH/NCBI	Bldg 38A/Room 8S808, 8600 Roc	dthierre@ncbi.nlm.nih	Thierry-Mieg	Danielle
6	1st if possible	Wei Shi	The Walter and Eliza Hall Institute of Medical Res	Bioinformatics Division, The Walte	shi@wehi.edu.au	Shi	Wei
7		Charles Wang	City of Hope National Medical Center	Functional Genomics Core, Beckr	cwang@coh.org	Wang	Charles
8		Gary P Schroth	Illumina Inc.	25861 Industrial Boulevard, Hayw	gschroth@illumina.co	Schroth	Gary P
9		Robert A Setterquist	Life Technologies Corporation	2130 Woodward Street, Austin, Te	bob.setterquist@lifete	Setterquist	Robert A
10		John F Thompson	Nabsys Inc.	60 Clifford Street, Providence, RI	thompson.john.f@gm	Thompson	John F
11		Wendell D Jones	Expression Analysis Inc.	4324 South Alston Avenue Durha	wjones@expressiona	Jones	Wendell D
12		Wenzhong Xiao	Harvard Medical School	Massachusetts General Hospital,	wxiao1@partners.org	Xiao	Wenzhong
13		Weihong Xu	Stanford University	Stanford Genome Technology Ce	weihongx@stanford.e	Xu	Weihong
14		Roderick V Jensen	Virginia Tech	Washington Street, MC 0910, Virg	rvjensen@vt.edu	Jensen	Roderick V
15		Reagan Kelly	FDA/NCTR	3900 NCTR Road, Jefferson, AR	reagan.kelly@fda.hhs	Kelly	Reagan
16		Joshua Xu	FDA/NCTR	3900 NCTR Road, Jefferson, AR	zhihua.xu@fda.hhs.g	Xu	Joshua
17		Ana Conesa	Centro de Investigación Príncipe Felipe (CIPF)	Computational Genomics Progran	aconesa@cipf.es	Conesa	Ana
18		Cesare Furlanello	Fondazione Bruno Kessler (FBK)	via Sommarive 18, 38123 - Trento	furlan@fbk.eu	Furlanello	Cesare
19		Hanlin Gao	City of Hope National Medical Center	DNA Sequencing/Solexa Core, Be	hgao@coh.org	Gao	Hanlin
20		Huixiao Hong	FDA/NCTR	3900 NCTR Road, Jefferson, AR	huixiao.hong@fda.hh	Hong	Huixiao
21		Nadereh Jafari	Northwestern University	Center for Genetic Medicine, Fein	n-jafari@northwesterr	Jafari	Nadereh
22		Stan Letovsky	SynapDx Corporation	4 Hartwell Place, Lexington, MA 0	sletovsky@synapdx.c	Letovsky	Stan
23		Yang Liao	The Walter and Eliza Hall Institute of Medical Res	Bioinformatics Division, The Walte	liao@wehi.edu.au	Liao	Yang
24		Fei Lu	SeqWright Inc.	2575 W. Bellfort #2001, Houston,	feilu@seqwright.com	Lu	Fei
25		Edward J Oakeley	Novartis Institutes for Biomedical Research (NIBF	Novartis Institutes for Biomedical	edward.oakeley@nov	Oakeley	Edward J
26		Zhiyu Peng	BGI-Guangzhou (and BGI-Shenzhen)	Main Building, Bei Shan Industrial	pengzhiyu@genomics	Peng	Zhiyu
27		Craig A Praul	The Pennsylvania State University	The Pennsylvania State University	cap142@psu.edu	Praul	Craig A
28		Javier Santoyo-Lope	Andalusian Human Genome Sequencing Center	Medical Genome Project (MGP), I	javier.santoyo@junta	Santoyo-Lop	Javier
29		Andreas Scherer	Australian Genome Research Facility Ltd. and Sp	The Walter and Eliza Hall Institute	Andreas.Scherer@sp	Scherer	Andreas
30		Tieliu Shi	East China Normal University	Center for Bioinformatics and Con	tieliushi@yahoo.com	Shi	Tieliu
31		Gordon K Smyth	The Walter and Eliza Hall Institute of Medical Res	Bioinformatics Division, The Walte	smyth@wehi.edu.au	Smyth	Gordon K
32		Frank Staedtler	Novartis Pharma AG	Novartis Institutes for Biomedical	frank.staedtler@nova	Staedtler	Frank
33		Peter Sykacek	Boku University Vienna	Chair of Bioinformatics, Boku Uni	Peter.Sykacek@boku	Sykacek	Peter
34		Xin-Xing Tan	SeqWright Inc.	2575 W. Bellfort #2001, Houston,	xxtan@seqwright.com	Tan	Xin-Xing
35		E Aubrey Thompson	Mayo Clinic	Department of Cancer Biology, Ma	Thompson.Aubrey@m	Thompson	E Aubrey
36		Jo Vandesompele	Biogazelle NV	Technologiepark 3, B-9052 Zwijna	jo.vandesompele@bio	Vandesomp	Jo
37		May D Wang	GeorgiaTech and Emory University	Department of Biomedical Engine	maywang@bme.gate	Wang	May D
38		Jian Wang	Eli Lilly and Company	Research Informatics, Lilly Corpor	jian.wang@lilly.com	Wang	Jian
39		Russell D Wolfinger	SAS Institute Inc.	SAS Campus Drive, Cary, NC 275	russ.wolfinger@sas.c	Wolfinger	Russell D
40		Jiri Zavadil	New York University Langone Medical Center	Department of Pathology, NYU Ge	zavadilj@iarc.fr	Zavadil	Jiri
41		Scott S Auerbach	NIH/NIEHS	National Institute of Environmenta	auerbachs@niehs.nih	Auerbach	Scott S

42	Wenjun Bao	SAS Institute Inc.	SAS Campus Drive, Cary, NC 27	wenjun.bao@sas.com	Bao	Wenjun
43	Hans Binder	University of Leipzig	Interdisciplinary Centre for Bioinfo	binder@izbi.uni-leipzi	Binder	Hans
44	Thomas Blomquist	University of Toledo Health Sciences Campus	Division of Pulmonary and Critical	Thomas.Blomquist@r	Blomquist	Thomas
45	Murray H Brilliant	Marshfield Clinic Research Foundation	Center of Human Genetics, 1000	BRILLIANT.MURRAY	Brilliant	Murray H
46	Pierre R Bushel	NIH/NIEHS	National Institute of Environmenta	bushel@niehs.nih.go	Bushel	Pierre R
47	Weimin Cai	Fudan University	School of Pharmacy, 826 Zhangh	weimincai@hotmail.c	Cai	Weimin
48	Jennifer G Catalano	FDA/CBER	Office of Cellular, Tissue, and Ge	jennifer.catalano@fda	Catalano	Jennifer G
49	Ching-Wei Chang	FDA/NCTR	3900 NCTR Road, Jefferson, AR	ching-wei.chang@fda	Chang	Ching-Wei
50	Tao Chen	FDA/NCTR	3900 NCTR Road, Jefferson, AR	tao.chen@fda.hhs.go	Chen	Tao
51	Geng Chen	East China Normal University	Center for Bioinformatics and Con	chengeng66666@gm	Chen	Geng
52	Marco Chierici	Fondazione Bruno Kessler (FBK)	via Sommarive 18, 38123 - Trento	chierici@fbk.eu	Chierici	Marco
53	Tzu-Ming Chu	SAS Institute Inc.	SAS Campus Drive, Cary, NC 27	tzu-ming.chu@sas.co	Chu	Tzu-Ming
54	Djork-Arné Clevert	Johannes Kepler University Linz	Johannes Kepler University Linz,	okko.clevert@gmail.c	Clevert	Djork-Arné
55	Youping Deng	Rush University Cancer Center	Department of Internal Medicine,	youping_deng@rush.	Deng	Youping
56	Adnan Derti	Novartis	Novartis Institutes for Biomedical	adnan.derti@novartis	Derti	Adnan
57	Viswanath Devanara	AbbVie Inc.	Global Pharmaceutical R&D, 32 K	viswanath.devanaray	Devanaraya	Viswanath
58	Zirui Dong	BGI-Shenzhen	Main Building, Bei Shan Industrial	dongzirui@genomics.	Dong	Zirui
59	Joaquin Dopazo	Centro de Investigación Príncipe Felipe (CIPF)	Department of Computational Ger	jdopazo@cipf.es	Dopazo	Joaquin
60	Tingting Du	Fudan University	Center for Pharmacogenomics, S	tingtingdu00@126.co	Du	Tingting
61	Hong Fang	FDA/NCTR	3900 NCTR Road, Jefferson, AR	hong.fang@fda.hhs.g	Fang	Hong
62	Yongxiang Fang	University of Liverpool	Centre for Genomic Research, Th	fangy@liv.ac.uk	Fang	Yongxiang
63	Mario Fasold	University of Leipzig	Interdisciplinary Centre for Bioinfo	mario@bioinf.uni-leip	Fasold	Mario
64	Anita Fernandez	Novartis Pharma AG	Novartis Institutes for Biomedical	anita.fernandez@nov	Fernandez	Anita
65	Matthias Fischer	University of Cologne	Department of Pediatric Oncology	matthias.fischer@uk-l	Fischer	Matthias
66	Pedro Furió-Tarí	Centro de Investigación Príncipe Felipe (CIPF)	Computational Genomics Progran	pfurio@cipf.es	Furió-Tarí	Pedro
67	James C Fuscoe	FDA/NCTR	3900 NCTR Road, Jefferson, AR	james.fuscoe@fda.hh	Fuscoe	James C
68	Stan Gaj	Maastricht University	Department of Toxicogenomics, P	s.gaj@maastrichtuniv	Gaj	Stan
69	Jorge Gandara	Cornell University	Department of Physiology, Biophy	jog2033@med.cornel	Gandara	Jorge
70	Huan Gao	BGI-Shenzhen	Main Building, Bei Shan Industrial	gaohuan@genomics.	Gao	Huan
71	Weigong Ge	FDA/NCTR	3900 NCTR Road, Jefferson, AR	weigong.ge@fda.hhs.	Ge	Weigong
72	Yoichi Gondo	RIKEN BioResource Center	3-1-1 Koyadai, Tsukuba, Ibarak, 3	gondo@brc.riken.jp	Gondo	Yoichi
73	Binsheng Gong	FDA/NCTR	3900 NCTR Road, Jefferson, AR	Binsheng.Gong@fda.	Gong	Binsheng
74	Meihua Gong	BGI-Shenzhen	Main Building, Bei Shan Industrial	gongmeihua@genom	Gong	Meihua
75	Zhuolin Gong	BGI-Shenzhen	Main Building, Bei Shan Industrial	gongzhuolin@genomi	Gong	Zhuolin
76	Bridgett Green	FDA/NCTR	3900 NCTR Road, Jefferson, AR	bridgett.green@fda.hl	Green	Bridgett
77	Chao Guo	City of Hope National Medical Center	Functional Genomics Core, Beckr	chguo@coh.org	Guo	Chao
78	Lei Guo	FDA/NCTR	3900 NCTR Road, Jefferson, AR	lei.guo@fda.hhs.gov	Guo	Lei
79	Li-Wu Guo	FDA/NCTR	3900 NCTR Road, Jefferson, AR	lwguo@yahoo.com	Guo	Li-wu
80	James Hadfield	University of Cambridge	Cancer Research UK Cambridge	james.hadfield@canc	Hadfield	James
81	Jan Hellemans	Biogazelle NV	Technologiepark 3, B-9052 Zwijna	Jan.Hellemans@Biog	Hellemans	Jan
82	Sepp Hochreiter	Johannes Kepler University Linz	Johannes Kepler University Linz,	hochreit@bioinf.jku.at	Hochreiter	Sepp
83	Meiwen Jia	Fudan University	Center for Pharmacogenomics, S	jmwcathy@gmail.com	Jia	Meiwen

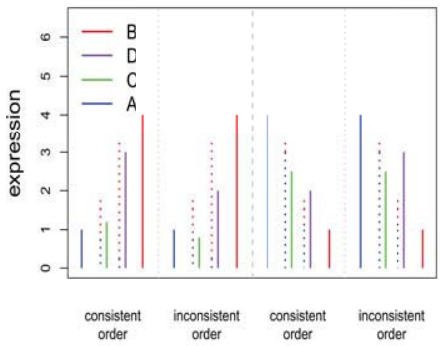
84	Min Jian	BGI-Shenzhen	Main Building, Bei Shan Industrial	jianm@genomics.org.	Jian	Min
85	Charles D Johnson	Texas A&M AgriLife Research	2123 TAMU, College Station TX 7	charlie@ag.tamu.edu	Johnson	Charles D
86	Suzanne Kay	University of Liverpool	Centre for Genomic Research, Th	skay@liv.ac.uk	Kay	Suzanne
87	Jos Kleinjans	Maastricht University	Department of Toxicogenomics, P	j.kleinjans@maastricht	Kleinjans	Jos
88	Samir Lababidi	FDA/CBER	WOC1 RM400S, HFM-210, 1401	samir.lababidi@fda.hl	Lababidi	Samir
89	Shawn Levy	HudsonAlpha Institute for Biotechnology	601 Genome Way, Huntsville, AL	slevy@hudsonalpha.c	Levy	Shawn
90	Quan-Zhen Li	University of Texas Southwestern Medical Center	6000 Harry Hines Boulevard/ND6.	quan.li@utsouthweste	Li	Quan-Zhen
91	Li Li	SAS Institute Inc.	SAS Campus Drive, Cary, NC 275	lili.li@sas.com	Li	Li
92	Peng Li	East China Normal University	Center for Bioinformatics and Con	ebio@foxmail.com	Li	Peng
93	Yan Li	FDA/NCTR	3900 NCTR Road, Jefferson, AR	yan.li@fda.hhs.gov	Li	Yan
94	Haiqing Li	City of Hope National Medical Center	Bioinformatics Core, Department	hali@coh.org	Li	Haiqing
95	Jianying Li	NIH/NIEHS; Kelly Government Solutions	National Institute of Environmenta	jianying.li@gmail.com	Li	Jianying
96	Shiyong Li	BGI-Shenzhen	Main Building, Bei Shan Industrial	lishiyong@genonics.c	Li	Shiyong
97	Simon M Lin	Marshfield Clinic Research Foundation	Biomedical Informatics Research	LINMD.SIMON@mcrf	Lin	Simon M
98	Francisco J López	Andalusian Human Genome Sequencing Center	Medical Genome Project (MGP), I	fjavier@bioinfomgp.or	López	Francisco J
99	Xin Lu	AbbVie Inc.	Global Pharmaceutical R&D, 1 Nc	xin.x.lu@abbott.com	Lu	Xin
100	Heng Luo	University of Arkansas at Little Rock	UALR/UAMS Joint Bioinformatics	hengluo88@gmail.co	Luo	Heng
101	Xiwen Ma	Eli Lilly and Company	Discovery Statistics, Lilly Corporat	ma_xiwen@lilly.com	Ma	Xiwen
102	Joseph Meehan	FDA/NCTR	3900 NCTR Road, Jefferson, AR	joe.meehan@fda.hhs	Meehan	Joseph
103	Dalila B Megherbi	University of Massachusetts at Lowell	CMINDS Research Center, Roon	dalila_megherbi@uml	Megherbi	Dalila B
104	Nan Mei	FDA/NCTR	3900 NCTR Road, Jefferson, AR	nan.mei@fda.hhs.gov	Mei	Nan
105	Bing Mu	City of Hope National Medical Center	Bioinformatics Core, Beckman Re	bmu@coh.org	Mu	Bing
106	Baitang Ning	FDA/NCTR	3900 NCTR Road, Jefferson, AR	baitang.ning@fda.hhs	Ning	Baitang
107	Akhilesh Pandey	Johns Hopkins University	McKusick-Nathans Institute of Ge	pandey@jhmi.edu	Pandey	Akhilesh
108	Javier Pérez-Florido	Andalusian Human Genome Sequencing Center	Medical Genome Project (MGP), I	jpflorido@bioinfomgp.	Pérez-Florid	Javier
109	Roger G Perkins	FDA/NCTR	3900 NCTR Road, Jefferson, AR	roger.perkins@fda.hh	Perkins	Roger G
110	Ryan Peters	Partek Inc.	624 Trade Center Boulevard, Sair	rpeters@partek.com	Peters	Ryan
111	John H Phan	GeorgiaTech and Emory University	Department of Biomedical Engine	jhphan@gatech.edu	Phan	John H
112	Mehdi Pirooznia	Johns Hopkins University	School of Medicine, Department o	pirooznia@gmail.com	Pirooznia	Mehdi
113	Feng Qian	FDA/NCTR	3900 NCTR Road, Jefferson, AR	feng.qian@fda.hhs.gc	Qian	Feng
114	Tao Qing	Fudan University	Center for Pharmacogenomics, S	qingtao1083@yahoo.(	Qing	Tao
115	Lucille Rainbow	University of Liverpool	Centre for Genomic Research, Th	lrainbow@liv.ac.uk	Rainbow	Lucille
116	Philippe Rocca-Serra	University of Oxford	Oxford e-Research Centre, Unive	proccaserra@gmail.c	Rocca-Serra	Philippe
117	Laure Sambourg	UJF-Grenoble 1 / CNRS / TIMC-IMAG UMR 552	UJF-Grenoble 1 / CNRS / TIMC-I	laure.sambourg@ima	Sambourg	Laure
118	Susanna-Assunta Sa	University of Oxford	Oxford e-Research Centre, Unive	sa.sansone@gmail.c	Sansone	Susanna-Assunta
119	Scott Schwartz	Texas A&M AgriLife Research	2123 TAMU, College Station TX 7	sschwartz@ag.tamu.e	Schwartz	Scott
120	Ruchir Shah	SRA International Inc.	2605 Meridian Parkway, Durham,	ruchir_shah@sra.com	Shah	Ruchir
121	Jie Shen	FDA/NCTR	3900 NCTR Road, Jefferson, AR	jie.shen@fda.hhs.gov	Shen	Jie
122	Todd M Smith	Geospiza Inc.	100 West Harrison, NT 330, Seatt	Todd.Smith@PERKIN	Smith	Todd M
123	Oliver Stegle	EMBL – Eureopan Bioinformatics Institute	EMBL Outstation - Hinxton, Europ	oliver.stegle@ebi.ac.u	Stegle	Oliver
124	Nancy Stralis-Paves	Boku University Vienna	Chair of Bioinformatics, Boku Uni	Nancy.Stralis@boku.a	Stralis-Pave	Nancy
125	Elia Stupka	San Raffaele Scientific Institute	Center for Translational Genomics	stupka.elia@gmail.co	Stupka	Elia

126		Yutaka Suzuki	The University of Tokyo	Department of Medical Genome S	ysuzuki@k.u-tokyo.ac	Suzuki	Yutaka
127		Lee T Szkotnicki	SeqWright Inc.	2575 W. Bellfort #2001, Houston,	lszkotnicki@seqwrigh	Szkotnicki	Lee T
128		Matthew Tinning	Australian Genome Research Facility Ltd.	The Walter and Eliza Hall Institute	Matthew.Tinning@AC	Tinning	Matthew
129		Bimeng Tu	BGI-Shenzhen	Main Building, Bei Shan Industrial	tubimeng@genomics.	Tu	Bimeng
130		Joost van Delft	Maastricht University	Department of Toxicogenomics, P	j.vandelft@maastricht	van Delft	Joost
131		Alicia Vela	Andalusian Human Genome Sequencing Center	Medical Genome Project (MGP), I	alicia.vela.boza@junt	Vela	Alicia
132		Elisa Venturini	NYU Langone Medical Center	Genome Technology Center, NYL	Elisa.Venturini@nyun	Venturini	Elisa
133		Stephen J Walker	Wake Forest University	Wake Forest Institute for Regener	swalker@wakehealth.	Walker	Stephen J
134		Liqing Wan	FDA/NCTR	3900 NCTR Road, Jefferson, AR	wanlq1114@gmail.co	Wan	Liqing
135		Wei Wang	Princeton University	Lewis-Sigler Institute for Integrativ	ww3@princeton.edu	Wang	Wei
136		Jinhui Wang	City of Hope National Medical Center	DNA Sequencing/Solexa Core, Be	jiwang@coh.org	Wang	Jinhui
137		Jun Wang	BGI-Shenzhen (and three other affiliations)	BGI-Shenzhen, Shenzhen, China;	Department of Biolog	Wang	Jun
138		Eric D Wieben	Mayo Clinic	Department of Biochemistry and M	Wieben.eric@mayo.e	Wieben	Eric D
139		James C Willey	University of Toledo Health Sciences Campus	Division of Pulmonary and Critical	james.willey2@utoled	Willey	James C
140		Po-Yen Wu	Georgia Institute of Technology	School of Electrical and Computer	leoboki@gmail.com; p	Wu	Po-Yen
141		Jiekun Xuan	FDA/NCTR	3900 NCTR Road, Jefferson, AR	j.xuan@yahoo.com	Xuan	Jiekun
142		Yong Yang	Eli Lilly and Company	Research Informatics, Lilly Corpor	yang_yong_yy@lilly.c	Yang	Yong
143		Zhan Ye	Marshfield Clinic Research Foundation	Biomedical Informatics Research	ye.zhan@mcrf.mfldcli	Ye	Zhan
144		Ye Yin	BGI-Shenzhen	Main Building, Bei Shan Industrial	yinye@genomics.org.	Yin	Ye
145		Ying Yu	Fudan University	Center for Pharmacogenomics, Sc	11111030027@fudan	Yu	Ying
146		Yate-Ching Yuan	City of Hope National Medical Center	Department of Molecular Medicine	yyuan@coh.org	Yuan	Yate-Ching
147		John Zhang	Systems Analytics Inc.	946 Great Plain Avenue, #125, Ne	johnz@systemsanaly	Zhang	John
148		Ke K Zhang	University of North Dakota School of Medicine	Department of Pathology, 501 N. (	ke.zhang@med.und.e	Zhang	Ke K
149		Wenqian Zhang	BGI-Shenzhen	Main Building, Bei Shan Industrial	zhangwenqian56@gn	Zhang	Wenqian
150		Wenwei Zhang	BGI-Shenzhen	Main Building, Bei Shan Industrial	zhangww@genomics.	Zhang	Wenwei
151		Yanyan Zhang	BGI-Shenzhen	Main Building, Bei Shan Industrial	zhangyanyan@genon	Zhang	Yanyan
152		Chen Zhao	East China Normal University	Center for Bioinformatics and Con	zhaochen_ny@yahoo	Zhao	Chen
153		Yiming Zhou	Digomics LLC	7 Franklin St Unit 2, Brookline, MA	zym@digomics.com;	Zhou	Yiming
154		Paul Zumbo	Cornell University	Department of Physiology, Biophy	paz2005@med.corne	Zumbo	Paul
155		Weida Tong	FDA/NCTR	3900 NCTR Road, Jefferson, AR	weida.tong@fda.hhs.g	Tong	Weida
156	**corresp.**	David P Kreil	Boku University Vienna & University of Warwick	Chair of Bioinformatics, Boku Univ	David.Kreil@boku.ac.	Kreil	David P
157	**corresp.**	Christopher E Mason	Cornell University	Department of Physiology, Biophy	chm2042@med.corne	Mason	Christopher E
158	**corresp.**	Leming Shi	FDA/NCTR and Fudan University	3900 NCTR Road, Jefferson, AR	leming.shi@gmail.com	Shi	Leming

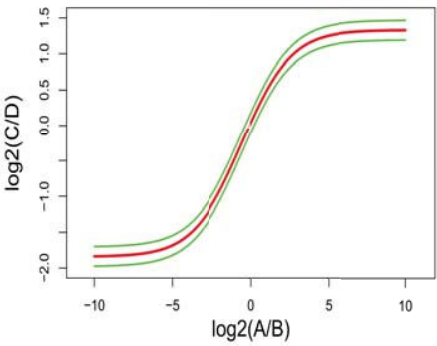
a



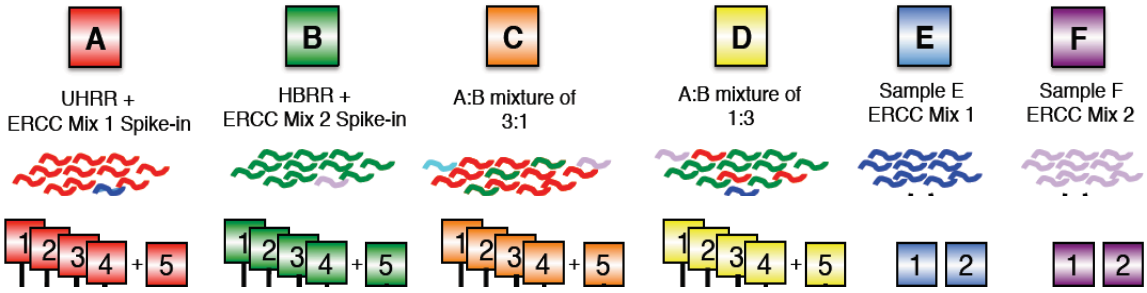
c



d



b





Illumina sites

- ***BGI***
- ***Cornell***
- ***Mayo Clinic***
- **City of Hope**
- **Novartis**
- **Australian Genome Research Facility**



library prep

barcode

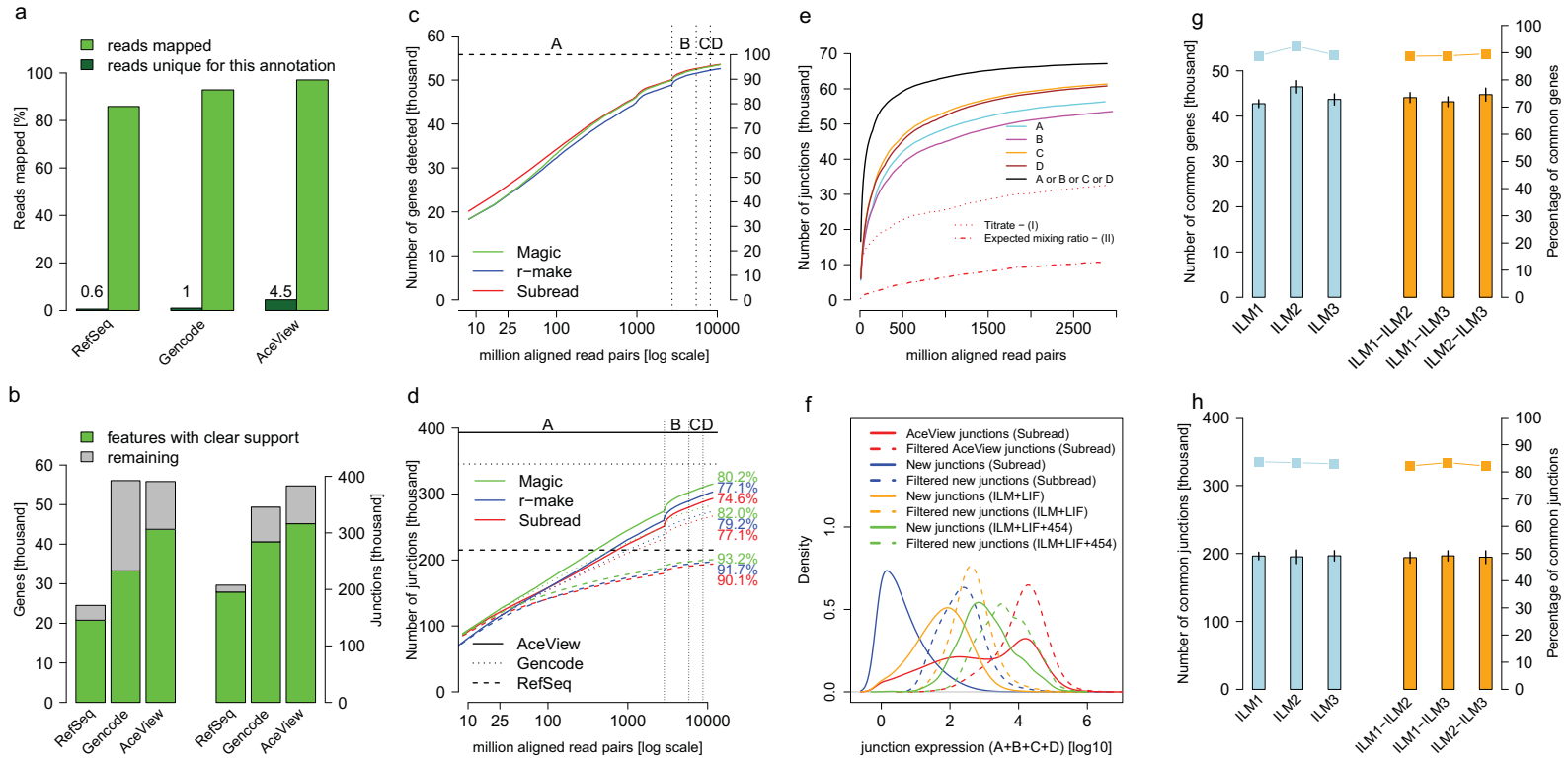


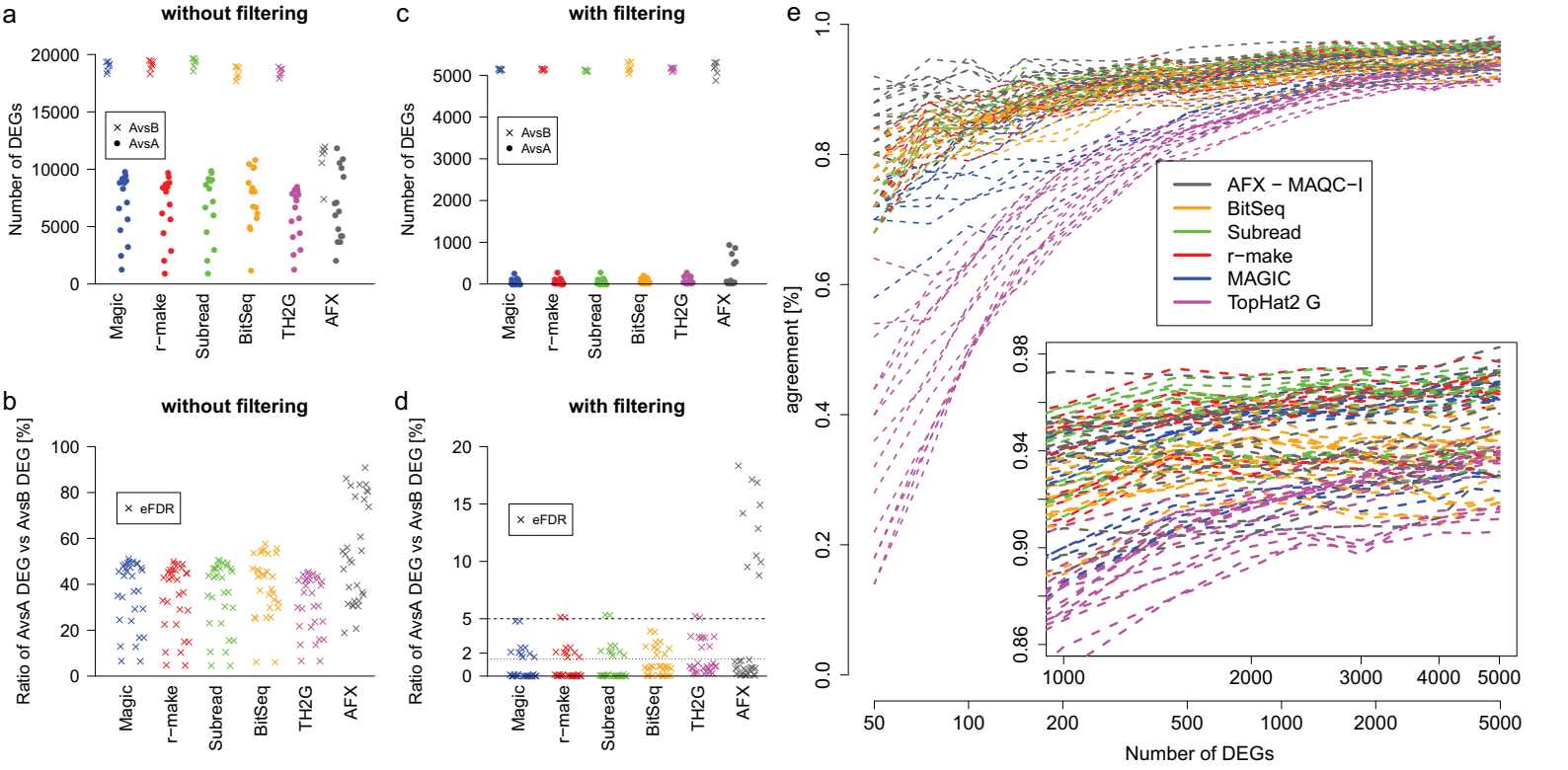
Life Technologies sites

- ***Northwestern***
- ***Penn State***
- ***SeqWright***
- **Liverpool**

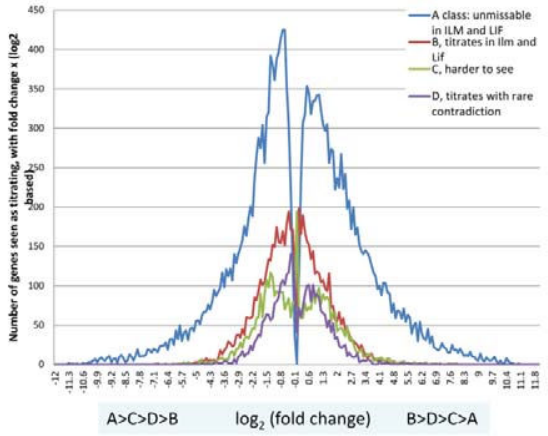


At each site: Pool & sequence site libraries (4x A..D + 2x E,F) + vendor library #5 (A..D)

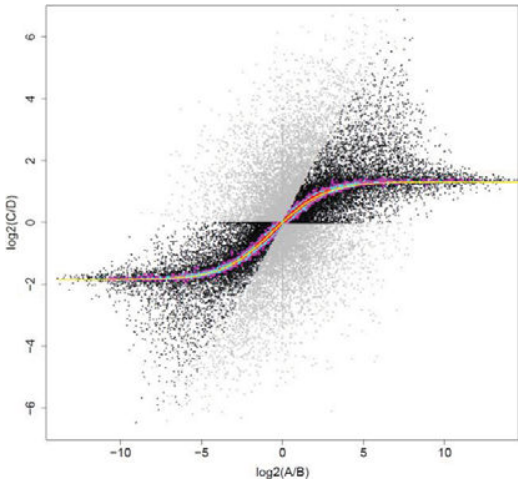

a
reads mapped
reads unique for this annotation
Reads mapped [%]
0.6
1
4.5
RefSeq
Gencode
AceView
b
features with clear support
remaining
Genes [thousand]
Junctions [thousand]
RefSeq
Gencode
AceView
RefSeq
Gencode
AceView
c
A
B
CD
Number of genes detected [thousand]
Magic
r−make
Subread
million aligned read pairs [log scale]
d
A
B
CD
Number of junctions [thousand]
Magic
r−make
Subread
AceView
Gencode
RefSeq
80.2%
77.1%
74.6%
82.0%
79.2%
77.1%
93.2%
91.7%
90.1%
million aligned read pairs [log scale]
e
Number of junctions [thousand]
A
B
C
D
A or B or C or D
Titrate − (I)
Expected mixing ratio − (II)
million aligned read pairs
f
AceView junctions (Subread)
Filtered AceView junctions (Subread)
New junctions (Subread)
Filtered new junctions (Subbread)
New junctions (ILM+LIF)
Filtered new junctions (ILM+LIF)
New junctions (ILM+LIF+454)
Filtered new junctions (ILM+LIF+454)
Density
junction expression (A+B+C+D) [log10]
g
Number of common genes [thousand]
Percentage of common genes
ILM1
ILM2
ILM3
ILM1−ILM2
ILM1−ILM3
ILM2−ILM3
h
Number of common junctions [thousand]
Percentage of common junctions
ILM1
ILM2
ILM3
ILM1−ILM2
ILM1−ILM3
ILM2−ILM3


a
without filtering
Number of DEGs
AvsB
AvsA
Magic
r−make
Subread
BitSeq
TH2G
AFX
c
with filtering
b
Ratio of AvsA DEG vs AvsB DEG [%]
eFDR
d
e
agreement [%]
AFX − MAQC−I
BitSeq
Subread
r−make
MAGIC
TopHat2 G
Number of DEGs

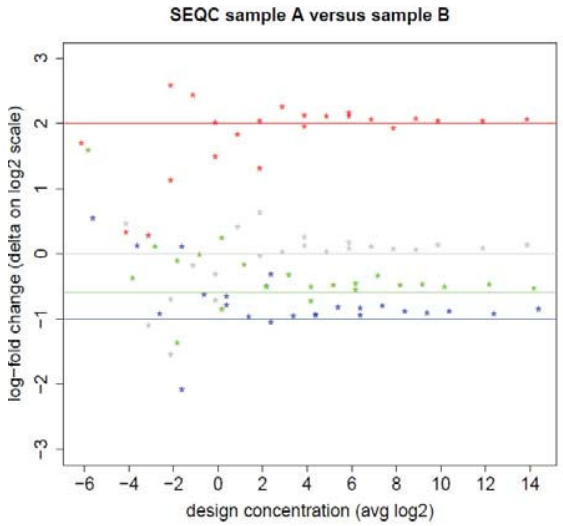
a


A class: unmissable in ILM and LIF
B, titrates in Ilm and Lif
C, harder to see
D, titrates with rare contradiction
Number of genes seen as titrating, with fold change x (log2 based)
A>C>D>B
log2 (fold change)
B>D>C>A

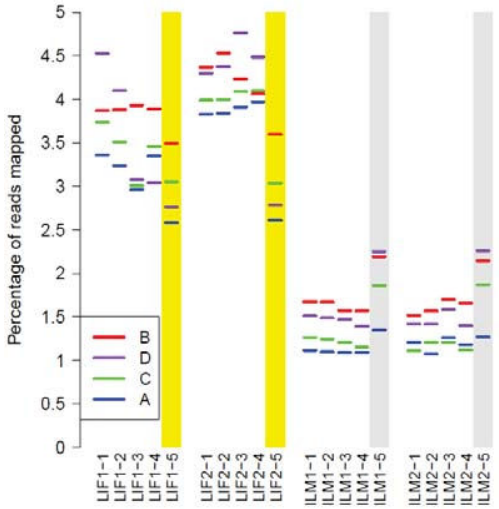
b

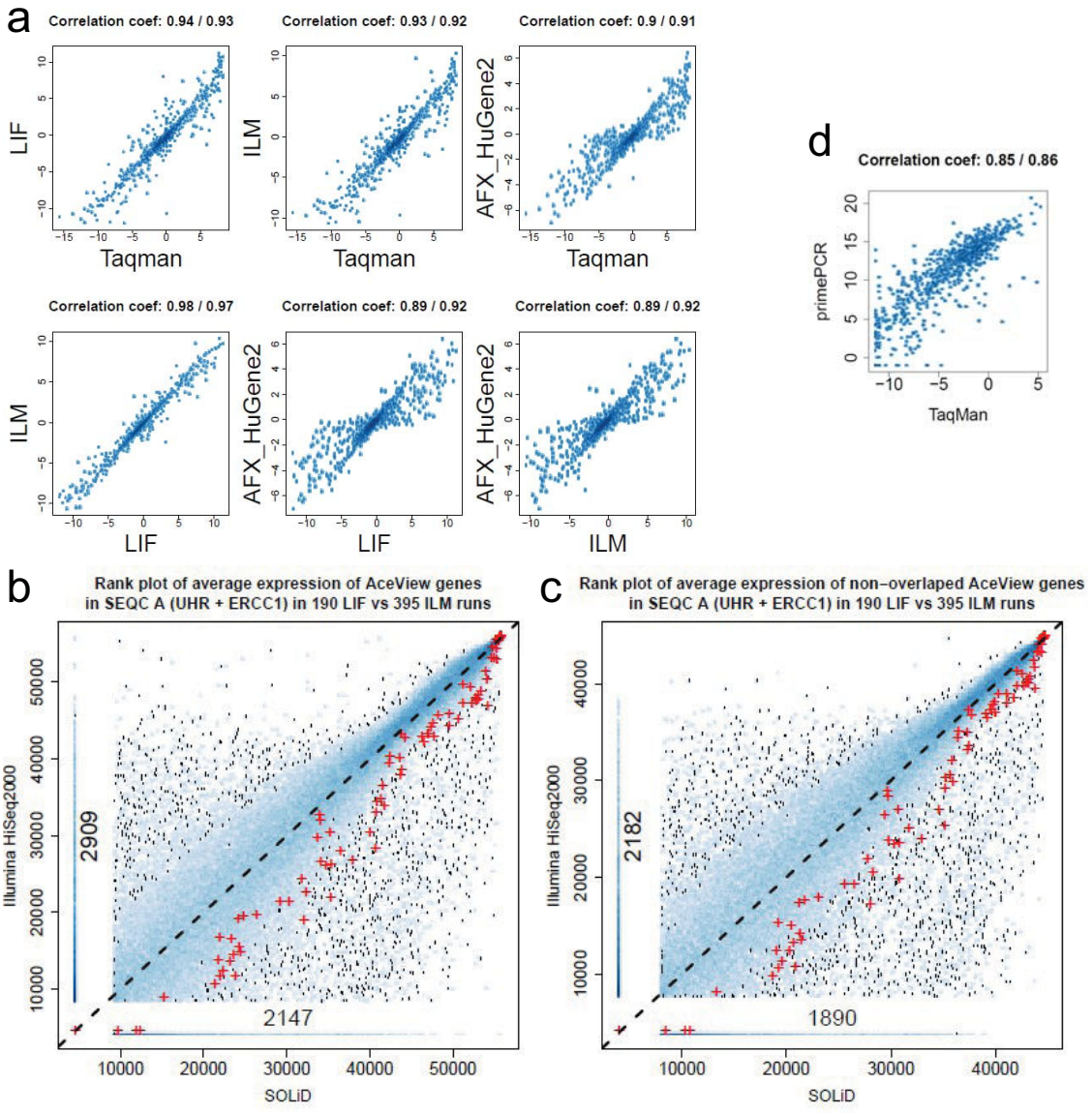

log2(C/D)
log2(A/B)

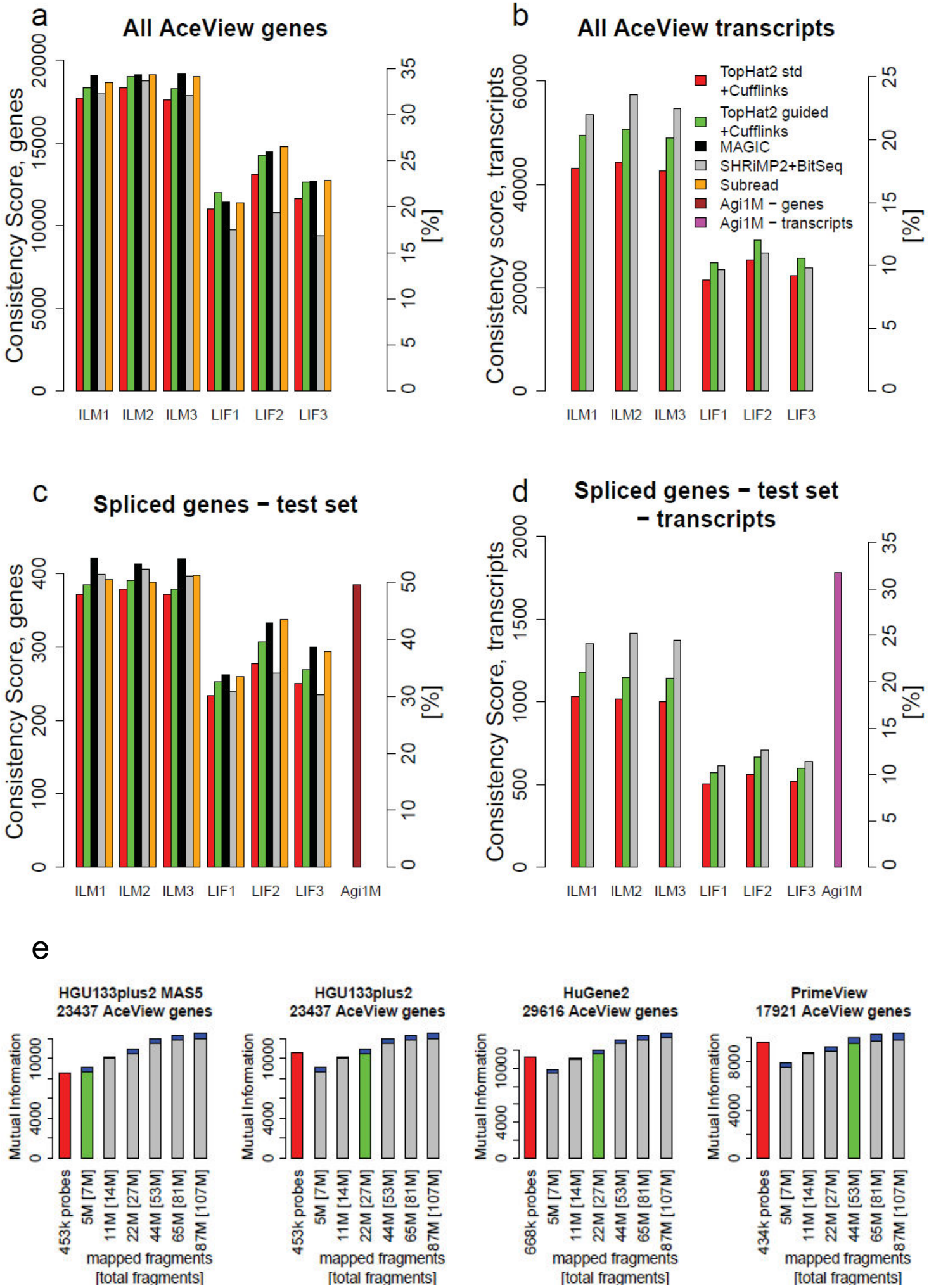
c


SEQC sample A versus sample B
log-fold change (delta on log2 scale)
design concentration (avg log2)

d


Percentage of reads mapped
B
D
C
A
LIF1-1
LIF1-2
LIF1-3
LIF1-4
LIF1-5
LIF2-1
LIF2-2
LIF2-3
LIF2-4
LIF2-5
ILM1-1
ILM1-2
ILM1-3
ILM1-4
ILM1-5
ILM2-1
ILM2-2
ILM2-3
ILM2-4
ILM2-5


a
Correlation coef: 0.94 / 0.93
Correlation coef: 0.93 / 0.92
Correlation coef: 0.9 / 0.91
LIF
ILM
AFX_HuGene2
Taqman
Correlation coef: 0.98 / 0.97
Correlation coef: 0.89 / 0.92
Correlation coef: 0.89 / 0.92
ILM
AFX_HuGene2
AFX_HuGene2
LIF
LIF
ILM
d
Correlation coef: 0.85 / 0.86
primePCR
TaqMan
b
Rank plot of average expression of AceView genes
in SEQC A (UHR + ERCC1) in 190 LIF vs 395 ILM runs
Illumina HiSeq2000
SOLiD
2909
2147
c
Rank plot of average expression of non−overlaped AceView genes
in SEQC A (UHR + ERCC1) in 190 LIF vs 395 ILM runs
Illumina HiSeq2000
SOLiD
2182
1890


a
All AceView genes
Consistency Score, genes
[%]
ILM1 ILM2 ILM3 LIF1 LIF2 LIF3
b
All AceView transcripts
Consistency score, transcripts
TopHat2 std +Cufflinks
TopHat2 guided +Cufflinks
MAGIC
SHRiMP2+BitSeq
Subread
Agi1M − genes
Agi1M − transcripts
c
Spliced genes − test set
Agi1M
d
Spliced genes − test set − transcripts
e
HGU133plus2 MAS5 23437 AceView genes
HGU133plus2 23437 AceView genes
HuGene2 29616 AceView genes
PrimeView 17921 AceView genes
Mutual Information
453k probes
668k probes
434k probes
5M [7M]
11M [14M]
22M [27M]
44M [53M]
65M [81M]
87M [107M]
mapped fragments
[total fragments]

Gènes significativement mutés chez 520 patientes atteintes du cancer de l'ovaire

TAB. 5.1 – **Intersection entre les gènes LoF des données TCGA et les gènes connus pour être impliqués dans le cancer.** Nous avons cherché parmi les 2 144 gènes LoF ceux présents également dans d'autres listes de gènes impliqués dans le cancer. La colonne "Gènes" contient tous les gènes de notre liste de candidats apparaissant au moins une fois dans une autre liste. La ligne Int. (intersection) donne le nombre des gènes présents dans notre liste et dans la liste concernée. La P-value (ligne "P") est le résultat du test exact de Fisher comparant les proportions de gènes cancer contenus dans notre liste de candidats et dans la liste "background". Pour une description des listes de gènes, voir section 4.7.

	Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
Nb gènes	3214	40	124	2273	125	184
Int.		12	32	503	31	62
P-valeur		0.000360	3.22e-07	2.94e-81	1.24e-06	1.07e-18
	TRIP6			X		
	ACTN4			X		
	NCLandSNORA75			X		
	PPARD			X		
	SLC35A4			X		
	PTGER3			X		
	MMP2			X		
	ERBB4			X		
	VPS33AandDIABLO			X		
	PINK1			X		
	UGCG			X		
	CD34			X		
	ITGA5			X		
	DOCK1			X		
	C17orf69andCRHR1			X		
	GNB2L1andSNORD95andSNORD96A			X		
	NBN			X		
	AKR1E2andAKR1C1andAKR1C3			X		
	PTTG1IP			X		
	EIF2AK2			X		
	TGFB2			X		
	KDR			X		
	RGS5			X		
	COL4A2			X		
	EHF			X		
	PACRG			X		
	ABCG2			X		
	CAPN10andGPR35			X		
	MMP7			X		
	AXIN1		X	X		X
	MDC1			X		
	SETD2		X			
	PSD3andRPL35P6			X		

Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
FLT1			X		
PPP2R4			X		
COL4A5			X		
C7orf68andMETTL2B			X		
RB1CC1			X		X
FAM24BandCUZD1			X		
TRIM33					X
TOPBP1			X		
LATS2			X		
CTNNBL1			X		
AES			X		
SKAP1			X		
ITLN1			X		
MET		X	X	X	X
NUP214			X		
TUBB2A			X		
AGR3			X		
NINL			X		
MAP3K3			X		
AXL			X		
PELP1			X		
ATF6BandTNXB			X		
OR7E38PandASNS			X		
F13A1			X		
BIK			X		
PLCG2			X		
NRP1			X		
PLK2			X		
CDH13			X		
USP15			X		
TSC22D1			X		
ACOT13			X		
EPS8			X		
BST2			X		
DUSP6			X		
INPP4B			X		
PTPN12			X	X	X
NAGLUandCOASY			X		
SCGB1A1			X		
EIF6andEDEM2			X		
MAP3K5			X		
RXRA			X		
NASP			X		
SLC23A2			X		
UGDH			X		
A2M			X		
ENPP2andCYCSP23			X		

Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
CEP290			X		
RFWD2			X		
ROCK2			X		
FUBP3			X		
STAG2		X			
ANXA3			X		
MSMBandNCOA4			X		X
ACACA			X		
EPHA2			X		
TEK			X		
CLIC4			X		
INTS6			X		X
SQSTM1			X		
TLE3			X		
KPNA2			X		
PDCD5			X		
SEPT10			X		
KLF8			X		
MACROD1			X		
SERPINA4andSERPINA5andSERPINA3			X		
PRKG1			X		
BIRC6			X		
PRDM2			X		
ANLN			X		
TACC3					X
EPC2			X		
CARD11		X			
RPL22andCHD5			X		
UCHL1			X		
DCAF6			X		
TSSC1			X		
PALB2			X	X	X
AFMIDandBIRC5			X		
PDCD6andAHRR			X		
PIK3CG			X		
SNRPN_			X		
UBE4AandATP5L			X		
SPOP		X			
EGFR		X	X	X	X
RAD23B			X		
VWA5A			X		
TYK2			X		
DICER1			X		X
DPP4			X		
NQO1			X	X	X
NCOR2			X		
SKP1			X		

Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
THBS4			X		
UBE2I			X		
SRP9andEPHX1			X		
ERCC6andPGBD3			X		X
MXI1				X	X
CNTN4			X		
GSTA2			X		
MYO6			X		
USP14			X		
EZH2		X	X		
APOE			X		
MSH6	X	X	X	X	X
NR4A1			X		
MUC16			X		
UBCandSCARB1			X		
MCM7			X		
ATP1B1			X		
ERBB2	X	X	X	X	X
FANCM			X		
SDC1			X		
BTN3A2			X		
ABI1			X		
CHST15			X		
CDK1			X		
KLRC4andKLRK1			X		
EP400			X		
PTK2			X		
IQGAP1			X		
TRDMT1			X		
PAQR8andEFHC1			X		
TNFSF10			X		
SORT1			X		
APLP2andST14			X		
MGAT5			X		
C1QTNF3andAMACR			X		
LATS1			X		
GNAQ		X	X		
CCDC46andAXIN2			X	X	X
SLC2A3			X		
PTHLH			X		
NCRNA00189andGAPDHP14andBACH1			X		X
CLPTM1L			X		
ALCAM			X		
COL18A1			X		
TAF1_			X		
GNL3andSNORD19B			X		
PDZRN3			X		

Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
SLC12A6			X		
NPM1		X	X		
ERBB3andPA2G4			X		
UBE2O			X		
MICB			X		
THRB			X		
AKAP9			X		
ALG11andUTP14C			X		
EDNRA			X		
NR5A2			X		
APBB3andSRA1			X		
IGF1R			X		
RIPK1			X		
ACVR1			X		
ICAM1			X		
RECK			X		
IGK_			X		
RAD18			X		
PCM1			X		X
ALDH1A1			X		
XPNPEP3andRBX1			X		
ATR	X		X		X
ABCC5			X		
HELLS			X		
RBBP7			X		
PTPRG			X		
NCOR1		X	X		
LTBP1			X		
TXNRD1			X		
DPYD			X		
ARHGEF2			X		
CFH			X		
CPEB1			X		
ETS2			X		
QKI			X		
THBS2			X		
KDM6A		X			
NCAM1			X		
USP33			X		
HOXC_			X		
EPHA5			X		
HSP90AB1			X		
DYNC1H1			X		
TRIOBP			X		
ERAP1			X		
ATP7A			X		
SYNE1			X		

Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
MFAP5			X		
IL6R			X		
PLD1			X		
CDK7			X		
LIG1			X		
CHI3L1			X		
JAG1			X		
ALOX15			X		
HSPA4			X		
TNFAIP8L2andSCNM1			X		
SUZ12			X		X
KIF1C			X		
MUTYH			X	X	X
KRT18			X		
NEBL			X		
CARKDandING1			X	X	X
PRKCD			X		
TGS1			X		
TP53INP2			X		
AKR1C2			X		
COL4A3			X		
MCC				X	X
BRD8			X		
BRCA1P1	X	X	X	X	X
XPC			X	X	
MRE11A			X		
FLVCR2andTTLL5andC14orf179			X		
ID2			X		
TMEM189-UBE2V1			X		
ZNF138			X		
WRN			X		
RAB5C			X		
FUBP1		X			
MSH2	X	X	X	X	X
MAP3K4			X		
TIAM1			X		
PKN1			X		
FN1			X		
ABCB1			X		
AHR			X		
PBRM1		X			
DLEC1			X		X
SRPX			X		
POLR2A			X		
KRT7			X		
SETD8			X		
ACVR2A			X		

Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
NR3C1			X		
YWHAZ			X		
POLE					X
IFITM2			X		
HDAC3			X		
FILIP1L			X		
WDR77andOVGP1			X		
TSPAN1			X		
PAXIP1			X		
TLR4			X	X	X
UIMC1			X		
KRT19			X		
SDHC					X
BTN2A2andBTN3A1			X		
CHFRandGOLGA3			X		
FAM175AandHELQ			X		
UBR5			X		
HFE			X		
EIF2AK1			X		
COL4A1			X		
UBQLN1			X		
VCAN			X		
WT1		X	X	X	X
INS-IGF2			X		
ILK			X		
PPP2R5D			X		
NCOA3			X		
CYP19A1			X		
VCAM1			X		
TP53	X	X	X	X	X
ENTPD4andLOXL2			X		
MLH3	X		X	X	X
OAS3			X		
MLL3		X			
CTSB			X		
CTBP2			X		
JUP			X		
C19orf2			X		
ZMYND11			X		
RGL3andEPOR			X		
RAPGEF1			X		
MAP3K8			X		X
RAF1			X		
STC1			X		
FURIN			X		
FGF13			X		
RAB13andJTB			X		

Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
SULF1			X		
SMARCA5			X		
WWC1			X		
CDC23			X		
HAAO			X		
CD248			X		
MAP3K7			X		
BCCIP			X		
IRGQandXRCC1			X		
TMPRSS3andTFF1			X		
MEST			X		
SULT1A1			X		
SPON1			X		
PAWR			X		
RAD51L3andRFFL	X		X	X	X
GTF2I			X		
ITGB1			X		
S1PR2andDNMT1		X	X		
KLK10			X		
PPP2R2B			X		
CUL2			X		
FHL2			X		
PDGFB			X		
TFE3					X
HEY2			X		
ITGAV			X		
ABCC3			X		
IRF7			X		
RBBP4			X		
ELAVL1			X		
PIAS4			X		
COX11			X		
SPAG9			X		
GPC3			X	X	X
PDGFRA		X	X	X	X
PDPK1			X		
KIF2AandIPO11andLRRC70			X		
SART3			X		
ROBO1			X		
DLC1			X	X	X
BARD1			X	X	X
HSD17B12			X		
THOC1			X		
ZFHX3					X
KALRN			X		
CCBE1			X		
XRCC5			X		

Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
ZNF19andZNF23			X		
SCNN1AandTNFRSF1A			X		
NF1	X	X	X		
MLL2		X			
YAP1			X		
ADAMTSL4			X		
EIF5A			X		
EGFLAM			X		
SHPRH			X		
SIK3			X		
NCF1			X		
CKAP5			X		
ADAM10			X		
ATXN3			X		
H19			X		X
FBXO5			X		
NFKB1			X		
ANK3			X		
ABCC1			X		
SIGLEC11andNUP62andIL4I1			X		
ZNF217			X		
LY75andCD302			X		
C20orf108			X		
MUC13			X		
GAS8			X		
CLIP1			X		
RNASEN			X		
CTSA			X		
NCAPD2andGAPDH			X		
TNKS			X		
NOTCH2		X			
IGF2R			X	X	X
CALD1			X		
PSIP1			X		
ATM	X	X	X	X	X
RNASET2andRPS6KA2			X		
KITLG			X		
FAM200BandBST1			X		
PFKFB3			X		
DKFZP686I15217andNQO2			X		X
DDR2			X		
CRTC1					X
DDR1			X		
PTGS1			X		
CEACAM1			X		
PMEPA1			X		
RBM3			X		

Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
SNRPD3andGGT1			X		
GOLGA5					X
EPAS1			X		
RBL2			X		
LMNA			X		
GTF2H4andVARS2			X		
ATF2andCHN1			X		
TNFRSF10C			X		
ARSGandPRKAR1A			X		X
HSD17B7			X		
KAT5			X		
CTHRC1			X		X
CSMD1			X		
LIMS1			X		
GALNT1			X		
SETBP1		X			
PIK3CD			X		
HSPA8_			X		
VAV2			X		
ABCA5			X		
CASP4			X		
PTCH1		X	X		X
ERCC3			X		
TP53BP1			X		
VNN2			X		
TMEM139andCASP2			X		
HMGA1			X		
BNC2			X		
TES			X		
HUS1andPKD1L1			X		
ETV1			X		
NAT1			X		
MCPH1			X		
ADAM15			X		
ADAMTSL3			X		
RPL21andSNORD102andSNORA27			X		
PLS3			X		
PARK2			X		X
WWOX			X	X	X
SLC9A3R2			X		
EPHB2				X	X
KAT2A			X		
ARID1B		X			
GSTM4andGSTM2andGSTM1			X		
RNF5			X		
PLOD2			X		
MCM8			X		

Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
ADRM1			X		
HSD17B4andFAM170A			X		
RAB31			X		
MRC1			X		
PPAP2AandRNF138P1			X		
CFLARandRNU7-45P			X		
PRCC					X
S100A3andS100A4			X		
HSPG2			X		
BRAF		X	X	X	X
TFF3			X		
SRSF3			X		
STAG3andPVRIG			X		
AKT3			X		
DAPK1			X		
IL1R1			X		
CDH5			X		
EHBP1					X
GPM6A			X		
PROCR			X		
AGA			X		
NPC1					X
MIB1			X		
PRKCA			X		X
TRIP10			X		
CRISP3			X		
LAPTM4B			X		
CES1			X		
HGF			X		
CCND3			X		
LAMC1			X		
USP1			X		
CTGF			X		
ELAC1andSMAD4		X	X	X	X
SEMA3C			X		
ANXA5			X		
ANGPT1			X		
STMN1			X		
KAT2B			X		
RAN			X		
MED23			X		
TWF2andTLR9			X		
SLC2A1			X		
EEF1A1			X		
BTAF1			X		
TCF25andMC1RandTUBB3			X		
SMAD5			X		

Gènes	Littérature	Vogelstein	Glad4U	ClinVar	OMIM
EPHA1			X		
RAD51L1			X		
MDM4			X		
RGS1			X		
PLXND1			X		
PTCH2					X
SUV420H2			X		
TNFSF12-TNFSF13			X		
PSMB9			X		
CASP3			X		
VAMP2andPER1			X		
MLLT4			X		
ANGPT2			X		
DYRK1B			X		
TRIM24			X		X
TRIP11			X		
PMS2	X		X	X	X
RPL4_			X		
PARP1	X		X		
MMP11			X		
SLC39A6			X		
ENTPD1andC10orf131andCC2D2B			X		
PLXNA3			X		
CDC25A			X		
FZR1			X		
SLC44A4			X		
ITGA4			X		