Analyse génomique en médecine de précision : Optimisations et outils de visualisation

Frederic Commo

NNT : 2015SACLS132

THESE DE DOCTORAT

DE L'UNIVERSITE PARIS-SACLAY

Laboratoire Inserm U981, Gustave Roussy

ECOLE DOCTORALE N° 582 CBMS

Cancérologie - Biologie - Médecine - Santé

Spécialité : Recherche clinique, innovation technologique, santé publique

Par

M. Frédéric Commo

Analyse génomique en médecine de précision : Optimisations & outils de visualisation

Thèse présentée et soutenue à Gustave Roussy, le 24 novembre 2015

Composition du jury :

Professeur Gilles Vassal	Gustave Roussy, Université Paris-Sud	Président
Professeur Richard Iggo	Institut Bergonié, Université de Bordeaux	Rapporteur
Docteur Aurélien de Reynies	Ligue Contre le Cancer, Université Paris Descartes	Rapporteur
Professeur Jean-Charles Soria	Gustave Roussy, Université Paris-Sud	Examinateur
Professeur Christophe Ambroise	UMR CNRS 8071, Université d'Evry Val d'Essone	Examinateur
Docteur Ivan Bièche	Institut Curie, Université Paris Descartes	Examinateur
Professeur Fabrice André	Gustave Roussy, Université Paris-Sud	Directeur de thèse

université PARIS-SACLAY | ÉCOLE DOCTORALE Cancérologie, biologie, médecine, santé (CBMS)

Titre : Analyse génomique en médecine de précision : Optimisations & outils de visualisation

Mots clé : Hybridation génomique comparative, aCGH, médecine de précision.

Résumé : Un nouveau paradigme tente de s'imposer en oncologie ; identifier les anomalies moléculaires dans la tumeur d'un patient, et proposer une thérapie ciblée, en relation avec ces altérations moléculaires. Nous discutons ici des altérations moléculaires considérées pour une orientation thérapeutique, ainsi que de leurs méthodes d'identification : parmi les altérations recherchées, les anomalies de nombre de copies tiennent une place importante, et nous nous concentrons plus précisément sur leur identification par hybridation génomique comparative (aCGH). Nous montrons, d'abord à partir de lignées cellulaires caractérisées, que l'analyse du nombre de copies par aCGH n'est pas triviale et qu'en particulier le choix de la centralisation peut être déterminant ; différentes stratégies de centralisation peuvent conduire à des profils génomiques différents, certains aboutissant à des interprétations erronées. Nous montrons ensuite, à partir de cohortes de patients, qu'une conséquence majeure est de retenir ou non certaines altérations actionnables dans la prise de décision thérapeutique. Ce travail nous a conduit à développer un workflow complet dédié à l'analyse aCGH, capable de prendre en charge les sources de données les plus courantes. Ce workflow intègre les solutions discutées, assure une entière traçabilité des analyses, et apporte une aide à l'interprétation des profils grâce à des solutions interactives de visualisation. Ce workflow, dénommé rCH, a été implémenté sous forme d'un package R, et déposé sur le site Bioconductor. Les solutions de visualisation interactives sont disponibles en ligne. Le code de l'application est disponible pour une installation sur un serveur institutionnel.

Title : Genomic analysis within precision medicine : Optimizations & visualization tools

Keywords : comparative genomic hybridization, aCGH, precision medicine.

Abstract: In oncology, a new paradigm tries to impose itself ; analyzing patient's tumors, and identifying molecular alterations matching with targeted therapies to guide a personalized therapeutic orientation. Here, We discuss the molecular alterations possibly relevant for a therapeutic orientation, as well as the methods used for their identification : among the alterations of interest, copy number variations are widely used, and we more specifically focus on comparative genomic hybridization (aCGH). We show, using well characterized cell lines, that identification of CNV is not trivial. In particular, the choice for centralizing profiles can be critical, and different strategies for adjusting profiles on a theoretical 2n baseline can lead to erroneous interpretations. Next, we show, using tumor samples, that a major consequence is to include, or miss, targetable alterations within the decision procedure. This work lead us to develop a comprehensive workflow, dedicated to aCGH analysis. This workflow supports the major aCGH platforms, ensure a full traceability of the entire process and provides interactive visualization tools to assist the interpretation. This workflow, called rCGH, has been implemented as a R package, and is available on Bioconductor. The interactive visualization tools are available on line, and are ready to be installed on any institutional server.

Remerciements

Je remercie chaleureusement le Professeur Fabrice André, mon directeur de thèse, pour avoir accepté et soutenu ce projet de thèse, sans réserve. Son savoir et sa passion pour la médecine sont communicatifs et incitent sans cesse à se dépasser.
Par ton aide et ton soutien financier, tu m'as offert l'extraordinaire opportunité de travailler durant 2 ans aux Etats-Unis et de m'enrichir d'une nouvelle culture scientifique. Je t'en serai éternellement reconnaissant.

Je remercie le Professeur Jean-Charles Soria pour son soutien et son aide qui ont été, pour moi, d'une valeur infinie.
Tes compétences et tes qualités humaines sont reconnues de tous, et je tire une immense fierté de l'amitié et la confiance dont tu m'honores depuis près de 20 ans. Je resterai pour longtemps ton débiteur.

Je remercie les Professeurs Gilles Vassal, Directeur de la Recherche Clinique à Gustave Roussy, et Alexander Eggermont, Directeur Général à Gustave Roussy, pour l'aide qu'ils ont apporté à ce projet.
Votre soutien et vos encouragements ont été indispensables à mon départ pour les Etats-Unis.

Je remercie tout particulièrement le Professeur Gilles Vassal de m'avoir fait l'honneur de présider mon jury de thèse. Je tiens, par ces quelques mots, à t'exprimer ma plus profonde gratitude.

Je remercie le Docteur Stephen Friend, Président et co-fondateur de Sage Bionetworks, Seattle. Son accueil, sa gentillesse et son soutien chaleureux m'ont été d'une aide plus que précieuse. Sa passion sans borne pour la Science, son dynamisme et son investissement au service des autres ne peuvent qu'impressionner, et poussent chacun à donner plus que le meilleur de soi.

Je tiens à remercier très chaleureusement Diane Gary, Directrice financière à Sage Bionetworks, Seattle. Sa disponibilité n'a d'égale que sa gentillesse, et son aide précieuse aura permis une transition efficace vers une nouvelle vie.

Je remercie vivement Justin Guinney, Bruce Hoff, Jay Hudgson, mon ami Brian Bot, ainsi tous les collaborateurs travaillant à Sage Bionetworks. Leur disponibilité et leur capacité à partager avec passion leur savoir ont été d'un grand secours. Ils ont tous très largement contribué à faire de ces 2 ans à Seattle un souvenir impérissable.

Un remerciement tout particulier à mon ami le docteur Charles Ferté, avec qui je garde en commun des souvenirs de Seattle. Merci pour m'avoir si souvent éclairé de tes compétences en oncologie et d'avoir apporté un regard critique sur mon travail.

Je remercie la société Roche, et plus particulièrement les docteurs Ariel Savina et Nathalie Varoqueaux, pour avoir cru en ce projet et participé à son financement. Ce soutien a été plus qu'appréciable.

Une pensée amicale et sincère aux amis que nous avons laissé à Seattle. Au plaisir de vous revoir et partager avec vous quelques bouteilles de vin, français bien sûr.

Je tiens également à exprimer ma profonde reconnaissance à M. Alain Donadey, responsable de la formation continue à l'Université de Technologie de Compiègne, ainsi qu'aux enseignants que j'ai eu l'honneur de côtoyer dans l'enceinte de cette prestigieuse Ecole. Je n'oublie pas ces années passées à l'UTC, et je sais tout ce que je vous dois.

Je remercie ma merveilleuse épouse, Cécilia. Toujours patiente, à la fois soutien et source d'inspiration, ton amour me pousse à me dépasser sans cesse, et ton sourire illumine mes jours.
Merci à nos enfants, Sasha et Solal, qui ont accepté de participer à cette aventure. J'espère qu'ils s'en seront enrichi autant que je l'ai été.

Je dédie cette thèse à mes grands-parents, Madeleine et Robert. Je sais à quel point ils auraient été fiers.

TABLE DES MATIERES

1 Le cancer, une maladie du génome

1.1 Origines

En 1902, le biologiste allemand Theodor Boveri (1862-1915) écrivait « […] malignant tumours might be the consequence of a certain abnormal chromosome constitution, which in some circumstances can be generated by multipolar mitoses. I had intended, even at that time, to put that assumption on a firmer footing in a separate article. But the scepticism with which my ideas were met when I discussed them with investigators who act as judges in this area induced me to abandon the project. »

Ces propos, rapportés par Henry Harris dans Journal of Cell Science en 2008 (Boveri 2008), reflètent la difficulté d'attribuer alors à des structures encore mal connues, les origines de maladies tout aussi mal comprises.

Une décennie plus tard, Thomas Morgan (1866-1945) conclura de ces travaux que les gènes sont portés par les chromosomes, puis Beadle (1909-1989) et Tatum (1909-1975) poseront le dogme « un gène, une protéine » en montrant, en 1941, que la modification d'un gène induit la modification d'une caractéristique (Beadle & Tatum 1941).

Les preuves de la relation entre gènes et carcinogénèse seront finalement apportées par la découverte des oncogènes et des gènes suppresseurs de tumeurs, ainsi que des effets de leurs altérations : en 1982, Tabin et collaborateurs mettent en évidence le pouvoir transformant et immortalisant des mutations du gène HRAS sur des fibroblastes en culture (Tabin et al. 1982), et en 1986, les pertes du gène RB1, conduisant à une perte de fonction, sont montrées comme associées au rétinoblastome (Friend et al. 1986).

Au cours de cette même décennie, puis dans les suivantes, d'autres altérations génétiques (mutations ou anomalies de nombre de copies) seront misent en évidence dans de nombreux types de cancers, et rapidement, des relations de

causalité seront fortement suspectées (Bishop 1991; Weinstein 2002, ainsi que Macconaill & Garraway 2010; Garraway & Lander 2013 pour des revues plus récentes).
Dans le même temps, le constat est fait des limites à l'orientation des patients vers des chimiothérapies standard : l'efficacité des traitements chez les patients atteints de cancers du poumon reste limitée (Schiller et al. 2002), et des bénéfices marginaux dans le traitement des mélanomes ne sont obtenus qu'au prix d'une toxicité accrue (Eggermont & Kirkwood 2004). La nécessité est avérée de comprendre les différences entre répondeurs et non-répondeurs.

Des découvertes de l'implication des gènes dans les processus tumoraux, et des constats d'échecs relatifs des traitements de référence, était née l'idée de projets de grande envergure pour l'exploration du génome. Ces projets auront pour but une meilleure connaissance des gènes, de leurs séquences et de leurs rôles dans la carcinogénèse et la progression tumorale.

1.2 Le Projet Génome Humain

Portée dans le même temps, et de manière indépendante, par plusieurs personnalités scientifiques, la volonté d'une exploration plus exhaustive et systématique du génome émerge à la fin des années 80 :

« If we wish to learn more about cancer, we must now concentrate on the celular genome. […] The classification of the genes will facilitate the identification of those involved in progression. […] Knowledge of the genes involved in progression would open new therapeutic approaches, which might lead to a general cancer cure if progression has common features in all cancers. » (Dulbecco 1986).

L'objectif est clair : cartographier le génome complet, et y localiser les gènes, fourniront les outils qui permettront de comprendre les mécanismes génétiques

impliqués dans la progression tumorale. Le choix technologique qui s'impose est le séquençage.

Le Projet Génome Humain sera lancé en 1988 sous la direction de the National Institutes of Health (NIH), et ce concomitamment à la création de the Human Genome Organization (HUGO) chargée de coordonner les efforts de 20 groupes de recherche présents à travers 7 pays.
Au cours de sa construction, le Projet Génome Humain adoptera explicitement 2 principes majeurs :

- *Ce projet collaboratif sera ouvert à toutes les nations.*
- *les données seront rendues publiques, le plus rapidement possible, et sans restriction.*

En avance sur le planning prévisionnel, la première séquence brute, couvrant 99.99% du génome humain, sera publiée en 2001 (Lander et al. 2001), et le projet officiellement achevé en 2003.

1.3 Les bases de données publiques

Sur la base des données de séquençage rendues disponibles par le Projet Génome Humain, d'autres projets vont naître, venant eux aussi enrichir les bases de données mises à la disposition de la communauté scientifique.

1.3.1 Bases de données tumorales

L'International Cancer Genome Consortium (ICGC) lance The Cancer Genome Atlas (TCGA) dont le but est, cette fois, d'étudier le génome d'un très large panel de tumeurs et d'en répertorier les anomalies génomiques (Hudson et al. 2010; Chin et al. 2011). Sans cesse enrichi par de nouveaux cas, le portail TCGA donne aujourd'hui accès aux profils génomiques (mutations, altérations de nombre de copies, expression génique, méthylations, microRNA et données cliniques) de plus de 11,000 échantillons tumoraux couvrant 33 types de cancers (Figure 1).

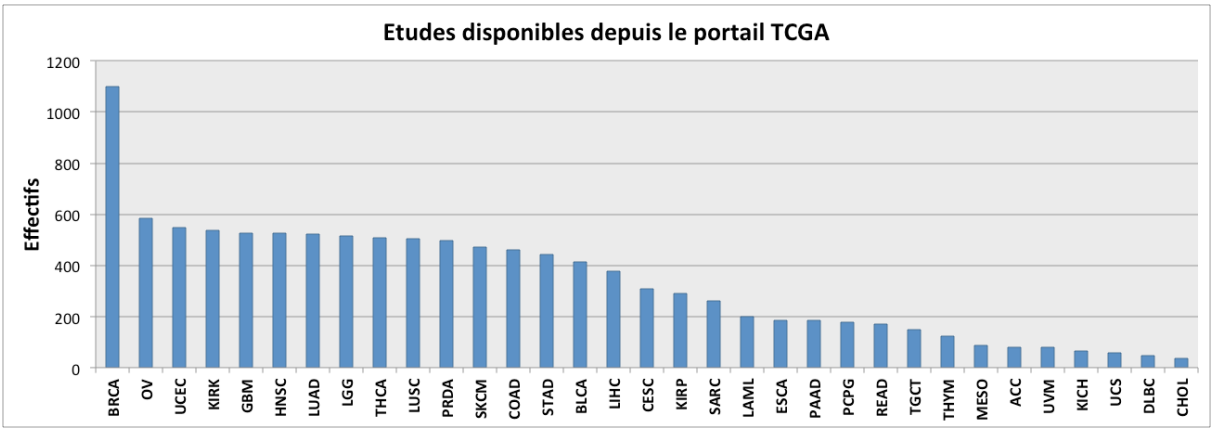


Figure 1 : Etudes disponibles depuis le portail du projet TCGA (type de tumeur et taille des cohortes). Source https://tcga-data.nci.nih.gov/tcga, 16/03/2015.

La mise à disponibilité de cette masse considérable de données a d'ores et déjà donné lieu à de très nombreuses publications ; certaines ont été soumises au nom du consortium lui-même et proposent une caractérisation exhaustive de plusieurs types de cancers (Collisson et al. 2014; Bass et al. 2014; Hoadley et al. 2014). De nombreux autres travaux se sont appuyés sur les données TCGA pour des développements méthodologiques (Wang et al. 2014), l'étude de l'impact des mutations ou de l'expression des gènes sur la biologie des tumeurs (Liu et al. 2014; Zhang et al. 2014), ou encore l'identification de nouvelles cibles thérapeutiques potentielles (Mikheev et al. 2014).

1.3.2 Bases de données cellulaires

Les lignées cellulaires ont aussi été largement étudiées, et les données moléculaires, ainsi que les sensibilités à de nombreux agents thérapeutiques, ont été mises à la disposition de la communauté scientifique. Ces études couvrent plus de 1300 lignées cellulaires.

Le premier de ces projets avait été initié par le National Cancer Institute (NCI), avec l'étude d'un panel de 60 lignées cellulaires. En plus des données moléculaires - analyse des profils génomiques et profils d'expression - cette étude a mis à disposition les résultats de sensibilité à l'exposition à plus de 40,000 composés chimiques (Shoemaker 2006).

Suivront 3 autres projets couvrant un nombre plus important de lignées, mais un nombre plus restreint de composés chimiques : The Cancer Cell Line Encyclopedia (CCLE), avec près de 1000 lignées cellulaires et 24 molécules (Barretina et al. 2012), le Cancer Genome Project (CGE), du Welcome Trust Sanger Institute, avec 732 lignées et 131 molécules (Garnett et al. 2012), et les données GlaxoSmithKline fournissant l'analyse génomique pour 318 lignées, mais sans tests de sensibilité à des composants chimiques (Kim et al. 2014) (Figure 2).

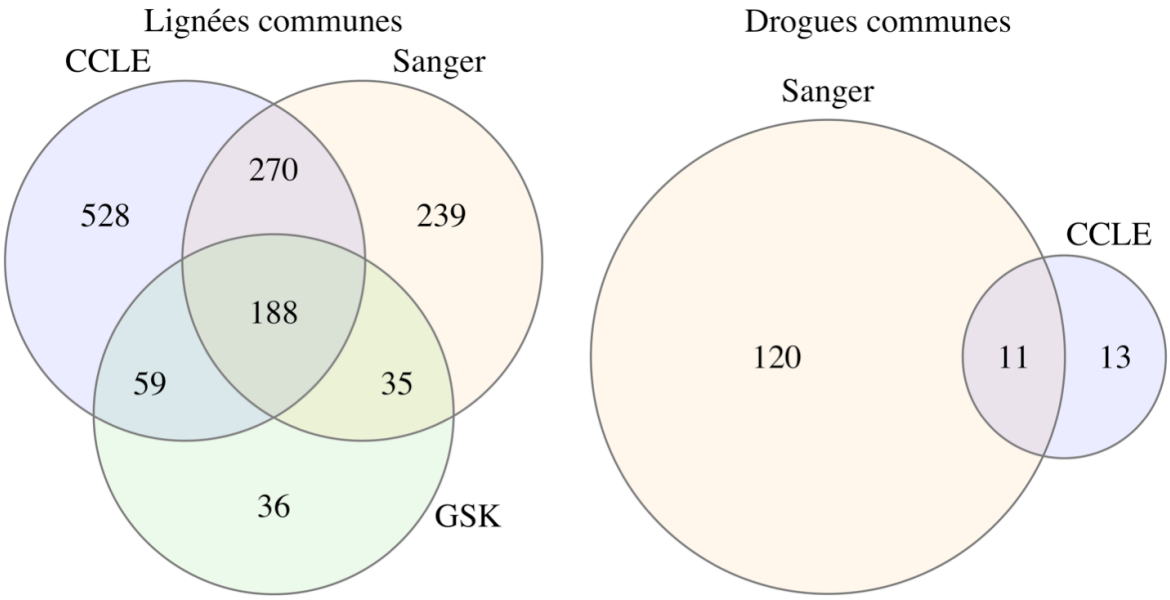


Figure 2 : Lignées cellulaires et agents thérapeutiques explorés dans les 3 études CCLE, Sanger et GSK. L'étude GSK n'a pas porté sur la réponse aux drogues, et le faible nombre de composés communs entre les séries CCLE et Sanger rend difficile leur comparaison pour ces données.

Toutefois, et malgré les efforts consentis, ces données de lignées cellulaires ne semblent pas avoir conduit aux résultats espérés, à savoir, générer et valider des hypothèses de travail sur les relations entre anomalies moléculaires et sensibilité aux agents chimiques ; si les données moléculaires (profils génomiques et expression génique) semblent reproductibles entre ces différentes séries, les

mesures de sensibilité in vitro ont donné des résultats peu concordants pour les lignées et molécules communes : ces résultats pourraient remettre en cause l'intérêt de telles données pour la génération d'hypothèses (Haibe-Kains et al. 2013; Weinstein & Lorenzi 2013).

2 Vers une médecine de précision en oncologie

2.1 Arguments

Parallèlement à l'engagement de ces projets, les constats étaient faits de la relative inefficacité des chimiothérapies (Schiller et al. 2002; Eggermont & Kirkwood 2004). Il devenait nécessaire d'identifier les mécanismes impliqués dans le développement et la progression tumorale, susceptibles être ciblés par des thérapies spécifiques (Sawyers 2004).

La nécessité de produire et fournir un accès à plus de données était aussi motivée par le besoin d'identifier des causes moléculaires pouvant expliquer les résistances aux thérapies standard.

Le principe de proposer des thérapies adaptées, dans des contextes moléculaires spécifiques, était appuyé pas la suspicion de l'existence de gènes 'driver', et par les succès de 2 prises en charges thérapeutiques, précurseurs des thérapies ciblées.

2.1.1 Les gènes 'drivers'

Avec la découverte des oncogènes et des gènes suppresseurs de tumeurs, avait émergé le concept de « gènes drivers » ; des gènes impliqués dans l'émergence de tumeurs et indispensables à leur progression (Bishop 1991).

Parmi les premiers oncogènes identifiés, les gènes Ras et Myc ont fait l'objet de recherches intensives.

En 1999, Chin et collaborateurs montraient que H-Ras, porteur de la mutation activatrice V12G, pouvait induire des mélanomes chez des souris transgéniques (Chin et al. 1999).

Dans le même temps était mis en évidence le rôle de Myc dans le développement de lymphomes et de sarcomes osteogéniques (Felsher & Bishop 1999; Jain et al. 2002).
Faits intéressants, ces études montraient également que l'inactivation de ces gènes entrainait une régression des proliférations tumorales.

Ces travaux semblaient donc indiquer qu'identifier les gènes drivers d'une tumeur, et cibler plus précisément leurs fonctions à l'aide d'inhibiteurs spécifiques, pouvait constituer une nouvelle approche pour la prise en charge des patients.
Cette hypothèse sera rapidement soutenue par l'identification, dans certaines pathologies tumorales, de nouvelles altérations moléculaires impliquant des dérégulations importantes, prédictives de réponses favorables à des thérapies spécifiques.

2.1.2 Le trastuzumab et la protéine Her2/neu

Le trastuzumab est un anticorps monoclonal ciblant la protéine Her2/neu, un récepteur tyrosine-kinase transmembranaire impliqué dans de nombreux processus cellulaires (prolifération, différenciation, apoptose,..).
Développé dans les années 90 (Carter et al. 1992), après que la surexpression de Her2/neu ait été identifiée comme associée à un mauvais pronostic dans les tumeurs du sein, cet agent thérapeutique a prouvé son efficacité pour le traitement des patientes Her2+ (Slamon et al. 2001; Baselga et al. 2005) (Figure 3).

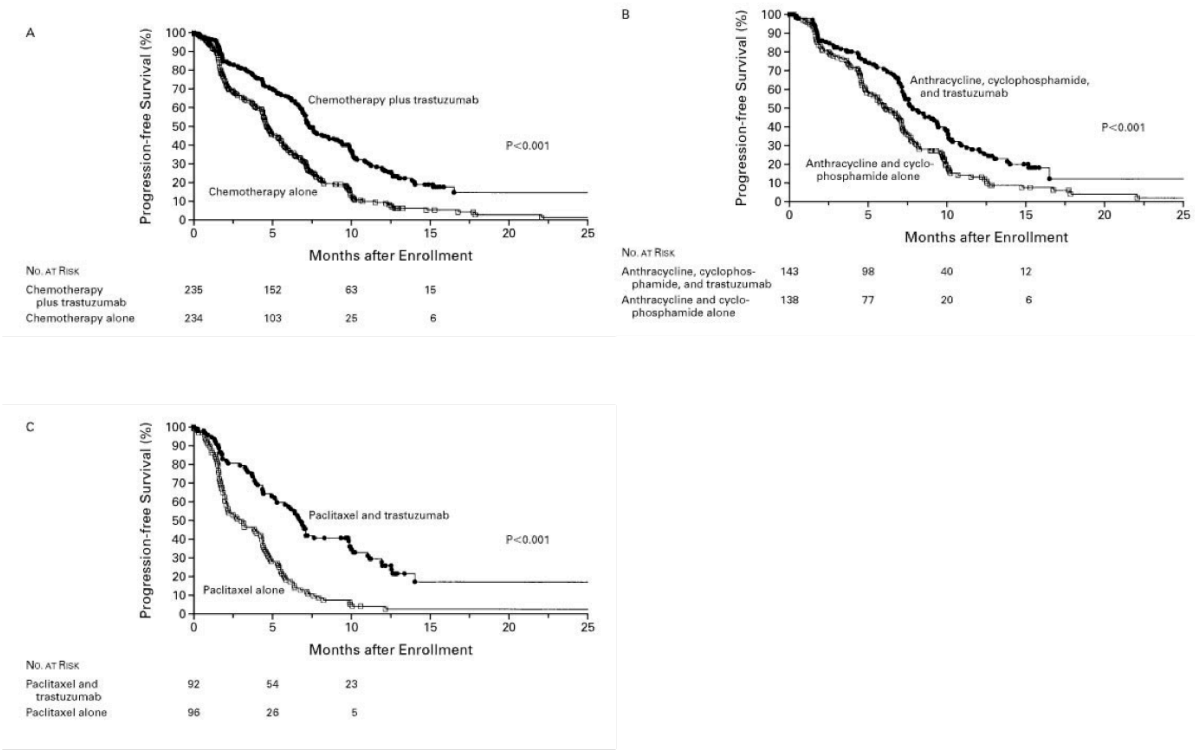


Figure 3 : Le trastuzumab, en association avec une chimiothérapie conventionnelle, accroit significativement la survie chez les patientes atteintes de tumeurs du sein surexprimant Her2/neu (Slamon et al. New Eng. J. Med. 2001).

Dans le même temps est développé un test diagnostic, l'hercepTest™, permettant une estimation fiable et rapide de l'expression de Her2/neu dans les tumeurs. Fait marquant de la réussite de cette stratégie de thérapie ciblée, le trastuzumab pour le traitement des tumeurs du sein Her2+, et l'hercepTest™ en tant que test diagnostic compagnon, seront approuvés aux Etats-Unis par la Food and Drug Administration (FDA).

2.1.3 L'imatinib et la protéine BRC-Abl

La protéine BCR-Abl résulte d'un remaniement chromosomique, la translocation t(9q34;22q11), se traduisant par la fusion de 2 gènes : BCR et Abl. Cette anomalie, appelée chromosome Philadelphie (Ph+), et initialement décrite dans les leucémies myéloïdes chroniques (LMC) (Rowley 1973), est responsable de la production d'une protéine de fusion dont l'activité kinase est bien supérieure à la protéine Abl native (Konopka et al. 1984; Sattler & Griffin 2001).

D'importants travaux sur des souris transgéniques ont montré que cette anomalie était un inducteur de transformation tumorale, et que cette transformation était réversible par inhibition du gène (Huettner et al. 2000).

La découverte de cette altération et de son implication dans la LMC ont conduit au développement de l'imatinib ; un inhibiteur de tyrosine kinase agissant par compétition avec l'ATP, source de phosphate pour les kinases, et ciblant principalement BCR-Abl. Les bénéfices du traitement des LMC/Ph+ par l'imatinib furent démontrés par plusieurs études (Druker et al. 2001; Vigneri & Wang 2001; O'Brien et al. 2003) (Figure 4), et l'utilisation de cet inhibiteur pour cette pathologie fut approuvée par la FDA en 2001. Plus tard, l'intérêt de cette molécule fut démontrée dans le traitement des tumeurs stromales gastro-intestinales (GIST) présentant des mutation du gène KIT ou du gène PDGFRA (Demetri et al. 2002; de Silva & Reid 2003).

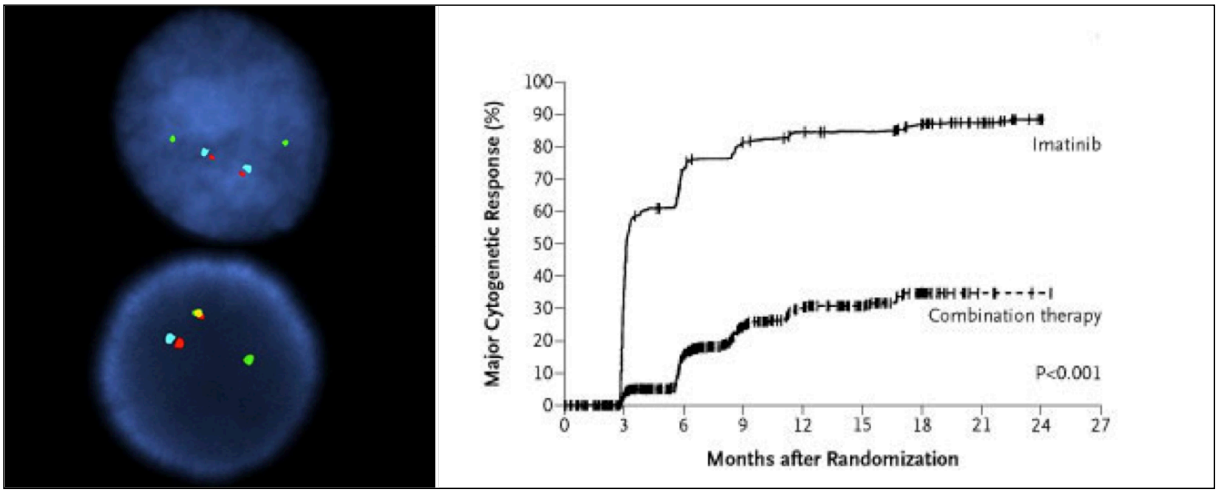


Figure 4 : La protéine de fusion BCR-Abl. Gauche : hybridation in situ fluorescente (fish) montrant la fusion des 2 gènes BCR et Abl (BCR : vert, Abl : rouge, fusion : orange, centromère : bleu). Droite : Réponse cytogénétique au traitement par l'imatinib chez les patients atteints de LMC, mesurée par la disparition de cellules Ph+ (O'brien et al. New Eng. J. Med. 2003).

2.1.4 Nouvelles données moléculaires

Avec le trastuzumab et l'imatinib, démonstration était faite des bénéfices de thérapies ciblées dans des contextes particuliers d'anomalies moléculaires. Les nouvelles données disponibles, mais aussi le développement de nouvelles molécules à actions ciblées, ont alors ouvert la voie à de nombreux projets de screening moléculaire, prospectifs ou rétrospectifs, visant à identifier de nouvelles cibles. Ceux-ci ont rapidement conduit à mettre en évidence de nouvelles anomalies, parfois observées avec des fréquences élevées dans certaines pathologies (voir Garraway, 2013 pour une revue) (Figure 5).

La présence de ces mutations, translocations, ou altérations du nombre de copies, ont fourni des pistes pour le développement de prises en charges thérapeutiques spécifiques (Jurgensmeier et al. 2014) (tableau 1) ; l'efficacité du vemurafenib dans le traitement des mélanomes avec mutation $BRAF^{V600E}$ (Flaherty et al. 2010), et les bénéfices du crizotinib dans les tumeurs pulmonaires non à petites cellules (NSCLC) et porteuses d'une translocation ALK/ROS1 (Kwak et al. 2010), en sont 2 exemples.

Les études rétrospectives ont également permis de spécifier des contextes moléculaires particuliers associés à de meilleures réponses, là où les traitements standard étaient peu satisfaisants. Ainsi dans les NSCLC, la présence de la mutation L858R dans le gène EGFR (substitution d'une leucine par une arginine en position 858) s'est avérée être un facteur favorable pour l'efficacité du gefitinib et de l'erlotinib, 2 inhibiteurs de tyrosine kinase (Sordella et al. 2004; Lynch et al. 2004; Pao et al. 2004). D'autres études, menées à partir de matériel d'archive, ont ensuite précisé le contexte des sensibilités aux anti-EGFR en montrant que des mutations dans le gène KRAS réduisaient significativement la probabilité de réponse au traitement par le cetuximab dans les tumeurs coliques (Amado et al. 2008; Karapetis et al. 2008; Cappuzzo, Varella-Garcia, et al. 2008).

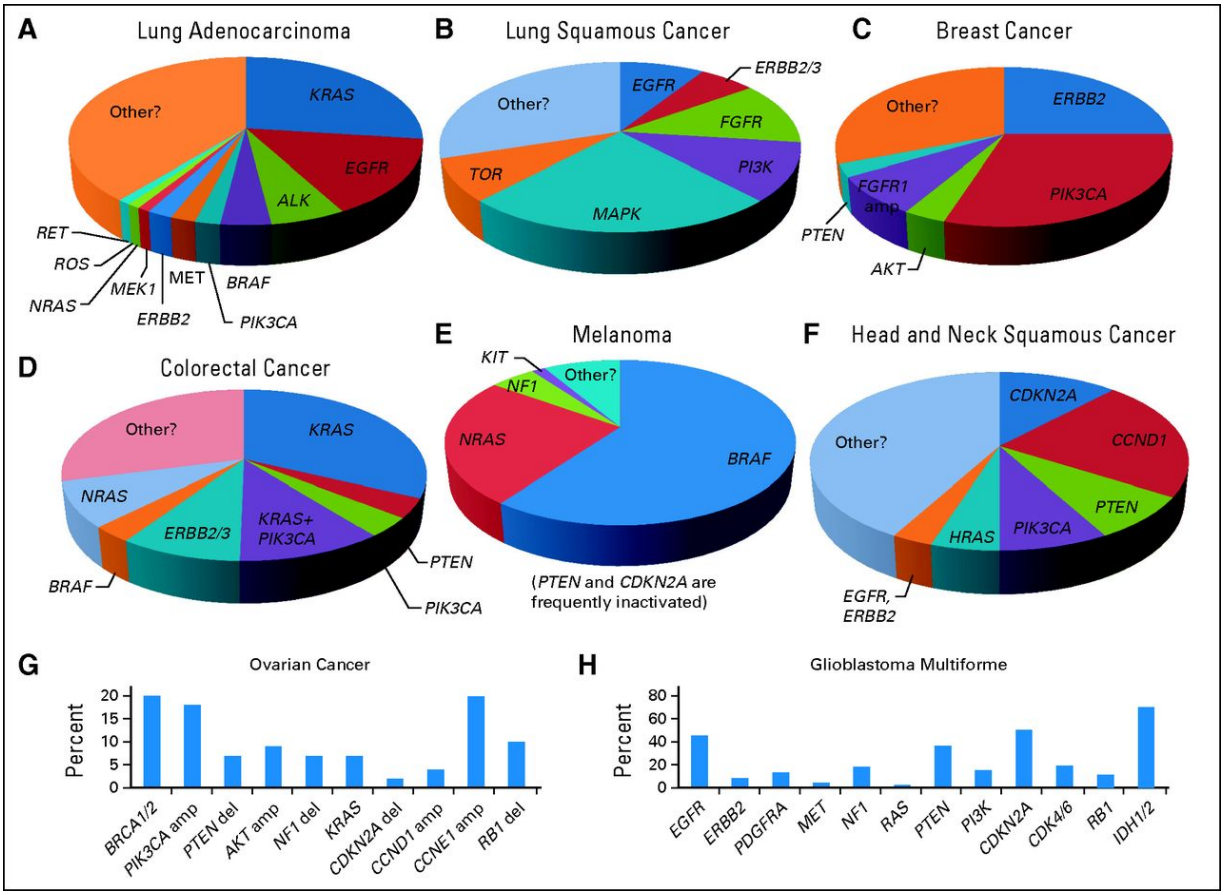


Figure 5 : Gènes actionnables connus à ce jour, et fréquence de leurs altérations dans différentes pathologies tumorales (Garraway et al., J. Clin. Oncol. , 2013)

Tableau 1 : relations entre altérations moléculaires et thérapies ciblées. D'après Jurgensmeier et al. (Jurgensmeier et al. 2014)

Target/pathway	Aberration type in solid tumors	Disease examples	Putative or proven drugs	Examples for drugs in clinical development
EGFR	Mutation Amplification	Lung cancer GBM	EGFR inhibitors	Erlotinib Gefitinib Afatinib AZD9291
HER2 (ERBB2)	Mutation Amplification	Breast cancer Gastric cancer Lung cancer	ERBB2/ERBB3 inhibitors	Lapatinib Neratinib Trastuzumab Trastuzumab-emtansine
ALK	Rearrangement Mutation	Lung cancer Neuroblastoma Colorectal cancer	ALK inhibitors	Crizotinib Ceritinib Alectinib
RET	Rearrangement Mutation	Thyroid cancer Lung cancer	RET inhibitors	Vandetanib Carbozantinib
ROS1	Rearrangement	Lung cancer	ROS1 inhibitors	Crizotinib
DDR2	Mutation	Lung cancer	DDR2 inhibitors	Dasatinib Nilotinib
FGFR1-4	Amplification Mutation	Lung cancer Gastric cancer Breast cancer Bladder cancer	FGFR inhibitors	Ponatinib Dovitinib BGJ398 AZD4547
MET/HGF	Mutation Amplification	Lung cancer Gastric cancer Colorectal cancer HCC RCC	Met inhibitors	Onartuzumab Crizotinib Foretinib INC280
KIT	Mutation	GIST Mastocytosis Melanoma	Kit inhibitors	Imatinib Nilotinib Sunitinib Dasatinib Ponatinib
PDGFRA and PDGFRB	Mutation Translocation	GIST Sarcoma GBM Leukemia Dermatofibrosarcoma protuberans	PDGFR inhibitors	Imatinib Sunitinib Ponatinib
KRAS, NRAS, HRAS (RAS-RAF-MEK)	Mutation	Most cancers, including colorectal cancer, lung cancers	Mek inhibitors	Trametinib Selumetinib
BRAF (RAS-RAF-MEK)	Mutation	Melanoma Colorectal cancer HCC	Braf inhibitors Mek inhibitors	Vemurafenib Dabrafenib Trametinib Selumetinib
PI3KCA (PTEN/PI3K/AKT/mTOR)	Mutation Amplification	Multiple, including: Breast cancer Colorectal cancer, GBM Lung cancer Endometrial	PI3K inhibitors AKT inhibitors	BKM-120 BEZ235 BYL719 GDC0941 GDC0032 MLN1117 AKT inhibitors (see below)
PIK3R1 (PTEN/PI3K/AKT/mTOR)	Mutation	Endometrial cancer Colorectal cancer	PI3K inhibitors	PI3K inhibitors (see above)
AKT1-3 (PTEN/PI3K/AKT/mTOR)	Mutation Amplification	Breast cancer Colorectal cancer Meningeal cancer Urinary tract cancers Endometrial cancer	PI3K inhibitors AKT inhibitors	GDC0086 MK2206 AZD5363 PI3K inhibitors (see above)
PTEN (PTEN/PI3K/AKT/mTOR)	Deletion/Mutation	Most cancers, including breast cancer Lung cancer Colorectal cancer	PI3K inhibitors/AKT inhibitors	PI3K and AKT inhibitors (see above)
mTOR (PTEN/PI3K/AKT/mTOR)	Mutation	Endometrial cancer RCC Colorectal cancer Lung cancer	mTOR inhibitors	Everolimus Temsirolimus MLN128 AZD2014 GDC00980 BEZ235
TSC1/2 (PTEN/PI3K/AKT/mTOR)	Mutation	Tuberous sclerosis Urinary tract cancers Endometrial cancer Cervical cancer HCCColorectal cancer	mTOR inhibitors	mTOR inhibitors (see above)
LKB1 (PTEN/PI3K/AKT/mTOR)	Mutation	Cervical cancer Small intestine cancer Lung cancer Skin cancer	mTOR inhibitors	mTOR inhibitors (see above)
SMO, PTCH1 (Hedgehog)	Mutation	Basal cell carcinoma Medulloblastoma Meningioma Breast cancer	Hedgehog inhibitors	Vismodegib
MDM2	Amplification	GBM Sarcoma	MDM2 inhibitors/antagonists disrupting p53-MDM2 interaction	RG7388
P53	Mutation	Most tumors	P53 activators	
NOTCH	Mutation Rearrangement	Breast cancer Lung cancer Ovarian cancer GBM H&N cancer	γ-secretase inhibitors ABs to notch receptors or ligands	MK0752 PF03084014 Demcizumab OMP59R5 OMP52M51 Enoticumab
CDKS	Amplification Mutation Rearrangement	Sarcoma Melanoma GBM	CDK inhibitors	Flavopiridol Palbociclib
CHK1/2	Mutation	Multiple tumors, including those listed below: Colorectal cancer Gastric cancer Endometrial cancer Breast cancer	CHK inhibitors	RG7741 LY2606368
AURKA (Aurora kinases)	Amplification	Multiple tumors	Aurora kinase inhibitors	Alisertib
ATR	Mutation Deletion	Gastric cancer Breast cancer Endometrial cancer	ATM inhibitors PARP inhibitors	No ATM inhibitors in clinical development PARP inhibitors (see below)
ATM	Mutation Deletion	Multiple tumors, including Breast cancer	ATR inhibitors PARP inhibitors	VX970 PARP inhibitors (see below)
BRCA1/2	Mutation	Breast cancer Ovarian cancer	PARP inhibitors	Olaparib Veliparib Rucaparib BMN673

2.1.5 Nouvelles technologies

Deux avancées technologiques majeures ont profondément modifié la recherche de ces altérations moléculaires ; le séquençage de 2nd génération (ou Next-Gen Sequencing, NGS) pour la recherche de mutations ou de translocations, et les microarrays à très haute densité pour la recherche d'anomalies de nombre de copies de gènes.

Les techniques NGS sont construites sur les bases de la méthode de séquençage développée par F. Sanger (Sanger et al. 1977), et sur les techniques d'amplification par PCR. Différents choix technologiques sont proposés : PCR en émulsion (Ion Torrent, Roche454) ou sur plaque (Illumina, Applied Biosystems), mais le principe général reste similaire : l'ADN est fragmenté par voie enzymatique, puis chaque fragment est séquencé de manière indépendante. Les multiples fragments de séquences obtenues sont ensuite alignés, puis comparés à des séquences de référence contenues dans les bases de données publiques (COSMIC, Sift, Polyphen,…). Ces techniques de séquençage massif permettent aujourd'hui d'analyser un génome entier, ou de restreindre l'exploration à un panel de gènes d'intérêts (voir Mardis 2011; MacConaill 2013 pour une revue des technologies NGS).

Les microarrays ont aussi bénéficié d'évolutions technologiques considérables ; en 1995, Shena et collaborateurs décrivaient une technique permettant d'évaluer simultanément l'expression de 45 gènes d'Arabidopsis thaliana, grâce à un nouveau système d'hybridation sur support de verre (Schena et al. 1995). Deux ans plus tard, Lashkari et collaborateurs étudiaient l'expression des gènes de Saccharomyces cerevisiae par hybridation de près de 5000 oligonucléotides déposés sur un même support de verre (Lashkari et al. 1997).
Les techniques actuelles de « printing » permettent de préparer des microarrays contenant jusqu'à plusieurs millions de séquences complémentaires. Ils peuvent être utilisés pour l'analyse de l'expression des gènes, ou l'exploration du génome

lui-même et la recherche d'anomalies de nombre de copies (amplifications ou délétions). Cette dernière approche est appelée array-based Comparative Genomic Hybridization (aCGH) et sera détaillée plus avant.

Là aussi, différents choix technologiques sont disponibles, mais le principe de l'hybridation sur lame reste relativement similaire ; il repose sur la capacité de 2 séquences d'acides nucléiques, complémentaires et antiparallèles, à s'apparier entre elles.
Pour l'analyse du nombre de copies, l'ADN est extrait du tissu à analyser, puis fragmenté à l'aide d'enzymes de restriction. Les fragments sont marqués à l'aide d'un traceur (biotine ou fluorochrome) avant d'être hybridés sur le support de verre. Les signaux sont quantifiés par un scanner, puis comparés à ceux d'un ADN normal, utilisé comme référence.
Les 2 plateformes les plus couramment utilisées proposent des stratégies différentes : Agilent utilise le principe de co-hybridation compétitive ; l'ADN analysé et un ADN normal de contrôle sont conjugués à l'aide de 2 fluorochromes différents (cyanine5 et cynanine3), puis co-hybridés sur le même microarray (Figure 6).
La plateforme Affymetrix propose une hybridation simple ; seul l'ADN analysé est couplé à un fluorochrome, puis hybridé. Dans ce dernier cas, l'analyse se fera par comparaison à un ADN de contrôle hybridé séparément, ou relativement à une base d'ADN virtuelle.

Les microarrays ont également été développés dans d'autres domaines comme l'analyse des méthylations de l'ADN, de l'expression des microARN, ou encore des protéines, mais ces applications sont au-delà des objectifs de ce travail.

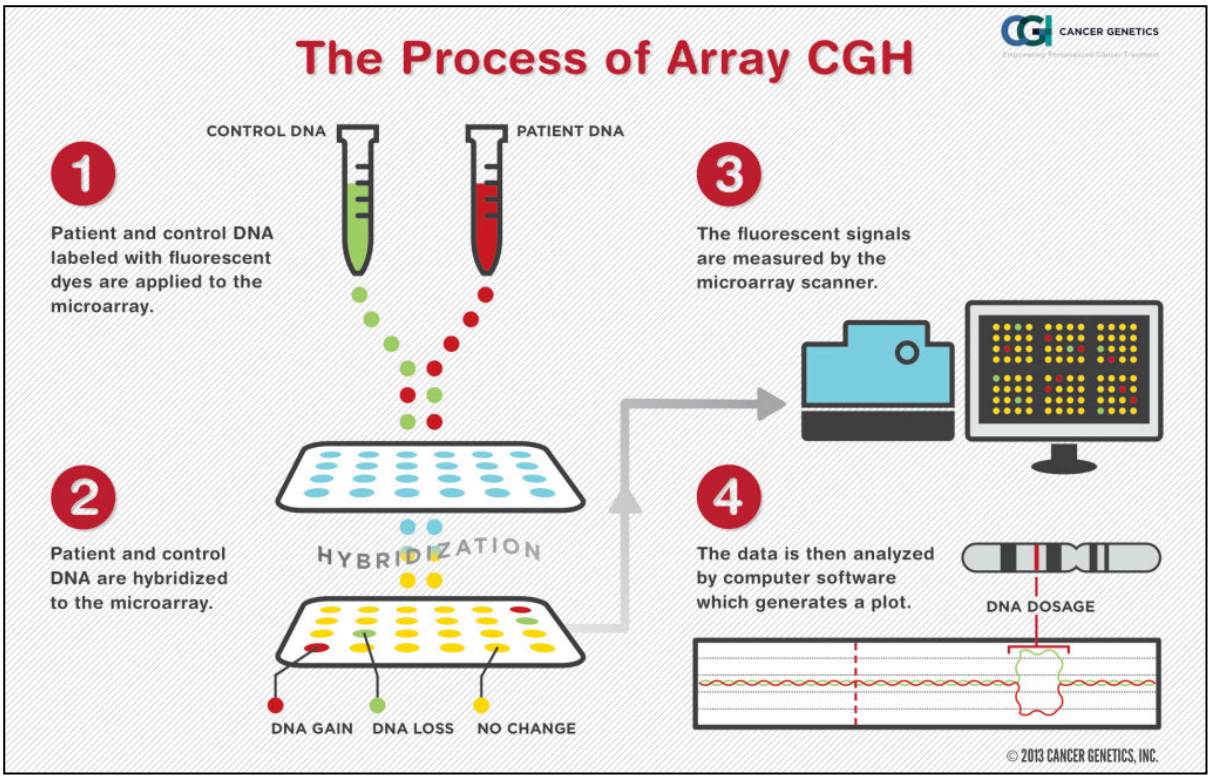


Figure 6 : Principe de l'hybridation génomique comparative (CGH) pour la détection d'anomalies de nombre de copies. Ici, l'ADN analysé et un ADN normal de contrôle sont marqués à l'aide de 2 fluorochromes (1), puis co-hybridés sur le microarray (2). Les signaux fluorescents sont lus à l'aide d'un scanner (3) avant d'être analysés par un programme informatique (4). Note : la technologie Affymetrix n'utilise pas la co-hybridation, seul l'échantillon analysé est hybridé.

2.2 Screening moléculaire pour la médecine de précision

2.2.1 Objectifs des screening moléculaires

Les oncologues disposent à présent de technologies de screening moléculaire, ainsi que d'un large panel d'altérations identifiées pour être en relation avec des sensibilités - ou des résistances - à des thérapies ciblées, c'est à dire spécifiquement dirigées contre un récepteur membranaire ou une protéine clé d'une voie de signalisation identifiée comme anormale sur la base d'analyses moléculaires.

Développer des essais de screening moléculaire, proposant d'analyser le profil moléculaire d'une tumeur pour orienter le patient vers une thérapie adaptée, est donc apparu comme une conséquence naturelle et attendue de ces avancées.
Cette approche personnalisée de la prise en charge des patients atteints de cancers, sur la base de profils moléculaires, est aujourd'hui le socle de nombreux essais cliniques, et la source d'une littérature scientifique extrêmement riche.

2.2.2 Essais de screening moléculaire en cours

De nombreux essais ont été initiés sur la base des connaissances actuelles entre anomalies moléculaires et thérapies ciblées. Ces programmes de screening moléculaire, pour une orientation thérapeutique ciblée, ont parfois fait des choix technologiques et stratégiques différents, et certains résultats ont d'ores et déjà été communiqués.

Le programme SAFIR01 (NCT01414933) s'est concentré sur les tumeurs du sein métastatiques. Cette étude a fait le choix de l'aCGH pour la détection d'anomalies de nombre de copies, et du séquençage pour la recherche de mutations dans un panel restreint de gènes d'intérêts : PIK3CA (exon 10 and 21) et AKT1 (exon 4). Cette étude a validé, dans ce contexte, la faisabilité d'une approche moléculaire pour l'orientation thérapeutique : au moins une altération moléculaire était identifiée chez 195 des 423 patientes (46%) incluses dans l'étude. Sur 55 patientes pouvant être orientées, 43 ont pu bénéficier d'une thérapie ciblée, et 13 d'entre elles (30%) ont montré une réponse objective ou une stabilisation de la maladie (André et al. 2014).

Le programme MOSCATO (Molecular Screening for Cancer Treatment Optimization, NCT01566019), s'appuie également sur l'aCGH et la recherche de mutations sur un panel de gènes d'intérêts (Hollebecque et al. 2013). Mais à la différence de SAFIR01, MOSCATO s'intéresse à la faisabilité et au bénéfice d'un choix thérapeutique orienté par le profil moléculaire, dans un contexte large de

tumeurs solides. Cette étude est, à ce jour, toujours en cours, mais les premiers résultats ont montré que des anomalies actionnables étaient identifiables chez 46% des patients biopsiés, et qu'un bénéfice thérapeutique était observé pour 47% de ceux ayant pu recevoir une thérapie ciblée.

D'autres programmes ont été initiés avec des choix stratégiques différents :
Le programme randomisé REMANUS04 a inclus plus de 300 patientes atteintes de cancer du sein, et a exploité l'analyse de l'expression génique par microarray pour guider le choix thérapeutique. Si ce programme a montré une relative faisabilité de cette technologie comme pratique de routine – des données microarray étaient exploitables pour 60% des patientes éligibles - l'analyse n'a pas montré de bénéfice significatif d'une orientation guidée par le profil d'expression, en comparaison du traitement standard (Pierga et al. 2012).

Les chercheurs canadiens du Princess Margaret Hospital–University Health Network se sont adressés aux tumeurs solides, mais en se focalisant sur la recherche de mutations intra-exoniques dans 19 gènes d'intérêt. Malgré un nombre relativement modeste de patients inclus (n = 51), cette étude a, elle aussi, démontré la faisabilité d'une exploration moléculaire des tumeurs dans une pratique de routine, et a validé, dans le même temps, plusieurs choix technologiques, comme l'utilisation de tissus inclus et fixés en paraffine pour la recherche de mutations (Tran et al. 2013).

Le programme IMPACT (Initiative for Molecular Profiling and Advanced Cancer Therapy, NCT00851032), mis en place au MD Anderson Cancer Center, s'est intéressé au profil moléculaire des tumeurs solides, mais n'a pas fait le choix des technologies à haut débit : seul un panel limité à 10 gènes était exploré pour la recherche de mutations par séquençage après PCR, la recherche de translocation ALK/ros1 était effectuée par hybridation fluorescente (FISH), et les pertes ou gains d'expression de PTEN, ER et HER2/neu étaient recherchés par

immunohistochimie. Bien que non randomisé, cet essai a comparé les réponses chez les patients orientés vers des thérapies ciblées à ceux non orientés, et a apporté des éléments en faveur d'un bénéfice d'une orientation thérapeutique guidée par le profil moléculaire (Tsimberidou et al. 2014).

Le programme BATTLE (Biomarker integrated Approaches of Targeted Therapy for Lung cancer Elimination, NCT00409968), également initié au MD Anderson Cancer Center, s'est focalisé sur les tumeurs pulmonaires non à petites cellules, et a exploré un choix très restreint de 7 gènes pouvant être associés à 4 thérapies ciblées. BATTLE propose une stratégie innovante pour l'orientation des patients : les patients pour lesquels une anomalie actionnable est identifiée sont d'abord orientés aléatoirement vers une des 4 thérapies ciblées, indépendamment des résultats moléculaires. Les réponses aux thérapies alimentent ensuite un algorithme décisionnel, utilisé pour orienter les patients suivants. Les nouveaux résultats enrichissent, à leur tour, l'algorithme de décision.

L'étude pilote MI-ONCOSEQ (Michigan Oncology Sequencing Project) a fait le choix du séquençage exhaustif de l'ADN et de l'ARN, pour rechercher plusieurs types d'altérations : mutations, gènes de fusion, anomalies de nombre, anomalies d'expression. Seuls 2 patients ont été inclus, et l'étude n'apporte pas de preuve quant au bénéfice du choix de thérapies ciblées sur la base des altérations identifiées. Cependant, les auteurs pointent la difficulté d'intégrer l'ensemble des anomalies moléculaires dans une prise de décision : anomalies ayant un impact inconnu sur la biologie, disponibilité des traitements (Roychowdhury et al. 2011).

Le programme multicentrique SHIVA (NCT01771458) initié à l'institut Curie, Paris, est un essai randomisé se concentrant sur les tumeurs avec métastases, sans spécificité de type histologique ou origine tissulaire, et réfractaires aux traitements standards.

SHIVA combine la recherche de mutations par séquençage sur une série restreinte de gènes, l'analyse des profils génomiques par aCGH, et l'exploration des récepteurs estrogènes, progestérone et androgènes, par immunohistochimie. L'étude identifie au moins une anomalie moléculaire actionnable chez 293 des 741 patients inclus (39.5%) mais ne montre aucun bénéfice significatif de l'orientation thérapeutique guidée, en comparaison des traitements standards (Tourneau et al. 2015).

De nombreux autres essais de screening ont été initiés, parfois sur des pathologies spécifiques, et avec des choix stratégiques différents pour la recherche d'altérations moléculaires (Bedard et al. 2013).

2.2.3 Limites aux essais de screening

IMPACT a montré que sur 1542 patients inclus, 1276 patients avaient pu bénéficié d'une analyse moléculaire, et qu'une anomalie avait été détectée chez 738 (57.8%). Cette anomalie était en relation avec une thérapie ciblée dans 143 cas, soit 19.3% des patients ayant une altération détectable, et 11.2% des patients ayant pu bénéficier d'une analyse moléculaire. En comparaison des patients ayant reçu un traitement standard, ceux ayant bénéficié d'une thérapie ciblée avaient une survie sans progression significativement plus longue ($p = 1^{e-3}$), bien que la médiane de gain soit modeste (11.4 contre 8.8 mois).

Similairement, dans SAFIR01, 407 des 423 patientes incluses ont pu être biopsiées, et cette biopsie était exploitable pour 283 (69.5%). Une anomalie moléculaire était détectée pour 195 patientes (47.9% des patientes biopsiées, et 68.9% des échantillons exploitables), et cette anomalie correspondait à une option thérapeutique pour 55 cas, soit 28.2% des patients avec une anomalie détectée, et 19.4% des patients ayant pu bénéficier d'une analyse moléculaire. 13 patients semblent avoir répondu, ou être resté stables, après un traitement ciblé, mais cet

essai n'ayant pas été randomisé, le bénéfice comparé au traitement standard n'a pu être mesuré (André et al. 2014).

De même, si l'étude MOSCATO (toujours en cours) a démontré une relative faisabilité de l'approche moléculaire, seuls 25 des 129 patients inclus (19.3%) ont bénéficié d'une thérapie ciblée, et 9 d'entre eux (7%) en ont tiré un bénéfice en terme de survie.

L'étude BATTLE a pu inclure 244 des 341 patients initialement recrutés (71.5%), et confirme les valeurs prédictives de meilleure survie pour certains marqueurs : mutation de EGFR pour le traitement par l'erlotinib, surexpression de VEGFR-2 pour le traitement par le vandetanib. Toutefois, les bénéfices en terme de survie, bien que significatifs, restent modestes, de l'ordre de 2 mois. D'autre part, l'algorithme adaptatif et l'absence de bras contrôle (traitement standard), rendent difficile la comparaison de ces résultats avec ceux d'autres études.

Ces résultats indiquent que, dans une pratique de routine, tous les patients ne bénéficient pas d'une analyse moléculaire. Lorsqu'elle est possible, cette analyse peut permettre une orientation thérapeutique, toutefois, seule une faible proportion de patients bénéficie d'une thérapie ciblée.
La difficulté d'accéder à des essais cliniques ciblés pour certaines anomalies moléculaires en est une cause possible (André et al. 2014). Mais, dans certains cas, les choix technologiques et les règles de décision pourraient être à reconsidérer.

Si des altérations moléculaires s'avèrent prédictives de l'efficacité de thérapies ciblées, certaines études ont aussi identifié certains contextes moléculaires dans lesquels plusieurs altérations sont apparues antagonistes : des mutations du gène KRAS et une amplification du gène MET chez les patient EGFR mutés réduisent les effets des anti-EGFR dans les tumeurs coliques et les NSCLC (Engelman et al.

2007; Cappuzzo, Varella-Garcia, et al. 2008). Dans ces cas d'antagonismes, il peut ne pas y avoir d'alternative thérapeutique.

De nombreuses études ont également pointé l'hétérogénéité tumorale et la sélection de nouvelles mutations comme des sources d'acquisition de résistances et des causes possibles d'échec des thérapies ciblées (Cottu et al. 2008; Arcila et al. 2011; Diaz et al. 2012; Shibata 2012; Bedard et al. 2013). Dans les cas d'antagonismes connus, il semble clair que ces altérations doivent être intégrées dans le panel d'anomalies recherchées. Mais la détection d'anomalies présentes dans des clones minoritaires pose le problème de la sensibilité des techniques de recherche, ainsi que des seuils à considérer pour retenir une altération comme devant être intégrée dans la décision thérapeutique. D'autre part, le choix d'un screening large, non restreint à des cibles - ou antagonismes – connues, pourrait permettre d'identifier d'autres associations ou sélections d'altérations responsables de l'échec des thérapies.

Enfin, les essais randomisés sont encore trop peu nombreux pour apporter une preuve indiscutable du bénéfice d'un choix thérapeutique guidé par le profil moléculaire : si les essais IMPACT et BATTLE montrent un bénéfice, même modéré, d'une prise en charge thérapeutique guidée par le screening moléculaire, ces résultats ne semblent pas être confirmés par d'autres études, comme l'essai SHIVA qui ne met en évidence aucun bénéfice d'une telle stratégie.

Certains oncologues, demeurant sceptiques, considèrent même que les informations cliniques pourraient rester un critère de choix pour la décision thérapeutique ; des orientations seulement prises sur la base d'anomalies moléculaires, pourraient priver des patients d'un traitement standard dont ils auraient potentiellement tiré bénéfice :

« Recently, the targeting of the PI3K/mTOR pathway using mTOR inhibitor in combination with exemestane has lead to a major improvement of clinical benefit in women with metastatic hormone receptor-positive and Her2neu negative breast cancer. It is noteworthy that this target was based on molecular research. The patients to benefit from this new treatment combination however were selected clinically: women who had developed endocrine resistance to prior endocrine treatment were selected and showed a remarkable clinical benefit rate. The clinical trials providing this evidence were successful without individual molecular identification of the treatment target. » Dr Dubsky.
Source : European Society for Medical Oncology. "First large scale trial of whole-genome cancer testing for clinical decision-making reported." ScienceDaily. ScienceDaily, 1 October 2012.
www.sciencedaily.com/releases/2012/10/121001084134.htm

« It is likely that the genetic and molecular abnormalities are very complex, and the assumption that we can implement a targeted approach because we can identify these abnormalities may put patients at risk of passing up standard treatments.. In some ways, it feels like we are on a rollercoaster that is a little out of control. »
Peter Ellis, MBBS, PhD, of McMaster University, Canada
Source :
http://www.ascopost.com/issues/january-15,-2013/french-investigators-prospectively-test-genomically-driven-treatment-in-metastatic-breast-cancer.aspx

3 Apports de l'aCGH en médecine de précision

3.1 Principe

L'analyse par aCGH a pour but d'analyser l'ADN d'un échantillon et d'y identifier des régions chromosomiques présentant des déséquilibres de nombre de copies. On fait généralement référence au résultat de l'analyse aCGH par le terme « profil génomique ».

3.1.1 Calcul des Log ratios relatifs

Quelque soit la technologie utilisée, hybridation compétitive ou hybridation simple, l'interprétation des signaux se fait toujours relativement à un ADN normal utilisé comme contrôle. Les signaux sont analysés après calcul des Log_2 ratios relatifs (LRR), comme suit :

$LRR_i = Log_2\left(\frac{S_{i,test}}{S_{i.ctrl}}\right)$, *où* $S_{i,test}$ *et* $S_{i,ctrl}$ *sont respectivement les signaux de la sonde i dans l'échantillon testé et dans l'ADN de contrôle.*

3.1.2 Segmentation

Une particularité majeure de l'aCGH est d'analyser des signaux liés par une relation physique : L'ADN est un support physique supposé intègre, au moins par parties. Des sondes hybridant des positions génomiques contigües, et présentes en même quantité, sont donc attendues avec des valeurs similaires, à la variabilité expérimentale près.

Pour mettre en évidence ces relations et identifier des anomalies de continuité, les LRR sont ordonnés selon leur position génomique, puis analysés à l'aide d'algorithmes de segmentation. Leur principe est d'identifier des points de cassure témoignant de changements de niveau dans les signaux. Ces points de cassure délimitent alors des segments à l'intérieur desquels le signal peut être résumé par la moyenne des LRR des sondes qu'ils contiennent.

Plusieurs méthodes de segmentation ont été développées, parmi lesquels on peut citer :

- *L'approche par modèle de Markov caché (ou Hidden Markov Models, HMM)* (Fridlyand et al. 2004) *estime les états cachés (nombre de copies) à partir de probabilités à priori et de matrices de transition (probabilité de passer d'un état à un autre) et d'émission (probabilité d'observer une valeur pour un état donné). Une fonction de pénalisation, généralement construite sur la log-vraisemblance, peut permettre d'optimiser le nombre d'états à considérer dans le modèle.*

- *L'algorithme GLAD (pour Gain and Loss Analysis of DNA)* (Hupé et al. 2004) *ramène la segmentation à un problème de régressions gaussiennes constantes locales. Les bornes des régions (ou segments) et leurs valeurs (les valeurs des régressions locales), sont déterminées par l'optimisation d'une fonction de vraisemblance pondérée.*
- *L'algorithme CBS (pour Circular Binary Segmentation)* (Olshen et al. 2004; Venkatraman & Olshen 2007) *est construit sur la recherche récursive de points de cassure par une méthode dérivée d'un l'algorithme de segmentation précédemment développé* (Sen & Srivastava 1975)*. Cette méthode et la statistique sur laquelle elle s'appuie seront plus largement décrites dans ce manuscrit.*
- *L'algorithme HaarSeg* (Ben-Yaacov & Eldar 2008) *implémente l'algorithme des ondelettes de Haar, développé pour des applications en analyse du signal* (Mallat 2008)*.*

Les performances de ces algorithmes ont été comparées à plusieurs reprises dans la littérature et ces études semblent donner la préférence à l'algorithme CBS (Willenbrock & Fridlyand 2005; Lai et al. 2005)*.*

3.1.3 Interprétation

Les régions anormales sont qualifiées de gagnées, ou amplifiées, lorsque leur nombre de copies est supérieur à 2 (nombre de copies attendu dans un ADN normal), et délétées lorsqu'il est inférieur à 2. Après calcul des ratios sonde-à-sonde, Log transformation et segmentation, les régions d'intérêt seront donc celles présentant un LRR > 0 pour les régions en gain, ou < 0 pour celles en perte (Figure 7).

L'analyse des gènes contenus dans des régions amplifiées ou délétées permet d'identifier des voies biologiques - ou pathways - potentiellement altérées, hyperactives dans le cas de gènes amplifiés, ou non fonctionnelles dans le cas de délétions.

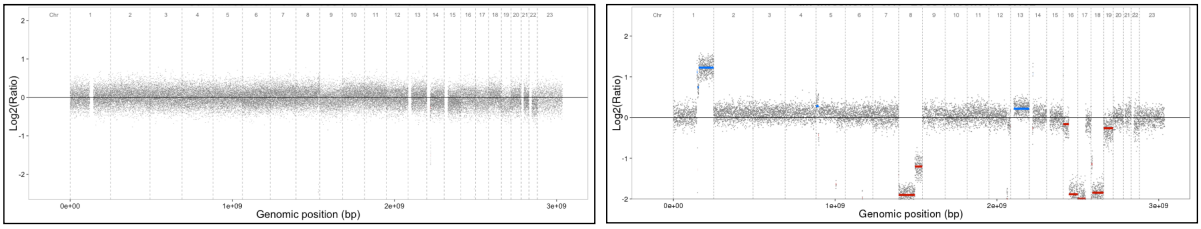


Figure 7 : Exemples de profils génomiques obtenus par aCGH. A gauche, aucune anomalie de nombre n'est détectée. A droite, l'analyse identifie plusieurs régions estimées comme gagnées (en bleu) ou délétées (en rouge). A noter que l'axe y représente les valeurs de ratios après Log_2 transformation, soit 0 = $Log_2(2/2)$.

En médecine de précision, la décision pour une orientation thérapeutique repose donc sur la présence de telles altérations de nombre dans le profil génomique de la tumeur d'un patient, et sur la disponibilité d'un traitement ciblant la fonction associée.

D'autres valeurs, telles que le nombre de points de cassure, peuvent aussi être extraites de ces profils pour participer à la construction d'indicateurs d'instabilité chromosomique (Russnes et al. 2010).

3.2 Avantages de l'aCGH en médecine de précision

Malgré l'intérêt grandissant pour le séquençage haut débit, et un nombre croissant de mutations génétiques identifiées, l'aCGH bénéficie encore de certains avantages :

- *Cette technologie est aujourd'hui mature, et les laboratoires en ont une bonne maîtrise technique.*
- *L'aCGH permet une analyse quantitative quasi exhaustive d'un ADN, pour un coût 4 à 5 fois inférieur à celui d'une quantification par séquençage.*
- *L'analyse des données pour l'obtention d'un profil génomique est extrêmement rapide, et ne requiert que peu de ressources informatiques.*

Il est intéressant de noter que, malgré l'intérêt croissant pour les techniques de séquençage à haut débit, les données microarray restent parmi les plus utilisées parmi les ressources TCGA disponibles (Figure 8).

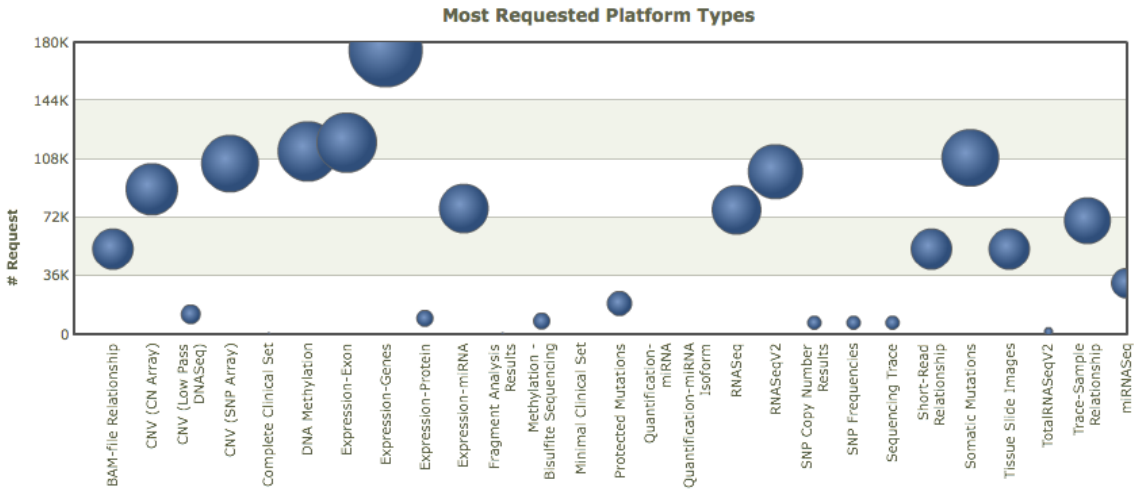


Figure 8 : Nombre de téléchargements des données publiques depuis le portail TCGA. Source https://tcga-data.nci.nih.gov/datareports/statsDashboard.htm, 16/03/2015.

Si le profil génomique permet de rechercher des régions, ou des gènes, montrant une altération du nombre de copies, il renseigne également sur la quantité de remaniements chromosomique présents. Cette forme d'instabilité génomique n'est pas, à ce jour, prise en compte pour l'orientation thérapeutique, mais de nombreuses études suggèrent une relation possible entre le nombre d'altérations, ou points de cassure dans l'ADN, et l'agressivité d'une tumeur (Russnes et al. 2010; Bonnet et al. 2012; Zheng et al. 2013; Vollan et al. 2015; How et al. 2015). Comprendre les phénomènes biologiques sous-jacents à l'apparition et la sélection de ces altérations pourrait permettre le développement de nouvelles prises en charge thérapeutiques (Lee et al. 2011; McGranahan et al. 2012).

3.3 Gènes d'intérêt et ambiguïtés pour la décision thérapeutique

Les amplifications du gène ERBB2 et leur relation avec une meilleure sensibilité aux inhibiteurs ont largement été démontrées. Comme rappelé plus haut, la protéine HER2/neu est parmi les premiers biomarqueurs identifiés, et son intérêt en décision thérapeutique est très largement reconnu, en particulier dans le cancer du sein (Slamon et al. 2001; Vogel et al. 2002; Piccart-Gebhart et al. 2005; Romond et al. 2005). Certaines études récentes semblent indiquer que des amplifications de ERBB2 pourraient aussi être un marqueur pronostic dans d'autres types de tumeurs, tels que les tumeurs coliques (Pectasides & Bass 2015).

Le gène c-MET est un proto-oncogène identifié au début des années 80 comme étant impliqué dans les proliférations tumorales (Cooper et al. 1984). Ce gène code pour une protéine à activité tyrosine kinase. Mettre en évidence une amplification de ce gène semble avoir des intérêts multiples : elle pourrait prédire une sensibilité à certains inhibiteurs de tyrosine kinases (Smolen et al. 2006), mais aussi une résistance aux anti-EGFR (Engelman et al. 2007). Cependant, une récente étude remet en cause l'intérêt de rechercher une amplification de MET dans une visée de décision thérapeutique : la forme phosphorylée de la protéine pourrait être mieux corrélée à son activation qu'une amplification ou une surexpression du gène (Watermann et al. 2015).

MDM2 est un proto-oncogène codant pour une protéine capable de bloquer l'action de certains gènes suppresseurs de tumeur tel que p53. Son amplification peut être retrouvée dans de nombreux types de tumeurs, mais la prévalence est particulièrement élevée dans les tumeurs des tissus mous (Momand et al. 1998). L'amplification de ce gène a été identifié comme un facteur péjoratif de survie (Forslund et al. 2008), et des inhibiteurs de l'interaction MDM2/p53 font l'objet de plusieurs essais cliniques (Iancu-Rubin et al. 2014; Zhao et al. 2015; Patnaik et al. 2015).

PTEN est un gène suppresseur de tumeur codant pour une protéine à activité phosphatase participant à la stabilité chromosomique, et antagoniste du réseau PIK3/AKT/mTOR. La perte de ce gène est un facteur péjoratif de survie (Hollander et al. 2011; Stern et al. 2015), et incite à orienter les patients vers des thérapies ciblant les protéines de ce réseau (Dillon & Miller 2014).

Le gène PIK3CA code pour une des protéines de la famille des phosphoinositide-3-kinases : un ensemble de protéines clé des cascades de signalisation intracellulaires, et impliquées dans la croissance, la prolifération, et la survie cellulaire. Les amplifications de PIK3CA ont été décrites comme associées aux proliférations tumorales, en particulier dans l'ovaire (Shayesteh et al. 1999). Cette amplification semble une possible cause de résistance aux chimiothérapies, voire aux thérapies ciblées anti-PIK3 dans les tumeurs de l'ovaire et du sein (Kolasa et al. 2009; Huw et al. 2013), mais a aussi été décrite comme un facteur prédictif de survie dans les tumeurs coliques (Jehan et al. 2009).

Le gène IGF1R code pour un récepteur tyrosine kinase appartenant à la famille des récepteurs insuline, et impliqué dans la prolifération tumorale (Larsson et al. 2005). Les amplifications de ce gène semblent être des évènements rares, mais pouvant être associés à une surexpression de la protéine (Ribeiro et al. 2014). Si les thérapies ciblées n'ont pas montré de réels bénéfices (Jin et al. 2013), des essais de phase 1, en combinaisons avec d'autres inhibiteurs, pourraient être plus prometteurs (Wilky et al. 2015).

Le gène FGFR1 code pour un récepteur de surface à activité kinase. Des études précliniques ont semblé indiquer que des amplifications de FGFR1 pouvaient être associées à une sensibilité aux inhibiteurs de FGFR, et que ces anomalies avaient une prévalence élevée dans les tumeurs du sein et les carcinomes squameux du poumon (Weiss et al. 2010; Dutt et al. 2011; Gozgit et al. 2012; Jain & Turner

2012). Si d'autres études n'ont pas confirmé l'intérêt de ces amplifications pour la décision thérapeutique, il est toutefois possible que le seuil définissant une amplification pertinente soit à spécifier (André et al. 2013; Wynes et al. 2014).

De même, les amplifications de EGFR pourraient être prédictives d'une réponse aux anti-EGFR (Ålgars et al. 2011). Mais là encore, ces résultats semblent controversés : des patients sans amplification pourraient aussi répondre à ces inhibiteurs (Chung et al. 2005), et d'autres études ne confirment pas a valeur prédictive de ce marqueur (Tsao et al. 2005) *; d'autres critères pourraient être nécessaires pour mieux anticiper la réponse à cette classe d'inhibiteurs ; il n'est pas exclu que la méthode la mieux adaptée pour rechercher une amplification EGFR soit à préciser (Sesboüé et al. 2012; Hutchinson et al. 2015).*

Le gène ESR1 code pour un récepteur aux estrogènes (RE), situé à la membrane des cellules. Complexé à son ligand, RE peut s'associer à d'autres complexes protéiques pour réguler l'expression d'autres gènes, ou activer des facteurs de transcriptions par l'intermédiaire de voies kinases. La valeur prédictive des amplifications de RE dans les hormonothérapies (Holst et al. 2007) motive la recherche de cette anomalie moléculaire. Cependant, des divergences existent quant au choix de la technique d'investigation à utiliser : la recherche par FISH semble plus sensible que l'analyse par aCGH (Yu & Shao 2011).

D'autres gènes pourraient, dans l'avenir, être aussi pris en considération, même en l'absence d'inhibiteurs spécifiques.
Il a été démontré que des amplifications du gène MYC - un oncogène dont la protéine interagit avec les réseaux PIK3/AKT et MAP kinases - ont un impact global sur la machinerie transcriptionnelle des cellules (Lin et al. 2012; Lovén et al. 2012). Les amplifications de MYC pourraient aussi être responsables de résistances à certaines thérapies ciblées comme la rapamycine, un anti-mTOR (Ilic et al. 2011).

Le gène cyclin-D1 (CCND1) est impliqué dans la régulation du cycle cellulaire, par l'intermédiaire de kinases cycline-dépendantes, CDK4/CDK8. Des amplifications de ce gène pourraient orienter vers des thérapies ciblant ces CDK (Musgrove et al. 2011; Rihani et al. 2015).

3.4 Limites à l'analyse aCGH en médecine de précision

Bien que possédant de nombreux avantages – coût, maîtrise technique, rapidité d'analyse – l'analyse de profils génomiques par aCGH a aussi des faiblesses :

3.4.1 Pas d'identification des anomalies équilibrées

Par construction, l'analyse par aCGH ne peut identifier que des anomalies de nombre : pertes ou gains de régions génomiques. Par nature, les altérations équilibrées, sans perte ou gain d'ADN, telles que les réarrangements et les fusions, ne sont pas mises en évidence par cette technique (Weiss et al. 1999).

3.4.2 Problème de mélange

L'ADN analysé est extrait à partir d'une biopsie tissulaire contenant de la tumeur, mais également du tissu normal. En outre, la tumeur elle-même n'est pas une entité homogène, mais est possiblement composée de plusieurs clones ne partageant pas tous les mêmes altérations, certaines pouvant être mutuellement exclusives (Das et al. 2014). Selon certains auteurs, des altérations présentes dans ces mosaïcismes ne sont détectables que lorsqu'elles sont portées par au moins 10 à 30% des cellules (Neill et al. 2010; Valli et al. 2011).

Ce problème du mélange de populations cellulaires - tissu normal, mosaïcisme - induit des effets de dilution pouvant impacter l'amplitude des signaux, et potentiellement induire des faux-négatifs : la non détection d'altérations présentes, mais à des niveaux trop faibles pour être observées.

3.4.3 Pas de seuils de décision clairs

La décision de retenir une anomalie de nombre de copies comme pertinente repose essentiellement sur l'amplitude des signaux, attendue comme

représentative du niveau d'amplification ou de perte d'une région, et par conséquent des gènes qu'elle contient. Or, il n'existe pas, à notre connaissance, de consensus sur les valeurs seuils de gain (ou perte) traduisant un réel effet biologique, et pouvant justifier une décision thérapeutique.

De même, la taille des segments anormaux, ainsi que leur incidence sur des dysfonctionnements potentiellement actionnables, est discutée : si certains auteurs préfèrent se concentrer sur des altérations de petite taille, dites focales (Mermel et al. 2011), d'autres ont montré que des altérations plus longues, incluant des régions non codantes et des promoteurs, pouvaient impacter l'expression des gènes (Miyaguchi et al. 2011).

L'absence de règles de décisions claires sur les valeurs de segments et leurs longueurs, pourrait être à l'origine des ambiguïtés concernant l'intérêt de considérer certains gènes tels que FGFR1 et EGFR, comme discuté précédemment.

3.4.4 Limites des outils existants

Plusieurs plateformes de microarray sont disponibles, et chaque fabricant propose des outils d'analyse et de visualisation, dédiés à leur propre technologie. Agilent propose Agilent CytoGenomics, un ensemble de solutions supportées par les principaux systèmes d'exploitation. Ces outils ne peuvent toutefois pas être installés sur un système linux. Les suites logicielles Affymetrix et Illumina, quant à elles, ne sont développées que pour les systèmes Windows.

Plusieurs auteurs ont proposé des outils d'analyse, indépendants et portables, mais ceux-ci sont parfois dédiés à des plateformes spécifiques, et peuvent manquer de flexibilité. En outre, aucun de ces outils ne propose de visualisations adaptées à la prise de décision en médecine de précision, et pouvant être partagées pour des discussions en comité scientifique.

CGHcall (Van De Wiel et al. 2007) combine une approche bayesienne et l'analyse de mélanges gaussiens pour estimer les probabilités à posteriori des altérations.

Les segments sont ensuite assignés à 3 ou 4 classes : perte, normal, gain, amplification, les 2 dernières pouvant être concaténées en une seule selon les spécifications de l'utilisateur.
Limites : CGHcall nécessite une cohorte d'échantillons pour redéfinir les régions de gains, amplification et pertes, et ne peut pas être appliqué sur un échantillon unique.

CGHnormaliter (van Houte et al. 2009) s'appuie sur une méthode itérative de normalisation des signaux : après segmentation, les valeurs de sondes définies comme 'normales' sont utilisées pour ajuster l'ensemble du profil avant une nouvelle segmentation. L'ensemble du processus est répété jusqu'à atteindre un critère d'arrêt, ou un nombre maximal d'itérations spécifié par l'utilisateur.
Limites : La méthode définissant des sondes dites 'normales' n'est pas explicitée. De plus, CGHnormaliter n'est implémenté que pour des données issues d'hybridations à 2 couleurs, e.g. données Agilent.dual-color.

GAP (Popova et al. 2009) recherche des clusters de valeurs dans un plan défini par les fréquences alléliques (B-allele frequency, BAF) en abscisse, et les LRR en ordonnée (Figure 9).

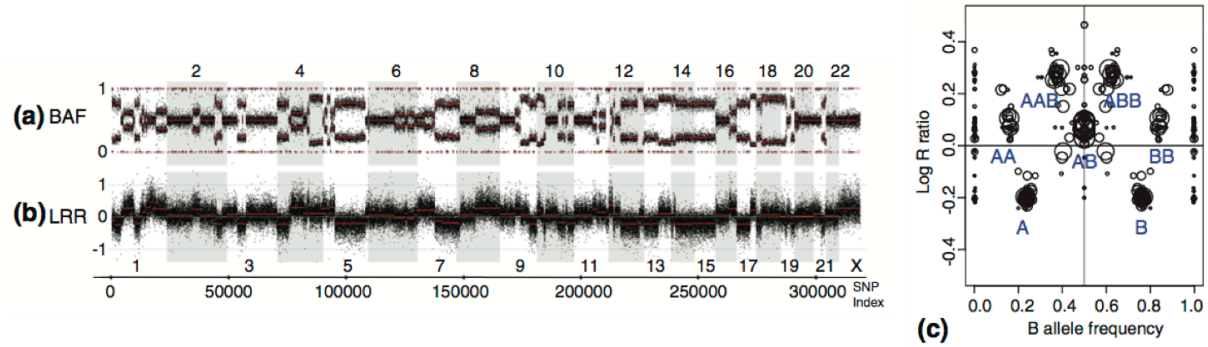


Figure 9 : dans GAP, les fréquences alléliques (a) et les LRR après segmentation (b) sont projetés dans un plan pour la recherche de clusters (c). Les clusters proches de BAF=0.5, LRR=0 sont assignés au statut 2-copies, hétérozygote AB. Les autres statuts sont déduits de cette position.

Les BAF, sont initialement calculés à partir des signaux des sondes SNP : $BAF_i = \frac{B_i}{A_i + B_i}$ *, où* A_i *et* B_i *sont respectivement les signaux des allèles A et B pour le snp i.*

GAP propose une modélisation des BAF prenant en considération de possibles effets de dilution dus à la présence de tissu normal dans l'échantillon :

$$BAF = \frac{(1-p).n_B^T + p.n_B^N}{(1-p).\left(n_B^T + n_A^T\right) + 2p}$$

où p est la proportion d'ADN normal dans l'échantillon - de fait, (1-p) représente la proportion d'ADN tumoral dans l'échantillon, n_A^T *et* n_B^T *sont respectivement les valeurs des allèles A et B dans la tumeur, et* n_B^N *la proportion de l'allèle B dans le tissu normal contaminant.*

De même, les LRR sont modélisés comme une fonction linéaire dont la pente est un coefficient de 'contraction' q, représentant les effets de dilution liés à la présence possible de clones ne partageant pas les mêmes altérations :

$LRR_n = q.\log_2\left(\frac{n}{2}\right) = q.LRR_n$ *, où n est le nombre de copies.*

Limites : construite sur les scores BAF, cette méthode n'est applicable que si ces valeurs peuvent être calculées, c'est à dire pour des microarrays proposant des sondes dédiées aux SNP. En outre, si le code est disponible sur demande, il n'est, à ce jour, pas formalisé sous la forme d'un package pouvant être distribué.

PAIR (Yang et al. 2013) propose également une recherche du niveau d'équilibre et une normalisation construites sur les pertes d'hétérozygoties estimées à partir des sondes SNP.

Limites : comme GAP, PAIR n'est applicable que sur des données issues de microarrays proposant des sondes SNP. De plus, cette méthode requiert les signaux obtenus à partir du tissu normal, apparié à l'échantillon analysé.

GISTIC (Mermel et al. 2011) est un algorithme développé par le Broad Institute, et largement exploité dans les analyses du TCGA. Cette approche itérative complexe vise à reconstruire la segmentation d'un ensemble de profils à partir d'un modèle de bruit, ce modèle étant optimisé à chaque itération. Après reconstruction, GISTIC estime la significativité d'altérations focales, définies comme des régions de forte amplitude, identifiées de manière récurrente au sein de la cohorte analysée.
Limite : cet algorithme est destiné à l'analyse de cohortes et semble difficile à évaluer dans le cadre de l'analyse de profils uniques.

popLowess (Staaf et al. 2007) propose une optimisation de centralisation par un ajustement des biais de cyanines, dans le cas d'hybridations utilisant 2 fluorochromes. Une régression locale est appliquée sur la population majoritaire, définie comme telle après une classification des LRR par la méthode des Kmeans à 3 centres. La médiane des valeurs ainsi ajustées, est ensuite utilisée pour corriger l'ensemble du profil.
Limites : cette méthode n'est applicable qu'aux hybridations utilisant 2 fluorochromes, e.g. plateforme Agilent, et se focalise sur la population majoritaire, alors considérée comme une population de référence à 2 copies.

4 Objectifs

Optimiser la paramétrisation des étapes nécessaires à l'analyse d'un profil génomique et intégrer ces solutions dans un workflow complet d'analyse aCGH dédié à la médecine de précision, portable et diffusable.

Ce workflow devra, en outre, être capable de prendre en charge les données issues des principales plateformes microarrays, et proposer des solutions facilitant la traçabilité, la diffusion des résultats, ainsi que leur discussion en comité pour une orientation thérapeutique, à travers des outils de visualisation flexibles et innovants.

5 Résultats

5.1 Le problème de la centralisation

Définir le niveau d'équilibre est essentiel à la décision puisque c'est à partir de cette ligne de base que les gains et pertes d'ADN seront estimés.
Plusieurs stratégies sont possibles pour cette étape, et certaines sont décrites dans les algorithmes précédemment présentés : une approche immédiate est de centrer l'ensemble des valeurs LRR sur leur médiane (Picard et al. 2005). Une autre approche consiste à évaluer la densité de distribution des LRR comme un mélange de populations gaussiennes, puis de choisir un pic de densité comme représentatif du niveau d'équilibre à 2 copies (Chen et al. 2008). Mais dans ce dernier cas, se pose le problème du choix. Ces auteurs proposent de considérer la population ayant le pic de plus haute densité comme représentative du niveau d équilibre. Cependant, si ce pic correspond aux valeurs les plus observées, dans les cas d'échantillons majoritairement aneuploïdes, cette région pourrait ne pas correspondre à un ratio neutre de 2/2, mais à un ratio relatif à la population de ploïdie plus élevé.

Dans le contexte des programmes de screening moléculaires, où l'analyse porte sur un profil unique, et la proposition d'orientation thérapeutique s'appuie sur l'identification d'anomalies actionnables, il nous a semblé opportun de souligner l'importance de la centralisation des profils, et l'effet de différentes stratégies sur la prise de décision.
Nous montrons, à partir de données issues de lignées cellulaires caractérisées pour leurs altérations génomiques, puis sur des données issues d'échantillons tumoraux analysés dans le contexte de programmes SAFIR-01 et MOSCATO, que le choix de la centralisation n'est pas un problème trivial, et qu'il peut avoir un impact important sur l'orientation des patients vers des thérapies ciblées (Commo et al. 2015).

5.1.1 Article : Impact of centralization on aCGH-based genomic profiles for precision medicine in oncology

Les méthodes et résultats supplémentaires sont présentés en annexe 2.

7. Therasse P, Arbuck SG, Eisenhauer EA et al. New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst 2000; 92: 205–216.
8. Brookmeyer R, Crowley J. A confidence interval for the median survival time. Biometrics 1982; 38: 29–41.
9. Larkin J, Del VM, Ascierto PA et al. Vemurafenib in patients with BRAF(V600) mutated metastatic melanoma: an open-label, multicentre, safety study. Lancet Oncol 2014; 15: 436–444.
10. Goldinger SM, Zimmer L, Schulz C et al. Upstream mitogen-activated protein kinase (MAPK) pathway inhibition: MEK inhibitor followed by a BRAF inhibitor in advanced melanoma patients. Eur J Cancer 2014; 50: 406–410.
11. Ackerman A, Klein O, McDermott DF et al. Outcomes of patients with metastatic melanoma treated with immunotherapy prior to or after BRAF inhibitors. Cancer 2014; 120: 1695–1701.
12. Ascierto PA, Simeone E, Sileni VC et al. Sequential treatment with ipilimumab and BRAF inhibitors in patients with metastatic melanoma: data from the Italian cohort of the ipilimumab expanded access program. Cancer Invest 2014; 32: 144–149.
13. Sosman JA, Kim KB, Schuchter L et al. Survival in BRAF V600-mutant advanced melanoma treated with vemurafenib. N Engl J Med 2012; 366: 707–714.
14. McArthur GA, Chapman PB, Robert C et al. Safety and efficacy of vemurafenib in BRAF(V600E) and BRAF(V600K) mutation-positive melanoma (BRIM-3): extended follow-up of a phase 3, randomised, open-label study. Lancet Oncol 2014; 15: 323–332.
15. Becker JC, Andersen MH, Hofmeister-Muller V et al. Survivin-specific T-cell reactivity correlates with tumor response and patient survival: a phase-II peptide vaccination trial in metastatic melanoma. Cancer Immunol Immunother 2012; 61: 2091–2103.
16. Hill GJ, Krementz ET, Hill HZ. Dimethyl triazeno imidazole carboxamide and combination therapy for melanoma. Cancer 1984; 53: 1299–1305.
17. Menzies AM, Haydu LE, Visintin L et al. Distinguishing clinicopathologic features of patients with V600E and V600K BRAF-mutant metastatic melanoma. Clin Cancer Res 2012; 18: 3242–3249.
18. Larkin J, Ascierto PA, Dréno B et al. Combined vemurafenib and cobimetinib in BRAF-mutated melanoma. N Engl J Med 2014; 371: 1867–1876.
19. Robert C, Karaszewska B, Schachter J et al. Improved overall survival in melanoma with combined dabrafenib and trametinib. NEJM 2014; 372: 30–39.
20. Long GV, Stroyakovskiy D, Gogas H et al. Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma. N Engl J Med 2014; 371:1877–1888.

Annals of Oncology 26: 582–588, 2015
doi:10.1093/annonc/mdu582
Published online 23 December 2014

Impact of centralization on aCGH-based genomic profiles for precision medicine in oncology

F. Commo[1,2,†], C. Ferté[1,2,3,†], J. C. Soria[2,3], S. H. Friend[1], F. André[2,3] & J. Guinney[1*]

[1]*Sage Bionetworks, Seattle, USA;* [2]*INSERM U981, Gustave Roussy, University Paris XI, Villejuif;* [3]*Department of Medical Oncology, Gustave Roussy, Villejuif, France*

Received 17 October 2014; revised 12 December 2014; accepted 16 December 2014

Background: Comparative genomic hybridization (CGH) arrays are increasingly used in personalized medicine programs to identify gene copy number aberrations (CNAs) that may be used to guide clinical decisions made during molecular tumor boards. However, analytical processes such as the centralization step may profoundly affect CGH array results and therefore may adversely affect outcomes in the precision medicine context.

Patients and methods: The effect of three different centralization methods: median, maximum peak, alternative peak, were evaluated on three datasets: (i) the NCI60 cell lines panel, (ii) the Cancer Cell Line Encyclopedia (CCLE) panel, and (iii) the patients enrolled in prospective molecular screening trials (SAFIR-01 $n = 283$, MOSCATO-01 $n = 309$), and compared with karyotyping, drug sensitivity, and patient-drug matching, respectively.

Results: Using the NCI60 cell lines panel, the profiles generated by the alternative peak method were significantly closer to the cell karyotypes than those generated by the other centralization strategies ($P < 0.05$). Using the CCLE dataset, selected genes (ERBB2, EGFR) were better or equally correlated to the IC50 of their companion drug (lapatinib, erlotinib), when applying the alternative centralization. Finally, focusing on 24 actionable genes, we observed as many as 7.1% (SAFIR-01) and 6.8% (MOSCATO-01) of patients originally not oriented to a specific treatment, but who could have been proposed a treatment based on the alternative peak centralization method.

*Correspondence to: Dr Justin Guinney , Sage Bionetworks, 1100 Fairview Ave. N., mail-stop M1-C108, Seattle, WA 98109, USA. Tel: +1-206-667-2146; Email: justin.guinney@sagebase.org

†These authors contributed equally to this work.

Annals of Oncology original articles

Conclusion: The centralization method substantially affects the call detection of CGH profiles and may thus impact precision medicine approaches. Among the three methods described, the alternative peak method addresses limitations associated with existing approaches.

Key words: precision medicine, comparative genomic hybridization, aCGH, targeted therapy

introduction

The detection of gene copy number aberrations (CNAs) by array-based comparative genomic hybridization (aCGH) is extensively used to decipher the molecular landscape of tumors in modern scientific programs [e.g. The Cancer Genome Atlas, Cancer Cell Line Encyclopedia (CCLE), the Cancer Genome Project] [1–3]. Combined with the identification of gene mutations, aCGH profiling is also part of precision medicine programs (e.g. SAFIR-01, MOSCATO-01, WINTHER) [4–6], guiding the prescription of a number of molecular targeted agents [7–9]. However, the rules to define amplifications from aCGH profiles are still unclear. Particularly, amplifications are related to signal magnitudes from a baseline, considered as a neutral (or 2*n*) copy numbers (CNs). Therefore, this baseline appears to be fundamental in genomic analysis of CN alterations, and may have profound consequences on the use of aCGH in precision medicine programs.

Regarding the aCGH analysis framework itself (supplementary Figure S1, available at *Annals of Oncology* online), most attention has been focused on the development of highly efficient segmentation algorithms, such as the circular binary segmentation (CBS) or hidden Markov models [10, 11], and on identifying significant regions of interest, such as GISTIC [12]. However, the importance of the centralization step is often underestimated. Its aim is to adjust the entire profile on a zero value, which is to facilitate comparisons across samples based using a neutral level (a normal 2 copies count), from which DNA fragments will be defined as gained or lost. Therefore, it is a crucial step that may affect decision-making criteria in the matching of genomic aberrations with targeted therapies.

A commonly used strategy consists in centralizing the LogR on their mean or their median [13]. Several more elaborated methods have been proposed, and are included in global analysis pipelines. CGHcall [14] uses a supplementary post-segmentation centralization. CGHnormaliter [15] performs an iterative normalization method, where the centralization and imbalances are optimized by a repeated two-step procedure. The popLowess algorithm suggests an efficient alternative for adjusting bias due to cyanines, when two-channel hybridizations are used [16]. Two other algorithms, PAIR and genome alteration print, dramatically differ from the others since they use snp probes signals to infer the tumor ploidy [17, 18]. Unfortunately, this latter information is not available on all the platforms (e.g. Agilent platforms), which precludes the use of these methods on all aCGH arrays. A detailed discussion of these methods is beyond the scope of this article and has been addressed comprehensively elsewhere [14–18].

An interesting approach models the LogR as a mixture of several Gaussian distributions: after estimating the parameters of the mixture, the mean of the highest peak, ~95% of the main density peak, is used as a centralization value [19]. Exploring the LogR densities gives a good overview of the different levels of imbalances. However, using this method for choosing the right profile centralization value appears frequently not trivial, since there is not always only one clear and unambiguous peak density choice (supplementary Figure S1, available at *Annals of Oncology* online). Moreover, the main density peak corresponds to the region of the most commonly observed values. In case of predominantly aneuploid samples, this region would not represent a neutral 2/2 copies ratio, but rather a higher ratio, relative to the main sample ploidy. For this reason, we are introducing a new central value estimator, which we called the alternative centralization. This approach, based on the LogR density analysis, uses a more flexible rule in order to capture the remaining 2-copies population, when exists, and use it as a possibly more accurate adjustment value.

By comparing this rule to standard approaches, we investigated in this work how different centralization methods influence on the genomic profiles, and thus impact the decision making in precision medicine programs. We first applied different centralization strategies on the NCI60 cell lines panel and evaluated each approach by comparing the corresponding genomic profiles with the expected values deduced from the karyotypes. Next, we processed a large panel of cell lines, labeled for drug sensitivity, and correlated gene CNs with drug sensitivities for their recognized companion actionable genes. Finally, we described the impact of these centralization methods on the identification of actionable genes using patients' data from two prospective molecular screening trials (SAFIR-01 and MOSCATO-01, NCT01414933, and NCT01566019, respectively).

patients and methods

karyotypes

The NCI panel karyotypes information was obtained from SKY/M-FISH and CGH Database [20, 21]. We generated genomic-like profiles from the karyotypic annotations as follows: for each cell line, each fully annotated segment count was used as an estimate for the corresponding region CN. Not fully annotated segments (missing start and/or end cytoband) were not considered. Data were then transformed into $Log_2(CN/2)$. The python script is available at http://nbviewer.ipython.org/gist/fredcommo/9334224. Among the NCI60 cell lines, 57 of 60 had both aCGH data and karyotype with sufficient information to reconstruct a profile. In case of replicates, only the best aCGH profile was considered (lowest derivative Log_2 Ratio spread).

cell lines panels and patients datasets

The NCI60 NimbleGen Whole Genome 385K microarray data were downloaded from Gene Expression Omnibus (id: GSE30291). These data represented 71 aCGH experiments carried out on 60 individual cell lines.

The aCGH SAFIR01 Affymetrix-snp6 CEL files ($n = 125$) and Agilent-4 × 180K FE files ($n = 158$) were downloaded from Synapse (https://www.synapse.org/#!Synapse:syn2286494) [4].

The MOSCATO-01 Agilent-4 × 180K FE files ($n = 309$) were downloaded from Gene Expression Omnibus (GEO id under process).

The CCLE Affymetrix snp6 CEL files and drug responses were downloaded from http://www.broadinstitute.org/ccle/home. We only considered the 487 cell lines for which responses for the 24 explored compounds were available.

processing the aCGH profiles

To estimate the NCI60 aCGH-based genomic profiles, Log_2(Cy3/Cy5) were computed from the provided paired files, after cyanine bias correction and GC% adjustment.

For the MOSCATO-01 Agilent FE files, LogR were computed using the two-channel intensities, after cyanine bias correction and GC% adjustment.

The SAFIR01 and the CCLE Affymetrix snp6 CEL files were preprocessed using the Affymetrix Genotyping Console, version 4.1.4.840.

In all cases, genomic profiles were generated from LogR, using the same pipeline.

expectation–maximization optimization

In each case, the LogR distribution was modeled as a mixture of several Gaussian variables, with potentially different mean and standard deviation. The parameters of the Gaussian mixture were estimated using an expectation–maximization (EM) algorithm [22] using the R package mclust [23]. Then, the centralization values were chosen using two different strategies, as follow: (i) maximal centralization: the centralization value was defined as the mean of the major density peak, (ii) alternative centralization: the centralization value was defined as the mean of the major-left peak, if its maximum density was at least 50% of the major peak, and at a distance of at least 0.14, in LogR. Our strategy was to increase the tolerance for choosing a minor population, when compared with the 95% threshold suggested in Chen et al. [19]: a lower threshold would catch neutral ratios related to 2-copies DNA segments, in case of a predominantly aneuploid sample. A distance of at least 0.14 from the maximum peak, was added as a supplementary criterion, and was deduced from preliminary tests more extensively described is supplementary Methods, available at *Annals of Oncology* online.

In parallel, LogR medians were considered as the centralization values.

In order to increase the efficacy of the EM algorithm on large sets of values, we applied a resampling strategy, as described in the supplementary Material, available at *Annals of Oncology* online, section (supplementary Methods, available at *Annals of Oncology* online, and https://github.com/fredcommo/EMnormalize for R code).

For each sample, LogR were adjusted by subtracting each centralization values, separately, and segmented using the CBS algorithm, with the appropriate parameters.

In order to minimize differences with karyotype-based profiles, and because of their low resolution compared with aCGH, segments with lengths lower than 200 markers were deliberately merged with the closest segment, considering the previous and next segment LogR value.

distances from karyotypes

As changing the correction value leads to a simple translation of the entire vector of values, comparing the centralization approaches using correlations between array-based genomic profiles and their corresponding karyotypes would not be an appropriate comparison. Instead, we computed gene-by-gene squared distances with the reconstructed karyotype profile, considered as a reference profile. Mean squared distances from karyotypes were then compared across the different centralization methods by using paired *t*-tests, after log-transformation.

correlations with drug responses

Spearman correlations between CNs and drug responses (active area scores) were computed for a selected panel of four actionable genes (ERBB2, EGFR, FGFR1, and MET), and their related inhibitor (lapatinib, erlotinib, TKI258, and PHA665752, respectively). Significance of differences between correlations was evaluated after Fisher *Z*-transformation of the correlation values.

decisions in patient cohorts

For the SAFIR-01 and the MOSCATO-01 data, we focused on the 24 actionable genes used in André et al. [4]. Since nearly all the actionable CN alterations today are amplifications, only amplifications were depicted here, and calls were defined according to the criteria previously published in this same paper.

results

comparison of the CGH profiled centralization methods using a panel of cell lines with known karyotypes (NCI60)

To investigate how the centralization impacts on the profiles accuracy, we first analyzed a panel of cell lines for which the karyotypes were available. When using LogR densities, an alternative choice for centralizing the profile occurred in 18/57 cases (31.6%), mainly on the 3n (57.9%) and the 4n (38.5%) cell lines. Conversely, an alternative adjustment was detected for only 2 of the 22 cell lines with 2n− to 2n+ ploidies (8.3%). No other choice than the maximum peak was observed for the unique 5n −/+ SF-295 cell line (supplementary Figure S2, available at *Annals of Oncology* online).

Focusing on the 18 cases where an alternative centralization was available, mostly the 3n and 4n cell lines, we observed that using the maximum peak or the LogR median for adjusting the profiles was unable to detect imbalances revealed by the karyotypes: in case of aneuploidy, entire chromosomes, or chromosome arms, in numbers corresponding to the main cell line ploidy on karyotypes, appeared as in neutral counts, i.e. 2 copies, on the genomic profiles. In these cases, 2-copy DNA regions on karyotypes appeared as lost on the same genomic profiles (supplementary Figure S3, available at *Annals of Oncology* online). In such cases, the alternative centralization resulted in more consistent profiles with the karyotypes, for 17/18 and 18/18 cell lines compared with the maximum peak and the median centralization, respectively. These results were confirmed by paired *t*-tests on mean squared distances between profiles and their corresponding reconstructed karyotype ($P = 1.13e-4$ and 6.35e−5 for the same comparisons, respectively) (Figure 1). Interestingly, the 5n−/+ SF-295 cell line did not show any alternative, as previously defined, for adjusting the profile. That said, none of the possible choices would have lead to a genomic profile consistent with this cell line karyotype (supplementary Figure S4, available at *Annals of Oncology* online).

comparison of the outputs of different centralization methods on large panels of cell lines labeled for drug sensitivity (CCLE)

Applying similar comparisons on the CCLE data, we first noted that of 995 cell lines' profiles, an alternative peak was detected in

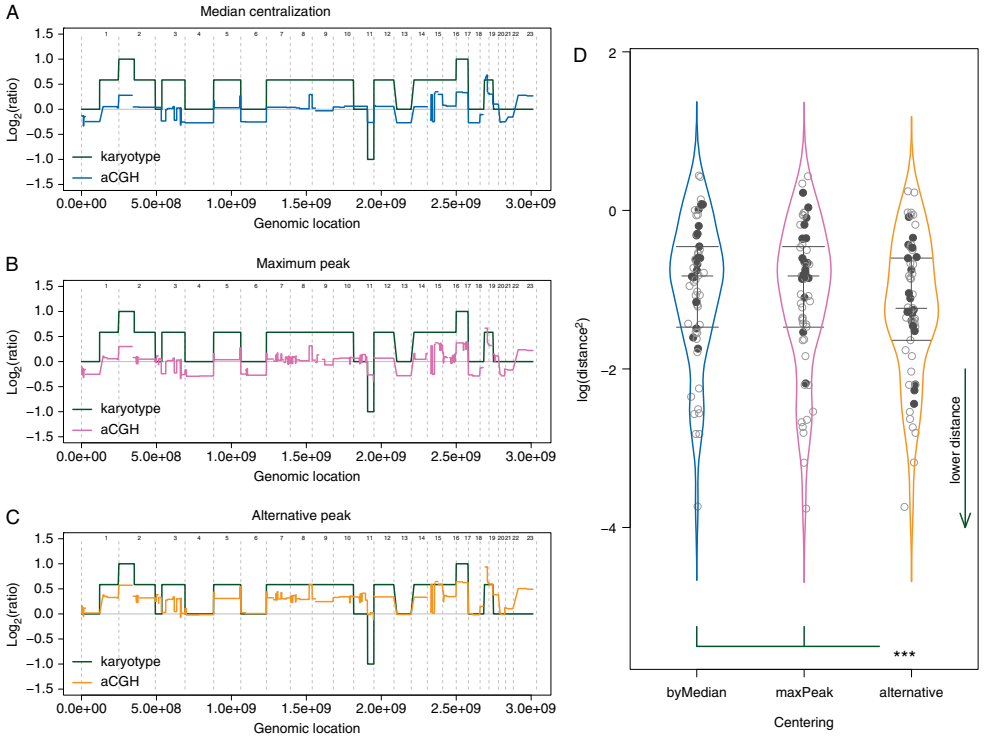
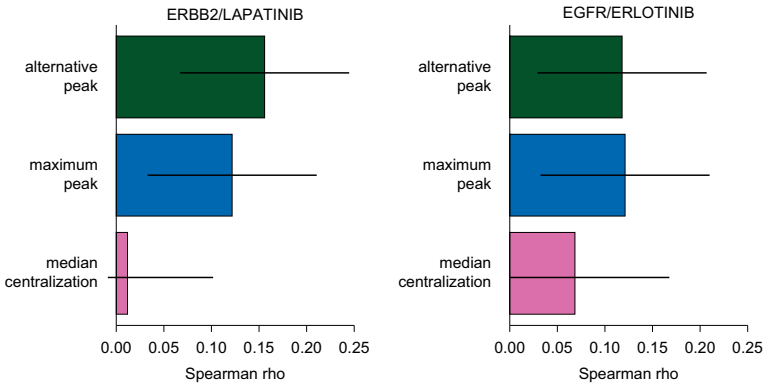


Figure 1. Distance from karyotypes. (A–C) In order to estimate the effect of the centralization methods, squared distances between genomic profiles (colored line) and karyotypes (green line) have been calculated. Distances are symbolized by the colored areas. (D) An alternative peak was available for 18 of the 57 analyzed cell lines (bold black points). Choosing the alternative peak for adjusting the genomic profiles significantly reduce the discrepancies with the corresponding karyotypes, compared with the other methods. $P = 1.13e{-}4$ and $6.35e{-}5$, compared with the maximum peak and the LogR median, respectively. Vertical colored curves represent the densities, and horizontal gray segments are the the Q25, Q50, and Q75 quantiles of each distribution.

160 cases (16%), and in 92 of the 487 sub-panel cell lines (18.9%) tested for drug sensitivities. To assess the impact of different centralization methods, we selected this latest sub-panel, and computed Spearman's correlations between centralized CN values and drug sensitivities. We focused on four genes for which the amplification is known to be associated with an increased sensitivity to the related inhibitor and are currently used in the clinic. For ERBB2 and lapatinib, the alternative and the maximum peak centralization both increased significantly the correlation when compared with the median method ($P = 0.012$ and $P = 0.043$, respectively). Further, regarding EGFR and erlotinib, both the maximum peak and the alternative peak tended to be associated with higher corrrelations when compared with the median centralization approach (Figure 2). No significant improvement was observed in correlations between FGFR1 and MET, and their respective inhibitors: all ρ values where lower than 0.1 for FGFR1, and close to 0 for MET, and P values were at least >0.27 in all centralization comparisons (supplementary Figure S5, available at *Annals of Oncology* online).

comparison of the outputs of different centralization methods on aCGH profiles from patients prospectively enrolled in precision medicine programs (SAFIR01 and MOSCATO-01)

Due to lower performances of the median centralization on cell lines, only the maximum peak and the alternative centralization were considered. The alternative peak detection method was applied on 283 breast metastasis samples from SAFIR01, and on 309 MOSCATO-01 tumor samples. In the SAFIR01 cohort, we observed an alternative centralization peak in 76 of the 283 profiles (26.9%), with similar proportions on both platforms: 31/125 profiles generated using Affymetrix (24.8%), and 45/158 generated using Agilent 4 × 180K (28.5%). Importantly, when applying the alternative centralization, an actionable amplification was detected in 20 of the 283 patients (7.07%), for whom no actionable trait was previously identified with the maximum peak method. Further, supplementary amplifications, not seen with the maximum peak centralization, were detected in 22 patients (7.8%).



ERBB2 Vs. Lapatinib

	Median	Maximum	Alternative
Rho	0.012	0.122	0.156

p-values	Maximum	Alternative
Median	0.043	0.012
Maximum		0.294

EGFR Vs. Erlotinib

	Median	Maximum	Alternative
Rho	0.068	0.121	0.118

p-values	Maximum	Alternative
Median	0.204	0.218
Maximum		0.480

Figure 2. Correlation between copy number variation and sensitivity to related inhibitors. The Spearman correlation between ERBB2 and lapatinib increased significantly when applying the maximum peak or the alternative peak centralizations, compared with the median value adjustment ($\rho = 0.122$, 0.156, and 0.012, respectively. $P = 0.043$ and 0.012, respectively). The alternative centralization even improved the correlation, but not significantly ($P = 0.294$). The same trend was observed for EGFR and erlotinib, even though none of the improvements appeared significant.

Similar results were obtained using the MOSCATO-01 cohort: an alternative peak was detected in 79 of 309 samples (25.6%). As a major consequence, an actionable amplification was detected using the alternative centralization in 21 patients (6.8%), while no aberration was previously identified using the maximum peak method, and supplementary amplifications were found in six patients (1.9%).

In both cohorts, the alternative centralization never missed any amplification detected using the maximum centralization method. In the two studies, the new (or possibly supplementary) actionable amplifications included the same genes, but with different frequencies. This can be due to the differences between the two cohorts: metastasis of breast tumors and metastasis from all type of tumors, in SAFIR01 and MOSCATO-01, respectively (Figure 3 and supplementary Table S6, available at *Annals of Oncology* online).

discussion

Array-based genomic profiling is widely used to estimate gains and losses of DNA segments and ultimately to guide the therapeutic decision in personalized medicine programs. Herein, we demonstrated the importance of the centralization step to determinate the LogR of the array signal intensities by comparing the effect of three different methods on several panels of cell lines and patients cohorts. To our knowledge, this is the first time that such a comparison between various centralization methods is carried out.

Using the NCI60 panel of cell lines for which the related karyotypes are known allowed us to prove that some centralization rules can lead to erroneous profiles. To note, his effect appears prominent in the aneuploidy setting, which is frequent in cancer. For instance, in cell lines with high ploidy, centering on the LogR median or on the highest density peak led to inappropriate values. In the latter setting normal 2-copy regions are estimated as losses; thus, amplifications are likely to be underestimated and deletions are likely to be overestimated. Though, even after the centralization adjustment, we did observe remaining discrepancies between the genomic profiles and karyotypes. Several reasons could be advocated to explain this such as technical issues like the adjustment step of DNA samples to a fixed quantity before being used could reduce ploidy differences between the sample and the reference used in the CGH array. Similar effects have been observed on gene expression analysis [24].

Second, using a large panel of cell lines (CCLE), we showed that applying the alternative centralization peak method led to a significantly improved correlation between ERBB2 CN and the sensitivity to lapatinib when compared with the median and the maximum peak methods. However, this had little impact on the correlation between EGFR and the sensitivity to erlotinib. To note, none of the centralization procedures tested lead to significant differences in correlations between FGFR1, and MET, with their respective known inhibitors. This latter result may be secondary to the fact that the drugs tested were relatively weak inhibitors of FGFR1 (TKI258) and MET (PHA665752), thus rendering the correlation hazardous. Further, we cannot formally exclude issues in

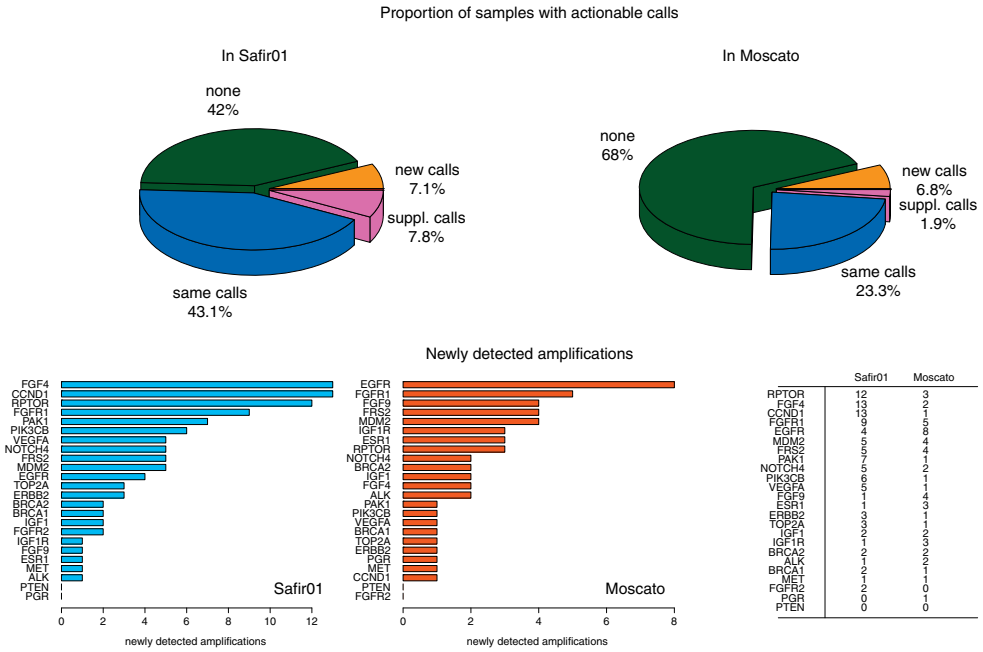


Figure 3. Effect of different genomic profiles centralization methods on possible therapeutic orientations. Top panel: Centralizing on an alternative peak leaded to identify actionable amplifications (new calls) in 20 more samples (7.1%) in the SAFIR01 data, and in 21 supplementary samples (6.8%) in MOSCATO, for whom no amplification was found by using the maximum peak for centralizing the genomic profiles. In 22 (7.8%) and 6 (1.9%) cases in SAFIR01 and MOSCATO, respectively, supplementary amplifications (sup. calls) were also identified, leading to supplementary options for a therapeutic decision making. For 42% and 68%, in Safir01 and Moscato, respectively, no therapeutic option appeared, and in 43.1% and 23.3%, the same actionable genes were identified with both methods. Bottom panel: FGF4, CCND1, and RPTOR were the most frequently newly detected amplified genes in Safir01, while EGFR and FGFR1, principally, were impacted by the centralization strategy in Moscato data (frequencies are summarized in the table, on right).

the estimation of the drug sensitivity in these large panels of cell lines, as this was previously noted before [25].

Finally, applying similar centralization comparisons on the SAFIR01 and the MOSCATO-01 data also unveiled ambiguities for determining the CGH calls in tumor samples, and thus the therapeutic decisions. Applying the alternative centralization method would have changed the decision for 20/283 (7.07%) and 21/309 (6.8%) patients, in SAFIR01 and MOSCATO-01, respectively, for whom a standard approach did not reveal any actionable gene amplification. Since array-based genomic profilings give a global overview of a diversity of events that occur in a tumor, and may provide possible misinterpretations, fluorescent *in situ* hybridizations may be considered as a necessary validation step. However, such verification is rarely performed because of evident cost and time issues, and thus reinforces the importance of the centralization step in the CGH profiling.

In this study, we showed that the centralization step is critical in the evaluation of gene copy number using CGH arrays and is susceptible to substantial effects on the decision-making criteria for patient treatment. Among the three different centralization methods tested, the alternative peak approach appears promising. Since centralization problems are linked with the tumor polidy variation, they are not likely to be restricted to only hybridization-based technologies and may also impact sequencing-based pipelines. Though, dedicated methods remain to be developed for the latter technologies.

funding

FC, CF, SF and JG were supported by the grant U54CA149237 from the Integrative Cancer Biology Program of the National Cancer Institute.

disclosure

The authors have declared no conflicts of interest.

references

1. Cline MS, Craft B, Swatloski T et al. Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser. Sci Rep 2013; 3: 2652.

2. Barretina J, Caponigro G, Stransky N et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012; 483 (7391): 603–607.
3. Garnett MJ, Edelman EJ, Heidorn SJ et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012; 483(7391): 570–575.
4. André F, Bachelot T, Commo F et al. Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIR01/UNICANCER). Lancet Oncol 2014; 15(3): 267–274.
5. Hollebecque A, Massard C, De Baere T et al. Molecular screening for cancer treatment optimization (MOSCATO 01): a prospective molecular triage trial—interim results. ASCO Annual Meeting, 2013: Abstract 2512.
6. WIN Consortium. http://www.winconsortium.org/ (30 October 2014, date last accessed)
7. Piccart-Gebhart MJ, Procter M, Leyland-Jones B et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med 2005; 353(16): 1659–1672.
8. Laurent-Puig P, Cayre A, Manceau G et al. Analysis of PTEN, BRAF, and EGFR status in determining benefit from cetuximab therapy in wild-type KRAS metastatic colon cancer. J Clin Oncol 2009; 27(35): 5924–5930.
9. André F, Bachelot T, Campone M et al. Targeting FGFR with dovitinib (TKI258): preclinical and clinical data in breast cancer. Clin Cancer Res 2013; 19(13): 3693–3702.
10. Marioni JC, Thorne NP, Tavaré S. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics 2006; 22(9): 1144–1146.
11. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 2007; 23(6): 657–663.
12. Mermel CH, Schumacher SE, Hill B et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 2011; 12(4): R41.
13. Picard F, Robin S, Lavielle M, Vaisse C, Daudin J-J. A statistical approach for array CGH data analysis. BMC Bioinformatics 2005; 6(1): 27.
14. Van De Wiel MA, Kim KI, Vosse SJ et al. CGHcall: calling aberrations for array CGH tumor profiles. Bioinformatics 2007; 23(7): 892–894.
15. Van Houte BPP, Binsl TW, Hettling H, Pirovano W, Heringa J. CGHnormaliter: an iterative strategy to enhance normalization of array CGH data with imbalanced aberrations. BMC Genomics 2009; 10(1): 401.
16. Staaf J, Jönsson G, Ringnér M, Vallon-Christersson J. Normalization of array-CGH data: influence of copy number imbalances. BMC Genomics 2007; 8(1): 382.
17. Yang S, Pounds S, Zhang K, Fang Z. PAIR: paired allelic log-intensity-ratio-based normalization method for SNP-CGH arrays. Bioinformatics 2013; 29(3): 299–307.
18. Popova T, Manié E, Stoppa-Lyonnet D et al. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. Genome Biol 2009; 10(11): R128.
19. Chen HI, Hsu FH, Jiang Y et al. A probe-density-based analysis method for array CGH data: simulation, normalization and centralization. Bioinformatics 2008; 24 (16): 1749–1756.
20. Knutsen T, Gobu V, Knaus R et al. The interactive online SKY/M-FISH & CGH database and the Entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence. Genes Chromosomes Cancer 2005; 44(1): 52–64.
21. NCI60 cell line panel Genetics. ftp://ftp.ncbi.nih.gov/sky-cgh/ESI/NCI60_cell_line_panel_Genetics_Branch_I.R.Kirsch.esi (1 October 2014, date last accessed).
22. Celeux G, Govaert G. Gaussian parsimonious clustering models. Pattern Recognit 1995; 28(5): 781–793.
23. Fraley C, Raftery AE. Model-based methods of classification: using the mclust Software in Chemometrics. J Stat Softw 2007; 18(6): 1–13.
24. Lovén J, Orlando DA, Sigova AA et al. Revisiting global gene expression analysis. Cell 2012; 151(3): 476–482.
25. Haibe-Kains B, El-Hachem N, Birkbak NJ et al. Inconsistency in large pharmacogenomic studies. Nature 2013; 504(7480): 389–393.

5.1.2 Résultats supplémentaires

En complément des résultats publiés, nous avons également comparé les profils des lignées du NCI60, après analyse par les différents algorithmes mentionnés plus haut. Les méthodes applicables à ces données Nimblegen étaient : cghNormaliter, cghCall, GISTIC et popLowess. A ces méthodes, nous avons ajouté, comme dans l'article, 3 stratégies de centrage : centrage médian, pic maximum et pic alternatif. Les distances quadratiques entre profils et caryotypes correspondants ont été comparées par des tests de Student appariés. Les résultats montraient que la méthode du pic alternatif réduisait significativement les erreurs de profil (Figure 10).

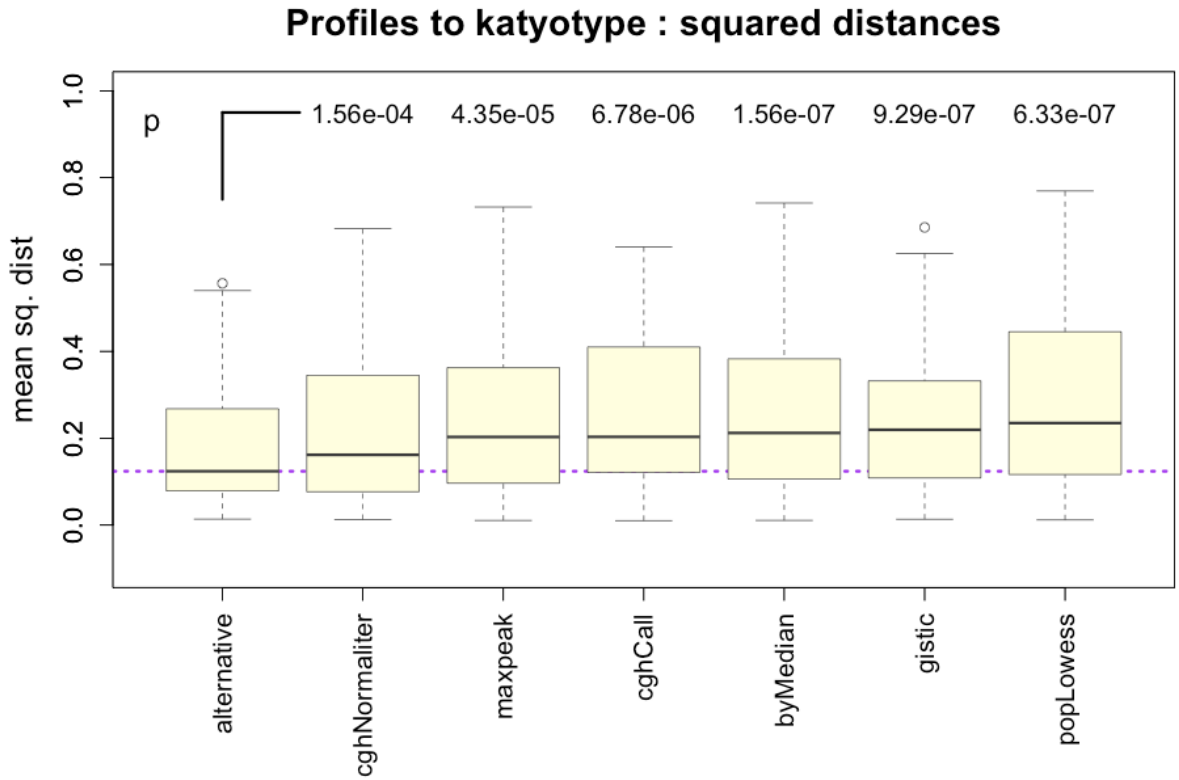


Figure 10: Distances aux caryotypes. La centralisation par la méthode du pic alternatif réduit significativement les distances entre profils génomiques et caryotypes.

5.1.3 Commentaires

Les lignées issues de la série NCI60 (Shoemaker 2006), et caractérisées pour leurs anomalies génomiques (Knutsen et al. 2005), nous ont permis de montrer l'influence de la centralisation sur la pertinence d'un profil génomique.

Des choix alternatifs pour la centralisation apparaissent dans plus de 30% des cas, et nous montrons, en particulier, que dans le cas d'aneuploïdies importantes et dominantes, une centralisation sur un pic majoritaire peut conduire à considérer des régions chromosomiques à 2n comme des régions perdues. Une autre conséquence est de potentiellement sous-estimer les gains et amplifications.

Dans la série plus large des lignées CCLE, les ambiguïtés de centralisation sont apparues dans 16% des cas. L'analyse de gènes particuliers, a également montré que le choix de centralisation pouvait avoir un impact sur la mesure de corrélation entre niveaux de gains et sensibilité à un inhibiteur spécifique : la corrélation entre niveaux de gains du gène ERBB2 et la sensibilité au lapatinib était significativement plus élevée lorsque la centralisation alternative était choisie, en comparaison d'une centralisation sur la médiane ($p = 0.012$). Une tendance similaire, mais non siginficative, était observée quant à la corrélation entre les anomalies de EGFR et la sensibilité à l'erlotinib, mais aucune influence du choix de centralisation n'était notée dans les corrélations entre FGFR1 et MET et leurs inhibiteurs respectifs, TKI258 et PHA665752 (au moins $p > .27$).
Il est toutefois possible que les mesures de sensibilités aux inhibiteurs dans la série CCLE soit une limite à des explorations fiables (Haibe-Kains et al. 2013).
En appliquant les mêmes stratégies sur les données issues des programmes de screening moléculaire SAFIR-01 et MOSCATO, nous avons observé que des centralisations alternatives pouvaient être choisies pour respectivement 26.9% et 25.6% des profils.
Une conséquence majeure d'un changement de stratégie de centralisation était de révéler de nouvelles altérations actionnables pour 69 des 592 patients (11.6%) analysés au cours des 2 études. Dans 41 cas (6.9%) une centralisation alternative révélait au moins une anomalie actionnable dans des profils pour lesquels aucune anomalie n'avait été détectée. Pour 28 autres cas (4.7%), des altérations supplémentaires étaient identifiées, ouvrant potentiellement la voie à des choix thérapeutiques supplémentaires.

Au cours de ce travail, nous avons montré que le choix de centralisation, dans les profils génomiques, est une étape critique pouvant avoir un impact direct sur la décision d'une orientation thérapeutique, dans le contexte des programmes de screening moléculaires. En l'absence d'une connaissance à priori des ploïdies des tumeurs, il est envisageable que certains profils conduisent à des interprétations erronées.
Nous n'avons pas, ici, évalué l'estimation des variations de nombre de copies à partir de données de séquençage, mais il est possible, compte tenu des nécessités de normalisation, que la ploïdie des tumeurs puisse avoir un impact similaire à celui observé dans l'analyse de profils génomiques par aCGH.

Il est possible que l'analyse de biopsies contenant une proportion significative de tissu normal puisse garantir d'avoir une population de cellules 2n suffisamment importante pour assurer une normalisation appropriée. Mais cette solution pourrait se faire au détriment de la sensibilité de détection des altérations.

Nos observations pourraient inciter à appuyer les décisions d'orientation sur une validation des altérations identifiées dans les profils génomiques : lorsqu'elles sont techniquement possibles, l'hybridation in situ fluorescente (FISH), ou l'immunohistochimie, pourraient être utilisées pour confirmer l'existence de variations de nombre de copies d'un gène d'intérêt, ou la surexpression de la protéine correspondante.

5.2 Implémentation d'un pipeline d'analyse

Afin de faciliter l'analyse des données génomiques, et leur interprétation, nous avons développé un package complet implémenté en langage R, nommé rCGH.
L'objectif était de créer un pipeline complet et portable, pouvant supporter les données produites à partir des principales plateformes de microarray, et pouvant assurer une traçabilité complète de l'analyse.
Avec ce package, nous avons également voulu fournir des outils de visualisation interactive, permettant d'affiner dynamiquement certains paramètres, et susceptible d'aider à l'interprétation des profils et à la prise de décision thérapeutique.

5.2.1 Lecture des données

Le pipeline rCGH supporte les données Agilent (microarray de 44K à 400K), les données Affymetrix issues de puces SNP6.0 et cytoScanHD, et peut accepter les données de microarrays « sur mesure », sous réserve du respect de certaines contraintes de format, comme indiqué dans la documentation de rCGH.
Chacune des plateformes produit des fichiers de type et contenu différents :

- *La plateforme Agilent produit des fichiers au format texte (.txt), contenant les mesures d'intensité de chacune des cyanines, ainsi que l'ensemble des indicateurs relatifs à la qualité des signaux.*
- *La plateforme Affymetrix produit des fichiers CEL binaires requérant des outils spécifiques de conversion.*

Affymetrix met à disposition de la communauté scientifique une suite de programmes d'analyse, Affymetrix Power Tools (APT), prenant en charge la lecture de ces fichiers binaires et leur conversion en fichiers .txt. Pour des raisons techniques, il ne nous a pas semblé judicieux d'intégrer APT au pipeline rCGH (dépendance aux systèmes d'exploitation, taille des fichiers), et cette conversion reste à la charge de l'utilisateur. Cependant, nous avons développé 2 packages R (SNP6 et cytoScan) dédiés à la conversion des fichiers CEL. Pour des raisons de taille, ces 2 packages n'ont pas été soumis à Bioconductor, mais sont disponibles sur demande.

5.2.2 Sauvegarde des informations et paramètres

Le package rCGH utilise massivement la programmation orientée objet.
A la lecture des données, les métadonnées associées aux fichiers, lorsqu'elles existent, sont sauvegardées dans un objet de classe 'rCGH', et peuvent être rappelées à tout moment par l'intermédiaire de fonctions spécifiques.
De la même manière, l'ensemble des paramètres ont leurs valeurs sauvegardées au fur et à mesure du processus d'analyse. Comme pour toute information stockée dans un objet 'rCGH', ces paramètres peuvent être rappelés via des fonctions dédiées.

5.2.3 Preprocessing

Le « preprocessing » consiste à appliquer un certain nombre de filtres se rapportant à la qualité de chaque spot du microarray : uniformité, circularité,..., ainsi qu'à appliquer certaines corrections de biais : contenu en GC, et biais de cyanines dans le cas de co-hybridation à 2 couleurs (Marioni et al. 2007). Une approche classique pour corriger les biais de cyanines, est d'utiliser une méthode de régression locale, ou LOESS (Cleveland & Devlin 1988) (Figure 11).

Brièvement, considérant f une fonction inconnue, pour chaque point x_0, $f(x_0)$ est estimée à partir des points x_i au voisinage de x_0, et pondérés par leur distance au point considéré. Une fonction de pondération couramment utilisée est une fonction cubique de la forme :

$$W(u) = \begin{cases} \left(1 - |u|^3\right)^3 & \text{si } |u| < 1 \\ 0 & \text{sinon} \end{cases}$$

Les pondérations sont donc calculées comme : $w_i(x_0) = W\left(\frac{x_i - x_0}{h(x)}\right)$, *avec h(x) un paramètre de distance, fixe, ou dépendant de x_0 pour assurer un nombre stable de valeurs de voisinage.*

Dans le cas d'une fonction de degré 2, f peut être approximée par :

$$f(x) \approx \beta_0 + \beta_1(x - x_0) + \frac{1}{2}\beta_2(x - x_0) \text{ pour } x \in [x_0 - h(x_0), x_0 + h(x_0)]$$

Compte tenu des pondérations, il est alors possible d'estimer $\hat{f}(x_0)$ *à partir du vecteur de paramètres* $\beta = (\beta_0, \beta_1, \beta_2)^T$ *minimisant :*

$$\hat{\beta} = \underset{\beta \in R^3}{\operatorname{argmax}} \sum_{i=1}^{n} w_i(x_0) \left[Y_i - \left(\beta_0 + \beta_1(x - x_0) + \frac{1}{2}\beta_2(x - x_0) \right) \right]^2$$

Il existe une solution analytique pour ce problème de minimisation :

$\hat{\beta} = (X^T W_i X)^{-1} X^T W_i Y$ *, où X,* W_i *et Y sont respectivement la variable prédictive, la matrice diagonale des* w_i*, et les valeurs prédites.*

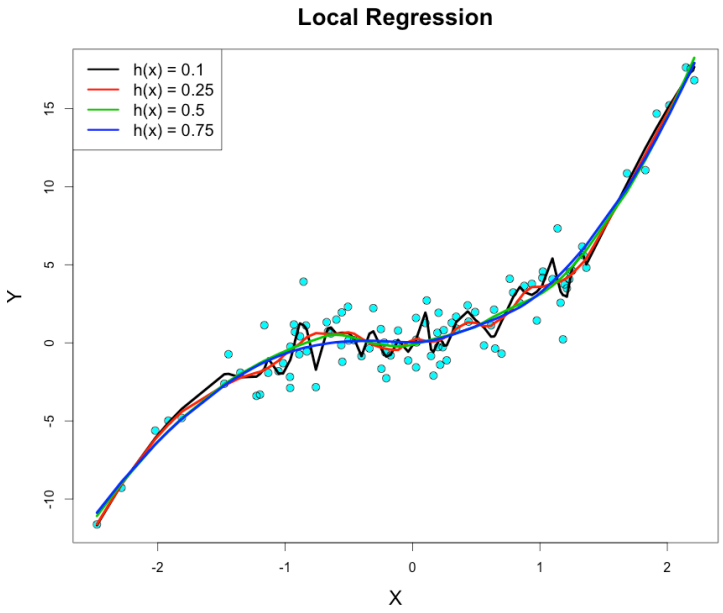


Figure 11 : Exemple de régression locale. L'estimation obtenue dépend fortement du nombre de points considérés au voisinage des points estimés : la fonction h(x) garantie d'utiliser un nombre stable de points, défini comme une proportion du nombre total de valeurs. Le code pour cette démonstration est fourni en annexe 4.

Dans le cas des données Affymetrix, SNP6.0 ou cytoScanHD, ces traitements sont pris en charge par APT.

Dans le cas de données Agilent, le pipeline rCGH intègre ces corrections et utilise la méthode des régressions locales pour ajuster ces biais.

La valeur d'intensité de chaque sonde est convertie en ratio d'intensité, considérant le signal de la sonde correspondante dans l'ADN de référence, puis transformée en Log_2(ratio) relatif (LRR). La valeur de chaque sonde S_i est donc calculée comme :

$LRR(S_i) = Log_2\left(\frac{S_{ti}}{S_{ni}}\right)$ *où S_{ti} et S_{ni} sont respectivement le signal de la sonde i dans la tumeur et dans l'ADN normal de contrôle.*

Les ajustements par régressions locales sont successivement appliqués pour la correction des biais de cyanines, puis pour la correction des biais liés aux contenus en bases GC (Figure 12).

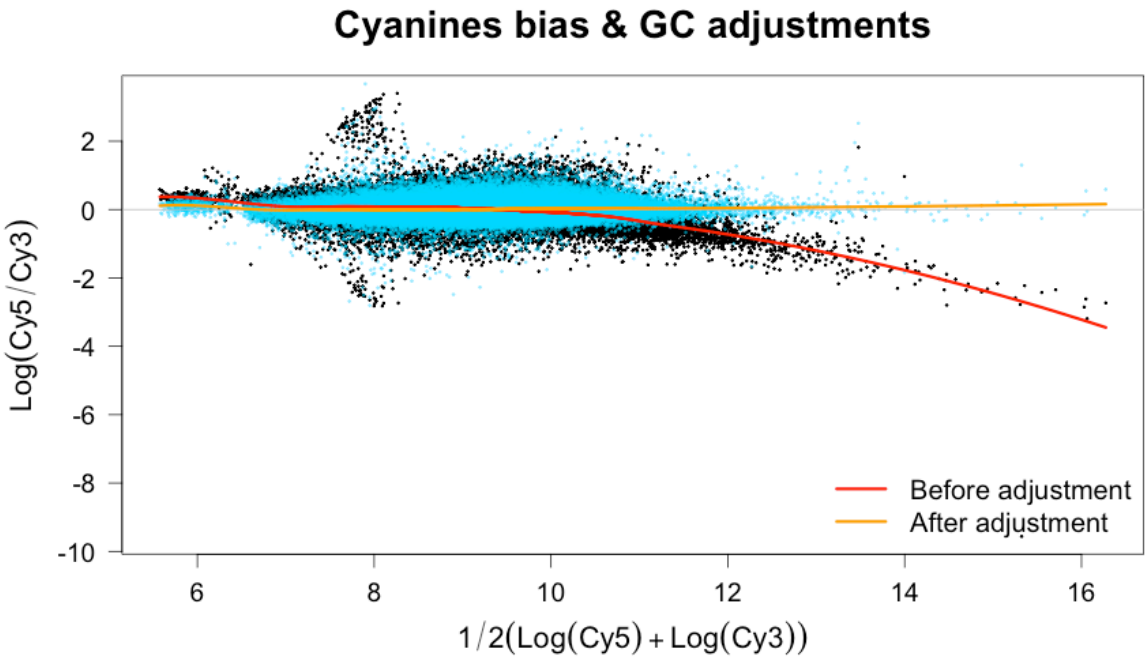


Figure 12 : Correction des biais. Le contenu des sondes en bases GC, ainsi que l'utilisation de 2 cyanines (Cy5 et Cy3) dans le cas de co-hybridations, peuvent générer des biais d'intensité. Ces biais peuvent être corrigés à l'aide de régressions locales (loess).

5.2.4 Réduction du bruit

Les données haut-débit généralisent le principe variabilité expérimentale : les signaux acquis sont la somme de 2 sources : le signal lui-même et un « bruit » expérimental. Dans le cas des microarrays, ces sources de bruit peuvent être liées à la qualité de l'ADN source, à la qualité de l'hybridation, mais aussi à la spécificité

des sondes elles-mêmes : certaines sondes peuvent hybrider des séquences partiellement complémentaires et générer un signal non spécifique.
Réduire le bruit est une étape délicate en ce qu'il peut être difficile d'en estimer la part dans le signal acquis.

Certains groupes proposent la recherche et la suppression de signaux dits « outliers » (Mermel et al. 2011) : selon une procédure de fenêtres glissantes de 11 marqueurs à gauche et à droite x_{i-} = (x_{i-11}, x_{i-10},…,x_{i-1}, x_{i+1}, x_{i+2},…,x_{i+11}), le point évalué x_i est défini comme outlier, et remplacé par médiane(x_{i-}), si :

$|x_i| > mediane(x_{i-}) + 1.5IQR(x_{i-})$, *où IQR($x_{i-}$) est la distance interquartile.*

Dans le pipeline rCGH, nous avons préféré intégrer une étape de réduction du bruit construite sur un algorithme d'expectation-maximisation (EM) (Dempster et al. 1977). Cette approche nous a paru moins coûteuse en temps de calcul, et moins stringente.

L'algorithme est décrit en annexe 5, et peut être résumé comme suit (Figure 13):
Soit un vecteur X = (x_1,…,x_n), considéré comme un mélange gaussien à K classes, tel que sa densité de probabilité est définie par :
$g(x,\Theta) = \sum_K \pi_k f(x,\theta_k)$ *où f est une densité de probabilité de paramètre* $\theta_k = \{\mu_k, \sigma_k\}$, *pour chaque* $k \in K$.

1) *Les paramètres sont initialisés aléatoirement.*
2) *L'étape E (expectation) estime la probabilité d'appartenance de chaque élément* $x_i \in X$ *à chaque classe* $k \in K$.
3) *Chaque x_i est assigné à sa classe de probabilité la plus élevée.*
4) *L'étape M (maximisation) optimise les nouveaux paramètres de chaque classe k.*

L'algorithme est répété jusqu'à atteindre un maximum d'itérations, où qu'une solution optimale soit obtenue. Généralement, les solutions sont évaluées sur les valeurs d'une fonction objective telle que la log-vraisemblance.

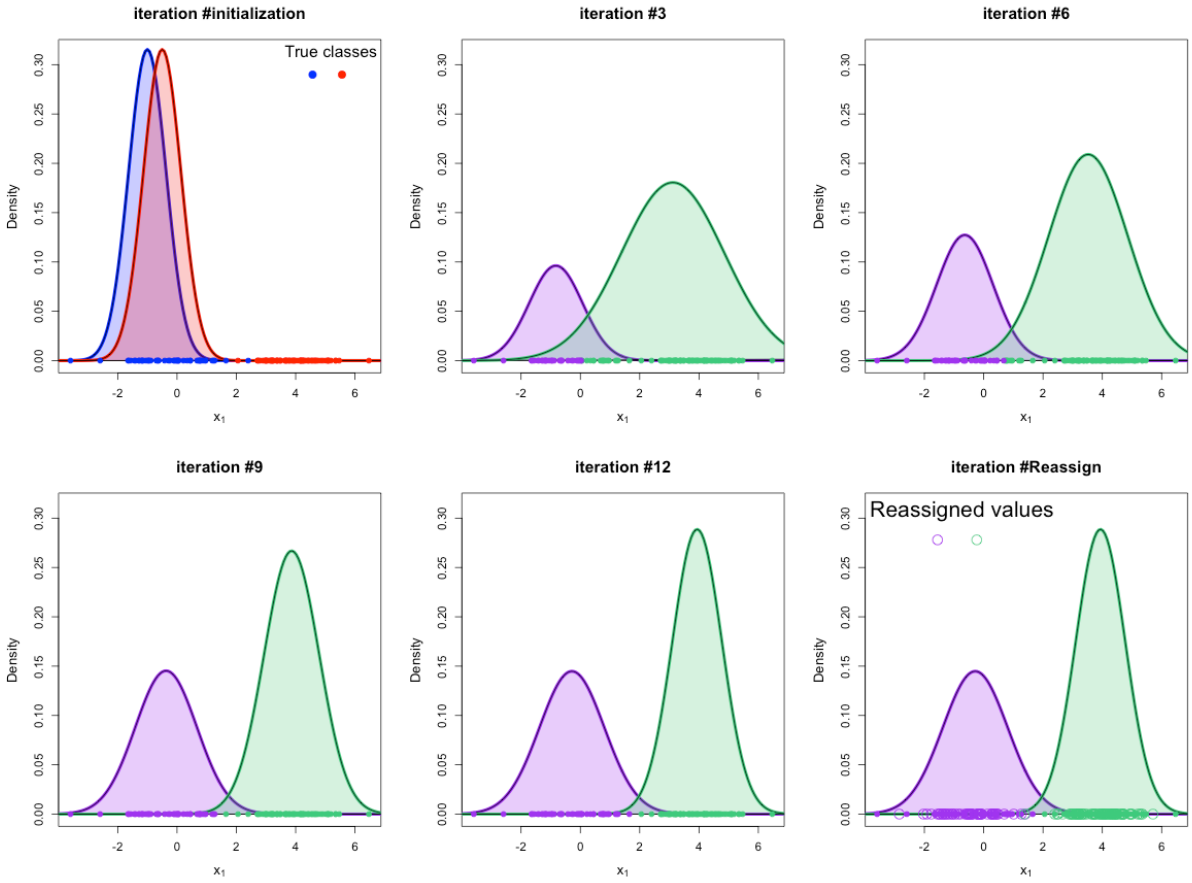


Figure 13 : Algorithme EM. Après initialisation par des valeurs aléatoires, les paramètres du mélange de densités sont estimés par la répétition des étapes E et M. Une fois la solution optimale atteinte, les valeurs initiales sont remplacées par des valeurs générées par chaque distribution, supprimant ainsi toute ou partie des valeurs extrêmes (iteration #Reassign).

Afin de réduire le bruit, nous appliquons cet algorithme sur des fenêtres non glissantes, comme suit :

$LRR_a = \left(LRR_i,...,LRR_j\right)$ *tel que a = 0, 1,…, e = 2.5*^{e}*3,* $i = \min\left(a.e+1, N\right)$ *et* $j = \min\left(i+e-1, N\right)$

Pour chaque LRR_a les valeurs sont réassignées, après estimation des paramètres du mélange : $LRR_{ai} \sim N(\mu_k, \sigma_k, 0.95)$ *si le point i appartient à la classe k.*

L'adoption de cette méthode à, en outre, permis de réduire significativement le bruit dans les données, mesuré ici par le derivative LogRatio spread (dLRs) (Figure 14).

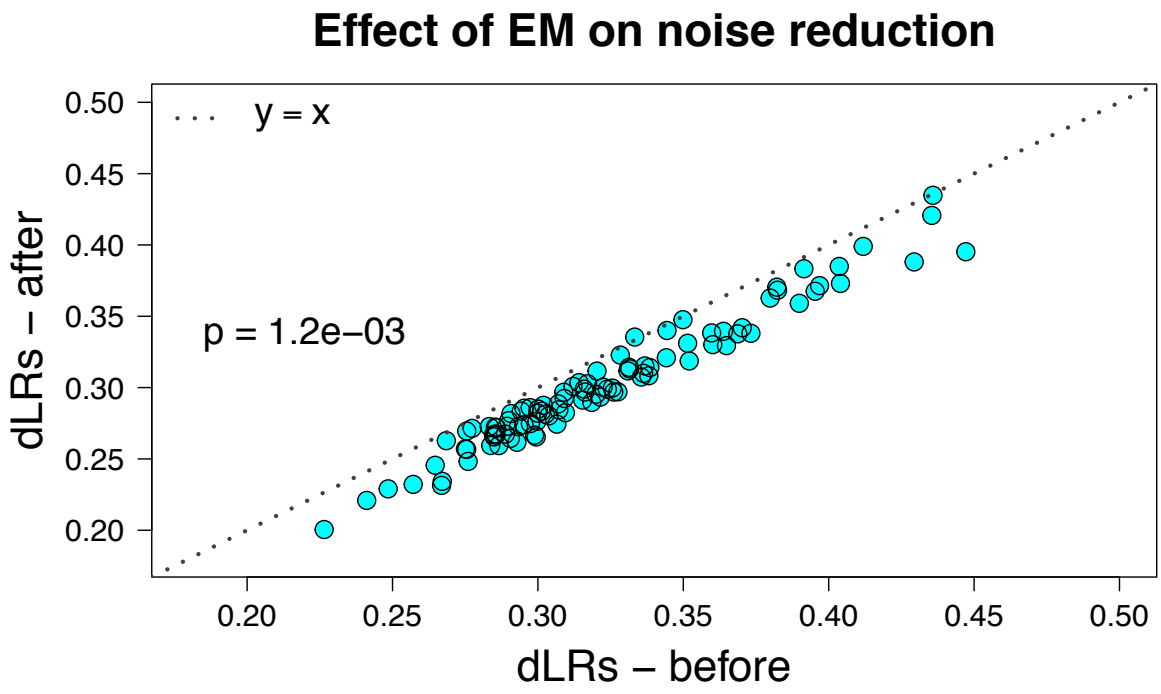


Figure 14 : L'application d'un algorithme EM pour réduire le nombre d'outliers permet de réduire significativement le bruit observé dans les données (p = 1.2^{e-3}).

5.2.5 Centrage des profils

Comme décrit dans la publication précédemment présentée, le centrage des profils génomiques utilise un algorithme EM permettant de modéliser les LRR comme un mélange de populations ayant des distributions gaussiennes à paramètres différents. Nous avons toutefois apporté certaines simplifications procurant un gain en temps de calcul, sans impacter l'estimation des paramètres des distributions : l'algorithme EM est appliqué à un échantillon unique de LRR, dont la taille peut être définie par l'utilisateur. Par défaut, n = 25^e3 sondes.

Après estimation des paramètres du mélange, les pics de densité sont comparés entre eux, et la règle de décision pour le choix d'une valeur de centrage se définit comme suit :

1) *Considérant* $\max\left(g\left(x,\Theta\right)\right)$*, le maximum de densité du mélange. Sont retenus les indices k tels que les hauteurs de densité des population k sont supérieures à* $\alpha.\max\left(g\left(x,\Theta\right)\right)$*, où* α *est un coefficient de proportionnalité, choisi entre 0 et 1 :*

$$K = \left\{k \mid \max\left(f_k\left(x,\theta_k\right)\right) \geq \alpha.\max\left(g\left(x,\Theta\right)\right)\right\},\ 0 \leq \alpha \leq 1$$

2) *La valeur de centrage, c, est alors définie comme la plus faible valeur parmi les moyennes des populations vérifiant le critère précédent.*

$$c = \underset{\mu}{\operatorname{argmin}}\left(\mu_{k|k\in K}\right)$$

Par défaut, α *= 0.5 ; un niveau neutre à 2 copies correspond donc à la population de plus faible moyenne, parmi celles ayant une hauteur de densité supérieure à 50% du maximum de densité du mélange gaussien (Figure 15).*

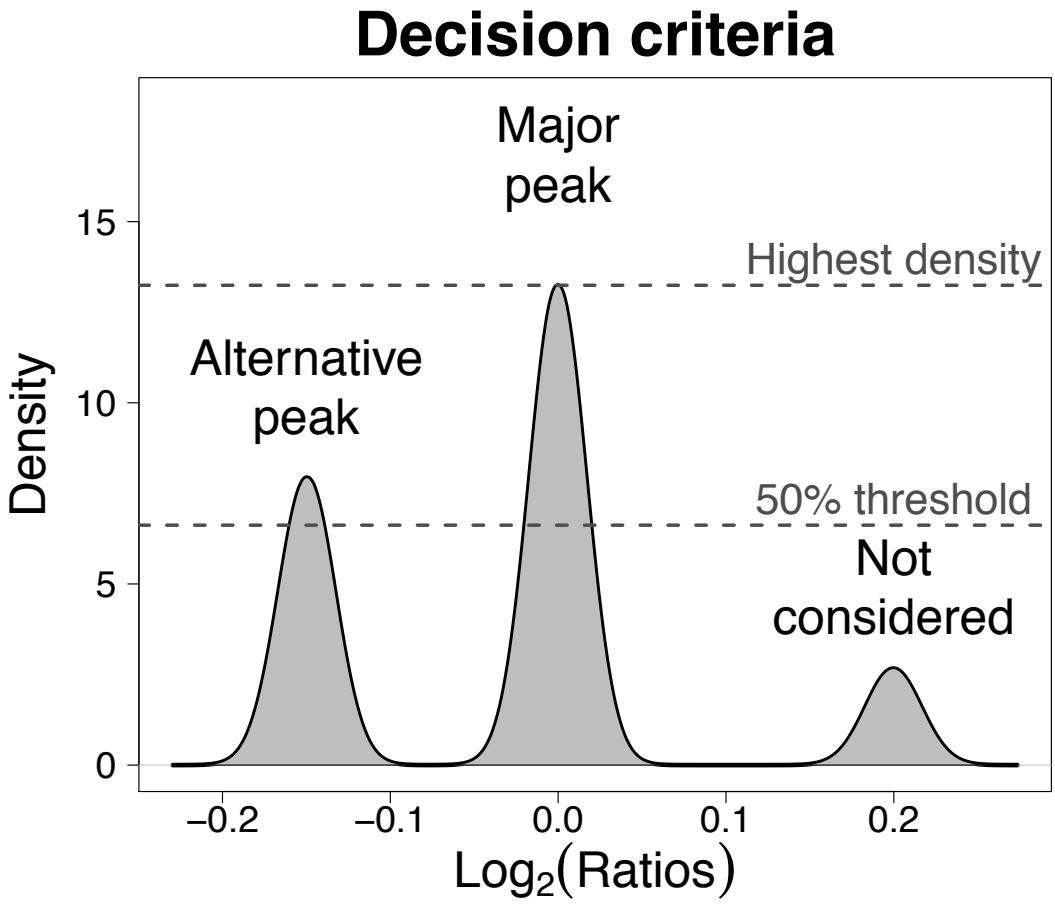


Figure 15 : Méthode de centrage. Par défaut, la population dont la hauteur du pic de densité est au moins 50% la hauteur du pic maximal est sélectionné pour représenter le niveau d'équilibre à 2 copies.

La valeur du paramètre α peut cependant être spécifiée par l'utilisateur, en fonction des stratégies adoptées : α = 1 entraine un centrage des données sur la population majoritaire, considérée comme représentant un niveau d'équilibre à 2 copies.

5.2.6 Segmentation

Comme mentionné précédemment, plusieurs techniques de segmentation ont été proposées, mais l'algorithme le plus fréquemment utilisé est le Circular Binary Segmentation, ou CBS (Olshen et al. 2004; Venkatraman & Olshen 2007). Cette méthode est adaptée des techniques de segmentations binaires (Sen & Srivastava 1975), et permet la détection de points de cassure multiples.

Son principe est le suivant :

Soit $X_1,...,X_n$ *les LRR de n sondes contigües, indexées par leur localisation chromosomique, et* $S_i = \sum_{k=1}^{i} X_k,\ 1 \le i \le n$ *les sommes partielles. Considérant les sondes* X_1 *et* X_n *raboutées, de manière à former un cercle, la statistique du ratio de vraisemblance testant l'hypothèse nulle, les moyennes de l'arc de* X_i *à* X_j *et de son complément sont égales, est donnée par :*

$$Z_c = \max|Z_{ij}|, \quad \text{où } Z_{ij} = \left(\frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i}\right) \cdot \left(\frac{1}{j-i} + \frac{1}{n-j+i}\right)^{-1/2}$$

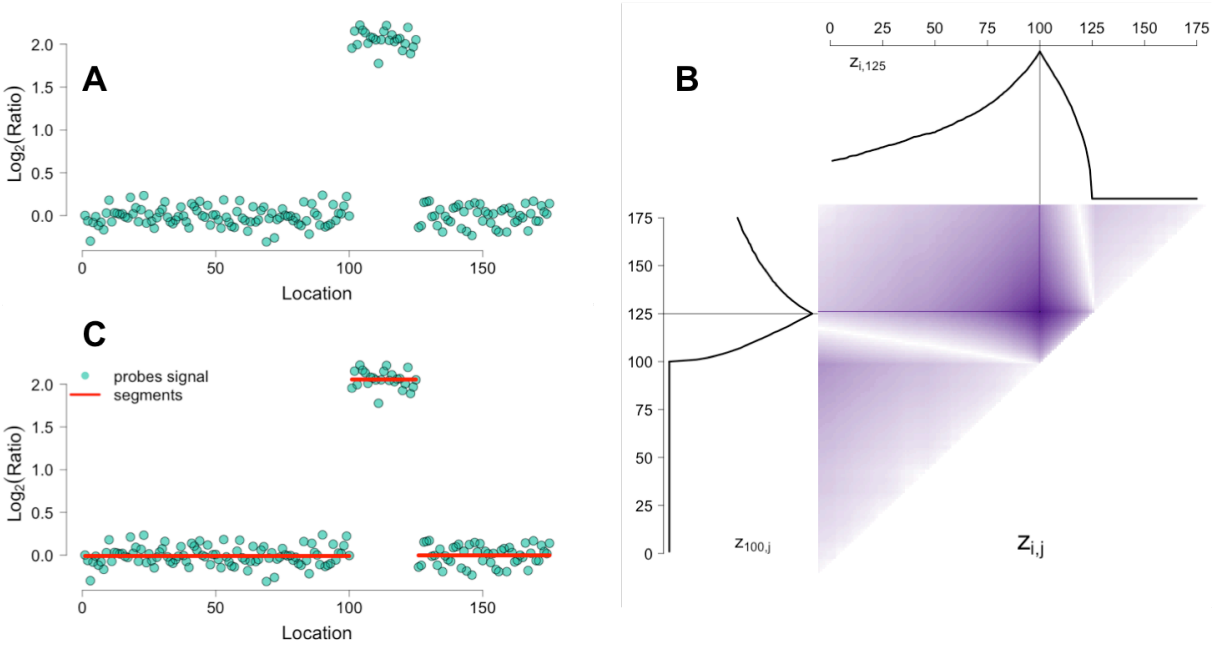


Figure 16 : L'algorithme CBS. les signaux ordonnés par leur position montrent une rupture de continuité aux positions 100 et 125 (B). L'algorithme CBS identifie un optimum aux coordonnées (100, 125) (B). Les signaux appartenant à une même région sont finalement résumés sous la forme d'un segment dont la valeur est la moyenne des sondes incluses dans cette région (C). Note : le code pour cette simulation est fourni en annexe 1.

Les études comparatives semblent donner la faveur à cette méthode (Willenbrock & Fridlyand 2005; Lai et al. 2005)*, ce que confirment nos propres analyses des performances sur des données simulées :*

Le package snapCGH (Marioni et al. 2006) *offre la possibilité de simuler des données et de comparer les performances de plusieurs algorithmes en mesurant les taux de vraies et fausses altérations. Nous avons comparé les performances des algorithmes GLAD, modèle de Markov caché, CBS et HaarSeg, sur 100 simulations.*

Si les temps de calcul sont largement en faveur de l'algorithme HaarSeg, les performances en terme de sensibilité (estimé par le taux de vraies altérations, TVA) et de spécificité (estimé par le taux de fausses altérations, TFA) se sont avérées à l'avantage de l'algorithme CBS (Figure 17). Il est toutefois à noter que les comparaisons ont été effectuées en utilisant les paramétrages par défaut des différents algorithmes. Chacune de ces méthodes accepte plusieurs arguments dont les valeurs peuvent impacter leurs performances.

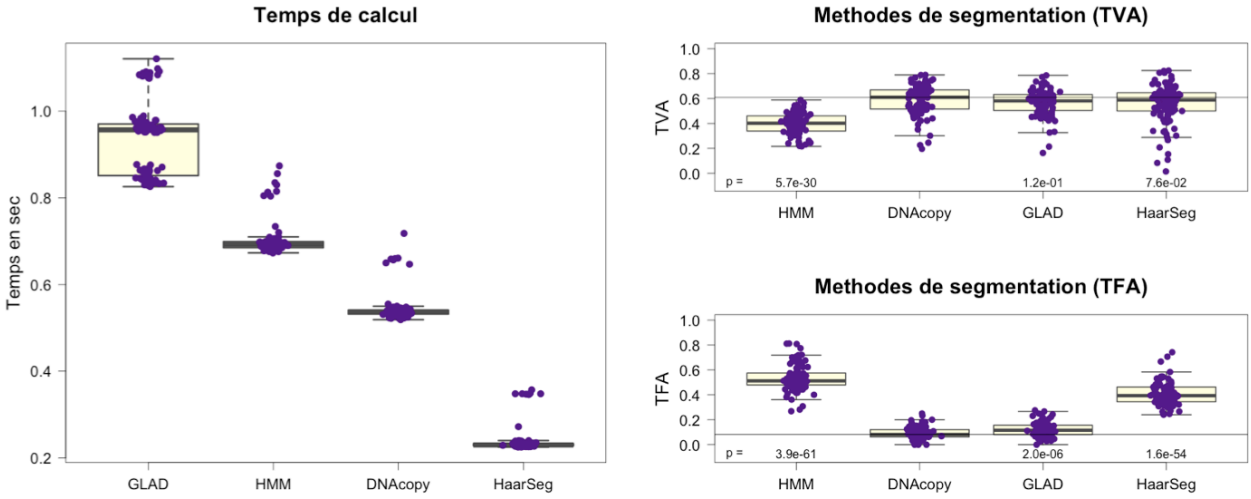


Figure 17: Comparaison des algorithmes de segmentation. A gauche, l'algorithme de Haar implémenté dans HaarSeg réduit les temps de calcul d'un facteur 2 (vs. DNAcopy) à 4 (vs. GLAD). A droite, en terme de sensibilité (TVA) et de spécificité (TFA), l'algorithme CBS implémenté dans DNAcopy apparaît significativement plus performant que les autres méthodes. Les p-values indiquées sont les résultats d'un test de Student entre chaque méthode et DNAcopy.

En conséquence de nos observations et des données de la littérature, nous avons fait le choix de l'algorithme CBS pour assurer la segmentation dans rCGH. Nous

avons cependant apporté certaines optimisations à l'algorithme initialement décrit par Olshen et al. et implémenté dans le package R DNAcopy (Venkatraman & Olshen n.d.).

DNAcopy propose plusieurs règles de décision permettant d'accepter, ou non, un point comme un point de cassure définissant un nouveau segment. Parmi ces règles, 'sdundo' nous a semblé la plus simple à paramétrer, et la mieux adaptée à l'analyse de profils génomiques.
Cette règle de décision s'appuie sur 2 paramètres devant être spécifiés à priori :

- *Le paramètre alpha, α, spécifie le seuil de significativité du test pour accepter un point comme un point de cassure.*
- *Le paramètre 'undo.sd' spécifie la différence minimale devant être observée entre les moyennes des LRR de 2 segments consécutifs pour les conserver disjoints. Il est exprimé en terme d'un nombre de déviations standard.*

Afin d'optimiser cette paramétrisation, nous avons analysé 100 profils génomiques issus de microarrays Agilent 4x180K (n = 50) et Affymetrix SNP6.0 (n = 50). La valeur de α a été fixée à $\alpha = 1^{e}$-6, et nous avons ajusté manuellement la valeur de 'undo.sd' de manière à obtenir des profils cohérents.
Ce test nous a permis de modéliser ce dernier paramètre comme une fonction du bruit observé dans les données, et exprimé par la déviation absolue à la médiane (MAD) (Figure 18), avec :

$MAD(x) = median|x - \tilde{x}|$, *où* $\tilde{x}$ *est la médiane du vecteur x.*

La valeur optimale de 'undo.sd' est alors estimée comme :

$undo.sd = 0.48\sqrt{MAD}$

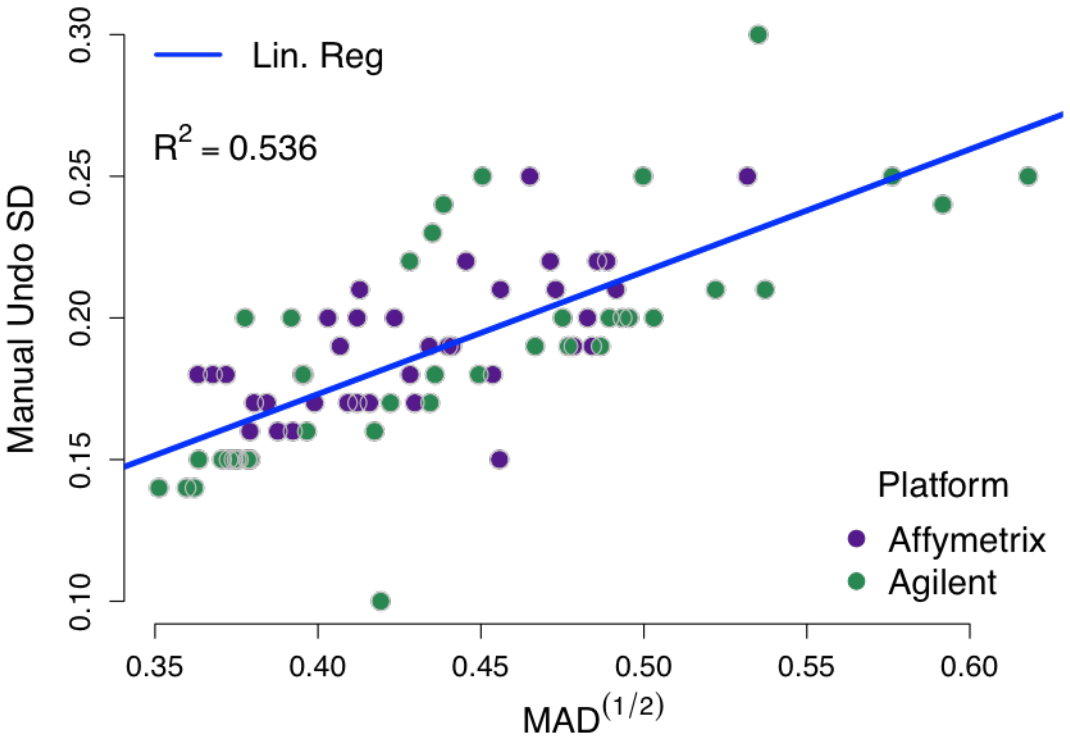


Figure 18 : l'analyse manuelle de 100 profils génomiques, générés sur Agilent 4x180 (n = 50) et Affymetrix SNP6.0 (n = 50) a permis de définir le paramètre undo.sd comme une fonction du bruit, estimé par la MAD.

L'intérêt de cette optimisation est d'affranchir l'utilisateur d'une paramétrisation arbitraire à priori. Au contraire, dans rCGH, cette paramétrisation peut être automatisée et guidée par la qualité des données observées. Pour des raisons de flexibilité, l'opérateur peut manuellement spécifier une autre valeur pour ce paramètre.

5.2.7 Visualisations statiques

Avec le pipeline rCGH, nous avons été particulièrement attentifs au développement d'outils de visualisation et de communication des résultats. Les outils de visualisation permettent, en outre, de re-spécifier manuellement les paramètres d'analyse.

5.2.7.1 Visualisation des densités

rCGH inclus, en premier lieu, une fonction permettant de visualiser le modèle de densité des LRR, et la population de valeurs ayant été utilisée pour le centrage du profil (Figure 19).

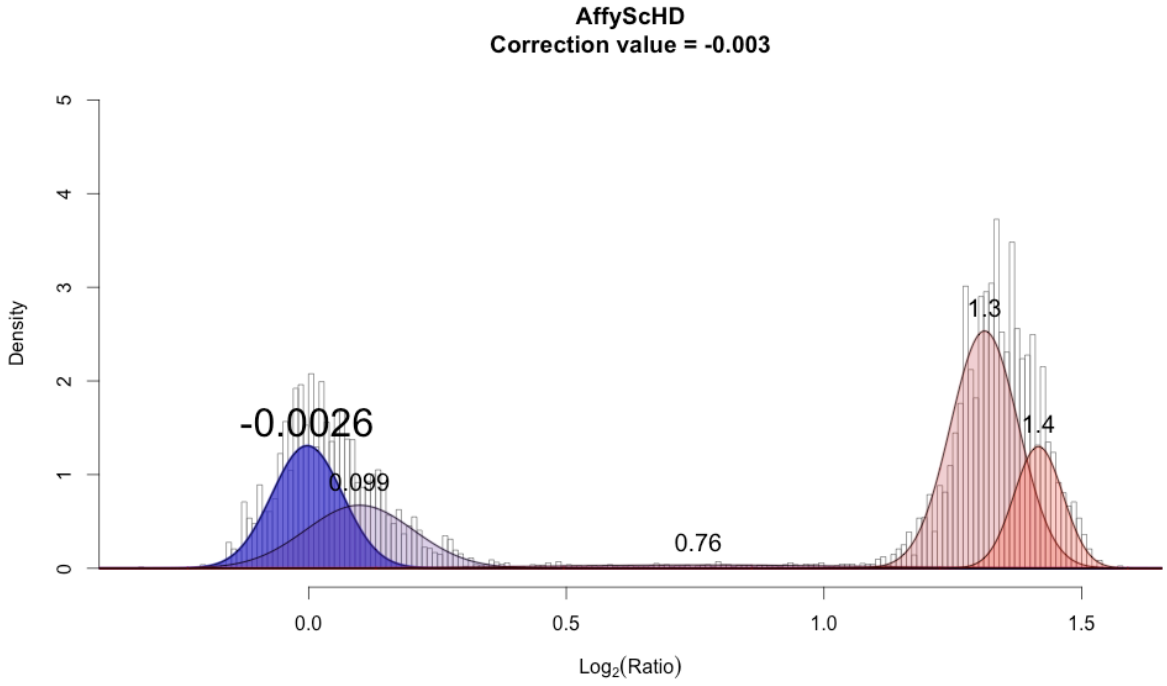


Figure 19 : dans rCGH, la fonction plotDensity() permet de visualiser le mélange de densités et la décision prise pour la centralisation (en gras).

5.2.7.2 Visualisation des profils

Pour la visualisation et l'interprétation d'un profil, plusieurs fonctions graphiques sont proposées (Figure 20):

1) *Le profil peut être visualisé seul à l'aide de la fonction plotProfile()*
2) *Le profil des déséquilibres alléliques (LOH), lorsque cette information est disponible, peut être généré à l'aide de la fonction plotLOH()*
3) *Un rapport graphique complet, combinant ces 2 visualisations, peut être généré à l'aide de la fonction multiplot().*

Ces fonctions graphiques autorisent également la localisation d'un ou plusieurs gènes, lorsque des symboles HUGO sont passés en argument.

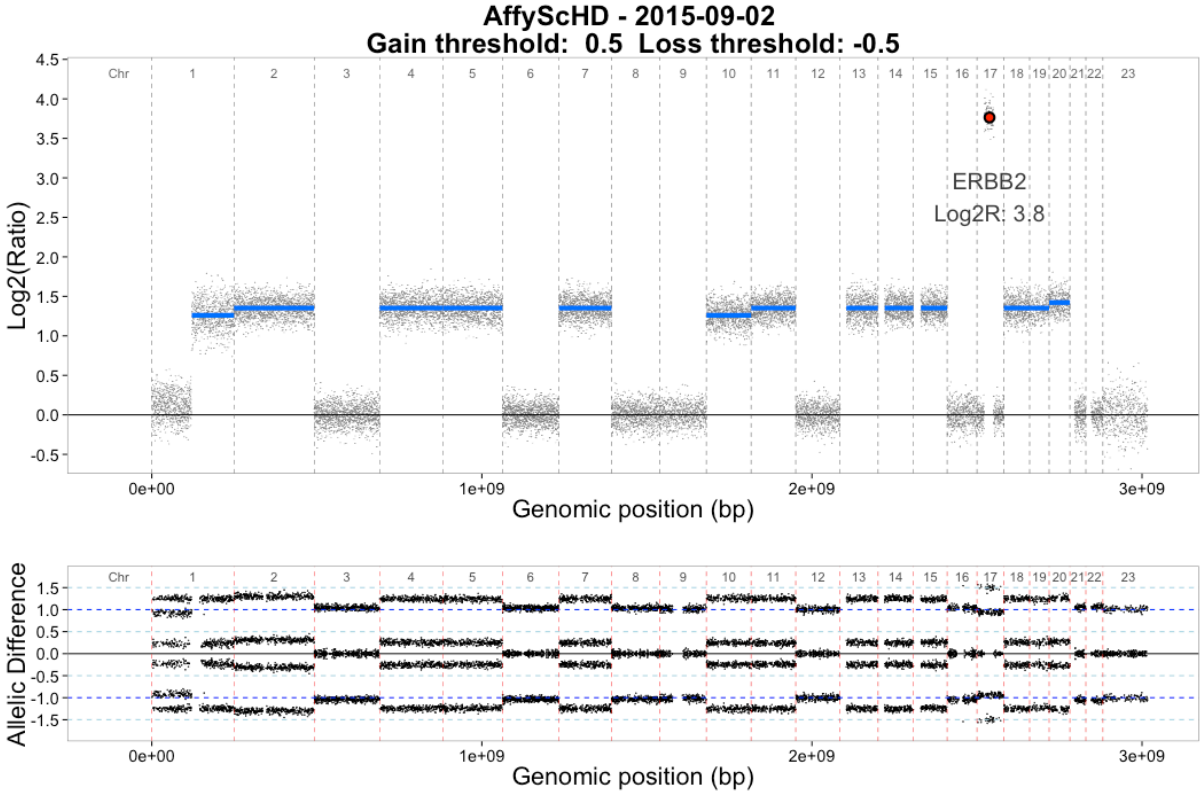


Figure 20 : la fonction multiplot() combine la vue du profil génomique et le profil des déséquilibres alléliques en un rapport graphique unique, sur lequel un ou plusieurs gènes d'intérêt peuvent être localisés. Chaque vue peut également être éditée individuellement à l'aide des fonctions plotProfile() et plotLOH().

5.2.8 Flexibilité

5.2.8.1 Paramétrage

Pour faciliter l'analyse des données, en particulier l'analyse en batch, les fonctions rCGH utilisent des arguments ayant des valeurs par défaut.

Toutefois, et afin de garder un maximum de flexibilité, tous ces arguments peuvent être spécifiés par l'utilisateur.

5.2.8.2 Taille des segments

En particulier, l'étape de segmentation permet de spécifier une taille minimale, exprimée en Kb, pour accepter un segment. Si la segmentation génère des segments inférieurs à cette limite, ils sont réassociés au segment contigu le plus proche en terme de LRR.

5.2.8.3 Centrage

Conscients des possibles ambiguïtés de centrage des profils, nous avons intégré à rCGH la possibilité d'ajuster le profil sur une autre valeur de centrage, sans pour autant devoir reprendre l'ensemble du processus.
Après analyse visuelle du profil, une fonction dédiée permet de définir une nouvelle valeur de centrage, estimée plus appropriée. Le profil est ajusté sur cette valeur, et le nouveau paramètre stocké pour traçabilité (Figure 21).

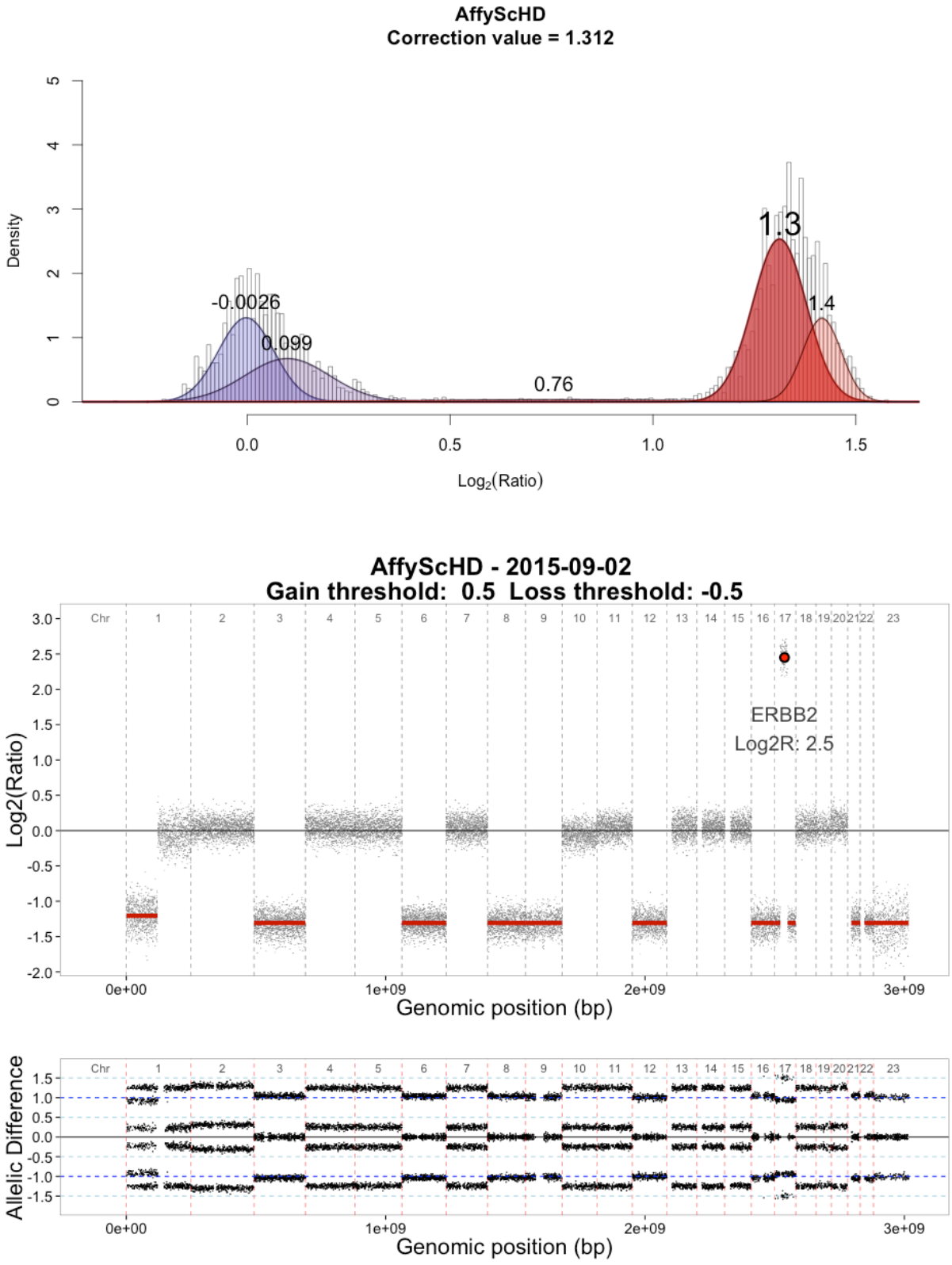


Figure 21: La fonction recenter() permet de choisir une nouvelle valeur de centralisation, sans avoir à reprendre le processus complet d'analyse.

5.2.9 Visualisation dynamique

Un des objectifs était de rendre plus flexible la lecture d'un profil et permettre certains ajustements, tout en ayant un contrôle visuel et dynamique sur les choix décisionnels.

Dans ce but, nous avons intégré au package rCGH une fonction de visualisation interactive : la fonction view() charge les données de segmentation stockées dans un objet 'rCGH', et les utilise pour reconstruire l'image du profil génomique et l'afficher dans un navigateur ; la liste des gènes, et leur valeur LRR dans le profil sont alors disponibles dans un second onglet (Figure 22).

Un panneau de commande permet d'interagir avec le profil, et propose à l'utilisateur des options d'affichage et d'ajustement du profil :

- *Afficher et localiser un gène d'intérêt.*
- *Afficher la totalité du profil, ou un chromosome unique.*
- *Modifier le code couleur des gains et pertes chromosomiques.*
- *Recentrer le profil.*
- *Supprimer les segments dont la longueur est inférieure à une valeur spécifiée. Dans ce cas, ces segments sont associés au segment le plus proche.*
- *Ajuster l'échelle de l'axe des ordonnées.*
- *Définir les valeurs seuils de gain et/ou perte chromosomique.*

Chaque ajustement - centralisation, suppression des segments courts – induit une mise à jour automatique des valeurs de gènes.

Enfin, le panneau de commande permet d'exporter et sauvegarder les modifications effectuées.

Ces solutions ont été développées sur la base du package shiny (Chang et al. 2015), permettant d'intégrer les langages html et CSS aux fonctions R.

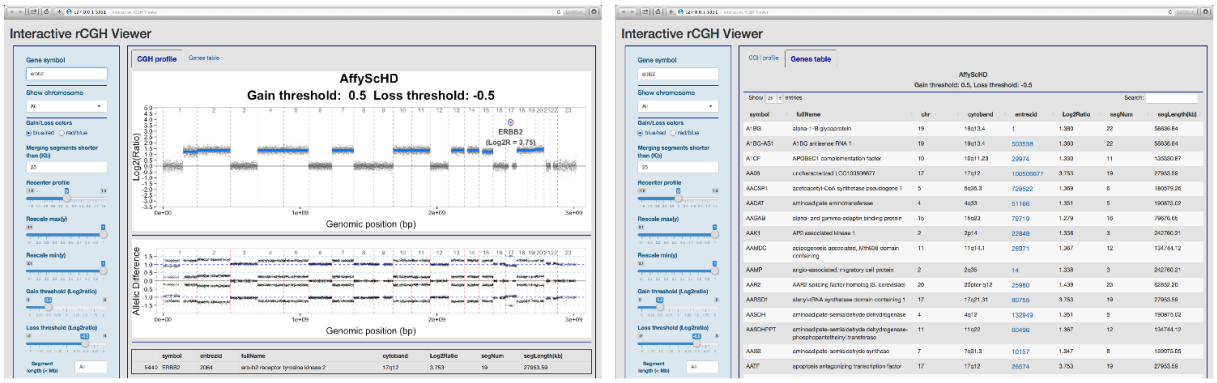


Figure 22 : le package rCGH propose une visualisation interactive des profils. La fonction view() utilise les données de segmentation pour reconstruire le profil génomique (gauche), et rendre accessible les valeurs des gènes (droite). Le panneau de contrôle (en bleu) permet d'afficher un gène d'intérêt, ajuster la vue, et partiellement redéfinir la segmentation.

5.2.10 Une version serveur

Les programmes de screening moléculaires pour la médecine de précision supposent une étape clé : la discussion, lors de comités scientifiques, des altérations génomiques présentes et leur pertinence pour une orientation thérapeutique.

En l'absence de règles de décision précises quant aux valeurs d'amplification et tailles d'amplicons à considérer, il nous a semblé opportun de développer des outils pouvant être partagés, et facilitant la discussion des profils génomiques en comité scientifique.

Pour ce faire, nous avons développé une version serveur reprenant les mêmes fonctionnalités que la fonction view() disponible dans le package rCGH.

Le serveur 'aCGH viewer' permet à tout utilisateur de visualiser, en ligne, un profil génomique à partir d'une simple table de segmentation téléchargée depuis un espace de stockage local. Ce serveur est cependant plus flexible en ce qu'il peut supporter des tables de segmentation générées par d'autres programmes que rCGH, sous réserve que ces tables respectent un format adapté.

Comme pour view(), la version en ligne permet d'interagir avec le profil, et d'exporter les modifications apportées (Figure 23).

Pour préserver la confidentialité des données, le serveur aCGH viewer a été implémenté de manière à minimiser le stockage des données téléchargées.

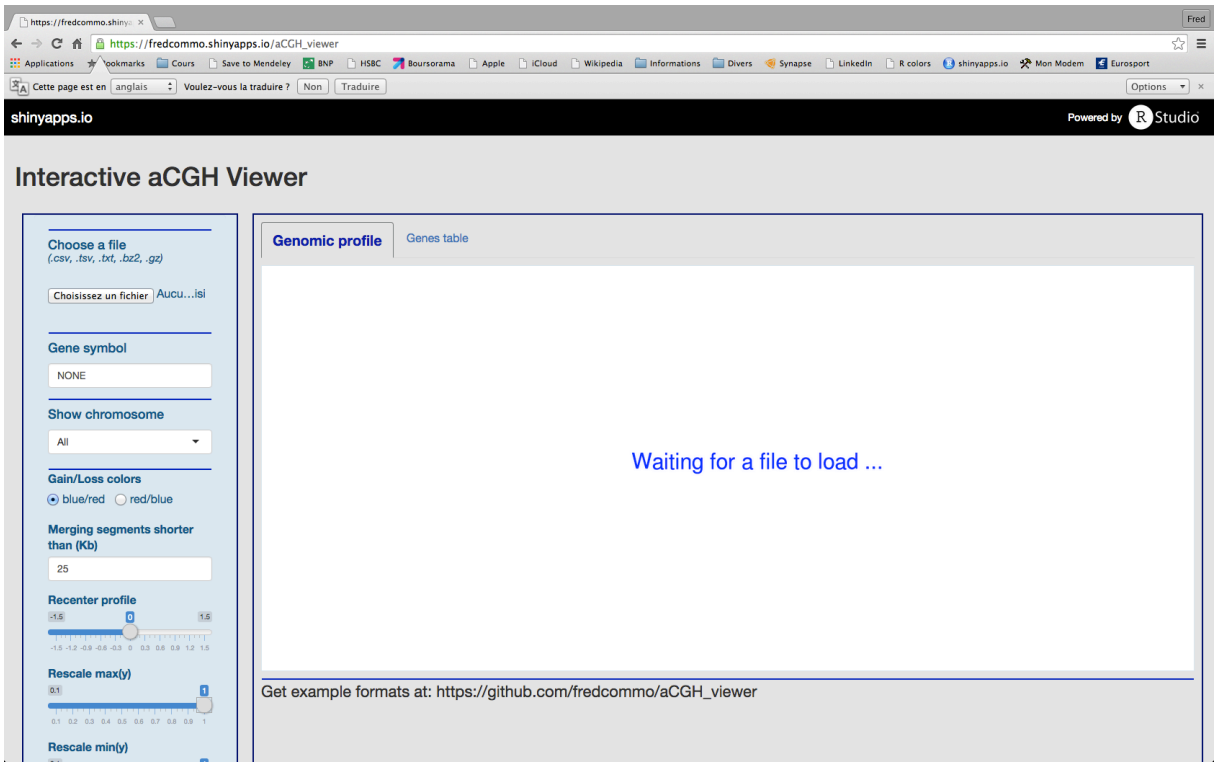


Figure 23 : Une version en ligne de visualisation interactive est disponible. Cette application permet de visualiser et interagir avec un profil génomique reconstruit à partir d'une simple table de segmentation. Les modifications peuvent être discutées lors de comités scientifiques, puis exportées.

'aCGH viewer' est une application gratuite en ligne accessible à l'adresse :
https://fredcommo.shinyapps.io/aCGH_viewer
L'outil peut aussi être installé sur tout autre serveur (au sein d'une institution par exemple), à partir du code mis à disposition sur github :
https://github.com/fredcommo/aCGH_viewer

5.2.11 Validation

Afin de valider notre package rCGH, nous avons utilisé les données de la série CCLE, contenant 995 profils de lignées cellulaires, générés à partir de puces Affymetrix SNP-6.0.

Les fichiers CEL ont été convertis à l'aide du package SNP6, et les fichiers générés ont été analysés à l'aide de rCGH. Les profils ainsi obtenus ont été comparés à ceux publiés par Barretina et col. (Barretina et al. 2012), et disponibles sur Gene Expression Omnibus (GEO : GSE36138) : 942 des 995 profils publiés étaient disponibles sous cette référence.

La correspondance entre profils appariés, CCLE et rCGH, a été mesurée par les corrélations de Pearson, les distances gène-à-gène, et par comparaison des corrélations de Pearson entre profils génomiques et expression des gènes.

5.2.11.1 Correspondance entre profils

La médiane des corrélations était de 0.913, indiquant une très bonne correspondance entre les segmentations des profils publiés et ceux obtenus par rCGH (Figure 24).

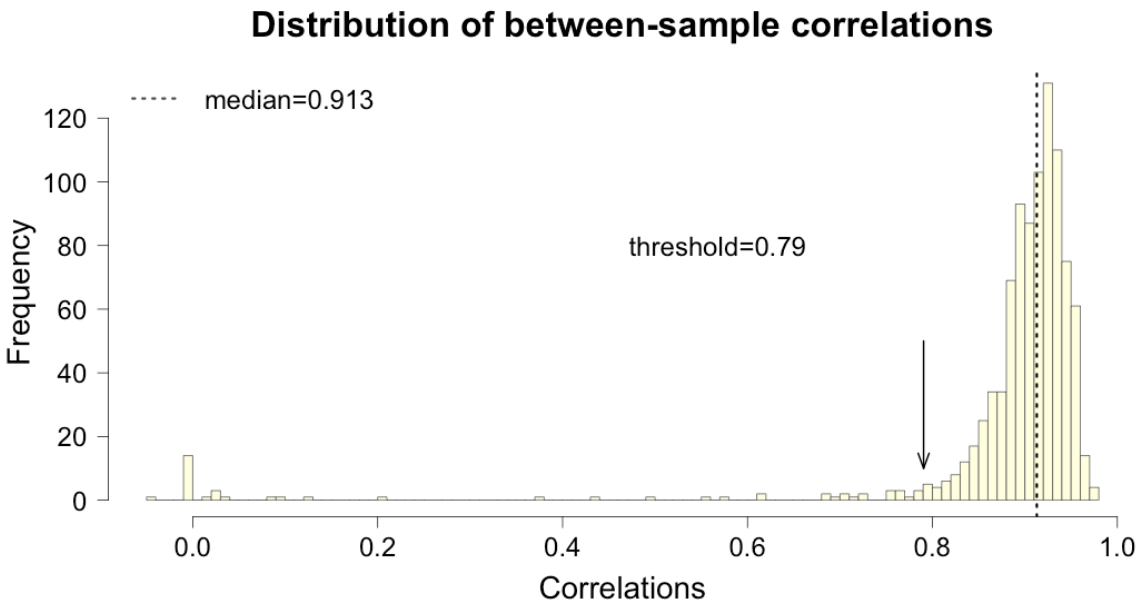


Figure 24 : Corrélations entre les profils CCLE publiés et ceux analysés par rCGH. La médiane des corrélations était de 0.913. Compte tenu du seuil de corrélation acceptable (rho = 0.79), 910 des 942 profils (94.7%) sont apparus comme bien corrélés.

Après transformation de Fisher des valeurs rho, $z = \frac{1}{2}\log\left(\frac{1+rho}{1-rho}\right)$*, nous avons utilisé le quantile* $q_{1e\text{-}2}(z) = 1.073$ *comme limite de corrélation acceptable. La*

transformation inverse, $rho = \frac{e^{2z}-1}{e^{2z}+1}$*, nous a permis de définir rho* = 0.79 comme valeur seuil de corrélation acceptable. Considérant ce seuil, 94.7% des profils montraient une bonne corrélation.*

Les profils peu corrélés étaient caractérisés par un faible nombre de segments de petite taille pouvant potentiellement être associés à du bruit.

5.2.11.2 Mesures des distances

Les mesures de distance donnent un meilleur reflet des différences liées au centrage, différentes centralisations ne correspondant qu'à une translation des valeurs, sans impact sur les corrélations entre profils.

Pour chaque profil apparié i, nous avons considéré $d_i = \frac{1}{n}\sum_{g=1}^{n}\left(y_{ig} - x_{ig}\right)$ *la moyenne des différences de signaux pour chaque gène g, g = 1,…,n.*

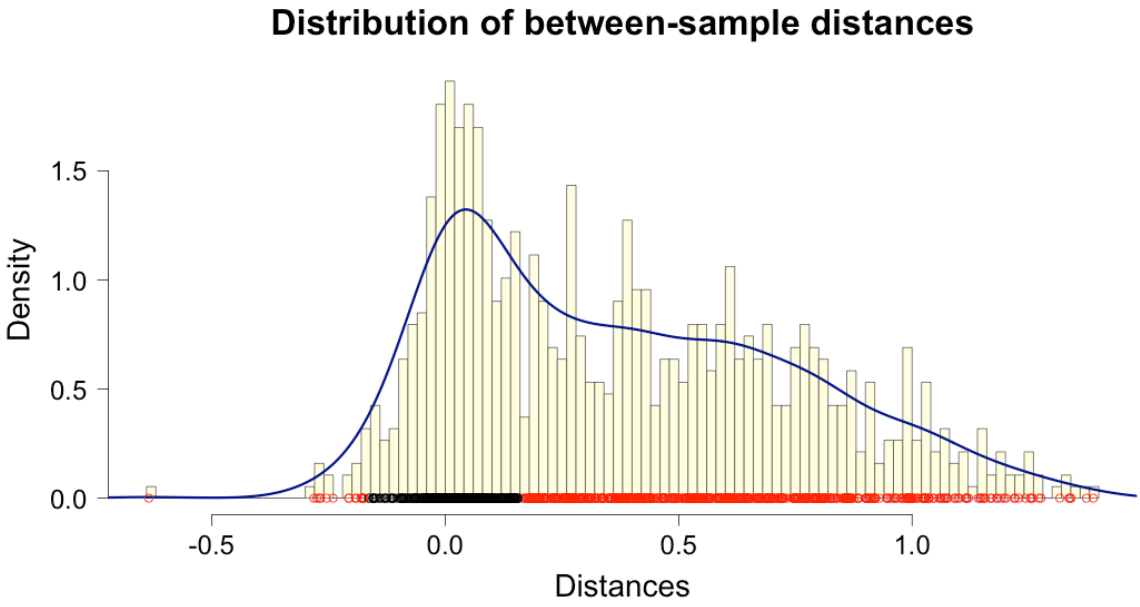


Figure 25: Distribution des distances. L'analyse des distances entre profils identifie deux situations : 36.2% des profils générés par rCGH montraient des différences négligeables avec les profils originaux (en noir), tandis que 63.8% montraient des différences liées à des choix différents de centralisation, ou à des différences d'amplitude.

La distribution de ces différences moyennes montrait 2 populations (Figure 25): pour 36.2% des profils, les différences moyennes entre gènes étaient proche de 0

(de -0.16 à 0.16). Pour les 63.8% restants, les différences étaient liées à des différences dans les choix de centralisation, mais aussi à des différences dans l'amplitude des signaux (Figure 26).

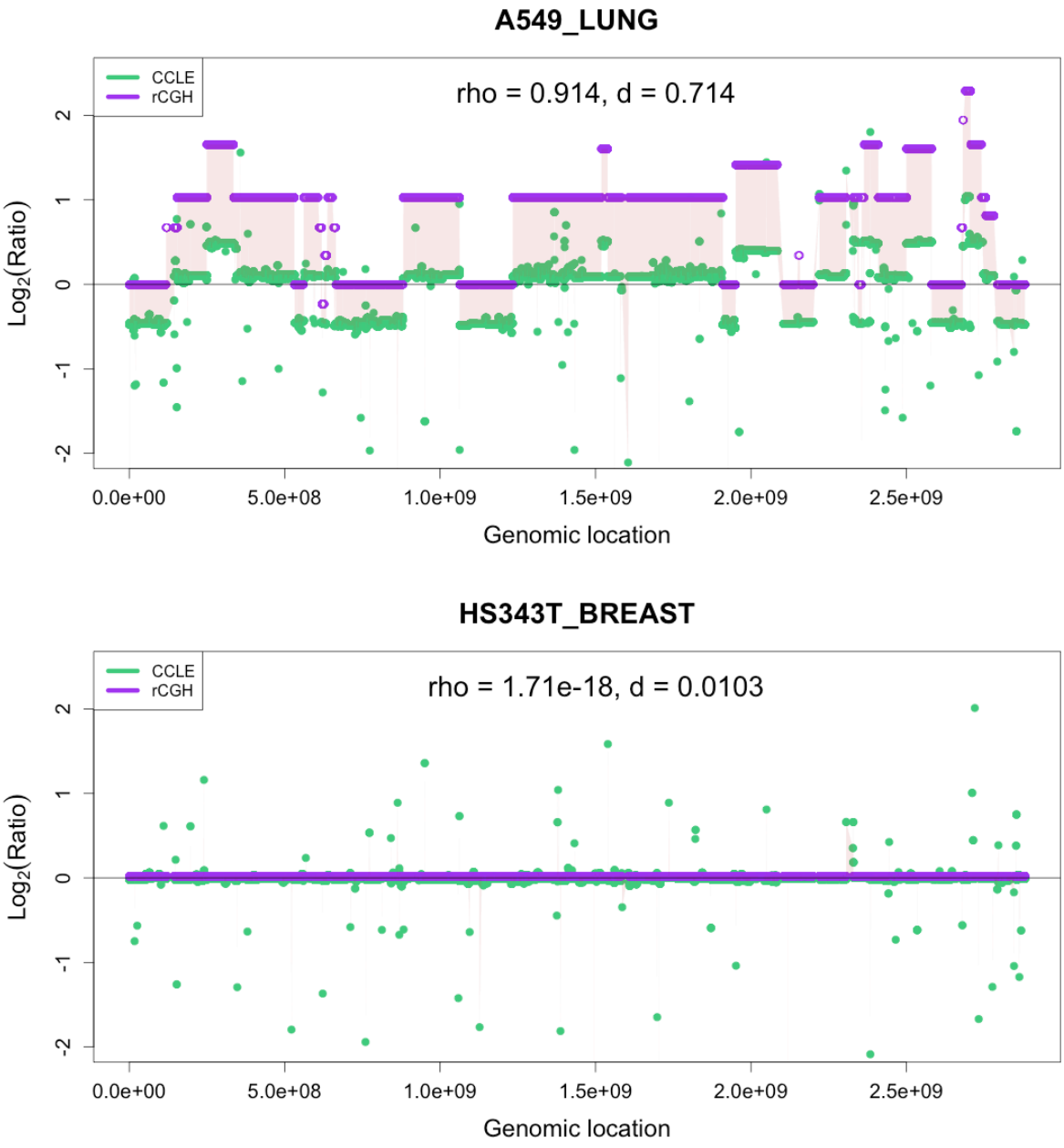


Figure 26: Corrélations et distances. En haut, la lignée A549 montre une très bonne corrélation, mais des différences importantes d'interprétation, liées à un centrage et des amplitudes différents. En bas, la lignée HS343T montre une corrélation proche de 0, en raison de la présence de signaux de très petite taille dans le profil original, pouvant être associés à du bruit.

5.2.11.3 Corrélations avec l'expression génique

Une dernière étape de validation a consisté à comparer les corrélations entre profils génomiques et expression des gènes (les valeurs d'expression génique pour la série CCLE sont disponibles sur GEO : GSE36133).

Pour chaque lignée cellulaire, son profil génomique, original et généré par rCGH, a été corrélé à l'expression des gènes (corrélation de Pearson). La correspondance entre les 2 vecteurs de corrélation a elle-même été estimée par une corrélation de Pearson.

Les 2 vecteurs de corrélations étaient extrêmement bien corrélés entre eux (rho = 0.994), indiquant que lorsque l'expression d'un gène apparaissait corrélée aux altérations génomiques dans CCLE, elle était détectée de même avec les profils rCGH (Figure 27).

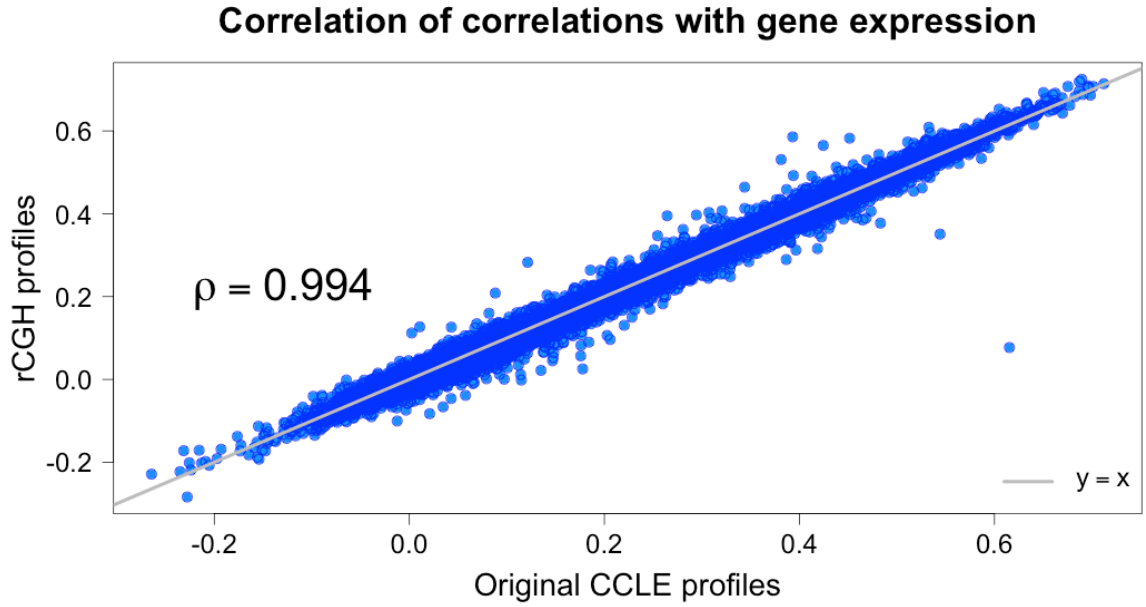


Figure 27: Corrélation des corrélations: les corrélations entre expression génique et profils génomiques montre une très bonne correspondance entre les données CCLE et l'analyse par rCGH.

Cependant, 53.7% des valeurs de corrélations observées dans CCLE étaient supérieures à celles observées avec les profils rCGH, et un test de Student apparié montrait une différence moyenne ente corrélations de 0.0018, significativement différente de 0 ($p < 1^{e-3}$).

Cette différence pourrait s'expliquer par un niveau de segmentation plus élevé dans les profils rCGH (figure 24), conduisant à une plus grande discrétisation des données. Cette discrétisation peut directement impacter et réduire la covariance, utilisée au numérateur dans le calcul de la corrélation de Pearson.

5.2.12 Commentaires

A notre connaissance, rCGH est le seul programme complet d'analyse aCGH pouvant supporter des données issues de plusieurs plateformes microarray, et proposant des visualisations interactives aussi développées.
Implémenté sous la forme d'un package R, rCGH est, en outre, portable et distribuable via Bioconductor : http://bioconductor.org/packages/rCGH/

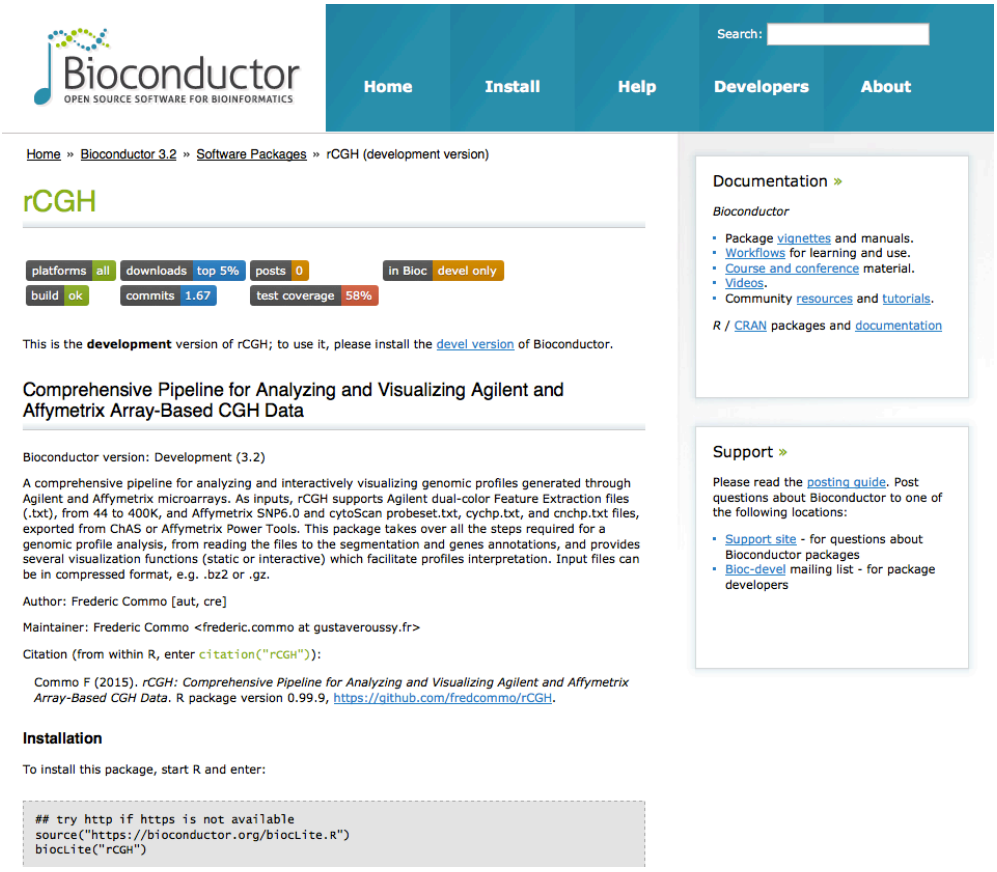
Avec rCGH, nous avons implémenté des solutions innovantes pour la réduction du bruit et le centrage des profils. Nous avons également apporté des optimisations à des solutions existantes comme l'algorithme CBS, utilisé pour la segmentation : rCGH propose un paramétrage facilité, guidé par la qualité des données.
Toutefois, et pour garder une totale flexibilité, tous les arguments utilisés dans les fonctions peuvent être spécifiés par l'utilisateur. L'ensemble des paramètres d'analyse sont sauvegardés pour assurer une totale traçabilité, et peuvent être rappelés à posteriori.

L'analyse des profils de près de 950 lignées cellulaires, et la comparaison des résultats avec les profils déjà publiés, a montré une très bonne performance du package rCGH en utilisation automatique, pré-paramétrée.

En nous concentrant sur une application en médecine de précision, nous avons également développé des solutions de visualisation interactives extrêmement avancées, permettant de manipuler et discuter un profil en comité scientifique, et pouvant faciliter la prise de décision.

Déployé sous la forme d'une application en ligne, cet outil de visualisation est à même de reconstruire un profil génomique à partir d'une simple table de segmentation. Elle supporte, en entrée, des données issues d'autres pipelines d'analyse : https://fredcommo.shinyapps.io/aCGH_viewer
Cette application peut également être installée sur un serveur, pour un usage interne à une institution. Le code et les instructions d'utilisation sont librement disponibles à l'adresse: https://github.com/fredcommo/aCGH_viewer

De futurs développements viseront à intégrer à rCGH la lecture de microarrays supplémentaires, tels que les nouvelles puces Affymetrix OncoScan.
Nous envisageons également de développer une nouvelle application de visualisation interactive, construite sur les langages javascript et D3js pour plus de fluidité et flexibilité.
Le package rCGH est aujourd'hui disponible sur Bioconductor :
http://bioconductor.org/packages/rCGH/



L'ensemble des fonctionnalités de rCGH sont décrites dans la vignette du package, en annexe 3.

Le code R complet est en libre accès à : https://github.com/fredcommo/rCGH

5.2.13 Publication

Une présentation du package rCGH a été récemment soumise à Bioinformatics pour publication.

Les résultats et méthodes supplémentaires sont proposées en annexe 6.

Bioinformatics

OXFORD UNIVERSITY PRESS | Bioinformatics

rCGH : a comprehensive array-based genomic profile platform for precision medicine

Journal:	*Bioinformatics*
Manuscript ID:	BIOINF-2015-1443
Category:	Applications Note
Date Submitted by the Author:	07-Sep-2015
Complete List of Authors:	Commo, Frederic; Gustave Roussy, UMR981 Guinney, Justin; Sage Bionetworks, Computational Biology Ferté, Charles; Gustave Roussy, Department of Medical Oncology Bot, Brian; Sage Bionetworks, Lefebvre, Celine; Gustave Roussy, UMR981 Soria, Jean-Charles; Gustave Roussy, Department of Medical Oncology André, Fabrice; Gustave Roussy, Department of Medical Oncology
Keywords:	Cancer, DNA, Genomics, Microarrays, Visualization

SCHOLARONE™ Manuscripts

rCGH : a comprehensive array-based genomic profile platform for precision medicine

Frederic Commo[1,2,*], Justin Guinney[2], Charles Ferté[1,2], Brian Bot[2], Celine Lefebvre[1], Jean-Charles Soria[1,3], Fabrice André[1,3]

[1]INSERM U981, Gustave Roussy, University Paris XI, Villejuif, France, [2]Sage Bionetworks, Seattle, WA, [3]Department of Medical Oncology, Gustave Roussy, Villejuif, France

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: We present rCGH, a comprehensive aCGH analysis workflow, integrating computational improvements, and functionalities specifically designed for precision medicine. rCGH supports the major microarray platforms, ensures a full traceability, and facilitates profiles interpretation and decision-making through sharable interactive visualizations.

Availability and implementation: The rCGH R package is available on bioconductor. The aCGH-viewer is available at https://fredcommo.shinyapps.io/aCGH_viewer, and the application implementation is freely available for installation at https://github.com/fredcommo/aCGH_viewer.

Contact: frederic.commo@gustaveroussy.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Precision Medicine and molecular screening programs aim at identifying individual cancer patients' molecular alterations, copy number alterations (CNA) or gene mutations, matching targeted therapies (Hollebecque *et al.*, 2013; André *et al.*, 2014). While Next-Generation Sequencing (NGS) is now widely used for identifying mutations, array-based Comparative Genomic Hybridization (aCGH) is a common platform for detecting CNAs (Laurent-Puig *et al.*, 2009; André *et al.*, 2013): the advantages of aCGH over NGS technologies include lower cost, rapid turn-around time, and lower computational overhead. In the context of precision medicine, CNA detection – alongside somatic mutation detection – is a critical component in the determination of clinically actionable genomic aberrations. Yet, significant technical challenges remain in the processing of aCGH data, and require new state-of-the-art tools for coordinating analysis and interpreting results.

2 Methods and Implementation

An aCGH analysis can be decomposed into 4 distinct phases (sup. figure 1): (1) Log_2 Relative Ratios (LRR) calculation (the sample DNA signals against a normal 2-copies DNA reference), (2) profile centralization, (3) profile segmentation, and (4) genomic profile interpretation to identify actionable genes affected by a CNA, in order to propose a matched therapeutic orientation.

The profile centralization defines a baseline - a neutral 2-copies level - from which CNA are estimated. We previously discussed the impact of centralization on aCGH analysis (Commo *et al.*, 2014), and rCGH implements the procedure described in the same paper. Briefly, the vector of LRRs is considered as a mixture of gaussian populations, and their respective proportion and parameters are estimated using an Expectation-Maximization (EM) algorithm. By default, the sub-population with a density peak higher than 50% of the highest density is considered as representing a neutral 2-copies state. Its mean is then used for centralizing the profile.

The segmentation step aims at identifying breakpoints in the LRRs continuity, each delimiting potentially gained or lost DNA segments. The rCGH segmentation relies on the Circular Binary Segmentation (CBS) (Olshen *et al.*, 2004), implemented in the *DNAcopy* R package. Although this algorithm is widely used (Willenbrock and Fridlyand, 2005), it suffers from several parameters to be specified *a priori*. In particular, the *'sdundo'* segmentation method mainly relies on two parameters: 1) a significance level α for the statistical test to accept points as breakpoints, and 2) the allowed difference between two consecutive segment means to keep them distinct (expressed in *DNAcopy* as a number of standard deviations, *undo.SD*). Instead of using arbitrary values, rCGH introduces a data-driven parameterization: given a fixed α, the corresponding optimal '*undo.SD*' value is estimated from the median absolute deviation (MAD), a widely used noise estimator (supplementary methods). This optimization greatly facilitates the use of this algorithm for routine practice, and standardizes the parameterization through a data-driven rule.

The final interpretation, and therefore decision-making regarding a therapeutic orientation, relies on the actionable genes status, defined with respect to gain/loss thresholds and alteration lengths.

Defining such thresholds is often arbitrary: 1) there is no consensus on which LRR values correspond to biologically relevant CNAs, 2) focal alterations, possibly referring to significantly recurrent alterations within a cohort (Mermel *et al.*, 2011; Yuan *et al.*, 2012), are not clearly defined when transposed to the interpretation of unique profiles.

rCGH provides an interactive visualization tools that allows the user to visualize and manipulate genomic profiles from within a web interface (Fig 1).

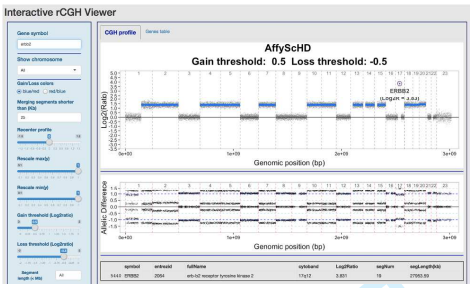


Fig1: Interactive visualization. The genomic profile, as well as the LOH profile (when available), are displayed on the *CGH profile* tab, while the gene values are accessible through the *Genes table* tab. The command panel (on left) can be used to display a gene of interest, to recenter the entire profile, and to specify several decision parameters (gain/loss threshold, segment length). Gene values are updated automatically, and profiles and table can be re-exported after modification.

Two primary perspectives are provided: visualization of CNAs along DNA strands or a gene-centric table. The latter includes gene specific LRR values and corresponding segment lengths. When available on microarrays, loss of heterozygosity (LOH) expressed as the A/B allelic difference is also provided. A command window provides control over display parameters including re-centering, merging short segments, and gain or loss thresholds. (see supplementary methods for a full description). Finally, both the genomic profile and the genes table can be re-exported, in ready-to-publish quality and xls format, respectively, including the changes applied on the profile.

3 Supported files

As inputs rCGH supports Agilent Human CGH data, from 44K to 400K arrays, and Affymetrix, SNP6 and cytoScanHD. All are provided in text format by platform specific softwares: standard Agilent text files are exported from Agilent Feature Extraction software (FE), while Affymetrix cychp.txt, cnchp.txt, or probeset.txt files are obtained by processing Affymetrix CEL files through ChAS or Affymetrix Power Tools (APT) softwares: both are freely available at http://www.affymetrix.com.

4 rCGH outputs

rCGH stores all the original and computed data, as well as the workflow parameters, to ensure traceability. Segmentation tables are of the same format as standard CBS segmentation outputs, completed with the segment lengths, and the within-segment LRR standard deviation.

5 Web Server version

Independently of rCGH, we have developed *aCGH-viewer*: an interactive visualization available as a free web application. Its implementation is freely available for installation on a server. As inputs, the application requires segmentation tables built through either *rCGH*, or any other workflow, provided the data is of the same form as the standard CBS outputs. This application allows a profile to be shared, discussed and annotated in tumor board committees, and finally saved for traceability.

6 Conclusion

In this work, we present the R package rCGH: a comprehensive aCGH analysis workflow, with features and functionalities particularly well adapted to precision medicine. rCGH ensures the traceability of the entire process, and provides interactive visualization tools allowing to better interpret - and potentially reprocess - genomic profiles. The rCGH workflow, and its web-server application are operating system independent, and can efficiently assist oncologists to discuss alterations in genomic profiles, and to identify matched therapeutic orientations.

7 Funding

This work was supported by the Integrative Cancer Biology Program of the National Cancer Institute (U54CA149237 to F.C., C.F., and J.G.), Unicancer, the ARC foundation, the Breast Cancer Research foundation, and Odyssea.

The authors declare that they have no competing financial interest for this work.

References

André,F. *et al.* (2014) Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIR01/UNICANCER). *Lancet Oncol.*, **15**, 267–74.

André,F. *et al.* (2013) Targeting FGFR with dovitinib (TKI258): preclinical and clinical data in breast cancer. *Clin. Cancer Res.*, **19**, 3693–702.

Commo,F. *et al.* (2014) Impact of centralization on aCGH-based genomic profiles for precision medicine in oncology. *Ann. Oncol.*

Hollebecque,A. *et al.* (2013) Molecular screening for cancer treatment optimization (MOSCATO 01): A prospective molecular triage trial—Interim results. In, *ASCO Annual Meeting.*, p. Abstract 2512.

Laurent-Puig,P. *et al.* (2009) Analysis of PTEN, BRAF, and EGFR status in determining benefit from cetuximab therapy in wild-type KRAS metastatic colon cancer. *J. Clin. Oncol.*, **27**, 5924–30.

Mermel,C.H. *et al.* (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, **12**, R41.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–72.

Willenbrock,H. and Fridlyand,J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 4084–91.

Yuan,X. *et al.* (2012) Comparative analysis of methods for identifying recurrent copy number alterations in cancer. *PLoS One*, **7**, e52516.

6 Discussion

Sur la base de nouvelles connaissances en génétique, et bénéficiant d'évolutions technologiques importantes, de nombreux essais de screening moléculaire ont été initiés au cours de cette décennie. Il s'agissait de valider une nouvelle approche pour la prise en charge des patients en oncologie : la médecine de précision.
Cette nouvelle stratégie personnalisée vise à identifier des altérations moléculaires, mutations, réarrangements ou anomalies de nombre de copies, puis orienter les patients vers des thérapies spécifiques, ciblant les gènes et fonctions identifiés comme altérés.

6.1 Des résultats mitigés

Si ces programmes ont en commun de mettre en évidence des altérations moléculaires chez un nombre conséquent de patients, leurs résultats peuvent diverger quant à la proportion de patients effectivement orientés vers des thérapies ciblées, et quant aux bénéfices d'une prise en charge guidée par les anomalies moléculaires ; ils se caractérisent par des choix technologiques et stratégiques différents.

Pour l'ensemble des programmes de screening moléculaire, une anomalie moléculaire, au moins, est détectée dans 39 à 70% des cas, ces proportions pouvant varier selon les études, en fonction des gènes explorés et des critères définissant une anomalie. Cependant, la proportion de patients effectivement orientés vers une thérapie ciblée, varie, quant à elle, de 22 (Safir01) à 100% (IMPACT). En terme de bénéfice, les programmes IMPACT et BATTLE montrent que la prise en charge personnalisée améliore, modestement mais significativement, la survie des patients (Kim et al. 2011; Tsimberidou et al. 2014), tandis que le programme SHIVA conclut à l'absence de bénéfice (Tourneau et al. 2015).

Ces disparités importantes peuvent s'expliquer par des choix stratégiques différents, mais aussi par la disponibilité des thérapies associées aux anomalies : IMPACT a rendu éligibles les patients ayant une des anomalies recherchées, tandis que André et collaborateurs (André et al. 2014) *explorent un panel plus large d'altérations, mais avancent la difficulté d'accéder à certaines thérapies ciblées comme étant un facteur limitant aux approches personnalisées.*

Les différences, tant dans les pathologies incluses que dans les anomalies explorées et les choix technologiques ou stratégiques, rendent toutefois difficile la comparaison des résultats présentés : IMPACT s'est concentré sur quelques gènes d'intérêt, et a préféré des technologies dédiées : séquençage de gènes ciblés, FISH et immunohistochimie, à l'utilisation de techniques à haut débit (Tsimberidou et al. 2012; Tsimberidou et al. 2014), tandis que d'autres, comme l''étude pilote MI-ONCOSEQ, ont fait le pari d'une analyse exhaustive de l'ADN et l'ARN (Roychowdhury et al. 2011). BATTLE se concentre sur les tumeurs du poumon non à petites cellules, considère 7 gènes et 4 thérapies ciblées, et introduit un algorithme adaptatif pour l'orientation des patients (Kim et al. 2011). SHIVA (Tourneau et al. 2015) intègre plus largement les tumeurs réfractaires aux thérapies standard, sans limite d'origine tissulaire ou type histologique, tout en se concentrant sur une liste limitée de gènes à explorer, et fait le choix d'une randomisation plus conventionnelle en comparant la survie chez les patients orientés vers une thérapie ciblée à ceux recevant un traitement standard.

Une autre conséquence des différents choix stratégiques, exploration exhaustive ou liste restreinte de gènes d'intérêts, est la difficulté d'intégrer l'ensemble des résultats dans une discussion d'orientation. Dans le premier cas, l'analyse peut identifier un grand nombre d'anomalies dont les effets biologiques sont mal connus, ou pour lesquelles aucune alternative thérapeutique ne peut être proposée (Roychowdhury et al. 2011). A l'inverse, une recherche ciblée semble simplifier l'interprétation des résultats, mais une conséquence d'une telle approche pourrait

être de réduire les capacités d'analyse à posteriori pour l'amélioration des algorithmes décisionnels, ainsi que pour l'identification de nouveaux biomarqueurs potentiels.

6.2 Définir les meilleures stratégies et règles de décision

Malgré des évolutions conséquentes en matière de séquençage, l'analyse de profils génomiques par aCGH fait toutefois partie intégrante de nombreux programmes de screening moléculaire : le projet TCGA pour la caractérisation de tumeurs (Hudson et al. 2010; Cline et al. 2013), ou les projets CCLE et CGE pour l'analyse moléculaire de lignées cellulaires (Barretina et al. 2012; Garnett et al. 2012), ont très largement utilisé cette technologie, et les données produites restent parmi les plus utilisées par la communauté scientifique.
En médecine de précision, cette technologie permet d'identifier, rapidement et à moindre coût, des altérations de nombre de copies portant sur des gènes dont les fonctions peuvent être ciblées par des thérapies spécifiques. Dans ce cadre, l'interprétation du profil et l'identification des régions d'intérêt font partie d'un processus de décision collégiale : un comité scientifique intégrera à la discussion pour une orientation thérapeutique, les gènes potentiellement actionnables dont le nombre de copies est altéré.

La décision de définir, ou non, un gène comme amplifié (ou délété) repose essentiellement sur la valeur de gain (ou perte), estimée par rapport à un l'ADN de référence, mais aussi sur la longueur du segment contenant ce gène. L'analyse vise souvent à privilégier les altérations de fortes amplitudes, localisées sur des régions restreintes en taille, dites « focales ». Cependant, il n'existe aucune règle formelle, ni consensus, permettant de définir ce qu'est une amplification focale et quelle valeur seuil de gain doit être considérée, et certaines incertitudes persistent quant à la pertinence de certains gènes.

Les amplifications du gène EGFR, lorsqu'elles sont estimées par FISH, ont été décrites comme prédictives d'une sensibilité aux thérapies anit-EGFR dans les cancers du colon sans mutation KRAS (Cappuzzo, Finocchiaro, et al. 2008; Ålgars et al. 2011). Cependant, il a été montré que EGFR pouvait ne pas être le seul marqueur capable de prédire une réponse à ces inhibiteur, des patients sans amplification pouvant aussi monter des réponses favorables à l'erlotinib (Chung et al. 2005). Tsao et collaborateurs, quant à eux, ne concluent pas à la pertinence de EGFR en tant que marqueur prédictif de survie, dans une étude sur des tumeurs du poumon pour lesquelles les mutations, expression génique et nombre de copies de EGFR était explorés (Tsao et al. 2005). Les choix technologiques pour la recherche d'altérations et les seuils de décision pourraient être, dans ce cas, à l'origine de conclusions ambigües (Sesboüé et al. 2012; Hutchinson et al. 2015).

Si des études précliniques ont semblé indiquer que des amplifications de FGFR1 pouvaient être associées à une sensibilité aux inhibiteurs de FGFR (Dutt et al. 2011), d'autres travaux n'ont pas confirmé l'intérêt de ces amplifications pour la décision thérapeutique. Il est toutefois possible que le seuil définissant une amplification pertinente soit là aussi à spécifier (André et al. 2013; Wynes et al. 2014).

La recherche d'amplifications de ESR1, rendant possible l'orientation des patientes atteintes de cancer du sein vers une hormonothérapie, est une autre illustration du poids des choix techniques et de la définition des seuils de décision. Yu et collaborateurs ont montré que la prévalence des amplifications de ESR1 était de l'ordre de 20% dans les tumeurs du sein analysées par FISH, mais inférieure à 5% (de 0 à 4.5%) dans les études ayant utilisé l'aCGH (Yu & Shao 2011). Pour note, considérant une valeur de gain de Log2Ratio > 1, et des amplicons de longueur < 10Mb, les amplifications de ESR1 étaient observées dans 4.6% des cas dans l'étude Safir-01.

Si les valeurs de gain restent à spécifier, la caractérisation des anomalies de nombre de copies par aCGH suppose d'abord de définir une ligne de base, un niveau d'équilibre à 2 copies, à partir de laquelle les gains et pertes de régions chromosomiques seront estimées.

Nous avons montré, en étudiant les profils génomiques de lignées cellulaires caractérisées, que le choix d'une valeur de centrage pouvait s'avérer complexe ; l'analyse de la distribution des valeurs LRR peut révéler plusieurs pics de densité, les plus importants représentant chacun une valeur possible de ratio équilibré. Le choix de la population majoritaire pour le centrage conduisant à ajuster le profil sur la ploïdie prédominante, dans les cas d'aneuploïdies, cette valeur peut ne pas correspondre à un ratio 2/2, où le dénominateur représente le nombre de copies attendues dans l'ADN normal de contrôle.

Le problème du choix de centralisation a relativement peu été évoqué dans la littérature :

Staaf et collaborateurs (Staaf et al. 2007) soulèvent ce problème, et proposent un meilleur ajustement des biais de cyanines par régression locale sur une sous-population de signaux. Ils suggèrent d'appliquer cette régression sur la population majoritaire, définie après une classification des LRR par la méthode des Kmeans à 3 centres. La médiane des valeurs ainsi ajustées, est ensuite utilisée pour corriger l'ensemble du profil. Plus tard, Chen et collaborateurs (Chen et al. 2008) introduisent l'analyse de la distribution des LRR comme un mélange de populations gaussiennes, et considèrent un seuil à 95% du pic majoritaire pour sélectionner une valeur de centrage.

Les résultats que nous avons observés sur les lignées caractérisées, ont cependant montré qu'un choix alternatif - une population dont le pic de densité est à 50% du pic majoritaire - conduit à des profils plus proches de la réalité. Nous montrons ensuite que des ambiguïtés de centrages peuvent également être observées sur des échantillons tumoraux : l'analyse des données aCGH des programmes SAFIR-01 (André et al. 2014) et MOSCATO (Hollebecque et al. 2013) a révélé que des centralisations alternatives pouvaient être choisies pour

respectivement 26.9% et 25.6% des profils. Regroupant les résultats des 2 études, une conséquence majeure d'un changement de stratégie de centralisation était de révéler de nouvelles altérations actionnables pour 69 des 592 patients (11.6%). Dans 41 cas (6.9%), une centralisation alternative révélait au moins une anomalie actionnable, dans des profils pour lesquels aucune anomalie n'avait été détectée, et dans 28 autres cas (4.7%), des altérations supplémentaires étaient identifiées, ouvrant potentiellement la voie à des choix thérapeutiques supplémentaires. Cette dernière situation pose aussi le problème de la priorisation des altérations, ou de leur intégration dans un algorithme décisionnel. Cet aspect a été abordé dans le programme BATTLE, et il est probable que des analyses rétrospectives permettrons d'optimiser ces règles de décision.

6.3 Développement d'une solution complète et flexible

Pour pallier l'absence de règle et faciliter la prise de décision, certains auteurs ont proposé différentes stratégies d'analyse permettant d'aider la prise de décision :

GAP (Popova et al. 2009) propose une alternative pertinente, construite sur la modélisation des Log_2Ratio et des fréquences alléliques, les deux étant ensuite combinés à une recherche de clusters après projection dans un plan. Un bénéfice de cette approche est que GAP propose des valeurs de gains exprimées en nombre de copies, plus faciles à interpréter et intégrer dans une discussion pour une décision thérapeutique. Mais bien que performante, cette méthode reste toutefois dédiée à l'analyse de microarrays comportant des sondes SNP, et fournissant une information sur les déséquilibres alléliques.

D'autres approches ont également été développées, mais elles ont généralement l'inconvénient d'être dédiées à des plateformes particulières :

PAIR (Yang et al. 2013), tout comme ASCAT (Van Loo et al. 2010), s'appuyant également sur les pertes d'hétérozygoties estimées à partir des sondes SNP, n'est applicable que sur des plateformes apportant cette information. CGHnormaliter (van Houte et al. 2009), utilisant sur une approche itérative de normalisation des signaux, n'est applicable qu'aux données issues d'hybridations à 2 couleurs.

D'autres programmes, tels que CGH Explorer (Lingjaerde et al. 2005) ou GISTIC (Mermel et al. 2011), sont dédiés à l'identification d'aberrations récurrentes au sein de cohortes de patients, et non à l'analyse de profils uniques.

Avec le package rCGH déposé sur Bioconductor, nous avons voulu développer un programme d'analyse à la fois flexible, portable et distribuable, capable d'accepter la majorité des plateformes aCGH utilisées : rCGH supporte actuellement les microarrays Agilent de 44 à 400K, ainsi que les microarrays Affymetrix SNP6.0 et cytoScan, et de futures implémentations permettront l'analyse de microarrays supplémentaires, tels que les Affymetrix oncoScan.
Utilisant largement la programmation orientée objet, rCGH permet le stockage des données originales et calculées, et assure la sauvegarde des paramètres d'analyse pour une parfaite traçabilité.
Ce package implémente de nouvelles solutions pour la réduction du bruit et l'estimation de la centralisation : en adoptant l'analyse de mélanges gaussiens, nous réduisons significativement le bruit dans le signal, ainsi que le nombre d'outliers. Cette approche est une alternative performante en terme de temps de calcul, comparée à l'identification par fenêtres glissantes adoptée dans d'autres procédures. De plus, nos travaux sur les lignées cellulaires caractérisées (Commo et al. 2015) *nous ont permis d'optimiser les règles de décision quant au choix d'une valeur d'ajustement définissant un niveau de base.*
Nous avons également apporté des optimisations à CBS, l'algorithme de segmentation le plus couramment utilisé (Venkatraman & Olshen 2007), en développant une paramétrisation simplifiée, guidée par la qualité des données.
Toutefois, nous avons fait le choix de laisser à l'utilisateur la possibilité de modifier les paramètres d'analyse, tout en contrôlant leurs effets à travers des outils de visualisation interactive. Ces outils n'ont aucun équivalent à notre connaissance.

Afin de faciliter la communication des résultats et la prise de décision, nous avons associé à ce package des outils de visualisation interactive, d'ores et déjà

disponibles en ligne, mais pouvant également être installés sur un serveur institutionnel, à partir du code que nous avons rendu public. Cette version serveur permet la diffusion et la manipulation de profils génomiques lors de comités scientifiques, et pourrait faciliter l'interprétation et la discussion des anomalies présentes.

7 Conclusion

Afin d'améliorer le traitement des données aCGH, et leur interprétation en médecine de précision, nous avons développé un workflow complet d'analyse, capable de prendre en charge les données issues des principales plateformes microarrays (Agilent, Affymetrix SNP6.0 et CytoScan). Ce workflow intègre plusieurs solutions facilitant la paramétrisation et l'automatisation des analyses, tout en assurant une traçabilité totale du processus.

Pour faciliter l'interprétation des profils génomiques et la prise de décision thérapeutique, nous avons associé à ce workflow des outils de visualisation interactive innovants, permettant d'affiner un profil génomique et de localiser les gènes d'intérêt. Une version serveur de ces outils a également été développée, afin de permettre la discussion des profils en comité de décision

L'ensemble du workflow est écrit en langage R et est disponible sous la forme d'un package sur le site Bioconductor. La version serveur est disponible en ligne, ou peut être installée sur un serveur local, à partir du code déposé sur github.

8 Références

Ålgars, A. et al., 2011. EGFR gene copy number assessment from areas with highest EGFR expression predicts response to anti-EGFR therapy in colorectal cancer. British journal of cancer, *105(2), pp.255–62.*

Amado, R.G. et al., 2008. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, *26(10), pp.1626–34.*

André, F. et al., 2014. Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIR01/UNICANCER). The lancet oncology, *15(3), pp.267–74.*

André, F. et al., 2013. Targeting FGFR with dovitinib (TKI258): preclinical and clinical data in breast cancer. Clinical cancer research : an official journal of the American Association for Cancer Research, *19(13), pp.3693–702.*

Arcila, M.E. et al., 2011. Rebiopsy of lung cancer patients with acquired resistance to EGFR inhibitors and enhanced detection of the T790M mutation using a locked nucleic acid-based assay. Clinical cancer research : an official journal of the American Association for Cancer Research, *17(5), pp.1169–80.*

Barretina, J. et al., 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, *483(7391), pp.603–7.*

Baselga, J. et al., 2005. Phase II study of efficacy, safety, and pharmacokinetics of trastuzumab monotherapy administered on a 3-weekly schedule. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, *23(10), pp.2162–71.*

Bass, A.J. et al., 2014. Comprehensive molecular characterization of gastric adenocarcinoma. Nature, *513(7517), pp.202–9.*

Beadle, G.W. & Tatum, E.L., 1941. Genetic Control of Biochemical Reactions in Neurospora. Proceedings of the National Academy of Sciences of the United States of America, *27(11), pp.499–506.*

Bedard, P.L. et al., 2013. Tumour heterogeneity in the clinic. Nature, *501(7467), pp.355–64.*

Ben-Yaacov, E. & Eldar, Y.C., 2008. A fast and flexible method for the segmentation of aCGH data. Bioinformatics (Oxford, England), *24(16), pp.i139–45.*

Bishop, J.M., 1991. Molecular themes in oncogenesis. Cell, *64(2), pp.235–48.*

Bonnet, F. et al., 2012. An array CGH based genomic instability index (G2I) is predictive of clinical outcome in breast cancer and reveals a subset of tumors without lymph node involvement

but with poor prognosis. BMC medical genomics, *5(1), p.54.*

Boveri, T., 2008. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. Journal of cell science, *121 Suppl , pp.1–84.*

Cappuzzo, F., Finocchiaro, G., et al., 2008. EGFR FISH assay predicts for response to cetuximab in chemotherapy refractory colorectal cancer patients. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO, *19(4), pp.717–23.*

Cappuzzo, F., Varella-Garcia, M., et al., 2008. Primary resistance to cetuximab therapy in EGFR FISH-positive colorectal cancer patients. British journal of cancer, *99(1), pp.83–9.*

Carter, P. et al., 1992. Humanization of an anti-p185HER2 antibody for human cancer therapy. Proceedings of the National Academy of Sciences of the United States of America, *89(10), pp.4285–9.*

Chang, W. et al., 2015. shiny: Web Application Framework for R (R package version 0.12.0).

Chen, H.-I.H. et al., 2008. A probe-density-based analysis method for array CGH data: simulation, normalization and centralization. Bioinformatics (Oxford, England), *24(16), pp.1749–56.*

Chin, L. et al., 1999. Essential role for oncogenic Ras in tumour maintenance. Nature, *400(6743), pp.468–72.*

Chin, L. et al., 2011. Making sense of cancer genomic data. Genes & development, *25(6), pp.534–55.*

Chung, K.Y. et al., 2005. Cetuximab shows activity in colorectal cancer patients with tumors that do not express the epidermal growth factor receptor by immunohistochemistry. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, *23(9), pp.1803–10.*

Cleveland, W.S. & Devlin, S.J., 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association, *83(403), pp.596–610.*

Cline, M.S. et al., 2013. Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser. Scientific reports, *3, p.2652.*

Collisson, E.A. et al., 2014. Comprehensive molecular profiling of lung adenocarcinoma. Nature, *511(7511), pp.543–50.*

Commo, F. et al., 2015. Impact of centralization on aCGH-based genomic profiles for precision medicine in oncology. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO, *26(3), pp.582–8.*

Cooper, C.S. et al., 1984. Molecular cloning of a new transforming gene from a chemically transformed human cell line. Nature, *311(5981), pp.29–33.*

Cottu, P.H. et al., 2008. Intratumoral heterogeneity of HER2/neu expression and its consequences for the management of advanced breast cancer. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO, *19(3), pp.595–7.*

Das, K. et al., 2014. Mutually exclusive FGFR2, HER2, and KRAS gene amplifications in gastric cancer revealed by multicolour FISH. Cancer letters, *353(2), pp.167–75.*

Demetri, G.D. et al., 2002. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. The New England journal of medicine, *347(7), pp.472–80.*

Dempster, A.P., Laird, N.M. & Rubin, D.B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, *39(1), pp.1–38.*

Diaz, L.A. et al., 2012. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. Nature, *486(7404), pp.537–40.*

Dillon, L.M. & Miller, T.W., 2014. Therapeutic targeting of cancers with loss of PTEN function. Current drug targets, *15(1), pp.65–79.*

Druker, B.J. et al., 2001. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. The New England journal of medicine, *344(14), pp.1031–7.*

Dulbecco, R., 1986. A turning point in cancer research: sequencing the human genome. Science (New York, N.Y.), *231(4742), pp.1055–6.*

Dutt, A. et al., 2011. Inhibitor-sensitive FGFR1 amplification in human non-small cell lung cancer. PloS one, *6(6), p.e20351.*

Eggermont, A.M.M. & Kirkwood, J.M., 2004. Re-evaluating the role of dacarbazine in metastatic melanoma: what have we learned in 30 years? European journal of cancer (Oxford, England : 1990), *40(12), pp.1825–36.*

Engelman, J.A. et al., 2007. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science (New York, N.Y.), *316(5827), pp.1039–43.*

Felsher, D.W. & Bishop, J.M., 1999. Reversible tumorigenesis by MYC in hematopoietic lineages. Molecular cell, *4(2), pp.199–207.*

Flaherty, K.T. et al., 2010. Inhibition of mutated, activated BRAF in metastatic melanoma. The New England journal of medicine, *363(9), pp.809–19.*

Forslund, A. et al., 2008. MDM2 gene amplification is correlated to tumor progression but not to the presence of SNP309 or TP53 mutational status in primary colorectal cancers. Molecular cancer research : MCR, *6(2), pp.205–11.*

Fridlyand, J. et al., 2004. Hidden Markov models approach to the analysis of array CGH data. Journal of Multivariate Analysis, *90(1), pp.132–153.*

Friend, S.H. et al., 1986. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. Nature, *323(6089), pp.643–6.*

Garnett, M.J. et al., 2012. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, *483(7391), pp.570–5.*

Garraway, L.A., 2013. Genomics-driven oncology: framework for an emerging paradigm. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, *31(15), pp.1806–14.*

Garraway, L.A. & Lander, E.S., 2013. Lessons from the cancer genome. Cell, *153(1), pp.17–37.*

Gozgit, J.M. et al., 2012. Ponatinib (AP24534), a Multitargeted Pan-FGFR Inhibitor with Activity in Multiple FGFR-Amplified or Mutated Cancer Models. Molecular Cancer Therapeutics, *11(3), pp.690–699.*

Haibe-Kains, B. et al., 2013. Inconsistency in large pharmacogenomic studies. Nature.

Hoadley, K.A. et al., 2014. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. Cell, *158(4), pp.929–44.*

Hollander, M.C., Blumenthal, G.M. & Dennis, P.A., 2011. PTEN loss in the continuum of common cancers, rare syndromes and mouse models. Nature reviews. Cancer, *11(4), pp.289–301.*

Hollebecque, A. et al., 2013. Molecular screening for cancer treatment optimization (MOSCATO 01): A prospective molecular triage trial—Interim results. In ASCO Annual Meeting. *p. Abstract 2512.*

Holst, F. et al., 2007. Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. Nature genetics, *39(5), pp.655–60.*

van Houte, B.P.P. et al., 2009. CGHnormaliter: an iterative strategy to enhance normalization of array CGH data with imbalanced aberrations. BMC genomics, *10(1), p.401.*

How, C. et al., 2015. Chromosomal instability as a prognostic marker in cervical cancer. BMC cancer, *15(1), p.361.*

Hudson, T.J. et al., 2010. International network of cancer genome projects. Nature, *464(7291), pp.993–8.*

Huettner, C.S. et al., 2000. Reversibility of acute B-cell leukaemia induced by BCR-ABL1. Nature genetics, *24(1), pp.57–60.*

Hupé, P. et al., 2004. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics (Oxford, England), *20(18), pp.3413–22.*

Hutchinson, R.A. et al., 2015. Epidermal growth factor receptor immunohistochemistry: new opportunities in metastatic colorectal cancer. Journal of translational medicine, *13, p.217.*

Huw, L.-Y. et al., 2013. Acquired PIK3CA amplification causes resistance to selective phosphoinositide 3-kinase inhibitors in breast cancer. Oncogenesis, *2*, *p.e83.*

Iancu-Rubin, C. et al., 2014. Activation of p53 by the MDM2 inhibitor RG7112 impairs thrombopoiesis. Experimental hematology, *42(2)*, *pp.137–45.e5.*

Ilic, N. et al., 2011. PI3K-targeted therapy can be evaded by gene amplification along the MYC-eukaryotic translation initiation factor 4E (eIF4E) axis. Proceedings of the National Academy of Sciences of the United States of America, *108(37)*, *pp.E699–708.*

Jain, M. et al., 2002. Sustained loss of a neoplastic phenotype by brief inactivation of MYC. Science (New York, N.Y.), *297(5578)*, *pp.102–4.*

Jain, V.K. & Turner, N.C., 2012. Challenges and opportunities in the targeting of fibroblast growth factor receptors in breast cancer. Breast cancer research : BCR, *14(3)*, *p.208.*

Jehan, Z. et al., 2009. Frequent PIK3CA gene amplification and its clinical significance in colorectal cancer. The Journal of Pathology, *219(3)*, *pp.337–346.*

Jin, M., Buck, E. & Mulvihill, M.J., 2013. Modulation of insulin-like growth factor-1 receptor and its signaling network for the treatment of cancer: current status and future perspectives. Oncology Reviews, *7(1)*, *p.3.*

Jurgensmeier, J.M., Eder, J.P. & Herbst, R.S., 2014. New Strategies in Personalized Medicine for Solid Tumors: Molecular Markers and Clinical Trial Designs. Clinical Cancer Research, *20(17)*, *pp.4425–4435.*

Karapetis, C.S. et al., 2008. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. The New England journal of medicine, *359(17)*, *pp.1757–65.*

Kim, E.S. et al., 2011. The BATTLE trial: personalizing therapy for lung cancer. Cancer discovery, *1(1)*, *pp.44–53.*

Kim, N., He, N. & Yoon, S., 2014. Cell line modeling for systems medicine in cancers (review). International journal of oncology, *44(2)*, *pp.371–6.*

Knutsen, T. et al., 2005. The interactive online SKY/M-FISH & CGH database and the Entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence. Genes, chromosomes & cancer, *44(1)*, *pp.52–64.*

Kolasa, I.K. et al., 2009. PIK3CA amplification associates with resistance to chemotherapy in ovarian cancer patients. Cancer biology & therapy, *8(1)*, *pp.21–6.*

Konopka, J.B., Watanabe, S.M. & Witte, O.N., 1984. An alteration of the human c-abl protein in K562 leukemia cells unmasks associated tyrosine kinase activity. Cell, *37(3)*, *pp.1035–42.*

Kwak, E.L. et al., 2010. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. The New England journal of medicine, *363(18)*, *pp.1693–703.*

Lai, W.R. et al., 2005. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics (Oxford, England), *21(19), pp.3763–70.*

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. Nature, *409(6822), pp.860–921.*

Larsson, O., Girnita, A. & Girnita, L., 2005. Role of insulin-like growth factor 1 receptor signalling in cancer. British journal of cancer, *92(12), pp.2097–101.*

Lashkari, D.A. et al., 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proceedings of the National Academy of Sciences of the United States of America, *94(24), pp.13057–62.*

Lee, A.J.X. et al., 2011. Chromosomal instability confers intrinsic multidrug resistance. Cancer research, *71(5), pp.1858–70.*

Lin, C.Y. et al., 2012. Transcriptional amplification in tumor cells with elevated c-Myc. Cell, *151(1), pp.56–67.*

Lingjaerde, O.C. et al., 2005. CGH-Explorer: a program for analysis of array-CGH data. Bioinformatics (Oxford, England), *21(6), pp.821–2.*

Liu, Y. et al., 2014. The tumor suppressor prostate apoptosis response-4 (Par-4) is regulated by mutant IDH1 and kills glioma stem cells. Acta neuropathologica.

Van Loo, P. et al., 2010. Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences of the United States of America, *107(39), pp.16910–5.*

Lovén, J. et al., 2012. Revisiting global gene expression analysis. Cell, *151(3), pp.476–82.*

Lynch, T.J. et al., 2004. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. The New England journal of medicine, *350(21), pp.2129–39.*

MacConaill, L.E., 2013. Existing and emerging technologies for tumor genomic profiling. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, *31(15), pp.1815–24.*

Macconaill, L.E. & Garraway, L.A., 2010. Clinical implications of the cancer genome. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, *28(35), pp.5219–28.*

Mallat, S., 2008. A Wavelet Tour of Signal Processing, *Academic Press.*

Mardis, E.R., 2011. A decade's perspective on DNA sequencing technology. Nature, *470(7333), pp.198–203.*

Marioni, J.C. et al., 2007. Breaking the waves: improved detection of copy number variation from

microarray-based comparative genomic hybridization. Genome biology, *8(10), p.R228.*

Marioni, J.C., Thorne, N.P. & Tavaré, S., 2006. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics (Oxford, England), *22(9), pp.1144–6.*

McGranahan, N. et al., 2012. Cancer chromosomal instability: therapeutic and diagnostic challenges. EMBO reports, *13(6), pp.528–38.*

Mermel, C.H. et al., 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology, *12(4), p.R41.*

Mikheev, A.M. et al., 2014. Periostin is a novel therapeutic target that predicts and regulates glioma malignancy. Neuro-oncology.

Miyaguchi, K. et al., 2011. Genome-wide integrative analysis revealed a correlation between lengths of copy number segments and corresponding gene expression profile. Bioinformation, *7(6), pp.280–4.*

Momand, J. et al., 1998. The MDM2 gene amplification database. Nucleic acids research, *26(15), pp.3453–9.*

Musgrove, E.A. et al., 2011. Cyclin D as a therapeutic target in cancer. Nature reviews. Cancer, *11(8), pp.558–72.*

Neill, N.J. et al., 2010. Comparative analysis of copy number detection by whole-genome BAC and oligonucleotide array CGH. Molecular cytogenetics, *3, p.11.*

O'Brien, S.G. et al., 2003. Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. The New England journal of medicine, *348(11), pp.994–1004.*

Olshen, A.B. et al., 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics (Oxford, England), *5(4), pp.557–72.*

Pao, W. et al., 2004. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proceedings of the National Academy of Sciences of the United States of America, *101(36), pp.13306–11.*

Patnaik, A. et al., 2015. Clinical pharmacology characterization of RG7112, an MDM2 antagonist, in patients with advanced solid tumors. Cancer chemotherapy and pharmacology.

Pectasides, E. & Bass, A.J., 2015. ERBB2 Emerges as a New Target for Colorectal Cancer. Cancer discovery, *5(8), pp.799–801.*

Picard, F. et al., 2005. A statistical approach for array CGH data analysis. BMC bioinformatics, *6(1), p.27.*

Piccart-Gebhart, M.J. et al., 2005. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. The New England journal of medicine, *353(16), pp.1659–72.*

Pierga, J. et al., 2012. A PROSPECTIVE RANDOMIZED TRIAL EVALUATING GENE EXPRESSION ARRAYS TO SELECT NEOADJUVANT CHEMOTHERAPY REGIMEN FOR OPERABLE BREAST CANCER: FIRST REPORT OF THE REMAGUS04 TRIAL. In ESMO Congress*. p. Abstract 2450.*

Popova, T. et al., 2009. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. Genome Biology, *10(11), p.R128.*

Ribeiro, T.C. et al., 2014. Amplification of the insulin-like growth factor 1 receptor gene is a rare event in adrenocortical adenocarcinomas: searching for potential mechanisms of overexpression. BioMed research international, *2014, p.936031.*

Rihani, A. et al., 2015. Inhibition of CDK4/6 as a novel therapeutic option for neuroblastoma. Cancer cell international, *15, p.76.*

Romond, E.H. et al., 2005. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. The New England journal of medicine, *353(16), pp.1673–84.*

Rowley, J.D., 1973. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature, *243(5405), pp.290–3.*

Roychowdhury, S. et al., 2011. Personalized oncology through integrative high-throughput sequencing: a pilot study. Science translational medicine, *3(111), p.111ra121.*

Russnes, H. et al., 2010. Genomic Architecture Characterizes Tumor Progression Paths and Fate in Breast Cancer Patients. Science translational medicine, *2(38), p.38ra47.*

Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, *74(12), pp.5463–7.*

Sattler, M. & Griffin, J.D., 2001. Mechanisms of transformation by the BCR/ABL oncogene. International journal of hematology, *73(3), pp.278–91.*

Sawyers, C., 2004. Targeted cancer therapy. Nature, *432(7015), pp.294–7.*

Schena, M. et al., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (New York, N.Y.), *270(5235), pp.467–70.*

Schiller, J.H. et al., 2002. Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. The New England journal of medicine, *346(2), pp.92–8.*

Sen, A. & Srivastava, M.S., 1975. On Tests for Detecting Change in Mean. The Annals of Statistics, *3(1), pp.98–108.*

Sesboüé, R. et al., 2012. EGFR alterations and response to anti-EGFR therapy: is it a matter of gene amplification or gene copy number gain? British journal of cancer, *106(2), pp.426–7; author reply 428.*

Shayesteh, L. et al., 1999. PIK3CA is implicated as an oncogene in ovarian cancer. Nature genetics, *21(1), pp.99–102.*

Shibata, D., 2012. Cancer. Heterogeneity and tumor history. Science (New York, N.Y.), *336(6079), pp.304–5.*

Shoemaker, R.H., 2006. The NCI60 human tumour cell line anticancer drug screen. Nature reviews. Cancer, *6(10), pp.813–23.*

de Silva, C.M. V & Reid, R., 2003. Gastrointestinal stromal tumors (GIST): C-kit mutations, CD117 expression, differential diagnosis and targeted cancer therapy with Imatinib. Pathology oncology research : POR, *9(1), pp.13–9.*

Slamon, D.J. et al., 2001. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. The New England journal of medicine, *344(11), pp.783–92.*

Smolen, G.A. et al., 2006. Amplification of MET may identify a subset of cancers with extreme sensitivity to the selective tyrosine kinase inhibitor PHA-665752. Proceedings of the National Academy of Sciences of the United States of America, *103(7), pp.2316–21.*

Sordella, R. et al., 2004. Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways. Science (New York, N.Y.), *305(5687), pp.1163–7.*

Staaf, J. et al., 2007. Normalization of array-CGH data: influence of copy number imbalances. BMC genomics, *8(1), p.382.*

Stern, H.M. et al., 2015. PTEN Loss Is Associated with Worse Outcome in HER2-Amplified Breast Cancer Patients but Is Not Associated with Trastuzumab Resistance. Clinical cancer research : an official journal of the American Association for Cancer Research, *21(9), pp.2065–74.*

Tabin, C.J. et al., 1982. Mechanism of activation of a human oncogene. Nature, *300(5888), pp.143–9.*

Tourneau, C. Le et al., 2015. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. The Lancet. Oncology.

Tran, B. et al., 2013. Feasibility of real time next generation sequencing of cancer genes linked to drug response: results from a clinical trial. International journal of cancer. Journal international du cancer, *132(7), pp.1547–55.*

Tsao, M.-S. et al., 2005. Erlotinib in lung cancer - molecular and clinical predictors of outcome. The New England journal of medicine, *353(2), pp.133–44.*

Tsimberidou, A.-M. et al., 2014. Personalized Medicine for Patients with Advanced Cancer in the Phase I Program at MD Anderson: Validation and Landmark Analyses. Clinical cancer research : an official journal of the American Association for Cancer Research, *20(18), pp.4827–36.*

Tsimberidou, A.-M. et al., 2012. Personalized medicine in a phase I clinical trials program: the MD Anderson Cancer Center initiative. Clinical cancer research : an official journal of the American Association for Cancer Research, *18(22), pp.6373–83.*

Valli, R. et al., 2011. Comparative genomic hybridization on microarray (a-CGH) in constitutional and acquired mosaicism may detect as low as 8% abnormal cells. Molecular cytogenetics, *4(1), p.13.*

Venkatraman, E.S. & Olshen, A., DNAcopy: A Package for Analyzing DNA Copy Data. Available at: http://www.bioconductor.org/packages/release/bioc/vignettes/DNAcopy/inst/doc/DNAcopy.pdf [Accessed March 19, 2015].

Venkatraman, E.S. & Olshen, A.B., 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics (Oxford, England), *23(6), pp.657–63.*

Vigneri, P. & Wang, J.Y., 2001. Induction of apoptosis in chronic myelogenous leukemia cells through nuclear entrapment of BCR-ABL tyrosine kinase. Nature medicine, *7(2), pp.228–34.*

Vogel, C.L. et al., 2002. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, *20(3), pp.719–26.*

Vollan, H.K.M. et al., 2015. A tumor DNA complex aberration index is an independent predictor of survival in breast and ovarian cancer. Molecular oncology, *9(1), pp.115–27.*

Wang, L. et al., 2014. Integrating multi-omics for uncovering the architecture of cross-talking pathways in breast cancer. PloS one, *9(8), p.e104282.*

Watermann, I. et al., 2015. Improved diagnostics targeting c-MET in non-small cell lung cancer: expression, amplification and activation? Diagnostic pathology, *10(1), p.130.*

Weinstein, I.B., 2002. Cancer. Addiction to oncogenes--the Achilles heal of cancer. Science (New York, N.Y.), *297(5578), pp.63–4.*

Weinstein, J.N. & Lorenzi, P.L., 2013. Cancer: Discrepancies in drug sensitivity. Nature.

Weiss, J. et al., 2010. Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. Science translational medicine, *2(62), p.62ra93.*

Weiss, M.M. et al., 1999. Comparative genomic hybridisation. Molecular pathology : MP, *52(5), pp.243–51.*

Van De Wiel, M. a et al., 2007. CGHcall: calling aberrations for array CGH tumor profiles. Bioinformatics, *23(7), pp.892–894.*

Wilky, B.A. et al., 2015. A phase I trial of vertical inhibition of IGF signalling using cixutumumab, an anti-IGF-1R antibody, and selumetinib, an MEK 1/2 inhibitor, in advanced solid tumours. British journal of cancer, *112(1), pp.24–31.*

Willenbrock, H. & Fridlyand, J., 2005. A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics (Oxford, England), *21(22), pp.4084–91.*

Wynes, M.W. et al., 2014. FGFR1 mRNA and protein expression, not gene copy number, predict FGFR TKI sensitivity across all lung cancer histologies. Clinical cancer research : an official journal of the American Association for Cancer Research, *20(12), pp.3299–309.*

Yang, S. et al., 2013. PAIR: paired allelic log-intensity-ratio-based normalization method for SNP-CGH arrays. Bioinformatics (Oxford, England), *29(3), pp.299–307.*

Yu, K.-D. & Shao, Z.-M., 2011. ESR1 gene amplification: another mechanism regulating the cellular levels of ERα. Nature reviews. Cancer, *11(11), p.823; author reply 823.*

Zhang, C. et al., 2014. High NR2F2 transcript level is associated with increased survival and its expression inhibits TGF-β-dependent epithelial-mesenchymal transition in breast cancer. Breast cancer research and treatment, *147(2), pp.265–81.*

Zhao, Y. et al., 2015. Small-molecule inhibitors of the MDM2-p53 protein-protein interaction (MDM2 Inhibitors) in clinical trials for cancer treatment. Journal of medicinal chemistry, *58(3), pp.1038–52.*

Zheng, S. et al., 2013. A survey of intragenic breakpoints in glioblastoma identifies a distinct subset associated with poor survival. Genes & development, *27(13), pp.1462–72.*

9 Table des Figures

10 Annexes

10.1 Annexe 1 : Simulation segmentation CBS, code R

```
##############################
# circular binary segmentation
# Simulation
##############################

require(gplots)
op <- par(no.readonly=TRUE)

# CBS statistics function
.test <- function(i, j, x){
    if(i >= j)
        return(0)
    si <- sum(x[1:i])
    sj <- sum(x[1:j])
    sn <- sum(x)
    den1 <- (sj-si)/(j-i)
    den2 <- (sn - sj + si)/(length(x) - j + i)
    num <- sqrt(1/(j-i) + 1/(length(x) - j + i))
    stat <- (den2 - den1)/num
    return(abs(stat))
}
#######################

# Simulate LRRs with 2 breakpoints at 100 and 125 (3 segments)
x <- c(rnorm(100, 0, .1), rnorm(25, 2, .1), rnorm(50, 0, .1))

# Plot simulated LRR
par(mar=c(4.2, 5, 4, 2), cex.axis=1.5, cex.lab=1.75, cex.main=2, las=1, bty="n")
plot(x,
     pch=19, cex=1.75,
     col=rgb(.2, .8, .7, .75),
     xlab="Location", ylab=expression(Log[2](Ratio)))
points(x, cex=1.8)
par(op)

# Compute statistics
z <- lapply(1:length(x), function(ii){
    tmp <- lapply(1:length(x), function(jj) .test(ii, jj, x))
    do.call(c, tmp)
})
z <- do.call(rbind, z)
mr <- which.max(apply(z, 1, max, na.rm=TRUE))
mc <- which.max(apply(z, 2, max, na.rm=TRUE))


# zprim <- rbind(z, rep(0, ncol(z)))
# zprim <- cbind(zprim, rep(0, nrow(zprim)))

n <- length(x)

# Main layout image
M <- matrix(4, 3, 3, byrow=TRUE)
M[1,3] <- 1
```

```
M[1, -3] <- 2
M[-1, 3] <- 3
lf <- layout(M) ; plot.new()

# top axis
par(mar=c(0, 0, 4, 0), cex.axis=1.75, bty="n")
plot(z[,mc], axes=FALSE, type="l", lwd=3, xlab="", ylab="")
axis(3, at=seq(0, length(x), by=25), labels=seq(0, length(x), by=25))
abline(v=mr)
legend("topleft", legend=expression(z["i,125"]), bty="n", cex=2)

# right axis
par(mar=c(4, 0, 0, 4), cex.axis=1.75, las=1, bty="n")
plot(x=-z[mr,], y=1:nrow(z), axes=FALSE, type="l", lwd=3, xlab="", ylab="")
axis(4, at=seq(0, length(x), by=25), labels=seq(0, length(x), by=25))
abline(h=mc)
legend("bottom", legend=expression(z["100,j"]), bty="n", cex=2)

# center image
par(mar=c(3, 0, 0,.35), bty="o")
image(
    seq(1, nrow(z)),
    seq(1, ncol(z)),
    z,
    col = colorpanel(1000, "white", "purple4"),
    axes=FALSE,
    xlab="", ylab=""
    )
points(mr, mc, pch=19, cex=.2)
segments(x0=mr, x1=1000, y0=mc, col="purple4")
segments(x0=mr, y0=mc, y1=n, col="purple4")
legend("bottom", legend=expression(z["i,j"]), bty="n", cex=3)

par(op)


# Final segmentation
n <- length(x)
s1 <- mean(x[1:mr]) ; s2 <- mean(x[(mr+1):mc]) ; s3 <- mean(x[(mc+1):n])

dev.off()
par(mar=c(4.2, 5, 4, 2), cex.axis=1.5, cex.lab=1.75, cex.main=2, las=1, bty="n")
pcol <- rgb(.2, .8, .7, .75)
plot(x,
     pch=19, cex=1.75,
     col=pcol,
     xlab="Location", ylab=expression(Log[2](Ratio)))
points(x, cex=1.8)
segments(x0=c(1, mr+1, mc+1), x1=c(mr, mc, n), y0=c(s1, s2, s3), col="red",
lwd=8)
legend("topleft", legend=c("probes signal", "segments"), cex=1.5,
       pch=c(19, -1), lwd=c(-1, 5), col=c(pcol, "red"), bty="n")
par(op)

# END SIMULATION
```

10.2 Annexe 2 : Méthodes supplémentaires de l'article : Impact of centralization on aCGH-based genomic profiles for precision medicine in oncology

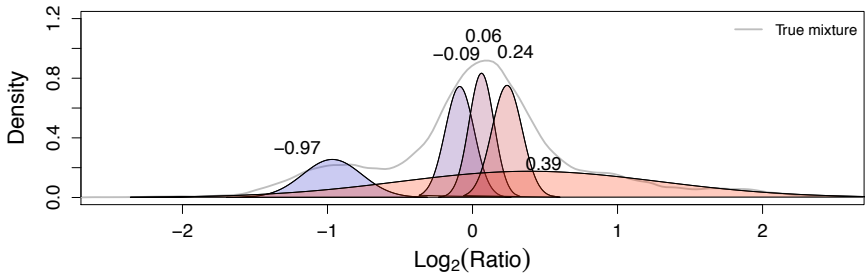
Impact of centralization on aCGH-based genomic profiles for precision medicine in oncology

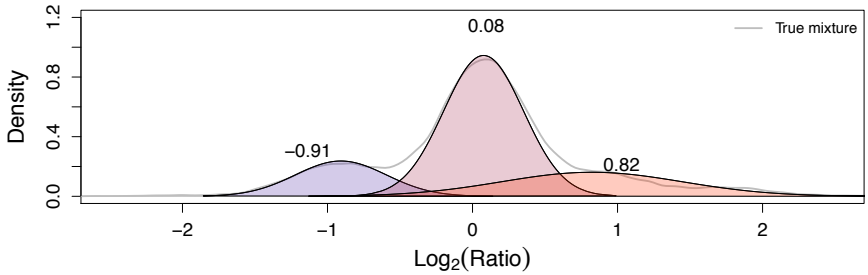
F. Commo[1,2,#], C. Ferté[1,2,3,#], JC. Soria[2,3], S.H. Friend[1], F. André[2,3], J. Guinney[1]

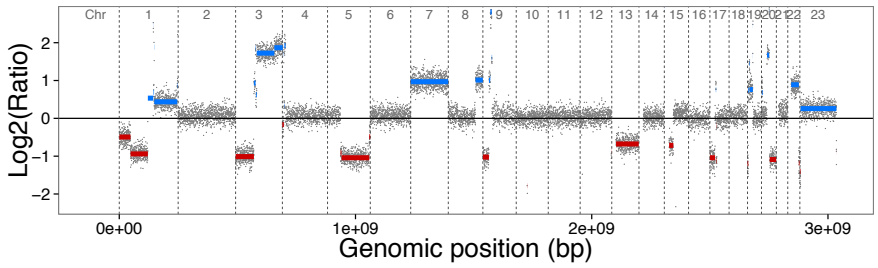
1) Sage Bionetworks, Seattle, WA, USA
2) INSERM U981, Gustave Roussy, University Paris XI, Villejuif, France
3) Department of Medical Oncology, Gustave Roussy, Villejuif, France
# These authors contributed equally to this work.

Supplementary methods and results:
The EM algorithm - EM stands for Expectation-Maximization - is well known and widely used for modeling mixture of normally distributed populations, and extracting their respective parameters. Unfortunately, EM does not perform well on very large and noisy vectors such as CGH Log_2(Ratios) computed from hundreds of thousands of probes. On such values, this method is time consuming, and not well adapted for detecting centralization peaks, since EM tends to over estimate (or underestimate) the number of sub-populations (figure 1).

Figure 1: The EM algorithm was applied on real data (Safir01 sample). On its original form (top), the method considers the central population as itself a mixture of 3 populations (means: -0.09, 0.062 and 0.23, respectively), which is probably true, but maybe not relevant in that context. Instead, the adapted resampling method (center) identifies one unique peak, and simplifies the decision for centering the entire vector.







To significantly reduce the computation time, as well as to improve the efficacy of the EM algorithm in that particular context, we adapted the algorithm as follow:

A Log_2(Ratios) is modeled as a mixture of Gaussian distributions, using 1e3 values randomly picked, instead of using the entire vector of values. The procedure is repeated 100 times, independently. Then, means and variances of each population in the mixture are averaged, considering only the models for which the number of groups corresponds to the median of groups detected over all the models.

2

The performances of the procedure were estimated on simulated Gaussian mixtures of various total lengths, *N from 1e3 to 1e6*, with 3 arbitrary components, C_i *i=1 to 3*, each with different means (*mu*), standard deviations (*sd*) and proportions (*p*):

$$C_1 \sim N(-0.58,\ 0.25);\ p_1 = 0.15$$
$$C_2 \sim N(0,\ 0.5);\ p_2 = 0.7$$
$$C_3 \sim N(0.58,\ 0.75);\ p_3 = 0.15$$

The choice of the different parameters was arbitrary, but consistent with mixtures observed in real situations.

This simulations showed that the EM computation time increases exponentially with the size of vector, while the resampling method increases almost linearly, and becomes advantageous on large data, namely N>2e5. This value roughly corresponds to the length of a vector of Log_2(Ratios) generated from Agilent 4x180K microarrays (figure2). Moreover, when N is large, N>2e5, the EM algorithm tends to split some of the sub-components itself into separate populations, and is unable to correctly estimate the mixture and its parameters. Under the same conditions, repeated random resamplings lead to better estimation of the mixture, close to the expected results (figure 3 & 4).

Figure 2: The EM computation time increases exponentially with the length of the vector (black line). The resampling approach is more stable, and becomes advantageous on very large data (N>2e5) (red line).

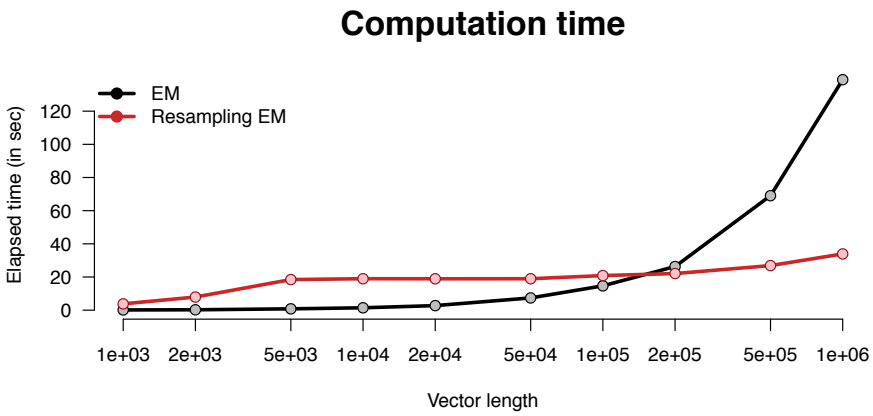


Figure 3: a mixture of 3 normally distributed vectors was simulated. When applying the original EM, the ability of detecting the right number of groups depends on the size of the entire vector (black line), while the resampling approach always detect the correct number of sub-populations, independently of the size of the data (red line).

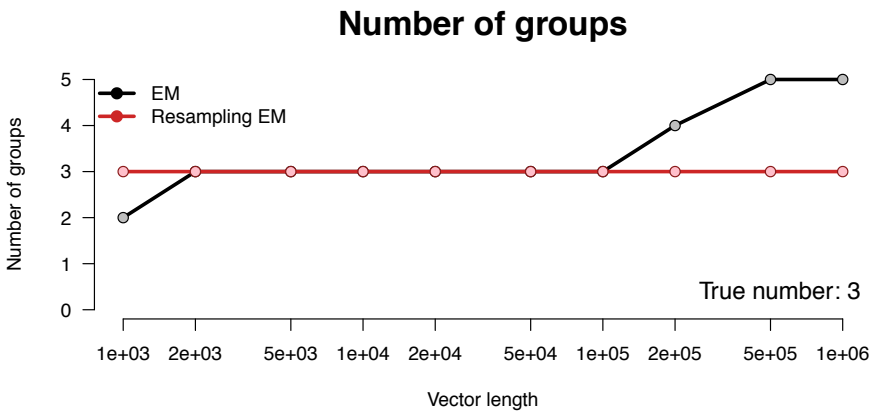
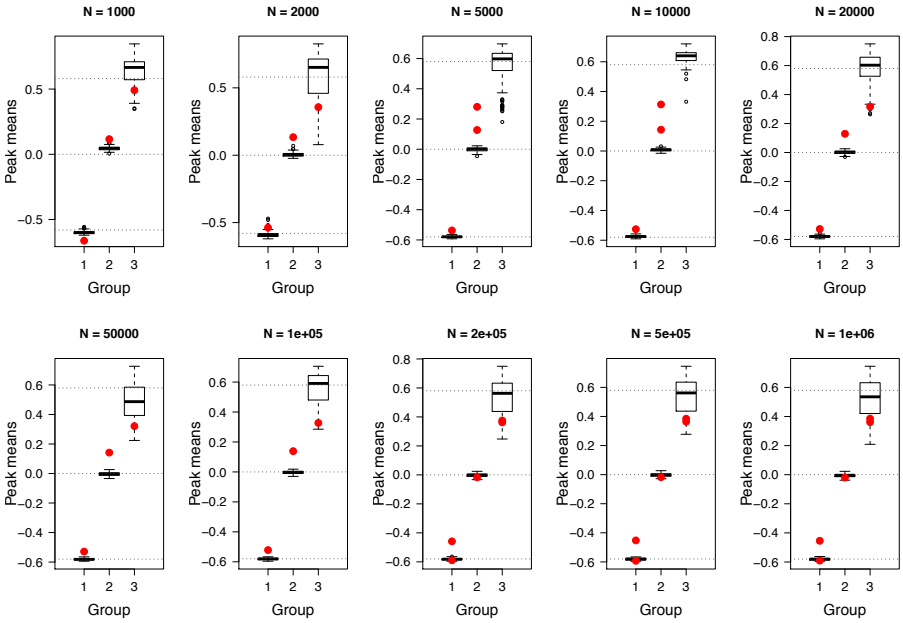


Figure 4: On the same 3-groups mixture simulation, the original EM fails to properly identify the means of each group (grey dash lines at -0.58, 0 and 0.58), while the resampling approach returns values close to the expected ones. Red dots are population means estimated with the original EM, boxes are the distributions of means across the replicated resampling method, bold lines are the medians of the

means for each sub-population. Note that, the EM estimated means were assigned to the group with the closest expected value.



Validation on NCI cell lines

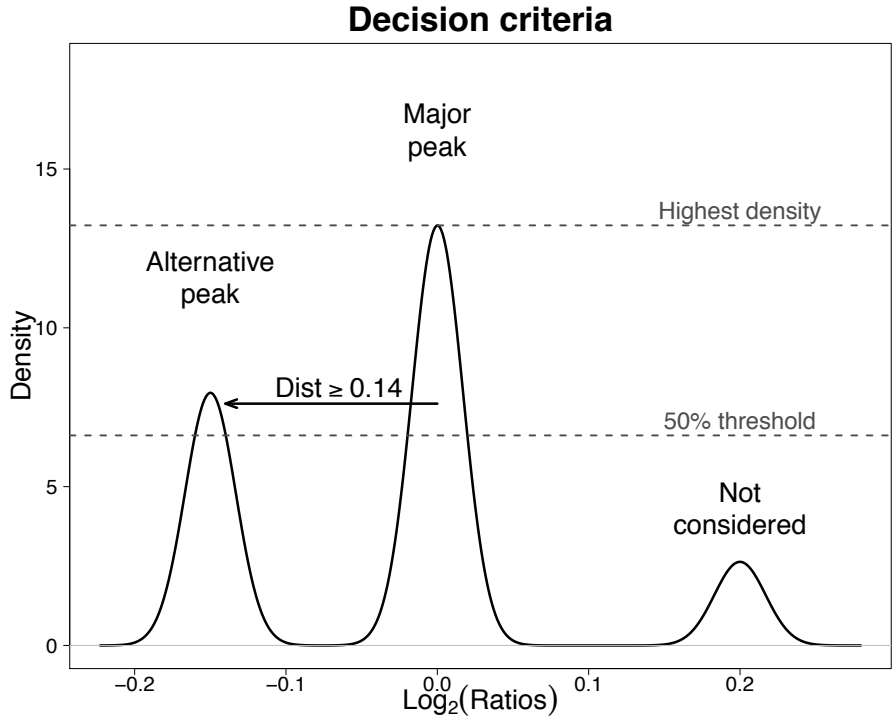
This approach was validated on the NCI data by estimating, for each sample, the reproducibility of the maximum peak detection using this procedure, on 1000 independent tests. The 2-standard deviation interval of the maximum peak values distribution was used to refine our definition of a valid alternative peak.

From the 60 individual cell lines represented by 72 CGH data, 3 were removed because of too few karyotypic information to generate a genomic-like profile. The remaining 57 cell lines were represented by 69 aCGH data, including replicated experiments.

Over all these remaining 69 experiments, the maximum peak values showed a 2-standard deviation (2SD) from 1.19e-3 to 1.38e-1. We then considered the largest observed 2SD interval of 0.14 as an additional constraint. Finally a relevant alternative centralization value was defined as the mean of a peak with a density

height of at least 50% the density height of the major peak, and at least at 0.14, in Log_2, from this major peak (figure 5).

Figure 5: Given the validation step on the NCI cell lines, a valid alternative peak was defined as a peak with a maximum density higher than 50% of the height of the main peak, and at least at 0.14, in Log_2 scale.



For 6 of the 57 different cell lines, 2 or more replicated aCGH experiments were available. In all the cases, but one, the peak estimations were similar across the replicated experiments. For one of the 4 MCF-7 replicates, an alternative peak was detected, but at a distance lower than 0.14. This profile had also a highest derivative log_2(Ratio) spread, compared to the 3 others (.256 vs. .252, 0.243 and 0.230, respectively).

Supplementary figure S1: General array-based genomic profiling workflow.
After hybridization, fluorescent signals are digitized and then mapped with the probes locations. LogR are computed against the reference signals (dual-color hybridization) or against an external reference (single-color hybridization). The centralization step adjust the LogR on their median, or on the maximum density value. Then, the segmentation step identifies the breakpoints. Calls rely on the segments magnitude with respect to the base line, considered as the neutral-copy level.

Supplementary figure S2: Frequency of ploidies in the NCI60 cell lines panel, and proportion of alternative peak identification.
The NIC60 cell line panel predominantly includes aneuploid cell lines; more than 50% of the cell lines are, at least, 3n (A). Interestingly, centralization alternatives occur in rare cases of 2n cell lines, while alternative options are extremely frequent in 3 and 4n cell lines (B). The unique case of a 5n-/+ cell line is specifically described in supplementary figure 2.

Supplementary figure S3: Consistency of genomic profiles according to the centralization methods: the A549 cell line
The A549 genomic profiles has been centered on LogR median (top panel), the maximum density peak (central panel), or the alternative LogR density peak (bottom panel), before being segmented using the CBS algorithm, with the same segmentation parameters. The corresponding centralization values are indicated in bold on each density plot (Centralization). When comparing with the corresponding karyotype, The 2 first methods adjust the entire profile on the main cell line ploidy, namely 3n, and lead to consider all of the 3-copy chromosomes (1p, 2, 3, 5, 7 to 10, 12, 14 and 16) as in normal count, while most of the 2-copy chromosomes are considered as lost (1p, 4, 6, 13, 19, 21 and 22). Adjusting on the alternative peak dramatically reduce such discrepancies, although errors persist on chr11. Notice that 2 supplementary copies of chr19 are located on chr15, according to the karyotype.

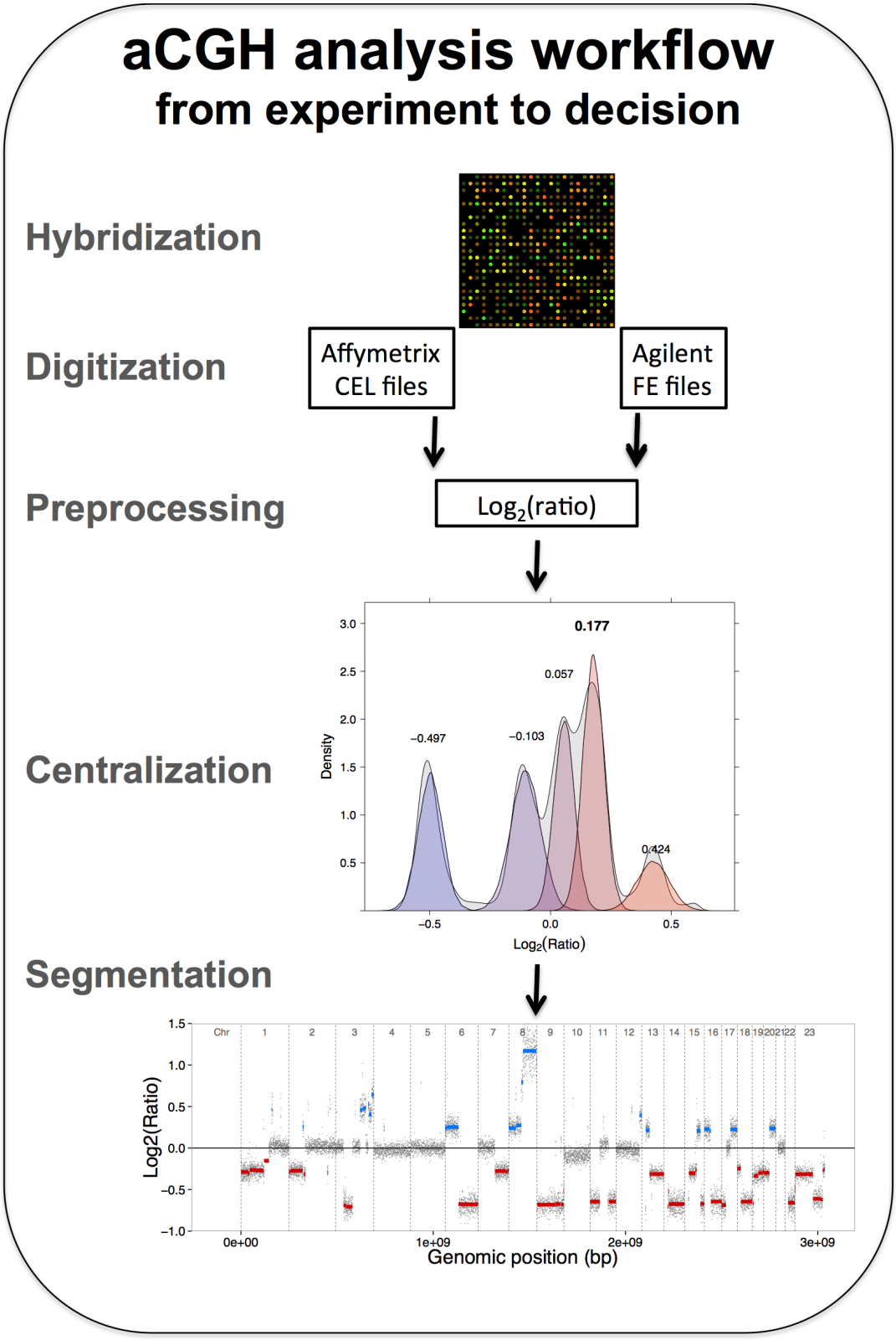
Supplementary figure S4: The 5n-/+ SF-295 cell line.
No alternative centralization is suggested when analyzing the LogR density as a mixture of Gaussian populations (left), and only the major density peak seemed to correspond to a sensible value for adjusting the genomic profile (center). However,

this choice leaded to an erroneous profile, given the karyotype (right); most of the 5-copy chromosomes were considered as in normal counts on the genomic profile, while 3-copy chromosomes (chr10 and 14) were considered as lost.

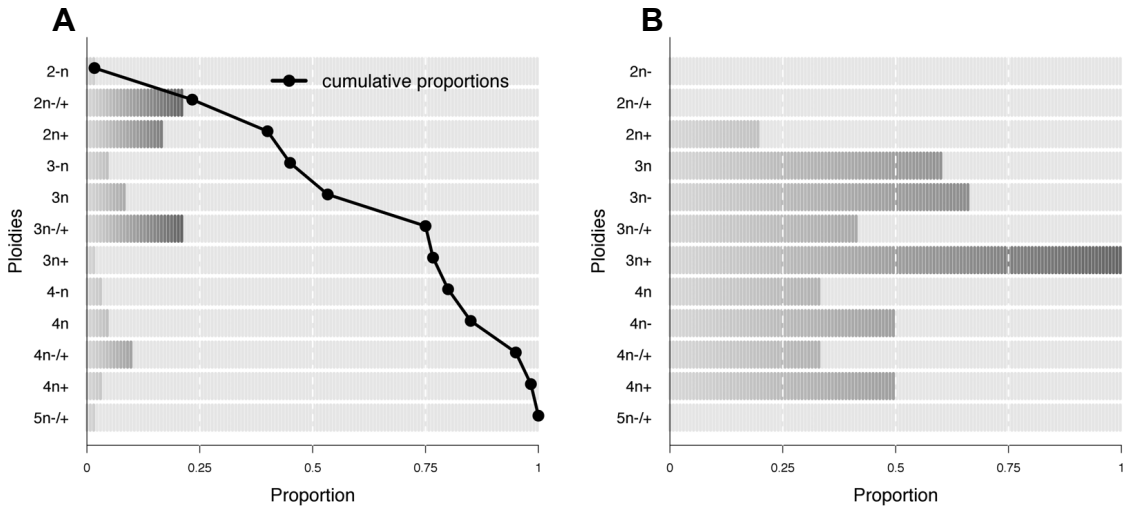
Supplementary figure S5: Correlation between copy number variation and sensitivity to related inhibitors.
The Spearman correlations of FGFR1 and MET, with their respective inhibitors in the CCLE data (TKI258 and PHA665752, respectively), are not significantly improved when changing the profile centralization strategy: *p* were at least greater than 0.27 in all comparisons. To note, the relatively weak correlations between these genes copy number variation and responses to their corresponding inhibitor.

Supplementary table S6: Amplification calls in SAFIR01 and MOSCATO-01, according to the centralization method.
Study, patient Id, and platforms are indicated in columns 1 to 3, respectively. Considering the same genes as in André *et al.*, and using the same decision rules, amplifications are indicated, for each centralization method, in the corresponding column: “none” means no amplification detected, and in bold, genes detected using one method only.

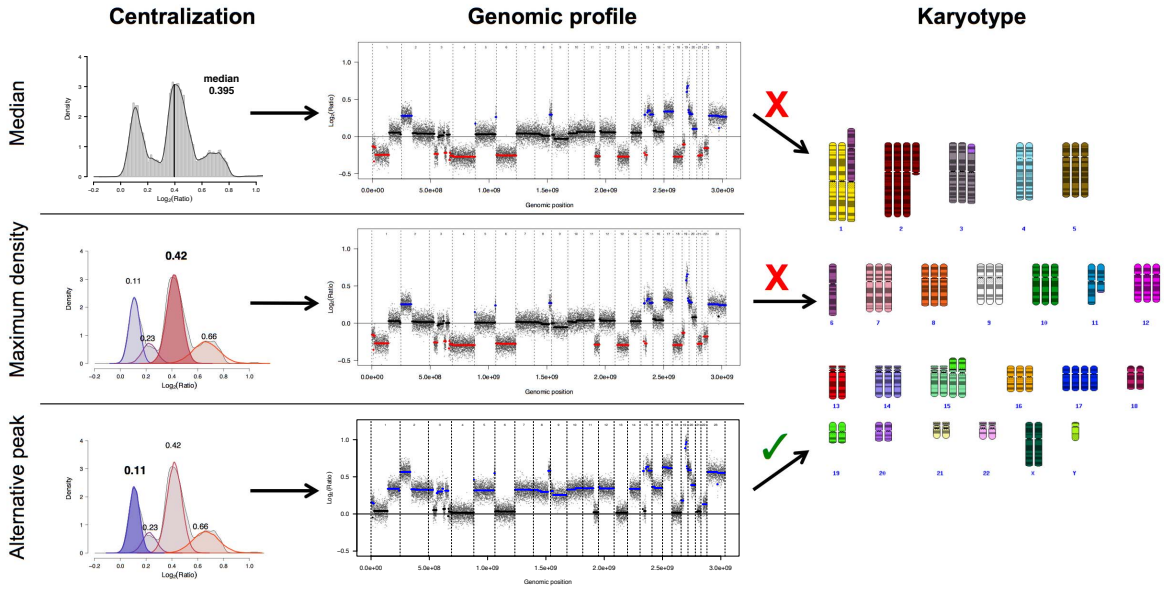
Supplementary figure S1



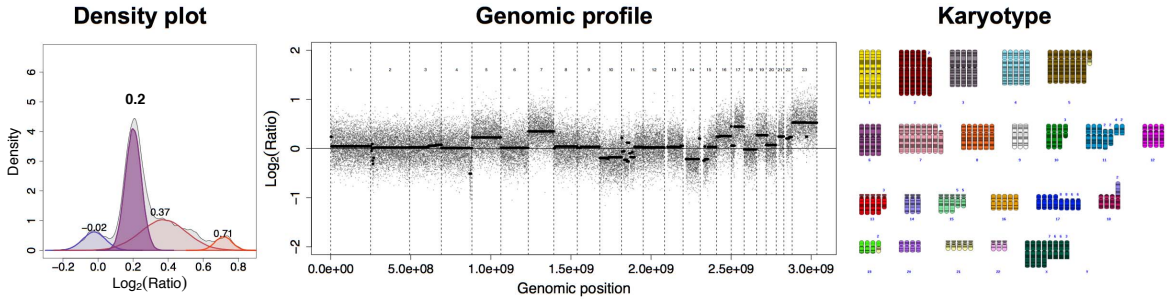
Supplementary figure S2



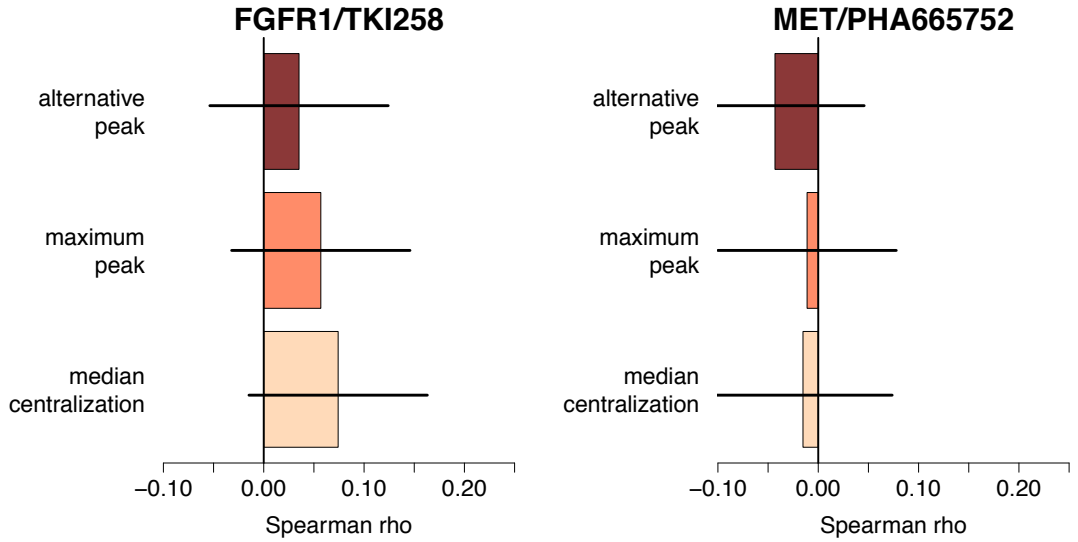
Supplementary figure S3



Supplementary figure S4



Supplementary figure S5



FGFR1 Vs. TKI258

	Median	Maximum	Alternative
Rho	0.074	0.057	0.035

p-values	**Maximum**	**Alternative**
Median	0.394	0.271
Maximum		0.367

MET Vs. PHA665752

	Median	Maximum	Alternative
Rho	-0.015	-0.011	-0.043

p-values	**Maximum**	**Alternative**
Median	0.475	0.332
Maximum		0.309

Supplementary table S6

			Amplifications	
Study	Patients index	Platform	maximum peak	Alternate peak
MOSCATO	M008	Agilent_4x180K	none	**FGF4**
MOSCATO	M028	Agilent_4x180K	none	**IGF1**
MOSCATO	M046	Agilent_4x180K	none	**EGFR**
MOSCATO	M070	Agilent_4x180K	none	**FGF9**
MOSCATO	M091	Agilent_4x180K	none	**ALK**
MOSCATO	M135	Agilent_4x180K	none	**FGFR1, PIK3CB**
MOSCATO	M168	Agilent_4x180K	none	**FRS2, MDM2**
MOSCATO	M220	Agilent_4x180K	none	**EGFR, NOTCH4**
MOSCATO	M258	Agilent_4x180K	none	**IGF1R**
MOSCATO	M263	Agilent_4x180K	none	**FGF9**
MOSCATO	M269	Agilent_4x180K	none	**BRCA2, FGF9**
MOSCATO	M283	Agilent_4x180K	none	**BRCA2, EGFR, IGF1R**
MOSCATO	M311	Agilent_4x180K	none	**FGFR1**
MOSCATO	M316	Agilent_4x180K	none	**EGFR**
MOSCATO	M358	Agilent_4x180K	none	**FRS2, IGF1, MDM2**
MOSCATO	M365	Agilent_4x180K	none	**CCND1, FGF4**
MOSCATO	M376	Agilent_4x180K	none	**EGFR, MET**
MOSCATO	M379	Agilent_4x180K	none	**RPTOR**
MOSCATO	M395	Agilent_4x180K	none	**EGFR, FGFR1**
MOSCATO	M438	Agilent_4x180K	none	**FGF9, RPTOR**
MOSCATO	M443	Agilent_4x180K	none	**FRS2, MDM2**
MOSCATO	M067	Agilent_4x180K	CCND1, FGF4, FGFR1, PAK1	CCND1, FGF4, FGFR1, PAK1, EGFR
MOSCATO	M118	Agilent_4x180K	CCND1, FGF4, PAK1	CCND1, FGF4, PAK1, ESR1, FGFR1, RPTOR
MOSCATO	M140	Agilent_4x180K	MET	MET, EGFR, ESR1, FRS2, MDM2
MOSCATO	M201	Agilent_4x180K	RPTOR	RPTOR, BRCA1, ERBB2, PGR, TOP2A
MOSCATO	M360	Agilent_4x180K	CCND1, FGF4	CCND1, FGF4, ALK, ESR1, IGF1R, NOTCH4, PAK1, VEGFA
MOSCATO	M412	Agilent_4x180K	FRS2, MDM2	FRS2, MDM2, FGFR1
SAFIR01	S001	Affymetrix_snp6	none	**VEGFA**
SAFIR01	S012	Agilent_4x180K	none	**PAK1**
SAFIR01	S018	Agilent_4x180K	none	**BRCA2, CCND1, FGF4, FGFR1, NOTCH4, PIK3CB, RPTOR, VEGFA**
SAFIR01	S078	Affymetrix_snp6	none	**RPTOR**
SAFIR01	S079	Agilent_4x180K	none	**EGFR**
SAFIR01	S082	Agilent_4x180K	none	**EGFR**
SAFIR01	S101	Affymetrix_snp6	none	**BRCA1, CCND1, ERBB2, FGF4, FRS2, IGF1, IGF1R, MDM2, RPTOR, TOP2A**
SAFIR01	S137	Agilent_4x180K	none	**PAK1**
SAFIR01	S164	Agilent_4x180K	none	**PAK1**
SAFIR01	S184	Affymetrix_snp6	none	**BRCA1, CCND1, ERBB2, FGF4, FRS2, MDM2, NOTCH4, PIK3CB, RPTOR, TOP2A, VEGFA**
SAFIR01	S195	Affymetrix_snp6	none	**CCND1, FGF4, FRS2, MDM2, RPTOR**
SAFIR01	S225	Agilent_4x180K	none	**ESR1, PIK3CB**
SAFIR01	S234	Agilent_4x180K	none	**CCND1, FGF4, FGFR1**
SAFIR01	S237	Agilent_4x180K	none	**CCND1, FGF4, FGFR1**
SAFIR01	S250	Agilent_4x180K	none	**FGFR1**
SAFIR01	S257	Affymetrix_snp6	none	**RPTOR**
SAFIR01	S268	Affymetrix_snp6	none	**ERBB2, RPTOR**
SAFIR01	S319	Affymetrix_snp6	none	**EGFR, NOTCH4**
SAFIR01	S367	Agilent_4x180K	none	**CCND1, FGF4**
SAFIR01	S396	Affymetrix_snp6	none	**FGFR2, RPTOR**
SAFIR01	S002	Affymetrix_snp6	ERBB2, PAK1, TOP2A	ERBB2, PAK1, TOP2A, CCND1, FGF4, PIK3CB
SAFIR01	S034	Agilent_4x180K	FGFR1	FGFR1, CCND1, FGF4, FGF9, MET, RPTOR
SAFIR01	S054	Agilent_4x180K	ERBB2, FGFR1	ERBB2, FGFR1, PIK3CB, RPTOR
SAFIR01	S063	Affymetrix_snp6	NOTCH4	NOTCH4, PAK1, TOP2A, VEGFA
SAFIR01	S066	Agilent_4x180K	CCND1, FGF4	CCND1, FGF4, PAK1
SAFIR01	S071	Agilent_4x180K	ESR1, FGFR1	ESR1, FGFR1, RPTOR
SAFIR01	S125	Affymetrix_snp6	FGFR2	FGFR2, ALK, CCND1, FGF4
SAFIR01	S144	Agilent_4x180K	ERBB2, PAK1, TOP2A	ERBB2, PAK1, TOP2A, FGF4, CCND1
SAFIR01	S165	Agilent_4x180K	FGFR1	FGFR1, RPTOR
SAFIR01	S173	Agilent_4x180K	ESR1, FGF4	ESR1, FGF4, CCND1, FGFR1
SAFIR01	S176	Affymetrix_snp6	CCND1, ERBB2, FGF4, PAK1	CCND1, ERBB2, FGF4, PAK1, FRS2, IGF1, MDM2, NOTCH4, VEGFA
SAFIR01	S179	Agilent_4x180K	CCND1, FGF4, IGF1R	CCND1, FGF4, IGF1R, PAK1
SAFIR01	S207	Affymetrix_snp6	ERBB2	ERBB2, FGFR1, PAK1
SAFIR01	S241	Affymetrix_snp6	CCND1, ERBB2, FGF4	CCND1, ERBB2, FGF4, FGFR1
SAFIR01	S245	Affymetrix_snp6	FGFR1	FGFR1, FGFR2, EGFR
SAFIR01	S310	Affymetrix_snp6	CCND1, FGF4	CCND1, FGF4, FGFR1
SAFIR01	S350	Affymetrix_snp6	CCND1, FGF4, FGFR1	CCND1, FGF4, FGFR1, FRS2, MDM2
SAFIR01	S364	Agilent_4x180K	CCND1, FGF4	CCND1, FGF4, FGFR1
SAFIR01	S393	Agilent_4x180K	FRS2, MDM2	FRS2, MDM2, CCND1, FGF4
SAFIR01	S395	Agilent_4x180K	CCND1, FGF4, PAK1	CCND1, FGF4, PAK1, NOTCH4
SAFIR01	S401	Affymetrix_snp6	NOTCH4	NOTCH4, PIK3CB
SAFIR01	S418	Affymetrix_snp6	ERBB2, PGR	ERBB2, PGR, BRCA2, CCND1, FGF4

10.3 Annexe 3 : Fonctionnalités du package rCGH

Comprehensive Pipeline for Analyzing and Visualizing Agilent and Affymetrix Array-Based CGH Data

Frederic Commo *
Inserm U981, Bioinformatics Group, Gustave Roussy, France

August 26, 2015

1 Introduction

Genomic profiling using array-based comparative genomic hybridization (aCGH) is widely used within precision medicine programs, in combination with DNA sequencing, to match specific molecular alterations (amplifications or deletions) with therapeutic orientations.
We present *rCGH*, a comprehensive array-based CGH analysis workflow, integrating functionalities specifically designed for precision medicine. *rCGH* ensures a full traceability by saving all the process parameters , and facilitates genomic profiles interpretation and decision-making through interactive visualizations.
rCGH supports Agilent (from 44K to 400K arrays), as well as Affymetrix, SNP6 and cytoScanHD arrays.

2 rCGH object structure

In order to store (or update) data, sample information, and the workflow parameters all along a genomic profile analysis process, `rCGH` objects are structured as follow:

- info: the sample information.
- cnSet: the full by-probe dataset.
- param: the workflow parameters, for traceability.
- segTable: the segmentation data.

All these slots are accessible through specific functions, as described in the next sections.
Notice that *rCGH* is a superclass designed for calling common methods. Depending on the type of array and the *read* functions used, the resulting objects will be assigned to classes *rCGH-Agilent*, *rCGH-SNP6*, or *rCGH-cytoScan*. These classes inherit from the superclass, and allow array-specific pre-parametrizations.

*frederic.commo@gustaveroussy.fr

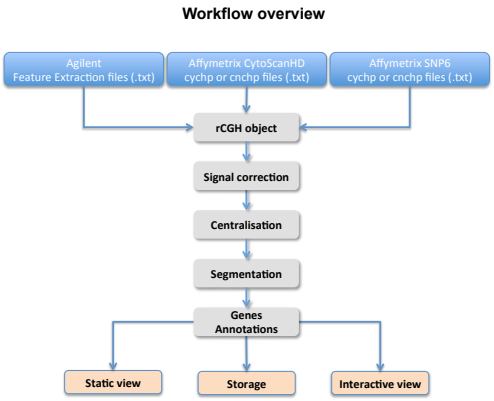


Figure 1: **rCGH workflow.** The global rCGH analysis workflow

3 rCGH functions

rCGH provides functions for each of the analysis steps, from reading files to visualizing genomic profiles. Several *get* functions allow the user to get access to specific results and workflow parameters, saved and stored at each step.

3.1 Reading files

Both Agilent Feature Exraction files (from 44K to 400K arrays), and Affymetrix SNP6 and cytoScanHD, data are supported.
However, and to keep more flexibility, Affymetrix CEL files have to be first read using ChAS or Affymetrix Power Tools (APT) [1], and then exported as cychp.txt or cnchp.txt files.
Notice that cnchp.txt files contain Allelic differences, that allow the loss of heterozygosity (LOH) to be estimated, while cychp.txt files do not.

Due to specific files structures, and since preambles may be missing (depending on ChAS and APT versions), *rCGH* as 3 specific read/build-object functions:

- `readAgilent()`: 44K to 400K FE (.txt) files.
- `readAffySNP6()`: cychp, cnchp and probeset (.txt) files, exported from SNP6.0 CEL.
- `readAffyCytoScan()`: cychp, cnchp and probeset (.txt) files, exported from CytoScanHD CEL.

Each of these functions take the file's path as the unique mandatory argument.
Optional arguments allow the user to save the following information: *sampleName, labName*:

```
> library(rCGH)
> filePath <- system.file("extdata", "Affy_cytoScan.cyhd.CN5.CNCHP.txt.bz2",
```

```
+     package = "rCGH")
> cgh <- readAffyCytoScan(filePath, sampleName = "CSc-Example",
+                         labName = "myLab")
```

```
> cgh

                                                    info
fileName          Affy_cytoScan.cyhd.CN5.CNCHP.txt.bz2
sampleName                                  CSc-Example
labName                                           myLab
usedProbes                                          snp
platform                                CytoScanHD_Array
barCode             @52082500958167113016424803602715
gridName               CytoScanHD_Array.na33.annot.db
scanDate                                    2015-01-22
programVersion                                   5.0.0
gridGenomicBuild                            hg19/GRCh37
reference           CytoScanHD_Array.na33.r1.REF_MODEL
analyseDate                                 2015-08-26
rCGH_version                                    0.99.9
```

In complement, any kind of useful annotation (logical, string or numeric) can be added, with `setInfo()`:

```
> setInfo(cgh, "item1") <- 35
> setInfo(cgh, "item2") <- TRUE
> setInfo(cgh, "item3") <- "someComment"
```

At any time, the full (or specific) annotations stored can be accessed:

```
> getInfo(cgh)

                                                    info
fileName          Affy_cytoScan.cyhd.CN5.CNCHP.txt.bz2
sampleName                                  CSc-Example
labName                                           myLab
usedProbes                                          snp
platform                                CytoScanHD_Array
barCode             @52082500958167113016424803602715
gridName               CytoScanHD_Array.na33.annot.db
scanDate                                    2015-01-22
programVersion                                   5.0.0
gridGenomicBuild                            hg19/GRCh37
reference           CytoScanHD_Array.na33.r1.REF_MODEL
analyseDate                                 2015-08-26
rCGH_version                                    0.99.9
```

```
item1                                              35
item2                                            TRUE
item3                                    someComment
```

```
> getInfo(cgh, c("item1", "item3"))
```

```
       item1         item3
        "35" "someComment"
```

3.2 Adjusting signals

When Agilent dual-color hybridization are used, GC content and the cy3/cy5 bias are necessary adjustments. `adjustSignal()` handle these steps before computing the $\log_2(RelativeRatios)$ (LRR). In both cases, a local regression (`loessFit`, R package *limma*) is used [2].
Note that by default, the cyanine3 signal is used as the reference. Use `Ref=cy5` if cyanine5 signal has to be used as the reference.
When Affymetrix cychp or cnchp files are used, these steps have already been done, and `adjustSignal()` simply rescale the LRR, when `Scale=TRUE` (default). As for Agilent data, some useful quality scores: the derivative Log Ratio Spread (dLRs) and the LRR Median Absolute Deviation (MAD), are stored in the object.

```
> cgh <- adjustSignal(cgh, nCores=1)
```

```
Log2Ratios QCs:
dLRs:  0.199
MAD: 0.24
```

```
Scaling...
Signal filtering...
Modeling allelic Difference...
```

3.3 Centering LRR

Centering LRR is a key step in the genomic analysis process since it defines the base line (the expected 2-copies level) from where gains ad losses are estimated. To do so, LRRs are considered as a mixture of several gaussian populations , and an expectation-maximization (EM) algorithm is used to estimate their parameters.
The centralization value is chosen according to the user specification: the mean of the sub-population with a density peak higher than a given proportion of the highest density peak [3]. The default value is 0.5. Setting `peakThresh = 1` leads to choose the highest density peak.
The `plotDensity()` function gives access to a graphical check on how the centralization step worked, and what LRR population has been chosen for centering the profile:

```
> # Resricted to 3 groups for the purpose of that demo.
> cgh <- EMnormalize(cgh, G = 3)

Smoothing param:  73
Analyzing mixture...
Merging peaks closer than 0.1 ...
Gaussian mixture estimation:
n.peaks =  3

Group parameters:
Grp 1:
prop:  0.391, mean:  0.035, Sd:  0.079, peak height:  1.968
Grp 2:
prop:  0.04, mean:  0.658, Sd:  0.407, peak height:  0.039
Grp 3:
prop:  0.569, mean:  1.33, Sd:  0.073, peak height:  3.096

Correction value:  0.035
Use plotDensity() to visualize the LRR densities.
```

```
> plotDensity(cgh)
```

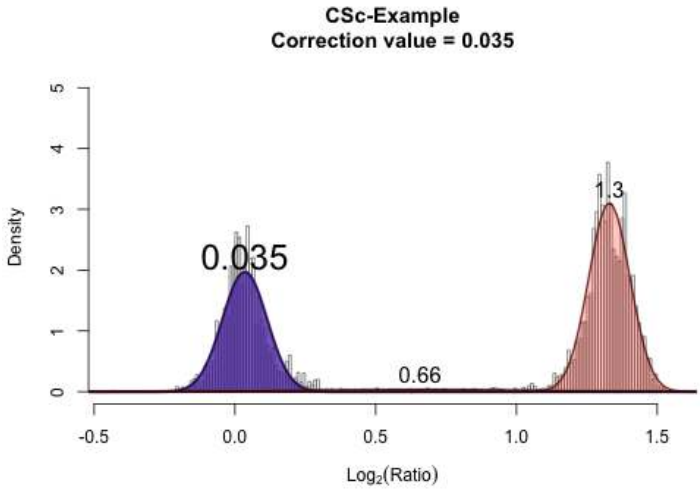


Figure 2: **plotDensity.** `plotDensity()` shows how *EM* models the *LRR* distribution, and what peak is chosen for centralizing the profile (in bold).

3.4 Segmenting

One possible strategy for segmenting the genome profile consists in identifying breakpoints all along the genome, when exist. These breakpoints define the DNA segments start and end positions. To do so, *rCGH* uses the Circular Binary Segmentation algorithm (*CBS*) [4] from the *DNAcopy* package [5]. All the steps are wrapped into one unique easy-to-use function, `segmentCGH()`. In order to faclitate its use, all the parameters but one are predefined: `UndoSD` is kept free. When this parameter is set to `NULL` (default), its optimal value is estimated directly from the values. However, the user can specify its own value, generaly from 0.5 to 1.5.
The resulting segmentation table is of the form of a standard *DNAcopy* output, plus additional columns:

- ID : sample Id.
- chrom : chromosome number.
- loc.start : segment start position.
- loc.end : segment end position.
- num.mark : number of markers within each segment.
- seg.mean : the mean LRR along each segment.
- seg.med : the median LRR along each segment.
- probes.Sd : the LRR standard deviation along each segment.

```
> cgh <- segmentCGH(cgh, nCores=1)

Computing LRR segmentation using UnodSD: 0.245
Merging segments shorter than 10Kb.
Number of segments:  26

> segTable <- getSegTable(cgh)
> head(segTable)

           ID chrom loc.start   loc.end num.mark seg.mean  seg.med probes.Sd
1 CSc.Example     1    882803 120345101      617   0.1452 -0.02710 0.6989930
2 CSc.Example     1 121155528 249198692      591   1.1393  1.20790 0.7024942
3 CSc.Example     2     15703 242775910     1316   1.2901  1.30565 0.4961236
4 CSc.Example     3     62614 197851260     1099  -0.0279 -0.02710 0.5173475
5 CSc.Example     4     46691 190921709     1041   1.3027  1.30565 0.4821446
6 CSc.Example     5    113577 180692833      985   1.3141  1.30565 0.4956702
```

Note that such data format allows GISTIC-compatible inputs to be exported [6].

3.5 Parallelization

rCGH allows parallelization within `EMnormalise()` and `segmentCGH()`, through `mclapply()` from R package *parallel*.
By default, `nCores` will be set to half of the available cores, but any value, from 1 to `detectCores()`, is allowed. However, this feature is currently only available on `Linux` and `OSX`: `nCores` will be auto-

matically set to 1 when a `Windows` system is detected.

3.6 Getting the by-gene table

The next step consists in getting access to the potentially altered genes. `byGeneTable()` extracts the list of genes included in each segment, and constructs a dataset, easy to export and to manipulate outside R. The final genes' list combines position information from *TxDb.Hsapiens.UCSC.hg19.knownGene*, and annotations from *org.Hs.eg.db*.

```
> #geneTable <- byGeneTable(cgh)
> geneTable <- byGeneTable(segTable)
```

```
Creating byGene table...
```

```
> head(geneTable, n=3)
```

```
  entrezid   symbol                       fullName cytoband chr chrStart
1        1     A1BG         alpha-1-B glycoprotein  19q13.4  19 58858172
2   503538 A1BG-AS1           A1BG antisense RNA 1  19q13.4  19 58859117
3    29974     A1CF APOBEC1 complementation factor 10q11.23  10 52559169
    chrEnd width strand Log2Ratio num.mark segNum segLength(kb) relativeLog
1 58874214 16043      -   1.30565      230     22      58836.84           0
2 58866549  7433      +   1.30565      230     22      58836.84           0
3 52645435 86267      -   1.30565      750     11     135330.87           0
  genomeStart
1  2718302494
2  2718303439
3  1732932312
```

3.7 Accessing the analysis parameters

For traceability and reproducibility, it may be useful to keep track to a profile analysis parameters. At each step, the workflow parameters, defined by default or specified by the user, are stored in a `params` slot. They are accessible at any time using `getParam()`.

```
> getParam(cgh)[1:3]
```

```
$ksmooth
[1] 73

$Kmax
[1] 20

$Nmin
```

```
[1] 160
```

3.8 Visualizing the genomic profile

In a context of Precision Medicine, visualizing and manipulating a genomic profile is crucial to interpret imbalances, to identify targetable genes, and to make decisions regarding a potential therapeutic orientation. In many situations, considering LOH can also help to better interpret imbalances.
rCGH provides 2 ways for visualizing a genomic profile: `plotProfile()`, `plotLOH()` and `multiplot()` are simple static ways to visualize a profile, possibly with some tagged gene, while `view()` is a more sophisticated and interactive visualization method, build on top of shiny. A control panel allows the user to interact with the profile, and to export the results.
Notice that `plotLOH()` and `multiplot()` are relevant only in case the allelic difference is available, namely when Affymetrix cnchp.txt files are used.

3.8.1 Static profile visualizations

`plotProfile()` allows the genomic profile visualization. Any gene(s) of interest can be added to the plot by passing a valid HUGO symbol. Other arguments can be used to color the segments according to specified gain/loss thresholds, or to change the plot title.
Two other static functions can be useful for reporting alterations: `plotLOH()` to analyse the LOH, and `multiplot()` to build a full report, including both the genomic profile and the LOH.

```
> multiplot(cgh, c("egfr", "erbb2"))
```

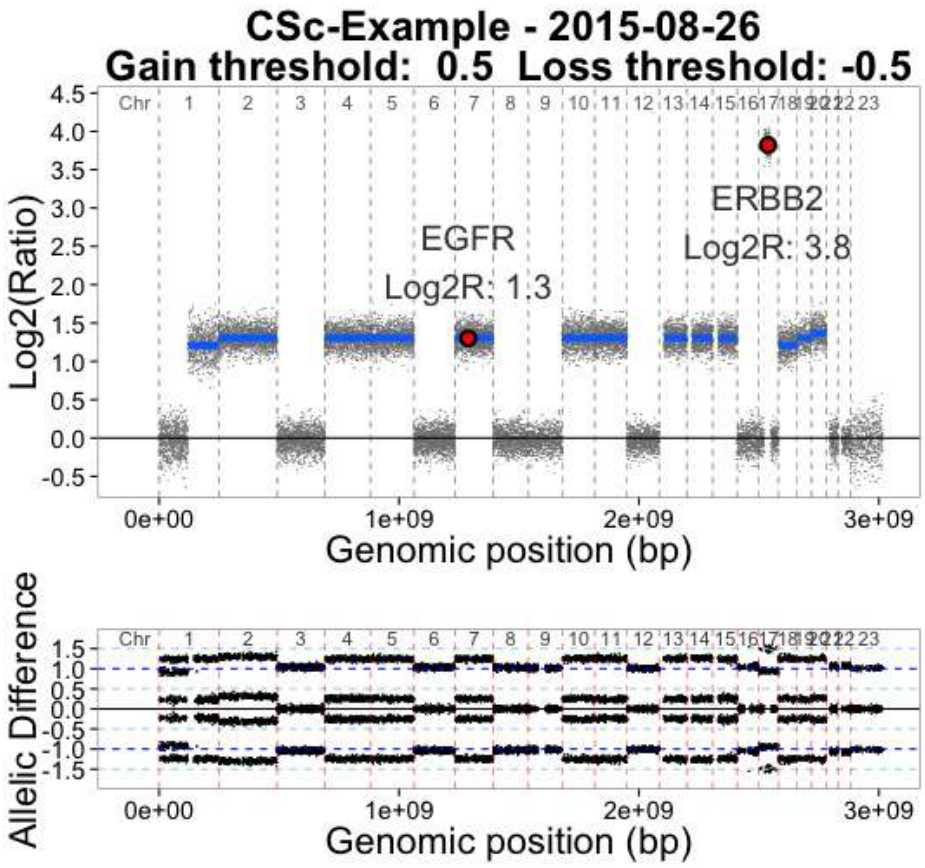


Figure 3: **Static views.** `multiplot()` provides static visualisations combining the genomic profile and the LOH.

3.8.2 Recentering

When the profile centering doesn't seem appropriate, `recenter()` allows the user to choose another centralization value. The new choice has to be specified as the peak index to use: peaks are indexed, from 1 to k (from left to right) as they appear on the density plot.

```
> # Recentering on peak #3
> recenter(cgh) <- 3

Profile recentered on:  1.33

> plotProfile(cgh, "erbb2")
```

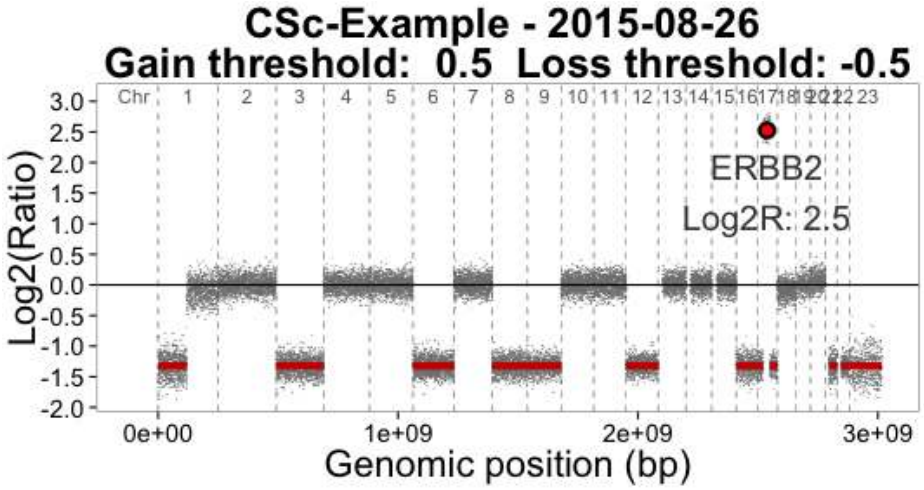


Figure 4: **Recentering.** By default, the EM-based normalization choose a possibly optimal peak to center the profile, but any other peak can be chosen, using `recenter()`.

3.8.3 Interactive visualization

The *view()* function provides a more flexible way for analyzing a genomic profile, and allows interactive graph manipulations through a control panel: defining the gain/loss thresholds, displaying a gene, resizing the y-axis, selecting one unique chromosome, and recentering the entire profile. Note that the *Genes table* is updated whenever changes are made through that control panel, e.g. selecting one unique chromosome on the graph filters the *Genes table* on that chromosome, simultaneously.
The Download buttons, *Plot*, *LOH* and *Table*, allow plots and gene table to be exported, as they have been modified.

The `view()` control panel:

- Gene Symbol : display any existing gene, providing its official HUGO symbol.
- Show chromosome : display the entire profile (default is 'All'), or one specific chromosome.
- Gain/Loss colors : choose blue/red or red/blue.
- Recenter profile : recenter the profile on-the-fly. Gene values are updated in the 'Genes table'.
- Merge segments... : merge segments shorter than the specified value, in Kb. Gene values are updated in the 'Genes table'.
- Recenter profile : recenter the profile on-the-fly. Gene values are updated in the 'Genes table'.
- Rescale max(y) : adjust the top y-axis (0¡y) using a proportion of the maximum value.
- Rescale min(y) : adjust the bottom y-axis (y¡0) using a proportion of the minimum value.
- Gain threshold (Log2ratio) : define the gain threshold. Segments higher than this value are colored according to the chosen color code, and the 'Genes table' is filtered, consequently.
- Loss threshold (Log2ratio) : same as 'Gain threshold' but for losses.
- Download - Profile : download the profile as it is displayed on the screen, including modifications.
- Download - LOH : download the LOH plot as it is displayed on the screen, including modifications.
- Download - Table : download the 'Genes table', including modifications.

```
> view(cgh)
```

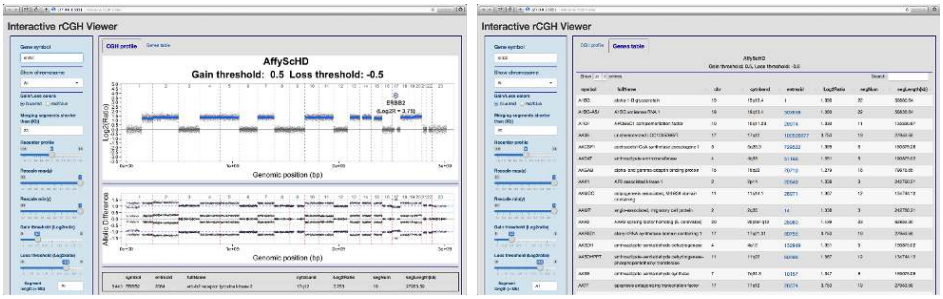


Figure 5: **Interactive profile.** The genomic profile is displayed in the first *CGH profile* tab (left). Several changes can be applied using the control panel (in blue). The list of genes is accessible through the *Genes table* tab (right). Both are updated simultaneously and can be exported, after modifications are applied.

4 Notes regarding the example files

In order to reduce the computation time, we provide subsets of real data for the 3 supported platforms:

```
> list.files(system.file("extdata", package = "rCGH"))
```

```
[1] "Affy_cytoScan.cyhd.CN5.CNCHP.txt.bz2"
[2] "Affy_snp6_cnchp.txt.bz2"
[3] "Agilent4x180K.txt.bz2"
```

comment:
In order to speed up demos, the provided example files contain only a subset of the original probes. Affymetrix example files (cytoScan and SNP6) only contain SNP probes. Setting `useProbes = "cn"` in `readAffy` functions should return an error.

5 Server version

A web browser version of the interactive visualization is available at
https://fredcommo.shinyapps.io/aCGH_viewer
As inputs, this application support the *rCGH* segmentation tables, or any segmentation table in the same format as the *CBS* outputs.
For more details about this application, or to install it on your own server, please visit
https://github.com/fredcommo/aCGH_viewer.

6 Session information

```
> toLatex(sessionInfo())
```

```
\begin{itemize}\raggedright
  \item R version 3.2.1 (2015-06-18), \verb|x86_64-apple-darwin13.4.0|
  \item Locale: \verb|C/fr_FR.UTF-8/fr_FR.UTF-8/C/fr_FR.UTF-8/fr_FR.UTF-8|
  \item Base packages: base, datasets, grDevices, graphics, methods,
    stats, utils
  \item Other packages: DBI~0.3.1, RSQLite~1.0.0, knitr~1.10.5,
    rCGH~0.99.9
  \item Loaded via a namespace (and not attached):
    AnnotationDbi~1.30.1, Biobase~2.28.0, BiocGenerics~0.14.0,
    BiocInstaller~1.18.4, BiocParallel~1.2.6, BiocStyle~1.6.0,
    Biostrings~2.36.1, DNAcopy~1.42.0, GenomeInfoDb~1.4.1,
    GenomicAlignments~1.4.1, GenomicFeatures~1.20.1,
    GenomicRanges~1.20.5, IRanges~2.2.4, MASS~7.3-40, R6~2.0.1,
```

```
    RCurl~1.95-4.6, RUnit~0.4.28, Rcpp~0.11.6, Rsamtools~1.20.4,
    S4Vectors~0.6.0, TxDb.Hsapiens.UCSC.hg19.knownGene~3.1.2,
    XML~3.98-1.2, XVector~0.8.0, aCGH~1.46.0, affy~1.46.1,
    affyio~1.36.0, biomaRt~2.24.0, bitops~1.0-6, cluster~2.0.1,
    colorspace~1.2-6, digest~0.6.8, evaluate~0.7, formatR~1.2,
    futile.logger~1.4.1, futile.options~1.0.0, ggplot2~1.0.1,
    grid~3.2.1, gtable~0.1.2, highr~0.5, htmltools~0.2.6, httpuv~1.3.2,
    labeling~0.3, lambda.r~1.1.7, lattice~0.20-31, limma~3.24.10,
    magrittr~1.5, mclust~5.0.1, mime~0.3, multtest~2.24.0,
    munsell~0.4.2, org.Hs.eg.db~3.1.2, parallel~3.2.1, plyr~1.8.3,
    preprocessCore~1.30.0, proto~0.3-10, reshape2~1.4.1,
    rtracklayer~1.28.6, scales~0.2.5, shiny~0.12.1, splines~3.2.1,
    stats4~3.2.1, stringi~0.5-2, stringr~1.0.0, survival~2.38-1,
    tools~3.2.1, xtable~1.7-4, zlibbioc~1.14.0
\end{itemize}
```

References

[1] URL: http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx.

[2] Smyth GK and Speed TP. Normalization of cdna microarray data. *Methods*, 31:265–273, 2003. URL: http://www.statsci.org/smyth/pubs/normalize.pdf.

[3] Commo F, Ferte C, Soria JC, Friend SH, Andre F, and Guinney J. Impact of centralization on acgh-based genomic profiles for precision medicine in oncology. *Ann Oncol.*, 2014.

[4] Venkatraman ES and Olshen AB. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 15(23):657–663, 2007.

[5] Venkatraman E. Seshan and Adam Olshen. *DNAcopy: DNA copy number data analysis*. R package version 1.40.0.

[6] Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, and Getz G. Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4):R41, 2011.

10.4 Annexe 4 : Algorithme LOESS

```
# LOESS demo
# FC 2015-09-02
# Ce code n'est pas optimisé pour sa vitesse d'execution
# et n'est propose quà titre indicatif.

op <- par(no.readonly = TRUE)

U <- function(X, x, a = .25){
    d <- sort(abs(X - x))
    q <- quantile(d, a)
    abs((X - x)/q)
}
W <- function(X, x, a = .25){
    u <- U(X, x, a)
    ifelse(u>1, rnorm(1, 0, 1e-9), (1-u^3)^3)
}
Fit <- function(X, Y, a = .25){
    sapply(X, function(x){
        wi <- W(X, x, a)
        B <- Solve(X, Y, wi)
        Pred(x, B)})

}
Solve <- function(X, Y, wi){
    dwi <- diag(wi)
    Xprim <- cbind(1, X, X^2)
    S <- solve(t(Xprim)%*%dwi%*%Xprim)
    S%*%t(Xprim)%*%dwi%*%Y
}
Pred <- function(x, B){
    t(B)%*%c(1, x, x^2)
}

set.seed(114)
n <- 100
X <- rnorm(n)
Y <- X^3 + X^2 + X + rnorm(n, 0, 1.5)

par(cex.main = 1.75, cex.lab = 1.5, ces.axis = 1.25)
plot(X, Y, main = "Local Regression", pch = 19, col = "cyan")
points(X, Y, cex = 1.25)
i = 1
A <- c(.1, .25, .5, .75)
for(a in A){
    message("a: ", a)
    fit <- Fit(X, Y, a)
    lines(sort(X), fit[order(X)], col = i, lwd = 4)
    i = i + 1
}
legend("topleft", legend = paste0( "h(x) = ", A), lwd = 3, col = 1:length(A),
cex = 1.25)
par(op)
```

10.5 Annexe 5 : Algorithme Expectation-Maximization (EM)

L'algorithme EM peut être utilisé pour estimer les paramètres d'un mélange de densités de probabilités, et assigner les observations à des classes :

Soit $x = \left(x_1,...,x_n\right)$, $x_i \in R^d$ *, la densité de probabilité, g, est définie par :*

$$g\left(x,\Theta\right) = \sum_K \pi_k f\left(x,\theta_k\right)$$

où f est une densité de probabilité de paramètre $\theta_k = \left\{\mu_k,\sigma_k\right\}$ *, pour chaque* $k \in K$ *.*

- L'étape E (expectation) estime la probabilité de chaque classe k conditionnée à chaque observation x_i *:*

$$p\left(k \mid x_i\right) = r_{ik} = \frac{\pi_k . f\left(x_i,\theta_k\right)}{\sum_L \pi_l . f\left(x_i,\theta_l\right)} \ , \ f\left(x,\theta_k\right) = N\left(\mu_k,\Sigma_k\right) \ , \ \sum_K \pi_k = 1$$

- L'étape C (classification) assigne chaque x_i *à la classe la plus probable :*

$$C_i = \underset{K}{\operatorname{argmax}}\left(r_{ik}\right)$$

- L'étape M (maximization) maximise les paramètres de chaque distribution, compte tenu des x_i *nouvellement assignés à chaque classe k :*

Estimation des proportions : $m_k = \sum_i r_{ik}$, $\pi_k = \frac{m_k}{\sum_K m_k}$

Estimation des centres : $\mu_k = \frac{1}{m_k}\sum_i r_{ik} . x_i$

Estimation des covariances : $\Sigma_k = \frac{1}{m_k}\sum_i r_{ik}\left(x_i - \mu_k\right)^t\left(x_i - \mu_k\right)$

Après initialisation des paramètres sur des valeurs aléatoires, chaque étape est répétée jusqu'à convergence vers une solution stable, ou jusqu'à atteindre un nombre maximal d'itérations. Une solution optimale peut être définie par l'atteinte d'un optimum d'une fonction objective telle que la log-vraisemblance, définie par :

$$L\left(\Theta,x\right) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} p\left(x_i \mid k\right).p\left(k\right)$$

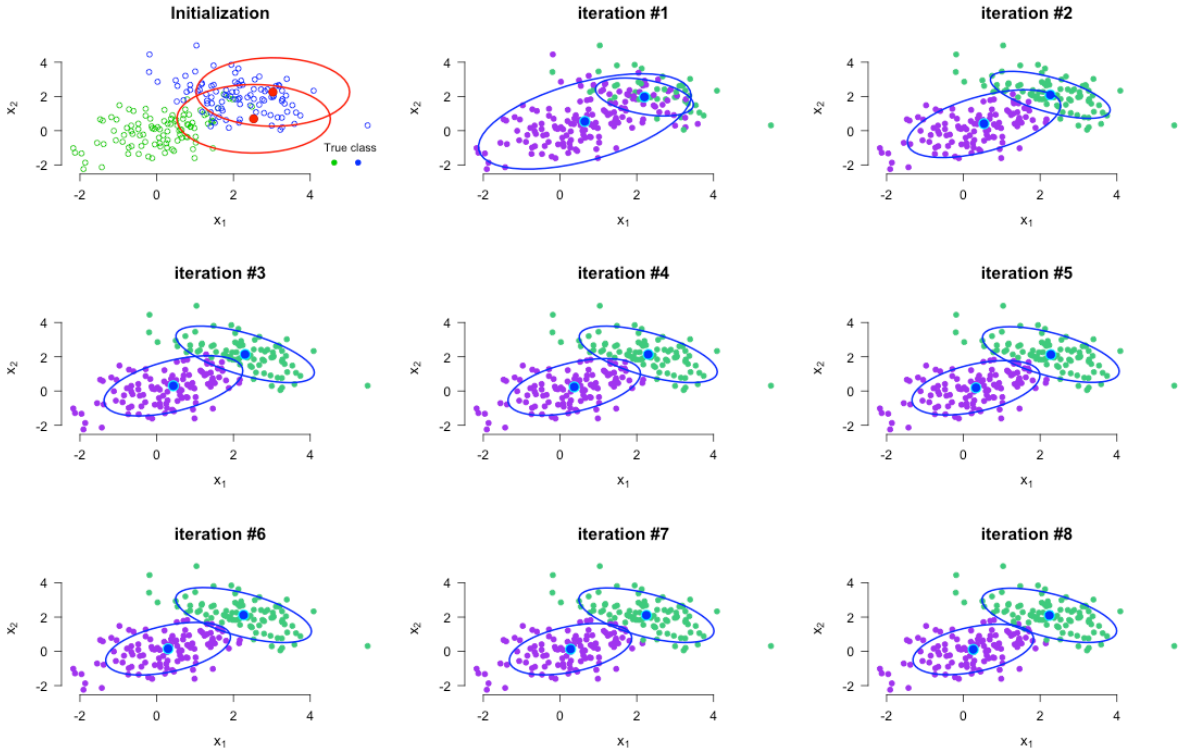
L'optimum du nombre K de classes peut également être recherché en optimisant une log-vraisemblance pondérée. Dans ce cas, des indicateurs tels que le critère d'information de Bayes (BIC) ou le critère d'information d'Akaike (AIC) peuvent être utilisés afin ajouter une pénalité relative au nombre de classes :

$$BIC : \max_K \left(L(\Theta, x) - \frac{1}{2} K \log(n) \right)$$

$$AIC : \min_K \left(2k - L(\Theta, x) \right)$$

Simulation :

Nous simulons, ci-dessous, un modèle de mélange de 2 populations, décrites par 2 variables : après initialisation des centres et variances (en haut à gauche), l'algorithme converge vers une solution optimale classant correctement 94% des observations (en bas à droite).



Code R, simulation de 2 populations.

```
################################
# Expectation-Maximisation (EM)
################################
# Required packages
library(MASS)
library(mvtnorm)
library(car)
################################
# HELPER FUNCTIONS
# E-step:
.g <- function(x, mu, sigma, p){
    d <- ncol(x)
    Sinv <- solve(sigma)
    M <- 1/(2*pi)^(d/2)*det(sigma)^(-1/2)*exp(-1/2*(x-mu)%*%Sinv%*%t(x-mu))
    return(diag(M)*p)
}
.ri <- function(x, p, mu, sigma){
    nk <- seq_len(length(p))
    L <- lapply(nk, function(ii) p[ii]*dmvnorm(x, mu[[ii]], sigma[[ii]]) )
    LS <- rowSums(do.call(cbind, L))
    if(any(LS==0)) LS[LS==0] <- 1e-6
    do.call(cbind, lapply(L, function(l) l/LS))
}

# C-step
.class <- function(ri){
    apply(ri, 1, which.max)
}

# M_step:
.mc <- function(ri, K){
    colSums(ri, na.rm=TRUE)
}
.p <- function(mc){
    mc/sum(mc)
}
.mu <- function(x, mc, ri, K){
    nk <- ncol(ri)
    #    xc <- lapply(mu, function(m) x - m )
    lapply(seq_len(nk), function(ii){
        if(sum(K==ii)>1)
            return(colMeans(x[K==ii,]))
        return(x[K==ii,])
        })
}
.sigma <- function(x, mc, ri, mu, K){
    nk <- ncol(ri)
    xc <- lapply(mu, function(m) x - m )
    lapply(seq_len(nk), function(ii){
        if(sum(K==ii)>1)
            return(var(x[K==ii,]))
        matrix(c(1,0,0,1),2,2)
    })
}
.trace <- function(M){
    sum(diag(M), na.rm=TRUE)
}
```

```
.ellipse <- function(m, s,...){
    r <- sqrt(.trace(s))*sqrt(qchisq(.95, length(m)))
e <- ellipse(m, s, r,...)
}
.plotModel <- function(x, K, mu, sigma, kk, cols = c("purple", "seagreen3",
"red")){
    plot(x, col=cols[factor(K)], pch=19,
         xlab=expression(x[1]), ylab=expression(x[2]),
         main=sprintf("iteration #%s", kk))
    lapply(mu, function(m){
        points(m[1], m[2], pch=19, cex=2, col="cyan")
    } )
    # draw confidence ellipse
    out <- lapply(seq_len(length(sigma)), function(ii){
        m <- mu[[ii]]
        s <- sigma[[ii]]
        e <- .ellipse(m, s, col="blue")
    })
}
.plotInit <- function(x, mu=mu.init, sigma=sigma.init, K=K0){
    nk <- length(mu)
    cols <- palette()[seq_len(nk)+2]
    plot(x, col=K+2,
         xlab=expression(x[1]), ylab=expression(x[2]),
         main="Initialization")
    out <- lapply(mu, function(m) points(m[1], m[2], pch=19, cex=1.5))
    out <- lapply(seq_len(nk), function(ii){
        e <- .ellipse(mu[[ii]], sigma[[ii]])
    })
    legend("bottomright", title="True class", legend=rep(" ", nk),
           pch=rep(19, nk), col=cols, ncol=nk, bty="n")
}
# END HELPER FUNCTIONS
################################
# MAIN FUNCTION
EM <- function(x, g = 2, mu=mu.init, sigma=sigma.init, maxiter = 8, Plot=TRUE){
    mu.init <- lapply(seq_len(g), function(ii) x[sample(nrow(x), 1),] )
    sigma.init <- lapply(seq_len(g), function(ii) matrix(c(1, 0, 0, 1), ncol(x),
ncol(x)) )

    p <- rep(1/g, g)

    # iterate
    for(kk in seq_len(maxiter)){
        ri <- .ri(x, p, mu, sigma);ri
        K <- .class(ri); K
        mc <- .mc(ri, K); mc
        p <- .p(mc); p
        mu <- .mu(x, mc, ri, K); mu
        sigma <- .sigma(x, mc, ri, mu, K); sigma
        if(Plot)
            .plotModel(x, K, mu, sigma, kk)
    }
    list(p=p, mu=mu, sigma=sigma, K=K)
}
L <- function(x, model){
    p <- model$p
    mu <- model$mu
    sigma <- model$sigma
```

```
    nk <- length(mu)
    l <- lapply(seq_len(nk), function(ii)
        p[ii]*dmvnorm(x, mu[[ii]], sigma[[ii]])
        )
    l <- do.call(cbind, l)
    l <- log(rowSums(l))
    sum(l)
}
# END MAIN
################################

# Start simulation
op <- par(no.readonly=TRUE)

# Simulation with 2 groups
set.seed(112234)
mu1 <- c(0, 0)
sigma1 <- matrix(c(1, 0.5, 0.5, 1),2,2)

mu2 <- c(2, 2)
sigma2 <- matrix(c(1,-.5, -.5, 1),2,2)

N <- 100
p0 <- rep(0.5, 2)
mu0 <- list(mu1, mu2)
sigma0 <- list(sigma1, sigma2)
K0 <- rep(c(1, 2), each=N)

x1 <- mvrnorm(N, mu1, sigma1)
x2 <- mvrnorm(N, mu2, sigma2)
x <- rbind(x1, x2)

g <- 2
mu.init <- lapply(seq_len(g), function(ii) x[sample(nrow(x), 1),] )
sigma.init <- lapply(seq_len(g), function(ii) matrix(c(1, 0, 0, 1), ncol(x),
ncol(x)) )

par(mfrow=c(3,3), las=1, bty="n", cex.axis=1.15, cex.lab=1.25, cex.main=1.5)
.plotInit(x, mu=mu.init, sigma=sigma.init, K=K0)
model <- EM(x, mu=mu.init, sigma=sigma.init)
par(op)

set.seed(Sys.time())

L(x, model)

finalK <- model$K
t <- table(finalK, K0)
message("Classification table:") ; t

# END SIMULATION
```

10.6 Annexe 6 : rCGH : a comprehensive array-based genomic profile platform for precision medicine. Résultats et méthodes supplémentaires.

rCGH : a comprehensive array-based genomic profile platform for precision medicine

Supplementary methods

Frederic Commo[1,2], Justin Guinney[2], Charles Ferté[1,2], Brian Bot[2], Celine Lefebvre[1], Jean-Charles Soria[1,3], Fabrice André[1,3]

1) INSERM U981, Gustave Roussy, University Paris XI, Villejuif, France
2) Sage Bionetworks, Seattle, WA
3) Department of Medical Oncology, Gustave Roussy, Villejuif, France

Workflow description

In order to keep a high level of flexibility, as well as to allow a better control of the entire process, we let the workflow decomposed in 5 five steps:

Object builders:
Since input files may not contain sufficient information in their preamble (depending on the software extraction version), we implemented three specific *read-file* functions, one for each of the supported type of array: Agilent (FE files), Affymetrix SNP6, and Affymetrix CytoScanHD (cychp.txt, cnchp.txt or probeset.txt). This step reads files, with respect to platform format specificities, saves array information, when exist, and renames each items in order to get a standardized output format. Note that any useful information can be added at any moment during the process.

Signal adjustment:
When Agilent dual-color hybridization are used, GC content and the cy3/cy5 bias are necessary adjustments, which have to be carried out before segmenting a genomic profile. This step takes care of these adjustments, before computing the LRR. In both cases, a local regression is applied[1]. Note that by default, the cyanine3 signal is used as the reference.
Since Affymetrix cychp or cnchp files contain already computed LRR, this signal adjustment step simply rescale the LRR values, if specified by the user, then stores the following quality scores: the derivative Log Ratio Spread (dLRs) and the LRR

Median Absolute Deviation (MAD). For all platforms, LRRs are finally smoothed to remove outliers, and to reduce the noise.

Profile centralization:

The centralization step is crucial since it defines a neutral level (potentially 2-copies) from which gains and losses will be estimated. As mentioned in the main manuscript, we implemented here the method we discussed in a previous paper. Briefly, the vector of LRR is considered as a mixture of several gaussian populations: $g(x,\Theta)=\sum_{k=1}^{K} p_k f(x,\theta_k)$ where p_k and θ_k are the proportion and the parameters distribution of the k^{st} population, respectively. The p_k and θ_k are estimated using an EM algorithm applied on multiple subsets of LRRs, and the mean of the lowest density higher than a given proportion (specified by the user) of the maximum density, when exists, is chosen as the centralization value. Here again, the user has full control on what should be the optimal centralization for a given profile (supplementary figure 1).

Segmentation:

This steps aims to identify breakpoints within the LRR continuity, each possibly defining the end of a DNA region and the start position of a new region with a different mean LRR value. As mentioned in the main manuscript, this step uses the CBS algorithm with slight modifications aiming to facilitate its use. The DNAcopy R package involves several parameters, which have to be set before the segmentation process (see DNAcopy R package for more details). We particularly focused on 2 parameters: the *alpha* value, which specifies a significance level for the test to accept change-points, and the *undo.SD*, which specifies the number of SDs between segment means to keep a split (when “sdundo” method is chosen as the “undo.split” procedure). We simplified the use of *DNAcopy* functions by embedding the different steps into one unique procedure, and by fixing *alpha = 1e-4*. Given this fixed alpha value, the user can either set a undo.SD value (typically between 0.5 and 1.5), or let rCGH estimating an optimal value from the MAD. The decision rule was constructed as follow:

80 genomic profiles (40 Agilent 180K, and 40 Affymetrix CytoScanHD) were manually checked in order to estimate their optimal undo.sd values, given *alpha = 1e-4*. In each case, we manually evaluated the consistency between the

segmentation and the LRR signals, given several undo.sd values. The optimal values were finally modeled as a function of MAD (supplementary figure 2):

$$undo.sd = 0.48 MAD^{\frac{1}{2}}$$

Genes table:

The goal of this final processing step is to list all the genes contained in each of the segments, to assign them the corresponding segment LRR value and length, then to build a table with all these information. To do so, Hg19 genome build, stored in a provided annotation file, is used. This file has been updated on October 28th 2013 by querying GeneBank with NCBI eUtils: http://www.ncbi.nlm.nih.gov/books/NBK25500/

Plot functions:

With this workflow we provide functions for several graphical outputs:

- The results of the EM centralization process can be displayed using plotDensity().
- plotProfile() displays a static visualization of the genomic profile.
- plotLOH() displays a static visualization of the LOH profile.
- multiplot() combines the genomic and the LOH profiles in a unique graphical report.
- view() open the profile in a web browser, and allows the user to interact with the graph in multiple ways through a control panel, as described below.

Interactive visualization functionalities:

The interactive visualization comes with a control panel and two tabs: the genomic profile is displayed in the first “CGH profile” tab, the gene table is available in “Gene table”.

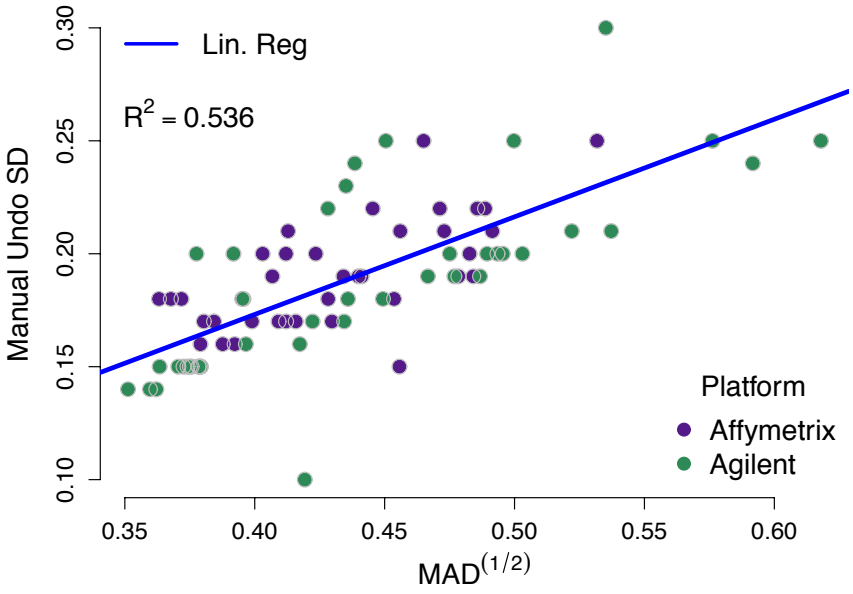
Below are described the functionalities available through the control panel.

- Gene Symbol: to display any existing gene, providing its official HUGO symbol.
- Show chromosome: to display the entire profile (default is 'All'), or one specific chromosome.

- Merging segments shorter than: to merge segments shorter than the size specified by the user, in Kb.
- Gain/Loss colors: to change the gain/loss colors. Segments are colored according to the specified thresholds.
- Recenter profile: to recenter the profile on-the-fly. The gene values will be also updated in the 'Gene table' tab.
- Rescale max(y): adjusts the top y-axis (0<y) using a proportion of the maximum value.
- Rescale min(y): adjusts the bottom y-axis (y<0) using a proportion of the minimum value.
- Gain threshold (Log2ratio): defines the gain threshold. Segments higher than this value are colored in 'gain' color, and the 'Gene table' is filtered, consequently.
- Loss threshold (Log2ratio): same as 'Gain threshold' for losses. Segments lower than this value are colored in 'loss' color, and the 'Gene table' is filtered, consequently.
- Download - Profile: to download the profile as it is displayed on the screen, including modifications.
- Download - LOH: to download the LOH plot as it is displayed on the screen, including modifications.
- Download - Table: to download the 'Gene table', including modifications.

Validation

To validate our workflow, we ran the analysis of 995 cell lines from the Cancer Cell Lines Encyclopedia[2] (CCLE), and compared our results with those already published and available at Gene Expression Omnibus (GEO, GSE36138). Affymetrix SNP6 CEL files were downloaded from http://www.broadinstitute.org/ccle/home, and processed using APT version 1.16.1: 942 out of the 995 CEL files matched the data available at GEO. The cychp.txt output files were processed using rCGH with the default parameters, and the newly generated profiles were compared with the corresponding GEO data by computing Pearson correlations on gene LRRs. We defined a high/low correlation threshold, as the quantile $q_{1e\text{-}2}$ of a normal distribution of same mean and standard deviation as the rho values, after Fisher transformation.

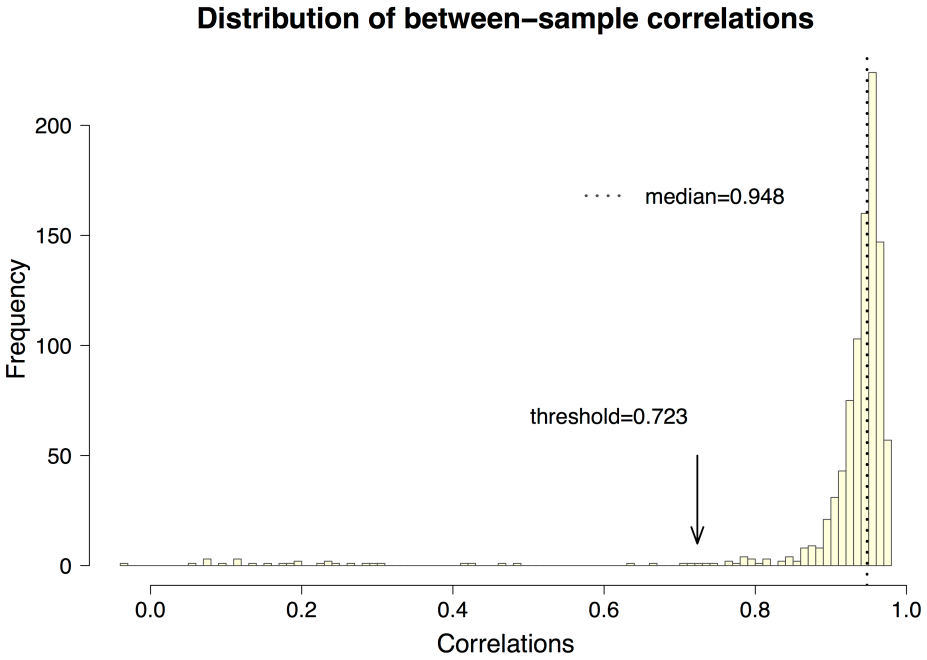


Since the major density peak may not always correspond to a neutral two-copies state, the centralization step allows a tolerance, expressed as a proportion of the highest density peak. When specified (default is 0.5), a peak with a density higher than proportion threshold can be used to centralize the entire profile.

Supplementary figure 2

Parameters optimization

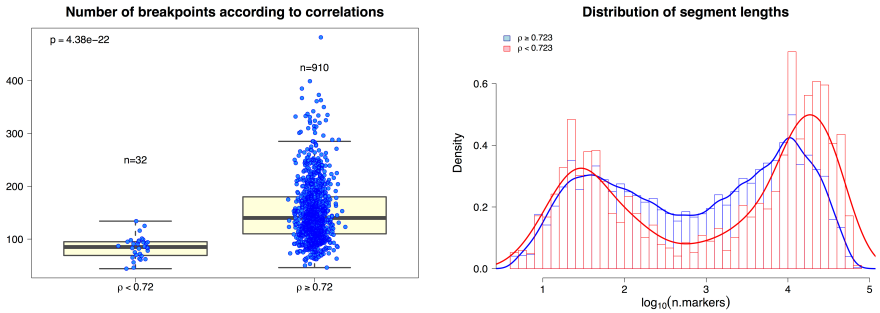
80 genomic profiles (40 Agilent 180K, and 40 Affymetrix CytoScanHD) were manually checked in order to estimate their optimal undo.sd values, given *alpha = 1e-4*. Optimal values were then modeled as a linear function of $MAD^{1/2}$.



Supplementary Figure 3

Between-profiles correlations

995 CCLE Affymetrix SNP6 CEL files were downloaded from Broad, then preprocessed using APT (version 1.16.1). Finally, cychp.txt files were used to reprocess the genomic profiles using rCGH workflow. Results were compared with the original data (GSE36138), using Pearson correlations of genes LRR, on 942 matched cell lines. The median of cell-cell profile correlations was 0.948, and 96.6% of the profiles (910/942) had a correlation greater than 0.723.



Supplementary Figure 4

Non-correlated profiles

When analyzing in details the 32 out of 942 samples with low correlations (rho < 0.723), we noted that 1) these profiles were characterized by a significantly lower number of breakpoints (p<2e-16) (left panel), and 2) breakpoints defined essentially very large and very short segments, while the correlated profiles showed copy number alteration of intermediate size (right panel). In case of short segments, it can be challenging to distinguish real altered DNA regions from noise, and profiles are generally more sensitive to small changes in analysis parameters. This may explain the discrepancies between the profiles generated through rCGH, and the original data on these particular cases.

Supplementary references

1. Smyth, G. K. & Speed, T. *Methods* **31,** 265–73 (2003).

2. Barretina, J. *et al. Nature* **483,** 603–7 (2012).