



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ludwig Institute for Cancer Research

Department of medical genetics

**COMPUTATIONAL AND STATISTICAL ANALYSIS OF COPY
NUMBER VARIANTS IN NORMAL AND CANCER GENOMES**

Thèse de doctorat ès sciences de la vie (PhD)

Présentée à la Faculté de biologie et médecine de l'Université de Lausanne par

ARMAND VALSESIA

Diplômé de l'Université Bordeaux I & II (France)

Master de Bioinformatique

Jury

Prof. Andreas Mayer, Président
Prof. Victor Jongeneel, Directeur de thèse
Prof. Sven Bergmann, co-Directeur de thèse
Prof. Nigel Carter, Expert
Prof. Alexandre Reymond, Expert

Lausanne, 2011



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président	Monsieur Prof. Andreas Mayer
Directeur de thèse	Monsieur Prof. Victor C. Jongeneel
Co-directeur de thèse	Monsieur Prof. Sven Bergmann
Experts	Monsieur Prof. Alexandre Reymond
	Monsieur Prof. Nigel Carter

le Conseil de Faculté autorise l'impression de la thèse de

Monsieur Armand Valsesia

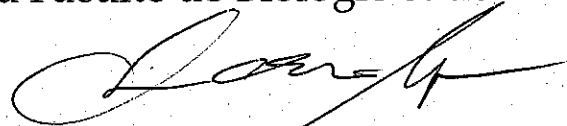
Master en bioinformatique Université de Bordeaux, France

intitulée

**COMPUTATIONAL AND STATISTICAL ANALYSIS
OF COPY NUMBER VARIANTS IN NORMAL AND CANCER GENOMES**

Lausanne, le 1 avril 2011

pour Le Doyen
de la Faculté de Biologie et de Médecine



Prof. Andreas Mayer

Abstract

Although the genomes from any two human individuals are more than 99.99% identical at the sequence level, some structural variation can be observed. Differences between genomes include single nucleotide polymorphism (SNP), inversion and copy number changes (gain or loss of DNA). The latter can range from submicroscopic events (CNVs, at least 1kb in size) to complete chromosomal aneuploidies. Small copy number variations have often no (lethal) consequences to the cell, but a few were associated to disease susceptibility and phenotypic variations. Larger re-arrangements (i.e. complete chromosome gain) are frequently associated with more severe consequences on health such as genomic disorders and cancer. High-throughput technologies like DNA microarrays enable the detection of CNVs in a genome-wide fashion. Since the initial catalogue of CNVs in the human genome in 2006, there has been tremendous interest in CNVs both in the context of population and medical genetics. Understanding CNV patterns within and between human populations is essential to elucidate their possible contribution to disease. But genome analysis is a challenging task; the technology evolves rapidly creating needs for novel, efficient and robust analytical tools which need to be compared with existing ones. Also, while the link between CNV and disease has been established, the relative CNV contribution is not fully understood and the predisposition to disease from CNVs of the general population has not been yet investigated.

During my PhD thesis, I worked on several aspects related to CNVs. As I will report in chapter 3, I was interested in computational methods to detect CNVs from the general population. I had access to the CoLaus dataset, a population-based study with more than 6,000 participants from the Lausanne area. All these individuals were analysed on SNP arrays and extensive clinical information were available. My work explored existing CNV detection methods and I developed a variety of metrics to compare their performance. Since these methods were not producing entirely satisfactory results, I implemented my own method which outperformed two existing methods. I also devised strategies to combine CNVs from different individuals into CNV regions.

I was also interested in the clinical impact of CNVs in common disease (chapter 4). Through an international collaboration led by the Centre Hospitalier Universitaire Vaudois (CHUV) and the Imperial College London I was involved as a main data analyst in the investigation of a rare deletion at chromosome 16p11 detected in obese patients. Specifically, we compared 8,456 obese patients and 11,856 individuals from the general population and we found that the deletion was accounting for 0.7% of the morbid obesity cases and was absent in healthy non-obese controls. This highlights the importance of rare variants with strong impact and provides new insights in the design of clinical studies to identify the missing heritability in common disease.

Furthermore, I was interested in the detection of somatic copy number alterations (SCNA) and their consequences in cancer (chapter 5). This project was a collaboration initiated by the Ludwig Institute for Cancer Research and involved other groups from the Swiss Institute of Bioinformatics, the CHUV and Universities of Lausanne and Geneva. The focus of my work was to identify genes with altered expression levels within somatic copy number alterations (SCNA) in seven metastatic melanoma cell lines, using CGH and SNP arrays, RNA-seq, and karyotyping. Very few SCNA genes were shared by even two melanoma samples making it difficult to draw any conclusions at the individual gene level. To overcome this limitation, I used a network-guided analysis to determine whether any pathways, defined by amplified or deleted genes, were common among the samples. Six of the melanoma samples were potentially altered in four pathways and five samples harboured copy-number and expression changes in components of six pathways. In total, this approach identified 28 pathways. Validation with two external, large melanoma datasets confirmed all but three of the detected pathways and demonstrated the utility of network-guided approaches for both large and small datasets analysis.

Résumé

Bien que le génome de deux individus soit similaire à plus de 99.99%, des différences de structure peuvent être observées. Ces différences incluent les polymorphismes simples de nucléotides, les inversions et les changements en nombre de copies (gain ou perte d'ADN). Ces derniers varient de petits événements dits sous-microscopiques (moins de 1kb en taille), appelés CNVs (copy number variants) jusqu'à des événements plus large pouvant affecter des chromosomes entiers. Les petites variations sont généralement sans conséquence pour la cellule, toutefois certaines ont été impliquées dans la prédisposition à certaines maladies, et à des variations phénotypiques dans la population générale. Les réarrangements plus grands (par exemple, une copie supplémentaire d'un chromosome appelée communément trisomie) ont des répercussions plus grave pour la santé, comme par exemple dans certains syndromes génomiques et dans le cancer. Les technologies à haut-débit telle les puces à ADN permettent la détection de CNVs à l'échelle du génome humain. La cartographie en 2006 des CNV du génome humain, a suscité un fort intérêt en génétique des populations et en génétique médicale. La détection de différences au sein et entre plusieurs populations est un élément clef pour élucider la contribution possible des CNVs dans les maladies. Toutefois l'analyse du génome reste une tâche difficile, la technologie évolue très rapidement créant de nouveaux besoins pour le développement d'outils, l'amélioration des précédents, et la comparaison des différentes méthodes. De plus, si le lien entre CNV et maladie a été établi, leur contribution précise n'est pas encore comprise. De même que les études sur la prédisposition aux maladies par des CNVs détectés dans la population générale n'ont pas encore été réalisées.

Pendant mon doctorat, je me suis concentré sur trois axes principaux ayant attiré aux CNV. Dans le chapitre 3, je détaille mes travaux sur les méthodes d'analyses des puces à ADN. J'ai eu accès aux données du projet CoLaus, une étude de la population de Lausanne. Dans cette étude, le génome de plus de 6'000 individus a été analysé avec des puces SNP et de nombreuses informations cliniques ont été récoltées. Pendant mes travaux, j'ai utilisé et comparé plusieurs méthodes de détection des CNVs. Les résultats n'étant pas complètement satisfaisant, j'ai implémenté ma propre méthode qui donne de meilleures performances que deux des

trois autres méthodes utilisées. Je me suis aussi intéressé aux stratégies pour combiner les CNVs de différents individus en régions.

Je me suis aussi intéressé à l'impact clinique des CNVs dans le cas des maladies génétiques communes (chapitre 4). Ce projet fut possible grâce à une étroite collaboration avec le Centre Hospitalier Universitaire Vaudois (CHUV) et l'Imperial College à Londres. Dans ce projet, j'ai été l'un des analystes principaux et j'ai travaillé sur l'impact clinique d'une délétion rare du chromosome 16p11 présente chez des patients atteints d'obésité. Dans cette collaboration multidisciplinaire, nous avons comparés 8'456 patients atteint d'obésité et 11'856 individus de la population générale. Nous avons trouvés que la délétion était impliquée dans 0.7% des cas d'obésité morbide et était absente chez les contrôles sains (non-atteint d'obésité). Notre étude illustre l'importance des CNVs rares qui peuvent avoir un impact clinique très important. De plus, ceci permet d'envisager une alternative aux études d'associations pour améliorer notre compréhension de l'étiologie des maladies génétiques communes.

Egalement, j'ai travaillé sur la détection d'altérations somatiques en nombres de copies (SCNA) et de leurs conséquences pour le cancer (chapitre 5). Ce projet fut une collaboration initiée par l'Institut Ludwig de Recherche contre le Cancer et impliquant l'Institut Suisse de Bioinformatique, le CHUV et les Universités de Lausanne et Genève. Je me suis concentré sur l'identification de gènes affectés par des SCNAs et avec une sur- ou sous-expression dans des lignées cellulaires dérivées de mélanomes métastatiques. Les données utilisées ont été générées par des puces ADN (CGH et SNP) et du séquençage à haut débit du transcriptome. Mes recherches ont montrées que peu de gènes sont récurrents entre les mélanomes, ce qui rend difficile l'interprétation des résultats. Pour contourner ces limitations, j'ai utilisé une analyse de réseaux pour définir si des réseaux de signalisations enrichis en gènes amplifiés ou perdus, étaient communs aux différents échantillons. En fait, parmi les 28 réseaux détectés, quatre réseaux sont potentiellement dérégulés chez six mélanomes, et six réseaux supplémentaires sont affectés chez cinq mélanomes. La validation de ces résultats avec deux larges jeux de données publiques, a confirmée tous ces réseaux sauf trois. Ceci démontre l'utilité de cette approche pour l'analyse de petits et de larges jeux de données.

Résumé grand public

L'avènement de la biologie moléculaire, en particulier ces dix dernières années, a révolutionné la recherche en génétique médicale. Grâce à la disponibilité du génome humain de référence dès 2001, de nouvelles technologies telles que les puces à ADN sont apparues et ont permis d'étudier le génome dans son ensemble avec une résolution dite *sous-microscopique* jusque-là impossible par les techniques traditionnelles de cytogénétique. Un des exemples les plus importants est l'étude des variations structurales du génome, en particulier l'étude du nombre de copies des gènes. Il était établi dès 1959 avec l'identification de la trisomie 21 par le professeur Jérôme Lejeune que le gain d'un chromosome supplémentaire était à l'origine de syndrome génétique avec des répercussions graves pour la santé du patient. Ces observations ont également été réalisées en oncologie sur les cellules cancéreuses qui accumulent fréquemment des aberrations en nombre de copies (telles que la perte ou le gain d'un ou plusieurs chromosomes). Dès 2004, plusieurs groupes de recherches ont répertorié des changements en nombre de copies dans des individus provenant de la population générale (c'est-à-dire sans symptômes cliniques visibles). En 2006, le Dr. Richard Redon a établi la première carte de variation en nombre de copies dans la population générale. Ces découvertes ont démontrées que les variations dans le génome était fréquentes et que la plupart d'entre elles étaient bénignes, c'est-à-dire sans conséquence clinique pour la santé de l'individu. Ceci a suscité un très grand intérêt pour comprendre les variations naturelles entre individus mais aussi pour mieux appréhender la prédisposition génétique à certaines maladies.

Lors de ma thèse, j'ai développé de nouveaux outils informatiques pour l'analyse de puces à ADN dans le but de cartographier ces variations à l'échelle génomique. J'ai utilisé ces outils pour établir les variations dans la population suisse et je me suis consacré par la suite à l'étude de facteurs pouvant expliquer la prédisposition aux maladies telles que l'obésité. Cette étude en collaboration avec le Centre Hospitalier Universitaire Vaudois a permis l'identification d'une délétion sur le chromosome 16 expliquant 0.7% des cas d'obésité morbide. Cette étude a plusieurs répercussions. Tout d'abord elle permet d'effectuer le diagnostique chez les enfants à naître afin de déterminer leur prédisposition à l'obésité. Ensuite ce locus implique une vingtaine de gènes. Ceci permet de formuler de nouvelles hypothèses de travail et d'orienter la

recherche afin d'améliorer notre compréhension de la maladie et l'espoir de découvrir un nouveau traitement. Enfin notre étude fournit une alternative aux études d'association génétique qui n'ont eu jusqu'à présent qu'un succès mitigé.

Dans la dernière partie de ma thèse, je me suis intéressé à l'analyse des aberrations en nombre de copies dans le cancer. Mon choix s'est porté sur l'étude de mélanomes, impliqués dans le cancer de la peau. Le mélanome est une tumeur très agressive, elle est responsable de 80% des décès des cancers de la peau et est souvent résistante aux traitements utilisés en oncologie (chimiothérapie, radiothérapie). Dans le cadre d'une collaboration entre l'Institut Ludwig de Recherche contre le Cancer, l'Institut Suisse de Bioinformatique, le CHUV et les universités de Lausanne et Genève, nous avons séquencés l'exome (les gènes) et le transcriptome (l'expression des gènes) de sept mélanomes métastatiques, effectués des analyses du nombre de copies par des puces à ADN et des caryotypes. Mes travaux ont permis le développement de nouvelles méthodes d'analyses adaptées au cancer, d'établir la liste des réseaux de signalisation cellulaire affectés de façon récurrente chez le mélanome et d'identifier deux cibles thérapeutiques potentielles jusqu'alors ignorées dans les cancers de la peau.

Remerciements

Je voudrais exprimer ma gratitude à mes directeurs de thèse les professeurs Victor Jongeneel et Sven Bergmann. Victor et Sven m'ont permis de travailler sur des projets fascinants, stimulants et à la pointe de la technologie. J'ai bénéficié tout au long de ma thèse d'une liberté exceptionnelle qui m'a permis d'essayer mes idées, bonnes ou mauvaises, d'apprendre de celles-ci et de progresser. Ils étaient également disponibles lorsque que j'avais besoin de conseils. Je me suis senti très privilégié de la confiance qu'ils m'ont accordée et d'avoir accès à des données de recherche clinique pour lesquelles de nombreux postdocs auraient été prêts à s'entretuer pour travailler dessus.

Je souhaiterais remercier Jacques Beckmann, directeur du département de génétique médicale et du service de génétique médicale du CHUV. Jacqui a été un mentor et a montré un grand intérêt dans mes travaux. J'ai aussi bénéficié de ces nombreux conseils tout au long de ma thèse.

Je suis tout particulièrement reconnaissant à Brian Stevenson pour toute sa supervision et ses conseils lors de mon doctorat. Brian a une double spécialité en biologie moléculaire et en bioinformatique. Cette rare expertise a été cruciale dans mes différents projets, en particulier pour mes travaux en génomique du cancer. J'ai énormément apprécié nos discussions sur les différents résultats qu'ils s'agissent de projets en commun ou de publications avec d'autres groupes. Ces débats m'ont permis de mieux comprendre la biologie, et surtout de développer un esprit très critique sur comment interpréter un résultat. Brian m'a aussi initié à la rédaction de façon claire, précise et concise de publications scientifiques. Loin d'être devenu un expert dans ce domaine, j'ai pu acquérir certaines bases qui s'affirmeront avec l'expérience.

Je suis également très reconnaissant à mon ami et collègue Zoltan Kutalik. Zoltan est un expert en mathématique et statistique, qui en plus d'être le fer de lance de nombreuses études internationales, trouve toujours le temps pour aider ses collègues. Son expertise, et pas seulement en mathématique, a été un moteur lors de ma thèse. Avec le recul, je m'aperçois à quel point j'ai pu apprendre ces trois

dernières années. Je réalise surtout que ces connaissances sont des graines que je continuerai de cultiver et de faire germer tout au long de ma carrière scientifique. J'ai aussi beaucoup apprécié la compagnie de Zoltan, lors de nos conférences à l'étranger et de nos réunions interminables. Merci pour ton aide précieuse et ta compagnie si agréable ! (J'espère que mes leçons pour survivre aux dîners de conférence te seront utiles)

Un grand merci à mon jury de thèse : Nigel Carter du Wellcome Trust Sanger Institute ; Alexandre Reymond et Andreas Mayer de l'université de Lausanne ; pour leur temps précieux et leur intérêt pour mes travaux.

Merci également à des collaborateurs proches, Christian Iseli, Donata Rimoldi, la famille Hor (Charlotte et Hyun, bonne chance à Barcelone !), Amalio Telenti, Medhi Tafti et Stylianos Antonarakis pour leur confiance en mes analyses, de nombreuses discussions stimulantes et des collaborations réussites.

Je voudrais aussi remercier tous mes collègues et amis du DGM (Aitana, Diana, Valérie, Gigi, Bastian, Gabor, Karen, Micha, Sascha, Jean, Danielle and Toby) et du SIB (Vassilios (mon professeur préféré de bouzouki !), Viviane (la meilleure chef du Valais !), Manu, Luli, Ioannis, Mark (et ses fantastiques whiskies !), Christian, Frédéric et de nombreux autres personnes qui se reconnaîtront). J'ai passé un très bon moment avec vous tous.

Un très grand merci à mes amis de la mafia bordelaise (Philippe, Irène, Alexandra, Yohan and Walid); le club des inoxydables (Séverine, Marc, Guigui, Laure, Pedro, Yannick, Thomas); mes amis (anciens) étudiants du SIB (Julien, Romain, Frédéric, Yannick), Manu et Josiane; Luli et Alix; Mathilde et Gaétan ; et bien d'autres encore qui se reconnaîtront aussi.

A ma famille française et grecque qui m'a soutenue toutes ces années. Merci à mes petites sœurs Aude et Anne-Cécile pour leurs corrections sur le résumé grand public. A mes parents et beaux parents, et à papi et mamie !

Je dédie cette thèse à Mado, qui m'a encouragée à postuler en thèse, qui m'a aidée et permis d'avancer pendant les moments difficiles. Avec tout mon amour.

Content

1	Introduction	12
1.1	Structural variation in the human genome	12
1.2	Mechanisms for CNV genesis.....	15
1.2.1	Non-allelic homologous recombination	15
1.2.2	Non-homologous end joining.....	17
1.2.3	Other CNV genesis mechanisms.....	18
1.3	Platforms for CNV detection.....	21
1.3.1	Microarray-based methods	22
1.3.2	Sequencing-based methods	28
1.3.3	Methods used in CNV validation	31
1.4	The clinical impact of copy number variants	35
1.4.1	CNVs and genomic disorder	35
1.4.2	CNVs and gene expression in the general population.....	37
1.5	Somatic copy number alterations in cancer.....	38
1.6	Scope of the thesis	41
1.6.1	Identification and validation of Copy Number Variants using SNP genotyping arrays from a large clinical cohort.....	41
1.6.2	Aetiology of CNVs in complex disease	42
1.6.3	Detection and impact of somatic copy number alterations in cancer.....	43
1.7	References	45
2	Methods.....	57
2.1	Normalization methods	57
2.1.1	Mean and median scaling	57
2.1.2	Linear Least Square Regression.....	59
2.1.3	Loess smoothing.....	60
2.1.4	Quantile-quantile normalization.....	61
2.2	Segmentation methods	63
2.2.1	Outlier-based detection	63
2.2.2	Recursive binary segmentation	65
2.2.3	Dynamic programming techniques	66
2.2.4	Linear piecewise regression	67
2.2.5	Hidden Markov Model	68
2.3	Multivariate and cluster analysis.....	71
2.3.1	Principal Component Analysis.....	71
2.3.2	Gaussian Mixture Models	73
2.3.3	K-means clustering.....	76
2.3.4	Hierarchical clustering	77
2.3.5	Self-organizing maps.....	81
2.4	References	83

3	Identification and validation of Copy Number Variants using SNP genotyping arrays from a large clinical cohort.....	86
3.1	Abstract	88
3.2	Author Summary	89
3.3	Introduction	89
3.4	Methods.....	92
3.4.1.	Ethics Statement.....	92
3.4.2.	CNV calling.....	92
3.4.3.	CNV merging	95
3.5	Results	96
3.5.1.	Identification of Copy Number Variants in CoLaus	96
3.5.2.	Comparison with known CNVs	99
3.5.3.	Validation with Illumina arrays.....	100
3.5.4.	Predicting relatedness between individuals based on their CNV profile	102
3.6	Discussion	104
3.6.1.	Properties of the PCA merging technique.....	104
3.6.2.	Comparison of the different CNV prediction methods	104
3.6.3.	Validation of CNVs in a large clinical cohort.....	105
3.6.4.	Conclusion and Perspectives	105
3.7	Authors and affiliations	106
3.8	Acknowledgements	107
3.9	References	107
4	Aetiology of CNVs in complex disease	110
4.1	Abstract	111
4.2	Methods summary	112
4.3	Results	113
4.4	Methods.....	120
4.5	Authors and affiliations	125
4.6	Acknowledgements	128
4.7	Author Contributions.....	130
4.8	Author Information	130
4.9	References	131
5	Detection and impact of somatic copy number alterations in cancer	135
5.1	Abstract	136
5.2	Introduction	137
5.3	Methods.....	138
5.3.1.	Melanoma samples, DNA and RNA extraction	138
5.3.2.	Cytogenetic and FISH analysis	139
5.3.3.	Comparative genomic arrays (CGH).....	139
5.3.4.	Single Nucleotide polymorphism arrays (SNP).....	140
5.3.5.	Transcriptome sequencing.....	140
5.3.6.	Detection of somatic copy number alterations with altered expression (SCNA-genes)	140
5.3.7.	Protein network-guided analysis of SCNA	140
5.3.8.	Pathway analysis	142
5.4	Results	142
5.4.1.	CGH and SNP arrays are required to comprehensively document somatic copy-number alterations in metastatic melanoma cell lines.....	142

5.4.2.	Few SCNA-genes are recurrent in different melanoma cell lines.....	147
5.4.3.	Pathways significantly enriched in SCNA-genes are recurrent in melanoma	151
5.5	Discussion	154
5.6	Accession numbers.....	158
5.7	Authors and affiliations	158
5.8	Acknowledgements	159
5.9	Author contributions	159
5.10	References	160
6	Outlook.....	165
6.1	Utility of methods for CNV analysis.....	165
6.1.1	Mining medical cohorts for rare CNVs.....	165
6.1.2	CNV-based genome-wide association studies	166
6.2	Perspectives in Obesity	167
6.2.1	Follow-up studies	167
6.2.2	Implication for medical genetics research.....	167
6.3	Perspectives in Melanoma.....	168
6.3.1	Investigation of the pathways enriched in SCNAs.....	168
6.3.2	Further characterization of FRS2 and EPHA3	169
6.3.3	Incorporating results from exome sequencing	170
6.3.4	Implications of our methodology for cancer genomics.....	170
6.4	The future of human genetics	171
6.5	References	173
Annexes.....		175

1 Introduction

1.1 Structural variation in the human genome

Genetic variation in the human genome takes many forms ranging from large chromosome anomalies (segmental aneuploidy) to single nucleotide polymorphisms (SNPs) (See definitions in Table 1). Deletion, insertion and duplication events giving rise to copy number variations (CNVs) have been found genome-wide in humans¹⁻⁸ and other species⁹⁻¹⁷. For historical reasons related to the resolution detection, CNVs are defined as events longer than 1kb; smaller events (100bp-1kb) are referred as indels.

Initial detection of copy number changes was made decades ago in human cytogenetics with karyotype and microscope observations. Chromosomal abnormalities in structure and copy number were associated to disease. A well known example is the Trisomy 21, also called Down syndrome, that was identified in 1959 by Jérôme Lejeune¹⁸. With the completion of the Human Genome Project in 2001 and the availability of clone libraries, new technologies such as DNA microarrays were developed and offered higher throughput and resolution to study *sub-microscopic* copy number changes. Rapidly, large and small copy number aberrations were detected in cancer cells using comparative genome hybridization arrays (CGH)¹⁹ then in 2004 independent groups reported structural variation in apparently-healthy individuals^{1,6}. Subsequently, in 2006, Redon et al. published the first genome-wide map of CNVs in the general population³ (see Figure 1.). This map showed that about 12% of the genome was covered with CNVs; re-estimation with higher resolution arrays reduced this fraction to 5%²⁰.

The observation that CNVs could occur both in normal and disease populations has opened a new chapter in human genomics. CNVs have been further explored within the human population in European²¹, African^{22,23}, and several Asian populations: Chinese²⁴, Japanese²⁵, Korean²⁶⁻²⁸. CNV comparisons have also been performed between human populations^{3,7,29-33} and across apes^{9,10,34,35}. CNVs have been shown to play an adaptive environmental role in the evolution of different populations. Remarkable examples include: the copy number of the amylase gene, correlated to starch diet in different human populations³⁶; the increase of copy number of the

CCL3L1 gene that protects against HIV infection ³⁵; and the evolution of olfactory receptor genes through CNV events ³⁷. Because CNVs constitute a major force in genetic diversity, with consequence in term of evolution and disease susceptibility, their identification and association to quantitative traits and clinical phenotypes, constitute a important and fascinating task.

Term	Description	Reference
Structural Variant	A genomic alteration (e.g., a CNV, an inversion that involves segments of DNA >1kb)	2
Copy number variant (CNV)	A duplication or deletion event involving >1 kb of DNA	
Copy number polymorphism (CNP)	CNV with frequency > 1% in a population	
Duplicon	A duplicated genomic segment > 1 kb in length with >90% similarity between copies	
Indel	Variation from insertion or deletion event involving <1kb of DNA	
Intermediate-sized structural variant	A structural variant that is \approx 8 kb to 40 kb in size. This can refer to a CNV or a balanced structural rearrangement (e.g., an inversion)	4
Low copy repeat	Similar to segmental duplication	38
Multisite variant	Complex polymorphic variation that is neither a PSV nor a SNP	39
Paralogous sequence variant	Sequence difference between duplicated copies (paralogs)	40
Segmental duplication (SD)	Duplicated region ranging from 1 kb upward with a sequence identity of >90%	40
Interchromosomal SD	Duplications distributed among non homologous chromosomes	
Intrachromosomal SD	Duplications restricted to a single chromosome	
Single nucleotide polymorphism (SNP)	Base substitution involving only a single nucleotide; \approx 10 million are thought to be present in the human genome at >1%, leading to an average of one SNP difference per 1250 bases between randomly chosen individuals	41
Mosaicism	Mosaicism is the presence of cells within an organism that have a different genetic composition despite deriving from a single zygote.	42
Aneuploidy	Abnormal number of chromosomes in a cell	
Segmental aneuploidy	Abnormal copy number for a portion of a chromosome	

Table 1 Structural variation terms - adapted from ⁵

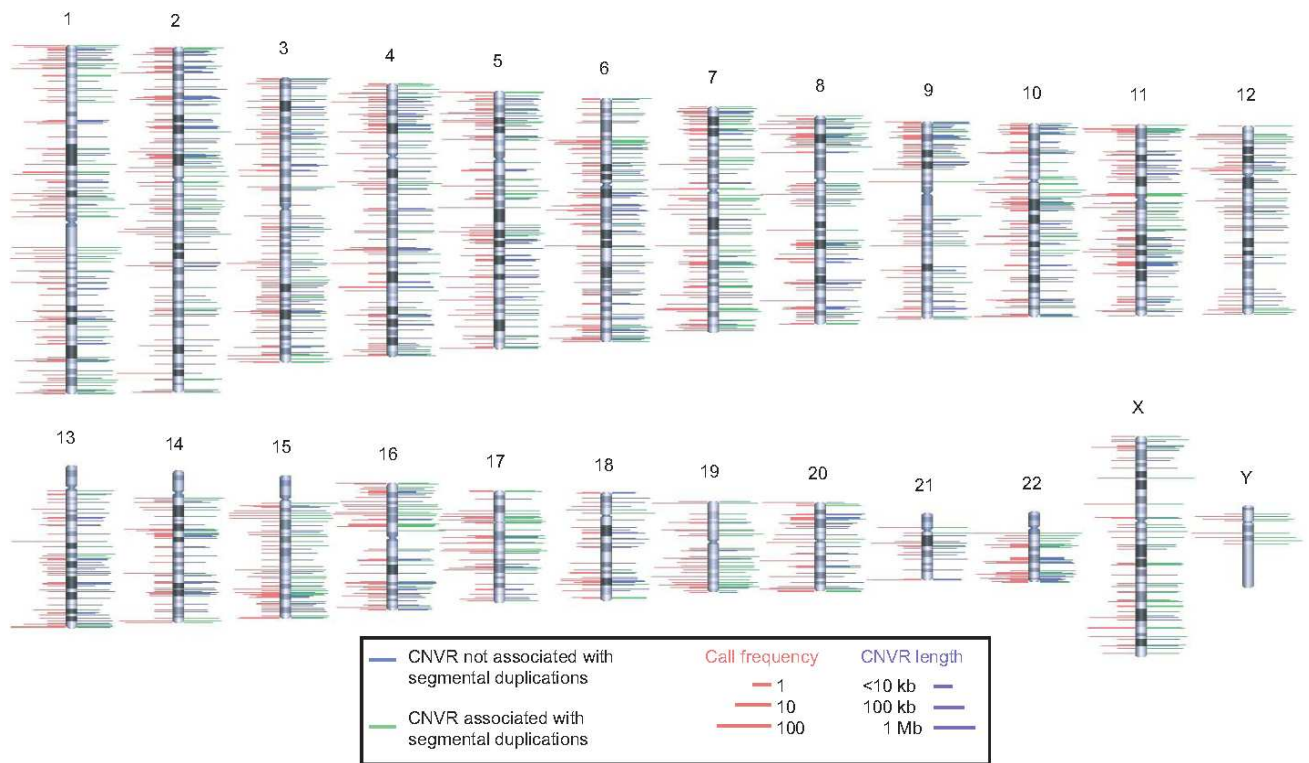


Figure 1 Initial genome wide map of CNV in the human genome - from ³

The chromosomal locations of 1,447 CNVRs are indicated by lines to either side of ideograms. Green lines denote CNVRs associated with segmental duplications; blue lines denote CNVRs not associated with segmental duplications. The length of right-hand side lines represents the size of each CNVR. The length of left-hand side lines indicates the frequency that a CNVR is detected (minor call frequency among 270 HapMap samples). When both platforms identify a CNVR, the maximum call frequency of the two is shown. For clarity, the dynamic range of length and frequency are log transformed (see scale bars).

1.2 Mechanisms for CNV genesis

Understanding the mechanisms of CNV genesis is important for the estimation of yet undiscovered CNVs in the genome and to improve our understanding about the distribution and evolution of CNVs in the genome. There are three major mechanisms: non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ) and until recently retrotransposon L1 elements^{43,44}. Both NAHR and NHEJ are involved in the repair of DNA stranded breaks. Several others mechanisms have been proposed: retrotransposons; non-B DNA structures; and Fork Stalling and Template Switching (FoSteS).

1.2.1 Non-allelic homologous recombination

Homologous recombination is a mechanism by which DNA is exchanged between two similar sequences (i.e. with nearly identical identity). This process is frequently used to repair double-stranded DNA breaks and also occur in meiosis where homologous chromosomes can form Holliday Junction structure resulting in cross-over or gene conversion (see Figure 2). Non-allelic homologous recombination is a process where non homolog sequences can exchange DNA^{38,45}, it occurs during Prophase I in meiosis (with a frequency between 10^{-6} and 5×10^{-5} per gamete⁴⁶) and it was shown that newborns carried less copy number changed due to NAHR than older individuals, suggesting that such rearrangements can accumulate in somatic cells during life⁴⁷. Rearrangements in germline cells are inheritable by definition, changes in somatic cells will either have no impact or be lethal to the cell; in very rare cases they can contribute to tumorigenesis. Several groups observed that CNVs were closely associated to segmental duplications (SDs)^{1,4,8}. Similar observations were made in genomic disorder studies, where the microdeletion or microduplication were clustered in the vicinity of SDs⁴⁸⁻⁵⁰ or Alu repeats^{51,52}. It was shown that SDs were in fact facilitating NAHR events leading to copy number changes⁴⁸, an illustration is given in Figure 3. NAHR is considered as a major force driving CNV genesis⁴⁵ and is more frequent than non-homology driven processes⁵. By contrast, the association between Alu repeats and CNVs has been minimized within recent studies^{32,53}.

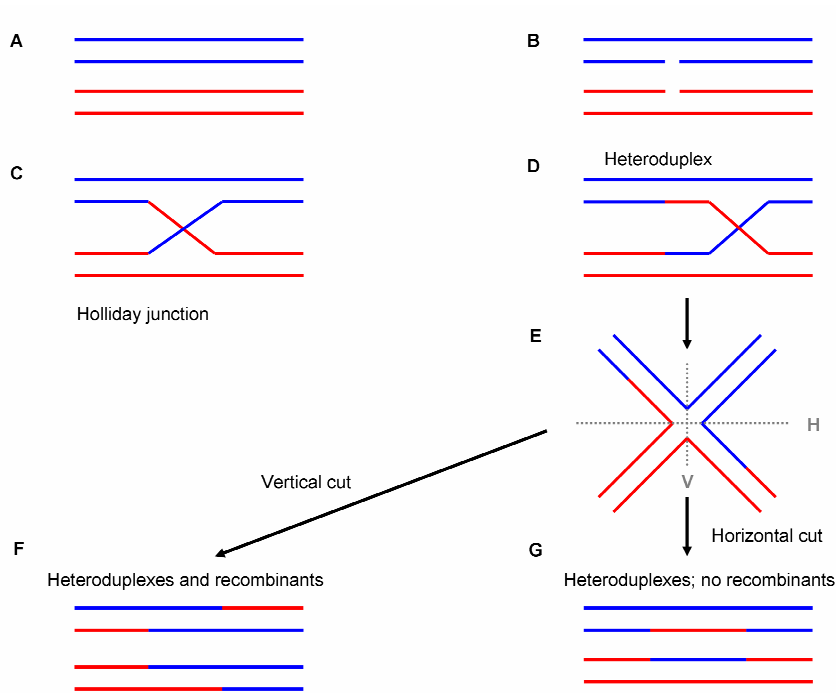


Figure 2 Holliday junction structure

A Two homologous chromosomes are paired, during meiosis, B DNA single strand break; C the broken strand cross and exchange DNA in a Holliday junction; D Heteroduplex region formed by branch migration; E simplified form of the Holliday junction leading to either cross-over (F) or gene conversion (G).

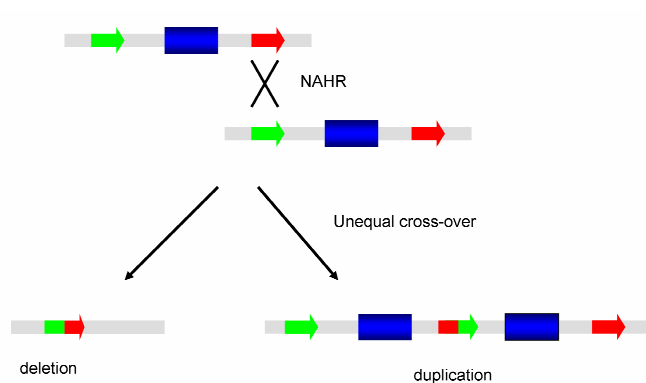


Figure 3 NAHR induced by segmental duplications

The segmental duplications indicated with the green and red arrows are mistaken for homologous alleles. Arrow heads indicates the direction of the duplication (here direct repeat). Here the cross-over leads to deleted and duplicated products.

1.2.2 Non-homologous end joining

Non-homologous end joining (NHEJ) is a pathway to repair double-stranded DNA breaks (DSB). DSBs arise when the replication fork encounters a nick or when DNA is damaged with ionizing radiation or reactive oxygen species. NHEJ is the major DSB repair pathway, because 1) it does not require a substrate with homology, as opposed to NAHR; and 2) it can operate during the cell cycle (NAHR can only operate during late S and during G2). NHEJ has been extensively reviewed by Lieber⁵⁴. When a DSB occurs, the damaged DNA is excised with nucleases, then new nucleotides are inserted with polymerases and finally ligases restore the phosphodiester bond (see Figure 4). This mechanism is rather imprecise and often lead to small deletions or duplications. Such errors have been implicated in genomic disorders^{55,56} and in cancer^{57,58}.

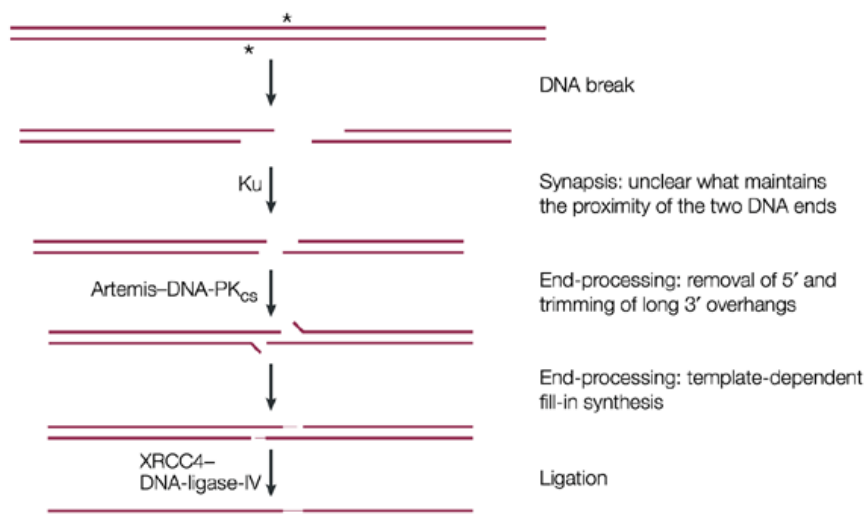


Figure 4 DSB repair via NHEJ - from⁵⁷

When a double-stranded DNA break (DSB) occurs, the ends must be held in proximity to allow subsequent repair steps to proceed and to align the two ends. This first step can be referred to as synapsis. Ku and the DNA-dependent protein kinase catalytic subunit (DNA-PK_{cs}) bind to DNA ends during this initial phase, although it is not clear how the synapsis occurs or what proteins specifically carry out this function. End-alignment can occur if there is terminal microhomology of, typically, 1–4 nucleotides between the two ends. This is an optional aspect, as non-homologous DNA end-joining (NHEJ) occurs regardless of microhomology. End-processing refers to the removal of DNA by the Artemis–DNA-PK_{cs} complex and the filling in of gaps by polymerases. Ligation is the final step, and it requires a ligatable nick on each strand. Ligation in NHEJ is done by the XRCC4 (X-ray cross complementation 4)–DNA-ligase-IV complex.

1.2.3 Other CNV genesis mechanisms

Retrotransposons

Long interspersed element-1 (L1) elements are the only active transposons in the human genome⁵⁹⁻⁶¹. The L1 family is estimated to contain about 600,000 copies, among which less than 100 have two open-reading frames (ORF): ORF1 encodes an RNA-binding protein and ORF2 encodes a protein with endonuclease and reverse-transcriptase activities. Integration of L1 element in the genome is made via a mechanism named target-primed reverse transcription (see Figure 5 and⁶²). The contribution of L1 elements to CNV has been reviewed in⁶³. From fosmid end-paired sequencing analysis, Kidd and colleagues found L1 elements accounted for about 30% of the detected indels⁶⁴. In the high-resolution CNV discovery project³², Conrad and colleagues found that small duplications were more likely to be induced by retrotransposition and VNTRs (variable number tandem repeats, these repeats can be mobilized by L1 elements) whereas longer duplications were more likely to be caused by NAHR re-arrangements. Disease like Haemophilia A was shown to be induced by de-novo insertion of L1 elements⁶⁵.

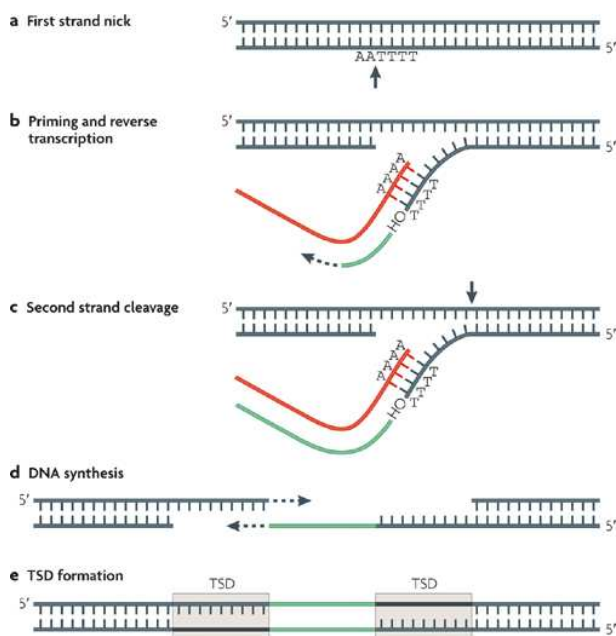


Figure 5 Integration of L1 element via target-primed reverse transcription - adapted from⁶²

During target-primed reverse transcription (TPRT), the L1 endonuclease cleaves the first strand of target DNA, generally between T and A at 5'-TTTTAA-3' consensus sites¹³³ (a). The free 3' hydroxyl (OH) generated by the nick is then used to prime reverse transcription of L1 RNA (red) by the L1 reverse transcriptase (b). The second strand of the target DNA is cleaved (c) and used to prime second-strand synthesis (d) through poorly understood mechanisms. Hallmarks of the integration process include frequent 5' truncations, the presence of an oligo(dA)-rich tail at the 3' end and target site duplications (TSDs) of between 2 and 20 base pairs in length (e).

Non-B DNA structures

Non canonical DNA structures (non-B DNA) have been studied over the past 35 years, with about one new structure type discovered every three years ⁶⁶. Specific sequence motifs like direct, mirror or inverted repeats can undergo structural modification from the canonical B DNA (right handed helical) structure to higher energy state non-B DNA structures. Non-B DNA examples include slipped hairpin loops, cruciforms, left-handed Z helices, triplexes and tetraplexes (Figure 6, see also ⁶⁶). These structures have been shown to coincide with deletion breakpoints ⁶⁷ and are thought to trigger genomic rearrangements through recombination-repair activities (NAHR).

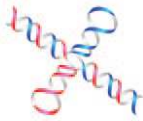
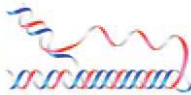
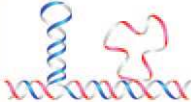
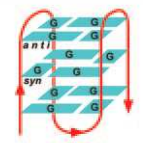

Name	Conformation	General Seq. Requirements	Sequence
Cruciform		Inverted Repeats	<pre> TCGGTACCGA AGCCATGGCT </pre>
Triplex		(R•Y) _n Mirror Repeats	<pre> AAGAGG GGAGAA TTCTCC CCTCTT </pre>
Slipped (Hairpin) Structure		Direct Repeats	<pre> TCGGTTCGGT AGCCAAGCCA </pre>
Tetraplex		Oligo (G) _n Tracts	AG ₃ (T ₂ AG ₃) ₃ single strand
Left-handed Z - DNA		(YR•YR) _n	CGCGTGCGTGTG GCGCACGCACAC

Figure 6 Non-B DNA conformations involved in rearrangements - adapted from.⁶⁶

In 2009, Conrad et al. found that two motifs of non-B DNA structures (G-quadruplexes and slipped DNA) were significantly over-represented in CNP breakpoints ³². It was also demonstrated that duplications were significantly more enriched in G-quadruplexes than deletions. These observations confirmed that some non-B conformation can indeed participate in genomic rearrangement and breakpoint formation.

Fork Stalling and Template Switching

Fork Stalling and Template Switching (FoSTeS) has first been proposed in 2007 by Lee and colleagues⁶⁸ as a novel replication-based mechanism to explain non-recurrent duplications of the dosage-sensitive proteolipid protein 1 (*PLP1*) gene, involved in the Pelizaeus-Merzbacher disease^{55,69} This replication-model contrasts with NAHR and NHJE which are recombination-based mechanisms. The proposed FoSTeS model was thought to be a consequence of a stall in the replication fork (see Figure 7). This stall can be resolved but with DNA lesions resulting from the genome instability near SDs. Following such stall, the lagging strand can switch, invade another replication fork and be replicated with the progression of this second fork (progression of this second fork depends of its leading strand and could be moving 5'->3' or 3'->5').

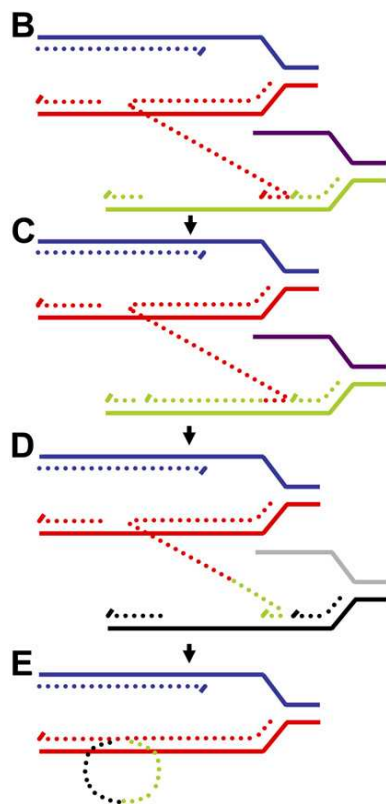


Figure 7 Illustration of the FoSTeS Model - adapted from⁶⁸
(B), one replication fork (dark blue and red, solid lines) with a lagging strand (red, dotted line) would invade a second fork (purple and green, solid lines), followed by (C) DNA synthesis (green, dotted line). After the fork disengages (D), the original fork (dark blue and red, solid lines) with its lagging strand (red and green, dotted lines) could invade a third fork (gray and black, solid lines). Dotted lines represent newly synthesized DNA. Serial replication fork disengaging and lagging strand invasion could occur several times before (E) resumption of replication on the original template.

1.3 Platforms for CNV detection

Gross copy number alterations were initially detected with karyotyping (Giemsa banding) (see Figure 8A) in the early days of cytogenetics. Several large-scale aberrations such as trisomy 21 and many cancer aberrations were identified before the development of techniques with higher resolutions. Fluorescence in situ hybridization (FISH) has increased this resolution, enabling the detection of *sub-microscopic* copy number changes that could not be detected with karyotyping (Figure 8B). Methods have continued to evolve with DNA microarrays and more recently ultra-high throughput sequencing (UHTS) technologies. Today, the most widely used techniques can be classified as amplification-based (e.g. polymerase chain reaction), hybridization based (e.g. FISH, CGH and SNP arrays) or sequencing-based. These techniques differ in precision, throughput and resolution. For example, FISH, Southern blot, and PCR-approaches (like long-range PCR) are work intensive and thus not suitable for CNV analysis in large cohorts. However these methods, as well as novel ones, are used for the CNV validation at targeted loci.

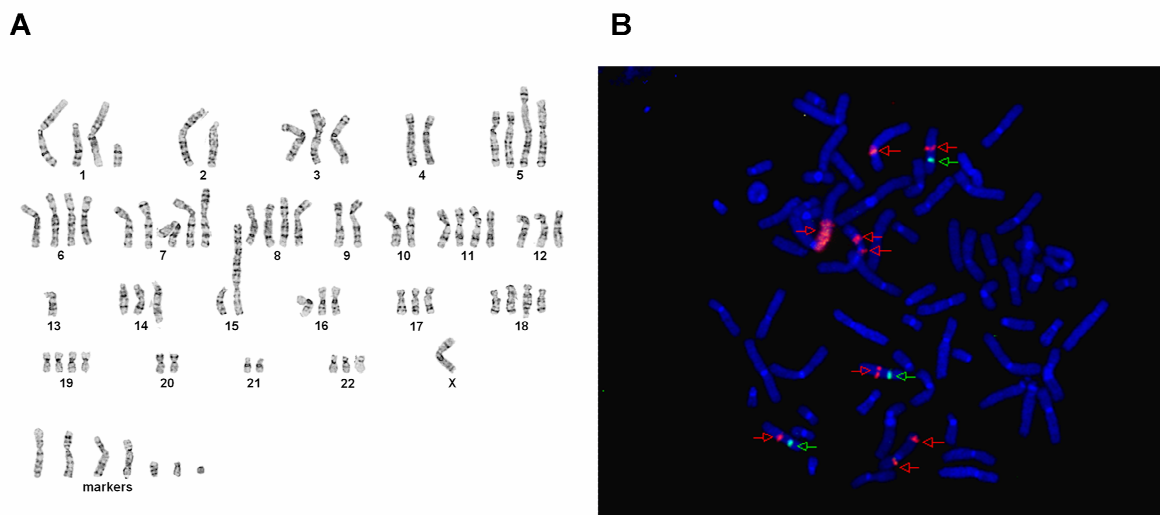


Figure 8 Karyotype and FISH

A Karyotype of a metastatic melanoma (Me275), here it has 76 chromosomes and 7 unclassified markers, **B** FISH analysis of the *MDM2* oncogene. *MDM2* probe is in red; centromere-specific probe is in green. Here the sample (Me275) has 8 copies on chromosome 12, and a large ectopic amplicon located on another chromosome.

1.3.1 Microarray-based methods

Comparative Genome Hybridization arrays

Comparative Genome Hybridization (CGH) is a method that compares the relative copy number of a test DNA with respect to a reference DNA ^{19,70,71}. Both DNAs are labelled with different dyes (a red and a green), then are hybridized competitively onto a chromosome (metaphase) spread. A ratio of relative copy number changes can then be measured; significant deviation from the baseline indicates copy number gain or loss with respect to the reference. With the completion of Human Genome Project, libraries of large-insert clones (i.e. bacterial artificial chromosome (BAC), cosmid or fosmid clones) were developed and spotted onto a microarray (slide) ⁷²⁻⁷⁴. Hybridization could then be performed onto such microarray with a higher throughput and resolution than on metaphases chromosomes (Figure 9). The length of the BAC clones ranged from 150 to 200kb and did not allow the detection of small CNVs (<50Kb). To overcome this limitation, oligonucleotide CGH arrays were developed, using spotted probes that were synthesized *in-situ* with 25-85 oligonucleotides. Current oligonucleotide CGH arrays have a median resolution of one probe every 2.1kb for Agilent 1M arrays and 1.1kb with Roche Nimblegen 2.1M arrays. CGH arrays have become a popular method for the identification of CNV both in tumours ⁷⁵⁻⁷⁷ and in clinical diagnosis ^{78,79} (see also Figure 10).

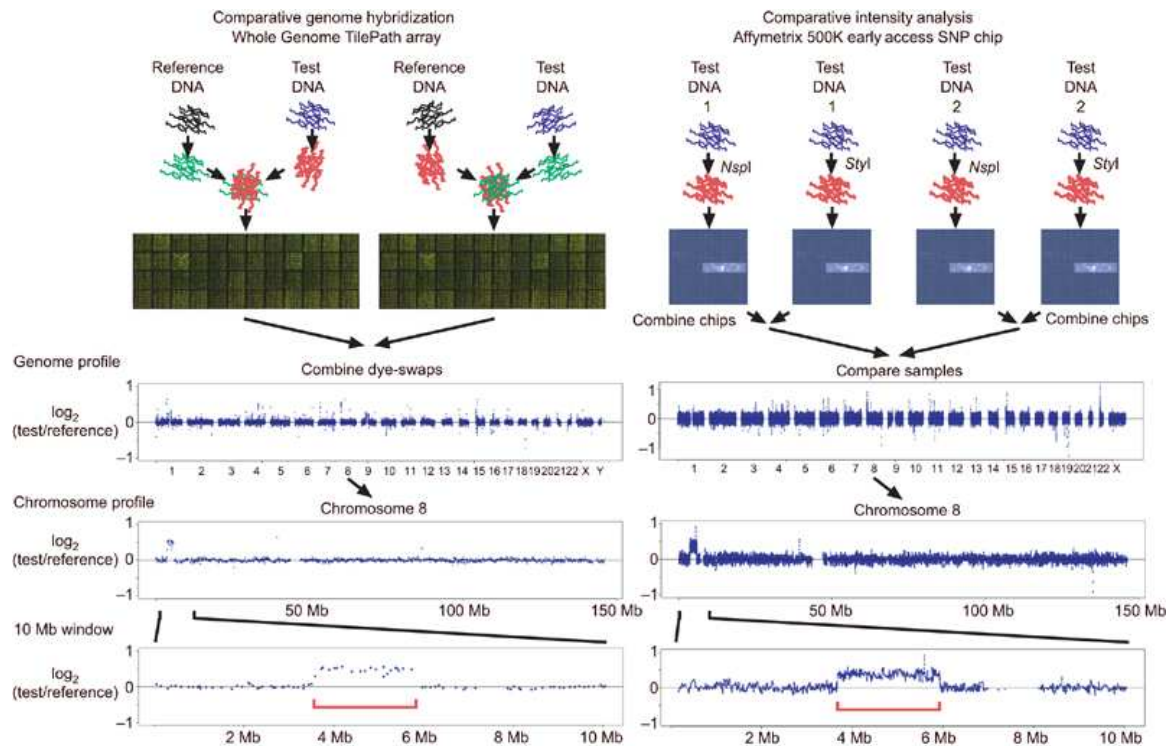


Figure 9 Protocol outline for two CNV detection platforms, adapted from ³
 Experimental procedures are displayed for comparative genome hybridization (WGTP array) and SNP comparative intensity analysis (500K EA Affymetrix SNP array). The genome profile shows the \log_2 ratio of copy number in these two genomes chromosome-by-chromosome. Below the genome profiles are expanded plots of chromosome 8, and a 10-Mb window containing a large duplication was identified on both platforms (as indicated by the red bracket).

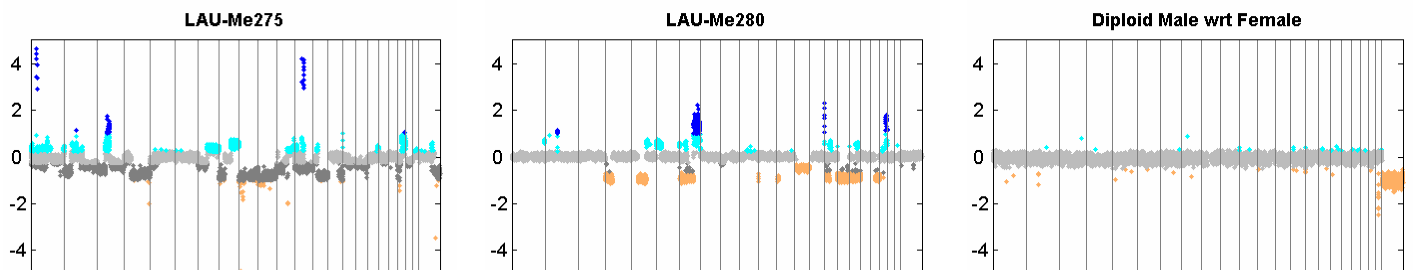


Figure 10 CNV predictions from CGH array
 The two first plots correspond to CGH copy number profile in human tumours (melanoma); the last plot to a (human) male individual hybridized with respect to a female reference. Chromosomes 1 to X are delimited with vertical gray bars, Y axis corresponds to \log_2 ratio of copy number, duplication (amplification) are shown in cyan (dark blue); copy neutral events in light and dark gray; and deletions in orange.

Single nucleotide polymorphism genotyping arrays

The Hapmap project has played a major role in the discovery and characterization of single nucleotide polymorphism (SNP) in four major human populations (European, African, Chinese and Japanese). This has generated tremendous interest in population genetics and medical genetics. Manufacturers like Affymetrix and Illumina design and commercialize DNA arrays to genotype SNPs. With Affymetrix arrays, DNA sequences are digested, ligated to adaptators, amplified and hybridized to the array (Figure 11A). DNA sequences hybridize to 25 mers probes that describe specific SNP allele. These hybridization intensities can be measured at each SNP allele and subsequently used to perform the genotyping call (Figure 11B). By contrast to CGH arrays where two individuals are hybridized on the chip; a single genome is assayed with SNP arrays. Data from Hapmap or any other dataset (with several experiments), can be used as a reference to call the SNP genotype. Illumina SNP array uses a slightly different technique: 1) DNA sequence hybridizes to a target probe; 2) the nucleotide corresponding to the SNP is added in a single base extension phase. A and T nucleotides are labelled with biotin whereas C and G are labelled with dinintrophenol. 3) Antibodies with fluorochrome are added, these recognize either the biotin or dinintrophenol and allow to detect which nucleotide has been incorporated. An illustration is given in Figure 12.

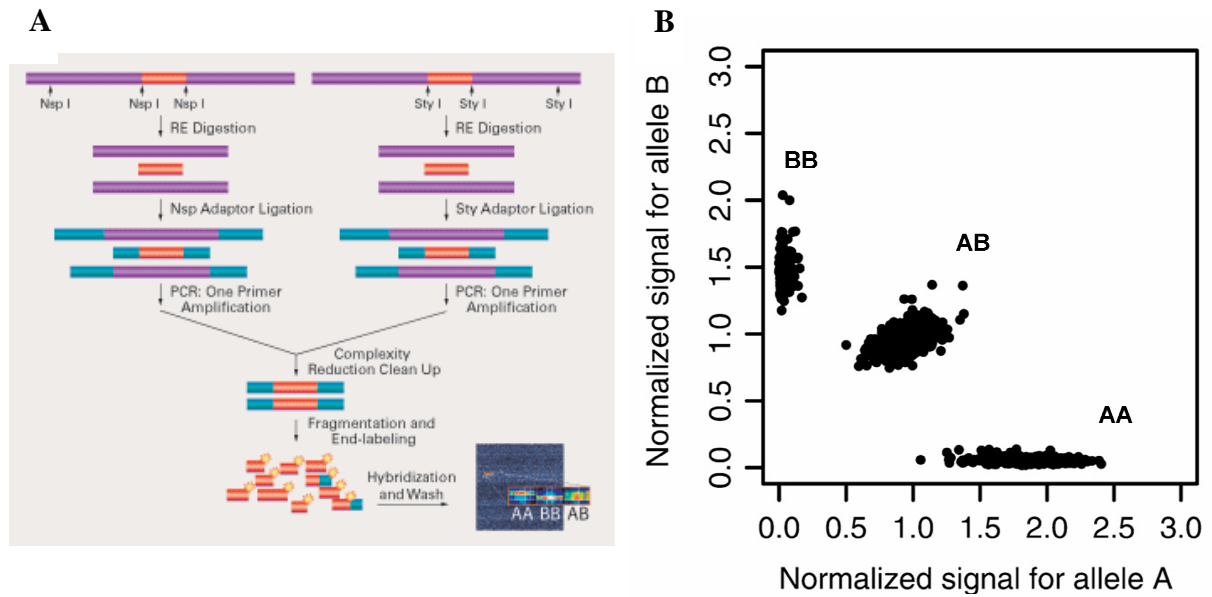


Figure 11 SNP genotyping analysis

A Affymetrix SNP array protocol (from www.affymetrix.com), **B** genotype calling based on clustering of allele-specific intensities (adapted from ⁸⁰)

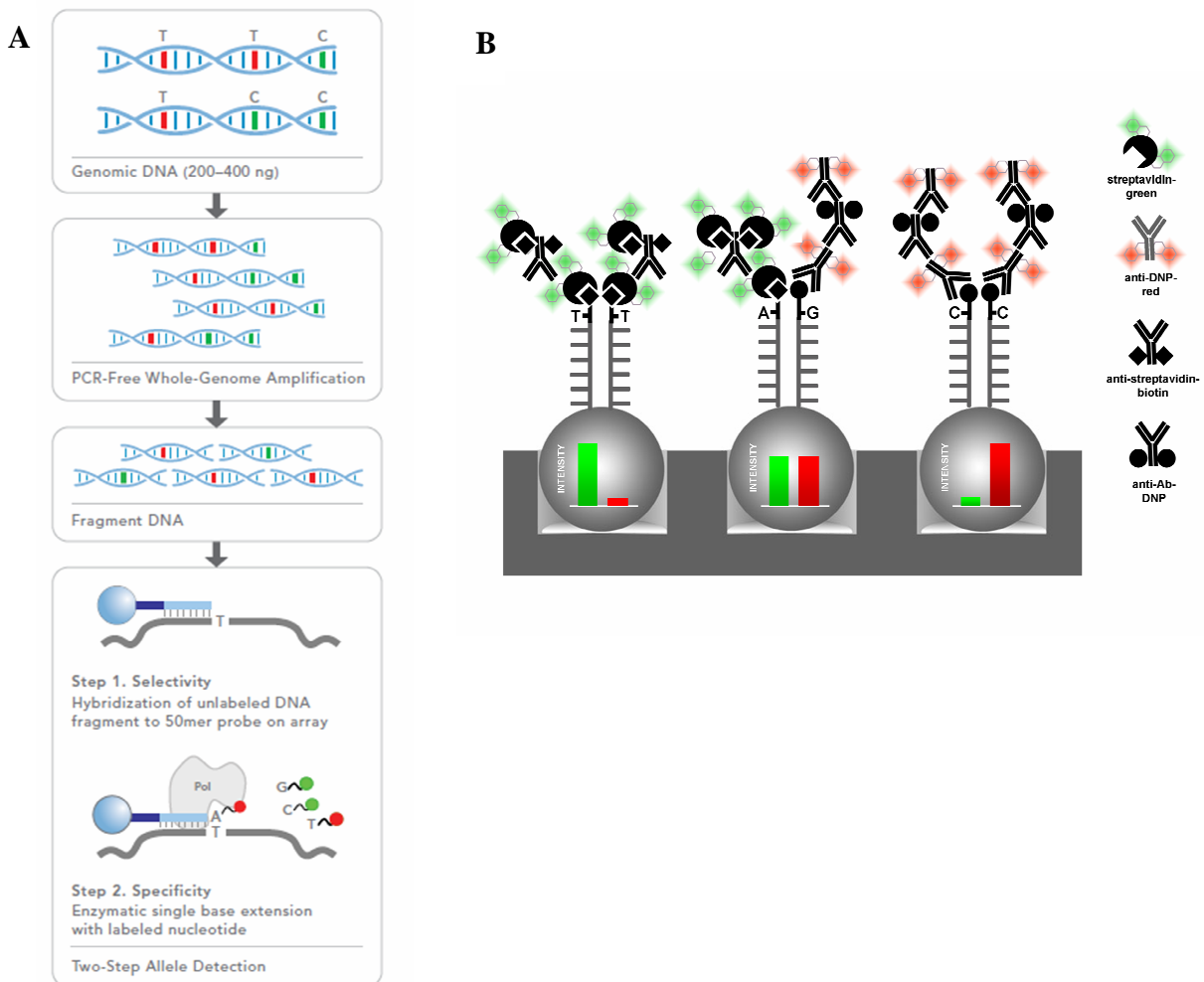


Figure 12 Illumina Infinium protocol - adapted from www.illumina.com

A Protocol from genome amplification to probe hybridization and primer extension **B** Immunohistochemical fluorescence detection

Based on the SNP genotype data, deletions were detected in the Hapmap population using Mendelian inconsistencies between trios⁸¹ and using clusters of genotyping errors or regions of SNPs that were not in Hardy-Weinberg equilibrium⁸². Although SNP genotyping arrays were not primarily designed for CNV analysis (only to call the three possible genotypes of SNPs), it is possible to obtain information on the copy number state by combining the intensities of the two alleles for a given SNP. This is similar to CGH analysis, where the comparison is made between the hybridization ratios from a test and reference sample. With SNP arrays, only one genome is hybridized per chip (Figure 9B) thus the reference can be a pool of external experiments and the copy number ratio can be computed as follow:

$$CNratio = \log_2 \left(\frac{S_A + S_B}{R_A + R_B} \right)$$

Where S refers to the intensity of the test sample (of an individual) and R to the (mean) intensity of the reference panel; A and B refer to the SNP alleles.

Subsequently similarly to CGH analysis, copy number changes can be identified by identifying significant deviation from the baseline CN ratio (Figure 9B). An alternative approach is to also consider the ratio of allele intensities. This ratio is close to 1 or 0 for homozygous SNP (AA or BB) and close to 0.5 for heterozygous (AB) (see Figure 11B). Identification of pattern from allelic intensities ratio helps to refine the CNV analysis. For example loss of heterozygosity (LOH) can reflect hemizygous deletion (if the copy number ratio is less than 0) or duplication with LOH (if the copy number ratio is >0). Clusters of noisy allelic ratios (i.e. aberrant fluctuation from 0 to 1) with very low CNratio can reflect homozygous deletion whereas allelic ratios ranging aberrantly from 0.3 to 0.6 with high CNratio can reflect allelic imbalance due to amplification. It has been demonstrated that incorporation of allelic ratios provides a powerful approach for the detection of CNVs both for tumor analysis^{83,84} and diploid sample analysis⁸⁵⁻⁸⁷. Figure 13 illustrates an analysis with a metastatic melanoma. The short arm of chromosome 7 (chr7p) was predicted as 3 copies (CN ratios were greater than 0 and with LOH) whereas chr7q was predicted with 5 copies.

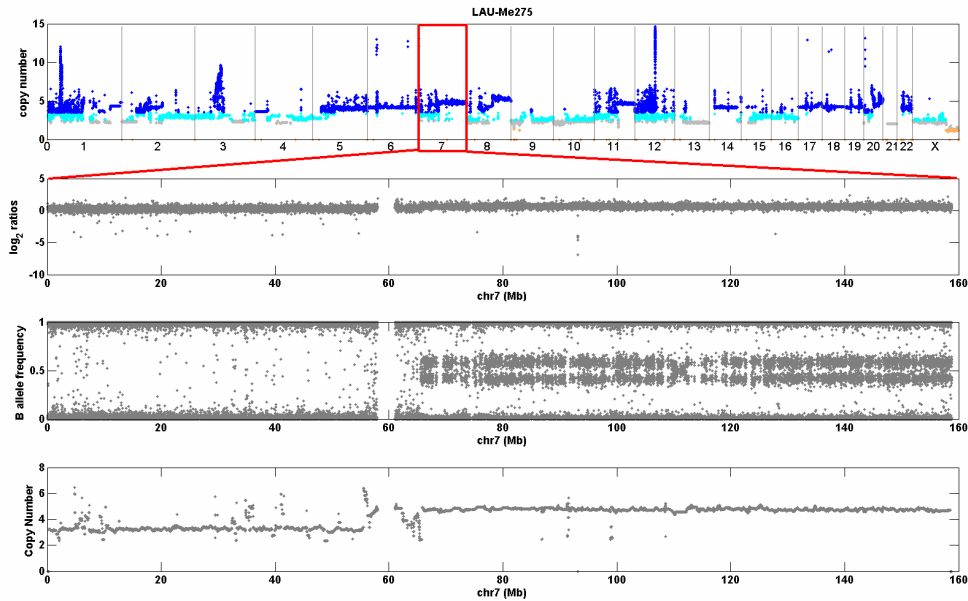


Figure 13 SNP array analysis

DNA from a metastatic melanoma (LAU-Me275) was hybridized to Illumina 1M SNP arrays. The top panel shows genome-wide copy number: dark blue indicates more than three copies; cyan: three copies; gray: copy neutral; orange : deletion. Subsequent panels show chromosome 7 with, from top to bottom: Hybridization log₂ ratio; B allele frequency; and copy number prediction.

Although CNVs can be mined from SNP array data, the analysis suffers from two major limitations: 1) the measured intensities are very noisy therefore additional normalization steps (such as GC-correction^{88,89}) are needed to improve the signal to noise ratio but can mask small copy number changes and 2) the uneven SNP density on the array challenges the CNV detection. Repeat-rich regions and regions within or close to segmental duplications are not covered on the array. These genomic regions are highly dynamic (prone to induce rearrangements) and thus likely to contain interesting CNVs. Also many interesting genes, recently associated with complex disease are not assayed.

To overcome this density limitation, Affymetrix in collaboration with the Broad Institute designed a new array combining both SNPs and *non-polymorphic* probes to cover CNV regions²⁰. The Affymetrix array (6.0) contains 906,000 SNPs and 946,000 *non-polymorphic* probes, with a median inter-marker distance less than 700 bases. Illumina provides a similar solution with the Illumina 1M-duo SNP array that has 1.2 million markers and a median inter-marker distance equals to 1.5kb. A new Illumina chip will be soon available with 2.5 million markers and inter-marker distance close to 630 bases.

1.3.2 Sequencing-based methods

Initial application with paired-end mapping approaches

With the availability of reference genomes^{90,91} discovery of structural variants became possible. However these genomes still contain genomic gaps and assembling of complex regions (near segmental duplications, repeats etc...) is not perfect. However better insights were provided with the development of newer sequencing generations and sequencing of additional individuals. Notably in 2005, Evan Eichler's group identified sequence variants by comparing fosmid end pairs of an individual to the reference genome sequence⁴. The principle is as follow: 1) the genomic sequence is fragmented and cloned into fosmid. 2) ends of the cloned sequence are sequenced using universal primers and are aligned to the reference genome 3) end pairs that are discordant in length or direction indicate respectively possible indels or inversion (Figure 14). Similar analyses were repeated with higher resolution in 2007⁹² and in 2008 with more genomes sequenced⁶⁴.

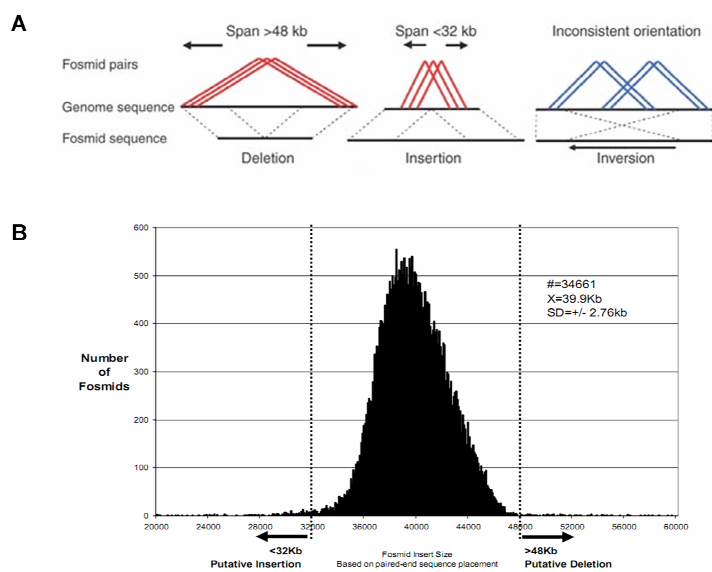


Figure 14 Detection of structural variation from fosmid end pair sequencing - adapted from⁴

A Methodology to define deletion, insertion from distance between fosmid end pairs and inversion from the orientation of fosmid end pairs **B** Distribution of fosmid insert size on chromosome 7.

Since, next generation sequencing (NGS) technologies have further evolved and are able to sequence millions of reads in parallel. The latest Illumina platform: HiSeq2000 is able to produce 200 GB of sequence with 100bp paired reads in less than 10 days. The massive amount of generated data has greatly stimulated the development of bioinformatics methods for structural variant analysis and now several approaches complementary to paired-end mapping have been proposed. These approaches include read-depth analysis, split-read strategies and sequence assembly comparisons. References to freely available tools are given in Table 2.

Methods	Algorithm	Reference
Paired-end mapping	Detection of discordant end-pairs	4,93,94
	Clustering of end-pairs	64,92,95-97
Read-depth analysis	Detection of local change points	98,99
	Detection of outliers compared to the read-depth baseline	100
	Event-wise testing	101
Split-read analysis	Identification of breakpoints with a pattern growth algorithm	102
Sequence assembly analysis	De novo assembly and comparison to reference genome	103,104
	Detection of small indels through local reassembly	105

Table 2 Algorithms for the detection of indels from NGS data

Read-depth and read-split approaches

The read-depth analysis investigates drop or increase of read coverage compared to an expected depth distribution (Figure 15). Mutual information about paired reads is used to improve the mapping quality and to detect complex and large rearrangements. However the analysis is challenged by repeat-rich regions due to mapping issues, and the breakpoint position is not precise. This contrasts with paired-end mapping (PEM) approaches, previously described, that enable precise breakpoint determination and perform well even in the presence of repetitive elements (LINE, SINE). However PEM approaches fail when both fosmid- or paired-ends map within repeats. Also the detection resolution is limited to the distance between end-pairs, therefore large (except deletion) or very small rearrangements cannot be detected. The split-read strategy entails in gapped-alignment of reads onto candidate breakpoints (Figure 15). The strategy used in the Pindel algorithm¹⁰² is to

detect paired-reads where only one end is uniquely mapped onto a reference genome. The assumption is that the second paired-read cannot be mapped, even with few mismatches allowed, because it corresponds to a deletion or insertion breakpoint. The mapped read is used as an anchor and knowing both a maximum event length and the direction to search for the unmapped read; alignment of the unmapped read can be performed either by splitting it in two (indicating a deletion event) or in three (indicating an insertion) fragments (Figure 15).

Providing a high sequencing depth, *denovo* assembly can be attempted using now standard tools like SOAPdenovo¹⁰³ or ABYSS¹⁰⁴. Once the genome has been assembled, sequence comparison can be made with the reference genome to identify deletions and insertions. The advantage of *denovo* assembling over PEM approaches is that deletion or insertion smaller than the paired-end insert size can be detected. But on the other hand, *denovo* assembling is only possible with high read-depth. When this criteria is not met, several experiments can be pooled together¹⁰⁶. Also *denovo* assembling is very difficult for repeat-rich regions.

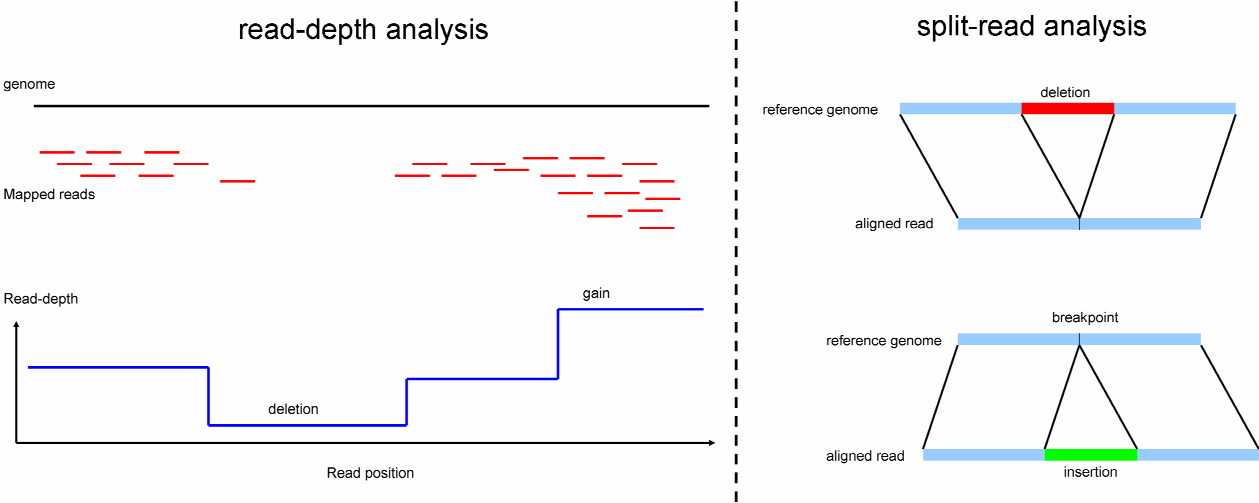


Figure 15 Read-depth and split-read analysis

The above techniques present different and complementary advantages, using several approaches in combination definitely empower the detection of different structural variations.

NGS offers several advantages over CGH and SNP arrays, in particular it allows detection of very small variants (indels, SNPs) and inversion. It can estimate exact breakpoint location and does not suffer from hybridization saturation allowing a better estimation of high copy numbers. Nevertheless the technique remains very expensive, the storage and computational analytical requirements are high and the computational methods still predict a significant fraction of false positives¹⁰⁶⁻¹⁰⁸, which can be controlled by using multiple algorithms^{106,109}.

1.3.3 Methods used in CNV validation

Validation constitutes an integral part of any study, especially in medical genetics where findings need to be validated (confirmed) and replicated (assessed in additional samples). In CNV studies, qualitative or quantitative estimation of copy number at targeted loci is crucial 1) to validate predictions from high-throughput platforms such as microarrays, 2) to estimate the true underlying copy number of a CNV region, 3) to fine-map the CNV boundaries and 4) to screen larger sample collections at a given locus. Classical molecular techniques include FISH, Southern blotting and long-range PCR. FISH was described previously. Southern blotting uses a labelled probe to bind specific sequence in genomic DNA. This DNA has been first digested with restriction enzyme, migrated onto electrophoresis gel and transferred to a membrane. The band intensity indicates the amount of DNA present. This technique is work-demanding, time consuming and cannot reveal mosaic changes. Long-range PCR enables the amplification of long DNA fragments (up to 50kb) but its protocol is highly demanding and involves a combination of modifications to standard conditions with a two-polymerase system. Because these three techniques require a high amount of work and offer a very limited throughput, these are not methods of choice for CNV validation. Instead other techniques with multiplexing capacities like quantitative real-time PCR, multiplex ligation-dependent probe amplification and multiple amplifiable probe hybridization are preferred, because of higher throughput possibilities.

Quantitative real-time PCR

Quantitative real-time PCR (qPCR) is the most commonly used approach for CNV validation ¹¹⁰. It is a reliable method for the detection of deletion and duplication at single loci and can be applicable to a large number of samples. However it is not suitable for precise copy number determination. During the qPCR experiment, the DNA amplification is monitored in real-time as a function of PCR cycles. Detection can be done using fluorescence dye (e.g. SYBR green) that emits fluorescence when incorporated in double stranded DNA. The fluorescence of the reporter dye increases as the PCR product accumulates with each successive cycle of amplification. A limitation is the low specificity, the dye will bind to any double stranded DNA. An alternative, but more expensive protocol, is to use probes (e.g. Taqman probes) that complement specifically the target DNA. The probe is an oligonucleotide with a reporter dye attached to the 5' and a quencher dye attached to the 3'. While the reporter is close to the quencher dye, only background fluorescence is measured. During the PCR, the probe complements the target DNA, the polymerase performs the primer extension and integrates the probe into the PCR product. Subsequently the 5' exonuclease activity of the polymerase cleaves the probe, which releases the reporter dye from the proximity of the quencher dye and enables the emission of the fluorescence signal (Figure 16). The process is repeated in every cycle and does not interfere with the accumulation of PCR product.

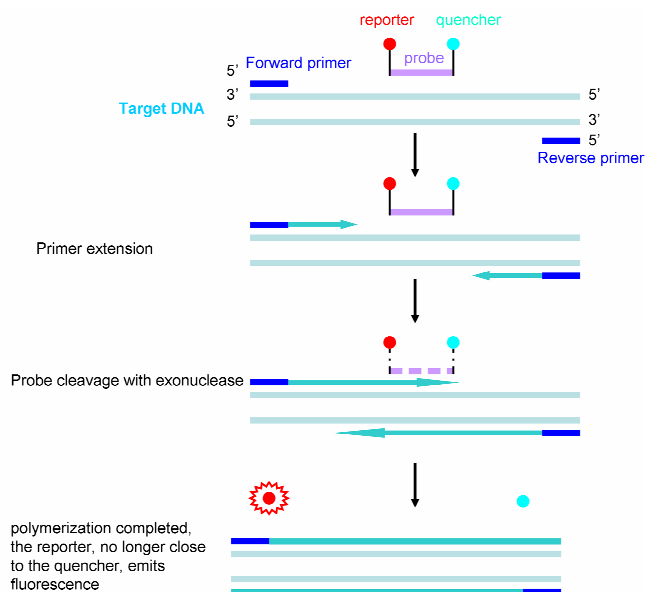


Figure 16 qPCR reaction with Taqman probes

Multiplex Quantitative Fluorescent Real-time PCR

Multiplex ligation-dependent probe amplification (MLPA) ¹¹¹ and multiple amplifiable probe hybridization (MAPH) ¹¹² provide an alternative to analyse simultaneously several genomic regions. MAPH is based on the hybridization of DNA probes to the target DNA, following stringent washing to only retain the hybridized probes, the probes are amplified using a common primer pair and quantified using capillary electrophoresis. MPLA is similar to MAPH, but requires the ligation of two adjacently hybridising oligonucleotides before a PCR product can be generated. Both techniques enable the interpretation of up to 50 loci in a single reaction and observation of mosaic copy number (with at least two consecutive probes). This has these techniques very attractive for CNV validation in a clinical setting (Figure 17).

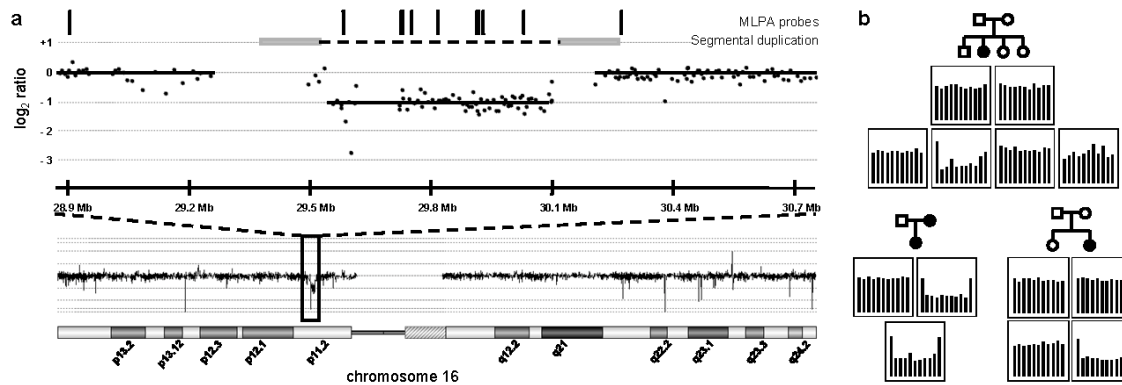


Figure 17 Identification of a deletion with CGH array and validation with MLPA - adapted from ¹¹³
a) CGH data showing the location of deletion on 16p11.2. The data show the log₂ intensity ratio for a deletion carrier compared to an undeleted control sample. Grey bars connected by a broken line denote the segmental duplication flanking the deletion region. Vertical bars indicate the positions of the probe pairs used for MLPA validation. **b)** MLPA validation of 16p11.2 deletions. Representative MLPA results are shown, illustrating one instance of maternal transmission and two instances of *de novo* deletions. Each panel shows the relative magnitude of the normalised, integrated signal at each probe location, in order of chromosomal position of the MLPA probe pairs as indicated in (a). Each panel corresponds to its respective position on the associated pedigree, as shown.

Newest high-throughput validation technologies

With the need for large-scale validation in SNP-based genome wide association studies, Illumina proposed to the scientific community custom SNP genotyping arrays to assay up to 200,000 SNPs associated to complex diseases (diabetes, obesity and cardiovascular disease), with a large throughput (24-sample format) and at reduced cost compared to whole genome genotyping arrays. Such custom-chips have been designed to screen candidate SNPs in metabolic and immune disease (referred as metabo- and immuno-chip).

A novel platform, called nanostring, has been recently proposed for multiplexed validation of CNV regions (up to 200 regions in a single reaction). This technology counts the number of probes carrying a colour-coded barcode that have hybridized to specific targeted regions (Figure 18). Initially developed for gene expression analysis¹¹⁴, this technology has been rapidly extended to copy number analysis. This technology enables sensitive detection of copy number because it does not rely on enzymes or amplification step and because the hybridization is performed in solution. According to the manufacturers specifications, it can detect copy number down to 2kb in size with a copy number range from 0 to 4. The latter implies the technology is of little interest in cancer genomics, where copy number within (focal) amplifications can reach values greater than 20. However because of the throughput (12 samples prepared and analysed in less than seven hours), the technology will prove useful for analysis of variation in the general population or disease cohorts.

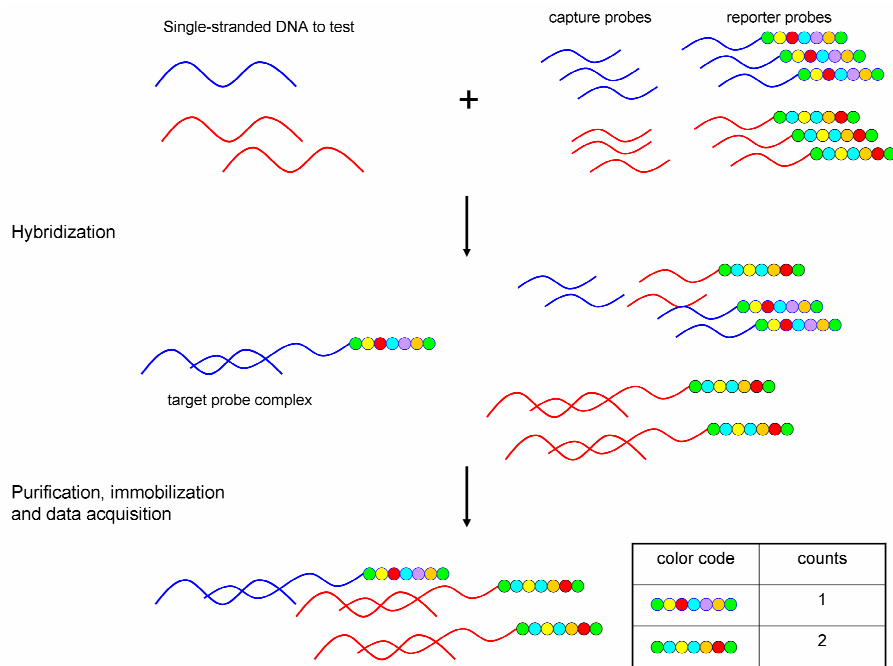


Figure 18 Nanostring technology overview

Both the capture and reporter probes will hybridize to a target single-stranded DNA and form a complex. After hybridization, probe excess is removed and complexes can be immobilized using the capture probe. Then image acquisition can be done using the colour-coded barcode attached to the reporter probe. Counts of barcode allow estimating the number of DNA that was captured.

1.4 The clinical impact of copy number variants

1.4.1 CNVs and genomic disorder

From their initial detection, CNVs were thought to lead to disease. The literature is enriched in examples of microduplication and microdeletion linked to genetic disorders. Some examples are listed in Table 3, the contribution of these CNVs is not yet fully understood and thus further studies are still ongoing to fine-map the CNV breakpoints and to detect the affected genes. Cataloguing candidate CNVs involved with disease is part of an international effort, DECIPHER¹¹⁵. This database aims at centralizing, providing information and facilitating data exchange between clinical laboratories. It also provides a number of bioinformatic tools to visualize genomic data (i.e. using the Ensembl browser) and to prioritize list of candidate genes using text mining techniques. At the end of November 2010, this database described 58 syndromes and more than 7000 patients. It constitutes an important step towards a better comprehension of the clinical impact of CNVs.

By contrast to the multitude (>700) of published SNP-based genome-wide association studies (GWAs), CNV-based GWAs have had a limited success so far¹¹⁶⁻¹²⁰. It should be noted that the results from Blauw et al. did not meet the authors'

criteria for replication. Wain et al. potentially identified four variants involved in sporadic amyotrophic lateral sclerosis, among which *EEF1D* was previously unknown. However the authors warn against the modest size effects detected and the suitability of the genotyping platform used. Craddock et al.¹¹⁷ concluded that it was unlikely that CNVs could explain the missing heritability in complex and common disease. However, the importance of rare CNVs emerged very recently with a few GWA¹²¹⁻¹²³ and several non-GWA studies^{113,124-128}.

Disease	Locus	Gene(s)	CNV	Reference(s)
Rare genomic disorders				
Cri du chat syndrome	5p15.2	Multiple genes	Deletion	129
Spinal muscular atrophy (SMA1-4)	5q12.2-q13.3	BIRC1, GTF2H2, SERF1A, SERF1B, SMN1, SMN2	Deletion	130
Williams-Beuren syndrome	7q11.23	Multiple genes	Deletion/duplication	131
CHARGE syndrome	8q12.1	CHD7	Deletion	132
Charcot-Marie-Tooth disease type 4B2	11p15.4	ADM, SBF2	Deletion	133
Prader-Willi and Angelman syndrome	15q11-q13	ATP10A, OCA2, OR4M2, OR4N4, UBE3A	Deletion	134
Smith-Magenis syndrome	17p11.2	ATPF2, COPS3, DRG2, MED9, NT5M, RAI1, SMCR8, SREBF1	Deletion	38
Charcot-Marie-Tooth disease type 1A	17p11.2	COX10, HS3ST3A1, PMP22, TEKT3, ZNF286	Deletion	49
Neurofibromatosis type 1	17q11.2	NF1	Deletion/duplication	135
Miller-Dieker lissencephaly syndrome	17p13.3	LIS1	Deletion	136
DiGeorge/Velocardiofacial syndrome	22q11.2	GGT2, GNB1L, HIC2, TBX1	Deletion/duplication	137
Pelizaeus-Merzbacher disease	Xq22	PLP1	Deletion/duplication	138
Common disorders				
HIV/AIDS susceptibility	17q11.2	CCL3L1	Deletion	35
Systemic lupus erythematosus	1q23	FCGR3B	Deletion	139-141
	6p21.3	C4	Deletion	142
Rheumatoid arthritis	17q11.2	CCL3L1	Duplication	143
Kawasaki disease	17q11.2	CCL3L1	Duplication	144
Crohn's disease	8p23.1	HBD-2	Deletion	145
	5q33.1	IRGM	Deletion	146
Psoriasis	8p23.1	HBD-2	Duplication	147

ANCA-associated vasculitis	1q23	FCGR3B	Deletion	140
	1q23	FCGR3B	Duplication	141
Atopic asthma	1q13.3	GSTM1	Deletion	148
	22q11.2	GSTT1	Deletion	149,150
Autism spectrum disorders	15q11-q13	Multiple genes	Deletion/duplication	151
	16p11.2	Multiple genes	Deletion/duplication	152
	22q13.3	SHANK3	Deletion	153
	Xp22.33	NLGN4	Deletion	154
	2p16.3	NRXN1	Deletion	155
Schizophrenia	2q34	ERBB4	Deletion	156
	5p13.3	SLC1A3	Deletion	157
	2q31.2	RAPGEF4	Deletion	158
	12.24	CIT	Deletion	159
Epilepsy	15q13.3	CHRNA7	Deletion	157,159
Parkinson's disease	4q22	SNCA	Duplication	160,161
Amyotrophic lateral sclerosis	5q12.2-q13.3	SMN1, SMN2	Deletion	162
Familial hypercholesterolemia	19p13.2	LDLR	Deletion/duplication	163

Table 3 CNV reported associated to rare and common disease adapted from ¹⁶⁴

1.4.2 CNVs and gene expression in the general population

Analogously to SNP-based analyses ¹⁶⁵⁻¹⁶⁸, efforts have focused on understanding whether CNVs could influence gene dosage in an individual from the general population. This is important both for our understanding of the predisposition to disease but also to understand “normal” phenotypic variation between individuals. Indeed, it has been demonstrated that both the copy number and position of CNVs affect the expression of nearby genes ^{12,169,170}. Scenarios for the effect of deletions and duplications are illustrated in Figure 19 and Figure 20, respectively. In addition, there has been growing evidence that CNVs could play a role in tissue-specific developmental constraints ^{12,15,171,172}. However the contribution of CNVs during development (i.e. impact on gene expression during morphogenesis) remains unknown.

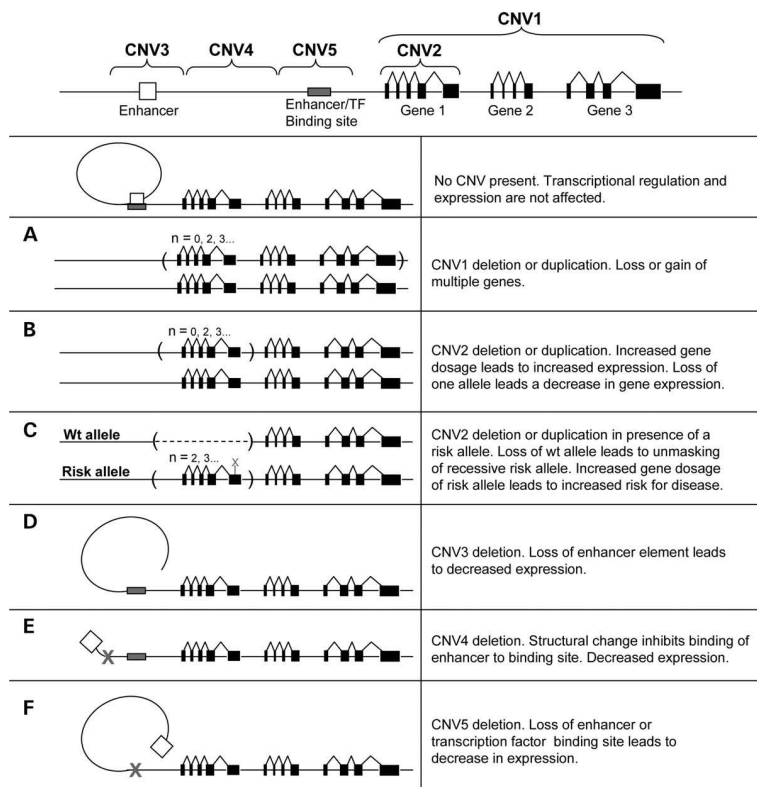


Figure 19 CNV influencing gene dosage and expression and disease - from ¹⁷³

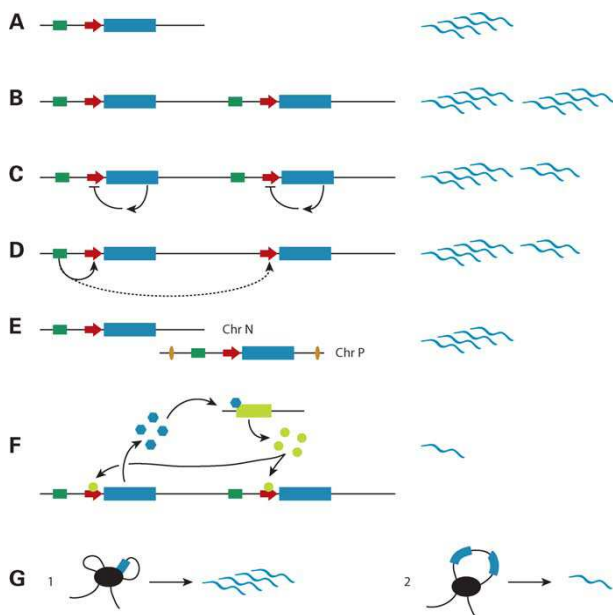


Figure 20 Duplication scenarios and their influence on gene expression - from ¹⁷⁴

(A) Single copy gene locus. The gene intron–exon region (blue box), the gene promoter (red arrow) and its enhancer (green box) are shown. Transcript levels are indicated schematically on the right. (B) Complete tandem duplication including the regulatory region. (C) Complete tandem duplication including the regulatory region of a gene under a compensatory mechanism. A negative feedback loop reduces the gene transcription level. (D) Complete tandem duplication excluding the regulatory region. The single enhancer only weakly influences expression of the second copy of the gene, which is expressed at a lower level. (E) Complete non-tandem duplication including the regulatory region. The duplicated locus maps to another chromosome region where a different chromatin context, e.g. insulators (yellow ellipses), modifies its expression level. (F) Immediate early gene model. In the presence of a duplication, the concentration of CNV gene product (blue hexagons) is sufficient to induce a repressor (light green box), the product of which (light green disks) blocks the expression of the CNV gene. (G) A tandem duplication (2) physically impairs the access of the CNV genes copies to the transcription factory where it should be transcribed (1).

1.5 Somatic copy number alterations in cancer

Somatic copy number alterations (SCNA) are typical of cancer cells^{76,175-180}. It has been demonstrated that SCNAs near oncogenes or tumour suppressor genes can affect gene expression levels or result in the expression of chimeric fusion genes^{77,179}. Alteration of the gene dosage and subsequent perturbation of pathways involved in cell proliferation, senescence or death has been shown in many cases to contribute significantly to the progression from the normal to the malignant state^{179,181-184}.

Challenges in cancer genomics differ greatly from medical genetics where patients or individuals are assumed to have a “mostly-diploid” genome (with the exception of specific trisomies like the Down syndrome). Cancer genomes accumulate point mutations and SCNAs which conspire to disrupt gene expression and the interplay between signalling pathways that control normal growth and tissue homeostasis^{99,107,108,175-177}. Only a small fraction among these aberrations contributes to the development from the normal to the malignant state. Such aberrations are referred as “cancer drivers” and should be distinguished from “passenger events”, which are the consequence of the stochastic accumulation of aberrations^{179,181-183}. The identification of potential cancer driver genes can be done by screening tumor collections for recurrent mutations or SCNAs^{177,185-189}. Then validation of the oncogenic (or tumor suppressor) function is usually achieved by following tumor progression when activating or repressing the gene in tumor models¹⁹⁰⁻¹⁹². Identification of cancer driver genes is important to understand the initial mechanisms leading to cancer, establish early diagnostic tests and possibly develop novel therapies.

The analysis of passenger mutations is very challenging, because in cancer genomes with deficiencies in DSB repair or mismatch repair pathways, there will be thousands of accumulated aberrations. Most of these aberrations will not give an advantage to the cancer cell (i.e. most of these will occur in intronic or intergenic regions, or be synonymous mutations). However several will occur in coding sequences, can result in translational frameshifts or stop codons; and be the consequence of premature protein termination or non-functional proteins. Even if these events are not initially responsible for the tumor development; there is still the

possibility that a small fraction can favour further tumor progression. In fact, once the cell has accumulated mutations that enable it to evade normal growth control (and enter a proliferative phase), stochastic accumulation of mutations (or SCNAs) can affect genes that will help the cell to proceed to an invasive phase. This scenario can be illustrated with melanomas, where the proliferative phase starts with the activation of the MAPK pathway but is not a sufficient event for malignant transformation¹⁹³⁻¹⁹⁵. Details about the progression of melanoma are given in the next section (Scope of the thesis). Following malignant transformation, cells continue to evolve and new “features” can be acquired from later mutations. These new features can create new sub-populations of cancer cells that can be resistant to treatment and lead to relapse. For example, melanogenesis is a metabolic pathway that converts L-tyrosine to melanin pigment. Melanin provides protection against UV radiation in normal melanocytes, but melanogenesis products generate immunosuppressive and mutagenic environments within the cell and can confer resistance to chemotherapy and radiotherapy in malignant melanomas^{196,197}. Consequently, inhibition of melanogenesis, which is not a driving event in melanoma malignancy, is under active research for treating melanoma^{196,198-200}.

To summarize, the comprehensive documentation of SCNAs and mutations with possible impact on gene expression and subsequent pathway regulation, will be useful to improve our knowledge of cancer driving events (giving rise to the initial malignant state), passenger events with possible contribution to malignancy and passenger events without contribution to the tumorigenesis.

1.6 Scope of the thesis

The aim of my thesis is to develop computational methods to detect CNVs from microarray data, and to investigate their possible clinical impact both in complex disease and cancer. This theme has been studied in the three following sections.

1.6.1 Identification and validation of Copy Number Variants using SNP genotyping arrays from a large clinical cohort

In the first section (chapter 3), I aim to mine CNVs from a population-based medical cohort named CoLaus. CoLaus: Cohorte Lausannoise includes more than 6,000 individuals, with age 35-75, from the Lausanne area. All individuals were genotyped on Affymetrix 500K SNP arrays, and extensive phenotypic data were collected by the Lausanne University Hospital (CHUV). These phenotypes include more than 150 measurements from anthropomorphic traits (height, weight, age, gender...) to metabolic measurements (cholesterol, glucose, HDL, LDL levels); as well as questionnaires (smoker-status, drugs prescription etc...). The initial goal of the study was to perform SNP-based genome-wide association studies to investigate cardiovascular predisposition. This was carried out by Dr. Zoltan Kutalik and Dr. Toby Johnson, both postdocs in Pr. Sven Bergman's group. At this time, the predisposition to disease due to CNVs was unknown and methods to analyse SNP arrays were limited to very few algorithms such as CNAT, GEMCA, dCHIP and CNAG²⁰¹⁻²⁰⁴. All but CNAT were limited to the Microsoft Windows operating system and could not scale with the analysis of a high number of samples. Other available methodologies were those used for CGH array analysis. These methods were not using allelic intensity ratios, a feature unique to SNP arrays that improves the CNV detection power⁸³. In addition, these methods were suitable for the analysis of BAC CGH arrays (with less than 30,000 clones) but could not scale to the analysis of half a million markers. Finally, there was no established gold standard for the comparison of different CNV methods.

In chapter 3, I present my work on the development of a novel method: GMM, which relies on Gaussian Mixture Modelling of copy number in the whole Lausanne population. I compare its sensitivity and specificity to three other CNV detection methods, notably by investigating the concordance in predicting CNVs in a *sub-sample* of individuals that were genotyped on the Illumina platform. I also describe

two merging strategies, which were applied to create a map of CNV regions and I devise a novel method to investigate the performance of different CNV detection methods using relatedness between individuals.

1.6.2 Aetiology of CNVs in complex disease

In this section, I describe my involvement in a collaborative effort on morbid obesity. This project was an international collaboration led by the CHUV and the Imperial College London. The aim was to investigate the penetrance of a 16p11 deletion detected in patients affected with developmental delays and/or obesity.

Obesity is a major problem in modern societies, in US more than 300,000 deaths per year can be attributed to obesity²⁰⁵. In 2005, more than 1.6 billion adults (with age greater than 15) were over-weights and 400 million adults were obese. According to the latest WHO projections for 2015, there will be 2.3 billion over-weights adults and more 700 million obese. Obesity has severe consequences on health such as increased risk for cardiovascular disease and type 2 diabetes^{206,207}. Lack of exercise and overweight can account for a third of cancers of the breast, colon, endometrial, kidney and oesophagus²⁰⁸⁻²¹⁰. It has been shown that obesity reduces fertility²¹¹ and life expectancy^{212,213}. Obesity is both a common and complex disease and its heritability is not yet fully understood^{214,215}. This may be due to the fact there is a strong environmental effect²¹⁶ and that the disease is involved with both genomic disorders (for e.g. Prader-Willi, Bardet-Biel and Cohen syndrome²¹⁷) and monogenic disorders. Examples of monogenic disorders include mutations in the melanocortin-4 receptor (*MC4R*)²¹⁸⁻²²⁰ and deficiencies in leptin^{221,222}, prohormone convertase-1 gene²²³ and proopiomelanocortin²²⁴. Both linkage analysis^{225,226} and SNP-based GWAs²²⁷⁻²³⁰ have been performed to understand the heritability and predisposition of the disease. These studies confirmed several genes such as *MC4R* and identified novel ones such as the *FTO* and *SH2B1* genes. However the implication of CNVs with obesity was not clear until recently^{113,125}.

In this chapter, I make a direct use of my map of variation in the Swiss population (see chapter 3) to assess the relevance and penetrance of a rare deletion detected in obese patients. Since the project was an international collaborative effort which led to publication in a high impact journal, my contributions are fully detailed.

1.6.3 Detection and impact of somatic copy number alterations in cancer

In chapter 3 and 4, I focused on the detection of CNVs in the general population, thus genomes that are assumed to be diploid. In this section, I was interested in the implications of somatic copy number aberrations (SCNAs) in highly aneuploid genomes.

This project was part of collaboration between the Ludwig Institute for Cancer Research, Swiss Institute of Bioinformatics, CHUV and universities of Lausanne and Geneva. The global aim of the project was to perform a comprehensive molecular profiling of seven metastatic melanoma cell lines, with matched donor controls using exome and transcriptome sequencing, methylation arrays, CGH and SNP arrays and karyotyping. I was strongly involved in all aspects of this collaboration (from the experiment design to data analysis and interpretation); my contributions are fully explained in chapter 5.

Melanoma was chosen for three reasons 1) the Ludwig Institute has both a demonstrated expertise in melanoma research and a very large sample collection that includes primary tumours and cell lines derived from metastases. Clinical records and matched donor controls are also available. 2) Melanoma is an highly aggressive form of cancer that leads to regional and distal metastases. Malignant melanomas are resistant to both *radio-* and *chemo-*therapy; and constitute the most lethal form of skin cancer (accounting for 80% of deaths), metastases are fatal within 5 years; only early diagnosis and surgical removal can provide a cure for the patient²³¹. 3) The genomes of melanoma undergo many re-arrangements^{232,233} which challenges the analysis and identification of novel genes that can be relevant to tumor progression. Therefore the development of robust computational methods, the thorough and comprehensive documentation of somatic aberrations and any novel hypothesis generated with the results will not only be of benefit to melanoma scientists and clinicians, but could potentially be of interest to the whole cancer community.

In melanoma the transformation from normal skin melanocytes to a proliferative cell is the result of somatic mutation in *BRAF* or *NRAS*, respectively in 50% and 20% of the cases. These mutations are mutually exclusive and lead to constitutive activation of serine–threonine kinases in the ERK–MAPK pathway²³⁴⁻²³⁶. Loss of tumor

suppressors *CDKN2A*^{237,238} and *PTEN*^{181,239} play an important role in melanoma tumorigenesis. *CDKN2A* encodes for *INK4A* (p16INK4A), a protein that blocks the cell cycle at the G1–S checkpoint by inhibition of *CDK4*, a cyclin dependent kinase. *PTEN* when expressed normally, down-regulates levels of phosphatidylinositol phosphate (*PIP3*), a growth factor which controls the activation of protein kinase B (*PKB*, also called *AKT*). Activation of *AKT* represses cell cycle inhibitors and inactivates apoptosis inducers. In the absence of *PTEN*, *PIP3* levels increase which in turn activates *AKT*. This *AKT* activation prolongs cell survival by repressing apoptosis and stimulate cell proliferation (for e.g. by increasing *CCND1* expression). Further tumor progression can be associated with decreased differentiation and decreased expression of melanoma markers regulated by *MITF*^{231,240}. Progression to the vertical-growth phase (invasion) and to subsequent metastases result from changes in cell adhesion (with perturbation of the cadherin, WNT and integrin signalling pathways)²⁴¹⁻²⁴³.

Despite these numerous candidate studies that have established the basis of melanoma development and progression, there are relatively few genome-wide analyses, compared to other cancers like breast cancer. Back in 2007, representative examples included an SCNA study²⁴⁴ that documented recurrent events in a large sample collection (>70) and a gene expression-based study that looked at pathways potentially perturbed with differentially expressed genes²⁴⁵. Only recently new studies were published: a second SCNA-based study²⁴⁶; a study that investigated gene fusion from RNA-seq data²⁴⁷ and another one which looked at somatic mutations from full genome sequencing¹⁸⁶.

In chapter 5, I describe my work on the detection of SCNAs in metastatic melanoma, I study the link between SCNAs and aberrant gene expression signature, and then I investigate commonalities between our samples and test my findings with external data²⁴⁴⁻²⁴⁶ both at the gene and pathway level.

1.7 References

1. Iafrate, A.J. et al. Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-51 (2004).
2. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat Rev Genet* **7**, 85-97 (2006).
3. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).
4. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-32 (2005).
5. Freeman, J.L. et al. Copy number variation: new insights in genome diversity. *Genome Res* **16**, 949-61 (2006).
6. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-8 (2004).
7. Jakobsson, M. et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998-1003 (2008).
8. Sharp, A.J. et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78-88 (2005).
9. Perry, G.H. et al. Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci U S A* **103**, 8006-11 (2006).
10. Perry, G.H. et al. Copy number variation and evolution in humans and chimpanzees. *Genome Res* **18**, 1698-710 (2008).
11. Lee, A.S. et al. Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet* **17**, 1127-36 (2008).
12. Henrichsen, C.N. et al. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet* **41**, 424-9 (2009).
13. Graubert, T.A. et al. A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* **3**, e3 (2007).
14. Guryev, V. et al. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet* **40**, 538-45 (2008).
15. Dopman, E.B. & Hartl, D.L. A portrait of copy-number polymorphism in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* **104**, 19920-5 (2007).
16. Fontanesi, L. et al. An initial comparative map of copy number variations in the goat (*Capra hircus*) genome. *BMC Genomics* **11**, 639 (2010).
17. Liu, G.E. et al. Analysis of copy number variations among diverse cattle breeds. *Genome Res* **20**, 693-703 (2010).
18. Lejeune, J., Gautier, M. & Turpin, R.A. Etude des chromosomes somatiques de neuf enfants mongoliens. *Comptes rendus de l'Académie des Sciences* **248**, 1721-1722 (1959).
19. Kallioniemi, A. et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818-21 (1992).
20. McCarroll, S.A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166-74 (2008).
21. Gayan, J. et al. Genetic structure of the Spanish population. *BMC Genomics* **11**, 326 (2010).
22. McElroy, J.P., Nelson, M.R., Caillier, S.J. & Oksenberg, J.R. Copy number variation in African Americans. *BMC Genet* **10**, 15 (2009).

23. Matsuzaki, H., Wang, P.H., Hu, J., Rava, R. & Fu, G.K. High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. *Genome Biol* **10**, R125 (2009).
24. Lin, C.H. et al. A large-scale survey of genetic copy number variations among Han Chinese residing in Taiwan. *BMC Genet* **9**, 92 (2008).
25. Takahashi, N. et al. Segmental copy-number variation observed in Japanese by array-CGH. *Ann Hum Genet* **72**, 193-204 (2008).
26. Kang, T.W. et al. Copy number variations (CNVs) identified in Korean individuals. *BMC Genomics* **9**, 492 (2008).
27. Jeon, J.P. et al. A comprehensive profile of DNA copy number variations in a Korean population: identification of copy number invariant regions among Koreans. *Exp Mol Med* **41**, 618-28 (2009).
28. Yim, S.H. et al. Copy number variations in East-Asian population and their evolutionary and functional implications. *Hum Mol Genet* **19**, 1001-8 (2010).
29. Li, J. et al. Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. *PLoS One* **4**, e7958 (2009).
30. Kato, M. et al. Population-genetic nature of copy number variations in the human genome. *Hum Mol Genet* **19**, 761-73 (2010).
31. Park, H. et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat Genet* **42**, 400-5 (2010).
32. Conrad, D.F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* (2009).
33. Ku, C.S. et al. Genomic copy number variations in three Southeast Asian populations. *Hum Mutat* **31**, 851-7 (2010).
34. Locke, D.P. et al. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res* **13**, 347-57 (2003).
35. Gonzalez, E. et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434-40 (2005).
36. Perry, G.H. et al. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**, 1256-60 (2007).
37. Hasin, Y. et al. High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS Genet* **4**, e1000249 (2008).
38. Lupski, J.R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14**, 417-22 (1998).
39. Fredman, D. et al. Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* **36**, 861-6 (2004).
40. Eichler, E.E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* **17**, 661-9 (2001).
41. The International HapMap Project. *Nature* **426**, 789-796 (2003).
42. Notini, A.J., Craig, J.M. & White, S.J. Copy number variation and mosaicism. *Cytogenet Genome Res* **123**, 270-7 (2008).
43. Kidd, J.M. et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**, 837-47 (2010).
44. Beck, C.R. et al. LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159-70 (2010).
45. Hastings, P.J., Lupski, J.R., Rosenberg, S.M. & Ira, G. Mechanisms of change in gene copy number. *Nat Rev Genet* **10**, 551-64 (2009).

46. Turner, D.J. et al. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet* **40**, 90-5 (2008).
47. Flores, M. Recurrent DNA inversion rearrangements in the human genome. *Proc Natl Acad Sci USA* **104**, 6099-6106 (2007).
48. Stankiewicz, P. & Lupski, J.R. Genome architecture, rearrangements and genomic disorders. *Trends Genet* **18**, 74-82 (2002).
49. Lupski, J.R. Charcot-Marie-Tooth disease: lessons in genetic mechanisms. *Mol Med* **4**, 3-11 (1998).
50. Stankiewicz, P. et al. Genome architecture catalyzes nonrecurrent chromosomal rearrangements. *Am J Hum Genet* **72**, 1101-16 (2003).
51. Matejas, V. et al. Identification of Alu elements mediating a partial PMP22 deletion. *Neurogenetics* **7**, 119-26 (2006).
52. Bauters, M. et al. Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. *Genome Res* **18**, 847-58 (2008).
53. Kim, P.M. et al. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res* **18**, 1865-74 (2008).
54. Lieber, M.R. The mechanism of human nonhomologous DNA end joining. *J Biol Chem* **283**, 1-5 (2008).
55. Inoue, K. et al. Genomic rearrangements resulting in PLP1 deletion occur by nonhomologous end joining and cause different dysmyelinating phenotypes in males and females. *Am J Hum Genet* **71**, 838-53 (2002).
56. O'Driscoll, M. & Jeggo, P.A. The role of double-strand break repair - insights from human genetics. *Nat Rev Genet* **7**, 45-54 (2006).
57. Lieber, M.R., Ma, Y., Pannicke, U. & Schwarz, K. Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol* **4**, 712-20 (2003).
58. Burma, S., Chen, B.P. & Chen, D.J. Role of non-homologous end joining (NHEJ) in maintaining genomic integrity. *DNA Repair (Amst)* **5**, 1042-8 (2006).
59. Smit, A.F. The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev* **6**, 743-8 (1996).
60. Goodier, J.L. & Kazazian, H.H., Jr. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**, 23-35 (2008).
61. Kazazian, H.H., Jr. & Moran, J.V. The impact of L1 retrotransposons on the human genome. *Nat Genet* **19**, 19-24 (1998).
62. Cordaux, R. & Batzer, M.A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703 (2009).
63. Zhang, F., Gu, W., Hurles, M.E. & Lupski, J.R. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**, 451-81 (2009).
64. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
65. Kazazian, H.H., Jr. et al. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164-6 (1988).
66. Bacolla, A. & Wells, R.D. Non-B DNA conformations, genomic rearrangements, and human disease. *J Biol Chem* **279**, 47411-4 (2004).
67. Bacolla, A. et al. Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc Natl Acad Sci U S A* **101**, 14162-7 (2004).

68. Lee, J.A., Carvalho, C.M. & Lupski, J.R. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**, 1235-47 (2007).
69. Inoue, K. et al. Proteolipid protein gene duplications causing Pelizaeus-Merzbacher disease: molecular mechanism and phenotypic manifestations. *Ann Neurol* **45**, 624-32 (1999).
70. Ylstra, B., van den Ijssel, P., Carvalho, B., Brakenhoff, R.H. & Meijer, G.A. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res* **34**, 445-50 (2006).
71. Carter, N.P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* **39**, S16-21 (2007).
72. Redon, R., Fitzgerald, T. & Carter, N.P. Comparative genomic hybridization: DNA labeling, hybridization and detection. *Methods Mol Biol* **529**, 267-78 (2009).
73. Redon, R., Rigler, D. & Carter, N.P. Comparative genomic hybridization: DNA preparation for microarray fabrication. *Methods Mol Biol* **529**, 259-66 (2009).
74. Fiegler, H. et al. Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res* **16**, 1566-74 (2006).
75. Kallioniemi, A. CGH microarrays and cancer. *Curr Opin Biotechnol* **19**, 36-40 (2008).
76. Bignell, G.R. et al. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res* **14**, 287-95 (2004).
77. Pinkel, D. & Albertson, D.G. Array comparative genomic hybridization and its applications in cancer. *Nat Genet* **37 Suppl**, S11-7 (2005).
78. Shaffer, L.G. & Bejjani, B.A. Medical applications of array CGH and the transformation of clinical cytogenetics. *Cytogenet Genome Res* **115**, 303-9 (2006).
79. Oostlander, A.E., Meijer, G.A. & Ylstra, B. Microarray-based comparative genomic hybridization and its applications in human genetics. *Clin Genet* **66**, 488-95 (2004).
80. Teo, Y.Y. et al. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23**, 2741-6 (2007).
81. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38**, 75-81 (2006).
82. McCarroll, S.A. et al. Common deletion polymorphisms in the human genome. *Nat Genet* **38**, 86-92 (2006).
83. LaFramboise, T. et al. Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Comput Biol* **1**, e65 (2005).
84. Attiyeh, E.F. et al. Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res* **19**, 276-83 (2009).
85. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-74 (2007).
86. Colella, S. et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* **35**, 2013-25 (2007).
87. Coin, L.J. et al. cnvHap: an integrative population and haplotype-based multiplatform model of SNPs and CNVs. *Nat Methods* **7**, 541-6 (2010).
88. Diskin, S.J. et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res* **36**, e126 (2008).

89. Marioni, J.C. et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol* **8**, R228 (2007).
90. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
91. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-51 (2001).
92. Korb, J.O. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420-6 (2007).
93. Chen, K. et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Meth* **6**, 677-681 (2009).
94. Korb, J. et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology* **10**, R23 (2009).
95. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
96. Lee, S., Hormozdiari, F., Alkan, C. & Brudno, M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat Meth* **6**, 473-474 (2009).
97. Hormozdiari, F., Alkan, C., Eichler, E.E. & Sahinalp, S.C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* **19**, 1270-8 (2009).
98. Chiang, D.Y. et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Meth* **6**, 99-103 (2009).
99. Campbell, P.J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-729 (2008).
100. Alkan, C. et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**, 1061-7 (2009).
101. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* **19**, 1586-92 (2009).
102. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-71 (2009).
103. Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**, 265-72 (2010).
104. Simpson, J.T. et al. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**, 1117-23 (2009).
105. Massouras, A. et al. Primer-initiated sequence synthesis to detect and assemble structural variants. *Nat Meth* **7**, 485-486 (2010).
106. Durbin, R.M. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
107. Pleasance, E.D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-6 (2010).
108. Pleasance, E.D. et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-90 (2010).
109. Sudmant, P.H. et al. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-6 (2010).
110. Higuchi, R., Dollinger, G., Walsh, P.S. & Griffith, R. Simultaneous amplification and detection of specific DNA sequences. *Biotechnology (N Y)* **10**, 413-7 (1992).

111. Schouten, J.P. et al. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* **30**, e57 (2002).
112. Armour, J.A., Sismani, C., Patsalis, P.C. & Cross, G. Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Res* **28**, 605-9 (2000).
113. Walters, R.G. et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* **463**, 671-5 (2010).
114. Geiss, G.K. et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol* **26**, 317-25 (2008).
115. Firth, H.V. et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* **84**, 524-33 (2009).
116. Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-72 (2010).
117. Craddock, N. et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713-20 (2010).
118. Wang, K. et al. A genome-wide association study on common SNPs and rare CNVs in anorexia nervosa. *Mol Psychiatry* (2010).
119. Blauw, H.M. et al. A large genome scan for rare CNVs in amyotrophic lateral sclerosis. *Hum Mol Genet* **19**, 4091-9 (2010).
120. Wain, L.V. et al. The role of copy number variation in susceptibility to amyotrophic lateral sclerosis: genome-wide association study and comparison with published loci. *PLoS One* **4**, e8175 (2009).
121. Grozeva, D. et al. Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Arch Gen Psychiatry* **67**, 318-27 (2010).
122. Prakash, S.K. et al. Rare Copy Number Variants Disrupt Genes Regulating Vascular Smooth Muscle Cell Adhesion and Contractility in Sporadic Thoracic Aortic Aneurysms and Dissections. *Am J Hum Genet* (2010).
123. Glessner, J.T. et al. A genome-wide study reveals copy number variants exclusive to childhood obesity cases. *Am J Hum Genet* **87**, 661-6 (2010).
124. Williams, N.M. et al. Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *Lancet* **376**, 1401-8 (2010).
125. Bochukova, E.G. et al. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**, 666-70 (2010).
126. Pagnamenta, A.T. et al. Rare familial 16q21 microdeletions under a linkage peak implicate cadherin 8 (CDH8) in susceptibility to autism and learning disability. *J Med Genet* (2010).
127. Mefford, H.C. et al. Copy number variation analysis in single-suture craniosynostosis: multiple rare variants including RUNX2 duplication in two cousins with metopic craniosynostosis. *Am J Med Genet A* **152A**, 2203-10 (2010).
128. de Cid, R. et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat Genet* **41**, 211-5 (2009).
129. Zhang, X. et al. High-resolution mapping of genotype-phenotype relationships in cri du chat syndrome using array comparative genomic hybridization. *Am J Hum Genet* **76**, 312-26 (2005).
130. Campbell, L., Potter, A., Ignatius, J., Dubowitz, V. & Davies, K. Genomic variation and gene conversion in spinal muscular atrophy: implications for disease process and clinical phenotype. *Am J Hum Genet* **61**, 40-50 (1997).
131. Scherer, S.W. et al. Human chromosome 7: DNA sequence and biology. *Science* **300**, 767-72 (2003).

132. Vissers, L.E., Veltman, J.A., van Kessel, A.G. & Brunner, H.G. Identification of disease genes by whole genome CGH arrays. *Hum Mol Genet* **14 Spec No. 2**, R215-23 (2005).
133. Senderek, J. et al. Mutation of the SBF2 gene, encoding a novel member of the myotubularin family, in Charcot-Marie-Tooth neuropathy type 4B2/11p15. *Hum Mol Genet* **12**, 349-56 (2003).
134. Horsthemke, B. & Wagstaff, J. Mechanisms of imprinting of the Prader-Willi/Angelman region. *Am J Med Genet A* **146A**, 2041-52 (2008).
135. Trovo-Marqui, A.B. & Tajara, E.H. Neurofibromin: a general outlook. *Clin Genet* **70**, 1-13 (2006).
136. Toyo-oka, K. et al. 14-3-3epsilon is important for neuronal migration by binding to NUDEL: a molecular explanation for Miller-Dieker syndrome. *Nat Genet* **34**, 274-85 (2003).
137. Carlson, C. et al. Molecular definition of 22q11 deletions in 151 velo-cardio-facial syndrome patients. *Am J Hum Genet* **61**, 620-9 (1997).
138. Edelmann, L. et al. A common molecular basis for rearrangement disorders on chromosome 22q11. *Hum Mol Genet* **8**, 1157-67 (1999).
139. Aitman, T.J. et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851-5 (2006).
140. Fanciulli, M. et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet* **39**, 721-3 (2007).
141. Willcocks, L.C. et al. Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake. *J Exp Med* **205**, 1573-82 (2008).
142. Yang, Y. et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am J Hum Genet* **80**, 1037-54 (2007).
143. McKinney, C. et al. Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. *Ann Rheum Dis* **67**, 409-13 (2008).
144. Burns, J.C. et al. Genetic variations in the receptor-ligand pair CCR5 and CCL3L1 are important determinants of susceptibility to Kawasaki disease. *J Infect Dis* **192**, 344-9 (2005).
145. Fellermann, K. et al. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet* **79**, 439-48 (2006).
146. McCarroll, S.A. et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* **40**, 1107-12 (2008).
147. Hollox, E.J. et al. Psoriasis is associated with increased beta-defensin genomic copy number. *Nat Genet* **40**, 23-5 (2008).
148. Piirila, P. et al. Glutathione S-transferase genotypes and allergic responses to diisocyanate exposure. *Pharmacogenetics* **11**, 437-45 (2001).
149. Ivaschenko, T.E., Sideleva, O.G. & Baranov, V.S. Glutathione- S-transferase micro and theta gene polymorphisms as new risk factors of atopic bronchial asthma. *J Mol Med* **80**, 39-43 (2002).
150. Brasch-Andersen, C. et al. Possible gene dosage effect of glutathione-S-transferases on atopic asthma: using real-time PCR for quantification of GSTM1 and GSTT1 gene copy numbers. *Hum Mutat* **24**, 208-14 (2004).

151. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445-9 (2007).
152. Szatmari, P. et al. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet* **39**, 319-28 (2007).
153. Marshall, C.R. et al. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* **82**, 477-88 (2008).
154. Weiss, L.A. et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**, 667-75 (2008).
155. Kumar, R.A. et al. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* **17**, 628-38 (2008).
156. Xu, B. et al. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* **40**, 880-5 (2008).
157. Stefansson, H. et al. Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232-6 (2008).
158. Walsh, T. et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539-43 (2008).
159. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237-41 (2008).
160. Singleton, A.B. et al. alpha-Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
161. Ibanez, P. et al. Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease. *Lancet* **364**, 1169-71 (2004).
162. Veldink, J.H. et al. Homozygous deletion of the survival motor neuron 2 gene is a prognostic factor in sporadic ALS. *Neurology* **56**, 749-52 (2001).
163. Wang, J., Ban, M.R. & Hegele, R.A. Multiplex ligation-dependent probe amplification of LDLR enhances molecular diagnosis of familial hypercholesterolemia. *J Lipid Res* **46**, 366-72 (2005).
164. Fanciulli, M., Petretto, E. & Aitman, T.J. Gene copy number variation and common human disease. *Clin Genet* **77**, 201-13 (2010).
165. Stranger, B.E. et al. Population genomics of human gene expression. *Nat Genet* **39**, 1217-24 (2007).
166. Kudaravalli, S., Veyrieras, J.B., Stranger, B.E., Dermitzakis, E.T. & Pritchard, J.K. Gene expression levels are a target of recent natural selection in the human genome. *Mol Biol Evol* **26**, 649-58 (2009).
167. Dimas, A.S. et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246-50 (2009).
168. Montgomery, S.B. et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-7 (2010).
169. Stranger, B.E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53 (2007).
170. Schuster-Bockler, B., Conrad, D. & Bateman, A. Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS One* **5**, e9474 (2010).
171. Cahan, P., Li, Y., Izumi, M. & Graubert, T.A. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nat Genet* **41**, 430-7 (2009).
172. Chaignat, E. et al. Copy number variation modifies expression time-courses. *Genome Res* (2010).
173. Feuk, L., Marshall, C.R., Wintle, R.F. & Scherer, S.W. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet* **15 Spec No 1**, R57-66 (2006).

174. Henrichsen, C.N., Chaignat, E. & Reymond, A. Copy number variants, diseases and gene expression. *Hum Mol Genet* **18**, R1-8 (2009).
175. Baudis, M. Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC Cancer* **7**, 226 (2007).
176. Mitelman, F., Johansson, B. & Mertens, F. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. (2010).
177. Beroukhi, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905 (2010).
178. Cowell, J.K. & Hawthorn, L. The application of microarray technology to the analysis of the cancer genome. *Current Molecular Medicine* **7**, 103-120 (2007).
179. Bignell, G.R. et al. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res* **17**, 1296-303 (2007).
180. Greenman, C.D. et al. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **11**, 164-75 (2010).
181. Stahl, J.M. et al. Loss of PTEN promotes tumor development in malignant melanoma. *Cancer Res* **63**, 2881-90 (2003).
182. Futreal, P.A. et al. A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183 (2004).
183. Lockwood, W.W. et al. DNA amplification is a ubiquitous mechanism of oncogene activation in lung and other cancers. *Oncogene* **27**, 4615-24 (2008).
184. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).
185. Taylor, B.S. et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11-22 (2010).
186. Bignell, G.R. et al. Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893-8 (2010).
187. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8 (2008).
188. Stratton, M.R., Wooster, R.W. & Futreal, P.A. The BRAF gene is frequently mutated in malignant melanoma. *J Drugs Dermatol* **3**, 573-5 (2004).
189. Santarius, T. et al. GLO1-A novel amplified gene in human cancer. *Genes Chromosomes Cancer* **49**, 711-25 (2010).
190. Zhou, D. et al. Mst1 and Mst2 maintain hepatocyte quiescence and suppress hepatocellular carcinoma development through inactivation of the Yap1 oncogene. *Cancer Cell* **16**, 425-38 (2009).
191. Tang, B. et al. Transforming growth factor-beta can suppress tumorigenesis through effects on the putative cancer stem or early progenitor cell and committed progeny in a breast cancer xenograft model. *Cancer Res* **67**, 8643-52 (2007).
192. Post, S.M. et al. A high-frequency regulatory polymorphism in the p53 pathway accelerates tumor development. *Cancer Cell* **18**, 220-30 (2010).
193. Pollock, P.M. et al. High frequency of BRAF mutations in nevi. *Nat Genet* **33**, 19-20 (2003).
194. Russo, A.E. et al. Melanoma: molecular pathogenesis and emerging target therapies (Review). *Int J Oncol* **34**, 1481-9 (2009).
195. Yazdi, A.S. et al. Mutations of the BRAF gene in benign and malignant melanocytic lesions. *J Invest Dermatol* **121**, 1160-2 (2003).
196. Slominski, A., Zbytek, B. & Slominski, R. Inhibitors of melanogenesis increase toxicity of cyclophosphamide and lymphocytes against melanoma cells. *Int J Cancer* **124**, 1470-7 (2009).
197. Wood, J.M. et al. What's the use of generating melanin? *Exp Dermatol* **8**, 153-64 (1999).

198. Slominski, A., Paus, R. & Mihm, M.C. Inhibition of melanogenesis as an adjuvant strategy in the treatment of melanotic melanomas: selective review and hypothesis. *Anticancer Res* **18**, 3709-15 (1998).
199. Riley, P.A. Melanogenesis and melanoma. *Pigment Cell Res* **16**, 548-52 (2003).
200. Pawelek, J., Korner, A., Bergstrom, A. & Bologna, J. New regulators of melanin biosynthesis and the autodestruction of melanoma cells. *Nature* **286**, 617-9 (1980).
201. Huang, J. et al. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics* **1**, 287-99 (2004).
202. Komura, D. et al. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res* **16**, 1575-84 (2006).
203. Nannya, Y. et al. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res* **65**, 6071-9 (2005).
204. Lieberfarb, M.E. et al. Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP. *Cancer Res* **63**, 4781-5 (2003).
205. Allison, D.B., Fontaine, K.R., Manson, J.E., Stevens, J. & VanItallie, T.B. Annual Deaths Attributable to Obesity in the United States. *JAMA: The Journal of the American Medical Association* **282**, 1530-1538 (1999).
206. Haslam, D.W. & James, W.P.T. Obesity. *The Lancet* **366**, 1197-1209 (2005).
207. Whitlock, G. et al. Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. *Lancet* **373**, 1083-96 (2009).
208. Vainio, H. & Bianchini, F. International Agency for Cancer handbook of cancer prevention, vol 6. Weight control and physical activity. *IARC* (2002).
209. Calle, E.E., Rodriguez, C., Walker-Thurmond, K. & Thun, M.J. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *N Engl J Med* **348**, 1625-38 (2003).
210. Reeves, G.K. et al. Cancer incidence and mortality in relation to body mass index in the Million Women Study: cohort study. *BMJ* **335**, 1134 (2007).
211. Green, B.B., Weiss, N.S. & Daling, J.R. Risk of ovulatory infertility in relation to body weight. *Fertil Steril* **50**, 721-6 (1988).
212. Janssen, I. & Mark, A.E. Elevated body mass index and mortality risk in the elderly. *Obes Rev* **8**, 41-59 (2007).
213. Breeze, E., Clarke, R., Shipley, M.J., Marmot, M.G. & Fletcher, A.E. Cause-specific mortality in old age in relation to body mass index in middle age and in old age: follow-up of the Whitehall cohort of male civil servants. *Int J Epidemiol* **35**, 169-78 (2006).
214. Bell, C.G., Walley, A.J. & Froguel, P. The genetics of human obesity. *Nat Rev Genet* **6**, 221-34 (2005).
215. Barsh, G.S., Farooqi, I.S. & O'Rahilly, S. Genetics of body-weight regulation. *Nature* **404**, 644-51 (2000).
216. Hill, J.O. & Peters, J.C. Environmental contributions to the obesity epidemic. *Science* **280**, 1371-4 (1998).
217. Gunay-Aygun, M., Cassidy, S.B. & Nicholls, R.D. Prader-Willi and other syndromes associated with obesity and mental retardation. *Behav Genet* **27**, 307-24 (1997).
218. Farooqi, I.S. et al. Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. *N Engl J Med* **348**, 1085-95 (2003).

219. Calton, M.A. et al. Association of functionally significant Melanocortin-4 but not Melanocortin-3 receptor mutations with severe adult obesity in a large North American case-control study. *Hum Mol Genet* **18**, 1140-7 (2009).
220. Branson, R. et al. Binge eating as a major phenotype of melanocortin 4 receptor gene mutations. *N Engl J Med* **348**, 1096-103 (2003).
221. Montague, C.T. et al. Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* **387**, 903-8 (1997).
222. Clement, K. et al. A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* **392**, 398-401 (1998).
223. Jackson, R.S. et al. Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. *Nat Genet* **16**, 303-6 (1997).
224. Krude, H. et al. Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. *Nat Genet* **19**, 155-7 (1998).
225. Norman, R.A. et al. Genomewide search for genes influencing percent body fat in Pima Indians: suggestive linkage at chromosome 11q21-q22. Pima Diabetes Gene Group. *Am J Hum Genet* **60**, 166-73 (1997).
226. Dong, C. et al. Possible genomic imprinting of three human obesity-related genetic loci. *Am J Hum Genet* **76**, 427-37 (2005).
227. Chambers, J.C. et al. Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nat Genet* **40**, 716-8 (2008).
228. Loos, R.J. et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* **40**, 768-75 (2008).
229. Meyre, D. et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat Genet* **41**, 157-9 (2009).
230. Renstrom, F. et al. Replication and extension of genome-wide association study results for obesity in 4923 adults from northern Sweden. *Hum Mol Genet* **18**, 1489-96 (2009).
231. Miller, A.J. & Mihm, M.C., Jr. Melanoma. *N Engl J Med* **355**, 51-65 (2006).
232. Becher, R., Gibas, Z., Karakousis, C. & Sandberg, A.A. Nonrandom chromosome changes in malignant melanoma. *Cancer Res* **43**, 5010-6 (1983).
233. Ozisik, Y.Y. et al. Cytogenetic findings in 21 malignant melanomas. *Cancer Genet Cytogenet* **77**, 69-73 (1994).
234. Omholt, K., Platz, A., Kanter, L., Ringborg, U. & Hansson, J. NRAS and BRAF mutations arise early during melanoma pathogenesis and are preserved throughout tumor progression. *Clin Cancer Res* **9**, 6483-8 (2003).
235. Davies, H. et al. Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res* **65**, 7591-5 (2005).
236. Albino, A.P. et al. Analysis of ras oncogenes in malignant melanoma and precursor lesions: correlation of point mutations with differentiation phenotype. *Oncogene* **4**, 1363-74 (1989).
237. Kamb, A. et al. Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nat Genet* **8**, 23-6 (1994).
238. Hussussian, C.J. et al. Germline p16 mutations in familial melanoma. *Nat Genet* **8**, 15-21 (1994).
239. Wu, H., Goel, V. & Haluska, F.G. PTEN signaling pathways in melanoma. *Oncogene* **22**, 3113-22 (2003).
240. Garraway, L.A. et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117-22 (2005).
241. Johnson, J.P. Cell adhesion molecules in the development and progression of malignant melanoma. *Cancer Metastasis Rev* **18**, 345-57 (1999).

242. Cowley, G.P. & Smith, M.E. Cadherin expression in melanocytic naevi and malignant melanomas. *J Pathol* **179**, 183-7 (1996).
243. Kuphal, S., Bauer, R. & Bosserhoff, A.K. Integrin signaling in malignant melanoma. *Cancer Metastasis Rev* **24**, 195-222 (2005).
244. Stark, M. & Hayward, N. Genome-wide loss of heterozygosity and copy number analysis in melanoma using high-density single-nucleotide polymorphism arrays. *Cancer Res* **67**, 2632-42 (2007).
245. Hoek, K. et al. Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas. *Cancer Res* **64**, 5270-82 (2004).
246. Gast, A. et al. Somatic alterations in the melanoma genome: a high-resolution array-based comparative genomic hybridization study. *Genes Chromosomes Cancer* **49**, 733-45 (2010).
247. Berger, M.F. et al. Integrative analysis of the melanoma transcriptome. *Genome Res* **20**, 413-27 (2010).

2 Methods

This chapter is about the main statistical methods that I used during my PhD. All the presented techniques or algorithms are general and can be applied to a wide range of analyses (regression analyses, prediction, clustering, signal segmentation etc). I decided to organize these techniques by themes, from micro-array normalization to CNV calling. I also included a section on multivariate and cluster analysis, topics that inspired me to design my own algorithms.

2.1 Normalization methods

Data normalization is a crucial aspect in any analysis, but in particular in the genomic era, where experiments need to be corrected for putative technical biases, as well as biological and experimental noise. Also experiments from different samples need to be calibrated with respect to each other to ensure that the results are comparable.

2.1.1 Mean and median scaling

The hybridization intensities measured by both CGH and SNP arrays can be combined into a hybridization ratio in order to infer CNVs (see Chapter 1). By definition, these ratios are expressed on the \log_2 scale, and reflect copy number changes between a test and a reference sample (either one sample or a pool of references). In an ideal, un-biased experiment, three copies measured with respect to a diploid locus, would have a \log_2 ratio equals to 0.58 ($=\log_2(3/2)$). Conversely a deletion (one copy with respect to two) will have a \log_2 ratio equal to -1. Yet, the observed \log_2 ratios usually do not assume exactly these values because of the noise in the measurements that causes signal variation from one chromosome to another. A simple correction can be a median subtraction (or centering) whereby the median ratio from each chromosome is subtracted from the ratio at each probe (Figure 1). This approach is more adequate than the mean subtraction, because the mean is sensitive to outliers whereas the median, by definition, ignore most of the values (Figure 2). In the past, median subtraction has been used to calibrate red and green intensities from two colour (gene expression) experiments ^{1,2}. However such centering does not consider local variations and the calibration is often in-sufficient, especially for aneuploid genome analysis (Figure 3).

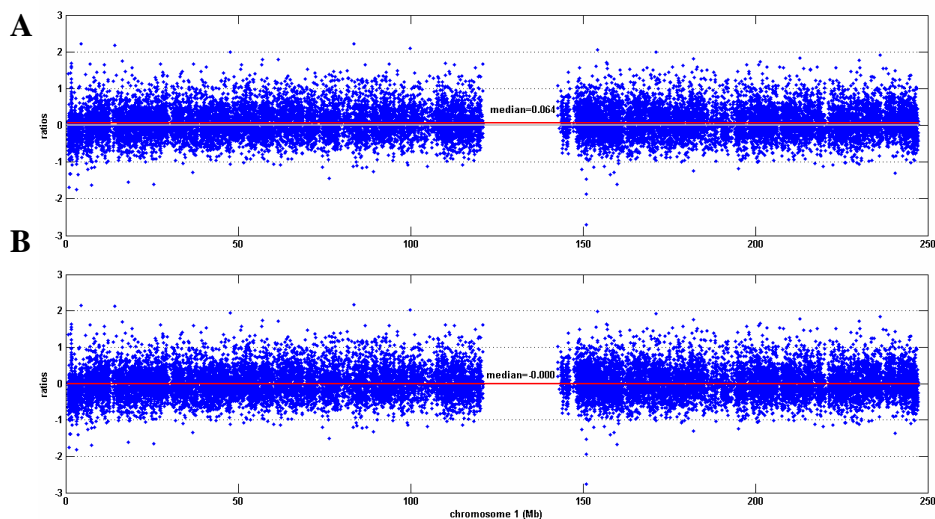


Figure 1 Median subtraction

A CGH ratios on chromosome 1, red line indicates the median and gray line $y=0$; **B** CGH ratios after median subtraction.

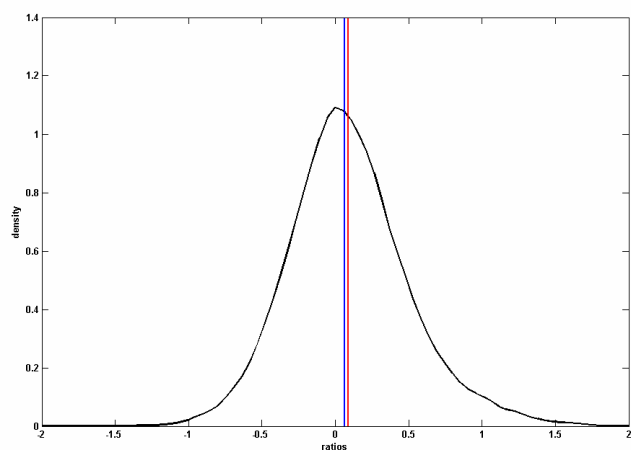


Figure 2 Mean and median estimation

Distribution of ratios from Figure 1, the red (blue) line corresponds to the mean (median). Neither the median nor the mean are centred on zero and the mean is slightly higher than the median.

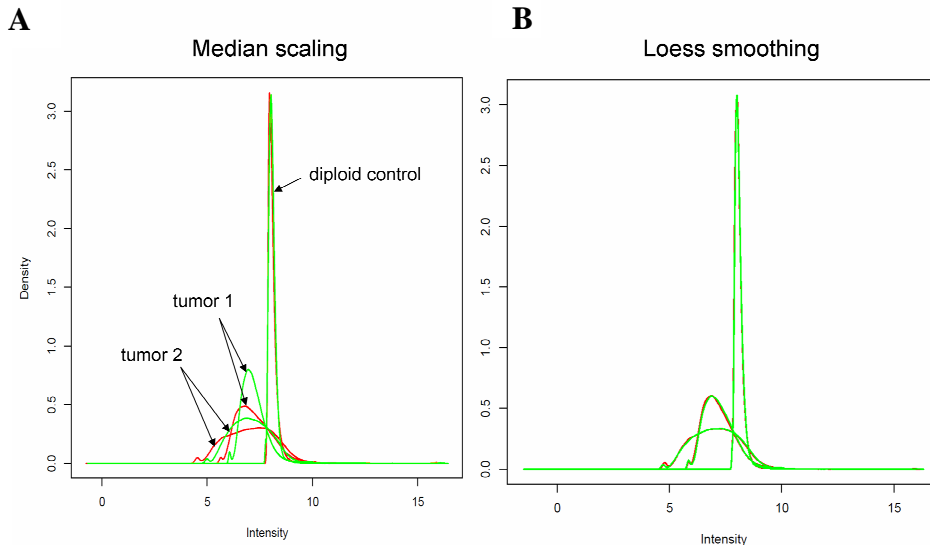


Figure 3 Calibration of red and green intensities in CGH experiments

A Median scaling is applied independently to the two intensity distribution for two tumor cell lines and a diploid control cell. The calibration of the red and green dye is only correct for the diploid genome **B** Calibration using Loess smoothing; the calibration is more robust than the median scaling.

2.1.2 Linear Least Square Regression

Linear regression is a simple but powerful method to test for a (linear) relationship between two variables X and Y . Assuming Y as the response variable and X as the variable that explains the response, linear regression fits a line between points from X and Y (Figure 4A). This can be written as: $Y = a + bX$ where a is the intercept and b the slope of the line. Both terms are estimated by the regression using a sample $(X_1, Y_1), (X_2, Y_2) \dots, (X_n, Y_n)$. The Least square method^{3,4} is used to find the best fit, and corresponds to finding the regression line where the sum of squares of vertical errors is minimized (Figure 4). Linear regression is usually formalized as: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where Y_i is the response variable and X_i the explanatory variable at a given point i ; both variables are known from the data. The model parameters are β_0 , and β_1 and are estimated with the regression; ε_i corresponds to the vertical error and is referred as error term or residual. This error term is assumed to be normally distributed and the $\sum \varepsilon_i^2$ is minimized by the least square method.

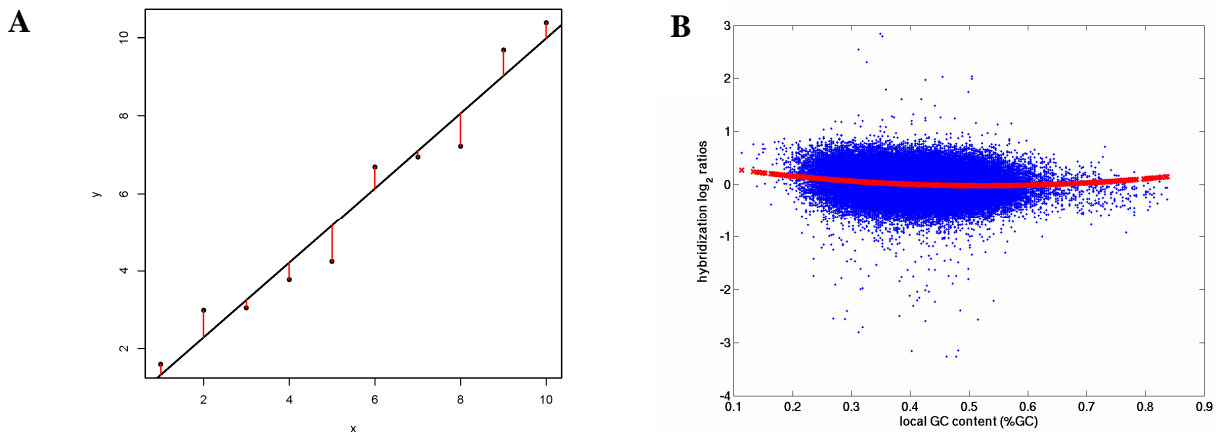


Figure 4 Linear regression

A Dots are the observed data, the black line was fitted with a linear regression and the red lines indicate the vertical errors which were minimized using the Least-square method. **B** Regression with quadratic effects to explain hybridization ratios (measured from a CGH array) as a function of GC content. The red cross symbols correspond to a fit $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$ where the response variable (Y) is the hybridization ratio and explanatory variable (X) the local GC content of a probe (from chromosome 1).

Linear regression is commonly used for genome-wide association studies for quantitative traits analysis (assuming the trait follows a normal distribution). Often the biological system is more complex and should be expressed as a non-linear model. For example, in micro-arrays, local GC content at each probe can affect the hybridization ratios⁵, such bias can be modelled (and subsequently corrected for) with a non-linear regression by adding quadratic effects in the regression formula: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$ with Y_i being the hybridization ratio and X_i the GC content in a window centred on probe i (see Figure 4B).

2.1.3 Loess smoothing

Locally weighted polynomial regression (Lowess), better known as Loess, was originally proposed by Cleveland⁶ and developed by Cleveland and Devlin⁷. A Loess fits a polynomial regression⁸ at each data point; using the evaluated point as response variable and points in the local neighbourhood as explanatory variables. The extent of the neighbourhood is called bandwidth or smoothing parameter, and its value is defined by the data analyst. Weights are applied in the regression with the weight least squares method^{3,4}. Data points close to the response variable are given higher weights than points further away. Loess is very flexible and only requires the specification of a bandwidth to partition the data for the regression. Loess has been extensively used in micro-array analysis, initially with the normalization of two dye

arrays (Figure 3B, ^{9,10}), and more recently to remove auto-correlation between probes (“spatial artefact”) (Figure 5; ¹¹).

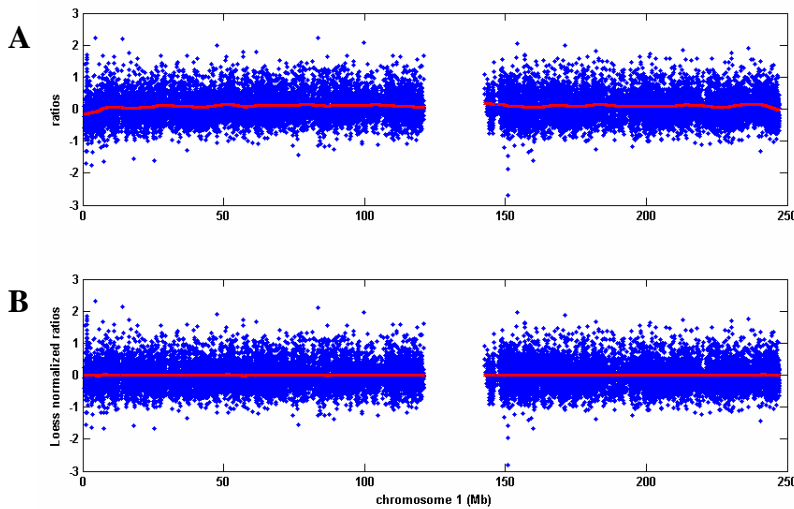


Figure 5 Loess smoothing

A Ratios as in Figure 1, red line indicates the Loess fit, **B** ratios were normalized using the Loess fit from A. A new Loess fit reveals almost a straight line close to 0, indicating a good normalization.

2.1.4 Quantile-quantile normalization

In micro-array analysis, it is common that the signal distribution from different experiments differ (Figure 6A). In most cases, distributions can be made comparable with simple median re-centering or with Loess smoothing. However when comparing samples that have different properties, e.g. tumours, different tissues or experiments made in different labs, the respective distributions cannot be made comparable using simple (linear) normalization. Quantile-quantile (QQ) normalization is a technique that forces two distributions to become identical in statistical properties (Figure 6B). QQ normalization sorts the values from two distributions (a test and a reference distribution), the highest value in the test distribution is re-attributed the value of the highest in the reference distribution, then the second highest in the test is recalibrated to the second highest in the reference and so on. This forces the values of the test distribution to be on the same quantile than those from the reference, while preserving the initial rank of the values from the test distribution. To normalize several distributions, the reference distribution can be determined by selecting points randomly from all distributions. QQ normalization is useful to satisfy normality assumptions for various test-statistics. However, it only maintains the rank of values,

so valuable information on skewness or multimodality of the distribution is lost. QQ normalization is extremely popular to normalize SNP or gene expression micro-arrays ¹². In genome-wide association studies, it can be used to transform a non-normally distributed phenotype to a normal distribution.

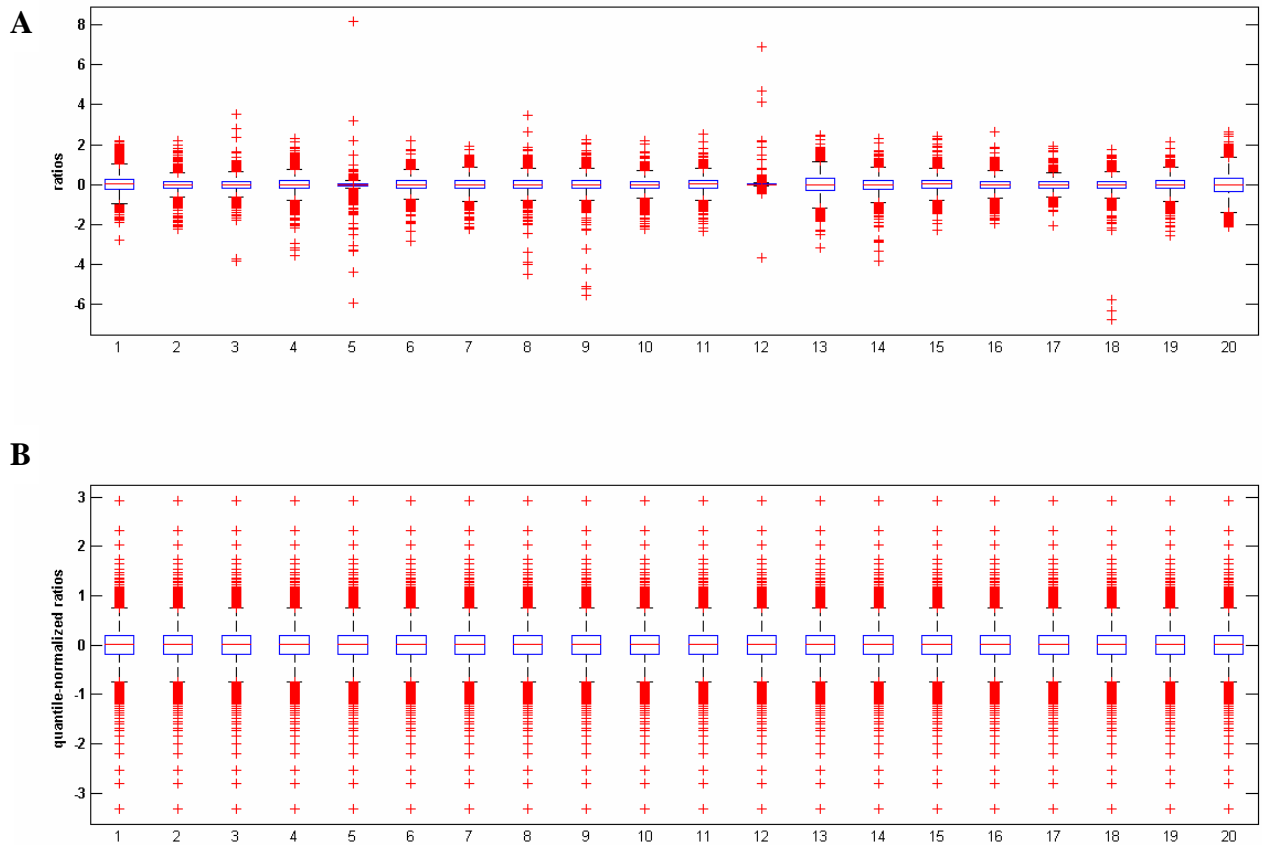


Figure 6 Quantile normalization

A Distribution of autosomal log₂ ratios from twenty CGH experiments, **B** log₂ ratios after quantile normalization.

2.2 Segmentation methods

Following normalization, a classical approach to detect CNV from CGH arrays, is to partition the hybridization ratio profile into consecutive segments that correspond to copy number changes. Such segments are constituted with probes having similar hybridization ratios. There is a plethora of published algorithms, each making use of different approaches and with different performances. In this section, I will explain the main segmentation approaches and highlight each with representative *state-of-art* and freely available algorithms.

2.2.1 Outlier-based detection

The goal of segmentation analysis is to differentiate between events that are significantly higher or lower than the baseline (or background) signal. A naïve approach consists of detecting outlier probes and merge them into regions. Outliers can be detected with several approaches; when the data are normally distributed, a straightforward approach is to use Z-score:

$$Z_i = \frac{X_i - \text{mean}(X)}{\text{std}(X)}$$

With X being the ensemble of data (the sample) and X_i a given point (i) of this sample. The Z-score reflects in how many units of standard deviation X_i is away from the sample mean. Assuming normally distributed data, a Z-score greater or equal to three will describe 0.1% of the data (Figure 7).

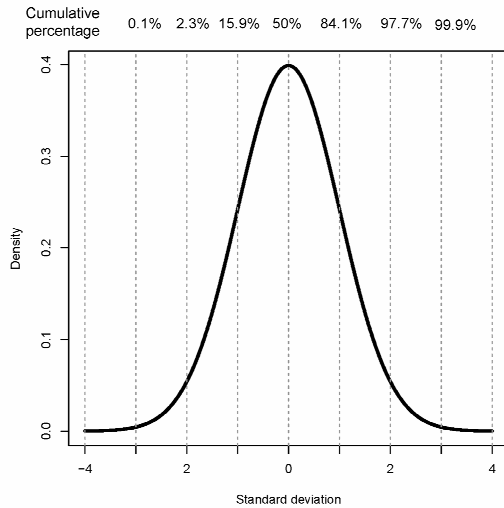


Figure 7 Normal Gaussian distribution with mean 0 and standard deviation 1

Since the assumption of normality does not often hold for genomic data, non-parametric approaches are more appropriate. An alternative to the Z-score could be

computed as: $\frac{X_i - \text{median}(X)}{MAD(X)}$

Where *MAD* corresponds to the median absolute deviation, a robust estimator of dispersion around the median. The *MAD* is computed as follow: $\text{median}(|X_i - \text{median}(X)|)$

MAD-based approaches have been used for the analysis of CGH¹³, providing a simple and fast method for CNV analysis. I also used a simple *MAD*-based scheme to analyse copy number aberrations in metastatic melanoma (see Chapter 5).

2.2.2 Recursive binary segmentation

Recursive binary segmentation consists of splitting a ratio profile into two segments and testing if the split result is significant in regions with different mean ratio. If it is, then a new split is attempted for one of these new segments and so on until no more split is possible. Testing the difference of two segments is done using permutation-based approaches. This segmentation procedure is well-known as the Circular Binary Segmentation proposed by Olshen et al. ^{14,15}. It has been applied to CGH analysis and is recognized as a *state-of-art* method in the community ^{16,17}. However due to the recursive nature of the algorithm and the permutations performed at each split, the method is computationally very intensive. To overcome this limitation, an improved CBS version has been subsequently proposed by the authors ^{14,15}. In this newer version, a stopping rule was added to stop the permutation procedure when strong evidence for differences between segments was found in early permutations. Despite this, it still remains computationally intensive but CBS remains a method of choice for CNV analysis (see Chapter 3 and 5).

2.2.3 Dynamic programming techniques

Dynamic programming is a method to solve optimization problems and consists of solving a complex problem by recursively breaking it into smaller problems until a solution can be found. Applications can be illustrated with algorithms to find the shortest path in a graph or in a scoring matrix. Bioinformatics has greatly benefited from dynamic programming, in particular in sequence analysis with the Smith-Waterman algorithm¹⁸ that finds the optimal local alignment with respect to a scoring matrix. The Smith-Waterman algorithm has been extended to CGH analysis by Price et al.¹⁹. The idea is to detect an initial segment made with adjacent, outlier probes and to expand the segment until no more expansion is possible. The detection of outliers is achieved via the median absolute deviation (*MAD*), described previously. Any probe with a ratio greater than $\alpha * MAD$ is considered as an outlier. This parameter α is a threshold provided by the data analyst. Adjacent outliers are combined in a segment, and a score is computed as $\sum_{i=n}^m X(i) - \alpha MAD$, where n is the first probe in the segment and m the last probe. The initial segment score is positive by definition. The best segment is determined using the Smith-Waterman principle: the score of the optimal segment cannot be improved by shrinking or expanding its boundaries. In practice, from the initial segment, adjacent probes are added iteratively and the score is updated. When the score becomes null or negative, the previous segment corresponded to the optimal solution. In addition to this scheme, the significance of the final segments can be controlled with a permutation-based procedure.

2.2.4 Linear piecewise regression

Copy number segments in the genome have two properties, 1) there are far less segments than assayed probes on the microarray and 2) the copy number at each segment is assumed to be discrete (0, 1, 2, 3..). Such properties cannot be measured directly from the hybridization ratios due to biological contamination (e.g. mixture of tumor and diploid cells) or technical noise inherent to the experiment. Nevertheless such segments can be estimated using piecewise linear regression. Piecewise linear regression is a form of regression that fits multiple linear models to the data (Figure 8A). Assuming there is one breakpoint c in the data, the model from a piecewise regression can be written as:

$$y = a_1 + b_1 x \text{ for } x \leq c$$

$$y = a_2 + b_2 x \text{ for } x > c$$

Such regression can also use a step function (Figure 8B), which provides a natural framework to identify segments of copy number. This technique has been further developed by Pique Regi et al. ²⁰, where segments are obtained with a piecewise-constant regression. The list of segments is controlled by Bayesian approaches to optimize the balance between the number of segments (the Bayesian prior parameter) and the regression fit (sum of residual squares), then a backward elimination procedure removes segments whose score (a *T-statistics*) is below a critical value. It has been demonstrated both by the authors ²⁰ and others ¹⁷ that this method (GADA) offers similar performances as the Circular Binary Segmentation (CBS, described previously). Moreover, GADA is two orders of magnitude faster than CBS and thus is a new method of choice for analysis of ultra-high resolution arrays ¹⁷.

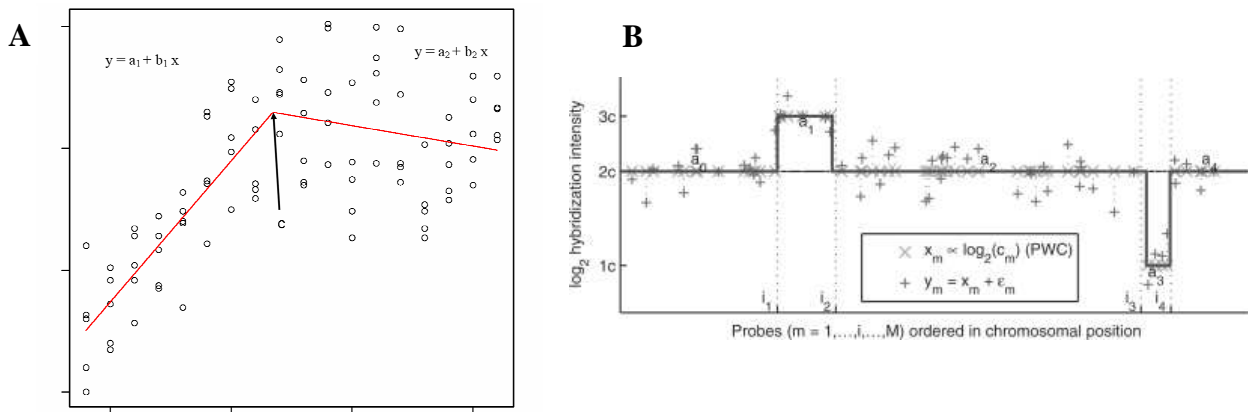


Figure 8 Piecewise regression
A Simple linear piecewise regression **B** (from Pique-Regi et al. Bioinformatics 2008) Piecewise constant regression on hybridization ratios as a function of their genomic position.

2.2.5 Hidden Markov Model

Hidden Markov Models (HMMs) are widely used in bioinformatics for gene prediction, motif search and sequence alignment²¹⁻²³. HMMs are extensions of Markov chains (or Markov Model) which consist in a sequence of states (s_1, s_2, \dots, s_k). A Markov chain starts at one state and moves from one state to another with a series of steps (Figure 9). If the chain is currently in a state s_i , the move to state s_j is happening with a probability (P_{ij}) which only depends on s_i and not from the previous states (s_1, \dots, s_{i-1}). The process can also remain in the same state with a probability (P_{ii}). These probabilities (P_{ii} and P_{ij}) are called transition probabilities.

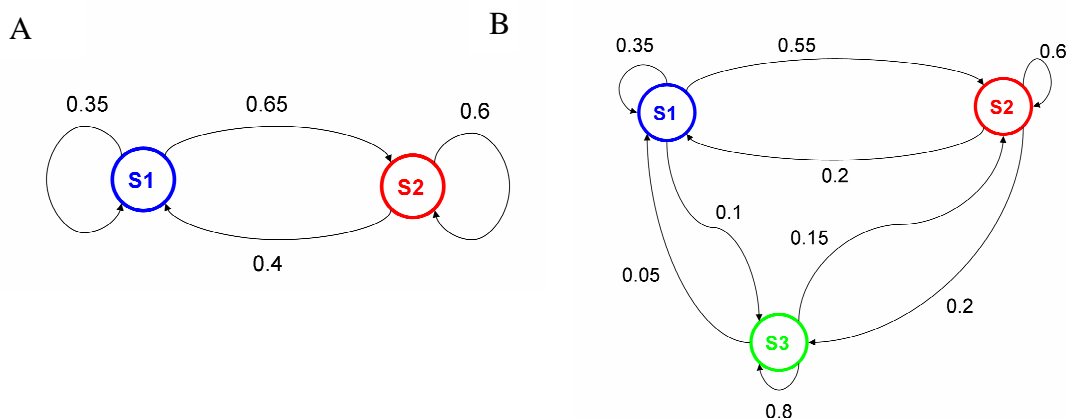


Figure 9 Markov Model
A two states Markov chain and **B** three states Markov chain. Transition probabilities are indicated near each edge.

In HMMs the state is not directly visible (i.e. hidden) but an output driven by the hidden state can be observed. In a Markov Chain the only parameters are the transition probabilities whereas HMMs are defined with transition probabilities and emission probabilities (also called output probabilities). The emission probability corresponds to the probability for an observation to be associated with a given state (Figure 10A). At a location l , an observation Y_l depends on its emission probability and on state S_l . State S_l depends on the transition probability from a state S_{l-1} (Figure 10B).

Formally an HMM can be defined as follow:

- 1) the number of state K in the HMM, the individual states are noted \mathbf{s} (s_1, s_2, \dots, s_K) and S_l correspond to a state at a location l with $1 \leq l \leq L$
- 2) the initial state distribution $\boldsymbol{\pi}=\{\pi_k\}$, with $\pi_k=P(s_1 = S_k)$ and $1 \leq k \leq K$
- 3) the state transition probabilities $\mathbf{a}=\{P_{ij}\}$ with $P_{ij}=P(S_{l+1} = s_j | S_l = s_i)$ and $1 \leq i, j \leq K$
- 4) and the emission probability density function \mathbf{b} . In special cases (e.g. CNV analysis), \mathbf{b} follows a normal distribution with mean μ_k and covariance matrix U_k : $\{b_k(\mathbf{Y})\} \sim N(\mathbf{Y}, \mu_k, U_k)$ where \mathbf{Y} corresponds to the vector of observations (Y_1, Y_2, \dots, Y_L) that is being modelled by the HMM.

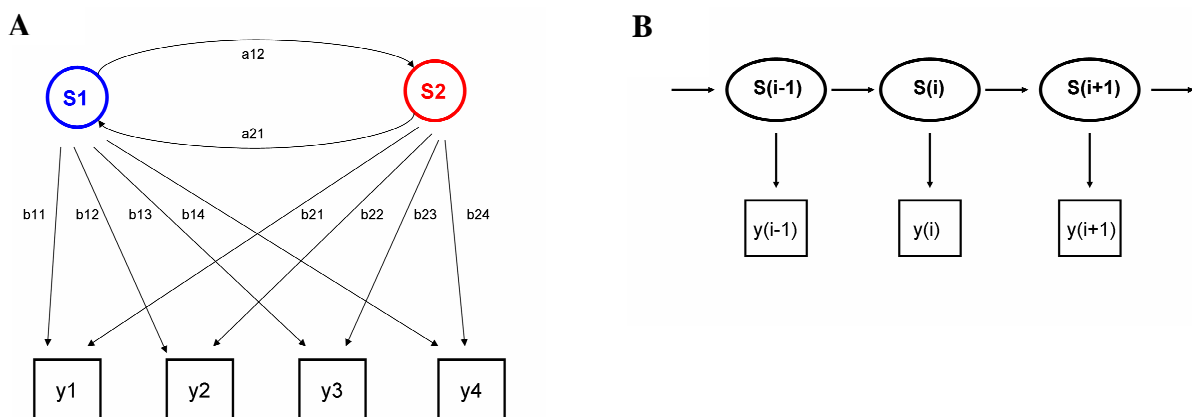


Figure 10 Hidden Markov Model

A Parameters of a Hidden Markov Model with two states ($S1$ and $S2$), four observations (\mathbf{y}), \mathbf{a} denotes state transition probabilities and \mathbf{b} emission probabilities. **B** HMM representation with S_i the state to be predicted having predicted S_{i-1} and with the observation y_i .

HMMs are very popular for CNV analysis since they were proposed by Fridlyand et al. ²⁴ because by essence they enable both segmentation and classification. In SNP and CGH arrays, probes from a same chromosome follow a Markov process because their position (ordered along the chromosome) is correlated with a local copy number state. The underlying copy number is the hidden state and can be modelled from the hybridization ratios (the observations) and also by knowing the corresponding transition probability. Many implementations of HMMs are available both for CGH ²⁴⁻²⁶ and SNP arrays ²⁷⁻³¹.

In CNV analysis, HMMs aim at classifying probes into a discrete and small number of states (e.g. homozygous deletion, hemizygous deletion, copy neutral, duplication and amplification). The optimal sequence state is derived using the Viterbi algorithm ³². This is achieved by predicting the most likely state S_i at each probe i . Subsequently parameters can be re-estimated to maximize the likelihood of the model using the Baum-Welch algorithm ³³ or the Expectation-Maximization algorithm ³⁴.

The challenge in using HMMs is to accurately estimate its parameters. It has been demonstrated ³⁵ that while the initial parameters $\boldsymbol{\pi}$ and \boldsymbol{a} (respectively the initial state distribution and the state transition probabilities) can be arbitrary decided, good initial estimation of the emission probability distribution (\boldsymbol{b}) is important. \boldsymbol{b} can be estimated either by training the HMM on known datasets ^{27,31} or by using Bayesian approaches ²⁶.

2.3 Multivariate and cluster analysis

Multivariate analysis is the simultaneous analysis of more than one variable. Bivariate analysis (for e.g. line fitting with linear regression, segmentation analysis etc...) is a special case of multivariate analysis where only two variables are analysed. Multivariate analysis is omnipresent in genomic data, for example when comparing several features (i.e. genes) across several samples (or conditions). Multivariate analysis is a very broad field, in this section I will only concentrate on methods that I used intensively in my analyses.

2.3.1 Principal Component Analysis

One-way PCA

Principal Component Analysis (PCA) ³⁶ is a linear algebraic technique which projects a data matrix onto a new subspace. Projection is done such that the largest variation can be explained along the new axes (Figure 11). These axes are called principal components (PC) and are orthogonal to each other. The type (or source) of variance explained differs from one PC to another. PCs are indexed (*PC1*, *PC2* etc..) according to the fraction of explained variance: *PC1* explains more than variance than *PC2*, *PC2* explains more than *PC3* etc...

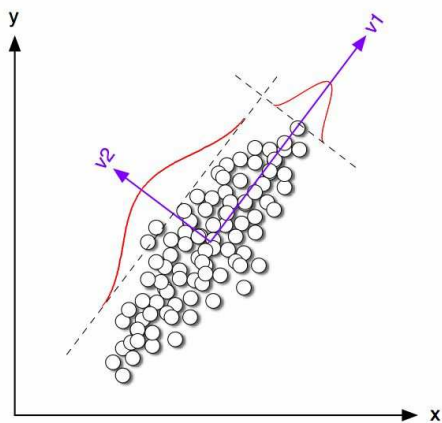


Figure 11 Principal component analysis - from Sven Bergmann, “Brief Introduction to Systems Biology” Scatter plot of x and y data, PCA analysis will project onto a new subspace with axes V1 and V2. Most of the variation will be explained by the V1 axis, a smaller (and orthogonal) fraction will be explained with V2.

The number of PCs is equal to the rank of the data matrix and PCs are the eigenvectors of the covariance matrix C computed as:

$$C = \frac{M^T * M}{n - 1}$$

Here M is the original data matrix, assumed to have a zero empirical mean (the sum of each column has a zero mean); n is the number of dimension (number of rows) and M^T is the transposed matrix of M . The eigenvectors of the C satisfy the following relation:

$$C * PC_i = \lambda_i * PC_i \quad \text{with } \lambda_i \text{ the } i\text{th eigenvalue of } C.$$

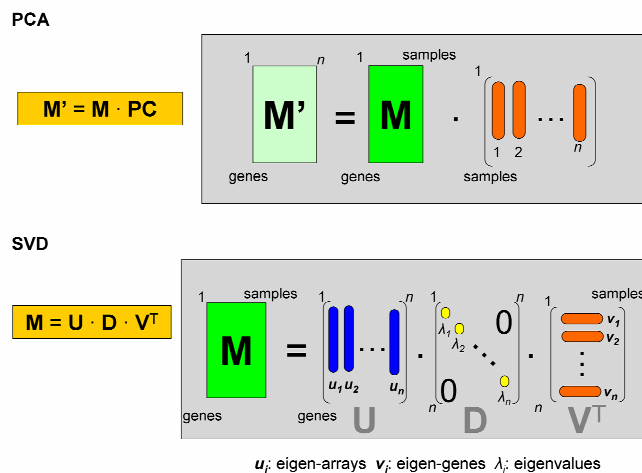


Figure 12 PCA and SVD decomposition - adapted from Sven Bergmann, “Brief Introduction to Systems Biology”

Singular Value Decomposition

PCs can also be obtained by Singular Value Decomposition (SVD, a two-way PCA), written as: $M = U * D * V^T$ where V^T correspond to the principal components (PCs) previously described (see also Figure 12).

Let M be a matrix of genes by samples (as frequently used in microarray analysis), then U is the PC matrix of “eigen-genes”, obtained from the eigenvectors of $C' = M * M^T$; V is the PC matrix of “eigen-arrays”, composed of eigenvectors of $C' = M^T * M$ and D is the diagonal matrix of eigenvalues (λ).

Applications of PCA techniques

PCA is extremely useful to reduce the complexity of a dataset (by reducing its number of dimension and using only PCs explaining most of the variance). It has a wide range of applications from data compression to computer vision and bioinformatics. In genomics, PCA is frequently used to check for batch effects (for e.g. to answer if most of the variance in a dataset is explained by (known or suspected) technical or experimental biases) (see Figure 13).

PCA has been recently used to investigate population stratification in genotyping cohorts. Novembre et al.³⁷ showed it was possible given the genotype of an individual to locate him (her) within 840 km from his/her reported origin (in 90% of the cases). Since then, genome-wide association studies routinely correct for population stratification using the main PCs from PCA decomposition of the genotype data. This corrects for putative biases due to the population structure.

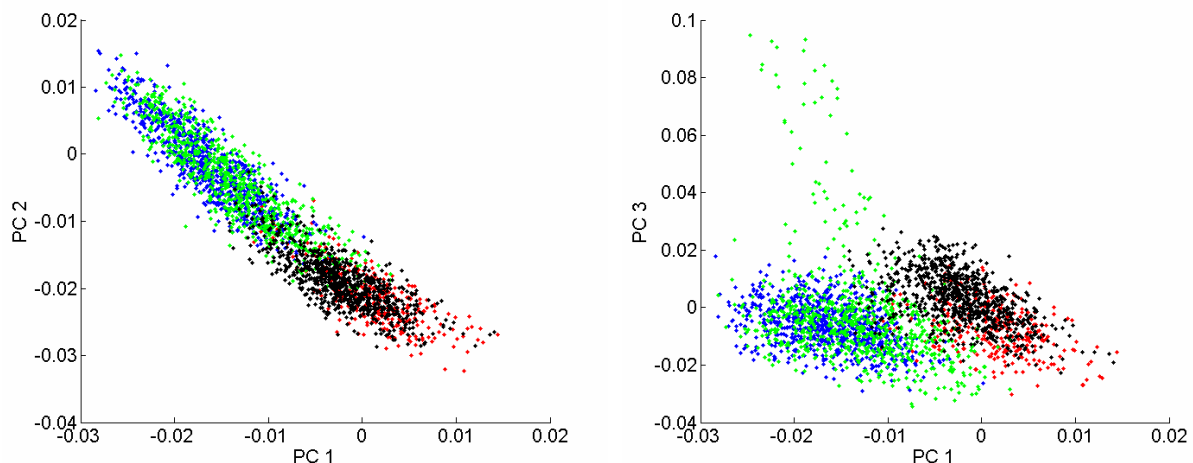


Figure 13 PCA analysis reveals batch effects

PCA analysis using the copy number (at each autosomal SNP) from 2654 male samples from CoLaus. Left panel shows PC1 and PC2; right panel shows PC1 and PC3. Each dot corresponds to a sample. Colours indicate the genotyping facility where a given sample was assayed. Samples processed at the blue and green facilities cluster differently than samples processed at the black and red facilities, suggesting a strong batch effect.

2.3.2 Gaussian Mixture Models

Gaussian Mixture Models (GMM) are composed by combining multivariate Gaussian distributions. Each distribution or component is defined by its mixture proportion (γ_k), mean (μ_k) and variance (σ_k). GMM are useful for unsupervised clustering and are also applicable to univariate data analysis. Estimation of the underlying components from a variable X is frequently made using the Expectation-Maximization (EM) algorithm³⁴ and requires an initial guess about parameters (mean and variance) of the underlying Gaussian components. The EM is an iterative algorithm with an expectation (E) step that estimates the likelihood of the data to come from a Gaussian component Y_k . Then a maximization (M) step updates parameters (μ_k, σ_k) to maximize the log-likelihood value determined in the E step (given the mixture proportions γ). Several E and M iterations are performed until the algorithm converges to an optimal solution (optimized likelihood). An example of GMM clustering with *in-silico* data using the EM algorithm is given in Figure 14A and an example applied to CNV analysis is shown in Figure 14B.

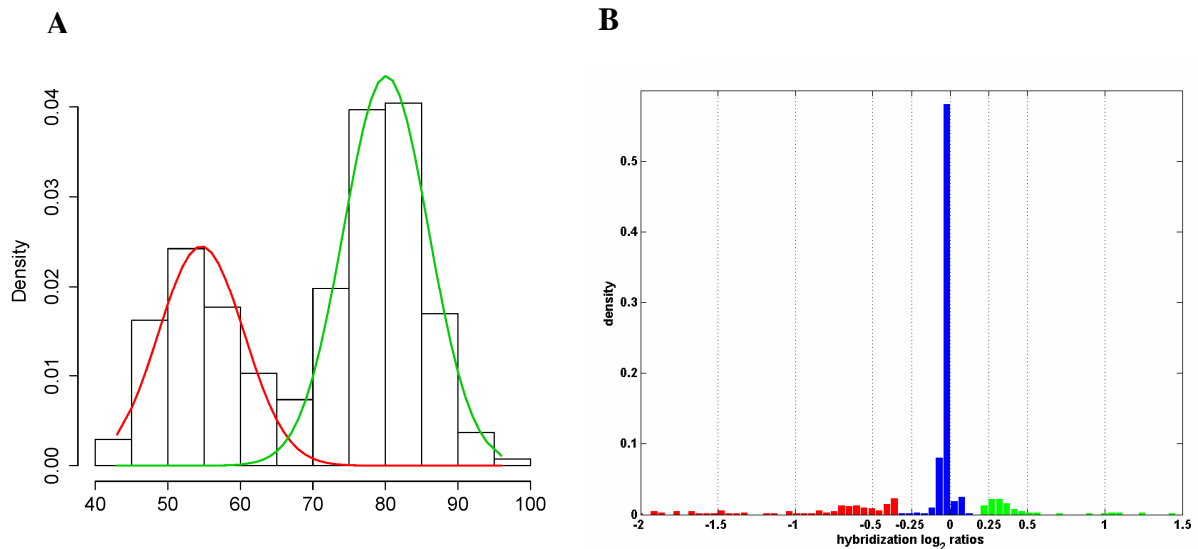


Figure 14 Gaussian Mixture distributions

A The histogram shows the observed data, the green and red distribution represent the two Gaussian components identified using an EM algorithm. **B** Modelling of copy number log₂ ratios: a three Gaussian Mixture was fitted to the data and revealed deletion, copy neutral and duplication components (as indicated in red, blue and green respectively). For display purpose, ratios are only shown from -2 to 1.5.

An alternative to the EM algorithm is to use Markov chain Monte Carlo simulations (MCMC) (e.g. with the Gibbs sampling algorithm^{38,39} that samples (draws) posterior probabilities (from a probability distribution) to explain a fit given the data and the parameters. Based on these posterior probabilities, the parameters are updated and a new sampling is done until probabilities reach equilibrium. MCMC is a very efficient and sophisticated method to estimate parameters in a model, but it is very computationally intensive (a large number of sampling is required to find the equilibrium).

The goodness of the GMM fit can be evaluated using the Bayes Information Criterion (BIC) or Akaike Information Criterion (AIC), that provide the likelihood of the fit given the data and given the number of components in the model. The BIC and AIC can be expressed as

- $AIC = 2k - 2 \ln(L)$
- $BIC = k \ln(n) - 2 \ln(L)$

With k being the number of parameters (for GMM, it is $3 * \text{the number of components} - 1$), L the maximum of the likelihood function, and n the number of observations.

AIC and BIC are very similar, but the BIC is slightly more conservative as it accounts for the sample size. Typically several fits with different k will be attempted; the model that minimizes the AIC or BIC will be considered as the best model.

2.3.3 K-means clustering

K-means is a clustering method which partitions the data into k clusters (Figure 15). K-means partitioning can also be achieved by the EM algorithm, as it is an iterative procedure that starts with initial guess about the centers (called the centroids) of each k clusters, then data points are assigned to the closest centroid (E step) then the centroid positions are refined to be the center of the assigned points (M step). These steps are repeated until the within-cluster sum of squares is minimized

($WCSS = \sum_{i=1}^k \sum_{x_j \in S_i} |x_j - \mu_i|^2$), where k is the number of clusters (S_1, S_2, \dots, S_k), x_j a point in cluster S_i and μ_i the mean of points in S_i .

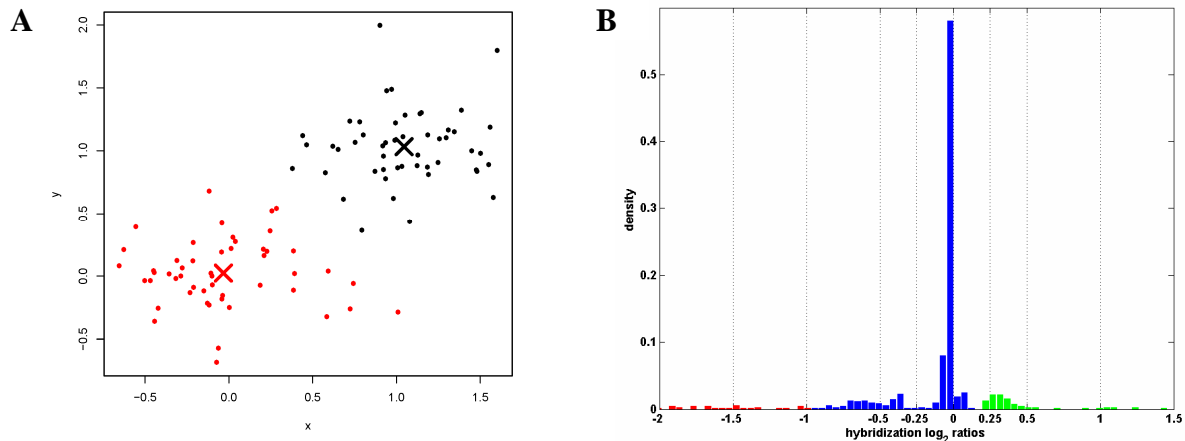


Figure 15 Kmeans clustering

A Kmeans clustering of two normally distributed variables x and y . Final centroids are marked with an X symbol, colours indicate which cluster the points were assigned to. **B** Clustering of copy number \log_2 ratios using Kmeans with $k=3$. Clusters revealed deletion, copy neutral and duplication components (as indicated in red, blue and green respectively). For display purpose, ratios are only shown from -2 to 1.5 . The data used are the same than those from **Figure 14B**, major differences with GMM modelling can be observed for the “deletion” component.

2.3.4 Hierarchical clustering

Hierarchical clustering is widely used in bioinformatics, in particular in microarray analysis to group genes and samples in a hierarchical manner. The output of hierarchical clustering is often a dendrogram (Figure 16). There are two types of strategies: divisive (top-down approach) where all observations start in a same and unique cluster, then this cluster is recursively split into smaller clusters, until the procedure reaches individual data points. The second strategy is called agglomerative (bottom-up approach): each observation starts as a singleton cluster, and then the most similar clusters are merged until there is only one cluster left.

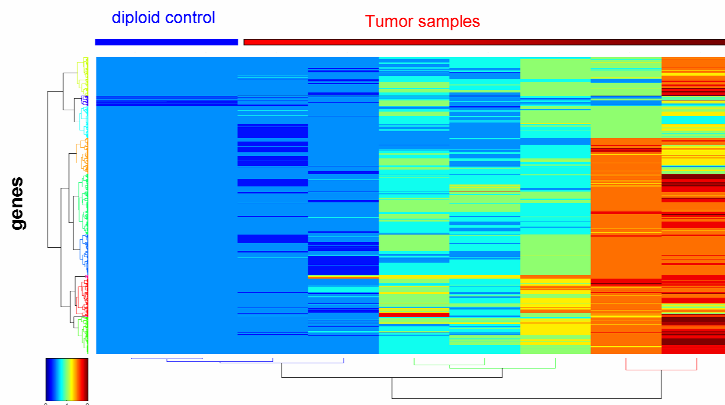


Figure 16 Hierarchical clustering

Hierarchical clustering based on gene copy number from two diploid genomes and seven aneuploid genomes. Genes are grouped based on their copy number profile across samples; similarly samples are grouped according to their ploidy status. Dendrograms indicates the hierarchy between the different elements. Colours in sub-trees represent clusters obtained after pruning the tree at a pre-defined height.

Distance metrics

An important step in any clustering is to compute a measure of similarity (or dissimilarity) between any two pairs of variables. This measure defines a distance matrix, with all possible pairwise distances. Several metrics are available to compute the distance between two variables \mathbf{a} and \mathbf{b} , each with n observations.

- Euclidean distance, $d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$

The Euclidean distance corresponds to the length of the line joining points \mathbf{a} and \mathbf{b} in the Euclidean space. It is the most frequently used distance in clustering.

- Manhattan distance, $d(a,b) = \sum_{i=1}^n |a_i - b_i|$

Manhattan, also known as Taxicab geometry, rectilinear distance, L_1 distance or city block, measures the distance along coordinate axes (e.g. like a taxi route in a building block from Manhattan).

- Mahalanobis distance, $d(a,b) = \sqrt{(a-b)^T C(a-b)}$

Here \mathbf{a} and \mathbf{b} are used in their vector form and C is the covariance matrix. The Mahalanobis distance takes into account the correlation between the variables and is often used to detect outliers (both in clustering and in linear regression).

- The Hamming distance, which is the number of positions that differs between two vectors (or strings) normalized by the length of these vectors
- Correlation (for e.g. Pearson) can also be used to defined a distance

$$d(a,b) = 1 - \frac{\sum_{i=1}^n (a_i - \mu_a)(b_i - \mu_b)}{\sqrt{\sum_{i=1}^n (a_i - \mu_a)^2} \sqrt{\sum_{i=1}^n (b_i - \mu_b)^2}} \text{ with } \mu_a \text{ and } \mu_b \text{ the mean of } \mathbf{a} \text{ and } \mathbf{b}.$$

Linkage methods

Once the matrix distance has been computed, rows and columns of this matrix can be merged into clusters: pairwise distances between clusters (or elements for the first merge) are computed, then the two clusters the closest to each other are merged together, and the procedure is re-iterated until there remains only one single cluster. This procedure is controlled with a linkage criterion:

- Complete linkage: distances between clusters are defined as the maximum distance between elements of each cluster
- Single linkage: as opposed to complete linkage, the cluster distances are defined as the minimum distance between clusters' elements
- Average, also referred as unweighted average distance (UPGMA) and corresponds to the mean distance between elements of each cluster
- Ward, clusters with the least increase in the total sum of squares are first merged together.

Evaluating the goodness of a tree

Both the choice of the distance metric and linkage method can lead to different dendrogram topology. Therefore as with regression and Gaussian Mixture Modelling, it is necessary to estimate the goodness of the model. In hierarchical clustering, the Cophenet correlation coefficient is useful to this matter. This coefficient corresponds to the correlation between distances from the final dendrogram and distances from the observations used to build the tree. It estimates how well the dendrogram fits the dissimilarities measured in the data and is defined as

$$c = \frac{\sum_{i < j} (Y_{ij} - \mu_y)(Z_{ij} - \mu_z)}{\sqrt{\sum_{i < j} (Y_{ij} - \mu_y)^2 \sum_{i < j} (Z_{ij} - \mu_z)^2}},$$

With Y_{ij} the distance between objects (nodes in the tree) i and j ; Z_{ij} is cophenetic distance between objects i and j (i.e. the height of the node at which these two objects are first joined); μ_y and μ_z correspond to the mean of \mathbf{Y} and \mathbf{Z} . A Cophenet correlation value close to one represents a perfect solution.

Extracting clusters from a tree

Hard clustering (i.e. identifying non-overlapping clusters) can be performed from a dendrogram. The most widely used approach consists in pruning the tree at an arbitrary height (see Figure 17). An alternative is to prune using an “inconsistency” coefficient. This coefficient characterizes the dissimilarity between a link in the dendrogram and its neighbours and is obtained by comparing the link height with the mean distance from other links at the same level in the hierarchy. High coefficient reflects smaller similarity between the objects connected by the link. This inconsistency-based approach relies on the cophenetic distance between sub-trees.

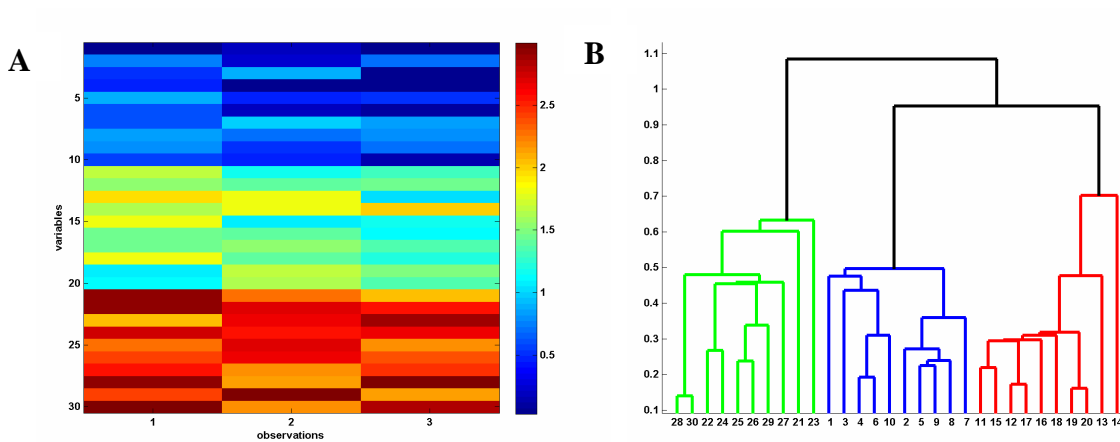


Figure 17 Hierarchical clustering and tree pruning

A Heatmaps showing 30 variables with 3 observations each, **B** Hierarchical clustering from data in A, using Euclidean distance and Single linkage. Y axis corresponds to Euclidean distance between variables. X axis display the label of the 30 variables from A. Coloured trees correspond to clusters obtained by pruning the dendrogram at 30% height.

2.3.5 Self-organizing maps

Self-organizing map (SOM) is a form of artificial neural network, where the network learns from the data and can perform discrete clustering on the same (learning) dataset. Once SOMs have been trained on a dataset, they can also be used for prediction (i.e. classification) on a new dataset. Artificial neural networks were inspired by biological nervous system. In statistics, neural networks are made with nodes (neurons) positioned in the space (a map). Such maps are usually a hexagonal or rectangular grid, made with neurons that can connect to their neighbours (Figure 18A). Connections between neurons are defined by their distance (Figure 18B) and by their weights computed from the data. The weight defines the neuron position in the map (Figure 18C) and is adjusted in a learning phase. This learning phase enables to re-organized neurons so that the ones with similar weights (and thus position) define the same cluster (Figure 18C and Figure 18D).

The most widely used feature in bioinformatics is the supervised clustering. In essence, SOMs with small number of neurons are similar to k-means clustering, with the major difference they do not requires specifying the number of clusters. This feature is particularly useful when the number of clusters is not known in advance and can range from a small to a .large number. In chapter 3, I used SOMs to merge SNPs with similar copy number into CNV regions. This merge was done on the whole cohort (more than 5,600 individuals) and was performed for both small regions (i.e. less than 50 SNPs) and longer regions. In this particular example, k-means clustering could have been performed but would have required iterations over different number of clusters and to use a metric (i.e. the BIC criteria, described previously for GMM clustering) to select the model with the best fit. However this k-means approach will still require initial guesses about the minimum and maximum number of expected clusters. By contrast, using a predefined grid (e.g. a 6*6 or 8*8 neuron grid) and by training it on the data, enables one to ignore how many clusters are expected. To ensure the learning phase is done using the “most relevant” information from the data, I performed a PCA analysis (described previously) to extract components explaining most of the variance. Then by training the SOM on the main PCs (i.e. those explaining at least 90% of the variance), I was able to cluster SNPs with similar copy number into CNV regions (see illustration in Figure 19). Although relatively

unused in statistical genetics, PCA and SOM combination has proved very powerful in other fields such as ECG analysis^{40,41}, forest fire risk classification⁴², text mining⁴³ and computer vision^{44,45}.

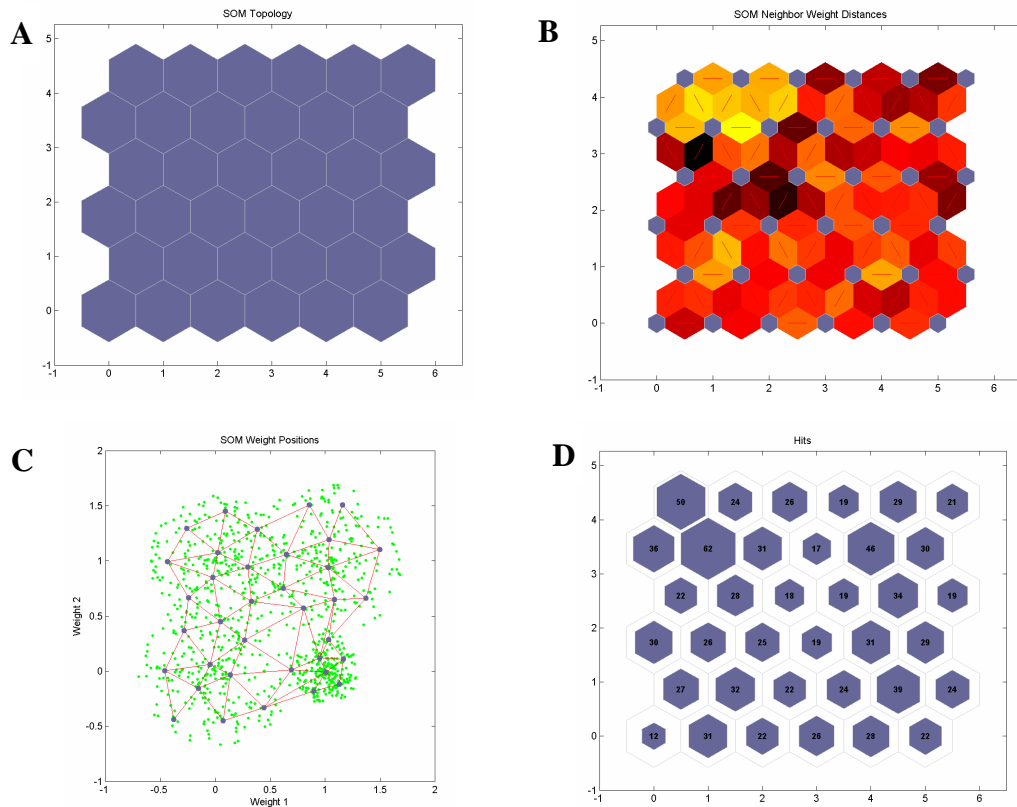


Figure 18 SOMs - adapted from the Matlab Neural Network toolbox

A SOM topology: a 6*6 grid. **B** Weight distances between neurons. Neurons are shown in blue, red lines indicate connections between neurons; darker (lighter) colours represent larger (smaller) distances between neurons. **C** Weight of each neuron (in blue), connections are shown in red, input data are shown in green. **D** Number of input data points attributed to each neuron.

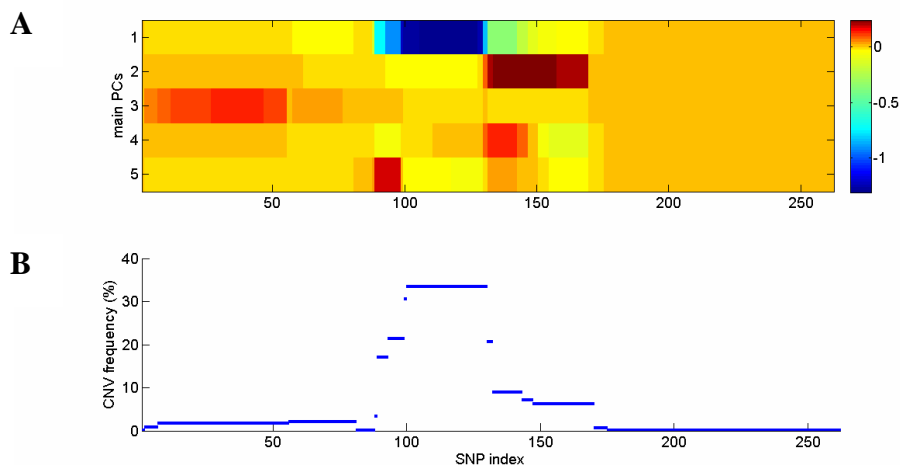


Figure 19 Merging SNPs into CNV regions using principal component analysis and SOMs

A shows a principal component analysis (PCA) on a local SNP window (chromosome 3 74.5-76.5Mb) across CoLaus individuals. The five main components are displayed on the Y axis and adjacent SNPs on the X axis. **B** shows CNV regions obtained from the SOM clustering of the main PCs. The Y axis represents CNV frequency in the CoLaus population (n≈5600).

2.4 References

1. Jazaeri, A.A. et al. Gene expression profiles of BRCA1-linked, BRCA2-linked, and sporadic ovarian cancers. *J Natl Cancer Inst* **94**, 990-1000 (2002).
2. Sotiriou, C. et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A* **100**, 10393-8 (2003).
3. Gergonne, J.D. The application of the method of least squares to the interpolation of sequences. *Historia Mathematica* **1**, 439-447 (1974).
4. Saville, D.J. & Wood, G.R. *Statistical methods : the geometric approach*, xv, 560 p. (Springer, New York, 1997).
5. Diskin, S.J. et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res* **36**, e126 (2008).
6. Cleveland, W.S. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* **74**, 829-836 (1979).
7. Cleveland, W.S. & Devlin, S.J. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association* **83**, 596-610 (1988).
8. Stigler, S.M. Gergonne's 1815 paper on the design and analysis of polynomial regression experiments. *Historia Mathematica* **1**, 431-439 (1974).
9. Smyth, G.K. & Speed, T. Normalization of cDNA microarray data. *Methods* **31**, 265-73 (2003).
10. Smyth, G.K., Yang, Y.H. & Speed, T. Statistical issues in cDNA microarray data analysis. *Methods Mol Biol* **224**, 111-36 (2003).
11. Marioni, J.C. et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol* **8**, R228 (2007).
12. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93 (2003).
13. Budinska, E., Gelnarova, E. & Schimek, M.G. MSMAD: a computationally efficient method for the analysis of noisy array CGH data. *Bioinformatics* **25**, 703-13 (2009).
14. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-72 (2004).
15. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-63 (2007).
16. Lai, W.R., Johnson, M.D., Kucherlapati, R. & Park, P.J. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763-70 (2005).
17. Conrad, D.F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* (2009).
18. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195-197 (1981).
19. Price, T.S. et al. SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res* **33**, 3455-64 (2005).

20. Pique-Regi, R. et al. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* **24**, 309-18 (2008).
21. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **14**, 755-63 (1998).
22. Eddy, S.R., Mitchison, G. & Durbin, R. Maximum discrimination hidden Markov models of sequence consensus. *J Comput Biol* **2**, 9-23 (1995).
23. Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. & Durbin, R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**, 320-2 (1998).
24. Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G. & Jain, A.N. Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis* **90**, 132-153 (2004).
25. Marioni, J.C., Thorne, N.P. & Tavare, S. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics* **22**, 1144-6 (2006).
26. Guha, S., Li, Y. & Neuberger, D. Bayesian Hidden Markov Modeling of Array CGH Data. *Journal of the American Statistical Association* **103**, 485-497 (2008).
27. Greenman, C.D. et al. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **11**, 164-75 (2010).
28. Colella, S. et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* **35**, 2013-25 (2007).
29. Coin, L.J. et al. cnvHap: an integrative population and haplotype-based multiplatform model of SNPs and CNVs. *Nat Methods* **7**, 541-6 (2010).
30. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-74 (2007).
31. Huang, J. et al. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics* **1**, 287-99 (2004).
32. Viterbi, A.J. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *Ieee Transactions on Information Theory* **13**, 260-+ (1967).
33. Baum, L.E., Petrie, T., Soules, G. & Weiss, N. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics* **41**, 164-171 (1970).
34. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological* **39**, 1-38 (1977).
35. Rabiner, L.R. A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition. *Proceedings of the Ieee* **77**, 257-286 (1989).
36. Pearson, K. {On lines and planes of closest fit to systems of points in space}. *Philosophical Magazine* **2**, 559-572 (1901).
37. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98-101 (2008).
38. Geman, S. & Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *Ieee Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741 (1984).
39. Casella, G. & George, E.I. Explaining the Gibbs Sampler. *The American Statistician* **46**, 167-174 (1992).
40. Stamkopoulos, T., Diamantaras, K., Maglaveras, N. & Strintzis, M. ECG analysis using nonlinear PCA neural networks for ischemia detection. *Ieee Transactions on Signal Processing* **46**, 3058-3067 (1998).

41. López-Rubio, E., Muñoz-Pérez, J. & Gómez-Ruiz, J.A. A principal components analysis self-organizing map. *Neural Networks* **17**, 261-270 (2004).
42. Annas, S., Kanai, T. & Koyama, S. Principal Component Analysis and Self-Organizing Map for Visualizing and Classifying Fire Risks in Forest Regions. *Agricultural Information Research* **16**, 44-51 (2007).
43. Marinai, S., Faini, S., Marino, E. & Soda, G. Efficient word retrieval by means of SOM clustering and PCA. *Document Analysis Systems Vii, Proceedings* **3872**, 336-347 (2006).
44. Lopez-Rubio, E., Ortiz-de-Lazcano-Lobato, J.M. & Lopez-Rodriguez, D. Probabilistic PCA Self-Organizing Maps. *Ieee Transactions on Neural Networks* **20**, 1474-1489 (2009).
45. Alba, J.L., Pujol, A. & Villanueva, J.J. Novel SOM-PCA network for face identification. *Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation Iv* **4479**, 186-194

3 Identification and validation of Copy Number Variants using SNP genotyping arrays from a large clinical cohort

Within the group of Pr. Sven Bergmann and in collaboration with the CHUV (Pr. Jacques Beckmann, Pr. Peter Vollenveider and Pr. Gérard Waeber) and GlaxoSmithKline (Dr. Vincent Moozer), I have been in charge of detecting CNVs from a large medical cohort named CoLaus. My contribution has varied from low-level analysis with the normalization of genotyping data to developing, applying and comparing CNV detection methods. With the exception of the Gaussian Mixture Model method, where I received help from Dr. Zoltan Kutalik, I developed and produced myself all the methodologies and results that are presented in this chapter. The first part of this chapter summarizes my work; the detailed analysis follows in the form of a manuscript that was submitted to an international peer-reviewed journal.

CoLaus is a population-based health survey to study the genetics of hypertension and cardiovascular disease ¹. More than 6,000 individuals (35-75 years old) from the Lausanne area participate in the study. Over 150 phenotypic measurements (e.g. blood pressure, lipid levels, metabolic traits...) have been collected at the CHUV; in addition, genotyping has been carried out on Affymetrix 500K SNP chips ². A number of SNP-based genome-wide association studies that employed the CoLaus data have already been reported ³⁻¹⁰. Although so far there is no evidence for common CNVs contributing significantly to the kind of clinical phenotypes measured in CoLaus phenotypes ¹¹, the number of rare CNVs and their contribution to clinical phenotype remains unclear. My main objective was to identify both common and rare CNVs in the CoLaus population and subsequently investigate their possible contribution to clinical phenotypes.

Since the publication of the first genome-wide CNV analysis in the general population ¹², there has been tremendous development of new methods for CNV analysis. However, to date, there are still no gold standards, especially for Affymetrix 500K arrays. At the beginning of my thesis, there were even fewer publicly available algorithms for analysing SNP arrays. Most of these methods have been developed and trained for CGH data, which are more reliable than SNP arrays for CNV detection. Among the SNP dedicated software (e.g. dChip ¹³, CNAG ¹⁴, GEMCA ¹⁵), only available for Windows operating system, none could scale for the analysis of a

very large dataset. Only CNAT ¹⁶ was available as UNIX binaries thus the computation could be distributed on nodes from the local high-performance computing center (Vital-IT), but there were few papers evaluating the CNAT performance.

Therefore I analysed the complete CoLaus dataset (n=5,612 individuals) with the two available CNAT implementations and tested their respective performance using technical replicates. These two methods were producing highly different results and their relative performance was not completely clear. In this context and with the help from Dr. Zoltan Kutalik, I developed a novel method based on a Gaussian Mixture Model. I also decided to use Circular Binary Segmentation ^{17,18}, a *state-of-the-art* method for CGH analysis. Based on the results from these four methods, I studied how the predicted CNVs coincide with previously reported variants. I also investigated the concordance in predicting CNVs in a *sub-sample* of individuals that were also genotyped on the Illumina platform. Finally I compared the sensitivity and specificity of the different approaches using related individuals. These validation metrics offer the advantage to being applicable to any other population-based cohort.

In addition to comparing CNV methods, I addressed the problem of integrating individual CNV predictions at the population-level (i.e. identifying copy number polymorphisms in the general population). To do so, I developed two procedures, a naïve approach, that relies on combining regions with identical CNV profiles, and a more elaborated one based on principal component analysis to combine markers that explain most of the copy number variance at a given locus.

Finally my work provides an extensive collection of both rare and common CNVs, which have recently played a key role in a study demonstrating the penetrance of rare variants in the etiology of morbid obesity (see Chapter 4).

3.1 Abstract

Genotypes obtained with Affymetrix 500K and Illumina 550K SNP arrays have been extensively used in many large case-control or population-based cohorts for SNP-based genome-wide association studies for a multitude of traits. Yet, these genotypes capture only a small fraction of the variance of the studied traits. Genomic structural variants (GSV) such as Copy Number Variation (CNV) may account for part of the missing heritability, but their comprehensive detection requires either next-generation arrays or sequencing. Sophisticated algorithms that infer CNVs by combining the intensities from SNP-probes for the two alleles can already be used to extract a partial view of such GSV from existing data sets.

Here we present several advances to facilitate the latter approach. First, we introduce a novel CNV detection method based on a Gaussian Mixture Model. Second, we propose a new algorithm for combining copy-number profiles from many individuals into consensus regions. We applied both our new methods as well as existing ones to data from 5,612 individuals from the *Cohorte Lausanne* who were genotyped on Affymetrix 500K arrays. We developed a number of procedures in order to evaluate the performance of the different methods. This includes comparison with previously published CNVs as well as using a replication sample of 239 individuals, genotyped with Illumina arrays. We also established a new evaluation procedure that employs the fact that related individuals are expected to share their CNVs more frequently than randomly selected individuals. The ability to detect both rare and common CNVs provides a valuable resource that will facilitate association studies exploring potential phenotypic associations with CNVs.

Our new methodologies for CNV detection and their evaluation will help in extracting additional information from the large amount of SNP-genotyping data on various cohorts and use this to explore structural variants and their impact on complex traits.

Availability: <http://www2.unil.ch/cbg/index.php?title=GMM>

3.2 Author Summary

The genomes of any two individuals can differ, for example, by single nucleotide polymorphisms (SNPs) or larger structural variants usually referred to as copy number variants (CNVs). SNPs have been intensively used to investigate potential links with disease but so far, collectively, explain only a small fraction of the genetic variance. The link between CNV and disease is well known but their contribution to complex disease is not yet fully understood. This is in part due to the fact, that genetic studies aiming to uncover such associations have been using SNP genotyping arrays which were not designed for CNV analysis. Yet some information on CNVs is hidden within these datasets. We have developed a novel CNV detection method and compared it with three other established methods using a variety of tests on a large genotyping clinical cohort. Our methods and testing procedures provide some new tools that will help for CNV analysis on existing cohorts and provide a first step to investigate the contribution of CNVs to common and complex diseases.

3.3 Introduction

Genetic variation in the human genome takes many forms ranging from large chromosome anomalies to single nucleotide polymorphisms (SNPs). Deletion, insertion and duplication events giving rise to copy number variations (CNVs) have been found genome-wide in humans^{12,19-25} and other species²⁶⁻²⁹. Genomic variants can impact both somatic and germ-line genetics. The link between CNVs and inherited diseases is now solidly established (e.g.³⁰⁻³²), and copy number plasticity is typical of cancer cells³³. Such genomic variability was identified more than a decade ago using array-based comparative hybridization^{34,35} and was known to exist for much longer from cytogenetic studies or Southern blots. It has been demonstrated that CNVs near oncogenes or tumor suppressor genes can affect gene expression levels or result in the expression of chimeric fusion genes^{35,36}. However, the number and positions of rare CNVs in the human genome are still likely to be underestimated and their contribution to common complex diseases such as diabetes or obesity is unclear. Very recent results demonstrate that rare variants can have very high penetrance in the etiology of morbid obesity³⁷.

CoLaus (Cohorte Lausannoise) is a population-based survey started in 2003 to study risk factors for hypertension and cardiovascular diseases ¹. 6,188 individuals (35-75 years old) from the Lausanne area in Switzerland participated in the study. All individuals were genotyped on Affymetrix 500K SNP chips, and a fraction of these were also genotyped on the Illumina platform ². A number of SNP-based genome-wide association studies (GWAS) that employed the CoLaus data have already been reported ^{3-7,10}. Although many other large cohorts including thousands of individuals have been genotyped for SNPs ^{3,7,8}, very few have reported CNV maps ^{38,39}.

It is important to emphasize that most SNP arrays used so far in GWAS of clinical cohorts were not designed for CNV (dosage) detection, but only to call the three possible genotypes of SNPs. Nevertheless, by combining the intensities of the two alleles for a given SNP, it is possible to also obtain information on the copy number state of the SNP locus. However, this is challenging for several reasons: Firstly, when analyzing very large datasets (with several thousands of individuals), it is likely that experiments were conducted at different times and/or by different laboratories, which often introduces strong batch effects for the raw intensities. Thus the first challenge in CNV calling is to ensure proper normalization of these raw data. Secondly, due to the large noise in the SNP probe intensities in these arrays (even after batch effects have been corrected for), the estimates of copy numbers for a given locus (SNP) are not very robust. Thus more reliable prediction can only be made by integration of intensities from several neighboring loci, a strategy that is employed by many different CNV detection methods ^{14-18,40}. However, this approach makes CNV detection difficult (and sometimes completely fails) in regions with low SNP density. Thirdly, while some methods take advantage of the signals from a single or a group of SNPs across the population to predict CNV regions for each individual ⁴¹⁻⁴³, there are very few methods to merge individual CNV predictions into regions at the population level: Redon et al. ¹² merged CNVs based on the extent of their overlap, whereas Itsara et al. ³⁸ manually annotated complex regions.

In the current study we followed two main goals: First we performed an extensive survey of candidate CNVs in the CoLaus cohort as detected by SNP genotyping microarrays. We provide a large dataset that can serve as a resource for other studies elucidating human structural variants, and for future association studies of

CNVs with the clinical phenotypes measured in CoLaus. Second since the methods for detecting individual CNV profiles and merging those into consensus regions have not yet been well established, we developed new algorithms for CNV calling and merging, and devised novel techniques to evaluate and compare them with existing methods. Specifically, we compared three existing CNV detection methods with our new method that uses a Gaussian Mixture Model to estimate the copy number dosage at each SNP of each individual. This new method was successfully applied to both Affymetrix and Illumina arrays; and is not restricted to SNP array analysis. We also developed two merging strategies, which were applied to create a map of CNV regions for each of the four CNV detection methods. We studied how CNVs predicted by the various algorithms coincided with previously reported variants. We also investigated the concordance in predicting CNVs in a subsample of individuals that were additionally genotyped on the Illumina platform. Finally we compared the sensitivity and specificity of the different approaches using related CoLaus individuals which are expected to share more CNVs than unrelated individuals. Based on these criteria, we demonstrated that our new method outperforms two established CNV detection algorithms and has higher sensitivity than a third method.

3.4 Methods

The implementation of the Gaussian Mixture Model is publicly available at <http://www2.unil.ch/cbg/index.php?title=GMM>. The algorithm has been implemented in Matlab, both the source code and a compiled version for UNIX 64-bit operating systems are available. The PCA-merging algorithm has also been written in Matlab and the source code is available upon request.

3.4.1. Ethics Statement

The CoLaus study was approved by the institutional review boards of the University of Lausanne, and written consent was obtained from all participants.

3.4.2. CNV calling

Copy Number Analysis Tool

We used the Affymetrix GeneChip Genotyping Analysis Software (GTYPE ²) to extract, normalize and summarize intensities for both alleles of each SNP. We normalized our data using a sketch-quantile distribution of 50k PM Probes and summarized the intensities using the plier method in RMA mode. (Detailed information can be found in the GTYPE manual.) We first normalized the CoLaus samples versus 30 unrelated CEU Hapmap ⁴⁴ individuals. Then we used the Affymetrix Copy Number Analysis Tool (CNAT ¹⁶) to attribute a copy number (CN) state to each SNP of all CoLaus individuals with the following encoding: 0 for homozygous deletion, 1 for hemizygous deletion, 2 for copy neutral, 3 for simple gain and 4 for multiple gains. It should be noted that such discrete copy number classification is relative to the median CN in the references. CNAT performs additional normalizations such as PCR bias correction; inter-array normalization when combining NSP and STY arrays; a Gaussian smoothing function to increase the signal-to-noise ratio; and combines allelic intensities into a CN ratio (CNR). CNAT has two HMM implementations (*CNAT.total* and *CNAT.allelic*), which mainly differ by the way they compute the CN ratios (equation 1 and 2).

$$CNR(CNAT.total) = \log_2\left(\frac{S_A + S_B}{R_A + R_B}\right) \quad (1)$$

$$CNR(CNAT.allelic) = \log_2\left(\frac{S_A}{R_A}\right) + \log_2\left(\frac{S_B}{R_B}\right) \quad (2)$$

In the above equations, S refers to the intensity of the test sample (of an individual) and R to the (mean) intensity of the reference panel; A and B refer to the SNP alleles.

The *CNAT.allelic* approach uses the sum of the logs of the allelic signals and is more sensitive to subtle allelic CN changes than *CNAT.total*.

Through QC analyses, we discovered an important batch effect related to the fact that samples were processed by four genotyping centers. Therefore we normalized data from each genotyping center independently and tested the improvement as a function of the number of references used (see Supplementary Data and Supplementary Figure 4). Although Affymetrix suggests that 25 samples are enough for normalization (see CNAT manual), we established that in the presence of strong experimental biases, using many more references performed significantly better (see Supplementary Methods). Thus we re-applied the two CNAT implementations to ratios normalized within each genotyping center and using 280 references, producing much more reliable results than the initial normalization (with 30 references).

Aroma normalization

In parallel to the normalizations performed using GTYPE, we normalized the CoLauS data with the Aroma.Affymetrix framework ⁴⁵. Normalizations were done independently for datasets from each genotyping center with at least 336 individuals (since the Aroma.Affymetrix requires a lot of I/O operations, which can cause a severe drop of the computational performance on shared-network discs, this number of references was decided for optimal computational performances while keeping this number large enough for batch effects correction (see Supplementary Methods). Normalization steps included Allelic Cross-talk calibration ^{46,47} to correct for differences between SNP alleles; intensity summarization using Robust Median Average and correction for any PCR amplification bias inherent to the Affymetrix SNP platform. To estimate the CNR for a given sample at a given SNP or CN probe, we

computed the \log_2 ratio of the normalized intensity of this probe divided by the median across all the samples from the same batch.

Circular Binary Segmentation

Circular Binary Segmentation (CBS) has been described as a state-of-the-art segmentation algorithm^{17,18}; it identifies change points using maximal t-statistics and assesses segment significance with permutations. We applied CBS on the CNRs as obtained by the Aroma.Affymetrix framework. The distribution of \log_2 ratios (Supplementary Figure 6), revealed that segments with \log_2 ratios greater than 0.25 or lower than -0.25 were outliers (i.e. ratios greater than 3rd quartile + 1.5 * interquartile range or lower than 1st quartile - 1.5 * interquartile range). A clustering using a three component Gaussian Mixture Model confirmed such data separation. Thus we decided to classify regions having a mean \log_2 ratio greater than 0.25 as gains and regions with mean \log_2 ratios lower than -0.25 as losses.

Gaussian mixture models

Raw copy number ratios were smoothed along physical position using Loess filtering with a 41-probe window size. Next, a four component Gaussian mixture model (one component for each of the following copy number states: deletion, copy-neutral, 1 and 2 additional copies) was fitted to the smoothed copy number ratios with a constraint on the differences between the mixture means. The means of the mixture components were decided not to be fixed as the population mean may not necessarily be two copies. Then, for each individual we determined the probabilities for each of these copy number states (see Supplementary Figure 7). The expected copy number was finally assigned as the weighted sum of individual dosage probabilities; for example a SNP with probabilities: 1% for CN=1, 9% for CN=2, 85% for CN=3 and 5% for CN=4, would have a CN dosage value equal to $2.94 (1*0.1 + 2*0.9 + 3*0.85 + 4*0.05)$.

Illumina CNV analysis

A subset of CoLaus individuals were analyzed on Illumina arrays (550K version 1 & 3, 1M⁴⁸). Intensities were normalized within BeadStudio using 120 Hapmap samples. Only SNPs that could be remapped to the 550K version 3 array (genome assembly build NCBI 36) were used for subsequent analysis. Only 239 samples with a genotyping call rate greater than 99.9% and whose QC metrics satisfied standard Illumina recommendations were used. To do the CNV calling, we applied our mixture Gaussian model (including the Loess filtering), then merged CNVs with the PCA approach (see below) and excluded any singleton regions.

3.4.3. CNV merging

Simple merge

Our raw CN data can be represented as a matrix where each element represents the Copy Number status for all individuals (rows) and all SNPs (columns). The “simple merge procedure” consists of combining adjacent SNPs that share the same CN profile across the whole population. This is equivalent to merging strictly identical SNP columns. To avoid creating CNV regions that would encompass long genomic regions with low SNP density, we applied the requirement that two SNPs in the same CNV region should not be further away than 500Kb from each other. This rule did not apply to regions where all SNPs were copy neutral.

PCA merge

The PCA merge is a novel merging algorithm for CNV profiles. It includes four steps: (1) The genome is partitioned into smaller regions, whose boundaries are a long stretch of SNPs in the diploid state; (2) For each of these regions, a principal component analysis is performed analyzing the regional (clipped) CNV profiles (Supplementary Figure 1); (3) Only the few largest components that explain at least 90% of the total variance are then used to train a self-organizing map (SOM) to cluster SNPs with similar variance together; (4) Strictly adjacent SNPs within a same cluster are merged into CNV regions.

3.5 Results

3.5.1. Identification of Copy Number Variants in CoLaus

To detect CNVs in CoLaus, we applied four different CNV detection algorithms to the data from 5,612 Caucasians generated with Affymetrix 500K microarrays: two implementations of the Copy Number Analysis Tool (CNAT¹⁶) that integrate the SNP intensities by summing their raw (*CNAT.total*) or log-transformed (*CNAT.allelic*) values; Circular Binary Segmentation (CBS^{17,18}) and our own algorithm based on a Gaussian Mixture Model, to which we refer subsequently as GMM. We restricted our analysis to autosomes allowing us to use a mixture of males and females as the reference panel. Using these four methods, we assigned copy number values to each SNP and each CoLaus individual. (The CBS method only returns segments and their mean signal intensity, which we used to identify SNPs within candidate regions for CNVs if the corresponding ratio was below (loss) or above (gain) a certain threshold, see Methods for more details.)

In a second step we attempted to reduce the complexity of these CNV profiles by merging adjacent SNPs that contained highly redundant information into CNV regions. The first method (“simple merge”) joins neighboring SNPs that have identical copy number values across all CoLaus participants. This simple approach already significantly reduced the number of SNPs (for example, it compresses 490K SNPs into 8,000 regions for *CNAT.total* and into 40K for *CBS*). However, this simple scheme leaves the boundaries of CNVs fragmented. Thus we devised a refined method, which is based on a principal component analysis (PCA). The PCA identifies orthogonal components explaining a significant (e.g. 90%) fraction of the variance that are subsequently used to cluster SNPs in CNV regions (see Methods for details).

Next, we excluded any CNV regions found in fewer than five individuals. We distinguish between Copy Number Polymorphisms (CNP, CNVs with a frequency greater than 1% in the population) and Copy Number Variant Regions (CNVRs, CNVs with population frequency below 1%). The numbers of CNPs and CNVRs predicted by the four different methods and the two merging methods are shown in Figure 1. *CNAT.total* and *CBS* are conservative methods that generate significantly fewer regions than *CNAT.allelic* and *GMM*. The simple merging procedure produces many small regions (<1kb or single SNPs) which are commonly integrated into fewer

larger regions with the PCA-based method. The PCA-based merging method is able to reduce the total number of regions by 35%, 53%, 67% and 70% for *GMM*, *CNAT.total*, *CNAT.allelic* and, *CBS*, respectively.

The fraction of the genome effectively covered by these regions is reported in Supplementary Table 1. Although *GMM* produces many more CNPs than the other methods, they only cover about 2.4% of the autosomes. *CNAT.allelic* predictions for CNPs cover 12.4% of the autosomes, while *CBS* and *CNAT.total* cover only 1.5% and 0.7% respectively. We also checked the coverage with rare variants (CNVRs), *GMM* had the lowest autosomal coverage of only 9.8%, whereas *CBS* had the highest with 42.4%. *CBS* predictions for CNPs are rather conservative in the sense that CNPs found with other methods are found for fewer individuals when using *CBS* (thus much higher genome coverage for CNVRs). Supplementary Figure 5 shows the CNV profile on chromosome 1 as predicted by the different methods and illustrates the limited ability of *CBS* to detect CNPs (despite using optimized thresholds when classifying *CBS* segments; see Methods for details).

We computed the intersection between the four methods using CN prediction from 60K independent autosomal SNPs (SNPs that were not in LD in the CEU population, see Supplementary Methods) (see the Venn diagrams in Supplementary Figure 8). Only 2.3% of the SNPs composing CNPs were validated with at least three methods (10% with at least two methods) (see Supplementary Table 4). By contrast, 23.5% of the SNPs in CNVRs were found in at least three methods and this number reached 55.3% for at least two methods. We also computed pair-wise comparison between the CNV methods (Supplementary Table 5). The maximal intersection between two methods is 47% and corresponds to the comparison between all CNVs from *GMM* and *CBS*. Such relatively low overlaps are not uncommon with CNV analysis from SNP genotyping arrays and underline the need for proper replication of any CNV predictions.

In order to evaluate the different detection and merging algorithms we compared the various outputs with some reference. In the following we compare the different methods using three different approaches: (i) A comparison with known CNVs from a public database, (ii) A cross-platform comparison using a subset of samples that were also genotyped on the Illumina platform, and (iii) similarity of related individuals with respect to their CNV profiles.

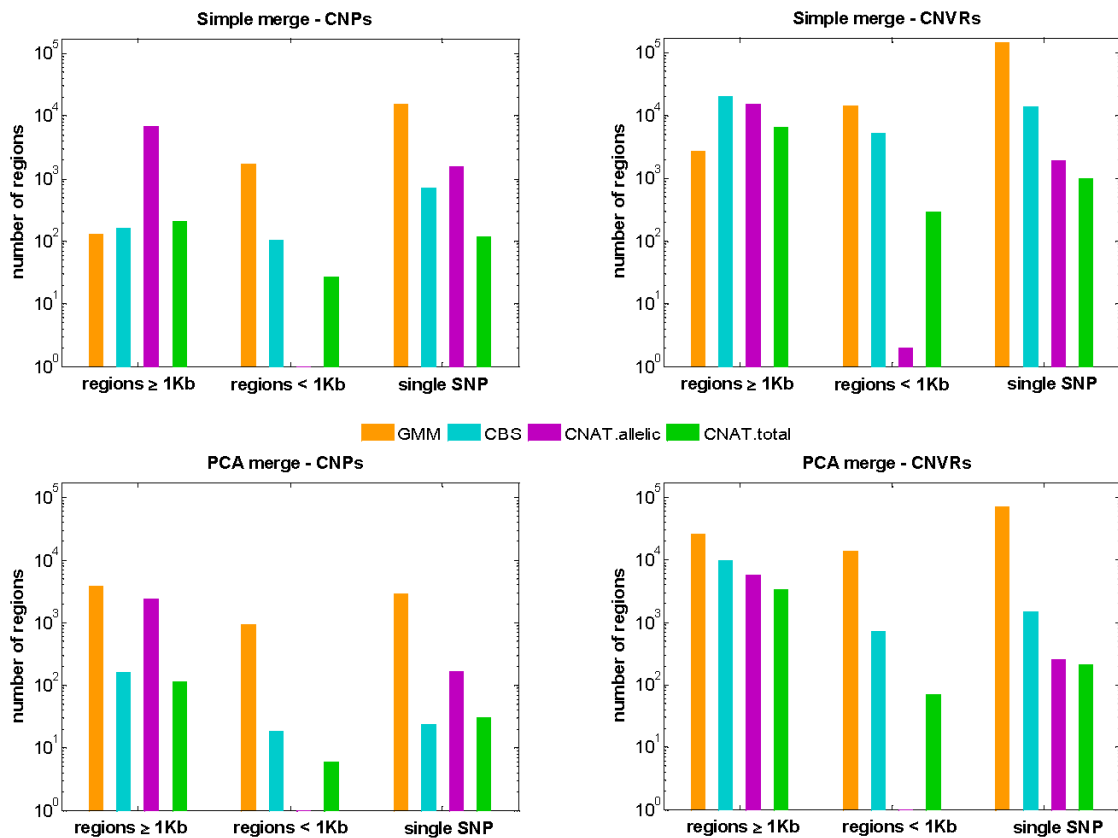


Figure 1 Counts of CNVs identified with the different methods

Copy number variants (CNVs) were detected with four different algorithms (see legend) using data generated by Affymetrix 500K SNP arrays for the Cohorte Lausanne ($n \approx 5600$). Adjacent SNPs with similar Copy Number profiles were merged into CNV regions using two different approaches: one based on principal component analysis (PCA, bottom panel) and a more simple approach that only merges SNPs with identical profiles (top panel). Copy number polymorphisms (CNPs, i.e. CNVs with population frequency above 1%) are shown on the left. Copy number variant regions (CNVRs, i.e. CNVs with population frequency below 1% but seen for at least five individuals) are shown on the right. In each plot, CNV counts are segregated according to their size.

3.5.2. Comparison with known CNVs

The Database of Genomic Variants (DGV ¹⁹) is a curated catalogue of structural variation in the human genome. We downloaded its content (release 7, March 2009) and kept only CNVs discovered from SNP or CGH arrays (BAC and ROMA arrays were excluded). We added to this dataset CNVs from the high resolution CNV project ¹¹. This combined dataset of “known” CNVs included 17,804 autosomal CNVs, whose size ranged from 1kb to 3Mb.

We then computed the overlap between this reference dataset and CNVs generated by each prediction method from the CoLaus data (Figure 2A). The overlap is reported as the Jacquard coefficient, which is the ratio between the size of the intersection and the union of two CNVs. A ratio close to one implies that the two CNVs have very similar boundaries; a ratio close to zero indicates a negligible overlap (or no overlap at all if the ratio is equal to zero) and intermediate values correspond to partial overlap (including the case where a small CNV is encompassed by a larger one). Since DGV contains CNVs from many fewer individuals than the CoLaus dataset, it was important to compare the distribution of overlaps with the CNVs generated by the different methods in a controlled setting. Therefore we computed for each method the expected overlap using reshuffled data from 1,000 permutations. Estimated *p*-values for observing more or less than expected CNVs with a given overlap are shown in Figure 2A (see Supplementary Table 2 for the corresponding *t*-statistics), and the relative excess of observed or expected counts is shown in Figure 2B. We observed that all prediction methods were enriched with respect to the controls for known CNVs (all Jacquard coefficient bins strictly above 50%) and depleted for novel CNVs (Jacquard coefficient of zero).

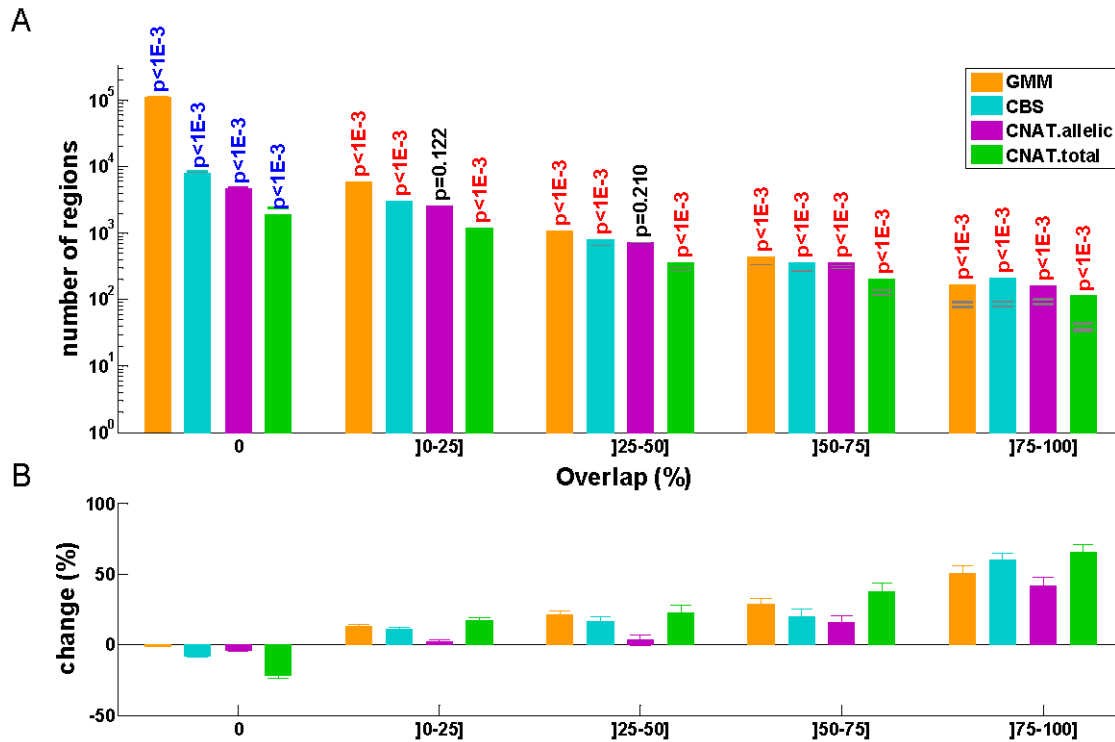


Figure 2 Overlap between CNVs identified from CoLaus and published CNVs

A) Counts of CNVs with different methods (see legend) are segregated according to their overlap with CNVs published in the Database of Genomic Variants. Overlap is measured by the Jacquard coefficient, i.e. the ratio between the intersection and the union of two groups of CNVs. Expected counts from (1000 times) reshuffled data are shown in gray (extending over one standard deviation). Estimated p-values are indicated for significant enrichment (red) or depletion (blue), with respect to these controls. Non significant p-values ($\alpha > 1\%$) are shown in black.

B) Percentage of changes between observed and expected counts from A. Error bars indicate \pm one standard deviation

3.5.3. Validation with Illumina arrays

DNA from a subset of 239 CoLaus individuals was assayed on the Illumina SNP platform, which uses a different technology from Affymetrix and also provides a different SNP content. In order to obtain a validation set of CNVs, we applied *GMM* and the PCA-based merging algorithm to these data. Note that *CNAT* is specifically designed for Affymetrix data so it could not be used here. To validate our CNV datasets as predicted from the Affymetrix arrays, we selected those CNVs containing at least one individual that had been assayed on the Illumina arrays. Next, we computed for the overlap between those selected Affymetrix CNVs and the validation CNV collection from the Illumina arrays (Figure 3).

From our overlap analysis, we found that *CNAT.allelic* predictions were not significantly different from random predictions (according to the controls using

reshuffled data). This indicates that *CNAT.allelic* is too permissive and that the vast majority of its predictions are likely to be false positives. In contrast, *CNAT.total* had a better specificity than *CNAT.allelic* but identified much fewer CNVs compared to other methods (*CBS* and *GMM*). Both *CBS* and *GMM* performed well (showing depletion of CNVs unique to the Affymetrix data and enrichment of common CNVs). Interestingly, *GMM* predicted many more CNVs than *CBS* and the bias with respect to predictions from reshuffled data was much stronger than for all the other methods (Supplementary Table 3). We also performed the above analyses independently for CNPs and CNVRs (both against DGV and the Illumina data, see Supplementary Figure 2) and arrived at the same conclusions.

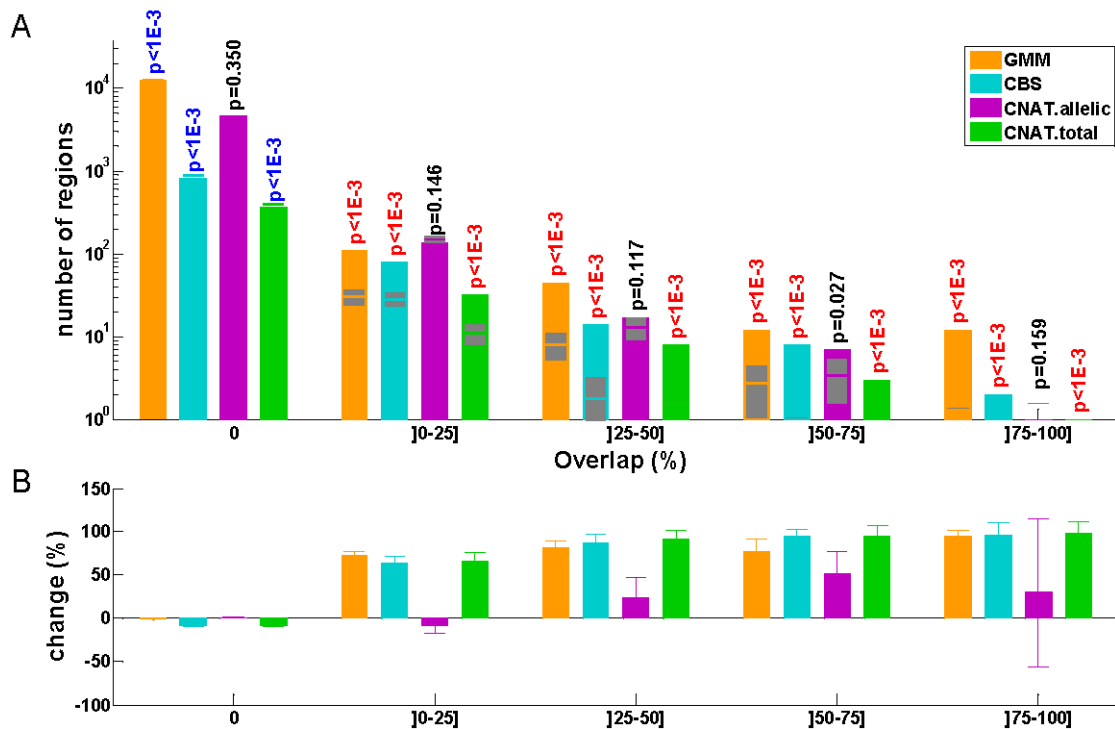


Figure 3 Overlap between CNVs identified from Affymetrix and Illumina data

A) Counts of CNVs identified with different methods (see legend) from Affymetrix data are segregated according to their overlap with CNVs identified from Illumina data. The Illumina panel includes a subset of 239 CoLaus individuals. Affymetrix-based CNVs, which did not include at least one individual from the Illumina panel, were excluded from the analysis. Overlap is measured by the Jacquard coefficient, i.e. the ratio between the intersection and the union of two groups of CNVs. Expected counts from (1,000 times) reshuffled data are shown in gray (extending over one standard deviation). Estimated p-values are indicated for significant enrichment (red) or depletion (blue), with respect to these controls. Non significant p-values ($\alpha > 1\%$) are shown in black.

B) Percentage of changes between observed and expected counts from A. Error bars indicate \pm one standard deviation

3.5.4. Predicting relatedness between individuals based on their CNV profile

Analysis of the CoLaus SNP-profiles revealed that five individuals had been genotyped twice and the cohort also included 157 pairs of first-degree relatives (either sibling or parent-offspring relationships). Using this information, we investigated whether predicting the relationship between these individuals would be feasible using exclusively their inferred CNP profiles. To this end we computed the Euclidean distance between the individuals belonging to 162 related pairs and between individuals in 2,000 randomly selected pairs. Knowing the true relationship status, we computed Receiver Operating Characteristic (ROC) curves for each CNV prediction method and for each merging approach (Figure 4). To evaluate the robustness of the ROC curves we reiterated the analysis 100 times with randomly chosen pairs of unrelated individuals.

All methods had significant prediction power with Area Under the Curve (AUC) values >0.5 . Only the relaxed CNV detection method *CNAT.allelic* did not show a significant difference between the PCA-based and the simple merging approach (both methods had an AUC ≈ 0.6). Interestingly for all other methods, there was a clear performance advantage of the PCA-based over the simple merging method. Also, these three CNV detection methods, post-processed with the PCA approach, performed better than *CNAT.allelic*. *GMM* and *CNAT.total* had the best AUC (0.71). We checked whether changing the CNV frequency filter and excluding small regions ($<1\text{kb}$) would improve the performance (Supplementary Figure 3). For all methods, there was no significant difference when excluding or keeping such small regions. For *CNAT.allelic*, there was some small improvement when increasing the filter on the CNV frequency, whereas there was no significant change for *CNAT.total* and *CBS*. Apparently, the rather small number of CNV predictions by *CNAT.total* are of good quality for predicting relatedness as reflected by the high AUCs (>0.7). Indeed, *GMM*, which is less conservative, profits from using a filter on CNV frequency significantly, improving its AUC. This improvement is particularly strong in combination with the PCA merge (giving an AUC up to 0.725, which is the best value we obtained across all methods (*CNAT.total* AUCs being slightly lower, see Supplementary Figure 3).

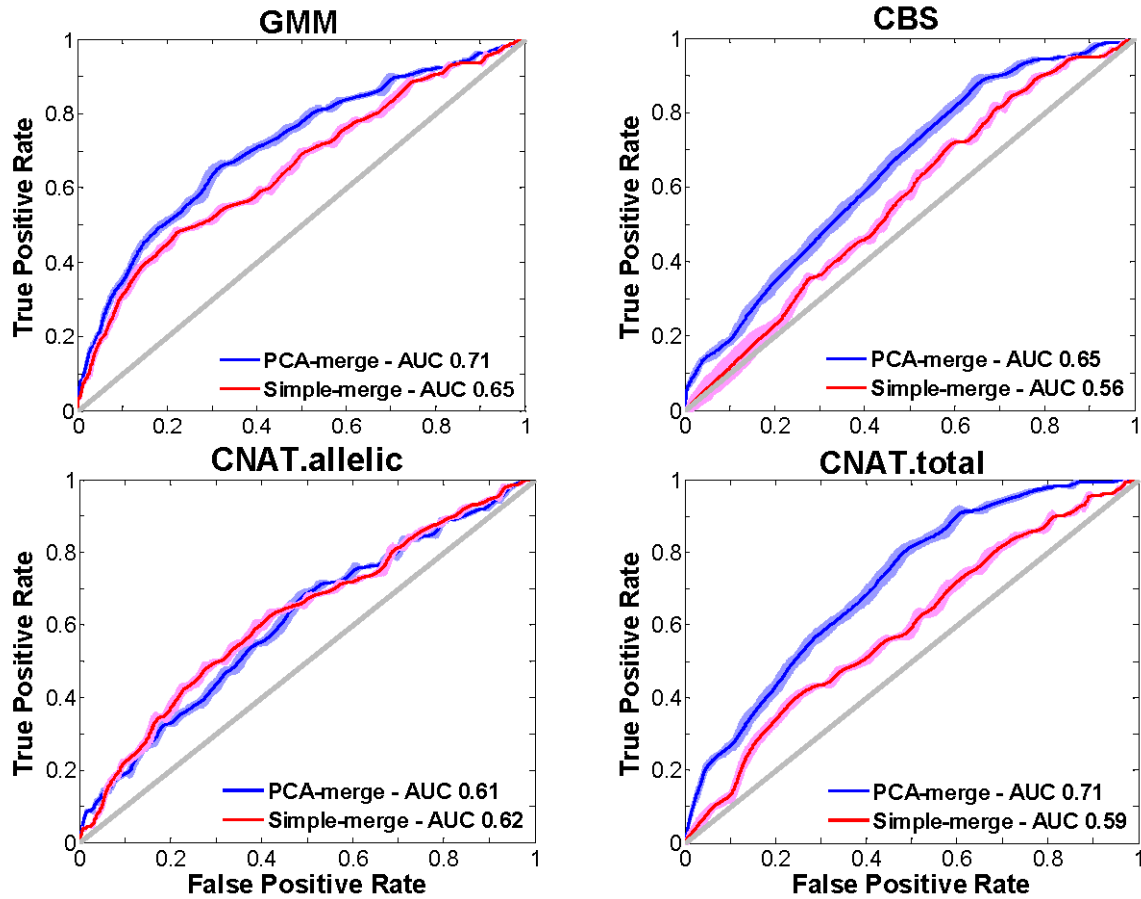


Figure 4 Performance for predicting relatedness based on CNP profiles generated by different methods

Each plot shows the Receiver Operator Characteristic (ROC) curve for predicting relatedness between individuals based on the similarity of their CNV profiles generated by different methods (CNV detection algorithms are indicated above each plot and merging procedures by colors). The analysis employed 162 pairs of individuals known to be related and 2,000 pairs of unrelated individuals. Curves were made with the mean (solid lines) +/- two standard deviation (light blue or light red surfaces) from 100 permutations. The Area Under the Curve (AUC) values are shown in the legends.

3.6 Discussion

In this work, we analyzed CNPs and rare CNV regions within the CoLaus population using four different copy number detection methods and applying two different merging procedures. We also devised various validation strategies to compare the performance of these methods.

3.6.1. Properties of the PCA merging technique

The simple merging approach is able to concatenate about half a million SNPs into a few thousands regions. Yet, this naïve technique leaves CNV edges fragmented into regions of few or even single SNPs. Therefore we developed a novel merging method, based on a PCA which, provides a strong improvement over the simple approach as it significantly reduces the number of single SNPs by re-attributing them to larger regions. Also, small regions (<1kb) were extended either by incorporating single SNPs or by merging them with other small regions.

3.6.2. Comparison of the different CNV prediction methods

We demonstrated that *CNAT.allelic* predicts a large number of CNVs. Yet only a relatively small fraction of these could be replicated, indicating that most of the predicted CNVs are likely to be false positives. This is also supported by the fact that CNV profiles generated by *CNAT.allelic* performed worse in predicting kinship. In contrast, *CNAT.total* appears to be overly conservative and is likely to miss subtle, but real CNV events. Based on our comparative analyses we find that CBS is a very efficient segmentation algorithm, confirming reports by several independent studies^{40,49,50}. Our *GMM* method also performs much better, both for sensitivity and specificity, than the two CNAT implementations. We also observed that our model was able to detect many more CNPs than CBS, suggesting higher sensitivity.

Currently our model only considers deletion, copy neutral, single copy or multiple copies. Since very few homozygous deletions were observed with other applied algorithms, we did not use such a dedicated component in our analysis. Nevertheless, our *GMM* implementation allows for such an extension.

3.6.3. Validation of CNVs in a large clinical cohort

Validation is an essential part of any CNV discovery project. PCR, Southern blot and many other targeted techniques are useful to predict accurately the copy number at a given locus, but low throughput is a severe limitation when large numbers of CNVs need to be validated. The Database of Genomic Variants is a valuable resource to reduce the fraction of CNVs to be further validated. Nevertheless, there still remains a significant fraction of novel CNVs. For such CNVs replicating a number of individuals (e.g. a few hundred) on an independent array platform is a viable option. With the recent reduction in the cost of microarrays, such large-scale replication now becomes affordable.

As a complement to replication experiments, one can take advantage of the relatedness between individuals. Deciphering relatedness (if not already known) can easily be achieved by clustering the SNP genotypes. Here we showed that assessing how well the relatedness can be predicted based on the CNV profiles is a powerful technique to gauge the quality of a CNV calling and merging method.

3.6.4. Conclusion and Perspectives

Our *GMM* and PCA merging algorithm are useful techniques to detect and merge CNVs. They have been successfully applied to a large clinical cohort. These techniques are not limited to data from SNP arrays, they require as input only a matrix of hybridization ratios (for the former) or copy number values (for the latter). Thus they can be applied to data from other platforms such as CGH arrays.

Despite significant improvements in CNV detection and analysis when using the most recent SNP arrays (e.g. new generation Affymetrix arrays^{41,42}), there are still many large medical cohorts where SNP data have been collected but CNV analysis has not been reported. This concerns both complex diseases (e.g.^{5,51-53}) and cancer (e.g.⁵⁴⁻⁵⁶). Hundreds of thousands of individuals have already been genotyped on 500K Affymetrix or 550K Illumina SNP chips, but the corresponding data have not been used for CNV analysis, simply because it is a much more challenging task due to the lack of well-established algorithms and protocols. We hope that the present work will make it easier for researchers to make better use of their data for CNV calling.

GWAS have demonstrated that the genetic variance cannot fully be attributed to SNPs. For example, for highly heritable traits such as height (with 13,665 individuals), SNPs only explain 3% of the variance ⁷. It has also been shown that, for common traits, the large fraction of heritability cannot be accounted for by CNPs ¹¹. Thus the identification of rare CNVs with stronger clinical impact, as we recently demonstrated for obesity ³⁷, is one of the most promising alternatives. Meta-analysis of existing cohorts for CNVs gives more power to detect rare CNVs because unique CNVs in a single cohort can then be supported by different cohorts. But such meta-analyses cannot be used to identify small variants due to the poor SNP density. In such cases, individuals with rare variants should be investigated further with higher density arrays or with genomic sequencing.

3.7 Authors and affiliations

Armand Valsesia^{1,2,3}, Zoltán Kutalik^{1,2}, Toby Johnson^{1,2,4,5}, Brian J. Stevenson^{2,3}, Dawn Waterworth⁶, Vincent Mooser⁶, Peter Vollenweider⁷, Gérard Waeber⁷, C. Victor Jongeneel^{2,3}, Jacques S. Beckmann^{1,8}, Sven Bergmann^{1,2*}

1. Department of Medical Genetics, University of Lausanne, Switzerland
2. Swiss Institute of Bioinformatics, Lausanne, Switzerland
3. Ludwig Institute for Cancer Research, Lausanne, Switzerland
4. University Institute of Social and Preventive Medicine, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland.
5. William Harvey Research Institute, Barts and The London, Queen Mary University of London, UK
6. Medical Genetics/Clinical Pharmacology and Discovery Medicine, GlaxoSmithKline, Philadelphia, Pennsylvania, USA
7. Department of Medicine, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland
8. Service of Medical Genetics, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland

* Corresponding author: Sven.Bergmann@unil.ch

3.8 Acknowledgements

We acknowledge Yolande Barreau, Mathieu Firmann, Vladimir Mayor, Anne-Lise Bastian, Binasa Ramic, Martine Moranville, Martine Baumer, Marcy Sagette, Jeanne Ecoffey, and Sylvie Mermoud for data collection. Part of the computations were performed on the Vital-IT cluster, we are grateful to the whole Vital-IT team, in particular to Ioannis Xenarios, Roberto Fabbretti and Volker Flegel. We also thank Bastian Peter for system administration and fulfilling our needs for storage. We would like to acknowledge Richard Redon, for his precious advice at early stage of the study.

3.9 References

1. Vollenweider, P. et al. [Health examination survey of the Lausanne population: first results of the CoLaus study]. *Rev Med Suisse* **2**, 2528-30, 2532-3 (2006).
2. Affymetrix. www.affymetrix.com.
3. Newton-Cheh, C. et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* (2009).
4. Kolz, M. et al. Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genet* **5**, e1000504 (2009).
5. Loos, R.J. et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* **40**, 768-75 (2008).
6. Prokopenko, I. et al. Variants in MTNR1B influence fasting glucose levels. *Nat Genet* **41**, 77-81 (2009).
7. Weedon, M.N. et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* **40**, 575-83 (2008).
8. Willer, C.J. et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* **41**, 25-34 (2009).
9. Yuan, X. et al. Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am J Hum Genet* **83**, 520-8 (2008).
10. Sandhu, M.S. et al. LDL-cholesterol concentrations: a genome-wide association study. *Lancet* **371**, 483-91 (2008).
11. Conrad, D.F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* (2009).
12. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).
13. Li, C. Automating dChip: toward reproducible sharing of microarray data analysis. *BMC Bioinformatics* **9**, 231 (2008).
14. Nannya, Y. et al. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res* **65**, 6071-9 (2005).
15. Komura, D. et al. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res* **16**, 1575-84 (2006).
16. Huang, J. et al. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics* **1**, 287-99 (2004).

17. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-72 (2004).
18. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-63 (2007).
19. Iafrate, A.J. et al. Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-51 (2004).
20. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat Rev Genet* **7**, 85-97 (2006).
21. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-32 (2005).
22. Sharp, A.J. et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78-88 (2005).
23. Jakobsson, M. et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998-1003 (2008).
24. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-8 (2004).
25. Freeman, J.L. et al. Copy number variation: new insights in genome diversity. *Genome Res* **16**, 949-61 (2006).
26. Perry, G.H. et al. Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci U S A* **103**, 8006-11 (2006).
27. Perry, G.H. et al. Copy number variation and evolution in humans and chimpanzees. *Genome Res* **18**, 1698-710 (2008).
28. Lee, A.S. et al. Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet* **17**, 1127-36 (2008).
29. Henrichsen, C.N. et al. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet* **41**, 424-9 (2009).
30. Lupski, J.R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* **1**, e49 (2005).
31. de Cid, R. et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat Genet* **41**, 211-5 (2009).
32. Beckmann, J.S., Estivill, X. & Antonarakis, S.E. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet* **8**, 639-46 (2007).
33. Cowell, J.K. & Hawthorn, L. The application of microarray technology to the analysis of the cancer genome. *Curr Mol Med* **7**, 103-20 (2007).
34. Kallioniemi, A. et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818-21 (1992).
35. Kallioniemi, A. CGH microarrays and cancer. *Curr Opin Biotechnol* **19**, 36-40 (2008).
36. Pinkel, D. & Albertson, D.G. Array comparative genomic hybridization and its applications in cancer. *Nat Genet* **37** **Suppl**, S11-7 (2005).
37. Walters, R.G. et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* **463**, 671-5 (2010).
38. Itsara, A. et al. Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* **84**, 148-61 (2009).
39. Shaikh, T.H. et al. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res* **19**, 1682-90 (2009).

40. Pique-Regi, R. et al. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* **24**, 309-18 (2008).
41. Korn, J.M. et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253-60 (2008).
42. McCarroll, S.A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166-74 (2008).
43. Pique-Regi, R., Ortega, A. & Asgharzadeh, S. Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA. *Bioinformatics* **25**, 1223-30 (2009).
44. The International HapMap Project. *Nature* **426**, 789-96 (2003).
45. Bengtsson, H. A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. *Tech Report, Department of Statistics, University of California, Berkeley* **745**(2008).
46. Bengtsson, H., Irizarry, R., Carvalho, B. & Speed, T.P. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**, 759-67 (2008).
47. Bengtsson, H., Ray, A., Spellman, P. & Speed, T.P. A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. *Bioinformatics* **25**, 861-7 (2009).
48. Illumina. www.illumina.com.
49. Lai, W.R., Johnson, M.D., Kucherlapati, R. & Park, P.J. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763-70 (2005).
50. Willenbrock, H. & Fridlyand, J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**, 4084-91 (2005).
51. van Es, M.A. et al. Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat Genet* **41**, 1083-7 (2009).
52. Soranzo, N. et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet* **41**, 1182-90 (2009).
53. Rivadeneira, F. et al. Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat Genet* **41**, 1199-206 (2009).
54. Gudmundsson, J. et al. Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat Genet* **40**, 281-3 (2008).
55. Eeles, R.A. et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* **41**, 1116-21 (2009).
56. Thomas, G. et al. Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* **40**, 310-5 (2008).

4 Aetiology of CNVs in complex disease

I have been heavily involved in a project centred on morbid obesity led by Pr. Jacques Beckmann and Pr. Philippe Froguel. This project was a collaborative effort between the CHUV, the Imperial College London, UNIL, SIB, University of Tartu, and numerous hospitals in Europe. The aim of the project was to investigate the prevalence and penetrance of a rare deletion found in obese patients.

Obesity is a major health problem in Western societies, it increases risk factor for type 2 diabetes, cardio-vascular problems, infertility, osteoarthritis, and several cancer such as breast, liver, pancreas, prostate, kidney ^{1,2}. The aim of this project was to investigate a rare, 600Kb-long deletion at chromosome 16p11.2 in obese patients. This locus at 16p11.2 is of the highest interest for three main reasons, 1) This locus is associated to schizophrenia and autism ³⁻⁵, 2) We initially detected this deletion in patients ascertained for developmental delays, following detailed clinical investigation we found all the affected patients were obese, suggesting the deletion could affect genes involved with both phenotypes. And 3) despite several candidates from genome-wide association studies (GWAs), these candidates account for little of the known heritability in obesity ⁶.

We thus compared CNVs from 8,456 obese patients and 11,856 individuals from the general population and we found that the deletion explained 0.7% of the morbidly obese patients. The odd ratios were highly significant for both obese patients (odds ratio 29.8, with 95% confidence interval 3.9-225) and morbidly obese patients (odds ratio 43, 95% confidence interval 5.6-329). This explained more than classical SNP-based GWAs and demonstrated the high penetrance of such rare variant. This also highlights an interesting strategy for identifying the missing heritability in obesity and other complex traits. The analysis of cohorts with extreme and well-documented phenotypes may offer increased power to detect rare variants with strong effects.

In this project, I was one of the main data analyst with Dr. Robin Walters, one of the lead authors on the publication. More specifically, we have meta-analysed CNVs from 8,456 obese patients and 11,856 individuals from the general population. Additionally I have been in charge of the analysis of 5,612 individuals from the

CoLaus cohort (see Chapter 2). It should be emphasized that CoLaus constituted about half of the controls used in the project. I was also involved in writing the results for a publication in Nature ⁷.

The following part includes our paper entitled “*A new highly penetrant form of obesity due to microdeletions on chromosome 16p11.2*”. This paper was published in Nature in February 2010. Detailed CoLaus analysis is available in Chapter 3, a method summary is also included in the paper and supplemental information can be found in the annexes.

4.1 Abstract

Obesity has become a major worldwide challenge to public health, owing to an interaction between the Western *obesogenic* environment and a strong genetic contribution ⁶. Recent extensive genome-wide association studies (GWASs) have identified numerous single nucleotide polymorphisms associated with obesity, but these loci together account for only a small fraction of the known heritable component ⁶. Thus, the ‘common disease, common variant’ hypothesis is increasingly coming under challenge ⁸. Here we report a highly penetrant form of obesity, initially observed in 31 subjects who were heterozygous for deletions of at least 593 kilobases at 16p11.2 and whose ascertainment included cognitive deficits. Nineteen similar deletions were identified from GWAS data in 16,053 individuals from eight European cohorts. These deletions were absent from healthy non-obese controls and accounted for 0.7% of our morbid obesity cases (body mass index (BMI) $\geq 40 \text{ kgm}^{-2}$ or BMI standard deviation score ≥ 4 ; $P=6.43 \times 10^{-8}$, odds ratio 43.0), demonstrating the potential importance in common disease of rare variants with strong effects. This highlights a promising strategy for identifying missing heritability in obesity and other complex traits: cohorts with extreme phenotypes are likely to be enriched for rare variants, thereby improving power for their discovery. Subsequent analysis of the loci so identified may well reveal additional rare variants that further contribute to the missing heritability, as recently reported for SIM1 ⁹. The most productive approach may therefore be to combine the ‘*power of the extreme*’ ¹⁰ in small, well-phenotyped cohorts, with targeted follow-up in case-control and population cohorts.

4.2 Methods summary

Obesity. Definitions for overweight, obesity and morbid obesity were based on previous studies ^{11,12}: for adults, BMI \geq 25, 30 and 40 kgm² respectively; for children, BMI respectively above the 90th, 97th centiles and at least four standard deviations above the mean, calculated according to their age and gender from a French reference population ^{13,14}.

Statistics. All reported statistical tests used Fisher's exact test ¹⁵, performed on contingency tables constructed for the number of subjects carrying or lacking a 16p11.2 deletion versus the obesity status or ascertainment of the individual. Because no homozygous deletions were observed, it was unnecessary to make a prior distinction between recessive, additive and dominant models of disease risk. Odds ratios and 95% confidence limits were calculated as described ¹⁶.

CNV discovery. Subjects ascertained for cognitive deficit/malformations with or without obesity were selected from those clinically referred for genetic testing; 16p11.2 deletions were identified in these individuals by standard clinical diagnostic procedures. Algorithmic analyses of GWAS data were performed variously using the cnvHap algorithm, a moving-window average-intensity procedure, a Gaussian mixture model, QuantiSNP, PennCNV, BeadStudio GTmodule, and Birdseed. When experimental validation was not possible, at least two independent algorithms were used for each data set.

4.3 Results

The extent to which copy-number variants (CNVs) might contribute to the missing heritability of common disorders is currently under debate⁸. Because most common simple CNVs are well tagged by single nucleotide polymorphisms (SNPs), it has recently been suggested that common CNVs are unlikely to contribute substantially to the missing heritability¹⁷. However, rare variants or recurring CNVs that have arisen on multiple independent occasions are unlikely to be captured by SNP tagging, and their identification will require alternative approaches.

We have previously proposed that cohorts with extreme phenotypes that include obesity may be enriched for rare but very potent risk variants^{10,11}. Here we investigate 312 subjects, from three centres in the UK and France, presenting with congenital malformations and/or developmental delay in addition to obesity as defined previously^{11,12} (see Methods). Known syndromes (for example, Prader–Willi and fragile X) were excluded. A combination of array comparative genomic hybridization (aCGH), genotyping arrays, quantitative PCR (qPCR) and multiplex ligation-dependent probe amplification (MLPA) was used to identify and confirm the presence of a heterozygous deletion on 16p11.2 in nine individuals (2.9%). These deletions, estimated to be a total of 740 kilobases (kb) in size (one copy of a segmental duplication plus 593 kb of unique sequences; Figure 1a), have previously been associated to varying extents with autism, schizophrenia and developmental delay¹⁸⁻²¹; however, the observed frequency of deletions in our cohort is appreciably higher than the reported frequencies in the cohorts from the previous studies (less than 1%), which did not include obesity as an inclusion criterion.

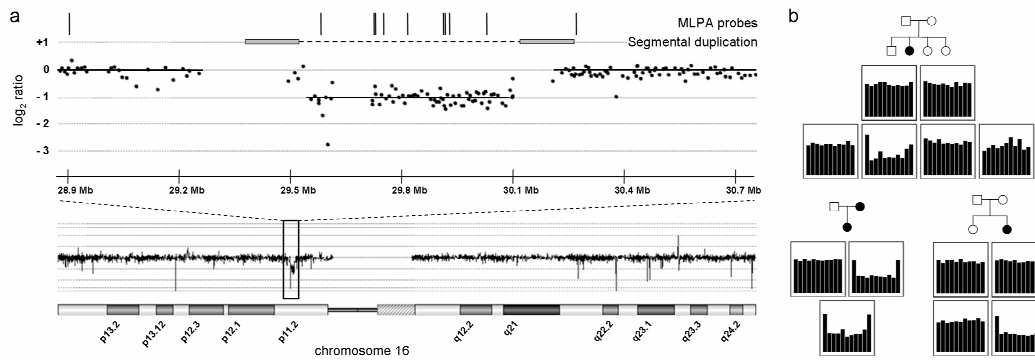


Figure 1 Identification and validation of deletions at 16p11.2

(a) aCGH data showing the location of the 16p11.2 deletion. The data show the log₂ intensity ratio for a deletion carrier compared to an undeleted control sample. Grey bars connected by a broken line denote the segmental duplication flanking the deletion region. Vertical bars indicate the positions of the probe pairs used for MLPA validation. Note that CGH and genotyping array probes targeted against segmental duplications may not accurately report copy number due to the increased number of homologous sequences in the diploid state. Genome coordinates are according to the hg18 build of the reference genome. (b) MLPA validation of 16p11.2 deletions. Representative MLPA results are shown, illustrating one instance of maternal transmission and two instances of de novo deletions. Genotyping data excluded the possibility of non-paternity. Full results for MLPA validation and inheritance analysis are shown in Supplementary Figure S1. Each panel shows the relative magnitude of the normalised, integrated signal at each probe location, in order of chromosomal position of the MLPA probe pairs as indicated in (a). Each panel corresponds to its respective position on the associated pedigree, as shown.

A parallel, independent survey of aCGH and SNP-CGH data from eight cytogenetic centres in France, Switzerland and Estonia, involving 3,947 patients with developmental delay and/or malformations but this time without selection for obesity, revealed 22 unrelated cases with similar deletions (0.6%). This is a frequency consistent with those found in the previous studies¹⁸⁻²¹, but is significantly lower than for the above cohort, which included only obese subjects ($P=2.2 \times 10^{-4}$, Fisher's exact test).

Analysis of the available clinical data for these 22 new carriers indicated that, in addition to the ascertained cognitive deficits or behavioural abnormalities (including hyperphagia, specifically identified in at least nine cases; see Supplementary Table 1), a 16p11.2 deletion gave rise to a strongly expressed obesity phenotype in adults, with a more variable phenotype in childhood. All four teenagers and adults carrying a deletion were obese, whereas child carriers were also frequently either obese (4 of 15) or overweight (2 of 15), a tendency that has previously been noted²¹; the very young (under 2 years old) were of normal weight. This age-dependent penetrance was observed in all instances of deletions for which phenotypic data were available,

whether from this study or from previously published reports²⁰⁻²⁵, and regardless of ascertainment (Figure 2; see Supplementary Tables 2 and 3).

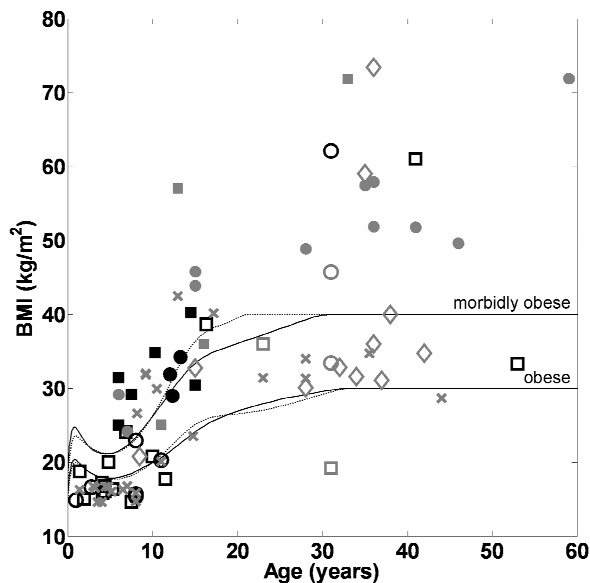


Figure 2 Dependence of BMI on age in subjects having a deletion at 16p11.2.

Data are for all individuals carrying a deletion for whom phenotypic data are available. Similar data from this study only are shown in Supplementary Figures S2 and S3. Lines denote the age- and gender-corrected thresholds (solid/broken – male/female) for obesity and morbid obesity. Symbols are: Square/circle – male/female; black/grey – ascertained/not ascertained for developmental delay; filled/open – ascertained/not ascertained for obesity; diamond – first-degree relative of proband; cross – previously published data²⁰⁻²⁵. The 31 year old male with BMI ~20 kg.m⁻² was diabetic based on fasting blood glucose >7 mmol/L.

Taken together, the data from these parallel studies suggest a possible direct association of deletions at 16p11.2 with obesity, distinct from their cognitive phenotype. Also identified in these cohorts were instances of the reciprocal duplication, which has also been implicated in neurodevelopmental disorders, but with a variable phenotype and lower penetrance^{19,20,22}. The frequency of the duplication in the two cohorts (12 of 4,183 (0.3%)) was consistent with previous reports for patients with cognitive deficits (0.3–0.7%)^{20,22}. Carriers of the duplication neither were obese nor had reported hyperphagia.

To investigate further the association of 16p11.2 deletions with obesity, and to estimate the extent to which it is observed independently of ascertainment for neurodevelopmental symptoms, we performed algorithmic and statistical analyses of genome-wide SNP genotyping data (see Table 1) from Swiss (CoLaus²⁶), Finnish (NFBC1966²⁷) and Estonian (EGPUT²⁸) general population cohorts (11,856 subjects in total), from child obesity and adult morbid obesity case-control cohorts^{11,29,30} (1,224 and 1,548 subjects, respectively), from an extreme early-onset obesity cohort (SCOOP, 931 subjects) and from 141 patients undergoing bariatric weight-loss surgery (see Methods); in total, we identified 17 instances of deletions (and four duplications) with no significant gender bias (Table 1). In addition, we identified two further unrelated carriers of a deletion from 353 members of 149 families with sibling pairs discordant for obesity (SOS Sib Pair Study³¹). When DNA was available for further analysis (15 of 19 samples), the presence of a deletion was validated by using MLPA (Figure 1b) or qPCR; the remaining deletions were validated by applying a second independent algorithm to the data. With the exception of a single individual who is apparently diabetic (fasting blood glucose more than 7 mM), all adult carriers of such deletions were obese, the majority being morbidly obese; similarly, each of the seven child or adolescent carriers had a BMI in the top 0.1% of the population range for their age and gender. None of the individuals ascertained on the basis of their obesity had any reported developmental delay or cognitive deficit; four subjects were reported as having hyperphagia.

Cohort	Deletions/Total				Total	Technology
	Lean/ Normal	Overweight	Obese	Morbidly Obese		
<i>Ascertained for cognitive deficits/malformations and obesity</i>						
Lille/Strasbourg ^a					8/279	qPCR, aCGH
London ^a					1/33	aCGH, MLPA
<i>Ascertained for cognitive deficits/malformations</i>						
French-Swiss cytogenetic clinical diagnostic group ^a					21/3870	aCGH, QMPSF, qPCR, FISH
Estonian cases of cognitive deficit ^a					1/77	Illumina CNV370-Duo, qPCR
<i>Ascertained for obesity</i>						
Swedish families with discordant siblings ^{b,d}	0/140	0/54	0/115	2/44	2/353	Illumina 610K-Quad, MLPA
French adult case-control ^b	0/669	0/174	-	4/705	4/1548	Illumina CNV370-Duo, MLPA
French child case-control ^c	0/530	0/51	1/260	3/383	4/1224	Illumina CNV370-Duo, MLPA
British extreme early-onset obesity ^c				3/931	3/931	Affymetrix 6.0, MLPA
French bariatric weight-loss surgery ^b	-	-	0/15	2/126	2/141	Illumina 1M-duo, MLPA
<i>Population cohorts (origin)</i>						
NFBC66 (Finnish) ^b	1/3148	0/1622	1/434	1/42	3/5246	Illumina CNV370-Duo
CoLaus (Swiss) ^b	0/2675	0/2049	0/830	0/58	0/5612	Affymetrix 500K
EGPUT (Estonian) ^b	0/412	0/358	1/213	0/15	1/998	Illumina CNV370-Duo, qPCR
<i>Total without ascertainment for cognitive deficits/malformations^d</i>	<i>1/7434</i>	<i>0/4254</i>	<i>3/1742</i>	<i>13/2260</i>		

Table 1 Frequency of detected 16p11.2 deletions in multiple cohorts

For each cohort, 16p11.2 deletions were identified and validated using the indicated technologies. Where full phenotypic data was available, members of cohorts were categorised according to the appropriate obesity criteria (see Supplementary Information): ^aNot categorised, complete phenotypic data not available. ^bBMI thresholds for overweight, obese, morbidly obese were $\geq 25 \text{ kg.m}^{-2}$, $\geq 30 \text{ kg.m}^{-2}$, $\geq 40 \text{ kg.m}^{-2}$ respectively. ^cBMI thresholds for overweight, obese, morbidly obese were the age- and gender-corrected 90th percentile, 97th percentile, +4 standard deviations above the mean, respectively. ^dDiscordant siblings not included in totals due to relatedness.

To enable sufficient statistical power to give robust conclusions, we combined data from the population and obesity cohorts in an overall case-control association analysis (the samples from sib-pair families were excluded to avoid complications due to their relatedness). In comparison with lean or normal weight subjects (see Table 1 and Methods), 16p11.2 deletions were associated with obesity ($P=5.8 \times 10^{-7}$, Fisher's exact test; odds ratio 29.8, 95% confidence limits 3.9 and 225) and morbid obesity ($P=6.4 \times 10^{-8}$; odds ratio 43.0, 95% confidence limits 5.6 and 329) at or near genome-wide levels of significance. Expanding the control group to include all non obese individuals increased the significance to $P=4.2 \times 10^{-9}$ (obese) and $P=6.1 \times 10^{-10}$ (morbidly obese).

Previous reports have indicated that these deletions are frequently not inherited from either parent but arise *de novo*, possibly by nonallelic homologous recombination between the more than 99% sequence-identical segmental duplications flanking the deleted region^{21,24}. Therefore, where possible we investigated the parents of carriers of deletions, identifying 11 cases of maternal transmission and 4 of paternal transmission. The available data showed that all first-degree relatives carrying a deletion were also obese (Supplementary Table 1). In ten instances the deletion was apparently *de novo* (see Figure 1b). Extrapolation to our full data set indicates that about 0.4% of all morbidly obese cases are due to an inherited 16p11.2 deletion. The frequency of *de novo* events is consistent with a previous report, in which ascertainment was for developmental delay and/or congenital anomalies²¹; by contrast, deletions are reported to be almost exclusively *de novo* in autistic subjects¹⁸⁻²⁰.

Although they may be heterogeneous in nature, these deletions are highly likely to be the causal variants, representing the second most frequent genetic cause of obesity after point mutations in *MC4R*^{32,33}. Their repeated *de novo* occurrence is likely to result in a lack of linkage disequilibrium with any other flanking variant—no consistent haplotype has been identified by analysis of the available surrounding genotypes. To assess the effect of a deletion on the expression of nearby genes (for example, the obesity GWAS-associated *SH2B1* locus 800 kb distant³⁴), we analysed available transcript data for subcutaneous adipose tissue samples from the discordant sibling cohort. Comparisons of the two subjects carrying a deletion with their corresponding

non-obese siblings, and with other obese and non-obese subjects (Supplementary Fig. 4 and Supplementary Tables 4 and 5), showed that many, although not all, transcripts from within the deletion had a markedly decreased abundance (0.4–0.7-fold). In contrast, no clear evidence was found for consistent cis effects of the deletion on the abundance of messenger RNAs encoded by genes flanking the deletion. In addition, global analysis of this data set has not identified any trans-acting expression quantitative trait loci either within or nearby the deletion.

Thus, although we cannot completely exclude the possibility that a 16p11.2 deletion affects the expression of nearby genes (for instance, its impact may be different in other tissues), the expression analysis described strongly indicates that the observed phenotypes are likely to be due to haploinsufficiency of one or more of the about 30 genes within the deleted region. Indeed, rather than being due to a single haploinsufficiency, the phenotype may well result from the deletion of multiple genes with an impact on pathways central to the development of obesity (see Supplementary Table 5). Functional network analysis of the deleted genes has led to the suggestion of a similar multigene effect for the cognitive phenotype¹⁸. The extent to which there is overlap between the genes involved in the obesity and cognitive phenotypes remains to be elucidated.

There is a strong correlation between developmental and cognitive disabilities and the prevalence of obesity: patients with autism or who have learning disabilities have a greatly increased risk of obesity³⁵, and the severely obese exhibit significant cognitive impairment³⁶. Possible explanations include a direct causal relationship between obesity and developmental delay, the involvement of the same or related regulatory pathways, or different outcomes of the same set of behavioural disorders with complex pleiotropic effects and variable ages of onset and expressivities. The higher frequency of 16p11.2 deletions in the cohort ascertained for both phenotypes (2.9%), compared with cohorts ascertained for either phenotype alone (0.4% and 0.6%, respectively), confirms their impact on both obesity and developmental delay, adding to the evidence that these two phenotypes may be fundamentally interrelated.

4.4 Methods

Obesity phenotype. We used previously defined criteria to define overweight, obesity, and morbid (class III) obesity^{11,12}. In adults, the thresholds were BMI \geq 25, 30 and 40 kg.m⁻², respectively. In children and adolescents, we used age-specific and sex-specific centiles of BMI, calculated from a French reference population^{13,14}, that approximately corresponded to these thresholds: overweight and obesity were defined by thresholds at the 90th and 97th centiles, respectively. Childhood morbid obesity was defined as BMI \geq 4 standard deviations above the age-specific and sex-specific mean, which corresponds to a BMI of 40 kg.m⁻² between the ages of 20 and 30 years for both men and women; this threshold was used in the recruitment of the SCOOP severe early-onset obesity cohorts¹². The age-specific and sex-specific thresholds used to define obesity and morbid obesity are shown in Figure 1 and Supplementary Figs 1 and 2. No carriers of a 16p11.2 deletion were reported to be taking atypical antipsychotics (known to be associated with weight gain).

Patient and population cohorts. Patients referred for cognitive delay and obesity: a group of 33 patients was selected from those referred for genetic testing at the North West Thames Regional Genetics Service, based at Northwick Park Hospital in Harrow, UK, with approval from the Harrow Research Ethics Committee. Inclusion was based on three criteria: mental retardation, dysmorphism, and a weight greater than the 97th centile for age and gender. Abnormal karyotype, fragile X and Prader–Willi syndrome had previously been excluded.

A second group of 279 French children were selected from those referred to two centres (Laboratoire de Diagnostic Génétique, Nouvel Hôpital Civil, Strasbourg, France, and Centre de Génétique Chromosomique, Hôpital Saint-Vincent de Paul, GHICL, Lille, France). Inclusion was based on obesity plus at least one Prader–Willi-like syndromic feature (neonatal hypotonia and difficulty to thrive, mental retardation, developmental delay, behavioural problems, skin picking, facial dysmorphism, hypogonadism or hypogonadism). Chromosomal abnormalities and Prader–Willi syndrome were excluded by karyotyping and DNA methylation analysis.

Patients referred for cognitive delay: patients with cognitive deficits are routinely referred to clinical genetics for aetiological work-ups including aCGH. We surveyed

seven cytogenetic centres in France and Switzerland, identifying 3,870 patients ascertained for developmental delay and/or malformations. Also included in the study was a further 77 patients, ascertained on similar criteria, who were referred to the Department of Genetics, University of Tartu, Tartu, Estonia. These analyses were performed for clinical diagnostic purposes, all available phenotypic data (weight and height) being those provided anonymously by the clinician ordering the analysis. Consequently, research-based informed consent was not required by the institutional review board that approved the study.

CoLaus: this prospective population cohort was described previously²⁶; 6,188 white individuals aged 35–75 years were randomly selected from the general population in Lausanne, Switzerland. These individuals underwent a detailed phenotypic assessment and were genotyped with the Affymetrix Mapping 500K array; 5,612 samples passed genotyping quality control. This study was approved by the institutional review boards of the University of Lausanne, and written consent was obtained from all participants. Because recruitment of this cohort required the ability to give informed consent, it is possible that the (statistically non-significant) lack of 16p11.2 deletions or duplications is due to an ascertainment bias. However, any such bias, if it exists, is very small and affects the identification of only one or two subjects carrying a deletion.

NFBC1966: the Northern Finland Birth Cohort 1966 is a prospective birth cohort of almost all individuals born in 1966 in the two northernmost provinces of Finland. Expectant mothers were enrolled, and clinical data collection took place prenatally, at birth, and at ages 6 months, 1 year, 14 years and 31 years. Biochemical and DNA samples were collected with informed consent at age 31 years. Genotyping with the Illumina Infinium 370cnvDuo array and phenotypic characteristics of the cohort were as described previously²⁷. Phenotypic and genotyping data were available for 5,246 subjects after quality control.

EGPUT: the Estonian Genome Project is a biobank coordinated by the University of Tartu (EGPUT)²⁸. The project is conducted in accordance with Estonian Gene Research Act, and all participants gave written informed consent. The cohort includes more than 39,000 individuals older than 18 years of age and reflects closely the age distribution in the Estonian population (33% male, 67% female; 83%

Estonians, 14% Russians, 3% other). Subjects are recruited by general practitioners and hospital physicians and are then randomly selected. Computer Assisted Personal interview (CAPI) was filled during 1–2 h at the doctor's office. The data included personal data (such as place of birth, place(s) of living and nationality), family history (four generations), educational and occupational history, lifestyle and anthropometric data. A total of 1,090 randomly selected subjects were genotyped with the Illumina 370cnvDuo array, 998 passing the required criteria (nationality, genotyping call rate and phenotype availability).

Case-control familial obesity: the adult-obesity case-control groups and the child-obesity case control groups were as published previously ¹¹, and were genotyped with the Illumina Human CNV370-duo array. In all, 643 children with familial obesity (BMI \geq 97th centile corrected for gender and age, at least one obese first-degree relative, age less than 18 years), 581 non-obese children (BMI \leq 90th centile), 705 morbidly obese adults with familial obesity (BMI \geq 40 kgm², at least one obese first-degree relative with BMI \geq 35 kg.m⁻², age \geq 18 years) and 197 lean adults (BMI \leq 25 kg.m⁻²) passed quality control; this cohort included a further 646 control subjects from the DESIR prospective cohort ²⁹ (age at examination \geq 45 years, normal fasting glucose in accordance with 1997 ADA criteria, BMI $<$ 27 kg.m⁻²) genotyped with the Illumina Hap300 array ³⁰. All participants or their legal guardians gave written informed consent, and all local ethics committees approved the study protocol.

Severe early-onset obesity cohort: the Genetics of Obesity Study (GOOS) cohort consists of more than 3,000 patients ascertained for severe obesity, defined as a BMI \geq 4 standard deviations above the age-specific and sex-specific mean, and onset of obesity before 10 years of age. In this study we selected a discovery set of 1,000 UK Caucasian patients from this cohort in whom developmental delay had been excluded by routine clinical examination by experienced physicians (this cohort is referred to as SCOOP). Mutations in *LEPR*, *POMC* and *MC4R* were excluded by direct nucleotide sequencing and a karyotype was performed. DNA samples were analysed with Affymetrix Genome-Wide Human SNP Array 6.0 by Aros, of which 931 passed quality control.

Bariatric surgery cohort: patients undergoing elective bariatric weight-loss surgery were recruited for the ABOS study at Lille Regional University Hospital. Genotyping

was performed with the Illumina Human 1M-duo array, and data from 141 adults passed quality control. All participants gave written informed consent, and the study protocol was approved by the local ethics committee.

Swedish discordant sibling cohort: the SOS Sib Pair Study cohort was as published previously³¹. It includes 154 nuclear families, each with BMI discordant sibling pairs (BMI difference $>10 \text{ kg.m}^{-2}$), giving a total of 732 subjects. Genotyping data with the Illumina 610K-Quad array was available for 353 siblings from 149 families. Expression data from subcutaneous adipose tissue (sampled after overnight fasting) were available for 360 siblings from 151 families. Subjects received written and oral information before giving written informed consent. The Regional Ethics Committee in Gothenburg approved the studies.

Statistical methods. In view of the low frequency of the 16p11.2 deletions, all reported statistical tests were conducted with Fisher's exact test¹⁵. This was applied to comparisons of separately ascertained cohorts or categories and was performed on contingency tables constructed for the number of subjects carrying or lacking a 16p11.2 deletion (zero or one copies, because no homozygous deletions were observed) versus the obesity status or ascertainment of the individual. Because no homozygous deletions were observed, it was unnecessary to make a prior distinction between recessive, additive and dominant models of disease risk. For overall analysis of the obesity risk resulting from a deletion, cohorts were pooled in accordance with their obesity status determined according to the criteria described above, and the described tests were then applied to the pooled data. Odds ratios and 95% confidence limits were calculated as described¹⁶.

CNV discovery and validation. Clinical identification of 16p11.2 deletions: all diagnostic procedures (aCGH, qPCR, QMPSF and FISH) were conducted in accordance with the relevant guidelines of good clinical laboratory practice for the respective countries. All rearrangements in probands were confirmed by a second technique, and karyotyping was performed in all cases to exclude a complex rearrangement.

cnvHap: CNVs were detected in the child/adult case-control, bariatric surgery, SOS sibpair and NFBC cohorts using the cnvHap algorithm (L.J.M.C., J. E. Asher, R.G.W.,

J.S.E.-S.M., A.J.d.S., R.S., D. J. Balding, P.F. and A.I.F.B., unpublished observations); this method is based on a hidden Markov model that models transitions between copy-number states at the haplotype level, improving sensitivity and accuracy by capturing linkage disequilibrium information between CNVs and SNPs. The compiled JAR and associated parameter files can be downloaded from <http://www.imperial.ac.uk/medicine/people/l.coin/>. Sample data from the algorithm applied to the NFBC cohort are illustrated in Supplementary Fig. 5a. After clustering of genotyping data with the internal Illumina BeadStudio cluster files, values for logR ratio (LRR) and B-allele frequency (BAF) were exported from each project and normalized: effects of percentage GC content on LRR were removed by regressing on GC and GC2, and wave effects³⁷ were removed by fitting a Loess function. Normalized data for probes within 2.5 megabases of the 16p11.2 deletion were analysed with cnvHap, and CNV calls intersecting the single-copy sequences within the deletion (chr16:29514353–30107356, build hg18) were extracted. 16p11.2 deletions were identified by a minimum 90% of probes within the deleted region being called as having a decreased copy number. All called 16p11.2 deletions were validated by direct analysis of LRR. Data for each probe were normalized by first subtracting the median value across all samples (so that the distribution of LRR for each probes was centred on zero), and then dividing by the variance across all samples (to correct for variation in the sensitivity of different probes to copy-number variation). The normalized data were then smoothed by application of a nine-point moving average and visualized graphically (see Supplementary Fig. 6); putative deletions were checked by subsequent manual confirmation of loss of heterozygosity across the entire region. Equally, all deletions called by this method were confirmed by cnvHap.

Gaussian mixture model: for the CoLaus cohort, raw genotyping data were normalized using the aroma.affymetrix framework³⁸. Normalization steps included allelic cross-talk calibration^{39,40}, intensity summarization using robust median average, and correction for any PCR amplification bias. Copy number (CN) ratios for a given sample, at a given SNP or CN probe, were computed as the log₂ ratio of the normalized intensity of this probe divided by the median across all the samples. CN ratios were subsequently smoothed by fitting a Loess function³⁷. CNV calling was performed with a new method based on a Gaussian mixture model (A.V., Z. Kutalik,

T. Johnson, B. J. Stevenson, C. V. Jongeneel, D.W., V.M., P.V., G.W., J.S.B. and S.B., unpublished observations). This Gaussian mixture model fits four components (deletion, copy neutral, one additional copy and two additional copies) to CN ratios. The final copy number at each probe location is determined as the expected (dosage) copy number. The method has been validated by comparing test data sets with results from the CNAT⁴¹ and CBS^{42,43} algorithms and by replicating a subset of CoLaus subjects on Illumina arrays. All calls at the 16p11.2 locus made by the highly stringent CBS algorithm were replicated by the Gaussian mixture model. Principal components analysis detected no significant batch effects. Sample data from the algorithm applied to the CoLaus cohort are illustrated in Supplementary Fig. 5b. PennCNV, QuantiSNP and Birdsuite: CNV discovery in the EGPOT cohort was performed with QuantiSNP⁴⁴, PennCNV⁴⁵ and BeadStudio GT module (Illumina). All analyses were conducted with the recommended settings, except changing EMitters to 25 and L to 1,000,000 in QuantiSNP. For PennCNV, the Estonian population-specific BAF file was used. Data from the SCOOP cohort were analysed with Affymetrix Power Tools and Birdsuite software⁴⁶. Multiplex ligation-dependent probe amplification (MLPA): MLPA was performed with standard methods⁴⁷ using reagents obtained from MRC-Holland. The SALSA MLPA kit P343-B1 Autism-1 probe mix was used, which contained nine probes within the deleted region on 16p11.2, plus one probe upstream and one downstream of this locus (see Figure 1a). MLPA products were separated with an AB3130 Genetic Analyser (Applied Biosystems) and outputs were analysed with GeneMarker software (Soft Genetics) and Microsoft Excel. Data normalization was performed by dividing the peak areas for each of the 11 test probes by the mean of 9 control probe peak areas. Normalized peak area data were then compared across the tested samples to determine which of them carried the 16p11.2 deletion.

4.5 Authors and affiliations

R. G. Walters^{1,2*}, S. Jacquemont^{3*}, A. Valsesia⁴⁻⁶, A. J. de Smith¹, D. Martinet³, J. Andersson¹, M. Falchi¹, F. Chen⁷, J. Andrieux⁸, S. Lobbens⁹, B. Delobel¹⁰, F. Stutzmann⁹, J. S. El-Sayed Moustafa¹, J.-C. Chèvre⁹, C. Lecoeur⁹, V. Vatin⁹, S. Bouquillon⁸, J. L. Buxton¹, O. Boute¹¹, M. Holder-Espinasse¹¹, J.-M. Cuisset¹², M.-P. Lemaître¹², A.-E. Ambresin¹³, A. Brioschi¹⁴, M. Gaillard³, V. Giusti¹⁵, F. Fellmann³, A. Ferrarini³, N. Hadjikhani^{7,16}, D. Champion¹⁷, A. Guilmatre¹⁷, A. Goldenberg¹⁸, N.

Calmels¹⁹, J.-L. Mandel¹⁹, C. Le Caignec^{20,21}, A. David²⁰, B. Isidor²⁰, M.-P. Cordier²², S. Dupuis-Girod²², A. Labalme²², D. Sanlaville^{22,23}, M. Béri-Dexheimer²⁴, P. Jonveaux²⁴, B. Leheup^{24,25}, K. Ounap²⁶, E. G. Bochukova²⁷, E. Henning²⁷, J. Keogh²⁷, R. J. Ellis²⁸, K. D. MacDermot²⁸, M. M. van Haelst²⁸, C. Vincent-Delorme²⁹, G. Plessis³⁰, R. Touraine³¹, A. Philippe³², V. Malan³², M. Mathieu-Dramard³³, J. Chiesa³⁴, B. Blaumeiser³⁵, R. F. Kooy³⁵, R. Caiazzo^{36,37}, M. Pigeyre³⁷, B. Balkau³⁸, R. Sladek^{39,40}, S. Bergmann^{4,6}, V. Mooser⁴¹, D. Waterworth⁴¹, A. Reymond⁴², P. Vollenweider⁴³, G. Waeber⁴³, A. Kurg⁴⁴, P. Palta⁴⁴, T. Esko^{45,46}, A. Metspalu^{45,46}, M. Nelis^{45,46}, P. Elliott², A.-L. Hartikainen⁴⁷, M. I. McCarthy^{48,49}, L. Peltonen⁵⁰⁻⁵², L. Carlsson⁵³, P. Jacobson⁵³, L. Sjostrom⁵³, N. Huang⁵⁰, M. E. Hurles⁵⁰, S. O’Rahilly²⁷, I. S. Farooqi²⁷, K. Mannik⁴⁴, M.-R. Jarvelin^{2,54,55}, F. Pattou^{36,37}, D. Meyre⁹, A. J. Walley¹, L. J. M. Coin², A. I. F. Blakemore¹, P. Froguel^{1,9} & J. S. Beckmann^{3,4}

*These authors contributed equally to this work.

1. Section of Genomic Medicine, Imperial College London, London W12 0NN, UK.
2. Department of Epidemiology and Public Health, Imperial College London, London W21PG, UK.
3. Service de Génétique Médicale, Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne, Switzerland.
4. Département de Génétique Médicale, Université de Lausanne, CH-1015 Lausanne, Switzerland.
5. Ludwig Institute for Cancer Research, Université de Lausanne, CH-1015 Lausanne, Switzerland.
6. Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland.
7. Brain Mind Institute, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.
8. Laboratoire de Génétique Médicale, Centre Hospitalier Régional Universitaire, 59000 Lille, France.
9. CNRS 8090-Institute of Biology, Pasteur Institute, 59800 Lille, France.
10. Centre de Génétique Chromosomique, Hôpital Saint-Vincent de Paul, GHICL, 59020 Lille, France.
11. Service de Génétique Clinique, Hôpital Jeanne de Flandre, Centre Hospitalier Universitaire de Lille, 59000 Lille, France.
12. Service de Neuropédiatrie, Centre Hospitalier Régional Universitaire, 59000 Lille, France.
13. Unité Multidisciplinaire de Santé des Adolescents, Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne, Switzerland.
14. Service de Neuropsychologie et de Neuroréhabilitation, Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne, Switzerland.
15. Service d’Endocrinologie, Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne, Switzerland.
16. Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, Massachusetts 02129, USA.

17. INSERM, U614, Faculté de Médecine, 76183 Rouen, France.
18. Service de Génétique, Centre Hospitalier Universitaire de Rouen, 76031 Rouen, France.
19. Laboratoire de Diagnostic Génétique, Nouvel Hôpital civil, 67091 Strasbourg, France.
20. Centre Hospitalier Universitaire Nantes, Service de Génétique Médicale, 44093 Nantes, France.
21. INSERM, UMR915, L'Institut du Thorax, 44007 Nantes, France.
22. Service de Génétique, Hospices Civils de Lyon, Hôpital de l'Hôtel Dieu, 69288 Lyon, France.
23. EA 4171, Université Claude Bernard, 69622 Lyon, France.
24. Laboratoire de Génétique, Centre Hospitalier Universitaire, Nancy Université, 54511 Vandoeuvre les Nancy, France.
25. EA4368 Medical School Nancy, Université Henri Poincaré, 54003 Nancy, France.
26. Department of Genetics, United Laboratories, Tartu University Children's Hospital, 50406 Tartu, Estonia.
27. University of Cambridge Metabolic Research Laboratories, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK.
28. North West Thames Regional Genetics Service, Northwick Park & St Marks Hospital, Harrow HA1 3UJ, UK.
29. Centre Hospitalier D'Arras, Génétique Médicale, 62000 Arras, France.
30. Service de Génétique Médicale, Centre Hospitalier Universitaire Clemenceau, 14033 Caen, France.
31. Centre Hospitalier Universitaire–Hôpital Nord, Service de Génétique, 42055 Saint Étienne, France.
32. Département de Génétique et INSERM U781, Université Paris Descartes, Hôpital Necker-Enfants Malades, 75015 Paris, France.
33. Service de Génétique Clinique, Centre Hospitalier Universitaire, 80054 Amiens, France.
34. Laboratoire de Cytogénétique, Centre Hospitalier Universitaire Caremeau, 30029 Nîmes, France.
35. Department of Medical Genetics, University Hospital & University of Antwerp, 2650 Edegem, Belgium.
36. INSERM U859, Biotherapies for Diabetes, 59045 Lille, France.
37. Université Lille Nord de France, Centre Hospitalier Universitaire Lille, 59037 Lille, France.
38. INSERM U780-IFR69, 94807 Villejuif, France.
39. McGill University and Genome Quebec Innovation Centre, Montreal H3A 1A4, Canada.
40. Department of Medicine and Human Genetics, McGill University, Montreal H3A 1B1, Canada.
41. Division of Genetics, GlaxoSmithKline, Philadelphia, Pennsylvania 19101, USA.
42. The Center for Integrative Genomics, University of Lausanne, CH-1015 Lausanne, Switzerland.
43. Department of Medicine, Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne, Switzerland.
44. Institute of Molecular and Cell Biology, University of Tartu, 51010 Tartu, Estonia.
45. Estonian Genome Project, University of Tartu, 50410 Tartu, Estonia.
46. Estonian Biocentre, 51010 Tartu, Estonia.

47. Department of Obstetrics and Gynaecology, University of Oulu, 90220 Oulu, Finland.
48. Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford OX3 7LJ, UK.
49. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.
50. Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK.
51. Institute of Molecular Medicine, Biomedicum, 00290 Helsinki, Finland.
52. Massachusetts Institute of Technology, The Broad Institute, Cambridge, Massachusetts 02142, USA.
53. Department of Molecular and Clinical Medicine and Center for Cardiovascular and Metabolic Research, The Sahlgrenska Academy, 413 45 Goteborg, Sweden.
54. Department of Child and Adolescent Health, National Public Health Institute, 90101 Oulu, Finland.
55. Institute of Health Sciences and Biocenter Oulu, University of Oulu, 90220 Oulu, Finland.

4.6 Acknowledgements

A.J.W., A.I.F.B. and P.F. are supported by grants from the Wellcome Trust and the Medical Research Council (MRC). J.S.B. is supported by a grant from the Swiss National Foundation (310000-112552). L.J.M.C. is supported by an RCUK Fellowship. S.J. is funded by Swiss National Fund 320030_122674 and the Synapsis Foundation, University of Lausanne. A.V. is funded by the Ludwig Institute for Cancer Research. S.B. is supported by the Swiss Institute of Bioinformatics. I.S.F. and M.E.H. are funded by the Wellcome Trust and the MRC. We thank the DHOS (Direction de l'Hospitalisation et de l'Organisation des Soins) from the French Ministry of Health for their support in the development of several array CGH platforms in France. We thank 'le Conseil Regional Nord Pas de Calais/ FEDER' for their financial support. Part of the CoLaus computation was performed at the Vital-IT center for high-performance computing of the Swiss Institute of Bioinformatics. The CoLaus authors thank Y. Barreau, M. Firmann, V. Mayor, A.-L. Bastian, B. Ramic, M. Moranville, M. Baumer, M. Sagette, J. Ecoffey and S. Mermoud for data collection. The CoLaus study was supported by grants from GlaxoSmithKline, the Faculty of Biology and Medicine of Lausanne and by the Swiss National Foundation (33CSCO-122661). K.M., A.K., T.E., M.N. and A.M. received support from targeted financing from Estonian Government SF0180142s08, and P.P. from SF0180026s09; and from the EU through the European Regional Development Fund. T.E., M.N. and A.M. received support from FP7 grants (201413 ENGAGE, 212111 BBMRI, ECOGENE (no. 205419, EBC)). The genotyping of the Estonian Genome Project samples were

performed in the Estonian Biocentre Genotyping Core Facility. The EGPOT authors thank V. Soo for technical help in genotyping. The Northwick Park authors acknowledge support from the NIHR Biomedical Research Centre Scheme and the Hammersmith Hospital Charity Trustees. Genome Canada and Genome Quebec funded the genotyping of DESIR subjects. Work on the SOS sib pair cohort was supported by grants from the Swedish Research Council (K2008-65X-20753-01-4, K2007-55X-11285-13, 529-2002-6671), the Swedish Foundation for Strategic Research to Sahlgrenska Center for Cardiovascular and Metabolic Research, the Swedish Diabetes Foundation, the Ake Wiberg Foundation, Foundations of the National Board of Health and Welfare, the Jeansson Foundations, the Magn Bergvall Foundation, the Tore Nilson Foundation, the Royal Physiographic Society (Nilsson-Ehle Foundation), VINNOVA-VINNMER, and the Swedish federal government under the LUA/ALF agreement. The DESIR study has been supported by INSERM, CNAMTS, Lilly, Novartis Pharma and Sanofi-Aventis, by INSERM (Réseaux en Santé Publique, Interactions entre les déterminants de la santé), by the Association Diabète Risque Vasculaire, the Fédération Française de Cardiologie, La Fondation de France, ALFEDIAM, ONIVINS, Ardix Medical, Bayer Diagnostics, Becton Dickinson, Cardionics, Merck Santé, Novo Nordisk, Pierre Fabre, Roche and Topcon. Northern Finland Birth Cohort 1966 (NFBC1966) was supported by the Academy of Finland (project grants 104781, 120315 and Center of Excellence in Complex Disease Genetics), University Hospital Oulu, Biocenter, University of Oulu, Finland, the European Commission (EURO-BLCS, Framework 5 award QLG1-CT-2000-01643), NHLBI grant 5R01HL087679-02 through the STAMPEED program (1RL1MH083268-01), NIH/NIMH (5R01MH63706:02), the ENGAGE project and grant agreement HEALTH-F4-2007-201413, and the MRC (studentship grant G0500539). The NFBC authors thank P. Rantakallio for the launch of NFBC1966 and initial data collection, S. Vaara for data collection, T. Ylitalo for administration, M. Koironen for data management, and O. Tornwall and M. Jussila for DNA biobanking.

4.7 Author Contributions

A.I.F.B., P.F., J.S.B. and L.J.M.C. designed and supervised the study. F.C., D.M., S.J., J.A. and S.B. coordinated and managed patient databases. R.G.W., A.V., A.J.d.S., C.L., F.S., F.C., J.-C.C., J.L.B., S.L., N.H. and J.S.E.-S.M. performed data analysis. A.J.d.S. conducted the MLPA analysis. J.A., M.F. and A.J.W. analysed expression data. A.-E.A., A.B., A.D., A.F., A.G., A.G., A.L., A.P., B.B., B.D., B.I., B.L., C.V.-D., C.L.C., D.C., D.M., D.S., F.F., G.M., G.P., J.-L.M., J.-M.C., J.A., J.C., K.M., K.D.M., K.O., M.M.v.H., M.-P.C., M.-P.L., M.P., M.B.-D., M.H.-E., M.M., N.C., O.B., P.J., R.C., R.E., R.F.K., R.T., S.D.-G., S.J., V.G. and V.M. supervised patient recruitment and performed cytogenetic analysis. A.-L.H., A.K., A.M., A.R., B.B., D.M., D.W., E.G.B., E.H., F.P., G.W., I.S.F., J.A., J.K., L.C., L.P., L.S., M.E.H., M.I.M., M.N., M.-R.J., N.H., P.E., P.J., P.P., P.V., R.S., S.B., S.O'R., T.E., V.M. and V.V. supervised cohort recruitment and/or conducted genotyping. R.G.W., S.J., A.V., A.J.d.S., L.J.M.C., A.I.F.B., P.F. and J.S.B. wrote and edited the manuscript and prepared figures. P.F. and J.S.B. contributed equally. All authors commented on and approved the manuscript.

4.8 Author Information

The expression microarray data for carriers of 16p11.2 deletions is deposited in Gene Expression Omnibus under accession number GSE19238. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.F. (p.froguel@imperial.ac.uk) or J.S.B. (jacques.beckmann@chuv.ch).

4.9 References

1. Haslam, D.W. & James, W.P. Obesity. *Lancet* **366**, 1197-209 (2005).
2. Calle, E.E., Rodriguez, C., Walker-Thurmond, K. & Thun, M.J. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *N Engl J Med* **348**, 1625-38 (2003).
3. McCarthy, S.E. et al. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* **41**, 1223-7 (2009).
4. Fernandez, B.A. et al. Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *J Med Genet* **47**, 195-203 (2010).
5. Weiss, L.A. et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**, 667-75 (2008).
6. Walley, A.J., Asher, J.E. & Froguel, P. The genetic contribution to non-syndromic human obesity. *Nat Rev Genet* **10**, 431-42 (2009).
7. Walters, R.G. et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* **463**, 671-5 (2010).
8. Manolio, T.A. Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).
9. Stutzmann, F. Loss-of-function mutations in SIM1 cause a specific form of Prader-Willi-like syndrome. *Diabetologia* **52**, S104 (2009).
10. Froguel, P. & Blakemore, A.I.F. The power of the extreme in elucidating obesity. *N. Engl. J. Med.* **359**, 891-893 (2008).
11. Meyre, D. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature Genet.* **41**, 157-159 (2009).
12. Farooqi, I.S. & O'Rahilly, S. Recent advances in the genetics of severe childhood obesity. *Arch. Dis. Child.* **83**, 31-34 (2000).
13. Poskitt, E.M. Defining childhood obesity: the relative body mass index (BMI). *Acta Paediatr.* **84**, 961-963 (1995).
14. Rolland-Cachera, M.F. Body mass index variations: centiles from birth to 87 years. *Eur. J. Clin. Nutr.* **45**, 13-21 (1991).
15. Fisher, R.A. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc. A* **85**, 87-94 (1922).
16. Woolf, B. On estimating the relation between blood group and disease. *Ann. Hum. Genet.* **19**, 251-253 (1955).
17. Conrad, D.F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* (2009).
18. Kumar, R.A. Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17**, 628-638 (2008).
19. Marshall, C.R. Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477-488 (2008).
20. Weiss, L.A. Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667-675 (2008).
21. Bijlsma, E.K. Extending the phenotype of recurrent rearrangements of 16p11.2: deletions in mentally retarded patients without autism and in normal individuals. *Eur. J. Med. Genet.* **52**, 77-87 (2009).
22. McCarthy, S.E. Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genet.* **41**, 1223-1227 (2009).

23. Ghebranious, N., Giampietro, P.F., Wesbrook, F.P. & Rezkalla, S.H. A novel microdeletion at 16p11.2 harbors candidate genes for aortic valve development, seizure disorder, and mild mental retardation. *Am. J. Med. Genet. A.* **143A**, 1462-1471 (2007).
24. Fernandez, B.A. Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *J. Med. Genet.* (2010).
25. Shimojima, K., Inoue, T., Fujii, Y. & Ohno, K. Yamamoto, T. A familial 593[thinsp]kb microdeletion of 16p11.2 associated with mental retardation and hemivertebrae. *Eur. J. Med. Genet.* **52**, 433-435 (2009).
26. Firmann, M. Prevalence of obesity and abdominal obesity in the Lausanne population. *BMC Cardiovasc. Disord.* **8**, 330 (2008).
27. Sabatti, C. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genet.* **41**, 35-46 (2009).
28. Nelis, M. Genetic structure of Europeans: a view from the north-east. *PLoS One* **4**, e5472 (2009).
29. Balkau, B., Eschwege, E., Tichet, J. & Marre, M. Proposed criteria for the diagnosis of diabetes: evidence from a French epidemiological study (D.E.S.I.R.). *Diabetes Metab.* **23**, 428-434 (1997).
30. Sladek, R. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-885 (2007).
31. Jernas, M. Regulation of carboxylesterase 1 (CES1) in human adipose tissue. *Biochem. Biophys. Res. Commun.* **383**, 63-67 (2009).
32. Yeo, G.S. A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nature Genet.* **20**, 111-112 (1998).
33. Vaisse, C., Clement, K., Guy-Grand, B. & Froguel, P. A frameshift mutation in human MC4R is associated with a dominant form of obesity. *Nature Genet.* **20**, 113-114 (1998).
34. Willer, C.J. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41**, 25-34 (2009).
35. Chen, A.Y., Kim, S.E., Houtrow, A.J. & Newacheck, P.W. Prevalence of obesity among children with chronic conditions. *Obesity (Silver Spring)* **18**, 210-213 (2010).
36. Boeka, A.G. & Lokken, K.L. Neuropsychological performance of a clinical sample of extremely obese individuals. *Arch. Clin. Neuropsychol.* **23**, 467-474 (2008).
37. Marioni, J.C. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol.* **8**, R228 (2007).
38. Bengtsson, H., Simpson, K., Bullard, J. & Hansen, K. aroma.affymetrix: a generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. *Department of Statistics, University of California, Berkeley, Technical Report 745* (2008).
39. Bengtsson, H., Irizarry, R., Carvalho, B. & Speed, T.P. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**, 759-67 (2008).
40. Bengtsson, H., Ray, A., Spellman, P. & Speed, T.P. A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. *Bioinformatics* **25**, 861-867 (2009).
41. Huang, J. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics* **1**, 287-299 (2004).
42. Olshen, A.B. & Venkatraman, E.S. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572 (2004).

43. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-63 (2007).
44. Colella, S. et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* **35**, 2013-25 (2007).
45. Wang, K. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665-1674 (2007).
46. Korn, J.M. et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253-60 (2008).
47. Schouten, J.P. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* **30**, e57 (2002).

5 Detection and impact of somatic copy number alterations in cancer

In collaboration between the Ludwig Institute for Cancer Research, Universities of Lausanne and Geneva, the Swiss Institute of Bioinformatics and the CHUV, I have been strongly involved in the comprehensive genomic profiling of metastatic melanoma. More specifically I have been interested in detecting somatic copy number aberrations (SCNAs), how these relate with the gene expression of affected genes and whether recurrent aberrations can be identified in melanoma, both at the gene and pathway level.

The project was centred on the molecular profiling of seven metastatic melanomas and their matched diploid controls with a series of experiments that included karyotyping, CGH and SNP arrays; methylation profiling; and exome and RNA sequencing. To derive a gene expression reference in melanoma samples, we also sequenced cDNA from a pool of normal human melanocytes.

Karyotype and CGH experiments were performed within the Service of Medical Genetics (CHUV); methylation and exon-capture arrays were done within the Department of Genetic Medicine and Development (Geneva); exome and RNA sequencing were outsourced to the Genomic Technology Facility (Lausanne) and University of Zurich; and SNP hybridizations were done at the Genomics Platform of the NCCR "Frontiers in Genetics" (Geneva). Samples, both melanoma cell lines derived from metastases and their matched controls (Peripheral Blood Lymphocytes (PBL) or Epstein-Barr virus transformed lymphoblastoid (EBV) cell lines) were available from the Ludwig Institute for Cancer Research. Melanocytes were a generous gift from Dr. Ghanem Ghanem (Institut Bordet, Belgium). Approval to use these samples for this project was given by the CHUV (Centre Hospitalier Universitaire Vaudois) ethical committee for clinical research.

In this chapter, I will concentrate on the work where I was the main actor. Notably I was in charge of analyzing SNP and CGH arrays for the detection of SCNAs, then to integrate SCNA with gene expression data. My contribution has ranged from developing and implementing algorithms to interpreting the biological results and

making appropriate decisions to refining both the experiments and the analysis. My work was supervised both by Dr. Brian Stevenson and Pr. Victor Jongeneel, with feedback from all other participants. Manual curation and final interpretation of pathways identified as enriched in SCNA was performed by me, with corroboration from Dr. Brian Stevenson and Dr. Donata Rimoldi. To validate the results, I performed a meta-analysis of two published, large melanoma datasets (Stark and Hayward 2007; Gast et al., 2010).

The project has benefited from the expertise of several collaborators. In particular, Dr. Donata Rimoldi (Ludwig Institute for Cancer Research) prepared cell lines and subsequent biological material for the experiments; Dr. Danielle Martinet (CHUV) interpreted karyotype data and performed CGH experiments; Dr. Christian Iseli (Ludwig Institute for Cancer Research) analysed the transcriptome sequencing data to derive raw tag counts at each transcript; and Dr. Mark Ibberson performed independent network analysis to evaluate different analysis methods.

My work shows that while few SCNA are recurrent between samples, these were associated essentially within the same signalling pathways. The results were replicated in the two external datasets, each ten-fold larger than our own melanoma collection. These results have two major implications 1) our analysis is useful to identify recurrent events from both large and small datasets and 2) the pathways and constituent genes we identified might offer new insights in the biology and treatment of melanoma. This can be exemplified with two genes involved in angiogenesis and migration (*FRS2* and *EPHA3*) that have been implicated in other cancers and should be investigated further as potential therapeutic targets in melanoma. This work is fully detailed in the next section and has been submitted for publication in a peer review journal.

5.1 Abstract

Cancer genomes frequently contain somatic copy number alterations (SCNA) that can significantly perturb the expression level of affected genes and thus disrupt pathways controlling normal growth. In melanoma, many studies have focussed on the copy number and gene expression levels of the *BRAF*, *PTEN* and *MITF* genes, but little has been done to identify new genes using these parameters at the genome-wide scale. Using karyotyping, SNP and CGH arrays, and RNA-seq, we have identified SCNA affecting gene expression ('SCNA-genes') in seven human metastatic melanoma cell lines. We showed that the combination of these techniques is useful to identify candidate genes potentially involved in tumorigenesis. Since few of these alterations were recurrent across our samples, we used a protein network-guided approach to determine whether any pathways were enriched in SCNA-genes in one or more samples. From this unbiased genome-wide analysis, we identified 28 significantly enriched pathway modules. Comparison with two large, independent melanoma SCNA datasets showed less than 10% overlap at the individual gene level, but network-guided analysis revealed 66% shared pathways, including all but three of the pathways identified in our data. Frequently altered pathways included WNT, cadherin signalling, angiogenesis and melanogenesis. Additionally, our results emphasize the potential of the *EPHA3* and *FRS2* gene products, involved in angiogenesis and migration, as possible therapeutic targets in melanoma. Our study demonstrates the utility of network-guided approaches, for both large and small datasets, to identify pathways recurrently perturbed in cancer.

Microarray and sequencing data were deposited in NCBI GEO and are available under accession number GSE23056.

5.2 Introduction

Somatic copy number alterations (SCNA) are a recurrent characteristic of malignant cancers¹⁻³. The amplification and subsequent over-expression or, conversely, deletion and loss of expression of key regulators of cell proliferation, senescence or death have been shown in many cases to contribute significantly to the progression from the normal to the malignant state⁴⁻⁷. Therefore, the discovery and characterization of chromosomal regions involved in SCNA and of the genes encoded in them has been a crucial contributor to our understanding of the molecular mechanisms of carcinogenesis.

The methods used to detect and characterize SCNA have evolved significantly over the last decades. Initial cytogenetic observations have been supplemented with Southern blots and quantitative PCR. Almost twenty years ago, the availability of BAC clones delineating a tiling path through the entire human genome made it possible to detect SCNA in a genome-wide fashion, but with limited resolution⁸. More recently, oligonucleotide-based arrays have enabled comparative genome hybridizations (CGH) at high resolution, and CGH has become the method of choice to detect copy-number variations⁹⁻¹¹. A recent SNP-based survey¹ of 3,131 copy-number profiles derived from over 26 different types of cancer has provided a dramatic illustration of the power of high-throughput techniques in distinguishing random alterations in the genome from those that may have a direct impact on tumorigenesis.

Genomic alterations in many tumors, especially at late stages in their development, are so extensive that the copy-number status of individual genes or chromosomal regions can vary over a very wide range of values. A mixture of chromosomal rearrangements and focal expansions can create genomic landscapes that are very difficult to analyze using standard CGH techniques. Moreover, the exact boundaries of SCNA or the expression status of the genes encoded within them are usually not known, precluding a thorough assessment of their impact on the phenotype of the cancer cells. It has recently been proposed that SNP arrays may be better suited for the determination of copy number states in tumor samples because the analysis of data derived from such arrays can make use of allelic imbalance information in addition to hybridization intensity¹².

In the present study, we analyzed the genome-wide copy-number status of seven highly aneuploid metastatic melanoma cell lines and determined the expression of their genes using a sequencing-based approach. We show that a combination of SNP-based and CGH arrays is necessary to obtain a reliable estimate of the true copy-number status of the entire genome in the face of extensive genomic instability, and that the combination of copy-number and expression status provides powerful clues as to the possible role of genes encoded within SCNA in tumorigenesis. Moreover, we show that a protein-based network-guided analysis of SCNA-affected genes with altered expression in our data and two published datasets^{13,14} identifies pathways commonly altered in melanoma.

5.3 Methods

5.3.1. Melanoma samples, DNA and RNA extraction

Melanoma cell lines were established from metastases from patients with cutaneous melanoma and were used at low passage (<10). Donor matched cells were either peripheral blood lymphocytes (PBL) or Epstein-Barr virus transformed lymphoblastoid (EBV) cell lines. EBV cell lines were karyotyped to ensure genome stability and diploidy. Approval to use these samples for this project was given by the CHUV (Centre Hospitalier Universitaire Vaudois) ethical committee for clinical research. Melanoma cell lines were cultured conditions in RPMI-1640 medium supplemented with 10% fetal calf serum (FCS), and no antibiotics. Human foreskin melanocytes were grown in HAM-F10 medium supplemented with 2% FCS, 5% MelanoMax supplement (Gentaur, Belgium) and 6 mM HEPES. EBV cell lines were cultured in IMDM /10% FCS medium. All cultures were without mycoplasma. DNA (Gentra kit, Qiagen) and RNA (guanidinium/cesium chloride gradient) isolation and karyotype preparations were performed from parallel cultures.

5.3.2. Cytogenetic and FISH analysis

Cytogenetic (GTG-banding) and fluorescence in situ hybridization (FISH) metaphase analyses of melanoma cell lines were performed using standard protocols. Dual color FISH was done using a commercially available set consisting of a locus-specific *MDM2* combined with a chromosome 12 centromeric probe (Kreatech Poseidon FISH probe) to distinguish aneuploidy of chromosome 12 and specific locus loss or gain. Chromosomes with homogeneously staining region (HSR) were identified with the analysis of FISH metaphases in inverted digital images. Copy number estimation of HSR was done using FISH interphases.

5.3.3. Comparative genomic arrays (CGH)

CGH arrays were processed according to the manufacturer's protocol (Agilent Technologies, Inc.) and as described in Martinet et al. ¹⁵.

The normalization and detection of copy number aberration is detailed in the Supplemental Information. In brief, signal intensities were normalized using three independent normalization schemes: Loess ¹⁶; PopLowess ¹⁷; and the statistical framework from Chen et al. ¹⁸.

Then probe-level data were segmented using Circular Binary Segmentation ^{19,20} and attributed a discrete copy number to segments using three independent methods 1) a naive scoring-based approach, where outliers relative to the chromosomal baseline are detected using a non-parametric score; 2) the MergeLevels method ²¹ and 3) our own classification algorithm based on Gaussian Mixture Model which models the observed distribution of intensity ratios as a combination of Gaussian distributions that can be subsequently classified into deletion (CN<2), copy neutral event (CN=2), duplication and amplification (CN=3 and CN≥4).

5.3.4. Single Nucleotide polymorphism arrays (SNP)

Illumina 1M SNP arrays

Genomic DNA from each of the 7 melanoma and their matched normal cells (either EBV cell line or PBL); as well as two control melanocytes were genotyped on the Illumina Infinium Human1M-Duo arrays. Aliquots of DNA (30 μ l at 50 ng/ μ l) for each sample were processed according to the manufacturer's protocol (Infinium HD Gemini protocol). Subsequently we used the OverUnder algorithm²² to correct the hybridization log ratios for polyploidy and to attribute a continuous copy number value to each SNP. We estimated that the window size parameter set to 201 SNPs, gave the highest reproducibility between technical replicates (Supplemental Fig. S5A and S5B).

Affymetrix 6.0 arrays

As part of the technical replicate design, we analyzed LAU-Me275 on Affymetrix 6.0 SNP arrays. The experiment was performed in accordance with the manufacturer's instructions. Normalization and copy number prediction were done using the PICNIC algorithm²³.

5.3.5. Transcriptome sequencing

The transcriptome from all seven melanoma as well as a pool of two melanocytes was sequenced using the Roche 454 Titanium technology. mRNA isolation and cDNA preparation were performed following the protocol used by Bainbridge et al.²⁴, with some modifications (See Supplemental Information). 3-5 μ g of cDNA were used for 454 libraries preparation, according to manufacturer's protocol. All experiments produced about 1M single end reads, with a median length of 367 nucleotides (interquartile range 265-436). We derived transcript tag counts using our own published methodology²⁵ (see also Supplemental Information). Using tag counts from the pool of melanocytes, we were able to derive a ratio of expression for each melanoma with respect to these control melanocytes.

5.3.6. Detection of somatic copy number alterations with altered expression (SCNA-genes)

We computed the median copy number at each Refseq gene. To overcome density limitations, we included SNPs that were within 2kb of the gene boundaries. For CGH arrays, we included probes within 3kb of the gene boundaries. We defined SCNA-genes as follows. A gene was flagged as within a focal amplification when its CN, as computed from SNP arrays, was ≥ 4 , the difference in CN relative to the chromosomal arm was ≥ 1 , the gene was diploid (CN=2) in the matched control cell line, and the expression in the melanoma cell line was at least 2-fold greater than that in the control melanocytes. For deletions, a gene needed to have CN < 2 , as detected by CGH, without expression in the melanoma cell line and CN=2 with detected expression in the melanocytes.

5.3.7. Protein network-guided analysis of SCNA

A non-redundant human protein interaction network was generated by combining iRefseq²⁶ and Pathway Commons²⁷ protein interaction databases with functional interactions from Panther pathways²⁸. The resulting network has 21,876 nodes and 376,528 edges and combines interaction data from 15 primary protein interaction databases (BIND, BioGRID, CORUM, DIP, HPRD, IntAct, MINT, MPact, MPPI, OPHID, Reactome, HumanCyc, Cancer Cell Map, IMID and NCI/Nature pathway interaction database). Using a walk trap community algorithm and permutation approaches (with n=1000), we were able to extract clusters of proteins from the network (Details are available in the Supplemental Information).

5.3.8. Pathway analysis

Significance of overlap between the modules and pathways from Panther, Kegg and MSigDB²⁸⁻³⁰ was calculated with an hypergeometric test, P-values were corrected for multiple comparisons by calculating the false discovery rate using the Benjamini and Hochberg procedure³¹.

To reduce the redundancy present in pathway database annotation, we first combined pathways that contained the same SCNA-genes from the same melanoma cell lines, then we reviewed the pathway list to either remove redundancies or un-merge unrelated pathways affected by similar genes. We also excluded KEGG “cancer” annotated pathways.

5.4 Results

5.4.1. CGH and SNP arrays are required to comprehensively document somatic copy-number alterations in metastatic melanoma cell lines

We analyzed seven low-passage melanoma cell lines that were established from metastases (Table 1) together with matched controls from the same patients (see Methods). Karyotyping of the melanoma cell lines revealed extreme levels of aneuploidy. For example, LAU-Me280, the most extensively deleted line, had a per cell content of 34 to 42 chromosomes (median: 40), whereas LAU-Me275, one of the most amplified melanomas, harbored 68 to 81 chromosomes (median 73.5) (Figure 1). Additionally, the presence of many unassigned chromosomal fragments (markers) made it difficult to determine the true level of aneuploidy.

Melanoma	Site	BRAF mutation	Number of chromosomes (karyotype)
LAU-Me280.R.LN	Lymph node	G593M, L597R	34-42
LAU-Me246.M1	Skin	V600E	45-82
LAU-T618A	Skin	wt but NRAS mutation (Q61R)	55-71
LAU-T50B	Skin	V600E	65-71
LAU-T149D	Visceral	V600E	68-81
LAU-Me275	Lymph node	V600E	68-81
LAU-Me235	Skin	K601E	73-103

Table 1 Melanoma cell lines

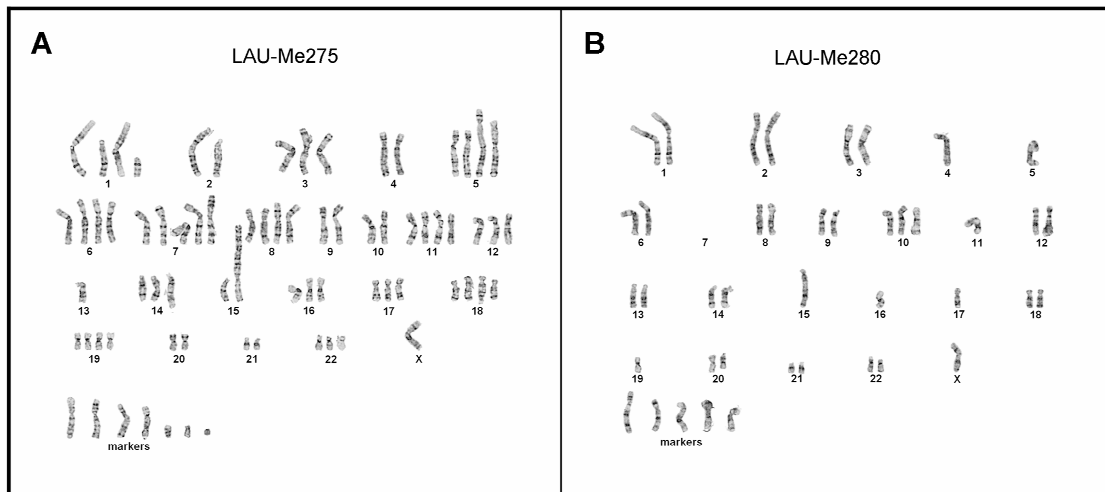


Figure 1 Karyotypes of two malignant melanomas

Representative karyotype (Giemsa stain) for LAU-Me275, one of the most hyperploid melanoma (here 76 chromosomes including 7 markers); and LAU-Me280, the most extensively deleted line (42 chromosomes including 5 markers).

In initial CGH (Agilent 244k) experiments we observed that in LAU-Me275, and other highly hyperploid cell lines, the hybridization ratios between cancer cells and matched controls did not reflect the chromosome-wide aberrations observed in the karyotypes. For example tetraploid regions were measured as triploid or less by the CGH arrays (see Supplemental Fig. S1). We considered whether this was due to the normalization protocol and subsequent segmentation analysis. Using technical replicates of LAU-Me275 DNA, we tested three independent normalization schemes, two of which were specifically developed for cancer genome analysis (see Methods) and found that the methodology proposed by Chen and colleagues¹⁸ was the most reproducible (Spearman correlation 0.96; Supplemental Fig. S2). We then partitioned the genome into regions reflecting copy number changes and assigned copy number using two independent classification methods (see Methods). Since neither of these classification methods gave entirely satisfactory results (see Supplemental Fig. S4 and Supplemental Information), we developed a Gaussian Mixture Model (GMM) approach that was highly reproducible based on a technical replicate analysis (Spearman correlation 0.9).

The GMM method found only 42 regions in the LAU-Me275 genome that were amplified (CN \geq 4; Table 2A). This number was less than expected based on the karyotype analysis, which documented a high number of arm-level chromosome amplifications (Figure 1). Thus, while CGH-based methods are well adapted to document differences in copy number status between the genomes of normal cells derived from different individuals, our results clearly show that they are inadequate to deal with the large-scale rearrangements and amplifications typical of hyperploid cancer cells. The most likely reason is that the total DNA content of cancer cells is too different from that of normal cells to allow a robust experimental normalization. Given this limitation, we asked whether SNP arrays might be better suited to detect chromosome-wide changes in a highly amplified genome.

We hybridized DNA from LAU-Me275 to Illumina 1M SNP arrays and analyzed the signals using the OverUnder algorithm²², which uses minor allele frequencies in heterozygous loci to improve copy number estimation. These results correlated well (Spearman correlation 0.77) with a technical replicate analyzed on the Affymetrix SNP platform (see Supplemental Fig. S5C and S5D), and indicated that 18,251 genes in the LAU-Me275 genome had a copy number of at least four (Table 2B). Within this group, 132 genes had undergone focal amplifications of at least 10-fold. These SNP-based results were more consistent with the karyotype observations. For example, CGH had predicted two copies of chromosome 7p (Chr7p) and three copies of Chr7q (Figure 2), while the SNP results indicated three and five copies, respectively (Figure 2 and Figure 3), which was more consistent with the cytogenetic data (Figure 1).

Therefore, based on our findings with LAU-Me275, we determined the SCNA in the six other melanoma cell lines using both CGH (Agilent 244k) and SNP (Illumina 1M) array platforms. Identification of genes within amplifications or deletions was determined as described in Methods, and the number of SCNA for each cell line is given in Table 2. In all cases CGH predicted more deletions than did SNP arrays, agreeing with our initial observations using LAU-Me275. Also, with the exception of the LAU-Me280 cell line, amplifications were better predicted by SNP arrays. This bias is evident in a graphical representation of the intersection between CGH and SNP predictions as presented in Supplemental Fig. S7. These results confirmed our conclusion that CGH is more suitable for detecting deletions while SNP arrays are better for identifying amplifications.

A

CGH arrays	LAU-Me280	LAU-Me246	LAU-T618A	LAU-T50B	LAU-T149D	LAU-Me275	LAU-Me235	Unique gene count
Deletion	3668	4281	986	3656	108	122	1059	10711
Arm-level amplification	222	0	549	99	998	42	0	1884
Focal amplification	0	0	0	26	379	0	4	409

B

SNP arrays	LAU-Me280	LAU-Me246	LAU-T618A	LAU-T50B	LAU-T149D	LAU-Me275	LAU-Me235	Unique gene count
Deletion	2294	3157	2	113	70	2	39	5544
Arm-level amplification	0	0	16584	1033	3477	16398	10384	19496
Focal amplification	213	0	978	438	894	1853	161	4055

Table 2 Number of genes affected by SCNA in seven melanoma cell lines

Number of genes affected by somatic deletions, arm-level amplifications (≥ 4 copies but < 1 copy above the chromosome arm baseline) and focal amplifications (≥ 4 copies and ≥ 1 copy above the chromosome arm baseline), as measured using SNP or CGH arrays.

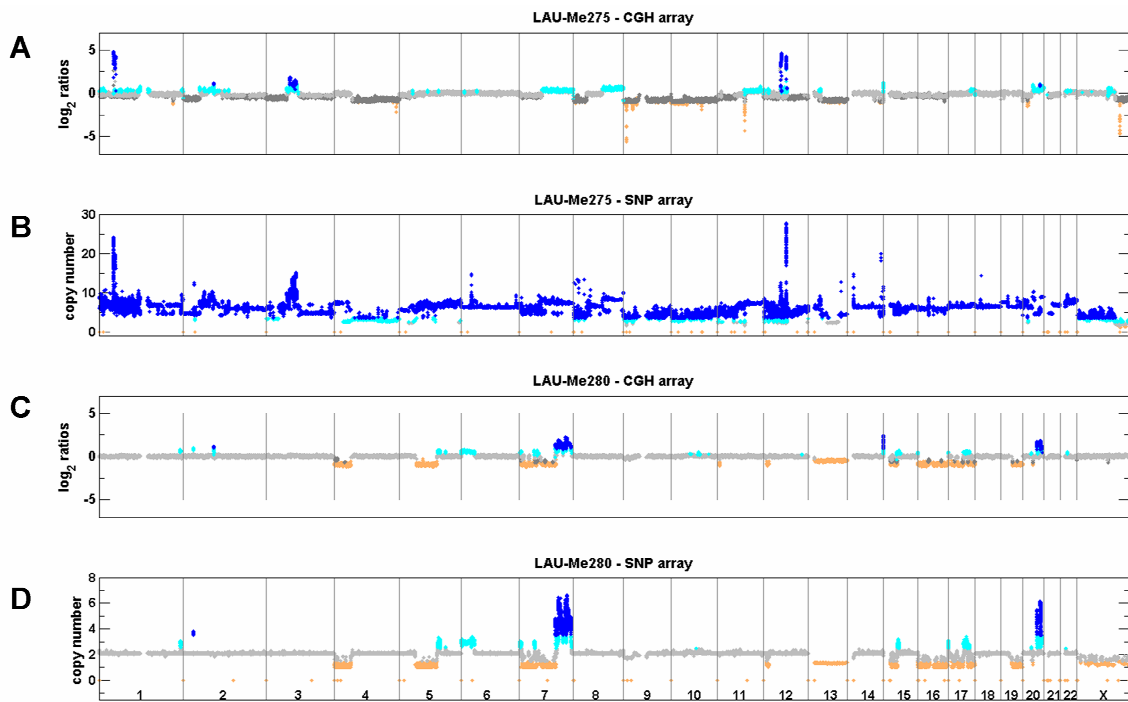


Figure 2 Copy number analysis using CGH and SNP arrays

A. and **B.** shows the analysis of LAU-Me275 on CGH and SNP arrays. **C.** and **D.** shows results for LAU-Me280. Probe/SNP are plotted as a function of their genomic position on the X axis. Y axis for CGH arrays corresponds to hybridization ratios. Y axis for SNP arrays corresponds to the predicted copy number. Colors indicate a copy number state (orange < 2 copies; gray = 2 copies; cyan = 3 copies; dark blue > 3 copies). Dark gray in the CGH panels indicates regions identified as diploid in the analysis, but where the karyotype analysis indicated copy neutral or deleted states, possibly due to cell heterogeneity.

5.4.2. Few SCNA-genes are recurrent in different melanoma cell lines

A potential problem with SCNA studies performed in isolation is that they cannot assess the expression status of the genes contained within the altered genomic regions. Amplified genes are not necessarily highly expressed, and the exact boundaries of deletions may or may not encompass a gene of interest. We reasoned that the combination of precise copy number determination and gene expression measurement would allow us to highlight with much higher confidence those genes whose expression is affected by SCNA (SCNA-genes) in the melanoma cell lines. We therefore analyzed gene expression in each of the melanoma cell lines by RNA-seq using the Roche/454 pyrosequencing method. Additionally, we performed RNA-seq on a pool of epidermal melanocytes to determine a reference level of expression for each gene in normal melanocytic cells (see Methods).

We first looked for genes within focal amplifications with at least two-fold over-expression relative to the reference melanocytes (Supplemental Table S1). Only *KIAA0090*, a protein coding gene of unknown function not previously associated with cancer, was affected in three melanomas (Supplemental Table S2). A further 56 genes were altered in two melanomas, but the only known cancer-related gene was *MDM2*, an oncogene previously demonstrated to be amplified in sarcoma, glioma, colorectal and other cancers including melanoma^{5,32}. In LAU-Me275, *MDM2* was 3.9-fold over-expressed relative to melanocytes and had a copy number greater than ten (as predicted by SNP array). By contrast, in LAU-T50B, *MDM2* was predicted by SNP array to be diploid (CN=2), and by CGH to be duplicated (CN=3). This potential difference between these cell lines is intriguing because they were derived from metastases surgically removed from the same patient at a 12 year interval. We therefore determined the copy number status of the *MDM2* gene in these two samples using fluorescent in situ hybridization (FISH). In LAU-Me275, the fluorescence signal indicated that at least 8 *MDM2* copies were present at the locus on Chr12 in addition to a homogeneously staining region on Chr5 (Figure 3A and Figure 3B) which is in agreement with results from the SNP arrays. In LAU-T50B, FISH revealed a total of four *MDM2* copies, two on Chr12 and two located on an unidentified chromosome (Figure 3C and Figure 2Figure 3D), which is higher than the

copy number estimated by CGH and SNP arrays. Re-investigation of the raw SNP data for LAU-T50B showed that there was indeed a small amplification signal at *MDM2*, but this had not been detected using our optimization parameters (see Supplemental Information). This highlights the challenge of determining optimal parameters that are usable on a genome-wide scale for all samples in a study.

We next derived a list of genes within deletions detected by CGH that were expressed in melanocytes but not in the melanoma cell lines (Supplemental Table S1). We reasoned that such genes are likely to be enriched for melanocyte functions that have been lost during tumorigenesis. The vast majority of such genes (554) were private to a single melanoma sample; seventy genes were shared by two samples; and only ten genes were shared by three melanomas: *ADAMTSL1*, *ARMC4*, *DLL1*, *HSD17B3*, *LOC441177*, *OSTCL*, *PARK2*, *PLXDC2*, *SLC24A2* and *ULBP3* (see Supplemental Table S2).

Altogether, we identified a total of 1,710 SCNA-genes affected by amplification or deletion (summarized in Supplemental Table S2; complete dataset in Supplemental Table S3). To determine the relevance to melanoma of this set of altered genes from our small sample set, we compared it to gene lists in two published studies that used larger melanoma collections^{13,14}. These two studies provided a list of genes recurrently affected by amplification or deletion in 76 (primary and metastatic) and 60 (metastatic) melanoma samples, respectively. Only 196 of our 1,089 amplified genes (p-value <0.001, Figure 4A), and 17 of our 634 genes within deletions (p-value <0.007, Figure 4B) were present in the Stark and Hayward or Gast et al. datasets. Surprisingly, the number of genes common to the two published gene sets was also small (27 amplified genes and 2 genes within deletions; p-values <0.005), and demonstrates the difficulty to identify commonly affected genes relevant to tumor progression even within larger melanoma collections.

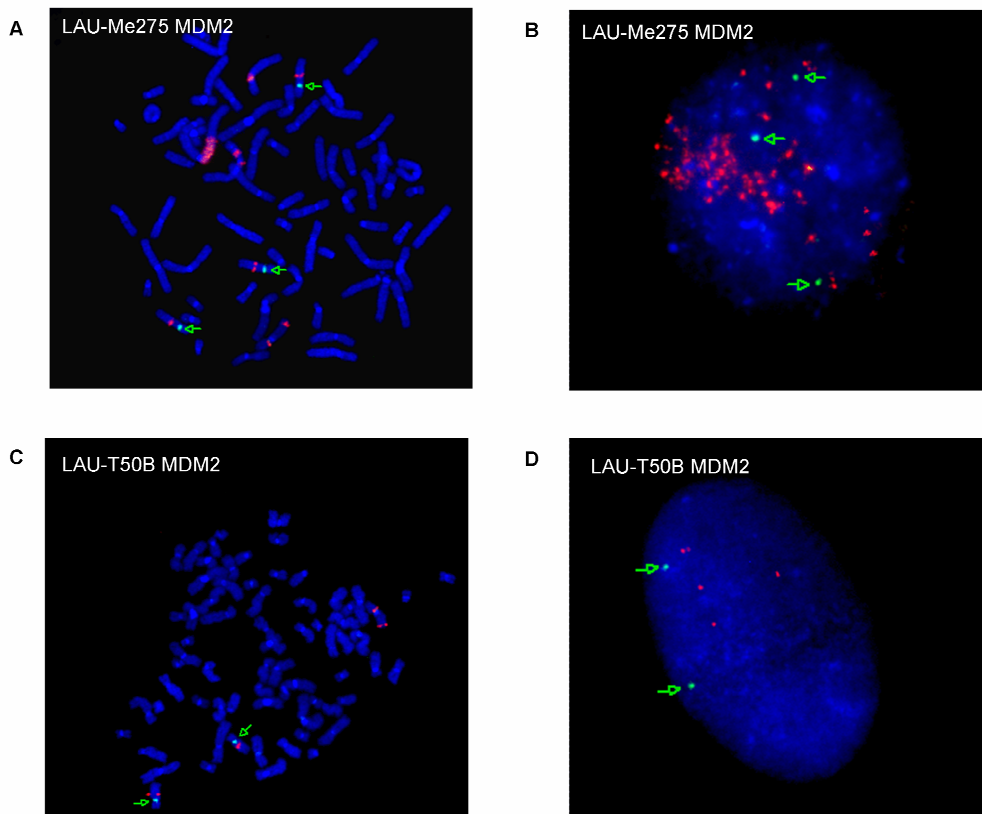


Figure 3 Determination of MDM2 copy number by FISH

The *MDM2* gene was assayed in two melanoma samples (LAU-Me275 and LAU-T50B) derived from the same patient. Panels A and C show a metaphase and B and D an interphase. *MDM2* probe is in red; centromere-specific probe is in green. FISH shows amplification for both LAU-Me275 (more than eight copies) and LAU-T50B (four copies). Metaphase-FISH helps to identify homogeneously staining regions and Interphase-FISH to estimate the copy number.

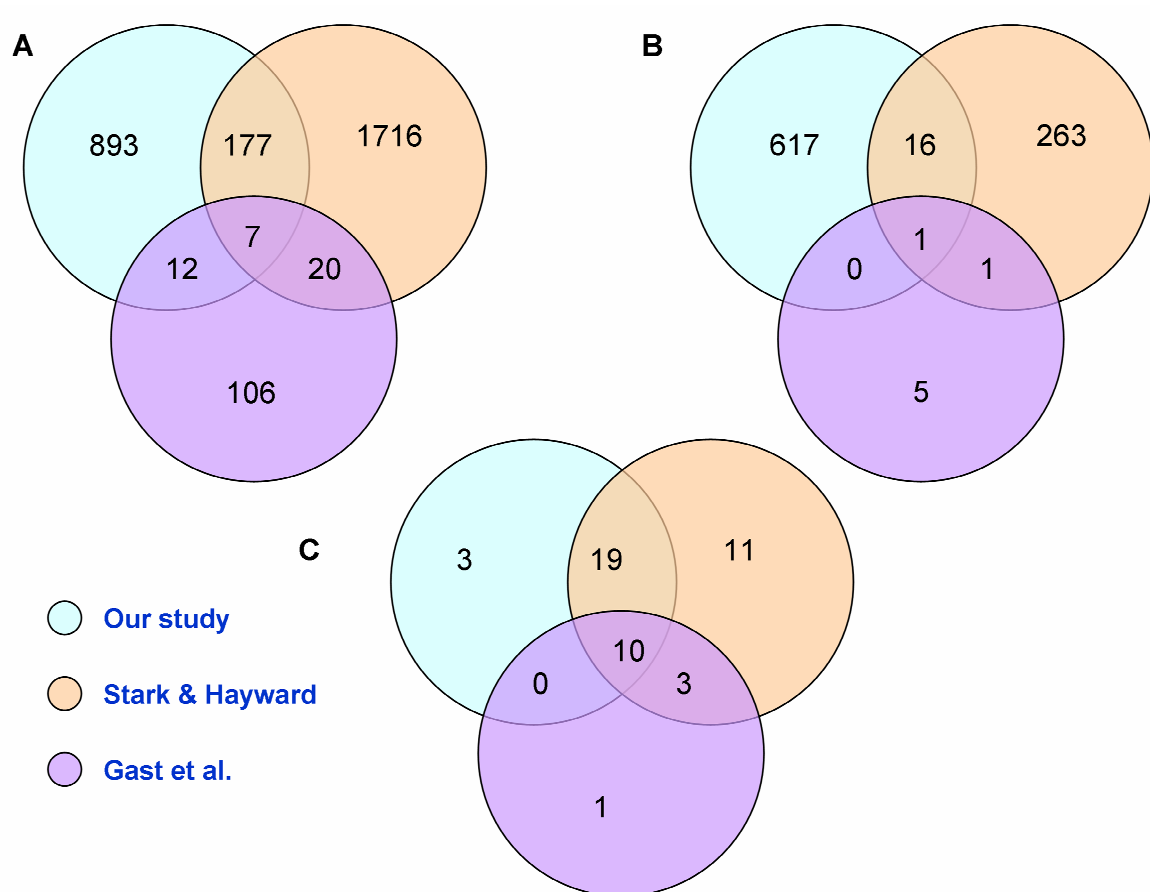


Figure 4 Intersection between our dataset and two published datasets of SCNA-genes and derived pathways

A. Intersection between amplified genes in published melanoma datasets (Stark and Hayward 2007; Gast et al., 2010) and our list of over-expressed genes within focal amplifications **B.** Intersection between genes within homozygous deletions from the Stark and Hayward and Gast et al. datasets and our list of non-expressed genes within deletions **C.** Intersection between pathways found significantly affected by SCNAs from our analysis of the three datasets.

5.4.3. Pathways significantly enriched in SCNA-genes are recurrent in melanoma

As described above, we found that very few of the SCNA-genes were altered in more than one melanoma cell line, which is not unexpected given the small number of samples in our dataset. An idea popular in the current literature is that signaling pathways, rather than individual genes, are recurrently perturbed in cancer³³. To determine whether the SCNA-genes from different melanoma cell lines shared membership of one or more cellular pathways, we investigated whether the proteins encoded by the SCNA-genes were connected in known human protein interaction networks (see Methods). Out of a total of 1,563 proteins analyzed, 377 (24%) were connected within the network, and clustering of the proteins based on the topology of the sub-networks identified 14 protein networks, or 'clusters', containing at least five significantly connected members. For each cluster, we identified genes belonging to known signaling and metabolic pathways including nine clusters that significantly overlapped known pathways (FDR \leq 0.05; listed in Supplemental Table S4). Following detailed manual annotation, the resulting pathways were ranked according to the number of contributing melanomas and to the number of SCNA-genes involved in the pathway. The pathways common to at least four melanoma samples are shown in Table 3 (for the complete list see Supplemental Table S4). Interestingly, the vast majority of the recurrent pathways we identified involve signal transduction and have been implicated in one or more cancer types. In addition, that we identified ten pathways common to at least five of the melanoma samples confirms the idea that protein network-guided analysis is a good method for detecting recurrently affected pathways in small datasets.

In our search for genes recurrently affected by SCNA, we found only ~10% overlap between our list of SCNA-genes and those derived from studies with much larger sample sizes^{13,14}. To determine if this was also true at the level of pathways, we performed a protein network-guided analysis as described above on each of these datasets using the published gene lists (neither study originally presented this type of analysis). Detailed annotation and comparison of the results for each dataset is given in Supplemental Table S4. In contrast to what we had found at the gene level, all but three of our pathway modules were also present among those identified from one or

both of the published gene datasets (Figure 4C). Ten pathway modules (angiogenesis, EGF, ERBB, integrin signaling, long term potentiation, MAPK, natural killer cell mediated toxicity, PDGF, regulation of actin cytoskeleton and VEGF) were common to all three datasets, and the combined overlap with our pathway dataset was 66%. Thus, the majority of pathways defined by SCNA affected genes in our melanoma samples were recurrent in the three datasets, whereas the individual genes were not.

An additional benefit of the protein network-guided approach is that it generates a list of genes affected by SCNA that contributed significantly to a given pathway (Supplemental Table S4). Although two-thirds of the pathways were common between our dataset and the published datasets, only two genes, *NRAS* and *BRAF*, were present in all three (Supplemental Table S5). Of the genes shared by two datasets, four were components of the angiogenesis pathway, including *EPHA3* and *FRS2*. We noted also that several members of the WNT (*WNT3A*, 4, 5B, 7A, 9A, 11, 16) or cadherin (*CDH2*, 4, 9, 12, 17, 18, 19) gene families were affected by SCNA in only one dataset, further reinforcing the idea that different genes can potentially alter the same pathway (WNT or cadherin) in different melanoma samples.

Pathway	#Melanomas	Genes	#genes
G protein signaling	6	ADORA1, ADRA2A, CHRM1, CHRM5, DRD2, GNAO1, GNB3, GNG4, HTR1F, OPRL1, PLCB2, RGS10, RGS11, RGS14, RGS19	15
WNT signaling (includes Apoptosis and Hedgehog signaling)	6	CDH19, CDH2, CDH4, DVL1, FRAT1, FZD8, PCDH17, PCDH9, SFRP1, WNT11, WNT16, WNT2B, WNT4, WNT5B	14
Cadherin signaling	6	ACTG2, CDH19, CDH2, CDH4, FZD8, PCDH17, PCDH9, WNT11, WNT16, WNT2B, WNT4, WNT5B	12
Melanogenesis	6	CAMK2A, CAMK2G, DVL1, FZD8, NRAS, WNT11, WNT16, WNT2B, WNT4, WNT5B	10
Angiogenesis	5	BRAF, DVL1, EFNB2, EPHA3, EPHB2, FGF1, FRS2, NRAS, PIK3R3, PRKCZ, SFRP1, WNT2B, WNT5B	13
Axon guidance (migration and adhesion)	5	CDK5, EFNB2, EPHA3, EPHB2, EPHB6, FES, NRAS	7
MAPK signaling	5	DUSP1, DUSP12, DUSP2, FGF1, FGF14, FGFR4, MAPK9	7
TGF beta signaling	5	ACVRL1, AMHR2, FOXH1, LEFTY1, SMAD9, TGFB1, TLL2	7
Alzheimer disease	5	CHRM1, CHRM5, PKN3, PRKCZ	4
FGF signaling	5	FGF1, FGF14, FGFR4, FRS2	4
Calcium signalling	4	CAMK2A, CAMK2G, CHRM1, CHRM5, GNAO1, GRIN2C, PRKCZ, RGS10, RGS11, RGS14, RGS19	11
Huntington_disease (vesicle-mediated transport)	4	ACTG2, CLTB, GRIN2B, GRIN2C, GRIN3A, KALRN	6
Neuroreceptor (Muscarinic, Metabotropic)	4	GRIN2B, GRIN2C, GRIN3A, KCNQ2, PKN3, PRKCZ	6
Cell cycle (G1 progression)	4	CCNA1, CDC20, CDC26, CDKN2B, HDAC1	5

Table 3 Pathways identified by network-guided analysis

5.5 Discussion

Our goal to identify somatic copy number aberrations in metastatic melanoma cell lines revealed extreme levels of aneuploidy characteristic of this cancer type^{34,35}, and complicated the application of standard CGH array protocols⁹⁻¹¹. Nevertheless, using our GMM method we were able to demonstrate that although CGH arrays fail to identify all large-scale amplifications, they are able to detect deletions very efficiently, including genes having lost expression compared to melanocytes (Table 2 and Supplemental Fig. S7). Conversely, SNP arrays, which measure hybridization intensities for both alleles at heterozygous loci, allow the consideration of an additional parameter (the so-called B-allele frequency) and greatly improve the measurement of DNA copies beyond the normal diploid complement (as implemented in the OverUnder algorithm, Attiyeh et al., 2009; see Supplemental Fig. S6). We did notice, however, that this algorithm systematically detected deletions located in sub-telomeric regions for both tumors and controls, which indicates a systematic bias and suggests that the algorithm is optimized to detect duplications and amplifications but not deletions. Therefore, it can be argued that CGH and SNP techniques should be combined to obtain a reliable assessment of all copy number states from deletion to high-level focal amplification.

To enrich for genes that might be involved in the oncogenic process, we focused on two groups: focally amplified genes that were over-expressed relative to melanocytes; and deleted genes with no expression in the melanoma cell lines, but that were expressed in normal melanocytes. In the first group, *MDM2*^{5,32} was the only cancer gene amplified and over-expressed in more than one melanoma sample. Comparison of genes amplified in our samples with published gene lists from two large melanoma studies (Stark and Hayward 2007; Gast et al., 2010) while revealing very little overlap (Figure 4) did identify *BRAF*, *MDM2*, and *NRAS*, genes known to be important in melanoma^{5,32,36-42}. In the second group, ten genes were deleted in three of the melanoma samples (Supplemental Table S2). These genes are located on Chr6q25, Chr6q27, Chr9 or Chr10p, consistent with previous observations that both arms of chromosomes 9 and 10 and Chr6q frequently undergo hemizygous deletion or copy neutral LOH in melanoma¹⁴. Of the ten genes, the Parkinson's disease-associated gene *PARK2* has been recently described as a tumor suppressor

gene in glioblastoma and other malignancies⁴³, while *DLL1*, *HSD17B3* and *ULBP* have been reported to be associated with cancer, although not as tumor suppressors⁴⁴⁻⁵⁰. Experimental investigation will be required to determine if any of these ten genes performs an anti-oncogenic function in melanoma cells. The only deleted gene common to our study and those of Stark and Hayward and Gast et al. was *PTEN*, a tumor suppressor gene already known to be deleted in melanoma^{7,51}.

In an alternative approach to detect recurrent events in these samples, we used a protein network-guided analysis^{33,52-56} to identify pathways affected by SCNA-genes in the seven melanoma cell lines. In contrast to the low level of recurrence in these melanoma samples at the individual gene level, we found that six pathways were shared by five of the samples, and four pathways (G protein, WNT, cadherin signaling and melanogenesis) were common to six (Table 3). Several of these pathways are highly relevant to melanoma (e.g. MAPK, cadherin and FGF signaling) and have also emerged from cDNA expression studies⁵⁷, lending support to our results. G proteins transduce signals from G protein-coupled receptors (GPCRs), the largest family of membrane receptors involved in signal transduction, and whose over-expression in tumors can contribute to tumor progression, angiogenesis and metastasis⁵⁸. Alteration of G proteins could impact the activities of GPCRs key to melanocytic cells, such as *MC1R* (melanocortin receptor), chemokine (e.g. *CXCR2*), and endothelin receptors.⁵⁹ The recent identification of activating mutations in two G protein alpha subunits, *GNAQ* and *GNA11*, in a large proportion of uveal melanomas^{60,61}, further underscores the relevance of this class of proteins to melanoma.

Although annotated as distinct pathways, WNT, cadherin signaling and melanogenesis shared six SCNA-genes in common (*FZD8* and several members of the WNT family). This may reflect interactions between these pathways, an interplay between the WNT and cadherin pathways is known to exist⁶², or may be a consequence of poor pathway annotation. The cadherin pathway controls cell-adhesion and plays a role in invasion and metastasis⁶³. WNT (and Hedgehog) control development and growth in the embryo; aberrant activation of their transcriptional components ultimately affects cell fate, proliferation, and migration⁶⁴⁻⁶⁶. The only common non-signaling pathway was melanogenesis. Melanoma develops from melanocytes, cells highly specialized in the synthesis of melanin

pigment, a process that requires a complex enzymatic machinery and unique organelle structures⁶⁷. Our pathway analysis predicted that melanoma SCNA affect melanogenesis. Loss of pigmentation in metastases compared to primary tumors is commonly observed in cutaneous melanoma, and although not completely understood, it can be brought about by different mechanisms, such as premature degradation of melanogenic proteins⁶⁸ or downregulation of *MITF* transcription program⁶⁹. Our study suggests that SCNA may also contribute to these alterations. An unexpected pathway that emerged from our analysis, and perhaps merits further exploration, is neurotransmission. These results suggest an involvement of neuronal pathways in melanoma, possibly related to the neural crest origin of melanocytes. Lending support to this hypothesis, the metabotropic glutamate receptor *GRM1* has recently been implicated in the development of spontaneous melanoma in a mouse model, and an autocrine glutamate/*GRM1* loop has been described in human melanoma⁷⁰.

Comparison of the pathways generated from SCNA-genes in our data and genes affected by copy number changes in two published datasets (Stark and Hayward¹⁴ and Gast et al.¹³) revealed a high level of overlap, much higher than we expected based on the number of commonly affected genes (Figure 4). An explanation for this outcome is that different genes within the same pathway are affected in different datasets, and the commonalities are apparent only at the pathway level. The number of affected genes in a given pathway would be expected to increase with increasing sample size, and this is largely the case between our data and those of Stark and Hayward, but not in the Gast et al dataset (Supplemental Table S4). The reason for the low number of SCNA affected genes and corresponding pathways in the latter case may be the high stringency criteria employed in their analysis¹³.

The angiogenesis pathway was one of ten common to all three datasets. Its up-regulation is a well-known hallmark of cancer⁷¹, and it has long been proposed as a target for therapeutic treatment^{72,73}. Activation signals for angiogenesis include vascular endothelial growth factor (VEGF) and acidic fibroblast growth factor (FGF), and both were in our list of significantly affected pathways (Table 3) and within our analysis of the Stark and Hayward (VEGF and FGF) and Gast et al (VEGF) datasets (Supplemental Table S4). Two genes in this pathway, *EPHA3* and *FRS2*, were

designated SCNA-genes in both our dataset and in Stark and Hayward, and were annotated as amplified, in skin-derived tumors, in the Cancer Genome Project dataset ^{5,74}.

In our analysis *EPHA3*, an ephrin tyrosine kinase receptor, was both focally amplified and over-expressed only in LAU-Me275. However, *EPHA3* was highly over-expressed in LAU-T149D and LAU-Me246 (Supplemental Table S3) and amplified in LAU-T618A (CN=6.4), LAU-Me235 (CN=4) and LAU-T50B (CN=4.2). *EPHA3* is recurrently mutated in adenocarcinoma ^{75,76} and has been implicated in renal carcinoma, glioblastoma, colorectal, breast and lung cancer ⁷⁶⁻⁸⁰. Mutations in *EPHA3* have been detected in melanoma ⁸¹, and several ephrin-derived peptide antigens (from *EPHA2*, *EPHA3* and *EPHB6*) can be recognized by cancer-specific cytotoxic T-cells ⁸²⁻⁸⁵. In addition, the feasibility of specific *EPHA3* targeting has been reported ⁸⁶. These observations indicate that *EPHA3* might be a promising target for therapeutic treatment in melanoma and other cancers.

FRS2, fibroblast growth factor receptor substrate 2, is an adaptor that acts downstream of a limited number of receptor tyrosine kinases, in particular FGF and neurotrophin receptors, RET and ALK, and plays a major role in tumorigenesis ⁸⁷. Dey and coworkers ⁸⁸ recently targeted the FGF receptors (FGFR) using tyrosine kinase inhibitors to decrease the activity of AKT and ERK kinases, inducing apoptosis in breast cancer cell lines. FGFR inhibition is highly relevant to melanoma, where autocrine stimulation via FGF2/FGFR1 constitutes a pivotal role in proliferation and survival ⁸⁹. *FRS2* has been suggested as a therapeutic target in cancer ⁹⁰ and because of its downstream activities to FGFR and other receptors, it might offer new insights in melanoma treatment. In our data *FRS2* was both focally amplified and over expressed in two melanoma samples (LAU-T149D and LAU-Me275) and amplified (CN=4) in two additional melanomas (LAU-T618A and LAU-Me235). Inspection of its amplification status in larger melanoma collections would be useful to confirm its potential role as a target of interest in melanoma.

In conclusion, we have identified SCNA-genes and pathways potentially altered in our metastatic melanoma samples and two published datasets (Stark and Hayward 2007; Gast et al., 2010) which should be investigated by screening larger tumor collections and in functional studies. Two SCNA-genes, *EPHA3* and *FRS2*, emerged

from our analysis as potential therapeutic targets. These genes were replicated in our analysis of the two published melanoma collections, have been extensively studied in other cancer types, and thus might offer new insights in the treatment of malignant melanoma.

5.6 Accession numbers

Microarray and sequencing data were deposited in NCBI GEO and are available under accession number GSE23056.

5.7 Authors and affiliations

Armand Valsesia^{1,2,3}, Donata Rimoldi¹, Danielle Martinet⁴, Mark Ibberson², Paola Benaglio³, Muriel Gaillard⁴, Mireille Pidoux⁴, Blandine Rapin⁴, Carlo Rivolta³, Ioannis Xenarios², Andrew J.G. Simpson⁵, Stylianos E. Antonarakis⁶, Jacques S. Beckmann^{3,4}, C. Victor Jongeneel^{1,2,7}, Christian Iseli^{1,2*}, and Brian J. Stevenson^{1,2*}

1. Ludwig Institute for Cancer Research, Lausanne, Switzerland
2. Swiss Institute of Bioinformatics, Lausanne, Switzerland
3. Department of Medical Genetics, University of Lausanne, Switzerland
4. Service of Medical Genetics, CHUV, Lausanne, Switzerland
5. Ludwig Institute for Cancer Research, New York, USA.
6. Department of Genetic Medicine and Development, University of Geneva, Switzerland
7. Institute for Genomic Biology and National Center for Supercomputing Applications, University of Illinois, USA

* Corresponding authors: Christian.Iseli@unil.ch, Brian.Stevenson@unil.ch

5.8 Acknowledgements

We thank Keith Harshman and Johann Weber (Genomic Technologies Facility, Lausanne) for helpful discussions; Anguraj Sadanandam (EPFL-ISREC, Lausanne) and Bob Strausberg (Ludwig Institute for Cancer Research, New York) for useful comments on the manuscript; Ghanem Ghanem (Institut Bordet, Belgium) for his generous gift of melanocytes; Patrick Descombes and the Genomics Platform of the NCCR "Frontiers in Genetics" for performing the SNP experiments; Sven Bergmann (Department of Medical Genetics, Lausanne) for access to his computing resources; and Bastian Peter for system administration. Part of the computation was performed on the cluster at the Vital-IT computing center, Lausanne. We are very grateful to Katja Muehlethaler (Ludwig Institute for Cancer Research, Lausanne) and Marianne Wicht, Nathalie Besuchet-Schmutz, and Anne-Claude Magnin (CHUV cytogenetic laboratory) for excellent technical assistance. This work was supported by the Ludwig Institute for Cancer Research Ltd.

5.9 Author contributions

AV, AJGS, BJS, CI, CVJ, DR, JSB, and SEA designed and supervised the study; AV designed algorithms and analyzed the data; CI analyzed transcriptome data; AV and MI performed the network-guided analysis; XI provided expertise for the network-guided analysis. AV, BJS and DR interpreted the data. DR managed and prepared melanoma samples; DM analyzed karyotype and FISH data; DM and JSB supervised CGH experiments; BR, MG and MP performed CGH, FISH and karyotype experiments; CR supervised SNP experiments; PB prepared materials for SNP experiments and transcriptome sequencing. AV, BJS, CVJ and DR wrote and edited the manuscript. All authors read and approved the manuscript.

5.10 References

1. Beroukhi, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905 (2010).
2. Baudis, M. Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC Cancer* **7**, 226 (2007).
3. Mitelman, F., Johansson, B. & Mertens, F. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. (2010).
4. Lockwood, W.W. et al. DNA amplification is a ubiquitous mechanism of oncogene activation in lung and other cancers. *Oncogene* **27**, 4615-24 (2008).
5. Futreal, P.A. et al. A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183 (2004).
6. Bignell, G.R. et al. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res* **17**, 1296-303 (2007).
7. Stahl, J.M. et al. Loss of PTEN promotes tumor development in malignant melanoma. *Cancer Res* **63**, 2881-90 (2003).
8. Kallioniemi, A. et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818-21 (1992).
9. Bignell, G.R. et al. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res* **14**, 287-95 (2004).
10. Pinkel, D. & Albertson, D.G. Array comparative genomic hybridization and its applications in cancer. *Nat Genet* **37 Suppl**, S11-7 (2005).
11. Kallioniemi, A. CGH microarrays and cancer. *Curr Opin Biotechnol* **19**, 36-40 (2008).
12. LaFramboise, T. et al. Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Comput Biol* **1**, e65 (2005).
13. Gast, A. et al. Somatic alterations in the melanoma genome: a high-resolution array-based comparative genomic hybridization study. *Genes Chromosomes Cancer* **49**, 733-45 (2010).
14. Stark, M. & Hayward, N. Genome-wide loss of heterozygosity and copy number analysis in melanoma using high-density single-nucleotide polymorphism arrays. *Cancer Res* **67**, 2632-42 (2007).
15. Martinet, D. et al. Subtelomeric 6p deletion: clinical and array-CGH characterization in two patients. *Am J Med Genet A* **146A**, 2094-102 (2008).
16. Smyth, G.K. & Speed, T. Normalization of cDNA microarray data. *Methods* **31**, 265-73 (2003).
17. Staaf, J., Jonsson, G., Ringner, M. & Vallon-Christersson, J. Normalization of array-CGH data: influence of copy number imbalances. *BMC Genomics* **8**, 382 (2007).
18. Chen, H.I. et al. A probe-density-based analysis method for array CGH data: simulation, normalization and centralization. *Bioinformatics* **24**, 1749-56 (2008).
19. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-72 (2004).
20. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-63 (2007).
21. Willenbrock, H. & Fridlyand, J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**, 4084-91 (2005).
22. Attiyeh, E.F. et al. Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res* **19**, 276-83 (2009).

23. Greenman, C.D. et al. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **11**, 164-75 (2010).
24. Bainbridge, M.N. et al. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**, 246 (2006).
25. Jongeneel, C.V. et al. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc Natl Acad Sci U S A* **100**, 4702-5 (2003).
26. Razick, S., Magklaras, G. & Donaldson, I.M. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9**, 405 (2008).
27. Bader, G., Cerami, E., Demir, E., Gross, B. & Sander, C.
<http://www.pathwaycommons.org>.
28. Mi, H., Guo, N., Kejariwal, A. & Thomas, P.D. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res* **35**, D247-52 (2007).
29. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
30. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
31. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289-300 (1995).
32. Oliner, J.D., Kinzler, K.W., Meltzer, P.S., George, D.L. & Vogelstein, B. Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature* **358**, 80-3 (1992).
33. Cerami, E., Demir, E., Schultz, N., Taylor, B.S. & Sander, C. Automated network analysis identifies core pathways in glioblastoma. *PLoS One* **5**, e8918 (2010).
34. Ozisik, Y.Y. et al. Cytogenetic findings in 21 malignant melanomas. *Cancer Genet Cytogenet* **77**, 69-73 (1994).
35. Becher, R., Gibas, Z., Karakousis, C. & Sandberg, A.A. Nonrandom chromosome changes in malignant melanoma. *Cancer Res* **43**, 5010-6 (1983).
36. Davies, H. et al. Mutations of the BRAF gene in human cancer. *Nature* **417**, 949-54 (2002).
37. Miller, A.J. & Mihm, M.C., Jr. Melanoma. *N Engl J Med* **355**, 51-65 (2006).
38. Bloethner, S. et al. Effect of common B-RAF and N-RAS mutations on global gene expression in melanoma cell lines. *Carcinogenesis* **26**, 1224-32 (2005).
39. Wellbrock, C. et al. V599EB-RAF is an oncogene in melanocytes. *Cancer Res* **64**, 2338-42 (2004).
40. Karasarides, M. et al. B-RAF is a therapeutic target in melanoma. *Oncogene* **23**, 6292-8 (2004).
41. Wellbrock, C. et al. Oncogenic BRAF regulates melanoma proliferation through the lineage specific factor MITF. *PLoS One* **3**, e2734 (2008).
42. Omholt, K., Platz, A., Kanter, L., Ringborg, U. & Hansson, J. NRAS and BRAF mutations arise early during melanoma pathogenesis and are preserved throughout tumor progression. *Clin Cancer Res* **9**, 6483-8 (2003).
43. Veeriah, S. et al. Somatic mutations of the Parkinson's disease-associated gene PARK2 in glioblastoma and other human malignancies. *Nat Genet* **42**, 77-82 (2010).
44. Ayyanan, A. et al. Increased Wnt signaling triggers oncogenic conversion of human breast epithelial cells by a Notch-dependent mechanism. *Proc Natl Acad Sci U S A* **103**, 3799-804 (2006).

45. Ma, D. et al. Aberrant expression of Notch signaling molecules in patients with immune thrombocytopenic purpura. *Ann Hematol* **89**, 155-61 (2010).
46. Bauer, S. et al. Activation of NK cells and T cells by NKG2D, a receptor for stress-inducible MICA. *Science* **285**, 727-9 (1999).
47. Waldhauer, I. & Steinle, A. NK cells and cancer immunosurveillance. *Oncogene* **27**, 5932-43 (2008).
48. Pende, D. et al. Major histocompatibility complex class I-related chain A and UL16-binding protein expression on tumor cell lines of different histotypes: analysis of tumor susceptibility to NKG2D-dependent natural killer cell cytotoxicity. *Cancer Res* **62**, 6178-86 (2002).
49. Montgomery, R.B. et al. Maintenance of intratumoral androgens in metastatic prostate cancer: a mechanism for castration-resistant tumor growth. *Cancer Res* **68**, 4447-54 (2008).
50. Yin, D. et al. High-resolution genomic copy number profiling of glioblastoma multiforme by single nucleotide polymorphism DNA microarray. *Mol Cancer Res* **7**, 665-77 (2009).
51. Wu, H., Goel, V. & Haluska, F.G. PTEN signaling pathways in melanoma. *Oncogene* **22**, 3113-22 (2003).
52. Klijn, C. et al. Identification of networks of co-occurring, tumor-related DNA copy number changes using a genome-wide scoring approach. *PLoS Comput Biol* **6**, e1000631 (2010).
53. Jones, S. et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801-6 (2008).
54. Menashe, I. et al. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. *Cancer Res* **70**, 4453-9 (2010).
55. Murohashi, M. et al. Gene set enrichment analysis provides insight into novel signalling pathways in breast cancer stem cells. *Br J Cancer* **102**, 206-12 (2010).
56. Heiser, L.M. et al. Integrated analysis of breast cancer cell lines reveals unique signaling pathways. *Genome Biol* **10**, R31 (2009).
57. Hoek, K. et al. Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas. *Cancer Res* **64**, 5270-82 (2004).
58. Dorsam, R.T. & Gutkind, J.S. G-protein-coupled receptors and cancer. *Nat Rev Cancer* **7**, 79-94 (2007).
59. Lee, H.J., Wall, B. & Chen, S. G-protein-coupled receptors and melanoma. *Pigment Cell Melanoma Res* **21**, 415-28 (2008).
60. Van Raamsdonk, C.D. et al. Mutations in GNA11 in Uveal Melanoma. *N Engl J Med* (2010).
61. Van Raamsdonk, C.D. et al. Frequent somatic mutations of GNAQ in uveal melanoma and blue naevi. *Nature* **457**, 599-602 (2009).
62. Heuberger, J. & Birchmeier, W. Interplay of cadherin-mediated cell adhesion and canonical Wnt signaling. *Cold Spring Harb Perspect Biol* **2**, a002915 (2010).
63. Cavallaro, U. & Christofori, G. Cell adhesion and signalling by cadherins and Ig-CAMs in cancer. *Nat Rev Cancer* **4**, 118-32 (2004).
64. Taipale, J. & Beachy, P.A. The Hedgehog and Wnt signalling pathways in cancer. *Nature* **411**, 349-54 (2001).
65. Lucero, O.M., Dawson, D.W., Moon, R.T. & Chien, A.J. A re-evaluation of the "oncogenic" nature of Wnt/beta-catenin signaling in melanoma and other cancers. *Curr Oncol Rep* **12**, 314-8 (2010).

66. Klaus, A. & Birchmeier, W. Wnt signalling and its impact on development and cancer. *Nat Rev Cancer* **8**, 387-98 (2008).
67. Raposo, G. & Marks, M.S. Melanosomes--dark organelles enlighten endosomal membrane transport. *Nat Rev Mol Cell Biol* **8**, 786-97 (2007).
68. Halaban, R. et al. Aberrant retention of tyrosinase in the endoplasmic reticulum mediates accelerated degradation of the enzyme and contributes to the dedifferentiated phenotype of amelanotic melanoma cells. *Proc Natl Acad Sci U S A* **94**, 6210-5 (1997).
69. Hoek, K.S. & Goding, C.R. Cancer stem cells versus phenotype-switching in melanoma. *Pigment Cell Melanoma Res* **23**, 746-59 (2010).
70. Namkoong, J. et al. Metabotropic glutamate receptor 1 and glutamate signaling in human melanoma. *Cancer Res* **67**, 2298-305 (2007).
71. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).
72. Folkman, J. Tumor angiogenesis: therapeutic implications. *N Engl J Med* **285**, 1182-6 (1971).
73. Kerbel, R.S. Tumor angiogenesis. *N Engl J Med* **358**, 2039-49 (2008).
74. Bignell, G.R. et al. Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893-8 (2010).
75. Ding, L. et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-75 (2008).
76. Davies, H. et al. Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res* **65**, 7591-5 (2005).
77. Pasquale, E.B. Eph receptors and ephrins in cancer: bidirectional signalling and beyond. *Nat Rev Cancer* **10**, 165-80 (2010).
78. Dalgliesh, G.L. et al. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463**, 360-3 (2010).
79. Sjoblom, T. et al. The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-74 (2006).
80. Stephens, P. et al. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat Genet* **37**, 590-2 (2005).
81. Balakrishnan, A. et al. Novel somatic and germline mutations in cancer candidate genes in glioblastoma, melanoma, and pancreatic carcinoma. *Cancer Res* **67**, 3545-50 (2007).
82. Chiari, R. et al. Identification of a tumor-specific shared antigen derived from an Eph receptor and presented to CD4 T cells on HLA class II molecules. *Cancer Res* **60**, 4855-63 (2000).
83. Tatsumi, T. et al. Disease stage variation in CD4+ and CD8+ T-cell reactivity to the receptor tyrosine kinase EphA2 in patients with renal cell carcinoma. *Cancer Res* **63**, 4481-9 (2003).
84. Alves, P.M. et al. EphA2 as target of anticancer immunotherapy: identification of HLA-A*0201-restricted epitopes. *Cancer Res* **63**, 8476-80 (2003).
85. Jin, M. et al. Erythropoietin-producing hepatocyte B6 variant-derived peptides with the ability to induce glioma-reactive cytotoxic T lymphocytes in human leukocyte antigen-A2+ glioma patients. *Cancer Sci* **99**, 1656-62 (2008).
86. Vearing, C. et al. Concurrent binding of anti-EphA3 antibody and ephrin-A5 amplifies EphA3 signaling and downstream responses: potential as EphA3-specific tumor-targeting reagents. *Cancer Res* **65**, 6745-54 (2005).
87. Gotoh, N. Regulation of growth factor signaling by FRS2 family docking/scaffold adaptor proteins. *Cancer Sci* **99**, 1319-25 (2008).

88. Dey, J.H. et al. Targeting fibroblast growth factor receptors blocks PI3K/AKT signaling, induces apoptosis, and impairs mammary tumor outgrowth and metastasis. *Cancer Res* **70**, 4151-62 (2010).
89. Wang, Y. & Becker, D. Antisense targeting of basic fibroblast growth factor and fibroblast growth factor receptor-1 in human melanomas blocks intratumoral angiogenesis and tumor growth. *Nat Med* **3**, 887-93 (1997).
90. Sato, T. & Gotoh, N. The FRS2 family of docking/scaffolding adaptor proteins as therapeutic targets of cancer treatment. *Expert Opin Ther Targets* **13**, 689-700 (2009).

6 Outlook

In this outlook, I summarize the utility of the methods I developed in the context of rare variant analysis, then discuss the implications and the new challenges that arose from the obesity and melanoma studies. Finally, I will expose my personal views about the future of medical genetics.

6.1 *Utility of methods for CNV analysis*

6.1.1 Mining medical cohorts for rare CNVs

In chapter 3, I presented novel methods to identify CNVs from Affymetrix 500k SNP arrays and to compare the predictions from different methods. Back in 2006, Affymetrix arrays were considered as a cutting-edge technology, today this platform is seen as obsolete due to its lack of coverage for dynamic genomic regions (i.e. repeat-rich regions, segmental duplications). To overcome this limitation, the Affymetrix generation (Affymetrix 6.0 SNP array) is a hybrid array that combines SNPs as well as probes covering known CNV regions. Nevertheless very large number of individuals (at least 100,000) have been genotyped on Affymetrix 500k arrays and extensive clinical phenotypic measurements have been recorded. These data have been extensively used in SNP-based genome-wide association studies but have definitely been not brought to their full potential for CNV analyses. Also recent development in CNV-detection algorithms were heavily focused on the newest chips. Thus generic algorithms, like GMM, are important to analyse the previous generation arrays. Analysis of several large genotyping cohorts (from these older arrays), could potentially be useful for rare variant discovery. There might be intermediate (>50Kb) or long (>100Kb) CNVs, with rare frequency but strong impact on clinical phenotypes that could be discovered and provide new insights about their contribution to disease. This can be exemplified with the obesity project where the identified rare variant was associated to the disease with a very high penetrance.

6.1.2 CNV-based genome-wide association studies

My PhD work has been useful to catalogue CNVs in the Swiss general population. This CNV dataset has already proved valuable in several studies (obesity, narcolepsy) and will continue to play an important role both in medical studies and in bioinformatics analyses. Notably, the exomes from 500 CoLaus individuals have been sequenced at the Wellcome Trust Sanger Institute and my set of CNV predictions will be used to tune and validate algorithms predicting indels from sequencing data. Then association studies on cardiovascular traits will be performed using this comprehensive set of CNVs (as predicted from sequencing data). Since CNV-based GWA using the CoLaus microarray CNV predictions have been so far unsuccessful, association with a more comprehensive CNV set, as detected from sequencing, might provide some new insights in cardiovascular disease.

I anticipate that more CNV-based association studies will be performed in the future. There was a trend to focus on SNPs so far, because SNP genotyping methods were more reliable than CNV detection methods; genotyping arrays were indeed designed for SNP analysis whereas there are some strong limitations (i.e. genome coverage, experimental noise) that challenge CNV detection. Now that we are revising our estimates about the fraction of CNVs to be discovered in the genome, I expect that microarray technology will evolve to include more markers that facilitate the identification of such CNVs. In addition, technologies like Nanostring seem very promising for CNV analysis at targeted loci in large number of samples, and we will probably witness many new CNV detection technologies being developed in the next few years. Similarly, new association methods (or strategies) could be developed and prove more powerful than the simple linear regression currently used in quantitative trait association studies. Statistical genetics is still a recent field, with less than 20 years expertise in handling large-scale datasets. In contrast, methods developed 40 years ago in econometrics only recently started to be used in Genetics. Thus it is reasonable to assume that more statistical methods will be borrowed from other fields (like electrical engineering, climatology, computer vision etc...) and applied to statistical genetics. This is even more likely considering genetics is a multidisciplinary area which has attracted over the past few years a plethora of scientists from different quantitative fields (e.g. physicists, mathematicians, computer scientists...).

6.2 Perspectives in Obesity

6.2.1 Follow-up studies

Our work on morbid obesity has led to three major questions to follow-up: 1) The deletion encompasses 28 genes. Since the very same locus was found associated to schizophrenia and autism, some of these genes may, individually or more likely collectively, contribute to these different disorders. A follow up study on gene expression is in progress for deletion- and duplication-carriers and diploid controls. This study is expected to highlight the genes that differ in expression between the three groups and could result in new hypothesis to be tested with the inactivation or activation of certain gene in model organisms (providing orthologous genes can be identified). 2) Another obesity-associated gene (*SH2B1*) is located 800Kb downstream from the 16p11 deletion. *SH2B1* encodes an adaptor protein involved in insulin and leptin signalling; and regulates the satiety effect. It will be interesting to study whether there is a *cis-effect* between *SH2B1* and the 16p11 genes. 3) About 110 individuals were found with a reciprocal duplication at 16p11. This duplication is significantly associated with leanness, which is very interesting as it suggests a mirror effect of the deletion. This aspect is also being investigated and we hope being able to publish these results very soon.

6.2.2 Implication for medical genetics research

Beyond the scope of obesity, this study represents a real breakthrough in medical genetics. It challenges the concept of genome-wide association studies; where modest effects are usually detected and where an association hit often does not imply causality. With this study, we showed that cohorts with extreme phenotypes can be enriched in rare variants with very high penetrance. To identify the missing heritability in complex disease, we proposed to identify rare variants in well-documented cohorts with extreme phenotypes, then to perform a follow-up of candidate variants in the general population. This approach might be more powerful than current scans in the general population, where very large sample size are required to detect reliable association signals.

6.3 Perspectives in Melanoma

6.3.1 Investigation of the pathways enriched in SCNAs

Our study of melanoma has been useful to establish a map of pathways enriched in SCNA and gene expression data. But enrichment, even if statistically significant, does not imply the pathway is genuinely perturbed. As such, our list of pathways deserves further exploration notably with functional studies. Among this list of pathways, we found WNT, angiogenesis, MAPK, FGF signalling and many other examples, which were previously well-known in melanoma development and progression. This definitely lends support to our results.

Among the pathways, for which the implication in melanoma is not yet fully understood, we identified neurotransmission and melanogenesis. The neurotransmission pathway may relate to the neural origin of melanocytes and could play a role in melanoma progression with the activation of other signalling pathways (via G-coupled receptors). Supporting this hypothesis, the glutamate receptor *MC1R*, has been shown to play a crucial role in the development of spontaneous melanoma in mice model ¹.

Melanogenesis is mediated by a network of complex metabolic pathways responsible for the secretion of melanin by melanocytes. Loss of pigmentation is common in metastatic melanoma compared to primary tumours. Melanogenesis products confer resistance to chemotherapy and radiotherapy in malignant melanomas ^{2,3}. All but one of the seven melanoma samples were obtained following treatment (mostly immunotherapy and chemotherapy; only one patient was treated with radiotherapy), thus assuming the pigmentation pathway is activated, there might have been a selection on a melanoma subpopulation resistant to these different treatments.

But only one sample is pigmented, so it not clear whether this pathway is indeed activated. Because accumulation of melanin is a complex and delicate process (the protein needs to be synthesized, post-processed, exported and accumulated), protein quantification analyses are needed to check whether melanin is indeed produced by the cell but cannot be exported. Additional experiments, such as spectrophotometry, could be used to determine the quantity of oxygen free radicals in the cells (these products are usually damaging for the cell, but were shown to interact

with chemotherapy agents in metastatic melanoma). If the pathway is not perturbed (or activated), then the SCNA-enrichment could be the consequence of an early event in the primary tumor (loss of function) or just be result of poor pathway annotation. However in the case this pathway is indeed active in these samples, the cell lines will be very valuable for drug design studies (i.e. to test new compounds inhibiting melanogenesis, this is a very active research topic in melanoma^{2,4-6}) and possibly to investigate the SCNA-affected genes (other than *MITF*) that could play a key role in the pathway destabilization.

6.3.2 Further characterization of *FRS2* and *EPHA3*

From our results, the products of two genes (*FRS2* and *EPHA3*) have emerged as potential therapeutic targets for melanoma. Both genes participate to the angiogenesis pathway and are frequently mutated in melanoma and other cancer types. We confirmed these genes were amplified in two large, external melanoma collections. Several inhibitors of angiogenesis already exists^{7,8}, but several angiogenesis factors can be expressed in melanoma and their expression vary greatly^{9,10}. Thus developing new inhibitors that could complement the action of existing ones may prove a more efficient way to treat patients. The feasibility of targeting these two genes has been previously demonstrated^{11,12}, but it may still be necessary to further characterizes these two genes in cancer model organisms. For example, the Ludwig Institute for Cancer Research (at the Brussels branch) recently established a mice melanoma model and performed gene expression profiling. Investigation of gene expression pattern under different conditions (i.e. treatment) in the model, might confirm the importance of *FRS2* and *EPHA3*.

6.3.3 Incorporating results from exome sequencing

The pathway map established from the SCNA study will be refined with the results from the exome sequencing of the same samples. We anticipate being able to strengthen connections within the detected pathways by the identification of additional genes and hope this will improve our understanding about how the different affected genes can potentially deregulate pathways.

In parallel to the analysis of somatic mutations, the genome from the matched diploid control has been sequenced and we are investigating recurrent germline mutations with possible impact on melanoma predisposition. Already, we identified some interesting candidates that are mutated in six of the patients and that are involved in DNA repair pathways. Validation of these genes in a larger (control) population is being pursued and should clarify the relevance of these candidates.

6.3.4 Implications of our methodology for cancer genomics

More generally, our study has several implications in cancer genomics. First, we demonstrate that combining platforms and selecting genes with both copy number alterations and perturbed expression is a means to enrich for putative candidate genes. Second, we show that gene alterations tend to be private (between our samples and across external melanoma collections) which challenges the identification of novel and recurrent candidate cancer genes. To overcome this limitation, we show that commonalities can be found at the pathway level with a network-guided approach. We demonstrate this approach is applicable to both large and small dataset analysis.

In the past, studies have focused on the identification of a single gene driving the cancer progression and very little has been done to identify groups of genes that collectively disrupt signalling pathways. Such network-guided approach could be applied to the analysis of primary tumours (or precursor cells) and could help to understand the pathways and their key regulators that contribute to tumorigenesis. For example, in melanoma research investigation of *BRAF* and *NRAS* mutations have been extensively done in nevi cells (which are considered as melanoma precursors). But to date, there is no report about comprehensive profiling at the gene and pathway level for these cells. In collaboration with dermatologists from the Bern

University hospital, a grant proposal has been submitted to carry such analyses in nevi cells.

6.4 The future of human genetics

The past few years have witnessed tremendous discoveries in the field of medical genetics. These discoveries were achieved with the advent of microarrays both CGH and SNP, and more recently with massively parallel sequencing. There are several examples with the Hapmap project ^{13,14}, the first map of CNV in the general population ¹⁵, the sequencing of several genomes ¹⁶, and the myriad of genome-wide association studies ¹⁷⁻¹⁹. Much more is to come with the release of the 1000 Genomes project data ^{20,21} and in the coming months of the 1000 Cancer Genomes project. These progresses have significantly expanded the list of loci associated to disease. For example, in 2007 the most important susceptibility gene for type 2 diabetes was *TCF7L2*, in 2008 the *FTO* gene was found from GWAs ²². Since 2008, 123 loci have been discovered ²³, their respective contribution is not yet fully understood and no gene was reported for nine of these loci. What I see as the most important challenge in the next decade(s) is to make sense of these GWA associations. Finding novel associations is necessary as it helps to find new biomarkers either for diagnosis or population stratification purpose. But efforts should also be focused on understanding the biology and the real contribution of these associated loci. This will involve both functional studies in model organisms and a better characterization of the effect of genetic variants to the phenotype (for example, through investigation of gene expression). Longitudinal experiments in carriers of the risk/protective variant will also be interesting to perform.

I anticipate that more and more sequencing analysis will be done in the coming years. Sequencing technologies are constantly evolving, each new generation producing even more and potentially longer reads, which facilitates the analysis. Targeted re-sequencing with sequence capture arrays or barcode technologies offers both the accuracy and high throughput needed for validation in large-scale project. A barrier to sequencing projects was (and is) the cost, but it is becoming more affordable every year. The Achon Genomics Xprize (<http://genomics.xprize.org>) is a \$10 million prize, awarded to the first team that can sequence 100 human genomes within ten days; with 98% coverage; less than one error in every 100,000 bases; and

for less than \$10,000 per genome sequenced. This prize will definitely stimulate innovation and help reducing running costs. It constitutes a fascinating challenge that will undoubtedly benefit to scientific community.

Sequencing has already started to give new insights in population genetics and in medical genetics. It will definitely help with orphan diseases, where there are no candidate genes and with very rare diseases, where it is difficult to collect and analyse a large number of samples.

In term of clinical diagnostics, it seems unlikely sequencing will be used as a routine because 1) the cost is still too large compared to microarray analysis and 2) the data interpretation is challenging. However, I think sequencing might be more commonly applied in clinical research to screen patients with atypical symptoms.

6.5 References

1. Namkoong, J. et al. Metabotropic glutamate receptor 1 and glutamate signaling in human melanoma. *Cancer Res* **67**, 2298-305 (2007).
2. Slominski, A., Zbytek, B. & Slominski, R. Inhibitors of melanogenesis increase toxicity of cyclophosphamide and lymphocytes against melanoma cells. *Int J Cancer* **124**, 1470-7 (2009).
3. Wood, J.M. et al. What's the use of generating melanin? *Exp Dermatol* **8**, 153-64 (1999).
4. Slominski, A., Paus, R. & Mihm, M.C. Inhibition of melanogenesis as an adjuvant strategy in the treatment of melanotic melanomas: selective review and hypothesis. *Anticancer Res* **18**, 3709-15 (1998).
5. Riley, P.A. Melanogenesis and melanoma. *Pigment Cell Res* **16**, 548-52 (2003).
6. Pawelek, J., Korner, A., Bergstrom, A. & Bologna, J. New regulators of melanin biosynthesis and the autodestruction of melanoma cells. *Nature* **286**, 617-9 (1980).
7. Mahabeleshwar, G.H. & Byzova, T.V. Angiogenesis in melanoma. *Semin Oncol* **34**, 555-65 (2007).
8. Rofstad, E.K. & HalsÅr, E.F. Vascular Endothelial Growth Factor, Interleukin 8, Platelet-derived Endothelial Cell Growth Factor, and Basic Fibroblast Growth Factor Promote Angiogenesis and Metastasis in Human Melanoma Xenografts. *Cancer Research* **60**, 4932-4938 (2000).
9. Mattei, S. et al. Expression of cytokine/growth factors and their receptors in human melanoma and melanocytes. *Int J Cancer* **56**, 853-7 (1994).
10. Westphal, J.R. et al. Angiogenic balance in human melanoma: expression of VEGF, bFGF, IL-8, PDGF and angiostatin in relation to vascular density of xenografts in vivo. *Int J Cancer* **86**, 768-76 (2000).
11. Vearing, C. et al. Concurrent binding of anti-EphA3 antibody and ephrin-A5 amplifies EphA3 signaling and downstream responses: potential as EphA3-specific tumor-targeting reagents. *Cancer Res* **65**, 6745-54 (2005).
12. Chiari, R. et al. Identification of a tumor-specific shared antigen derived from an Eph receptor and presented to CD4 T cells on HLA class II molecules. *Cancer Res* **60**, 4855-63 (2000).
13. The International HapMap Project. *Nature* **426**, 789-796 (2003).
14. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).
15. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).
16. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
17. Craddock, N. et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713-20 (2010).
18. Dupuis, J. et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* **42**, 105-16 (2010).
19. Sabatti, C. et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* **41**, 35-46 (2009).
20. Durbin, R.M. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
21. Sudmant, P.H. et al. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-6 (2010).
22. Prokopenko, I., McCarthy, M.I. & Lindgren, C.M. Type 2 diabetes: new genes, new understanding. *Trends Genet* **24**, 613-21 (2008).

23. Hindorff LA, J.H., Hall PN, Mehta JP, Manolio TA. . A Catalog of Published Genome-Wide Association Studies. www.genome.gov/26525384

Annexes

Annexe I: CoLaus Supplementary Information.....	176
Supplementary methods	176
Normalization scheme to account for experimental biases	176
Intersection between CNV detection methods	177
Evaluation of the Gaussian Mixture Model using simulated data.....	177
Supplementary figures and tables	178
References	190
Annexe II: Obesity Supplementary Information.....	191
Annexe III: Melanoma Supplementary Information.....	210
Supplementary Methods.....	210
CNV analysis from CGH arrays.....	210
Transcriptome analysis.....	212
Protein network-guided analysis	212
References	213

Annexe I: CoLaus Supplementary Information

Supplementary methods

Normalization scheme to account for experimental biases

Principal Component Analysis applied to the CN status of SNPs across CoLaus individuals revealed four distinct clusters of individuals, which corresponds to the four genotyping centers. To correct this batch effect, we performed normalization within each center and used an increasing number of randomly chosen samples (with equal proportions of males and females).

By running the normalization twice on the same individuals but with 2 independent reference panels; we were able to compute the distance between the same individuals (in the two normalization runs) and to compare it to the distance between random defined a score G as:

$$G = \frac{|\mu(d1) - \mu(d2)|}{\sqrt{std(d1)^2 + std(d2)^2}}$$

where $d1$ is the Euclidean distance between CNV profiles of the same individual, but with respect to two different reference panels for normalization (Supplementary Figure 4). For comparison we also computed $d2$, the Euclidean distance between random pairs of individuals; here μ is the geometric mean and std is the standard deviation.

This score G measures how well the distribution of distances between pairs of replicates separate from the distance distribution of unrelated pairs. It can be seen as an indicator of goodness for a given normalization and is useful to rank normalizations using different number of references. As the number of references increases, the distance between pairs of replicates should become smaller and the separation between replicates and unrelated samples should increase.

We tested normalizations with 30, 120, 200 and 280 references. As expected, the resulting G scores indicated that the normalization improves significantly with the number of references (Supplementary Figure 4). Using 280 references is significantly better than using only 30 references. Using even more references would not significantly improve the normalization; therefore we decided to use 280 references to keep compute times reasonable.

Intersection between CNV detection methods

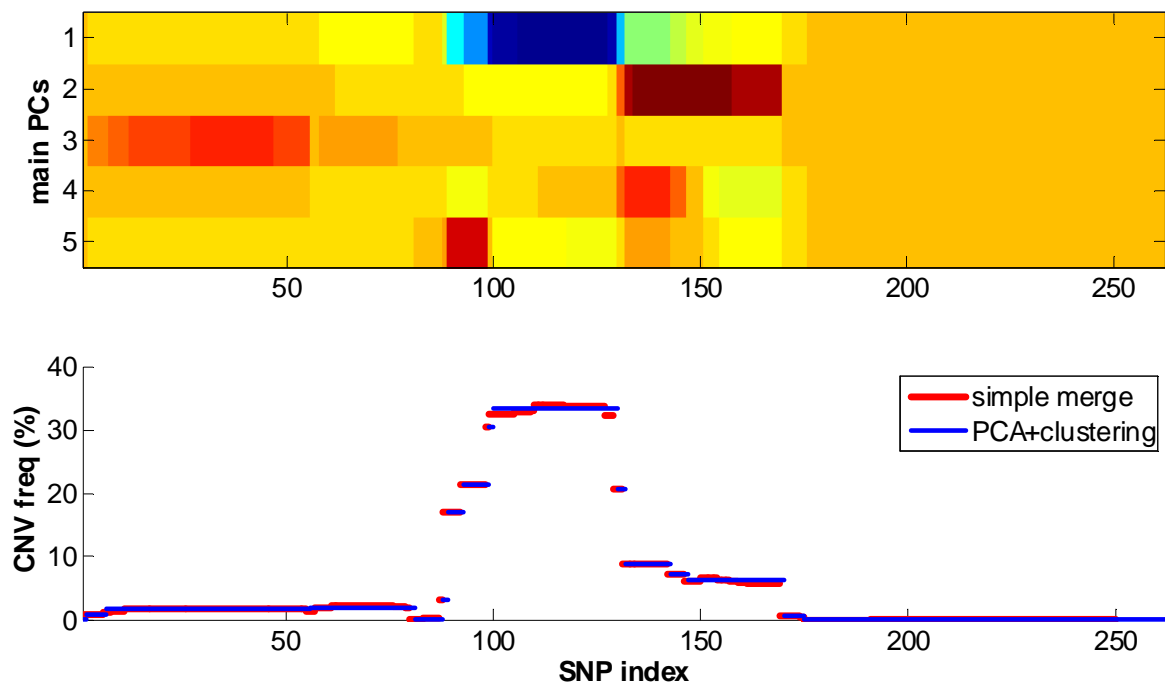
Overlap-based approaches are useful to compare the CNV predictions between two distinct methods. But it becomes tedious when comparing results between more than two methods. A way to simplify this problem is to use a common number of elements to compare between the different methods. The full autosomal dataset was subjected to “LD based pruning”, as implemented in plink [1]. We used a sliding window of 100 SNPs, sliding along in 20 SNP increments. SNPs with a variance inflation factor (VIF) greater than 1.2 were pruned from each window. This procedure identified around 60’000 autosomal SNPs. Then, for each method, we created a binary SNP vector, indicating which SNPs were found variant or copy neutral. This facilitates the comparison and the number of SNPs found variant in one or more methods can easily be computed. We used such approach both for pair-wise comparison and for generating 4-way Venn diagrams.

Evaluation of the Gaussian Mixture Model using simulated data

To evaluate the true and false positive rates of our Gaussian Mixture Model, we generated an artificial dataset composed of 5600 individuals, in which the true copy number status was defined and different levels of Gaussian noise was added. This approach is similar to other studies ([2,3]). In more detail, our artificial dataset consists of a distribution of \log_2 ratios ($n=5600$), where a predefined number of individuals should reflect a deletion state (\log_2 ratio = -1); all remaining individuals are copy neutral thus \log_2 ratios = 0. We then added Gaussian noise to the true copy numbers and tested the ability of our model to correctly detect the underlying copy number state. Supplementary Figure 9 illustrates one test where half of the population is expected to have a deletion.

Supplementary figure 10 shows the performance of our model as a function of the number of individuals sharing the CNV. Each colored curve corresponds to a different level of noise in the data (from low : $\sigma=0.1$ to high : $\sigma=0.6$). This analysis was repeated 50 times for each point of the curves (mean and standard deviation are shown in the figure). The high TPR and very low FPR demonstrate the good performance of our model.

Supplementary figures and tables



Supplementary figure1: Merging SNPs into CNVs using principal component analysis

Top plot shows a principal component analysis (PCA) on a local SNP window (chromosome3:74.5-76.5Mb) across CoLaus individual. The main components are on Y axis and adjacent SNPs are on X axis. The bottom plot shows in red regions obtained from simple merge and in blue, regions from the PCA merge. The Y axis represents CNV frequency in the CoLaus population ($n \approx 5600$)

	PCA merged		Simple merge	
	CNPs	CNVRs	CNPs	CNVRs
GMM	2.4	9.86	0.02	0.38
CBS	1.54	42.43	0.09	24.23
CNAT.allelic	12.4	30.88	8.08	21.19
CNAT.total	0.73	12.71	0.15	7.79

Supplementary table1: genome coverage of CNVs identified by different methods

CNV detection methods are shown as rows, the merging approach as columns. Distinction is made between CNPs (i.e. CNVs with population frequency above 1%) and CNVRs (i.e. CNVs with population frequency below 1% but seen for at least five individuals). The coverage is expressed as the % of the autosomes (there are no predictions for sex chromosomes).

	GMM	CBS	CNAT.allelic	CNAT.allelic
0	-12.21	-10.26	-3.62	-10.42
]0-25]	8.79	6.27	1.20	6.36
]25-50]	7.53	4.75	0.81	4.12
]50-75]	7.25	4.15	3.17	6.25
]75-100]	9.48	13.75	7.03	11.43

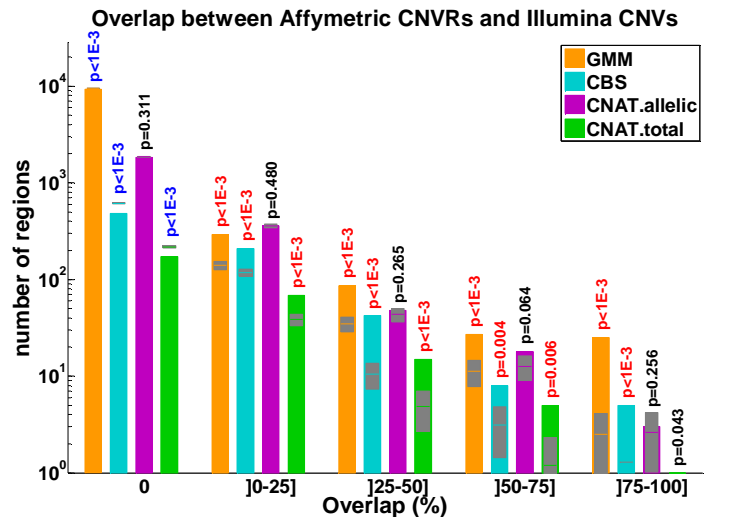
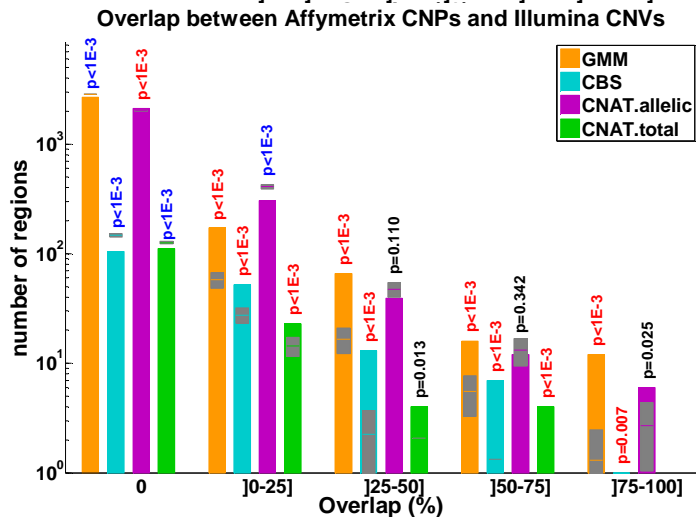
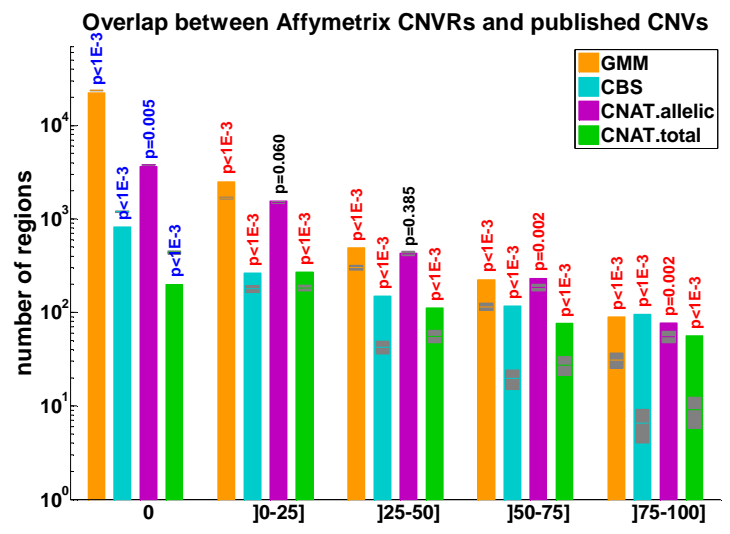
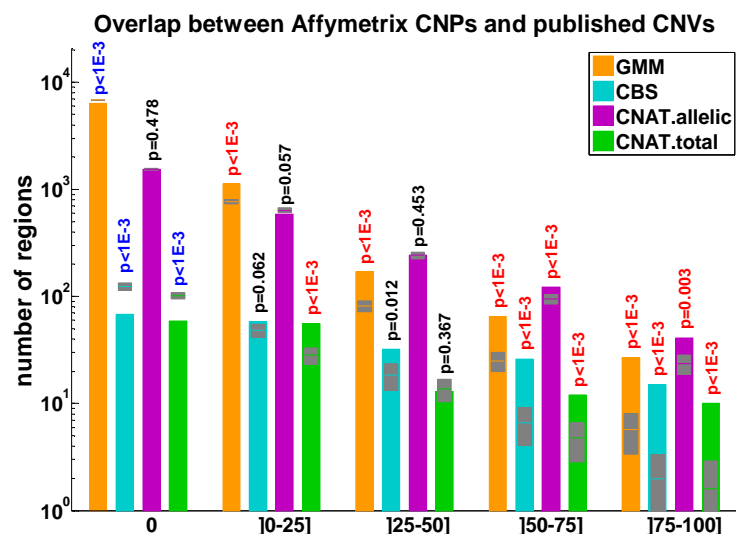
Supplementary table2: T statistic values for overlap between CNVs identified from CoLaus and published CNVs

The T statistic is computed from the difference between observed overlap and expected counts normalized by the standard deviation of expected counts. Expected counts are inferred from the overlap between reshuffled data (n=1000) and published CNVs. T statistics greater than 2.58 are significant with $\alpha=1\%$. Positive (negative) T statistics indicates enrichment (depletion) with respect to the expected counts.

	GMM	CBS	CNAT.allelic	CNAT.allelic
0	-17.68	-12.67	0.31	-9.69
]0-25]	12.14	9.78	-1.01	6.81
]25-50]	12.81	8.45	1.06	8.59
]50-75]	5.49	12.28	1.95	7.78
]75-100]	15.14	6.13	0.34	7.18

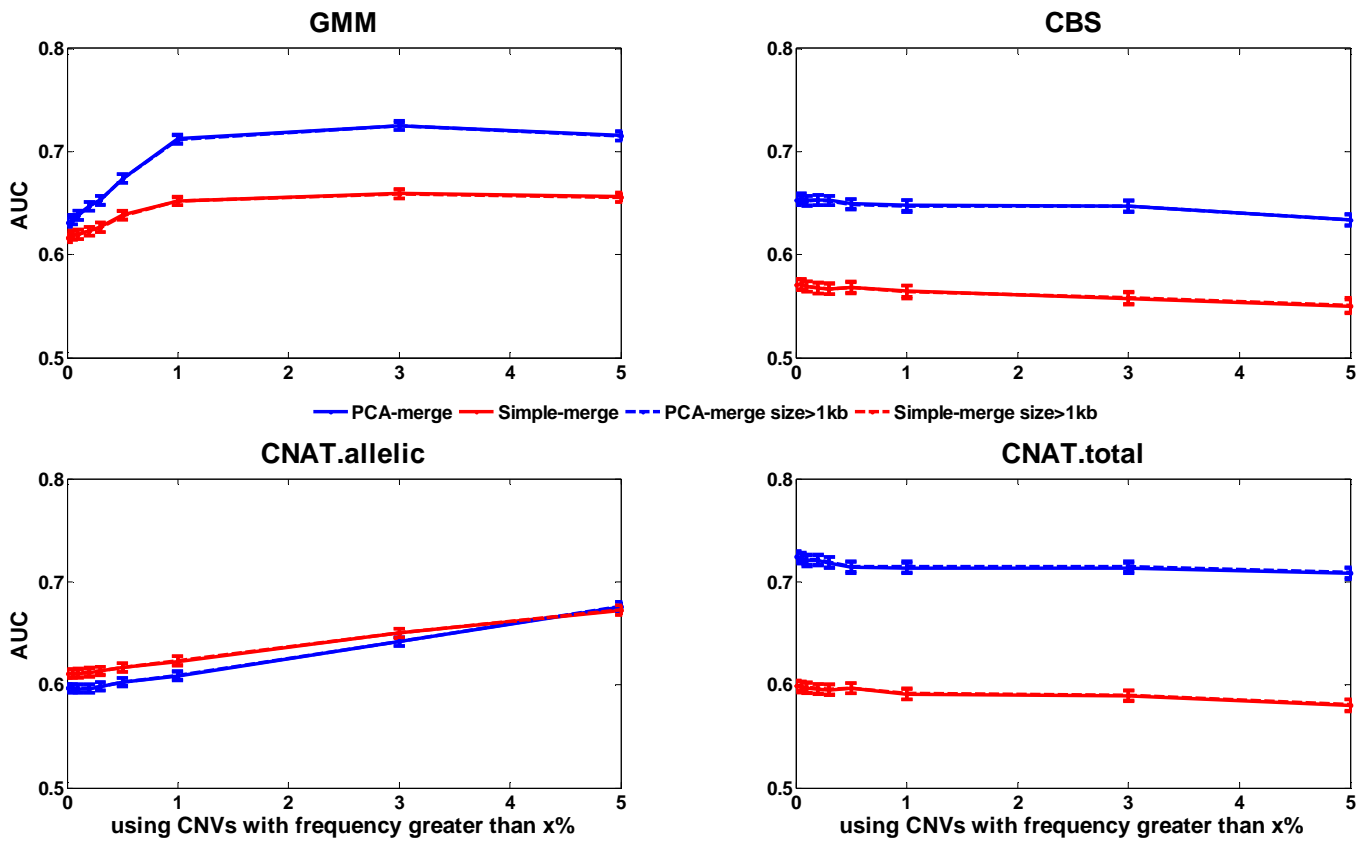
Supplementary table 3: T statistic values for overlap between CNVs identified from Affymetrix and Illumina data

The T statistic is computed from the difference between observed overlap and expected counts normalized by the standard deviation of expected counts. Expected counts are inferred from the overlap between reshuffled data (n=1000) and CNVs identified on Illumina. T statistics greater than 2.58 are significant with $\alpha=1\%$. Positive (negative) T statistics indicates enrichment (depletion) with respect to the expected counts.



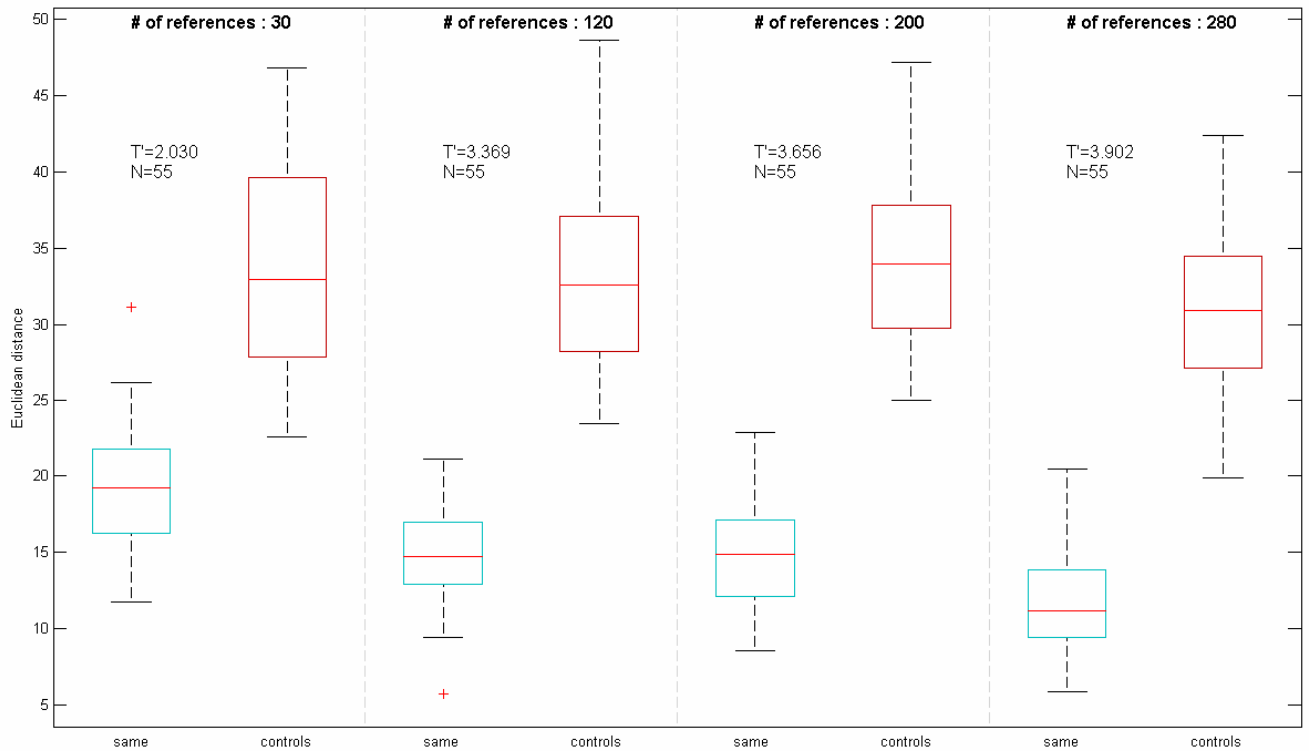
Supplementary figure 2: Overlap between CNPs/CNVRs and published CNVs/Illumina CNVs

These figures are similar to figures two and three. Each plot title indicates the overlap analysis. Overlap is measured by the Jacquard coefficient, i.e. the ratio between the intersect and the union of two CNVs. Expected counts from reshuffled data (n=1000) are shown in gray (extending over one standard deviation). Estimated p-values are indicated for significant enrichment (red) or depletion (blue), with respect to these controls. Non significant p-values ($\alpha > 1\%$) are shown in black.



Supplementary figure 3: Performance for predicting relatedness based on CNV profiles generated by different methods

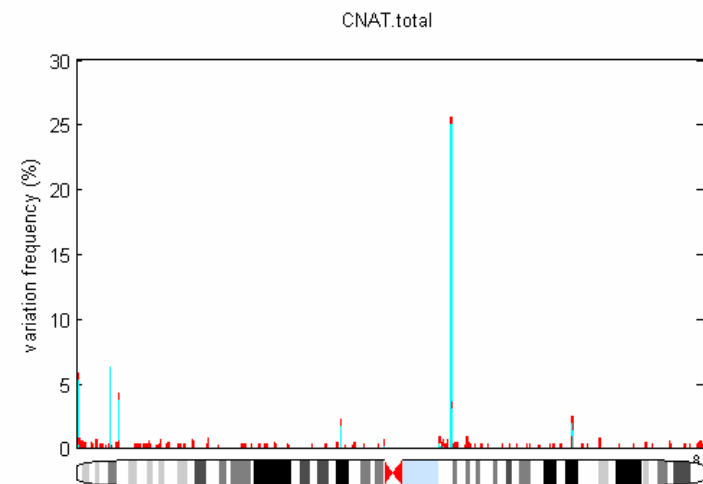
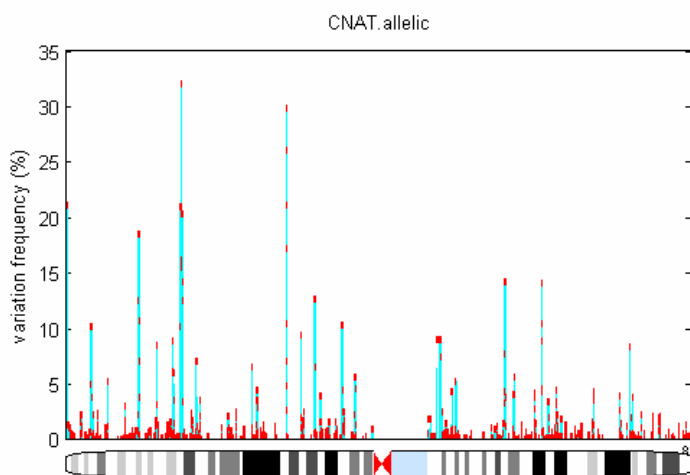
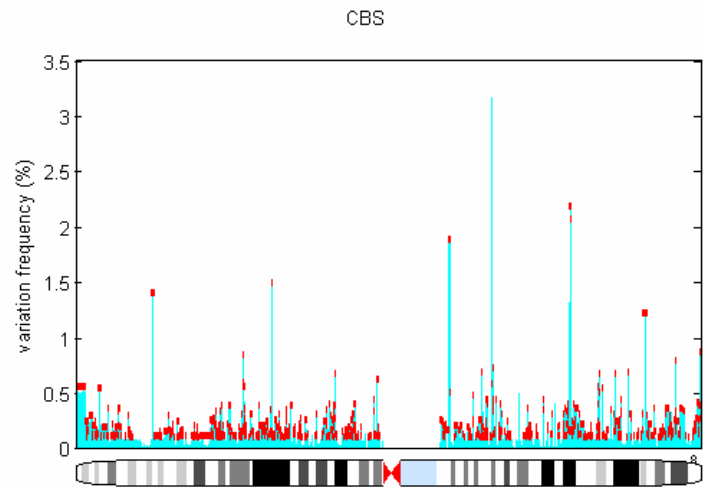
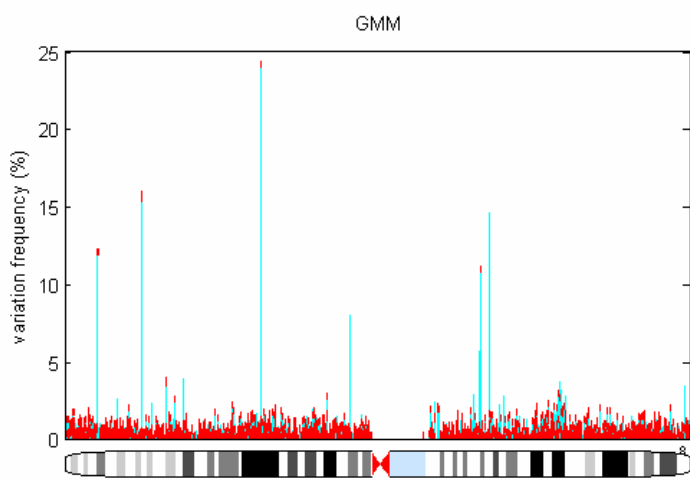
Each plot shows the Area Under the Curve (AUC) (Y axis) for predicting relatedness between individuals as a function of CNV frequency (X axis). CNV detection algorithms are indicated on top and merging procedure by colors. Predictions made with all CNV regions irrespective of their length are shown as straight lines and predictions using only CNV regions with length greater than 1kb are represented with dashed line (both solid and dash lines overlap each other). Curves were made with the mean from n=100 permutations, +/- one standard deviation around the mean is shown by the thickness of the square points. The analysis employed 162 pairs of individuals known to be related and 2000 pairs of unrelated individuals.



Supplementary figure 4: Improvement of the normalization as a function of the reference panel size

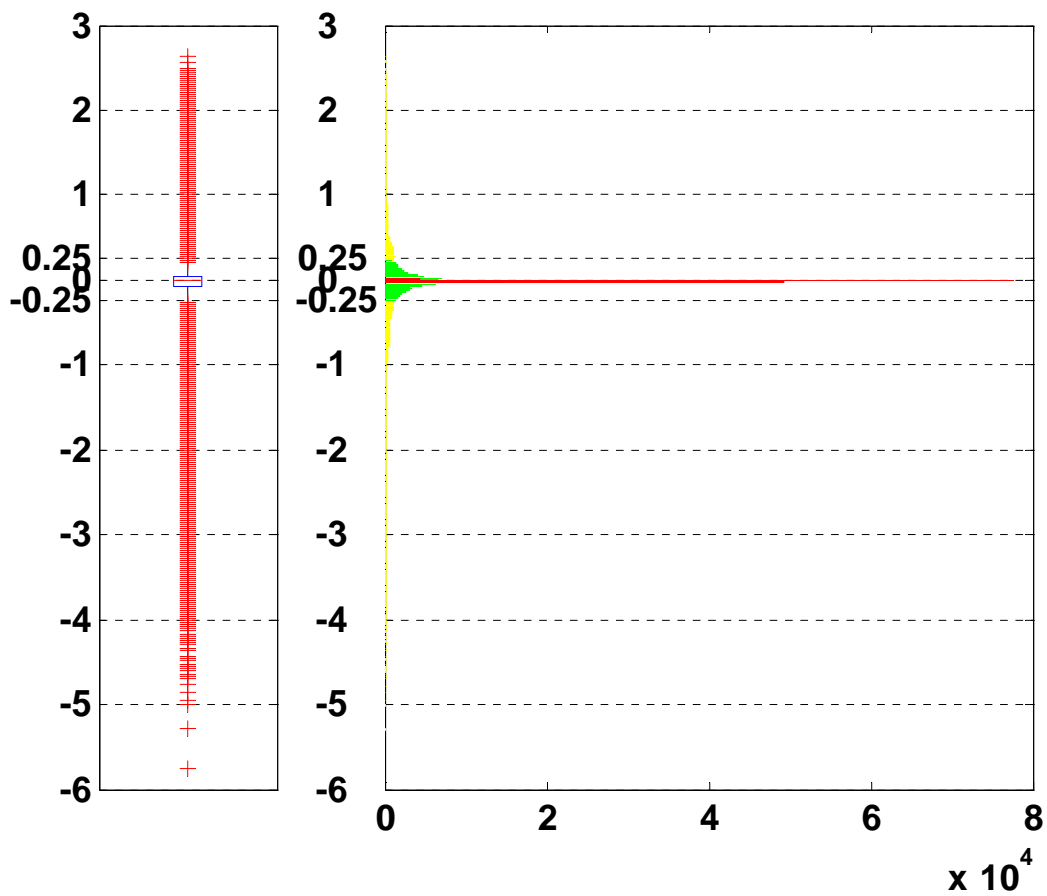
In cyan are shown the distances (n=55) between the CNV profiles as predicted by two independent reference panel (having the same size) for a same individual. In red are the distances between unrelated individuals predicted by these two reference panels. Different size of reference panel have been tested (30,120,200 and 280). The T' score is an estimate of the separation between pairs of identical individuals (same) and controls and is computed as:

$$T' = \frac{abs(geomean(same) - geomean(controls))}{\sqrt{std(same)^2 + std(controls)^2}}$$



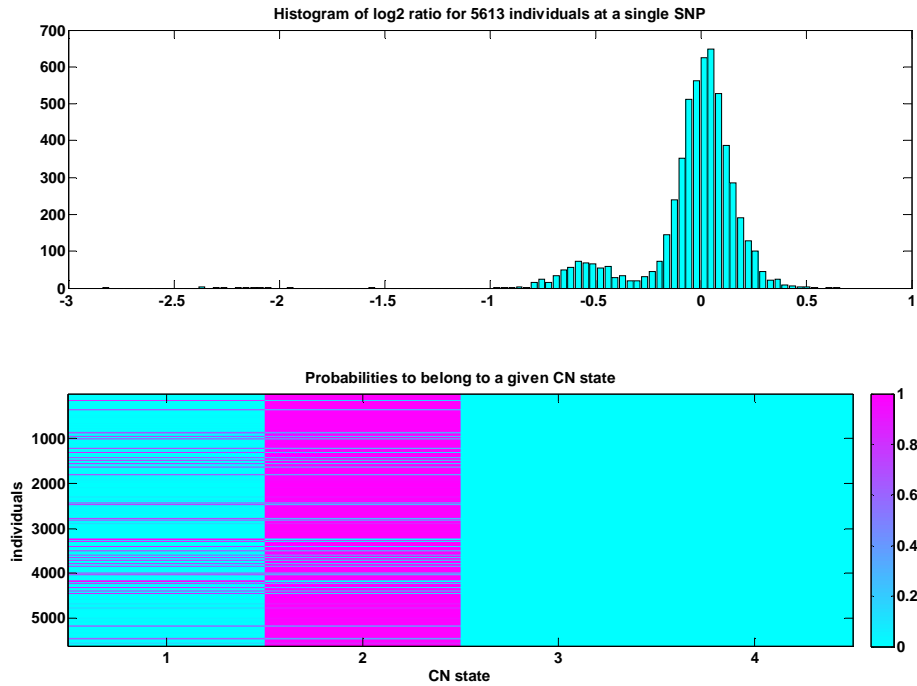
Supplementary figure 5: CNV frequency on chromosome 1

Each plot shows the CNV prediction from a CNV detection algorithm (title indicates the method name). Each cyan bars is a SNP on chromosome 1, the bar height represent the % of CoLaus individuals found as having this SNP as variant (deleted or duplicated). In red are regions defined by the PCA merge, the height shows the maximal CNV frequency found in this region. Only regions with CNV frequency greater than 0.1% are plotted.



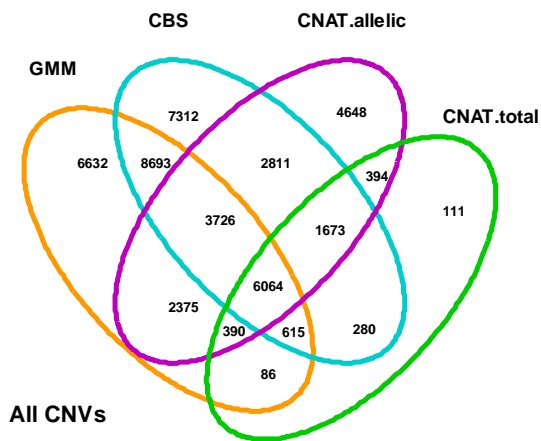
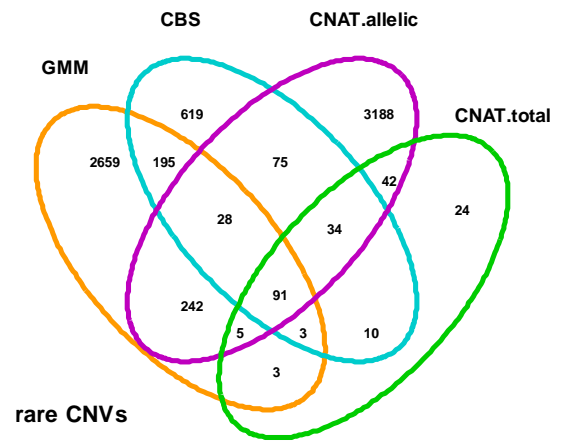
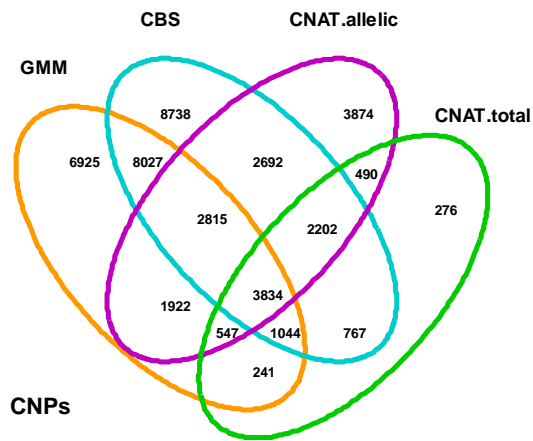
Supplementary figure 6: Distribution of CBS segment mean ratios

Distribution of all CBS segments mean \log_2 ratios, detected in the 5612 CoLaus individuals. Left panel shows a boxplot representation. Median is -0.007 (and mean -0.0637), inter-quartile 0.114 and 25th and 75th percentiles are respectively -0.082 and 0.032. Right panel shows an histogram representation, data have been clustered using a 3 components Gaussian Mixture Model (color indicates data points from a same cluster).



Supplementary figure 7: CNV calling using Gaussian Mixture Model

Top panel, distribution of log2 ratios for all 5613 CoLaus individuals at a single SNP. Bottom panel, matrix of probabilities for CoLaus individuals to have the queried SNP in a deletion state (CN=1), copy neutral (CN=2), with one or additional copy (CN=3 and CN=4). The CN dosage value can be computed as $\sum_{i=1}^4 P_i * C_i$ where P_i represents the probability at an individual j and a SNP k to be in the copy number state C_i .



Supplementary figure 8: Intersection between CNVs methods

Intersection between the different CNV detection methods using the CNV status at 60k autosomal and independent SNP (i.e. SNPs that are not in LD in the CEU population). The percentage of SNPs found variant in several methods is indicated in Supplementary Table 4.

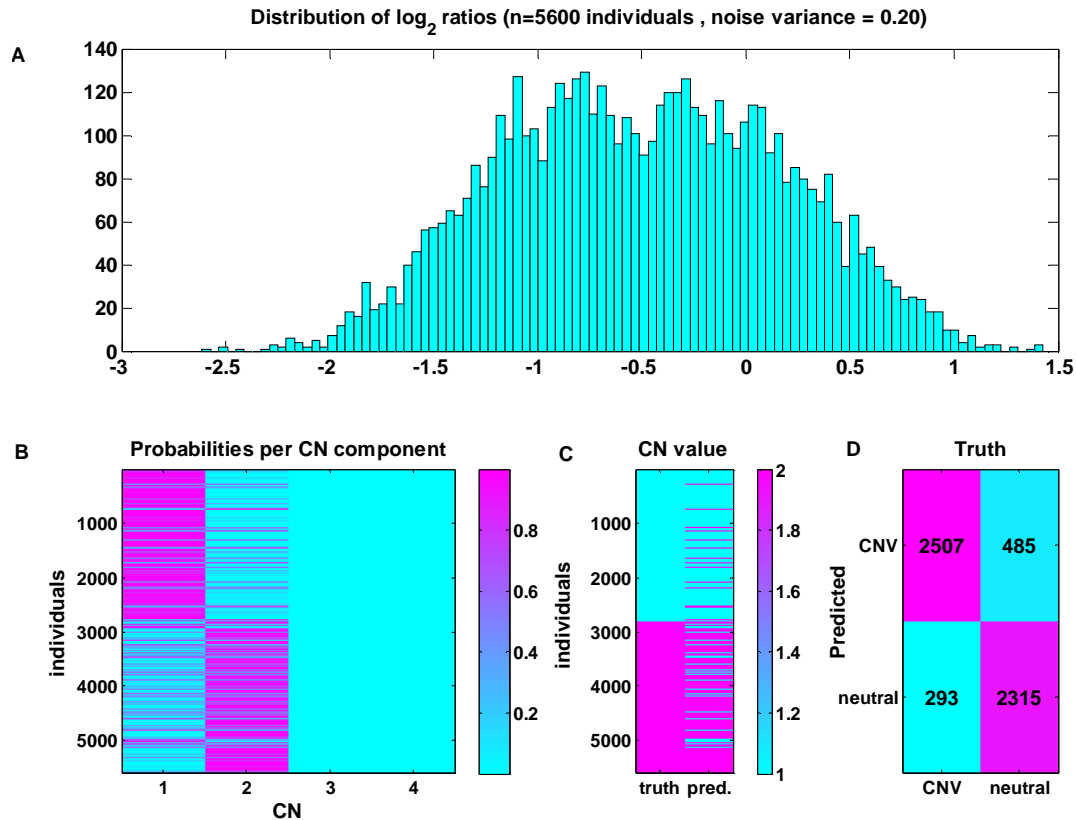
	CNPs	CNVRs	All CNVs
% of SNPs found variant in at least 3 methods	2.3	23.5	27.21
% of SNPs found variant in at least 2 methods	10	55.3	59.17

Supplementary table 4: Percentage of SNPs found copy number variant by several methods

CNPS	GMM	CBS	CNAT.allelic	CNAT.total
GMM	3226 (100.00%)	317 (8.00%)	366 (5.58%)	102 (3.06%)
CBS		1055 (100.00%)	228 (5.03%)	138 (12.22%)
CNAT.allelic			3705 (100.00%)	172 (4.59%)
CNAT.total				212 (100.00%)
CNVRs	GMM	CBS	CNAT.allelic	CNAT.total
GMM	25355 (100.00%)	15720 (39.54%)	9118 (26.34%)	5666 (19.48%)
CBS		30119 (100.00%)	11543 (31.24%)	7847 (24.78%)
CNAT.allelic			18376 (100.00%)	7073 (34.16%)
CNAT.total				9401 (100.00%)
ALL CNVs	GMM	CBS	CNAT.allelic	CNAT.total
GMM	28581 (100.00%)	19098 (46.97%)	12555 (32.95%)	7155 (23.05%)
CBS		31174 (100.00%)	14274 (36.62%)	8632 (26.84%)
CNAT.allelic			22081 (100.00%)	8521 (36.77%)
CNAT.total				9613 (100.00%)

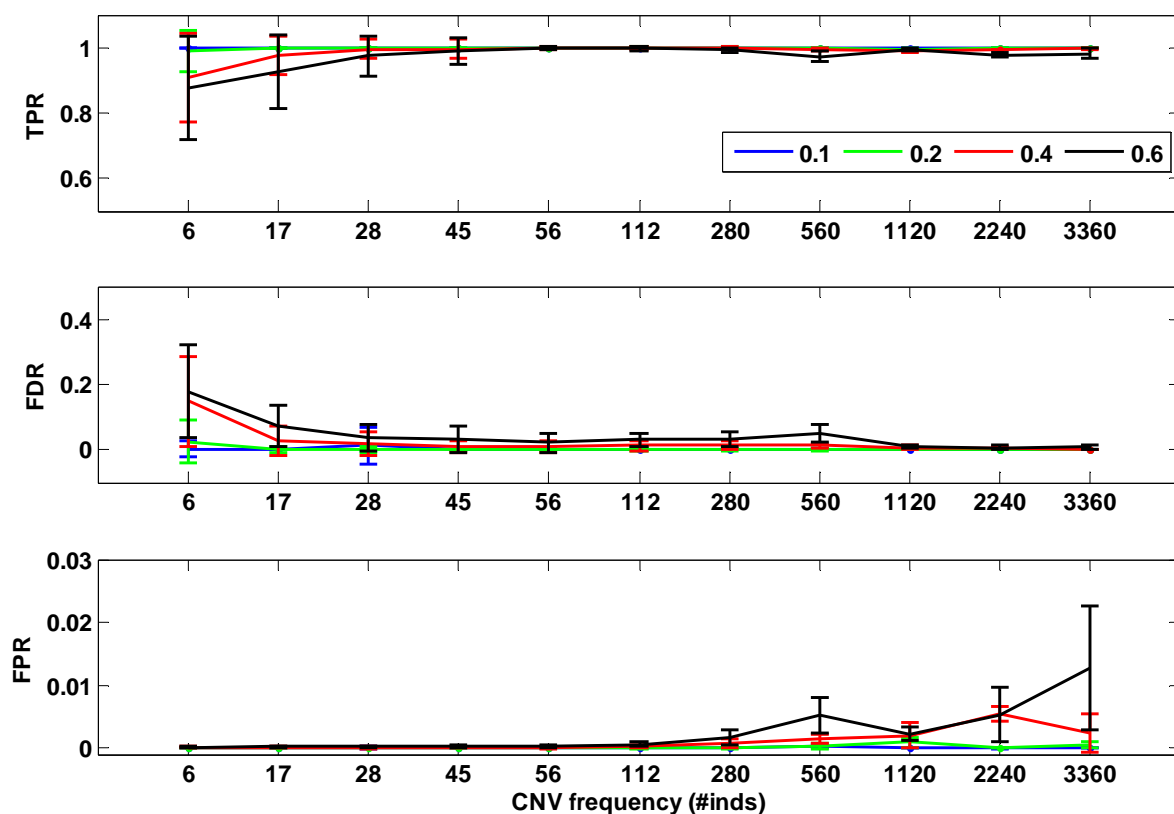
Supplementary table 5: Pairwise comparison between CNV detection methods

Pairwise comparison between the different detection methods using the CNV status at 60k autosomal and independent SNPs (i.e. SNPs that are not in LD in the CEU population). Intersection is reported as the number of SNPs found variant in both methods, the percentage, indicated between the parentheses, corresponds to the intersection divided by the union of SNPs found variant in any of the two methods. The number of variant SNPs in each method (independently of any other methods) is indicated as self-self comparison.



Supplementary Figure 9: Testing the Gaussian Mixture Model on simulated data

A) Distribution of simulated \log_2 ratios (n=5600) with Gaussian noise ($\sigma=0.2$). Half of the individuals have an underlying \log_2 ratio = -1 (deletion) and the remaining \log_2 ratio = 0 (copy neutral) B) Probabilities, as computed by our Gaussian Mixture Model, for each individual (Y axis) to belong to a given copy number state (X axis) C) Comparison between the true and predicted copy number state of each individual D) Contingency table containing the number of individuals correctly or incorrectly predicted



Supplementary Figure 10: Performance of the Gaussian Mixture Model

Panels show respectively from top to bottom, the True Positive, False Discovery and False Positive rates, when predicting deletions (with predefined frequency, see X axis) in a population of $n=5600$ individuals. The different curves correspond to simulated data with a given Gaussian noise (see graph legend). Each point is the mean from 50 resamplings, error bars correspond to one standard deviation.

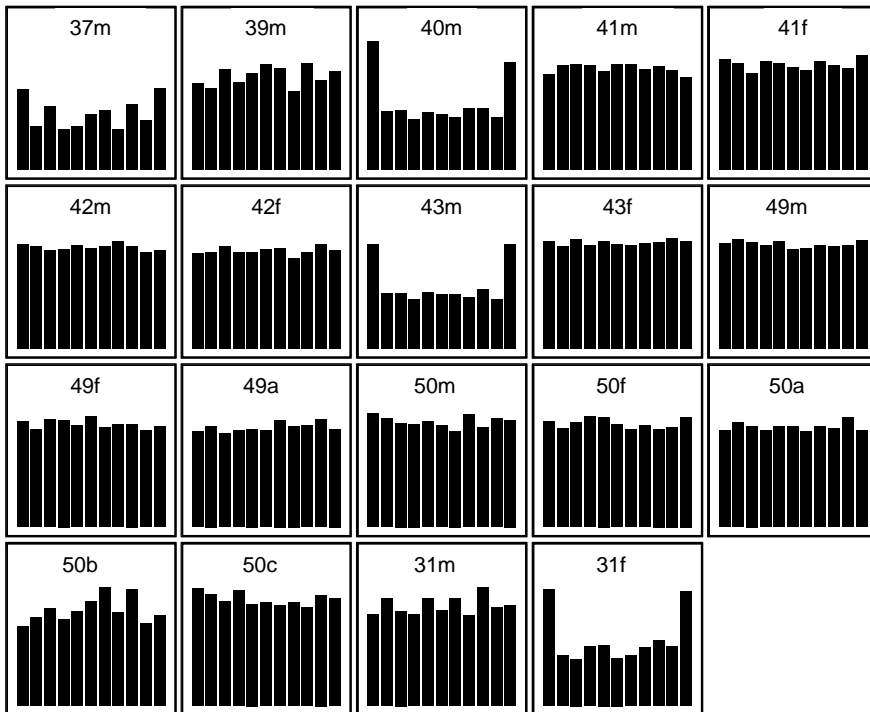
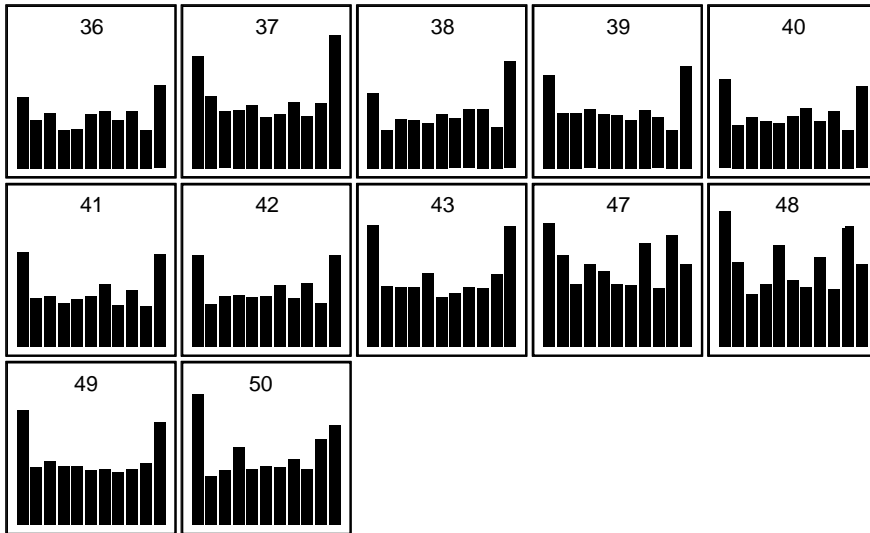
References

1. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* 81: 559-575.
2. Lai WR, Johnson MD, Kucherlapati R, Park PJ (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* 21: 3763-3770.
3. Willenbrock H, Fridlyand J (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* 21: 4084-4091.

Annexe II: Obesity Supplementary Information

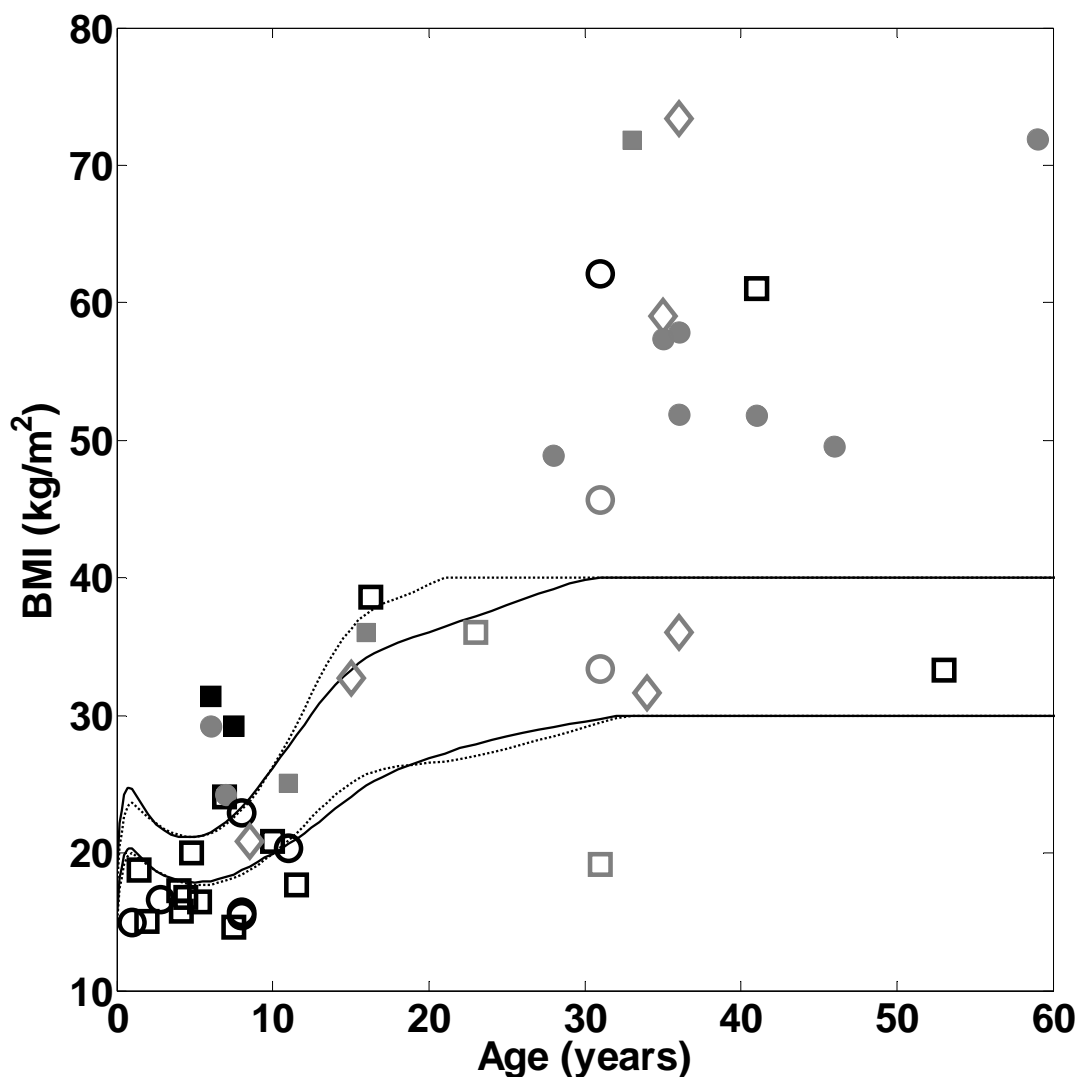
Supplementary Figure S1

Validation of 16p11.2 deletions by MLPA and determination of their modes of inheritance. MLPA was carried out using 9 probe pairs within and 2 lying outside the deletion (one to each side), as shown in Figure 1, together with 9 control (nominally copy number invariant) probe pairs. Panels show the relative magnitude of the normalised, integrated signal at each probe location in order of chromosomal location. Where DNA was available, samples were analysed if they were identified from GWAS data as carrying a deletion at 16p11.2 (top) or if they were a first degree relative of a proband (bottom). Labels correspond to the case ID of the proband as shown in Table S2; f = father; m = mother; a-c = siblings.



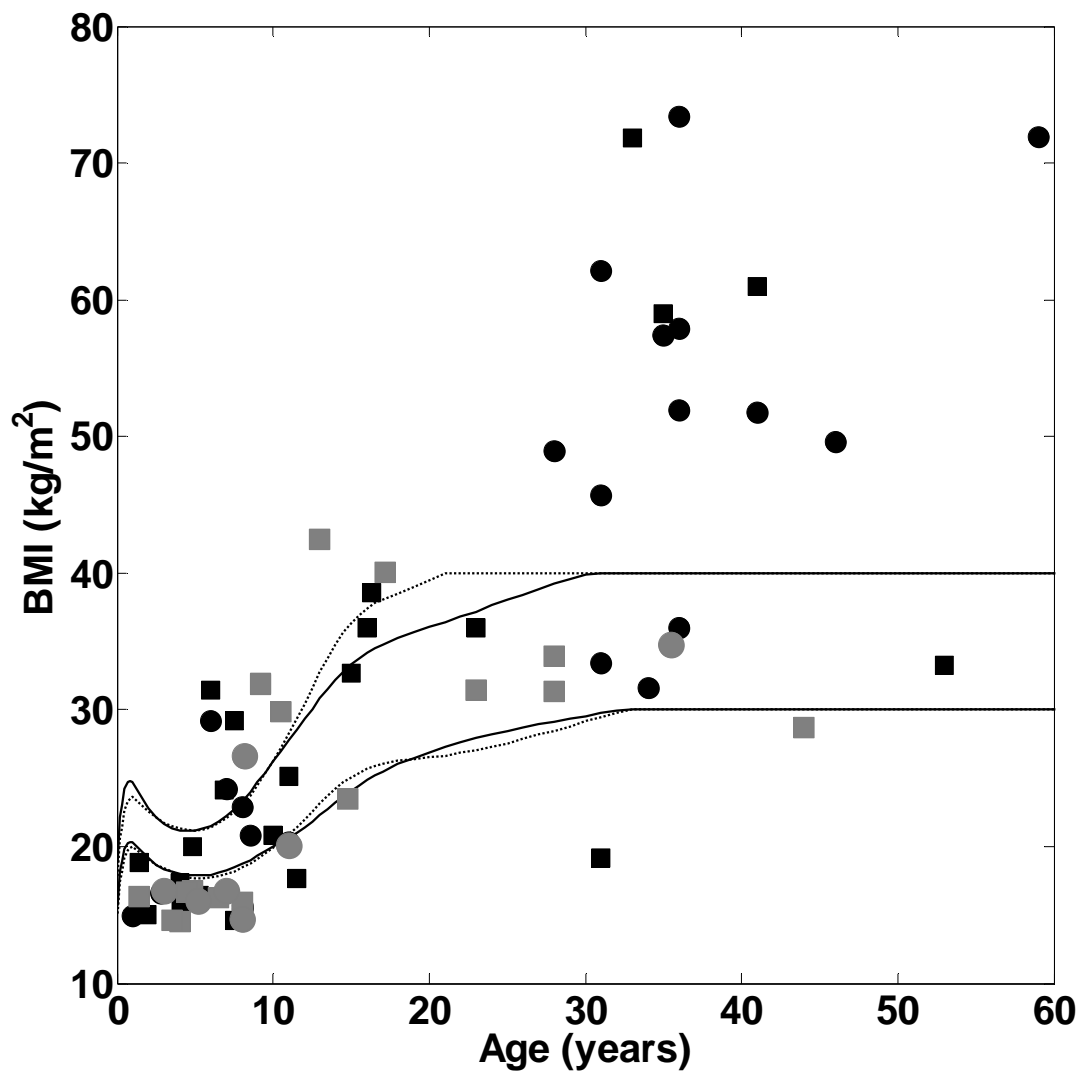
Supplementary Figure S2

Dependence of BMI on age in subjects having a deletion at 16p11.2. Data are shown for all individuals identified in this study as having a deletion at 16p11.2, for whom phenotypic information is available. Lines denote the age- and gender-corrected thresholds (solid/broken – male/female) for obesity (adults – BMI ≥ 30 kg.m⁻², children $\geq 97^{\text{th}}$ percentile) and morbid obesity (adults – BMI ≥ 40 kg.m⁻², children Z-BMI ≥ 4). Symbols are as follows: Square/circle – male/female; black/grey – ascertained/not ascertained for developmental delay; filled/open symbols – ascertained/not ascertained for obesity; grey diamonds – first-degree relative of a proband. Individuals from general population are shown as open grey circles or squares.



Supplementary Figure S3

Dependence of BMI on age in subjects having a deletion at 16p11.2. Data are shown for all individuals identified in this study as having a deletion at 16p11.2, for whom phenotypic information is available. Lines denote the age- and gender-corrected thresholds (solid/broken – male/female) for obesity (adults – BMI ≥ 30 kg.m⁻², children $\geq 97^{\text{th}}$ percentile) and morbid obesity (adults – BMI ≥ 40 kg.m⁻², children Z-BMI ≥ 4). Symbols are as follows: Square/circle – male/female; black – all probands and relatives identified in this study (see Figure S1); grey – subjects from other studies.



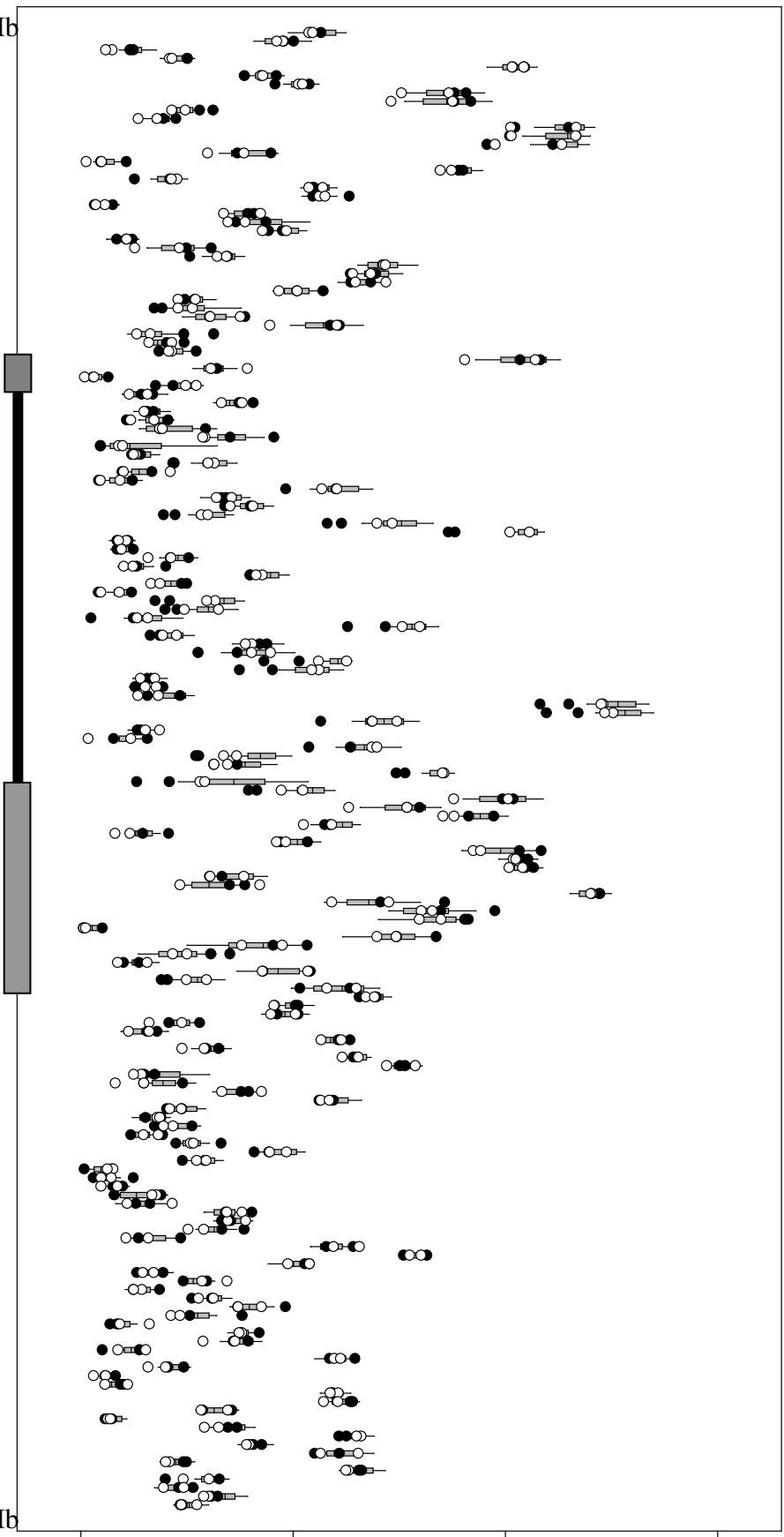
Supplementary Figure S4

Transcript levels for genes within and nearby 16p11.2 deletions. Expression data for adipose tissue from the SOS Sib Pair cohort were analysed for probes detecting transcripts for genes lying within the interval chr16:28.4–31.0Mb (see Supplementary Table S4 for details). Transcript levels in the two individuals carrying a deletion of 16p11.2 (black symbols) are plotted alongside those for their non-obese siblings (white). Also shown are box plots summarising the data for the other 157 obese subjects from this study, indicating the 10th, 25th, 50th, 75th and 90th percentiles for each transcript. The positions of the the 16p11.2 deletion and the flanking segmental duplications relative to the transcripts are indicated by a solid line and grey bars at the left axis.

Within the deleted region, there is a consistent reduction in expression in the subjects carrying a deletion, relative to both their siblings and to other obese subjects. In contrast, although CNVs have been shown to have the potential to affect expression of neighbouring genes up to 0.5Mb distant^{31,32}, no such clear and consistent differences in transcript levels are observed for the genes lying nearby, outside the region of the 16p11.2 deletion.

28.4Mb

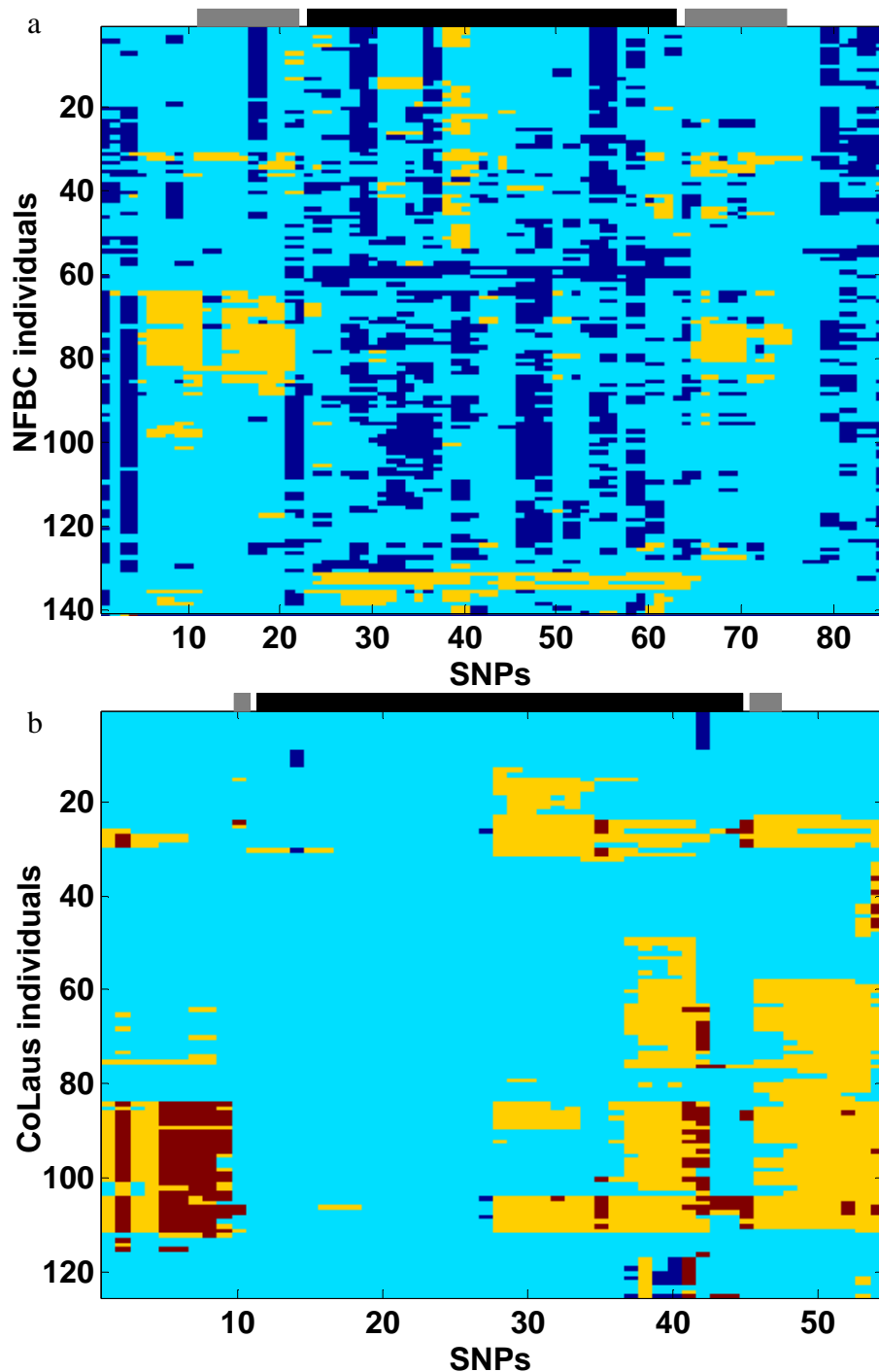
31.0Mb



Relative transcript abundance

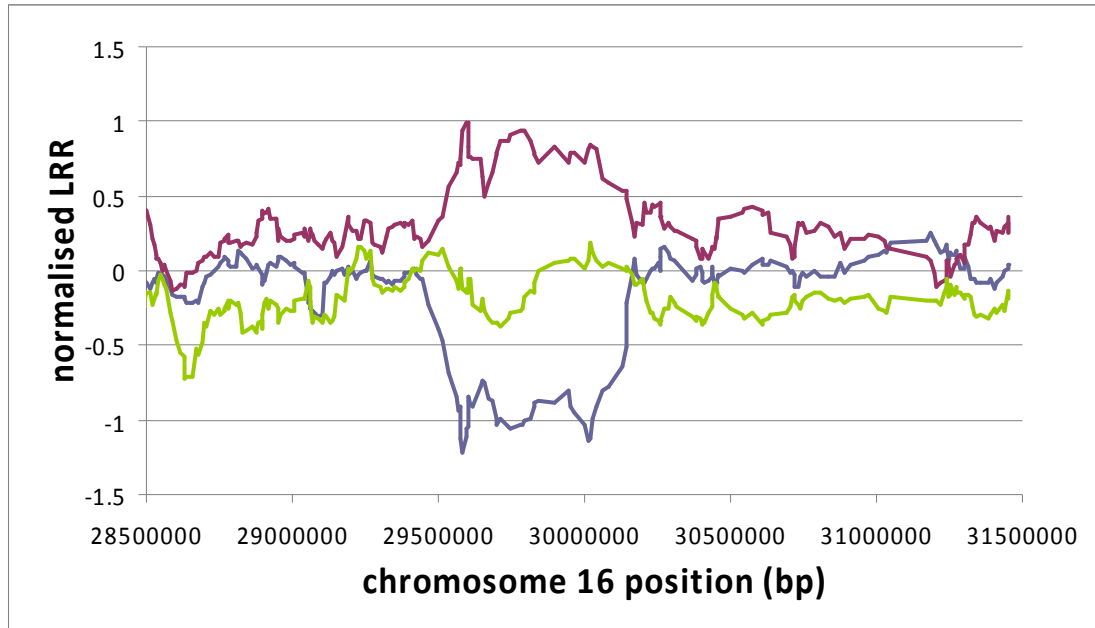
Supplementary Figure S5

Graphical representation of the output of CNV discovery algorithms. Copy number calls at SNPs within and surrounding the deleted region (black bar) and its flanking segmental duplications (grey bars) are shown as follows: blue – 1 copy; cyan – 2 copies (i.e. no aberration); yellow – 3 copies; red – 4 copies. (a) *cnvHap* output for the NFBC cohort, showing all individuals with at least 10 aberrant probes within the deletion; (b) Gaussian Mixture Model output for the CoLaus cohort, showing all individuals with at least 1 aberrant probeset; this method has been validated by comparing test datasets with results from the *CNAT*³³ and *CBS*^{34,35} algorithms. The different patterns for the two methods reflect the locations interrogated by the respective platforms, and also the respective sensitivities of the platforms and algorithms to copy number variation.



Supplementary Figure S6

Validation of deletion calls from Illumina genotyping data. LogR ratio (LRR) data exported from Illumina BeadStudio was normalised with respect to the median and variance for each probe, and smoothed by averaging over a 9-point moving window. Example data are shown for samples from the NFBC cohort that had normal copy number (green) and which carried a deletion (blue) or duplication (purple).



Supplementary Table S1

Cognitive/behavioral symptoms observed in carriers of 16p11.2 deletions ascertained for developmental delay, and for affected relatives. 'No data' indicates that this phenotype was not assessed. NA – not applicable due to age of the patient.

case ID	Age	Mental retardation	Language	Hyperphagia	ASD	Other Behavioral symptoms
1	4.1	Borderline-mild	Language delay	yes	no	no data
2	16.3	Executive function deficits	Language deficit	yes	no, social cognition deficit	Shyness, obsessive compulsive disorder
3	5.3	No	Echolalia		yes	Stereotypes, hyperactivity
4	31	no	no data	yes	no data	no data
5	8	No	Dysphasia	fluctuating	no	Hyperactivity
6	41	Mild	Language delay	severe	no data	no data
7	10	Mild-moderate	Language delay	mild	no data	no data
8	11	Mild, IQ 72-49	mild language delay	yes	no	anxiety
9	11.5	Mild	Language delay	no	yes	Oppositional, aggressivity, stereotypes
10	1.4	NA	NA	no	no	no
11	2.8	Mild to moderate, global delay	Language delay	yes	no data	Hyperactivity
12	53	Mild to moderate	no data	no	no	no data
13	1	Developmental age: 5 months	NA	no	NA	NA
14	4.4	Borderline, IQ 73	Language delay	no	no	Hyperactivity
15	4.8	Mild	Language delay	no	yes	no data
16	6.9	Mild	Language delay	yes	no	Oppositional, aggressivity
17	7.5	No, IQ 77-89	Language delay	no	no	Anxiety, hyperactivity
18	1.9	moderate	Language delay	no	NA	none / NA
19	4.4	Moderate, IQ 57	Severe language delay	no	yes	Repetitive, restricted behavior

20	8	Borderline, VIQ: 78, PIQ: 96	Language delay	no	no	Attention deficit
21	4	Moderate	Severe language delay	no	no	Hyperactivity, temper tantrums, oppositional
22	8	no data	no data	no	no	Attention deficit, mutism
63	15	Mild, VIQ: 67, PIQ: 61		yes	no	Hyperactivity, aggressivity, anxiety
64	36	Borderline		yes	no	no data

Supplementary Table S2

Obesity characteristics of carriers of 16p11.2 deletions. The basis for ascertainment and all available data for gender, age and BMI are shown for each subject identified as carrying a deletion at 16p11.2. Also shown are the methods used to identify the deletion and for its validation, and the inheritance of the deletion as inferred from the results of analysis of parental DNAs where these were available.

Ascertainment	case ID	gender	location	Age (years)	BMI	inheritance	CNV detection platform	
							detection	validation
Developmental Delay	1	M	Estonia	4.1	15.8	de novo	Illumina Human CNV370-Duo	qPCR
	2	M	Lausanne	16.3	38.6	inherited (mother)	aCGH Agilent 244k	none
	3	M	Lille	5.3	16.4	inherited (mother)	aCGH Agilent 44K	qPCR
	4	F	Lille	31	62.1	probably inherited ^a	aCGH Agilent 44K	qPCR
	5	F	Lille	8	22.9	de novo	aCGH Agilent 44K	qPCR
	6	M	Lille	41	61		aCGH Agilent 44K	qPCR
	7	M	Lille	10	20.8	inherited (father)	aCGH Agilent 44K	qPCR
	8	F	Lille	11	20.3	inherited (mother)	aCGH Agilent 44K	qPCR
	9	M	Lille	11.5	17.7		aCGH Agilent 44K	qPCR
	10	M	Lille	1.4	18.8	inherited (father)	aCGH Agilent 44K	qPCR
	11	F	Lyon	2.8	16.6	inherited (father)	aCGH Agilent 105K	qPCR
	12	M	Lyon	53	33.3		aCGH Agilent 105K	qPCR
	13	F	Nancy	1	14.9	de novo	aCGH Agilent 105K	qPCR
	14	M	Nancy	4.4	16.8	de novo	aCGH Agilent 105K	qPCR
	15	M	Nancy	4.8	20.0	inherited (mother)	aCGH Agilent 105K	qPCR
	16	M	Nantes	6.9	24.1	inherited (mother)	aCGH Agilent 44K	FISH
	17	M	Nantes	7.5	14.6	inherited (mother)	aCGH Agilent 44K	FISH
	18	M	Nantes	1.9	15.0	inherited (mother)	aCGH Agilent 44K	FISH
	19	M	Paris	4.4	16.0	de novo	FISH	none

	20	F	Rouen	8	15.4		QMPSF	FISH
	21	M	Rouen	4	17.3	de novo	QMPSF	FISH
	22	F	Rouen	8	15.7	inherited (mother)	QMPSF	FISH
Obesity & Developmental Delay	23	M	Lille	6	31.4		qPCR	aCGH Agilent 44K
	24	M	Lille	10.3	34.8		qPCR	aCGH Agilent 44K
	25	F	Lille	12	31.9		qPCR	aCGH Agilent 44K
	26	M	Lille	14.5	40.2		qPCR	aCGH Agilent 44K
	27	F	Lille	13.3	34.2		qPCR	aCGH Agilent 44K
	28	M	Lille	13			qPCR	aCGH Agilent 44K
	29	M	Lille	6	25.0		qPCR	aCGH Agilent 44K
	30	F	Nîmes	12.3	29.0		qPCR	aCGH Agilent 44K
	31	M	London	7.5	29.2	inherited (father)	aCGH Agilent 185K	none
General Population	32	M	Estonia	23	36		Illumina Human CNV370-Duo	qPCR
	33	F	Finland	31	33.4		Illumina Human CNV370-Duo	multiple algorithms
	34	F	Finland	31	45.7		Illumina Human CNV370-Duo	multiple algorithms
	35	M	Finland	31	19.2		Illumina Human CNV370-Duo	multiple algorithms
Adult Obesity	36	F	Lille	28	48.9		Illumina Human CNV370-Duo	MLPA
	37	M	Lille	33	71.8	inherited (mother)	Illumina Human CNV370-Duo	MLPA
	38	F	Lille	41	51.8		Illumina Human CNV370-Duo	MLPA
	39	F	Lille	36	57.9		Illumina Human CNV370-Duo	MLPA
Childhood Obesity	40	M	Lille	16	36.0	inherited (mother)	Illumina Human CNV370-Duo	MLPA
	41	F	Lille	7	24.2	de novo	Illumina Human CNV370-Duo	MLPA
	42	F	Lille	6	29.2	de novo	Illumina Human CNV370-Duo	MLPA
	43	M	Lille	11	25.1	inherited (mother)	Illumina Human CNV370-Duo	MLPA
	44	F	Cambridge	15	43.9		Affymetrix 6.0	MLPA

	45	M	Cambridge	13	57.0		Affymetrix 6.0	MLPA
	46	F	Cambridge	15	45.8		Affymetrix 6.0	MLPA
Obesity Bariatric Surgery	47	F	Lille	46	49.6		Illumina Human 1M-Duo	MLPA
	48	F	Lille	59	71.9		Illumina Human 1M-Duo	MLPA
Obesity Discordant Siblings	49	F	Gothenburg	36	51.9	de novo	Illumina Human 610K-Quad	MLPA
	50	F	Gothenburg	35	57.4	de novo	Illumina Human 610K-Quad	MLPA
Proband Relative	51	F	Lausanne	36	73.4	Mother of 2	aCGH Agilent 244K	none
	52	F	Lille			Mother of 3	MLPA	none
	53	M	Lille	35	59	Brother of 4 ^a	aCGH Agilent 44K	qPCR
	54	M	Lille			Father of 7	qPCR	none
	55	F	Lille	42	34.7	Mother of 8	aCGH Agilent 44K	qPCR
	56	F	Lille	8.5	20.8	Sister of 8	aCGH Agilent 44K	qPCR
	57	M	Lille			Father of 10	MLPA	none
	58	M	Lyon	37	31.1	Father of 11	qPCR, 3 primer pairs	none
	59	F	Nancy	28	30.1	Mother of 15	qPCR	none
	60	F	Nantes	32	32.8	Mother of 16	FISH	none
	61	F	Nantes	34	31.6	Mother of 17	FISH	none
	62	F	Nantes			Mother of 18	FISH	none
	63	M	Rouen	15	32.7	Brother of 22	QMPSF	FISH
	64	F	Rouen	36	36	Mother of 22	QMPSF	FISH
	65	M	London			Father of 31	aCGH Agilent 244K	MLPA
	66	F	Lille			Mother of 37	MLPA	none
67	F	Lille			Mother of 40	MLPA	none	
68	F	Lille			Mother of 43	MLPA	none	

^aThe proband's brother has the deletion, but both parents are deceased so inheritance cannot be confirmed. One instance has been reported⁴ of presumed germ-line mosaicism in which a deletion was found in two siblings but neither parent.

Supplementary Table S3

Obesity phenotype of carriers of 16p11.2 deletions from other publications, as included in Figure 2.

Publication	Patient ID	gender	Age (years)	BMI
<i>Bijlsma et al.</i> ¹¹	Case 1	M	44.0	28.7
	Case 2	M	17.2	40.1
	Case 3	F	8.2	26.6
	Case 6	F	7.0	16.8
	Case 8	F	11.0	20.1
	Case 10	F	8.0	14.7
	Case 11	M	4.0	14.6
	Case 13	M	4.8	16.7
<i>Fernandez et al.</i> ¹⁴	Proband 2	M	13.0	42.5
	Proband 3	M	4.5	16.7
	Patient 3b	M	3.5	14.6
	Patient 3c	F	35.5	34.7
<i>Ghebranious et al.</i> ¹³	Twin1	M	28.0	31.3
	Twin2	M	28.0	34.0
<i>McCarthy et al.</i> ¹²	CHOP1	F	3.0	16.8
	CHOP4	M	14.8	23.5
	03C18520	M	23.0	31.4
	AU041905	M	8.0	15.9
<i>Shimojima et al.</i> ¹⁵	-	M	3.2	16.5
<i>Weiss et al.</i> ¹⁰	Pt1	M	6.5	16.3
	Pt3	M	1.4	16.3
	Pt4	M	9.2	31.8
	Pt5	M	9.2	32.0
	Aut1	F	5.2	16.0
	Aut2	M	10.5	29.9

Supplementary Table S4

Details of probes analysed in the course of expression analysis (Supplementary Figure S4), listing the Affymetrix probe identification code, the gene whose transcript is listed as being detected by the probe, and the chromosomal coordinate (build hg18) for the start of that gene.

Probe ID	Gene	Coordinate	Probe ID	Gene	Coordinate
209275_s_at	CLN3	28396101	242414_at	QPRT	29582101
210859_x_at	CLN3	28396101	1559584_a_at	C16orf54	29661285
220023_at	AC138894.3	28413494	214142_at	ZG16	29697091
1552995_at	IL27	28418184	202183_s_at	KIF22	29709542
209230_s_at	NUPR1	28456107	216969_s_at	KIF22	29709542
221822_at	CCDC101	28472748	207824_s_at	MAZ	29725356
48117_at	CCDC101	28472748	212064_x_at	MAZ	29725356
207122_x_at	SULT1A2	28510765	228798_x_at	AC009133.1	29729246
211385_x_at	SULT1A1	28524404	218300_at	C16orf53	29734786
238995_at	SULT1A1	28524404	227192_at	C16orf53	29734786
217314_at	AC145285.2	28618998	202180_s_at	MVP	29739230
200647_x_at	EIF3S8	28630283	201253_s_at	CDIPT	29777179
210949_s_at	EIF3S8	28630283	240537_s_at	AC120114.2	29782656
215230_x_at	EIF3S8	28630283	218720_x_at	SEZ6L2	29789981
201806_s_at	ATXN2L	28741821	223458_at	SEZ6L2	29789981
207798_s_at	ATXN2L	28741821	233337_s_at	SEZ6L2	29789981
201113_at	TUFM	28761233	238406_x_at	SEZ6L2	29789981
238190_at	TUFM	28761233	1553997_a_at	ASPHD1	29819201
209322_s_at	SH2B1	28782579	214993_at	ASPHD1	29819201
40149_at	SH2B1	28782579	221889_at	KCTD13	29825158
205444_at	ATP2A1	28797305	45653_at	KCTD13	29825158
219057_at	RABEP2	28823244	238142_at	KCTD13	29825158
74694_s_at	RABEP2	28823244	224981_at	TMEM219	29880852
77508_r_at	RABEP2	28823244	204877_s_at	TAOK2	29892723
206398_s_at	CD19	28850761	204878_s_at	TAOK2	29892723
212808_at	NFATC2IP	28869814	204986_s_at	TAOK2	29892723
212809_at	NFATC2IP	28869814	204504_s_at	HIRIP3	29911812
217526_at	NFATC2IP	28869814	227286_at	INO80E	29914532
217527_s_at	NFATC2IP	28869814	205744_at	DOC2A	29924336
229235_at	NFATC2IP	28869814	1557162_at	C16orf92	29942156
223173_at	SPNS1	28893597	227781_x_at	FAM57B	29943249
209881_s_at	LAT	28903648	200966_x_at	ALDOA	29971945
211005_at	LAT	28903648	214687_x_at	ALDOA	29971945
216902_s_at	AC009093.1	28993664	208932_at	PPP4C	29994812
216908_x_at	AC009093.1	28993664	207684_at	TBX6	30004583
243124_at	AC009093.1	28993664	215122_at	TBX6	30004583
221184_at	AC009093.2	29167674	223179_at	YPEL3	30011136
241644_at	BOLA2	29365833	232077_s_at	YPEL3	30011136
215299_x_at	SULT1A4	29374628	219722_s_at	GDPD3	30023632
1558044_s_at	AC009086.2	29471943	212046_x_at	MAPK3	30032927
1558534_at	AC009086.2	29471943	209083_at	CORO1A	30102393
237464_at	AC009086.2	29471943	209836_x_at	BOLA2B	30111740
1568964_x_at	SPN	29581801	203615_x_at	SULT1A3	30113244

206056_x_at	SPN	29581801	209607_x_at	SULT1A3	30113244
206057_x_at	SPN	29581801	210580_x_at	SULT1A3	30113244
216981_x_at	SPN	29581801	218317_x_at	SULT1A3	30113244
204044_at	QPRT	29582101	222094_at	SULT1A3	30113244
233334_x_at	SULT1A3	30113244	235950_at	ZNF688	30488529
211996_s_at	AC106782.7	30141697	213525_at	AC002310.1	30491072
214035_x_at	AC106782.7	30141697	1554769_at	ZNF785	30497795
214870_x_at	AC106782.7	30141697	1554770_x_at	ZNF785	30497795
215920_s_at	AC106782.7	30141697	242272_at	ZNF785	30497795
215921_at	AC106782.7	30141697	227294_at	ZNF689	30521380
221501_x_at	AC106782.7	30141697	227445_at	ZNF689	30521380
238449_at	AC106782.7	30141697	1559397_s_at	PRR14	30569724
1557987_at	AC106782.7	30141697	218714_at	PRR14	30569724
215123_at	AC106782.7	30141697	45687_at	PRR14	30569724
215002_at	AC106782.7	30141697	218255_s_at	FBRS	30577790
235060_at	AC106782.7	30141697	242217_s_at	FBRS	30577790
235167_at	AC106782.7	30141697	238771_at	FBRS	30577790
238341_at	AC106782.7	30141697	1552630_a_at	SRCAP	30617031
242114_at	AC106782.7	30141697	1569138_a_at	SRCAP	30617031
231989_s_at	AC106782.4	30186315	212275_s_at	SRCAP	30617031
244766_at	AC106782.4	30186315	213667_at	SRCAP	30617031
210396_s_at	AC106782.8	30204010	215053_at	SRCAP	30617031
202257_s_at	CD2BP2	30269588	38766_at	SRCAP	30617031
202256_at	CD2BP2	30269588	203709_at	PHKG2	30667092
220947_s_at	TBC1D10B	30275923	231300_at	C16orf93	30676254
205163_at	MYLPF	30293613	206845_s_at	RNF40	30681100
227552_at	37135	30296955	239801_at	RNF40	30681100
227470_at	ZNF48	30313934	1556368_at	RNF40	30681100
219781_s_at	ZNF771	30326236	1556369_a_at	RNF40	30681100
218069_at	DCTPP1	30342520	213196_at	ZNF629	30697271
200961_at	SEPHS2	30362453	219072_at	BCL7C	30752874
1554240_a_at	ITGAL	30391484	206813_at	CTF1	30811875
213475_s_at	ITGAL	30391484	1553586_at	NCRNA00095	30841418
218916_at	ZNF768	30442826	228277_at	FBXL19	30841893
206180_x_at	ZNF747	30449189	221864_at	ORAI3	30867888
228856_at	ZNF747	30449189	213202_at	SETD1A	30876116
238606_at	ZNF747	30449189	222817_at	HSD3B7	30904020
239774_at	ZNF747	30449189	230691_at	STX1B	30908078
57516_at	ZNF764	30472586	203530_s_at	STX4	30951820
222120_at	ZNF764	30472586	229395_at	STX4	30951820
213527_s_at	ZNF688	30488529	219047_s_at	ZNF668	30979672
213529_at	ZNF688	30488529	204876_at	ZNF646	30993265
235951_s_at	ZNF688	30488529	214226_at	AC135050.2	31002259

Supplementary Table S5

Details of genes lying within the deleted region at 16p11.2. Gene name, coordinates and strand of protein coding region are according to genome build hg18. Protein function descriptions are based on GeneCards entries (<http://www.genecards.org/>) or from the indicated references. Change in expression is given as the mean transcript level (all probes) in the 2 deletion carriers relative to obese (normal/lean) subjects (data as in Supplementary Figure S4). Possible functional relevance to obesity (bold type) or developmental delay/ cognitive deficit (italics) is as indicated. The first three pairs of genes lie within the segmental duplications.

Gene name	CDS start	CDS end	Strand	Change in Expr ⁿ	Protein function	Refs
BOLA2 BOLA2B	2936583 3 3011179 6	2937378 6 3011261 5	- -	0.6 (0.8)	Possibly involved in cell proliferation or cell-cycle regulation	
GIYD1 GIYD2	2937337 6 3011290 6	2937704 1 3011628 8	+ +	-	GIY-YIG domain containing	
SULT1A 4 SULT1A 3	2937390 2 3011955 0	2938380 1 3012274 2	+ +	1.0 (1.5)	Induced in response to fasting or as a result of a defect in leptin signalling <i>Catalyzes the sulfate conjugation of phenolic monoamine neurotransmitters</i>	36
SPN	2958255 0	2958375 3	+	1.1 (1.1)	Sialophorin, CD43. Activator of JNK1 and MAPK3 signalling	37-39
QPRT	2959801 9	2961623 3	+	1.2 (1.3)	<i>Catabolism of quinolinate, a neural excitotoxin and NMDA receptor agonist</i>	40
C16orf54	2966309 8	2966377 3	-	0.5 (0.8)		
MAZ	2972552 3	2972856 4	+	0.7 (0.8)	<i>Interacts with SP1 in regulating transcription of serotonin receptor gene HTR1A</i>	41
PRRT2	2973187 6	2973346 0	+		Proline-rich transmembrane protein	
C16orf53	2973534 7	2973857 6	+	0.8 (0.8)		
MVP	2974937 1	2976681 1	+	0.5 (0.8)	Regulates cytoplasmic localisation of PTEN	42
CDIPT	2977801 0	2978167 9	-	0.4 (0.5)	Phosphatidylinositol synthesis	
SEZ6L2	2979052 0	2981784 1	-	0.9 (0.9)	<i>Seizure-related. May contribute to specialized ER function in neurons</i>	
ASPHD1	2981979 3	2982471 9	+	1.0 (1.1)	Aspartate beta-hydroxylase domain containing	
KCTD13	2982569 3	2984485 5	-	0.6 (0.6)	Similar to TNFAIP1, a mediator of insulin resistance in rodent obesity models	
TMEM21 9	2988196 5	2989036 7	+	0.6 (0.7)	Transmembrane protein	

TAOK2	2989659 4	2990680 2	+	0.8 (0.8)	Activates JNK1 and MAPK3 pathways via the upstream MKK3 and MKK6 kinases	
HIRIP3	2991202 8	2991442 7	-	0.6 (0.5)	Possibly functions in some aspects of chromatin and histone metabolism	
INO80E	2991513 2	2992426 4	+	0.5 (0.5)	INO80 complex subunit E	
DOC2A	2992500 7	2992904 4	-	1.0 (1.0)	<i>Possibly involved in Ca²⁺-dependent neurotransmitter release</i>	
C16orf92	2994217 6	2994304 9	+	1.1 (1.1)		
FAM57B	2994400 4	2994934 9	-	0.9 (0.8)		
ALDOA	2998607 6	2998903 4	+	0.5 (0.6)	Fructose-bisphosphate aldolase A	
PPP4C	2999519 9	3000388 4	+	0.7 (0.8)	Regulates JNK1 signalling	
TBX6	3000504 6	3001001 5	-	1.0 (1.0)	Transcription factor involved in regulation of early developmental processes	
YPEL3	3001153 1	3001419 0	-	0.6 (0.6)	Possibly involved in proliferation and apoptosis in myeloid precursor cells	
GDPD3	3002369 3	3003230 0	-	0.8 (0.9)	Glycerophosphodiesterase domain	
MAPK3	3003565 8	3004203 1	-	0.7 (0.7)	ERK1. Multiple roles in proliferation and differentiation of preadipocytes	43
CORO1A	3010403 1	3010778 6	+	0.3 (0.5)	Coronin. Actin binding protein	

Supplementary references

31. Merla *et al.* *Am. J. Hum. Genet.* **79**, 332-341 (2006).
32. Henrichsen, C.N. *et al.* *Nat. Genet.* **41**, 424-429 (2009).
33. Huang, J. *et al.* *Hum. Genomics* **1**, 287-299 (2004).
34. Olshen, A.B. *et al.* *Biostatistics* **5**, 557-572 (2004).
35. Venkatraman, E.S. & Olshen, A.B. *Bioinformatics* **23**, 657-663 (2007).
36. Li, J.-Y. *et al.* *J. Biol. Chem.* **277**, 9069-9076 (2002).
37. Cho, J.Y. *et al.* *Exp. Cell Res.* **290**, 155-167 (2003).
38. Mattioli, I. *et al.* *Blood* **104**, 3302-3304 (2004).
39. Fierro N.A. *et al.* *J. Immunol.* **176**, 7346-7353 (2006).
40. Guillemin, G.J. & Brew, B.J. *Redox Rep.* **7**, 199-206 (2002)
41. Parks, C.L. & Shenk, T. *J. Biol. Chem.* **271**, 4417-30 (1996).
42. Minaguchi, T. *et al.* *Cancer Res.* **66**, 11677-11682
43. Bost, F. *et al.* *Biochimie* **87**, 51-56 (2005).

Annexe III: Melanoma Supplementary Information

Supplementary Methods

CNV analysis from CGH arrays

Hybridization signals were extracted using the Feature Extraction software (v.9.5.3.1) and normalized using three independent methods: 1) the local weighted polynomial regression (Loess (Smyth and Speed 2003)), widely used for the analysis of diploid genomes; 2) the PopLowess method proposed by Staaf et al (Staaf et al. 2007), where normalization is applied to population of probes that have been clustered in a deletion, copy neutral or duplication bin; and 3) the more elaborated framework from Chen et al (Chen et al. 2008), which combines several approaches to calibrate channels from all arrays and to centralize the copy number ratios.

After normalization, we segmented the log₂ ratios using Circular Binary Segmentation (Olshen et al. 2004; Venkatraman and Olshen 2007) (with parameters `undo.splits="sdundo"`, `undo.SD=2`, `nperm=10000` and `alpha=0.01`) and attributed a discrete copy number to segments using three independent methods:

1) Scoring-based approach.

We computed a score *S* defined as :

$$S = \frac{r - \text{median}(R)}{\text{mad}(R)}$$

where *R* is the log₂ ratio for a chromosome, *r* the median log₂ ratio for a CBS segment, and *mad* the median absolute deviation, a robust estimator of dispersion around the median. This score reflects how significant a segment is compared to the chromosome baseline, segments with *S* <-4 were classified as CN=0; *S* <-2 as CN=1; *S* >2 as CN=3 and *S* >4 as CN=4.

2) The MergeLevels method from Willenbrock and Fridlyand (Willenbrock and Fridlyand 2005). This procedure is used to effectively remerge similar segments and to produces a new segment level that is used for classification into deletion and duplication events. This procedure has been used in several CNV detection frameworks (Diaz-Uriarte and Rueda 2007; Budinska et al. 2009).

3) Classification based on Gaussian Mixture Model (GMM).

GMM fit Gaussian Model on the log₂ ratios from CBS segments to identify Gaussian components in the distribution (Supplemental Fig. S3). Several models (with different numbers of components) are fitted using an expectation maximization algorithm (Dempster et al. 1977), only the model that minimizes the Bayesian Information Criteria (Schwarz 1978) is kept for subsequent segment clustering. The cluster with the median log₂ ratio the closest to zero is assumed to reflect copy neutral events (CN=2). The right-hand side cluster (with positive ratios) is assumed to reflect duplication events (CN=3), any additional clusters with higher ratios are classified as amplification events (CN≥4).

An interesting property of the GMM is its ability to detect copy neutral events due to cell heterogeneity. We initially thought the GMM clusters with negative ratios would only reflect deletion events. In fact, karyotype analysis revealed that the component left of the diploid state was reflecting mostly diploid events and few deletions. For example, in LAU-Me275, Chr4q and Chr10q both had mean log₂ ratios close to -0.8 (see Fig. 2) and were diploid in 12 and 13 karyotype spreads, respectively (out of 19 spreads), duplicated in 6 and 5 spreads and deleted in one. By contrast, in LAU-Me280, Chr13q had a mean log₂ ratio close to -0.5, was deleted in 10 out of 15 karyotype spreads and its corresponding GMM component was not the adjacent neighbor of the diploid component. This demonstrates that setting thresholds on log ratios is not appropriate and that the ratios should be modeled within a sample and not across samples. In this case, statistical decomposition of the data (e.g. using GMM) is helpful to distinguish between genuine copy number events and loci that are mostly diploid in a cell population but can undergo sporadic copy number events. Based on this observation, we assigned the cluster left of the diploid component as copy neutral, and any cluster with a more negative ratio was classified as a deletion event (CN<2).

Transcriptome analysis

cDNA preparation

mRNA isolation and cDNA preparation were performed following the protocol used by Bainbridge et al (Bainbridge et al. 2006), with some modifications. Specifically, mRNA was purified from 300-500 µg of total RNA from each sample using the µMACS mRNA Isolation Kit (Miltenyi Biotec, Bergisch Gladbach, Germany), inclusive of the optional DNase I treatment (2 U of DNase I, RNase free, Roche, Mannheim, Germany) during the purification. The quality and amount of purified mRNA were determined using a Bioanalyzer 2100 (RNA Nano assay, Agilent Technologies, Basel, Switzerland). cDNA was prepared from mRNA (2-5µg), using the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen Life Technologies, Carlsbad, CA), and quantified with Quant-iT PicoGreen (Invitrogen). The resulting 3-5 µg of cDNA were used for Roche 454 library preparation, according to the manufactures' procedures. RNA preparations from two normal melanocyte cell lines were pooled together and purified as a single sample.

Sequence analysis

Sequences from the Refseq, GeneBank RNA and ESTs databases, as well as 454 reads obtained from the melanoma and melanocyte cell line cDNAs were aligned to the the GRCh37 assembly of the human genome using SIBsim4, a modified version of sim4 (Florea et al. 1998). For sequences that matched to multiple locations, only the best alignment was kept. Finally Refseq mRNAs were used to annotate these unique transcripts and to compute a sequence tag count per transcript and per sample analyzed.

Protein network-guided analysis

We mapped SCNA onto our non-redundant human protein interaction network and identified connected components using the RBGL package (Gentleman et al. 2004). Putative functional clusters within these networks were calculated using the walk trap community algorithm from the igraph package (Pons and Latapy 2005; Csárdi and Nepusz 2006). For community detection a random walk path length of 3 was used and functional clusters were extracted where the modularity was maximal. Clusters with less than 5 nodes were filtered out and not used for further analysis. To test the significance of the clustering, a permutation test was performed by re-calculating the clustering of 1000 random networks generated from the original subnetwork. The resulting random networks had randomized edges but the same degree distribution as the original subnetwork.

References

- Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V et al. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**: 246.
- Budinska E, Gelnarova E, Schimek MG. 2009. MSMAD: a computationally efficient method for the analysis of noisy array CGH data. *Bioinformatics* **25**(6): 703-713.
- Chen HI, Hsu FH, Jiang Y, Tsai MH, Yang PC, Meltzer PS, Chuang EY, Chen Y. 2008. A probe-density-based analysis method for array CGH data: simulation, normalization and centralization. *Bioinformatics* **24**(16): 1749-1756.
- Csárdi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* 1695.
- Dempster AP, Laird NM, Rubin DB. 1977. Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological* **39**(1): 1-38.
- Diaz-Uriarte R, Rueda OM. 2007. ADaCGH: A parallelized web-based application and R package for the analysis of aCGH data. *PLoS One* **2**(1): e737.
- Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8**(9): 967-974.
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**(10): R80.
- Olshen AB, Venkatraman ES, Lucito R, Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**(4): 557-572.
- Pons P, Latapy M. 2005. Computing communities in large networks using random walks. *Computer and Information Sciences - Iscis 2005, Proceedings* **3733**: 284-293.
- Schwarz G. 1978. Estimating Dimension of a Model. *Annals of Statistics* **6**(2): 461-464.
- SIBsim4. <http://sibsim4.sourceforge.net/>.
- Smyth GK, Speed T. 2003. Normalization of cDNA microarray data. *Methods* **31**(4): 265-273.
- Staaf J, Jonsson G, Ringner M, Vallon-Christersson J. 2007. Normalization of array-CGH data: influence of copy number imbalances. *BMC Genomics* **8**: 382.
- Venkatraman ES, Olshen AB. 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**(6): 657-663.
- Willenbrock H, Fridlyand J. 2005. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**(22): 4084-4091.

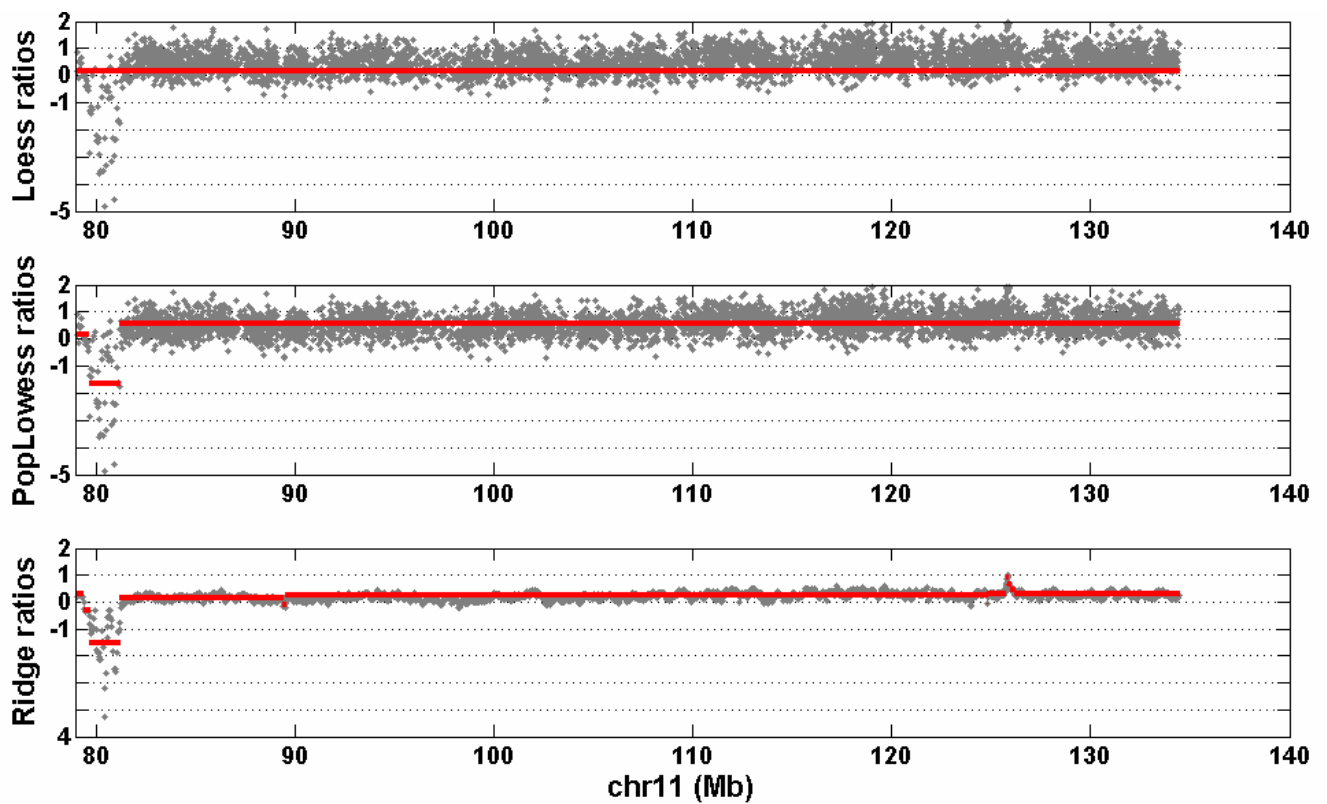
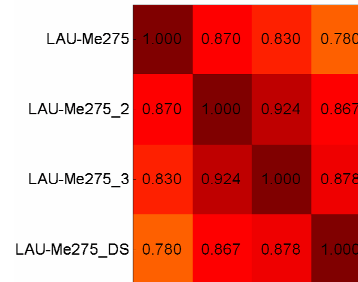
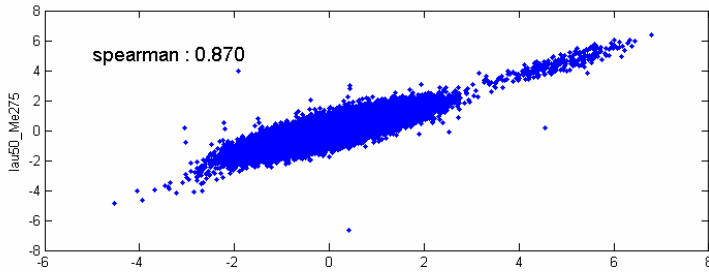


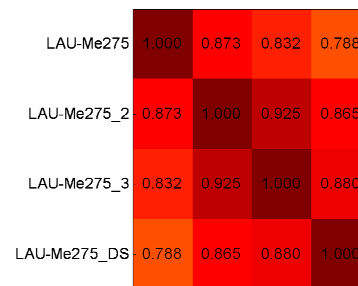
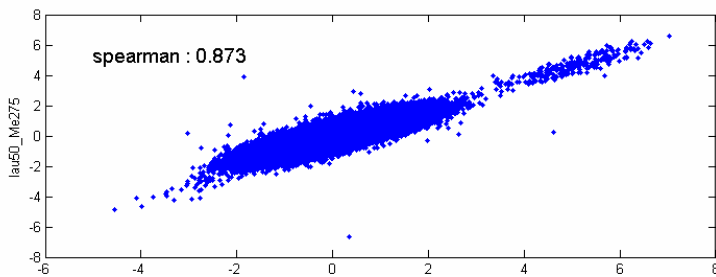
Figure S1. CGH hybridization ratio in a tetraploid region in LAU-Me275.

Each plot shows the hybridization \log_2 ratio at each CGH probe (in gray) obtained using three normalization methods. Ridge refers to the framework from Chen et al. Red segments were obtained using Circular Binary Segmentation. Karyotype analysis of LAU-Me275 revealed 11q amplification ($CN \geq 4$), so the expected CGH log ratio would be two. Here the ratios obtained from three different normalizations failed to reflect the amplification (both Loess and Ridge were close to 0; PopLowess was close to 0.6 indicating 3 copies).

LOESS



POPLOWESS



RIDGE

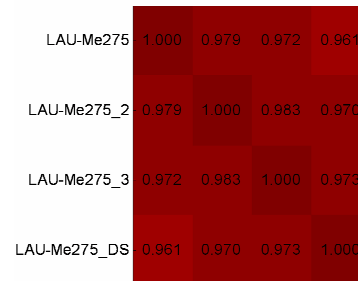
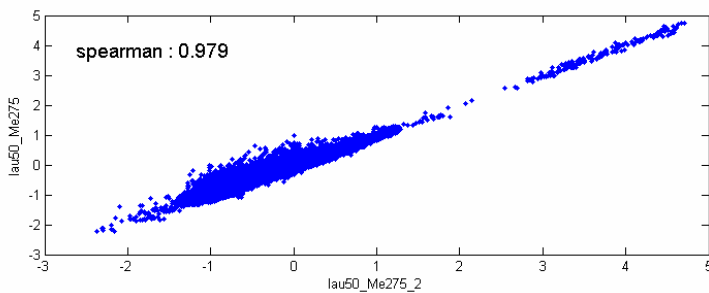


Figure S2. Correlation between replicates using different normalization schemes.

The scatter plots illustrate the correlation at each CGH probe between two replicates. The heatmaps show the correlation for each pairs of replicates. The normalization method is indicated in each plot title. RIDGE refers to the framework from Chen et al.

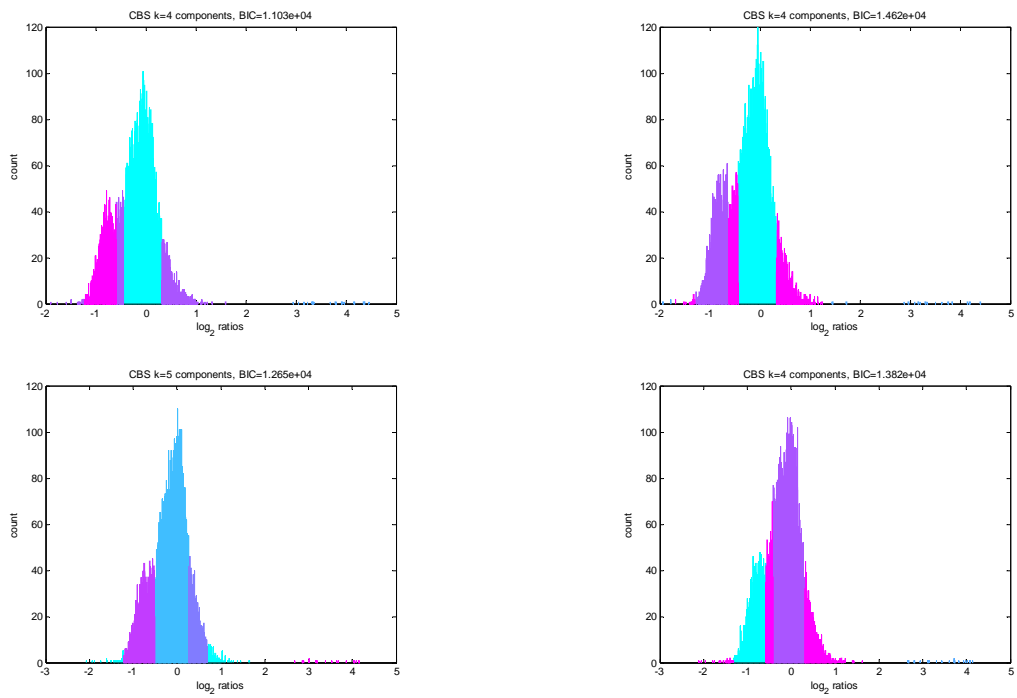


Figure S3. Gaussian Mixtures identified in four replicates from LAU-Me275.

Each histogram shows the distribution of CBS segment log₂ ratios, colors highlight the Gaussian components. The number of components identified and the Bayes Information Criterion are indicated in each figure title.

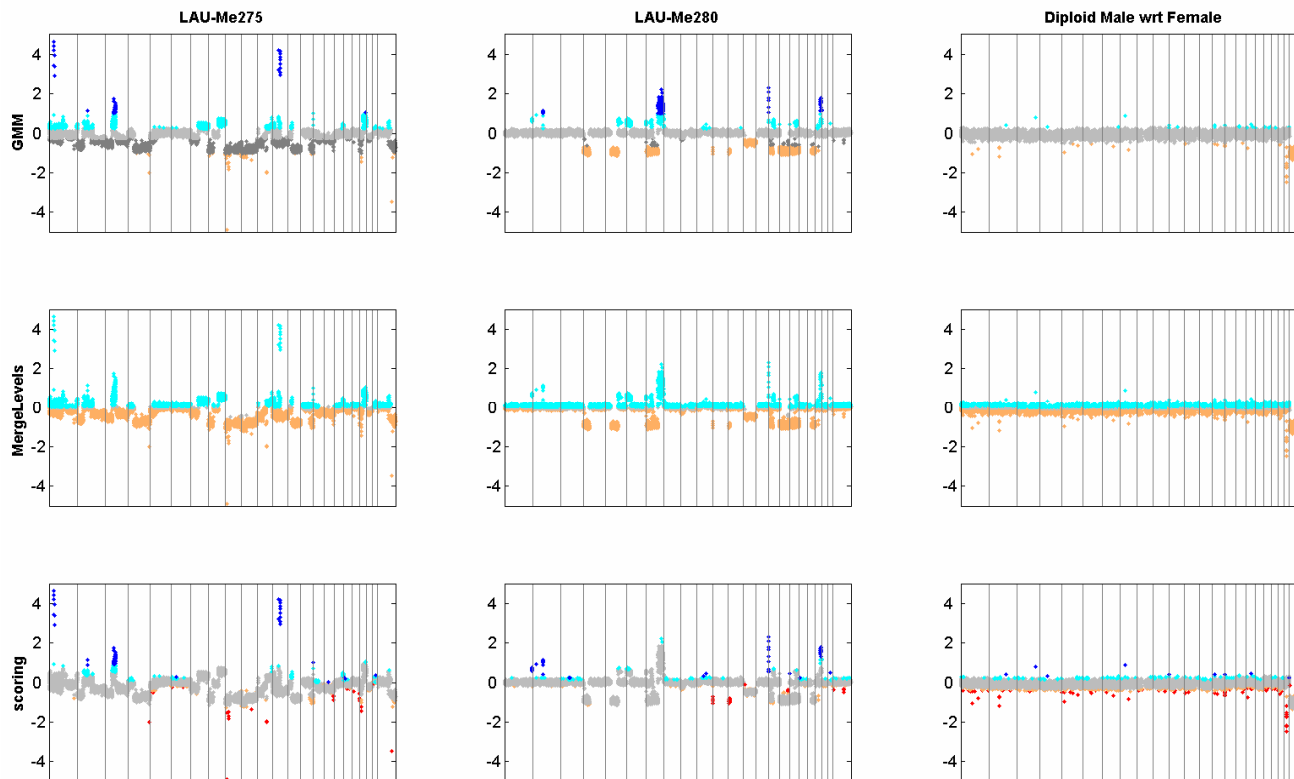


Figure S4. Comparison of CNV detection algorithm on CGH data.

Panels are, from left to right: a melanoma with large deletions (LAU-Me280); a melanoma with large amplifications (LAU-Me275); and a control EBV cell line (male) hybridized using a pool of female references. From top to bottom: CNV classification (following CBS segmentation) using 1) Gaussian Mixture Model (GMM), 2) MergeLevels, 3) the scoring-based approach. Each dot corresponds to a CGH probe with its genomic position on the X axis and its log₂ ratio of hybridization on the Y axis. Colors indicate the copy number state : orange ≤ 1 copy gray = 2 copies, cyan = 3 copies and dark blue more than 3 copies. For the scoring approach distinction is made between 1 copy (orange) and 0 copy (red).

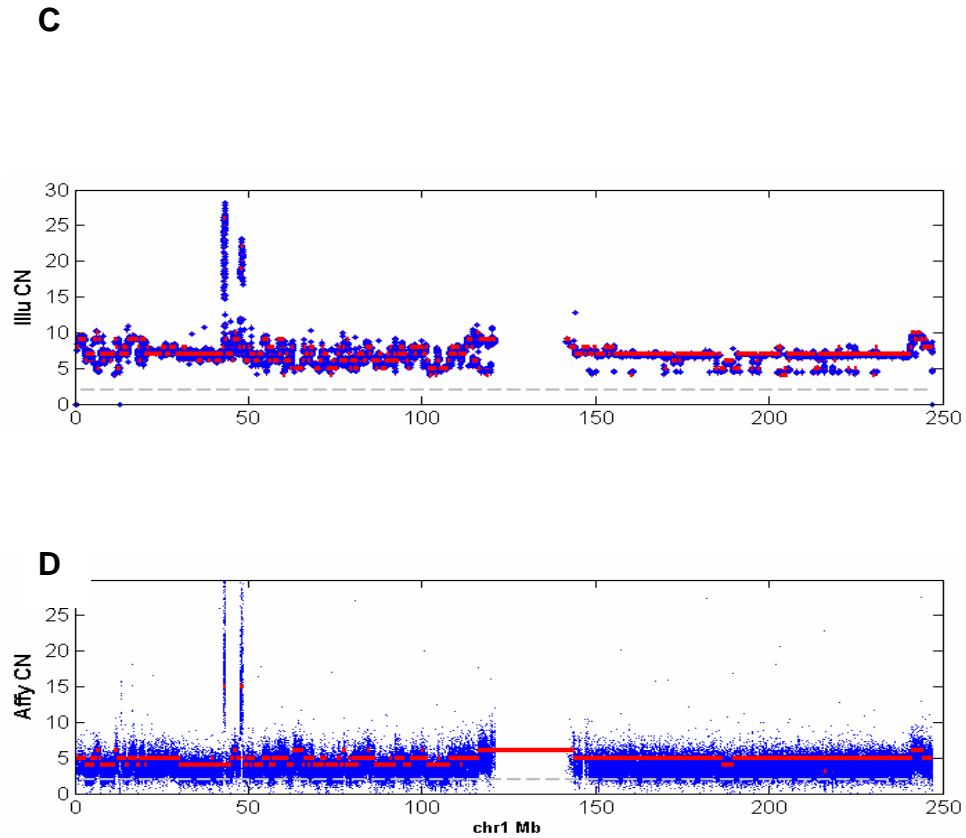
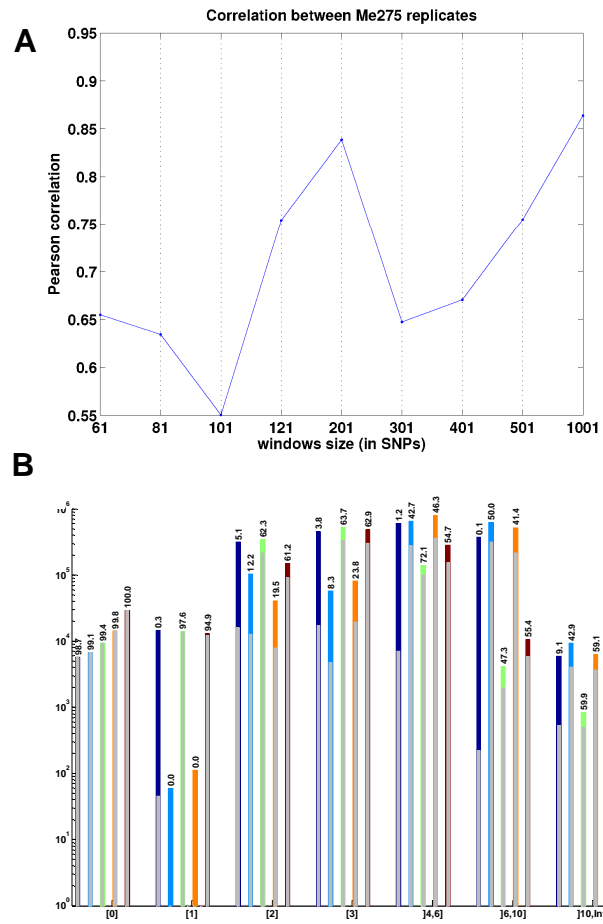


Figure S5. Optimization of Illumina analysis and comparison with Affymetrix prediction in LAU-Me275.

A. Pearson correlation between SNP CN, as a function of OverUnder window size. **B.** Copy number concordance at each SNP for different window sizes of OverUnder. Colors indicate window size parameters, the bar height indicates the total number of SNPs (in log10 scale) found in one replicate. The gray bar indicates the intersection between two technical replicates. The percentage of concordance (number of SNPs found with the same copy number bin in both replicates / total number of SNPs from this given copy number bin in the first replicate) is shown on top of each bar. **C.** Copy number

prediction on chromosome 1 using OverUnder with a window size of 201 SNPs. **D.** Copy number prediction on chromosome 1 using an Affymetrix 6.0 array (with the PICNIC algorithm).

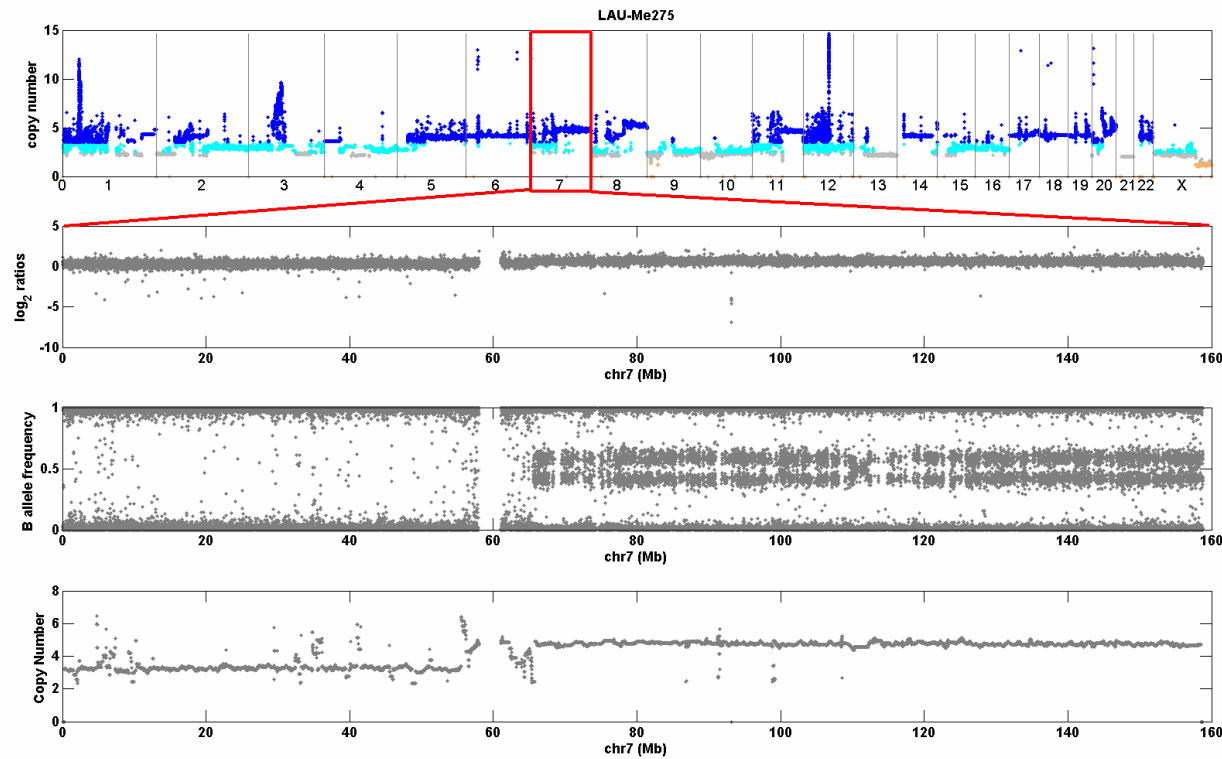


Figure S6. Copy number analysis using Illumina SNP arrays.

DNA from LAU-Me275 was hybridized to Illumina SNP arrays, and the data were analyzed using the method of Attiyeh et al. The top panel shows genome-wide copy number: dark blue indicates more than three copies; cyan:three copies; gray:copy neutral; orange : deletion. Subsequent panels show chromosome 7 with, from top to bottom: Hybridization \log_2 ratio; B allele frequency; and copy number prediction.

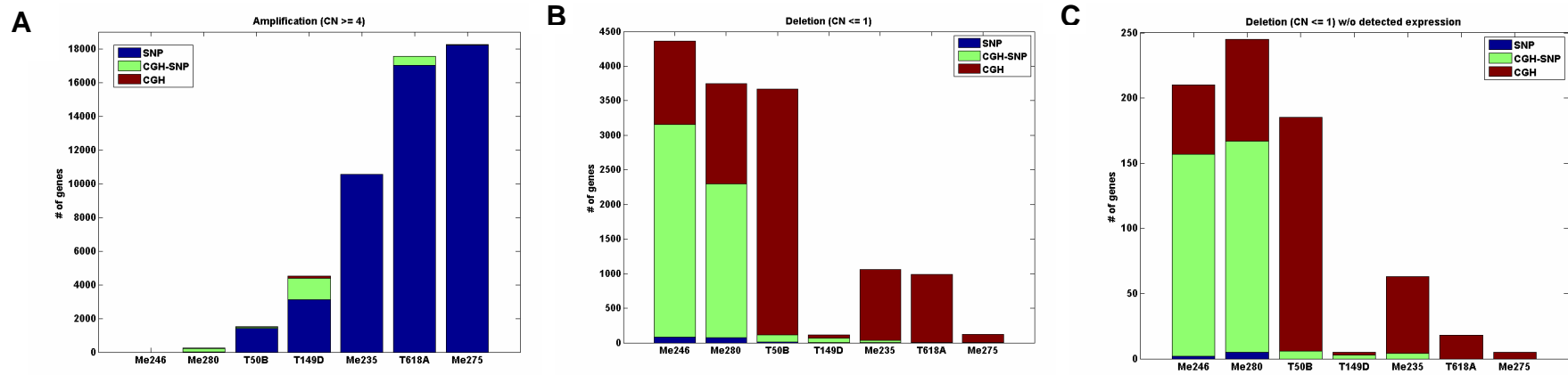


Figure S7. Intersection between CGH and SNP predictions.

A. Intersection between CGH and SNP predictions for genes with more than 4 copies. **B.** Intersection for genes within deletions. **C.** Intersection for genes within deletions for which expression was not detected.

Figure S8. Copy number prediction from CGH and SNP arrays, LOH prediction from SNP arrays.

Supplied as a PDF file.

	SNP arrays							
	LAU-Me280	LAU-Me246	LAU-T618A	LAU-T50B	LAU-T149D	LAU-Me275	LAU-Me235	unique gene count
Focal Amplification	213	0	978	438	894	1853	161	4055
Focal Amplification with 2xOE	85	0	227	106	202	502	25	1089
Arm-level Amplification	0	0	16584	1033	3477	16398	10384	19496
Arm-level Amplification with 2xOE	0	0	2988	263	915	3566	1778	6007
Deletion	2294	3157	2	113	70	2	39	5544
Deletion w/o expression in melanoma but some in melanocytes	167	157	0	6	3	0	4	333
	CGH arrays							
	LAU-Me280	LAU-Me246	LAU-T618A	LAU-T50B	LAU-T149D	LAU-Me275	LAU-Me235	unique gene count
Focal Amplification	0	0	0	26	379	0	4	409
Focal Amplification with 2xOE	0	0	0	6	129	0	1	136
Arm-level Amplification	222	0	549	99	998	42	0	1884
Arm-level Amplification with 2xOE	92	0	148	32	398	29	0	689
Deletion	3668	4281	986	3656	108	122	1059	10711
Deletion w/o expression in melanoma but some in melanocytes	240	208	18	185	5	5	63	634

Supplementary Table 1. Count of genes affected by SCNA.

Table S2. Processed list of SCNA genes in all seven melanoma cell lines.
Supplied as an Excel file.

Table S3. Genomic and transcriptomic data for SCNA-genes in all seven melanoma cell lines.
Supplied as an Excel file.

Table S4. List of pathways significantly enriched in SCNA, and pathway comparison between three melanoma datasets.
Supplied as an Excel file.

Table S5. List of genes contributing to pathway enrichment in three melanoma datasets
Supplied as an Excel file.