Identification and validation of copy number variants from high-throughput microarrays

Mid-thesis report
6th November 2009

Armand Valsesia

University of Lausanne
Ludwig Institute for Cancer Research
Swiss Institute of Bioinformatics

Supervisors:
Professor C. Victor Jongeneel
Professor Sven Bergmann

Table of contents

1 Abbreviations

AUC	Area Under the Curve
CBS	Circular Binary Segmentation
CGH	Comparative Genome Hybridization
CNAT	Copy Number Analysis Tool
CN	Copy number
CNP	Copy number Polymorphism
CNV	Copy Number Variation
CT	Cancer-Testis genes
EBV	Epstein-Barr virus
FISH	Fluorescence in situ hybridization
HMM	Hidden Markov Model
LOESS	Local Weighted Polynomial Regression
FPR	False Positive Rate
PBL	Peripheral Blood Leukocyte
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
QC	Quality Control
ROC	Receiver Operating Characteristics
SNP	Single Nucleotide Polyphormism
TPR	True Positive Rate

2 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 the humans ([1-8]) and other mammals ([9-12]). These genomic variants can impact on both somatic and germline genetics. The link between CNVs and inherited diseases has been previously established ([13-15]). Copy number plasticity is typical of cancer cells ([16]). Such genomic aberrations were identified already more than a decade ago using array-based comparative hybridization ([17], see also [18]). It has been demonstrated that CNVs near oncogenes or tumour suppressor genes can affect gene expression levels or result on the expression of chimeric fusion genes ([18, 19]). CNVs can affect gene expression in apparently healthy individuals ([20]) and initial evidence has also been provided that CNVs could shape tissue transcriptomes on a global scale ([12]).

2.1 Method for CNV detection

CNV detection can be made using different experimental techniques. Classical methods in cytogenetic such as karyotype were used to detect whole-chromosome aneuploidy but did not allow identification of CNVs, which were beyond their detection capacities. FISH and PCR-based approaches are reliable techniques to identify CNVs on small loci. Although these techniques are routinely used for CNV validation, they lack the throughput required for genome-wide analysis.

Tremendous progress has been made with the advent of micro-arrays ([21], [17]) that now enable to interrogate with up to several millions of probes the genome for copy number change. One can notably differentiate between array comparative genome hybridization (CGH) and Single Nucleotide Polymorphism (SNP) arrays.

A CGH experiment consists in hybridizing in competition two genomes: a test and a reference. Each genome is labelled with a different dye (red or green), subsequently to hybridization on the array, a ratio green versus red is computed. Significant shift from the baseline (unit ratio or zero log ratio) reflects copy number changes. CGH is now a *well-established* technique for CNV analysis and is used routinely for clinical diagnosis.

With SNP arrays, DNA usually corresponding to a single genome is hybridized and allele-specific intensities are quantified. By combining the intensities of the two probes for a given SNP and comparing to the same SNP from other arrays, it is possible to obtain also information on the copy number state. It is important to emphasize that most SNP arrays used so far for genotyping clinical

cohorts were not designed for CNV (dosage) detection, but only to call the three possible genotypes of SNPs. CNV analysis from SNP arrays is challenging for several reasons: Firstly, when analysing very large datasets, it is very likely that experiments were conducted at different times and/or by different laboratories, which often introduces severe batch effects for the raw intensities. Thus the first challenge in CNV calling is to ensure proper normalization of the raw data. Secondly, due to the large noise in the SNP 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. So more reliable prediction can only be made by integration of intensities from several neighbouring loci. Indeed this strategy is employed by many different CNV detection methods ([22-27]). 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 ([28-30]), there are very few methods to merge individual CNV predictions into regions at the population level: Redon et al. ([3]) merged CNVs based on the extent of their overlap, whereas Itsara et al. ([31]) manually annotated complex regions.

2.2 Association between CNV and clinical phenotypes

Recently numerous SNP-based genome wide association studies have provided new insights to complex and metabolic diseases [32-36]. Nevertheless, even in a very large cohort, one can usually explain only a small fraction of the genetic variance (i.e. 12%) for quantitative trait such as height [35]. Copy number variation (CNV) is the most frequent structural variation in the human genome and encompasses more nucleotides than SNPs. But to date, there are very few published results regarding their association to diseases [37]. There are many reasons; Firstly, as explained previously, methods for CNV detection are not as robust as methods for SNP genotyping and the resolution on SNP arrays (such as Affymetrix 500K and Illumina 550K) provides a low coverage for complex genomic regions, which challenges the analysis. Second, the methodology to associate CNVs to phenotype is still a topic under active research. And last, analysing rare CNVs is a challenging task, as their number can be very large and it is hard to distinguish them from false positives.

2.3 Research outline

2.3.1 Detection, integration and validation of CNVs from a very large clinical cohort

In collaboration with the Lausanne hospital (CHUV), we have access to a large clinical cohort (CoLaus). The aim of this study has two main goals: First we provide an extensive survey of candidate CNVs which 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 this cohort. Second, -since the methods for detecting individual CNV profiles and merging those into consensus regions have not yet been well established-, we also developed new algorithms for these goals, and devised novel techniques to evaluate and compare them with existing methods.

2.3.2 CNV profiling of metastatic melanoma

In collaboration between the Ludwig Institute for Cancer Research, Universities of Lausanne and Geneva, the Swiss Institute of Bioinformatics and the CHUV, we are performing a comprehensive genomic profiling of metastatic melanoma. More specifically we are interested in recurrent copy number aberrations and how these relate with gene expression of affected genes. We have analysed a collection of seven metastatic melanomas with matched control cell lines from the same patients. All samples were analysed with karyotyping, CGH and SNP arrays. To date, the transcriptome of two melanomas has been sequenced with ultra-high throughput technology (Roche 454). More samples will be sequenced, including a melanocyte that will provide a baseline of transcript levels in normal cells. Also we plan to do methylation analysis and exome sequencing (using sequence capture arrays and Solexa sequencing).

2.3.3 Studying the CN polymorphism of Cancer-Testis genes

Cancer-testis (CT) genes are expressed only in cells of the germ line in normal individuals, but re-expressed in a number of cancers, where some of them are thought to contribute to the malignant phenotype. Spontaneous immune responses to CT gene products are commonly found in cancer patients, and CT-derived peptides are increasingly being used in cancer immunotherapy. More than half of all CT genes are located on the human X chromosome, mostly in segmentally duplicated regions, and appear to be under strong diversifying evolutionary pressure. Because of their association with segmental duplications, very little is known about their genetic diversity in human populations. We have designed a custom CGH array that will be of use to derive the copy number status of cancer-testis in our collection of metastatic melanoma.

3 First part: Detection, integration and validation of CNVs from a very large clinical cohort

CoLaus is a population-based health survey to study the genetics of hypertension and cardiovascular disease [38]. More than 6000 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 ([39]).

A number of SNP-based genome-wide association studies that employed the CoLaus data have already been reported ([32-36, 40-42]). Although so far there is no evidence for common CNVs contributing significantly to the kind of clinical phenotypes measured in CoLaus phenotypes ([43]), the number of rare CNVs and their contribution to clinical phenotype remains unclear. We aim at identifying both common and rare CNVs in the CoLaus population and subsequently investigate the association with the CoLaus phenotypes.

Although there has been tremendous development of new methods for CNV analysis, there is no gold standard, especially for Affymetrix 500K arrays. At the beginning of this project, there were few publicly available algorithms for analysing SNP arrays. Most of the methods have been developed and trained for CGH data, which are much more reliable than SNP arrays for CNV detection. Among the SNP dedicated software (e.g. dChip [44], CNAG [25], GEMCA [27]), only available for Windows operating system, none could scale for the analysis of a very large dataset. Only CNAT ([26]) 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 its performance. In this context, we developed our own method, that is based on a Gaussian Mixture Model and we compared it to three existing CNV detection methods. We also developed two merging strategies, which were applied to create a map of CNV regions. In this report, we study how CNVs predicted by the various algorithms coincide with previously reported variants. We also investigate the concordance in predicting CNVs in a *sub-sample* of individuals that were also genotyped on the Illumina platform. Finally we compare the sensitivity and specificity of the different approaches using related individuals.

3.1 Material and methods

3.1.1 Detection of Copy Number Variants

Affymetrix Copy Number Analysis Tool

We used the Affymetrix Copy Number Analysis Tool (CNAT [26]) to attribute a copy number status to each SNP of each CoLaus individuals. CNAT uses intensities, as normalized by GTYPE ([39]), performs additional array normalization (such as PCR bias correction), combines intensities in a copy number ratio (see below) then uses an Hidden-Markov Model (HMM) to predict a copy number state at each SNP. CNAT has two HMM implementations (*CNAT.total* and *CNAT.allelic*). *CNAT.total* computes the CNratio as the log of the allele signal sum, whereas *CNAT.allelic* approach uses the sum of the log allelic signal.

$$CNratio(total) = \log_2(\frac{S_A + S_B}{R_A + R_B}) \qquad CNratio(allelic) = \log_2(\frac{S_A}{R_A}) + \log_2(\frac{S_B}{R_B})$$

In the equations, S and R refer to the test sample (individual) and a reference panel respectively; A and B refer to the SNP alleles. According to the GTYPE manual, *CNAT.total* is optimized for noise reduction but allelic subtle changes like hemizygous deletion (CN=1) or single copy gain (CN=3) will be harder to detect.

Correction of batch effects

By doing a Principal Component Analysis on the CN status of SNPs (as predicted by *CNAT.allelic*) across CoLaus individuals, we found that individuals clustered into 4 distinct groups, which corresponded to four independent genotyping centers (see Annexes for details). To correct this batch effect, we performed normalization within each center and used, as references, 280 randomly chosen samples (with equal proportion of males and females).

Aroma normalization

In parallel to CNAT, we normalized the CoLaus data with the Aroma.affymetrix framework [45]. Normalization was done within each genotyping center and with at least 336 individuals. 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 CN ratios for a

given sample at a given SNP, 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) is a *state-of-art* segmentation algorithm ([22, 23]); it identifies change points using a maximal *T-statistics* and assesses segment significance with permutations. We applied CBS on the CN ratios as normalized by the Affymetrix.Aroma framework. CBS only segments the ratios and does not perform CNV classification into deletion, duplication or diploid events. After inspection of the distribution of $\log_2$ ratios, we decided to classify into gains regions having a $\log_2$ ratio greater than 0.25 and into losses regions with $\log_2$ ratios lower than -0.25.

Gaussian Mixture Model

Raw copy number ratios were smoothed along physical position using Loess filtering with a 41-probe window size. Next, four component Gaussian mixture model (one component for each of the following copy number states: deletion, copy-neutral, 1 and 2 additional copy) was fitted to the smoothed copy number ratios with a constraint on the difference between the mixture means. Then, for a given individual we determined the probabilities for each copy number state. The copy number was finally determined as the expected copy number (dosage). I.e. a SNP with probabilities: 1% for CN=1, 9% for CN=2, 85% for CN=3 and 5% for CN=4, would have a dosage value equal to 2.94 (1*0.01 + 2*0.09 + 3*0.85 + 4*0.05).

3.1.2 Integrating CNVs from CoLaus individuals into Copy Number Polymorphism

Simple merge

The data can be represented as a matrix of individuals by SNPs, where each element is the Copy Number status. The “simple merge procedure” consists of combining adjacent SNPs that share the same CN prediction 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 a 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 we developed. It first partitions the genome into smaller regions, whose boundaries are a long stretch of SNPs in the diploid state. Then for each of these regions, it performs a principal component analysis of SNP data across individuals (Figure 1). Only components that explain most of the variance (e.g. 90%) are used to train a self-organizing map (SOM) to cluster SNPs with similar variance. Strictly adjacent SNPs within a same cluster are then merged into CNV regions.

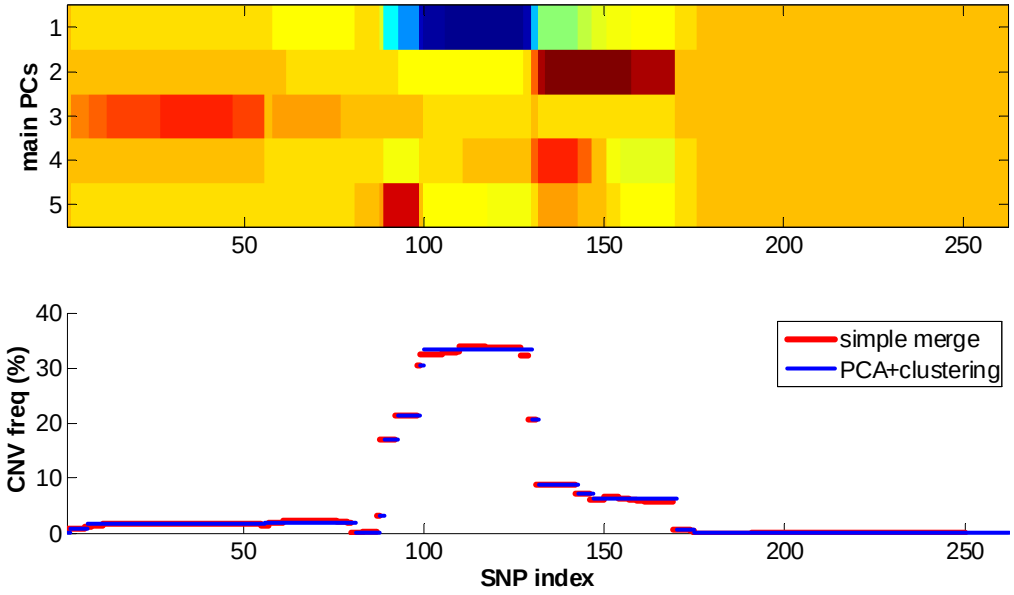

Figure 1 Merging SNPs into CNV regions using principal component analysis

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≈5600)

Replication on Illumina arrays

A subset of CoLaus individuals were analysed on the Illumina arrays (550K version 1 & 3, 1M [48]). Intensities were normalized within BeadStudio using 120 Hapmap samples as references. Only SNPs that could be remapped to the 550K version 3 array (genome assembly build36) 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 and excluded any unique regions.

3.2 Results

3.2.1 Identification of Copy Number Variant in CoLaus

To detect CNVs in the Cohorte Lausanne, we applied four different CNV detection algorithms to the 5600 Affymetrix 500k SNP array profiles: the two CNAT, CBS and our own algorithm based on a Gaussian Mixture Model. We restricted our analysis to autosomes allowing us to use a mixture of males and females as reference panel. Since we used copy number ratios as normalized by the Affymetrix.Aroma framework prior to applying CBS and our mixture model, we refer to latter methods as, respectively, *AROMA.CBS* and *AROMA.mix*.

In a second step we attempt to reduce the complexity of these CNV profiles by merging adjacent SNPs that contain highly redundant information into CNV regions. The first method (called "simple merge") joins neighbouring SNPs that take identical copy number values across all CoLaus participants. This simple approach already reduced significantly the number of SNPs by combining them into regions (i.e. by about 10 fold for CNAT.allelic, 20 fold for CNAT.total), irrespectively of whether these regions are CNVs or copy neutral. Nevertheless this approach is extremely stringent and many of such CNV regions are fragments of single CNVs with noisy boundaries. Thus we devised a second method that employs a principal component analysis (PCA) of such regions and only merges such fragments into segments that explain a significant amount of the variation (i.e. 90%) at the population level (see Methods for details).

Subsequently, we excluded any CNV regions found in fewer than five individuals. We distinguish between Copy Number Polymorphisms (CNPs, 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 by the four different methods and the two merging methods are shown in Figure 2. *CNAT.total* and *AROMA.CBS* are conservative methods that generate significantly less regions than *CNAT.allelic* and *AROMA.mix*. The simple merging procedure produces lots of small regions (<1kb or single SNPs) which are commonly integrated into much fewer regions with the PCA-based method. The PCA-based method is able to reduce the total number of regions by 35%, 70%, 67% and 53% for *AROMA.mix*, *AROMA.CBS*, *CNAT.allelic* and *CNAT.total*, respectively.

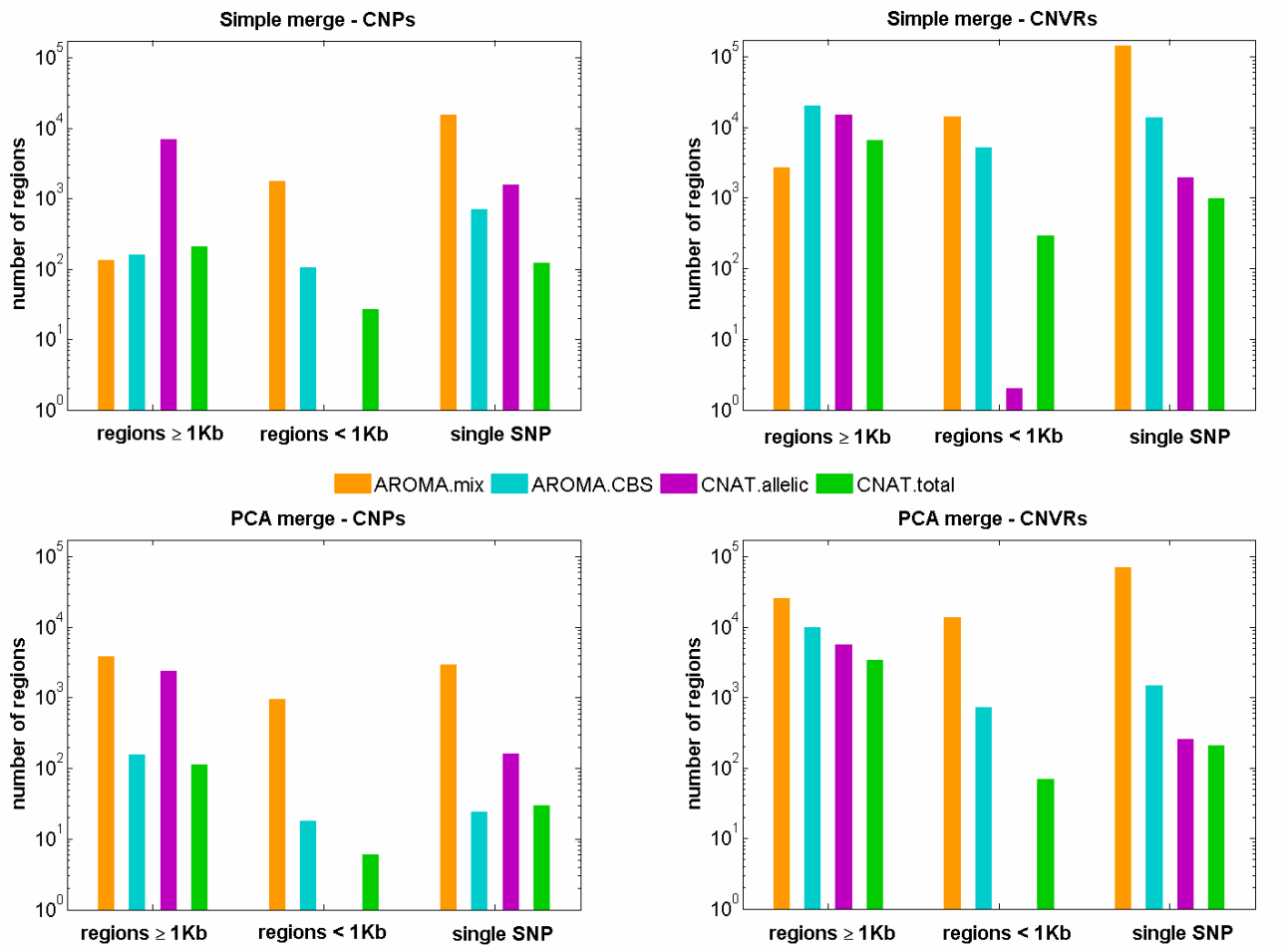

Figure 2: counts of CNVs identified with different methods

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≈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.

The fraction of the genome effectively covered by regions, as obtained from the PCA-merge, is reported in Table 1. Although *AROMA.mix* has much more CNPs than the other methods, they only cover about 2.4% of the autosomes. *CNAT.allelic* predictions for CNPs cover for 12.4% of the autosomes, *AROMA.CBS* and *CNAT.total* for 1.5 and 0.7% respectively. We also checked the coverage with rare variants (CNVRs), *AROMA.mix* had the lowest autosomal coverage only 9.8% whereas *AROMA.CBS* had the highest with 42.4%.

	CNPs	CNVRs
AROMA.mix	2.4	9.86
AROMA.CBS	1.54	42.43
CNAT.allelic	12.4	30.88
CNAT.total	0.73	12.71

Table 1: genome coverage of CNVs identified by different methods and merged using the PCA-based approach
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).

3.2.2 Comparison with known CNVs

The Database of Genomic Variants (DGV [1]) is a curated catalogue of structural variation in the human genome. We downloaded its content (release 7, March 2009) and only kept CNVs discovered from SNP or CGH arrays (BAC and ROMA arrays were excluded). We added to this dataset CNVs from individuals of European ancestry that were reported by Itsara et al. ([31]) This combined dataset of "known" CNVs included 17804 autosomal CNVs, whose size ranged from 1kb to 3Mb.

We then computed the overlap between CNVs generated by each prediction methods and this reference dataset (Figure 3). We report this overlap as the Jacquard coefficient, which is the ratio between the intersection and the union of two CNVs. A ratio close to one implies that the two CNVs have very similar boundaries; a ratio equal to zero indicates no overlap 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 much less individuals than the CoLaus dataset, it was important to compare the distribution of overlaps with the CNV generated by the different methods in a controlled setting. Therefore we computed for each method the expected overlap using reshuffled data (from n=1000 permutations). Estimated p-values for observing more or less CNVs with a given overlap are shown in Figure 3 (see Table 2 for the corresponding *t*-statistics). We observed that all prediction methods were enriched (with respect to the controls) for known CNVs (see all Jacquard coefficient bins above 25%) and depleted for novel CNVs (Jacquard coefficient bin 0%).

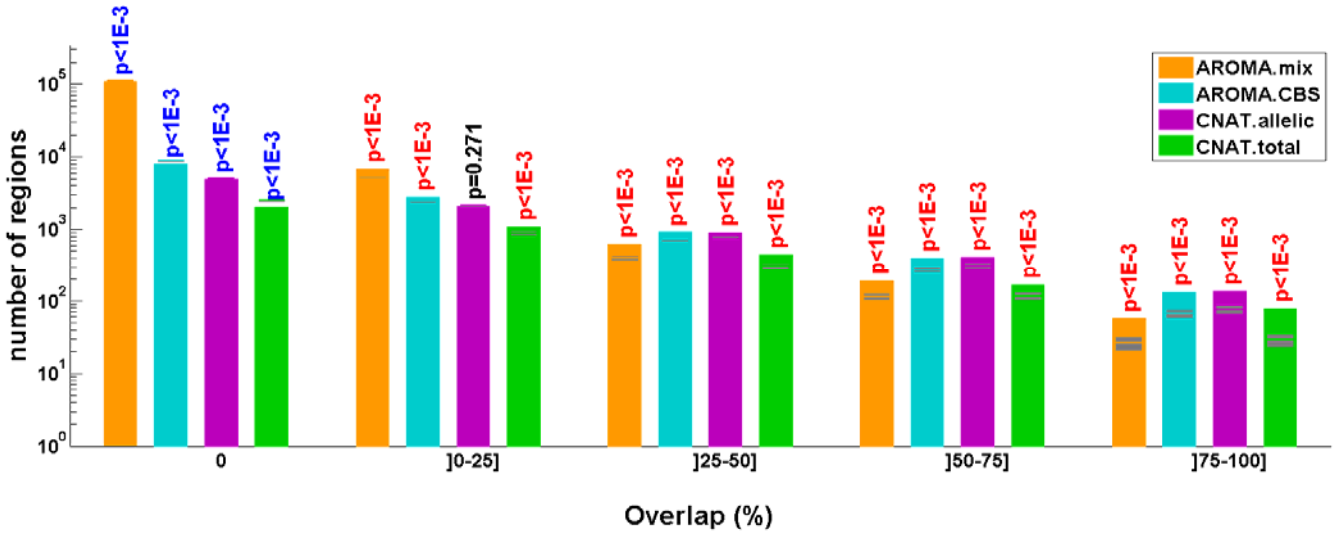

Figure 3 Overlap between CNVs identified from CoLaus and published CNVs

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 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 (at α=1%) are shown in black.

	AROMA.mix	AROMA.CBS	CNAT.allelic	CNAT.total
0	-19.16	-10.78	-3.60	-9.69
]0-25]	16.62	6.15	-0.61	5.50
]25-50]	11.42	7.09	3.12	7.56
]50-75]	7.33	7.52	5.61	4.49
]75-100]	6.43	8.11	6.94	9.54

Table 2 *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 at α=1%. Positive (negative) T statistics indicates enrichment (depletion) with respect to the expected counts.

3.2.3 Validation with Illumina arrays

Using Affymetrix CNVs that included at least one individual probed on the Illumina arrays, we checked the fraction that could be replicated (Figure 4).

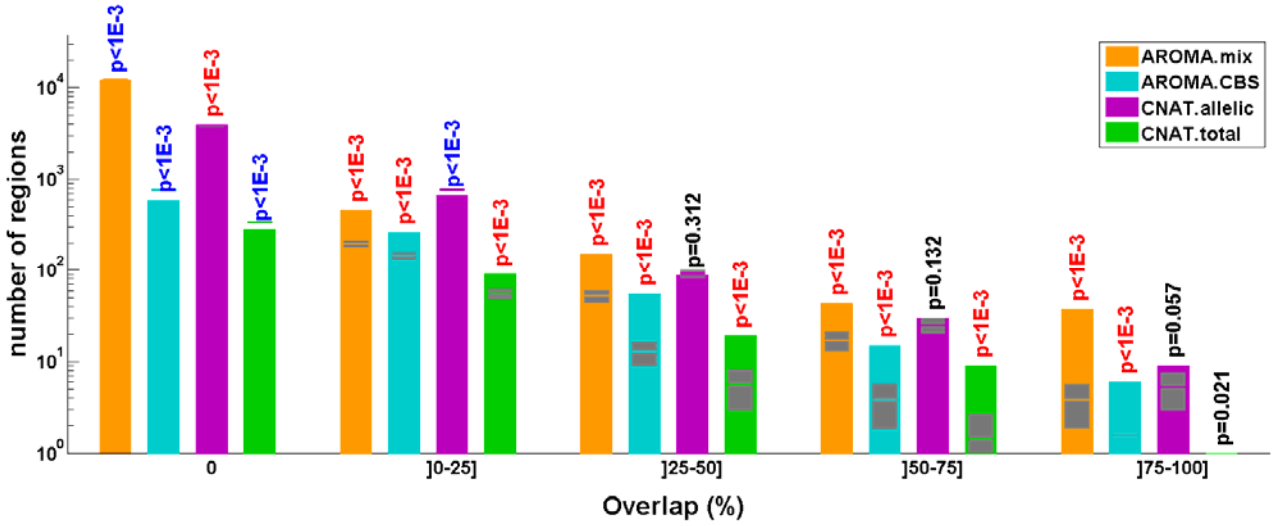

Figure 4 Overlap between CNVs identified from Affymetrix and Illumina data

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 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 (at α=1%) are shown in black.

CNAT.allelic was significantly enriched for CNVs that were not called on the Illumina platform and generated as many CNVs with an overlap of greater than 25% as would have been expected by chance (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* identified less CNVs that were not seen using the Illumina data indicating much better specificity. However, not a single region with an overlap greater than 75% was generated pointing to poor sensitivity. Both *AROMA.CBS* and *AROMA.mix* performed well (showing depletion of CNVs unique to the Affymetrix data and enrichment of confirmed CNVs). Interestingly, *AROMA.mix* predicted many more CNVs than *AROMA.CBS* and the difference with respect to predictions from reshuffled data was much stronger than for all the other methods (Table 3). We also performed the above analyses independently for CNPs and CNVRs (data not shown) arriving at the same results.

	AROMA.mix	AROMA.CBS	CNAT.allelic	CNAT.total
0	-21.30	-13.68	3.60	-7.56
]0-25]	16.01	10.18	-4.20	5.46
]25-50]	13.34	11.80	-0.49	5.21
]50-75]	6.45	5.96	1.00	6.17
]75-100]	17.53	6.15	1.61	1.35

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 α=1%. Positive (negative) T statistics indicates enrichment (depletion) with respect to the expected counts.

3.2.4 Predicting relatedness between individuals based on their CNV profile

Analysis of the CoLaus SNP-profiles revealed that five individuals had been genotyped twice and it also included 157 pairs of first-degree relatives (either sibling or parent-offspring relationships). Using this information, we investigated whether predicting relationship between these individuals would be feasible using exclusively their inferred CNV profiles. To this end we computed the Euclidean distance between all 162 pairs of related and between a *sub-sample* of 2000 unrelated pairs. Knowing the true relationship status, we computed ROC curves for each CNV prediction methods and each merging approach (figure 5). To evaluate the robustness of the ROC curves we reiterated the analysis 100 times choosing randomly the pairs of unrelated individuals.

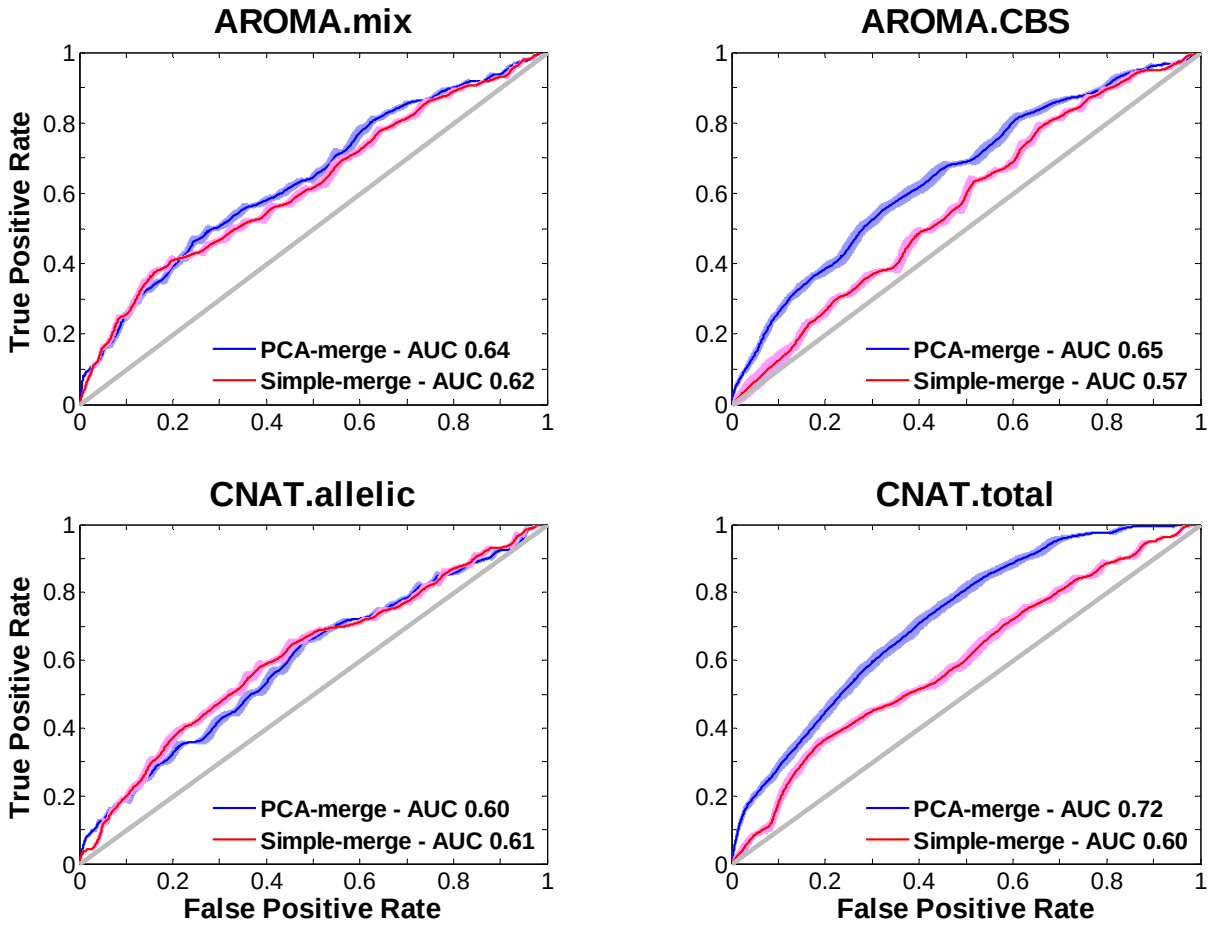

Figure 5 Performance for predicting relatedness based on CNV profiles generated by different methods

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 on top and merging procedure by colour, see legend). The analysis employed 162 pairs of individuals known to be related and 2000 pairs of unrelated individuals. Curves correspond to the mean (solid lines) +/- two standard deviation (light blue or light red surfaces) from n=100 permutations. The Area Under the Curve (AUC) values are shown in the legends, value above (below) 0.5 indicates a better (worse) performance than a random predictor.

All the prediction methods had significant prediction power with Area Under the Curve (AUC) values >0.6. The less conservative CNV detection methods *CNAT.allelic* and *AROMA.mix* did not show a significant difference between the PCA-based and the simple merging approach. However, for the more conservative methods *CNAT.total* and *AROMA.CBS* there was a clear advantage in performance of the PCA-based over the simple merging methods. We checked whether filtering for rare CNVs and excluding small regions (<1kb) would improve the performance (Figure 6). For all methods, there was no significant difference when excluding or keeping such small regions. For *CNAT.allelic,* there was some small improvement when filtering for rare CNVs. For *AROMA.mix* filtering rare CNVs (with frequency < 1%) significantly improved the AUC. This improvement is particularly strong in combination with the PCA merge (giving AUC up to 0.725, which is the best value we obtained across all methods).

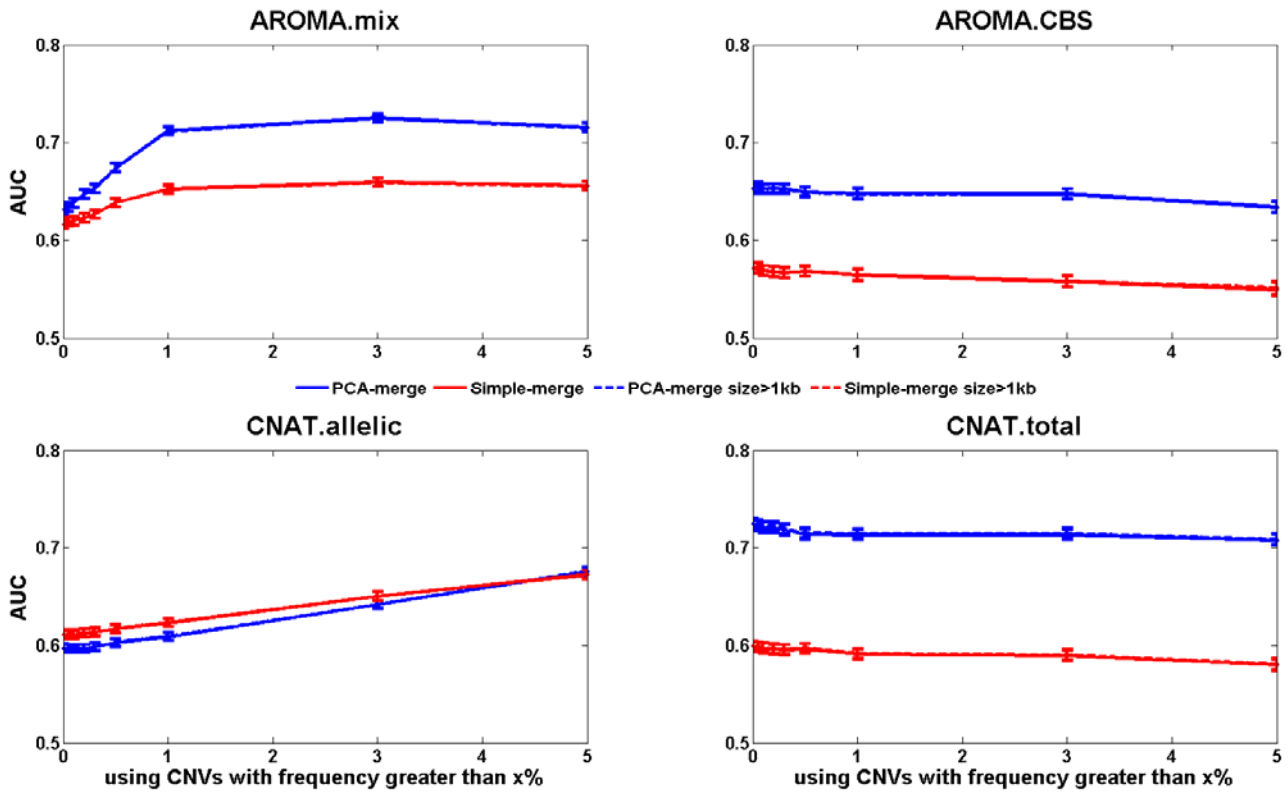

Figure 6 Performance for predicting relatedness based on CNV profiles generated by different methods

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 colours. 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.

3.3 Discussion and perspectives

3.3.1 Properties of the PCA merging technique

The simple merging approach, a naïve technique, is able to concatenate about half a million SNPs into about 40k genomic regions for *AROMA.CBS* and 8k for *CNAT.total*. However this approach leaves CNV edges fragmented into regions as small as single SNPs. Therefore we have developed a novel merging technique, which is purely data driven: only (orthogonal) components which explain most of the variance are used to define CNV regions. By contrast to *overlap-based* approach that is often used to merge predictions into CNV regions, there is no need for *ad-hoc* thresholds. This PCA-based method provides a significant improvement over the simple approach. For conservative CNV detection methods (i.e. *AROMA.CBS)*, it reduces the number of regions by 56% and for relaxed methods like *AROMA.mix*, the reduction was 35%. The PCA merge was able to significantly reduce the number of single SNPs by re-attributing them to existing regions. Similarly, small regions (<1kb) were extended either by incorporating single SNPs or by merging them with other small regions.

3.3.2 Comparison of the different CNV prediction methods

We demonstrated that *CNAT.allelic* predicted by far the most CNVs, but that a relatively small fraction of these could be replicated and therefore 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* appeared to be overly conservative and is likely to miss subtle, but real CNV events. CBS is a very efficient segmentation algorithm, as confirmed by the good replication of its predictions. Our Gaussian Mixture Model, *AROMA.mix*, is also performing much better, both for sensitivity and specificity, than the two CNAT implementations. *AROMA.mix* also has an increased sensitivity with respect to *AROMA.CBS*, it was able to find 50% more CNPs, covering 2.4% of the autosomes, whereas CNPs detected with *AROMA.CBS* only covered 1.5%.

Yet, it is an open question whether these latter methods are indeed more sensitive to capture smaller CNVs. This is very difficult to evaluate from both Affymetrix 500K arrays and Illumina 550K arrays, because the SNP density is not sufficient to assess whether a CNV composed by few SNPs (i.e. smaller than 5 SNPs), is indeed a true positive. However with newer and higher-density arrays (Affymetrix 6.0 or Illumina 1M), identifying smaller CNVs is indeed easier. Finding extremely rare and/or small CNVs is definitely the next challenge for finding causal variant to diseases. The High Resolution CNV discovery project ([43]) recently published important results on this matter. In this

survey, 40 individuals have been analysed on Nimblegen arrays with more than 42M probes and it was estimated that about 95% of the common CNVs (with frequency greater than 5%) could be discovered, with a size down to 500bp. More than 12000 CNV regions were identified, but only 30 candidate loci that could influence disease susceptibility. The authors conclude that, for complex traits, the variance missed by SNP genome wide association studies, will not be accounted by common CNVs. Therefore this leaves open the identification of either common CNVs with very small effect size or rare/small CNVs with stronger clinical impact.

3.3.3 Improving our Gaussian Mixture Model

AROMA.mix makes prediction at each SNP without considering predictions previously made at neighbouring SNPs from a same individual; so with noisier dataset, it will be much more affected with outliers. In contrast *CNAT* smooth the input ratios before analysing them and predictions are made using a Hidden Markov Model. Therefore *CNAT* takes into account the CN state of adjacent SNPs which protects from local fluctuations in the data. CBS also smoothens the ratios, and change points are robustly identified using permutations. One way to improve predictions from our Gaussian Mixture Model would be 1) for higher density arrays, to increase the window size used by the loess, prior to CNV calling; or 2) using a sliding window approach to either remove outliers or replace their values with the median ratios in the window.

Our model can shift the separation between the Gaussian components using an optimization algorithm. Such optimization is bounded to a maximal number of fits and number of iterations per fit, increasing these limits will lead to more accuracy but this will increase the run time significantly.

Currently our model only considers deletion, copy neutral, single copy or multiple copies. Since very few homozygote deletions were observed with other applied algorithms, we did not incorporate such dedicated component in our analysis. Nevertheless, our Gaussian Mixture Model implementation allows such extension.

3.3.4 Validation of CNVs in a large clinical cohort

Validation is an essential part of any CNV discovery project. PCR, Southern and many other targeted techniques are useful to predict accurately the copy number at a given locus, but the throughput is a severe limitation when large numbers of CNVs need to be validated. The Database of Genomic Variant is a valuable resource to reduce the fraction of CNVs to be further validated. Nevertheless for a very large cohort, there will still remain a consequent fraction of novel CNVs. Therefore replicating a number of individuals (i.e. a few hundreds) on an independent array platform

is needed. With the recent reduction in the cost of microarrays, such replication now becomes affordable to any large cohort analysis. Yet it is still an interesting and open question, whether replicating a number of individuals on a higher-resolution array is a better strategy than replicating even more individuals on either a similar resolution or on a targeted array.

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.

3.3.5 Conclusion and Perspectives

Our Gaussian Mixture model and our 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 bound to SNP arrays, they only require an input matrix of hybridization ratios (for the former) or copy number values (the latter). Thus they can be applied to other platforms such as CGH arrays.

Based on the analysis of a 6000 individual strong cohort, we have comprehensively documented the genome for both CNPs and rare CNV regions. This CNV resource is already being use in a clinical context. To date, many clinical diagnosis laboratories rely on the content of DGV, however this database is regularly updating its content to remove newly discovered pathogenic CNVs (and despite the fact that individuals in DGV are reported to be healthy, there is no certainty based on neither clinical examinations nor family history). The CoLaus survey integrates numerous clinical parameters which permits a distinction between “healthy” and “apparently healthy” individuals. Therefore it brings an added value and when a variant is found in any disease cohort (i.e. diabetes, obesity and narcolepsy), investigations can be carried out to check the CNV frequency of a given variant in a healthy cohort, thus helping both clinicians and scientists to decide about the need for further follow up.

Moreover, CoLaus provides more than 150 clinical phenotypes, such as metabolic measurements (triglyceride, HDL, LDL...), blood pressure related measurements, physical characteristic (weight, height, sex...) along with many other diverse phenotype (education levels, smoking status, treatment taken...). Our CNV map will be useful for investigations of these phenotypes and we hope to bring more clarification regarding the link between CNVs and complex disease, in particular about cardio-vascular diseases.

4 Second part: CNV profiling of metastatic melanoma

Melanoma are malignant tumours arising from pigmentation skin cells (melanocytes); they can lead to regional and distant metastases. Melanoma are responsible for more than 48000 deaths per year in US. Many mutations in tumour suppressor genes have been identified in melanoma [49, 50]. As part of collaboration between the Ludwig Institute for Cancer Research, Universities of Lausanne and Geneva and the CHUV, we are performing a comprehensive genomic profiling of melanoma. This project includes 1) karyotype, CGH and SNP arrays to study genomic rearrangements; 2) to study methylation pattern using oligonucleotide arrays; 3) to search mutations in protein-coding genes by sequence capture and sequencing and 4) to identify aberrant splicing by transcriptomic profiling using ultra-high throughput sequencing.

4.1 Material and methods

4.1.1 Metastatic melanoma

Our dataset includes six metastatic melanomas, of which two were taken from the same patient before and after treatment. In addition, a replicate was made with a few passages away from its derived melanoma. For all melanoma, either control EBV or PBL cells are available. Melanoma were selected based on experimental evidence for either low or moderate Cancer-Testis genes expression. We complemented our dataset with two melanocytes cell lines. Approval to use these samples for our project was given by the CHUV ethical committee for clinical research.

4.1.2 CGH computational analysis

All melanoma were analysed on Agilent 244k arrays. Experiments were conducted at the Service of Medical Genetics at the Lausanne Hospital (CHUV). Melanoma were hybridized in competition with their control PBL cells (derived from the same patient). Raw intensities were corrected for background intensities ([51]); within-array normalization and dye bias correction were done using a local weighted polynomial regression, Loess ([52]); and hybridization log2 ratios were segmented using the CBS algorithm [22, 23]). Other normalization scheme were tested and included the popLowess method ([53]) and a normalization scheme based on a ridge regression ([54]).

4.1.3 Illumina SNP array analysis

CNV analysis

We used the methodology, named OverUnder, developed by Attiyeh et al ([55]), which corrects hybridization ratios for aneuploidy and uses a classification table based on both corrected hybridization ratios and ratios of allelic intensities to predict continuous Copy Number values at each SNP.

Defining a map of recurrent rearrangements

From the CN predictions with SNP arrays, we defined regions of recurrent rearrangements as follow: 1) Only consider SNPs which have been observed as amplified in at least 6 out of 7 melanomas; 2) Merge adjacent SNPs into regions and exclude any regions < 50Kb in size and 3) Repeat steps one and two for deletions.

4.1.4 Transcriptomic analysis

Two melanoma transcriptomes have been sequenced on the Roche 454 sequencing technology. Reads were aligned and mapped to known transcripts in our group (Ludwig Institute for Cancer Research) by Dr. Christian Iseli.

4.2 Results

4.2.1 Comparative Genome Hybridization analysis

Karyotype analysis revealed the genome-wide amplification status. An example is given in Figure 7, where this melanoma has 91 chromosomes.

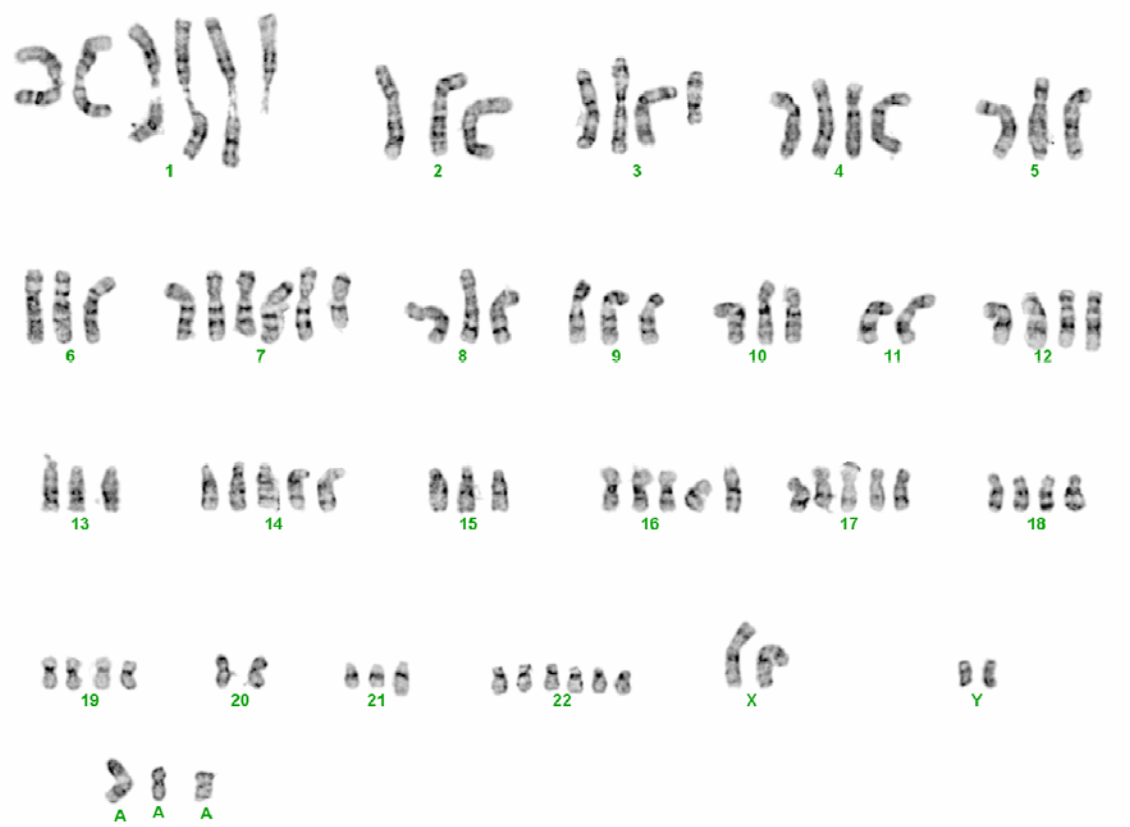

Figure 7 Karyotype for the most amplified melanoma (Me275)

From our CGH analyses, we observed that the hybridization ratios (Figure 8) did not reflect the aneuploidy reported from karyotype analyses (Figure 7). We rationalized that the normalization ([52]), well-established to study diploid genome, was probably not adequate for cancer genome. Thus we applied different normalization scheme developed for aneuploid genome [53, 54]. Nevertheless none of these different normalizations, although they were significantly improving the signal to noise ratio, allowed deciphering the true copy number baseline. We could only infer aberrant amplifications (e.g. >10 copies) that were significantly higher than the already amplified chromosome (e.g. ~6 copies). In the CGH experimental protocol, equimolar DNA concentration for both test and reference genome are used. Thus, we hypothesize that it artificially equalizes an aneuploid genome (i.e. a melanoma) with the reference diploid genome (the matched EBV cell line). As a consequence, the melanoma chromosomal baseline is erroneously observed as diploid and so only local amplification or deletion can be detected.

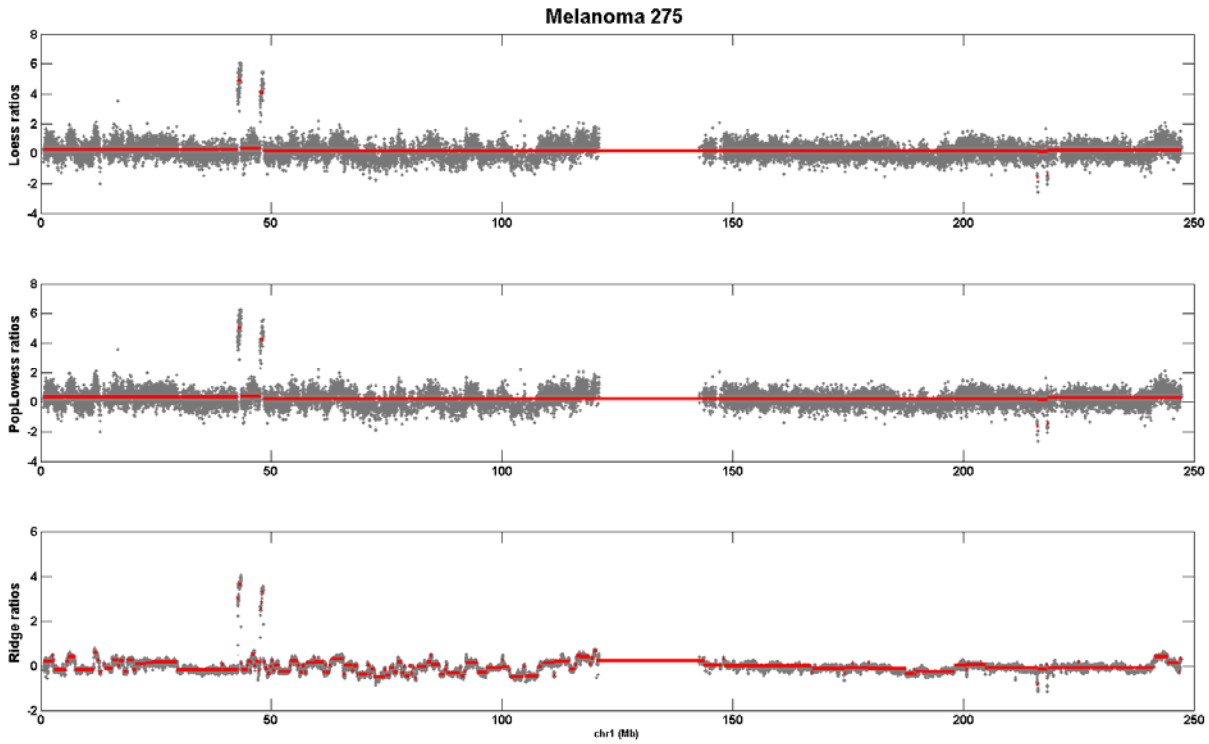

Figure 8 CGH profile for a metastatic melanoma

Gray dots correspond to CGH probe plotted along their physical position on chromosome 1(x axis) and their hybridization log2 ratios (y axis). Each plot corresponds to a different normalization of the log2 ratios. Red segments correspond to significant change-points in the ratios. From 18 karyotyping analyses, we observed that median number of copies of 1p and 1q were respectively 3 and 3.5 (with a range 1-4 and 2-6).

4.2.2 High-resolution SNP array analysis

Although SNP arrays are single-channel experiments (only one genome is hybridized), we observed similar limitation than with the CGH, when analyzing CNV based on the hybridization profile only. However SNP arrays provide information about allele-specific hybridization therefore allelic imbalance, which reflects CNVs, can be detected and used to refine the CNV predictions. Figure 9A illustrates a hemizygous deletion on 19q found in a melanoma but not in its control EBV. Figure 9B, shows locus 7q31-34 known as frequently amplified in melanoma. It is important to note, that in figure 9B, a prediction based only on hybridization log ratios would not have detected such amplification. However from the B Allele Frequency pattern, the OverUnder algorithm was able to detect this event which ranges between 3 and 6 copies.

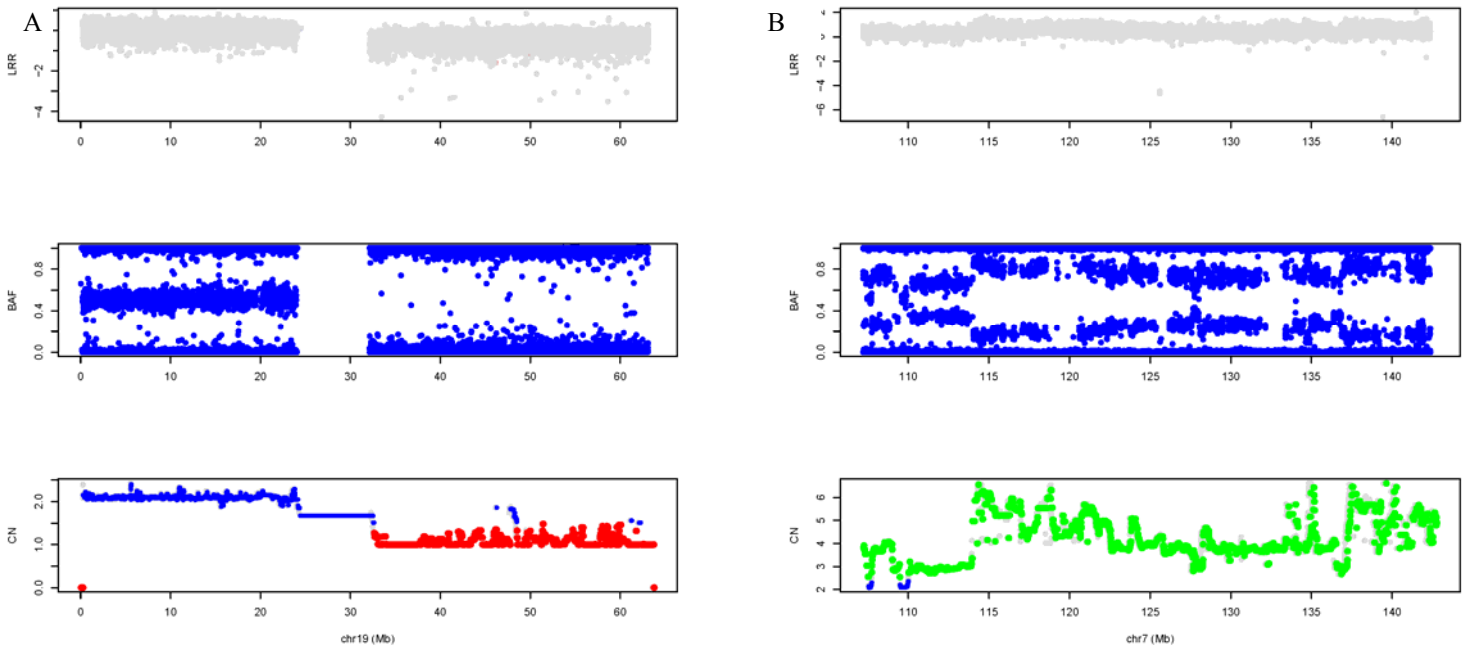

Figure 9 CNV analyses from SNP arrays

Top panel show hybridization log2 ratios, middle panel the B Allele Frequency (BAF), which is the ratios of allele-specific intensities; and the bottom plot is the copy number values inferred by the OverUnder algorithm (with blue for copy neutral SNP, red for deletion and green amplification). The algorithm searches for aberrant BAF, such as loss of heterozygosity supported by lower hybridization ratios due to large hemizygous deletion (as in A) or intermediate BAF values (i.e. 0.3 or 0.6) which reflect allelic copy number imbalances (as in B).

OverUnder was optimized using the replicated melanoma and comparison between EBV and PBL cell lines from the same patient. Optimal results were obtained using a genomic window of 101 SNPs (Figure 10A). Figure 10B illustrates the strong correlation (>0.97) between CNV profiles from the two replicates of a melanoma.

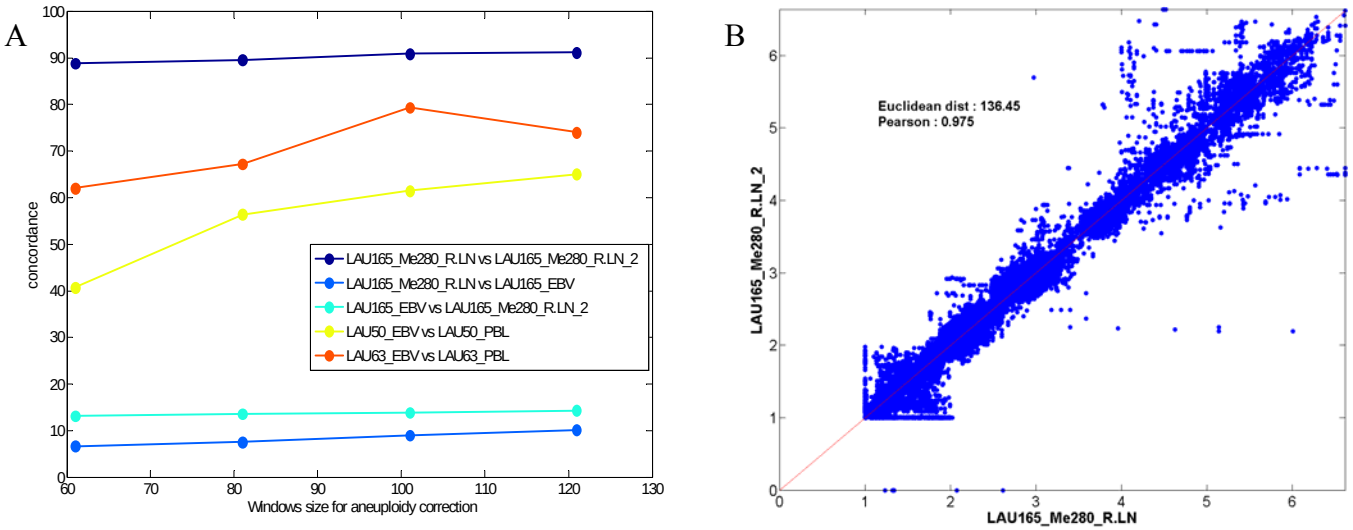

Figure 10 OverUnder parameter optimization and quality control

A) Concordance in CNV prediction as a function of genomic window size. Concordance is computed as the fraction of SNPs having the same copy number values in both samples. Pairwise comparison included the same melanoma replicated twice (dark blue); EBV and PBL cell lines from the same patient (orange and yellow), melanoma and EBV from the same patient (cyan and blue). B) Correlation between the two replicates of Melanoma Me280. Each point is a SNP with its CN value as predicted in both samples. The plot includes more than 1.1M data points, predictions were made using OverUnder with a window size of 101 SNPs.

Replication on another SNP platform

The most amplified melanoma (Me275) was also analysed on the Affymetrix platform. Using the PICNIC algorithm ([56]) dedicated to Affymetrix 6.0 arrays, we predicted the copy number at each SNP and compared these results to the ones from the Illumina array (Figure 11). We found an overall good correlation (Pearson correlation >0.77) between CNV profiles as predicted with the two SNP platforms.

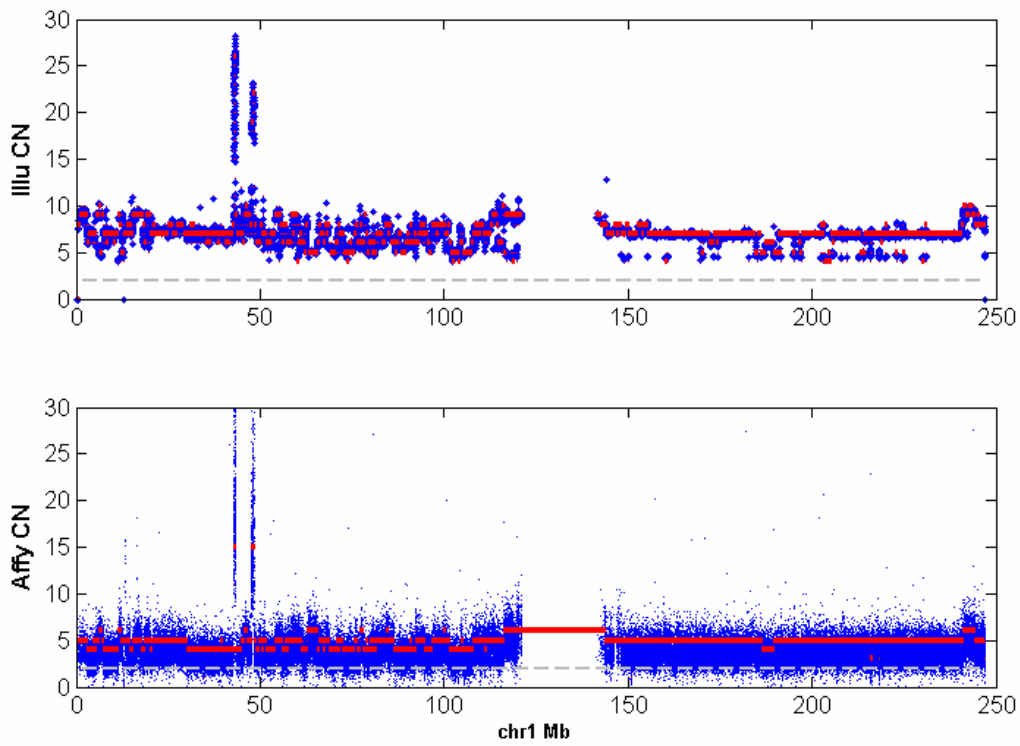

Figure 11 CN profile on Illumina and Affymetrix SNP arrays

Top plot, CN values at each SNP, as predicted on Illumina 1M array with the OverUnder algorithm. Bottom plot, CN values predicted on Affymetrix 6.0 array using the PICNIC algorithm. Dashed lines indicate CN=2 (diploid state). Red segments correspond to genome regions with similar CN values.

4.2.3 Correlation between CNVs and Transcriptome data

We investigated whether there was any correlation between CNVs and transcriptome levels. We first computed the median CN at each Refseq gene that was successfully mapped during transcriptome analysis. Then we discretized these CN values into the following bins: deletion (CN<2), diploid (CN=2), duplication (CN≥3 and < 5), amplification (CN≥5 and < 10) and very high amplification (CN≥10). Next, we checked for any significant difference between discretized CNV bins and transcript levels (tag count). We found significant differences between transcript levels of diploid regions and (highly) amplified regions (Figure 12), implying that highly amplified genes where also highly expressed. With the most amplified melanoma (Me275), significant differences were found for amplifications greater than 10 copies, whereas for the second most amplified sample (Me235), significant differences were already detected from 3 copies.

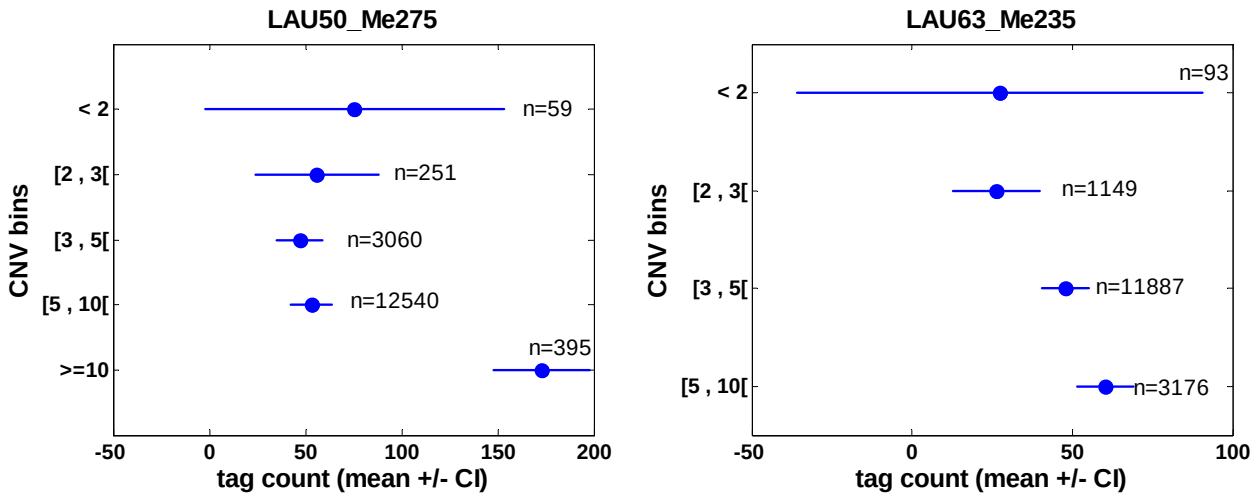

Figure 12 ANOVA analyse between transcript levels and genes discretized CN values

The two plots correspond to a distinct melanoma (name is indicated in the title). Both samples correspond to the two most aneuploid melanomas. Gene CN values were discretized into bins (Y axis) and compared to their transcript tag count (X axis). Blue dots represent the mean tag count at a given CNV bin, the blue bars indicates the mean 95% confidence interval and the numbers indicates the number of data points in each group. Two groups are significantly different when their confidence intervals are strictly non-overlapping.

4.2.4 Recurrent re-arrangements in melanoma samples

Using SNP-based predictions, we derived a map of recurrent rearrangements. (Figure 13).

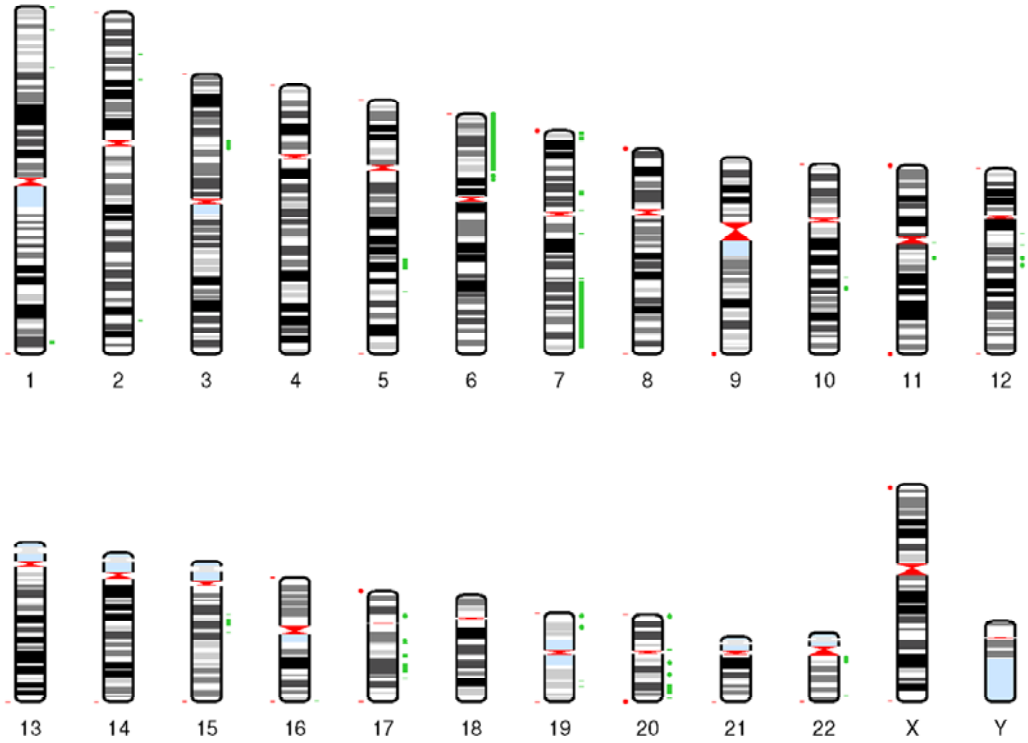

Figure 13 Recurrent rearrangements found in metastatic melanoma

Deletions are indicated in red and amplification in green. Each region was found in at least 6 out of 7 melanomas and region length is greater than 50kb.

This genome-wide map includes 44 recurrent deletions and 96 recurrent amplifications. We then manually compared these regions with the CGH profiles. All telomeric deletions (except the ones on chromosome 10) could not be detected from CGH hybridization ratios. The large amplifications on chromosomes 6, 7, 15, 17, 19, 20 and 22 were confirmed on Agilent, although some were found at a lower frequency. We also checked 56 amplifications (from chromosome 1 to 7) and 23 deletions (chromosome 1 to 11), see table 4.

Regions	**validated**	**Found with lower frequency**	**Miscall***	**neutral**	**total**
Amplifications	38 (62.3%)	6 (9.8%)	4 (6.6%)	13 (21.3%)	61 (100%)
Deletions	7 (33.3%)	9 (42.9%)	4 (19%)	1 (4.8%)	21 (100%)

Table 4 Number of SNP-based recurrent re-arrangements found in Agilent CGH results

* miscall means discrepancies between the copy number prediction, i.e. the same sample would be found as deleted in a platform and amplified in the other platform

The OverUnder is a highly sensitive algorithm: more than 60% of the amplifications were validated, but only 33% of the deletions were confirmed. Also a significant number of events were found neutral or at a lower frequency (i.e. a CNV found in only two samples from CGH arrays as opposed to at least six samples with SNP arrays). Our fine-tuning was done using a replicated melanoma whose genome is mostly deleted (39 chromosomes in average, of which 6 chromosomes had only one arm). Samples with large deletions tend to be easier to analyse compared to samples having higher aneuploidies, which have a significant higher noise. We are currently replicating Me275 on both SNP and CGH arrays, which will enable to fine-tune our analysis pipelines and will greatly benefit to the results.

1027 and 85 genes were overlapping, respectively, the recurrent amplifications and deletions. We downloaded 125 Refseq genes, from the Atlas of Genetics and Cytogenetics in Oncology and Haematology ([57]), that were associated with melanoma. This list included all genes from the Wnt and MAPK pathways. Only 7 of the 1027 recurrently amplified genes were present in this list, and none for the deletions. However all 125 genes were found amplified in at least 2 melanomas (see table 5) and 48 were found deleted in a single melanoma. 8 genes were found amplified in 6 melanomas, one of them was in a region with size lower than 50Kb, and thus was not included in our map of recurrent amplifications. All these demonstrate that our criteria for defining recurrent re-arrangements are very conservative. These criteria were established stringent on purpose, to detect only the most frequent (and largest) amplified or deleted regions. After pipeline optimization using the Me275 replication experiments, we should be able to relax these criteria.

	Count of genes found amplified or deleted in melanoma							
Number of melanoma	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Number of amplified genes	0	0	3	13	34	67	8	0
Number of deleted genes	76	49	0	0	0	0	0	0

Table 5 Count of Melanoma associated genes found amplified or deleted in our samples

4.3 Conclusion and perspectives

4.3.1 Limitations and challenges in analysing highly amplified genomes

Our analysis of metastatic melanoma revealed the limitations of karyotype, CGH and SNP analysis. The karyotype gives a global estimation of the aneuploidy of each chromosome. Such technique is work intensive because of the sample heterogeneity, many replicates are needed. Also not every marker can be attributed with certainty to its corresponding chromosome. As a result, the karyotype under-estimates the aneuploidy. Nevertheless, this technique is crucial to reveal whole-chromosome aneuploidy when both CGH and SNP arrays fail. We observed that CGH ratios can be accurately segmented into regions which reflect different copy number events. However in the presence of complete chromosomal amplification, the hybridization ratios are aberrantly centered on zeros and, so far, we did not find a reliable method to estimate the underlying copy numbers. Using information about allelic imbalances, SNP arrays overcome the limitation of the hybridization ratios and enable copy number estimation.

4.3.2 Replication design

To improve our CNV detection pipelines, we are analysing two replicates of Me275 on both CGH and SNP arrays. These experiments will enable to fine-tune our detection algorithms and to better compare across platforms.

4.3.3 Ultra-high throughput sequencing data

With the two sequenced melanoma transcriptomes, we already established the correlation between high transcript levels and genomic amplification. Melanocyte cell lines are being sequenced then will be used as a baseline of gene expression and will help to detect genes specifically expressed in cancer. We plan to investigate the copy number status of such genes and check whether these have a common function, map to a same pathway or Gene Ontology category. Sequence capture and exome sequencing will also be performed in the coming months. Such data will complement our CNV analysis and provide valuable information regarding somatic point mutations.

5 Third part: Characterization of the CN polymorphisms of cancer-testis genes

CT genes are normally expressed in testis, brain and placenta; but also are aberrantly expressed in many tumour types. Although the function of these genes remains unclear, experimental data indicate that some are involved in the meiotic process ([58]). CTs are mostly located on chromosome X, in highly dynamic regions such as low-complexity regions and segmental duplications. As a consequence the probe coverage as provided by different vendors is low. Thus we developed a pipeline to design probes with reliable hybridization properties, which was used to create the first CT custom chip (an Agilent 4*44K array).

5.1 Material and methods

5.1.1 Location of CT genes

We aligned all known transcript sequences (from EMBL) to the Human genome (build 36) and created clusters of transcripts mapping to a same genomic region. Using unique CT names, as downloaded from Hoffmann et al. ([59]) and complemented with the latest CT candidates (AKAP4, OTOA, RHOXF2, NXF3, IL13RA2 and PRAME), we extracted the relevant transcript clusters and determined their genomic locations. Extending such CT clusters by 1.5Kb upstream and downstream and merging all overlapping ones, produced a final list of 152 CT regions with a total length of ~3.9Mb.

5.1.2 Custom CT-chip design

We queried the Agilent probe database (at this time, it contained more than 24.5M probes), 25759 probes were available for 110 or our CT regions (42 CT regions did not have any coverage). We created all possible 25mers combinations, excluded any of those having more than 21 hits on the genome. Then we derived 60 mers probes, by using remaining adjacent 25mers probes. Finally, we kept a total of 53467 probes matching our CT regions. Using all Agilent "best" (aka "Similarity Filter") probes for chromosome X (totalizing more than 836000 probes), we derived reference distribution for GC content, melting temperature, enthalpy and entropy, and hairpin structures. It has been demonstrated decades ago, with primer design, that GC content and melting temperature are essential parameters. One may also consider enthalpy and entropy which reflect the affinity and stability of the complex DNA-DNA during hybridization ([60]). Hairpin structures, which occur in palindrome sequences and leads to the formation of a loop, need to be accounted as they will affect

the hybridization. We excluded any of our designed probes containing at least a single hairpin structure and selected only probes that were within 1.5 standard deviation from the distribution of Agilent probes for all other metrics (Figure 14). In total 30916 out of 53467 probes passed all the QC steps.

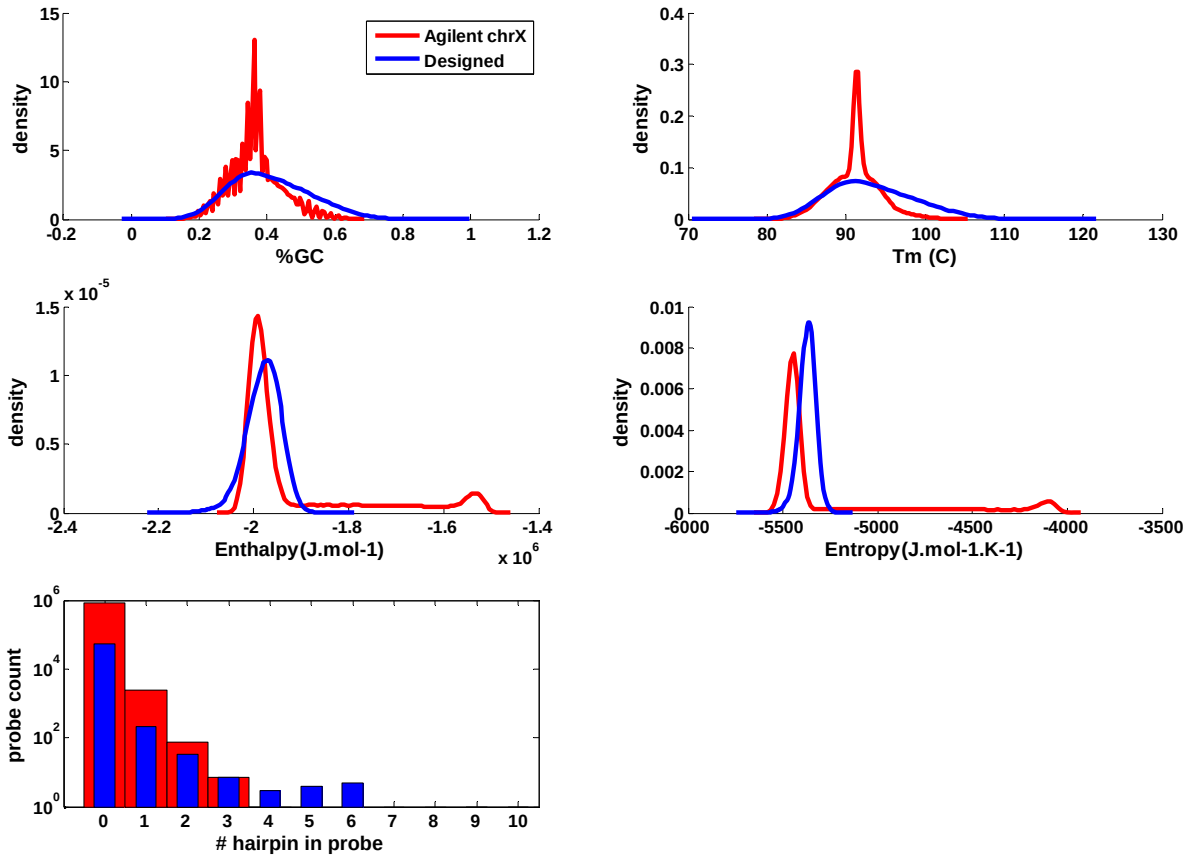

Figure 14 Comparison of Agilent chromosome X probes with all our in-silico designed probes (only controlled for genome match)
A) density distribution of GC content (expressed in %) for Agilent chromosome X best probes (red) and our designed probes (blue), B) melting temperature, C) enthalpy and D) entropy values for probe-DNA hybridization in standard conditions (Sodium concentration 1mol/L , oligo concentration ~0.1mol/L) E) histogram of hairpin structures predicted in probe sequences

Finally we created a 4*44K array design, where we placed the 25759 Agilent "best" probes for CT regions and 17343 of our designed probes (out of the 30916 "good" probes). We selected probes first for regions with lower probe coverage until a minimal probe density was fulfilled (or that no more probes were available for such region) before moving to regions with higher probe density. And we ensured that selected probes were equally spaced thus avoiding having all probes to be located either at the 3' or the 5' end of the region of interest.

All our designed probes were 60 mers, as well as most of Agilent probes (figure 15A and B). Only 10 CT regions were covered with only one probe (Figure 15C) and three CT regions did not have any coverage, as opposed to 42 uncovered regions using only Agilent probes. After investigation, these three regions corresponded to duplicated CT regions containing only pseudo genes; and their respective copy, with annotated Refseq gene, were covered by our custom probes.

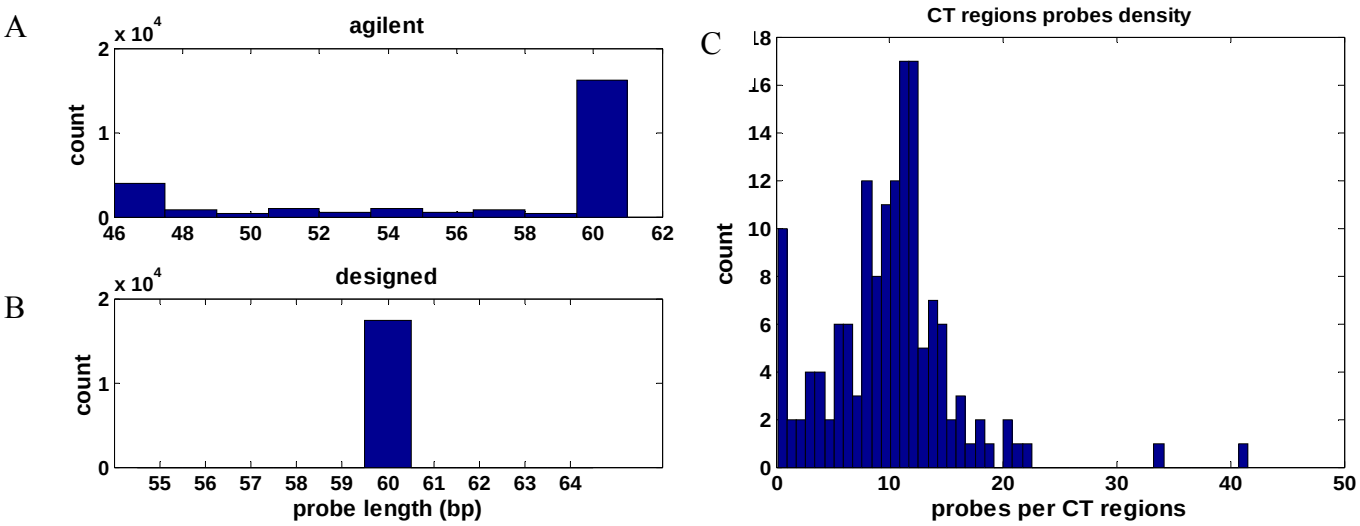

Figure 15 A and B Histogram of probe length, respectively, for selected Agilent probes and our designed probes; C: Histogram of probe density per CT regions

5.1.3 CNV analysis

As done for the melanoma sequencing project, we corrected background intensities with the normexp ([51]) method, normalized the hybridization ratios using popLowess ([53]), then segmented the signal using CBS [22, 23]). Analysis parameters were optimized using a replicated melanoma.

5.2 Results

Using our custom CT chip, we analysed the same seven melanomas from our sequencing project (see part two) with respect to their matched control cell lines. Figure 16 illustrates CGH results for two melanomas at the GAGE (G antigen) cluster. The locus contains a genomic gap between 49.12 and 49.17Mb. Red regions correspond to significant change-points in copy number. Although the GAGE genes 12C,12D,12E,12F (after 49.15Mb) seem to be present in similar copy numbers, the pattern of variation is much more complex for GAGE12J, 13, 2E, 2D…(before 49.1Mb). Interestingly, the two melanomas have a similar variation pattern.

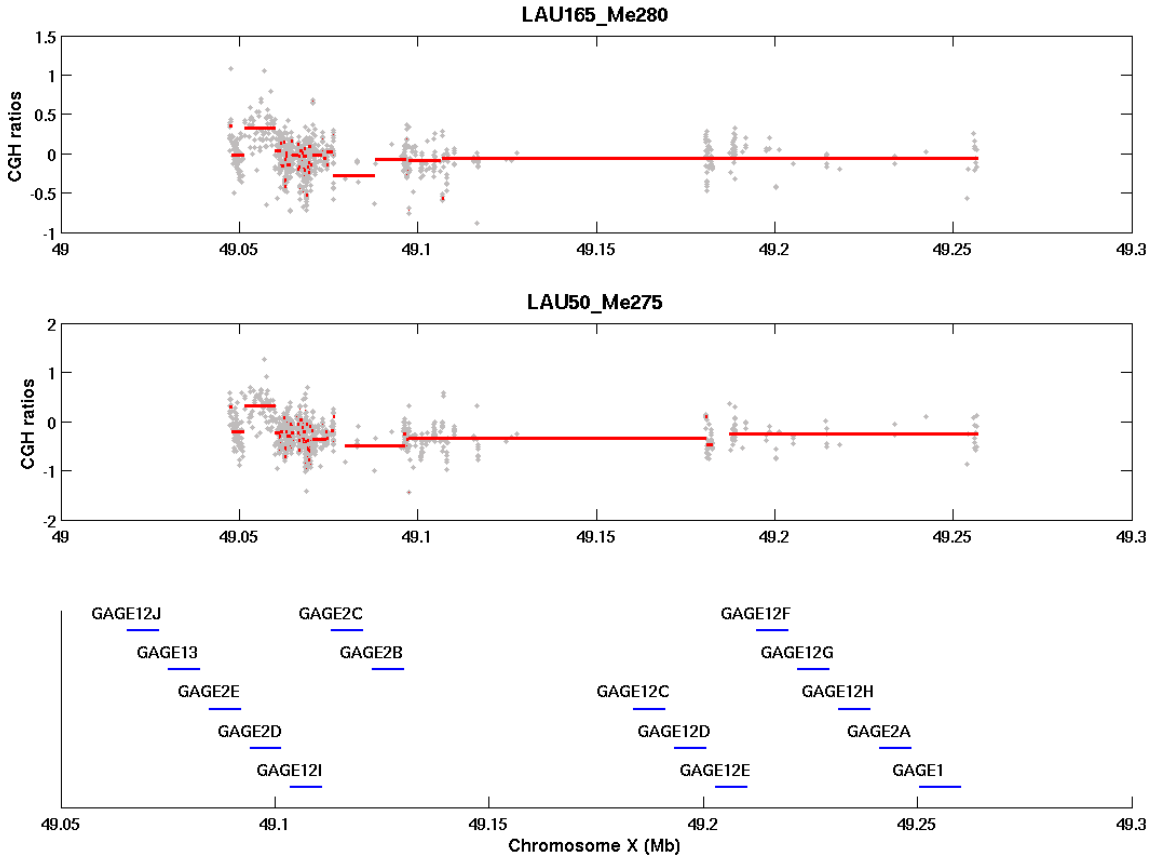

Figure 16 CGH profile for the GAGE cluster in two metastatic melanoma
The top and middle plots show the normalized hybridization ratios from two CGH experiments (two distinct melanomas hybridized with respect to their matched control cell line). Each gray dot is a probe on the array; red regions correspond to the segmentation of the probe-level signal. Bottom plot displays members of the GAGE cluster.

5.3 Conclusion and perspectives

CT analysis is challenging because these genes are located in complex genomic regions (low complexity regions, segmental duplications, near genomic gaps…) and because the samples to analyse are highly amplified which bring array platforms to their limitations. The resolution on micro-arrays was not sufficient for CNV analysis so we designed the first custom CT chip.

Based on knowledge and methods applied to the Melanoma Sequencing Project (Part 2), deriving copy number from the segmented signal on our custom CT array will follow. Moreover since our melanoma samples are being characterized with other techniques (i.e. RNA and exome sequencing), we will be able to complement our analysis and better characterize CT genes.

We initially budgeted for five chips and used only two as a pilot study. So once analysis pipelines are fully optimized, we will be able to extend our study to more tumour samples that are available within the Ludwig Institute for Cancer Research.

6 Future directions

CNV-based genome wide association studies

We have successfully developed methods to normalize, detect copy number variation and combine individual CNV profile into polymorphisms. We have established a catalogue of both common and rare CNVs in a control population. Such resource is of particular interest for follow-up of clinical CNVs in healthy population. Currently our results are used as control for morbid obesity and narcolepsy. Moreover our database of CNVs will be used for genome-wide association with the CoLaus clinical phenotypes, in particular with metabolic measurements and blood pressure related phenotypes.

Melanoma CNV profiling

Data for this project are still being generated. Notably replication experiments will be crucial to fine-tune CNV detection algorithms on both CGH and SNP arrays and to compare the performance from both platforms. Transcriptome sequencing is still in progress, in particular the sequencing of melanocytes, will provides us with a baseline of gene expression in normal cells. Sequence capture and exome sequencing are in progress. Methylation analysis is being performed at the University of Geneva, in Pr. Stylianos Antonarakis lab. We also plan to perform FISH experiments to validate the CNV status of candidate genes. By the end of this project, we anticipate to have a comprehensive profiling at the genomic, transcriptomic and epigenetic levels in melanoma. Regions of interest would be investigated in a larger melanoma dataset, available at the Ludwig Institute for Cancer Research.

Cancer-testis analysis

The custom-chip provides us with coverage to investigate the CNV status of CT genes, which, currently, was not possible with any other array platform. The methodology developed and the knowledge gained with the melanoma project, will be very helpful to infer the copy number from hybridization ratios. Sequence analysis from the Melanoma CNV profiling project, will also be useful to validate CNV prediction from our custom CGH arrays. We anticipate being able to clarify the copy number status of CT genes in tumour genomes. Such results may trigger interest from other groups at the Ludwig Institute for Cancer Research.

We will also investigate the CT status in healthy population using our custom array. DNA is already available at the Ludwig Institute for Cancer Research and with a collaboration involving Dr. Carlo Rivolta at the Department of Medical Genetics from the University of Lausanne. In addition, we plan to retrieve data from whole genome sequencing trace archives and to detect CNVs in CT genes from the coverage depth.

7 Acknowledgments

CoLaus project

We are grateful to Pr. Peter Volleinder (CHUV), Pr. Gérard Waeber (CHUV) and Dr. Vincent Mooser (GSK), principal investigators of the CoLaus project for giving us access to their data; to GlaxoSmithKline for performing all genotyping experiments (both Affymetrix and Illumina) and to the CHUV for collecting clinical phenotypes. We also thank Dr. Zoltan Kutalik, in the group of Pr. Sven Bergmann, for useful discussion and contribution to the implementation of the Gaussian Mixture Model. We thank Dr. Toby Johnson for his precious expertise and suggestions. We are also thankful to Pr. Jacques Beckmann and other members of the Computational Biology Group for feedback all along this project. Part of the computation has been performed at the Vital-IT high performance computing center at the Swiss Institute of Bioinformatics.

Melanoma sequencing project and Cancer-Testis project

We are thankful to the Ludwig Institute for Cancer Research for collecting, preparing and managing all melanoma samples, in particular, we acknowledge Dr. Donata Rimoldi. We also thank Dr. Brian Stevenson and Dr. Christian Iseli for precious support and discussion regarding the project design and computational analyses. We are also grateful to Dr. Danielle Martinet and the Service of Medical Genetic at the CHUV, for realizing karyotype and CGH analysis; to Dr. Carlo Rivolta, Ms. Paola Benaglio and Frontiers in Genetics for SNP array experiments; to the Lausanne Microarray facility for performing the custom CGH experiments.

8 References

1. Iafrate, A.J., et al., *Detection of large-scale variation in the human genome.* Nat Genet, 2004. **36**(9): p. 949-51.
2. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. **7**(2): p. 85-97.
3. Redon, R., et al., *Global variation in copy number in the human genome.* Nature, 2006. **444**(7118): p. 444-54.
4. Sharp, A.J., et al., *Segmental duplications and copy-number variation in the human genome.* Am J Hum Genet, 2005. **77**(1): p. 78-88.
5. Tuzun, E., et al., *Fine-scale structural variation of the human genome.* Nat Genet, 2005. **37**(7): p. 727-32.
6. Freeman, J.L., et al., *Copy number variation: new insights in genome diversity.* Genome Res, 2006. **16**(8): p. 949-61.
7. Jakobsson, M., et al., *Genotype, haplotype and copy-number variation in worldwide human populations.* Nature, 2008. **451**(7181): p. 998-1003.
8. Sebat, J., et al., *Large-scale copy number polymorphism in the human genome.* Science, 2004. **305**(5683): p. 525-8.
9. Perry, G.H., et al., *Hotspots for copy number variation in chimpanzees and humans.* Proc Natl Acad Sci U S A, 2006. **103**(21): p. 8006-11.
10. Perry, G.H., et al., *Copy number variation and evolution in humans and chimpanzees.* Genome Res, 2008. **18**(11): p. 1698-710.
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, 2008. **17**(8): p. 1127-36.
12. Henrichsen, C.N., et al., *Segmental copy number variation shapes tissue transcriptomes.* Nat Genet, 2009. **41**(4): p. 424-9.
13. Lupski, J.R. and P. Stankiewicz, *Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes.* PLoS Genet, 2005. **1**(6): p. e49.
14. de Cid, R., et al., *Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis.* Nat Genet, 2009. **41**(2): p. 211-5.
15. Beckmann, J.S., X. Estivill, and S.E. Antonarakis, *Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability.* Nat Rev Genet, 2007. **8**(8): p. 639-46.
16. Cowell, J.K. and L. Hawthorn, *The application of microarray technology to the analysis of the cancer genome.* Curr Mol Med, 2007. **7**(1): p. 103-20.
17. Kallioniemi, A., et al., *Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.* Science, 1992. **258**(5083): p. 818-21.
18. Kallioniemi, A., *CGH microarrays and cancer.* Curr Opin Biotechnol, 2008. **19**(1): p. 36-40.
19. Pinkel, D. and D.G. Albertson, *Array comparative genomic hybridization and its applications in cancer.* Nat Genet, 2005. **37 Suppl**: p. S11-7.
20. Stranger, B.E., et al., *Relative impact of nucleotide and copy number variation on gene expression phenotypes.* Science, 2007. **315**(5813): p. 848-853.
21. Fiegler, H., et al., *Accurate and reliable high-throughput detection of copy number variation in the human genome.* Genome Research, 2006. **16**(12): p. 1566-1574.
22. Olshen, A.B., et al., *Circular binary segmentation for the analysis of array-based DNA copy number data.* Biostatistics, 2004. **5**(4): p. 557-72.
23. Venkatraman, E.S. and A.B. Olshen, *A faster circular binary segmentation algorithm for the analysis of array CGH data.* Bioinformatics, 2007. **23**(6): p. 657-63.
24. Pique-Regi, R., et al., *Sparse representation and Bayesian detection of genome copy number alterations from microarray data.* Bioinformatics, 2008. **24**(3): p. 309-18.

25. Nannya, Y., et al., *A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays.* Cancer Res, 2005. **65**(14): p. 6071-9.
26. Huang, J., et al., *Whole genome DNA copy number changes identified by high density oligonucleotide arrays.* Hum Genomics, 2004. **1**(4): p. 287-99.
27. Komura, D., et al., *Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays.* Genome Res, 2006. **16**(12): p. 1575-84.
28. Korn, J.M., et al., *Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs.* Nat Genet, 2008. **40**(10): p. 1253-60.
29. McCarroll, S.A., et al., *Integrated detection and population-genetic analysis of SNPs and copy number variation.* Nat Genet, 2008. **40**(10): p. 1166-74.
30. Pique-Regi, R., A. Ortega, and S. Asgharzadeh, *Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA.* Bioinformatics, 2009. **25**(10): p. 1223-30.
31. Itsara, A., et al., *Population analysis of large copy number variants and hotspots of human genetic disease.* Am J Hum Genet, 2009. **84**(2): p. 148-61.
32. Loos, R.J., et al., *Common variants near MC4R are associated with fat mass, weight and risk of obesity.* Nat Genet, 2008. **40**(6): p. 768-75.
33. Newton-Cheh, C., et al., *Genome-wide association study identifies eight loci associated with blood pressure.* Nat Genet, 2009.
34. Prokopenko, I., et al., *Variants in MTNR1B influence fasting glucose levels.* Nat Genet, 2009. **41**(1): p. 77-81.
35. Weedon, M.N., et al., *Genome-wide association analysis identifies 20 loci that influence adult height.* Nat Genet, 2008. **40**(5): p. 575-83.
36. Willer, C.J., et al., *Six new loci associated with body mass index highlight a neuronal influence on body weight regulation.* Nat Genet, 2009. **41**(1): p. 25-34.
37. Barnes, C., et al., *A robust statistical method for case-control association testing with copy number variation.* Nat Genet, 2008. **40**(10): p. 1245-52.
38. Vollenweider, P., et al., *[Health examination survey of the Lausanne population: first results of the CoLaus study].* Rev Med Suisse, 2006. **2**(86): p. 2528-30, 2532-3.
39. Affymetrix, *www.affymetrix.com.*
40. Kolz, M., et al., *Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations.* PLoS Genet, 2009. **5**(6): p. e1000504.
41. Yuan, X., et al., *Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes.* Am J Hum Genet, 2008. **83**(4): p. 520-8.
42. Sandhu, M.S., et al., *LDL-cholesterol concentrations: a genome-wide association study.* Lancet, 2008. **371**(9611): p. 483-91.
43. Conrad, D.F., et al., *Origins and functional impact of copy number variation in the human genome.* Nature, 2009.
44. Li, C., *Automating dChip: toward reproducible sharing of microarray data analysis.* BMC Bioinformatics, 2008. **9**: p. 231.
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, 2008. **745**.
46. Bengtsson, H., et al., *Estimation and assessment of raw copy numbers at the single locus level.* Bioinformatics, 2008. **24**(6): p. 759-67.
47. Bengtsson, H., et al., *A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods.* Bioinformatics, 2009. **25**(7): p. 861-7.
48. Illumina, *www.illumina.com.*

49. Goldberg, E.K., et al., *Localization of multiple melanoma tumor-suppressor genes on chromosome 11 by use of homozygosity mapping-of-deletions analysis.* Am J Hum Genet, 2000. **67**(2): p. 417-31.
50. Miele, M.E., et al., *A human melanoma metastasis-suppressor locus maps to 6q16.3-q23.* Int J Cancer, 2000. **86**(4): p. 524-8.
51. Ritchie, M.E., et al., *A comparison of background correction methods for two-colour microarrays.* Bioinformatics, 2007. **23**(20): p. 2700-7.
52. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
53. Staaf, J., et al., *Normalization of array-CGH data: influence of copy number imbalances.* BMC Genomics, 2007. **8**: p. 382.
54. Chen, H.I., et al., *A probe-density-based analysis method for array CGH data: simulation, normalization and centralization.* Bioinformatics, 2008. **24**(16): p. 1749-56.
55. 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, 2009. **19**(2): p. 276-83.
56. Greenman, C.D., et al., *PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data.* Biostatistics, 2009.
57. Huret, J.L., P. Dessen, and A. Bernheim, *Atlas of Genetics and Cytogenetics in Oncology and Haematology, year 2003.* Nucleic Acids Res, 2003. **31**(1): p. 272-4.
58. Simpson, A.J., et al., *Cancer/testis antigens, gametogenesis and cancer.* Nat Rev Cancer, 2005. **5**(8): p. 615-25.
59. Hofmann, O., et al., *Genome-wide analysis of cancer/testis gene expression.* Proc Natl Acad Sci U S A, 2008. **105**(51): p. 20422-7.
60. Le Novere, N., *MELTING, computing the melting temperature of nucleic acid duplex.* Bioinformatics, 2001. **17**(12): p. 1226-7.
61. *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
62. *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.

9 Annexes

9.1 Detection and correction of batch effects in CoLaus

We used the Affymetrix GeneChip Genotyping Analysis Software [39] 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 [39]). We first normalized the CoLaus individuals versus 30 unrelated CEU Hapmap ([61, 62]) individuals.

By doing a Principal Component Analysis on the CN status of SNPs (as predicted by CNAT.allelic, details are below) across CoLaus individuals, we found that individuals clustered into 4 distinct groups, which corresponded to four independent genotyping centers (see Figure I). To correct this batch effect, we performed normalization within each center and used an increasing number of randomly chosen samples (with equal proportion of males and females).

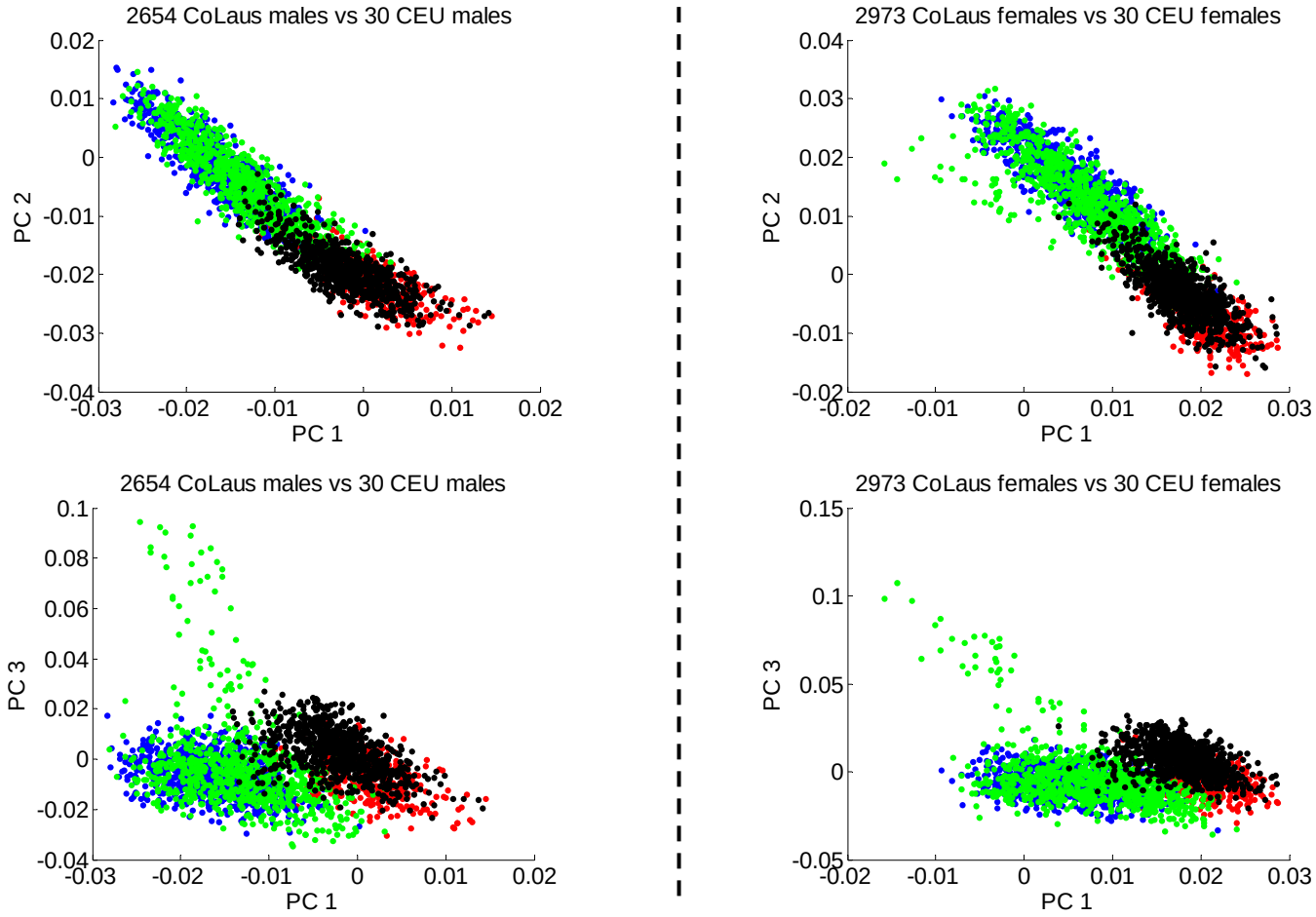

Figure I: principal component analysis on the CNV profile of the CoLaus individuals. Using normalization with 30 unrelated CEU individuals leads to differences between CoLaus males and females. Also differences between individuals genotyped in different centers are observed (the blue and green dots clearly separate from the red and black dots).

By running the normalization twice on the same individuals but with two 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 pairs of individuals (Figure II). We tested normalization with 30, 120, 200 and 280 references and observed that the normalization improves significantly with the number of references. Using 280 references is significantly better than using only 30 references. Using even more references (i.e. 300, data not shown) could still improve a little the normalization but we decided to use 280 references for computational reasons.

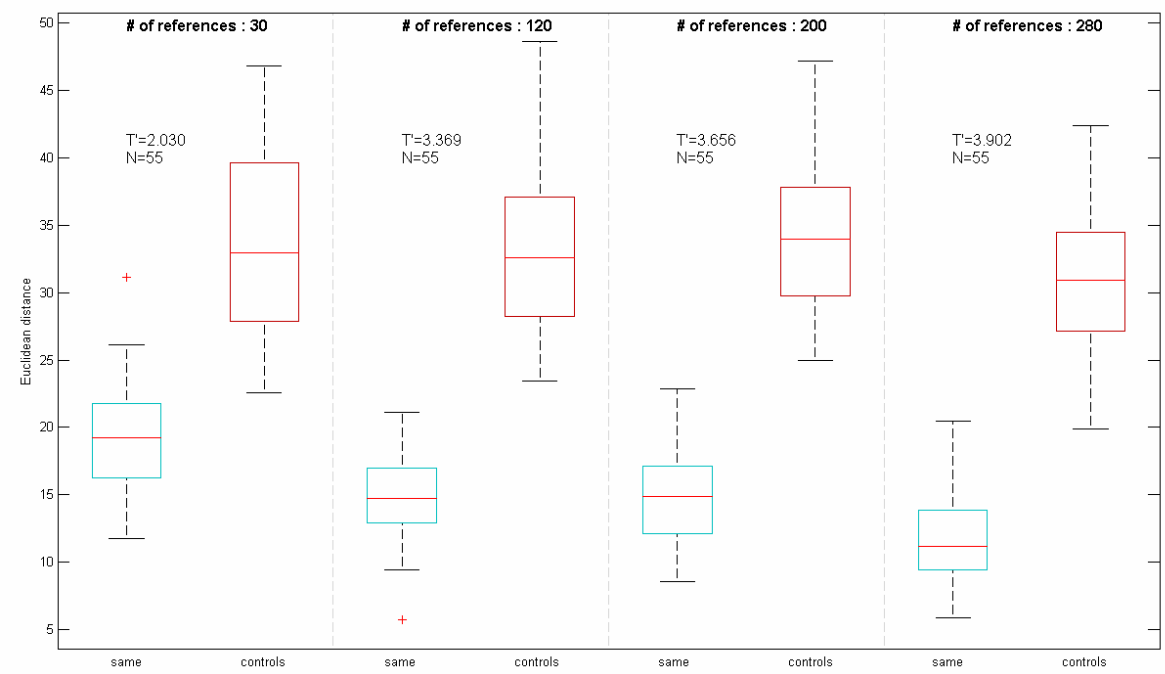

Figure II Improvement of the normalization as a function of the reference panel size
In cyan are shown the distances (n=55) between the CNV profiles as predicted by two independent reference panel (having the same size) for a same individual. In red are the distances between unrelated individuals predicted 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:

$$T' = \frac{|\,mean(same) - mean(controls)\,|}{\sqrt{std(same)^2 + std(controls)^2}}$$

where mean is the geometric mean and std the standard deviation