UNIL | Université de Lausanne

SIB

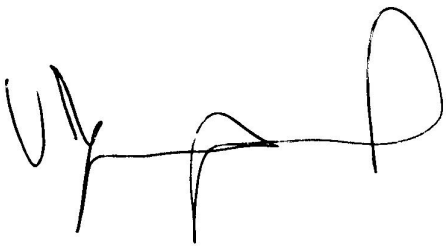
LUDWIG INSTITUTE FOR CANCER RESEARCH

Armand Valsesia

PhD student – University of Lausanne

2nd year report

Lu et approuvé le 15 juillet 2009



Detection and assessment of copy number variants in control population and disease cohort

Summary

Structural variation ranges from single nucleotide polymorphisms (SNP) to large genomic rearrangements [1, 2]. Recent and numerous SNP-based genome wide association studies have provided new insights to complex and metabolic diseases [3-7]. Nevertheless, in a very large cohort, one can explain at best 12% of the genetic variance for quantitative trait such as height [6]. 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 [8]. There are many reasons; one is methods for CNV detection are not as robust as methods for SNP genotyping. CNV analysis is very sensitive to technical effects, such as experimental batch effects. Also the methodology to associate CNVs to phenotype is still a topic under active research. Moreover analysing rare CNVs is a challenging task, as their number can be very large and it is hard to distinguish them from false positives.

This report summarizes my work on several projects, where I developed methods to detect CNVs, to combine these CNVs at the population level and to assess the different prediction methods. The main projects I have been working on are the following:

1) The CoLaus cohort (Cohorte Lausannoise) a 6000 individual population from apparently healthy individuals genotyped on Affymetrix 500K arrays and for which more than 150 phenotypic measurements are available. Using this large dataset, we developed a new detection method based on Mixture Gaussian models and we compare it to other existing detection methods. Previously we developed a simple but efficient method to integrate individual CNV predictions and have now extended this in a more sophisticated approach.

2) The Melanoma Sequencing Project, a collection of 7 metastatic melanomas with matched control cell lines from the same patients. These samples are being assessed by comparative genome hybridization (CGH) and karyotype analysis; SNP and methylation arrays; and ultra-high throughput sequencing at genomic and transcript levels. We have already established analysis pipelines for both CGH and SNP arrays and are investigating a strategy to define regions of recurrent genomic aberrations.

3) Studying the CN status of cancer-testis genes. Since these genes are insufficiently covered by probes in commercially available arrays, we developed a pipeline to design probes and created the first custom CT-array.

Table of contents

CNV analysis in very large cohorts

CoLaus (Cohorte Lausannoise) is a population based health examination survey started in 2003 to study the genetics of hypertension and cardiovascular disease [9]. More than 6000 individuals (35-75 years old) from the Lausanne area participate in the study. Over 159 phenotypic measurements have been collected by the CHUV; in addition, genotyping has been carried out on Affymetrix 500K SNP chips[10]. The CoLaus dataset provides a unique opportunity to develop and assess computational method for CNV detection, then to develop methods for association and apply them on the list of phenotypes.

Detecting Copy Number Variants

Analysis pipelines based on the CNAT package [11] have been previously used [12] to predict CNVs for each CoLaus individuals. In the mean time, other methods have been released in the CNV community [13, 14]. To compare with some of these recent development, we completely renormalized the CoLaus dataset with the Aroma framework [15]. Although this framework requires enormous computational resources (both in term of disc usage and CPU), it was of interest to compare with our previous results because the authors demonstrated the Aroma normalization to perform slightly better than the CNAT package. Subsequently to this renormalization, we used the Circular Binary Segmentation algorithm [16, 17], a state-of-art segmentation algorithm.
In addition, we developed a new CNV detection algorithm based on Gaussian Mixture Model. This new method is unique in the sense it performs a probabilistic calling. Each probed SNP is compared to the signal from the whole CoLaus population and is attributed probabilities for being deleted (CN = 0 or 1), copy neutral (CN=2), simple copy (CN=3) or multiple copies (CN>3). The underlying copy number can be defined by using the dosage value of all these probabilities. 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*0.1 + 2*0.9 + 3*0.85 + 4*0.05).
In total, we have four set of predictions for the whole CoLaus: two from the CNAT implementations (CNAT_Allelic and CNAT_Total) and two using CN ratios normalized by the Aroma package: CBS and our new CNV calling method.

Deriving Copy Number Polymorphism from Copy Number Variants

We previously addressed the problem to integrate variants from many individuals into a consensus CNV map [12]. Currently, scientists merge CNVs based on ad-hoc thresholds on the CNV reciprocal overlap. We have developed a simple merging scheme which allowed to reduce greatly the data complexity and did not require any prior knowledge. But such merging scheme had the disadvantages of producing too many small regions especially for noisy datasets and/or frequent polymorphic loci.
We solved this issue, by developing a novel merging algorithm which 1) partition the genome into smaller regions, whose boundaries are a long stretch of SNPs in the diploid state; 2) then for each of these regions, perform a principal component analysis of SNP data across individuals [Figure 1]. Only components that explain most of the variance are used to cluster SNPs into CNV regions. From ~490k autosomal SNPs analysed with a relaxed algorithm (CNAT_Allelic), the simple merging scheme was creating ~39k regions whereas the novel merging algorithm creates now ~20k regions.

Assessing different prediction methods

Overlap with the Database of Genomic Variants

The Database of Genomic Variants (DGV) [1, 18] 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 completed this dataset by CNVs from European ancestry individuals as found by Itsara et al. [19] Then we computed overlap between CNVs of each prediction methods versus this reference CNV set. As a control, we shuffled positions of our predictions and recomputed the overlap between these ñrandom CNVsò and the reference set. Repeating this a thousand times, allows us to check the performance of the different prediction methods.
We demonstrated that our novel prediction method had the highest number of predictions and was significantly enriched for known CNVs. Moreover this novel method had a number of new CNVs that was significantly less than one can expect if these predictions were random [Figure 2A]. On the contrary, the CNAT_Allelic method had an enrichment of know CNVs but its fraction of novel CNVs was not different from the expectation of a random classifier.

Replication on an independent array platform

312 CoLaus individuals were assayed on the Illumina SNP platform (550K and 1M chip) [20], after QC, we only kept 250 arrays having a call rate > 99.9% and satisfying the Illumina standard QC metrics. We performed CNV calling using our Gaussian Mixture model, applied our PCA-merging approach and only kept CNV regions present in at least 5 individuals thus creating a high confidence CNV dataset.
We used this dataset to check what fraction of novel CNV regions, with respect to DGV, would replicate. Our Gaussian Mixture calling method was the only method to be significantly enriched for CNVs that replicate on the Illumina arrays [Figure 2B]. Strikingly, CNAT_Allelic had significantly less CNVs that replicate than by chance, implying that a random predictor would perform much better. This situation was the same whether looking at copy number polymorphisms (regions with a frequency greater than 1%) or looking at rare CNV regions (frequency < 1%) [Data not shown]. There was no statistical difference between prediction from CNAT_Total and CBS with respect to their control sets.

Using CNV profile to predict relatedness

By chance, family-related individuals were included in the CoLaus study. These were identified based on the similarity of their genotype and having plausible difference between their birth date.
In total, there are 5 pairs of individuals that were inadvertently sampled twice and 157 pairs with sibling or offspring relationship. Using this information, we could check whether predicting relationship between individuals using their CNV profile was possible. By computing Euclidean distance between pairs of related and unrelated individuals, then thresholding these distances and checking with the ñtruthò, we built ROC curves (Receiver Operating Curves) for each CNV prediction methods and each merging approach. All the prediction methods had significant prediction power (Area Under the Curve ~0.7); though no distinction was observed between the merging approaches [Figure 3].

Conclusion

We have developed an innovative merging method which does not require ad-hoc thresholds, greatly reduces the data complexity while keeping relevant information explaining the statistical data variation. The major benefit compared to ñclassical merge-by-overlap approachò is that much finer information in highly rearranged regions is kept. The merge is fully driven by components explaining most of the variance, therefore outlier CNVs will not cause over-merging. Also it aligns each region across the whole population thus making it easier for subsequent association tests. We devised new testing strategies in the context of very large cohorts and believe such knowledge would benefit to other research groups.
We have also implemented a novel CNV prediction method which differs from other Gaussian mixture methods by its probabilistic output. This novel method provides more predictions and performs better than the 3 other tested methods. Based on predictions from our method, we already started doing association with the CoLaus clinical phenotypes and we are currently exploring which association strategy would perform best.

Melanoma Sequencing Project

Melanomas are malignant tumours arising from pigmentation skin cells (melanocytes); they can lead to regional and distant metastasis. Melanomas are responsible for more than 48000 deaths per year in US. Many mutations in tumour suppressor genes have been identified in melanomas [21, 22]. As part of a collaboration between the Ludwig Institute for Cancer Research, universities of Lausanne and Geneva and the CHUV, we plan to perform a comprehensive genomic profiling of melanomas. 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. Samples available for this study are 6 metastatic melanomas and their matched control cells (either a PBLs or EBV cell lines derived from blood). We have an additional metastatic melanoma, taken after treatment of one of these 6 patients and we have 2 normal melanocytes.

Comparative Genome Hybridization analysis

All melanomas were analysed by Agilent 244k arrays [23], using its matched PBL or EBV cell line as a reference. Karyotype analysis revealed the genome-wide amplification status. We observed discrepancies with the hybridization ratios from CGH and applied different normalization scheme developed for aneuploid genome [24-26]. 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 (i.e. 10 copies) that were significantly higher than the already amplified chromosome (i.e. 6 copies). A hypothesis would be that when such amplications are occurring genome-wide, the CGH hybridization saturates. The amount of tumour DNA is so high that it hybridizes and outcompetes with the reference diploid genome.

High-resolution SNP array analysis

Our observations on CGH arrays were also seen by Attiyeh et al. [27] In this recent study, the authors developed a classification algorithm that considers both hybridization ratios and allelic imbalance from SNP arrays. According to their results, the authors demonstrate a major improvement in copy number classification when other methods relying only on hybridization ratios are massively underpowered. We applied the same algorithm to our SNP arrays and observed good data reproducibility between replicates and a better concordance with the karyotype compared to the CGH results.

Defining regions of recurrent rearrangements

To find genomic regions with recurrent rearrangements, we calculated for each SNP of each melanoma, a score Tôdefined as:

$$T'_{stat(i)} = \frac{CN_i - median(CN_{chr})}{MAD(CN_{chr})}$$

Where MAD is the median absolute deviation, a robust estimator of the dispersion around the median; CN_i is the copy number of a SNP_i, median(CNchr) the median of the copy number of all SNPs on this chromosome.
Such score allows extracting SNPs having a copy number significantly different from the chromosomal baseline of a given melanoma [Figure 4]. Using a threshold for statistical significance, then requiring the same SNPs to be seen significant in a minimal number of melanomas, allows defining regions of recurrent rearrangements. We are currently investigating what thresholds are the most adequate to find regions of recurrent rearrangements. Then we will analyse the affected genes, in particular checking prior knowledge, doing pathway and Gene Ontology analysis; and correlation with the preliminary results from the transcriptome sequencing data.

Analysing the copy number status of cancer-testis genes

As part of the Melanoma Sequencing Project, we plan to investigate the copy number status of cancer-testis genes. Cancer-testis (CT) genes are normally expressed in testis, brain and placenta. They 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.
CT 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 pretty low.
In a first step, we set up an in-house pipeline to design probes that would cover the CT regions. Emphasis was put on designing 60 mer probes with good hybridization properties (such as Tm, Enthalpy and Entropy, GC content etcé) while avoiding the cross-hybridization inherent to these complex genomic regions. This allowed us to design a CT custom Agilent chip (4*44K probes) [23] and to perform the experiments with all our 7 melanomas. CNV analysis might be challenging given the CGH saturation observed previously in highly amplified samples. However, from preliminary normalizations, it seems the experiments have worked and we hope being able to identify the high amplification events.

Conclusion

Data are still being generated for this project, only SNP and CGH experiments have been completed. Information from transcriptome sequencing will be important for correlation with the copy number aberrations. Exome sequencing will be very useful to check the copy number and to identify both germ-line and somatic point mutations. Integration of all the genomic, transcriptomic and epigenetic data will definitely be an interesting challenge, improve our knowledge about melanomas and should provide insights for molecular therapies.

Future directions

We demonstrated that our new CNV detection method was performing much better than other existing ones. Our novel merging algorithm is also of interest for integrating CNVs in control population or in disease cohort. We are optimistic that these methodologies and our results will prove useful to other projects and other groups.
Using our CNV predictions, we already started doing association with the CoLaus clinical phenotypes. We will try devising new association methods that could take advantage of our CNV probabilistic calling framework.
In parallel, more data are being generated for the Melanoma Sequencing Project. These data will be used to continue the genomic profiling of melanomas. We will also work on establishing strategies to integrate the genomic, transcriptomic and epigenetic data generated by the project. Then we plan to identify genes that are specific to the melanoma aberrations and which might be driver cancer-mutation. Additionally, we will analyse the custom CGH experiments to derive the copy number status of cancer-testis genes and check how these CN status correlate with gene expression levels as obtained from transcriptome sequencing.

Annexes

Figure 1: A) principal component analysis (PCA) on a local SNP window (chr3:74.5-76.5Mb) across CoLaus individual B) red regions = CNV regions from merging adjacent SNPs having the same CNV profile in the CoLaus population, blue = CNV regions obtained from clustering the main PCA components. Y axis represents frequency of CNV in the CoLaus population (N=5612)

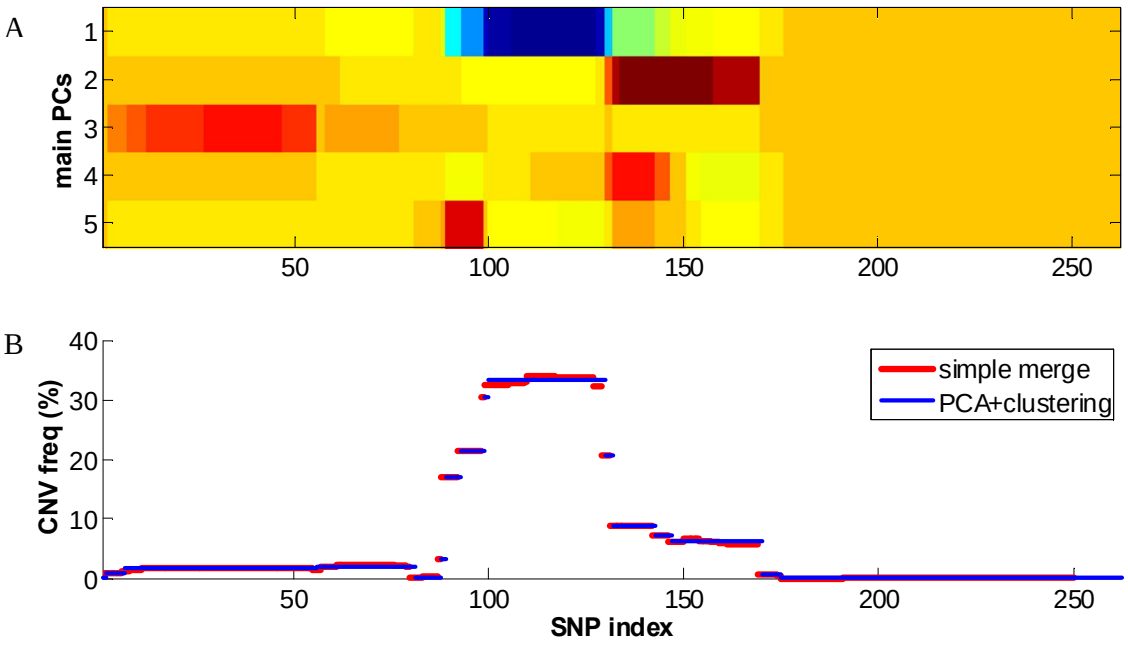
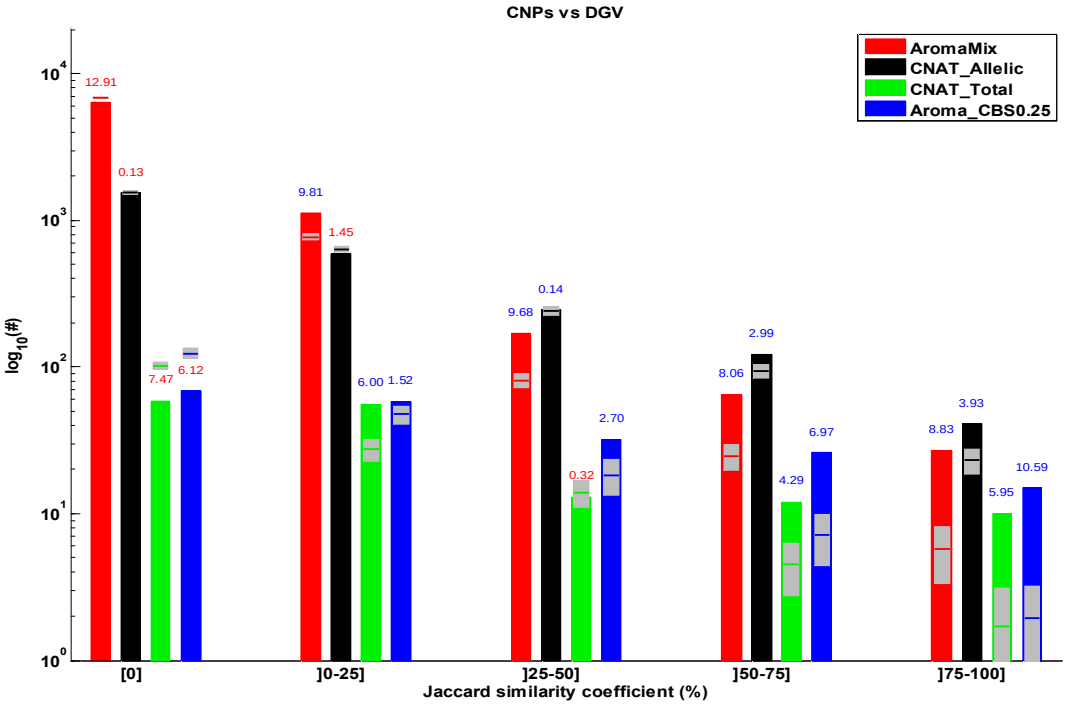


Figure 2: A) Overlap between CoLaus Copy Number Polymorphisms versus DGV. Bins corresponds to reciprocal overlap, count is in log10 scale. B) Overlap between CNPs, having less than 50% reciprocal overlap with DGV, versus the CoLaus CNVs detected on Illumina. Gray area correspond to standard deviation around the mean overlap value for 1000 sets of control CNVs. Numbers above each bar, correspond to T statistic between observed and control overlap (T values equals to 1.98; 2.58 or 3 correspond respectively to statistical significance at 5%, 1% or 0.1%). Red (blue) numbers indicate observed overlap is more (less) than mean control overlap.

A



B

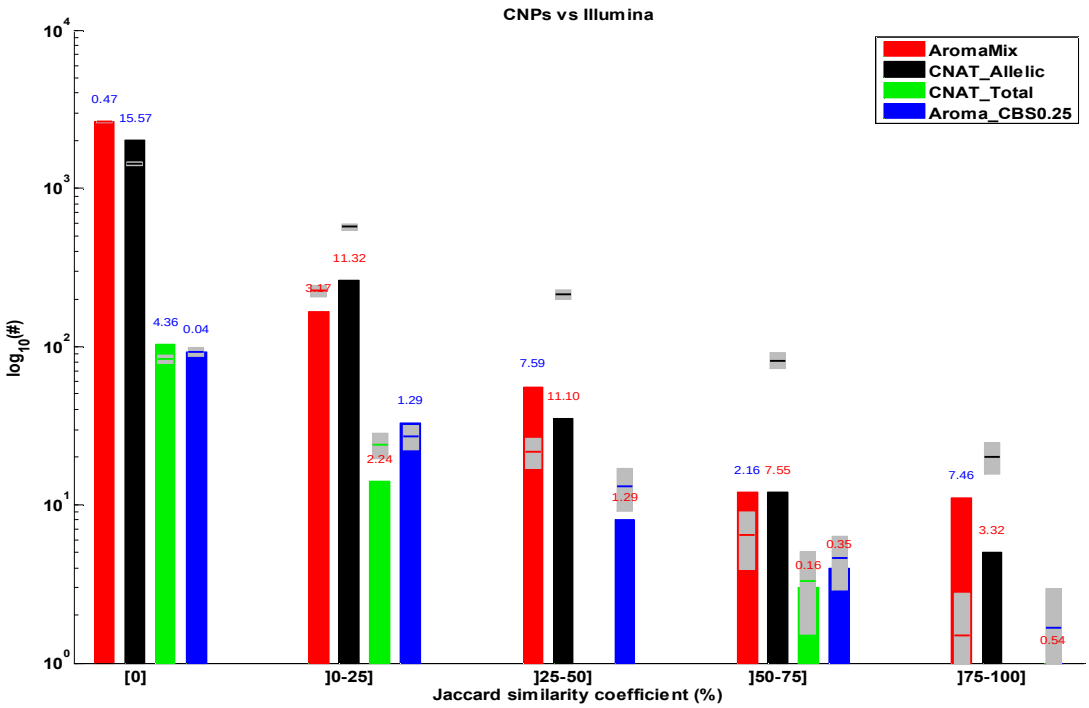


Figure 3: ROC curves when predicting individual relatedness using CNV profile as predicted by our Mixture Gaussian model. No difference was found whether using regions from simple or PCA merges. Both merging approaches have prediction power (Area Under the Curve (AUC) > 0.5)

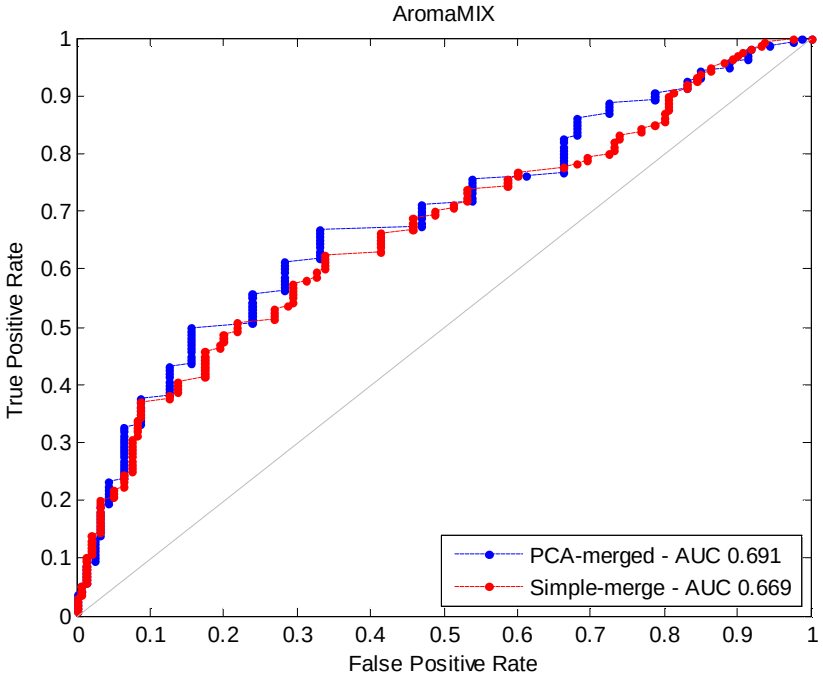
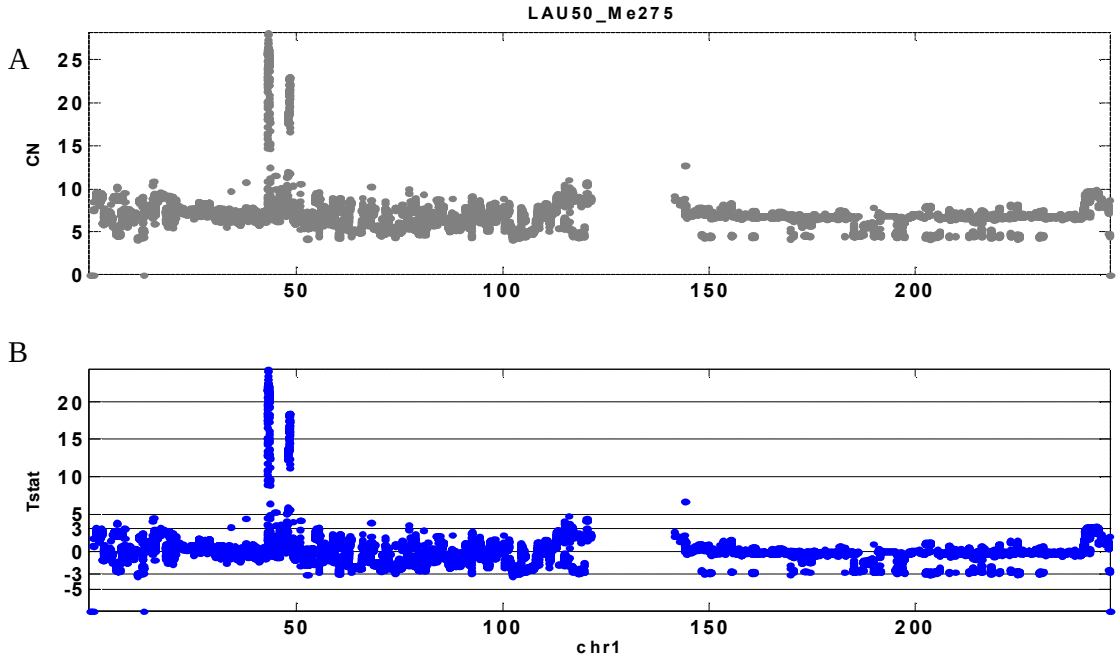


Figure 4: A) Copy Number (CN) from SNP arrays for a highly amplified melanoma. Y axis is the copy number and X axis physical position on chromosome 1 B) Tô statistic (Y axis) derived from CN. High/low score (.i.e. > 5 or < -5) indicates statistical difference with respect to the chromosomal baseline.



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. 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.
4. Newton-Cheh, C., et al., *Genome-wide association study identifies eight loci associated with blood pressure.* Nat Genet, 2009.
5. Prokopenko, I., et al., *Variants in MTNR1B influence fasting glucose levels.* Nat Genet, 2009. **41**(1): p. 77-81.
6. Weedon, M.N., et al., *Genome-wide association analysis identifies 20 loci that influence adult height.* Nat Genet, 2008. **40**(5): p. 575-83.
7. 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.
8. 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.
9. 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.
10. Affymetrix, *www.affymetrix.com.*
11. 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.
12. Valsesia, A., *1st Year PhD report.* 2008.
13. Bengtsson, H., et al., *Estimation and assessment of raw copy numbers at the single locus level.* Bioinformatics, 2008. **24**(6): p. 759-67.
14. 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.
15. 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**
16. 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.
17. 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.
18. DGV, *Database of Genomic Variants* http://projects.tcag.ca/variation/
19. 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.
20. Illumina, *www.illumina.com.*
21. 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.
22. 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.
23. Agilent, *www.agilent.com.*
24. 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.
25. 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.
26. Staaf, J., et al., *Normalization of array-CGH data: influence of copy number imbalances.* BMC Genomics, 2007. **8**: p. 382.
27. 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.