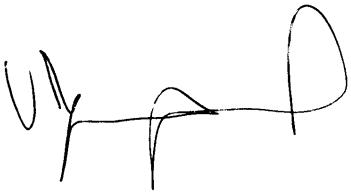
Armand Valsesia

PhD student – University of Lausanne

3rd year report

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

Lu et approuvé,
Professor C. Victor Jongeneel, Directeur de thèse



Professor Sven Bergmann, Directeur de thèse

CNV analysis in very large cohort

Previously [1], we developed a novel algorithm to detect copy number variant (CNV) in a large medical cohort (CoLaus, n~6000 individuals [2]). This novel method is based on Gaussian Mixture Models (GMM) and provides a probabilistic CNV calling. This method has been compared to three other established methods: CNAT with its two implementations [3] and the Circular Binary Segmentation algorithm (CBS, [4,5]). We confirmed the high sensitivity of CBS, and also showed the increased performance of our method with respect to the CNAT implementations. We also developed a merging strategy to combine CNV predictions at the population level using Principal Component Analysis and Self-Organizing Maps. A manuscript describing both the GMM and the merging strategy has been submitted to a peer-reviewed journal. Meanwhile the GMM code has been packaged, documented and compiled. Both executables and source code are freely available at http://www2.unil.ch/cbg/index.php?title=GMM. The package for the merging procedure is in progress, with some major modifications; notably to replace the Self-Organising Maps with other clustering techniques such as K-means and Hierarchical clustering, which are much less computationally intensive. Code to perform the post-processing of GMM predictions is also in preparation and will be added to the GMM package.

Our GMM package has been put to use within several collaborations. For example, in a project led by Prof. Jacques Beckmann (UNIL, CHUV) and Prof. Philippe Froguel (Imperial College London), we identified a rare deletion in patients affected with both morbid obesity and cognitive deficiencies. Our GMM predictions in the CoLaus population have proved to be very useful to clarify the penetrance of this variant. In fact, this variant accounts for 0.7% of the morbid case (odds ratio=43, P value = 6.4×10^{-8}). These results have been published in Nature [6] and a follow up study is in progress to clarify the expression pattern of the genes affected by the deletion. In addition, our results demonstrate the importance in common disease of rare variants with strong effects. Given that common variants are unlikely to explain the missing heritability in complex disease [7,8] screening existing cohorts for rare CNVs is of the highest interest.

Melanoma Sequencing project

In a collaboration involving the University of Lausanne and Geneva, the Ludwig Institute for Cancer Research, the Swiss Institute of Bioinformatics and the CHUV, we are performing a comprehensive molecular profiling of seven metastatic melanoma cell lines and their patient-matched control cells. This study aims at identifying somatic variants (copy number aberrations and mutations), genes with altered expression (over expression or down regulation) and genes with methylation status that differ from control cells (derived from the same patient). To achieve these objectives, the project combines karyotyping, array hybridization (CGH and SNP arrays) and RNA-seq analyses.

Comparison between CGH and SNP platforms

In our preliminary analysis [1], we observed that in highly polyploid samples, the hybridization ratios between cancer cells and matched controls did not reflect the chromosome-wide aberrations observed in the karyotypes. For example, tetraploid regions were measured as triploid or less by the CGH arrays. We asked whether this was due to the ratio normalization protocol and subsequent segmentation analysis. So we tested several normalization frameworks and applied two independent segmentation methods. Since neither of these methods gave entirely satisfactory results, we developed our own method, based on Gaussian Mixture Models. These Gaussian Mixture Models differ in implementation from those employed in the CoLaus study, because here the number of Gaussian components is not fixed in advance, but estimated from the data. Then a decision algorithm is employed to decipher the copy number state of each component. Based on technical replicates, we found that our method was outperforming the initial methods. Despite our improved analysis, the number of aberrations detected was less than expected based on the karyotype analysis. Thus, while CGH-based methods are well adapted to document differences in copy number status between the genomes of normal cells derived from different individuals, our results indicate that they are inadequate to deal with the large-scale rearrangements and amplifications typical of cancer cells. The most likely reason is that the total DNA content of cancer cells is too different from that of normal cells to allow a robust experimental normalization. Given such obvious limitations, we next asked whether SNP arrays were better than CGH arrays at detecting chromosome-wide changes in a highly amplified genome.

In contrast SNP-based predictions were much more accordant with the expectations based on chromosome-wide changes observed in karyotype experiments. SNP arrays are indeed a method of choice for the detection of amplifications because they benefit from the information about allelic imbalance that greatly improve the predictions of copy number [9,10]. However CGH arrays were able to detect significantly more deletions than SNP arrays, therefore it can be argued that CGH and SNP techniques should be combined to obtain a reliable assessment of all copy number states from homozygous deletion to high-level focal amplification.

Identifying genes with both somatic copy number aberrations and altered expression

We computed the median copy number at each Refseq gene, and then identified somatic copy number alterations (SCNA) as follows: A gene was flagged as within a focal amplification when its CN, as computed from SNP arrays, was ≥ 4, the difference in CN relative to the chromosomal arm was ≥ 1.5 and the gene was diploid (CN=2) in the matched control cell. For homozygous deletion, a gene needed to have CN=0, as detected from CGH, without detected expression in the melanoma and also to have CN=2 with detected expression in the melanocytes.

We reasoned that the combination of precise copy number determinations and gene expression measurements would allow us to flag with much higher confidence those genes whose expression is affected by SCNA in the melanoma cell lines. We therefore analyzed genome-wide gene expression in each of the metastatic melanoma cell lines by RNA-seq using the Roche/454 pyrosequencing method. Additionally, we performed RNA-seq on a pool of normal melanocytes to determine a reference level of expression for each gene in this cell type. We used our list of genes within focal amplifications as predicted from SNP arrays and with at least two-fold over-expression in the affected melanoma relative to normal melanocytes, and added those genes within homozygous deletions detected from CGH arrays that had also lost expression. In total, we identified 1,526 SCNA genes across the seven melanomas.

Identifying pathways significantly perturbed in seven metastatic melanoma cell lines

Very few of the SCNA genes were present in more than one melanoma cell line, which is not unexpected given the small number of samples. A current idea in literature [11,12,13,14,15] is that signalling pathways, rather than individual genes, are recurrently perturbed in cancer. To investigate the possibility that SCNA genes from different melanoma cell lines shared

membership of one or more signalling pathways, we devised a metric combining copy number status, expression level and participation in known networks of protein-protein interactions. Such a metric would indicate which pathways were significantly affected by deletions and amplifications in our melanoma samples and therefore were potentially relevant to malignant transformation.

We investigated whether the proteins encoded by the SCNA genes were connected in any of the known human protein interaction networks. Out of a total of 1,424 proteins analyzed, 309 (22%) were found to be connected within the network and were clustered in 17 sub-networks. Among these sub-networks, only nine (accounting for 63 genes) has significantly overlapped with known pathways. These pathways are highly relevant to cancer; a few examples include the cell cycle, the WNT and MAPK signalling pathways as well as the angiogenesis.

Fifty of our 63 genes were also observed in the much larger survey of cancer-associated SCNA carried out in the Cancer Genome Project [16,17]. Our results confirm that genes known to be altered in melanoma, such as *KRAS* and *BRAF*, are commonly affected by SCNA. We therefore feel confident that in spite of the small number of samples that we analyzed our results will be able to inform subsequent studies.

The methylation and exome-sequencing data of our melanoma initiative are still being generated. Once these results are available, we will be able to refine our integrative analysis, hopefully confirm the identified pathways and improve our knowledge about the different mechanisms that can deregulate pathways in cancer.

References

1. Valsesia (2009) Mid-thesis report.
2. Firmann M, Mayor V, Vidal PM, Bochud M, Pecoud A, et al. (2008) The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC Cardiovasc Disord 8: 6.
3. Huang J, Wei W, Zhang J, Liu G, Bignell GR, et al. (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. Hum Genomics 1: 287-299.
4. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657-663.
5. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5: 557-572.
6. Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature 463: 671-675.
7. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464: 713-720.
8. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, et al. Origins and functional impact of copy number variation in the human genome. Nature 464: 704-712.
9. Attiyeh EF, Diskin SJ, Attiyeh MA, Mosse YP, Hou C, et al. (2009) Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. Genome Res 19: 276-283.
10. LaFramboise T, Weir BA, Zhao X, Beroukhim R, Li C, et al. (2005) Allele-specific amplification in cancer revealed by SNP array analysis. PLoS Comput Biol 1: e65.
11. Klijn C, Bot J, Adams DJ, Reinders M, Wessels L, et al. Identification of networks of co-occurring, tumor-related DNA copy number changes using a genome-wide scoring approach. PLoS Comput Biol 6: e1000631.
12. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, et al. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321: 1801-1806.
13. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, et al. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer Res 70: 4453-4459.
14. Murohashi M, Hinohara K, Kuroda M, Isagawa T, Tsuji S, et al. Gene set enrichment analysis provides insight into novel signalling pathways in breast cancer stem cells. Br J Cancer 102: 206-212.
15. Heiser LM, Wang NJ, Talcott CL, Laderoute KR, Knapp M, et al. (2009) Integrated analysis of breast cancer cell lines reveals unique signaling pathways. Genome Biol 10: R31.
16. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, et al. Signatures of mutation and selection in the cancer genome. Nature 463: 893-898.
17. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. (2004) A census of human cancer genes. Nat Rev Cancer 4: 177-183.