UNIL | Université de Lausanne

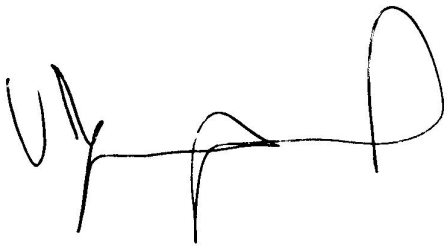
SIB

LUDWIG INSTITUTE FOR CANCER RESEARCH

Armand Valsesia

PhD student – University of Lausanne

1st year report



Lu et approuvé
C. Victor Jongeneel, Directeur de thèse

Preamble

The PhD subject, as proposed by Pr. Victor Jongeneel, is centered on the analysis of the copy number polymorphisms of cancer testis genes and their association with cancer. Cancer Testis (CT) genes are mainly expressed in the germ line and in various types of tumors. To date, 70 CT families have been identified. The biological function for most CT genes remains unclear, although experimental data indicate that some are involved in the meiotic process.

Improving our understanding of CT genes is crucial as they are targets of choice for cancer immunotherapy and for the development of cancer vaccines. We propose to study the polymorphisms of CT genes and to investigate the relationship between CT variation and the development of cancer. The first step would be to identify copy number polymorphism in known CT genes, the second to derive the status of CT genes in normal and cancerous patients then to investigate the link between CT polymorphism and oncogenesis.

Since little is known about the copy number polymorphism of CT genes, the first step is to constitute a polymorphism dataset. Through a collaboration with both the CHUV and GlaxoSmithKline; and supervised by Pr. Sven Bergmann (thesis co-director) and by Pr. Jacques Beckmann; I have access to the CoLaus dataset, a population based health survey of about 6000 individuals from the Lausanne area. In this report, I am presenting my work on cataloguing rare and common copy number variants in the Swiss population.

Background

Genetic variation in the human genome takes many forms ranging from large chromosome anomalies to single nucleotide changes (SNPs). Deletion, insertion and duplication, termed copy number variations (CNVs) have been found genome wide in the human genome [1-6] and in other genomes such as primates [7, 8].

A CNV is commonly defined as a DNA segment longer than 1Kb that is present at variable copy number compared to a reference genome [9, 10]. The link between CNVs and disease has been previously established [11]. Copy number plasticity is typical of cancer cells [12]. Such genomic aberrations were identified decades ago using array-based comparative hybridization [13]. It has been demonstrated that CNVs near oncogenes or tumor suppressor genes can affect the gene expression level or result on the expression of chimeric fusion genes [14, 15]. Cataloguing CNVs involved with clinical outcome is part of the international effort DECIPHER [17]. However the number of CNVs and positions in the human genome are still underestimated and their contribution to complex diseases such as heart disease is unclear.

CoLaus (Cohorte Lausannoise) is a population based health examination survey started in 2003 to study the genetics of hypertension and cardiovascular disease [18]. More than 6000 individuals (35-75 years old) from the Lausanne area participated 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 [19].

One of our primary interests in analysing CoLaus is to establish a comprehensive CNV map in order to carry out association studies to determine their impact on the etiology of common disorders. The CNV research community agrees that many CNVs are to be discovered and their frequency in different population to be estimated. We take the opportunity with our large population dataset to estimate the fraction of the genome that is effectively variant, to look at both *rare* and *common* CNVs and compare them to each other. We also investigate how to estimate CNV boundaries with unsupervised approaches.

Results

CNV analysis pipeline

A complete bioinformatics pipeline has been set up to perform the CNV analysis of the Affymetrix 500K platform. This pipeline can be decomposed into three major steps:

1. Normalizing and summarizing the probe hybridization intensities with the Affymetrix GeneChip® Genotyping Analysis Software [20] software.
2. CNV calling using the Copy Number Analysis Tool (CNAT) [21]. This tool includes additional normalization and smoothing procedures before applies an Hidden Markov Model (HMM) segmentation algorithm that categorizes each SNPs of a given individual as being either deletion (homozygous or hemizygous), copy neutral (non variant) or gain (simple gain or multiple gain).
3. Combining the CNV profile of all individuals into a map of common and rare variants at the population level.

Considering the amount of individuals to be analysed, considerable effort has been put to parallelize as much as much possible the heavy computational steps (steps 1 and 2) on the high performance computing cluster Vital-IT [22]. Provided there is enough storage (which can be a limitation), analysis of the complete cohort can be done in less than a week. The pipeline also permits to do genotype calling (using either the DM or the BRLMM algorithms [21]). Step 3 has been implemented in the Matlab programming language [23] is extremely fast to run (less than 1 hour).

Improving the normalization

A CNV in a test individual is relative to a reference genome, when doing CNV analysis of tumours the reference can be from a normal cell from the same individual. Using a pool reference makes sense when analysing healthy individuals. Affymetrix recommendations are to use 25 individuals as references. Therefore we initially carried out the analysis with 30 unrelated CEU references[23]; gender matched with the test individuals. We demonstrated that such sample size is too small and leads to strong biases (figure 1). Another technical problem was that CoLaus individuals have been genotyped in four different centres.

By randomly selecting equal proportion of CoLaus males and females for each genotyping centre, we accounted for the bias due to technical differences between centres and corrected completely for the gender differences. But this implied we could no longer predict CNVs on chromosome X due to the gender mixed reference panel. After testing, we

found that using 280 references was producing significantly better results than the initial 30 references (figure 2).

Combining CNV from individuals at the population level

A recurrent problem for CNV discovery is in combining information from individual level CNVs into a population consensus. So far researchers have been using ad-hoc methods such as when two CNVs overlap each other by a minimal threshold then they are merged together into a CNV region. But in a large population, one should only consider highly confident CNVs and have some prior knowledge about the population in order to set the appropriate threshold and avoid over-estimating the CNV regions.

A solution is to split CNVs into consensus segments and to retain the frequency information for each segment. Such merging procedure (figure 3) reduced the data complexity from 500k SNPs to about 21k and 39k regions for the *CNAT total* and *allelic* approaches, respectively (Figure 4). These regions can be stratified by variation frequency (by counting the number of distinct individuals that have a copy number state different from the diploid case).

Based on the variation frequency, we were able to stratify regions. The spectrum of variation frequency indicated that *rare* variant regions (less than 1% frequency) were more numerous than *common* regions (>1%). Interestingly the *rare* fraction was comparable between the two CNAT approaches. However *CNAT allelic* was able to detect more *common* variant regions (covering 7.7% of the genome as opposed to less than 1% with CNAT *total*).

We observed that long variants tend be *rare* (figure 4). The average size of *rare* CNVs is 42 and 36kb for *CNAT allelic* and *total* respectively, whereas *common* CNVs are 27kb and 12kb long. Such observations can be explained by the fact that long CNVs will more likely affect the phenotype, by affecting gene regulatory elements or disrupting a gene. Therefore those long CNVs are more likely to undergo strong purifying selection pressure.

Testing CNV algorithms

To test CNV discovery algorithm, we developed a powerful test that uses relatedness between individuals. Given that family-related individuals are available in the cohort, one can check whether significant distinction on the CNV profile is made between pairs of related individuals compared to random pairs. CNAT implements two methods to call CNVs, the *total* and the *allelic* approach. Total is optimized to reduce the noise but is biased toward the copy neutral state whereas allelic is optimized for this bias but at the cost an increase in the false positives rate.

We observed that *allelic* made significant distinction between related and unrelated pairs whereas total was not able to distinguish at all between these two distributions (figure 5).

This demonstrates that *allelic* produces CNV calls that better reflect the relatedness of the sampled individuals and therefore is likely to produce more reliable results than *CNAT total* approach.

Conclusion

We have catalogued a high resolution map of CNV in the Swiss population. We found 8436 and 356 *common* regions (regions > 1% frequency) with *CNAT allelic* and *total,* respectively (*total* being a subset of *allelic*). We found that *rare* regions (frequency <1% and present in more than 2 individuals) are more numerous than *common* regions (22710 and 12595 regions for *CNAT allelic* and *total* respectively).
Moreover *rare* regions tend to be longer than *common* regions. Depending on which functional genomic elements (gene promoters, enhancers, transcription factorsé) are contained within *rare* regions. These might undergo strong selection pressure. Regions that either confer selective advantage or remain neutral may be fixed in the population. A population genetics study using different ethnicity and comparative genomics will give more power to explain the history of *common* variations and possibly to date when the variation was ñfixedò in the population.

Our merging procedure produces conservative CNV regions and reduces greatly the data complexity. It does not require using any ad-hoc thresholds and is platform-independent thus can be applied to any other SNP arrays. We are now developing a more sophisticated approach whose preliminary results indicate a significant improvement for complex and common regions.

In the absence of proper technical replicates, identifying related individuals based on their CNVs provides a powerful and innovative test to assess CNV discovery algorithms.
Base on this approach, we demonstrated that the *CNAT allelic* performs better than *CNAT total*. This is important because only the *CNAT total* approach is implemented on the Windows CNAT GUI. In contrast *CNAT allelic* is available via UNIX command line.

Since both chip platforms and methods evolve rapidly, benchmark reviews become very rapidly obsolete. It is therefore important that we continue our efforts to develop validation tests and filtering procedures. Also in the context of very large cohort, analysing and validating *rare* CNVs is definitely not a trivial task. Very likely, it may be necessary to process them differently from the *common* CNVs. We will investigate applying more stringent criteria such as keeping the intersection of *total* and *allelic rare* CNVs.

Taking into account the *common* regions, we found that more than 7% of the genome is variant. This alone demonstrates that there are a lot of *common* CNVs to be discovered. Recent studies are also converging to the same conclusion and higher resolution platforms such as Affymetrix 6.0 are being developed to interrogate the genome even more comprehensively. Since CNVs encompass more nucleotides than SNPs and SNPs explain so far but a small fraction of the phenotypic variation, one can investigate though genome wide association studies the impact of CNVs on disease. Once we completed a high-quality CNV map, we plan to perform GWA studies on the Colaus phenotypes and clarify the link between CNVs and cardiovascular diseases.

Perspectives and future work

Extending the Copy Number Polymorphism catalog

In addition to the CoLaus project, CNV analysis is in progress on the EuroCHAVI cohort, an HIV-infected cohort composed with about 1200 individuals genotyped on the Illumina 550K platform. Also in collaboration with Pr. Mehdi Tafti, we are mining for CNVs in a case and control study including 500 narcoleptic patients and about 250 controls. Such studies will greatly complement the copy number polymorphism list from the CoLaus study and be of use to study CT genes, in particular the ones located on the autosomes.

A custom chip covering CT genes for CNV and transcriptomic analyses

Because SNP coverage on chromosome X is poor on the current arrays, we have little power to detect CNVs on X. Since many CT genes are present on this chromosome, we are designing a custom oligonucleotide chip to perform CNV analysis. The aim is to cover all known CT genes at high resolution with 60 mer probes.
This work is done in collaboration with Dr. Brian Stevenson and Dr. Christian Iseli. Such chip will be very useful to analyse both normal and cancer cell lines available within the Ludwig Institute. Then comparing the polymorphisms of normal and cancerous cell lines will help to find variants that might be correlated with oncogenesis.
By covering different gene transcripts, it will be possible to quantify the amount of transcripts (reversed to cDNA) that hybridize to the chip and thus to study the transcriptomic profile of CTs. Combining CNV and transcriptomic profiles of CTs will help to derive their status in healthy and cancerous patients.

Comprehensive genomic profiling of melanomas

As part of a big collaboration, headed by Dr. Christian Iseli, between the Ludwig Institute, CHUV, Universities of Lausanne and Geneva; we plan to perform a comprehensive genomic profiling of melanomas.
This project includes edge cutting genomic methodologies such as 1) array CGH, SNP arrays and karyotyping to study genomic rearrangement; 2) to study methylation pattern using oligonucleotide arrays; 3) to search for mutations in protein-coding genes by sequence capture and 4) to identify aberrant splicing by transcriptomic profiling using ultra high-throughput sequencing.
My contribution in this ambitious project will be on analysing micro-arrays (both CGH and SNP arrays) and data from sequencing for structural variations. We expect that this integrated analysis will identify novel genetic alterations associated with melanoma and potentially provides new insight for molecular therapies.

Annexes

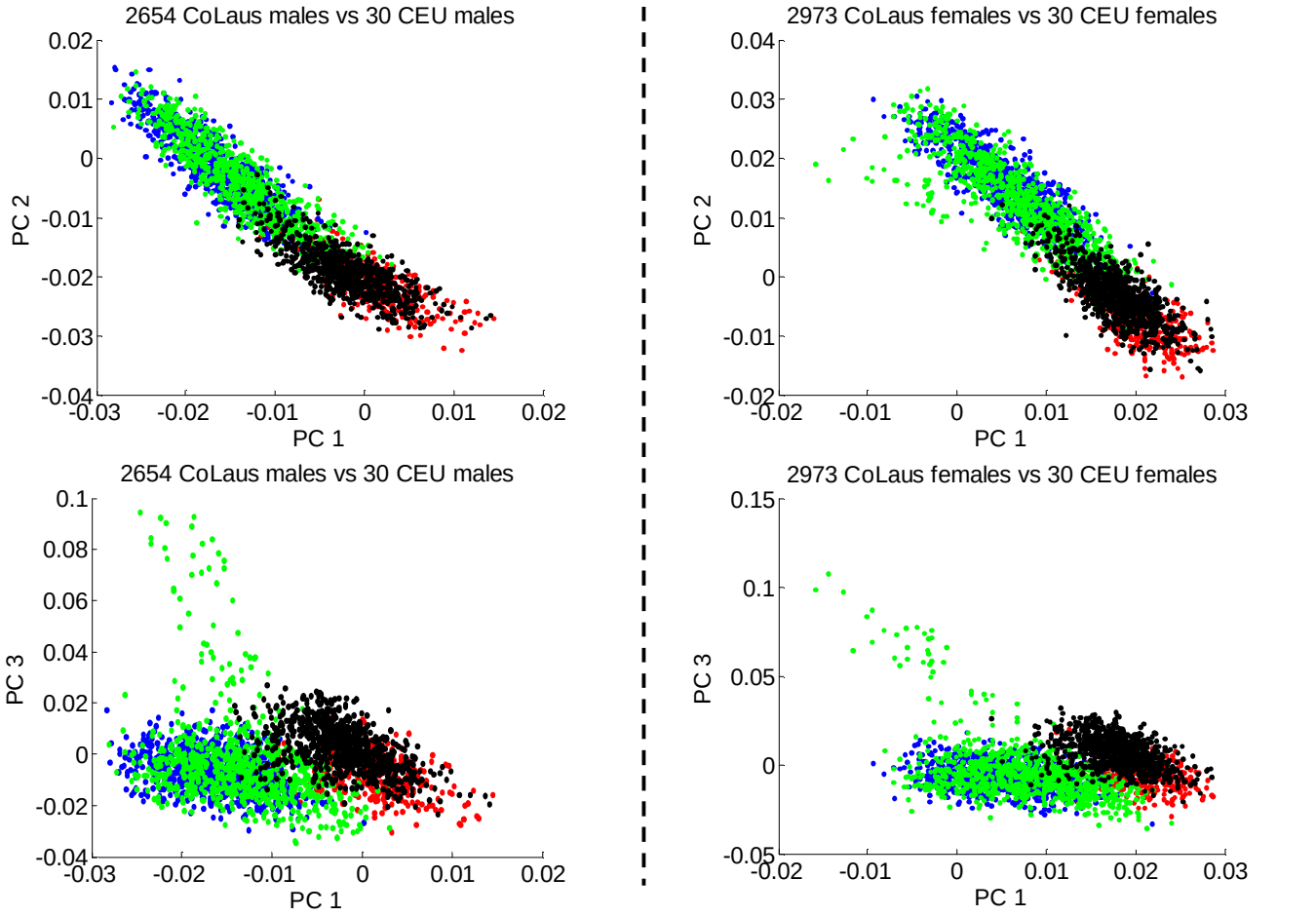


Figure 1: principal component analysis on the CNV profile of the CoLaus individuals.
Using normalization with 30 unrelated CEU individuals leads to differences between CoLaus males and females. Also differences between individuals genotyped in different centers are observed (the blue and green dots clearly separate from the red and black dots).

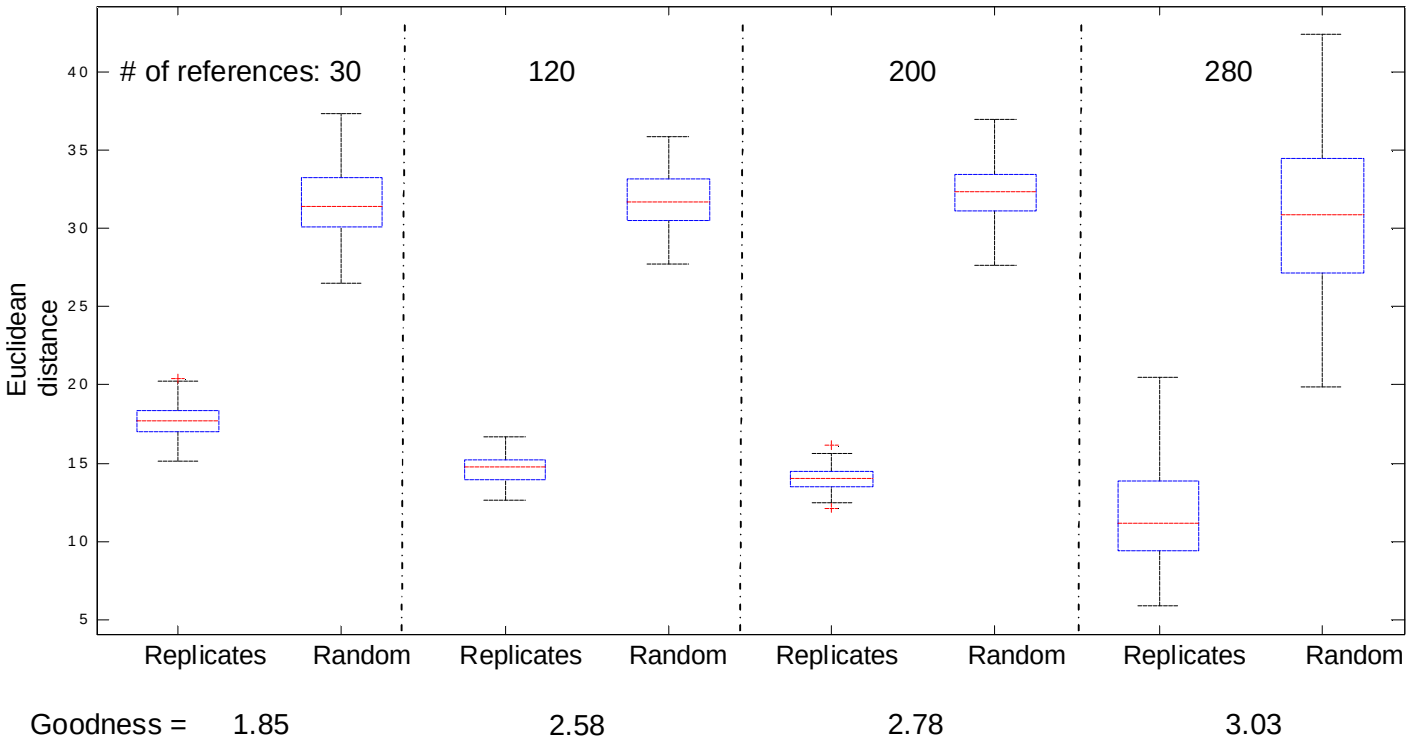


Figure 2: Normalization using different number of references (from 30 references to 280 with equal proportion of males and females) from a single genotyping centre (centre1 where 615 CoLaus individuals have been analysed). The box plots represent the distribution of Euclidean distances between pairs of the same individuals normalized with two independent reference panels having the same number of references (ñReplicatesò group) and distance between random pairs of individuals (ñRandomò group).
To compare the 30,120 and 200 reference panels, the distance has been calculated from 10 iterations using 55 individuals. The distance distribution for the 280 reference panel has been calculated from 55 individuals only once (no other iteration was possible). This explains why the variance is bigger in this panel.

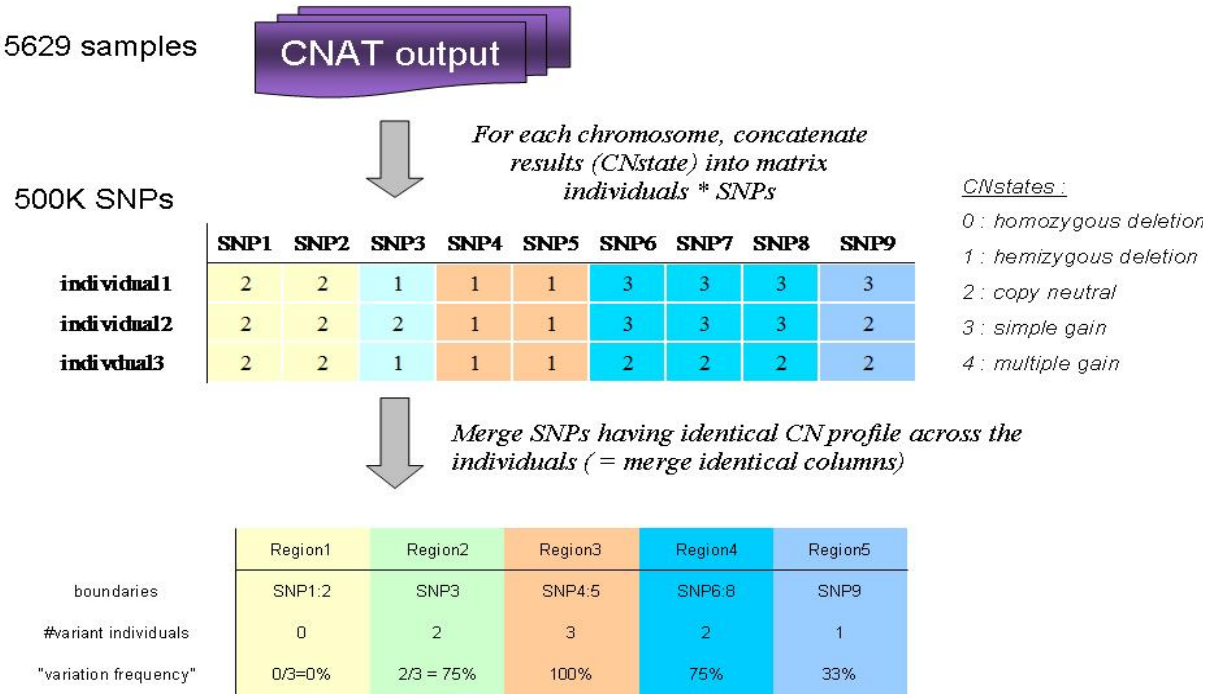
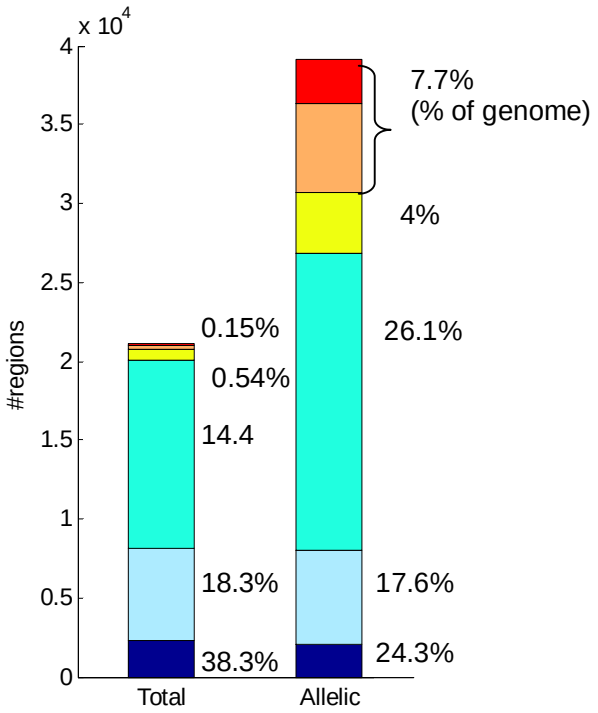


Figure 3: Workflow to merge SNP CN prediction from individuals into regions of variation
Regions composed only by SNPs in copy neutral states are designed as monomorphic regions.
The variation frequency is expressed as the percentage of individuals having non-copy neutral SNPs in a given region divided by the total number of individuals.



	meanSize (Kb)	
frequency	**Total**	**Allelic**
>=5%	13	28
]1, 5[	12	27
]0.5, 1]	24	31
]0.05, 0.5]	36	42
]0, 0.05]	93	89
0	516	354

Figure 4: number of regions (stratified by their population frequency) found by the CNAT approaches *total* and *allelic*. The percentage next to the bars correspond to the fraction of the genome covered by the ñcommonò, ñrareò, ñvery rareò and ñmonomorphicò regions.
The table gives the mean region length in Kb for each strata of frequency

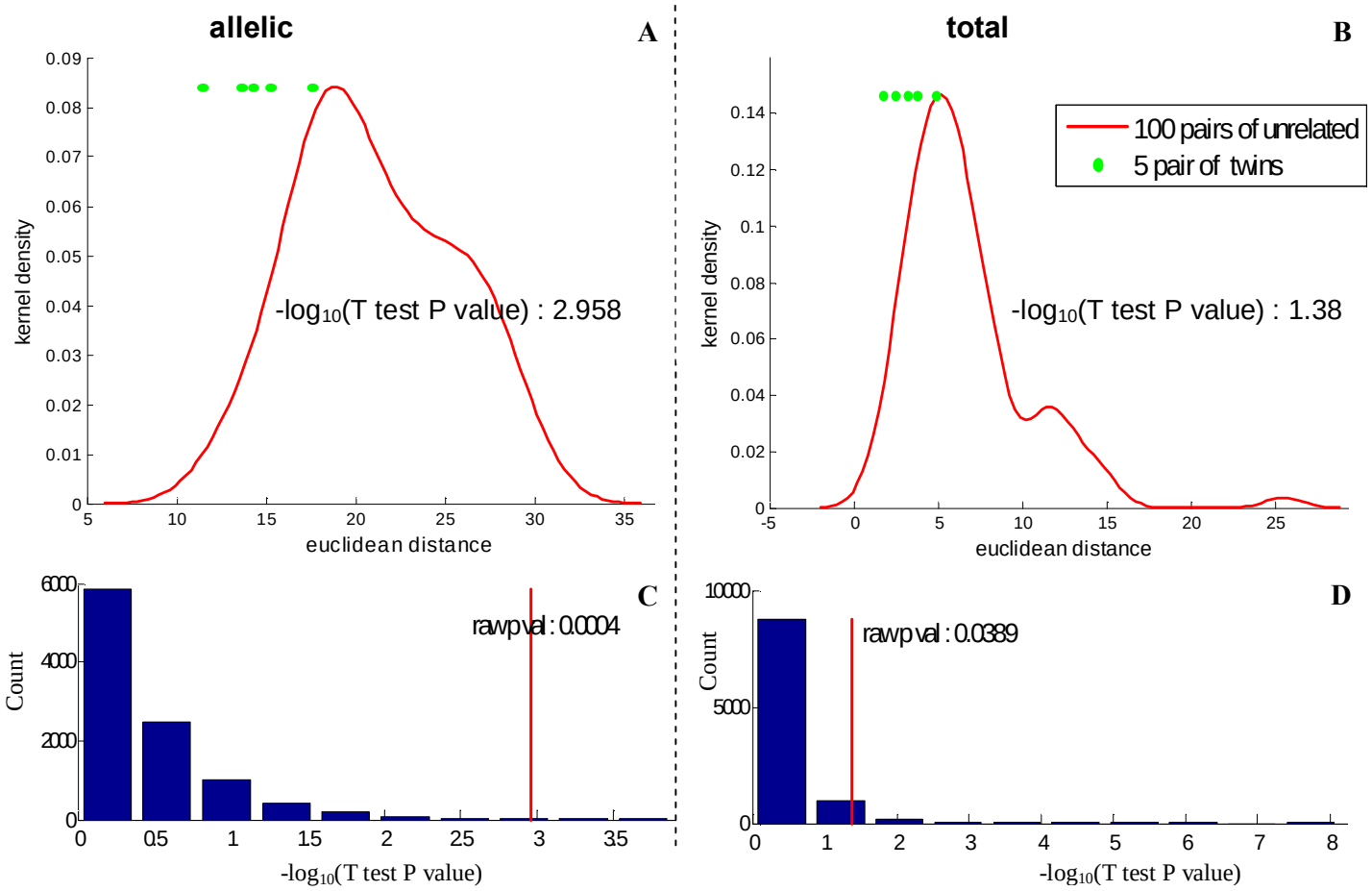


Figure 5: distinguishing between related and unrelated using the CNV profile

A, B: pairwise distance between pairs of individuals defined as Euclidean distance using the copy number state at each SNP. Pairwise distances have been computed for the 5 pairs of replicated individuals (also called twins) and for 100 pairs of unrelated individuals.
Based on a T test between these two distributions, *allelic* demonstrates significant distinction between related and unrelated individuals ($-\log_{10}$ P value =2.958) whereas *total* is not able to distinguish ($-\log_{10}$ P value = 1.38).
C, D: in blue, empirical distribution of ï $\log_{10}$ P value from T test between 5 values selected randomly versus the remaining 100 values. The raw P value indicates how robust the observed P value (in red) is. This raw P value is equal to the number empirical P values (from the blue histograms) that are better than the observed P value, divided by the number of tests. The *allelic* raw P value is 0.0004 and confirms observation made in A is robust.

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. Sebat, J., et al., *Large-scale copy number polymorphism in the human genome.* Science, 2004. **305**(5683): p. 525-8.
3. 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.
4. Tuzun, E., et al., *Fine-scale structural variation of the human genome.* Nat Genet, 2005. **37**(7): p. 727-32.
5. Freeman, J.L., et al., *Copy number variation: new insights in genome diversity.* Genome Res, 2006. **16**(8): p. 949-61.
6. Redon, R., et al., *Global variation in copy number in the human genome.* Nature, 2006. **444**(7118): p. 444-54.
7. 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.
8. 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.
9. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. **7**(2): p. 85-97.
10. Feuk, L., et al., *Structural variants: changing the landscape of chromosomes and design of disease studies.* Hum Mol Genet, 2006. **15 Spec No 1**: p. R57-66.
11. Lupski, J.R. and P. Stankiewicz, *Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes.* PLoS Genet, 2005. **1**(6): p. e49.
12. 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.
13. Kallioniemi, A., et al., *Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.* Science, 1992. **258**(5083): p. 818-821.
14. Pinkel, D. and D.G. Albertson, *Array comparative genomic hybridization and its applications in cancer.* Nat Genet, 2005. **37 Suppl**: p. S11-7.
15. Kallioniemi, A., *CGH microarrays and cancer.* Curr Opin Biotechnol, 2008. **19**(1): p. 36-40.
16. Butler, M.G., et al., *Array comparative genomic hybridization (aCGH) analysis in Prader-Willi syndrome.* Am J Med Genet A, 2008. **146**(7): p. 854-60.
17. DECIPHER. https://decipher.sanger.ac.uk/.
18. 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.
19. Affymetrix. www.affymetrix.com.
20. GTYPE. http://www.affymetrix.com/products/software/specific/gtype.affx.
21. CNAT. http://www.affymetrix.com/support/technical/software_downloads.affx.
23. HAPMAP. www.hapmap.org.