Global variation in copy number in the human genome

Redon et. al. Nature 444:444-454 (2006)

12.03.2007

Tarmo Puurand

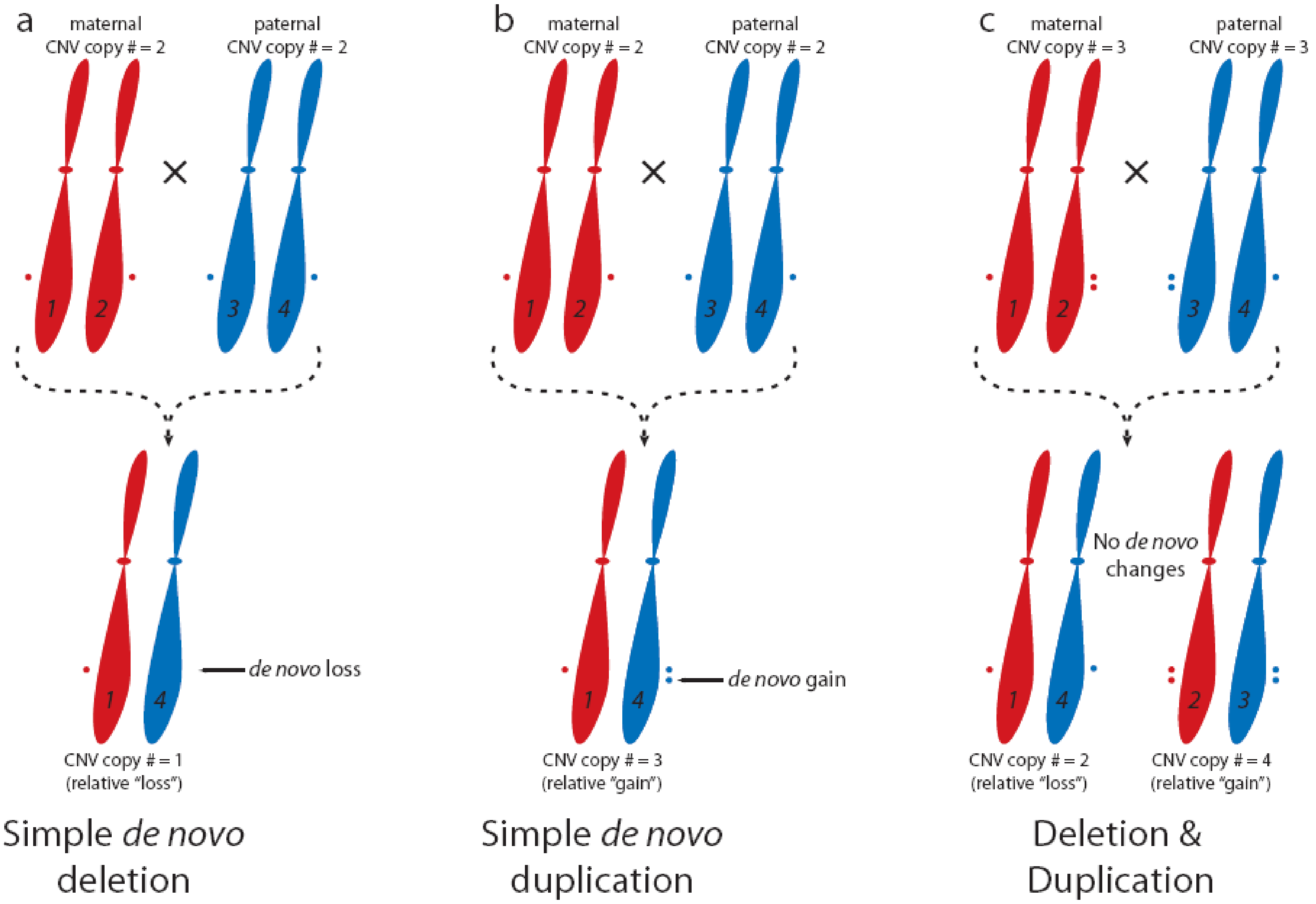
Study

- 270 individuals (HapMap collection)
- Affymetrix 500K
- Whole Genome TilePath (WGTP)
- 1447 CNVRs, 360 Mbp, 12% of the genome
- sum(length(CNVRs)) > sum(SNPs) per genome

CNVs

- Deletions
- Insertions
- Duplications
- Complex multi-site variants

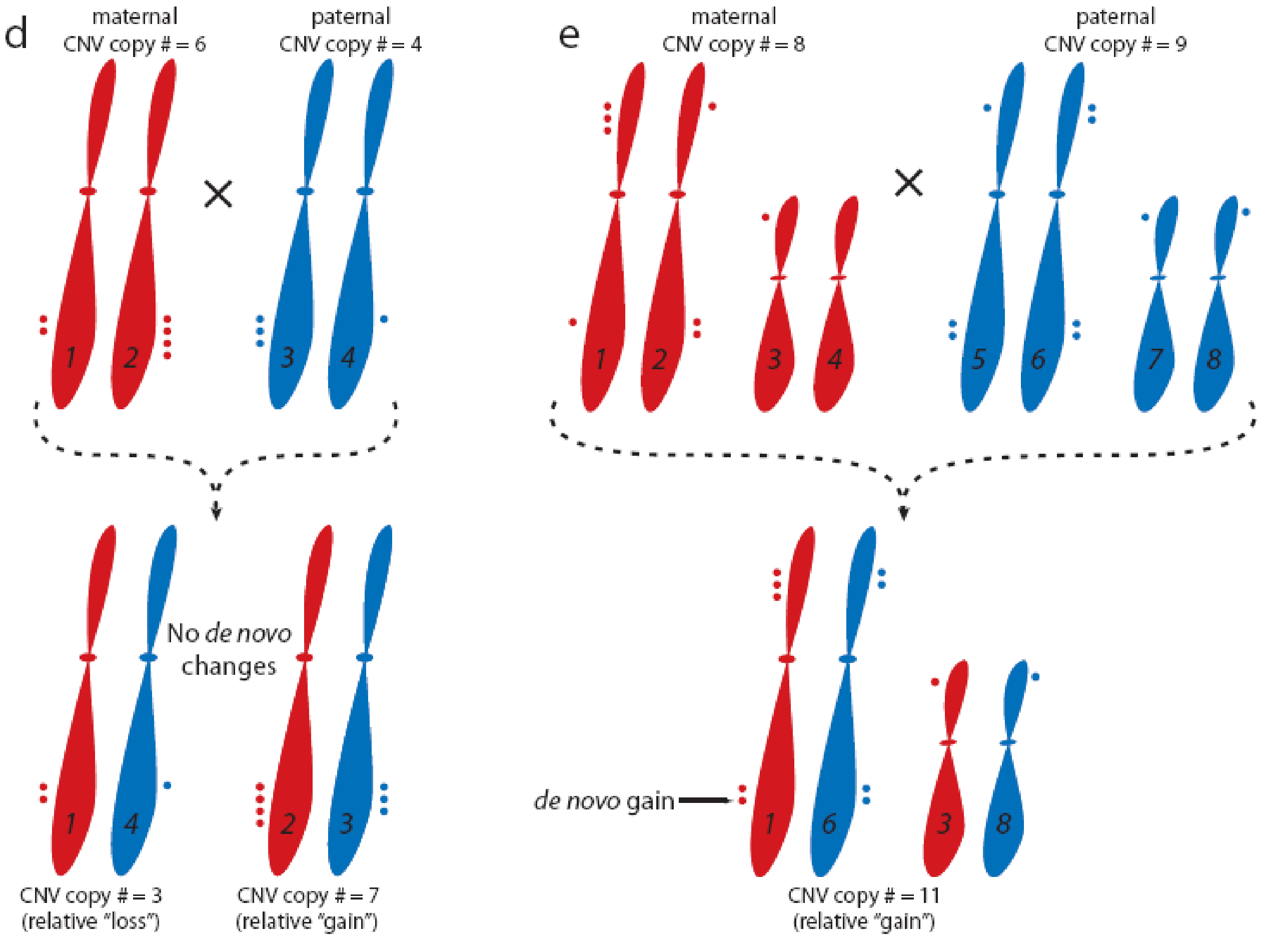
- In this study over 1 kb in coparision with a reference genome



Simple *de novo* deletion

Simple *de novo* duplication

Deletion & Duplication



Multi-allelic variant
(2-8 copies per individual)

Complex CNV with
de novo gain

Technology platforms

- Copparative analysis of:
- 1. Affymetrix GeneChip Human Mapping 500K, 474642 SNPs
- 2. Hybridisation with Whole Genome TilePath (WGTP) array, 26574 large-inserting clones, 93,7% euchromatic DNA

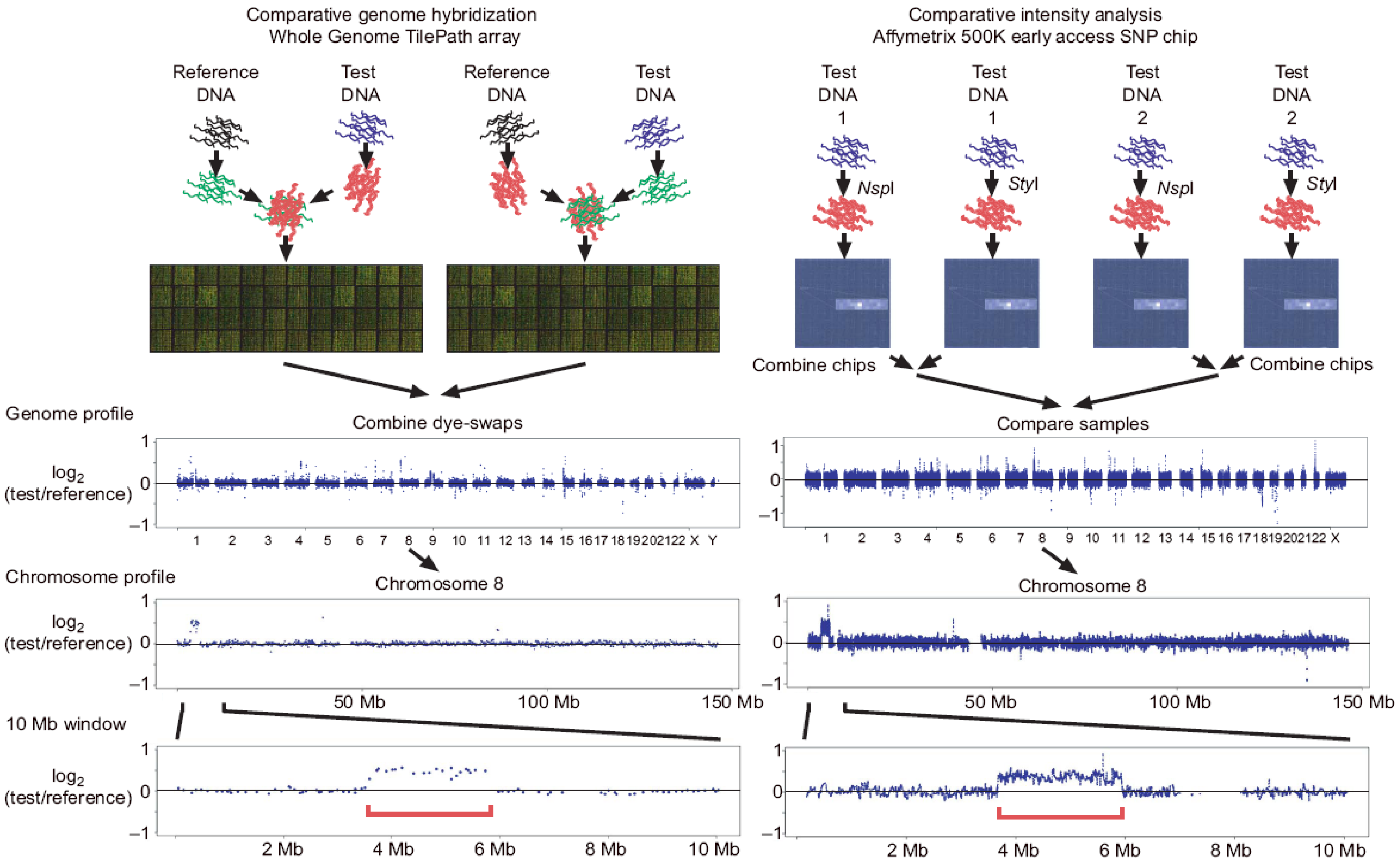
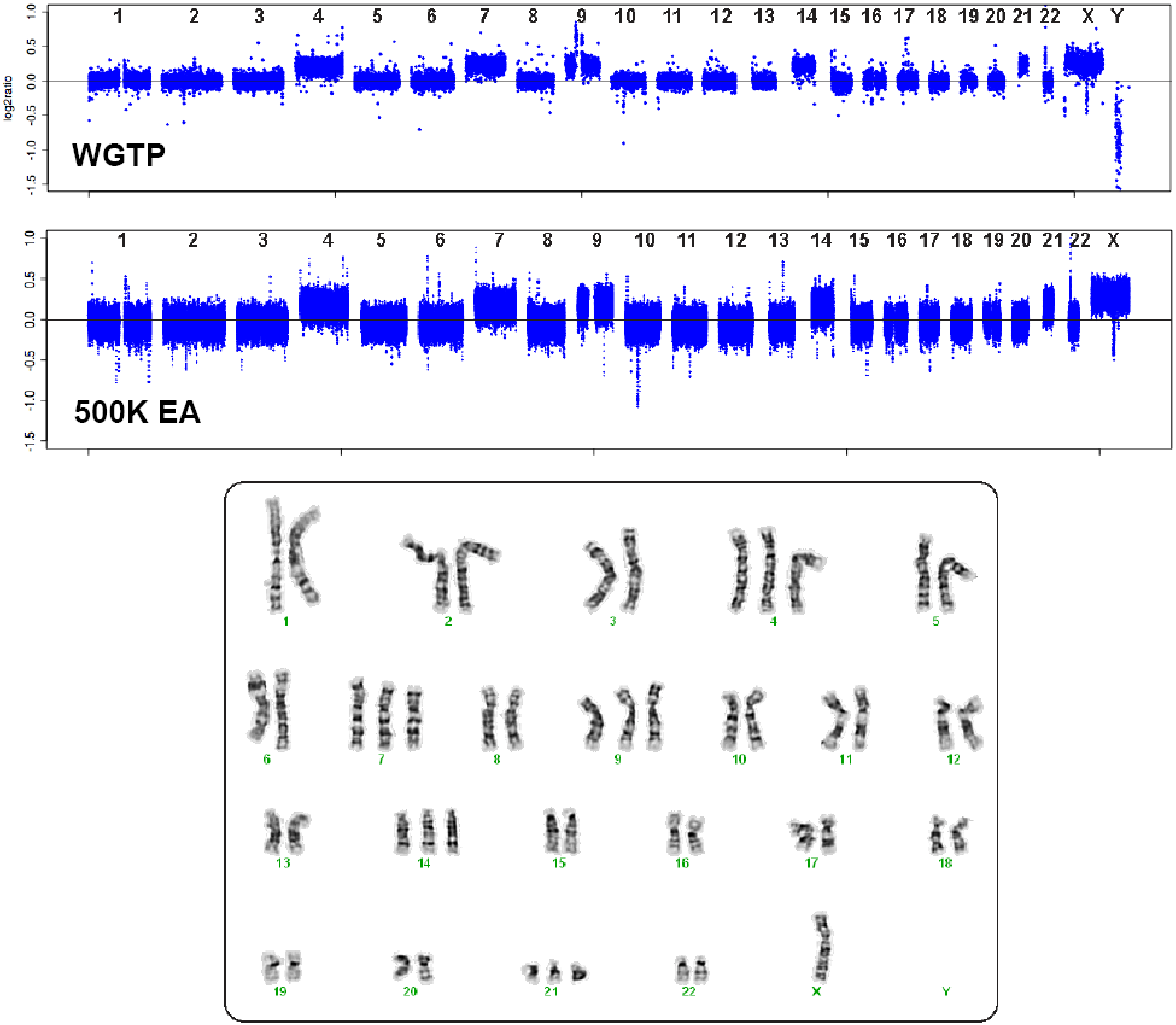


Figure 1 | Protocol outline for two CNV detection platforms. The experimental procedures for comparative genome hybridization on the WGTP array and comparative intensity analysis on the 500K EA platform are shown schematically (see Supplementary Methods for details), for a comparison of two male genomes (NA10851 and NA19007). The genome profile shows the $\log_2$ ratio of copy number in these two genomes chromosome-by-chromosome. The 500K EA data are smoothed over a five-probe window. Below the genome profiles are expanded plots of chromosome 8, and a 10-Mb window containing a large duplication in NA19007 identified on both platforms (indicated by the red bracket).

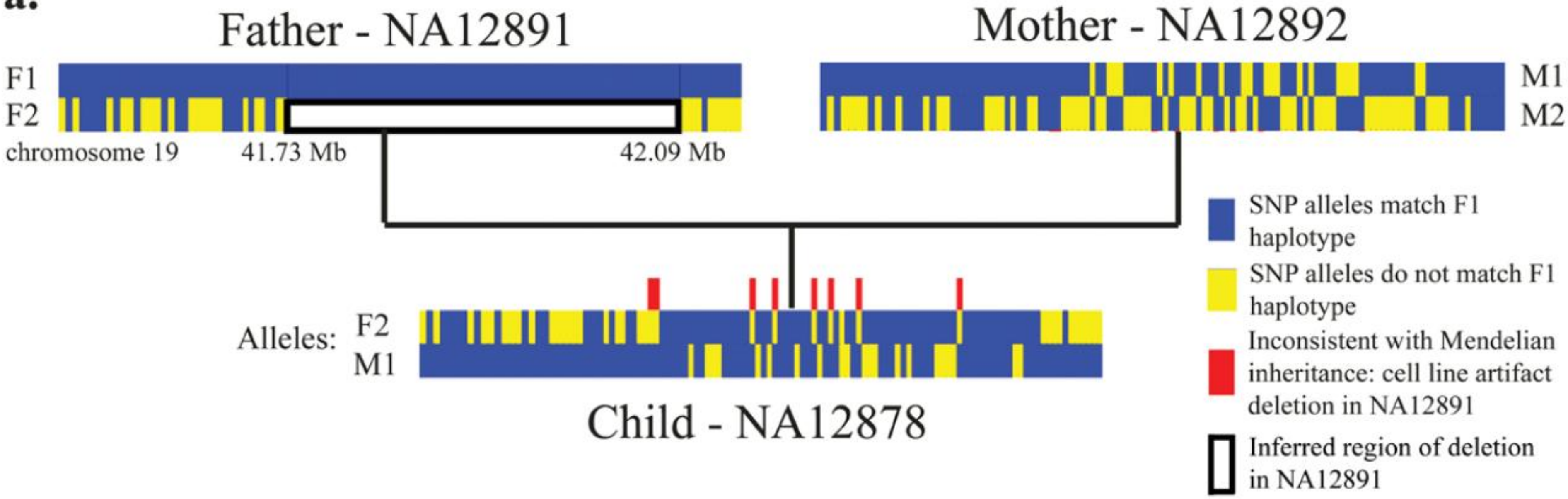
Quality control

- Repeated experiments- 82 individuals on the WGTP and 15 individuals on 500K
- Assessed by standard deviation among $\log_2$ ratios of autosomal probes (after normalisation and filtering for cell-line artefacts)
- To train threshold parameters- 203 CNVs based by NA10851 and NA15510
- DNA- Epstein-Barr-virus transformed lymphoblastoid cell lines.
- Karyotype all 268 cell lines, 30 of these with abnormalities.
- Most common chr9, chr12 and chrX trisomies, mosaic trisomy of chr12
- Experiments triplication for 10 individuals



WGTP
log2ratio
1.0
0.5
0.0
-0.5
-1.0
-1.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
500K EA
1.0
0.5
0.0
-0.5
-1.0
-1.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
1 2 3 4 5
6 7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 X Y

a.


Father - NA12891
F1
F2
chromosome 19
41.73 Mb
42.09 Mb
Mother - NA12892
M1
M2
Alleles:
F2
M1
Child - NA12878
SNP alleles match F1 haplotype
SNP alleles do not match F1 haplotype
Inconsistent with Mendelian inheritance: cell line artifact deletion in NA12891
Inferred region of deletion in NA12891

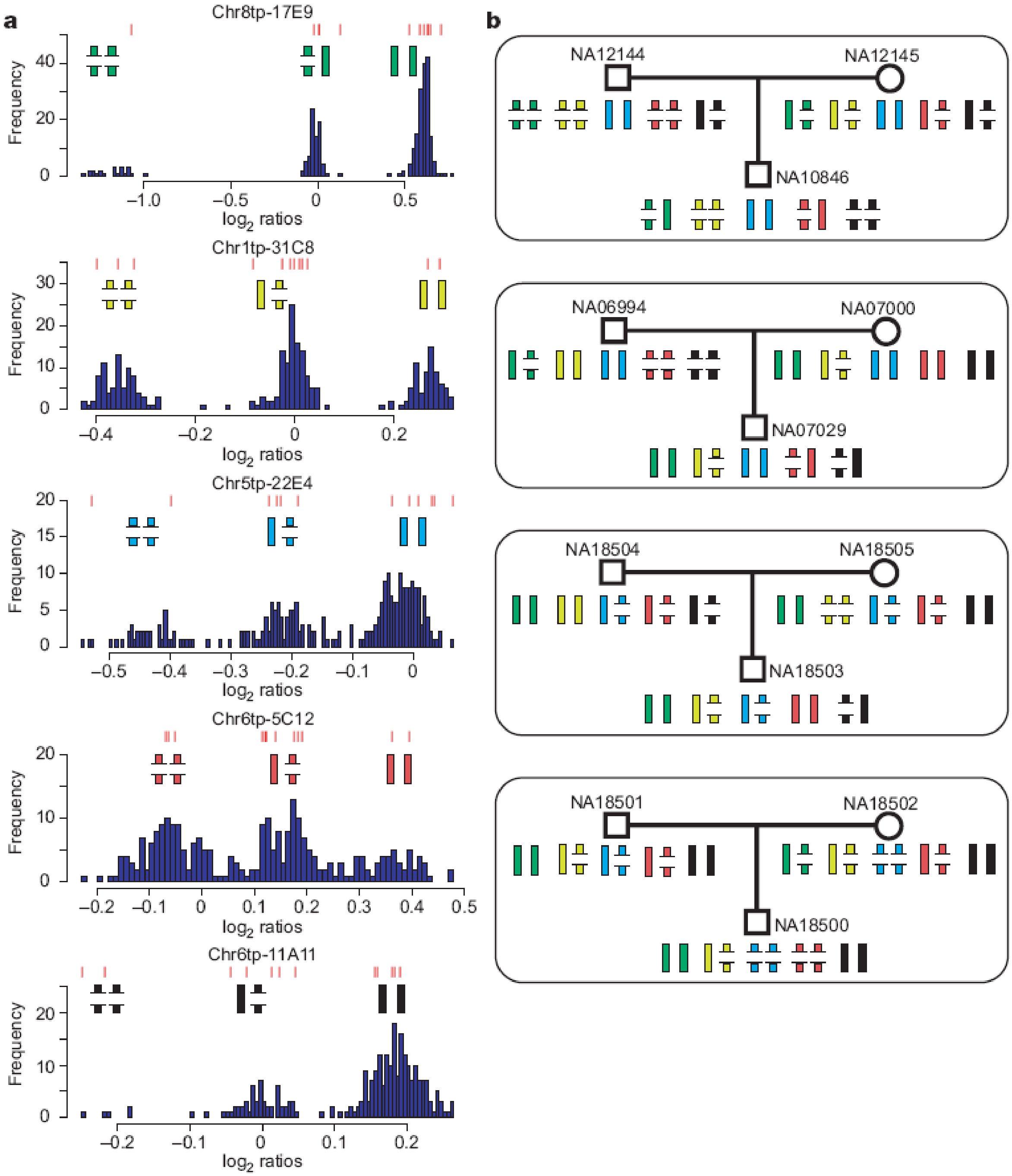


Figure 2 | Heritability of five CNVs in four HapMap trios. a, The distribution of WGTP $\log_2$ ratios at five CNVs with genotype information. Each histogram of $\log_2$ ratios in 270 HapMap individuals exhibits three clusters, each corresponding to a genotype of a biallelic CNV, with the two alleles depicted by broken and complete bars, representing lower and higher copy number alleles, respectively. Red lines above each histogram denote $\log_2$ ratios in the 12 individuals represented in **b**. **b**, Mendelian inheritance of five CNVs in four parent–offspring trios. The individual CNVs were genotyped from WGTP clones: green, Chr8tp-17E9; yellow, Chr1tp-31C8; blue, Chr5tp-22E4; red, Chr6tp-5C12; black, Chr6tp-11A11.

A genome-wide map of CNVs

- CNV detection per experiment was 70 and 24 (WGTP and 500K respectively)
- WGTP detects in both, test and reference
- 500K detect only in single genome
- Median size 228 kb (WGTP) and 81 kb
- Detection of 913 CNVs in WGTP platform and 980 CNVs in 500K, in total 1447 CNVs

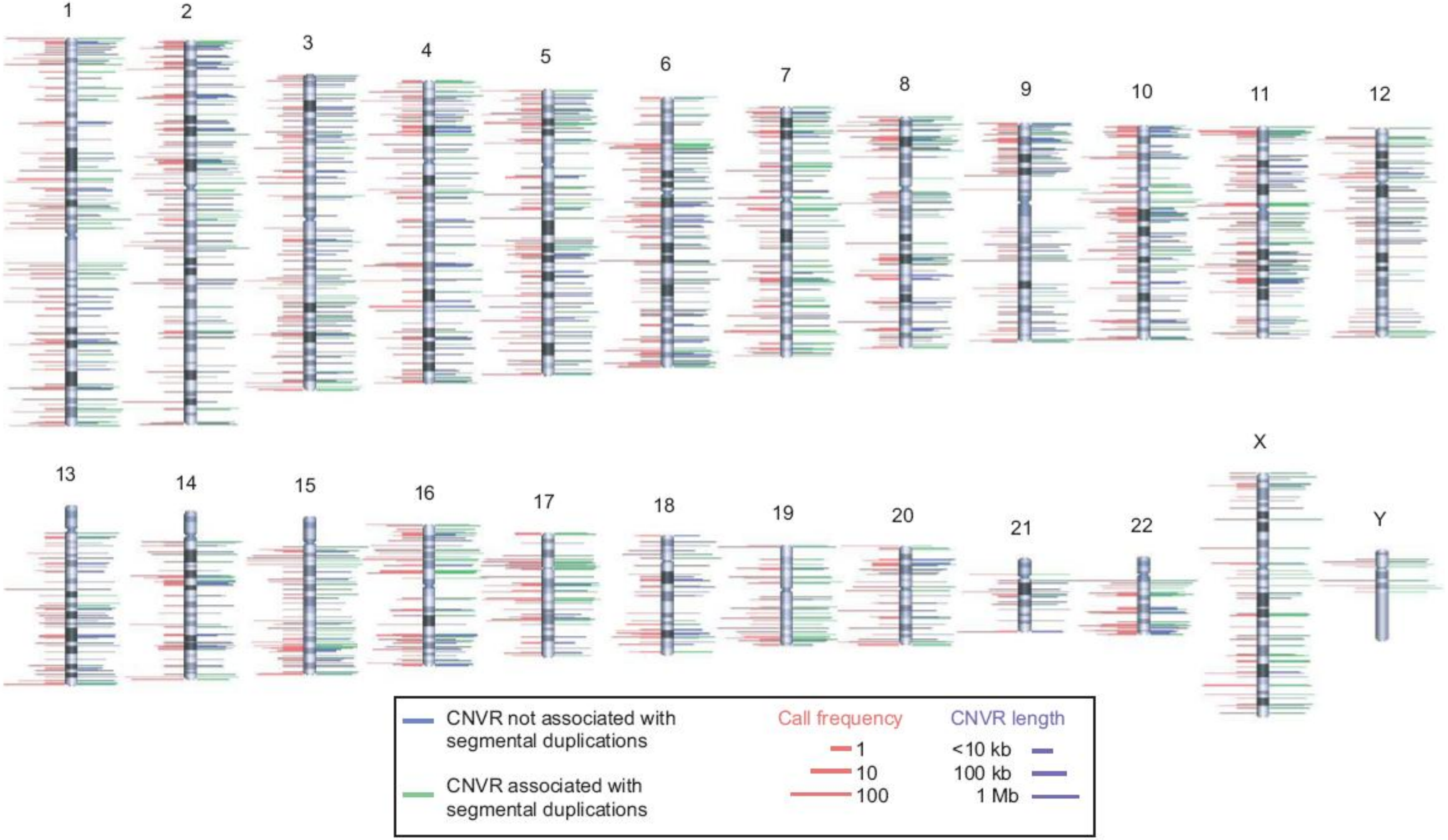
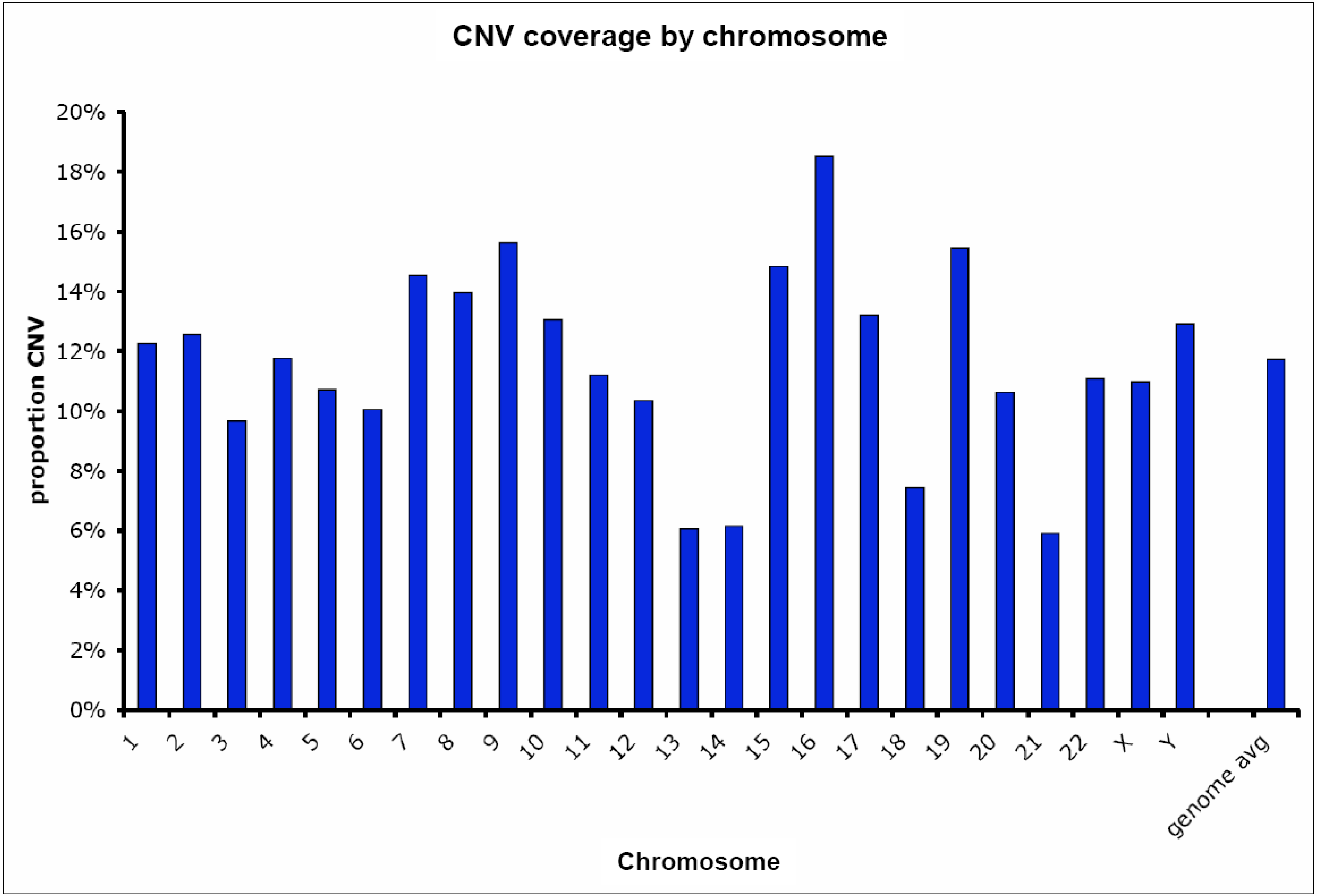


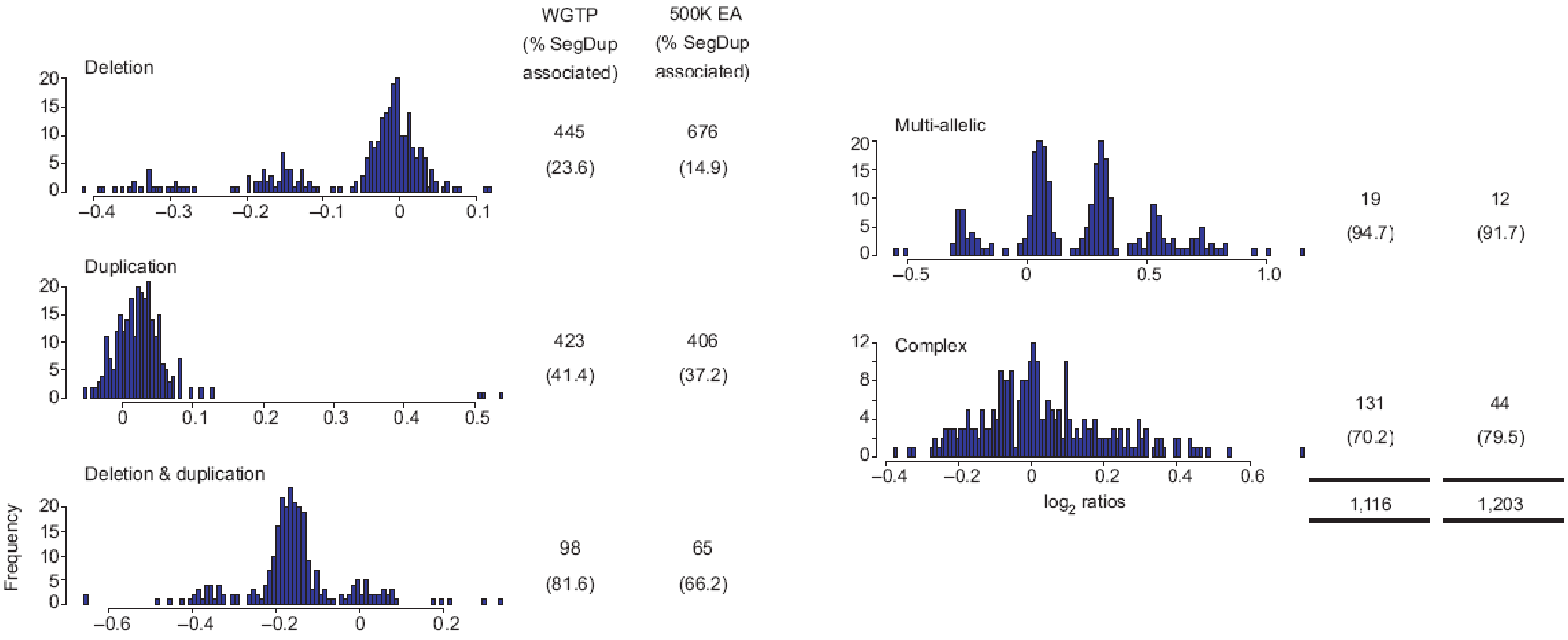
Figure 4 | Genomic distribution of CNVRs. The chromosomal locations of 1,447 CNVRs are indicated by lines to either side of ideograms. Green lines denote CNVRs associated with segmental duplications; blue lines denote CNVRs not associated with segmental duplications. The length of right-hand side lines represents the size of each CNVR. The length of left-hand side lines indicates the frequency that a CNVR is detected (minor call frequency among 270 HapMap samples). When both platforms identify a CNVR, the maximum call frequency of the two is shown. For clarity, the dynamic range of length and frequency are log transformed (see scale bars). All data can be viewed at the Database of Genomic Variants (http://projects.tcag.ca/variation/).


CNV coverage by chromosome
proportion CNV
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
genome avg
Chromosome

Gaps

- 164 out of 345 gaps in the built 35 assembly flanked or overlapped by CNVs

Classes of CNVs



	WGTP (% SegDup associated)	500K EA (% SegDup associated)
Deletion	445 (23.6)	676 (14.9)
Duplication	423 (41.4)	406 (37.2)
Deletion & duplication	98 (81.6)	65 (66.2)
Multi-allelic	19 (94.7)	12 (91.7)
Complex	131 (70.2)	44 (79.5)
	1,116	1,203

CNV formations

- 24% of 1447 CNVs are associated with segmental duplications
- 1. rearrangemets generated by non-allelic homologous recombination
- 2. not all annotated seqmental duplications are fixed in humans, but are CNVs
- 121 (500K) and 223 (WGTP) CNVs contains 100 bp similarities either end

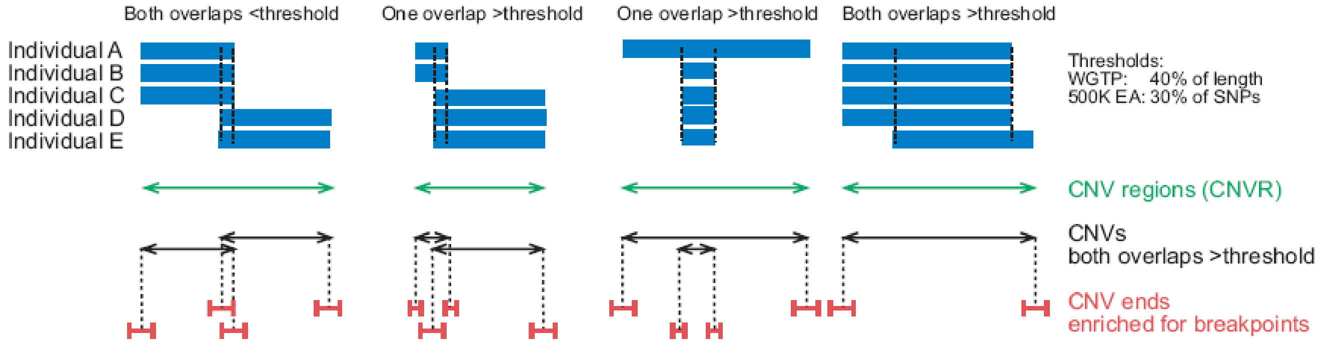


Figure 3 | Defining CNVRs, CNVs and CNV ends. Overlapping CNVs called in five individuals are shown schematically for four loci (in blue); dashed lines indicate overlap. Copy number variable regions (CNVRs) represent the union of overlapping CNVs (in green). Independent juxtaposed CNVs (in black) are identified by requiring that only individual-specific CNVs that overlap by more than a threshold proportion be merged. Intervals encompassing CNV breakpoints (in red) are defined using platform-dependent criteria (Supplementary Methods), and contain a significant paucity of recombination hotspots[76,77] (Supplementary Table 13), which results from the enrichment of segmental duplications within which fewer inferred recombination hotspots reside.

Genomic impact of CNV

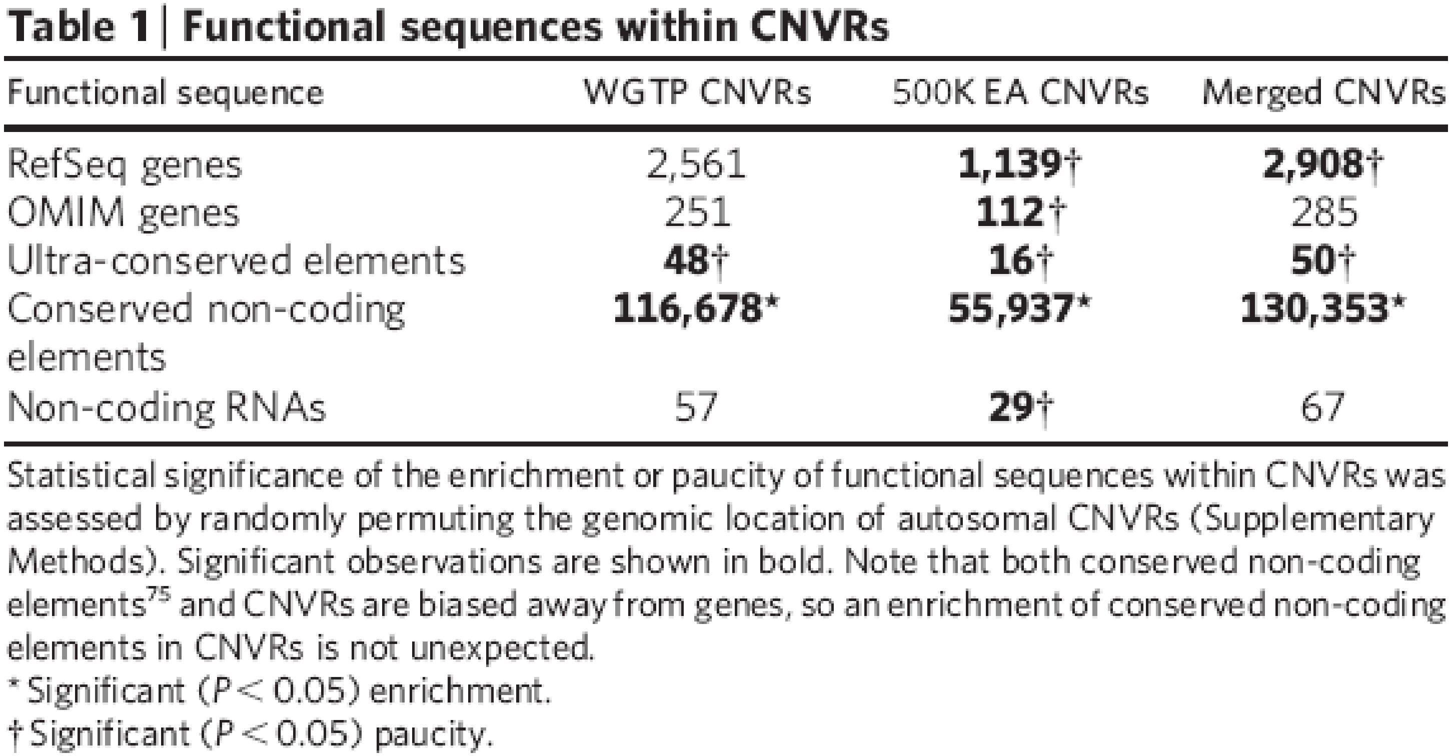


Table 1 | Functional sequences within CNVRs

Functional sequence	WGTP CNVRs	500K EA CNVRs	Merged CNVRs
RefSeq genes	2,561	**1,139†**	**2,908†**
OMIM genes	251	**112†**	285
Ultra-conserved elements	**48†**	**16†**	**50†**
Conserved non-coding elements	**116,678***	**55,937***	**130,353***
Non-coding RNAs	57	**29†**	67

Statistical significance of the enrichment or paucity of functional sequences within CNVRs was assessed by randomly permuting the genomic location of autosomal CNVRs (Supplementary Methods). Significant observations are shown in bold. Note that both conserved non-coding elements[75] and CNVRs are biased away from genes, so an enrichment of conserved non-coding elements in CNVRs is not unexpected.

* Significant ($P < 0.05$) enrichment.

† Significant ($P < 0.05$) paucity.

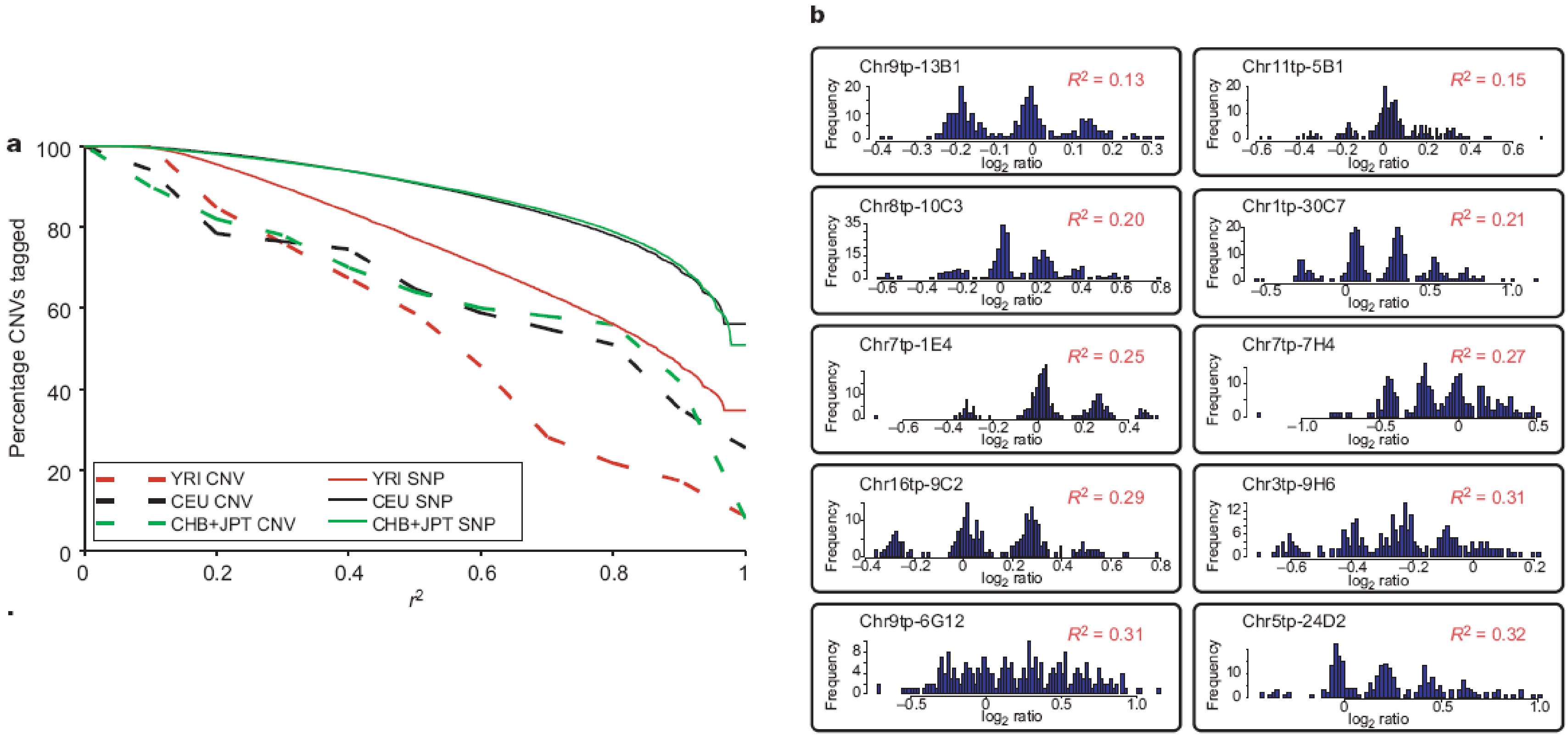
LD around CNVs

- Indirect methods to identify causative variants, such as co-segregation of linked markers in families and genetic association with markers in linkage disequilibrium with the causative variant, are considered to be blind to the nature of the underlying mutation. This raises the question of whether SNP-based whole-genome association studies have the same power to detect disease-related CNVs as for disease-related SNPs.
- Recent studies of linkage disequilibrium around CNVs have produced conflicting evidence as to the degree to which CNVs are ‘tagged’ by neighbouring SNPs.
- Comparing the proportion of variants tagged by a neighbouring SNP with an arbitrary threshold of r2.0.8 shows that whereas 75–80% of Phase I SNPs in non- African populations were tagged, only 51% of CNVs were tagged in the same populations.
- We considered three explanations for these observations of lowerapparent linkage disequilibrium around CNVs than SNPs.

LD around CNVs

- First, some duplications might represent transposition events that would generate linkage disequilibrium around the (unknown) acceptor locus but not the donor locus. One of the genotyped CNVs is known to be a duplicative transposition, but evidence from de novo pathogenic duplications strongly suggests a preference for tandem, rather than dispersed, duplications, regardless of whether the duplication is caused by non-allelic homologous recombination.
- Second, some CNVs might undergo recurrent mutations or reversions, especially tandem duplications which are mechanistically prone to unequal crossing over, causing reversions back to a single copy. However, duplications were not in lower linkage disequilibrium with flanking SNPs than were deletions.
- Finally, we considered that CNVs might occur preferentially in genomic regions with lower densities of SNP genotypes in HapMap Phase I. We found that CNVs are enriched within segmentally duplicated regions of the genome, in which thereis a paucity of genotyped SNPs owing to technical difficulties.

Patterns of linkage disequilibrium between CNVs and SNPs



Population clustering from CNV genotypes.

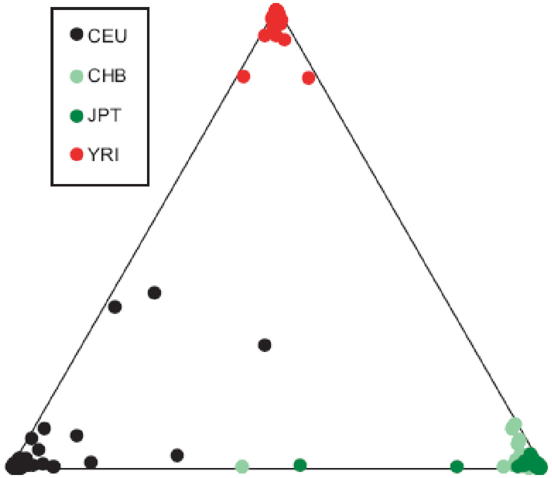
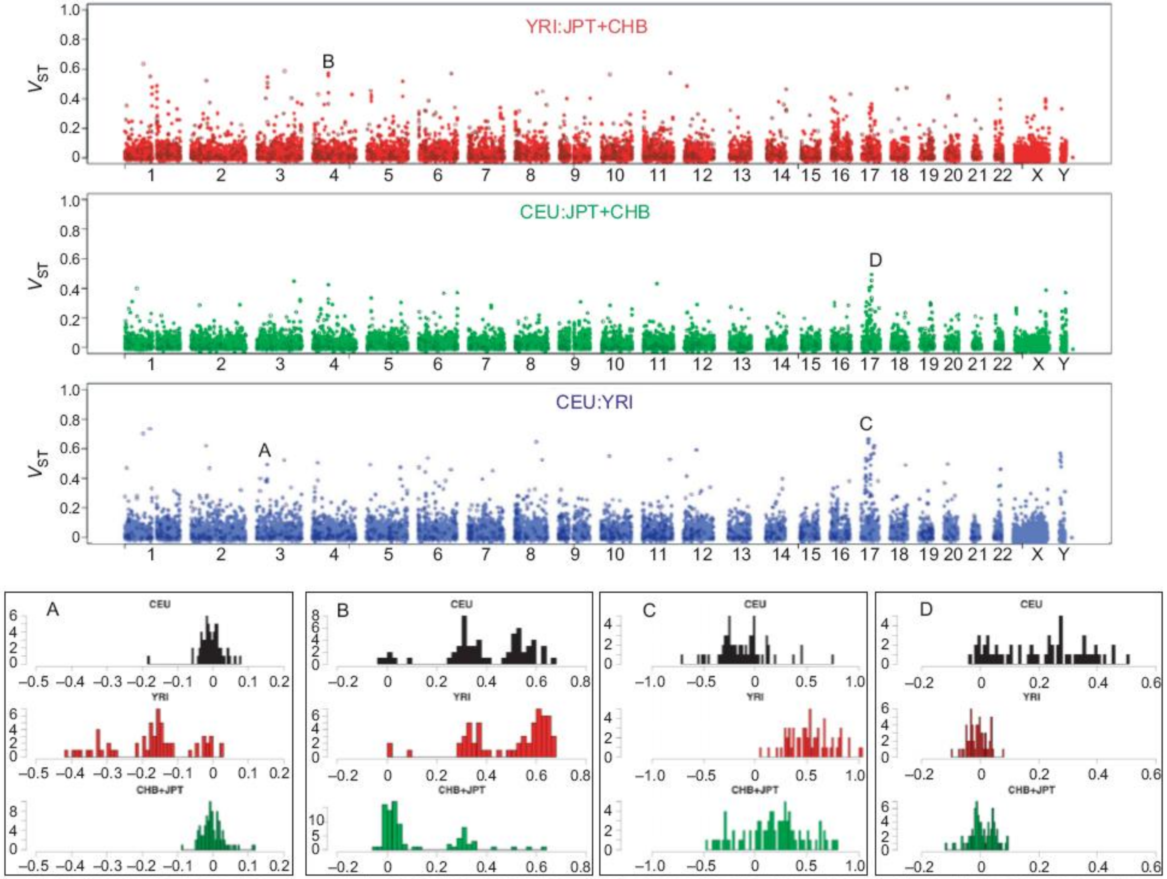


Figure 7 | Population clustering from CNV genotypes. A triangle plot showing the clustering of 210 unrelated HapMap individuals assuming three ancestral populations ($k = 3$). The proximity of an individual to each apex of the triangle indicates the proportion of that genome that is estimated to have ancestry in each of the three inferred ancestral populations. The clustering together of most individuals from the same population near a common apex indicates the clear discrimination between populations obtained through this analysis. The clustering was qualitatively similar to that obtained previously with a similar number of biallelic *Alu* insertion polymorphisms on different African, European and Asian population samples[60].

- A range of polymorphisms, including SNPs, microsatellites and Alu insertion variants, has been used to investigate population structure. To demonstrate the utility of copy number variation genotypes for population genetic inference we performed population clustering on 67 genotyped biallelic CNVs.

Population differentiation for copy number variation



Chr3: 46,770,243–46,947,493

Chr4: 69,126,468–69,296,476
UGT2B17

Chr17: 31,583,743–31,651,855
CCL3L1

Chr17: 41,635,804–41,722,491
Near *MAPT*

Discussion

- CNV assessment should now become standard in the design of all studies of the genetic basis of phenotypic variation, including disease susceptibility. Similarly important will be CNV annotation in all future genome assemblies.
- Our analysis of linkage disequilibrium between CNVs and SNPs gives us limited optimism that CNVs influencing risk to complex disease will be detected by such approaches. The tag SNPs that we have identified for specific CNVs can be used as proxies for these CNVs. Moreover, CNV-specific genotyping assays can be developed for CNVs for which tag SNPs are not readily identifiable but whose proximity to candidate genes warrants further characterization.
- Extrapolation based on existing data suggests that smaller deletions (<20 kb) are much more frequent than larger deletions (>20 kb), and the same may be true for duplications.
- CNV calls have been released at the Database of Genomic Variants (http://projects.tcag.ca/variation/) integrated with all other CNV data.