

Supplementary Methods

DNA samples and cell lines

Unless otherwise noted, all samples were obtained from the Coriell Institute for Medical Research. DNA sample NA10851 (cell line GM10851; 46,XY) was used as a hybridization control in all array-CGH experiments. NA10851 and NA15510 (GM15510; 46,XX) (Tuzun et al. 2005) were used as controls for FISH and CNV validation experiments. Both cell lines were karyotyped and found to have a normal chromosomal complement. The HapMap DNA used in most array and validation experiments was from the same preparation lot, with the exception of 53 Yoruban HapMap samples for which different lots were used for 500K EA and WGTP experiments (Supplementary Table 5). For HapMap karyotyping the majority of cell lines were obtained from D. Altshuler (Broad Institute).

WGTP Clone selection and verification

A total of 26,678 large insert clones were selected from the published “Golden Path” to cover the human genome in tiling path resolution. Clones were screened for T1 phage and *Pseudomonas* contamination and verified by finger printing and end sequencing. Clone information can be obtained from http://www.ensembl.org/Homo_sapiens/cytoview.

Clone positions were mapped onto the reference sequence (NCBI Build 35) using known accessions and end sequences. End sequencing verified the mapping position of 20,967 clones. 5,611 clones however failed to provide end sequence that could be mapped onto the reference sequence but were taken forward for array construction and further validation. The final validated set consists of 26,574 clones.

Preparation of clones and array spotting for WGTP

Large-insert clone DNA was isolated as described previously (Marra et al. 1997) (Humphray et al. 2001) and diluted to a final concentration of 1 ng/μl. For array construction, clone DNA was amplified in three separate DOP-PCR reactions using primers DOP1, DOP2 and DOP3 as described (Fiegler et al. 2003). After combining the appropriate DOP-PCR amplified products, a secondary PCR reaction using a 5'-amine modified primer designed to match the 10 bases at the 5'-end of each DOP-PCR primer was performed. 29 μl of 4 x microarray spotting buffer (1 M sodium phosphate buffer, pH 8.5, 0.001% sarkosyl) was then added to 88 μl of PCR

products and filtered by centrifugation at 2000 rpm for 10 min through 0.22 µm Millipore Multiscreen®-GV filter plates (Millipore).

Arrays were printed in a HEPA-filtered and humidity-controlled environment (40% to 45% RH) onto CodeLink™ activated slides (GE Healthcare UK Limited, UK) using MicroGrid 610 robots equipped with tungsten 10K split pins (Genomic Solutions) and stored desiccated at room temperature until use.

WGTP array hybridization

Test and reference DNA samples were differentially labeled using the Bioprime labeling kit (Invitrogen Carlsberg, CA, USA) with modifications of the nucleotide mix. Briefly, a 260 µl reaction is set up containing 300 ng of DNA and 120µl of 2.5x random primer solution. After denaturing the DNA for 10 minutes at 100°C, 30 µl of 10x dNTP mix (1 mM dCTP, 2 mM dATP, 2 mM dGTP, and 2 mM dTTP in TE buffer), 3 µl of 1 mM Cy5-dCTP or Cy3-dCTP (NEN Life Science Products), and 6 µl of Klenow fragment were added on ice to a final reaction volume of 300µl. The reaction is incubated at 37°C overnight and stopped by adding 30 µl of stop buffer supplied in the kit. Unincorporated nucleotides are removed by use of Microcon YM-300 Filter Devices (Millipore Co.), according to the suppliers' instructions.

Hybridizations were carried out on a Tecan HSTM Hybridization Station (Tecan Group Ltd.) using 63x20 mm chambers. Cy3 and Cy5 labeled DNAs were combined, precipitated together with 270 µg of human Cot1 DNA (Roche Diagnostics Ltd., UK) and resuspended in 165 µl of hybridization buffer (50% formamide, 5% dextran sulfate, 0.1% Tween 20, 2x SSC and 10 mM Tris/HCl, pH 7.4, 10 mM Cysteamine). Pre-hybridization solution was prepared simultaneously by precipitating 100 µl of herring sperm DNA (10 mg/ml, Sigma Aldrich, UK) and resuspending in 165 µl of hybridization buffer.

The prehybridization and hybridization solutions were then denatured for 10 min at 72°C. The prehybridization solution was injected into the Tecan chamber following instructions displayed on the station. During prehybridization (45 min at 37 °C), the hybridization solution was incubated at 37°C. Hybridization was carried out for 21 hours at 37°C with medium agitation frequency. Slides were washed with PBS/Tween 20/2mM cysteamine (wash time 0.30 min, soak time 0.30 min, 15 Cycles at 37°C), 0.1 x SSC (wash time 1.00 min, soak time 2.00 min, 5 Cycles at 54°C), PBS/Tween 20/2mM cysteamine (wash time 0.30 min, soak time 0.30 min, 10 Cycles

at 23°C) and HPLC water (wash time 0.30, soak time 0.00, 1 Cycle at 23°C) before drying for 2.30 min using nitrogen gas. All experiments were performed in duplicate with DNA labeling color reversal (dye swap).

WGTP data analysis

Array images were acquired using an Agilent laser scanner (Agilent Technologies, UK). Fluorescence intensities and log₂ ratio values were extracted using Bluefuse software (Bluegenome Ltd). Spots with low signal intensities ("amplitude"<100 in both channels) or inconsistent fluorescence patterns ("confidence" < 0.5 or "quality" = 0) were excluded before normalizing all log₂ ratio values by blocks (sub-arrays).

Fusion of dye-swap results and subsequent analyses were performed using custom Perl scripts. The median of all ratio values was calculated chromosome by chromosome for each individual hybridization. Each ratio was then normalized by the corresponding chromosomal median. The ratios of each clone in the two dye swap hybridizations were then averaged if replicate ratios differed by less than 50% (i.e. less than a difference of 0.585 on the log₂ scale).

The 68.2th percentile of the absolute values for all combined ratios was then calculated chromosome by chromosome as an estimation of the standard deviation (SDe). Clones reporting replicates different by more than eight times the SDe were excluded from further analysis.

Dye-swap experiments were accepted for CNV calling only if the following criteria were fulfilled: (1) Global SDe < 0.06; (2) Global clone exclusion rate < 10%; (3) Clone exclusion rate per individual chromosome < 20%. Clones contained in chromosomes with a corresponding chromosomal median of > 0.1 or < -0.1 were flagged and excluded from CNV calling. Thus, chromosomes X and Y in female versus male results, as well as chromosomal artifacts (aneusomies or very large imbalances) were automatically excluded from calling. See (Fiegler 2006) for more details.

Array images, raw intensities and normalized log₂ ratios can be downloaded from <http://www.sanger.ac.uk/humgen/cnv/data/>.

500K EA Array Content and Genome Coverage

The 500K EA (Early Access) arrays, a pre-commercial version of the GeneChip® Human Mapping 500K Array Set, contain 534,500 SNPs on two arrays and are used in conjunction with the whole genome sampling assay (WGSA). Each array is designed to interrogate SNPs residing on PCR amplicons of 100 – 1100 bp in length using either Nspl or Styl restriction enzyme digested, adaptor-ligated genomic DNA template. Each SNP is interrogated by 24 features, with six features each for the perfect match (PM) probes and the mismatch (MM) probes. To minimize the impact of oligonucleotide probes that may cross-hybridize to multiple independent DNA target sequences, probes were removed whose central 21 bases perfectly matched additional locations in the human reference genome (build 35). Probes corresponding to Nspl or Styl restriction fragments in which the enzyme recognition site contains a SNP were also removed. These steps trimmed the total content to 474,642 SNPs (88.8% of the original total) with a relatively minor effect on genome coverage.

Hybridization of HapMap samples onto 500K EA arrays using WGSA

DNA was purchased from Coriell Institute for Medical Research (Camden, NJ) as described in the main text. For target preparation prior to hybridization, 250ng of genomic DNA was digested with either Nspl or Styl (New England Biolabs, Ipswich, MA), followed by ligation with Nspl or Styl specific adapters (Affymetrix Inc., Santa Clara, CA). The ligated DNA was diluted 4 fold, and then PCR amplified using a primer designed to the adapter DNA. The PCR reactions were purified using a Qiagen MinElute 96 UF PCR Purification Plate (Qiagen Inc., Valencia, CA), and 60ug of this purified product was fragmented using .0.25 units of DNase I. The fragmented DNA was labeled using 0.5714mM DLR (Affymetrix Inc., Santa Clara, CA) and 105 U of terminal deoxy-nucleotidyl transferase (Tdt) for 2 hours at 37°C. Hybridization onto the 250K Nspl and 250K Styl EA arrays and subsequent steps were performed exactly as described by the manufacturer (www.affymetrix.com). Genotype calls were generated using the DM (Dynamic Modeling) calling algorithm with cutoff p-value 0.33 (Di et al. 2005). Intensity information for each probe set was extracted using Affymetrix software.

Data Analysis for 500K EA CNV detection

Algorithm Overview. The algorithm used to call CNVs using the 500K EA platform was developed to accurately define CNV regions using a large set of reference samples and is described in detail in a separate publication (Komura 2006). The algorithm contains three major parts: 1) Intensity pre-processing using an improved

version of Genomic Imbalance Map (GIM) (Ishikawa et al. 2005), including probe selection, noise reduction, normalization, and intensity ratio adjustment based on affinity differences between alleles of a SNP, 2) CNV extraction, which identifies CNVs from all pair-wise comparisons using a modified SW-ARRAY, and 3) A copy number inference step which utilizes signal ratios and SNP information to more precisely define CNV boundaries and the copy number within each region.

Intensity pre-processing and normalization. The intensity pre-processing and normalization steps are completed using the intensity ratio's from pair-wise comparisons separately for the 250K Nspl and 250K Styl data sets prior to merging. Signal variation due to GC content of the probes and restriction fragments, the difference in intensity of different SNP alleles, as well as other probe and fragment characteristics were normalized using the GIM algorithm (Ishikawa et al. 2005). The algorithm was applied separately to each sample pair, each array (Nspl and Styl), each genotype combination, and each restriction enzyme recognition site. The median ratio was set to one separately by dividing all the ratios with the median ratio for normalization. Following normalization, information from the Nspl and Styl arrays was merged to generate one file per sample.

Preliminary CNV extraction from multiple samples. CNV regions are first defined from the normalized intensity ratio's for each pair-wise comparison using SW-ARRAY. For any two DNA samples, a CNV was called if 4 SNPs on 3 restriction fragments showed an intensity ratio above 1.12 for insertions, or 0.89 for deletions, and a p-value of ≤ 0.01 (based on 5000 permutations of the data). CNVs were called from all pair-wise comparisons between the 270 HapMap samples (269 comparisons for each sample). The total CNV regions were then inferred by merging and summarizing all pair-wise comparisons. The 'CNV density' is defined as the detection rate of copy number differences between the test sample and the 269 sample reference set, and during the summarization step, only CNVs passing a 10% density cut-off were included for further analyses. Subsequent analyses were done separately for each CNV region.

Copy Number Inference and Identification of Diploid Samples. In order to identify gains vs losses, each sample's CNV copy number was calculated by comparison to only diploid samples. The diploid group was initially identified by a maximum clique algorithm as the largest group with the same copy number. To confirm that the 2 copy group was identified correctly, we used SNP genotypes to calculate the level of

heterozygosity and the A/B ratio for each CNV region with the assumption that single copy losses should be homozygous while three copy number regions should show heterozygous A-B ratios significantly different from one. For regions that did not satisfy these assumptions, the largest group remaining after removing the previously defined set was reselected as the diploid group.

Boundary assignment. After the diploid group was defined, CNVs and densities (the fraction of times a CNV is called when compared to the reference set) were calculated again using only the diploid group as the reference set. The boundaries were defined with varying confidence based on the density determinations. For example, a 90% boundary indicates the outermost inclusive SNP positions for a CNV in at least 90% of the pair-wise comparisons. The 10% boundaries are less stringent, and indicate the outermost inclusive SNP positions as called in at least 10% of the comparisons. Both 10% and 90% boundaries are listed in the Supplementary Tables.

Detection of Homozygous Deletions. SNP genotypes in regions of homozygous deletion are often not called. Therefore, in order to accurately detect this class of putative CNVs, a separate algorithm was developed based on the discrimination score for each SNP. The discrimination score measures the ability to discriminate between the perfect match and mismatch probes for any given SNP, and is typically very low in regions that are homozygous deleted. Stretches of probes with low discrimination scores were detected using SW-ARRAY and tested using homozygous deletions generated artificially by enzymatic digestion of DNA.

Merging of CNVs between platforms

CNV regions from each sample identified using the two platforms were merged into a single list of non-redundant CNV regions if they overlapped, regardless of the size of the overlap or the frequency with which the CNV is called on either platform. The borders listed are the outermost borders defined by either platform (the 10% density border was used for the 500K EA CNVs). Independent CNVs or CNV events were defined as intervals sharing at least 30% of SNPs for 500K EA or 40% of length for WGTP for both sample CNVs being compared. An interval containing a CNV end on 500K EA was defined as the interval between the 90% border and the closest SNP outside of the 10% border (maximum distance of 65kb). An interval containing a CNV end on WGTP was defined as being centered on the outside end of the outer clone within the CNV, and extending in either direction by either 3/8 of the length of the CNV or 65 kb, whichever was the shorter.

500K EA and WGTP data quality assessment

Quantitative PCR was used to independently test sets of CNVs called from replicate experiments with NA15510 (non-HapMap individual analysed for structural variation by fosmid read-pair analysis) and NA10851 (HapMap CEU trio offspring). With 500K EA, each pair-wise comparison for 3 replicate experiments resulted in an average proportion of CNV calls that are false positives of 6.1% (2 out of 33 CNV calls), and a false negative percentage of 18% (7 validated CNVs not called in any one replicate/38 total validated CNVs in the three replicates). Using the HapMap population as a reference with NA15510, the average proportion of CNV calls that were false positives was 2.3% (0.33 out of 15 CNV calls), and the false negative rate was 24% (3.3 out of 14.7). With WGTP, an average proportion of CNV calls that were false positives of 5% (3.4 out of 68.2) and a false negative percentage of 37.8% (58 non called out of 154 tested) in 5 replicate experiments (Supplementary Table 1) was found.

We also tested 50 CNV regions (25 from each platform) that were identified as singletons from only one platform, and 14 singletons called as CNVs from both platforms (Supplementary Table 4). All 14 singleton CNVs identified by both platforms were verified as true positives, while 38 out of 50 singleton CNVs called by only one platform were confirmed. Thus the false positive rate for singletons found by only one platform, which makes up one third of the CNVs represented in our data set, is 24%.

For reproducibility studies, 10 HapMap DNAs were analyzed in triplicate. On 500K EA, they were compared to the 270 HapMap samples: on average, 80% of CNVs called from these samples were called in all three replicates, 10% were called twice, and 10% were called in one out of the three replicates. On WGTP array, the three replicate experiments were made at different dates and using different batches of arrays. For each cell line, they were ranked by the number of called CNVs and then categorized in three sets A, B and C (with experiments A reporting with the highest number of calls, experiments C the lowest number of calls). On average, 73% of the CNVs called in experiment C were called in A and B, 14% were called only in A or only in B and 13% neither in A nor in B²⁷ (Supplementary Table 6).

Quantitative validation of CNVs

Experimental validation of CNV calls was performed using one or more of the following techniques: quantitative typing by real-time PCR using the ABI Prism 7500 Sequence Detection System (PE Applied Biosystems) or a Bio-Rad iCycler Thermal Cycler, quantitative multiplex PCR of short fluorescent fragments, or mass spectrometry. In all of these approaches, a test was considered confirmatory if a paired t-test based on > 3 replicates gave rise to a p-value of less than 0.05. Computational validation was based on overlap with entries in the Database of Genomic Variants <http://projects.tcag.ca/variation/>).

Quantitative PCR. Sequences for all primer pairs can be found in Supplementary Table 4. Real-time PCR reactions were performed using the ABI Prism 7500 Sequence Detection System (PE Applied Biosystems) and followed the manufacturer's guidelines and cycling conditions. For normalization, a VIC-labeled TaqMan probe to the RPPH1 locus (RNA moiety of RNase P) was used (PE Applied Biosystems). At least three replicate reactions were run for each primer pair and the comparative C_T method (User Bulletin #2; Applied Biosystems) was used to calculate the fold change at each locus between the test and reference samples. In addition, a t-test based on the ΔC_t values was used to determine the statistical significance of the result. All results that showed a fold change less than 0.9 or greater than 1.10 as well as a p-value < 0.05 were considered to be significant.

Mass Spectrometry validation. The determination of allele frequencies in test and reference samples was based on MOLDI-TOFF MassSpectroscopy of allele-specific primer extension products (MassArray Sequenom Inc., San Diego, California). All assays for the PCR and associated extension reactions were performed as suggested by the manufacturer. For each SNP position, differences in allele frequencies between a test and reference sample were calculated. For CNVs in samples with heterozygous genotypes, allele dosage ratios were compared to normal reference heterozygotes with no CNVs. DNA from individuals with homozygous genotypes in the CNV region were mixed with references homozygous for the alternate allele, and its allelic dosage ratio was compared with heterozygous, diploid references. The appropriate mixture procedure in these samples was verified by distinct CNV-free loci. In all cases, a p value > 0.05 (t-test) was considered significant.

PCR validation of homozygous deletions. Homozygous deletions were verified using PCR, with 34 rounds of amplification. One to three samples called for the deletion were examined along with 3 diploid reference samples. Visual inspection of agarose gels stained with ethidium bromide was used to call the presence or absence of the PCR product.

Fluorescence in situ hybridization and karyotyping

Lymphoblastoid cells were cultured at 37°C in RPMI 1640 medium supplemented with 10% Fetal Bovine Serum and 1% L-glutamine (200 mM), exposed for 15 minutes to Colcemid (final concentration 0.05ug/ml) and harvested according to standard cytogenetic protocol, using hypotonic 0.075M KCl and Carnoy's fixative. Metaphase preparations were made by dropping the fixed cell suspension onto pre-cleaned slides. Subsequent steps for FISH and band calling have been described (Iafrate et al. 2004). We examined a minimum of 10 metaphases for each experiment by multicolor fluorescence microscopy and imaging using Cytovision software (Applied Imaging, Inc., Santa Clara, CA, USA).

SNP genotyping failures

SNP content in CNVs. The redundant-unfiltered set of HapMap SNPs was downloaded from the HapMap website (<http://www.hapmap.org/>). In this set of SNPs, one can find SNPs that have fulfilled the Quality Control criteria (termed QC+ from now on) together with those that have not (termed QC-). The assignment of the QC-status to a given SNP is given according to the following criteria: QC-p => pass rate < 80%; QC-d => >1 duplicate discrepancy; QC-h => Hardy-Weinberg p-value < 0.0001; QC-m => >1 Mendel inconsistencies; QC-s => fail-flagged by submitter.

In order to compare the relative amount of SNPs within copy number variant regions (CNVs), we performed permutation tests. We measured every type of SNP within the identified CNVs (i.e. observed value). In each permutation we randomly assigned the position of the set of CNVs along the chromosome (excluding the centromeres and avoiding overlap between simulated segments). Each experiment consisted of n= 1000 random permutations. The comparison between the observed value and the mean of the 1000 replicates represents the trend (enrichment or impoverishment) while the significance is calculated as the number of times that the value in the

replicate equals or exceeds the observed value divided by the total number of permutations+1. Bonferroni correction was applied taking into account that seven different comparisons were performed with the same set of simulated data (corrected p-value = 0.05, /#comparisons = 0.00714).

Comparison of SNP failures in CNV gains and losses. It is known that CNV losses are enriched in Mendelian errors in trios and in SNPs not in HW equilibrium within a population. To test whether CNV gains and losses have different patterns of QC+ and QC- SNPs we looked at the events called in the different individuals. We calculated the percentage of each type of SNP (number of QCtype / total number of SNPs) in such events. The rate of percentages is calculated to compare gains and losses. Under the null hypothesis the rate equals 1. To test this, we assume that the number of SNPs is distributed according to a Poisson and we use a general linear regression model (including the total number of SNPs as an offset) to compare the two poisson distributions.

We have performed the analysis of all SNP types at the same time (multivariate approach) using a principal component analysis. We summarize the results in a plot where the squares with the QC labels represent the projected position of the original variables in the two principal components, and tiny circles represent individuals' projections in the same factors. The ellipses show the inertia of points belonging to each group (i.e., gains and losses) indicating whether the principal component analysis is able to discriminate between CNV gains and losses. The analyses were performed using the *ade4 R-package*.

The global analysis of SNPs in CNVs called by the two platforms reveals that the 500K EA platform is better able to discriminate between CNV gains and losses based on their SNP content (Supplementary Figure 13). This is probably due to the better refinement of CNV boundaries achieved by the 500K EA platform.

Genotype calls

Sixty seven non-redundant common (minor allele frequency greater than 5%) biallelic CNVs suitable for genotyping were identified manually using different procedures on the WGTP and 500K EA platforms after automated clustering methods proved not to be robust. Thirty-six common biallelic CNVs that could be genotyped were identified on the WGTP platform and thirty-one on the 500K EA platform. Two procedures were used to cluster intensity ratios into genotypes: Kmeans and Partitioning Around

Medoids (PAM) (Kaufman 1990). For CNVs genotyped using WGTP data, genotypes were generated for the reference individual (NA10851) by adding a \log_2 ratio of zero for this individual into each clustering. In four instances, two WGTP clones detect the same CNV and were better able (higher genotype pass rate) to cluster genotypes using bivariate clustering, than univariate clustering based on a single clone (e.g. Supplementary Figure 14). Low quality genotypes were identified and excluded by calculating either the relative Euclidean distance ($D_{\text{next_nearest}}/D_{\text{nearest}}$) between nearest clusters (Kmeans), or the silhouette width (PAM), with minimal threshold values of 1.7 and 0.5 respectively. These threshold values were determined empirically to generate genotypes of similar quality to SNPs, as judged by the Mendelian error rate in CEU and YRI trios (Conrad et al. 2006).

Population genetic and statistical analysis

Pairwise LD (r^2) was estimated between a CNV and all filtered non-redundant Phase I HapMap SNPs in an interval stretching from 500 kb 5' to 500 kb 3' to the likely maximal extent of the CNV using HaploView (Barrett et al. 2005). The default quality control parameters of HaploView were used (e.g. Hardy-Weinberg Equilibrium p value >0.001), which excluded 2/67 CNVs from the LD analyses. Population clustering was performed using STRUCTURE (Pritchard et al. 2000). F_{ST} was estimated using the method of Weir and Cockerham (Weir and Cockerham 1984). V_{ST} estimates the proportion of variance of the quantitative intensity ratios (from both WGTP and 500K EA data) attributable to variation between populations, as opposed to variation within populations, by considering $(V_T - V_S)/V_T$ where V_T is the variance in \log_2 ratios apparent among all the unrelated individuals from both populations and V_S is the average of the variance in \log_2 ratios within each population, weighted for population size. Validating V_{ST} against F_{ST} via the 67 genotyped biallelic CNVs confirmed that the two measures of population differentiation are very highly correlated ($R^2=0.91$, Supplementary Figure 15). The large size of the CNVs and the uncertainty in the SNP genotyping calls within the CNV precluded a standard calculation of Extended Haplotype Homozygosity (EHH) and REHH (Sabeti et al. 2002; Walsh et al. 2006). We therefore coded each genotyped CNV as a SNP, located this at each boundary of the CNV, and then performed an independent REHH analysis of HapMap Phase I SNPs 500 kb upstream and downstream of the CNV using the program Sweep (<http://www.broad.mit.edu/mpg/sweep/resources.html>).

References:

- Barrett, J.C., B. Fry, J. Maller, and M.J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**: 263-265.
- Conrad, D.F., T.D. Andrews, N.P. Carter, M.E. Hurler, and J.K. Pritchard. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38**: 75-81.
- Di, X., H. Matsuzaki, T.A. Webster, E. Hubbell, G. Liu, S. Dong, D. Bartell, J. Huang, R. Chiles, G. Yang, M.M. Shen, D. Kulp, G.C. Kennedy, R. Mei, K.W. Jones, and S. Cawley. 2005. Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics* **21**: 1958-1963.
- Fiegler, H., P. Carr, E.J. Douglas, D.C. Burford, S. Hunt, C.E. Scott, J. Smith, D. Vetrie, P. Gorman, I.P. Tomlinson, and N.P. Carter. 2003. DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer* **36**: 361-374.
- Fiegler, H.e.a. 2006. Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Research* **in press**.
- Humphray, S.J., S.J. Knaggs, and I. Ragoussis. 2001. Contiguation of bacterial clones. *Methods Mol Biol* **175**: 69-108.
- Iafate, A.J., L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, and C. Lee. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36**: 949-951.
- Ishikawa, S., D. Komura, S. Tsuji, K. Nishimura, S. Yamamoto, B. Panda, J. Huang, M. Fukayama, K.W. Jones, and H. Aburatani. 2005. Allelic dosage analysis with genotyping microarrays. *Biochem Biophys Res Commun* **333**: 1309-1314.
- Kaufman, L.a.R., P.J. 1990. *Finding groups in data: an introduction to cluster analysis* Wiley, New York.
- Komura, D.e.a. 2006. Genome-Wide Detection of Human Copy Number Variants using High Density DNA Oligonucleotide Arrays. *Genome Research* **in press**.
- Marra, M.A., T.A. Kucaba, N.L. Dietrich, E.D. Green, B. Brownstein, R.K. Wilson, K.M. McDonald, L.W. Hillier, J.D. McPherson, and R.H. Waterston. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res* **7**: 1072-1084.
- Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945-959.
- Sabeti, P.C., D.E. Reich, J.M. Higgins, H.Z. Levine, D.J. Richter, S.F. Schaffner, S.B. Gabriel, J.V. Platko, N.J. Patterson, G.J. McDonald, H.C. Ackerman, S.J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E.S. Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832-837.
- Tuzun, E., A.J. Sharp, J.A. Bailey, R. Kaul, V.A. Morrison, L.M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M.V. Olson, and E.E. Eichler. 2005. Fine-scale structural variation of the human genome. **37**: 727-732.
- Walsh, E.C., P. Sabeti, H.B. Hutcheson, B. Fry, S.F. Schaffner, P.I. de Bakker, P. Varilly, A.A. Palma, J. Roy, R. Cooper, C. Winkler, Y. Zeng, G. de The, E.S. Lander, S. O'Brien, and D. Altshuler. 2006. Searching for signals of evolutionary selection in 168 genes related to immune function. *Hum Genet* **119**: 92-102.
- Weir, B.S. and C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **47**: 264-279.

