Section 2: Supplementary Methods for "Origins and functional impact of copy number variation in the human genome"

Donald F. Conrad[1], Dalila Pinto[2], Richard Redon[1,3], Lars Feuk[2,4], Omer Gokcumen[5], Yujun Zhang[1], Jan Aerts[1], T. Daniel Andrews[1], Chris Barnes[1], Peter Campbell[1], Tomas Fitzgerald[1], Min Hu[1], Chun Hwa Ihm[5], Kati Kristiansson[1], Daniel MacArthur[1] Jeff MacDonald[2], Ifejinelo Onyiah[1], Andy Wing Chun Pang[2], Sam Robson[1], Armand Valsesia[1], Klaudia Walter[1], John Wei[2], Wellcome Trust Case Control Consortium, Chris Tyler-Smith[1], Nigel P. Carter[1], Charles Lee[5], Steve Scherer[2,6], Matthew E. Hurles[1]

[1] Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

[2] The Centre for Applied Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children, MaRS CentreEast Tower, 101 College Street, Room 14-701, Toronto, Ontario M5G 1L7, Canada.

[3] Inserm UMR915, L'institut du thorax, Nantes, France

[4] Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden

[5] Department of Pathology, Brigham and Womens Hospital and Harvard Medical School, Boston, MA, USA

[6] Department of Molecular and Medical Genetics, Faculty of Medicine, University of Toronto M5S 1A8, Canada.

September 28, 2009

Contents

1 CNV Discovery

1.1 Samples

For the discovery phase, we analysed 20 CEU HapMap samples, 20 YRI HapMap samples and one Polymorphism Discovery Resource sample for CNVs by array-CGH (Table 2.1). DNAs used for discovery were first screened on a lower resolution tiling-BAC array to exclude any with large-scale somatic chromosomal artifacts.

Sample ID	Test/Reference	Population	Sex	Trio member
NA12156	Test	CEU	Female	mother
NA12878	Test	CEU	Female	child
NA12239	Test	CEU	Female	mother
NA11993	Test	CEU	Female	mother
NA12004	Test	CEU	Female	mother
NA12006	Test	CEU	Female	mother
NA11995	Test	CEU	Female	mother
NA12044	Test	CEU	Female	mother
NA06985	Test	CEU	Female	mother
NA10851	Reference	CEU	Male	child
NA15510	Test	PDR	Female	-
NA18517	Test	YRI	Female	mother
NA19129	Test	YRI	Female	child
NA19240	Test	YRI	Female	child
NA18505	Test	YRI	Female	mother
NA18502	Test	YRI	Female	mother
NA19099	Test	YRI	Female	mother
NA18523	Test	YRI	Female	mother
NA18508	Test	YRI	Female	mother
NA18858	Test	YRI	Female	mother
NA18861	Test	YRI	Female	mother
NA07045	Test	CEU	Female	mother
NA11931	Test	CEU	Female	mother
NA12489	Test	CEU	Female	mother
NA12749	Test	CEU	Female	mother
NA12828	Test	CEU	Female	mother
NA12776	Test	CEU	Female	mother
NA11894	Test	CEU	Female	mother
NA12287	Test	CEU	Female	mother
NA07037	Test	CEU	Female	mother
NA12414	Test	CEU	Female	mother
NA18511	Test	YRI	Female	mother
NA18909	Test	YRI	Female	mother
NA19114	Test	YRI	Female	mother
NA19147	Test	YRI	Female	mother
NA19190	Test	YRI	Female	mother
NA18907	Test	YRI	Female	mother
NA18916	Test	YRI	Female	mother
NA19108	Test	YRI	Female	mother
NA19257	Test	YRI	Female	mother
NA19225	Test	YRI	Female	mother

Table 2.1: **Description of samples used in discovery experiment.** Abbreviations: CEU, CEPH Europeans from Utah; YRI, Yoruba from Idaban Nigeria, PDR, Polymorphism Discovery Resource. Reference: common reference sample used on all CGH hybridizations.

1.2 Array design

The CNV discovery platform consisted of 20 Nimblegen HD2 chips, each chip containing 2.1M probes. Each chip is further subdivided into 3 equal-sized subarrays containing about 726k probes. The probes were designed according to Nimblegen's standard protocol, with the exception that probes with up to 100 close matches in the genome were included, allowing greater coverage of segmentally duplicated regions. The final design provides 1 probe per 56bp median density across the genome. The layout of the entire chipset is included as Appendix A at the back of this document.

For quality control purposes, three sets of special probes were used for this experiment. Over 1400 exons of known dosage sensitive genes were identified, and a single probe placed in each exon. This set of control probes was printed on each subarray. Second, each successive chip overlaps the previous one by about 14,000 probes, equivalent to the average number of probes per megabase. Third X-linked probes were printed on each subarray, which allowed empirical measurement of experimental dose-response for each of our male-female cohybridizations.

1.3 Overview of CNV map construction

In brief, the construction of the CNV map entailed the following steps:

- QC of hybridizations.
- Normalize log2 ratios.
- Segment log2 ratios from each sample. This was done running the GADA algorithm (Pique-Regi et al., 2008) using the options "-M 10 -T 10 -a 2.5".
- Filter non-CNV segments using intensity thresholds. Merge remaining CNV "calls" within each sample. Adjacent calls of the same direction (gain or loss) are merged if both: the distance between calls is less than 10kb, and the distance between calls is less than 10% of the size of the largest of the two calls.
- Post-calling QC: remove sub-arrays with oversegmentation, split calls spanning gaps and centromeres. Relabel calls using probe midpoints.
- Merge CNV calls into CNV regions (CNVRs) and loci using hierarchical clustering.

Full details on algorithms, parameter settings for various steps, and choice of QC statistics can be found in the following sections.

1.4 QC of hybridizations

The quality of all subarrays was measure by two summary statistics, mad.d1r and the log2 ratio of x-linked probes. The threshold for selecting an experiment for re-analysis was:

- mad.d1r> 0.23
- sd/mad.d1r < 1.8
- median $X < 0.35$ [median (X controls) median (autosomal controls)]
- X response < 2 [median X / (mad.d1r of all test probes)]

and 81 chips were selected for repeating once, and 5 selected to be done twice. The statistic "mad.d1r" is defined as the median of absolute differences in copy number measurement between neighbouring probes. In the R language it would be calculated as `median(abs(diff(x)))` .

1.5 Normalization

The normalization pipeline begins with the q-spline normalized data provided by Nimblegen; q-spline normalization transforms the red and green channel data from a single array experiment to the identical distribution. Log2 ratios are then obtained at each probe position as Cy3/Cy5. In-house, we corrected for GC effects by fitting a model with linear and quadratic effects of GC content to the log2 ratios, separately for each subarray. We take the GC percentage in a 300bp window centered on each probe as our data for this analysis, using NCBI36 as our reference genome sequence. Finally, long-range spatial auto-correlation in log2 ratios (the 'wave effect') is modeled and removed using the method described in Marioni et al. (2007).

All of the data in this dataset are generated from female samples co-hybridized with a common male reference. For X-linked probes outside of PAR1 and PAR2, we take a slightly different normalization approach, to remove the effect of the reference sample. The raw data for non-PAR1/PAR2 X-linked probes are separated from all other probes and normalized as above (q-spline, GC, wave). Following this, the population median log2 ratio at each probe is calculated, and this value is subtracted from each probe in turn.

1.6 Background on segmentation algorithms

We initially explored two calling algorithms, CBS and GADA, as potential single-sample calling algorithms (Venkatraman and Olshen, 2007; Pique-Regi et al., 2008). GADA is a newly described method that relies upon a novel piecewise-constant vector representation of the intensity data to facilitate extremely fast matrix-based breakpoint finding. The method consists of two primary steps. The first step is a Bayesian learning process which generates a list of candidate breakpoints and segment means while trying to strike an optimal balance between model fit (measured as residual sum of squares) and model sparseness (the number of breakpoints). The Bayesian learning process is driven by a prior parameter, a, which summarizes the user's prior belief in the appropriate degree of segmentation (larger a leads to greater segmentation). After this initial segmentation process, a "t" statistic is calculated

for each segment, which is a function of the segment mean and variance. The second step is then a backwards elimination process which removes breakpoints with a level of significance (t statistic) less than some user-defined threshold, T.

Under the null hypothesis that a segment is copy normal, the t statistic should be a draw from a standard normal distribution. In algorithm comparisons using real (Affymetrix 500K, Illumina 550) and simulated data, Pique-Regi et al. (2008) found that a critical value of $t = 4.8$ provided comparable results to a CBS analysis with $\alpha = 0.01$. Based on the relative speed and performance of the two methods we elected to use GADA for the final CNV segmentation.

1.7 Data-driven intensity thresholding

The most basic implementation of GADA can produce a set of segments but does not classify segments into gains and losses. Our approach to the problem is to select intensity thresholds above/below which a segment is considered a CNV. Based on analysis of X-chromosome probes on each subarray, we estimated that a typical log2 ratio for a deletion heterozygote will be -0.55. Using an additive background model for the log2 ratios, ie.

$$y = log_2(\frac{a+c}{b+c})$$

where a is the intensity of the target, b the intensity of the reference, and c the background, we used the X-chromosome observations to find $c = 1.125$. This value of c was then used to calculate expected log2 ratios for other critical relative copy number comparisons (Table 2.2).

Relative copy number	expected log2 ratio
1/2	-.55
3/4	-.31
4/5	-.25
2/2	0
5/4	.25
4/3	.31
3/2	.40
4/2	.71

Table 2.2: **Estimated log2 ratios including additive background, calculated for several canonical relative copy numbers**

We appraised the usefulness of the numbers calculated in Table 2.2 by examining the distribution of segment log2 ratios from a whole-genome GADA analysis. In Figure 2.1, we can see discrete peaks near -.55 and .71 that appear to correspond to deletion heterozygotes and duplication homozygotes, respectively. There is no clear peak for duplication heterozygotes;

this could indicate a flaw in the use of an additive background model, or it could reflect the greater difficulty of calling such events. A key consideration was that we wanted to call some fold-change CNVs, such as might be observed at duplication of DNA segment already present in 4 copies. Combining analyses of replicate experiments and predicted dose-response for various classes of CNV, we settled on simple intensity thresholds of -.25 for losses and 0.1 for gains.

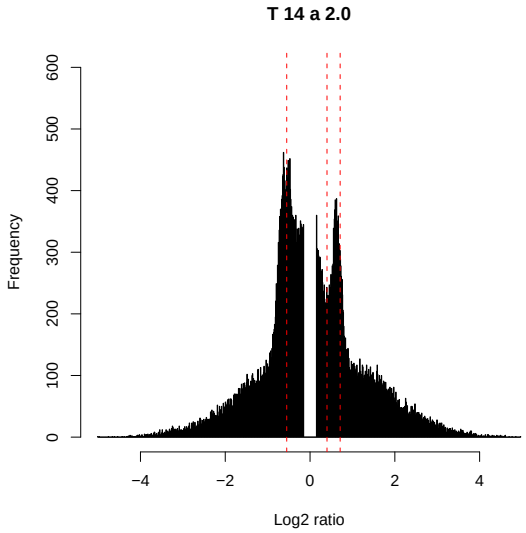


Figure 2.1: Distribution of segment log 2 ratios from whole-genome GADA analysis of 40 samples. The expected location of log2 ratios corresponding to 1/2, 3/2 and 4/2 relative copy number comparisons are marked with vertical red lines, left to right. Note that the data appear to be a mixture of CNV at unique sequence (producing discrete peaks corresponding to a small number of relative copy number states) and CNV at high-copy number sequence (that produces a normal distribution of relative copy number).

1.8 Post-calling QC and processing

We have found that one of the most sensitive metrics for detecting a poor CNV experiment is the number of CNV calls made in that experiment (Figure 2.2). Let n_{ij} be the number of CNV calls made on subarray i in individual j. We removed 104 subarrays where $n_{ij} > 3 * MAD_i + m_i$, where MAD_i and m_i are the mean absolute deviation and median number of calls on subarray i, respectively. In support of the belief that these "oversegmented" arrays represent technical artifacts, we observed that such subarrays tend to come in bunches of 3 within an individual (corresponding to a single chip), and 56 of 104 subarrays removed came from only 4 individuals.

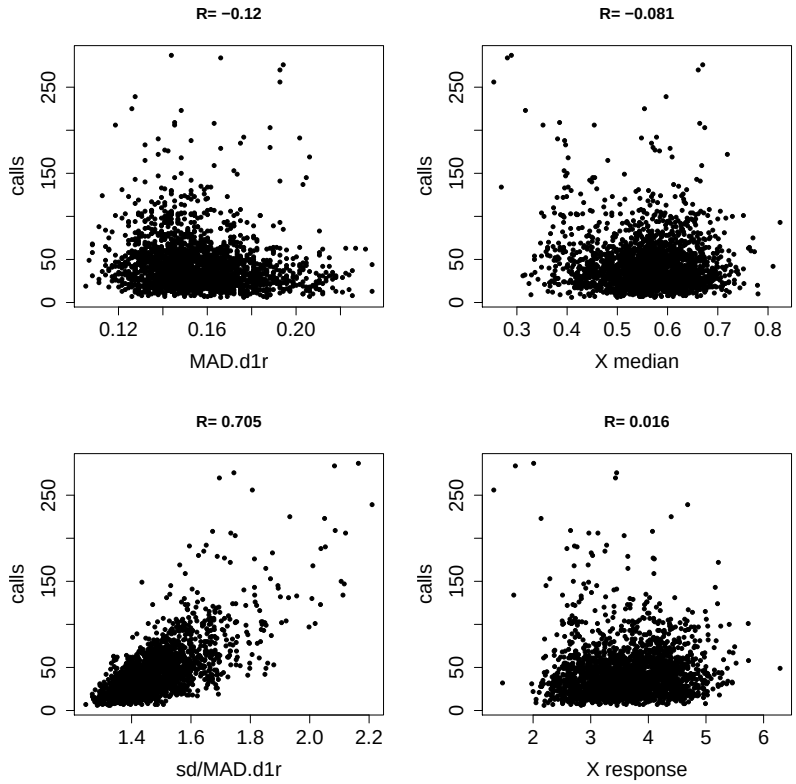


Figure 2.2: **Relationship between number of calls per subarray and 4 QC statistics.** The Pearson correlation between the number of calls per subarray and the QC statistic is written above each plot.

In addition, there were 511 calls that span gaps or centromeres in the reference sequence. Each of these calls were split into two daughter calls, and daughter calls were retained if their length $>$ 400bp.

1.9 CNV region definitions

As segmentation is done one sample at a time, we required an approach to identify and combine calls from different indiviudals that correspond to the same mutation event. While one could envision statistical approaches to the problem we chose to implement an algorithmic approach predicated on hierarchical clustering. At the top level of the hierarchy, all contiguous bases overlapping at least 1bp of a CNV call are merged into a "CNV region" (CNVR). Within each CNVR we further define CNVs with the following algorithm:

1. Calculate reciprocal overlap (RO) between all remaining calls.

2. Identify pair of calls with greatest RO. If RO $>$ threshold, merge and create a new CNV (CNV). If not, exit.

3. Continue adding unclustered calls to the CNV, in order of best overlap. In order to add a call, the new call must have > threshold to all calls within CNV to be added. When no additional calls may be added, move to next step.

4. If calls remain, return to 1. Otherwise exit.

We used a reciprocal overlap threshold value of 0.51 for constructing CNVs. CNV events are labeled in a way that indicates their CNVR membership. For example, the CNV name "CNVR45.2" refers to the second event in CNVR45.

1.10 Analysis of replicate pairs

As the experimental unit in these repeat hybridizations is a chip (consisting of 3 subarrays), some subarrays were replicated despite passing the QC on the raw hyb data. In total there were 42 replicate subarray experiments that passed our QC process; these were generated from 14 individuals and represented 29 different subarrays (that is, 13 subarrays were replicated on multiple individuals). In theory it is possible to analyze these replicate experiments to gain a better understanding of the relationship between calling parameters and CNV call reproducibility.

To gain a better understanding of the impact of calling parameters on segmentation performance, we ran GADA on the entire data set over a grid of $a(0.2, 0.7, \ldots, 2.7)$ and $T(1, 2, \ldots, 17)$. For each combination of parameter settings we calculated numerous summaries of the data which are illustrated in the following figures. In the range of T that we would plausibly consider ($T = 6 \ldots 10$) the value of a doesn't actually make much of an impact on the final results with our data.

The summaries considered can be loosely split into those measuring statistical aspects of calling (false negative rate, false positive rate, false discovery rate, overlap in calls between replicate samples), summaries of the extent of CNV (proportion of probes called in CNVs, projected number of calls per experiment, proportion of singleton calls) and call properties (median length of calls, median intensity of calls, ratio of gains to losses).

We examined a few ways of estimating the false positive and negative rates of CNV calling using replicate pairs; one of the main considerations is whether to consider CNVR or probe as the unit of testing. In Figures 2.3 and 2.4 we report the Jaccard measure of overlap between replicate pairs measured on both the probe level and the CNVR level:

$$\frac{A \cap B}{A \cup B}$$

False positive rate, negative rate and FDR are all estimated using three data points: the Jaccard overlap between replicates (probe level version), the average proportion of probes called as CNV across replicates, and the (unknown) average proportion of sequence contained in CNV within a single genome. We have arbitrarily used 1.5% as our estimate of this last quantity, which we feel is a conservative number; the estimated false positive rate will be underestimated if this number is too large.

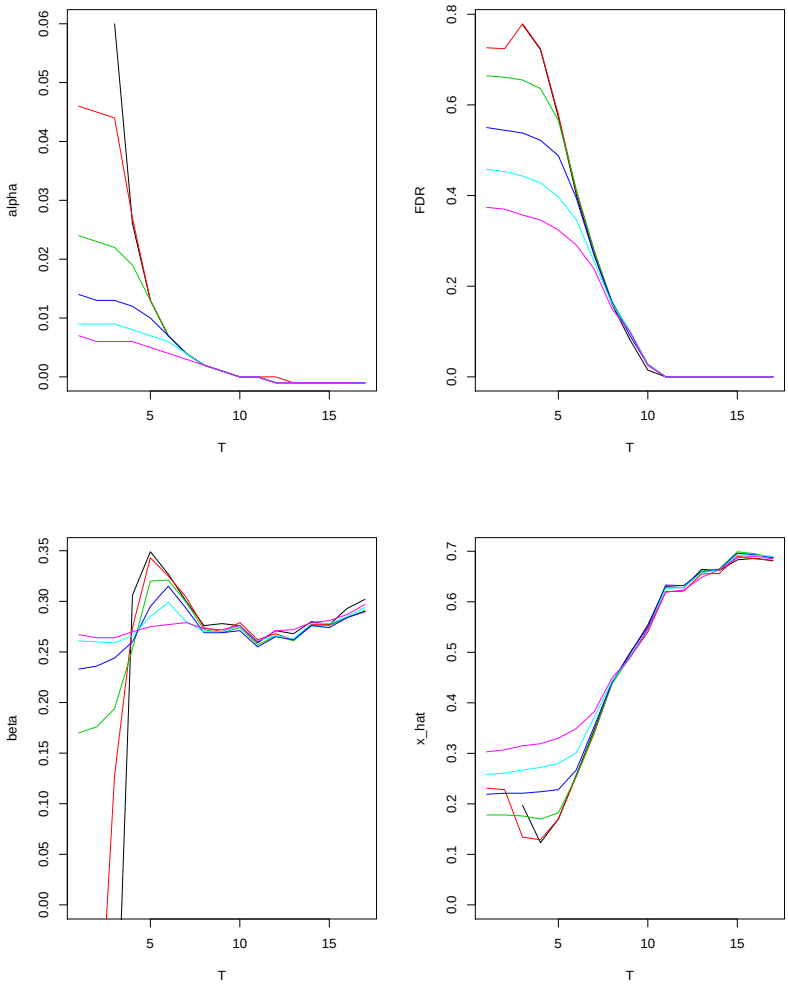


Figure 2.3: **Summary statistics of replicate pairs 1.** Each colored line represents a different value of a. Clockwise, from upper left: false-positive rate (alpha), false discovery rate, jaccard overlap on CNVR basis, false-negative rate (beta).

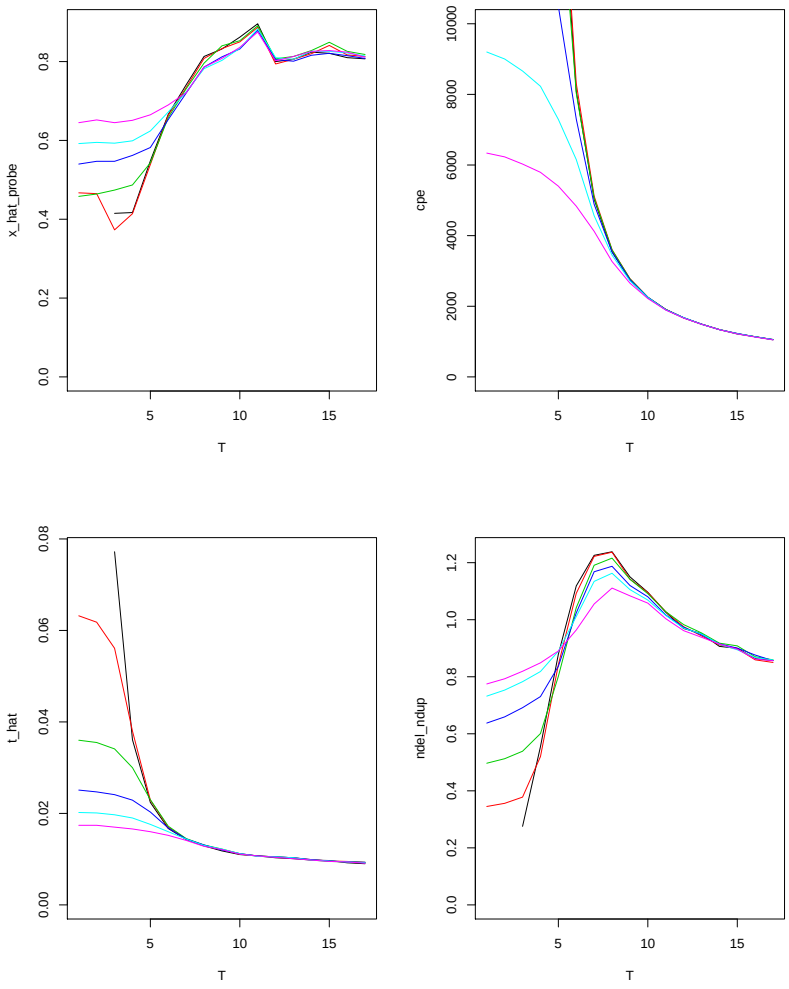


Figure 2.4: **Summary statistics of replicate pairs 2.** Each colored line represents a different value of a. Clockwise, from upper left: Jaccard overlap on probe basis, estimated number of genome-wide calls per experiment (co-hybridization), ratio of gains/loses, percent of probes inside a CNV call.

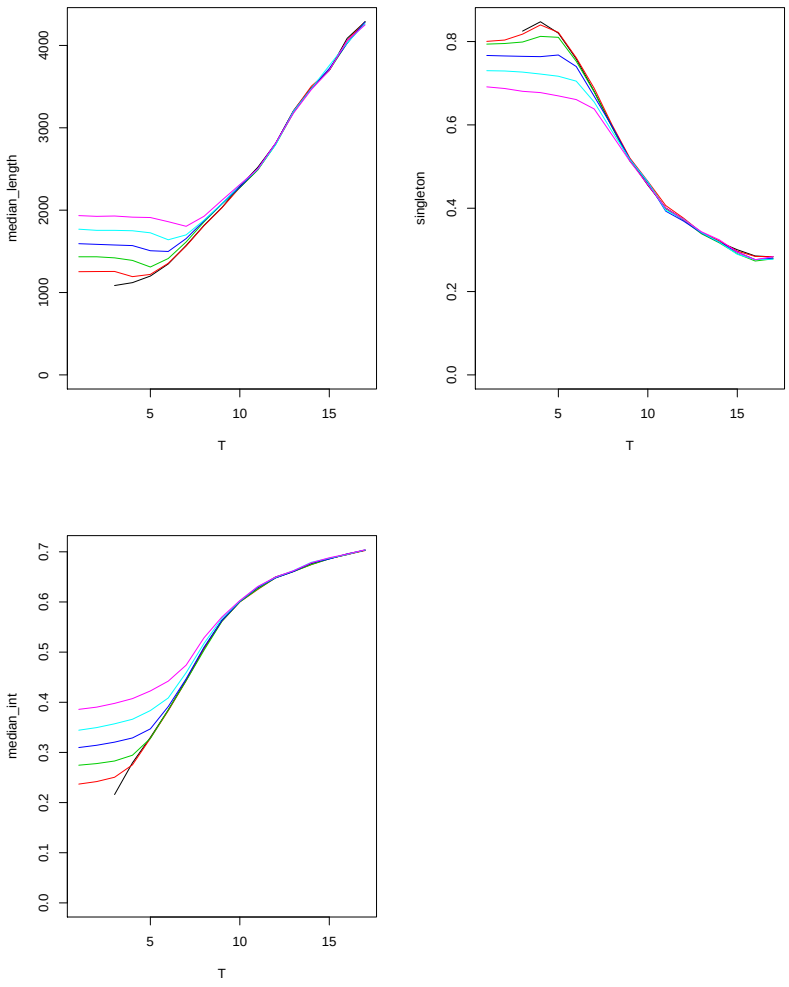
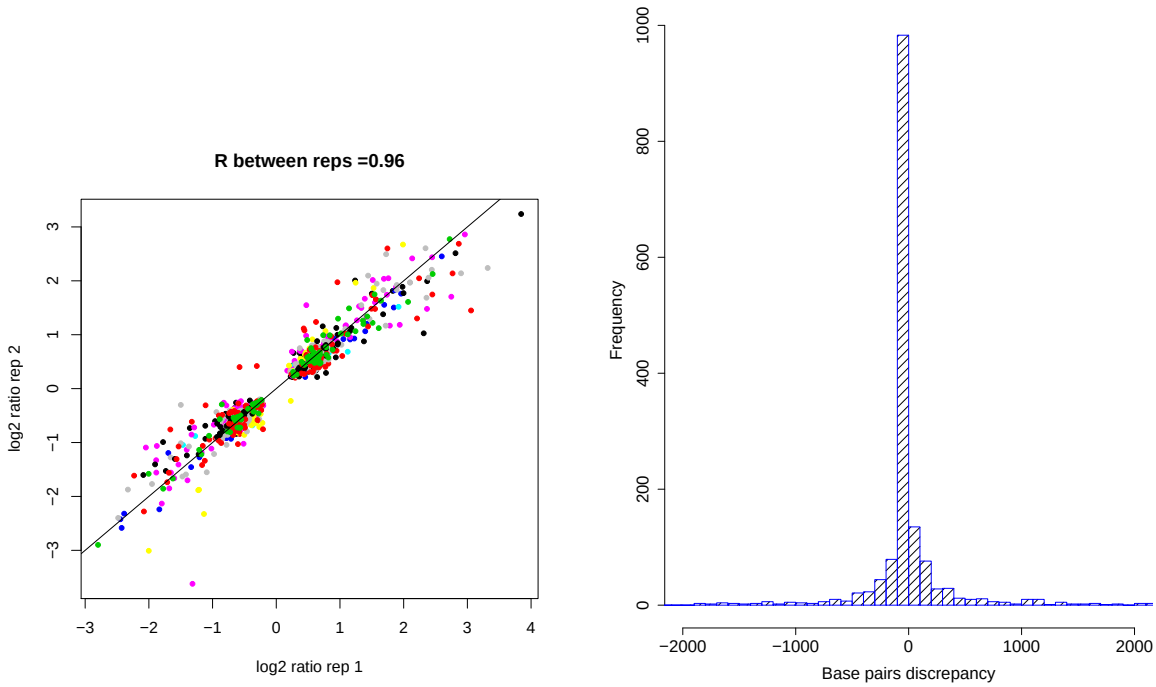


Figure 2.5: **Summary statistics of replicate pairs 3.** Each colored line represents a different value of a. Clockwise, from upper left: median length of CNV call, percent of CNVRs with a single CNV call, median intensity of CNV call.

Additional summaries of reproducibility. Other measures of reproducibility one could examine, to better understand the quality of the data, are the agreement in breakpoint position and log2 ratio of a CNV call made in both replicates. A conservative estimate of breakpoint reproducibility puts the variation in breakpoint position at around 7.9% ($T = 8$) or 3.9% ($T = 14$) of the total possible CNV length (in base pairs). Twenty-one percent ($T = 8$) to 24% ($T = 14$) of calls had perfect breakpoint agreement between replicates. Agreement in log2 ratio between replicated CNV calls appears extremely high with a Pearson correlation of 0.96 (Figure 2.6).



(a) (b)

Figure 2.6: **Agreement in log2 ratio and breakpoints for CNV calls made in replicate experiments.** (a) The correlation in log2 ratio for approximately 1000 calls made in both replicates, average across all samples, was 0.96. Data points are colored by sample of origin. (b) Distribution of difference in breakpoint location for CNVs called in both replicates. Both results from segmentation with GADA parameters of $T = 8, a = 1.7$.

Interaction between data quality and calling stringency One interesting observation is that data quality appears to have a non-linear affect on GADA CNV calling as the calling parameters are relaxed. We ran the whole-genome segmentation analysis on all 40 samples for values $T = 8, 10, 14$ and $a = 2.0$. One simple QC check is to calculate the number of calls per chromosome, separately for each sample (Figure 2.7). We see that the number of calls per chromosome increases quite rapidly with decreasing call stringency for a small number of outlier samples.

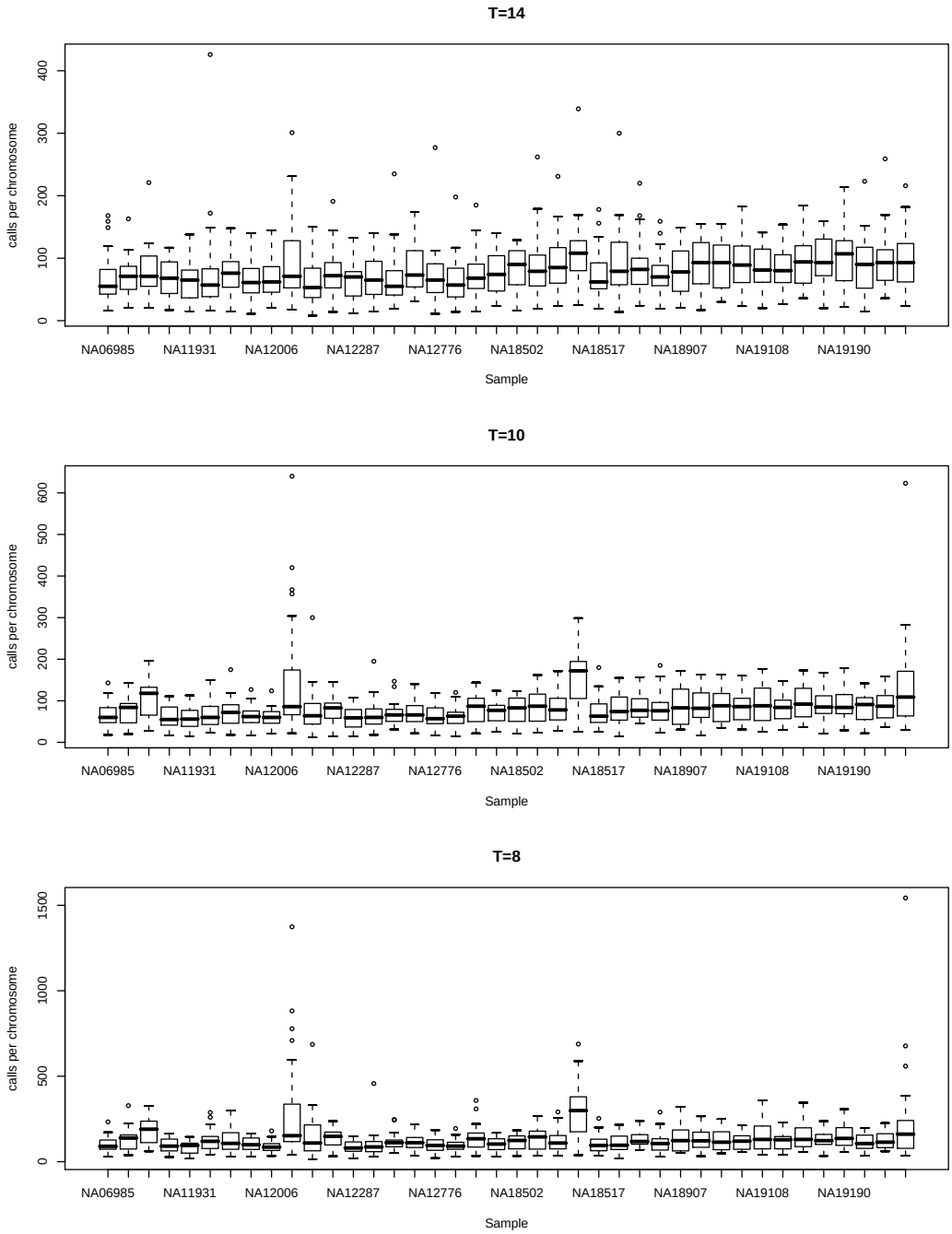


Figure 2.7: **Calls per chromosome, for three different set of segmentation results.** There appear to be some samples (it would be more accurate to say "subarrays") that "blow up" as the calling becomes more permissive.

1.11 Analysis of paired end sequence data from NA15510

For one sample, NA15510, CNV calls have been made using sequencing-based approaches (Fosmid end sequence (FES) and paired-end mapping (PEM)) that represent a fairly orthogonal source of information on our call quality and power (Tuzun et al., 2005; Korbel et al., 2007). We restricted our analysis to autosomal deletions in NA15510 relative to the reference (102 from FES, 282 in PEM), as insertions detected by the paired-end methodology may not be easily identified in the array data.

The analysis consisted of segmenting the entire autosomal genome for NA15510 over a range of a and T values, and counting the proportion of FES and PEM calls also called on the 42M platform (Figure 2.8). In order to declare that two variants detected on different platform correspond to the same event we require at least 40% reciprocal overlap between them. The striking impression that one gets from examining these results is that the maximum number of replicated calls, for both sets, is achieved at very conservative parameter settings (the smallest value of T examined is 24). For instance, over the range of $T = 24$ to $T = 5$, the number of replicated FES calls increased from 47 to 62. Overall a greater percentage of FES calls are replicated than PEM.

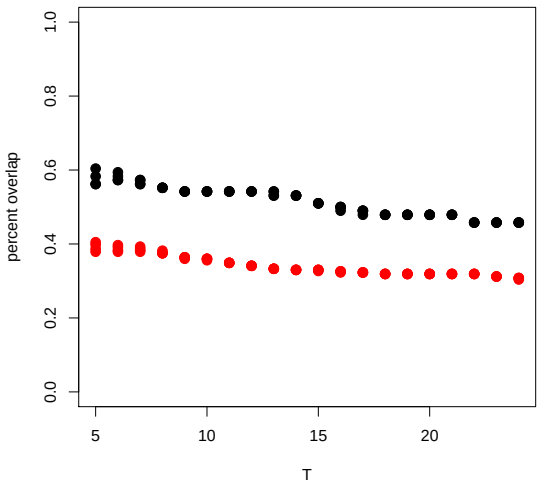


Figure 2.8: **Overlap between NA15510 42M calls and Tuzun FES (black dots) or Korbel PEM (red dots) calls, as function of T.** The y-axis indicates the proportion of 102 deletions identified by FES or 282 deletion identified by PEM that overlap calls made in 42M discovery data from sample NA15510.

The end-sequence based discovery strategies have a different power profile than array-based methods; they are especially advantageous at detecting rearrangements that involve extremely high copy number elements such as SINEs and LINEs. We manually curated the 47 FES calls that were not replicated for the value of $T = 8$, and identified two distinct sets

Location	Identified with Korbel PEM	Notes
chr1:216,247,751-216,282,164	Korbel	LINE, 1 gain called
chr2:19,051,097-19,091,018	No Korbel	large segdups
chr3:188,056,986-188,080,903	Korbel	SVA, 1 deletion called
chr5:103,880,569-103,907,829	Korbel	LINE
chr6:150,023,750-150,043,744	No Korbel	many SINES
chr7:4,584,757-4,607,795	No Korbel	HERVK, 1 deletion called
chr7:57,462,039-57,487,506	No Korbel	large segdups
chr7:91,033,112-91,073,941	No Korbel	LINEs
chr7:96,289,288-96,332,046	Korbel	LINE
chr8:126,663,820-126,670,399	Korbel	single LINE element
chr8:135,141,298-135,188,617	Korbel	single LINE element
chr10:6,433,069-6,462,822	Korbel	single LINE element
chr11:1,855,962-1,897,796	Korbel	repeat rich
chr12:13,429,399-13,466,004	No Korbel	LINE??
chr12:57,006,943-57,032,578	Korbel	HERVK, 1 loss called
chr14:23,475,038-23,545,053	No Korbel	2 losses called
chr14:23,539,508-23,574,552	No Korbel	3 losses called
chr15:52,995,999-53,034,494	Korbel	LINE
chr19:8,237,615-8,276,742	No Korbel	2 gains called, segdup

Table 2.3: **Summary of Tuzun deletion calls not validated with 42m discovery data**

of events. On one hand, there are 28 FES calls that correspond to regions of high frequency CNV in our data; these are likely to be places were the reference CNV genotype is affecting our ability to identify an event in NA15510. In support of this, 11/28 of these CNVRs contain both gains and losses, while such gain/loss loci represent fewer than 10% of CNVRs in the entire dataset. On the other hand, there are a set of 19 deletions called by FES (Table 2.3). Comparison with the PEM dataset suggests that a.) they are often replicated by PEM and cannot be solely explained as false positive by FES; b.) the breakpoints are often localized to the edges of LINEs and segmental duplications, and the deleted region is either not covered by probes on the 42M platform or the fold-change is too shallow to be called by single-sample segmentation.

In summary the results suggest that we come close to 100% power for calling deletions identified with FES, when we consider false-negatives due to reference effects and experimental limitations in highly repetitive sequence contexts.

2 CNV Genotyping

2.1 Array Design

Within the context of a CNV association study conducted by the Wellcome Trust Case Control Consortium (WTCCC), a CNV-typing array was designed by the WTCCC in a collaboration with the other co-authors of this paper in which a preliminary version of our discovery data was shared at an early stage. The array used the Agilent CGH-platform and comprised 105,000 long oligonucleotide probes. Full details of the array design will be provided elsewhere (manuscript in preparation), but in brief, it included 10,819 loci discovered from our CNV discovery experiment and 375 CNVs described in other discovery projects (Table 2.4).

The genotyping array design incorporated 9,722 loci (89.5%) from a preliminary set of 10,865 candidate CNV loci from the CNV discovery experiment. This set of loci includes all of the CNVs discovered in the CEU samples, and all of the CNVs discovered in 2 or more YRI individuals, but only a subset (53%) of the loci discovered in a single YRI individual. Upon generating a refined set of 11,700 candidate CNV loci from the same underlying data, we identified 10,819 (92.5%) of these loci as having probes on the genotyping array, and we used this later set of CNV definitions in downstream analyses.

Affymetrix 6.0 CEU CNVs are those CNVs observed in 2 or more unrelated CEU individuals in McCarroll et al. (2008) that were not also among the 42M CNV discovery set. Novel Sequence Insertions were selected as follows: 186 Novel Sequences identified and validated as being polymorphic by array-CGH by Kidd et al. (2008), and 106 Novel Sequences identified from alignment of the Venter genome sequence Levy et al. (2007) against NCBI36. Selection and analysis of WTCCC loci will be described in a forthcoming manuscript.

Name	Number of Loci	Source
42M CNV discovery	10,819 (9,722)	This study
Novel sequence insertions	292	Kidd, et al. 2008; Levy, et al. 2007
Affymetrix 6.0 CEU	83	McCarroll, et al. 2008
WTCCC loci	1,530	Manuscript in preparation

Table 2.4: Content on 105K CNV genotyping array

2.2 Samples

We screened the 270 samples genotyped as part of the PhaseI and II International HapMap Project (International HapMap Consortium, 2005), as well as the second plate of YRI samples and second plate of CEU samples included in the Phase III HapMap, amounting to a grand total of 450 samples. Sample NA15510 from the Polymorphism Discovery Resource was also included. The reference DNA for these experiments was actually an equimolar pool

of DNAs from 10 European samples (9 males, 1 female) obtained from the European Collection of Cell Cultures. Their IDs are C2078, C2141, C2153, C2173, C2175, C2188, C2159, C2184, C2142, C2151.

2.3 Pre-calling QC

The HapMap samples were typed on three days: 9/26, 10/27 and 12/08 2008, over a total of 12 batches (hyb plates). We used several statistics provided by Agilent Feature Extraction software for QC, including the Red Channel signal and "DLRSpread". The metric DLRSpread is computed as: DLRSpread = IQR(dLR) / 4*erfinv(0.5), where dLR is an array of differences between log ratios of adjacent probes, erfinv is the Inverse Error Function and IQR is Inter Quartile Range. Ninety-four samples were selected for repeat on the basis of DLRSpread and Red Channel signal. The median reduction of DLR spread for the repeated samples was to 60% of the original. Over 95% of the repeated samples had DLRSpread less than 0.3.

Prior to genotype calling, we searched for individuals carrying cell-line artifacts, in the form of partial or whole-chromosome aneuploidies. For each normalization (details below) we tabulated the 1st, 2nd, and 3rd quartile of the intensity distribution for each sample's data, chromosome-by-chromosome. Individual chromosomes that were outliers (+/- 4 median absolute deviations from the population median) in the population distribution for any of these quantiles were then replaced with missing data. The chromosomes censored are listed in Appendix B at the end of this document.

2.4 Normalization

All probe-level data from the genotyping array was initially normalized using the default settings of Agilent's feature extraction software. We developed 4 different post-processing routines for creating univariate summaries of the probe-level data from each CNV (Table 2.5). In brief, each post-processing ("norm") consists of a normalization step and probe summary step. Normalization 1 consists of quantile normalization of the log2 ratio from all probes across all samples. Normalization 2 involved adding an *ad hoc* batch correction at the probe level. This batch effect doesn't correspond to any experimental unit in our genotyping pipeline, but a subset of samples, predominately CEU, show systematic intensity bias at a set of high-GC content loci. We suspect this may represent a labeling problem due to an epigenetic change in some cell lines. This bias is apparent in principle components analysis (it is the first principle component (PC) when looking at intensities, and the third PC when looking at genotypes). We therefore used a linear model relating probe intensity to both PC1 of the intensity data and PC3 of the norm1 genotype data loading to regress out this effect. Normalization 3 was formulated to separate the signals of overlapping CNVs in complex regions of the genome. We developed an algorithm to identify the most unique regions of CNVs belonging to CNVRs with multiple events, and we used a principal components probe summary to further boost the major signal of variation at the locus (Barnes et al., 2008).

name	processing	summary
norm 1	log2(Cy5/Cy3), quantile normalized across samples	mean
norm 2	norm1 + PC correction	mean
norm 3	norm1, unique regions	PC1
norm 4	Cy5/Cy3	mean

Table 2.5: **Post-processing algorithms used for the 105K genotyping data.** See main text for details.

Definition of new loci. Some of the CNV regions on the array consist of multiple overlapping events. This complexity could derive from repeat mutations, or it could be noise in the segmentation process, which leads to incorrect fragmentation of a single CNV. To maximize the likelihood of obtaining quality genotypes from each CNV region, we defined a "meta locus" for each complex CNV that included all probes from the CNV region; these are referred to in the genotyping data as "_full" loci, e.g. "CNVR123_full". This boosts the number of loci, derived from the discovery project and that we attempted to genotype, above the 10819 that we originally defined.

2.5 Genotype calling and QC

After post-processing of the data each CNV is represented by a vector of N numbers, $\vec{x}$, that contains all the information used for assigning CNV genotypes to samples. The data at each locus is modeled as a mixture of normal densities, where each genotype cluster is parameterized by a mean, variance and frequency. We used the EM algorithm, a standard statistical technique for fitting such models that has enjoyed success in CNV analysis (Barnes et al., 2008; Korn et al., 2008). Models were fitted to all 4 post-processed datasets using priors on mean locations appropriate to the scale of that dataset. These prior means were determined manually by inspection of the data. As in (Korn et al., 2008), we place s "pseudopoints" at the prior mean of each cluster in the "M" step; we find that with the scale data we were using (450 samples) 1 point was often enough to significantly improve the frequency of correct solutions. Five models were fitted for each locus, corresponding to $\{2\}$, $\{1,2\}$,$\{0,1,2\}$, $\{2,3\}$ and $\{2,3,4\}$ copy number clusters; each model thus specified the number of genotype clusters and a prior on the cluster location. We also used a shrinkage step to stabilize the variances.

For any given model there are K clusters with frequencies $q_1, ..., q_j ..., q_k$. The algorithm iterates between estimation and maximization. The E-step estimates the probability that the ith sample belongs to cluster j,

E-Step:

$$P(z_i = j) = \frac{N(x_i; \mu_j, \sigma_j) * q_j}{\sum_{i=1}^{K} N(x_i; \mu_j, \sigma_j)}$$

where $N(x; \mu, \sigma)$ represents the normal pdf with mean μ and standard deviation σ evaluated at x. The M-Step finds maximum likelihood estimates for the cluster means and variances, and mixture proportions, conditional on the probability distribution of cluster membership across samples.

M-Step:

$$q_j = \frac{\sum_{i=1}^{N} P(z_i = j)}{N}$$

$$\mu_j = [(1/(s + q_j) * \sum_{i=1}^{N} P(z_i = j) * x_i] + [s/(s + q_j)\mu_j^p]$$

$$\sigma_j^2 = (1/q_j) * \sum_{i=1}^{N} *(x_i - \mu_j)^2$$

$$\sigma_j^2 = (\frac{1}{1+r}) * \sigma_j^2 + (\frac{r}{1+r}) * \hat{\sigma}^2$$

where $\hat{\sigma}^2$ is the mean of the cluster variances; we found $r = .1$ to be a generally useful setting.

Each model was fit using 5 independent starts of the EM and the highest likelihood run (of the 5) was retained. The best-fitting model was selected by Bayesian Information Criterion (BIC) and carried forward for manual curation. This manual curation was done using the program GASSS, which allows joint visualization of the 4 normalizations as well as manual editing of the number and location of genotype clusters. GASSS was developed internally within our group, by Jan Aerts, Don Conrad and Matt Hurles, and programmed by Jan Aerts in the Processing language (`www.processing.org`). The output of the manual curation process is a single set of intensities and models for all CNVs on the genotyping chip. Of the 13,007 total CNVs that we attempted genotyping, 8038 (60%) polymorphic CNVs were carried through from manual curation at the operators' discretion.

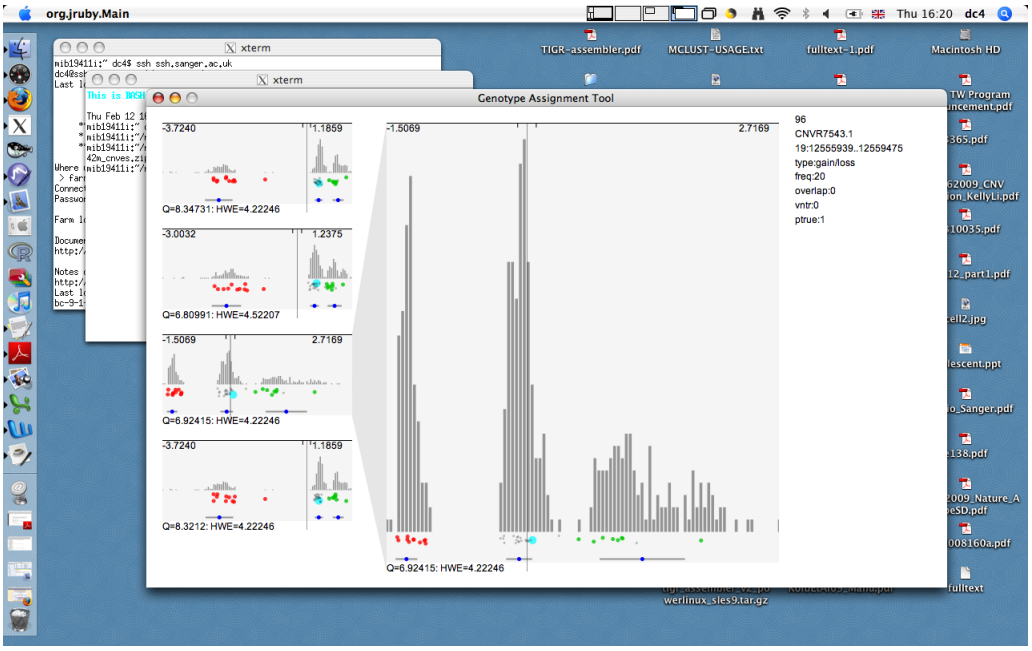


Figure 2.9: **Manual curation with Genotype Assignment Software (GASSS)).**

Manual curation using GASSS was run by two operators, so we analyzed the properties of the resulting datasets produced by each operator. The proportion of CNVs carried through was similar for each operator (65% and 59%), as was the distribution of normalizations (Table 2.6). The main goal of the curation process was to select the best normalization for each locus, but manual editing of the model fit did happen for 23.9% and 27.8% of loci for Operator 1 and 2, respectively.

Normalization	Operator 1	Operator 2
Norm 1	757 (16%)	475 (13.7%)
Norm 2	1994 (43.7%)	1451 (41.2%)
Norm 3	356 (7.8%)	621 (17.8%)
Norm 4	1455 (31.8%)	929 (26.8%)

Table 2.6: GASSS operator characteristics: Normalization usage.

After manual curation a final round of model-fitting was done which differed slightly from before. In this round two sets of model fits were generated: in the first, the cluster means are fixed to the values assigned during manual curation; in the second the cluster means from GASSS are used as priors but the means are allowed to be updated. The number of clusters at each locus are fixed to the value specified by GASSS curation. CNVs were genotyped using these final fitted models, with the requirement that each individual have a posterior probability $> 90\%$ of cluster membership in order to be called. A small number of model

fits didn't lead to multiple copy number states being called (due to poor clustering quality) and removal of these loci brought the total number down to 7,827 genotyped loci which were then passed on to QC.

These genotypes were subjected to a round of QC using criteria analogous to those applied to the Phase I HapMap (International HapMap Consortium, 2005). CNVs were flagged QC- if $> 10\%$ of the genotypes at the locus were missing data (QC-m), if the chi-square statistic for goodness of fit to the expected genotype counts under Hardy-Weinberg Equilibrium > 10 in at least one population (QC-h), if two or more Mendelian errors were observed at the locus (QC-). A total of 6487 CNVs passed QC as 1340 (17%) of CNVs failed QC.

XY homologies. Segmental duplications between the sex chromosomes and autosomes can potentially generate false-positive CNV calls due to duplication shadowing. We conducted a genomewide association study on all autosomes using the Agilent data in order to identify such CNVs, fitting a logistic regression of sex on CNV intensity. The distribution of the standardized regression coefficient from this procedure suggested a threshold for 10 standard deviations for declaring an XY homology, and 296 CNVs genomewide fell above this threshold (Figure 2.10).

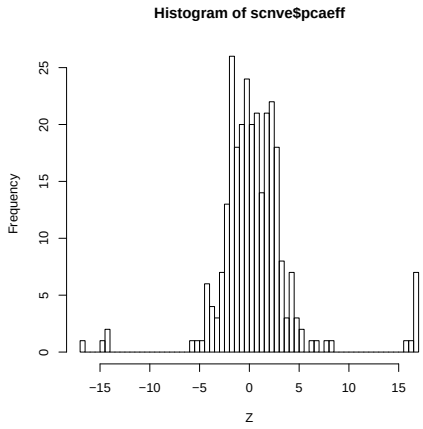


Figure 2.10: **Distribution of standardized regression coefficient of sex on CNV intensity.** Data generated using all CNVs from chromosome 22.

Redundant loci. Finally, we removed CNVs that appeared to represent a duplicate of another genotyped CNV. Such a situation could arise, for instance, where the CNV map showed evidence of overlapping CNVs with distinct breakpoints but a) there was in fact only one mutation segregating or b) genotyping probes could not be designed that could distinguish the two events. After filtering on possibly redundant CNVs we were left with 5238 QC+ loci.

QC Flag	Number of Loci
QC+	6487
QC-e	66
QC-h	32
QC-hi	23
QC-i	15
QC-m	1204

Table 2.7: Tabulation of QC flags across all genotyped loci. See text for description of flags.

We assessed the quality of the genotypes by measuring the concordance of genotype calls in repeat experiments; concordance between replicates ranged from 99.76% to 99.92% (Table 2.8. The Mendelian error rate was measured in 114 trios at non-X, bi-allelic loci with $< 10\%$ missing data, and we found that it was in a range comparable to some SNP genotyping platforms that were used in the Phase I HapMap (median .25%, range 0-2.1%).

Sample	Complete Data	QC+
NA15510	99.77	99.85
NA18537	99.23	99.81
NA18923	99.25	99.92
NA19096	99.12	99.85
NA12878 (4 repeats)	NA	99.76 (mean)

Table 2.8: Average concordance for replicate experiments at all genotyped loci (left) and at QC+ loci (right).

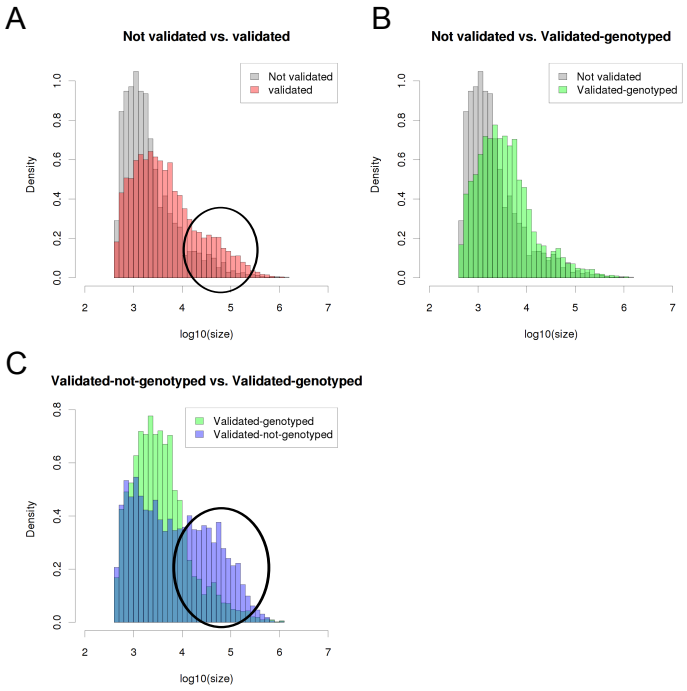


Figure 2.11: **Size distribution for the various CNV datasets generated in different stages of the study.** The distribution of CNV length is shown for the various breakdowns of (A) validated vs. non-validated, (B) validated-genotyped vs. non-validated, and (C) validated-genotyped vs. validate-non-genotyped. There is a subset of large CNVs that have not been genotyped though they are part of the validated set. Closer inspection of these CNVs indicates that ungenotyped CNVs larger than 10Kb are predominantly "gain" calls from the CGH experiment (60% "gain" vs. 23% of "loss" and 17% of "gain+loss") and associated with segmental duplications as well as with repeat elements such as simple repeats (ie. using Tandem Repeat Finder).

2.6 Absolute Copy Number / Ancestral State Assignment

The problem of ancestral state assignment is related to the determination of absolute copy number, and thus the analyses were done in parallel. We ran 1 chimp (Clint) on the 42M platform as well as 8 chimps on the Agilent 105K genotyping chip. Our original plan was to use the chimp genotype calls to define the human ancestral state at each CNV. As the analysis progressed, the patterns of polymorphism in the chimp data and inspection of the

chimp genome sequence made it clear that there would frequently be independent variants segregating in both chimp and human genealogies and multiple mutational events might separate human and chimp. Additionally, in the case of two-component CNVs, there is the possibility for high-frequency derived alleles that the ancestral state is not represented by a human genotype. Consequently we regard the absolute copy model to be a better guide than chimp ancestral state.

Chimp genotyping All of the human QC+ genotyping models were re-mapped to a single normalized dataset, log2(red/green). Divergence-corrected chimp intensity values were mean-summarized for each CNV, and then genotyped using the human genotyping models. For three-component CNVs, ancestral state was called only if 5 non-missing chimp calls were identical. Seventeen percent (about 900/5238) of loci show high likelihood of polymorphism in chimpanzee.

Modeling absolute copy number The Agilent 105K array is a comparative genome hybridization (CGH) platform which requires a reference sample for each experiment. Instead of using reference DNA from a single individual we used a pooled reference consisting of an equimolar mix of 9 males and 1 female from the European Collection of Cell Cultures (ECACC). The consequence of this pooled reference is that the cluster means and variances for each genotype is a function of the underlying genotype frequency (Figure 2.12). If we have a bi-allelic locus with integer copy number C_0, C_1, C_2, that are present with frequencies f_0, f_1, f_2; then the cluster mean for the integer copy number 'n' should be

$$M_n = n/(f_0C_0 + f_1C_1 + f_2C_2) \tag{1}$$

And the variance might be

$$V_n \propto n/(f_0V_0 + f_1V_1 + f_2V_2)$$

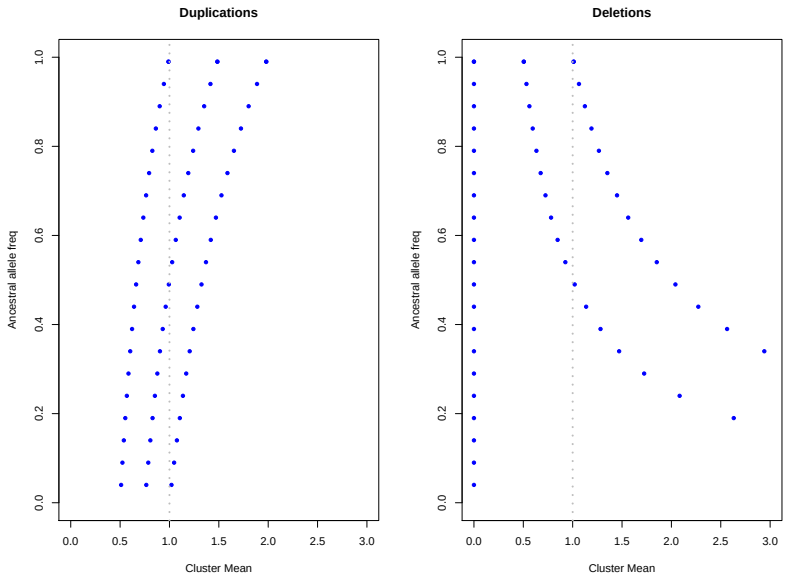


Figure 2.12: **Expected cluster locations as function of CNV allele frequency for bi-allelic CNVs.** The expected intensity ratios (Cy5/Cy3), calculated by Eq. 1 for all three genotypes (x-axis) are plotted as a function of the ancestral allele frequency (y-axis), for duplications (left) and deletions (right). Notice that the range of deletions and duplications is overlapping.

Our strategy for assigning absolute copy number is to use as much information as possible: we use the allele frequency (a, the cluster means for both the ratioed (Cy5/Cy3) as well as single channel (Cy5) intensities, and the relative spacing of the cluster means. The red channel intensity alone potentially contains useful information about absolute copy number that is lost when compared to the reference. We treat the problem of assigning absolute copy number as a model selection problem. Given the number of components at the locus, we identify the maximum likelihood model (m) out of a set of predefined models:

- for a 2 component locus, there are 4 models: 0,1, 1,2, 2,3, 3,4
- for a 3 component locus, 3 models: 0,1,2, 1,2,3, 2,3,4
- and for > 3 components, 3 models: 0,1,2,$>$ 2, 1,2,3,$>$ 3, 2,3,4,$>$ 4

We consider the cluster means (both ratioed and single-channel) and the ratio of adjacent cluster means to be normally distributed conditional on absolute copy number. The mean of these cluster distributions are a function of allele frequency (as explained above). The variance is modeled to be constant for duplications, and a linear function of allele frequency for deletions. These variances are estimated in an iterative fashion, beginning with a clustering of outliers, and then re-estimated after the first set of absolute copy number assignments.

Data. For notation, let $\vec{f}$ be the k cluster frequencies at the locus, $\vec{x}$ be the means of the k cluster log2 ratios, $\vec{r}$ be the $k-1$ ratios of adjacent cluster means $(x_1/x_2, ..., x_{k-1}/x_k)$. These are all observed quantities.

Model Notation. A full table of symbols used in notation is given in Table 2.9. The absolute copy number likelihood is formulated as a function of (maybe unknown) parameters. Each model m is defined by a number, k_m, of components and the absolute copy number of those components: $c_1, ..., c_i, ...c_{k_m}$. Therefore $\vec{\mu^l}$ and $\vec{\sigma^l}$ are the set of k_m log2 ratio means and standard deviations for model m, $\vec{\mu^{cy5}}$ and $\vec{\sigma^{cy5}}$ the set of red channel means and standard deviations; $\vec{\mu^r}$ and $\vec{\sigma^r}$ the means and standard deviations of adjacent clusters.

We model the cluster log2 ratio means simply as a function of allele frequency and absolute copy number, as in Equation 1. Using the current notation then given a model m,

$$\mu_i^l = \frac{c_i}{\sum_i^{k_m} f_i c_i} \tag{2}$$

Although we are using the intensity data from all populations to fit this model, we only use CEU data to estimate the f_i's as our reference pool is thought to be comprised of only European individuals.

Given these expected μ^l's we can then calculate expected values of μ^r as $\mu_i^r = \frac{\mu_{c_i}^l}{\mu_{c_{i+1}}^l}$.

The values for the μ_i^{cy5}'s, are not obtainable from first principles. We manually clustered a subset of loci that were obvious deletions and duplications and used these prior cluster assignments to calculate μ_i^{cy5} and σ_i^{cy5} for copy number $i = 0, 1, ..., 6$. Values for σ_i^r for $i = 0, ..., 5$ and σ_i^l for $i = 2, ..., 6$ were calculated in the same way.

A preliminary analysis of deletion loci suggested that the variance in the distribution of cluster means increases as a function of deletion allele frequency. We therefore use a linear model for the σ_i^l in deletion models. The model parameters were estimated using an iterative process of absolute copy number assignment and model fitting. The final variance models used were

$$\sigma_0^l = 0.11$$

$$\sigma_1^l = 0.06 + 1.38x^2$$

$$\sigma_2^l = 0.03 + 2.36x^2$$

where x is deletion allele frequency.

Using the parameters and data defined above, the likelihood of model m, given the data, is

$$L(m|\vec{f}, \vec{x}, \vec{r}) = \prod_i^{k_m} N(x_i; \mu_{c_i}^l, \sigma_{c_i}^l) \prod_i^{k_m} N(y_i; \mu_{c_i}^{cy5}, \sigma_{c_i}^{cy5}) \prod_i^{k_m - 1} N(r_i; \mu_{c_i}^r, \sigma_{c_i}^r) \tag{3}$$

Data	Description
$\vec{f}$	cluster freqs
$\vec{x}$	cluster means (Cy5/Cy3)
$\vec{y}$	cluster means (Cy5)
$\vec{r}$	ratio of adjacent cluster means (Cy5/Cy3)
Model Parameters	**Description**
m	model type
k	number of clusters in model
c_i	absolute copy number of ith cluster
$\mu^l_{c_i}, \sigma^l_{c_i}$	mean, standard deviation of cluster means for copy number c_i, (Cy5/Cy3)
$\mu^{Cy5}_{c_i}, \sigma^{Cy5}_{c_i}$	mean, standard deviation of single-channel cluster means for copy number c_i, (Cy5)
$\mu^r_{c_i}, \sigma^r_{c_i}$	mean, standard deviation of $\mu^l_{c_i}/\mu^l_{c_i+1}$

Table 2.9: Description of symbols used in modeling absolute copy number.

3 Validation

3.1 Comparison of discovery and genotyping data

All 41 samples from the 42M discovery experiment were also typed on the 105K genotyping chip. Intensity data from probes measuring the same CNV should be correlated between the two experiments; however, we don't expect intensities to be correlated at the location of a false positive CNV. Using this principle we devised a method to estimate the false discovery rate (FDR) of the CNV calls made in the discovery phase.

Pearson correlations were calculated on the summarized data (normalization 2) for all 10819 autosomal CNVs with probes on both platforms. These were then transformed into standard normal variates using Fisher's z transformation (Figure 2.13). The resulting data can be modeled as a mixture of two Gaussians, one representing the distribution of correlations for false positive sites, and one the distribution of correlations for true positives. The expectation-maximization (EM) algorithm was used to estimate means and variances for each distribution, and the mixture proportion. The false discovery rate as measured by the mixture proportion was 15.3%, however, this does not consider that a proportion of CNVs with probes on the Agilent array will falsely report negative due to erroneous probe placement (see below).

Using the fitted model, we calculated for each CNV the probability that it belongs to the false positive cluster, pr($Z = 1$), and the probability that it belongs to the true positive cluster, pr($Z = 2$). We refer to this latter probability as "pp_true". If we removed all CNVs with pr($Z = 1$) >pr($Z = 2$), we would remove 1945 events, and there would be an estimated

314/8874 (3.5%) false discovery rate in the remaining CNVs.

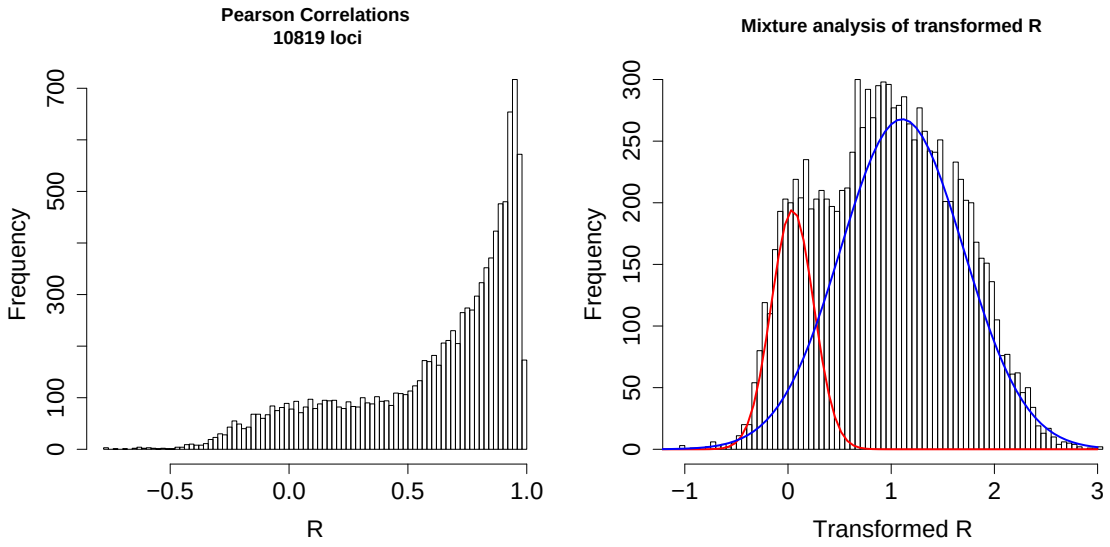


Figure 2.13: **Validation of CNVs discovered from 42M CGH data using 105K CGH data.** On the left, distribution of Pearson correlation coefficient for Agilent genotyping signal and 42M discovery data signal across 10819 loci. The data can be modeled as a mixture of two distributions: correlations from false positive loci, and correlations from true positive loci.The right panel displays a mixture model fit to Fisher z-transformed data from left panel.

This FDR estimate is actually a concordance metric that conflates false positives from the 42M discovery with false negatives on the Agilent 105K array. One potential source of low concordance is the probe placement on the 105K array. More conservative probe design filters on the 105K array mean that if probes are not well spread throughout the discovered CNV region they may lie outside the true CNV boundaries and not show any CNV signal accordingly. We formulated two statistics, "edge", and "spread" to measure the extent of this phenomenon (Figure 2.14). "Edge" is the proportion of probes in a CNV that are within the first and last 10% of the CNV. "Spread" is the distance from the start of the first probe to the end of the last probe within a CNV, scaled to the length of the CNV (ie. the maximum spread is 1).

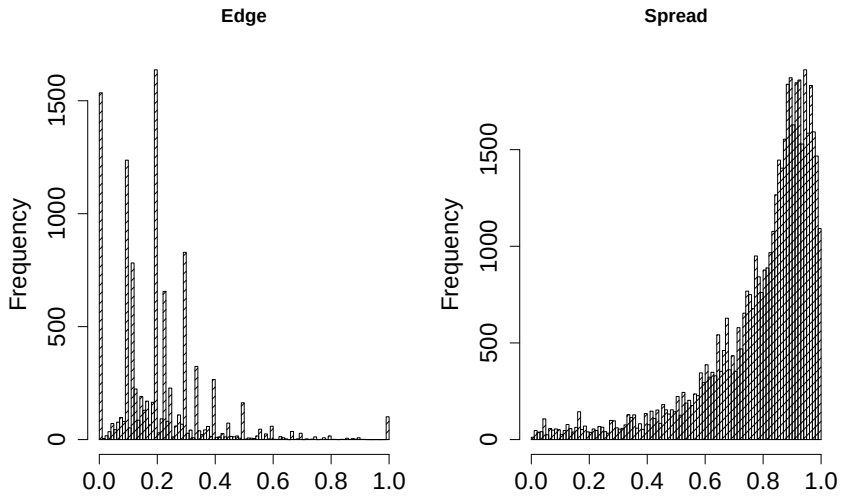


Figure 2.14: **Two statistics summarizing probe placement within CNVs.** Left, the "edge", or proportion of probes in the first and last 10% of the CNV. Right, "spread" the distance between the first and last probe within the CNV, scaled by the length of the CNV.

The overall impact of the probe placement on validation rate can be gauged by looking at validation rate across deciles of edge and spread (Figure 2.15). There is a clear deficit of validated calls in the smallest decile of spread and the top decile of edge (all probes clustered in outer 10%). One can make heuristic arguments based on these numbers that approximately 2% of loci are not validated due to issues with probe placement, although we cannot say definitively which CNVs are the false negatives.

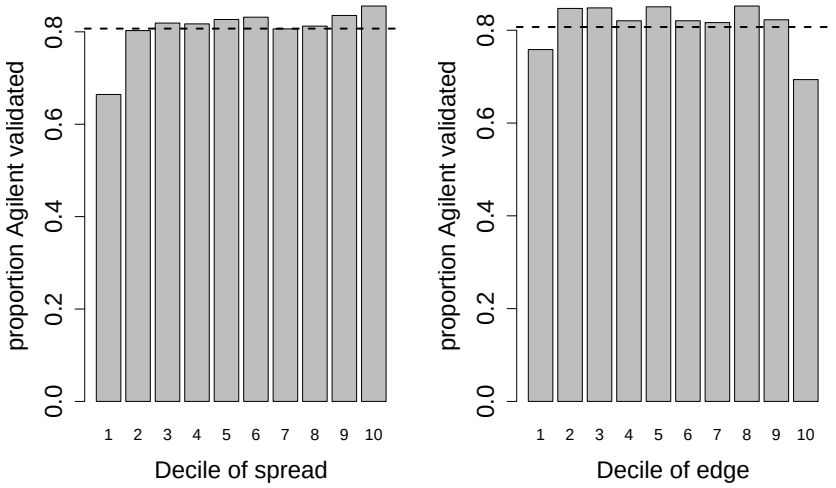


Figure 2.15: **Impact of probe placement on validation.** The proportion of CNVs validated by Agilent is plotted by deciles of the spread and edge statistics. There are clear deficits of validation in the bottom decile of spread and top decile of edge.

3.2 Prior datasets

We compiled a number of datasets from high-resolution experiments that would be useful for validating calls made in the current project.

- Validated sites of structural variation from Kidd et al. (2008) were downloaded from the Eichler lab website on 21 January 2009. Data were converted from NCBI35 to NCBI36 using liftOver, all but 18 converted successfully.
- Indels and structural variants from 4 published individual genome sequences: Venter Levy et al. (2007), Watson Wheeler et al. (2008), NA18507 Bentley et al. (2008), and the Asian sequenced by Wang et al. (2008).
- McCarroll et al. (2006); Conrad et al. (2006); Mills et al. (2006)
- McCarroll et al. (2008) 1318 sites of genotypable CNVs from HapMap.
- 342 CNVs with sequenced breakpoints from NA18505 and NA15510 (Kim et al., 2008).
- deSmith et al. (2007),3554 CNV calls from Perry et al. (2008).

There is uncertainty in the location of the breakpoint for every dataset. The rules for deciding if two variants correspond to one another may have a strong influence on the validation rate. We have looked at both 51% and 80% reciprocal overlap.

3.3 Validation summary

The final validated dataset was created by retaining CNVs that met at least one of the following three criteria: the pp_true of the CNV > 0.9, the CNV had > 0.6 reciprocal overlap with a call from one of the external datasets, or QC+ genotypes were generated for the CNV. There was an additional filter that all loci must have an average GC content less then .65 in order to be considered validated; this final filter removes 365 CNVs that would otherwise be validated (many of which are VNTRs). We also removed CNVRs overlapping Immunoglobulin loci likely to be somatic events. In total 8599 CNVs were validated, representing 43911 calls.

3.4 Q-PCR

Introduction. In collaboration with Applied Biosytems (AB), TaqMan assays for 103 randomly chosen CNV regions from our pre-validated set were selected from pre-designed TaqMan Copy Number Assays. All 41 samples from the discovery phase (40 targets plus reference sample), were analyzed using these assays. TaqMan assays for each CNVR target were selected from the Applied Biosystems pre-designed TaqMan Copy Number Assays (P/N 4400293). One assay was selected for all the CNVRs except that 2 assays were selected for 7 of them. Among the 110 selected TaqMan copy number assays, about half of them targeted gene regions and the other half targeted non-gene regions.

Methods. All the assays, ordered from appliedbiosystems.com, were first validated with a panel of 92 genomic DNAs, which were purchased from Coriell and composed of Caucasians and African Americans. Validated TaqMan assays were then run on all 41 discovery samples, plus one of two additional Coriell samples (NA17210 or NA17144). The FAM dye-based TaqMan copy number assays, designed to detect the target of interest, and the VIC dye-based RNaseP TaqMan Copy Number Reference Assay (P/N 4403328, from Applied Biosystems) were run in a duplex real-time PCR reaction. The final assay condition was 10 ng of genomic DNA, 1X TaqMan probe/primer mix in 1X TaqMan Genotyping Master Mix (P/N 4371357, from Applied Biosystems) in a 10 μl reaction with quadruplicates on 384-well plates. PCR reactions were incubated in an Applied Biosystems 7900HT SDS instrument for 2 minutes at 50C, 10 minutes at 95C, and followed by 40 cycles of 15 seconds at 95C and 60 seconds at 60C. Real-time data was collected and processed by the SDS 2.3 software. The SDS output files were analyzed by CopyCaller, the free data analysis software from Applied Biosystems for the TaqMan Copy Number Assay. The relative quantification analysis with a reference sample was performed to calculate estimated copy numbers of each sample for the target of interest. The HapMap samples NA10851 (default) or NA15510 were used as a reference, whereas the Coriell samples NA17210 or NA17144 was used as a reference sample only in the situation where both NA10851 and NA15510 give 0 or non 2 copy numbers.

Comparison with absolute copy number predictions. The copy number estimates of 60 validated non-complex CNVs (i.e., the breakpoints of which do not overlap with other CNVs among individuals) were highly concordant (99.99%) between AB predictions and array-based predictions (Figure 2.16), except for one locus (CNVR2217). Even though the relative estimations of array based and TaqMan based predictions for the copy number state of this locus are highly correlated, the former predicts the absolute copy number to vary between 2 and 4 among samples, whereas the latter predicts a variation between 0 and 2 (data not shown).

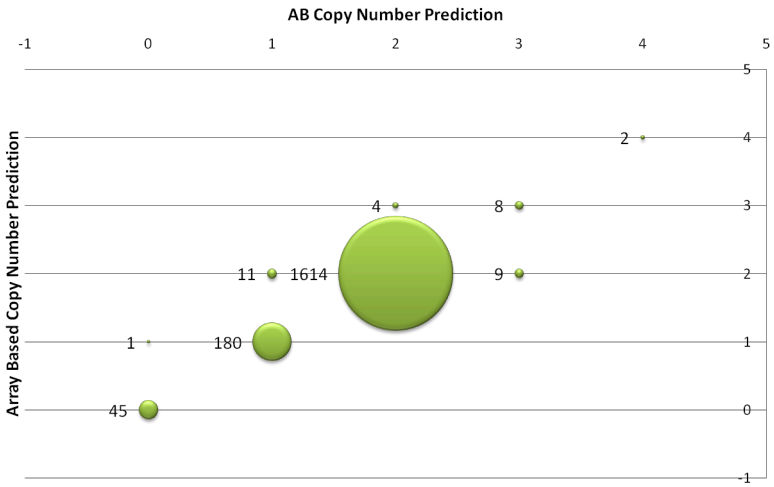


Figure 2.16: **Validation of absolute copy number estimates.** This scatterplot shows the concordance between estimates of absolute copy number derived from the 105K genotyping data (y-axis) and TaqMan (x-axis). The size of each point is proportional to the number of observations, which is printed immediately to the left in each case. The data represented by the figure are individual copy number estimates for 60 CNVs made on 41 samples (with a smaller number of observations due to missing data).

3.5 Mass Spectrometry

Introduction. The MassARRAY copy number variation (CNV) method combines real-competitive PCR (rcPCR) with MassEXTEND procedures and matrix-assisted laser des-

orption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) (Ding and Cantor, 2003; Elvidge et al., 2005; Oeth et al., 2003). Following the design of assays, genomic template is spiked with a synthetic DNA molecule (competitor) or with chimpanzee DNA, which matches the sequence of the targeted human assay region in all positions except a single base and serves as an internal standard. The two distinct sequences are targeted to the two different alleles. As a result, tools for primer extension may be applied that are the same as those used for SNP allele frequency analysis, and the ratio of peak areas associated with each allele can be used to quantitatively determine the number of copies of the wild-type template vs. the known number of copies spiked into each reaction on a per-locus basis.

Alignment processing and extraction of loci. The human genomic sequence for each of the CNV regions were extracted from the UCSC Genome Browser and aligned with the chimpanzee genomic sequence to identify regions of similarity. All sequence differences between human and chimpanzee were categorized based on the position of difference in the alignment, alleles in human and chimpanzee, direction of alignment, etc. Using these annotations, the human sequence in each region was masked at the locations of differences. Loci were formatted as SNPs for Sequenom assay design, with the first allele being human and the second chimpanzee, irrespective of alignment direction; alignment differences in the sequence flanks had already been masked in step above. As a quality control, all masked region sequences were inspected and compared to unmasked sequences. The same procedure as above was used for the 5 control regions received.

Loci prescreening and processing assay design. All aligned human-chimpanzee loci were processed through the web-based Sequenom Assay Design tools; ProxSNP and PreXtend. ProxSNP is used to identify the location of variant bases within a given proximity of the specified human-chimpanzee loci. PreXtend is used to select PCR primers and verify the uniqueness of the primer pair. Processed loci for each genomic sub-section were sorted by chromosome locations, which allowed selection of assays evenly across the region. Assays that had SNPs defined as chimpanzee deletions (i.e. [N/-]) were excluded from further consideration. Eight independent assays were selected across each individual sub-section. All control non-variant region assays were selected with the same processes used for genomic CNV regions.

Assay design. The initial study plan called for each multiplex to consist of 4 assays from each of three regions of interest to be grouped with 4 control assays (4 assays X 3 ROI + 4 control assays). The multiplexed grouping of CNV region assays as well as the control assays formed the SNP group file. The rcPCR assay design format was used for all assays. This design parameter utilizes a wildcard designation ([human/chimpanzee/*]) for the competitor allele. During the assay design process, the wildcard was translated into the appropriate base for mass detection.

Genomic templates and primers. Genomic templates were serially diluted from 40ng/ul to 5ng/ul. Stock concentrations of all samples were measured by NanoDrop according to the manufacturers protocol. All primers and competitor oligos were ordered from IDT Technologies. Competitive templates were HPLC purified and independently quantitated with a NanoDrop to verify concentration before usage. Each DNA sample was titrated from 40ng/ul to 5ng/ul against a fixed amount of either synthetic competitor oligo (4500 haploid genome equivalents) or against 15ng of chimpanzee DNA, respectively.

Sequenom iPLEX reaction. The genomic template/competitor mixture was rcPCR amplified and subjected to a post-PCR shrimp alkaline phosphatase (SAP) enzyme treatment to dephosphorylate any remaining unincorporated nucleotides. After inactivation of the alkaline phosphatase, a primer extension cocktail was added. The rcPCR products from the competitor and genomic template then served as templates for MassEXTEND. The primer extension products were purified through the addition of clean resin and then dispensed on a SpectroCHIP, a chip array that was preloaded with the components needed for MALDI-TOF MS sample preparation.

Data analysis. We defined the wild-type genome frequency as (signal intensity of wild-type allele) / (signal intensity of wild-type allele + signal intensity of the competitor allele), and the QGE software exported frequencies were used for this study (SEQ, 2005; Mistro, 2005). According to the competitive binding assumption, the frequency and competitor concentration should follow the relationship:

$$f = \frac{1}{1 + 10^{(log_{10}EC50 - log_{10}C)}}$$

where f is the wild type genome frequency, C is the competitor concentration, and $EC50$ is the genomic template concentration where the concentrations of the wild type genomic template and competitor template are equal (SEQ, 2005; Mistro, 2005; Oeth et al., 2003). Given this relationship, for each assay and sample, we could easily estimate the genomic template concentration by doing a non-linear regression (Mistro, 2005). To control for sample to sample (i.e. well to well) loading variation, all copy number measurements were normalized against copy number measurements from the assays targeting non-variant regions of the genome from the same well. For the ease of the comparison, we also expressed or plotted these normalized copy numbers as relative ratio between the observed samples vs. a reference sample (the HapMap sample NA11931 was used for most of the plexes; samples NA11995, NA12004, NA12044, NA18505, NA12006 were used in order, in case NA11931 was not available due to amplification failure). All the statistical analyses were carried out using R (R Development Core Team, 2008).

No.	Nimblegen 42M					Agilent		SQNM			Concordance
#	CNVE ID	Reasons for selection	cytoband	Target-Region (chr. start-end)	Size (bp)	CN_type	CN	Ref Sample	# assays/ locus	CN_type	(#errors/ samples tested)
1	CNVR305_full	large CNV; various genes: SYT6\HIPK1\MAGI3\PHTF1\PTPN22\BCL2L15\AP4B1\OLFML3 \DCLRE1B\RSBN1	1p13.2	113684475-114002640	318,165	Dels	1,2	NA11931	3	Dels	1/41
1	CNVR305_full	large CNV; various genes: SYT6\HIPK1\MAGI3\PHTF1\PTPN22\BCL2L15\AP4B1\OLFML3 \DCLRE1B\RSBN1	1p13.2	114654475-114692640	38,165	Dels	1,2	NA11931	4	Dels	1/41
2	CNVR2217.1	high VST=0.704; *PDLIM3*-intronic	4q35.1	186678922-186681050	2,128	Dels	0,1,2	NA12004	3	Dels	2/41
3	CNVR2445_full	Known *GHR* exon3-deletion polymorphism; OMIM	5p12	42664005-42667027	3,022	Dels	0,1,2	NA11931	4	Dels	0/41
4	CNVR2906.1	high VST=0.419; *C6orf142*-intronic	6p12.1	54037093-54042000	4,907	Dels	0,1,2	NA11993	4	Dels	1/41
5	CNVR3107_full	high VST=0.261; *ESR1*-intronic; OMIM	6q25.1	152431681-152433972	2,292	Dels	0,1,2	NA11931	3	Dels	0/41
6	CNVR3928.1	common CNV in CEU; *EYA1*-intronic; OMIM:gene associated to branchio-oto-renal syndrome); dosage-sensitive gene	8q13.3	72377264-72380227	2,964	Dels	0,1,2	NA11931	3	Dels	1/41
7	CNVR4739.1	MBL2-exonic; OMIM:deficiencies associated with susceptibility to autoimmune and infectious diseases	10q21.1	54196670-54199100	2,430	Dels	1,2	NA11931	1	Dels	1/41
8	CNVR6315.1	high VST=0.417; ATP10A-intronic; Imprinted	15q12	23482489-23483487	998	Dels	0,1,2	NA11931	2	Dels	3/41
9	CNVR6782.1	high VST=0.426; CNTNAP4-intronic	16q23.1	75097034-75101030	3,996	Dels	0,1,2	NA11931	4	Dels	0/41
10	CNVR8300_full	DMD-intronic; OMIM	Xp21.1	32897162-32898244	1,082	Dels	0,1,2	NA11931	2	Dels	0/41
11	CNVR8499_full	high VST=0.383; TMLHE-intronic	Xq28	154443364-154450773	7,409	Dels	0,1,2	NA11931	2	Dels	0/41
12	CNVR8422.1	X-linked; *AKAP14*-exonic	Xq24	118920043-118938368	18,325	Dels	0,1,2	NA11931	6	Dels	0/41
13	CNVR1668_full	TP63-intronic; dosage-sensitive gene	3q28	190845902-190853732	7,831	Dels	0,1,2	NA11931	2	Dels	0/41
14	CNVR6359.1	PLCB2-exonic; OMIM:Taste transduction - PLCβ2 knockouts have a complete loss of sweet, amino acid, and bitter taste responses	15q15.1	38376756-38377676	920	Dups	2,3	NA11931	2	Dups	0/41*
15	CNVR8324.1	X-linked; various genes: SSX6\SPACA5B\ZNF630\ZNF182\SPACA5	Xq11.23	47802588-47815345	12,757	Dels	0,1,2	NA11931	4	Dels	6/41 **

*not variable in agilent
** a bit noisey in agilent, there are two clear, tight clusters for males and females, but a number of samples falling in between these clusters

Figure 2.17: **Validation of biologically interesting loci by mass spectrometry.** This table summarizes the results obtained for 15 CNV loci experimentally confirmed on three platforms, Nimblegen 42M CGH, Agilent 105K CGH and Sequenom Mass Spec (SQNM). We observed that for loci where both SQNM and 105K CGH platforms showed copy number variation the concordance was high (0-2 discrepancies). We also observed the absolute copy number for a locus could vary between different methodologies/platforms when the same reference sample was not used. For example, after examining each locus carefully, we could determine that a CNV with CN levels [0, 1, 2] in the Agilent platform could easily represent a CNV with CN levels [0, 2, 4] in the SQNM platform, as was the case for loci CNVR2906.1 and CNVR8499_full.

3.6 Genotyping on Illumina Platform

The raw intensity data from the Human660W genotyping platform were normalized with Illumina's standard normalization method in the BeadStudio software (Framework version 3.1.3.0, Genotyping Module version 3.3.4.). The normalization algorithm was applied on the

sub-bead pool level and consisted of five main steps; outlier removal, background estimation, rotational estimation, shear estimation, and scaling estimation (Illumina's Genotyping Data Normalization Methods, Pub. No. 970-2006-010). The steps are designed to adjust for global intensity differences, channel-dependent background, and to scale the data. Raw genotyping data was available for 285 samples (including replicates). After excluding replicate samples, and 28 samples with a lower overall call rate in BeadStudio, 242 samples had data for CNV genotyping analysis. These samples had a call rate $> 99\%$ in BeadStudio. The normalized X and Y values, which correspond to normalized signals from alleles A and B for the specific probe, were used as input values in subsequent analyses.

We used CNVtools Barnes et al. (2008) to summarize signal intensity data and to assign samples to discrete copy number classes in 6,236 unique CNV regions. CNVtools performs a principal component analysis (PCA) on a matrix of normalized signal intensities and clusters the result of the PCA; these clusters represent the copy number assignments for the samples. Output from the PCA procedure can then be used in linear discriminant analysis (LDF) to further improve the clustering of the data. For each CNV, we compared the PCA and LDF clustering results and chose the method with higher cluster separation parameter 'Q'. We applied quality control for the raw genotyping results, and included to the final QC+ CNV set only the CNVs, which had; genotyping success rate of at least 90 %, clustering quality Q$>$ 4, Hardy-Weinberg equilibrium test statistic 15 or less (applied only for bi-allelic CNVs and calculated within populations), and less than 3 Mendelian errors in the HapMap trios. In addition, we checked for residual correlation within a copy number cluster by examining signal intensity correlations between the Human660W and the Agilent 105k genotyping experiments. Finally, we checked all QC+ CNV cluster plots manually, to identify any additional problems with the CNVtools copy number assignments.

4 Genomic Overlaps

4.1 CNV overlap with genomic features

Intersection analysis of CNVs with gene annotations and other datasets from a variety of biological databases and published studies was done using custom Perl, python and R scripts (Figure 2.18). To avoid a bias towards genes with multiple annotated transcripts, we also generated a merged RefSeq gene set (downloaded from UCSC on April 30, 2009) that combines all transcripts for each gene. Briefly, transcripts were merged when they overlapped at their mapped genomic positions on the same strand and referenced the same gene identifier (NCBI Entrez ID), taking the minimum and maximum boundaries to define the merged gene structure. There were 35 genes that mapped to more than one chromosome, and in few cases to more than two chromosomes. For example, FAM138A, has been mapped to chromosomes 1, 9 and 19, resulting in three gene models, while other genes in the pseudo autosomal regions (PAR) have been mapped to both chr.X and chr.Y. For our analyses, we maintained each copy as they likely represent distinct paralogous transcripts sharing 100%

sequence similarity. Transcripts that mapped to _hap, _random chromosomes or chrUn, and mitochondria and the Y chromosome were removed from the dataset. Other datasets were obtained and processed as indicated in Figure 2.18. Where needed, feature positions were lifted over to the hg18 genome assembly.

Features were considered as intersecting CNVs if they overlapped by 1 or more bases on either strand. Enrichment or impoverishment of CNV overlaps with different classes of genomic features was further assessed in a permutation analysis where CNVs were randomly reshuffled across the genome, maintaining the size distribution and the number of CNVs per chromosome. In each iteration, the chromosome number for all CNVs was kept the same while the positions were randomized within chromosome boundaries. P-values for enrichment or impoverishment were calculated as the fraction of permutations where the number of overlaps was respectively greater or smaller than the observed value in 1000 permutations.

Dataset	Source
DGV Entries	http://projects.tcag.ca/variation/v7, Mar2009, NCBI 36 (hg18)
DGV_BAC	DGV including BAC studies, DGV v.7, Mar2009, NCBI 36 (hg18)
DGV_noBAC	DGV -BAC studies excluded, DGV v.7, Mar2009, NCBI 36 (hg18)
RefSeq Transcripts, hg18	http://genome.ucsc.edu (May 1st, 2009)
RefSeq Genes, hg18	Non-redundant gene set; 'merged RefSeq' was also generated (supplementary methods)
miRNA	http://microrna.sanger.ac.uk/sequences/; Release/Version 11.0, (Apil 11th 2008), mirBASE { PIMD:17991681}
Promoter	Promoter defined as +/- 500 bp from the TSS
CpG Islands	http://hgdownload.cse.ucsc.edu/downloads.html#human DNA regions >500 bp with a GC content >55% and observed CpG/expected CpG of 0.65 {PMID:11891299}.
Enhancer Elements	VISTA, http://enhancer.lbl.gov/ (negative enhancers were removed)
Ultra Conserved Elements	UCE - elements perfectly conserved over 200 bp or more, Supp Table (http://www.soe.ucsc.edu/~jill/ultra.html) {PIMD:15131266}
Imprinted Genes	Otago, http://igc.otago.ac.nz/home.html
UCSC Duplications	http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/genomicSuperDups.txt.gz; v. created 03-Aug-2006, downloaded 03-Sept-2008
WSSD Duplications	http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/; added Celera Dup positive
Recombination Hotspots	http://www.hapmap.org/downloads/recombination/
OMIM Disease Genes	http://www.ncbi.nlm.nih.gov/Omim/getmorbid.cgi
Repeat Masker	SINE; LINE; LTR; DNA element;Low complexity;Satellite; Other; TRF_SimpleRepeat; http://hgdownload.cse.ucsc.edu
AR/AD Genes	Hand-curated list of 980 gene-phenotype pairs of the OMIM database with phenotypic information about the mode of inheritance and age at onset (hOMIM) {PMID: 18571414}
Dosage Sensitive Genes associated with disease	105K Oligo Array- Baylor's Disorder List (http://www.bcm.edu/geneticlabs/tests/new.html)
Decipher Syndromes	https://decipher.sanger.ac.uk/application/
Pharmaco-genes	Drug or xenobiotics metabolizing enzymes (DMEs or XMEs)- Integrated list from Goldstein2004 and PMID:18032438 [Hernandez-Boussard et al.2008, Pharmacogenetics Knowledge Base (PharmGKB; http://www.pharmgkb.org/)]
Cancer Genes	Cancer Gene Census list, (Dec. 16 2008) [http://www.sanger.ac.uk/genetics/CGP/Census/:CancerGeneCensus_Table_1_full_2008-12-16.xls]
GWAS	http://www.genome.gov/gwastudies/
Levy *et al.* (Venter genome)	{PMID:17803354} (only entries ≥ 400bp used)
Wheeler *et al.* (Watson genome)	{PMID:18421352} (only entries ≥ 400bp used)
Bentley *et al.* (Yoruban genome)	{PMID:18987734}
Wang *et al.* (Asian genome)	{PMID:18987735}
Mills *et al.*	{PMID:16902084} (only entries ≥ 400bp used)
Korbel *et al.*	{PMID:17901297} (only entries ≥ 400bp used)

Figure 2.18: **Reference datasets used in the CNV overlap analyses.**

4.2 Functional classification of genes intersected by CNV loci

To identify enriched functional annotation gene categories, we used the Database for Annotation, Visualization and Integrated Discovery (DAVID; v. April2008) (Huangda et al., 2009). Entrez gene identifiers were used for the upload format. The 3,340 RefSeq genes that were overlapped by 8,599 validated CNVs were tested for enrichment of biological process gene ontology (GO) terms, KEGG, BBID and BIOCARTA pathways, further clustering into function-related gene groups. Cutoffs for enrichment were set at an FDR-corrected EASE score of 0.05, with a minimum of 2 genes per functional category. To check for consistency with an independent classification tool we also used BABELOMICS (v.3.2), which was further used to annotate loci with impoverished gene ontology/functional classes (Al-Shahrour et al., 2006, 2005). The highest-scoring enriched categories were seen by two methods, and are listed in Table SIV together with impoverished categories, after correction for multiple testing.

We primarily identified ontology gene categories that were enriched or impoverished in validated genotyped CNV loci when compared to the genomic background. We further explored enrichment and impoverishment for the loci subgroups deletions, duplications, multiallelic, common and rare.

5 Mutation Mechanisms

5.1 VNTRs, NAHR

VNTRs were identified as all CNVs with greater than 50% of their bases contained within Tandem Repeat Finder annotation. Non-allelic homologous recombination (NAHR) was ascribed as a mutation process using two different approaches. First all CNVs were identified that had one edge within one half of a segmental duplication pair, and the other edge within the other duplication of the same pair. As the segdup definition only identifies homologous sequences larger than 1kb we added a second, more flexible approach to allow for NAHR mediated by shorter sequences. Breakpoint regions were constructed for each CNV using an algorithm that searches for the modal edges in the set of calls comprising that CNV. `Vmatch` sequence analysis software (`www.vmatch.de`) was used to identify the longest stretch of perfect homology between the two breakpoint regions for each CNV. CNVs with at least 20bp of perfect match homology between breakpoint regions were deemed likely to have formed by NAHR. In order to be classified as NAHR, a CNV must have less than 70% of sequence contained in VNTR annotation. This classification leads to 39 CNVs being classified as both VNTR and NAHR.

5.2 Non-B DNA structures

Based on 30 years of nucleic acid research 10 distinct non-B DNA structures have been identified. The formation of these structures is largely driven by the primary sequence and

thus the genome can be annotated with locations likely to form non-B DNA. The sequence motifs of interest are tandem repeats, mirror repeats, inverted repeats and G quartets.

Five Non-B structures were classified with the following rules:

- **Z DNA:** direct tandem repeats where sequence is $(\mathrm{GY.RC})_n$ with $n \geq 6$;
- **triplexes:** mirror repeats + direct tandem repeats composed exclusively of Rs or Ys. Intermolecular triplexes were identified as the sequence $(\mathrm{G.C})_n$ with $n \geq 12$. Intramolecular triplexes: $(\mathrm{G.C})_{12-13}\mathrm{N}_{4-6}(\mathrm{G.C})_{12-13}$, $(\mathrm{G.C})_{14-15}\mathrm{N}_{2-7}(\mathrm{G.C})_{14-15}$, $(\mathrm{G.C})_{16-17}\mathrm{N}_{0-8}(\mathrm{G.C})_{16-17}$, $(\mathrm{G.C})_{18-19}\mathrm{N}_{0-9}(\mathrm{G.C})_{18-19}$.
- **cruciforms:** inverted repeats + direct tandem repeats composed exclusively of Rs or Ys.
- **slipped DNA:** direct tandem repeats where sequence is $(\mathrm{A.T})_n$ with $n \geq 12$.
- **tetraplex structure:** G quartets.

Whole genome annotations of these non-B DNA forming structures were generated using software kindly provided by Jack Collins of NCI (available from `http://ncisgi.ncifcrf.gov/~collinsj/pgms/`).

5.3 Motif discovery

Our characterization of breakpoint sequences up to this point involved testing hypotheses about the enrichment of specific sequences suggested by prior research. We also took a hypothesis-free approach to characterizing CNV breakpoints by trying to identify any small motifs over-represented in our set of breakpoints sequences. A simple algorithm was used to define breakpoint regions for each CNV; the target size for each breakpoint region was 550bp, but in the case of small CNVs the window was necessarily smaller. In order to maximize the chance of identifying a novel rearrangement process, we focused on CNVs without any obvious etiology: we removed VNTRs and CNVs flanked by segdups or long stretches of perfect homology. Each sequence was then processed with "dust" (Tatusov, RL and Lipman, DJ. unpublished), removing homopolymer and simple sequence repeats; only sequences with greater than 20 non-missing bases were retained for further analysis, leaving 6516 sequences. After these processing steps, 80% of breakpoint sequences were 550bp in length, with 192 smaller than 550bp and 596 sequences > 1kb.

If one were to test for over-representation of a given motif by exhaustively enumerating all motifs, it would not be possible to test for motifs of even moderate size (> 10bp) as the search space grows exponentially with motif length. Instead we used a stochastic search approach implemented by nestedMICA (Down and Hubbard, 2005). Briefly, this approach uses a novel Bayesian hill-climbing algorithm called "nested sampling" to simultaneously learn multiple motifs represented in a set of input sequences (first proposed by John Skilling at Cambridge University, http://www.inference.phy.cam.ac.uk/bayesys/). The input sequences

are modeled as a hidden Markov Model; this model has hidden states corresponding to the background genomic sequence (without motifs), a user-defined number, 'n', of motif states, and silent states not responsible for modelling any sequence. Inference essentially amounts to fitting parameters for the most likely n motifs, which results in n position weight matrices (PWMs) of variable length.

We ran nestedMICA with parameter settings that would identify the top 20 most enriched motifs, from 6-13bp in size: "nminfer -threads 4 -checkpointInterval 1000 -distributed -port 1024 -mixtureUpdate weakResample -revComp". The algorithm ran for 80,000 updates, and convergence was confirmed by visual inspection of likelihoods. The resulting PWMs are presented in Supplementary Notes.

Interpretation of the resulting motifs is not a trivial task. We compiled list of about 200 common transposable element sequences and low complexity DNA motifs (e.g. simple repeats, homopolymers, etc.) and scanned this list with PWMs for each motif and recovered all motif hits with bits-sub-optimal score > -5. We identified one 13mer with perfect hits to many SVA and Alu subfamily consensus sequences. We mapped the location of this hit on Alu secondary structure and noticed that it resides within a conserved SRP9/14 recognition site on the left monomer (Figure 2.19).

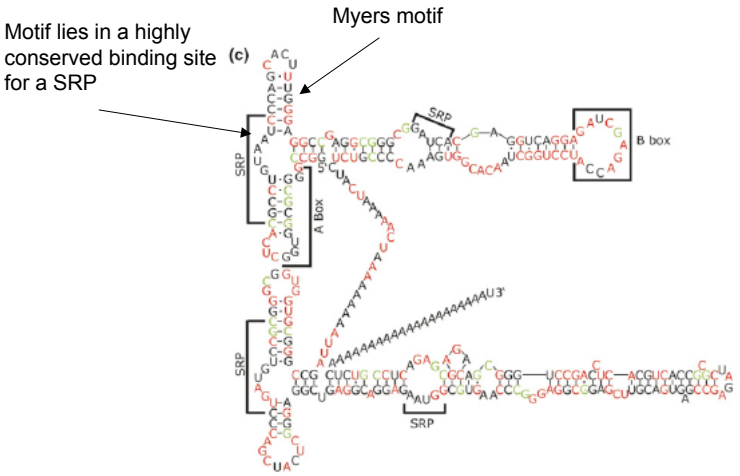


Figure 2.19: **Mapping of breakpoint motif on *Alu* secondary structure.** The motif match begins at position 25 of this AluY consensus, reading from the figure "GUAAUC..". We have also labeled the location of a close match to the Myers recombination hotspot motif (Myers et al., 2008). This figure is adapted from (Mills et al., 2007). Base colorings are from their analysis and do not pertain to ours.

5.4 Genomic annotations used

In the next section, we describe our analysis of motifs at CNV breakpoints. In addition to testing motifs producing B-DNA structures (listed above) and the novel motifs identified by nestedMICA, we also examined two types of annotation (breakpoint annotation and CNV annotation) derived from the UCSC genome browser tables and other sources. Here we try to list these sources using the names presented in Figure 3a of the main text. All annotations were with respect to NCBI36.

I. Breakpoint Annotations. These are annotations that are added to a CNV if one or more of the elements fall within one or both of the breakpoint regions of that CNV (unless otherwise specified).

- Simple repeats (UCSC, only those < 100bp).
- dbSNP indels, from dbSNP build 129, only those $< 500bp$.

- segdup_pairs. This is derived from UCSC table "Segmental Dups". A CNV is annotated with this feature only if the left edge falls into one half of a segdup pair, and the right edge falls into the other half of the same pair.
- CpG islands (from UCSC).
- degen myers - We annotated this motif on the genome ourself using the "degenerate" 13bp form of the hotspot motif reported in Myers et al. (2008).

II. CNV annotations. These are annotations that are added to a CNV if the entire CNV region has 50% reciprocal overlap with the

- Simple repeats (UCSC, only those > 100bp).
- Uniq_segdups (UCSC table "Segmental Dups", summarized to remove redundant events)
- Pseudogenes (from the Gerstein lab (`www.pseudogene.org`).

5.5 Hypothesis testing

We tested two sets of hypotheses: 1. Are any motifs of interest (identified as *a priori* candidates or through nestedMICA analysis) overrepresented at the breakpoints of CNVs, 2. Do duplications and deletions differ in their motif enrichment?

To test for breakpoint enrichment it is necessary to have a null distribution for comparison. We decided to use sequence closely flanking each CNV to construct this null as it should be fairly well matched for possible confounders like GC content. The other issue is the scale at which the motifs are enriched. For some mutation processes the motif by definition will contain the breakpoint (eg. segmental duplications and NAHR), but one could imagine that the enrichment may be more diffuse for other motifs.

We created data sets that we refer to as "aligned feature summaries" for each motif of interest. Each summary is parameterized by a "window" size, w, and a number of bins, n. A grid of n bins of size $\frac{w}{n}$ is constructed to the left and right of each of the k CNVs. Likewise we construct a grid of $\frac{w}{n}$bp bins internal to each CNV, the number of bins depending on CNV size but not exceeding n. In the case that the CNV size is greater than $w$bp, equal sized grids of $\frac{n}{2}$ bins are constructed internally from the left and right breakpoints. The result of this griding process is a set of (at most) $3n$ bins centered on each CNV. We tabulate the density of the motif in each of the $3n$ bins for each CNV and then average across CNVs. If D_{ij} is the density of the motif for bin i at CNV j, the total average density in the ith bin is

$$D_i = \frac{\sum_{j=1}^{k} D_{ij}}{k}$$

To assess the significance of observed patterns of motif density, we use bootstrap resampling to create 95% confidence intervals on the observed motif densities in each bin. We create 1000 datasets of k CNVs each using sampling with replacement and find the 25th and

975th ordered value of each D_i across all samples. Ninety-five percent confidence intervals created this way suggested that slipped DNA, G quadruplexes, hotspot motif, dbSNP indels, CpGs, simple repeats > 100bp, segmental duplications were enriched at CNV breakpoints, while cruciforms, Z-DNA, and triplexes were not enriched.

We next tested the hypothesis that deletion and duplication breakpoints are associated with distinct sequence features. Multiallelic events were treated as duplciations. A dataset was created of 615 duplications and 2532 deletions for which the breakpoint regions were 550bp on both the left and right edge. For each motif in our list of candidates, we scored each CNV as spanning no copies of the motif (0) or at least one copy (1). The proportion of duplications and deletions containing each motif was tabulated and the statistical significance of the differences in proportion for each motif was assessed by permuting deletion/duplication labels across CNVs. In all cases of significant difference ($p < .05$) the difference was an enrichment of motif usage in duplications over deletions. Segmental duplication pairs and G-quadruplexes (6.5 fold and 1.75 fold enriched) were highly significant ($p < .001$), while short indels were mildly enriched ($p < .05$).

A similar analysis was conducted looking at the sequence content of entire CNVs; instead of scoring motif placement in breakpoint regions, we scored the proportion of deletions and duplications with > 80% reciprocal overlap with various genomic annotations. Segmental duplications (20 fold), CpGs (4 fold), and VNTRs over 100bp (3 fold) were enriched in duplications over deletions.

5.6 Dispersed Duplications

There are a very small number of inter-chromosomal duplications that are known to be polymorphic in man (Wong et al., 1990; Doggett et al., 2006); it is thus of great interest to identify additional cases. Such knowledge will aid in the interpretation of GWAS results, and improve our understanding of mutation mechanisms.

Genomic overlaps. We hypothesized that polymorphic inter-chromosomal duplications longer than 1kb in the reference genome sequence will be annotated as segmental duplications, and that we might be able to identify a subset of such variants that are also present in our CNV map. We downloaded the "Segmental Duplication" track from NCBI36 and identified interchromosomal segmental duplication pairs where both members showed > 90% reciprocal overlap with a CNV in our map.

We obtained a set of 23 validated, unpublished germline interchromosomal CNVs from the Cancer Genome Project (Campbell et al., 2008). Cross-referencing these with our CNV map we were able to identify 4 polymorphic inter-chromosomal duplications; these were validated by PCR genotyping of each duplication in the Plate 1 HapMap samples (Supplementary Notes).

Greater consideration of the possible mechanisms leading to an inter-chromosomal duplication suggested several additional strategies to identify polymorphic retrogenes. In the case of a processed pseudogene, only the exonic sequence will be copy number variable. As the

exons are typically much smaller than the introns, and we may have only 1 (or no) probes in any given exon of a gene, we don't expect the entire gene to show CNV in the case of a processed pseudogene.

Sequence analysis. We used DNA sequence analysis to identify 14 strong candidates of retroposition in our CNV map. Briefly, we looked for a poly-A tail at one end of the CNVR and target site duplications flanking the CNVR, as well as a clear homology match of the CNV sequence to somewhere else in the genome.

in silico **splicing.** We created a set of "spliced" CGH data for each gene in RefSeq; this consisted of creating one vector of all exonic data for each gene and each individual. We calculated two statistics on each combination of gene and sample: "slope" and "t-test". "Slope" starts by fitting a regression of probe position vs. log2 ratio for just the exonic probes (with estimated regression coefficient $\hat{\beta}_1$) and intronic probes ($\hat{\beta}_2$). Under the null of no retrogene $\hat{\beta}_1 = \hat{\beta}_2$; our statistic is $\frac{\hat{\beta}_1 - \hat{\beta}_2}{SE(\hat{\beta}_2)}$. The "t-test" statistic is just the t-test statistic for difference in mean between the intronic probes and exonic probes,

$$t = \frac{\hat{\mu_1} - \hat{\mu_2}}{\hat{\sigma}_{pooled}\sqrt{\frac{2}{n}}}$$

which should be equal to 0 under the null. We retained all genes that had greater than 5 t-tests with Bonferroni-corrected significance. The goal here is to identify polymorphic processed pseudogenes not detected on the discovery array. Therefore we removed all significant genes that spanned a CNV already in the discovery map.

Inter-chromosomal LD. Finally, we looked for signatures of inter-chromosomal LD between CNV genotypes and SNPs using the Phase II HapMap data (release 23a). Briefly, we used a published genetic map (Frazer et al., 2007) to define hotspot intervals around each CNV (the interval from the nearest hotspot 5' of the CNV to the nearest hotspot 3' of the CNV). We then found the maximum Pearson correlation between the CNV genotypes and a) SNPs within the hotspot interval and b) SNPs outside the hotspot interval. This resulting data was used to construct a list of "R^2 differences" ($R2D = R^2_{in} - R^2_{out}$). After filtering our list to CNVRs with $R2D > 0$, we binned each CNVR into four categories:

- A: Multiple SNPs with high R^2_{out} on the same chromosome, but different from the R^2_{in} chromosome.
- B: A single high R^2_{out} SNP on a different chromosome from the R^2_{out} chromosome.
- C: Multiple SNPs on more than one chromosome.
- D: Single or Multiple SNPs all on the same chromosome as the R^2_{out} chromosome.

We retained 58 likely inter-chromosomal events in which the $R2D > .5$ in at least one population, the CNV was called a duplication or multi-allelic event, and was in categories A or B.

6 Analysis of ascertainment

Previously published population studies of CNV have only ascertained events from the largest end of the CNV size spectrum. We are interested to know the proportion of common variants (MAF $> 5\%$) that we have ascertained in the discovery phase of the current project. Our data is the CNV map from the discovery phase and the site frequency spectrum (SFS) of CNV from the genotyping phase.

To explore this issue we used a Beta-Binomial framework for modeling the process of sampling CNVs from the underlying population; the Beta distribution models the complete frequency distribution of CNVs in the CEU population and the CNVs present in our discovery sample are binomial samples from that frequency distribution. However, there are two additional layers of sampling involved here (Figure 2.20). First, our discovery array has incomplete power to detect CNVs; the CNV map we construct in the discovery phase represents only a subset of all copy number variation present in the discovery samples. Second, while we included all CEU CNVs detected in the discovery phase on the 105K genotyping chip, we were only able to genotype a subset of these leading to another level of sampling. Our ultimate goal is to estimate parameters of the underlying Beta distribution, which will required us to model the various sampling processes leading to the genotype set.

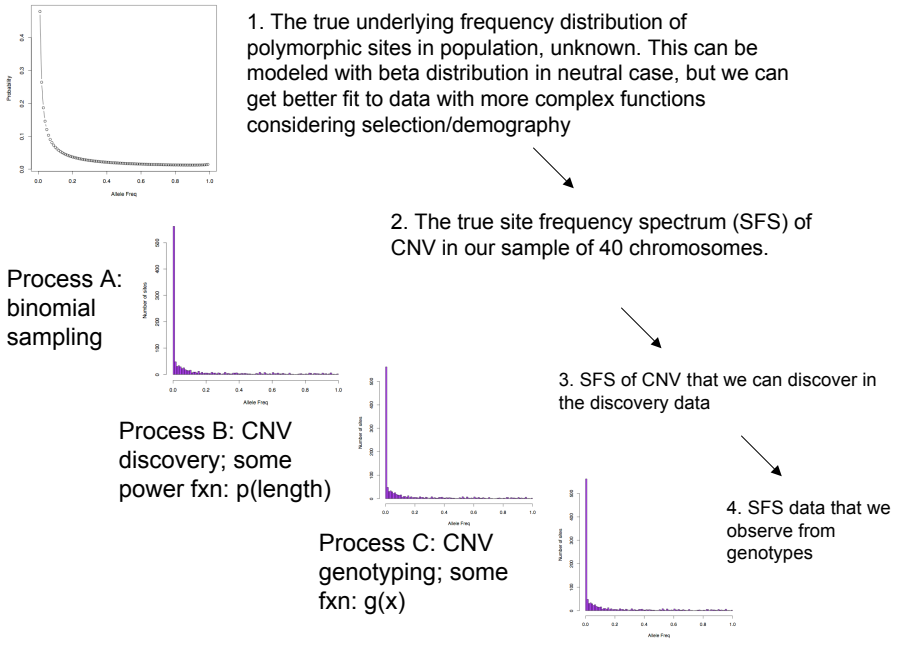


Figure 2.20: **Sampling processes generating genotype data.** There are 3 sampling processes leading to the set of genotype data analyzed in this section. Our goal is to make inferences on the underlying CNV frequency distribution using the discovery and genotype data.

We consider size to be the primary characteristic that influences CNV detection power, and we model our power to detect a variant in a single experiment, $p_1(l)$, as a simple linear function of length. To construct this power curve we compare the genotyping data and discovery CNV calls for the 21 samples typed on both platforms. Homozygous CNV genotypes are far less common than CNV heterozygotes, and we make the conservative assumption that power to detect a sample that has 2 alleles different from the reference will be the same as a sample with 1 allele different from the reference.

For each genotyped CNV, we estimate the false negative rate (FNR) in the discovery data as the proportion of samples with a genotype call different from the CGH reference individual but not called. We define power as 1-FNR (Figure 2.21).

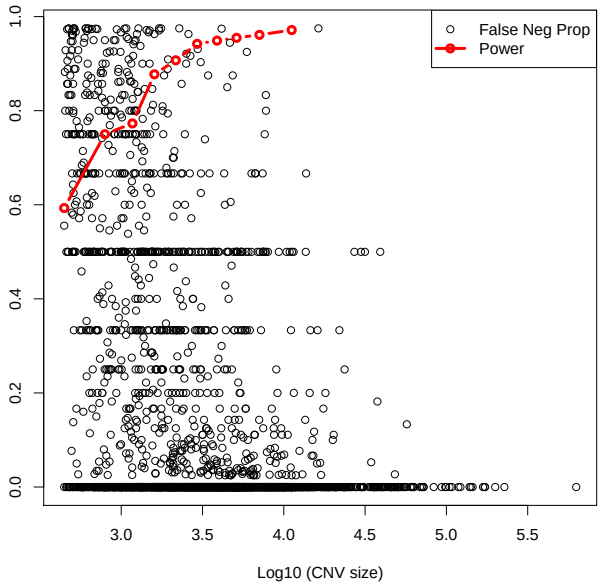


Figure 2.21: **False negative rates and power as function of CNV size.** A black point representing the false negative proportion for each genotyped CNV is plotted against CNV size. The power function, constructed by binning CNVs into deciles of length, is plotted as a red line using the same scale.

This type of power estimate is the probability of detecting a variant in a single sample; however we are interested in the detection probability of a locus. The power of detecting a locus is a function of the per sample power, $p_1(l)$ and the allele frequency, f, at that locus; we denote this power as $p_2(l, f)$. More explicitly,

$$p_2(l,f) = 1-(1-p_1(l)^a)*(1-f)^2+(1-(1-p_1(l)^b)*2f(1-f)+(1-(1-p_1(l)^c)*f^2) \quad (4)$$

where a, b, and c represent the expected numbers of samples without the minor allele, not heterozygous, and without the major allele, respectively. The formula can be understood to represent a weighted average of three power estimates, one for each possible genotype of the CGH reference individual. An example locus power curve is given in Figure 2.22.

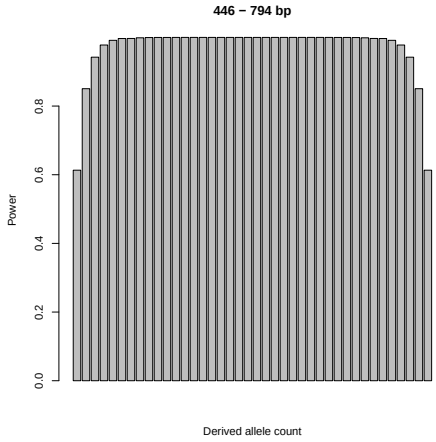


Figure 2.22: **Power curve to detect a CNV locus in a sample of 40 chromosomes.** Each bar corresponds to the power to detect a locus in the CGH discovery experiment, as a function of derived allele count; going from 1 copy on the left to 40 copies on the right. The power indicated on y-axis is calculated with Eq. 4. using CNV data from the bottom decile of the CNV length distribution (446-794bp).

To model the sampling of loci introduced by genotyping, we make the assumption that probability of genotyping success is equal for all CNVs in the discovery set. Therefore there is a constant loss of observations in each bin of the SFS, $1/g$, which we estimate as 4978/8599=0.59. Note that this does not take into account redundancy in the validated CNVs that has been eliminated in the QC+ genotyping set.

Next we divided all CNVs into deciles of length, $l_i, ..., l_{10}$, and constructed power functions for each decile, $p_2(l_i, x)$. Let N_{x_i} be the number of CNVs of frequency x in decile i. We combine the loss of loci due to genotyping, the locus specific power curve and the observed SFS data to estimate the true SFS in the discovery samples:

$$E[X = x] = \sum_{i=1}^{10} *N_{x_i} * 1/p_2(l_i, x) * g \tag{5}$$

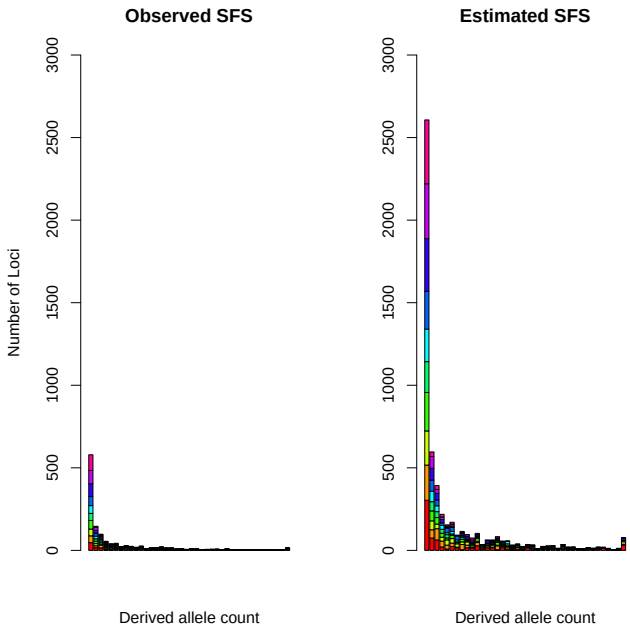


Figure 2.23: **Observed and estimated SFS of CNVs.** On left, the actual SFS observed by genotyping the 20 samples used in the discovery phase. Each frequency bin is divided into deciles of CNV length. Coloring of deciles is consistent across bins. On right is the analogous SFS plotted for the estimated SFS calculated by Eq. 5.

Based on these calculations there are 5500 potentially assay-able CNVs > 450bp segregating in the CEU discovery samples, 5000 (91%) of which were detected by CGH (Figure 2.23).

This addresses the sampling due to discovery power, but there is the additional issue of sampling from the population; we have not yet considered that our discovery sample is a subset of 40 chromosomes from the entire European "population". Under neutrality, the expected frequency distribution can be modeled by a beta distribution of the form

$$f(x) = \frac{x^{(\frac{\theta}{2}-1)} * (1-x)^{(\frac{\theta}{2}-1)}}{B(\frac{\theta}{2}, \frac{\theta}{2})}$$

where θ is the population scaled mutation rate and $B(a, b)$ is the beta function (Wright, 1951). We fit this neutral beta density to our expected SFS calculated in Eq. 5, obtaining an MLE for the whole-genome θ. Assuming a total assay-able genome size of 3Gb our estimate translates to a base-pair $\theta = 4.3 \times 10^{-7}$.

Considering just CNVs greater than 5% MAF, the expected number of sites with exactly i copies of the derived allele in a sample of n chromosomes is

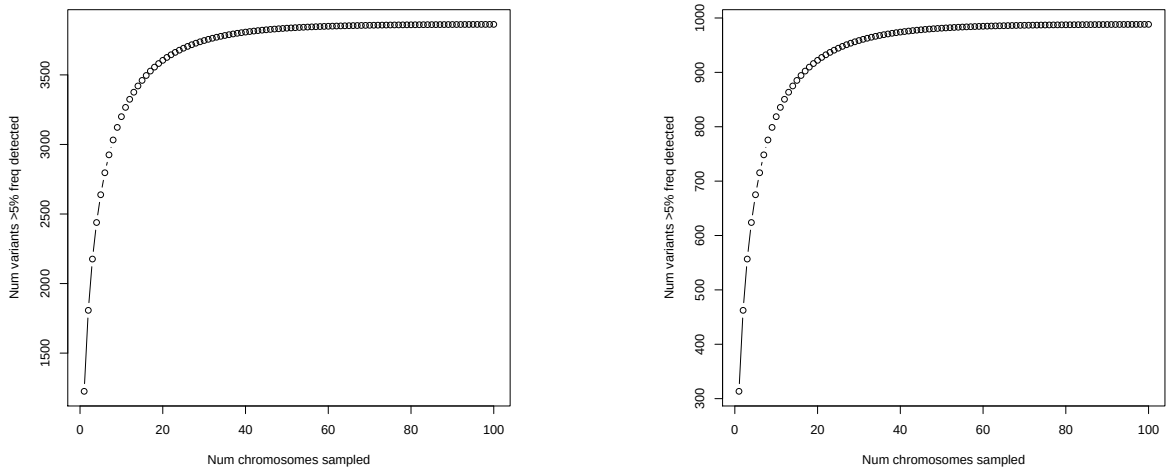
$$F(i, n) = \int_{0.05}^{0.95} \binom{n}{i} x^i (1-x)^{n-i} f(x) dx$$

and the total number of sites where we expect to observe a variant of 5% or greater is

$$\sum_{i=1}^{n} F(i,n). \tag{6}$$

In Figure 2.24a we plot Eq. 6 over a grid of sample sizes; from analysis of the asymptotic behaviour of the function we estimate that there are 3797 CNVs $> 5\%$ MAF and > 450bp in the European population from which our discovery sample was drawn. We estimate that we have genotyped 939 (25%) of these.

After consideration of the power curve in Figure 2.21, we believe that all CNVs greater than 3kb that are present in the discovery sample have been identified. We present a parallel analysis focusing on just these variants for comparison in Figure 2.24b. We estimate that there are 971 variants $>$ 3kb in size, $> 5\%$ MAF and that we have genotyped 426 (44%) of these.



(a) $>$ 450 bp

(b) $>$ 3kb

Figure 2.24: **Expected number of sites with minor allele freq $> 5\%$ as a function of number chromosomes sampled for (a) all CNVS $>$ 450bp long and (b) all CNVS $>$ 3kb long.**

7 Mutation rate

We used the ascertainment-adjusted data described above to formulate population-genetic estimates of the autosomal CNV mutation rate. First, we use the number of ascertainment-adjusted segregating sites in the CEU samples, 5437, as input to the Watterson etstimator of the population-scaled mutation rate (Watterson, 1975). Assuming an effective population size of 10,000, this translates into an estimate of .03175 CNV events per haploid autosomal

genome per generation. Fitting a Beta distribution corresponding to the expected stationary allele frequency distribution to the adjusted SFS, described above, yields a value of θ corresponding to .035 *de novo* CNVs per genome per generation. Both of these estimate this agrees with what has been observed in the few studies of large *de novo* events which have produced observations of one de novo event per $1 - 10\%$ of trios examined (Marshall et al., 2008; Sebat et al., 2007).

8 LD Analyses

8.1 Phasing

We investigated several methods for phasing CNV genotype calls onto haplotypes defined by HapMap SNPs, including PHASE 2.1, fastPHASE 1.2, and BEAGLE 3.0.3 (Marchini et al., 2006; Scheet and Stephens, 2006; Browning and Browning, 2007). BEAGLE 3.0.3 was the best fit to our need of an accurate phasing algorithm that uses trio information, provides imputation, and is computationally efficient.

Imputation is a particularly important aspect of phasing for our data, because genotyping accuracy for CNVs is still much lower than that for SNPs. Our strategy to improve the overall call quality of the CNV genotypes is to combine highly accurate SNP genotype calls with a model of haplotype structure to impute CNV status for individuals with noisy or ambiguous intensity data.

To assess the performance of BEAGLE, we conducted a series of analyses using the Phase II HapMap CEU genotype calls and our CNV genotype calls from chromosome 22; in this section we will simply use "genotypes" to refer to this combined set unless the label "CNV" or "SNP" is explicitly given. We chose to restrict our analyses to chromosome 22 as, in addition to defining a small data set, it has one of the highest recombination rates of all chromosomes, and thus may represent a lower bound on expected imputation and phasing accuracy.

First we used the family structure of the CEU trios to determine the phase of our genotype calls at all sites that can be unambiguously determined. Genotypes were phased with BEAGLE 3.0.3 using the "trios=" option and default settings. Sites of data missing in the empirically phased haplotypes (due to phase ambiguity) were set to missing data in the BEAGLE-phased haplotypes as well.

The following procedure was run for both sets of haplotypes. All monomorphic sites were removed, and 500 loci were selected at random from the remaining data. For each of these 500 "target" loci, a window of 200kb was defined that centered on the start of the locus, and pairwise r^2 was calculated between the "target" and all variants inside the window, producing a set of 127,000 r^2 values.

Comparison of these paired r^2 values reveals that the model-based phasing implemented in BEAGLE 3.0 is extremely accurate (Figure 2.25). Eighty-three percent of all r^2 pairs are identical, and 86.3% of r^2 pairs with both $r^2 > 0.5$ are identical. One important analysis in

the paper is the identification of tag SNPs for the CNVs on the genotyping array. Among the 2768 variant pairs with $r^2 > .85$ in the empirical haplotype set, only 10 (0.3%) of these have $r^2 < .7$ in the model-based haplotype set.

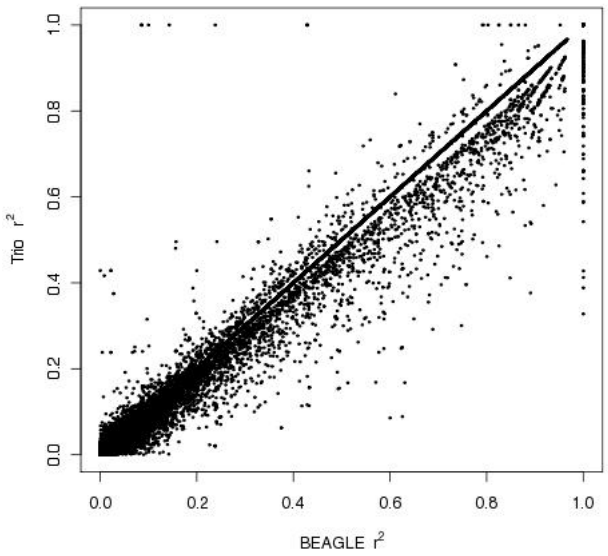


Figure 2.25: **Comparison of r^2 estimates from empirical and model-based phasing.**

To measure the imputation accuracy of BEAGLE 3.0.3 we created artificial missing data by masking a subset of the unphased genotypes and then counted concordance between the true and imputed genotypes for these masked observations. In detail, we masked 10% of the genotypes of each parent in the CEU trio sample. The entire CEU data (parents and offspring) were phased twice, once with and once without using the family information. Imputation accuracy averaged across all loci and samples was extremely high: the median genotype concordance rate was 99.66%, and the minimum was 99.2%. Curiously, the trio information only added an extremely small boost to imputation accuracy, changing concordance rates at the fourth decimal place (hundreths of percents). Computation time was cut in half when using trio information (from 7 minutes to 3.5 minutes for 55000 markers).

8.2 Tagging

CNVs with integer-value copy number were integrated into genotypes from HapMap release 23a. All genotypes in the integrated set of SNPs and CNVs leading to a Mendelian error in one of the CEU or YRI trios were replaced with missing data (in the Mendelian error trio, only), and all variants with minor allele frequency $< 5\%$ were removed (on a by-population basis). After this cleaning step, genotypes were phased using BEAGLE 3.0.3

with the "trios=" setting for CEU and YRI samples, and "unphased=" setting for the JPT+CHB samples.

The 15 June, 2009 release of the NHGRI GWAS association table was used to identify GWAS hit SNPs in high LD with HapMap CNVs. Two sets of correlations were estimated for each population: Pearson correlation between CNV intensities and hit-SNP genotypes and r^2 between phased CNV and SNP alleles.

For all tagging analyses, we used a dynamic window size to search for tag SNPs to each CNV. The left end of the window was defined as the first recombination hotspot from the Oxford genetic map 5' from the left end of the CNV, or 100kb from the left end of the CNV, whichever was larger; the right end was defined likewise in the 3' direction.

9 Selection analyses

9.1 Population differentiation

We calculated V_{st} for using the log2 ratios for each probe on the 42M discovery array, omitting individual NA15510 (who is of unknown ancestry) and including a datapoint for NA10851 (the Caucasian reference individual). V_{st} calculations were the same as described in (Redon et al., 2006).

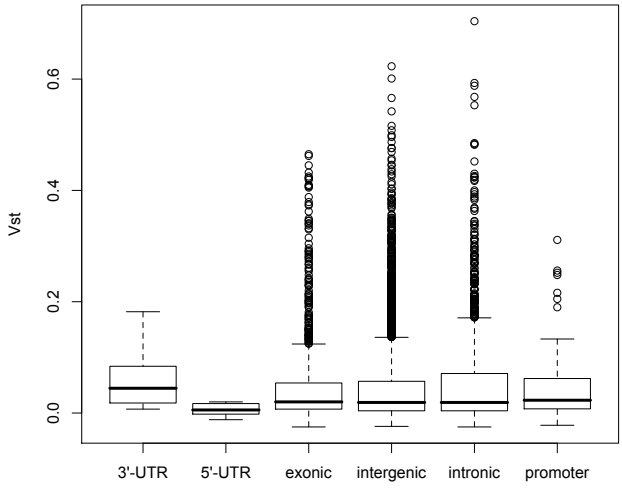


Figure 2.26: **Population differentiation does not depend on functional content of CNV.**

9.2 Frequency spectrum

The site frequency spectrum of variation (SFS) in a sample of chromosomes contains information about the history of mutation, demography and selection in the genealogy of those chromosomes. We used the Poisson Random Field (PRF) methodology implemented by the program `prfreq` (Boyko et al., 2008) to model the site frequency spectrum of CNVs ascertained in the discovery phase and thus make inferences on the population genetic forces acting on CNV.

A central strength of the discovery project is that it allows careful consideration of the ascertainment of our variants. The models implemented in `prfreq` are predicated on unbiased ascertainment of variation in the sample under study.

The extent to which CNV conforms to the standard neutral model is a fundamental unanswered question, as is the relative impact on fitness of SNPs and CNVs. A typical population genetics convention is to describe the reproductive disadvantage associated with samples that are heterozygous s, or homozygous, $2s$, for a mutant allele. We wanted to test the hypothesis that the scaled selection coefficient $\gamma = 4N_e s$ is equal to 0 for various classes of CNV, and compare estimates of γ for CNV and SNPs.

In order to accurately resolve the various forces acting on the SFS it is necessary to consider demography and mutation as well. Due to their ease of genotyping, multitude and lower mutation rate, SNPs are far better than CNVs for characterizing demography. We chose to use the demographic model fitted to sequencing data from 20 Europeans that was reported in (Boyko et al., 2008). We therefore restrict our analyses in this section to CNV ascertained from the 20 CEU samples in our discovery panel.

We fitted the "single point mass" model of γ for to the ascertainment-adjusted SFS for 3 categories of CNV: exonic, intronic, and intergenic. A maximum likelihood estimate of γ was obtained by exploring over a grid of γ values using multinomial likelihoods. The likelhood of the data at this MLE is L_γ. To test the null hypothesis of $\gamma = 0$ we calculate the likelihood of the data at $\gamma = 0$ (call it L_0) and compare the likelihood ratio test statistic, $2\log(L_o/L_\gamma)$, to a chi-square distribution with 1 degree of freedom.

In Figure 5a of the main text, we plot the expected SFS for the different CNV classes. Here, in Figure 2.27 and Figure 2.28, we present the observed SFS and ascertained-adjusted SFS for each class.

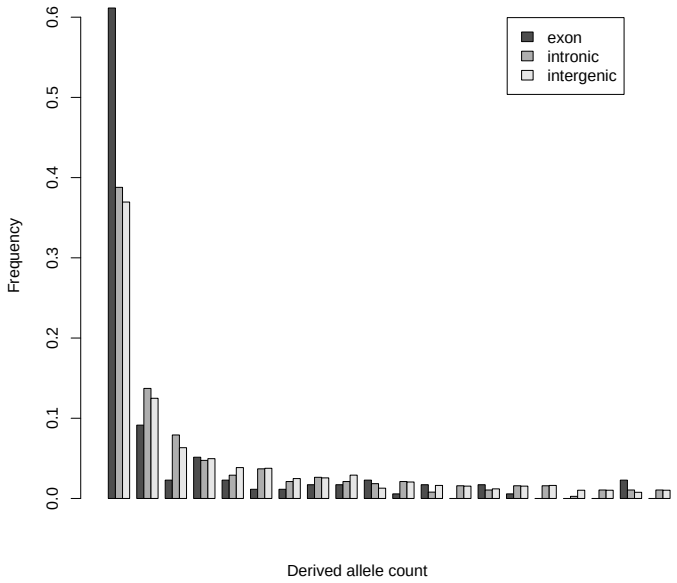


Figure 2.27: **Raw SFS.**

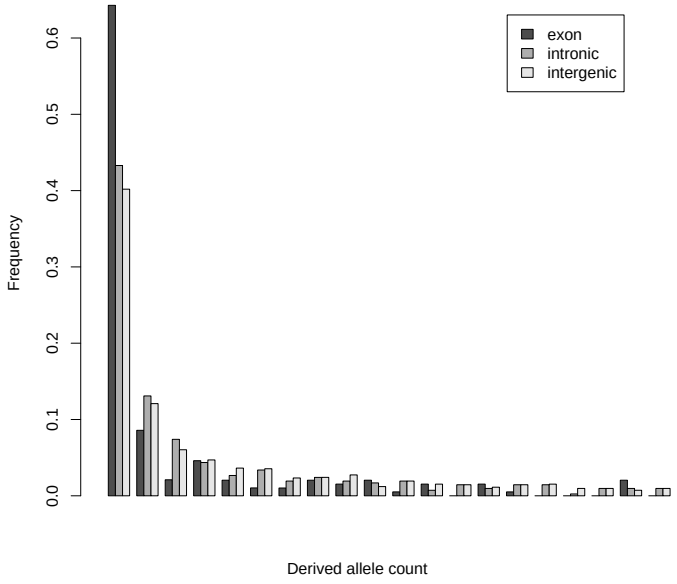


Figure 2.28: **Ascertainment-corrected SFS.**

9.3 Haplotype-based statistics

A separate dataset of phased haplotypes was constructed for calculating haplotype-based selection statistics. Prior to phasing, all SNPs with a MAF < 0.05 and SNPs not present in all 3 populations were removed.

Assignment of ancestral states was made using data from dbSNP to find chimp/orang/macaque orthologous alleles where possible. In the small number of cases where outgroup data was not available the YRI major allele was taken as the ancestral state. Calculation of raw iHS and XP-EHH scores was performed using C programs kindly provided by Joe Pickrell and Sridhar Kudaravalli (U. Chicago).

Unstandardized iHS scores were standardized by binning all variants (CNVs and SNPs) in 20 equally-spaced bins of derived allele frequency and performing a Z transformation on the data in each bin separately (following (Voight et al., 2006)). An interesting open question is wether the strength of selection differs between CNVs and SNPs; key to answering this question will be carefully understanding differences in ascertainment between the two forms of variants in any given dataset. In Figures 2.29 and 2.30 we compare the distribution of XPEHH and iHS scores for SNPs and CNVs.

We observed that the distributions of XPEHH for CNVs and SNPs the CEU-ASN distributions are similar, but for CEU-YRI and ASN-YRI they are quite different. For example, the P values by t-test are:
ceu vs. yri: 3.11×10^{-49}
asn vs. yri: 6.66×10^{-46}
ceu vs. asn: 0.62

In the significant cases, XPEHH values for CNVs are more negative than for SNPs, which in some convoluted way is telling us that the ratio of YRI/non-YRI homozygosity is greater around CNVs than for SNPs.

For standardized iHS P values for different means between SNPs and CNVs are non-significant for CEU and ASN, and $< 3 \times 10^{-6}$ for YRI. CNVs are more negative, which in this context suggest less HH on CNV haplotypes than SNP haplotypes (our iHS signs are flipped compared to Voight et al). We are not at all confident that these differences are meaningful, and further investigation is warranted into possible biases in the calculation of iHS/XPEHH related to variant "type" (ie. CNV or SNP).

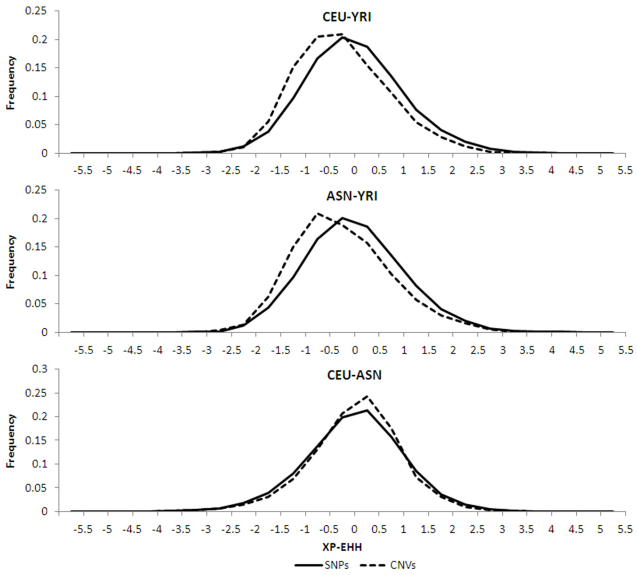


Figure 2.29: **Comparison of XP-EHH scores for CNVs and SNPs.**

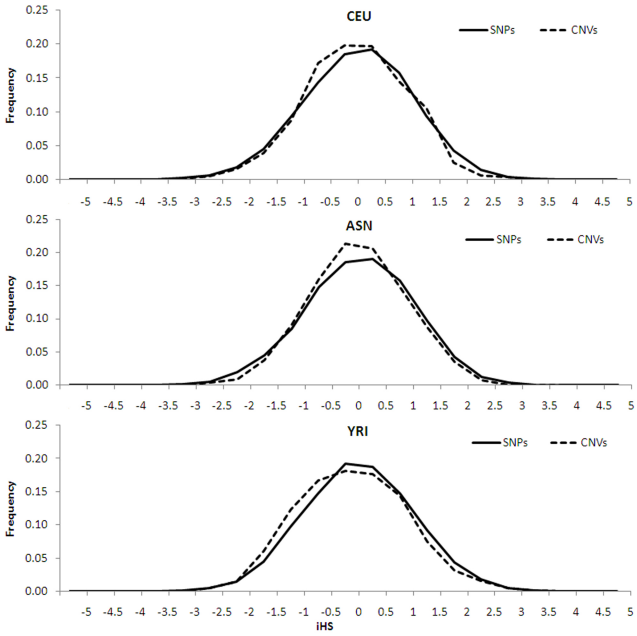


Figure 2.30: **Comparison of iHS scores for CNVs and SNPs.**

References

F. Al-Shahrour, P. Minguez, J. M. Vaquerizas, L. Conde, and J. Dopazo. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res*, 33:W460–4, 2005.

F. Al-Shahrour et al. BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res*, 34:W472–6, 2006.

C. Barnes et al. A robust statistical method for case-control association testing with copy number variation. *Nat Genet*, 40:1245–52, 2008.

D. R. Bentley et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–9, 2008.

A. R. Boyko et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, 4:e1000083, 2008.

S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81:1084–97, 2007.

P. J. Campbell et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, 40:722–9, 2008.

D. F. Conrad, T. D. Andrews, N. P. Carter, M. E. Hurles, and J. K. Pritchard. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet*, 38:75–81, 2006.

A. J. deSmith et al. Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum Mol Genet*, 16:2783–94, 2007.

C. Ding and C. R. Cantor. A high-throughput gene expression analysis technique using competitive pcr and matrix-assisted laser desorption ionization time-of-flight ms. *Proc Natl Acad Sci USA*, 100:3059–64, 2003.

N. A. Doggett et al. A 360-kb interchromosomal duplication of the human HYDIN locus. *Genomics*, 88:762–71, 2006.

T. A. Down and T. J. Hubbard. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res*, 33:1445–53, 2005.

G. P. Elvidge, T. S. Price, L. Glenny, and J. Ragoussis. Development and evaluation of real competitive PCR for high-throughput quantitative applications. *Anal Biochem*, 339: 231–41, 2005.

K. A. Frazer et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–61, 2007.

W. Huangda, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4:44–57, 2009.

International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437: 1299–320, 2005.

J. M. Kidd et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453:56–64, 2008.

P. M. Kim et al. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res*, 18:1865–74, 2008.

J. O. Korbel et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318:420–6, 2007.

J. M. Korn et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*, 40:1253–60, 2008.

S. Levy et al. The diploid genome sequence of an individual human. *PLoS Biol*, 5:e254, 2007.

J. Marchini et al. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet*, 78:437–50, 2006.

J. C. Marioni et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol*, 8:R228, 2007.

C. R. Marshall et al. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet*, 82:477–88, 2008.

S. A. McCarroll et al. Common deletion polymorphisms in the human genome. *Nat Genet*, 38:86–92, 2006.

S. A. McCarroll et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*, 40:1166–74, 2008.

R. E. Mills et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*, 16:1182–90, 2006.

R. E. Mills, E. A. Bennett, R. C. Iskow, and S. E. Devine. Which transposable elements are active in the human genome? *Trends Genet*, 23:183–91, 2007.

A. Del Mistro. Calculating ec50 measurements using massarray. Technical report, SEQUENOM, 2005.

S. Myers, C. Freeman, A. Auton, P. Donnelly, and G. McVean. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet*, 40:1124–9, 2008.

P. Oeth, D. Correll, and C. Jurinke. Quantitative gene expression using competitive pcr and massarray. Technical report, SEQUENOM, 2003.

G. H. Perry et al. The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet*, 82:685–95, 2008.

R. Pique-Regi et al. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics*, 24:309–18, 2008.

R. Redon et al. Global variation in copy number in the human genome. *Nature*, 444:444–454, 2006.

P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78:629–44, 2006.

J. Sebat et al. Strong association of de novo copy number mutations with autism. *Science*, 316:445–9, 2007.

MassARRAY QGE Software User's Guide. SEQUENOM, 2005.

E. Tuzun et al. Fine-scale structural variation of the human genome. *Nat Genet*, 37:727–32, 2005.

E. S. Venkatraman and A. B. Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23:657–63, 2007.

B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A map of recent positive selection in the human genome. *PLoS Biol*, 4:e72, 2006.

J. Wang et al. The diploid genome sequence of an Asian individual. *Nature*, 456:60–5, 2008.

G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7:256–76, 1975.

D. A. Wheeler et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452:872–6, 2008.

Z. Wong, N. J. Royle, and A. J. Jeffreys. A novel human DNA polymorphism resulting from transfer of DNA from chromosome 6 to chromosome 16. *Genomics*, 7:222–34, 1990.

S. Wright. The general structure of populations. *Ann Eugenics*, 15:323–54, 1951.

A Chip Design

Subarray	Chromosome	start	end	nprobes
1	1	475	50709675	715628
2	1	50709740	98108291	715628
3	1	98108371	167887865	715628
4	1	166969764	214451478	715628
5	1	214451493	247199522	486747
5	2	321	13525448	211749
5	1_random	721	1663161	17130
6	2	13525508	61054388	715628
7	2	60130420	112298804	715628
8	2	112298874	159094964	715628
9	2	159094989	206387661	715628
10	2_random	21	185461	2517
10	3	35001	9635816	148082
10	2	205538945	242751081	565027
11	3	9635841	58089043	715628
12	3	58089103	109394892	715628
13	3	108551013	156060868	715628
14	4	191	3788423	55682
14	3	156060908	199446763	650573
14	3_random	16	749178	9371
15	4	3788463	54403787	715628
16	4	53433714	102075332	715628
17	4	102075392	149508703	715628
18	4	149508773	191262999	636740
18	4_random	51	842554	9953
18	5	64840	4266715	68933
19	5	3410784	53774183	715628
20	5	53774243	101836135	715628
21	5	101836220	149483558	715628
22	6	5001	14966339	228940
22	5_random	11	143239	1884
22	5	148564081	180837747	484802
23	6	14966429	66433733	715628
24	6	66433803	114882138	715628
25	6	113990973	161301866	715628
26	6	161301921	170896931	150611
26	6_random	36	1875467	28251
26	7	34011	35840154	536764

Table 2.10: **Layout of probe content for 42M discovery chip design**

Subarray	Chromosome	start	end	nprobes
27	7	35840209	89208956	715628
28	7	88309601	136525724	715628
29	8	46	24580495	376985
29	7_random	66	549184	7170
29	7	136525734	158821274	331471
30	8	24580565	76247638	715628
31	8	75353347	122880886	715628
32	8	122880926	146273053	361503
32	8_random	46	943728	6813
32	9	461	23010846	347310
33	9	23010931	90815313	715628
34	9	89838829	138192328	715628
35	9_random	377	1146355	9362
35	9	138192343	140273167	31023
35	10	50046	49461074	675241
36	10	49461129	97318956	715628
37	11	50031	7976157	119694
37	10_random	36	113198	1399
37	10	96254670	135374563	594533
38	11	7976222	59706518	715628
39	11	59706608	108056759	715628
40	11	106997474	134451799	422588
40	11_random	46	215199	1841
40	12	17346	19710657	291197
41	12	19710702	70436337	715628
42	12	70436377	118929957	715628
43	13	17918001	52120092	512456
43	12	117963955	132289468	203171
44	13	52120439	98008972	715628
45	13_random	6	186783	2830
45	13	98009052	114127916	245392
45	14	18070191	50012244	467404
46	14	48964632	97018727	715628
47	14	97018742	106360491	143517
47	15	18260026	57641897	572110
48	15_random	1	784271	8799
48	16	48	4845208	67243
48	15	57641972	100338484	639584
49	16	3749927	63858540	715628
50	16_random	51	105389	1137

Table 2.10: **Layout of probe content for 42M discovery chip design**

Subarray	Chromosome	start	end	nprobes
50	16	63858580	88822197	366179
50	17	16	24980012	348310
51	17	24980087	76245228	715628
52	18	633	43385437	641021
52	17_random	41	2617363	26067
52	17	75301540	78654689	48538
53	18	43385497	76117073	511827
53	19	11046	16954743	203727
53	18_random	16	4181	72
54	19_random	66	301747	3040
54	19	16955393	63806569	509071
54	20	8016	13395778	203515
55	21	9719783	13692255	11968
55	20	12525859	62435704	703659
56	22	14430011	27555339	185829
56	21	13692355	46944152	510708
56	21_random	1	1679594	19089
57	22_random	21	257239	3048
57	22	27555434	49591382	316279
57	M	1	16480	299
57	X	62	27551770	395999
58	X	26641154	82175954	715628
59	X	82175999	132113557	715628
60	X	132113732	154912731	337210
60	X_random	6	1719088	17967
60	Y	62	57771911	360443

Table 2.10: **Layout of probe content for 42M discovery chip design**

B Aneuploidies

	name	chr	median	q25	q75	norm
1	NA10842	1	0.90	0.74	1.26	norm4
2	NA06991	1	0.96	0.83	1.12	norm4
3	NA19139	1	1.04	0.78	1.35	norm4
4	NA10842	2	0.88	0.72	1.20	norm4
5	NA10843	2	-0.05	-0.16	0.06	norm1
6	NA12234	2	0.94	0.80	1.12	norm4
7	NA19222	2	0.09	-0.02	0.21	norm1
8	NA11832	2	0.95	0.79	1.17	norm4
9	NA12875	2	1.07	0.92	1.26	norm4
10	NA19139	2	1.05	0.79	1.38	norm4
11	NA19160	2	1.06	0.91	1.23	norm4
12	NA10842	3	0.87	0.73	1.18	norm4
13	NA06997	3	1.03	0.81	1.35	norm4
14	NA19139	3	1.09	0.85	1.41	norm4
15	NA10842	4	0.86	0.71	1.18	norm4
16	NA18540	4	0.13	0.00	0.26	norm1
17	NA19139	4	1.10	0.83	1.43	norm4
18	NA10842	5	0.91	0.72	1.35	norm4
19	NA10843	5	0.09	-0.01	0.19	norm1
20	NA18506	5	1.06	0.85	1.29	norm4
21	NA18608	5	1.06	0.84	1.35	norm4
22	NA18991	5	0.96	0.79	1.17	norm4
23	NA19139	5	1.06	0.79	1.39	norm4
24	NA10842	6	0.89	0.72	1.25	norm4
25	NA10843	6	-0.05	-0.16	0.06	norm1
26	NA12234	6	0.95	0.81	1.13	norm4
27	NA19178	6	-0.04	-0.19	0.10	norm1
28	NA18973	6	0.95	0.77	1.18	norm4
29	NA19139	6	1.07	0.81	1.41	norm4
30	NA10842	7	0.89	0.72	1.26	norm4
44	NA10847_2	9	0.05	-0.07	0.18	norm1
45	NA10847_3	9	0.04	-0.09	0.18	norm1
31	NA10843	7	-0.05	-0.16	0.06	norm1
32	NA18540	7	0.11	-0.02	0.24	norm1
33	NA07019	7	0.94	0.78	1.15	norm4
34	NA18506	7	1.06	0.83	1.33	norm4
35	NA19139	7	1.04	0.77	1.38	norm4
36	NA10842	8	0.90	0.73	1.26	norm4

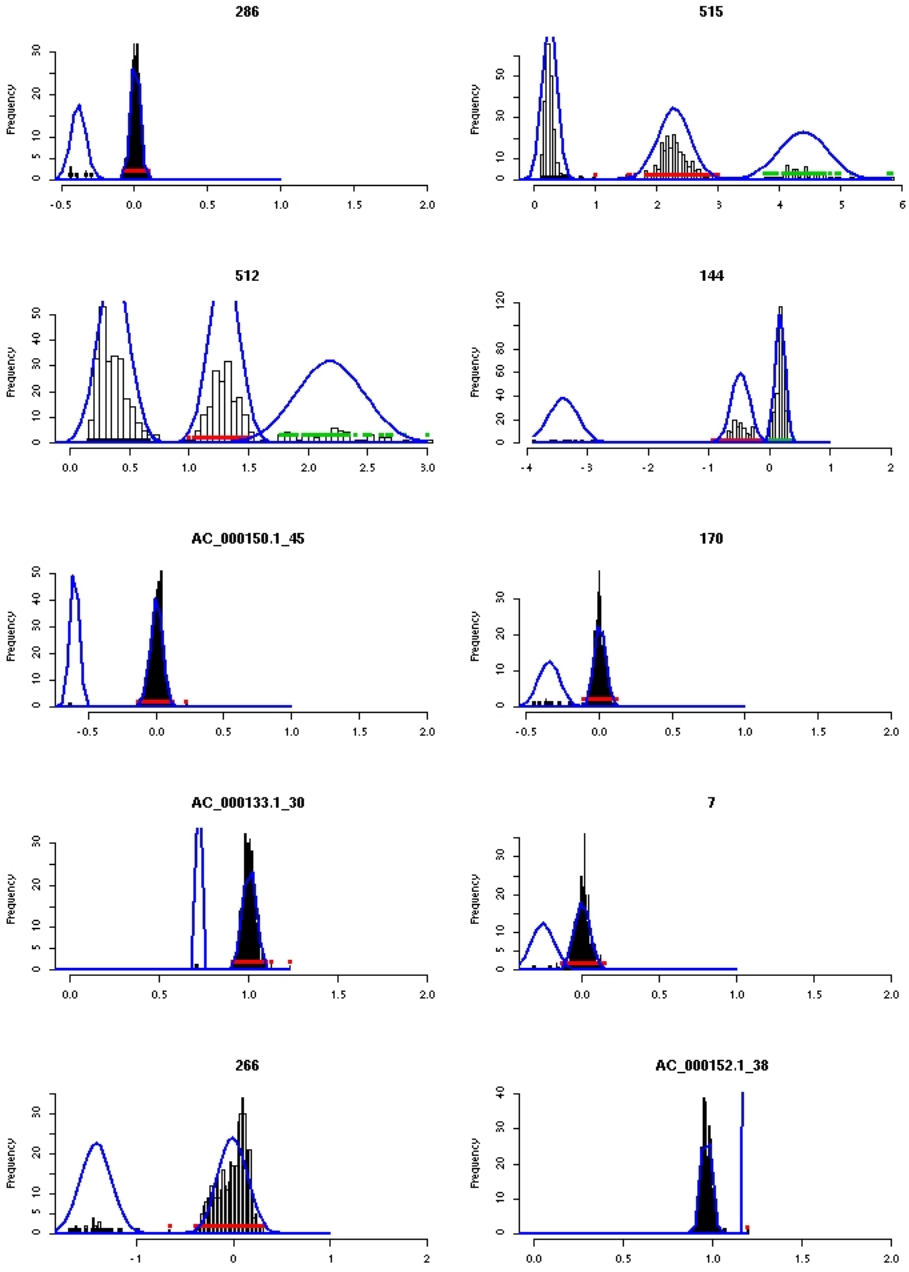
37	NA10843	8	-0.05	-0.15	0.06	norm1
38	NA12145	8	-0.04	-0.15	0.08	norm1
39	NA12234	8	0.95	0.81	1.14	norm4
40	NA19139	8	1.07	0.82	1.38	norm4
41	NA19160	8	1.07	0.90	1.23	norm4
42	NA10842	9	0.93	0.76	1.29	norm4
43	NA10843	9	0.09	-0.01	0.19	norm1
46	NA12248	9	0.06	-0.05	0.18	norm1
47	NA12249	9	-0.04	-0.17	0.07	norm1
48	NA12341	9	-0.04	-0.17	0.08	norm1
49	NA12814	9	0.08	-0.05	0.20	norm1
50	NA12878_1	9	0.03	-0.09	0.19	norm1
54	NA18540	9	0.12	-0.01	0.26	norm1
55	NA18563	9	0.03	-0.10	0.18	norm1
56	NA18940	9	-0.04	-0.17	0.08	norm1
57	NA18953	9	-0.05	-0.17	0.08	norm1
58	NA18999	9	-0.04	-0.17	0.08	norm1
59	NA19000	9	0.03	-0.11	0.19	norm1
60	NA19005	9	0.03	-0.09	0.18	norm1
61	NA19208	9	0.13	-0.00	0.25	norm1
62	NA19238	9	-0.04	-0.17	0.08	norm1
63	NA19249	9	0.03	-0.10	0.18	norm1
64	NA18857	9	1.07	0.89	1.26	norm4
65	NA19139	9	1.06	0.81	1.37	norm4
66	NA10842	10	0.93	0.75	1.37	norm4
67	NA10843	10	0.09	-0.01	0.18	norm1
68	NA18540	10	0.96	0.81	1.14	norm4
69	NA06997	10	1.01	0.78	1.32	norm4
70	NA18506	10	1.04	0.83	1.28	norm4
71	NA19139	10	1.04	0.79	1.35	norm4
72	NA10842	11	0.91	0.74	1.28	norm4
73	NA10843	11	-0.04	-0.14	0.08	norm1
74	NA12234	11	0.94	0.81	1.12	norm4
75	NA12274	11	0.11	-0.03	0.26	norm1
76	NA18924	11	0.08	-0.03	0.19	norm1
77	NA19222	11	0.09	-0.03	0.21	norm1
78	NA19139	11	1.07	0.81	1.40	norm4
79	NA07435	12	0.12	0.02	0.22	norm1
80	NA10842	12	0.02	-0.11	0.15	norm1
81	NA10843	12	0.10	0.01	0.19	norm1
82	NA11992	12	0.04	-0.07	0.16	norm1
83	NA12057	12	0.04	-0.07	0.14	norm1

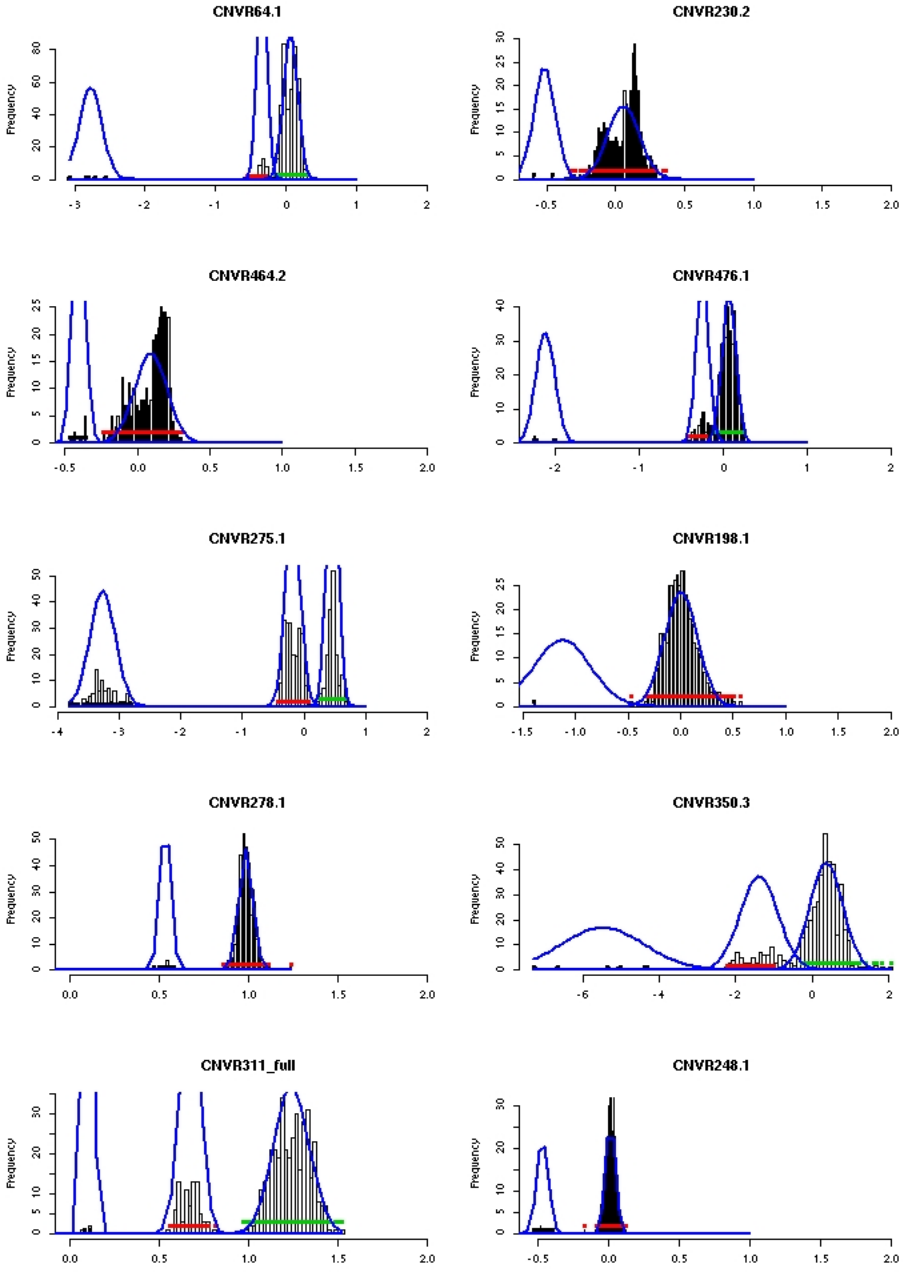
84	NA12248	12	0.07	-0.04	0.19	norm1
85	NA12274	12	-0.02	-0.16	0.09	norm1
86	NA12348	12	0.04	-0.07	0.15	norm1
87	NA12739	12	0.35	0.24	0.47	norm1
88	NA18540	12	-0.03	-0.15	0.08	norm1
89	NA19193	12	0.17	0.04	0.32	norm1
90	NA18522	12	1.05	0.86	1.26	norm4
91	NA19139	12	1.02	0.77	1.34	norm4
92	NA10842	13	0.90	0.73	1.28	norm4
93	NA10843	13	-0.04	-0.15	0.07	norm1
94	NA11918	13	-0.06	-0.17	0.06	norm1
95	NA11992	13	0.07	-0.05	0.19	norm1
96	NA06997	13	1.05	0.82	1.39	norm4
97	NA19139	13	1.11	0.85	1.43	norm4
98	NA06984	14	0.06	-0.10	0.26	norm1
99	NA07051	14	0.03	-0.11	0.18	norm1
100	NA10842	14	0.92	0.73	1.42	norm4
101	NA12154	14	0.04	-0.09	0.18	norm1
102	NA12248	14	0.07	-0.06	0.19	norm1
103	NA12341	14	0.04	-0.10	0.22	norm1
104	NA12829	14	0.04	-0.09	0.18	norm1
105	NA18540	14	0.07	-0.04	0.20	norm1
106	NA18990	14	0.04	-0.08	0.19	norm1
107	NA19098	14	1.06	0.88	1.26	norm4
108	NA10843	15	0.09	-0.03	0.20	norm1
109	NA11891	15	0.22	0.08	0.39	norm1
110	NA10842	16	1.02	0.79	1.52	norm4
111	NA10851	16	1.07	0.87	1.28	norm4
112	NA11832	16	1.02	0.83	1.27	norm4
113	NA12005	16	1.05	0.85	1.29	norm4
114	NA18571	16	1.05	0.85	1.28	norm4
115	NA18973	16	1.05	0.86	1.30	norm4
116	NA10842	17	1.04	0.79	1.61	norm4
117	NA10851	17	1.09	0.89	1.30	norm4
118	NA12005	17	1.06	0.86	1.31	norm4
119	NA18571	17	1.07	0.85	1.33	norm4
120	NA18948	17	1.08	0.90	1.30	norm4
121	NA18973	17	1.07	0.86	1.34	norm4
122	NA18991	17	1.07	0.87	1.31	norm4
123	NA06984	18	-0.07	-0.23	0.08	norm1
124	NA10842	18	0.90	0.74	1.32	norm4
125	NA10843	18	-0.04	-0.14	0.08	norm1

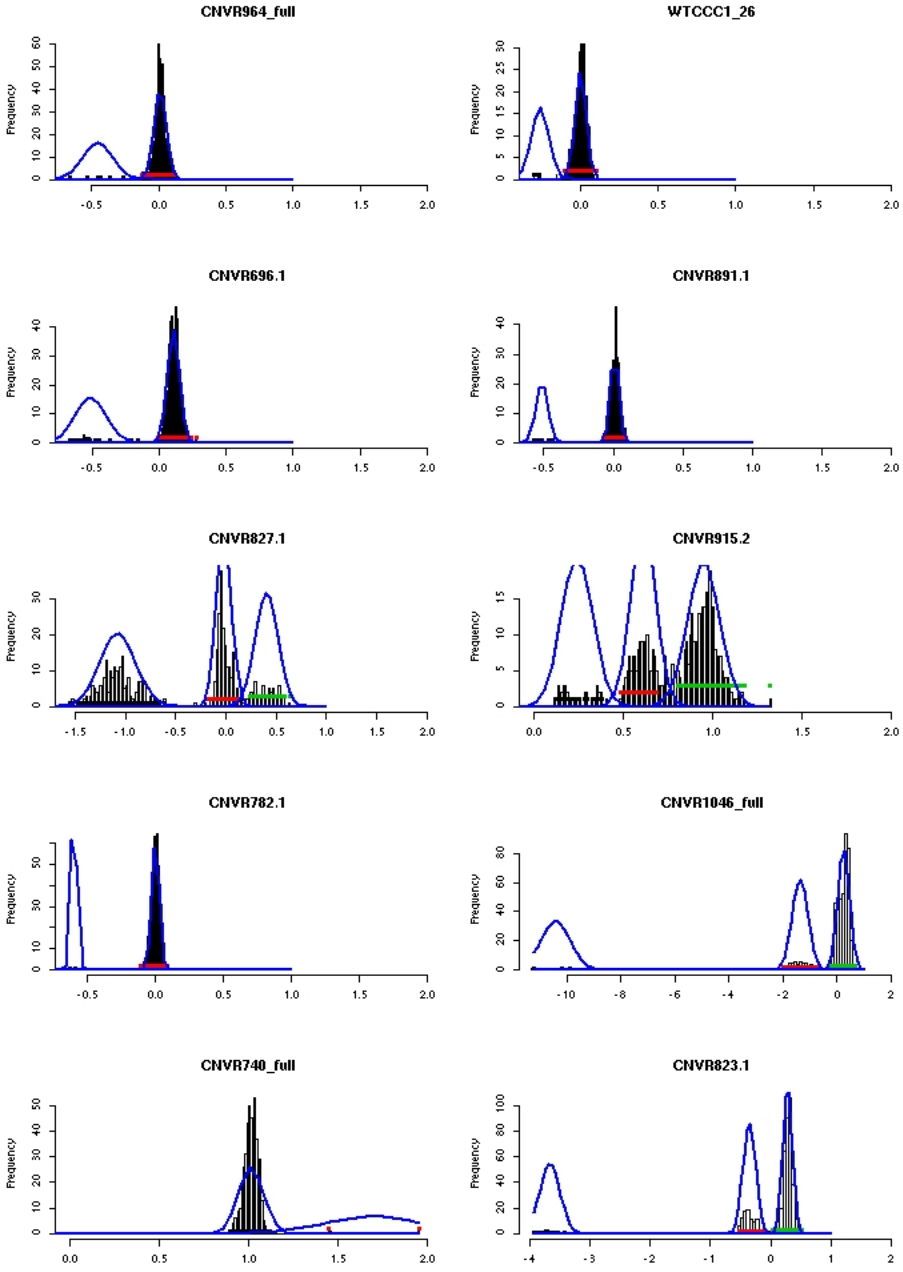
126	NA10856	18	0.06	-0.07	0.20	norm1
127	NA11918	18	-0.04	-0.15	0.08	norm1
128	NA18972	18	-0.04	-0.15	0.09	norm1
129	NA06997	18	1.05	0.80	1.39	norm4
130	NA18973	18	0.95	0.79	1.20	norm4
131	NA10842	19	1.07	0.80	1.61	norm4
132	NA12057	19	1.07	0.91	1.24	norm4
133	NA12145	19	0.09	-0.03	0.21	norm1
134	NA12234	19	1.07	0.89	1.29	norm4
135	NA07019	19	1.06	0.87	1.31	norm4
136	NA10835	19	1.11	0.91	1.29	norm4
137	NA10851	19	1.11	0.89	1.34	norm4
138	NA10859	19	1.05	0.87	1.28	norm4
139	NA11832	19	1.10	0.88	1.34	norm4
140	NA11882	19	1.09	0.92	1.24	norm4
141	NA12005	19	1.08	0.86	1.36	norm4
142	NA12342	19	1.05	0.88	1.25	norm4
143	NA18537_1	19	1.06	0.90	1.25	norm4
144	NA18571	19	1.08	0.86	1.34	norm4
145	NA18948	19	1.09	0.91	1.32	norm4
146	NA18960	19	1.08	0.89	1.30	norm4
147	NA18973	19	1.09	0.88	1.39	norm4
148	NA18991	19	1.08	0.88	1.31	norm4
149	NA10842	20	0.99	0.78	1.54	norm4
150	NA19139	20	0.99	0.76	1.30	norm4
151	NA10842	21	0.88	0.71	1.29	norm4
152	NA10843	21	-0.06	-0.17	0.06	norm1
153	NA12234	21	0.01	-0.12	0.18	norm1
154	NA18540	21	0.08	-0.05	0.20	norm1
155	NA06997	21	1.01	0.79	1.35	norm4
156	NA19103	21	1.05	0.84	1.28	norm4
157	NA10842	22	1.01	0.76	1.51	norm4
158	NA11832	22	1.13	0.87	1.40	norm4
159	NA19238	23	1.56	1.11	1.87	norm4
161	NA10861	23	1.58	1.21	1.81	norm4
162	NA10863	23	1.58	1.13	1.79	norm4
163	NA12156	23	1.59	1.10	1.82	norm4
164	NA12239	23	1.57	1.15	1.77	norm4
165	NA18576	23	1.56	1.14	1.85	norm4

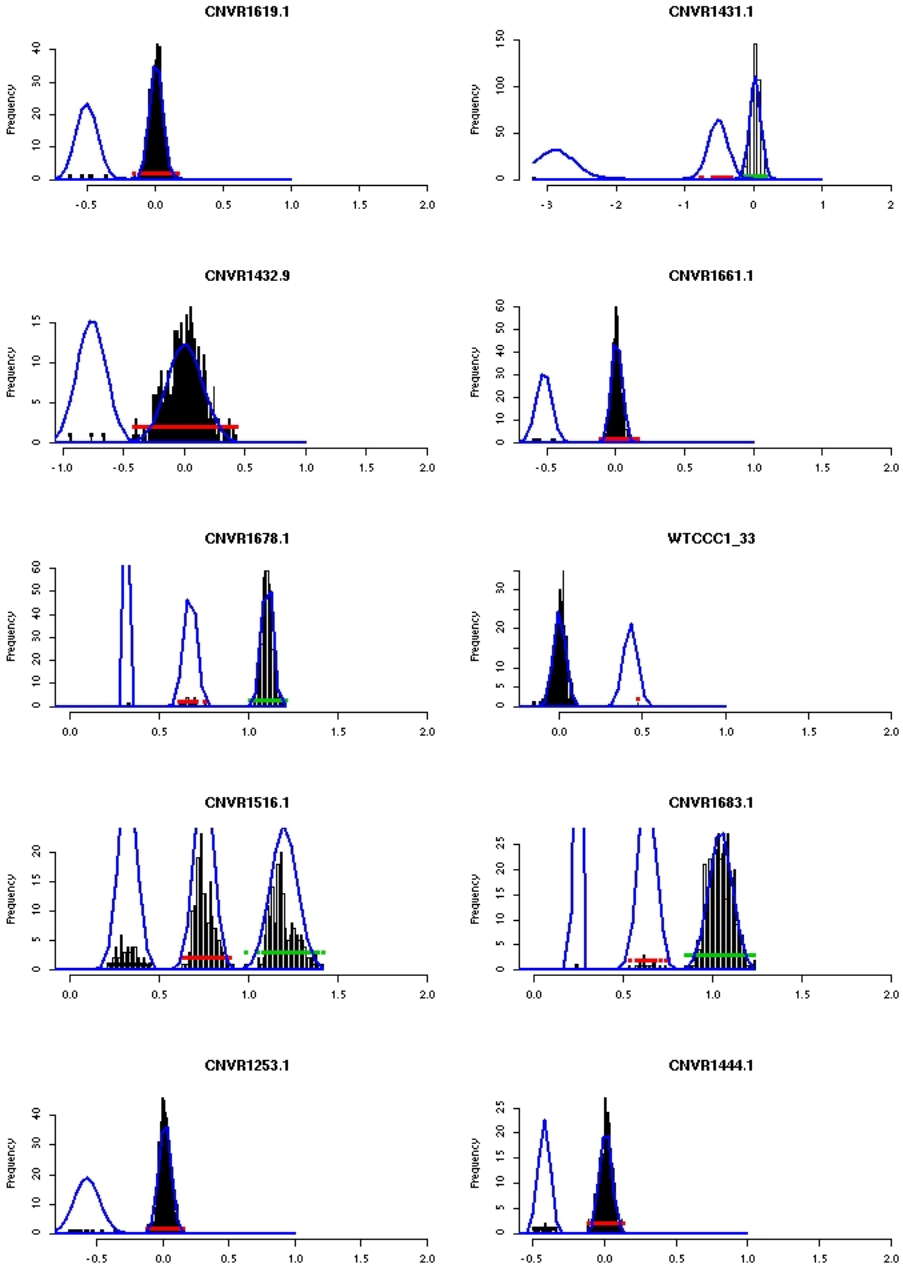
C Genotype Cluster Plots

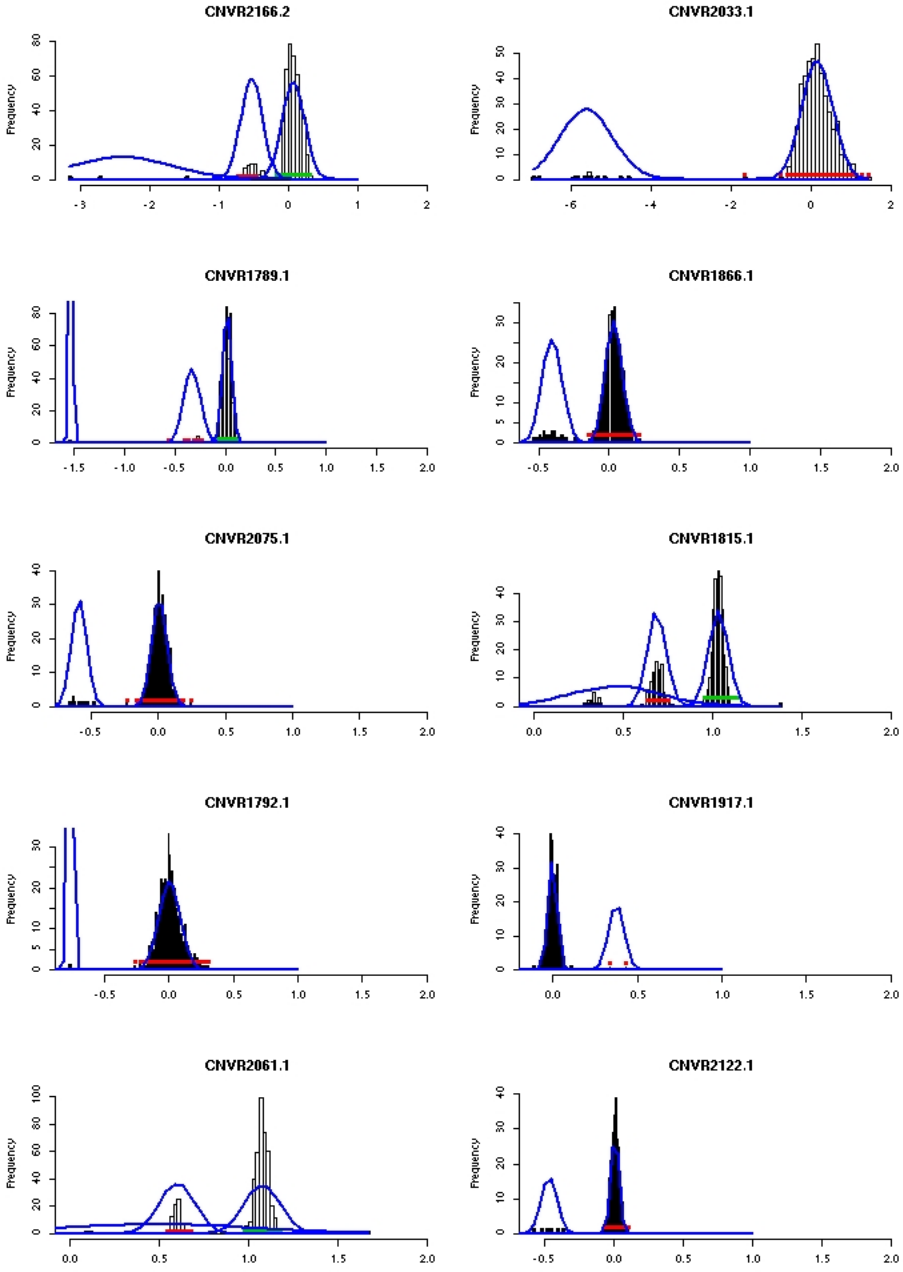
On the following pages, we present cluster plots for the final genotyping models used in our analyses. For each chromosome we have selected at random 10 CNVs to plot. Each plot is labelled with its CNV name. The data shown within each plot is the normalized, un-logged Cy5/Cy3 intensity ratios (histogram), the mixture model fit (blue lines), and the genotype calls (colored dots: black, red, green, etc.). The first set of 10 plots come from the novel insert sequences.

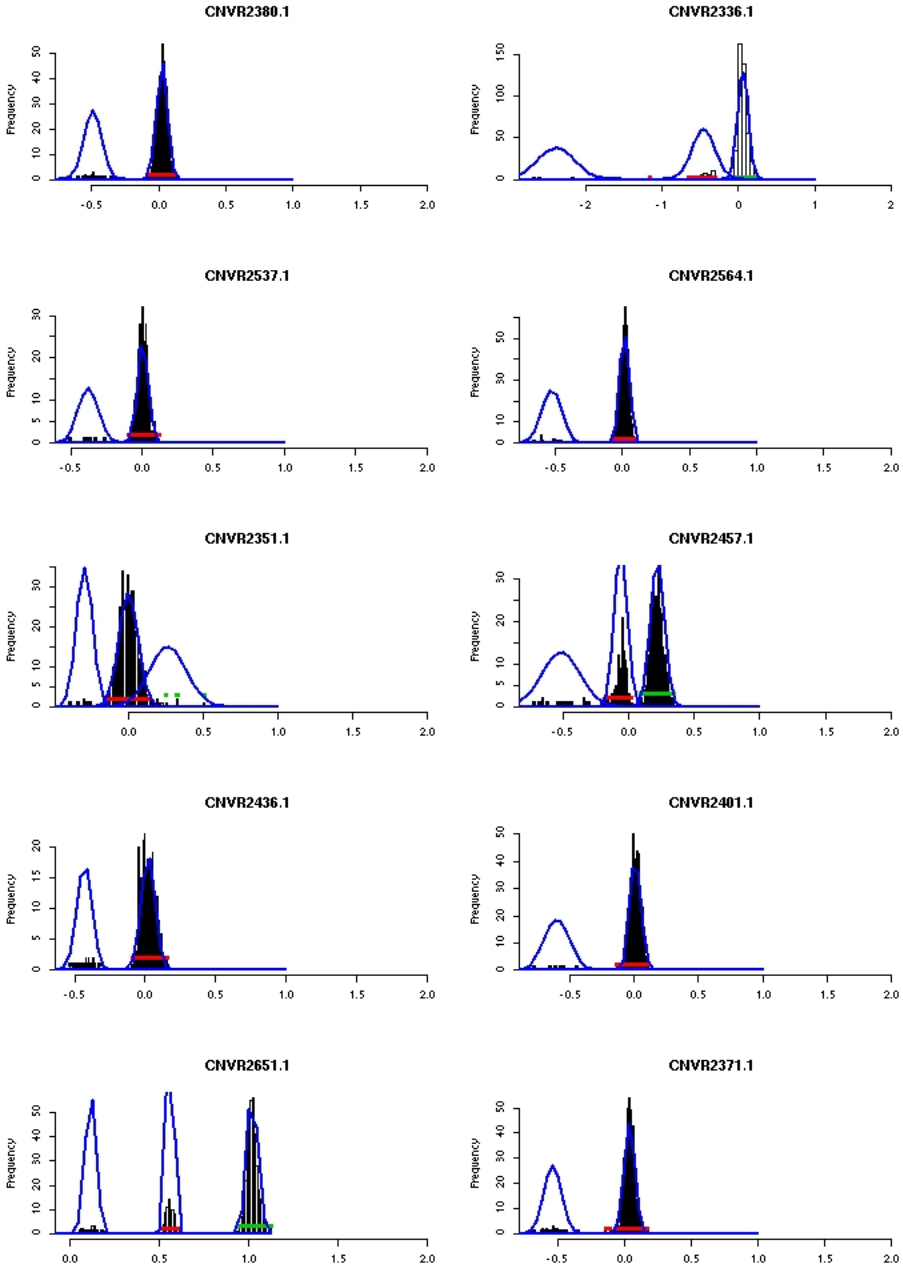

286
Frequency
515
Frequency
512
Frequency
144
Frequency
AC_000150.1_45
Frequency
170
Frequency
AC_000133.1_30
Frequency
7
Frequency
266
Frequency
AC_000152.1_38
Frequency

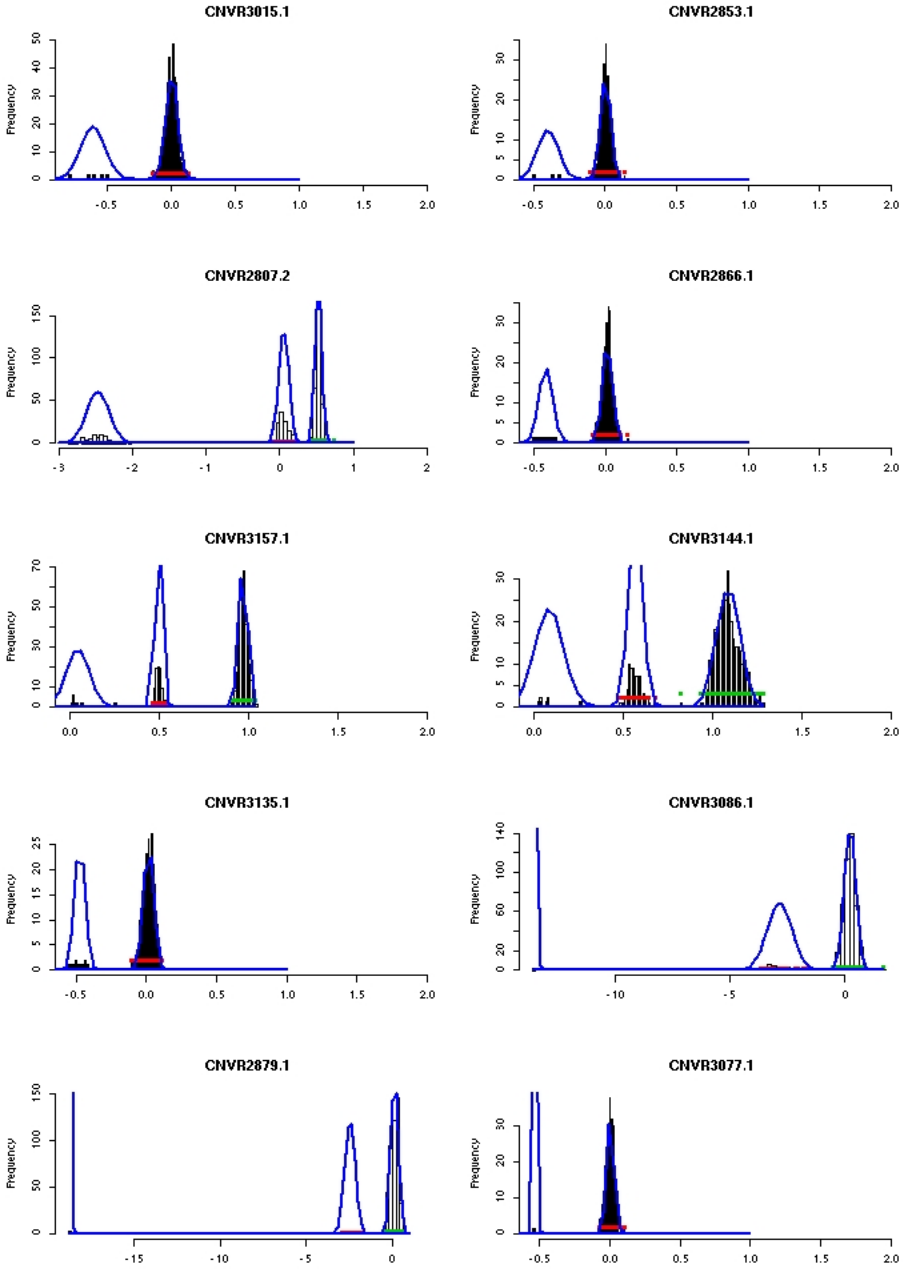

CNVR64.1
CNVR230.2
CNVR464.2
CNVR476.1
CNVR275.1
CNVR198.1
CNVR278.1
CNVR350.3
CNVR311_full
CNVR248.1
Frequency

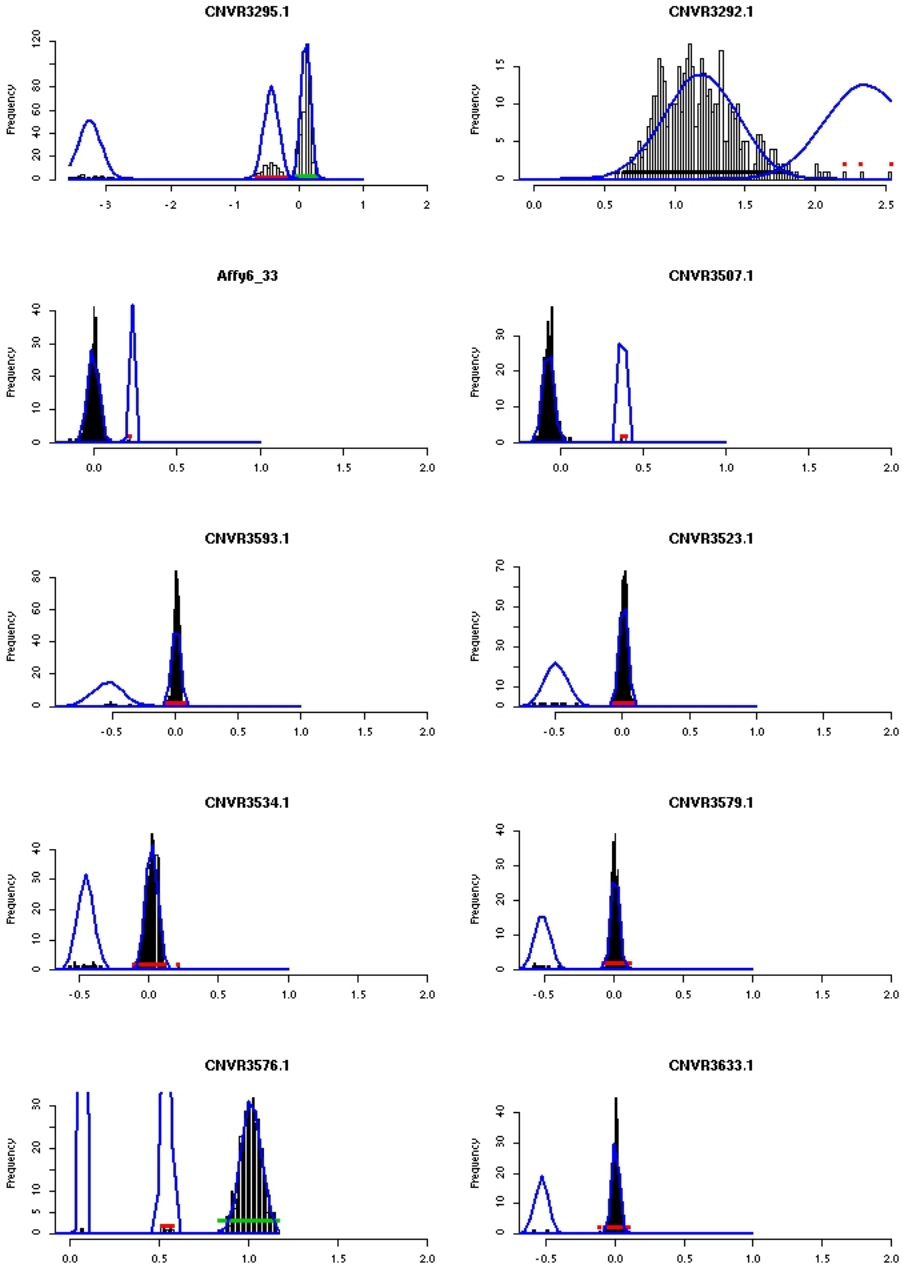

CNVR964_full
WTCCC1_26
CNVR696.1
CNVR891.1
CNVR827.1
CNVR915.2
CNVR782.1
CNVR1046_full
CNVR740_full
CNVR823.1
Frequency

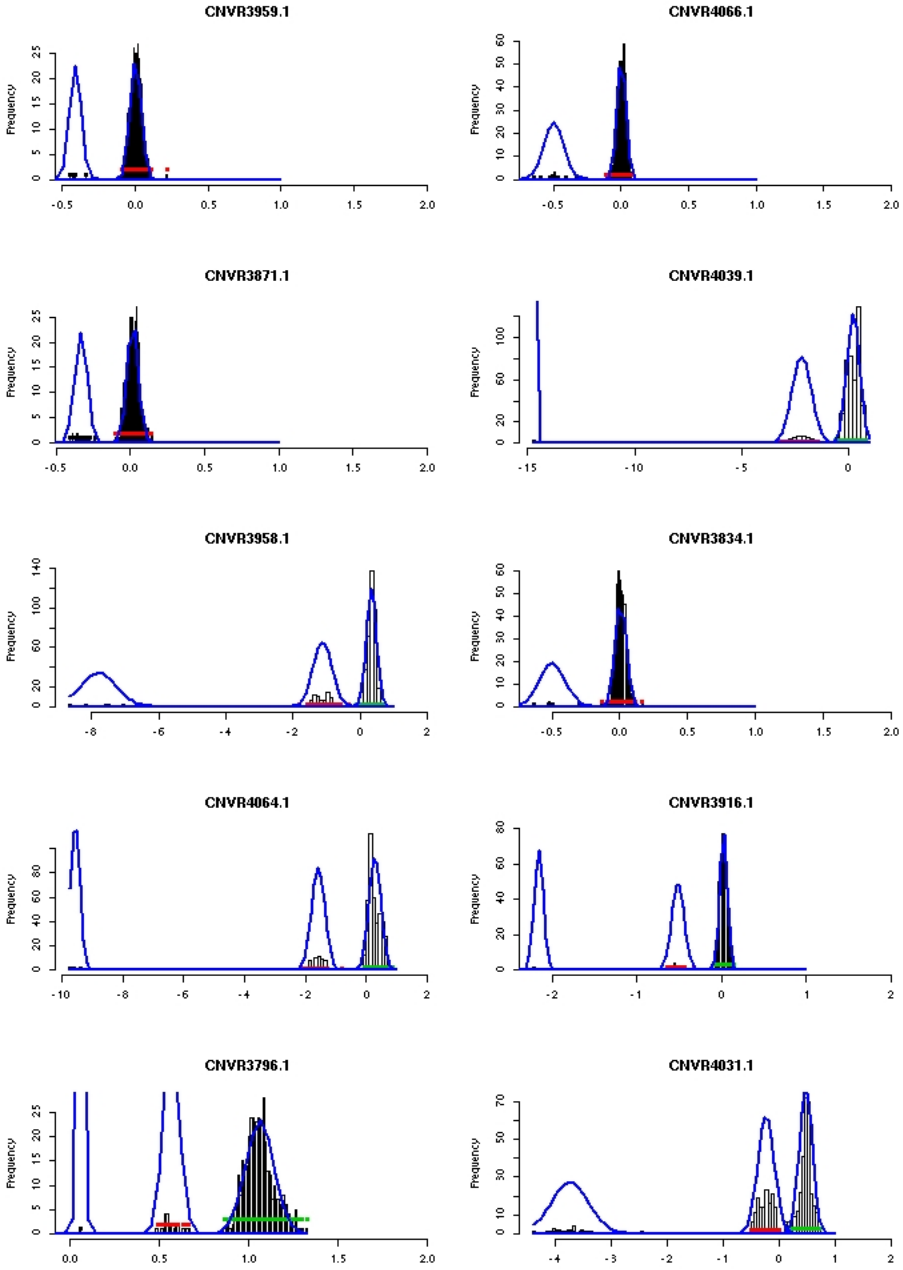

CNVR1619.1
CNVR1431.1
CNVR1432.9
CNVR1661.1
CNVR1678.1
WTCCC1_33
CNVR1516.1
CNVR1683.1
CNVR1253.1
CNVR1444.1
Frequency

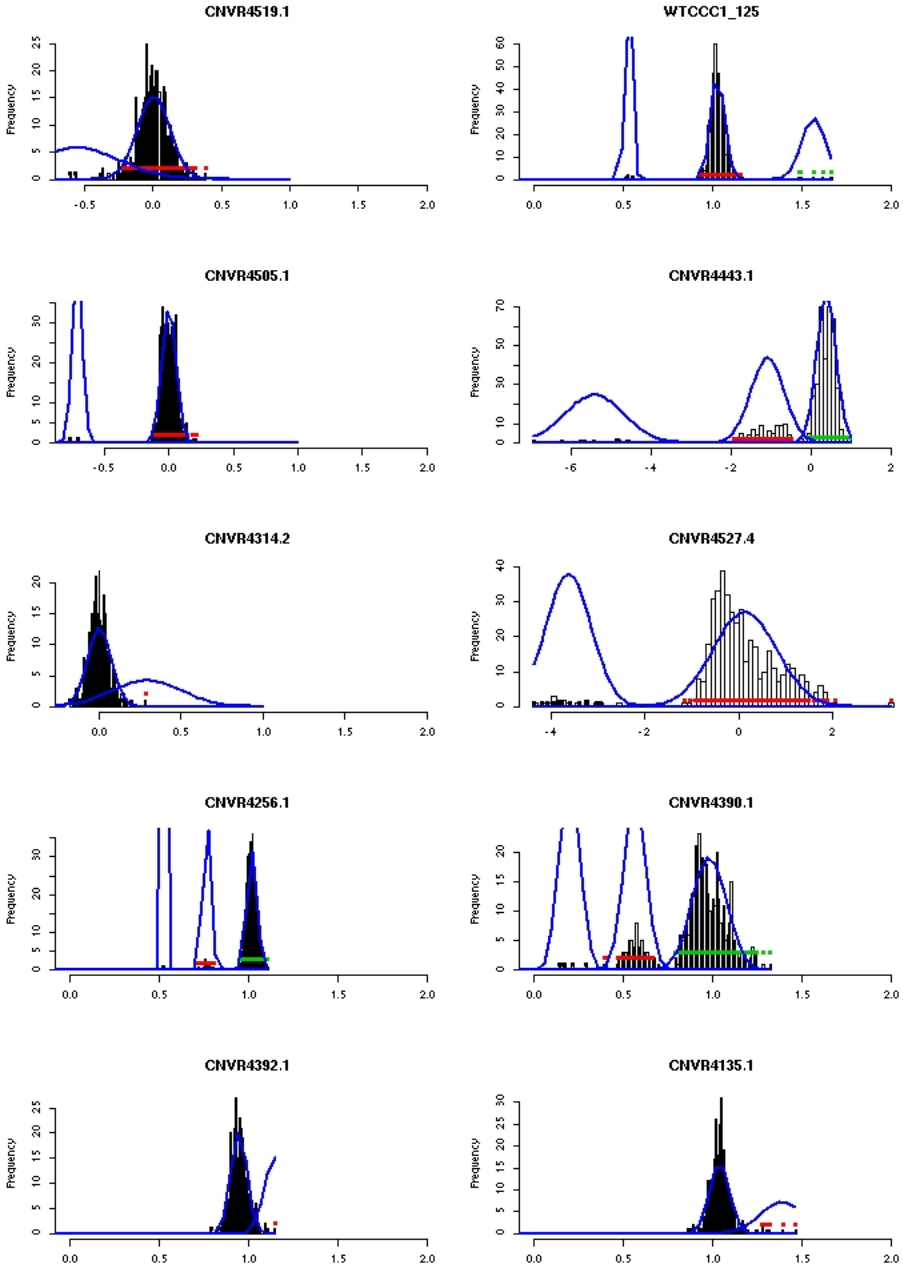

CNVR2166.2
CNVR2033.1
CNVR1789.1
CNVR1866.1
CNVR2075.1
CNVR1815.1
CNVR1792.1
CNVR1917.1
CNVR2061.1
CNVR2122.1
Frequency

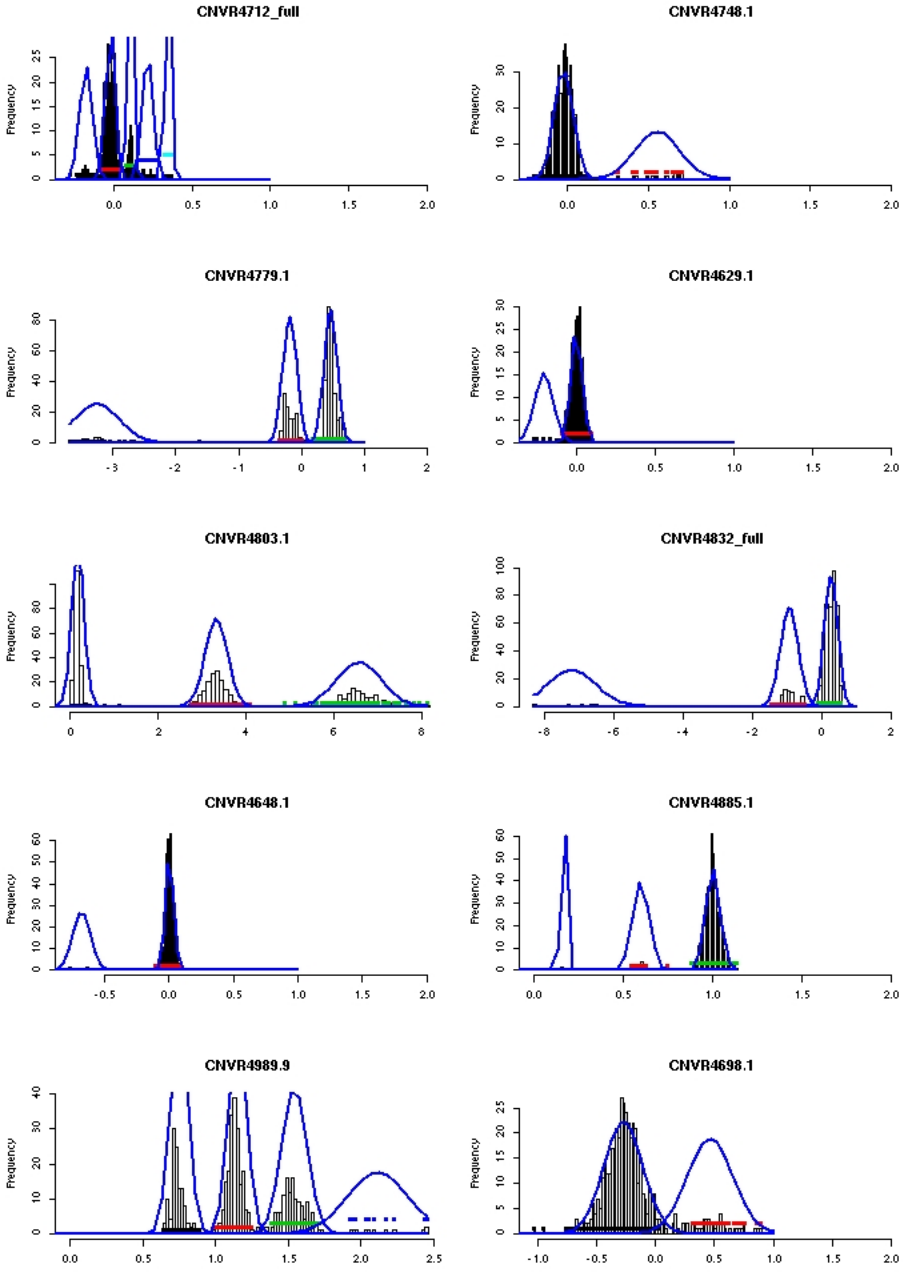

CNVR2380.1
CNVR2336.1
CNVR2537.1
CNVR2564.1
CNVR2351.1
CNVR2457.1
CNVR2436.1
CNVR2401.1
CNVR2651.1
CNVR2371.1
Frequency

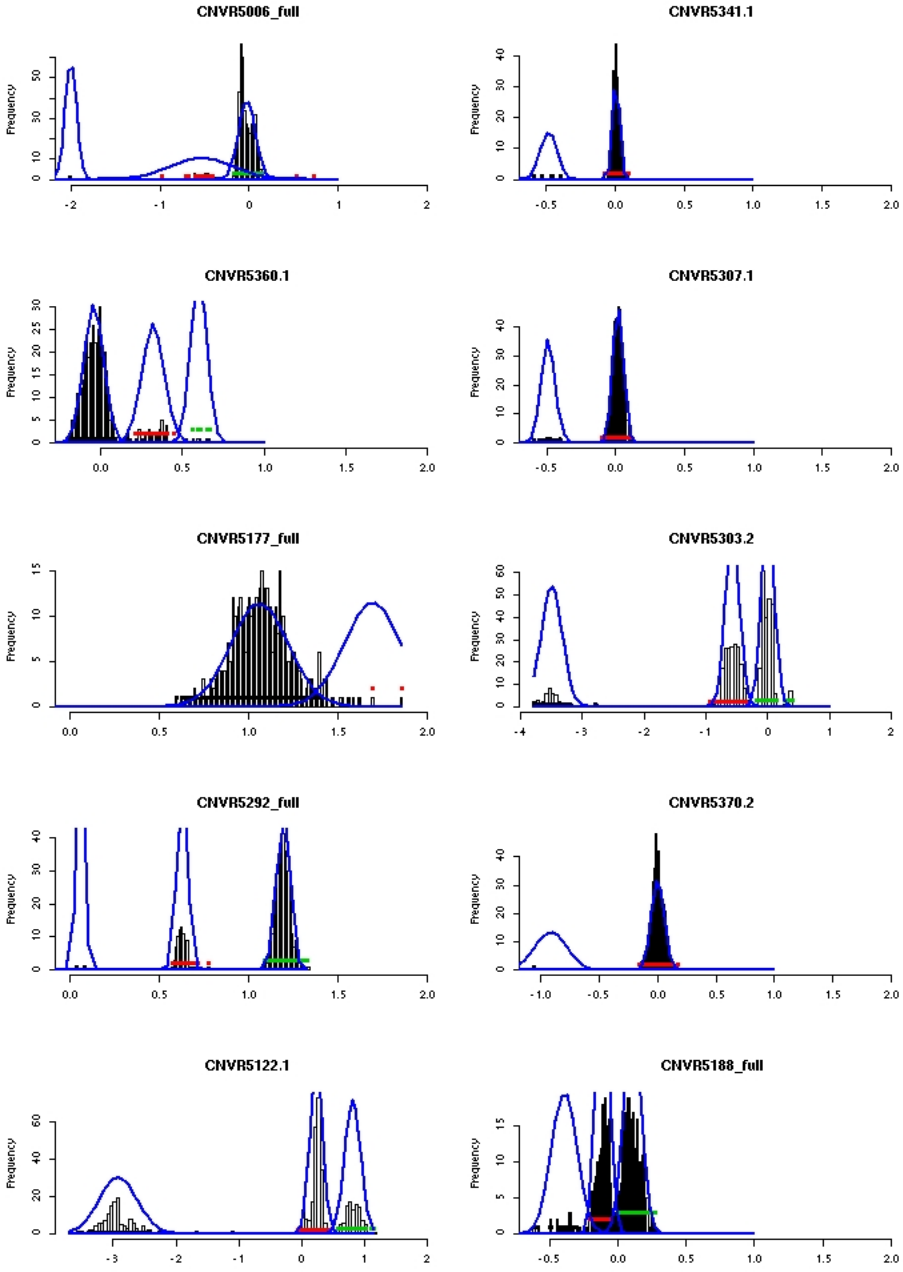

CNVR3015.1
CNVR2853.1
CNVR2807.2
CNVR2866.1
CNVR3157.1
CNVR3144.1
CNVR3135.1
CNVR3086.1
CNVR2879.1
CNVR3077.1
Frequency

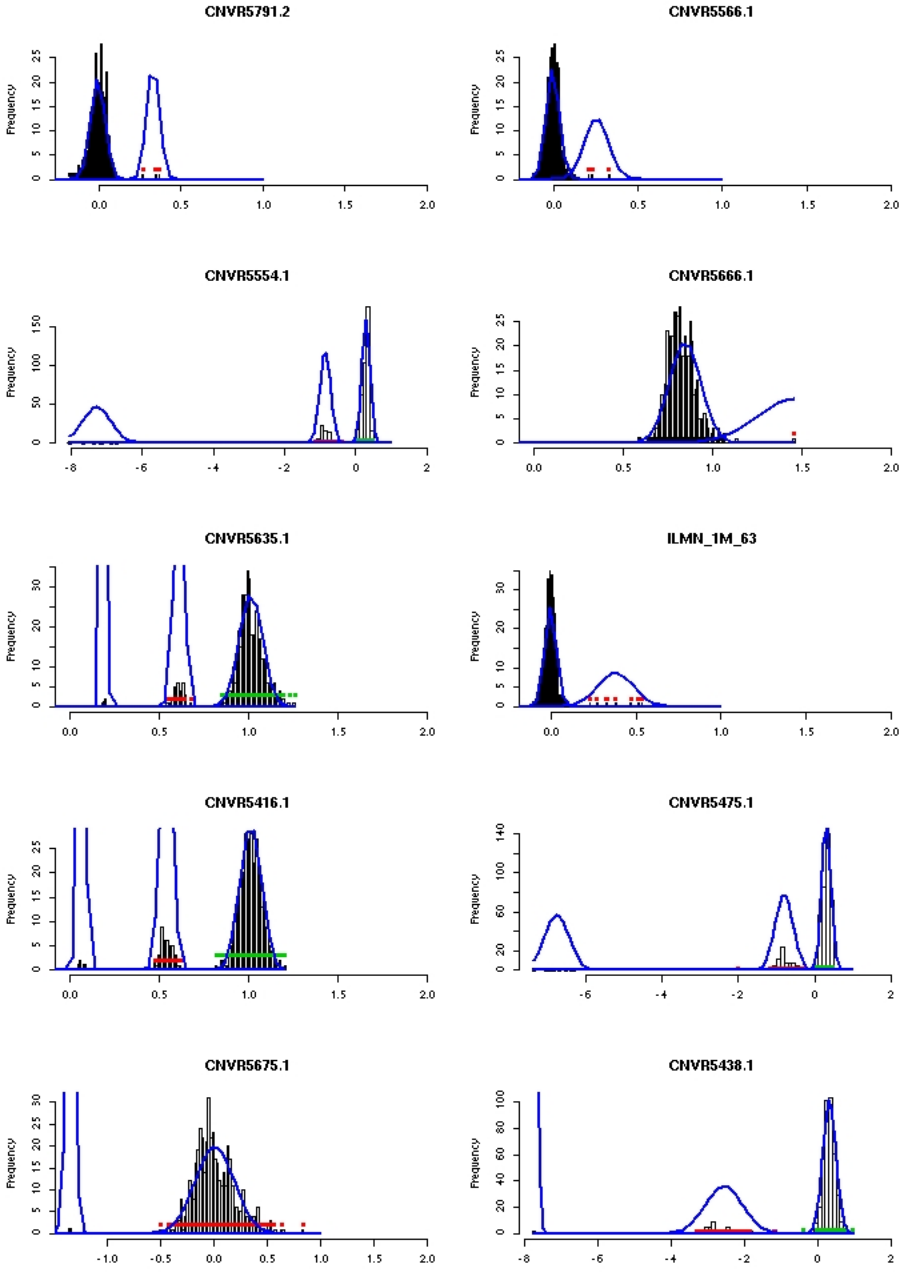

CNVR3295.1
CNVR3292.1
Affy6_33
CNVR3507.1
CNVR3593.1
CNVR3523.1
CNVR3534.1
CNVR3579.1
CNVR3576.1
CNVR3633.1
Frequency

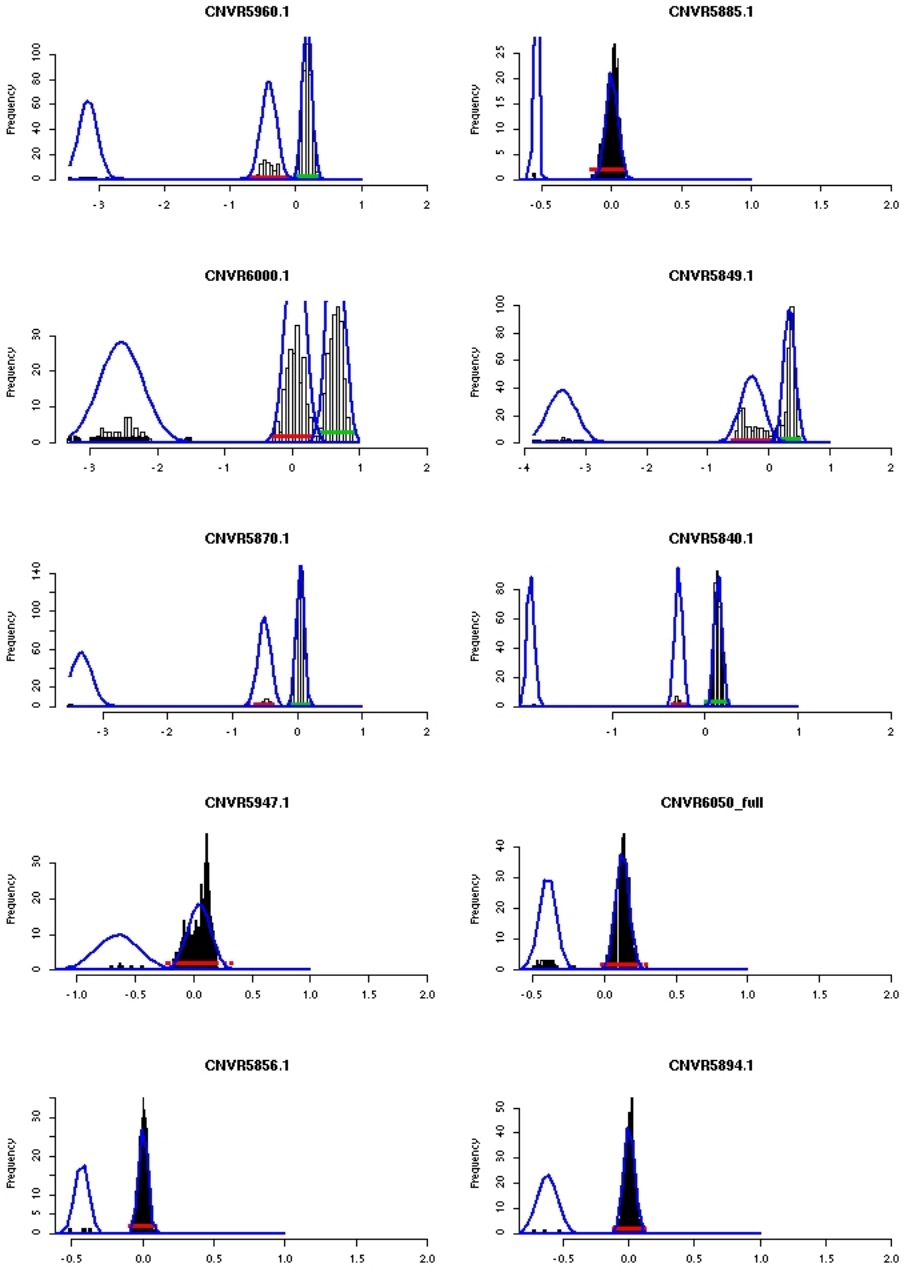

CNVR3959.1
CNVR4066.1
CNVR3871.1
CNVR4039.1
CNVR3958.1
CNVR3834.1
CNVR4064.1
CNVR3916.1
CNVR3796.1
CNVR4031.1
Frequency

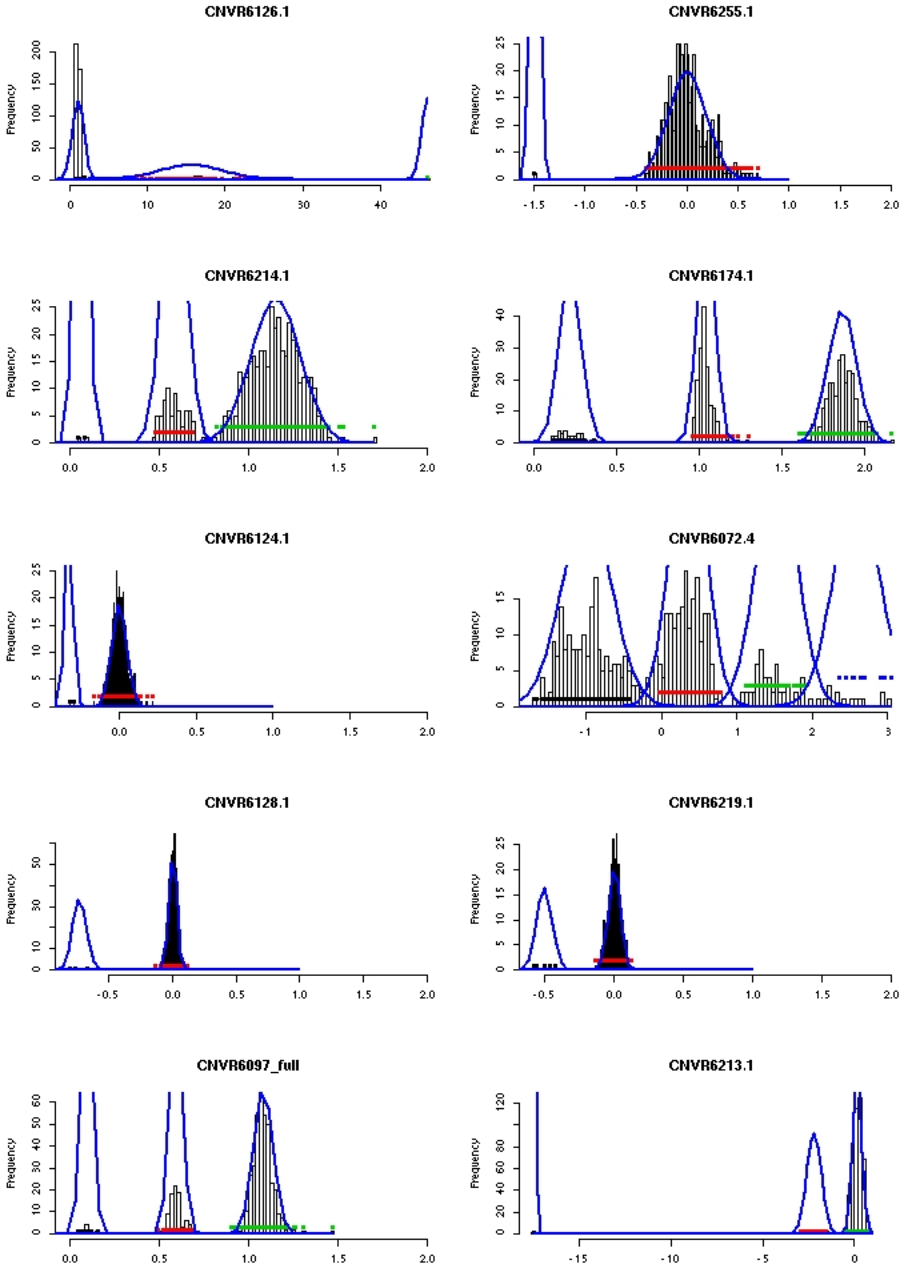

CNVR4519.1
WTCCC1_125
CNVR4505.1
CNVR4443.1
CNVR4314.2
CNVR4527.4
CNVR4256.1
CNVR4390.1
CNVR4392.1
CNVR4135.1
Frequency

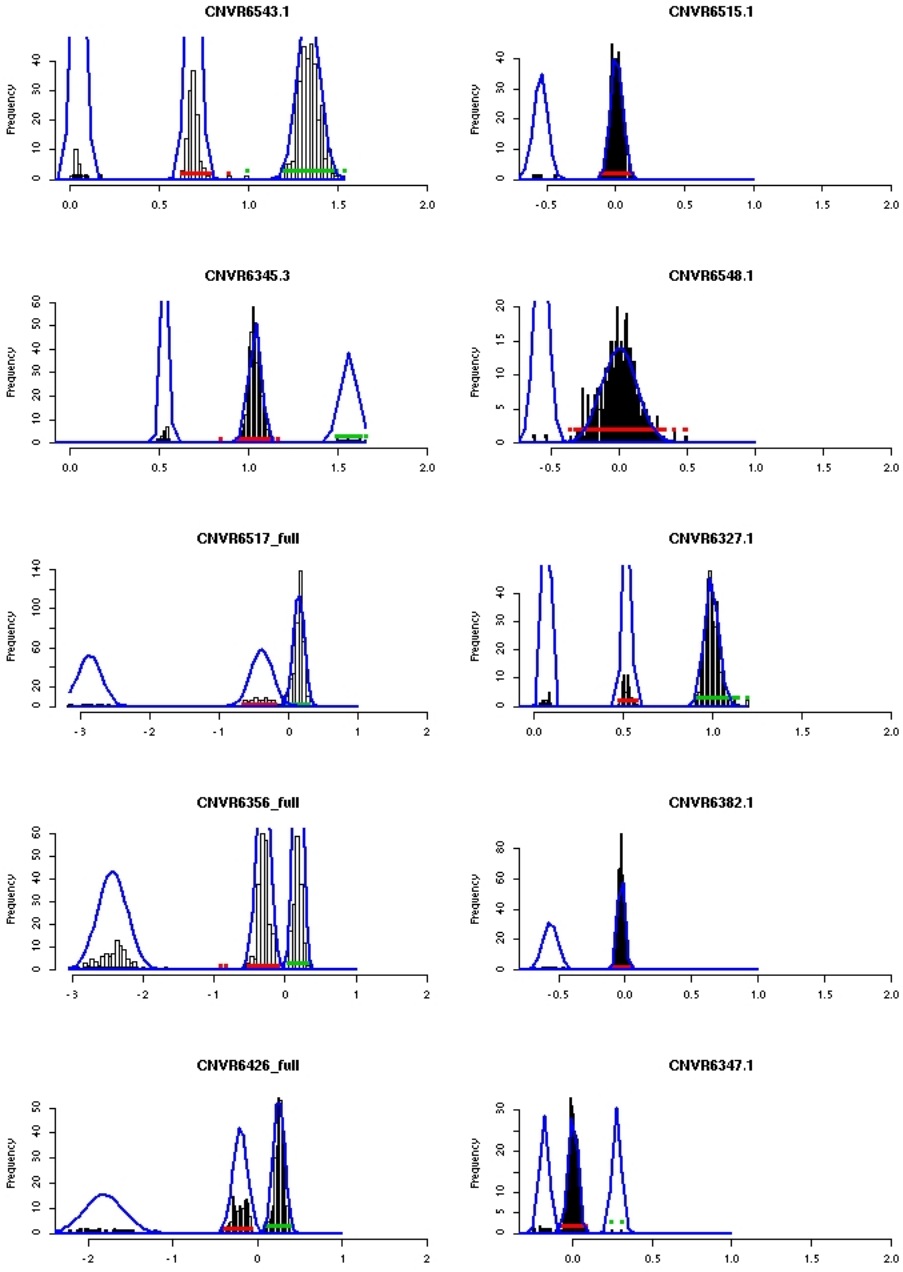

CNVR4712_full
CNVR4748.1
CNVR4779.1
CNVR4629.1
CNVR4803.1
CNVR4832_full
CNVR4648.1
CNVR4885.1
CNVR4989.9
CNVR4698.1
Frequency

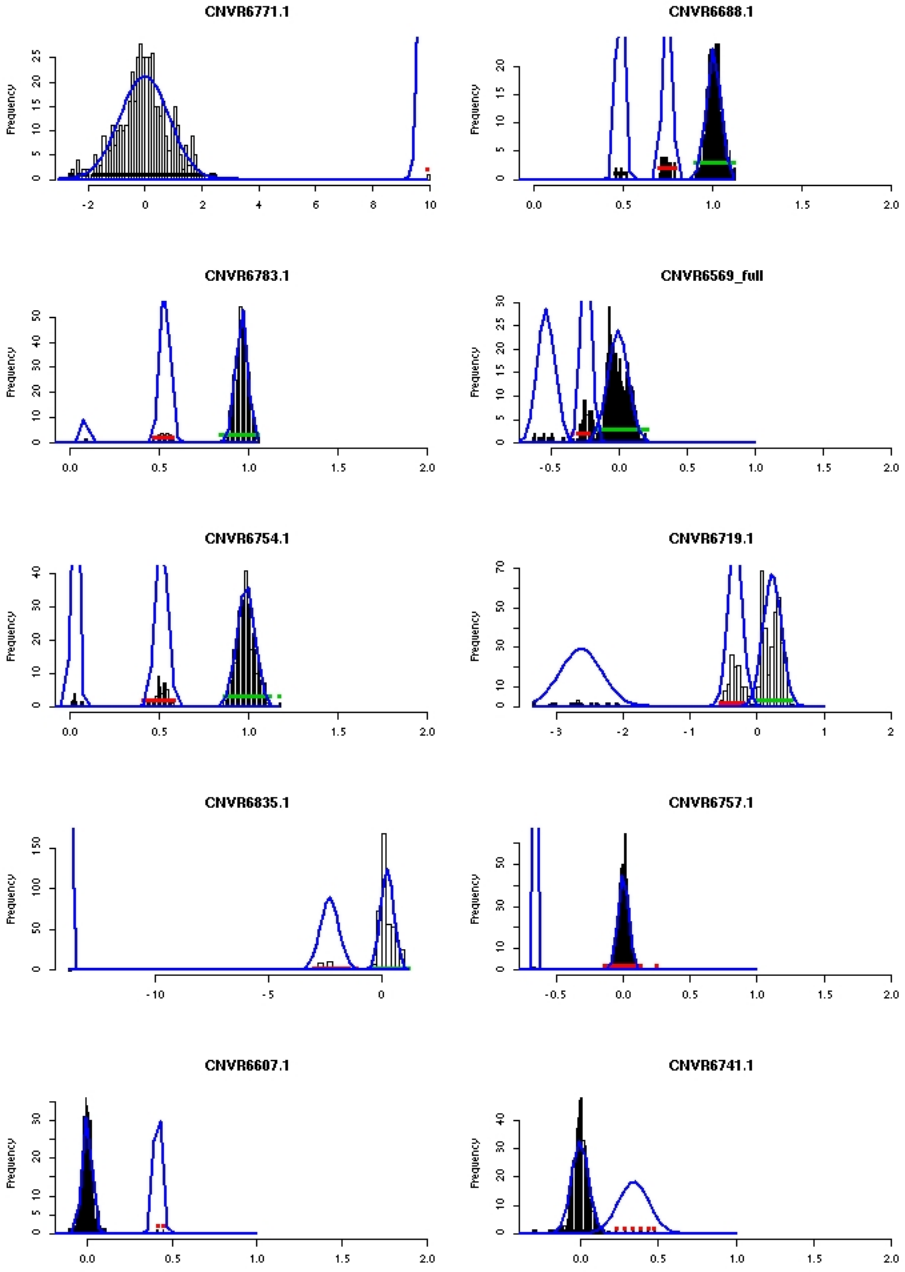

CNVR5006_full
CNVR5341.1
CNVR5360.1
CNVR5307.1
CNVR5177_full
CNVR5303.2
CNVR5292_full
CNVR5370.2
CNVR5122.1
CNVR5188_full
Frequency

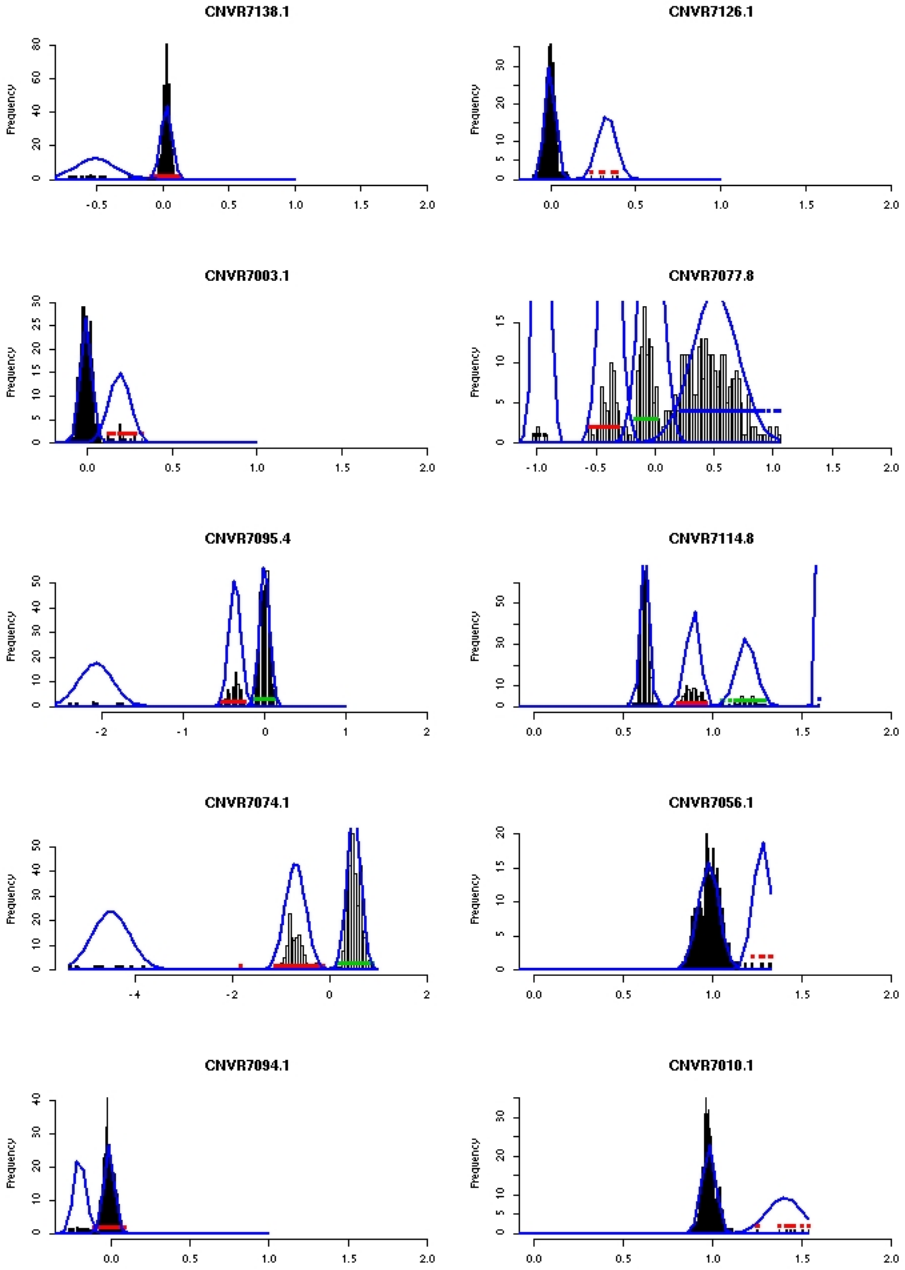

CNVR5791.2
CNVR5566.1
CNVR5554.1
CNVR5666.1
CNVR5635.1
ILMN_1M_63
CNVR5416.1
CNVR5475.1
CNVR5675.1
CNVR5438.1
Frequency

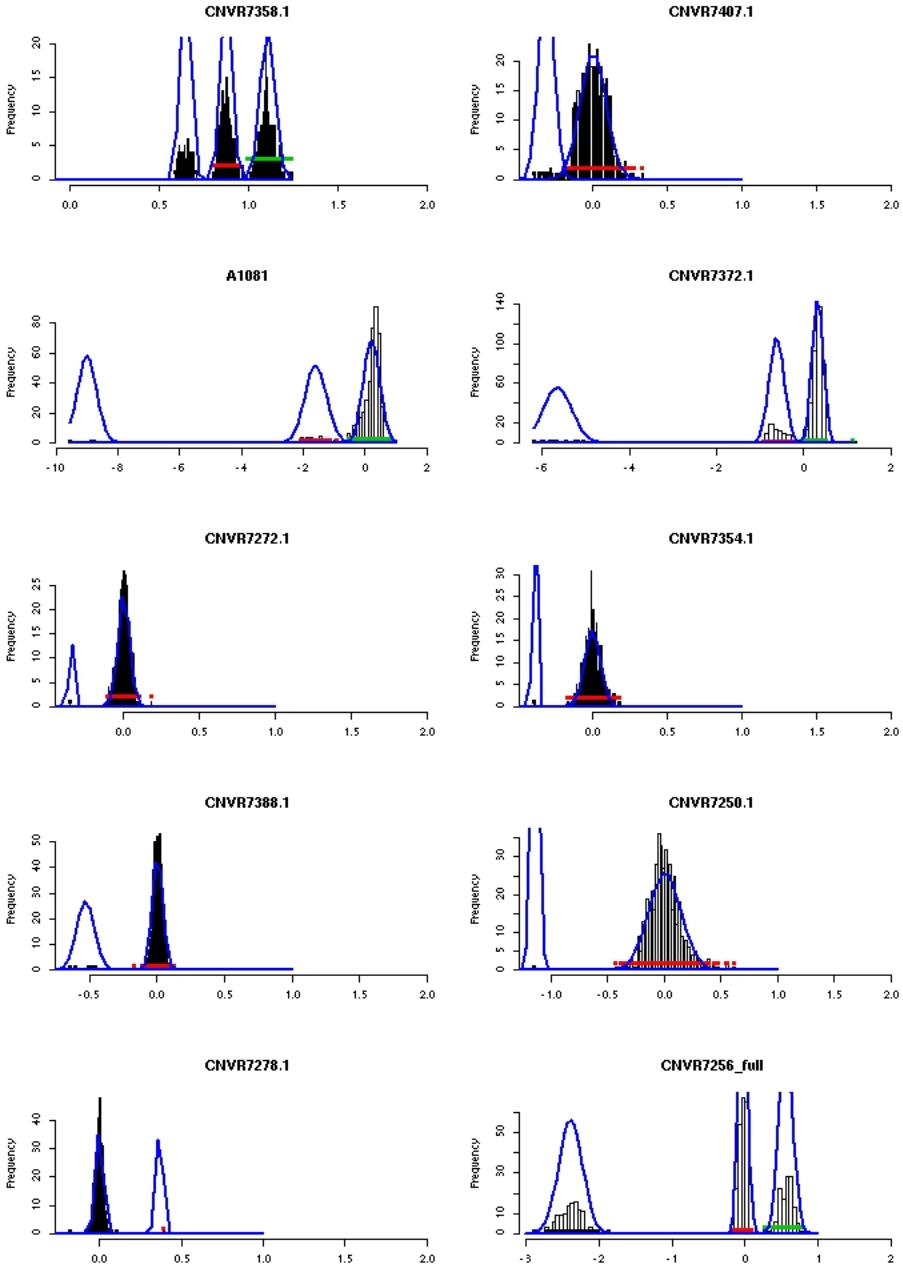

CNVR5960.1
CNVR5885.1
CNVR6000.1
CNVR5849.1
CNVR5870.1
CNVR5840.1
CNVR5947.1
CNVR6050_full
CNVR5856.1
CNVR5894.1
Frequency

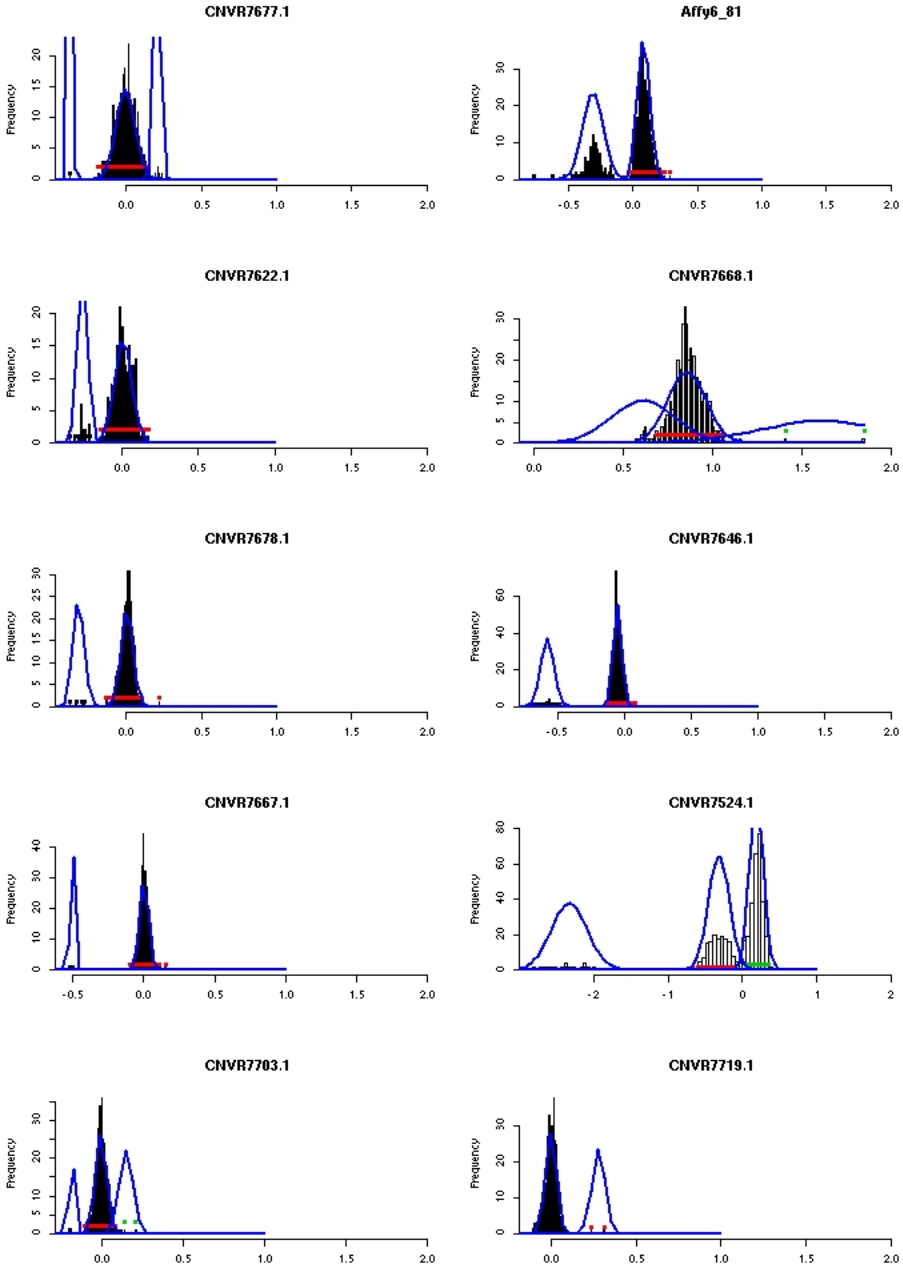

CNVR6126.1
CNVR6255.1
CNVR6214.1
CNVR6174.1
CNVR6124.1
CNVR6072.4
CNVR6128.1
CNVR6219.1
CNVR6097_full
CNVR6213.1
Frequency

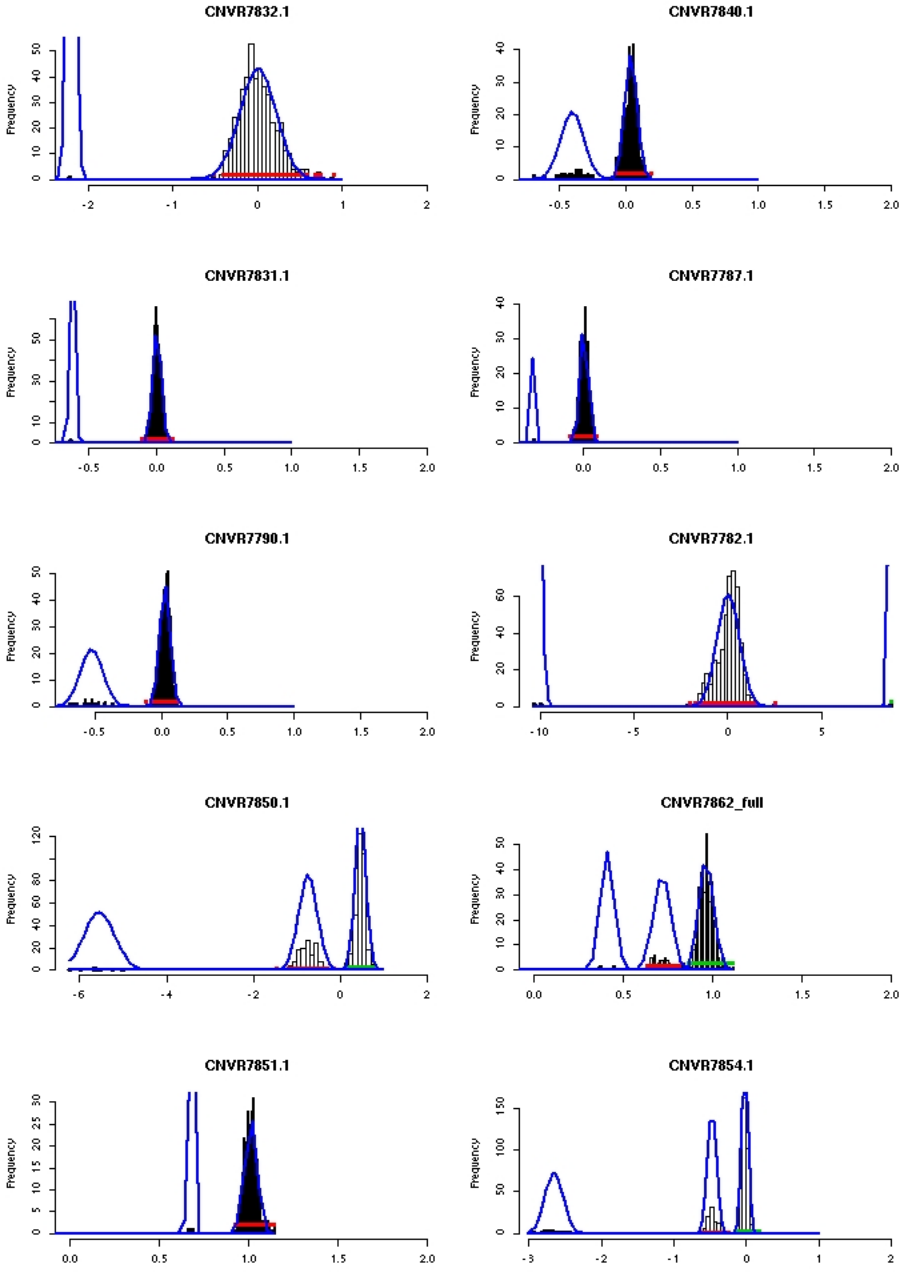

CNVR6543.1
CNVR6515.1
CNVR6345.3
CNVR6548.1
CNVR6517_full
CNVR6327.1
CNVR6356_full
CNVR6382.1
CNVR6426_full
CNVR6347.1
Frequency

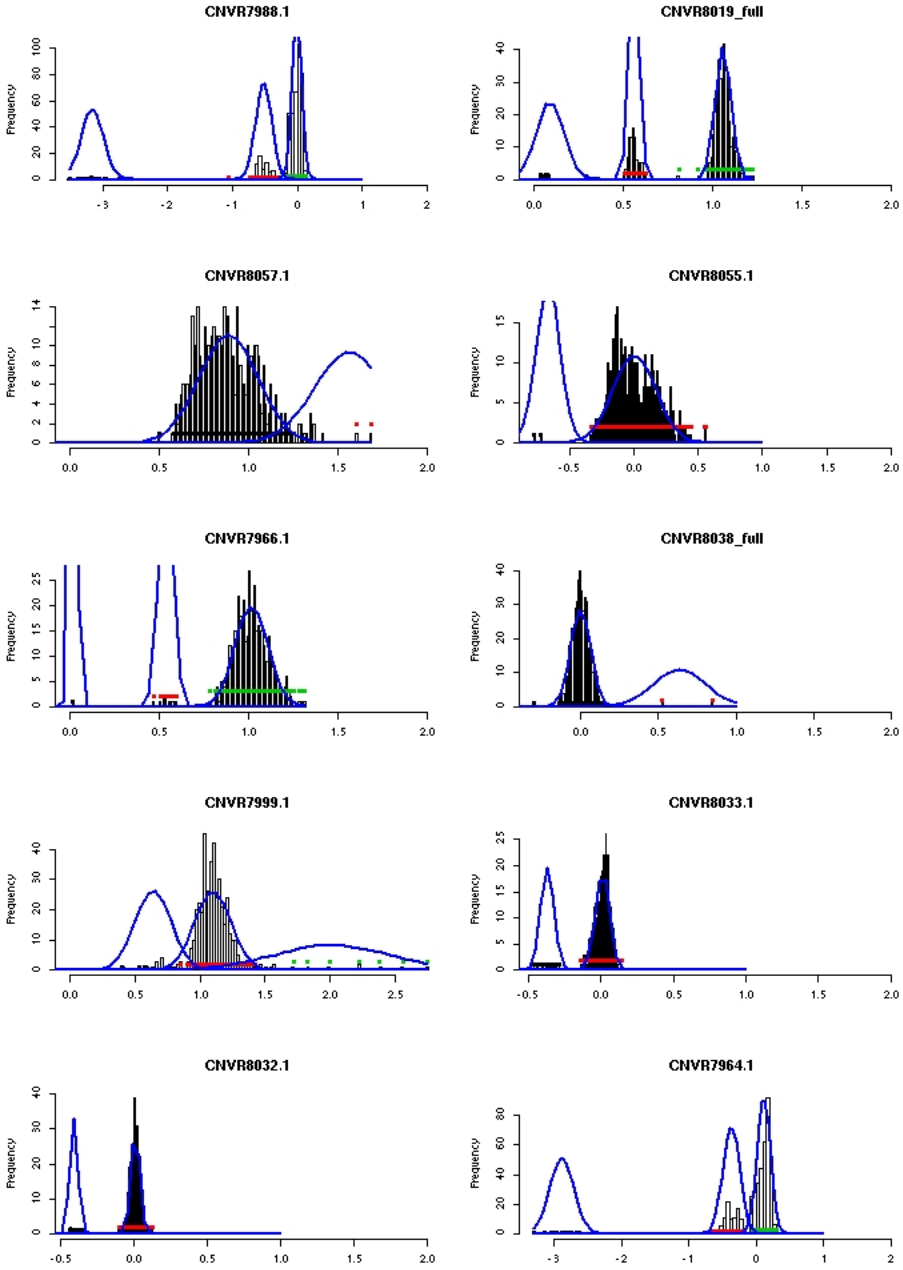

CNVR6771.1
CNVR6688.1
CNVR6783.1
CNVR6569_full
CNVR6754.1
CNVR6719.1
CNVR6835.1
CNVR6757.1
CNVR6607.1
CNVR6741.1
Frequency

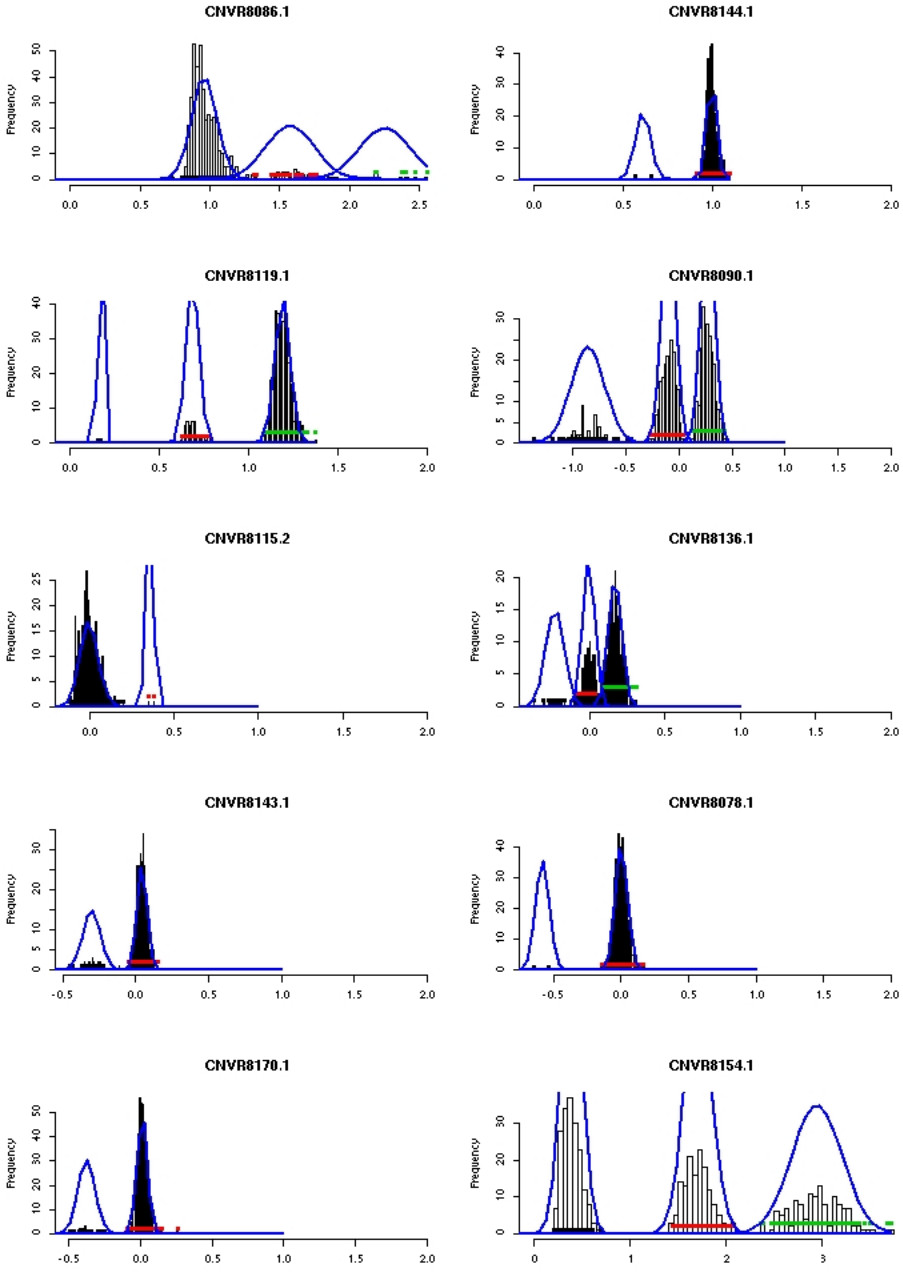

CNVR7138.1
CNVR7126.1
CNVR7003.1
CNVR7077.8
CNVR7095.4
CNVR7114.8
CNVR7074.1
CNVR7056.1
CNVR7094.1
CNVR7010.1
Frequency

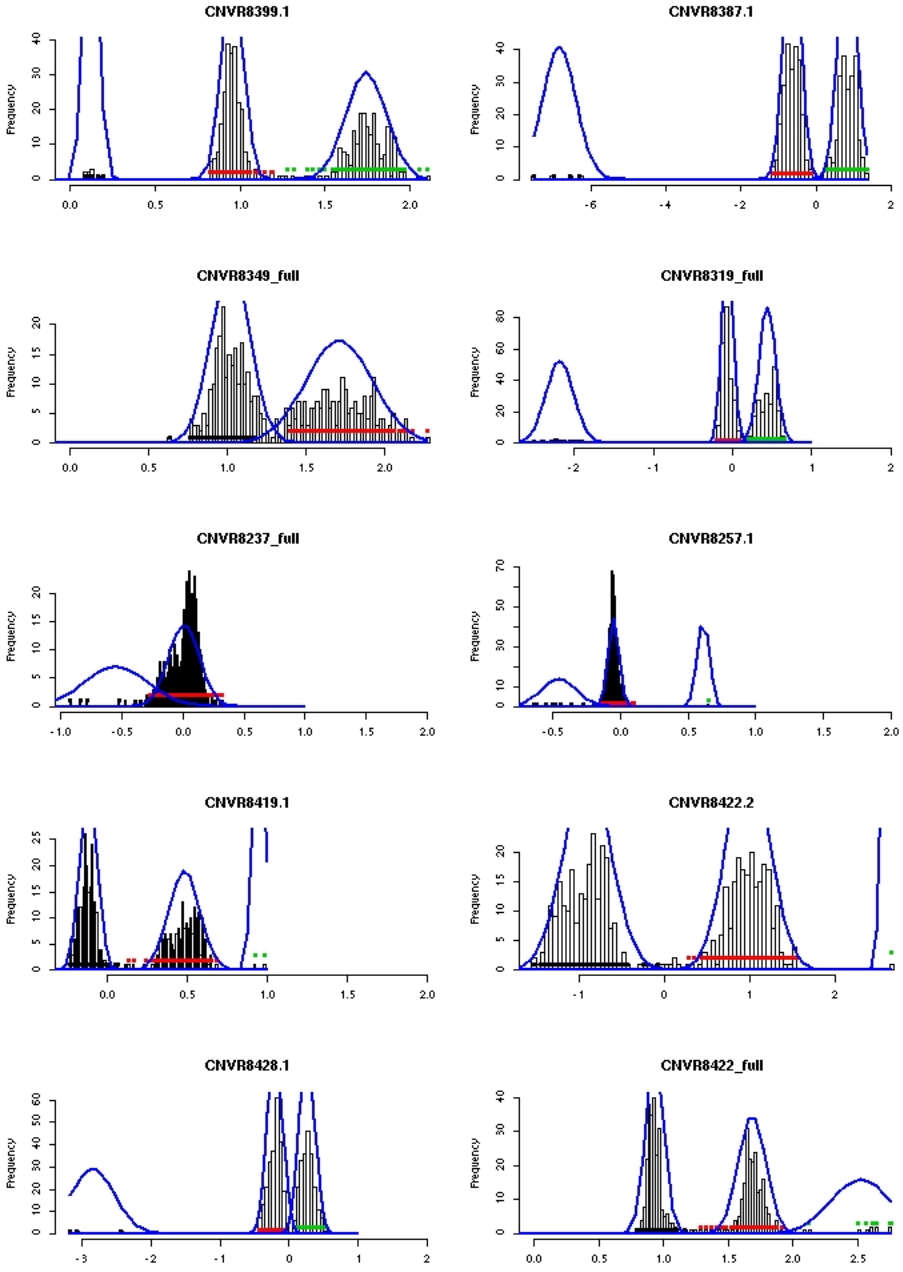

CNVR7358.1
CNVR7407.1
A1081
CNVR7372.1
CNVR7272.1
CNVR7354.1
CNVR7388.1
CNVR7250.1
CNVR7278.1
CNVR7256_full
Frequency


CNVR7677.1
Affy6_81
CNVR7622.1
CNVR7668.1
CNVR7678.1
CNVR7646.1
CNVR7667.1
CNVR7524.1
CNVR7703.1
CNVR7719.1
Frequency


CNVR7832.1
CNVR7840.1
CNVR7831.1
CNVR7787.1
CNVR7790.1
CNVR7782.1
CNVR7850.1
CNVR7862_full
CNVR7851.1
CNVR7854.1
Frequency


CNVR7988.1
CNVR8019_full
CNVR8057.1
CNVR8055.1
CNVR7966.1
CNVR8038_full
CNVR7999.1
CNVR8033.1
CNVR8032.1
CNVR7964.1
Frequency


CNVR8086.1
CNVR8144.1
CNVR8119.1
CNVR8090.1
CNVR8115.2
CNVR8136.1
CNVR8143.1
CNVR8078.1
CNVR8170.1
CNVR8154.1
Frequency


CNVR8399.1
CNVR8387.1
CNVR8349_full
CNVR8319_full
CNVR8237_full
CNVR8257.1
CNVR8419.1
CNVR8422.2
CNVR8428.1
CNVR8422_full
Frequency