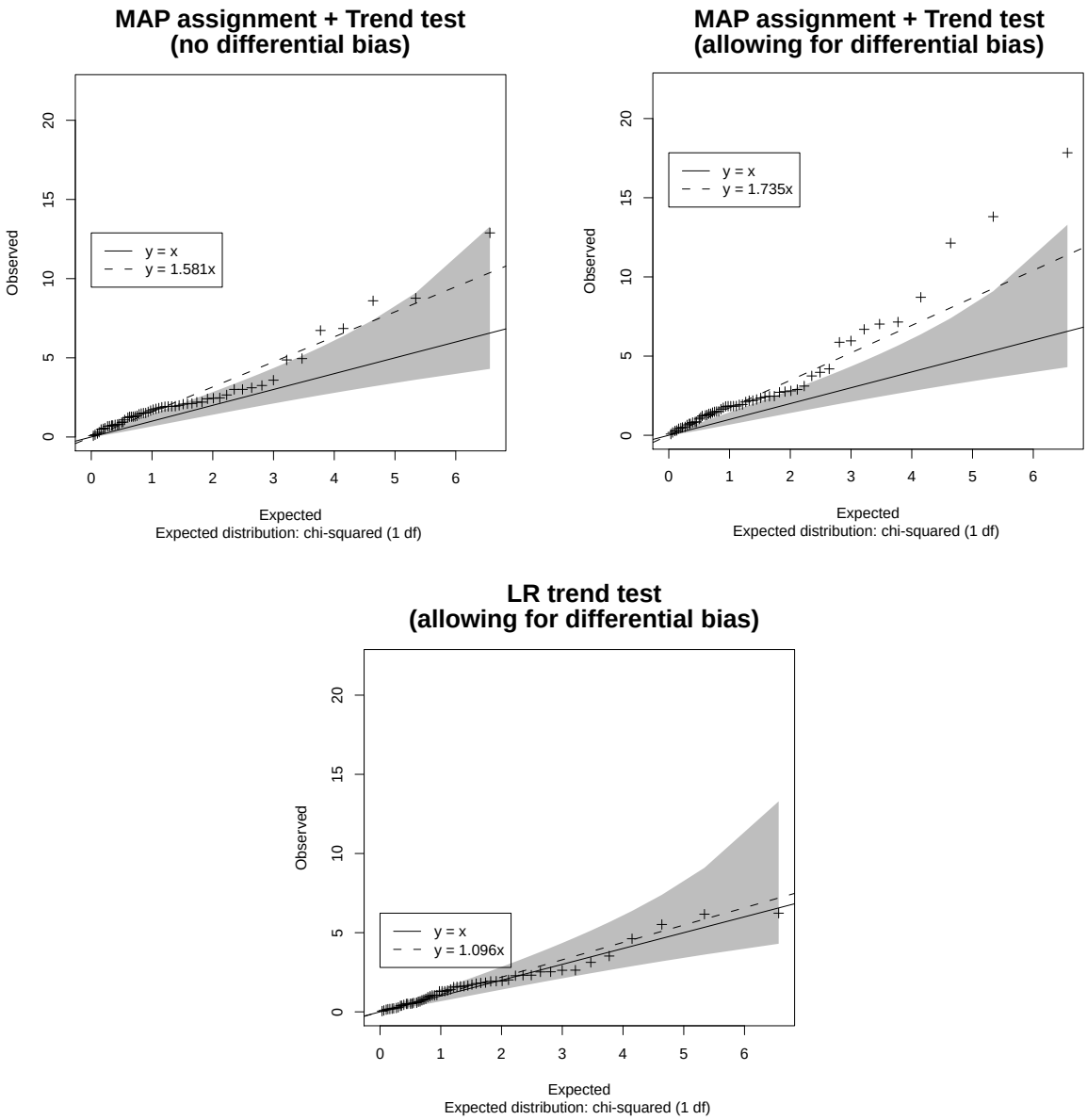
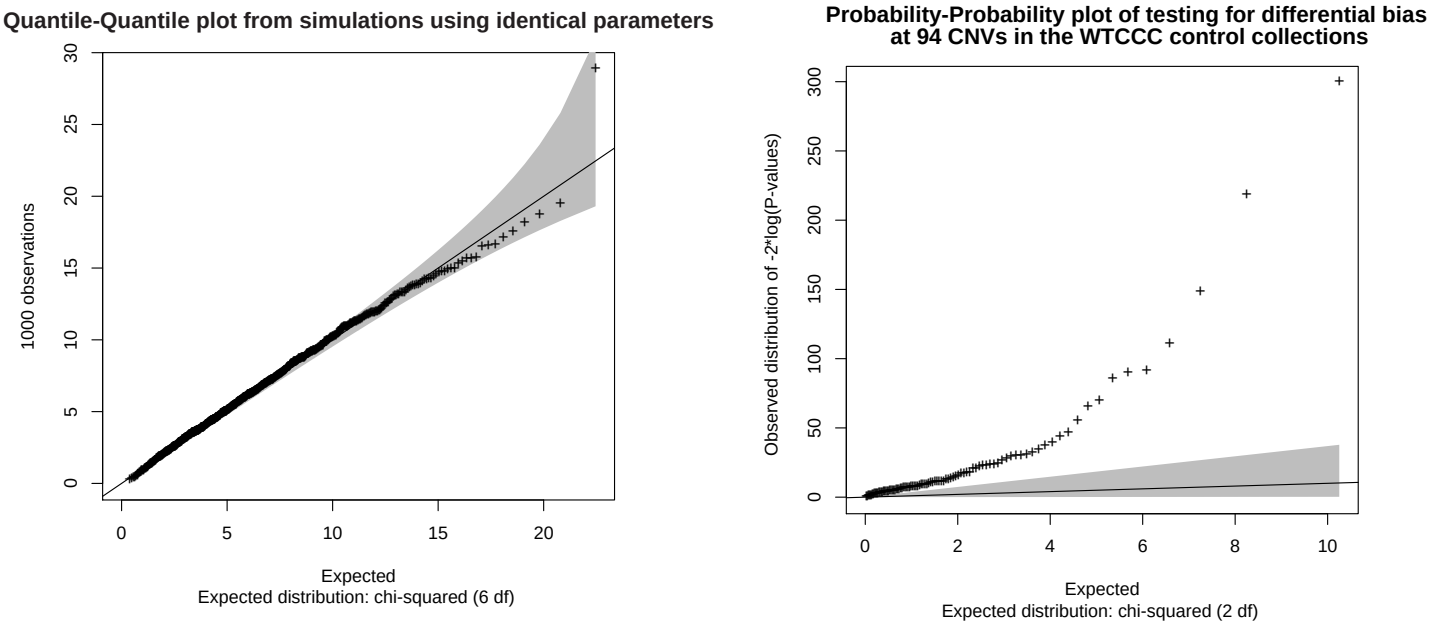
Supplementary Figure 1



Supplementary Figure 1. Quantile-Quantile plot of control/control association tests on empirical data on 95 CNVs in Wellcome Trust Case Control Consortium data

A. the QQ plot resulting from a Cochran-Armitage trend test of the mixture model assignment, not allowing for differential bias; B. the QQ plot resulting from a Cochran-Armitage trend test of the mixture model assignment, allowing for differential bias; C. the QQ plot resulting from the LR trend test. The grey area represents the 95% confidence interval on each quantile.

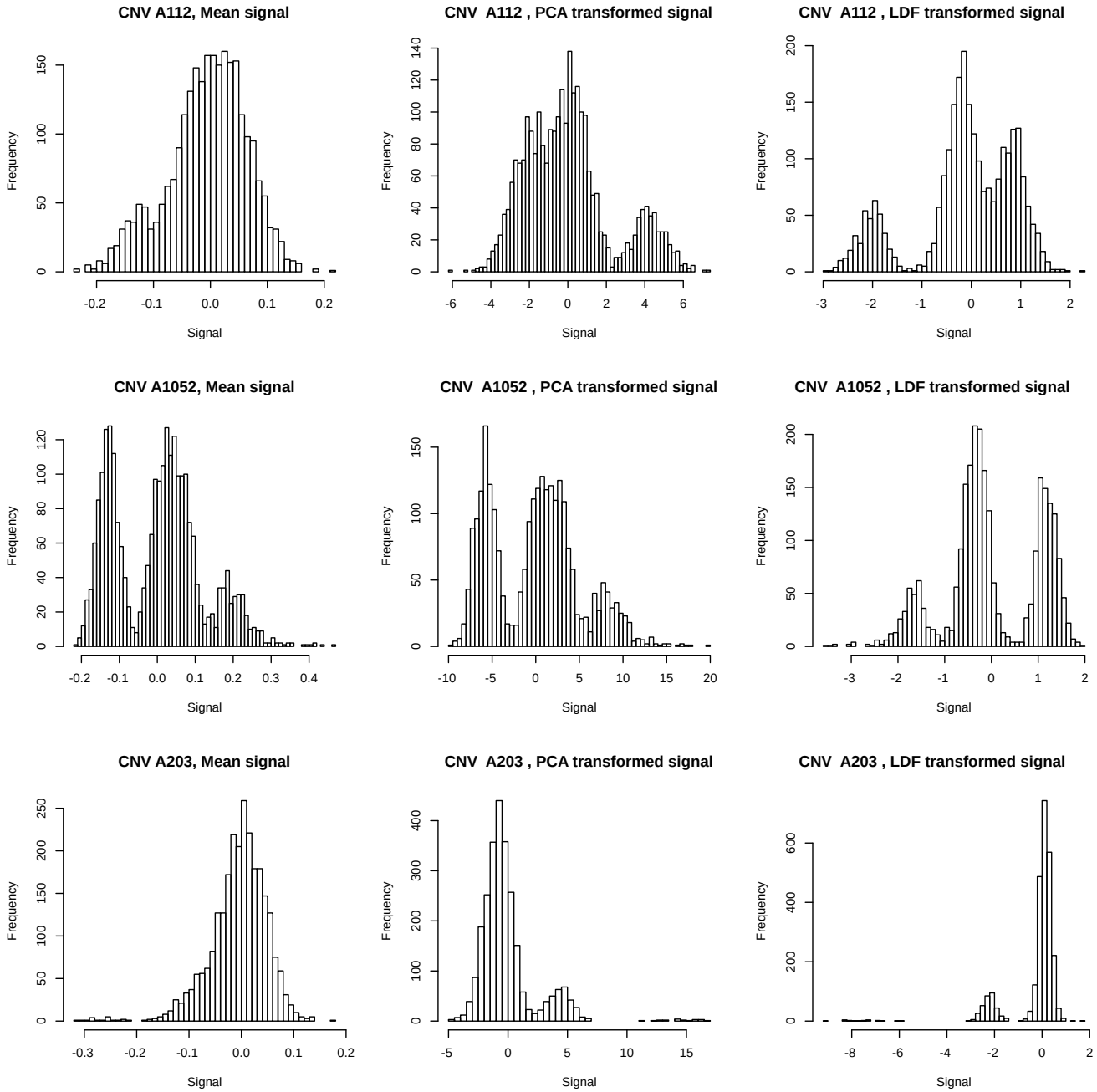
Supplementary Figure 2



Supplementary Figure 2. Quantile-Quantile plots of differential bias test statistics in simulations and real data.

Panel A shows a Quantile-Quantile plot of the differential bias test statistic (6df) in 1000 simulations of two samples drawn from the same underlying measurement distribution. Panel B shows a Probability-Probability plot comparing the observed vs. expected distribution when testing for differential bias between the two WTCCC control cohorts at 94 CNVs. Under the null hypothesis the clustering parameters are identical for both cohorts, and under the alternative these parameters are allowed to differ. A P-value is obtained for each of the 94 CNVs, and this figure shows the distribution of –2log(P), which under the null should be distributed as chi-square with 2df. The differing numbers of copy number classes between the 94 CNVs means that test statistics themselves are not directly comparable.

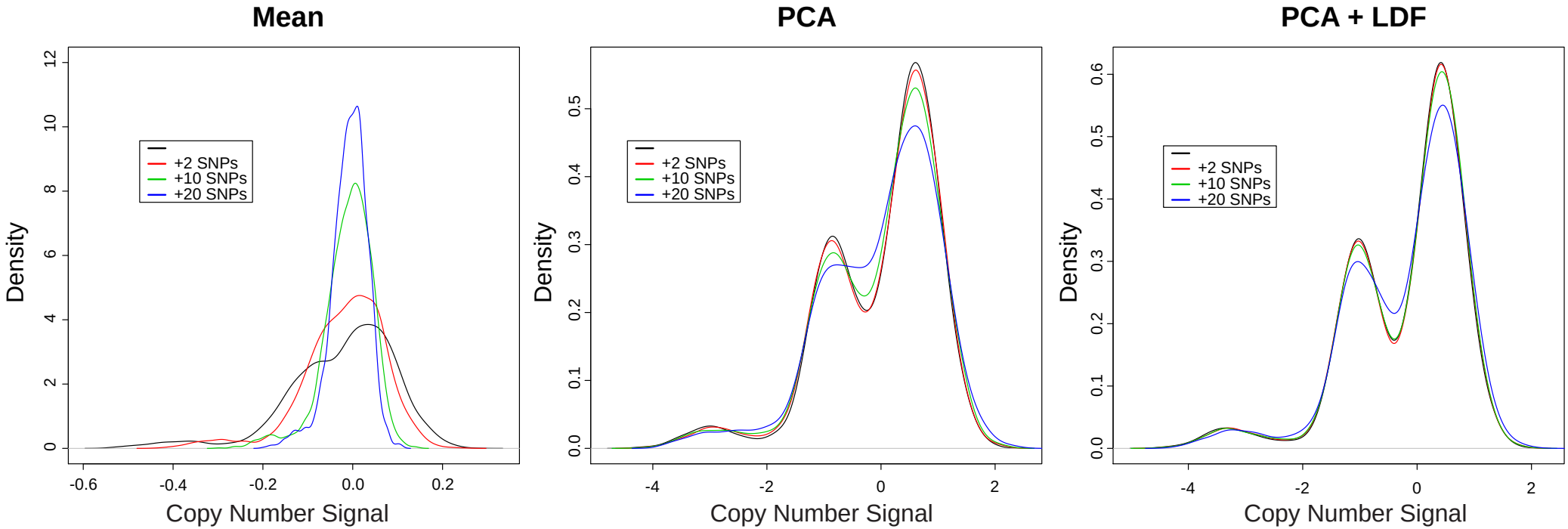
Supplementary Figure 3



Supplementary Figure 3. Histograms of copy number signal using different probe summary methods

Histograms of copy number signal are plotted for three different probeset summary methods: mean, Principal Component Analysis (PCA) and Linear Discriminant Function (LDF), see Methods for details, for three CNVs from the WTCCC data (A112, A1052 and A203, described in Supplementary Table 3). The sign of the PCA and LDF transformed measurements are arbitrary, and so cluster positions sometimes appear to be reversed.

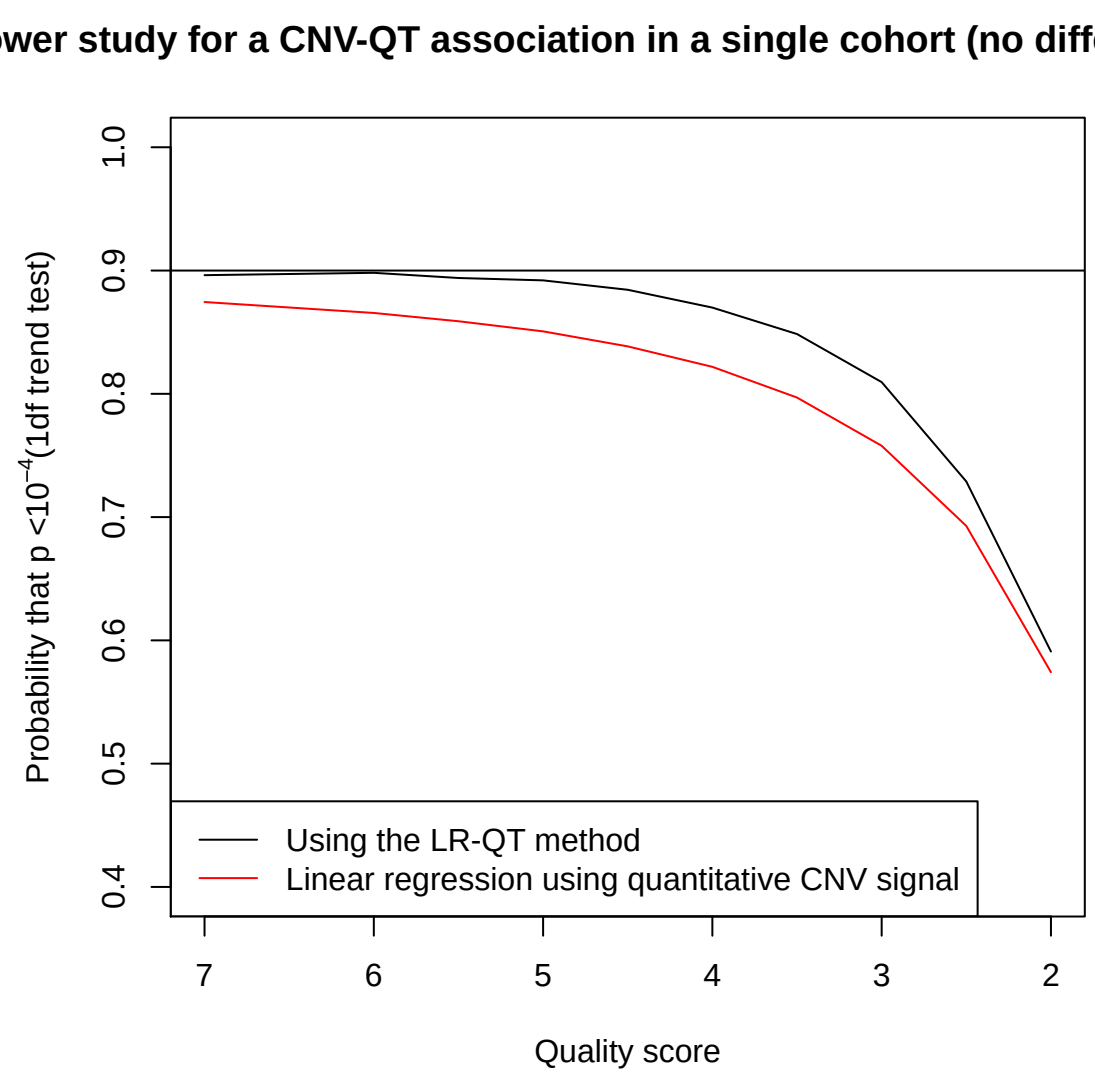
Supplementary Figure 4



Supplementary Figure 4. Impact of over-specification of CNV boundaries on different probe summary methods.

The three panels show the performance of three different probe summary methods as the boundaries of a CNV from the WTCCC data are overestimated by 1,5 and 10 SNPs on both flanks. Panel A shows the impact of over-specification on the arithmetic mean, Panel B shows the impact on the first principal component, and Panel C shows the impact on the Linear Discriminant Function applied to the mixture model classification of the first principal component, as described in the main text.

Supplementary Figure 5



Supplementary Figure 5. Statistical power of the LR-Quantitative Trait test, relative to linear regression. Power study using simulated data to measure the statistical power to detect correlations between DNA copy number and quantitative traits. We simulated a cohort of 2,000 individuals with a bi-allelic CNV in Hardy-Weinberg equilibrium with minor allele frequency of 30%. We also assumed a linear relationship between number of copies and the average quantitative trait measurement, and Gaussian noise around the quantitative trait measurements in each copy number class. The model for gene expression and the effect of the number of copies has been set to provide 90% power when the DNA copy number is known with certainty. The black line uses our LR-QT method. The red line directly correlates the measurement from the CNV assay with the gene expression measurement in a linear regression.

Supplementary Table.
Ninety-five CNVs tested for association between WTCCC control groups

CNV_ID	chr	start(b35)	end(b35)	LR trend test statistic	association p value
W254	1	73355013	73623441	0.2181867	0.64
W463	1	162356681	162426576	0.5721205	0.449
W482	1	165856060	166007288	2.280705	0.131
W504	1	172968878	173023892	0.04139505	0.839
W523	1	177914644	178027769	9.41E-06	0.998
W547	1	182904835	182937524	0.1943785	0.659
W795	2	2575107	2597548	1.060549	0.303
W798	2	2803328	2844314	0.08718403	0.768
W803	2	3000334	3188301	1.933595	0.164
W862	2	14651055	14660690	0.5115853	0.474
W942	2	37832592	37916099	0.2338394	0.629
W1045	2	63221880	63851068	0.2838643	0.594
W1103	2	79384352	79388559	0.05961949	0.807
W1117	2	82106021	82234972	0.01199373	0.913
W1118	2	82287535	82436383	0.2272044	0.634
W1626	2	239264981	239292492	2.528389	0.112
W1944	3	100210983	100225648	1.036198	0.309
W1977	3	113585255	113596024	1.297669	0.255
W2033	3	131233075	131291453	0.400479	0.527
W2044	3	134222823	134288270	0.01438993	0.905
W2137	3	162880168	163020601	0.2169759	0.641
W2142	3	166523818	166555708	0.01525146	0.902
W2272	4	6531318	6541270	0.2006228	0.654
W2280	4	9856891	10083263	1.785735	0.181
W2433	4	42965948	43048682	0.1067171	0.744
W2454	4	52539564	52753517	0.3228591	0.57
W2664	4	120566529	120872359	2.311583	0.128
W2755	4	153147391	153154372	1.822218	0.177
W2960	5	15706765	15776346	1.3418	0.247
W3207	5	97076449	97125076	0.07117197	0.79
W3210	5	98136658	98375543	6.23012	0.0126
W3306	5	126463761	126602292	2.018718	0.155
W3444	5	165703458	165716625	2.631765	0.105
W3708	6	45106893	45458085	0.04450124	0.833
W3754	6	67065532	67106127	0.72747	0.394
W3784	6	77496587	77558299	0.2436023	0.622
W3786	6	79036117	79083405	0.04823872	0.826
W3789	6	79620938	79839756	0.0915497	0.762
W3948	6	130589849	130691297	0.638522	0.424
W4305	7	64001544	64687746	0.07033807	0.791
W4341	7	77504262	77609125	0.0268241	0.87
W4355	7	82208167	82258900	0.1671222	0.683
W4569	7	149650799	149790679	1.303119	0.254
W4888	8	93077468	93173097	0.1586687	0.69
W4907	8	98797002	98919923	0.4462098	0.504

W5041	8	137757137	137933062	0.1778802	0.673
W5042	8	138362332	138406796	1.418262	0.234
W5055	8	141689895	142036454	0.1371476	0.711
W5157	9	21673289	21807777	0.5046897	0.477
W5224	9	38761831	44108554	0.5120674	0.474
W5293	9	87186120	87198911	6.167063	0.013
W5351	9	101793026	101801974	0.6557536	0.418
W5379	9	107801610	107808792	0.06216589	0.803
W5576	10	13096593	13104229	0.981545	0.322
W5669	10	27262326	27270681	1.923495	0.165
W5701	10	33292269	33351038	1.658113	0.198
W5883	10	90363825	90523215	0.02438729	0.876
W5885	10	90786131	90797223	1.845579	0.174
W5898	10	93335238	93339994	4.616483	0.0317
W5930	10	101645317	101754574	0.1928718	0.661
W6210	11	24596631	24596789	0.1540546	0.695
W6368	11	76445890	76470079	0.1000863	0.752
W6462	11	99017466	99070312	3.525258	0.0604
W6489	11	107461287	107885084	0.1798609	0.671
W6641	12	7854479	8279923	0.5014894	0.479
W6740	12	30247605	30260529	2.624342	0.105
W6747	12	31130867	31301551	0.03462076	0.852
W6755	12	33176320	33224334	0.007211411	0.932
W7275	13	69627005	69687992	1.949261	0.163
W7763	14	92665344	92699712	0.02776828	0.868
W7789	14	97677436	97717303	0.0066519	0.935
W7821	14	105149735	106287351	0.3778723	0.539
W7824	15	22580329	22580344	0.3909269	0.532
W7891	15	41619215	41834930	2.315479	0.128
W8177	16	21441805	22620480	0.9271343	0.336
W8199	16	28023752	28173286	1.301272	0.254
W8203	16	34307201	34618468	1.722144	0.189
W8438	17	16512229	16732685	0.2836425	0.594
W8458	17	25937741	26289747	0.09414661	0.759
W8495	17	40343050	40347575	2.522201	0.112
W8497	17	41134385	41357489	1.615035	0.204
W8498	17	41518415	42144468	0.5719503	0.449
W8641	17	73152020	73160154	1.561511	0.211
W8760	18	33555925	33610573	0.441116	0.507
W8843	18	54081842	54089919	0.01045802	0.919
W8847	18	55028517	55034846	0.1699787	0.68
W8883	18	60511319	60569424	0.7729019	0.379
W8893	18	64897804	64908767	0.5571776	0.455
W9097	19	59939919	60040503	1.577734	0.209
W9319	20	52488072	52491840	0.5234058	0.469
W9337	20	55212415	55229082	3.12407	0.0771
W9342	20	55984705	55996322	5.517019	0.0188
W9407	21	18975170	19002805	1.608739	0.205
W9622	22	41570644	41842773	1.042722	0.307

W9624	22	42193069	42445743	0.8273893	0.363

Supplementary Methods

Normalization

Each Affymetrix 500k probeset is summarized into allelic intensities by taking the median of the perfect match probes for the A allele and B alleles respectively. Using heterozygotes with high posterior probabilities from the WTCCC analysis [1] for which the intensity for the A and B alleles is assumed to be equal, a locus dependent correction factor is calculated, ϕ, which accounts for systematic differences in A and B allele intensities. The A and B allele intensities are then combined to give a log diploid copy number intensity $I = \log_2[A+\phi B]$. Quantile Normalization [2] is then performed on the I values separately for the *Nsp*I and *Sty*I arrays. The intensity at a locus is known to be dependent on both the restriction fragment length and GC content [3]. These are corrected for using cubic polynomials. The intensity values are then median normalized with respect to the median at that locus within a 96-well microtitre plate. The resultant value gives the diploid copy number $\log_2$ ratio, R. There is also known to be a non-linear autocorrelation in the R values across the genome. A loess correction on a sample by sample basis is performed to remove this bias [4]. This pipeline was implemented in C++ utilizing the ROOT framework [5].

Testing for differential errors

The likelihood ratio testing framework provides a natural way to test for the differences in measurement characteristics between cases and controls. Under the null hypothesis, the signal model described in Methods is fitted constraining the component means and variances to be the same in both groups. Under the alternate hypotheses the means, variances or both are allowed to differ between groups. If a CNV has n components, the likelihood ratio test statistic in the presence of no differential bias is distributed as χ^2 with n, n or 2n degrees of freedom, respectively.

Probe weighting

The signal x is a composite of measurements from a collection of probesets, each one being a normalised allele sum intensity for one SNP within the presumed CNV (see above). The simplest way to construct a composite signal is to sum these measurements. However an improvement is to use the first principal component; this down-weights measurements not highly correlated with the remainder. Having fitted the Gaussian mixture model as described above, a further improvement can be obtained by using the first Fisher linear discriminant function, with the calculations modified to allow for the uncertainty in group assignment. If there are M probesets contributing to the score, we have an $S \times M$ matrix of signal intensities, U and, following an initial fit of the Gaussian mixture model, we have the $S \times (N+1)$ matrix of posterior copy number probabilities, P. We then find the first canonical correlation *i.e.* we compute the values of n and h which maximize the correlation between Un and Ph. The first canonical variate Un then provides our improved composite score. These calculations are carried out by a standard function in R. When a composite signal is calculated from only two probesets, rather than use the probe-weighting method described above we simply take the mean of the two probesets.

The signal-to-noise ratio, Q

In simulations, the component Gaussian distributions for signal intensity conditional upon underlying copy number had equal standard deviation and equally spaced means. The signal-to-noise ratio, Q, is then defined as the ratio of the separation between adjacent means divided by the within-component standard deviation. For real data, the position is more complicated since spacings and standard deviations can vary and, consequently, the signal-to-noise ratio varies. While it would be possible to derive an overall measure of the clustering quality from properties of the likelihood function, this would not serve our requirement for an easily understood measure and we have elected to define Q as a simple averaged value of the signal to noise ratios between adjacent copy numbers, with weights reflecting the frequency of copy numbers. Precisely, summing over all pairs of adjacent clusters (i,j):

$$Q = \frac{\sum_{i,j} \alpha_i \alpha_j \frac{|\mu_i - \mu_j|}{\sigma_{ij}}}{\sum_{i,j} \alpha_i \alpha_j}$$

where α_i is the relative frequency of cluster *i*, μ_i the center of the cluster and $\sigma_{ij} = \frac{\alpha_i \sigma_i + \alpha_j \sigma_j}{\alpha_i + \alpha_j}$ is a weighted standard deviation for the pair (*i,j*).

Simulations

The six general methods of copy number association testing outlined in the main text are examined for type I error rates over a range of clustering qualities and in the presence of two different forms of bias. For five of these methods, statistical tests of trend are applied, which are expected to be asymptotically distributed as χ^2 with one degree of freedom.

Simulated Gaussian mixtures with three components are generated in R using the package `nor1mix`. The mixture probabilities follow Hardy-Weinberg equilibrium with a minor allele frequency, *f*, of 0.3. The means and variances of the components are given by μ=(2,3,4) and $\sigma^2 = \sigma_F^2(1, s, s^2)$ where *s* is a scale factor equal to 1.5 and σ_F^2 is the variance of the first component, which can be expressed in terms of Q,μ,f,s.

To investigate the effect of clustering quality on the Type I error, two groups each containing 2000 individuals are sampled from ten mixtures with *Q* ranging from 7.5 to 3.0. To simulate the effect of differential mean, two parent distributions are constructed with *Q* fixed at 4.5. The first of these has component means located at μ and the second has component means located at $\mu' = \mu + \Delta\mu$ with $\Delta\mu$ taking ten values ranging from 0 to 0.18. A group containing 2000 individuals is sampled from each of these parent distributions. Differential variance is simulated in an analogous way with parent distributions constructed with *Q* fixed at 4.5, the first having component variances given by σ^2 and the second by $\sigma'^2 = \sigma^2 \times \Delta\sigma^2$ with $\Delta\sigma^2$ taking ten values ranging from 0 to 1.8. A group containing 2000 individuals is sampled from each of these parent

distributions.

For each different hypothetical association test (of which there are thirty in total) 1000 pairs of distributions are simulated and a quantile-quantile plot produced. To quantify the over-dispersion a straight line is fit to the first 90% of values with the intercept constrained to be zero. The value λ is given by the fitted gradient of the line.

References

1. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
2. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93 (2003).
3. Nannya, Y. et al. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res* **65**, 6071-9 (2005).
4. Marioni, J.C. et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridisation. *Genome Biology* **8**, R228 (2007).
5. Brun, R. & Rademakers, F. ROOT - An Object Oriented Data Analysis Framework. *Nucl. Inst. & Meth. in Phys. Res A* **389**, 81-86 (1997).