Correspondence

Whole genome–amplified DNA: insights and imputation

Yik Y Teo, Michael Inouye, Kerrin S Small, Andrew E Fry, Simon C Potter, Sarah J Dunstan, Mark Seielstad, Inês Barroso, Nicholas J Wareham, Kirk A Rockett, Dominic P Kwiatkowski & Panos Deloukas

Supplementary figures and text:

Supplementary Table 1. Description of the datasets included in this study.

Supplementary Table 2. Analysis of SNPs within regions of segmental duplications.

Supplementary Table 3. Coverage calculations with extended threshold.

Supplementary Figure 1. Performance of ϕ29MDA DNA compared to genomic DNA for a typical underperforming SNP.

Supplementary Figure 2. The relative performance of amplified DNA to genomic DNA.

Supplementary Figure 3. Mean ratios of variances against mean % GC content of probe sequences for SNPs on the

Affymetrix array.

Supplementary Figure 4. Coverage of the genome for the SNP arrays as a function of pairwise *r*2.

Supplementary Figure 5. SNP data recovery when missing genotypes are imputed as a function of the initial call rate.

Supplementary Methods

Supplementary Table 1. Description of the datasets included in this study

Cohort	Number of subjects	Genotyping platform	Types of DNA	Population	Genotyping laboratory
58C	1502*	Affymetrix 500K	Genomic	British	Affymetrix
58C	1438*	Illumina 550K	Genomic	British	Wellcome Trust Sanger Institute
ML	278†	Affymetrix 500K	ϕ29MDA	Gambian	Affymetrix
ML	2288†	Illumina 650Y	ϕ29MDA	Gambian	Wellcome Trust Sanger Institute
OBC	2198	Affymetrix 500K	ϕ29MDA	British	Affymetrix
TB	461	Affymetrix 500K (*Nsp*I only)	Genomic	Vietnamese	Genome Institute of Singapore
TB	56	Affymetrix 500K (*Nsp*I only)	ϕ29MDA	Vietnamese	Genome Institute of Singapore

*1,402 individuals were genotyped on both the Affymetrix 500K and Illumina 550K platforms.
†278 individuals were genotyped on both the Affymetrix 500K and Illumina 650K platform

Supplementary Table 2. Analysis of SNPs within regions of segmental duplications.

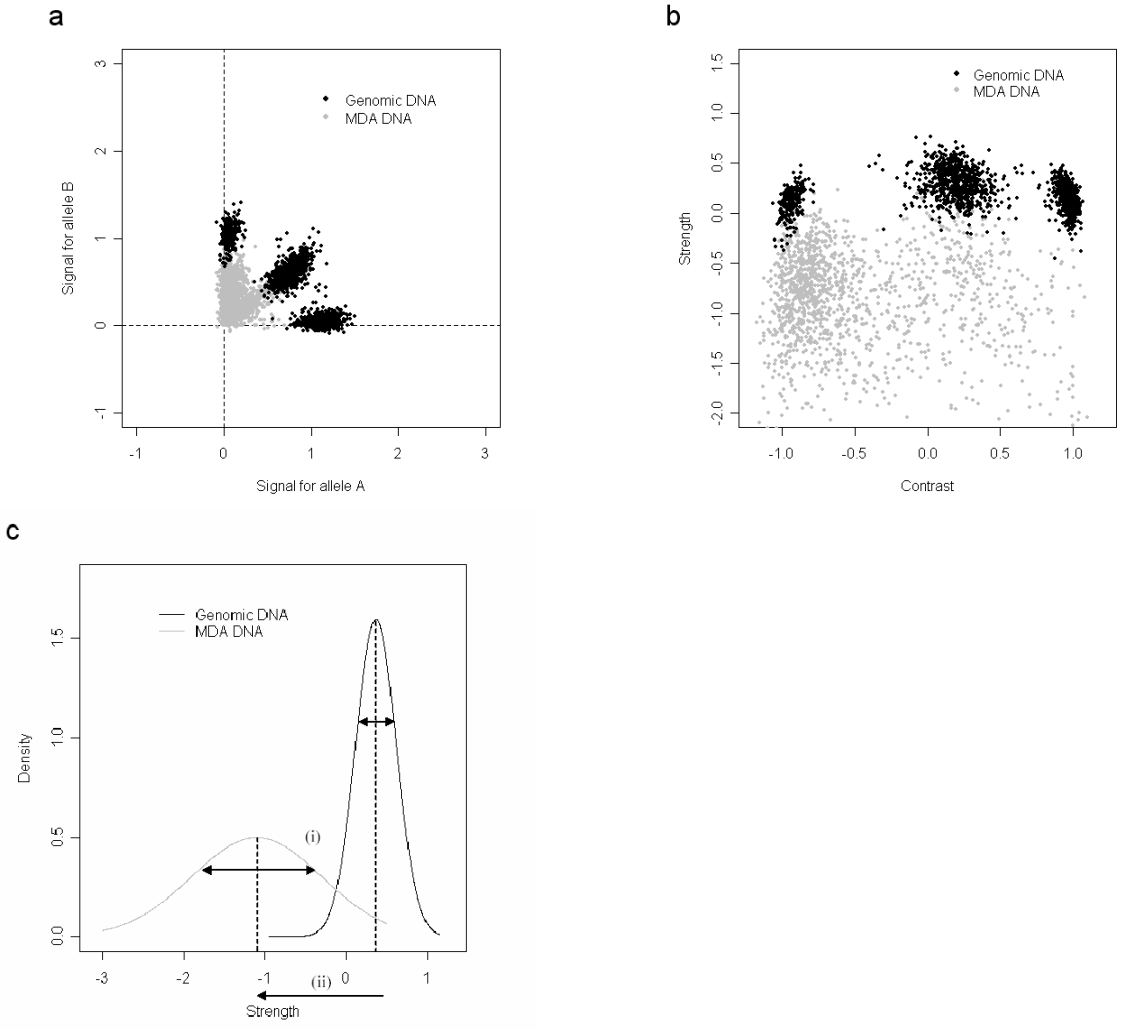
Platform	# SNPs	Genomic DNA		WGA DNA	
		Call rate (%)	# SNPs with < 95% call-rate	Call rate (%)	# SNPs with < 95% call-rate
Affymetrix 500K					
- within seg. dup.	7,710	97.8%	1,072 (13.9%)	95.0%	2,756 (35.7%)
- outside seg. dup.	492,858	98.7%	29,771 (6.0%)	96.3%	109,674 (22.3%)
Illumina*					
- within seg. dup.	5,440	97.6%	461 (8.5%)	92.2%	1,811 (33.3%)
- outside seg. dup.	549,708	98.5%	18,936 (3.4%)	95.0%	102,356 (18.6%)

* SNPs common to both the HumanHap550 and HumanHap650Y arrays.

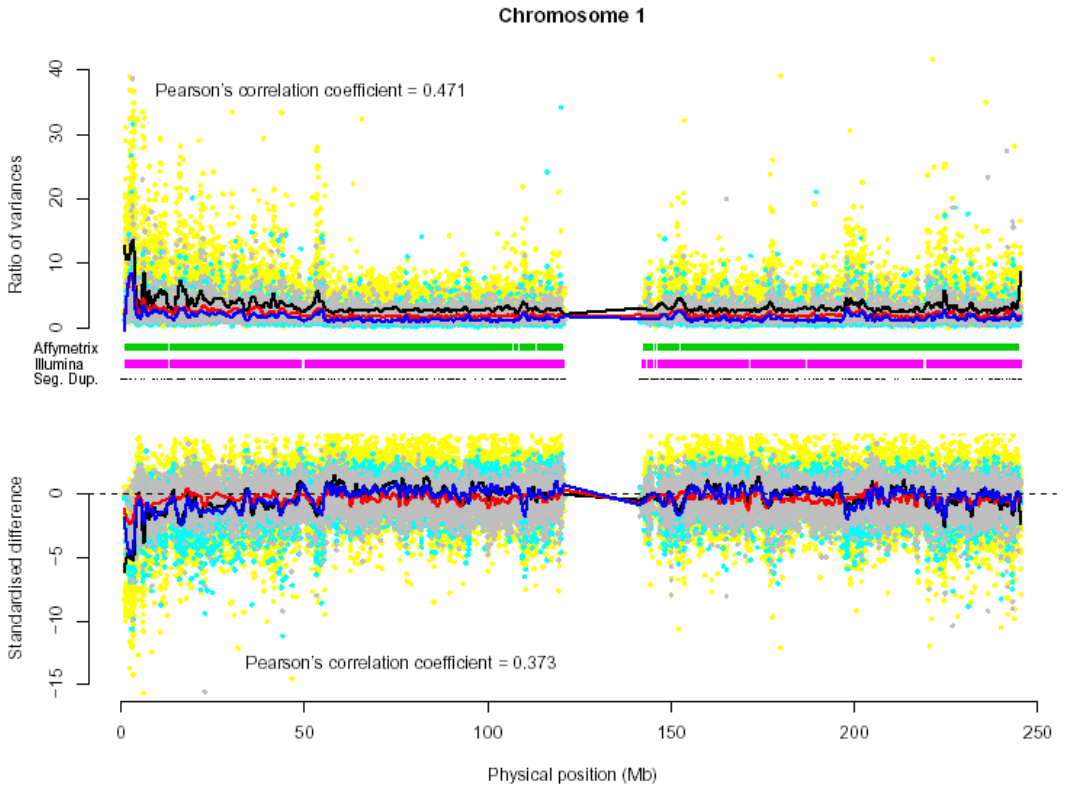
Supplementary Table 3. We extend the coverage calculations for different rates of missingness beyond the manuscript's adopted call rate threshold of 0.95 (less than 5.0% missingness for each SNP). The resultant coverage is calculated at call rate thresholds of 0.90 and 0.97. Genome coverage is calculated at a pairwise tagging r^2 of at least 0.8. We also explored the use of a novel calling algorithm for Illumina platforms which explicitly handles WGA DNA (Teo et al. 2007), and report the resultant genomic coverage for the same data. Numbers in brackets denote the difference between the actual coverage and the benchmark coverage.

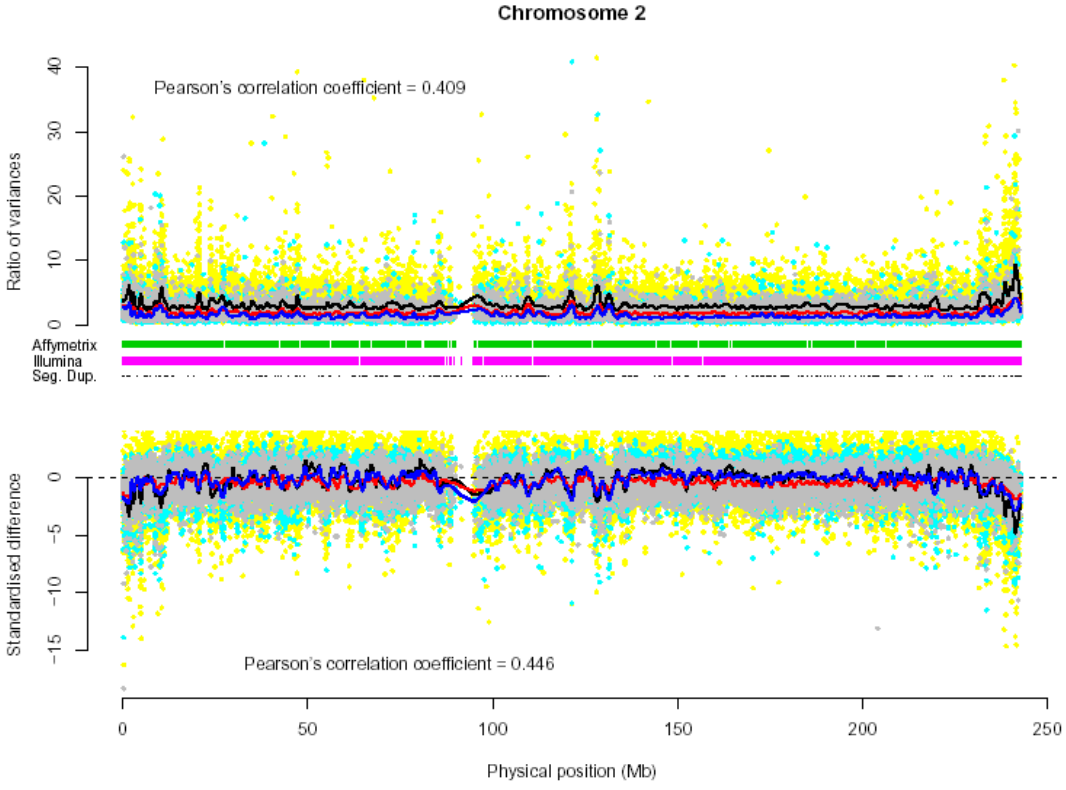
	Benchmark Coverage	**Call Rate Threshold**		
		0.90	**0.95**	**0.97**
CEU				
Affy 500K	60.6	58.5 (-2.1)	54.1 (-6.5)	47.9 (-12.8)
Illumina 650K (GenCall)	81.3	76.8 (-4.5)	73.2 (-8.1)	68.3 (-13.0)
Illumina 650K (Illuminus)	81.3	81.0 (-0.3)	79.9 (-1.2)	76.0 (-5.1)
CHB + JPT				
Affy 500K	63.0	60.8 (-2.1)	56.3 (-6.6)	50.1 (-12.9)
Illumina 650K (GenCall)	80.8	76.8 (-4.0)	73.4 (-7.5)	68.7 (-12.1)
Illumina 650K (Illuminus)	80.8	80.5 (-0.4)	79.4 (-1.4)	75.8 (-5.0)
YRI				
Affy 500K	37.2	34.9 (-2.3)	30.6 (-6.6)	25.1 (-12.1)
Illumina 650K (GenCall)	54.5	48.2 (-6.3)	44.1 (-10.4)	38.8 (-15.7)
Illumina 650K (Illuminus)	54.5	53.9 (-0.6)	52.5 (-2.0)	47.8 (-6.7)

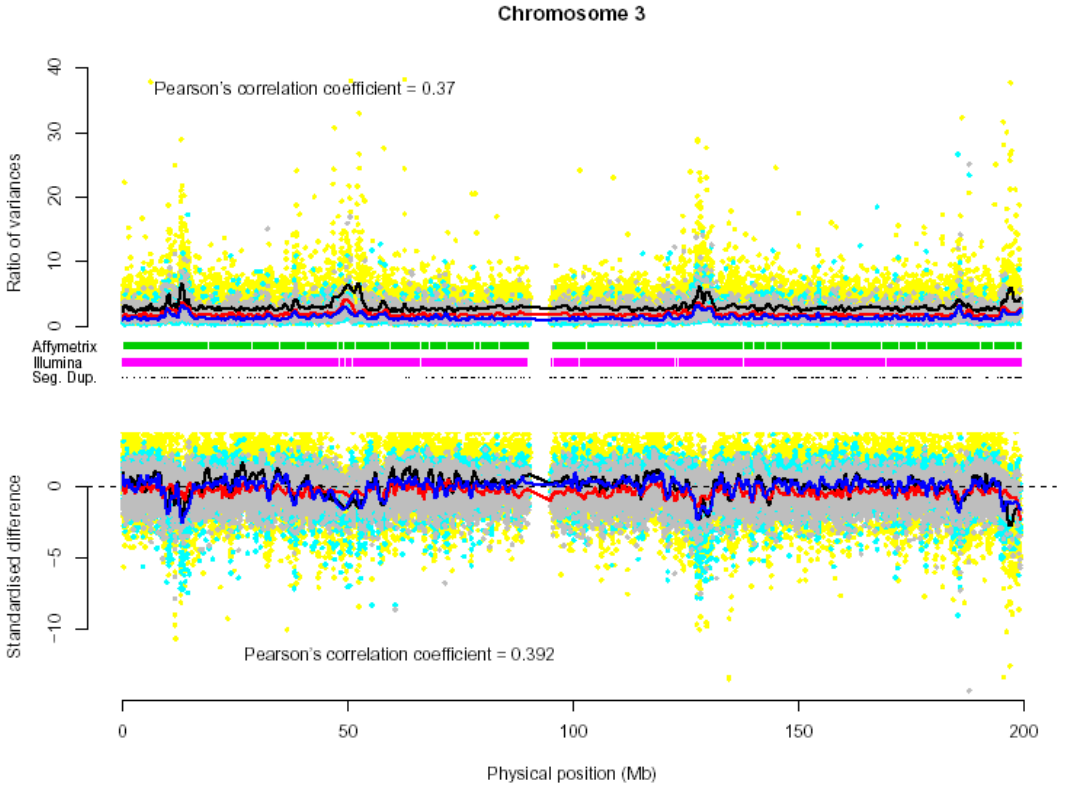
Supplementary Figure 1. Performance of ϕ29MDA DNA compared to genomic DNA for a typical underperforming SNP. (**a**) Clusterplot for a SNP on the signal intensity. (**b**) Clusterplot for the same SNP on the strength-contrast scale. (**c**) A comparison of the hybridization intensities for genomic and amplified DNA, where two effects are observed: (i) an increase in the variability of the hybridization intensities; (ii) a reduction in the mean hybridization intensity; for amplified DNA. For bulk of the SNPs which perform properly, the genotype clusters for genomic and amplified DNA are generally distinct and resemble the dark circles in (**a**).

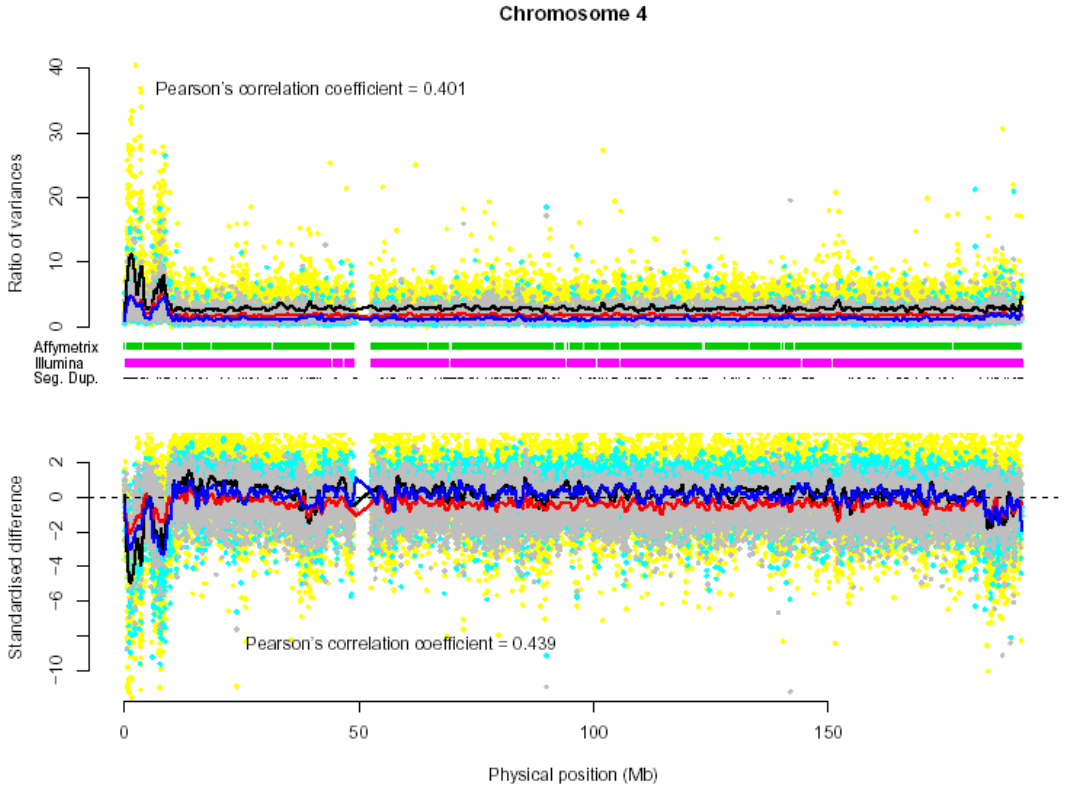


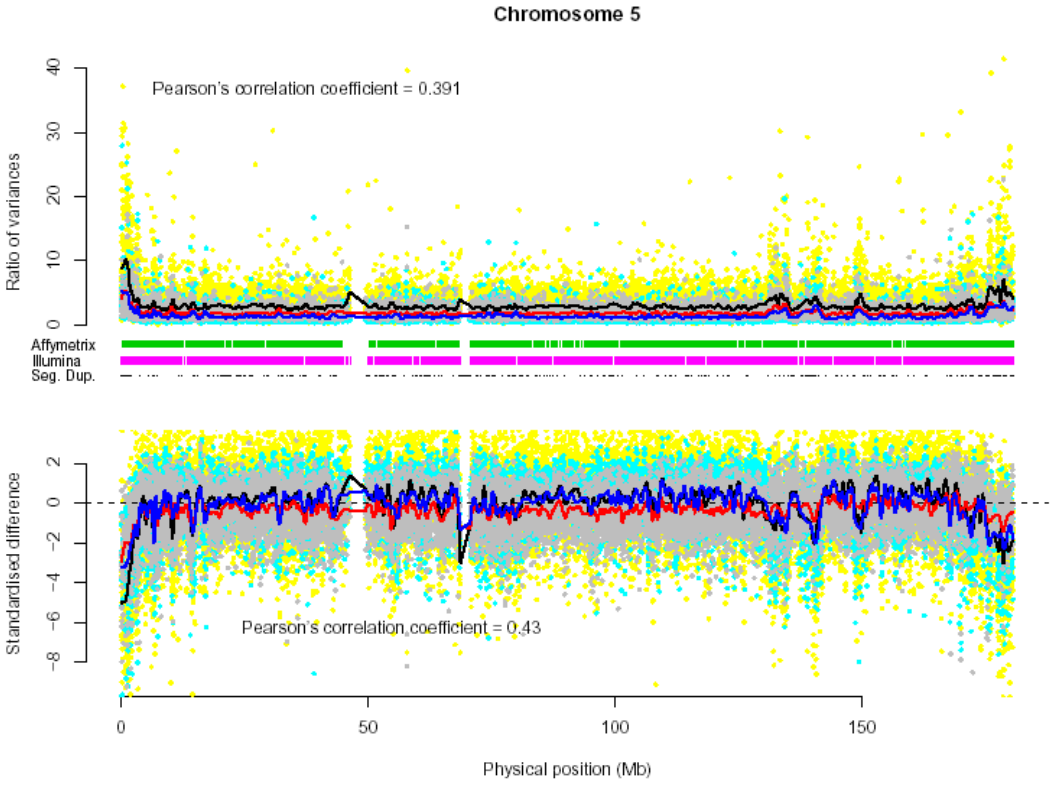
Supplementary Figure 2. The relative performance of amplified DNA to genomic DNA, as quantified by two measures: (i) the ratio of hybridization strengths; (ii) the standardized difference in mean hybridization strength. Each plot shows data from three comparisons of amplified DNA to genomic DNA – Affymetrix: TB (cyan dots and blue lines); Affymetrix: OBC-58C (grey dots and red lines); Illumina: ML-58C (yellow dots and black lines). Pearson's correlation coefficient is calculated to quantify the correlation between the TB and OBC-58C comparisons using the Affymetrix data. Lines below the upper plots indicate regions where SNPs on the platform have call rates <95.0%. The dashes in black indicate regions of segmental duplications. Plots are arranged in chromosomal order.

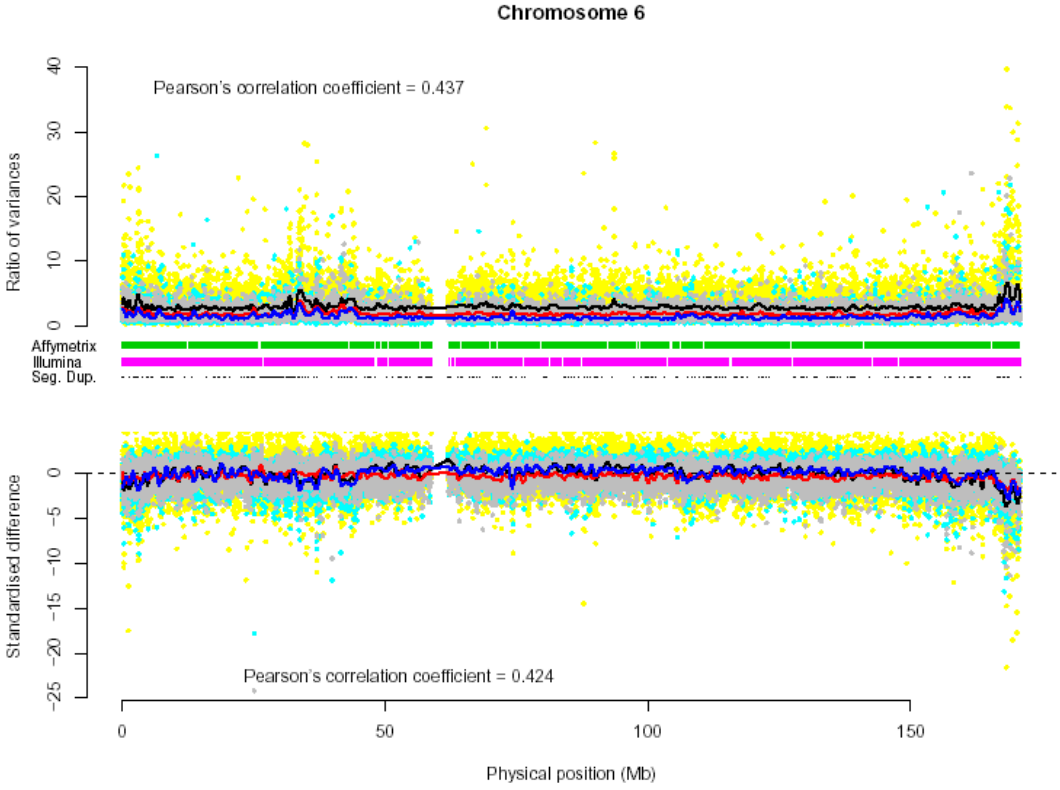

Chromosome 1
Pearson's correlation coefficient = 0.471
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.373
Physical position (Mb)

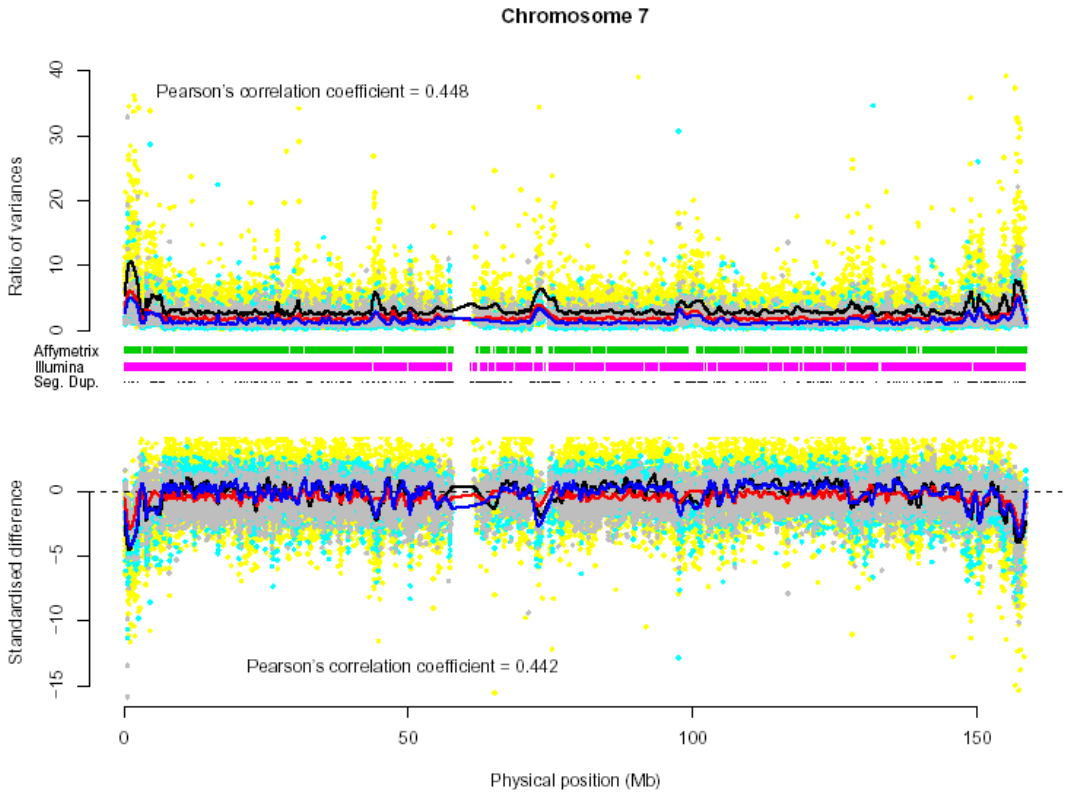

Chromosome 2
Pearson's correlation coefficient = 0.409
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.446
Physical position (Mb)

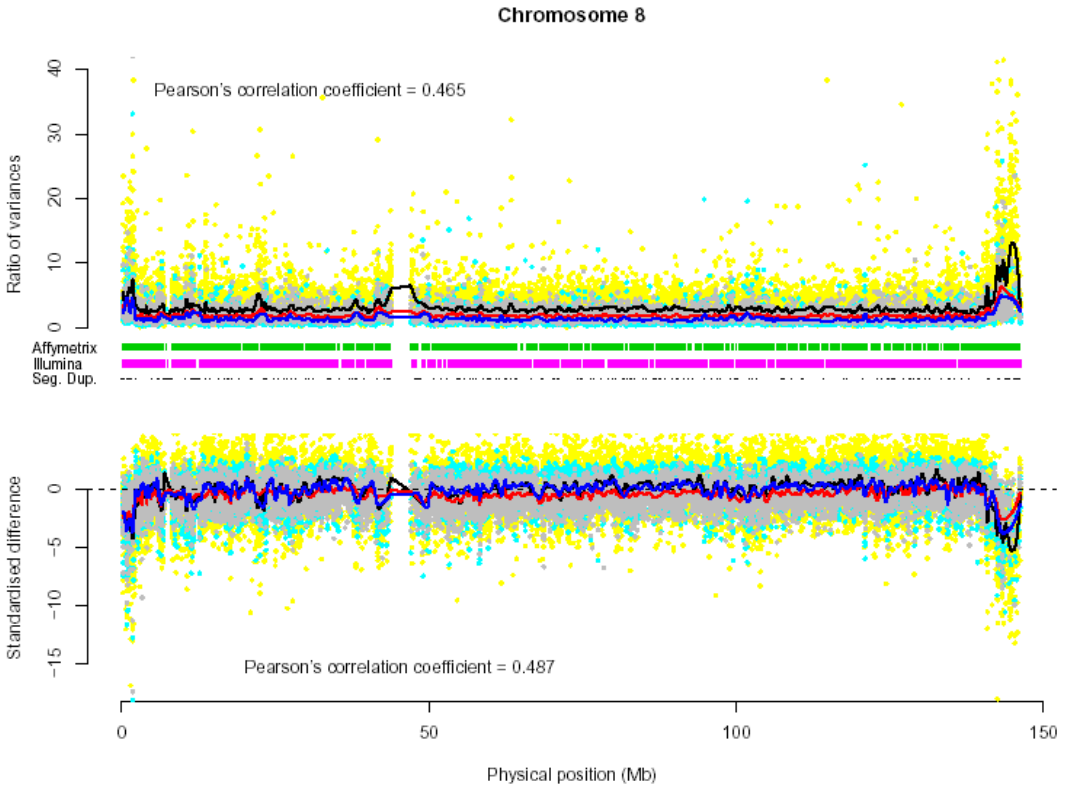

Chromosome 3
Pearson's correlation coefficient = 0.37
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.392
Physical position (Mb)

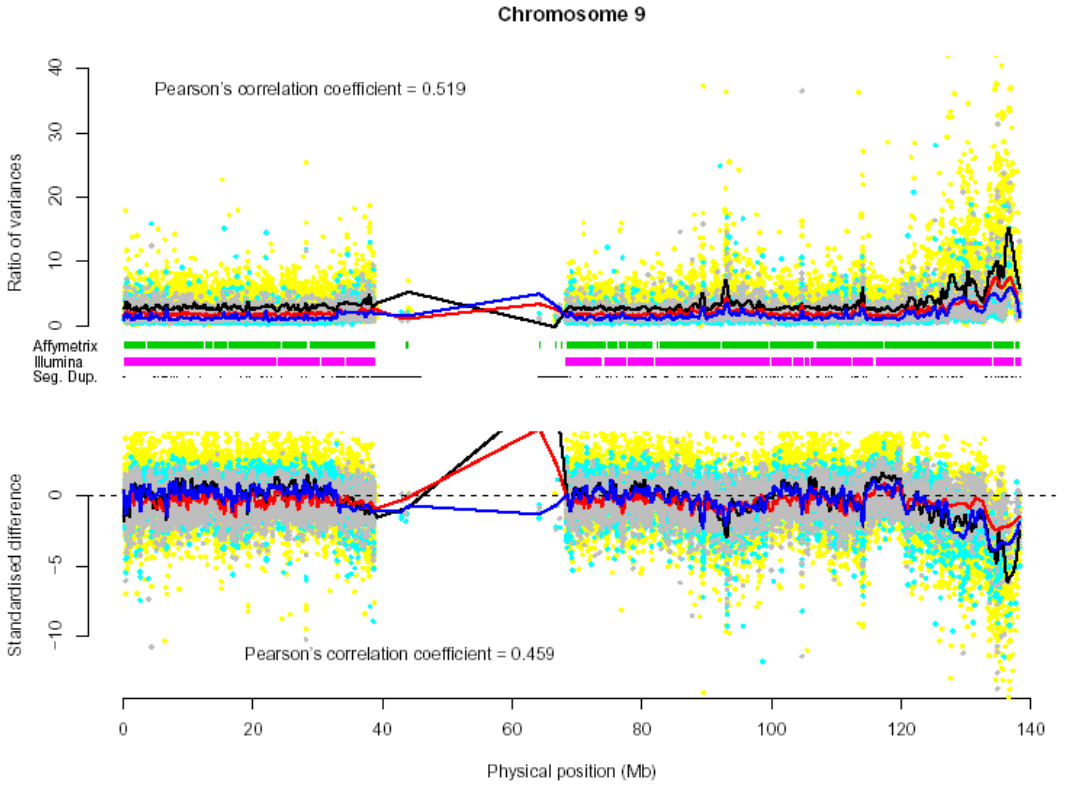

Chromosome 4
Pearson's correlation coefficient = 0.401
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.439
Physical position (Mb)

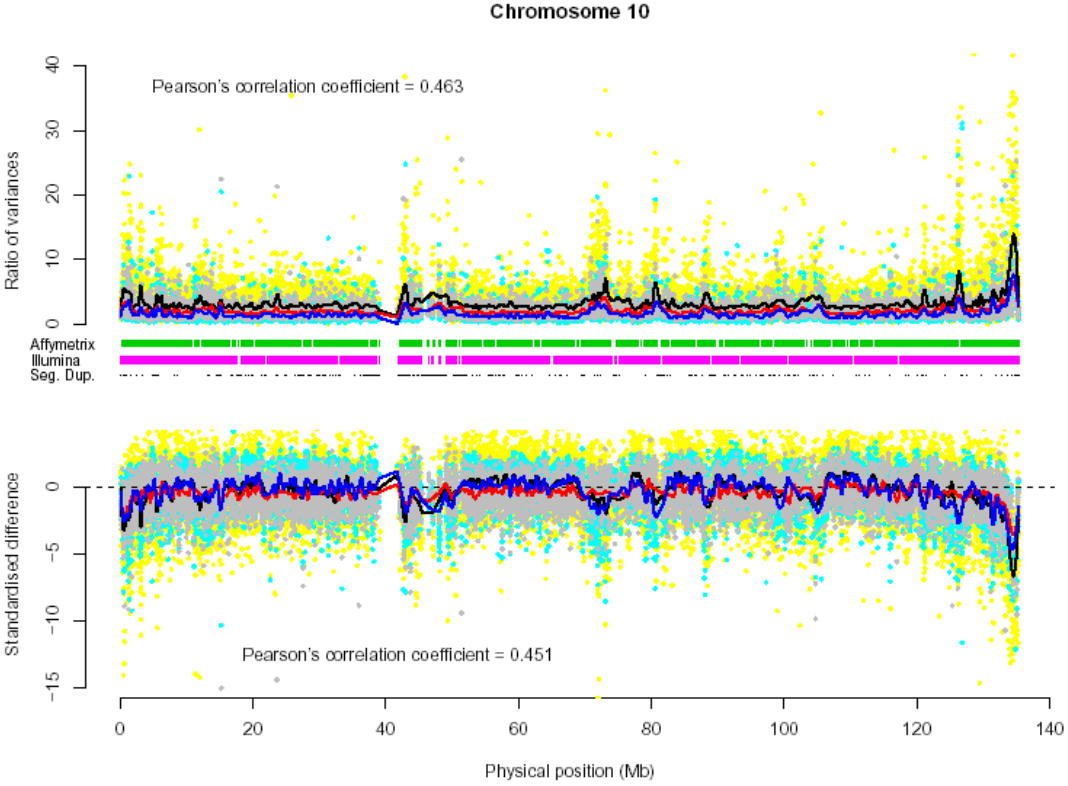

Chromosome 5
Pearson's correlation coefficient = 0.391
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.43
Physical position (Mb)

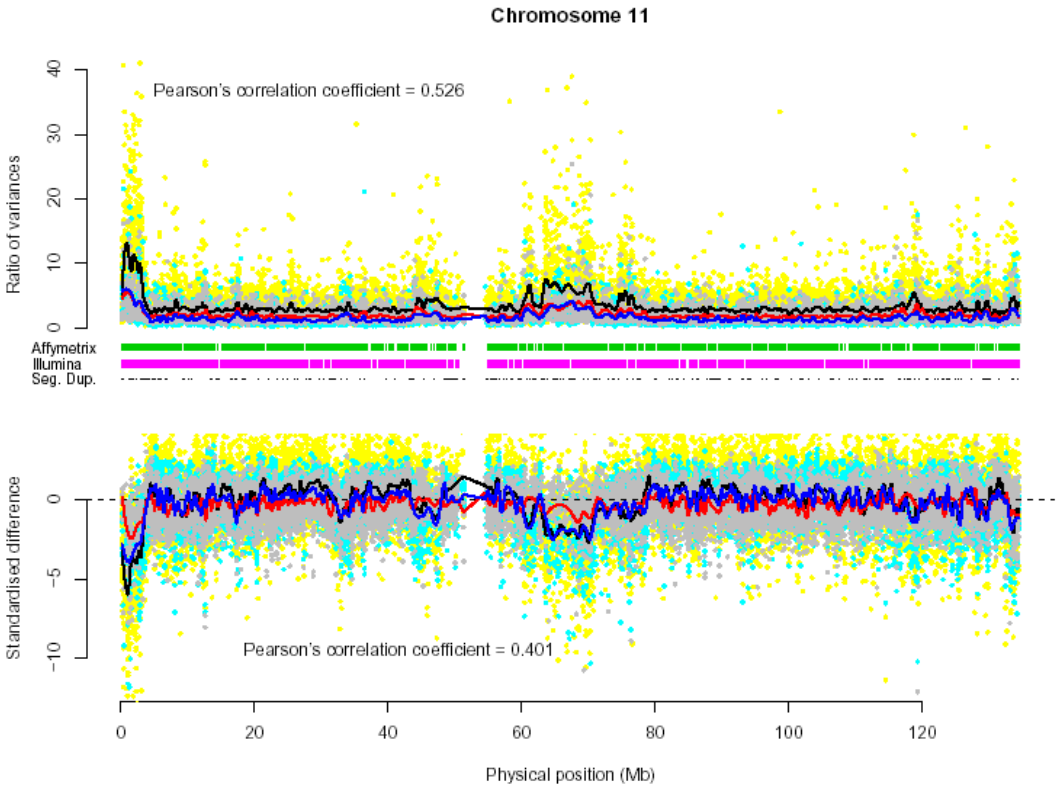

Chromosome 6
Pearson's correlation coefficient = 0.437
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.424
Physical position (Mb)

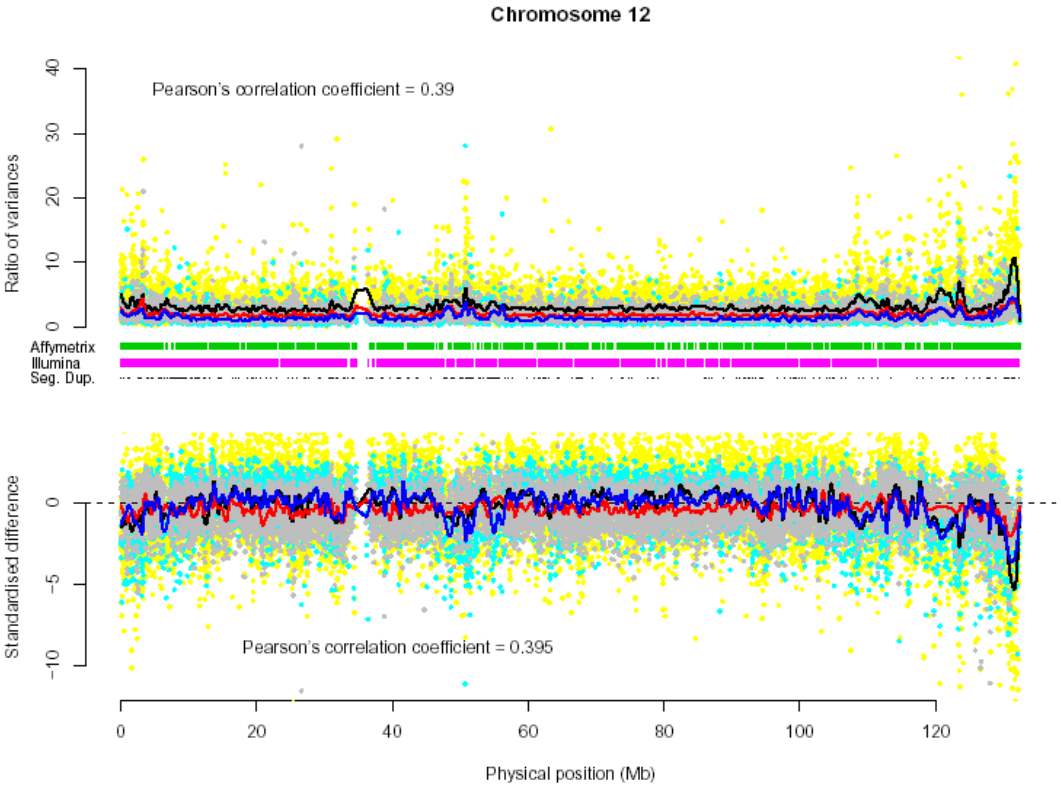

Chromosome 7
Pearson's correlation coefficient = 0.448
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.442
Physical position (Mb)

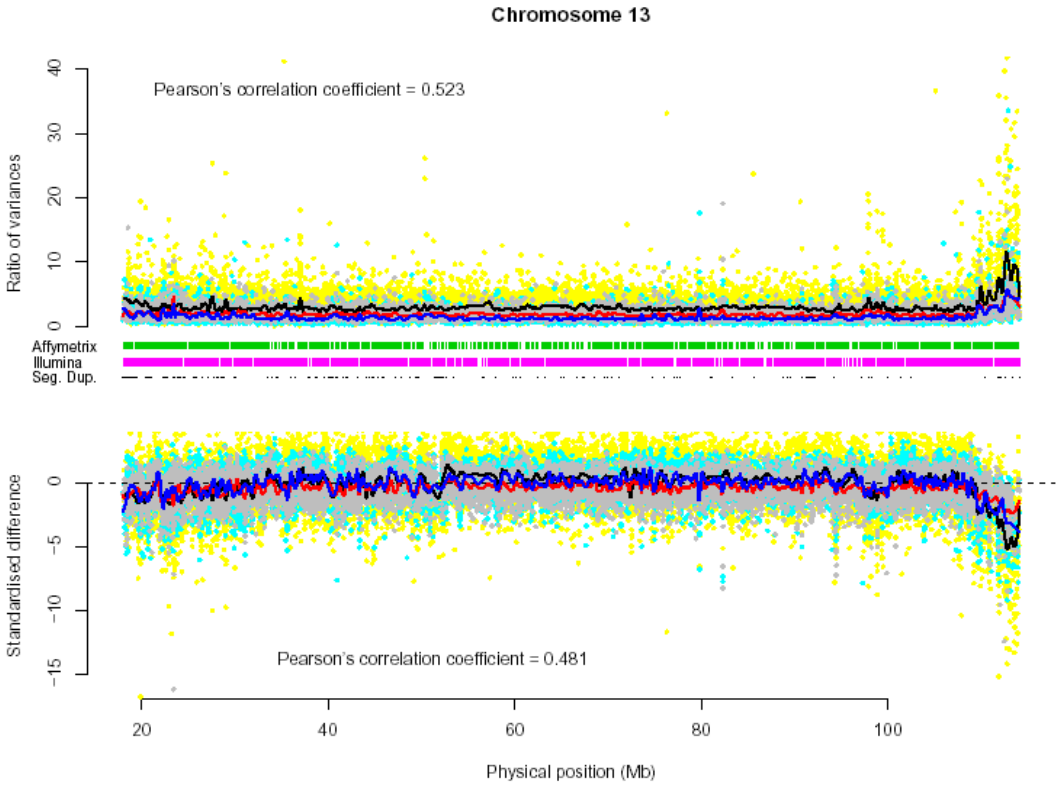

Chromosome 8
Pearson's correlation coefficient = 0.465
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.487
Physical position (Mb)

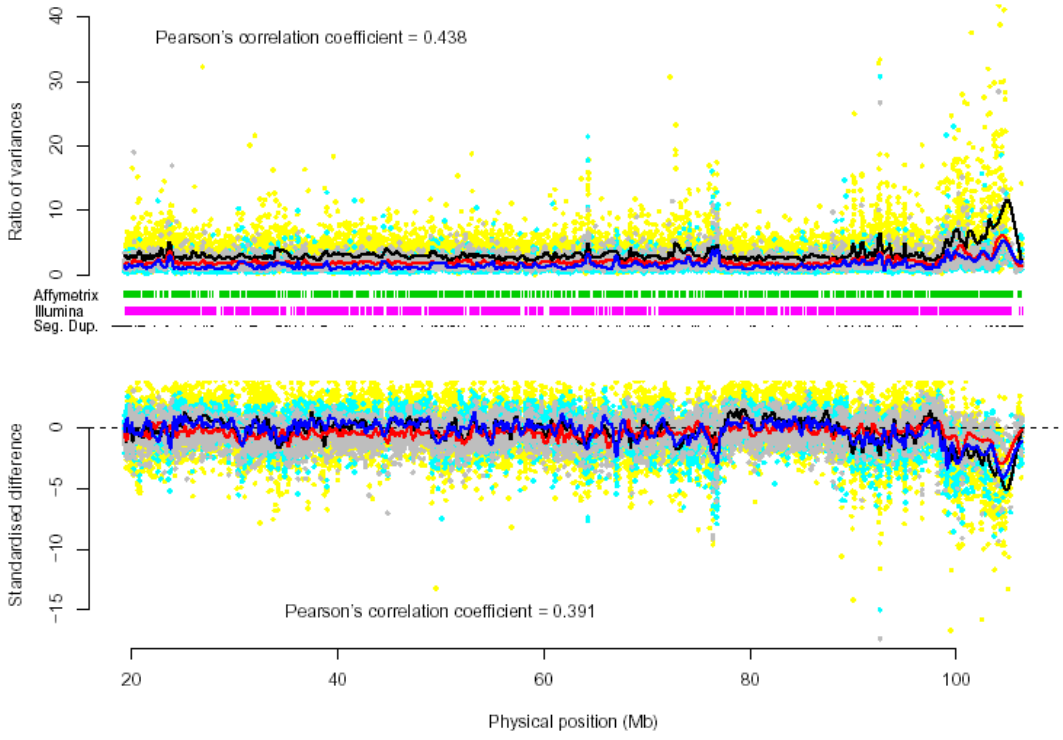

Chromosome 9
Pearson's correlation coefficient = 0.519
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.459
Physical position (Mb)

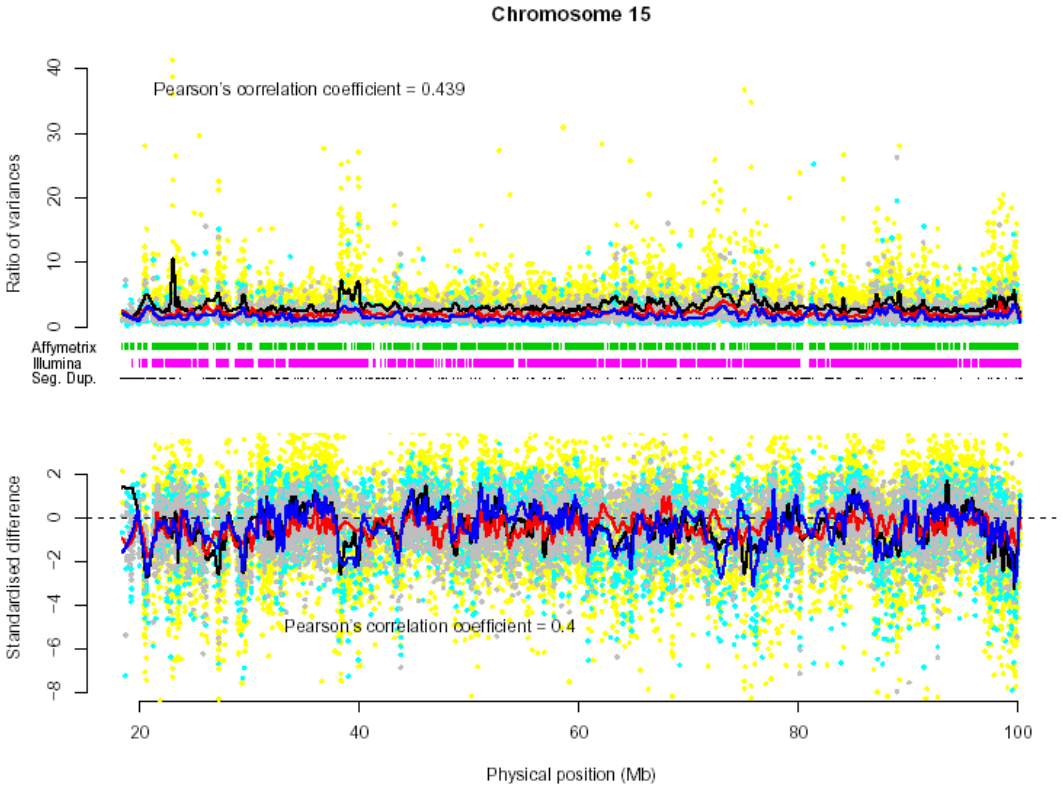

Chromosome 10
Pearson's correlation coefficient = 0.463
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.451
Physical position (Mb)

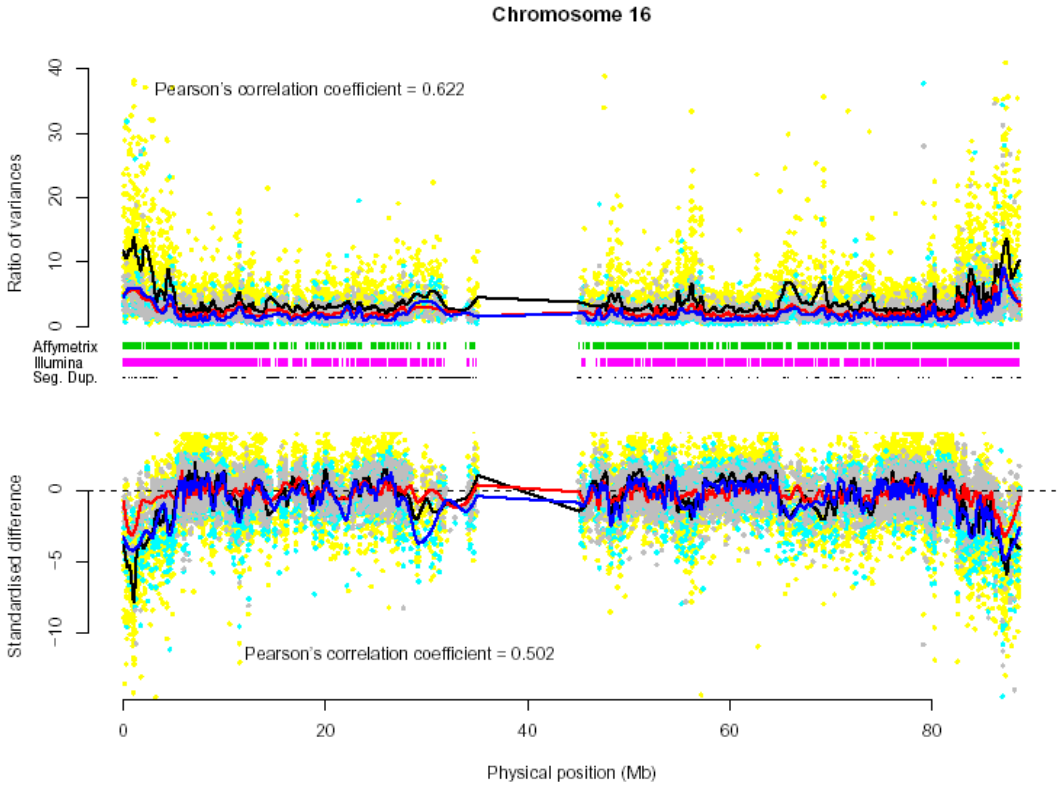

Chromosome 11
Pearson's correlation coefficient = 0.526
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.401
Physical position (Mb)

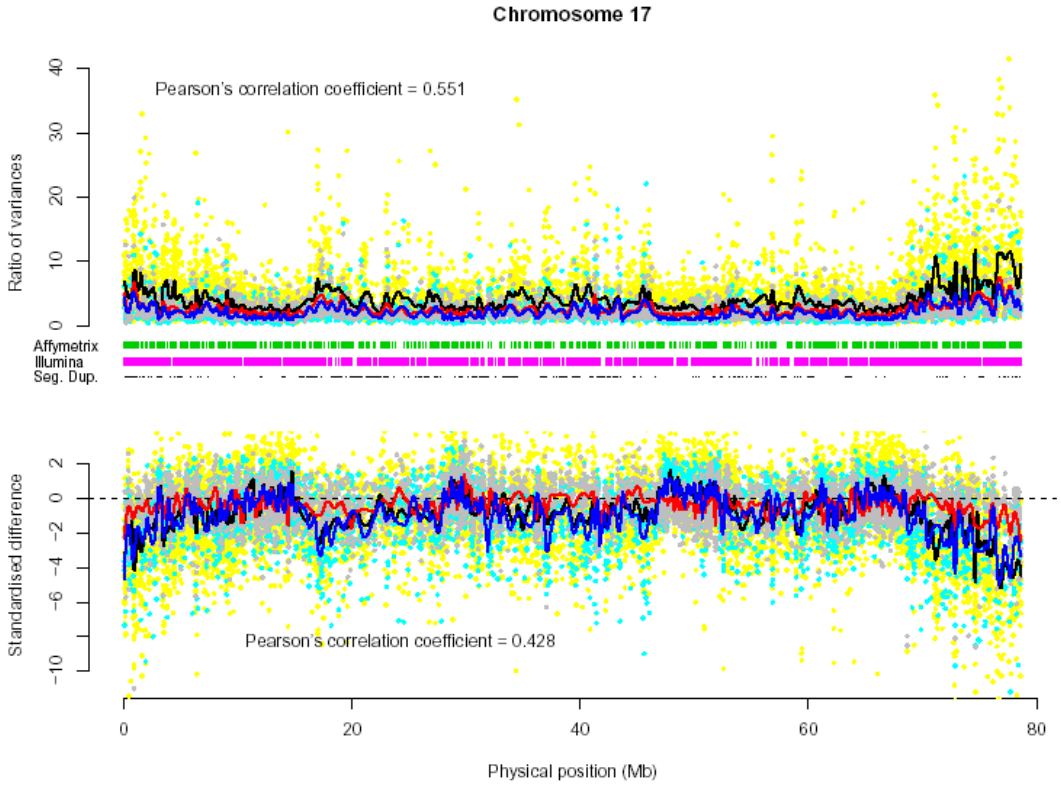

Chromosome 12
Pearson's correlation coefficient = 0.39
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.395
Physical position (Mb)

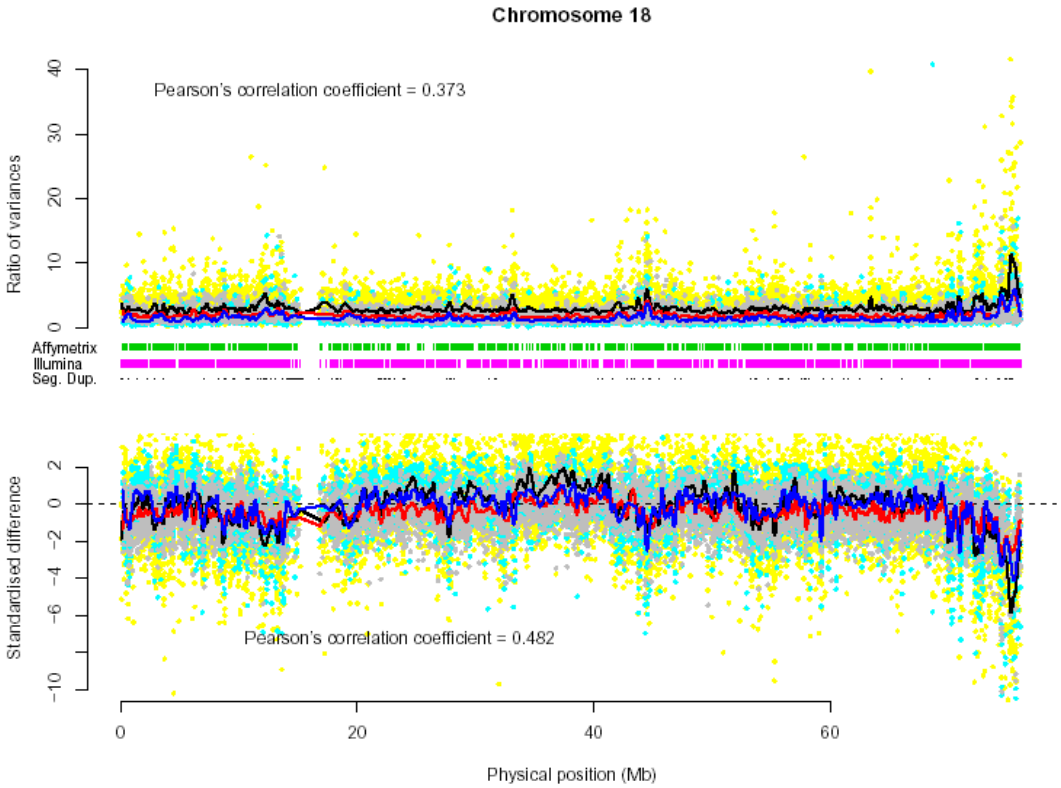

Chromosome 13
Pearson's correlation coefficient = 0.523
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
0
−5
−10
−15
Pearson's correlation coefficient = 0.481
20
40
60
80
100
Physical position (Mb)

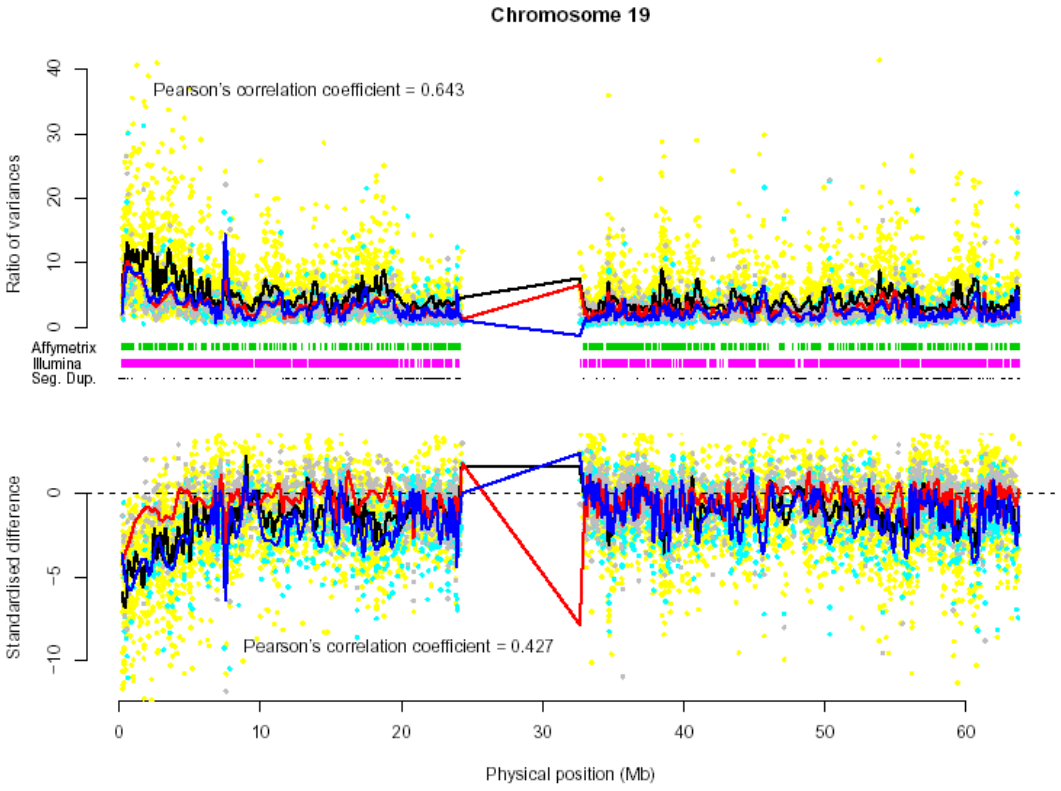

Chromosome 14
Pearson's correlation coefficient = 0.438
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
0
−5
−10
−15
Pearson's correlation coefficient = 0.391
20
40
60
80
100
Physical position (Mb)

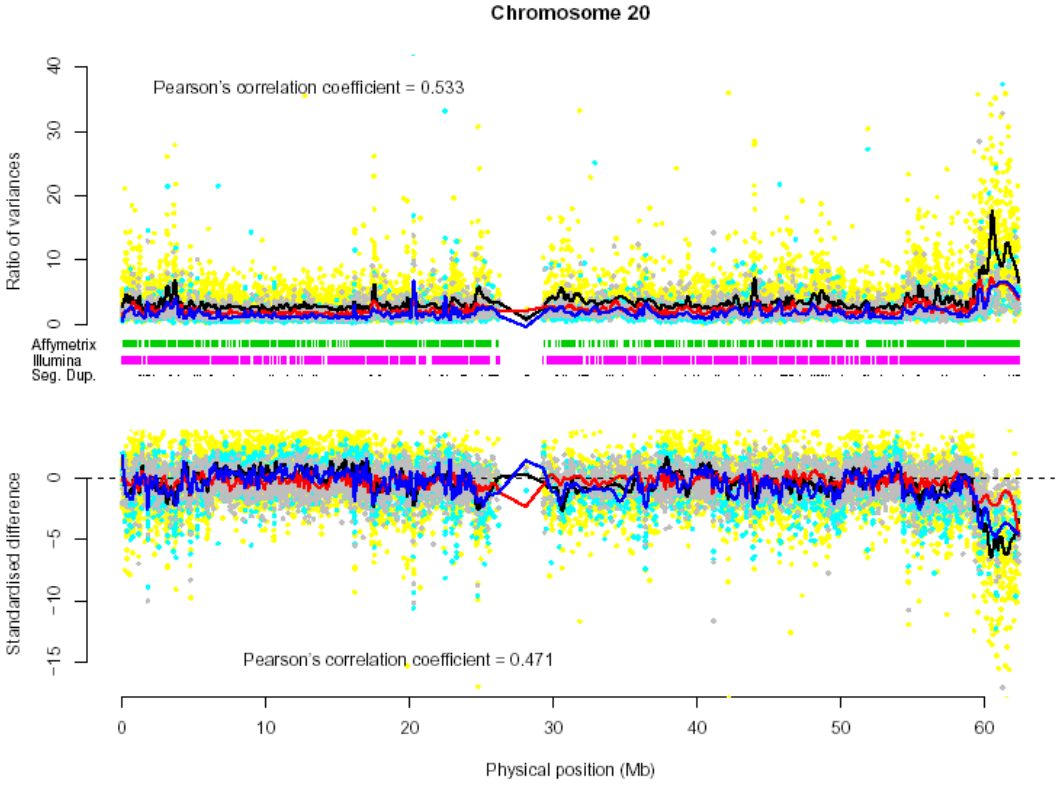

Chromosome 15
Pearson's correlation coefficient = 0.439
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.4
Physical position (Mb)

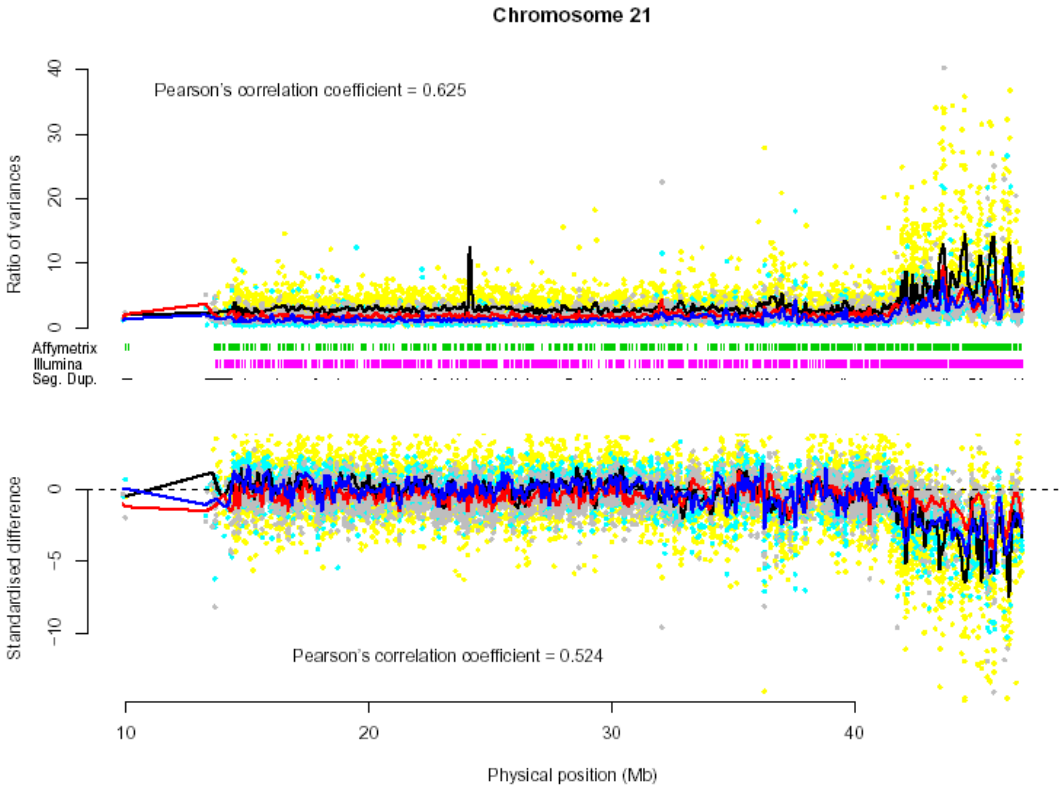

Chromosome 16
Pearson's correlation coefficient = 0.622
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.502
Physical position (Mb)

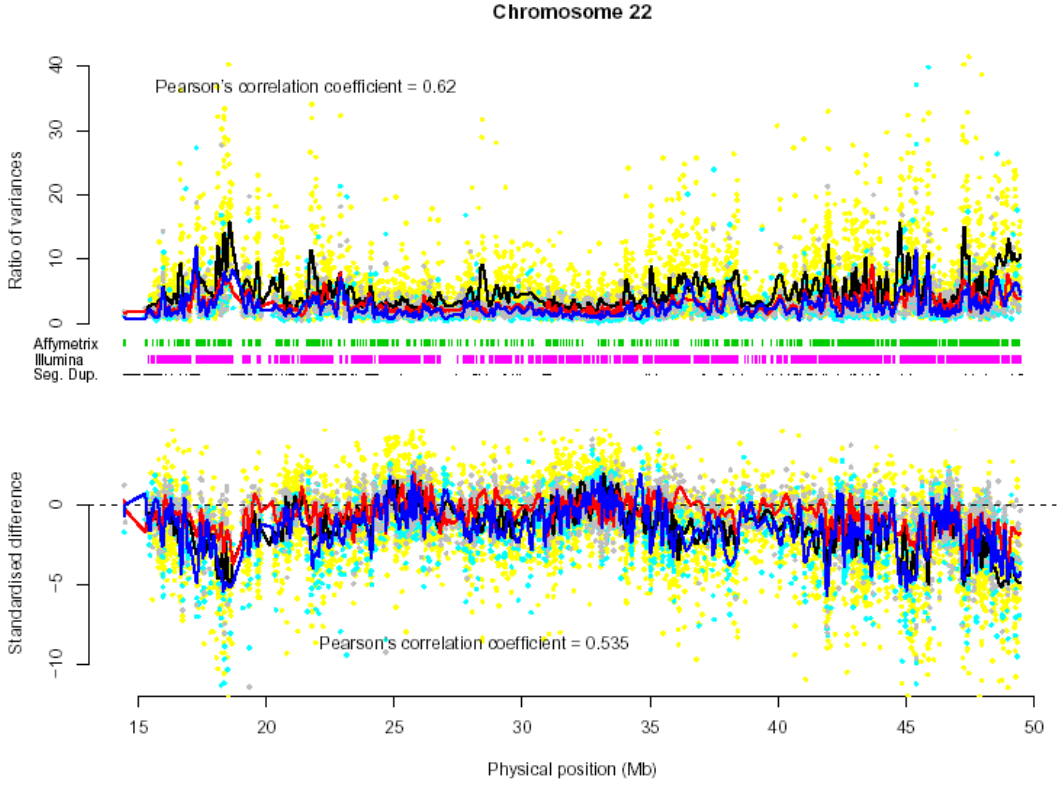

Chromosome 17
Pearson's correlation coefficient = 0.551
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.428
Physical position (Mb)

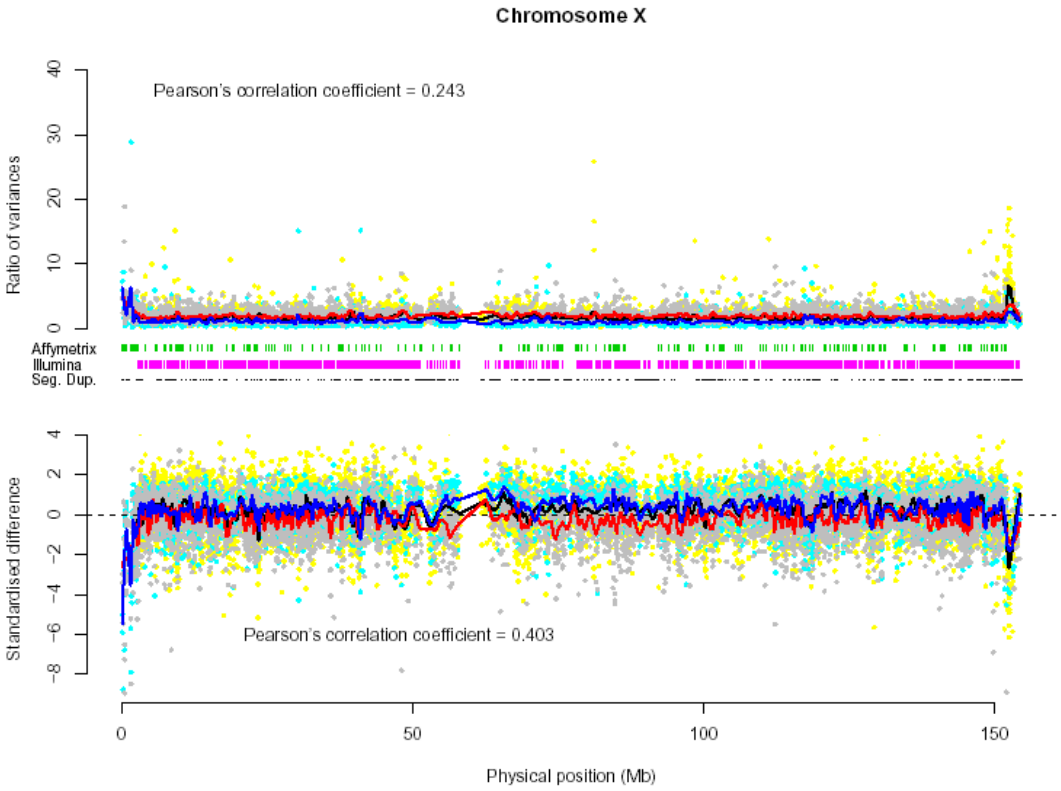

Chromosome 18
Pearson's correlation coefficient = 0.373
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.482
Physical position (Mb)


Chromosome 19
Pearson's correlation coefficient = 0.643
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.427
Physical position (Mb)

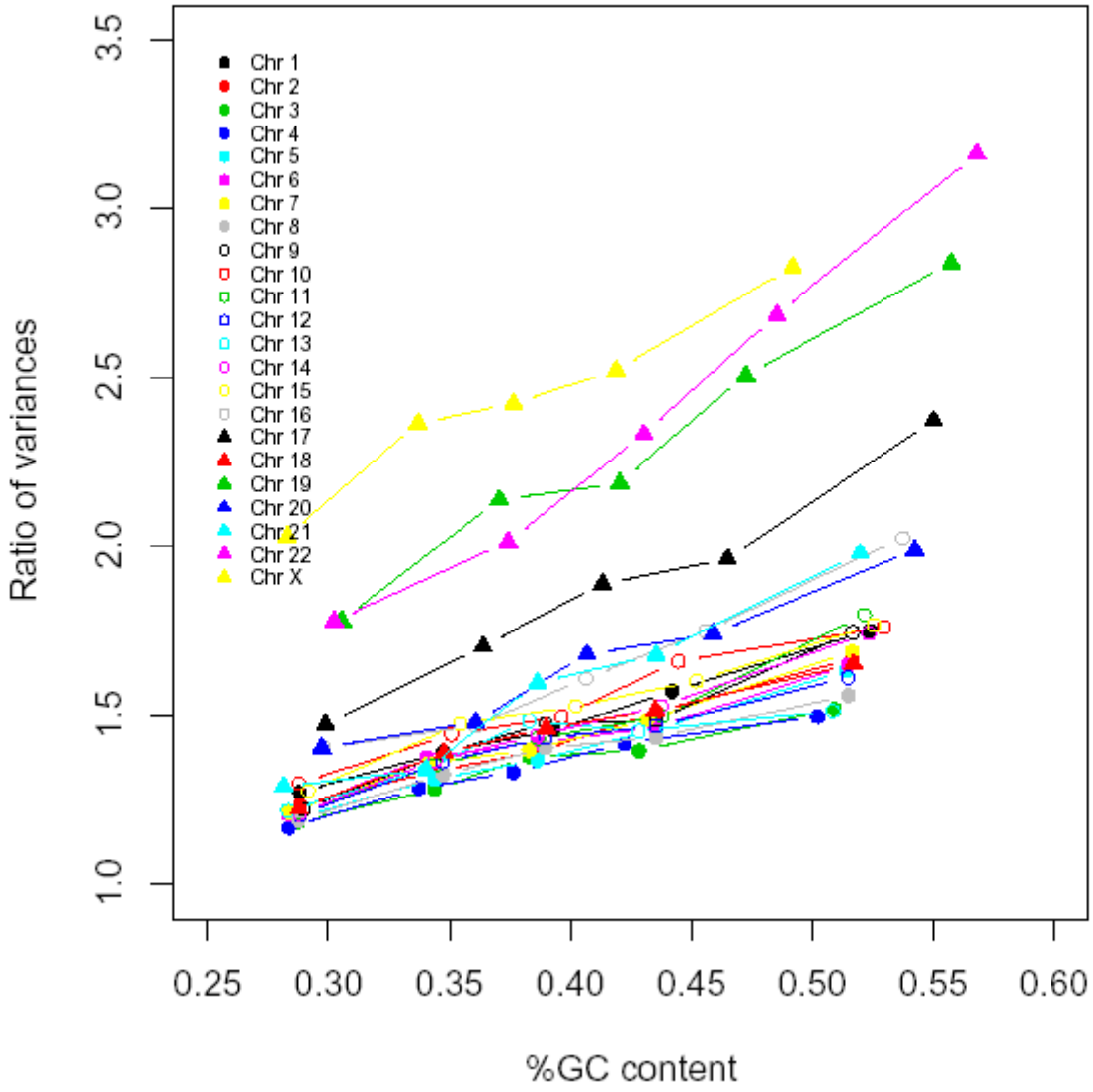

Chromosome 20
Pearson's correlation coefficient = 0.533
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.471
Physical position (Mb)


Chromosome 21
Pearson's correlation coefficient = 0.625
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.524
Physical position (Mb)

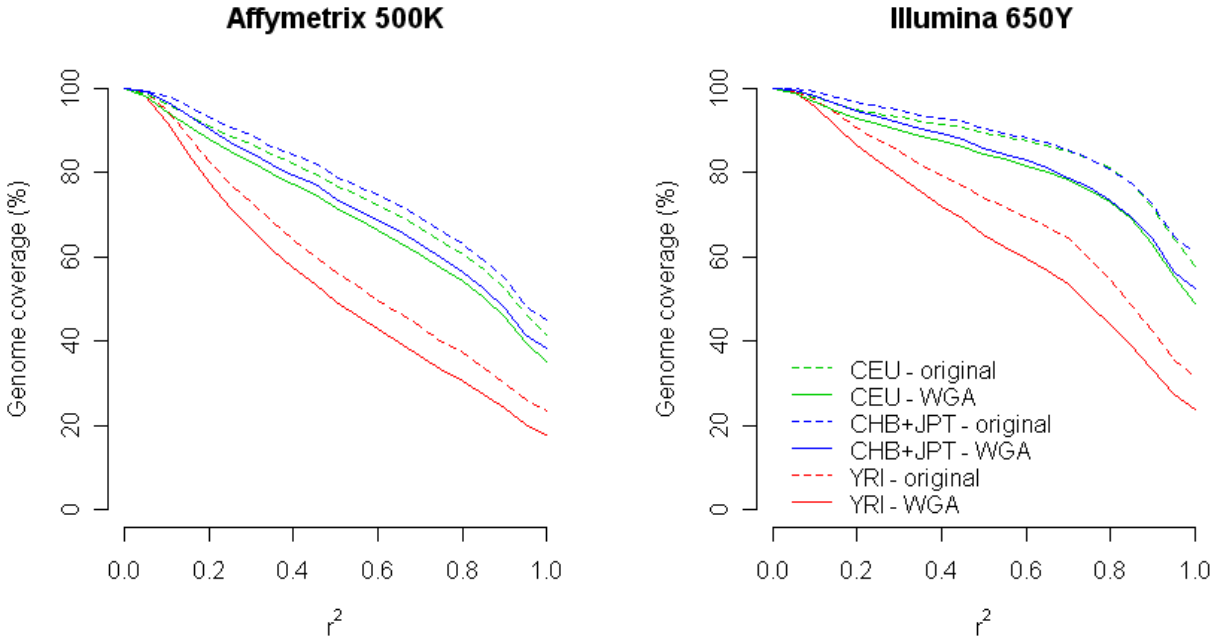

Chromosome 22
Pearson's correlation coefficient = 0.62
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.535
Physical position (Mb)


Chromosome X
Pearson's correlation coefficient = 0.243
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
−8
−6
−4
−2
0
2
4
Pearson's correlation coefficient = 0.403
0
50
100
150
Physical position (Mb)

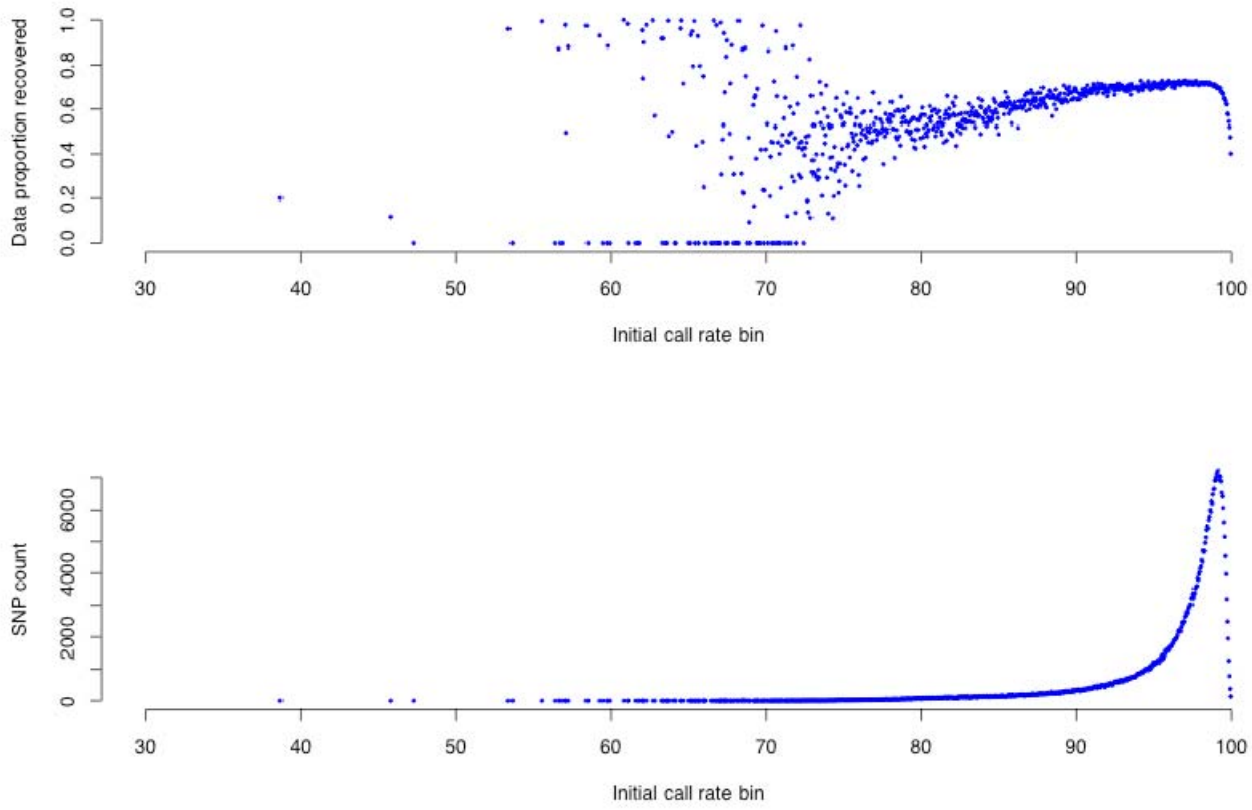
Supplementary Figure 3. Mean ratios of variances against mean % GC content of probe sequences for SNPs on the Affymetrix array. For each chromosome, the data has been divided into quintiles based on the GC content and the mean ratio of variances for each quintile is calculated.



Supplementary Figure 4. Coverage of the genome for the SNP arrays as a function of pairwise r^2. Dashed lines represent the original coverage for the genotyping platform as assessed by single marker tagging; solid lines represent the effective coverage for the genotyping platform after removing SNPs with more than 5.0% missing genotypes. Lines are represented for each of the three HapMap panels: CEPH individuals of European ancestry (green); Han Chinese from Beijing and Japanese individuals from Tokyo (blue); Yoruba people from Ibadan, Nigeria (red).



Supplementary Figure 5. SNP data recovery when missing genotypes are imputed as a function of the initial call rate. Data proportion recovered is calculated as (CRi – CRw) / (1 – CRw) where CRi denotes the SNP call rate after imputation and CRw the call rate before imputation. Initial SNP call rates have been partitioned into 0.01 bins with each bin's data proportion recovery averaged across the number of SNPs. As shown, variance in the data proportion recovered increases as the number of SNPs populating each call rate bin decreases.



Supplementary Methods

Data sets

Genotypic data were collected as part of four separate studies. DNA samples included 517 individuals from a Vietnamese study on tuberculosis (TB), 1,538 individuals from the 1958 British Birth Cohort which included all births from the United Kingdom during one week in 1958 (58C), 2,198 control individuals from the Norfolk area in the United Kingdom who have been recruited as part of an EPIC study on obesity (OBC), and 2,288 individuals from a Gambian study on malaria (ML). Please see **Supplementary Table 1** for details. The 58C Affymetrix data was obtained from the Wellcome Trust Case Control Consortium (WTCCC 2007). The 58C and OBC samples have been genotyped on both the *Nsp*I and *Sty*I arrays of the Affymetrix GeneChip 500K set while the TB samples have been genotyped only on the *Nsp*I array. 1,438 individuals from the 58C have also been genotyped on the Illumina HumanHap550 BeadChip array, of which 1,402 samples overlap with the 58C WTCCC dataset. All ML individuals have been genotyped on the Illumina HumanHap650Y BeadChip array, while 278 of these individuals have also been genotyped on the Affymetrix GeneChip.

Genotyping on the Affymetrix arrays took place in two separate genotyping facilities: the 58C and OBC samples were genotyped at the Affymetrix genotyping laboratories in San Francisco, while the TB samples were genotyped at the Genome Institute of Singapore (GIS). The genotyping of the Illumina arrays took place at the Wellcome Trust Sanger Institute (WTSI) in Hinxton, UK. In comparisons between the 3 different cohorts, only data from the *Nsp*I array is used for the Affymetrix experiment, corresponding to 262,264 SNPs. For the Illumina experiment, the set of SNPs which are common on both the HumanHap550 and HumanHap650Y arrays are extracted. This corresponds to 553,595 SNPs. Evaluation of the coverage for the Affymetrix platform uses data from the OBC and 58C cohorts, while the ML and 58C cohorts are used for the Illumina platform. As males have only one copy of chromosome X, the extent of hybridization on any SNP which is not on the pseudo-autosomal region of chromosome X is expected to be lowered, and thus all analyses involving chromosome X are performed using female samples only.

Laboratory protocol and DNA quality assessment

Samples sent to Affymetrix (58C, ML, and OBC cohorts) and samples run on the Illumina microarrays at the WTSI (58C and ML cohorts) followed the sample handling procedures outlined in the WTCCC (2007). Briefly, samples collections were requested at a DNA concentration of 100 ng/μl in deep 96-well plates, each with a unique barcode. Upon receipt, samples were assayed in triplicate by Picogreen, checked for degradation on a 0.75% agarose gel, and genotyped for up to 38 SNPs via the Sequenom MassExtend and/or iPLEX assay. The latter of which served to experimentally validate the provided gender and act as a molecular fingerprint through the genotyping pipeline. Samples with DNA concentrations greater than or equal to 50 ng/μl, showing limited or no degradation, >60% success rate for assayed Sequenom markers, and gender marker agreement were selected for genomic or WGA genotyping on genome-wide microarrays. Instead of pre-selection Sequenom typing, Taqman assays at two loci were performed on samples genotyped at the GIS.

All DNA collections that underwent whole genome amplification followed the procedure of multiple displacement amplification (MDA) with REPLI-g^{TM} 625S reagents based on instructions from the manufacturer (MSI Inc, New Haven). The ML and OBC cohorts were amplified at Geneservice Ltd (Cambridge, UK) while the TB collection was amplified at the GIS. All WGA DNA was then re-assessed with Picogreen, normalized to 250 ng/μl, and run on 0.75% agarose gels to filter those which experienced degradation post-amplification.

Genotyping

Affymetrix genotyping was performed using the GeneChip 500K at the Affymetrix Services Lab as outlined by the WTCCC (2007). Briefly, each plate was processed together, and each sample was digested in two aliquots of 250 ng, by the *Nsp*I and *Sty*I enzymes respectively; this is followed by ligation of an adaptor, fragmentation, and labeling (Matsuzaki et al. 2004). Each enzyme preparation is then hybridized to its corresponding SNP array (262,000 and 238,000 SNPs for the *Nsp*I and *Sty*I respectively). Samples were then called with the Affymetrix Dynamic Model algorithm (DM, Di et al. 2005) and repeated if failing a 93.0% call rate threshold (at an individual genotype score cutoff of 0.33). Successful completion and delivery of samples entailed a DM call rate >93.0%, with >90.0% concordance for 50 SNPs common to both the *Nsp*I and *Sty*I arrays, and >70.0% identity to their WTCCC Sequenom genotypes. Genotyping at the GIS was performed using the same Affymetrix protocol and was initiated only on the *Nsp*I array.

Illumina genotyping using both the BeadArray 550K and 650K SNP microarrays was performed as per the Illumina Infinium II system (Gunderson, et al. 2006). This system uses single base extension biochemistry once the input DNA is initially whole genome amplified, fragmented, denatured, and then hybridized to the microarray. The process is automated using a Tecan GenePaint system while workflow and sample tracking are handled by the Laboratory Information Management System (LIMS). To identify repeats, samples were loaded and initially called within the Illumina BeadStudio software using the automated, proprietary GenCall algorithm; DNAs which exhibited a call rate <94.0% (at a GC score cutoff of 0.20) were queued for repeat. These samples were sorted by call rate, the lowest performing of which were re-genotyped until it was financially impractical to continue. Duplicate samples were then filtered by the criteria of highest call rate.

Data pre-processing

The raw data output from the Affymetrix genotyping consist of measures of probe hybridization intensities. For each individual at each SNP, there are either 6 or 10 probe quartets. Each probe quartet consists of four probe cells which assay for a perfect match or a mismatch to a specific 25-base oligonucleotide sequence for each of the two possible alleles (generically denoted *A* and *B*). These hybridization intensities need to undergo a pre-processing phase to combine the information across the probe quartets to yield a pair of coordinates corresponding to the signal strength for each of the two possible alleles. Our initial pre-processing phase is similar to that adopted by the WTCCC (2007). Briefly, quantile normalization against a reference intensity distribution is applied to all the data to minimize chip-to-chip variability and the logarithms are taken to reduce skewness. To minimize the variation of the signals due to the different cohorts, the reference distribution is obtained from the Affymetrix data on the 269 HapMap individuals. Suppose $Y_{il} = (Y_{il}^{(PA)}, Y_{il}^{(MA)}, Y_{il}^{(PB)}, Y_{il}^{(MB)})$ denote the vector of log-normalized intensities for probe quartet *i* on SNP *l* for an individual, we make the following transformations:

$$Y_{il}^{(A)} = Y_{il}^{(PA)} - \frac{\left(Y_{il}^{(MA)} + Y_{il}^{(MB)}\right)}{2} \text{ if } Y_{il}^{(PA)} \geq Y_{il}^{(MA)} \text{ and zero otherwise;}$$

$$Y_{il}^{(B)} = Y_{il}^{(PB)} - \frac{\left(Y_{il}^{(MA)} + Y_{il}^{(MB)}\right)}{2} \text{ if } Y_{il}^{(PB)} \geq Y_{il}^{(MB)} \text{ and zero otherwise.}$$

These quartet-specific signals for the alleles are pooled across all the probe quartets to yield a pair of signal coordinates corresponding to the two alleles: $\left(s_l^{(A)}, s_l^{(B)}\right) = \left(\frac{1}{n_l}\sum_{i=1}^{n_l} Y_{il}^{(A)}, \frac{1}{n_l}\sum_{i=1}^{n_l} Y_{il}^{(B)}\right)$, where $n_l \in \{6, 10\}$. We refer to $s_l^{(A)}$ and $s_l^{(B)}$ as the signals for alleles *A* and *B* respectively at SNP *l*. We further define a corresponding measure of signal strength as the logarithm of the sum of the signals, excluding any individuals at the particular SNP where the signals yield a non-positive sum. In addition, we define the contrast as $\sinh^{-1}\left(\frac{s_l^{(A)} - s_l^{(B)}}{s_l^{(A)} + s_l^{(B)}}\right)$. The strength and the contrast can be respectively interpreted as the equivalent of *r* and θ in polar coordinates for the allelic signals $\left(s_l^{(A)}, s_l^{(B)}\right)$. For the Illumina genotyping, the raw fluorescence intensities are self-normalized by the BeadStudio software which performs a 6-degree of freedom affine transformation (Peiffer et al. 2006) to yield pairs of signals which are directly equivalent to the allelic signals $\left(s_l^{(A)}, s_l^{(B)}\right)$. We perform identical transformations of the allelic signals as the Affymetrix data to obtain the signal strengths and contrasts.

Quantification of differential hybridization

Two measures are used to quantify the difference in performance of ϕ29MDA DNA and genomic DNA:

$$\textit{Ratio of strength variance} = \frac{Var(Strength)_{MDA}}{Var(Strength)_{genomic}}$$

$$\textit{Standardized difference of mean strength} = \frac{Mean(Strength)_{MDA} - Mean(Strength)_{genomic}}{\sqrt{Var(Strength)_{genomic}}}.$$

The ratio of strength variance measures the increase in variability of the strength of ϕ29MDA DNA to the strength of genomic DNA, whereas the standardized difference of mean strength effectively assesses the extent of the change in the hybridization signal for ϕ29MDA DNA compared to genomic DNA. The standard deviation of the strength for genomic DNA is used rather than the classical measurement of standard deviation for the means of two independent samples since the interest here is primarily in the change in the strengths of whole genome amplified DNA as compared to genomic DNA.

Automated genotyping and identifying underperforming SNPs

The definition of underperforming SNPs depends on the criterion used. In this paper, we have chosen the extent of missing genotypes for each SNP as a measure of poor performance. Our observation that the amount of missing data is an effective surrogate for poor performing SNPs in an association study is consistent with similar assessments made by groups conducting genome-wide association studies (WTCCC 2007, Rioux et al. 2007, Gudmundsson et al. 2007, Yeager et al. 2007, Saxena et al. 2007, Scott et al. 2007), and SNPs with call-rates <95.0% are often discarded from further statistical analyses. The call-rates for SNPs depend on the stringency of the threshold used in the automated genotype assignment procedure, where there is a trade-off between the fidelity of the

genotype assignments and call-rates. For the samples genotyped on the Affymetrix 500K Array set, the genotypes are called using the BRLMM algorithm (Affymetrix 2006). We used the Affymetrix recommended threshold of 0.50 for the ratio of the Mahalanobis distance between the two most likely genotype clusters to assign a call and samples with a Mahalanobis distance ratio of greater than 0.50 are assigned a NULL genotype. Genotypes for samples assayed on the Illumina platforms are assigned using GenCall – a proprietary calling algorithm designed by Illumina for their BeadStudio software. A GC score filter of 0.2 is used to threshold the confidence score associated with each assigned genotype, and a NULL genotype is assigned if the confidence score is less than 0.2. To control for chip and laboratory failures, only runs with > 70.0% call rate are considered for further analysis.

Inter-platform genotyping accuracy

As a quality control step to confirm the accuracy of the Illumina and Affymetrix genotyping platforms for genomic and WGA DNA, we performed a test of concordance on all shared-platform samples from the 58C and ML cohorts. Across a panel of 84,496 SNPs which are common to all of the Affymetrix 500K, Illumina 550K, and Illumina 650K microarrays, we assessed 1402 58C samples and 278 ML samples which have been typed on the Affymetrix 500K and either the Illumina 550K or the 650K microarrays. When disregarding any comparison which contained a null call on either platform, we observed a genotype concordance of 99.6% for the genomic 58C cohort at call rates of 99.6% and 98.7% for Illumina and Affymetrix, respectively. For the whole genome amplified ML cohort, we observe a genotype concordance of 98.1% at call rates of 93.8% and 97.0%.

Assessing GC content

For SNPs on the Affymetrix array, every probe cell within a quartet assay a unique 25-base sequence, and the four probes differ at a single interrogation base. To increase the reliability of the hybridization, multiple probe quartets with the interrogation position placed in different locations of the sequence are used. Each probe quartet may have a different interrogation position by shifting the sequence up or down stream of the SNP site, referred to as the different degree of offset (of -4 to +4 bases from the position of the SNP). We define the GC content of a probe as the percentage of G and C bases in the 25-base sequence. As we average over the probe quartets to obtain the allelic signals at each SNP, we similarly define the GC content of a SNP as the average GC content across all the probe quartets (please see **Figure S2**). For the Illumina array, each probe is a unique 50-mer sequence immediately adjacent to the SNP, and the GC content for each SNP is similarly defined as the percentage of G and C bases on the probe.

Evaluating coverage

Coverage is assessed by single marker tagging: a SNP is considered to be tagged if the pairwise r^2 between this SNP and another is greater than some predetermined threshold. In line with established criteria for evaluating coverage (International HapMap Consortium 2007), the maximum allowed physical distance between the tag SNP and a SNP in interrogation is 200kb, and only SNPs with minor allele frequency > 5.0% are considered. We adopt an established measure (Barrett & Cardon 2006) of genome-wide coverage as

$$\frac{\left(\frac{L}{R-T}\right)(G-T)+T}{G},$$

with *T* denoting the size of the tag set, *R* as the size of the reference set (from the HapMap), *L* as the number of SNPs in LD with a SNP in the tag set, and *G* as the number of common SNPs in the genome which is estimated at 7.5 million for individuals with European ancestries. We have used the same *G* for the analysis as the resultant coverage calculations are fairly robust to the choice of *G*.

Imputation

Genotype imputation is performed on all the ML and OBC samples using the program IMPUTE (Marchini 2007). HapMap Phase II autosomal haplotypes for the YRI and CEU have been used to perform the imputation. IMPUTE estimates the posterior probabilities of the three valid genotypes for each individual at each SNP. We assigned the genotype call with the maximum posterior probability if this probability is at least 0.90, and assigned a missing (NULL) genotype if the maximum posterior probability is < 0.90. Only SNPs with less than 5.0% missing data for the imputed genotypes are included in the analysis. Data recovery was calculated as

$$\frac{CRi - CRw}{1 - CRw}$$

with CRi denoting the call rate of the SNP after imputation, CRw as the call rate without imputation.

Statistical analysis

A paired sample t-test is used to compare the per-SNP call rates between genomic DNA and ϕ29MDA DNA. Local polynomial regressions are fitted using the loess function in R, with the degree of smoothing fixed by specifying the span to be 0.01. To investigate the trend of the effect that the GC content of the Affymetrix probes has on the ratio of variances, the Affymetrix data for each chromosome is divided by the quintiles of GC content and the mean GC content for each quintile is calculated. Pearson's correlation coefficient is used to quantify the correlations of: (i) the ratios of strength variances; (ii) the standardized differences of mean strengths, between the TB and the ML-58C Affymetrix data. The statistical significance for the Pearson's correlation coefficient ρ calculated from *L* SNPs is approximated from the Student's *t*-distribution with $L - 2$ degrees of freedom and a test statistic of

$$\frac{\rho}{\sqrt{(1-\rho)^2/(n-2)}}.$$

Supplementary References

Affymetrix Inc. BRLMM: an improved genotype calling method for the GeneChip Human Mapping 500K array set. http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf (2006).

Barrett, J.C. and Cardon, L.R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**, 659–662 (2006).

Di, X. *et al.* Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics* **21(9):** 1958-1963 (2005).

Gudmundsson, J. *et al*. Genome-wide association study identifies a second prostate cancer susceptibility variantat 8q24. *Nat. Genetics* **39**, 631–637 (2007).

Gunderson, K.L., Kuhn, K.M., Steemers, F.J., Ng, P., Murray, S.S., Shen, R. Whole-genome genotyping of haplotype tag single nucleotide polymorphisms. *Pharmacogenomics* **7**(4): 641-648 (2006).

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851 – 861 (2007).

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).

Matsuzaki, H. *et al*. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* **1:** 104-105 (2004).

Peiffer, D.A. *et al*. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* **16:** 1136–1148 (2006).

Rioux, J.D. *et al*. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy n disease pathogenesis. *Nat. Genetics* **39**, 596–604 (2007).

Saxena, R. *et al*. Genome-wide association analysis identifies loci for Type 2 diabetes and triglyceride levels. *Science* [epub Apr 26 2007].

Scott, L.J. *et al*. A genome-wide association study of Type 2 diabetes in Finns detects multiple susceptibility variants. *Science* [epub Apr 26 2007].

Teo, Y.Y. *et al*. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* [epub Sep 10] (2007).

The Wellcome Trust Case Control Consortium. Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

Yeager, M. *et al*. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genetics* **39**, 645–649 (2007).

ACKNOWLEDGEMENTS
We thank D Strachan and WL McArdle for samples from the British 1958 Birth Cohort collection, funded by the UK Medical Research Grant G0000934 and the Wellcome Trust grant 068545/Z/02. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113. We acknowledge A Green, R Gwilliam, TT Nguyen, E Png, A Richardson and all members of the DNA/Genotyping Facilities and System Support teams at the Sanger Institute for assistance in generating and handling genotype data respectively. The work of YYT, MI, KSS, AEF, SCP, IB, KAR, DPK and PD is supported by the Wellcome Trust. In addition, YYT, KSS, AEF, KAR and DPK also acknowledge support from the Grand Challenges in Global Health, and the UK Medical Research Council. MS acknowledges support from the Agency for Science Technology and Research Singapore.

AUTHOR CONTRIBUTIONS
SJD, MS, IB, NJW, KAR provided DNA samples; SCP coordinated the data processing; YYT, KSS and MI performed the analyses; YYT and MI designed the project and wrote the paper, with contributions from KSS and AEF; DPK and PD jointly directed the project.