Algorithms and analysis of human disease genomics

Inouye, M.

3

WHOLE GENOME-AMPLIFIED DNA: INSIGHTS AND IMPUTATION

Yik Y. Teo[1,2,7], Michael Inouye[2,7], Kerrin S. Small[1,2], Andrew E. Fry[1], Simon C. Potter[2], Sarah J. Dunstan[3], Mark Seielstad[4], Ines Barroso[5], Nicholas J. Wareham[6], Kirk A Rockett[1], Dominic P. Kwiatkowski[1,2], and Panos Deloukas[2]

[1] Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, United Kingdom

[2] Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom

[3] Oxford University Clinical Research Unit, Hospital for Tropical Diseases, Ho Chi, Minh City, Vietnam

[4] Genome Institute of Singapore, Agency for Science, Technology and Research, 60 Biopolis Street, Singapore

[5] Metabolic Disease Group, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, United Kingdom

[6] MRC Epidemiology Unit, Strangeways Research Laboratories, Worts Causeway, Cambridge CB1 8RN, United Kingdom

[7] These authors contributed equally

Nature Methods. 2008 Apr;5(4):279-80

Whole genome–amplified DNA: insights and imputation

To the editor: Genome-wide association studies (GWAS) have enabled a considerable portion of the human genome to be scanned for genetic variants associated with disease etiology. Such large-scale investigations depend on DNA samples of high biological integrity and quality. As clinical DNA is often available in limited quantities, *in vitro* reproduction of quality template DNA using whole-genome amplification is necessary. The most widely used technique is multiple displacement amplification with ϕ29 polymerase[1] (ϕ29MDA). Earlier studies do not offer a detailed map of how robust genome-wide panels of 300,000 to 1 million single-nucleotide polymorphisms (SNPs) will perform with ϕ29MDA[2–3]. We performed a meta-analysis of 6,541 DNA samples to assess the extent of information lost (**Supplementary Table 1** online), and investigated genotype imputation for recovering the statistical power and genomic coverage of GWAS (**Supplementary Methods** online).

As previously seen with array-CGH[3], we observed that ϕ29MDA led to differential rates of hybridization compared to genomic DNA, especially in telomeric regions, on both Affymetrix and Illumina arrays (**Fig. 1a** and **Supplementary Fig. 1** online). This correlated to the G+C content of the SNP oligonucleotide probes (**Supplementary Fig. 2** online) and to the presence of segmental duplications (**Supplementary Table 2** online). In the context of a GWAS, this results in a proportion of SNPs with lower signal strength and increased variability. often resulting in overlapping or heavily scattered genotype clusters for both the allelic signal and strength-contrast scales (**Supplementary Fig. 3** online). This increases the uncertainty when assigning genotypes and lowers call rates. The average call rates for SNPs on the Affymetrix array with ϕ29MDA-amplified (Norfolk European Prospective Investigation of Cancer (EPIC); OBC) and genomic (1958 British Birth Cohort; 58C) DNA samples (written informed consent was obtained from all subjects), were 96.3% and 98.7%, respectively. The corresponding call rates for the Illumina arrays for a ϕ29MDA-amplified (ML) and genomic (58C) DNA samples were 95.9% and 98.5%, respectively. Furthermore, missing data were distributed nonrandomly across the genome resulting in considerable decreases in genomic coverage when we applied GWAS SNP quality control. By excluding SNPs with call rates < 95.0%, the Affymetrix 500K array experienced a 6.5% drop in coverage from 60.6% to 54.1% when measured on the HapMap population of European ancestry (CEU) at a correlation threshold of 0.8. The Illumina 650Y, a chip with high tag-SNP content, had an 8.1% decrease (81.3% to 73.2%; **Supplementary Table 3** and **Supplementary Fig. 4** online).

Most low call rate SNPs contain samples with signal intensities found in overlapping genotype clusters (**Supplementary Fig. 1**). Although the use of custom calling algorithms can potentially alleviate this[4], genotype imputation provides a highly promising solution for analyzing regions with sufficient linkage disequilibrium by statistically inferring missing genotypes with high accuracy[5] (**Supplementary Fig. 5** online). Expanding this strategy to a genome-wide scale, we imputed the regions of data loss for the OBC cohort and assessed genome coverage and data recovery. Using a probability threshold of 0.90, imputation of all missing genotypes for OBC samples recovered an additional 2.4% of the original ϕ29MDA dataset, giving an overall call rate of 98.7% across all SNPs. This is comparable to the performance for the 58C. Typically, one can expect to recover 60% of a SNP's missing genotypes if the initial call rate is >75.0% (**Fig. 1b**). Imputation rescued 328 (14.9%) samples and 80,613 (16.7%) SNPs. The recovery of SNPs increased genome coverage for the Affymetrix 500K (measured by the HapMap CEU population at a pairwise $r^2 > 0.8$) from 54.1% to 59.7% (with benchmark coverage at 60.6%), while the recovered samples allowed greater power in a GWAS.

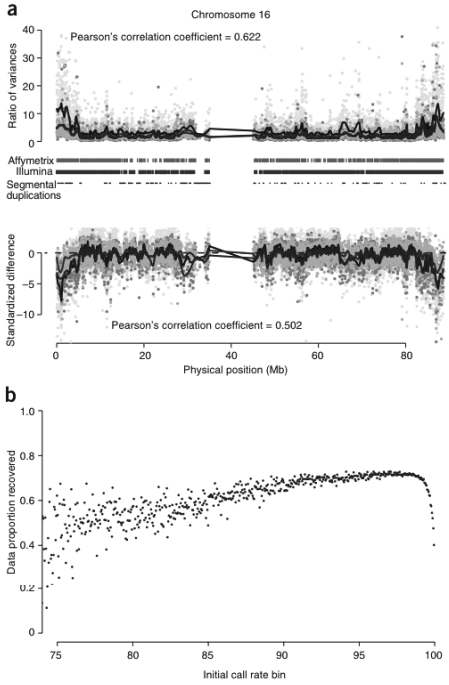


Figure 1 | Missingness and imputation of amplified DNA on chromosome 16. (**a**) The relative performance of amplified DNA to genomic DNA, as quantified by the ratio of hybridization strengths and the standardized difference in mean hybridization strength. Each plot shows data from three separate pairs of comparisons between amplified and genomic DNA. Affymetrix: TB (cyan dots and blue lines); Affymetrix: OBC-58C (gray dots and red lines); Illumina: ML-58C (yellow dots and black lines). The line for each comparison pair is obtained by a local polynomial regression fit to the observed 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. (**b**) The expected proportion of missing genotypes which can be recovered using the program IMPUTE[5] as a function of the initial call rate. Initial SNP call rates have been partitioned into 0.01 bins with each bin's data proportion recovery averaged across the number of SNPs.

CORRESPONDENCE

Although variation in linkage disequilibrium between SNPs across different populations and the SNP content of genotyping arrays may affect the performance of imputation, the statistical inference of missing genotypes presents a powerful solution for genetic studies constrained by severely limited quantities of DNA.

Note: Supplementary information is available on the Nature Methods website.

Yik Y Teo[1,2,7], Michael Inouye[2,7], Kerrin S Small[1,2], Andrew E Fry[1], Simon C Potter[2], Sarah J Dunstan[3], Mark Seielstad[4], Inês Barroso[5], Nicholas J Wareham[6], Kirk A Rockett[1], Dominic P Kwiatkowski[1,2] & Panos Deloukas[2]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [2]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [3]Oxford University Clinical Research Unit, Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam. [4]Genome Institute of Singapore, Agency for Science, Technology and Research, 60 Biopolis Street, 138672 Singapore. [5]MRC Epidemiology Unit, Strangeways Research Laboratories, Worts Causeway, Cambridge CB1 8RN, UK. [6]Metabolic Disease Group, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [7]These authors contributed equally to this work.
e-mail: panos@sanger.ac.uk

1. Dean, F.B. *et al. Proc. Natl. Acad. Sci.* **99**, 5261–5266 (2002).
2. Paez, J.G. *et al. Nucleic Acids Res.* **32**, e71 (2004).
3. Lage, J.M. *et al. Genome Res.* **13**, 294–307 (2003).
4. Teo, Y.Y. *et al. Bioinformatics* **23**, 2741–2746 (2007).
5. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. *Nat. Genet.* **39**, 906–913 (2007).

Supplementary Material

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 **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 ϕ29 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.

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.

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.

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

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

Table ST1. Analysis of SNPs within regions of segmental duplications.

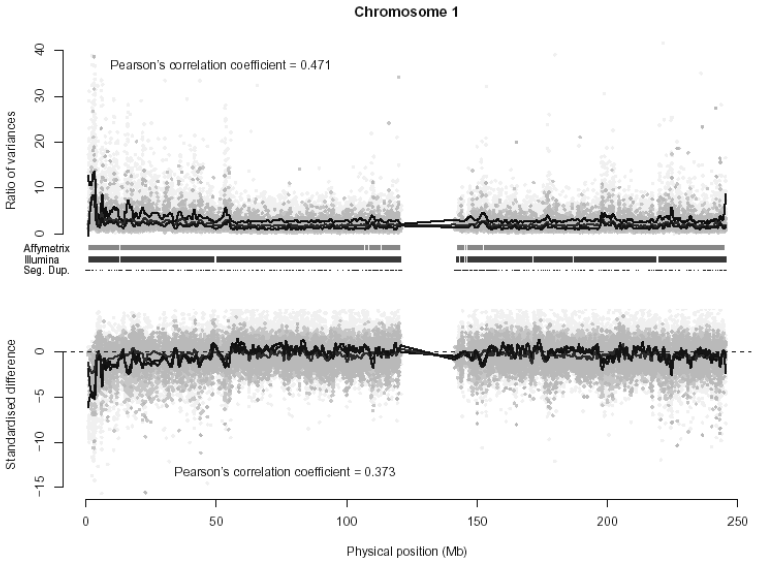
Platform	# SNPs in seg. dup.	Genomic DNA		WGA DNA	
		Call rate (%)	# SNPs with < 95% call-rate	Call rate (%)	# SNPs with < 95% call-rate
Affymetrix 500K	7,710	97.8%	1,072 (13.9%)	95.0%	2,756 (35.7%)
Illumina*	5,440	97.6%	461 (8.5%)	92.2%	1,811 (33.3%)

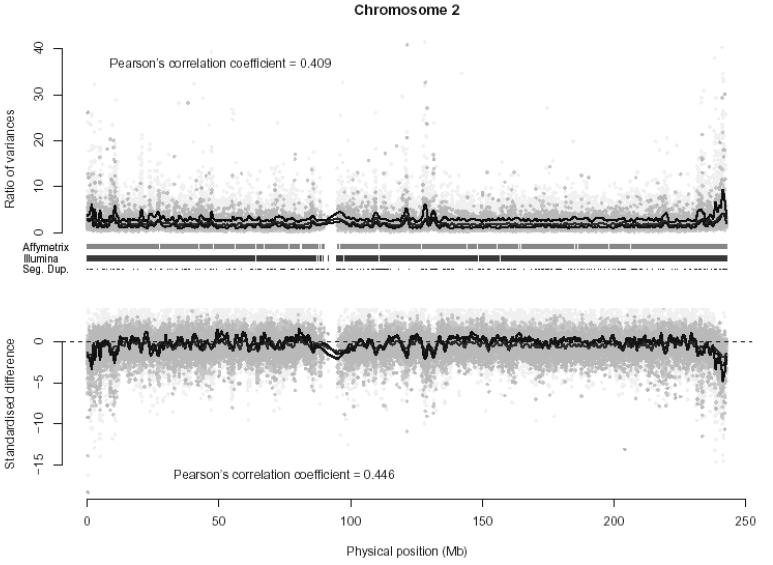
* SNPs common to both the HumanHap550 and HumanHap650Y arrays.

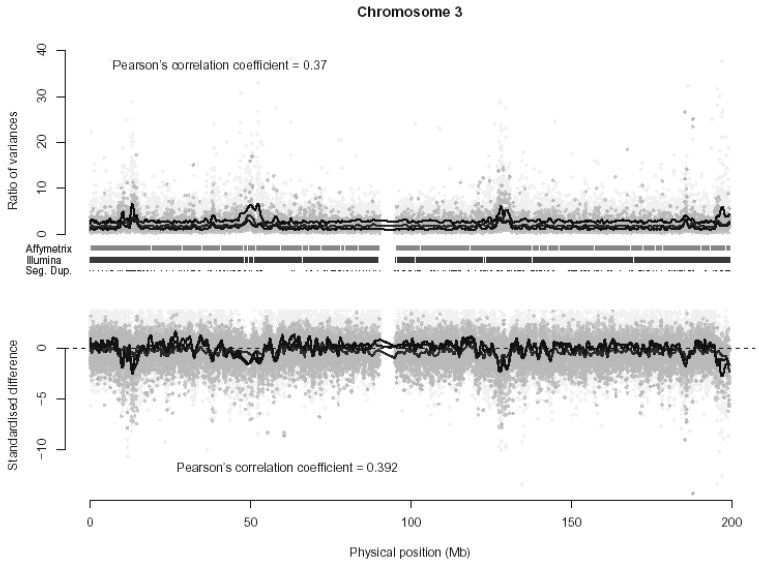
Table ST2. We extend the coverage calculations for different rates of missingness beyond the manuscript's adopted call rate threshold of 0.95 (less than 5.0% missingness for each SNP). The resultant coverage is calculated at call rate thresholds of 0.90 and 0.97. Genome coverage is calculated at a pairwise 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.

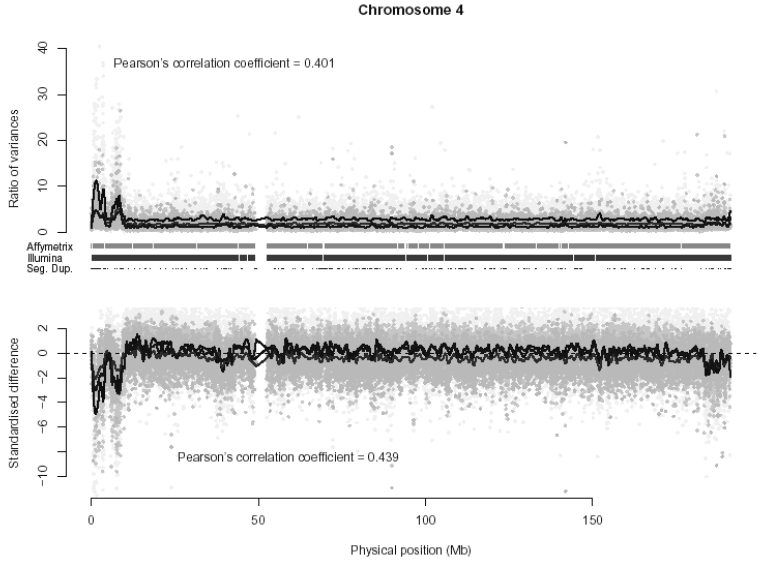
	Benchmark	**Call Rate Threshold**		
	Coverage	**0.90**	**0.95**	**0.97**
CEU				
Affy 500K	60.6	58.5 (-2.1)	54.1 (-6.5)	47.9 (-12.8)
Illumina 650K (GenCall)	81.3	76.8 (-4.5)	73.2 (-8.1)	68.3 (-13.0)
Illumina 650K (Illuminus)	81.3	81.0 (-0.3)	79.9 (-1.2)	76.0 (-5.1)
CHB + JPT				
Affy 500K	63.0	60.8 (-2.1)	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)

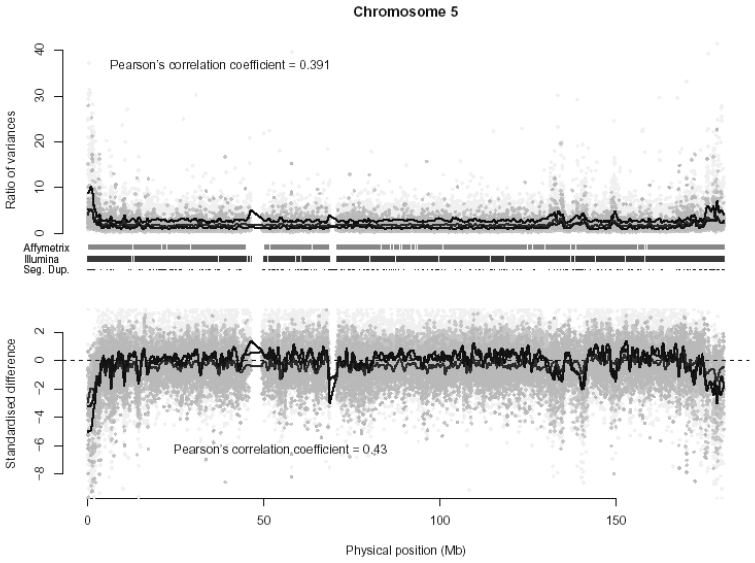
Figure S1. The relative performance of amplified DNA to genomic DNA, as quantified by two measures: (i) the ratio of hybridization strengths; (ii) the standardized difference in mean hybridization strength. Each plot shows data from three comparisons of amplified DNA to genomic DNA – Affymetrix: TB (cyan 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.

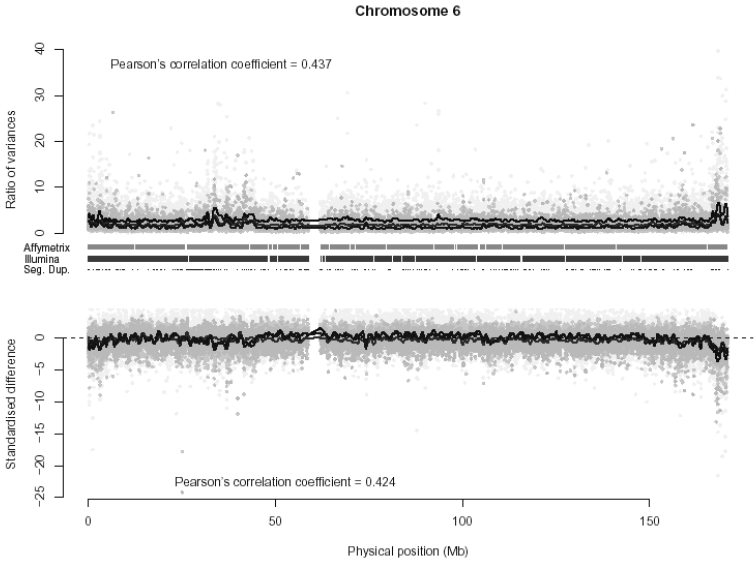

Chromosome 1
Pearson's correlation coefficient = 0.471
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.373
Physical position (Mb)

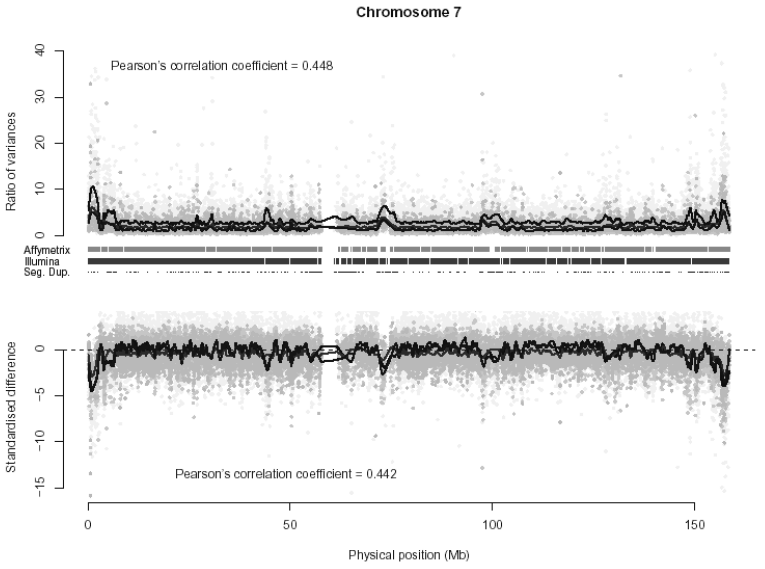

Chromosome 2
Pearson's correlation coefficient = 0.409
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.446
Physical position (Mb)

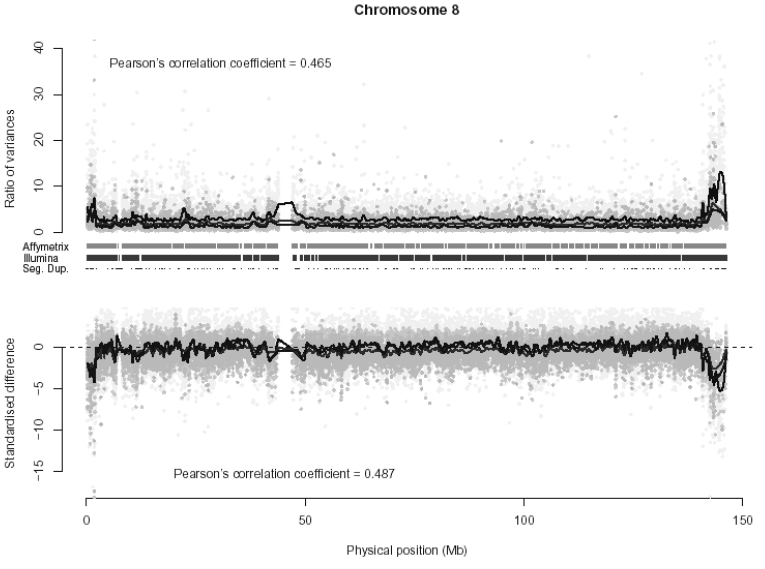

Chromosome 3
Pearson's correlation coefficient = 0.37
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.392
Physical position (Mb)

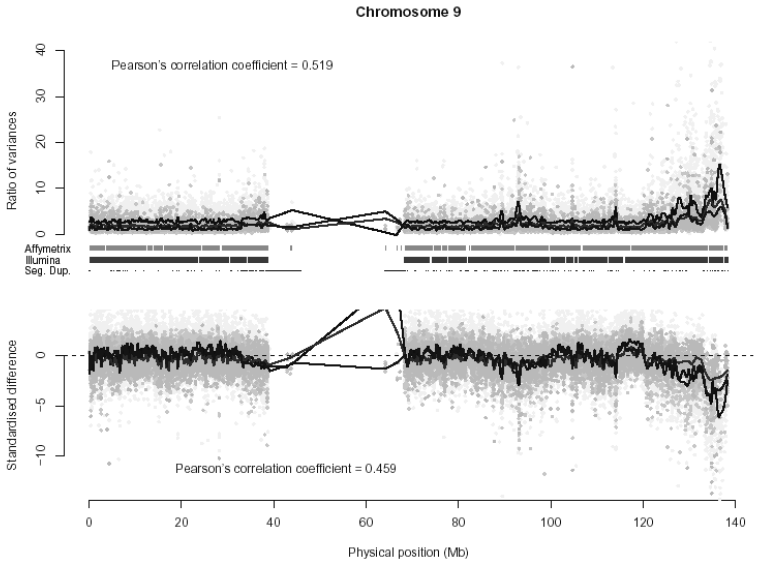

Chromosome 4
Pearson's correlation coefficient = 0.401
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.439
Physical position (Mb)

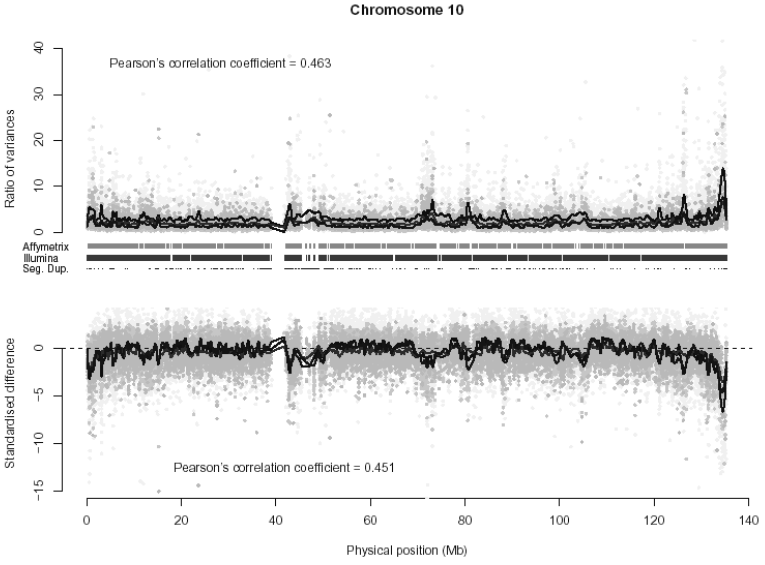

Chromosome 5
Pearson's correlation coefficient = 0.391
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.43
Physical position (Mb)

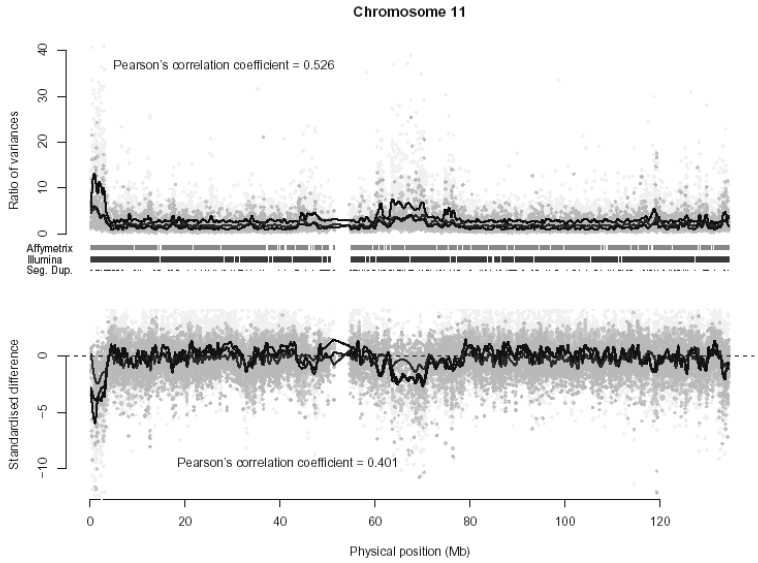

Chromosome 6
Pearson's correlation coefficient = 0.437
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.424
Physical position (Mb)

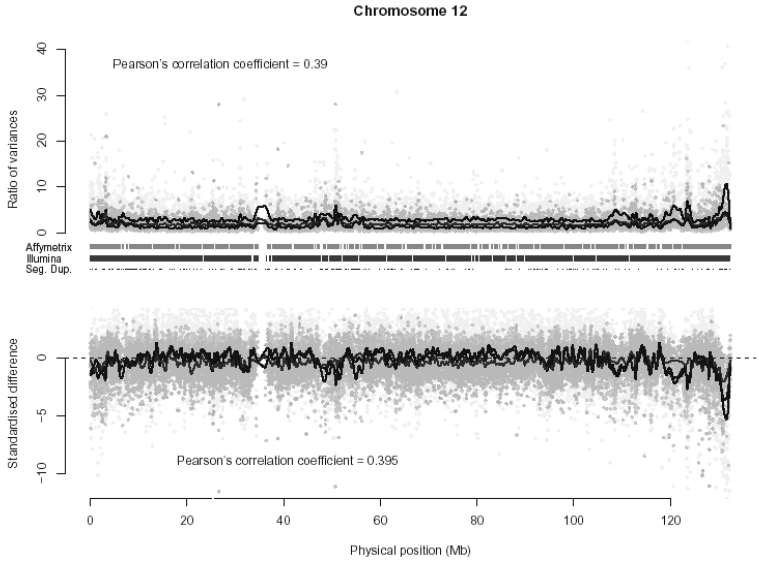

Chromosome 7
Pearson's correlation coefficient = 0.448
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
0
−5
−10
−15
Pearson's correlation coefficient = 0.442
0
50
100
150
Physical position (Mb)

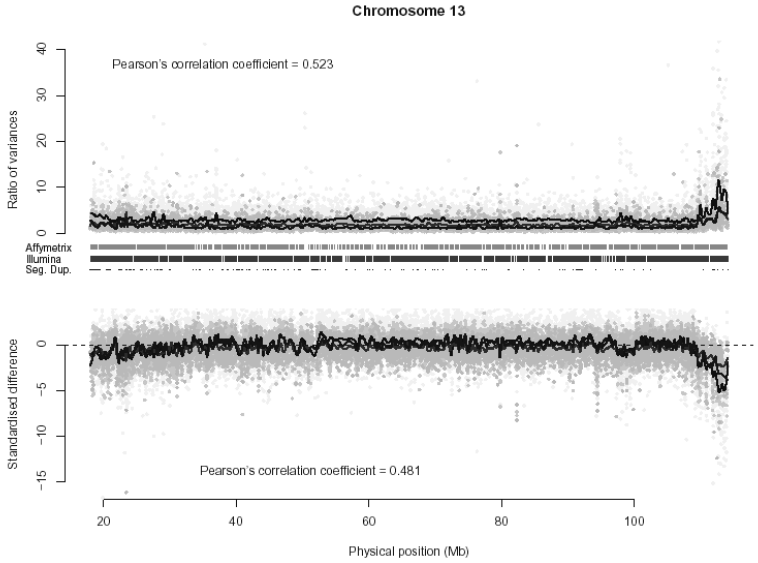

Chromosome 8
Pearson's correlation coefficient = 0.465
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
0
−5
−10
−15
Pearson's correlation coefficient = 0.487
0
50
100
150
Physical position (Mb)

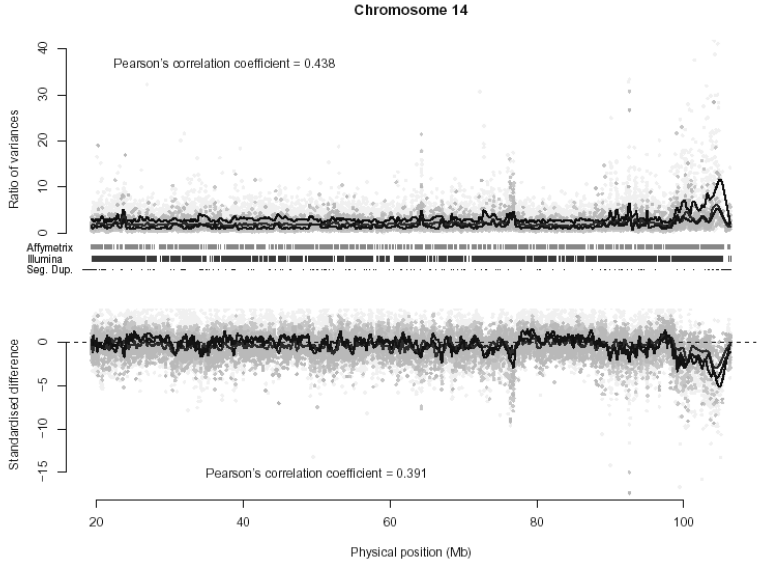

Chromosome 9
Pearson's correlation coefficient = 0.519
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
−10
−5
0
Pearson's correlation coefficient = 0.459
0
20
40
60
80
100
120
140
Physical position (Mb)

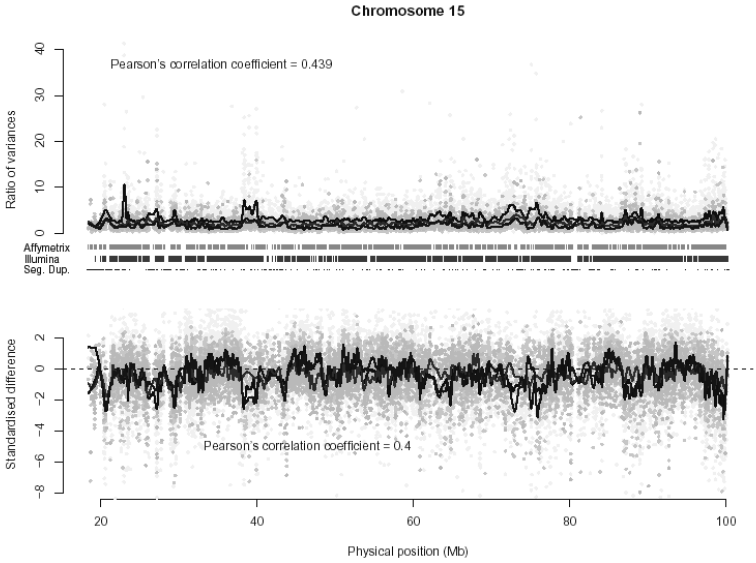

Chromosome 10
Pearson's correlation coefficient = 0.463
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
−15
−10
−5
0
Pearson's correlation coefficient = 0.451
0
20
40
60
80
100
120
140
Physical position (Mb)

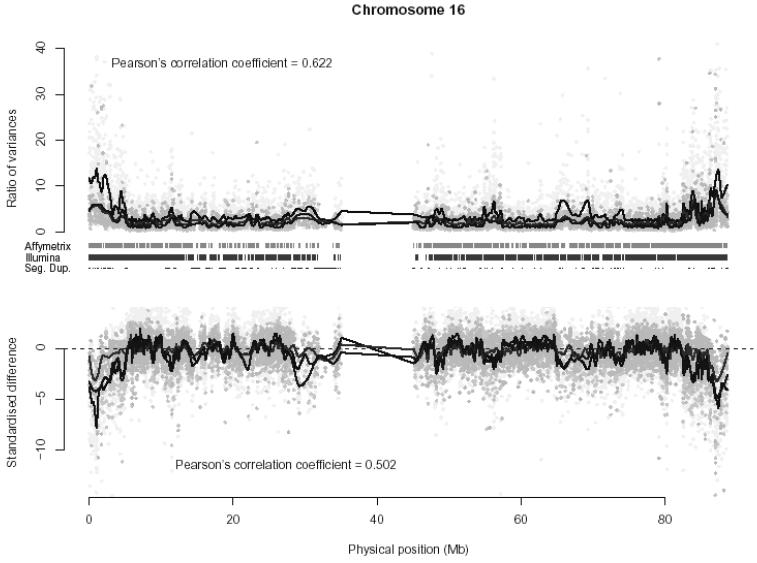

Chromosome 11
Pearson's correlation coefficient = 0.526
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
0
−5
−10
Pearson's correlation coefficient = 0.401
0
20
40
60
80
100
120
Physical position (Mb)

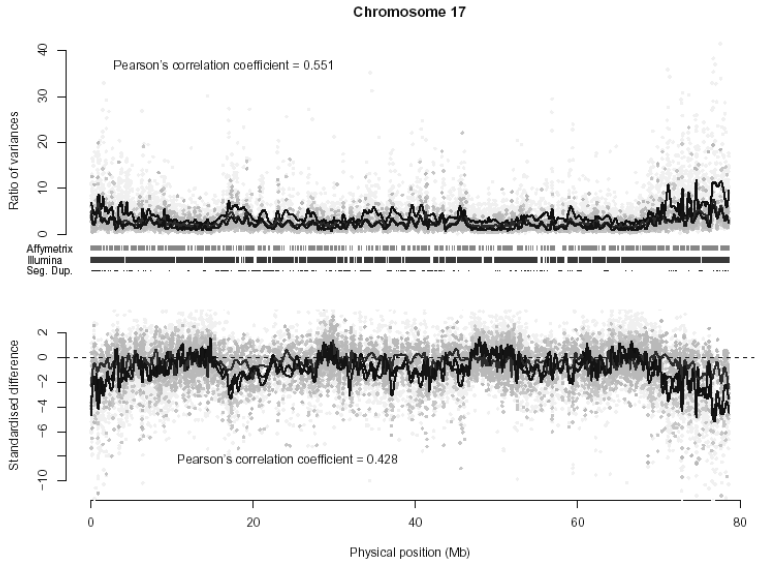

Chromosome 12
Pearson's correlation coefficient = 0.39
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
0
−5
−10
Pearson's correlation coefficient = 0.395
0
20
40
60
80
100
120
Physical position (Mb)

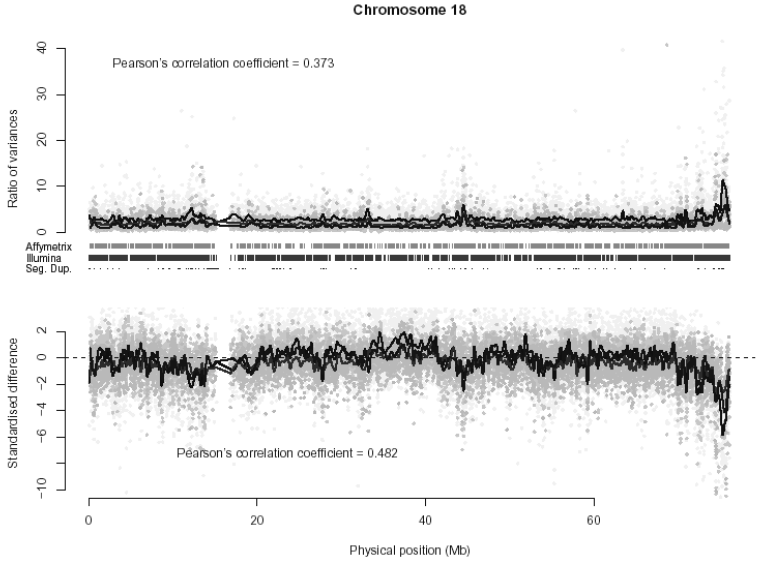

Chromosome 13
Pearson's correlation coefficient = 0.523
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
0
−5
−10
−15
Pearson's correlation coefficient = 0.481
20
40
60
80
100
Physical position (Mb)

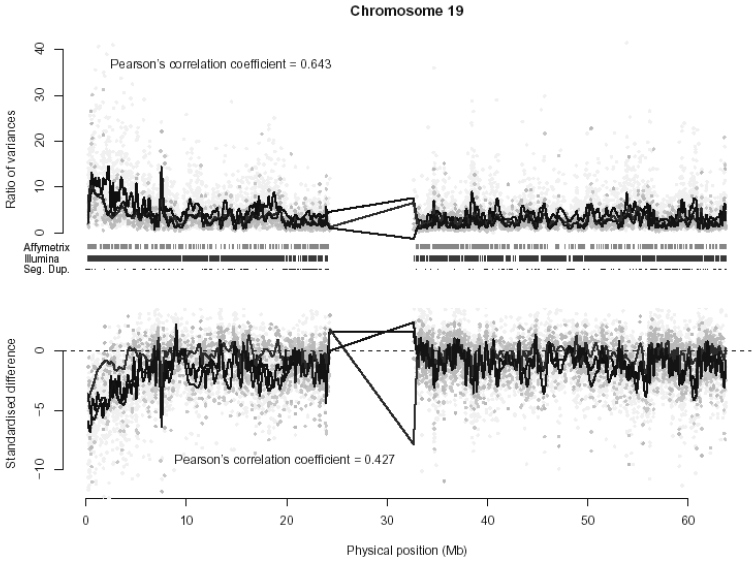

Chromosome 14
Pearson's correlation coefficient = 0.438
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
0
−5
−10
−15
Pearson's correlation coefficient = 0.391
20
40
60
80
100
Physical position (Mb)

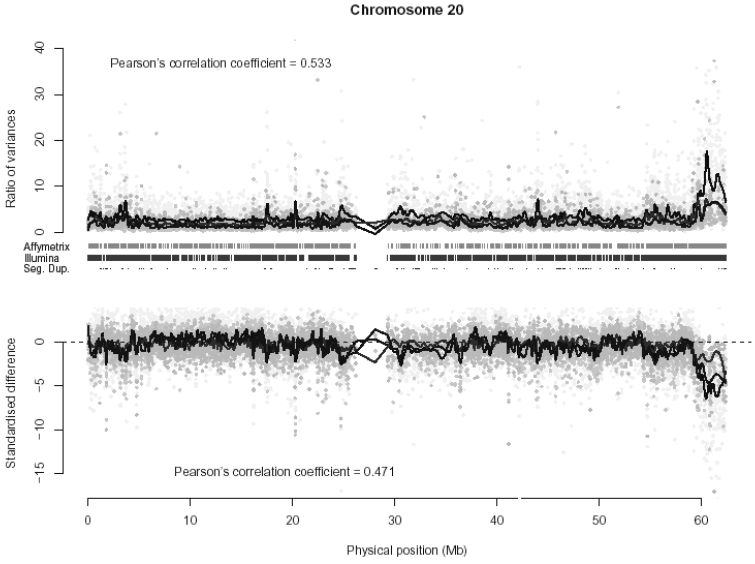

Chromosome 15
Pearson's correlation coefficient = 0.439
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.4
Physical position (Mb)

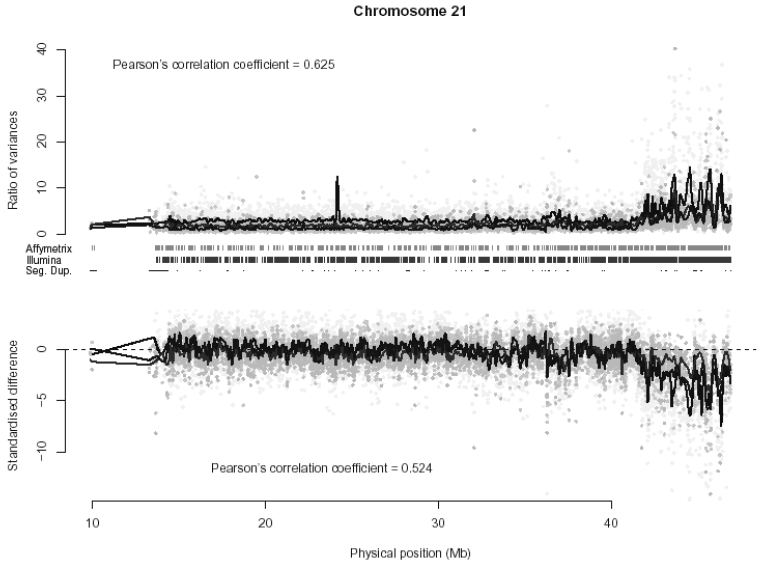

Chromosome 16
Pearson's correlation coefficient = 0.622
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.502
Physical position (Mb)

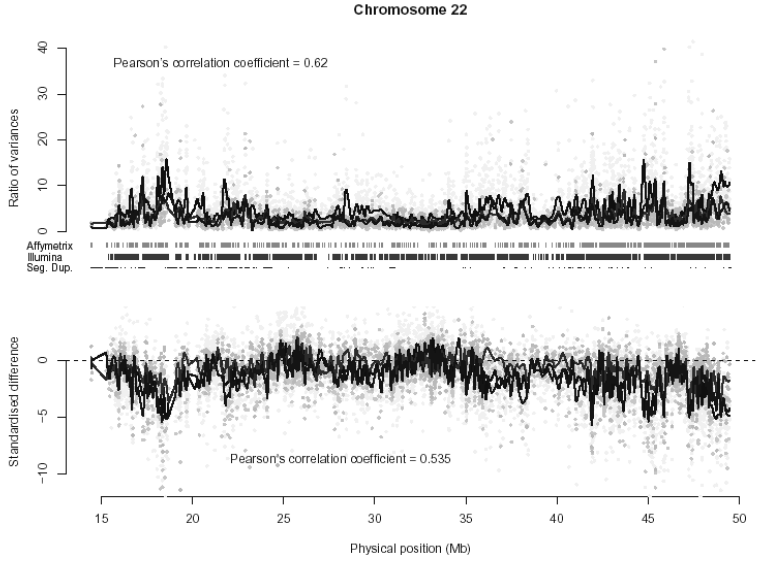

Chromosome 17
Pearson's correlation coefficient = 0.551
Ratio of variances
Affymetrix
Illumina
Seg. Dup.
Standardised difference
Pearson's correlation coefficient = 0.428
Physical position (Mb)


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
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
−10
−5
0
Pearson's correlation coefficient = 0.524
10
20
30
40
Physical position (Mb)


Chromosome 22
Pearson's correlation coefficient = 0.62
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
−10
−5
0
Pearson's correlation coefficient = 0.535
15
20
25
30
35
40
45
50
Physical position (Mb)

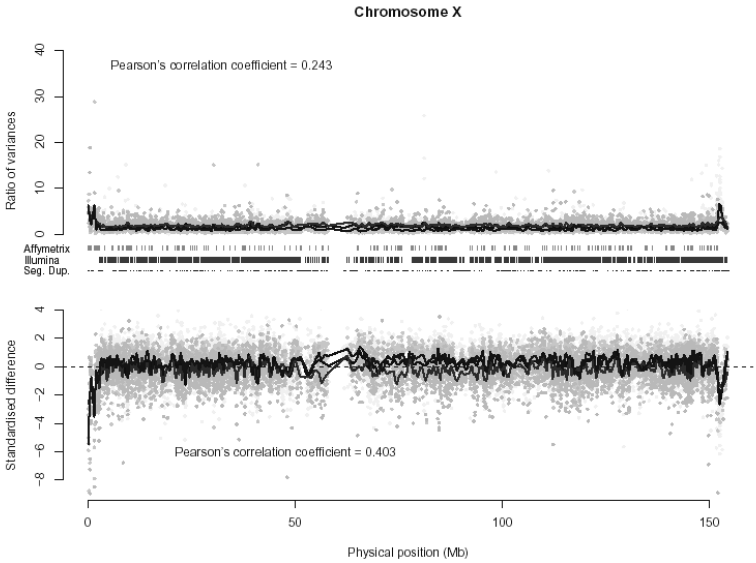

Chromosome X
Pearson's correlation coefficient = 0.243
Ratio of variances
0
10
20
30
40
Affymetrix
Illumina
Seg. Dup.
Standardised difference
−8
−6
−4
−2
0
2
4
Pearson's correlation coefficient = 0.403
0
50
100
150
Physical position (Mb)

Figure S2. 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.

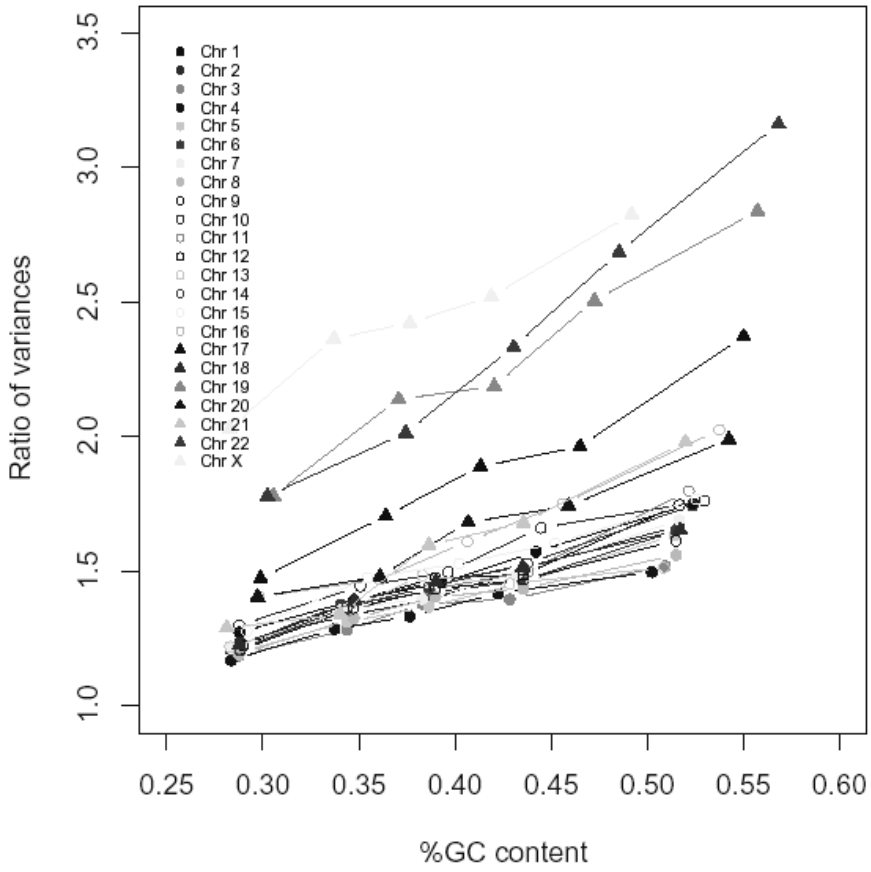


Figure S3. An example of the use of imputation in recovering the performance of a SNP with low call rate. (**a**) Clusterplot of a SNP on the Affymetrix platform on chromosome 19 for 278 ML individuals with whole genome amplified DNA, where dots in red correspond to the AA genotype; green to the AB genotype; blue to the BB genotype; and dots in grey represent individuals where valid genotypes failed to be assigned after thresholding the posterior probabilities of the genotypes. This SNP is designated as the focal SNP. (**b**) The LD map of common SNPs found within 200kb of the focal SNP, calculated from the HapMap YRI. The dotted line represents the position of the focal SNP, and the clusterplots of the SNPs represented by the dots in red are shown in (**c**). These 9 SNPs have well-separated clusters and have LD > 0.2 with the focal SNP. (**d**) The clusterplot for the focal SNP after the missing genotypes have been imputed using the program IMPUTE.

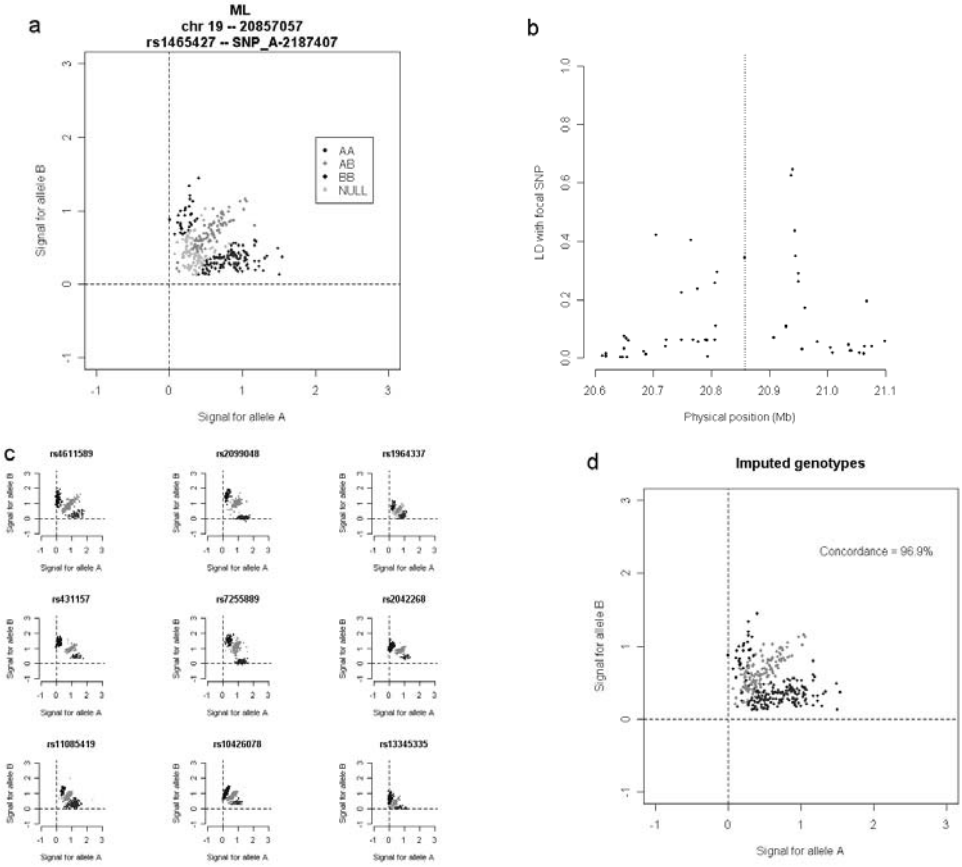
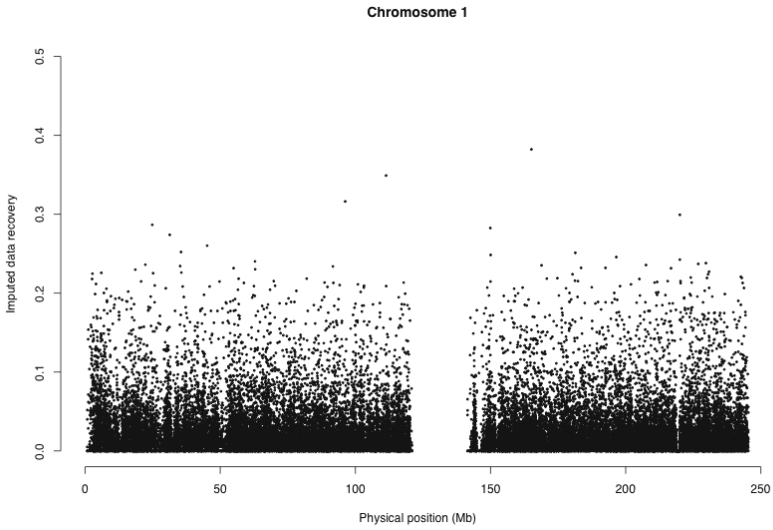
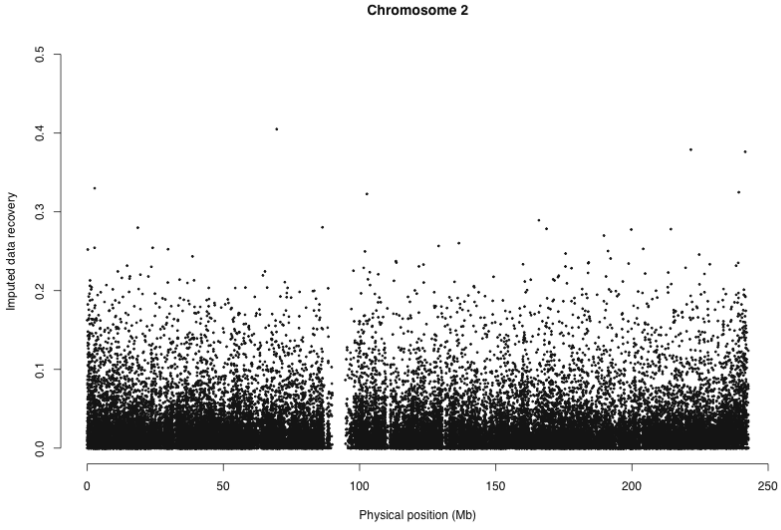
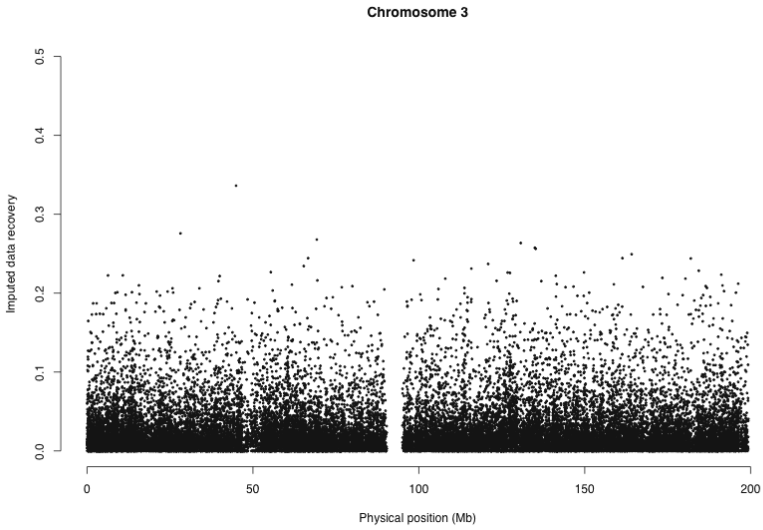
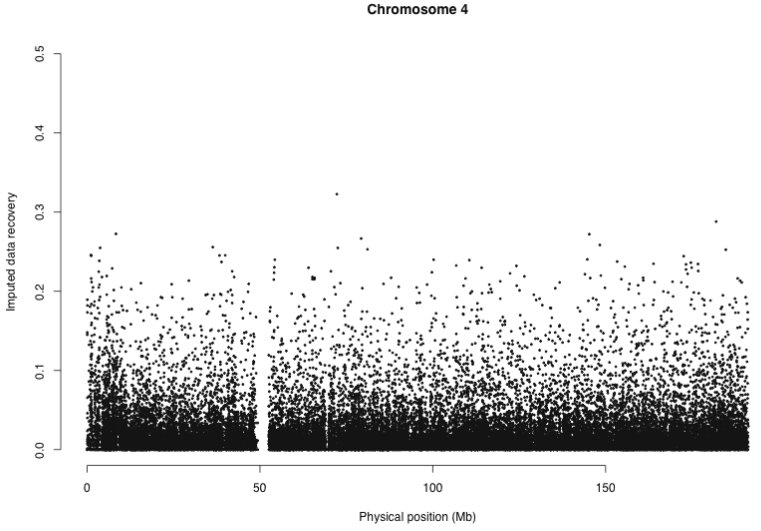


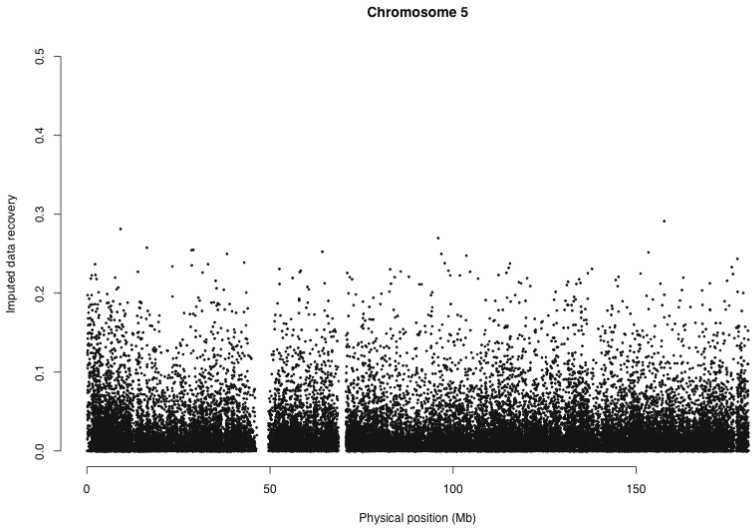
Figure S4. Genomic view of per-SNP data recovery when missing genotypes are imputed. The program IMPUTE was used to fill in data loss on the Affymetrix 500K microarray for all autosomal SNPs present in HapMap phase II. Imputed data recovery is the calculated on a per-SNP basis as the proportion of genotypes successfully imputed (>0.90 posterior probability threshold) in addition to those already successfully called. All chromosomes follow.

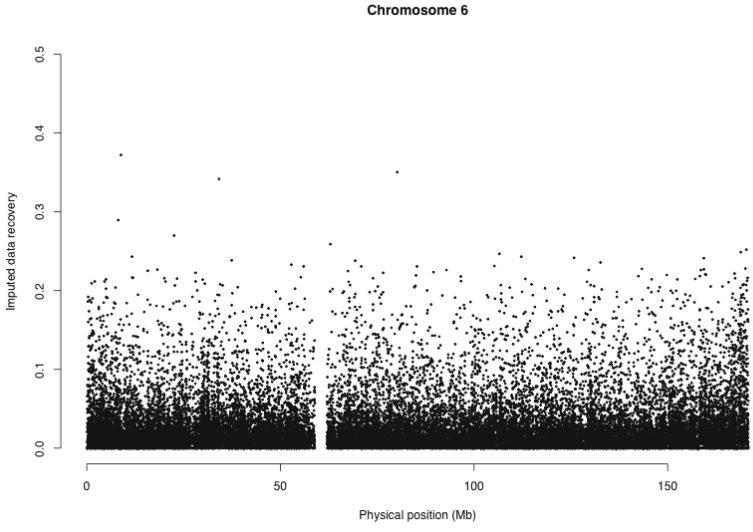

Chromosome 1
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
50
100
150
200
250
Physical position (Mb)

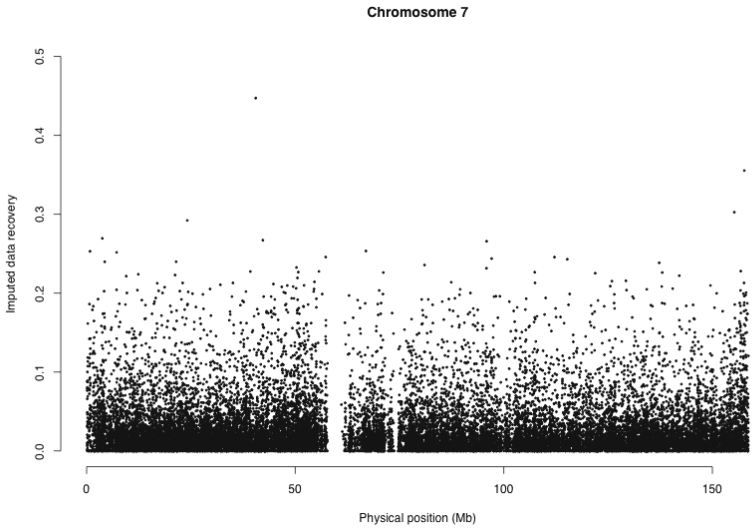

Chromosome 2
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
50
100
150
200
250
Physical position (Mb)

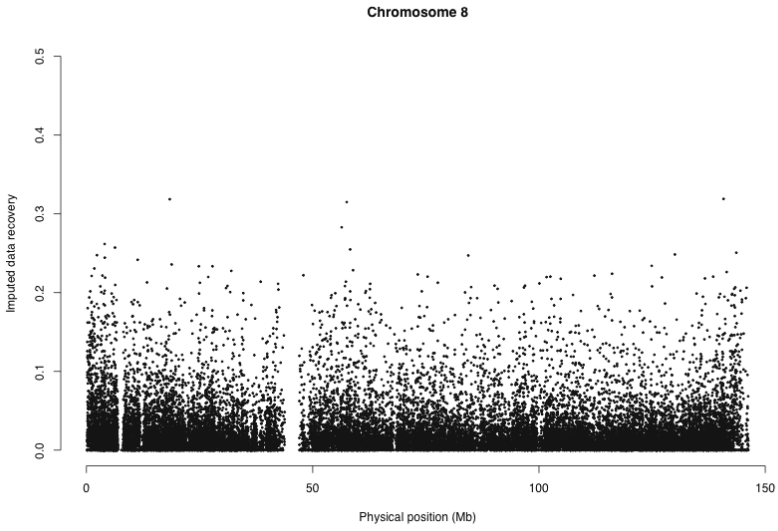

Chromosome 3
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
50
100
150
200
Physical position (Mb)

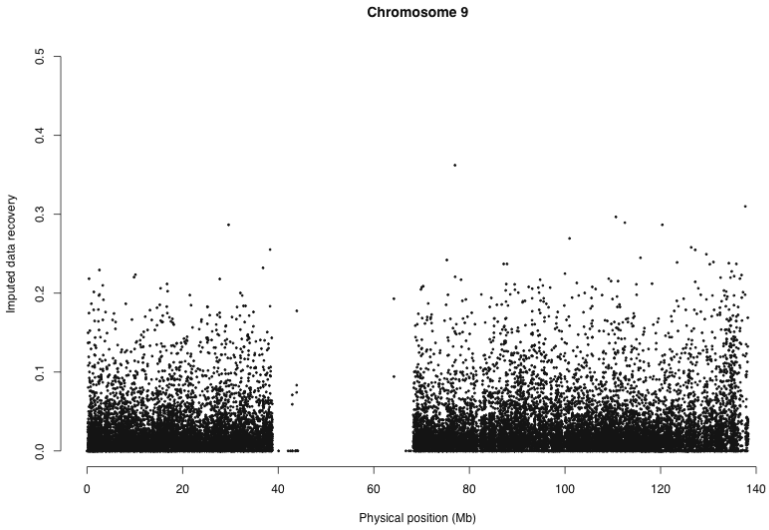

Chromosome 4
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
50
100
150
Physical position (Mb)

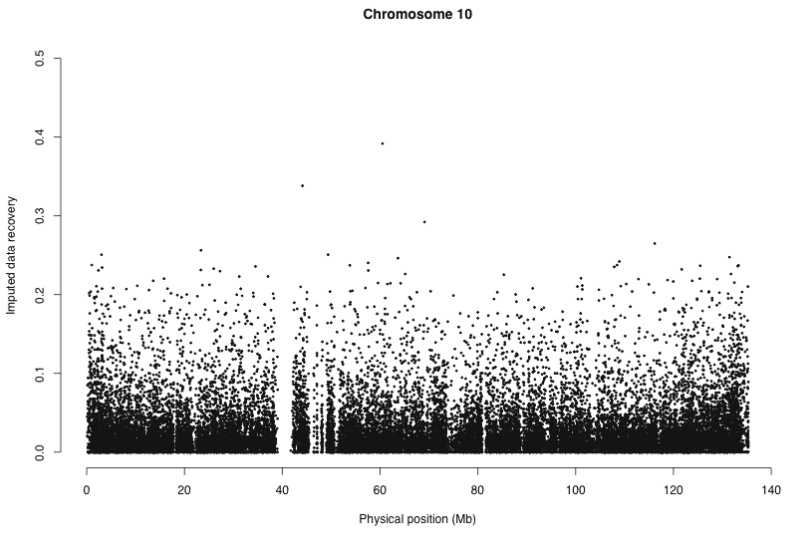

Chromosome 5
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
50
100
150
Physical position (Mb)

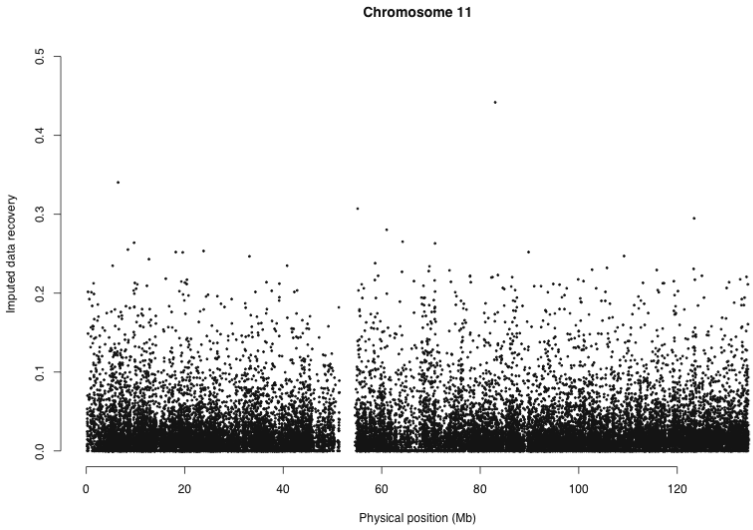

Chromosome 6
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
50
100
150
Physical position (Mb)

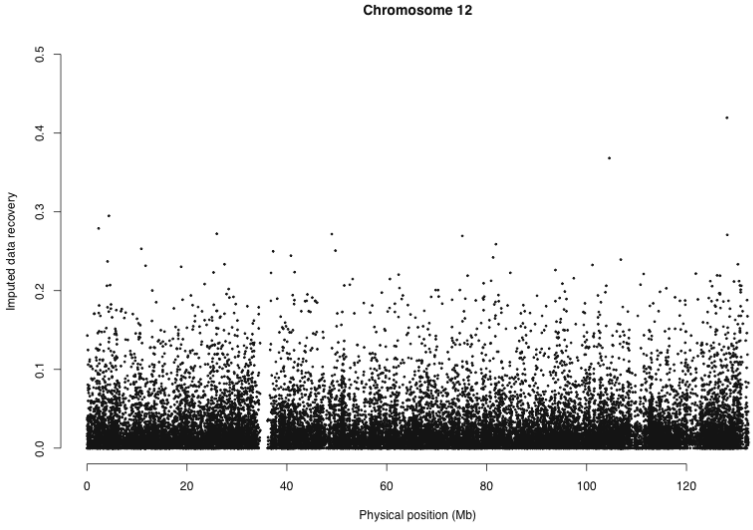

Chromosome 7
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
50
100
150
Physical position (Mb)

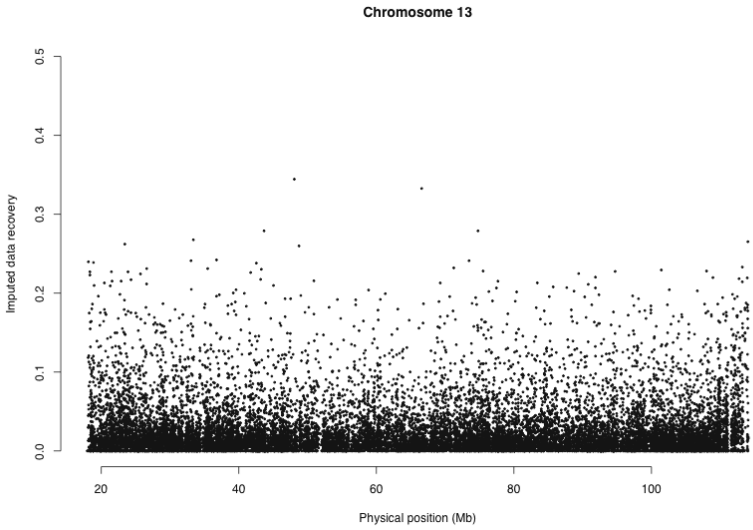

Chromosome 8
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
50
100
150
Physical position (Mb)

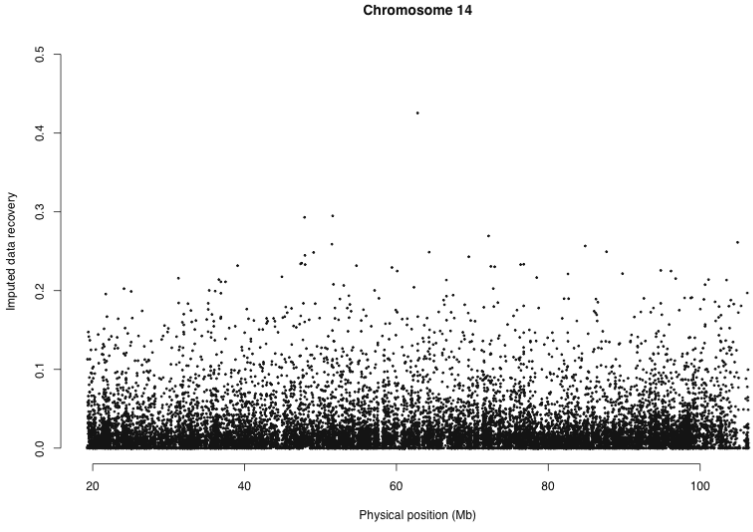

Chromosome 9
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
100
120
140
Physical position (Mb)

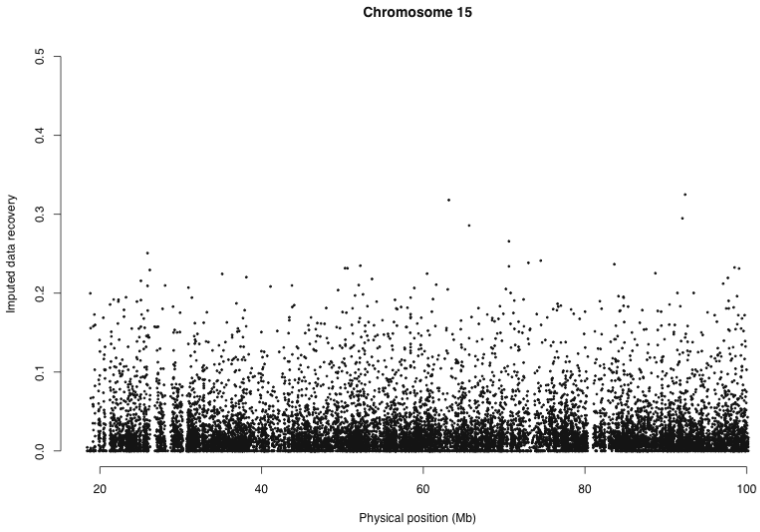

Chromosome 10
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
100
120
140
Physical position (Mb)

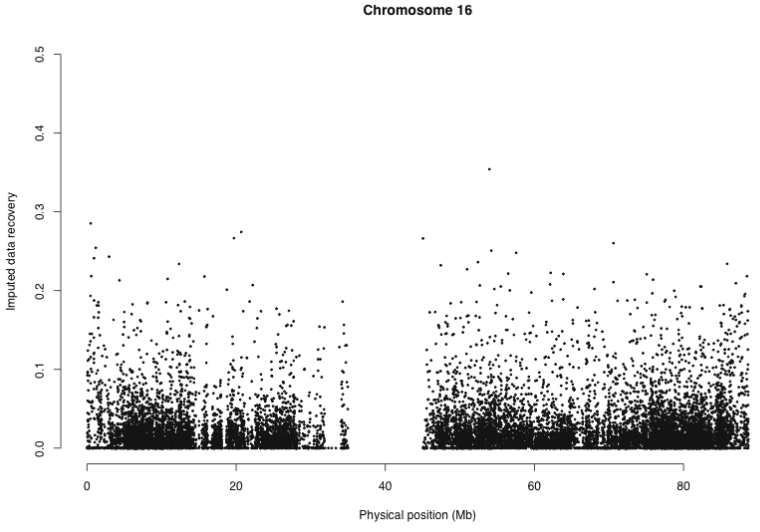

Chromosome 11
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
100
120
Physical position (Mb)

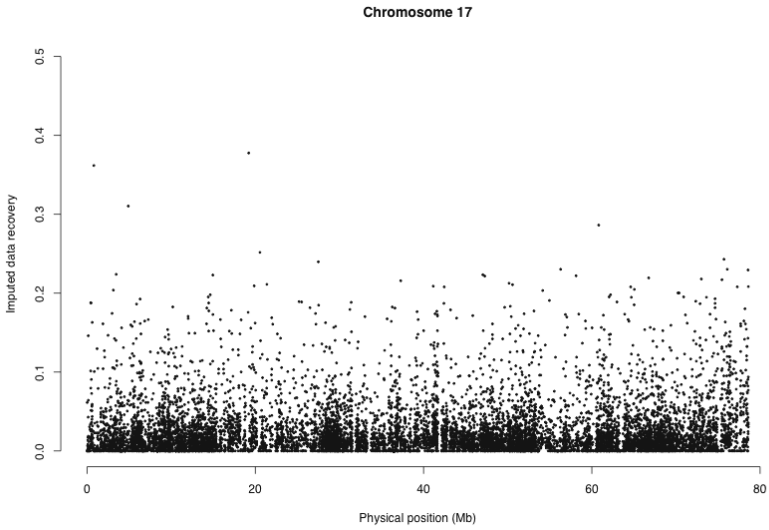

Chromosome 12
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
100
120
Physical position (Mb)

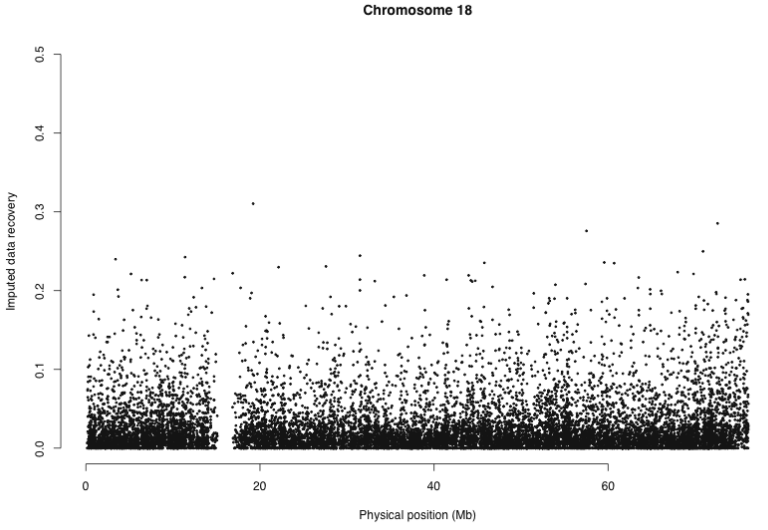

Chromosome 13
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
20
40
60
80
100
Physical position (Mb)

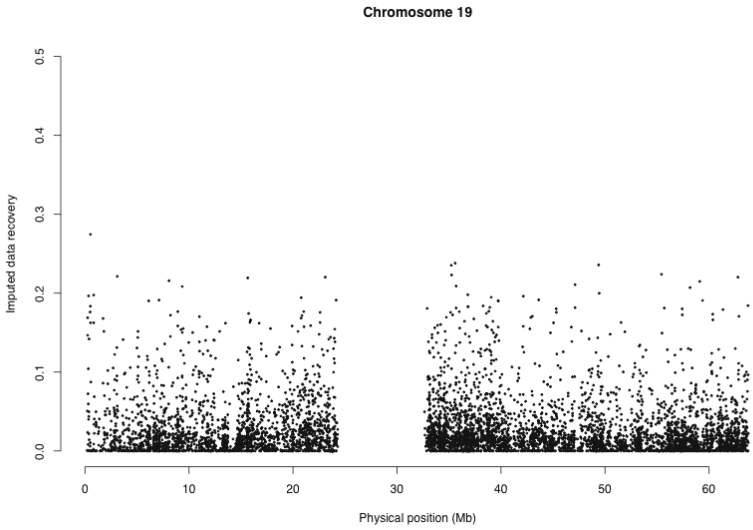

Chromosome 14
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
20
40
60
80
100
Physical position (Mb)

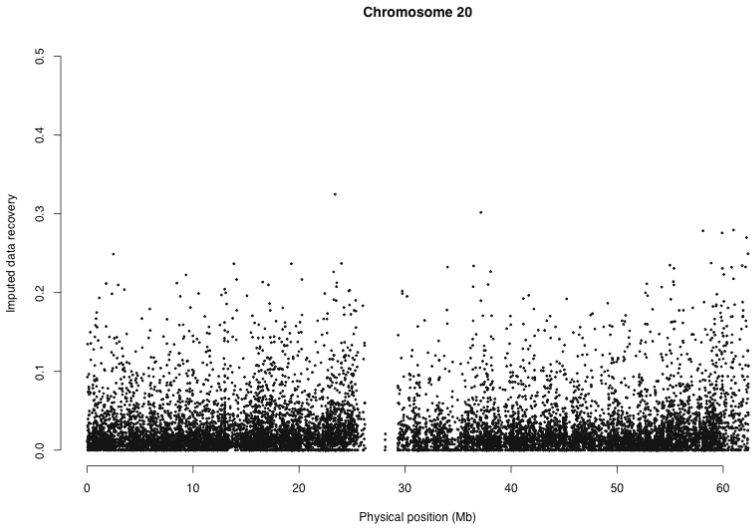

Chromosome 15
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
20
40
60
80
100
Physical position (Mb)

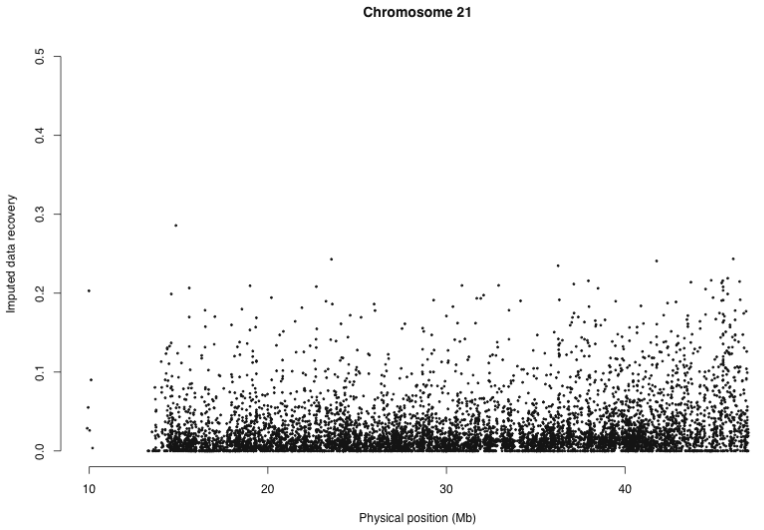

Chromosome 16
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
Physical position (Mb)

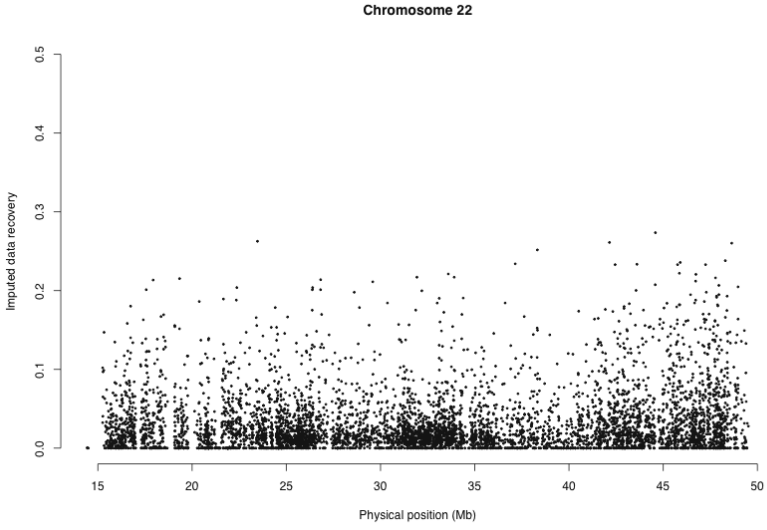

Chromosome 17
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
Physical position (Mb)


Chromosome 18
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
Physical position (Mb)


Chromosome 19
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
10
20
30
40
50
60
Physical position (Mb)


Chromosome 20
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
0
10
20
30
40
50
60
Physical position (Mb)


Chromosome 21
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
10
20
30
40
Physical position (Mb)


Chromosome 22
Imputed data recovery
0.0
0.1
0.2
0.3
0.4
0.5
15
20
25
30
35
40
45
50
Physical position (Mb)

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).

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).

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).