Vol 447|7 June 2007|doi:10.1038/nature05911 nature

ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

To be discussed in SMG

November 23, 2007

Editor's Summary

7 June 2007
Spreading the genomic net

With the advent of many more markers in the human genome, **it has become possible to search for genes associated with human disease without having to narrow down candidate regions of the genome first.** In a ground-breaking publication, the Wellcome Trust Case Control Consortium reports an exciting **genome-wide association study** of some 17,000 individuals for seven common familial diseases. The analysis confirms previously identified loci and provides strong evidence for many novel disease susceptibility genes.

News and Views: Genomics: Guilt by association

In a tour-de-force demonstration of feasibility, **a consortium of 50 research teams uses 500,000 genetic markers from each of 17,000 individuals to identify 24 genetic risk factors for 7 common human diseases.**

WTCCC Summary

- 7 major diseases, e.g. Bipolar, Hypertension
- SNPs: Affymetrix 500K
- Samples:
  - ~2,000 cases for each of the 7 diseases
  - ~3,000 shared controls
- Analyses:
  - much effort on QC
  - simple trend and genotype tests
  - novel imputation method
- Results:
  - 24 independent association signals at $p < 5x10^{-7}$;
    - almost all true positives based on previous&replication studies
    - Some of the loci confer risk for multiple diseases studies
  - 58 additional loci at $10^{-5} < p < 5x10^{-7}$

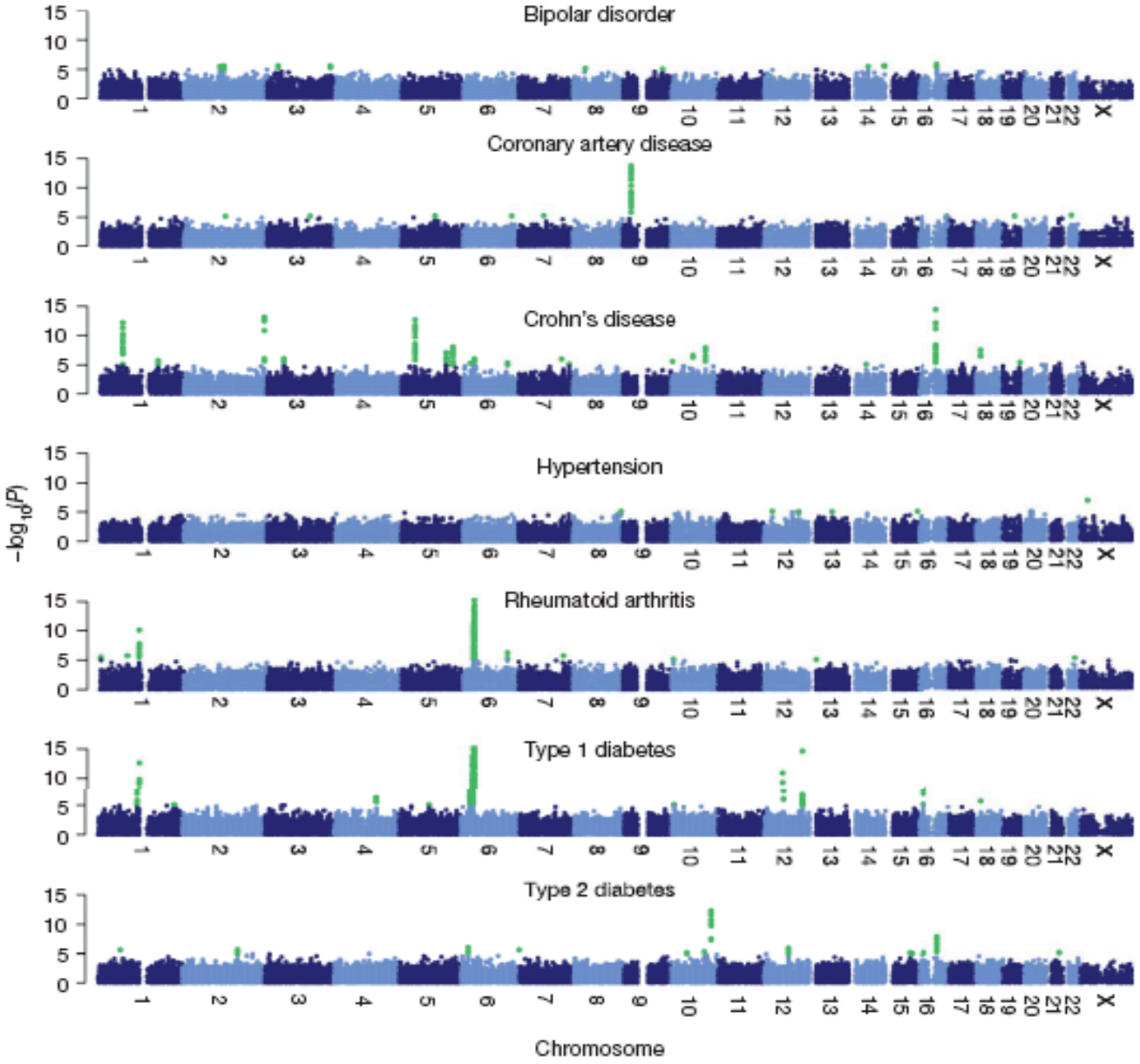


Figure 4 | Genome-wide scan for seven diseases. For each of seven diseases $-\log_{10}$ of the trend test P value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome. Chromosomes are shown in alternating colours for clarity, with P values $<1 \times 10^{-5}$ highlighted in green. All panels are truncated at $-\log_{10}(P\text{value}) = 15$, although some markers (for example, in the MHC in T1D and RA) exceed this significance threshold.

Table 3 | Regions of the genome showing the strongest association signals

Collection	Chromosome	Region (Mb)	SNP	Trend *P* value	Genotypic *P* value	$\log_{10}$(BF), additive	$\log_{10}$(BF), general	Risk allele	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
					Standard analysis								
BD	16p12	23.3–23.62	rs420259	2.19×10^{-04}	6.29×10^{-08}	1.96	4.79	A	G	2.08 (1.60–2.71)	2.07 (1.6–2.69)	0.282	0.248
CAD	9p21	21.93–22.12	rs1333049	1.79×10^{-14}	1.16×10^{-13}	11.66	11.19	C	C	1.47 (1.27–1.70)	1.9 (1.61–2.24)	0.474	0.554
CD	1p31	67.3–67.48	rs11805303	6.45×10^{-13}	5.85×10^{-12}	10.07	9.41	T	T	1.39 (1.22–1.58)	1.86 (1.54–2.24)	0.317	0.391
CD	2q37	233.92–234	rs10210302	7.10×10^{-14}	5.26×10^{-14}	11.11	11.28	T	C	1.19 (1.01–1.41)	1.85 (1.56–2.21)	0.481	0.402
CD	3p21	49.3–49.87	rs9858542	7.71×10^{-07}	3.58×10^{-08}	4.24	5.22	A	A	1.09 (0.96–1.24)	1.84 (1.49–2.26)	0.282	0.331
CD	5p13	40.32–40.66	rs17234657	2.13×10^{-13}	1.99×10^{-12}	10.41	9.89	G	G	1.54 (1.34–1.76)	2.32 (1.59–3.39)	0.125	0.181
CD	5q33	150.15–150.31	rs1000113	5.10×10^{-08}	3.15×10^{-07}	5.36	5.01	T	T	1.54 (1.31–1.82)	1.92 (0.92–4.00)	0.067	0.098
CD	10q21	64.06–64.31	rs10761659	2.68×10^{-07}	1.75×10^{-06}	4.69	4.13	G	A	1.23 (1.05–1.45)	1.55 (1.3–1.84)	0.461	0.406
CD	10q24	101.26–101.32	rs10883365	1.41×10^{-08}	5.82×10^{-08}	5.91	5.48	G	G	1.2 (1.03–1.39)	1.62 (1.37–1.92)	0.477	0.537
CD	16q12	49.02–49.4	rs17221417	9.36×10^{-12}	3.98×10^{-11}	8.93	8.47	G	G	1.29 (1.13–1.46)	1.92 (1.58–2.34)	0.287	0.356
CD	18p11	12.76–12.91	rs2542151	4.56×10^{-08}	2.03×10^{-07}	5.42	5.00	G	G	1.3 (1.14–1.48)	2.01 (1.46–2.76)	0.163	0.208
RA	1p13	113.54–114.16	rs6679677	4.90×10^{-26}	5.55×10^{-25}	22.36	21.99	A	A	1.98 (1.72–2.27)	3.32 (1.93–5.69)	0.096	0.168
RA	6	MHC	rs6457617*	3.44×10^{-76}	5.18×10^{-75}	74.84	73.18	T	T	2.36 (1.97–2.84)	5.21 (4.31–6.30)	0.489	0.685
T1D	1p13	113.54–114.16	rs6679677	1.17×10^{-26}	5.43×10^{-26}	23.07	22.83	A	A	1.82 (1.59–2.09)	5.19 (3.15–8.55)	0.096	0.169
T1D	6	MHC	rs9272346*	2.42×10^{-134}	5.47×10^{-134}	141.9	142.2	A	G	5.49 (4.83–6.24)	18.52 (27.03–12.69)	0.387	0.150
T1D	12q13	54.64–55.09	rs11171739	1.14×10^{-11}	9.71×10^{-11}	8.89	8.24	C	C	1.34 (1.17–1.54)	1.75 (1.48–2.06)	0.423	0.493
T1D	12q24	109.82–111.49	rs17696736	2.17×10^{-15}	1.51×10^{-14}	12.53	11.88	G	G	1.34 (1.16–1.53)	1.94 (1.65–2.29)	0.424	0.506
T1D	16p13	10.93–11.37	rs12708716	9.24×10^{-08}	4.92×10^{-07}	5.15	4.70	A	G	1.19 (0.97–1.45)	1.55 (1.27–1.89)	0.350	0.297
T2D	6p22	20.63–20.84	rs9465871	1.02×10^{-06}	3.34×10^{-07}	4.15	3.98	C	C	1.18 (1.04–1.34)	2.17 (1.6–2.95)	0.178	0.218
T2D	10q25	114.71–114.81	rs4506565	5.68×10^{-13}	5.05×10^{-12}	10.14	9.43	T	T	1.36 (1.2–1.54)	1.88 (1.56–2.27)	0.324	0.395
T2D	16q12	52.36–52.41	rs9939609	5.24×10^{-08}	1.91×10^{-07}	5.35	5.05	A	A	1.34 (1.17–1.52)	1.55 (1.3–1.84)	0.398	0.453
					Multi-locus analysis								
T1D	4q27	123.26–123.92	rs6534347	4.48×10^{-07}	1.83×10^{-06}	5.15	4.69	A	A	1.30 (1.10–1.55)	1.49 (1.25–1.78)	0.351	0.402
T1D	12p13	9.71–9.86	rs3764021	7.19×10^{-05}	5.08×10^{-08}	2.12	4.55	C	T	1.57 (1.38–1.79)	1.48 (1.25–1.75)	0.467	0.426
					Sex differentiated analysis								
RA	7q32	130.80–130.84	rs11761231	3.91×10^{-07}	1.37×10^{-06}	-	-	G	A	1.44 (1.19–1.75)	1.64 (1.35–1.99)	0.375	0.327
					Combined cases								
RA+T1D	10p15	6.07–6.17	rs2104286	5.92×10^{-08}	2.52×10^{-07}	5.26	4.45	T	C	1.35 (1.11–1.65)	1.62 (1.34–1.97)	0.286	0.245

Regions with at least one SNP with a *P* value of less than 5×10^{-7} for our primary analyses. The $\log_{10}$ value of the Bayes factor (BF) for the bayesian analysis corresponding to the trend and genotypic tests is also given. Region marks the boundaries of signal defined by recombination and return of test statistics to background levels. The minor allele is defined in the controls and its frequency in that group as well as the case sample is reported. MAF, minor allele frequency. Cluster plots for each SNP have been inspected visually, and are shown in Supplementary Fig. 10. Positions are in NCBI build-35 coordinates *Multiple SNPs in the MHC region are significant, we report the most extreme.

Relevant Statistical Issues

Focus of the Discussion

- **Data Quality Control (QC)**
- **Basic Analysis**
- **Assess Significance**
- **Power and Sample Size**
- **QQ-plot**
- **Missing Data Imputation**

Others

- SNP calling algorithms
- Shared Controls
- Population Stratification
- Bayes Factor
- Multiple Correlated Phenotypes
- Replications
- Winner's Curse, Selection Bias
- Multi-stage Design
- Gene-based Analyses
- Multiple Correlated Phenotypes
- Multiple Datasets
- Family-based vs. Case-Control
- Many more …

lysis. In addition to our main association results, we address several of these issues below, including the choice of controls for genetic studies, the extent of population structure within Great Britain, sample sizes necessary to detect genetic effects of varying sizes, and improvements in genotype-calling algorithms and analytical methods.

Analyses Flowchart

- Shared controls
- Genotype calling
- **QC of samples & SNPs**
- A number of association tests
- **Interpretation of the results**
- Recommendations and guidelines

Shared Controls

- Two sources
  - 58C: 1,480 from the 1958 British Cohort
  - UKBS: 1,458 recruited as part of this project

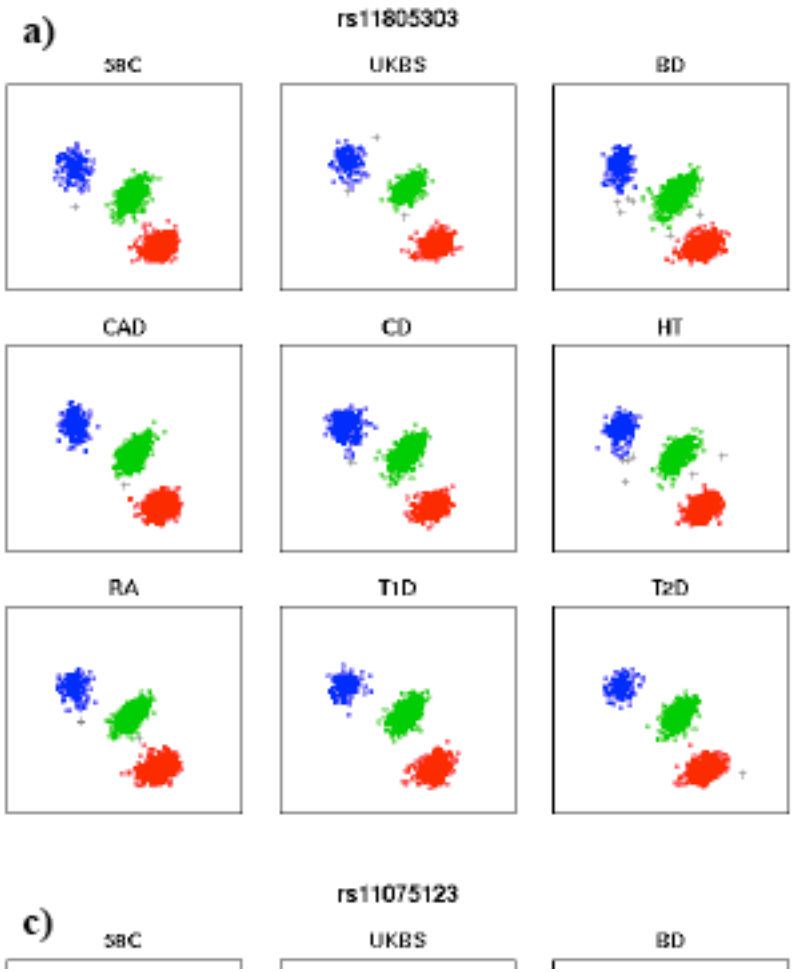
- Purpose of using two control groups: assess possible
  - Ascertainment bias
  - Genotyping error (differ in DNA collection and preparation)

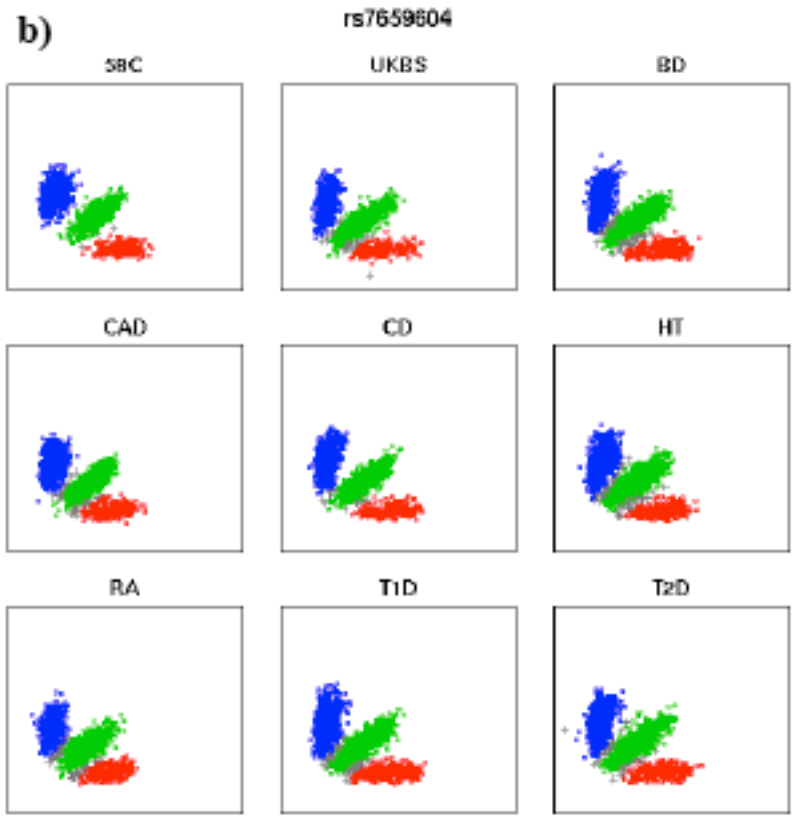
Allele/Genotype Calling

- Can easily lead to spurious associations if there is a systematic differences between cases and controls
- Two steps
  - Normalization of the signals
  - Genotype calling (WTCCC: CHIAMO)
    - Bayesian hierachical 4-class (AA/AB/BB/00) mixture model
    - Cluster plots: ensure correct genotype counts (more confident about a true association)
      - Clear, distinct clusters on the plot
      - Consistency between the clusters and actual genotype calling

a) A SNP with well-separated genotype cloud

b) A more challenged SNP





Supplementary Figure 17 | The genotyping calling challenge. Scatter plots of normalized probe intensities

Analyses Flowchart

- Shared controls
- Genotype calling
- **QC of samples & SNPs**
- A number of association tests
- Interpretation of the results
- Recommendations and guidelines

Quality Control (QC)

- Aim: exclude poor quality data without removing genuine associations
- Both Samples & SNPs
- Methods to detect them
  - Common practice: separately for samples & SNPs
  - Ideally: jointly
- What to do after error detection
  - Common practice: discard using cut-points
  - Ideally: incorporate data quality measures directly into the association tests
- Empirical evidence
  - MANY things can go wrong!
  - Time consuming!

QC - Samples

WTCCC: 6 filters

- Missingness
- Heterozygosity (contamination)
- External Discordance
- Non-European Ancestry
- Duplicates
- Cryptic Relatedness

- 16,988 samples
- 809 excluded
- 16,179 remained (95.2%)

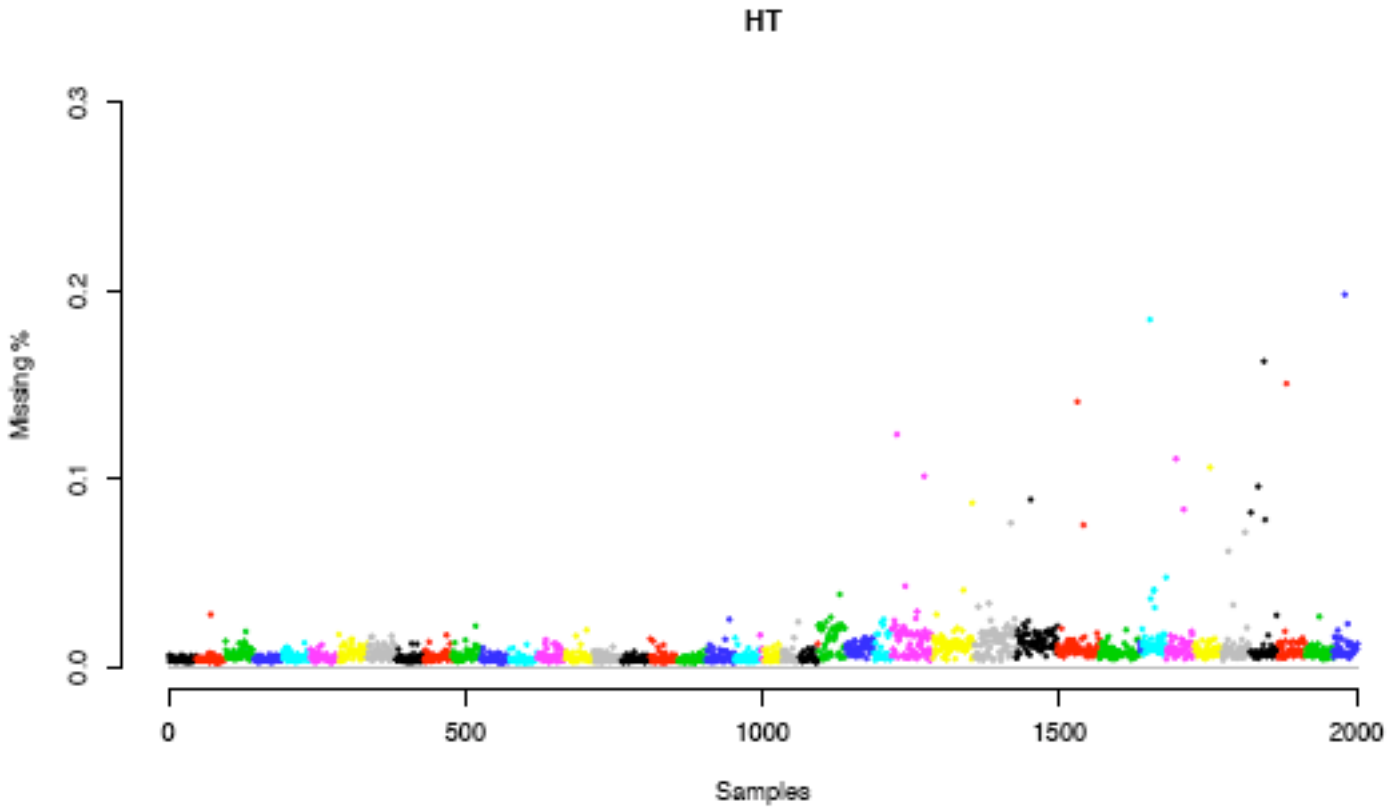
Collection	Missingness	Heterozygosity	External discordance	Non-European ancestry	Duplicate	Relative	Total
58C	9	0	4	6	4	1	**24**
UKBS	8	0	5	14	0	15	**42**
BD	30	0	0	9	77	13	**129**
CAD	41	1	0	13	2	5	**62**
CD	43	4	6	54	131	18	**256**
HT	29	0	0	2	6	11	**48**
RA	47	1	0	26	53	9	**136**
T1D	7	2	1	18	6	3	**37**
T2D	36	1	0	11	16	11	**75**
Total	**250**	**9**	**16**	**153**	**295**	**86**	**809**

Supplementary Table 4 | Exclusion summary by collection. Six filters were applied for sample exclusion: 1. SNP call rate < 97% (missingness). 2. Heterozygosity > 30% or < 23% across all SNPs. 3. External discordance with genotype or phenotype data. 4. Individuals identified as having recent non-European ancestry by the Multidimensional Scaling analysis (see Methods). 5. Duplicates (the copy with more missing data was removed) 6. Individuals with too much IBS sharing (>86%); likely relatives. Where individuals could be excluded for more than one reason, they appear in the leftmost such column.

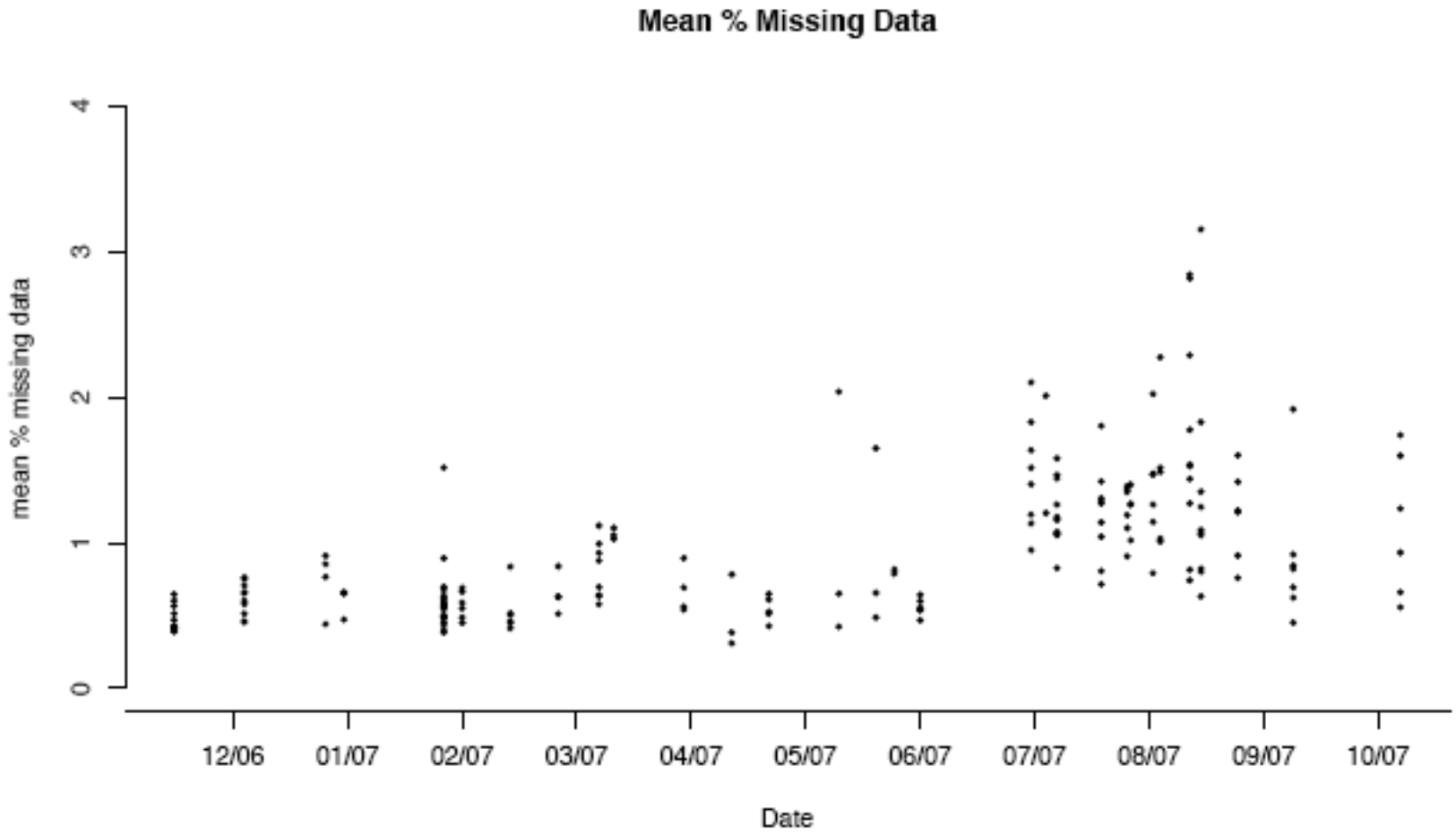
QC Sample - Missingness

- Average SNP call rate per sample
- Indicator of low DNA quality
- Metric: # of “null” SNPs / total # of SNPs
- WTCCC cut-point: < 97% (250 samples)
- Most sample with low rates of missing data

  mu: 0.00925; sd: 0.0187
- Interesting: trend over time, e.g. Hypertension

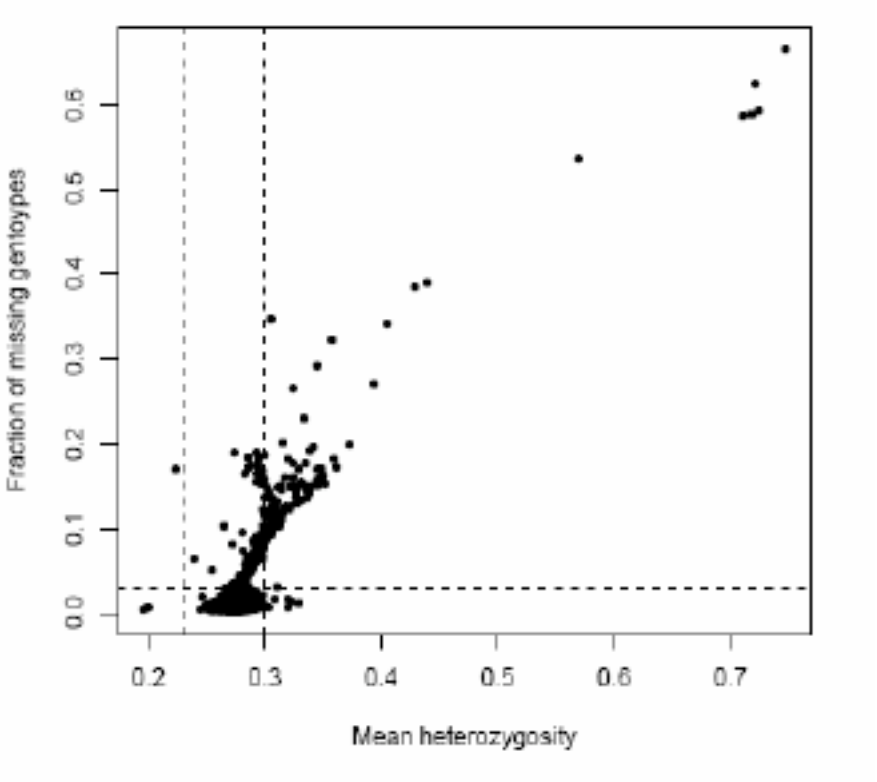


Supplementary Figure 25 | Missing genotypes per collection over time. The plots show the percentage of missing data for each individual in the unfiltered dataset for each collection. The individuals are coloured by plate, and plates of individuals are arranged left to right in ascending order of the time the plate was shipped to Affymetrix. The plot titled "Mean % Missing Data" shows the mean percentage of missing data for each plate of individuals in the unfiltered dataset. The plates of individuals are arranged by the date the plate was shipped to Affymetrix.


Mean % Missing Data
mean % missing data
0
1
2
3
4
12/06
01/07
02/07
03/07
04/07
05/07
06/07
07/07
08/07
09/07
10/07
Date

QC Sample - Heterozygosity

- Proportion of heterozygous SNPs per sample
- Excess heterozygosity indicate contamination
- Metric: # of heterozygous SNPs / total # of SNPs with data
- WTCCC cutpoint: > 30% or < 23%
  - Empirical thresholds
  - 6 > 30% and 3 < 23% (excluding those already excluded for missingness)
  - Why 30% and 23%?



Supplementary Figure 18 | Individual missing data and heterozygosity. Scatter plot of the proportion of SNPs called heterozygote (x-axis) against the proportion called missing at a posterior probability threshold of 0.9 (y-axis) for each individual in the study. Dotted lines delimit the threshold used for exclusion of individuals from further analysis.

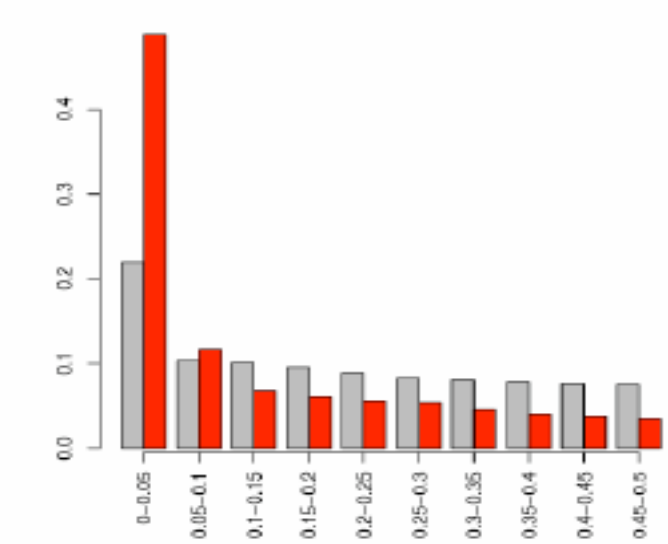
Why 30% and 23%?

- Prob (heterozygous) = $2*p*(1-p) = 2*(p-p^2)$
- Average over all SNPs

Average Heterozygosity = $2*(mu_p-(var_p-mu_p^2))$

- Assume
  - MAF: 0.025 0.075 0.125 0.175 0.225 0.275 0.325 0.375 0.425 0.475
  - Proportion: 0.220 0.110 0.010 0.098 0.096 0.094 0.092 0.091 0.090 0.090
  - $mu_p = 0.224$; $var_p = 0.024$

Average Heterozygosity = 0.299



Supplementary Figure 1 | Minor allele frequency (MAF) spectrum of SNPs included and excluded from the study. Red bars show the proportion of SNPs excluded from the study (see text and Methods) in 10 MAF bins. Grey bars show the frequency spectrum of included SNPs.

QC Sample - External Discordance

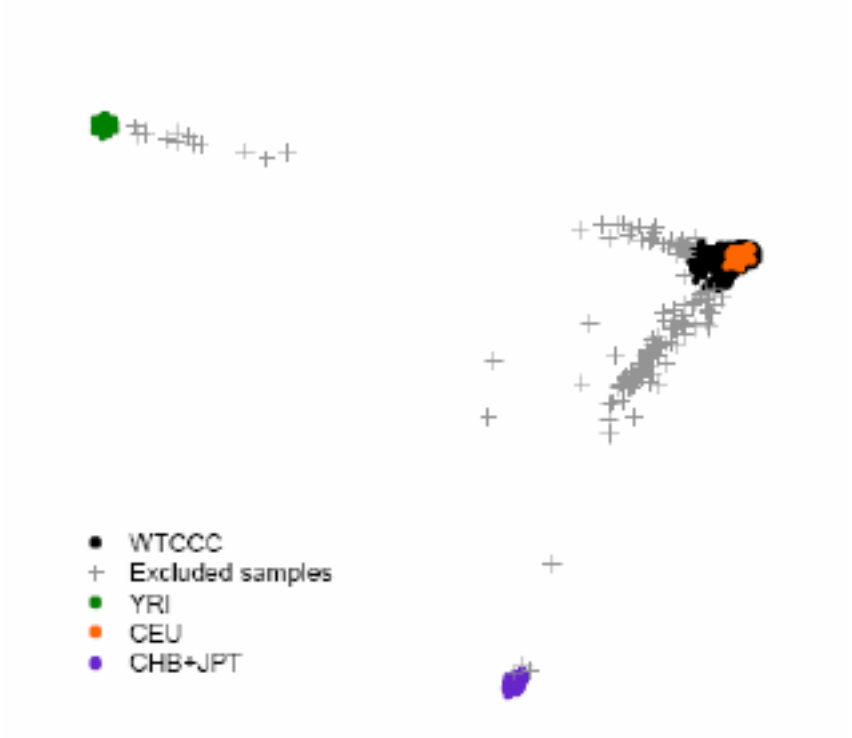
- 16 samples
- Discrepancies between WTCCC information and external identifying information, e.g.
  - genotypes from another experiment
  - blood type
  - incorrect disease status
- Would be study-specific, e.g. sex, age, etc

QC Sample - Non-European Ancestry

- 153 samples
- Multi-Dimensional Scaling (MDS)

  a set of related statistical techniques often used in data visualization for **exploring similarities or dissimilarities in data**. An MDS algorithm starts with a matrix of item-item similarities, then assigns a location of each item in a low-dimensional space

  – Select a set of 71,458 SNPs with $r^2 < 0.2$
  – Compute average IBS statistic (proportion of IBS sharing per sample pair) between each pair of individuals within each sample collection along with 270 HapMap samples
  – Project the matrix of pair-wise IBS values onto the MDS axes
  – **Exclusion of the 153 samples resulted in a substantial reduction in estimates of over-dispersion in test statistic distribution**



Supplementary Figure 5 | Multidimensional Scaling (MDS). WTCCC and HapMap samples plotted for the first two principal components obtained by multidimensional scaling of a matrix of pairwise IBS values between samples. Samples near the YRI cluster were subsequently identified in sample records as Afro-Caribbean; the large cluster one-third of the way between CEU and CHB+JPT were subsequently identified as South Asian (India/Pakistan). Samples showing evidence of non-European ancestry were excluded from further analyses (grey crosses).

QC Sample - Duplicates, Cryptic Relatedness

- All belong to “Pedigree” Errors
- Empirical evidence from many GWA studies:
  - Rate of such errors can be quite high
  - Putatively unrelated cases/controls can be first-degree relatives, e.g. sibs
- Impact: PCA, association test statistics
- WTCCC
  - a simple IBS counting methods
    - 295 duplicated (>99% IBS identity)
    - 86 related (86%-98% IBS identity)
  - straightforward, but low power

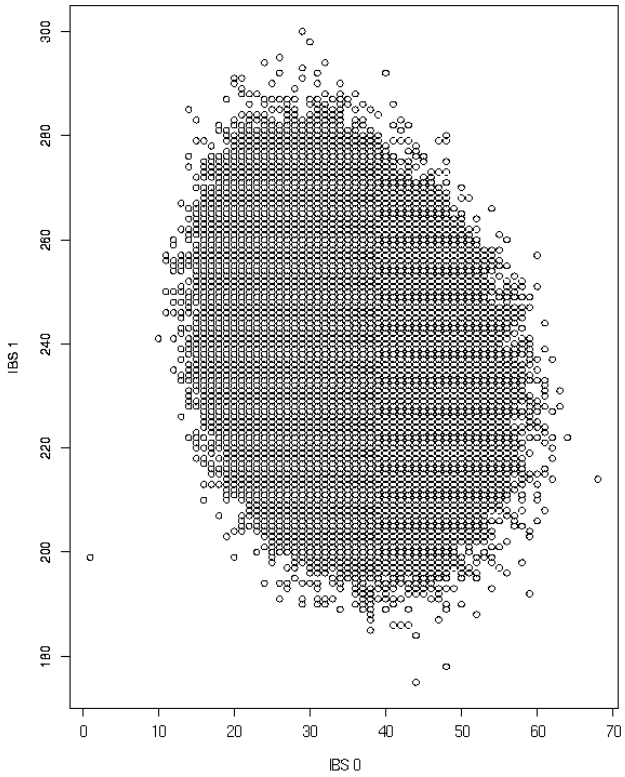
QC Sample - Duplicates, Cryptic Relatedness

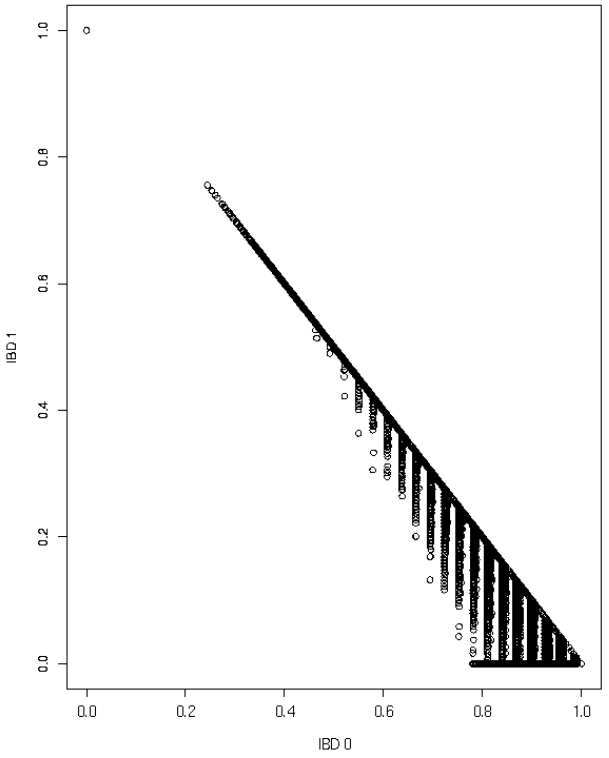
- Alternative methods
- **PLINK**
  - A method of moment
  - Fast, less powerful, sensitive to misspecification of allele frequency
- **PREST**:
  - Likelihood approach, MLE, EM
  - A bit slower (1.5 hrs for 1 million pairs), more powerful, robustness to allele frequency?
- Multiplicity Issues:
  - # of pairs in the order of $n^2/2$
  - e.g. ~8 million relative pairs for 2000/2000 cases/controls
  - Some will look related by chance

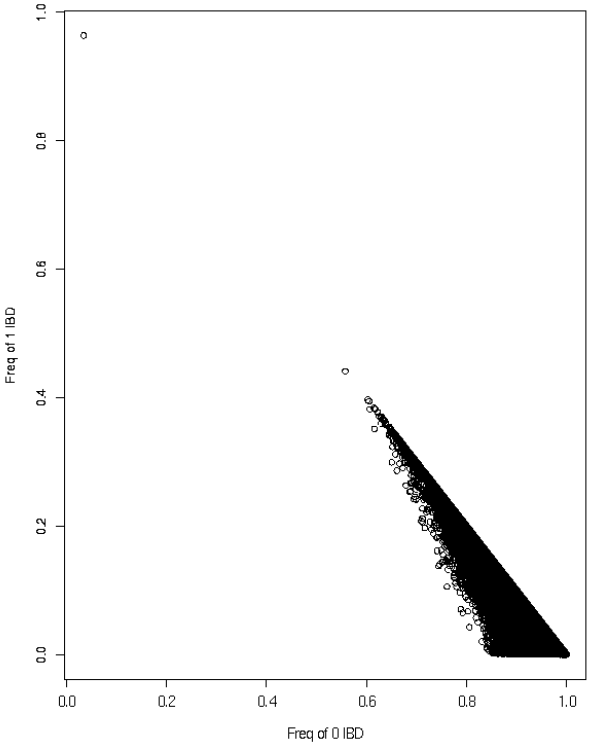
A preliminary study

- ~1,400 individuals, ~1 million pairs; ~ 750 SNPs
- unrelated pair IBD distribution: (p0, p1, p2) = (1, 0, 0)

- Least powerful: IBS1 vs. IBS 0
- Less powerful: PPLINK IBD1 vs. IBD 0
- Most powerful: PREST IBD 1 vs. IBD 0

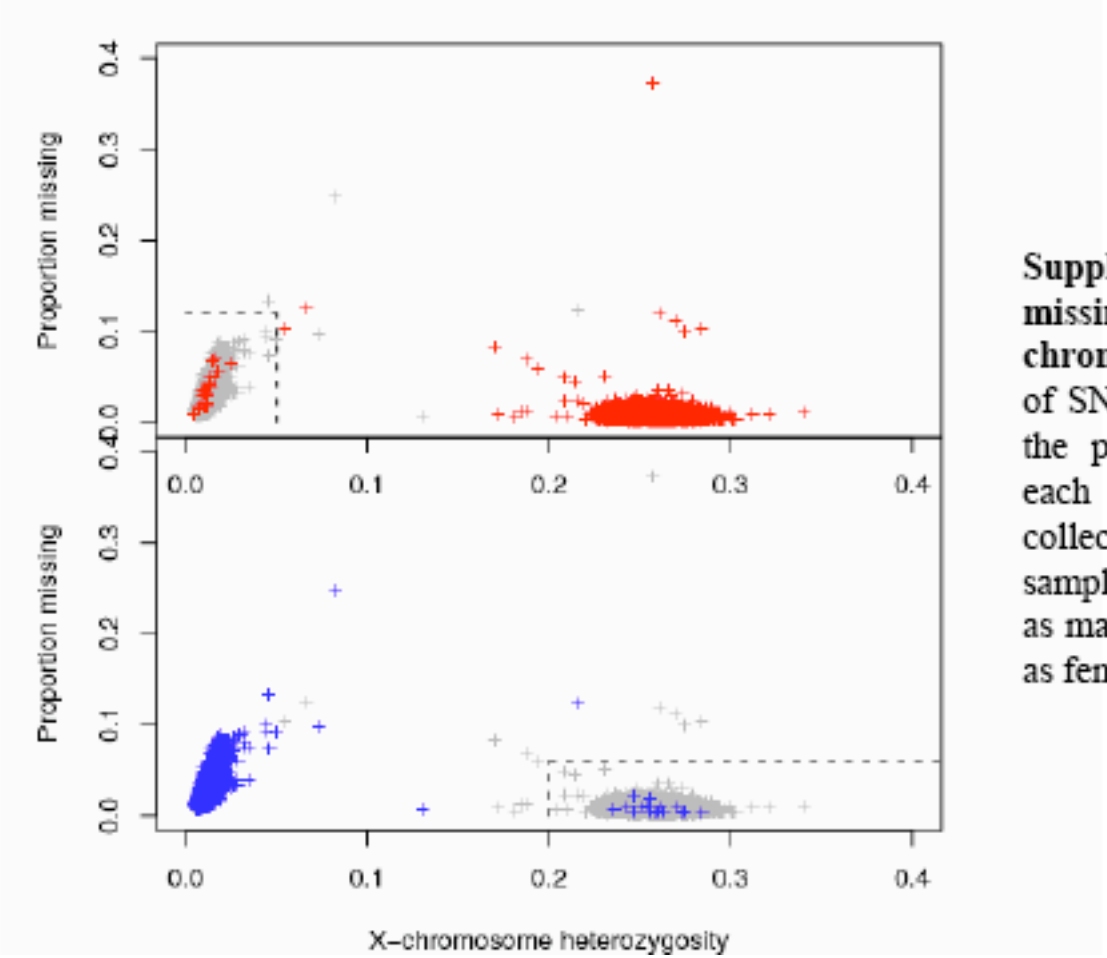






More QC - Samples

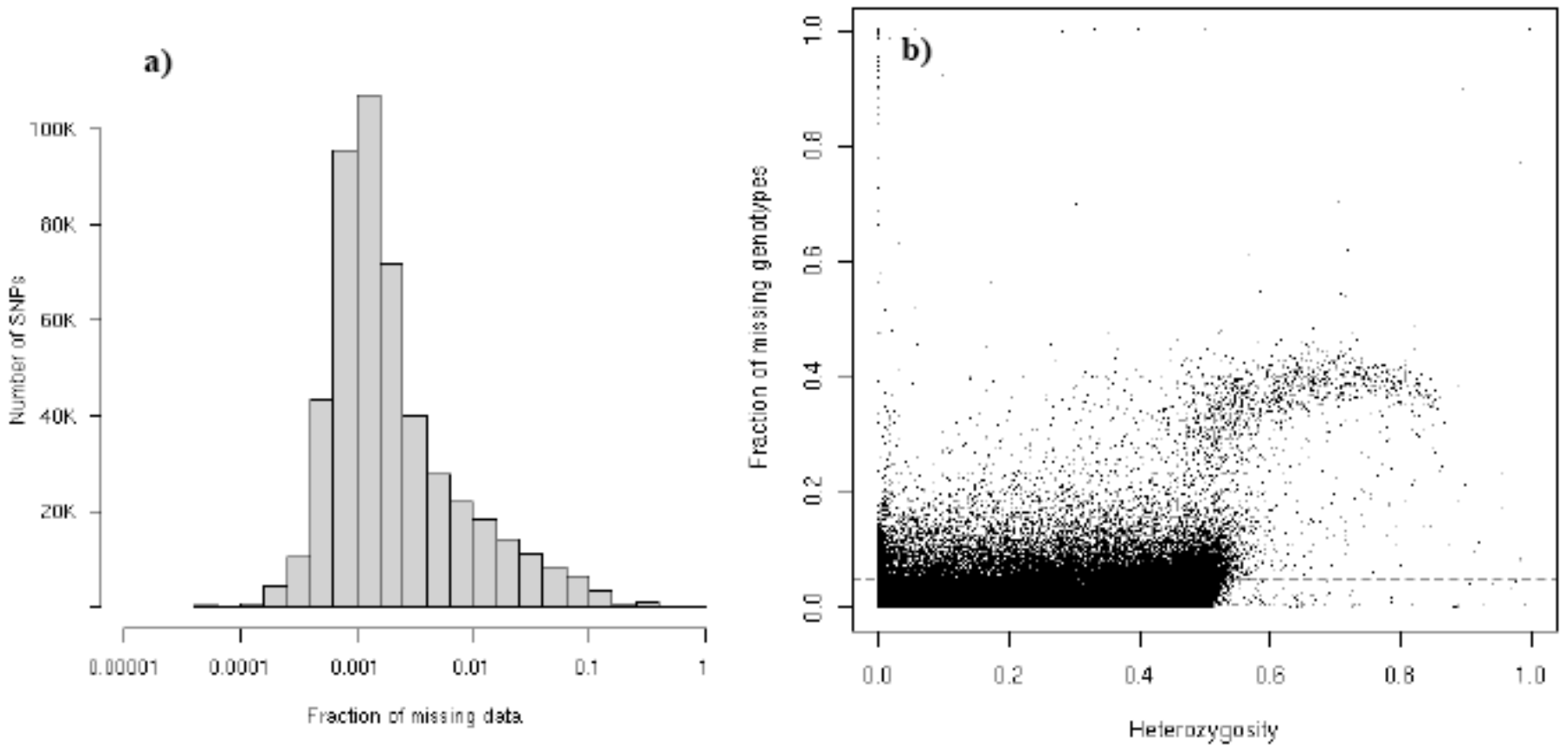
- Sex discrepancies
  - WTCCC: X-chromosome heterozygosity
  - Problematic ones were fed back to disease groups for verification
  - ~80 samples were not corrected but left in the study



Supplementary Figure 21 | Individual missing data and heterozygosity on the X chromosome. Scatter plot of the proportion of SNPs called heterozygote (x-axis) against the proportion called missing (y-axis) for each individual in the study. For each collection the individuals are plotted twice; samples whose gender were initially reported as male are coloured blue and those reported as female are coloured red.

QC - SNPs

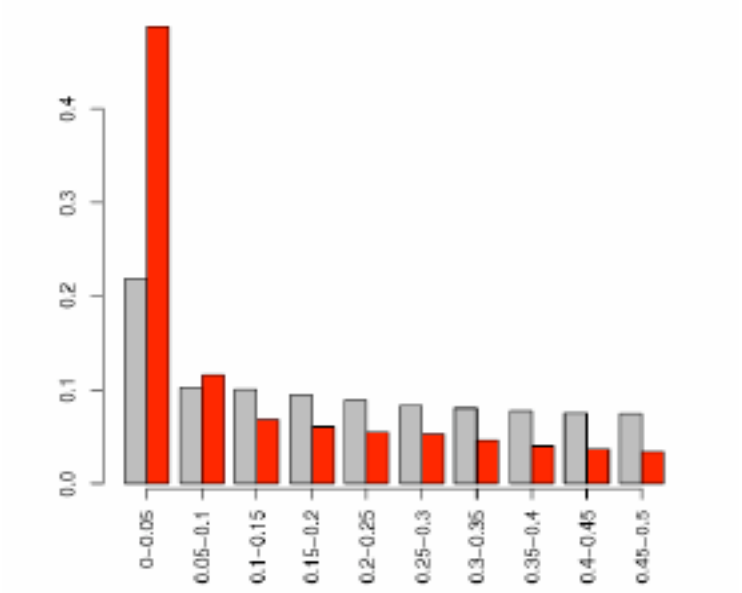
- Two factors: platform and genotype calling algorithm
- WTCCC: 3 filters
  - A study-wise **missing data rate (depends on the MAF):** >1% for SNPs with MAF < 5% and >5% for other SNPs (26,567 SNPs excluded)
  - **HWE exact test** p-value < 5.7×10^{-7} in the combined set of 2,938 **controls** (4,351 SNPs excluded)
  - **Association tests between the two control groups** p-value < 5.7×10^{-7} (either the 1 d.f. trend or 2 d.f. genotype) (93 SNPs excluded)
- WTCCC choice of cut-points: based on the empirical distributions of the statistics



Supplementary Figure 19 | Missing data and heterozygosity per SNP. a) Histogram of proportion of individuals called missing for each SNP (i.e with posterior probability < 0.9) b) Scatter plot of the proportion of individuals called heterozygote (x-axis) against the proportion called missing at a posterior probability threshold of 0.9 (y-axis) for each SNP assayed. The dotted line shows the threshold over which a SNP was excluded from further analyses.

QC - SNPs

- WTCCC SNP QC results
  - 500,568 SNPs
  - 31,011 excluded
  - 469,557 remained (93.8%)
    - 392,575 with MAF > 1%
    - 45,106 with MAF < 0.1%



Supplementary Figure 1 | Minor allele frequency (MAF) spectrum of SNPs included and excluded from the study. Red bars show the proportion of SNPs excluded from the study (see text and Methods) in 10 MAF bins. Grey bars show the frequency spectrum of included SNPs.

QC - Summary

- WTCCC filtering
  - **6 filters on samples**: missingness, heterozygosity, external discordance, MDS (non-European ancestry), duplicate, relatives - **95.2% samples remained**
  - **3 filters on SNPs**: missingness (depends on MAF), HWE exact test, association tests between two control groups - **93.8% SNPs remained**
- WTCCC approach: flagging
  - apply **relatively light QC filters**
  - then **subject all apparently associated SNPs to visual inspection** of genotype cluster plots (around 100 plots were assessed per disease)

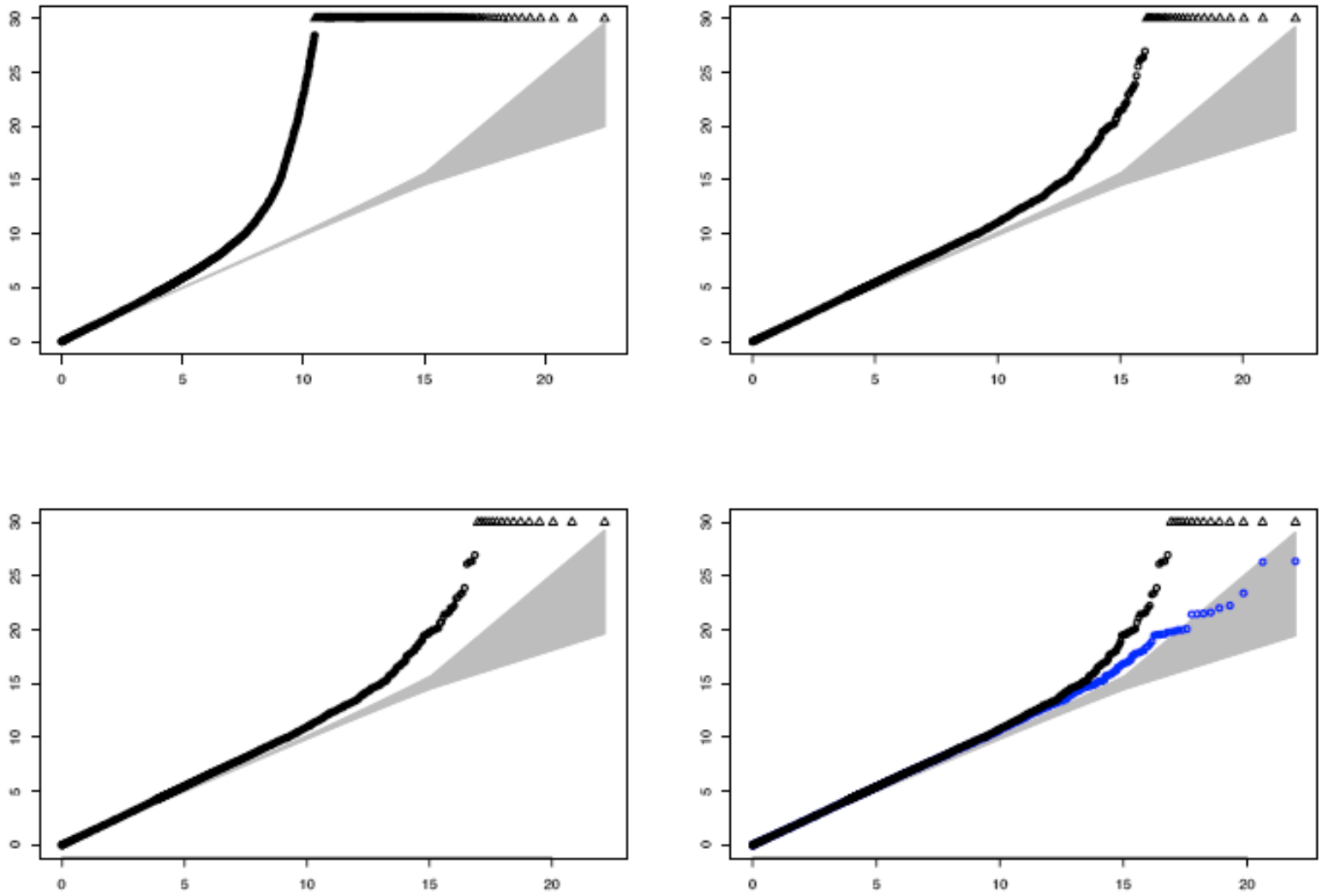
QC Summary

- Strike a balance
  - Stringency: may discard true signals **or generate spurious association with differential missingness**
  - Leniency: swamp the true signals with spurious association due to poor genotype calling

- QC can distort the association results
  - The large number of SNPs in the study means that even a small error rate or a small amount of poor quality data can lead to non-trivial numbers of false positives or negatives

	1000 SNPs	500K SNPs
Error Rate	Discarded SNPs	Discarded SNPs
.1%	1	500
1%	10	5000

  - QC does not end here: all significant association results are subject to visual inspection of cluster plots for the SNPs. (WTCCC: 638 SNPs were removed)



Supplementary Figure 20 | Quantile-quantile plots at four different stages of filtering. Each panel shows a QQ plot for the trend test results in T2D for the following subsets of SNPs (λ estimates for each subset in parentheses – see Methods for details), observed test statistics (y-axis) >30 are shown as triangles: top left, all SNPs (λ = 1.17); top right, SNPs passing standard project filter described in text and having minor allele frequency > 1% (λ = 1.09); bottom left, those passing the previous filter but excluding SNPs for which visual inspection of cluster plots revealed poor genotype calls (λ = 1.09); bottom right is as bottom left, but the plot which excludes regions with strong evidence for association is superimposed in blue, as in Main Figure 3.

Analyses Flowchart

- Shared controls
- Genotype calling
- QC of samples & SNPs
- **A number of association tests**
- Interpretation of the results
- Recommendations and guidelines

Association Tests

- PCA
- Trend test
- Genotype test
- Sex-differentiated tests
- Combine cases from ≥ 2 diseases
- Multi-locus method: Imputation
- X chromosome analysis
- Bayesian version of the above

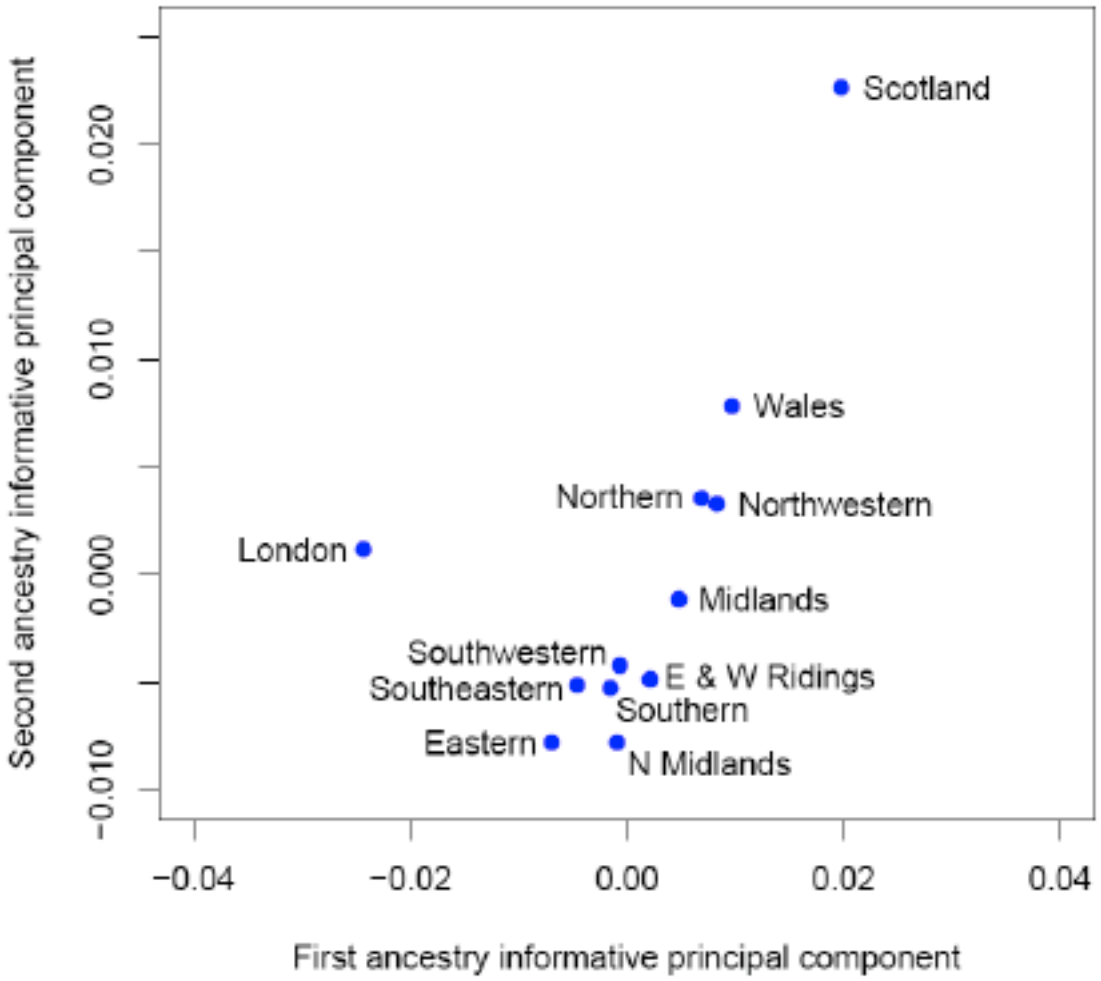
Association Tests - PCA

- Adjust for potential population stratification (connection with MDS?)
- PCA used 197,175 SNPs to reduce effect of LD

  (a clever way of doing it: separately for even- and odd-numbered chromosomes)
- Yet, 4 out 6 PC picked up local LD rather than genome-wide structure
- The remaining two show the same geo graphical trend from NW to SE
- WTCCC: PCA did not have noticeable effects
  - Main results reported without adjusting for PCA
  - But SNPs in regions showing strong geographical differentiation were flagged

	BD	CAD	CD	HT	RA	T1D	T2D
No covariates	1.11	1.07	1.11	1.06	1.03	1.05	1.08
PCA covariates	1.09	1.06	1.07	1.07	1.03	1.05	1.06

Supplementary Table 6 | Estimated over-dispersion of tests for association. Values greater than one indicate that the distribution of the test statistic is shifted towards larger values relative to the expected Chi-squared distribution. Values are given for the trend test (equivalent to a score test in an additive logistic regression model with no covariates), and for a test based on an additive logistic regression model in which the two ancestry informative principal components were included as covariates. Note that inclusion of the principal components does reduce overdispersion, but only by a small amount. The SNPs used in making these estimates were those that passed the filter described in the legend for Figure 3.



Supplementary Figure 8 | UK geographic population structure. Means, by geographical region, of the two principal components we judged to be informative of ancestral population admixture. Note that the means reflect the geographical configuration of the regions to some extent, confirming that the principal components are informative about geographical population structure. However, the distributions of principal component scores for individuals from each region overlap very extensively (data not shown), indicating that the population structure is very weak.

Association Tests - Trend and Genotype

- Trend test
  - 1 d.f.
  - Equivalent to a score test in an additive logistic regression model
- Genotype test
  - 2 d.f.
  - Fewer model assumptions
- WTCCC: similar results, a few differences

  (Table 3, 21 SNPs with either $p < 5 \times 10^{-7}$, 3 differ)

Association Tests - Sex-differentiated

- Perform the above two tests **separately for males and females**
- **Combine the results by adding the chi-squared statistics** leading to a 2 d.f. trend test and 4 d.f. genotype test
- Sensitive to association with different magnitude/and or direction in the two sexes
- Less powerful than the simple test when effect sizes does not vary with sex
- Table 3, 1 SNP < $5x10^{-7}$

Association Tests - ≥ 2 diseases

- Look for loci with effects in more than one disease
- Three new case groups
  - CAD+HT+T2D (metabolic and cardiovascular phenotypes with potential aetiological overlap, e.g. involving detects in insulin action)
  - RA+T1D (already known to share common loci)
  - CD+RA+T1D (all autoimmune diseases)
- Perform the above association tests for each group and the shared controls
- Table 3, 1 SNP with $p < 5x10^{-7}$ for RA+T1D

Association Tests - Imputation

- Multi-locus method utilizing LD
- Allow for association analyses on SNPs that are in the HapMap reference samples, but were not genotyped in WTCCC
- Genotypes of 2,193,483 SNPs were imputed for the WTCCC samples
- *"Genotypes are then tested for association **in the same way** (? More on this later) as SNPs genotyped in the project"*
- Table 3, 2 SNPs with $p < 5x10^{-7}$

Association Tests - Others

- X chromosome analysis
- Bayesian version of the above
  Similar results (Supp. Figure 22)

Association Tests - Summary

- A number of tests could be performed
- Which one(s) better?
  - Robustness to QC assumptions
  - Robustness to model assumptions
  - Asymptotic properties (extremely relevant given the large number of SNPs, MAF)
  - Power

Analyses Flowchart

- Shared controls
- Genotype calling
- QC of samples & SNPs
- A number of association tests
- **Interpretation of the results**
- Recommendations and guidelines

Interpretation of Results

- Flagging (further QC) is important
  - WTCCC: all signals are subject to visual inspection of the genotype cluster plots among others (Around 100 cluster plots were inspected for each disease, 638 SNPs were removed)
- Multiple hypothesis testing
  - Proper threshold to control false positives
  - WTCCC: 5×10^{-7} (debatable)
  - FWER vs. FDR
  - Utilize prior information
    - Roeder et al. (2006, AJHG, Weighted FDR)
    - Sun et al. (2006, Genet Epi, Stratified FDR)

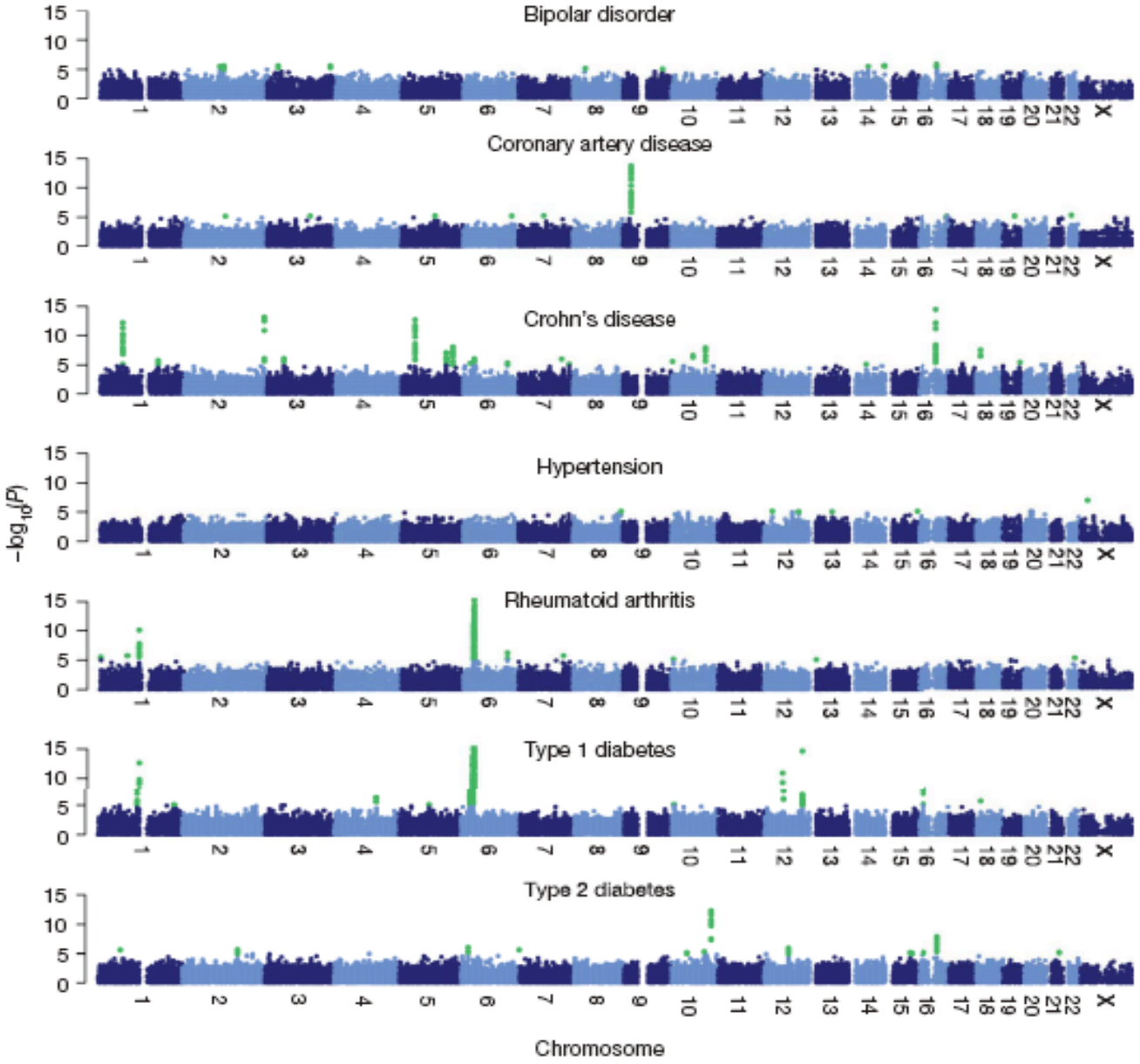
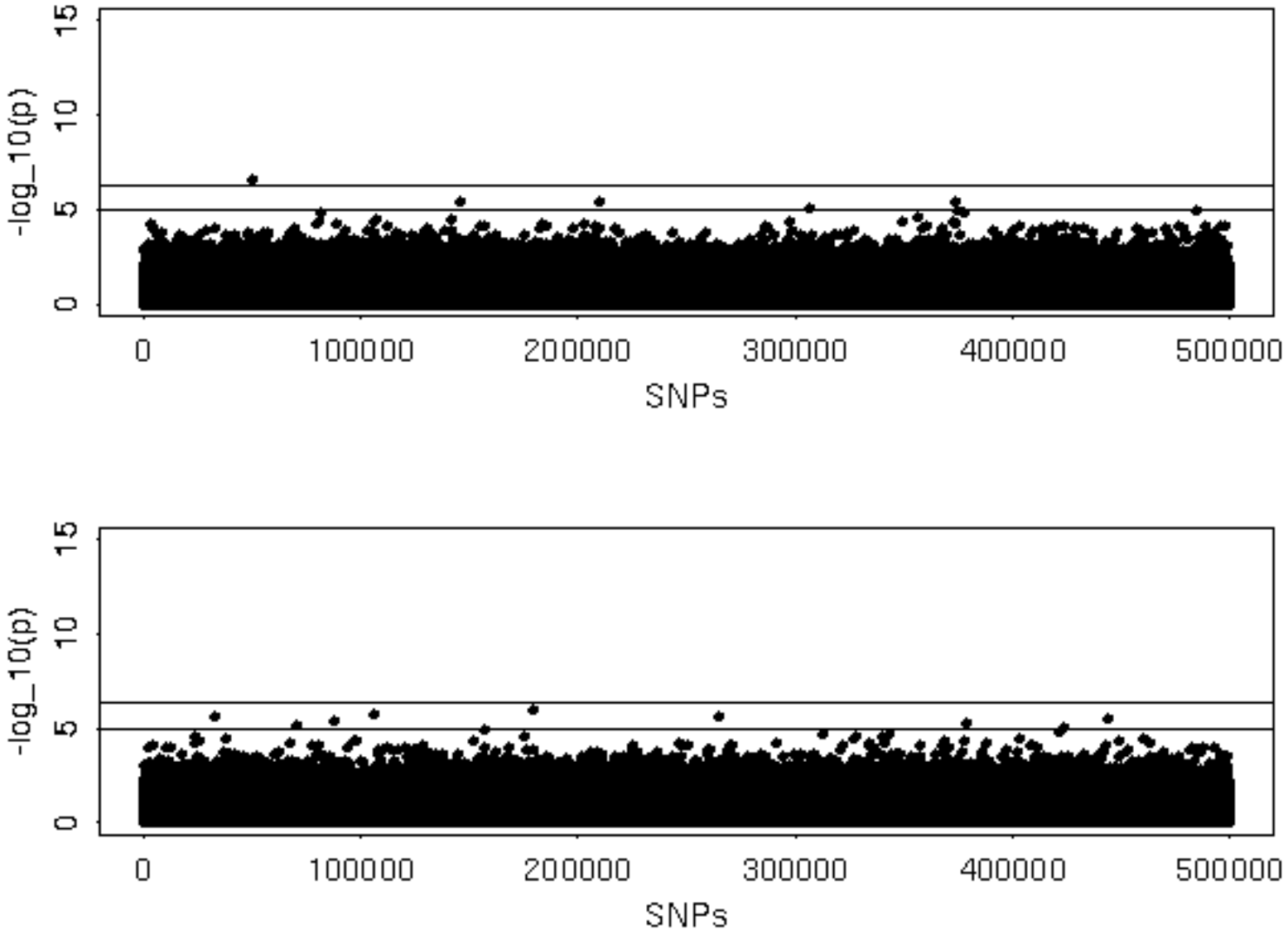


Figure 4 | Genome-wide scan for seven diseases. For each of seven diseases $-\log_{10}$ of the trend test P value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome. Chromosomes are shown in alternating colours for clarity, with P values $<1 \times 10^{-5}$ highlighted in green. All panels are truncated at $-\log_{10}(P\text{value}) = 15$, although some markers (for example, in the MHC in T1D and RA) exceed this significance threshold.

Signal plots of indep. 500K SNPs under the null (10^{-5}, $5x10^{-7}$; 2 indep. runs)



Interpretation of Results

- Multiple signals in a region
  - *"most susceptibility loci should show elevated signals at multiple nearby SNPs unless the r.f. in the region is very high or the SNP density very low."*

    "This effect should decline with …decreasing LD but also depends on factors like MAF."

    "A single elevated signals is often the spurious results of data or analysis artifacts, such as miscalled genotypes at a SNP"
  - In all WTCCC cases except one, nearby correlated SNPs also showed a strong signal (Figure 4)

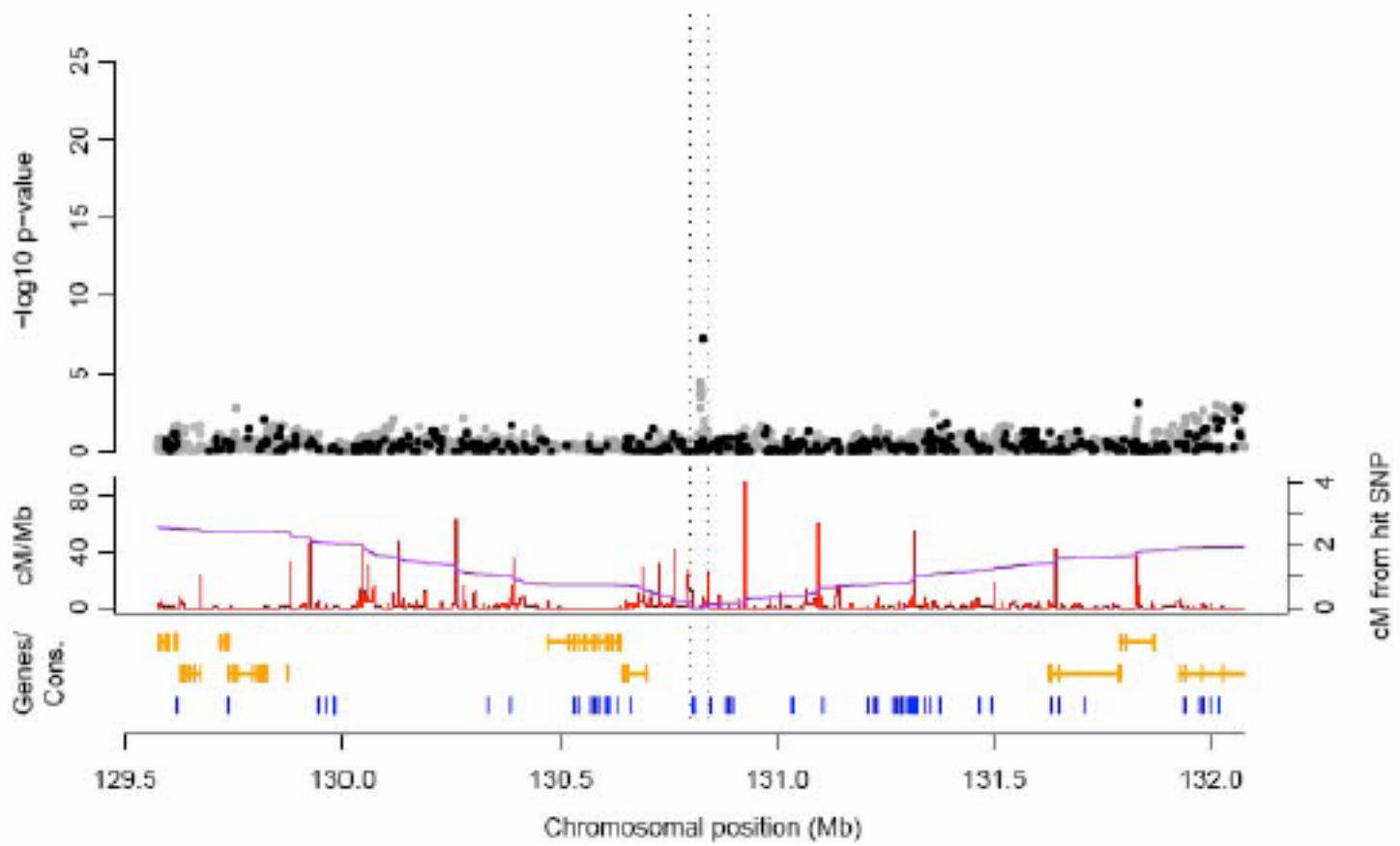
Interpretation of Results

- Replication: where do previously indicated/confirmed SNPs/loci stand?
  - WTCCC: all or most of the loci with $p < 5x10^{-7}$ are either known or have been subsequently confirmed (Table 2)
  - Indication of power
  - Failure to replicate a prominent association signal in previous studies?
    - Cannot conclusively exclude any give gene
    - Consequence of several factors: e.g. coverage, low power (small sample size, modest effect sizes)

Interpretation of Results

- Association ≠ Causality
  - A recent NIH FOA calls for *"statistical analyses…[that provides clues about the specific functional variants in these sets [containing both associated and causal variants]".*
  - Some previous work
    - Sun et al. (2002, AJHG, Binary Trait)
    - Blangero et al. (2005, Human Biology, QTL)

Interpretation of Results - Signal Plots



- Top: $-\log_{10}(p)$ for the trend or genotypic tests with the smallest p, gray dots: imputed SNPs
- Red: Fine-scale recomb. in cM/Mb (cM: expected # crossovers/100): high = recomb. hot spots
- Purple: cumulative distance in cM from the hit SNP at either side
- Bottom: known genes and sequence conservation in 17 vertebrates (more details here in the text)

- Vertical dotted lines: chosen to coincide with loci where test statistics returned to background levels; if this coincides with a recombination hotspot, as it often does, chose that as the boundary

Power Estimates

- Using a p-value threshold $5x10^{-7}$
- Averaged across SNPs with MAF > 5%
- Using 2,000 cases and 3,000 controls
- Power ~ 43% with RR 1.3
- Power ~ 80% with RR 1.5

Relative Risk	1.3	1.5	1.7
Power (p-value threshold $1x10^{-6}$)	0.461	0.813	0.91
Power (p-value threshold $5x10^{-7}$)	0.429	0.798	0.902

Supplementary Table 2 | Power of study design. Estimate of the power of the study, with 3,000 controls and 2,000 cases using for SNPs above 5% MAF in HapMap. See Methods section for details.

Size Does Matter

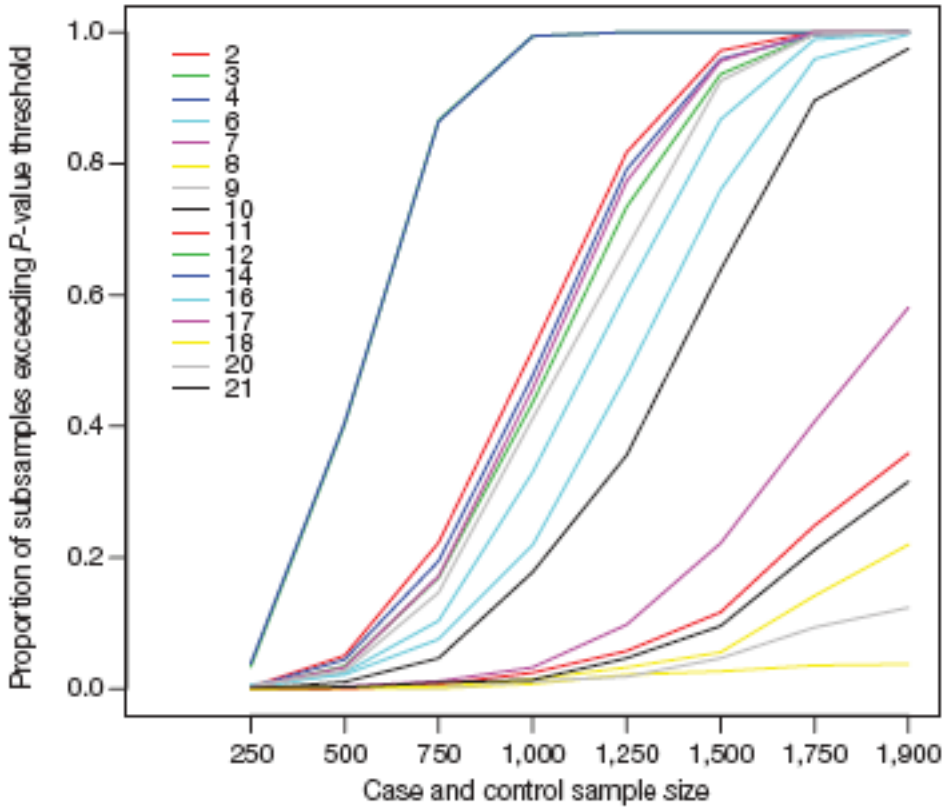


Figure 6 | Strong associations in subsamples of our data. For the 16 SNPs in Table 3 (outside the MHC) with *P* values for the trend test below 5×10^{-7}, we randomly generated 1,000 subsets of our full data set corresponding to case-control studies with different numbers of cases, and the same number of controls (*x* axis). The *y* axis gives the proportion of subsamples of a given size in which that SNP achieved a *P* value for the trend test below 5×10^{-7}. SNPs are numbered according to the row in which they occur in Table 3 (so that, for example, the CAD hit is numbered 2, and the *TCF7L2* hit on chromosome 10 for T2D is numbered 20).

All about QQ-Plots

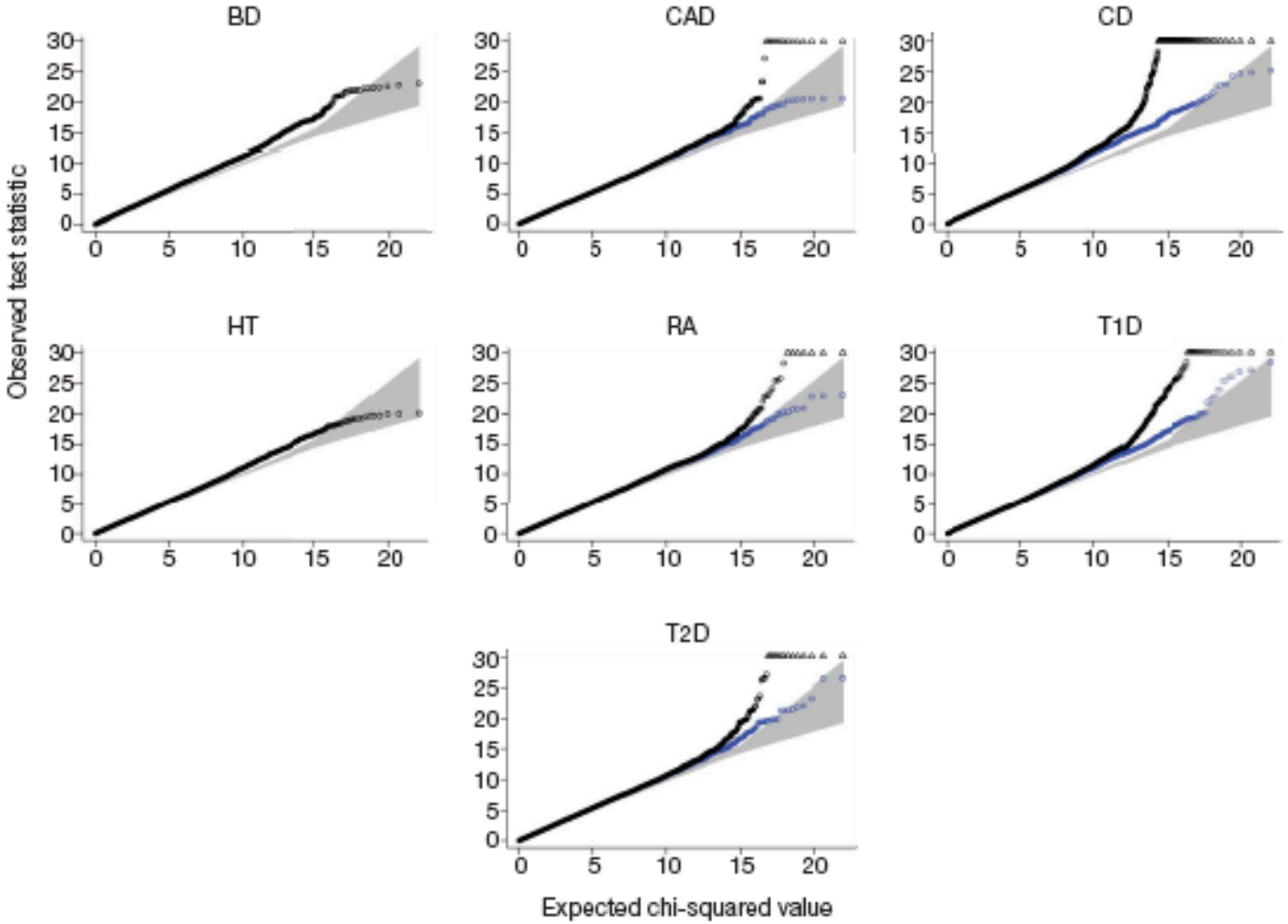
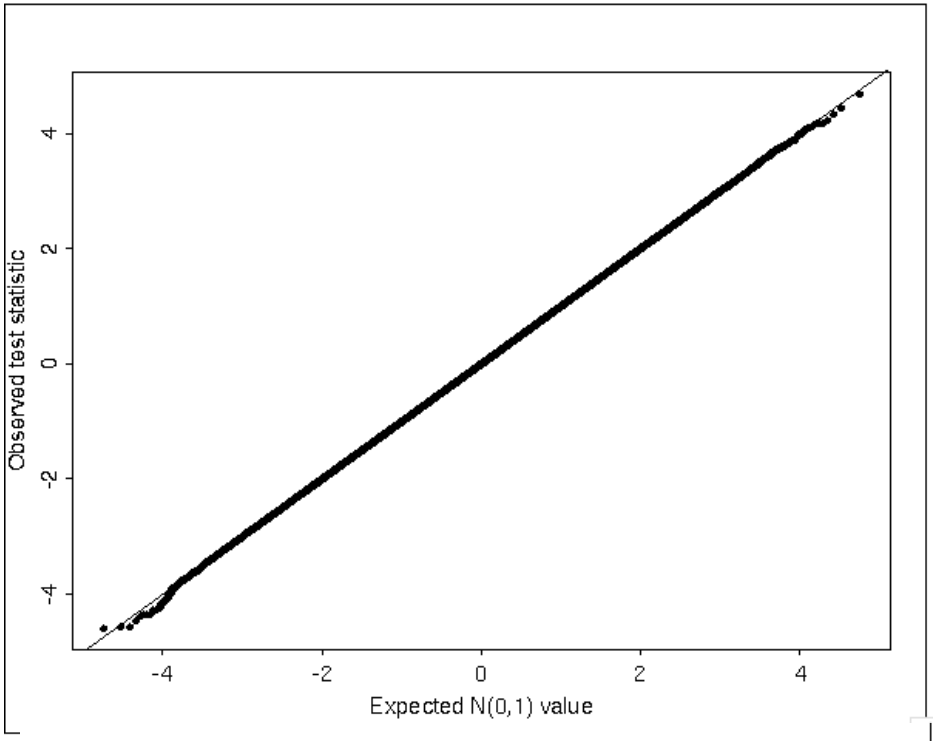


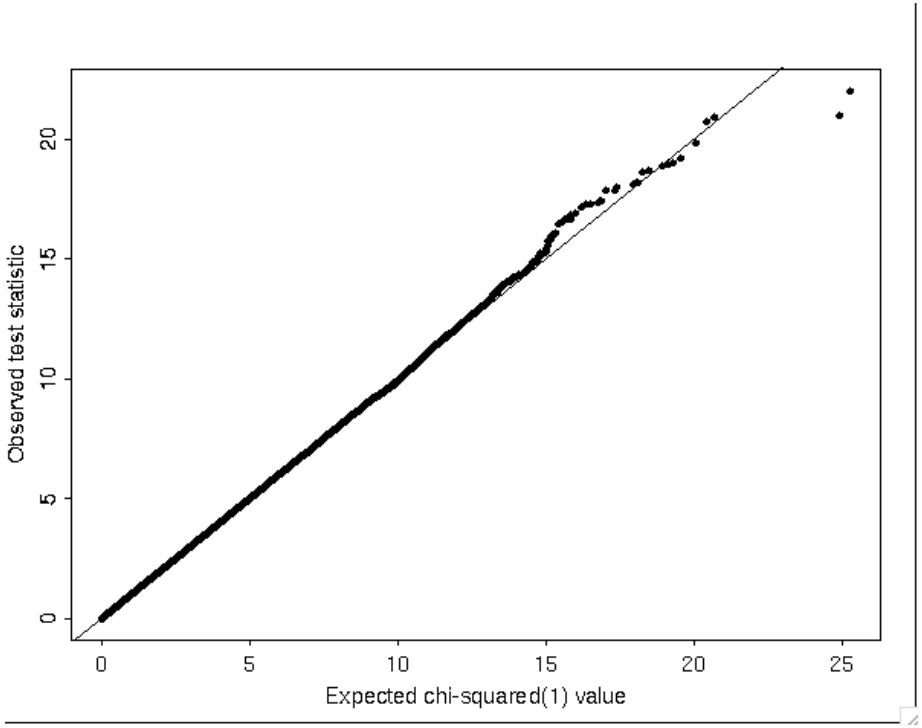
Figure 3 | Quantile-quantile plots for seven genome-wide scans. For each of the seven disease collections, a quantile-quantile plot of the results of the trend test is shown in black for all SNPs that pass the standard project filters, have a minor allele frequency >1% and missing data rate <1%. SNPs that were visually inspected and revealed genotype calling problems were excluded. These filters were chosen to minimize the influence of genotype-calling artefacts. Each quantile-quantile plot shown in black involves around 360,000 SNPs. SNPs at which the test statistic exceeds 30 are represented by triangles. Additional quantile-quantile plots, which also exclude all SNPs located in the regions of association listed in Table 3, are superimposed in blue (for BD, the exclusion of these SNPs has no visible effect on the plot, and for HT there are no such SNPs). The blue quantile-quantile plots show that departures in the extreme tail of the distribution of test statistics are due to regions with a strong signal for association.

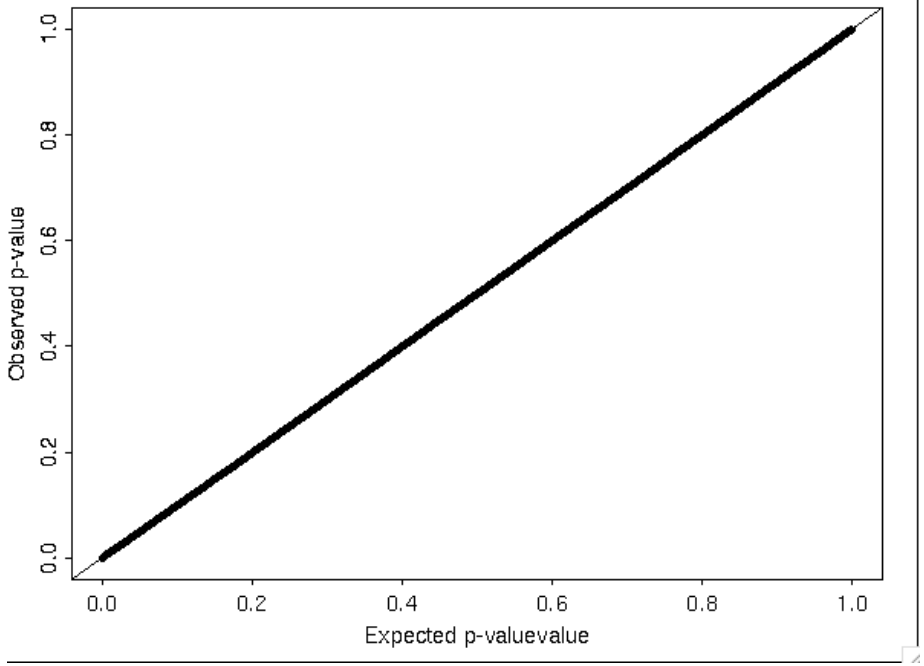
QQ-plot, **Effect of Scaling**

Same dataset with 500K SNPs (under the null, independent)

- **N(0,1)**
- **$N(0,1)^2$ = chisq with 1 d.f.**
- **Corresponding P-values**





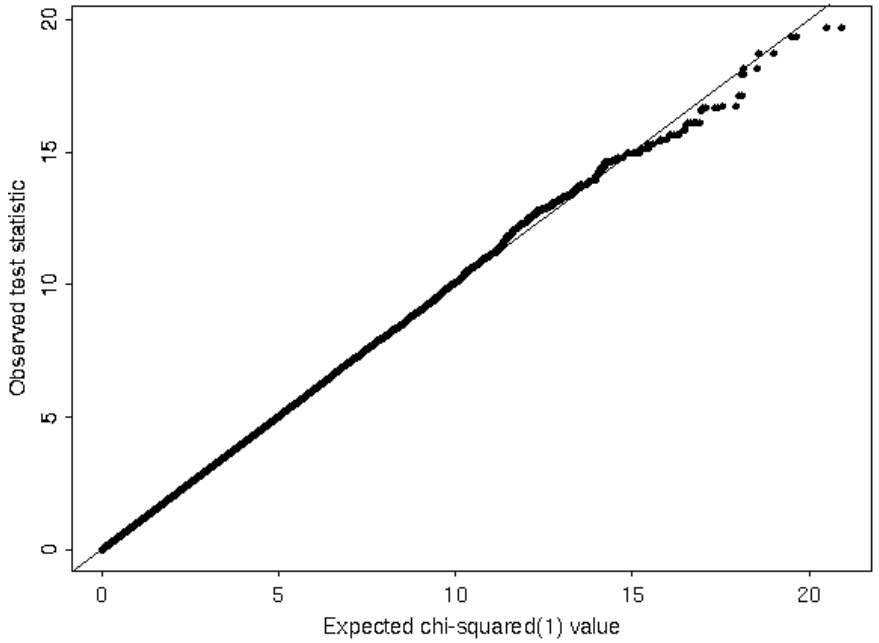


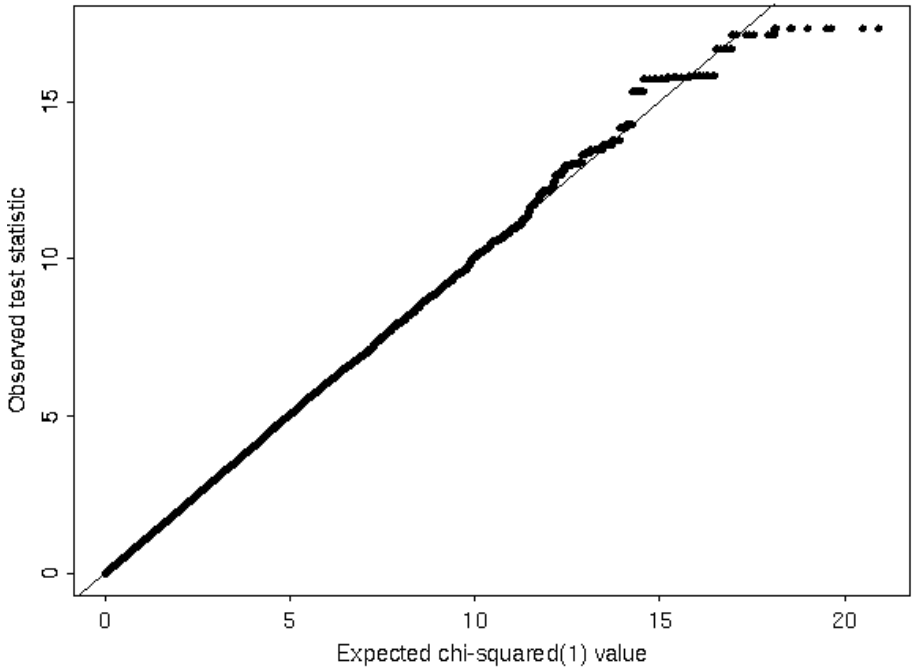
QQ-plot,**Effect of Correlation**

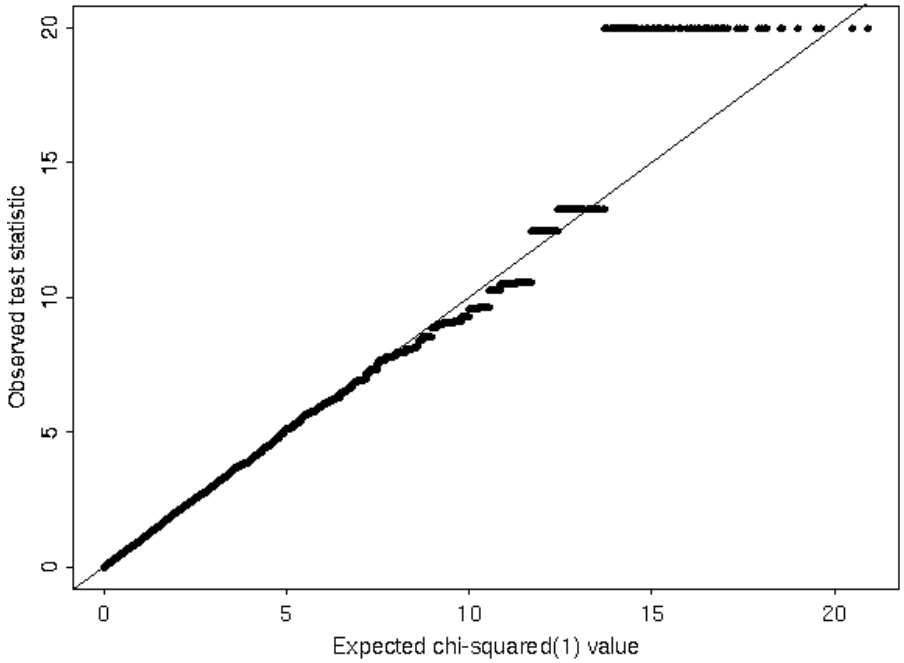
Total 500K SNPs under the null
Assume strong LD so that

- 250K indep SNPs,**each duplicated 2 times**
- 50K indep SNPs, each duplicated **10 times**
- 5K, indep SNPs, each duplicated **100 times**

- QQ-plot of the above
  correlated N(0,1)^2 vs. Indep. N(0,1)^2



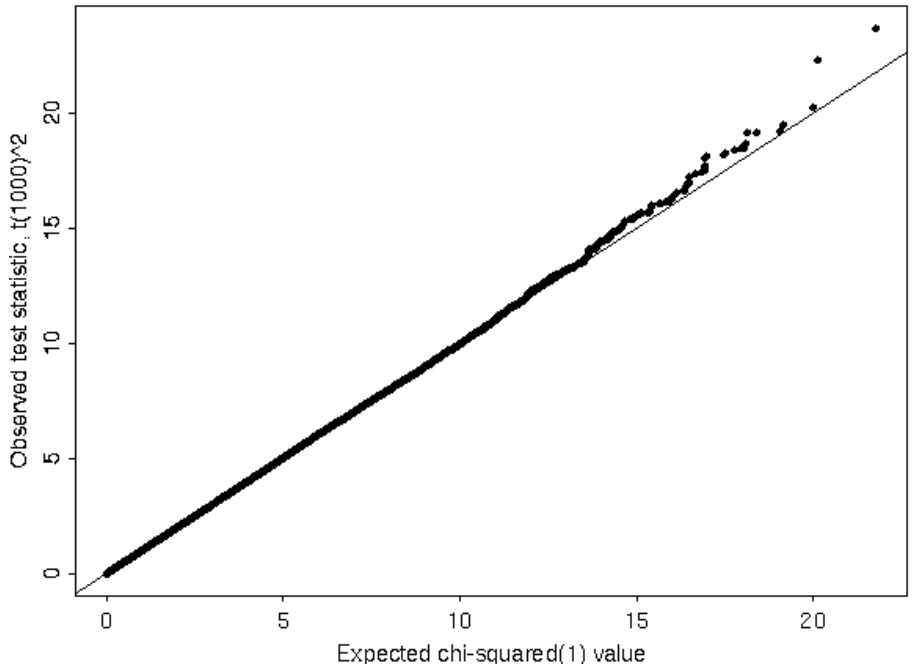


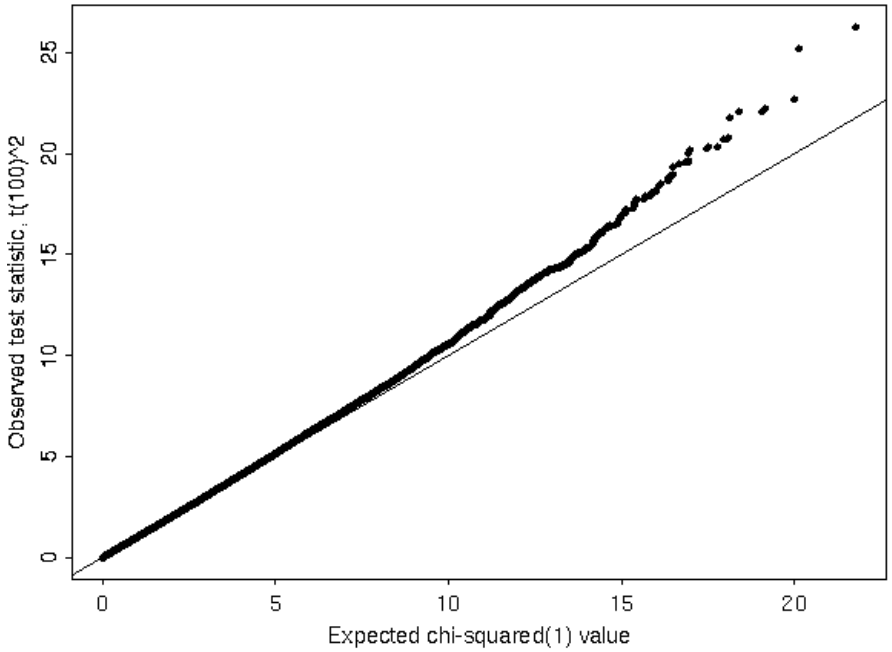


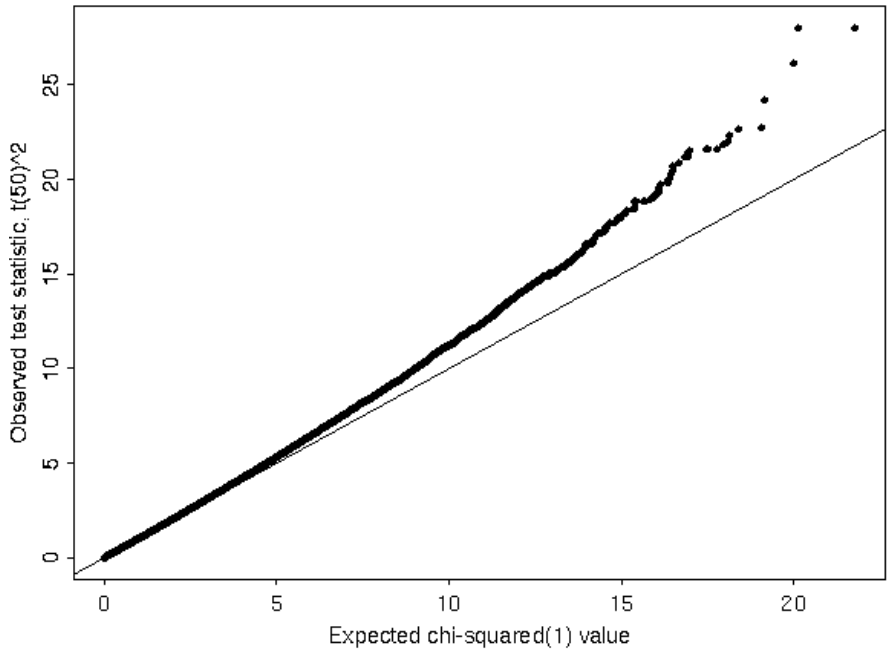
QQ-plot,**Effect of Model Misspecification**

Case I: asymptotic theory breaks down, e.g. normal approximation may not hold in the far tail (with 500K tests, some will be in the far tail!)

- 500K indep. SNPs under the null
- test statistic: **t with 1000, 100 or 50 d.f.**
- QQ-plot of **t^2 vs. $N(0,1)^2$=chisq(1)**





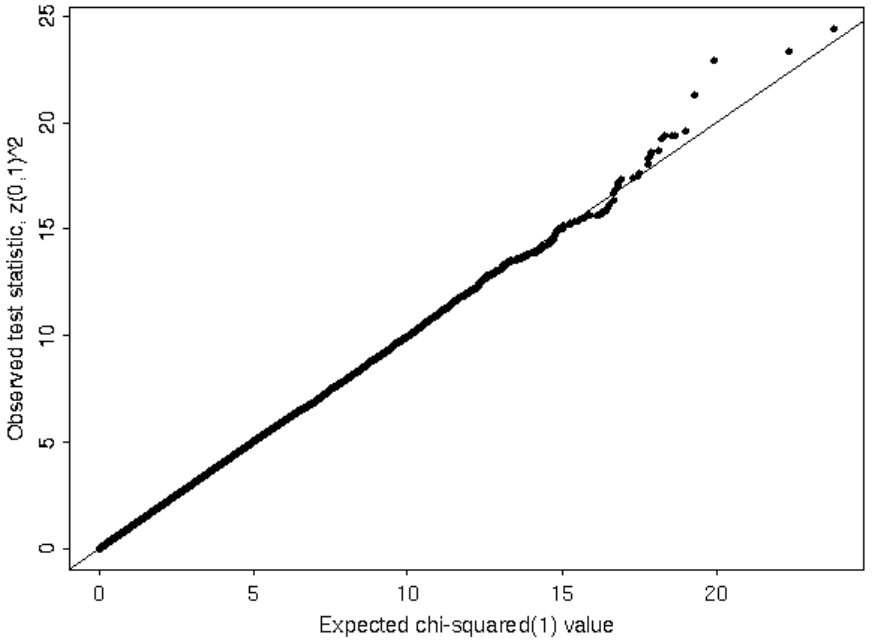


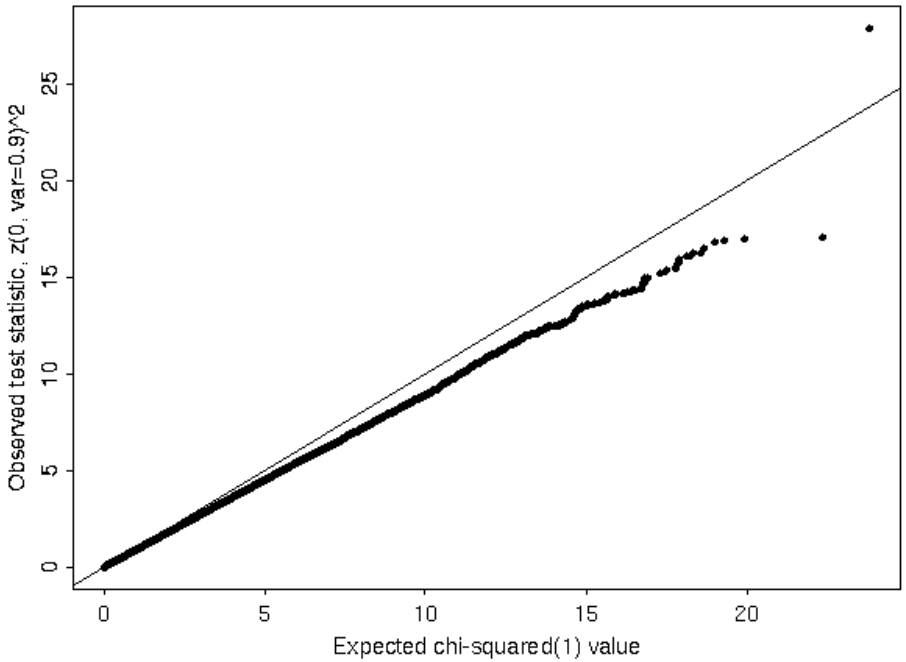
QQ-plot,**Effect of Model Misspecification**

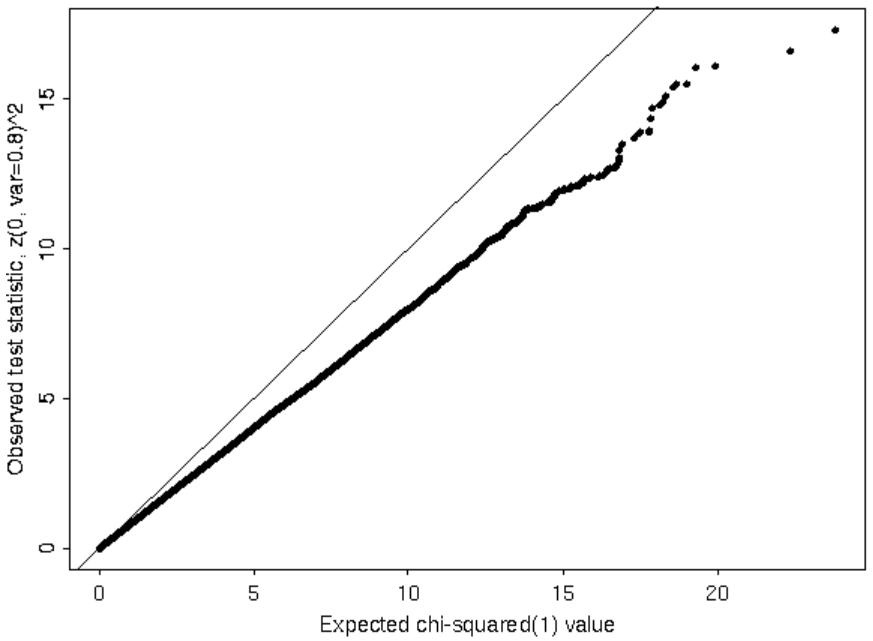
Case II: misspecification of variance, e.g. incorrect imputation methods may overestimate the true variance of the test statistics (more on next slide)

- 500K indep. SNPs under the null
- test statistic: **Z ~ N(0, var), var=1, 0.9, 0.8** (equivalent to 1-proportion of missing data in the example below)

- QQ-plots of **Z^2 vs. $N(0,1)^2$=chisq(1)**







Missing Data Imputation

- Déjà vu: recall linkage setting
  - Schork NJ, Greenwood TA (2004) Inherent bias toward the null hypothesis in conventional multipoint non-parametric linkage analysis. Am J Hum Genet 74:306-316
  - Many Letters to the Editor: missing information reduces power of detecting linkage but does not necessarily lead to invalid linkage tests
  - Proper test statistics must account for/model the missing information: **imputed data are NOT exactly the same as observed data**, the amount of departure depends on the amount of missing information

Missing Data Imputation

- A simple coin-tossing example
  - Null: $\pi = \pi_0 = 0.5$ (proportion of heads)
  - X = number of heads out of n tosses
  - Under the null,

$$S = \text{sqrt}\,(n/\pi_0*(1-\pi_0))*(X/n-\pi_0) \sim N(0, 1)$$

- Suppose data of some tosses were not observed: n_2 = n*prop.missing
  - X_1 = number of heads out of n_1
  - X_2 = **number of heads out of n_2 imputed under the null = $n_2 * \pi_0$**
  - What is the distribution of

$$S_{impute} = \text{sqrt}(n/\pi_0*(1-\pi_0))*((\mathbf{X_1+X_2})/n-\pi_0) \ ?$$

Missing Data Imputation

$$S_{impute} = \text{sqrt}(n/\pi_0*(1-\pi_0))*((\mathbf{X_1+X_2})/n-\pi_0)$$

$$\mathbf{S_{impute} \sim N(0,\ var = 1-n_2/n = 1-prop.missing \leq 1)}$$

- **Assuming $S_{impute} \sim N(0, 1)$ will overestimate the true variance of the test statistic,** and the amount of overestimation is proportional to the amount of missing data or missing information
- Related to Var(E[S|Data]) = Var(S) - E[Var(S|Data)]
- WTCCC: *"[Imputed] Genotypes are then tested for association* ***in the same way [figuratively speaking, not exactly the same]*** *as SNPs genotyped in the project" (*qq-plots of the test statistics of the 2,193,483 imputed SNPs are not directly provided)

Analyses Flowchart

- Shared controls
- Genotype calling
- QC of samples & SNPs
- A number of association tests
- Interpretation of the results
- **Recommendations and guidelines**

General Recommendations

The first relates to the importance of **careful quality control**. In such large data sets, small systematic differences can readily produce effects capable of obscuring the true associations being sought

The criteria used for SNP exclusion need to strike a compromise between stringency and leniency

As such, systematic visual inspection of cluster plots for SNPs of interest remains an integral part of the quality control process. The potential for population structure to undermine inferences

The potential for **population structure** to undermine inferences in case-control association studies has long been debated but limited empirical data have been available to assess the issue.

Our study highlighted several loci, some known and some new, which demonstrate substantial geographical variation in allele frequencies across Britain (Table 1), most probably due to natural selection in ancestral populations.

Outside these loci, the effects of population structure are relatively minor, and do not represent a major source of confounding, provided that individuals with appreciable non-European ancestry are excluded.

Although these conclusions may not generalize to studies in other locations, this finding reinforces the logistical and economic benefits of the case-control design over alternatives (such as family-based association studies).

Our study allowed us to address another important methodological issue: the adequacy, or otherwise, of using a **common set of controls**, rather than a sample recruited explicitly for use with a defined disease sample.

It is often assumed that failure to match cases and controls for socio-demographic variables will lead to substantial inflation of the type I error rate.

Our study demonstrates that, within the context of large-scale genetic association studies, for British populations at least, this concern has been overstated.

A related argument against use of population controls relates to the perceived impact of misclassification bias when a proportion of controls meet the criteria used to define cases. However, the consequent loss of power is modest unless the trait of interest is very common.

Given the above, the present study provides a compelling case for both the suitability and efficiency of the common control design in Britain and warrants its serious consideration elsewhere.

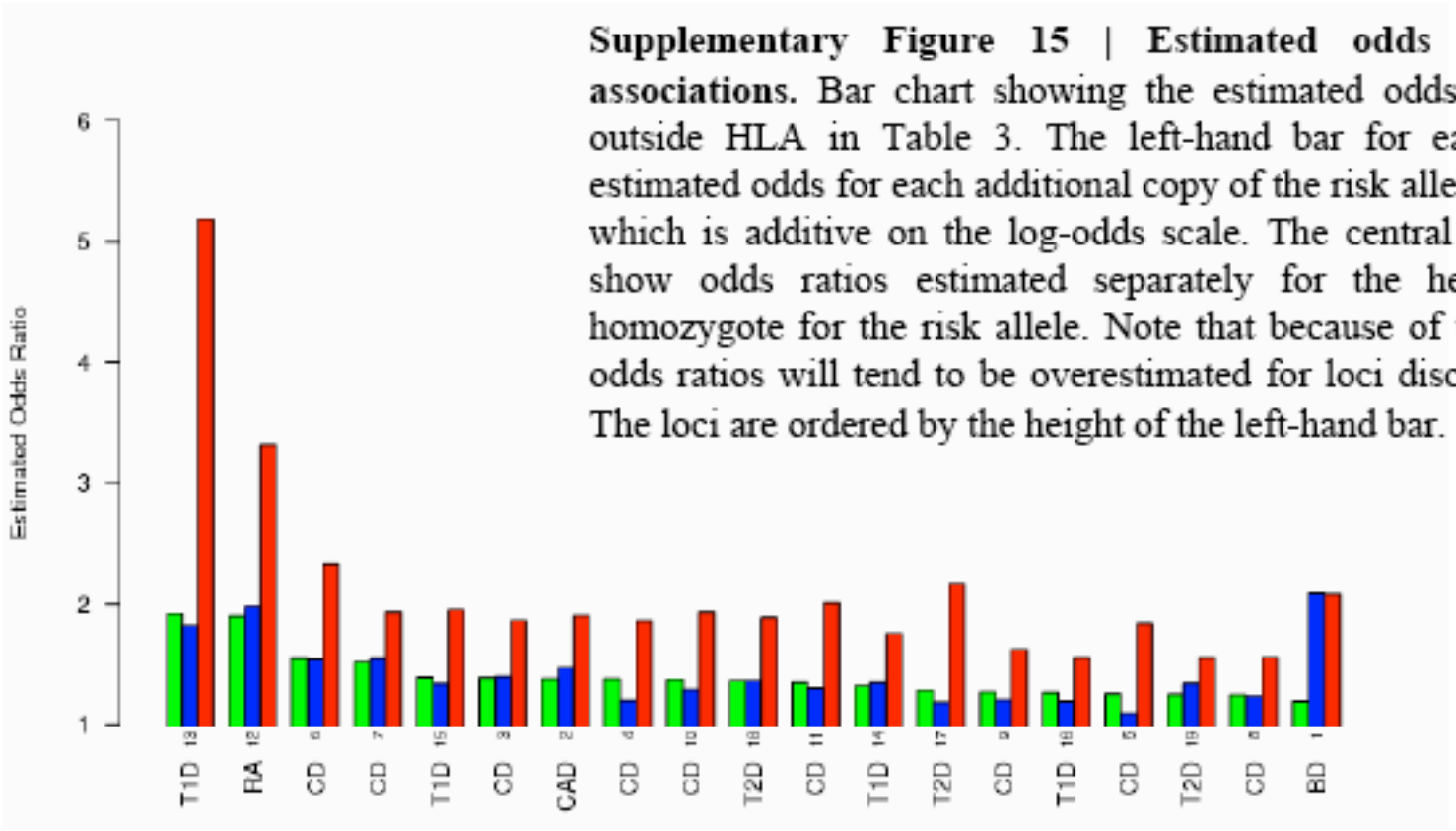
Finally, in failing to detect significant differences in performance between the epidemiological sample (58C) and that derived from blood donors (UKBS), we validate the use of the latter samples for cost-effective, large-scale control DNA provision

The novel variants we have uncovered are characterized by **modest effect size** (that is, per-allele **ORs between 1.2 and 1.5**) **and even these estimates are likely to be inflated [due to Winner's Curse/Selection Bias**].

We identified no additional common variants of very large effect (akin to HLA in T1D: Supplementary Fig. 15).

The observed distribution of effect sizes is consistent with models based on theoretical considerations and empirical data from animal models87,115,116 that suggest that, **for any given trait, there will be few (if any) large effects, a handful of modest effects and a substantial number of genes generating small or very small increases in disease risk.**

Winner's Curse/Selection Bias



Supplementary Figure 15 | Estimated odds ratios for strong associations. Bar chart showing the estimated odds ratio for the SNPs outside HLA in Table 3. The left-hand bar for each SNP shows the estimated odds for each additional copy of the risk allele in a disease model which is additive on the log-odds scale. The central and right-hand bars show odds ratios estimated separately for the heterozygote and the homozygote for the risk allele. Note that because of the "winner's curse" odds ratios will tend to be overestimated for loci discovered in our study. The loci are ordered by the height of the left-hand bar.

Methods for bias correction

- Sun and Bull (2005, Genet Epi, Resampling-based)
- Zollner and Pritchard (2007, AJHG, Likihood-based)