Shared and Distinct Genomics of Chronic Thromboembolic Pulmonary Hypertension and Pulmonary Embolism

Dr James Liley, *et al.*

ONLINE DATA SUPPLEMENT

Supplementary Methods

Sample details

We recruited CTEPH patients with western and central European ancestry from five European and one United States specialist pulmonary hypertension centres: Bad Nauheim (Kerckhoff Heart and Lung Centre, Bad Nauheim, Germany); Papworth (Royal Papworth Hospital, Cambridge, UK), Imperial (Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK), Leuven (KU Leuven - University of Leuven, Leuven, Belgium), San Diego (University of California, San Diego, USA), Vienna (Medical University, Vienna, Austria). CTEPH was diagnosed using international criteria (27) and patients were excluded if they had other major contributing factors to their pulmonary hypertension. Cases were recruited between 2011 and 2017. Centres supplied all available bio-banked samples that had been consented for genomic studies and were suitable for DNA extraction. Clinical details of samples are shown in table S1.

Cases	**Pre-QC**	**Final**
N	2467	1907
Male	49.7; 10.5 %	49.1; 13.1 %
Age	65.4 (53-74); 24 %	66 (54-74); 17 %
Height	172 (160-180); 86 %	172 (160-180); 85 %
Weight	80 (72-94); 86 %	80.7 (72-95); 85 %
MPAP	45 (36-53); 74 %	45 (36-52); 70 %
PVR	652 (360-930); 75 %	649 (370-930); 71 %
CI	2.4 (2-2.9); 96 %	2.4 (2-2.9); 95 %
CO	4.18 (3.3-5.1); 69 %	4.2 (3.3-5.1); 67 %
PCWP	11 (8-14); 72 %	10 (8-13); 70 %
6MWD	313 (210-390); 75 %	314 (210-390); 73 %
NYHA class	3 (3-3); 70 %	3 (3-3); 68 %

Controls	**Pre-QC**	**Final**
N	13247	10363
Male (%)	48.2; 0.034 %	48.7; 0.019 %
Age	45 (35-53); 88 %	45 (35-53); 86 %

Table S1: Clinical characteristics of case and control samples, format median (IQR) where appropriate. MPAP: mean pulmonary artery pressure; PVR: pulmonary vascular resistance; CI: cardiac index; CO: cardiac output; PCWP: pulmonary capillary wedge pressure; 6MWD: six-minute walk distance; NYHA: New York Heart Association class. Percentages after values give the proportion of missing or ambiguous values. Data on height, weight, MPAP, PVR, CI, CO, PCWP, 6MWD and NYHA class were available only for patients from UK clinics

In our discovery phase, we compared UK- and California- sourced CTEPH cases to 5984 healthy controls from the UK 1958 birth cohort and UK Blood Service. These samples were originally genotyped on the Affymetrix Axiom Genome-Wide CEU 1 Array, and we re-genotyped 1533 controls on the Illumina HumanOmniExpress Exome-8 v1.2 BeadChip which was used for cases.

In our replication phase, we compared non-UK non-California samples with 6717 UK- and European- samples from a recent GWAS on eosinophilic granulomatosis with polyangiitis (28). Although cases used in the replication dataset were exclusively non-UK, we found that inclusion of UK-sourced controls did not worsen inflation, so we did not restrict control samples to those not from the UK.

Genotyping, quality control and imputation

As above, our cohort consisted of Illumina-typed cases and controls and Affymetrix-typed cases which we genotyped and imputed, UK-based Affymetrix-typed controls which were previously genotyped, but we imputed, and Affymetrix-typed UK- and Europe- based controls which were previously genotyped and imputed. We were able to split the discovery phase into two separate analyses by platform type, but this was not possible in the replication phase as all controls were genotyped on an Affymetrix platform. Our quality control procedures diverted slightly between the discovery and replication phase.

Illumina samples were genotyped in four separate batches, and Affymetrix cases in a fifth. Genomic DNA was extracted and from whole blood or buffy coat fractions and quantified with ultraviolet-visible spectrophotometry (LGC, Hoddesdon, Herts, UK). DNA was normalised to a concentration of 50 *ng*/μ*L* and a total volume greater than 4 μ*L* (total DNA > 200 *ng*), which was required for the DNA microarray. Each batch of micro-array intensity data was normalised and clustered. Genotypes were called independently using Illumina GenomeStudio (v2.0) or the Affymetrix Genotyping Console (4.0). Samples containing more than 1 % missing genotypes were removed and SNPs were re-clustered. SNPs with poor clustering quality scores (GenTrain score (< 0.7) or clustering separation score (< 0.5)) were excluded following re-clustering. Genotyping procedures for the UK 1958 Birth cohort and UK NBS controls chip used in the discovery cohort are described in (25) and for controls used in the replication cohort in (28). We removed samples with heterozygosity rate more than 3 standard deviations from the batch mean or disparate reported and inferred sex. Across all samples including those genotyped, we assessed relatedness and removed one of any pair with > 30 % identity-by-descent, ensuring the absence of first-degree relatives in the dataset.

We then added two further batches: Affymetrix controls from the 1958 birth cohort, and Affymetrix controls from the UK NBS. Within each batch, we removed SNPs with minor allele frequency < 1 %, SNPs deviating from Hardy-Weinberg equilibrium with $p < 1 \times 10^{-5}$, and SNPs with missingness > 2 % or differential missingness between cases and controls ($p < 0.05$, Bonferroni corrected). We removed samples of divergent ancestry (separating by discovery and replication cohorts), assessed using principal components derived from the 1000 Genomes project (see section below).

We then combined all Illumina samples and all Affymetrix samples into separate combined batches for imputation, and imputed combined batches separately to genome-wide cover (Haplotype Reference Consortium (r1.1)) using the Sanger Imputation Server (9,10), pre-phasing with EAGLE2. Imputation details for replication controls are described in (28). We retained imputed variants with an INFO score of > 0.5 and a minor allele frequency of > 1 % in all three datasets.

We then separated all samples to be used in the replication phase. We combined remaining discovery-phase samples into two cohorts by genotyping platform (Illumina/Affymetrix). Since cases and controls in the replication phase were genotyped and imputed separately, and had somewhat different geographic distributions, we imposed further quality control measures on this cohort. We again removed SNPs with differential missingness between cases and controls ($p < 0.05$, Bonferroni corrected), removed SNPs for which any difference was detectable between batches ($p < 1 \times 10^{-6}$), removed SNPs with even slightly differing allele frequencies between UK controls in the discovery phase and UK controls in the replication phase ($p < 0.005$).

We then formed three separate cohorts for association testing. We split the discovery cohort by platform (Illumina/Affymetrix) but were unable to do this for the replication cohort, since all control samples were genotyped on an Affymetrix platform, so combined all replication case samples into a single cohort.

Assessment of divergent ancestry

Principal component analysis using a set of independent directly genotyped SNPs was used to identify samples with outlying ancestry. This was done separately in the discovery cohort (four Illumina batches, combined Affymetrix samples) and with all samples combined in the replication cohort. Samples were initially excluded if they did not cluster with super-populations from 1000 genomes data (29) PCA was then repeated, and samples that did not cluster with 1000 genomes European populations were excluded. Samples were excluded on the basis of distance from the relevant cluster median in standard-deviation units. Thresholds for exclusion were decided visually from each plot, but in no case were samples included if they were more than 3 standard deviations from the median on either the first or second principal component. In the replication cohort, in order to reduce genomic inflation to a reasonable level, we additionally excluded samples for whom the first five principal component values were at a Mahalanobis distance of > 10 from the population mean. Plots are shown in Supplementary Figure 6. Some residual differences can be seen between cases and controls in the replication cohort. Analyses were conducted in R using the snpRelate package (30).

Statistical analysis

We assessed association between cases and controls using a logistic regression for each cohort. In the discovery analyses, we used ten principal components as covariates, in which principal components were derived from genotyped SNPs only. In the replication cohort, in order to manage a wider geographical diversity, we used five principal component covariates derived from both our samples and European samples from the 1000 genomes data, as above.

We did not adjust for age or sex in our logistic regression models. Neither age nor sex can be associated with (autosomal) genotype, so neither can act as a confounder in our analysis. In some cases, adjustment for age or sex may strengthen GWAS associations due to younger individuals in the control cohort who would become cases were they older or of a different sex. However, given the rarity of CTEPH, this is unlikely in our case.

We evaluated genomic inflation in sets of p-values derived from each study. We evaluated both direct genomic inflation, termed λ, and the effective genomic inflation had the same allele frequencies been observed in a study of 1000 cases and 1000 controls, termed $\lambda_{1000,}$ defined as follows, in which n_0 is the number of control samples and n_1 the number of case samples:

$$\lambda = 1 + (\lambda - 1)\frac{\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}{\left(\frac{1}{1000} + \frac{1}{1000}\right)}$$

The genomic inflation factor for the replication cohort was moderately large (λ=1.21, λ_{1000} = 1.16) but we were unable to reduce it by inclusion of further covariates or by use of a linear mixed-model (BOLT-LMM (31)) in place of logistic regression. We thus simply corrected for inflation in each cohort by scaling χ^2 statistics (12).

We combined the two sets of p-values from the discovery cohorts into an overall discovery p-value, and all three sets of p-values into a set of meta-analysed p-values, using a standard z-score meta-analysis. Our criteria for genome-wide association are described in the results overview section above.

Levered analysis

Since CTEPH is phenotypically similar to DVT and PE, we expected that it would share some genetic associations. This enabled us to use results from large GWAS on PE and DVT to 'lever' our analysis on CTEPH to improve our power to detect CTEPH associations. Roughly, we do this by concentrating attention on variants more strongly associated with DVT or PE.

Define H^0_{CTEPH} as a null hypothesis of non-association of a variant of interest with CTEPH. The p-value in our CTEPH GWAS p_{CTEPH} gives us some information as to whether H^0_{CTEPH} holds. We may also be able to glean some information about H^0_{CTEPH} holds by considering the association of that variant with some other disease, measured by a p-value p_{OTHER} from an association study on other on separate samples. This will only be useful if the diseases tend to share the same associations. We use a procedure which both assesses degree of association sharing and tests assocaition with CTEPH in one, involving a quantity termed the conditional false discovery rate, or cFDR (14,32,33). In our case, the 'other' phenotype is PE, giving p-values $p_{OTHER} = p_{PE}$ (or p_{DVT}). We consider values (p_{CTEPH}, p_{PE}) as samples from the bivariate random variable (P_{CTEPH}, P_{PE})

A routine analysis rejecting H^0_{CTEPH} whenever $p_{CTEPH} < 5 \times 10^{-8}$ corresponds to a rejection subregion of the sample space of the (P_{CTEPH}, P_{PE}): specifically, the regions to the right of the dotted black lines in Figure 4a,4b. The cFDR replaces this with a data-driven rejection region (the regions to the right of the solid black lines in Figures 4a, 4b), which approximates the most powerful possible such region (14). It is roughly equivalent to firstly restricting attention to only SNPs for which $P_{PE} < \alpha$ for some α, concentrating associations with CTEPH.

We can then estimate the joint distribution of p-values for both CTEPH and DVT under the null hypothesis for H^0_{CTEPH} and integrate this over these data-driven rejection regions, giving 'v-values', which behave like p-values in having uniform distributions under H^0_{CTEPH}. These v-values can be thought of as p-values against H^0_{CTEPH} 'adjusted' for the additional information learned from the set of p-values for DVT association.

Differential effect sizes between CTEPH, DVT and PE

To determine whether the observed differential effect sizes at *F5* and *HLA-DRA* between CTEPH, DVT and PE reached significance (red lines on Figure 4a, 4b) we considered a null hypothesis that the underlying odds ratios of these variants were the same in both diseases.

If n_1, n_0, m_0, m_1, μ_1, μ_0 represent case/control numbers, observed case/control minor allele frequencies and population case/control minor allele frequencies respectively for a SNP of interest, then the Z score is approximated by

$$Z \approx \frac{\log\ (OR)}{SE\{\log\ (OR)\}}$$

$$\approx \frac{\log\left(\frac{m_1(1-m_0)}{m_0(1-m_1)}\right)}{\frac{1}{\sqrt{2}}\sqrt{\frac{1}{n_0 m_0} + \frac{1}{n_0(1-m_0)} + \frac{1}{n_1 m_1} + \frac{1}{n_1(1-m_1)}}}$$

$$\approx \frac{m_1 - m_0}{\sqrt{\mu_0(1-\mu_0)}}\sqrt{\frac{2n_0 n_1}{n_0 + n_1}}$$

assuming n_0 and n_1 are large, $\mu_0 \approx \mu_1$ and the SNP is diploid. Thus

$$E(Z) \approx \frac{\mu_1 - \mu_0}{\sqrt{\mu_0(1-\mu_0)}}\sqrt{\frac{2n_0 n_1}{n_0 + n_1}}$$

$$var(Z) \approx \frac{var(m_1 - m_0)\ 2n_0 n_1}{\mu_0(1-\mu_0)\ \ n_0 + n_1}$$

$$\approx 1$$

Variance in Z is due to random variance in the study population, and we assume it is independent between studies on independent traits. Thus, denoting n_1^{CTEPH}, n_0^{CTEPH} , n_1^{PE} , n_0^{PE} as case/control numbers in GWAS on CTEPH and PE respectively, under a null hypothesis that the effect size of the SNP is identical in both diseases, the joint distribution of Z scores (Z_{CTEPH}, Z_{PE}): will be bivariate normal with mean on a line through the origin with gradient

$$\sqrt{\frac{n_0^{PE} n_1^{PE}}{n_0^{PE} + n_1^{PE}}}\sqrt{\frac{n_0^{CTEPH} + n_1^{CTEPH}}{n_0^{CTEPH} n_1^{CTEPH}}}$$

and unit variance I_2. A multivariate normal with unit variance is invariant under rotation, so given n SNPs, the probability that at least one pair of Z-scores is at distance greater than D from the mean line is approximately

$$2n\Phi(-D)$$

where Φ is the Gaussian CDF function. Dotted lines on Figures 4a, 4b show distances D such that: under a null hypothesis that the effect size of all SNPs is identical for both diseases, the probability of at least one of the n SNPs reaching genome wide significance for either disease lying outside the dotted lines is < 0.05,

This is somewhat conservative, since Z scores are dependent due to linkage disequilibrium and the effective number of independent SNPs is less than n. Correspondingly, shortcomings of this approach include the possibility that geographic origin can affect relative effect sizes between GWAS and magnitude of linkage disequlibrium between SNPs, potentially confounding the relationship between different disease pathologies and different observed effect sizes in GWAS.

Rationale for tiers of association and choice of p-value threshold

We chose to reject null hypotheses of genetic non-association for variants which had either a meta-analysis p-value or a v-value less than 5×10^{-8}. This value is generally taken as an industry standard in genome-wide association studies, based on a Šidak correction to control family-wise error rate (FWER) at 5 % across a million independent common variants (34). Even after stringent quality control, it remains possible that confounding from population structure or batch effects could lead to a false positive rate exceeding 5 % in variants affected by this confounding. Since such confounding would be more likely to differentially affect either the discovery or replication cohort, we additionally required that in order to reject the null hypothesis, variants had to have a consistent direction of effect in both the Illumina- and Affymetrix- subcohorts of the discovery cohort and in the replication cohort. We also defined higher 'tiers' of association (as defined in the main manuscript) for variants for which we additionally had stronger evidence of a consistent effect in both the discovery and replication cohorts. Although variants in all three tiers reach a reasonable threshold for genome-wide significance, we have greater confidence that variants in tier 1 and tier 2 were not rejected due to confounding affecting only one cohort.

We investigated whether the genome-wide association threshold of 5×10^{-8} was appropriate in our case, since in practice, individual genotyping platforms have fewer than a million independent sites. We also aimed to assert that our approach (in which we rejected the null hypothesis for a SNP if the p-value *or* the v-value was less than the threshold) would not lead to an FWER in excess of 5 %.

We estimated the effective number of independent variants in our dataset as follows. for 500 simulations, we repeatedly ran a logistic regression, restricted to our final quality-controlled set of SNPs and a single chromosome (chromosome 10), and using a random phenotype (that is, we simulated under the null hypothesis). We also found the minimum v-value when conditioning on a second random phenotype. For each simulation i, we examined set $\{p_i\}$ of meta-analysed p-values and the set $\{v_i\}$ of v-values attained in each simulation, and recorded the minimum p- or v- value from the simulation as m_i=min($\{p_i\}$,$\{v_i\}$).

Given the set of such minimum values across 500 simulations, the log-likelihood for an effect number n_{10} of independent SNPs on chromosome 10 was

$$l(n_{10}) = \sum_{i=1}^{500} f_{\beta(n_{10},1)}(m_i)$$

where $f_{\beta(n_{10},1)}(m_i)$ is the density at m_i of a Beta distribution with parameters n_{10} and 1. since if we generate independent p-values for n SNPs for which the null hypothesis holds, the density of the minimum p-value p is $f_{\beta(n,1)}(p)$.

We found a maximum-likelihood estimate of 48941 effective independent variants amongst our 211705 total quality-controlled SNPs on chromosome 10. Presuming that the ratio of total variants to independent variants is roughly conserved across chromosomes, we estimate that our total dataset of 4090493 quality-controlled SNPs corresponds to approximately 945624 total independent variants (that is, slightly under a million). Under a Šidak correction to control FWER at 5 %, this suggests use of a threshold of 5.4×10^{-8} on the minimum p-value or v-value to reject the null hypothesis for a SNP.

It is reassuring that this threshold is slightly looser than the industry standard. However, we chose to revert to the stricter threshold of 5×10^{-8}, in the interests of consistency with the standard.

Power for tier 3 association

To approximate power to reject a variant at tier 3 significance given a z-score z_{other} at that variant for DVT or PE, we refer to the relevant plot in Figure 4a, 4b. The Z-score that must be obtained for CTEPH in order to reject the null hypothesis for CTEPH equivalent to a p-value $< 5 \times 10^{-8}$ corresponds to the x-co-ordinate of the intersection of the horizontal line at z_{other} with either the dotted or solid black lines (whichever x-intersection gives the smaller value). The minimum odds-ratio resulting in the requisite Z-score given can be routinely computed given the minor allele frequency and study size.

Assessment of batch effects

Samples were genotyped in several separate procedures (batches), and between-batch differences (batch effects) could have led to false positive results. The distribution of cases and controls across batches is shown in table S2. The absence of both control and case samples in some batches meant that such batch effects could not be directly differentiated from true case/control differences, and that batch numbers could not be included as covariates in the GWAS analysis.

	Dis. Illu. cont.	**Dis. Illu. case.**	**Dis. Aff. cont.**	**Dis. Aff. case.**	**Rep. cont.**	**Rep. case**	**All**
Batch 1	1492	369	0	0	0	0	1861
Batch 2	0	68	0	0	0	208	276
Batch 3	0	319	0	0	0	205	524
Batch 4	0	213	0	0	0	54	267
Batch Aff.	0	0	0	177	0	294	471
NBS	0	0	1293	0	0	0	1293
1958BC	0	0	2713	0	0	0	2713
Eur. cont	0	0	0	0	4865	0	4865
All	1492	969	4006	177	4865	761	12436

Table S2. Distribution across batches for cases and controls in the discovery and replication phases. Batches 1-4 used Illumina chips; all other batches used Affymetrix. Genotyping of the final three batches was performed by external groups. cont = control, rep = replication

The three areas of concern were 1. that in the Illumina-genotyped part of the discovery phase, batches 2-4 contained only cases; 2. that in the Affymetrix-genotyped part of the discovery phase, cases and controls were genotyped in separate batches; 3. that in the replication phase, cases and controls were genotyped in separate batches; and 4. that controls in the discovery phase were partially sourced from blood bank samples, which may drive the ABO association through differential distribution of ABO groups.

We address these problems by showing that at our discovered associations, allele frequencies are generally consistent across batches, allowing for case-control status. We also demonstrate that on a genome-wide scale, inter-batch effects are not detectable for each analysis (see section 'Between-batch comparisons'). We acknowledge that the presence of batch effects cannot be definitively ruled out, particularly for the Affymetrix-genotyped part of the discovery phase and for the replication phase.

Allele frequency at genome-wide associations

We computed allele frequencies across each batch for each genome-wide association in table 1, separating by case/control status. Across these nine associations allele frequencies in batches were generally consistent (Supplementary Figure 7).

We also note that the association at the ABO locus (chromosome 9) is not driven by the blood bank-sourced cohort (NBS); allele frequencies for the peak variant are consistent in the NBS cohort and 1958BC cohort, the latter of which, as a birth cohort, can be considered an unbiased population sample. Indeed, allele frequencies are consistent for the NBS and 1958BC cohort for all associations.

Between-batch comparisons

Where possible, we analysed whether allele frequencies differed systematically across batches within one of the three case/control comparisons. We compared allele frequencies amongst cases in batches 1-4 for the Illumina-genotyped discovery phase, amongst NBS and 1958BC controls in the Affymetrix-genotyped discovery phase (Supplementary Figure 8), amongst all batches (including between Affymetrix and Illumina) amongst cases in the replication phase, and between controls in the replication phase and NBS and 1958BC controls.

We compared allele frequencies at all variants using Fisher's exact test, and assessed whether the distribution of resultant p-values differed from the expected distribution of p-values should the observed batches represent identically-genotyped truly random samples from a common population. We removed 48 variants for which an allelic difference between batches had a p-value $< 1 \times 10^{-6}$.

We found that our results were consistent with equality of underlying allele frequency between each pair of compared batches using Q-Q plots of log-p values (Supplementary Figure 8). Moreover, in all cases, inflation of test statistics was < 1, and (after removing variants as above), there was no evidence of a surplus of low p-values, in that all Q-Q plots lay below the X-Y line.

Amongst our claimed CTEPH associations (Table 1), no variant reached a p-value $< 1 \times 10^{-3}$ in any comparison (Bonferroni-corrected significance threshold 4×10^{-4}). We were thus otherwise unable to detect any systematic or specific differences between batches after accounting for case/control status. We concluded that systematic batch effects were unlikely to be present, and SNP-specific batch effects were unlikely to be causing false-positive associations.

STREGA guidelines statement

Item	Item number	STROBE guideline	Extension for Genetic Association Studies (STREGA)	Reference section
Title and Abstract	1	(a) Indicate the study's design with a commonly used term in the title or the abstract		Title
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found		Abstract
Introduction				
Background rationale	2	Explain the scientific background and rationale for the investigation being reported		Introduction, paragraphs 1-2
Objectives	3	State specific objectives, including any pre-specified hypotheses	State if the study is the first report of a genetic association, a replication effort, or both	Introduction, paragraphs 3-4
Methods				
Study design	4	Present key elements of study design early in the paper		Results: methods overview: all paragraphs. Figure 1.
Setting	5	Describe the setting, locations and relevant dates, including periods of recruitment, exposure, follow-up, and data collection		Introduction: paragraph 4 Supplementary Methods: sample details: all paragraphs
Participants	6	(a) *Cohort study:* give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case–control study*: give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale	Give information on the criteria and methods for selection of subsets of participants from a larger study, when relevant	Results: Methods overview: paragraph 2 Supplementary Methods: sample details: paragraph 1

Item	**Item number**	**STROBE guideline**	**Extension for Genetic Association Studies (STREGA)**	**Reference section**
		for the choice of cases and controls *Cross-sectional study:* give the eligibility criteria, and the sources and methods of selection of participants		
		(b) *Cohort study:* for matched studies, give matching criteria and number of exposed and unexposed *Case–control study*: for matched studies, give matching criteria and the umber of controls per case		Results: Methods overview: paragraph 2
Variables	7	(a) Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	(b) Clearly define genetic exposures (genetic variants) using a widely-used nomenclature system. Identify variables likely to be associated with population stratification (confounding by ethnic origin)	RSID and CHR/BP used, with GRCh build specificied (eg Table 1). Supplementary Methods: Assessment of divergent ancestry Supplementary Methods: Statistical methods.
Data sources/ measurement	8[a]	(a) For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	(b) Describe laboratory methods, including source and storage of DNA, genotyping methods and platforms (including the allele calling algorithm used, and its version), error rates and call rates. State the laboratory/center where genotyping was done. Describe comparability of laboratory methods if there is more than one group. Specify whether genotypes were assigned using all of the data from the study simultaneously or in smaller batches	Results: Methods overview: paragraphs 2-4 Supplementary Methods: Genotyping, quality control and imputation: all paragraphs

Item	**Item number**	**STROBE guideline**	**Extension for Genetic Association Studies (STREGA)**	**Reference section**
Bias	9	(a) Describe any efforts to address potential sources of bias	(b) For quantitative outcome variables, specify if any investigation of potential bias resulting from pharmacotherapy was undertaken. If relevant, describe the nature and magnitude of the potential bias, and explain what approach was used to deal with this	Supplementary Methods: Assessment of batch effects: all paragraphs
Study size	10	Explain how the study size was arrived at		Introduction: paragraph 3
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	If applicable, describe how effects of treatment were dealt with	Not applicable
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding	State software version used and options (or settings) chosen	Supplementary Methods: Genotyping, quality control and imputation: paragraphs 2-5 Supplementary Methods: assessment of divergent ancestry.
		(b) Describe any methods used to examine subgroups and interactions		Supplementary Methods: Batch effects: all paragraphs
		(c) Explain how missing data were addressed		Supplementary Methods: Genotyping, quality control and imputation: paragraph 4
		Cohort study: if applicable, explain how loss to follow-up was addressed *Case–control study*: if applicable, explain how matching of cases and controls was addressed *Cross-sectional study*: if applicable, describe analytical methods taking account of sampling strategy		Supplementary Methods: Sample details: all paragraphs Supplementary Methods: Batch effects: Allele frequency at genome-wide associations: paragraph 2

Item	Item number	STROBE guideline	Extension for Genetic Association Studies (STREGA)	Reference section
		(e) Describe any sensitivity analyses		Supplementary Methods: Batch effects: Allele frequency at genome-wide associations
			(f) State whether Hardy–Weinberg equilibrium was considered and, if so, how	Supplementary Methods: Genotyping, quality control and imputation: paragraph 3
			(g) Describe any methods used for inferring genotypes or haplotypes	Supplementary Methods: Genotyping, quality control and imputation: paragraph 4
			(h) Describe any methods used to assess or address population stratification	Supplementary Methods: Assessment of divergent ancestry: all paragraphs Supplementary Methods: Statistical analysis: all paragraphs
			(i) Describe any methods used to address multiple comparisons or to control risk of false-positive findings	Results: Methods overview: paragraphs 5-7 Supplementary Methods: Statistical analysis: all paragraphs Supplementary Methods: Levered analysis: all paragraphs
			(j) Describe any methods used to address and correct for relatedness among subjects	Supplementary Methods: Genotyping, quality control and imputation: paragraph 2
Results				
Participants	13[a]	(a) Report the numbers of individuals at each stage of the study—e.g., numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analyzed	Report numbers of individuals in whom genotyping was attempted and numbers of individuals in whom genotyping was successful	Results: Methods Overview: paragraph 2 Results: GWAS on CTEPH: paragraph 1
		(b) Give reasons for non-participation at each stage		Supplementary Methods: Genotyping, quality control and imputation: all paragraphs

Item	Item number	STROBE guideline	Extension for Genetic Association Studies (STREGA)	Reference section
				Supplementary Methods: Assessment of divergent ancestry: all paragraphs
		(c) Consider use of a flow diagram		Figure 1
Descriptive data	14[a]	(a) Give characteristics of study participants (e.g., demographic, clinical, social) and information on exposures and potential confounders	Consider giving information by genotype	Table S1
		(b) Indicate the number of participants with missing data for each variable of interest		Supplementary Methods: Genotyping, quality control and imputation: paragraphs 2,3,5
		(c) *Cohort study*: summarize follow-up time, e.g., average and total amount		Not applicable
Outcome data	15[a]	*Cohort study*: report numbers of outcome events or summary measures over time	Report outcomes (phenotypes) for each genotype category over time	Not applicable
		Case–control study: report numbers in each exposure category, or summary measures of exposure	Report numbers in each genotype category	Table 1 (MAF) Supplementary Methods: Assessment of batch effects: Allele frequency at genome-wide associations
		Cross-sectional study: report numbers of outcome events or summary measures	Report outcomes (phenotypes) for each genotype category	
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95 % confidence intervals). Make clear which confounders were adjusted for and why they were included		Table 1 Figures 2-3 Supplementary Figures 4-5 Supplementary Methods: Statistical analysis Supplementary Methods: Assessment of divergent ancestry: all paragraphs

Item	Item number	STROBE guideline	Extension for Genetic Association Studies (STREGA)	Reference section
				Supplementary Methods: Statistical analysis: all paragraphs
		(b) Report category boundaries when continuous variables were categorized		Not applicable
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period		Not applicable
			(d) Report results of any adjustments for multiple comparisons	Results: Methods overview: paragraphs 5-7 Results: Co-analysis with DVT and PE Supplementary Methods: Statistical analysis: all paragraphs Supplementary Methods: Levered analysis: all paragraphs
Other analyses	17	(a) Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses		Results: Co-analysis with DVT and PE Results: Comparison with IPAH Results: Comparison of DVT, PE and CTEPH
			(b) If numerous genetic exposures (genetic variants) were examined, summarize results from all analyses undertaken	Results: GWAS on CTEPH Results: Comparison of DVT, PE and CTEPH Results: Comparison with IPAH
			(c) If detailed results are available elsewhere, state how they can be accessed	Please contact authors
Discussion				
Key results	18	Summarize key results with reference to study objectives		Discussion: paragraphs 1,2
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or		Discussion: paragraph 5

Item	Item number	STROBE guideline	Extension for Genetic Association Studies (STREGA)	Reference section
		imprecision. Discuss both direction and magnitude of any potential bias		Supplementary Methods: Assessment of batch effects: all paragraphs
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence		Discussion: paragraph 2
Generalizability	21	Discuss the generalizability (external validity) of the study results		Discussion: paragraph 4
Other information				
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based		Acknowledgements

Supplementary Figures

Supplementary Figure 1: Power to reject a null hypothesis of CTEPH non-association at tier 1 or 2 significance for a range of of minor allele frequencies in controls. For instance, if a SNP has a MAF of 0.4 in controls, and has an odds ratio of 1.4 between CTEPH cases and controls, we have approximately 20 % power to detect it at tier 1, and 40 % power to detect it at tier 2. Power calculations take account of the meta-analytic structure.

Supplementary Figure 2: Power to reject a null hypothesis of CTEPH non-association at tier 1 significance for a range of minor allele frequencies in controls and odds ratios. Colours correspond to power. As expected, we have greater power to detect variants at higher MAF and higher odds ratios. Power calculations take account of the meta-analytic structure.

Supplementary Figure 3: Power to reject a null hypothesis of CTEPH non-association at tier 2 significance for a range of minor allele frequencies in controls and odds ratios. Colours correspond to power. As expected, we have greater power to detect variants at higher MAF and higher odds ratios. Power calculations take account of the meta-analytic structure.

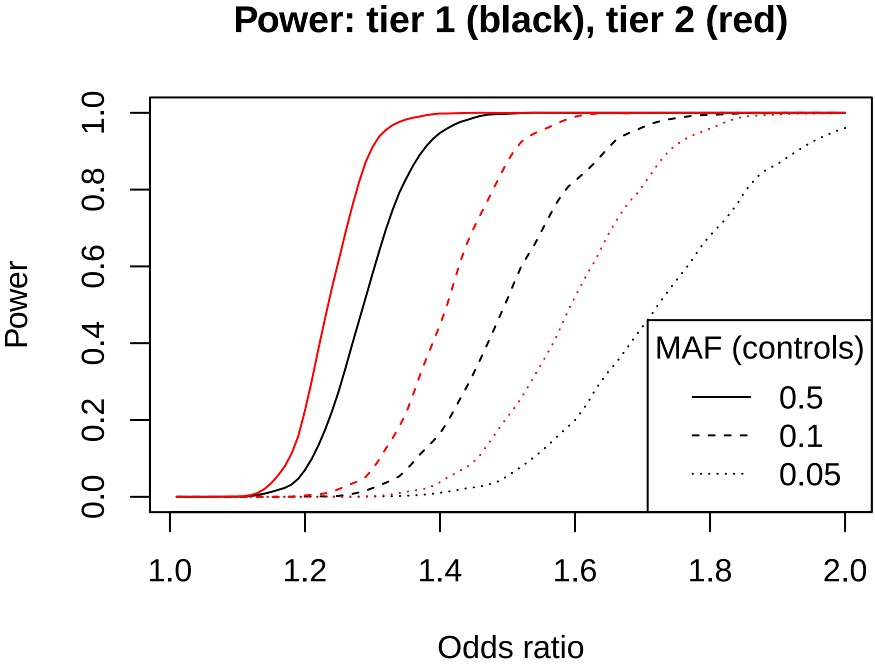
Supplementary Figure 4: Manhattan plot of $-\log_{10}(p)$-values from analysis of discovery cohort only. Points higher up correspond to variants more strongly associated with CTEPH. Variants reaching genome-wide significance ($P_{CTEPH} < 5 \text{ x } 10^{-8}$) are marked in black, and variants discovered using co-analysis with PE are marked in blue, both labelled with the likely associated gene. Note that most variants do not reach genome-wide significance when analysis is restricted to the discovery cohort. The black horizontal line denotes genome-wide significance ($p = 5 \times 10^{-8}$). Values of $-\log_{10}(p)$ larger than 16 are truncated to 16

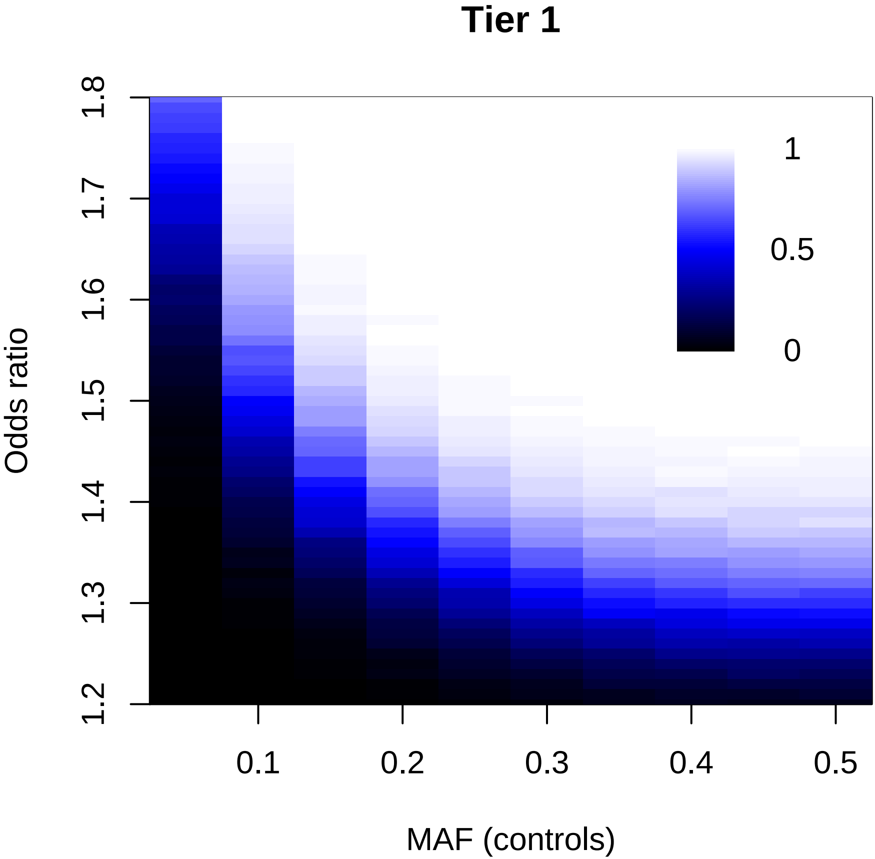
Supplementary Figure 5: Manhattan plot of $-\log_{10}(p)$-values from analysis of replication cohort only. Points higher up correspond to variants more strongly associated with CTEPH. Variants reaching genome-wide significance ($P_{CTEPH} < 5 \text{ x } 10^{-8}$) are marked in black, and variants discovered using co-analysis with PE are marked in blue, both labelled with the likely associated gene. Note that most variants do not reach genome-wide significance when analysis is restricted to the replication cohort. The black horizontal line denotes genome-wide significance ($p = 5 \times 10^{-8}$). Values of $-\log_{10}(p)$ larger than 16 are truncated to 16

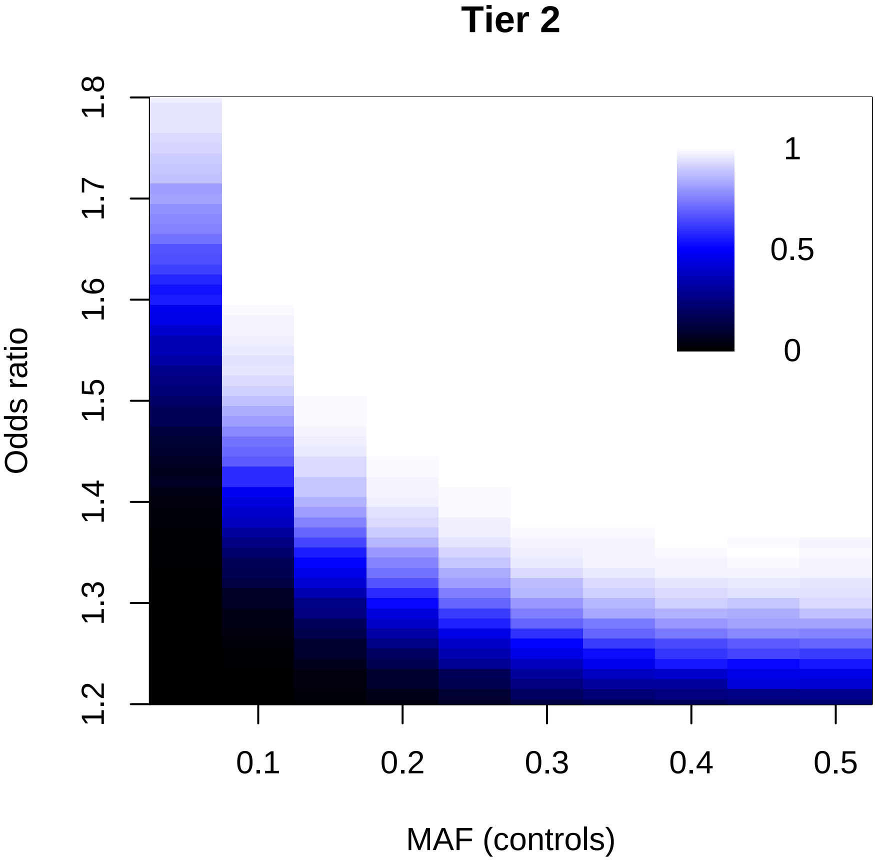
Supplementary Figure 6: Principal components of genetic samples combined with 1000 Genomes (1KG) samples. Leftmost plots show principal components including all 1KG samples, middle plots including all European 1KG samples, and rightmost plots including all European 1KG samples after exclusions. Black lines indicate exclusion boundaries. Cases are marked in black, and controls in red. Some cases and controls can be seen to cluster with East Asian or African 1KG samples, and some are widely aberrant and isolated (likely due to widespread genotyping errors).

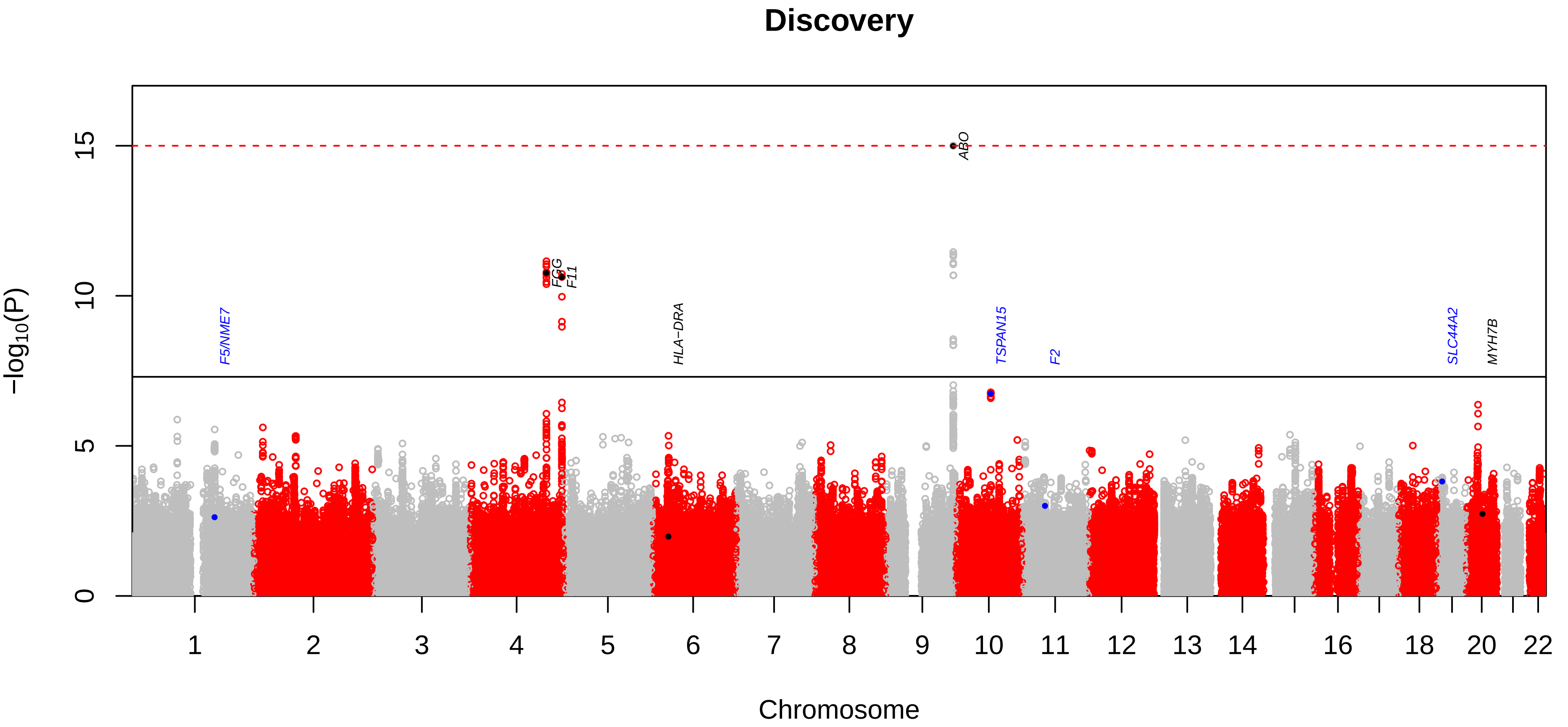
Supplementary Figure 7. Allele frequencies across batches at peak SNPs in Table 1. Horizontal lines show average allele frequencies, and vertical lines show 95 % confidence intervals. In general, observed allele frequencies in cases and controls are consistent with equal underlying case and control allele frequencies in each batch.

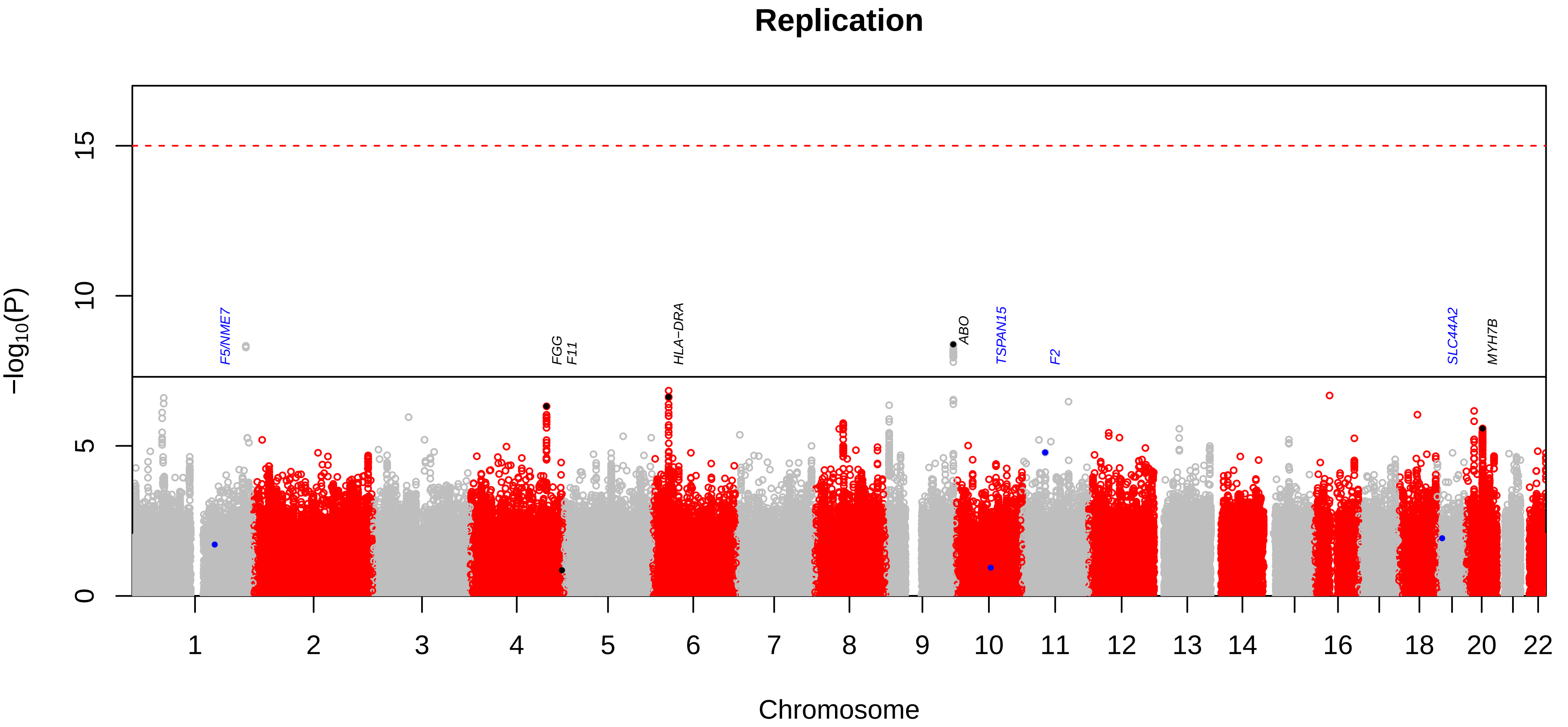
Supplementary Figure 8. Q-Q plot for genome-wide p-values for between-batch comparisons. In each case, we compare the allele frequency for each variant in our final dataset between two batches, and compute p-values using Fisher's exact test. We then consider the distribution of $-\log_{10}(p)$ values against the distribution of $-\log_{10}(p)$ values we would expect to see if there were no differences in underlying allele frequency between batches. If any p-values are lower than what would be expected in this case, they would correspond to the black line lying above the red X-Y line. Figures on the graph show 'inflation', analogous to genomic inflation; values above 1 indicate that p-values are generally lower than expected. Since all black lines stay below the X-Y line, and all between-batch inflation values are less than 1, we conclude that our data (following quality control) show no evidence of between-batch differences in allele frequency.

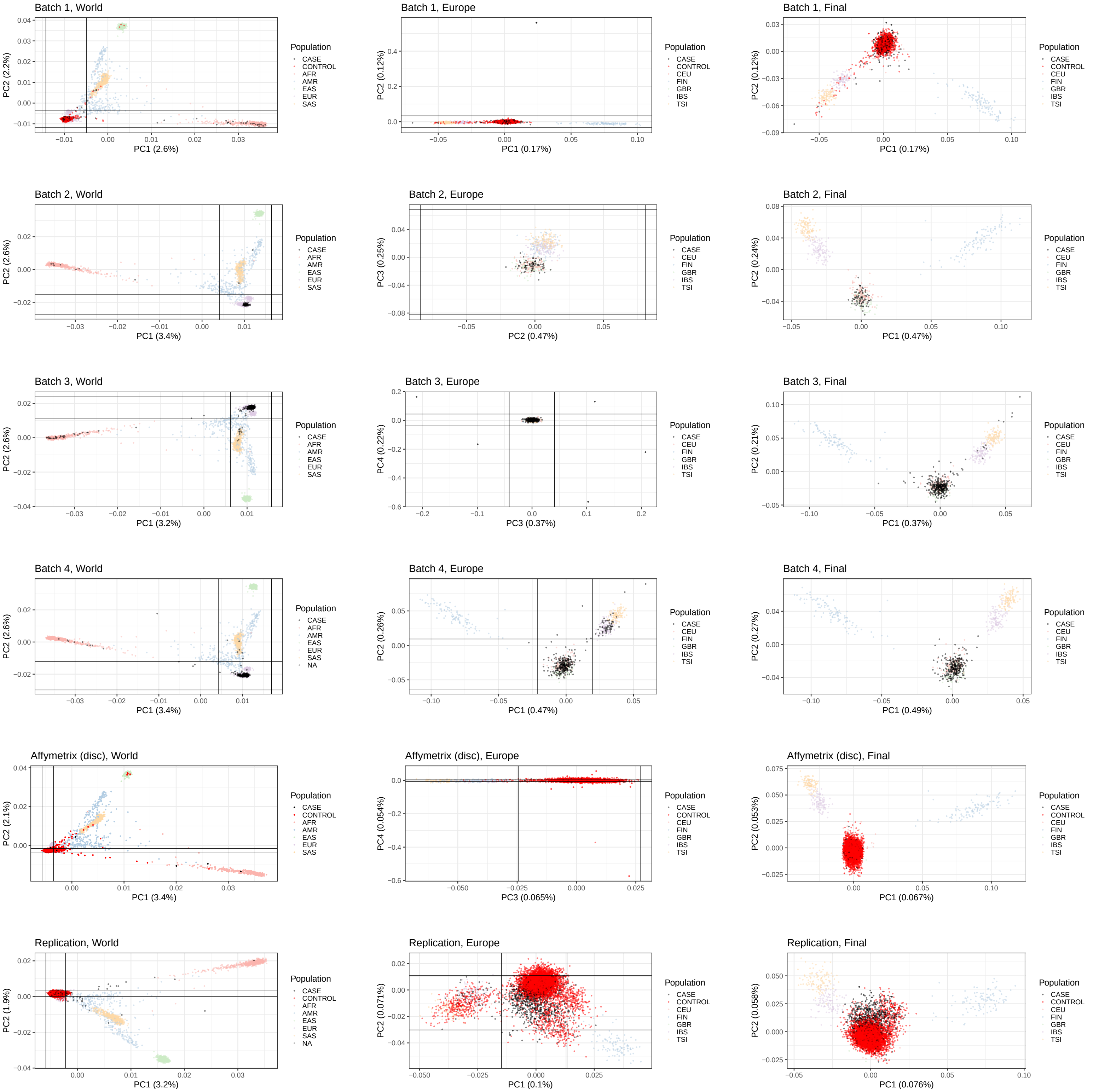

Power: tier 1 (black), tier 2 (red)
Power
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.2
1.4
1.6
1.8
2.0
Odds ratio
MAF (controls)
0.5
0.1
0.05

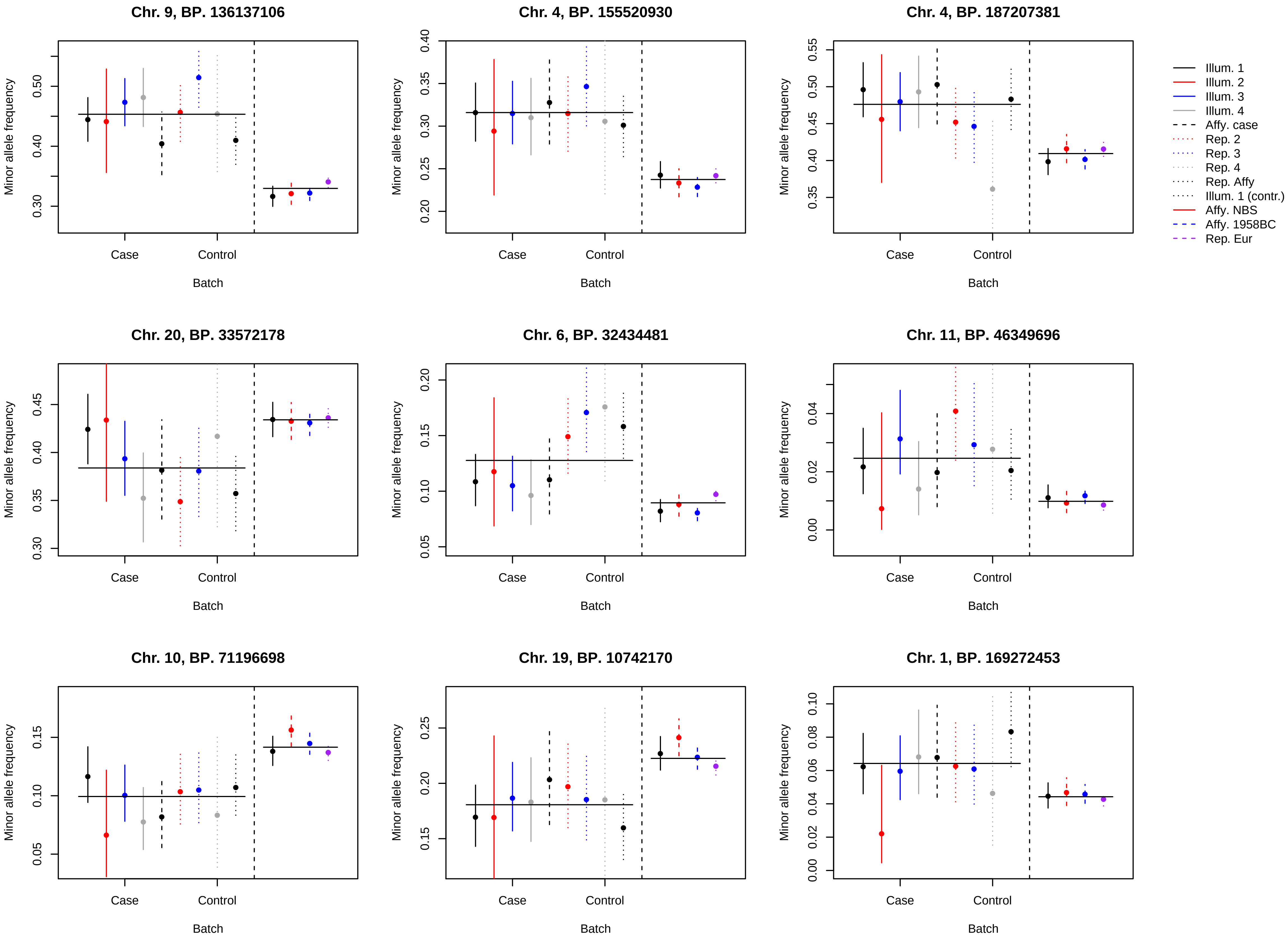

Tier 1
Odds ratio
1.8
1.7
1.6
1.5
1.4
1.3
1.2
0.1
0.2
0.3
0.4
0.5
MAF (controls)
1
0.5
0

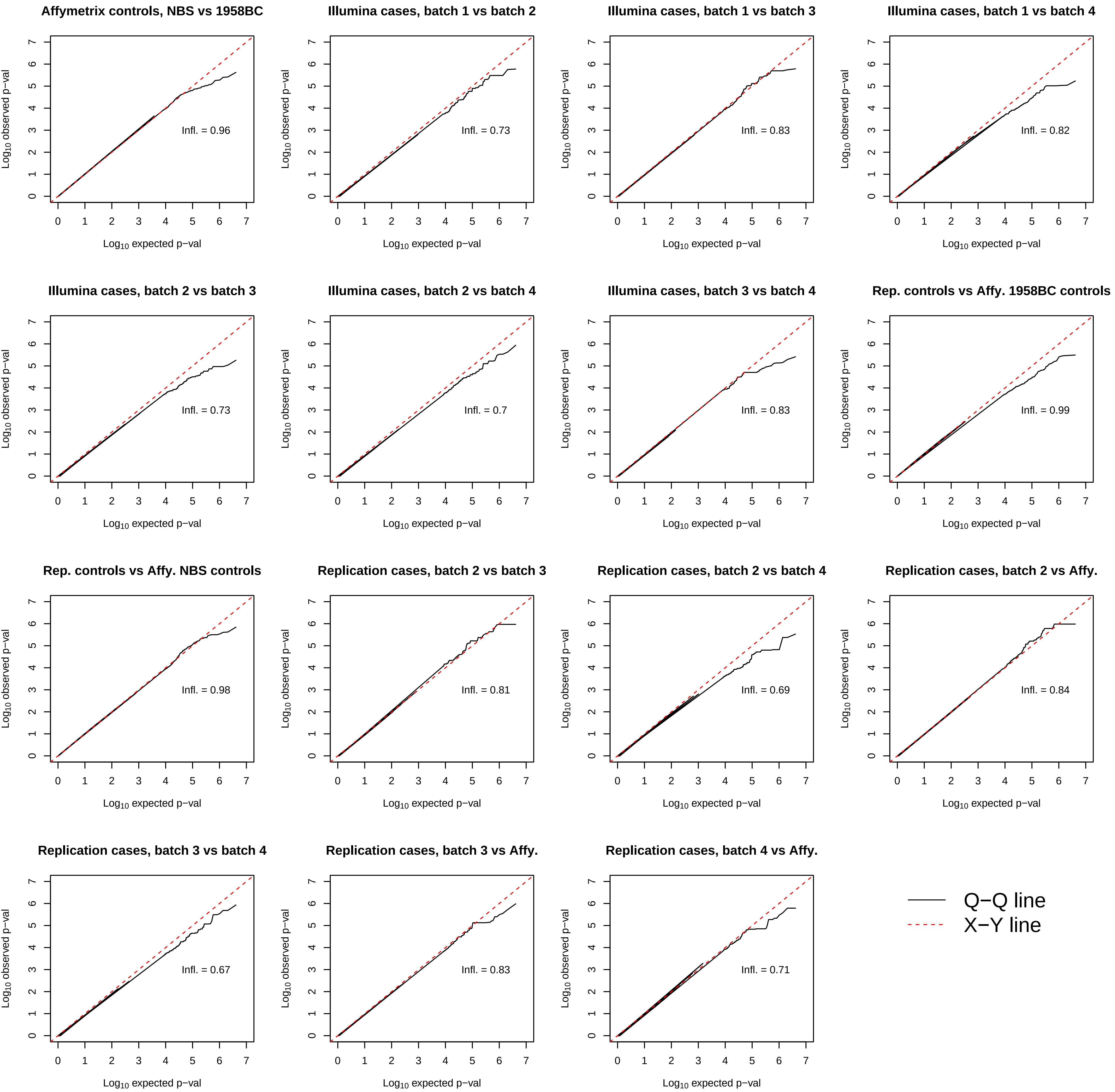

Tier 2
1
0.5
0
1.8
1.7
1.6
1.5
1.4
1.3
1.2
Odds ratio
0.1
0.2
0.3
0.4
0.5
MAF (controls)


Discovery
−log10(P)
0
5
10
15
F5/NME7
FGG
F11
HLA−DRA
ABO
TSPAN15
F2
SLC44A2
MYH7B
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
18
20
22
Chromosome


Replication
−log10(P)
0
5
10
15
F5/NME7
FGG
F11
HLA−DRA
ABO
TSPAN15
F2
SLC44A2
MYH7B
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
18
20
22
Chromosome


Batch 1, World
PC1 (2.6%)
PC2 (2.2%)
−0.01 0.00 0.01 0.02 0.03
−0.01 0.00 0.01 0.02 0.03 0.04
Population
CASE
CONTROL
AFR
AMR
EAS
EUR
SAS
Batch 1, Europe
PC1 (0.17%)
PC2 (0.12%)
−0.05 0.00 0.05 0.10
0.0 0.2 0.4
Population
CASE
CONTROL
CEU
FIN
GBR
IBS
TSI
Batch 1, Final
PC1 (0.17%)
PC2 (0.12%)
−0.05 0.00 0.05 0.10
−0.09 −0.06 −0.03 0.00 0.03
Population
CASE
CONTROL
CEU
FIN
GBR
IBS
TSI
Batch 2, World
PC1 (3.4%)
PC2 (2.6%)
−0.03 −0.02 −0.01 0.00 0.01
−0.02 0.00 0.02
Population
CASE
AFR
AMR
EAS
EUR
SAS
Batch 2, Europe
PC2 (0.47%)
PC3 (0.25%)
−0.05 0.00 0.05
−0.08 −0.04 0.00 0.04
Population
CASE
CEU
FIN
GBR
IBS
TSI
Batch 2, Final
PC1 (0.47%)
PC2 (0.24%)
−0.05 0.00 0.05 0.10
−0.04 0.00 0.04 0.08
Population
CASE
CEU
FIN
GBR
IBS
TSI
Batch 3, World
PC1 (3.2%)
PC2 (2.6%)
−0.03 −0.02 −0.01 0.00 0.01
−0.04 −0.02 0.00 0.02
Population
CASE
AFR
AMR
EAS
EUR
SAS
Batch 3, Europe
PC3 (0.37%)
PC4 (0.22%)
−0.2 −0.1 0.0 0.1 0.2
−0.6 −0.4 −0.2 0.0 0.2
Population
CASE
CEU
FIN
GBR
IBS
TSI
Batch 3, Final
PC1 (0.37%)
PC2 (0.21%)
−0.10 −0.05 0.00 0.05
−0.05 0.00 0.05 0.10
Population
CASE
CEU
FIN
GBR
IBS
TSI
Batch 4, World
PC1 (3.4%)
PC2 (2.6%)
−0.03 −0.02 −0.01 0.00 0.01
−0.02 0.00 0.02
Population
CASE
AFR
AMR
EAS
EUR
SAS
NA
Batch 4, Europe
PC1 (0.47%)
PC2 (0.26%)
−0.10 −0.05 0.00 0.05
−0.05 0.00 0.05
Population
CASE
CEU
FIN
GBR
IBS
TSI
Batch 4, Final
PC1 (0.49%)
PC2 (0.27%)
−0.10 −0.05 0.00 0.05
−0.04 0.00 0.04
Population
CASE
CEU
FIN
GBR
IBS
TSI
Affymetrix (disc), World
PC1 (3.4%)
PC2 (2.1%)
0.00 0.01 0.02 0.03
0.00 0.02 0.04
Population
CASE
CONTROL
AFR
AMR
EAS
EUR
SAS
Affymetrix (disc), Europe
PC3 (0.065%)
PC4 (0.054%)
−0.050 −0.025 0.000 0.025
−0.6 −0.4 −0.2 0.0
Population
CASE
CONTROL
CEU
FIN
GBR
IBS
TSI
Affymetrix (disc), Final
PC1 (0.067%)
PC2 (0.053%)
0.00 0.05 0.10
−0.025 0.000 0.025 0.050 0.075
Population
CASE
CONTROL
CEU
FIN
GBR
IBS
TSI
Replication, World
PC1 (3.2%)
PC2 (1.9%)
0.00 0.01 0.02 0.03
−0.04 −0.02 0.00 0.02
Population
CASE
CONTROL
AFR
AMR
EAS
EUR
SAS
NA
Replication, Europe
PC1 (0.1%)
PC2 (0.071%)
−0.050 −0.025 0.000 0.025
−0.04 −0.02 0.00 0.02
Population
CASE
CONTROL
CEU
FIN
GBR
IBS
TSI
Replication, Final
PC1 (0.076%)
PC2 (0.058%)
−0.05 0.00 0.05 0.10
−0.025 0.000 0.025 0.050
Population
CASE
CONTROL
CEU
FIN
GBR
IBS
TSI


Chr. 9, BP. 136137106
Chr. 4, BP. 155520930
Chr. 4, BP. 187207381
Chr. 20, BP. 33572178
Chr. 6, BP. 32434481
Chr. 11, BP. 46349696
Chr. 10, BP. 71196698
Chr. 19, BP. 10742170
Chr. 1, BP. 169272453
Minor allele frequency
Batch
Case
Control
0.30
0.40
0.50
0.20
0.25
0.30
0.35
0.40
0.35
0.40
0.45
0.50
0.55
0.30
0.35
0.40
0.45
0.05
0.10
0.15
0.20
0.00
0.02
0.04
0.05
0.10
0.15
0.15
0.20
0.25
0.00
0.02
0.04
0.06
0.08
0.10
Illum. 1
Illum. 2
Illum. 3
Illum. 4
Affy. case
Rep. 2
Rep. 3
Rep. 4
Rep. Affy
Illum. 1 (contr.)
Affy. NBS
Affy. 1958BC
Rep. Eur


Affymetrix controls, NBS vs 1958BC
Infl. = 0.96
Illumina cases, batch 1 vs batch 2
Infl. = 0.73
Illumina cases, batch 1 vs batch 3
Infl. = 0.83
Illumina cases, batch 1 vs batch 4
Infl. = 0.82
Illumina cases, batch 2 vs batch 3
Infl. = 0.73
Illumina cases, batch 2 vs batch 4
Infl. = 0.7
Illumina cases, batch 3 vs batch 4
Infl. = 0.83
Rep. controls vs Affy. 1958BC controls
Infl. = 0.99
Rep. controls vs Affy. NBS controls
Infl. = 0.98
Replication cases, batch 2 vs batch 3
Infl. = 0.81
Replication cases, batch 2 vs batch 4
Infl. = 0.69
Replication cases, batch 2 vs Affy.
Infl. = 0.84
Replication cases, batch 3 vs batch 4
Infl. = 0.67
Replication cases, batch 3 vs Affy.
Infl. = 0.83
Replication cases, batch 4 vs Affy.
Infl. = 0.71
Log_{10} observed p−val
Log_{10} expected p−val
0
1
2
3
4
5
6
7
Q−Q line
X−Y line