# DISTMIX2: robust imputation of summary statistics for cosmopolitan cohorts using a large and diverse reference panel

Chris Chatzinakos[1*], Donghyung Lee[2], Cai Na[3], Vladimir I. Vladimirov[1], Bradley T. Webb[1], Brien P. Riley[1], Jonathan Flint[4], Kenneth S. Kendler[1] and Silviu-Alin Bacanu[1]

[1]Department of Psychiatry, Virginia Commonwealth University, Richmond Virginia, United States of America

[2]The Jackson Laboratory for Genomic Medicine, Farmington Connecticut, United States of America

[3]Welcome Trust Centre for Human Genetics, Oxford, United Kingdom

[4]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, United States of America

* chris.chatzinakos@vcuhealth.org

## ABSTRACT

Methods for direct imputation of summary statistics, e.g. our group's DISTMIX tool, were shown to be practically as accurate as genotype imputation method, while incurring orders of magnitude lower computational burden. Given that such imputation needs a precise estimation of linkage disequilibrium (LD) for mixed ethnicity (cosmopolitan) cohorts, there is a great need i) for much larger and diverse panels and ii) to estimate the ethnic composition of the cohort, e.g. the weights for subpopulations in the diverse panel. Unfortunately, DISTMIX and its main competitors are largely using a very small reference panel of ~2,500 subject coming from the 1000 Genome (1KG) Project. DISTMIX computed the ethnic weights of a cohort based on in-cohort allele frequency (AF)

estimates. Unfortunately, due to privacy issues, most genome wide association studies (GWAS) largely stopped providing cohort AFs. Thus, to accurately estimate the LD needed for an exhaustive analysis of cosmopolitan cohorts, we propose DISTMIX2. When compared to DISTMIX and its competitors, the proposed method adds a i) much larger and diverse reference panel and ii) novel estimation for weights of ethnic mixture based solely on Z-scores (when AFs not available). To build a larger and more diverse reference panel, we use the publicly and privately available data to obtain a 33,000 (33K) panel which includes ~11K Han Chinese. The proposed method of estimating ethnic weights adequately controls the Type I error rates, especially when the subpopulations in the study are well represented in the reference panel. However, using naive pre-estimated weights incurs a much higher false positive rate. We apply DISTMIX2 to the GWAS summary statistics from the Psychiatric Genetic Consortium (PGC). Our method which uncover signals in numerous new regions, with most of these findings coming from the rarer variants.

## Author summary

By predicting summary statistics at unmeasured genetic variants, direct imputation is a promising method for enhancing the resolution of genetic studies. However, for a better prediction of statistics at unmeasured variants, there is a need to address two urgent issues. First, there is a need for very large and diverse reference panels that greatly improve on the mostly European ones having ~2,500 (2.5K) subjects. We address this shortcoming by building a large and diverse reference panel (33K subjects, ~20K Europeans and ~11K Asians). Second, there is a need to estimate the ethnic composition

of the study cohort, even when they do not report in-cohort allele frequency for genetic variants. We solve this issue by using a novel method that uses only Z-scores, which are easily computed from reported summary statistics. Our method that implements the two above solutions i) adequately controls the false positive rate and ii) provides much improved resolution when compared to methods based on older reference panels. Practical application to reported summary statistics from studies of psychiatric disorders greatly increase the number of regions harboring signals. Most of these findings are associated with rarer variants that could not be robustly assessed using smaller panels.

## Introduction

Genotype imputation [1-4] methods are commonly used to increase the genomic resolution for large-scale multi-ethnic meta-analyses [5-9] by predicting genotypes at unmeasured markers based on cosmopolitan reference panels, e.g. 1000 Genomes (1KG) [10]. However, genotypic imputation is computationally burdensome and require access to subject level genetic data, which is harder and slower to get than summary statistics.

To overcome these limitations researchers proposed summary statistics based imputation methods, e.g. DIST [11] and ImpG [12]. These methods can directly impute summary statistics (two-tailed Z-scores) for unmeasured SNPs from genome-wide association studies (GWASs) or called variants from sequencing studies. The methods were shown to i) substantially reduce the computational burden and ii) be practically as accurate as commonly used genotype imputation methods. These methods were

successfully applied in gene-level joint testing of functional variants (Lee et al., 2014) and functional enrichment analyses (Pickrell, 2014). However, these first direct imputation methods were only amenable for imputation in ethnically homogeneous cohorts.

To accommodate cosmopolitan cohorts, DIST method was extended [13] to **D**irectly **I**mputing summary **S**tatistics for unmeasured SNPs from **M**ixed ethnicity cohorts (DISTMIX). It i) predicted a study's proportions (weights) of ethnicities from a multi-ethnic reference panel based only on cohort allele frequencies (AFs) for (common) Single Nucleotide Polymorphisms (SNPs) from the studied cohort or taking prespecified ethnic weights, ii) computed ethnicity-weighted correlation matrix based on the estimated/prespecified weights and genotypes of ethnicities from the reference panel and, then, iii) using the weighted correlation matrix for accurate imputation.

Unfortunately, lately two issues occurred in practical applications of DISTMIX. First, due to privacy concerns [14], cohort AFs are lately only rarely provided. Second, similar to its competitors, the method relied on 1KG reference panel which was both small and European centric, while many meta-analyses have non-trivial fractions of non-European subjects [6, 15]. Since its publishing larger reference cohorts were sequenced and published, e.g. Haplotype Reference Consortium (HRC) [16] and CONVERGE [5]. CONVERGE complements nicely HRC due to consisting of >11K Han Chines subjects.

Consequently, in DISTMIX2, we address the above shortcomings by including two critical components. First, we provide a novel method to accurately estimate ethnic weights of the cohort which uses only summary statistics, e.g. Z-scores. Second, we build a larger, more diverse reference panel with 33K subjects, which combines the subjects from the

publicly available part of HRC with CONVERGE. Subsequently, we apply the method to Psychiatric Genetics Consortium (PGC) data and uncover many possible new signals.

## Results

For Illumina 1M SNPs [17] that were masked, and then imputed (see Method evaluation section), DISTMIX2 with our novel automatic ethnic weight detection (see Method section), controls the false positive rates at or below nominal threshold, even at very low type I error, e.g. $10^{-6}$ (Text S1, Fig S1 in SI). $R^2$ between true values and estimated one is practically above $0.92$ for our five simulated mixed-cohort scenarios (Text S1, Figs S2-S6 in SI). Also, DISTMIX2 imputed statistics have very good mean squared error (RMS) (Text S1, Figs S7-S11 in SI). For the above three measurements (size of the test, $R^2$ and RMS) the setting of 250Kb for the length of the predicted window was the least precise, while 500Kb and 1000Kb had practically identical precision.

For rare and very rare variants, the size of the test is up to 300-1000X higher than the nominal one and even up to 5000-10000X for cohorts that have large fractions of subpopulations that are underrepresented in the reference panel (e.g. Americans, Africans etc.), especially for the setting Minor Allele Frequency (MAF), $0.05\%$<MAF<$0.5\%$ and Information (Info), Info<$0.2$ (Text S1, Figures S12-S47 in SI).

For the “nullified” data sets, e.g. those obtained from real data sets by substituting the study Z-scores by their expected quantile under the null hypothesis ($H_0$) (Method evaluation section and Text S1, Figs S42-S48 in SI), DISTMIX2 controls reasonably well

the size of the test - up to 20X higher than the nominal rate (even for SNPs with low MAFs and low Info).The minimum GWAS p-values for the nullified data sets that were imputed ranged between $8.13 * 10^{-7}$ and $1.11 * 10^{-11}$. By fitting a normal distribution to $-log_{10}$(minimum p-values), we estimate the mean to be $8.655$ and the standard deviation to be $1.172$. Using as criterion the conservative three standard deviations above the mean, we obtain from these realistic data a $12.17$ upper bound for the $-log_{10}$ (minimum p-values). I.e. in DISTMIX2 practical applications [Psychiatric Genetics Consortium (PGC) traits], a conservative threshold for significance is $10^{-12}$, regardless of imputation Info and SNP MAF. *Consequently, in all applied analyses in this paper we add this very stringent threshold for DISTMIX2 imputed summary statistics*. [Using as criterion the even more conservative five standard deviations above the mean (the very conservative Chebyshev inequality for the upper bound of the p-value of exceeding this threshold= $\frac{1}{5^2} = 0.04$), we obtain a $14.515$ upper bound for the $-log_{10}$ (minimum p-values), i.e. a super-conservative significance threshold of $3 * 10^{-15}$.]

For the practical applications to PGC traits (Table 1), we construct Manhattan plots for all autosome chromosomes (1-22) and, individually, for chromosomes harboring novel signals (defined as imputed SNPs with statistically significant p-values that are at least 250Kb away from the reported GWAS signal) (Fig. 1-2, Text S2, Fig. S49-S57 in SI). For all Manhattan plots we draw two dash lines denoting statistical significance signals. The red line is the default genome-wide threshold of $p = 5 * 10^{-8}$, which is applicable to signals from measured SNPs and common imputed SNPs with high Info values. The purple line at $p = 10^{-12}$ is the threshold to be used for rare/very rare variants and/or

variants with low information; it corresponds to the above mentioned upper bound for nullified data. As an illustration, we present Schizophrenia Manhattan plot for all chromosomes and only for chromosome 12 (Fig. 1 and Fig. 2).

**Table 1. Real dataset description.**

| Trait | Trait Abbreviation | Dataset Description |
|---|---|---|
| Schizophrenia | SCZ | PGC2 SCZ [6] |
| Attention Deficit Hyperactivity Disorder | ADHD | PGC ADHD [18] |
| Autism | AUT | PGC AUT [19] |
| Bipolar | BIP | PGC BIP [20] |
| Eating Disorders | EAT | PGC EAT [21] |
| Major depression disorder | MDD | PGC MDD [22] |

These applications of DISTMIX2 to PGC data sets suggests the existence of numerous new signals, most associated with rare and very rare SNPs (see Table 2) (for all signals see SE, Excel file). For instance in chromosome 12 for PGC schizophrenia (rs143374), with MAF=$0.0007$, Info=$0.245$ and p-value=$9.26 * 10^{-46}$ the magnitude of the p-value suggest that this signal is likely not to be an artifact (above the more stringent threshold), in chromosome 11 for ADHD (rs5681132) where the MAF=$0.0004$, the Info= $0.018$ and p-value=$7.40 * 10^{-16}$ (above the more stringent threshold), in chromosome 22 for AUT (rs1380986), with MAF=$0.0006$, Info=$0.498$ and p-value=$8.01 * 10^{-15}$ (above the more stringent threshold), in chromosome 7 for BIP (rs76350051), with MAF= $0.0004$, Info=$0.04$ and p-value=$2.47 * 10^{-37}$ (above the more stringent threshold), in chromosome 8 for EAT (rs78958069), with MAF=$0.0002$, Info=$0.005$ and p-value=$4.17 * 10^{-10}$ (above the default threshold) and in chromosome 12 for MDD (rs567868887), with MAF=$0.0009$, Info=$0.28$ and p-value=$1.57 * 10^{-55}$ (above the more stringent threshold).

When imputing in parallel SNPs regions of 40 Mbp, the analysis of each data set had a running time of less than 5 days on a cluster node with 4x Intel Xeon 6 core 2.67 GHz.

**Table 2. Best three signals for each PGC dataset. Bolded red entries correspond to the most stringent threshold of $p < 3 * 10^{-15}$, not bolded red to the second most conservative threshold $3 * 10^{-15} < p < 10^{-12}$ and not bolded blue $10^{-12} < p < 5 * 10^{-8}$.**

| Trait | rs_id | chr | bp | p-val | Info | MAF |
|---|---|---|---|---|---|---|
| **ADHD** | **rs568113293** | **11** | **54,899,533** | $\mathbf{7.40 * 10^{-16}}$ | **0.0189** | **0.00049** |
| | rs544637819 | 3 | 15,310,737 | $1.78 * 10^{-14}$ | 0.1543 | 0.00171 |
| | chr6:30450452 | 6 | 30,450,452 | $6.44 * 10^{-13}$ | 0.0698 | 0.00151 |
| **AUT** | rs138098629 | 22 | 36,584,165 | $8.01 * 10^{-15}$ | 0.4980 | 0.00063 |
| **BIP** | **rs76350051** | **7** | **64,164,245** | $\mathbf{2.42 * 10^{-37}}$ | **0.0417** | **0.00046** |
| | **rs138549126** | **3** | **52,592,843** | $\mathbf{6.65 * 10^{-16}}$ | **0.073** | **0.00052** |
| | **rs149257260** | **15** | **71,600,045** | $\mathbf{1.40 * 10^{-15}}$ | **0.4246** | **0.00017** |
| **EAT** | rs78958069 | 8 | 43,539,021 | $4.17 * 10^{-10}$ | 0.005 | 0.0002 |
| | rs144485994 | 20 | 4,963,320 | $5.18 * 10^{-9}$ | 0.15 | 0.0001 |
| **MDD** | **rs567868887** | **12** | **31,931,432** | $\mathbf{1.57 * 10^{-55}}$ | **0.2800** | **0.00098** |
| | **rs112241719** | **11** | **111,514,969** | $\mathbf{8.14 * 10^{-45}}$ | **0.4900** | **0.00025** |
| | **rs182264017** | **1** | **188,992,506** | $\mathbf{5.05 * 10^{-44}}$ | **0.2775** | **0.00035** |
| **SCZ** | **rs559199817** | **3** | **17,267,731** | $\mathbf{1.30 * 10^{-87}}$ | **0.0213** | **0.00073** |
| | **rs143337489** | **12** | **11,2089,686** | $\mathbf{9.26 * 10^{-46}}$ | **0.2464** | **0.00019** |
| | **rs193224736** | **16** | **8,593,132** | $\mathbf{3.79 * 10^{-21}}$ | **0.28476** | **0.00018** |

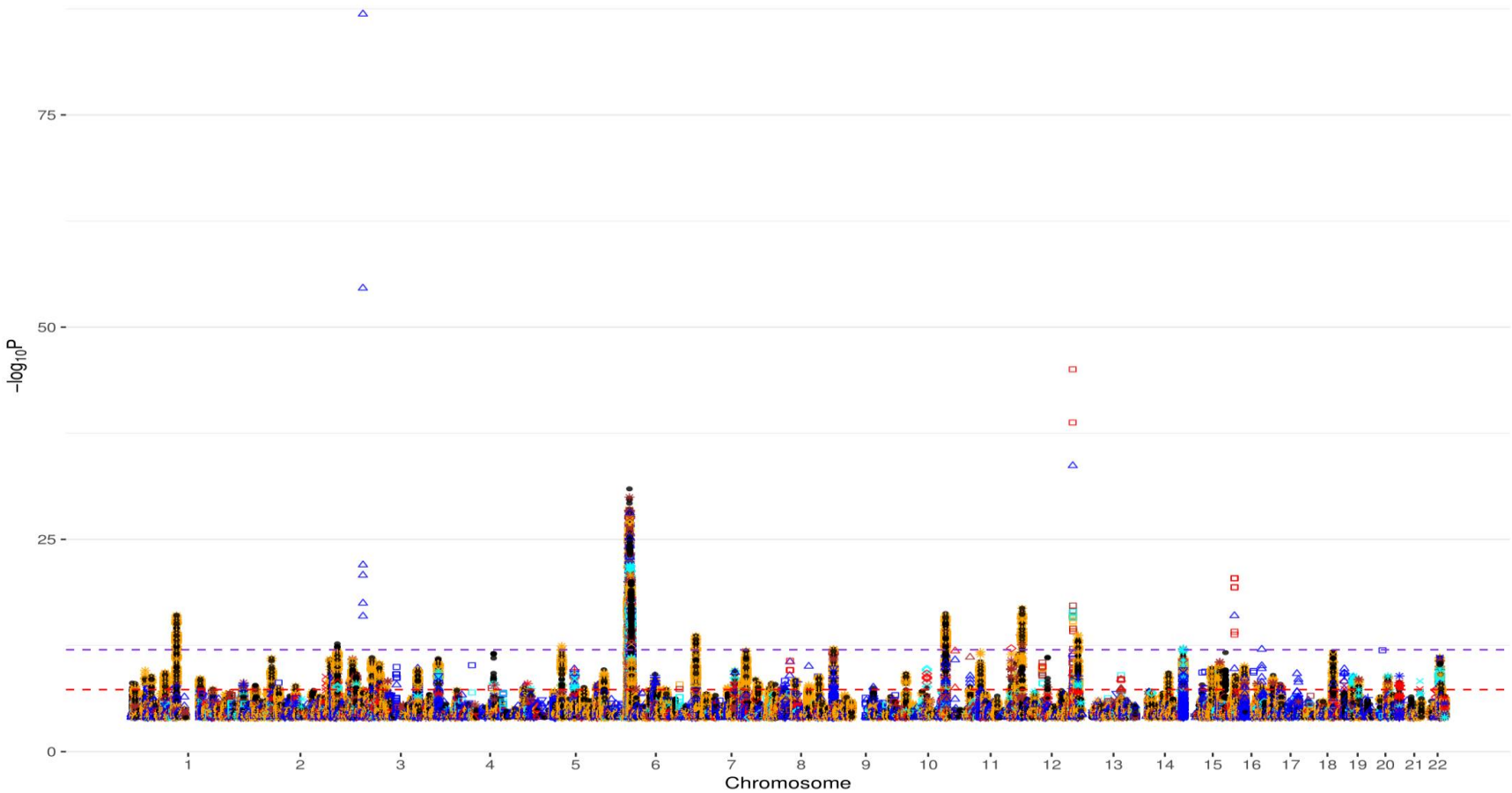


**Fig. 1 Manhattan plot for chromosomes 1-22 for SCZ. ● denotes reported signals from the original GWAS and the remain symbols and colors denote DISTMIX2 imputed signals. Among imputed signals blue denotes info<0.2, red denotes 0.2<info<0.4, cyan denotes 0.4<info<0.6, brown denotes 0.6<info<0.8, orange denotes info>0.8, □ denotes MAF <0.05%, △ denotes 0.05%<MAF<0.5%, ▽denotes 0.5%<MAF<1%, + denotes 1%<MAF<2%, ◊ denotes 2%<MAF<5%, $x$ denotes 5%<MAF<10% and ✳ denotes 10%<MAF<50%. The red line is the default genome-wide threshold of $p = 5 * 10^{-8}$, which is applicable common SNPs with moderate to large Info values. The purple line at $p = 10^{-12}$ is the threshold to be used for rare and/or low Info variants.**

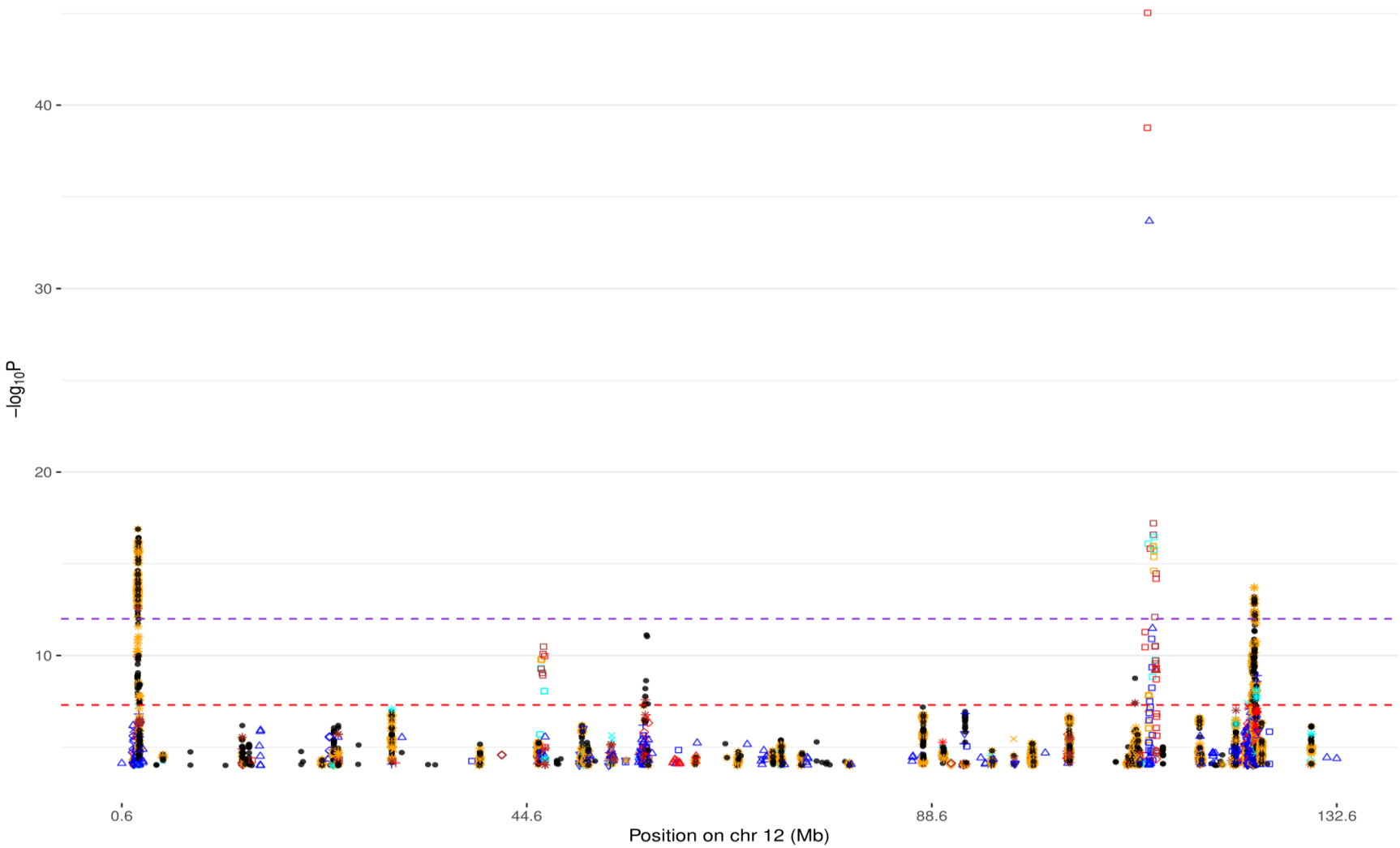


**Fig 2. Manhattan plot for chromosome 12 for SCZ (see Fig 1. for background).**

## Method evaluation

To estimate the accuracy and false positive rates of DISTMIX2, for five different cosmopolitan studies scenarios we simulated (under $H_0$: no association between trait and variants) 100 cosmopolitan cohorts of 10,000 subjects for autosomal SNPs in Ilumina 1M panel [13] using 1KG haplotype patterns (Text S1, Table S1 in SI). The subject phenotypes were simulated independent of genotypes as a random Gaussian sample. SNP phenotype-genotype association summary statistics were computed from a correlation test.

The accuracy of the procedure was assessed by masking $5\%$ of the SNPs (Experiment 1, Table 3). Subsequently, the true values and the imputed values at these masked SNPs were used to computed i) their correlation and ii) the mean squared error of the imputation. We assess these measures at four different levels of MAF. To compare the Type I error rate of our proposed method, DISTMIX2, we estimated the relative Type I error (the empirical divided by the nominal Type I error rate) as a function of the nominal Type I error rate, for the same four MAF levels for all the cohorts. Finally, for all the combinations between MAFs and Info we performed DISTMIX2 analyses with three different parameters for the length of the predicted window (the length of the predicted window is also the minimum number of the measured SNPs).

**Table 3. Experiment 1 parameter settings.**

| MAF levels | Panel | Window length |
|---|---|---|
| MAF<5% | 1K | 250Kb |
| 5% <MAF< 10% | 33K | 500Kb |
| 10%<MAF<20% | | 1000Kb |
| 20%<MAF<50% | | |

To assess the reliability of DISTMIX2 results for rare and very rare variants, for the above cohorts, we also estimate the size of the test for DISTMIX2 for very low MAFs (rare variants), (Experiment 2, Table 4). The size of the test is assessing for 5 imputation Info intervals and 6 MAF intervals.

**Table 4. Experiment 2 variable parameter settings. Fixed parameters for this experiment: 33K panel and 500Kb window length.**

| MAF levels | Info levels |
|---|---|
| 0.05% <MAF< 0.5% | Info< 20% |
| 0.5% <MAF< 1% | 20%<Info<40% |
| 1% <MAF<2% | 40%<Info<60% |
| 2% <MAF< 5% | 60%<Info<80% |
| 5% <MAF< 10% | Info>80% |
| 10% <MAF< 50% | |

However, given that 1) the simulated cohorts might not reflect real data and 2) these data sets do not have the sample sizes needed to detect very rare SNPs (e.g. MAFs $< 0.05\%$), which is important for DISTMIX2 inference in practical applications, we used real data sets to create so-called nullified data sets (Experiment 3, Table 5). These nullified data are based on 20-real and mostly Caucasian GWAS SCZ, ADHD, AUT, MDD and sixteen GWAS meta-analyses that are not yet publicly available. This approximation for null data is obtained by substituting the expected quantile of the Gaussian distribution for the (ordered) Z-score (see also S4 in SI). We note that, while the quantile estimation adjusts the noncentrality parameter (enrichment) of the statistics to zero, it does not change the order of the statistics. One effect of this fact is that imputing statistics within/near the peak signals in original GWASs might result in somewhat increased false positive rates and, thus, the genome-wide false positive rates might appear to be moderately inflated. Thus,

adjusting imputed statistics post-factum for the false positive increase observed in these nullified data is likely to yield conservative inference.

**Table 5. Experiment 3 variable parameter settings. Fixed parameters for this experiment: 33K panel and 500Kb window length.**

**.**

| **MAF levels** | **Info levels** |
|---|---|
| MAF< 0.05% | Info< 20% |
| 0.05% <MAF< 0.5% | 20% <Info< 40% |
| 0.5% <MAF< 1% | 40% <Info< 60% |
| 1% <MAF< 2% | 60% <Info< 80% |
| 2% <MAF< 5% | Info> 80% |
| 5% <MAF< 10% | |
| 10% <MAF< 50% | |

## Practical Applications

We applied DISTMIX2 to some of the psychiatric summary datasets available for download from Psychiatric Genetics Consortium (PGC- http://www.med.unc.edu/pgc/), i.e. schizophrenia (SCZ), attention deficit hyperactive disorder (ADHD), autism (AUT), eating disorder (EAT), bipolar (BIP) disorder and major depressive disorder (MDD) (see Table 1 for references). Based on the results from simulations under the null hypothesis (Experiment 1), for all these practical applications we used a) the larger 33K size panel and b) a length of the predicted of 500Kb.To improve the imputation of the unmeasured SNPs for SCZ, we denote as “measured SNPs” only those with very high information (Info>0.997). For the ADHD, AUT, BIP and MDD data sets, because the imputation information is not available, we accept as measured SNPs the set consisting of the

intersection between SNPs in each GWAS and the above SCZ's "measured" SNPs. Where available (e.g. MDD) we also filtered out SNPs with effective sample sizes below the maximum.

## Discussion

DISTMIX2, is a software/method for "off-the-shelf" direct imputation of the unmeasured SNP statistics in cosmopolitan cohorts. The main features of the updated version are 1) a much larger (33K subjects) and more diverse (includes ~11K Han Chinese) reference panel and 2) a novel procedure for estimating the ethnic composition of the cohort without the need for AF information. Its application to PGC data sets provides numerous new signal regions, most harboring rarer variants.

Due to our reassignment of subjects to subpopulations when constructing the 33K reference panel, the naive assignment of the pre-estimated weights to only specific subpopulations from the reference panel that are considered the closest ones to the perceived cohort composition, can greatly increase of the type I error (false positives). For that reason, when AF is not available, we recommend to the users to provide continental cohort weights (i.e. European [EUR], East Asian [ASN], South Asian [SAS], African [AFR] and America native [AMR]) and our software automatically will allocate these meta-weights to the most likely within-continent subpopulations. However, when AF is available there is no need to provide this additional information.

DISTMIX2 maintains the type I error reasonably accurately even for low MAFs and low Info variants, especially for mostly European (East Asian) cohorts that are

overrepresented in our reference panel. When MAF>5% (common variants), DISTMIX2 for all the levels of the information, appears to maintain the false positive rates to at most an order of magnitude higher than the nominal ones. For imputed variants (especially rarer or with lower Info) in study of Europeans, preliminary results from nullified data suggest that a conservative threshold for significance can be set at $p = 10^{-12}$; a very conservative one is $p = 3 * 10^{-15}$. Simulation results suggest that, when a larger part of study cohort consist of subpopulations underrepresented in our reference panel, it is reasonable to lower the genome-wide significant threshold for p-value of imputed variants by a factor of ~10,000.

The length of the prediction window (250Kb, 500Kb, 1000Kb) is an important design parameter due to its implications to speed and precision. Simulations results suggest that, while the accuracies for 500Kb and 1000Kb estimates are very close, the computational burden increases ~2.5 times for the 1000kb window. For that reason, we recommend that researchers use a 500Kb prediction window.

While mentioned only briefly in this manuscript, for practical application we use as "measured" SNP in the input summary statistic file only the GWAS SNPs reported to have close to perfect information and/or effective sample size. Our approach is rooted in preserving the cardinal assumption, of our and all but one other imputation methods [23], that the LD between SNP Z-scores is very well approximated by the LD of the same SNPs in the reference panels. It is well known that when there are non-negligible missing rates for the variant pair this assumption is not met [23]. While the LD of Z-scores can be

estimated by making reasonably realistic assumptions about co-missingness patterns of such SNP pairs, to avoid even the rarer circumstances in which these assumptions might not be met, we decided to avoid such an approach. Consequently, we employed (and recommend) the conservative approach of deeming as measured only SNPs with close to perfect imputation information and/or effective sample sizes in the original GWAS.

In practical applications, the very low MAF and Info for some SNP can cause up to 4 orders of magnitude inflation in false positive rates. While signals for rarer SNPs can be viewed as much “softer” signals than the ones associated with common and high Info variants, the very low p-values for some of them suggest that most of these signals are likely to be real. This suggestion is enhanced by the fact that, to avoid the pitfalls of estimating covariances from just very few minor alleles we did not include in the imputation panel SNPs that don not have at least i) 20 minor alleles in the Europeans or East Asians or ii) 5 in all other continental groups. Nonetheless, we recognize that signals for these SNPs should be treated with more skepticism than the more common/higher Info variants and subjected to the most stringent wet-lab validations.

## Method

### Larger and more diverse reference panel

To facilitate imputation of rarer variants, the current version uses the 33,000 subjects (33K) as reference panel. It consists of 20,281 Europeans, 10,800 East Asians, 522 South Asians, 817 Africans and 533 Native Americans (Text S3, Table S2 in SI). The reference panel includes the publicly available 22,691 subjects from Haplotype Reference

Consortium (HRC) and 10,262 CONVERGE. For CONVERGE subjects, we used we used Province to divide them into 4 population (CNE, CCE, CSE and CCS). HRC subjects coming from the small Orkney (ORK) island provided the basis for an extra European population, i.e. ORK. Subjects from 1KG in HRC sample, CONVERGE and ORK along with their a) population label b) first 20 ancestry principal components were used to train a quadratic discriminant model for predicting population label from principal components. Subsequently, to have more homogeneous populations in the panel, all available subjects were assigned(reassigned) population labels based on model prediction. Consequently, a subject might be re-assigned to a different (but related) population.

Finally, our reference panel contains twenty-six million SNPs. To have reasonably accurate SNP LD estimators, we eliminate the rarest SNPs which did not have at least i) 20 alleles in European or East Asian superpopulations or ii) 5 in African, South Asian and America native superpopulations.

**Converge haplotypes**

**DNA sequencing.** DNA was extracted from saliva samples using the Oragene protocol. A barcoded library was constructed for each sample. Sequencing reads obtained from Illumina Hiseq machines were aligned to Genome Reference Consortium Human Build 37 patch release 5 (GRCh37.p5) with Stampy (v1.0.17) [24] [21] [21] [21] [21] [1] [5][2] using default parameters, after filtering out reads containing adaptor sequencing or consisting of more than 50% poor quality (base quality <= 5) bases. Samtools (v0.1.18)

[25] was used to index the alignments in BAM format [25] and Picardtools (v1.62) was used to mark PCR duplicates for downstream filtering. The Genome Analysis Toolkit's (GATK, version 2.6). Base quality score recalibration (BQSR) was then applied to the mapped sequencing reads using BaseRecalibrator in Genome Analysis Toolkit (GATK, basic version 2.6) [26] with the known insertion and deletion (INDEL) variations in 1000 Genomes Projects Phase 1 [27] and known single nucleotide polymorphisms (SNPs) from dbSNP (v137, excluding all sites added after v129) excluded from the empirical error rate calculation. GATKlite (v2.2.15) was then used to output sequencing reads with the recalibrated base quality scores while removing reads without the "proper pair" flag bit set by Stampy (1-5% of reads per sample) using the --read_filter ProperPair option (if the "proper pair" flag bit is set for a pair of reads, it means both reads in the mate-pair are correctly oriented, and their separation is within 5 standard deviations from the mean insert size between mate-pairs).

**Variant calling, imputation, and phasing**

Variant discovery and genotyping (for both SNPs and INDELs) at all polymorphic SNPs in 1000G Phase1 East Asian (ASN) reference panel[28] was performed simultaneously using post-BQSR sequencing reads from all samples using the GATK's UnifiedGenotyper (version 2.7-2-g6bda569). Variant quality score recalibration (VQSR) was then performed with GATK's VariantRecalibrator (v2.7-4-g6f46d11) in SNP variant calls using the SNPs in 1000 Genomes Phase 1 ASN Panel [27] as the known, truth and training sets. A sensitivity threshold of 90% to SNPs in the 1000G Phase1 ASN panel was applied for SNP selection for imputation after optimizing for Transition to Transversion (TiTv) ratios

in SNPs called. Genotype likelihoods (GLs) were calculated at selected sites using a sample-specific binomial mixture model implemented in SNPtools (version 1.0), and imputation was performed at those sites without a reference panel using BEAGLE (version 3.3.2) [29]. The second round of imputation was performed with BEAGLE on the same GLs, but only at biallelic SNPs polymorphic in the 1000G Phase 1 ASN panel using the 1000G Phase 1 ASN haplotypes as a reference panel. The genotypes derived from Beagle imputation were phased using Shapeit (version 2, revision 790) [30]. Genetic maps were obtained from the Impute2 [31] website. Chromosomes 13 - 22 and X were phased using 12 threads and default parameters. Chromosomes 1-12 were phased using 12 threads in four chunks that overlap by 1MB. The phased chunks were ligated together using ligateHAPLOTYPES, available from the Shapeit website. A final set of allele dosages and genotype probabilities was generated from these two datasets by replacing the results in the former with those in the latter at all sites imputed in the latter. We then applied a conservative set of inclusion threshold for SNPs for genome-wide association study (GWAS): a) p-value for violation HWE > $10^{-6}$, b) Info score > 0.9, c) MAF in CONVERGE > 0.5% to arrive at the final set of 6,242,619 SNPs. Details can be found in [15].

**Automatic detection of cohort composition**

Our group has previously described, in DISTMIX paper [13], a method to estimate the ethnic composition when the cohort allele frequencies (AF) are available. However, lately some consortia do not provide such measure; they often provide only the Caucasians AF.

*Consequently, there is a great need for a method to estimate the ethnic composition of the cohort even when no AFs are provided*.

Below is the theoretical outline of such method. Suppose that the cohort genotype is a mixture of genotypes coming out from the $k$ ethnic groups from the reference panel. The $G_{ij}$ denotes the genotype for the $i$-th subject at the $j$-th SNP which belongs to the $l$-th group, let $p_j^{(l)}$ be the frequency of the reference allele frequency for this SNP in the $l$-th group. Let $G'_{ij} = \frac{G_{ij} - 2\,p_j^{(l)}}{\sqrt{2p_j^{(l)}\,(1-p_j^{(l)})}}$ be the normalized genotype, i.e. the transformation to a variable with zero mean and unit variance. Near $H_0$, SNP Z-score statics $Z_j's$ have the approximately the same correlation structure as the genotypes used to construct it, $G_{*j}$'s, and, thus, the same correlation structure as its transformation, $G'_{*j}$'s. However, given that both $G'_{*j}'s$ and $Z_j's$ have unit variance, it follows that the two have the same covariance (i.e. not only the same correlation) structure. Therefore, for any $s \geq 1$ $E(Z_j\ Z_{j+s}) = E(G'_{*j}\ G'_{*\,j+s})$, which, due independence of genotypes in different ethnic groups becomes:

$E(Z_j\ Z_{j+s}) = \sum_{l=1}^{k} w^{(l)}\ E[G'_{*j}{}^{(l)}\ G'_{*(j+s)}{}^{(l)}] = \sum_{l=1}^{k} w^{(l)}\ Cor(G'_{*j}{}^{(l)}, G'_{*(j+s)}{}^{(l)})$ (1),

where $w^{(l)}$ is the expected fraction of subjects from the entire cohort that belong to the $l$-th group.

While $Cor(G'_{*j}{}^{(l)}, G'_{*(j+s)}{}^{(l)})$ is unknown, it can be easily approximated using their reference panel counterparts. Thus, the weights, $w^{(l)}$, can be simply estimated by simply

regressing the product of product of reasonably close SNP $Z$-scores, $Z'_j\ Z'_{j+s}$, on correlations between normalized genotypes at the same SNP pairs for all subpopulations in the reference panel. To increase bias power, we chosen the parameter $s$, such as to maximize the variance of the within panel ethnic group correlations while keeping $j+s$-th SNP no more than 50Kb away from $j$ –th SNP. Because some GWAS might have numerous large signals, e.g. latest height meta-analysis [6, 32], a more accurate estimation of the weights is very likely to be obtained by substituting expected gaussian quantiles for $Z'_j$ (see **Nonparametric robust estimation of weights subsection**).

Due to the strong LD among SNPs, the estimation of the correlation using all SNPs in a genome might lead to a poor regression estimate in (1). To avoid this, we sequentially split GWAS SNPs into 1000 non-overlapping SNP sets, e.g. first set consists of the 1st, 1001st, 2001st, etc. map ordered SNPs in the study. The large distances between SNPs in the same set make them quasi-independent which, thus, improves the accuracy of the estimated correlation. $W = (w^{(l)})$ is subsequently estimated as the average of the weights obtained from the 1000 SNP sets. Finally, we set to zero the negatives weights and normalize the remaining weights to sum to 1 [33]. This method should be even more useful when we already know the approximate continental (EUR, ASN, SAS, AFR and AMR) weights (as estimated from study information) but it is not always clear how these proportions should be allocated among continental subpopulations. This further apportioning of continental weights is likely to be extremely important when the GWAS cohorts contain many admixed populations, e.g. African Americans and American native populations. Consequently, when continental proportions are provided by the users, we

use our automatic detection to distribute these weights to the most likely subpopulations in the reference panel. To eliminate unforeseen artifacts, we strongly recommend to the users to provide continental proportions when AFs are not available.

### Nonparametric robust estimation of weights

To estimate robust weights and to avoid false positives we apply a two-step, robust algorithm to the $Z$-scores of the SNPs. **First,** let $Z_\sigma = (z_{\sigma_1}, z_{\sigma_2}, \dots, z_{\sigma_m})$, where $\sigma$ indicates the permutation of indices of Z-scores, $Z$, for the $m$ SNPs, that orders these statistics in increasing order. **Second,** $z_i' = \Phi^{-1}(\frac{\sigma_i}{m+1})$, where $\Phi^{-1}$ is the inverse normal cumulative distribution function. Subsequently, these transformed risk scores are used for computing ethnic weights.

## Software and data availability

DISTMIX2 is freely available for academic use at https://github.com/Chatzinakos/DISTMIX2. The DISTMIX2 executable requires only the GWAS summary statistics from the user. The reference panel also available at the same repo.

## Supporting information

**SI. Text and Figures.**
(PDF)

**SE. Table with the significant signals for the real applications.**
(EXCEL)

## REFERENCES


1. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84(2):210-23.
2. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529.
3. Nicolae DL. Testing untyped alleles (TUNA)-applications to genome-wide association studies. Genet Epidemiol. 2006;30(8):718-27.
4. Servin B, Stephens M. Imputation-based analysis of association studies: Candidate regions and quantitative traits. PLoS Genet. 2007;3:1296-308.
5. Consortium C. Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature. 2015;523(7562):588-91.
6. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013;45(10):1150-9.
7. Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA, et al. Genome-wide association study identifies five new schizophrenia loci. Nat Genet. 2011;43(10):969-76.
8. Sklar P, Ripke S, Scott LJ, Andreassen OA, Cichon S, Craddock N, et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet. 2011;43(10):977-83.
9. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445(7130):881-5.
10. Abecasis GR, Auton A, Brooks LD, Depristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
11. Lee D, Bigdeli TB, Riley BP, Fanous AH, Bacanu SA. DIST: direct imputation of summary statistics for unmeasured SNPs. Bioinformatics. 2013;29(22):2925-7.
12. Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Quantity Biology,Cornell University Library. 2013.
13. Lee D, Bigdeli TB, Williamson VS, Vladimirov VI, Riley BP, Fanous AH, et al. DISTMIX: Direct imputation of summary statistics for unmeasured SNPs from mixed ethnicity cohorts. Bioinformatics. 2015.
14. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008;4(8):e1000167.
15. Cai N, Bigdeli TB, Kretzschmar WW, Li YH, Liang JQ, Hu JC, et al. Data Descriptor: 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. Sci Data. 2017;4.
16. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1279-83.
17. Marenne G, Rodriguez-Santiago B, Closas MG, Perez-Jurado L, Rothman N, Rico D, et al. Assessment of copy number variation using the Illumina Infinium 1M SNP-array: a comparison of methodological approaches in the Spanish Bladder Cancer/EPICURO study. Hum Mutat. 2011;32(2):240-8.
18. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, et al. Discovery Of The First Genome-Wide Significant Risk Loci For ADHD. bioRxiv. 2017.
19. Buxbaum JD, Daly MJ, Devlin B, Lehner T, Roeder K, MW S. The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. Neuron. 2012;76(6):1052-6.
20. Psychiatric GCBDWG. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet. 2011;43(10):977-83.
21. Duncan L, Yilmaz Z, Gaspar H, Walters R, Goldstein J, Anttila V, et al. Significant Locus and Metabolic Genetic Correlations Revealed in Genome-Wide Association Study of Anorexia Nervosa. Am J Psychiat. 2017;174(9):850-8.
22. Major Depressive Disorder Working Group of the Psychiatric GC, Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, et al. A mega-analysis of genome-wide association studies for major depressive disorder. Mol Psychiatry. 2013;18(4):497-511.

23. Rueger S, McDaid A, Kutalik Z. Evaluation and application of summary statistic imputation to discover new height-associated loci. PLoS Genet. 2018;14(5):e1007371.
24. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011;21(6):936-9.
25. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
26. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-+.
27. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.
28. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
29. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084-97.
30. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype Estimation Using Sequencing Reads. Am J Hum Genet. 2013;93(4):687-96.
31. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529.
32. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46(11):1173-86.
33. Chatzinakos C, Lee D, Webb BT, Vladimirov VI, Kendler KS, Bacanu S-A. JEPEGMIX2: improved gene-level joint analysis of eQTLs in cosmopolitan cohorts. Bioinformatics. 2017:btx509-btx.