European Journal of Human Genetics (2016) 24, 761–766

www.nature.com/ejhg

ARTICLE

FAPI: Fast and accurate *P*-value imputation for genome-wide association study

Johnny SH Kwan[1,5], Miao-Xin Li*[,1,2,3,4,5], Jia-En Deng[1] and Pak C Sham*[,1,2,3,4]

Imputing individual-level genotypes (or genotype imputation) is now a standard procedure in genome-wide association studies (GWAS) to examine disease associations at untyped common genetic variants. Meta-analysis of publicly available GWAS summary statistics can allow more disease-associated loci to be discovered, but these data are usually provided for various variant sets. Thus imputing these summary statistics of different variant sets into a common reference panel for meta-analyses is impossible using traditional genotype imputation methods. Here we develop a fast and accurate *P*-value imputation (FAPI) method that utilizes summary statistics of common variants only. Its computational cost is linear with the number of untyped variants and has similar accuracy compared with IMPUTE2 with prephasing, one of the leading methods in genotype imputation. In addition, based on the FAPI idea, we develop a metric to detect abnormal association at a variant and showed that it had a significantly greater power compared with LD-PAC, a method that quantifies the evidence of spurious associations based on likelihood ratio. Our method is implemented in a user-friendly software tool, which is available at http://statgenpro.psychiatry.hku.hk/fapi.
European Journal of Human Genetics (2016) **24,** 761–766; doi:10.1038/ejhg.2015.190; published online 26 August 2015

INTRODUCTION

Genome-wide association studies (GWAS) are now widely used to investigate associations of common genetic variants with various human traits and diseases.[1,2] Although >50 million genetic variants have been discovered in the human genome and cataloged in dbSNP,[3] no more than a few millions of them are usually typed in a GWAS (often using high-density commercial genotyping arrays). As obtaining whole-genome sequencing data of GWAS samples is still costly, imputing the individual-level genotypes at untyped common variants in a GWAS (or genotype imputation) is common in GWAS and multiple genotype imputation methods have been developed, including IMPUTE2,[4] MACH,[5] and BEAGLE.[6–8] For a comprehensive review of traditional imputation approaches, see the work by Nothnagel *et al*[9] as well as by Marchini and Howie.[10] As most GWAS now release their summary statistics, a meta-analysis of these GWAS is attractive.[11] However, these summary statistics were provided for different sets of common variants owing to the type of genotyping platforms and reference panel used for genotype imputation (such as HapMap2[12] *versus* 1000G[13]). Therefore, imputing these summary statistics of different variant sets into a common reference panel for meta-analyses is impossible using these traditional genotype imputation methods.

As researchers are mostly interested in assessing the evidence of disease association at untyped variants but not individual-level genotypes in most GWAS, the traditional two-step imputation approach (ie, genotype imputation followed by association analysis) can be simplified to a single-step approach (ie, imputing the significance of association or *P*-values at untyped common variants directly) what we call the *P*-value imputation. As the *P*-values of associations of two variants in strong linkage disequilibrium (LD) must be correlated, one may impute the *P*-values of untyped variants from the *P*-values of nearby typed variants. This allows imputation of summary statistics from different variant sets into a common reference panel without the need of raw genotype data.

Here we determine the relationship between LD r-squared and the covariance of normal test statistics (ie, *z*-scores) at any two common variants by simulation and curve fitting. Then we compute the distribution of the normal test statistic of an untyped common variant conditional on the *P*-values at neighboring typed variants and the observed LD structure using a multivariate normal distribution. We compare the speed as well as accuracy of our method, called fast and accurate *P*-value imputation (FAPI) with IMPUTE2 with prephasing,[4] a leading method in traditional genotype imputation, in GWAS data sets of simulated traits as well as schizophrenia. Furthermore, based on the FAPI idea, we develop a metric to detect abnormal association at a variant using the *P*-values at neighboring variants in LD and compare it with LD-PAC,[14] a method that quantifies the evidence of spurious associations based on likelihood ratio, using simulations.

MATERIALS AND METHODS

Fast and accurate *P*-value imputation

Relationship between LD r-squared and covariance of normal test statistics. Genotypes of two biallelic common variants were simulated for various sample sizes (4000 and 10 000 for a quantitative trait, as well as 2000 cases/2000 controls and 5000 cases/5000 controls for a binary trait) for a range of LD (in

[1]Department of Psychiatry, University of Hong Kong, PokFuLam, Hong Kong; [2]Centre for Genomic Sciences, University of Hong Kong, PokFuLam, Hong Kong; [3]State Key Laboratory for Cognitive and Brain Sciences, University of Hong Kong, PokFuLam, Hong Kong; [4]Centre for Reproduction, Development and Growth, University of Hong Kong, PokFuLam, Hong Kong
*Correspondence: Dr M-X Li, Department of Psychiary, The University of Hong Kong, Room 1-05H, 1/F, The Hong Kong Jockey Club Building for Interdisciplinary Research, 5 Sassoon Road, PokFuLam, Hong Kong. Tel: +852 2831 5105; Fax: +852 2855 1345; E-mail: mxli@hku.hk
or Professor PC Sham, Department of Psychiatry and Centre for Genomic Sciences, The University of Hong Kong, Room 6-05A, 6/F, The Hong Kong Jockey Club Building for Interdisciplinary Research, 5 Sassoon Road, PokFuLam, Hong Kong. Tel: +852 2831 5425; Fax:+852 2818 5653; E-mail: pcsham@hku.hk
[5]These authors contributed equally to this work.
Received 11 November 2014; revised 8 June 2015; accepted 3 July 2015; published online 26 August 2015

terms of r^2, from 0 to 1) and allele frequencies (from 0.05 to 0.95), under Hardy–Weinberg equilibrium. We then performed an association analysis for each of the two variants to obtain two normal test statistics (ie, *z*-scores). We repeated this procedure 10 000 times for each set of parameters and computed the covariance of the normal test statistics of the two variants empirically under each scenario. We approximated the covariance of the normal test statistics with a polynomial of r^2 in R and considered the most parsimonious model that maximized adjusted R^2.

Mean and variance of normal test statistics of untyped variants conditional on the normal test statistics of neighboring typed variants in LD and the covariance of normal test statistics of all variants. Assume that there are n untyped variants and m typed variants and their normal test statistics are x_{untyped} and x_{typed}, respectively, the mean vector of these $m+n$ statistics μ and the covariance matrix of these statistics $\sum$ can be partitioned as follows:

$$\mu = \begin{bmatrix} \mu_{\text{untyped}} \\ \mu_{\text{typed}} \end{bmatrix} \text{with sizes} \begin{bmatrix} n \\ m \end{bmatrix}$$

$$\sum = \begin{bmatrix} \Sigma_{\text{untyped}\cdot\text{untyped}} & \Sigma_{\text{untyped}\cdot\text{typed}} \\ \Sigma_{\text{typed}\cdot\text{untyped}} & \Sigma_{\text{typed}\cdot\text{typed}} \end{bmatrix} \text{with sizes} \begin{bmatrix} n\times n & n\times m \\ m\times n & m\times m \end{bmatrix}$$

where μ equals 0 under the null and $\sum$ is approximated using a polynomial of pair-wise r^2 between the corresponding variants determined previously. The distribution of x_{untyped} given $x_{\text{typed}} = a$ (where a is obtained from the probit function in case only *P*-values are available) follows a multivariate normal distribution with estimated mean

$$\bar{\mu} = \mu_{\text{untyped}} + \sum\nolimits_{\text{untyped}\cdot\text{typed}} \sum\nolimits^{-1}_{\text{typed}\cdot\text{typed}} \left(a - \mu_{\text{typed}}\right)$$
$$= \sum\nolimits_{\text{untyped}\cdot\text{typed}} \sum\nolimits^{-1}_{\text{typed}\cdot\text{typed}} a$$

and estimated covariance matrix

$$\bar{\sum} = \sum\nolimits_{\text{untyped}\cdot\text{untyped}} - \sum\nolimits_{\text{untyped}\cdot\text{typed}} \sum\nolimits^{-1}_{\text{typed}\cdot\text{typed}} \sum\nolimits_{\text{typed}\cdot\text{untyprd}'}$$

(Note that given μ^-, we can compute the *P*-values for untyped variants from the cumulative distribution function).

Comparison with IMPUTE2 and DIST. We compared the imputation performance of FAPI with IMPUTE2 (with prephasing), a leading method in genotype imputation, and DIST,[15] a similar algorithm that imputes summary statistics for untyped common variants, using the genotype data of the 1958 Birth Cohort and National Blood Services ($n = 2938$) in the Wellcome Trust Case Control Consortium 1 (WTCCC1). We used chromosome 1 data only and removed variants with minor allele frequencies $<5\%$, Hardy–Weinberg *P*-value <0.001, and/or missing rate $>1\%$. We then randomly simulated a binary trait (with an equal number of cases and controls) and a normally distributed trait, masked the data of one out of every five variants, and re-imputed their *P*-values for each trait using three different analyses:

(1) IMPUTE2: Prephasing of genotypes of the unmasked variants using SHAPEIT v2.5, followed by genotype imputation at the masked variants using IMPUTE v2.3.2 and association test at the masked variants using the imputed genotypes in SNPTEST v2.5;[16]
(2) DIST: Association test at the unmasked variants in PLINK v1.07,[17] followed by *P*-value imputation at the masked variants using DIST v1.0.0; and
(3) FAPI: Same as DIST above but with DIST v1.0.0 replaced by FAPI.

We measured their running time and maximum RAM usage on a 3.47 GHz Intel processor. In addition, we assessed their accuracy by comparing the imputed *P*-values with the actual *P*-values for the subset of randomly masked variants analyzed by all three methods. 1000G European Phase I interim data was used as reference, and badly imputed genotypes and *P*-values with a score <0.6 were removed.

To demonstrate that FAPI also works outside of simulated data, we also analyzed a previously published real in-house GWAS data set of schizophrenia[18] using FAPI. We performed the same quality-control procedure in this data set as in the simulated data.

Detecting abnormal associations

Assume that a variant has an observed *P*-value p_o (corresponding to a normal test statistic x_o) and m neighboring variants with known normal test statistics or *P*-values. Based on the FAPI idea, we can construct the conditional normal distribution, $N(\theta, \sigma^2)$, which has a cumulative function $F(X)$, of the test statistic at that variant given the normal test statistics of the m variants (which can be transformed from *P*-values by a probit function in case only *P*-values are available). We compute a metric, $Prob_C$, which is given by $F(x_o)$ if $x_o < \theta$; otherwise $1 - F(x_o)$. When the variant and its m neighboring variants are in strong LD, x_o is expected to be close to θ with high confidence. If genotyping errors or other artifacts are present at that variant, the observed *P*-value p_o will deviate from expected and this leads to a small $Prob_C$.

Comparison with LD-PAC. We compared the $Prob_C$ criterion of detecting abnormal associations with LD-PAC,[14] a method that quantifies the evidence of spurious associations based on likelihood ratio, using computer simulations. Consider a disease-causing bi-allelic variant in LD with five other bi-allelic variants and their genotypes are under Hardy–Weinberg equilibrium. Given the allele frequencies of these variants, say 0.4, a variety of pairwise LD coefficient r among these loci, disease prevalence (5%), inheritance model (multiplicative), and relative allelic risk, the genotypes and case–control status of a population of 5 million individuals were simulated similarly as above. A sample of 2000 cases and 2000 controls were randomly drawn without replacement from the population and with a genotyping error rate of ω%; ω% of the genotypes at the risk locus in each simulated data set were altered. Allelic association test was then used to compute the association *P*-value at the risk locus in the sample. A hypothesis test using the proposed metric was used to assess whether the computed *P*-value is different from expected (ie, imputed *P*-value in FAPI). Under the null hypothesis of no abnormal associations, the observed *z*-statistic of a variant should come from the conditional distribution derived from the *P*-values of neighboring typed variants, and a large deviation of the observed *z*-statistic from expected indicates abnormal association. We used $Prob_C \leq 0.025$, which is expected to give a type-I error rate of 5%, to reject the null hypothesis that the observed association is normal. We repeated this procedure to generate 1000 data sets and counted the proportion of rejected hypotheses. A relative allelic risk of 1.0 and 1.2, as well as a genotype error rate of 0, 10, 30, and 50%, were considered in the simulations. The same simulation procedure was also conducted to evaluate the performance of LD-PAC[14] with default parameters except that the allelic association was performed by another method[19] as suggested by the LD-PAC authors.

RESULTS

Fast and accurate *P*-value imputation

We introduced a FAPI method to directly infer the association *P*-values at untyped common variants from the association *P*-values of neighboring typed variants in LD (see Materials and methods). Briefly, we obtained the conditional means and (co)variances of the normal test statistics at untyped variants given the observed normal test statistics (transformed from the observed association *P*-values in case only *P*-values are available) at the typed variants and the known covariance among the normal test statistics of typed and untyped variants, assuming that the test statistics follow a multivariate normal distribution (see Materials and methods). The *P*-values at the untyped variants can then be obtained based on the conditional distribution of the normal test statistics.

We found from simulations that the covariance of normal test statistics of any two common variants can be approximated by a fourth-order polynomial of squared genotype correlation (r^2) of the two common variants, when we considered the most parsimonious model that maximized adjusted R^2 (see Figure 1b and Supplementary Table S1). It should be noted that Figure 1b includes variants with a range of allele frequencies from 0.05 to 0.95, strongly suggesting that the approximation is independent of allele frequencies. This

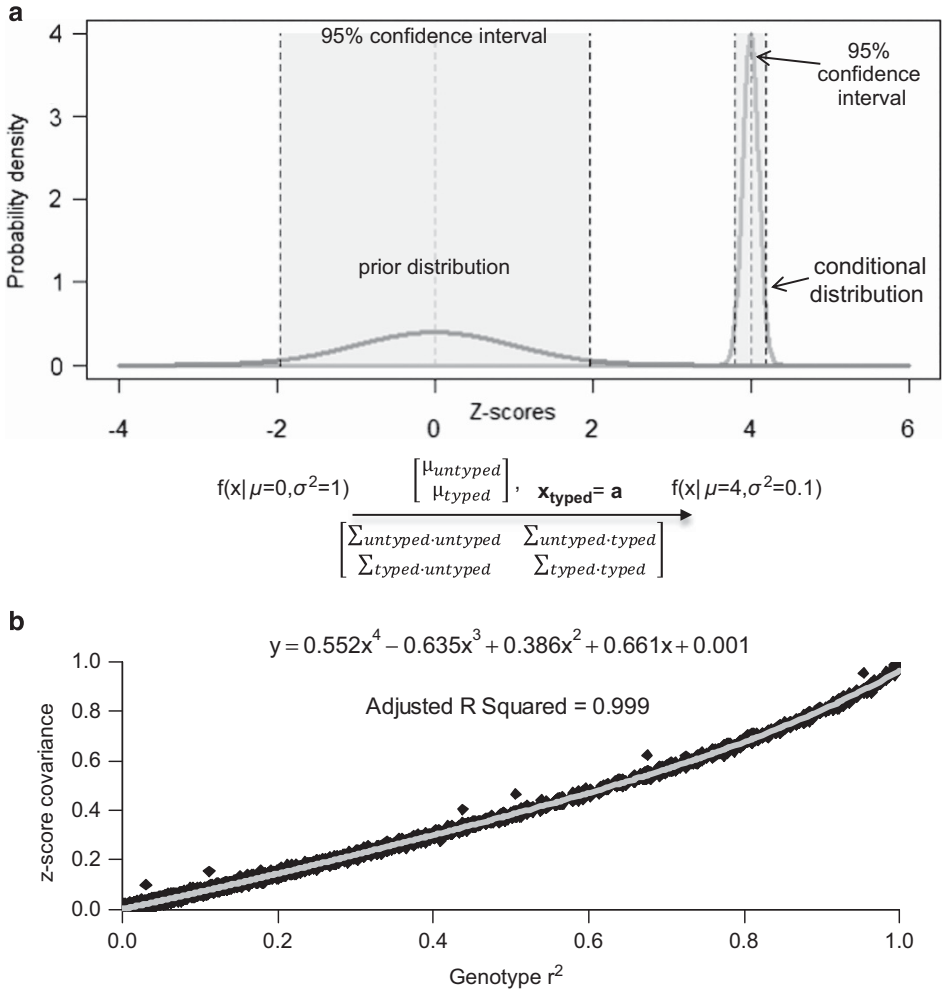

Figure 1 Diagrams illustrating the FAPI approach. (**a**) Prior and posterior distributions of a standard normal test statistic (ie, *z*-score) at an untyped variant. The prior distribution of a *z*-score of an untyped variant follows a standard normal distribution. Given the *z*-scores of its neighboring typed variants in strong LD, the conditional distribution of the *z*-score of the untyped variant has a variance <1 and thus a narrower 95% confidence interval. (**b**) The relationship between genotype correlation and covariance of *z*-scores by curve fitting.

approximation works for both binary and quantitative traits as well as for different sample sizes (see Supplementary Figure S1).

Comparison with IMPUTE2 and DIST

We compared the imputation performance of FAPI with IMPUTE2 (with prephasing), a leading method in genotype imputation, and DIST,[15] a similar algorithm that imputes summary statistics for untyped variants, using the chromosome 1 genotype data (number of sample = 2938; number of variants = 28 369) provided by the WTCCC1. We randomly simulated a binary trait (with an equal number of cases and controls) and a normally distributed trait, masked the data of one out of every five variants, and re-imputed their *P*-values for each trait using the three imputation approaches (See Materials and methods).

We also assessed the accuracy of the three approaches by comparing their imputed *P*-values with the actual *P*-values for the subset of masked variants analyzed by all three methods (Figure 2). As expected, *P*-values from IMPUTE2 were the most accurate among the three, with a correlation of 0.98 with the actual *P*-values. Nearly 86% and all of the IMPUTE2-imputed *P*-values lie within 0.1 and 0.5 $\log_{10}$ units of their corresponding actual *P*-values, respectively. FAPI was similar in accuracy compared with IMPUTE2, as indicated by a correlation of 0.96 of the imputed *P*-values with the actual ones. About 76% and all of its imputed *P*-values lie within 0.1 and 0.5 $\log_{10}$ units of their corresponding actual *P*-values, respectively. The accuracy of the imputed *P*-values was similar in a real data set with a real phenotype (see Supplementary Figure S2). DIST performed the worst as the imputed values were modestly correlated with the actual ones (around 0.72 for both traits) and even seriously underestimated (as seen from the large number of points below the diagonal in the scatterplots). Only around 50, 92, and 98% of its imputed *P*-values could lie within 0.1, 0.5, and 1 $\log_{10}$ units of their corresponding actual *P*-values, respectively.

Table 1 shows the running time and maximum RAM usage of all three methods measured on a 3.47 GHz Intel processor. FAPI was the fastest and finished the task in 10 min. The running time of DIST almost doubled that of FAPI, whereas the traditional method was the slowest and ran 200 times slower than others. In terms of RAM usage, DIST was the most memory efficient and used 1 G of RAM, whereas FAPI used 8 G and was the least efficient.

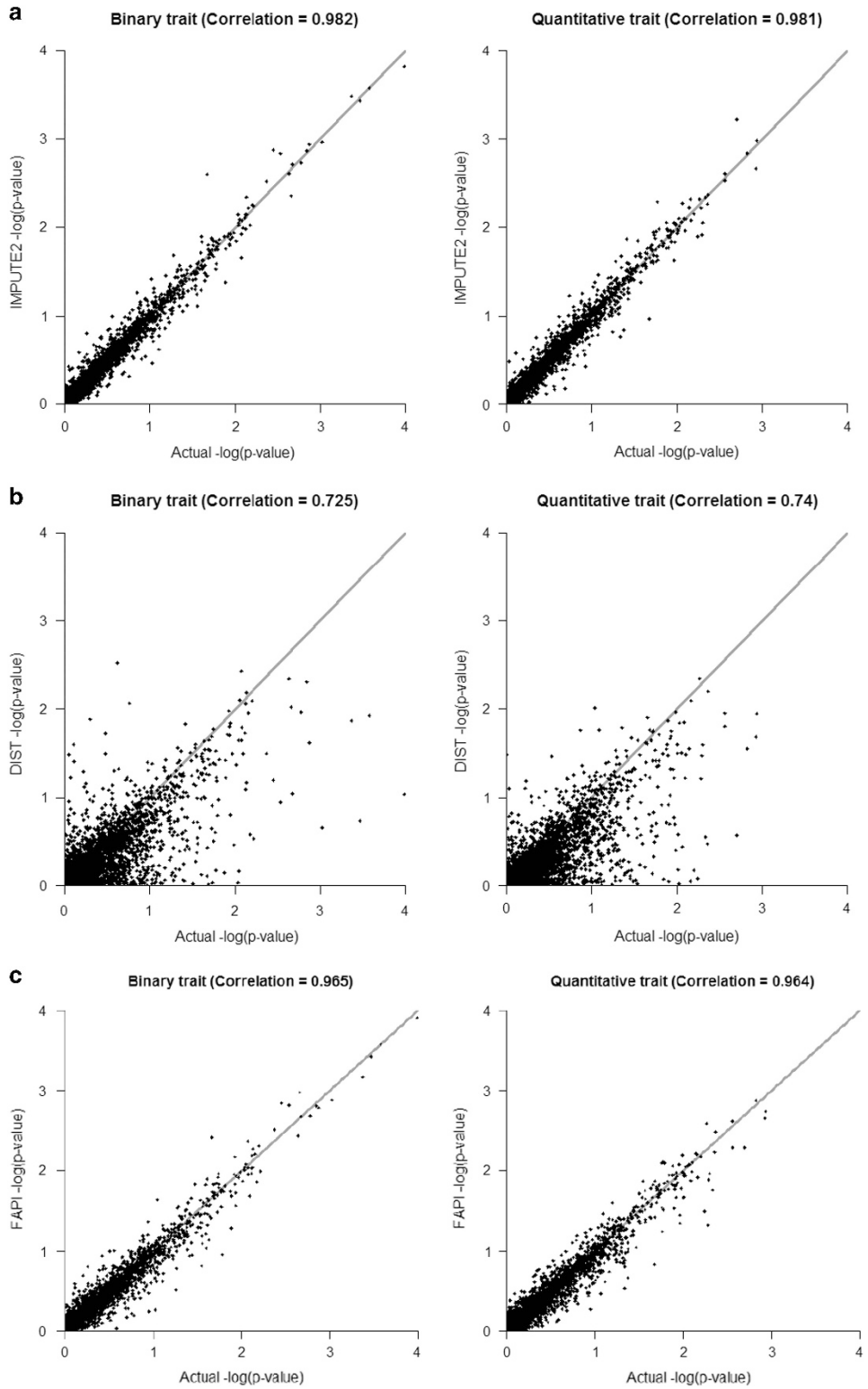

Figure 2 Scatterplots of actual $-\log_{10}$(*P*-value) against imputed $-\log_{10}$(*P*-value) by (**a**) IMPUTE2; (**b**) DIST; and (**c**) FAPI. The genotype data on chromosome 1 in WTCCC1 was used (number of sample = 2938). We randomly simulated a binary trait (with an equal number of cases and controls) as well as a normally distributed trait, masked the data of one out of every five variants and re-imputed their *P*-values for each trait by each of the three imputation approaches.

Detecting abnormal associations

We also derived a metric $Prob_C$ to detect abnormal association at a variant using the association *P*-values at its neighboring variants in LD. In brief, a conditional normal distribution of its normal test statistic is constructed based on the *P*-values of its neighboring variants in LD using the FAPI idea and the observed test statistic, z_o, at that variant is

examined for deviation from the mean of the distribution using the cumulative distribution function (see Materials and methods). A small $Prob_C$ suggests that the observed test statistic (or association *P*-value) is different from expected and thus may require further attention.

We compared our approach of detecting abnormal associations with LD-PAC,[14] a method that quantifies the evidence of spurious associations based on likelihood ratio, using simulations. We simulated a series of situations in which genotyping errors were introduced at various rates in a sample of 2000 cases and 2000 controls to see whether the metric can detect the abnormal associations introduced (see Materials and methods). A $Prob_C$ threshold of 0.025 was used to reject the null hypothesis of no abnormal associations, while LD-PAC was run with the default parameters. Table 2 shows the type-I error rates and powers of our approach and LD-PAC. When no genotyping error was introduced, at most 5 and 2.5% of the associations were flagged as abnormal by FAPI and LD-PAC, respectively, no matter whether a null or risk locus was examined. However, for a genotyping error rate of 10%, our approach had around 20–40% power of detecting the abnormal associations when a null locus was simulated and around 50–70% power when a risk locus was simulated, whereas LD-PAC only had around 3–5% in both scenarios. The power of FAPI increased further when the genotyping error rate increased, but there was little increase in power for LD-PAC. The conclusion is independent of sample size.

Implementation

All the methods proposed in this paper are implemented in a software tool freely available at http://statgenpro.psychiatry.hku.hk/fapi. The software tool has three major functions:

(1) FAPI: imputing association *P*-values at untyped variants;
(2) $Prob_C$: detecting associations with serious genotyping errors; and
(3) M-A: performing a meta-analysis of GWAS summary statistics at both untyped and typed variants by imputing them into a common reference panel and combining them using a sample-size weighted combination method.

It can run on most operating systems and automatically checks for updates and downloads necessary resources before performing the user-specified functions. For a GWAS with around 360 000 typed variants, imputation of association *P*-values at 5.7 million untyped variants found in the 1000G Project using FAPI requires ~72 min CPU time and 10 GB of memory on a single 3.1 GHz Intel processor.

DISCUSSION

In this paper, we propose a method, FAPI, for assessing the associations at untyped common variants that is fast, accurate, and independent of raw genotype data. In our simulations, it finished imputing chromosome 1 based on HapMap data in less than an hour, but the same task required 3 days for IMPUTE2 even with prephasing. At the same time, the calls produced by FAPI were unbiased and had little loss in accuracy compared with those produced by IMPUTE2. In addition, based on the FAPI idea, we develop a metric to detect abnormal associations at a target variant using the association *P*-values at neighboring variants in LD. Our method was shown to have a significantly greater power compared with LD-PAC, a method that quantifies the evidence of spurious associations based on likelihood ratio. These methods are implemented in a user-friendly software tool freely available at http://statgenpro.psychiatry.hku.hk/fapi.

Compared with DIST, which also models summary statistics of untyped variants using a multivariate normal distribution, FAPI has some important advantages. First, DIST wrongly approximates the variance–covariance matrix of the summary statistics by the genotype correlation given, whereas FAPI accurately models the relationship

Table 1 Running time and maximum RAM usage of three imputation methods to finish an imputation task

	IMPUTE2	*DIST*	*FAPI*
Running time	30 h (SHAPEIT v2.5)	0.5 min (PLINK v1.07)	0.5 min (PLINK v1.07)
	+37 h (IMPUTE v2.3.2)+5 h (SNPTEST v2.5)	+18 min (DIST v1.0)	+9 min (FAPI)
	≈72 h	≈0.3 h	≈0.2 h
Maximum RAM usage	5 G	1 G	8 G

Table 2 Type-I error rates and powers of the $Prob_C$ approach and LD-PAC

RAR	*LD*[a]	*Genotyping error rate*							
		0%		*10%*		*30%*		*50%*	
		$Prob_C$	*LD-PAC*	$Prob_C$	*LD-PAC*	$Prob_C$	*LD-PAC*	$Prob_C$	*LD-PAC*
1	0.7	0.020	0.019	0.180	0.028	0.503	0.040	0.690	0.046
	0.8	0.018	0.013	0.251	0.028	0.592	0.041	0.768	0.048
	0.9	0.017	0.010	0.407	0.028	0.727	0.040	0.856	0.048
1.2	0.7	0.051	0.021	0.494	0.047	0.909	0.134	0.969	0.368
	0.8	0.036	0.025	0.569	0.049	0.929	0.135	0.976	0.365
	0.9	0.025	0.024	0.703	0.046	0.959	0.123	0.986	0.353

Abbreviation: RAR, relative allelic risk. A $Prob_c$ threshold of 0.025 was used to reject the null hypothesis of no abnormal associations, while LD-PAC was run with the default parameters. The disease model was multiplicative with allele frequency 0.4 and prevalence 0.05.
[a]LD r with neighboring variant.

using an empirical fourth-order polynomial function. Second, DIST accepts only *z*-scores as inputs, whereas FAPI accepts *P*-values and so is more flexible. Finally, FAPI provides more useful functions derived from the imputation function (including reliability checking for *P*-values and meta-analysis) to facilitate genetic association analysis.

As FAPI relies on LD, so it certainly does not work for low frequency and rare variants that do not exhibit strong LD with either rare or common variants. Also, the results must be sensitive to the provided LD structure, which can be quite different among various reference data sets. The simple 'best-match' strategy, which requires the ancestry of the GWAS sample matches closely with that of the reference panel, is what we recommend at the moment. However, sometimes a clear perfect match may not be available, especially for GWAS in admixed populations, and a popular strategy in genotype imputation to deal with this is to use a 'cosmopolitan' reference panel including as many haplotypes as possible. This strategy is able to work in genotype imputation as genotype imputation can search for haplotype segments shared between each GWAS sample and a reference panel of densely typed individuals within a limited genomic region only. But whether this will work in FAPI requires further investigation.

This fast and easy imputation tool may encourage many more studies to make use of publicly available data for meta-analysis studies of many complex diseases. Because of too many missing genotypes according to different genotyping arrays, genotype imputation with raw genotypes is often a prelude to a meta-analysis study. However, because of the difficulties in sharing raw genotypes in practice, this procedure is often very inefficient. With FAPI, one only needs the association *P*-values, which can even often be directly downloaded from public domains. With the integrated reference data from HapMap or 1000 Genomes Projects on FAPI, the imputation and meta-analysis can be quickly performed on a desktop computer with ordinary configurations. In this way, more reliable genetic risk factors will be suggested to explain part of the missing heritability for a variety of complex diseases.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was funded by Hong Kong Research Grants Council GRF HKU 768610M, HKU 776412M, and HKU 777511M; Hong Kong Research Grants Council Theme-Based Research Scheme T12-705/11; European Community Seventh Framework Programme Grant on European Network of National Schizophrenia Networks Studying Gene-Environment Interactions (EU-GEI); the Hong Kong Health and Medical Research Fund 01121436 and 02132236; the HKU Seed Funding Programme for Basic Research 201302159006; the HKU Small Project Funding 201309176244; and The University of Hong Kong Strategic Research Theme on Genomics.

1 Barsh GS, Copenhaver GP, Gibson G, Williams SM: Guidelines for genome-wide association studies. *PLoS Genet* 2012; **8**: e1002812.
2 Bush WS, Moore JH: Chapter 11: genome-wide association studies. *PLoS Comput Biol* 2012; **8**: e1002822.
3 Sherry ST, Ward MH, Kholodov M *et al*: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001; **29**: 308–311.
4 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**: e1000529.
5 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010; **34**: 816–834.
6 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**: 210–223.
7 Browning SR: Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet* 2006; **78**: 903–913.
8 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**: 1084–1097.
9 Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A: A comprehensive evaluation of SNP genotype imputation. *Hum Genet* 2009; **125**: 163–171.
10 Marchini J, Howie B: Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11**: 499–511.
11 Zeggini E, Ioannidis JP: Meta-analysis in genome-wide association studies. *Pharmacogenomics* 2009; **10**: 191–201.
12 The International HapMap Consortium: The International HapMap Project. *Nature* 2003; **426**: 789–796.
13 The 1000 Genomes Project Consortium, Abecasis GR, Altshuler D *et al*: A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.
14 Han B, Hackel BM, Eskin E: Postassociation cleaning using linkage disequilibrium information. *Genet Epidemiol* 2011; **35**: 1–10.
15 Lee D, Bigdeli TB, Riley BP, Fanous AH, Bacanu SA: DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics* 2013; **29**: 2925–2927.
16 Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007; **39**: 906–913.
17 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
18 Wong EH, So H-C, Li M *et al*: Common variants on Xq28 conferring risk of schizophrenia in Han Chinese. *Schizophr Bull* 2014; **40**: 777–786.
19 Han B, Kang HM, Seo MS, Zaitlen N, Eskin E: Efficient association study design via power-optimized tag SNP selection. *Ann Hum Genet* 2008; **72**: 834–847.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)