

## Leveraging ancestry to improve causal variant identification in exome sequencing for monogenic disorders

Robert Brown<sup>1</sup>, Hane Lee<sup>2</sup>, Ascia Eskin<sup>3</sup>, Gleb Kichaev<sup>1</sup>, Kirk E. Lohmueller<sup>1,4</sup>, Bruno Reversade<sup>5,6,7</sup>, Stanley F. Nelson<sup>2,3</sup>, Bogdan Pasaniuc<sup>1,2,3</sup>

<sup>1</sup>Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles California, USA

<sup>2</sup>Department of Pathology and Laboratory Medicine, Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, USA

10 <sup>3</sup>Department of Human Genetics, Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, USA

<sup>4</sup>Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, California, United States of America

<sup>5</sup>Institute of Medical Biology, Human Genetics and Embryology Laboratory, A\*STAR, Singapore 138648, Singapore

<sup>6</sup>Institute of Molecular and Cellular Biology, A\*STAR, Singapore 138673, Singapore

<sup>7</sup>Department of Pediatrics, National University of Singapore, Singapore 119260, Singapore

### Abstract

Recent breakthroughs in exome sequencing technology have made possible the identification of many causal variants of monogenic disorders. Although extremely  
20 powerful when closely related individuals (e.g. child and parents) are simultaneously sequenced, exome sequencing of individual only cases is often unsuccessful due to the large number of variants that need to be followed-up for functional validation. Many approaches remove from consideration common variants above a given frequency threshold (e.g. 1%), and then prioritize the remaining variants according to their allele frequency, functional, structural and conservation properties. In this work, we present methods that leverage the genetic structure of different populations while accounting for the finite sample size of the reference panels to improve the variant filtering step. Using simulations and real exome data from individuals with monogenic disorders, we show that our methods significantly reduce the number of variants to be followed-up (e.g. a 36%

30 reduction from an average 418 variants per exome when ancestry is ignored to 267 when  
ancestry is taken into account for case-only sequenced individuals). Most importantly our  
proposed approaches are well calibrated with respect to the probability of filtering out a  
true causal variant (i.e. false negative rate, FNR), whereas existing approaches are  
susceptible to high FNR when reference panel sizes are limited.

## Introduction

Vast decreases in the cost of exome sequencing have allowed for major advancements in  
the identification of causal variants for rare monogenic traits and disorders<sup>1-4</sup>. Although  
each individual carries 20,000-24,000 single nucleotide variants in their exome that differ  
from the human reference genome, most of these variants are common in the population or  
40 do not have a damaging effect and therefore are unlikely to explain a rare monogenic trait.  
Finding causal variants for monogenic traits through exome sequencing follows a two-step  
approach. First, variants that are too common to be consistent with the prevalence of a rare  
disorder are discarded<sup>2</sup>. Variants that remain under consideration are then prioritized based  
on frequency, functional, structural and conservation properties<sup>5; 6</sup>, with more recent  
approaches using cross species comparisons<sup>7</sup> or a combination of scores from several  
stand-alone methods or other data sources<sup>8-11</sup>. When pedigrees or cohorts of patients (with  
the same disorder) and their close relatives (e.g. parents or siblings) are sequenced, this  
two-step approach has proven to be extremely powerful in refining the list of prioritized  
variants to just a few variants<sup>2-4; 12-16</sup>. However, when only the case individual is available  
50 for sequencing, the number of variants that are left for follow-up in functional analysis is  
often on the order of hundreds of plausible variants<sup>10; 17; 18</sup>, thus making it difficult to  
identify the causal variant(s).

In this work we present methods that leverage population structure (i.e. the variability in variant frequencies across populations) to improve the performance of exome sequencing studies of monogenic traits. Although it is commonly accepted in studying all types of disease that large well-matched control cohorts are important in limiting false positives<sup>19</sup>, monogenic disease studies often estimate variant frequencies across large databases of human variation at the level of continental ancestry (e.g. the Exome Variant Server<sup>20</sup> European or African American data) instead of reference panels more finely tuned to the ancestry of the case individual(s); this is especially true when it is not practical to obtain a well-matched control group. Here, we investigate the use of matched allele frequency estimates (typically at the level of a country) to the ancestry of the sequenced individual<sup>21; 22</sup>. Since rare variants tend to be present in only a few closely related populations and absent from the rest<sup>23-27</sup>, the frequency estimates of alleles present in a given population will show a downward bias if estimated across individuals of multiple ancestries. That is, a variant might appear rare (<1%) across many populations, when in reality it is only rare in most populations and less rare or even common (>1%) in a few (see Supplemental Figure 1). As an example, consider variant rs17046386, it is generally rare or non-existent in non-Africans and present in Africans and those of African descent (see Figure 1). Based on European reference panels (or a global ancestry-unaware panel), this variant would not be excluded even though it's relatively high frequency in Africans makes it unlikely to be pathogenic.

The limited size of existing reference panels, especially when defining ancestry at the level of a country, induces significant statistical variance in the allele frequency estimates that must be accounted for. For example, a variant with a true frequency of 0.5%

has an 8% chance of being observed with frequency  $>1\%$  (and thus discarded) in a reference panel of 100 individuals as compared to only a 0.1% chance of being disregarded when frequencies are estimated over 1,000 individuals. To account for this effect, we introduce an false negative rate (FNR) filtering technique that protects against filtering a potential causal variant at a pre-specified significance level.

Starting from the 1000 Genomes<sup>24</sup> and the NHLBI Exome Sequencing Project (ESP) Exome Variant Server (EVS)<sup>20</sup> data sets, we use simulations to show that correctly matching reference panels to the country-level ancestry of the sequenced individual reduces the number of candidate causal variants from 418 to 351 in case-only simulations. Furthermore, by comparing frequencies across multiple populations, more variants can be removed (267 remaining if removing variants above the 5% FNR threshold in any population). In addition to case-only simulations, we simulate trio data for dominant and recessive diseases and show that ancestry aware filtering yields similar improvement under these scenarios. We validate our approach using exome-sequencing data from 20 real individuals with monogenic disorders for which the causal variants are known. In this data, without ever filtering out the true causal variant, our approach successfully reduces the mean number of heterozygous variants to be functionally tested from 750 to 604 (FNR $<5\%$ ) when only matching to one population and to 435 (a 42% reduction) when leveraging all population data. Our results demonstrate that existing filtering pipelines for exome sequencing studies of monogenic traits can be significantly improved by taking ancestry into account. Finally, our results suggest that utilizing narrowly defined ancestry matched reference panels (i.e. at the country level) overcomes the reduction in performance due to higher statistical noise from the smaller panels.

## Methods

### 100 Datasets

The 1000 Genomes Project<sup>24</sup> has produced a public catalog of human genetic variation through sequencing individuals from different populations across the world. We use the 1000 Genomes individuals (1,092 in total) to evaluate the effectiveness of various filtering approaches (we removed the IBS from simulations due to the small number of individuals). Following the commonly accepted assumption that 85% of causal variants for monogenic traits are found in the exome<sup>29</sup>, we restrict our analysis only to variants found in the coding regions of autosomal chromosomes. For admixed individuals we matched ancestry locally using the local ancestry calls provided by the 1000 Genomes Project (the consensus calls of four different local ancestry inference methods<sup>30-33</sup>). Damaging scores for each single nucleotide variant were computed using the KGGSeq software with default parameters<sup>10</sup> that combines the functional annotation scores in the dbNSFP<sup>34</sup> database v2.0.

The NHLBI Exome Sequencing Project (ESP) Exome Variant Server (EVS) has released allele counts on 4,300 European-Americans and 2,203 African Americans<sup>20</sup>. PolyPhen2 scores are also provided for missense variants; the probably and possibly damaging predicted variants were used in the EVS-based analyses. We simulate different size reference panels using a binomial sampling with the allele frequencies estimated across all the European (or African American) EVS data as the true frequency.

To compare results seen in simulations to real data, we used exomes of 101 individuals with self-reported countries of origin including Turkey, Jordan, Tunisia, Egypt, Israel, Iran, Syria and Palestine. We grouped these individuals into a single supplemental reference population for estimating best matching allele frequencies. Of the 101

individuals, 20 had identified causal variants of monogenic disorders. Nine individuals were known to harbor heterozygous variants in genes causing autosomal dominant disorders, 10 individuals had homozygous variants, and one individual had potential compound heterozygous variants in genes causing an autosomal recessive disorder. We filtered out all variants except those with damaging annotations: splice acceptors, stop gains, frame shifts, stop losses, initiator codon changes, inframe insertions, inframe deletions, missense variants, splice region variants and KGGSeq predicted damaging variants. For consistency, the individual being examined was removed from the rest of the  
130 data that served as a reference panel. For comparison purposes, we included these individuals when estimating average frequencies across all populations in the 1000 Genomes project for the real data only.

### **False negative rate estimation**

We estimate the probability of filtering out a true causal variant (false negative rate) at a given frequency threshold as a function of a given reference panel and the maximum true allele frequency of the causal variant. The filtering threshold can be adjusted in order to provide a desired FNR. Let  $t$  be the nominal frequency threshold that is used for filtering. We define the corresponding FNR at this threshold as:

$$FNR(t) = \frac{\int_0^{\max(f_c)} fP(f_{ref,N} > t | f)P(f)df}{\int_0^{\max(f_c)} fP(f)df}$$

140 where  $f$  is the frequency of the variant in the population,  $\max(f_c)$  is the maximum assumed frequency of the causal variant in the population,  $P(f)$  is the proportion of variants with frequency  $f$  in the population and  $P(f_{ref,N} > t | f)$  is the probability that a variant with

frequency  $f$  is observed at a frequency greater than  $t$  in the reference panel of  $N$  individuals randomly drawn from the population.

The FNR computation at a threshold  $t$  requires knowledge about the distribution of variants across all frequencies in the population; this can be estimated from population genetic theory under various demographic assumptions<sup>25; 35-39</sup> or empirically from the data. In this work, we estimate the distribution  $P(f)$  from reference panel allele counts and perform the above integration across the observed site frequency spectrum as follows:

$$150 \quad FNR(t) = \frac{\sum_{f_i \leq \max(f_c)} f_i P(f_{ref,N} > t | f_i) P'(f_i)}{\sum_{f_i \leq \max(f_c)} f_i P'(f_i)}$$

Here  $f_i$  represents each of the unique allele frequencies observed in the reference panel of  $N$  individuals and  $P'(f_i)$  represents the proportion of variants in the reference panel that have estimated frequency  $f_i$ . The function sums over the discrete frequency values less than the  $\max(f_c)$ .  $P(f_{ref,N} > t | f_i)$  is modeled as a binomial draw with the frequency of success equal to  $f_i$  and the number of draws equal to the number of allele counts ( $2N$ ). Since the integration is over a discrete space we calculate the probability that the number of success is greater than the threshold times  $2N$ . The  $\max(f_c)$  must also be larger than  $(2N)^{-1}$  for the FNR estimate to function. Having estimated a FNR for all possible filtering thresholds  $t$ , we propose to filter variants using the minimum frequency threshold  $t$  such that  $FNR(t) <$   
160 5%.

### **Leveraging population structure for improved filtering**

We compared different filtering approaches using simulated exomes from the 1000 Genomes data. For non-admixed individuals (e.g. not African-Americans) we use simulations where four haplotypes are drawn from a specific 1000 Genomes population

and then paired to form two parental genomes. An offspring genome was then simulated from the parental genomes by passing variants according to Mendelian inheritance under the assumption that all sites are independent. We filtered out variants that do not result in an amino acid change or do not create or remove a stop codon. We compare three different scenarios likely to be encountered in a clinical diagnostic setting with 40 simulated  
170 offspring genotypes per scenario for each 1000 Genomes population (not including the IBS). The Case-Only scenario assumes there is no information on parental genotypes and there is no known mode of inheritance. The Trio-Dominant scenario assumes that both parental exomes are sequenced and the offspring and one of the parents has the disorder. The Trio-Recessive scenario assumes that both parents are exome sequenced and heterozygous for the causal allele and that the offspring had two copies of the causal allele. We repeated the same analysis only using the variants for which there is a predicted damaging KGGSeq score (the variants in dbNSFP).

The cosmopolitan filtering approaches (*NoAncestry,  $f > 1\%$*  and *NoAncestry,  $FNR < 5\%$* ) estimate allele frequencies and FNRs across all 1000 Genomes individuals. The  
180 key intuition behind these approaches is that statistical noise is decreased with large reference panels at the cost of ignoring population structure. *NoAncestry,  $f > 1\%$*  filters out variants with allele frequency  $> 1\%$  without regard for the FNR, whereas *NoAncestry,  $FNR < 5\%$*  filters out variants above a threshold determined to ensure a desired FNR. We propose two ancestry-aware filtering approaches. The first (*PopMatched,  $FNR < 5\%$* ) uses only the reference individuals from the sub-continental population (country-level, see 1000 Genomes<sup>24</sup>) that best matches the sequenced individual. The second approach (*AllPop*) uses data across multiple sub-continental populations, by simply requiring that variants not



be filtered out at each population's FNR <5% threshold in each of a randomly chosen set of sub-continental populations. Sets sizes ranged between none and all the 1000 Genomes populations. The intuition behind this approach is that a variant common in at least one population is unlikely to be causal for monogenic disorders.

For admixed populations (MXL, PUR, CLM and ASW) we only assessed the Case-Only scenario and used the genotypes of the real admixed individuals from 1000 Genomes. In each individual at loci that are homozygous for African, European or Native American ancestry, we used continental allele frequency estimates obtained by averaging across the CEU, FIN, GBR and TSI for European frequencies, the YRI and LWK for African frequencies and the JPT, CHB and CHS for Native American frequencies. In local ancestry heterozygous regions we used a 50-50 weighting of the matching continental frequencies. In order to determine the FNR threshold, we first calculated the FNR threshold for the CEU, JPT and YRI and used the maximum threshold of those three populations. This is an overestimate of the true threshold because the allele frequencies will be downwardly biased and there is higher confidence in the allele frequency estimates due to larger reference panel sizes.

## Results

### Modeling statistical uncertainty increases filtering efficacy

We use simulations from the European-American Exome Variant Server (EVS)<sup>20</sup> dataset to assess the increase in performance of FNR-based filtering as a function of panel size in a homogenous population. Figure 2a shows the threshold on the observed frequency as a function of reference panel size such that a FNR of 5% is maintained under different maximum frequencies of the causal variant ( $\max(f_c)$ ). As expected, the frequency threshold

that maintains a 5% FNR increases as reference panel sizes decrease (see Figure 2a). As the reference panel size is increased the filtering threshold that maintains a 5% FNR approaches  $\max(f_c)$ .

Figure 2b shows the average number of damaging or possibly damaging variants for follow-up (according to the EVS Polyphen2<sup>40</sup> annotation scores) below the <5% FNR threshold. There is a diminishing return in filtering efficacy for reference panels larger than 500 individuals. This shows that as the assumed maximum frequency of the true causal variant ( $\max(f_c)$ ) decreases, the 5% FNR threshold and number of variants for follow-up per individual also decrease.

220        Next, we investigated the effect of different maximum allele frequencies of the causal variant ( $\max(f_c)$ ) and reference panel sizes on frequency-based and FNR-based filtering methods. Table 1 shows that for small reference panels (e.g. ~100 reference individuals, approximately the size of a 1000 Genomes country-level population) the frequency-based approach is mis-calibrated with respect to the probability of filtering out the true causal variant. Although the approach that maintains a proper  $\text{FNR} < 5\%$  significantly increases the number of variants for follow-up from 298.0 to 724.1 on average, this is necessary as it reduces the FNR from 25% for the frequency-based approach to 5% for the FNR-based approach. In contrast, when reference panels are large, the frequency-based approach is too conservative (due to the lack of a FNR calculation)  
230        leading to an increased number of follow-up variants. For example, the  $\text{FNR} < 5\%$  approach reduces the number of variants for follow-up from 311 to 150 variants, under the assumption that the  $\max(f_c)$  is 0.1% and 2500 reference individuals. Qualitatively similar

results were observed when simulating exomes from the EVS African-American data (see Supplemental Figure 2).

## Leveraging ancestry to increase filtering performance

We next assessed the performance of filtering with or without including ancestry information to account for the highly structured nature of rare variants<sup>23-25; 41</sup> in simulations of non-admixed individuals starting from the 1000 Genomes data. We compared several methods for filtering variants in exome studies under Case-Only, Trio-Dominant and Trio-Recessive scenarios. Under all disease architecture and trio scenarios the methods that take ancestry into account outperform methods that do not (Table 2). The *PopMatched*, *FNR* <5% and *NoAncestry*, *FNR* <5% each filter based on a threshold determined to ensure a FNR of at most 5%. The difference is that the *NoAncestry* method uses the entire 1000 Genomes data set as a single reference population whereas *PopMatched* uses only the 1000 Genomes individuals from the same population as the simulated case individual as references. The population matching decreases the number of variants by at least 15% under the Case-Only, Trio-Dominant and Trio-Recessive scenarios; this is true when only using KGGSeq predicted damaging variants as well (Table 2). This improvement comes regardless of the fact that the <5% FNR filtering threshold in the *PopMatched* method (average filtering threshold is 2.2% and always >2%) is on average twice that of that *NoAncestry* method (filtering threshold is 1.1%) due to the significantly reduced reference panel size (average of 93 individuals per 1000 Genomes population). This demonstrates that the benefit of better population matching outweighs the cost of higher statistical noise resulting from the small reference panels. The Trio-Dominant scenario has approximately half as many variants for follow-up as the Case-Only scenario (288 compared to 582 for

the *PopMatched*,  $FNR < 5\%$  method, Table 2), this would be expected assuming one parent were also affected. The Trio-Recessive scenario, simulated without inbreeding, shows very few variants for follow-up ( $< 6$  under all scenarios and filtering methods), this is expected when looking for rare variants appearing homozygotically in an individual (Table 2).

260 We observe that the *NoAncestry*,  $FNR < 5\%$  approach leads to a slightly increased number of variants that need to be functionally followed up over the *NoAncestry*,  $f > 1\%$  approach (Table 2). The increase in the number of variants is necessary to attain a correct 5% FNR rate (*NoAncestry*,  $f > 1\%$  attains a FNR of 6.0%).

Finally, we assessed a method that filters variants observed in any population above each population's 5% FNR threshold in addition to the best-matched population (*AllPop*). All FNR-based approaches assume that  $\max(f_c) < 1\%$  in all populations. The *AllPop* approach does not account for multiple testing, but does demonstrate that as more population are randomly sampled, many variants can be eliminated due to high frequency in some populations that are inconsistent with causal allele frequency assumptions. This method shows a 40% reduction from the *NoAncestry*,  $FNR < 5\%$  method in the Case-Only scenario (Table 2). In Figure 3, we investigated how the number of follow-up variants decreases as a function of the number of reference populations available in the Case-Only scenario. With data from more populations available, there is a greater chance of observing that a given variant is common in at least one population, and is thus unlikely to cause a rare disorder. This demonstrates that there are a significant number of variants that are common in some populations but rare in the population of the case individual. Careful selection of a few populations genetically distant from a case individual's population can

270

further reduce the number of variants remaining beyond what is obtainable through only the *PopMatched*,  $FNR < 5\%$  method and will reduce the effects of multiple testing.

280 Finally, we observe a similar pattern of improved performance when restricting to functionally damaging variants as predicted by KGGSeq (Table 2). Taken together, this shows that in all the aforementioned scenarios the improvements of ancestry-aware filtering do not come exclusively from variants that would have been filtered later due to non-damaging predictions, but rather, from variants removed proportionally from damaging and non-damaging predicted variants.

### **Ancestry-aware filtering in admixed individuals**

Results above were obtained using individuals of homogeneous ancestry. We extend our approach to populations of admixed ancestry (e.g. African Americans) by considering their local ancestry structure. Individuals with recent ancestry from multiple  
290 continents have genomes that are a mosaic of segments each originating from different ancestral populations. We incorporate the local ancestry structure in the filtering step with the *PopMatched-LA*,  $FNR < 5\%$  approach that matches reference panels by ancestry according to each site in an individual's genome (see Methods). This significantly lowers the number of variants for follow-up in the admixed populations as compared to the local ancestry naïve method (*PopMatched*,  $FNR < 5\%$ ) (see Figure 4). When using information from all populations in the 1000 Genomes dataset, there is improvement for all admixed populations over the method that ignores ancestry (*NoAncestry*,  $FNR < 5\%$ ) (see Figures 4 and Supplemental Figure 3). For example, in African American individuals we observe a reduction from 668 to 592 variants from just matching the local ancestry to continental  
300 populations as compared to using all 1000 Genomes data with a  $FNR < 5\%$ . There is

significant diversity within subpopulations of Mexico that may make finding well-matching reference panels difficult<sup>42</sup>. While correct matching has been shown to be very beneficial, incorrect matching may increase the number of variants remaining for follow-up. This could explain why the *PopMatched-LA, FNR <5%* approach performs worse for the MXL than the *PopMatched, FNR <5%* approach .

### **Analysis of 20 exomes of individuals with monogenic traits**

To examine the performance of the different filtering strategies when applied to actual data, we used the data from 20 real exome sequenced individuals with monogenic disorders where the causal variants have been previously identified. We assumed a  
310 maximum causal allele frequency of 1% for all cases because there was no prevalence data<sup>2</sup>. For all modes of inheritance, the number of variants in an individual for follow-up after filtering was lower when filtering with the *PopMatched, FNR <5%* and *AllPop* approaches as opposed to the *NoAncestry, f>1%* approach that does not account for the FNR (See Supplemental Table 1). For example, using our approach only 604.8 (434.6 if using all populations' data) variants need to be followed-up for a dominant disorders as compared to 749.7 for the *No Ancestry, f>1%* approach (see Table 4). Importantly, for all of the filtering schemes, the true causal variant identified in these real individuals was never filtered out. This demonstrates population matching for filtering allele frequencies significantly reduces the number of variants remaining for follow-up analysis, while still  
320 maintaining an appropriate false negative rate. It also shows that when researchers can safely assume a causal allele frequency <1% in many populations, they can substantially further reduce the number of variants remaining for follow-up analysis. This real data analysis shows the importance of being able to first filter with a well-matched reference

panel, even of just 100 individuals. In Supplemental Table 1 we report the variants remaining in each individual along with country data, reported inheritance pattern, the zygosity of the causal variant and assumed disorder architecture.

## Discussion

Exome sequencing for rare monogenic disorders has proven to be very useful for discerning the causal genes for these traits. Although extremely powerful when closely related individuals are simultaneously sequenced, current pipelines for exome sequencing of a single individual often yield too many variants to be tractable for functional follow-up. In this work, we introduce approaches that account for the finite sample size of the existing reference panels used in filtering while jointly modeling the ancestry to improve the filtering step. Both the real data analysis of 20 exomes of individuals with known monogenic disorders and the simulations show that our approaches reduce the number of variants that need to be further investigated, thus increasing the effectiveness of identifying causal variants using exome sequencing of unrelated individuals. This work demonstrates that in a clinical setting, even a small reference panel of 100 individuals from a well-matched population can have significant impact on the filtering ability of a case individual.

The current bottleneck in using population structure to help identify rare variants is the limited size of the reference panels for specific narrowly defined populations. Current reference panels show that at the 1% allele frequency level that there are significant differences between populations. However, it is difficult to assess allele frequencies below this level due to reference panel size for specific populations. This demonstrates the need for larger reference panels of more narrowly defined populations (not just continental level) in order to fully assess the structure of rare variation. As the panel sizes increase,

researchers studying monogenic disorders will be able to use smaller maximum causal allele frequencies when appropriate for the trait of interested. Furthermore, the larger reference panels will also lower filtering thresholds.

350           With the increasing availability of public databases it may be possible to obtain accurate estimates of disorder prevalence across populations. Our proposed approaches could be further extended to take this information into account by using different thresholds on the maximum frequency of causal variants ( $\max(f_c)$ ) across populations. We leave that as ongoing and future work.

## Supplemental Data

The supplemental data contains three figures and one table.

## Acknowledgements

This work is supported in part by the National Institutes of Health (R03-CA162200, R01-GM053275 to BP and T32-HG002536 to RB). AE and SFN are supported by the  
360   Genomics/Informatics Core of the UCLA MUSCULAR DYSTROPHY CORE CENTER from NIAMS (P30AR057230). B.R. is a fellow of the Branco Weiss Foundation and an A\*STAR and EMBO Young Investigator. This work was funded by a Strategic Positioning Fund for Genetic Orphan Diseases and an inaugural A\*STAR Investigatorship from the Agency for Science, Technology and Research in Singapore. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.



## Web Resources

370 We provide publicly available software implementing our approach:

<http://bogdan.bioinformatics.ucla.edu/software/>

Figure 1 was generated using the following website with data from the Human Genome

Diversity Panel<sup>43; 44</sup>:

<http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/>

## References

1. Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2011). Unlocking Mendelian disease using exome sequencing. *Genome Biol* 12, 228.
2. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12, 745-755.
- 380 3. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42, 30-35.
4. Ku, C.S., Naidoo, N., and Pawitan, Y. (2011). Revisiting Mendelian disorders through exome sequencing. *Hum Genet* 129, 351-370.
5. Ng, P.C., and Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* 7, 61-80.
6. Nishino, J., and Mano, S. (2013). The Number of Candidate Variants in Exome Sequencing for Mendelian Disease under No Genetic Heterogeneity. *Comput*
- 390 *Math Methods Med* 2013, 179761.
7. Robinson, P., Kohler, S., Oellrich, A., Wang, K., Mungall, C., Lewis, S.E., Washington, N., Bauer, S., Seelow, D.S., Krawitz, P., et al. (2013). Improved exome prioritization of disease genes through cross species phenotype comparison. *Genome Res*.
8. Gonzalez-Perez, A., and Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, *Condel*. *Am J Hum Genet* 88, 440-449.
9. Lopes, M.C., Joyce, C., Ritchie, G.R., John, S.L., Cunningham, F., Asimit, J., and Zeggini, E. (2012). A combined functional annotation score for non-synonymous
- 400 variants. *Hum Hered* 73, 47-51.
10. Li, M.X., Kwan, J.S., Bao, S.Y., Yang, W., Ho, S.L., Song, Y.Q., and Sham, P.C. (2013). Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet* 9, e1003143.
11. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46, 310-315.

12. Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 42, 790-793.
- 410
13. Sobreira, N.L., Cirulli, E.T., Avramopoulos, D., Wohler, E., Oswald, G.L., Stevens, E.L., Ge, D., Shianna, K.V., Smith, J.P., Maia, J.M., et al. (2010). Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet* 6, e1000991.
14. Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* 369, 1502-1511.
15. Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328, 636-639.
- 420
16. Bilguvar, K., Ozturk, A.K., Louvi, A., Kwan, K.Y., Choi, M., Tatli, B., Yalnizoglu, D., Tuysuz, B., Caglayan, A.O., Gokben, S., et al. (2010). Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* 467, 207-210.
17. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335, 823-828.
- 430
18. MacArthur, D.G., and Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. *Hum Mol Genet* 19, R125-130.
19. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature* 508, 469-476.
20. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: <http://evs.gs.washington.edu/EVS/>) (Accessed August 2013).
21. Moore, C.B., Wallace, J.R., Wolfe, D.J., Frase, A.T., Pendergrass, S.A., Weiss, K.M., and Ritchie, M.D. (2013). Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. *PLoS Genet* 9, e1003959.
- 440
22. Wang, X., Zhu, X., Qin, H., Cooper, R.S., Ewens, W.J., Li, C., and Li, M. (2011). Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics* 27, 670-677.
23. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493, 216-220.
24. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65.
- 450
25. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., and Bustamante, C.D. (2011). Demographic history and rare allele

- sharing among human populations. *Proc Natl Acad Sci U S A* 108, 11983-11988.
26. Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337, 100-104.
- 460 27. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64-69.
28. (2003). The International HapMap Project. *Nature* 426, 789-796.
29. Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 106, 19096-19101.
30. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28, 470 1359-1367.
31. Churchhouse, C., and Marchini, J. (2013). Multiway admixture deconvolution using phased or unphased ancestral panels. *Genet Epidemiol* 37, 1-12.
32. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am J Hum Genet* 93, 278-288.
33. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 5, e1000519.
- 480 34. Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. *Hum Mutat* 34, E2393-2402.
35. Kimura, M. (1979). The neutral theory of molecular evolution. *Sci Am* 241, 98-100, 102, 108 passim.
36. Kimura, M. (1983). *The neutral theory of molecular evolution.* (Cambridge Cambridgehire ; New York: Cambridge University Press).
37. Polanski, A., and Kimmel, M. (2003). New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* 165, 427-436.
- 490 38. Marth, G.T., Czabarka, E., Murvai, J., and Sherry, S.T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166, 351-372.
39. Boyko, A.R., Williamson, S.H., Indap, A.R., Degenhardt, J.D., Hernandez, R.D., Lohmueller, K.E., Adams, M.D., Schmidt, S., Sninsky, J.J., Sunyaev, S.R., et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 4, e1000083.

40. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* Chapter 7, Unit7 20.
- 500 41. Casals, F., Hodgkinson, A., Hussin, J., Idaghdour, Y., Bruat, V., de Maillard, T., Grenier, J.C., Gbeha, E., Hamdan, F.F., Girard, S., et al. (2013). Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet* 9, e1003815.
42. Moreno-Estrada, A., Gignoux, C.R., Fernandez-Lopez, J.C., Zakharia, F., Sikora, M., Contreras, A.V., Acuna-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S., et al. (2014). Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* 344, 1280-1285.
43. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008).  
510 Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100-1104.
44. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. *Science* 298, 2381-2385.

		100 Reference Individuals			2500 Reference Individuals		
Max True Frequency	Method	Threshold	Number of Variants for Follow-up	Probability of Filtering True Causal	Threshold	Number of Variants for Follow-up	Probability of Filtering True Causal
1.00%	( $f > 1\%$ )	1.0%	298.0	25.4%	1.0%	310.9	2.2%
	FNR<5%	6.5%	724.1	4.6%	0.9%	298.3	4.8%
0.10%	( $f > 1\%$ )	1.0%	298.0	12.1%	1.0%	310.9	0.0%
	FNR<5%	1.5%	356.0	4.3%	0.1%	149.8	3.6%
0.05%	( $f > 1\%$ )	1.0%	298.0	8.0%	1.0%	310.9	0.0%
	FNR<5%	1.5%	356.0	1.9%	0.1%	141.2	2.3%

520 Table 1. Method comparisons for different reference panel sizes and maximum causal allele frequencies. We compare two methods. The first is a method ( $f > 1\%$ ) that filters out any variants at an observed frequency  $>1\%$  ignoring the statistical noise on the frequency estimates (and thus the FNR). The second is a method (FNR<5%) that filters out variants if observed above a threshold frequency guaranteed to provide less than a 5% chance of filtering out the true causal variant. At small reference panel sizes it is critical to incorporate statistical noise from the reference panel to not over-filter the true causal variants. Conversely, with large reference panels, a hard 1% frequency filter is too conservative and significantly increases the number of variants remaining for follow-up analysis.

530

	Number of variants for follow-up			Number of variants for follow-up, with KGGSeq variants		
	Case-Only	Trio-Dom	Trio-Rec	Case-Only	Trio-Dom	Trio-Rec
<i>NoAncestry, <math>f &gt; 1\%</math></i>	678.9	334.7	5.3	409.5	202.4	2.9
<i>NoAncestry, FNR &lt; 5%</i>	695.7	342.7	5.6	418.2	206.5	3.1
<i>PopMatched, FNR &lt; 5%</i>	581.9	288.2	3.1	350.6	173.8	1.7
<i>AllPop</i>	415.9	207.8	1.6	267.3	133.3	<del>5.40</del>

Table 2. Average number of variants that remain for follow-up post-filtering in simulations of non-admixed individuals. All FNR approaches assume the maximal causal variant frequency of 1%. Poorly matched reference panels greatly affect the number of variants for follow-up analysis more so than accounting for increased statistical error from smaller reference panels. The top method, *No Ancestry,  $f > 1\%$*  has an increased FNR (6.0%) relative to all the other methods. Case-Only represents filtering with exome data from a single individual and makes no assumptions about disease architecture. Trio-Dom assumes the case individual and both parental exomes are sequenced and that only one parent has the dominant disorder. Trio-Rec assumes there is exome data from the case individual and both parents and that both parents are carriers of one copy of the causal variant but do not have the recessive disorder.

1000 Genomes Population (Number of Individuals)	<i>NoAncestry, FNR &lt;5%, (s.d.)</i>	<i>PopMatched, FNR &lt;5%, (s.d.)</i>	<i>AllPop, (s.d)</i>
ASW* (61)	668.2 (75.9)	591.5 (39.9)	398.8 (40.2)
CEU (85)	298.9 (32.4)	336.2 (36.2)	221.4 (26.7)
CHB (97)	320.7 (31.4)	298.7 (34.2)	234.0 (27.4)
CHS (100)	326.8 (18.2)	295.0 (24.1)	238.1 (15.2)
CLM* (60)	342.0 (39.2)	438.0 (40.0)	242.0 (20.5)
FIN (93)	288.5 (25.1)	270.2 (36.9)	179.3 (21.2)
GBR (89)	286.6 (27.0)	309.8 (33.1)	206.4 (22.5)
JPT (89)	342.1 (22.6)	302.0 (25.9)	237.4 (18.5)
LWK (97)	814.5 (38.3)	507.9 (36.6)	388.7 (29.0)
MXL* (66)	332.9 (25.1)	446.6 (37.2)	232.7 (22.6)
PUR* (55)	353.3 (48.5)	434.3 (44.9)	250.2 (29.0)
TSI (98)	315.6 (24.6)	336.0 (26.7)	238.9 (21.4)
YRI (88)	757.7 (29.9)	482.5 (33.5)	345.4 (20.6)

Table 3. Different levels of genetic diversity across populations induce a variation in the average number of variants remaining for follow-up in an individual. The highest number of variants remaining for follow-up is seen in African populations (YRI and LWK) as well as African-Americans (ASW); this is consistent with these populations have the greatest amount of genetic diversity. \* denotes admixed populations where a local ancestry aware method was utilized (see Methods).

Method	Recessive (#cases=10)	Dominant (#cases=9)	Compound Heterozygous (#cases=1)
<i>NoAncestry, <math>f &gt; 1\%</math></i>	57.7 (34.8)	749.7 (91.0)	604
<i>PopMatched, <math>FNR &lt; 5\%</math></i>	40.1 (32.5)	604.8 (107.1)	426
<i>AllPop</i>	27.3 (20.2)	434.6 (57.1)	352

Table 4. Average number of variants that remain for follow-up post-filtering in real exome studies of 20 individuals with Mendelian disorders. None of the filtering approaches removed the true casual variants from consideration. Across all disorder architectures, we observe a significant decrease in the number of variants that need to be followed up if ancestry is incorporated in the filtering step. Parentheses denote standard deviations.

570 Variants were eliminated from consideration as potentially true causal variants if they are not annotated as damaging (see Methods) and if they are not observed twice if the disorder is assumed to be autosomal recessive or at least once if it is assumed to be dominant (heterozygous) or compound heterozygous.



## Titles and Legends to Figures

Figure 1. Geographic distribution of rs17046386 across the Human Genome Diversity

Panel CEPH data. The minor allele is rare in non-African populations, but not rare in

580 African populations.

Figure 2. Reference panel size impacts the efficacy of filtering in exome sequencing in

European simulations from the EVS data. Figure 2a shows the threshold on the variant

frequency needed to achieve a 5% FNR for various assumptions about the maximum

frequency of the causal variant in the population (from 0.001 to 0.01). Figure 2b displays

the number of variants that remain to be followed up post-filtering at a 5% FNR rate. As

expected with larger reference panel sizes, the estimated frequency from the reference

panel becomes more accurate making the 5% FNR threshold converge to the maximum

assumed frequency of the causal variant ( $f_M$ ) which in turn increases the efficacy of

590 filtering. We observe limited gains in accuracy for reference panels over 500 individuals.

Similar results are obtained for simulations of African Americans (see Supplemental Figure

2).

Figure 3. Estimates of the number of case-only variants for follow-up for the *AllPop*

method for various numbers of additional comparison populations. Starting from the results

of the *PopMatched*,  $FNR < 5\%$  approach (shown at x-axis=0 in the plot) additional

randomly chosen populations are added to the reference panels. Increasing the amount of

information about various populations can further reduce the number of variants for follow-

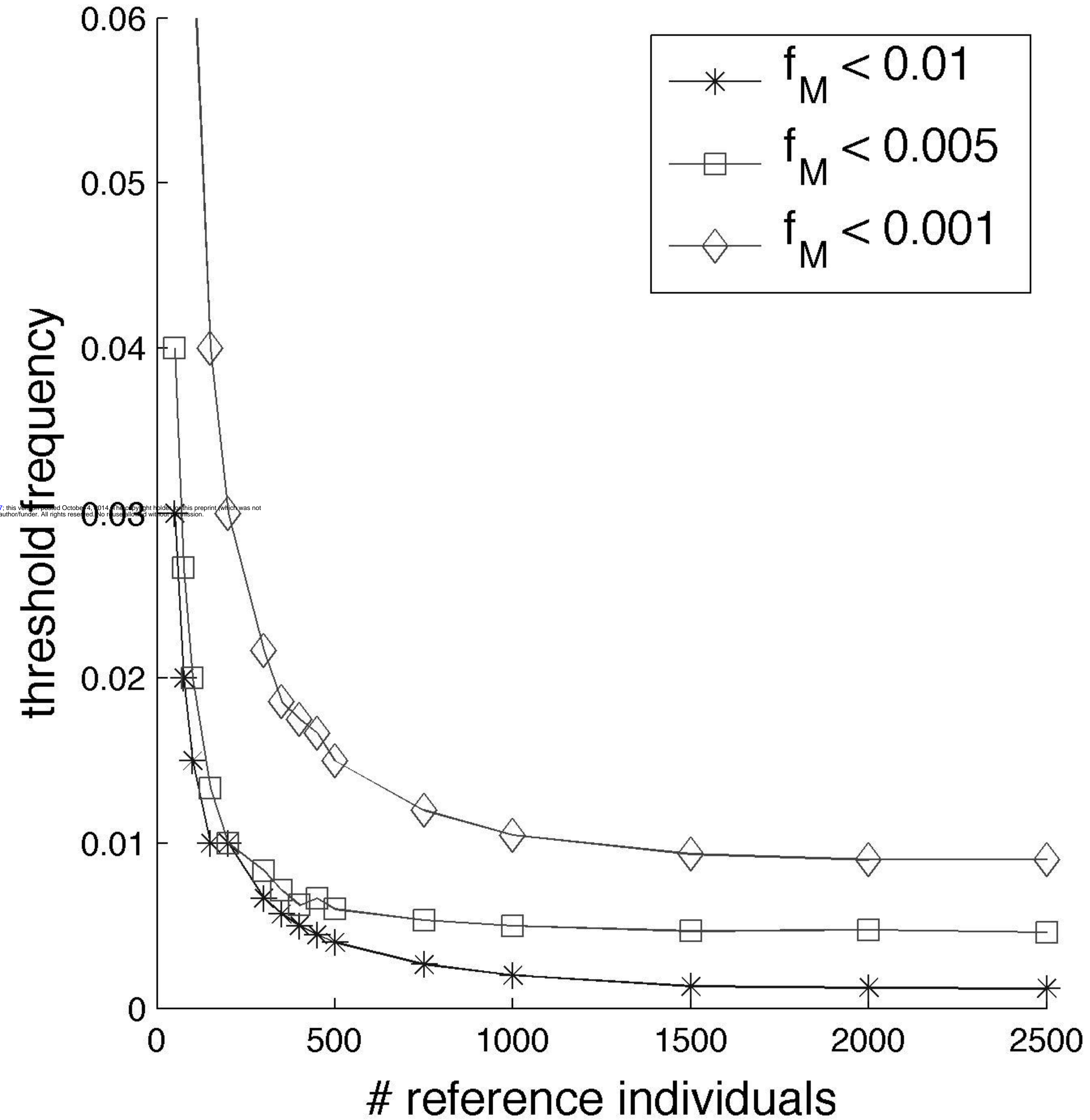
up analysis thus increasing the efficacy of filtering.

600

Figure 4. Population matching using local ancestry information improves performance over local ancestry naïve population matching in admixed populations. The MXL has worse local ancestry aware performance as compared to local ancestry naïve performance likely due to reference panels being poorly matched given the high levels of diversity in Mexicans.



a



b

