Title

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

Permalink

https://escholarship.org/uc/item/0551429t

Journal

European Journal of Human Genetics, 24(1)

ISSN

1018-4813

Authors

Brown, Robert
Lee, Hane
Eskin, Ascia
et al.

Publication Date

2016

DOI

10.1038/ejhg.2015.68

Peer reviewed

European Journal of Human Genetics (2016) 24, 113–119
© 2016 Macmillan Publishers Limited All rights reserved 1018-4813/16
www.nature.com/ejhg

ARTICLE

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

Robert Brown*,[1], Hane Lee[2], Ascia Eskin[3], Gleb Kichaev[1], Kirk E Lohmueller[1,4], Bruno Reversade[5,6,7], Stanley F Nelson[2,3] and Bogdan Pasaniuc*,[1,2,3]

Recent breakthroughs in exome-sequencing technology have made possible the identification of many causal variants of monogenic disorders. Although extremely powerful when closely related individuals (eg, child and parents) are simultaneously sequenced, sequencing of a single case is often unsuccessful due to the large number of variants that need to be followed up for functional validation. Many approaches filter out common variants above a given frequency threshold (eg, 1%), and then prioritize the remaining variants according to their functional, structural and conservation properties. Here we present methods that leverage the genetic structure across different populations to improve filtering performance while accounting for the finite sample size of the reference panels. We show that leveraging genetic structure reduces the number of variants that need to be followed up by 16% in simulations and by up to 38% in empirical data of 20 exomes from individuals with monogenic disorders for which the causal variants are known.

European Journal of Human Genetics (2016) **24,** 113–119; doi:10.1038/ejhg.2015.68; published online 22 April 2015

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.[1–4] Although each individual carries 20 000–24 000 single-nucleotide variants, most are common in the population and are unlikely to explain a rare monogenic trait. Variants that are too common to be consistent with the prevalence of a rare disorder are removed from consideration[2] and the remaining variants are prioritized based on functional, structural and conservation properties.[5–7] Recent prioritization approaches use cross-species comparisons[8] or a combination of scores from several stand-alone methods for increased performance.[9–12] Although such techniques are very powerful when family data are available,[2–4,13–17] hundreds of variants often remain for follow-up validation when only a single case individual is sequenced.[11,18,19]

Variant filtering in exome-sequencing studies is usually performed using frequencies that are estimated across large databases of human variation either by ignoring ancestry, or by matching at the level of continental ancestry (eg, the Exome Variant Server (EVS)[20])[21] thus ignoring sub-continental ancestry. Although F_{ST} values calculated within continental populations are usually low (mostly due to the dependency of F_{ST} on allele frequency),[22,23] detectable population structure still exists.[24,25] Population genetic models predict that rare variants show greater clustering within continental populations than more common variants[26] and empirical studies support this prediction.[27–33] Therefore, a variant might appear rare ($<1\%$) when its frequency is estimated across many populations, when in reality it is only rare in most populations and less rare or even common ($>1\%$) in a one or more clustered sets of populations (see Supplementary Figure 1). For example, variant rs17046386 is common in Africans (therefore unlikely to be pathogenic) and generally rare or absent in non-Africans[34,35] (see Figure 1). However, this variant would not be discarded in the filtering step based on frequency estimates from European reference panels thus increasing the validation burden in the subsequent steps. In addition, the limited size of existing reference panels, especially when defining ancestry at the level of a country (often <100 individuals), induces significant statistical variance in allele frequency estimates that needs to be accounted for (eg, a variant with true frequency of 0.5% has 9.0% probability of being observed with a frequency $>1\%$ in a sample of 100 individuals and thus erroneously discarded).

In this work we investigate the use of sub-continental allele frequencies (typically estimated at the level of a country[36,37]) for a discrete frequency-based filtering step in exome-sequencing studies but such ideas can also be applied to general statistical methodologies aimed at finding causal genes in exome scans. We propose approaches that leverage the frequency estimates across all sub-continental populations in filtering while accounting for the statistical noise introduced by the smaller number of individuals used to estimate frequencies. We use simulations starting from the 1000 Genomes,[28]

[1]Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, CA, USA; [2]Department of Pathology and Laboratory Medicine, Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA; [3]Department of Human Genetics, Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA; [4]Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, CA, USA; [5]Institute of Medical Biology, Human Genetics and Embryology Laboratory, A*STAR, Singapore, Singapore; [6]Institute of Molecular and Cellular Biology, A*STAR, Singapore, Singapore; [7]Department of Pediatrics, National University of Singapore, Singapore, Singapore
*Correspondence: R Brown, Bioinformatics Interdepartmental Program, University of California, Los Angeles, 10833 Le Conte Ave, CHS 33-365, Los Angeles 90095-7088, CA, USA. Tel: +310 825 3069; Fax: +310 825 4846; E-mail: rpb2103@ucla.edu
or Dr B Pasaniuc, Department of Pathology & Laboratory Medicine, University of California, Los Angeles, 10833 Le Conte Ave, CHS 33365, Los Angeles 90095-7088, CA, USA. Tel: +310 825 3291; Fax: +310 825 4846; E-mail: pasaniuc@ucla.edu
Received 31 August 2014; revised 1 March 2015; accepted 10 March 2015; published online 22 April 2015

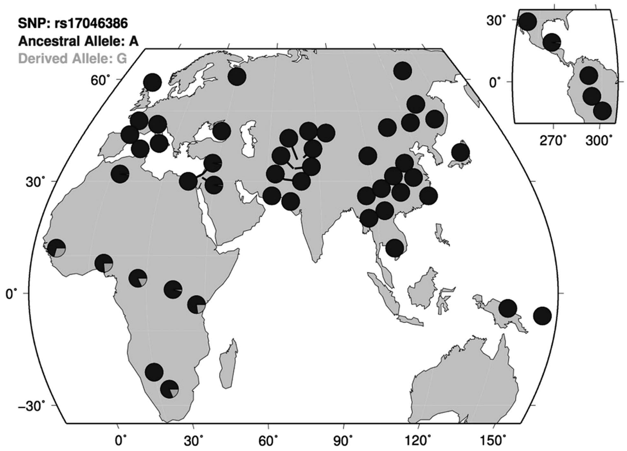

Figure 1 Geographic distribution of rs17046386 across the Human Genome Diversity Panel CEPH data. The minor allele is rare in non-African populations, but not rare in African populations.

the NHLBI GO Exome Sequencing Project Exome Variant Server (EVS)[20] and the ClinVar[38] data to show that our approach improves efficacy of filtering (eg, a reduction of 16% in the number of variants to be followed up in case-only simulations). Importantly, we show that the standard approach that ignores statistical noise in the allele frequency estimation is miscalibrated with respect to the false-negative rate (FNR) (ie, the probability of filtering out a true causal variant). Finally, we validate our approach using exome-sequencing data from 20 real individuals with monogenic disorders for which the true causal variants are known. Here we successfully reduce the number of variants to be functionally tested (a 38% reduction from 750 to 468 in the heterozygous case), while never discarding the known causal variants. Our results show that existing filtering pipelines for studies of monogenic traits can be significantly improved by incorporating ancestry while accounting for statistical noise in the filtering step. Interestingly, utilizing sub-continental population reference panels overcomes the reduction in performance due to higher statistical noise from the smaller panels.

MATERIALS AND METHODS

Data sets

The 1000 Genomes Project[28] has produced a public catalog of human genetic variation through sequencing from several populations: Han Chinese in Beijing (CHB), Japanese in Tokyo (JPT), Southern Han Chinese (CHS), Utah Residents with Northern and Western European ancestry (CEU), Toscani in Italia (TSI), Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian population in Spain (IBS), Yoruba in Ibadan (YRI), Luhya in Webuye (LWK), Americans of African Ancestry in SW USA (ASW), Mexican ancestry from Los Angeles (MXL), Puerto Ricans from Puerto Rico (PUR) and Colombians from Medellin (CLM). We use the 1000 Genomes data (with the exception of IBS individuals, only 14 in total) to evaluate the effectiveness of various filtering approaches. As the vast majority of causal variants for monogenic traits are located in the exome,[39] we restrict our analysis to coding regions of autosomal chromosomes. For admixed individuals we downloaded and used the 1000 Genomes Project local ancestry calls (the consensus calls from four inference methods[40–43]). Damaging scores for each single-nucleotide variant were estimated using KGGSeq with default parameters[11] that combines the functional scores from dbNSFP[44] v2.0.

The EVS has released allele counts from 4300 European-Americans and 2203 African Americans[20] along with PolyPhen2 scores for missense variants and we used those in our analyses. A set of 1395 pathogenic variants (as reported by multiple submitters) was obtained from the ClinVar database[38] (accessed 4 December 2014).

To compare 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 population for estimating best matching allele frequencies. Of the 101 individuals, 9 were known to harbor heterozygous variants in genes causing autosomal dominant disorders, 10 had homozygous variants and 1 had compound heterozygous variants in a gene causing an autosomal recessive disorder (see Supplementary Information and Supplementary Tables S1 and S2).

FNR estimation

We estimate the probability of filtering out a true causal variant (FNR) at a given frequency threshold as a function of a reference panel and the maximum true allele frequency of the causal variant. The filtering threshold can be adjusted 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:

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

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 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[29,45–49] 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:

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

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. 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 (2*N*). As the integration is over a discrete space we calculate the probability that the number of success is greater than the threshold times 2*N*. We propose to filter variants using the minimum frequency threshold *t* such that FNR(*t*) < 0.05. If multiple populations are used in filtering (ie, removing variants that are common in any population, see below), we employ a Bonferroni correction for the threshold; that is, we require FNR(t) < 0.05/K in each of the *K* considered populations.

Leveraging population structure for improved filtering

We simulate individuals with monogenic disorder by drawing two individuals from a specific 1000 Genomes population and then simulating an offspring assuming Mendelian inheritance and independence between SNPs. We compare three possible disease scenarios (Case-Only, Trio-Dominant and Trio-Recessive) using 40 simulated individuals per scenario and population. The Case-Only scenario assumes there is no information on parental genotypes. 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 has two copies of the causal allele. Prior to frequency filtering, we remove all variants that do not result in an amino-acid change or do not create or remove a stop codon. In addition, we remove variants inconsistent with the disease scenario.

We consider multiple frequency-filtering approaches. The NoAncestry, *f* > 1% and NoAncestry, FNR < 5% approaches estimate allele frequencies and FNRs across all 1000 Genomes individuals. The key intuition here is that statistical noise is decreased with large reference panels but at the cost of ignoring population structure. NoAncestry, *f* > 1% filters out variants with allele frequency >1% without regard for the FNR; NoAncestry, FNR<5% filters out variants above a threshold determined to ensure a 5% FNR. The PopMatched, FNR<5% approach uses only the reference individuals from the sub-continental population (country-level, see 1000 Genomes[28]) of the simulated individual. The AllPop, FNR <5% approach filters out variants observed in any population above a conservative Bonferroni-corrected FNR <5% threshold. We assume that populations are independent and set the desired FNR for each population to 0.05 divided by the number of populations used in filtering (eg, 14 for simulation results). MaxPopFreq filters variants if observed above 1% allele frequency in any 1000 Genomes continental population and is similar to a strategy implemented by ANNOVAR[50] that filters variants if observed above 1% in any 1000 Genomes continental population or the EVS European or African-American populations.

For admixed populations we only simulated the Case-Only scenario by using the genotypes of real admixed individuals from 1000 Genomes as case individuals. In addition to methods above, we considered a method that utilizes local ancestry calls (PopMatched-LA, FNR<5%). In each individual at loci that are homozygous for African, European or Native American ancestry, we used the corresponding continental allele frequency estimates obtained by averaging across all 1000 Genomes individuals from a given continent. In local ancestry heterozygous regions we used a 50–50 weighting of the matching continental frequencies. We use the maximum continental FNR-based frequency threshold from the African, European and Asian continents as the filtering threshold.

RESULTS

Modeling statistical uncertainty increases filtering efficacy

We use simulations from the European-American EVS[20] data set to assess filtering based on an FNR as compared with the standard approach of ignoring statistical noise in the allele frequency estimates. We use simulations of various reference panel sizes created with binomial sampling from the frequencies estimated across all the European (or African-American) EVS data. As expected, the frequency threshold that maintains a 5% FNR increases as reference panel size decreases (see Figure 2a). As the maximum frequency of the true causal variant (max(f_c)) decreases the number of variants for follow-up per individual also decreases thus increasing filtering performance (see Figure 2b). Overall, we find a diminishing return in performance for reference panels larger than 500 individuals.

Next, we investigated the FNR attained by the standard approach that ignores statistical noise and filters based on the mean frequency estimate. At small reference panels the standard approach is miscalibrated attaining an FNR close to 25% thus removing causal variants from consideration (see Table 1). In contrast, the approach that maintains an FNR<5% significantly increases the number of variants for follow-up from 298 to 724 on average; this is necessary as it reduces the FNR to the desired 5% (see Table 1). With large reference panels the frequency-based approach is conservative (FNR ~0%) yielding twice as many variants for follow-up than the FNR-based approach if the maximum causal frequency is 0.1%. Qualitatively similar results were observed for simulations from the EVS African-American data (see Supplementary Figure 2).

Leveraging ancestry to increase filtering performance

Next, we assessed the performance of filtering with or without accounting for the highly structured nature of rare variants.[27–29,51] Using simulated exome data we investigated the efficacy of filtering across a wide range of methods and sequencing studies. When comparing the methods that do not leverage ancestry, 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 increased number of variants is necessary to attain a correct 5% FNR rate (NoAncestry, $f>1\%$ attains an FNR of 6%). The MaxPopFreq approach yields the fewest number of variants by filtering at 1% frequency in any continental population, but has a 50% probability of filtering out the true causal variant (FNR=50%) (see Table 2).

Among all methods that maintain an FNR <5% and use ancestry information, we observe that the method that incorporates data across all populations (AllPop, FNR<5%) attains the best performance across all simulated scenarios (an average 16% reduction across all populations from the NoAncestry, FNR<5% method in the Case-Only scenario, see Table 2). The improvement is likely due to variants common in at least one population that are filtered out as unlikely to be pathogenic. This benefit from assaying variants across many populations comes even at the expense of multiple testing correction (a Bonferroni adjustment is made to the FNR required in each individual population resulting in an average filtering threshold of 3.2%). This demonstrates that the benefit of better population matching outweighs the cost of higher statistical noise from the small reference panels. The greatest improvement from population matching comes with the African populations where there is a 26% decrease in the number of variants remaining for follow-up (see Table 3).

We also investigated other types of clinical scenarios. As expected, the Trio-Dominant scenario has approximately half as many variants for follow-up as the Case-Only scenario (see Table 2). The Trio-Recessive scenario, simulated without inbreeding, shows less than six variants remaining for all scenarios and methods (see Table 2). Finally, we observe a similar pattern of improved performance when also filtering non-damaging variants as predicted by KGGSeq (Table 2). Therefore, improvements of ancestry-aware filtering do not come preferentially from variants with non-damaging predictions.

Ancestry-aware filtering in admixed individuals

We extend our approach to admixed individuals (eg, African Americans) with genetic ancestry from multiple continents. We incorporate the local ancestry structure in the filtering step with the PopMatched-LA, FNR <5% approach that matches reference panels according to the ancestry at each site in an individual's genome. This significantly lowers the number of variants for follow-up in the admixed populations as compared with the standard local ancestry naive method (PopMatched, FNR <5%) (see Figure 3). For example, in African-American individuals we observe a reduction from 664 to 487 variants from just matching the local ancestry to continental populations as compared with using all 1000 Genomes data with an FNR <5%. When using information from all populations in the 1000 Genomes data set, there is improvement for all admixed populations over the method that ignores ancestry (NoAncestry, FNR<5%) (see Figure 3).

Ancestry-aware filtering in ClinVar data

In the simulations above, we have made the assumption that the frequency distribution of causal variants is well approximated using the rare variants in our data, which may not hold in practice. To investigate deviations from this assumption, we filter the set of ClinVar pathogenic SNPs according to our methods. We find that the AllPop, FNR<5%, PopMatched, FNR<5% and NoAncestry, $f>1\%$ approaches filtered out 42 (3.0% FNR), 18 (1.3% FNR) and 38 (2.7% FNR) of the 1395 variants, respectively. This shows that all approaches are conservative with respect to FNR and suggests that by approximating the distribution of frequencies at causal variants from real data we do not artificially increase the FNR in empirical data.

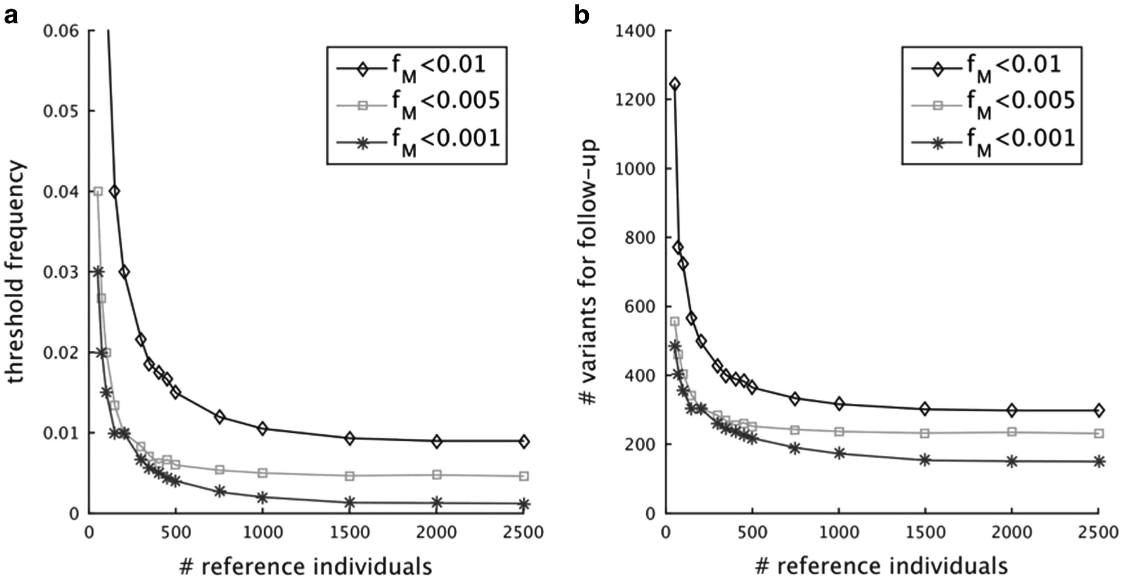

Figure 2 Reference panel size impacts the efficacy of filtering in exome sequencing in European simulations from the EVS data. We simulated reference panels at various sizes using a binomial sampling from the EVS frequencies. (**a**) Shows the threshold on the variant frequency needed to achieve a 5% FNR for various assumptions about the maximum frequency of the causal variant in the population (from 0.001 to 0.01). (**b**) 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 frequencies from the reference panel become 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 filtering. We observe limited gains in accuracy for reference panels over 500 individuals. Similar results are obtained for simulations of African Americans (see Supplementary Figure 2).

Table 1 Method comparisons for different reference panel sizes and maximum causal allele frequencies

		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%

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.

Table 2 Average number of variants that remain for follow-up post-filtering in simulations of non-admixed individuals

	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*
NoAncestry, *f*>1%	679.3	330.5	5.2	410.5	200.4	2.8
NoAncestry, FNR<5%	702.1	346.4	5.9	422.5	208.8	3.1
MaxPopFreq	358.3	176.6	1.0	235.4	115.9	1.0
PopMatched, FNR<5%	675.4	332.2	4.2	400.4	196.8	2.2
AllPop, FNR <5%	570.1	279.2	3.1	353.7	173.6	1.7

All FNR approaches assume the maximal causal variant frequency of 1%. *NoAncestry, f>1%* and *MaxPopFreq* have increased FNRs of 6% and 50% respectively. The *AllPop, FNR<5%* approach outperforms all other FNR-based approaches. The *PopMatched, FNR<5%* approach is the second best performing FNR-based approach demonstrating that the improvements from better population matching outweigh the effects of increased statistical noise from smaller reference panels.

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

1000 genomes population (number of individuals)	*NoAncestry, FNR <5%, (SD)*	*PopMatched, FNR <5%, (SD)*	*AllPop, FNR<5%, (SD)*
ASW[a] (61)	645 (78)	487 (33)	514 (53)
CEU (85)	311 (38)	393 (43)	302 (40)
CHB (97)	321 (33)	323 (35)	282 (32)
CHS (100)	322 (16)	317 (17)	282 (16)
CLM[a] (60)	335 (44)	377 (32)	309 (29)
FIN (93)	289 (19)	312 (28)	264 (18)
GBR (89)	293 (29)	355 (40)	286 (30)
JPT (89)	344 (25)	341 (34)	295 (26)
LWK (97)	833 (32)	610 (31)	605 (29)
MXL[a] (66)	312 (27)	392 (34)	308 (24)
PUR[a] (55)	353 (52)	369 (37)	321 (39)
TSI (98)	326 (27)	386 (31)	321 (27)
YRI (88)	765 (26)	566 (23)	547 (23)

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. These populations also show the greatest benefit from better population matching and from applying the *AllPop, FNR<5%* approach.
[a]Denotes admixed populations where results from the *PopMatched-LA, FNR<5%* approach are reported.

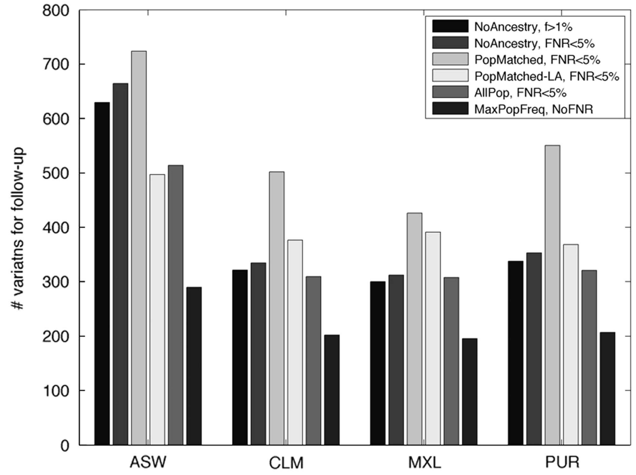

Figure 3 Population matching using local ancestry information improves performance over local ancestry naive population matching in admixed populations. The *PopMatched, FNR<5%* approach performs poorly because the admixed reference panel sizes are much smaller than non-admixed reference panels leading to increased filtering thresholds. The *AllPop, FNR<5%* outperforms all other FNR-based approaches.

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 of the 101 real exome sequenced individuals who had monogenic disorders where the causal variants have been previously identified. We assumed a maximum causal allele frequency of 1% for all cases because there was no prevalence data.[2] 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, FNR <5% approaches as opposed to the NoAncestry, $f>1\%$ approach that does not account for the FNR (See Supplementary Table 1). 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. The 101 individuals form a supplemental population with the test individual held out. We included these individuals when estimating average frequencies across all populations in the 1000 Genomes Project for the real data. Using our AllPop, FNR <5% approach only 468 variants need to be followed up for dominant disorders as compared with 750 for the NoAncestry, $f>1\%$ approach (see Table 4). The true causal variant identified in these individuals was never filtered out. This demonstrates that using multiple population frequency estimates significantly reduces the number of variants remaining for follow-up analysis, while still maintaining an appropriate FNR. In Supplementary Table 1 we report the variants remaining in each individual along with country data, presumed inheritance pattern, the zygosity of the causal variant and disease. Removing outliers based on PCA from the 101 self-reported Middle Eastern individuals makes no significant difference in the number of variants remaining for follow-up (see Supplementary Figure S3 and Supplementary Table S3).

DISCUSSION

In this work, we introduce approaches that account for the finite sample size of the existing reference panels used in filtering while leveraging sub-continental ancestry to improve the filtering step in monogenic disease studies. 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 efficacy.

Our methods are limited in that they do not account for the cases where a second genetic or environmental factor is required for the phenotype to appear and this increases the risk of filtering the true causal if the second factor is rare in some populations. Errors in variant calling in reference populations may also falsely elevate the

Table 4 Average number of variants that remain for follow-up post-filtering in real exome studies of 20 individuals with monogenic disorders

Method	*Recessive (#cases= 10)*	*Dominant (#cases= 9)*	*Compound heterozygous (#cases= 1)*
NoAncestry, $f>1\%$	57.7 (34.8)	749.7 (91.0)	604
PopMatched, FNR<5%	40.1 (32.5)	604.8 (107.1)	426
AllPop, FNR <5%	29.2 (21.4)	467.7 (61.5)	370

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 SDs. 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.

frequency of a true causal variant and so using multiple technologies for estimating allele frequencies would be a best practice. While our work is presented for use with exome-sequencing studies, its central idea will be extendable to whole genome sequencing, as rare variation both in and out of the exome will show population clustering.

The current bottleneck in using population structure to help identify rare variants is the limited size of the reference panels for narrowly defined sub-continental populations. Large databases such as the EVS could increase their impact if they could report sub-continental allele frequencies in addition to just European and African-American allele frequencies. Recent projects such as the UK10K[52] will be extremely valuable as it is a large reference panel of a specific population. The ALFRED database[53] will also be a very valuable resource for cross-population work with monogenic diseases when it can provide sequencing level data. Founder populations (eg, Amish or Iceland) where some non-causal variants are pulled to high frequency may be powerful in eliminating non-causal variants if the disease is rare or not present in the founder population.[54,55] Tools such as Kaviar[56] will allow researchers to quickly search these emerging sources of population frequency data. Finally, a Bayesian approach to integrate cross-population prevalence, allele frequencies, annotation and functional data in a filter-free probabilistic manner is possible but left to explore in future work.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

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 Genomics/Informatics Core of the UCLA MUSCULAR DYSTROPHY CORE CENTER from NIAMS (P30AR057230). BR 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: We provide publicly available software implementing our approach: http://bogdan.bioinformatics.ucla.edu/software/. Figure 1 was made on this website with Human Genome Diversity Panel data: http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/. 1000 Genomes local ancestry calls are available at: http://www.1000genomes.org/phase1-analysis-results-directory.

1 Gilissen C, Hoischen A, Brunner HG, Veltman JA: Unlocking Mendelian disease using exome sequencing. *Genome Biol* 2011; **12**: 228.
2 Bamshad MJ, Ng SB, Bigham AW *et al*: Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011; **12**: 745–755.
3 Ng SB, Buckingham KJ, Lee C *et al*: Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010; **42**: 30–35.
4 Ku CS, Naidoo N, Pawitan Y: Revisiting Mendelian disorders through exome sequencing. *Hum Genet* 2011; **129**: 351–370.
5 Ng PC, Henikoff S: Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* 2006; **7**: 61–80.
6 Nishino J, Mano S: The number of candidate variants in exome sequencing for mendelian disease under no genetic heterogeneity. *Comput Math Methods Med* 2013; **2013**: 179761.
7 Javed A, Agrawal S, Ng PC: Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat Methods* 2014; **11**: 935–937.
8 Robinson P, Kohler S, Oellrich A *et al*: Improved exome prioritization of disease genes through cross species phenotype comparison. *Genome Res* 2013; **24**: 340–348.
9 Gonzalez-Perez A, Lopez-Bigas N: Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 2011; **88**: 440–449.
10 Lopes MC, Joyce C, Ritchie GR *et al*: A combined functional annotation score for non-synonymous variants. *Hum Hered* 2012; **73**: 47–51.
11 Li MX, Kwan JS, Bao SY *et al*: Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet* 2013; **9**: e1003143.
12 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014; **46**: 310–315.
13 Ng SB, Bigham AW, Buckingham KJ *et al*: Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 2010; **42**: 790–793.
14 Sobreira NL, Cirulli ET, Avramopoulos D *et al*: Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet* 2010; **6**: e1000991.
15 Yang Y, Muzny DM, Reid JG *et al*: Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* 2013; **369**: 1502–1511.
16 Roach JC, Glusman G, Smit AF *et al*: Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 2010; **328**: 636–639.
17 Bilguvar K, Ozturk AK, Louvi A *et al*: Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* 2010; **467**: 207–210.
18 MacArthur DG, Balasubramanian S, Frankish A *et al*: A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 2012; **335**: 823–828.
19 MacArthur DG, Tyler-Smith C: Loss-of-function variants in the genomes of healthy humans. *Hum Mol Genet* 2010; **19**: R125–R130.
20 Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, USA. URL http://evs.gs.washington.edu/EVS/. (accessed August 2013).
21 MacArthur DG, Manolio TA, Dimmock DP *et al*: Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014; **508**: 469–476.
22 Jakobsson M, Edge MD, Rosenberg NA: The relationship between F(ST) and the frequency of the most frequent allele. *Genetics* 2013; **193**: 515–528.
23 Bhatia G, Patterson N, Sankararaman S, Price AL: Estimating and interpreting FST: the impact of rare variants. *Genome Res* 2013; **23**: 1514–1521.
24 Novembre J, Johnson T, Bryc K *et al*: Genes mirror geography within Europe. *Nature* 2008; **456**: 98–101.
25 Yang WY, Novembre J, Eskin E, Halperin E: A model-based approach for analysis of spatial structure in genetic data. *Nat Genet* 2012; **44**: 725–731.
26 Mathieson I, McVean G: Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 2012; **44**: 243–246.
27 Fu W, O'Connor TD, Jun G *et al*: Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 2013; **493**: 216–220.
28 Abecasis GR, Auton A, Brooks LD *et al*: An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.
29 Gravel S, Henn BM, Gutenkunst RN *et al*: Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA* 2011; **108**: 11983–11988.
30 Tennessen JA, Bigham AW, O'Connor TD *et al*: Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 2012; **337**: 64–69.
31 Myles S, Davison D, Barrett J, Stoneking M, Timpson N: Worldwide population differentiation at disease-associated SNPs. *BMC Med Genomics* 2008; **1**: 22.
32 Zawistowski M, Reppell M, Wegmann D *et al*: Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. *Eur J Hum Genet* 2014; **22**: 1137–1144.
33 Nelson MR, Wegmann D, Ehm MG *et al*: An abundance of rare functional variants in 202 drug target genes sequenced in 14 002 people. *Science* 2012; **337**: 100–104.
34 Li JZ, Absher DM, Tang H *et al*: Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008; **319**: 1100–1104.
35 Rosenberg NA, Pritchard JK, Weber JL *et al*: Genetic structure of human populations. *Science* 2002; **298**: 2381–2385.
36 Moore CB, Wallace JR, Wolfe DJ *et al*: Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. *PLoS Genet* 2013; **9**: e1003959.
37 Wang X, Zhu X, Qin H *et al*: Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics* 2011; **27**: 670–677.
38 Landrum MJ, Lee JM, Riley GR *et al*: ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014; **42**: D980–D985.
39 Choi M, Scholl UI, Ji W *et al*: Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA* 2009; **106**: 19096–19101.
40 Baran Y, Pasaniuc B, Sankararaman S *et al*: Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 2012; **28**: 1359–1367.
41 Churchhouse C, Marchini J: Multiway admixture deconvolution using phased or unphased ancestral panels. *Genet Epidemiol* 2013; **37**: 1–12.
42 Maples BK, Gravel S, Kenny EE, Bustamante CD: RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* 2013; **93**: 278–288.
43 Price AL, Tandon A, Patterson N *et al*: Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 2009; **5**: e1000519.
44 Liu X, Jian X, Boerwinkle E: dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat* 2013; **34**: E2393–E2402.
45 Kimura M: The neutral theory of molecular evolution. *Sci Am* 1979; **241**: 98–100, 102, 108 passim.
46 Kimura M: *The Neutral Theory Of Molecular Evolution*. Cambridge Cambridgeshire; New York: Cambridge University Press, 1983.
47 Polanski A, Kimmel M: New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* 2003; **165**: 427–436.
48 Marth GT, Czabarka E, Murvai J, Sherry ST: The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 2004; **166**: 351–372.

49 Boyko AR, Williamson SH, Indap AR *et al*: Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 2008; **4**: e1000083.
50 Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010; **38**: e164.
51 Casals F, Hodgkinson A, Hussin J *et al*: Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet* 2013; **9**: e1003815.
52 Muddyman D, Smee C, Griffin H, Kaye J: Implementing a successful data-management framework: the UK10K managed access model. *Genome Med* 2013; **5**: 100.
53 Cheung KH, Miller PL, Kidd JR, Kidd KK, Osier MV, Pakstis AJ: ALFRED: a Web-accessible allele frequency database. *Pac Symp Biocomput* 2000; 639–650.
54 Shifman S, Darvasi A: The value of isolated populations. *Nat Genet* 2001; **28**: 309–310.
55 Lim ET, Wurtz P, Havulinna AS *et al*: Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* 2014; **10**: e1004494.
56 Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC: Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* 2011; **27**: 3216–3217.

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