Bioinformatics, 31(22), 2015, 3577–3583
doi: 10.1093/bioinformatics/btv457
Advance Access Publication Date: 6 August 2015
Original Paper

Genome analysis

Strategies to improve the performance of rare variant association studies by optimizing the selection of controls

Na Zhu[1], Verena Heinrich[1], Thorsten Dickhaus[2], Jochen Hecht[3], Peter N. Robinson[1], Stefan Mundlos[1,4], Tom Kamphans[5] and Peter M. Krawitz[1,4,*]

[1]Institute of Medical Genetics and Human Genetics, Charité Universitätsmedizin Berlin, 13353 Berlin, Germany, [2]Institute for Statistics, University of Bremen, 28344 Bremen, Germany, [3]Berlin-Brandenburg Center for Regenerative Therapies (BCRT), 13353 Berlin, Germany, [4]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany and [5]GeneTalk, 13189 Berlin, Germany

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on May 25, 2015; revised on July 30, 2015; accepted on July 30, 2015

Abstract

Motivation: When analyzing a case group of patients with ultra-rare disorders the ethnicities are often diverse and the data quality might vary. The population substructure in the case group as well as the heterogeneous data quality can cause substantial inflation of test statistics and result in spurious associations in case-control studies if not properly adjusted for. Existing techniques to correct for confounding effects were especially developed for common variants and are not applicable to rare variants.

Results: We analyzed strategies to select suitable controls for cases that are based on similarity metrics that vary in their weighting schemes. We simulated different disease entities on real exome data and show that a similarity-based selection scheme can help to reduce false positive associations and to optimize the performance of the statistical tests. Especially when data quality as well as ethnicities vary a lot in the case group, a matching approach that puts more weight on rare variants shows the best performance. We reanalyzed collections of unrelated patients with Kabuki make-up syndrome, Hyperphosphatasia with Mental Retardation syndrome and Catel–Manzke syndrome for which the disease genes were recently described. We show that rare variant association tests are more sensitive and specific in identifying the disease gene than intersection filters and should thus be considered as a favorable approach in analyzing even small patient cohorts.

Availability and implementation: Datasets used in our analysis are available at ftp://ftp.1000genomes.ebi.ac.uk./vol1/ftp/

Contact: peter.krawitz@charite.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In genome wide association studies (GWAS), spurious associations from population substructure are known pitfalls, as statistical tests in case-control studies assume that differences in the compositions of genotypes are solely due to their difference in disease status and not to any difference in genetic background. Especially in large association studies comprising thousands of individuals, the assumption of population homogeneity is often violated. This can increase the familywise error rate (FWER) which is the probability to report at least one false association. An inappropriate genetic matching of cases and controls in the presence of population substructure can lead to inflation, unless properly accounted for in the analysis. Over the recent years several effective methods have been developed to control the inflation of test statistics in association studies. These can be subdivided into approaches that select suitable controls for cases prior to the statistical test such as genotype-based matching (Guan *et al.*, 2009) or a stratification score-based matching (Epstein *et al.*, 2012) and strategies that correct the *P*-values afterwards due to differences in the group substructure such as genomic control (Devlin and Roeder, 1999), principal component analysis (Price *et al.*, 2006) or other variance component models (Devlin and Roeder, 1999; Epstein *et al.*, 2012; Guan *et al.*, 2009; Kang *et al.*, 2010; Price *et al.*, 2006). Most of these methods have been extensively studied in association studies that work with common variants in complex diseases.

However, as the genetic risk modifiers that were identified in many GWAS do explain the disease susceptibility only to a limited degree, it was suggested to extend the search to rare variants to find the missing heritability (Eichler *et al.*, 2010; Zuk *et al.*, 2014).

The statistical tests that have to be applied for rare variants differ substantially from those, which are suitable for common variants. Like in other studies, we will use in hereinafter the abbreviations CVAS and RVAS to make a clear distinction between common and rare variant association studies (Bansal *et al.*, 2010; Zuk *et al.*, 2014). Over the last few years many new statistical approaches have been developed or adapted to analyze associations for collections of rare variants. These methods can be divided into two main types, namely multiple-marker tests and univariate tests that are applied to genotype data that are combined within a region of interest, e.g. a gene (Asimit and Zeggini, 2010).

The issue of appropriate methods for multiple testing corrections has been extensively discussed in CVAS and it took several years to agree on 5×10^{-8} as an acceptable significance threshold for indefinitely dense maps of common single nucleotide polymorphisms (SNPs) (Dudbridge and Gusnanto, 2008). For RVAS, a careful analysis of significance testing has just begun and is an ongoing field of research (Bush and Moore, 2012; Sham and Purcell, 2014). In the meantime, computationally intensive permutation tests can be used to correct the *P*-values empirically and to compare different test statistics (Sham and Purcell, 2014).

GWAS for common variants in complex diseases usually require hundreds of samples for sufficient power. For rare variants that are functionally relevant and thus exhibit higher effect sizes, significant associations might already be detectable for smaller case groups (Asimit and Zeggini, 2010; Zuk *et al.*, 2014). However, the limited number of cases that are available for the analysis of ultra-rare diseases will remain a challenge and in such studies ranking genes by *P*-values may serve as a stopgap solution to identify promising disease loci in the genome.

In this article, we analyze how test statistics perform on exome datasets of individuals from diverse ethnic backgrounds and varying data quality. We focused especially on strategies for selecting controls and studied how this may influence the disease gene discovery. We show that an appropriate choice of controls is of utmost interest when studying individuals with rare disorders, as these collections show usually a substantial population substructure due to diverse ethnicities.

2 Material and Methods

We worked with exome data of healthy, unrelated individuals that had been sequenced with the approval of the ethical Board of the Charité over the recent 5 years in Berlin (referred to as BER cohort) as well as with publicly available exome data of the 1000 genomes project (referred to as 1KGP cohort) (Clarke *et al.*, 2012; Genomes Project Consortium *et al.*, 2010, 2012). All participants or their caregivers gave their written informed consent for their clinical records to be used in this study.

As cryptic relatedness is a known confounder in case-control association studies (Voight and Pritchard, 2005), we first removed related individuals. Then we used PLINK (Prucell *et al.*, 2007) to compute the kinship coefficients of all sample pairs and excluded individuals that showed cryptic relatedness in an iterative manner as described in Turner *et al.* (2011). This procedure yielded 207 unrelated individuals in the BER cohort and 2289 unrelated individuals in the 1KGP cohort (Supplementary Fig. S1).

Similarity metrics: We conducted our rare variant association studies on case-control groups comprising different numbers of individuals N. In our simulations we considered scenarios, where the number of affected individuals in the case group, N_a, is equal to the number of unaffected individuals in the control group, N_u, as well as scenarios with $N_a \leq N_u$. In general, a distance matrix $d_{i,j} = d(x_i,\ x_j)_{i,j=1...N}$ can be computed on any pool of samples, where x_i is a vector of genotype data for individual i. Several similarity functions have been suggested in the statistical genetics literature that put stronger emphasis on different aspects of the underlying data, such as rare or common variants. In this work, we focus on three different functions that differ in the weighting scheme W of genotypes and that we will explain in the following. The distance between two individuals is computed by considering all genomic positions k at which other alleles than the one of the reference sequence have been called:

$$d_{ij} = 1 - \frac{1}{c}\sum_k I_{ij}(k) * W(k)$$

The function $I_{ij}(k)$ indicates whether the genotypes at a certain position k of the genome agree in individuals i and j:

$$I_{ij}(k) = \ I\Big(x_i(k), x_j(k)\Big) = \begin{cases} 1, \ \ if \ \ x_i(k) == x_j(k) \\ 0, \ \ if \ \ x_i(k) \neq x_j(k) \end{cases}$$

The weight $W(k)$ is a function of the genotype frequencies and $c = \sum_k W(k)$ a normalizing constant that assures that $d_{ij} \in [0, 1]$, with $d = 0$ if and only if x_i equals x_j. By changing the weighting function $W(k)$ the distance function will change into different similarity-metrics, which we will discuss in the following. If genotype frequencies are ignored, $W(k) = 1$ for all k, the mere number of agreeing genotypes between two samples is quantified. In the literature this metric is often referred to as identity by state, *IBS*, that is similar to the hamming distance. A stronger weight on rare genotypes can be achieved by using the inverse of the frequency, f, that was determined in large population genetics studies, $W(k) = 1/f(x_i(k))$. The metric that results from this weighting function, W^1, was shown to be especially suitable for measuring the quality of variant call sets of exome data (Heinrich *et al.*, 2013). A similarity metric that uses stronger weights for shared rare alleles has also been studied by Guan *et al.* (2009) and showed best case-control group matched stratification in association studies that were based on SNP data.

Another weighting scheme and the corresponding metric, W^2, that we used in our current study is based on Shannon's concept of information content. Here, the genotype at a specific position k of the genome is treated as the outcome of a discrete random variable and the information content of this position is defined as the entropy over the probability mass function of the frequencies of all the genotypes l that have been observed for this position, $W^2\ (k) = -\sum_l f(k,l) \log f(k,l)$. For a biallelic position of the genome at which the genotypes are in Hardy–Weinberg equilibrium, the weight is maximal for an allele frequency of 0.5 and thus a common variant contributes in average more to the distance than a rare variant. A visualization of the similarity between all analyzed samples is shown in Supplementary Figure S2 for all three weighting schemes.

All three similarity-metrics (*IBS*, W^1 and W^2) that we studied serve two purposes: They may be used to identify the next nearest neighbors for any sample and can thus be used for matching controls. Second, they may be used to characterize the population substructure in the case- and control-group after the individuals have been assigned. Many approaches to account for the effect of population substructure in CVAS have been described. Since most multivariate procedures require invertibility of kinship matrices, these approaches usually work with matrices that are based on IBS, or versions thereof which are weighted by allelic frequencies such as W^1 and W^2 (Nievergelt *et al.*, 2007; Schaid, 2010).

Simulated case-control-groups: All case-control groups of varying sizes were set up as follows: After choosing randomly N_a individuals from a pool of unrelated individuals (e.g. BER or 1KGP) for the case group, N_u individuals were selected from the remaining pool for the control group either also randomly or most similar to the individuals in the case group. We also studied the effect of increasing sizes of control groups at a fixed case group size, $N_a \leq N_u$, by adding stepwise additional samples from the remaining pool, either randomly or similarity-matched. Once a sample had been selected as the next similar individual to a case, it was removed from the pool.

After selecting the samples for the simulated association study we added pathogenic mutations that had already been reported in the literature to the exomes of the case group (Supplementary Table S2). We studied altogether eight different monogenic disorders, three dominant disorders (Noonan syndrome, Neurofibromatosis type I and Kabuki make-up syndrome) and five recessive disorders (Mabry syndrome, which is also known as Hyperphosphatasia with Mental Retardation syndrome, HPMRS, Cystic Fibrosis, Tay–Sachs syndrome and hereditary non-syndromic hearing loss, Catel–Manzke syndrome). Five of the analyzed disorders were caused by mutations in single genes (*NF1*, *KMT2D*, *CTFR*, *HEXA*, *TGDS*) and the other half was heterogeneous (Supplementary Table S1). For the recessive disorders we added either a pathogenic allele in a homozygous state to the exome or two pathogenic alleles in a compound heterozygous state. For disorders with locus heterogeneity we used data summarizing the prevalence of pathogenic mutations in the known disease genes from GeneReviews and the diagnostic laboratory of the Charité.

Statistical analysis: All the sequence variants of the individuals were annotated on the functional gene level with Jannovar and filtered for non-synonymous single nucleotide variants, SNVs (Jager *et al.*, 2014). As all analyzed disorders are ultra-rare, we removed alleles with a population frequency above 0.001.

We examined the remaining variants by different statistical analysis strategies for rare variants (Asimit and Zeggini, 2010; Bansal *et al.*, 2010). As an example for an univariate marker test we used the Cochran–Armitage Trend Test, CATT (Asimit and Zeggini, 2010). For CATT we collapsed the genotype frequencies of all rare, non-synonymous variants within a gene between cases and controls in a 2 × 3 contingency table. As a multivariate marker test we used CMC that was suggested by Li and Leal (2008).

We used Receiver Operating Characteristic(ROC)curves for power and FWER comparisons at different significance levels. For a comparison of the test statistics we corrected the *P*-values empirically by label permutations and based the figures of the results section on the test statistic that performed best (Sham and Purcell, 2014). As an additional method to evaluate the performance of the RVAS setup we assessed the probability that a disease gene has the highest test statistic compared with all other genes.

3 Results

3.1 Defining rare sequence variants

When conducting a RVAS a suitable definition of rareness is required. However, the rareness of a sequence variant depends crucially on the choice of the reference population. In many NGS studies the frequency of a variant or a genotype is based on how often it was observed in individuals that were sequenced in the 1000 genomes project, 1KGP (Genomes Project Consortium *et al.*, 2010, 2012). Currently, high quality genotype data of 2535 individuals of 26 different populations from all over the world is available from the project site. We studied systematically the effect on the filtering process when genotype data of certain populations were included or excluded in the definition of the allele frequencies. Besides the populations of the 1KGP we also analyzed a cohort of over 200 unrelated, healthy individuals that were subjected to exome sequencing over the recent years at the Charité, University Hospital Berlin, and that we refer to as BER. First, we filtered for SNVs, that were predicted to cause missense or nonsense amino acid substitutions on the protein level in a consensus target region of the exome comprising 28 Mb. All European populations of the 1KGP, as well as the BER cohort showed a comparable distribution of such variant calls. Exemplarily in Figure 1 these distributions are depicted for 88 individuals from Great Britain, GBR, 99 individuals from Finland, FIN,

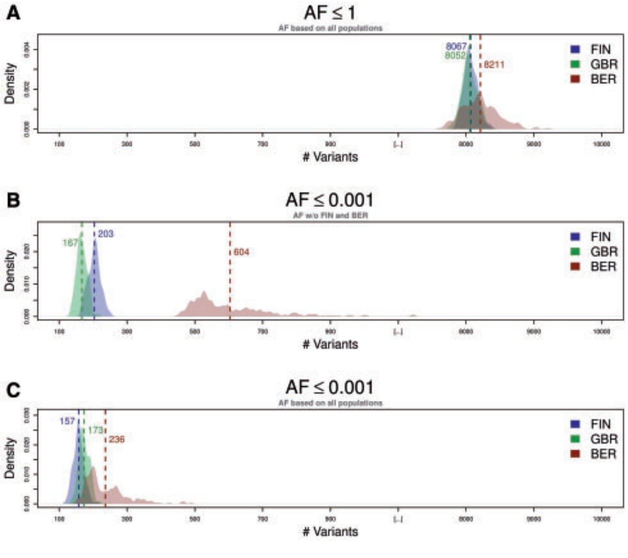

Fig. 1. Distribution of rare variants in individuals from different populations and sequencing studies. (**A**) The number of non-synonymous sequence variants in the consensus coding DNA sequences is comparable for individuals from different population backgrounds. (**B**) However, the definition of the allele frequencies, AF, that are required for computing the subset of rare variants, has a great influence. The profile of rare variants depends on the specific subpopulation as well as on the data quality. If individuals from a subpopulation are filtered for rare variants and the subpopulation is not considered in computing the allele frequencies the number of rare variants is considerably higher (FIN and BER in B). (**C**) A lower mean data quality and more genotyping errors in the BER cohort result in a substantially larger number of singletons explaining the higher mean number of variants after filtering

and more than 200 individuals from the Berlin metropolitan area, BER. Prior to filtering the mean for all three cohorts is around 8000 variant calls (Fig. 1A). The variance for the BER cohort is markedly larger, mainly due to more heterogeneous data quality and higher population substructure than for a single subpopulation in the 1KGP, due to diverse immigration background in Berlin.

In a second filter step we applied an allele frequency cutoff of 0.001, which is commonly used in rare variant association studies. However, depending on the definition of allele frequency there are considerable differences in the number of variants passing this filter. If the entire pool of samples that is used to compute the allele frequencies comprises ~1000 individuals, a cutoff of 0.001 is basically a filter for singletons. That means a variant that is present in two individuals will not pass. When an individual is filtered for rare variants without considering further cases of the same population in the computation of the allele frequency, the mean number of variants after filtering is markedly higher (FIN and BER in Fig. 1B) compared with individuals of GBR where all variants of that population have already been used to compute the allele frequencies. In Figure 1C the distributions for individuals of all three populations are shown when all individuals of the 1KGP as well as of the BER were considered in defining the allele frequencies. In contrast to Figure 1B the expected mean of rare variants passing the filter is now considerably lower for an individual from FIN as all the population-specific variants are now accounted for. Interestingly the expected number of rare variants in a random individual from FIN is even smaller than for a random individual from GBR, as the FIN population is more homogeneous and thus less rare variants remain in the lower frequency range. This is in agreement with the findings of the 1000 genomes consortium that showed a higher proportion of shared alleles and a higher median length of haplotype identity for FIN in comparison to GBR (Genomes Project Consortium *et al.*, 2012).

The proportion of variants passing the frequency filter in an individual of BER is also markedly lower when considering the rare variants of that cohort in the frequency definition. However, lower data quality and consequently more genotyping errors, as well as more diverse population backgrounds result in more singletons that are not filtered out by a 0.001 frequency cutoff.

3.2 Influence of data quality on RVAS performance

As already shown in the previous section, the genotyping quality, the population background and the definition of the allele frequencies affect the rare variant count in an individual. We hypothesized that these factors will also influence the performance of RVAS and therefore studied different compositions of case-control groups systematically, while the definition of allele frequencies that is based on all individuals from the BER and 1KGP cohort was kept fixed. First, we restricted the selection of cases and controls to individuals from the BER cohort. In this sample collection there is substantial population substructure due to migration to Berlin from different countries. In addition, variability in data quality also plays a major role, as exome genotyping quality improved over the recent years substantially. After selecting cases randomly from the BER pool we spiked in pathogenic mutations of a monogenic disorder.

We tested different strategies of matching controls that have already been shown to reduce spurious associations in GWAS on common variants in complex diseases (Guan *et al.*, 2009; Nievergelt *et al.*, 2007). For all different rare monogenic disorders that we studied we found a similar qualitative behavior for the selection strategies. Whenever we discuss the different effects of the matching approaches, we will focus on the heterogeneous recessive disorder HPMRS for the sake of simplicity. Whenever we want to discuss the differences of the single disorders, we will work with the similarity metric W^1, that uses a higher weight for rare alleles.

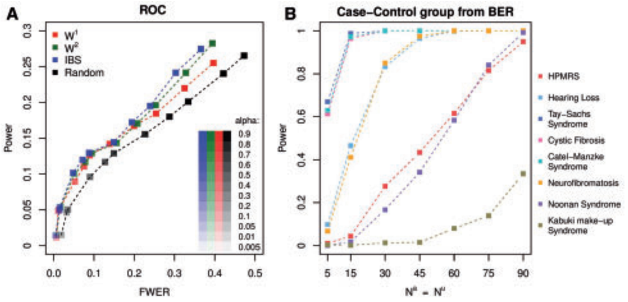

Fig. 2. ROC curves for different selection strategies in RVAS and power for different monogenic disorders. (**A**) Case and control groups of size 5 were simulated by choosing individuals from the BER cohort either randomly or matched by their similarity. Three different metrics were used to infer kinship matrices, a simple identity by state matrix, IBS, and two allele-frequency weighted matrices (W^1: high weight on rare variants, W^2: high weight on common variants). Pathogenic mutations of Mabry syndrome (HPMRS) were spiked into the rare variant sets of individuals of the case group and permutation-based *P*-values from CATT were computed for every gene. The value pairs for power and FWERs were plotted for a range of significance levels. The ROC curves show a lower FWER at comparable power for similarity matched control groups indicating that performance is improved if population substructure is accounted for. (**B**) The detection power increases for growing group sizes. It is more difficult to detect significant associations for dominant disorders that are heterogeneous (Noonan syndrome) or disease genes that are large and highly variable (Kabuki make-up syndrome)

For most simulations the performance was superior if the control group was set up with the closest unrelated matches of the BER cohort. The receiver operating characteristics, ROC, for three different similarity-matching approaches show lower FWER at comparable significance levels and detection power of the disease compared with a simulations where controls were randomly chosen (Fig. 2A).

The power increases for all different diseases that we analyzed when the size of the case-control-groups is growing (Fig. 2B). However, there are substantial differences in the detection power depending on the mode of inheritance, the heterogeneity of the disorder and the variability of the genes. In average it is more likely to detect a significant association for recessive disorders (Tay–Sachs syndrome, Cystic Fibrosis, Catel–Manzke syndrome, HPMRS, autosomal recessive non-syndromic hearing loss), than dominant disorders (Neurofibromatosis, Kabuki make-up syndrome, Noonan syndrome). Furthermore significant associations are more readily detected for disorders that are caused by pathogenic mutations in a single gene (*HEXA* in Tay–Sachs syndrome, *TGDS* in Catel–Manzke syndrome, *CFTR* in cystic fibrosis and *NF1* in Neurofibromatosis) than in heterogeneous disorders such as HPMRS, Noonan syndrome or hearing loss. The power of the RVAS was the lowest for the Kabuki make-up syndrome as the disease-causing gene *KMT2D* is large and highly variable, which also results in many rare variants in the controls.

For almost all rare disorders it is challenging if not impossible to build up case groups that are large enough for significant associations. As an alternative readout for performance we therefore looked at the probability that a true disease gene will yield the lowest *P*-value. We found that the probability to rank the true disease gene at the top increases with growing case-control group sizes for all different population subsets (Fig. 3).

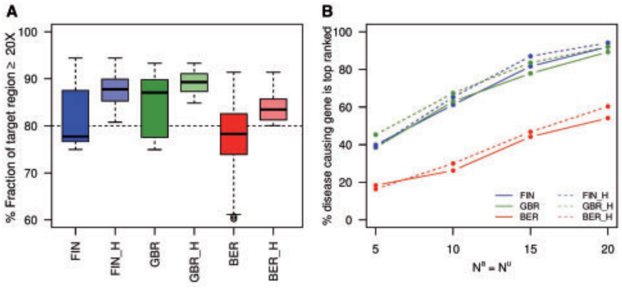

Fig. 3. Influence of population substructure and data quality on disease gene identification. (**A**) The percentage of the target region in a sample that is covered by more than 20 reads is a good indicator for quality. The larger this percentage is, the lower are the expected false positive and false negative error rates for genotyping. BER is the cohort with most low quality datasets. (**B**) The probability to rank the disease-causing gene at the top position increases as the case-control group size increases in all cohorts. The BER subgroup which has the highest population substructure and most low-quality datasets shows the lowest performance. Restricting the analysis to datasets in which more than 80% of the target region are covered by more than 20 reads (*_H) improves the performance of RVAS

We hypothesized that excluding samples of lowest quality could further improve the performance. We thus computed the mean percentage of the exome target region that was covered by more than 20 reads and chose 80% as a cutoff for quality. Limiting the RVAS to a better subset increases the probability to detect the disease causing gene in all cohorts. We assumed that the similarity metrics match the control group not only by ethnical background but also their data quality. By this means we should be able to select high quality controls for a high quality case automatically without detailed information about the coverage.

For BER we could see a major improvement in performance when controls were similarity matched (Fig. 4A and C). The metric that overweighs rare variants shows the most prominent gain in performance. As this metric is especially sensitive to data quality differences as shown in Heinrich *et al.* 2013, it is most effective in BER, which is a cohort with large variability in data quality (Heinrich *et al.*, 2013). To a smaller extent this can also be seen in the 1KGP data that also show differences in coverage and exhibit technology specific error signatures (Moore *et al.*, 2013; Nothnagel *et al.*, 2011).

Interestingly an optimum is not achieved if the maximum available number of controls is selected. Instead the similarity matching strategy will ensure that the last suitable controls will not be selected and spurious associations can thus be reduced.

3.3 Matching controls in ethnically diverse case groups

Case groups for ultra-rare disorders are often extremely small and the individuals are from different ethnicities. In Ehmke *et al.* (2014) for instance three of the individuals affected by Catel–Manzke syndrome originated from Northern Germany, the fourth individual was of British descent and the fifth individual came from Cameroon. While the size of the case group might be small due to the rareness of the disease in Mendelian disorders, there are no such limitations for the control group. In GWAS of common variants it has been shown that expanding the control group size may increase the power (Zhuang *et al.*, 2010). We therefore analyzed the probability to have the disease gene at the first rank, when gradually including additional controls while keeping the case group fixed at size 5 or 15 (Fig. 4). As in CVAS, the disease gene is more likely to be at the top position with a growing number of controls. However, in our RVAS analysis the maximum is not necessarily reached if all remaining individuals from the pool were added as controls. This is especially true if the controls were added based on their similarity to the cases. In the BER cohort we had an optimum in the performance when around 60 most similar controls were added (Fig. 4A and C). The probability of identifying the disease gene was also considerably higher for all similarity-based selection strategies of the controls compared with a random setup of this group. The probability for identifying the disease gene decreases again if controls are included that present an unsuitable match with respect to data quality or ethnic background. The tipping point depends also on the substructure of the available data pool. When we performed the same simulations on a subset of the 1KGP project that has a similar population substructure as the BER cohort we saw a better performance overall due to higher average data quality and a maximum for similarity matched control groups around 120 samples (Fig. 4B).

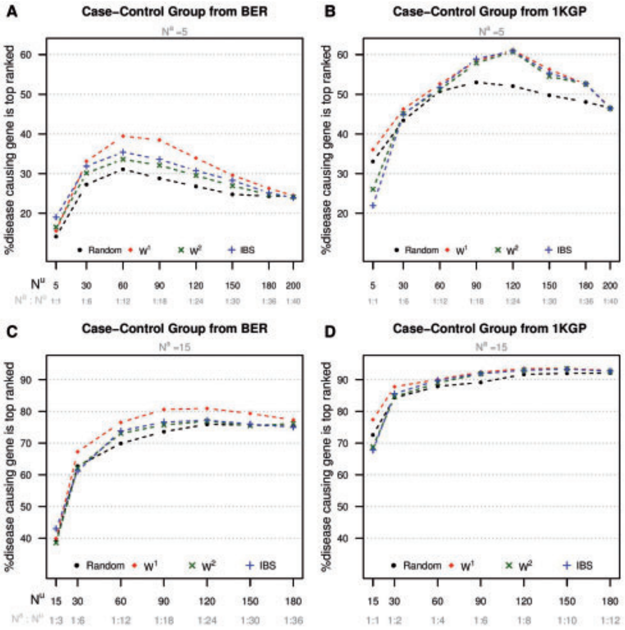

Fig. 4. Performance of rare variant associations with increasing control group size. The simulations were performed on exome data from two different pools the BER cohort (**A** and **C**) and a subgroup of the 1KGP of equal size (**B** and **D**). The subgroup of the 1KGP was also matched to the BER cohort based on W^2 similarity to minimize differences in performance that are due to distinct population substructure between these pools. The size of the case group is kept fixed at 5 (A and B) or 15 (C and D) individuals with pathogenic mutations for HPMRS. For both case group sizes the probability increases that the disease gene has the highest CATT test statistic if additional controls are included. All strategies that select the controls by similarity (W^1, W^2 and *IBS*) outperform a random selection of controls. In the BER cohort, for which the data quality is more heterogeneous than in the 1KGP, the best strategy to select the controls is W^1. This similarity metric is especially sensitive to data quality. The effect of a similarity-matched control group on performance is more prominent for the smaller case group size $N^a = 5$, where an optimum is reached when 60–120 suitable controls are selected

3.4 Disease gene identification in three case studies

As real case scenarios we tested the performance of RVAS in the disease gene identification of three monogenic disorders that were recently resolved, Kabuki make-up syndrome, HPMRS and Catel–Manzke syndrome (Ehmke *et al.*, 2014; Krawitz *et al.*, 2013; Ng *et al.*, 2010). For all three disorders, the disease gene was identified

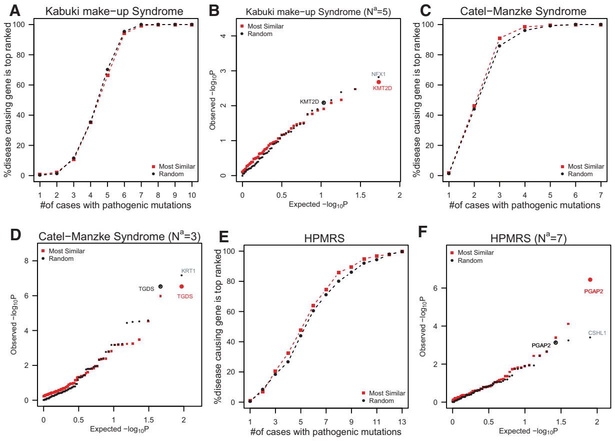

Fig. 5. RVAS for three recently resolved monogenic disorders. Kabuki make-up syndrome, Mabry syndrome, and Catel–Manzke syndrome are Mendelian disorders for which pathogenic mutations have been identified in new disease genes by intersections of candidate variant sets in case groups of unrelated, affected individuals comprising 10, 7 or 13 cases. For fewer cases the approach of an intersection filter would have been inconclusive. By a RVAS approach with 40 controls, however, the probability of identifying the disease gene in such cohorts is still considerable even when the number of cases with pathogenic mutations in these cohorts is reduced markedly (**A, C** and **E**). Additionally, a selection of similarity-matched controls may also help to reduce spurious associations effectively: The QQ plots (**B**, **D** and **F**) show the observed versus the expected *P*-values for instances of the RVAS simulations where 5, 3 and 7 individuals had pathogenic mutations in the disease genes *KMT2D, TGDS* and *PGAP2* and showed the lowest *P*-value only when similarity-matched controls were used. When random controls are used, the *P*-values of highly variable genes such as *NFX1*, *KRT1* or *CSHL1* are inflated

by analyzing cohorts of unrelated, affected individuals with an approach that is commonly referred to as intersection filtering (Gilissen *et al.*, 2012; Robinson *et al.*, 2011). In the case of Kabuki make-up syndrome 10 individuals were analyzed, the HPMRS cohort comprised 13 samples and the disease gene in Catel–Manzke syndrome was identified in a case group of size 7. The example of the study of the Kabuki make-up syndrome shows also the limitations of the conventional intersection approach: First, the authors considered genes that showed rare non-synonymous variants in all of the affected individuals. This filtering criterion was fulfilled by 34 genes and didn't allow the disease gene identification. Only when the authors restricted their set of candidate variants to loss-of-function mutations, they could pinpoint a single gene, *KMT2D*, that showed such mutations in at least 7 out of 10 patients.

We hypothesized that a RVAS should be more straightforward and also more sensitive for disease gene identification and reanalyzed the Kabuki make-up cohort by subsequently increasing the number of pathogenic variants from 1 to 10. With the same set of pathogenic variants as previously published, a RVAS approach using 40 similarity matched controls would have ranked the disease gene at the top position in almost 100% of our simulations when at least 6 out of 10 individuals had a pathogenic non-synonymous mutation (Fig. 5A).

Also for the other two disorders, Catel–Manzke syndrome and HPMRS the disease gene identification is highly effective even if the fraction of samples with pathogenic mutations is smaller than in the initial study. Especially the disease gene *TGDS* in Catel–Manzke syndrome, has such a low variability, that it can readily be identified with as little as four affected samples (Fig. 5C). Spurious associations often occur for highly variable genes, for example, genes from the mucin family or for genes that show a higher rate of calling artifacts e.g. due to pseudogenes, such as in *KRT1*. Interestingly, the false positive error resulting from such genes can also be reduced by using a similarity-matched setup of the control group.

4 Discussion

The increasing availability of high-throughput sequencing technology has revealed the cause of multiple rare monogenic disorders over the recent years. Many of these discoveries could be made because classical linkage analysis in large affected pedigrees could limit the genomic search space. However, it seems that many of the low hanging fruits have been picked by now and that collections of unsolved cases are building up that require different analysis strategies.

In some of the rare disorders a single pathogenic allele does not need to be fully penetrant and might require additional modifiers to develop the phenotype. In autism spectrum disorders it seems to be the mutational load of functionally compromising variants in several genes that seems to matter (Krumm *et al.*, 2015). Especially in these instances case group association studies that focus on rare variants, might help to find further disease-related loci.

Although GWAS have been a big success in identifying target genes for complex disorders there are some key differences when it comes to rare. Mathieson and McVean (2012) showed that population substructure from rare variants is systematically different from stratification that is due to common variants. They demonstrated that existing methods to correct for confounding effects in CVAS cannot account for the substructure introduced by rare variants.

As correction of substructure is not feasible in RVAS, strategies to identify suitable controls seem to be a promising alternative. In our work we therefore focused on methods that match controls based on their similarity. For associations studies that build on SNP data there has already been pioneering work by Guan *et al.* (2009) on genotype-based matching approaches and the metrics that we analyze for selecting controls also build on these efforts. We studied systematically how NGS data quality, as well as different population backgrounds influence the chances of identifying the disease gene in small case groups of individuals with ultra-rare. We also found that in RVAS a similarity metric that puts a stronger weight on sharing rare alleles shows the best performance in the statistical analysis, especially if the data quality is heterogeneous. We also showed that the profiles of rare variants between different European populations and different sequencing studies do vary considerably, which has important consequences for the setup of RVAS. Whenever ancestral components are potential confounders, stratification-based matching approaches such as described by Epstein *et al.* (2012) should also be considered.

From a statistical point of view it is often not optimal to include as many controls as possible, if they are not matching due to data quality or ethnicity. The best performance may be achieved if samples that are prone to cause spurious associations are not included. Whenever there is high variability in the quality of the datasets in the case group a matching approach that is also sensitive to the rate of genotyping errors seems to be the method of choice. However, there is no simple rule of thumb for a selection strategy that applies to any case cohort. In some datasets simulations with known pathogenic mutations such as described in this work might help to decide whether sequencing of additional controls would improve the statistical power.

Acknowledgements

We thank two anonymous reviewers for helpful comments on a previous version of this article.

Funding

This work was supported by Deutsche Forschungsgemeinschaft grants to P.M.K. (KR3985/1-1) and S.M. (SFB 665). N.Z. received funding from China Scholarship Council.

Conflict of Interest: none declared.

References

Asimit,J. and Zeggini,E. (2010) Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.*, **44**, 293–308.

Bansal,V. *et al.* Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.*, **11**, 773–785.

Bush,W.S. and Moore,J.H. (2012) Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.*, **8**, e1002822.

Clarke,L. *et al.* (2012) The 1000 Genomes Project: data management and community access. *Nat. Methods*, **9**, 459–462.

Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Dudbridge,F. and Gusnanto,A. (2008) Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.*, **32**, 227–234.

Ehmke,N. *et al.* (2014) Homozygous and compound-heterozygous mutations in TGDS cause Catel-Manzke syndrome. *Am. J. Hum. Genet.*, **95**, 763–770.

Eichler,E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

Epstein,M.P. *et al.* (2012) Stratification-score matching improves correction for confounding by population stratification in case-control association studies. *Genet. Epidemiol.*, **36**, 195–205.

Genomes Project Consortium. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Genomes Project Consortium. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Gilissen,C. *et al.* (2012) Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.*, **20**, 490–497.

Guan,W. *et al.* (2009) Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genet. Epidemiol.*, **33**, 508–517.

Heinrich,V. *et al.* (2013) Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects. *Genome Med.*, **5**, 69.

Jager,M. *et al.* (2014) Jannovar: a java library for exome annotation. *Hum. Mutat.*, **5**, 548–555.

Kang,H.M. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.

Krawitz,P.M. *et al.* (2013) PGAP2 mutations, affecting the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation syndrome. *Am. J. Hum. Genet.*, **92**, 584–589.

Krumm,N. *et al.* (2015) Excess of rare, inherited truncating mutations in autism. *Nat. Genet.*, **47**, 582–588.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Mathieson,I. and McVean,G. (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.*, **44**, 243–246.

Moore,C.B. *et al.* (2013) Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. *PLoS Genet.*, **9**, e1003959.

Ng,S.B. *et al.* (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.*, **42**, 790–793.

Nievergelt,C.M. *et al.* (2007) Generalized analysis of molecular variance. *PLoS Genet.*, **3**, e51.

Nothnagel,M. *et al.* (2011) Technology-specific error signatures in the 1000 Genomes Project data. *Hum. Genet.*, **130**, 505–516.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Purcell,S. *et al.* (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.*, **81**, 559–575.

Robinson,P.N. *et al.* (2011) Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clin. Genet.*, **80**, 127–132.

Schaid,D.J. (2010) Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. *Hum. Hered.*, **70**, 109–131.

Sham,P.C. and Purcell,S.M. (2010) Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.*, **15**, 335–346.

Turner,S. *et al.* (2011) Quality control procedures for genome-wide association studies. *Curr. Protoc. Hum. Genet.* Chapter 1:Unit 1. 19.

Voight,B.F. and Pritchard,J.K. (2005) Confounding from cryptic relatedness in case-control association studies. *PLoS Genet.*, **1**, e32.

Zhuang,J.J. *et al.* (2010) Optimizing the power of genome-wide association studies by using publicly available reference samples to expand the control group. *Genet. Epidemiol.*, **34**, 319–326.

Zuk,O. *et al.* (2014) Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. USA*, **111**, E455–E464.