Friedrichs et al. BMC Genetics 2016, 17(Suppl_2):8
DOI 10.1186/s12863-015-0313-x

BMC Genetics

PROCEEDINGS Open Access

Filtering genetic variants and placing informative *priors* based on putative biological function

Stefanie Friedrichs[1], Dörthe Malzahn[1], Elizabeth W. Pugh[2], Marcio Almeida[3], Xiao Qing Liu[4,5] and Julia N. Bailey[6,7*]

From Genetic Analysis Workshop 19
Vienna, Austria. 24-26 August 2014

Abstract

High-density genetic marker data, especially sequence data, imply an immense multiple testing burden. This can be ameliorated by filtering genetic variants, exploiting or accounting for correlations between variants, jointly testing variants, and by incorporating informative *priors*. *Priors* can be based on biological knowledge or predicted variant function, or even be used to integrate gene expression or other omics data. Based on Genetic Analysis Workshop (GAW) 19 data, this article discusses diversity and usefulness of functional variant scores provided, for example, by PolyPhen2, SIFT, or RegulomeDB annotations. Incorporating functional scores into variant filters or weights and adjusting the significance level for correlations between variants yielded significant associations with blood pressure traits in a large family study of Mexican Americans (GAW19 data set). Marker rs218966 in gene *PHF14* and rs9836027 in *MAP4* significantly associated with hypertension; additionally, rare variants in *SNUPN* significantly associated with systolic blood pressure. Variant weights strongly influenced the power of kernel methods and burden tests. Apart from variant weights in test statistics, *prior* weights may also be used when combining test statistics or to informatively weight *p* values while controlling false discovery rate (FDR). Indeed, power improved when gene expression data for FDR-controlled informative weighting of association test *p* values of genes was used. Finally, approaches exploiting variant correlations included identity-by-descent mapping and the optimal strategy for joint testing rare and common variants, which was observed to depend on linkage disequilibrium structure.

Background

With the availability of very dense genetic marker data sets, such as sequence data, even large association studies can become underpowered. This raises the need to filter, or prioritize, or jointly test genetic variants.

Filters or *priors* on genes may be derived from methylation or expression data if available in the same individuals. Alternatively, one may use external information. Recently, multiple annotation tools have become available using several databases and algorithms that predict functional effects of genetic variants. Commonly used are, for example, ANNOVAR (Annotate Variation) [1], VariantTools [2], PolyPhen [3], SIFT (Sorting Intolerant From Tolerant) [4], ENCODE (Encyclopedia of DNA Elements) [5], RegulomeDB [6], CADD (Combined Annotation-Dependent Depletion) [7], or Gerp++ [8]. Tools like ANNOVAR additionally provide variant annotation to genes and to regions such as conserved regions among species, predicted transcription factor binding sites, and segmental duplication regions. Many of the above-listed tools also provide information on regulatory elements that control gene activity. This article demonstrates that functional scores can contribute to the success of association studies. Simultaneously, functional scores may differ substantially between databases and prediction tools as they can be based on different functional aspects.

* Correspondence: JBailey@mednet.ucla.edu
Stefanie Friedrichs and Dörthe Malzahn share first authorship.
[6]Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, USA
[7]Epilepsy Genetics/Genomics Laboratory, West Los Angeles Veterans Administration, Los Angeles, CA, USA
Full list of author information is available at the end of the article

Additionally, variant annotations to chromosomal positions continue to be updated with the National Center for Biotechnology Information (NCBI) [9] human genome build as standard. Furthermore, variants can be annotated to genes based on different sources, such as ENSEMBL [10], Vega [11], GENCODE [12], and many more. Researchers also use a variety of definitions of flanking regions. Finally, genes may be grouped by function or biological pathway, again with substantial variability between data bases such as KEGG [13], Biocarta [14], or Pathway Interaction Database [15]. This article discusses approaches that filtered or prioritized genetic variants, regions, or genes. Pathway-based approaches, although also incorporating filters or *priors*, are discussed separately by Kent [16].

Many researchers filter genetic variants. The simplest forms of filters are minor allele frequency (MAF), candidate genes or variants, or considering the exome. Filters and statistical models are chosen to increase the power under a hypothetical disease model. The advent of sequencing renewed interest in disease mechanisms less frequent but more penetrant than common single nucleotide polymorphisms (SNPs) of genome-wide association studies (GWAS). This led, for example, to screening for recessive variants by examining runs of homozygosity [17, 18]. When multiple rare causal variants cluster within a gene, identity-by-descent (IBD) mapping may be more powerful than single-locus association testing [19]. IBD mapping can be used in 2-step approaches. For example, Balliu et al [20] identified regions where hypertension cases shared more segments of IBD than controls in one part of the sample. They modeled aggregate effects of each of these regions on blood pressure (BP) in the sample remainder. Aggregation tests are used especially for testing rare single-nucleotide variants (SNVs). Aggregation tests are burden tests, variance-component tests, or a combination of both, such as SKAT-O (optimal unified sequence kernel association test) (see, eg, Lee et al [21] for a review). Kernel-based approaches (see Schaid [22] for a review) such as the sequence kernel association test (SKAT) [23] are variance-component tests. Examples of genetic burden tests are T5, combined multivariate collapsing (CMC) [24], or C-α [25]; see also Santorico et al [26]. Aggregation tests can prioritize SNVs by weighting minor allele dosages in the test statistic. Typical weights account for MAF, but may also incorporate putative functional relevance of SNVs [27, 28]. Moreover, weights may be used to combine aggregation test statistics [21, 29, 30], and one may weight p values while controlling the false discovery rate (FDR) [31, 32]. For example, GWAS p values may be weighted based on functional annotations. For aggregation tests on genes, p value weights can be utilized to integrate gene expression or other omics data [33].

This article summarizes contributions of the Genetic Analysis Workshop (GAW) 19 group on filtering variants and placing informative *priors* (Tables 1 and 2). These investigations found that improving SNV grouping or selection can noticeably increase power. Moreover, including functional scores or gene expression data as filters or weights on variants, genes, or when combining test statistics assisted in detecting associations. Some contributions also exploited SNV correlations to increase power or improved the multiple-testing adjusted significance threshold by accounting for SNV correlations.

Materials

Analyzed data were provided by GAW 19 and included a family sample (n = 959) with extended pedigrees of Mexican Americans from the San Antonio Family Heart Study (SAFHS) and the San Antonio Family Diabetes/ Gallbladder Study (SAFDS/ SAFGS) [34]. The family sample also contained 103 unrelated sequenced subjects; 259 subjects had gene expression data. This study was designed to identify low-frequency or rare variants influencing susceptibility to type 2 diabetes (T2D) as part of the T2D Genetic Exploration by Next-generation sequencing in Ethnic Samples (T2D-GENES) Consortium. Phenotypes included real and simulated longitudinal systolic (SBP) and diastolic blood pressure (DBP) and hypertension (HT) status. Available were sequence for 464 pedigree members and GWAS SNPs for all 959 subjects. Additionally, all subjects were imputed to sequence based on original genotypes and familial relationships [34]. Approaches described herein mostly analyzed imputed dosages to avoid missing genotypes and to maximize sample size. Zhang et al [28] analyzed the GAW19 sample of 1943 independent Hispanic subjects with whole exome sequence. This sample had been ascertained by T2D status. However, GAW19 provided real and simulated cross-sectional BP traits instead [35], using the same trait-simulation model as for the family study.

All approaches described herein are nonlongitudinal analyses of BP traits (SBP, DBP, or HT) in relation to minor allele dosages of sequence SNVs or genome-wide SNPs.

Methods

Statistical methods employed by this group (see Table 1) to incorporate filters or informative *priors* are mostly based on regression models [27, 30, 33, 36, 37]; one is also based on counting methods [28]. Analyses of family data adjusted for familial dependence based on the kinship matrix. They included the familial covariance in a linear mixed model [27, 30, 36] or transformed the trait to a conditionally independent surrogate variable [33]. Analyses of independent subjects accounted for population structure (cryptic relatedness and admixture) [37] by using the programs Eigensoft [38] and Admixture [39].

Annotating genetic variants for location and function

A variety of freely available genetic databases and highly developed software tools support annotation of location

Table 1 Statistical tests and analyzed data

Marker data	Data set	Statistical tests	Covariates	Trait(s)
Almeida et al [36]				
Sequence	Family study	Single-variant regression in SOLAR	Smoking, BP medication, PC1-3, sex, age, age^2, sex*age, sex*age^2	Real SBP and DBP at first time point, own simulated trait for H_0
Liu et al [37]				
Chr3: GWASmp and sequence	Unrelated individuals (from family study)	Regress pairwise DBP residual difference and sum on IBD sharing status; sequence data analyses by SKAT-O	Sex, age, smoking, PC 1-3	Real DBP at first time point
Kim and Wei [27]				
Sequence	Family study	Informative SNV weights in burden test T5 and SKAT; with R: seqMeta	Age, sex, smoking, BP medication	Real SBP at earliest available measurement
Zhang et al [28]				
Exome sequence	Unrelated individuals (large Hispanic sample)	LRT, C-*α*, CMC on informatively weighted SNV burden	None	Simulated HT status; real SBP, DBP with cutoffs for case-control status
Malzahn et al [30]				
Sequence and GWASmp	Family study	SKAT with R (coxme, kinship2, QuadCompForm); strategies for joint testing of rare and common SNVs	Sex, age, sex*age; subjects not on BP medication	Real and simulated SBP at first time point
Ho et al [33]				
Sequence and GWASmp	Family study, including gene expression data	Seq-aSum-VS burden test; regression on gene expression data; gene set enrichment analysis	PC1-3	Average real SBP and DBP

BP blood pressure, *Chr* Chromosome, *CMC* Combined multivariate collapsing, *DBP* diastolic blood pressure, *GWASmp* genome-wide association study marker panel, *HT* hypertension, *IBD* identity-by-descent, *LRT* likelihood ratio test, *PC* principal component, *SBP* systolic blood pressure, *SKAT* sequence kernel association test, *SNV* single nucleotide variant, *Seq-aSum-VS* sequential sum

and biological function of SNVs. In our group, SNV locations were obtained by ANNOVAR [28, 36] or determined based on reference data, for example, from the Genome Reference Consortium [40] or the International Haplotype Map (HapMap) Consortium [41] [30, 37]. Reference data were also used to determine linkage disequilibrium (LD) blocks [30] with Haploview [42].

Kim and Wei [27] and Almeida et al [36] used functional annotations from ENCODE, PolyPhen or PolyPhen2, and SIFT, while Liu et al [37] used CADD. In contrast, Zhang et al [28] annotated putative protein binding sites based on 2 different algorithms using random forest classifiers [43].

Filtering genetic variants

Not all areas of the genome were studied. Some researchers filtered the data prior to analyses. Zhang et al [28] investigated exome sequence and Almeida et al [36] molecularly functional nonsynonymous SNVs predicted by PolyPhen and SIFT. Liu et al [37] examined IBD sharing regions on chromosome 3. Malzahn et al [30] considered gene-containing LD blocks for selected candidate genes. Ho et al [33] analyzed rare SNV burden in genes containing less than 50 and more than 1 rare SNV (MAF <0.01).

Accounting for correlations between genetic variants

An important difference between methods is that variant correlations can either be a nuisance or may be used to increase power. For example, IBD mapping exploits variant correlations. IBD mapping can be more powerful than single-locus association testing when multiple causal rare variants cluster within a gene [19]. Therefore, Liu et al [37] tested the relationship between IBD sharing status and trait differences and sums for pairs of individuals. Moreover, the power of kernel methods such as SKAT may be increased through the exploitation of variant correlations [44]. This ability can be utilized fully by analyzing LD blocks [30]. On the other hand, single-locus methods need to account for variant correlations to appropriately correct the significance level for multiple testing. Hence, Almeida et al [36] determined the effective number of independent tests by extreme value theory based on replicates of a simulated unassociated trait.

Correcting the significance level for the number of independent tests

The significance level used with multiple testing is always an issue as too conservative a correction will cause false negatives and not correcting enough will cause false positives.

Table 2 Filters, *priors*, and findings

Filter	*Prior*	Conclusions	Annotation
Almeida et al [36]			
Functional annotation, LD-corrected effective number of tests	None	LD-correction in WGS reduces multiple-testing burden by 85 %, significant associations: *PFH14* with SBP, *MAP4* with DBP	Location: ANNOVAR; functional annotation: PolyPhen, SIFT
Liu et al [37]			
IBD sharing	None	No significances, *ZPLD1* had strongest evidence	IBD mapping: BEAGLE; functional annotation: CADD
Kim and Wei [27]			
Sliding window on MAF ≤5 % SNVs	*SNV-weights:* based on MAF or regulatory importance	Significant association: *SNUPN*	Functional annotation: ENCODE, RegulomeDB, PolyPhen2
Zhang et al [28]			
Genes, exome-sequence	*SNV-weights:* up-weight protein binding sites, apply direction weights	Top-ranked genes differ between weighted burden tests LRT, C-α, CMC; but good overlap with literature	ANNOVAR, variant tools; random forest classifiers assign SNVs to protein binding sites; DSSP, PSAIA, DOMINO
Malzahn et al [30]			
Gene covering LD-blocks	*SNV-weights:* using MAF *Overall weight:* on rare SNV variance component in SKAT	SKAT: power depends on SNV weights, exploiting LD is very beneficial, optimal strategy for joint testing rare and common SNVs depends on LD structure	Haploview with HapMap data for LD-calculation
Ho et al [33]			
Rare SNVs in genes with >1 and <50 rare SNVs (MAF < 0.01)	*p value weights:* improve gene ranking	Power of burden tests improved by incorporating phenotype associated gene expression into *p* value weights	Genes: hg19; GO biological process categories

CADD combined annotation dependent depletion, *DBP* diastolic blood pressure, *DOMINO* database of domain–peptide interactions, *DSSP* define secondary structure of proteins, *ENCODE* encyclopedia of DNA elements, *GO* gene ontology, *IBD* identity-by-descent, *LD* linkage disequilibrium, *MAF* minor allele frequency, *PSAIA* protein structure and interaction analyzer, *SBP* systolic blood pressure, *SIFT* sorting intolerant from tolerant, *SKAT* sequence kernel association test, *SNV* single nucleotide variant, *WGS* whole genome sequence

Almeida et al [36] adjusted the significance level for single locus analyses by estimating the number of independent tests [45]. A total of 1000 replicates of a quantitative phenotype with no genetic effects were simulated and tested on whole genome sequence data, using linear mixed models in SOLAR (Sequential Oligogenic Linkage Analysis Routines) [46]. The smallest p value per simulation run was extracted. The density of these 1000 extremely small p values was fitted to a theoretical beta distribution $beta(1,n_e)$ where n_e is the effective number of independent tests [47]; yielding the adjusted significance level $a^* = \frac{0.05}{n_e}$. This procedure was applied to both whole genome sequence and functional nonsynonymous SNVs.

Identity-by-descent mapping

IBD mapping aims to detect loci sharing ancestral segments in unrelated individuals. In particular, unrelated subject-pairs with smaller trait differences are expected to share significantly more rare causative variants than pairs with larger trait differences. Liu et al [37] estimated IBD sharing segments with BEAGLE [48]. The squared trait difference (D) and squared trait sum (S) for trait DBP between pairs of unrelated subjects was regressed on IBD sharing status. This yielded parameter estimates for slopes $(\hat{\beta}_S, \hat{\beta}_D)$ and variances (σ^2_S, σ^2_D), which were combined into an overall slope estimate $\hat{\beta} = \left(\frac{\sigma^2_D}{\sigma^2_S+\sigma^2_D}\right)\hat{\beta}_s + \left(\frac{\sigma^2_S}{\sigma^2_S+\sigma^2_D}\right)\hat{\beta}_D$. Linkage was tested with test statistic $t = \frac{\hat{\beta}}{SE(\hat{\beta})}$ under the null hypothesis of an overall slope of zero [37]. The significance threshold for non-independent pairs was estimated by permutation procedure.

Priors on genes and variants

Genetic *priors* can be incorporated by variant weights in aggregation tests such as burden tests or SKAT [21]. Burden tests collapse minor allele dosages x_{ik} of a set of $i = 1, \ldots, m$ variants into a burden score $s_k = \sum_{i=1}^{m} \omega_i x_{ik}$ per individual k using a priori specified variant weights ω_i. One tests trait association with genetic burden s_k. Although burden tests are powerful when causal SNVs have the same effect direction, SKAT is more powerful when effect directions differ or if many noncausal SNVs are included in testing [21, 49]. SKAT is based on an underlying Bayesian model that estimates a random effect per SNV [50]. Specified is a kernel matrix of genetic

between-subject similarity and this kernel constitutes a *prior* on genetic model space [51]. SNV weights are incorporated in the kernel (see, eg, Malzahn et al [30]).

Typically, rarer SNVs get assigned more weight to counterbalance their reduced power compared to more frequent SNVs. Used are, for example, weights $\omega_j = \frac{1}{\mathrm{MAF_j}\left(1-\mathrm{MAF_j}\right)}$ [52], inverse MAF weights $\omega_j = \frac{1}{\mathrm{MAF_i}}$, or *beta*-weights such as $\omega_j = b(\mathrm{MAF_i})$ [23], where b is the probability density function of a $beta(1, 25)$ random variable. Malzahn et al [30] compared the power of SKAT when using different SNV weights and different kernel functions that either allow or do not allow for SNV interactions in the genetic model. Alternatively, SNV weights may be based on regulatory importance [27] or protein binding effects [28].

Incorporating functional information into variant weights

Kim and Wei [27] categorized SNVs according to RegulomeDB and PolyPhen2 functional relevance scores. SNV weights were defined based on $f(s) = S^2$ where s equaled the reverse order of categories, namely s = 6, 5, 4, 3, 2, 1 for category 1 ("most likely affecting binding and expression") to category 6 ("not functionally relevant"). Kim and Wei [27] tested rare SNVs jointly, in sets defined by sliding windows of 4 kb size, for association with SBP. They compared the power of SNV weighting schemes in SKAT ($\omega_j = \sqrt{f(s_j)}$ versus $\omega_j = b(\mathrm{MAF_j})$), and burden test T5 ($\omega_j = f(s_j)$ versus $\omega_j = \frac{1}{\mathrm{MAF_j}\left(1-\mathrm{MAF_j}\right)}$). SKAT and T5 provide analytical asymptotically exact p values with good small sample size behavior.

Zhang et al [28] used a likelihood ratio test (LRT) [53] to test if the proportion of subjects with an informatively weighted minor allele burden exceeding a given threshold differed between HT cases and controls. P values were obtained by permutation procedure. SNV weights ω_i accounted for putative effect direction and distinguished between functional SNVs in binding-sites ($|\omega_i| = 10$), not in binding-sites ($|\omega_i| = 5$), and nonfunctional SNVs ($|\omega_i| = 1$). The informatively weighted LRT was compared with C-α and CMC burden tests.

Optimal joint testing of rare and common variants

When not filtering for rare or common SNVs, optimal joint testing of both becomes an issue. Suppose, one computed 2 SKAT statistics, $\mathrm{Q_{rare}}$ and $\mathrm{Q_{common}}$, separately on rare SNVs and common SNVs, in the same region of interest, for the same trait, based on the same genetic null model. As SKAT is a variance-component test, combining $\mathrm{Q_{rare}}$ and $\mathrm{Q_{common}}$ [29]

$$\mathrm{Qws} = (1-\lambda)\cdot\mathrm{Q_{rare}} + \lambda\cdot\mathrm{Q_{common}} \quad (1)$$

weights the rare SNV variance-component by overall a priori weight (1-λ) relative to the common SNV variance-component (see Ionita-Laza et al [29] and Malzahn et al [30] for choices of λ). The weighted sum test (1) is another way of structuring a *prior* in SKAT. Note that $\mathrm{Q_{rare}}$ and $\mathrm{Q_{common}}$ may use different kernel functions or different SNV weights. Malzahn et al [30] compared this form of joint testing of rare and common SNVs with the default choice of entering *all* SNVs with appropriate weights into a *single* kernel. Exact p values for SKAT and weighted sum test (1) were obtained by Davies method [54]. Another investigated alternative was Fisher pooling of the correlated p values resulting from the separate rare SNV and common SNV SKAT statistics. Fisher pooling accounted for correlations by Satterthwaite approximation and Brown's method ([55]; see also [29, 30]).

Note that analogously to equation (1), SKAT-O combines SKAT and burden tests with statistic $Q = (1-\rho)Q_{SKAT} + \rho Q_{burden}$ where $0 \le \rho \le 1$ [56].

Informed *p* value weighting for genes

Ho et al [33] obtained gene-wise p values, p_g, for association of average BP T with rare SNV burden s_g in genes g that had more than 1 and less than 50 rare SNVs (MAF <0.01)

$$T \sim b_{s,g}\cdot s_g \quad (2)$$

Restricting the number of rare SNVs avoids collapsing too many null variants. Ho et al [33] used the sequential sum test [57], which data-adaptively assigned SNV weights $\omega_i = 0, 1, -1$. Earlier, Genovese et al [31] and Roeder and Wasserman [32] had proven that informative weighting of p values $\frac{p_g}{v_g}$ with weights $v_g > 0$, $\bar{v}_g = 1$ maintains proper FDR control; where $\frac{p_g}{v_g} \le \alpha_{FDR}$ means significance. Ho et al [33] determined such weights v_g as follows. They tested if rare minor allele burden s_g^* (with SNV weights $\omega_i = 1$, for simplicity) also associated with gene expression E_g

$$E_g | T \sim b_{E,g}\cdot s_g^* + c\cdot T \quad (3)$$

and further if gene expression E_g associated with trait value T

$$T | s_g^* \sim b_{T,g}\cdot E_g + d\cdot s_g^* \quad (4)$$

Association tests (2) to (4) were made conditionally independent by adjusting test (3) for trait value T and test (4) for rare minor allele burden s_g^* (Fig. 1). P value weights $v_g = v_g^*\bar{v}_g^*$ were derived as $v_g^* = max\left(\left(\frac{\hat{b_{E,g}}}{SE\left(\hat{b_{E,g}}\right)}\right)^2 \times \left(\frac{\hat{b_{T,g}}}{SE\left(\hat{b_{T,g}}\right)}\right)^2\right)$ where the maximum was over all gene expression measurements and $\bar{v}_g^*$ was the average of all v_g^*.

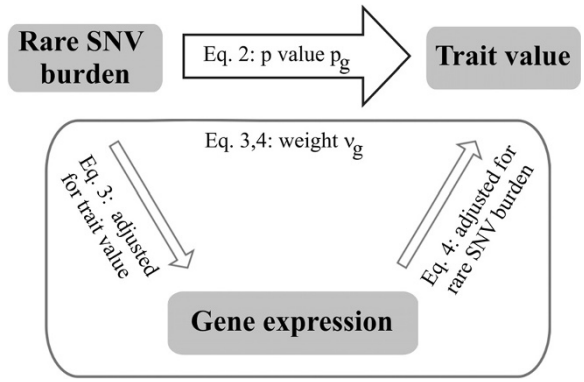

Fig. 1 Informed *p* value weighting for genes based on conditionally independent associations between rare variant burden, gene expression, and trait. The *p* value weight v_g was defined as the product of the association strengths of rare SNV burden with gene expression and gene expression with trait value

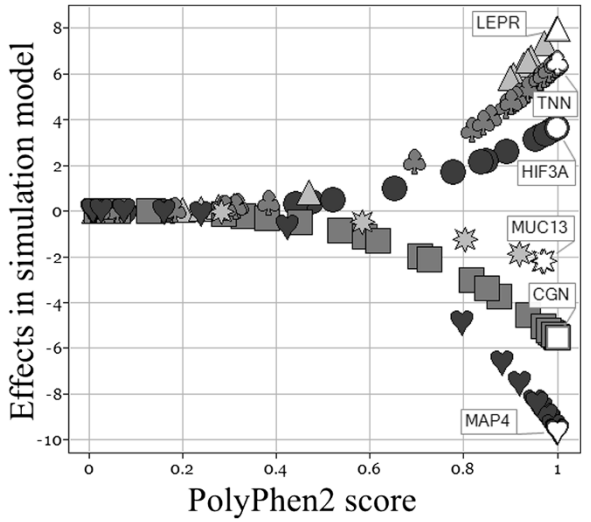

Fig. 2 SNV effect sizes on GAW19 simulated DBP increase with increasing PolyPhen2 scores. Depicted are 6 genes with a range of SNV effect sizes that could be simultaneously displayed. Symbols depict SNVs in the same gene: LEPR (▲), TNN (♣), HIF3A (●), MAP4(♥), MUC13(✳), CGN(■)

Results and discussion

The results for this GAW19 working group varied widely as a result of the different objectives of each contributor. Table 2 provides a brief summary of specific results.

Under H_0, extreme p values follow a beta distribution [47]. Almeida et al [36] reported that the beta distribution provided an excellent fit to determine the effective number of independent tests n_e for n single-locus tests. For whole genome sequence, $\frac{n_e}{n} = 15\%$; that is, accounting for LD reduced the multiple-testing burden by 85 %. However, significant associations could only be found when LD-correcting the significance level after a priori reducing sequence data based on functional annotations. Then 2 SNPs were detected: rs218966 in gene *PHF14* associated with SBP and rs9836027 in *MAP4* associated with DBP.

Liu et al [37] scanned chromosome 3 (GWAS data) for IBD sharing segments that associated with DBP. No genome-wide significance was found. However, several risk variants were detected in the region of gene *ZPLD1* by using CADD functional scores and sequence for the most promising region at 3q12.3.

In the GAW19 trait simulation model, SNV effect sizes were based on PolyPhen2 functional prediction scores (Fig. 2) [35]. In Figs. 2 and 3, displayed SNV effects, PolyPhen2 scores, and the assignment to positions and genes (NCBI build37, human genome build 19) came from the simulation answers. To illustrate differences between functional annotations, SIFT scores (and rs-numbers) were added by annotating sequence (variant call format [vcf] files) with ANNOVAR and merging vcf files and simulation answers by chromosome and position. RegulomeDB scores were merged by dbsnp138 rs-identifier. Furthermore, functional scores were transformed to have the same directionality (Fig. 3). Different functional annotations focus on different information about SNVs and only annotate selected SNVs. PolyPhen2 and SIFT both annotate nonsynonymous coding SNVs by a metric score that can be categorized to distinguish benign mutations from damaging ones affecting protein function. Nevertheless, PolyPhen2 and SIFT scores differ to a substantial extent in value and category (Fig. 3a). RegulomeDB annotates regulatory SNVs by an ordinal score ranging from the highest evidence (eQTL, expression quantitative trait locus) to the lowest. Figure 3c illustrates that some SNVs were rated to affect gene expression and transcription factor binding (RegulomeDB scores 1 to 5) but not the protein function (scored "benign" by PolyPhen2). For *simulated* BP, SIFT and RegulomeDB annotations yield mismatched filters or *priors* whenever they deviate from the PolyPhen2 score used to simulate SNV effects. For example, SIFT annotated some SNVs with large effects in gene *TNN* as benign mutations (Fig. 3b) and only few SNVs in associated genes were rated to be of regulatory importance (Fig. 3d). Nevertheless, for *real* SBP, several multiple-testing adjusted significant windows (2 with SKAT, 4 with burden test T5) were only found when including RegulomeDB scores as variant weights for rare SNV analysis [27]. One of these regions contained *SNUPN* [27] which is a novel finding not previously reported to associate with BP. T5 and SKAT maintained the nominal significance level on simulated unassociated trait Q1 also when incorporating RegulomeDB scores into variant weights [27]. Kim and Wei [27] and Zhang et al [28]

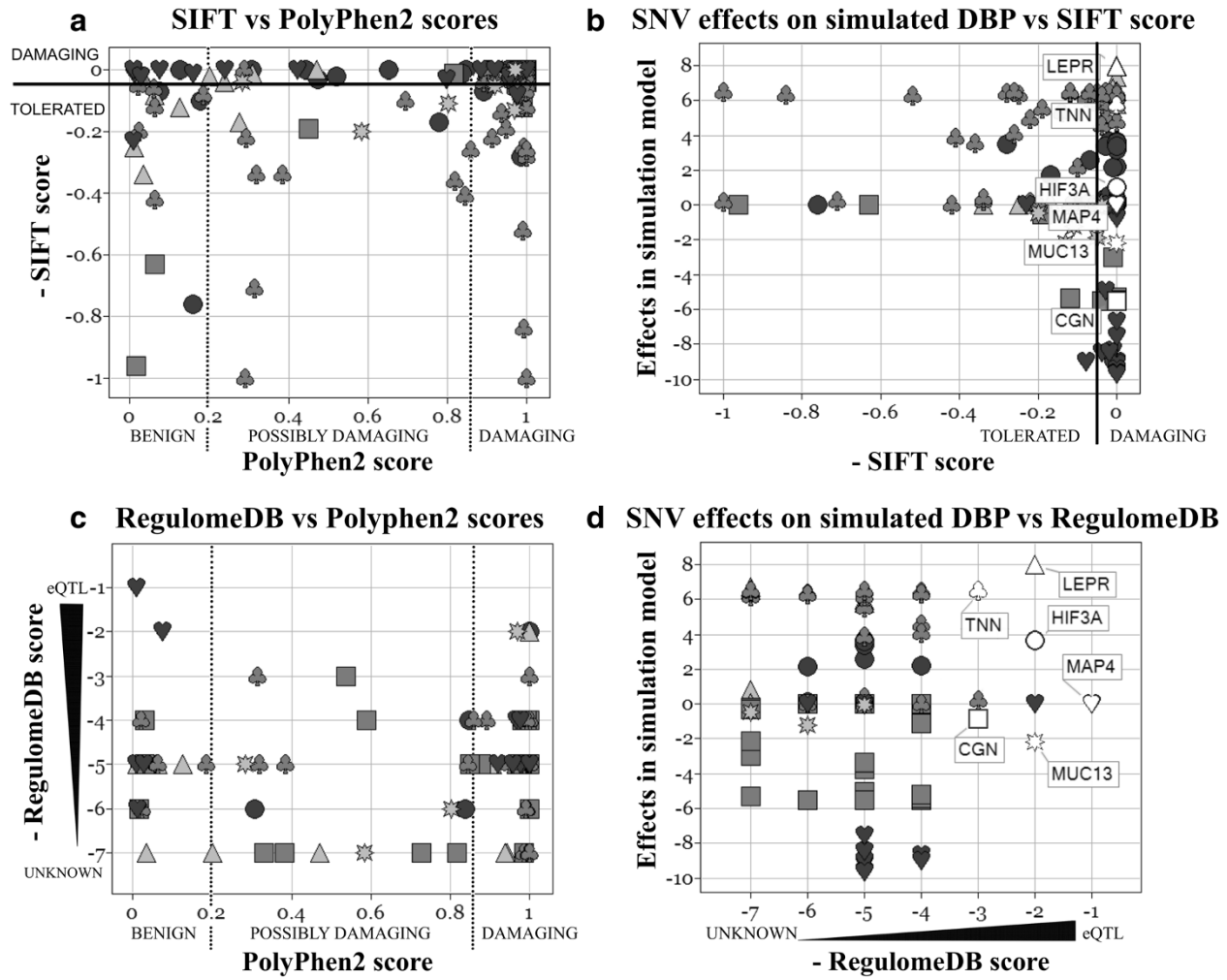

Fig. 3 Comparison between the PolyPhen2, SIFT, and RegulomeDB functional prediction scores. *Left column*: Correlation of PolyPhen2 functional prediction scores with (**a**) SIFT or (**c**) RegulomeDB scores. Functional scores were transformed to have the same directionality. Nonsynonymous coding SNVs that alter the protein function should receive a PolyPhen2 score of 1 and a SIFT score of 0. Scores are metric and can be categorized as displayed. RegulomeDB annotates regulatory SNVs by an ordinal score ranging from the highest evidence (eQTL, expression quantitative trait locus) to the lowest. *Right column*: Filters or *priors* based on (**b**) SIFT or (**d**) RegulomeDB functional scores are partially mismatched on GAW19 simulated DBP. Symbols depict SNVs in the same gene: LEPR (▲), TNN (♣), HIF3A (●), MAP4(♥), MUC13(✳), CGN(■)

both recommended using relatively big differences in SNV weights distinguishing functional from nonfunctional SNVs. Zhang et al [28] observed that different burden tests with functionally informative SNV weights yielded different top ranked genes. Although no gene was significant, many of them had been reported in the BP literature before. For SKAT, Malzahn et al [30] found that variant weights, but not kernel choice, had a strong influence on power, for rare as well as common SNVs. Kernel methods may gain power by exploiting SNV correlations. This can be utilized fully by analyzing LD blocks [30]. LD structure also influenced which strategy yielded the best joint test of rare and common SNVs with SKAT [30].

When using gene expression data to informatively weight gene-wise p values for association of rare SNV burden with BP [33], 153 genes (out of 6118) reached nominal significance (weighted $p \leq 0.05$). P value weights were determined such that evidence for phenotype associated gene expression lowered burden test p values. As no gene reached multiple-testing adjusted significance, Ho et al [33] used gene set enrichment analysis as aggregation test to relate the 153 top genes to biological pathways.

Conclusions

All analyses presented herein used a cross-sectional design by analyzing trait data of the first examination, the first available examination, or longitudinally averaged traits. This mainly contributed to differences in sample

size and trait variability. Furthermore, analyzing trait values at different time points may affect the marginal effect of genes that interact with age.

Including biological knowledge increased the power of association studies performed in our GAW group; especially filtering variants based on putative functional relevance. *Prior* weights can be included at different stages of the testing procedure. They can be incorporated into the test statistic of SKAT or burden tests, used when combining test statistics, or applied to association test p values. Selecting variant-sets also should take genetic structures into consideration, such as LD or IBD sharing. Moreover, the effective number of independent tests can be determined relatively easily by extreme value theory. This enables appropriate adjustment of the significance level for multiple testing to avoid an overly conservative approach. Ideally, variant grouping and selection, inclusion of biological information, and significance level adjustment can be applied simultaneously. Strategies like these are useful in increasing power in analyses of highly dense genetic data sets.

Filtering variants clearly boosted power in the discussed studies. However, filtering might also lose information. Functional scores such as PolyPhen2, SIFT, CADD, or RegulomeDB differ as they focus on different information about SNVs. Moreover, appropriateness of functional scores for a considered trait is a priori unknown. Hence, one is well advised to use and combine multiple functional annotations into a single filter or *prior*. This is feasible as functional annotations yield strong filters that greatly reduce the SNV space.

Competing interests
The authors declare they have no competing interests.

Authors' contributions
SF and DM wrote the manuscript. EWP contributed the comparison between the PolyPhen2, SIFT, and RegulomeDB functional annotation scores. All authors critically reviewed the manuscript for important intellectual content and interpretation of findings. All authors read and approved the final manuscript.

Acknowledgements
The authors would like to thank Zheyang Wu and Peng Wei for their comments and suggestions, as well as the GAW organizers for all their efforts. SF and DM were supported by the Deutsche Forschungsgemeinschaft (DFG, grant Research Training Group "Scaling Problems in Statistics" RTG 1644; grant Klinische Forschergruppe (KFO) 241: TP5, BI 576/5-1). EWP and JNB acknowledge support by National Institutes of Health (NIH) grants (HHSN268201200008I, R01 NS055057). XQL was supported by the University of Manitoba start-up funds. T2D-GENES is supported by NIH grants U01 DK085524, U01 DK085501, U01 DK085526, U01 DK085584 and U01 DK085545, the SAFHS by grant P01 HL045222, the SAFDS by grant R01 DK047482, and the SAFGS by grant R01 DK053889. Genetic Analysis Workshop 19 was supported by NIH grant R01 GM031575.

Declarations
This article has been published as part of *BMC Genetics* Volume 17 Supplement 2, 2016: Genetic Analysis Workshop 19: Sequence, Blood Pressure and Expression Data. Summary articles. The full contents of the supplement are available online at www.biomedcentral.com/bmcgenet/supplements/17/S2. Publication of the proceedings of Genetic Analysis Workshop 19 was supported by National Institutes of Health grant R01 GM031575.

Author details
[1]Department of Genetic Epidemiology, University Medical Center, Georg-August University Göttingen, Göttingen, Germany. [2]Center for Inherited Disease Research, Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA. [3]South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley, Brownsville, TX, USA. [4]Department of Obstetrics, Gynecology, and Reproductive Sciences, Department of Biochemistry and Medical Genetics, Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada. [5]Children's Hospital Research Institute of Manitoba, Winnipeg, MB, Canada. [6]Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, USA. [7]Epilepsy Genetics/Genomics Laboratory, West Los Angeles Veterans Administration, Los Angeles, CA, USA.

Published: 3 February 2016

References
1. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164–e164.
2. San Lucas FA, Wang G, Scheet P, Peng B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. Bioinformatics. 2012;28:421–2.
3. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit 7.20.
4. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.
5. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014;111:6131–8.
6. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22:1790–7.
7. 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;46:310–5.
8. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6:e1001025.
9. NCBI: National center for biotechnology information search database. http://www.ncbi.nlm.nih.gov/.
10. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):D662–9.
11. Harrow JL, Steward CA, Frankish A, Gilbert JG, Gonzalez JM, Loveland JE, et al. The vertebrate genome annotation browser 10 years on. Nucleic Acids Res. 2014;42:D771–9.
12. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res. 2012;22:1760–74.
13. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
14. Nishimura D. BioCarta. Biotech Softw Internet Rep. 2001;2:117–20.
15. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. Nucleic Acids Res. 2009;37:D674–9.
16. Kent Jr JW. Pathway-based analyses. BMC Genet. 2015;16 Suppl 3:S5.
17. Gibson J, Morton NE, Collins A. Extended tracts of homozygosity in outbred human populations. Hum Mol Genet. 2006;15:789–95.
18. Hildebrandt F, Heeringa SF, Rüschendorf F, Attanasio M, Nürnberg G, Becker C, et al. A systematic approach to mapping recessive disease genes in individuals from outbred populations. PLoS Genet. 2009;5:e1000353.
19. Browning SR, Thompson EA. Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics. 2012;190:1521–31.
20. Balliu B, Uh HW, Tsonaka R, Boehringer S, Helmer Q, Houwing-Duistermaat JJ. Combining information from linkage and association mapping for next-generation sequencing longitudinal family data. BMC Proc. 2014;8 Suppl 1:S34.
21. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95:5–23.

22. Schaid DJ. Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum Hered. 2010;70:109–31.
23. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). Am J Hum Genet. 2011;89:82–93.
24. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83:311–21.
25. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS Genet. 2011;7:e1001322.
26. Santorico SA, Hendricks AE. Progress in methods for rare variant association. BMC Genet. 2015;16 Suppl 3:S7.
27. Kim T, Wei P. Incorporating ENCODE information into association analysis of whole genome sequencing data. BMC Proc. 2015;9 Suppl 8:S34.
28. Zhang D, Cui H, Korkin D, Wu Z. Incorporation of protein binding effects into likelihood ratio test for exome sequencing data. BMC Proc. 2015;9 Suppl 8:S37.
29. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013;92:841–53.
30. Malzahn D, Friedrichs S, Bickeböller H. Comparing strategies for combined testing of rare and common variants in whole sequence and genome-wide genotype data. BMC Proc. 2015;9 Suppl 8:S36.
31. Genovese CR, Roeder K, Wasserman L. False discovery control with p-value weighting. Biometrika. 2006;93:509–24.
32. Roeder K, Wasserman L. Genome-wide significance levels and weighted hypothesis testing. Stat Sci. 2009;24:398–413.
33. Ho YY, Guan W, Basu S. Powerful association test combining rare variant and gene expression using family data from genetic analysis workshop 19. BMC Proc. 2015;9 Suppl 8:S33.
34. Almasy L, Dyer TD, Peralta JM, Jun G, Wood AR, Fuchsberger C, et al. Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. BMC Proc. 2014;8 Suppl 1:S2.
35. Blangero J, Teslovich TM, Sim X, Almeida MA, Jun G, Dyer TD, et al. Omics-squared: human genomic, transcriptomic and phenotypic data for Genetic Analysis Workshop 19. BMC Proc. 2015;9 Suppl 8:S2.
36. Almeida M, Blondell L, Peralta J, Kent JW, Jun G, Teslovich TM, et al. Independent test assessment using the extreme value distribution theory. BMC Proc. 2015;9 Suppl 8:S32.
37. Liu X-Q, Fazio J, Hu PZ, Paterson AD. Identity-by-descent mapping for diastolic blood pressure in unrelated Mexican Americans. BMC Proc. 2015;9 Suppl 8:S35.
38. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2:e190.
39. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
40. GRC: The Genome Reference Consortium. http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/.
41. The International HapMap Consortium. The international HapMap project. Nature. 2003;426:789–96.
42. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.
43. Sikić M, Tomić S, Vlahovicek K. Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. PLoS Comput Biol. 2009;5(1):e1000278.
44. Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SL, Peyser P, et al. SNP set association analysis for familial data. Genet Epidemiol. 2012;36:797–810.
45. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genome wide association scans. Genet Epidemiol. 2008;32:227–34.
46. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 1998;62:1198–211.
47. Sidak Z. Rectangular confidence regions from means of multivariate normal distributions. J Am Stat Assoc. 1967;62:626–33.
48. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81:1084–97.
49. Chen H, Malzahn D, Balliu B, Li C, Bailey JN. Testing genetic association with rare and common variants in family data. Genet Epidemiol. 2014;38 Suppl 1:S37–43.
50. Liu D, Lin X, Ghosh G. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007;63:1079–88.
51. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge: MIT Press; 2006.
52. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5:e1000384.
53. Chen Y-C, Carter H, Parla J, Kramer M, Goes FS, Pirooznia M, et al. A hybrid likelihood model for sequence-based disease association studies. PLoS Genet. 2013;9:e1003224.
54. Davies RB. Algorithm as 155: the distribution of a linear combination of chi-2 random variables. J R Stat Soc: Ser C: Appl Stat. 1980;29:323–33.
55. Brown MB. A method for combining non-independent, one-sided tests of significance. Biometrics. 1975;31:987–92.
56. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13:762–75.
57. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. Genet Epidemiol. 2011;35:606–19.