Published in final edited form as:

Genet Epidemiol. 2014 September ; 38(0 1): S5–S12. doi:10.1002/gepi.21819.

Local and Global Ancestry Inference, and Applications to Genetic Association Analysis for Admixed Populations

Timothy A. Thornton^{1,*} and Justo Lorenzo Bermejo²

¹Department of Biostatistics, University of Washington, Seattle, WA, USA

²Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany

Abstract

Genetic association studies in recently admixed populations offer exciting opportunities for the identification of variants underlying phenotypic diversity. At the same time, genetic heterogeneity due to population admixture has to be accounted for in order to ensure validity of association tests. The whole genome sequence data and the genome-wide single nucleotide polymorphism chip data for Mexican American individuals provided by the Genetic Analysis Workshop 18 (GAW18) presents a unique opportunity to evaluate and compare methods for the statistical analysis of admixed genetic data. We summarize here the five contributions from the GAW18 Admixture group. Although group members considered a variety of research topics, the general theme was inference and consideration of ancestry admixture in genetic analyses. The topics considered can be grouped into three categories: (1) global and local ancestry inference and estimation; (2) association and admixture mapping; and (3) genotype imputation in admixed samples. We describe the approaches that were used and the most relevant findings from individual contributions. We also provide insight into the strengths and limitations of the state-of-the-art methods considered for genetic analyses in admixed populations.

Keywords

admixture; association; local ancestry; genotype imputation; GWAS

INTRODUCTION

While genetic association studies to identify variants that influence complex traits have primarily focused on populations of European ancestry, several recent studies investigate populations with admixed ancestry. Here we define recently admixed populations to be populations with ancestry derived within the last few hundred years from two or more progenitor groups that were reproductively isolated. A glossary of relevant terms and key concepts for genetic analysis in admixed populations is given in Table I. The two largest minority groups in the United States, African Americans and Hispanics, are examples of recently admixed populations.

*Correspondence to: Dr. Timothy A. Thornton, Department of Biostatistics, University of Washington, BOX 357232, Seattle, WA 98195, USA, (206) 543-8004, Fax: (206) 543-3286, tathornt@u.washington.edu.

Ancestry differences among sampled individuals from admixed populations can be a confounder in genetic association studies. It is well known that failure to appropriately account for population structure due to ancestry admixture can lead to both spurious association (increased type-I error rates) as well as reduced power (inflated type-II error rates). The heterogeneous genomes of individuals from admixed populations, however, may provide advantages over genetic association analyses in homogenous populations, including the possibility of gene mapping by admixture linkage disequilibrium, i.e., admixture mapping, and improved sequencing of the human genome by considering the inherited variation patterns created by population admixture, as has been recently demonstrated [Genovese et al., 2013].

The analysis of whole genome sequence (WGS) data and genome-wide single nucleotide polymorphism (SNP) chip data from Mexican American pedigrees provided by the Genetic Analysis Workshop 18 (GAW18) presents a number of challenges due to both ancestry admixture as well as relatedness among individuals. Mexican Americans are known to be a multi-way admixed population who descend from a combination of European, Native American and West African progenitors [Johnson et al. 2011; Manichaikul et al. 2012]. The five contributions to the GAW18 Admixture group considered a wide variety of topics under the general theme of inferring, accounting for, and incorporating ancestry admixture in genetic analyses. The authors considered admixture not only as a potential confounder but also exploited differential ancestry for improved association mapping and more accurate genotype imputation. In this summary paper we describe methods and analyses conducted by the Admixture group, as well as the most relevant findings of the contributors, providing insight into their strengths and limitations.

METHODS

GAW18 Data

The GAW18 dataset consists of 1,043 individuals from 20 large multi-generational Mexican American pedigrees. Whole genome sequence (WGS) and dense SNP genotype data for odd numbered autosomes were available. The number of individuals in the 20 pedigrees with available genotype data ranged from 22 to 86 individuals, and both real and simulated phenotype data on systolic blood pressure (SBP), diastolic blood pressure (DBP), and hypertension were provided for the GAW18 sample individuals. A detailed description of the data can be found elsewhere [Blangero et al., 2013].

Members of the Admixture group made different choices in the genomic markers used for analyses. Two groups [Thornton et al., 2013; Culverhouse et al., 2013] used SNP genotype data across all of the odd autosomes. The remaining three groups restricted their analyses to chromosome 3, the designated chromosome for GAW18 participants who analyzed a single chromosome. In particular, Chen et al. [2013] analyzed SNP genotype data, Yorgov et al. [2013] analyzed sequence data, and Huang and Tseng [2013] used both sequence and SNP genotype data from chromosome 3.

Heterogenous choices were also made by group members regarding the selection of individuals to be included in the analyses. Thornton et al. [2013], Culverhouse et al. [2013],

and Yorgov et al. [2013] performed analyses using all pedigree samples with available genotype data as well as subsets of unrelated individuals, Chen et al. [2013] selected a set of unrelated individuals from the sample for genetic analysis, and Huang and Tseng [2013] considered a subset of related and unrelated individuals from the sample.

Complex trait association and/or admixture mapping was a focus for three contributors, and both real and simulated phenotype data were analyzed. Chen et al [2013] and Yorgov et al [2013] analyzed real SBP and DBP phenotype data. Yorgov et al. additionally analyzed a simulated Q1 phenotype that was not influenced by genetic factors. Thornton et al. [2013] analyzed a simulated DBP phenotype.

The Admixture group members performed a variety of genetic analyses with the GAW18 data. The contributions can be grouped into three categories according to the following broad themes: (1) global and local ancestry inference and estimation, (2) association and/or admixture mapping, and (3) genotype imputation in admixed samples. An overview of the Admixture group contributions are presented in Table II and summarized below.

Global Ancestry Estimation

The chromosomes of an individual with admixed ancestry represent a mosaic of chromosomal blocks from the ancestral populations, and the overall genetic ancestry, or global ancestry of an individual, has previously been defined as the relative proportion of ancestral blocks from each contributing population across the chromosomes [Tang et al., 2005]. Thornton et al. [2013] estimated proportional European, African, Native American, and East Asian ancestry for all genotyped individuals from the 20 pedigrees in GAW18 using the ADMIXTURE software program [Alexander et al., 2009]. The estimation of proportional ancestry was supervised, based on SNP data for odd numbered autosomes, where the CEU and YRI samples of release 3, phase III of the International Haplotype Map Project (HapMap) [International Hapmap 3 Consortium, 2010] were used as surrogates for European and African ancestry, and surrogates for Native American and East Asian ancestry were obtained from the Human Genome Diversity Project (HGDP) [Li et al., 2008]. The Native American ancestry proxies from HGDP were the combined samples from the Americas (Surui, Maya, Karitiana, Pima, and Colombian samples). Chen et al. [2013] performed a similar supervised analysis with the ADMIXTURE but with (1) the analysis restricted to chromosome 3 and applied to a subset of unrelated individuals, and (2) proportional ancestries estimated for European, Native American, and African populations. Yorgov et al. [2013] estimated global ancestry proportions for all individuals by averaging supervised local ancestry estimates over chromosome 3 markers. Thornton et al. [2013] also performed an unsupervised individual ancestry estimation analysis with ADMIXTURE for the GAW18 individuals, where reference population samples from HapMap and HGDP were not included in the analysis.

Local Ancestry Estimation

Local ancestry is defined as the genetic ancestry of an individual at a particular chromosomal location, where an individual can have 0, 1 or 2 copies of an allele derived from each ancestral population. Two group members [Chen et al., 2013; Yorgov et al., 2013]

estimated local ancestry on chromosome 3 using the LAMP-LD software program [Baran et al., 2012]. Both Yorgov et al. [2013] and Chen et al. [2013] used the HGDP Native American samples and the HapMap CEU and YRI samples as the reference population panels for local ancestry estimation, and the SHAPEIT software was used for phasing the reference population samples. Yorgov et al. [2013] performed the analysis with 40,098 chromosome 3 markers to allow the LAMP-LD method to leverage the structure of linkage disequilibrium. Chen et al. [2013] additionally evaluated the LAMP [Sankararaman et al., 2008] and MULTIMIX [Churchhouse and Marchini, 2013] methods. Phased reference panel data was used with the MULTIMIX software, and local ancestry estimates when using phased and unphased genotype data for the GAW18 sample individuals were compared. For the local ancestry analysis with the LAMP software, Chen et al [2013] identified a set of 522 ancestry informative markers (AIMs) that were in low linkage disequilibrium (LD) by using genotype and allele frequency data from Hapmap CEU and YRI for European and African ancestries, respectively, as well as allele frequencies for the Mayan (MAY) and Pima (PMA) samples from the Allele FREquency Database (ALFRED) (<http://alfred.med.yale.edu/alfred/AboutALFRED.asp>) for Native American ancestry.

Principal Components Analysis with Pedigrees

Principal components analysis (PCA) has been the prevailing approach in recent years for inferring population structure. PCA has been shown to account for population structure in samples with unrelated individuals, and the EIGENSTRAT method of Price et al. [2006], where principal components corresponding to the highest eigenvalues are included as covariates in the subsequent association analysis, has been widely applied to genome-wide association studies for protection against confounding due to ancestry differences among sample individuals. Culverhouse et al. [2013] evaluated the performance of PCA when applied to the GAW18 pedigrees and investigated the structure that is reflected by the top principal components when all individuals are given equal weights in the analysis, and when weights are proportional to the inverse of the pedigree size, so that all families equally contribute to the genetic variation.

For PCA in nuclear families, Zhu et al. [2008] proposed a method for obtaining ancestry informative principal components (PCs) that performs a PCA using only the genotyped parents and then uses the weights of the SNPs from the PCA to obtain PCs for the offspring. Thornton et al. [2013] extended the Zhu et al. [2013] approach and developed the R-PCA method for ancestry informative PCA with general pedigrees. R-PCA performs a PCA using all pedigree founders who have SNP genotype data available so that the principal components are informative for ancestry. The SNP weights from the PCA with the pedigree founders are then used to compute PCs for all non-founders who are genotyped in the pedigree. The R-PCA method was applied to the GAW18 pedigrees, and the top principal components were compared to the top principal components from EIGENSTRAT and to global ancestry estimates from the previously described supervised ADMIXTURE analysis of Thornton et al. [2013].

Association and Admixture Mapping

Global and local ancestry estimates were used for complex trait association and admixture mapping by three group contributors. Thornton et al. [2013] performed an association analysis with a simulated DBP phenotype and SNP genotype data from the odd numbered chromosomes using the EMMAX method [Kang et al., 2010] that implements linear mixed model approach for association testing, accounting for both pedigree and global population substructures with an empirical covariance matrix. The authors compare their association analysis with EMMAX to a linear regression association analysis with the PLINK software [Purcell et al., 2007] where the top 10 PCs from a standard PCA were included as covariates.

Chen et al. [2013] used local ancestry estimates for admixture mapping with the real SBP and DBP phenotype data. Chen et al. [2013] performed the admixture mapping analysis using a linear regression model for the identification of SNPs with unusual deviations of European or Native American ancestry, relative to global ancestry estimates from these populations on chromosome 3. In addition to the admixture mapping analysis, Chen et al. [2013] also performed an association analysis with the blood pressure phenotypes, where a linear regression analysis was used to detect association between each of the SNPs and the SBP and DBP phenotypes, with local ancestry at a SNP included as a covariate to correct for population stratification.

Yorgov et al. [2013] performed association and admixture mapping of the real SBP and DBP phenotypes on chromosome 3 with the sequencing data. The following four nested linear regression models were considered, where all of the models included global ancestry on chromosome 3 as a covariate: model 1 (null model); model 2 included the SNP genotype as a predictor; model 3 included estimates of local European, Native American, and African ancestry at a SNP as predictors; and model 4 included both genotype and local ancestry at the SNP as predictors, which allows for the population specific genetic association effects at a SNP based on local ancestry. Wald and likelihood ratio tests were used to test association (model 2 versus model 1); admixture (model 3 versus model 1); association adjusted for admixture (model 4 versus model 3); and admixture and/or association (model 4 versus model 1).

Genotype Imputation in Admixed Populations

Imputation is often used to infer genotypes at untyped markers in association studies. Widely used genotype imputation algorithms, such as IMPUTE [Howie et al., 2009], MACH [Li et al., 2010], and fastPHASE [Scheet and Stephens, 2007], match genotype data at typed SNPs in a sample of individuals to haplotypes from suitable reference panels. Huang and Tseng [2013] examined the accuracy of genotype imputation in admixed population samples using different reference panels with the IMPUTE2 software [Howie et al., 2012]. A subset of 345 individuals with both WGS and SNP genotype data on chromosome 3 was chosen for the imputation analysis. SNPs on chromosome 3 that are represented in the 1000 Genomes Project (1kGP) [The 1000 Genomes Project Consortium, 2010] but are not typed in the GAW18 SNP data were imputed. The following subsets from the 1kGP were used as reference panels: (1) all 1,094 individuals; (2) 120 randomly selected

individuals; (3) 246 individuals with African ancestry; (4) 286 individuals with Asian ancestry; (5) 381 individuals with European ancestry; and (6) 181 individuals from the Americas (comprised of Colombian, Mexican, and Puerto Rican samples). A seventh reference panel consisting of 119 individuals from the GAW18 sample who were not included in the imputation study sample was also evaluated. Genotype imputation accuracy for the seven reference panels was assessed by comparing the concordance between imputed genotypes and genotypes from the WGS data.

RESULTS

Global Ancestry Estimation

The supervised ADMIXTURE ancestry analysis by Thornton et al. [2013] revealed that most of the GAW18 sample individuals have European and Native American derived ancestry, but with quite variable ancestry proportions. Proportional European ancestry ranged from 0% to 96%, with a 45% mean and a 14% standard deviation. Proportional Native American ancestry ranged from 0% to 84%, with a 50% mean and a 14% standard deviation. These proportions are consistent with previous reports for Mexican American samples [Thornton et al., 2012; Manichaikul et al. 2012]. The African and East Asian ancestry proportions in GAW18 were modest, with corresponding mean proportions of 4% and 1%, respectively. Figure 1 shows a plot of the clustering results from an unsupervised ADMIXTURE analysis of GAW 18, HapMap CEU, HapMap YRI, and HGDP Native American samples, assuming that there are three ancestral populations. The vast majority of GAW18 individuals fall between the HapMap CEU cluster and the HGDP Native American cluster but with proportional ancestry being quite variable. The individual ancestry results from the unsupervised analysis with reference population samples included are complementary to those provided by the supervised analysis of GAW18 with the ADMIXTURE software. Thornton et al. [2013] also found that an unsupervised ADMIXTURE analysis that did not include reference population samples performed poorly due to the relatedness in the sample, where proportional ancestry estimates were largely reflecting membership to the pedigrees contributing the largest groups of genotyped relatives in the sample.

Yorgov et al. [2013] used global ancestry proportions for chromosome 3 to evaluate local ancestry estimates produced over the set of all genotype individuals in GAW18. Proportional Native American, European, and African ancestries were reported to have means of 45%, 49%, and 6%, respectively, for chromosome 3, comparable to the global estimates from the supervised analysis using all odd autosomes given above.

Local Ancestry Estimation

Chen et al. [2013] found that LAMP-LD and MULTIMIX software programs, both of which use dense SNP genotype data and incorporate LD when estimating local ancestry, perform better than the LAMP software, which relies on a pre-defined set of AIMs that are in low LD. When averaging local ancestry based on LAMP, LAMP-LD, and MULTIMIX across chromosome 3, LAMP-LD estimates showed the highest correlation (0.989) with global ancestry estimates on chromosome 3 from a supervised ADMIXTURE analysis. LAMP-LD

and MULTIMIX resulted in discordant ancestry inferences for 18% of the SNPs with unphased genotype data for the sample individuals. The two methods inferred a similar number of generations since admixture for the Mexican American sample individuals, between 10 and 12, based on the number of inferred ancestry blocks, for which the authors found to be consistent with previous reports on Hispanic populations. However, when using phased genotype data for the GAW18 individuals with MULTIMIX, the number of generations since admixture was overestimated, which may be an artifact attributable to the uncertainty introduced by the genotype phasing of admixed samples.

Yorgov et al. [2013] reported that when using denser marker sets, LAMP-LD produced close to the number of ancestry blocks and global ancestry proportions that would be expected for Mexican populations. They also found that global ancestry estimates were stable when using sparser AIMs. However, there were too few ancestry blocks for the admixed sample individuals when using AIMs, even with appropriate adjustment of the parameters used in the ancestry analysis with the LAMP-LD software program.

Principal Components Analysis with Pedigrees

Culverhouse et al. [2013] found that when all individuals receive the same weights in the PCA analysis, the top three PCs separate members of three of the largest pedigrees from the other samples. For the PCA based on proportional weighting of families, the top two PCs separated four families from the rest of the sample, where one of pedigrees was also identified as an outlier in the unweighted PCA, and the three smallest pedigrees in GAW18 were the other outliers.

Thornton et al. [2013] also found that the top PCs largely reflected pedigree structure with a standard PCA, and that using the top PCs as surrogates for ancestry was not appropriate. The top PCs from R-PCA, however, reflected ancestry and not pedigree structure. Both Native American and European ancestry estimates from the supervised ADMIXTURE analysis have a correlation of 0.92 with the top PC from R-PCA. The variances explained by the top 10 PCs from R-PCA in a linear regression model for East Asian and African ancestry were reported as 0.74 and 0.69, respectively.

Association and Admixture Mapping

Thornton et al. [2013] found that the association analysis of the first simulated replicate of DBP across the odd autosomes using PLINK, where with the top PCs from a PCA with the EIGENSOFT software were included as covariates, lead to systematically inflated p-values. The genomic control inflation factor [Devlin and Roeder, 1999], λ , was 1.3 as a result of unaccounted population and pedigree structure. In contrast, $\lambda = 0.97$ for the association analysis with the EMMAX, indicating that the method is slightly conservative for the GAW18 data. EMMAX identified genome-wide significant associations for SNPs in the MAP4 gene on chromosome 3. This gene is causal for the simulated DBP phenotype.

Chen et al. [2013] found that an admixture mapping analysis for detecting association with the real DBP and SBP phenotypes and local ancestry was underpowered due to the small

number of unrelated individuals in their sample. There were no significant associations detected with DBP and SBP after adjustment for local ancestry.

With a combined test for admixture and association for the real DBP phenotype data and using an empirical significance threshold to adjust for multiple testing, Yorgov et al. [2013] identified a significant association with SNP rs12639065, located in an intergenic region between the LSM3 and SLC6A genes on chromosome 3. No significant SNPs were identified with the admixture mapping test, the association test, and the association test with adjustment for admixture for the DBP and SBP phenotypes. The authors additionally used simulated data sets for a trait not influenced by the genotype and verified that their method has the appropriate type I error rates. Yorgov et al. [2013] concluded from their analyses that combining admixture and association mapping signals is a promising approach for identifying variants for complex traits.

Genotype Imputation in Admixed Populations

Huang and Tseng [2013] identified the cosmopolitan reference panel containing all population samples from the 1kGP to be optimal, in terms of having both high genotype imputation accuracy and low missing genotype call rates, for genotype imputation in GAW18 with the IMPUTE2 software. They also found that a larger-sized reference panel can reduce imputation error and missing genotype, but the improvement can be limited. Indeed, when comparing the cosmopolitan reference panel consisting of all 1,094 1kGP samples to the panel consisting of 181 sample individuals from the Americas, genotype imputation error rates and missing genotype call rates were comparable. They also found that reference panels from 1kGP that did not include samples from the Americas resulted in substantially higher imputation error rates compared to the two reference panels that included these samples. Using reference panels from 1kGP consisting of single ancestral populations, e.g., the African, European, and Asian reference panels, resulted in poor genotype imputation quality for the admixed GAW18 samples. Interestingly, the reference panel consisting of admixed individuals from GAW18 that closely matched the ancestry of the sample individuals had higher imputation accuracy than all of the 1kGP reference panels considered, but this panel also resulted in higher missing genotype call rates.

DISCUSSION

The Admixture group members at GAW 18 considered a variety of topics for genetic analyses in admixed populations, including global and local ancestry inference, complex trait mapping, and genotype imputation. While the five contributions summarized here had different aims, a commonality of inference and consideration of ancestry in genetic analyses was identified.

Genotype data from appropriate reference population samples can improve ancestry inference in samples from admixed populations, and three contributions [Thornton et al., 2013; Chen et al., 2013; Yorgov et al., 2013] used populations samples from HapMap and HGDP as surrogates for European, African, Native American, and Asian ancestry for proportional ancestry estimation of the GAW18 sample individuals. Thornton et al. [2013] showed that in the absence of reference population samples, individual ancestry estimates

with the ADMIXTURE software can be seriously confounded in the presence of relatedness, but that reliable estimates can be obtained in related admixed samples when appropriate surrogates for ancestry are included in the analysis.

Local ancestry can be estimated using AIMs or high density SNPs, and both Chen et al. [2013] and Yorgov et al. [2013] compared inference on local ancestry when using the two types of marker sets. Chen et al. found that the LAMP-LD software, which models the LD of dense SNP sets, outperforms the LAMP method that relies on low-LD AIMs. Yorgov et al. [2013] reported that their local ancestry analysis using LAMP-LD with high-density SNPs produced close to the expected number of ancestry blocks that would be expected for Mexican populations, while the LAMP-LD analysis with a sparse set of AIMs produced too few ancestry blocks. Chen et al. [2013] also compared local ancestry estimates when using phased versus unphased genotypes with the MULTIMIX program, and found that estimates for local ancestry were more reliable when using unphased genotypes. This finding could be attributed to phasing uncertainty in admixed samples that is not accounted for in local ancestry estimation.

Two of the contributions applied PCA to the pedigrees in order to examine population structure and found that the top principal components reflected family structure, suggesting that the widely used PCA approach for population structure inference may not be appropriate for related samples [Thornton et al. 2013; Culverhouse et al.; 2013]. Thornton et al. [2013] proposed the R-PCA method for ancestry informative PCA with general pedigrees, and demonstrated that the top PCs from this method were highly correlated with estimated proportional ancestry from a supervised analysis with the ADMIXTURE software.

For complex-trait mapping in unrelated GAW18 individuals, Chen et al. [2013] reported that admixture mapping was underpowered, and Yorgov et al. [2013] found that simultaneously testing for admixture and association, while allowing for heterogeneous genetic effects, may improve the power to detect causative variants compared to association testing or admixture mapping alone. For association testing within admixed pedigree samples, Thornton et al. [2013] demonstrated that the EIGENSTRAT approach of incorporating the top 10 principal components in a linear regression model does not appropriately account for both population and pedigree structure in a sample. In contrast, the EMMAX method, a linear mixed-effects model approach that uses an empirical covariance matrix, was appropriately calibrated in this setting.

Huang and Tseng [2013] compared the genotype imputation results for the GAW18 admixed samples with the IMPUTE2 software using different reference panels. They identified the cosmopolitan reference panel consisting of all sample individuals in 1kGP to be optimal, where imputation with this reference panel resulted in both high genotype imputation accuracy and low missing genotype call rates for the GAW18 samples. They also found that using a reference panel consisting of GAW18 individuals that closely matched the admixed ancestry of the sample individuals had the highest genotype imputation accuracy among all reference panels considered, but it also resulted in higher missing genotype call rates compared to the composite 1kGP reference panel. A number of factors can impact genotype imputation quality in admixed populations [Hancock et al., 2012], in

addition to the choice of reference panels, and a more extensive study investigating genotype imputation accuracy of rare versus common variants in GAW18 as well as comparing the imputation performance of different software programs will provide greater insight.

In summary, the general theme of the contributions from the Admixture Group was inferring, accounting for, and incorporating ancestry admixture. Despite the diversity of methods and applications considered, a common finding was that genetic analyses in admixed populations substantially benefit from appropriate proxies for the ancestral populations. This applies to both local and global ancestry estimation, admixture mapping, and genotype imputation. The sparsity of adequate reference population samples - for example, specific reference panels for Native American subpopulations - however, is currently a limitation in genetic analyses of populations with admixed ancestry. As public resources of genotype and sequence data from diverse ancestral populations expands in parallel with continued development of powerful and computationally efficient statistical methods, the future of genetic studies in admixed populations is promising.

Acknowledgments

The authors thank the GAW18 Admixture Group members for their contributions and assistance; Dr. Andrew Patterson, for discussion and critical comments; and two anonymous reviewers, for helpful comments. This study was supported in part by the National Institutes of Health (NIH) grant K01 CA148958 (to T.T.). The Genetic Analysis Workshop 18 was supported by NIH grant R01 GM031575.

References

- Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*. 2009; 19:1655–1664. [PubMed: 19648217]
- Baran Y, Pasaniuc B, Sankararaman S, Torgerson GD, Gignoux C, et al. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*. 28:1359–1367. [PubMed: 22495753]
- Chen M, Yang C, Li C, Hou L, Chen X, Zhao H. Admixture Mapping Analysis in the context of GWAS with GAW18 data. *BMC Proceedings*. 2013 in press.
- Churchhouse C, Marchini J. Multi-way admixture deconvolution using phased or unphased ancestral panels. *Genet Epidemiol*. 2013 in press.
- Culverhouse RC, Hinrichs AL, Suarez BK. Identifying cryptic population structure in multi-generation pedigrees in a Mexican American sample. *BMC Proceedings*. 2013 in press.
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999; 55:997–1004. [PubMed: 11315092]
- Genovese G, Handsaker RE, Li H, Altemose N, Lindgren AM, Chambert K, Pasaniuc B, Price AL, Reich D, Morton CC, Pollak MR, Wilson JG, McCarroll SA. Using population admixture to help complete maps of the human genome. *Nat Genet*. 2013; 45:406–414. [PubMed: 23435088]
- Hancock DB, Levy JL, Gaddis NC, Bierut LJ, Saccone NL, Page GP, Johnson EO. Assessment of genotype imputation performance using 1000 Genomes in African American studies. *PLoS One*. 2012; 7:e50610. [PubMed: 23226329]
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009; 5:e1000529. [PubMed: 19543373]
- Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*. 2012; 44:955–959. [PubMed: 22820512]

- Huang GH, Tseng YC. Genotype imputation accuracy with different reference panels in admixed populations. *BMC Proceedings*. 2013 in press.
- International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010; 467:52–58. [PubMed: 20811451]
- Johnson NA, Coram MA, Shriver MD, Romieu I, Barsh GS, London SJ, Tang H. Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genetics*. 2011; 7:e1002410. [PubMed: 22194699]
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*. 2010; 42:348–354. [PubMed: 20208533]
- Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. 2008; 319:1100–1104. [PubMed: 18292342]
- Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*. 2010; 34:816–834. [PubMed: 21058334]
- Manichaikul A, Palmas W, Rodriguez CJ, Peralta CA, Divers J, Guo X, Chen WM, Wong Q, Williams K, Kerr KF, Taylor KD, Tsai MY, Goodarzi MO, Sale MM, Diez-Roux AV, Rich SS, Rotter JJ, Mychaleckyj JC. Population structure of hispanics in the United States: the Multi-Ethnic Study of Atherosclerosis. *PLoS Genetics*. 2012; 8:e1002640. [PubMed: 22511882]
- Price AL, Patterson NH, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006; 38:904–909. [PubMed: 16862161]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet*. 2007; 81:559–575. [PubMed: 17701901]
- Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*. 2006; 78:629–644. [PubMed: 16532393]
- Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating Local Ancestry in Admixed Populations. *Am J Hum Genet*. 2008; 82:290–303. [PubMed: 18252211]
- Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: Analytical and study design considerations. *Genet Epidemiol*. 2005; 28:289–301. [PubMed: 15712363]
- The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467:1061–1073. [PubMed: 20981092]
- Thornton T, Conomos M, Sverdlov S, Marchani EE, Cheung C, Glazner C, Lewis S, Wijsman EM. Estimating and adjusting for ancestry admixture in statistical methods for relatedness inference, heritability estimation, and association testing. *BMC Proceedings*. 2013 in press.
- Yorgov D, Edwards KL, Santorico SA. Use of admixture and association for detection of quantitative trait loci in the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples (T2D-GENES) study. *BMC Proceedings*. 2013 in press.
- Zhu X, Li S, Cooper RS, Elston RC. A unified association analysis approach for family and unrelated samples correcting for stratification. *Am J Hum Genet*. 2008; 82:352–365. [PubMed: 18252216]

Figure 1. Individual-Ancestry Clustering Results for GAW18, HapMap CEU, HapMap YRI, and HGDP Native American samples

Estimates for proportional ancestry were calculated from an unsupervised structure analysis with the ADMIXTURE software program assuming three populations. Each point shows the mean estimated ancestry for an individual. For a given individual, proportional ancestry values from the three populations are given by the distances to each of the three sides of the equilateral triangle. The HGDP Native Americans are the samples from the Americas (Surui, Maya, Karitiana, Pima, and Colombian samples).

Table I**Fundamental concepts in association analyses of admixed populations**

Ancestry estimation: The genetic origin of the individuals in a population can be estimated by using genetic markers. For example, ADMIXTURE is a software tool for maximum likelihood estimation of individual ancestries relying on multiple single nucleotide polymorphisms. The ancestry estimation is called supervised when it considers the genotypes of individuals with known ancestry. The ancestry estimation is called unsupervised when no individuals with known ancestry are included.

Ancestry informative marker (AIM): a characteristic, usually a genetic marker, which shows strong differences among populations.

Cryptic population substructure: Population substructure is often hidden. Genetic studies should routinely examine cryptic population substructure since it can inflate the variance of statistics of genetic association.

Genetic principal component analysis: a method to examine the cryptic relatedness of genotypes assumed to be uncorrelated. A variance-covariance matrix of genetic similarity is built based on individual genotypes, and then investigated using principal component analysis.

Hardy-Weinberg equilibrium (HWE): A population is said to be in HWE if the observed genotype frequencies are in agreement with the allele frequencies. Ancestry admixture may reduce the proportion of heterozygous genotypes in a population and originate departures from HWE, but the relationship between genotype and allele frequencies is also distorted by non-random mating, inbreeding and selection, and testing for HWE has low statistical power.

Imputation of genotypes: The linkage disequilibrium patterns in the human genome can be exploited to estimate unknown genotypes. Usually, a panel of reference with densely genotyped individuals is combined with a study where fewer variants have been genotyped. After phasing genotypes in the reference panel and in the study sample, haplotypes are compared to predict (impute) unobserved genotypes. A reference panel that resembles the ancestry of the study sample can increase imputation accuracy.

Table II

Overview of GAW18 Admixture Group contributions

Contribution	Aims	Sample	Genetic Data	Chromosomes Analyzed	Phenotypes	Analyses
Chen et al. [2013]	Complex Trait Mapping, Proportional Ancestry Estimation	Subset of Unrelated Individuals	SNPs	Chromosome 3	Diastolic and Systolic Blood Pressure	Local and Global Ancestry Estimation, Admixture Mapping, Association Testing
Culverhouse et al. [2013]	Population Structure Inference	Pedigrees, Subset of Unrelated Individuals	SNPs	Odd Numbered Autosomes	-	Principal Components Analysis
Huang and Tseng [2013]	Evaluation of Genotype Imputation Accuracy with Different Reference Panels	Subset of Related and Unrelated Individuals	Sequences and SNPs	Chromosome 3	-	Genotype Imputation
Thornton et al. [2013]	Complex Trait Mapping, Proportional Ancestry Estimation, Population Structure Inference	Pedigrees, Subset of Unrelated Individuals	SNPs	Odd Numbered Autosomes	Simulated Diastolic Blood Pressure	Global Ancestry Estimation, Principal Components Analysis, Association Testing
Yorgov et al. [2013]	Complex Trait Mapping, Proportional Ancestry Estimation	Pedigrees, Subset of Unrelated Individuals	Sequences	Chromosome 3	Diastolic and Systolic Blood Pressure	Local Ancestry Estimation, Admixture Mapping, Association Testing