VCFLIB: an ensemble of methods for variant manipulation and population genetics

Zev N. Kronenberg[1,☯], Erik Garrison[2,☯], Mark Yandell[3,4], Mike Shapiro[5,3], Gabor Marth[3,4], Richard Durbin[1], Evan E. Eichler[1,*], with the VCFLIB Consortium[¶]

1 Department of Genome Sciences, University of Washington, Seattle, WA, USA
2 Wellcome Trust Sanger Institute, Cambridge, UK
3 Department of Human Genetics, University of Utah, Salt Lake City, UT, USA
4 Ustar Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA
5 Department of Biology, University of Utah, Salt Lake City, UT, USA

☯These authors contributed equally to this work.
‡These authors also contributed equally to this work.
¤a Insert current address of first author with an address update
†Deceased
¶Membership list can be found in the Acknowledgments section.
*** zevk@uw.edu**

Abstract

Introduction

Genetic variation is commonly represented in the Variant Call Format (VCF) [3]. VCF files are flat-text, human-readable, and provide a flexible way to annotate genetic information. When compressed, VCF files can be indexed and queried by genomic position. One or more genomes can be defined in the VCF genotype fields, which is important for population-level sequencing experiments. For these reasons, VCF has been widely adopted for large-scale genetic studies including the The 1000 Genomes Project [1], ExAC [4], GoNL [2].

To the uninitiated, the VCF format, while easy to comprehend can be difficult to parse and manipulate. There is an ongoing need for methods that can facilitate rapid genetic analyses of VCF data. VCFLIB, VCFTools [3], GATK, and BioConductor are few software packages that are commonly used, each fulfilling a niche.

This article is milepost for VCFLIB development and showcases the population genetic analyses that can be done with the toolkit. In the Design and implementation section we described the three attributes of VCFLIB: a C++ library, an ensemble of programs for modifying VCF files, and a toolkit for population genetics (The Genotype Phenotype Association Toolkit: GPAT). In the results section we apply VCFLIB to the 1000 Genomes Project data to quantify patterns of genetic variation within and between populations. Lastly, we briefly discuss our goals for future VCFLIB development.

Design and Implementation

The VCFLIB library

ERIK G.

The Genotype Phenotype Association Toolkit (GPAT)

The genotype-phenotype association toolkit was originally designed to map the genetic basis of phenotypic variation, but it has grown into a fully functional population genetic toolkit. The GPAT association test, pFst, has been successfully used to identify genetic variants associated with phenotypic variation in domestic pigeons [12], and *Tetrahymena* [8]. GPAT's descriptive statistics have also been applied to human [13] and pigeon [7] populations.

Fst methods

Weir and Cockerhams estimator of F_{ST} is implemented in GPAT's "wcFst" [9]. To reduce spurious F_{ST} signals, sites with less than five individuals in either population are not scored. Values of F_{ST} will range from slightly negative to one; negative values can be treated as zero.

An independent bayesian method of F_{ST} is provided in GPAT [**?**]. The method, implemented in "bFst" provides a way to add confidence intervals to F_{ST} point estimates. The posterior distribution is calculated using MCMC over 50,000 iterations per site. For this reason, "bFst" is only well suited for small genomic regions.

Demining the start and end of a Fst peak, genomic island, is a subjective and difficult task. To standardize this procedure "segmentFst" was implemented in GPAT. The heuristic scans the output of "wcFst" with a ten SNV window measuring the number of Fst values that are above a user-defined threshold. If the low-high Fst ratio is above two the window is recursively extended. This heuristic, unlike a sliding window, does not suffer from the goldilocks principle.

Permutation provides a way to measure the empirical probability of a genomic island. The average F_{ST} from a window defined by ?smoother? or "segmentFst" can be quantified against other contiguous windows with the same number SNVs using "permuteGPATwindow." The program adds the number of permutations and the number of matched windows with a higher average FST value.

Δ allele frequency (AF) and association testing

For simple phenotypic traits quantifying the Δ allele frequency between individuals with the phenotype and those without can be sufficient for genotype-phenotype association. A likelihood test for Δ AF is implemented in "pFst" for both pooled and genotypic data Eq (1). In Eq (1), B represented the binomial distribution parameterized by the number of trials n, the number of successes n and the probability of success, p. The D statistic is chi-squared distributed with one degree of freedom.

$$D = -2 * ln(\frac{B(n_c, k_c, p_c)}{B(n_t, k_t, p_t) * B(n_b, k_b, p_b)}) \quad (1)$$

For "pFst", n corresponds to the number of callable alleles, k is the number on non-reference alleles, and p is the bounded allele frequency (0.00001-0.99999). The subscripts denotes the group membership for the target (t), background (b) and both combined (c). If genotype likehoods are provided pFst weights each allele proportionally to the genotype likelihood, rather than their direct count. For pooled datasets, with more than one biological replicate in the target and background, the model substitutes the binomial distribution for the beta. The methods of moments used to estimate α, and

β, the two parameters of the beta distribution. The likelihood ratio-test implemented in "pFst" has been described several times [10, 11], but few practical implementations exist.

Haplotype methods (for phased variants)

GPAT provides several methods for quantifying haplotype structure at a locus. The extended haplotype homozygosity score measures the haplotypic diversity; a value of zero means all haplotypes are unique and a value of one means all haplotypes are identical. The "sequenceDiversity" program quantifies Eq (2) across a fixed number of SNVs (user defined), where n is the number of chromosomes (2x the number of genotypes), i is the index for each unique haplotype in a fixed window and x_i is the number of i haplotypes.

$$EHH = \frac{\sum_i \binom{x_i}{2}}{\binom{n}{2}} \tag{2}$$

$$EHH_c = \frac{\sum_{ic} \binom{x_{ic}}{2}}{\binom{n_c}{2}} \tag{3}$$

The integrated haplotype score (iHS) measures the relative decay of EHH_c between the alternative and reference core haplotypes, Eq (4). At a single SNV (core haplotype), EHH_c equals one and n_c is the number of reference or alternative alleles. The trapezoid rule is used to integrate EHH_c with respect to genetic distance. GPAT's implementation of iHS uses Plink formatted genetic maps(cite). If no genetic map is provided a fixed value will be assumed. GPAT also provides a tool ("normalize-iHS") to normalize iHS by mean and standard deviation, binned by allele frequency. Comparison of GPAT iHS and other tools can be seen in Fig S1 [5] [6].

$$iHS = ln(\frac{\int EHH_a}{\int \mathrm{EHH}_r}) \tag{4}$$

Figure 1. Using VCFLIB to find population stratified loci in the One Thousand Genomes Project data. The Fst scatter plots (A-C) show genomic position on the x-axis and Weir and Cockerham's Fst on the y-axis. The vertical bands delineate regions of high Fst defined by segmentFst. The color of the bands denotes the empirical significance of the region compared to the rest of the genome determined with permuteSmooth. A: EDAR is outlier in the CEU-CHB comparison because EDAR is under selection in CHB. B: OCA2 and HERC2 are outliers in the CEU-CHB comparison because they are under selection in CEU. The segmentation of the region is broken by segmental duplications between 28-29 Mb. C: KCNQ5 is an outlier in the CEU-YRI comparison. D: KCNQ5 shows decreased heterozygosity at it's 3' (73.9Mb) end in CEU (red) compared with YRI (blue). E: The GPAT workflows used for the Fst analyses.

Results

As a demonstration of the power and flexibility of VCFLIB we performed genome-wide selection scans on the Phase III One Thousand Genomes Project data. To discover genes that may have been targets of natural selection we applied within and between population metrics. We measured population stratification between Northern Europeans (CEU), Southern Han Chinese (CHB), and Yoruba (YRI) using Weir and Cockerham's Fst. To detect within-population selection we applied the integrated haplotype score (iHS). These analyses serve as a good control for VCFLIB methods as

Figure 2. Using VCFLIB to identify patterns of haplotype diversity consistent with natural selection. A: The GPAT workflows used for the haplotype analysis. B: Genome-wide iHS Manhattan plot for Norther Europeans (CEU). The x-axis is an index and the y-axis is the average absolute iHS within a 100Kbs sliding window. The repeating color palette delineates chromosomes. Several iHS peaks are annotated with the closes gene or a dash if there was not a gene nearby. The window that overlaps lactase has the highest genome-wide average iHS. C: The lactase haplotypes present in CEU and YRI. Each row is a single haplotype and each column is a position where there is a non-reference allele (red). Fewer unique haplotypes can be seen in CEU compared to YRI. The trapezoid denotes where the haplotppe plots are in relation to lactase. D: The Extended Haplotype Homozygosity (EHH) decay for rs3754686(A/G). The position is shown on the x-axis and EHH for the derived (orange) and ancestral allele (green) is shown on the y-axis.

many studies have perviously identified genes under selection in CEU, CHB, and YRI. The VCLIB workflows used are shown in Fig. 1A and Fig. 2E.

The three pairwise Fst analyses resulted in over 30 million scored genomic positions, which is prohibitively large for manual outlier detection. We applied our segmentation algorithm (implemented in segmentFst) to grow regions that had 10 Fst values greater than 0.6. To control for the number of observations in each region and to determine empirical outliers we ran a permutation test for one million iterations (implemented in permuteSmooth)(supporting table XXX). There were 3,790, 1,826, and 665 regions with high Fst values for the CHB-YRI, CEU-YRI and CEU-CHB comparisons. These regions overlapped XXX genes in CEU-CHB, XXX genes in CEU-YRI, and XXX genes in CHB-YRI. Classic examples of stratified loci including *EDAR* and *OCA2/HERC2* can be seen in Fig. 1A-B [cite,cite].

Segmentation of iHS data resulted in 1500, 1600, and 1937 regions in CHB, CEU, and YRI (segmentation threshold : 3). For all three analyses a total of 1,885 unique genes were discovered. In the CEU population lactase (LCT) was amounts the highest scored regions shown in Fig. 2B. The haplotype diversity around LCT is noticeably different (Fig. 2C-D).

Availability and Future Directions

VCFLIB is publicly available at (`https://github.com/ekg/vcflib`VCFLIB) and additional documentation can be found at `https://github.com/zeeev/vcflib/wiki`.

Areas of continued development include extending documentation, unit testing, support for new VCF specifications, and additional population genetic metrics.

Supporting Information

S1 Snakemake

Fst analyses. The snakemake file for the CEU-CHB Fst analysis.

S2 Snakemake

iHS analyses. An archive of the snakemake file, the config file and region file used to run the CEU iHS analyses.

S1 Fig

Comparison of iHS methods. Scatter plots comparing VCFLIB's implementation of iHS to SELSCAN or Prichard's iHS. The iHS data are from a two megabase window around *LCT* in CEU.

S1 Table

Regions with high Fst values. The CEU-YRI, CEU-CHB, and CHB-YRI regions are in three different excel sheets. The region in GRCh37 coordinates, the average Fst values, and the empirical significance are listed as column headers.

Acknowledgments

We would like to thank the countless number of people who have contributed suggestions, created bug reports, and modified the VCFLIB codebase.

The VCFLIB Consortium

EJ Osborne[1], Brett Kennedy[1], Daniel Ence[1], Travis Collier[1], EJ Osborne[1], ...

References

1. Sudmant, Peter H., et al. "An integrated map of structural variation in 2,504 human genomes." Nature 526.7571 (2015): 75-81.
2. Genome of the Netherlands Consortium. "Whole-genome sequence variation, population structure and demographic history of the Dutch population." Nature Genetics 46.8 (2014): 818-825.
3. Danecek, Petr, et al. "The variant call format and VCFtools." Bioinformatics 27.15 (2011): 2156-2158.
4. Exome Aggregation Consortium (ExAC), Cambridge, MA, http://exac.broadinstitute.org/, Feb 2016.
5. Szpiech, Zachary A., and Ryan D. Hernandez. "selscan: an efficient multithreaded program to perform EHH-based scans for positive selection." Molecular biology and evolution 31.10 (2014): 2824-2827.
6. Voight, Benjamin F., et al. "A map of recent positive selection in the human genome." PLoS Biol 4.3 (2006): e72.
7. Shapiro, Michael D., et al. "Genomic diversity and evolution of the head crest in the rock pigeon." Science 339.6123 (2013): 1063-1067.
8. Galati, Domenico F., et al. "DisAp-dependent striated fiber elongation is required to organize ciliary arrays." The Journal of cell biology 207.6 (2014): 705-715.

   bfst Holsinger, Kent E., Paul O. Lewis, and Dipak K. Dey. "A Bayesian approach to inferring population structure from dominant markers." Molecular Ecology 11.7 (2002): 1157-1164.
9. Weir, Bruce S., and C. Clark Cockerham. "Estimating F-statistics for the analysis of population structure." evolution (1984): 1358-1370.

10. Kim, Su Yeon, et al. “Design of association studies with pooled or un?pooled next?generation sequencing data.” Genetic epidemiology 34.5 (2010): 479-491.

11. Li, Heng. “A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.” Bioinformatics 27.21 (2011): 2987-2993.

12. Domyan, Eric T., et al. ”Epistatic and combinatorial effects of pigmentary gene mutations in the domestic pigeon.” Current Biology 24.4 (2014): 459-464.

13. Barber, Matthew F., and Nels C. Elde. ”Escape from bacterial iron piracy through rapid evolution of transferrin.” Science 346.6215 (2014): 1362-1366.