Bioinformatics, 32(14), 2016, 2224–2226
doi: 10.1093/bioinformatics/btw147
Advance Access Publication Date: 18 March 2016
Applications Note

OXFORD

Genetics and population analysis

Ferret: a user-friendly Java tool to extract data from the 1000 Genomes Project

Sophie Limou[1,*], Andrew M. Taverner[1,2] and Cheryl A. Winkler[1]

[1]Molecular Genetic Epidemiology Section, Basic Research Laboratory, Basic Science Program, NCI, Leidos Biomedical Research, Inc., Frederick National Laboratory, Frederick, MD 21702, USA and [2]Quantitative and Computational Biology program, Princeton University, Princeton, NJ 08544, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on August 14, 2015; revised on February 19, 2016; accepted on March 14, 2016

Abstract

Summary: The 1000 Genomes (1KG) Project provides a near-comprehensive resource on human genetic variation in worldwide reference populations. 1KG variants can be accessed through a browser and through the raw and annotated data that are regularly released on an ftp server. We developed Ferret, a user-friendly Java tool, to easily extract genetic variation information from these large and complex data files. From a locus, gene(s) or SNP(s) of interest, Ferret retrieves genotype data for 1KG SNPs and indels, and computes allelic frequencies for 1KG populations and optionally, for the Exome Sequencing Project populations. By converting the 1KG data into files that can be imported into popular pre-existing tools (e.g. PLINK and HaploView), Ferret offers a straightforward way, even for non-bioinformatics specialists, to manipulate, explore and merge 1KG data with the user's dataset, as well as visualize linkage disequilibrium pattern, infer haplotypes and design tagSNPs.

Availability and implementation: Ferret tool and source code are publicly available at http://limousophie35.github.io/Ferret/.

Contact: ferret@nih.gov

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

1 Motivation

The 1000 Genomes (1KG) Project is an international consortium to establish a human genome map of genetic variations (SNPs, indels and CNVs) with a minor allele frequency as low as 0.1–0.5% in the coding regions and 1% in the rest of the genome in multiple reference populations (1000 Genomes Project Consortium *et al.*, 2012). This publicly available, near-comprehensive catalogue of human genome diversity contributes to improve the human reference sequence, and furthers understanding of the evolutionary history of populations and of the genetic contribution to diseases and traits. The 1KG variants can be directly queried through the 1KG browser, however the process to access some information can become tedious, requiring a good knowledge of the website architecture and many clicks. Raw and annotated data are regularly released on the 1KG ftp server, but manipulating these large data files requires advanced bioinformatics skills. Finally, 1KG offers the *VCF to PED converter* to transform SNP genotype data of a locus from the 1KG .vcf files into files that can be loaded into the HaploView visualization tool (Barrett *et al.*, 2005).

We developed Ferret as a user-friendly tool to easily extract genotype data from 1KG SNPs and indels into files that can be imported into popular pre-existing tools, while also computing allelic frequencies. Ferret unique features are to: (i) handle SNP and indel genotype data; (ii) extract data from a locus, gene(s) or SNP(s) of interest; (iii) calculate SNP, indel and CNV allele frequencies in 1KG and Exome Sequencing Project (Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, http://evs.gs.washington.edu/EVS/) (Dorschner *et al.*, 2013) populations; (iv) create PLINK (Purcell *et al.*, 2007) and HaploView output files; and (v) provide a user-friendly tool for the non-bioinformatics

specialists. Ferret is therefore a straightforward program that permits easy manipulation, exploration and visualization of 1KG genotype data with well-known pre-existing tools.

2 Implementation and Java interface

We developed a Java algorithm to extract 1KG SNP and indel genotype data and to compute SNP, indel and CNV allelic frequencies from a user-friendly interface (Fig. 1 and Supplementary Fig. S1). The user has to specify the following inputs:

(1) Genetic region(s):

a. **Locus** defined by chromosome number, start and stop positions.
b. **Gene(s)** name or ID (e.g. CCR5 or 1234 for the chemokine C-C motif receptor 5 gene) can be entered manually in the search box separated by a comma, or a file containing a list of genes can be uploaded (.csv, .tab, .tsv and .txt file formats accepted).
c. **SNP(s)** rs number (e.g. 562091107 for the CCR5 rs562091107 32-bp deletion) can be entered manually in the search box separated by a comma, or a file containing a list of SNP rs numbers can be uploaded (.csv, .tab, .tsv and .txt file formats accepted). Optionally, the user can request to include surrounding variants in a certain base-pair window range around the SNP(s) of interest.

(2) 1KG population(s): it can be any combination of the 26 1KG reference populations.

Optionally, the user can request to output the allele frequencies in African and European American populations from the Exome Sequencing Project (ESP). Finally, the user can pick the name and location of the files that will be generated by Ferret.

If the genetic region input is not a locus, Ferret retrieves the chromosome coordinates of the gene(s) or SNP(s) from NCBI (NCBI Resource Coordinators, 2014) Entrez Gene E-utilities (Maglott *et al.*, 2011) or dbSNP (Sherry *et al.*, 2001), respectively.

From the chromosome coordinates, Ferret extracts the 1KG genotype data using tabix (Li, 2011), parses the .vcf subset file to create .map, .ped and .info files (for file format description, see (Barrett *et al.*, 2005; Purcell *et al.*, 2007)), and computes allelic frequencies in the selected 1KG and ESP populations, which are recapitulated in the .frq file.

Fig. 1. Ferret screenshot

In compliance with the .map/.ped/.info file formats, we exclude multi-allelic variants and CNVs. In order to include small biallelic indel variants, which are not supported either by these file formats, we artificially transform them into A/T SNPs while coding the allele correspondence in the variant ID so they can be handled by pre-existing tools. Our variant ID reports the rs number when available (e.g. rs199824195), and the chromosome position otherwise (e.g. chr3_46414861). For small indels, the variant ID also reports the alleles (e.g. indel_rs136158_AT/A or indel_chr3_46415285_GC/G) and Ferret converts the genotype outputs by coding as 'A' the first allele and as 'T' the second allele (e.g. AT converted into A and A into T for a AT/A indel, and GC converted into A and G into T for a GC/G indel).

By default, Ferret extracts information from 1KG Phase 3 (ftp://ftp.1000genomes.ebi.ac.uk/Vol03214/ftp/release/20130502/) using the hg19/GRCh37 human genome coordinates. Alternatively, the user can opt for the 1KG Phase 1 version or for the hg38/GRCh38 human genome version. Ferret can also output allelic frequencies without returning the genotype data, and the user can set a frequency threshold for variants to output. Finally, more advanced users can export data into a .vcf file.

3 Performance and example of usage

We have tested Ferret on general examples and monitored runtime and output file size (see Supplementary Information). Ferret takes approximately 3–4 min to generate files for a 100 kb locus or for 10 SNPs, and 35min for a 1Mb locus or for 100 SNPs (Supplementary Figs S2 and S3). Runtimes are much more variable for genes (between 16 and 55 min for 10 genes), as gene size and variant density can be highly variable (Supplementary Table S1). The reference population(s) size is the main factor impacting output file size (Supplementary Fig. S4): for a 100 kb locus, 1 MB disc space would be occupied for the CEU individuals ($n = 99$) while 28 MB would be required for all 1KG populations ($n = 2504$); for a 1Mb locus, 14 and 311 MB disc space would be necessary for CEU and all individuals, respectively.

After retrieving the genotype information through Ferret, the user can easily manipulate the .map/.ped data with a pre-existing tool such as PLINK to be merged with the user's dataset, to perform association or population analyses, to infer haplotypes and impute missing genotypes, etc. The .ped/.info data can be imported into a pre-existing visualization tool such as HaploView to access linkage disequilibrium pattern and metrics, haplotype estimation, tagSNP design for customized genotyping arrays, or any other applications. The .vcf file can be merged with the user's dataset, or used to filter out specific variants using VCFtools (Danecek *et al.*, 2011), or to annotate variants (e.g. gene location, functional prediction, GWAS catalog) using Annovar (Wang *et al.*, 2010).

Funding: This work has been funded in whole or in part with federal funds from the Frederick National Laboratory for Cancer Research, National Institutes of Health [contract HHSN26120080001E] and by the Intramural Research Programs of Frederick National Laboratory, Center for Cancer Research, National Institutes of Health. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

Acknowledgements

The authors thank Anna Purtscher and Uma Mudnuri for their technical assistance in developing some parts of the code. We also would like to acknowledge Victor David, George Nelson and Nicolas Vince for feedback.

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium *et al*. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Barrett,J.C. *et al*. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Danecek,P. *et al*. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Dorschner,M.O. *et al*. (2013) Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am. J. Hum. Genet.*, **93**, 631–640.

Li,H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.

Maglott,D. *et al*. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.

NCBI Resource Coordinators (2014) Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res.*, **42**, D7–D17.

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

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Wang,K. *et al*. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.