Bioinformatics, 2017, 1–3
doi: 10.1093/bioinformatics/btx165
Advance Access Publication Date: 27 March 2017
Applications Note

OXFORD

Genetics and population analysis

EthSEQ: ethnicity annotation from whole exome sequencing data

Alessandro Romanel[1,*], Tuo Zhang[2,3], Olivier Elemento[2,4] and Francesca Demichelis[1,4,*]

[1]CIBIO, University of Trento, Trento, Italy, [2]Caryl and Israel Englander Institute for Precision Medicine, New York Presbyterian Hospital-Weill Cornell Medicine, New York, NY, USA, [3]Genomics Core Facility and [4]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on November 5, 2016; revised on February 25, 2017; editorial decision on March 20, 2017; accepted on March 24, 2017

Abstract

Summary: Whole exome sequencing (WES) is widely utilized both in translational cancer genomics studies and in the setting of precision medicine. Stratification of individual's ethnicity is fundamental for the correct interpretation of personal genomic variation impact. We implemented EthSEQ to provide reliable and rapid ethnicity annotation from whole exome sequencing individual's data, validated it on 1000 Genome Project and TCGA data (2700 samples) demonstrating high precision, and finally assessed computational performances compared to other tools. EthSEQ can be integrated into any WES based processing pipeline and exploits multi-core capabilities.
Availability and Implementation: R package available at github.com/aromanel/EthSEQ and CRAN repository.
Contact: alessandro.romanel@unitn.it or f.demichelis@unitn.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Interrogation of the entire coding genome for germline and somatic variations through Whole Exome Sequencing (WES) is rapidly becoming a preferred approach for the exploration of large cohorts (such as The Cancer Genome Atlas initiative) especially in the context of precision medicine programs (Beltran *et al.*, 2015). In this setting, the estimation of individual's ethnical background is fundamental for the correct interpretation of variant association studies and of personal genomic variations importance (Petrovski and Goldstein, 2016; Price *et al.*, 2006; Spratt *et al.*, 2016; Zhang *et al.*, 2016). To enable effective annotation of individual's ethnicity and improve downstream analysis and interpretation of germline and somatic variations, we developed EthSEQ, a tool that implements a rapid and reliable pipeline for ethnicity annotation from WES data. The tool can be used to annotate ethnicity of individuals with germline WES data available and can be integrated in any WES-based processing pipeline. EthSEQ also exploits multi-core technologies when available.

2 Approach

EthSEQ provides an automated pipeline, implemented as R package, to annotate the ethnicity of individuals from WES data inspecting differential SNPs genotype profiles while exploiting variants covered by the specific assay. As input the tool requires genotype data at SNPs positions for a set of individuals with known ethnicity (the *reference model*) and either a list of BAM files or genotype data of individuals with unknown ethnicity. EthSEQ then annotates the ethnicity of each individual using an automated procedure (Supplementary Fig. S1a) and returns detailed information about individual's inferred ethnicity, including aggregated visual reports.

The *reference model* builds on genotype data of individuals with known ethnicity; 1000 Genome Project individuals data is here used to construct platform-specific reference models relying on the most conserved ethnic groups EUR (Caucasian), AFR (African), EAS (East Asian) and SAS (South Asian) for multiple WES designs: Agilent HaloPlex, Agilent SureSelect and Roche Nimblegen (**Supplementary**

Methods). More generally, given a set of genomic regions and genotype data of a set of individuals annotated for ethnicity, a procedure to automatically generate a reference model is also provided by EthSEQ. The *target model* is created either from the input list of individual's germline BAM files that are genotyped at all reference model's positions using the genotyping module of ASEQ (Romanel *et al.*, 2015) (depth of coverage $\geq$ 10X and read/base mapping qualities $\geq$ 20 here required by default to guarantee confident genotype calls) or from genotypes provided as input to EthSEQ in VCF format.

Principal component analysis (PCA) is next performed by means of *SNPRelate* R package (Zheng *et al.*, 2012) on aggregated target and reference models genotype data; only SNPs that satisfy user-defined call rate are retained. The space defined by the first two PCA components is then automatically inspected to first generate the smallest convex sets identifying the ethnic groups described in the *reference model* and next to annotate the ethnicity of the individuals of interest (Supplementary Fig. S1b). Individuals positioned inside an ethnic group set (or intersecting group sets) are annotated with the corresponding ethnicity and labeled with INSIDE. For individuals positioned outside all ethnic group sets, the relative contribution of each group is computed through the distances from the centroids using the procedure described in Supplementary Figure S2, and top ranked contributing groups are reported (labeled CLOSEST).

To better discern ethnicity annotations across ancestrally close groups within a study cohort (for instance Ashkenazi and Caucasians), a multi-step inference procedure is provided. Given a tree of ethnic group sets such that sibling nodes have non-intersecting ethnic groups and child nodes have ethnic groups included in the parent node ethnic groups, ethnicity of individuals is inferred by a pre-order traversal of the tree. At each node with ethnic groups S, annotations resulted from the analysis of the parent node is refined by reducing both *reference* and *target* models on individuals with annotations in S only. Global annotation of all individuals is updated throughout the tree traversal.

3 Performances and results

Performances of EthSEQ ethnicity inference method were tested for precision, computational time and dependence on SNP set size on two main datasets, 1000 Genomes Project genotype data and germline samples TCGA data.

Initial precision tests utilized 1000 Genomes Project data; we randomly divided 2096 individuals into reference and target model groups while preserving the ethnic groups proportions, and ran EthSEQ relying only on SNPs in WES platform-specific captured regions (Supplementary Methods). Analyses were performed using reference models either built considering major ethnic groups annotations (EUR, AFR, EAS and SAS) or considering annotations for all the corresponding 21 populations reported in the 1000 Genome Project. In the first case, individuals' ethnicities were all correctly classified (100% precision and more than 97% of the individuals annotated with the INSIDE label) (Supplementary Fig. S3, Table S1). When the fine-grained annotation was used, ethnicity inference reached a precision of 92.2% with the multi-step refinement analysis. For instance, when considering European individuals only that includes 5 populations, precision reached 94% (Supplementary Methods for details).

Finally, EthSEQ performances were compared to LASER 2.0 *trace* module (Wang *et al.*, 2015) performances on the same data. Results in the PCA space were highly concordant (Supplementary Figs S4 and S5), but EthSEQ was up-to 10X faster using a single core and up-to 18X faster when exploiting parallel computation (Supplementary Fig. S6) on multi-individual analyses (Supplementary Methods for details).

Further, EthSEQ was ran on germline WES data from 604 TCGA (cancergenome.nih.gov) individuals with reported interview-based race classification (as per TCGA nomenclature, race is annotated as 513 *White*, 42 *Black or African American* and 49 *Asian*). 505 *White* individuals were annotated by EthSEQ as EUR, 37 *Black or African American* individuals as AFR and 48 *Asian* individuals as EAS or SAS for an overall precision of 97.7%. EthSEQ results were compared to results from fastSTRUCTURE tool (Raj *et al.*, 2014) fed with genotype data generated by EthSEQ pre-processing module. For 594 individuals (98.3%) the two analyses inferred the same major ethnic contribution. Both tools inferred 5 individuals originally annotated as *Black or African American* in TCGA as admixed with a major Caucasian contribution, one originally annotated as *Asian* as non-admixed Caucasian, and two originally annotated as *White* as major African contribution (see Supplementary Methods and Table S1). In terms of tool specific results, 4.6% of individuals were inferred as admixed by fastSTRUCTURE that explained the TCGA dataset population structure with 3 clusters achieving a precision of 98%; 7.9% of individuals were inferred as CLOSEST by EthSEQ with the majority with SAS main contributions, not captured by fastSTRUCTURE, and secondary African contribution correctly detected by EthSEQ when above 15%. EthSEQ analysis resulted 3.2X faster.

The effectiveness of the multi-step refinement analysis was recently proven in a precision medicine setting study (Zhang *et al.*, 2016) were ethnicity based stratification was key to interpret the relevance of germline cancer-associated variants. Specifically, our analysis ruled out the possibility that the high fraction of germline cancer-associated variants observed in the clinical cohort of 343 patients with metastatic tumors (Beltran *et al.*, 2015) was due to the presence of Ashkenazi inheritance (Carmi *et al.*, 2014) shown to carry high percentage of cancer-associated variants. Provided an Agilent HaloPlex reference model including Ashkenazi genome data (Carmi *et al.*, 2014) the identification of Ashkenazi individuals required the multi-step analysis to precisely discern them from the ancestrally close European individuals; 29.7% of Ashkenazi individuals were identified confirming the anticipated fraction of about 30% based on an internal cancer registry.

To measure the impact of the number of available SNPs on EthSEQ precision, we extended the performance analyses by randomly down-sampling the number of SNPs both in the 1000 Genome Project and in the TCGA dataset (Supplementary Methods). Supplementary Figure S7 shows that using the multi-step refinement analysis, 2000 SNPs are sufficient to reach more than 98% precision. Overall, this data indicates that EthSEQ is also amenable to targeted sequencing NGS data.

4 Conclusions

We presented EthSEQ, a rapid, reliable and easy to use R package to annotate individuals ethnicity from WES data. EthSEQ can be used to process single sample or multi-sample datasets, provides a large variety of pre-computed platform-specific reference models, a simple and transparent mode to generate ethnicity annotations starting from a list of BAM files and can be easily integrated into any WES based processing pipeline also exploiting multi-core capabilities.

Funding

This work has been supported by the Prostate Cancer Foundation Challenge Award 2014 (F.D., A.R.), the Caryl and Israel Englander Institute for Precision Medicine, New York and the European Research Council ERCCoG648670 (F.D.).

Conflict of Interest: none declared.

References

Beltran,H. *et al*. (2015) Whole-exome sequencing of metastatic cancer and biomarkers of treatment response. *JAMA Oncol*., **1**, 466–474.

Carmi,S. *et al*. (2014) Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat. Commun*., **5**, 4835.

Petrovski,S. and Goldstein,D.B. (2016) Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol*., **17**

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

Raj,A. *et al*. (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, **197**, 573–589.

Romanel,A. *et al*. (2015) ASEQ: fast allele-specific studies from next-generation sequencing data. *BMC Med. Genomics*, **8**, 9.

Spratt,D.E. *et al*. (2016) Racial/ethnic disparities in genomic sequencing. *JAMA Oncol*., **2**, 1070.

Wang,C. *et al*. (2015) Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *Am. J. Hum. Genet*., **96**, 926–937.

Zhang,T. *et al*. (2016) Germline Variants and Secondary Findings in a Cancer Precision Medicine Cohort. In: *Laboratory Investigation*. Nature Publishing Group 75 Varick St, 9TH Flr, New York, NY 10013-1917 USA, pp. 461A–461A.

Zheng,X. *et al*. (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.