Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: X XXXX XXXX
Applications Note

Subject Section

AKT: Ancestry and Kinship Toolkit

Rudy Arthur [1,*], Ole Schulz-Trieglaff [1], Anthony J. Cox [1] and Jared O'Connell [1,*]

[1] Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Essex CB10 1XL, United Kingdom

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Ancestry and Kinship Toolkit (AKT) is a statistical genetics tool for analysing large cohorts of whole-genome sequenced samples. It can rapidly detect related samples, characterise sample ancestry, detect IBD segments, calculate correlation between variants, check Mendel consistency and perform data clustering. AKT brings together the functionality of many state-of-the-art methods, with a focus on speed and a unified interface. We believe it will be an invaluable tool for the curation of large WGS data-sets.
Availability: The source code is available at `https://illumina.github.io/akt`
Contact: joconnell@illumina.com, rudy.d.arthur@gmail.com

1 Introduction

As whole genome sequencing (WGS) costs decrease, it is becoming common to have re-sequencing data for cohorts of thousands of individuals (Taylor *et al.*, 2015; Lek *et al.*, 2015; Gudbjartsson *et al.*, 2015). Such large cohorts will often have cases of sample duplication, cryptic relatedness and heterogeneous ancestry. These data sets require careful curation before further analysis. In the DNA-microarray world, a range of high-quality tools are available to perform principal component analysis (PCA), kinship coefficient calculation and other routine quality control analyses (Purcell *et al.*, 2007; Manichaikul *et al.*, 2010; Yang *et al.*, 2011). While the algorithms implemented in such tools remain relevant, they require custom formats that are not well suited to WGS data. The conversion between the standard WGS format (VCF/BCF) and these custom formats can be time consuming and error prone for end-users. Additionally, the larger number of rare variants and false-positives in WGS data require some care to handle correctly.

In this note we present *AKT*, a software suite designed to perform routine analyses of large re-sequencing data sets. We envision AKT being applied to large multi-sample BCFs to identify related samples, detect sample swaps and ascertain the spectrum of ancestry in a cohort. Our focus is on speed and simplicity with the hope that this toolkit can become a standard part of the bioinformatician's arsenal when investigating large cohorts of WGS samples. AKT is freely available under the GPLv3 license. It is implemented in C++ using HTSlib (Li *et al.*, 2009) for fast reading of VCF/BCF files and the Eigen matrix library for matrix manipulations (http://eigen.tuxfamily.org).

2 Methods

AKT follows the popular bioinformatics convention of combining many sub-functions into a single binary, analyses are run via: *akt subcommand input.bcf*. Many of the algorithms we describe do not require the entire dense set of variants that will be present in a WGS cohort. Indeed, some of the estimators assume variants are in linkage equilibrium. A standard way of achieving this is to thin variants. We provide this thinning functionality, but this involves decompressing an entire BCF which is time consuming. For example, the final release of the 1000 Genomes Project (1000GP) has 84.8 million variants (The 1000 Genomes Project Consortium, 2015). Our preferred approach is to provide AKT with a well behaved sparse set of common SNPs, AKT can then use tabix indexing (Li, 2011) which substantially reduces file reading time. We distribute an appropriate site-only VCF with AKT.

Fast principal component analysis PCA is a common method to detect and classify ancestry. Plotting the first few principal components will identify large population structure present in a cohort. Reducing a large genotype matrix with M markers and N samples to principal components requires calculating its singular value decomposition (SVD). Exact SVD is quite slow, $O(MN^2)$ when $M{>}N$. However it is often sufficient to compute an inexact SVD to obtain the first and most important principal components (Patterson *et al.*, 2006). We use the RedSVD implementation (Tropp *et al.*, 2009), a very fast approximate SVD algorithm. We also provide options to compute the exact SVD using the Jacobi algorithm and to project samples onto pre-computed principal components.

Kinship coefficients and average IBD sharing Estimating the proportion of the genome that is identical-by-descent (IBD) between two samples allows us to ascertain the degree of relatedness between them or to check if the samples are duplicates. We calculate the same IBD estimators used in PLINK which require population allele frequencies. Users can either estimate frequencies from their data or provide pre-computed frequencies from a reference panel such as 1000GP. The latter option can be especially useful when sample sizes are small.

Detecting cryptic pedigrees We implement a similar routine to Staples *et al.* (2014). First order relationships (parent-child and sibling) have IBD patterns which allow easy classification. When both parents in a nuclear family are assayed, pedigrees can be reconstructed unambiguously. In cases where only one parent is assayed, a parent-child relationship can be established but not which sample is the parent. Grandparent-grandchild relationships and sibling relationships (when no parents are assayed) can also be detected.

Detecting segments shared IBD Average IBD sharing may be too coarse a metric for distantly related samples. The ability to detect segments of the genome that are shared IBD between two samples is useful. We follow the approach of Loh *et al.* (2015), with a focus on speed rather than detecting the smallest possible IBD segments. We first detect long runs of genotypes with IBS$>$0 and extend these seed regions in an error-tolerant way to give candidate IBD segments. We assign a score to each potential IBD segment which is the sum of the log odds ratio of the genotypes assuming IBD versus non-IBD. We can then filter on segment length and score to detect regions of IBD sharing between samples.

Other functions We also include code for data clustering using k++-means, Gaussian mixtures and density based methods (Rodriguez and Laio, 2014); calculation of LD metrics including correlation and LD score (Bulik-Sullivan *et al.*, 2015); transforming principal component projections to ancestry fractions (Zheng X., 2016) as well as profiling of Mendelian inheritance patterns for pedigrees.

3 Results

We demonstrate the speed and ease-of-use of AKT on publicly available 1000GP data. We test AKT on two data-sets:

- 1000GP phase 3 release (2504 unrelated samples, 84.8M variants)
- 433 high-coverage samples (including 129 trios and 9 duos, 34.4M variants).

The first data set is perhaps the most commonly analysed WGS cohort, the latter allows us to evaluate the pedigree analysis components of AKT.

Table 1 lists the commands used and respective timings. On the larger N=2504 dataset, PCA took 63 seconds (single threaded) and kinship coefficient calculation took 47 seconds (20 threads). Re-constructing pedigrees on the smaller data set, took 45 seconds. All trios were correctly reconstructed and all parent-child relationships were identified (we acknowledge this is a fairly straightforward example). Profiling Mendel error rates on these pedigrees across all 34.4M sites took 310 seconds (single threaded).

4 Conclusion

AKT is a powerful tool for for bioinformaticians who routinely deal with large numbers of WGS samples. AKT will help in cases where meta-data about the samples may be missing or unreliable, allowing easy inference of ancestry and relatedness from the data itself. We expect to expand the functionality of AKT with time, however the software will already enable rapid and accurate curation of WGS data. This short analysis gives a feel for the power and speed of AKT for some common problems.

Funding: All authors are employees of Illumina Inc., a public company that develops and markets systems for genetic analysis, and receive shares as part of their compensation.

Table 1. Timing results for a subset of AKT functionality. Analysis was performed on the 1000 Genomes Phase 3 BCF (n2504.bcf) and on a separate set of 433 high-coverage samples (n433.bcf). Where appropriate, we perform analysis using a thinned list of 17535 common SNPs (snps.vcf.gz). Results are all single-threaded, except for kin which used 20 threads. Intel Xeon E5-2670 CPUs were used.

Algorithm	Command line	Time (s)
PCA	akt pca -R snps.vcf.gz n2504.bcf > n2504.pca	62.6
kinship	akt kin -n 20 -R snps.vcf.gz n2504.bcf > n2504.kin	47.3
Finding cryptic pedigrees	akt kin -n 20 -R snps.vcf.gz n433.bcf > n433.kin akt relatives n433.kin -p n433.fam	43.6 <1
Mendel profile	akt mendel n433.bcf -p n433.fam	310

References

Bulik-Sullivan *et al.* (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature*, **47**(3), 291–295.

Gudbjartsson *et al.* (2015). Large-scale whole-genome sequencing of the icelandic population. *Nature genetics*, **47**(5), 435–444.

Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Fennell, T., O'Donnell-Luria, A., Ware, J., Hill, A., Cummings, B., *et al.* (2015). Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*, page 030338.

Li, H. (2011). Tabix: fast retrieval of sequence features from generic tab-delimited files. *Bioinformatics*, **27**(5), 718–719.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Loh, P., Pier Francesco Palamara, F., and Price, A. (2015). Fast and accurate long-range phasing and imputation in a UK biobank cohort. *bioRxiv*, page 028282.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**(22), 2867–2873.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet*, **2**(1), e190.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M., and Sham, P. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, **81**.

Rodriguez, A. and Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, **344**(6191), 1492–1496.

Staples, J., Qiao, D., Cho, M. H., Silverman, E. K., Nickerson, D. A., and Below, J. E. (2014). PRIMUS: Rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *Am J Hum Genet.*, **95**(5), 553–564.

Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J.-B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., *et al.* (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature genetics*, **47**(7), 717–726.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.

Tropp, A., Halko, N., and Martinsson, P. (2009). Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. Technical report, Technical report.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, **88**(1), 76–82.

Zheng X., B. S. W. (2016). Eigenanalysis of snp data with an identity by descent interpretation. *Theoretical Population Biology*, **107**, 65–476.