The American Journal of Human Genetics, Volume *100*

Supplemental Data

Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with *Peddy*

Brent S. Pedersen and Aaron R. Quinlan

Supplementary Materials

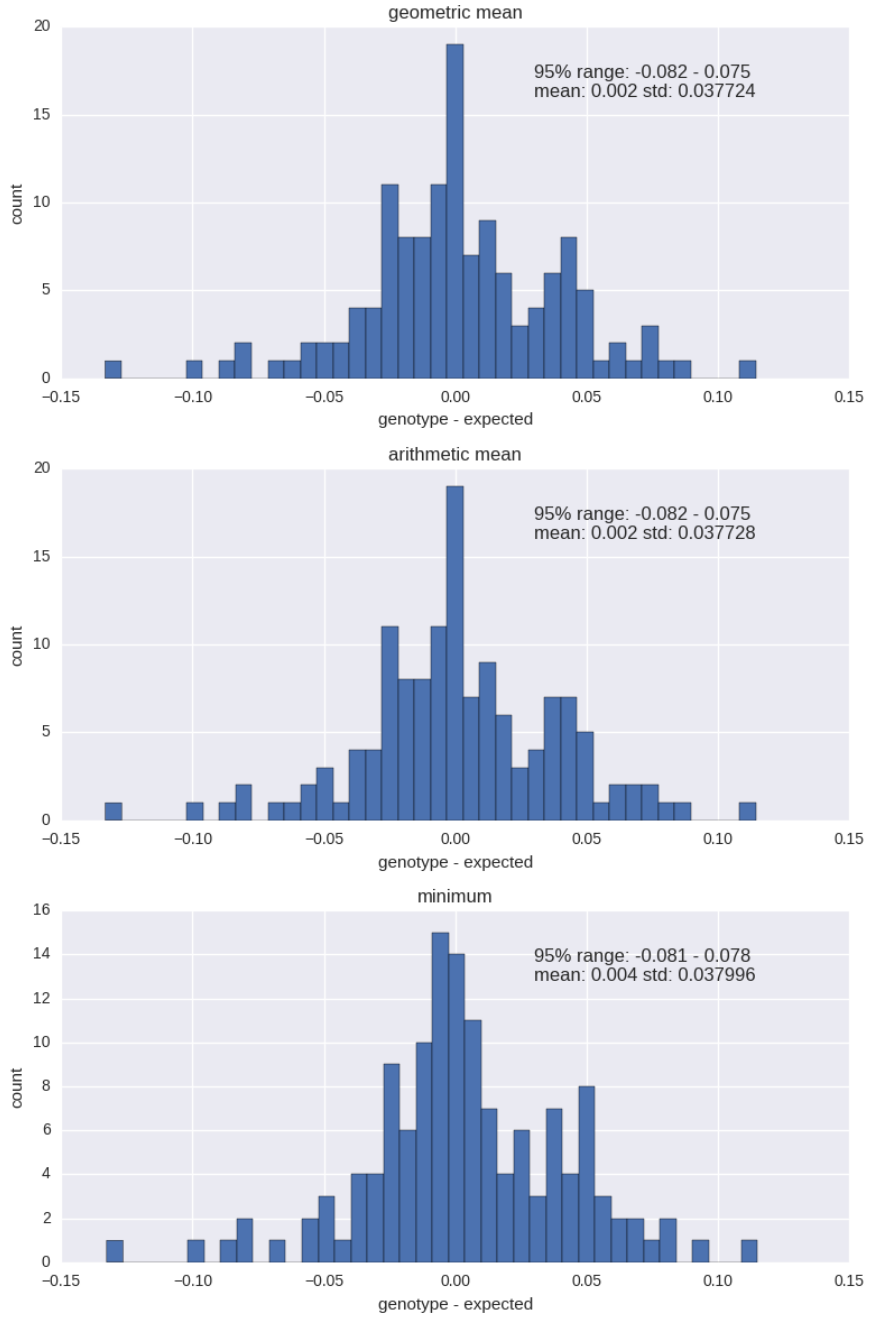

Figure S1. Comparison of methods for relatedness calculation. Here we illustrate the difference between the expected relatedness (from the relationship known by the pedigree) and the relatedness calculated from the observed genotypes. We have chosen to use the geometric mean to calculate the relatedness coefficient in the top panel. KING uses the mean (panel 2) to measure *within* family relatedness and the minimum (panel 3) to measure *between* family relatedness. Given that in some cases, the family may be mis-specified, we use the geometric mean to avoid bias. Here we show that the choice matters little for the sites we sampled by peddy, but the geometric mean has a low bias and a smaller 95% interval. The use of minimum has the largest bias and the largest 95% interval.

Assuming minimal bias during DNA library preparation, the ratio of sequence alignments harboring the alternate allele is expected to follow a binomial distribution (p~0.5) for all sites at which an individual is heterozygous (**Figure S2A**, bottom panel). Substantial deviation from this expectation is potential evidence for either aberrantly low average sequencing depth or contamination with DNA from other individuals in the DNA library (**Figure S2A**, top panel). *Peddy* measures the inter-decile range (10th to 90th percentile; IDR) of alternate allele ratios from heterozygous genotypes

as a statistic to summarize the degree to which the binomial expectation is violated for each individual (**Figure S2B**). Individuals with potential contamination will have substantially more heterozygous genotypes than other individuals and will have a higher alternate allele ratio IDR. In contrast, individuals conceived from consanguineous parents will have substantially fewer heterozygous genotypes, reflecting a higher degree of homozygosity.

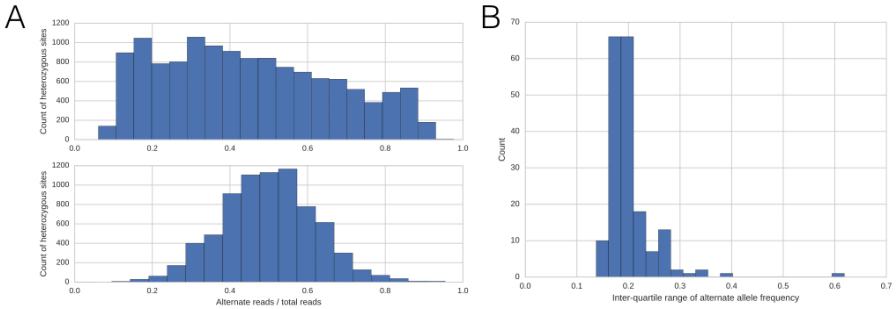

Figure S2. Inter-decile range of fraction of alternate reads. The top panel in A shows a sample with a large inter-decile range while the bottom panel shows the distribution of a good-quality sample with a lower range. The distribution of all samples is shown in Figure 1B, where we can clearly see the outlier at the far right.

Web resources

Software Availability: https://github.com/brentp/peddy
Demonstration (Chrome suggested): http://peddy.readthedocs.io/en/latest/_static/ceph.html
cyvcf2: github.com/brentp/cyvcf2
htslib: github.com/samtools/htslib

References

1. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–575. PMCID: PMC1950838

2. Andrews S. FastQC: A quality control tool for high throughput sequence data. Reference Source. 2010;

3. Miller CA, Qiao Y, DiSera T, D'Astous B, Marth GT. bam.iobio: a web-based, real-time, sequence alignment file inspector. Nat Methods. 2014 Dec;11(12):1189. PMCID: PMC4282680

4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–2079. PMCID: PMC2723002

5. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov 15;26(22):2867–2873. PMCID: PMC3025716

6. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–2158. PMCID: PMC3137218

7. Heinrich V, Kamphans T, Mundlos S, Robinson PN, Krawitz PM. A likelihood ratio based method to predict exact pedigrees for complex families from next-generation sequencing data. Bioinformatics [Internet]. 2016 Aug 26; Available from: http://dx.doi.org/10.1093/bioinformatics/btw550 PMID: 27565584

8. Staples J, Ekunwe L, Lange E, Wilson JG, Nickerson DA, Below JE. PRIMUS: improving pedigree reconstruction using mitochondrial and Y haplotypes. Bioinformatics. 2016 Feb 15;32(4):596–598. PMID: 26515822

9. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011 Mar 1;27(5):718–719. PMCID: PMC3042176

10. Eberle MA, Fritzilas E, Krusche P, Kallberg M, Moore BL, Bekritsky MA, Iqbal Z, Chuang H-Y, Humphray SJ, Halpern AL, Kruglyak S, Margulies EH, McVean G, Bentley DR. A reference dataset of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree [Internet]. 2016 May. Available from: http://biorxiv.org/lookup/doi/10.1101/055541

11. Blue E.M., Brown L.A., Conomos M.P., Kirk J.L., Nato A.Q., Popejoy A.B., Raffa J., Ranola J., Wijsman E.M., Thornton T. Estimating relationships between phenotypes and subjects drawn from admixed families. BMC Proc.

12. Halko N, Martinsson P-G, Tropp JA. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions [Internet]. arXiv [math.NA]. 2009. Available from: http://arxiv.org/abs/0909.4061

13. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74. PMCID: PMC4750478

14. Boada R, Janusz J, Hutaff-Lee C, Tartaglia N. The cognitive phenotype in Klinefelter syndrome: a review of the literature including genetic and hormonal factors. Dev Disabil Res Rev. 2009;15(4):284–294. PMCID: PMC3056507