Verifying Identity of DNA and RNA Specimens Through SNP Analysis

Kayla Schimke[1], Samuel Nkrumah[1], Alden Huang[2], Hane Lee[3], Stanley Nelson[3,4]

1. BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA,
2. Institute for Precision Health, David Geffen School of Medicine, UCLA,
3. Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, UCLA,
4. Department of Human Genetics, David Geffen School of Medicine, UCLA

Background

Clinical molecular laboratories have incorporated next-generation sequencing of genomic DNA as an essential component in the diagnosis of rare congenital disorders. Given that 50-75% of disease causing variants occur in non-coding regions [1], there has been an increased interest in also performing transcriptomic profiling in patient samples to identify genetic variants that may lead to regulatory and/or splicing defects, further inform variant interpretation and ultimately, improve diagnostic yield. Therefore, reliable methods to unequivocally identifying individuals from multi-omics data are required.

Here, we describe a computational workflow to generate informative SNP panels for the molecular fingerprinting samples from whole-genome (WGS), whole-exome (WES), and transcriptome (RNA-seq) sequencing. We evaluated the performance of our SNP panels on select patient samples submitted for clinical sequencing at the UCLA Undiagnosed Diseases Network clinical site. Our data was comprised of 191 samples with genomic sequencing, either WGS or WES, and matched RNA-seq data from either whole blood (N=151) or fibroblasts (N=108). Our method is computationally efficient, outperforms existing panels, and accurately distinguishes between even closely-related individuals.

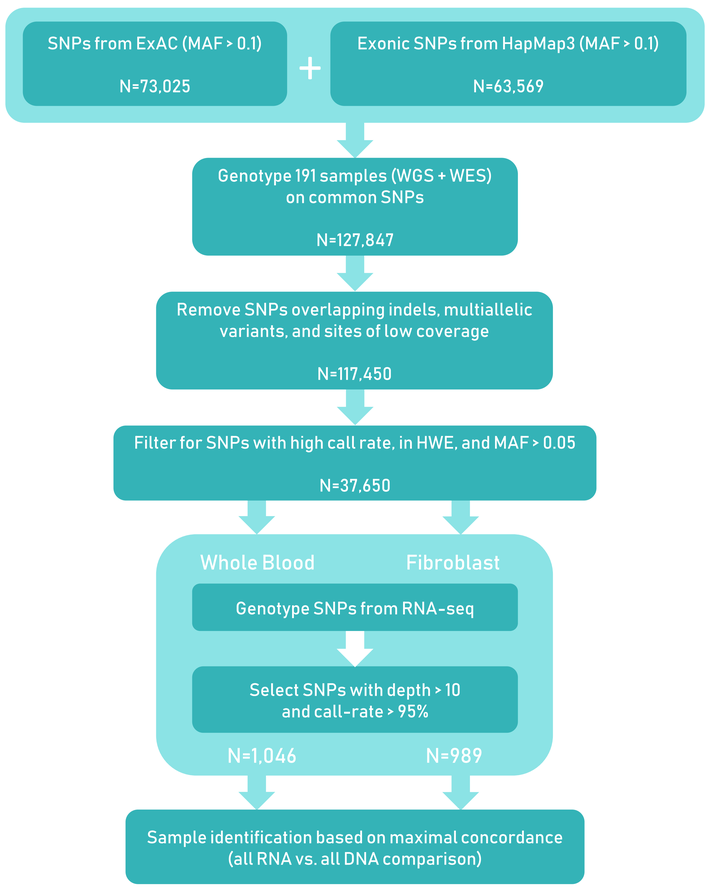

Figure 1. Overview of SNP selection to generate fingerprint panels for comparing sequencing samples from blood RNA, fibroblast RNA, and DNA.

Results

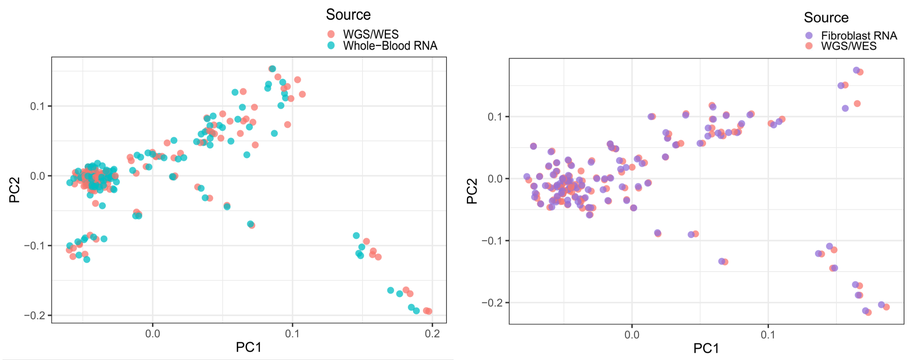

Figure 2. Principal component analysis of SNP calls from both RNA and DNA
PCA analysis was conducted separately for blood (left) and fibroblasts (right) along with matching DNA samples using all SNPs from each respective fingerprinting panel (Fig 1).

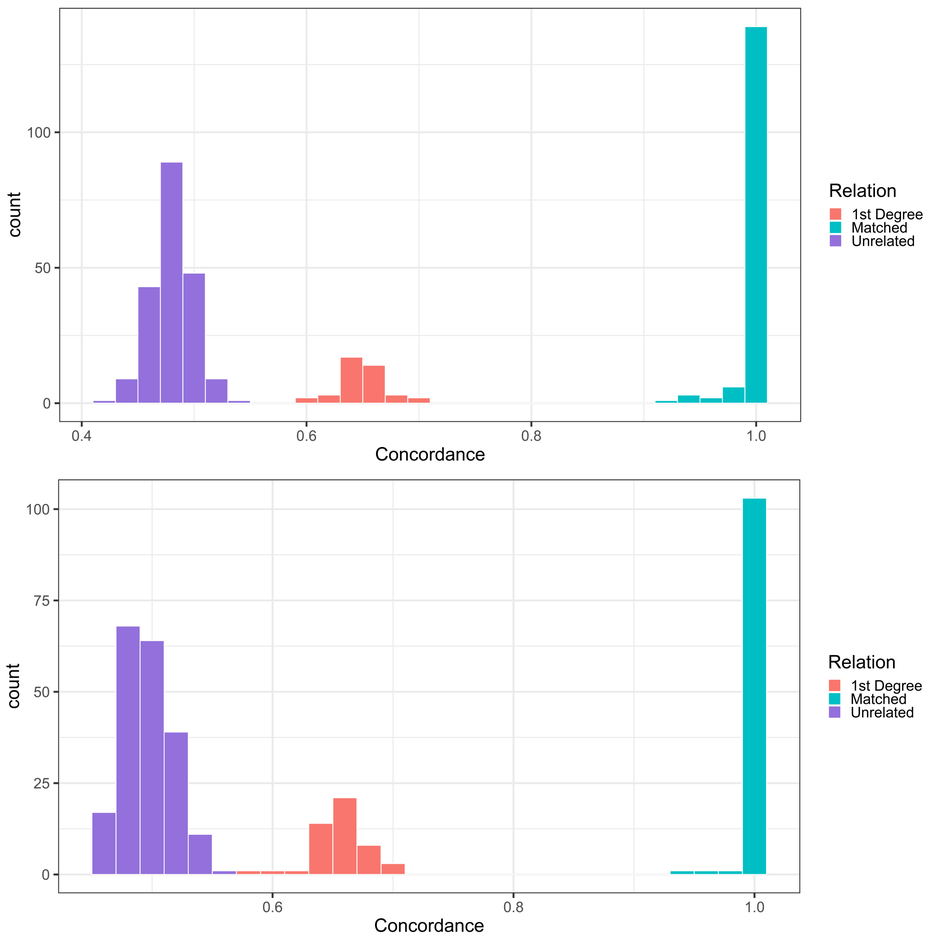

Figure 3. Performance comparison of fingerprinting panel
To evaluate the discriminatory ability of our fingerprinting panels, we compared the concordance rate of SNP calls from RNA-seq in blood (top) or fibroblasts (bottom) with those derived from WGS/WES for unrelated, first-degree relative, and matched sample pairs. Concordance was determined as the number of matching calls between RNA-seq and WGS/WES divided by the total number of RNA-seq calls. Unrelated pairs of samples were selected randomly.

Computational Methods

WGS (30X) or WES (150X) was performed using genomic DNA extracted from whole blood. RNA-seq was conducted using total RNA extracted from either whole-blood or skin fibroblast. Alignment was completed using BWA-mem (WGS/WES) or STAR (RNA-seq). Genotyping was performed using GATK. Statistical analysis of genotypes was conducted using PLINK and R.

RNA Source	SNP Panel Size	Mean Call Rate	Mean Concordance
Whole-Blood	1045	1031.1 ± 33.2	0.996 ± 0.01
Fibroblast	989	977.4 ± 27.7	0.991 ± 0.03

Table 1. Call rate and concordance among matched RNA/DNA samples

	Sites Considered	Sites Called	Concordance	Runtime
GATK Best Practices [2]	All	102,473	0.833	14hr
HapMap Panel	76,514	11,814	0.865	15min
Fingerprinting Panel	989	987	0.998	1min

Table 2. Comparison of runtime and performance of different SNP panels in a single representative sample in fibroblasts

Conclusion

We developed a computational method to establish unique SNP panels for individual tissues which greatly minimizes runtime for variant calling and more accurately identifies samples. Our method does so by only genotyping the most reliably called SNPs specific to each tissue, using the filters described in Figure 1. These fingerprinting panels demonstrated superior performance compared to using either all SNPs called from RNA-seq or known, common SNPs alone (Table 2) and other previously established panels [3]. We generated the concordance values between each genomic and transcriptomic sample for the blood specimens and recorded the IDs for the pairings that returned the highest concordance. Across our dataset, we observed six pairings with mismatched IDs which indicates a sample swap, proving that we can accurately determine mismatches and verify sample identity.

References

[1] L. S. Kremer *et al.*, "Genetic diagnosis of Mendelian disorders via RNA sequencing," *Nat Commun*, vol. 8, p. 15824, Jun. 2017.
[2] "GATK | Doc #3891 | Calling variants in RNA-seq." [Online]. Available: https://software.broadinstitute.org/gatk/documentation/article.php?id=3891. [Accessed: 26-July-2018].
[3] S. Yousefi *et al.*, "A SNP panel for identification of DNA and RNA specimens," *BMC Genomics*, vol. 19, Jan. 2018.

Acknowledgements

We would like to thank UCLA BIG Summer for making it possible to conduct this research and the Nelson Lab for providing the guidance and atmosphere needed to complete this project.