

# Whole Genome Sequencing of a family trio using DNA extracted from saliva

M. Tayeb, R. Iwasiow and S. Rabuka  
DNA Genotek Inc., Ottawa, Canada

## Introduction

In recent years there have been significant advancements in Next Generation Sequencing (NGS) technology and reduced cost. This combination has made NGS a more practical and economical tool for studies requiring larger numbers of samples. Saliva collected using Oragene® self-collection kits is a non-invasive alternative method to blood samples for collecting large amounts of high quality genomic DNA, as demonstrated through its utility in numerous published GWAS studies<sup>1,2, 3</sup>. Oragene devices contain a stabilizing reagent that ensures DNA stability during ambient temperature transport and long-term storage. The extracted DNA is of high quantity and quality.

Previously, we have demonstrated that saliva DNA performs similarly to DNA from blood in whole exome and other targeted NGS applications<sup>4, 5</sup>. In this study we investigated the suitability of DNA extracted from saliva collected using Oragene for Whole Genome Sequencing. To evaluate the performance of the saliva samples we compared the sequencing metrics from a family trio of saliva samples against the metrics from a publicly available data set from a family trio sequenced using DNA from blood.

## Materials and methods

### Sample collection and DNA extraction

- Three saliva samples (2 mL) collected from a family trio of donors consisting of a mother, father and male child.
- Saliva was collected according to the instructions provided in the Oragene self-collection kit (Figure 1).
- DNA was extracted from Oragene/saliva samples using prepIT®-L2P according to DNA Genotek protocol PD-PR-006.
- DNA was quantified using the Quant-iT™ PicoGreen® kit (Invitrogen).

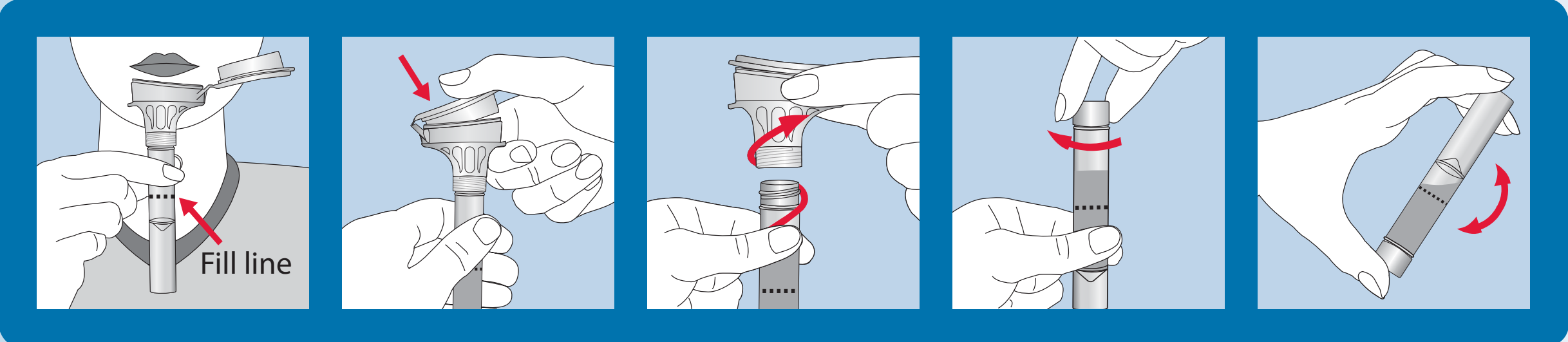


Figure 1: Oragene collection instructions

### DNA sequencing

Library preparation, sequencing and assembly were performed at Complete Genomics Inc.

### Data analysis

- A previously published data set from a family trio sequenced using blood samples was used for comparison (NA12877, NA12889 and NA12890).
- Alignment and variant calling was performed by Complete Genomics using their standard analysis pipeline.
- All data was processed through the kGAP genome analysis pipeline (Knome Inc.) for variant annotation and generation of statistics related to variant calls and classes.
- Knome's comparison algorithm was used to filter and classify variants based on genomic context.
- knomeVARIANTS, a variant query tool, was used to identify variants called specific to the blood- and saliva-derived data to determine the rate of concordance.

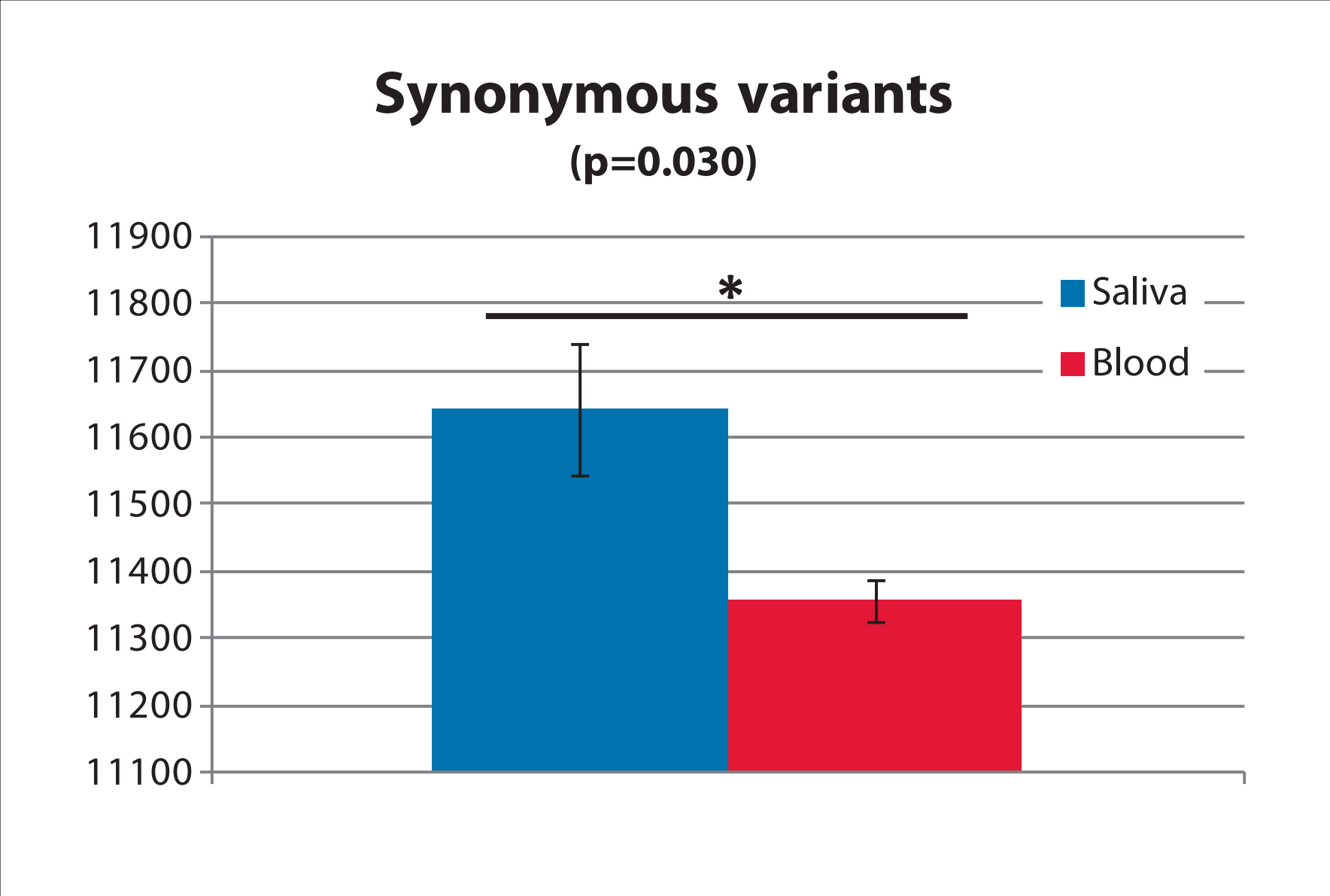
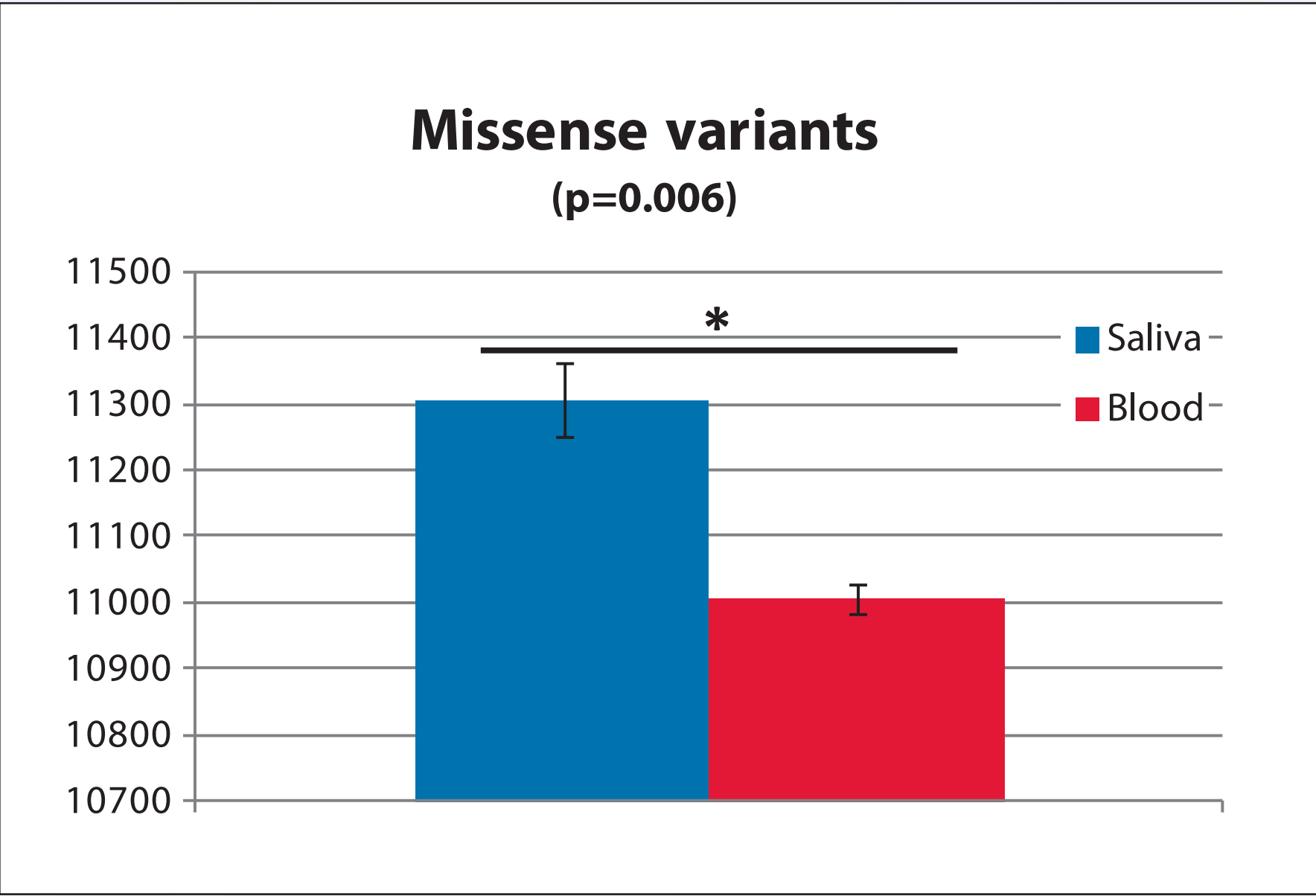
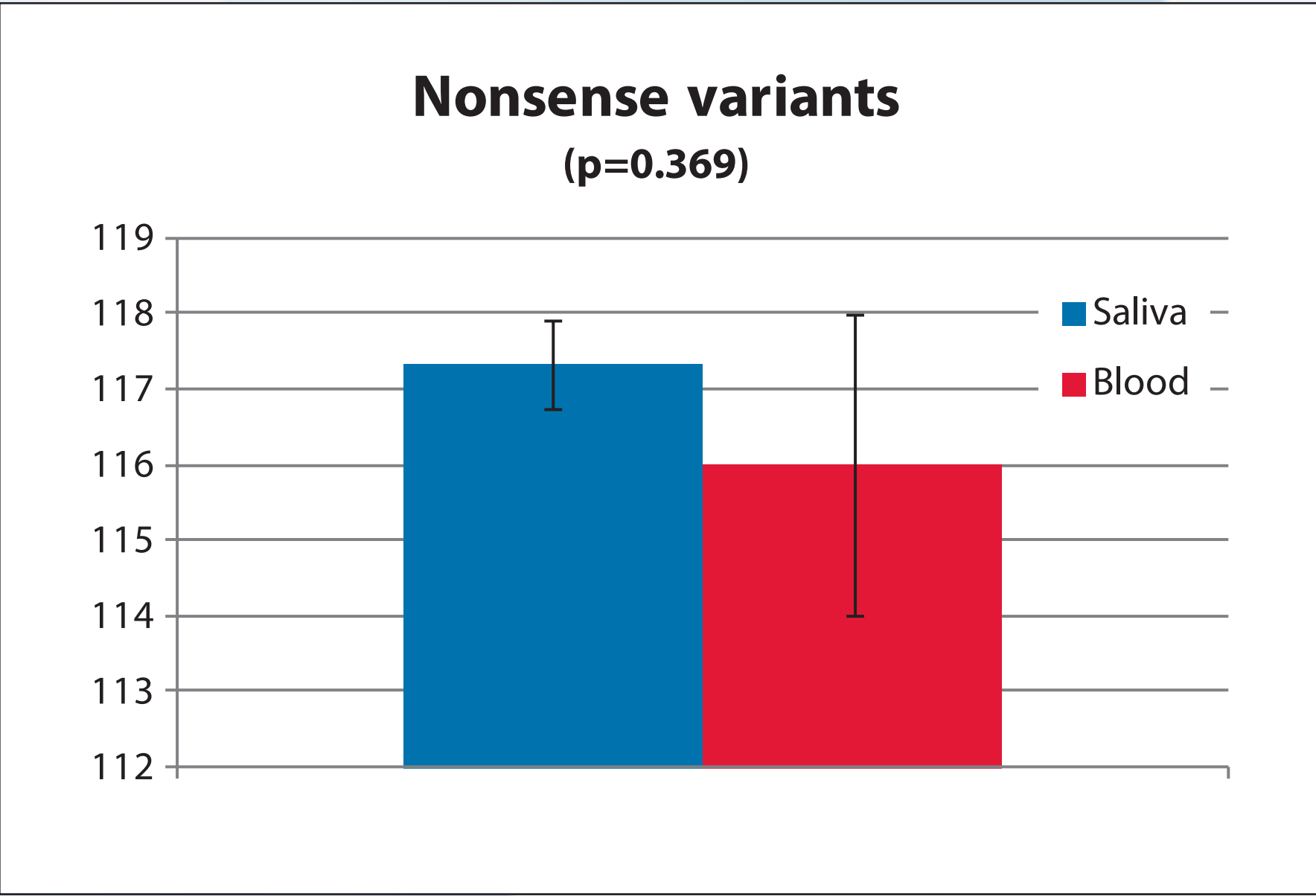
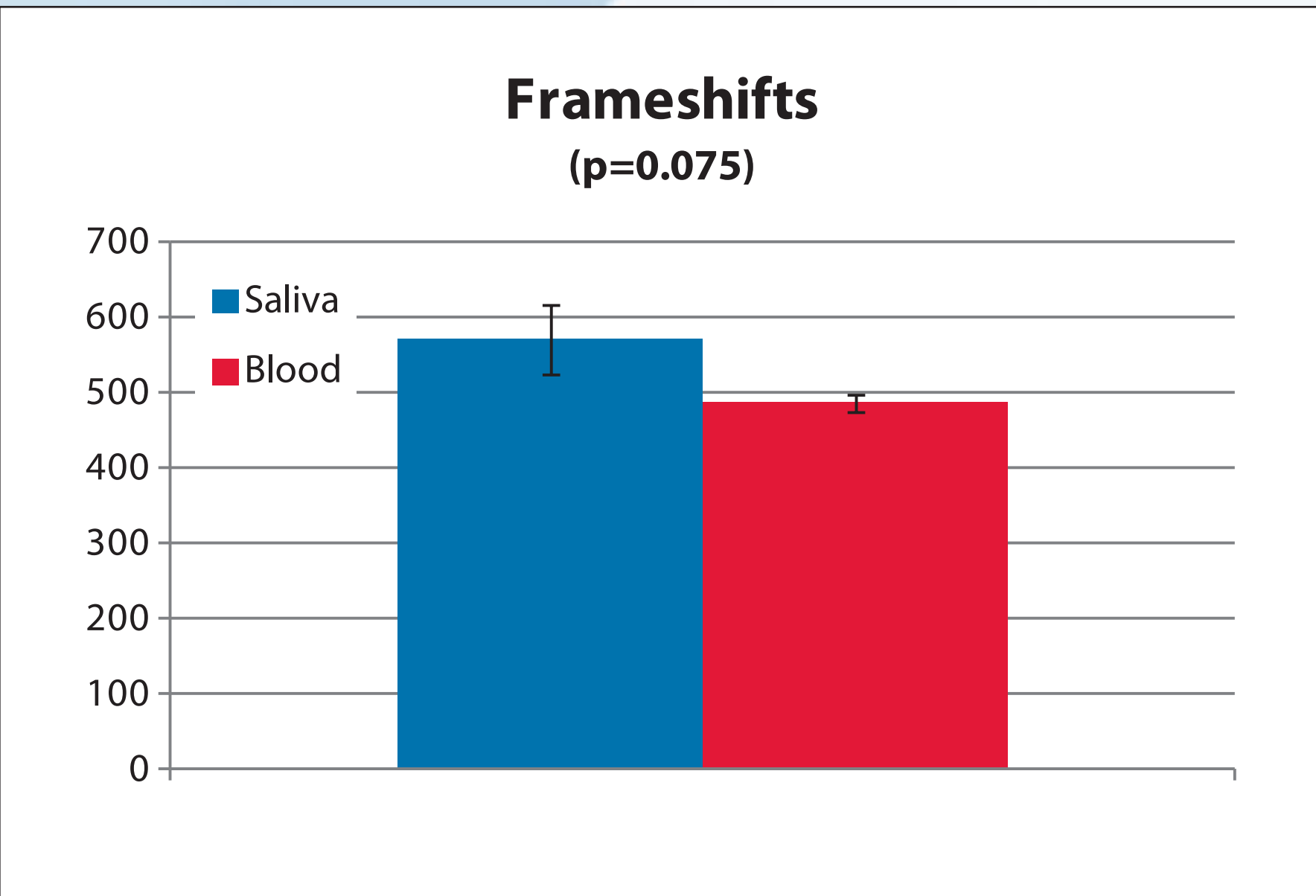
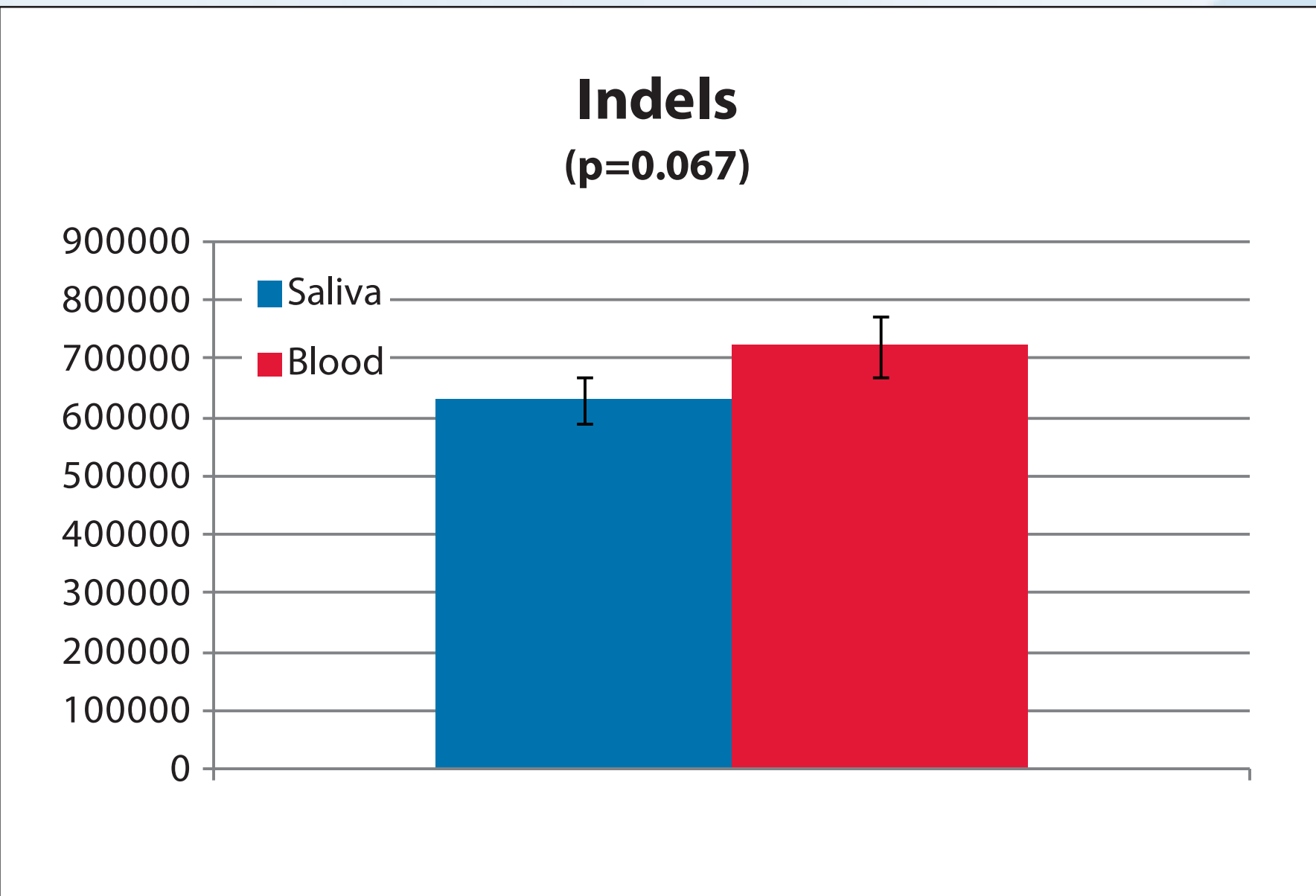
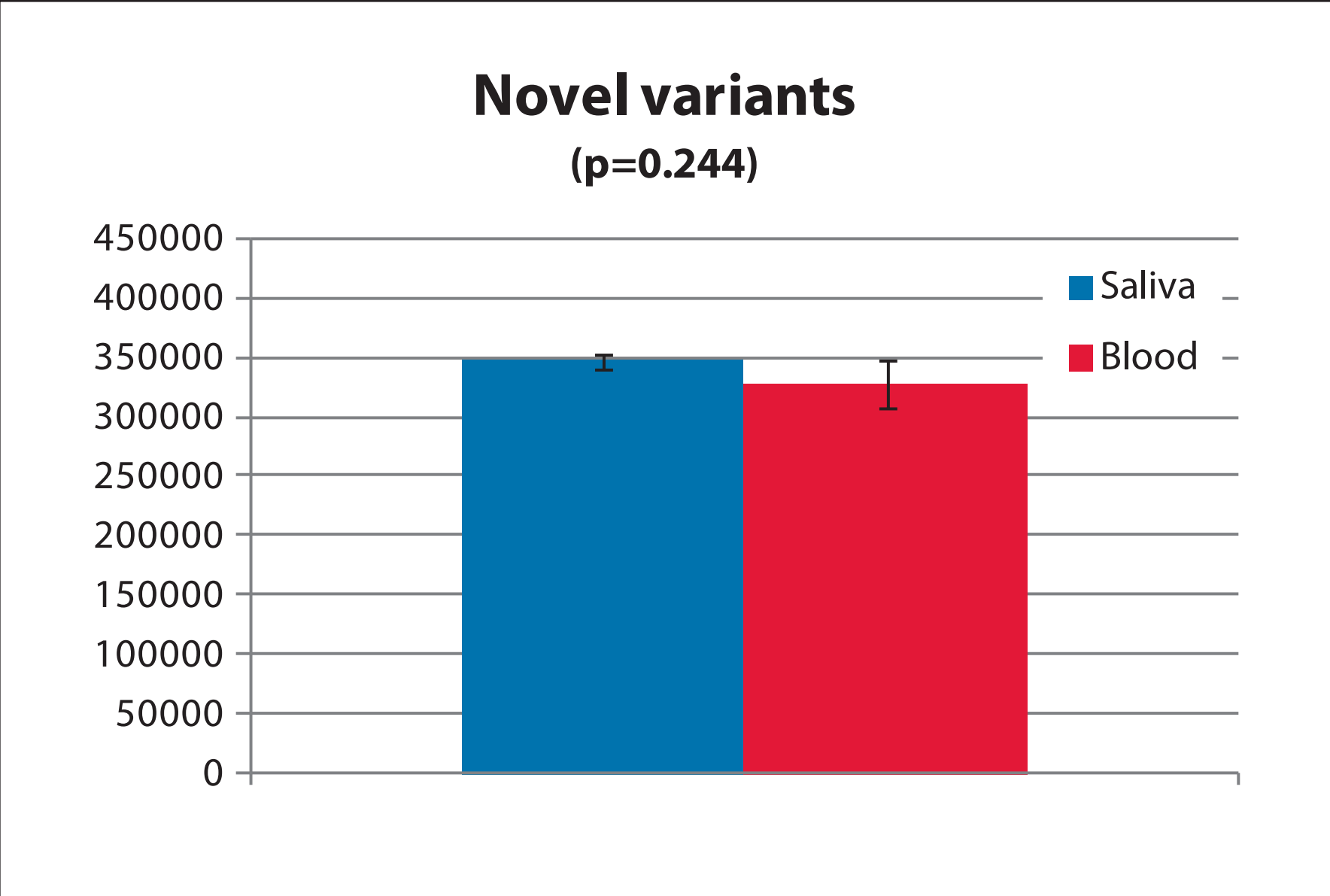
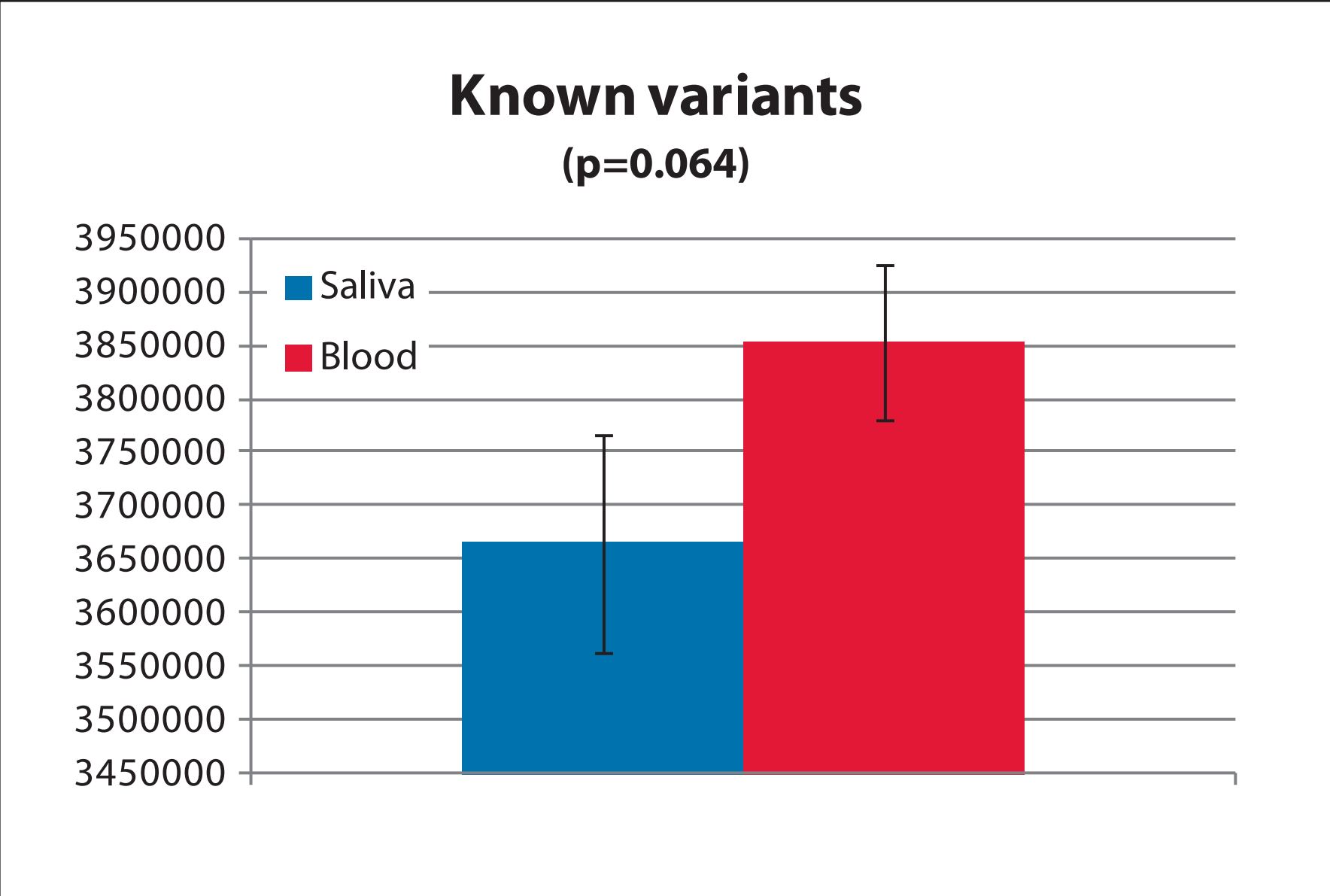
## Results

|                         | Saliva        | Blood         |         |
|-------------------------|---------------|---------------|---------|
| Mean coverage depth     | 52x           | 79x           | p-value |
| Mean total bases        | 2,737,249,262 | 2,777,478,494 | 0.029   |
| Mean heterozygous bases | 2,447,590     | 2,513,720     | 0.177   |

Table 1: Sequencing statistics for saliva and blood trios

The public blood trio samples were sequenced to a higher mean coverage depth than the saliva samples however this difference in sequencing depth resulted in only a **1.4% reduction in the number of bases called in saliva samples** compared to the blood sample. There was no statistically significant difference in the number of heterozygous sites called (Table 1).

Using a list of 51,583 homozygous, reference-mismatching variants present in all 69 normal individuals included in the CGI public genomes (<http://www.completegenomics.com/public-data/>), we evaluated the concordance of saliva and blood samples. It is likely that the mismatches contained in this list represent either errors in the reference sequence or rare reference alleles. **The saliva genomes were 99.7% concordant with the 69 CGI control genomes across these sites** (data not shown).



| Variant category | Saliva | Blood |
|------------------|--------|-------|
| Rare paternal    | 20072  | 21714 |
| Rare maternal    | 20305  | 21993 |
| Rare de novo     | 769    | 642   |
| Novel paternal   | 78746  | 76079 |
| Novel maternal   | 74959  | 78114 |
| Novel de novo    | 27408  | 16336 |

Table 2: Inherited and de novo variants among trios

## Discussion

Due to differences in sequencing depth between blood (79x) and saliva (52x) there is an inherent bias towards blood DNA. In spite of this bias, only a 1.4% reduction in the number of bases sequenced was observed. Additionally, the majority of the variant categories examined did not show a statistically significant difference in number of variants called between blood and saliva. For categories where differences were seen (missense and nonsense variants) it is likely that both differences between the sample donors and the coverage bias were contributing factors.


Maternally- and paternally-inherited variants were comparable between the saliva and blood trios, varying less than 8%. Again, due to the differing sequencing coverage and genomic variation between the saliva and blood donors, it is difficult to determine the effect of sample type on the differences observed.

Concordance of 99.7% was observed between saliva and blood samples across a set of ~50,000 previously-identified, reference-mismatching sites.

Saliva collected using Oragene provided DNA of sufficient quality and quantity for Whole Genome Sequencing. Data obtained from the family trio indicated excellent sequencing performance of all three samples. These results confirm that saliva collected using Oragene is an excellent source of genomic DNA for Next Generation Sequencing. Future experiments performed by DNA Genotek will include a direct head-to-head comparison using saliva and blood from the same donors to more completely examine sample type effects.

## References

- <sup>1</sup> Cancer Epidemiol Biomarkers Prev; 19(3) March 2010. Saliva-Derived DNA Performs Well in Large-Scale, High-Density Single-Nucleotide Polymorphism Microarray Studies. Bahlo et al.
- <sup>2</sup> Nature Genetics; 44(6) May 2012. Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. Jin et al.
- <sup>3</sup> Genomics; 98(2) Aug 2011. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Hoffman et al.
- <sup>4</sup> Saliva samples collected and stabilized with Oragene-DNA are a reliable source of DNA for Next Generation Sequencing. DNA Genotek. MK-00014.
- <sup>5</sup> Saliva collected using the Oragene family of products is a reliable source of DNA for HLA typing using Next Generation Sequencing. DNA Genotek. MK-00111.

 We would like to thank Devin Locke and Ari Kiirikki at Knome Inc for their kind assistance with analysis and interpretation of the sequencing data using their custom informatics tools.