

# Evaluation of methodologies for the analysis of human exomes using DNA extracted from saliva

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

## Introduction

Non-invasive sample collection has been demonstrated to dramatically increase donor compliance<sup>1</sup>. Saliva collected with Oragene® provides a non-invasive alternative to blood samples for collecting large amounts of high quality genomic DNA that is suitable for array based GWAS and Next Generation Sequencing studies. Oragene offers a non-invasive collection method and also contains a stabilizing reagent that ensures the sample is of high quality and allows long-term storage at ambient temperature. For these reasons, DNA isolated from saliva collected using the Oragene self-collection kit has been used in many large scale epidemiological array based GWAS studies<sup>2,3,4</sup>. In recent years it has become more practical and economical to analyze samples using Next Generation Sequencing technologies, in particular Whole Exome Sequencing.

In this study we extracted DNA from 7 year old Oragene/saliva samples stored at room temperature (~23°C) and evaluated the data from both the Illumina® HumanExome v1.1 array and Whole Exome Sequencing on the Illumina HiSeq 2000 after enrichment using the Agilent SureSelect Human All Exon v4+UTRs 71Mb Kit.

## Methods

### Collection and storage

Saliva samples were collected from 8 consented donors in 2006. Two millilitres (2 mL) of saliva was collected using the Oragene self-collection kit. After collection, the samples were heated for 1 hour at (50°C) and then stored at room temperature (~23°C) in the original collection tube for 7 years, until full purification in 2013.

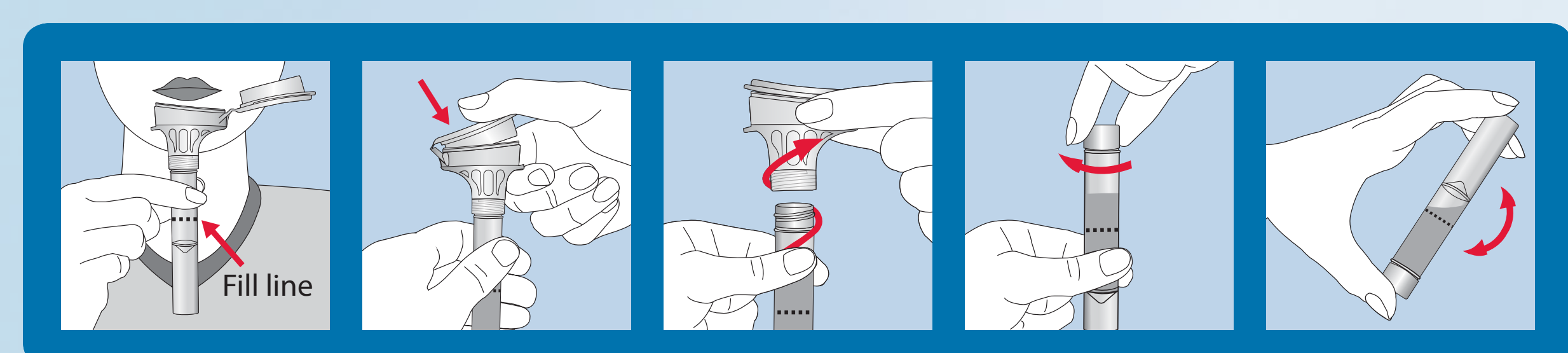


Figure 1: Oragene collection instructions

### Purification

All samples were purified using the prepIT®-L2P DNA extraction kit from DNA Genotek (protocol PD-PR-006). The kit uses a proprietary solution to remove inhibitors followed by alcohol precipitation of DNA. For all Oragene collected samples an aliquot of 500 µL was purified and eluted in 50 mL TE buffer.

### Quality control

Purified DNA was assessed using 4 different methods. First, the sample was quantified using PicoGreen® to accurately quantify the amount of DNA present. Next, the A<sub>260</sub>/A<sub>280</sub> ratio was measured using a NanoDrop® spectrophotometer and the integrity of the DNA was assessed using agarose gel electrophoresis. Approximately 100 ng of DNA as determined by PicoGreen was loaded per sample on the 0.8% agarose gel. Finally, bacterial DNA content was assessed using an in-house developed qPCR method (protocol PD-PR-065).

### Exome arrays

Samples were processed by Affiliated Genetics Inc. on the Illumina HumanExome v1.1 arrays in accordance to Illumina protocols.

### Exome sequencing

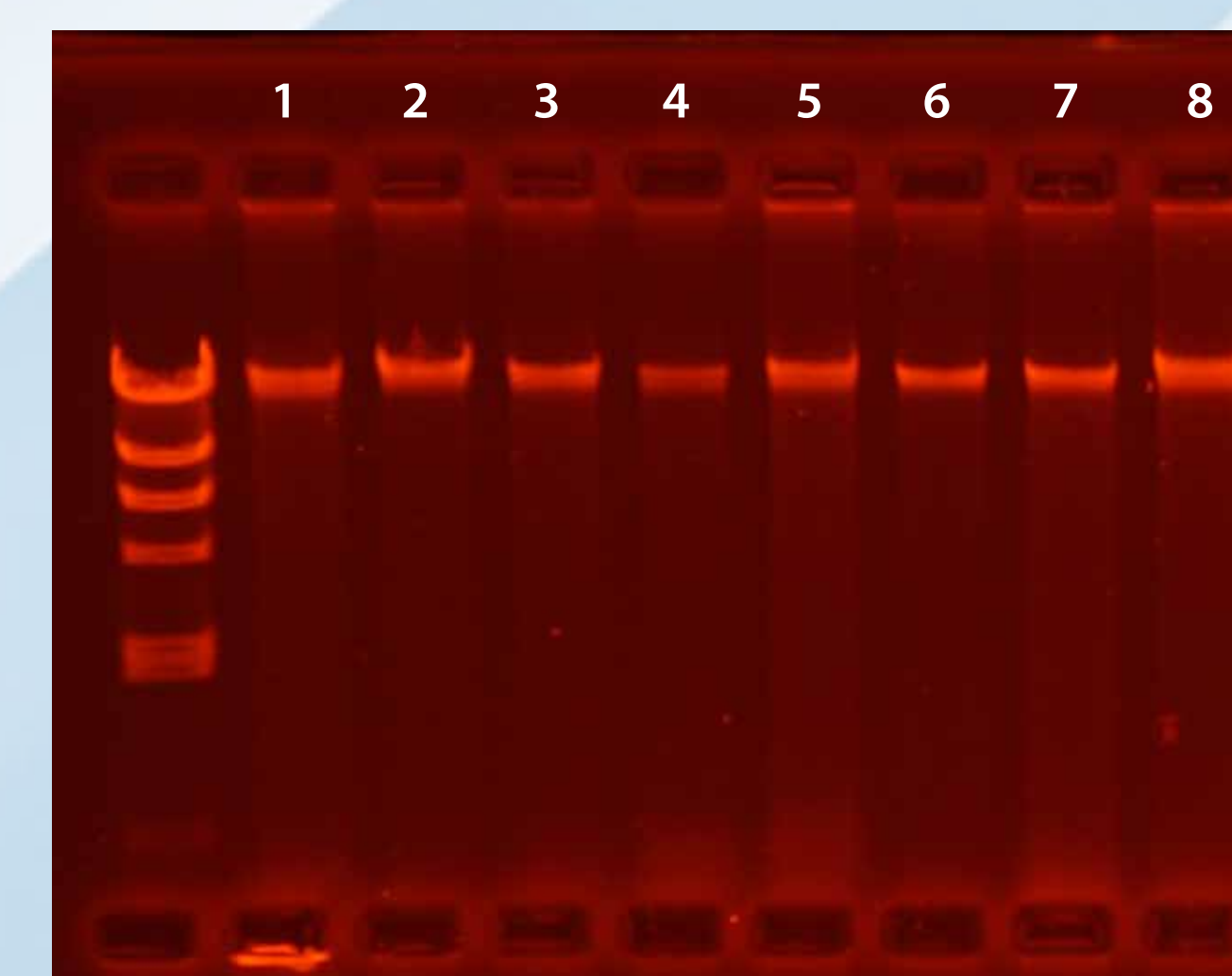
Library preparation and sequencing was performed at Expression Analysis. The DNA was enriched for the exome using the Agilent SureSelect Human All Exon v4+UTRs 71Mb Kit. The enriched library was sequenced on an Illumina HiSeq 2000 to a mean depth of 119x.



## Results

After 7 years room temperature storage (~23°C) the Oragene/saliva samples were purified. The purified DNA was of high yield, quality and molecular weight.

Sample ID	Collection date	Total yield (µg)	Concentration (ng/µL)	A <sub>260</sub> /A <sub>280</sub>	% Bact
1	2006	94.5	236.2	1.86	17%
2	2006	28.6	71.5	1.86	30%
3	2006	52.3	130.7	1.85	18%
4	2006	52.3	130.8	1.86	14%
5	2006	61.7	154.3	1.92	37%
6	2006	192.1	480.2	1.95	45%
7	2006	70.3	175.6	1.84	10%
8	2006	66.0	164.9	1.87	36%



Each sample was barcoded and 4 samples were run per lane on the Illumina HiSeq 2000. The average sequencing yield per sample was 12.5 Gb, with 98% of sequences prior to clipping aligned to the human hg19 reference.

Sample	Sequencing yield (Mb)	% Align genome	Insert mean	Mean quality	% Duplication	Mean depth
1	11610	97.70	223	36.5	0.41	111
2	11145	97.57	246	36.4	0.43	104
3	11936	97.94	220	36.6	0.37	114
4	13989	97.86	234	36.6	0.32	131
5	13399	97.60	265	36.3	0.28	124
6	13428	96.89	256	36.3	0.34	125
7	12271	98.07	213	36.5	0.35	119
8	12718	97.69	246	36.6	0.30	121

The call rates on the Illumina HumanExome v1.1 array ranged between 99.81% and 99.94%. Similarly, based on sequencing results we observed between 99.71% and 99.84% coverage of the Agilent SureSelect Human All Exon v4+UTRs 71Mb Kit. We observed, on average, 76% of sequenced bases within the captured exon regions.

Sample	Coverage (%)	# Variants (in target)	Het SNPs in target	Hom SNPs target	Indels in target	kbases not in Exon	kbases in Exon	Proportion in Exon (%)	Array call rates (%)
1	99.80	67500	42294	20899	4307	2,338,203	7,896,136	77.2	99.94%
2	99.71	64602	40651	19869	4082	2,440,522	7,422,073	75.3	99.88%
3	99.73	68642	43090	21083	4469	2,311,256	8,130,362	77.9	99.90%
4	99.75	69488	43574	21499	4415	2,863,802	9,338,541	76.5	99.91%
5	99.84	69410	43388	21493	4529	3,228,553	8,836,106	73.2	99.94%
6	99.83	69497	43565	21462	4470	3,135,419	8,939,596	74.0	99.90%
7	99.82	67223	41809	21158	4256	2,324,367	8,465,341	78.5	99.81%
8	99.81	69332	43596	21289	4447	2,925,642	8,657,029	74.7	99.89%

The Illumina HumanExome v1.1 array contains 242,901 markers of which 201,756 overlap with content located on the Agilent SureSelect Human All Exon v4+UTRs 71Mb Kit. After filtering the sequencing data for Q>20 we observed a concordance >99.2% between the two technologies across all samples. Filtering for quality > Q30 had no significant impact on concordance. As the data was further filtered to consider depth of coverage we observed increased concordance, >99.7%. We did not observe any increase in concordance when filtering at higher depths of coverage, 100x.

Sample	Concordance no filtering	Depth > 20	Depth > 30	Depth > 50	Depth > 100	Qual > 20	Qual > 30
1	99.29%	99.47%	99.64%	99.77%	99.76%	99.29%	99.30%
2	99.24%	99.42%	99.62%	99.78%	99.76%	99.24%	99.24%
3	99.34%	99.50%	99.68%	99.79%	99.78%	99.34%	99.34%
4	99.44%	99.57%	99.69%	99.80%	99.79%	99.44%	99.45%
5	99.40%	99.54%	99.69%	99.78%	99.77%	99.40%	99.40%
6	99.31%	99.46%	99.62%	99.73%	99.71%	99.31%	99.31%
7	99.22%	99.37%	99.54%	99.67%	99.66%	99.22%	99.22%
8	99.35%	99.50%	99.67%	99.75%	99.78%	99.35%	99.35%

## Highlights

- Saliva samples stored in Oragene for 7 years exhibit high yields of high molecular weight human genomic DNA.
- Saliva samples collected using Oragene are an excellent source of gDNA for array-based and Whole Exome Sequencing studies.
- The >99.7% concordance between array and exome sequencing results indicates that Oragene/saliva samples are a reliable source of gDNA which can be safely stored for years at room temperature with no impact on genotyping results.

## References

- <sup>1</sup> *Cancer Epidemiol Biomarkers Prev*; 15(9) Sep 2006. Quality and quantity of saliva DNA obtained from the self administered oragene method—a pilot study on the cohort of Swedish men. Rylander-Rudqvist et al.
- <sup>2</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>3</sup> *Nature Genetics*; 44(6) May 2012. Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. Jin et al.
- <sup>4</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.