illumina®

Improved Human Mitochondrial DNA Analysis using Next-generation Sequencing and Cloud-based Computing and Storage

Carey Davis, Kevin Rhodes, Nathalie Mouttham, Anthony Rensfield, Aprajita Mathur, Dan Sa, Van Le-Pham, Amanda Young, Joseph Varlaro, John Walsh

SAMPLE PREP

SEQUENCE (MISEQ FGx®)

PROCESS (mtDNA VARIANT PROCESSOR)

ANALYZE (mtDNA VARIANT ANALYZER)

REPORT (EXCEL)

BaseSpace®
Genomics Cloud Computing
BaseSpace®
Genomics Cloud Computing
MiSeq FGx
Welcome
Variant Summary Report

INTRODUCTION

Mitochondrial DNA (mtDNA) analysis has been applied in a number of fields [1-4]. Since mtDNA is found in higher copy number per cell than nuclear DNA, it has been an invaluable genetic marker in forensics with samples that are often limited in quantity and quality [5-7].

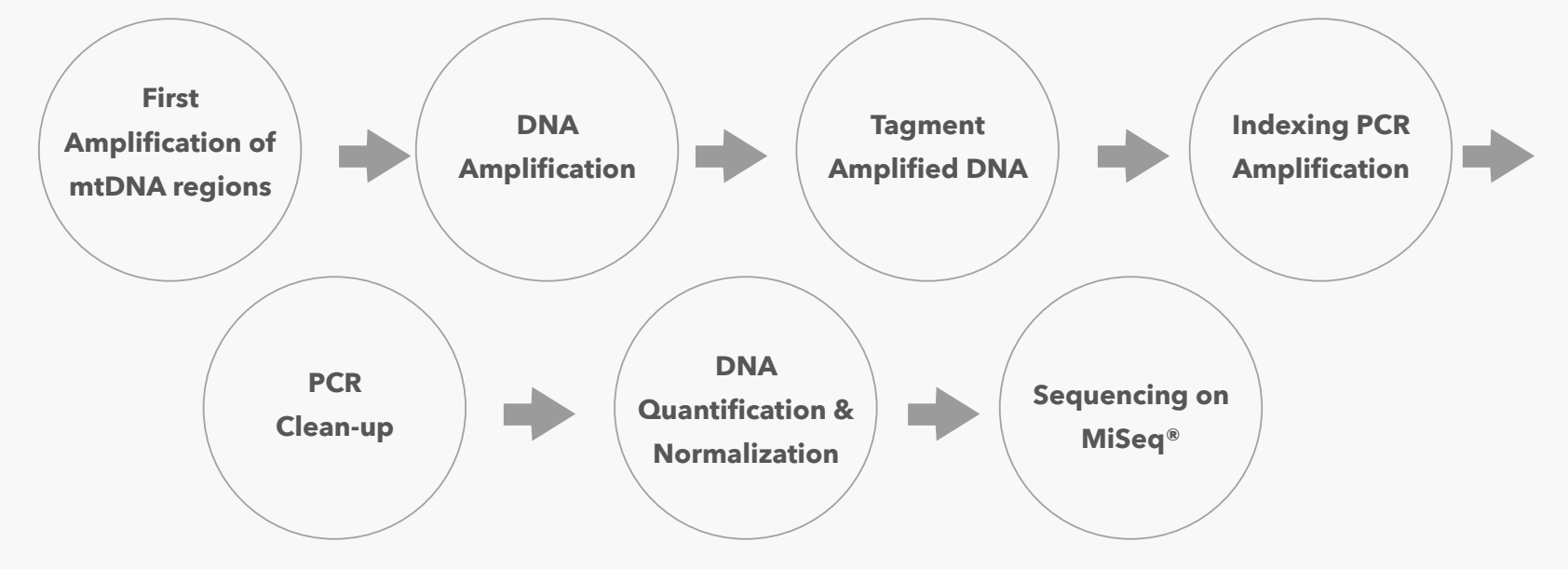
Massively parallel sequencing (MPS) is a high throughput technology that can rapidly generate high quality sequence from targeted areas of the human genome. The technology has reached a level of robustness such that it can be considered a viable approach to analyze challenging forensic samples. Illumina previously released two protocols for sequencing mtDNA (whole mt genome and d-loop region) based on Nextera® XT DNA Sample Preparation Kit and the MiSeq FGx™ platform, and published findings in King et al. [8].

Now, Illumina is adding analysis workflow applications on Illumina's cloud-computing environment, BaseSpace™, specifically suited to the mtDNA samples of the Forensic Genomics community. After streaming sequencing data to BaseSpace from your MiSeq FGx™, the data is readily available for data analysis and visualization with the apps. The mtDNA Variant Processor is used to align and variant call the sample data using a workflow optimized for mtDNA samples and the Forensic Genomics community.

Following processing, the results are viewed in the mtDNA Variant Analyzer which provides a modern user interface that enables the user. Multiple samples can be inspected in aggregate, or compared to each other with a responsive experience. When the analyst's trained eye disagrees with the computed call for a locus, the call can be changed and saved back to the cloud. An output report summarizes the variants across samples as well as providing more detailed information per sample when necessary.

The mtDNA work presented herein represents an end-to-end solution for the preparation, analysis, visualization, and reporting supporting today's forensic scientists.

SAMPLE PREP



First Amplification of mtDNA regions → DNA Amplification → Tagment Amplified DNA → Indexing PCR Amplification → PCR Clean-up → DNA Quantification & Normalization → Sequencing on MiSeq®

Figure 1. High-level user workflow from sample preparation to reporting of mtDNA samples

Samples

* Buccal swabs from volunteers

* Extracted using QIAamp® DNA Investigator kit (Qiagen, Hilden, Germany) according to manufacturer's instructions

* Quantity, size and quality of libraries assessed with Agilent DNA 1000 kit on the Agilent 2100 Bioanalyzer system

* Normalized PCR amplification products prepared using the Nextera® XT DNA Sample Preparation kit

* Samples were indexed to allow for pooling and demultiplexing

* Sequencing performed on a MiSeq FGx with MiSeq Reagent Kit v3

* Data streamed to BaseSpace cloud computing platform during sequencing

D-Loop Protocol

* Human mtDNA D-Loop Hypervariable Region Guide [i]

* Sequenced using a 2 x 151 cycle run with dual index reads

Amplicon Start	Amplicon End
29	285
172	408
15997	16236
16159	16401

Table 1. Amplicon locations for the D-Loop Protocol

Whole Genome Protocol

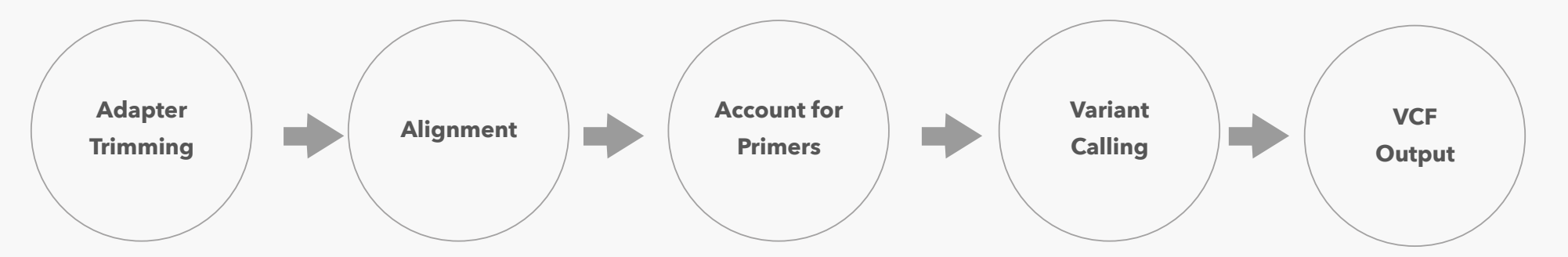
* Human mtDNA Genome Guide [ii]

* Sequenced using a 2 x 251 cycle run with dual index reads

Amplicon Start	Amplicon End
9397	1892
15195	9796

Table 2. Amplicon locations for the Whole Genome Protocol

mtDNA VARIANT PROCESSOR



Adapter Trimming → Alignment → Account for Primers → Variant Calling → VCF Output

Figure 2. Workflow of the bioinformatic data analysis pipeline employed by the mtDNA Variant Processor

Alignment and variant calling was performed using a bioinformatics pipeline developed and optimized for mtDNA samples from the Forensic Genomics community. This workflow is available as the mtDNA Variant Processor app on Illumina's BaseSpace cloud computing platform. The app supports the processing of both the mtDNA D-Loop and the whole genome samples prepared as specified above. Additionally, custom primers can be used if provided during analysis set-up.

Nextera® Adapter Trimming

* Nextera® adapters are removed from the forward and reverse reads

* Trimming is repeated until no more than 3 adapter bases found on the end of the read

* Reads with excessive trimming or result in very short amplicons are discarded entirely

Alignment

* Alignment is performed with BWA-MEM [9]

* Parameters optimized for single nucleotide homopolymeric regions (i.e. C-stretches)

* Circular alignment is handled across the origin by identifying the true start and end of reads

* Indels are realigned to improve alignment and shift to a 3' alignment, particularly important in C-stretches

Account for Primers

* The supplied manifest is used to identify the primers and amplicons

* Primer contributions are removed from the reads for accurate variant calling

Variant Calling

For each position, data is piled up to aid in identifying a consensus call for the position

* Bases are quality score filtered prior to use in calling

* A minimum read count is used to avoid positions with practically no coverage or noise

* A score is calculated for each called position that accounts for the proportion of reads below BaseQ, MapQ, and Analysis Treshold

* Calls and reporting of alleles are strictly threshold based:

- Percentage of reads for a nucleotide exceeds the Analysis Threshold (AT), it is reported in the output
- Percentage of reads for a nucleotide exceeds the Interpretation Threshold (IT), it is used in the call

* Ambiguous indels are reported when there is significant data (exceeds the IT) that supports the indel and doesn't support it (See Figure)

Code	Definition
A	Adenine
C	Cytosine
T	Thymine
G	Guanine
R	A and G
Y	C and T
S	G and C
W	A and T
K	G and T
M	A and C
B	C and G and T
D	A and G and T
H	A and C and T
V	A and C and G
N	A and C and G and T
-	Deletion
	No Call

Table 3. IUPAC nucleotide codes for unambiguous consensus calls

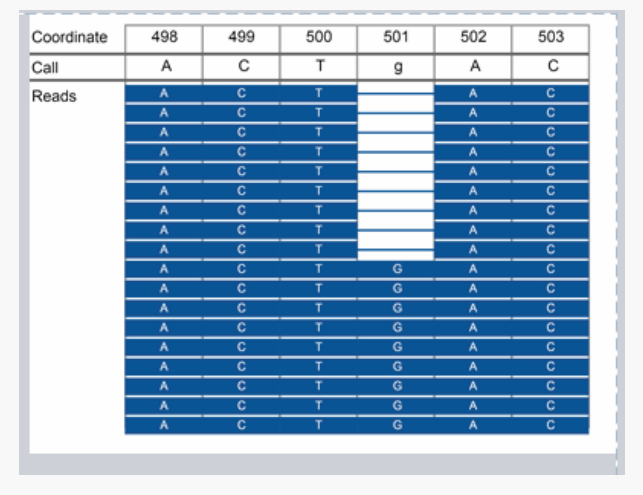


Figure 3. An example illustrating the presence of an ambiguous deletion at position 501. There are an equal number of reads supporting a deletion call and a reference call of G so we report an ambiguous call of g as the consensus.

Code	Definition
a	Ambiguous indel with adenine
c	Ambiguous indel with cytosine
t	Ambiguous indel with thymine
g	Ambiguous indel with guanine
r	Ambiguous indel with A and G
y	Ambiguous indel with C and T
s	Ambiguous indel with G and C
w	Ambiguous indel with A and T
k	Ambiguous indel with G and T
m	Ambiguous indel with A and C
b	Ambiguous indel with C and G and T
d	Ambiguous indel with A and G and T
h	Ambiguous indel with A and C and T
v	Ambiguous indel with A and C and G
n	Ambiguous indel with A and C and G and T

Table 4. Codes for ambiguous consensus calls of insertions and deletions

Output

* VCF file formatted with extra data for use with the mtDNA Variant Analyzer

* BAM file for investigation of the alignment with third party tools (e.g. IGV)

mtDNA VARIANT ANALYZER

The results of the mtDNA Variant Processor are designed to enable data viewing and interrogation in the mtDNA Variant Analyzer BaseSpace app. The interface of the app was specifically designed to simplify navigation throughout the genome while providing all relevant information to understand the data. Additionally, the mtDNA Variant Analyzer app enables simultaneous data analysis across runs and projects for the side-by-side comparison of results collected over time or within a run.

Genome Visualizations

* Coverage is shown across the relevant area of the genome and can be zoomed in or out

* Variant and interpretation threshold warnings are shown contextually against the genome

Comparison to rCRS

* Each sample's consensus genome is displayed against the rCRS genome

* Navigation options allow to skip to just hotspots (e.g. variants, interpretation threshold warnings, modified calls)

* Different view filters enable the restriction of the view to just hotspots or the D-loop

Comparison to other samples

* A sample can be selected as the reference and viewed against all other samples loaded

* Comparison of two samples will show the coverage and call differences for inspection

* All results that don't match the reference sample can be quickly hidden from view

Position level information

* For each called position, the coverage, score, and nucleotide distribution can be inspected

* When comparing samples, the reference and selected sample are shown together

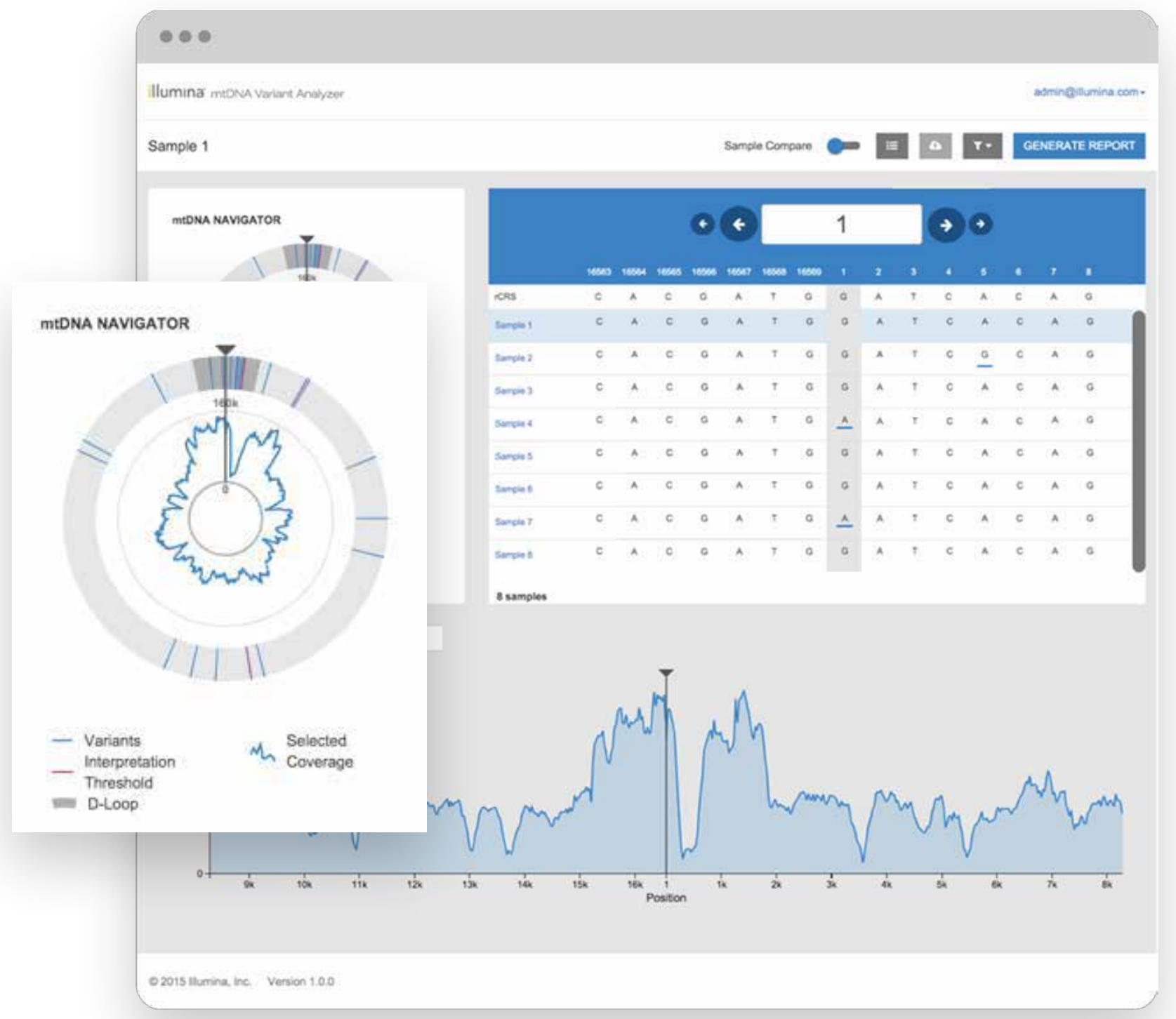
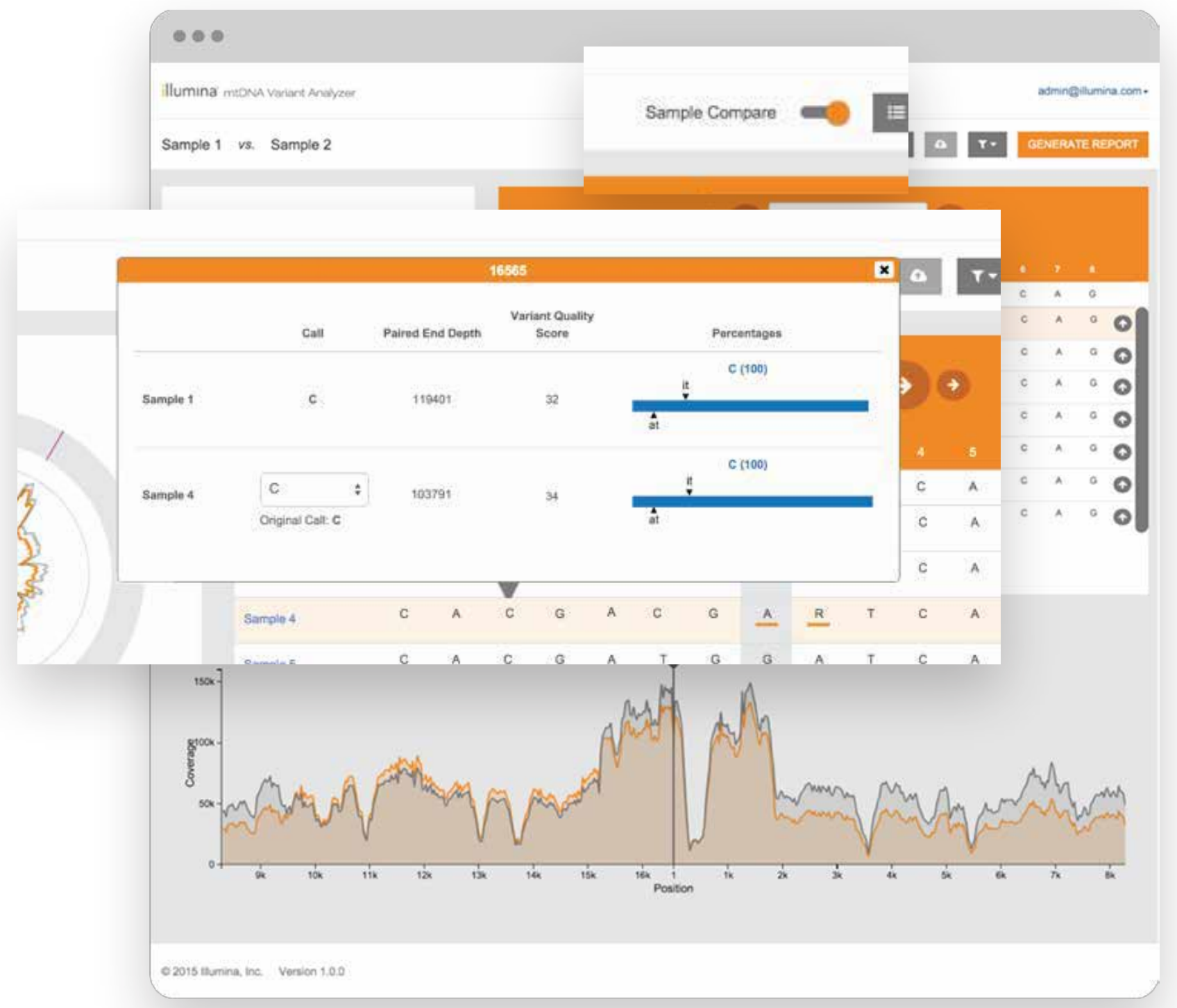


Figure 4. The mtDNA Variant Analyzer with the mtDNA Navigator enlarged for emphasis. The mtDNA Navigator gives visual indications of variants, differences, interpretation threshold indicators, and coverage across the genome.



illumina mtDNA Variant Analyzer
Sample 1 vs. Sample 2
Sample Compare
GENERATE REPORT
Position
© 2015 Illumina, Inc. Version 1.0.0

Figure 5. The mtDNA Variant Analyzier in the Sample Compare mode with a position's details enlarged for emphasis. Information about coverage, score, call, thresholds used for analysis, and the distribution of alleles at each position are available for inspection. When comparing to other samples, this information is provided for both the reference and selected sample.

REFERENCES

1] S. Bannwarth, V. Procaccio, A.S. Lebre, C. Jardel, A. Chaussenot, C. Hoarau, et al. Prevalence of rare mitochondrial DNA mutations in mitochondrial disorders. J Med Genet. 50 (2013) 704-14.
[2] J. Nunnari, A. Suomalainen. Mitochondria: in sickness and in health. Cell. 148 (2012) 1145-59.
[3] T. Kivisild, M. Reidla, E. Metspalu, A. Rosa, A. Brehm, E. Pennarun, et al. Ethiopian Mitochondrial DNA Heritage: Tracking Gene Flow Across and Around the Gate of Tears. The American Journal of Human Genetics. 75 (2004) 752-70.
[4] M. Richards, V. Macaulay, A. Torroni, H.-J. Bandelt. In search of geographical patterns in European mitochondrial DNA. The American Journal of Human Genetics. 71 (2002) 1168-74.
[5] P. Gill, P.L. Ivanov, C. Kimpton, R. Piercy, N. Benson, G. Tully, et al. Identification of the remains of the Romanov family by DNA analysis. Nat Genet. 6 (1994) 130-5.
[6] M.M. Holland, T.J. Parsons. Mitochondrial DNA sequence analysis-validation and use for forensic casework. Forensic Sci Rev. 11 (1999) 21-50.
[7] M.R. Wilson, J.A. DiZinno, D. Polanskey, J. Replogle, B. Budowle. Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Leg Med. 108 (1995) 68-74.
[8] J. L. King, B. L. LaRue, N. Novroski, M. Stoljarova, S. B. Seo, X. Zeng, et al., High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq, Forensic Sci. Int. Genet. 12 (2014) 128-135.
[9] Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

LINKS

[i] Illumina http://support.illumina.com/downloads/human_mtdna_d_loop_hypervariable_region_guide_15034858.ilmn
[ii] http://support.illumina.com/downloads/human_mtdna_genome_guide_15037958.html