A tale of two platforms: An evaluation of the Roche GS Junior and Illumina® MiSeq next-generation sequencing instruments for forensic mitochondrial DNA analysis

Brittania J. Bintz, M.S.; Erin S. Burnside, M.S.; Mark R. Wilson, Ph.D.

Forensic Science Program, Department of Chemistry and Physics – Western Carolina University

ABSTRACT

Next-generation sequencing (NGS) refers to a suite of technologies that enable cost-effective, rapid generation of large amounts of detailed sequence information from clonal populations of individual template molecules. These methods are proving to be particularly well-suited for mitochondrial DNA analysis, and may provide forensic DNA analysts with a powerful tool that enables deconvolution of mtDNA mixtures. Recently, Illumina® has been working with members of the community to establish a human mtDNA forensic genomics consortium (HGC) for concerted evaluation of NGS methods for potential use in mtDNA casework and databasing. In June, a set of samples was prepared consisting of quantified buccal extracts from two donors, as well as a series of mixtures of the buccal extracts at defined ratios (5, 2, 1 and 0.5%). This sample set has been distributed to participating HGC laboratories for sequencing on multiple NGS platforms including the Ion PGM™, Roche GS Junior, and Illumina® MiSeq, to enable a cross-laboratory comparison of sequencing methods using identical samples. In our laboratory, the samples were sequenced on both the Roche GS Junior, and Illumina® MiSeq NGS platforms. Libraries from hypervariable regions 1 and 2 (HV and HV2) were sequenced on the Roche GS Junior using an amplicon library preparation approach where PCR primers were designed to include required adaptors and multiplexing indices. For sequencing on the Illumina® MiSeq, libraries were prepared using Nextera® XT in which two large amplicons covering the whole mtGenome as well as HV1 and HV2 amplicons were randomly fragmented, and adapters and indices incorporated enzymatically. The resulting data was analyzed using CLC Genomics Workbench software and variant calls were compared. The Illumina® MiSeq resulted in significantly higher coverage across all positions sequenced, giving rise to higher certainty with low-level variant calls. Further, the MiSeq allowed for detection of minor variants in all mixtures where the majority of minor variants were undetected in the 0.5% mixture with the Roche GS Junior. Finally, data from the MiSeq showed lower background noise overall, especially in homopolymeric regions when compared to data from the GS Junior. The Illumina® MiSeq offers a streamlined enzymatic library preparation approach, higher-throughput and more accurate variant detection and basecalling than the Roche GS Junior. As a result, we feel that the MiSeq is better suited for forensic mtDNA analysis in both casework and databasing laboratories.

SAMPLE SET

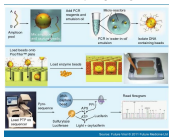
Buccal swabs (x20) were obtained from two donors (001-CF30 and 003-CM54) whose whole mtGenome had been previously characterized in our laboratory using Sanger methods. The swabs were collected according to approved IRB protocol.

DNA from each set of 20 buccal swabs was extracted independently using the Qiagen DNA Investigator kit surface and buccal swab protocol. A single RB was extracted alongside each set. The resulting extracts from each donor were pooled to create a large volume master extract. Pooled samples were quantified in quintuplicate using a human mtDNA specific real-time PCR assay developed by Mark Kavlick at the F.B.I.[1] Quintuplicate values were averaged after outliers were removed. Averages were used to prepare mixtures of donors in defined ratios of 5, 2, 1 and 0.5%. Donor 003-CM54 was used as the minor contributor in all mixtures. Final mixtures and sole source samples were quantified in quintuplicate using both the Quantifiler® Human kit, and the human mtDNA specific real-time PCR assay mentioned above. All quantified samples were distributed to HGC laboratories for sequencing as requested.

001 – CF30 Sole Source	003 – CM54 Sole Source	5% 003-CM54	2% 003-CM54	1% 003-CM54	0.5% 003-CM54	Reagent Blanks	9947A Positive Control

SEQUENCING CHEMISTRIES

Roche 454 Pyrosequencing



http://img.medscape.com/article/738/786/738786-fig2.jpg

Illumina® Reversible Terminator Sequencing

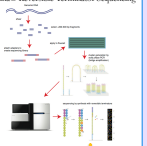
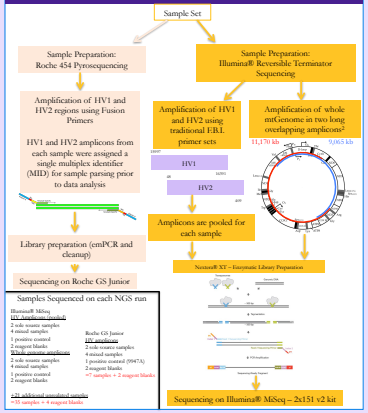


Figure 1: *Chemistry and workflows of leading NGS instruments.*

On the left, Roche 454 Pyrosequencing chemistry is presented. Initially, DNA target regions are amplified using PCR with a high-fidelity polymerase, and modified primers containing 454 adapter sequences and multiplexing barcodes 5' of a template specific sequence. Multiplexing barcodes enable post-run sample parsing, while adapters allow for clonal amplification on the surface of a bead, which is then loaded onto a PicoTiter™ plate (PTP) where sequencing chemistry occurs. Native dNTPs are flowed iteratively across the surface of the PTP. Release of PPi following incorporation initiates an enzymatic cascade that ends in the ATP catalyzed conversion of luciferin to oxyluciferin, and the production of light. The amount of light produced is directly related to the number of nucleotides in the region being sequenced.

On the right, Illumina® chemistry is shown. DNA target regions are amplified using traditional PCR, and a high-fidelity polymerase. The resulting amplicons are fragmented enzymatically using the Illumina® Nextera® XT tagmentation kit, which also enables incorporation of sequencing adaptors and multiplexing barcodes during a limited cycle PCR step. Libraries are then added to an optically transparent flowcell, where they bind to anchored oligonucleotides complementary to their incorporated adaptor sequences. Clusters consisting of clonal populations of individual template molecules are generated on the instrument. Sequencing takes place after cluster generation when modified dNTPs with base specific fluors, and blocking groups on their 3' -OH groups are flowed across the flowcell. After each cycle, LED lights excite nucleotide specific fluors, and the software images the surface of the flowcell to catalogue the specific base incorporated at each cluster.

NGS LIBRARY PREPARATION



Samples Sequenced on each NGS run

Illumina® MiSeq
HV Amplicons (pooled)
2 sole source samples
4 mixed samples
1 positive control
2 reagent blanks
Whole genome amplicons
2 sole source samples
4 mixed samples
1 positive control (9947A)
2 reagent blanks
+7 samples + 2 reagent blanks

Roche GS Junior
HV amplicons
2 sole source samples
4 mixed samples
1 positive control (9947A)
2 reagent blanks
+7 samples + 2 reagent blanks

+40 additional unrelated samples
+51 samples + 4 reagent blanks

DATA ANALYSIS – CLC GENOMICS WORKBENCH

CLC Genomics Workbench v6.5 was used for all data analyses. Raw data files (.SFF files) from the Roche GS Junior were uploaded into CLC Genomics Workbench, and demultiplexed using the software. Fastq files demultiplexed during secondary analysis on the Illumina® MiSeq were also uploaded. All sample files were analyzed using the same pipeline. The data was initially mapped to the revised Cambridge reference sequence (rCRS)[3] using a local alignment option. Variant calling was performed using the quality-based variant detection method with a **0.1% minor variant detection threshold** to enable capture of a majority of sites showing variability. Resulting variant tables were exported as tab delimited text files, and uploaded into the Galaxy[4] open-source cloud computing environment. A count application within Galaxy was applied to the data set to assist with categorization of unexpected variants. Full analysis parameters, and raw data files are available upon request.

EXPECTED VARIANT FREQUENCIES – SOLE SOURCE SAMPLES

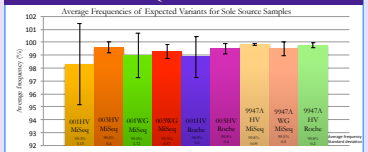
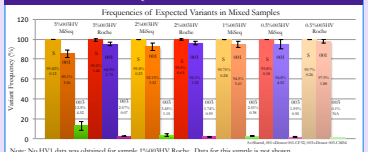


Figure 2: *Bar graph showing average frequencies for expected variants in HV1 and HV2, detected in sole source samples.* The expected variants are those that were previously detected for each donor using Sanger sequencing, and are expected be detected at or near a frequency of 100% unless otherwise noted. Averages of reported frequencies and accompanying standard deviations were calculated for the expected variants of each donor. Error bars represent standard deviations of frequency averages. Donor 001-CF30 is expected to possess a total of 9 variants from the rCRS across the HV region, with varying levels of heteroplasmy at position 16093 (data not shown). This heteroplasmy is contributing to higher standard deviation values for all donor 001 samples. Donor 003-CM54 was expected to possess a total of 14 variants from the rCRS in the HV region. Differences in alignments were often observed at positions 150C, 152T, 16193C and 16195T, expected variants for donor 003-CM54 (figure 4). When this was true, frequencies for these variants were not used to calculate frequency averages and standard deviations (omitted: 150/152-003HV MiSeq 150/152, 16193/16195-003HV Roche). The 9947A control was expected to possess 5 variants from the rCRS, with no apparent heteroplasmy in the HV region.

EXPECTED VARIANT FREQUENCIES IN HV REGION – MIXED SAMPLES



Note: No HV1 data was obtained for sample 1%003HV Roche. Data for this sample is not shown.

Figure 3: *Bar graph showing average frequencies for expected variants in HV1 and HV2, detected in mixed samples.* The expected variants are those that were previously detected for each donor using Sanger sequencing. All variants shared between donors are expected to be detected at or near a frequency of 100%, barring haplotypmatic sites. Conversely, depending on the mixture, the donor specific variant frequencies change (95:5; 98.2, 99.1, and 99:50.5% respectively) with donor 003-CM54 representing the minor contributor in all mixtures. In the HV region, donor 001-CF30 and donor 003-CM54 share a total of 5 variants. In addition, donor 001-CF30 possesses 4 unshared variants from the rCRS, and donor 003-CM54 possesses 8. Similar to the data presented in figure 2, heteroplasmy for donor 001-CF30 at position 16093, is contributing to higher than expected standard deviations in frequency averages. Also, frequencies for variants with observed alignment anomalies were omitted from calculations in this data set (omitted: positions 150C/152 from all donor 003 samples and positions 16193/16195 at 5%003HV and 1%003HV data generated with the MiSeq).

In all MiSeq HV sample data, frequencies of minor contributor variants were higher than expected. This was not observed in data generated using the Roche GS Junior. This was also observed to a much lesser extent in whole genome data also generated using the Illumina® MiSeq (see figure 4). This is likely a result of differences in library preparation, since the same samples were used for all sequencing runs. Additionally, in the 0.5% mixture, all minor contributor variants dropped out (were below the 0.1% minor variant frequency threshold) except 1. This is a direct result of the higher coverage achieved with the Illumina® MiSeq, and shows that the MiSeq is more sensitive than the Roche GS Junior.

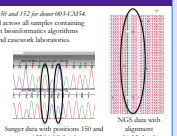
UNEXPECTED VARIANTS IN HV DATA – SUMMARY

	001HV		001HV		3%003HV		2%003HV		1%003HV		0.5%003HV		9947AHV	
	M	R	M	R	M	R	M	R	M	R	M	R	M	R
A>G, G>A	112	56	92	82	94	53	92	57	107	NA	102	51	63	30
C>T, T>C	117	66	118	73	99	67	107	71	108	NA	103	72	78	46
Transversions	5	15	3	2	3	0	0	2	2	NA	4	1	2	10
Insertions	10	16	10	17	15	15	14	17	14	NA	15	13	9	11
Deletions	15	31	14	37	13	28	14	37	15	NA	13	27	9	27
MNVs (multi-nucleotide variants)	5	5	1	5	1	7	0	5	1	NA	2	5	1	1
Total	264	187	238	216	225	170	227	187	247	NA	239	167	162	125
Average Coverage	2,366	121	10,252	121	10,919	122	10,967	126	11,127	NA	14,181	111	8,109	118

Table 1: *Summary of characterized unexpected variants from HV1 and HV2 samples in Roche GS Junior and Illumina® MiSeq data.* All unexpected variants were counted using Galaxy open-source cloud computing software. The resulting counts were manually categorized and tabulated. In all cases, when compared to the average coverage calculated for each sample, the number of unexpected variants in Illumina® MiSeq data sets. This indicates that the Roche GS Junior yields more noise than the Illumina® MiSeq. However, additional research is needed to further characterize the unexpected variants as noise. The 9947A control, derived from a human cell line possessed fewer numbers of unexpected variants when compared to the samples obtained from buccal swabs.

ALIGNMENT DIFFERENCES

Figure 4: *Alignment differences in Sanger and NGS data at positions 150 and 152 for donor 003-CM54.* Differences in alignment at positions 150 and 152 were detected across all samples containing DNA from donor 003-CM54. Alignment discrepancies between bioinformatics algorithms must be considered when implementing NGS into databasing and casework laboratories.



WHOLE GENOME SAMPLES – COVERAGE MAPS

Figure 5: *Coverage maps for Illumina® MiSeq whole genome data (donors 001-CF30 and 003-CM54) obtained from Illumina® MiSeq Reporter secondary analysis software.* Coverage across the entire mtGenome was achieved using a long PCR approach developed in our laboratory. Q-scores are high (≥30) throughout the length of the genome.

UNEXPECTED VARIANTS IN WHOLE GENOME DATA - SUMMARY

	001WG	003WG	3%003WG	2%003WG	1%003WG	0.5%003WG	9947AWG
A>G, G>A	3384	4064	3428	3341	3269	1775	2072
C>T, T>C	3330	4010	3477	3400	3255	1776	3116
Transversions	47	324	67	51	89	14	78
Insertions	40	48	45	42	54	36	23
Deletions	210	516	219	180	269	118	205
MNVs (multi-nucleotide variants)	26	26	29	27	59	29	20
Total	7057	8788	7263	7041	6955	3748	6412
Average Coverage	15,477	9,020	10,274	9,270	14,326	14,151	13,294

Table 2: *Summary of unexpected variant characterizations for whole genome samples from Illumina® MiSeq data.* Average coverage for whole genome samples is high despite being sequenced in the same run with 7 samples consisting of pooled HV amplicons, 21 unrelated samples consisting of 2 pooled amplicons, and controls. The types of variants seen across the whole mtGenome sequenced using the Illumina® MiSeq are consistent. No whole genome data was obtained from the Roche GS Junior, so no direct instrumentation comparisons can be made with this data.

EXPECTED VARIANT FREQUENCIES – WHOLE GENOME DATA

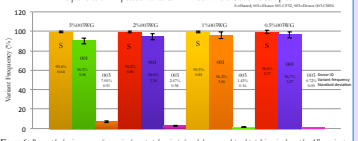


Figure 6: *Bar graph showing average frequencies for expected variants for whole genome data, detected in mixed samples.* All variants shared between donors are expected to be detected at a frequency of 100%. However, depending on the mixture, the donor specific variant frequencies will change (95:5; 98.2, 99.1, and 99:50.5% respectively) with donor 003-CM54 representing the minor contributor in all mixtures. Frequency averages and standard deviations were calculated for the expected variants of each donor. Error bars represent standard deviations of frequency averages. Across the genome, donor 001-CF30 and donor 003-CM54 share a total of 21 variants. In addition, donor 001-CF30 possesses 7 unshared variants from the rCRS, and donor 003-CM54 possesses 18. Frequencies for variants with observed alignment anomalies were omitted from calculations in this data set as well (omitted: 150/152 in all samples; 16193/16195-5%003WG). Furthermore, several data points were not detected in 0.5%003WG. A large portion of data is missing from this set. More research is being conducted to determine the cause. Overall, the frequencies of the expected variants seem to more accurate when using our whole genome approach, than when using an small amplicon approach (figure 3).

CONCLUSIONS

- The Illumina® MiSeq allows for rapid enzymatic library preparation, higher throughput, and longer read-lengths than the Roche GS Junior.
- Much higher coverage is achieved per sample when using the Illumina® MiSeq versus the Roche GS Junior, even with additional whole genome samples included in the MiSeq run. As a result, we feel that the MiSeq can result in higher sensitivity depending on the design of the run, and lead to detection of minor variants at a lower frequency than the GS Junior.
- Additionally, it appears that the Roche GS Junior is more prone to noise than the Illumina® MiSeq, especially in areas containing homopolymers. However, this observation is not novel. Additional work is being done to reanalyze the data from the Roche GS Junior with filtering mechanisms designed to remove sequencing artifacts associated with homopolymers. Also, unexpected variants which were labeled as "noise" for this work, are being further studied to identify positions that may be consistent with true biological variation.
- The types of unexpected variants observed in data obtained using both instruments is consistent across samples sequenced.
- Overall, we feel that the Illumina® MiSeq is well-suited for forensic mtDNA analysis though additional validation studies are required to move this technique into the crime laboratory.

REFERENCES

1. M.F. Kavlick, H.S. Lawrence, R.T. Merritt, C. Fisher, A. Isenberg, J.M. Robertson, B. Budowle. Quantification of human mitochondrial DNA using synthesized DNA standards. J Forensic Sci. 2011:56(6):1457-63.
2. R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genetics. 1999:23:147.
3. H. Stawski, B.J. Bintz, E.S. Burnside, M.R. Wilson. Preparing whole genome human mitochondrial libraries using Illumina® Nextera® XT. Proceedings of AAFS. 2013.
4. Public Galaxy Instance. [http://usegalaxy.org].

ACKNOWLEDGMENTS

The authors of this work wish to gratefully acknowledge Illumina® for their continuing support. Special thanks to Cynde Holt, Joe Varlaro, Joe Walsh, Kathryn Stevens, Carey Davis and Tom Richardson. We also wish to thank Western Carolina University for supporting some of this work.

CONTACT

Brittania J. Bintz, M.S.
Forensic Science Program
Western Carolina University
111 Memorial Drive
NSB #231
Cullowhee, NC 28723
bbintz@wcu.edu
828.227.3680