Mitochondrial Mixture Analysis with Pacific Biosciences Data Using NextGENe®LR Software

November 2020

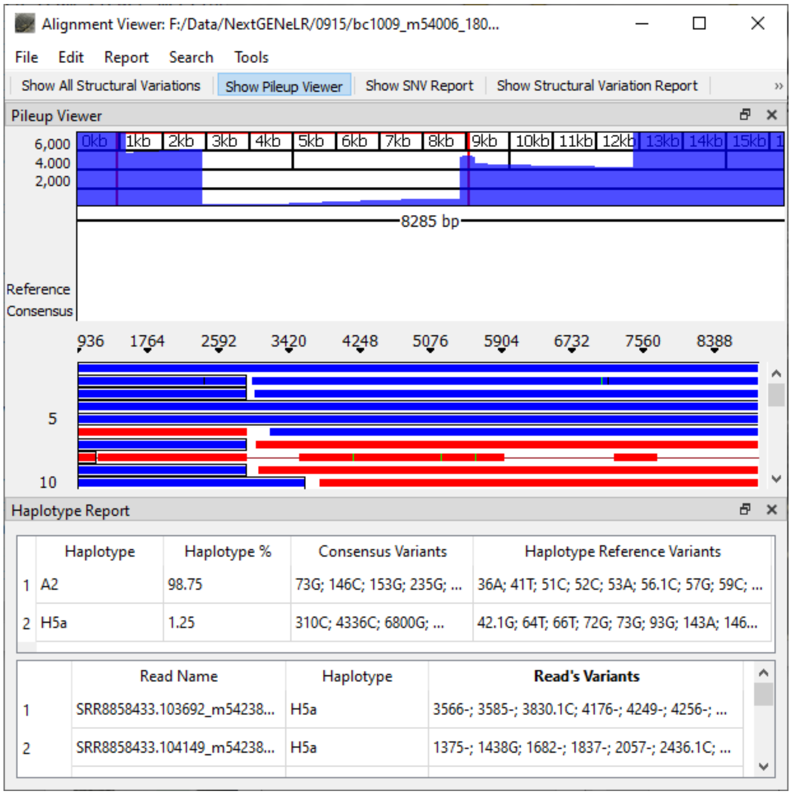
Megan McCluskey, Jacie Wu, Lidong Luo, Jonathan Liu

Introduction

Mitochondrial DNA haplotyping can be utilized as a valuable tool for studying phylogeny (1), phenotypic information related to disease (2), and for identification in forensic applications (3). Mitochondrial DNA is particularly useful for forensic applications when working with samples that have been degraded (4) or contain mixtures (5). Mitochondrial DNA is helpful for identification from degraded samples as each cell can contain thousands of copies of the mitochondrial genome. Forensic identification from mixed DNA samples generally presents a greater challenge than identifying individual samples. However, the numerous copies of mitochondrial DNA in each cell also lends itself to deciphering mixed samples by allowing low copy number specimens to be observed.

Pacific Biosciences SMRT (single molecule, real-time) sequencing, produces reads of dozens of kilobases in length, enabling the sequencing of up to the entire mitochondrial genome in one pass (6). These long read lengths facilitate phasing of variants to allow for haplotyping of the mitochondrial genome (7). Since variants within the same read can be concluded to be from the same mtDNA molecule, utilizing individual reads spanning the full mitochondrial genome can provide a straight-forward determination of a complete haplotype and therefore to identify the mixture of mitochondrial samples.

NextGENeLR allows the accurate haplotyping of long read data such as data from Pacific BioSciences RS, Sequel, and Sequel II Systems by implementing accurate alignment and comparing aligned reads to reference haplotypes. NextGENeLR's algorithms also facilitate distinguishing distinct haplotypes from mixed samples, identifying major and minor haplotypes present. Haplotypes can also be discerned from mixture samples, with detection as low as 0.1% frequency.



	Haplotype	Haplotype %	Consensus Variants	Haplotype Reference Variants
1	A2	98.75	73G; 146C; 153G; 235G; ...	36A; 41T; 51C; 52C; 53A; 56.1C; 57G; 59C; ...
2	H5a	1.25	310C; 4336C; 6800G; ...	42.1G; 64T; 66T; 72G; 73G; 93G; 143A; 146...

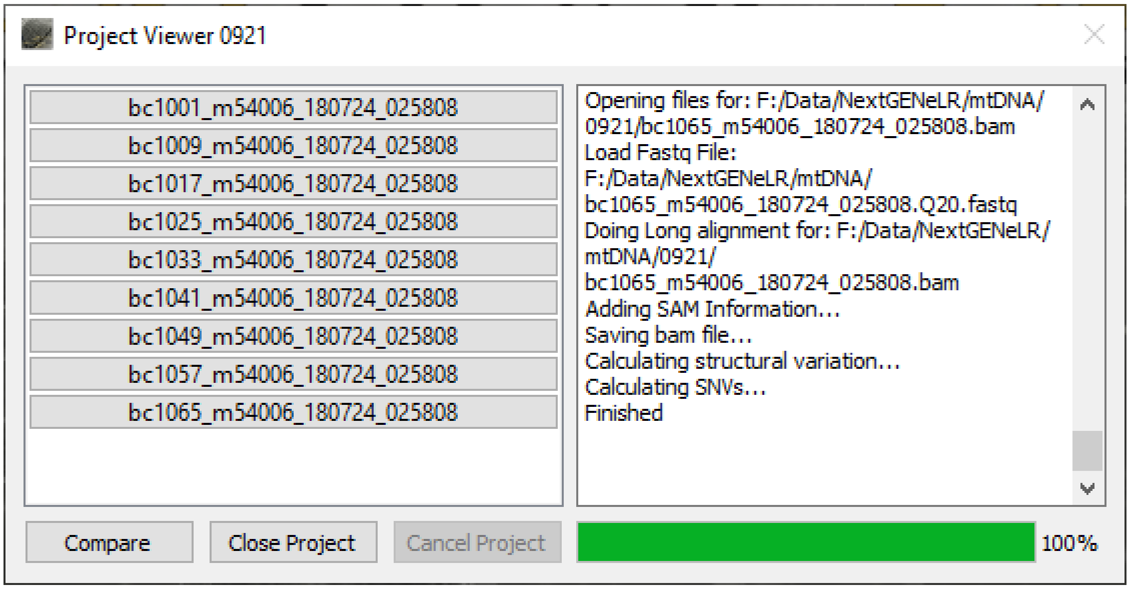
	Read Name	Haplotype	**Read's Variants**
1	SRR8858433.103692_m54238...	H5a	3566-; 3585-; 3830.1C; 4176-; 4249-; 4256-; ...
2	SRR8858433.104149_m54238...	H5a	1375-; 1438G; 1682-; 1837-; 2057-; 2436.1C; ...

Figure 1: NextGENeLR Alignment Viewer with Haplotype Report displayed. The Haplotype Report for this sample indicates a mixture is present with the minor haplotype present at 1.25% frequency. The bottom section of the report lists the haplotype and variants present for each individual read in the sample.

Method

NextGENeLR employs a novel algorithm for accurate and fast alignment of long read sequences to a reference. For mitochondrial DNA analysis, a chrM fasta reference file of the Cambridge Reference Sequence is used. The reference sequence is indexed, with homopolymer sequences compressed. Reads are similarly indexed and then compared to the reference index table. Matching sections are identified and then a local alignment is performed to complete alignment of the reads. High and low frequency variants are identified and major and minor variant lists are saved. The variant lists are compered to the mitochondrial ancestor haplotype database from MitoMap, which is built-in to NextGENeLR software. The best matching haplotype(s) are then reported. Each read is also assigned a haplotype based on the best match to the reported haplotypes. Haplotypes at frequencies as low as 0.1% can be reported.

Multiple samples can be loaded in batch to quickly process each sample consecutively. Individual results are generated for each sample, and these results are saved together as a project that facilitates viewing the results for each sample.


Project Viewer 0921
bc1001_m54006_180724_025808
bc1009_m54006_180724_025808
bc1017_m54006_180724_025808
bc1025_m54006_180724_025808
bc1033_m54006_180724_025808
bc1041_m54006_180724_025808
bc1049_m54006_180724_025808
bc1057_m54006_180724_025808
bc1065_m54006_180724_025808
Compare
Close Project
Cancel Project
100%

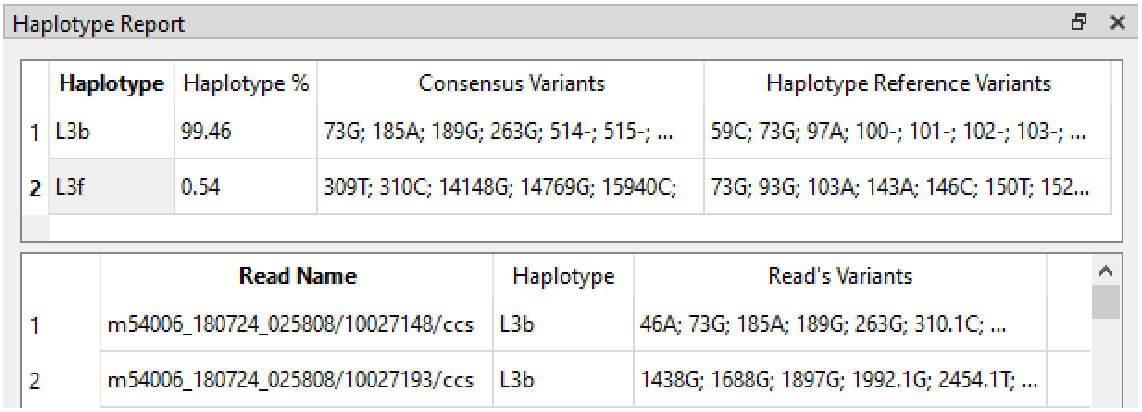
```
Opening files for: F:/Data/NextGENeLR/mtDNA/
0921/bc1065_m54006_180724_025808.bam
Load Fastq File:
F:/Data/NextGENeLR/mtDNA/
bc1065_m54006_180724_025808.Q20.fastq
Doing Long alignment for: F:/Data/NextGENeLR/
mtDNA/0921/
bc1065_m54006_180724_025808.bam
Adding SAM Information...
Saving bam file...
Calculating structural variation...
Calculating SNVs...
Finished
```

Figure 2: NextGENeLR Project Viewer is displayed once an analysis is completed. Results for all loaded samples can be easily accessed and reviewed by clicking the button with the sample name.

Results

Following alignment NextGENeLR outputs a BAM file which can be reviewed using the built-in Alignment Viewer. By clicking "Show Haplotype Report" the Haplotype Report is displayed at the bottom of the Viewer. Above the Haplotype Report a Pileup Viewer is displayed showing the coverage distribution at the top with the read pileup shown below.

The Haplotype Report consists of two components. The upper section provides the Haplotypes detected for the sample. For each haplotype, the percentage of reads in the sample that match this haplotype is shown, along with a list of the variants observed in the consensus for the sample as well as the list of variants for the haplotype reference. The lower section provides a list of all reads included in the sample, displaying the assigned haplotype and a list of the variants present for the read.



Haplotype Report

	Haplotype	Haplotype %	Consensus Variants	Haplotype Reference Variants
1	L3b	99.46	73G; 185A; 189G; 263G; 514-; 515-; ...	59C; 73G; 97A; 100-; 101-; 102-; 103-; ...
2	L3f	0.54	309T; 310C; 14148G; 14769G; 15940C;	73G; 93G; 103A; 143A; 146C; 150T; 152...

	Read Name	Haplotype	Read's Variants
1	m54006_180724_025808/10027148/ccs	L3b	46A; 73G; 185A; 189G; 263G; 310.1C; ...
2	m54006_180724_025808/10027193/ccs	L3b	1438G; 1688G; 1897G; 1992.1G; 2454.1T; ...

Figure 3: Haplotype Report shows the detected haplotypes as well as the haplotype matched to each individual read. The example above shows a mixture sample with haplotypes present at 99.46% and 0.54% frequencies.

Discussion

Using NextGENeLR with Pacific Biosciences long read data provides a powerful solution for mitochondrial DNA haplotyping, particularly for the more challenging case of identifying haplotypes from mixtures. This can be a valuable tool for forensic identification.

NextGENeLR can also be used for detection of SNV, Indel and structure variants from the mitochondrial genome. The very long read lengths, covering up to the full genome, eliminates false positive SNVs that can be detected using Illumina short reads due to similar sequence of other genomic regions.

References

1. Lopopolo et al. (2016). A study of the peopling of Greenland using next generation sequencing of complete mitochondrial genomes. American Journal of Physical Anthropology. 2016;161(4):698–704. doi: 10.1002/ajpa.23074.

2. Shen L, Wei J, Chen T, He J, Qu J, He X et al. Evaluating mitochondrial DNA in patients with breast cancer and benign breast disease. *J Cancer Res Clin Oncol* 2011; **137**: 669–675.

3. Amorim A, Fernandes T, and Taveira N. Mitochondrial DNA in human identification: a review. PeerJ. 2019; **7**: e7314.

4. Templeton et al. (2013). DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. Investigative Genetics. 2013;4(1):1–13. doi: 10.1186/2041-2223-4-26.

5. Churchill et al. (2018). Massively parallel sequencing-enabled mixture analysis of mitochondrial DNA samples. International Journal of Legal Medicine. 2018;132(5):1263–1272. doi: 10.1007/s00414-018-1799-3.

6. Vossen R.H.A.M., Buermans H.P.J. (2017) Full-Length Mitochondrial-DNA Sequencing on the PacBio RSII. In: White S., CantsilierisS. (eds) Genotyping. Methods in Molecular Biology, vol1492. Humana Press, New York, NY [doi: 10.1007/978-1-4939-6442-0_12].

7. Weerts, M.J.A., Timmermans, E.C., Vossen, R.H.A.M. *et al.* (2018). Sensitive detection of mitochondrial DNA variants for analysis of mitochondrial DNA-enriched extracts from frozen tumor tissue. *Sci Rep* **8,** 2261.