appliedbiosystems

MVC: an integrated mitochondrial variant caller for forensics

Roth C[1], Parson W[2], Strobl C[2], Lagacé R[4], Short M[1]
Human Identification, Thermo Fisher Scientific, South San Francisco, CA, USA

ABSTRACT

Mitochondrial DNA (mtDNA) sequencing is a valuable forensics tool in determining the source of DNA obtained from degraded, damaged or small biological samples. However, despite the recent advances in Massively Parallel Sequencing (MPS), the analysis of mtDNA has remained challenging: alignments can be misleading and may result in incorrect variant calls, and contamination from nuclear DNA from small samples may obscure true variants. In this poster, we discuss an integrated approach to solving these issues in a new mito variant caller that integrates various sources of knowledge about mtDNA, including phylotree, EMPOP and NUMTs statistics, that avoids many of the pitfalls of standard algorithms. The knowledge is integrated into the alignment algorithm itself to distinguish true variant calls from NUMT contamination and sequencing artefacts.

INTRODUCTION

The traditional Sanger-type sequencing (STS) method for establishing mtDNA haplotypes tends to be laborious and expensive. The emergence of Massively Parallel Sequencing (MPS) technologies allows for a quicker and more cost effective analysis. However, traditional algorithms and tools that have been used in the past for MPS sequence analysis have not been optimized for mtDNA and show several deficiencies for mtDNA variant calling, in particular in the forensics context:

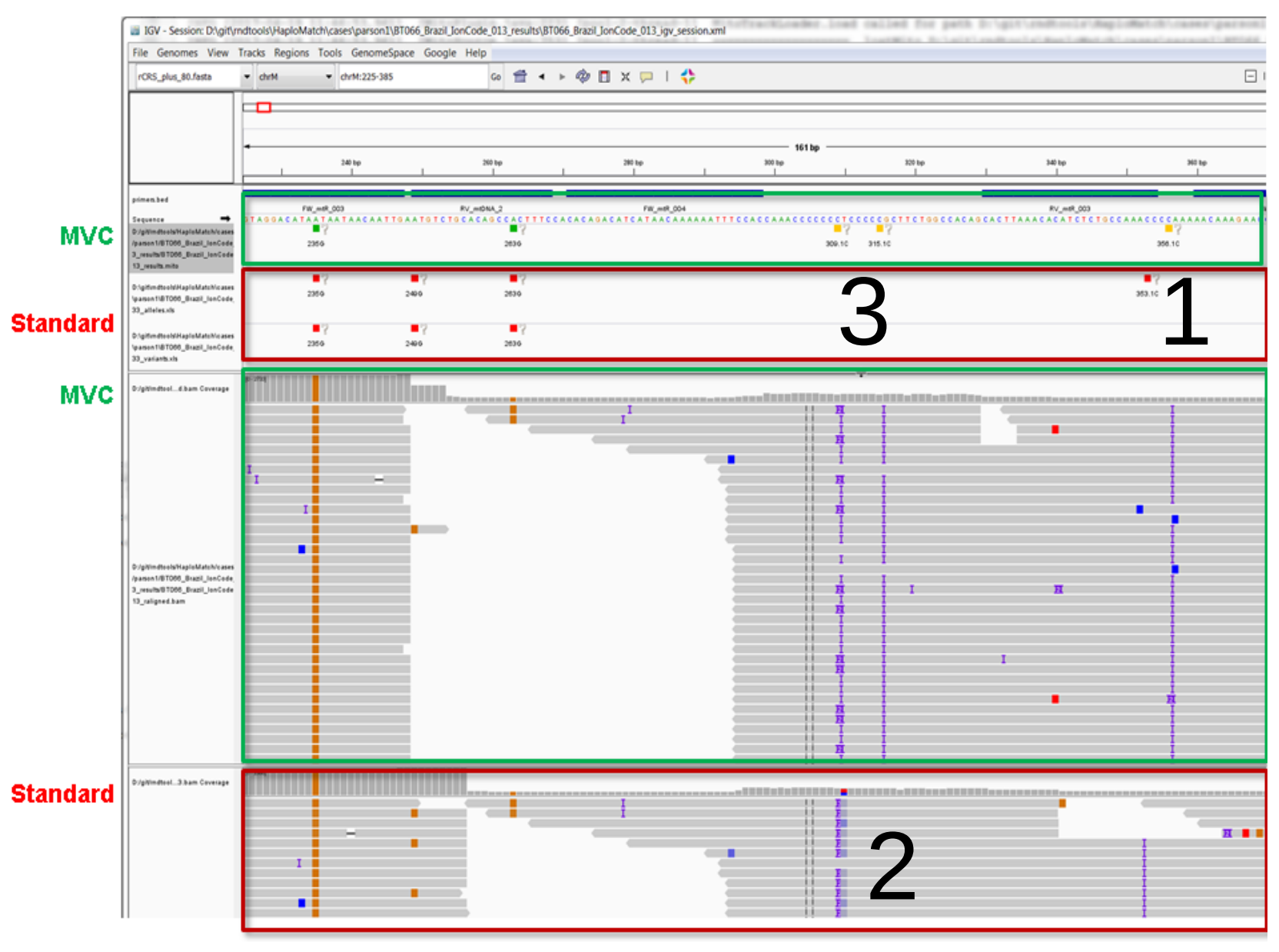


Figure 1. Examples of challenges with current methods

- The alignments are often incorrect and do not reflect the evolutionary phylogeny of mtDNA. Example: Instead of an insertion of two bases at 309, there should be an insertion at 309 and one insertion at 315 (Figure 1, 2/3)
- Variant calls do not adhere to the forensics nomenclature and standards. For instance 315.1C should be used instead of 315 INS
- Insertions and deletions should be right shifted as much as possible (Figure 1, 1). For instance instead of 311.1C, the insertion should be 315.1C
- Existing tools are not able to detect NUMT and other types of contamination. Example: Figure 2 shows an example of a region with NUMT contamination (IGV)

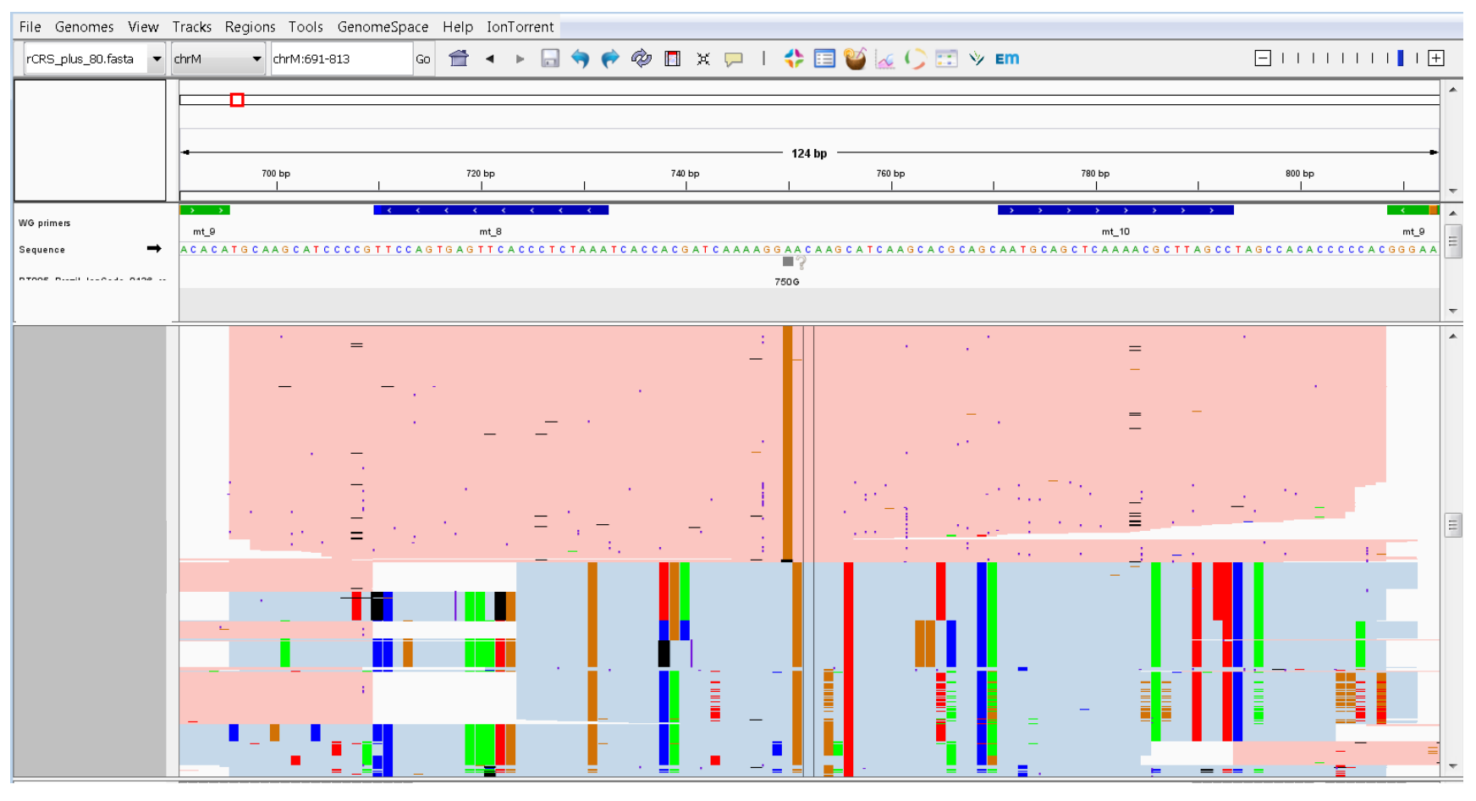
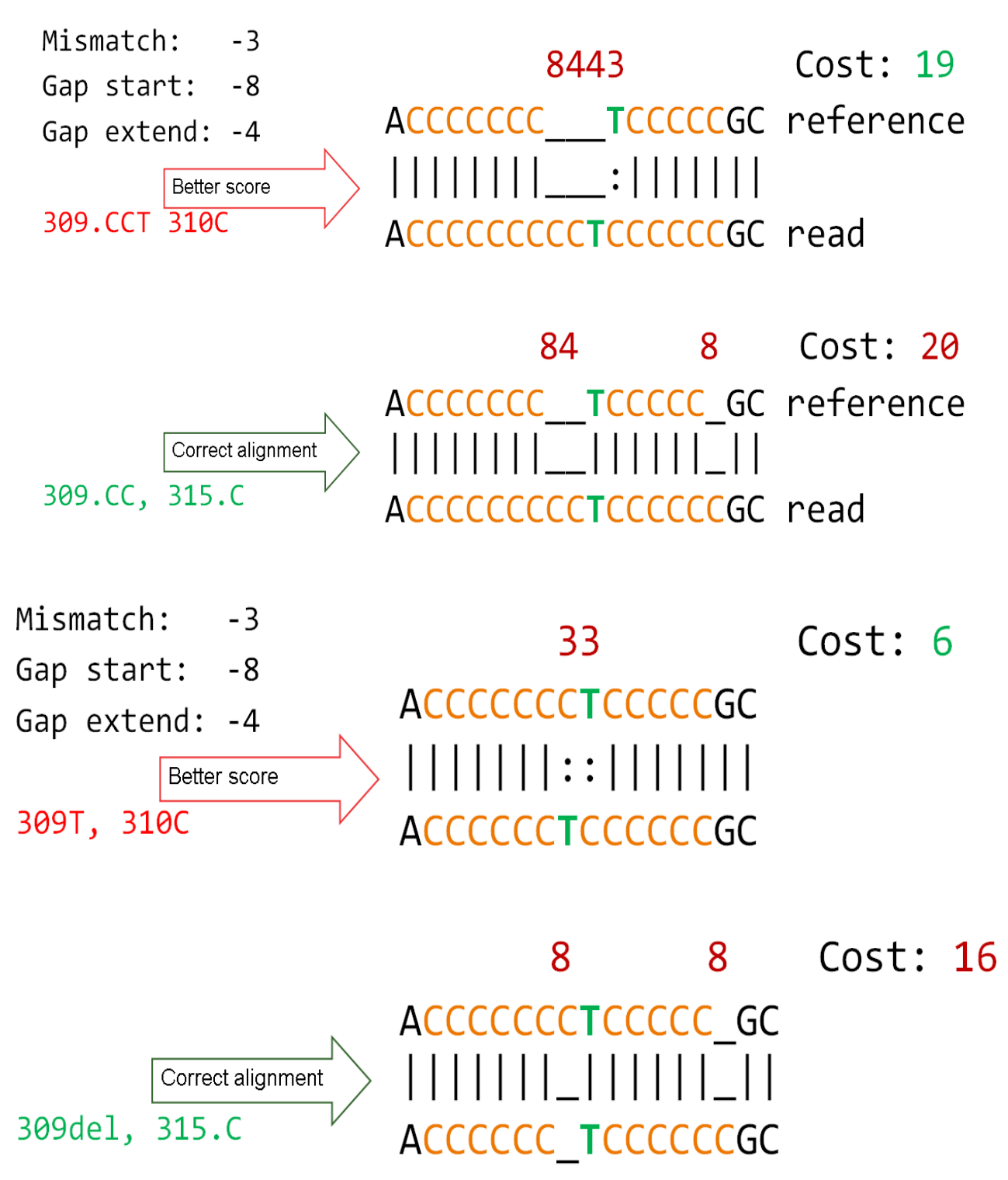


Figure 2. Examples of NUMT contamination: the reddish reads on top are from the sample, whereas the reads at the bottom are nuclear contamination

MATERIALS AND METHODS

To understand the problem, we need to take a closer look at how traditional alignment algorithms work. For MPS in particular, in many cases the focus is on speed due to the amount of data involved and so heuristics are used that do not necessarily compute the optimum solution. For mtDNA however, this is a minor concern as the mtDNA genome is much smaller (approx. 16kB), which allows us to focus on quality and not on throughput, and so it is feasible to compute the optimum alignment. The Smith-Waterman algorithm (Figure 4) is commonly used for this purpose. Yet even in this case, the solution is often not correct for mtDNA. Normally, a fixed cost is defined for a match, mismatch, insertion and deletion (indel). In most cases, an affine gap cost is used, which means that a new indel normally costs more than the extension of an indel (3).



```
Mismatch:   -3              8443        Cost: 19
Gap start:  -8         ACCCCCCC___TCCCCCGC reference
Gap extend: -4         ||||||||___:|||||||
            Better score
309.CCT 310C           ACCCCCCCCCTCCCCCCGC read

                            84      8    Cost: 20
                       ACCCCCCC__TCCCCC_GC reference
            Correct alignment
                       ||||||||__||||||_||
309.CC, 315.C          ACCCCCCCCCTCCCCCCGC read

Mismatch:   -3              33          Cost: 6
Gap start:  -8         ACCCCCCCTCCCCCGC
Gap extend: -4         |||||||::|||||||
            Better score
309T, 310C             ACCCCCCTCCCCCCGC

                            8       8    Cost: 16
                       ACCCCCCCTCCCCC_GC
            Correct alignment
                       |||||||_||||||_||
309del, 315.C          ACCCCCC_TCCCCCCGC
```

Figure 3. Examples of alignments using traditional scoring schemes that lead to wrong results

Figure 3 shows an example of an alignment around position 310 (T) of the mtDNA, using a mismatch cost of -3, an indel start cost of -8 and a gap extension cost of -4. If we compute the optimal alignment using dynamic programming, the alignment on top has the lowest cost. This solution however is wrong, even though it is clearly the optimal solution, and instead the alignment at the bottom would be the correct solution.

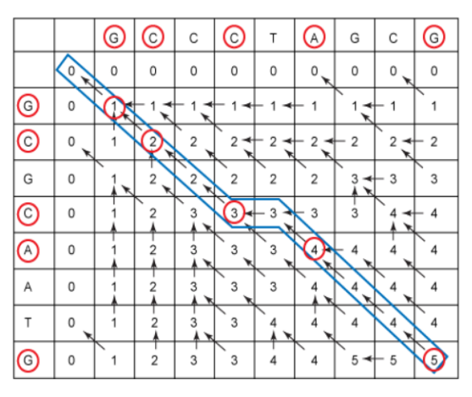


Figure 4. An illustrative example of how an optimal alignment can be computed using a standard dynamic programming approach

Instead of employing a fixed cost for matches, mismatches and indels, our algorithm adjusts for context relevant biological information, including EMPOP and phylotree data, based on the matching haplo group. This way, we are able to compute the correct solution at the bottom of figure 3.

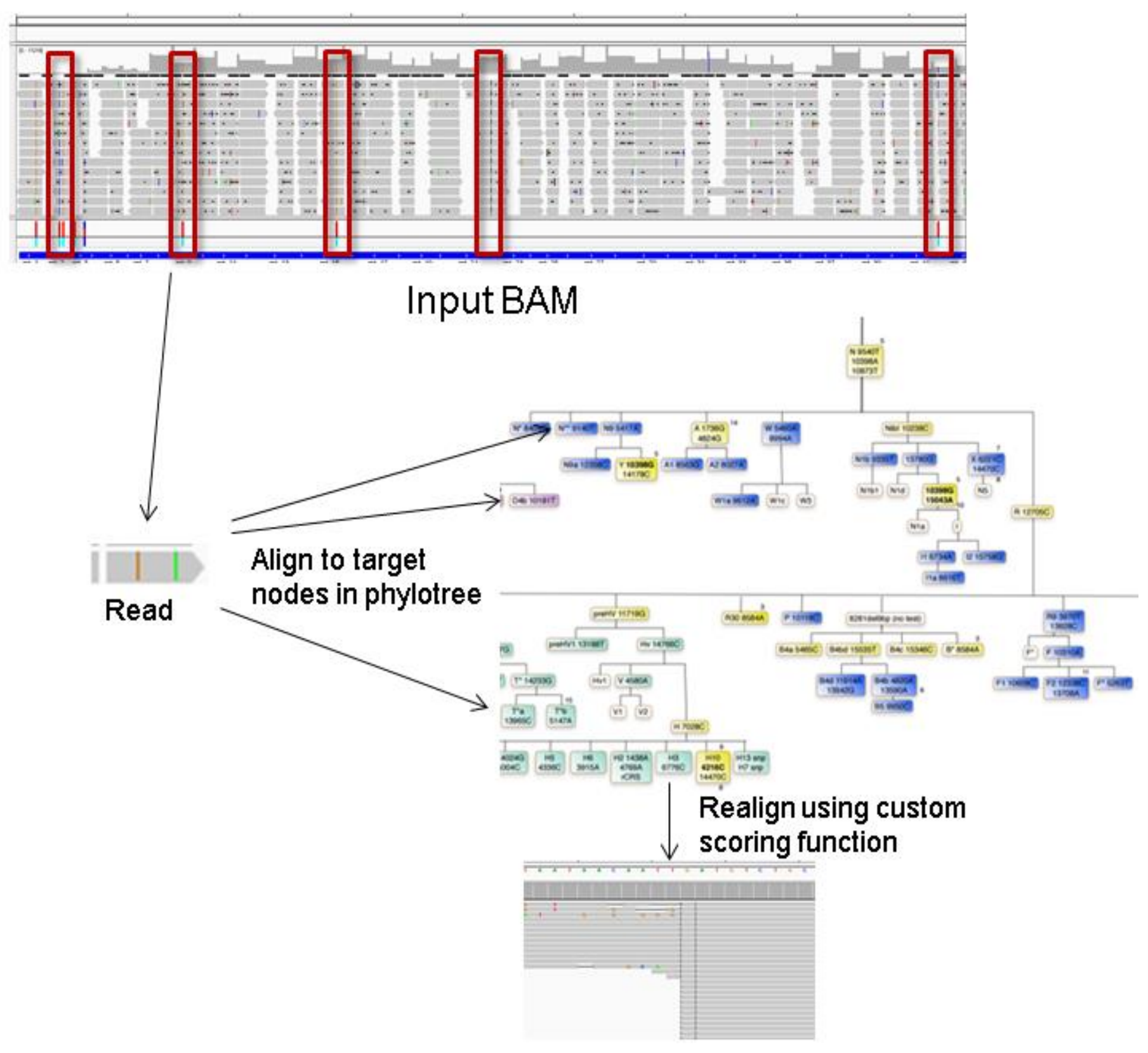

Input BAM
Read
Align to target nodes in phylotree
Realign using custom scoring function

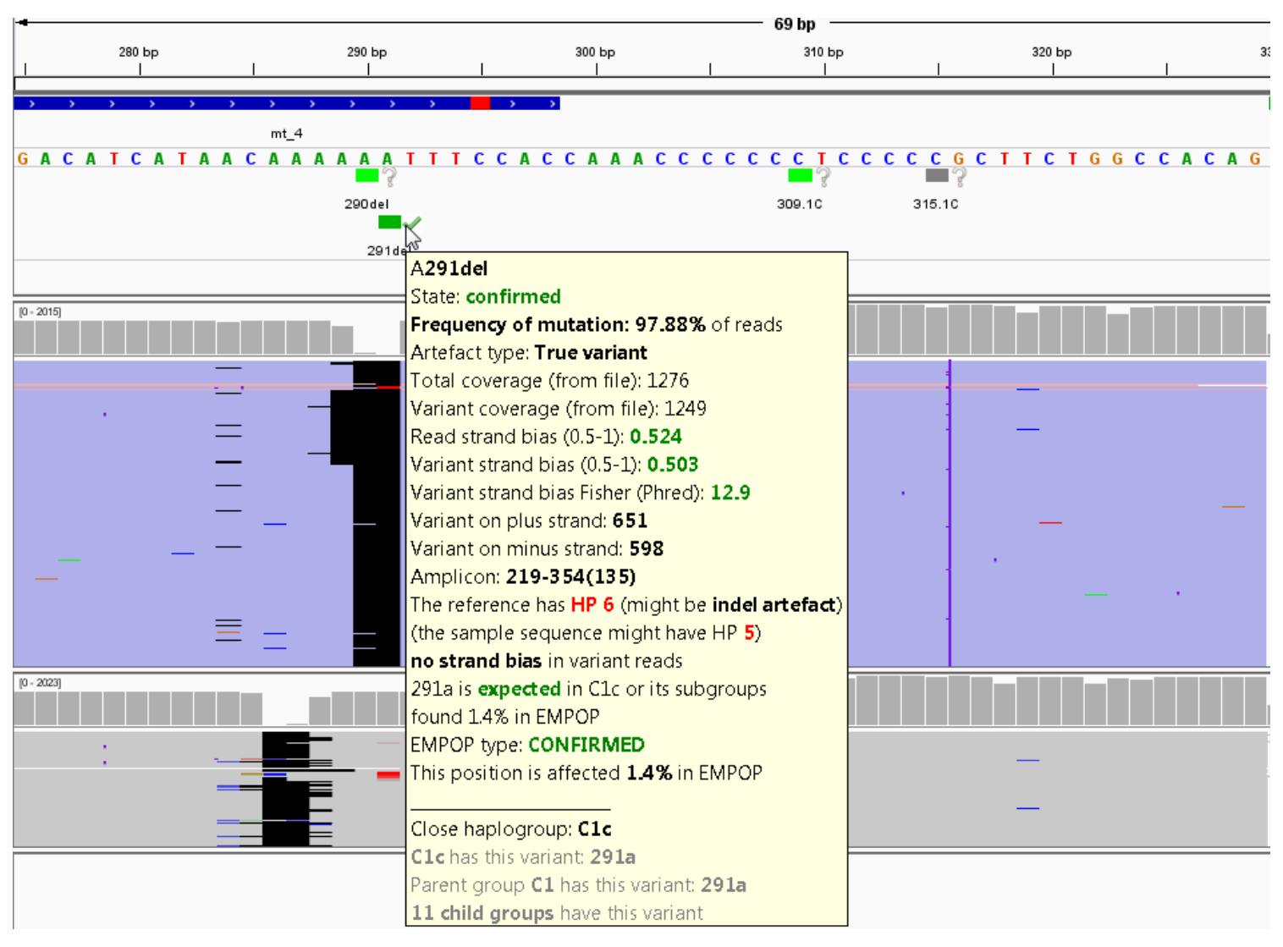
Figure 5. The algorithm takes an input bam file and recomputes the alignments based on biologically relevant information

The approach utilizes artificial reference sequences based on the available biological information to determine the biological context of the reads, which is then applied to determine the correct weighting based on the phylogenetic context. The reads may then be aligned to the reference genome.

Contamination detection may also be applied to avoid calling variants that are due to NUMTs or other types of contamination. One element of this detection is the database of known NUMT positions and statistical calculations to infer NUMTs even if a position is not a known NUMT artefact.

To call the variants and determine the quality of the variants, multiple pieces of evidence are analyzed, such as strand bias, the phylogenetic context of the variant, coverage statistics, sequencing quality, nearby correlated variants and other biologically relevant information.

RESULTS



A291del
State: confirmed
Frequency of mutation: 97.88% of reads
Artefact type: True variant
Total coverage (from file): 1276
Variant coverage (from file): 1249
Read strand bias (0.5-1): 0.524
Variant strand bias (0.5-1): 0.503
Variant strand bias Fisher (Phred): 12.9
Variant on plus strand: 651
Variant on minus strand: 598
Amplicon: 219-354(135)
The reference has HP 6 (might be indel artefact)
(the sample sequence might have HP 5)
no strand bias in variant reads
291a is expected in C1c or its subgroups
found 1.4% in EMPOP
EMPOP type: CONFIRMED
This position is affected 1.4% in EMPOP

Close haplogroup: C1c
C1c has this variant: 291a
Parent group C1 has this variant: 291a
11 child groups have this variant

Figure 6. Screen shot of mito IGV that shows information relevant to a particular variant call. The variants 309.1C and 315.1C have been correctly identified, and the deletion at 290 has been shifted to the right side as expected.

The algorithm has been implemented in Java^{TM} 1.8 and can run on any operating system that supports the Java^{TM} VM. IGV has been extended to support the visualization and editing of mito variant calls. Figure 8 shows examples of variants in the mito track. It can be seen that the issues that were observed with other tools are no longer present, for instance the insertions are correctly placed at position 309.1C and 315.1C. Also, the deletion hat 290 is shifted to the right most position as expected. Figure 7 shows how only the correct variant at 750G was called, whereas the other positions marked in purple have been identified as NUMT contamination.

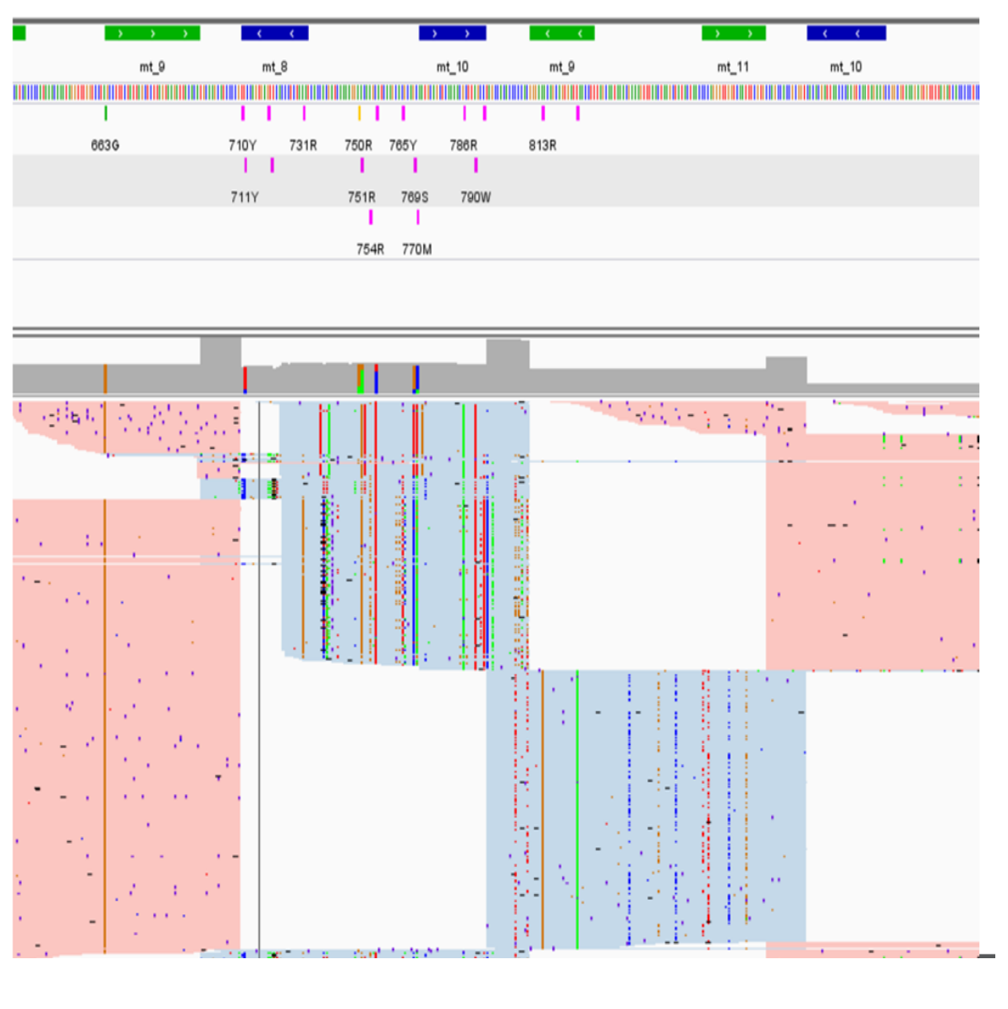


Figure 7. Example of how contamination was identified (colored in light blue), and how the variant 750G was still identified.

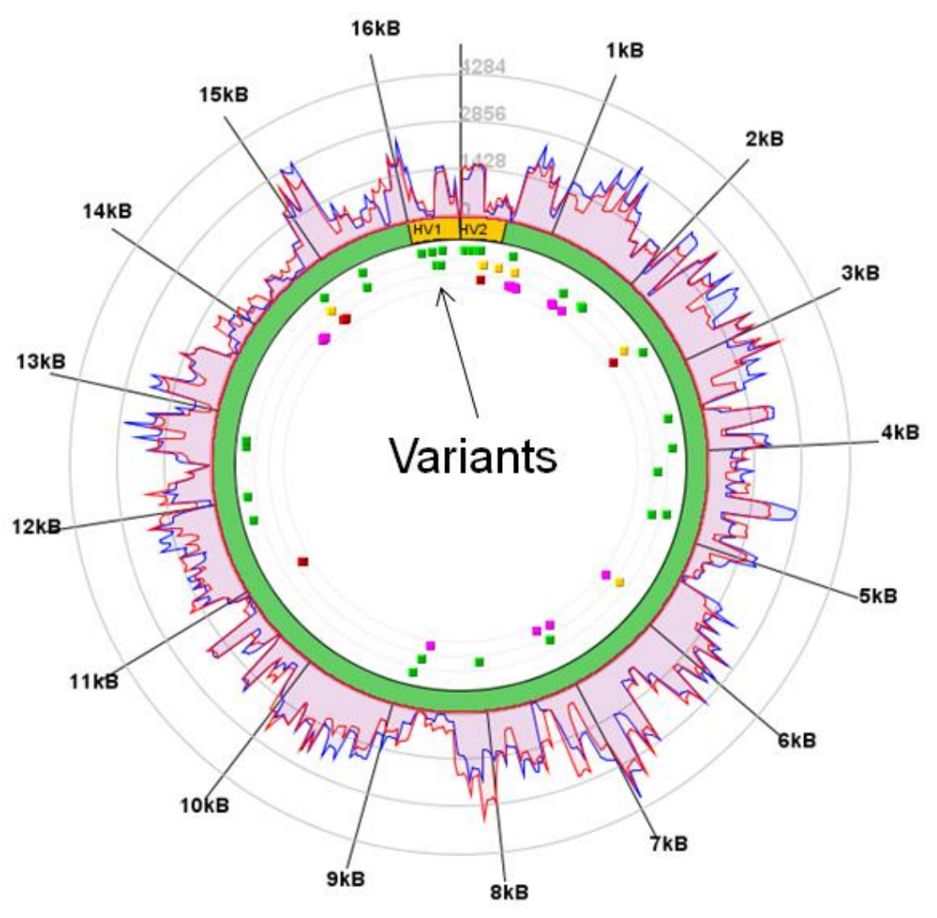
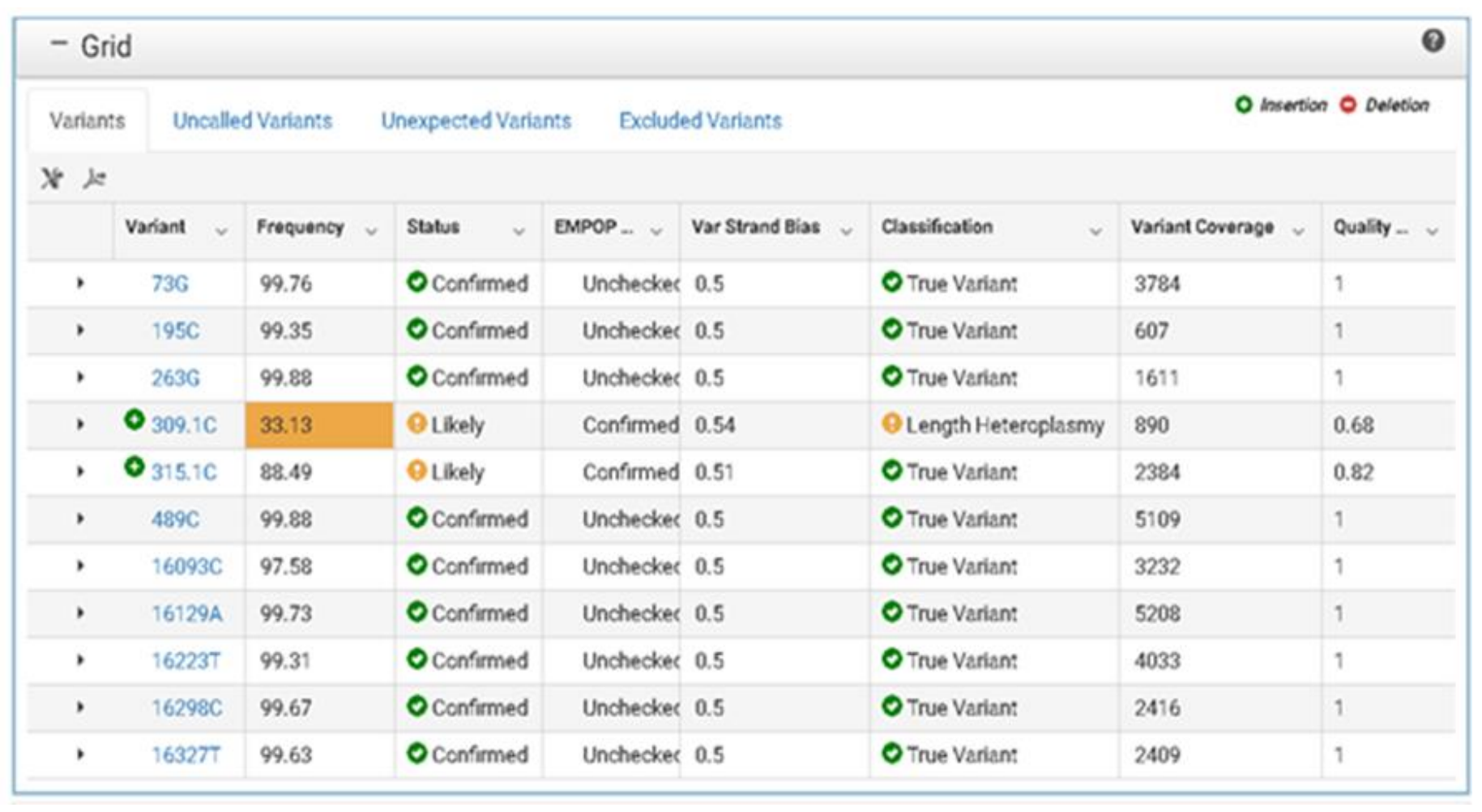


Figure 8. Interactive circular plot implemented both in IGV and Converge^{TM}. Clicking on a variant shows the details in IGV.

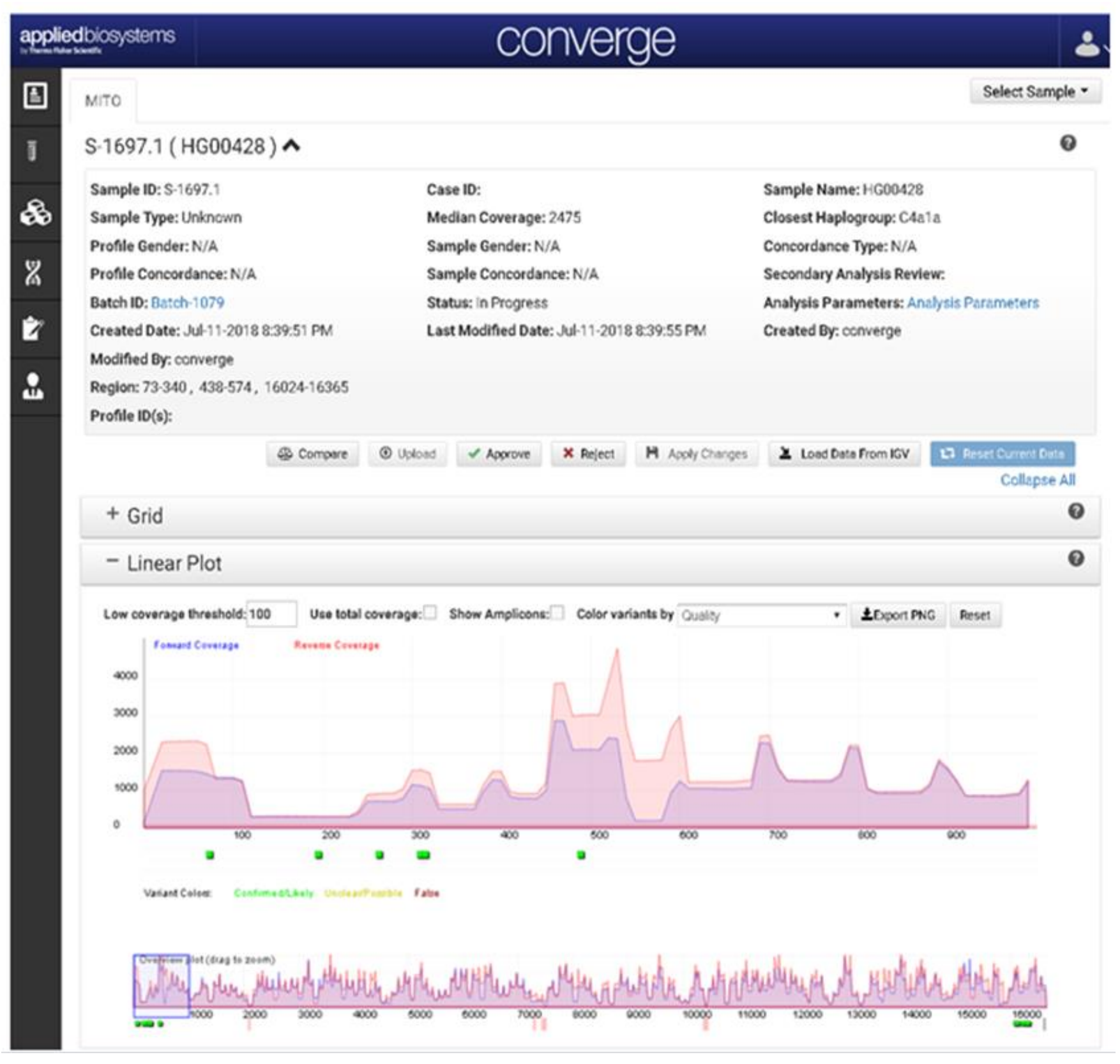


	Variant	Frequency	Status	EMPOP ...	Var Strand Bias	Classification	Variant Coverage	Quality ...
▸	73G	99.76	Confirmed	Unchecke	0.5	True Variant	3784	1
▸	195C	99.35	Confirmed	Unchecke	0.5	True Variant	607	1
▸	263G	99.88	Confirmed	Unchecke	0.5	True Variant	1611	1
▸	309.1C	33.13	Likely	Confirmed	0.54	Length Heteroplasmy	890	0.68
▸	315.1C	88.49	Likely	Confirmed	0.51	True Variant	2384	0.82
▸	489C	99.88	Confirmed	Unchecke	0.5	True Variant	5109	1
▸	16093C	97.58	Confirmed	Unchecke	0.5	True Variant	3232	1
▸	16129A	99.73	Confirmed	Unchecke	0.5	True Variant	5208	1
▸	16223T	99.31	Confirmed	Unchecke	0.5	True Variant	4033	1
▸	16298C	99.67	Confirmed	Unchecke	0.5	True Variant	2416	1
▸	16327T	99.63	Confirmed	Unchecke	0.5	True Variant	2409	1

Figure 9. Example of the variant grid in Converge^{TM} that shows relevant information for each called variant

The mito variant caller has been integrated into Converge^{TM} enterprise application and allows a user to analyze mtDNA samples with the Precision ID mtDNA Whole Genome and Control Region Panels. The algorithm can be launched and scheduled directly on the torrent server through a plugin.

The integration into Converge^{TM} allows all relevant associated data to be centrally managed, where it can be linked to a Case. Figure 10 shows an example of a table with called variants. Both the table and also the plot in Figure 11 is linked to mito IGV, which allows the inspection of the underlying data such as the recomputed alignments.



S-1697.1 (HG00428)

Sample ID: S-1697.1
Sample Type: Unknown
Profile Gender: N/A
Profile Concordance: N/A
Batch ID: Batch-1079
Created Date: Jul-11-2018 8:39:51 PM
Modified By: converge
Region: 73-340 , 438-574 , 16024-16365
Profile ID(s):

Case ID:
Median Coverage: 2475
Sample Gender: N/A
Sample Concordance: N/A
Status: In Progress
Last Modified Date: Jul-11-2018 8:39:55 PM

Sample Name: HG00428
Closest Haplogroup: C4a1a
Concordance Type: N/A
Secondary Analysis Review:
Analysis Parameters: Analysis Parameters
Created By: converge

Low coverage threshold: 100 Use total coverage: □ Show Amplicons: □ Color variants by Quality

Figure 10. The results page of one sample in Converge^{TM} that shows the interactive coverage plot

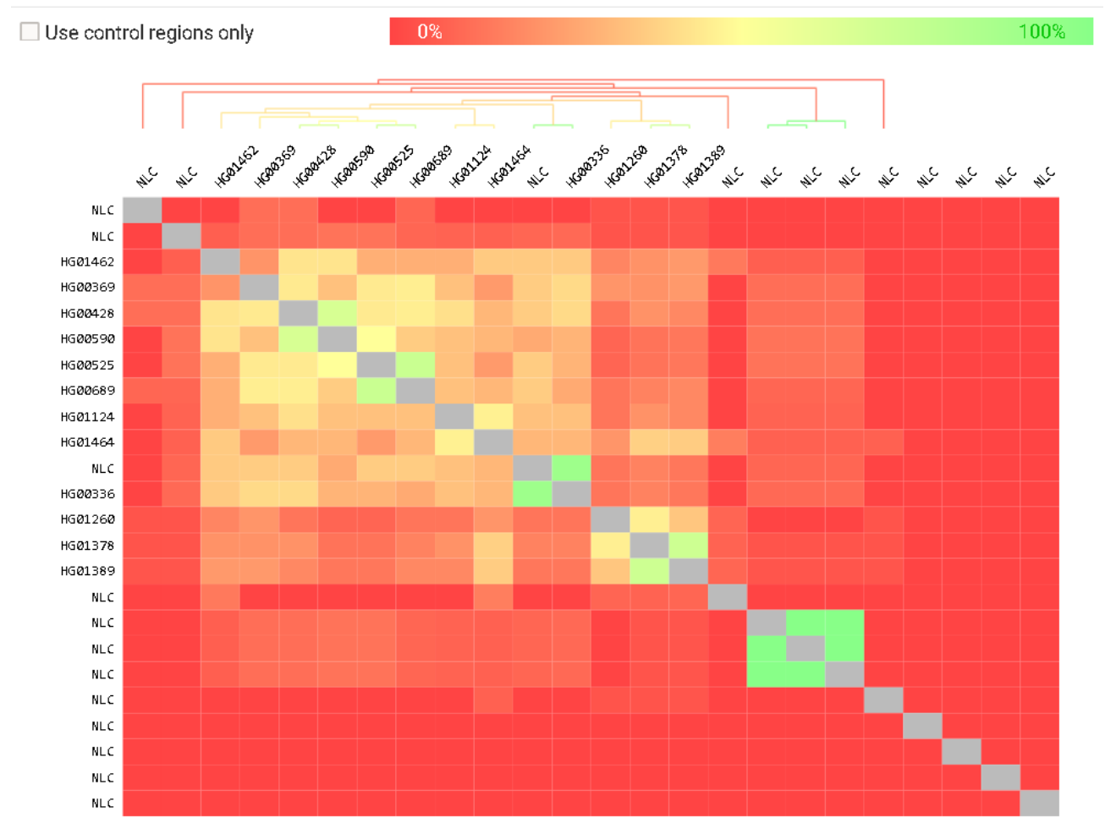


Figure 11. The software allows the comparison of multiple samples with each other, for instance in the form of a heat map

REFERENCES

1. Parson, W and Dür, A (2007) EmPOP—a forensic mtDNA database. *Forensic science international. Genetics. 1. 88-92. 10.1016/j.fsigen.2007.01.018.*
2. Parson, W, Gusmao, L, Hares, DR et al (2014) DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. *Forensic Sci Int Genet* 13:134–142.
3. Andrews R.M., Kubacka I., Chinnery P.F., Lightowlers R.N., Turnbull D.M., Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. 1999;23:147. [PubMed]
4. Carracedo A, Butler JM, Gusmão L, Linacre A, Parson W, Roewer L, Schneider PM (2013) New guidelines for the publication of genetic population data. FSI:Genetics 7(2), 217-220
5. Smith TF, Waterman MS: Identification of common molecular subsequences. J Mol Biol 1981, 147: 195–197. 10.1016/0022-2836(81)90087-5

TRADEMARKS/LICENSING

For Research, Forensic, or Paternity Use Only. Not for use in diagnostic procedures.

ThermoFisher SCIENTIFIC