Bionano Solve Theory of Operation: Bionano EnFocus™ FSHD Analysis

Document Number: 30321

Document Revision: C

Table of Contents

Legal Notice

For Research Use Only. Not for use in diagnostic procedures.

Revision History

Revision	Notes
C	• Added timing for analysis on Bionano Access server • Added text on limitations of mosaicism simulation

Introduction

Facioscapulohumeral Muscular Dystrophy (FSHD) is one of the most common forms of muscular dystrophy. FSHD symptoms include progressive muscular degeneration, weakness, and atrophy, with variation in the phenotype among affected individuals. There are no approved therapies, but physiotherapy may alleviate the symptoms.

FSHD can be inherited and impact multiple members of a family as an autosomal dominant genetic disease. Non-inherited FSHD (presumably due to *de novo* or somatic mutations) also occurs. FSHD involves a retrogene DUX4, which is normally not expressed. Abnormal expression of DUX4 in skeletal muscle causes FSHD. There are two FSHD subtypes, both involving abnormal DUX4 expression, but they differ in the underlying mechanism by which DUX4 expression is activated.

Genetic testing is the most reliable way to confirm a diagnosis. The contraction of the D4Z4 tandem repeat on chromosome region 4q35 on a permissive haplotype 4qA is diagnostic for FSHD Type 1, which accounts for 95% of cases. It is currently primarily assayed by Southern blot.

Bionano Genomics has developed an FSHD analysis workflow that offers several advantages and is based on optical mapping data collected on the Saphyr Genome Imaging instrument. Based on specific labeling and mapping of ultra-high molecular weight DNA in nanochannel arrays, optical mapping provides a high-resolution analysis of the D4Z4 repeat array.

Briefly, the molecules aligning to regions of interest are extracted and assembled. The resulting consensus maps are used for the Bionano EnFocus™ FSHD Analysis. The D4Z4 repeat regions in chromosomes 4 and10 are sized, and the permissive and non-permissive haplotypes (4qA and 4qB) assigned. Additional structural variants and copy number gains and losses are noted in the proximity of the D4Z4 repeat array on chromosome 4 and of the SMCHD1 gene on chromosome 18. The analysis data can be imported into Bionano Access, a graphical user interface tool for visualization and curation. Access can generate a summary of the results in pdf and in json format.

Note that the method described cannot detect single-nucleotide variants that do not impact sequence motif sites and may miss small variants with potential functional impacts. Also, the pipeline only supports Bionano's DLE-1 labeling enzyme.

Analysis workflow

Local assembly of regions of interest

The FSHD analysis pipeline first performs a local assembly of regions of interest by selecting molecules that align to those regions and assembling only those molecules and subsequently analyzing the resulting genome maps in the chr4 and chr10 D4Z4 regions in order to size the repeats and assign haplotypes to the alleles. Specifically, we extract molecules from chr4: 187.2-190.2 Mbp and from chr10: 126.0-133.8 Mbp. Additional selected regions of the genome are also assembled and analyzed as part of the quality-control process (discussed in a subsequent section; Table 1).

The local assembly workflow is similar to the standard Bionano *de novo* whole-genome assembly workflow, but the key difference here is that the reference is used as a guide and that only targeted regions are assembled. The local assembly is performed using parameters optimized for effective assembly of the D4Z4 repeat regions. This workflow significantly reduces the assembly time and is effective in assembling the complex targeted regions.

Identification of maps of interest and chromosome assignment

The pipeline identifies maps aligning to the chr4 or chr10 D4Z4 region for FSHD analysis. There is partial homology between chr4 and chr10, and both contain a D4Z4 array. A map may align to both chr4 and chr10. The pipeline tries to assign the maps to the correct chromosomes using the following criteria.

If a map aligns to both chromosomes, the pipeline keeps the alignment with the highest confidence score (negative logarithm with base 10 of the p-value of the alignment). If a map aligns equally well to chr4 and chr10 (based on confidence), the pipeline keeps the alignment with the most label matches beyond the D4Z4 repeat region. If the number of matches is the same, the pipeline assigns the map to the chromosome which has only one alignment. The map would be filtered out if there is still ambiguity (for example if both chromosomes have the same number of alignments).

Haplotype assignment

The pipeline then assigns haplotype A or B to each map spanning the D4Z4 repeat array. Similar to the Smith-Waterman algorithm used for local sequence alignment, the pipeline uses a dynamic-programming algorithm to assess similarity between the reference haplotype-specific intervals and the intervals in the assembled maps. The pipeline performs a local alignment for each possible haplotype separately. If there are at least two matching intervals, the pipeline looks at the scores of the two alignments and assigns the haplotype based on the higher-scoring alignment. If there are less than two matching intervals, the pipeline assigns the haplotype as “unknown”.

D4Z4 repeat count estimation

After identifying maps that are relevant to D4Z4 analysis, the pipeline uses those maps for sizing the repeat arrays. While the DLE-1 enzyme does not directly label the individual D4Z4 units, the pipeline estimates the repeat array lengths based on the interval between labels flanking the D4Z4 arrays. There are expected offsets between the flanking labels and the actual repeat starts and ends; these are pre-determined based on analysis of the reference and applied automatically by the pipeline. The offsets for the D4Z4 array on chromosomes 4 and 10 are determined and used in a similar fashion.

The same offset is used for the repeat start for both A and B haplotypes. However, haplotype-specific offsets for the repeat end are required due to a difference in the location of the flanking label on the two haplotypes.

The pipeline uses the following formula to compute the number of D4Z4 repeat units (N):

$N = (\Delta P - D_I - D_{A|B})/S$

where N is the number repeat units, ΔP is the size of the interval between the flanking labels, D_1 is the offset for repeat start, D_A and D_B are the haplotype-specific offsets for the repeat end, and S is the expected size of a D4Z4 repeat unit (3.3 kbp). Whether D_A or D_B is used depends on the haplotype assignment.

Note that only maps with more than 15 unique labels before the start of D4Z4 repeat array in chr4 or chr10 are included in the final output. These maps may contain all the expected labels for either haplotype (presumably fully assembled to the end of the chromosome) or may lack some or all of the haplotype-specific labels, if the maps are truncated before the repeat array ends (presumably because molecules did not fully span the repeat array and the haplotype-specific labels). If the repeat array is truncated, the haplotype is assigned to be “unknown”, but the pipeline would make a lower-bound estimate of the repeat count (reported as, for example, “>= N”) by using the 4qA offsets.

Assessment of molecule support

The pipeline analyzes the molecule-map alignment to assess the amount of molecule support for a given map. The number of molecules spanning across repeat start and end provide supporting evidence for the repeat count estimation. This information is output in the final report.

Quality control

Inferred sex of sample

The whole genome copy number pipeline is run as part of the analysis pipeline, and it produces information about specific regions of interest (4q35 and SMCHD1, for example). The pipeline also outputs the sample’s inferred sex information in the final report. It checks for whether there is non-trivial coverage of chrY. If there is, the sex is inferred to be male, and female if otherwise. The pipeline does not handle more complex sex chromosome configurations. If external data is available, one could compare the inferred sex with the external data and check

for consistency.

Assessment of molecule quality

The pipeline collects data on molecule alignment quality to the reference. To ensure that the molecule quality is sufficient for downstream analyses, it requires that the map rate be at least 70%, the effective coverage be at least 75X, and the molecule N50 be at least 200 kbp.

Selection and assessment of stable regions

To assess consensus map level quality, the pipeline analyzes regions of the genome that are deemed stable (Table 1) based on the hg38 reference.

Selection of stable regions

One region per autosome (for a total of 22 regions) was selected based on analysis of 58 *de novo* assemblies of Bionano human control samples. The consensus map-to-reference alignment for the controls was analyzed. For each reference interval and for each sample, the absolute percent difference between the interval length of the reference and that of a given map was computed. The mean absolute percent difference across controls for each interval was then computed and sorted. After excluding regions with insufficient data or too many alignments, the regions with the lowest mean absolute percent differences, assumed to be the most stable, were selected.

Assessment of stable regions

When analyzing a sample of interest, the FSHD pipeline assembles molecules from the stable regions in Table 1, and the resulting consensus maps and consensus map-to-reference alignment are analyzed in a similar fashion. The pipeline expects the consensus maps to be consistent with the reference for the selected regions. Based on expected sizing errors, the absolute percent differences between the map and the reference should not exceed 1.2%. The pipeline requires that at least 90% of the regions be under this threshold.

Table 1. List of stable regions based on hg38 coordinates included in the quality control assessment.

Chr	Coordinates		Chr	Coordinates
1	222,324,492 - 222,349,194		**12**	25,901,387 - 25,914,482
2	203,598,419 - 203,624,005		**13**	26,774,581 - 26,796,446
3	31,786,552 - 31,805,963		**14**	49,469,153 - 49,487,044
4	159,395,859 - 159,416,605		**15**	60,014,272 - 60,041,969
5	37,318,756 - 37,335,731		**16**	77,498,326 - 77,517,842
6	53,262,654 - 53,282,806		**17**	1,377,309 - 1,389,336
7	26,961,779 - 26,970,051		**18**	12,367,665 - 12,396,092
8	121,580,578 - 121,595,557		**19**	13,335,300 - 13,361,530
9	116,981,783 - 117,012,768		**20**	47,394,428 - 47,417,351
10	62,135,760 - 62,157,297		**21**	37,268,614 - 37,282,468
11	78,075,503 - 78,100,220		**22**	38,154,243 - 38,163,301

Data summary

The pipeline compiles the intermediate data for each map into final results for reporting. It checks whether the repeat region contains potential variants, flags truncated maps, and removes truncated maps if they are the partial results of fully assembled ones. Finally, it gathers all the data necessary for Bionano Access to visualize the maps, to highlight the repeat regions and haplotype labels, and to generate the final FSHD analysis report. The key data files are compressed into a zip file, and results are summarized in a JSON file, both of which are automatically transferred into Bionano Access. Access can then generate a PDF report based on the results. See Bionano Access Software User Guide (PN 30142) for more detail.

The JSON file is also available for download and direct import into the customer's reporting tool for parsing and presentation. For more information about the JSON file, see Bionano EnFocus™ FSHD JSON File Format Specification Sheet (PN 30322) available at the Bionano Genomics support website.

If enabled in Bionano Access in the report generation configuration, the pipeline would note additional structural variants from the SV detection module and large copy number gains and losses from the copy number analysis module (typically above 500 kbp) in the proximity of the D4Z4 repeat array on chr4 (within 1 Mbp of the start of the array). Copy number gains and losses in the proximity of the SMCHD1 gene on chr18 (2.66-2.81 Mbp) would also be noted.

Performance summary

In total, we analyzed 30 samples expected to be FSHD-positive; 28 of them had repeat contractions on the 4qA haplotype. The two samples that did not show repeat contractions were run twice; the results were consistent. It is likely that they in fact did not have repeat contractions, and that the reported FSHD-like phenotypes may be mediated by a different mechanism.

In Table 2, we highlight results from 12 FSHD-positive cell lines that we obtained from the Coriell repository. All were expected to contain repeat contractions on the 4qA haplotype. Six of the cell lines were run in triplicate for reproducibility analysis. We detected the expected repeat contractions in all 12 samples. In all cases, repeat counts from the pipeline were within two units of the expected counts. The haplotype assignments were consistent with the annotation and among the triplicates. However, the haplotype was not assigned for the long chr4 allele in GM16354.

Table 2. Results from FSHD analyses of Coriell cell lines. Annotation data were incomplete for some samples. Differences within one unit are considered consistent and not highlighted.

Sample	Annotation	Run 1	Run 2	Run 3
GM16250	4A-6U/4B	Consistent	Consistent	Consistent
GM16337	4A-5U/4A	Consistent	Consistent	Consistent
GM16348	4A-4U/4B	Consistent	Consistent	Consistent
GM16354	4A-9U/4A	32U (haplotype not called)*	Consistent	Consistent
GM17868	5U/31U	Consistent	Consistent	Consistent
GM18027	3U/27U	Consistent	Consistent	Consistent
GM16283	4A-6U/4A	Consistent	Did not re-run	
GM16334	4A-5U/4A	Consistent		
GM16420	4A-6U/4A	Consistent		
GM17724	6U/18U	8U/18U		
GM17898	4U/9U	6U/10U		
GM17939	3U/33U	Consistent		

*Haplotypes may not be called for longer alleles; see the FAQ below for more information.

Control samples with no reported FSHD-like phenotypes from the San Diego Blood Bank and 1000 Genomes Project were also analyzed. Figure 1 shows the repeat count distribution of those controls. 2 out of 58 had 10-unit repeats on 4A. 10 units is considered borderline; in one study[1], ~7% of control samples had 8-11 units.

[1] Butz et el. J Neurol (2003) surveyed 39 unrelated FSHD patients and 102 healthy controls using Southern blot.

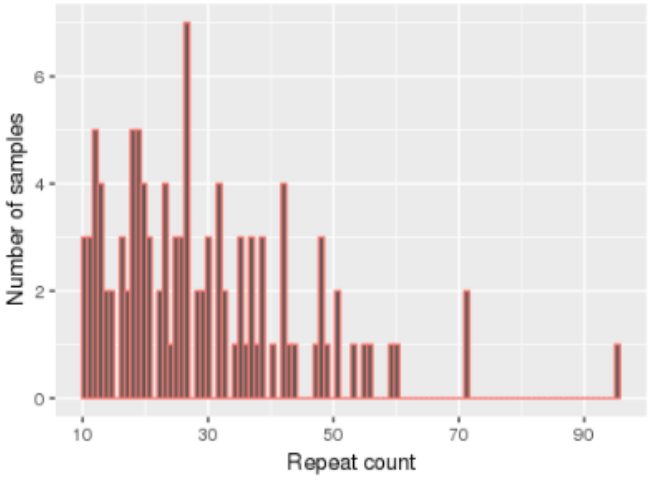

Figure 1. D4Z4 repeat count distribution in 58 control samples.

The runtime performance for analysis and quality control of an EnFocus™ FSHD dataset set up using the EnFocus™ FSHD run template in Bionano Access is typically 90 minutes on a Saphyr Compute Server and 3.5 hours on the Bionano Access Server.

EnFocus™ FSHD Analysis Report

Bionano Access generates a PDF summary report that contains the key results from the FSHD analysis pipeline. The first page of the report is a summary page that contains basic information about the sample being analyzed and the main findings on the D4Z4 repeat region in chr4 and chr10. An example of this summary page is shown in Figure 2.

The summary page is followed by detailed results pages where each page shows an assembled Bionano map that contains the D4Z4 repeat in either chr4 or chr10. The maps shown correspond to entries in the results table on the first page. The molecules that support the assembled maps are also shown below the maps as supportive evidence. An example map with a disease repeat contraction allele (i.e. a contracted repeat array on a 4qA haplotype) is shown in Figure 3. As a comparison, a map with a normal-sized repeat array on a 4qB haplotype is shown in Figure 4. Maps from the homologous D4Z4 region on chr10 are also shown (Figure 5).

Sometimes, when the size of the repeat array is long (typically more than 30 units), there may not be enough long molecules to span the full D4Z4 array and the haplotype-specific labels. As a result, the consensus map may not have the full D4Z4 repeat array. In such cases, the FSHD analysis pipeline would provide an estimate on the lower bound on the repeat array size. An example is shown in Figure 6.

Typically, two distinct alleles of D4Z4 repeat region on chr4 or chr10 would be assembled. However, mosaic repeat alleles have been reported, and more than two alleles may be assembled. Figure 7 shows an example of mosaicism where two normal alleles and one contracted allele were assembled in chr4.

Bionano EnFocus™ FSHD Analysis Report

Experiment information

Sample name: GM16250
Enzyme used: DLE-1
Instrument serial number: SAPHYR_F12
Chip ID: 3RSBCYWNPMKXRNWU (Flowcell 2)
Run ID: 4ba6a250-c593-41fe-b8bf-fd56ecee9e33
Date of data collection: 2019-07-29 10:20:39 AM
Version of ICS software: ICS 4.8.19085.2

Overall sample quality metrics

Inferred sex of sample: male
Assessment of molecule quality: PASS
Assessment of stable regions: PASS

Analysis information

Analysis performed: Bionano EnFocus™ FSHD Analysis
Job ID: 12345
Job name: FSHD analysis run 1
Operator name: Tom Wang
Date of analysis: 2019-12-10 10:58
Version of assembly pipeline: Bionano Solve 3.5
Version of FSHD analysis pipeline: Bionano EnFocus™ FSHD Analysis 1.0

Detailed results

Chr	Map ID	Calculated repeat count (units)	Haplotype	Repeat-spanning coverage (X)
4	22	5	4qA	27
4	271	17	4qB	24
10	12	6	10qA	23
10	250	15	10qA	25

Figure 2. Example summary page from EnFocus™ FSHD analysis report.

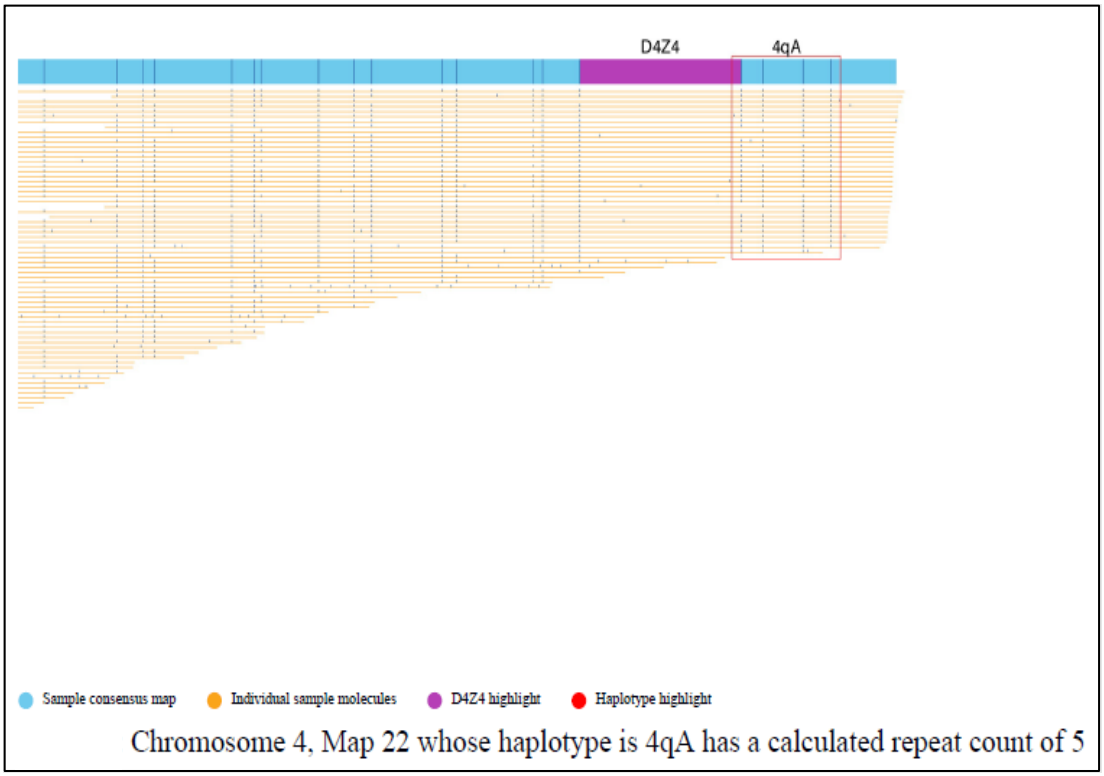

Figure 3. An example of a map with a contracted repeat array on a 4qA haplotype.

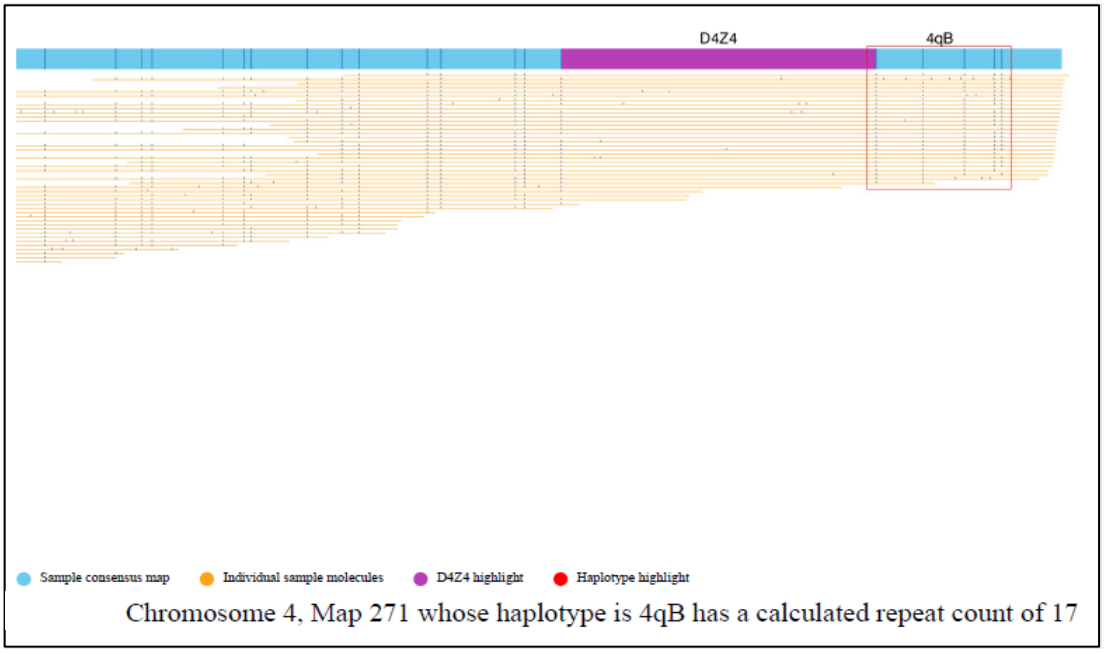

Figure 4. An example of a map with a normal repeat array (more than 10 units) on a 4qB haplotype.

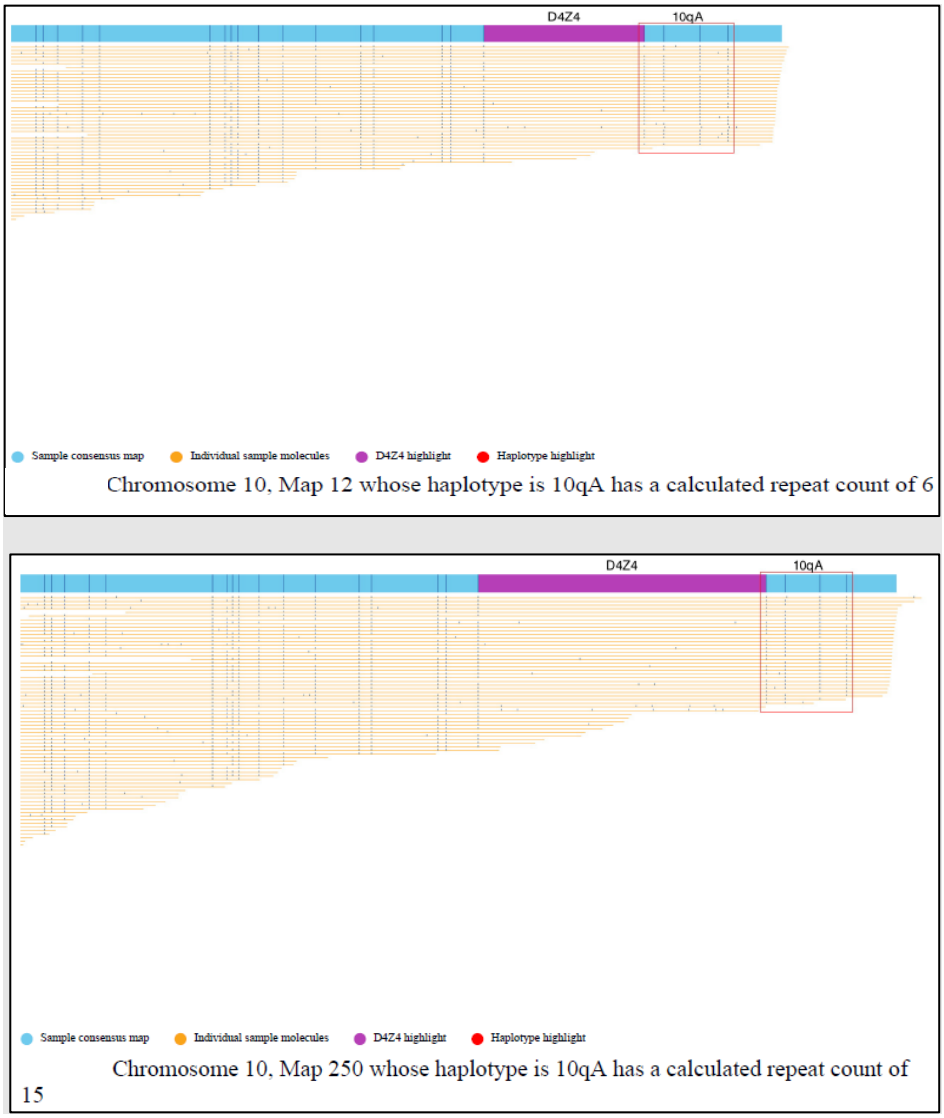

Figure 5. Examples of the assembled D4Z4 repeat regions on chr10

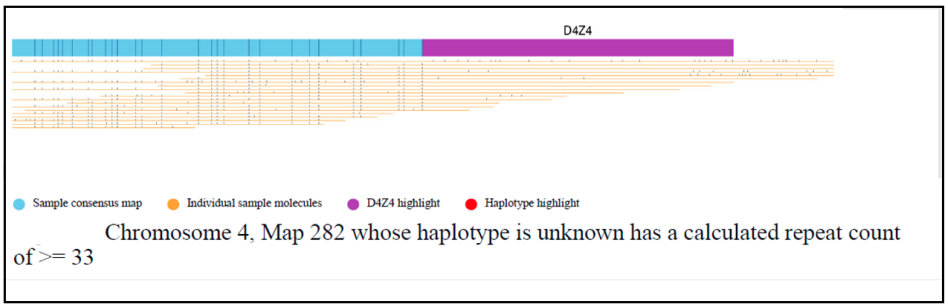

Figure 6. An example of a map with a truncated D4Z4 repeat. As a result, a lower bound on the repeat array size was estimated from the data.

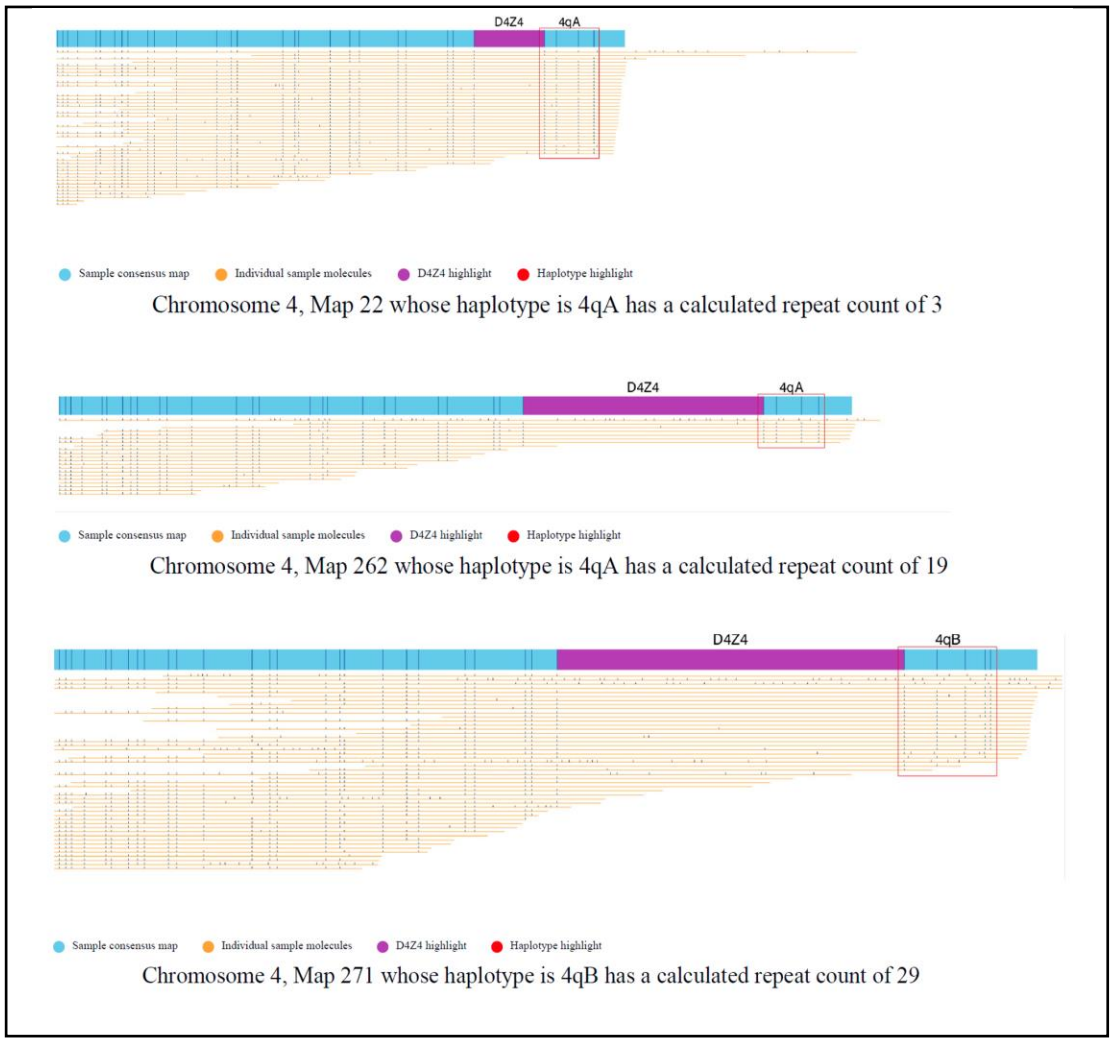

Figure 7. Example of mosaicism in the D4Z4 region on chr4.

Mosaicism detection

To estimate the sensitivity for detecting mosaic alleles, we started with four Coriell samples with known FSHD-related FSHD repeat contractions (Table 3). These samples were mixed *in vitro* with NA12878 such that the contraction alleles would be at 25%, 12.5, and 6.25% allele fractions. 1.3 Tbp of data were collected for each of the mixtures and analyzed with EnFocus FSHD Analysis Pipeline. The contraction alleles were detected consistently at as low as 12.5% allele fraction. The contraction alleles were detected at 6.25% allele fraction for all but one sample.

Although performance in this test is excellent, this is a limited assessment of synthetic “mosaic” mixtures that may not represent all true mosaic cases. Bionano is making no claims regarding performance for detection and characterization of D4Z4 loci.

Table 3. Coriell cell lines used for mosaicism detection analysis.

Sample	# of known repeats (U)	Run in reproducibility experiments as triplicates
GM16250	5	Yes
GM17724	8	No
GM16354	8	Yes
GM16348	3	Yes

Table 4. Summary of detection results. Only alleles expected in the disease samples are considered in the following table. The NA12878 background/spike-in alleles are ignored.

Sample	Allele fraction (%)	# of repeats detected on contraction allele (U)	
		At standard coverage (400 Gbp)	At high coverage (1.3 Tbp)
GM16250	50	5	5
	25	5	5
	12.5	5	5
	6.25	5	5
GM17724	50	8	8
	25	8	8
	12.5	8	8
	6.25	8	8
GM16354	50	8	8
	25	8	8
	12.5	8	8
	6.25	Not detected	Not detected
GM16348	50	3	3
	25	3	3
	12.5	3	3
	6.25	3	3

FAQs

1. How does data quality impact FSHD analysis results?

As discussed in the quality control section, the pipeline looks at three specific criteria at the molecule quality level (map rate, molecule N50 > 150 kbp, and effective coverage) and also checks the consensus map quality. Having sufficiently long molecules and sufficient coverage ensures that the repeats can be fully spanned, and that haplotypes can be assigned. It also helps ensure that the map-level errors are low.

2. What is the sensitivity to detect mosaic repeat contraction alleles?

Preliminary analyses showed that the FSHD pipeline has some sensitivity to detect such mosaic repeat contraction alleles. However, a full validation is needed to determine the limit of detection. The pipeline makes no assumption on the expected number of alleles during assembly. If there are sufficient molecules to form a consensus map, additional alleles may be assembled.

3. Why are some repeat counts prefixed by a “>=” sign?

In some cases, there may not be molecules fully spanning the repeats. These typically involve unusually long repeats (larger than 30 units). If a partial repeat is assembled, the pipeline tries to measure the length of the partial repeat and output a lower-bound estimate.

4. Why is the haplotype unknown?

In some cases, there may not be molecules spanning across the haplotype-specific labels. These typically involve unusually long repeats (larger than 30 units). It is possible that the repeat array is assembled without the haplotype-specific labels. The haplotype would be assigned as “unknown”, even though the repeat count or at least its lower bound could be measured.

5. How does de-duplication work?

The assembly pipeline sometimes generates maps that contain redundant D4Z4 information and maps with partial repeats if there are no molecules spanning to repeats.

If there are maps with *the same repeat counts and haplotypes*, the pipeline picks one representative one (with highest coverage).

If there are *at least two* full-repeat maps, the pipeline keeps the largest truncated map if it has a higher repeat count than the rest. If there is *one* full-repeat map, the pipeline takes largest truncated map regardless of the repeat count. If there are *no* full-repeat maps, the pipeline keeps the two largest truncated maps regardless of repeat counts. Users can then manually inspect truncated maps.

Technical Assistance

For technical assistance, contact Bionano Genomics Technical Support.

You can retrieve documentation on Bionano products, SDS's, certificates of analysis, frequently asked questions, and other related documents from the Support website or by request through e-mail and telephone.

Type	Contact
Email	**support@bionanogenomics.com**
Phone	**Hours of Operation:** **Monday through Friday, 9:00 a.m. to 5:00 p.m., PST** **US: +1 (858) 888-7663**
Website	**www.bionanogenomics.com/support**

Bionano Genomics, Inc.
9540 Towne Centre Drive, Suite 100
San Diego, CA 92121