Resolution of the D4Z4 repeat responsible for facioscapulohumeral muscular dystrophy with HiFi sequencing

Poster #8175T

Xiao Chen[1], Richard J. L. F. Lemmers[2], Zev Kronenberg[1], Joseph M. Devaney[3], Jessica Noya[3], April S. Berlyoung[3], Shamila Yusuff[3], Solomon Lynch[3], Keith Nykamp[3], Amanda S. Lindy[3], Egor Dolzhenko[1], Silvère M. van der Maarel[2], , Michael A. Eberle[1]

1. PacBio, Menlo Park, CA, USA 2. Department of Human Genetics, Leiden University Medical Center, Netherlands. 3. GeneDx, Gaithersburg, MD, USA

Introduction on D4Z4 and Kivvi

The D4Z4 repeat is a macrosatellite repeat that contains some of the most difficult-to-resolve medically relevant alleles in the human genome. D4Z4 has a repeat unit of 3.3 kb (encoding the *DUX4* gene) that is repeated between 1-100 times per allele. Facioscapulohumeral muscular dystrophy (FSHD) is caused by ectopic expression of *DUX4*, mediated by contraction of D4Z4 to less than 11 copies (FSHD1, 95% of FSHD cases) or hypomethylation of D4Z4 induced by mutations in the epigenetic machinery such as *SMCHD1* (FSHD2, 5% of FSHD cases). D4Z4 is present on both chromosome 4 and chromosome 10, while *DUX4* can only be expressed from the permissive A haplotype that usually occurs on chromosome 4. The D4Z4 repeat size is conventionally measured by Southern blot analysis, which is low-throughput and can be confounded by genetic polymorphism. Due to the large repeat size, high polymorphism and homology issues, D4Z4 remains difficult to resolve by sequencing technologies and population analysis is lacking. Here we present a computational tool, Kivvi, to resolve D4Z4 using PacBio HiFi whole-genome sequence data.

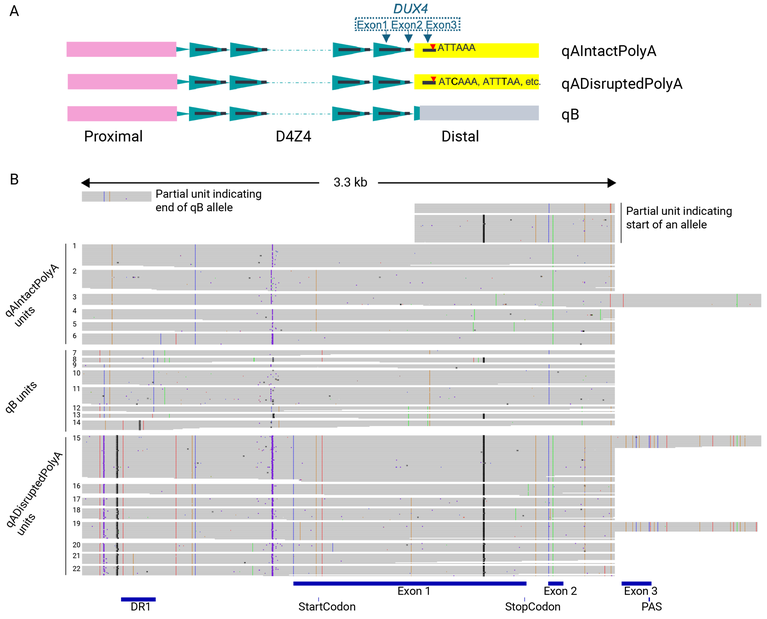


Figure 1. Overview of D4Z4 and Kivvi. A. Each repeat unit of D4Z4 encodes the first and second exon of *DUX4*. Proximal to D4Z4, chromosome 4 and chromosome 10 share sequence homology for a ~42 kb region. The third exon of *DUX4* is located in the sequence distal to D4Z4 and contains the somatic polyadenylation signal (PAS), denoted by the red triangle. D4Z4 alleles can be classified into three types based on the distal region: qA with an intact polyA signal (qAIntactPolyA), qA with a disrupted polyA signal (qADisruptedPolyA) and qB, which completely lacks the third exon and polyA signal. **B.** HiFi reads are realigned against one D4Z4 unit, plus a short region distal to D4Z4, including the PAS. Read segments are classified into unique repeat units based on sequence variants, and these unique repeat units are labeled with distinct indices (1-22 in this example sample HG02071). All alleles start with a partial unit, indicated by the shorter read segments at the top right. All qB alleles end with a partial unit indicated by the shorter read segments at the top left. All qA alleles (both qAIntactPolyA and qADisruptedPolyA) continue to the distal sequence containing the third exon. Repeat units are assembled into alleles, which are then assigned to chromosomes 4 or 10 based on the proximal sequences.

Validation of Kivvi calls for qAIntactPolyA alleles

We compared Kivvi results against pulse field gel electrophoresis (PFGE) in 54 control samples from the 1000 Genomes Project (1kGP) and against optical genome mapping (OGM) in 10 FSHD1 patient samples.

Table 1. Validation of D4Z4 calls against PFGE-Southern blot and Optical Genome Mapping.

Concordance is calculated as (fully assembled with correct size + partially assembled and correctly classified)/total alleles.

Dataset	Allele size (RU)	Total alleles	Fully assembled with correct size	Partially assembled and correctly classified[1]	Inconclusive	Concordance
1kGP	1-10	1	1	0	0	1
	>10	61	15	43	3[2]	0.951
	total	62	16	43	3	0.952
FSHD1[3]	1-10	10	10	0	0	1
	>10	4	0	2	2	0.5
	total	14	10	2	2	0.857

[1]Correctly classified as contracted (1-10RU) or >10RU alleles. [2]The three inconclusive alleles are reported as >=9RU, >=9RU and >6RU.
[3]This dataset has lower data quality leading to lower performance in longer alleles.

Differential methylation signals between FSHD1 samples and samples with *SMCHD1* variants

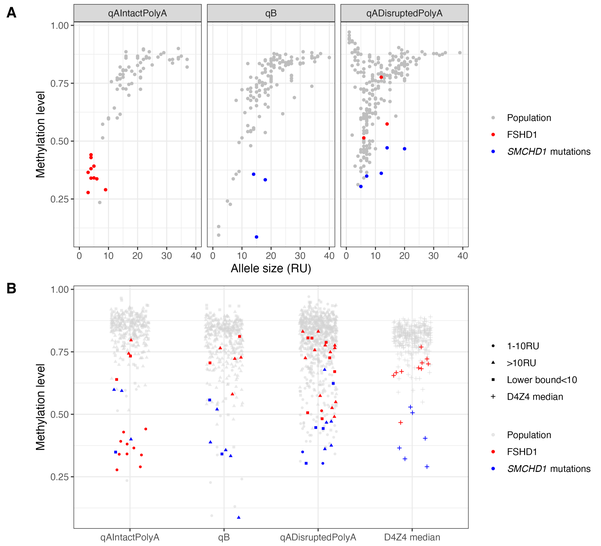


Figure 2. Size and methylation levels of D4Z4 alleles in population samples, FSHD1 patient samples and individuals with *SMCHD1* pathogenic variants. A. Distribution of allele size and methylation levels in fully assembled alleles. **B.** Methylation levels of all D4Z4 alleles, including partially assembled alleles. The data points marked as "lower bound<10" indicate a partial allele where the reported lower bound of allele size is shorter than 10 and thus cannot be classified into 1-10RU or >10RU. The last column plots the median methylation level of all D4Z4 units for each sample.

Sequence analysis across populations

We analyzed 14,325 D4Z4 repeat units from 791 fully assembled D4Z4 alleles and performed principal component analysis (PCA) on these repeat units based on their sequence variants. PCA separated repeat units into three clusters that correspond to qAIntactPolyA, qB and qADisruptedPolyA. Hybrid repeat units exist between qADisruptedPolyA and qAIntactPolyA (gray-bordered box).

The three types of repeat units are differentially distributed against different genetic backgrounds defined by chromosomes and proximal haplotypes across populations, allowing us to identify hybrid repeat arrays and translocation alleles.

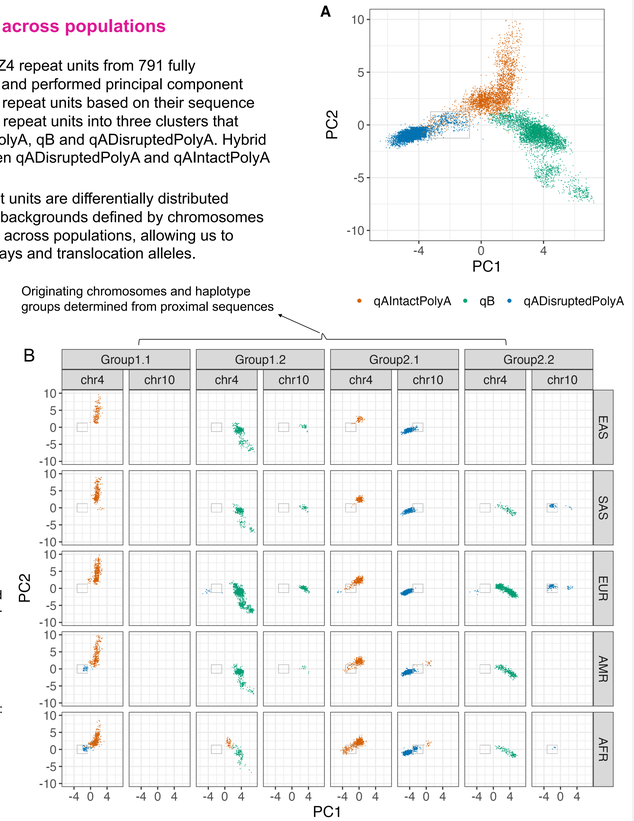


Figure 3. Sequence variations of D4Z4 repeat units and upstream regions across populations.

A. Principal component analysis separates D4Z4 repeat units into three clusters. Every repeat unit on a qAIntactPolyA, qB or qADisruptedPolyA allele is labeled as qAIntactPolyA, qB or qADisruptedPolyA, respectively. The gray-bordered box denotes the qAIntactPolyA-qADisruptedPolyA hybrid zone.

B. Same PCA plot as in A, showing repeat units grouped by population, chromosome and upstream haplotypes. EAS: East Asians, SAS: South Asians, EUR: Europeans, AMR: Admixed Americans, AFR: Africans.

Conclusion

We developed Kivvi for genotyping D4Z4 using HiFi WGS sequencing data and validated Kivvi against gold standard assays for FSHD diagnostics.

We showed differential methylation signals between FSHD1 and candidate FSHD2 samples.

Across 601 individuals from five ancestral populations, we identified common haplotypes of D4Z4 alleles, hybrid repeat units, hybrid repeat arrays, and translocation alleles.

Kivvi brings us one step closer to the consolidation of multiple FSHD assays into a single workflow and will help facilitate the discovery of novel genetic modifiers of FSHD through population-scale studies.

Preprint

Kivvi on GitHub

Research use only. Not for use in diagnostic procedures. ©2026 Pacific Biosciences of California, Inc. ("PacBio"). All rights reserved. Information in this document is subject to change without notice. PacBio assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of PacBio products and/or third-party products. Refer to the applicable PacBio terms and conditions of sale and to the applicable license terms at pacb.com/license. Pacific Biosciences, the PacBio logo, PacBio, Circulomics, Omniome, SMRT, SMRTbell, Iso-Seq, Sequel, Nanobind, Revio, Revio SPRQ, Kinnex, PureTarget, SPRQ, Vega and Vega SPRQ are trademarks of PacBio.