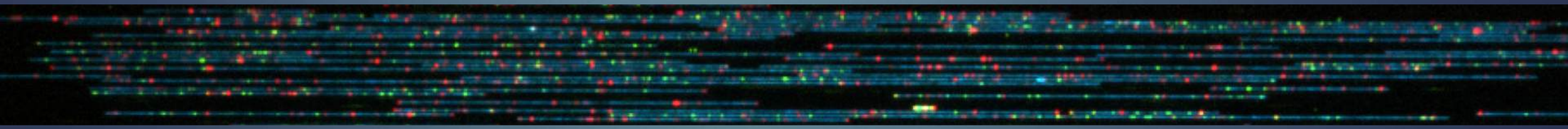
Applications of Optical Genome Mapping



Hayk Barseghyan, PhD
Research Faculty
Center for Genetic Medicine Research
Children's National Medical Center

October 1st, 2019

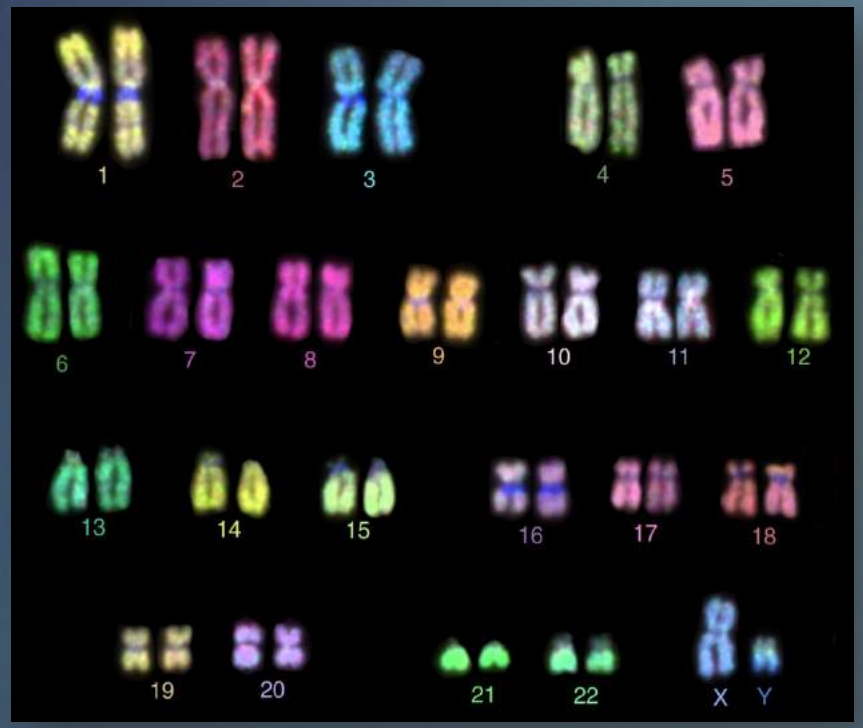
Overview

- Part 1:
  - Structural variant identification via optical mapping
- Part 2:
  - Investigation of haplotype specific epigenetic regulation
- Part 3:
  - *De novo* genome assembly

Part 1:

Structural variant identification via optical mapping

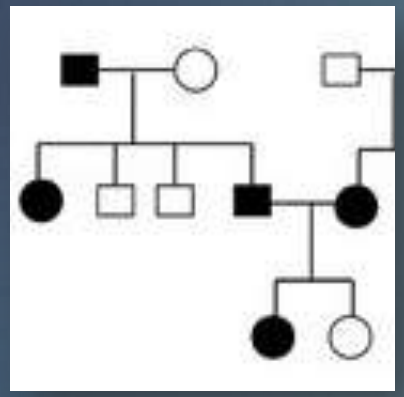
Genetic Diseases

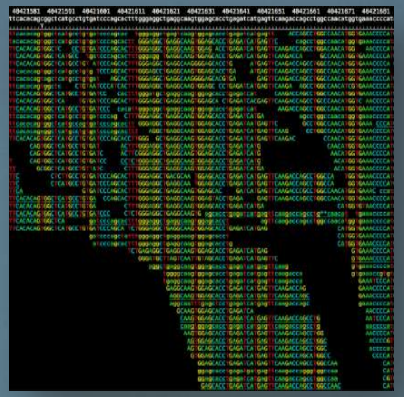


3.2 billion bases (ATGC), ~20,000 genes

- There are >6,000 different genetic disorders
- Affect millions of people around the world
- Difficult to diagnose or predict the age of onset
- Not enough tests performed globally

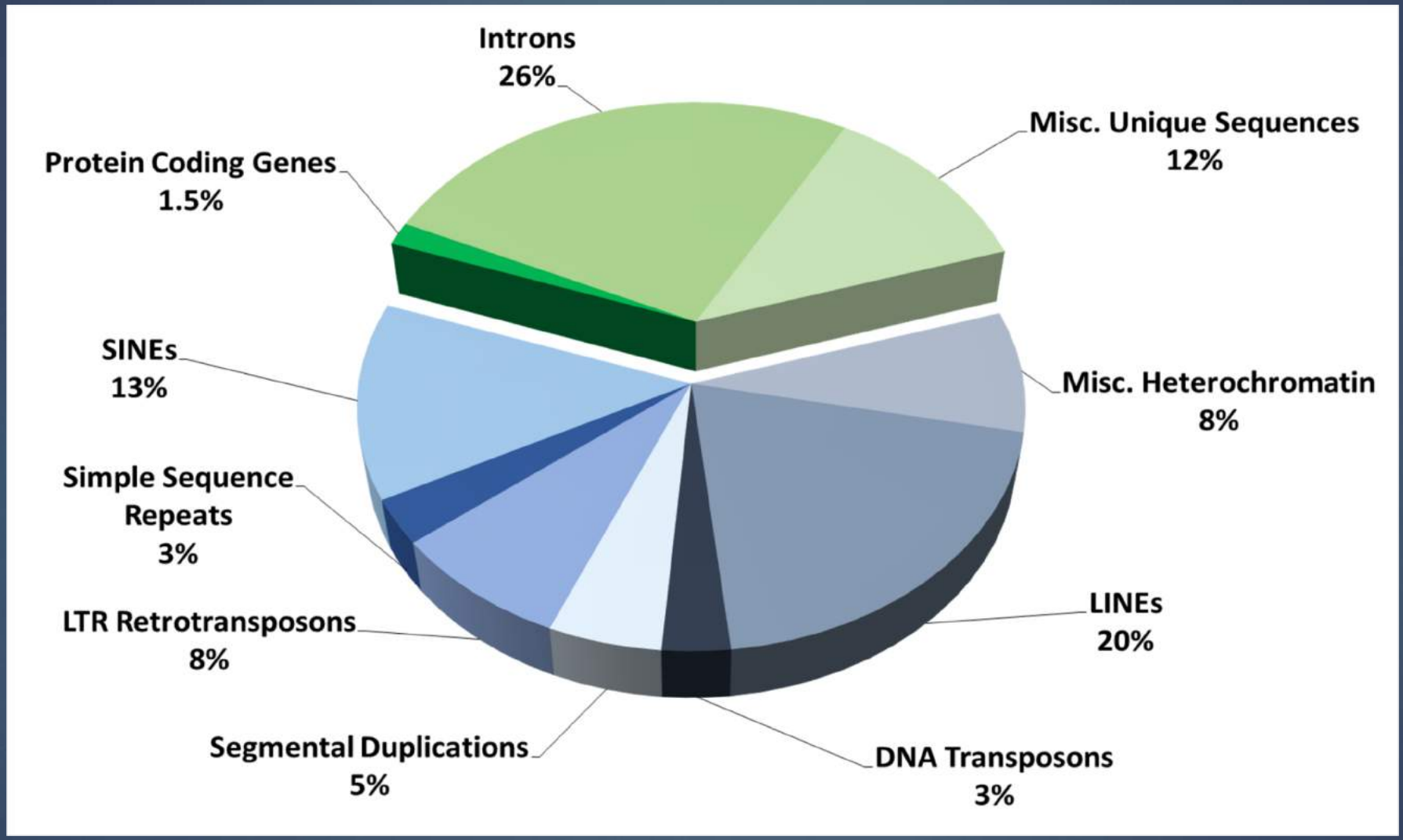
Precision Health





- SNVs small INDELs/CNVs
- Misses on identification of large SVs

Human Genome Composition



Human Genome Project

Genomic Technologies Compared

Illumina

~150-300 bp
SNV
SV

$1000

PacBio

~5-50 kbp
SNVs
SV

$5000

Oxford Nanopore

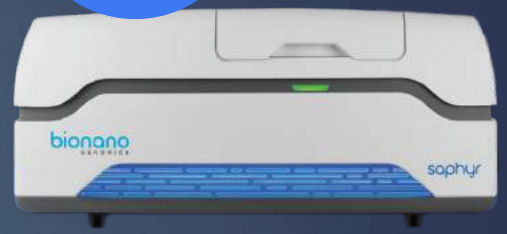
~10-50 kbp (no limit)
SNVs
SV

$4000

Agilent aCGH

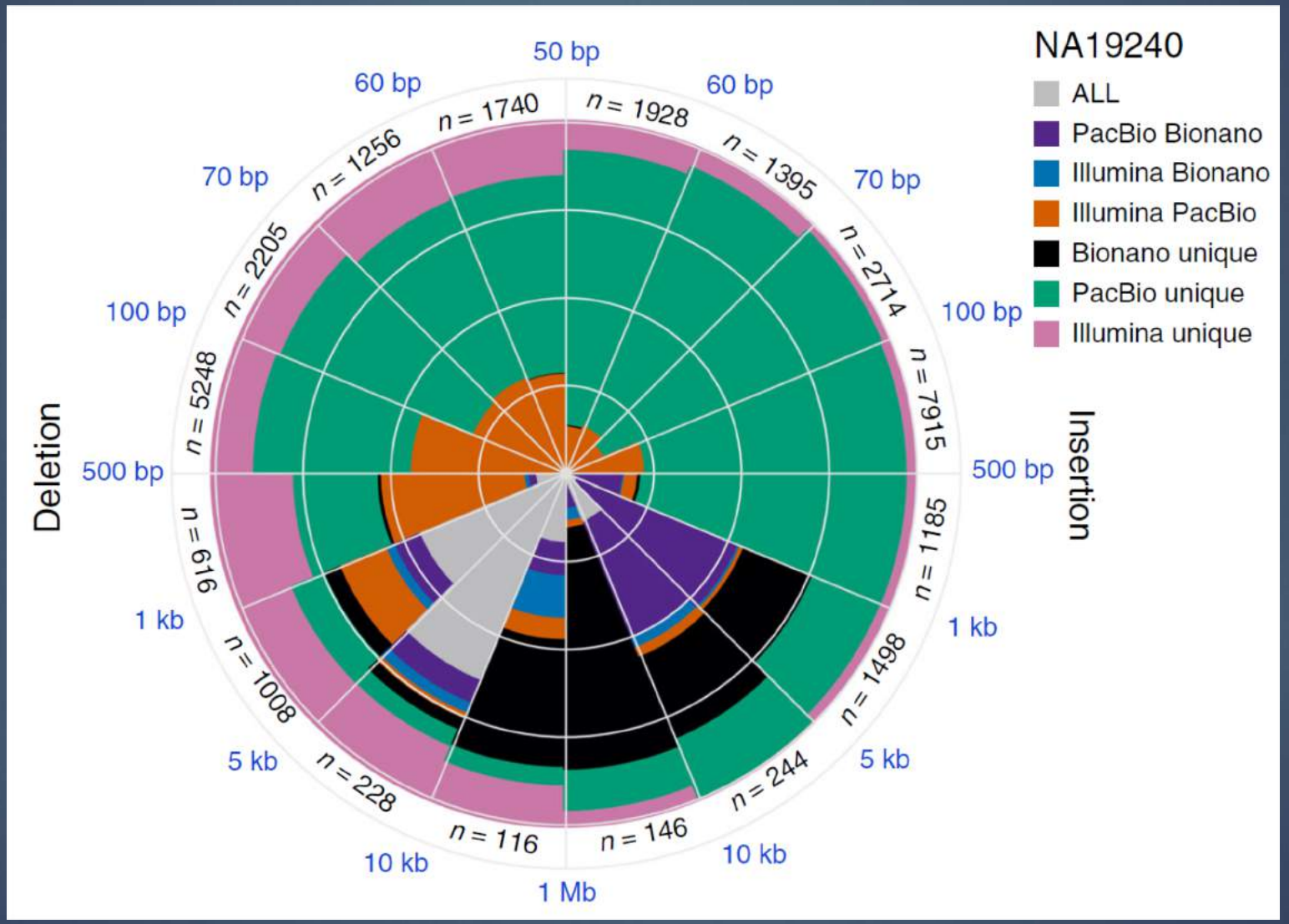
Hybridization
SNV
CNV
> 30kbp
$500

Bionano



~ 300 kbp (no limit)
SNV
SV
> 500 bp
$1000

Structural Variation Identification by Technology

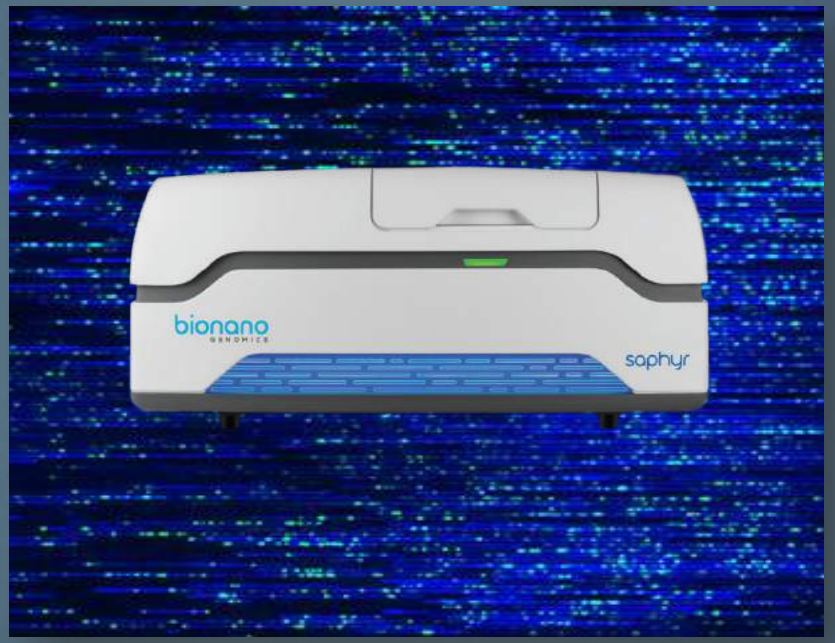


Chaisson M, 2019

Optical Genome Mapping (OGM)

DNA molecules
native-state (Mb size)

labeled in specific
sequence motifs



produces high-
resolution images of
long DNA molecules

generates optical
map of the genome

repeat expansions, deletions,
insertions, inversions, translocations
and complex rearrangements

90% sensitivity for large
structural variants >1.5kb

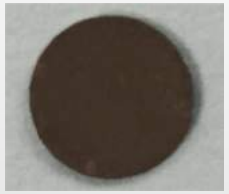
Standard OGM Methodology

High Molecular Weight DNA

1. Bio-Rad CHEF Genomic DNA Plug Kits

WBC or Cultured cells →  → Lysis

2. Circulomics HMW DNA Extraction Kit

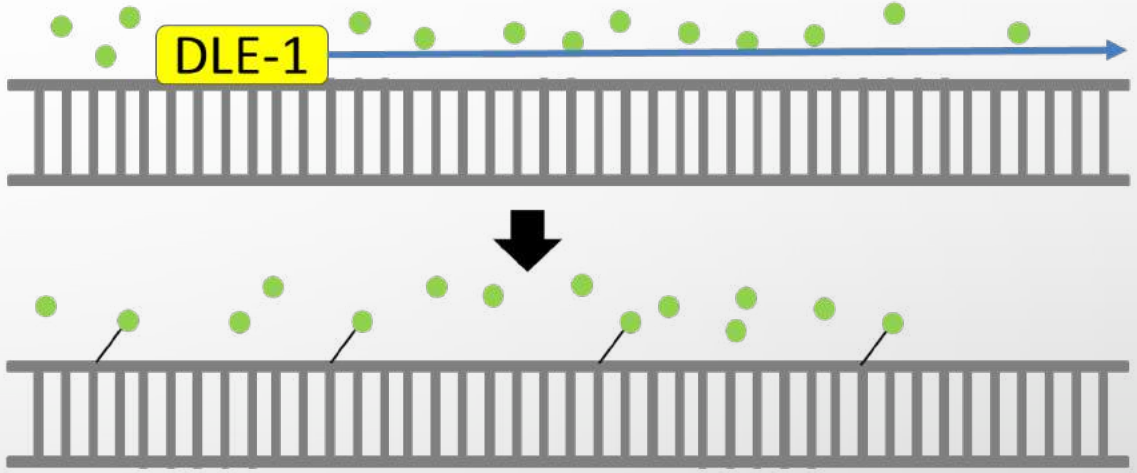
WBC or Cultured cells → Lysis → 

3. Bionano Genomics

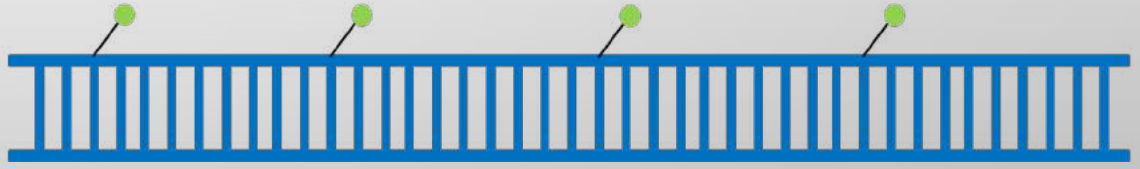
Modified versions of 1 and 2

DLS (Direct-Label, Stain)

1. DLE-1 enzyme recognizes CTTAAG sites and covalently attaches fluorescent labels



2. Staining of the backbone



DNA Loading/Analysis

Raw Optical Images

↓

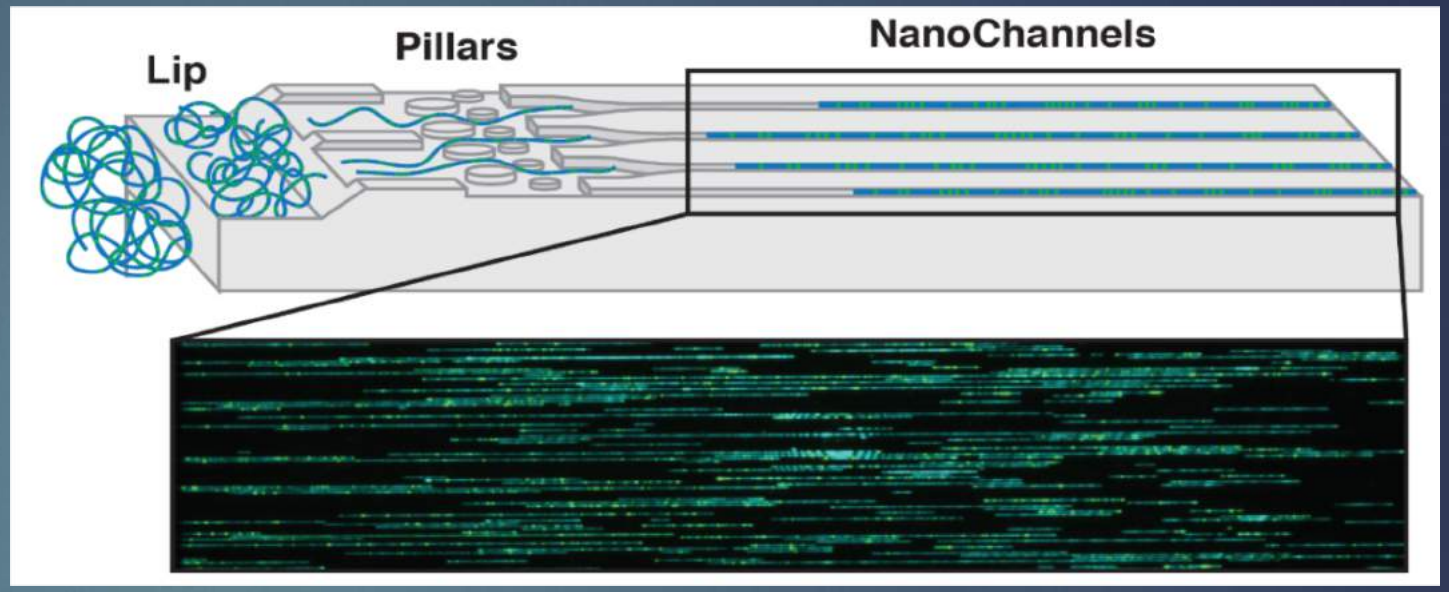
Conversion of images into strings retaining label patterns

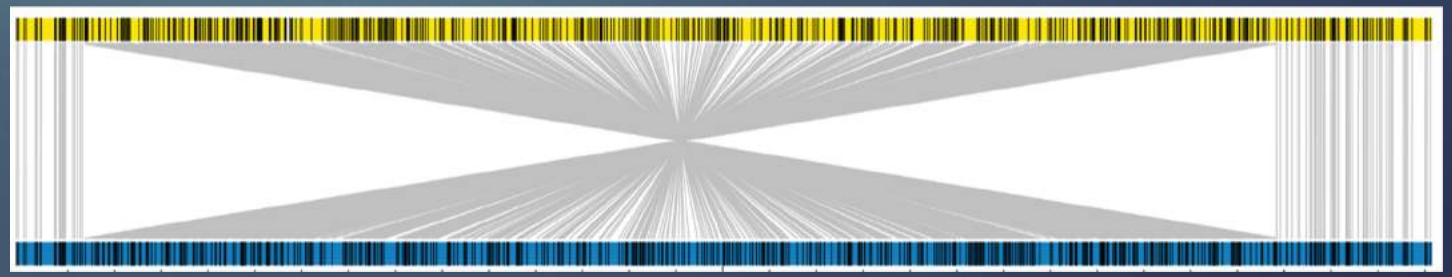
↓

De novo genome assembly

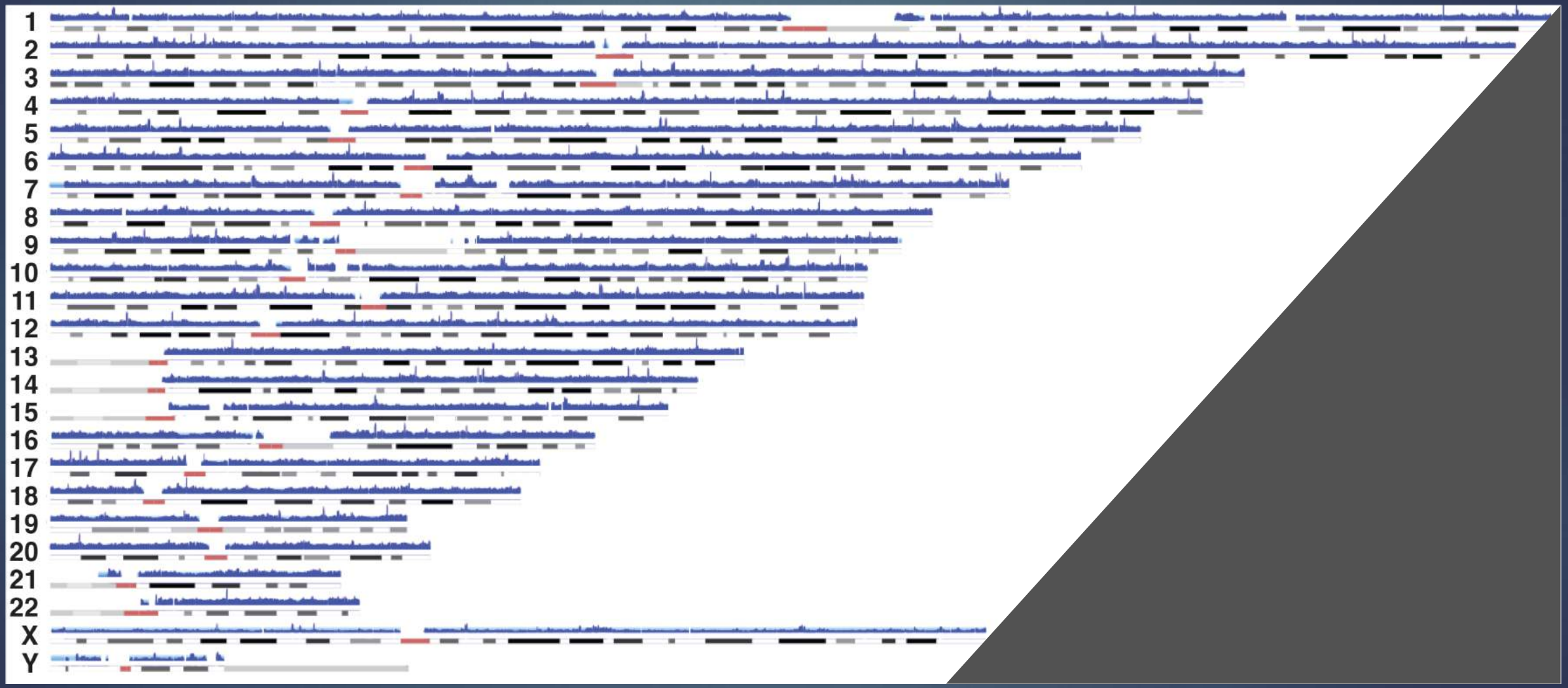
↓

Structural Variant Calling





GENOME COVERAGE



Distribution of Identified SVs

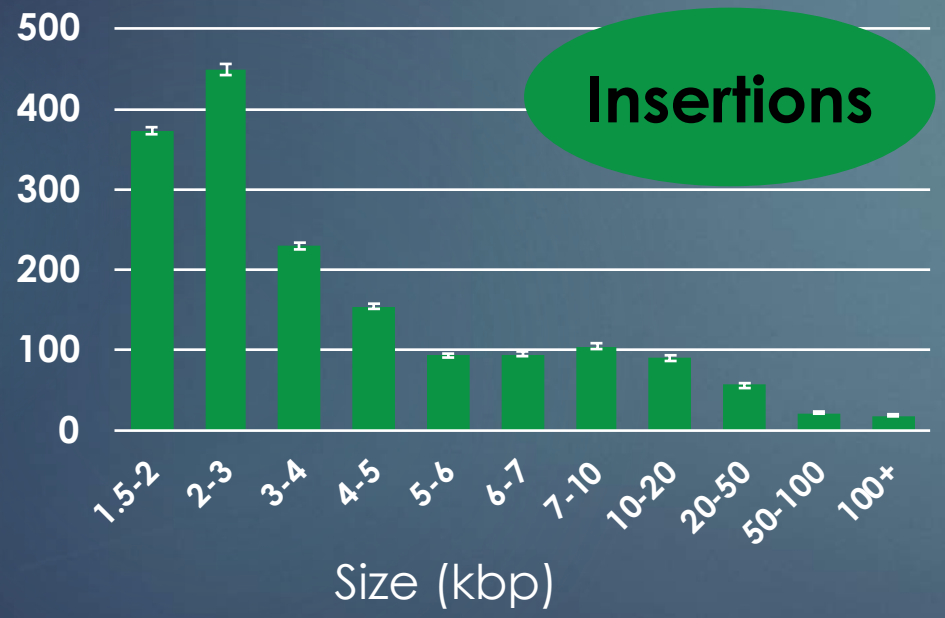
~1600 Insertions

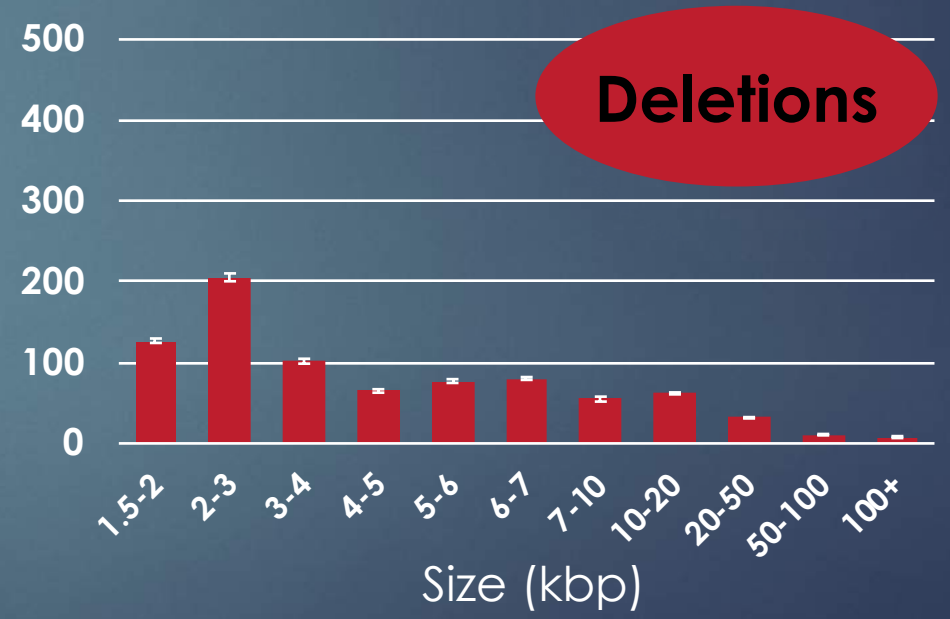
~440 Duplications

~800 Deletions

~40 Inversions

0-2 Translocations





Duchenne Muscular Dystrophy

https://www.statnews.com/2015/11/19/podcast-duchenne-biomarin-sarepta/

characterized by muscle degeneration

diagnosed predominantly in males 2-4 yrs age due to muscle weakness

X-linked recessive disorder

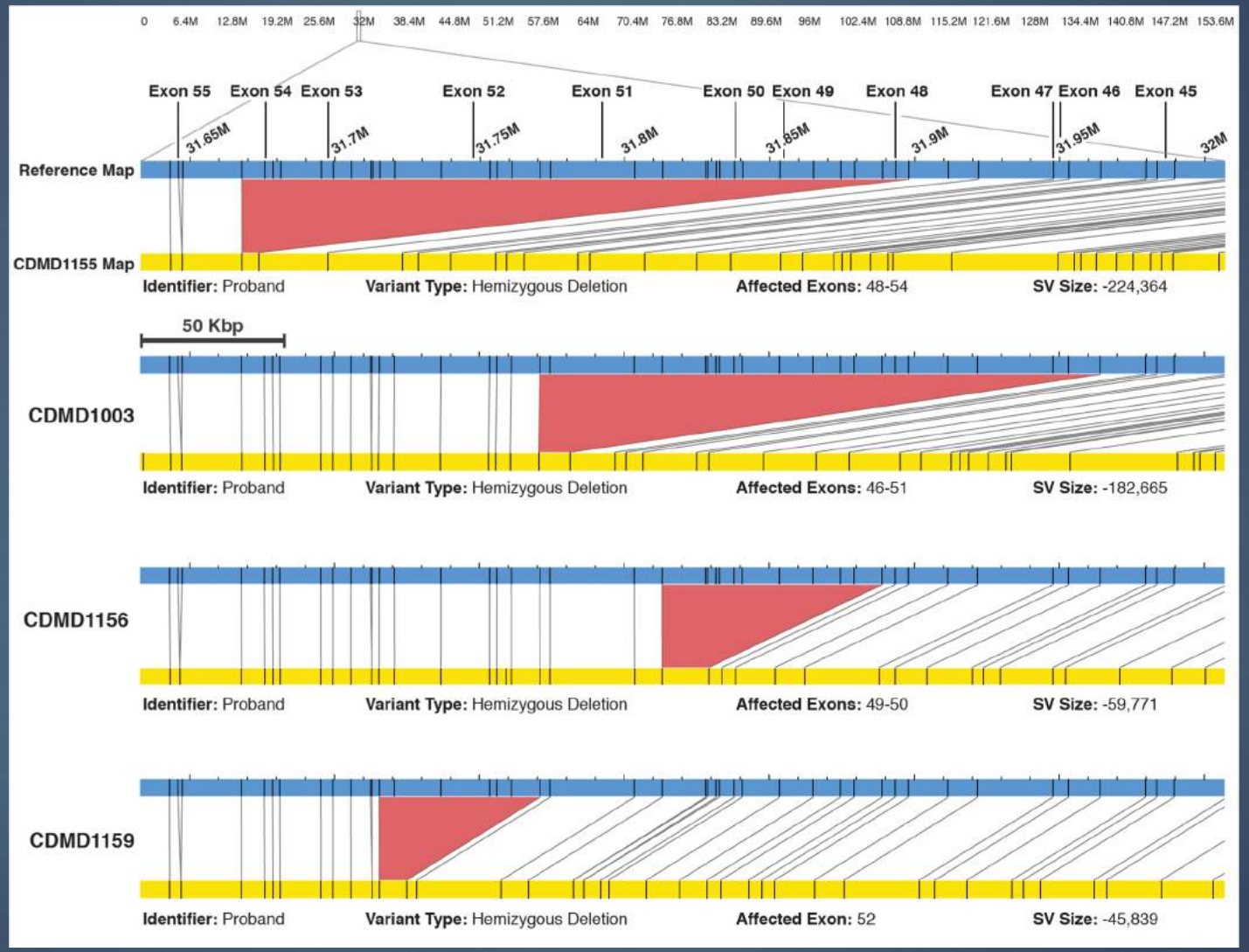
frequency 1:5000 males

caused by mutations in *DMD* gene Xp21

important for muscle cell fiber membrane maintenance

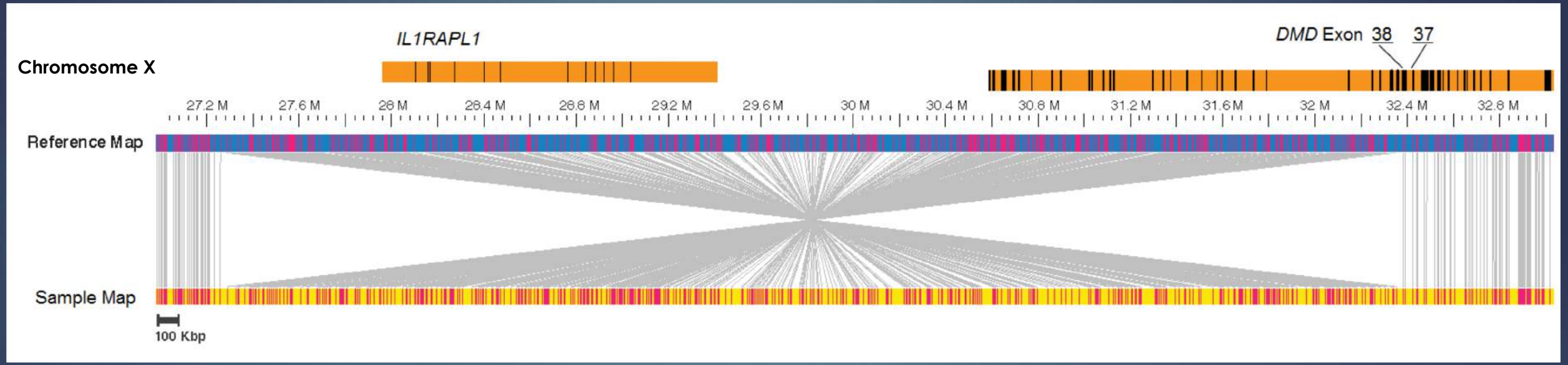
average life expectancy - 26

OM identifies hemizygous multi-exon deletions in DMD patients



Barseghyan H, et al., Genome Med 2017 Oct 25;9(1):90.

OM identifies large inversion in *DMD* gene

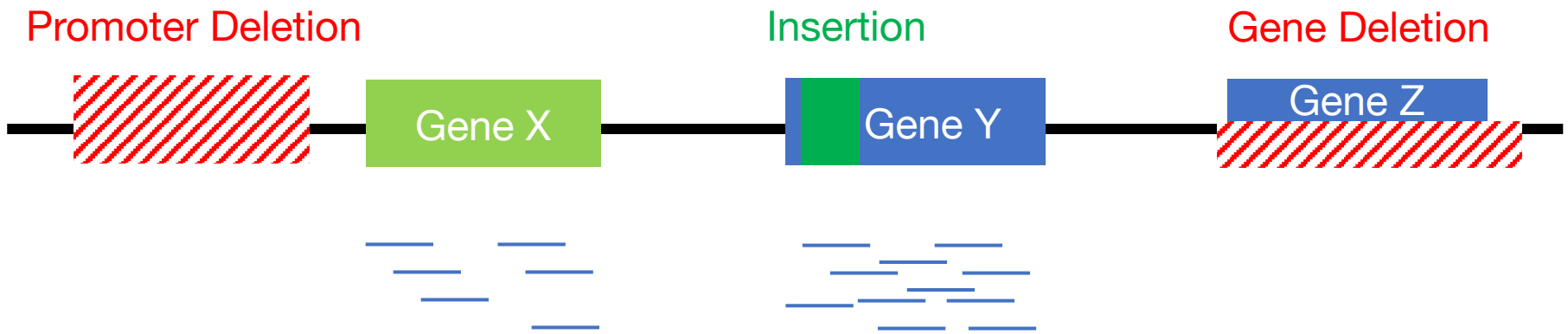


Barseghyan H, et al., Genome Med 2017 Oct 25;9(1):90.

OGM in Genetics

- **An Integrative Framework For Detecting Structural Variations In Cancer Genomes**
  - *Nature Genetics* volume 50, pages1388–1398 (2018)
- **Clinical application of single-molecule optical mapping to a multigeneration FSHD1 pedigree**
  - https://doi.org/10.1002/mgg3.565
- **Genome maps across 26 human populations reveal population-specific patterns of structural variation**
  - Nature Communications volume 10, Article number: 1025 (2019)
- **Long-read single-molecule maps of the functional methylome**
  - Sharim et al Genome Research 2019

nanotatoR Structural Variant Annotation



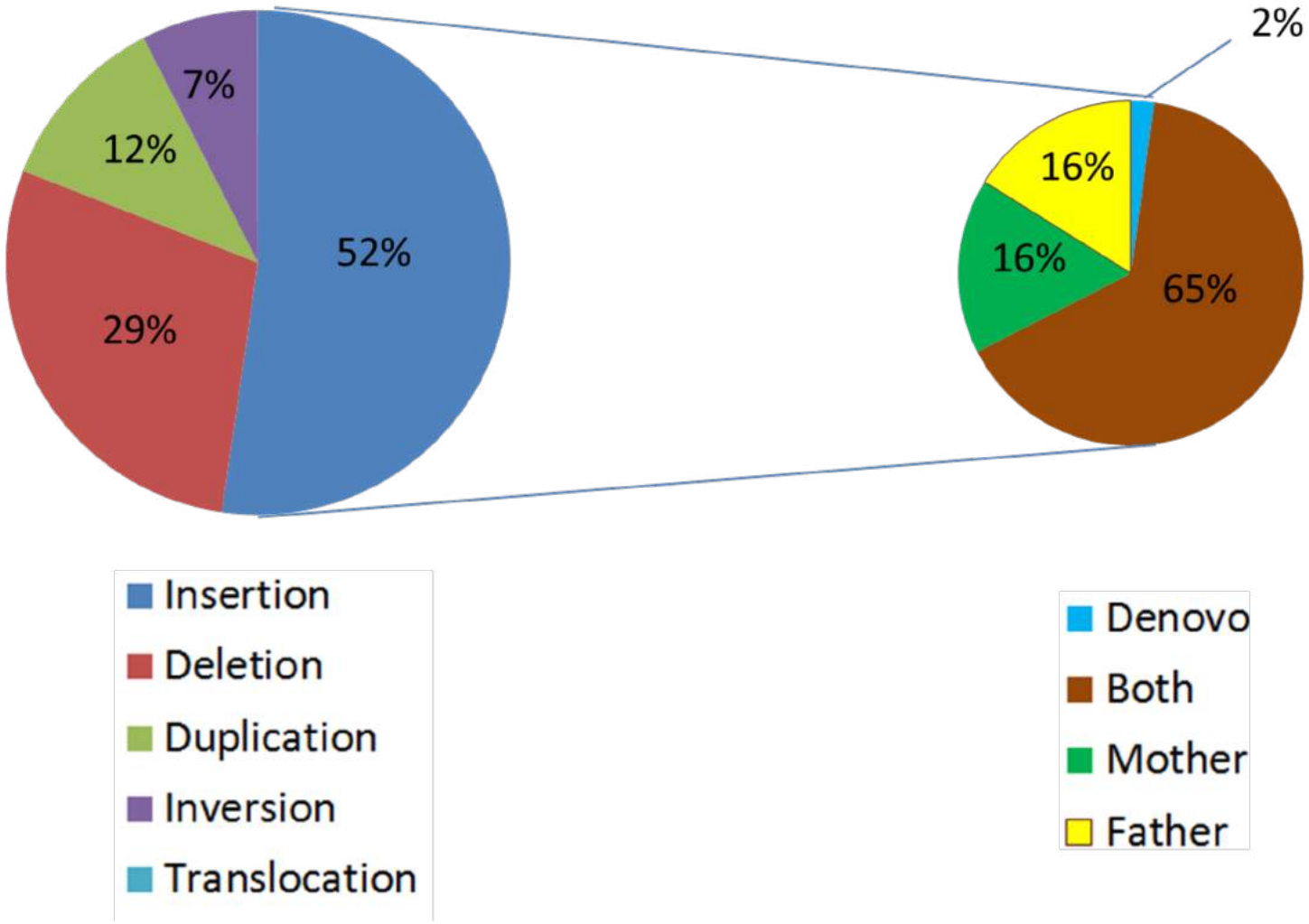
SV Type	Overlapping Genes	Nearest Genes	Population Frequency	Cohort Frequency	RNA-Seq Proband	RNA-Seq Father	RNA-Seq Mother	Phenotype
Del	-	Gene X	0.013%	0.8 %	2	26	27	Myopia
Ins	Gene Y	Gene Z	0.1%	0.2 %	40.5	38	43	Autism

- Integration of genomic, transcriptomic data and public repositories
- Variant classification and prioritization

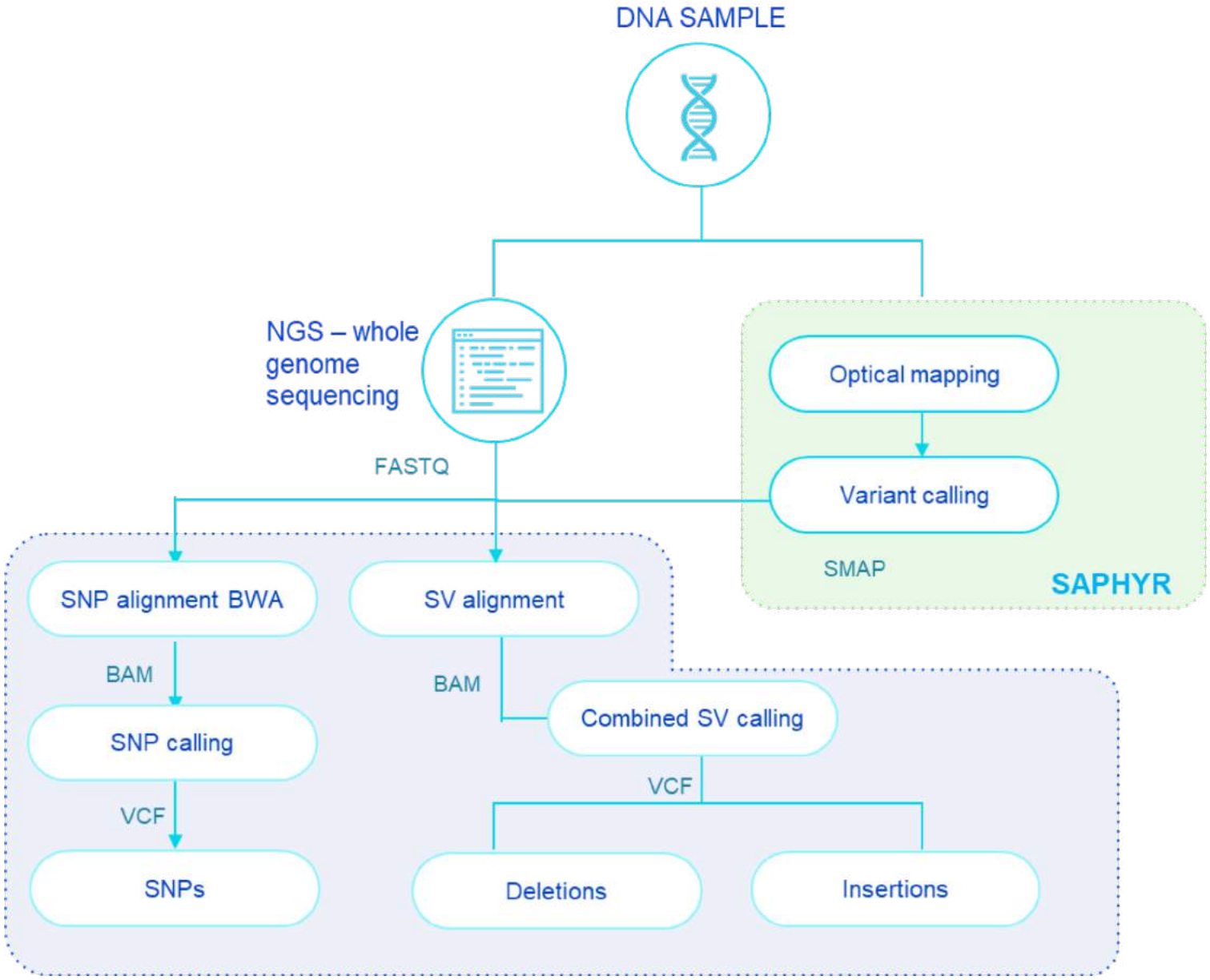
https://github.com/VilainLab/nanotatoR

Manuscript in preparation

Number and Distribution of Identified SVs



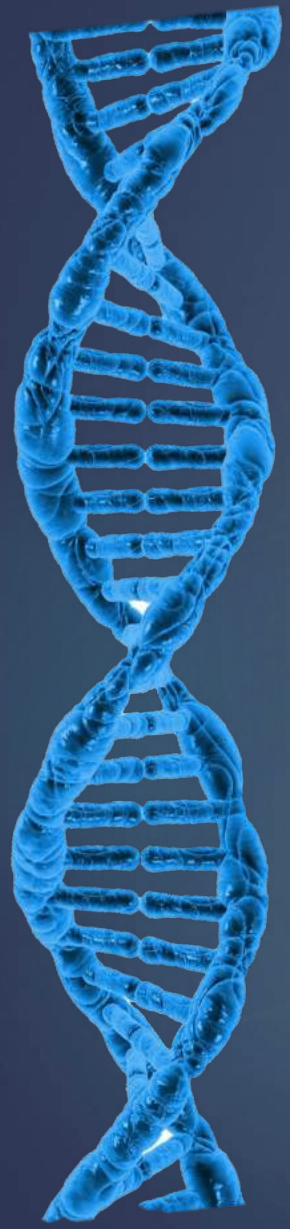
Integration of Optical Mapping and Short-Read Sequencing



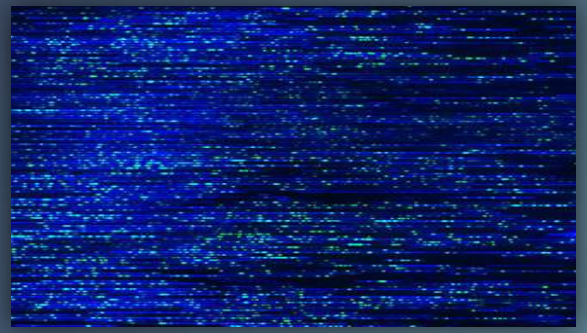
CNMC
Genoox
Bionano

SV Type	Breakpoint Position Error	
	OM Breakpoints	Refined Breakpoints
Deletions	6,367 bp	8 bp
Insertions	4,313 bp	22 bp

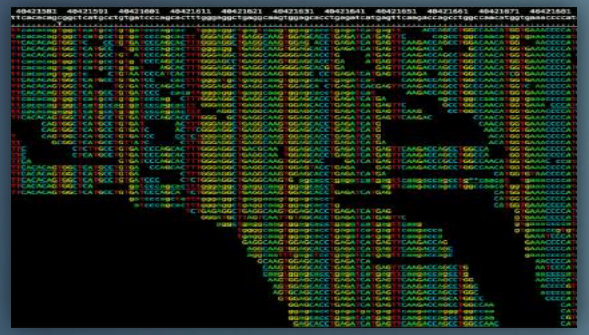
Comprehensive Variant Discovery



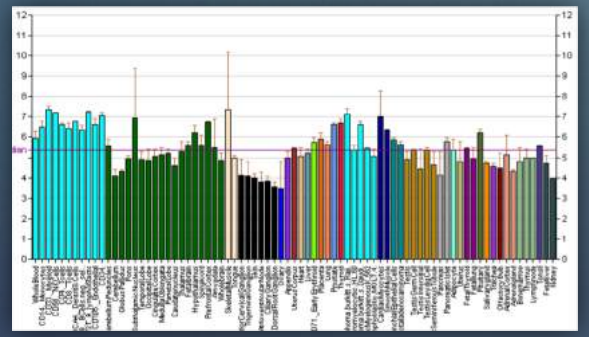
Map



Sequence



Expression



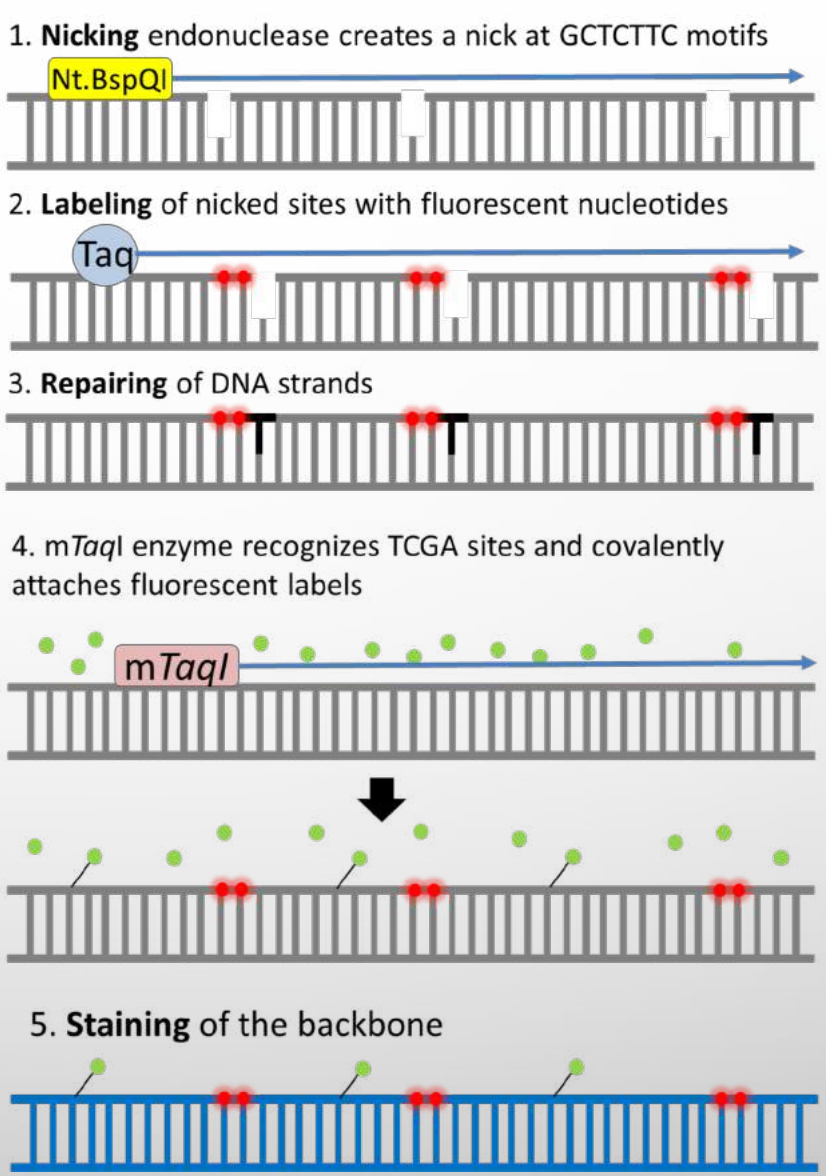
- Integration of NGS/NGM technologies
- Identification of SNVs; INDELs; CNVs; SVs
- Corresponding effect on gene expression

Part 2:

Investigation of haplotype specific epigenetic regulation

Dual Labeling Methodology

- **1st Label – RED** (BspQI or BssSI)
  - Used for genome assembly
- **2nd Label – GREEN** (m*TaqI*)
  - Used for quantification of non-Methylation
- **Stain – BLUE**
  - Used for molecule sizing

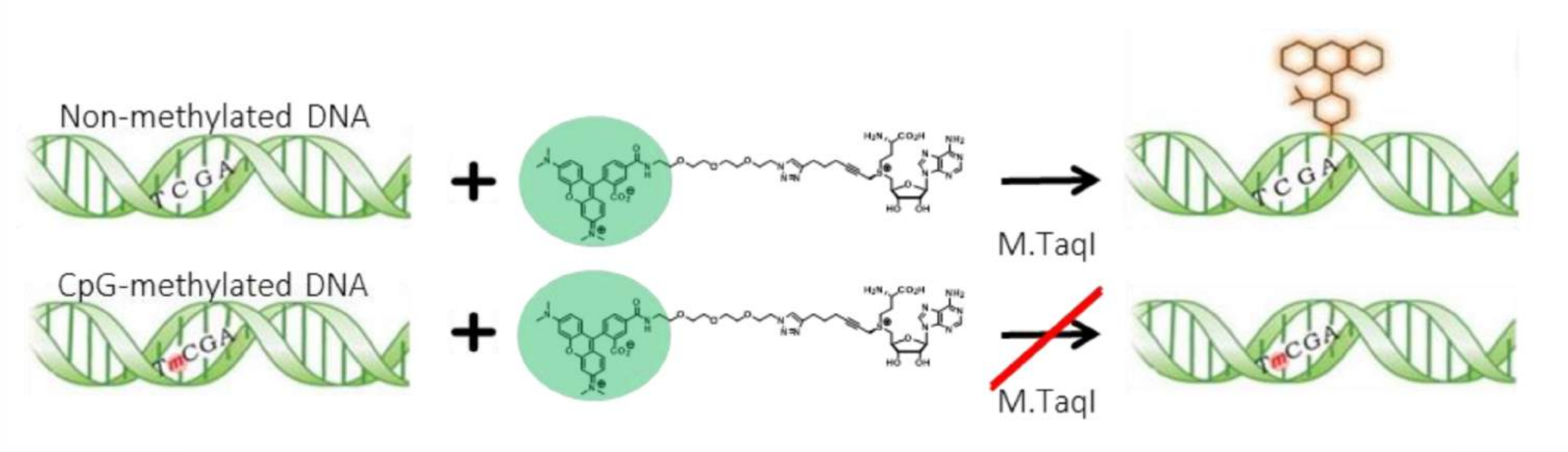


Genetic Level

epiGenetic Level

Sharim H, et al., Genome Res. 2019 Apr;29(4):646-656.

Dual Labeling Methodology

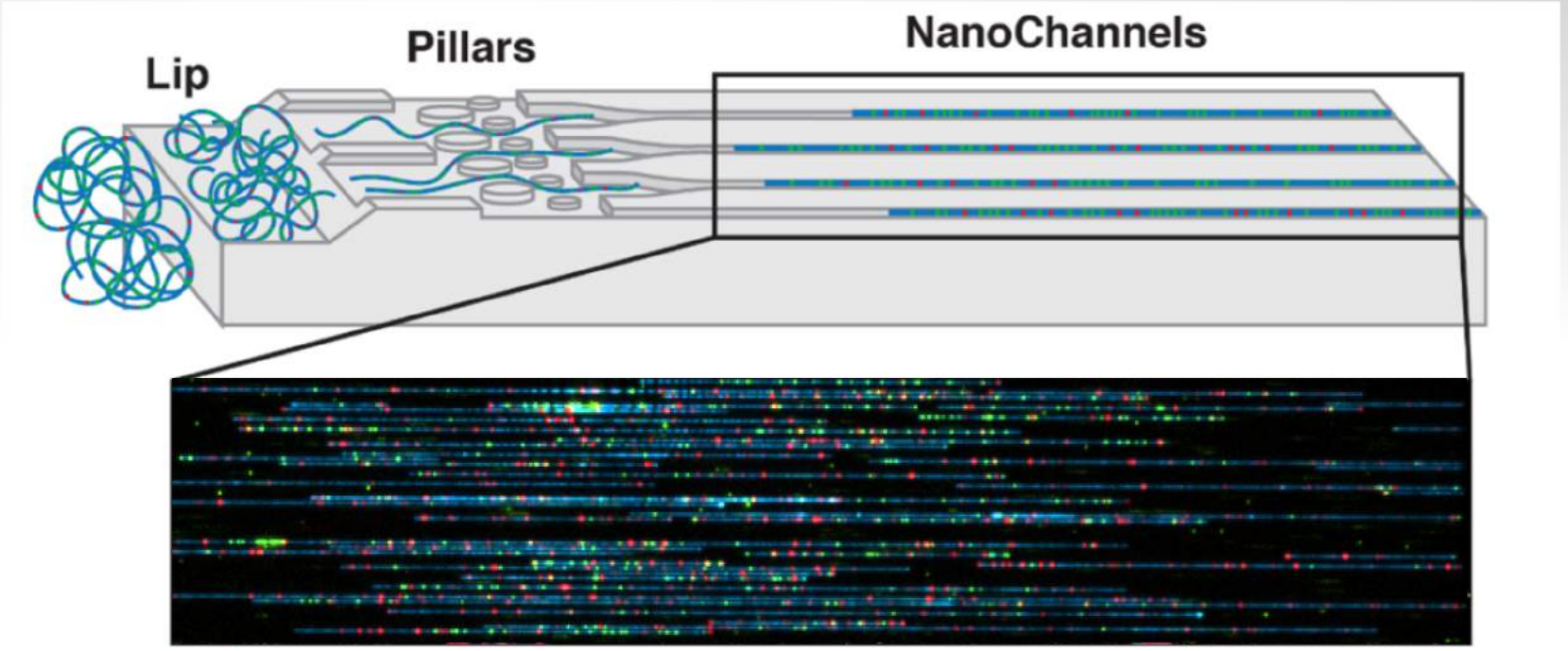


- M.TaqI generally methylates the adenine at TCGA sites; however, it can be tricked to incorporate a fluorophore (synthetic cofactor analog) instead of a methyl group
- The reaction is blocked when the nested CpG dinucleotide in the recognition site is methylated or hydroxymethylated
- The method acts as a fluorescent reporter for non-methylated CpGs within TCGA sites
- Labeling efficiency – 90%; specificity – 99.9%

Sharim H, et al., Genome Res. 2019 Apr;29(4):646-656.

DNA Loading/Analysis

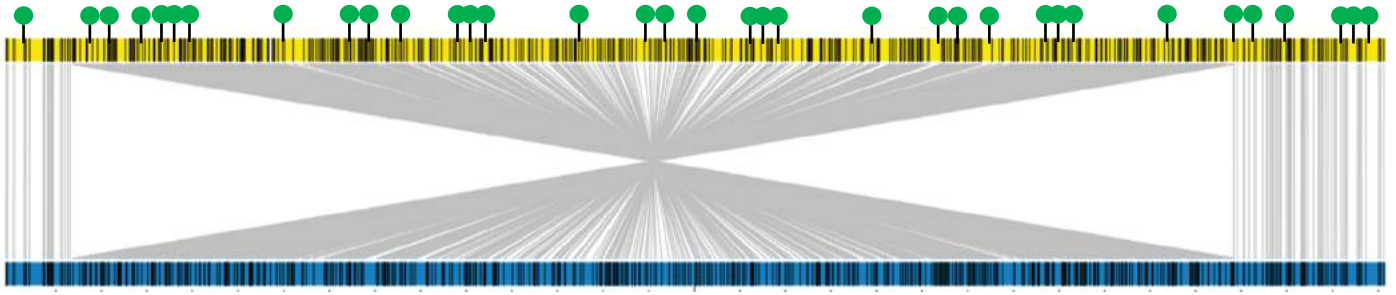
Raw Images
(.tiff)



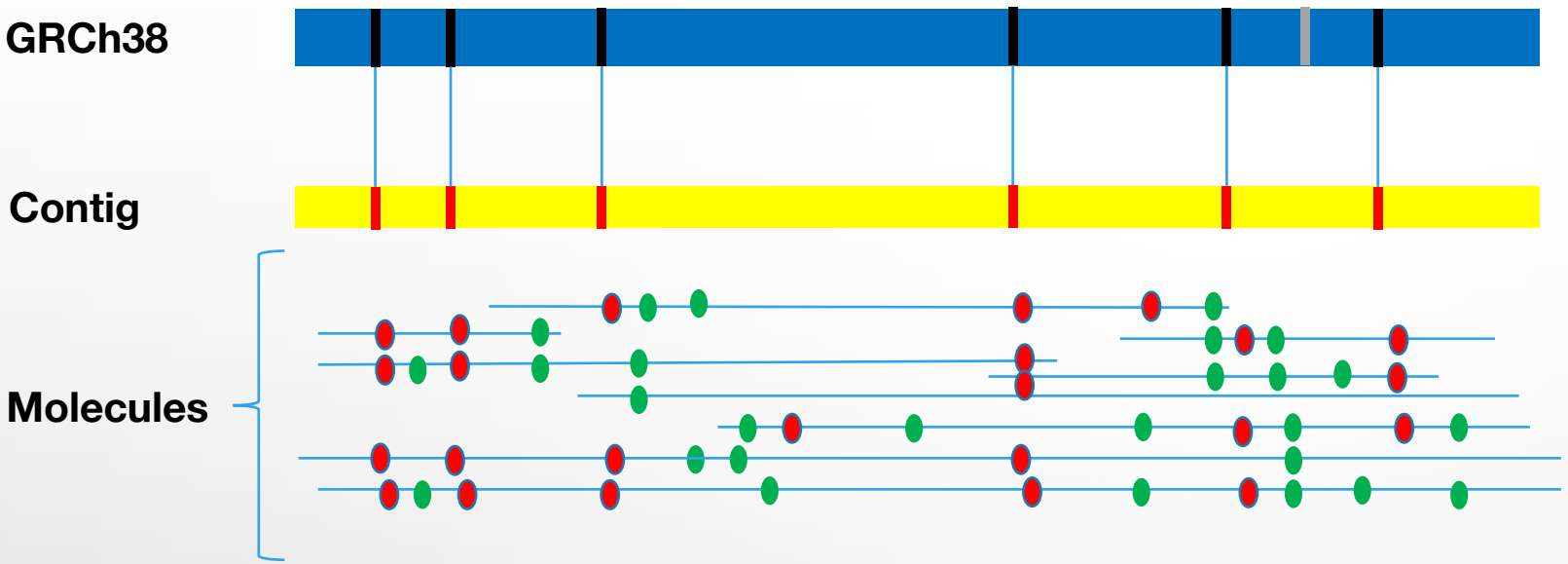
Conversion of images into text files
(.bnx)

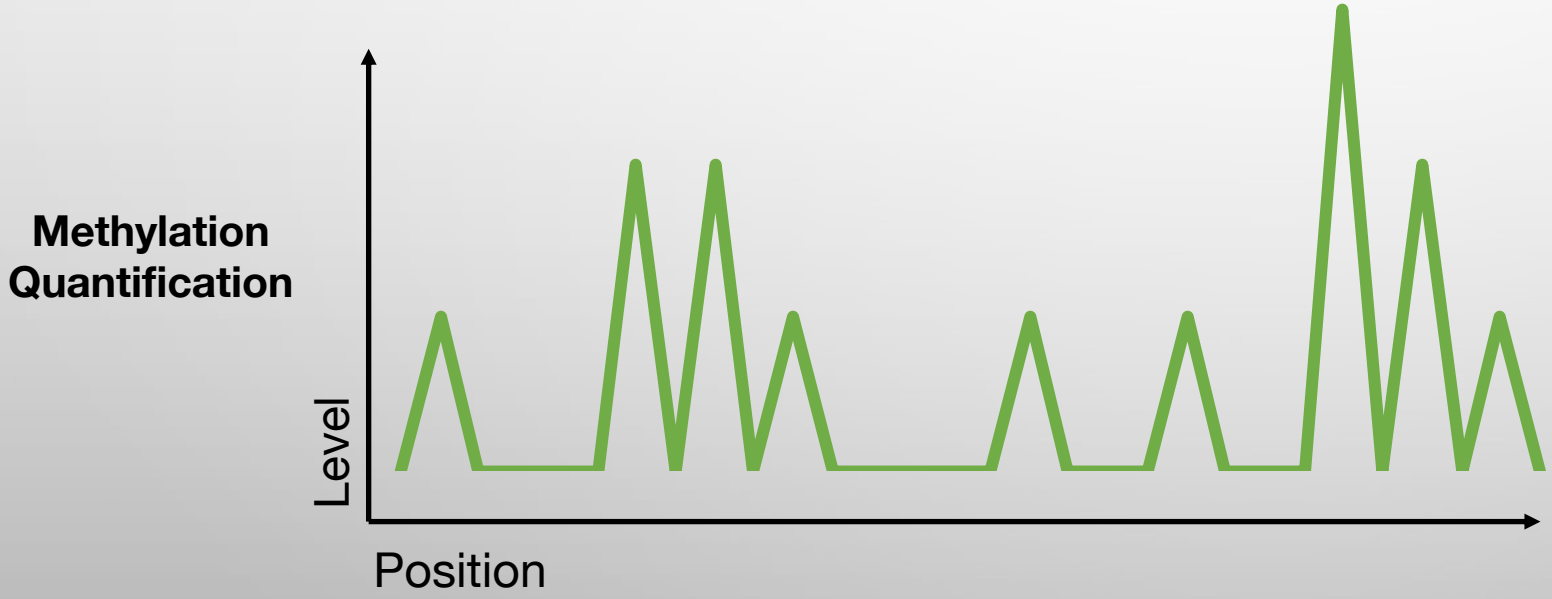
Genome assembly

Identification of structural
variants and methylation profiles
(.cmap)



methometR: OM Methylation Quantifier





- Reference genome nick sites aligned with contig nick sites (black with red)
- Contig CMAP containing consensus nick position
- Molecule CMAP containing nick and non-Me positions

methometR:

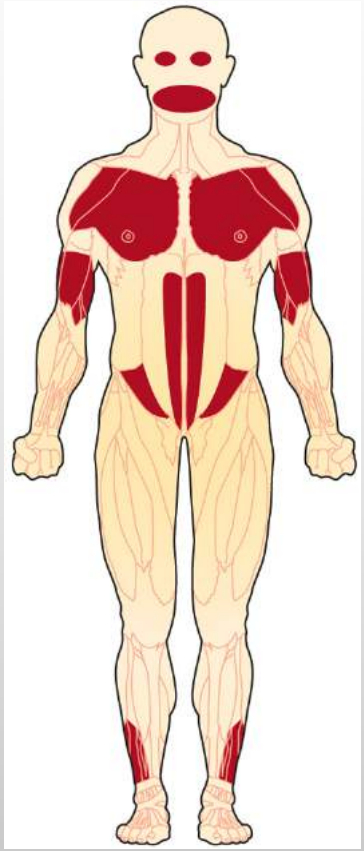
1) Translates non-Me label positions from molecules to contigs with ~ RefGen coordinates
2) Estimates non-Me levels by label count/spanning molecules

Advantages of Dual Label OGM (DL-OGM)

1) Haplotype specific methylation patterns
   - ➢ Study imprinting

2) Quantification of methylation in repetitive regions
   - ➢ Study effects of methylation in repetitive elements

Facioscapulohumeral Muscular Dystrophy

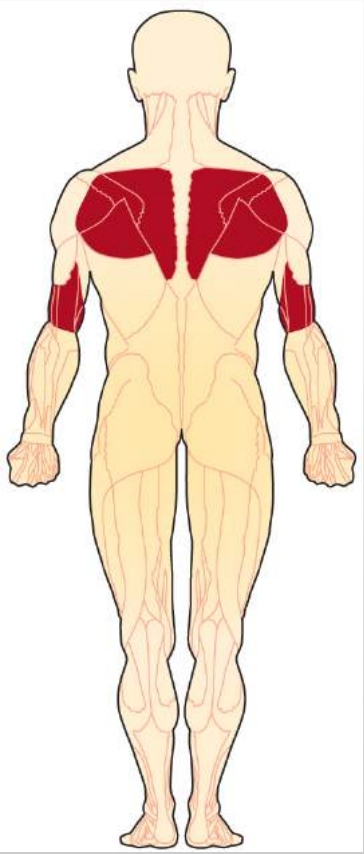


characterized by muscle degeneration

face, shoulder blades and upper arms

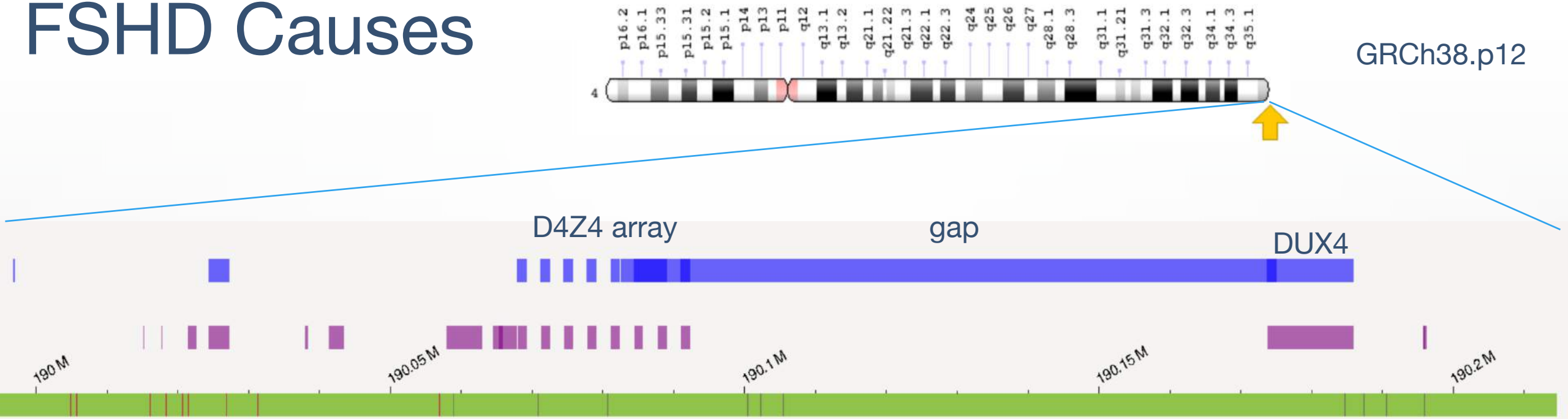
age of onset 20

prevalence of 1:7,000-20,000



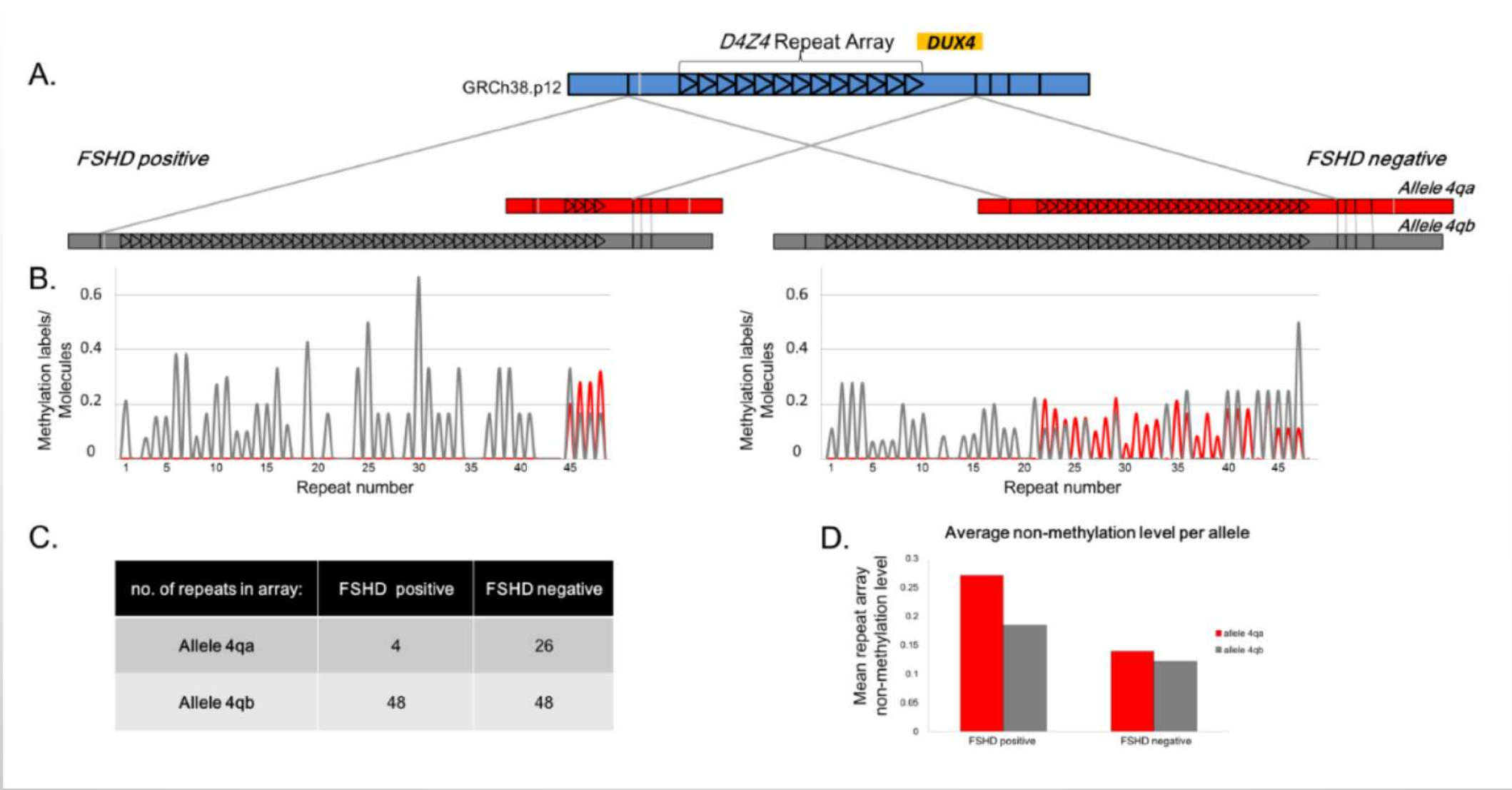
www.mda.org

FSHD Causes



- FSHD1: 95% - contraction of D4Z4 repeats (3.3kbp)
  - Normal 11-100; Pathogenic 1-10 repeats
- pLAM – PolyAdenilation sequence 4qA has a functional seq and 4qB doesn't
- FSHD2: 5% - loss of epiG repression (hypomethylation) 80% cases mutation in *SMCHD1*
- Expression of the active DUX4 protein influences the activity of other genes in muscle cells
  - Exact function is unknow

FSHD Case



no. of repeats in array:	FSHD positive	FSHD negative
Allele 4qa	4	26
Allele 4qb	48	48

Sharim H, et al., Genome Res. 2019 Apr;29(4):646-656.

Part 3:

De novo genome assembly

De Novo Genome Assembly

Bionano

> 150 kbp

+

Oxford Nanopore

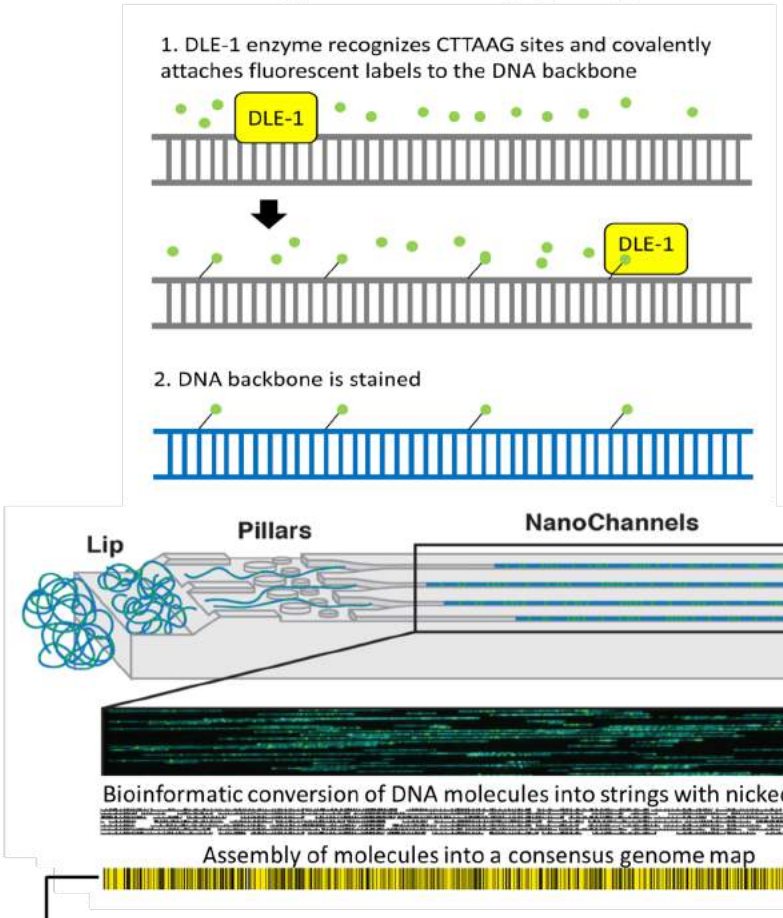
> 1 kbp

or

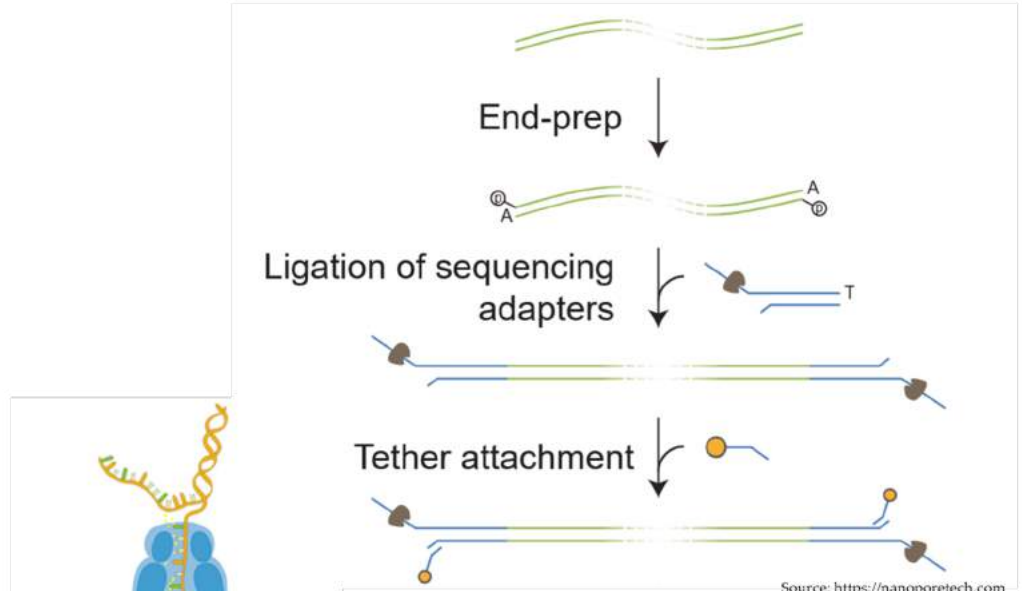
PacBio

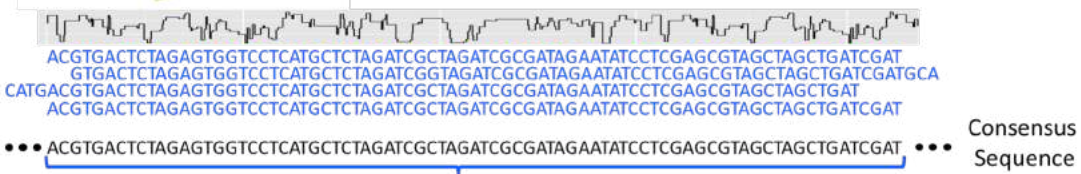
>1 kbp

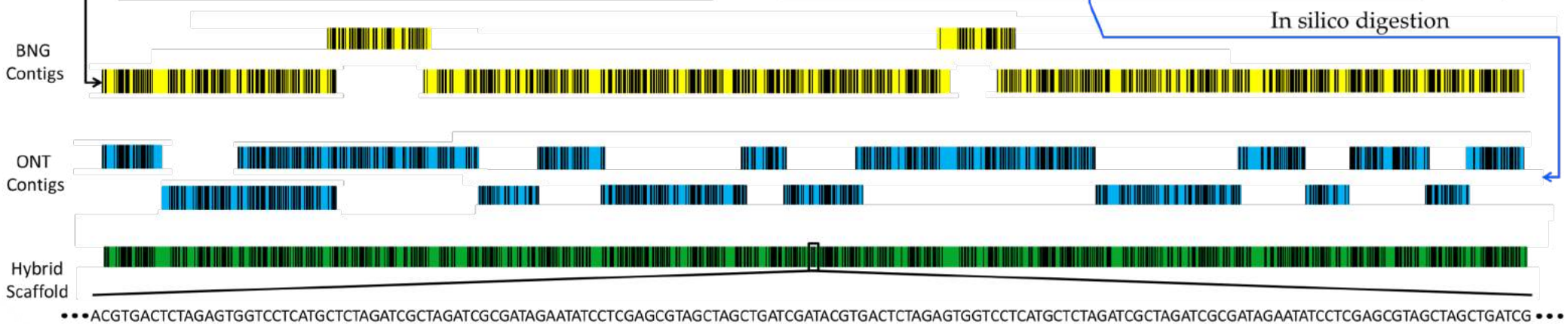
Optical Mapping



Long-Read Sequencing







Thank you!

Children's Bioinformatics Unit
Eric Vilain, MD, PhD
Lisa Guay-Woodford, MD
Hiroki Morizono, PhD
Susan Knoblach, PhD
Seth Berger, MD, PhD
Wei Li, PhD
Kazue Hashimoto-Torii, PhD
Michael Keller, PhD
Marius George Linguraru, PhD
Payal Banerjee, MS
Surajit Bhattacharya, PhD
Hayk Barseghyan, PhD

Funding:
DC IDDRC DISTRICT OF COLUMBIA INTELLECTUAL AND DEVELOPMENTAL DISABILITIES RESEARCH CENTER
U54 HD090257