CRISPR/Cas9 Enrichment and Long-read WGS for Structural Variant Discovery

PacBio CoLab Session

October 20, 2017

PACBIO SMRT SEQUENCING

Sequel System

Long Reads
average 10 to 15 kb

High Consensus Accuracy
random errors produce QV50 consensus

Uniform, Unbiased Coverage
no GC% or sequence complexity bias

Epigenetic Characterization
simultaneous detection of DNA methylation

APPLICATIONS OF SMRT SEQUENCING

Sequel System

***De novo* genome assembly**

Full isoform sequencing

Epigenetic characterization

Minor variant discovery

Structural variant discovery

Targeted sequencing

APPLICATIONS OF SMRT SEQUENCING

Sequel System

***De novo* genome assembly**

Full isoform sequencing

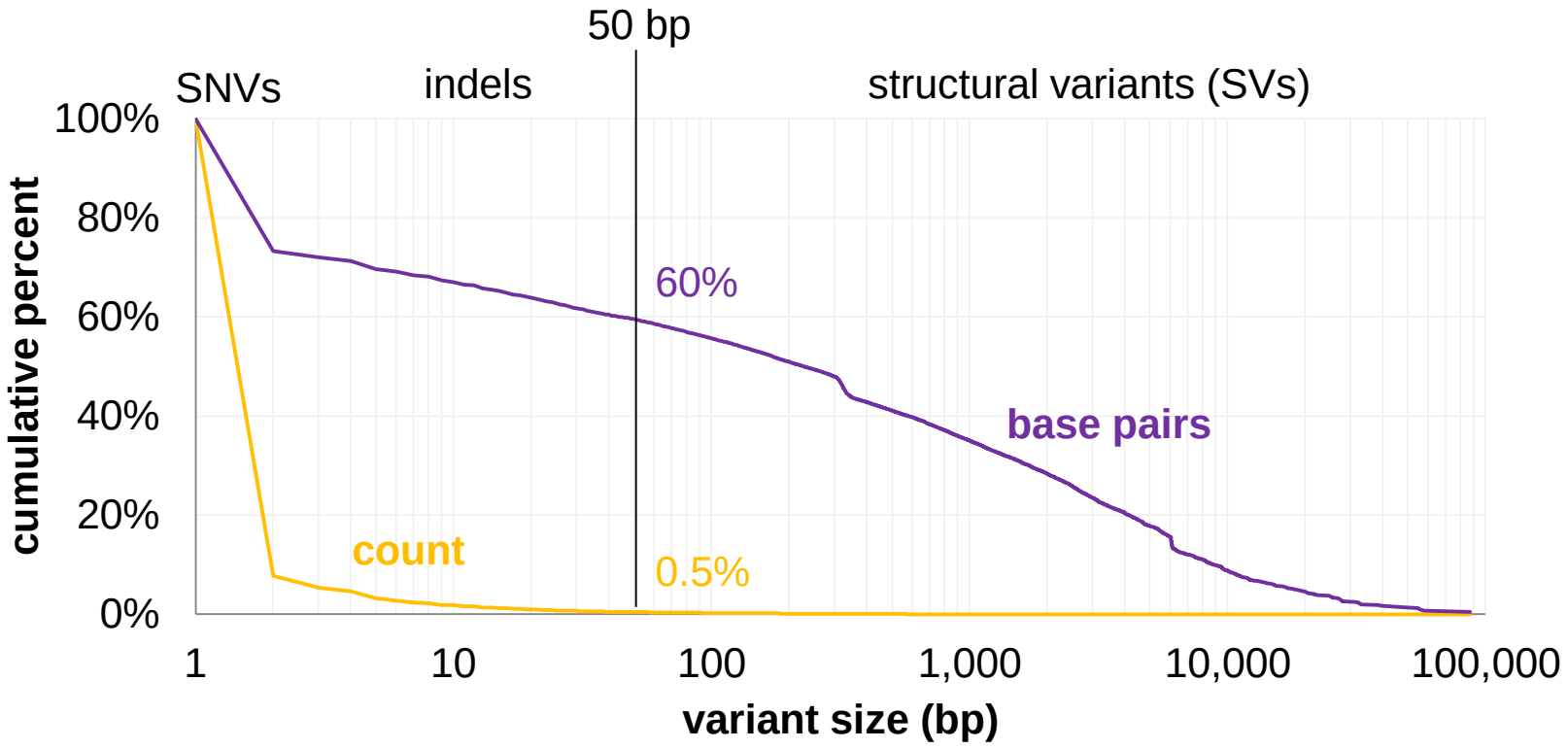
Epigenetic characterization

Minor variant discovery

Structural variant discovery

Targeted sequencing

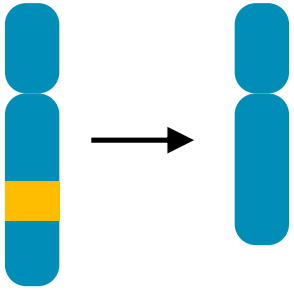
VARIATION IN A HUMAN GENOME – HG00733



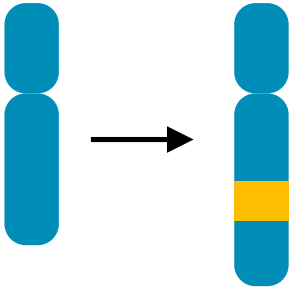
Chaisson et al. (2017) *bioRxiv*. doi:10.1101/193144.

TYPES OF STRUCTURAL VARIATION

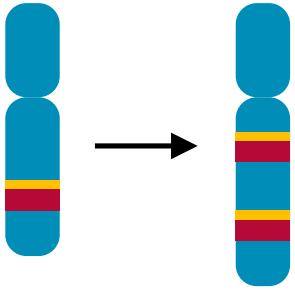
deletion



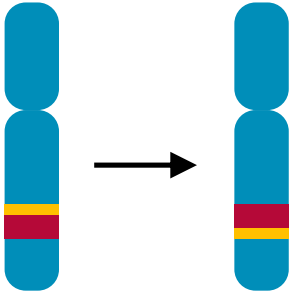
insertion



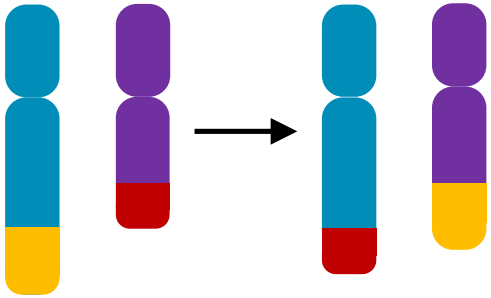
duplication



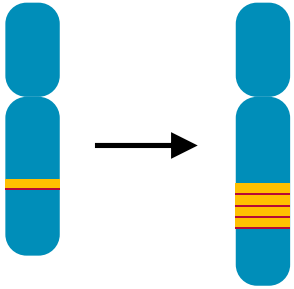
inversion



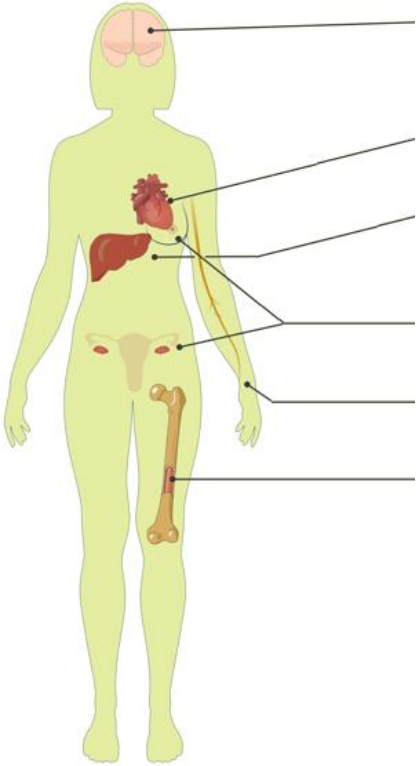
translocation



repeat expansion



STRUCTURAL VARIANTS AND DISEASE



Schizophrenia

Carney complex

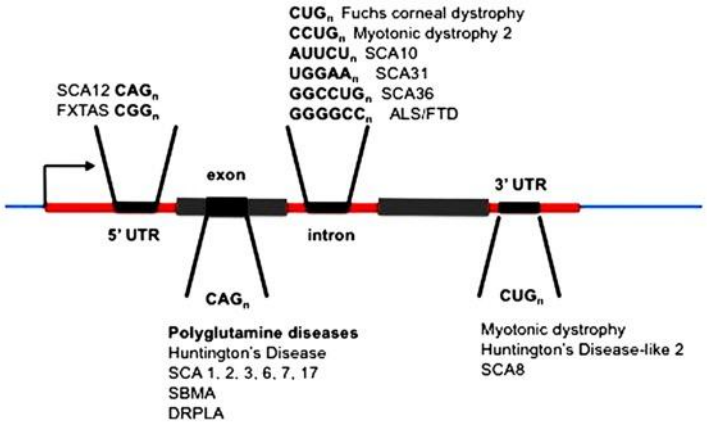
Poor drug metabolism

Breast & ovarian cancer

Neurofibromatosis

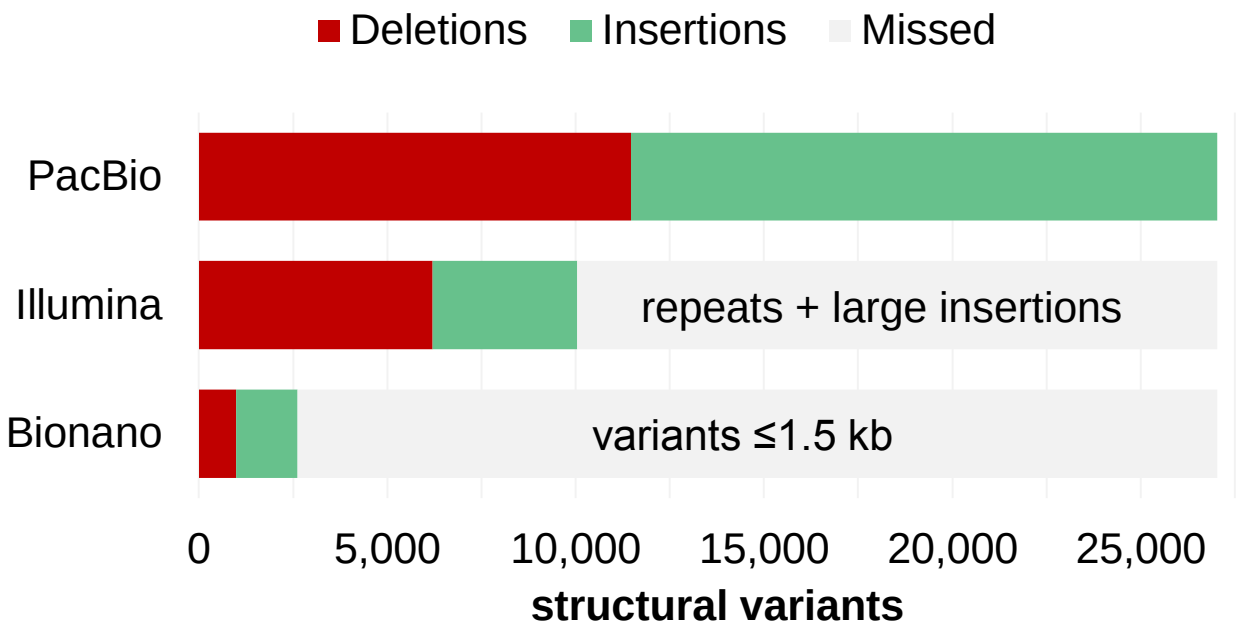
Chronic myeloid leukemia

repeat expansion disorders



http://www.pacb.com/wp-content/uploads/Structural-Variation-Infographic.pdf
Richards et al. (2013) *Front Mol Neurosci*. doi:10.3389/fnmol.2013.00025.

TECHNOLOGY TO DETECT STRUCTURAL VARIANTS



Chaisson et al. (2017) *bioRxiv*. doi:10.1101/193144.

“A move forward to **full-spectrum SV detection** ... will increase the diagnostic yield in patients with genetic disease, SV-mediated mutation, and repeat expansions.”

Chaisson et al. (2017) *bioRxiv*. doi:10.1101/193144.

PacBio Long-Read WGS for Structural Variant Discovery

Targeted Enrichment without Amplification and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions

PacBio Long-Read WGS for Structural Variant Discovery

Targeted Enrichment without Amplification and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions

FOR MORE INFORMATION – PACB.COM/SV

PACBIO®

PRODUCTS + SERVICES | RESEARCH FOCUS | APPLICATIONS | SMRT SCIENCE | SUPPORT | COMPANY

APPLICATIONS

STRUCTURAL VARIATION

WHOLE GENOME SEQUENCING

- Human Whole Genome Sequencing
- Plant and Animal Whole Genome Sequencing
- Microbial Whole Genome Sequencing
- Structural Variation

TARGETED SEQUENCING

CALLING ALL VARIANT TYPES

Structural variation accounts for most of the base pairs that differ between two human genomes, and causes many genetic disorders. The ability to study structural variants, in addition to smaller single nucleotide variants and indels, is critical to understanding how genetic variation impacts health and disease in the era of Precision Medicine.

WGS FOR STRUCTURAL VARIANT DISCOVERY

	Structural Variants	Small Variants
Sequencing	PacBio Sequel System	Short-read NGS
Read Mapping	NGMLR	BWA
Variant Calling	pbsv	GATK
Visualization	IGV 2.4	IGV

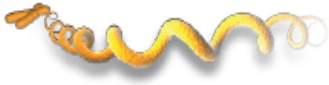
WGS FOR STRUCTURAL VARIANT DISCOVERY

	Structural Variants	Small Variants
Sequencing	PacBio Sequel System	Short-read NGS
Read Mapping	NGMLR	BWA
Variant Calling	pbsv	GATK
Visualization	IGV 2.4	IGV

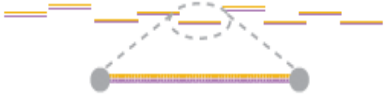
SEQUENCING

Library Preparation

5 µg DNA



20 kb shear
+ damage repair



SMRTbell adapter ligation

15 kb size selection

SEQUENCING

Sequencing

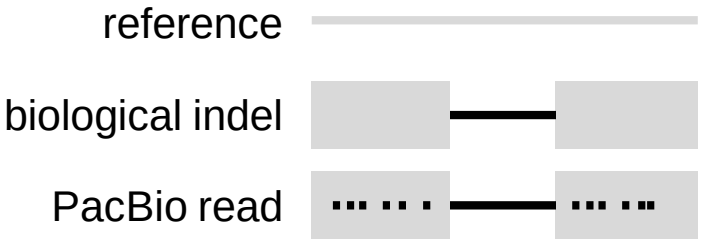
polymerase binding

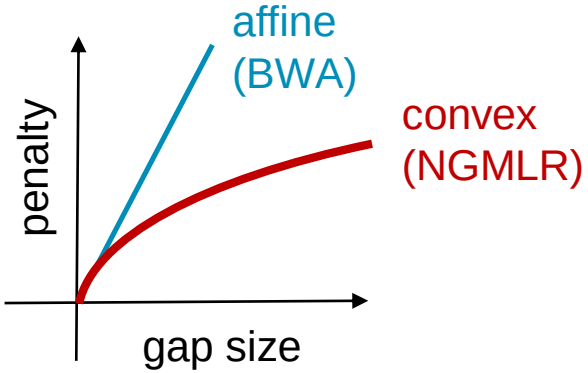
Sequel System
(5 Gb per SMRT Cell)

WGS FOR STRUCTURAL VARIANT DISCOVERY

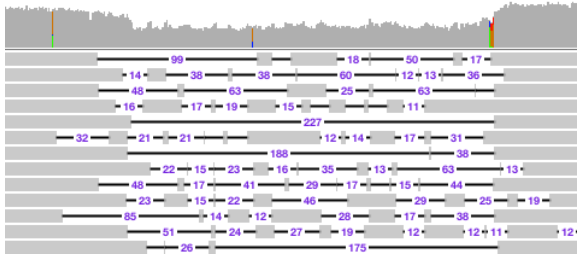
	Structural Variants	Small Variants
Sequencing	PacBio Sequel System	Short-read NGS
Read Mapping	NGMLR	BWA
Variant Calling	pbsv	GATK
Visualization	IGV 2.4	IGV

READ MAPPING

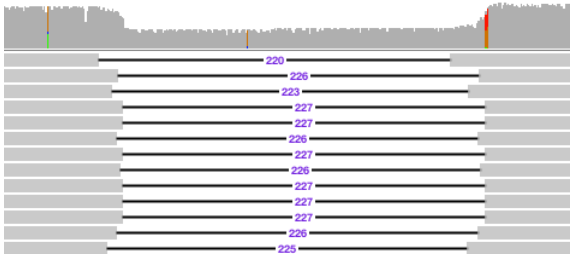




BWA



NGMLR



Sedlazeck et al. (2017) *bioRxiv*. doi:10.1101/169557.

WGS FOR STRUCTURAL VARIANT DISCOVERY

	Structural Variants	Small Variants
Sequencing	PacBio Sequel System	Short-read NGS
Read Mapping	NGMLR	BWA
Variant Calling	pbsv	GATK
Visualization	IGV 2.4	IGV

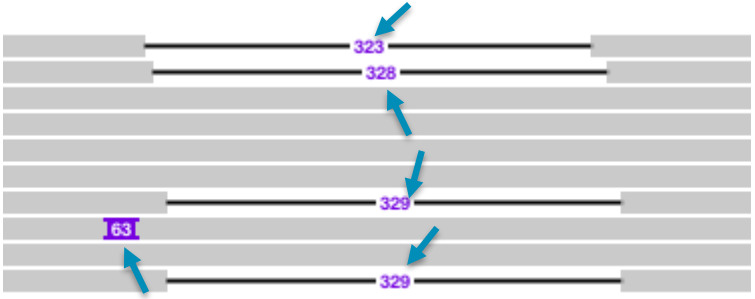
VARIANT CALLING

FIND SV SIGNATURES	CIGAR D & I ≥50 bp
CLUSTER SV SIGNATURES	nearby with similar sequence
FILTER	≥2 and ≥20% reads support
SUMMARIZE INTO SV	consensus of supporting reads
GENOTYPE	supporting reads / covering reads

http://pacb.com/sv

VARIANT CALLING

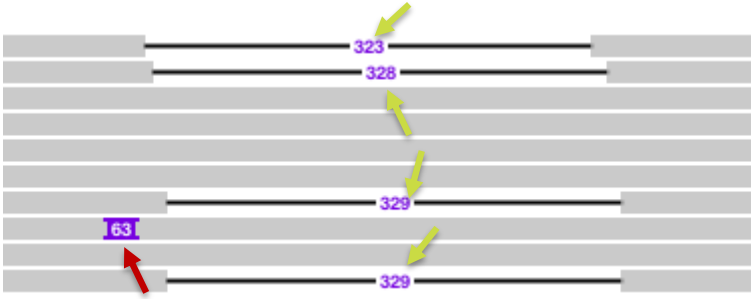
FIND SV SIGNATURES	CIGAR D & I ≥50 bp
CLUSTER SV SIGNATURES	nearby with similar sequence
FILTER	≥2 and ≥20% reads support
SUMMARIZE INTO SV	consensus of supporting reads
GENOTYPE	supporting reads / covering reads



http://pacb.com/sv

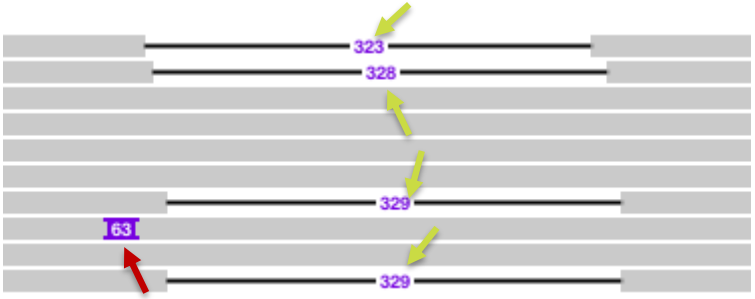
VARIANT CALLING

FIND SV SIGNATURES	CIGAR D & I ≥50 bp
CLUSTER SV SIGNATURES	nearby with similar sequence
FILTER	≥2 and ≥20% reads support
SUMMARIZE INTO SV	consensus of supporting reads
GENOTYPE	supporting reads / covering reads



VARIANT CALLING

FIND SV SIGNATURES	CIGAR D & I ≥50 bp
CLUSTER SV SIGNATURES	nearby with similar sequence
FILTER	**≥2 and ≥20% reads support**
SUMMARIZE INTO SV	consensus of supporting reads
GENOTYPE	supporting reads / covering reads



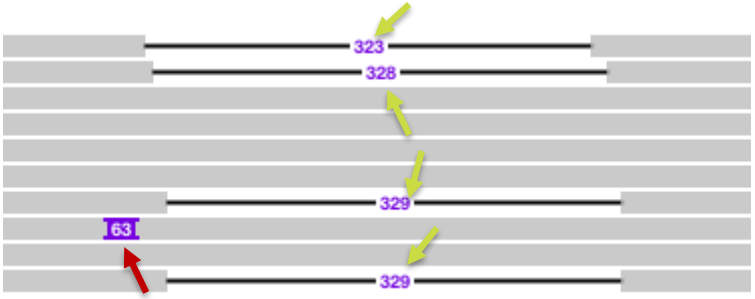
1 of 10

4 of 10

http://pacb.com/sv

VARIANT CALLING

FIND SV SIGNATURES	CIGAR D & I ≥50 bp
CLUSTER SV SIGNATURES	nearby with similar sequence
FILTER	≥2 and ≥20% reads support
SUMMARIZE INTO SV	consensus of supporting reads
GENOTYPE	supporting reads / covering reads



1 of 10

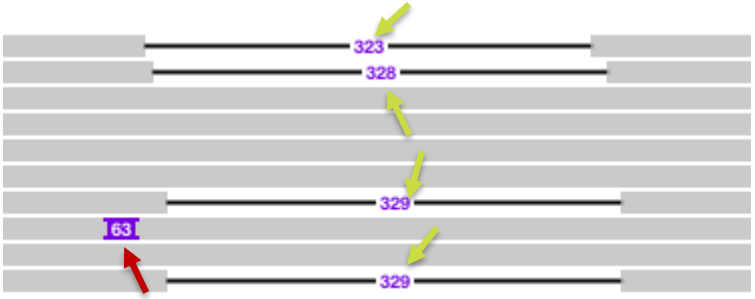
4 of 10

329 bp deletion

http://pacb.com/sv

VARIANT CALLING

FIND SV SIGNATURES	CIGAR D & I ≥50 bp
CLUSTER SV SIGNATURES	nearby with similar sequence
FILTER	≥2 and ≥20% reads support
SUMMARIZE INTO SV	consensus of supporting reads
GENOTYPE	supporting reads / covering reads



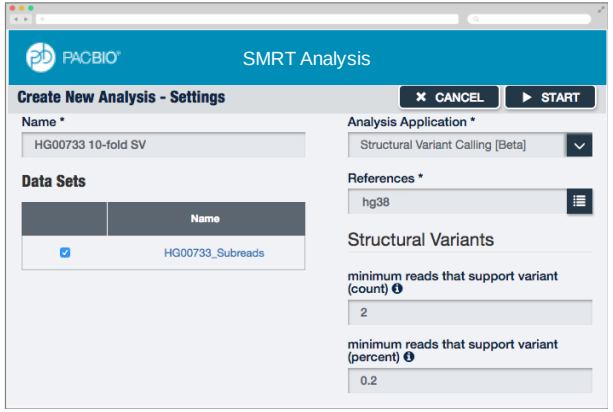
1 of 10

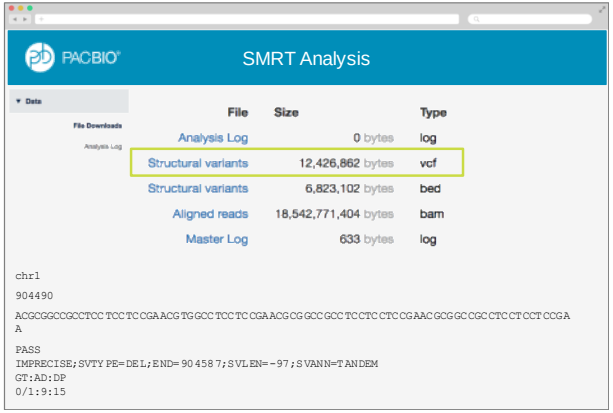
4 of 10

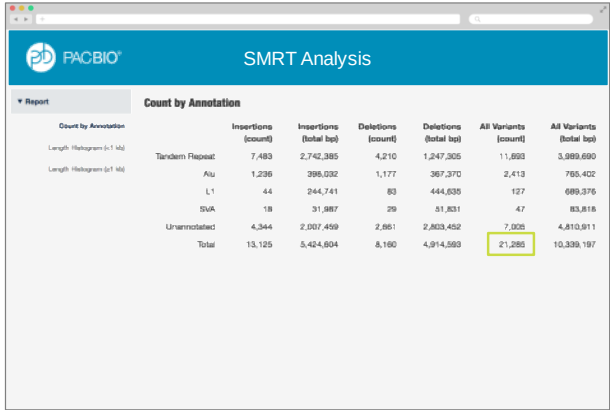
329 bp deletion

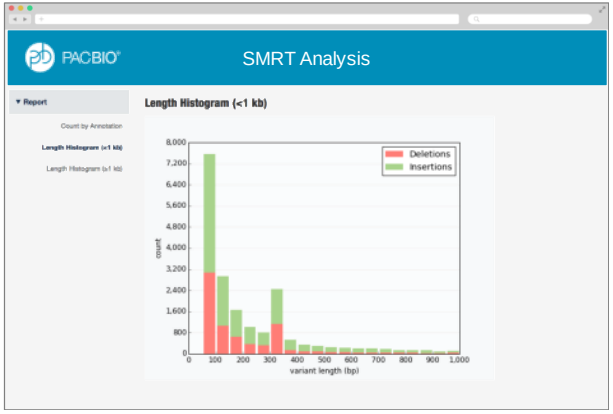
heterozygous (4 of 10)

VARIANT CALLING





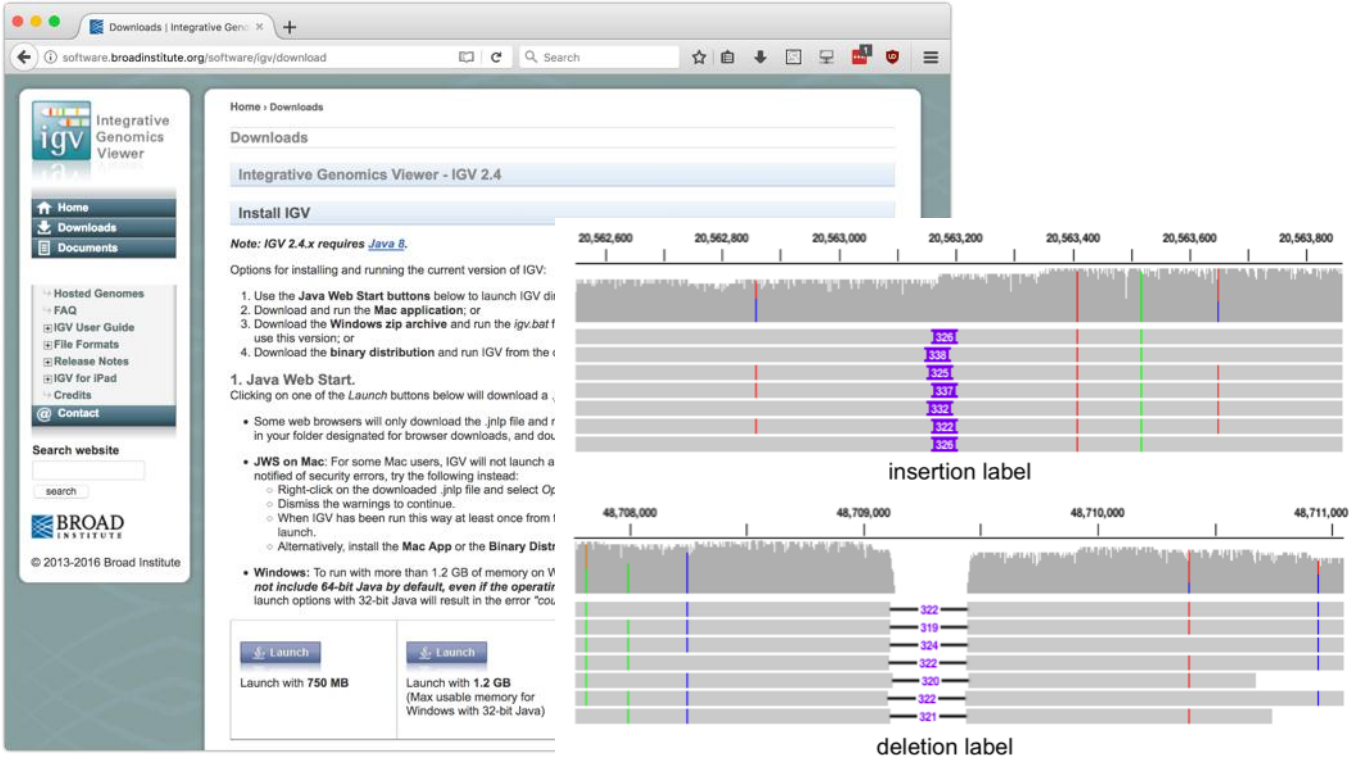




WGS FOR STRUCTURAL VARIANT DISCOVERY

	Structural Variants	Small Variants
Sequencing	PacBio Sequel System	Short-read NGS
Read Mapping	NGMLR	BWA
Variant Calling	pbsv	GATK
Visualization	IGV 2.4	IGV

VISUALIZATION



insertion

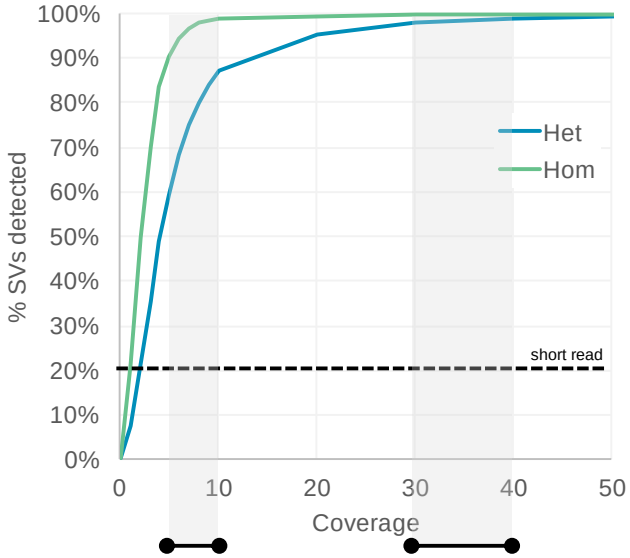
deletion

Robinson et al. (2011) *Nature Biotechnology*. doi:10.1038/nbt.1754.

WGS FOR STRUCTURAL VARIANT DISCOVERY

	Structural Variants	Small Variants
Sequencing	PacBio Sequel System	Short-read NGS
Read Mapping	NGMLR	BWA
Variant Calling	pbsv	GATK
Visualization	IGV 2.4	IGV

HOW MUCH TO SEQUENCE?



Human HG00733
Sequel System
211 Gb (70-fold)

5- to 10-fold
optimal tradeoff of
cost vs. performance

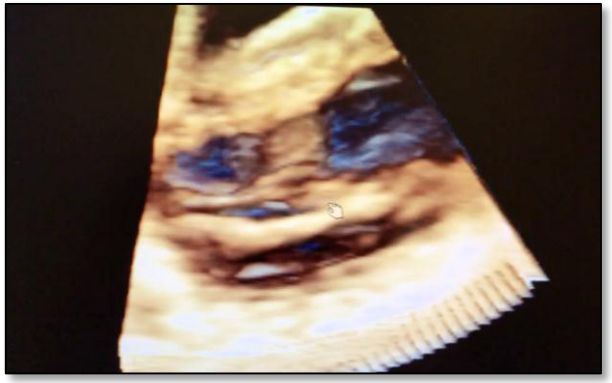
disease gene discovery;
population characterization

30- to 40-fold
saturate discovery

de novo variant discovery

CLINICAL CASE HISTORY

7 yrs	left atrial myxoma resection, atrial repair
10 yrs	testicular mass, right orchiectomy
13 yrs	pituitary tumor
16 yrs	recurrence of myxomata, resection, adrenal microadenoma
18 yrs	recurrence of ventricular myxomata, resection, VT
19 yrs	ACTH-independent Cushing's disease, thyroid nodules
21 yrs	transphenoidal resection of pituitary
present (26 yrs)	recurrence of myxomata, consideration for heart transplant



→ genetics suggests Carney complex

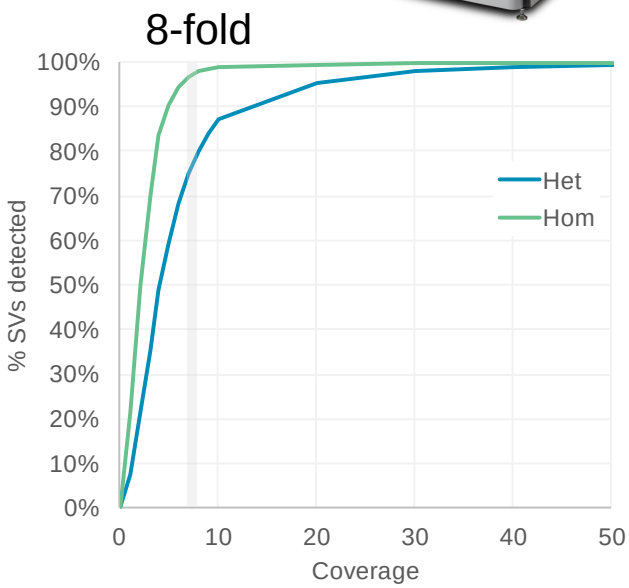
PRKAR1A testing negative

→ short-read whole genome sequencing negative

Merker et al. (2017) *Genetics in Medicine*. doi:10.1038/gim.2017.86.

EVALUATING STRUCTURAL VARIANTS

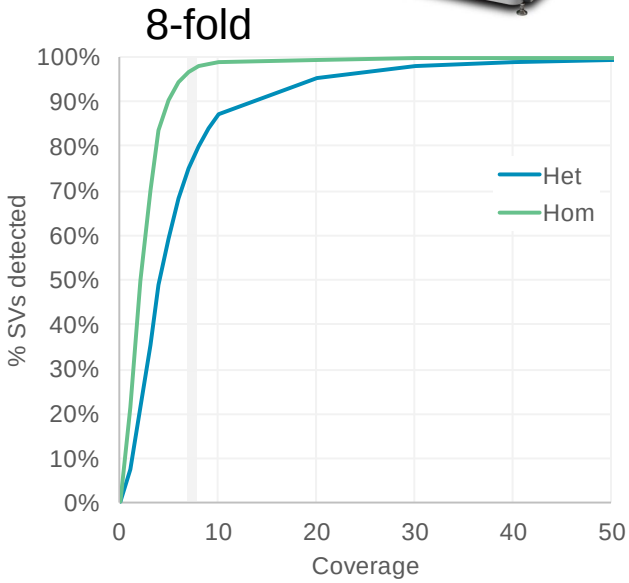
	Deletions	Insertions
Initial call set	6,971	6,821



Merker et al. (2017) *Genetics in Medicine*. doi:10.1038/gim.2017.86.

EVALUATING STRUCTURAL VARIANTS

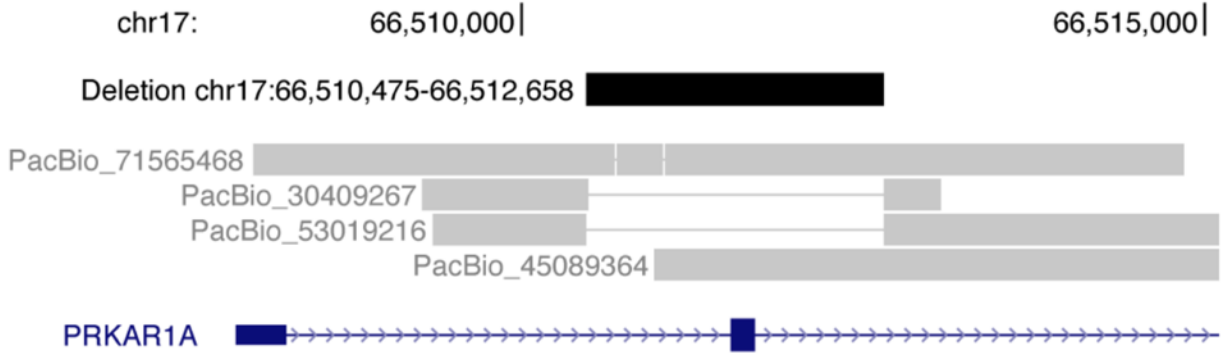
	Deletions	Insertions
Initial call set	6,971	6,821
Not in segdup	5,893	6,254
Not in NA12878 “healthy” control	2,476	3,171
Overlaps RefSeq coding exon	39	16
Gene linked to some disease in OMIM	3	3



Merker et al. (2017) *Genetics in Medicine*. doi:10.1038/gim.2017.86.

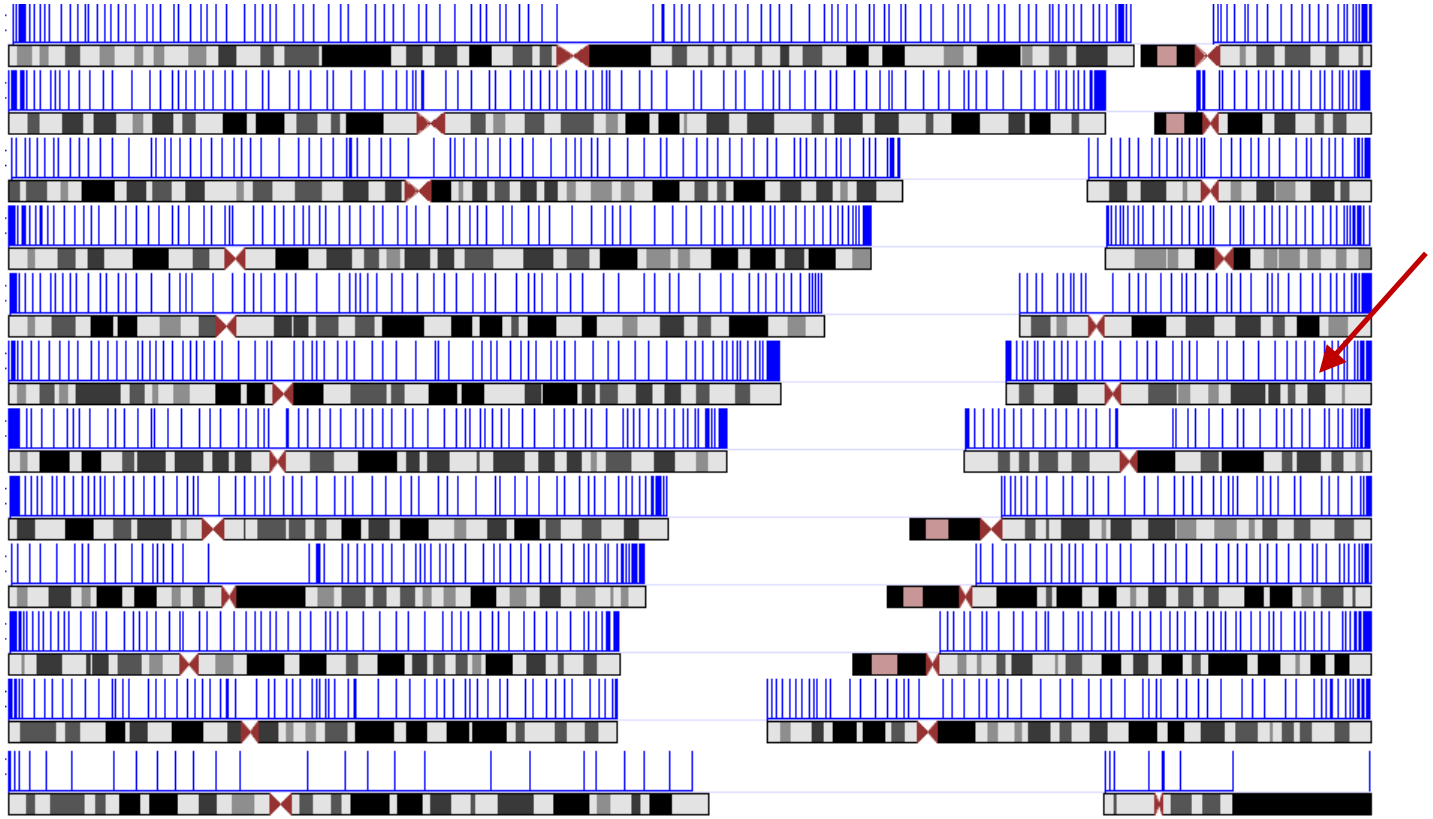
HETEROZYGOUS 2.2 KB DELETION IN *PRKAR1A*

PacBio discovery



Sanger confirmation

Merker et al. (2017) *Genetics in Medicine*. doi:10.1038/gim.2017.86.



PacBio Long-Read WGS for Structural Variant Discovery

Targeted Enrichment without Amplification and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions

REPEAT EXPANSION DISEASES

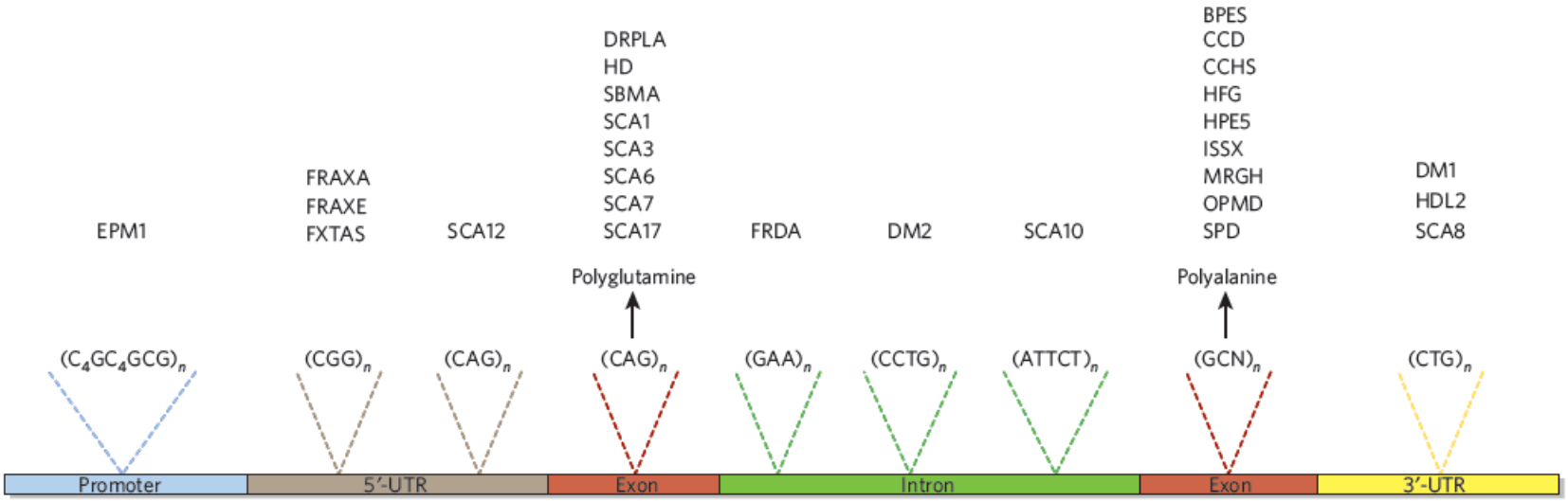
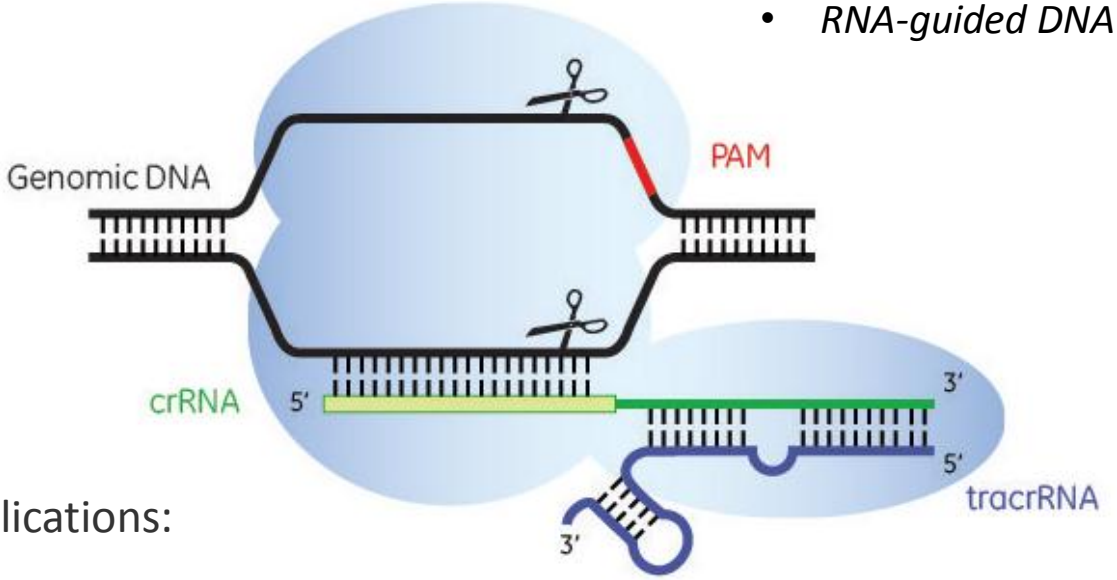


Figure 1 | Location of expandable repeats responsible for human diseases. The sequence and location within a generic gene of expandable repeats that cause human diseases are shown, and the associated diseases are listed. BPES, blepharophimosis, ptosis and epicanthus inversus; CCD, cleidocranial dysplasia; CCHS, congenital central hypoventilation syndrome; DM, myotonic dystrophy; DRPLA, dentatorubral–pallidoluysian atrophy; EPM1, progressive myoclonic epilepsy 1; FRAXA, fragile X syndrome; FRAXE, fragile X mental retardation associated with *FRAXE* site; FRDA, Friedreich's ataxia; FXTAS, fragile X tremor and ataxia syndrome; HD, Huntington's disease; HDL2, Huntington's-disease-like 2; HFG, hand–foot–genital syndrome; HPE5, holoprosencephaly 5; ISSX, X-linked infantile spasm syndrome; MRGH, mental retardation with isolated growth hormone deficiency; OPMD, oculopharyngeal muscular dystrophy; SBMA, spinal and bulbar muscular atrophy; SCA, spinocerebellar ataxia; SPD, synpolydactyly.

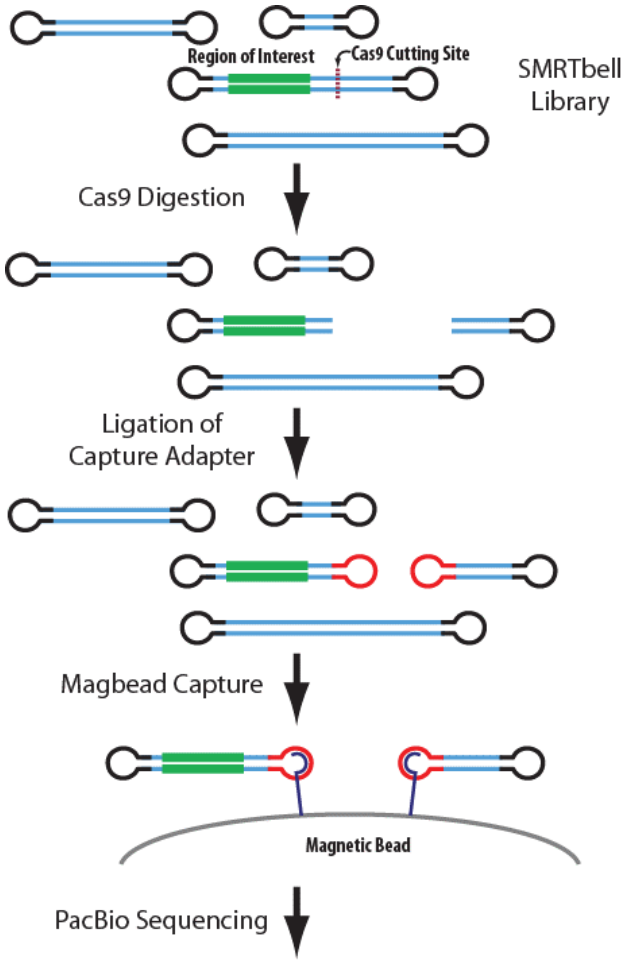
CRISPR/CAS9 SYSTEM

- *Bacterial Adaptive Immunity*
- *RNA-guided DNA Endonuclease*

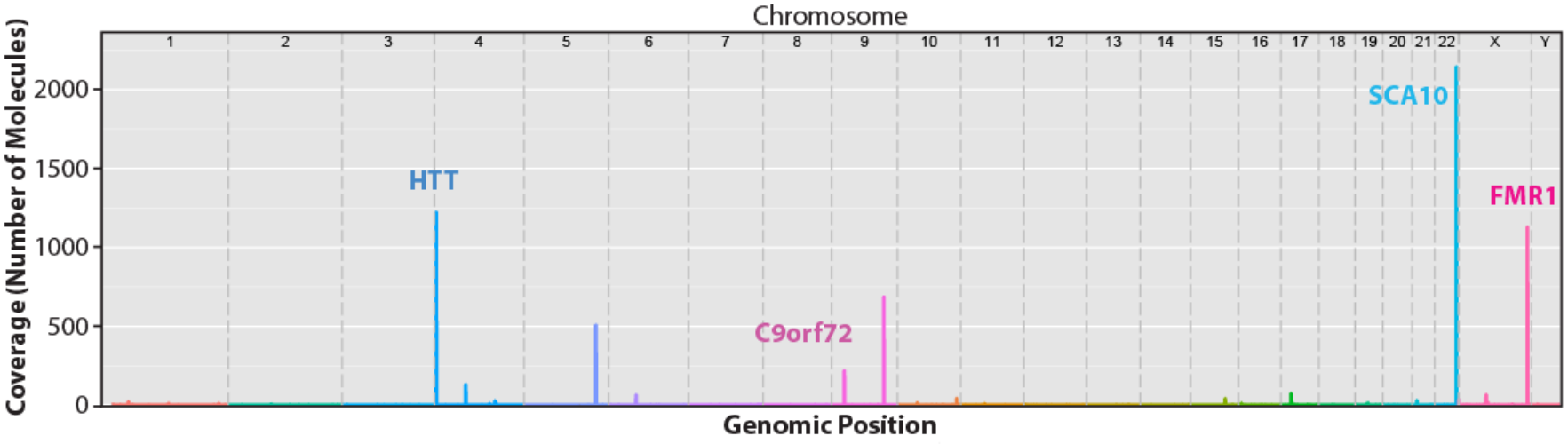


Some *in vivo* applications:
- Gene silencing
- Homology-directed repair
- Transient gene silencing or transcriptional repression
- Transient activation of endogenous genes
- Transgenic animals and embryonic stem cells

PCR-FREE TARGET ENRICHMENT VIA CAS9



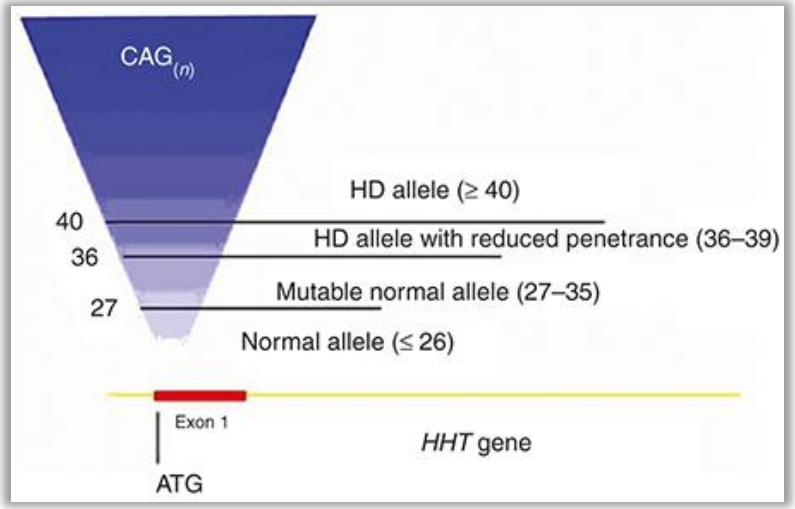
COVERAGE ACROSS THE GENOME

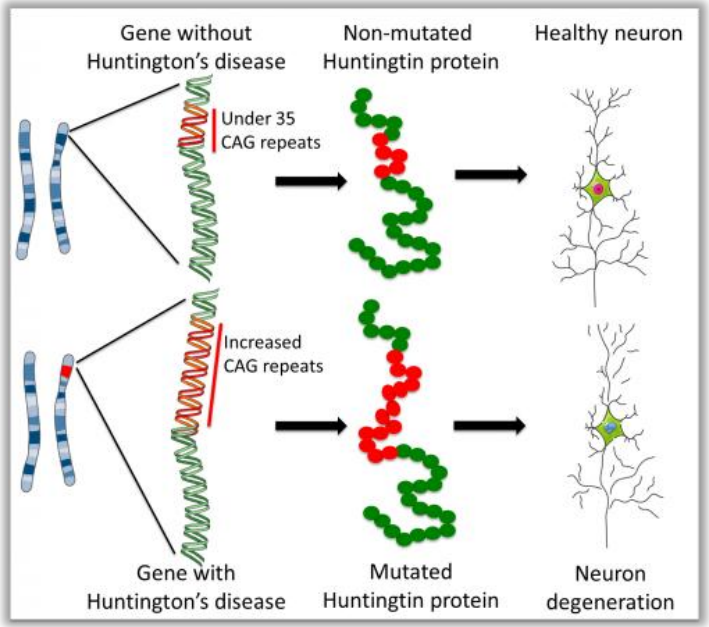


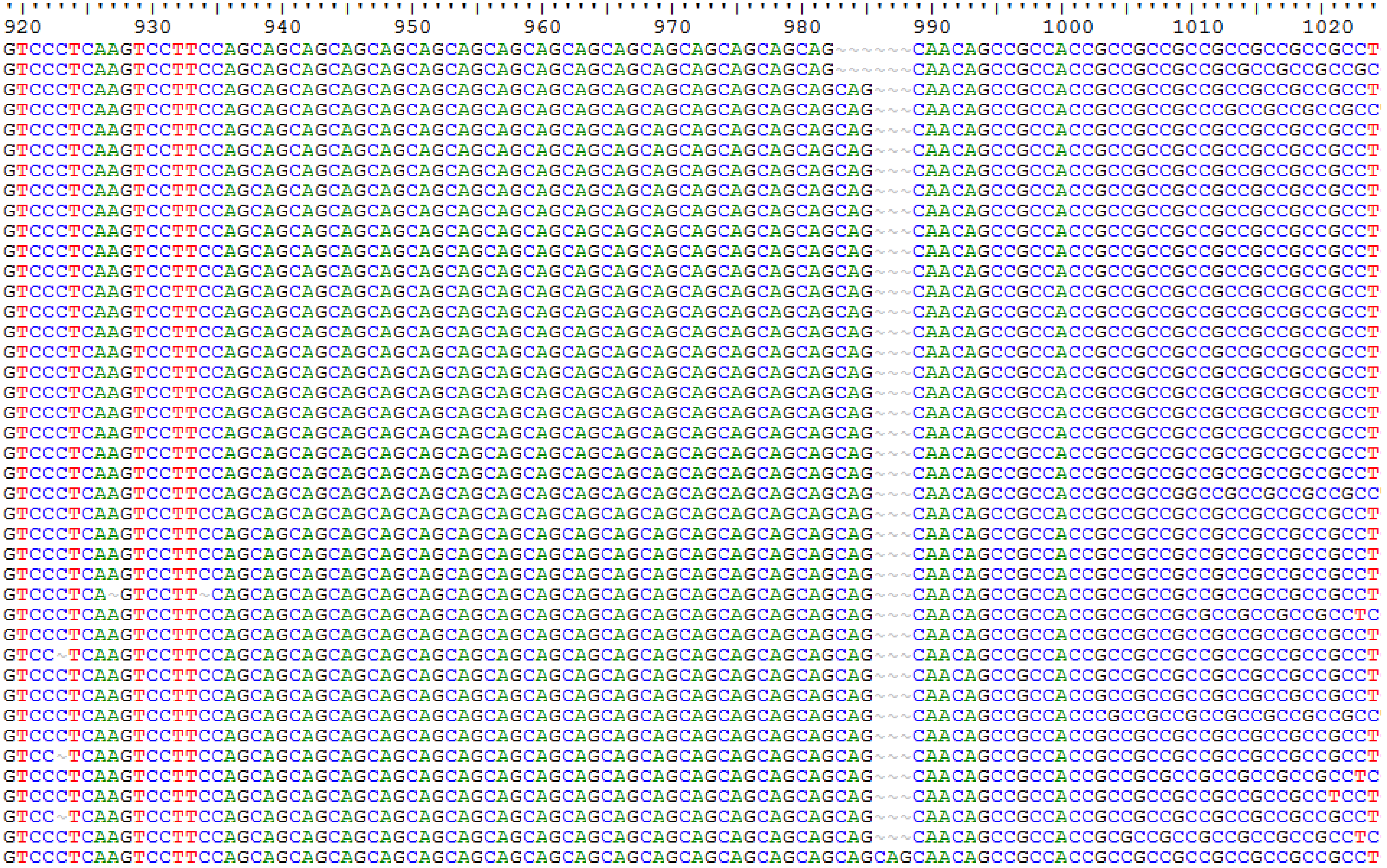
1 SMRT Cell (PacBio RS II)

HUNTINGTON'S DISEASE (HD)

- Autosomal dominant neurodegenerative genetic disorder
- Caused by an expansion of a CAG triplet repeat stretch in the Huntingtin (HTT) gene
  - polyglutamine tract

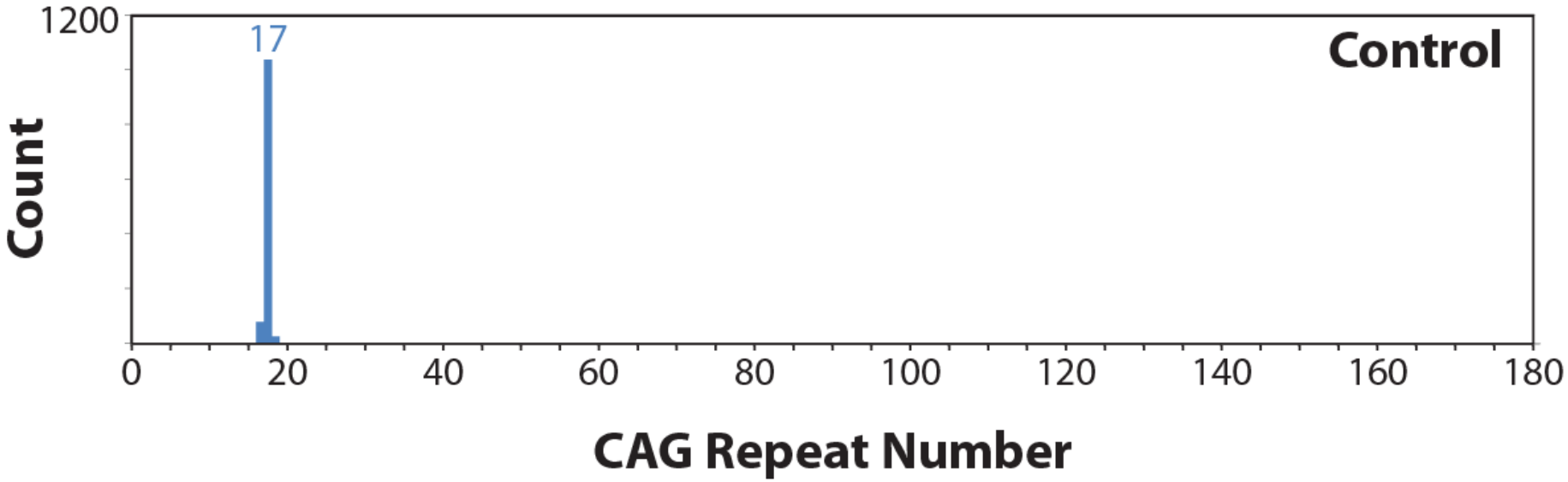




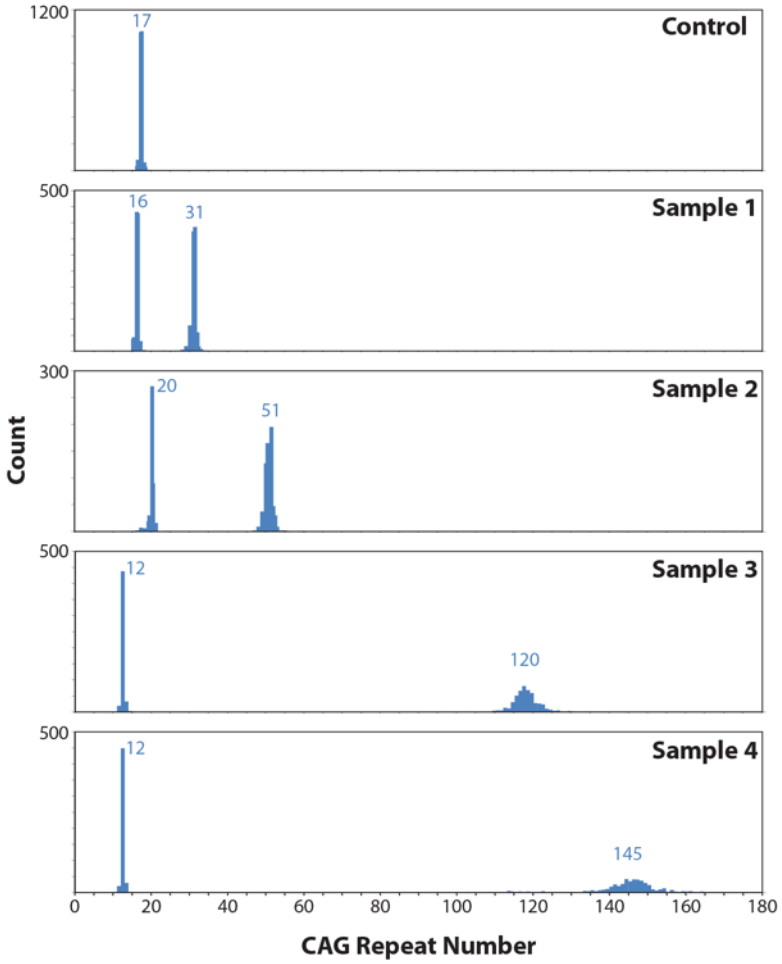


```
920       930       940       950       960       970       980       990       1000      1010      1020
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~~~~CAACAGCCGCCACCGCCGCCGCCGCGCCGCCGCCGC
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCCGCCGCCGCCGCC
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGGCCGCCGCCGCCGCC
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCA~GTCCTT~CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCGCCGCCGCCGCCTC
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCC~TCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCCGCCGCCGCCGCCGCCGCCGCC
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCC~TCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCGCCGCCGCCGCCGCCTC
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCTCCT
GTCC~TCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG~~~CAACAGCCGCCACCGCGCCGCCGCCGCCGCCGCCTC
GTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCT
```

CAG REPEAT COUNTS

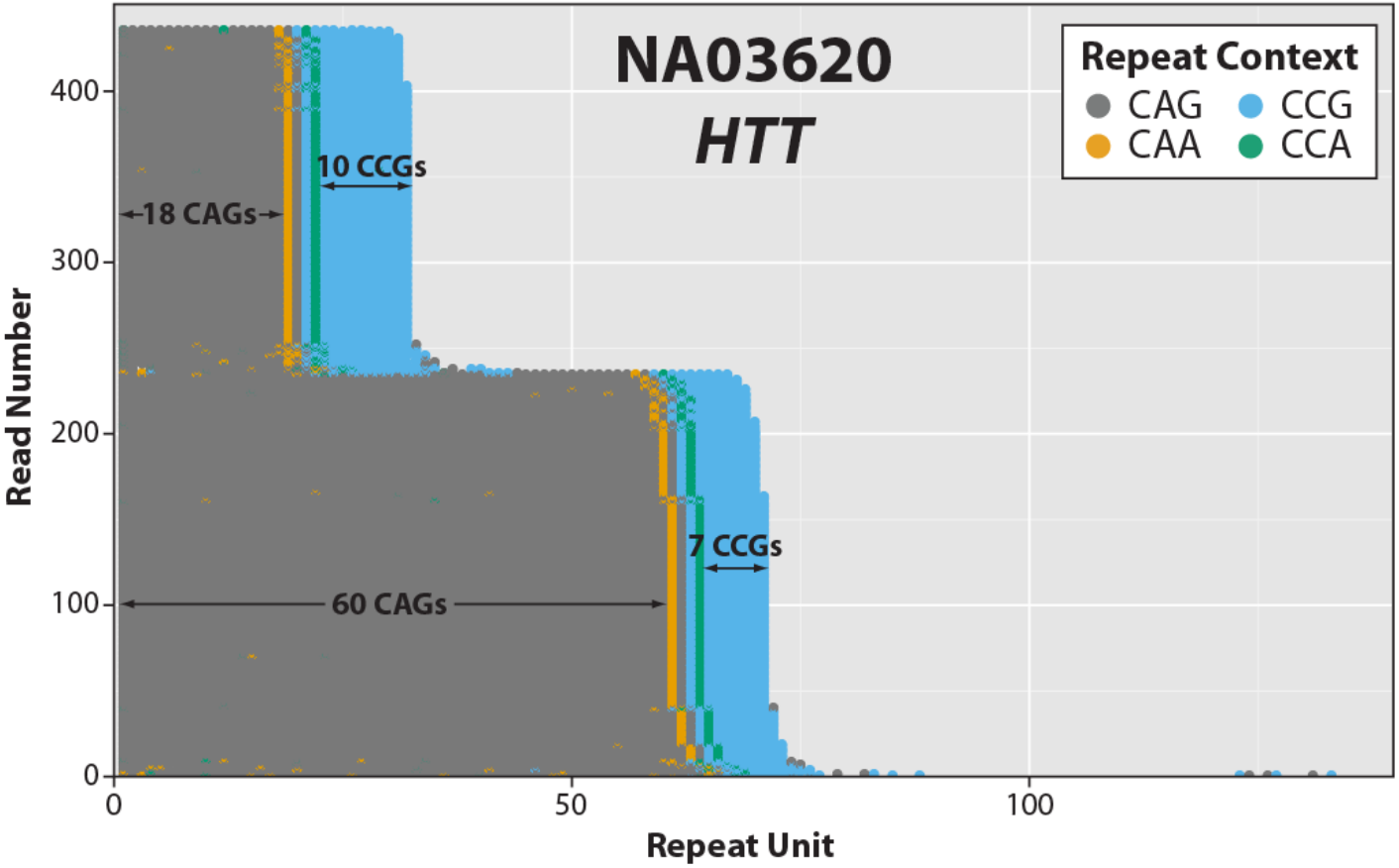


CAG REPEAT COUNTS IN HD PATIENTS



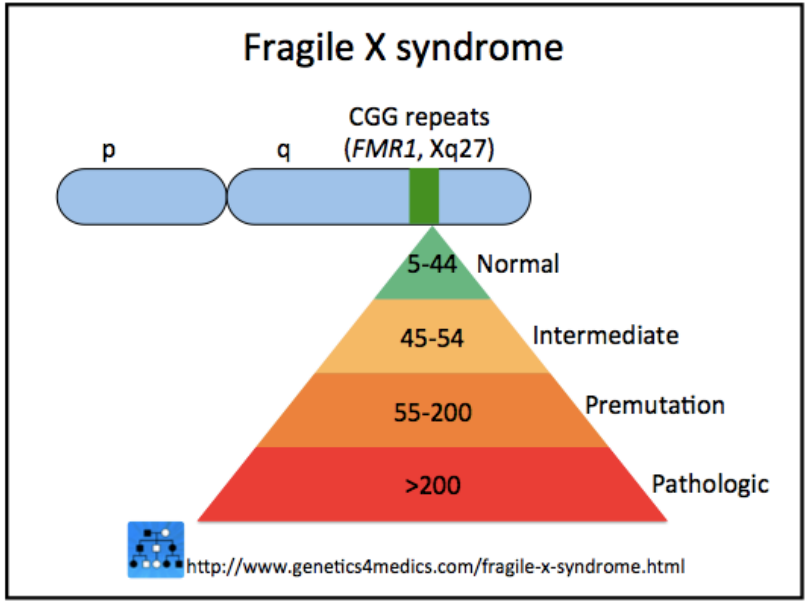
- Widening repeat number distribution at the mutated allele is biological
- Obtained roughly equal number of sequenced molecules for normal and mutated alleles

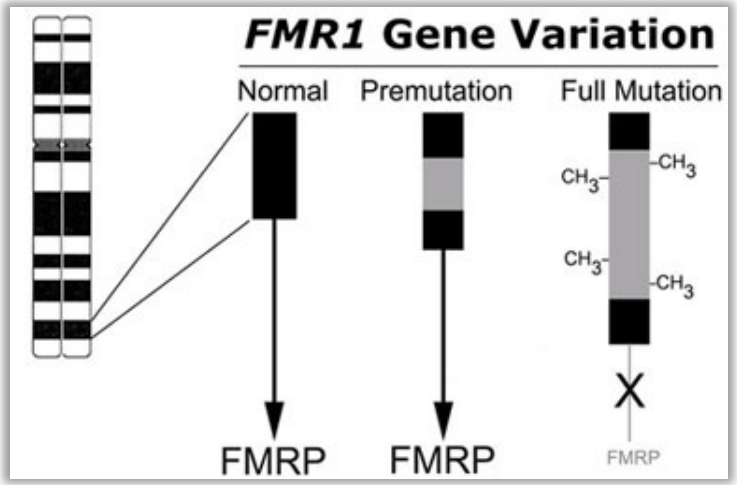
Samples obtained from Vanessa Wheeler (Harvard Medical School)


NA03620
HTT
Repeat Context
CAG
CCG
CAA
CCA
10 CCGs
18 CAGs
7 CCGs
60 CAGs
Read Number
Repeat Unit
0
100
200
300
400
0
50
100

FRAGILE X SYNDROME

- Most common heritable form of cognitive impairment
- Caused by expansion of a CGG trinucleotide repeat in the 5' UTR of the FMR1 gene



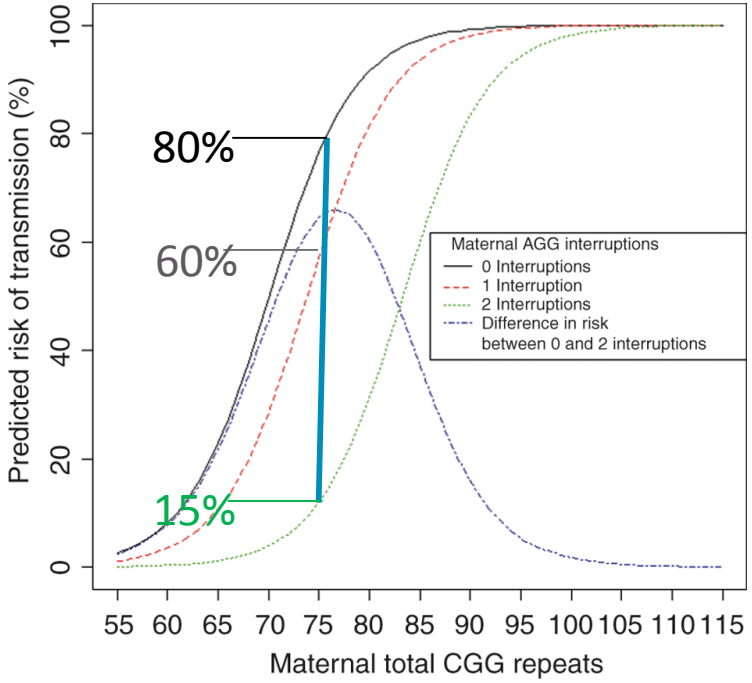


fraxa.org

AGG "INTERRUPTIONS" REDUCE THE CHANCES OF PRE- TO FULL MUTATION TRANSMISSION

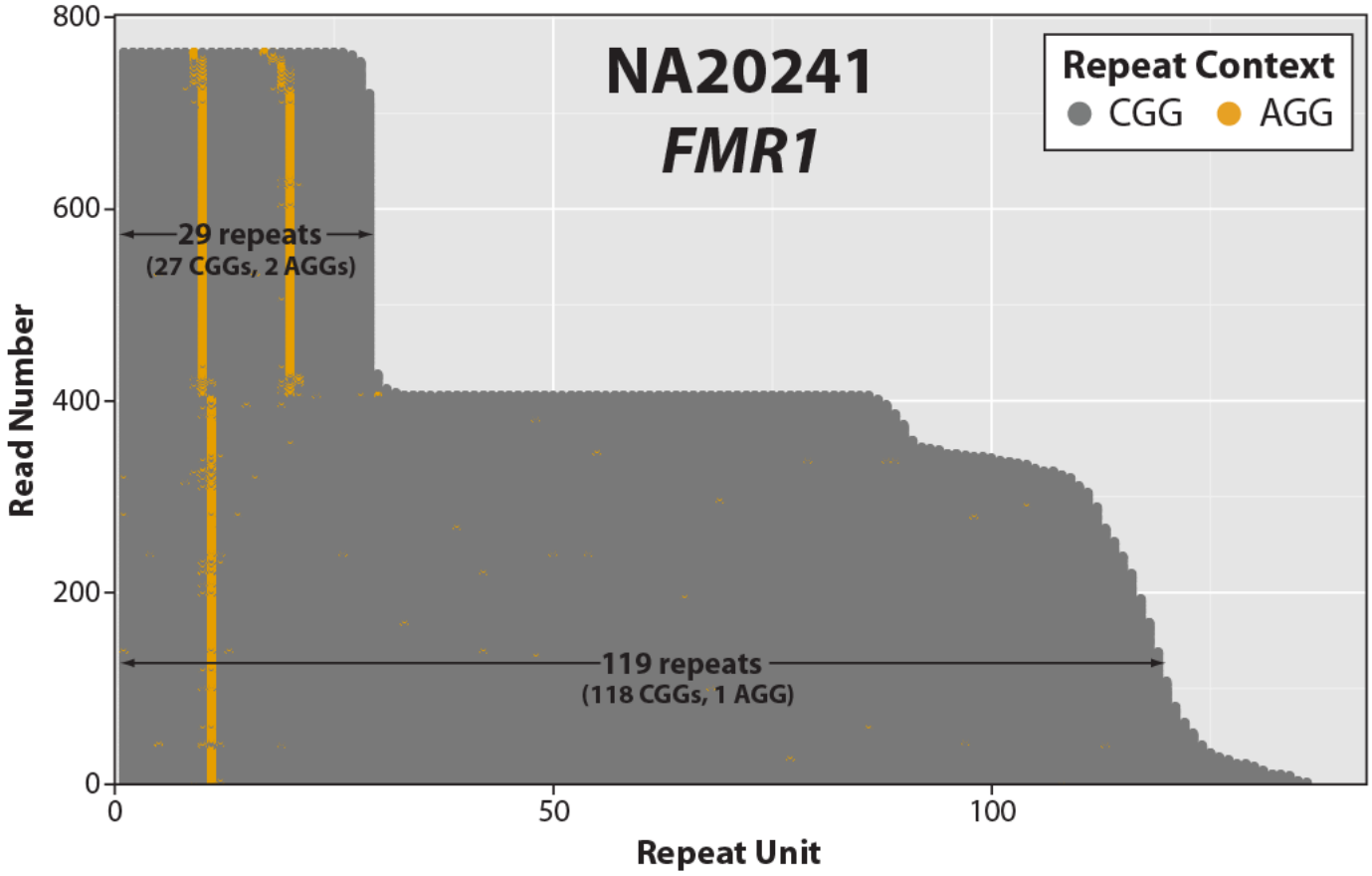
…CGG CGG CGG CGG *AGG* CGG…

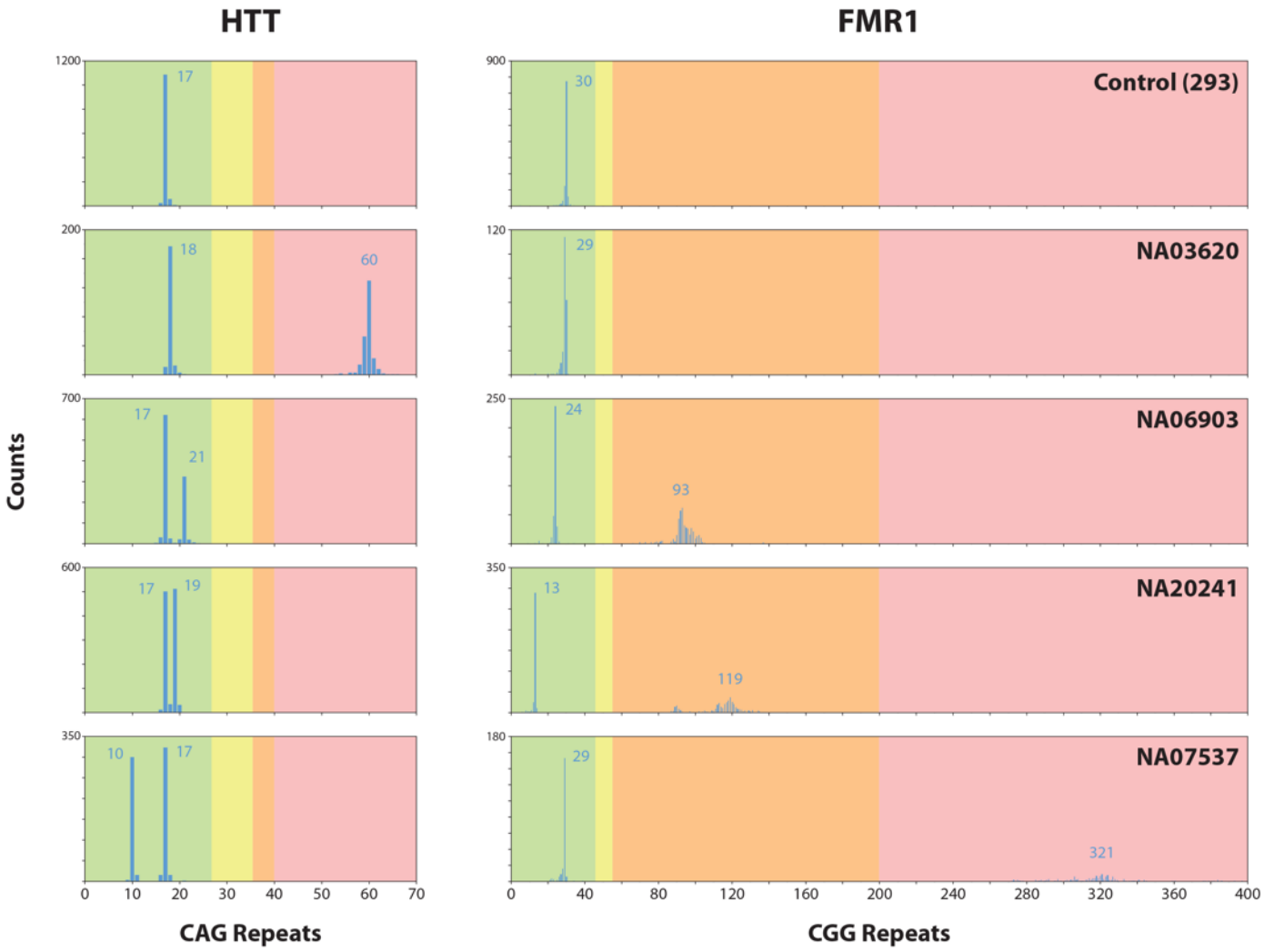
- Difference in risk is greatest near 75-80 CGG repeats
- Having full sequence information is medically relevant



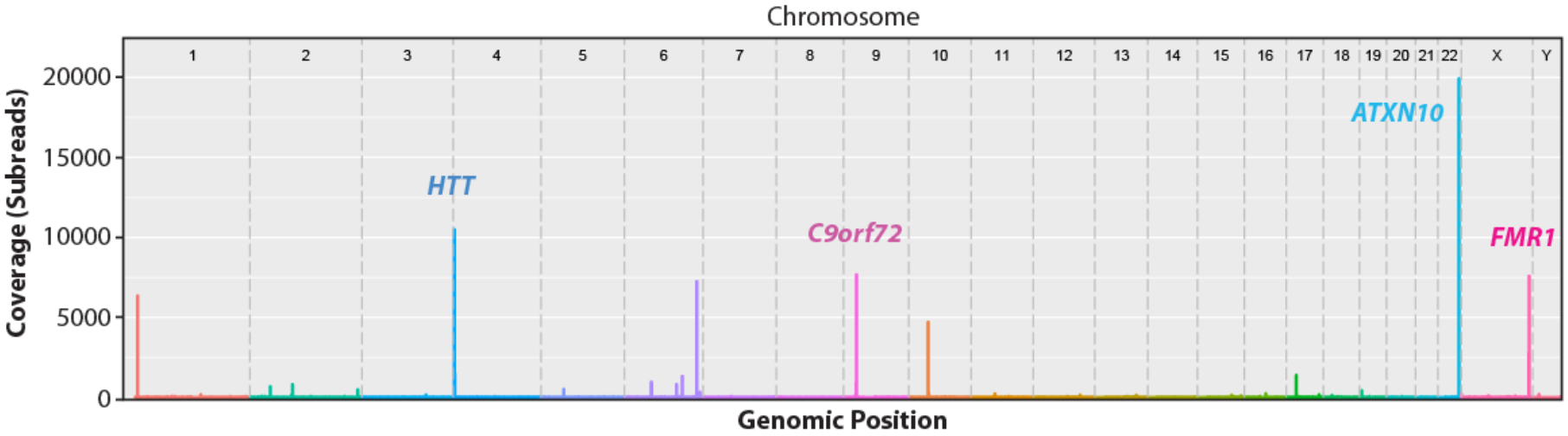
2 …CGG CGG CGG CGG *AGG* CGG CGG CGG CGG CGG CGG CGG CGG CGG *AGG* CGG …
1 …CGG CGG CGG CGG *AGG* CGG CGG CGG CGG CGG CGG CGG CGG CGG CGG CGG …
0 …CGG CGG CGG CGG CGG CGG CGG CGG CGG CGG CGG CGG CGG CGG CGG CGG …

Yrigollen et al. (2012) Genet Med 14:729–736


NA20241
FMR1
Repeat Context
CGG
AGG
29 repeats
(27 CGGs, 2 AGGs)
119 repeats
(118 CGGs, 1 AGG)
Read Number
Repeat Unit
0
200
400
600
800
0
50
100


HTT
FMR1
Counts
Control (293)
NA03620
NA06903
NA20241
NA07537
1200
17
900
30
200
18
60
120
29
700
17
21
250
24
93
600
17
19
350
13
119
350
10
17
180
29
321
0
10
20
30
40
50
60
70
0
40
80
120
160
200
240
280
320
360
400
CAG Repeats
CGG Repeats

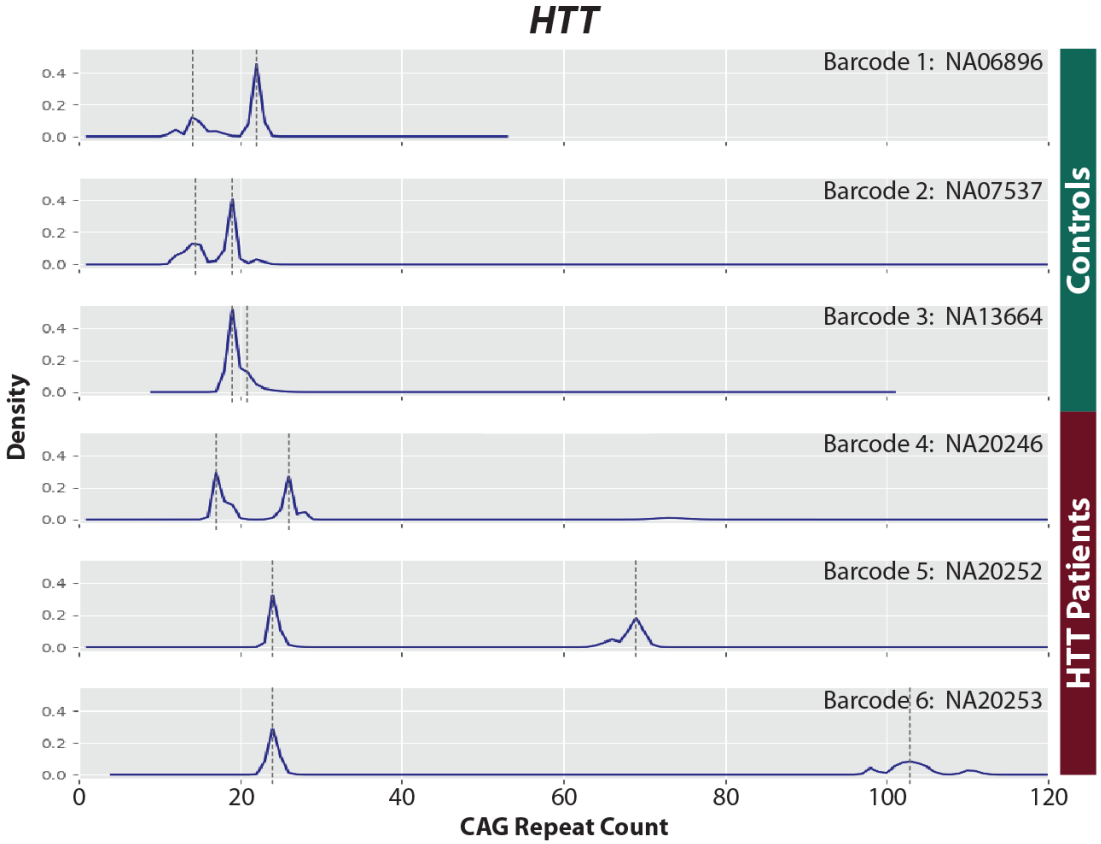
SUBREAD COVERAGE ON THE SEQUEL SYSTEM



1 Sequel SMRT Cell 1M

MULTIPLEXED SAMPLES ON THE SEQUEL SYSTEM

CAG Repeat Counts from 3 Controls and 3 HD Patients



CONCLUSION

Amplification-free enrichment with CRISPR/Cas9 and SMRT Sequencing achieves the base-level resolution required to understand the underlying biology of repeat expansion disorders

- Target any hard-to-amplify genomic region regardless of sequence context
- Avoid PCR bias and PCR errors
- Accurately sequence through long repetitive and low-complexity regions
  - Count repeats and identify sequence interruptions
- Detect sample mosaicism

PACBIO®

www.pacb.com