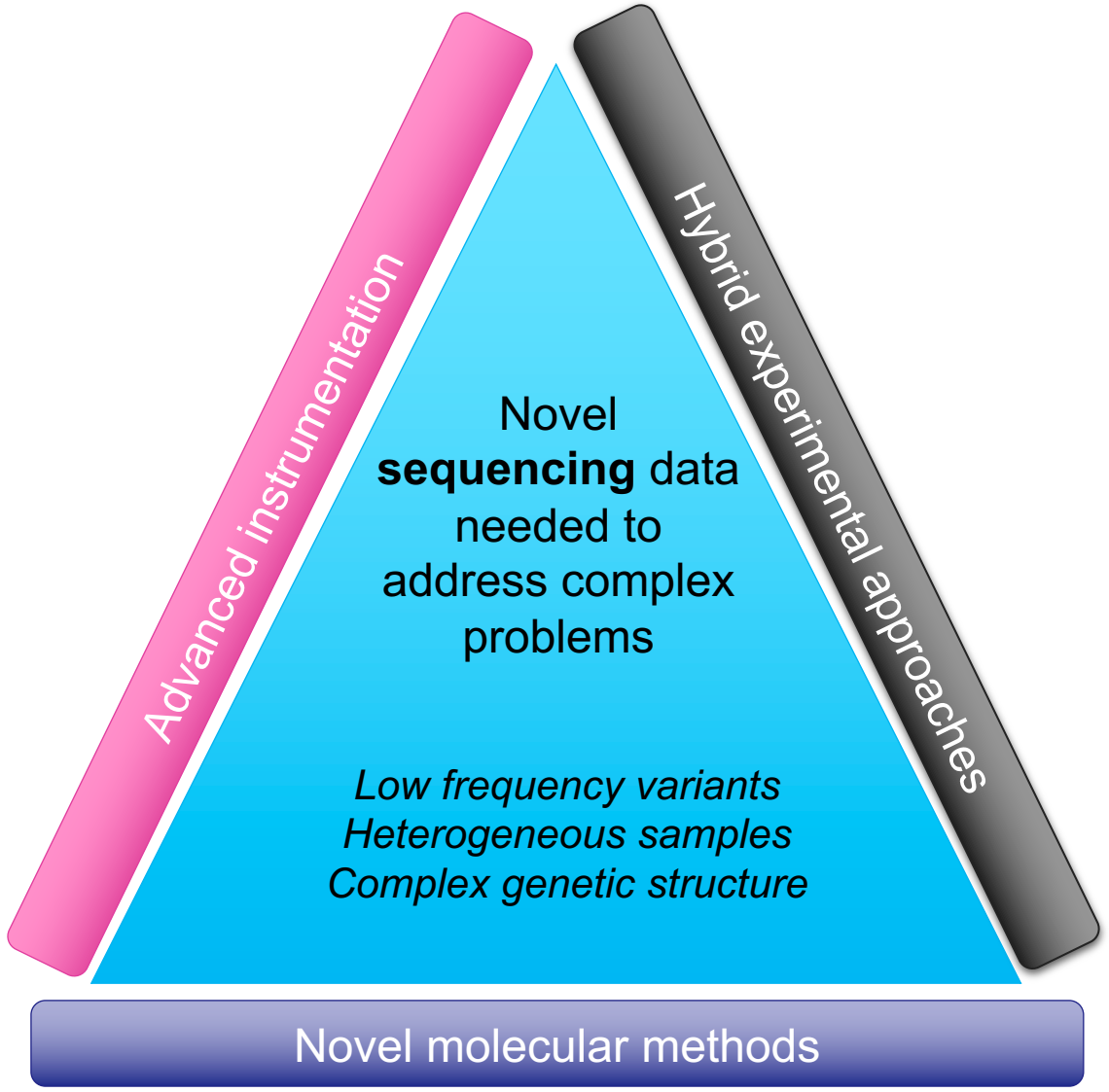
Sequel 2: Enabling SMRT Sequencing for All

September 26th, 2019

Melissa L. Smith, Ph.D.
Assistant Professor
Assistant Director of Technology Development
Icahn School of Medicine at Mount Sinai
Dept. of Genetics & Genomics Sciences
Icahn Institute for Data Science & Genomic Technology

Mount Sinai

ISMMS Technology Development Approach & Toolbox



Multiplatform NGS:

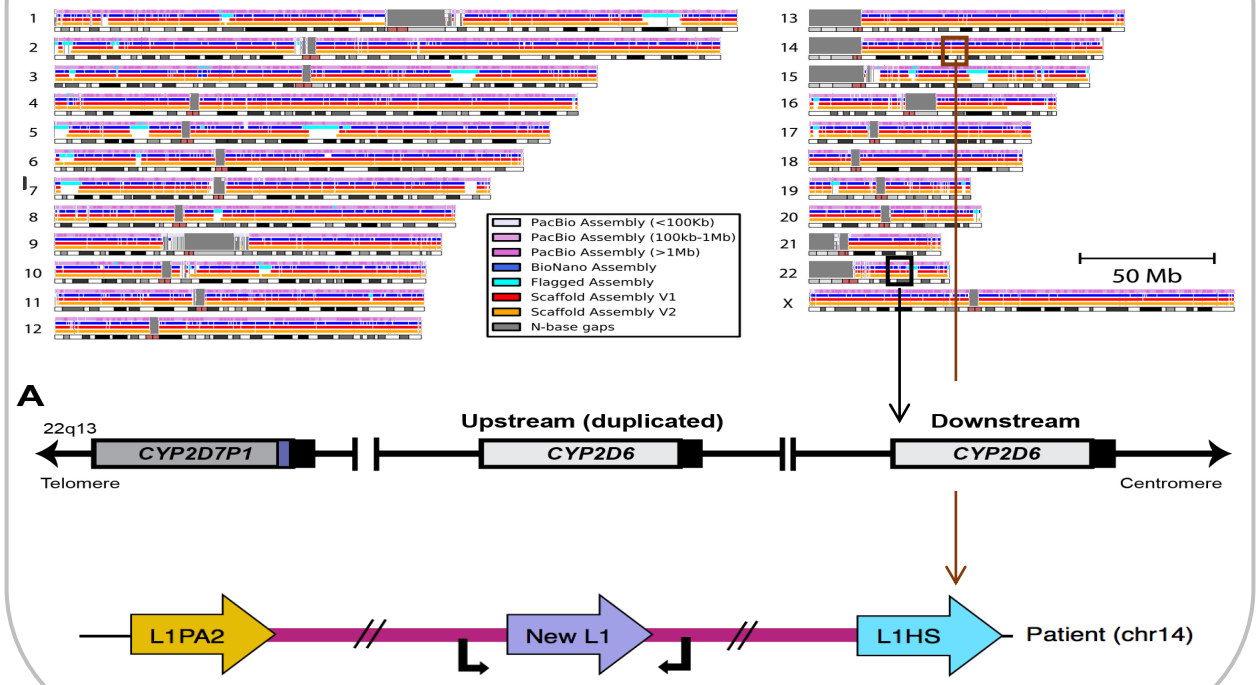
- 1 PacBio Sequel 2 System
- 3 PacBio Sequel 1 Systems
- 2 PacBio RSII Systems
- 1 ILMN NovaSeq 6000
- 3 ILMN Next550
- 2 ILMN MiSeqs
- 2 ILMN MiniSeqs
- 1 Illumina HiSeq 2500
- 3 Ion S5XL
- 2 Ion PGM
- 3 Ion Chefs
- Thermo-Fisher ddPCR
- 1 BioNano Irys
- Oxford Nanopore MinIONs
- 1 Oxford Nanopore GridION

Single Cell Characterization:

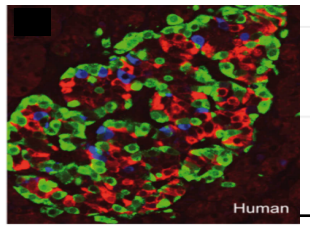
- 4 10X Genomics Chromium(s)
- 1 BLI Beacon
- 1 Mission Bio Tapestri
- 1 Celsee Genesis
- 1 ILMN/BioRad ddSeq

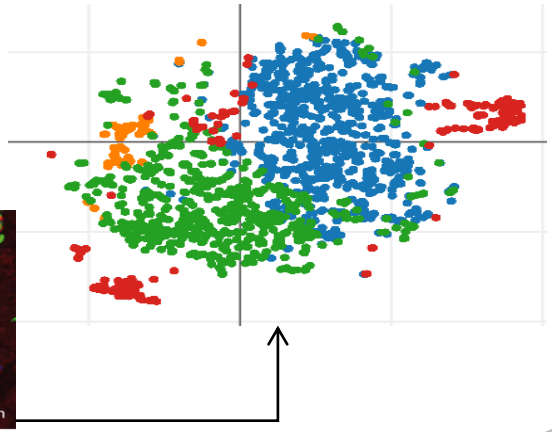
(Unique) Expertise: Utilizing Long Reads to Enable Translational Research

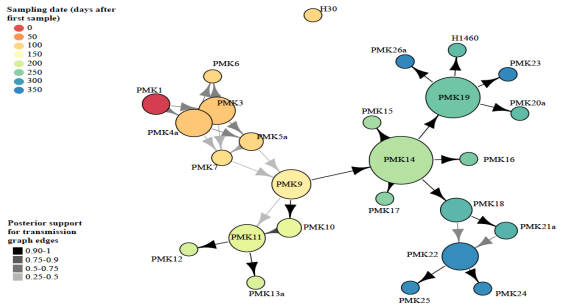
Addressing structure & targeting comprehensive diagnostics



Rapid validation of low frequency variants in somatic or complex tissue





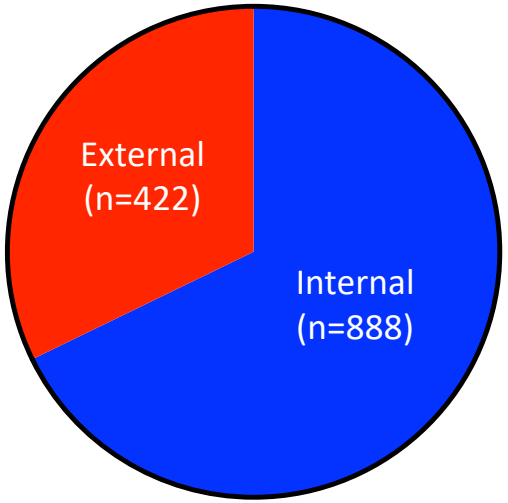


Infectious Disease & Immune Repertoire Profiling and Surveillance

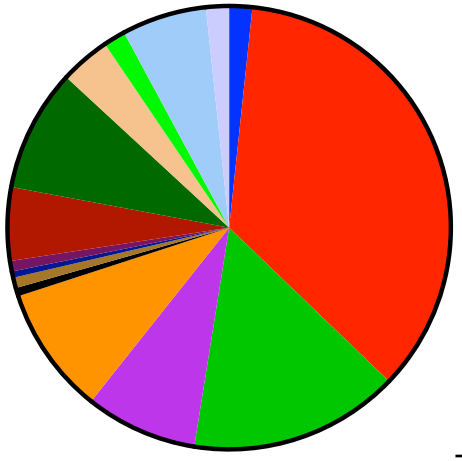
- Additional applications supporting many niche research applications not listed in categories above

2018 SMRTcells: by Disease Focus & Application (1310 total) + Sequel2

Customer/Collaborator Base:

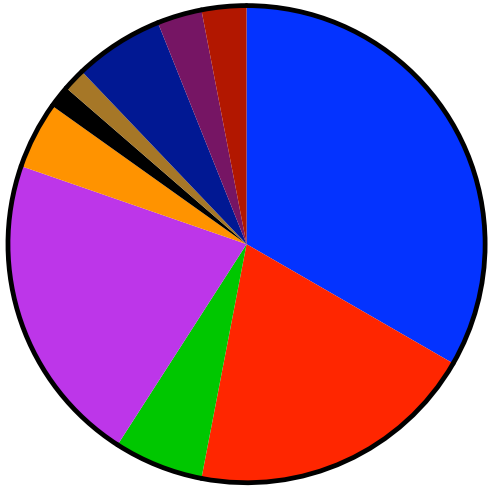


Disease focus:



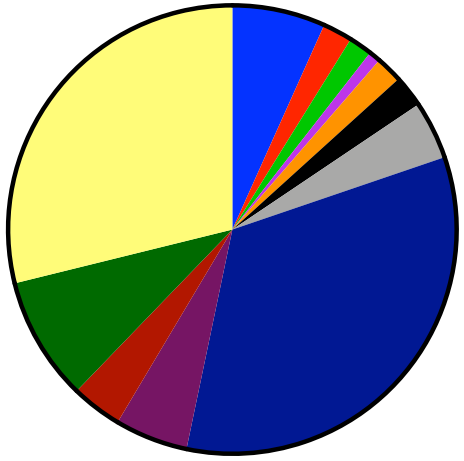
- Clinical (22)
- Infectious Disease (465)
- Cancer (201)
- Microbiology (107)
- Ecology (122)
- Gene Fusions (8)
- Gene Panels (10)
- Bioremediation (6)
- Developmental Biology (10)
- Neurology (70)
- Immunology (117)
- Metagenomics (48)
- Repeat Expansion Disorders (21)
- Reference Genomes (81)
- N/A (22)

SEQUEL2 (Jan 2019-now; n=66)



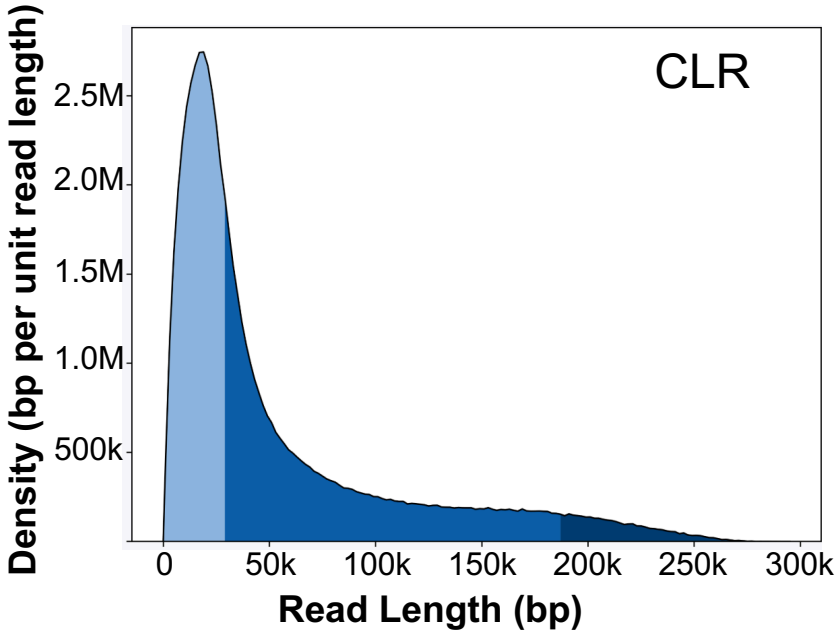
- CLR WGS (22)
- HiFi (13)
- 2-plex CLR (4)
- IsoSeq (14)
- sc IsoSeq (3)
- Target Enrichment (1)
- CATCH (1)
- Multiplexed Amplicons (4)
- Microbial WGS (2)
- Metagenomics (2)

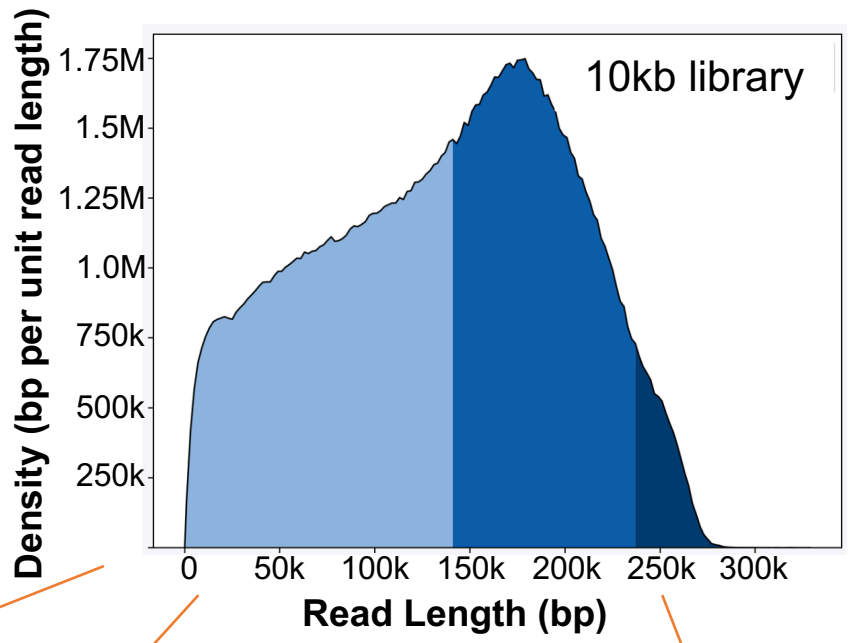
Application:



- Custom Amplicon (88)
- Cas9 Enrichment (28)
- Clinical Amplicon (22)
- Epigenetics (11)
- Fosmids (25)
- Immune Profiling (30)
- IsoSeq (55)
- Microbial WGS (439)
- Oligo Capture (69)
- Shotgun Metagenomics (48)
- Targeted IsoSeq (117)
- Whole Genome Sequencing (378)

Chemistry and system improvements on SEQ2 enable diverse applications





CLR WGS (SEQ2)

Avg Yield: 80-120 Gb
(~30-fold human genome)

FALCON for diploid aware assembly

Large SV detection/characterization

HiFi WGS (SEQ2)

Avg Total Yield: 214 Gb
Avg Q30 CCS Yield: 15 Gb
(5-6-fold hu genome)

HiFi-based genome assembly forthcoming

Q30 accuracy, SNVs

IsoSeq (SEQ1/2)

Avg Yield SEQ1: 30 Gb
Avg Yield: SEQ2: 204 Gb

Between 20k-140k HQ isoforms called/sample

SEQ2 enables single cell isoform sequencing

Targeted (SEQ1/2)

Avg Yield SEQ1: 25-35 Gb
SEQ2: 200-350 Gb

CCS enables HQ variant detection (SNV & SV)

Amplicons, Oligo capture, Cas9-based

Targets range in size from 2-20kb

Targeting disease-associated genes with long read sequencing

Loci with known disease association requiring advanced NGS

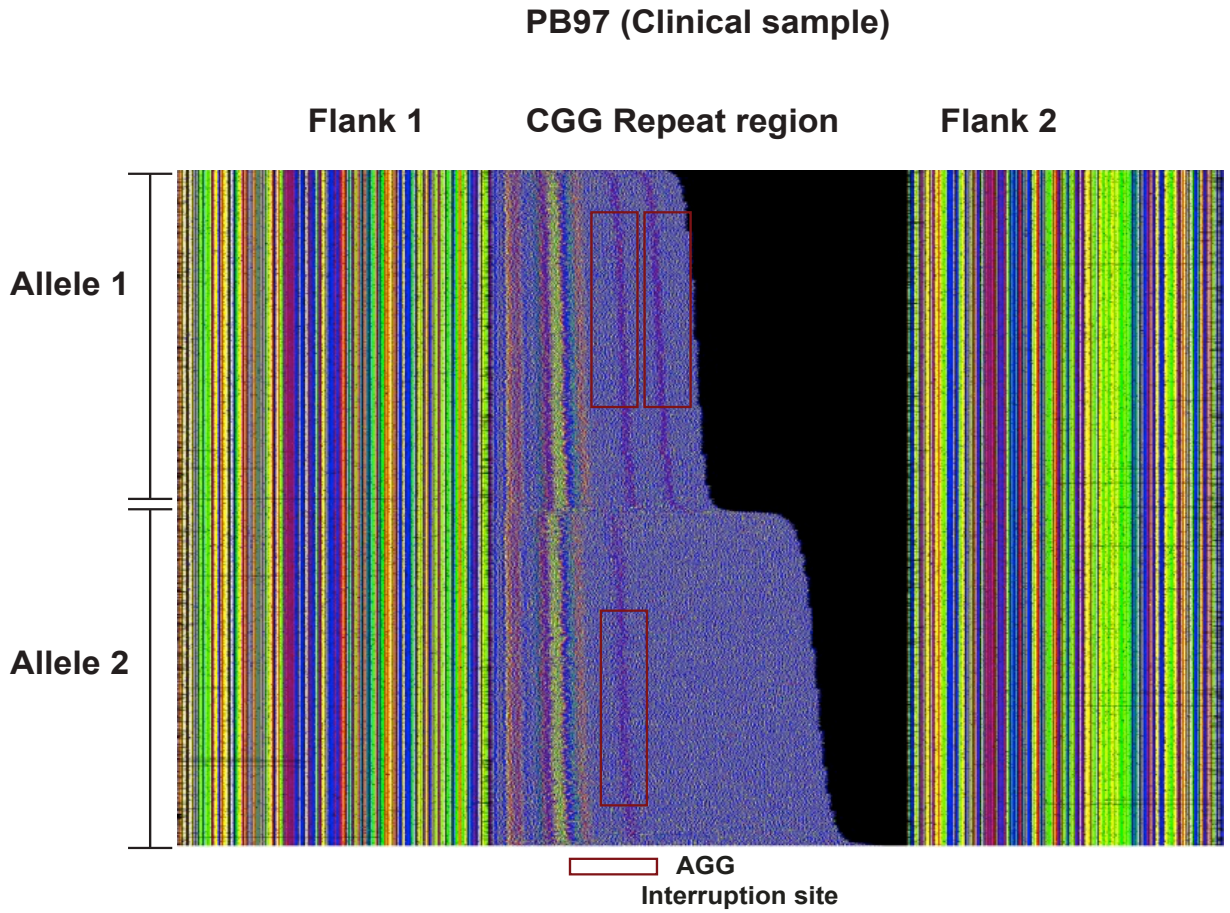
Disease Association	Gene(s)	Incl. Intronic Regions	Normal Repeat #	Pathogenic Repeat #
PKU	PAH	121525	NA	NA
Non-PKU HPA	PAH	121525*	NA	NA
PKU	GTPCH1	61144	NA	NA
PKU	PTPRS	182308	NA	NA
PKU	DHPR (QDPR)	51973	NA	NA
PKU	PCD (PCBD1)	6504	NA	NA
CF	CFTR	202881	NA	NA
CF	TGFBR1	50174	NA	NA
CF	CEACAM6	16784	NA	NA
CF	CEACAM3	15222	NA	NA
3-MCC	MCCC1	100857	NA	NA
3-MCC	MCCC2	71416	NA	NA
MCAD	ACADM	63224	NA	NA
VLCAD	ACADVL	8148	NA	NA
SCAD	ACADS	14273	NA	NA
C3	PCCA	441417	NA	NA
C3	PCCB	87590	NA	NA
C3	MUT	32831	NA	NA
C3	MMAA	41772	NA	NA
C3	MMAB	18424	NA	NA
C3	MMACHC	11014	NA	NA
C3	MMADHC	18182	NA	NA
C3	LMBRD1	121309	NA	NA
Krabbe	GALC	155845	NA	NA
Krabbe	PSAP	35071	NA	NA
Pompe	GAA	18323	NA	NA
Gaucher	GBA	10414	NA	NA
Fabry	GLA	10122	NA	NA
NPA/B	SMPD1	4573	NA	NA
DRPLA	ATN1	17858	7 to 25	49 to 88
SBMA	AR	185996	11 to 24	40 to 62
SCA1	ATXN1	462379	6 to 39	39 to 83
SCA2	ATXN2	147462	15 to 29	34 to 59

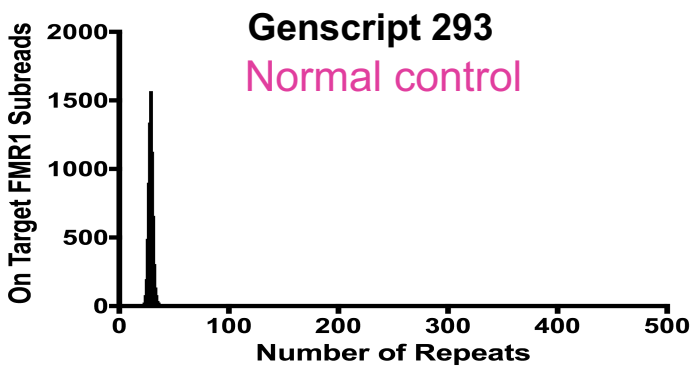
Disease Association	Gene(s)	Incl. Intronic Regions	Normal Repeat #	Pathogenic Repeat #
SCA3	ATXN3	48069	13 to 36	55 to 84
SCA6	CACNA1A	417548	4 to 16	21 to 30
SCA7	ATXN7	138905	4 to 35	34 to 300
SCA17	TBP	18567	25 to 44	45 to 66
DM1	DMPK	12835	5 to 37	>50 to 2000
DM2	CNBP	14438	<27	75 to 11,000
EPM1	CSTB	3933	2 to 3	30 to 75
FXS	FMR1	39176	6 to 52	55 to 2000
FRAXE MR	AFF2 or FMR3	500054	6 to 25	>200
FRA12A MR	DIP2B	243682	6 to 23	unknown
FRDA	FXN	64919	7 to 22	66 to 900
SCA10	ATXN10	173509	10 to 29	280 to 4500
SCA8	ATXN8 or ATXN8(	24333	6 to 37	107 to 250
HDL-2	JPH3	96321	<50	>50
SCA12	PPP2R2B	496411	<66	>66
HD	HTT	169268	10 to 35	>35
SMA	SMN1	19654	1 to 5	0
SMA	SMN2	19654*	0 to 5	NA
FXTAS	FMR1	39176*	6 to 52	55 to 2000
HLA	HLA-B	3316	NA	NA
Celiac disease	HLA-DQA1	11190*	NA	NA
Celiac disease	HLA-DQB1	11190	NA	NA
Narcolepsy	HLA-DQB1	8916	NA	NA
Anklyosing spondy	HLA-B	3316*	NA	NA
6-Mercaptopurines	TPMT	26763	NA	NA
Irinotecan	UGT1A1	13051	NA	NA
Warfarin	CYP2C9	50732	NA	NA
Warfarin	VKORC1	5138	NA	NA
Tricyclic Antidepre	CYP2D6	4407	NA	NA
Abacavir	HLA-B	3316*	NA	NA
Carbamamazepin	HLA-B	3316*	NA	NA
Ximelagatran	HLA-DRB1	34102	NA	NA
Ximelagatran	HLA-DQA1	22353	NA	NA

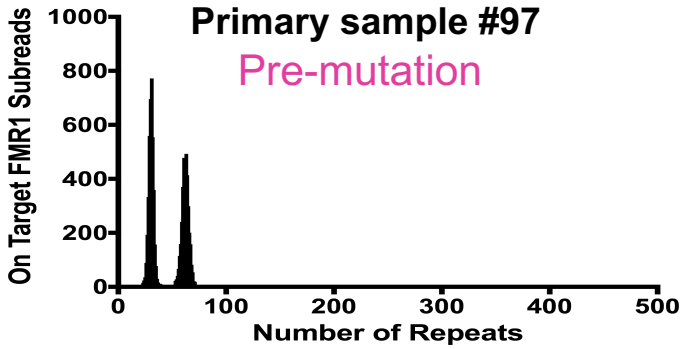
We need to gain a better understanding of pseudogenes, repeat expansions, and polymorphic loci for disease diagnostics

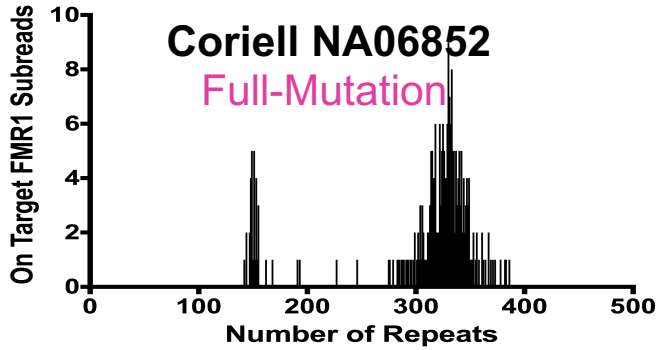
Every individual has ~1000 SVs that are > 2500bp in length, requiring long read sequencing

Long read, NoAmp SMRT sequencing resolves FRAX expansions





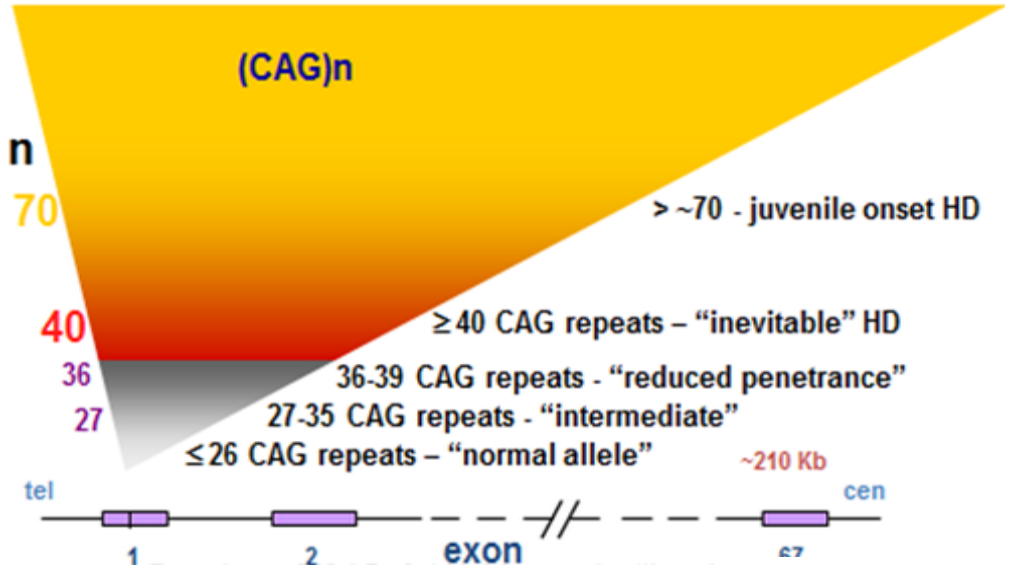




Proper phasing of heterozygous events can often be missed with traditional methods.

What about repeat interruptions and methylation?

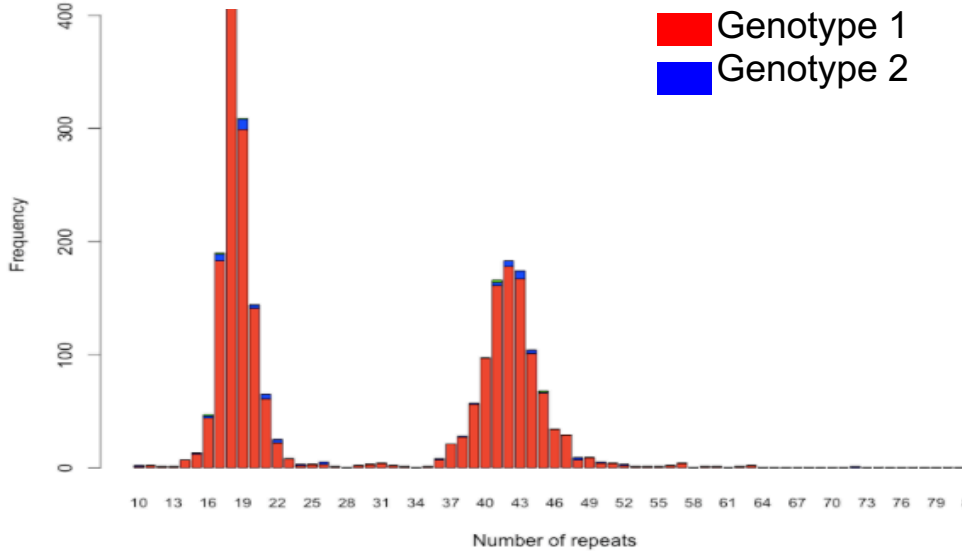
Targeted IsoSeq of HTT locus; phasing drug targets with severity loci



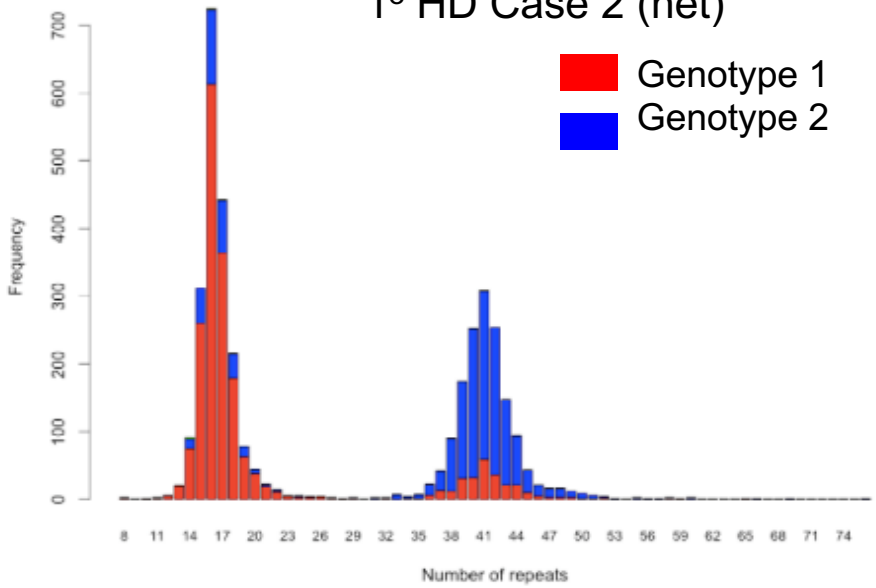
Fahn, S. et al, *Principles & Practice of Movement Disorders, Chapter 1*, 2011.

From ~210kb (DNA) → ~9-10kb (RNA)

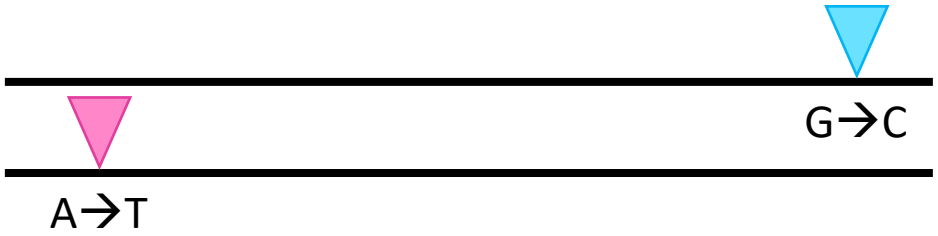
1º HD Case 1 (hom)



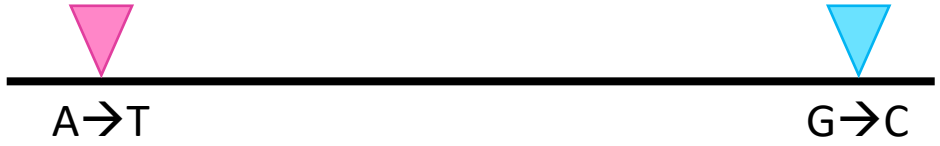
1º HD Case 2 (het)



Targeted IsoSeq to determine orientation of SNVs: *cis* vs *trans*



vs.



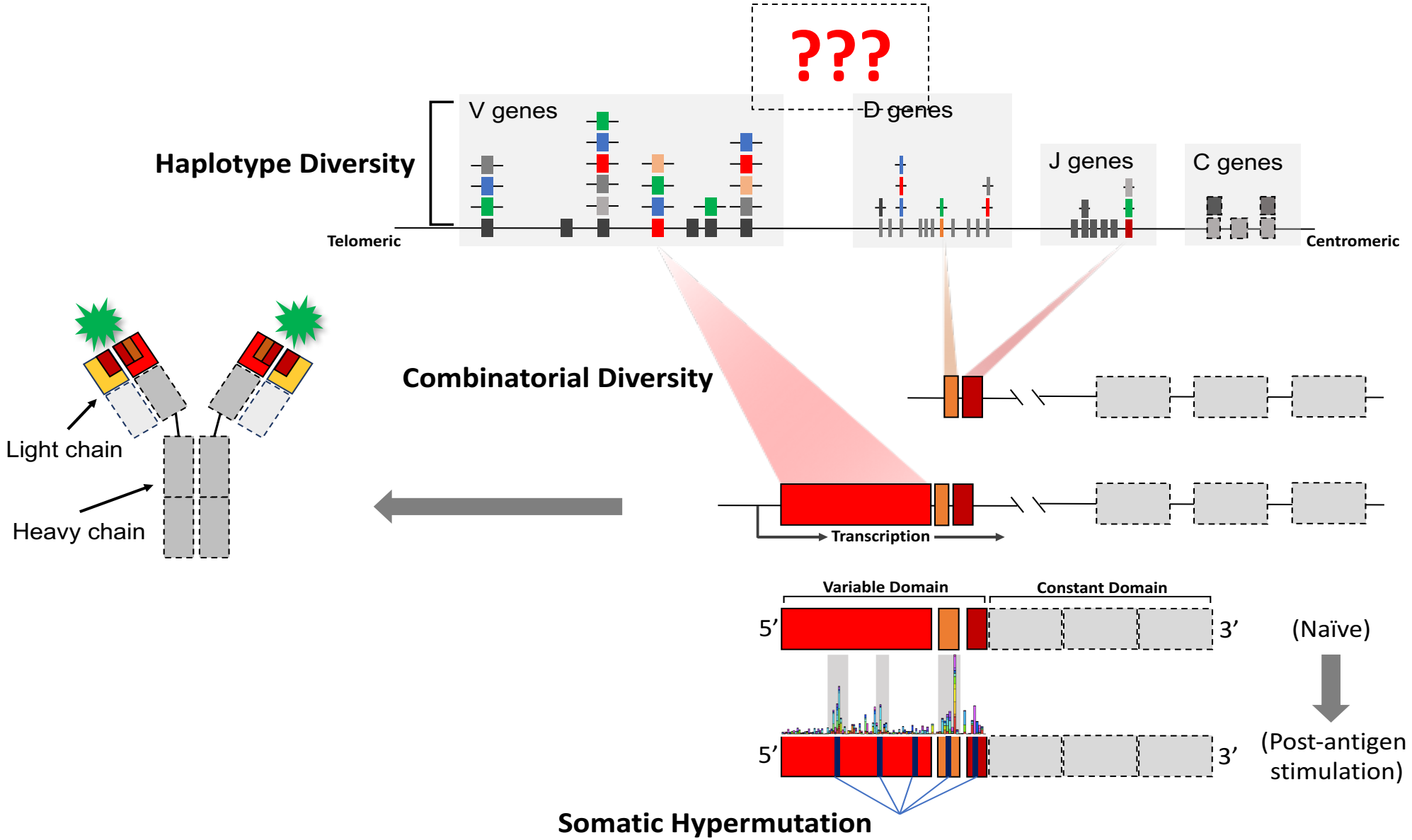
								A	B	C	D
Cell Line A	Gene A			Sample Variants				Haplotype %			
	Codon	AA	Position	AA	Codon	%	Coverage	66.9	21.6	6.1	5.4
	CCT	P	123	R	CGT	29	3710				
	CAT	H	4567	R	CGT	32	7573				

Primary Sample #1								A	B	C	D
SNP1/ SNP2	GeneA			Sample Variants				Haplotype %			
	Codon	AA	Position	AA	Codon	%	Coverage	54.4	45.1	0.3	0.1
	GAG	E	123	K	AAG	46	21258				
	ACC	T	4567	A	GCC	45	20520				

Primary Sample #2								A	B	C	D
SNP3/ SNP4	GeneB			Sample Variants				Haplotype %			
	Codon	AA	Position	AA	Codon	%	Coverage	62.8	35.7	0.8	0.7
	GAG	E	789	K	AAG	62	18186				
	ATG	M	1112	L	TTG	63	21161				

Characterizing diversity in the immunoglobulin loci

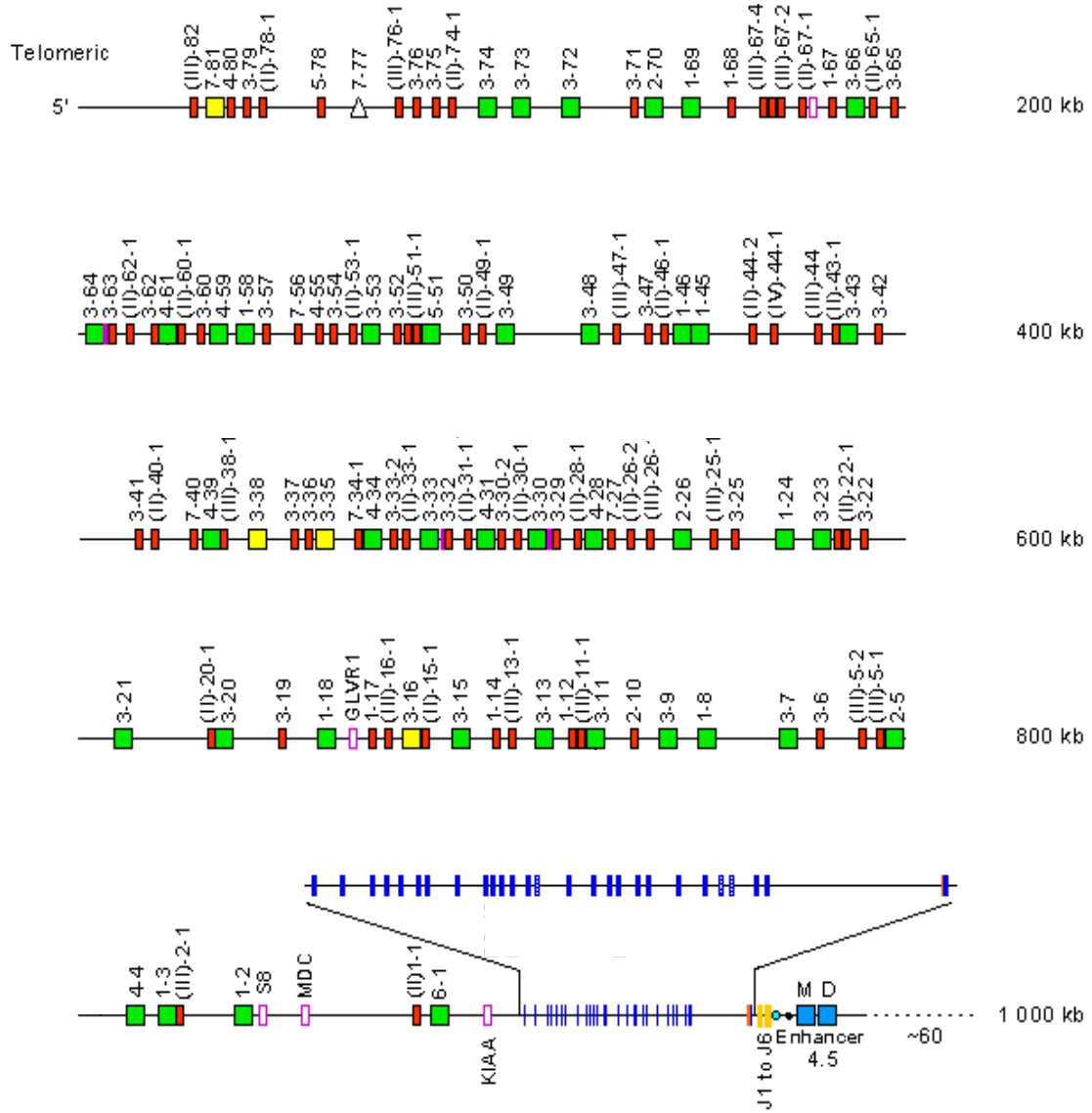
Multiple mechanisms underlie the diversity of the antibody repertoire



(Watson, Glanville, Marasco, Trends Immuno 2017)

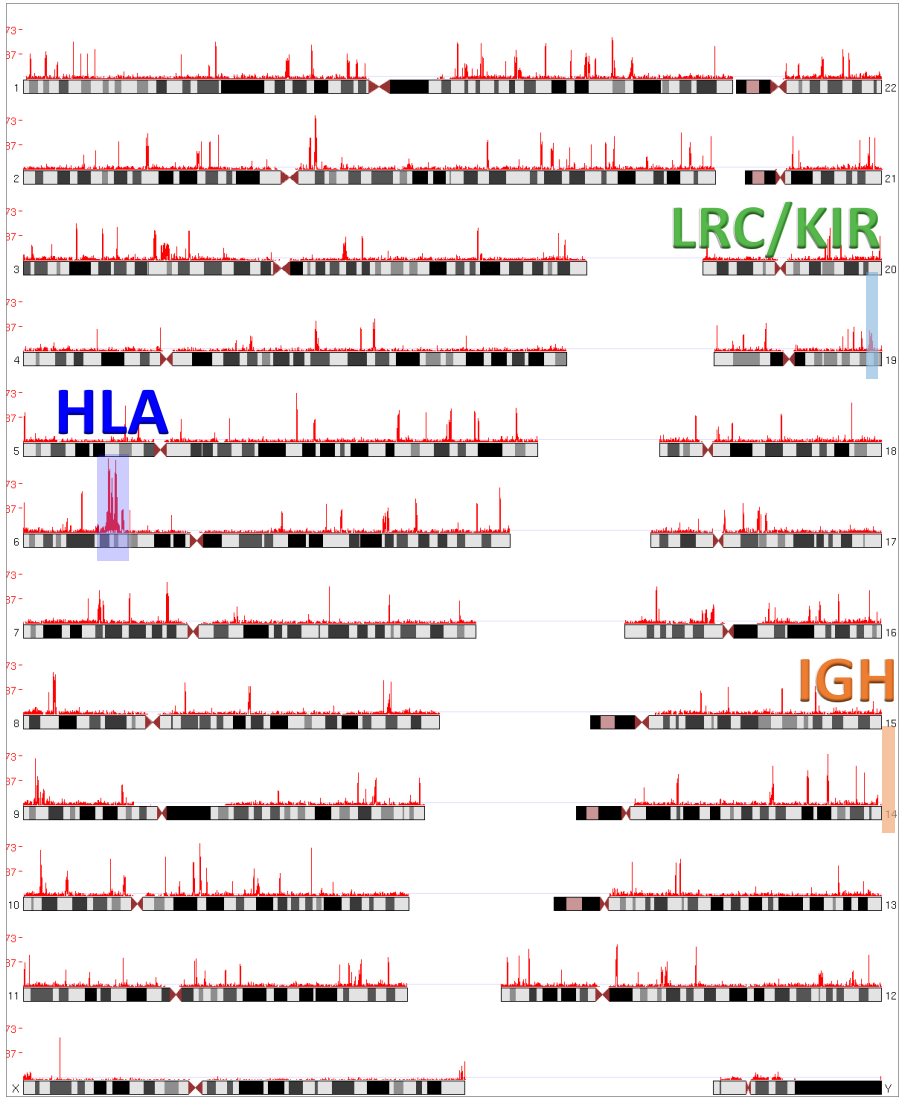
Genetic variation at the immunoglobulin loci is complex & poorly understood

- >50% of IGH sequence falls within a segmental duplication.
- CNVs are prevalent in IGH
  - ~50% of functional/ORF IGHV genes vary in copy number
- Currently, the full ~1MB IGH V, D and J regions (excluding IGHC) has only been sequenced twice.
  - Many genes thought not to be represented in current assemblies



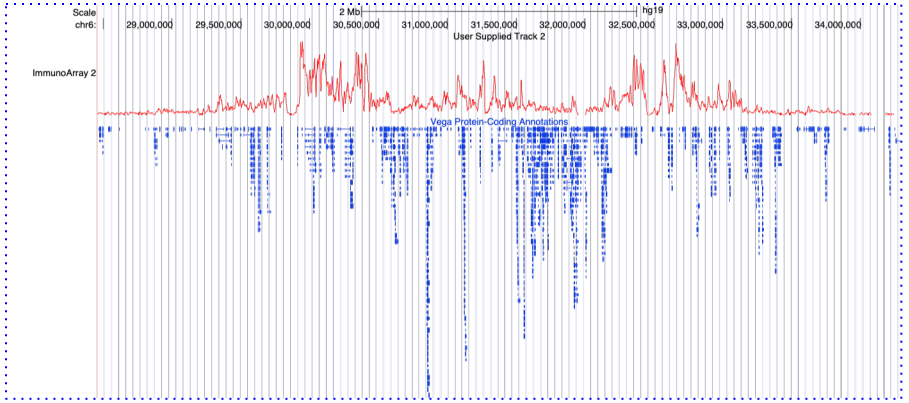
(IMGT, 2001)
(Watson, CT, et al., AJHG 2013)

Incomplete genomic resources has impacted association studies



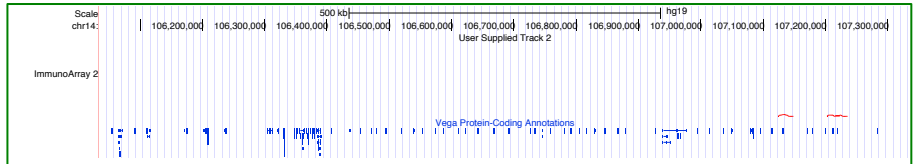
HLA (6Mb)

SNP total = 9,319;
Mean SNP#/10Kb = 16.28



IGH (1.3 Mb)

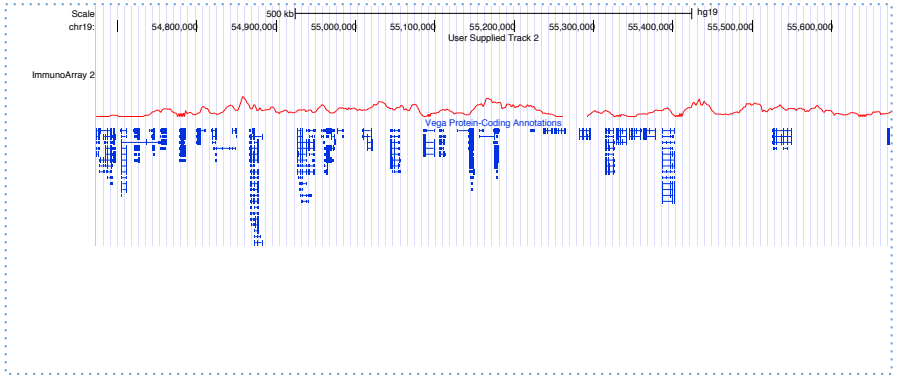
SNP total = 5;
Mean SNP#/10Kb = 0.03



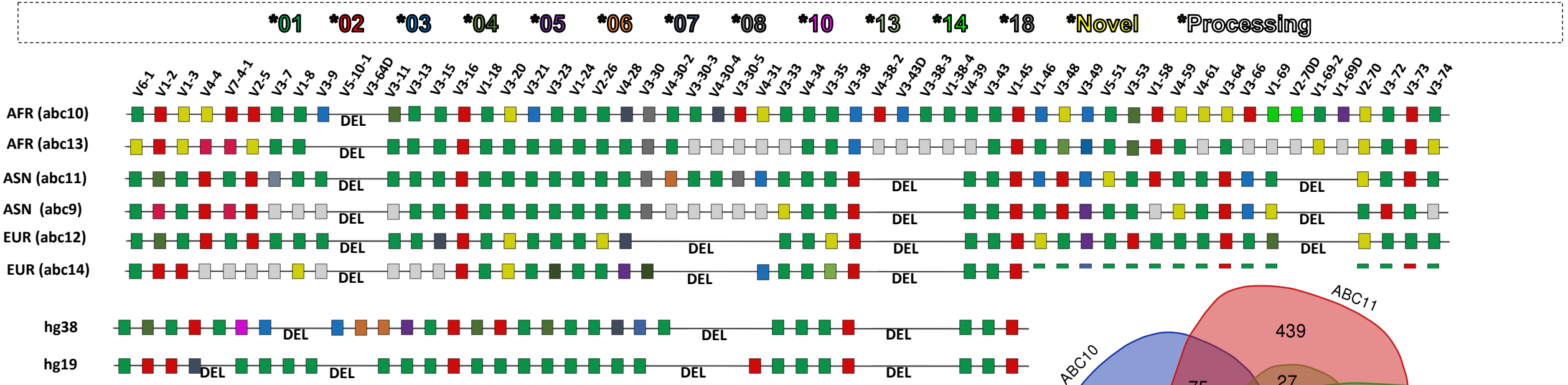
Food for thought: Between just 2 haplotypes in IGH, we have observed >2,800 SNPs!!!

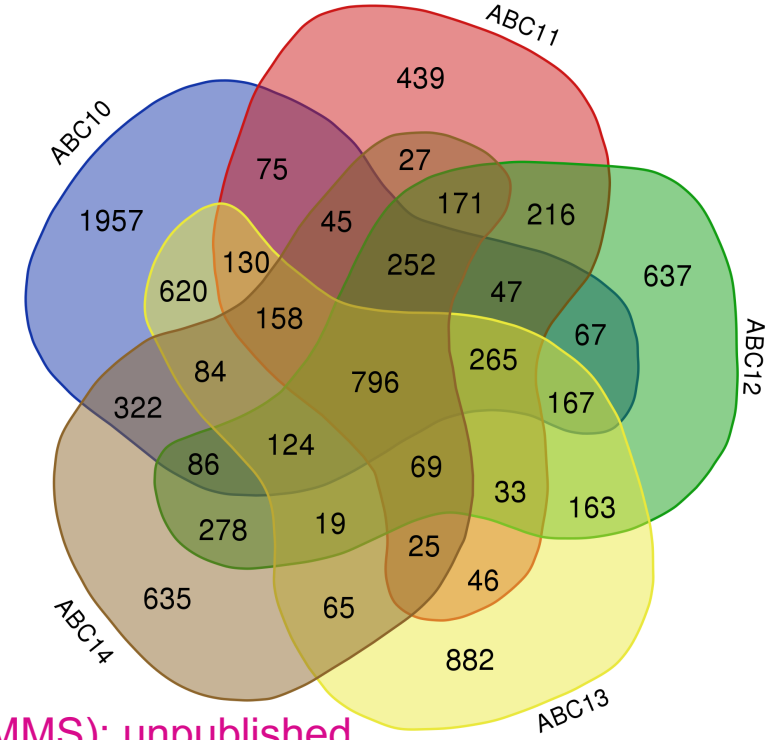
LRC/KIR (1 Mb)

SNP total = 818;
Mean SNP#/10Kb = 8.3



New fosmid-derived IGH references show extensive diversity

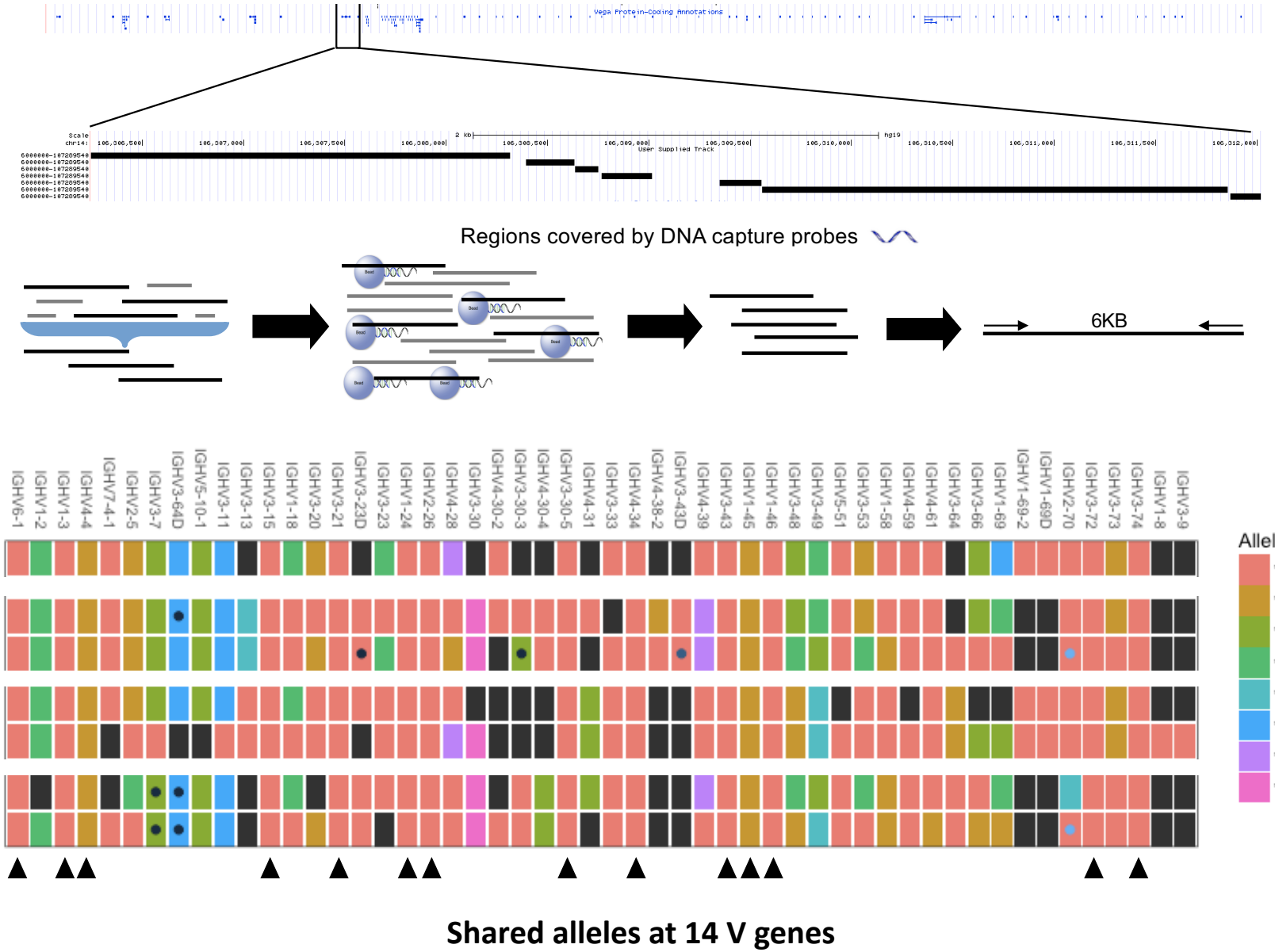




ABC10 (**AFR**) – **14** non-IMGT alleles
ABC13 (**AFR**) – **11** non-IMGT alleles
ABC11 (**ASN**) – **5** non-IMGT alleles
ABC9 (**ASN**) – **4** non-IMGT alleles
ABC12 (**EUR**) –**10** non-IMGT alleles
ABC14 (**EUR**) – **2** non-IMGT alleles

Collaboration with Corey Watson (University of Louisville), Andy Sharp & Ali Bashir (ISMMS); unpublished

Using diverse reference data to develop and effective IGH genotype assay

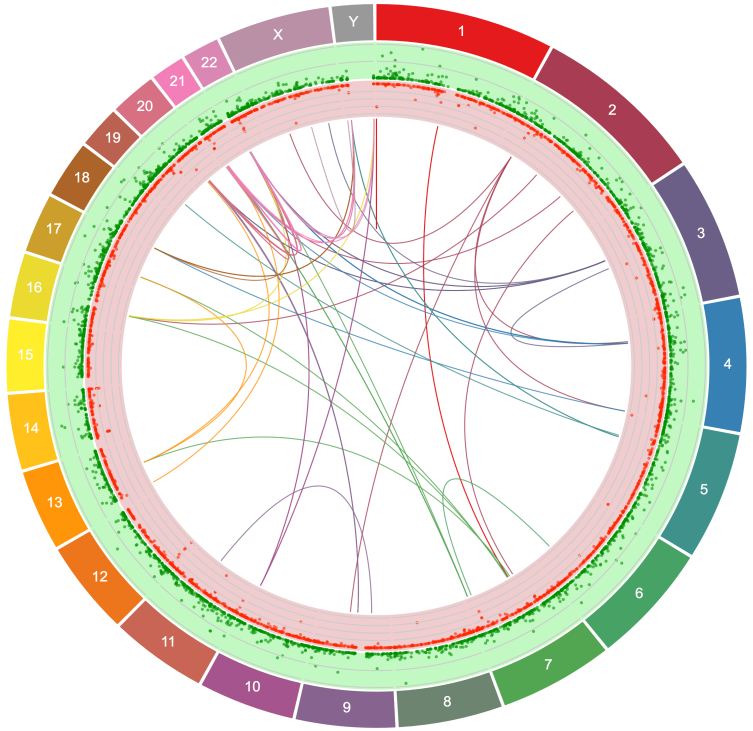


Shared alleles at 14 V genes

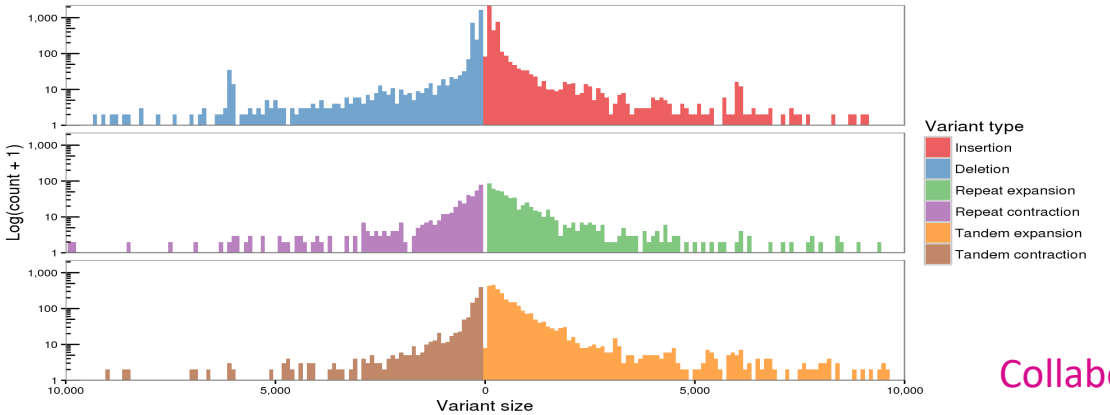
Collaboration with Corey Watson (University of Louisville); unpublished

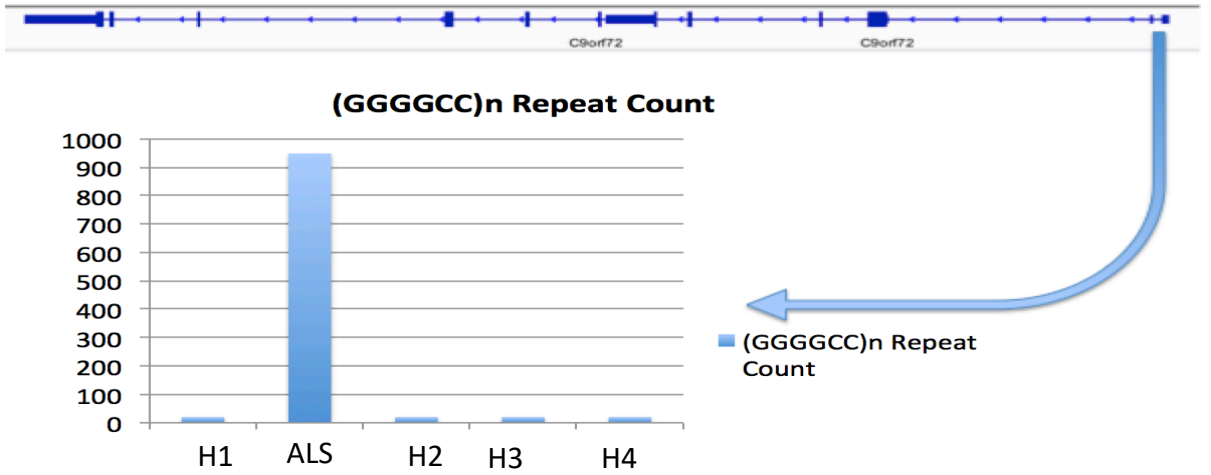
Long read WGS for structural variant detection

Sequel1 WGS: ALS c9orf72+ case with PB-SV calls



Plot Stats: SV > 50 bp Shown

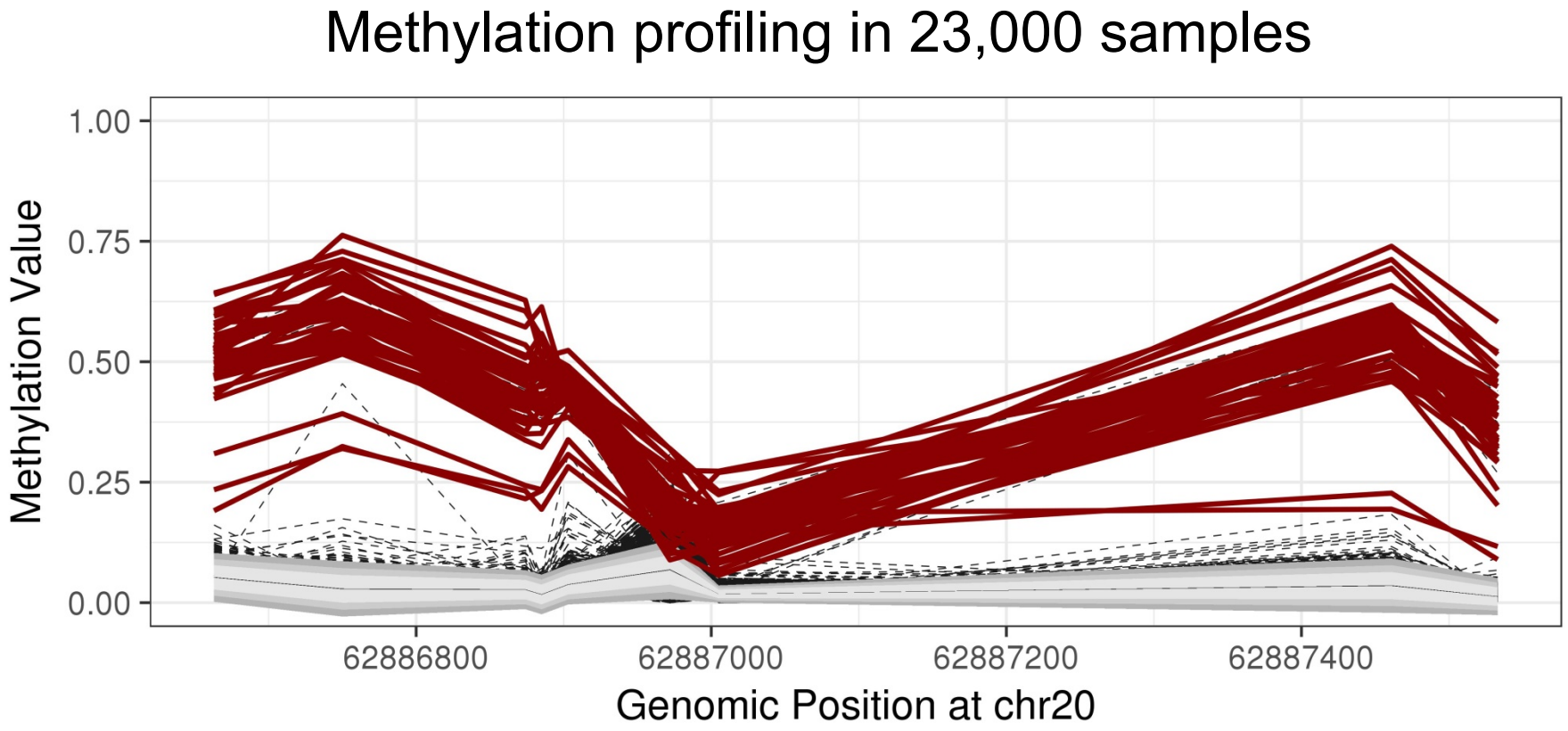




Longest G4C2 repeat ever sequenced

Collaboration with Jamie Gregory (NYGC) & Neil Schneider (Columbia); unpublished

Strong hypermethylation in PCMTD2 – caused by CGG repeat expansion?



Approx 1 per 1,000 samples show hypermethylation of *PCMTD2* promoter, a region that contains a highly polymorphic CGG repeat

Data kindly provided by Dr. Andrew Sharp (ISMMS), unpublished

Sequel 2 low coverage CLR WGS resolves novel CGG repeat expansion

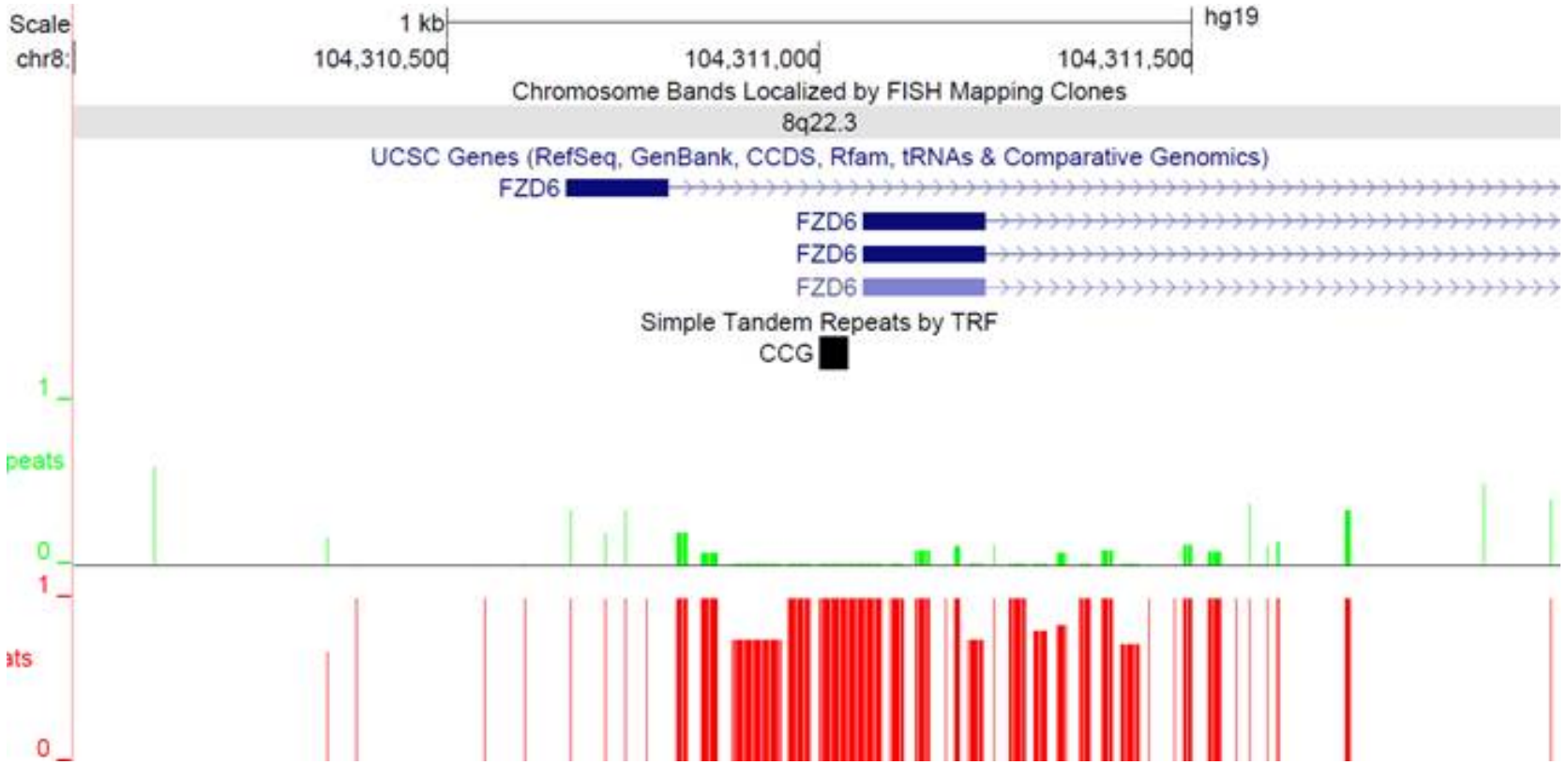
CLR gDNA libraries were duplexed and sequenced on a single SMRT cell 8M

SMRTcell output ~100 Gb; generating ~15-fold coverage per genome

SMRT sequencing data showed insertions of 1.8-2.7 kb (625-839 copies of novel CGG repeat), overlapping hypermethylated region

Data kindly provided by Dr. Andrew Sharp (ISMMS), unpublished

Orthogonal validation of CGG repeat and methylation with nanopore



CGG repeats were orthogonally validated using Oxford Nanopore MinION sequencing

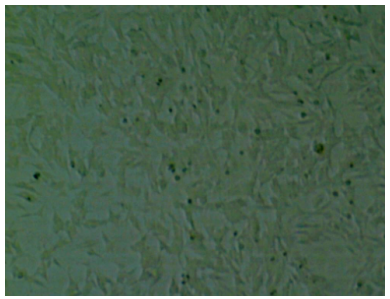
Direct methylation signatures could be observed on the allele carrying the expanded CGG repeat (red); not the normal allele (green)

Data kindly provided by Dr. Andrew Sharp (ISMMS), unpublished

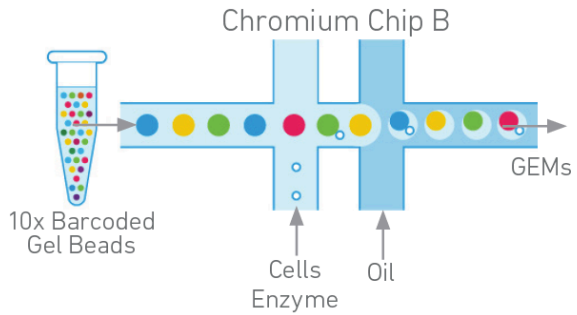
Looking Ahead: Resolving Isoform Diversity at the Single Cell Level

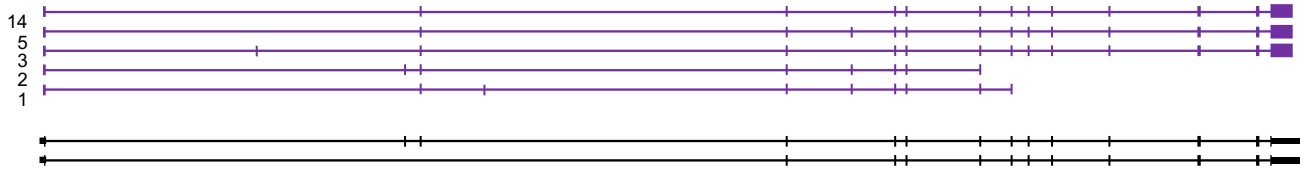
Merging long read sequencing with single cell transcriptomics: scIsoSeq

SH-SY5Y
cell culture

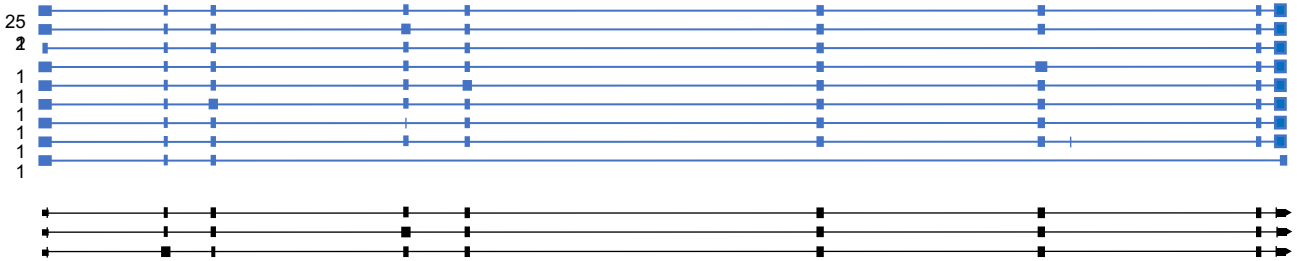


Single cell partitioning
on 10X instrument

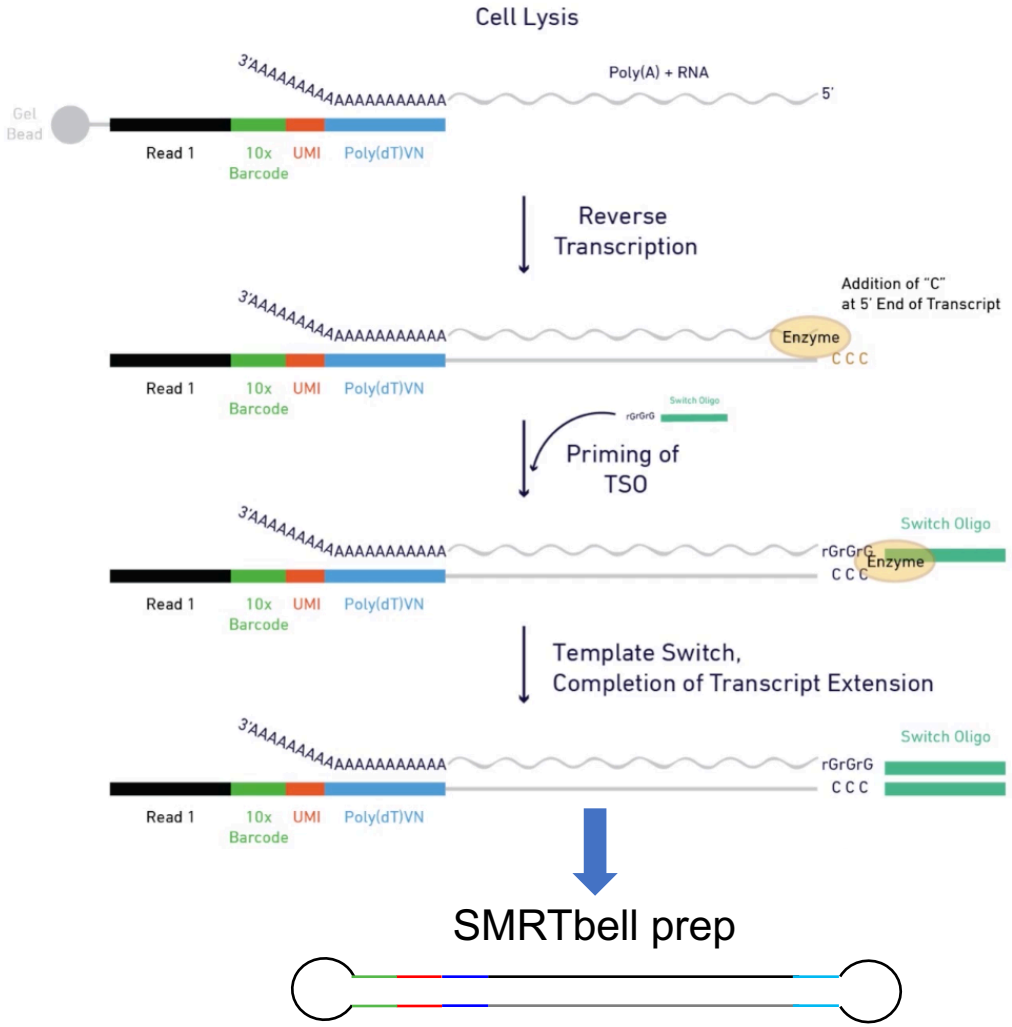




NCBI RefSeq Annotation of *SEPT7*



NCBI RefSeq Annotation of
SMARCB1



SEP7: encodes a cytoskeletal GTPase that contributes to neuronal dendritic spine maturation

SMARCB1: encodes a component of neuron-specific chromatin remodeling complex, nBAF, playing a critical role in dendritic arborization during neuronal development

Pricing single molecule whole genome sequencing to be truly accessible

RS2

~1 Gb per SMRTcell
(850 Mb expected avg)

Multiple human genomes sequenced

HG005 (2.1 chemistry)

Total: 46 SMRT cells (60-fold coverage)

Raw Cost of Sequencing: **~$42k**

Sequel 2

~100-350 Gb per SMRTcell 8M
(100-300 Gb expected)

Multiple human genomes sequenced

NA12878 (XL-C2 and P5/C3 chemistry)

Total: 1,013 SMRT cells (46-fold coverage)

Raw Cost of Sequencing (not including libraries/instrumentation): **~$150k**

Sequel 1

~10-35 Gb per SMRTcell 1M
(10-40 Gb expected)

Multiple human genomes sequenced

HG007 (CLR; v1 chemistry): 2 SMRTcells (82-fold coverage)

Raw Cost of Sequencing: **$2600**

HG007 (HiFi; v1 chemistry): 4 SMRTcells (21-fold coverage)

Raw Cost of Sequencing: **$5200****

Conclusions and Looking Ahead

- Biomedical and translational research can be complemented & enhanced using genomics approaches of all types (and combined!)
- Long, single molecule reads provide a unique tool to provide phased, single molecule resolution of complex regions of genomes
- Ongoing chemistry, platform and sample preparation improvements are enabling the accessibility of single molecule sequencing for all applications
- 7 word challenge: Long reads resolve complexity underlying disease etiology

melissa.smith@mssm.edu

OUR TEAM

Bobby Sebra
Kristin Beaumont
Michael Beaumont

Illumina Library Prep and Sequencing
James Powell
Diane Castillo
Swaroop Bose
Nataly Fishman
Kent Madrid
Irene Salib
Juan Soto
Konstantinos Vlachos
Marissa Zyndorf

Ion Torrent
Maya Strahl
Ethan Ellis
Hanane Arib

Single Cell
Kimaada Allette
Arpit Dave
Ethan Ellis*
Ching He
Sanjana Schroff

Long Read
Gintaras Deikus
Matt Emery
Nancy Francoeur
Irina Oussenko
Zelda Salfati
D'andra Howell
Ethan Ellis*

Bioinformatics/Analysis
Hardik Shah
Mondale Wang
Ayesha Rasool
David Sachs
Shwetha Sridhar
Peter Warburton

COLLABORATORS from Today's Slides

Corey Watson
Oscar Rodriguez
William Gibson
Ali Bashir
Hemali Phatnani
Jaime Gregory
Neil Schneider

And many others….

Andrew Sharp

National Institute of Allergy and Infectious Diseases

R24 AI138963
R21 AI142590
R01 AI121285
R21 AI117407