

From: [Mark Pratt](#)
To: [John West](#)
Subject: Lab projects
Date: Monday, August 6, 2012 11:40:50 AM
Attachments: [LabUpdate_120802.pptx](#)

Hi John -

I'm attaching Shujun's slides on lab R&D around accuracy. Sorry about the size but it includes a lot of read stacks. This doesn't include production and/or diversity panels but is more focused on specific problems. On the last page is a grid where we evaluate the basic approaches. The only one that we are planning to drop from the agenda is the low concentration phasing since there are others doing this (GNOM and Moleculo).

The price for the used CHEF system was \$1000.

^^

Personalis Wet Lab Update

Shujun Luo
08/02/2012

Goals

- Expanding Repeat Diseases (small, challenging, high value set)
 - Characterize achievable performance with contemporary technologies (v3 chemistry & lobSTR)
 - Develop custom sequencing libraries or methods as necessary
- Difficult motifs (Homopolymers, STRs)
 - Identify and develop challenge sets
 - Develop methods to improve performance
- SD/Compression
 - Use library methods to reduce mapping and assembly degeneracy
 - Use library methods to improve haplotype performance in degenerate regions
- Phasing
 - Link heterozygous loci across genes
 - Develop target set (if necessary)
- Improve performance on interpretable content
 - Customized libraries and sample preparation methods



Grand Entry



R&D Lab

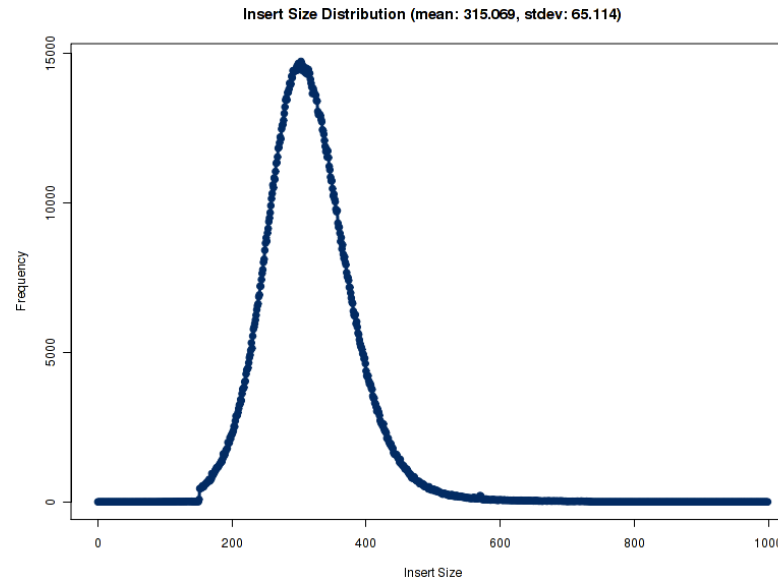
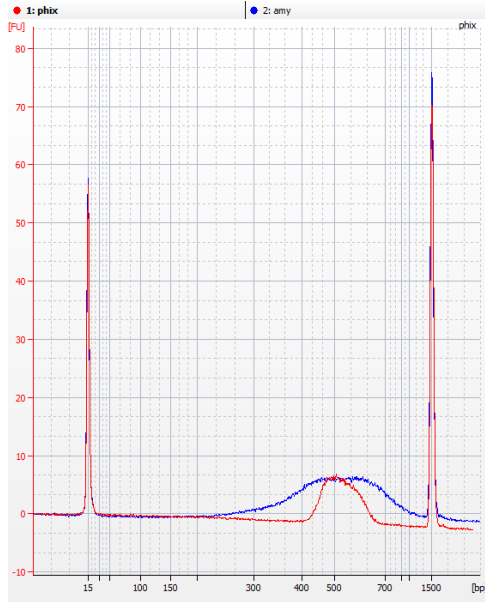


PCR Room



Ghost Room
Personalis EX2075

Sequencing Platform Performance Bench Mark



John's analysis:

R1 of 2x300:

Median # bases $\geq$ Q25: 258

Median # bases $\geq$ Q30: 243

Best read out of 2,500:

299 bases $\geq$ Q25

295 bases $\geq$ Q30

90th percentile:

282 bases $\geq$ Q25

271 bases $\geq$ Q30

R2 of 2x300 :

Median # bases $\geq$ Q25: 202

Median # bases $\geq$ Q30: 187

Best read out of 2,500:

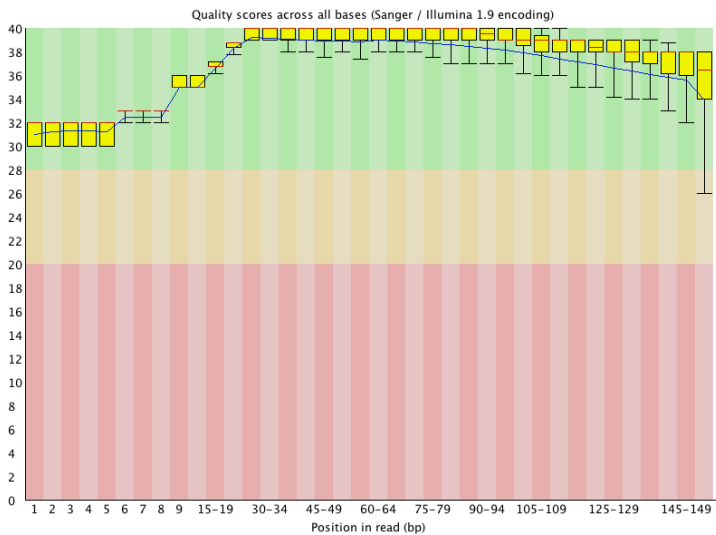
251 bases $\geq$ Q25

233 bases $\geq$ Q30

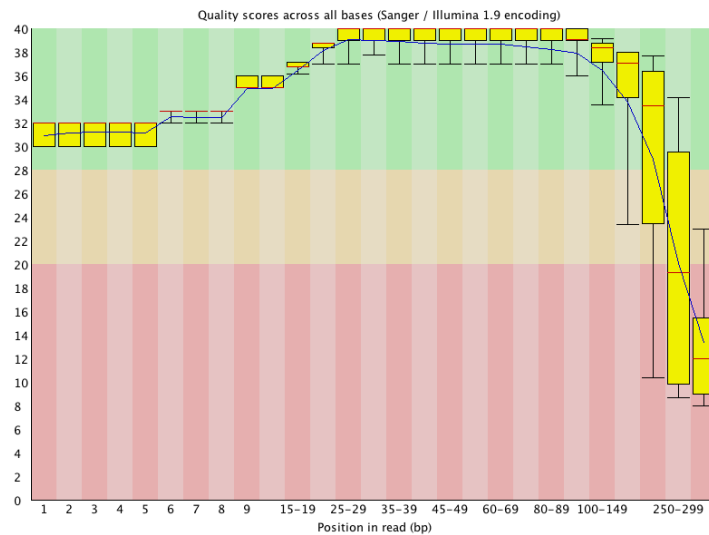
90th percentile:

225 bases $\geq$ Q25

210 bases $\geq$ Q30



2x150



2x300

Personalis EX2075

Current Status on 19 ExpandRepeat Diseases

---data from Amy ERX069505 (2x100)

Chr	Start	End	Disease	Nan's Lobstr v1									shujun's eye				
				STR	Period	Reference copy	Allelotyp e	Coverage	Supporting reads	Conflicting reads	All alleles	Score	STR	Period	Reference copy	Allelotype	note
chr1	208440810	208440838		AGC	3	9.3	39,39	1	1	0	39:1	1					
chr12	7045879	7045938	Dentatorubral-pallidoluyisian atrophy (DRPLA))	AGC	3	19.7	0,0	16	16	0	0:16	1	CAG/CAA	3	19	19,?	good coverage, could be only one allele
chr12	112036753	112036823	Spinocerebellar ataxia 2 (SCA2)	AGC	3	23.3	0,0	9	8	1	-3:1/0:8	1.36E-10	CTG	3	23	22,23	lobstr misses one allele
chr13	70713483	70713517		ACT	3	11.3	6,12	2	2	0	6:1/12:1	1					
chr13	70713515	70713561	Spinocerebellar ataxia 8 (SCA8)	AGC	3	15.3	0,0	1	1	0	0:1	1	CTG	3	15	15,16	lobstr misses one allele
chr14	92537354	92537396	Machado-Joseph disease (SCA3)	AGC	3	14	27,30	10	10	0	27:6/30:4	1	CTG	3	8	>14	no cross over, lobstr should be correct, hugeseq didn't call long insertion well
chr16	87637888	87637935	Huntington disease-like 2 (HDL20)	AGC	3	15.3	0,6	12	12	0	0:8/6:4	1	TGC	3	14	14,16	agree with lobstr
chr19	13318672	13318712	Spinocerebellar ataxia 6 (SCA6)	AGC	3	13.3	-6,-3	16	16	0	-6:12/-3:4	1	TGC	3	13	11,12	agree with lobstr
chr19	13318970	13319016		AGG	3	15.3	-6,0	5	5	0	-6:3/0:2	1					
chr19	46273462	46273524	Myotonic dytroph 1 (DM1)	AGC	3	20.7	-21,-21	9	9	0	-21:9	1	CAG	3	20	13,5	lobstr misses one allele
chr21	45196323	45196360	Myoclonic epilepsy	no call									CGCGGGGCGGGG	12	3	2,3	
chr22	46191234	46191304	Spinocerebellar ataxia 10 (SCA10)	no call									ATTCT	5	14	13,?	good coverage, could be only one allele
chr3	63898360	63898392	Spinocerebellar ataxia 7 (SCA7)	AGC	3	10.7	0,6	10	10	0	0:5/6:5	1	CAG	3	10	10,12	agree with lobstr
chr3	128891419	128891502	Myotonic dytroph 2 (DM2)	AGGC	4	20.8	-24,-24	1	1	0	-24:1	1	CAGG	4	20	15,?	disagree with lobstr
chr3	128891538	128891576		AC	2	19	6,18	9	8	1	6:6/14:1/18:2	1					
chr4	3076603	3076667	Huntington disease	AGC	3	21.3	-9,-3	10	10	0	-9:5/-3:5	1	CAG	3	21	20,18	agree with lobstr
chr4	3076665	3076695		CCG	3	10	0,9	11	11	0	0:4/9:7	1					
chr5	146258290	146258322	Spinocerebellar ataxia 12 (SCA12)	AGC	3	10.7	0,12	25	25	0	0:15/12:10	1	GCT	3	10	10,14	agree with lobstr
chr6	16327864	16327955	Spinocerebellar ataxia 1 (SCA1)	no call									CTG	3	29	?	no cross over, good coverage, could benefit from longer read
chr6	170870994	170871105	Spinocerebellar ataxia 17 (SCA17)	no call									CAG	3	38	?	no cross over, good coverage, could benefit from longer read
chr9	71702832	71702873	Friedrich ataxia (FRDA)	AAG	3	13.7	0,0	26	26	0	0:26	1	CTT/CTC	3	14	14,?	good coverage, could be only one allele
chrX	146993040	146993079		AAACC	5	7.8	0,0	9	9	0	0:9	1					
chrX	146993554	146993629	Fragile-X site A (FRAXA)	no call									CGG	3	20	?	no cross over, poor coverage
chrX	147582124	147582273	Fragile X site E (FRAXE)	no call									GCC	3	15	15,?	ok coverage, may have another allele
chrX	66765158	66765261	Kennedy disease	no call									CAG	3	22	20,24	lobstr misses both alleles

Personalis EX2075

Longer Read Length Increases the Possibility of Detection

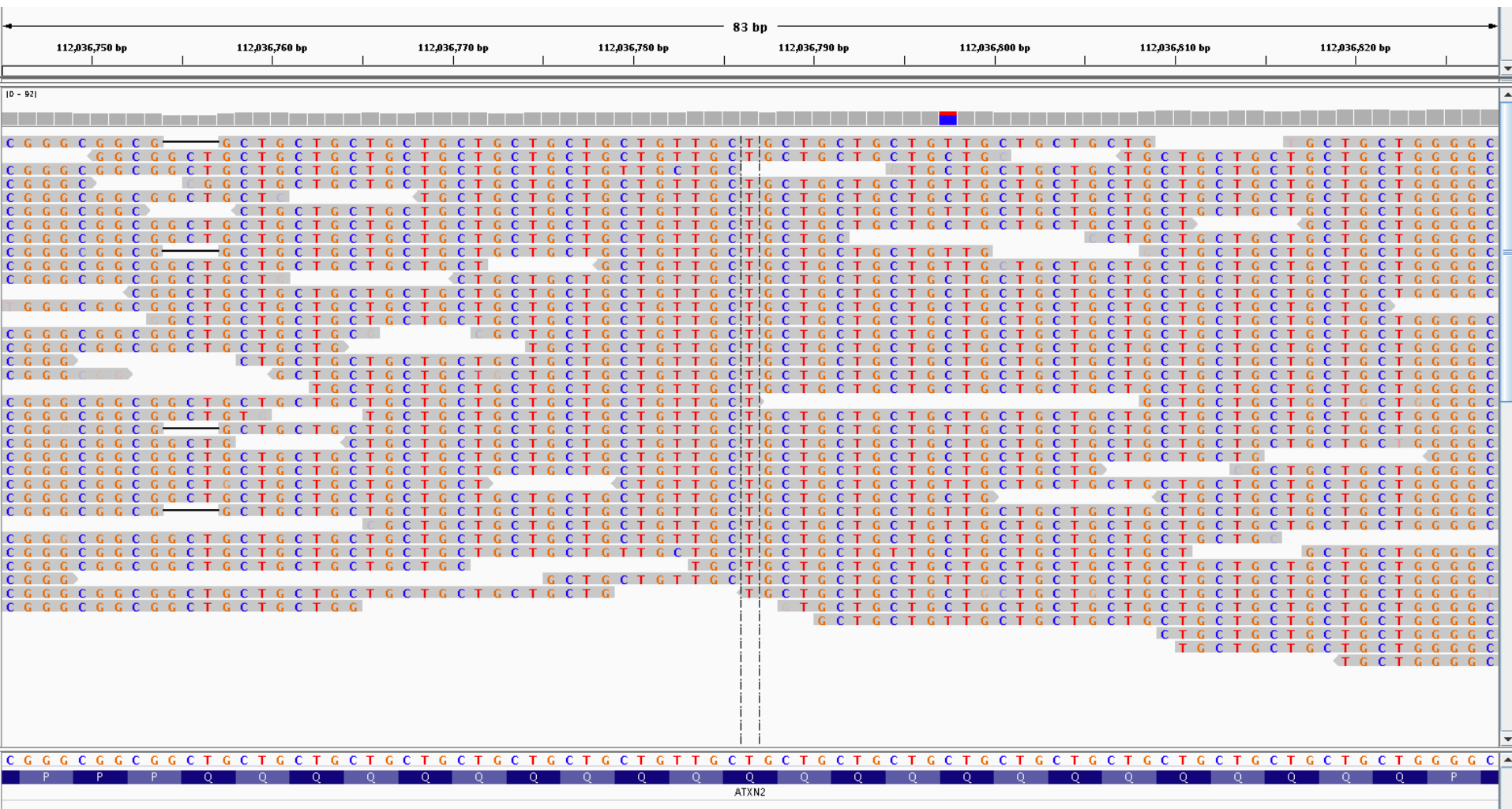
Chr	Disease	Coord	Gene	region	Description	read length for stable	*aligned frag allowance for 100bp	*aligned frag allowance for 150bp	*aligned frag allowance for 200bp	P(N>2,RL=100) at 30x	P(N>2,RL=150) at 30x	P(N>2,RL=200) at 30x
chr12	Spinocerebellar ataxia 2 (SCA2)	12q24	ATXN2	coding	(CAG)n (stable 15-24)	72	25	75	125	0.9953	1.000	1.000
chr12	Dentatorubral-pallidoluysian atrophy (DRPLA)	12p13.31	DRPLA	coding	(CAG)n (stable 7-34)	102	-5	45	95	NaN	1.0000	1.000
chr13	Spinocerebellar ataxia 8 (SCA8)	13q21	?	UTR	(CTG)n (stable 16-34)	102	-5	45	95	NaN	1.0000	1.000
chr14	Machado-Joseph disease (SCA3)	14q32.1	ATXN3	coding	(CAG)n (stable 13-36)	108	-11	39	89	NaN	0.9999	1.000
chr16	Huntington disease-like 2 (HDL20)	16Q24.3	JPH3	variably spliced exon	(CTG)n (stable 7-28)	84	13	63	113	0.9008	1.0000	1.000
chr19	Myotonic dystrophy 1 (DM1)	19q13	DMPK	3' UTR	(CTG)n (stable 5-37)	111	-14	36	86	NaN	0.9988	1.000
chr19	Spinocerebellar ataxia 6 (SCA6)	19p13	CACNA1A	coding	(CAG)n (stable 4-17)	51	46	96	146	1.000	1.0000	1.000
chr21	Myoclonic epilepsy	21q22.3	CSTB	promotor	(CCCCCGCCCCGCG)n (stable 2-3)	39	58	108	158	1.0000	1.0000	1.000
chr22	Spinocerebellar ataxia 10 (SCA10)	22q13.31	ATXN10	intron 9	(ATTCT)n (stable 10-20)	100	-3	47	97	NaN	1.0000	1.000
chr3	Myotonic dystrophy 2 (DM2)	3q21.3	ZNF9	intron 1	(CCTG)n (stable 10-26)	104	-7	43	93	NaN	1.0000	1.000
chr3	Spinocerebellar ataxia 7 (SCA7)	3p14.1	ATXN7	coding	(CAG)n (stable 4-35)	105	-8	42	92	NaN	1.0000	1.000
chr4	Huntington disease	4p16.3	HD (HTT)	coding	(CAG)n (stable 6-34)	102	-5	45	95	NaN	0.9999	1.000
chr5	Spinocerebellar ataxia 12 (SCA12)	5q32	PPP2R2B	promotor	(CAG)n (stable 7-45)	135	-38	12	62	NaN	0.8743	1.000
chr6	Spinocerebellar ataxia 1 (SCA1)	6p23	ATXN1	coding	(CAG)n (stable 6-38)	114	-17	33	83	NaN	0.9995	1.000
chr6	Spinocerebellar ataxia 17 (SCA17)	6q27	TBP	coding	(CAG)n (stable 24-42)	126	-29	21	71	NaN	0.9866	1.000
chr9	Friedrich ataxia (FRDA)	9q13	FXN	intron 1	(GAA)n (stable 6-32)	96	1	51	101	0.0369	1.0000	1.000
chrX	Fragile-X site A (FRAXA)	Xq27.3	FMR1	5' UTR	(CGG)n (stable 6-54)	162	-65	-15	45	NaN	NaN	1.000
chrX	Fragile X site E (FRAXE)	Xq28	FMR2	5' UTR	(CCG)n (stable 4-39)	117	-20	30	80	NaN	0.9988	1.000
chrX	Kennedy disease	Xq12	AR	coding	(CAG)n (stable 9-35)	105	-8	42	92	NaN	1.0000	1.000

*: Minimal 3bp flanking

Still need to include sampling for det

Personalis EX2075

Personalis EX2075



Chr12:SCA2

Personalis EX2075



Chr13:SCA8

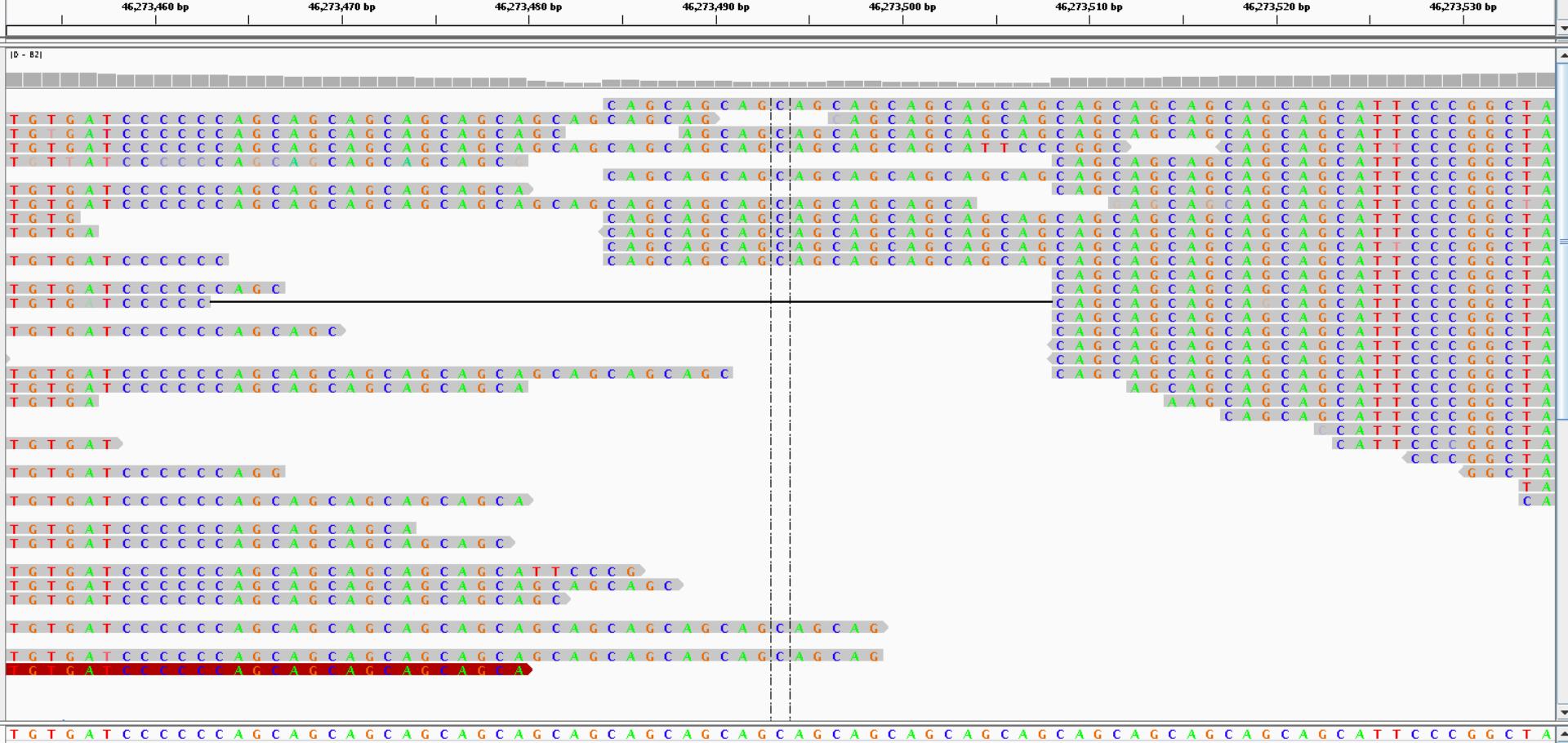
Personalis EX2075



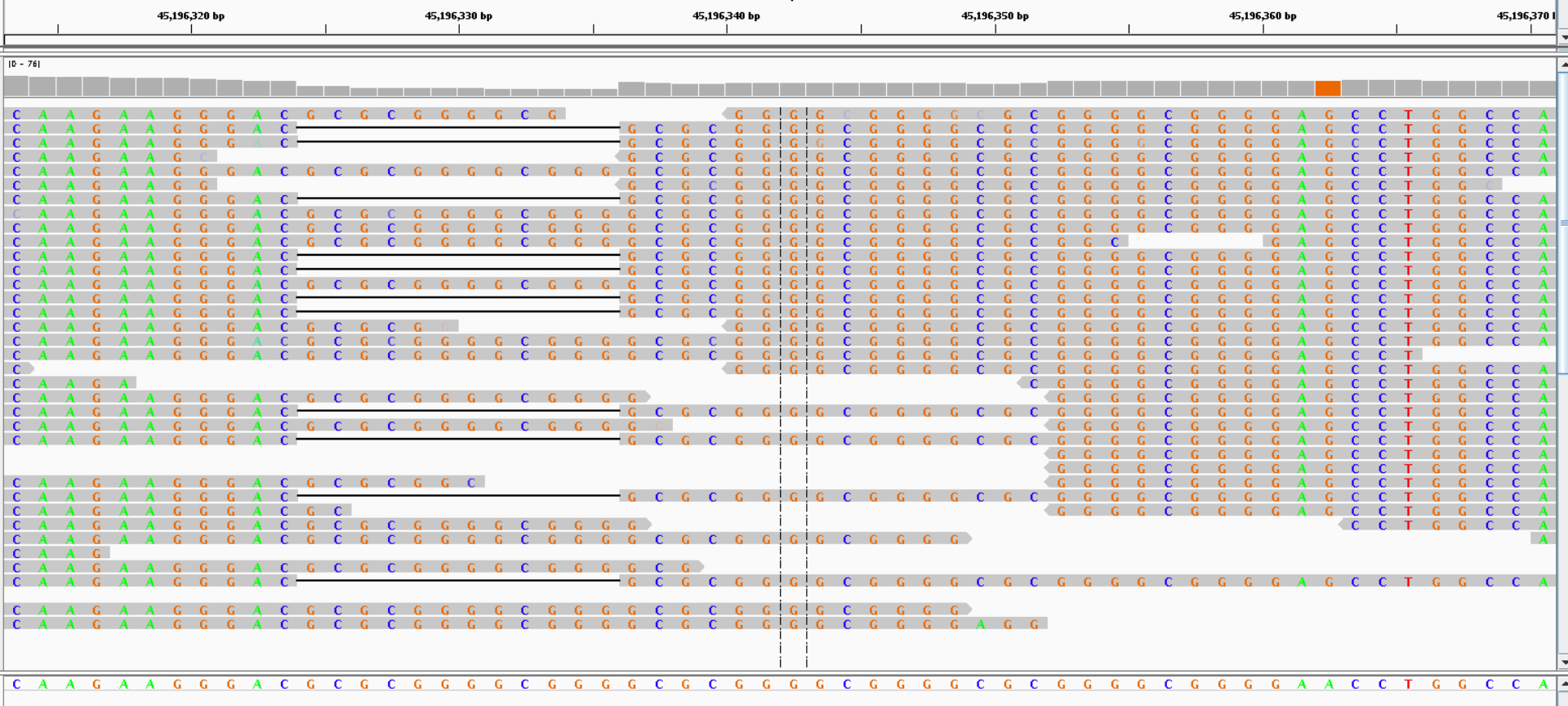
Chr14:SCR3

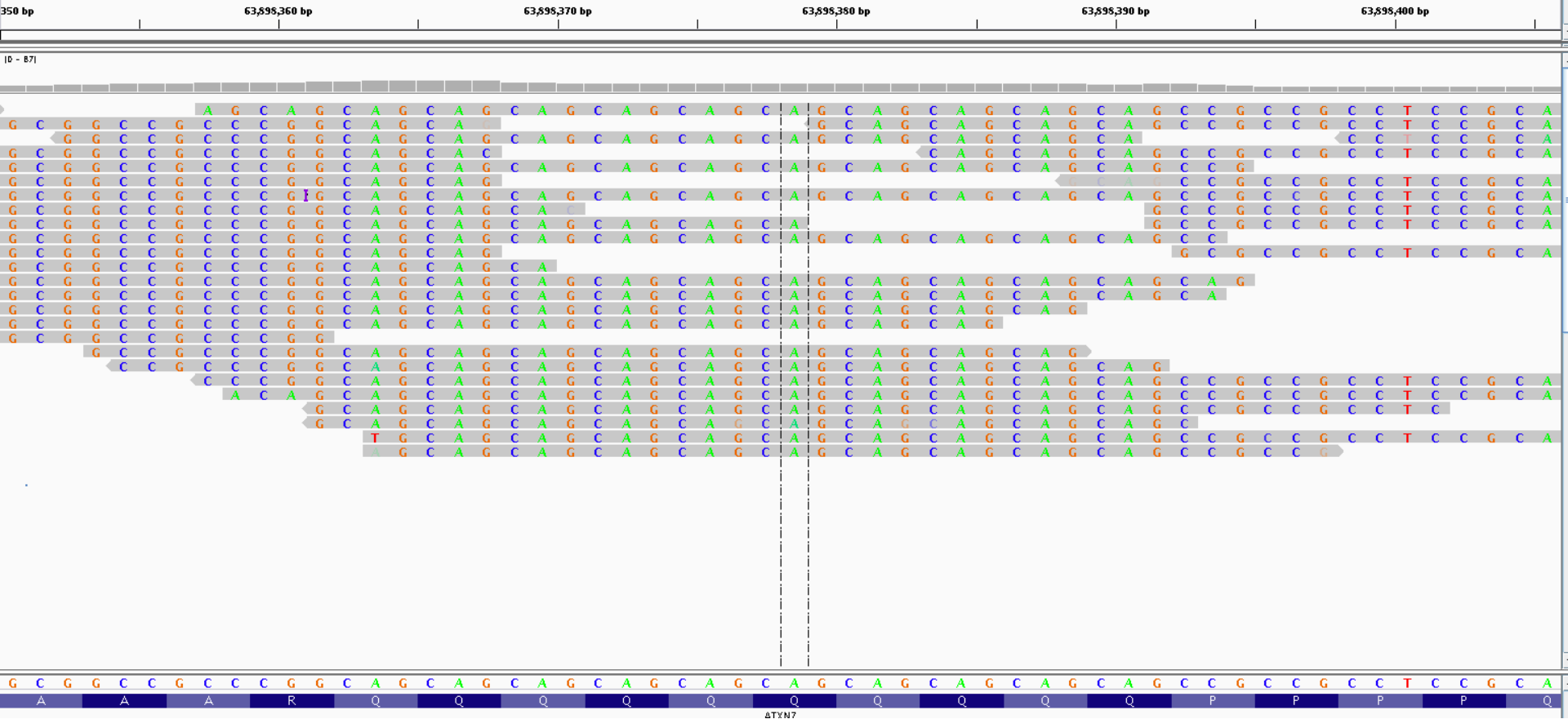
Personalis EX2075

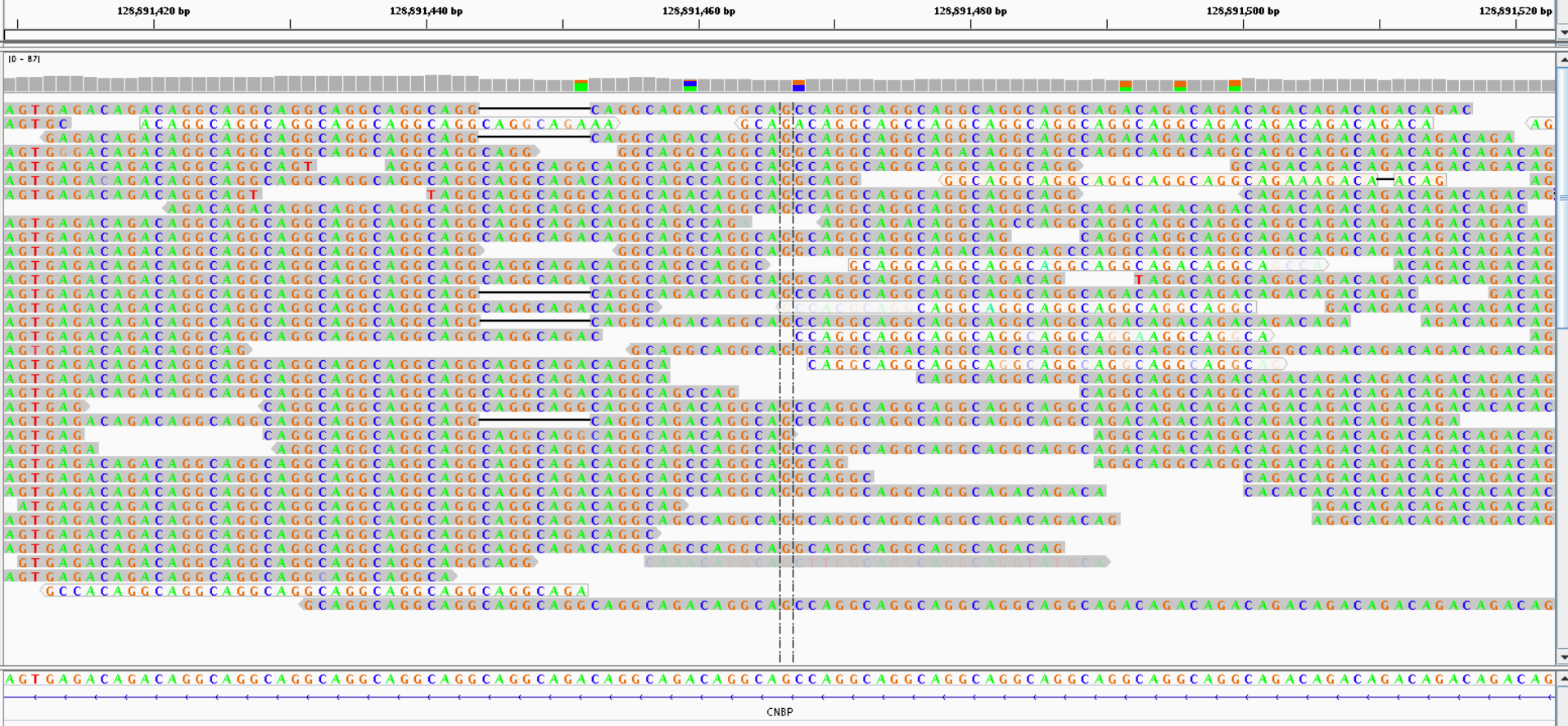
Personalis EX2075



Chr19:DM1

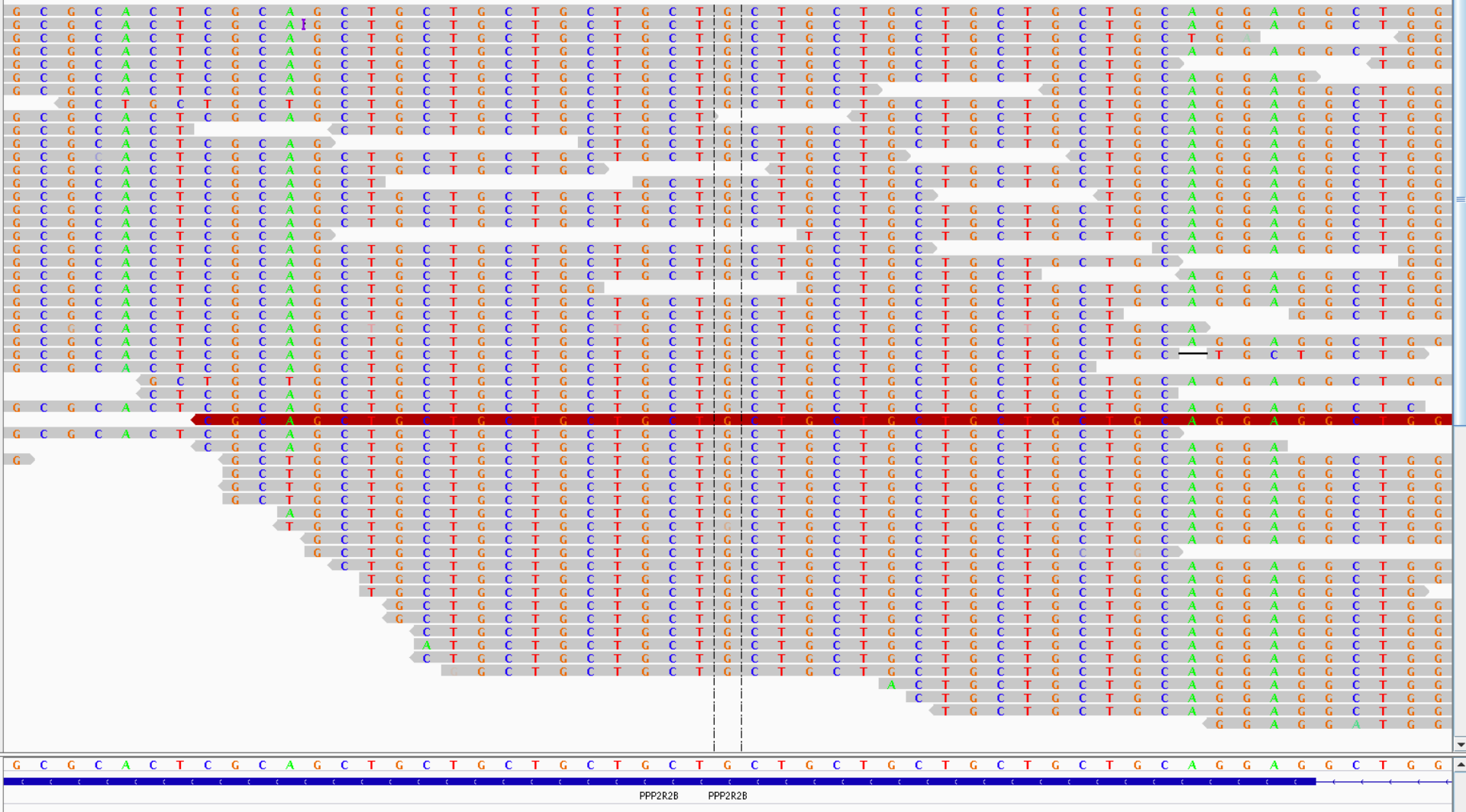






Chr3:DM2

Personalis EX2075

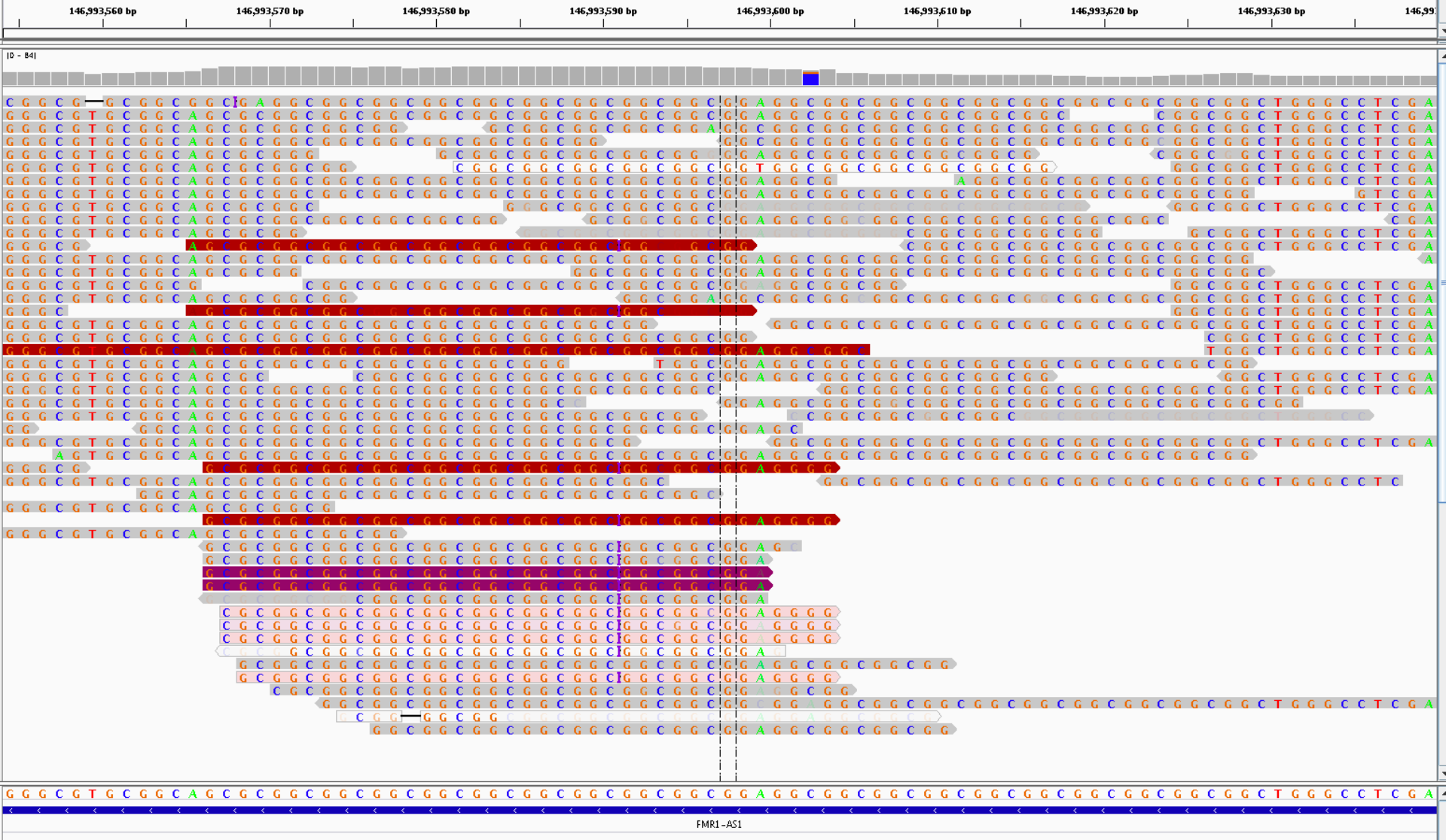


Chr5:SCA12

Personalis EX2075

Personalis EX2075





Personalis EX2075

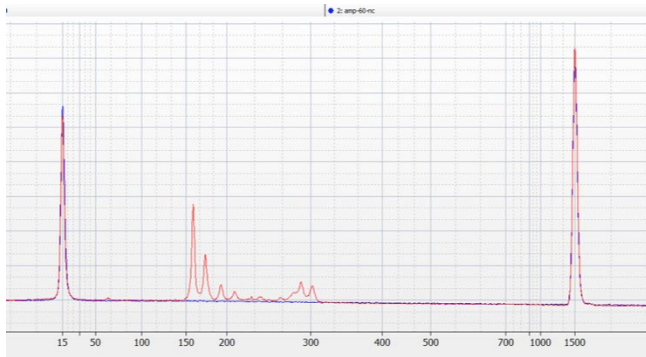
Conclusions

- Newer chemistry appears to be able to read through most of these repeats (unlike older chemistry). The problem is getting enough sampling of spanning reads since these regions are approaching current read lengths.
- With 150bp or 200bp, we should be able to detect stable alleles for all the disease regions from one end to the other end.
- FMR1 is so far the most problematic one, with poor coverage (mostly truncated reads).
- Exploring MiSeq long reads but this will only be commercially feasible if we can make a targeted long-read library or perform on whole genome with longer HiSeq reads.

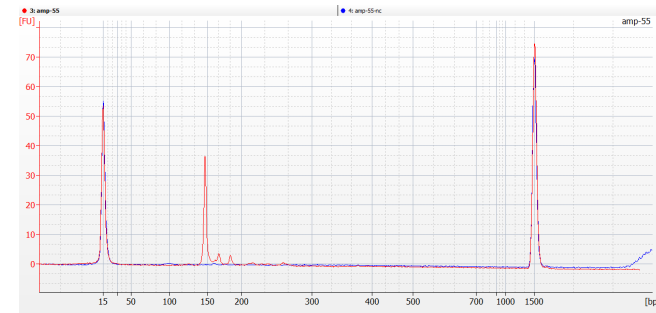
HTT(Huntington) PCR

Primer 2

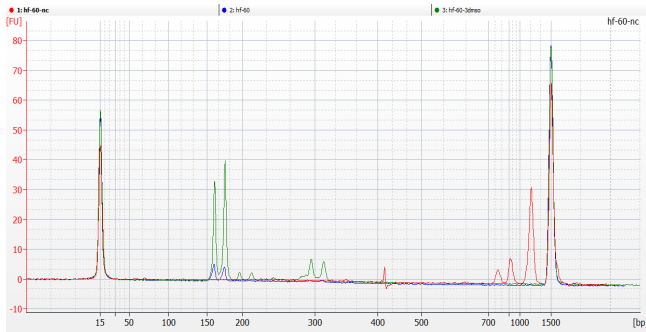
Primer 1

[illegible]

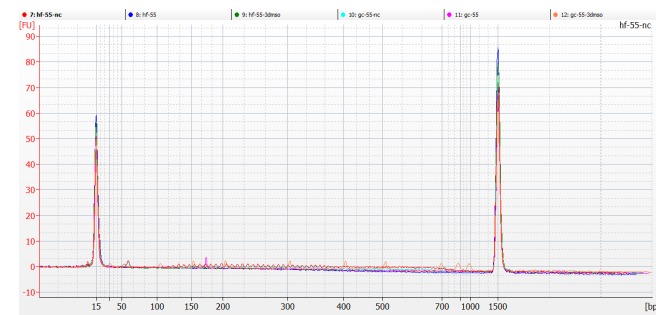
Primer1
taq



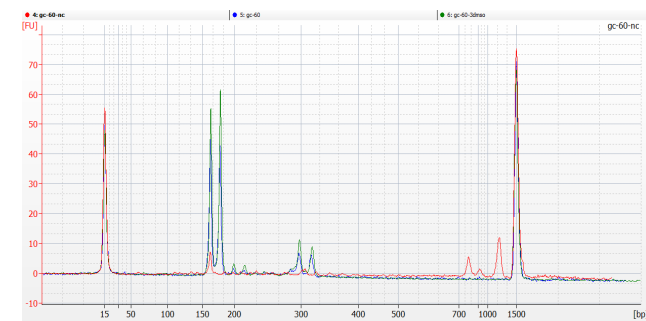
Primer2
taq



Primer1
Phu-hf



Primer2
Phu-hf-gc



Primer1
Phu-gc

Personalis EX2075

FMR1 (fragile X) PCR

GCGGGCCGGGGGTTTCGGCCTCAGTCAGGCGCTCAGCTCCGTTTCGGTTTCACTTCCGGTGGAGGGCCGCCTCTGAGCGGGCGGCGGGCCGACGGC
GAGCGCGGGCGGCGGCGGTGACGGAGGCGCGCTGCCAGGGGGCGTGCGGCAGCGCGGCGGCGGCGGCGGCGGCGGCGGCGGAGGCGG
CGGCGGCGGCGGCGGCGGCGGCGGCTGGGCCTCGAGCGCCGCAGCCACCTCTCGGGGCGGGCTCCCGGCGCTAGCAGGGCTGAAGAGAAGA
TGGAGGAGCTGGTGGTGGAAGTGCGGGGCTCCAATGGCGCTTCTACAAGGTACTTGGCTCTAGGGCAGGCCCATCTTCGCCCTTCCTCCCTCCCT
TT

Red and underlined primers failed in both Taq and Phusion, will try blue primer with betaine.

Difficult Motifs

CTTAAAAAAAAAAAAATCCAGC
CGAGAATTTTTTTTTTTTTTAGG

CTTAAAAAAAAAAAAAAAAAAAAAAAAATCCAGC
CGAGAATTTTTTTTTTTTTTTTTTTTTTTTAGG

CTTGGGGGGGGGGGGGTCCAGC
CGAGAACCCCCCCCCCCCCAGG

TGGTTAAGCAACACAGAAGGGT CTTAAAAAAAAAAAAATCCT AACTAGGCATATGTGCATCTGGAGC
CGAACCAATTCGTTGTGTCTTC CCAGAATTTTTTTTTTTTTTAGGATTG ATCCGTATACACGTAGACC

TGGTTAAGCAACACAGAAGGGT CTTAAAAAAAAAAAAAAAAAAAAAAAAATCCT AACTAGGCATATGTGCATCTGGAGC
CGAACCAATTCGTTGTGTCTTC CCAGAATTTTTTTTTTTTTTTTTTTTTTTTTTTTAGGATTG ATCCGTATACACGTAGACC

TGGTTAAGCAACACAGAAGGGT CTTGGGGGGGGGGGGGTCCT AACTAGGCATATGTGCATCTGGAGC
CGAACCAATTCGTTGTGTCTTC CCAGAACCCCCCCCCCCCCAGGATTG ATCCGTATACACGTAGACC

CTTACACACACACACACACACACACACACACACACA CACACACACACACACACACACACACACATCCAGC
CGAGAATGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG TGTGTGTGTGTGTGTGTGTGTGTGTGTAGG

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCG
GTTTCGTCTTCTGCCGTATGCTCTAGC[aaagtg](#)CACTGACCTCAAGTCTGCACA**CGAGAAGGCTAG**

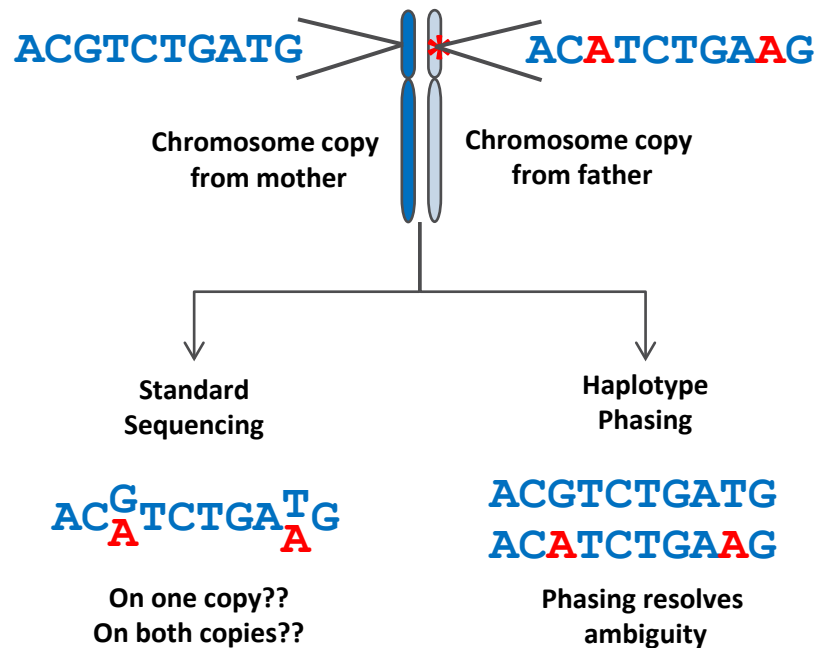
Long Range Haplotyping using Dilution and Amplification

Shujun Luo

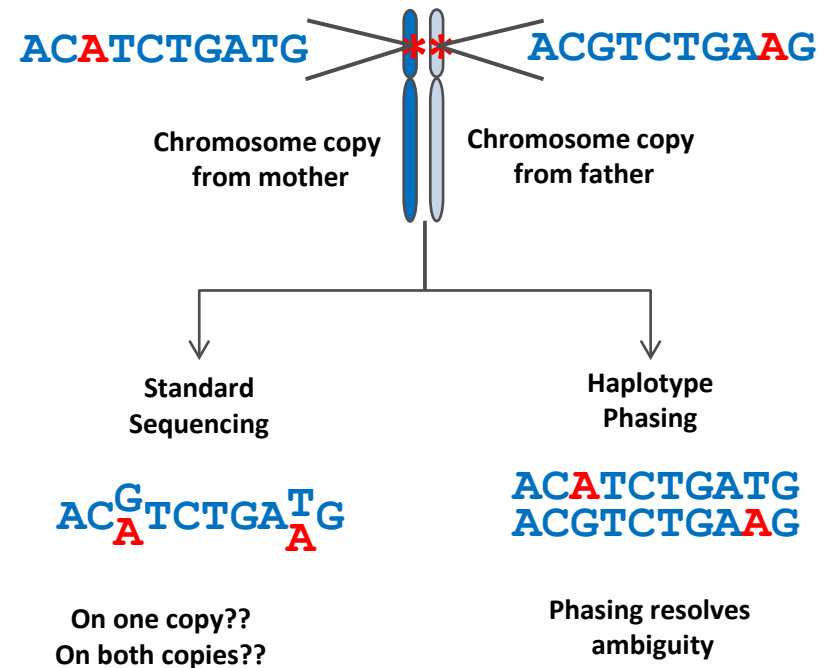
Mark Pratt

Why haplotype – do we need to identify case I vs case II?

Case I: Single copy of Gene carries “risk” alleles



Case II: Both copies of Gene carry “risk” alleles



Incidence of Multi-SNP Genes

- NA12878 A/B Nonsynonymous SNPs:
 - 1,024 RefSeq Genes
 - 232 Mendelian Tier 1 Genes
 - 834 Mendelian Tier 2 Genes
 - 9 PharmGKB VIP Genes
- PharmGKB – 14 Genes, 27 Transcripts
- Regulome1 - 6,077 Genes, 9,253 Transcripts
- Regulome2 - 19,292 Genes, 26,439 Transcripts
- Varimed – 21 Genes, 35 Transcripts

Includes Alt/Alt but shouldn't

Potential methods for phasing/haplotyping

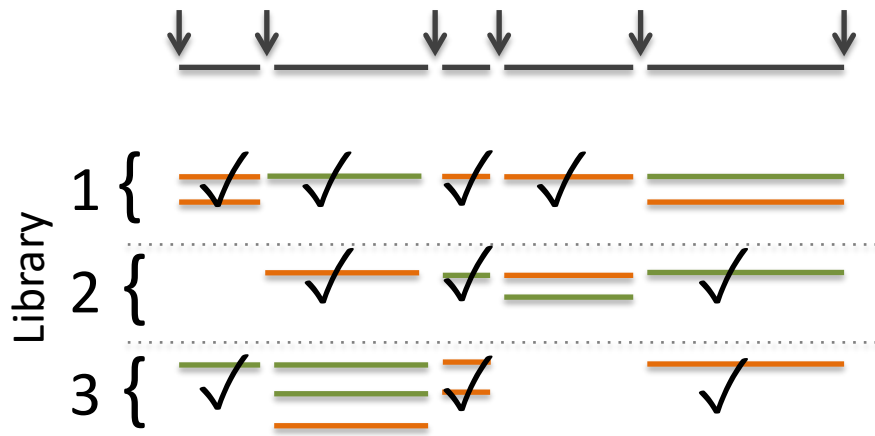
- ▶ Statistical imputation and phasing
 - Population-based linkage disequilibrium and pedigree analysis
 - Former unlikely to be accepted by researchers / FDA
 - Later unlikely to be commercially viable
- ▶ Sequencing of large insert fosmid clone pools
 - Very labor intensive
- ▶ Somatic cell hybrid construction to convert a diploid cell into haploid cell lines
 - Very labor intensive
- ▶ Isolation of individual chromosomes through chromosome microdissection or microfluidic devices
 - Requires high level expertise
- ▶ Very long reads – e.g. ONT
 - Accuracy and cost are big factors
- ▶ Read pairs
 - Inefficient unless the read length increases drastically (while not compromising error)

High concentration/Low concentration dilution

- ▶ High concentration dilution:
 - ~1x coverage of genome per reaction
 - Select haplotype within defined RE fragments reducing likelihood of mislinking separate fragments
 - Require smaller number of libraries (~10) to sample entire genome
- ▶ Low Concentration dilution:
 - ~0.1x coverage of genome per reaction
 - Likely to have SD/compression regions in different reactions allowing for separate alignment or assembly
 - Requires larger numbers of libraries (~100) to cover genome

Restriction Fragment Libraries (Mark to edit)

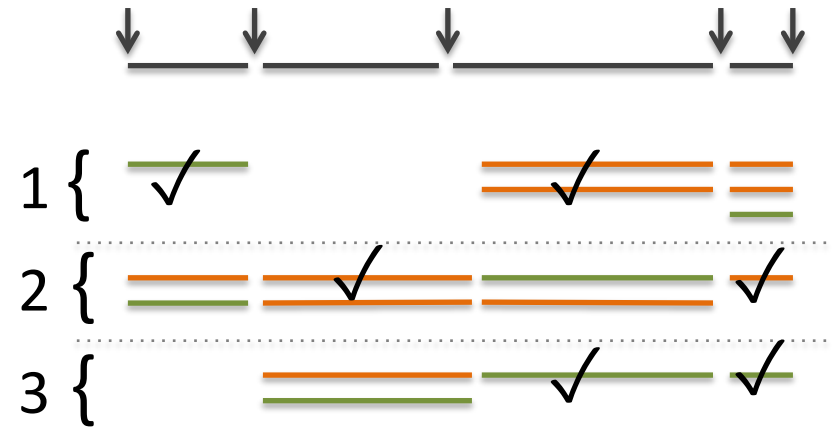
Restriction A libraries



Phased A Segments



Restriction B libraries



Phased B Segments

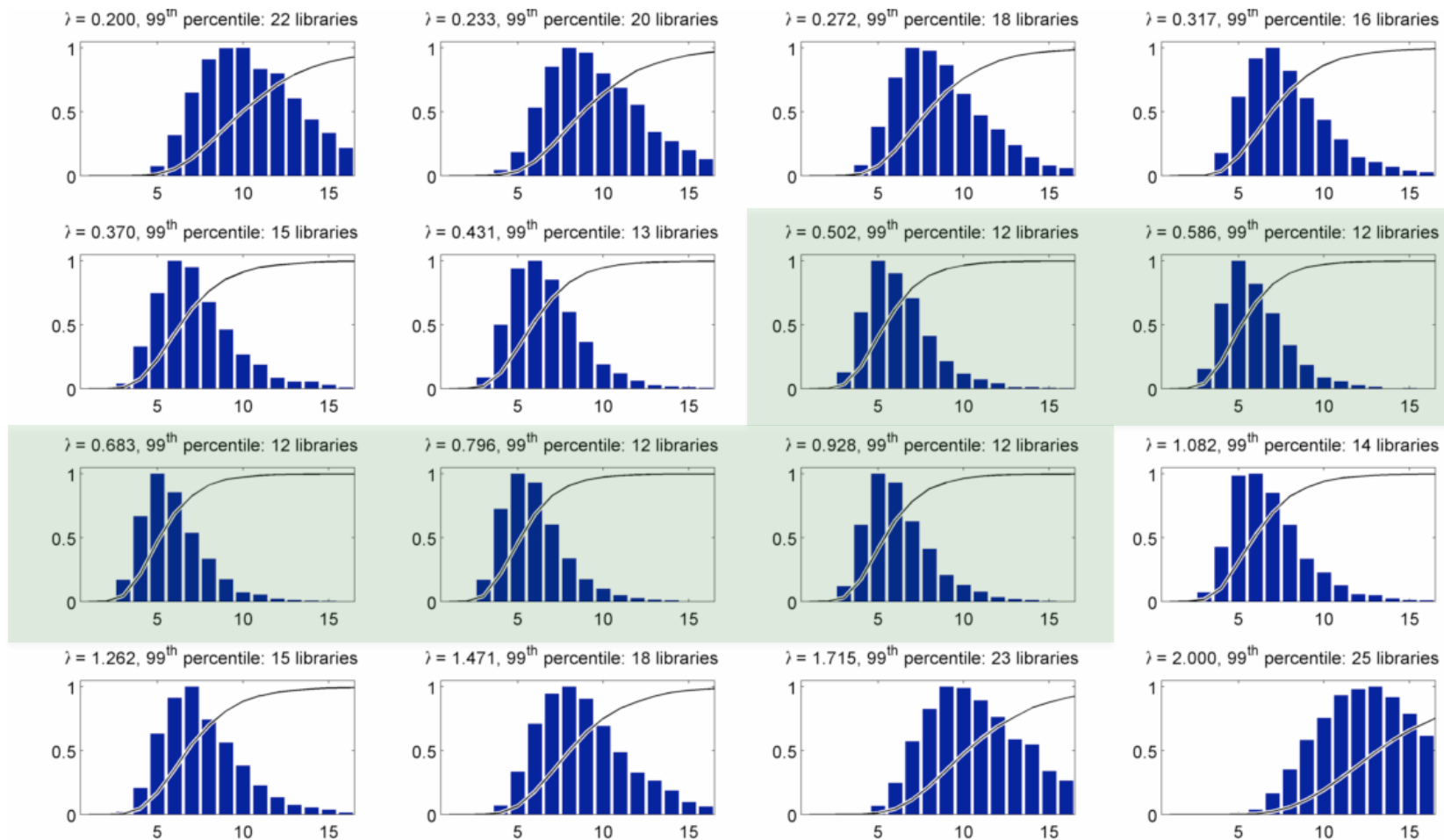


Linked Segments



Single Haplotype: ✓

Concentration and library multiplicity to achieve 99% fragment phasing



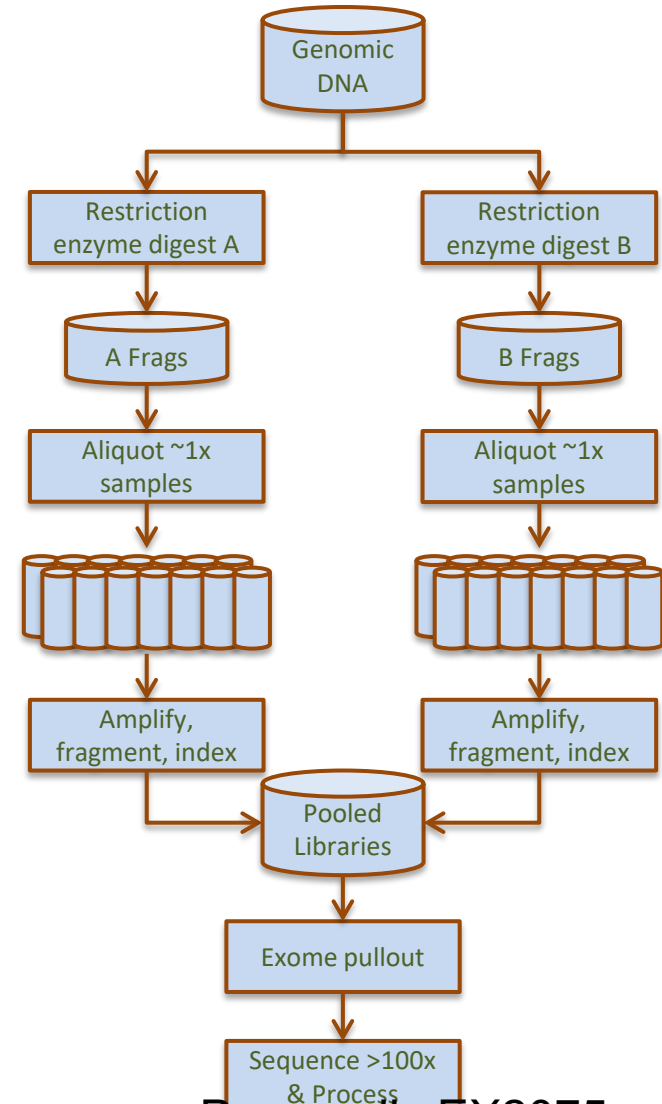
At concentrations 0.5-1.0x “coverage” it will take approximately 12 libraries to see 99% of fragments in a single molecule situation.

Personalis EX2075

Proposed high concentration Phasing Sample Preparation

1. Extract long fragment, whole genome DNA
2. Process 2 aliquots using restriction enzyme A and B to fragment.
3. Further aliquot into ~12 libraries, each with a target density such that each fragment is represented at ~1x-fold
4. Prepare sequencing libraries (amplify, fragment, index)
5. Pool all libraries and pullout exome
6. Sequence at ~5-10x per library, 100-200x net (ensure sensitivity for zygosity detection)
7. Process with normal exome pipeline
8. Custom post-process to determine phasing
 - a. Identify restriction spans with strong evidence of single haplotype segments (requires modest sensitivity at several heterozygous sites)
 - b. Link alleles on identified single haplotype segments
 - c. Link haplotype restriction spans using alternating (A,B) restriction spans

Restriction cutting produces fixed endpoints to fragments and allows for much higher concentration and therefore fewer libraries because this method is much less likely to inadvertently link copies by chance.



Proposed Proof of concept:

- ~1Mb Duchenne Muscular Dystrophy (DMD) region on X chromosome
 - Two male Coriell DNAs (NA18507 + HG10377) were first sequenced to identify true haplotypes. Samples are mixed to simulate a diploid “XX” genome then sequenced to identify SNPs and sequenced using the dilution-amplification method
- 4.6Mb Human Leukocyte Antigen (HLA, Chr6) region
 - Genomic DNA from COX and PGF cell lines, both of which are homozygous across the HLA locus due to identity-by-descent, mixed at 1:1 ratio to simulate a heterozygous HLA locus and sequenced using the dilution-amplification method
- Amy
 - Locate Broad, Yale, OpGen results for comparison

Data from CGI (Low concentration)

- Starting material: 10-20 cells worth of DNA -- isolate long DNA 50-500kb DNA
- 384 dilutions
 - only 10% of the genome should be represented in each aliquot
 - ~30 diluted DNA copies – probably >150X sequencing
- Haplotyped 91~97% of heterozygote SNPs
- **1 in 10 MB error** (by filtering on haplotype)

Cost

High concentration direct sequencing:	\$9,200
Library: \$50 (cost/library) x 12 (dilution) x 2 (enzyme)	\$1,200
Sequencing (HiSeq)	\$8,000
High concentration exome pull-down sequencing:	\$3,300
Library: \$50 (cost/library) x 12 (dilution) x 2 (enzyme) + \$100 (pull down)	\$1,300
Sequencing (MiSeq)	\$2,000
Low concentration direct sequencing	\$12,800
Library: \$50 (cost/library) x 96 (dilution)	\$4,800
Sequencing (HiSeq)	\$8,000
Low concentration exome pull-down sequencing	\$6,900
Library: \$50 (cost/library) x 96 (dilution) + \$100 (pull down)	\$4,900
Sequencing (MiSeq)	\$2,000

Potential Risks

- Currently the MDA amplification step appears to have extreme biases
- Haplotype switching errors
- Requires substantial bioinformatics support
 - assign zygosity by bin and library
 - link phased segments
 - develop output format, etc.

RFLP-Seq

Shujun Luo

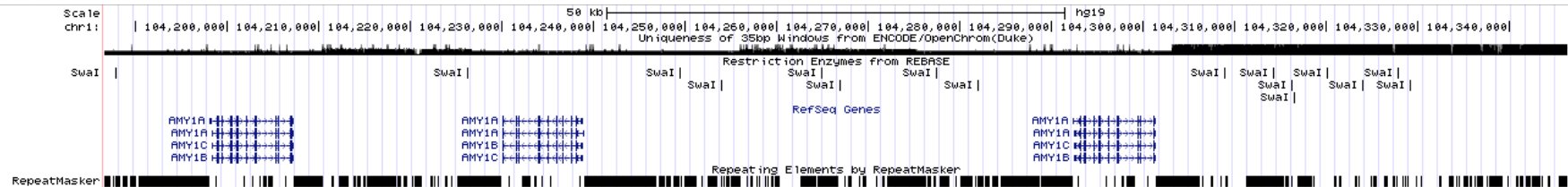
Mark Pratt

John West

Goal

- To resolve SD by physically separating DNA fragments, assuming each copy of the SD falls into different bins (by length of restriction enzyme digested fragments).

Example---3 copies of Amylase

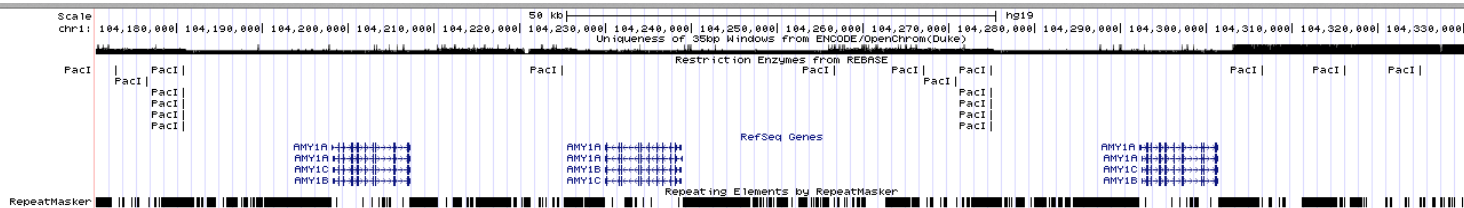


SwaI

38383 bp

23266 bp

27019 bp

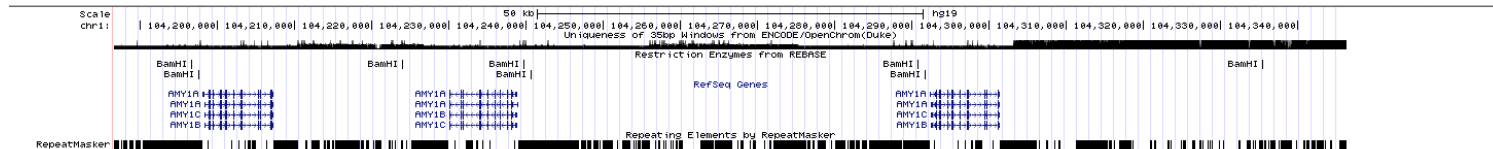


PacI

44105 bp

31780 bp

31637 bp



BamHI

26399 bp

15698 bp

43676 bp

Of the 3 enzymes showing here, BamHI is best at distinguishing the 3 copies; Swal is ok if size selection is controlled very well; PacI can't distinguish rep2 and rep3.

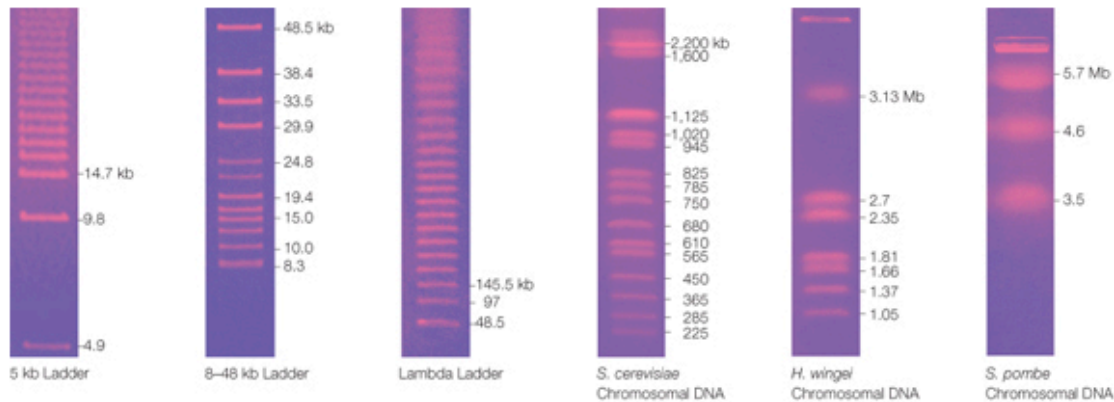
It would need combination of different enzymes to capture different SD, and still it is possible there some SDs can't be distinguished with the chosen enzymes.

Personalis EX2075

CHEF Mapper

PFGE Standards Selection Guide

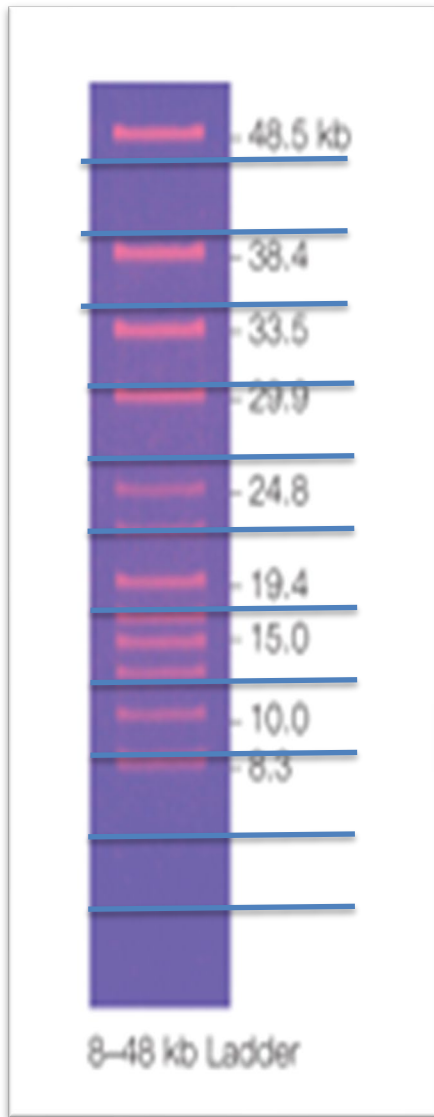
[Close]



Comparison of Pulsed Field Standards

	Range	Contents	Amount	Number of Applications
5 kb ladder	4.9–120 kb	Concatemers of pBR328	20 µg in 200 µl	20–25
8–48 kb ladder	8.3–48.5 kb	Mixed digest of phage λ	25 µg in 125 µl	125
Lambda(λ) ladder	0.05–1 Mb	Concatemers of phage λcl857Sam7	5 agarose blocks	25–40
<i>S. cerevisiae</i>	0.225–2.2 Mb	<i>Saccharomyces cerevisiae</i> chromosomal DNA	5 agarose blocks	25–40
<i>H. wingei</i>	1–3.1 Mb	<i>Hansenula wingei</i> chromosomal DNA	5 agarose blocks	25–40
<i>S. pombe</i>	3.5–5.7 Mb	<i>Schizosaccharomyces pombe</i> chromosomal DNA	5 agarose blocks	25–40

Project Planning



Total Libraries = # section x # enzyme

Cost

CHEF Mapper System	\$30,000
Direct sequencing	\$10,160~13,760
Library: \$60 (cost/library) x 12 (gel sections) x 3~8 (enzyme)	\$2160~5760
Sequencing (HiSeq)	\$8,000
exome pull-down sequencing:	\$4260~7860
Library: \$60 (cost/library) x 12 (gel sections) x 3~8 (enzyme) + \$100 (pull down)	\$2260~5860
Sequencing (MiSeq)	\$2,000

Potential Risks

- Need large amount of starting material.
- Difficult to scale up.
- Useful for refining reference, however hard to apply to individual genome.
- Some SD may not be distinguished with the chosen enzymes.
- Requires significant bioinformatics support
 - Implement bin assignment prior probabilities in alignment or assembly

Summary of Lab Based Accuracy Options

	Random Error	Mapping Error	SD & CMP	SV	UNERD "VNTR"	phasing	Cost (R&D)	Cost (Prod)	Development Effort
Long Paired Reads <i>Improving mapping and phasing by using longer paired reads</i>		+	+	+	++	+	Low	0	- Experiment with long reads on MiSeq <u>Wait</u> for ILMN to produce HiSeq kits - Modest sample prep development for appropriate library sizing - Modest bioinformatics support for improved longer read alignments - High level of bioinformatics support to leverage for SD, compressions and SVs.
Fused Reads <i>Improving mapping and phasing by using longer reads</i>	+	+		+		-	Low	Low*	- Modest sample prep development for appropriate sizing - Modest bioinformatics support for longer single fragment alignment
RFLP binning <i>Resolve long range mapping degeneracies by size binning RE frags</i>		++	+++	+			High	High	- Substantial setup costs - Substantial lab effort per sample - High level of low level bioinformatics support to integrate into alignment and assembly (e.g. mapping priors)
Phasing by low concentration dilution	+	++	++	+		++	High	Med	- Substantial development on unbiased amplification - Substantial bioinformatics support to identify and link fragments - Work with outside vendor (e.g. Moleculo)
Phasing / RF binning at high concentration	+	+	+	+		++	High	Low	- Substantial development on unbiased amplification - Substantial bioinformatics support to identify and link fragments

* Opportunity cost relative to using longer inserts to link larger structures and loss of overlap region for coverage