The state of the art in genetic variant identification

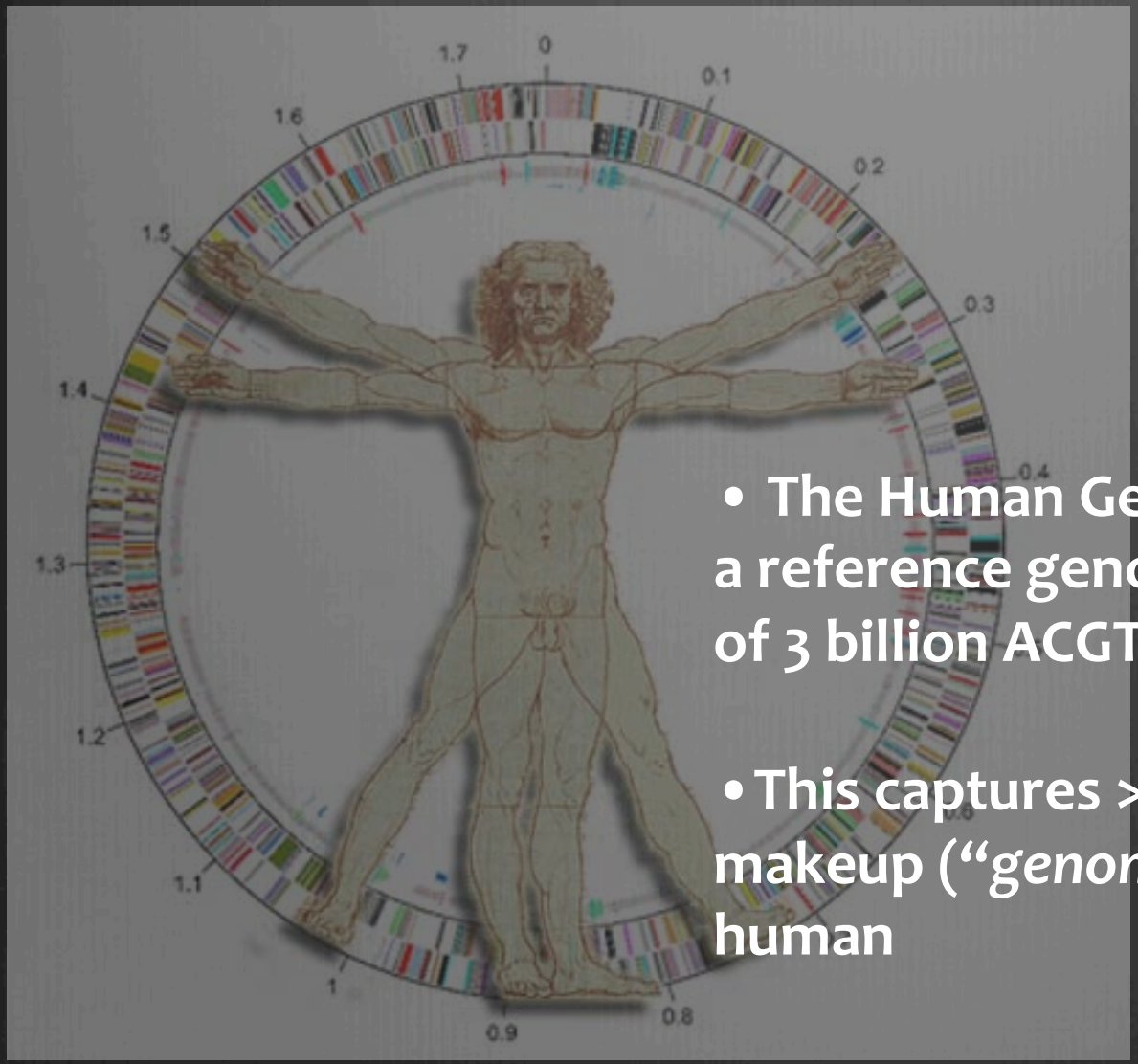
Andrew Farrell
USTAR Center for Genetic Discovery
Department of Human Genetics
University of Utah School of Medicine

CSHL Sequencing Course
2017

Planned new material/changes

- Add more detail/ de novo variant calling results to RUFUS
- More detail on Graphite
- Some detail on VG (ask Erik for a presentation for materials?)
- Add one slide about 10X data, and variant calling
- Add IOBIO content… Andrew's new variant caller?
- Add at least one slide about PacBio, Oxford Nanopore

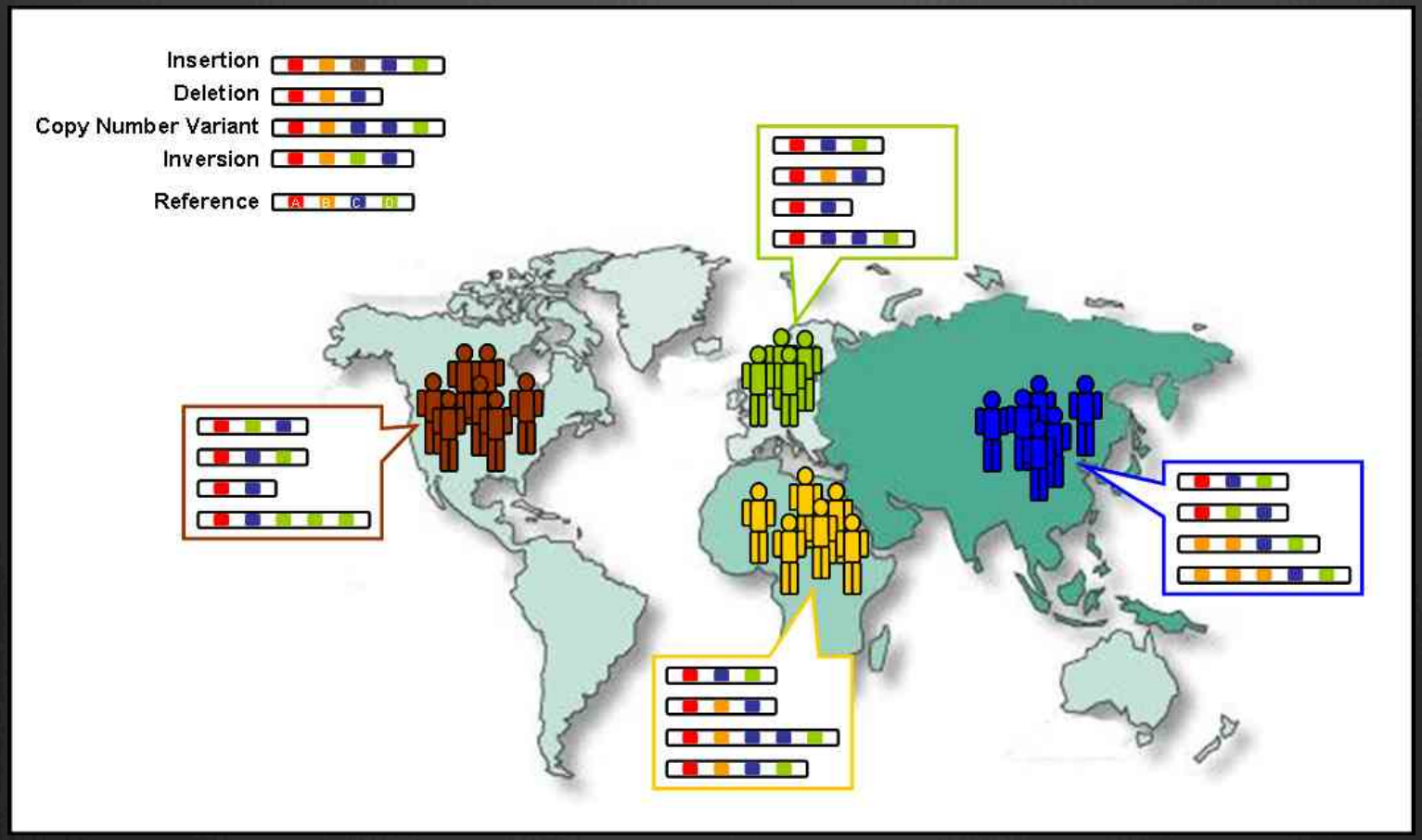
Genetic variations



- The Human Genome Project produced a reference genome sequence (a series of 3 billion ACGT nucleotides)

- This captures >99% of the genetic makeup ("*genome*") of every single human

Human genetic diversity is substantial

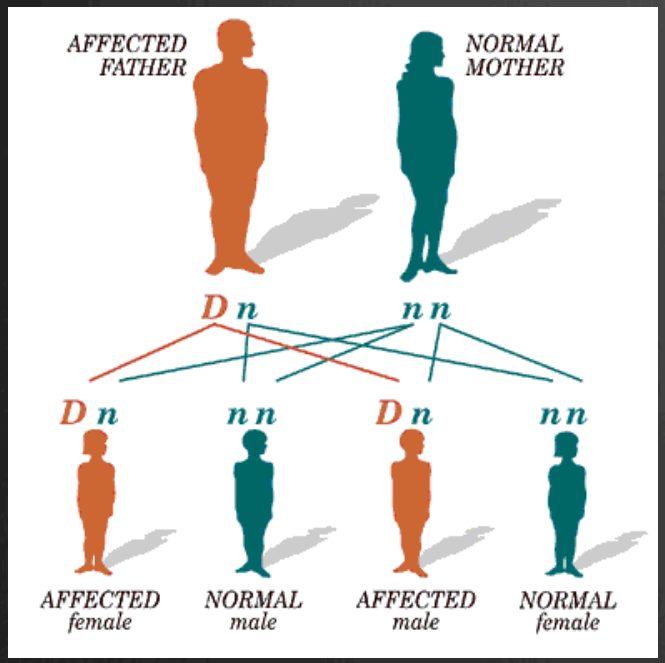


- Average heterozygosity (the rate of polymorphic differences between two human chromosomes) is ~10E-3 i.e. roughly 1 in 1,000 bp

Genetic variations are important

phenotypic differences

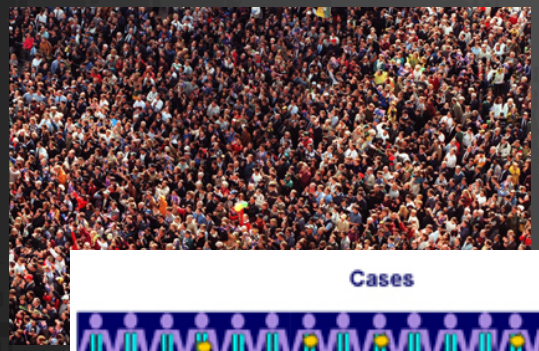


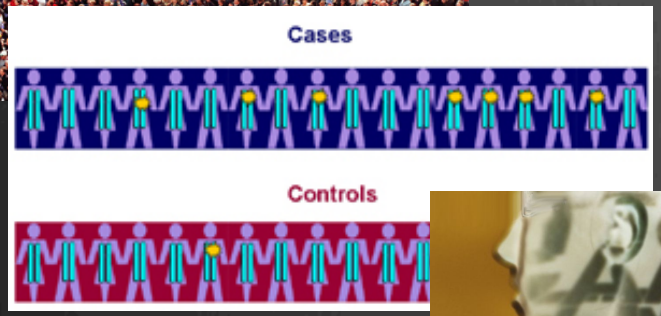
heritable diseases

demographic history

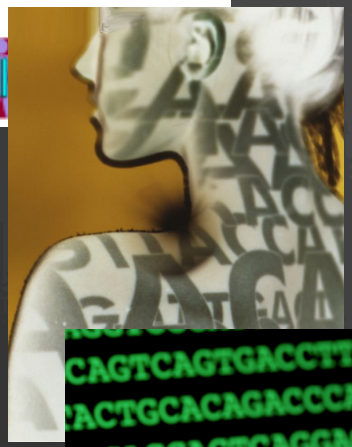
Variants impact many areas of science & medicine



Population genomics



Medical genomics



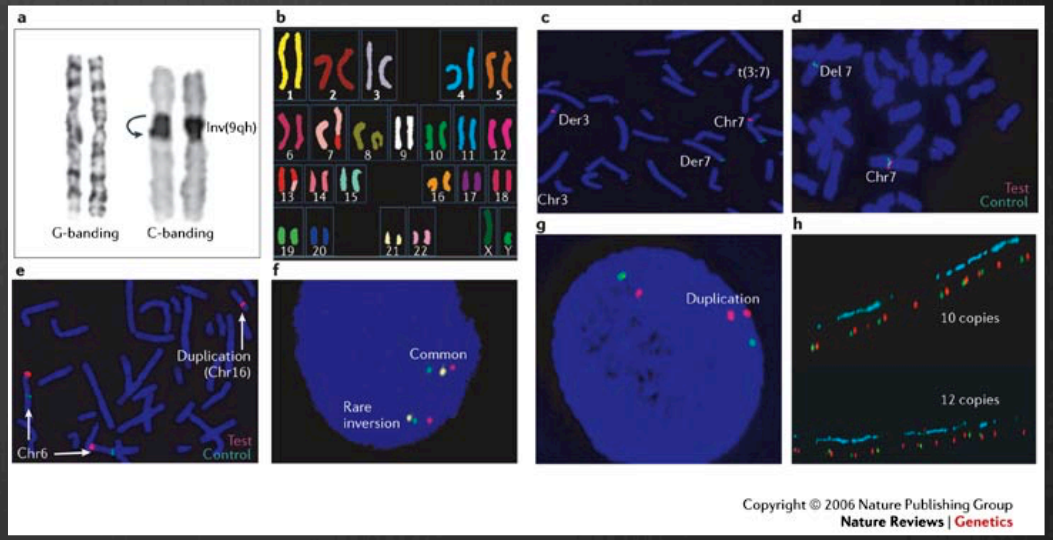
Personal genomics

Clinical genomics

Genetic variations come in different shapes & sizes

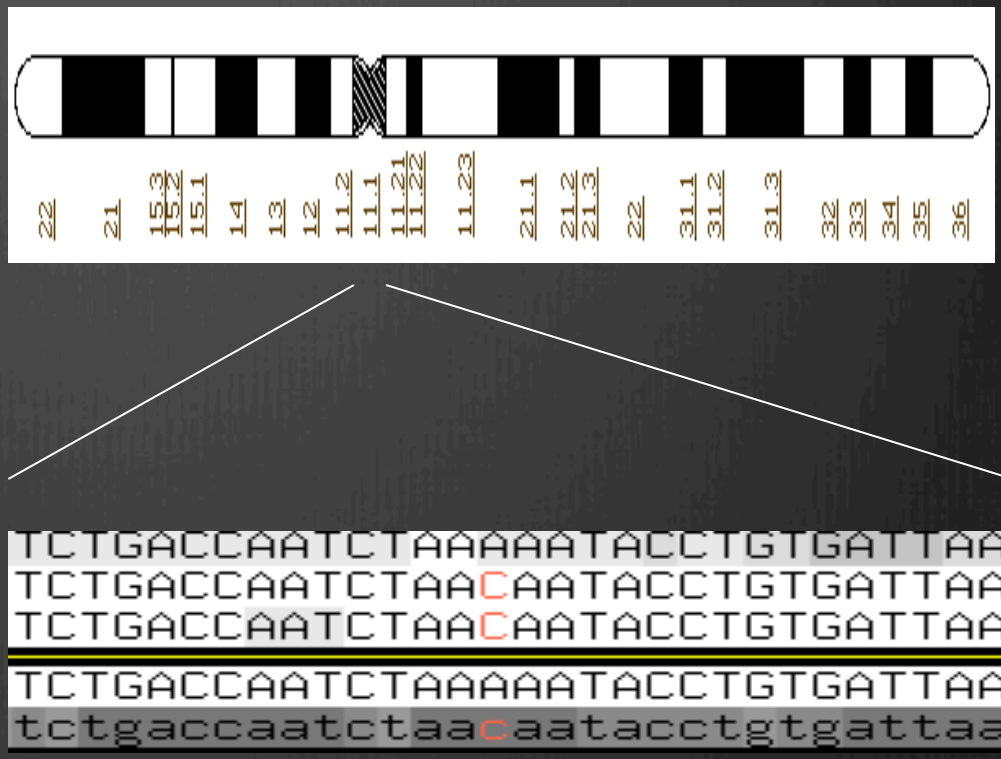
- Small (single-nucleotide) variants are most prevalent: ~4 million in each human genome!

- Larger "structural" variants are also abound



The fundamentals of variant discovery

Variant discovery requires comparative analysis of multiple sequences from the same region of the genome

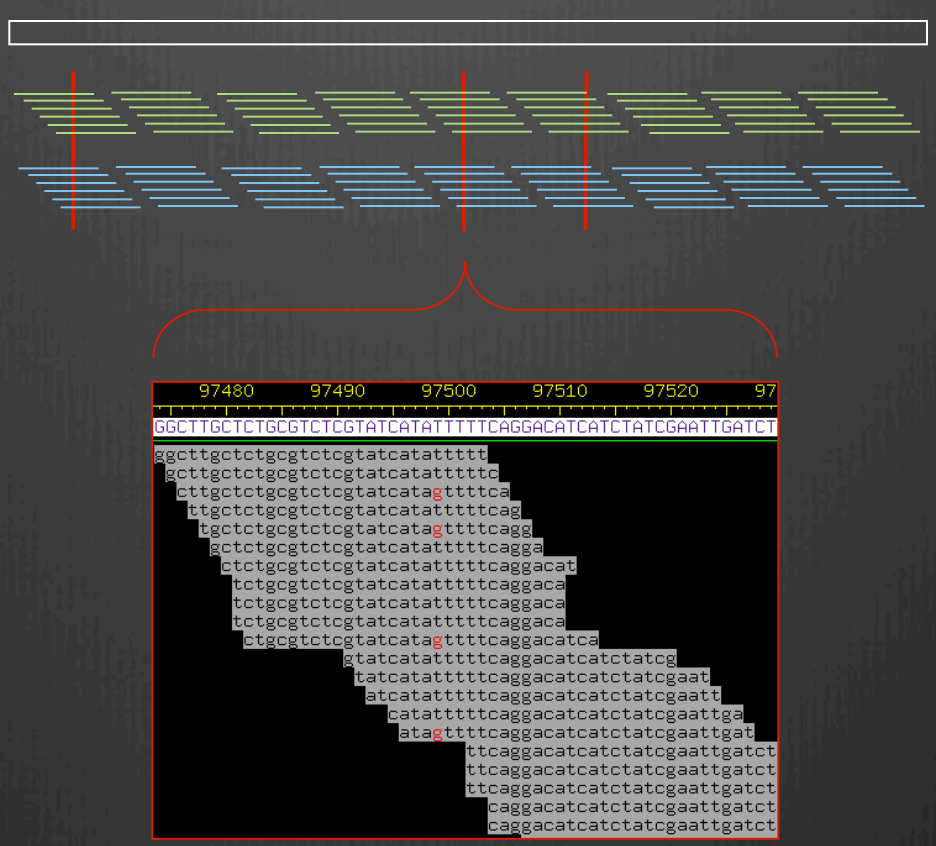


Reference-guided variant discovery

Variant discovery

reference sequence

sequencing reads



SNP calling


97490
97500
97510
975
TGCGTCTCGTATCATATTTTTCAGGACATCATCTATCC
tgcgtctcgtatcatattttt
tgcgtctcgtatcatatttttc
tgcgtctcgtatcatagttttca
tgcgtctcgtatcatatttttcag
tgcgtctcgtatcatagttttcagg
tgcgtctcgtatcatatttttcagga
tgcgtctcgtatcatatttttcaggacat
tgcgtctcgtatcatatttttcaggaca
tgcgtctcgtatcatatttttcaggaca
tgcgtctcgtatcatatttttcaggaca
tgcgtctcgtatcatagttttcaggacatca
gtatcatatttttcaggacatcatctatcg

SNP calling: what goes into it?

Base qualities

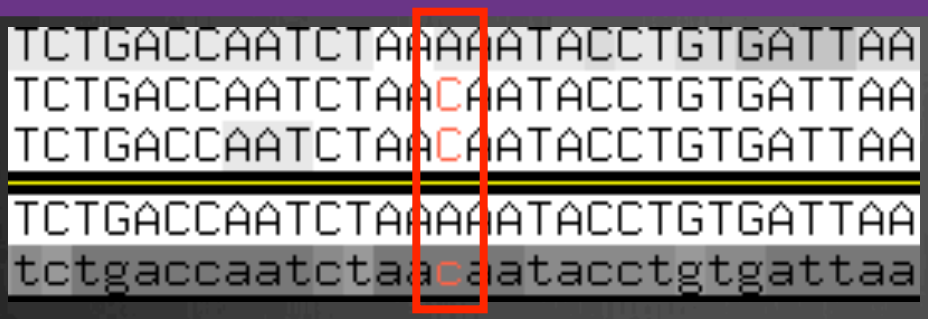
sequencing error

true polymorphism

Base coverage

Prior expectation

Bayesian SNP calling



polymorphic permutation

monomorphic permutation

Bayesian posterior probability

Base call + Base quality

Expected polymorphism rate

$$P(SNP) = \sum_{all\ variable\ S} \frac{\frac{P(S_1 \mid R_1)}{P_{Prior}(S_1)} \cdot \ldots \cdot \frac{P(S_N \mid R_N)}{P_{Prior}(S_N)} P_{Prior}(S_1,\ldots,S_N)}{\sum_{S_{i_1}\in[A,C,G,T]} \ldots \sum_{S_{i_N}\in[A,C,G,T]} \frac{P(S_{i_1} \mid R_1)}{P_{Prior}(S_{i_1})} \cdot \ldots \cdot \frac{P(S_{i_N} \mid R_1)}{P_{Prior}(S_{i_N})} P_{Prior}(S_{i_1},\ldots,S_{i_N})}$$

Base composition

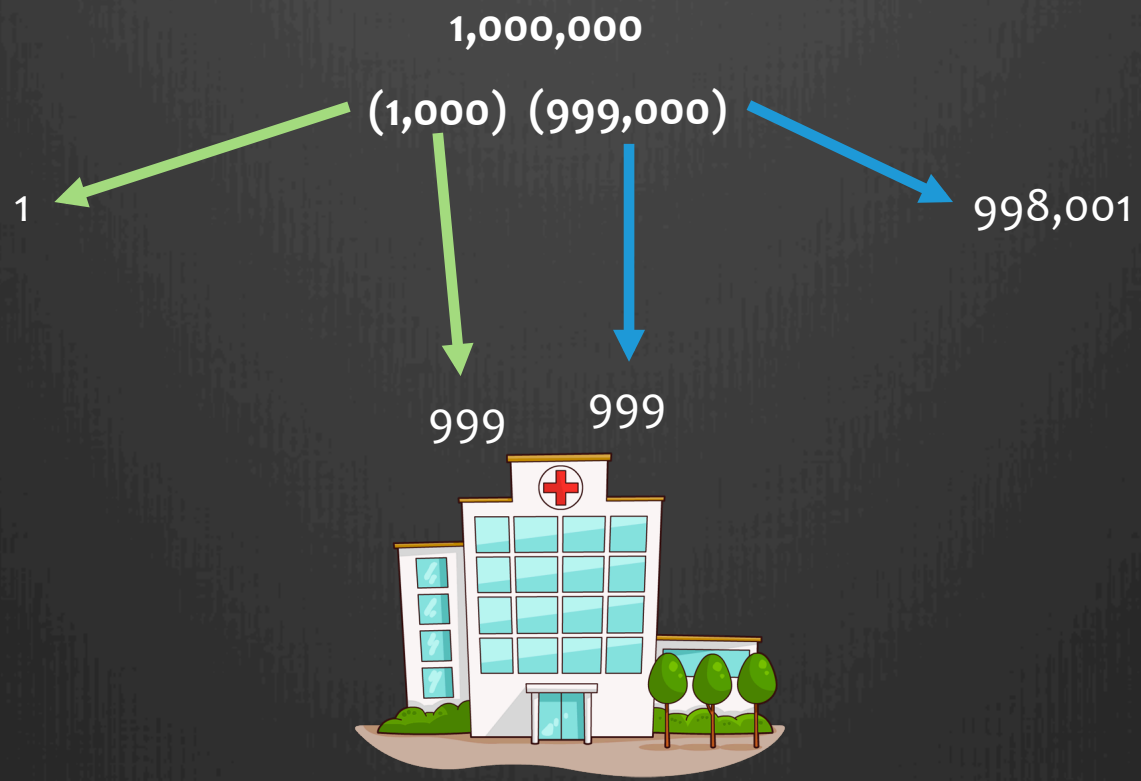
Depth of coverage

Bayes Example

Lets say you have a test for a disease that is 99.9% accurate

You get a positive result, what are the chances you have the disease?

What if I tell you this disease occurs in 1/1,000 people? Does that change your guess?

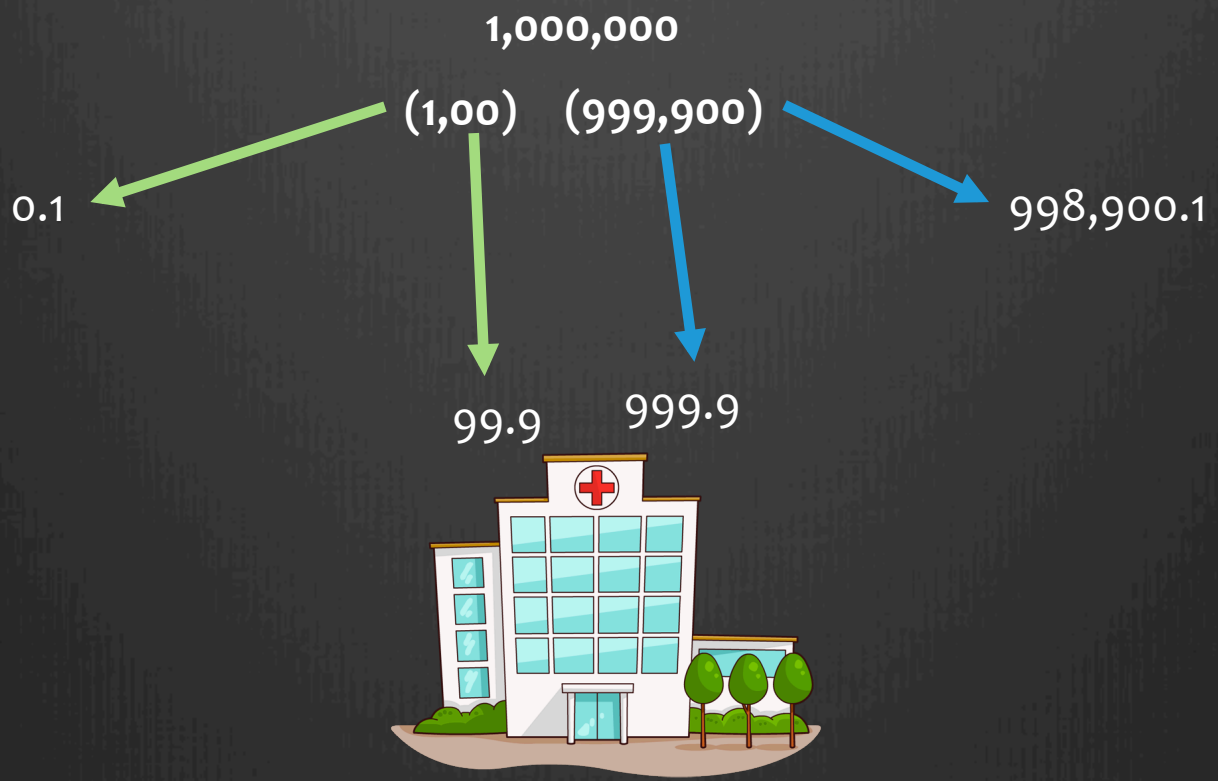


Bayes Example

Lets say you have a test for a disease that is 99.9% accurate

You get a positive result, what are the chances you have the disease?

What if I tell you this disease occurs in 1/10,000 people? Does that change your guess?



Bayes Example

Lets say you have a test for a disease that is 99.9% accurate

You get a positive result, what are the chances you have the disease?

P(sick | +test)

$$P(+test \mid sick) = \frac{P(sick \mid +test)\, P(sick)}{P(Any\ Positive)}$$

P(sick | +test) = 99.9% = 0.999

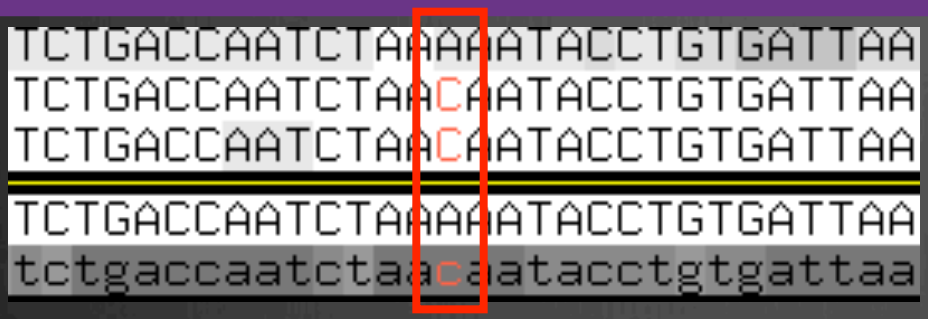
P(sick) = 1/1,000 = 0.001

$$P(Any\ Positive) = \frac{999+999}{1{,}000{,}000} = 0.001998$$

$$P(+test \mid sick) = \frac{(0.999)\ \ (0.001)}{(0.001998)} = 0.5$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Bayesian SNP calling



polymorphic permutation

monomorphic permutation

Bayesian posterior probability

Base call + Base quality

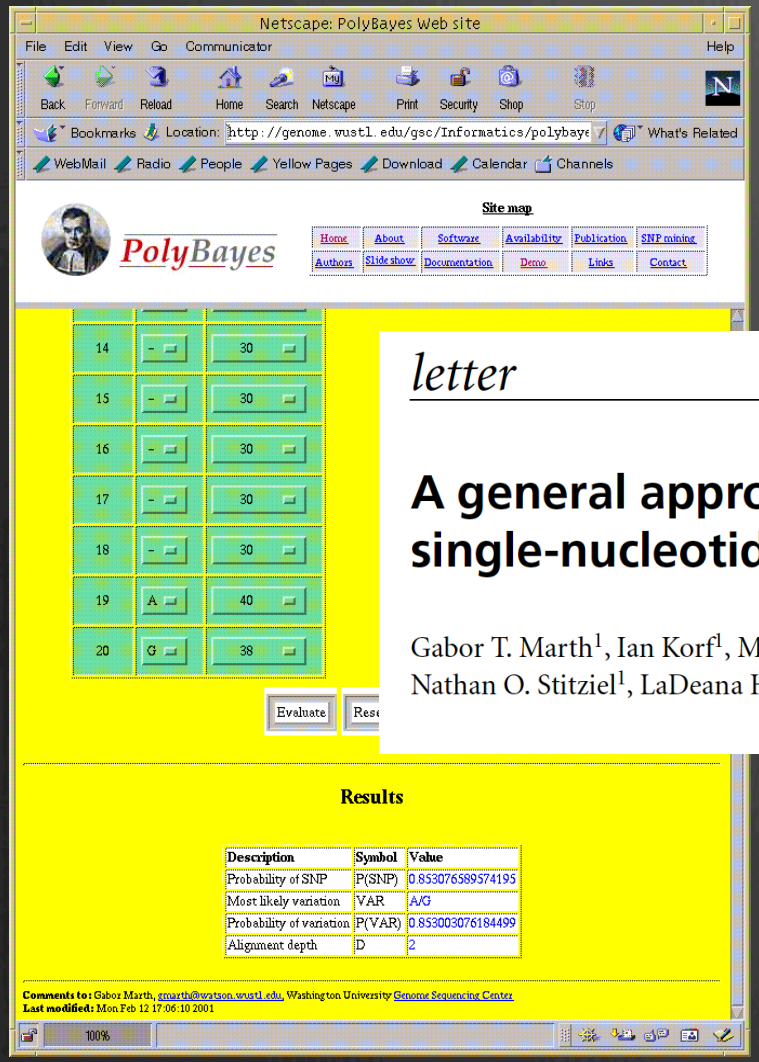
Expected polymorphism rate

$$P(SNP) = \sum_{all\ variable\ S} \frac{\frac{P(S_1 \mid R_1)}{P_{Prior}(S_1)} \cdot \ldots \cdot \frac{P(S_N \mid R_N)}{P_{Prior}(S_N)} P_{Prior}(S_1,\ldots,S_N)}{\sum_{S_{i_1}\in[A,C,G,T]} \ldots \sum_{S_{i_N}\in[A,C,G,T]} \frac{P(S_{i_1} \mid R_1)}{P_{Prior}(S_{i_1})} \cdot \ldots \cdot \frac{P(S_{i_N} \mid R_1)}{P_{Prior}(S_{i_N})} P_{Prior}(S_{i_1},\ldots,S_{i_N})}$$

Base composition

Depth of coverage

PolyBayes: the first statistically rigorous variant detection system



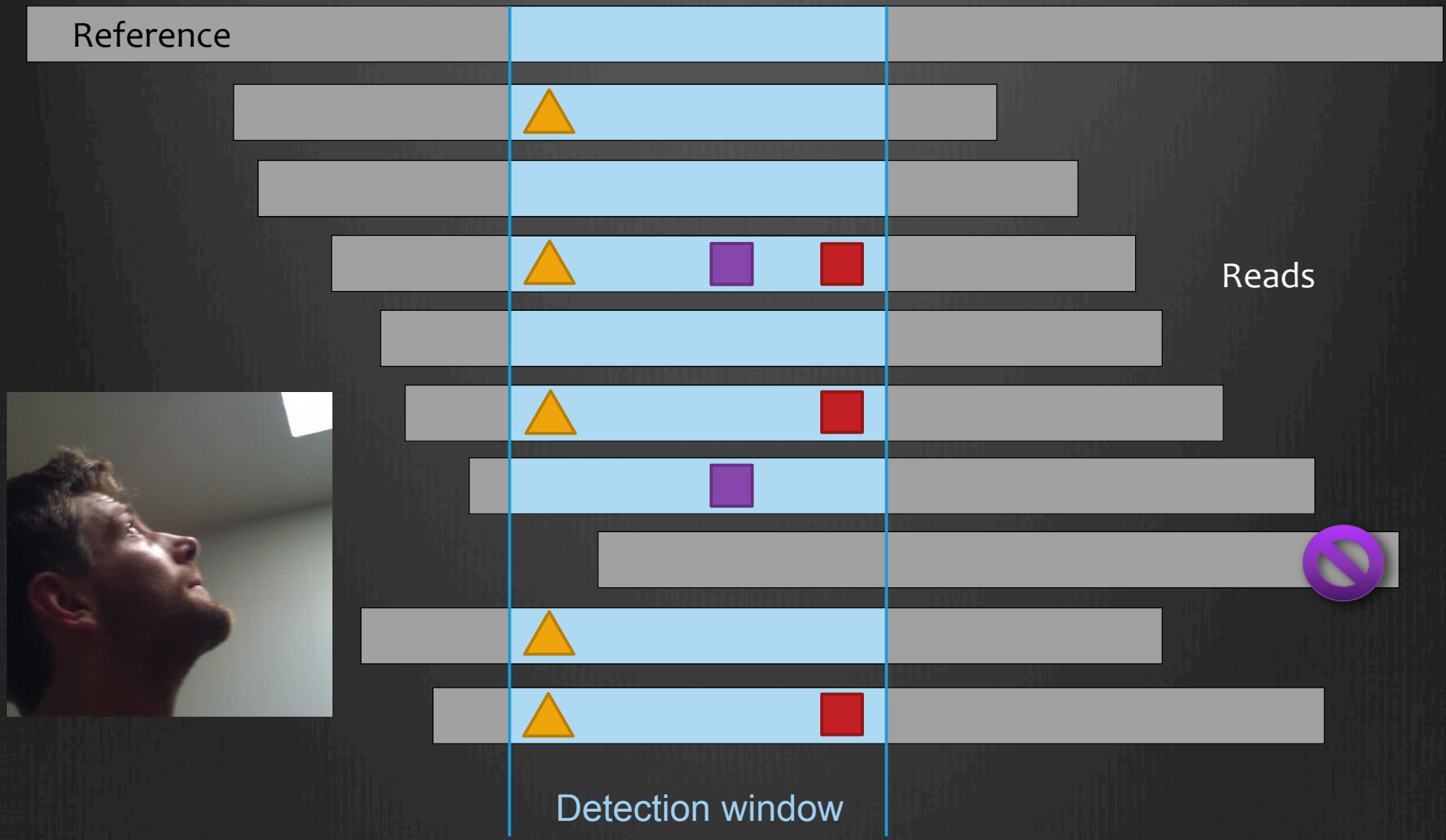
letter © 1999 Nature America Inc. • http://genetics.nature.com

A general approach to single-nucleotide polymorphism discovery

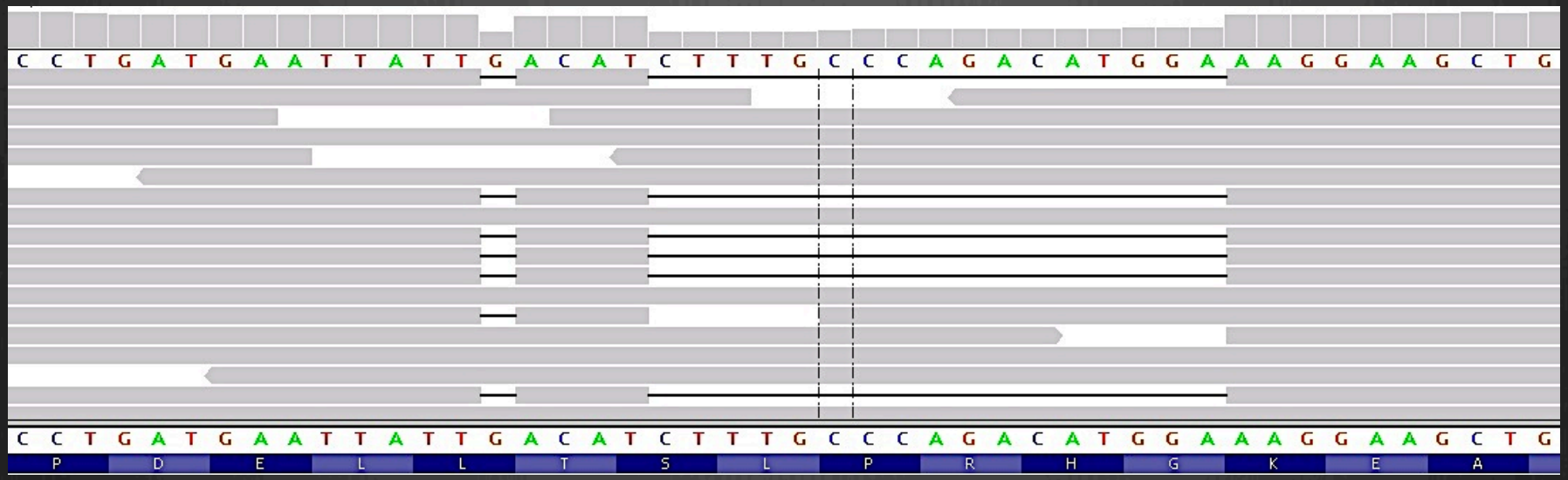
Gabor T. Marth[1], Ian Korf[1], Mark D. Yandell[1], Raymond T. Yeh[1], Zhijie Gu[2], Hamideh Zakeri[2], Nathan O. Stitziel[1], LaDeana Hillier[1], Pui-Yan Kwok[2] & Warren R. Gish[1]

http://bioinformatics.bc.edu/~marth/PolyBayes/

The FreeBayes local haplotype caller

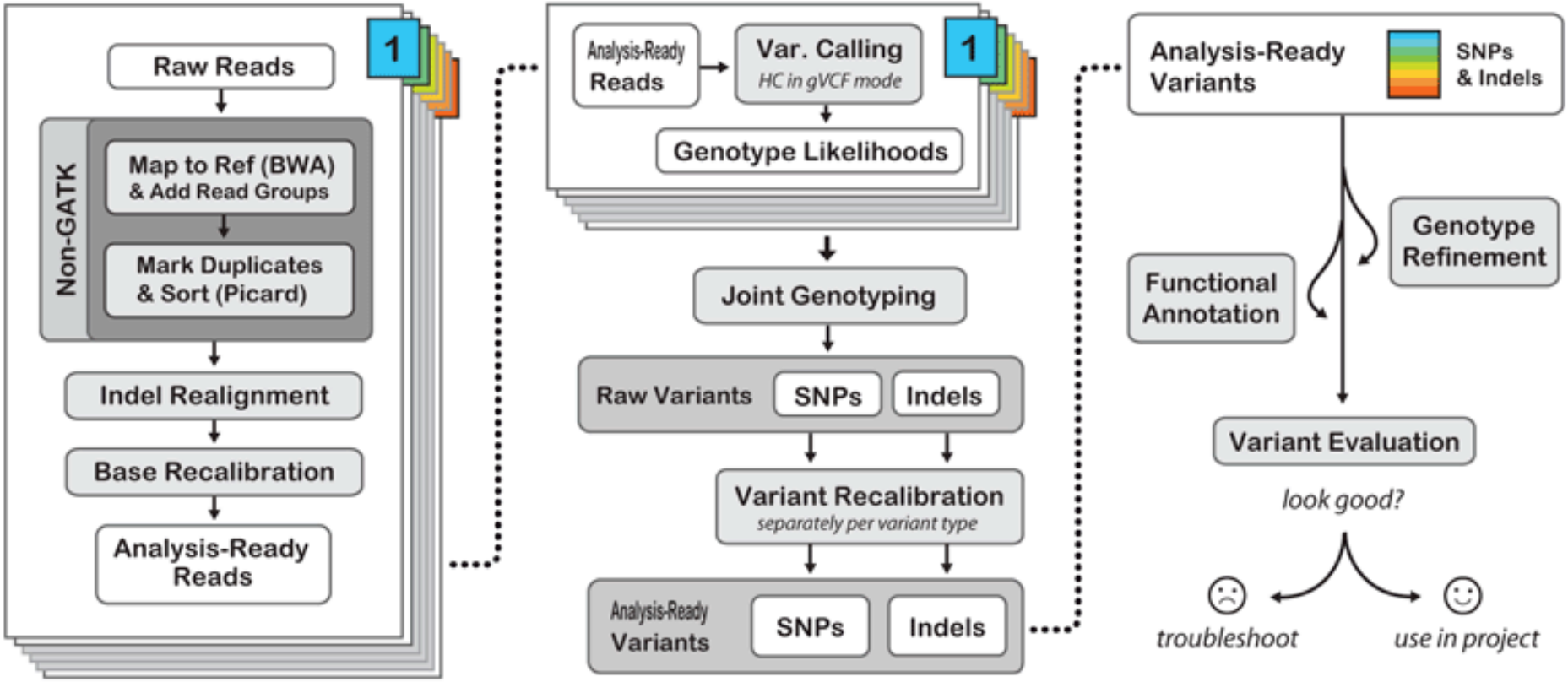


Able to identify compound hets



- Two apparent frameshift deletions in the CASP8AP2 gene (one 17 bp, one 1 bp) on the same haplotype
- Overall effect is **in-frame deletion** of six amino acids

The GATK variant caller



GATK vs freebayes via Genome in a Bottle

Colby Chiang

NA12878

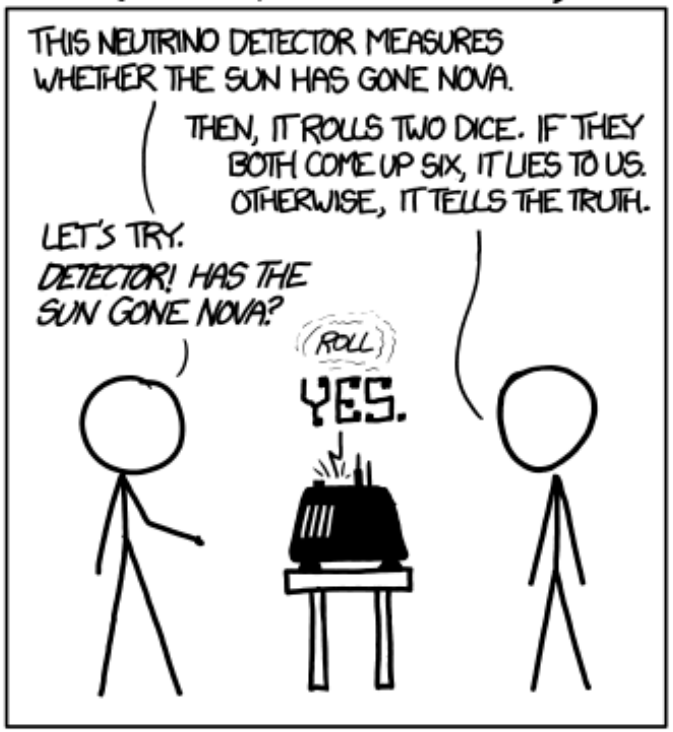
1 Complete Genomics
1 SOLiD WGS
1 454 WGS
1 Ion Exome
2 Illumina exomes
8 Illumina WGS

→ 3.5M SNPS
0.5M INDELs

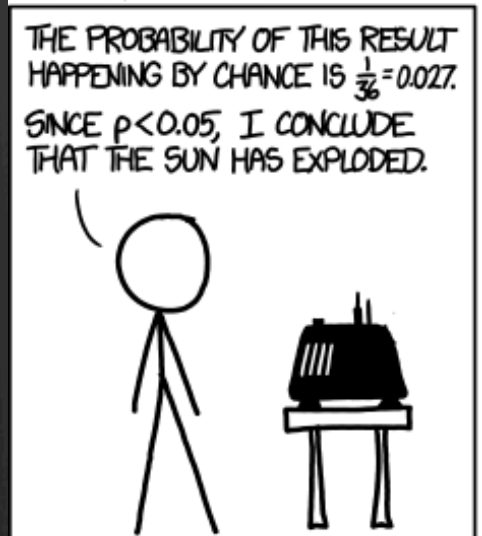
	SNPs		INDELs	
	GATK	freebayes (q≥250)	GATK	freebayes(q≥0)
Sensitivity	0.976	0.986	0.857	0.91
Specificity	0.999	0.999	0.999	0.999
FDR	0.008	0.022	0.01	0.021

http://genomeinabottle.org/

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)



FREQUENTIST STATISTICIAN:



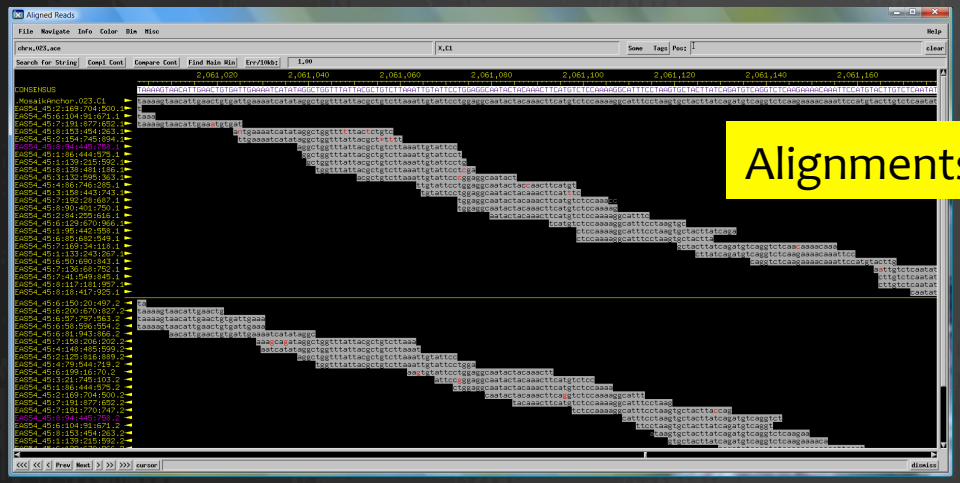
BAYESIAN STATISTICIAN:

Finer points

Data formats

```
@IL11_266:1:1:395:231/1
CCAACCACAACACAAAAAACACAAGCAAC
+
@@AAAAA?<>@@>?:475;A6?384,>5
@IL11_266:1:1:399:301/1
CAAAAAAAAAAAGAAGTACGAGATACGACACATCAC
+
;@AAAA>5;>@C67'&2?&7<&7&@1/1408=19::
```

Reads: FASTQ

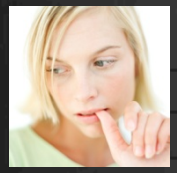


Alignments: SAM/BAM

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002
20	14370	rs6054257	G	A	29	0	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1	1|0:48:8
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3	
20	1110696	rs6040355	A	G,T	67	0	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6	
20	1230237	.	T	.	47	0	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7	
20	1234567	microsat1	G	D4,IGA	50	0	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

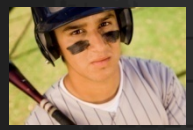
Variants: VCF

Joint calling



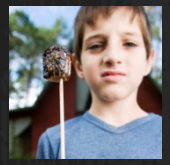
a
a
c
c

$P(B_1=aacc|G_1=aa)$
$P(B_1=aacc|G_1=cc)$
$P(B_1=aacc|G_1=ac)$



a
a
a
a
c

$P(B_i=aaaac|G_i=aa)$
$P(B_i=aaaac|G_i=cc)$
$P(B_i=aaaac|G_i=ac)$



c
c
c
c

$P(B_n=cccc|G_n=aa)$
$P(B_n=cccc|G_n=cc)$
$P(B_n=cccc|G_n=ac)$

"genotype likelihoods"

P(SNP)

$Prior(G_1,..,G_i,..,G_n)$

$P(G_1=aa|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_1=cc|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_1=ac|B_1=aacc; Bi=aaaac; B_n=cccc)$

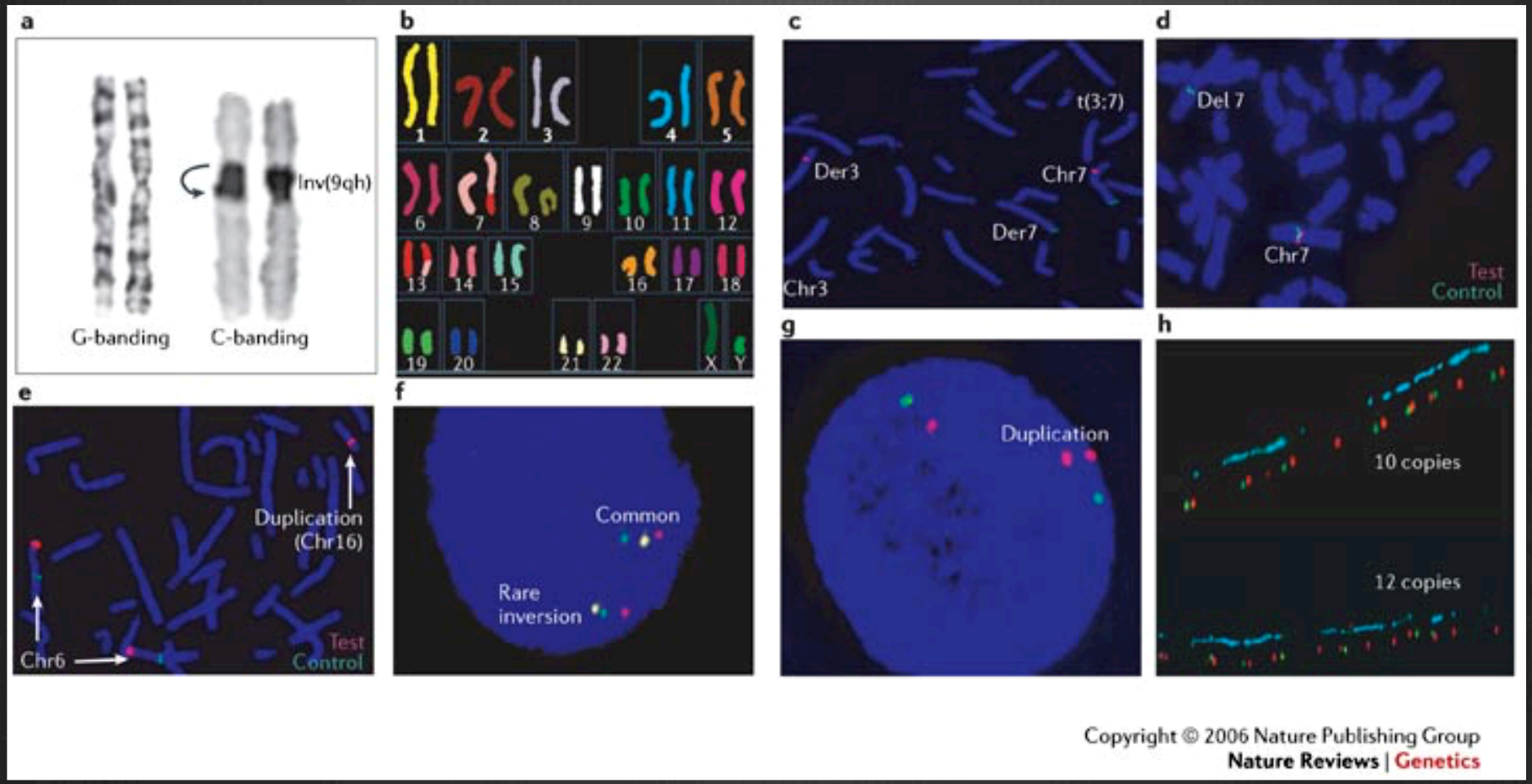
$P(G_i=aa|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_i=cc|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_i=ac|B_1=aacc; Bi=aaaac; B_n=cccc)$

$P(G_n=aa|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_n=cc|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_n=ac|B_1=aacc; Bi=aaaac; B_n=cccc)$

"genotype probabilities"

Detecting structural variations

Structural variations require different approaches

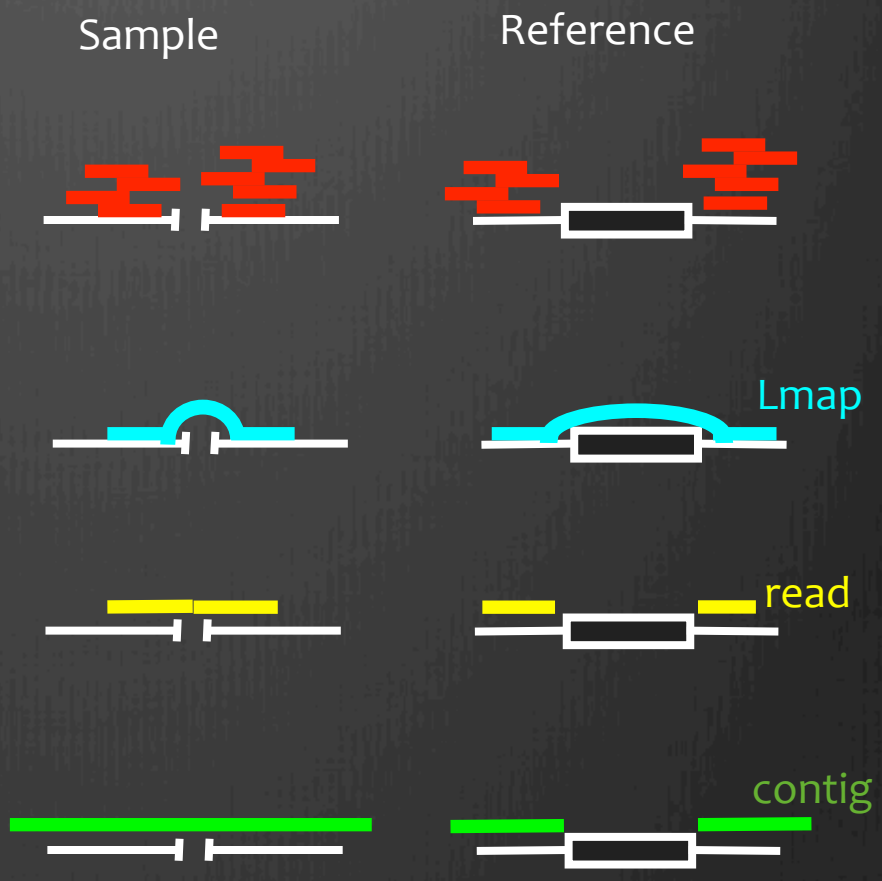


Feuk *et al.* Nature Reviews Genetics, 2006

Short variants are detected directly from read “multiple alignments”

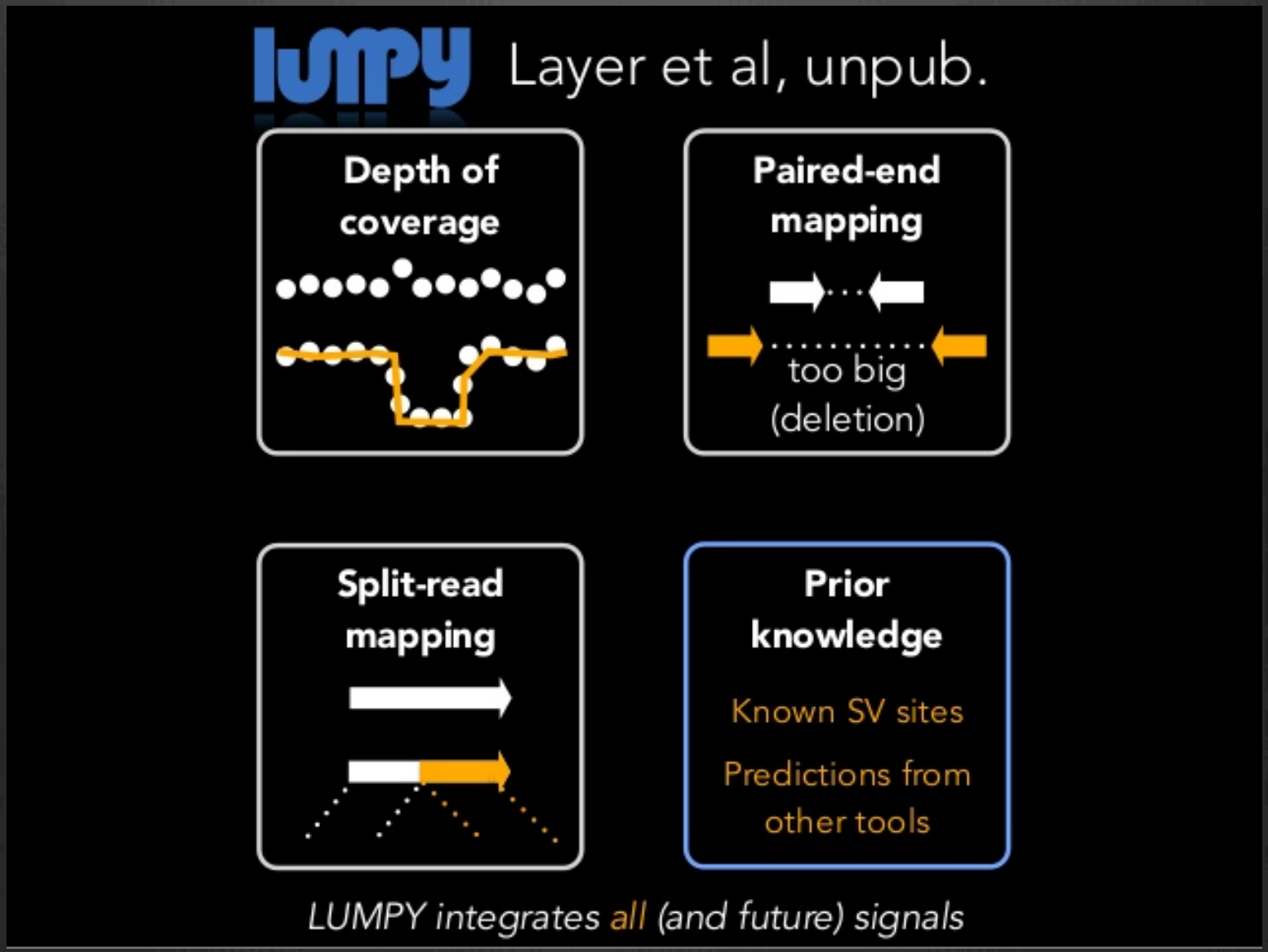
Structural variant use mapping patterns, rather than base alignments

- Read Depth:
  good for big CNVs
- Paired-end:
  all types of SV
- Split-Reads
  good break-point resolution
- *De novo* Assembly

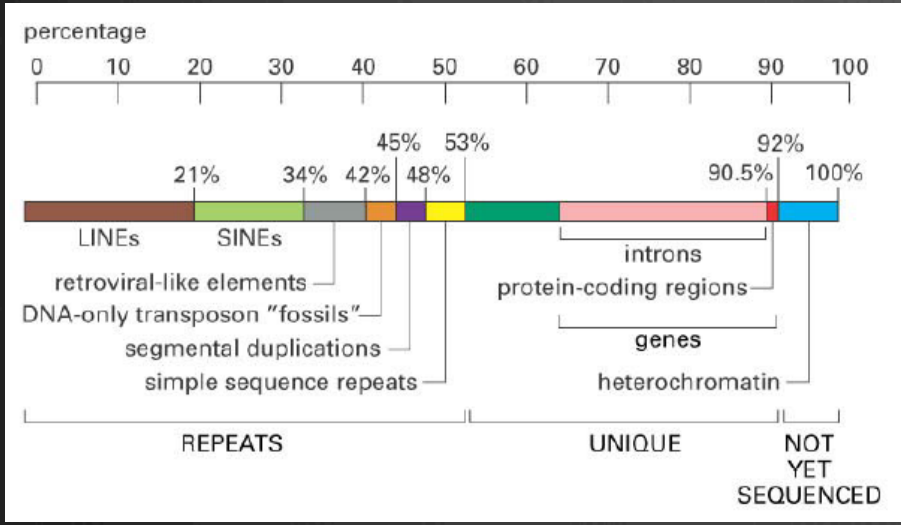


Chip Stewart

The Lumpy SV caller (Layer, Quinlan)

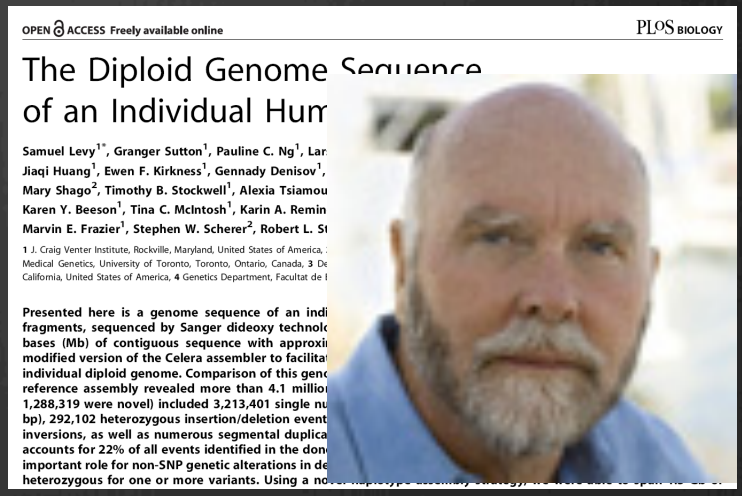


Mobile Element Insertions (MEIs)



- MEs ~half of the human genome
- Many are polymorphic
- Some elements (Alu subclasses, L1s, SVAs) are still active

- ~2,000 events present between the two reference genomes
- population segregation properties largely unknown


OPEN ACCESS Freely available online

PLOS BIOLOGY

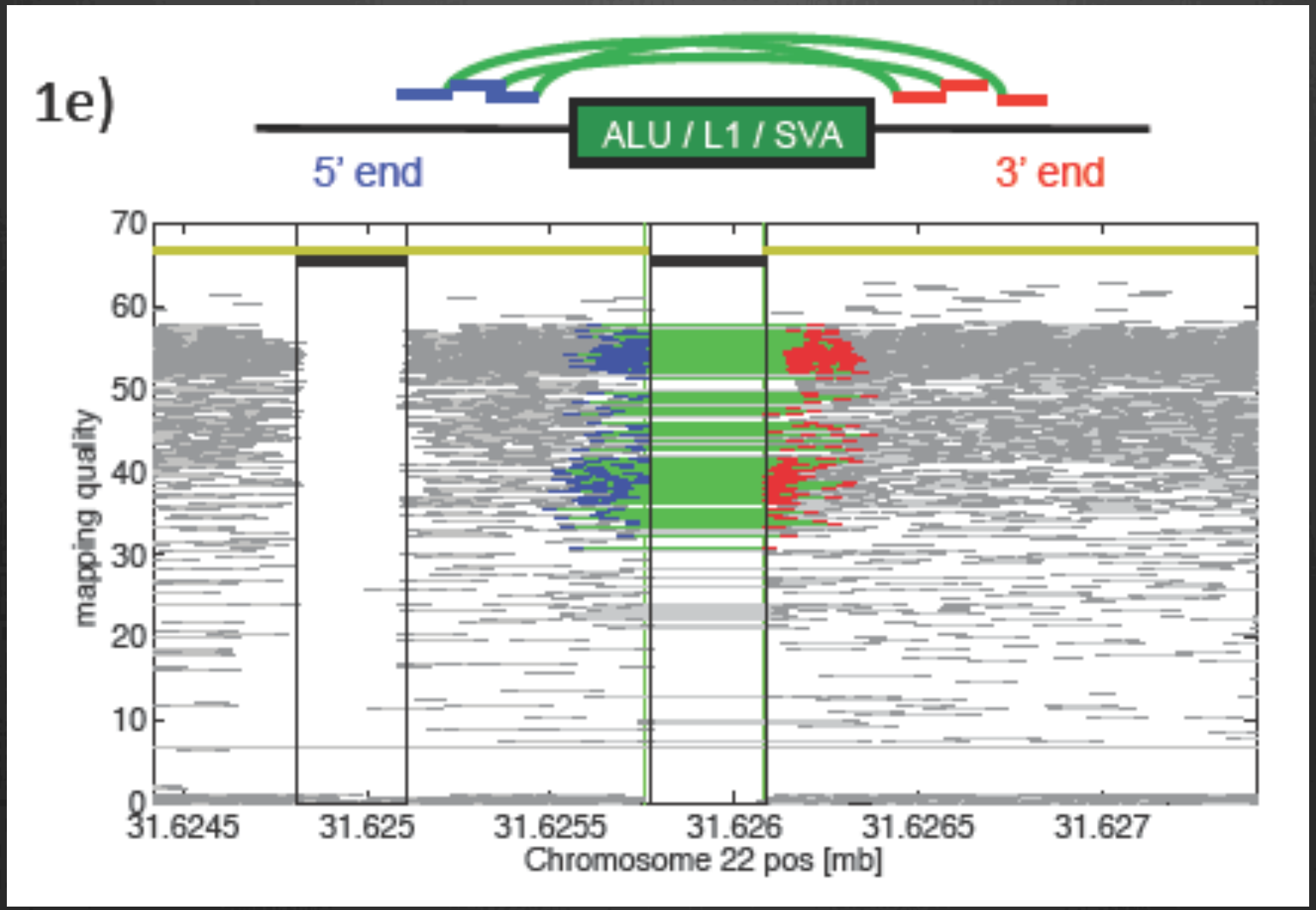
The Diploid Genome Sequence of an Individual Hum

Samuel Levy[1*], Granger Sutton[1], Pauline C. Ng[1], Lar
Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1],
Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamou
Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remin
Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. St

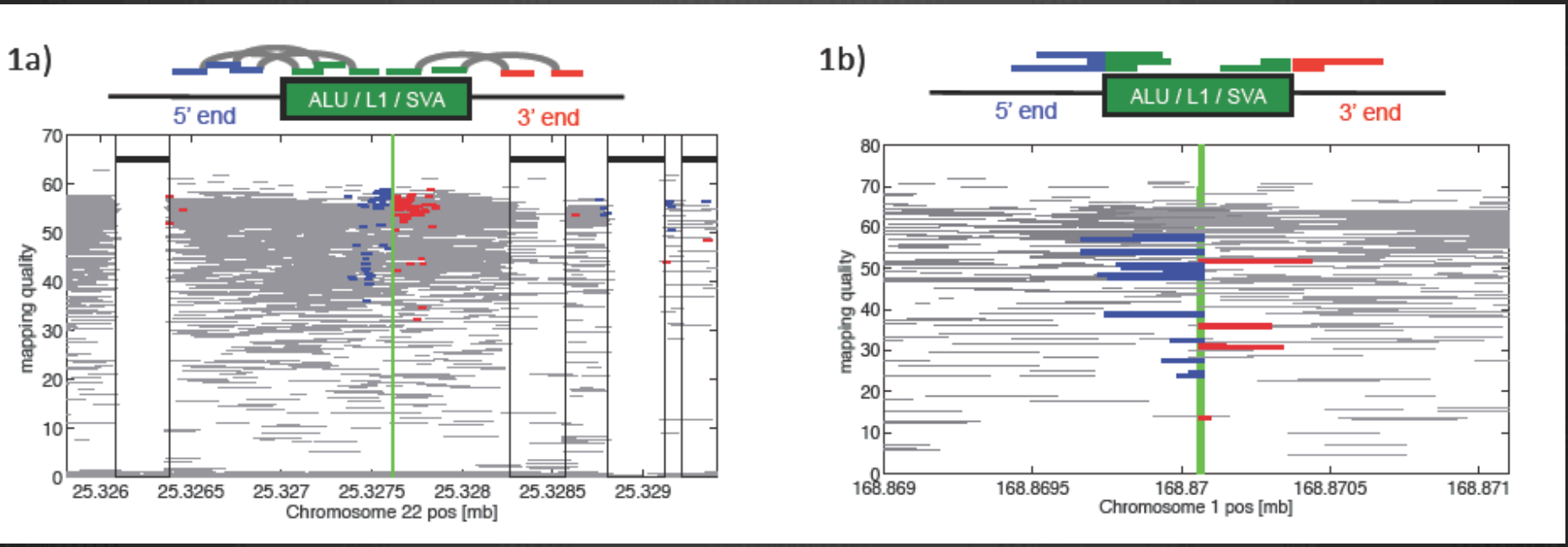
1 J. Craig Venter Institute, Rockville, Maryland, United States of America,
Medical Genetics, University of Toronto, Toronto, Ontario, Canada, 3 De
California, United States of America, 4 Genetics Department, Facultat de

Presented here is a genome sequence of an indi
fragments, sequenced by Sanger dideoxy technol
bases (Mb) of contiguous sequence with approxi
modified version of the Celera assembler to facilita
individual diploid genome. Comparison of this gen
reference assembly revealed more than 4.1 millio
1,288,319 were novel) included 3,213,401 single nu
bp), 292,102 heterozygous insertion/deletion event
inversions, as well as numerous segmental duplica
accounts for 22% of all events identified in the don
important role for non-SNP genetic alterations in de
heterozygous for one or more variants. Using a

Detection previously restricted to ref. insertions

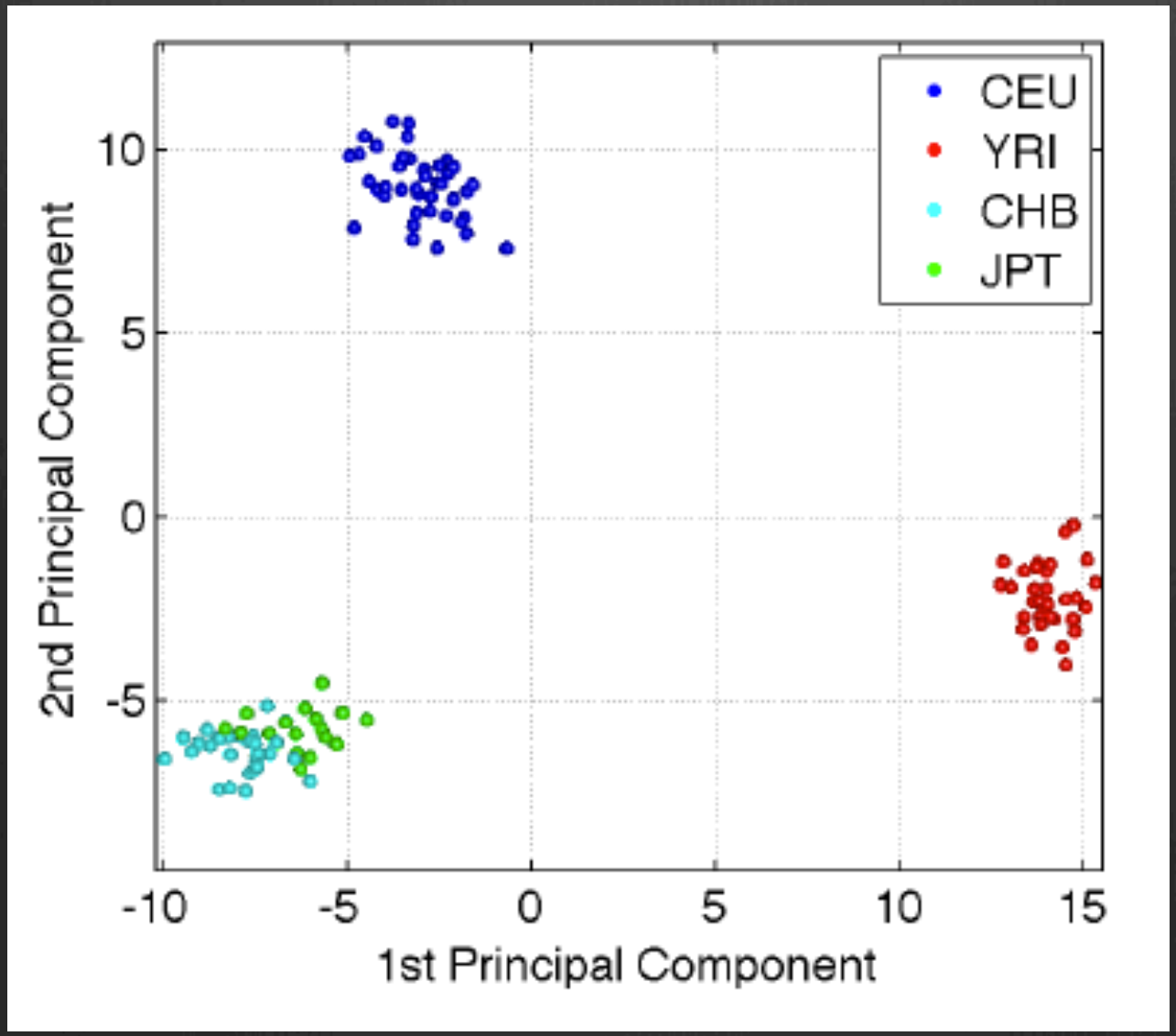


Direct detection of ME insertion events

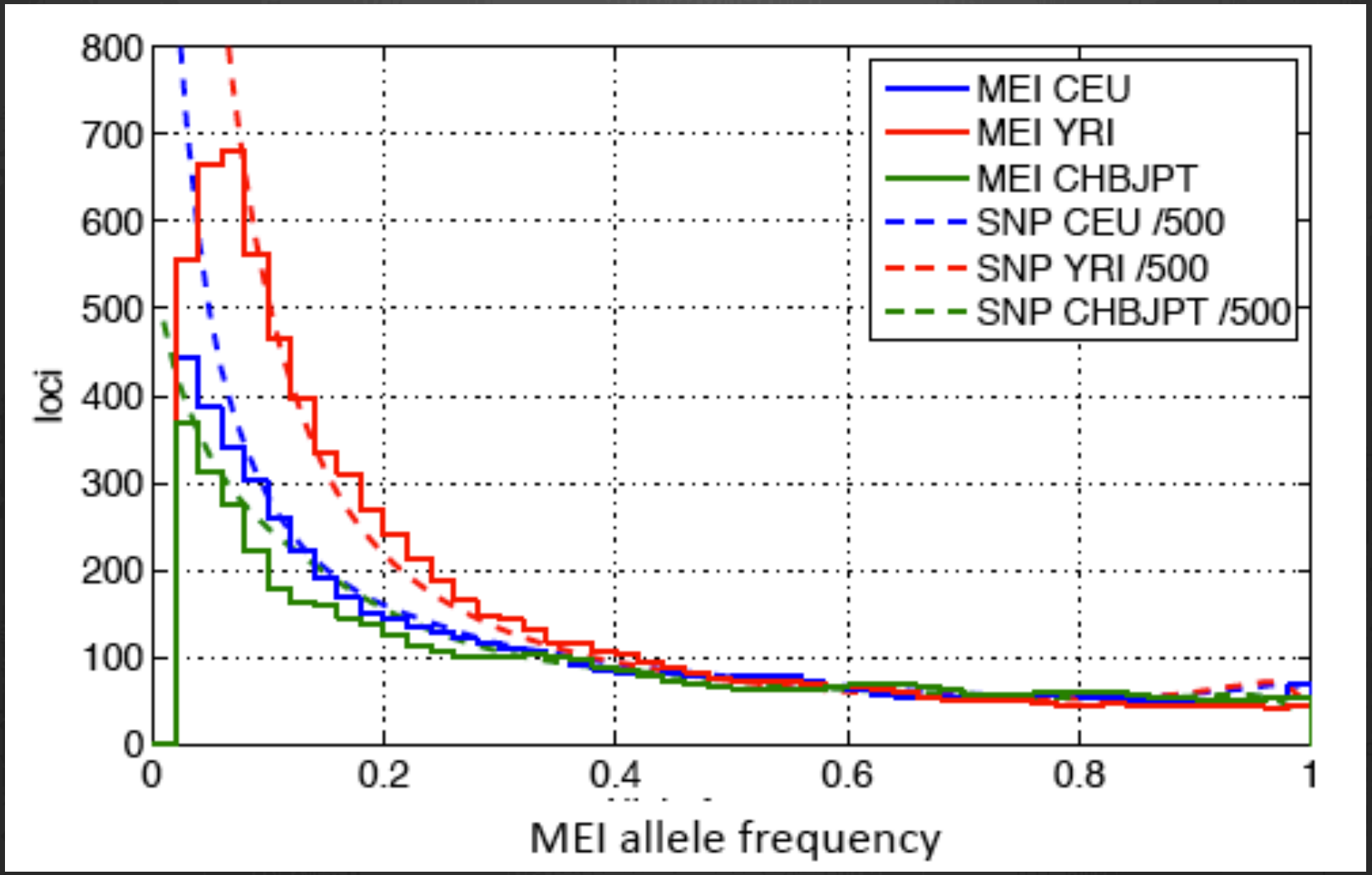


Stewart et al. PLoS Genetics, 2011

MEIs cluster according to populations (like SNPs)



MEI allele frequency also similar to SNPs

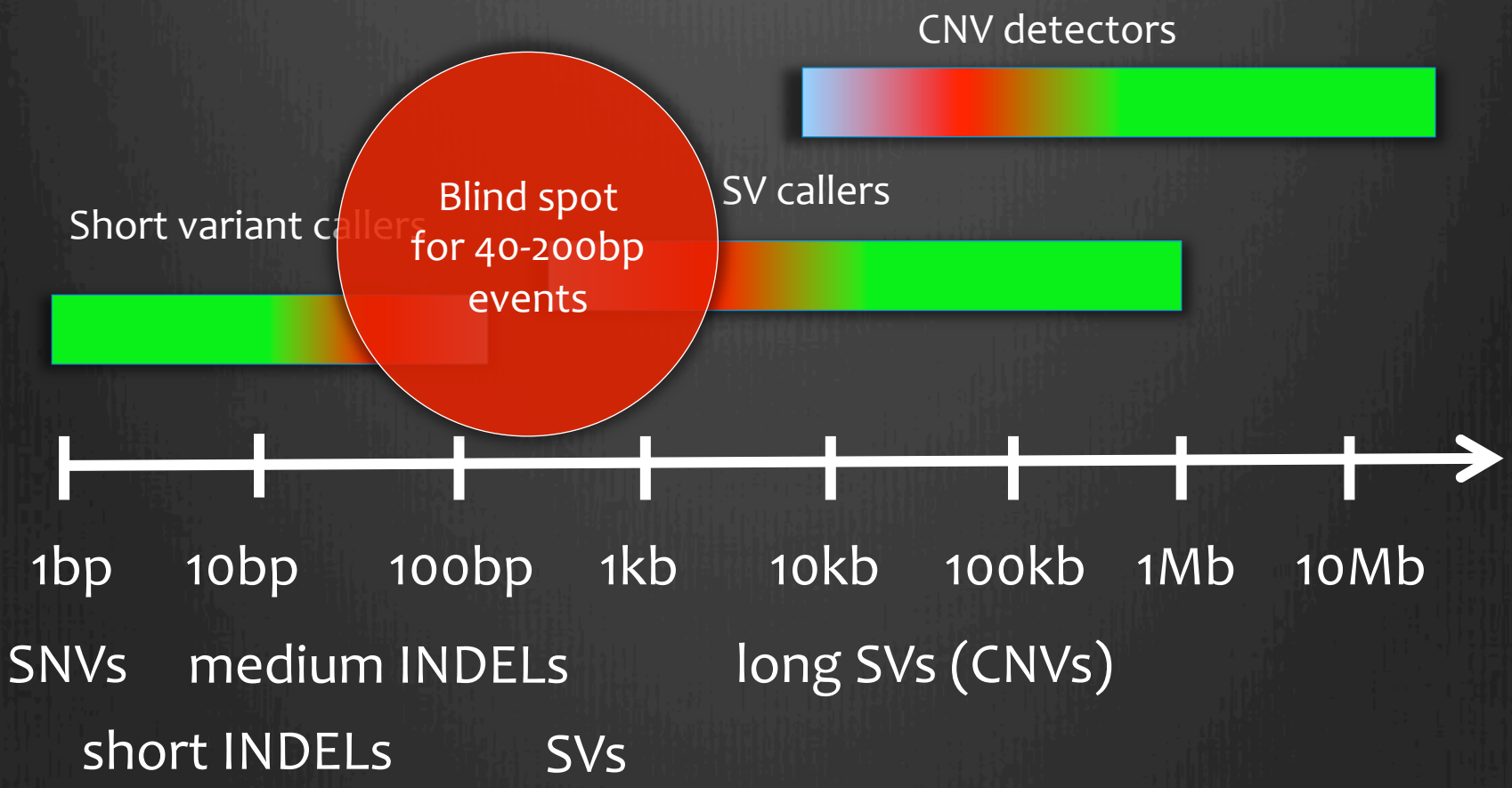


Chip Stewart

Reference-free variant detection

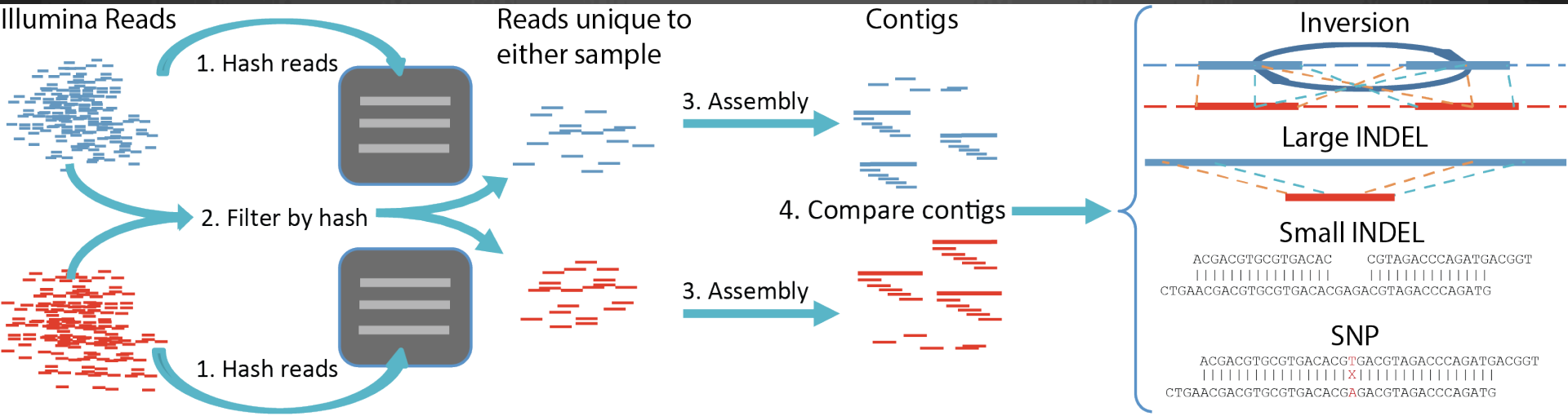
Reference-guided methods have limitations for highly diverged variants

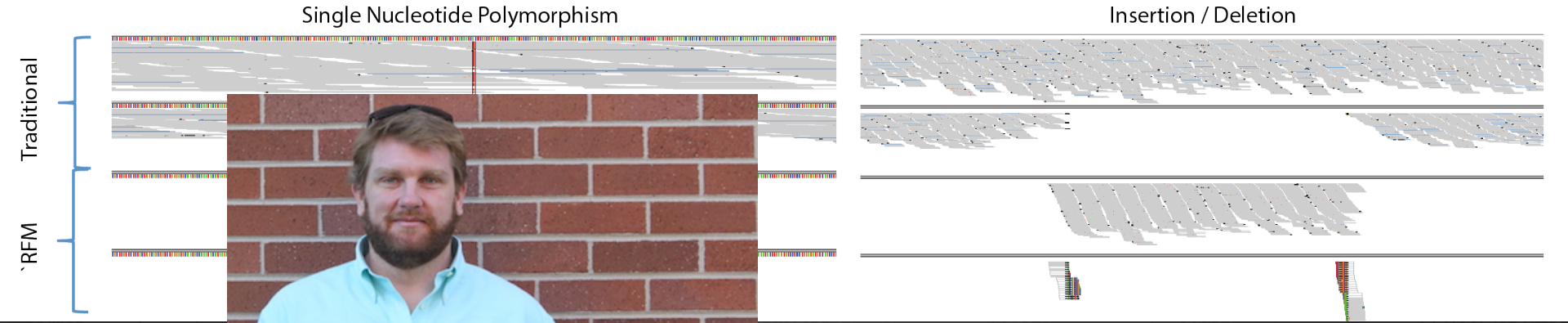
Detection sensitivity low for medium-length events



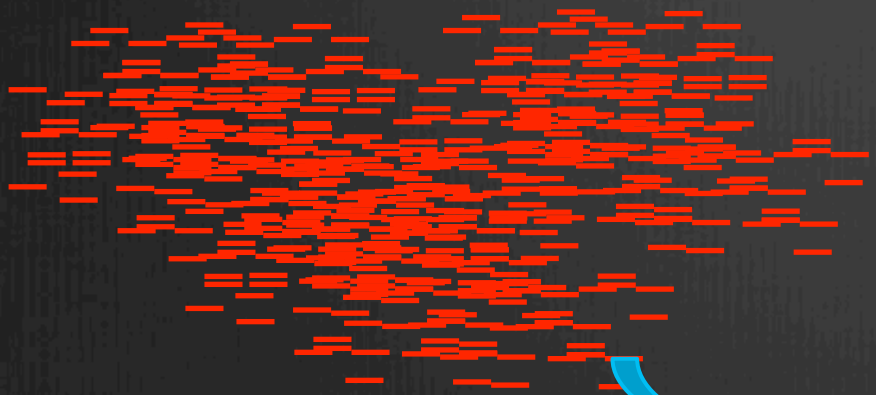
Reference-free methods compare genomes directly (without first aligning data to a reference sequence)

rufus


Illumina Reads
1. Hash reads
2. Filter by hash
1. Hash reads
Reads unique to either sample
3. Assembly
3. Assembly
Contigs
4. Compare contigs
Inversion
Large INDEL
Small INDEL
ACGACGTGCGTGACAC
CGTAGACCCAGATGACGGT
CTGAACGACGTGCGTGACACGAGACGTAGACCCAGATG
SNP
ACGACGTGCGTGACACGTGACGTAGACCCAGATGACGGT
CTGAACGACGTGCGTGACACGAGACGTAGACCCAGATG


Single Nucleotide Polymorphism
Insertion / Deletion
Traditional
`RFM

1: Use reads to create “hash tables”

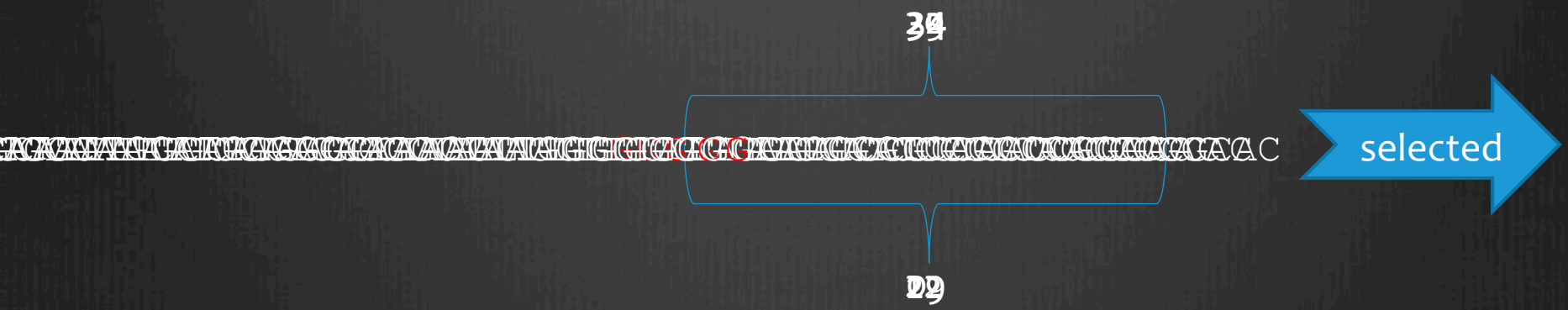


46	AGTCGTCGCCTCTTGCTT
62	TTGAGGAGTCATCTCGAC
59	AAAGCTATGGCAGATATC
58	GTCAAGAGAGAGAAAAGG
61	AACAGGATGAAACCCGGA
59	ACAGAGCGCCGCGACGGC
74	AGAACCACAGTTCGAGCG
78	GTCTTTCTCCTTCTCTTC
56	GTGCAAACTGCGCAGACT
54	GAACACGATGGGCGGAGG
64	GCTAAGAGAGAGAAAAGG
⋮	⋮

46	AGTCGTCGCCTCTTGCTT
62	TTGAGGAGTCATCTCGAC
59	AAAGCTATGGCAGATATC
58	GTCAAGAGAGAGAAAAGG
61	AACAGGATGAAACCCGGA
59	ACAGAGCGCCGCGACGGC
74	AGAACCACAGTTCGAGCG
78	GTCTTTCTCCTTCTCTTC
56	GTGCAAACTGCGCAGACT
54	GAACACGATGGGCGGAGG
64	GCTAAGAGAGAGAAAAGG
⋮	⋮

True Variants

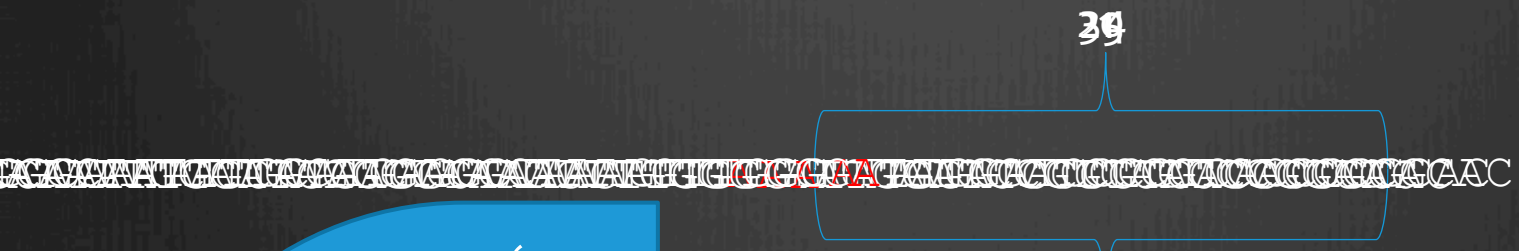
46	AGTCGTCGCCTCTTGCTT
62	TTGAGGAGTCATCTCGAC
59	AAAGCTATGGCAGATATC
58	GTCAAGAGAGAGAAAAGG
61	AACAGGATGAAACCCGGA



46	AGTCGTCGCCTCTTGCTT
62	TTGAGGAGTCATCTCGAC
59	AAAGCTATGGCAGATATC
58	GTCAAGAGAGAGAAAAGG
61	AACAGGATGAAACCCGGA

Sequencing Error

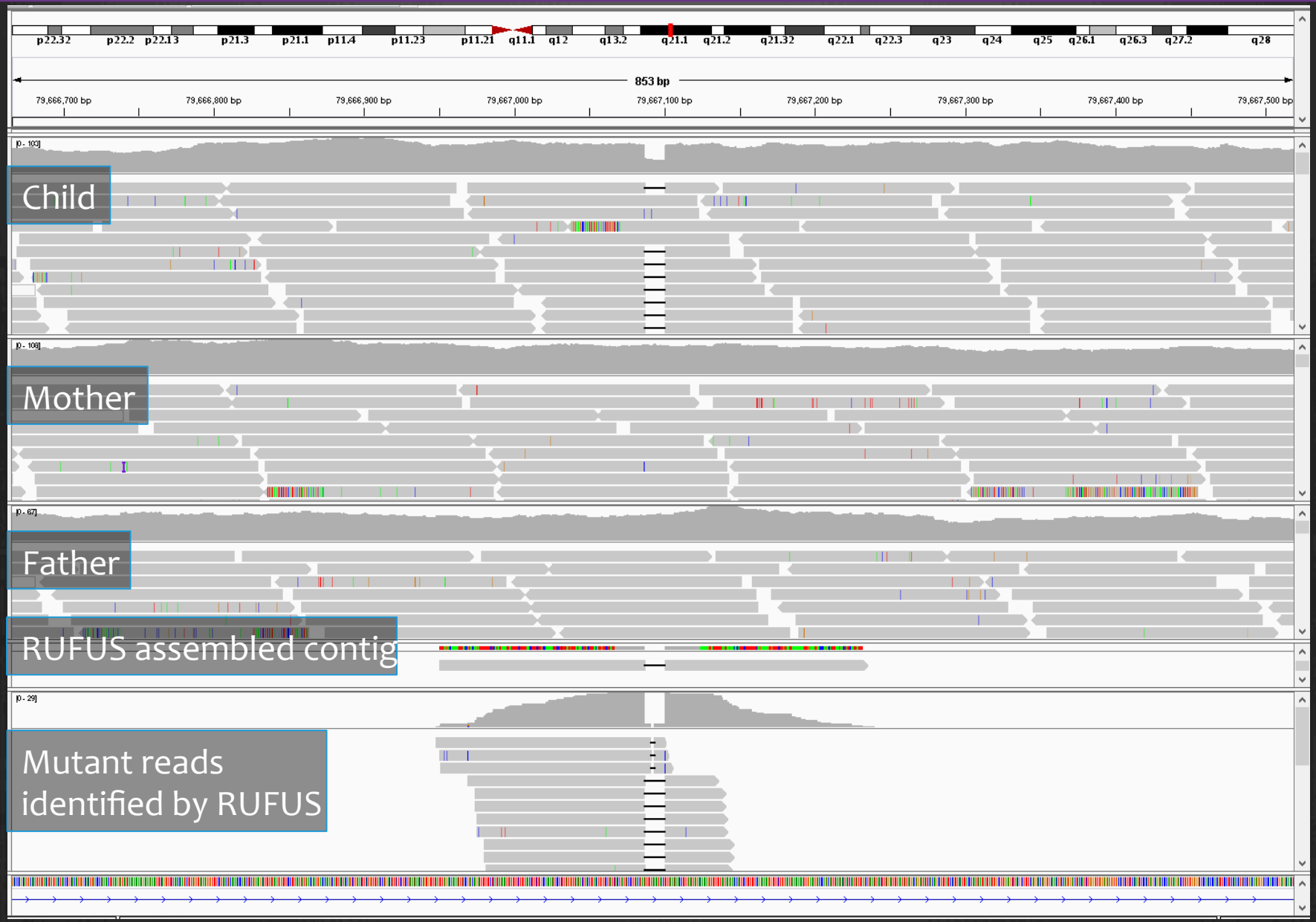
```
46    AGTCGTCGCCTCTTGCTT
62    TTGAGGAGTCATCTCGAC
59    AAAGCTATGGCAGATATC
58    GTCAAGAGAGAGAAAAGG
61    AACAGGATGAAACCCGGA
```



Sequence Error & Uninformative Reads

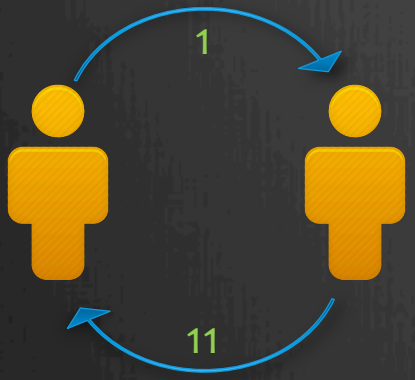
```
6     AGTCGTCGCCTCTTGCTT
62    TTGAGGAGTCATCTCGAC
59    AAAGCTATGGCAGATATC
58    GTCAAGAGAGAGAAAAGG
61    AACAGGATGAAACCCGGA
```

A *de novo* deletion detected by Rufus

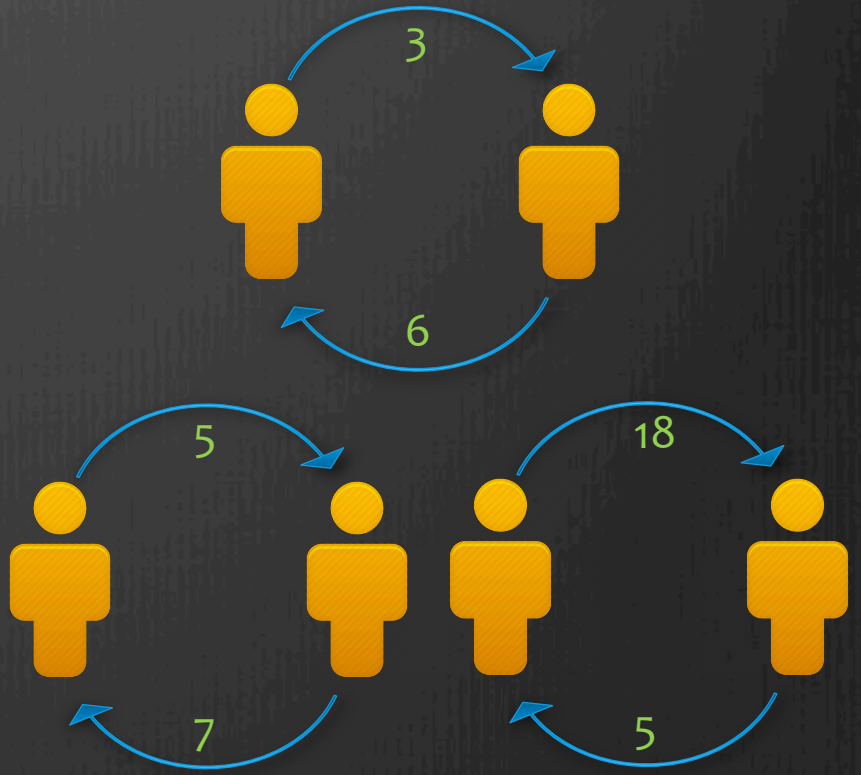


High specificity for variation in MZ twin studies

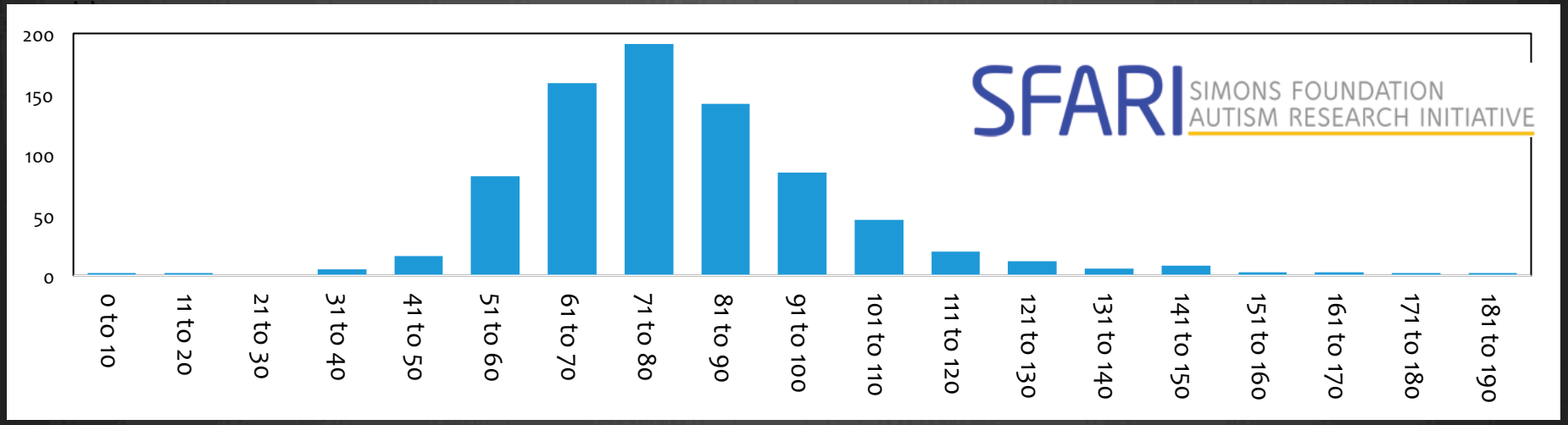
60X Illumina Sequencing

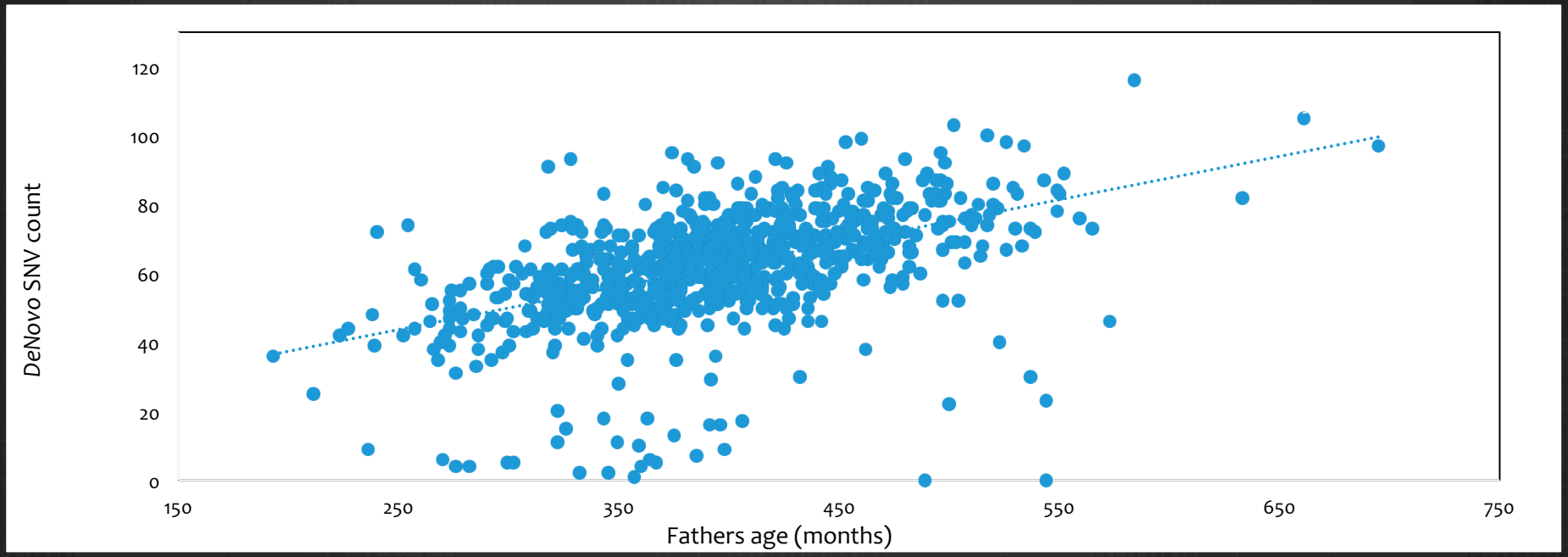


30X Illumina sequencing
+
1000G k-mer filter

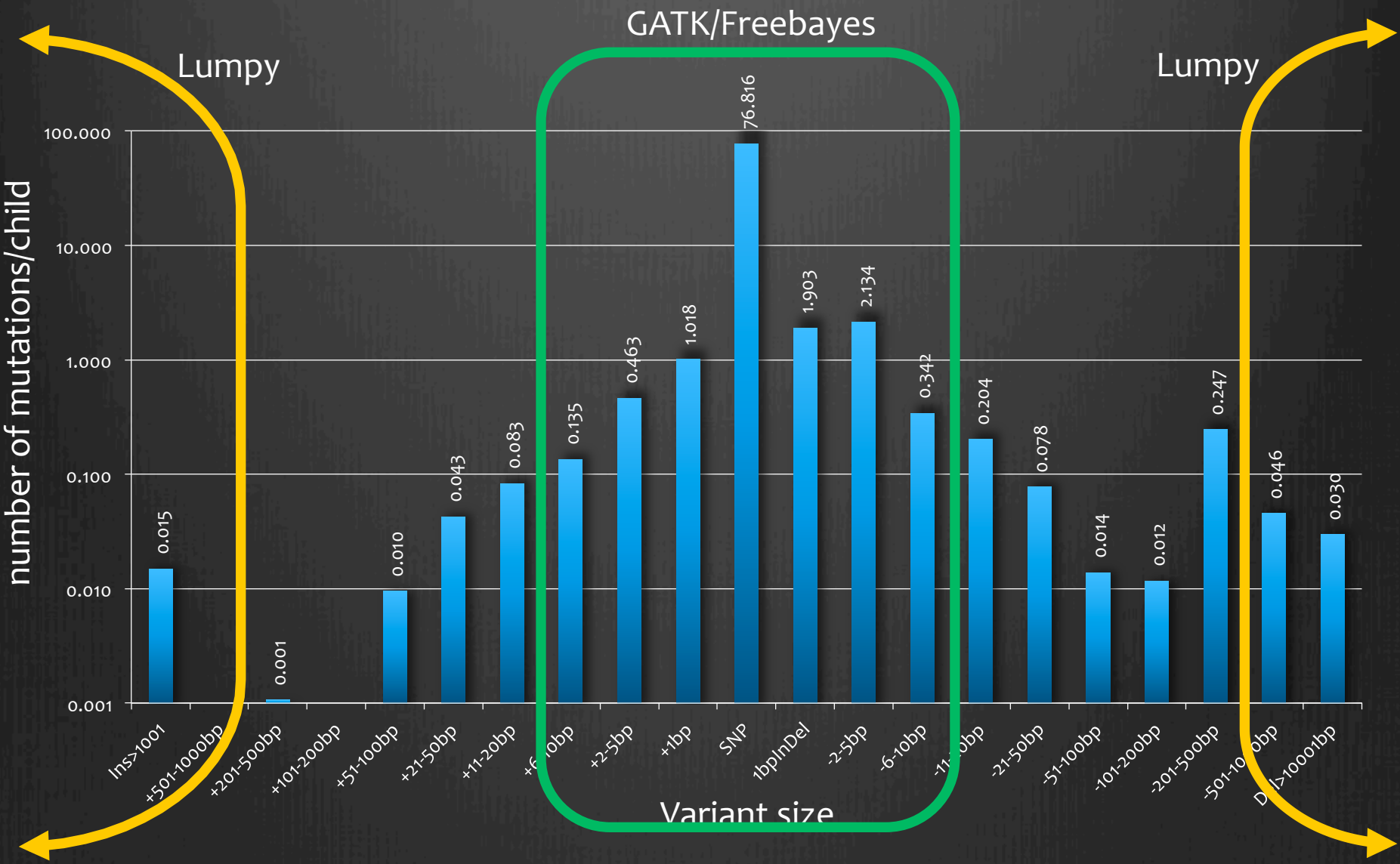


Application to 500 SFARI quartets





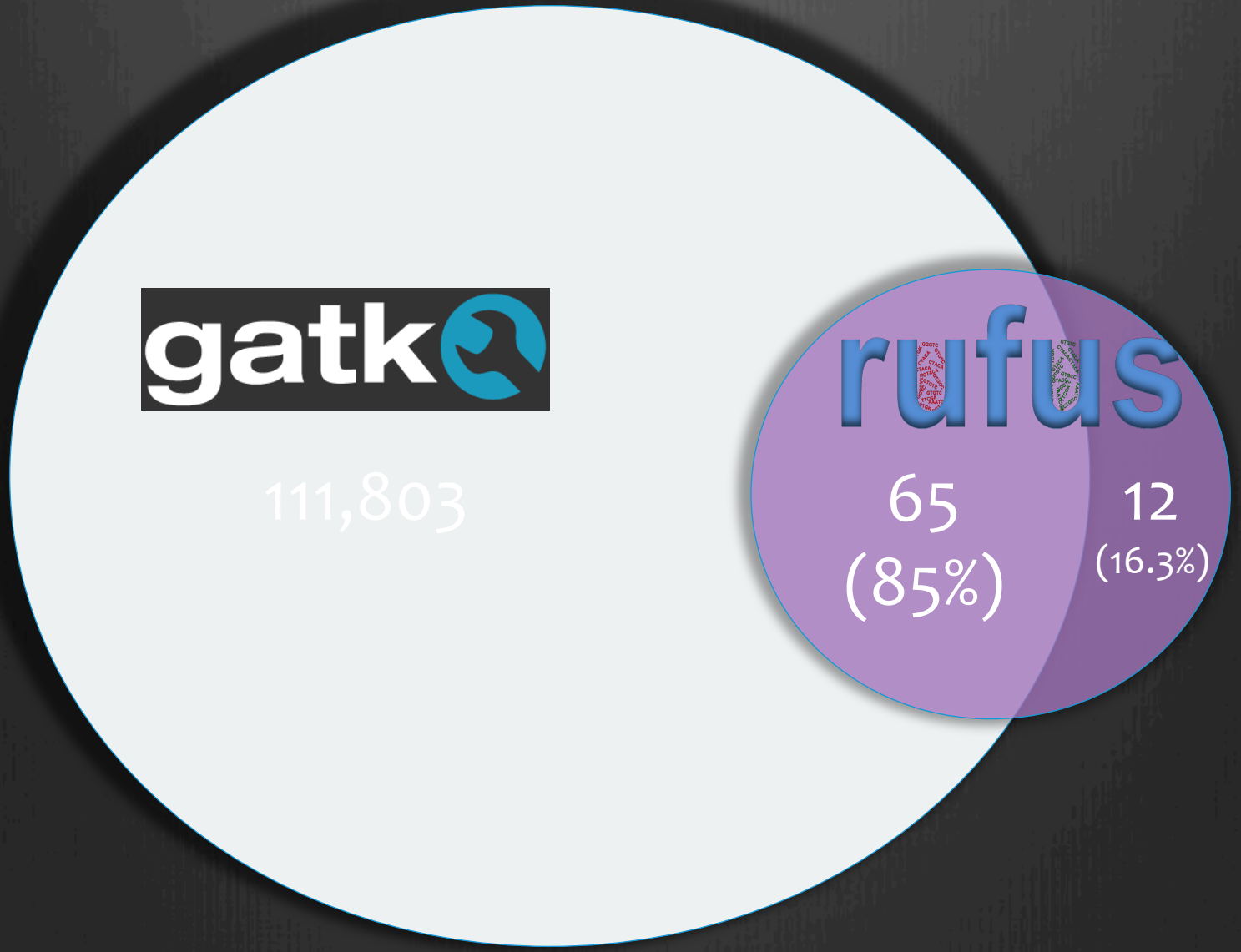
Detecting *de novo* variants of all sizes



Clinical *de novo* variant detection

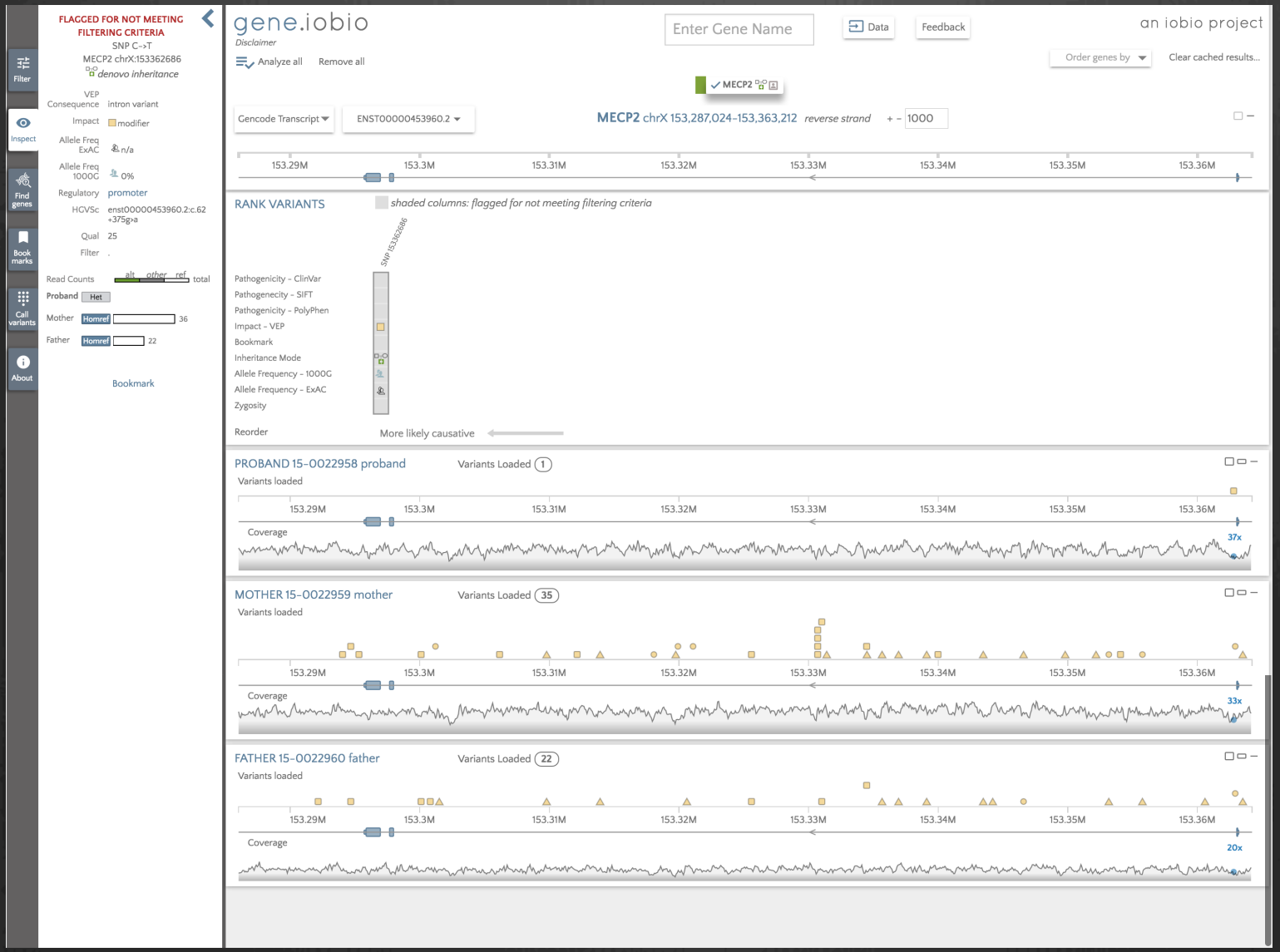
- 15 EIEE (early infantile epileptic encephalopathy) patients sequenced to 60X whole-genome coverage
- Severe seizures in the first weeks/months of life
- Incidence between 1/50,000 and 1/100,000
- Known genetic causes indicate a dominant disease, often caused by de novo mutations
- "Channelopathy": known EIEE genes include: SCN1A, SCN2A, SCN8A, KCNQ2

Typical *de novo* detection results

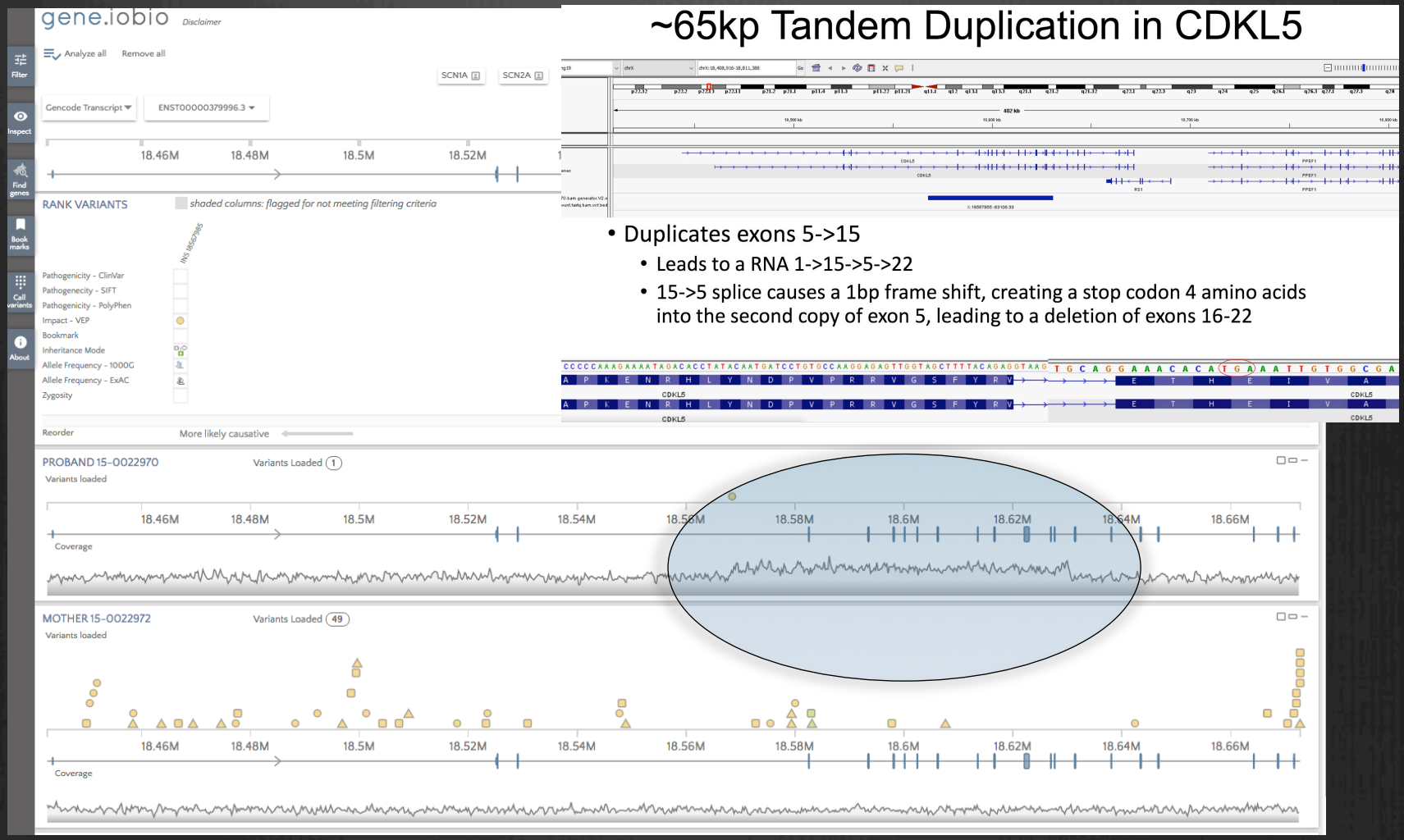


Many false positives, even in coding regions

Confidently detecting non-coding *de novos*



Confidently detecting *de novo* SVs




gene.iobio
MAD RIVER MT


That's all Folks!