

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



**USTAR Center for
Genetic Discovery**

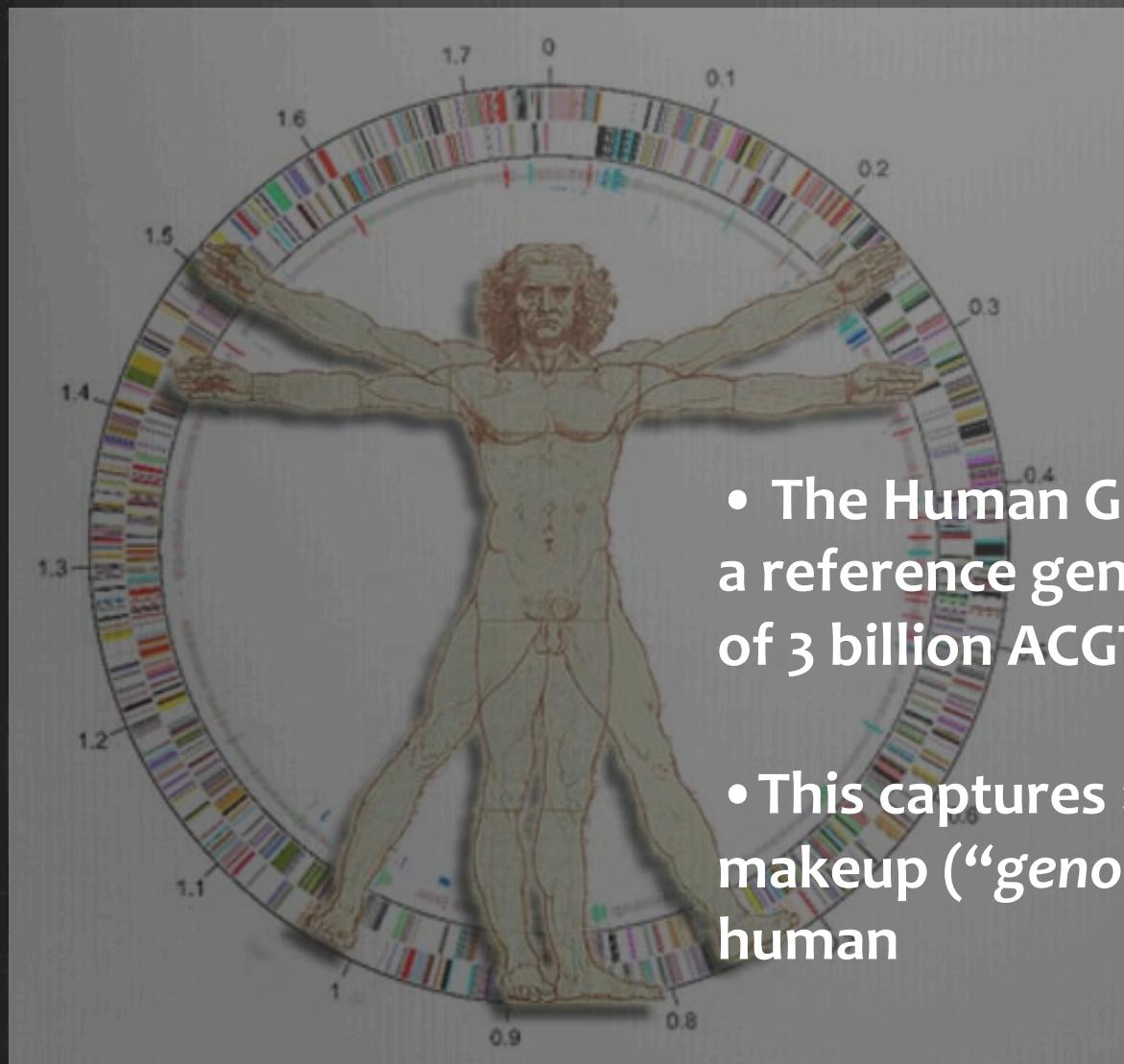
CSHL Sequencing Course
2017

Planned new material/changes

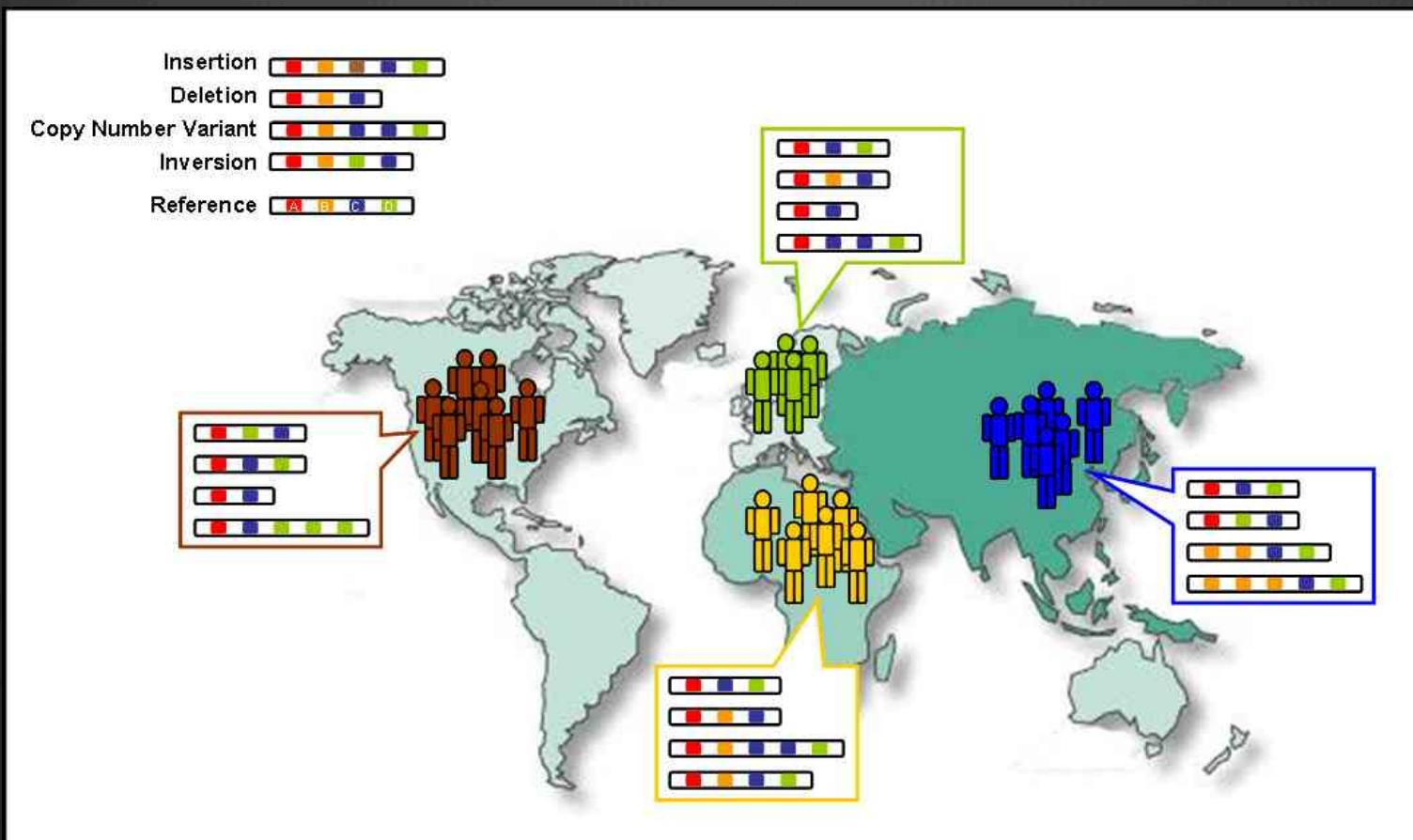
- Add more detail/ de novo variant calling results to RUFUS
- More detail on Graphite
- Some detail on VG (ask Erik for a presentation for materials?)
- Add one slide about 10X data, and variant calling
- Add IOBIO content... Andrew's new variant caller?
- Add at least one slide about PacBio, Oxford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

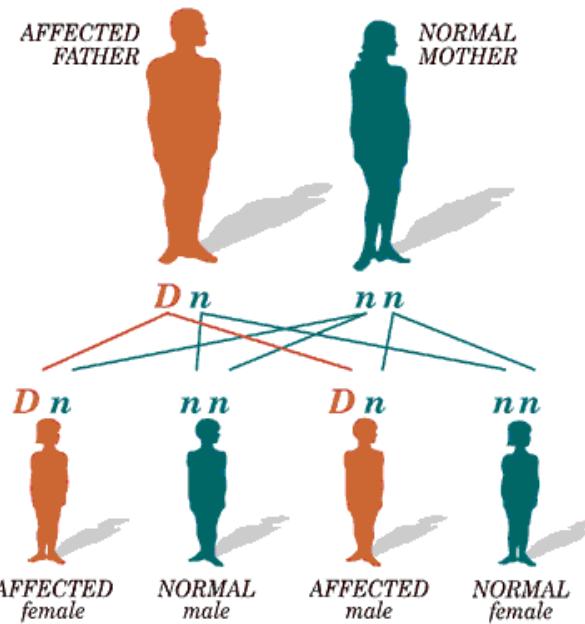


Human genetic diversity is substantial



- Average heterozygosity (the rate of polymorphic differences between two human chromosomes) is $\sim 10^{-3}$ i.e. roughly 1 in 1,000 bp

Genetic variations are important



phenotypic
differences



heritable diseases



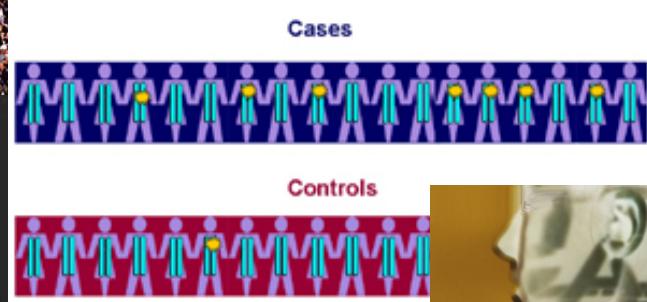
demographic history



Variants impact many areas of science & medicine



Population genomics



Medical genomics



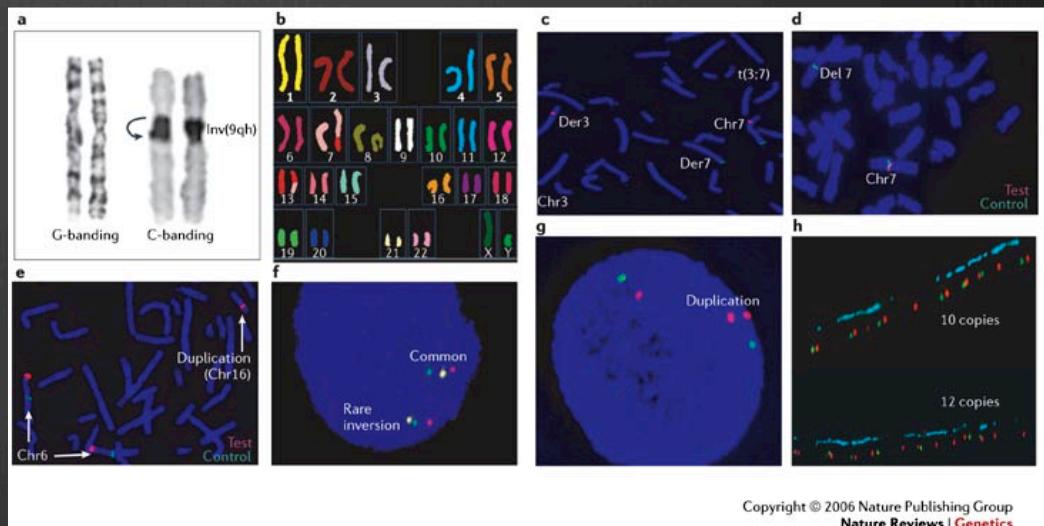
Personal genomics

Clinical genomics

Genetic variations come in different shapes & sizes

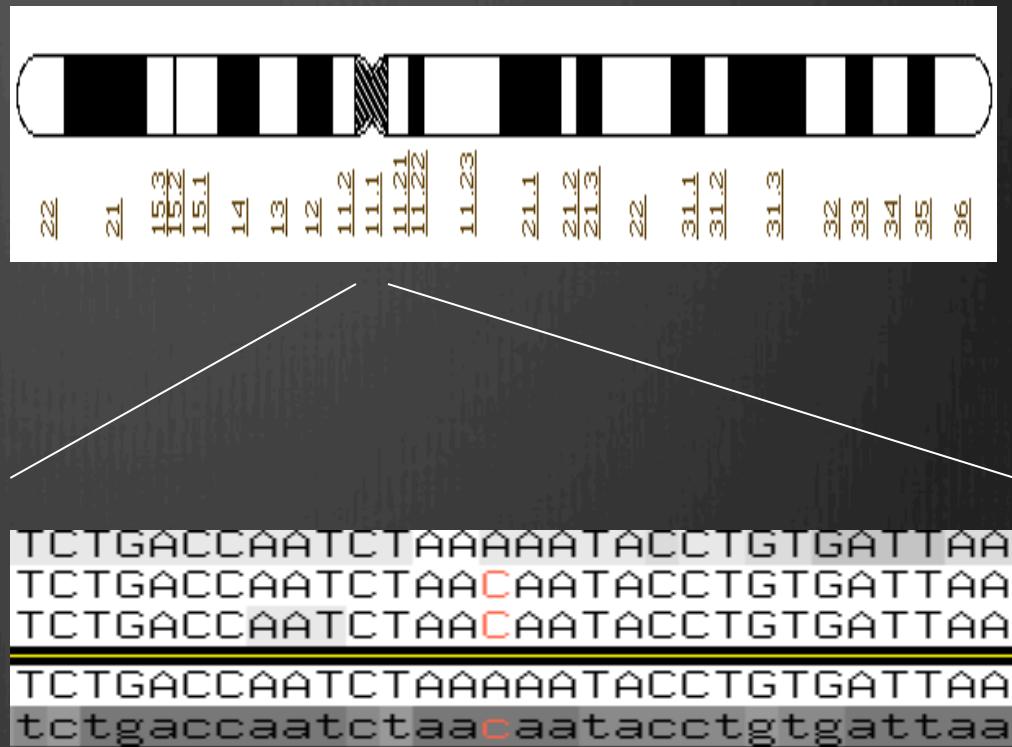
- Small (single-nucleotide) variants are most prevalent:
~4 million in each human genome!
- Larger “structural” variants are also abound

TCTGACCAATCTAAAAATACCTGTGATTAA
TCTGACCAATCTAA**C**AATACCTGTGATTAA
TCTGACCAATCTAA**C**AATACCTGTGATTAA
TCTGACCAATCTAAAAATACCTGTGATTAA
tctgaccaatcta**a**caatac**c**tgtgattaa



The fundamentals of variant discovery

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

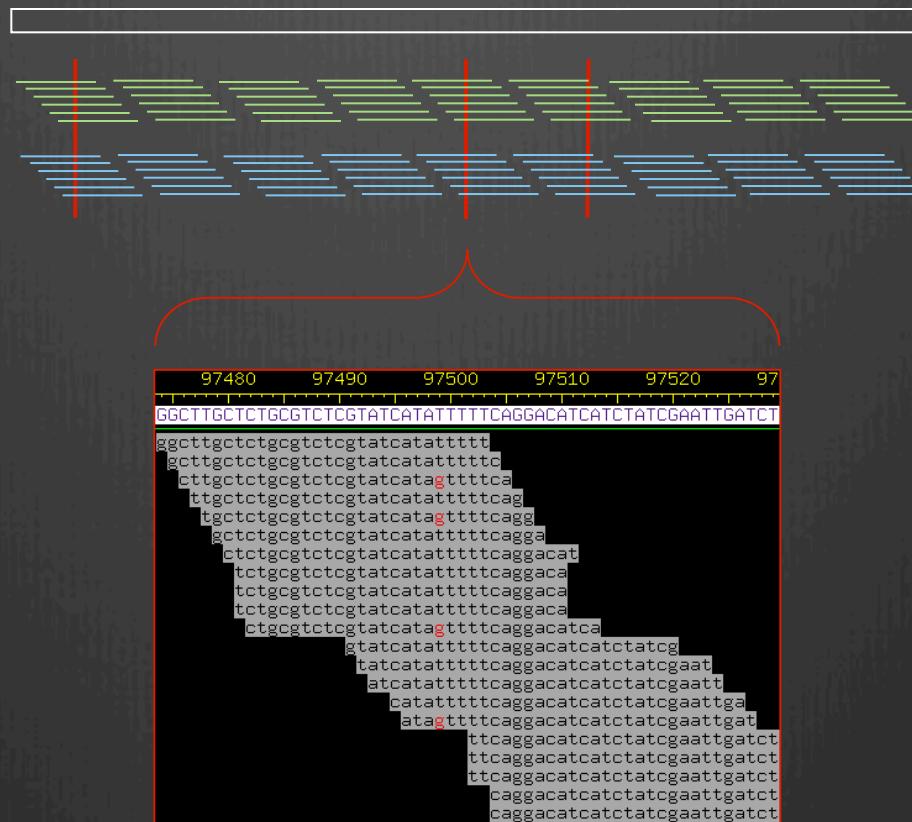


Reference-guided variant discovery

Variant discovery

reference sequence

sequencing reads



SNP calling

97490 97500 97510 97520

GCGTCTCGTATCATATTTCAGGACATCATCTATCG

cgcggtctcgatcatatttt
cgcggtctcgatcatattttc
cgcggtctcgatcatagtttca
cgcggtctcgatcatattttcag
cgcggtctcgatcatagtttcagg
cgcggtctcgatcatattttcagga
cgcggtctcgatcatattttcaggacat
cgcggtctcgatcatattttcaggaca
cgcggtctcgatcatattttcaggaca
cgcggtctcgatcatattttcaggaca
cgcggtctcgatcatagtttcaggacatca
gtatcatattttcaggacatcatctatcg

SNP calling: what goes into it?

Base qualities

TGAAA**Agg**AATT

TGAAA**t**GAATT

TTGAT**CCCTGT**

TTGATT**CCTGT**

sequencing error

true polymorphism

TCTGACCAATCTAAAAATACCTGTGATTAA
TCTGACCAATCTAA**C**AATA**C**ACCTGTGATTAA
TCTGACCAATCTAA**C**AATA**C**ACCTGTGATTAA
TCTGACCAATCTAAAAATACCTGTGATTAA
tctgaccaatctaa**c**aata**c**acctgtgattaa

Base coverage

Prior expectation

Bayesian SNP calling

TCTGACCAATCTAAAAAATACCTGTGATTAA
 TCTGACCAATCTAAACAAATACCTGTGATTAA
 TCTGACCAATCTAAACAAATACCTGTGATTAA

 TCTGACCAATCTAAAAAATACCTGTGATTAA
 tctgaccaatctaacaataaacctgtgattaa

A	C	G
A	C	G
A	C	G
A	C	G
A	C	G



polymorphic
permutation

monomorphic
permutation

Expected polymorphism rate

Bayesian
posterior
probability

Base call +
Base quality

$$P(SNP) = \sum_{all \ variable \ S}$$

$$\frac{P(S_1 | R_1) \cdot \dots \cdot P(S_N | R_N)}{P_{Prior}(S_1) \cdot \dots \cdot P_{Prior}(S_N)}$$

$$P_{Prior}(S_1, \dots, S_N)$$

$$\sum_{S_{i_1} \in [A,C,G,T]} \dots \sum_{S_{i_N} \in [A,C,G,T]} \frac{P(S_{i_1} | R_1)}{P_{Prior}(S_{i_1})} \cdot \dots \cdot \frac{P(S_{i_N} | R_1)}{P_{Prior}(S_{i_N})} P_{Prior}(S_{i_1}, \dots, S_{i_N})$$

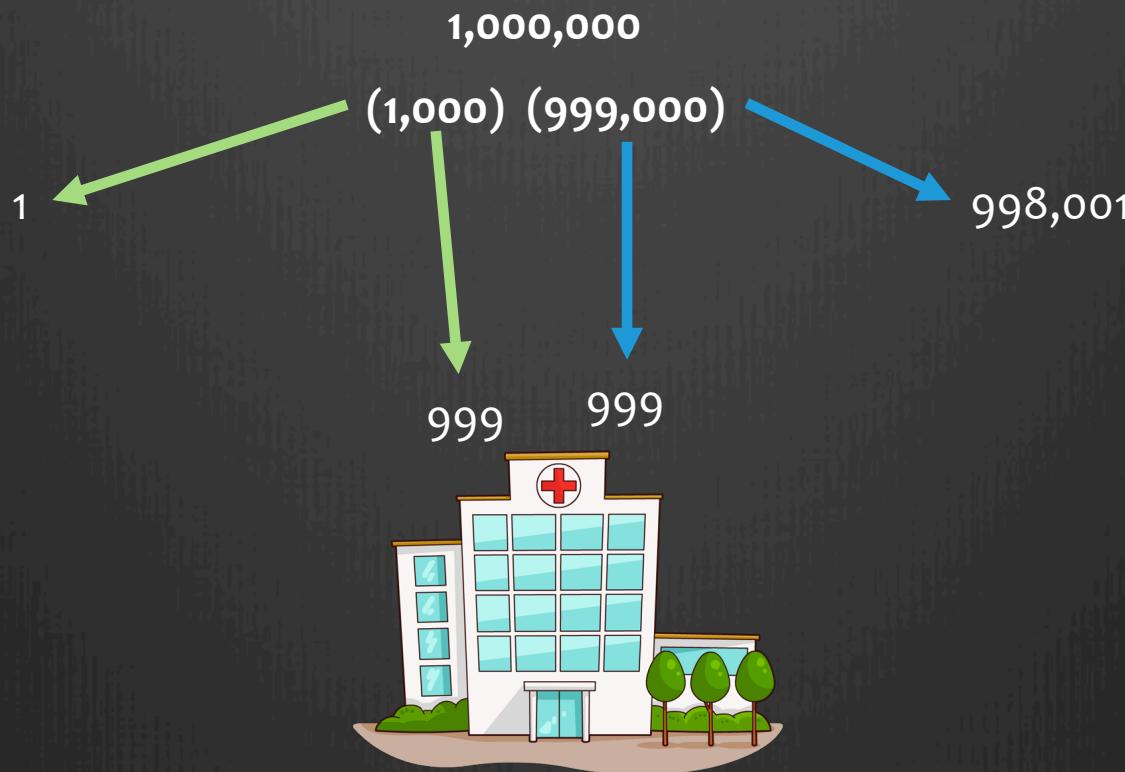
Base composition Depth of coverage

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?

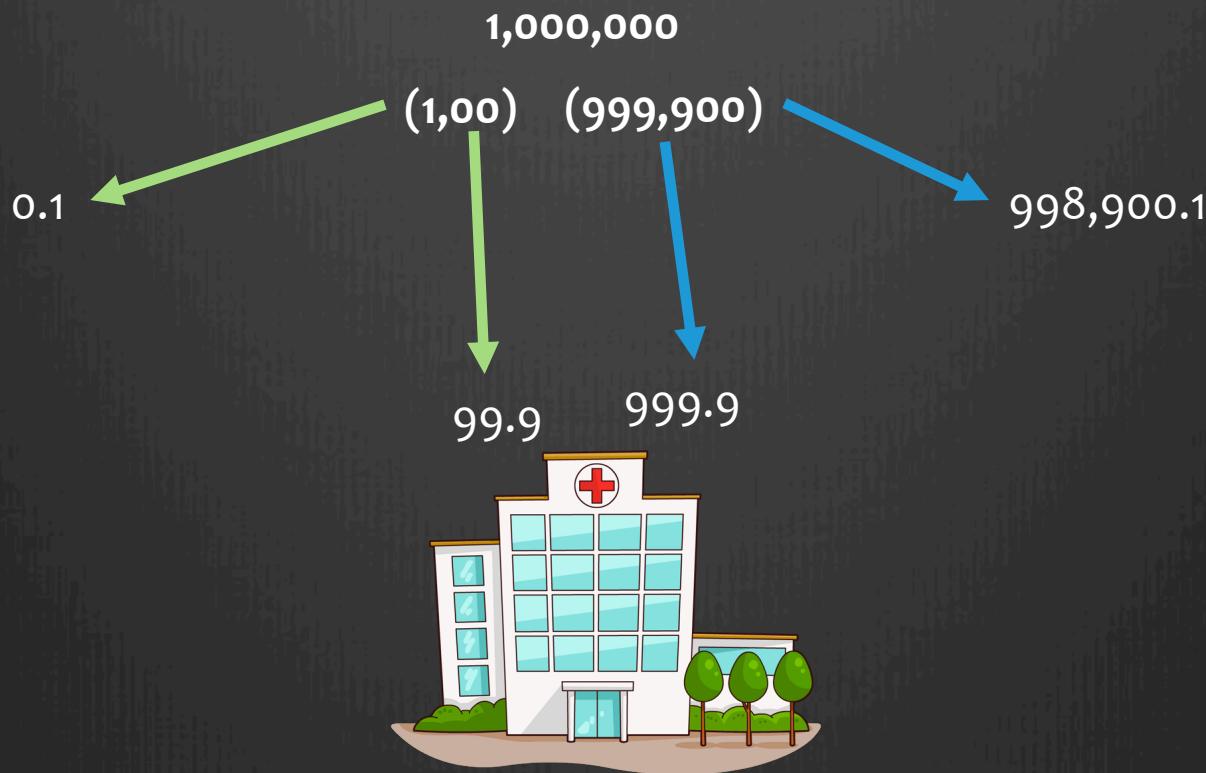


Bayes Example

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?

$$P(\text{sick} | +\text{test})$$

$$P(\text{sick} | +\text{test}) = 99.9\% = 0.999$$

$$P(\text{sick}) = 1/1,000 = 0.001$$

$$P(+\text{test} | \text{sick}) = \frac{P(\text{sick} | +\text{test}) P(\text{sick})}{P(\text{Any Positive})}$$

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

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

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

Bayesian SNP calling

TCTGACCAATCTAAAAAATACCTGTGATTAA
 TCTGACCAATCTAAACAAATACCTGTGATTAA
 TCTGACCAATCTAAACAAATACCTGTGATTAA

 TCTGACCAATCTAAAAAATACCTGTGATTAA
 tctgaccaatctaacaataaacctgtgattaa

A	C	G
A	C	G
A	C	G
A	C	G
A	C	G



polymorphic
permutation

monomorphic
permutation

Expected polymorphism rate

Bayesian
posterior
probability

Base call +
Base quality

$$P(SNP) = \sum_{all \ variable \ S}$$

$$\frac{P(S_1 | R_1) \cdot \dots \cdot P(S_N | R_N)}{P_{Prior}(S_1) \cdot \dots \cdot P_{Prior}(S_N)}$$

$$P_{Prior}(S_1, \dots, S_N)$$

$$\sum_{S_{i_1} \in [A,C,G,T]} \dots \sum_{S_{i_N} \in [A,C,G,T]} \frac{P(S_{i_1} | R_1)}{P_{Prior}(S_{i_1})} \cdot \dots \cdot \frac{P(S_{i_N} | R_1)}{P_{Prior}(S_{i_N})} P_{Prior}(S_{i_1}, \dots, S_{i_N})$$

Base composition Depth of coverage

PolyBayes: the first statistically rigorous variant detection system

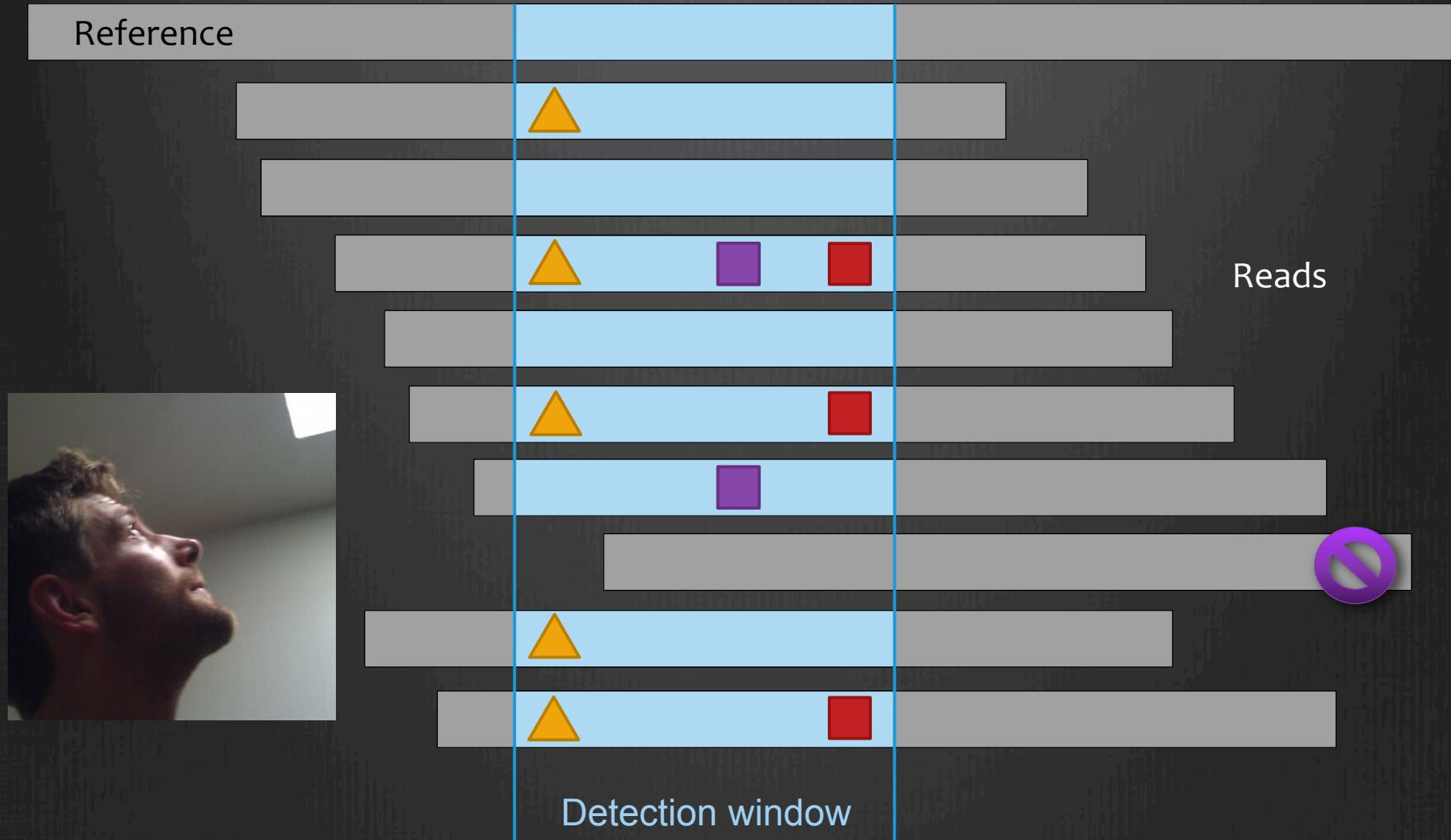
The screenshot shows the PolyBayes web site running in a Netscape Communicator browser window. The title bar reads "Netscape: PolyBayes Web site". The menu bar includes File, Edit, View, Go, Communicator, Help, Back, Forward, Reload, Home, Search, Netscape, Print, Security, Stop, Bookmarks, Location (<http://genome.wustl.edu/gsc/Informatics/polybayes/>), What's Related, WebMail, Radio, People, Yellow Pages, Download, Calendar, Channels. A "Site map" link is visible. The main content area features a portrait of Gabor T. Marth and the text "PolyBayes". Below it is a grid for "letter" analysis with rows numbered 14 to 20. The grid contains various symbols (e.g., -, A, G) and numbers (e.g., 30, 40, 38). Buttons for "Evaluate" and "Resu..." are at the bottom. To the right, a "letter" logo and copyright information ("© 1999 Nature America Inc. • <http://genetics.nature.com>") are displayed. The main title "A general approach to single-nucleotide polymorphism discovery" is centered. Below it is a list of authors: Gabor T. Marth¹, Ian Korf¹, Mark D. Yandell¹, Raymond T. Yeh¹, Zhijie Gu², Hamideh Zakeri², Nathan O. Stitzel¹, LaDeana Hillier¹, Pui-Yan Kwok² & Warren R. Gish¹. The "Results" section displays a table:

Description	Symbol	Value
Probability of SNP	P(SNP)	0.853076589574195
Most likely variation	VAR	A/G
Probability of variation	P(VAR)	0.853003076184499
Alignment depth	D	2

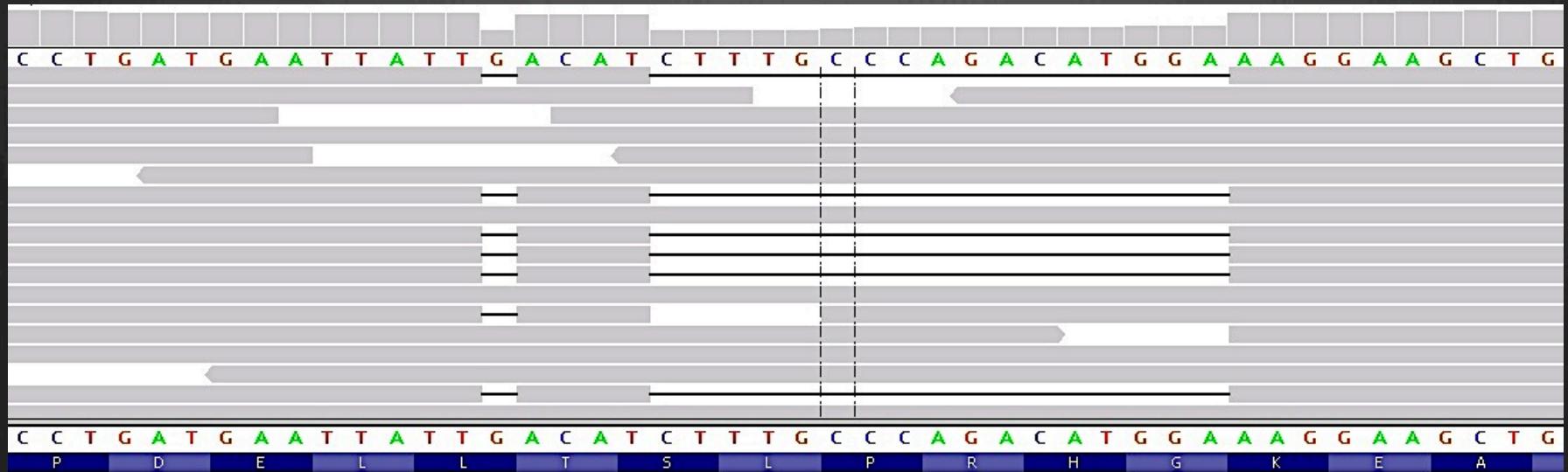
Comments to Gabor Marth, gmarth@genome.wustl.edu, Washington University Genome Sequencing Center.
Last modified: Mon Feb 12 17:06:10 2001

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

The FreeBayes local haplotype caller

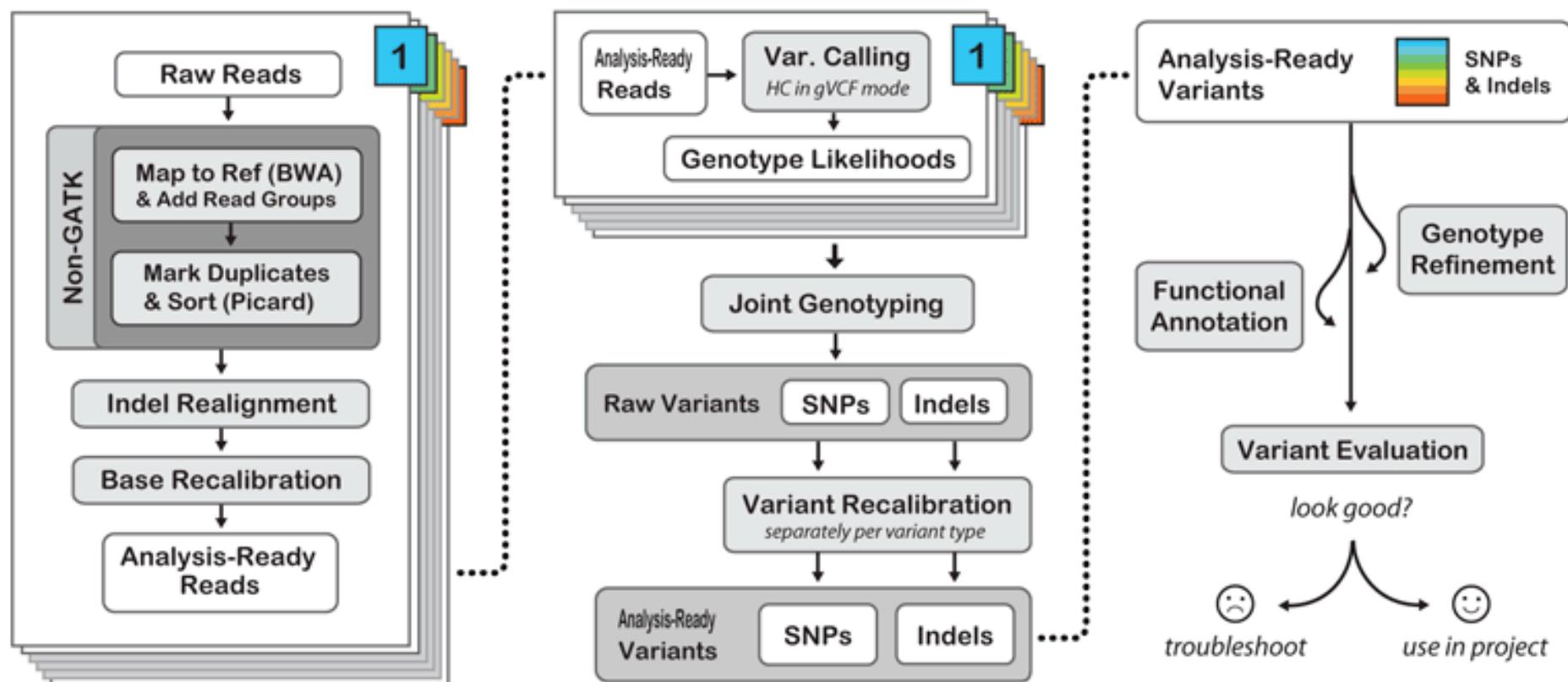


Able to identify compound hets



- ⌚ 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	
	<u>GATK</u>	<u>freebayes (q≥250)</u>	<u>GATK</u>	<u>freebayes (q≥0)</u>
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

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

THIS NEUTRINO DETECTOR MEASURES
WHETHER THE SUN HAS GONE NOVA.

(THEN, IT ROLLS TWO DICE. IF THEY
BOTH COME UP SIX, IT LIES TO US.
OTHERWISE, IT TELLS THE TRUTH.

LET'S TRY.

DETECTOR! HAS THE
SUN GONE NOVA?



FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT
HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.
SINCE $p < 0.05$, I CONCLUDE
THAT THE SUN HAS EXPLODED.



BAYESIAN STATISTICIAN:

BET YOU \$50
IT HASN'T.

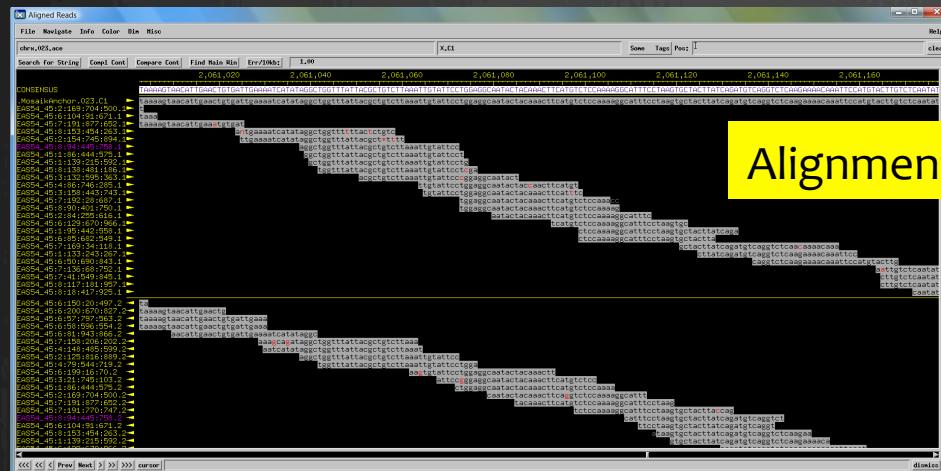


Finer points

Data formats

```
@IL11_266:1:1:395:231/1
CCAACCACAAACACAAAAAACACAAGCAACGACCC
+
@AAAAAA?<>@@>?:475;A6?384,>53>14<>
@IL11_266:1:1:399:301/1
CAAAAAAAAAAGAAGTACGAGATACGACACATCAC
+
;@AAAAA>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 0:48:1:51.51	1 0:48:8:51.51
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

$$\left. \begin{array}{l} P(B_i=aac|G_i=aa) \\ P(B_i=aac|G_i=cc) \\ P(B_i=aac|G_i=ac) \end{array} \right\}$$

$$\left. \begin{array}{l} P(G_i=aa|B_1=aacc; B_i=aaaaac; B_n=cccc) \\ P(G_i=cc|B_1=aacc; B_i=aaaaac; B_n=cccc) \\ P(G_i=ac|B_1=aacc; B_i=aaaaac; B_n=cccc) \end{array} \right\}$$



a

a

a

c

$$\left. \begin{array}{l} P(B_i=aaaa|G_i=aa) \\ P(B_i=aaaa|G_i=cc) \\ P(B_i=aaaa|G_i=ac) \end{array} \right\}$$

Prior($G_1, \dots, G_i, \dots, G_n$)

$$\left. \begin{array}{l} P(G_i=aa|B_1=aacc; B_i=aaaaac; B_n=cccc) \\ P(G_i=cc|B_1=aacc; B_i=aaaaac; B_n=cccc) \\ P(G_i=ac|B_1=aacc; B_i=aaaaac; B_n=cccc) \end{array} \right\}$$



c

c

c

c

$$\left. \begin{array}{l} P(B_n=cccc|G_n=aa) \\ P(B_n=cccc|G_n=cc) \\ P(B_n=cccc|G_n=ac) \end{array} \right\}$$

$$\left. \begin{array}{l} P(G_n=aa|B_1=aacc; B_i=aaaaac; B_n=cccc) \\ P(G_n=cc|B_1=aacc; B_i=aaaaac; B_n=cccc) \\ P(G_n=ac|B_1=aacc; B_i=aaaaac; B_n=cccc) \end{array} \right\}$$

“genotype
likelihoods”

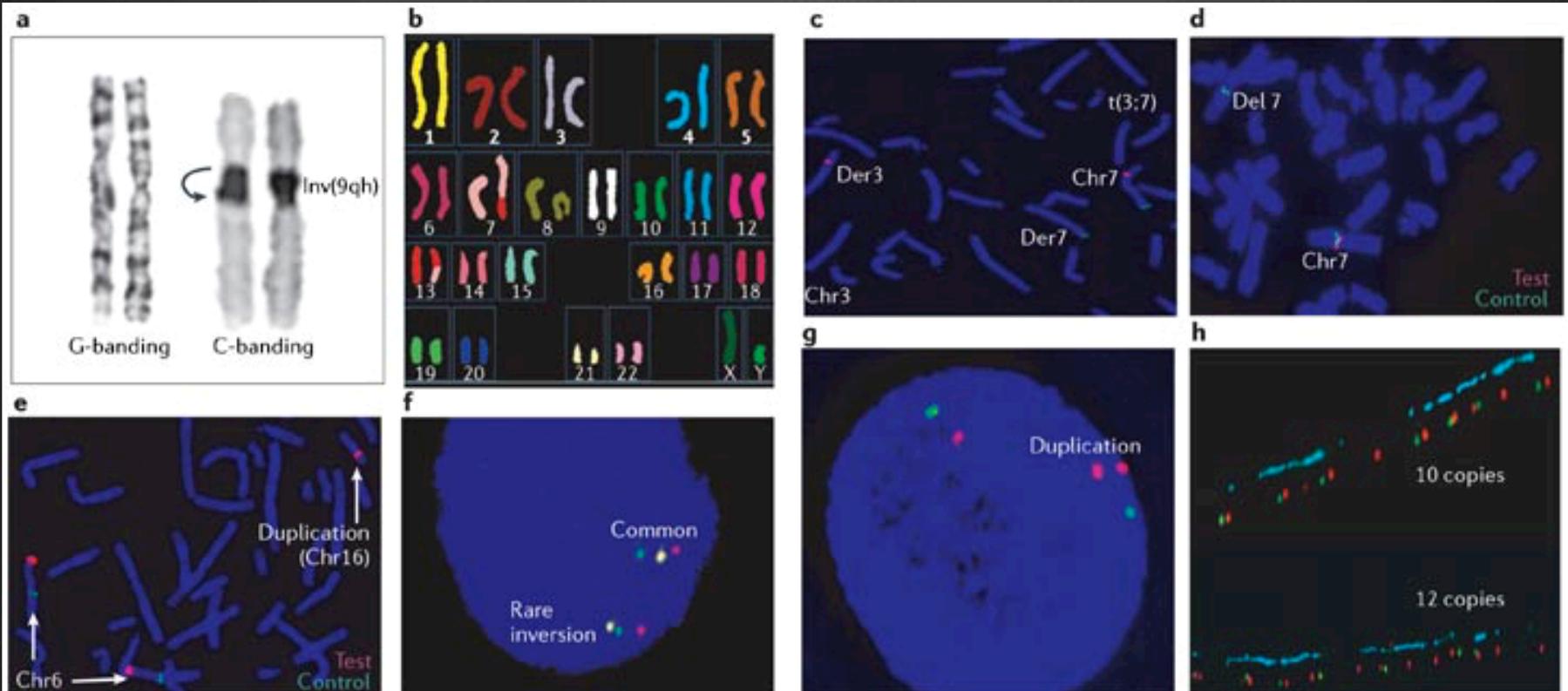
“genotype
probabilities”



$$P(\text{SNP})$$

Detecting structural variations

Structural variations require different approaches



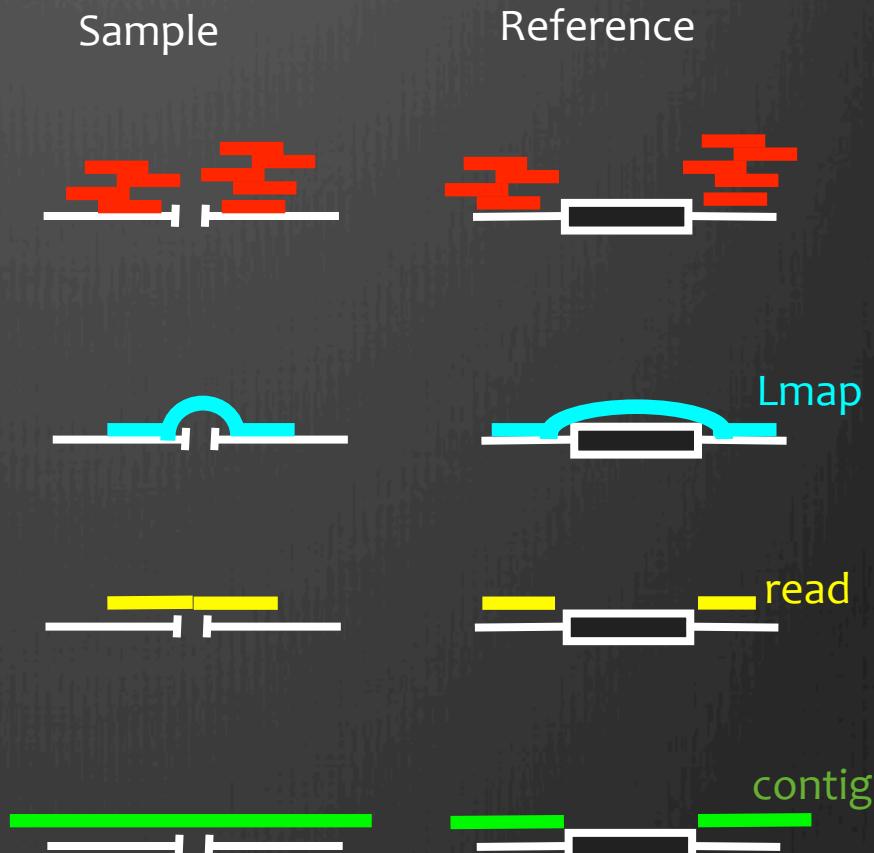
Copyright © 2006 Nature Publishing Group
Nature Reviews | Genetics

Short variants are detected directly from read “multiple alignments”

```
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTCAGGG*TCTCC*ATAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTCAGGG*TCTCC*ATAAAGAT
*tt*act*gtaatggaataactcatgaagtgttaaggggctaaaaagaagcctccggcctt
GTT*ACT*GtcGTTGT*AA*TACTCC*a*cgatgtCTTCAGGG*tctcc*atAAAGat
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTCAGGG*TCTCC*ATAAAGAT
tgt*act*gaagttgc*aa*tactCc*a*cgATGTcttcAGGG*TCTcc*aATAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTCAGGG*TCTCC*ATAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CAATGTCTTCAGGG*TCTCC*ATAAAGAT
gttaact*gtcgttgt*aaatactcc*aacgatgtCTtTcaggg*TctcccataAAagat
GTT*a*t*gTCGTTGT*AA*TACTCC*A*CAATGTCTTCAGGG*TCTCC*ATAAAGAT
gtt*act*gTCGTTgttaa*tactccc*a*cgatgTCTttcaggg*TCTcccataaaagat
gtt*act*gtcgttgt*aa*aactccc*a*caatgtcttcaggg*tctcc*ataAAagat
GTT*Act*gtcgtTGt*aa*tacTcc*a*caatgtcttcAGgg*tcTCC*ATAAAGAT
gtt*act*gtcgttGt*AA*TAcTcc*A*CGATgtcttcaggg*TCTcc*aATAAAGAT
gtt*acg*atcggtGt*aa*taatcc*a*cgatGtctgtcaggG*Tctcc*ataaaagat
GTt*actcgacgttgt*aa*tacTcc*a*caatgtctatCAGGG*TCTCC*ATAAAGAT
gtt*aat*gtcgttgt*aaatactccc*a*caatgtCTttcaggggtctcccataaaagat
Gtt*act*gTCGTTGt*aa*tatccc*a*caatGTCTTCAGGG*TCTCC*ATAAAGAT
GTT*ACT*GTCGTTGT*AA*TACTCC*A*CGATGTCTTCAGGG*TCTCC*ATAAAGAT
gtt*act*gTCgttgt*AA*TACTCc*a*cAtgtcttcaggg*tctcc*ATAaaaGAT
```

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



The Lumpy SV caller (Layer, Quinlan)

lumpy

Layer et al, unpub.

Depth of coverage



Paired-end mapping



Split-read mapping



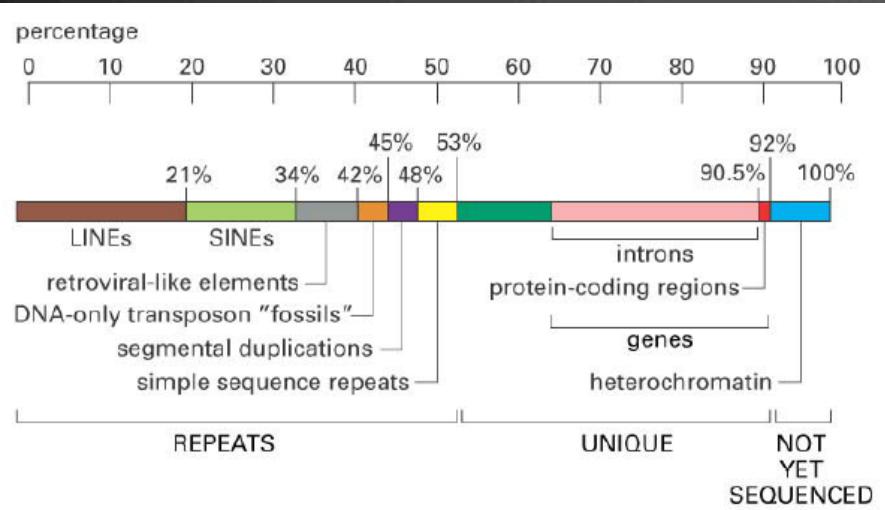
Prior knowledge

Known SV sites

Predictions from other tools

LUMPY integrates **all** (and future) signals

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 Human

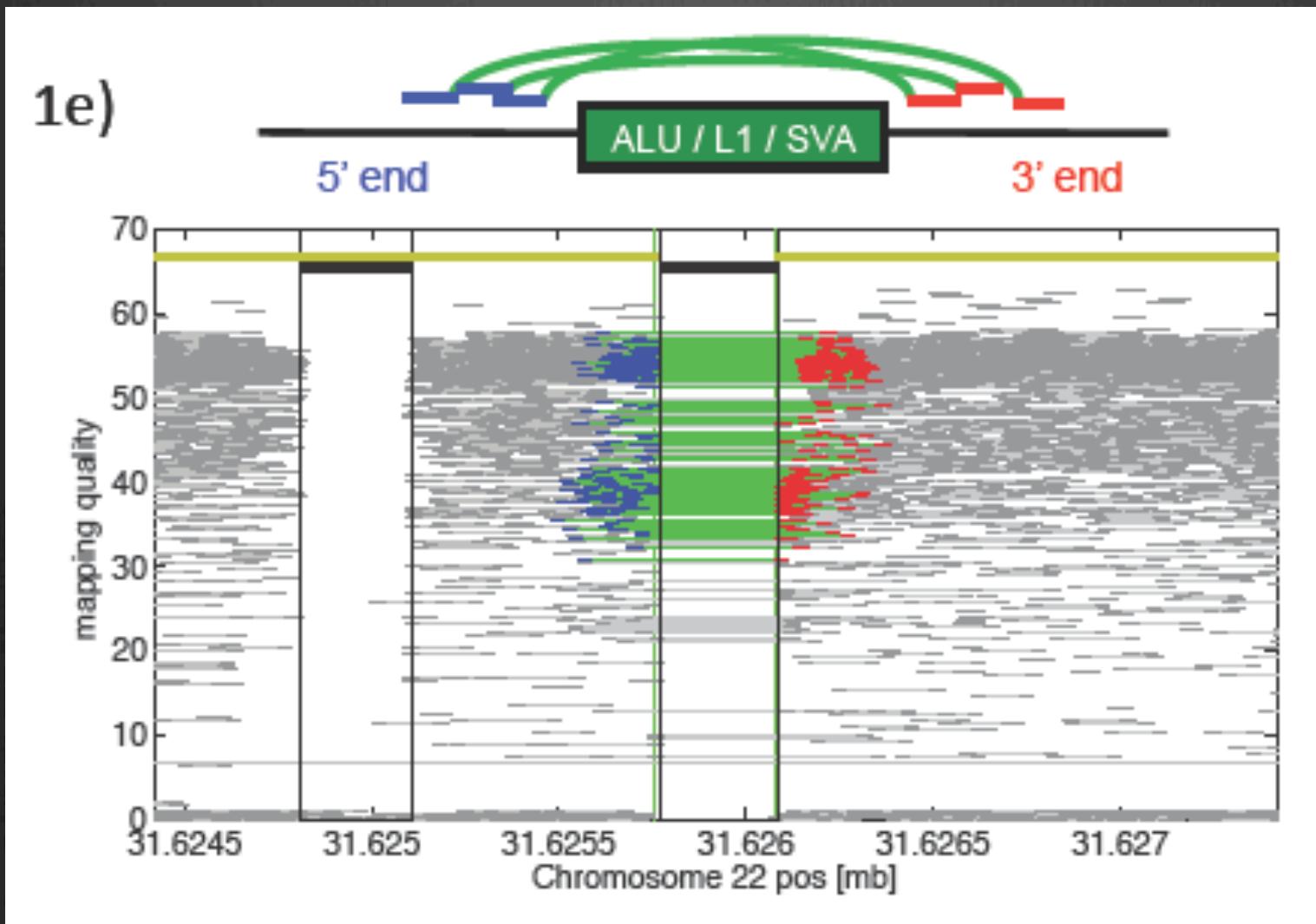
Samuel Levy^{1*}, Granger Sutton¹, Pauline C. Ng¹, Lan Jiaqi Huang¹, Ewen F. Kirkness¹, Gennady Denisov¹, Mary Shago², Timothy B. Stockwell¹, Alexia Tsiamou Karen Y. Beeson¹, Tina C. McIntosh¹, Karin A. Remin Marvin E. Frazier¹, Stephen W. Scherer², Robert L. St

1 J. Craig Venter Institute, Rockville, Maryland, United States of America, 2 Medical Genetics, University of Toronto, Toronto, Ontario, Canada, 3 De California, United States of America, 4 Genetics Department, Facultat de I

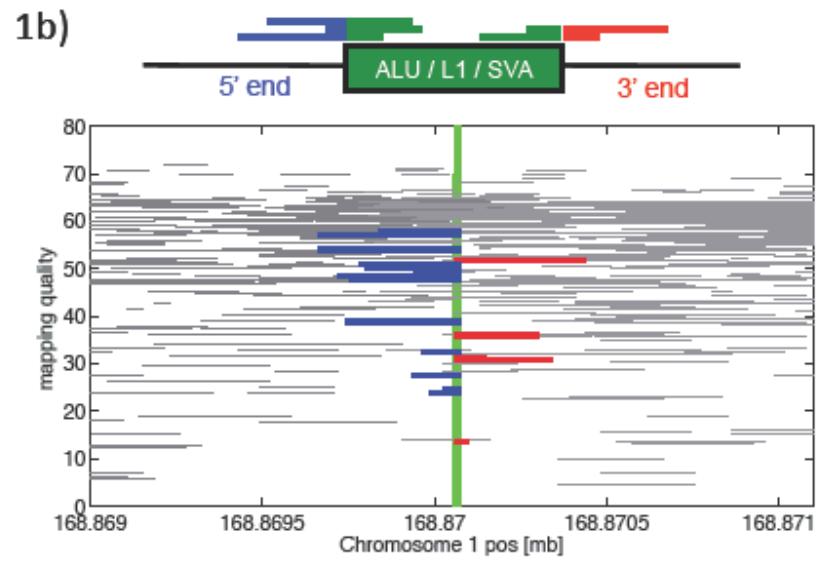
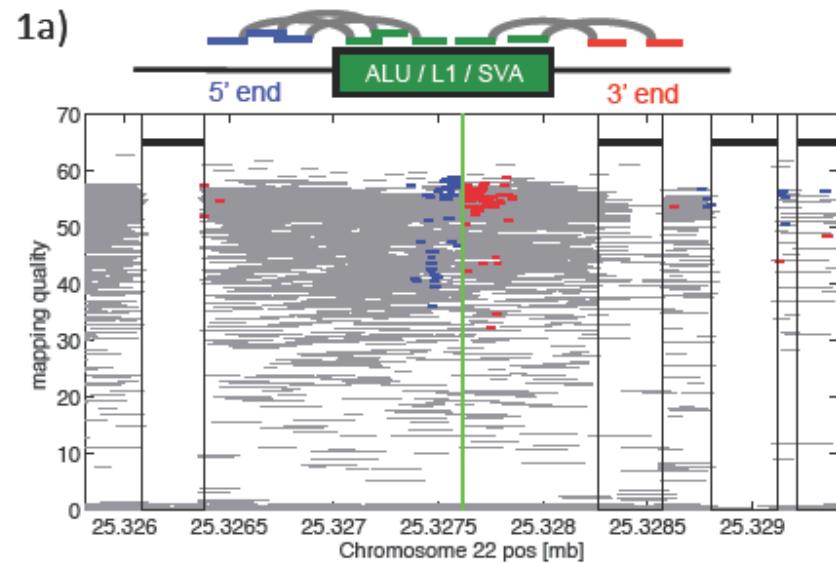
Presented here is a genome sequence of an individual fragments, sequenced by Sanger dideoxy technology bases (Mb) of contiguous sequence with approximately modified version of the Celera assembler to facilitate individual diploid genome. Comparison of this genome reference assembly revealed more than 4.1 million 1,288,319 were novel) included 3,213,401 single nucleotide, 292,102 heterozygous insertion/deletion events inversions, as well as numerous segmental duplications account for 22% of all events identified in the donor important role for non-SNP genetic alterations in de



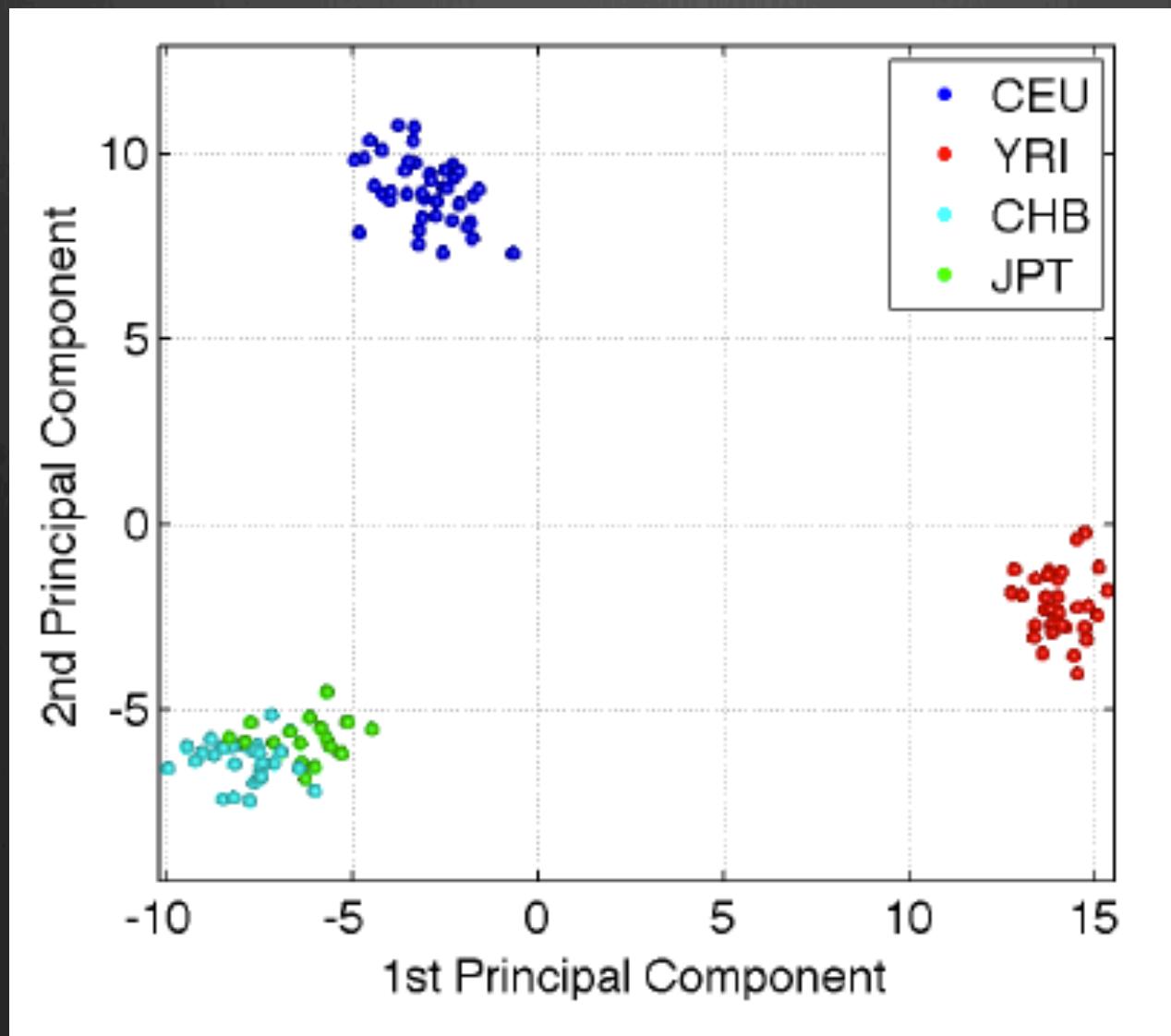
Detection previously restricted to ref. insertions



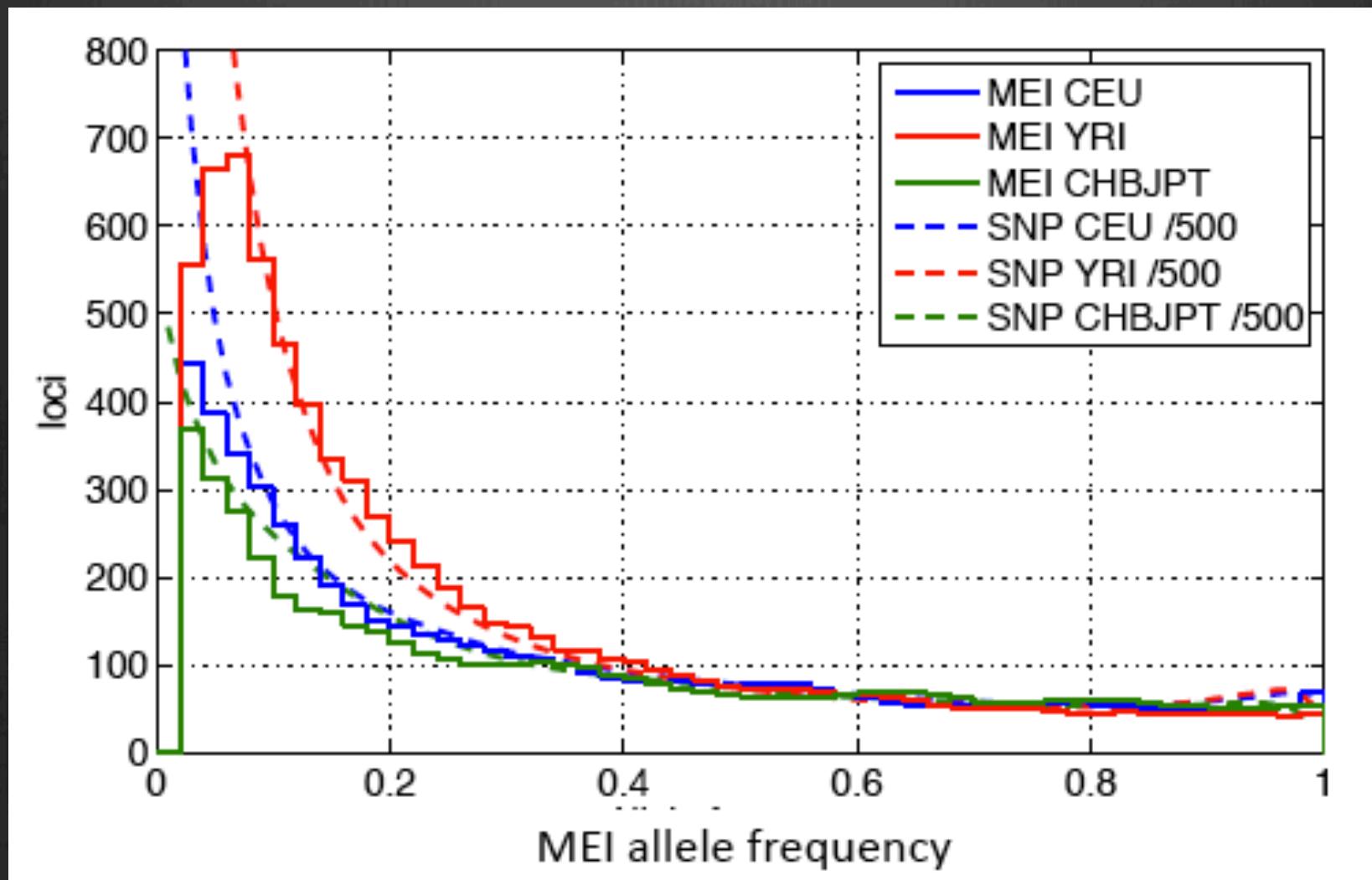
Direct detection of ME insertion events



MEIs cluster according to populations (like SNPs)



MEI allele frequency also similar to SNPs

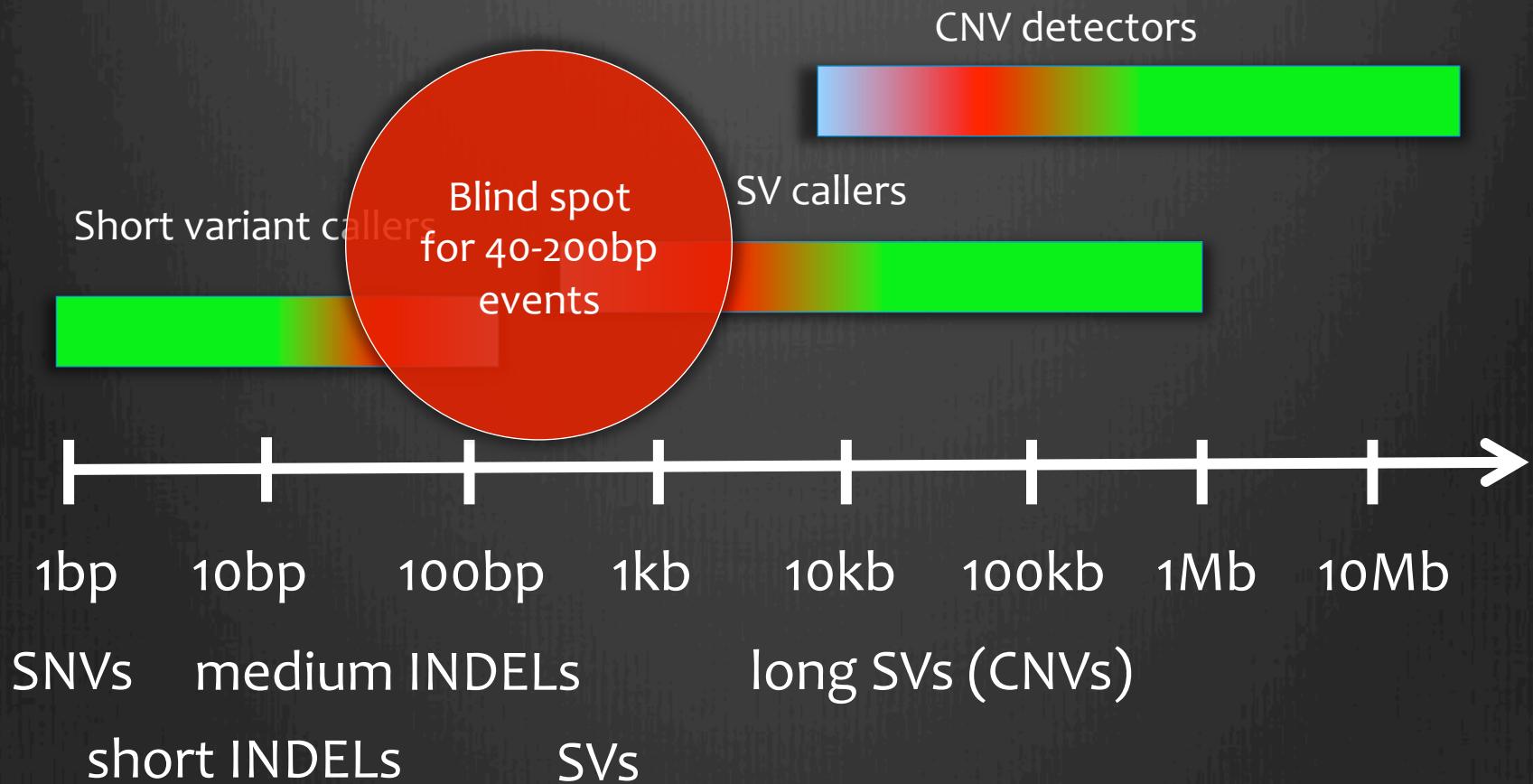


Reference-free variant detection

Reference-guided methods have limitations for highly diverged variants

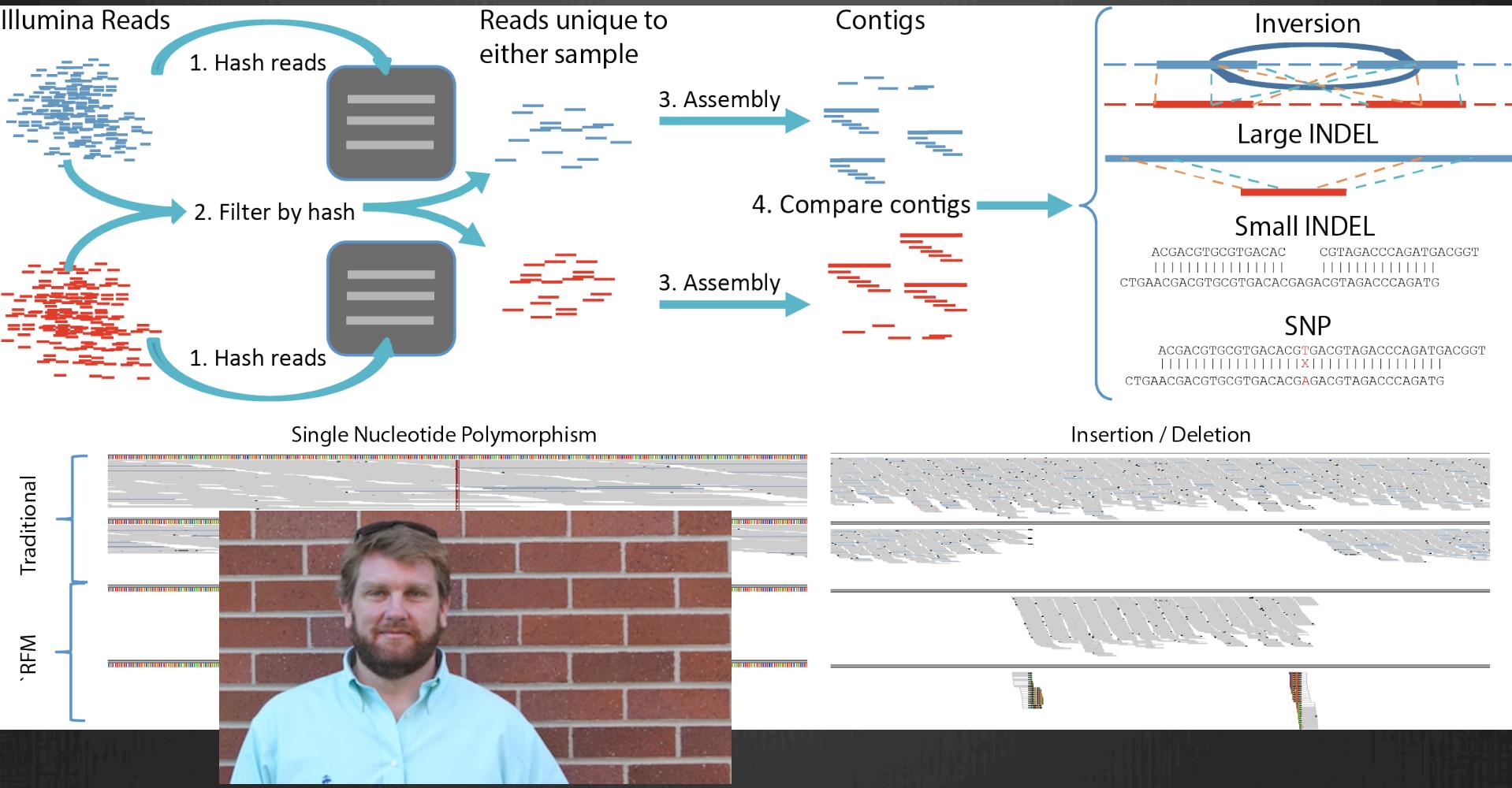


Detection sensitivity low for medium-length events

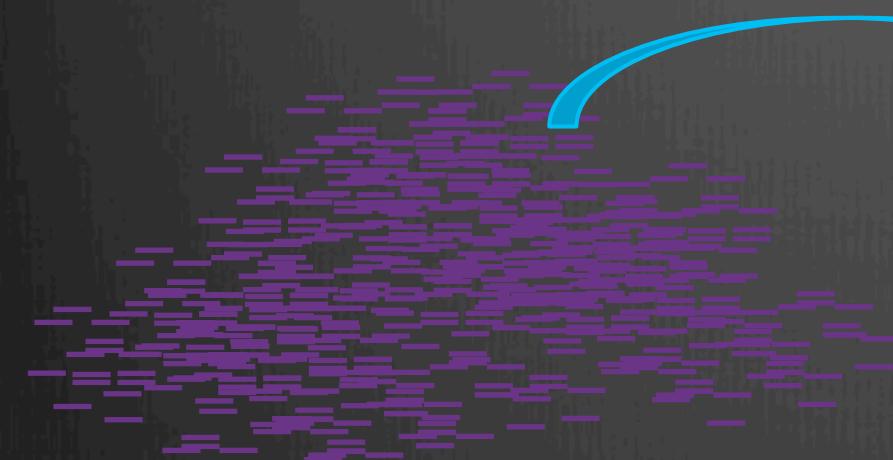


Reference-free methods compare genomes directly
(without first aligning data to a reference sequence)

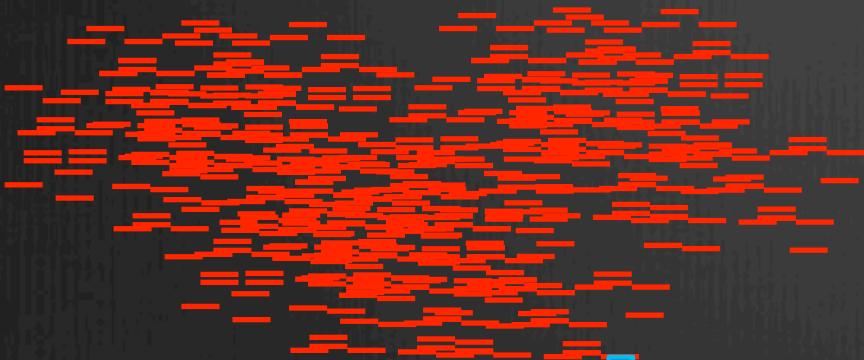




1: Use reads to create “hash tables”



46	AGTCGTCGCCTTGCTT
62	TTGAGGAGTCATCTCGAC
59	AAAGCTATGGCAGATATC
58	GTCAAGAGAGAGAAAAGG
61	AACAGGATGAAACCCGGA
59	ACAGAGCGCCGCGACGGC
74	AGAACACACAGTTCGAGCG
78	GTCTTCTCCTTCTCTTC
56	GTGCAAATGCGCAGACT
54	GAACACGATGGGCGGAGG
64	GCTAAGAGAGAGAAAAGG
⋮	⋮
⋮	⋮



46	AGTCGTCGCCTTGCTT
62	TTGAGGAGTCATCTCGAC
59	AAAGCTATGGCAGATATC
58	GTCAAGAGAGAGAAAAGG
61	AACAGGATGAAACCCGGA
59	ACAGAGCGCCGCGACGGC
74	AGAACACACAGTTCGAGCG
78	GTCTTCTCCTTCTCTTC
56	GTGCAAATGCGCAGACT
54	GAACACGATGGGCGGAGG
64	GCTAAGAGAGAGAAAAGG
⋮	⋮
⋮	⋮

True Variants

46	AGTCGTGCCTCTTGCTT
62	TTGAGGAGTCATCTCGAC
59	AAAGCTATGGCAGATATC
58	GTCAAGAGAGAGAAAAGG
61	AACAGGATGAAACCCGGA

34

...GACACATTAATGAGAACACACACATATGCGGCGCAGACCTTACACACCGAAC

selected

39

46	AGTCGTGCCTCTTGCTT
62	TTGAGGAGTCATCTCGAC
59	AAAGCTATGGCAGATATC
58	GTCAAGAGAGAGAAAAGG
61	AACAGGATGAAACCCGGA

Sequencing Error

46	AGTCGTCGCCTCTTGCTT
62	TTGAGGAGTCATCTCGAC
59	AAAGCTATGGCAGATATC
58	GTCAAGAGAGAGAAAAGG
61	AACAGGATGAAACCCGGA

34

...TACATGATGAAACCCGGAACAGGATGAAACCCGGA

Sequence Error
&
Uninformative Reads

20

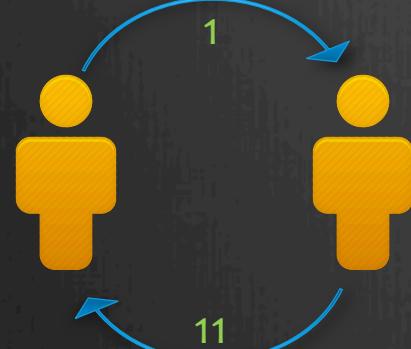
6	AGTCGTCGCCTCTTGCTT
62	TTGAGGAGTCATCTCGAC
59	AAAGCTATGGCAGATATC
58	GTCAAGAGAGAGAAAAGG
61	AACAGGATGAAACCCGGA

A *de novo* deletion detected by Rufus

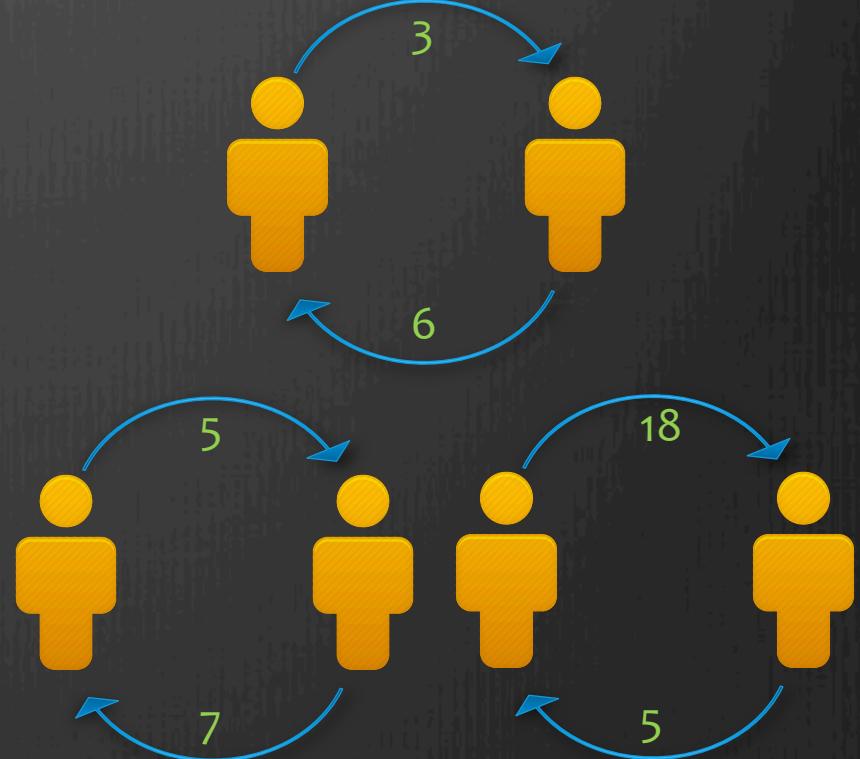


High specificity for variation in MZ twin studies

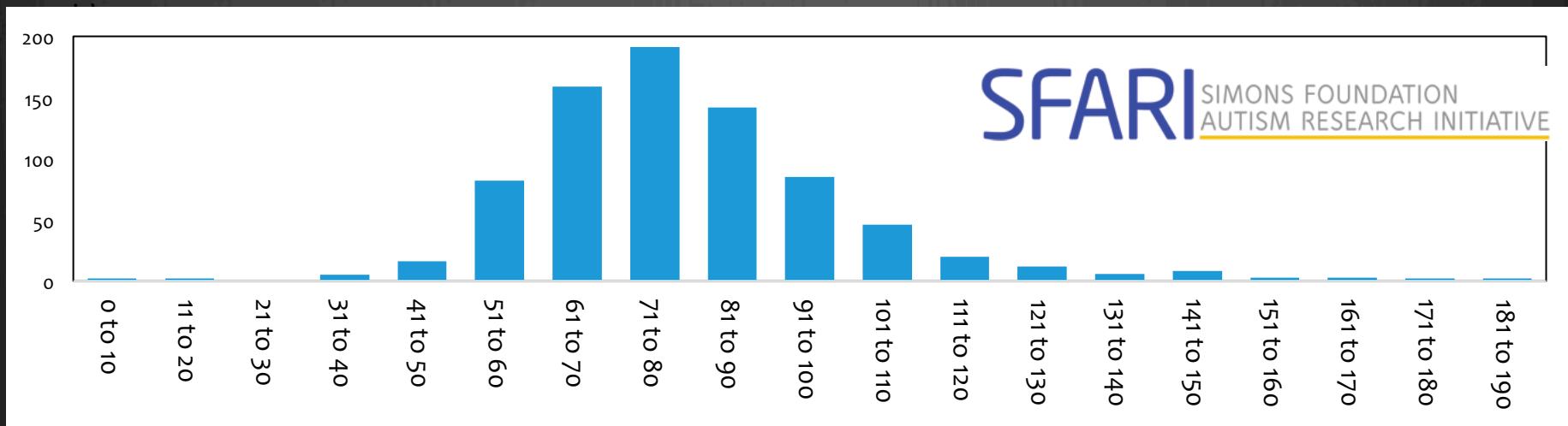
60X Illumina Sequencing



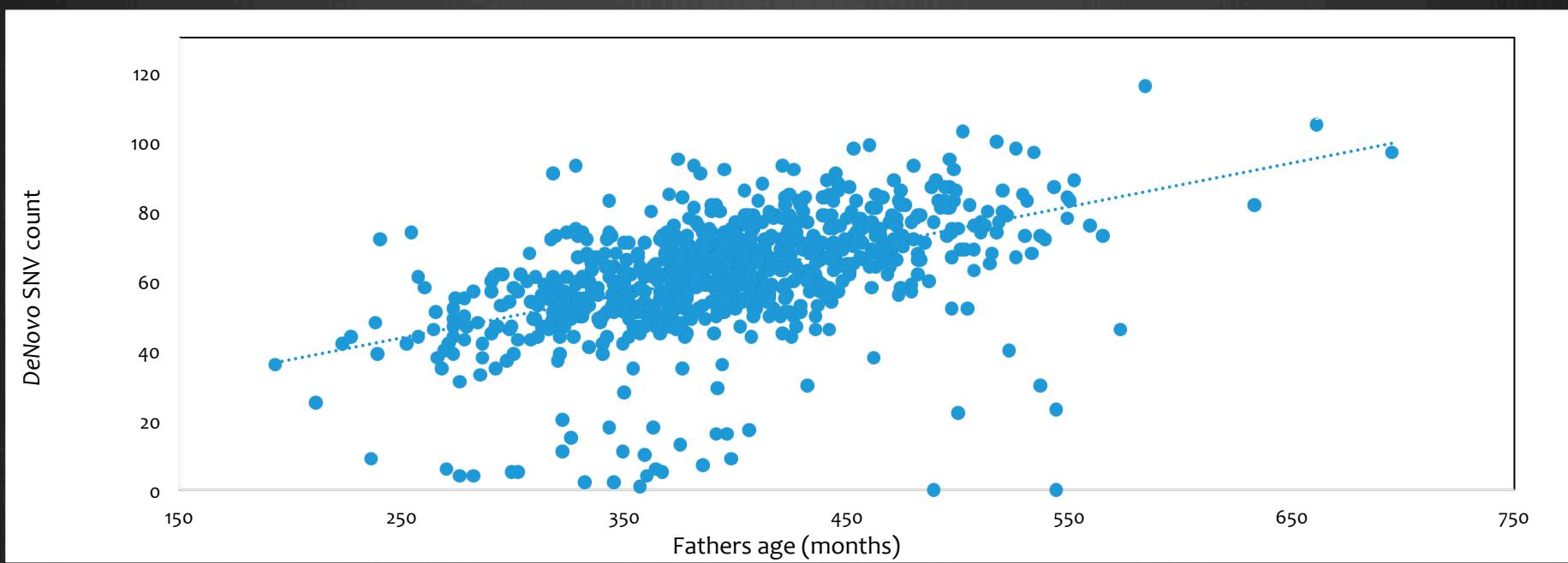
30X Illumina sequencing
+
1000G k-mer filter



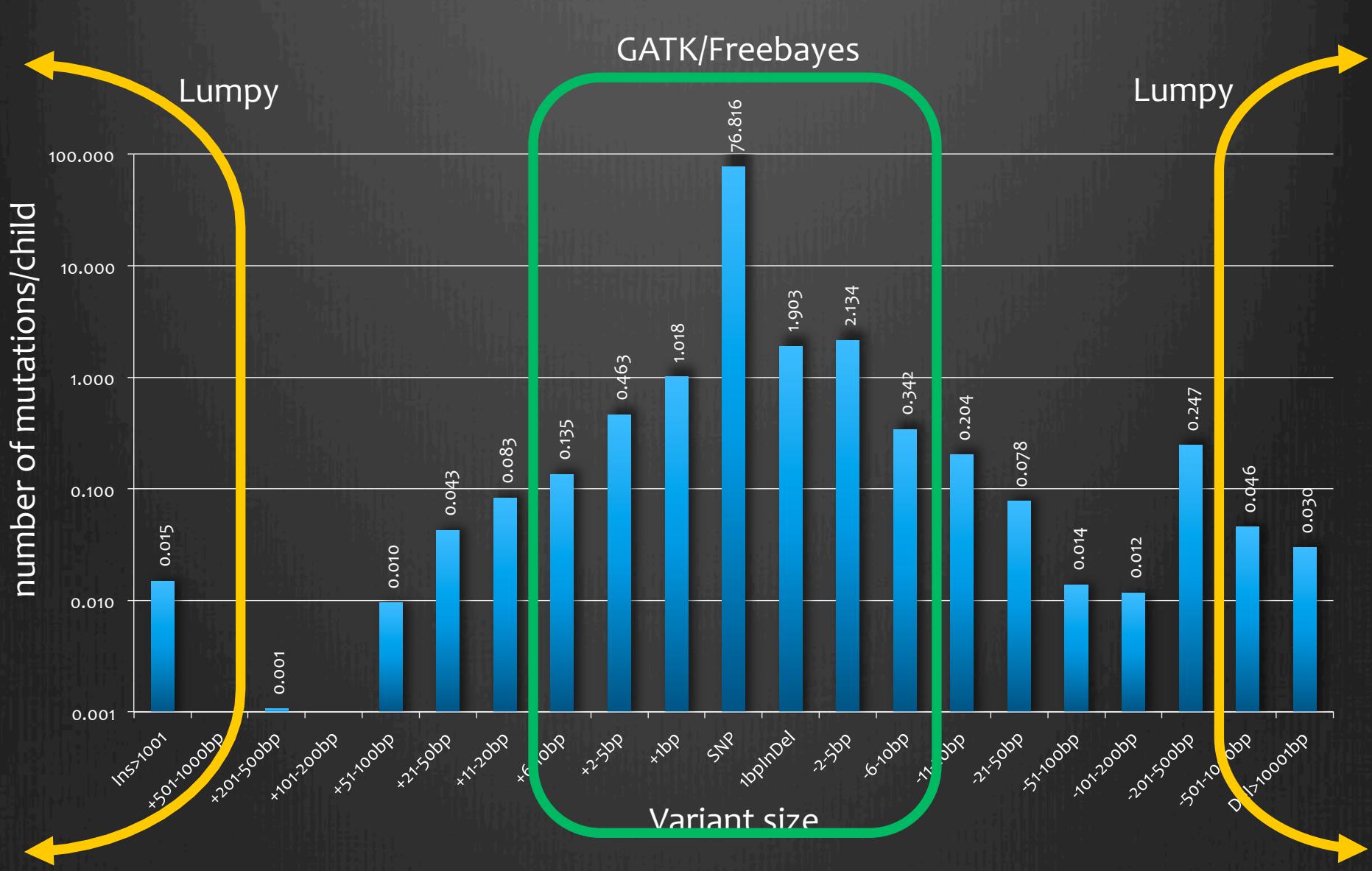
Application to 500 SFARI quartets



SFARI SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE



Detecting *de novo* variants of all sizes



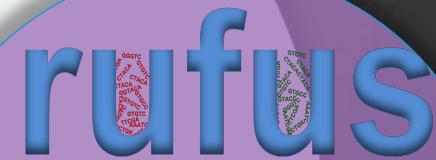
Clinical *de novo* variant detection

- 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



111,803



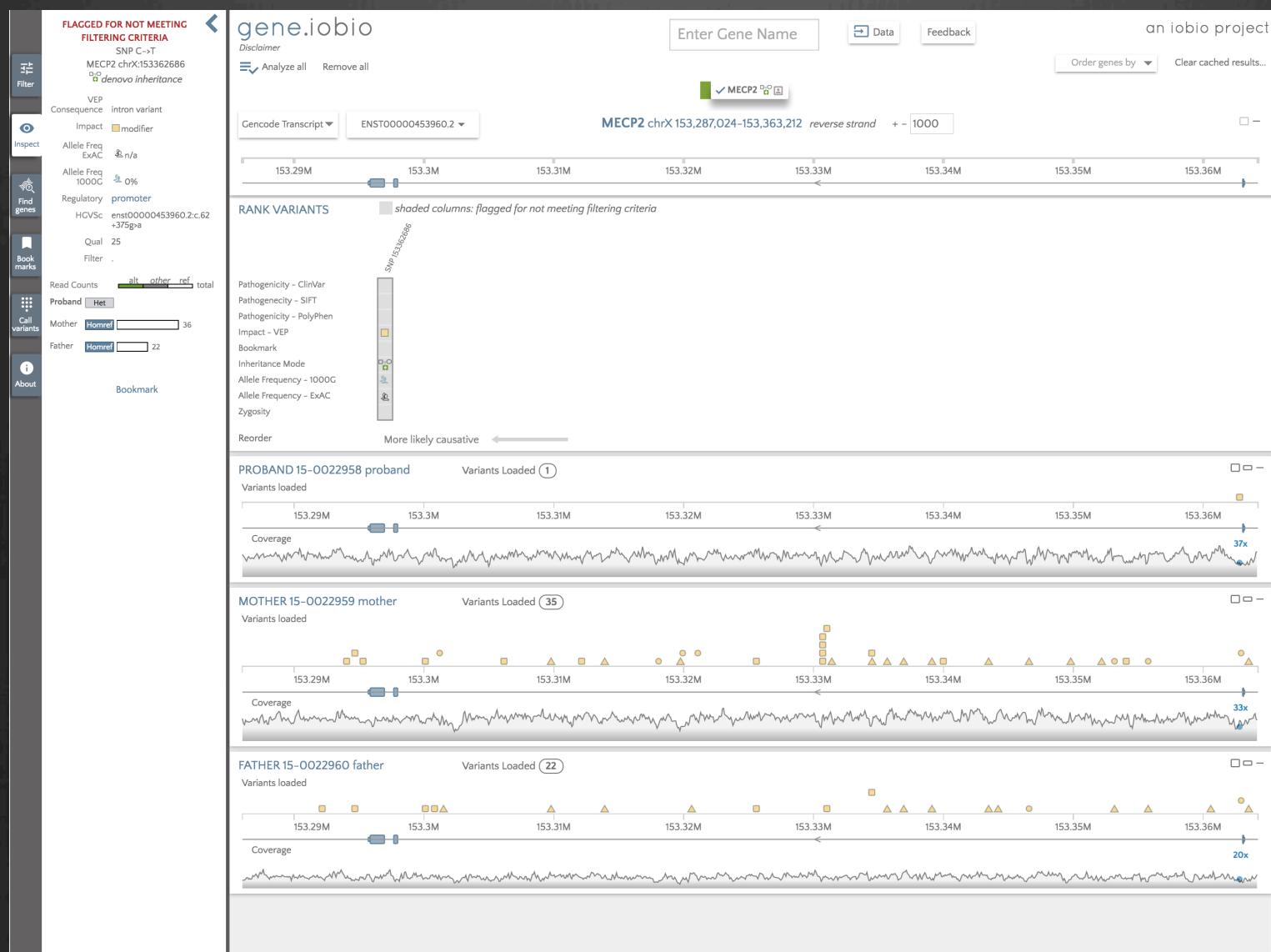
65
(85%)

12
(16.3%)

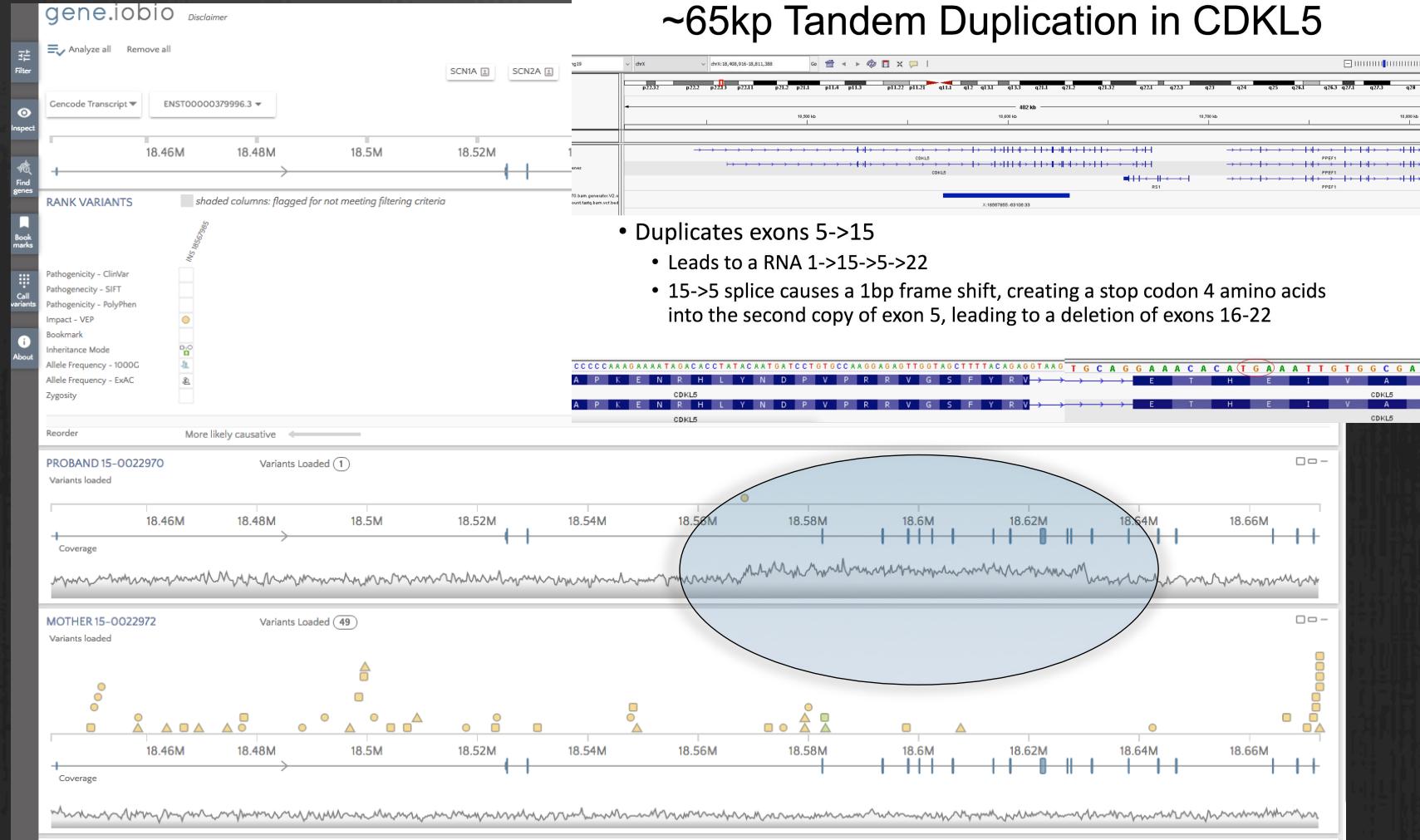
Many false positives, even in coding regions



Confidently detecting non-coding de novos



Confidently detecting de novo SVs





That's all Folks!