Using high-throughput DNA sequencing and bioinformatics to search for disease mutations

Vikas Bansal, Ph.D.
Department of Pediatrics

MED263, March 14th 2017

High throughput DNA sequencing

2005

Roche 454
100 million bp

2008

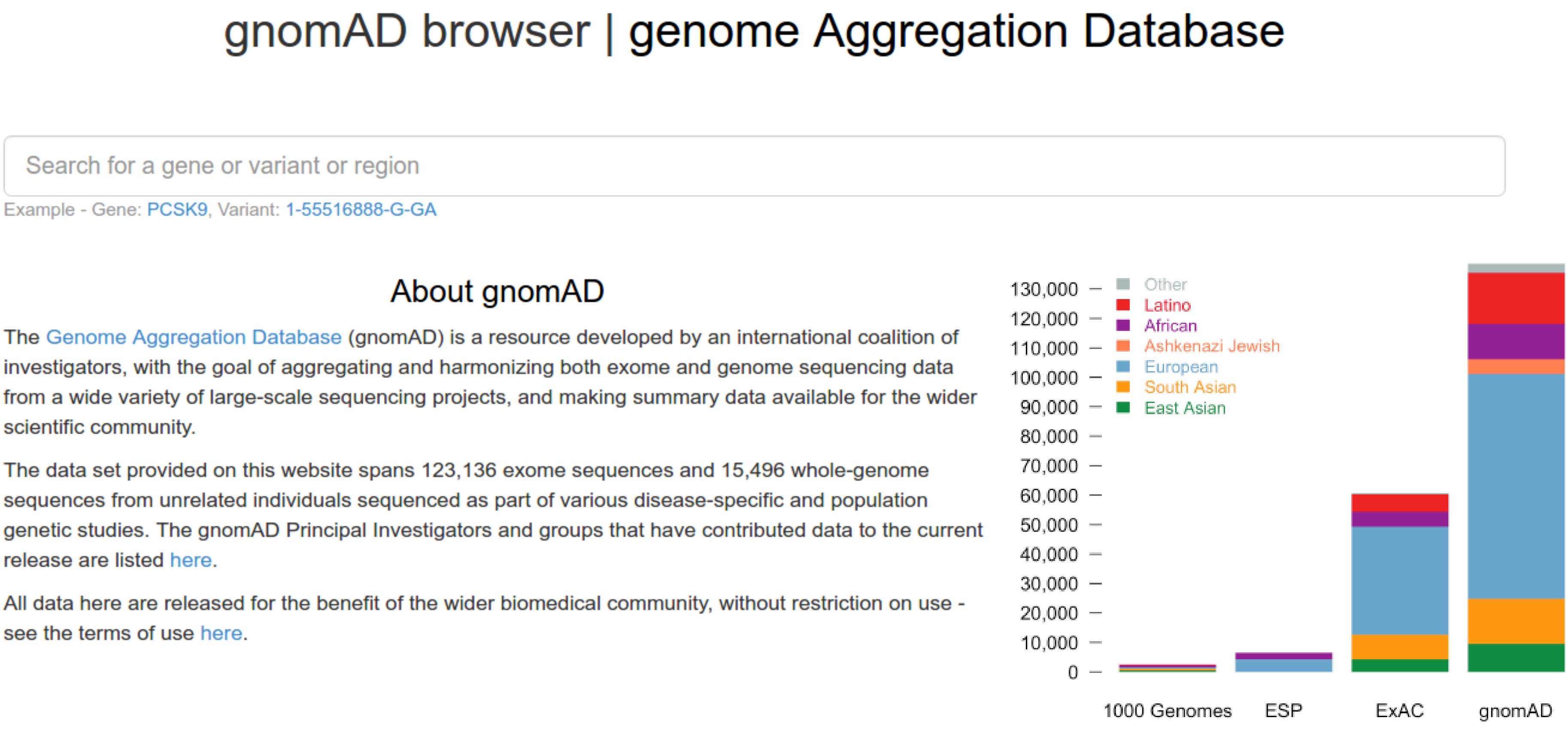
Solexa instrument
1 Gb per run

2015

Illumina HiSeq
600-1000 Gb per run

- 10,000 fold increase in throughput of sequencing technologies
- **Sequencing of human genomes has become routine**

Sequence data from > 100,000 individuals



gnomAD browser | genome Aggregation Database

Search for a gene or variant or region

Example - Gene: PCSK9, Variant: 1-55516888-G-GA

About gnomAD

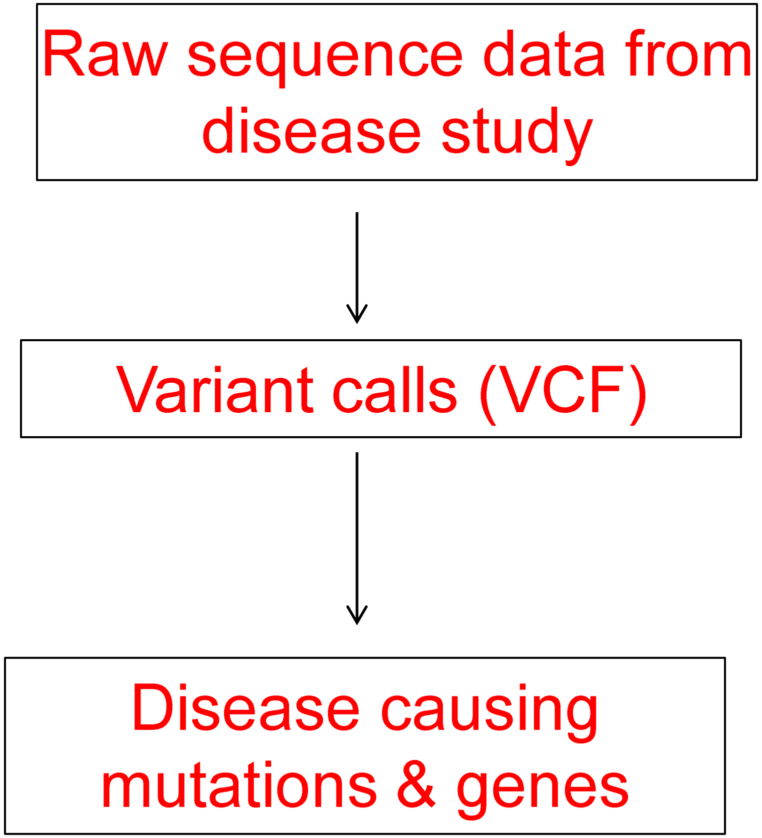
The Genome Aggregation Database (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.

The data set provided on this website spans 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The gnomAD Principal Investigators and groups that have contributed data to the current release are listed here.

All data here are released for the benefit of the wider biomedical community, without restriction on use - see the terms of use here.

http://gnomad.broadinstitute.org

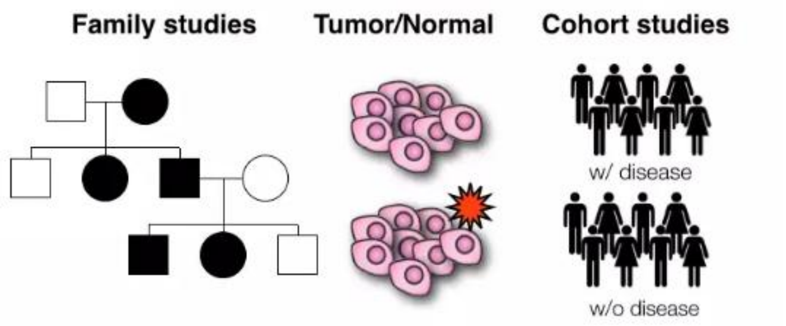
DNA sequencing to understand disease

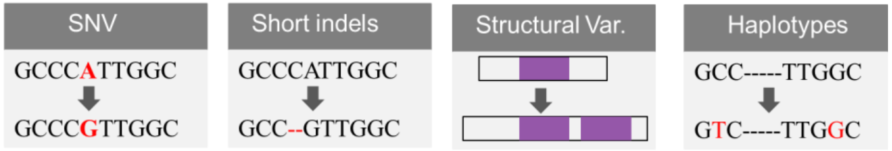


Raw sequence data from disease study

Variant calls (VCF)

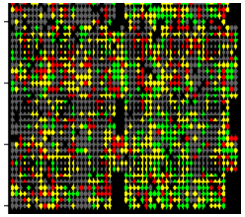
Disease causing mutations & genes





Variant annotation

Model organism gene knockout



Population	Allele Frequency
Other	0.001104
European (Non-Finnish)	0.0008406
African	9.647e-05
South Asian	6.056e-05

Figure from https://gemini.readthedocs.io/

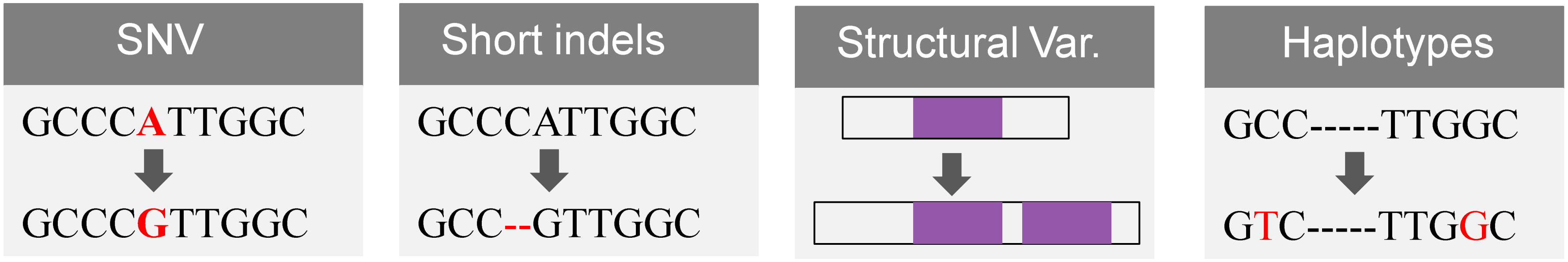
1. Discovery of sequence variants

Raw sequence data from WGS or WES

↓

Aligned sequence reads

↓



SNV

GCCCATTGGC

GCCCGTTGGC

Short indels

GCCCATTGGC

GCC--GTTGGC

Structural Var.

Haplotypes

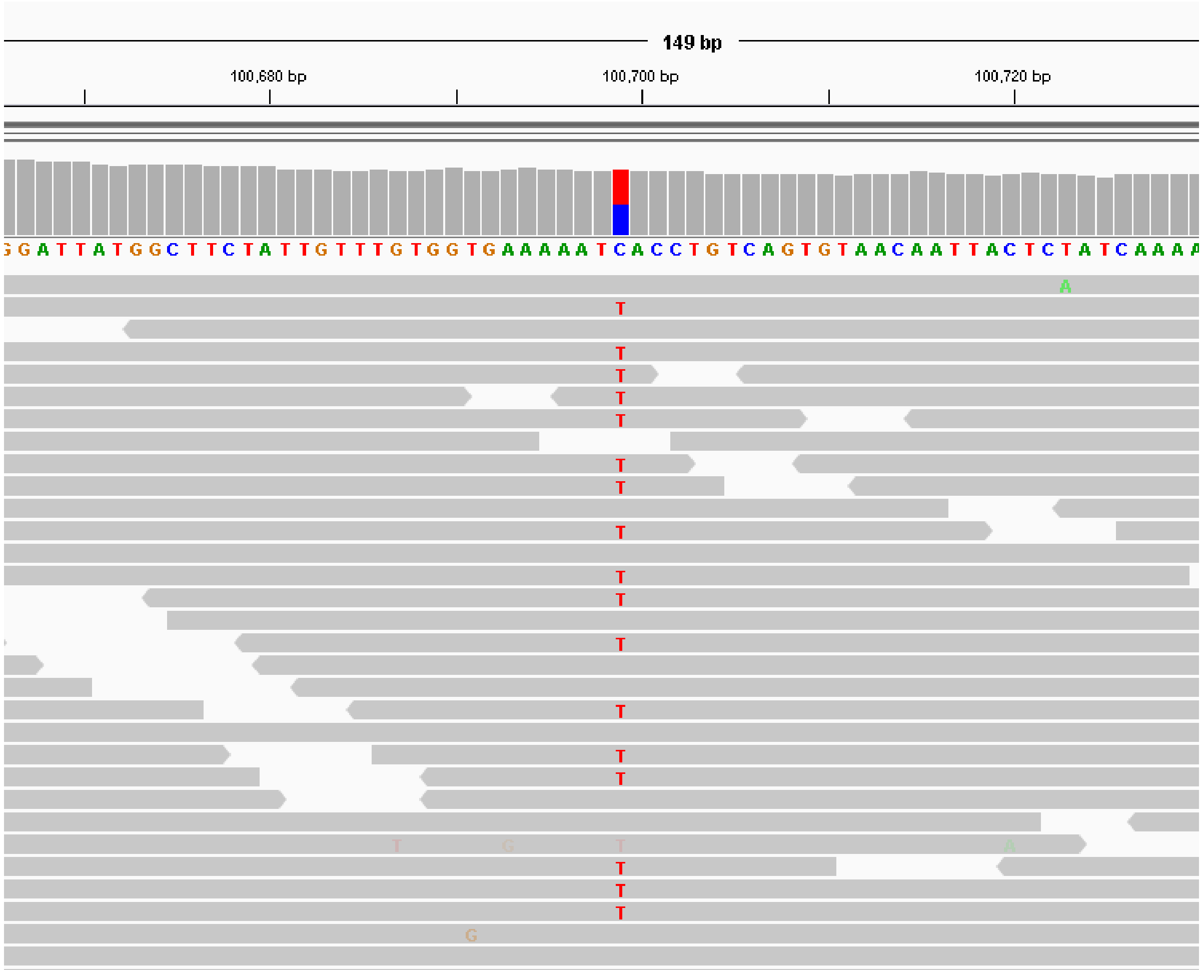
GCC-----TTGGC

GTC-----TTGGC

Goal: Identify differences between the sequenced genome (represented in the form of short reads) and a ‘reference’ genome

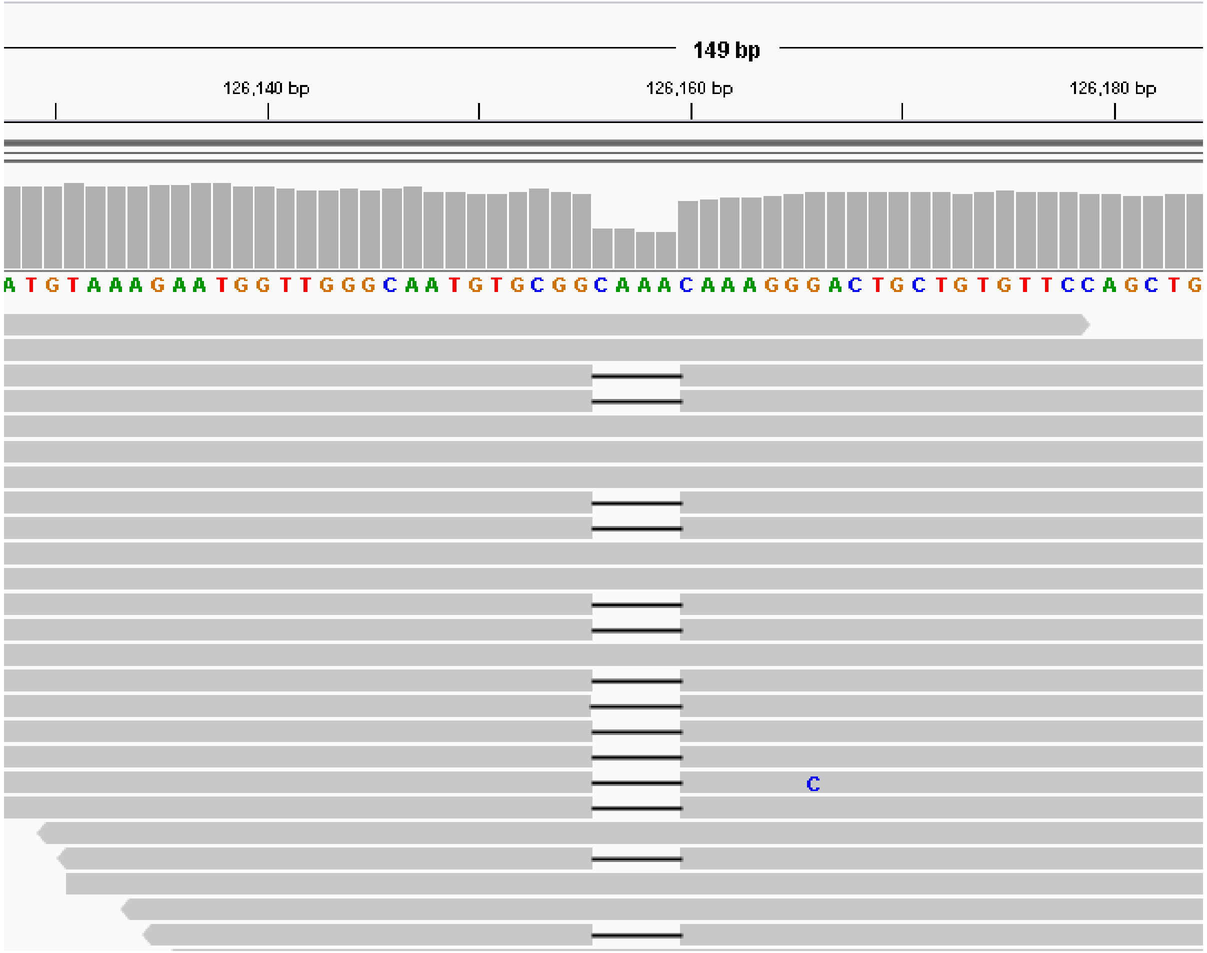
SNVs from aligned reads

reference genome

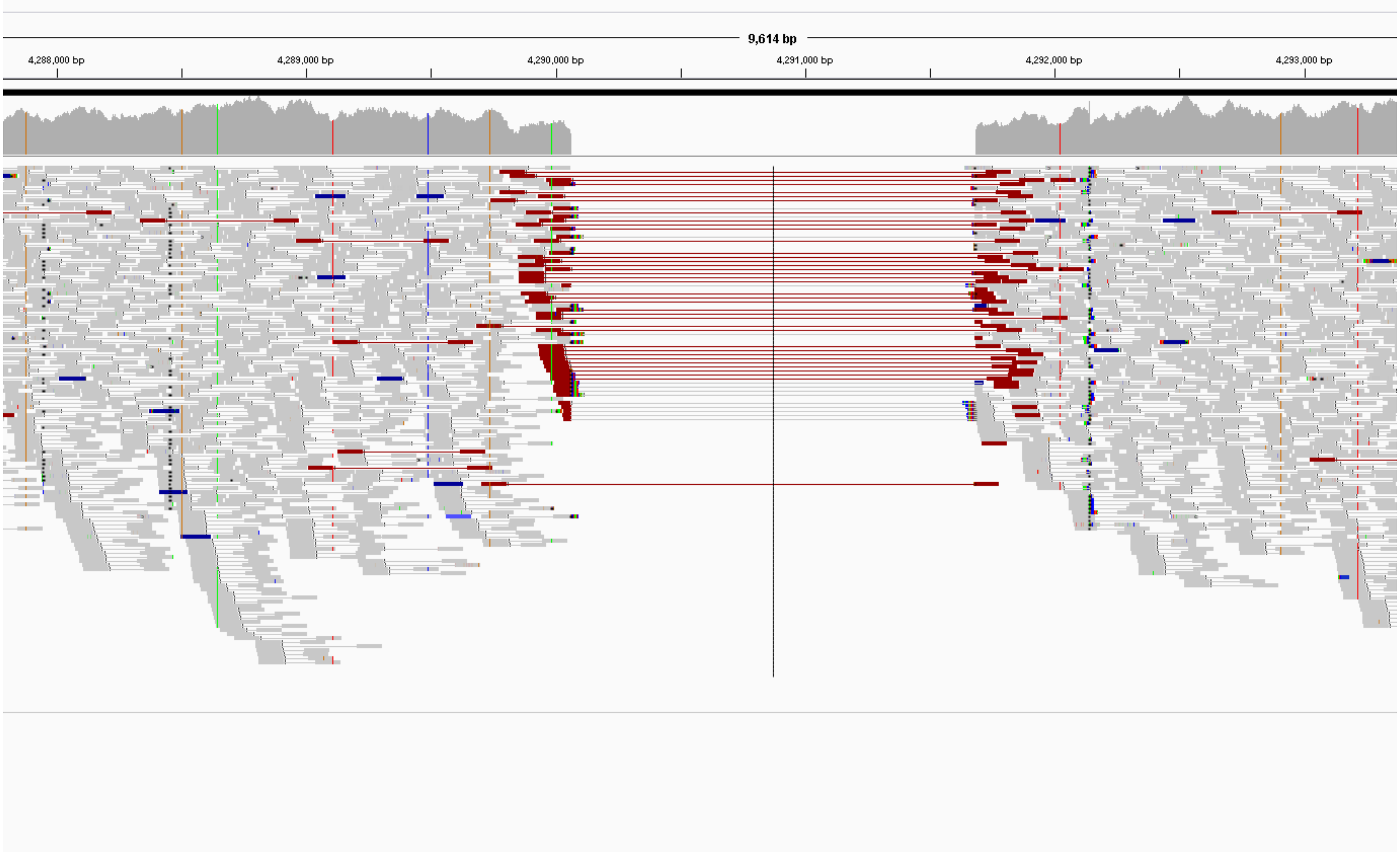

149 bp
100,680 bp
100,700 bp
100,720 bp
GGATTATGGCTTCTATTGTTTGTGGTGAAAAATCACCTGTCAGTGTAACAATTACTCTATCAAAA
A
T
G

aligned reads

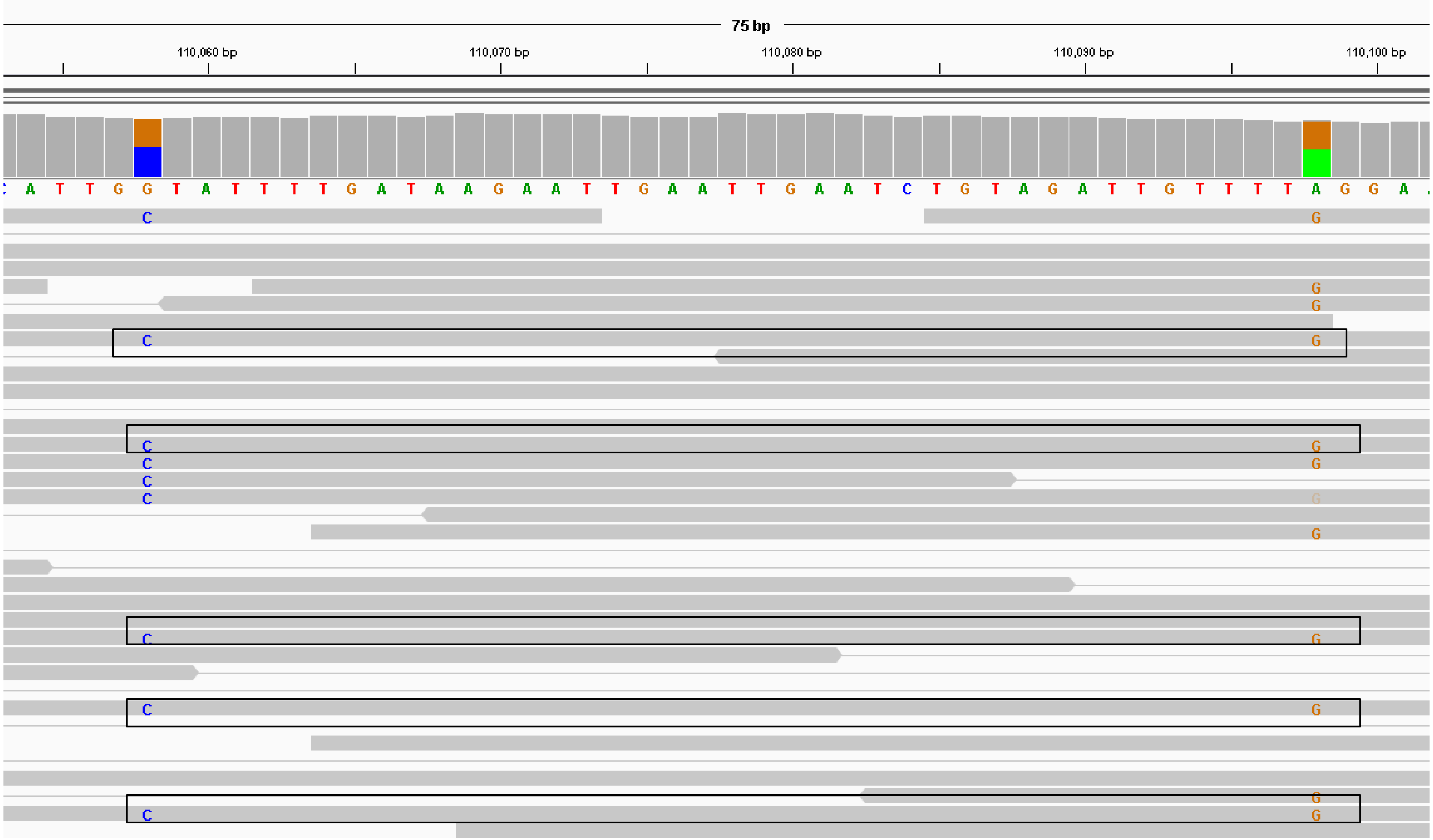
Small indels from aligned reads



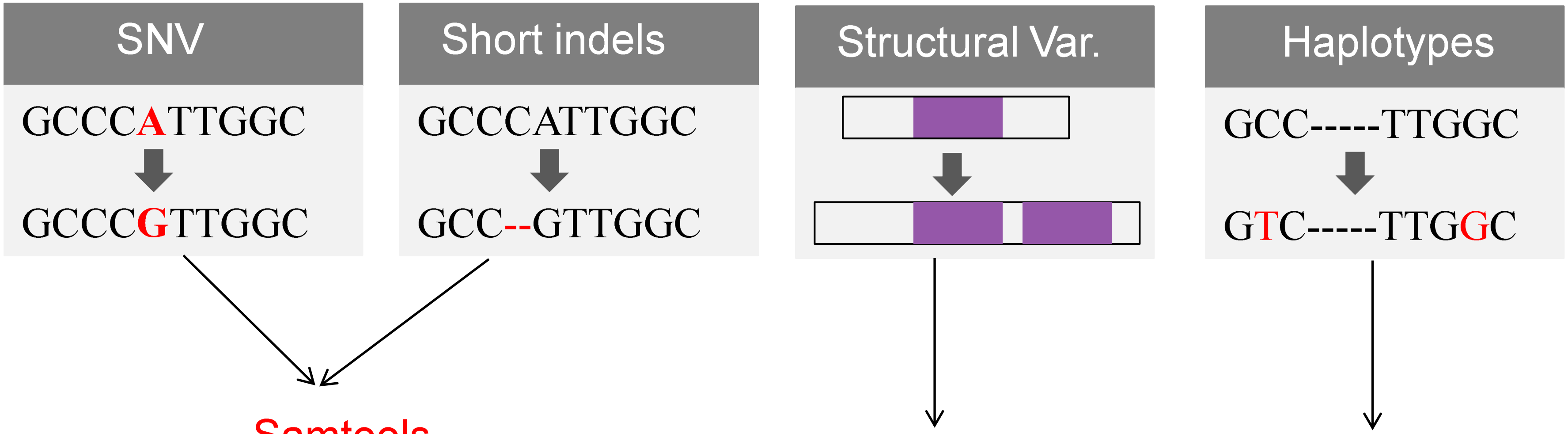
Long deletions from aligned reads



Phasing of heterozygous variants


75 bp
110,060 bp
110,070 bp
110,080 bp
110,090 bp
110,100 bp
A T T G G T A T T T T G A T A A G A A T T G A A T T G A A T C T G T A G A T T G T T T T A G G A
C
G

1. Discovery of genetic variants



SNV

GCCCATTGGC

GCCCGTTGGC

Short indels

GCCCATTGGC

GCC--GTTGGC

Structural Var.

Haplotypes

GCC-----TTGGC

GTC-----TTGGC

Samtools
GATK
Platypus
FreeBayes
Strelka (Tumor)
Mutect (Tumor)

GenomeStrip
LUMPY
Pindel
BreakDancer
Delly

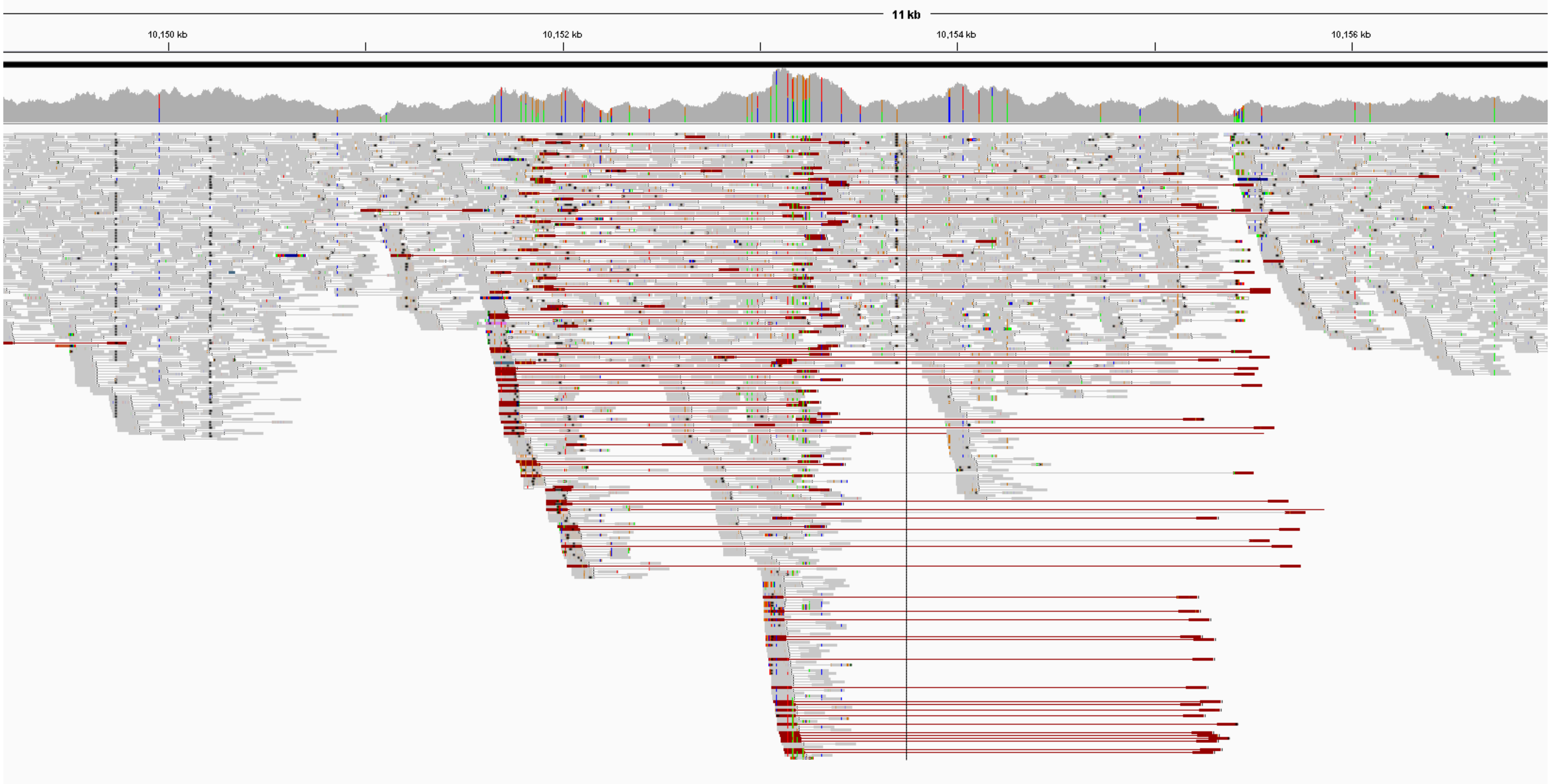
HapCUT
RefHap
phASER

Challenges in variant detection

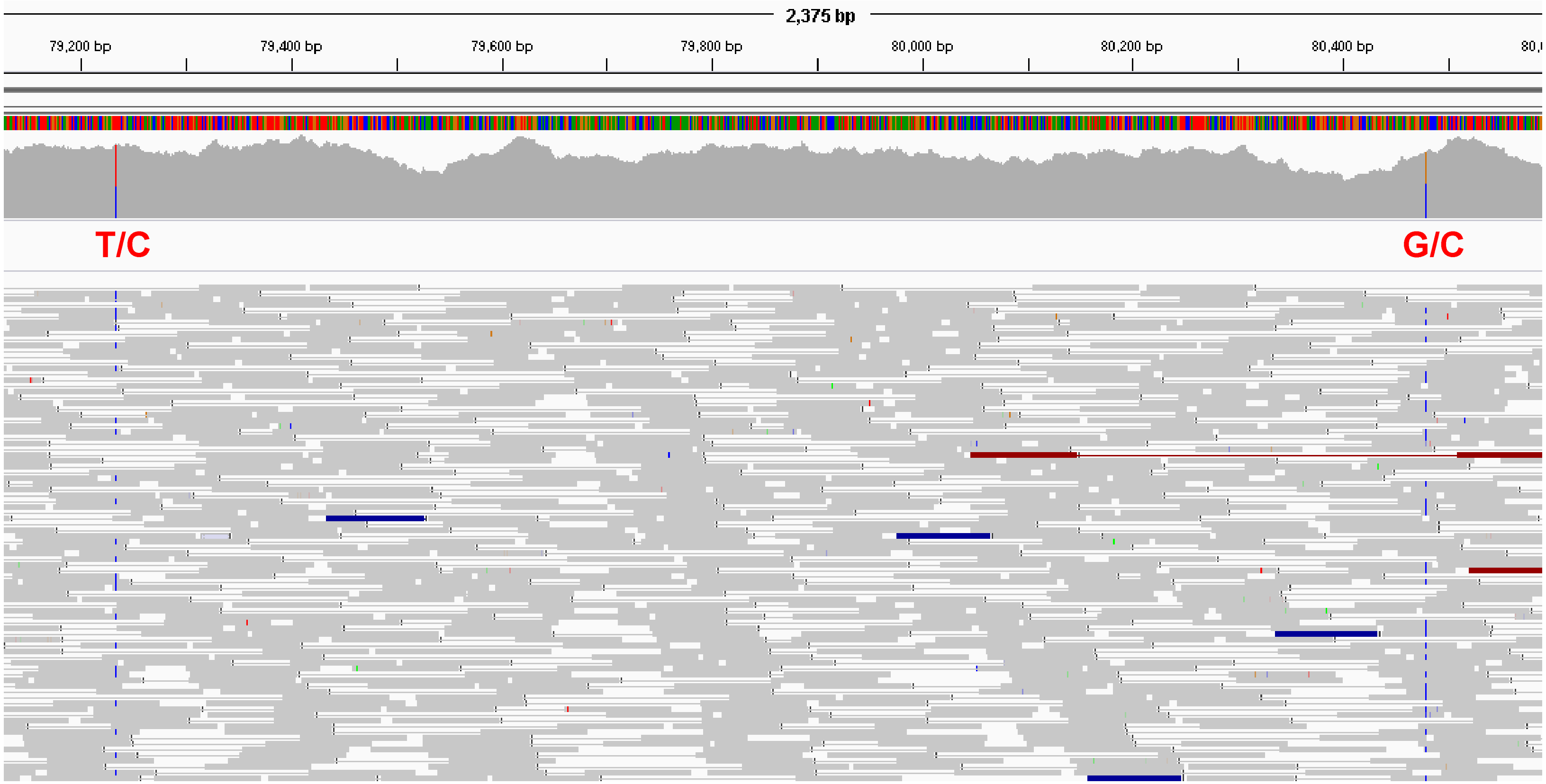
- Systematic sequencing and alignment errors
- Indels and structural variants over-represented in repeats and low-complexity sequence

- Limitations of short read lengths
  - Structural variants
  - Haplotypes
  - Variants in duplicated genes

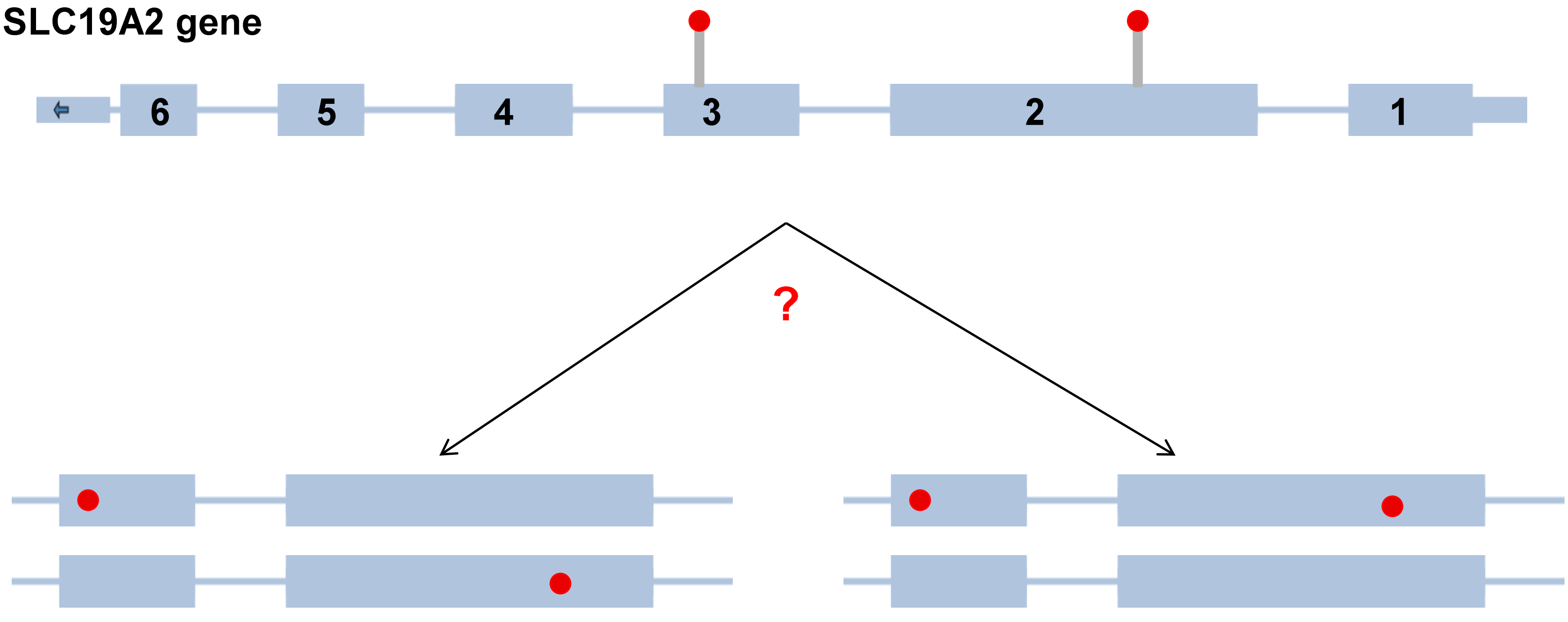
Complex signal at loci with structural variation



Distant heterozygous variants cannot be phased



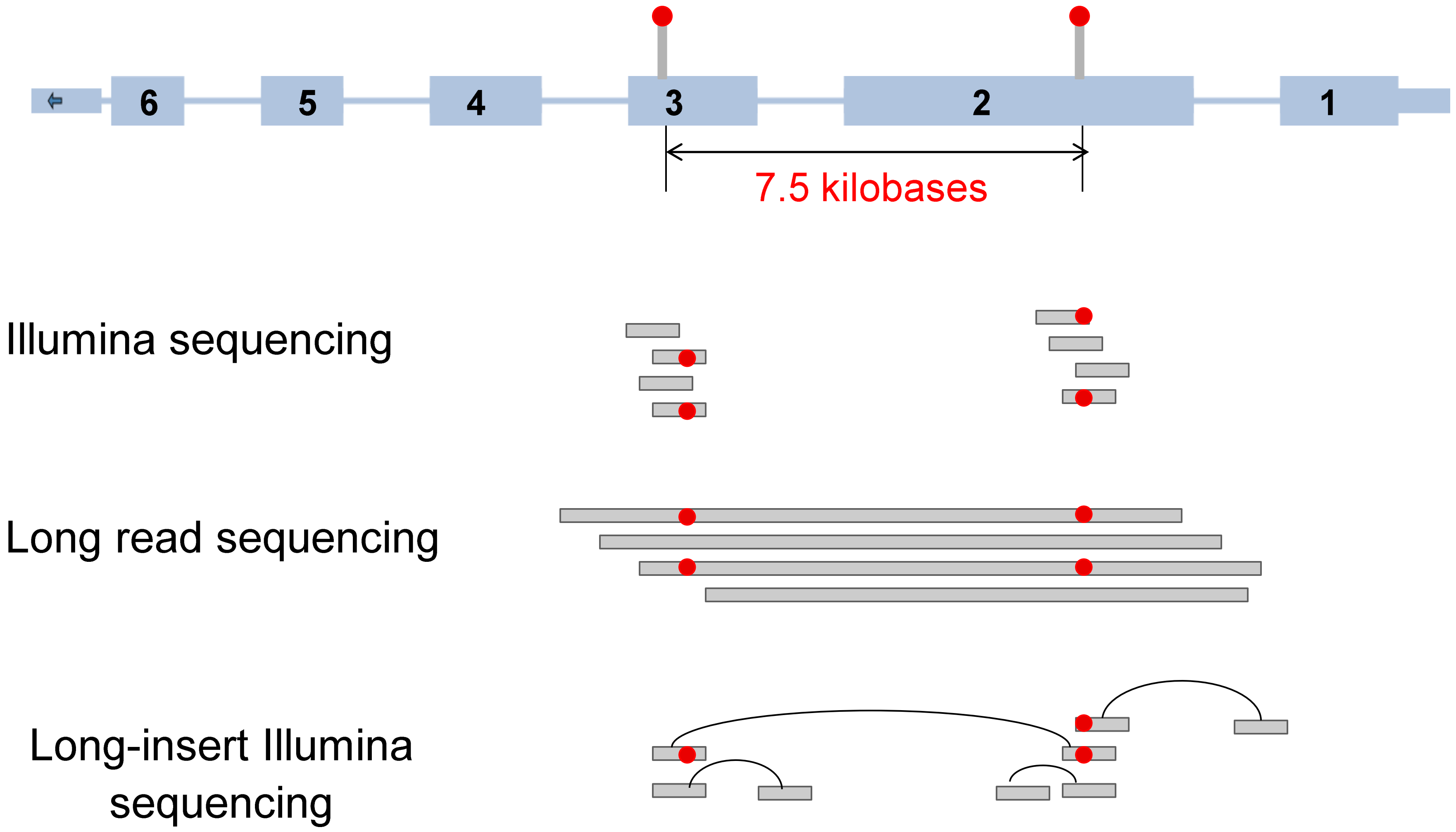
Why do we need phasing ?



Thiamine-responsive Megalobastic anemia

No disease

Haplotyping requires long range information

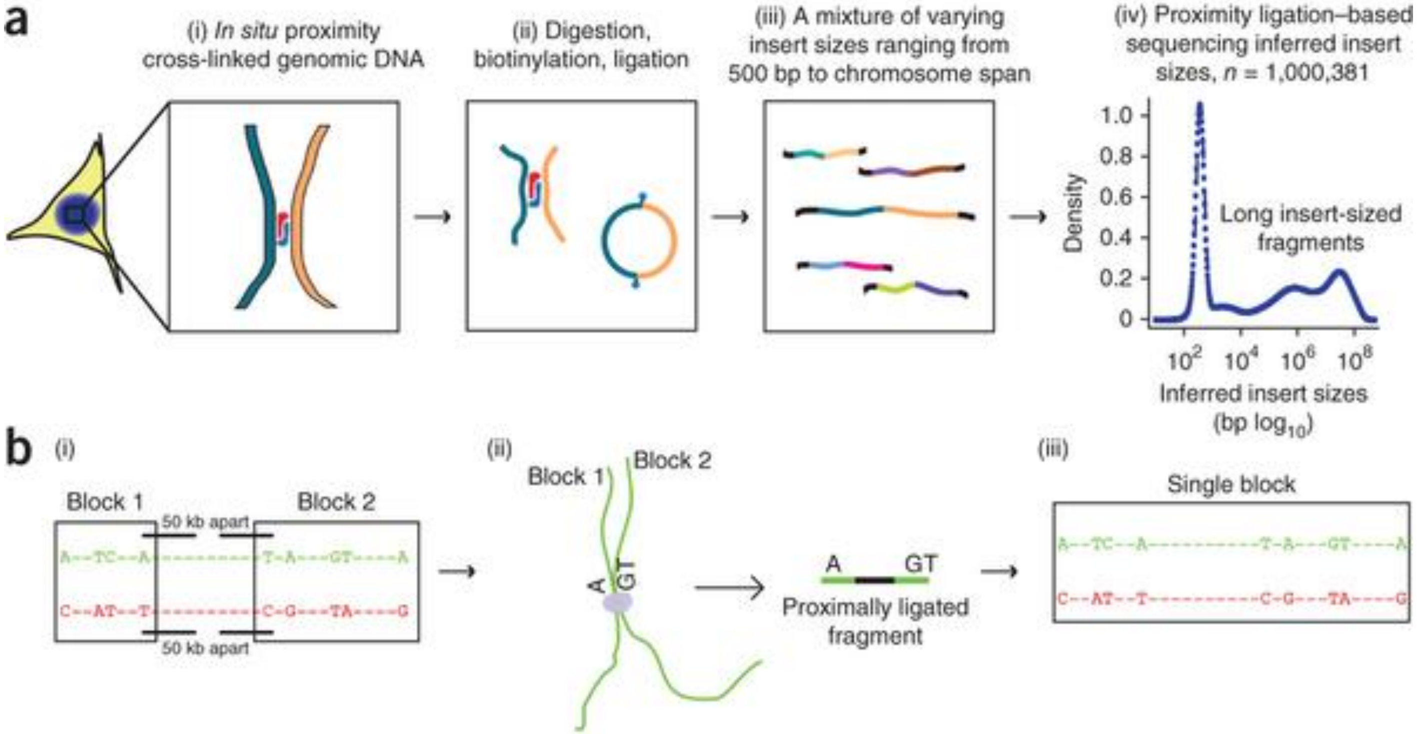


Illumina sequencing

Long read sequencing

Long-insert Illumina sequencing

Chromosomal-span haplotypes from proximity-ligation sequencing


a
(i) *In situ* proximity cross-linked genomic DNA
(ii) Digestion, biotinylation, ligation
(iii) A mixture of varying insert sizes ranging from 500 bp to chromosome span
(iv) Proximity ligation–based sequencing inferred insert sizes, n = 1,000,381
Density
1.0
0.8
0.6
0.4
0.2
0
Long insert-sized fragments
10^2
10^4
10^6
10^8
Inferred insert sizes (bp $\log_{10}$)
b
(i)
Block 1
Block 2
50 kb apart
A--TC--A---------T-A---GT----A
C--AT--T---------C-G---TA----G
50 kb apart
(ii)
Block 1
Block 2
A
GT
A
GT
Proximally ligated fragment
(iii)
Single block
A--TC--A----------T-A---GT----A
C--AT--T----------C-G---TA----G

- Feasible to assemble accurate, chromosomal-span haplotypes from Illumina short reads using HapCUT

S. Selvaraj, J. Dixon, V. Bansal, B. Ren, Nature Biotech. 2013

Variants in duplicated genes cannot be detected using short reads

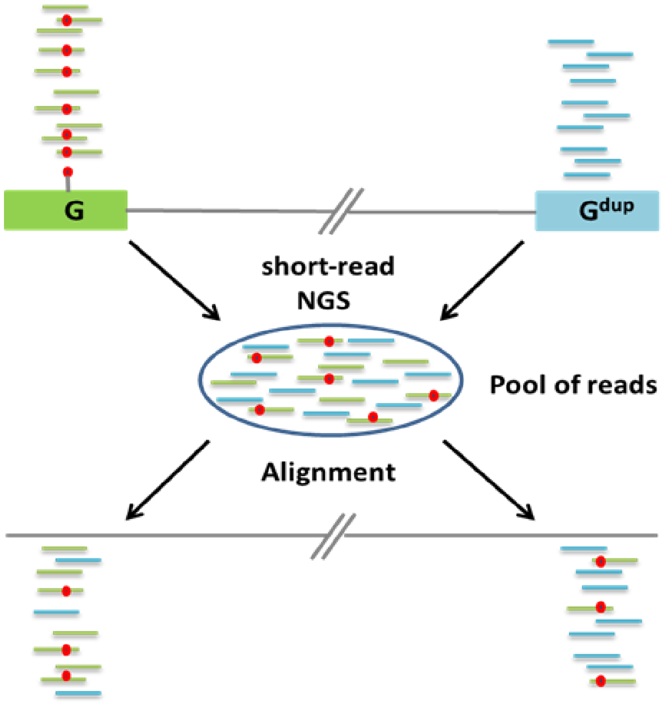


Fig. 1: NGS with short reads cannot identify variants in duplicated genes with high sequence homology. A variant is located in a gene 'G' (that has a duplicated copy G^{dup}). Reads with the variant allele align equally well to both G and G^{dup} masking the correct location of the variant.

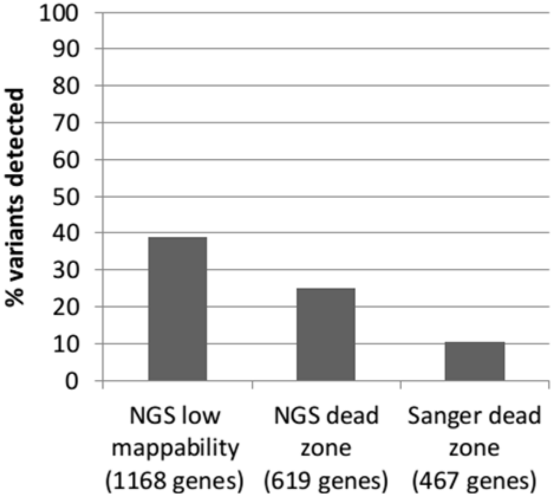
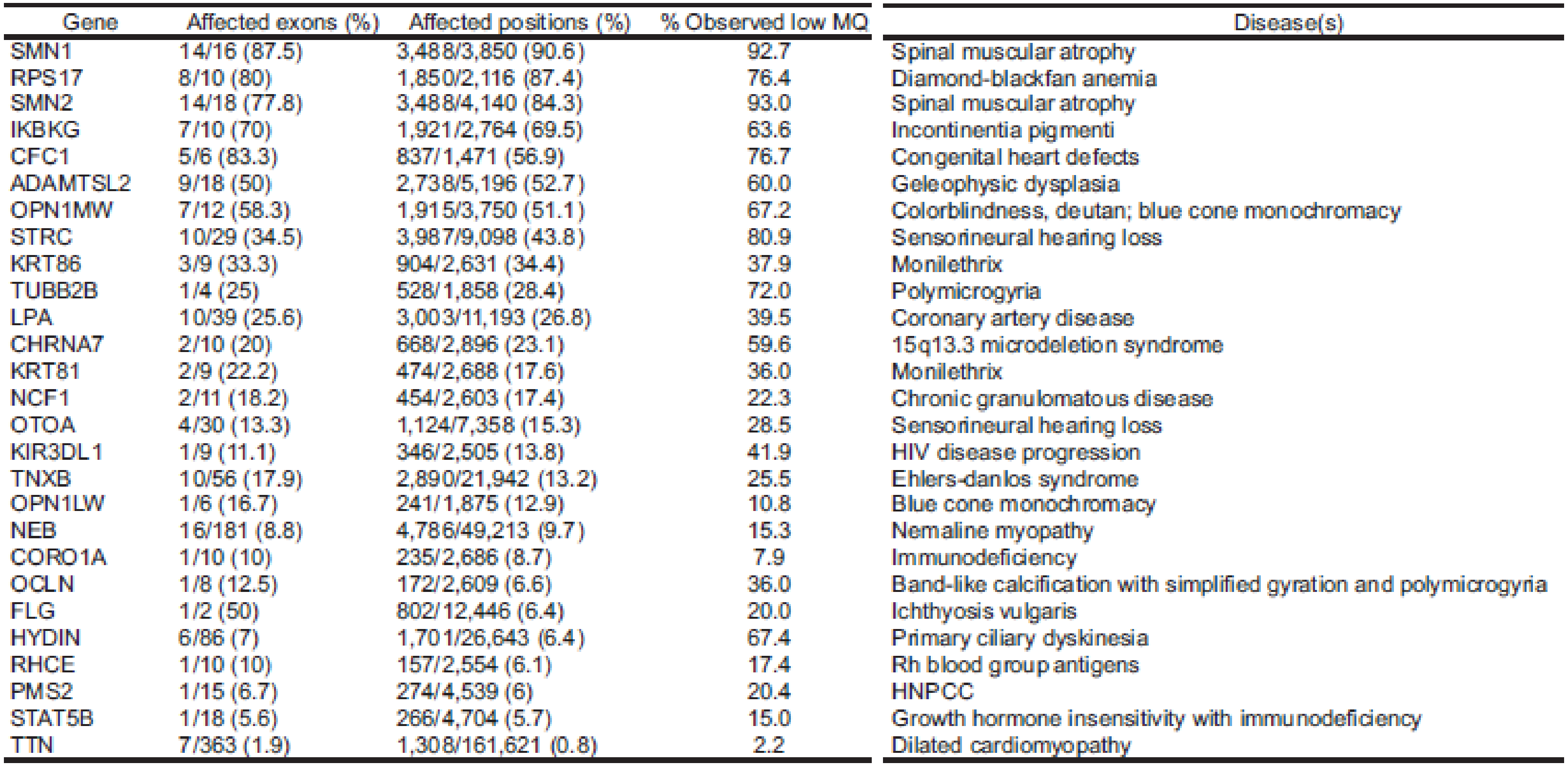


Fig 2: Low power to detect variants in genes with high sequence homology in the human genome using short NGS reads. Variants and 100 bp paired-end reads (60x coverage) were simulated in exons with HSH, reads were aligned to the reference human genome and variants identified using the GATK toolkit. The fraction of variants that could be detected in exons with low mappability using 250 bp reads was less than 40% and reduced to 10% in exons which overlap a 1000 bp region with perfect homology to other loci (Sanger dead zone).

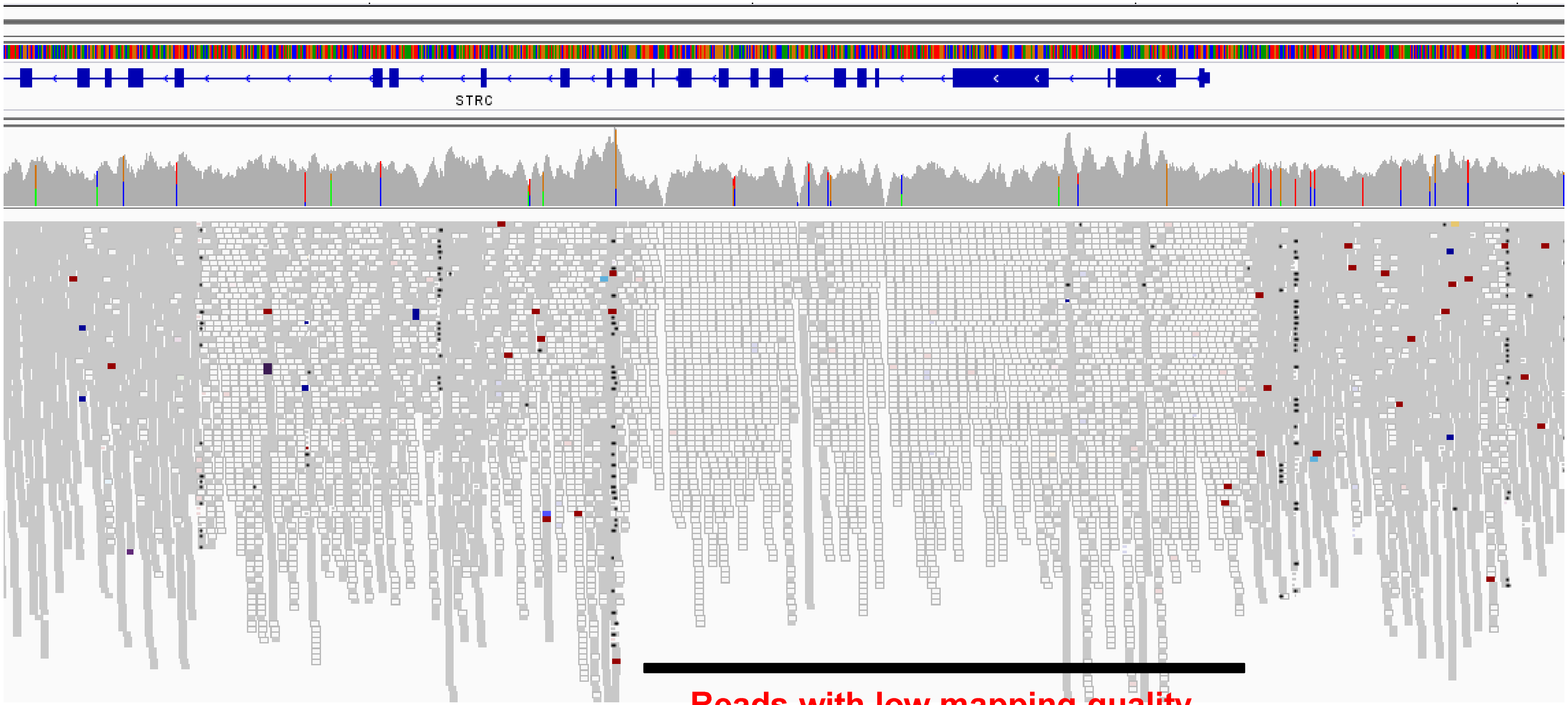
> 100 genes with high medical relevance are problematic for short reads



Gene	Affected exons (%)	Affected positions (%)	% Observed low MQ	Disease(s)
SMN1	14/16 (87.5)	3,488/3,850 (90.6)	92.7	Spinal muscular atrophy
RPS17	8/10 (80)	1,850/2,116 (87.4)	76.4	Diamond-blackfan anemia
SMN2	14/18 (77.8)	3,488/4,140 (84.3)	93.0	Spinal muscular atrophy
IKBKG	7/10 (70)	1,921/2,764 (69.5)	63.6	Incontinentia pigmenti
CFC1	5/6 (83.3)	837/1,471 (56.9)	76.7	Congenital heart defects
ADAMTSL2	9/18 (50)	2,738/5,196 (52.7)	60.0	Geleophysic dysplasia
OPN1MW	7/12 (58.3)	1,915/3,750 (51.1)	67.2	Colorblindness, deutan; blue cone monochromacy
STRC	10/29 (34.5)	3,987/9,098 (43.8)	80.9	Sensorineural hearing loss
KRT86	3/9 (33.3)	904/2,631 (34.4)	37.9	Monilethrix
TUBB2B	1/4 (25)	528/1,858 (28.4)	72.0	Polymicrogyria
LPA	10/39 (25.6)	3,003/11,193 (26.8)	39.5	Coronary artery disease
CHRNA7	2/10 (20)	668/2,896 (23.1)	59.6	15q13.3 microdeletion syndrome
KRT81	2/9 (22.2)	474/2,688 (17.6)	36.0	Monilethrix
NCF1	2/11 (18.2)	454/2,603 (17.4)	22.3	Chronic granulomatous disease
OTOA	4/30 (13.3)	1,124/7,358 (15.3)	28.5	Sensorineural hearing loss
KIR3DL1	1/9 (11.1)	346/2,505 (13.8)	41.9	HIV disease progression
TNXB	10/56 (17.9)	2,890/21,942 (13.2)	25.5	Ehlers-danlos syndrome
OPN1LW	1/6 (16.7)	241/1,875 (12.9)	10.8	Blue cone monochromacy
NEB	16/181 (8.8)	4,786/49,213 (9.7)	15.3	Nemaline myopathy
CORO1A	1/10 (10)	235/2,686 (8.7)	7.9	Immunodeficiency
OCLN	1/8 (12.5)	172/2,609 (6.6)	36.0	Band-like calcification with simplified gyration and polymicrogyria
FLG	1/2 (50)	802/12,446 (6.4)	20.0	Ichthyosis vulgaris
HYDIN	6/86 (7)	1,701/26,643 (6.4)	67.4	Primary ciliary dyskinesia
RHCE	1/10 (10)	157/2,554 (6.1)	17.4	Rh blood group antigens
PMS2	1/15 (6.7)	274/4,539 (6)	20.4	HNPCC
STAT5B	1/18 (5.6)	266/4,704 (5.7)	15.0	Growth hormone insensitivity with immunodeficiency
TTN	7/363 (1.9)	1,308/161,621 (0.8)	2.2	Dilated cardiomyopathy

Mandelkar et al. Genetics in medicine 2016

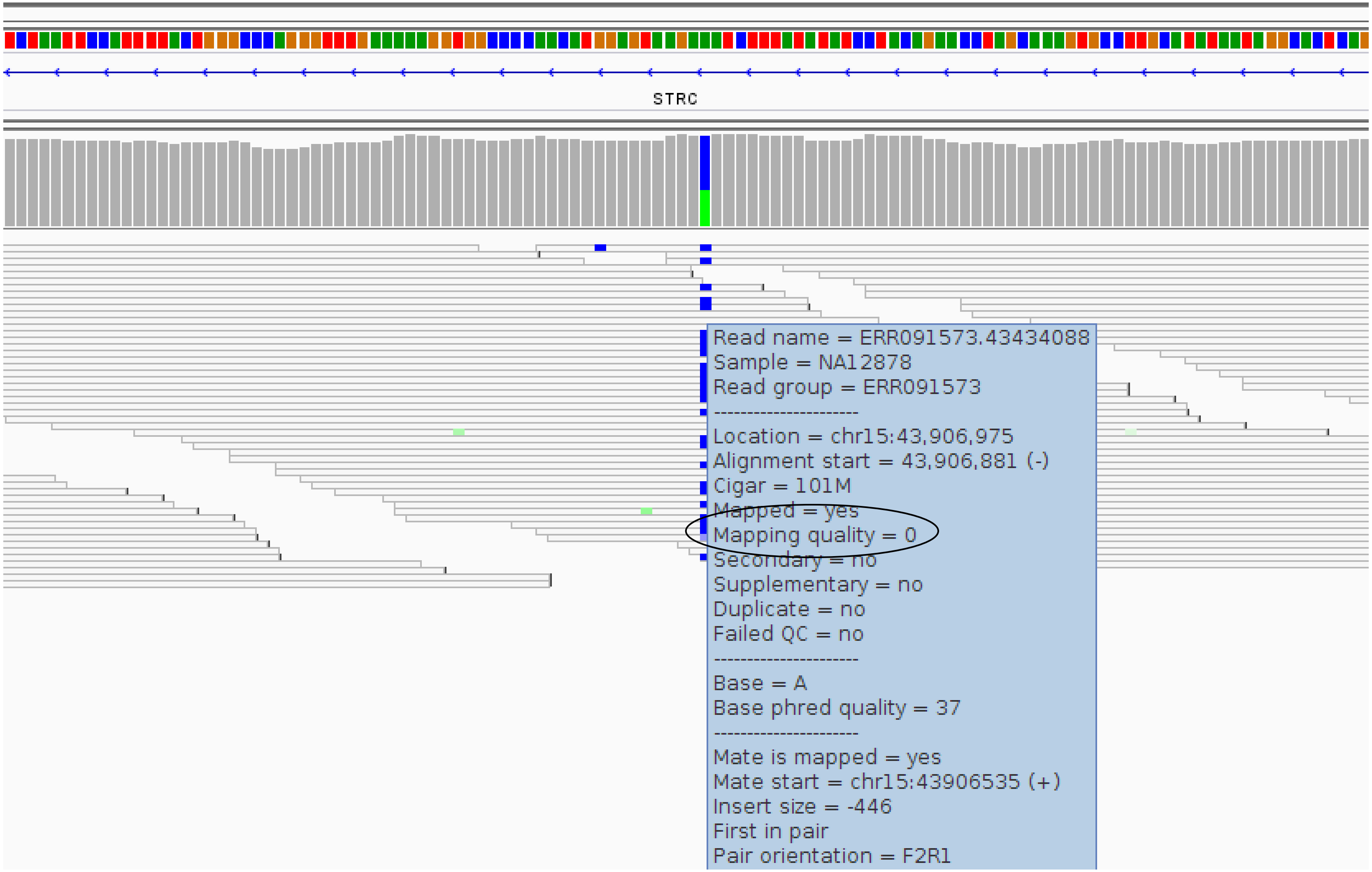
STRC gene: exons 1-15 are duplicated with 99% homology



Reads with low mapping quality

- Bi-allelic mutations in STRC cause sensorineural hearing loss

Good coverage but reads have low mapping quality



Read name = ERR091573.43434088
Sample = NA12878
Read group = ERR091573

Location = chr15:43,906,975
Alignment start = 43,906,881 (-)
Cigar = 101M
Mapped = yes
Mapping quality = 0
Secondary = no
Supplementary = no
Duplicate = no
Failed QC = no

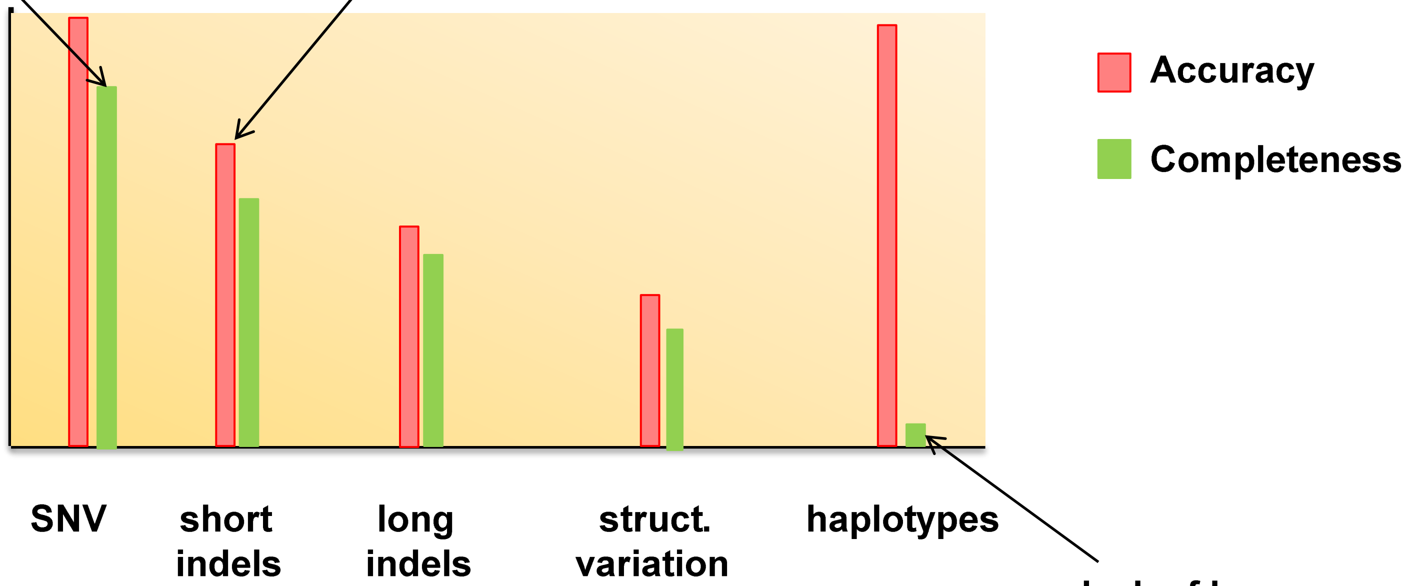
Base = A
Base phred quality = 37

Mate is mapped = yes
Mate start = chr15:43906535 (+)
Insert size = -446
First in pair
Pair orientation = F2R1

Whole-genome Illumina sequencing is incomplete

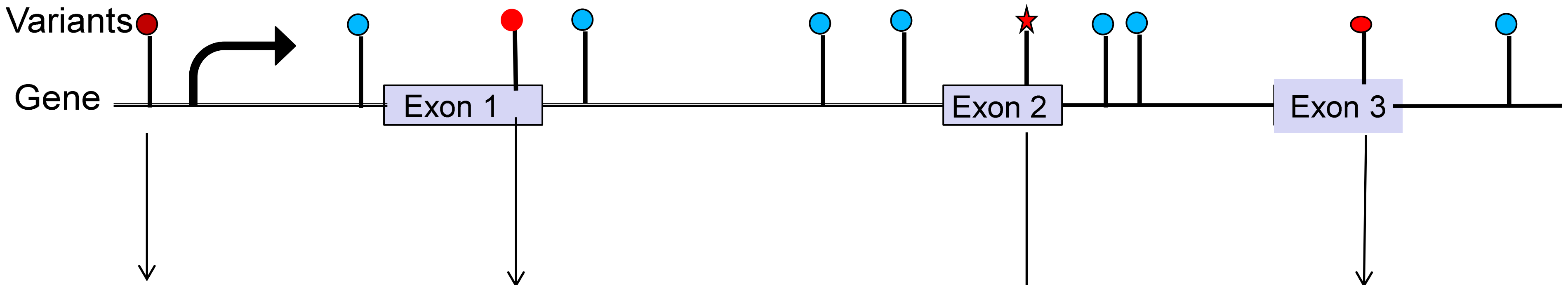
~10% variants missed in duplicated regions of genome

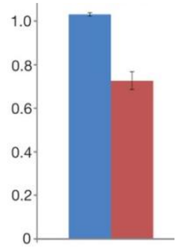
low accuracy (60-80%) for short indels



lack of long-range haplotype information

2. Interpretation of variants

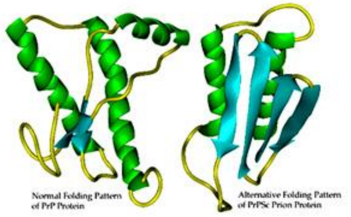




Promoter mutation reduces mRNA expression

Missense mutation with no impact

Stop gain mutation leads to truncated protein sequence

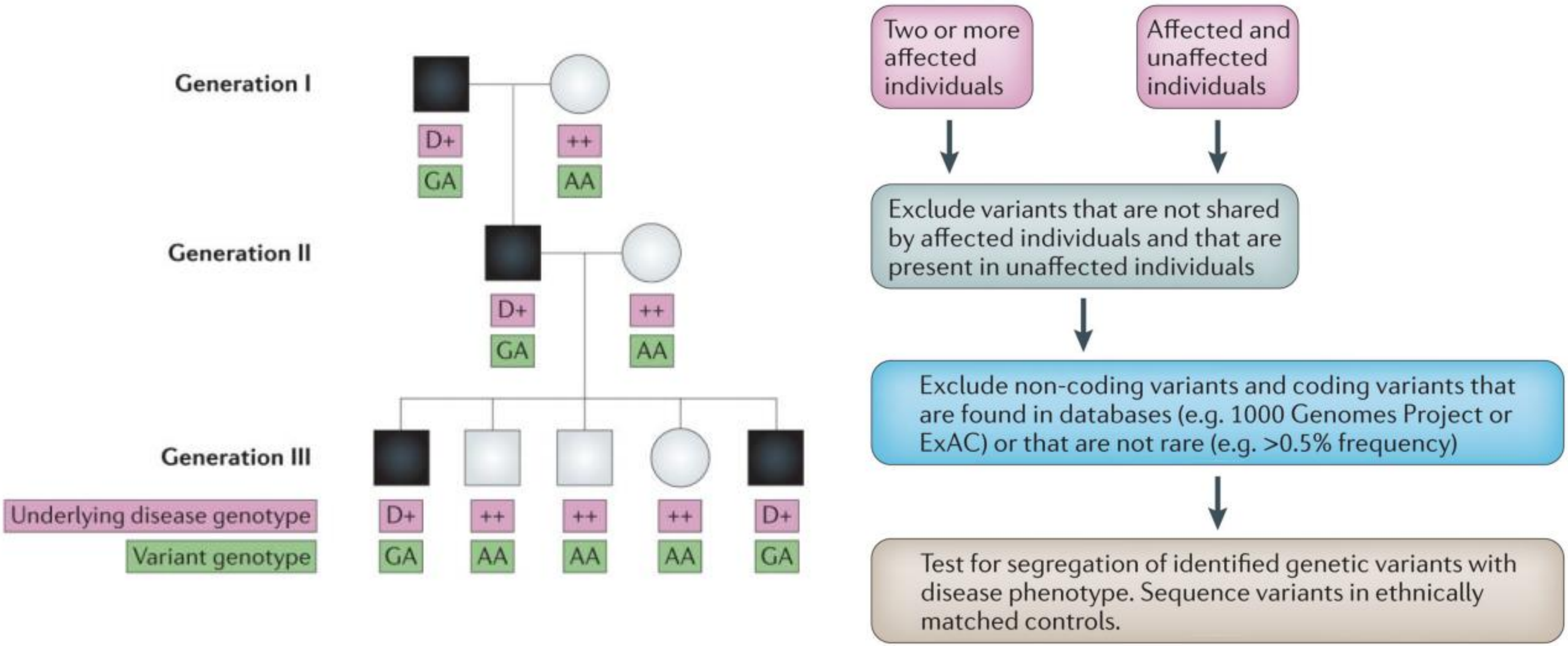

Normal Folding Pattern of PrP Protein
Alternative Folding Pattern of PrPSc Prion Protein

misfolded protein due to missense mutation

Different approaches for finding disease causing mutations

1. Family data
2. Multiple unrelated individuals
3. N=1: prioritization using population data
4. Integrating genetic, gene-expression and model organism data
5. Combining DNA-seq and RNA-seq from a single individual

1. Family data



Two or more affected individuals

Affected and unaffected individuals

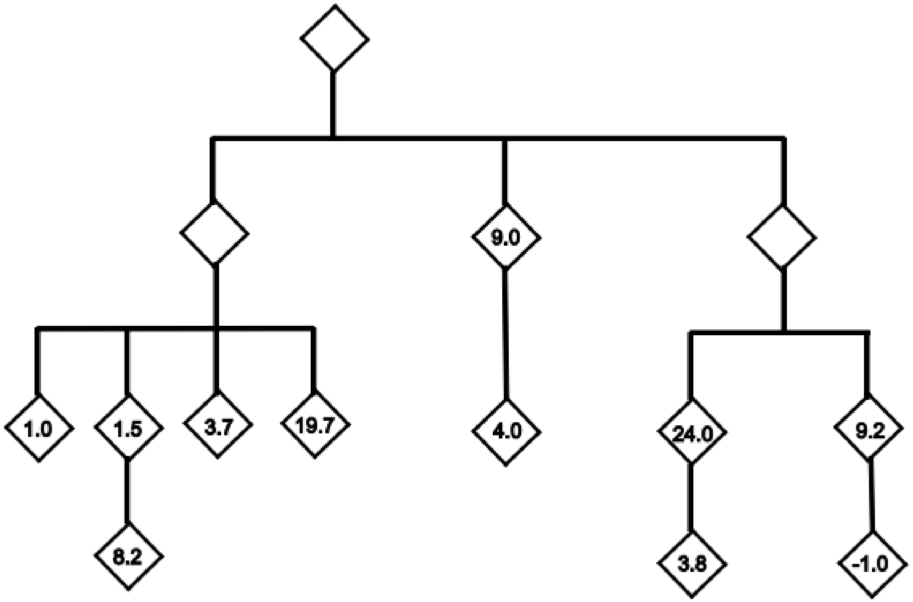
Exclude variants that are not shared by affected individuals and that are present in unaffected individuals

Exclude non-coding variants and coding variants that are found in databases (e.g. 1000 Genomes Project or ExAC) or that are not rare (e.g. >0.5% frequency)

Test for segregation of identified genetic variants with disease phenotype. Sequence variants in ethnically matched controls.

Ott et al. Nat Rev. Genetics 2015

1. Family data: Hypertriglyceridemia



- 5-generation family with 121 individuals
- Linkage mapping using genotyping arrays
- Exome sequencing of 16 individuals
- Two linkage peaks: chromosome 7 & 17

Rosenthal et al. AJHG 2013

Variants under linkage peaks

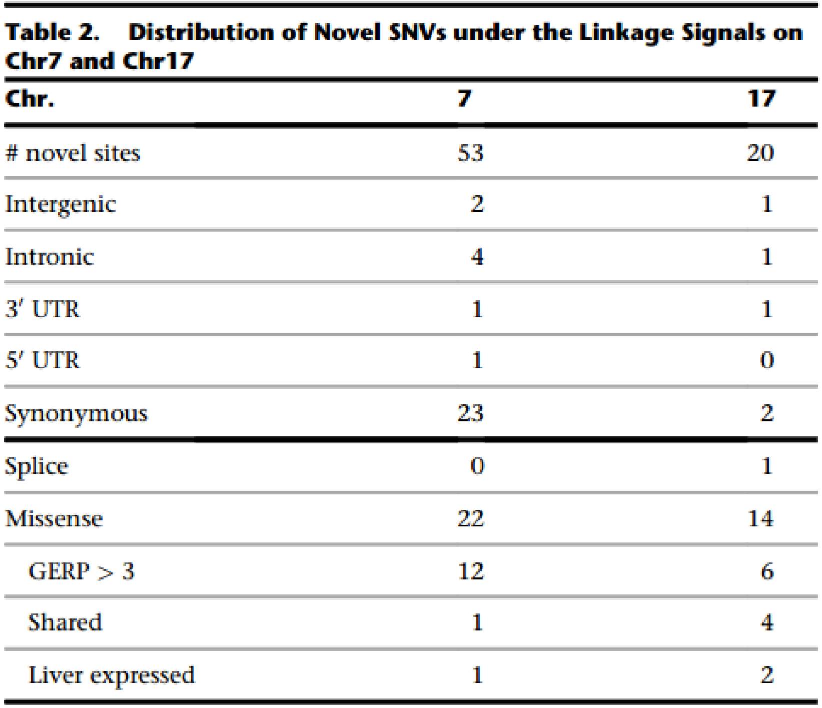


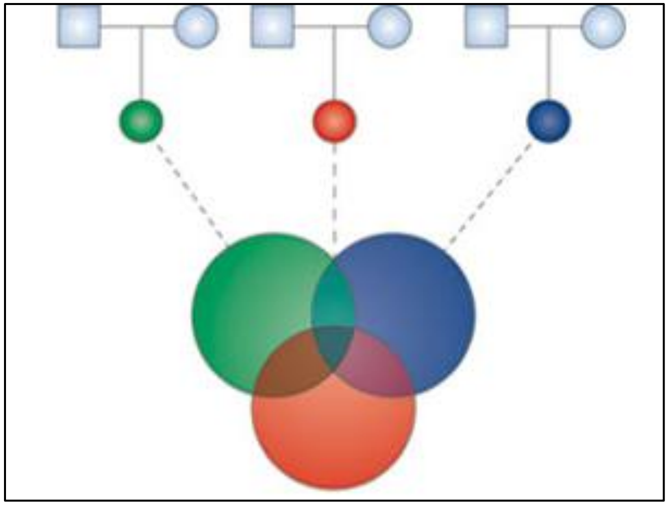
Table 2. Distribution of Novel SNVs under the Linkage Signals on Chr7 and Chr17

Chr.	7	17
# novel sites	53	20
Intergenic	2	1
Intronic	4	1
3′ UTR	1	1
5′ UTR	1	0
Synonymous	23	2
Splice	0	1
Missense	22	14
GERP > 3	12	6
Shared	1	4
Liver expressed	1	2

- Tyr125Cys mutation in SLC25A40 explains chr7 peak
- Pop. Freq = 0.00006
- Additional evidence that SLC25A40 mutations affect cholesterol levels
- Variant in PLD2 cosegregates with high TG but unlikely to be causal

Rosenthal et al. AJHG 2013

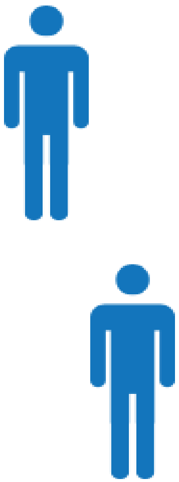
2. Multiple unrelated individuals



- Different mutations present in the same gene in multiple unrelated individuals
- mutations have very low population allele frequency and are deleterious
- Power depends on number of individuals with disease and genetic heterogeneity

Figure from Bamshad et al. Nat. Rev Gen. 2011

2. Multiple unrelated individuals



unrelated individuals
with same phenotype

sequence DNA

CCAGTCGATCAGGCA TGATCGAAC
GTGTAGCACCCAAAGTA
TGCGATCGATA CGTGTAGTAGTT
GTGTAGCACGACTAGCA
GTGCACACCATTTTCAG
GTGACGACACTCGATCGA

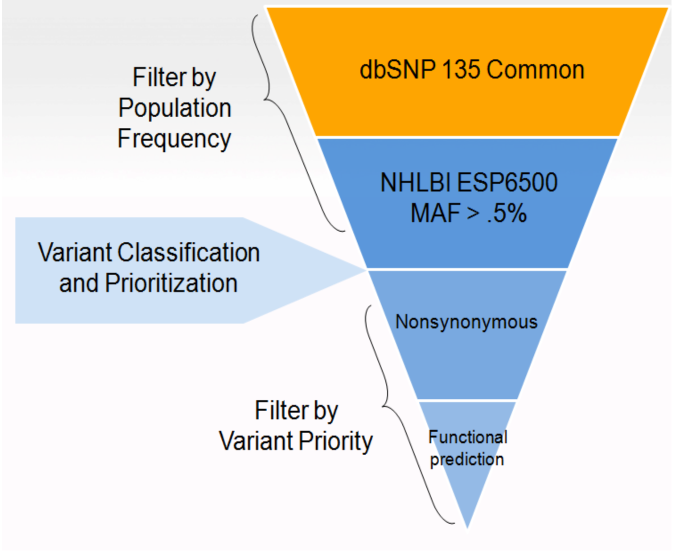
millions of short reads

align to reference

human genome

identify variants

45,000 coding variants

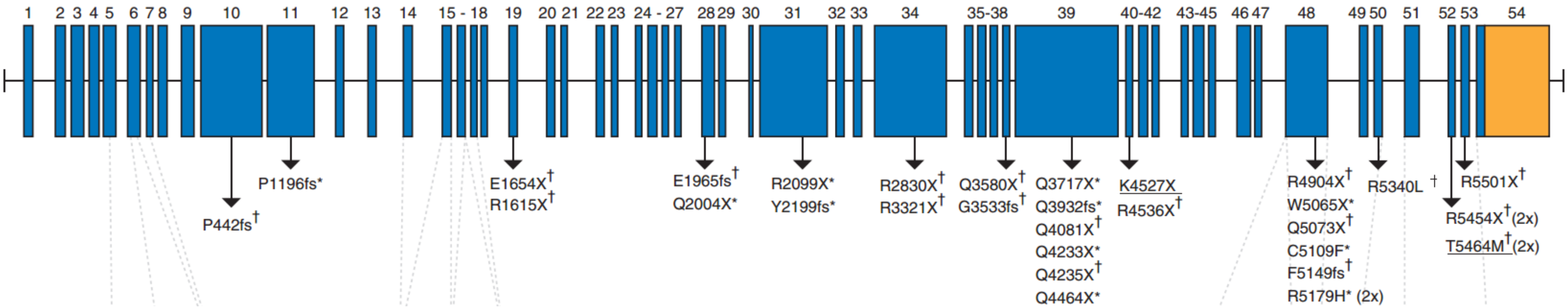


~100 candidate genes per individual

1-2 shared genes

2. Multiple unrelated individuals: Kabuki syndrome

- Multiple malformation syndrome first described in 1981
- 7/10 patients with loss-of function mutations in **MLL2**
- 54 exon gene that regulates DNA methylation



- Mutations detected in 26/43 additional patients

Ng. et al. Nat. Genetics 2010

3. N=1: prioritization using population data

- 100-200 candidate variants or genes per individual
- How to prioritize further ?

- Use gene-level constraint in population data
  - If mutations in a gene cause severe disease, such mutations likely to be depleted in healthy individuals

Disease mutations and fitness

Mutations causing rare disease have negative fitness effects (less likely to reproduce)

Mutation less likely to be transmitted to next generation compared to a neutral mutation

Negative selection -> such mutations less likely to be observed in the normal population

Model

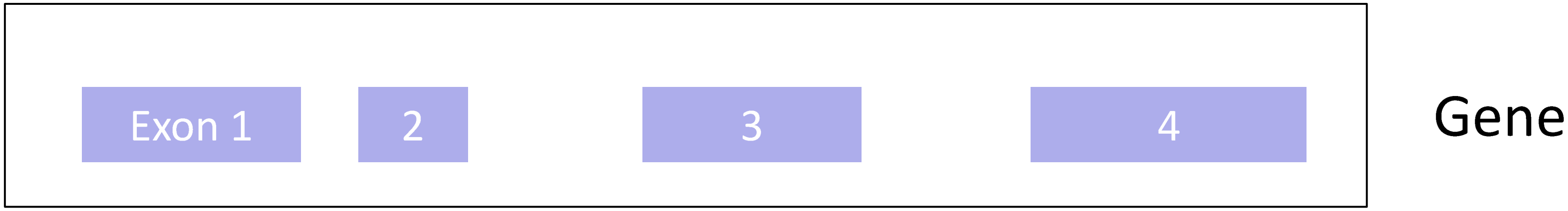
For any gene G

- Let p(LoF) denote the probability of observing loss-of-function mutations in the gene
- Let p(S) be the probability of observing silent mutations

If loss-of-function mutations cause a disease that reduces fitness:

- Obs(LoF) << Exp(LoF) and Obs(S) ~ Exp(S)
- Silent mutations are mostly neutral

Mutation probabilities per gene

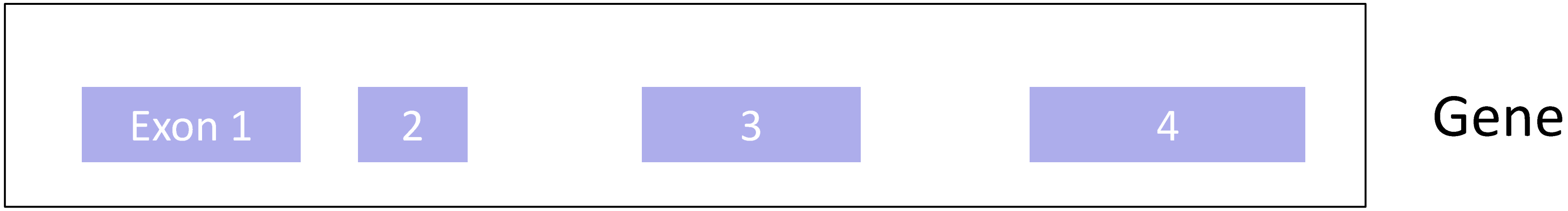


Tri-nucleotide mutation rates

ACG -> ATG	5.6×10^{-9}
CGA -> CTA	8.7×10^{-10}
.	
.	
.	
.	
GCT -> GTT	2.4×10^{-8}

silent	867	2.1×10^{-7}
missense	1784	5.3×10^{-7}
stop-gain	123	4.2×10^{-8}

Expected vs observed mutation counts



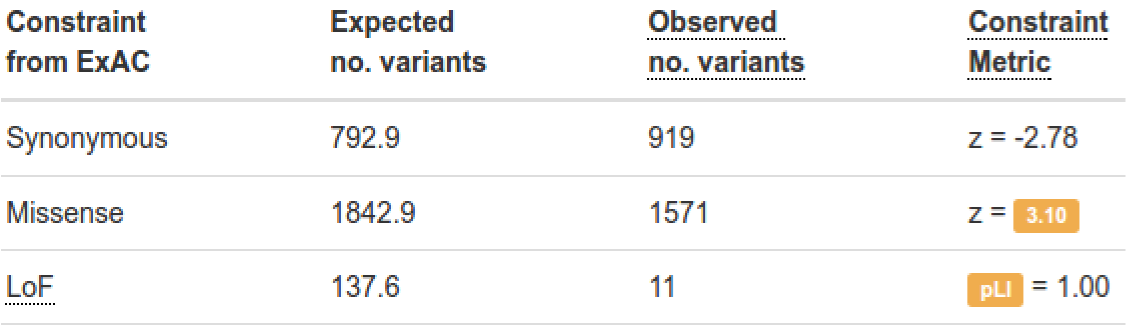
Tri-nucleotide mutation rates

ACG -> ATG	5.6×10^{-9}
CGA -> CTA	8.7×10^{-10}
.	
.	
.	
.	
GCT -> GTT	2.4×10^{-8}

silent	867	2.1×10^{-7}	64
missense	1784	4.6×10^{-7}	131
stop-gain	**123**	$\mathbf{4.2 \times 10^{-8}}$	**2**

Expected/Observed = 6.4 for stop-gain

MLL2 is among top 2% of genes in human genome ranked by LoF constraint



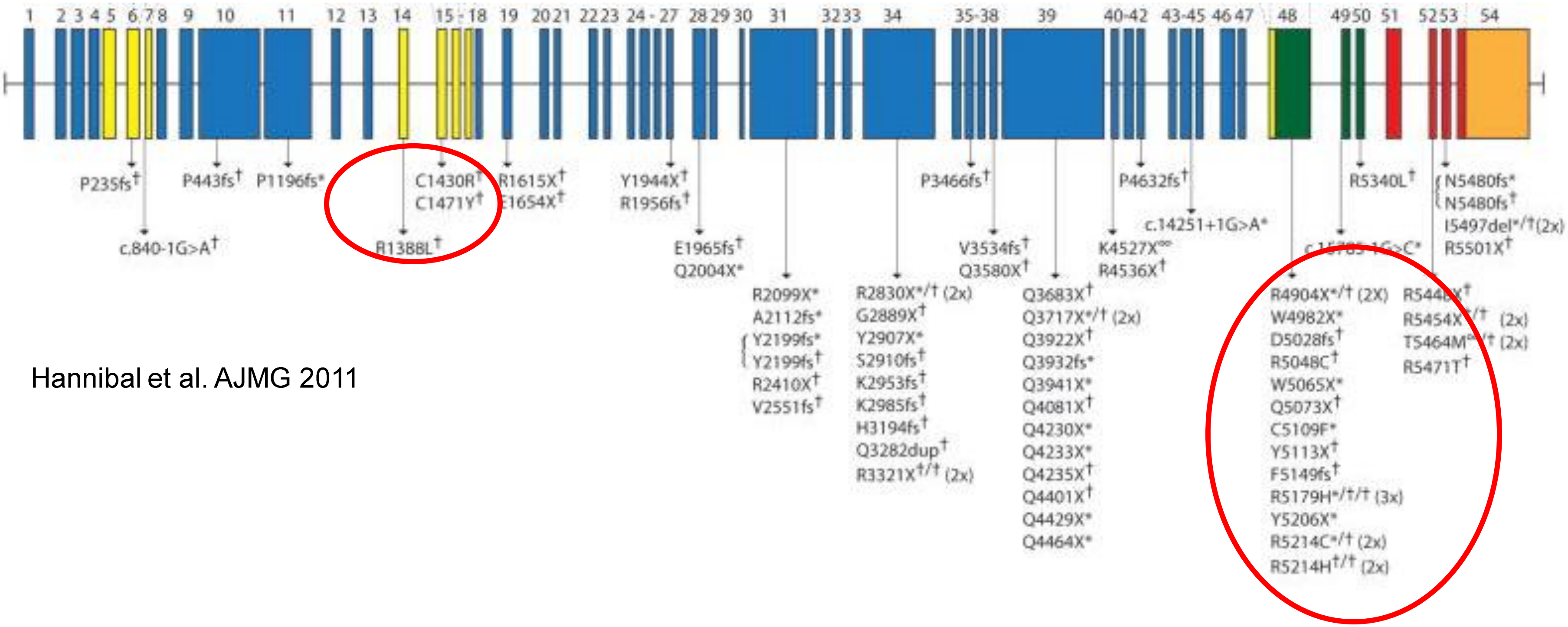
Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Synonymous	792.9	919	z = -2.78
Missense	1842.9	1571	z = 3.10
LoF	137.6	11	pLI = 1.00

Exome data from 65,000 individuals

- More than 3000 genes have pLI > 0.9
- Doesn’t imply causality but useful for prioritization
- If mutation is ‘de novo’, more likely to be pathogenic

http://exac.broadinstitute.org/gene/ENSG00000167548

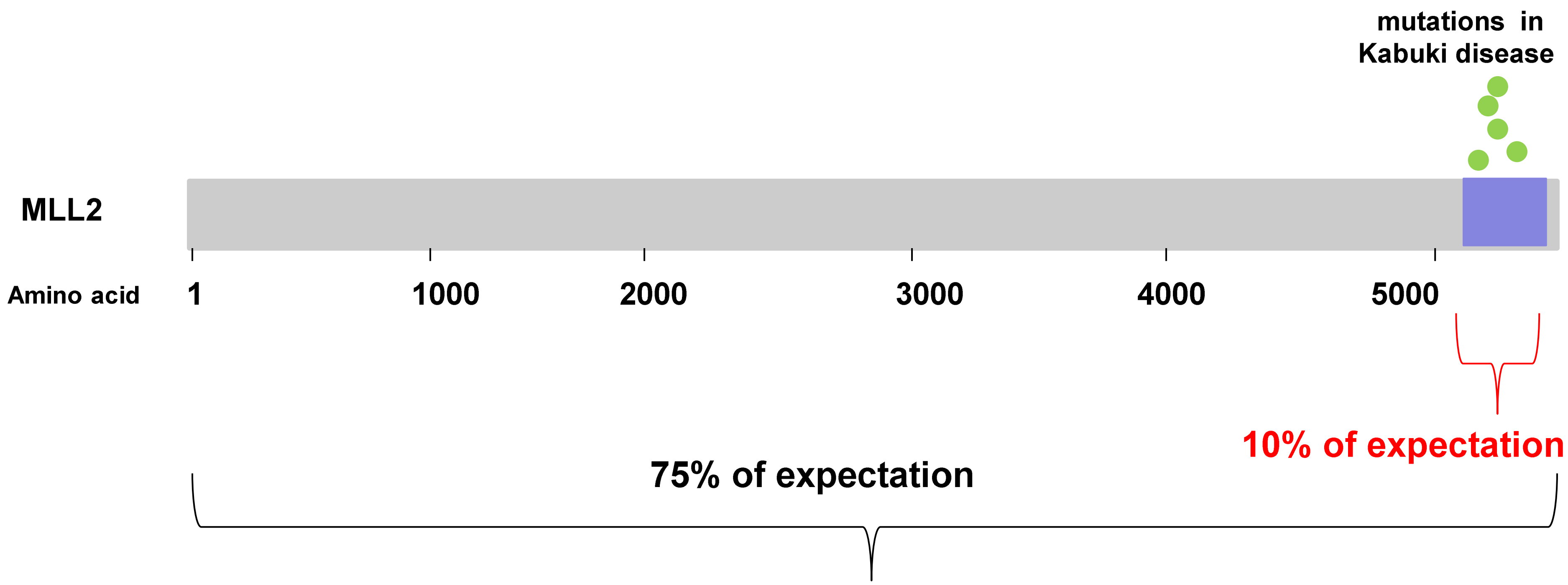
Missense mutations in MLL2



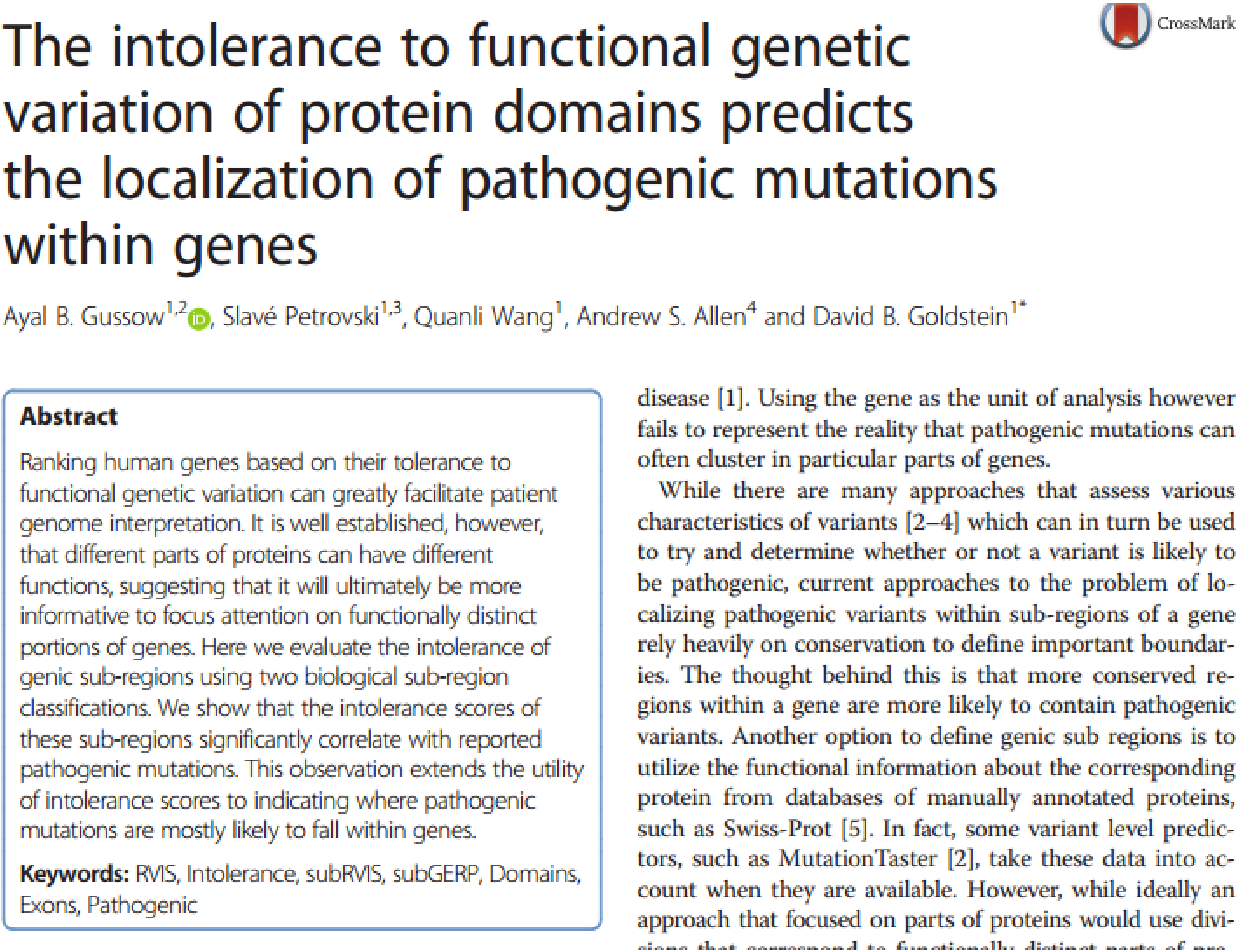
Hannibal et al. AJMG 2011

- Missense mutations in some exons cause Kabuki syndrome
- 1/120 individuals in population carriers of missense mutations

Prioritizing missense mutations in MLL2



- Significantly lower frequency of missense mutations in 5340-5537 region of MLL2 protein using ExAc data



The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes

Ayal B. Gussow[1,2], Slavé Petrovski[1,3], Quanli Wang[1], Andrew S. Allen[4] and David B. Goldstein[1*]

Abstract

Ranking human genes based on their tolerance to functional genetic variation can greatly facilitate patient genome interpretation. It is well established, however, that different parts of proteins can have different functions, suggesting that it will ultimately be more informative to focus attention on functionally distinct portions of genes. Here we evaluate the intolerance of genic sub-regions using two biological sub-region classifications. We show that the intolerance scores of these sub-regions significantly correlate with reported pathogenic mutations. This observation extends the utility of intolerance scores to indicating where pathogenic mutations are mostly likely to fall within genes.

Keywords: RVIS, Intolerance, subRVIS, subGERP, Domains, Exons, Pathogenic

disease [1]. Using the gene as the unit of analysis however fails to represent the reality that pathogenic mutations can often cluster in particular parts of genes.

While there are many approaches that assess various characteristics of variants [2–4] which can in turn be used to try and determine whether or not a variant is likely to be pathogenic, current approaches to the problem of localizing pathogenic variants within sub-regions of a gene rely heavily on conservation to define important boundaries. The thought behind this is that more conserved regions within a gene are more likely to contain pathogenic variants. Another option to define genic sub regions is to utilize the functional information about the corresponding protein from databases of manually annotated proteins, such as Swiss-Prot [5]. In fact, some variant level predictors, such as MutationTaster [2], take these data into account when they are available. However, while ideally an approach that focused on parts of proteins would use divisions that correspond to functionally distinct parts of pro-

4. Integrating genetic, gene-expression and model organism data

affected individuals and related individuals

sequencing

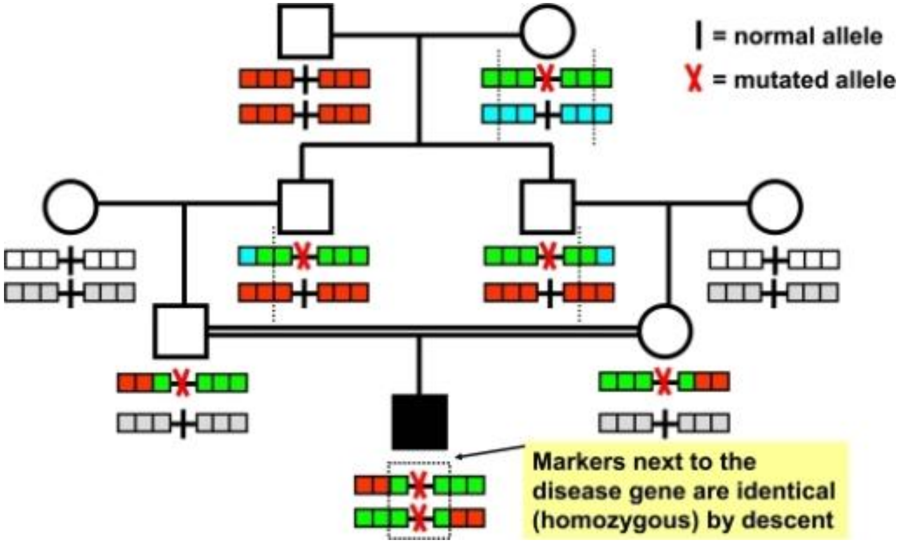
Variant calling

Homozygosity mapping

Variants shared by affected Individuals in HBD regions

Retain variants in genes expressed in disease relevant tissues

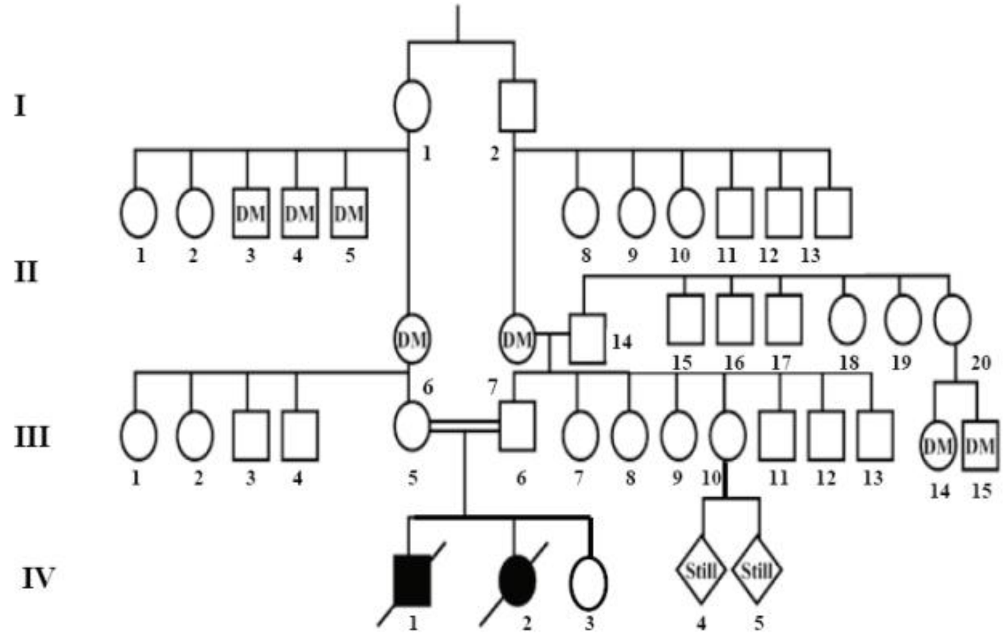
Validation by phenotyping of mouse knockout



Markers next to the disease gene are identical (homozygous) by descent

Hildebrandt et al, Plos. Gen 2009

Mitchell-Riley syndrome



- Neonatal diabetes, diarrhoea, intestinal atresia in two individuals from consanguineous family

Smith et al. Nature 2010

Homozygosity mapping

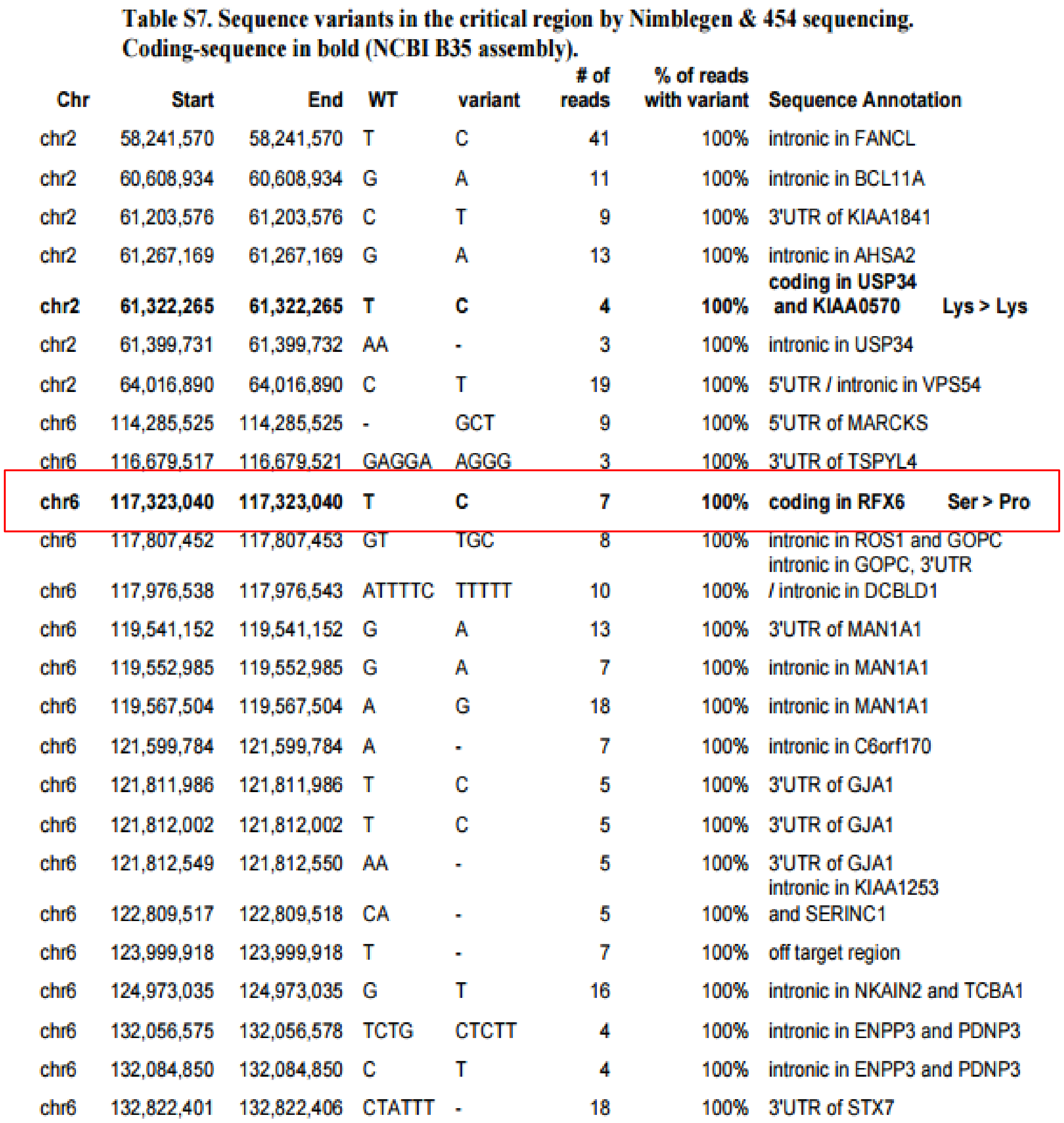


Table S7. Sequence variants in the critical region by Nimblegen & 454 sequencing. Coding-sequence in bold (NCBI B35 assembly).

Chr	Start	End	WT	variant	# of reads	% of reads with variant	Sequence Annotation	
chr2	58,241,570	58,241,570	T	C	41	100%	intronic in FANCL	
chr2	60,608,934	60,608,934	G	A	11	100%	intronic in BCL11A	
chr2	61,203,576	61,203,576	C	T	9	100%	3'UTR of KIAA1841	
chr2	61,267,169	61,267,169	G	A	13	100%	intronic in AHSA2	
chr2	**61,322,265**	**61,322,265**	**T**	**C**	**4**	**100%**	**coding in USP34 and KIAA0570**	**Lys > Lys**
chr2	61,399,731	61,399,732	AA	-	3	100%	intronic in USP34	
chr2	64,016,890	64,016,890	C	T	19	100%	5'UTR / intronic in VPS54	
chr6	114,285,525	114,285,525	-	GCT	9	100%	5'UTR of MARCKS	
chr6	116,679,517	116,679,521	GAGGA	AGGG	3	100%	3'UTR of TSPYL4	
chr6	**117,323,040**	**117,323,040**	**T**	**C**	**7**	**100%**	**coding in RFX6**	**Ser > Pro**
chr6	117,807,452	117,807,453	GT	TGC	8	100%	intronic in ROS1 and GOPC	
chr6	117,976,538	117,976,543	ATTTTC	TTTTT	10	100%	intronic in GOPC, 3'UTR / intronic in DCBLD1	
chr6	119,541,152	119,541,152	G	A	13	100%	3'UTR of MAN1A1	
chr6	119,552,985	119,552,985	G	A	7	100%	intronic in MAN1A1	
chr6	119,567,504	119,567,504	A	G	18	100%	intronic in MAN1A1	
chr6	121,599,784	121,599,784	A	-	7	100%	intronic in C6orf170	
chr6	121,811,986	121,811,986	T	C	5	100%	3'UTR of GJA1	
chr6	121,812,002	121,812,002	T	C	5	100%	3'UTR of GJA1	
chr6	121,812,549	121,812,550	AA	-	5	100%	3'UTR of GJA1	
chr6	122,809,517	122,809,518	CA	-	5	100%	intronic in KIAA1253 and SERINC1	
chr6	123,999,918	123,999,918	T	-	7	100%	off target region	
chr6	124,973,035	124,973,035	G	T	16	100%	intronic in NKAIN2 and TCBA1	
chr6	132,056,575	132,056,578	TCTG	CTCTT	4	100%	intronic in ENPP3 and PDNP3	
chr6	132,084,850	132,084,850	C	T	4	100%	intronic in ENPP3 and PDNP3	
chr6	132,822,401	132,822,406	CTATTT	-	18	100%	3'UTR of STX7	

Gene expression analysis

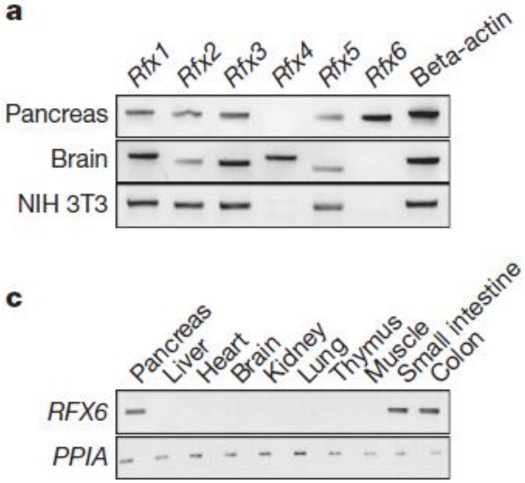


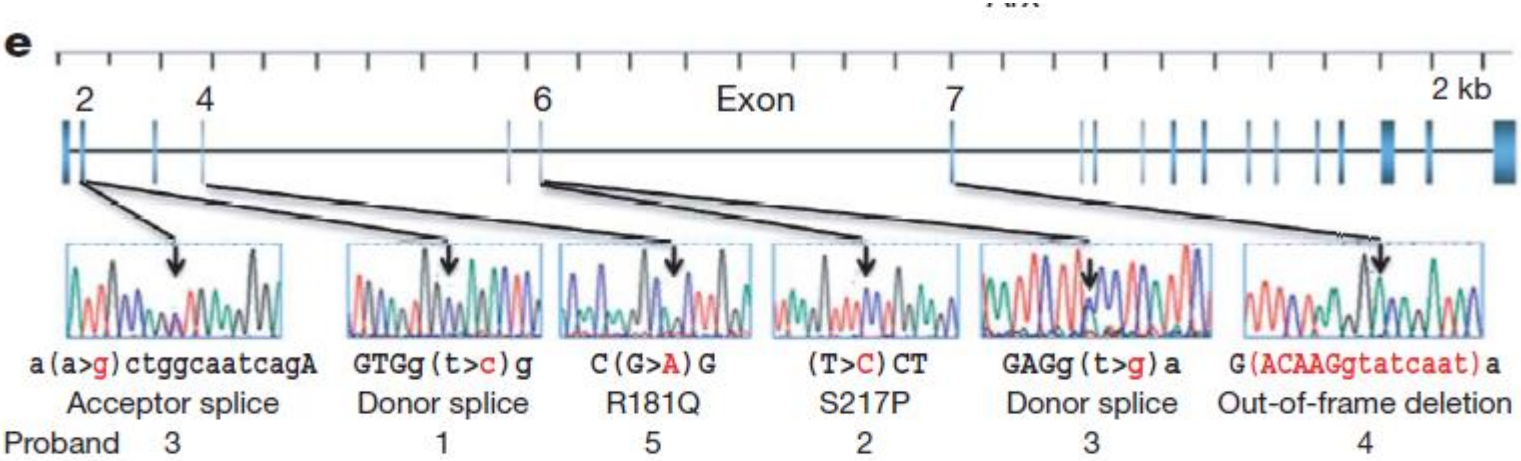
Figure 1 | Expression of *Rfx6* in mice and human

Mouse knockout phenotype

RFX6-null mice lack pancreatic islet cells, have intestinal atresia and fail to survive

Smith et al. Nature 2010

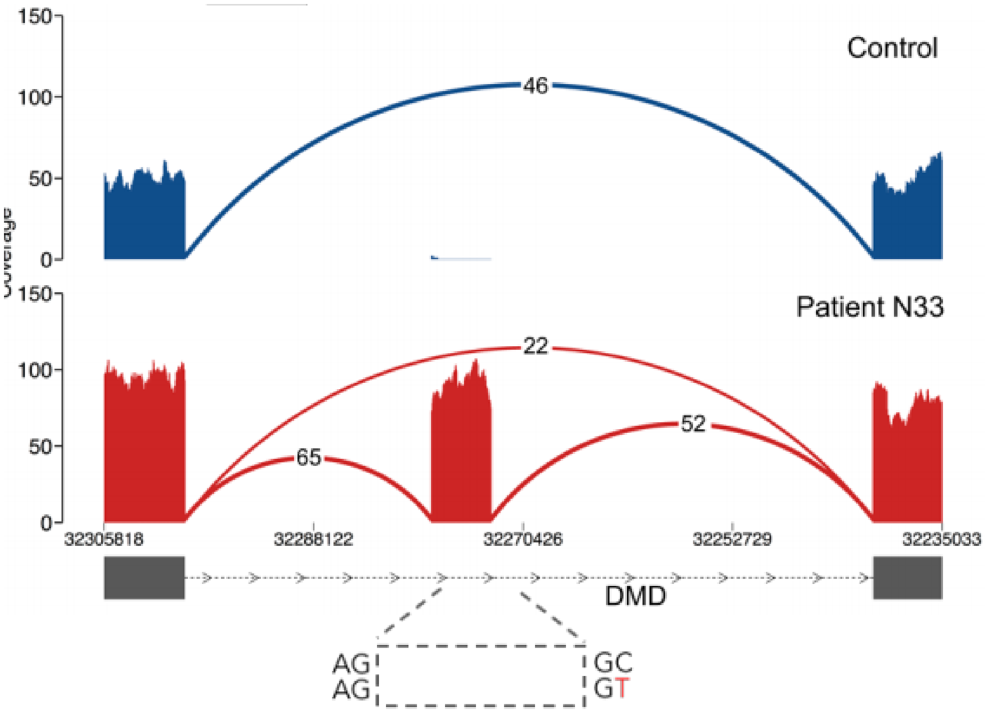
Additional mutations in RFX6 confirm association



- Gene could also have been identified by whole-exome sequencing of the six “unrelated individuals” with Mitchell-Riley syndrome

Smith et al. Nature 2010

5. Combining DNA-seq and RNA-seq in a single individual



- Mutation activates cryptic splice site and pseudo-exon added to transcript
- Difficult to predict using computational tools
- DNA and RNA-seq data on individual(s) with phenotype can identify causal variant

http://biorxiv.org/content/biorxiv/early/2016/09/08/074153.full.pdf

Prioritizing variants and genes for disease

1. **Variant annotation:** How deleterious is the mutation

   PolyPhen/SIFT
   CADD score

2. **Familial segregation**: how well variant segregates with phenotype in family data

   Gemini

3. **Gene-level constraint**: human population data

   ExAc database

4. **Gene expression** : is the expression of the gene high in or limited to disease relevant tissues

   GTEX database

5. **Model organism data:** does loss of gene or mutation lead to similar phenotype

   Mouse Phenotyping Consortium

6. **Statistical association:** does gene contain mutations in multiple affected individuals