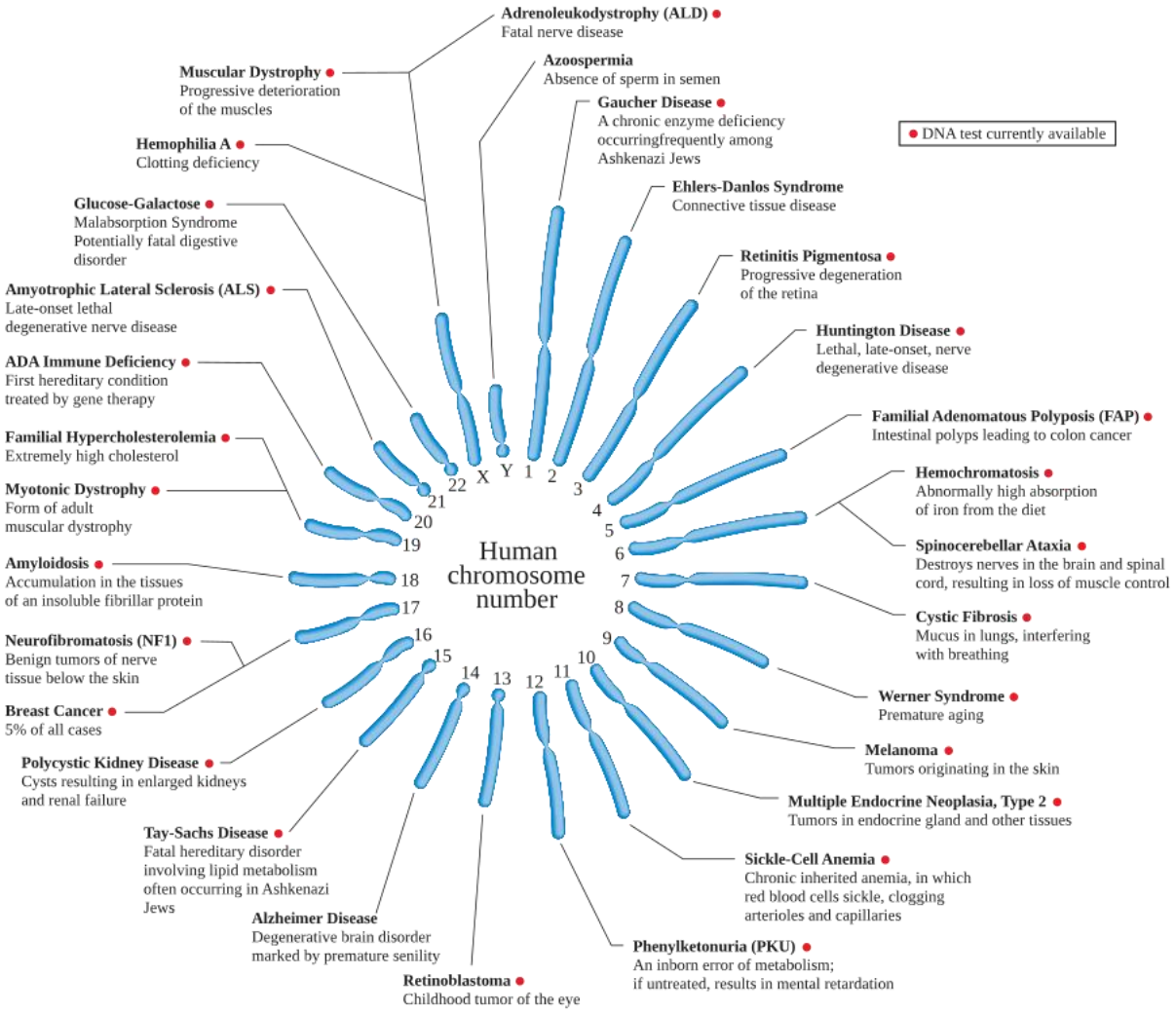
Using high-throughput DNA sequencing and bioinformatics to search for disease mutations

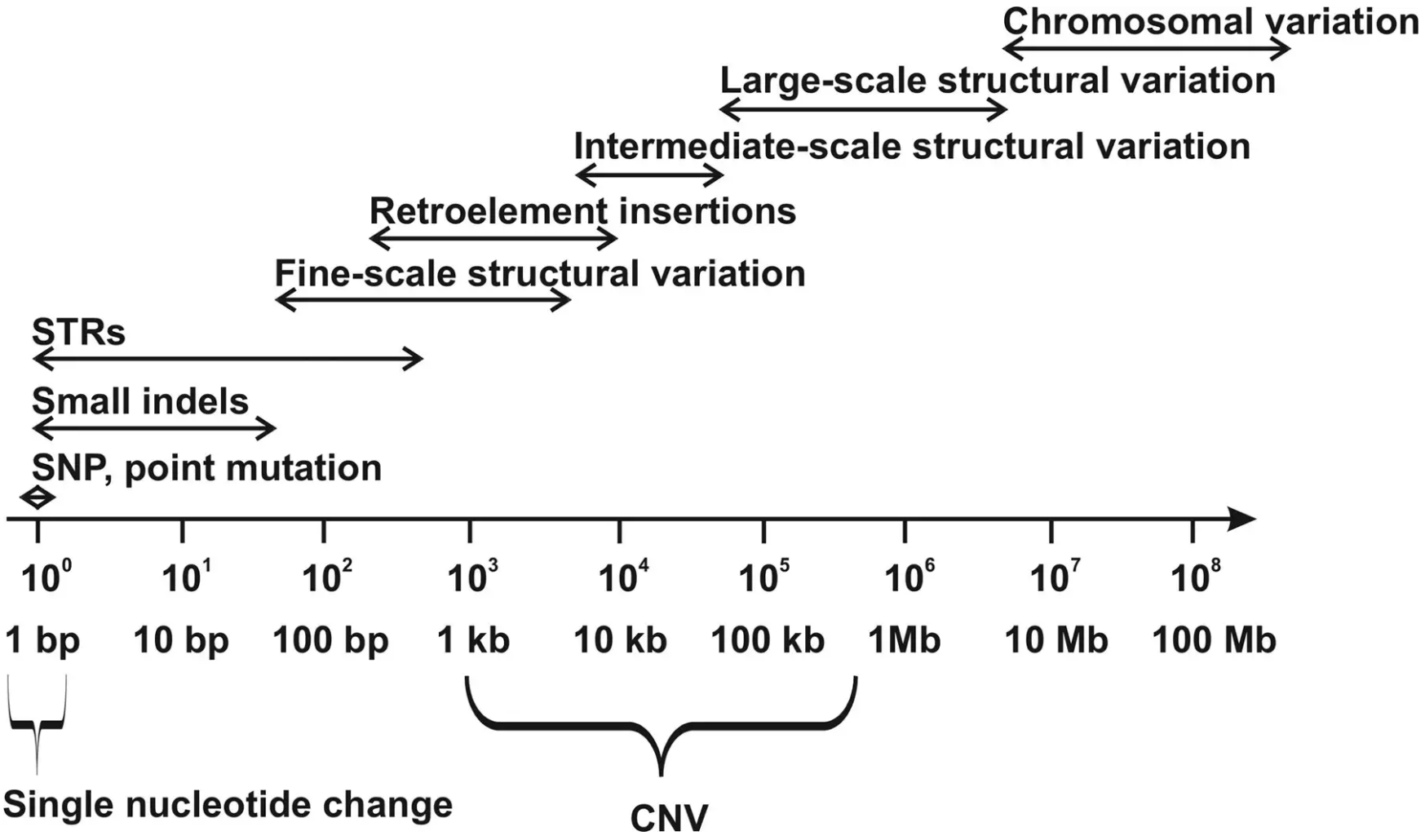
Vikas Bansal, Ph.D.
Department of Pediatrics

MED263
February 13th 2018

Human Genetic Disorders

- **> 7000 genetic disorders**
- **Genes found for 4300 rare diseases**





http://circ.ahajournals.org/content/115/24/3130/tab-figures-data

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
Reads of length 100-250 bases

- Sequencing can “potentially” detect variation of all types and sizes
  - 1 bp deletion….. single exon deletion …..chromosome deletion

DNA sequencing to understand disease

Individuals with/without disease (DNA)

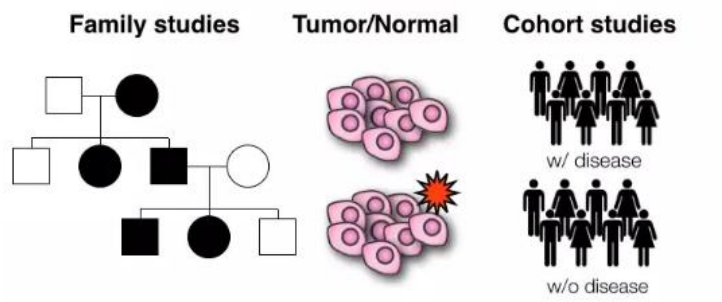
1

Raw sequence data

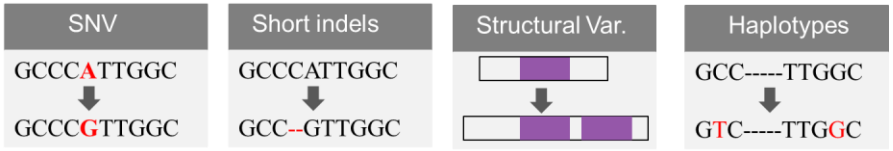
Variant calls (VCF)

2

Disease causing variants and genes

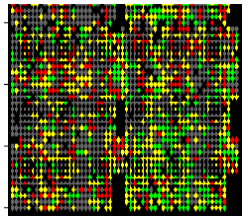


ACGTAGCTAGCGC
GCATCGACGAGA GTAGTAGTAGTGT
TCGATGGATC CCTTGAGTTAGCC



Variant annotation

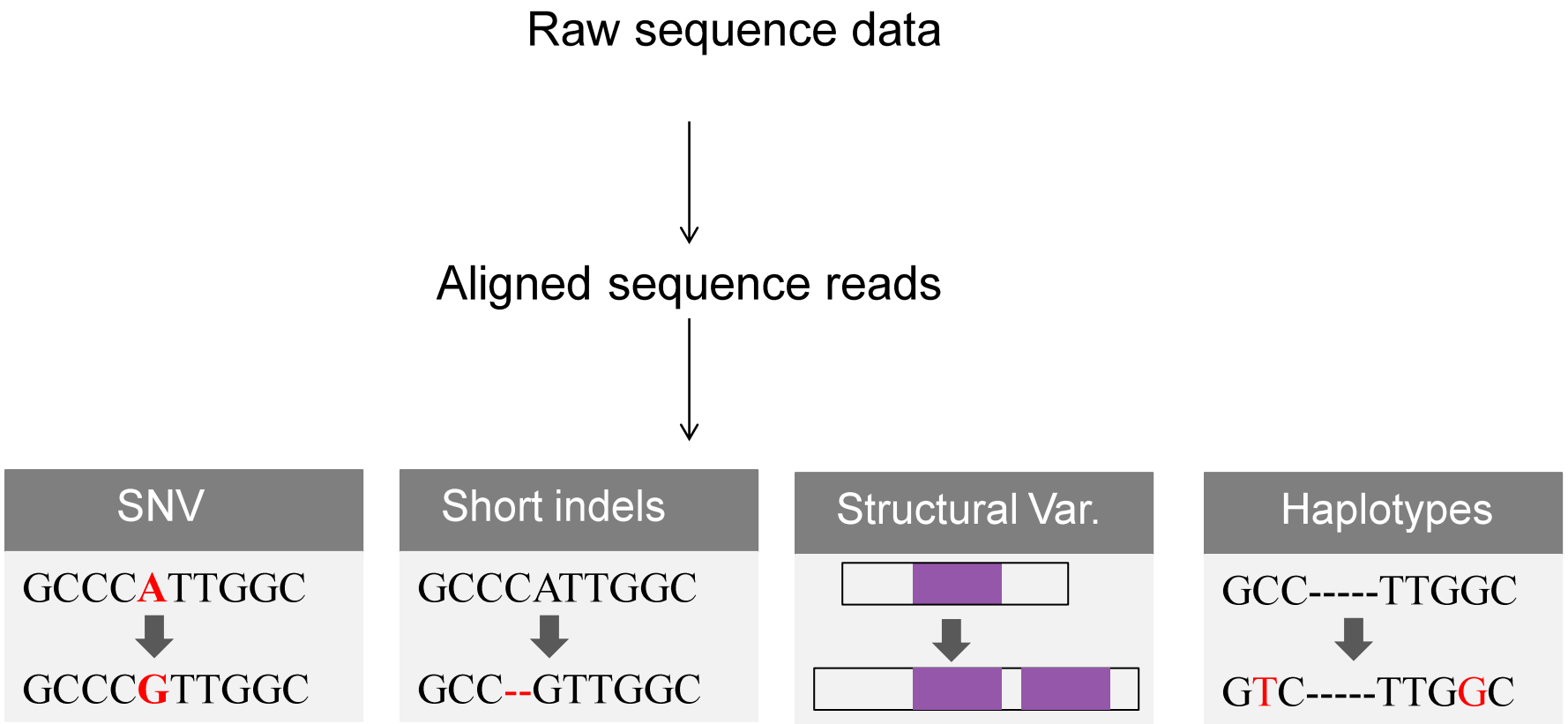
Model organism gene knockout



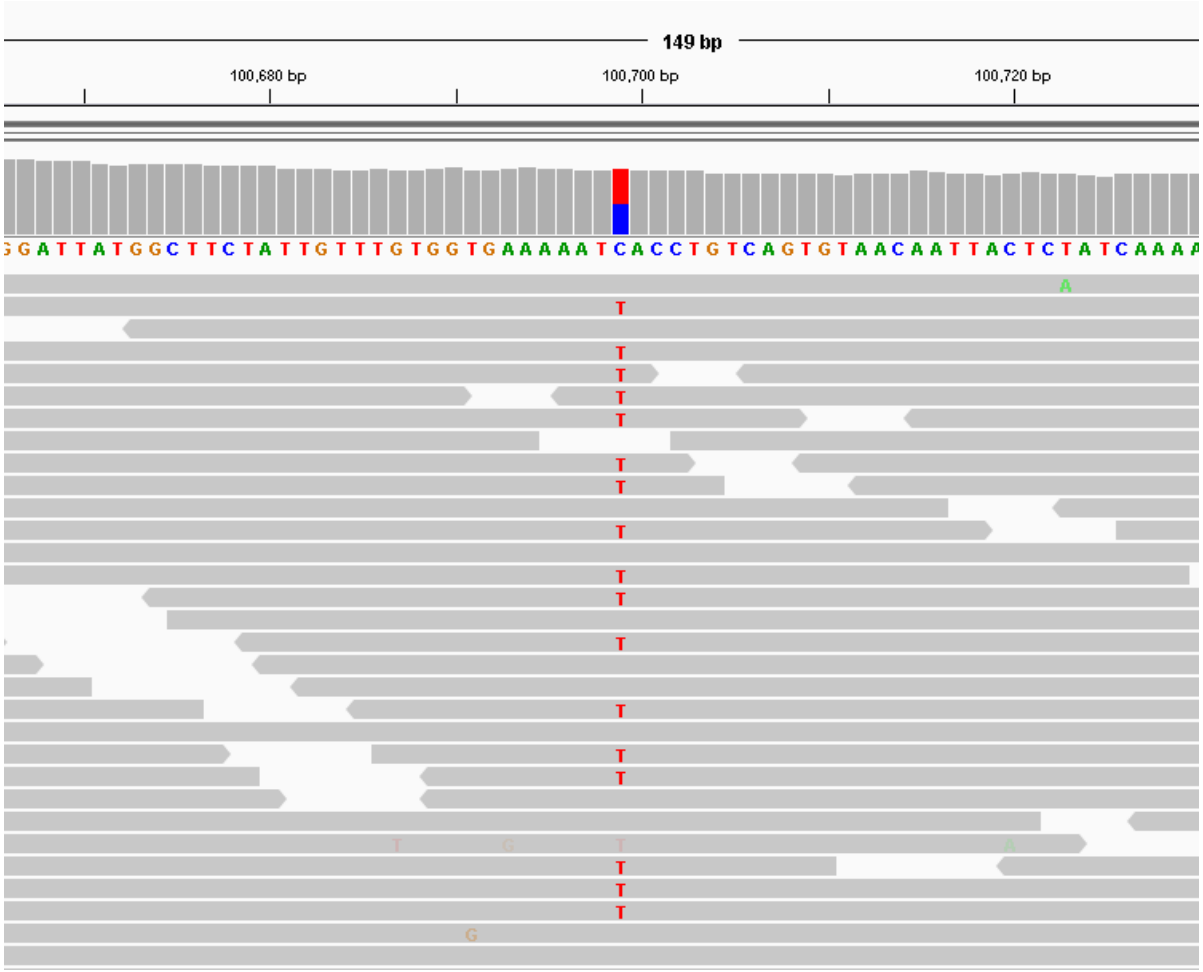
Population	Allele Frequency
Other	0.001104
European (Non-Finnish)	0.0008406
African	9.647e-05
South Asian	6.056e-05

1. Discovery of sequence variants

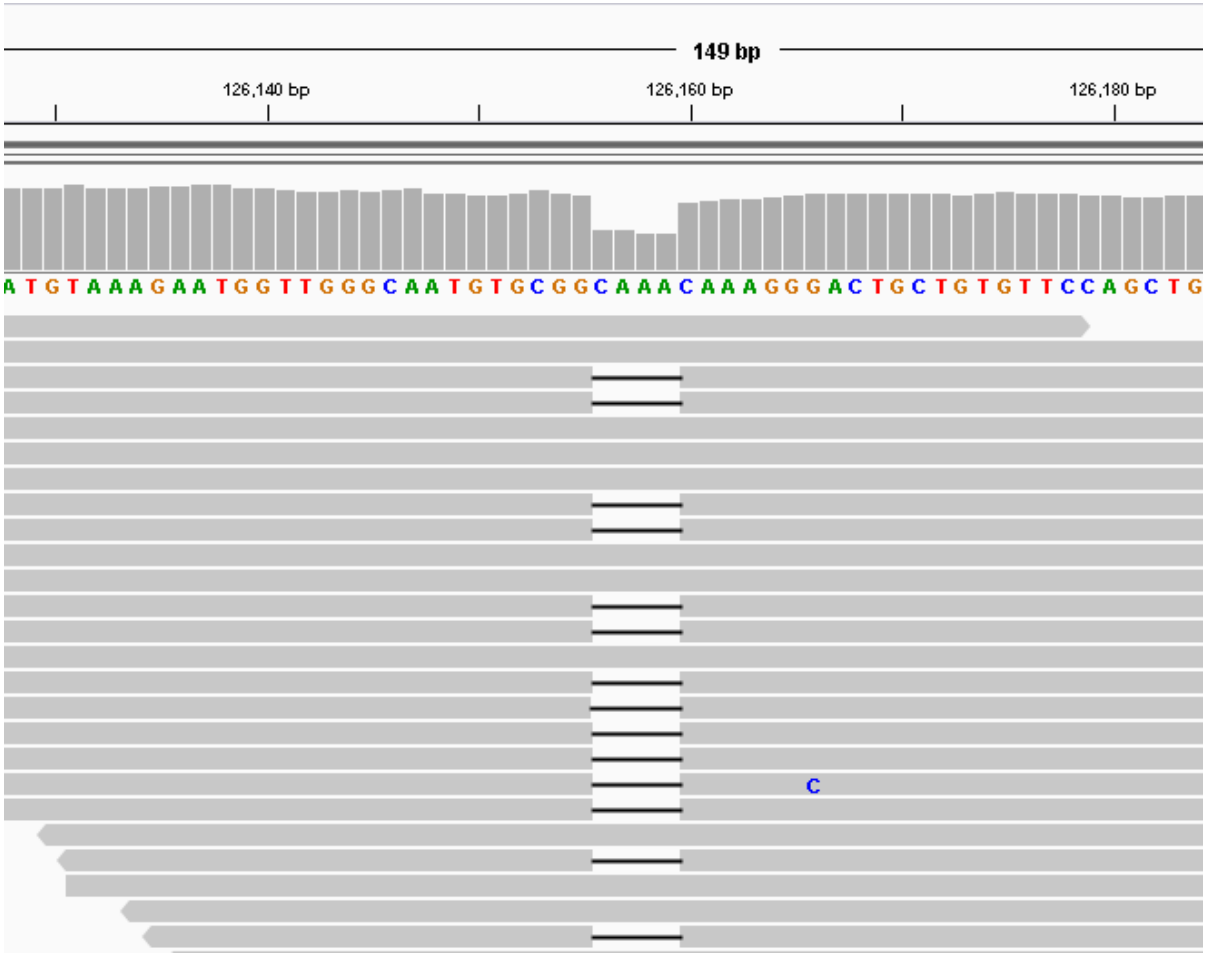
Goal: Identify differences between the sequenced genome (represented in the form of short reads) and a ‘reference’ genome



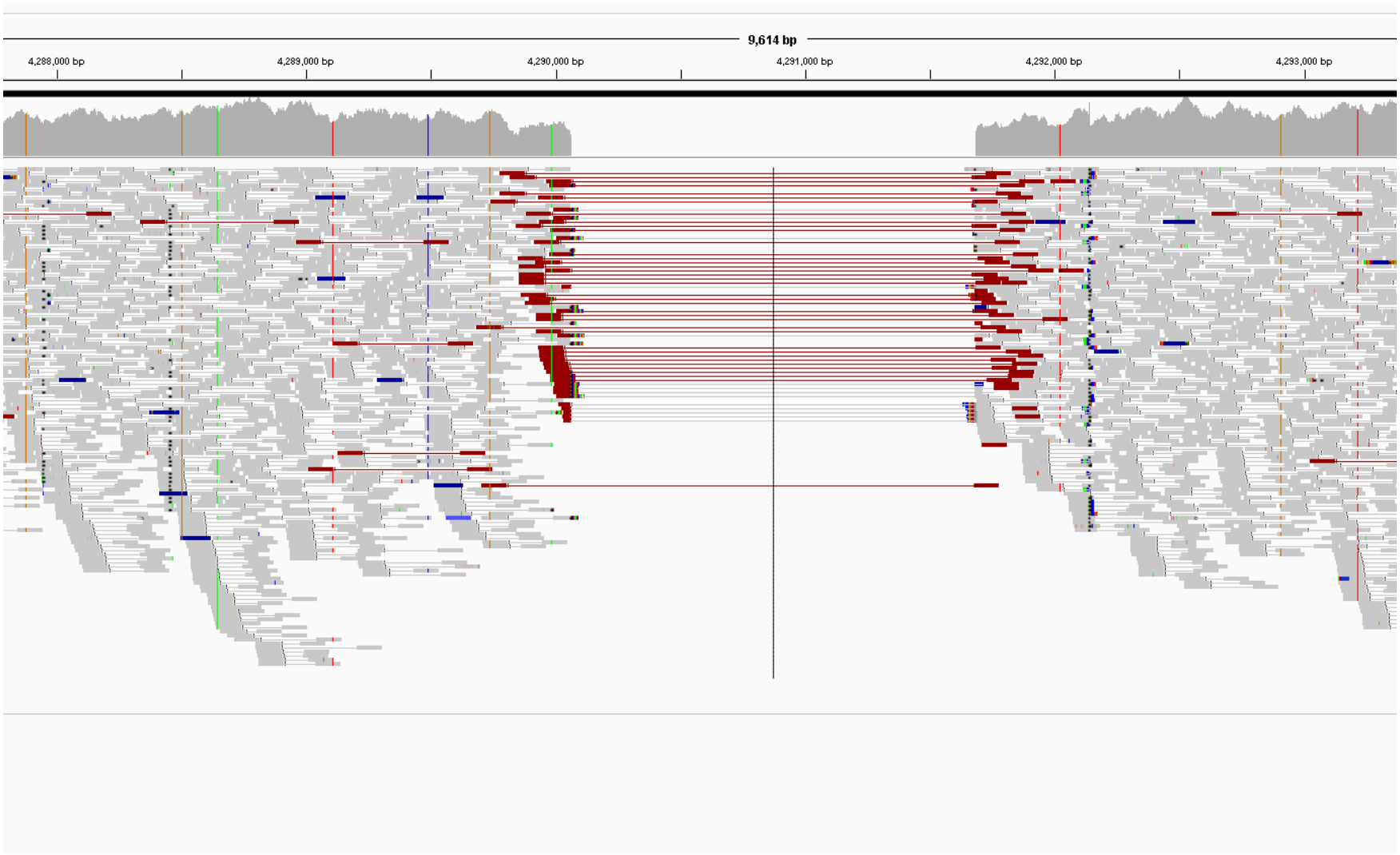
SNVs from aligned reads



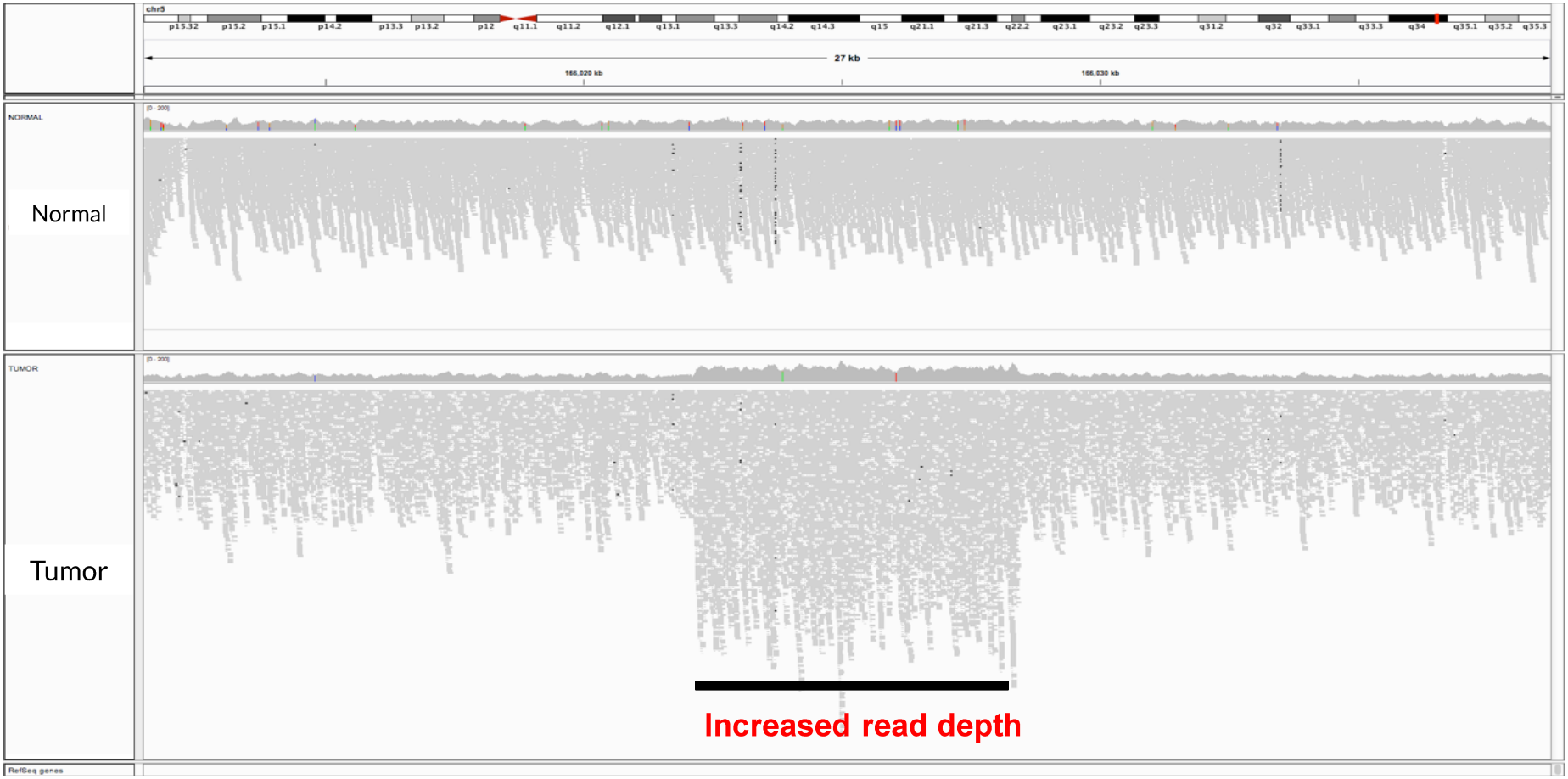
Small indels from aligned reads



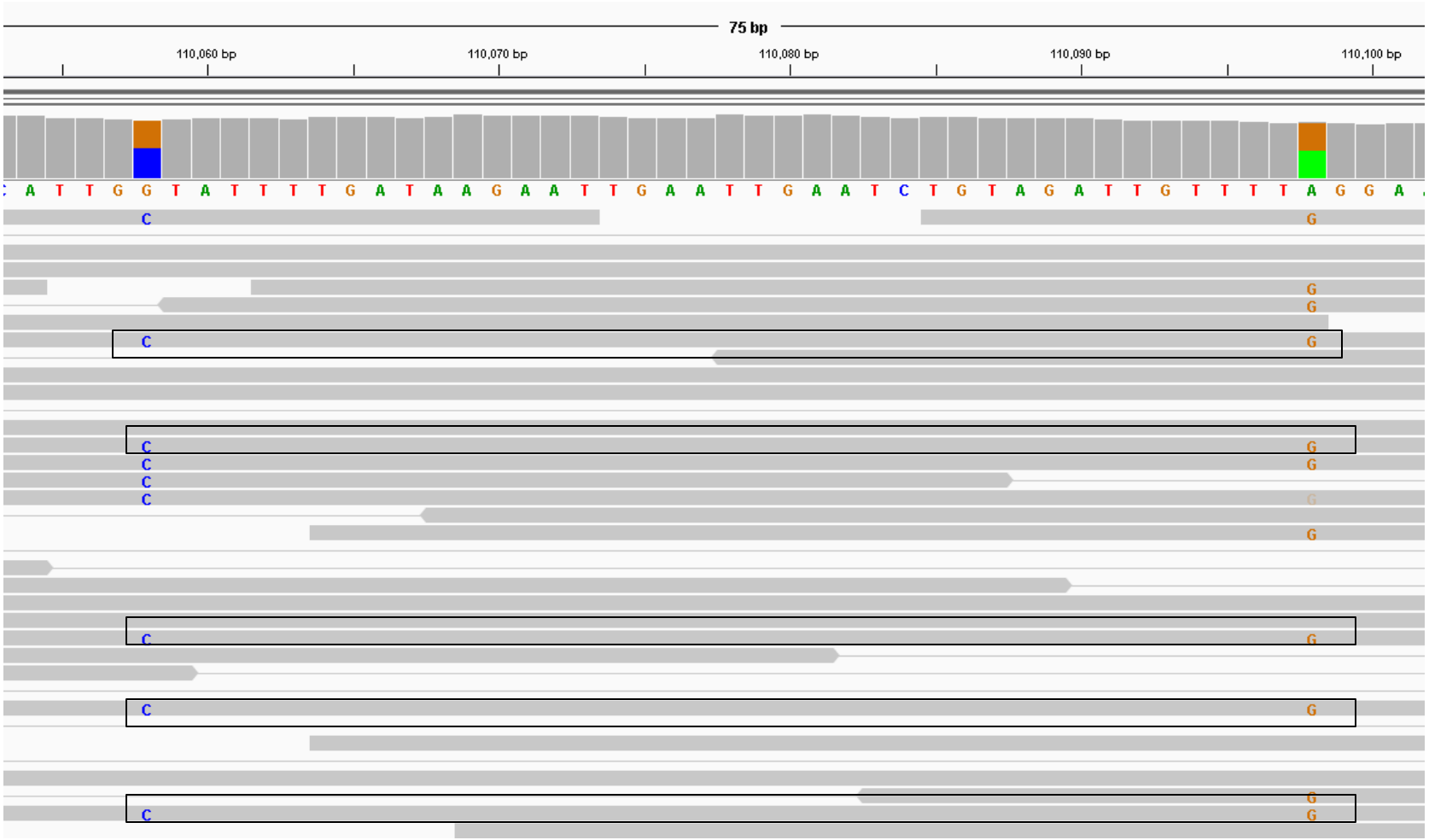
Long deletions from aligned reads



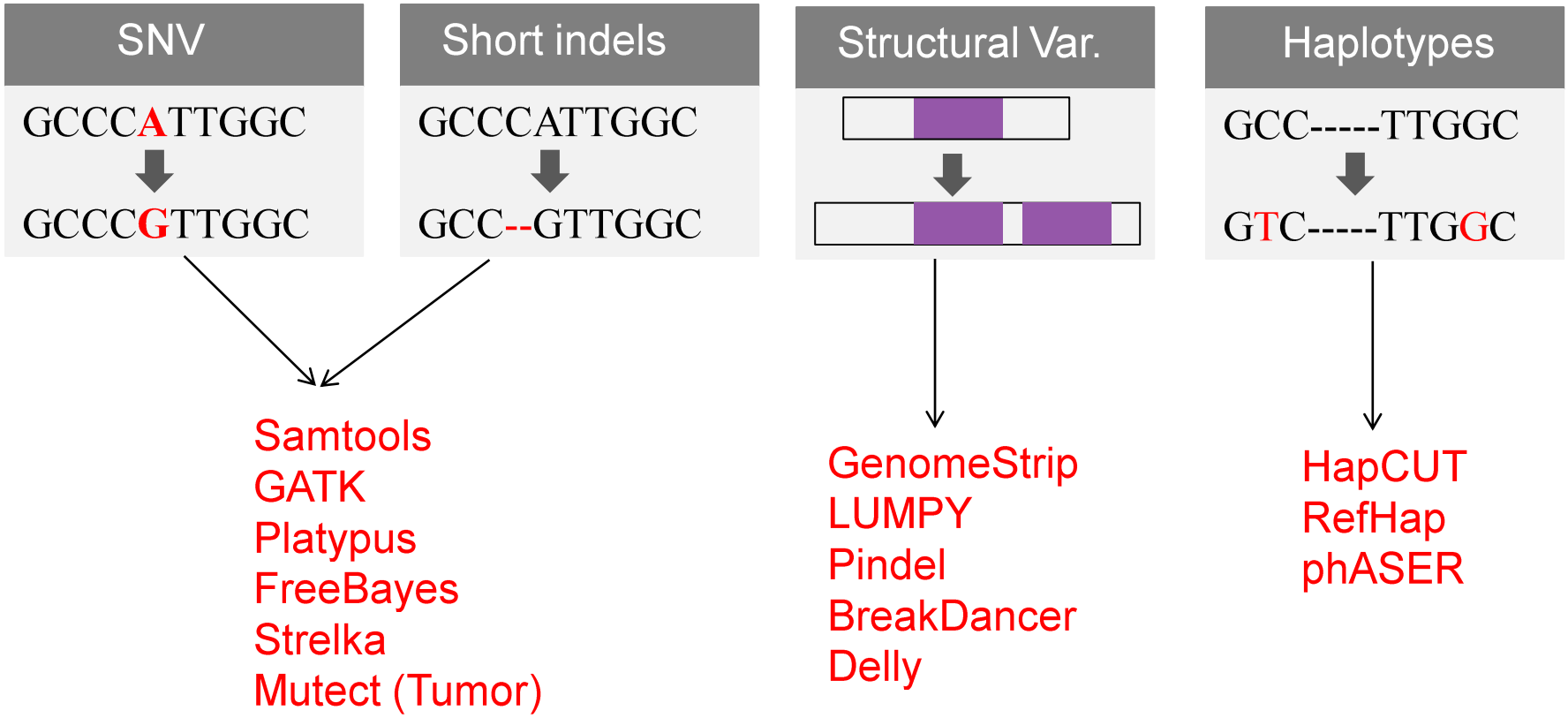
Copy number changes from read-depth



Phasing of heterozygous variants



1. Tools for variant calling



Question

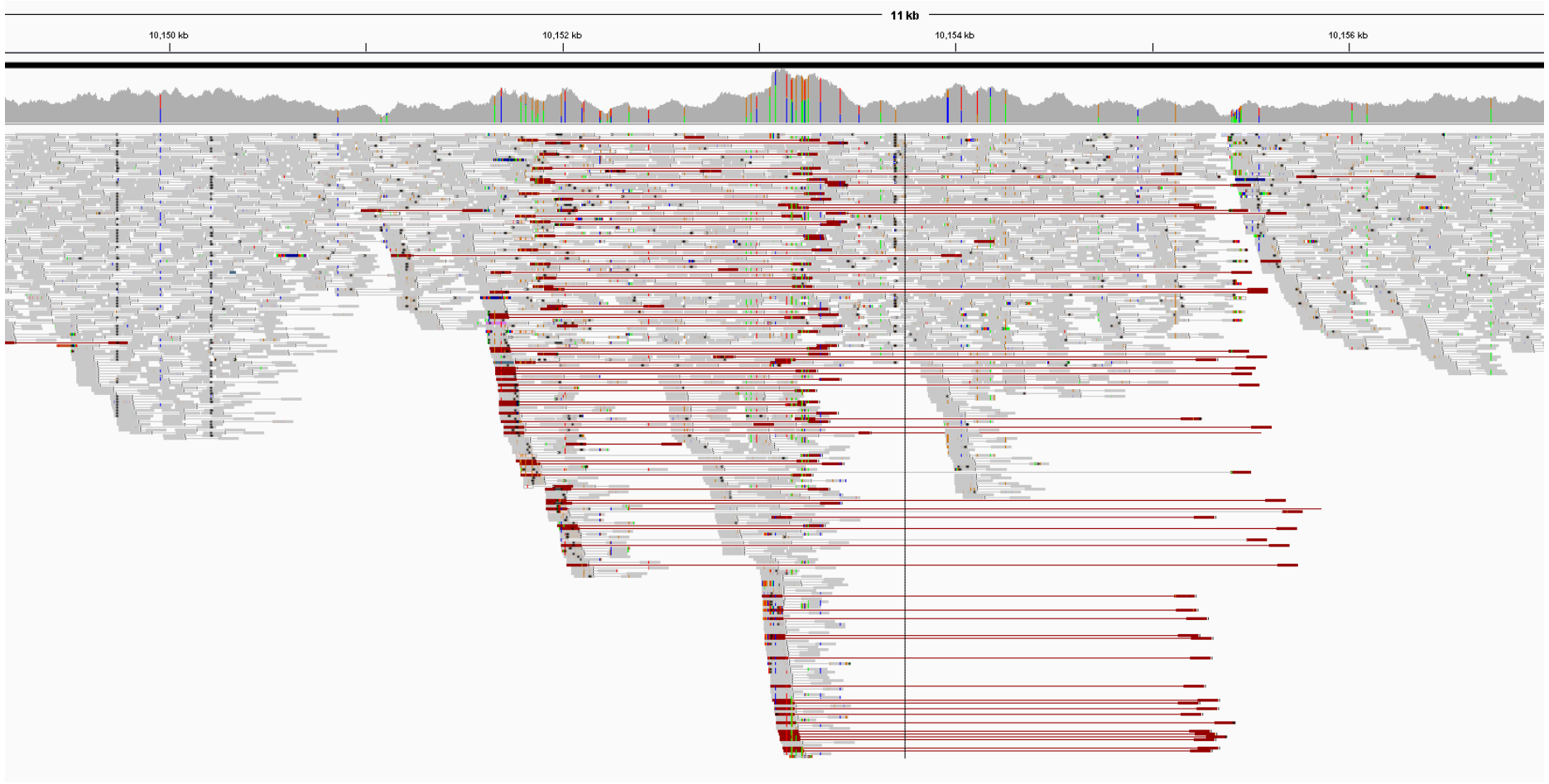
Can whole-chromosome duplications be detected using NGS ?

Challenges in variant detection

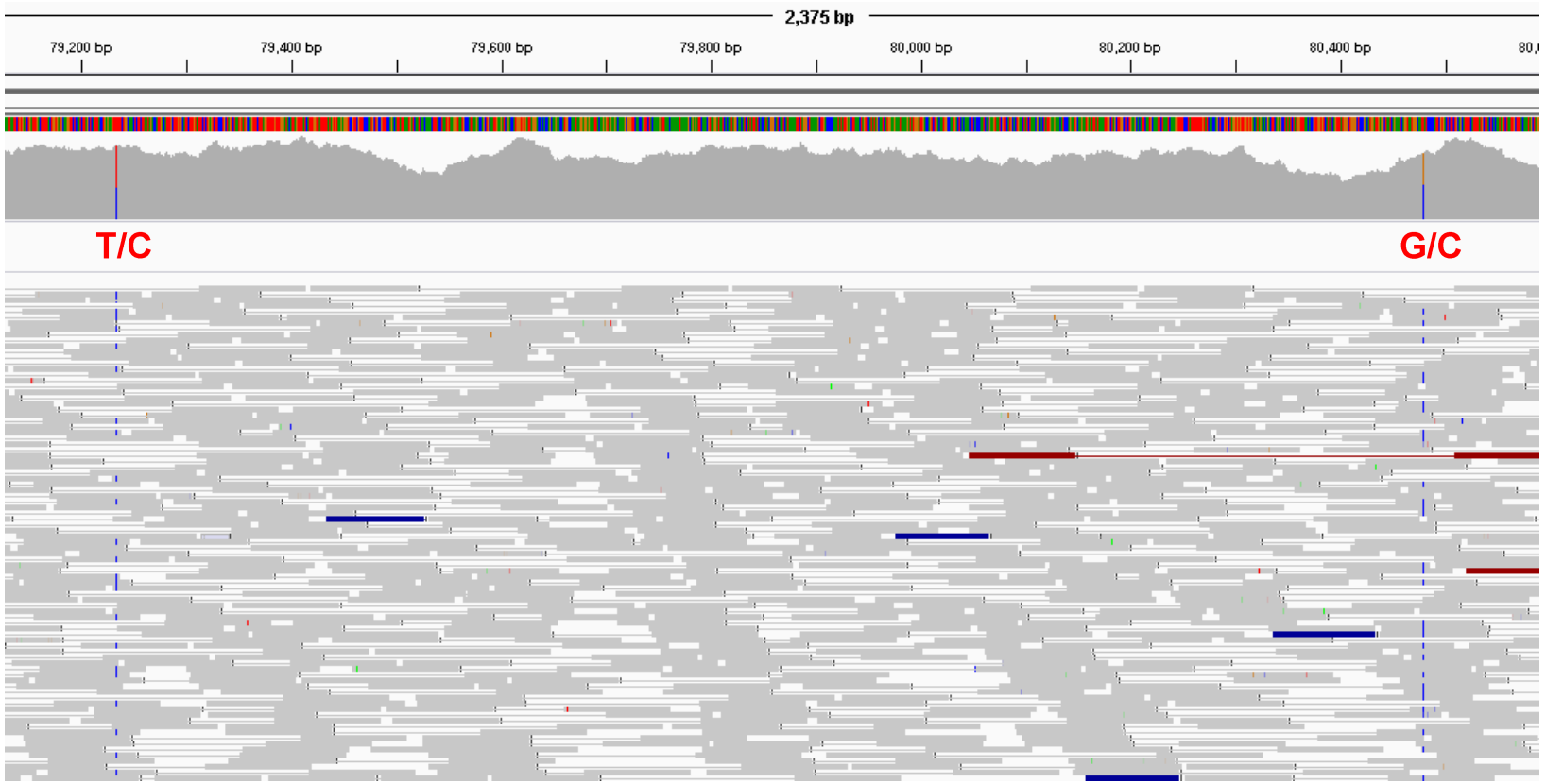
- Systematic sequencing and alignment errors
- Indels and structural variants over-represented in repeats and low-complexity sequence

- Limitations of short read lengths
  1. Structural variants require long-range information
  2. Haplotype information lost during fragmentation
  3. 8-10% of the genome not uniquely mappable

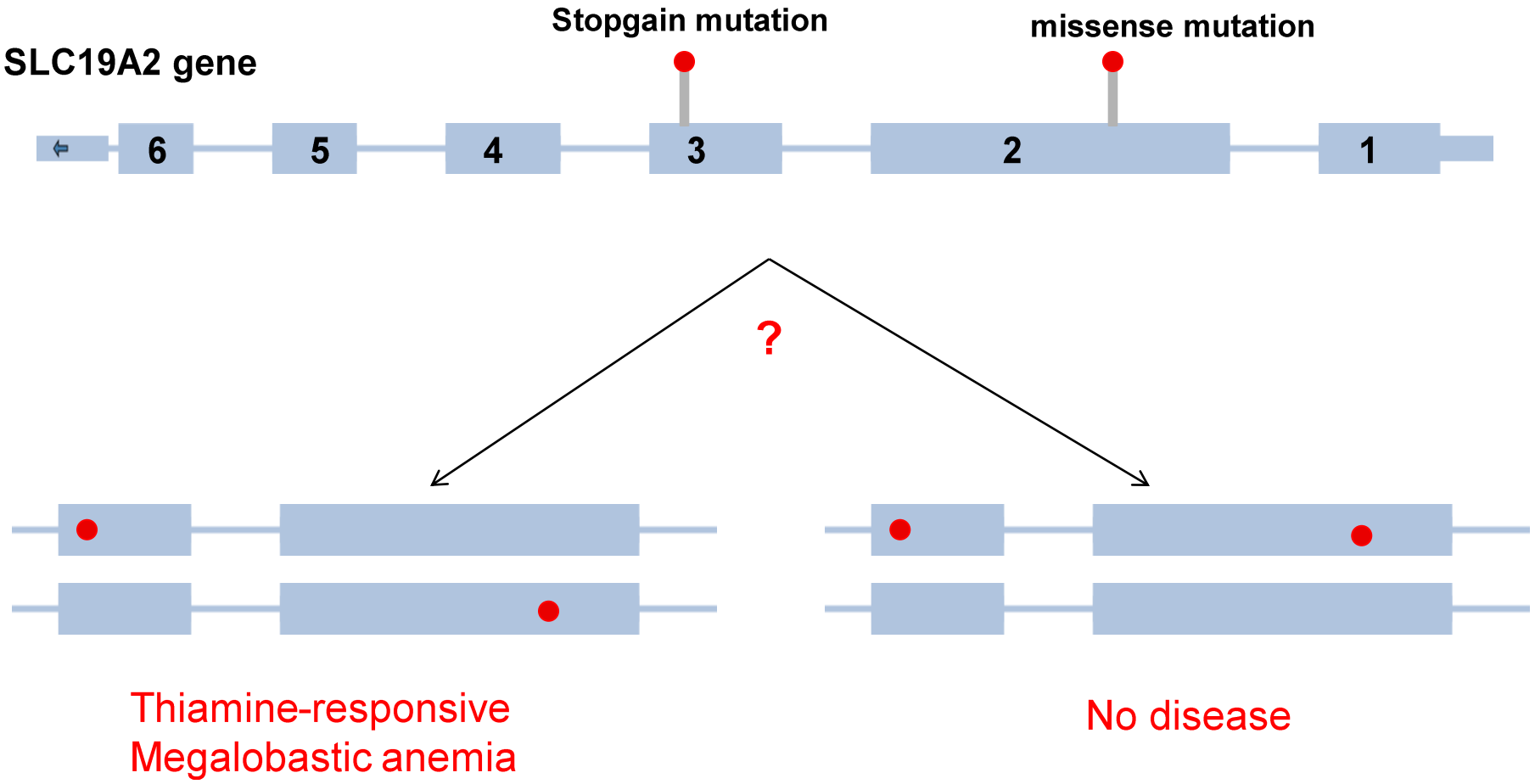
Complex signal at loci with structural variation



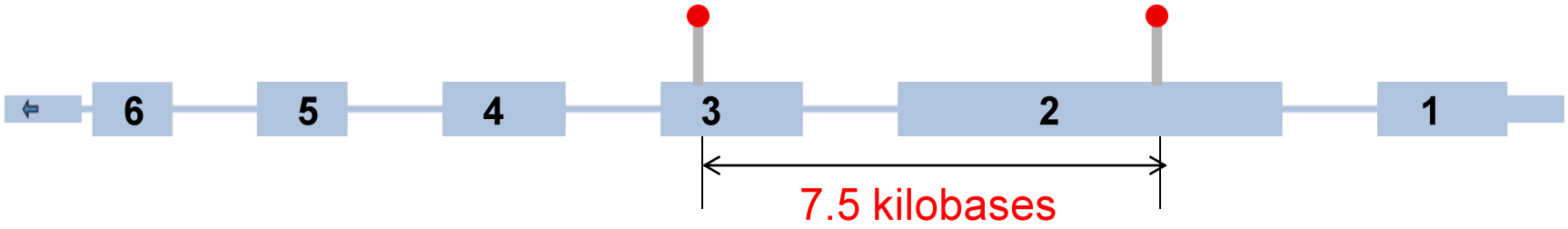
Distant heterozygous variants cannot be phased



Long-range phasing is important

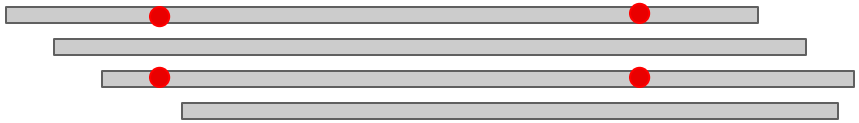


Long-range phasing

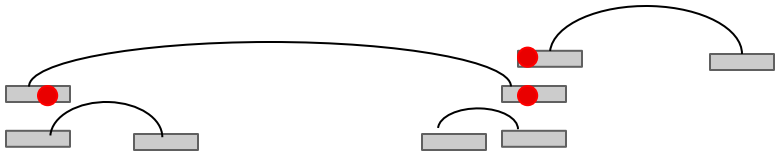


Illumina sequencing

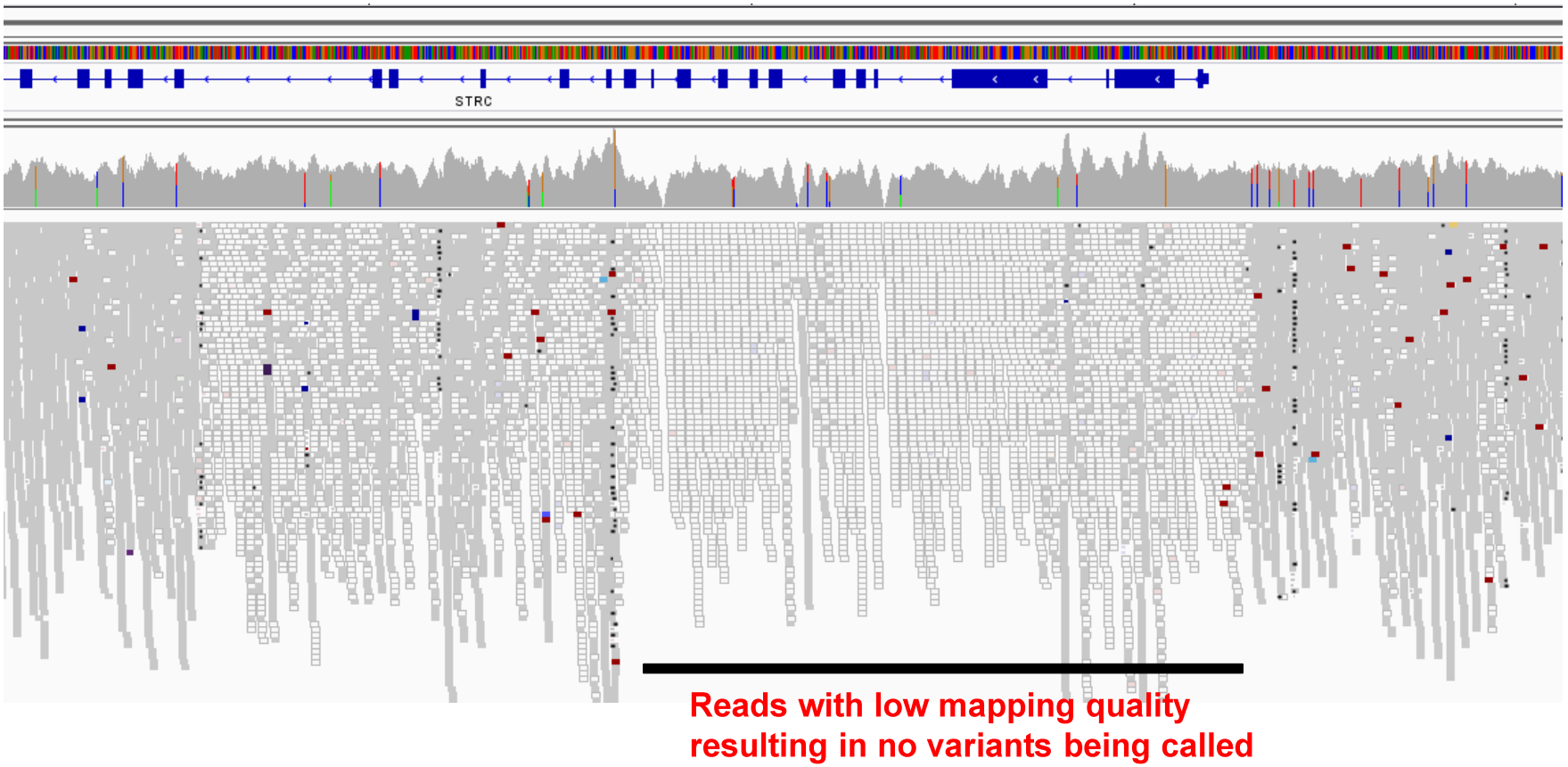
Long read sequencing



Long-insert Illumina sequencing



Variants in duplicated genes cannot be detected using short reads



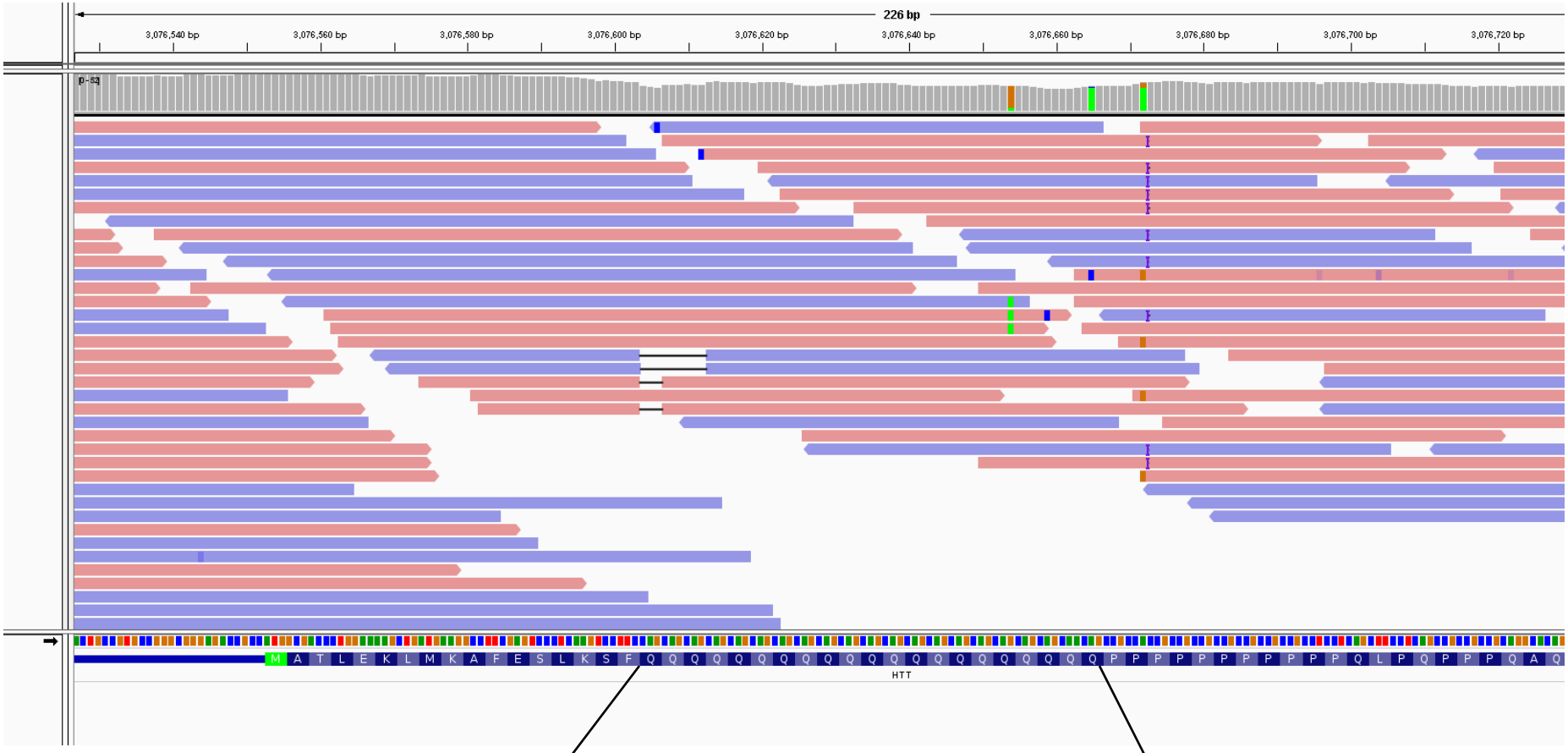
- Bi-allelic mutations in STRC cause sensorineural hearing loss

100+ genes relevant to human disease are problematic for short reads

Gene	Total Exons	# homologous exons	# of homology hits	Main disease	Onset	Prevalence	Evidence level (3,2,1,0)
GBA	12	5	1	Gaucher Disease	pediatric	1/855 (Ashkenazi Jews)	3
CYP21A2	10	8	1	Congenital adrenal hyperplasia	pediatric	~1/10000	3
VWF	52	6	2	von Willebrand disease	pediatric-adult	1/100-1/1000	3
PMS2	15	5	1	Lynch Syndrome	adolescence-adulthood	1/440	3
SMN1	9	9	1	Spinal Muscular Atrophy	pediatric	1/10,000	3
SMN2	9	9	1	Spinal Muscular Atrophy	pediatric	1/10,000	3
HBA1	3	2	1	Hemoglobinopathy			3
HBA2	3	2	1	Hemoglobinopathy			3
STRC	29	23	1	Autosomal recessive sensorineural hearing loss	pediatric	1/1000	3
CFC1	6	6	1	Congenital heart defects	pediatric	~1/100	3
HYDIN	86	78	2	Ciliary dyskinesia, primary, 5	pediatric	1/16,000	3
OTOA	28	8	1	Autosomal recessive sensorineural hearing loss	pediatric	1/1000	3
IKBKG	10	8	1	Incontinentia Pigmenti	pediatric	unknown	3
ABCC6	31	9	2	Pseudoxanthoma elasticum	variable	1/25,000 to 1/100,000	3

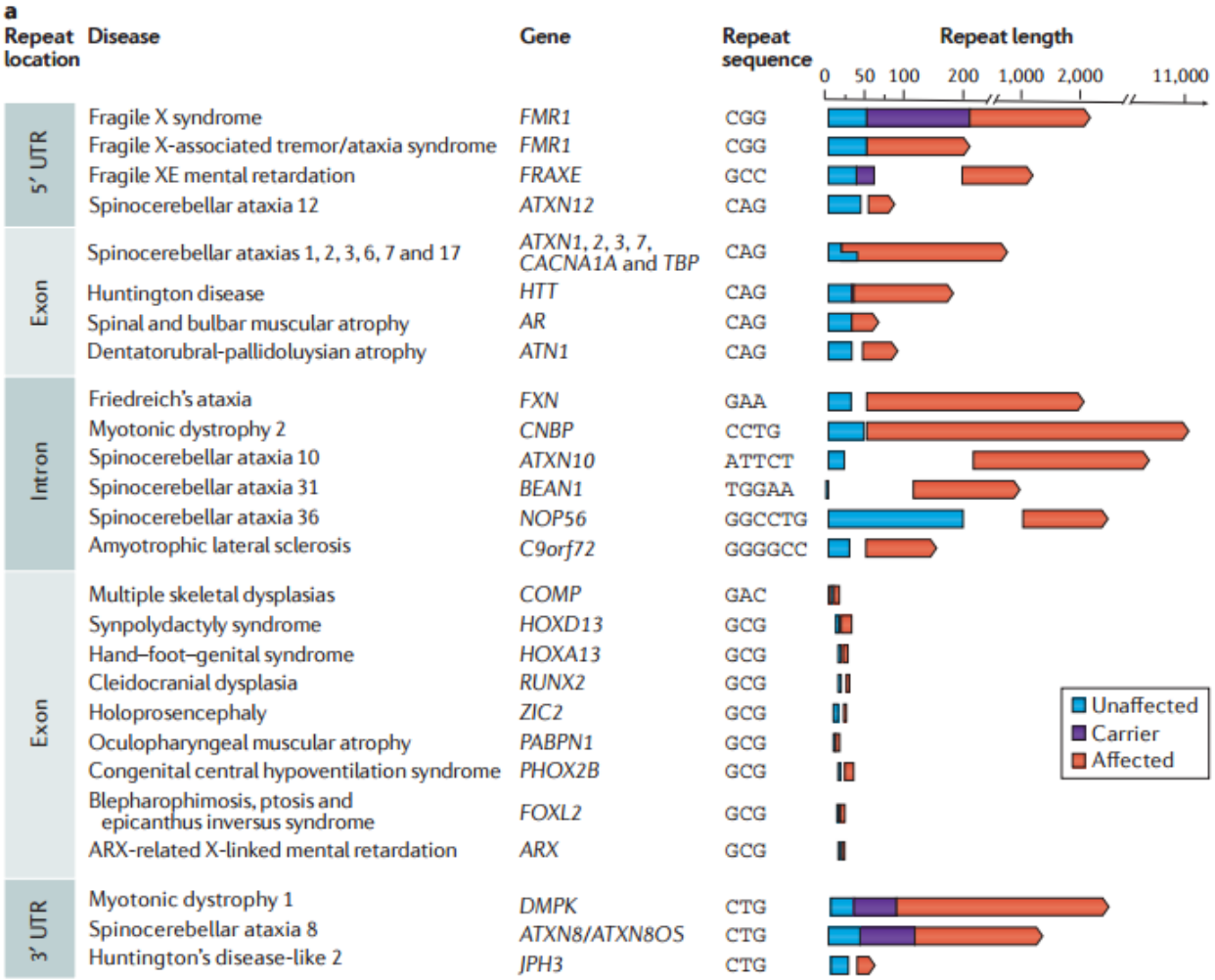
Mandelkar et al. Genetics in medicine 2016

Short tandem repeats (STRs)



CAG.CAG.CAG…….CAG (20-35 copies)

STR expansions cause many diseases



Ashley NRG 2017

2. Finding causal variant or gene

- Finding disease causing variants is trying to find a ‘needle in a haystack’
- large number of candidate variants
- False positives

1. **Linkage mapping:** identifying region(s) of the genome that harbor disease associated variants
2. **Gene function**: which genes are functionally relevant to the disease phenotype
3. **Variant impact**: which variants are likely to be deleterious to protein function, gene expression, etc

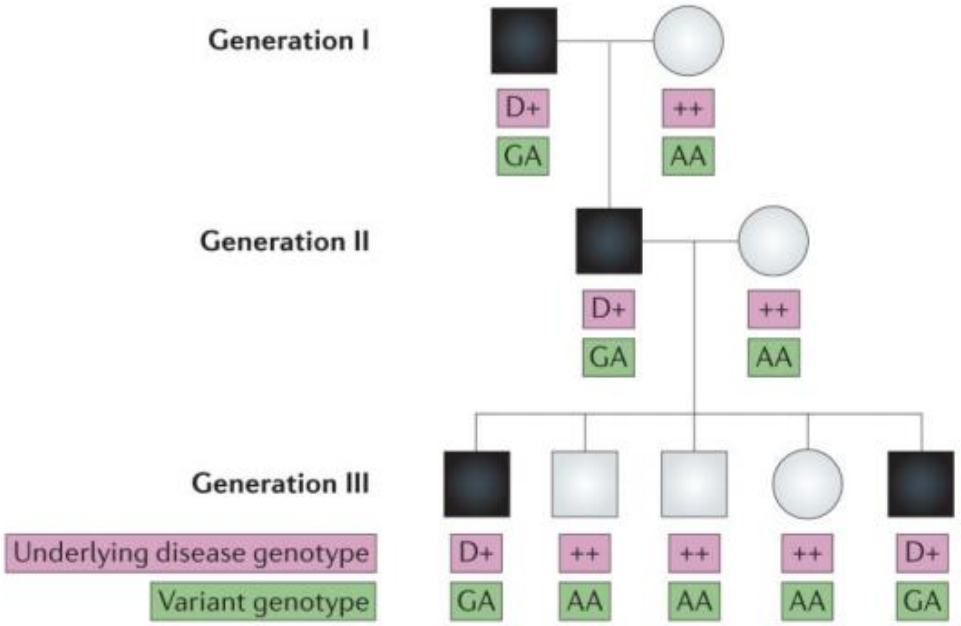
Different approaches for finding disease causing mutations (rare disease)

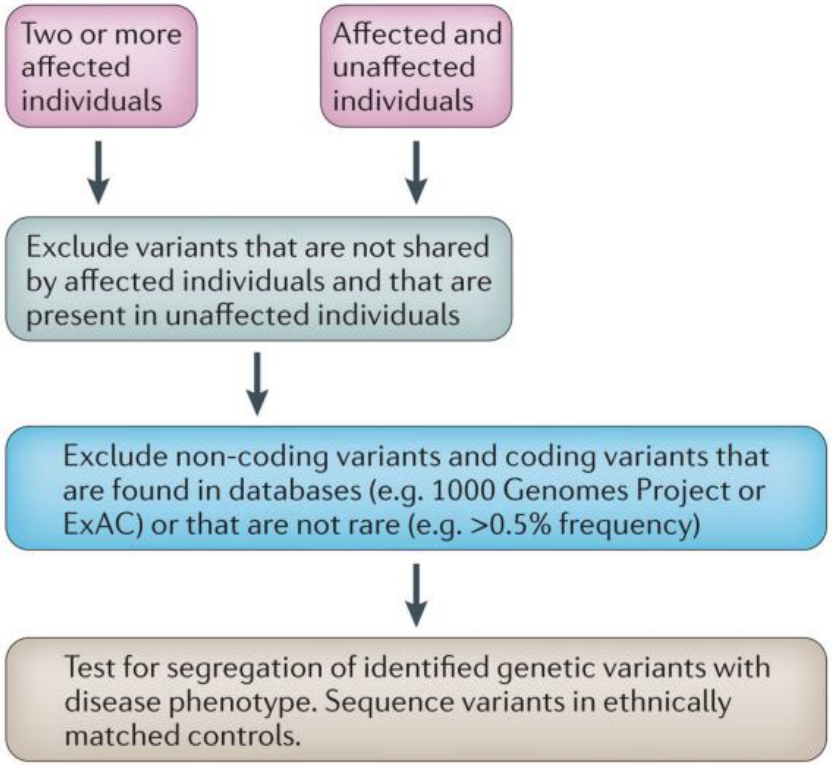
1. Multiple related individuals (family)
2. Multiple unrelated individuals

3. Single individual: prioritization using population data
4. Single individual: integrating DNA-seq and RNA-seq data

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

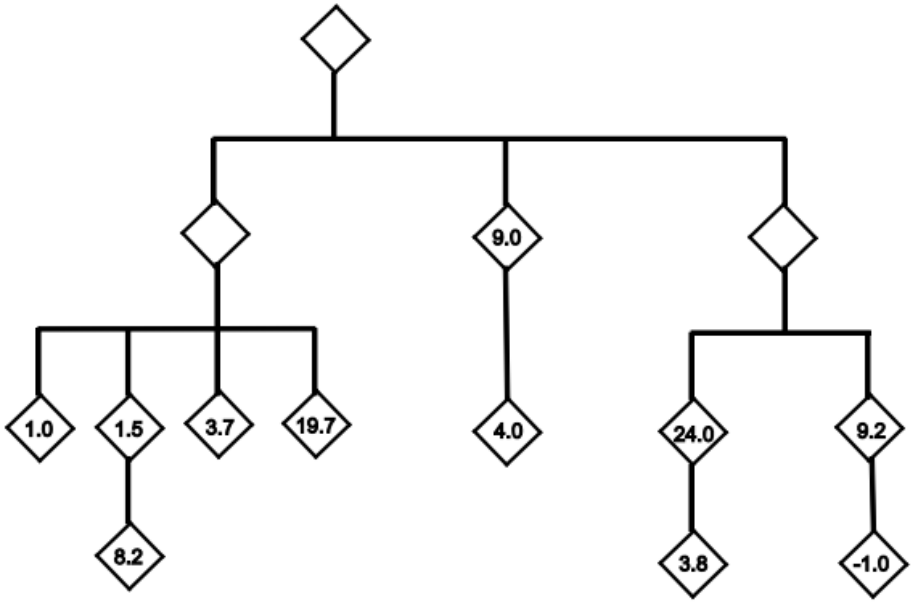
1. Family data





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

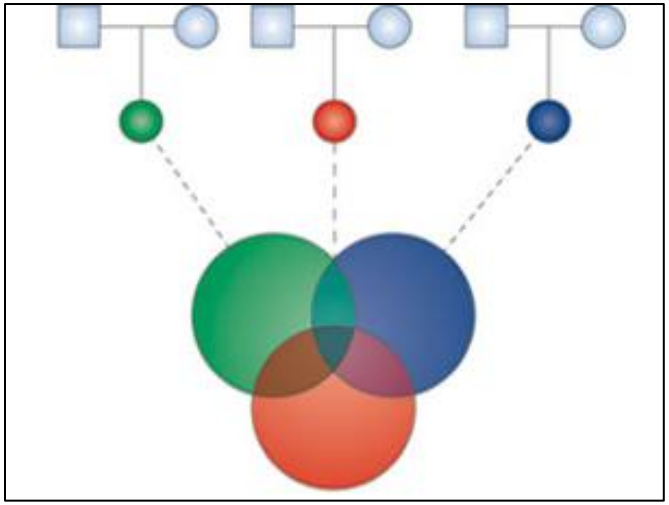
Variants under linkage peaks

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 co-segregates with high TG but unlikely to be causal

2. Multiple unrelated individuals



- Different (or identical) 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

2. Multiple unrelated individuals

unrelated individuals
with same phenotype

sequence DNA

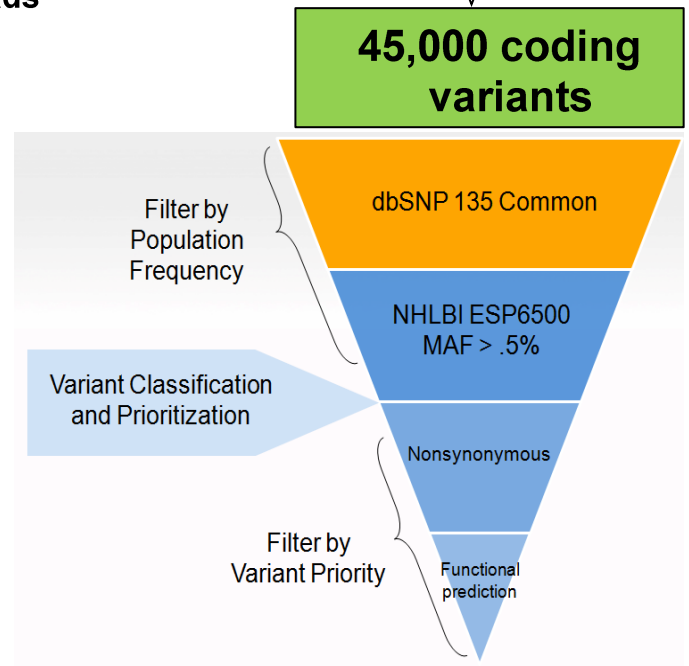
CCAGTCGATCAGGCA TGATCGAAC
GTGTAGCACCCAAAGTA
TGCGATCGATA CGTGTAGTAGTT
GTGTAGCACGACTAGCA
GTGCACACCATTTTCAG
GTGACGACACTCGATCGA

millions of short reads

align to reference

human genome

identify variants

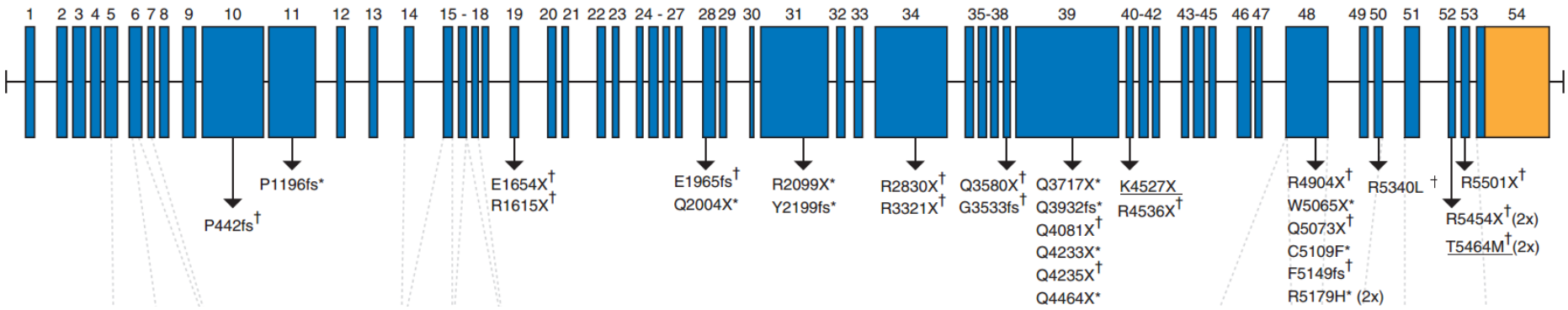


~100 candidate genes per individual

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

Number of genes common to any subset of *x* affected individuals

The number of genes with at least one non-synonymous variant (NS), splice-site acceptor/donor variants (SS) or coding indel (I) are listed under various filters. Variants were filtered by presence in dbSNP or 1000 genomes ("Not in dbSNP129 or 1000 genomes") and control exomes ("Not in control exomes") or both ("Not in either"); control exomes refer to those from 8 Hapmap3 4 FSS3, 4 Miller2 and 10 EGP samples. The number of genes found using the union of the intersection of x individuals is given.

a. Subset analysis (any *x* of 10)	1	2	3	4	5	6	7	8	9	10
NS/SS/I	12,042	8,722	7,084	6,049	5,289	4,581	3,940	3,244	2,486	1,459
Not in dbSNP129 or 1000 genomes	7,419	2697	1057	488	288	192	128	88	60	34
Not in control exomes	7,827	2865	1025	399	184	90	50	22	7	2
Not in either	6,935	2227	701	242	104	44	16	6	3	1
Is loss-of-function (nonsense/frameshift indel)	753	49	7	3	2	2	1	0	0	0

Ng. et al. Nat. Genetics 2010

Multiple related individuals (family)

Disease variant(s) shared across all individuals with disease

Multiple unrelated individuals

Different variants but in the same gene(s)

Question: which approach is more powerful ?

3. Single individual: 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 depleted in the normal population

Model

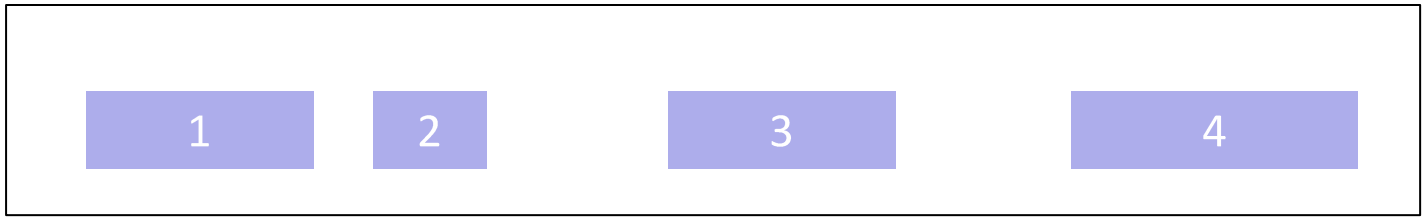
For any gene G

- 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



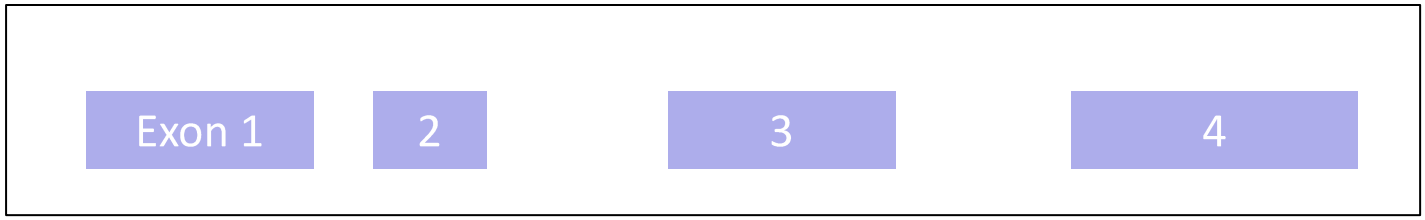
Exons

Tri-nucleotide mutation rates

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

Mutation type	sites	total probability
silent	867	2.1×10^{-7}
missense	1784	5.3×10^{-7}
stop-gain	123	4.2×10^{-8}

Expected vs observed mutation counts

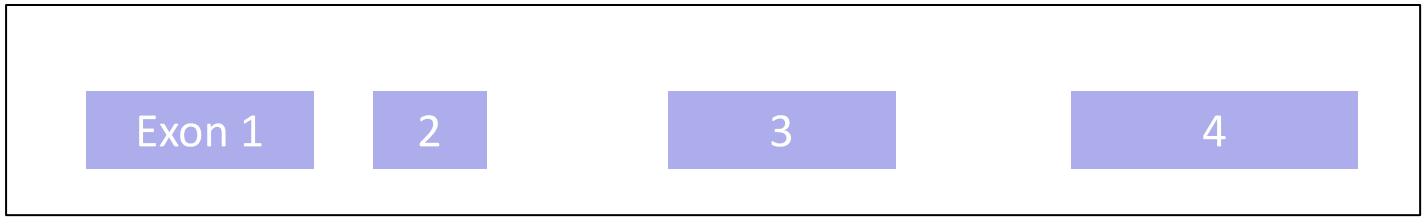


Tri-nucleotide mutation rates

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

Mutation type	sites	total probability	observed mutations
silent	867	2.1×10^{-7}	64
missense	1784	4.6×10^{-7}	131
stop-gain	**123**	**4.2×10^{-8}**	**2**

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}

Mutation type	sites	total probability	observed mutations
silent	867	2.1×10^{-7}	64
missense	1784	4.6×10^{-7}	131
stop-gain	**123**	**4.2×10^{-8}**	**2**

Expected/Observed = 6.4 for stop-gain

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

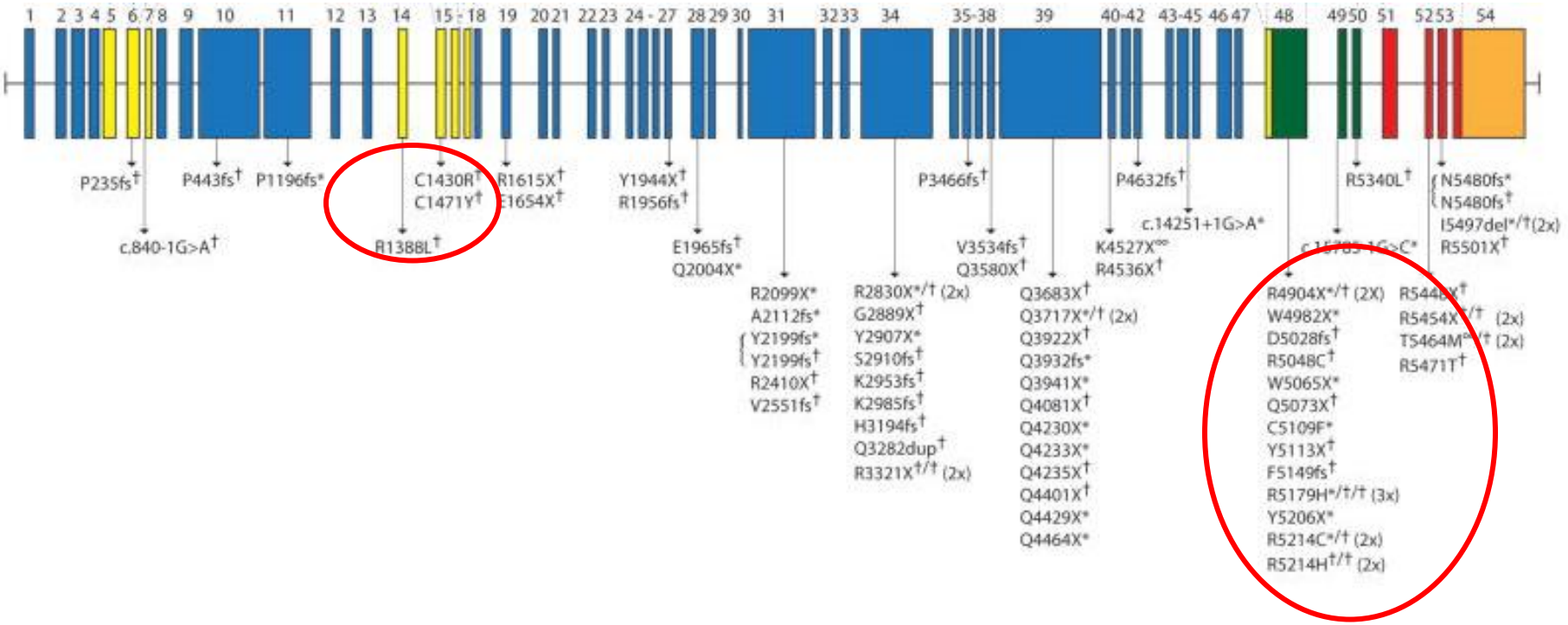
Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Synonymous	792.9	919	z = -2.78
Missense	1842.9	1571	z = 3.10
LoF	137.6	11	pLI = 1.00

Exome data from 65,000 individuals

- More than 2600 genes have high (> 0.95) LoF-intolerance
- Doesn't imply causality but useful for prioritization
- If mutation is 'de novo', more likely to be pathogenic

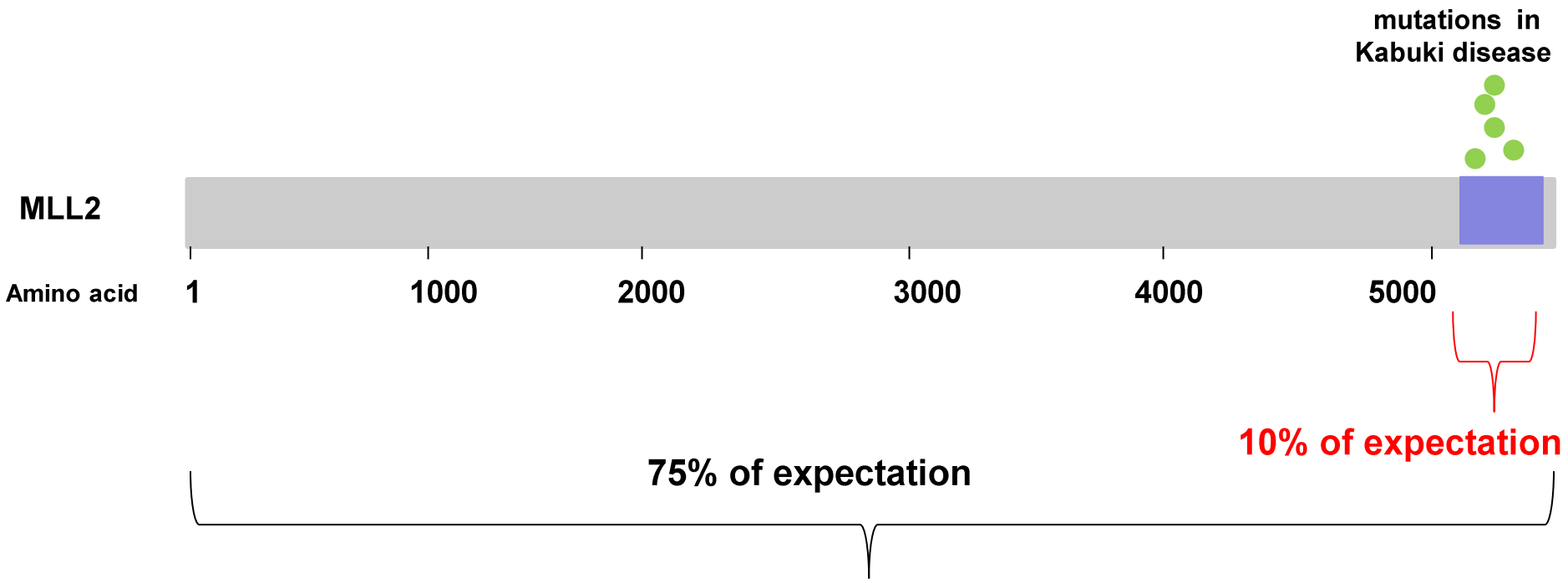
http://exac.broadinstitute.org/gene/ENSG00000167548

Missense mutations in MLL2



- Missense mutations in some exons cause Kabuki syndrome (prevalence = 1/32000 births)
- 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

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

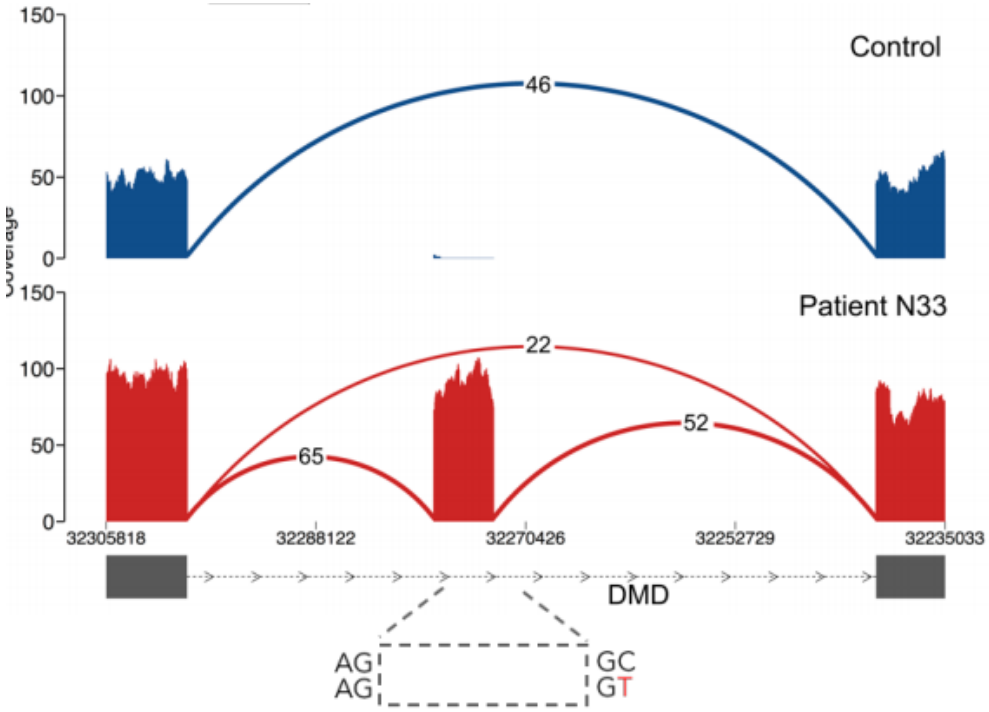
Ranking human genes based on their tolerance to functional genetic variation can greatly facilitate patient genome interpretation. It is well established, however, that different parts of proteins can have different functions, suggesting that it will ultimately be more informative to focus attention 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 divi-

4. Integrating DNA-seq and RNA-seq from 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

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

affected individuals and related individuals

sequencing

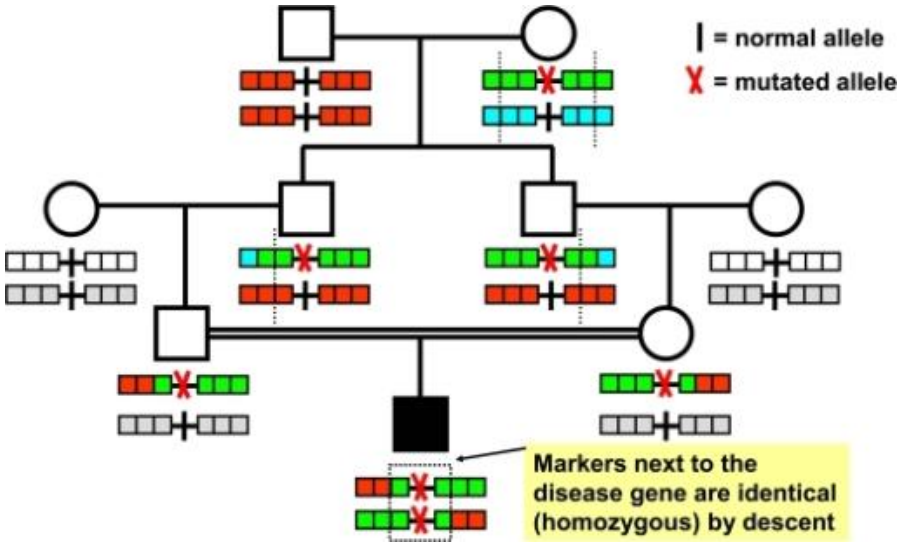
Variant calling

Homozygosity mapping

Variants shared by affected Individuals in HBD regions

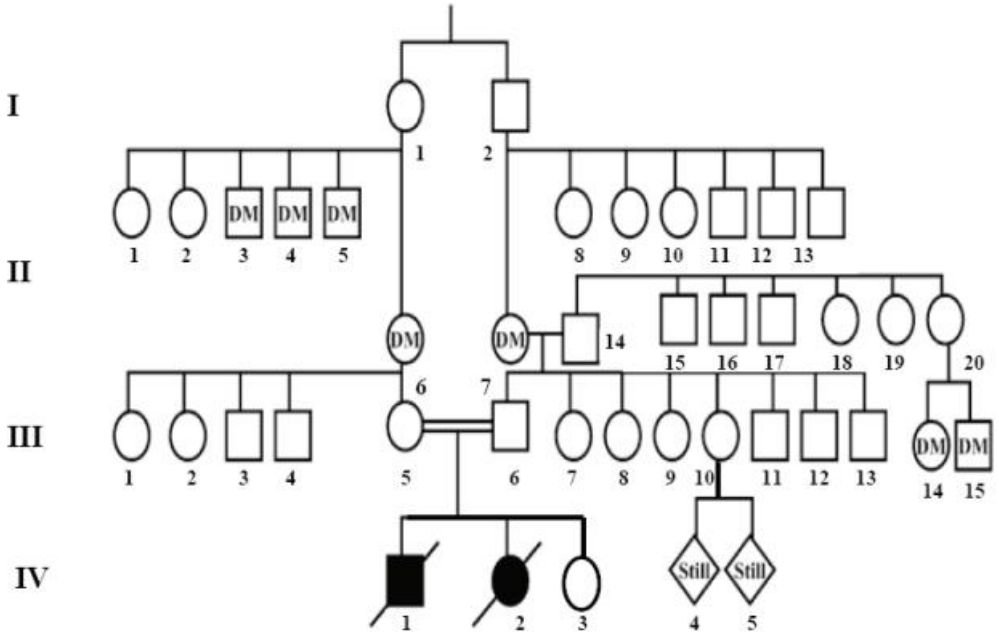
Retain variants in genes expressed in disease relevant tissues

Validation by phenotyping of mouse knockout



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

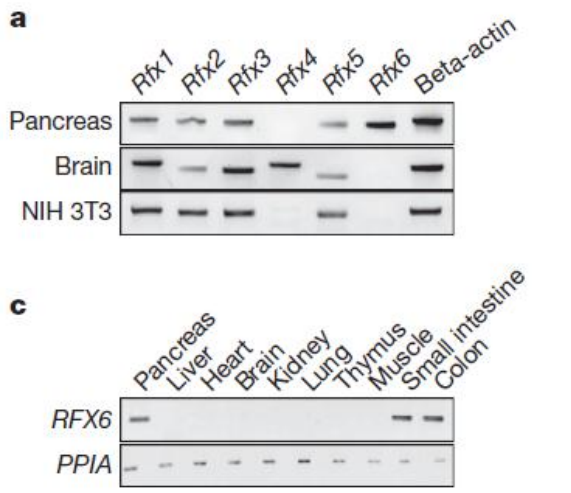


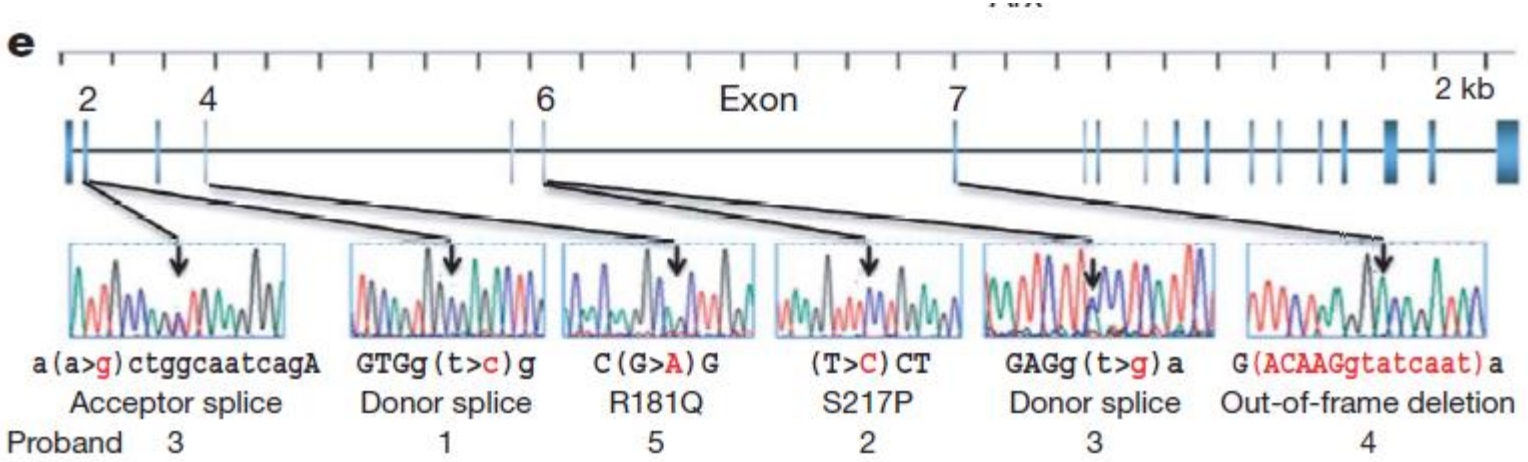
Figure 1 | Expression of *Rfx6* in mice and human

Mouse knockout phenotype

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

Smith et al. Nature 2010

Additional mutations confirm RFX6 as the causal gene



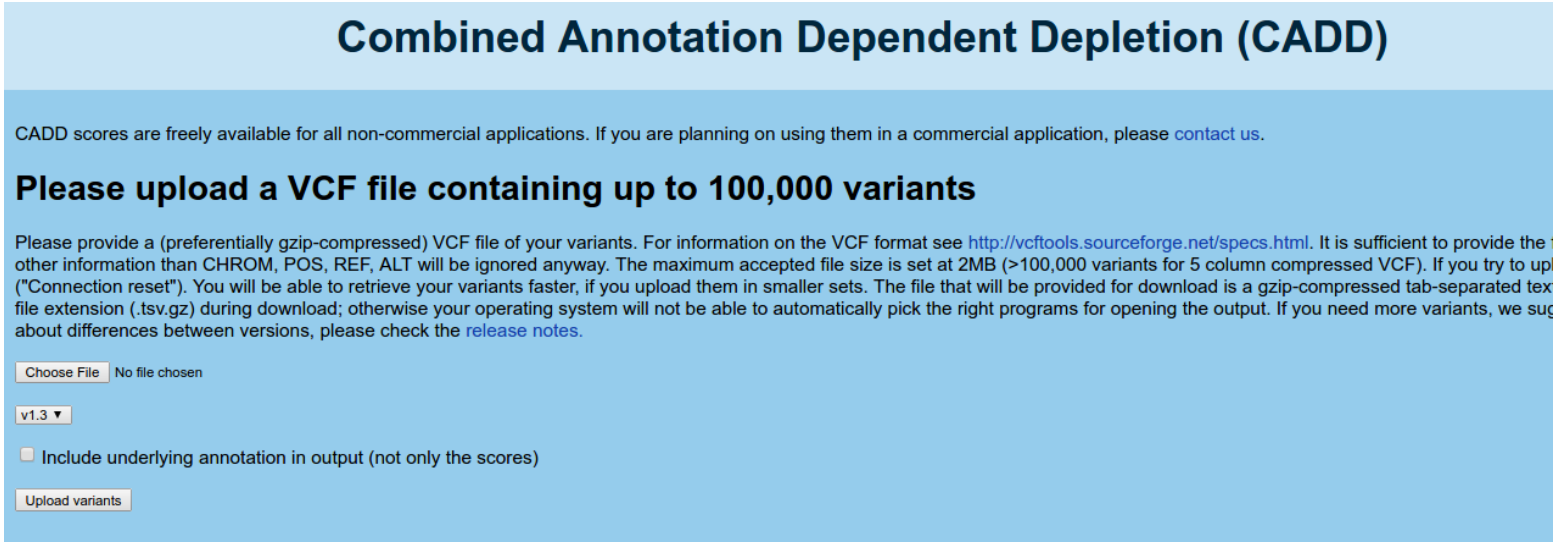
- Gene could also have been identified by whole-exome sequencing of the six “unrelated individuals” with Mitchell-Riley syndrome

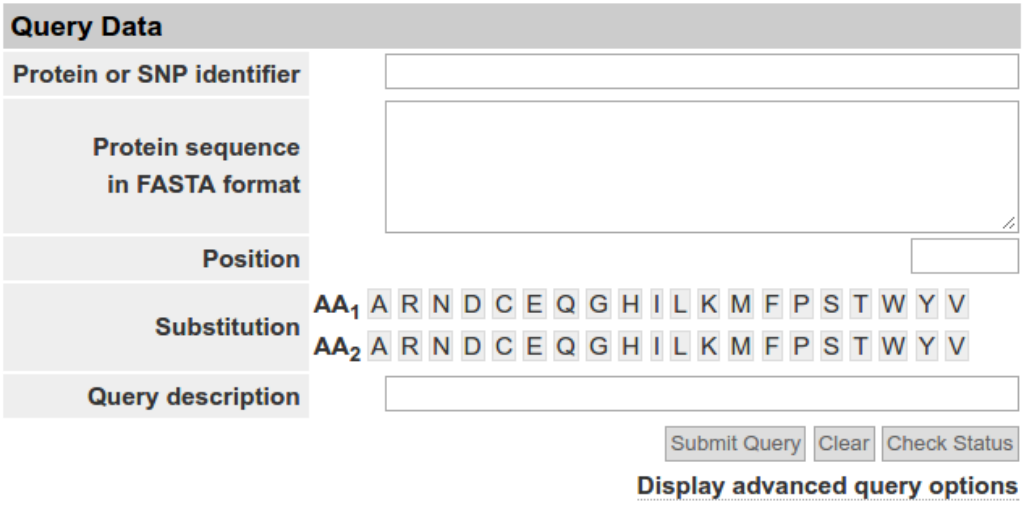
Smith et al. Nature 2010

Prioritizing variants and genes

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

Variant annotation

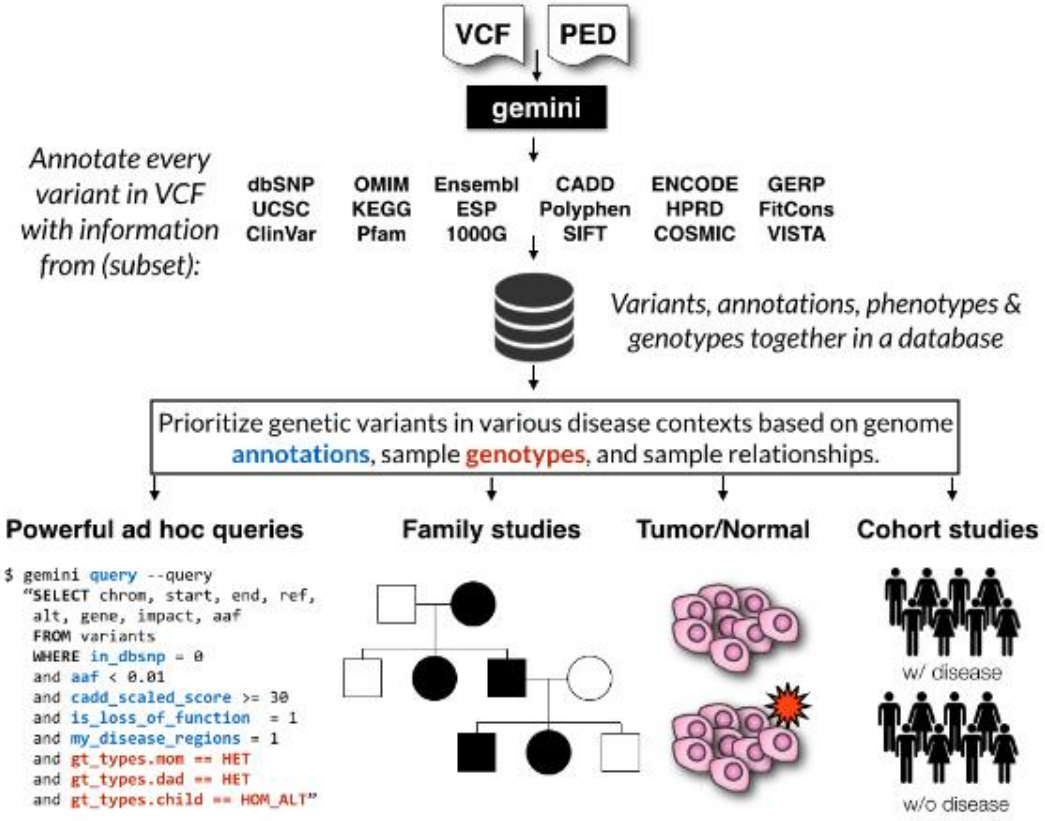




http://genetics.bwh.harvard.edu/pph2/

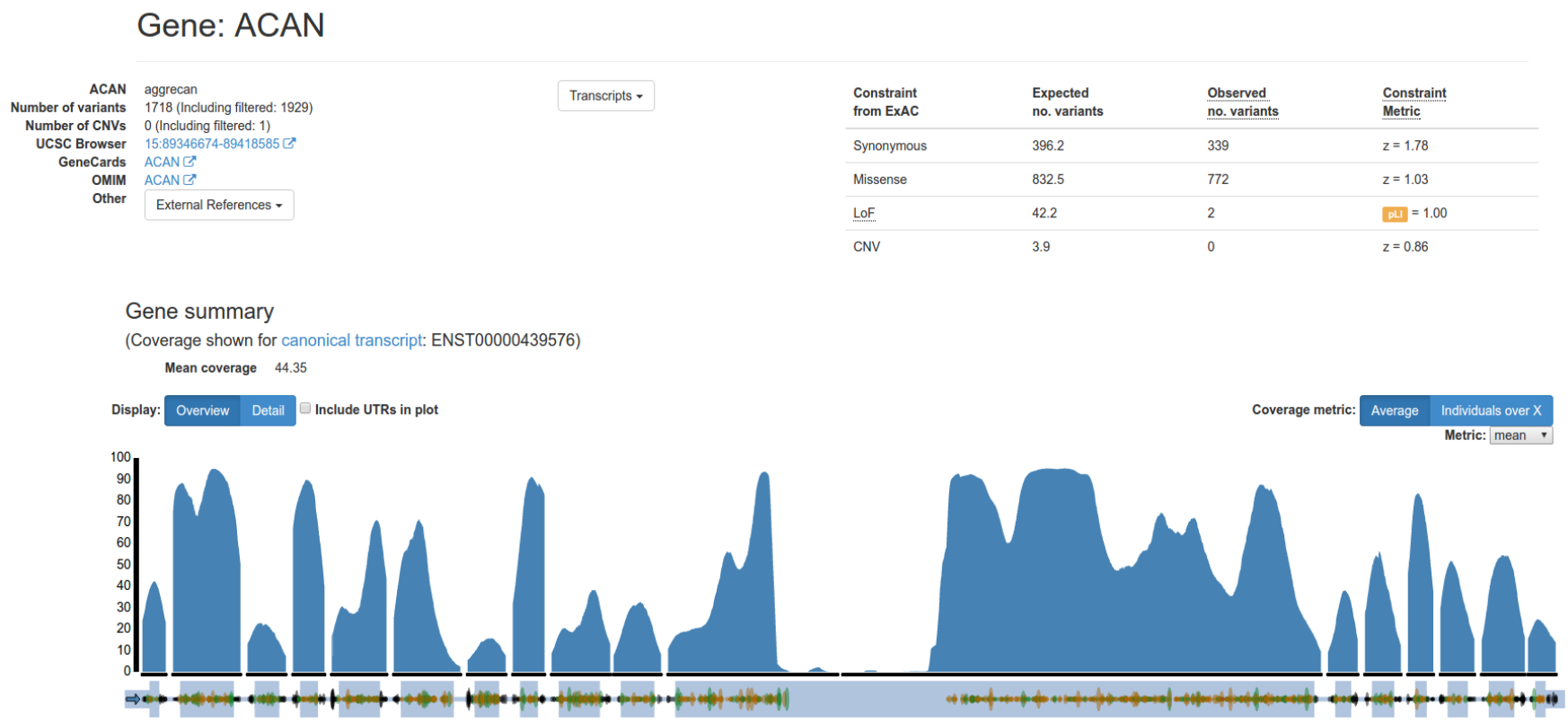
http://annovar.openbioinformatics.org/en/latest/

GEMINI: *a flexible framework for exploring genome variation*



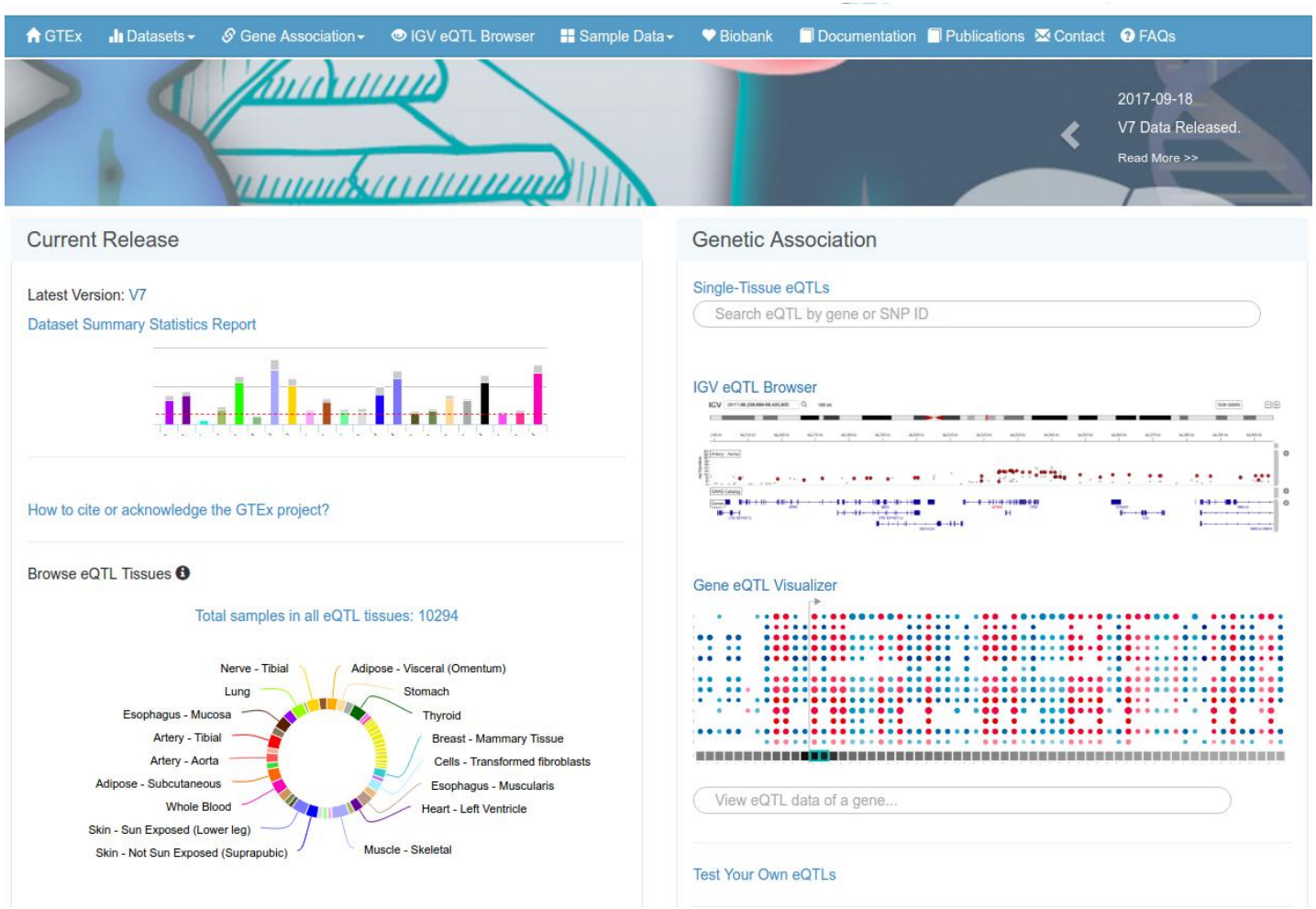
https://gemini.readthedocs.io/en/latest/

ExAC variant browser



http://exac.broadinstitute.org/gene/ENSG00000157766

GTEX (gene expression) portal



https://gtexportal.org/home/

Summary

- Comprehensive variant detection is important for finding disease causing mutations
  - indels, SV and haplotypes are challenging

- Variant interpretation is context-dependent and discovery of disease associated mutations requires integration of different data-types
- Bioinformatics is key

Questions ?