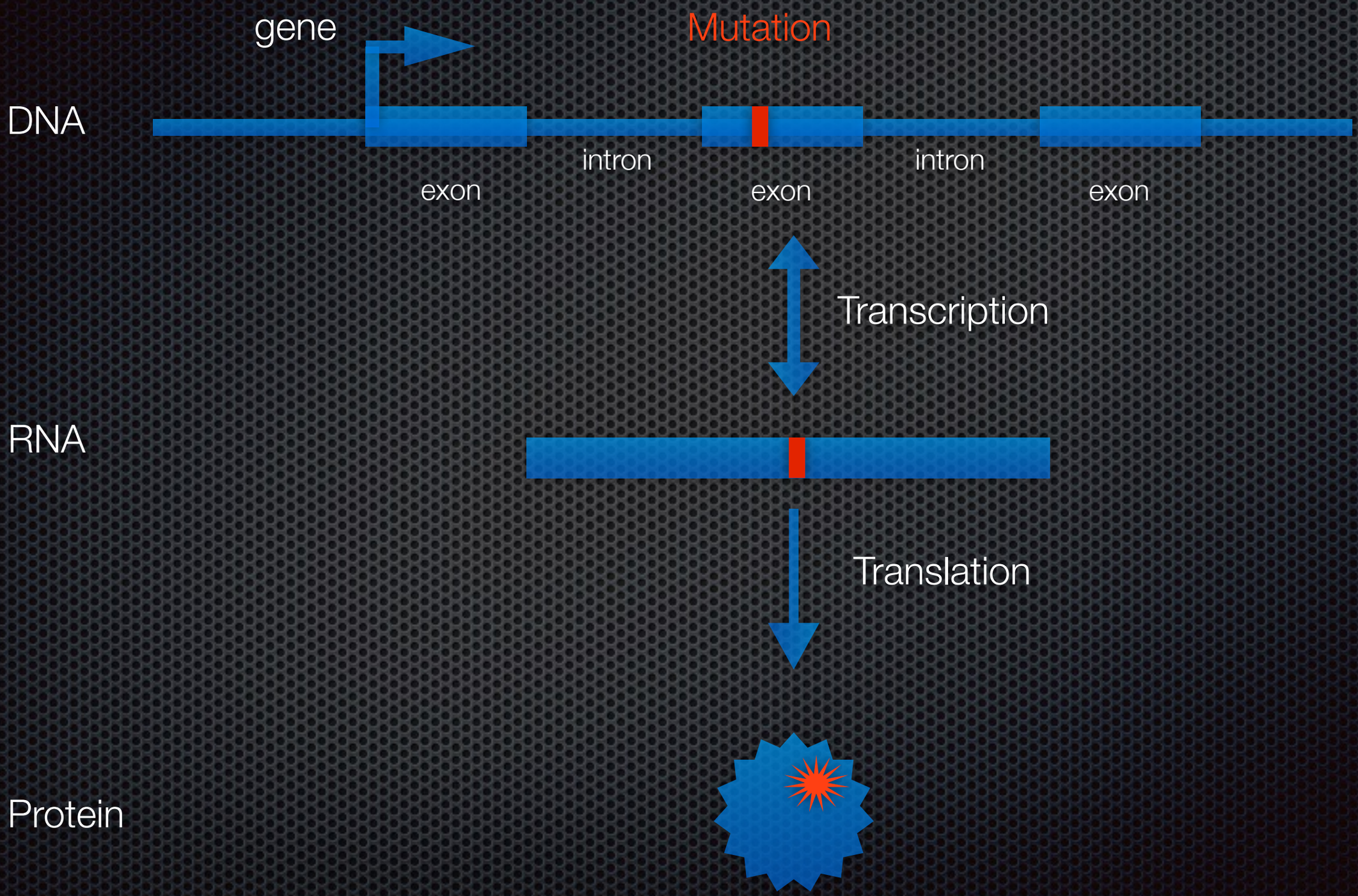
HTS and medical genetics

Finding mutations which cause disease

DNA - RNA - protein



Aim

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;;9;7;;.7;39333
```

FASTQ format

R|G

Compare to reference

3 billion reads

3 billion bases

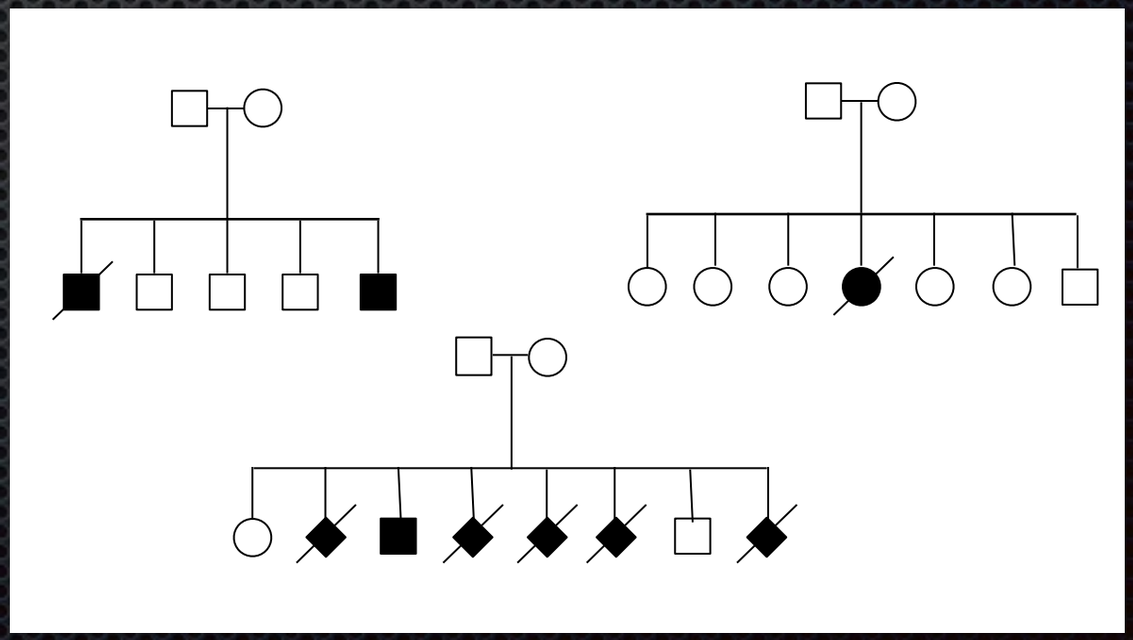
1 mutation

Genetic disease: the challenge

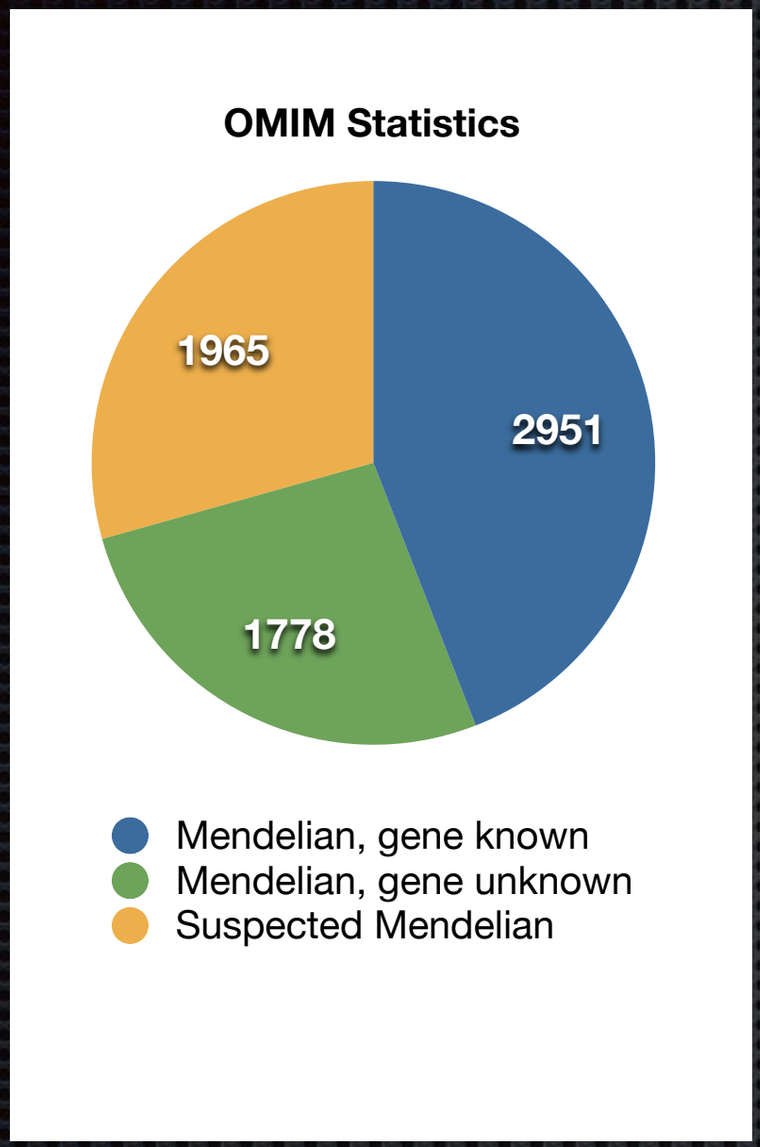
- Single gene (monogenic) disorders
- Approximately 50% of children with congenital syndromes/mental retardation do not receive a firm etiological diagnosis
- Many of these children have disorders that are primarily genetic in origin.
- Many disorders are individually rare and difficult to recognize even for experts.
- Clinical phenotypes can be non-specific and variable.
- Many phenotypes are genetically heterogeneous.

- Human genome is (quite) big
  - ~23 000 genes
  - ~3 billion bases pairs

- How can we (rapidly) identify a mutation causing disease?



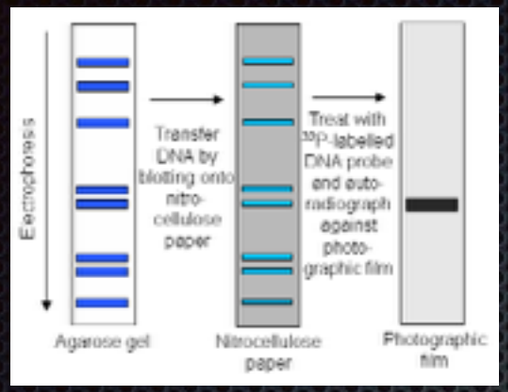
Mendelian disease in man



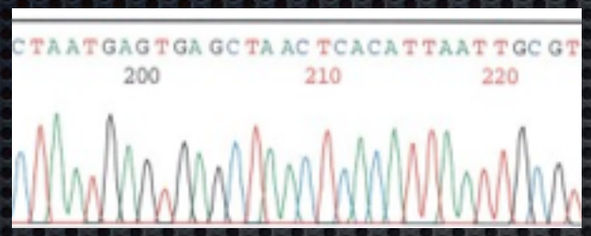
- 2951 of the well-characterized phenotypes registered in OMIM have a known molecular basis
- 3743 registered phenotypes with known or suspected Mendelian basis, no associated gene has been identified

- ***Improve speed/cost of diagnosis of known genetic disorders***
- ***Improve speed/cost of identification of the cause new/suspected genetic disorders***

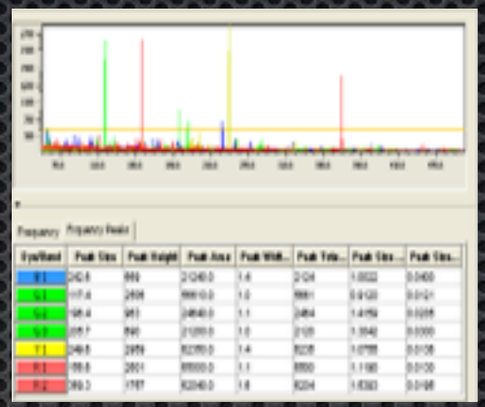
Methods for identifying variants/aberrations



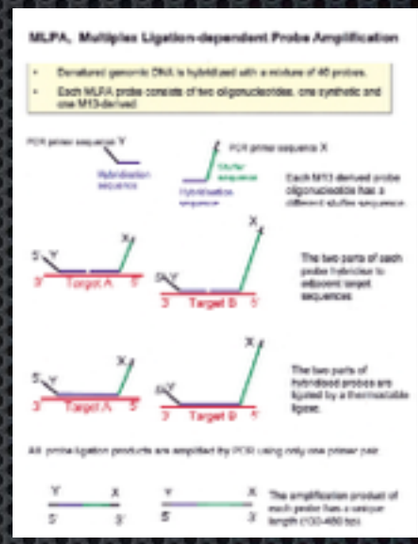
Southern blotting



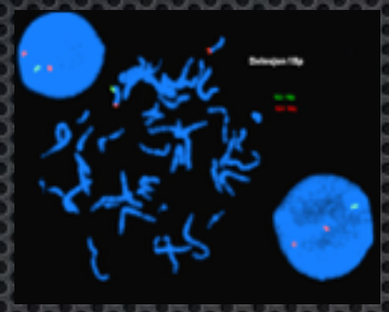
DNA (Sanger) sequencing



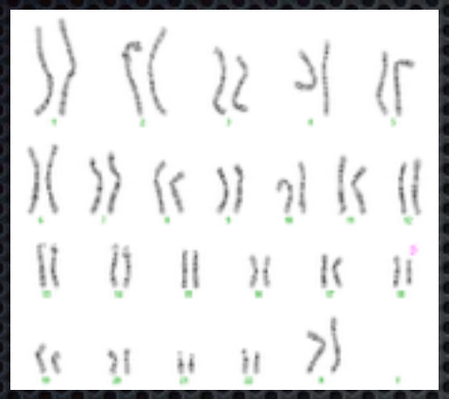
Fragment analysis



MLPA

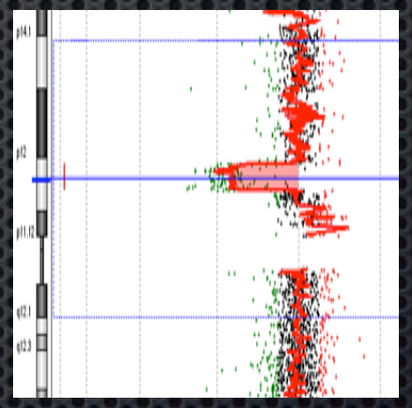


FISH



Karyotyping

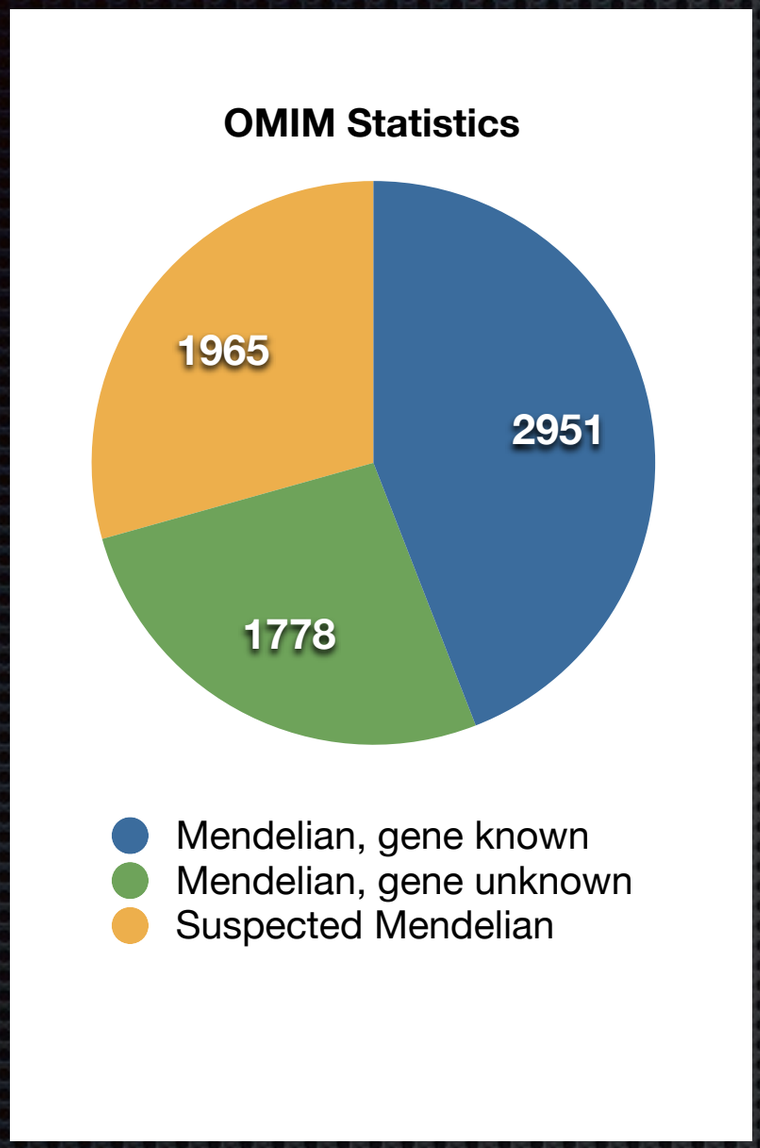
SNP array



Array CGH

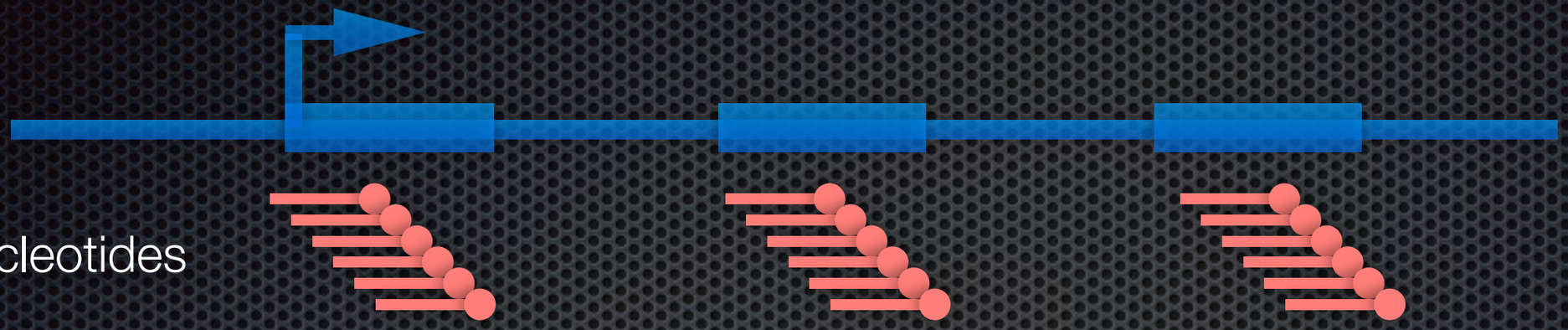
- Low throughput (limited number of loci per run)
- Detect specific types of variation

Mendelian disease in man

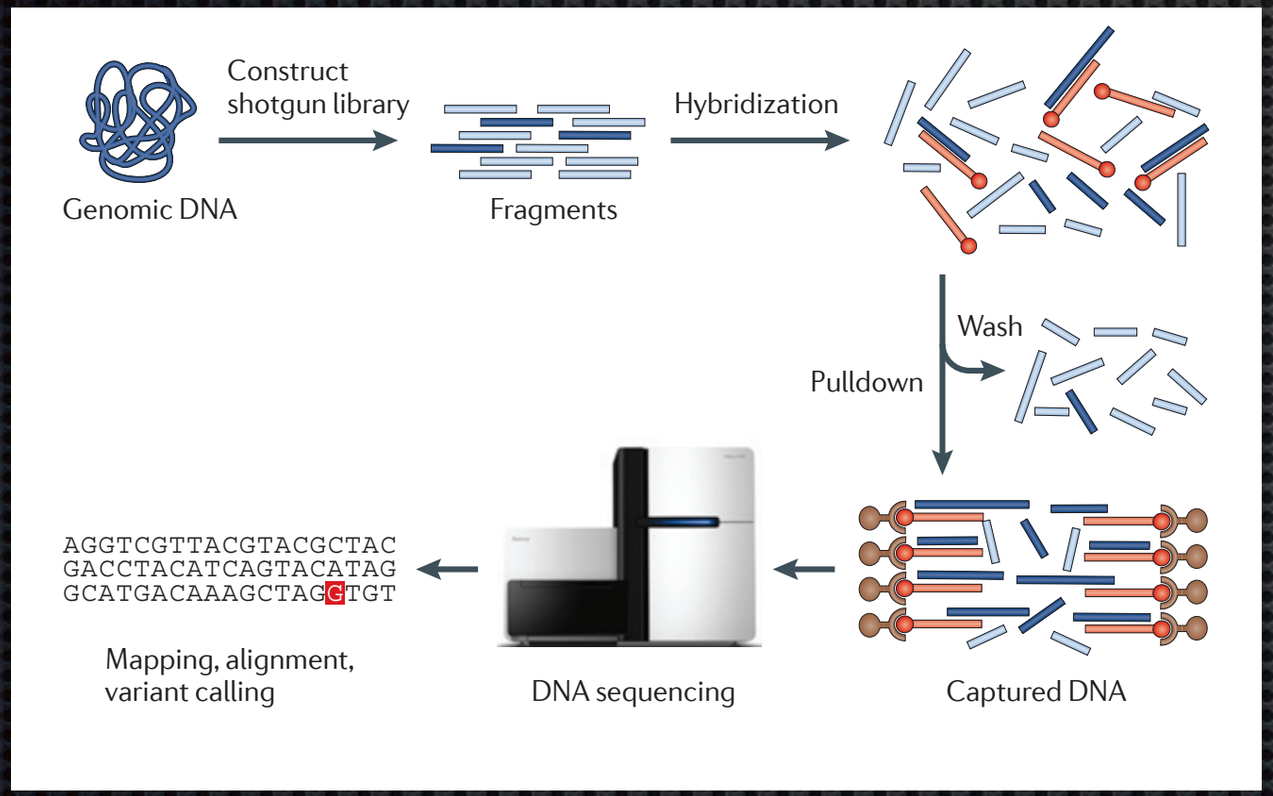


- 2951 of the well-characterized phenotypes registered in OMIM have a known molecular basis
- 3743 registered phenotypes with known or suspected Mendelian basis, no associated gene has been identified
- Protein coding regions of the human genome (the exome) constitute approximately 1.5% of the total, but harbour ~85% of the mutations with large effects on disease-related traits
- ***Exome sequencing***
- ***HTS in research and routine diagnostics?***

Exome sequencing



oligonucleotides



- Design exome capture array
- Make sequence library from patient DNA
- Hybridize and capture
- Sequence

Aim

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;;9;7;;.7;39333
```

FASTQ format

R|G

Compare to reference

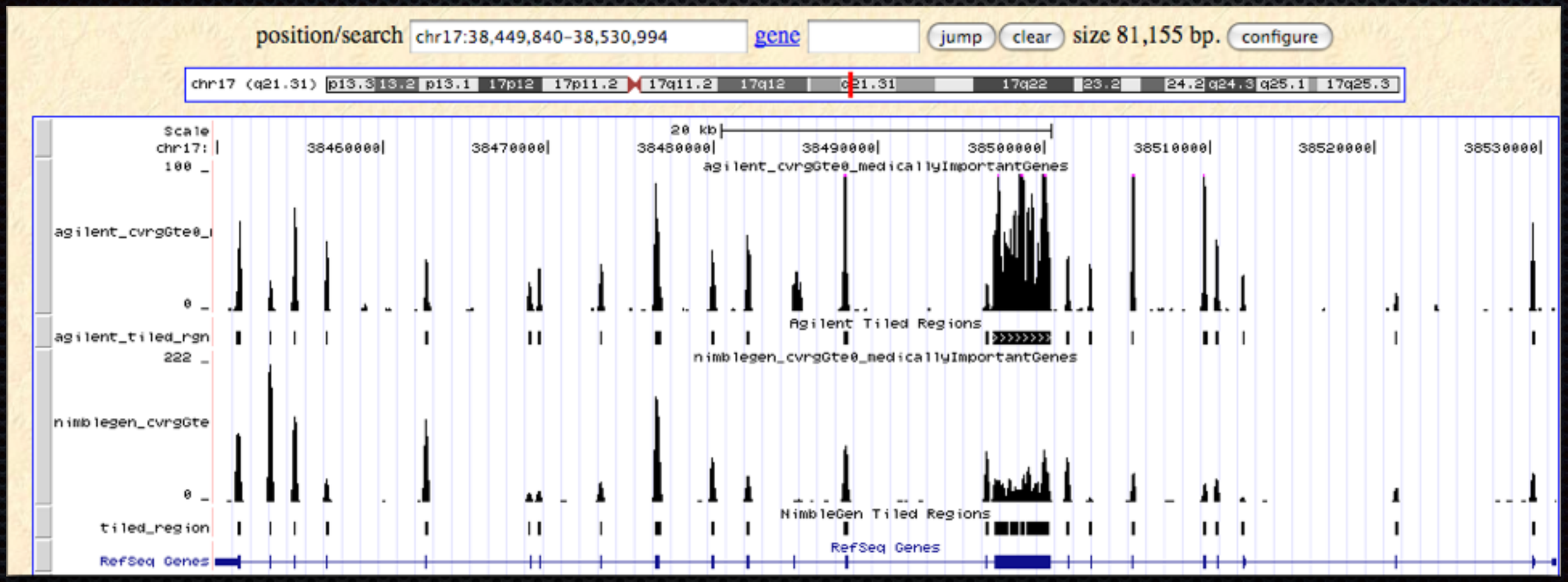
3 billion reads

3 billion bases

1 mutation

Aligning sequence

Exome sequencing data

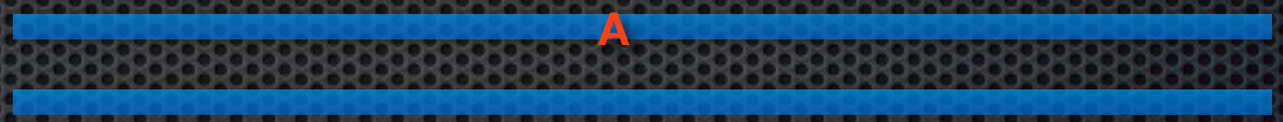


Types of genetic variation

SNP/SNV

Homozygous

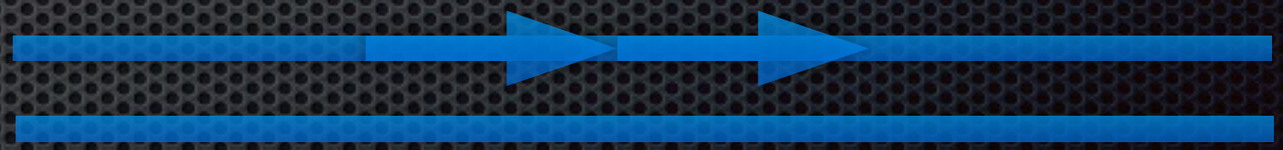
Heterozygous



CNV/indel

Deletion
(hemizygous)

Duplication



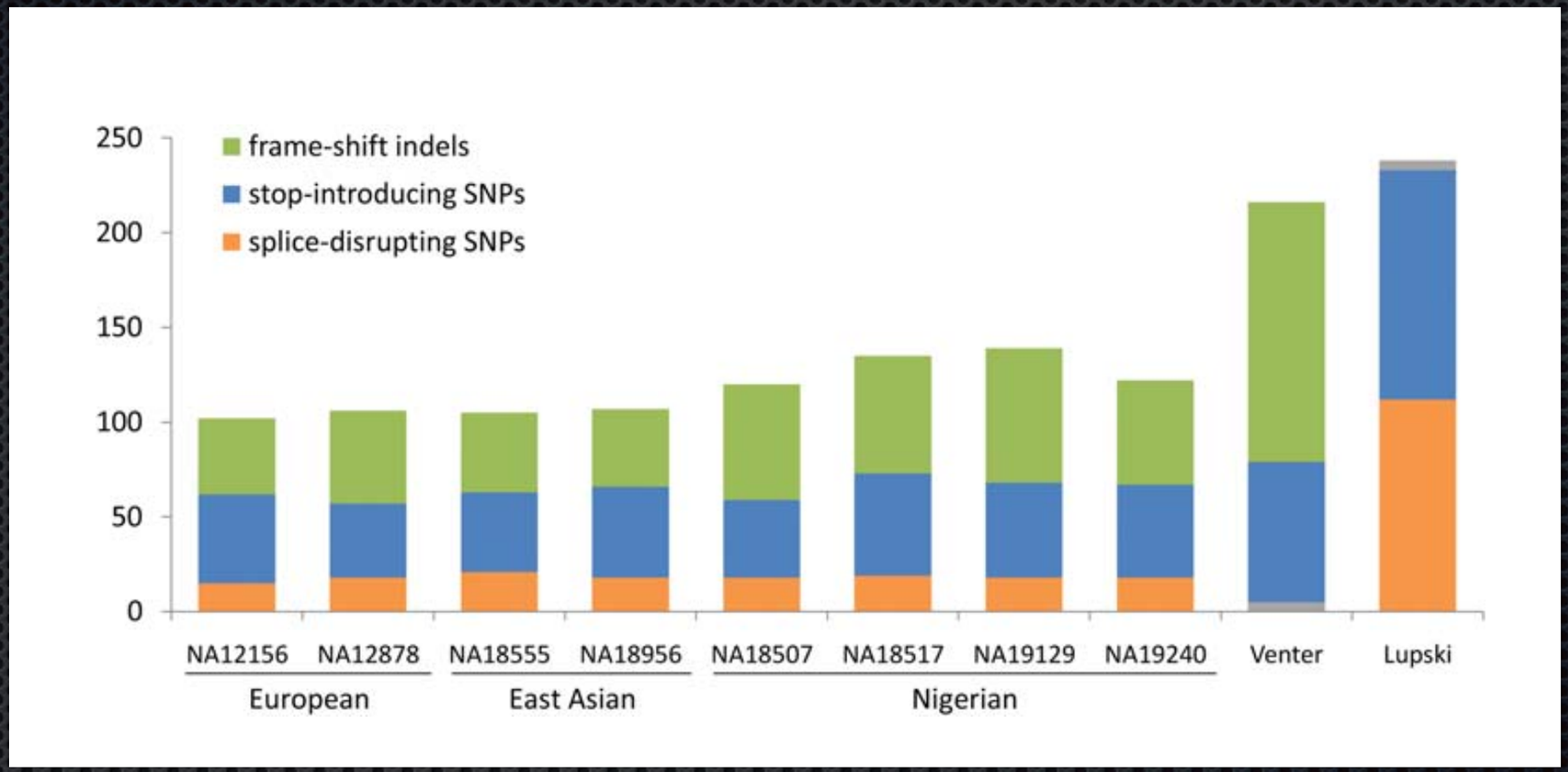
SNP/SNV: single nucleotide polymorphism/variant
CNV: copy number variant
indel: insertion/deletion

Effect of variants

- missense - change amino acid
- nonsense - premature stop codon
- frameshift - shift the codon frame
- Indels - multiple effects

What's in an exome?

> 20 000 variants



~100 loss-of-function variants

MacArthur & Tyler-Smith Hum Mol Genet 19 (2010).

Finding causitive mutations is not easy

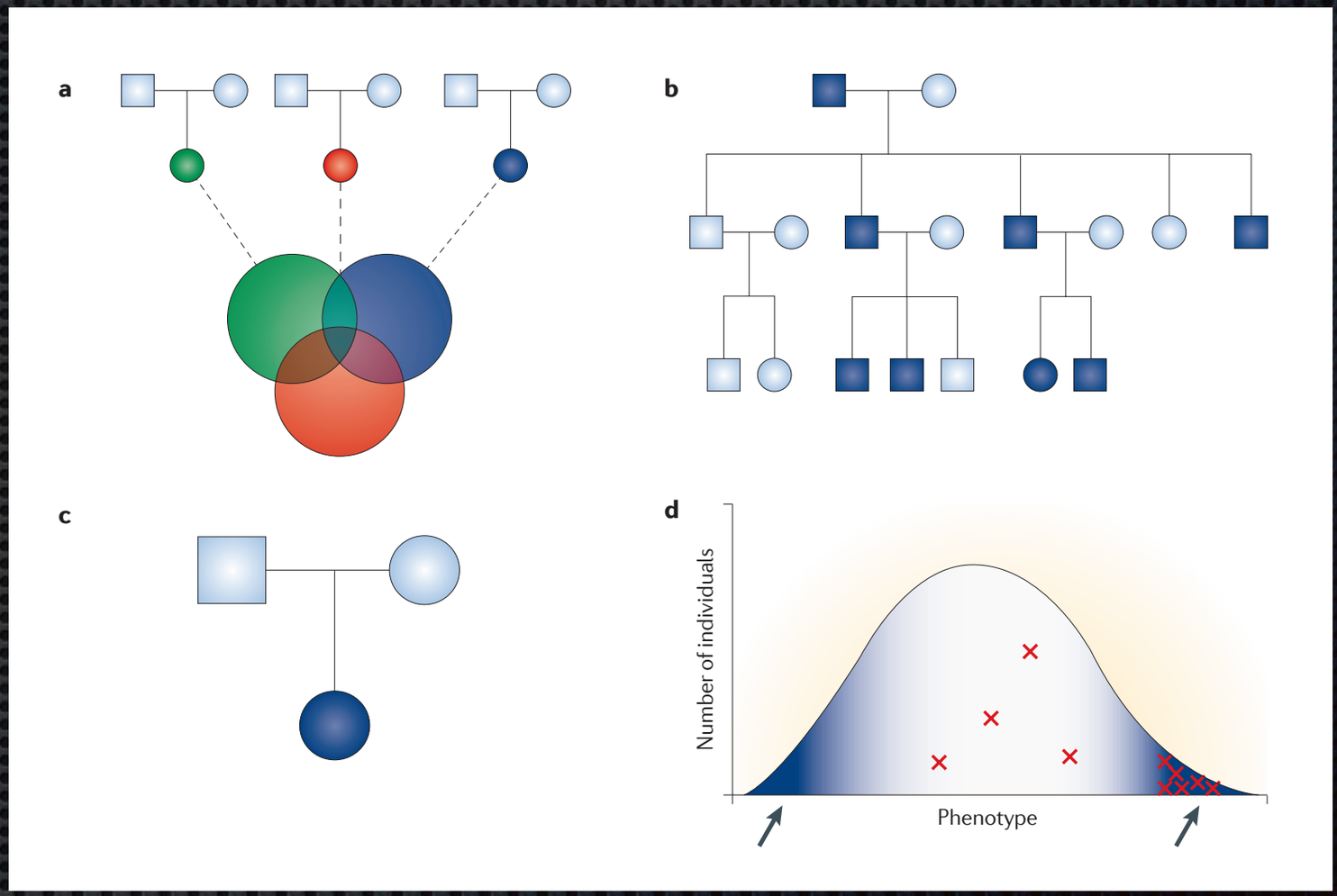
- **Many, many variants will be found**
- Which variants are deleterious?
- Novel? (dbSNP, 1000genomes, HGMD)
- Synonymous/non-synonymous?
- Conserved?
- Alter protein structure?

SNPnexus
PolyPhen2
MutationTaster
ANNOVAR
SeattleSeq Annotation

This is the hard part

Strategies to identify mutations

- multiple individuals same disease
- large multigenerational pedigrees
- de novo mutations
- population frequency for complex diseases



Bamshad et al. Nature Reviews Genetics Nov 2011

Family data - Shendure table

more exomes

stricter criteria

Table 3 Number of candidate genes identified based on different filtering strategies

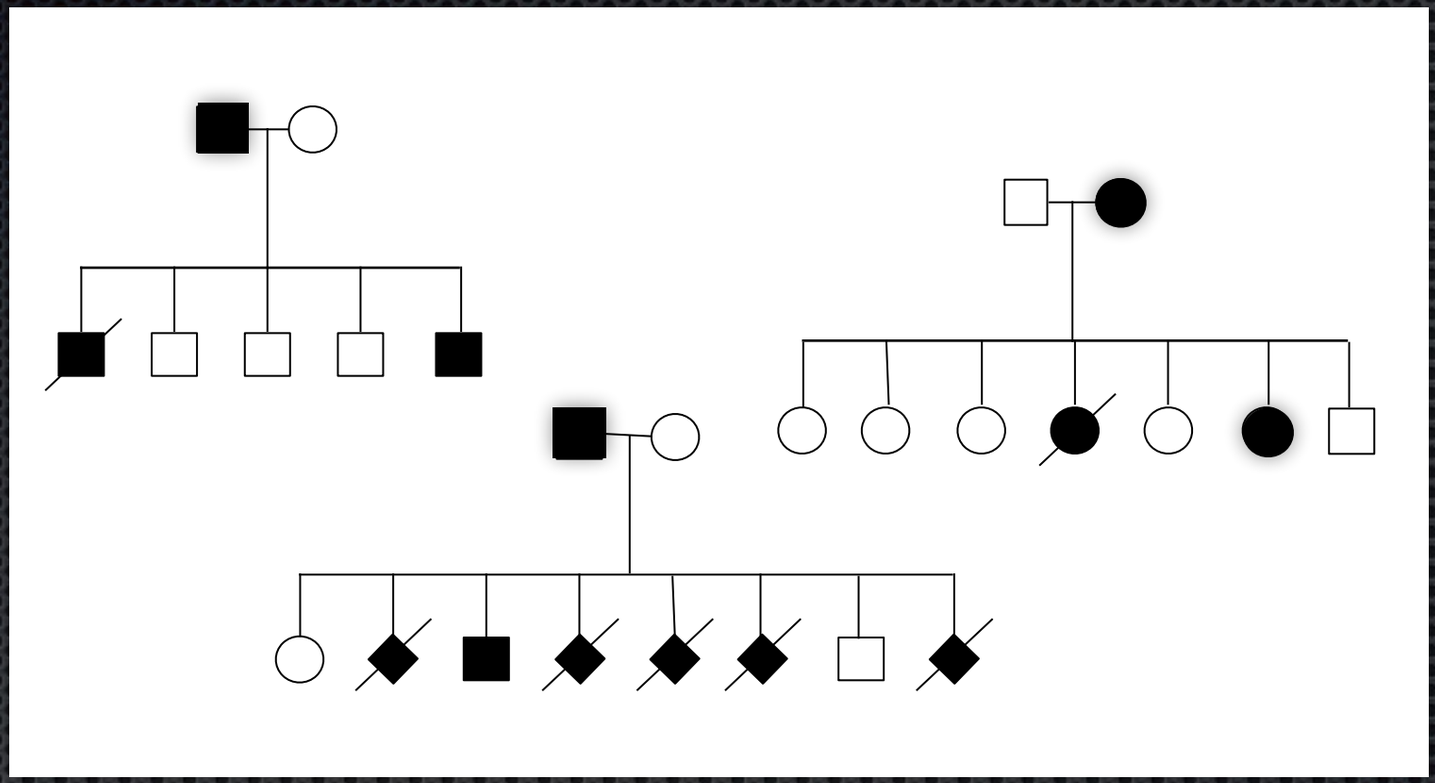
	Number of affected exomes			Subsets of 3 exomes		Subsets of all 4 exomes		
Dominant model	1	2	3	Any 1	Any 2	Any 1	Any 2	Any 3
NS/SS/I	4,645–4,687	3,358–3,940	2,850–3,099	6,658	4,489	6,943	5,167	3,920
Not in dbSNP129	634–695	136–369	72–105	1,617	274	1,829	553	172
Not in HapMap 8	898–979	161–506	55–117	2,336	409	2,628	835	222
Not in either	453–528	40–228	10–26	1,317	109	1,516	333	44
Predicted damaging	204–284	10–83	3–6	682	37	787	126	11
Recessive model								
NS/SS/I	2,780–2,863	1,993–2,362	1,646–1,810	4,097	2,713	4,293	3,172	2,329
Not in dbSNP129	92–115	30–53	22–31	226	61	270	90	42
Not in HapMap 8	111–133	13–46	5–13	329	32	397	75	19
Not in either	31–45	2–9	2–3	100	6	121	14	4
Predicted damaging	6–16	0–2	0–1	35	2	44	4	1

- Comparing two exomes identifies ~20 000 SNPs
- Which is the causal variant?
- In a family, compare more exomes

Example

Finding the mutation in ARAS syndrome

ARAS syndrome



- Born blind, due to anopthalmia/ micropthalmia
- Early onset (infant) neurodegenerative disease
- MRI shows brain atrophy
- Often fatal early

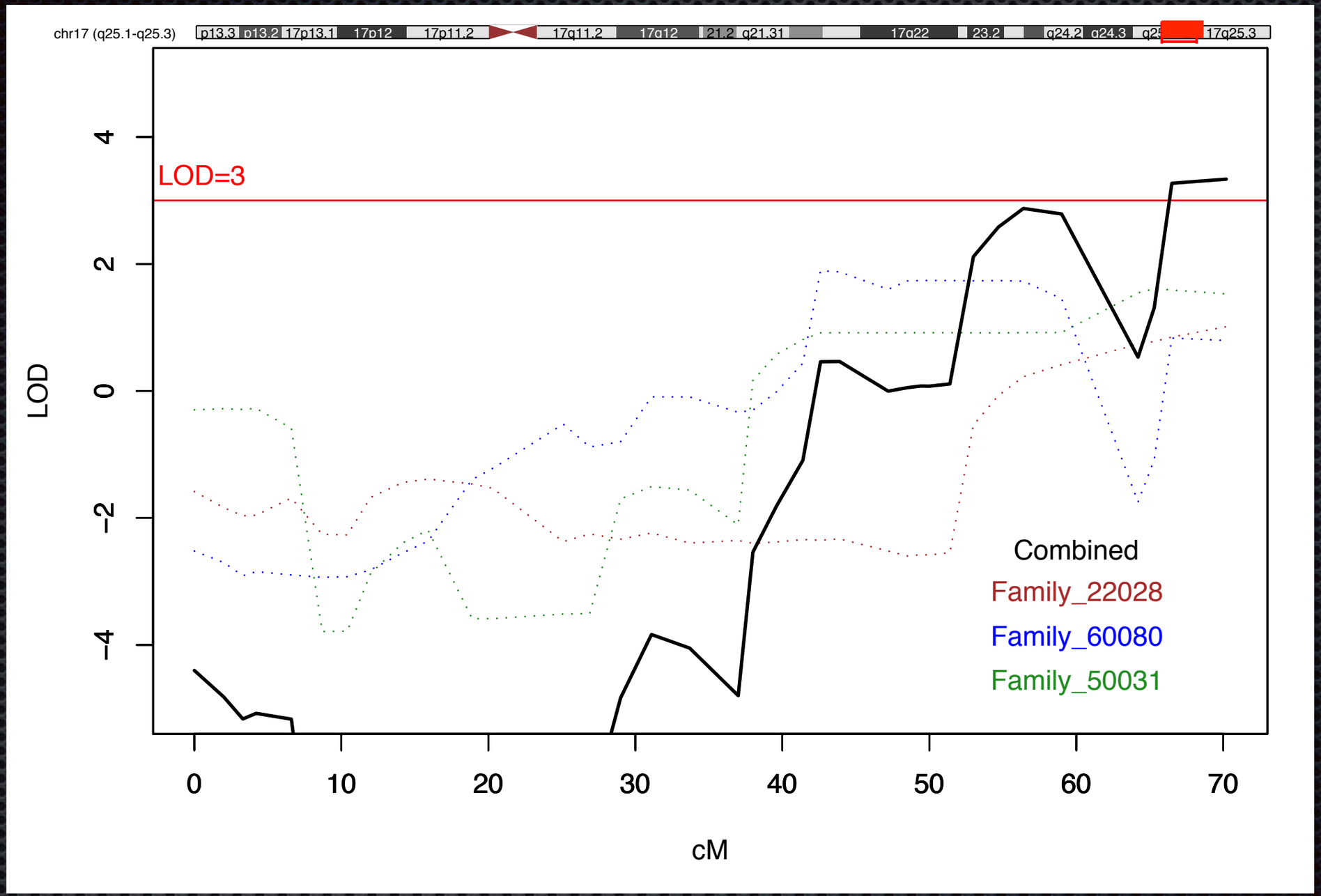
Diagnostic test?

Linkage analysis

- Mendelian disorder
- Multiple pedigress
- Dominant

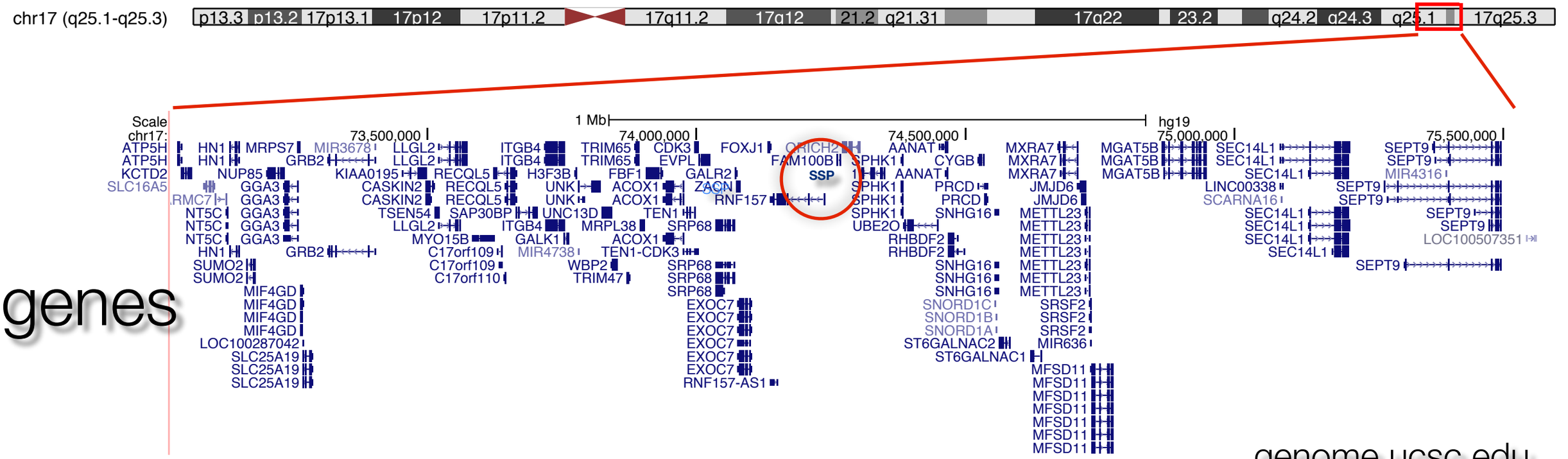
- Linkage analysis
  - Identify region of genome with disease locus
  - Genetic markers (known) and disease locus (unknown) are co-inherited

Linkage results



- Linkage to chromosome 17q
- 2.4 Mb region (large)

Candidate region



genome.ucsc.edu

- 72 genes
- Candidate genes?
  - Function?
  - Expression pattern?
- *SSP* and others

What are our options?

- Sanger sequencing of candidate genes?
  - Potentially simple and quick
  - Potentially slow and expensive
- Capture/sequencing region under linkage peak?
  - Good chance of success (can capture exons+++)
  - Design and test new capture array - expensive/time
- Exome sequencing?
  - Off-the-shelf reagents available, comprehensive exons
  - Mutation may not be exonic

Resequencing: mutation detection

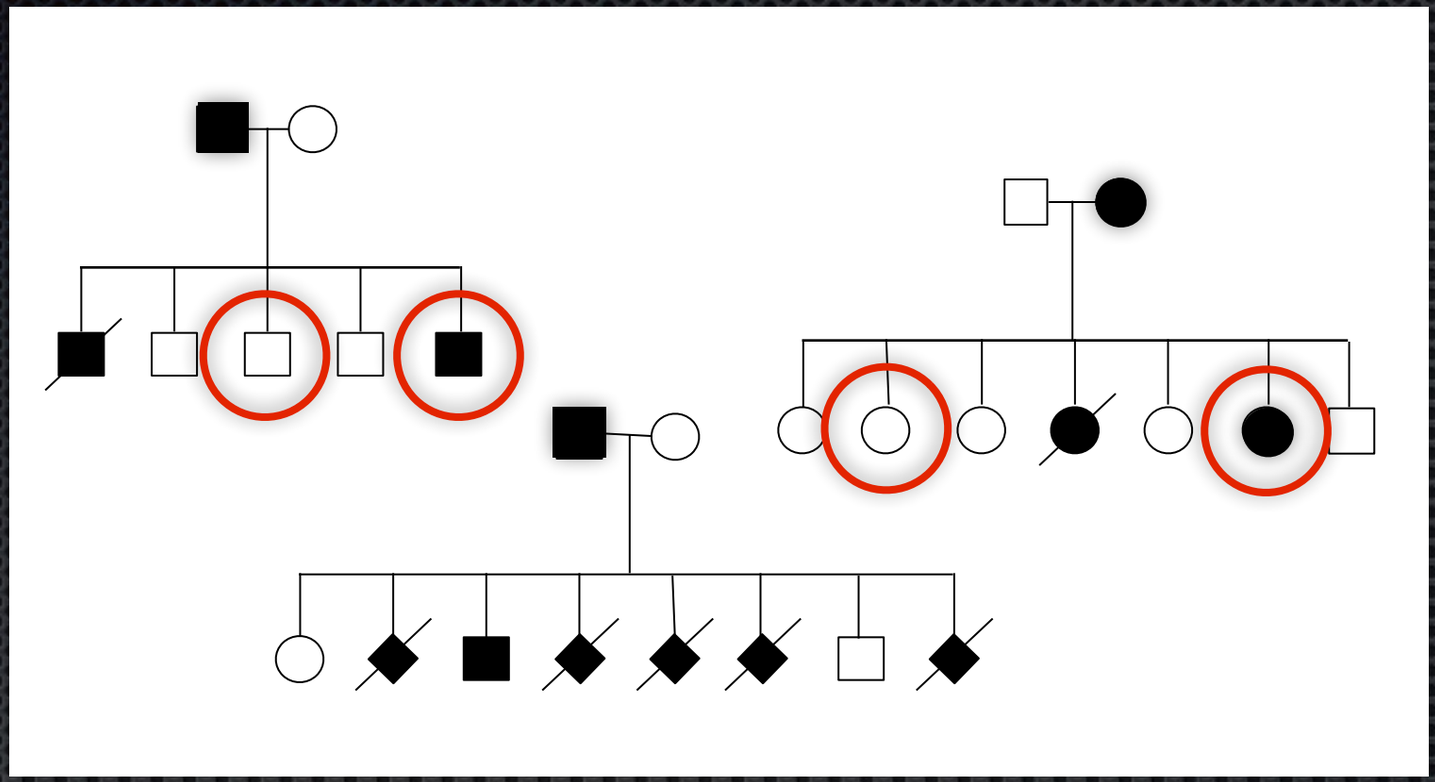
Genomic region unknown

- Rare Mendelian disorders
- Sequence capture - exome
- RNAseq

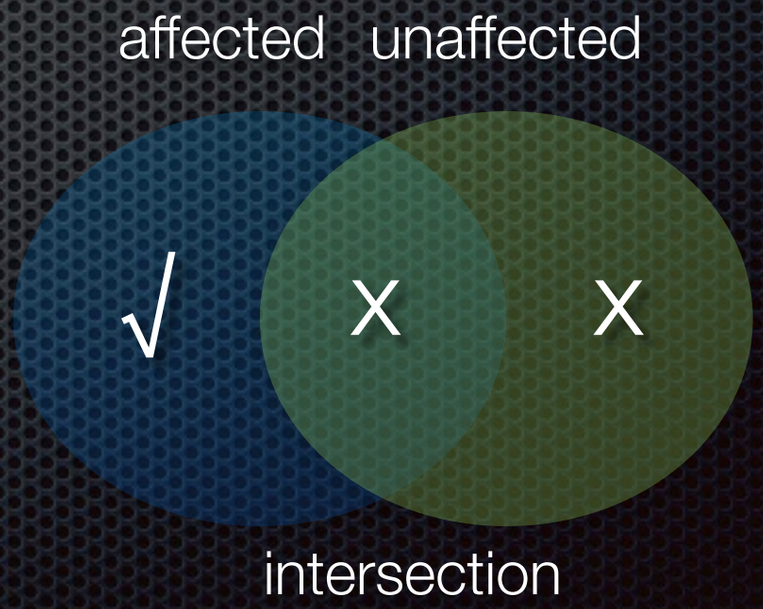
Genomic region known

- Linkage peak
- GWAS
- Sequence capture
  - Region of interest
  - **Exome**

Exome sequencing



- 4 individuals
- 2 affected, 2 unaffected
- Identify variants only in affected



Aim

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;39333
```

FASTQ format

R|G

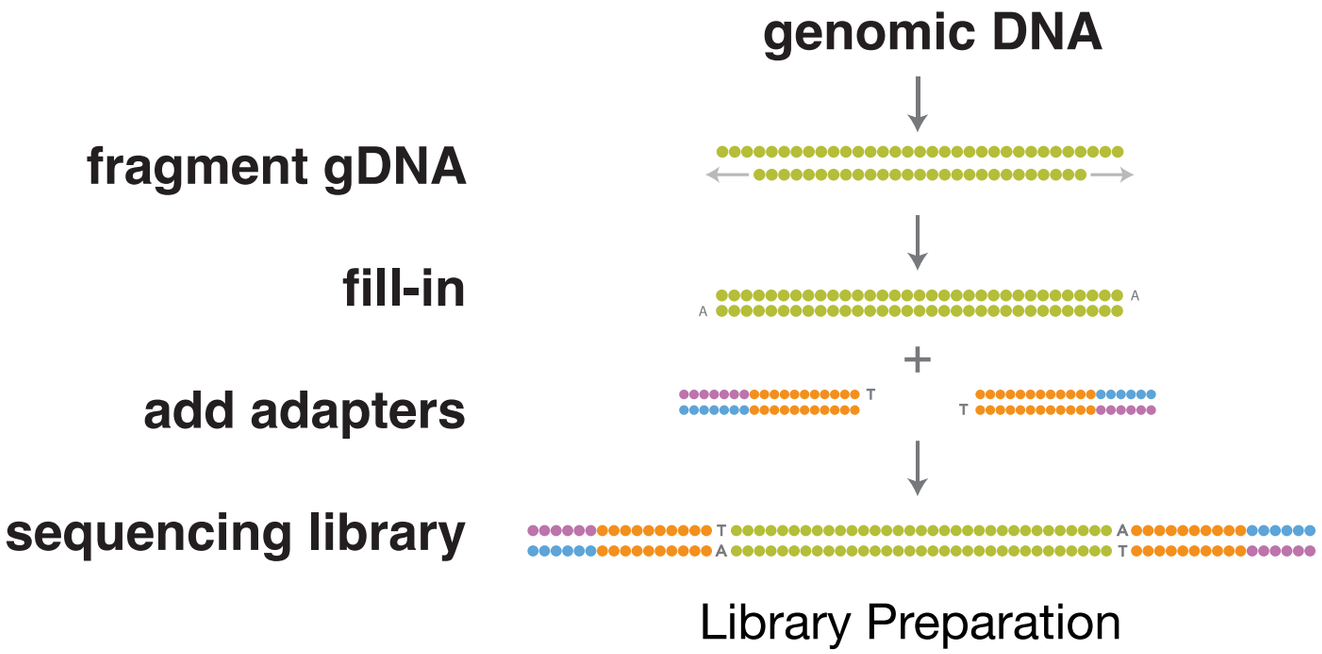
Compare to reference

Sequence

Mutation

Exome capture in 4 easy steps

1. Library preparation



2. Sequence capture

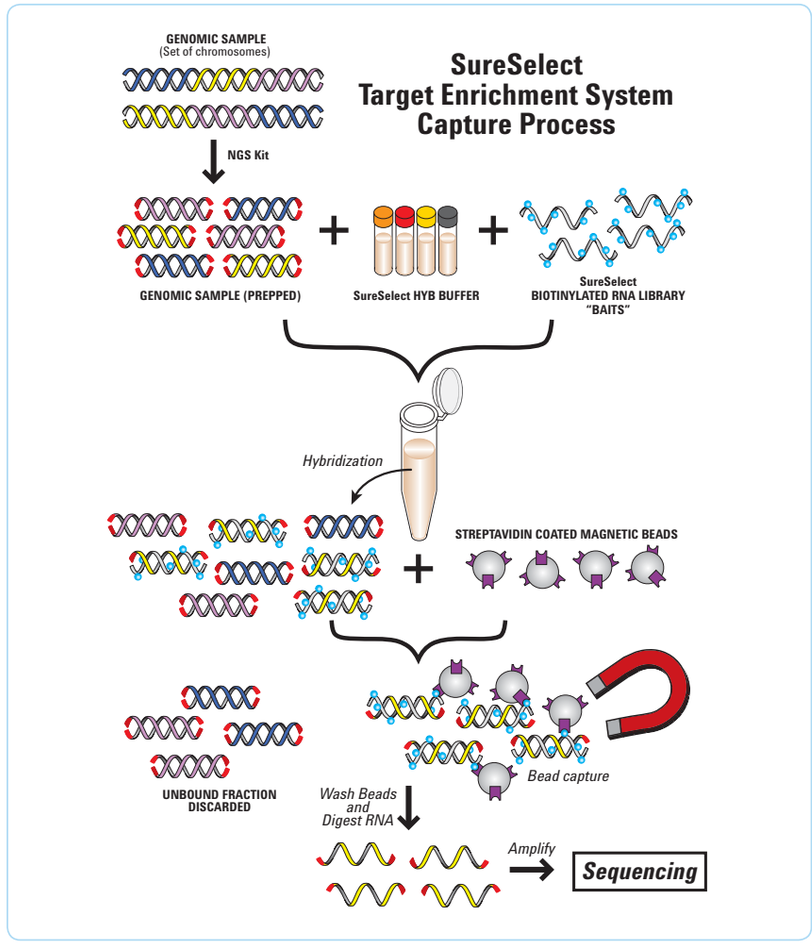
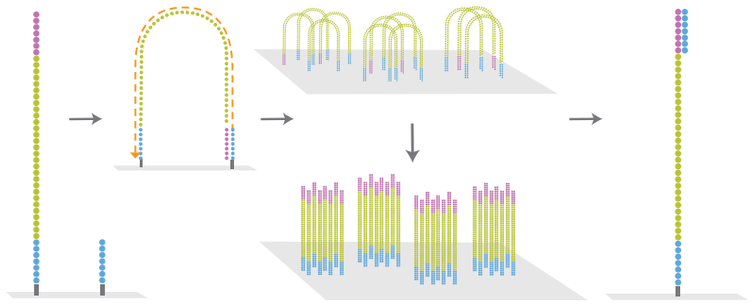


Figure 1. SureSelect Target Enrichment System Workflow

3. Illumina sequencing



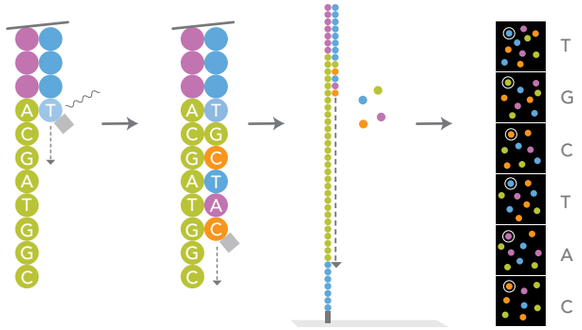
Cluster sequence library

HiSeq2000

cBot

Cluster Generation

Sequencing by Synthesis



Sequencing by synthesis

4. Analysis

Align reads to reference genome

▼

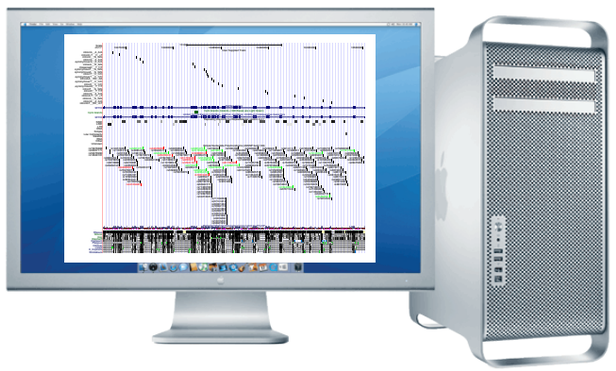
Call variants

▼

Filter variants

▼

View



Software

Step	Software	Link
QC/preprocessing	FastQC	http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/
	FASTX-Toolkit	http://hannonlab.cshl.edu/fastx_toolkit/
Aligning	Novoalign	http://www.novocraft.com
	BWA	http://bio-bwa.sourceforge.net/
Variant calling	Samtools	http://samtools.sourceforge.net/
	VCFtools	http://vcftools.sourceforge.net/
Variant annotation	SeattleSeq Annotation	http://gvs.gs.washington.edu/SeattleSeqAnnotation/
Data viewing	IGV	http://www.broadinstitute.org/software/igv/
	UCSC Browser	http://genome.ucsc.edu/
Misc	tabix	http://samtools.sourceforge.net/tabix.shtml
	Perl	http://www.perl.org/
	R	http://www.r-project.org/

Resequencing steps

- Capture
- Sequence
- Align to reference
- Call variants
- Filter
- View

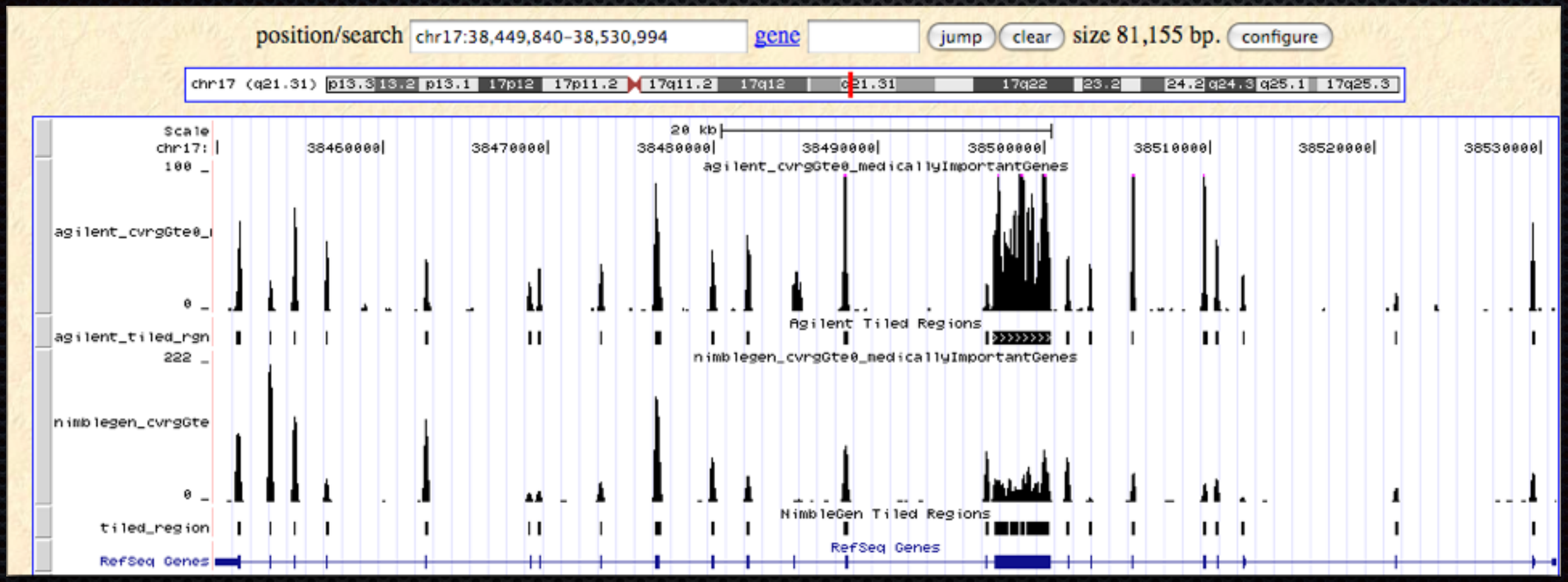
} The hard part

ExomeSeq results

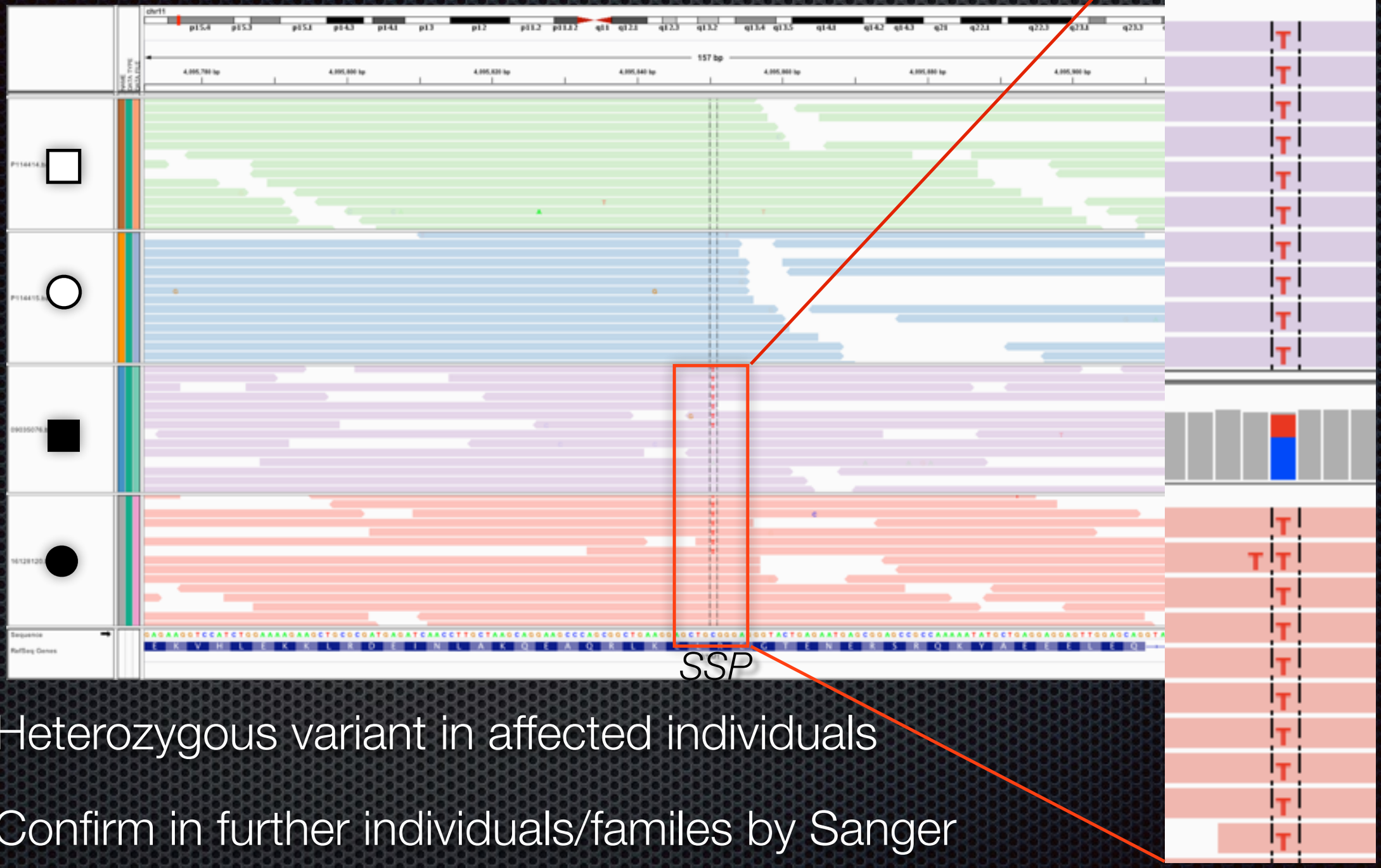
	■	□	●	○
All variants	21456	10987	4356	2489
Novel	1345	547	245	148
Missense	567	296	143	67
Damaging	298	139	56	22
Highly conserved	143	30	9	**2**

- Very good candidate variants
- How can we confirm these?
- Identify correct variant?

Exome sequencing data



Data viewer - IGV



- Heterozygous variant in affected individuals
- Confirm in further individuals/familes by Sanger
- Mutation in *SSP* causes ARAS syndrome

Genetic diagnosis and rare diseases

2009	2013
Confirm diagnosis	Confirm or clarify diagnosis
Test series of single genes	1000s genes tested at once
Often international labs	Carried out in Norway
Very expensive	Costs decreasing
Time-consuming (years)	Fast (weeks)

A diagnostic revolution

(common/rare)

Summary

- High-throughput sequencing
  - Dramatic increase in sequence production
  - Many applications on one platform
  - Field new and moving very quickly
  - Diagnostic (exome) sequencing in place
  - Huge impact on human/medical genetics

- Challenges/opportunities
  - Data storage/backup/distribution
  - Data analysis
  - Whole-genome sequencing?

People

NSC

Siri Okkenhaug
Magnus Leithaug
Sari Thiele
Kristine Fjelland
Monica Solbakken
Rune Moe
Tim Hughes
Ave Tooming-Klunderud
Morten Skage
Gregor Gilfillan
Ying Sheng
Lex Nederbracht
Sissel Jentoft
Robert Lyle
Kjetill Jakobsen
Dag Undlien

MedGen OUS

Hanne Sorte
Gregor Gilfillan
Martin Hammerø
Yvan Strahm
Beate Skinningsrud
Trine Prescott
Asbjørg Stray-Pedersen
Olaug Rødningen
Robert Lyle
Dag Undlien

post@sequencing.uio.no
Robert.Lyle@medisin.uio.no