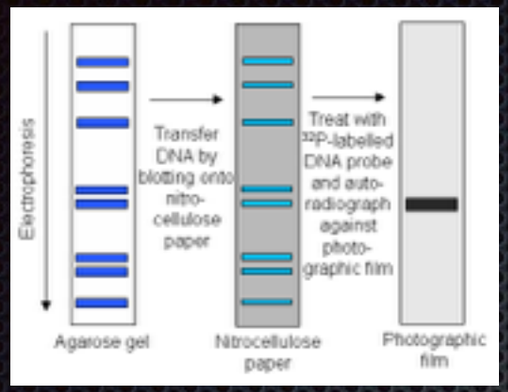
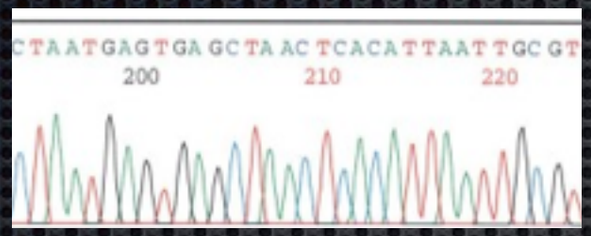
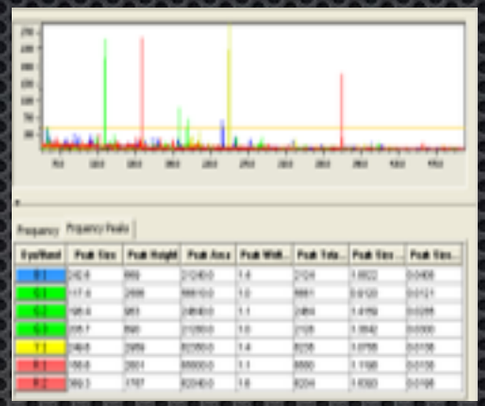
Methods for identifying variants/aberrations



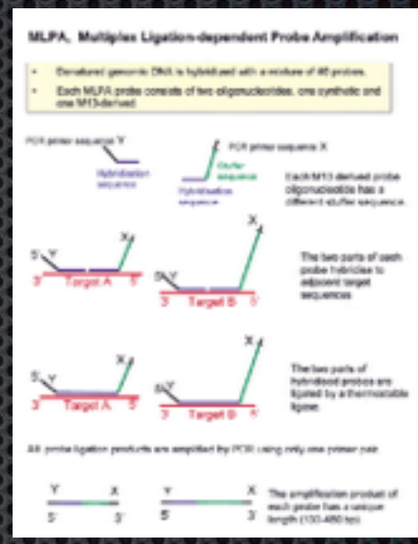
Southern blotting



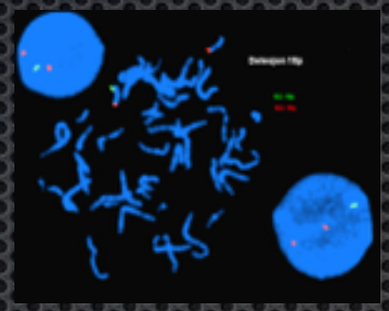
DNA (Sanger) sequencing



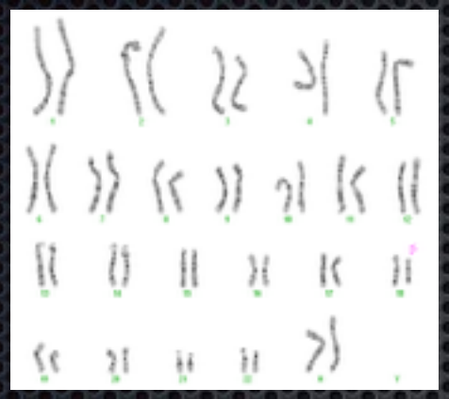
Fragment analysis



MLPA

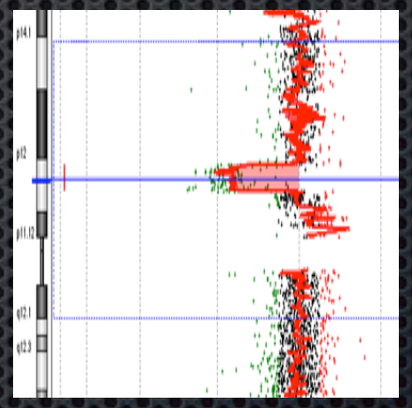


FISH



Karyotyping

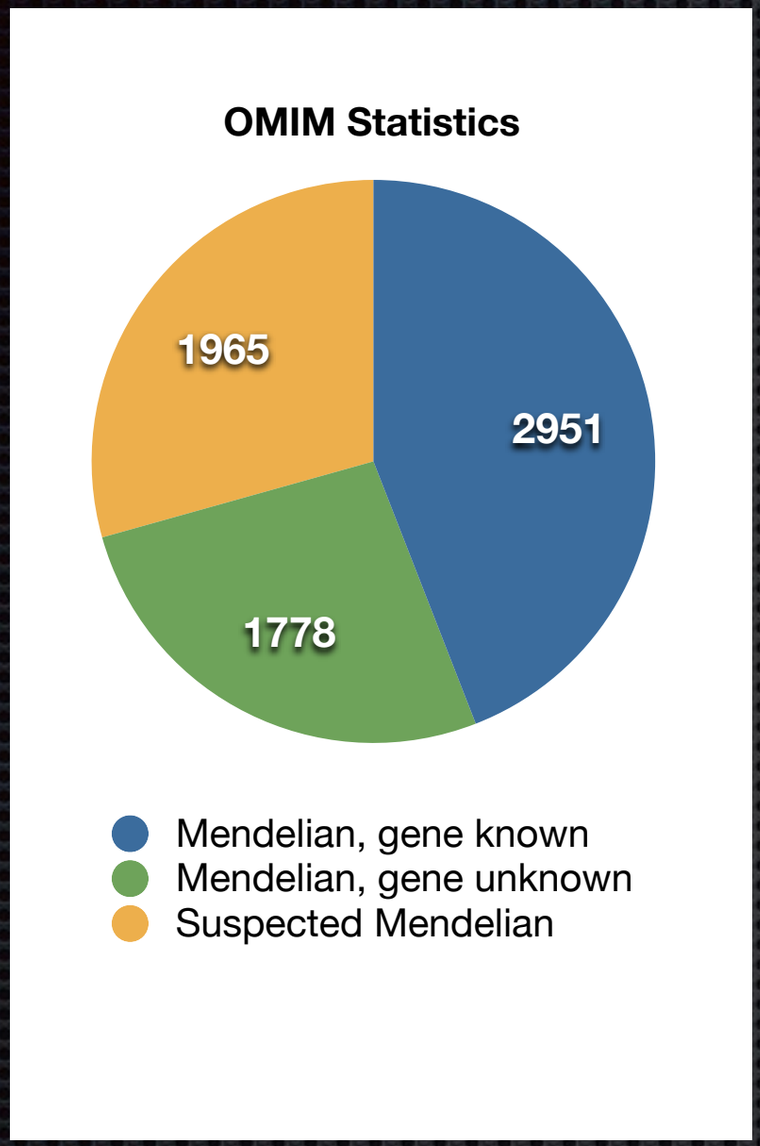
SNP array



Array CGH

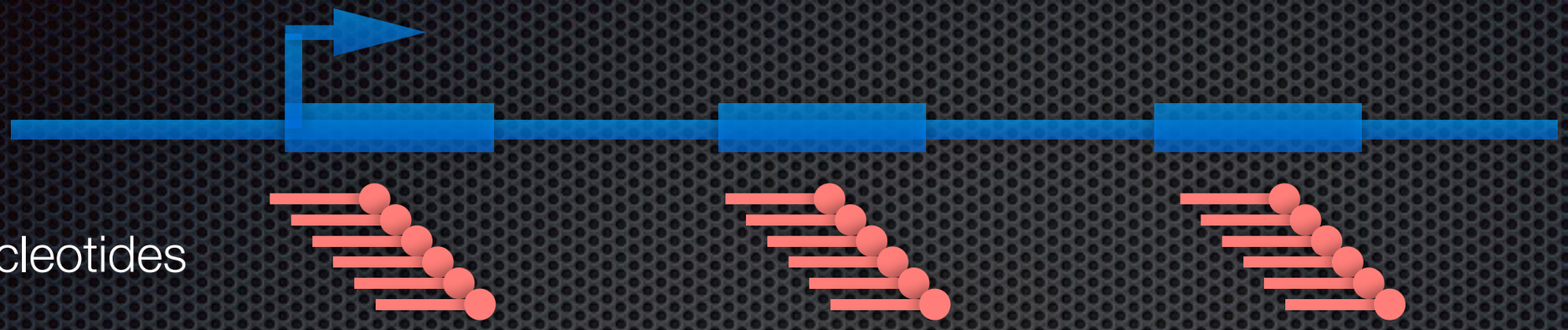
- Low throughput (limited number of loci per run)
- Detect specific types of variation

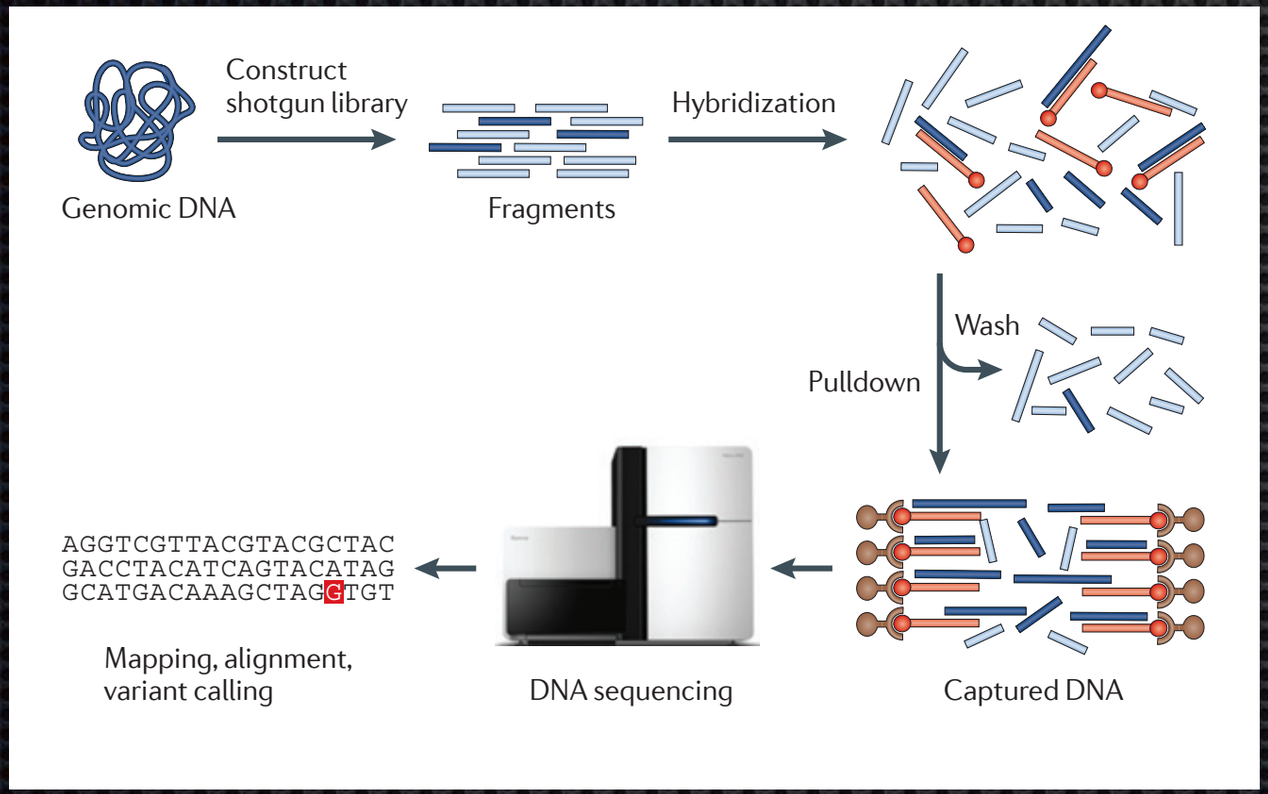
Mendelian disease in man



- Protein coding regions of the human genome (**the exome**) constitute approximately 1.5% of the total
- Exome contains ~85% of the mutations with large effects on disease-related traits
- ***Exome sequencing***
- ***HTS in research and routine diagnostics?***

Exome sequencing





- 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

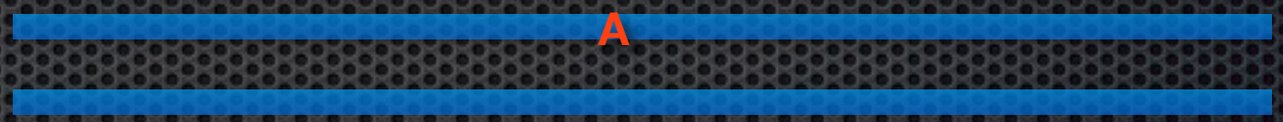
Mapping/aligning sequence

Types of genetic variation

SNP/SNV

Homozygous

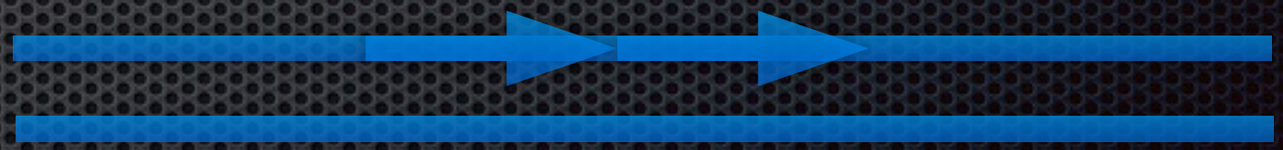
Heterozygous



CNV/indel

Deletion
(hemizygous)

Duplication



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

Effect of variants

- silent - no amino acid change
- missense - change amino acid
- nonsense - premature stop codon
- frameshift - shift the codon frame
- Indels - multiple effects

Codon table

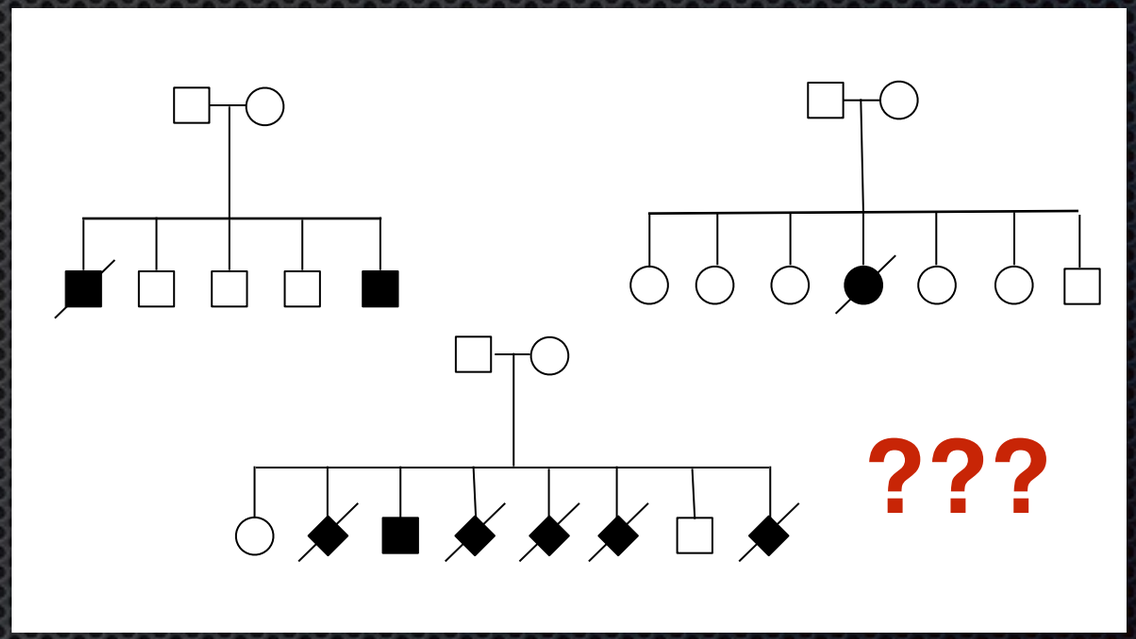
1st letter	Second Letter: U		Second Letter: C		Second Letter: A		Second Letter: G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Effect of variants

- silent - no amino acid change
- missense - change amino acid
- nonsense - premature stop codon
- frameshift - shift the codon frame
- Indels - multiple effects

What's in an exome?

- Compare 2 exomes
- ~20 000 variants
- ~10 000 silent
- ~10 000 missense
- ~100 nonsense
  - loss-of-function variant



Which variants cause disease?

- Many, many variants will be found
- Which variants are deleterious?
- Novel, low frequency (dbSNP, 1000genomes, HGMD, ESP6500 etc.)
- Synonymous/non-synonymous?
- Conserved?
- Alter protein structure?
- Biology?

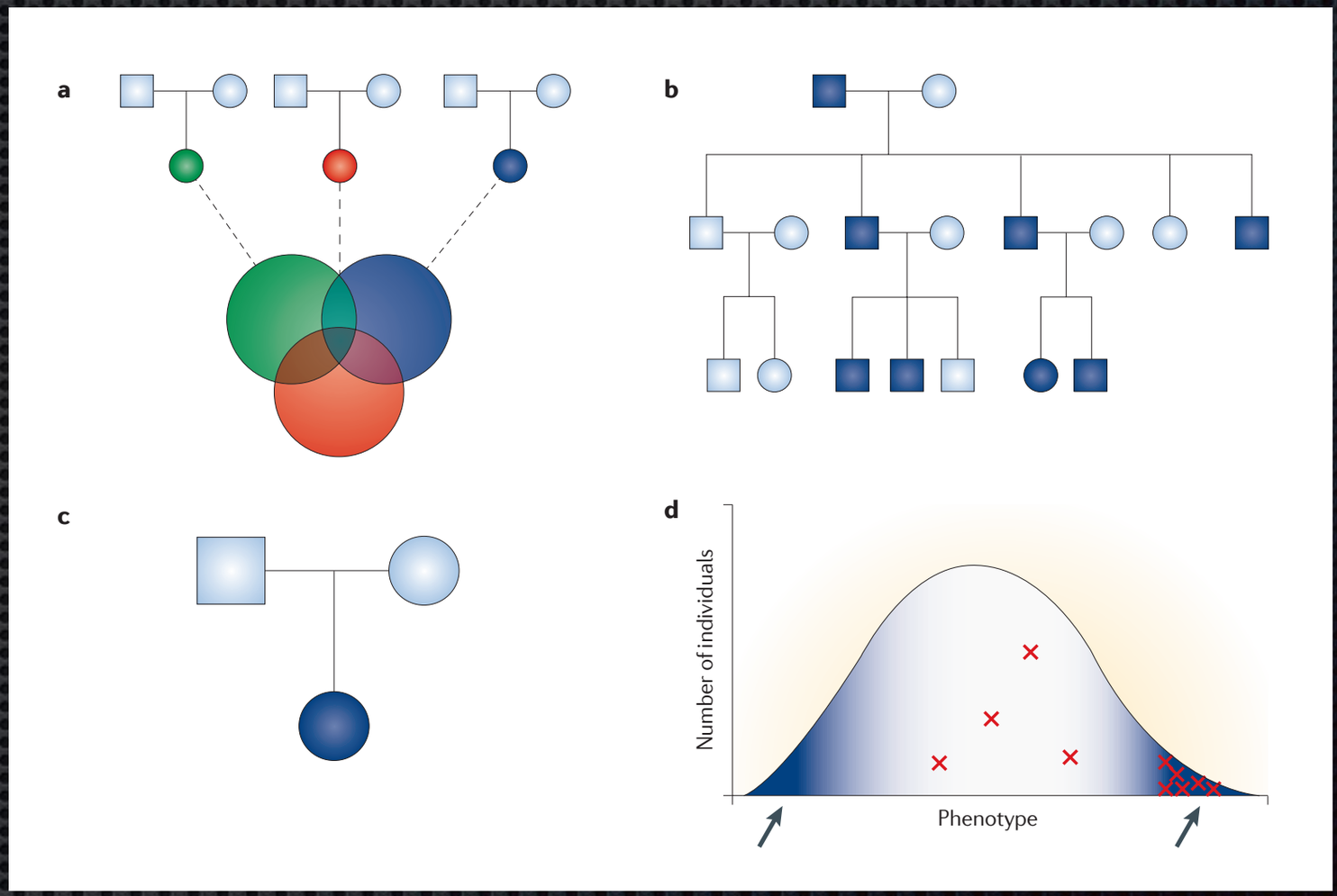
Bioinformatic tools

SNPnexus
PolyPhen2
MutationTaster
ANNOVAR
SIFT
SeattleSeq Annotation

This is the hard part

Genetics strategies to identify mutations

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

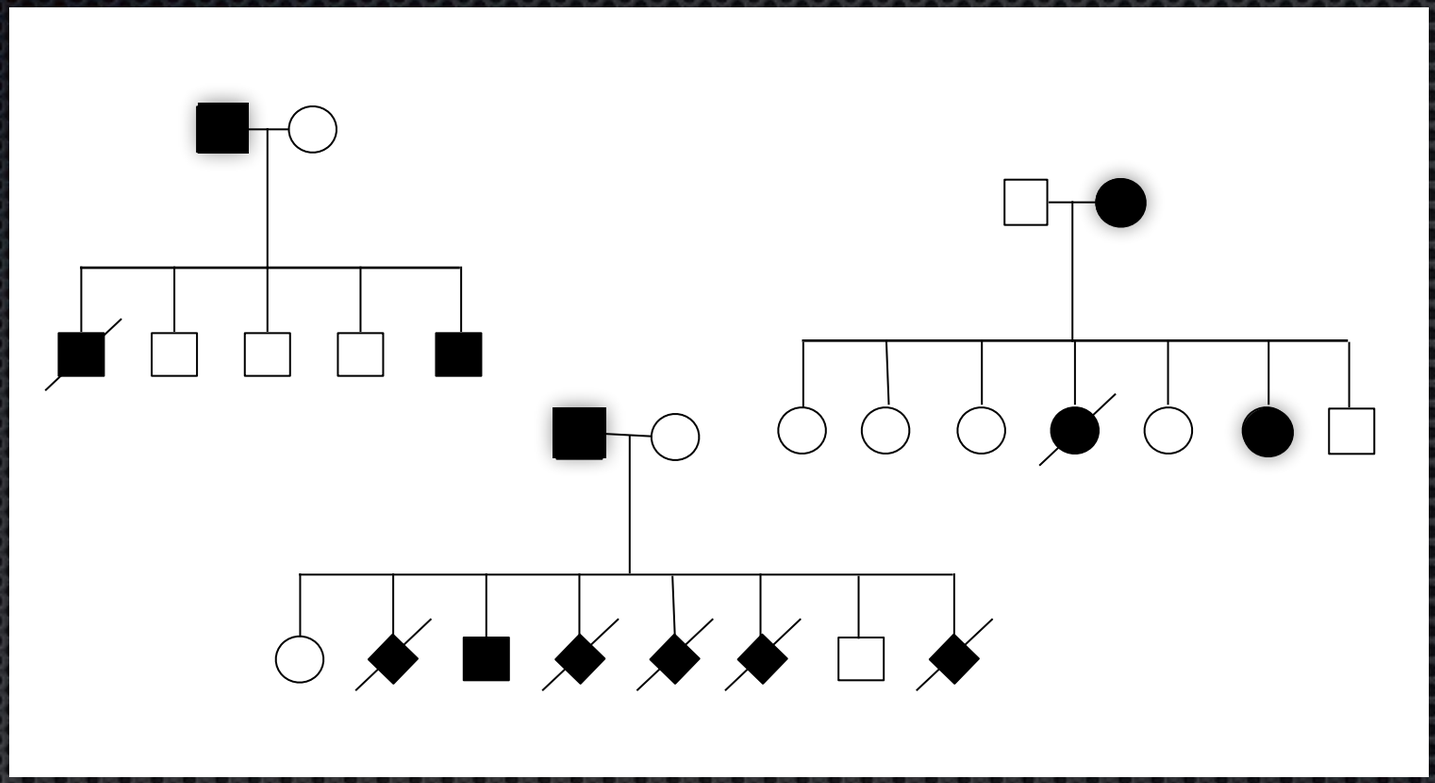


Bamshad et al. Nature Reviews Genetics Nov 2011

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?

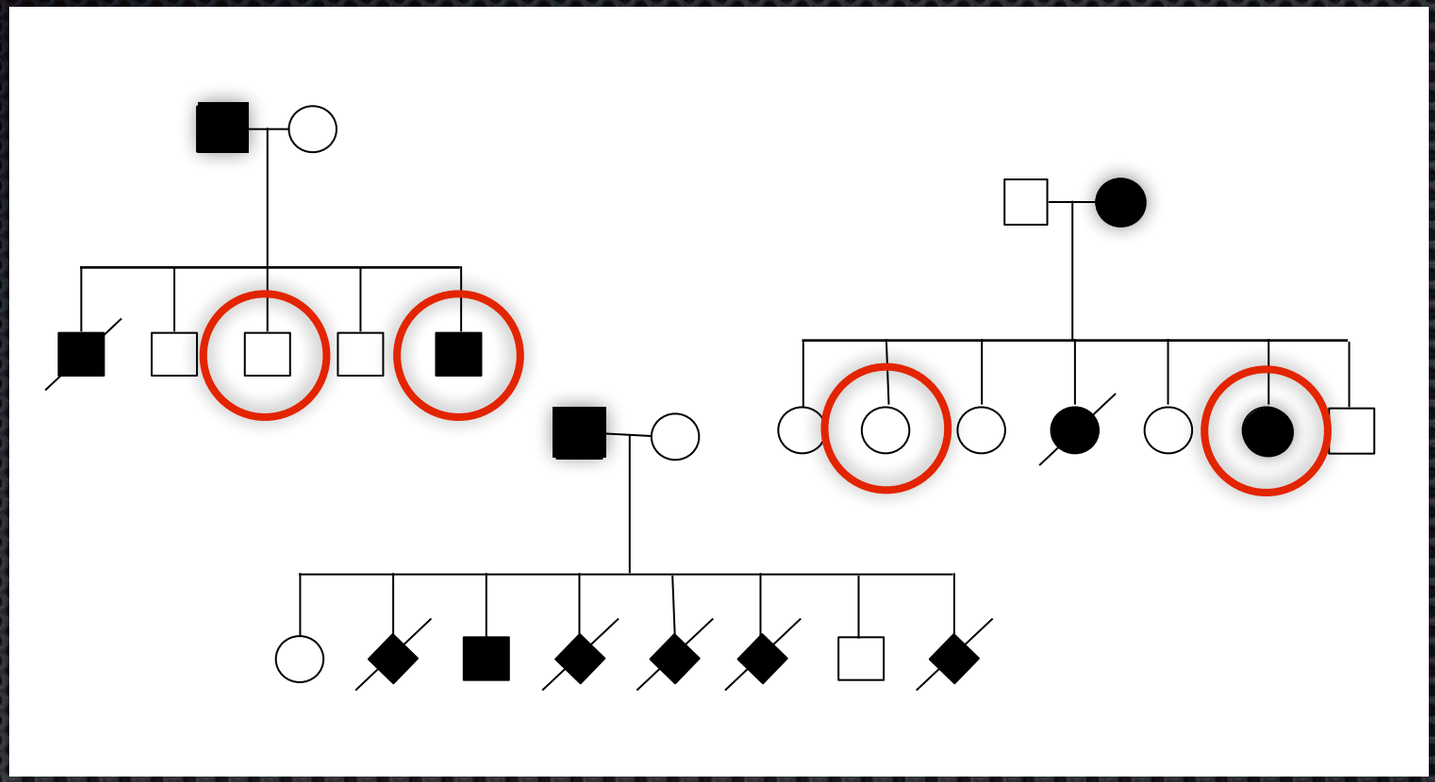
Previously…

- Mendelian disorder
- Multiple pedigress

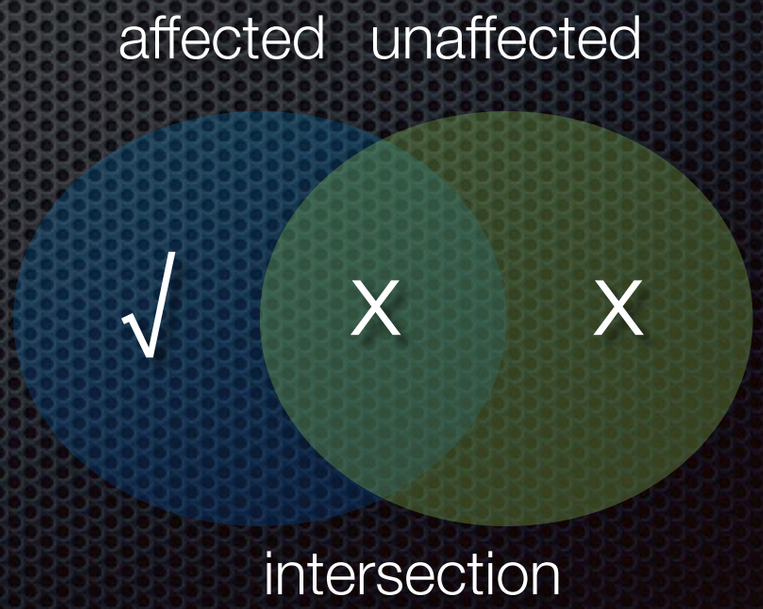
- Linkage analysis
  - Identify region of genome with disease locus

- Months - years to identify disease

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

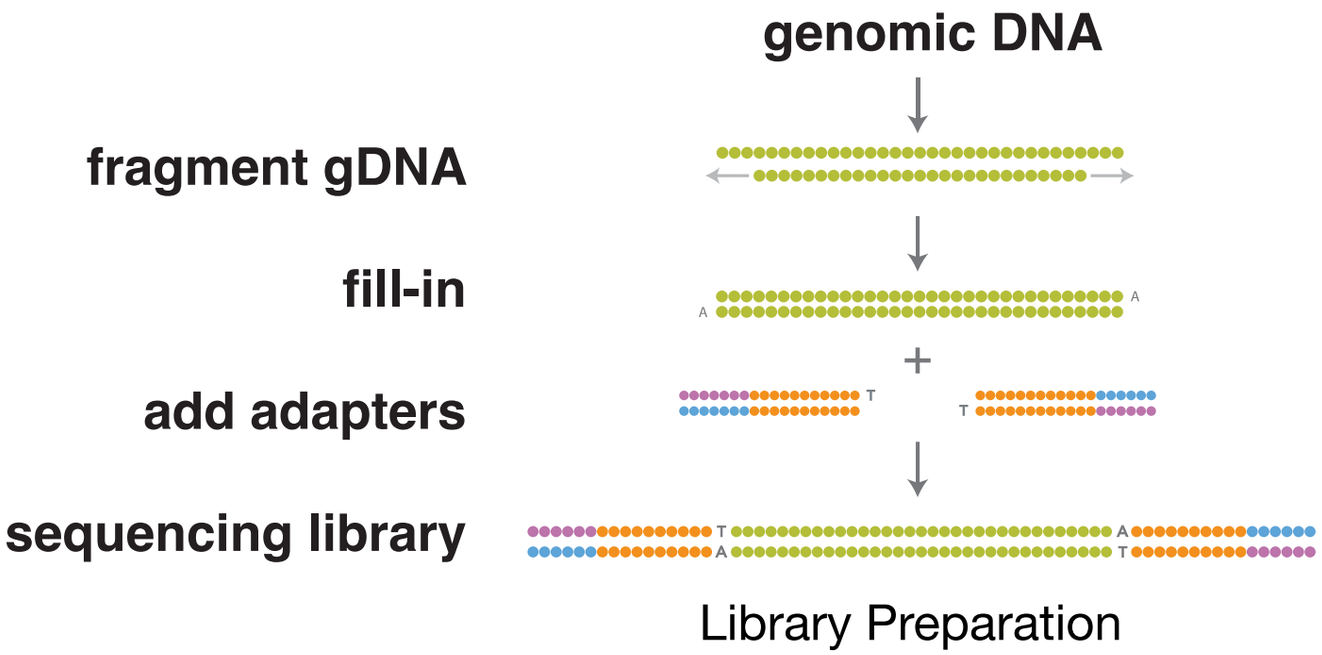
Compare to reference

Sequence

Mutation

Exome capture in 4 easy steps

1. Library preparation



Library Preparation

2. Sequence capture

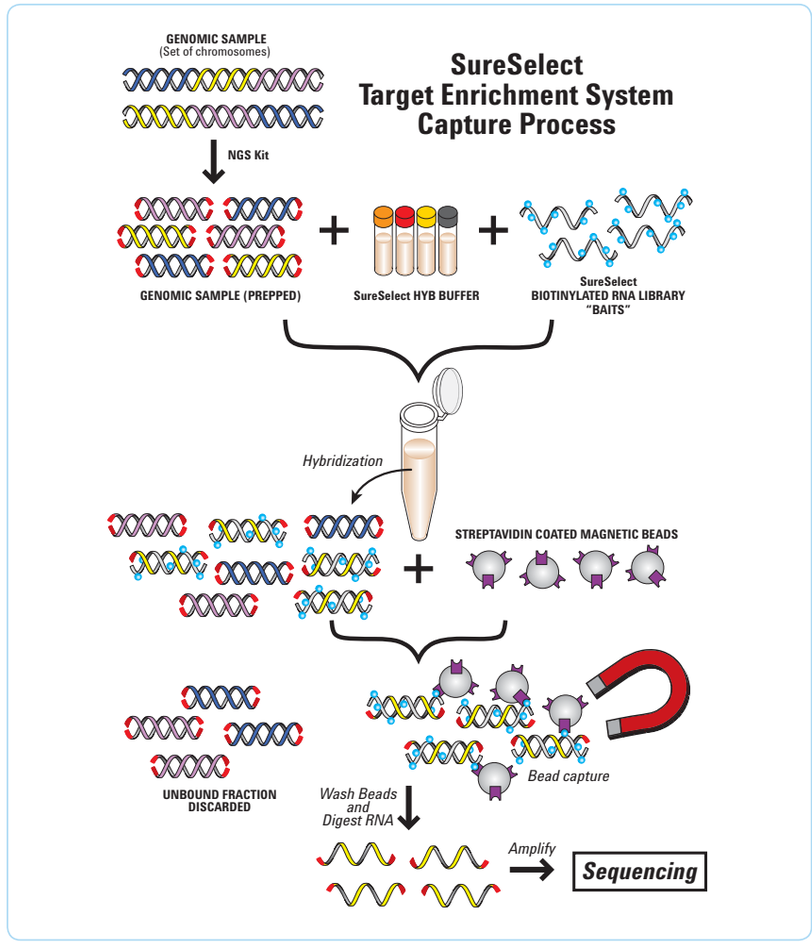
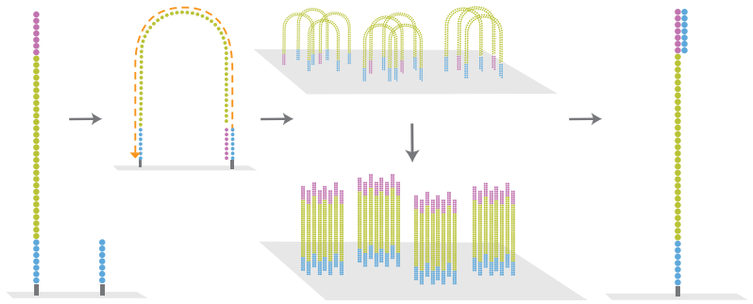


Figure 1. SureSelect Target Enrichment System Workflow

3. Illumina sequencing



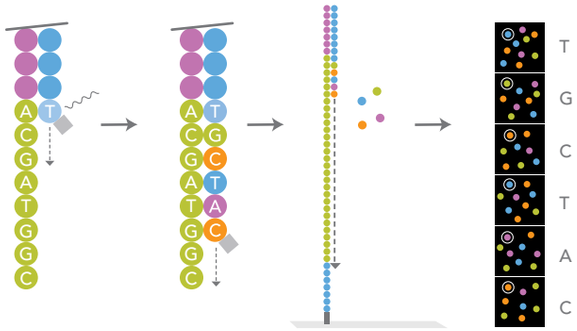
Cluster sequence library

HiSeq2000

cBot

Cluster Generation

Sequencing by Synthesis



Sequencing by synthesis

4. Analysis

Align reads to reference genome

▼

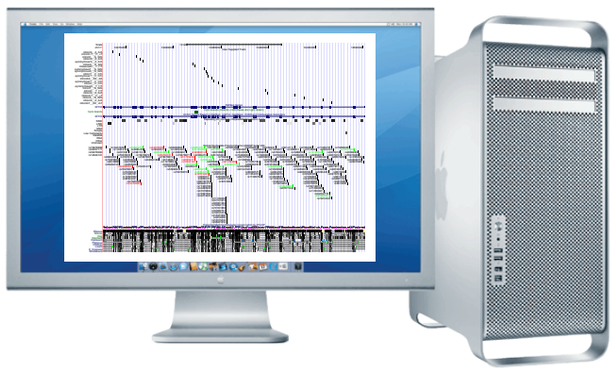
Call variants

▼

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
- Identify disease-causing candidates

} The hard part

- View