Understanding the science and technology of whole genome sequencing

Dag Undlien

Department of Medical Genetics

Oslo University Hospital

University of Oslo

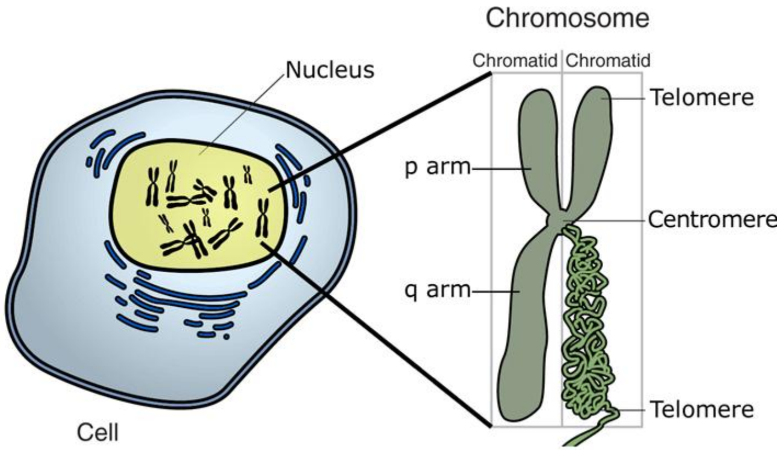
and

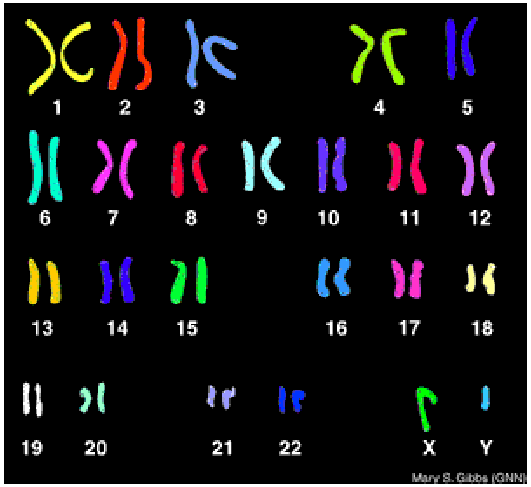
The Norwegian Sequencing Centre

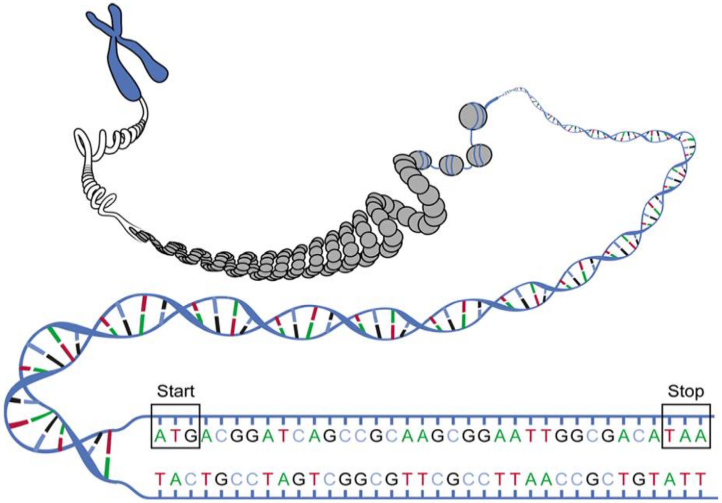
d.e.undlien@medisin.uio.no

DNA sequencing technology

DNA






Start
Stop
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

- 4 bases - A, G , C, T

GATC
CTAG

- Human ***genome*** ~3 billion bases

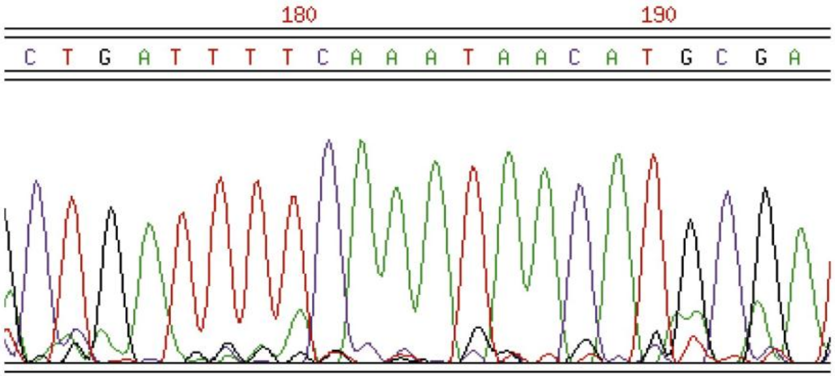
- How can we read the sequence of bases?

DNA sequencing

Mimic the way DNA replication occurs in living cells in a test tube.

Monitor the chemical reactions that take place

- Sanger sequencing
- Detect nucleotide extension with radioactivity or fluorescence
- Accurate but slow



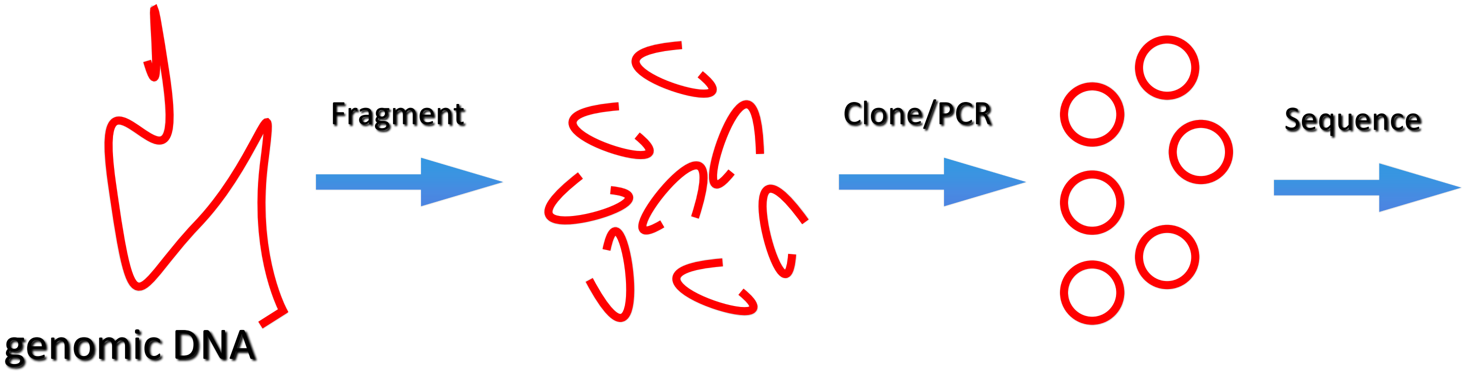
High-throughput sequencing

also known as next generation sequencing, deep sequencing, massively parallel sequencing………

Sequencing: old and next

LTS (Sanger)

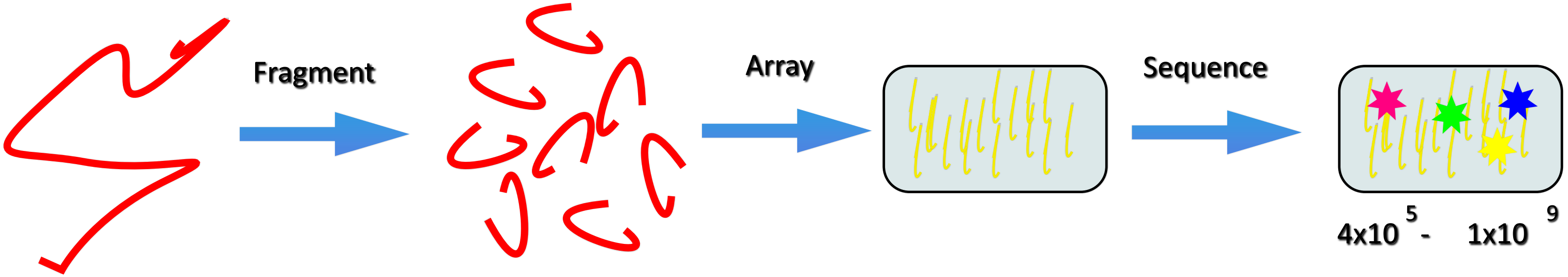
Molecules sequenced



1, 48, 96...

...unless you have a lot of machines

HTS (High-throughput sequencing)



$4x10^{5}$ - $1x10^{9}$

...on one machine

Massively parallel

Illumina sequence data

Random DNA library of short fragments ~300 bp

3 billion DNA sequence reads

50, 100 bp long

Single-end reads ←

Paired-end reads ↔

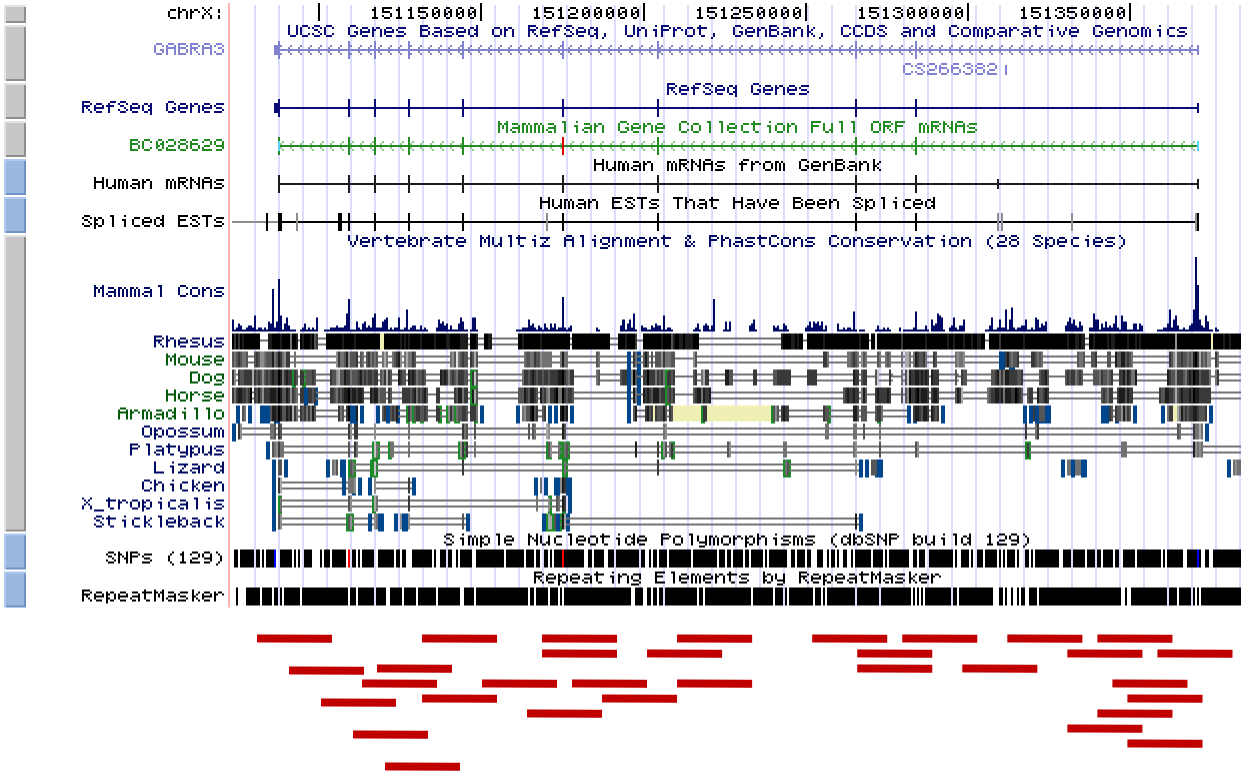
Run time: 1-9 days

Data volume: 1-500 GB

Sequencing platforms

Platform	454	Illumina HiSeq	Illumina MiSeq	PacBio	Ion Torrent
System cost	-	--	++	---	+++
Prep	-	+	++	+	+
Running cost	--	+	++	++	++
Run time	10 hours	1-9 days	27 hours	2 hours	2 hours
Read accuracy	99%	98%	98%	87%	98.8%
Read number	100000	3000000000	3500000	75000	6 x 10^6
Read length	400 bp	2x100	2x150	~2700 (10kb)	2-400 bp
Output	35 Mb	600 Gb	>1 Gb	90 Mb	>1 Gb

HT sequencing - commonalities

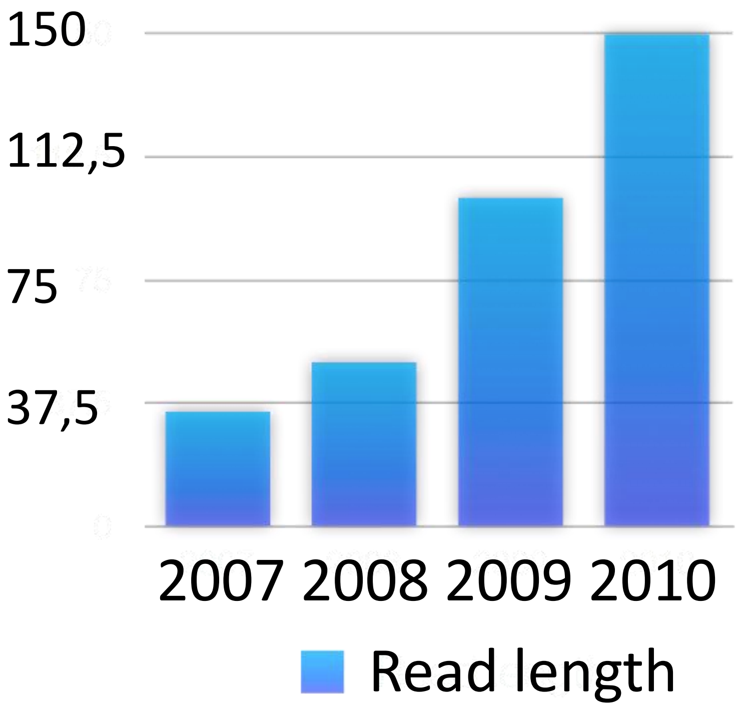

chrX:
151150000
151200000
151250000
151300000
151350000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
GABRA3
CS266382
RefSeq Genes
RefSeq Genes
BC028629
Mammalian Gene Collection Full ORF mRNAs
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Rhesus
Mouse
Dog
Horse
Armadillo
Opossum
Platypus
Lizard
Chicken
X_tropicalis
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 129)
SNPs (129)
Repeating Elements by RepeatMasker
RepeatMasker

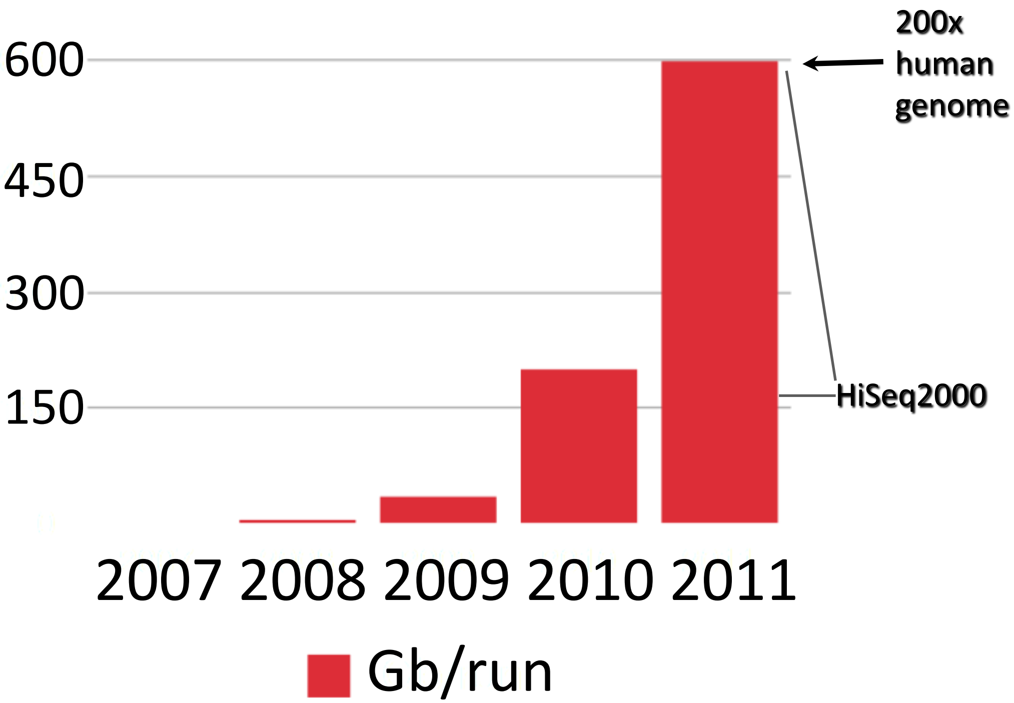
- Many short sequence fragments – assembly
- ”Stochastic” how many times a given sequence is sequenced
- The technology can be used for quantitative investigations
  - E.g. CNVs, gene expression
  - *Well suited for many non-genetic studies*

High throughput sequencing - applications

Platform	454	Illumina HiSeq	Illumina MiSeq*	PacBio	Ion Torrent
Resequencing	-	+++	++	+	+++
de novo	+++	+	+	+++	+++
metagenomics	+++	++	+	++	+++
mRNA	++	+++	++	++	++
miRNA	-	+++	+++	-	-
ChIP	-	+++	++	-	-
DNA meth	-	+++	+	???	-

Illumina throughput





- Human genome 3 billion bases - 3 Gb
- Illumina HiSeq 600 Gb per run
- 200 x human genome per run

So what?

Parameter	ABI 3100	ABI 3730	Illumina HiSeq
Read length	~700	~700	100 (x2)
Reads per run	16	96	3000000000
Run time	2 hours	30 minutes	9 days
Time for 1x human genome (3 Gb)	120 years	15 years	1 hour

HTS and medical genetics

- Finding mutations which cause disease

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.

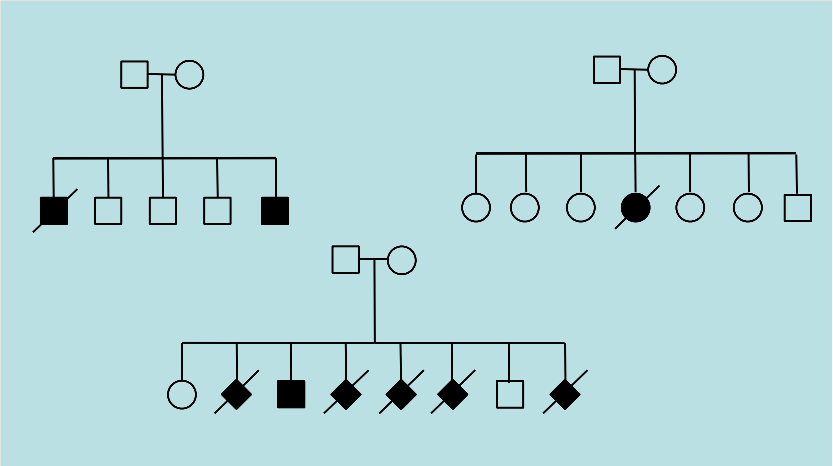
Many disorders are individually rare and difficult to recognize even for experts.

Clinical phenotypes can be non-specific and variable.

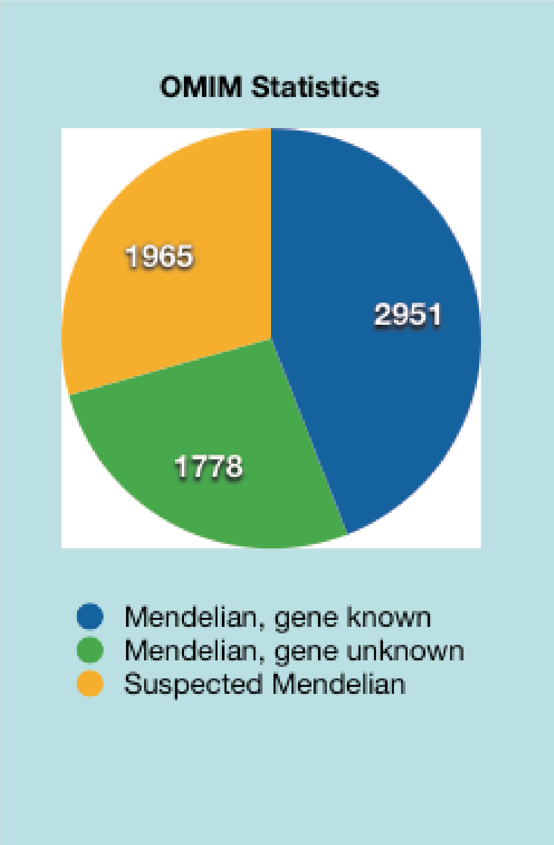
Many phenotypes are genetically heterogeneous.

- **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

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

Compare to reference

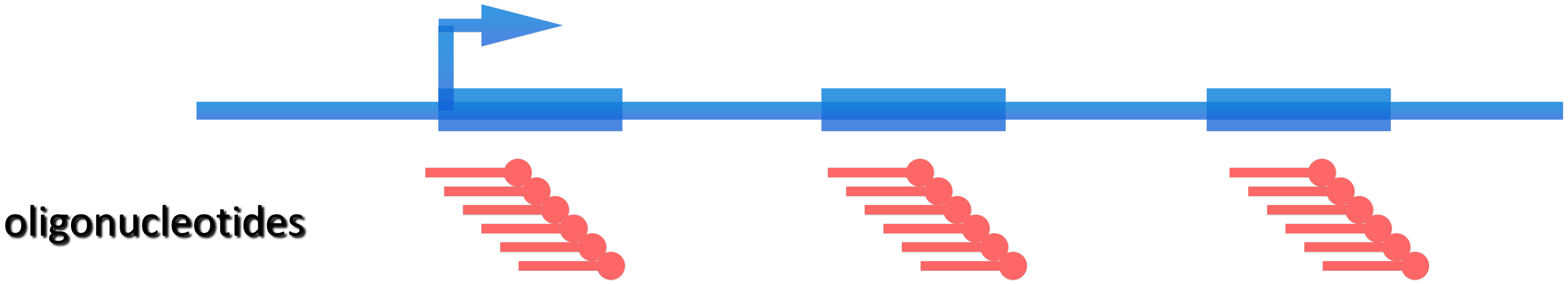
R|G

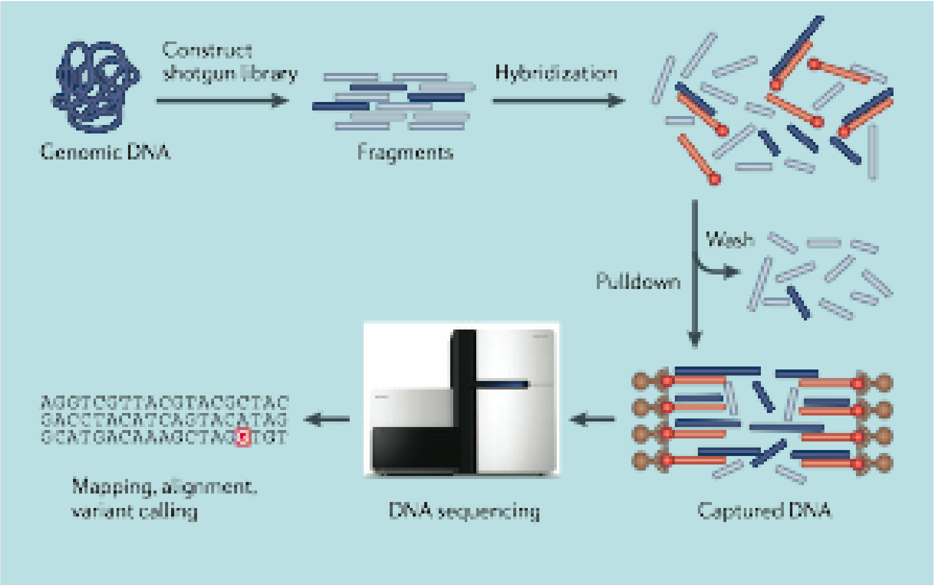
3 billion reads

3 billion bases

1 mutation

Exome sequencing – still more common than whole genome





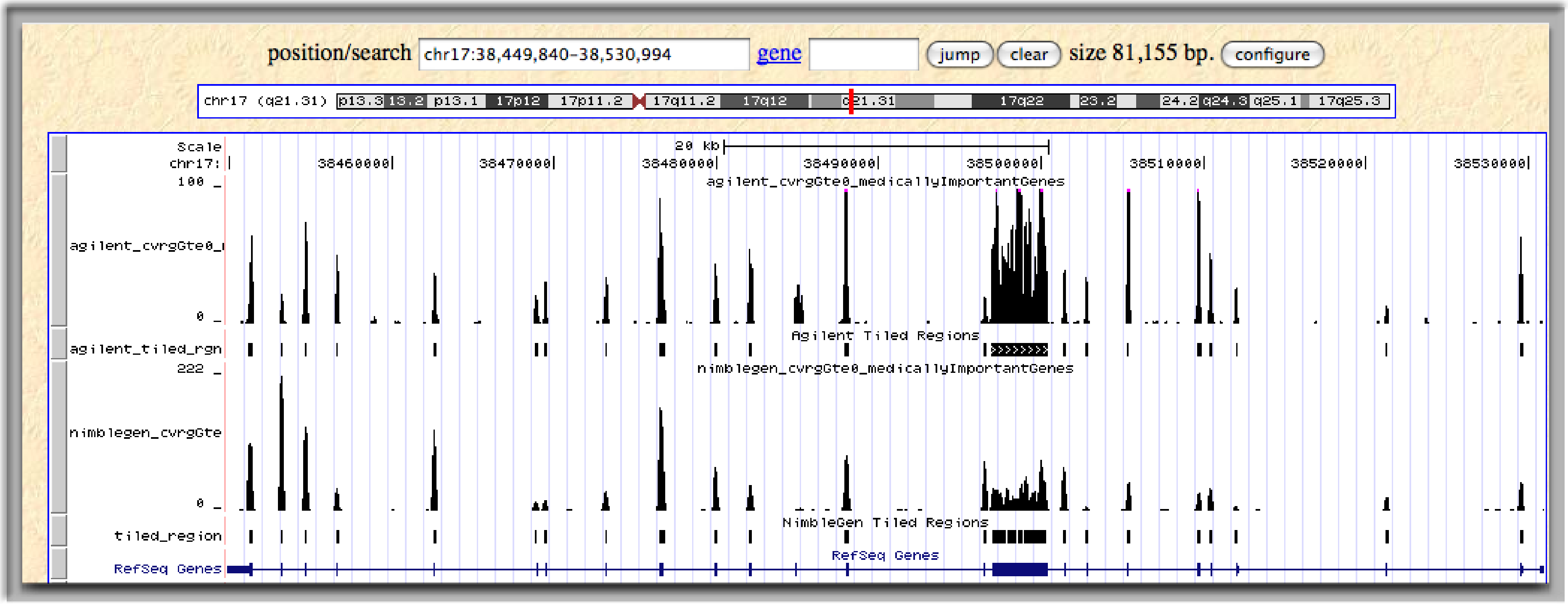
Design exome capture array

Make sequence library from patient DNA

Hybridize and capture

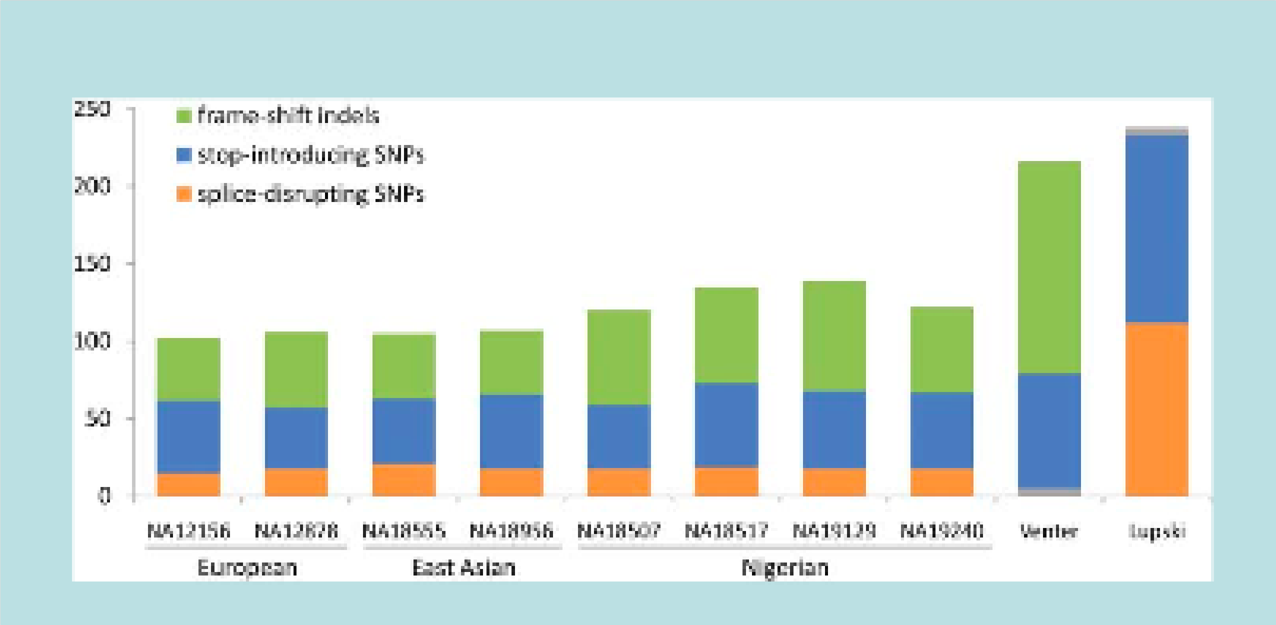
Sequence

Exome sequencing data



What’s in an exome?

> 20 000 variants



Many loss-of-function variants

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

Mardis *Genome Medicine* 2010, 2:84
http://genomemedicine.com/content/2/11/84

Genome Medicine

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*

Many, many variants will be found

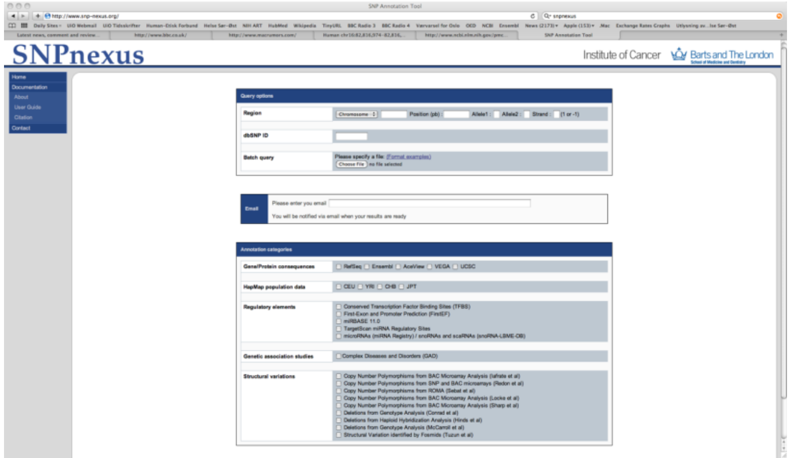
Which variants are deleterious?

Novel? (dbSNP, 1000genomes, HGMD)

Synonymous/non-synonymous?

Conserved?

Alter protein structure?



Query options

Region: Chromosome, Position (pb): , Allele1: , Allele2: , Strand: (1 or -1)

dbSNP ID

Batch query: Please specify a file: (Format examples) no file selected

Email: Please enter you email
You will be notified via email when your results are ready

Annotation categories

Gene/Protein consequences: □ RefSeq □ Ensembl □ AceView □ VEGA □ UCSC

HapMap population data: □ CEU □ YRI □ CHB □ JPT

Regulatory elements:
□ Conserved Transcription Factor Binding Sites (TFBS)
□ First-Exon and Promoter Prediction (FirstEF)
□ miRBASE 11.0
□ TargetScan miRNA Regulatory Sites
□ microRNAs (miRNA Registry) / snoRNAs and scaRNAs (snoRNA-LBME-DB)

Genetic association studies: □ Complex Diseases and Disorders (GAD)

Structural variations:
□ Copy Number Polymorphisms from BAC Microarray Analysis (Iafrate et al)
□ Copy Number Polymorphisms from SNP and BAC microarrays (Redon et al)
□ Copy Number Polymorphisms from ROMA (Sebat et al)
□ Copy Number Polymorphisms from BAC Microarray Analysis (Locke et al)
□ Copy Number Polymorphisms from BAC Microarray Analysis (Sharp et al)
□ Deletions from Genotype Analysis (Conrad et al)
□ Deletions from Haploid Hybridization Analysis (Hinds et al)
□ Deletions from Genotype Analysis (McCarroll et al)
□ Structural Variation identified by Fosmids (Tuzun et al)

SNPnexus
PolyPhen2
MutationTaster
ANNOVAR
SeattleSeq Annotation

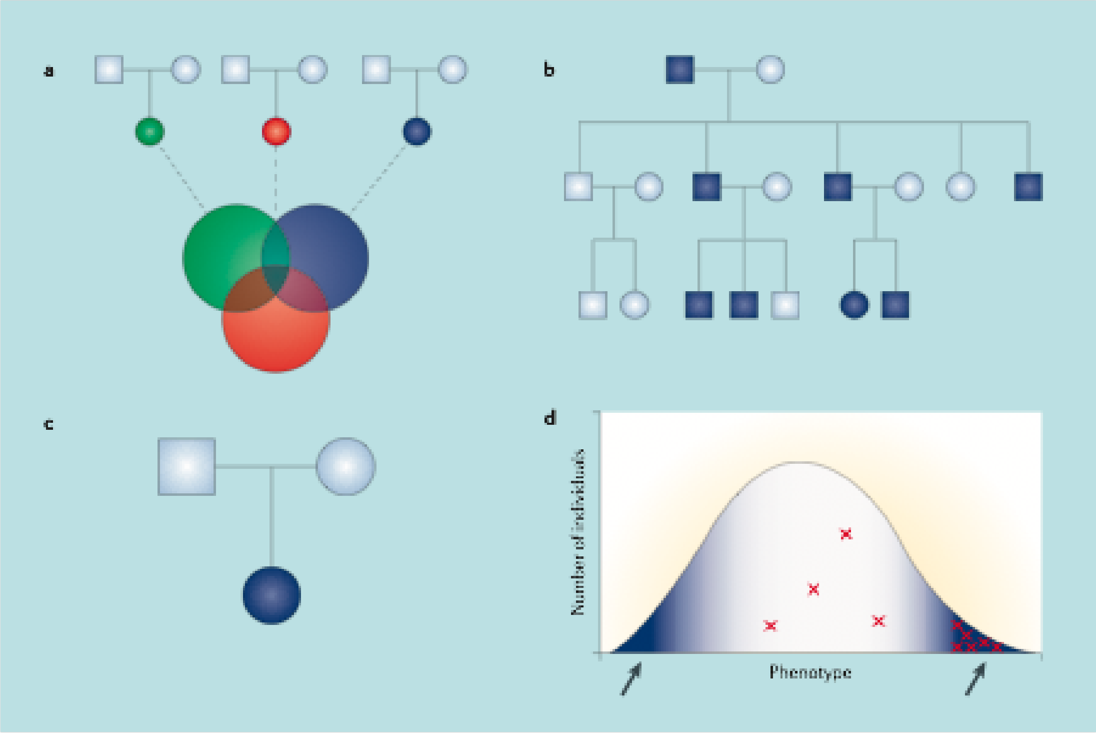
Strategies to identify mutations

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

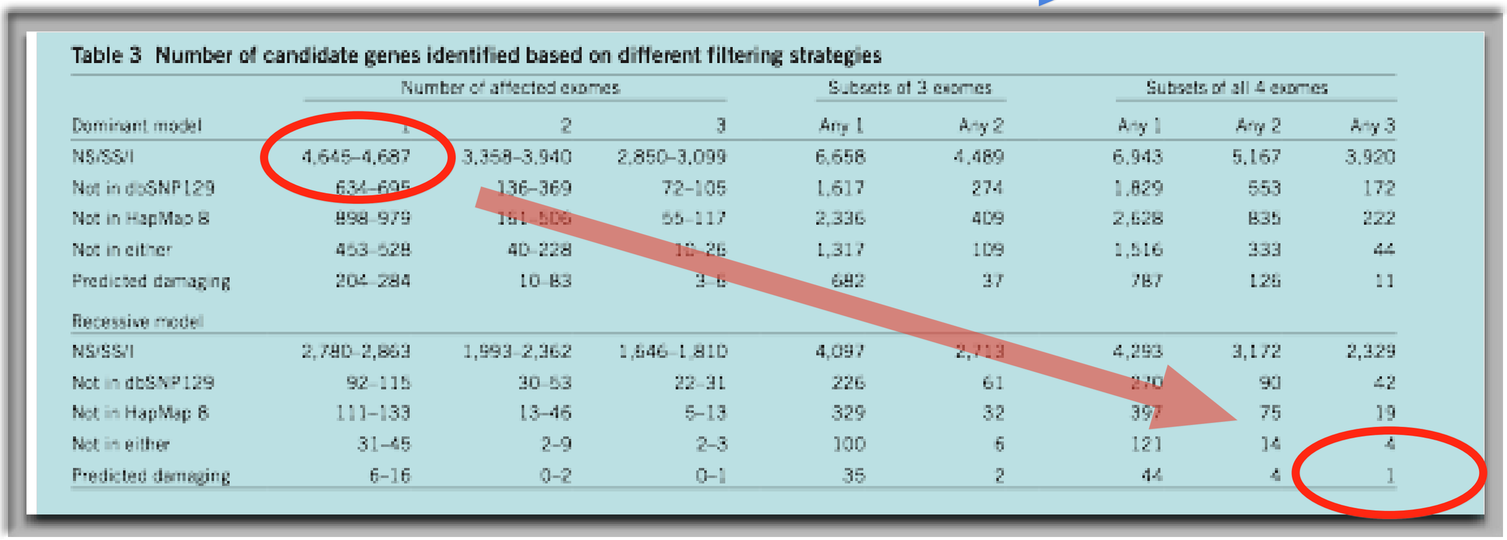


Table 3 Number of candidate genes identified based on different filtering strategies

	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	181–506	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

Genetic diagnosis and rare diseases

2009

Confirm diagnosis

Test series of single genes

Often international labs

Very expensive

Time-consuming (years)

2012

Confirm or clarify diagnosis

1000s genes tested at once

Costs decreasing

Fast (weeks)

A diagnostic revolution

The future?
Oxford Nanopore MinION

disposable device

plug directly into computer USB port

compatible with blood, serum and environm

~1 Gb sequence

~$900

Personal genomics?

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?
- Incidental findings