

Applications

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

*projected: Q4 2011-Q2 2012

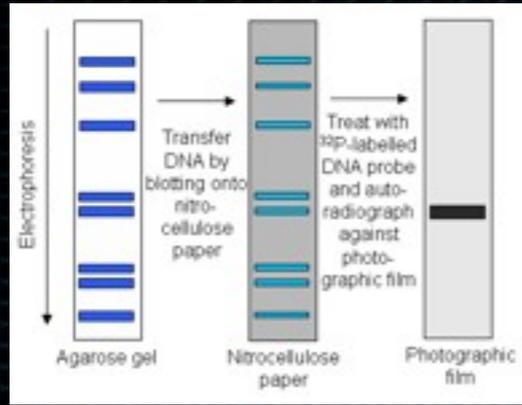
NSC platforms - applications



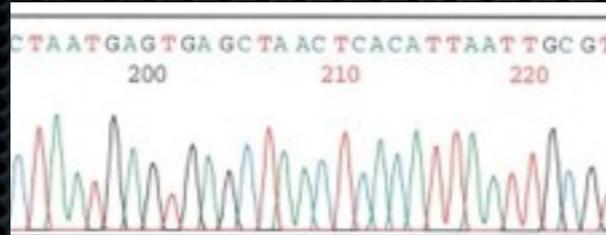
Platform	454	Illumina HiSeq	Illumina MiSeq*	PacBio*	Ion Torrent*
Resequencing	-	+++	++	+	+++
de novo	+++	+	+	+++	+++
metagenomics	+++	++	+	++	+++
mRNA	++	+++	++	++	++
miRNA	-	+++	+++	-	-
ChIP	-	+++	++	-	-
DNA meth	-	+++	+	???	-

HTS and medical genetics

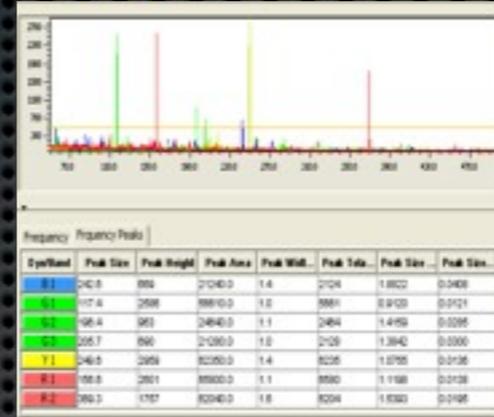
Methods for identifying variants/aberrations



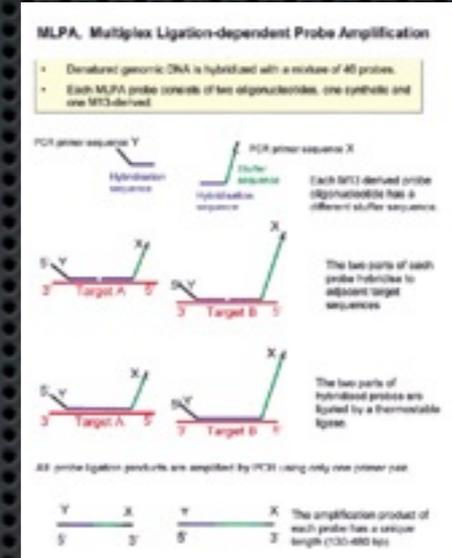
Southern blotting



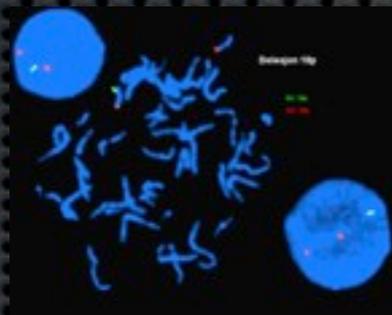
DNA (Sanger) sequencing



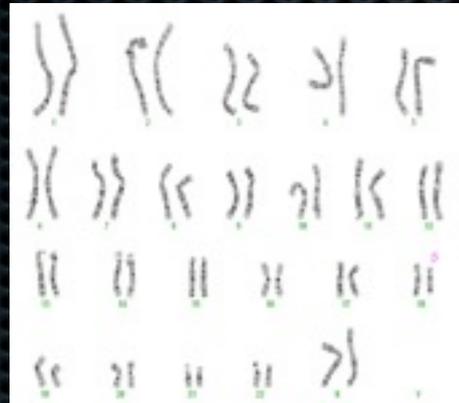
Fragment analysis



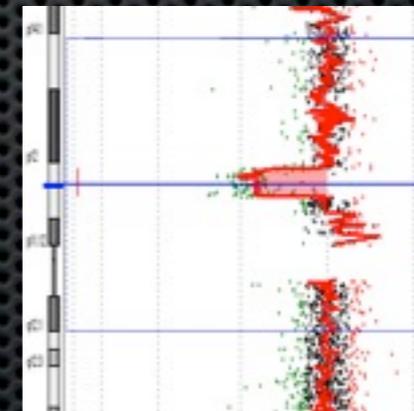
MLPA



FISH



Karyotyping



Array CGH



SNP array

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

How soon will high-throughput sequencing replace these techniques?

Variation	aCGH	SNP array	HTS
SNP	-	-	+
indel	-	-	+
CNV	+	+	+
Non-balanced chromosomal aberrations	+	+	+
Balanced chromosomal aberrations	-	-	+
UPD	-	+	+
Regions of homozygosity	-	+	+
Trinucleotide repeats	-	-	+?

Resequencing - 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

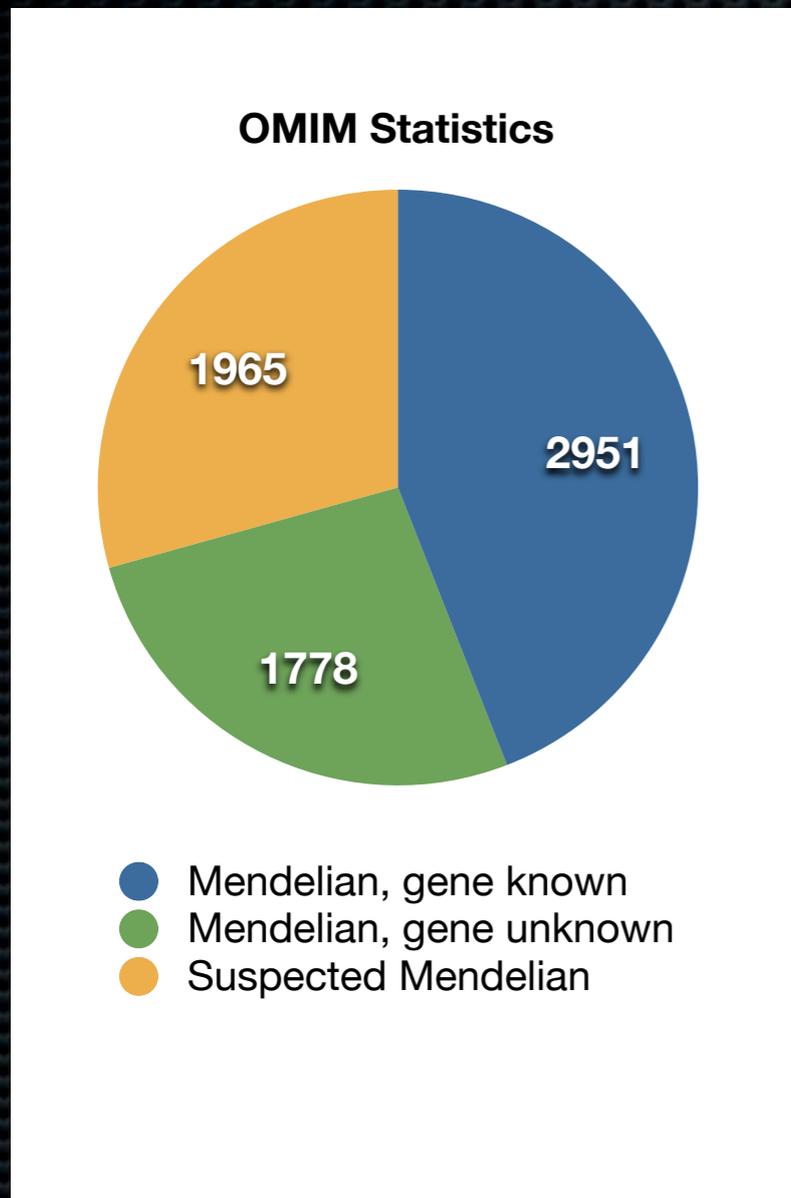
Sequence

Mutation

Resequencing

- ✦ Compare test sequence to a reference sequence
 - ✦ Mendelian (linkage)
 - ✦ Association studies
 - ✦ Exome sequencing
- ✦ Identify genetic variation
 - ✦ Single-nucleotide polymorphisms (SNPs)
 - ✦ Insertions/deletions
 - ✦ Copy-number variation (CNVs)

Mendelian disease in man



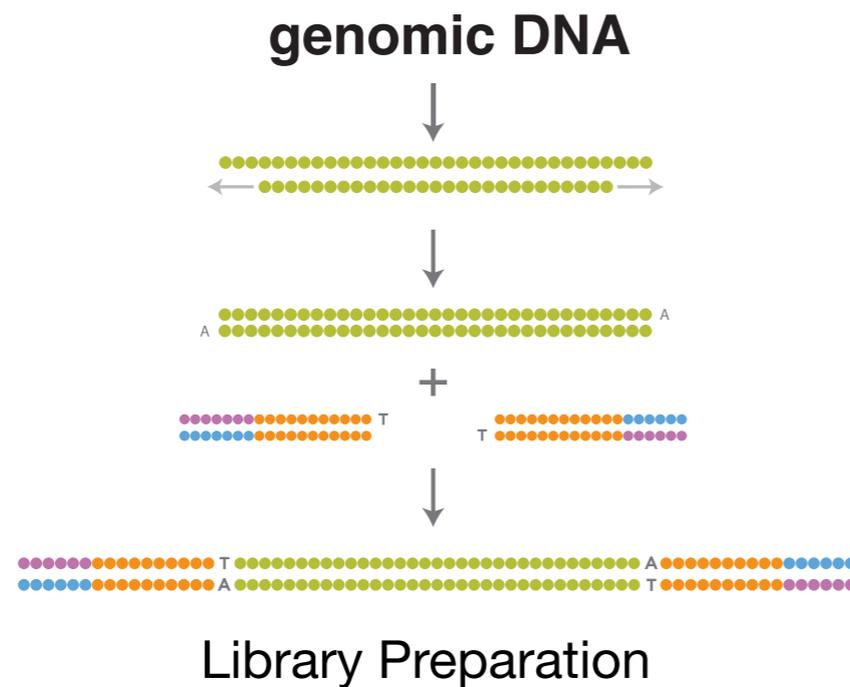
- 2951 of the well-characterized phenotypes registered in OMIM have a known molecular basis
- 3743 registered phenotypes with known or suspected Mendelian basis, no associated gene has been identified
- Protein coding regions of the human genome (the exome) constitute approximately 1.5% of the total,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

Exome sequencing

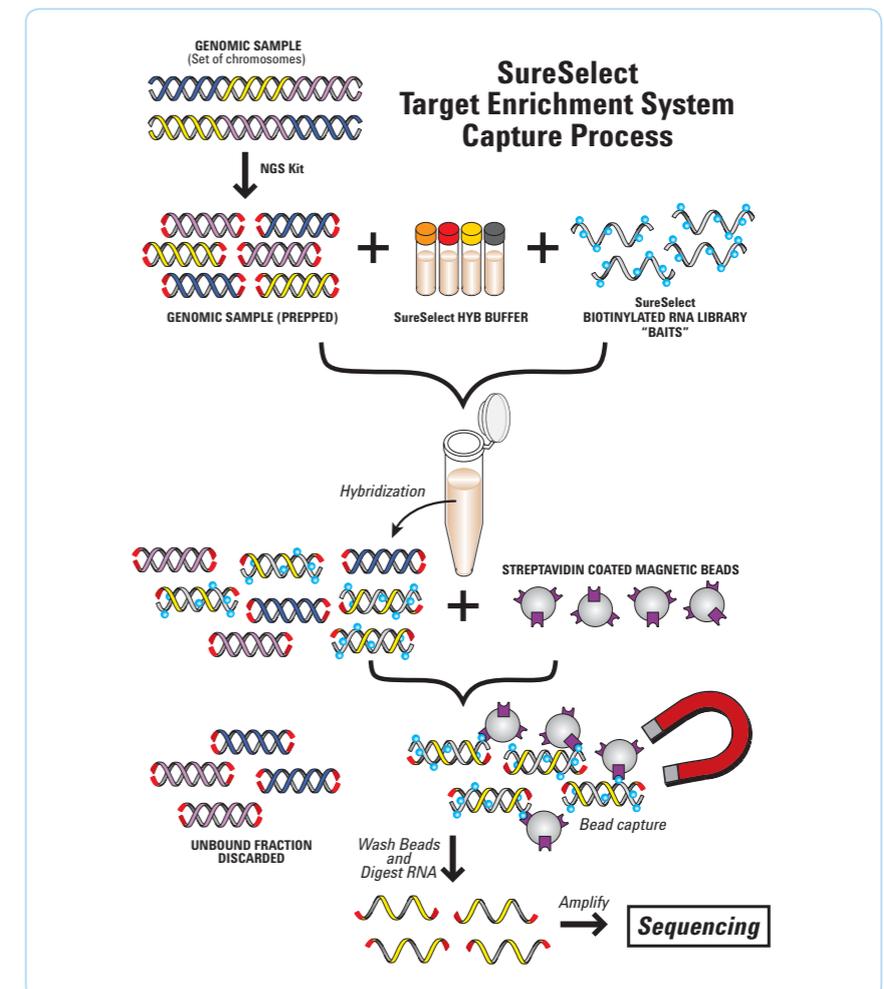
4 easy steps

1. Library preparation

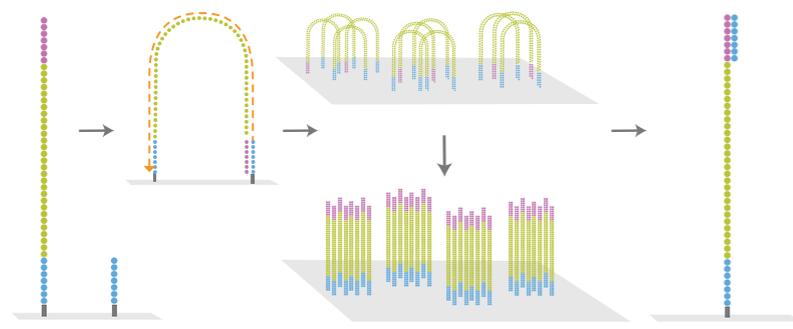
fragment gDNA
fill-in
add adapters
sequencing library



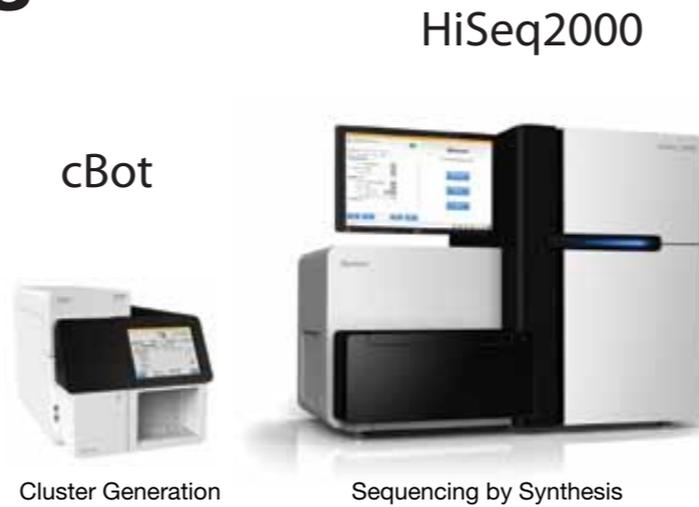
2. Sequence capture



3. Illumina sequencing

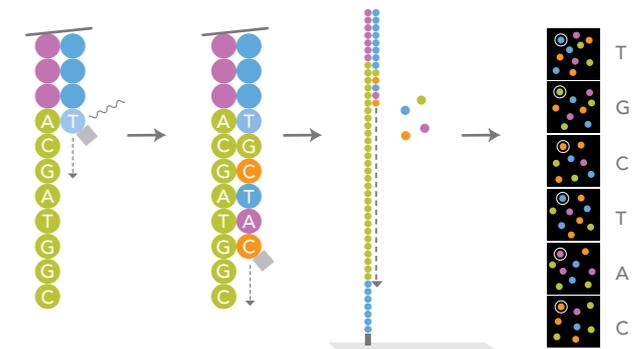


Cluster sequence library



Cluster Generation

Sequencing by Synthesis



Sequencing by synthesis

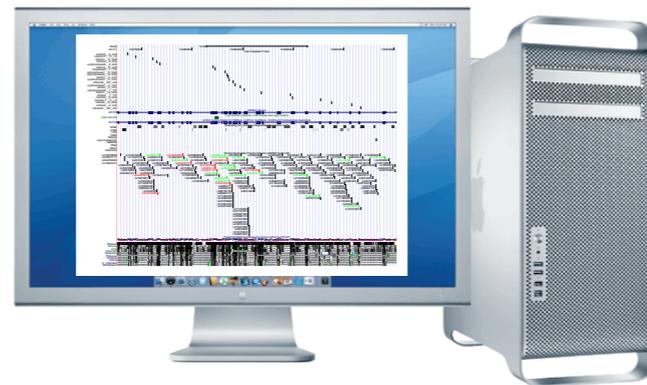
4. Analysis

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/

Bioinformatics solutions

Software/list - SEQwiki

http://seqanswers.com/wiki/Software/list

Below is (one of many possible) dynamic tables of software data, created from pages in the wiki. To add a package to the list, use the following form:

new package name

Name	Description	Bio Tags	Tags	Features	Language	Licence	OS
AB Large Indel Tool	Identifies deviations in clone insert size that indicate intra-chromosomal structural variations compared to a reference genome.	InDel discovery	Copy number			GPL	
AB Small Indel Tool	The SOLiD™ Small Indel Tool processes the indel evidences found in the pairing step of the SOLiD™ System Analysis Pipeline Tool (Corona Lite).	InDel discovery				GPL	
ABBA	Assembly Boosted By Amino acid sequence is a comparative gene assembler, which uses amino acid sequences from predicted proteins to help build a better assembly		Assembly				
ABYSS	ABYSS is a de novo sequence assembler that is designed for short reads.	De-novo assembly	Assembly	MPI		Free to academics	POSIX
ALEXA-Seq	Alternative Expression Analysis by massively parallel RNA sequencing	RNA-Seq Quantitation				GPLv3	
ALLPATHS	De novo assembly of whole-genome shotgun microreads.	De-novo assembly					
Alta-Cyclic	Alta-Cyclic is a Illumina Genome-Analyzer (Solexa) base caller.		Basecaller				
AMOS	AMOS is a de novo sequence assembler.		Assembly				Linux
ANNOVAR	ANNOVAR is a software tool for identifying genetic variants from high-throughput sequencing data	Genomics Genetics	Annotation	Variant Prioritization	annotation	free to academia	Linux Windows MacOS
ArrayStar	ArrayStar is an easy-to-use gene expression analysis software package that offers powerful visualization and statistical tools to help you analyze your microarray data.	Gene Expression Analysis	Differential expression	gene ontology analysis	Statistics	Available as a standalone system a network license	Windows 7 Windows Vista Windows XP SP2 Mac OS X 10.6 with Parallels Desktop
ATAC	ATAC is a computational procedure for comparative mapping between two genome assemblies, or between two different genomes.						
Atlas-SNP2	Atlas-SNP2 is a SNP discovery tool developed for next generation sequencing platforms	SNP discovery				No redistribution but otherwise free	Unix

Alignment

Variant calling

Filtering

Viewing

NovoAlign

BWA

Bowtie

Tophat

MAQ

SAMtools

GATK

Annovar

Perl

shell

UCSC (bed, gff, wig)

IGV

<http://seqanswers.com/wiki/Software/list>

Resequencing - 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

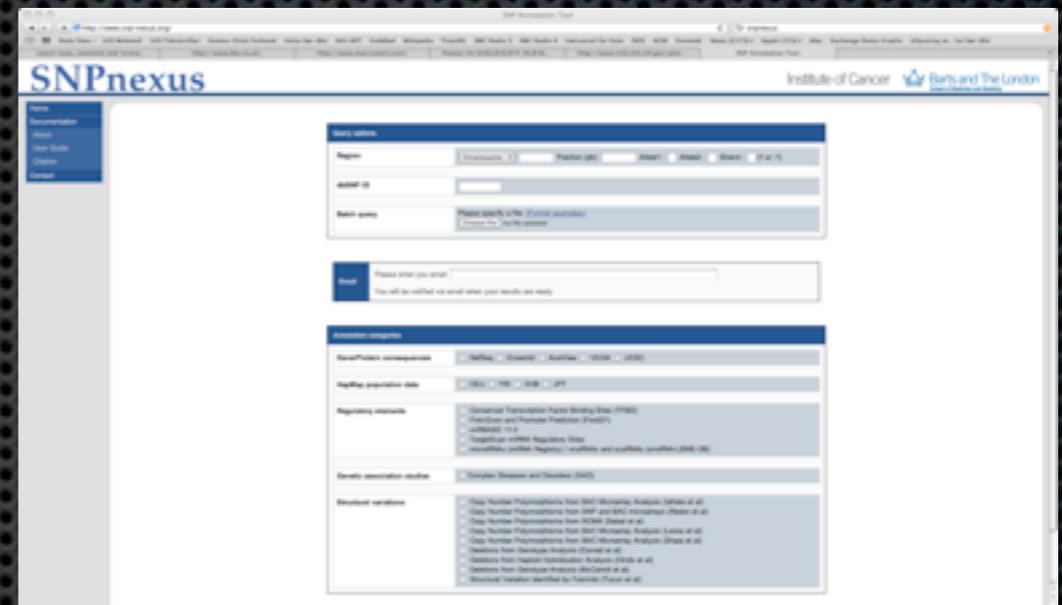
Compare to
reference

Sequence

Mutation

ExomeSeq - finding mutations

- ❖ ~ **20 000 variants will be found**
- ❖ Which variants are deleterious?
- ❖ Novel? (dbSNP, 1000genomes, HGMD)
- ❖ Synonymous/non-synonymous?
- ❖ Conserved?
- ❖ Alter protein structure?

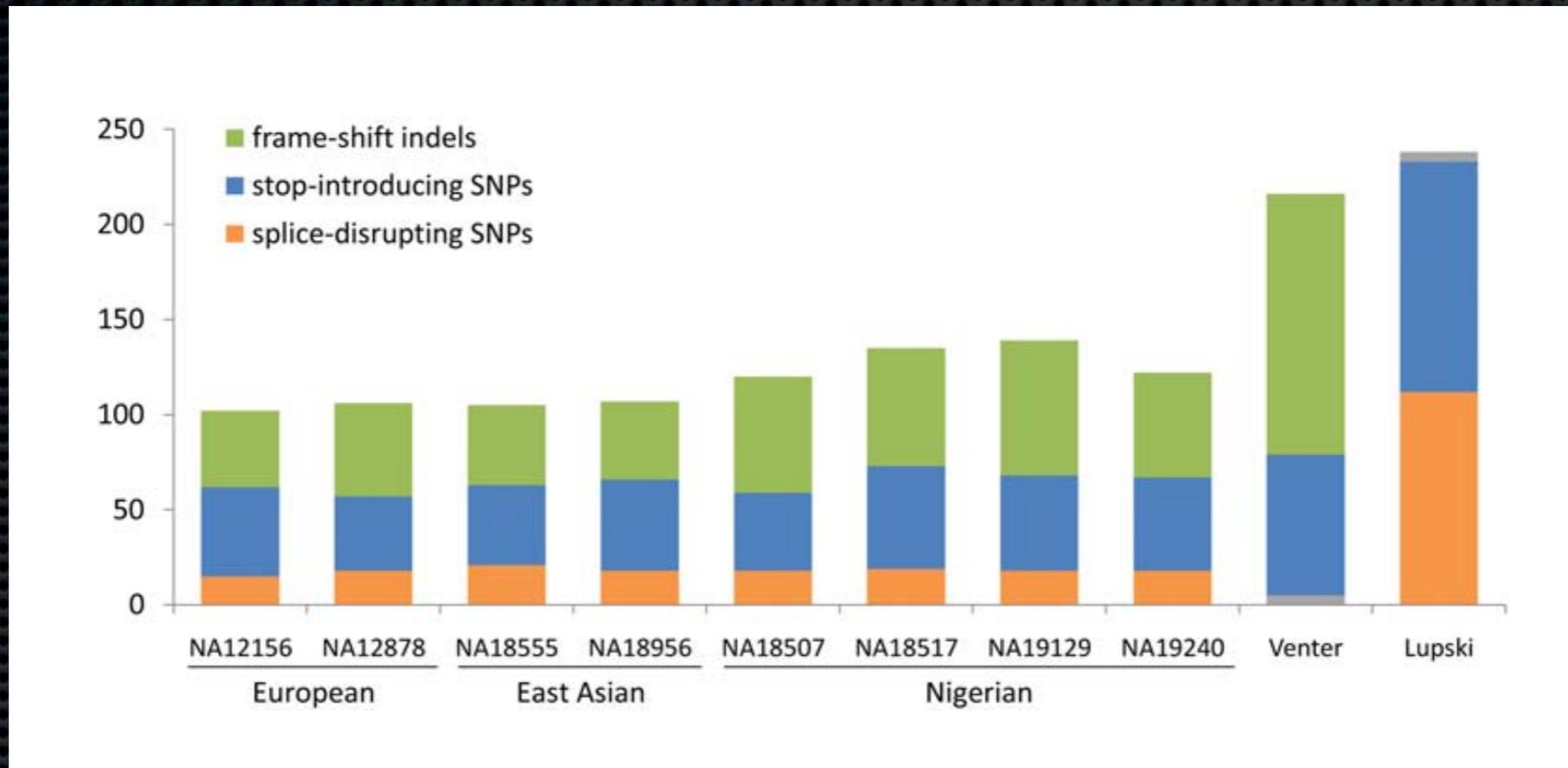


SNPnexus
PolyPhen2
MutationTaster
ANNOVAR
SeattleSeq Annotation

This is the hard part

What's in an exome?

> 20 000 variants



Many loss-of-function variants

1000 genomes data - everyone has 200-250 nonsense variants

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		
	1	2	3	Any 1	Any 2	Any 1	Any 2	Any 3
Dominant model								
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 ~22 000 SNPs
- ✦ Which is the causal variant?
- ✦ In a family, compare more exomes

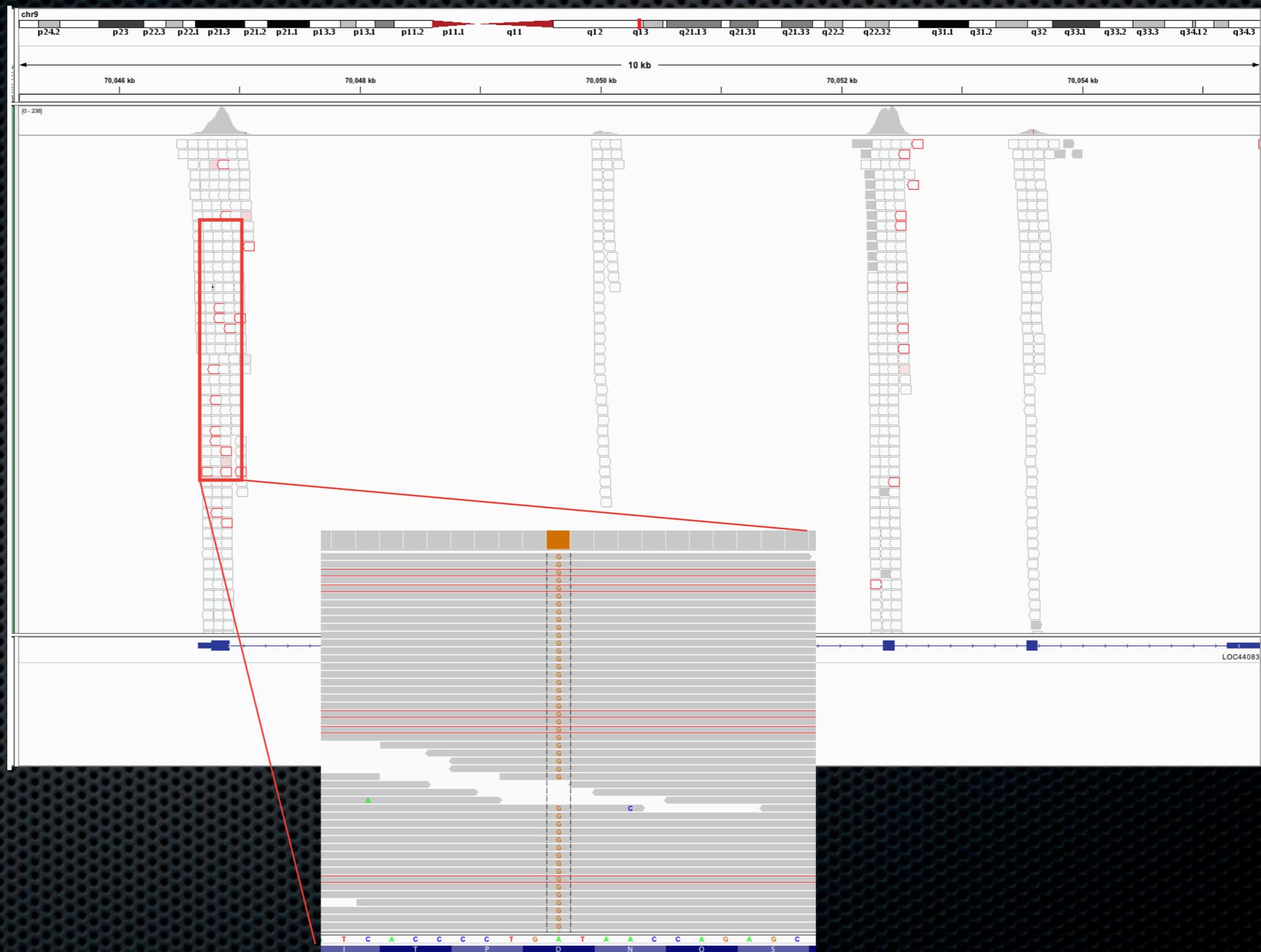
Checking variants - IGV

Chromosome

Coverage

Mapped reads

Gene



Disease genes found by ExomeSeq

Disorder	Inheritance	Gene identified	Scope	References
Congenital chloride diarrhea	Recessive	<i>SLC26A3</i>	Exome	Choi <i>et al.</i> [16]
Miller syndrome	Recessive	<i>DHODH</i>	Exome	Ng <i>et al.</i> [14]
Charcot-Marie-Tooth neuropathy	Recessive	<i>SH3TC2</i>	Genome	Lupski <i>et al.</i> [20]
Metachondromatosis	Dominant	<i>PTPN11</i>	Genome	Sobreira <i>et al.</i> [23]
Schinzel-Giedion syndrome	Dominant	<i>SETBP1</i>	Exome	Hoischen <i>et al.</i> [29]
Nonsyndromic hearing loss	Recessive	<i>GPSM2</i>	Exome	Walsh <i>et al.</i> [69]
Perrault syndrome	Recessive	<i>HSD17B4</i>	Exome	Pierce <i>et al.</i> [25]
Hyperphosphatasia mental retardation syndrome	Recessive	<i>PIGV</i>	Exome	Krawitz <i>et al.</i> [68]
Sensenbrenner syndrome	Recessive	<i>WDR35</i>	Exome	Gilissen <i>et al.</i> [26]
Cerebral cortical malformations	Recessive	<i>WDR62</i>	Exome	Bilguvar <i>et al.</i> [70]
Kaposi sarcoma	Recessive	<i>STIM1</i>	Exome	Byun <i>et al.</i> [71]
Spinocerebellar ataxia	Dominant	<i>TGM6</i>	Exome	Wang <i>et al.</i> [72]
Combined hypolipidemia	Recessive	<i>ANGPTL3</i>	Exome	Musunuru <i>et al.</i> [40]
Complex I deficiency	Recessive	<i>ACAD9</i>	Exome	Haack <i>et al.</i> [52]
Autoimmune lymphoproliferative syndrome	Recessive	<i>FADD</i>	Exome	Bolze <i>et al.</i> [73]
Amyotrophic lateral sclerosis	Dominant	<i>VCP</i>	Exome	Johnson <i>et al.</i> [74]
Nonsyndromic mental retardation	Dominant	Various	Exome	Vissers <i>et al.</i> [31]
Kabuki syndrome	Dominant	<i>MLL2</i>	Exome	Ng <i>et al.</i> [30]
Inflammatory bowel disease	Dominant	<i>XIAP</i>	Exome	Worthey <i>et al.</i> [18]
Nonsyndromic mental retardation	Recessive	<i>TECR</i>	Exome	Caliskan <i>et al.</i> [75]
Retinitis pigmentosa	Recessive	<i>DHDDS</i>	Exome	Züchner <i>et al.</i> [56]
Osteogenesis imperfecta	Recessive	<i>SERPINF1</i>	Exome	Becker <i>et al.</i> [53]
Dilated cardiomyopathy	Dominant	<i>BAG3</i>	Exome	Norton <i>et al.</i> [24]
Hajdu-Cheney syndrome	Dominant	<i>NOTCH2</i>	Exome	Simpson <i>et al.</i> [76]
Hajdu-Cheney syndrome	Dominant	<i>NOTCH2</i>	Exome	Isidor <i>et al.</i> [77]
Skeletal dysplasia	Recessive	<i>POP1</i>	Exome	Glazov <i>et al.</i> [78]
Amelogenesis	Recessive	<i>FAM20A</i>	Exome	O'Sullivan <i>et al.</i> [80]
Chondrodysplasia and abnormal joint development	Recessive	<i>IMPAD1</i>	Exome	Vissers <i>et al.</i> [80]
Progeroid syndrome	Recessive	<i>BANF1</i>	Exome	Puente <i>et al.</i> [81]
Infantile mitochondrial cardiomyopathy	Recessive	<i>AARS2</i>	Exome	Götz <i>et al.</i> [82]
Sensory neuropathy with dementia and hearing loss	Dominant	<i>DNMT1</i>	Exome	Klein <i>et al.</i> [49]
Autism	Dominant	Various	Exome	O'Roak <i>et al.</i> [32]

Genetic diagnosis

2009	2011
Diagnosis	Diagnosis
Test series of single genes	Test exome
Often international labs	In house
Very expensive	Cheap (and getting cheaper)
Time-consuming (years)	Fast (weeks)

A diagnostic revolution
(common/rare)

Summary

- ✦ High-throughput sequencing
 - ✦ Dramatic increase in sequence production
 - ✦ Many new technologies
 - ✦ Field new and moving very quickly
 - ✦ Many applications on one platform
 - ✦ Huge impact on human/medical genetics
- ✦ Bioinformatics challenges/opportunities
 - ✦ Data storage/backup/distribution
 - ✦ Data analysis