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



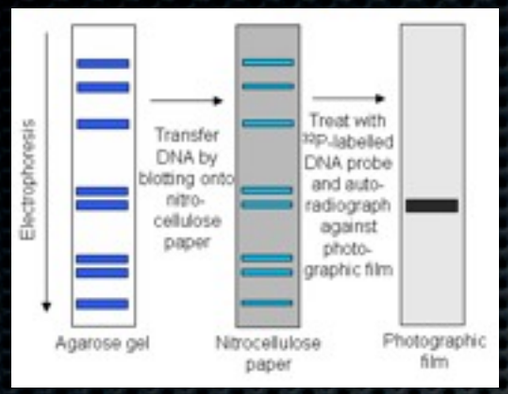


Platform	454	Illumina HiSeq	Illumina MiSeq*	PacBio*	Ion Torrent*
Resequencing	-	+++	++	+	+++
de novo	+++	+	+	+++	+++
metagenomics	+++	++	+	++	+++
mRNA	++	+++	++	++	++
miRNA	-	+++	+++	-	-
ChIP	-	+++	++	-	-
DNA meth	-	+++	+	???	-

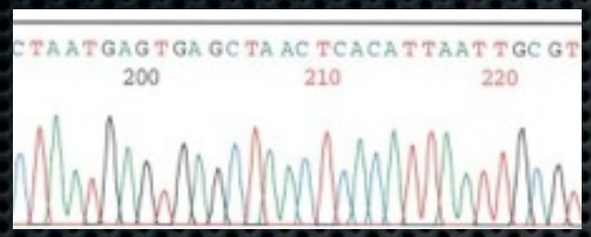
*Q4 2011

HTS and medical genetics

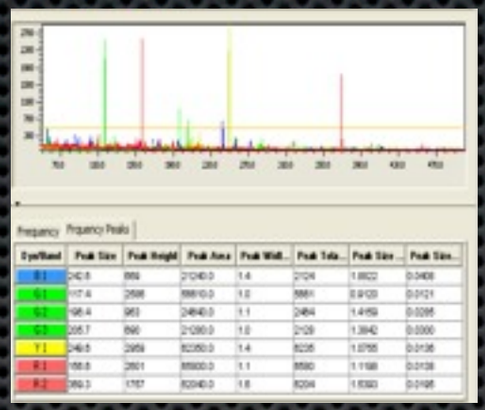
Methods for identifying variants/aberrations



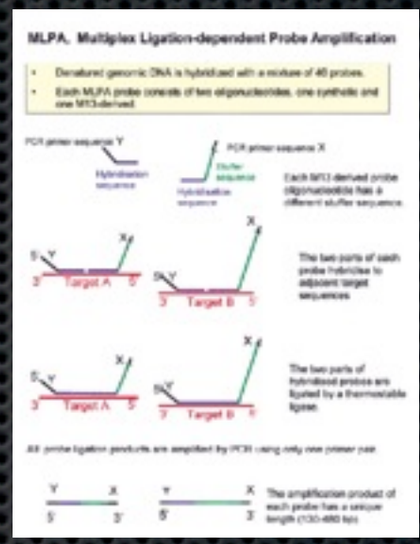
Southern blotting



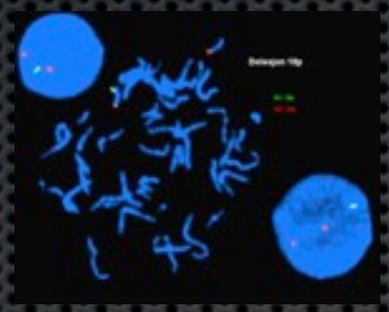
DNA (Sanger) sequencing



Fragment analysis



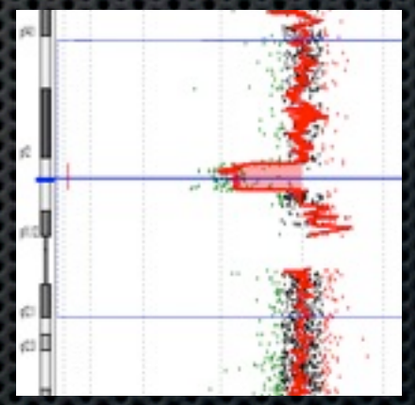
MLPA



FISH

Karyotyping

SNP array



Array CGH

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

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

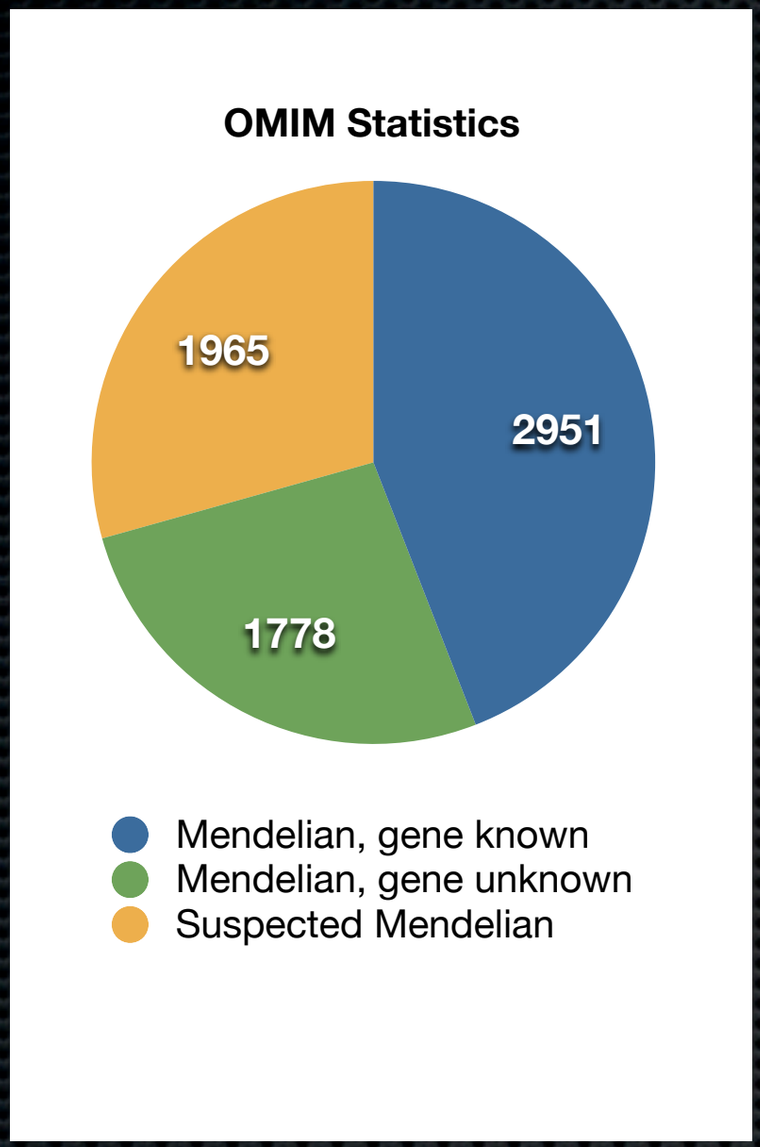
Sequence

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

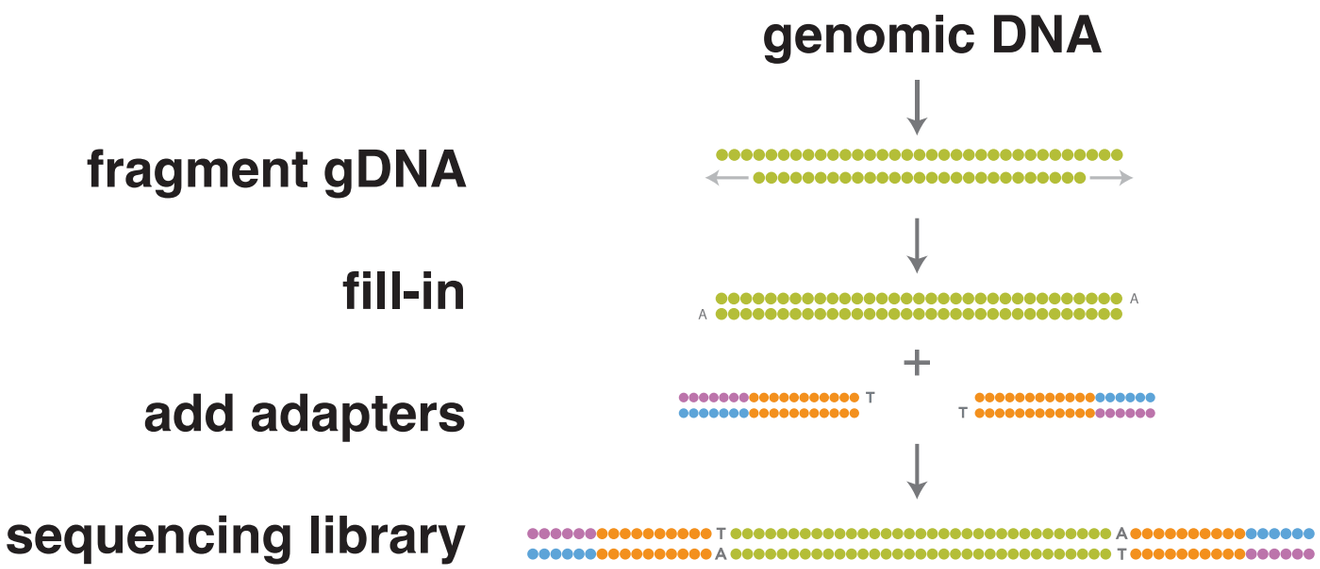


- 2951 of the well-characterized phenotypes registered in OMIM have a known molecular basis
- 3743 registered phenotypes with known or suspected Mendelian basis, no associated gene has been identified
- Protein coding regions of the human genome (the exome) constitute approximately 1.5% of the total, 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



Library Preparation

2. Sequence capture

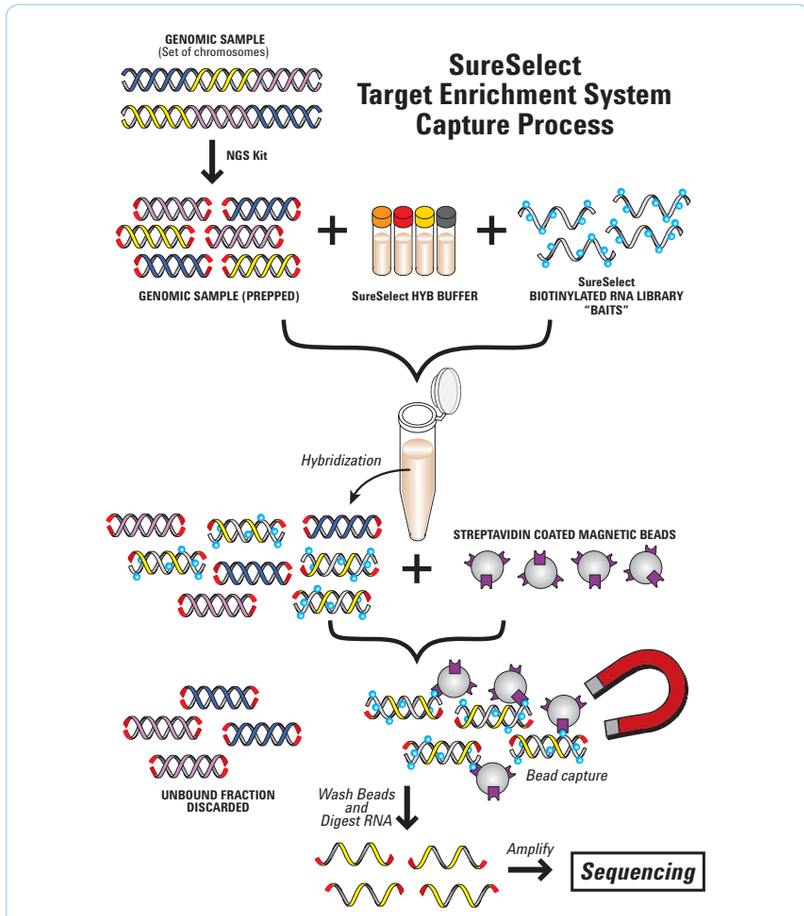
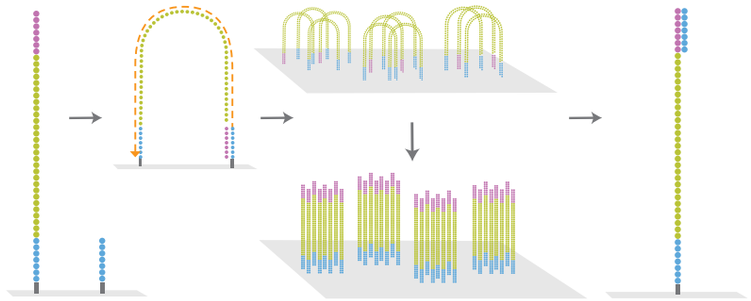


Figure 1. SureSelect Target Enrichment System Workflow

3. Illumina sequencing



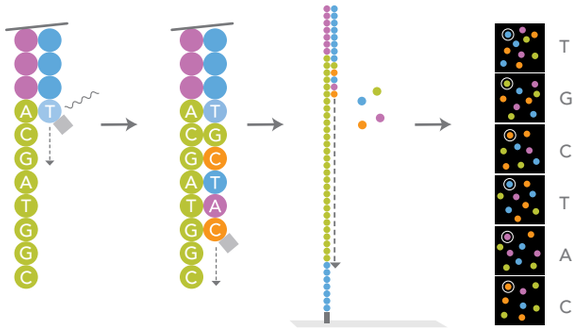
Cluster sequence library

HiSeq2000

cBot

Cluster Generation

Sequencing by Synthesis



Sequencing by synthesis

4. Analysis

Align reads to reference genome

▼

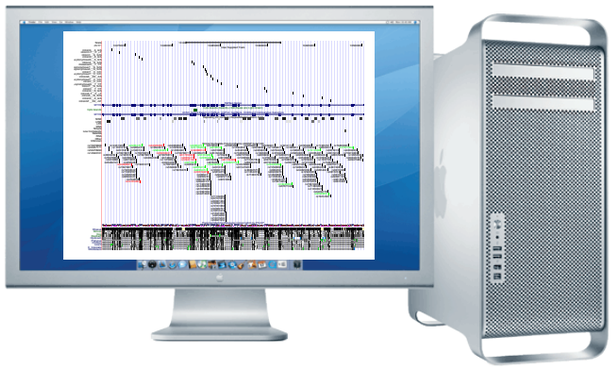
Call variants

▼

Filter variants

▼

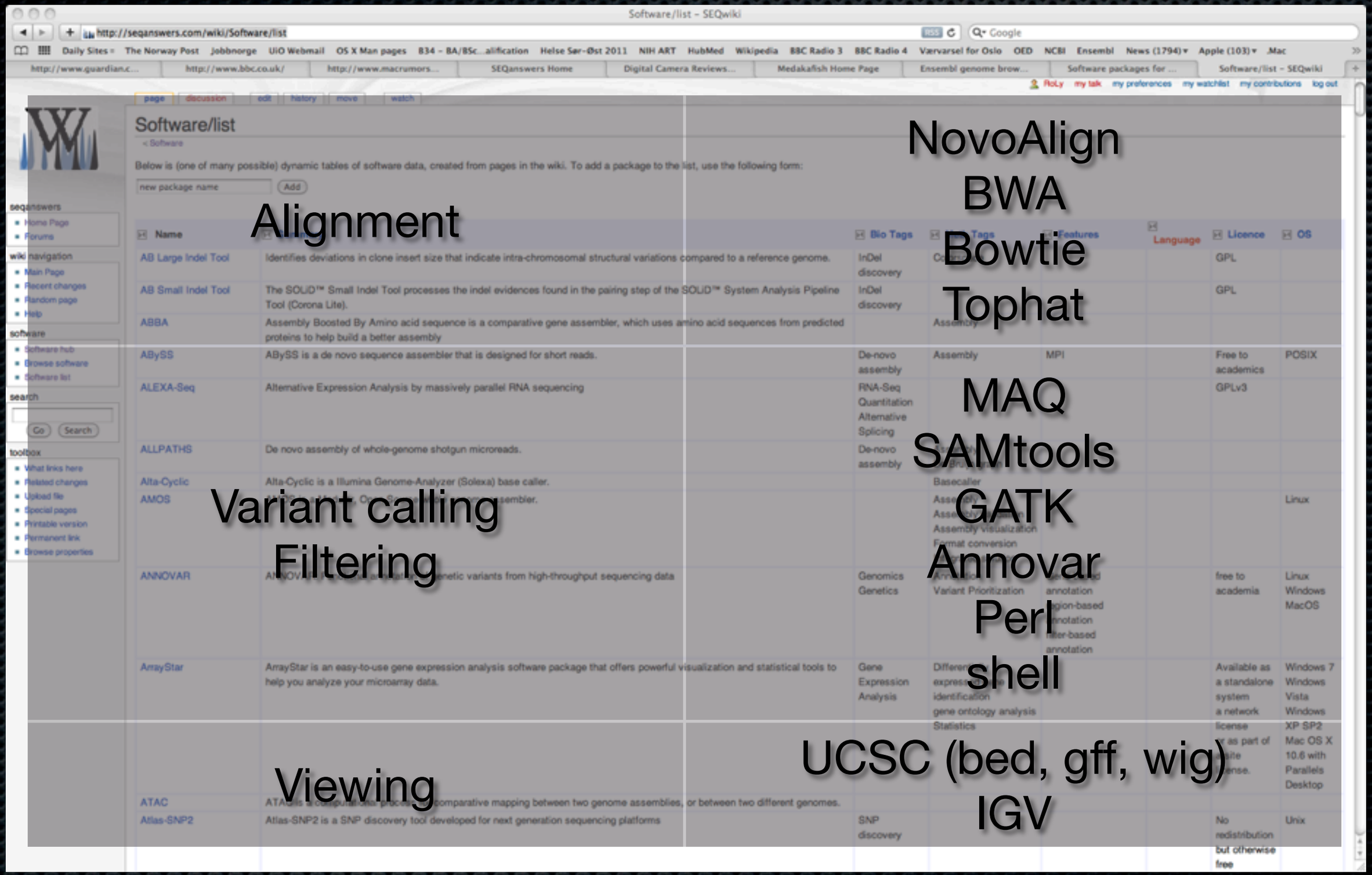
View



Software

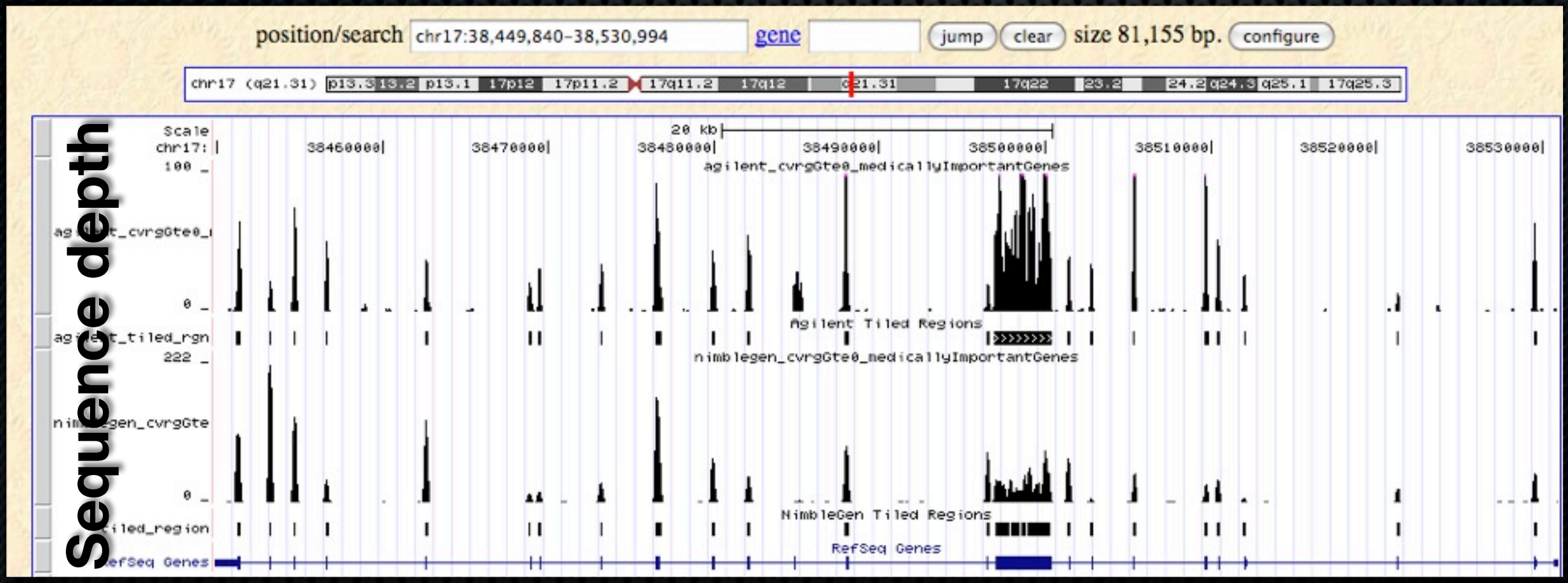
Step	Software	Link
QC/preprocessing	FastQC	http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/
	FASTX-Toolkit	http://hannonlab.cshl.edu/fastx_toolkit/
Aligning	Novoalign	http://www.novocraft.com
	BWA	http://bio-bwa.sourceforge.net/
Variant calling	Samtools	http://samtools.sourceforge.net/
	VCFtools	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/

Bioinformatics solutions



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

Exome sequencing data



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

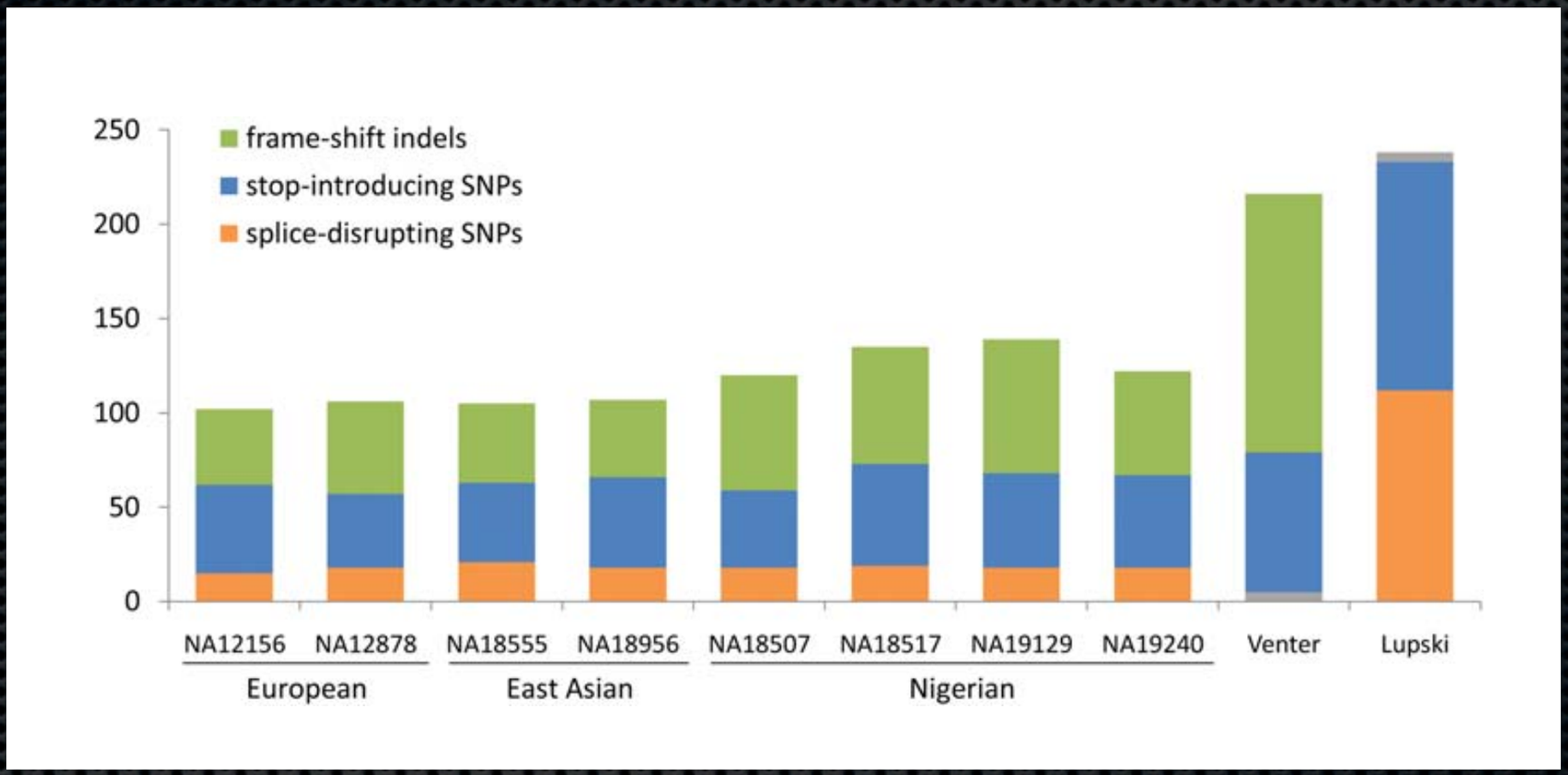
- **~ 20 000 variants will be found**
- Which variants are deleterious?
- Novel? (dbSNP, 1000genomes, HGMD)
- Synonymous/non-synonymous?
- Conserved?
- Alter protein structure?

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

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

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 ~22 000 SNPs
- Which is the causal variant?
- In a family, compare more exomes

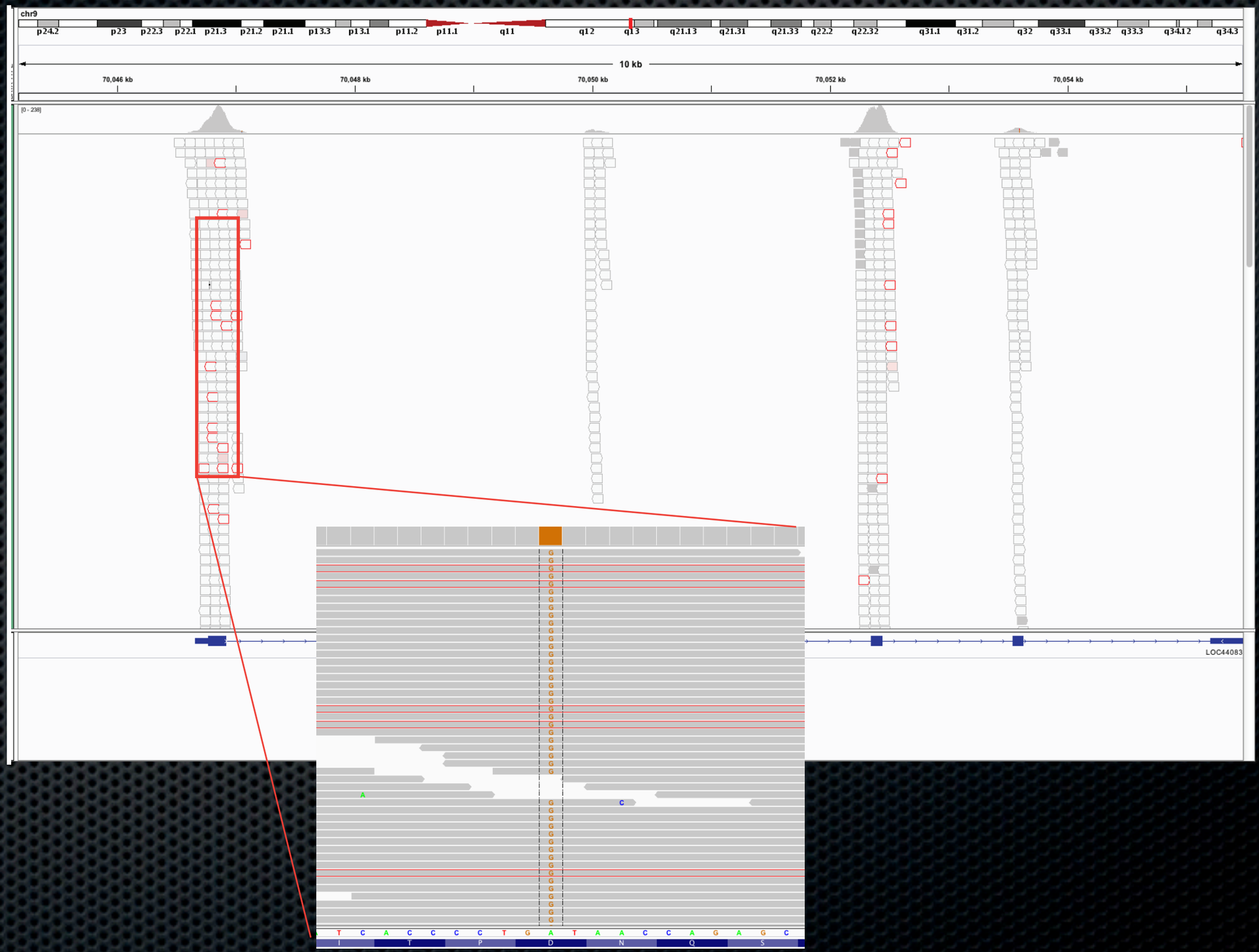
Checking variants - IGV

Chromosome

Coverage

Mapped reads

Gene



Disease genes found by ExomeSeq

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

Gilissen et al. Genome Biology (October)

Genetic diagnosis

2009	2011
Diagnosis	Diagnosis
Test series of single genes	Test exome
Often international labs	In house
Very expensibe	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