High-throughput sequencing

Robert Lyle

Department of Medical Genetics

Oslo University Hospital

Robert.Lyle@medisin.uio.no

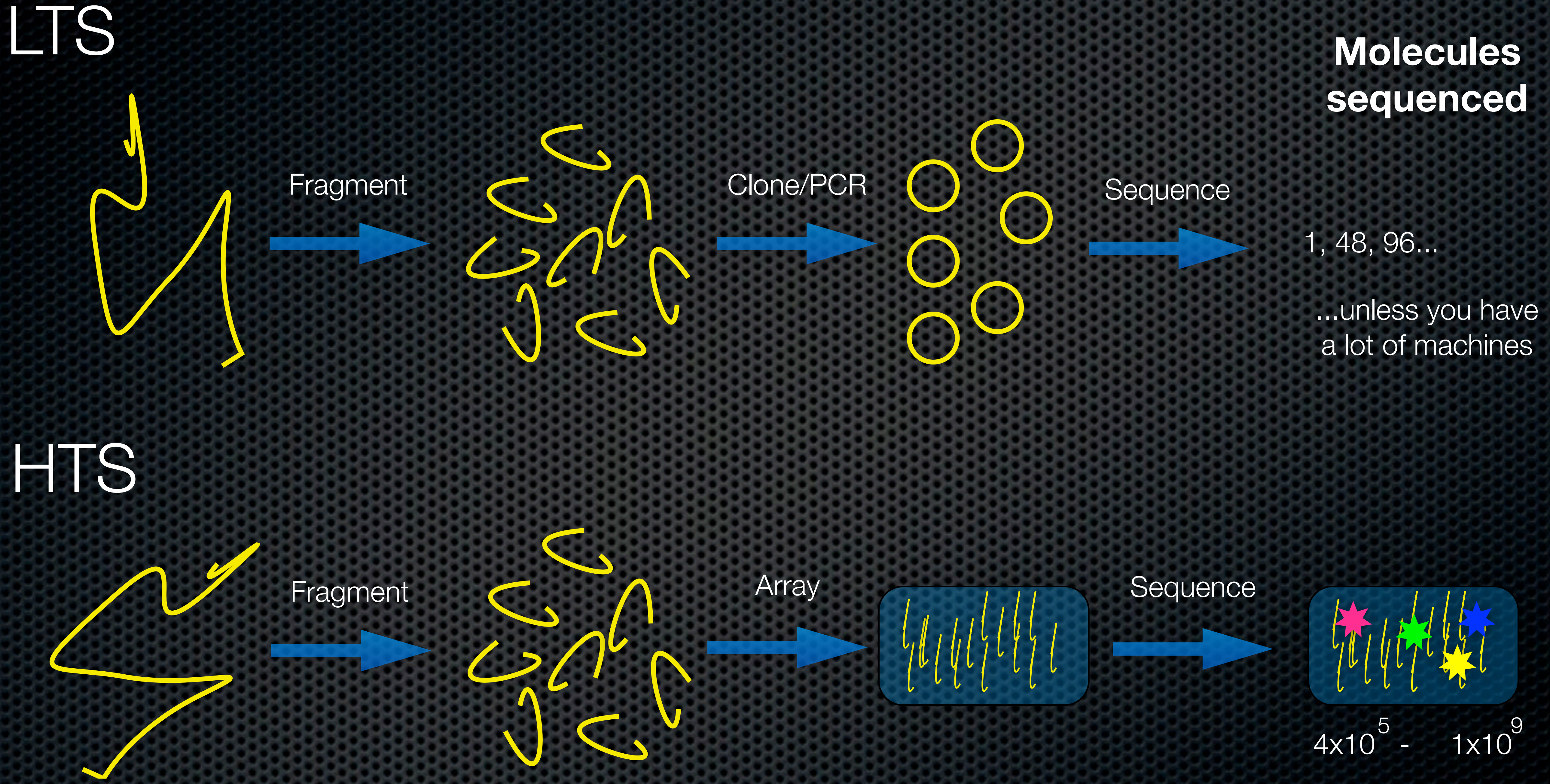
Overview

- Technology
- Data and analysis
- Applications

Technology

Sequencing past, present and future

Sequencing: old and next



LTS

Fragment

Clone/PCR

Sequence

Molecules sequenced

1, 48, 96...

...unless you have
a lot of machines

HTS

Fragment

Array

Sequence

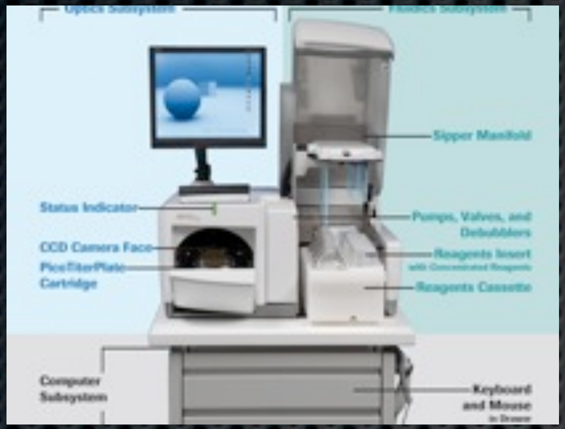
$4x10^5$ - $1x10^9$

...on one machine

Massively parallel

HTS systems available

454



Roche

Solexa

Illumina

SOLiD

ABI

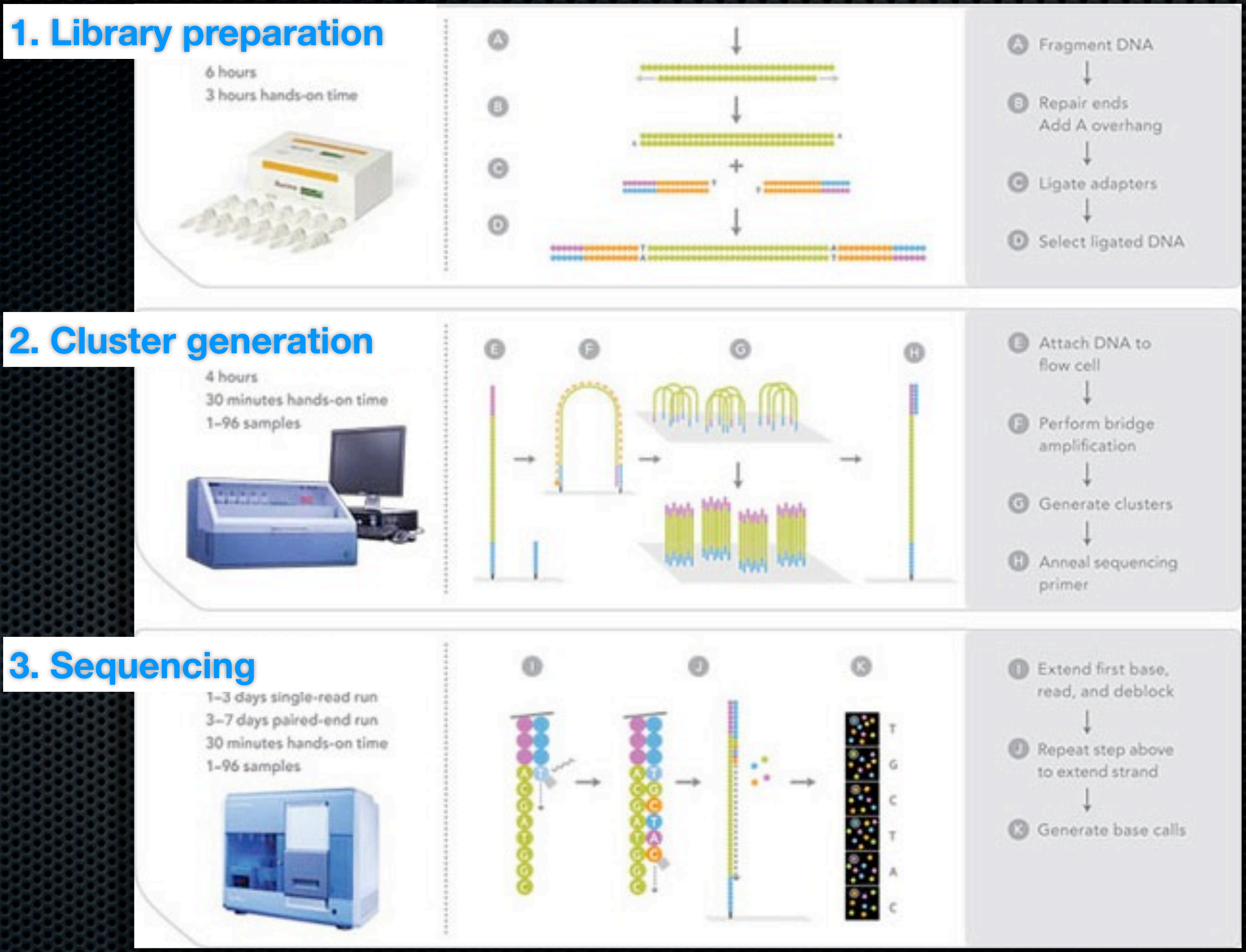
HeliScope

Helicos

Others in 2011

(Pacific BioSciences, Ion Torrent)

Illumina sequencing technology



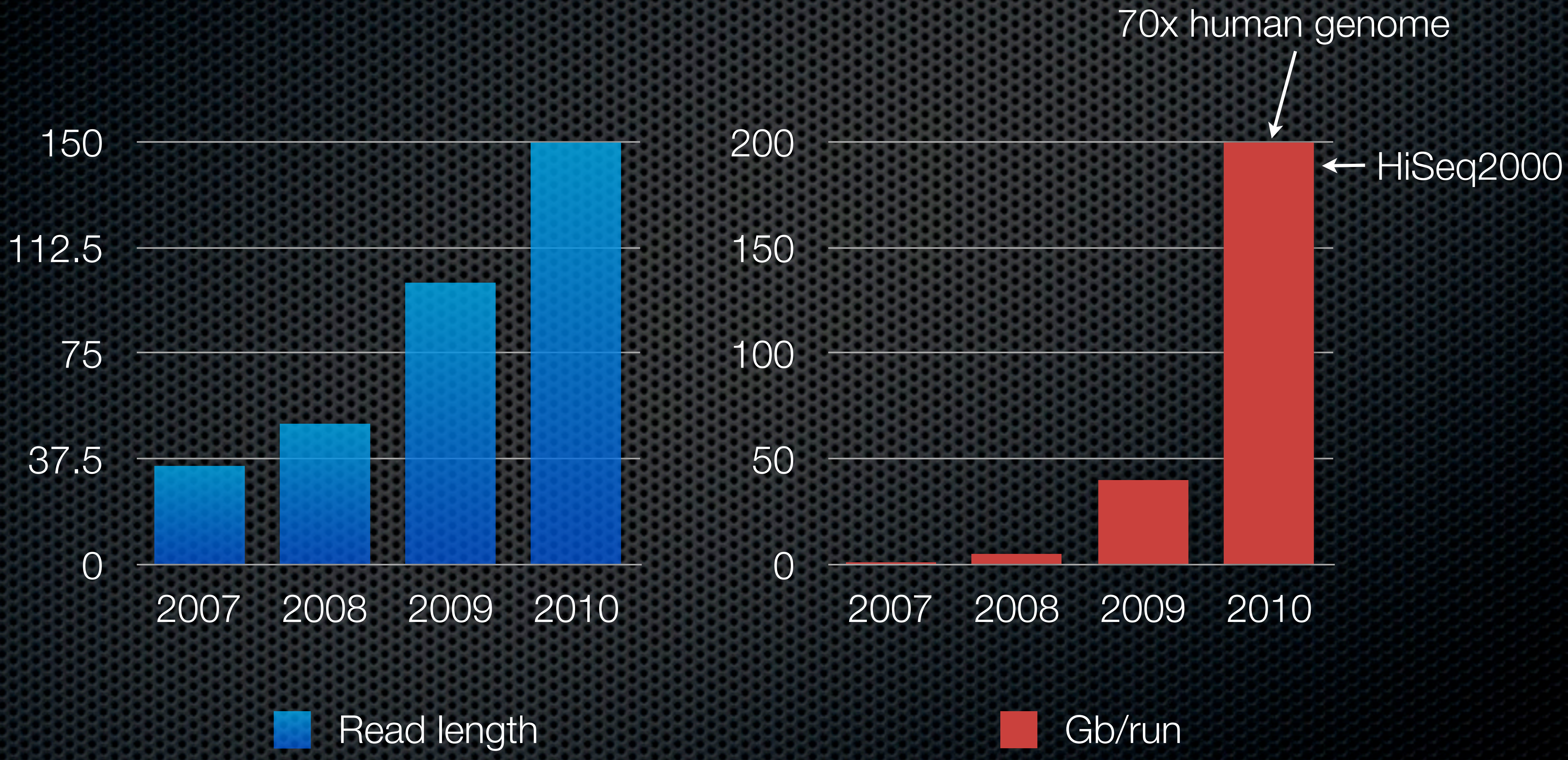
1. Library preparation

2. Cluster generation

4 hours
30 minutes hands-on time
1–96 samples
E
F
G
H
E Attach DNA to flow cell
F Perform bridge amplification
G Generate clusters
H Anneal sequencing primer

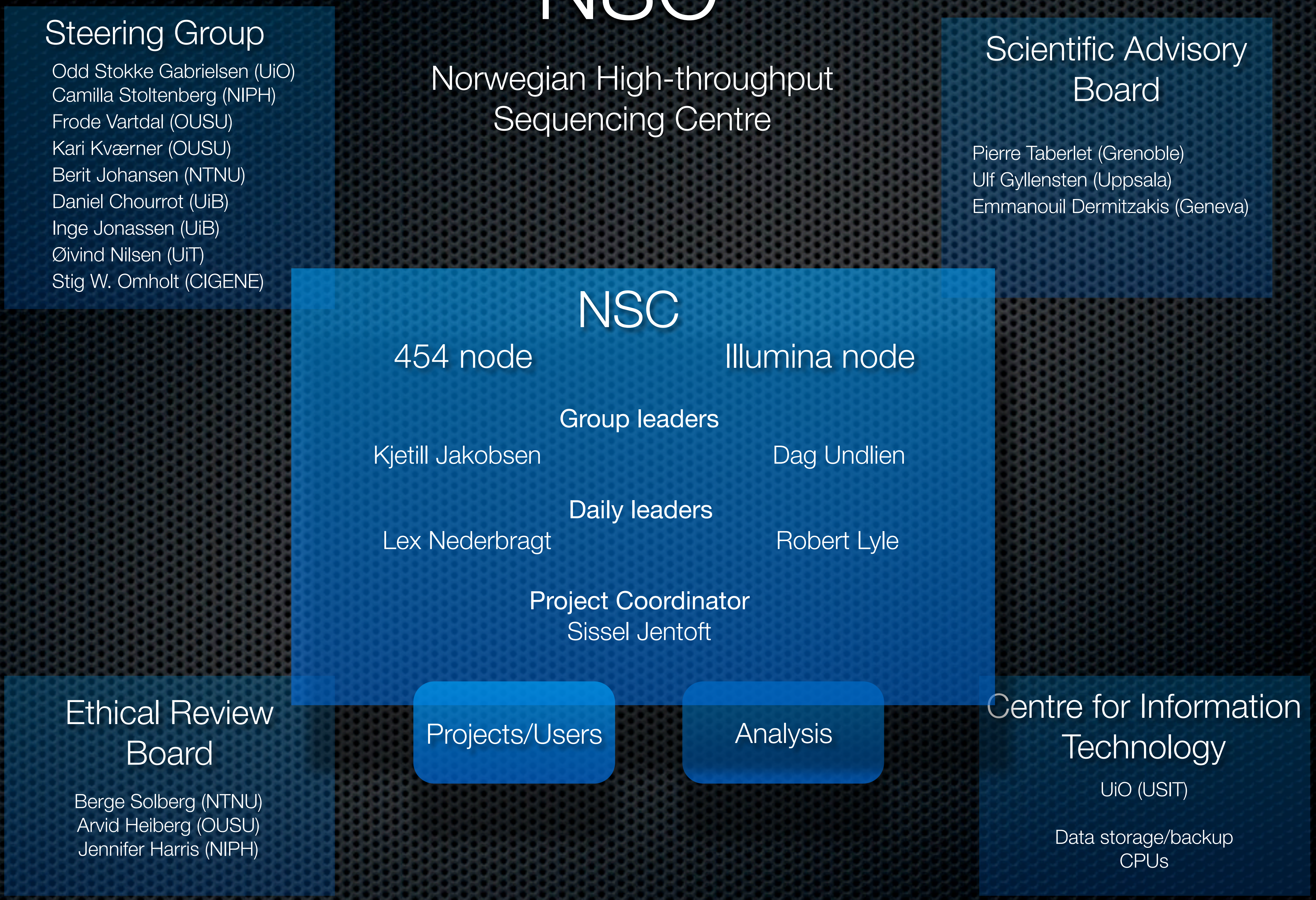
3. Sequencing

Illumina throughput



NSC

Norwegian High-throughput Sequencing Centre



Steering Group

Odd Stokke Gabrielsen (UiO)
Camilla Stoltenberg (NIPH)
Frode Vartdal (OUSU)
Kari Kværner (OUSU)
Berit Johansen (NTNU)
Daniel Chourrot (UiB)
Inge Jonassen (UiB)
Øivind Nilsen (UiT)
Stig W. Omholt (CIGENE)

Scientific Advisory Board

Pierre Taberlet (Grenoble)
Ulf Gyllensten (Uppsala)
Emmanouil Dermitzakis (Geneva)

NSC

454 node	Illumina node

Group leaders

Kjetill Jakobsen	Dag Undlien

Daily leaders

Lex Nederbragt	Robert Lyle

Project Coordinator
Sissel Jentoft

Projects/Users

Analysis

Ethical Review Board

Berge Solberg (NTNU)
Arvid Heiberg (OUSU)
Jennifer Harris (NIPH)

Centre for Information Technology

UiO (USIT)

Data storage/backup
CPUs

User contact

http://www.sequencing.uio.no

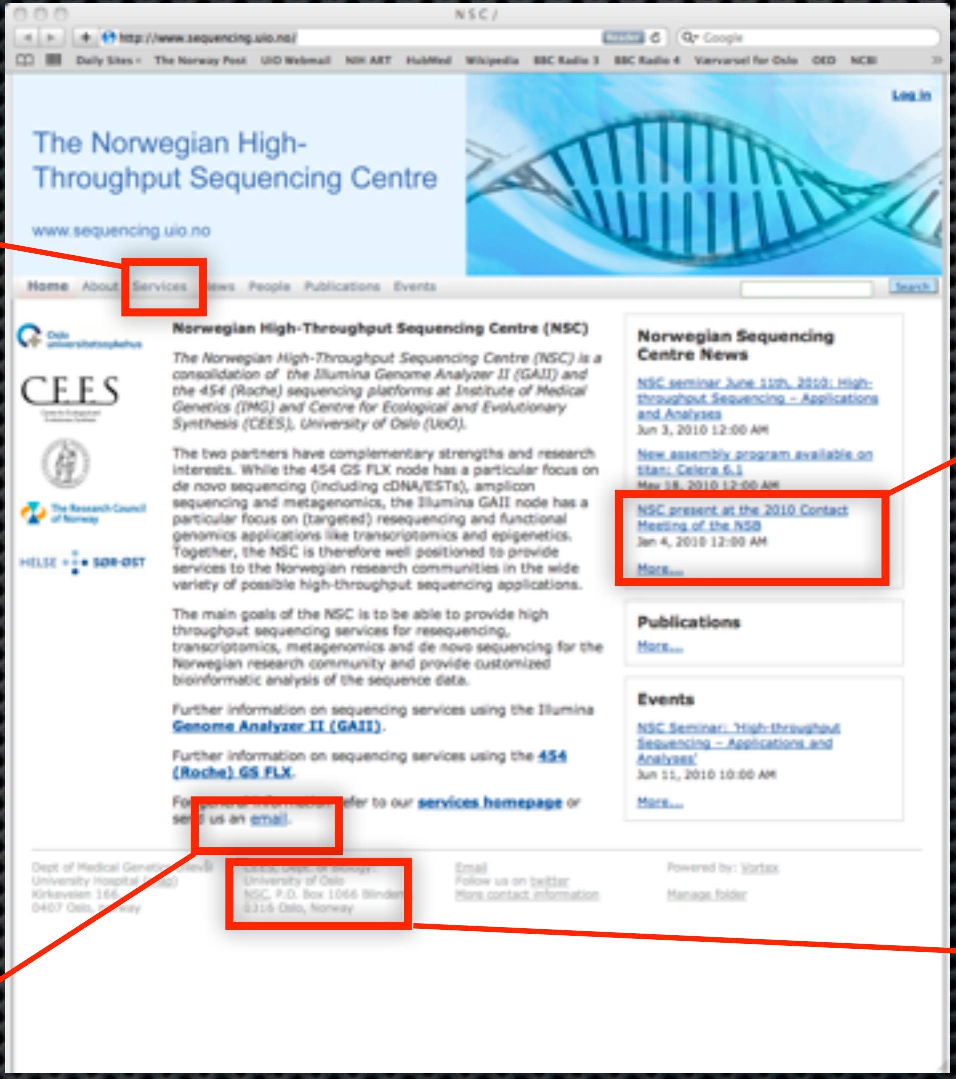
Services

454

Illumina

Sample delivery form
FAQ

post@sequencing.uio.no



The Norwegian High-Throughput Sequencing Centre

www.sequencing.uio.no

Norwegian High-Throughput Sequencing Centre (NSC)

The Norwegian High-Throughput Sequencing Centre (NSC) is a consolidation of the Illumina Genome Analyzer II (GAII) and the 454 (Roche) sequencing platforms at Institute of Medical Genetics (IMG) and Centre for Ecological and Evolutionary Synthesis (CEES), University of Oslo (UoO).

The two partners have complementary strengths and research interests. While the 454 GS FLX node has a particular focus on de novo sequencing (including cDNA/ESTs), amplicon sequencing and metagenomics, the Illumina GAII node has a particular focus on (targeted) resequencing and functional genomics applications like transcriptomics and epigenetics. Together, the NSC is therefore well positioned to provide services to the Norwegian research communities in the wide variety of possible high-throughput sequencing applications.

The main goals of the NSC is to be able to provide high throughput sequencing services for resequencing, transcriptomics, metagenomics and de novo sequencing for the Norwegian research community and provide customized bioinformatic analysis of the sequence data.

Further information on sequencing services using the Illumina **Genome Analyzer II (GAII)**.

Further information on sequencing services using the **454 (Roche) GS FLX**.

For general information refer to our **services homepage** or send us an email.

Norwegian Sequencing Centre News

NSC seminar June 11th, 2010: High-throughput Sequencing - Applications and Analyses
Jun 3, 2010 12:00 AM

New assembly program available on titan: Celera 6.1
May 18, 2010 12:00 AM

NSC present at the 2010 Contact Meeting of the NSB
Jan 4, 2010 12:00 AM

More...

Publications

More...

Events

NSC Seminar: 'High-throughput Sequencing - Applications and Analyses'
Jun 11, 2010 10:00 AM

More...

Dept of Medical Genetics, Ullevål
University Hospital (Ullevål)
Kirkeveien 166,
0407 Oslo, Norway

CEES, Dept. of Biology,
University of Oslo
NSC, P.O. Box 1066 Blindern
0316 Oslo, Norway

Email
Follow us on twitter
More contact information

Powered by: Vortex

Manage folder

National conferences
NSB: talks, booth
NSHG, seminars etc.

Contact

Illumina platform

Instruments	Illumina GAIIx (2) (HiSeq2000)
People	1 Daily leader 1 PostDoc 2 Technicians 2 Bioinformaticians
Data storage	~60 TB local NorStore Secure storage...

Platform services

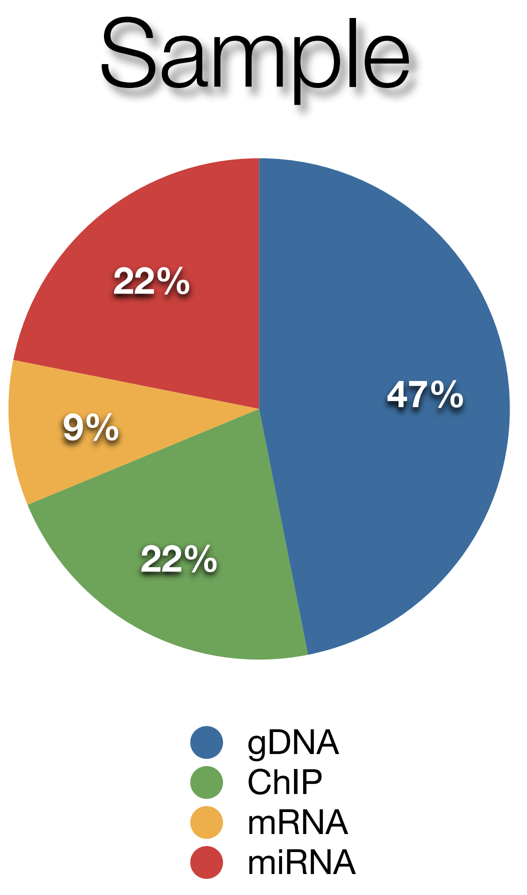
User	Sample	DNA, RNA
Platform	Sequencing	QC library preparation sequencing
	Costs	Illumina reagents QC reagents 20% platform fee (No staff/platform costs)
	Bioinformatics	Basic run information, QC Alignment to reference genome ?

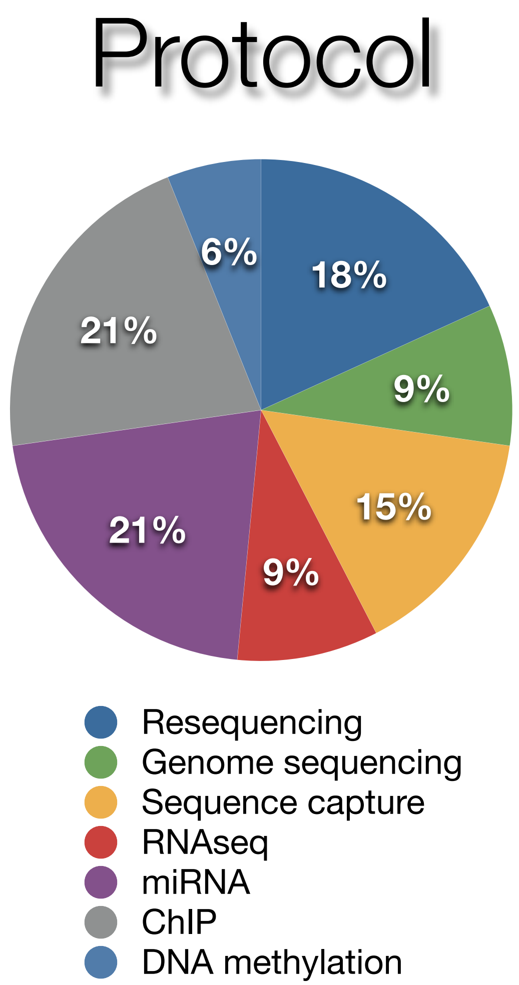
Applications run on Illumina node

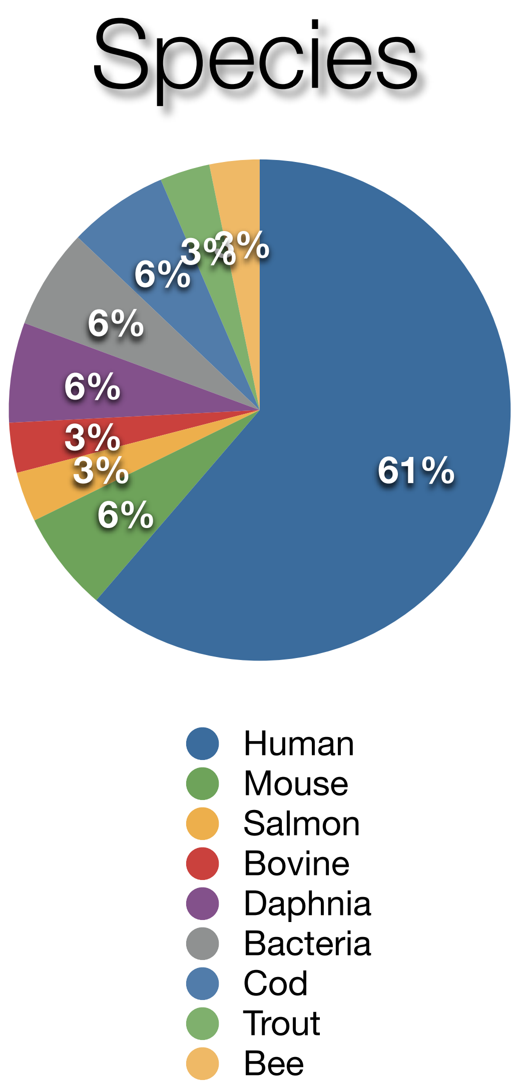
Application	Project	Sample	Protocol
Resequencing	whole genome linkage/association mutation detection	Genomic DNA	sequence capture, exome sequencing
de novo sequencing	metagenomics new species	Genomic DNA	SE, PE, mate-pair
Expression	transcriptome miRNA	mRNA, miRNA	RNAseq, miRNA
Epigenetics	DNA methylation chromatin structure	Genomic DNA	Bisulphite sequencing (RRBS), ChIP, MeDIP

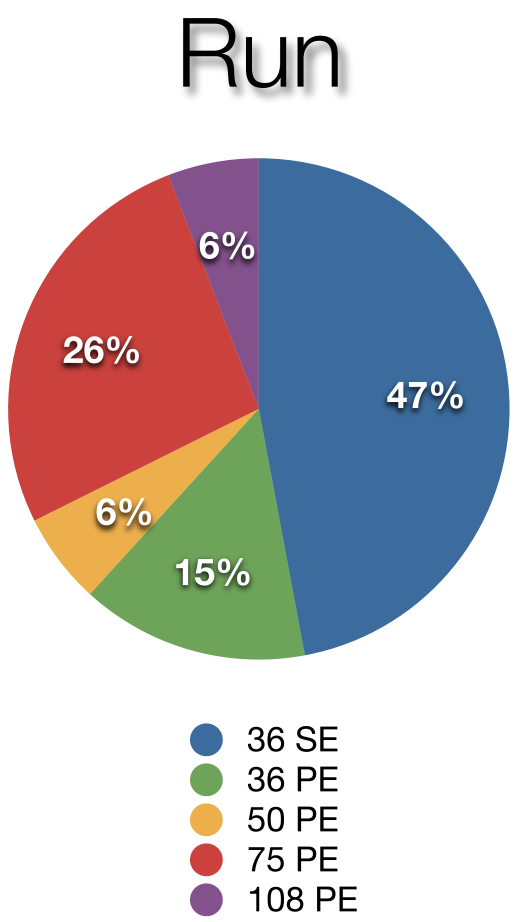
1x36 bp -> 2x108 bp

Runs overview



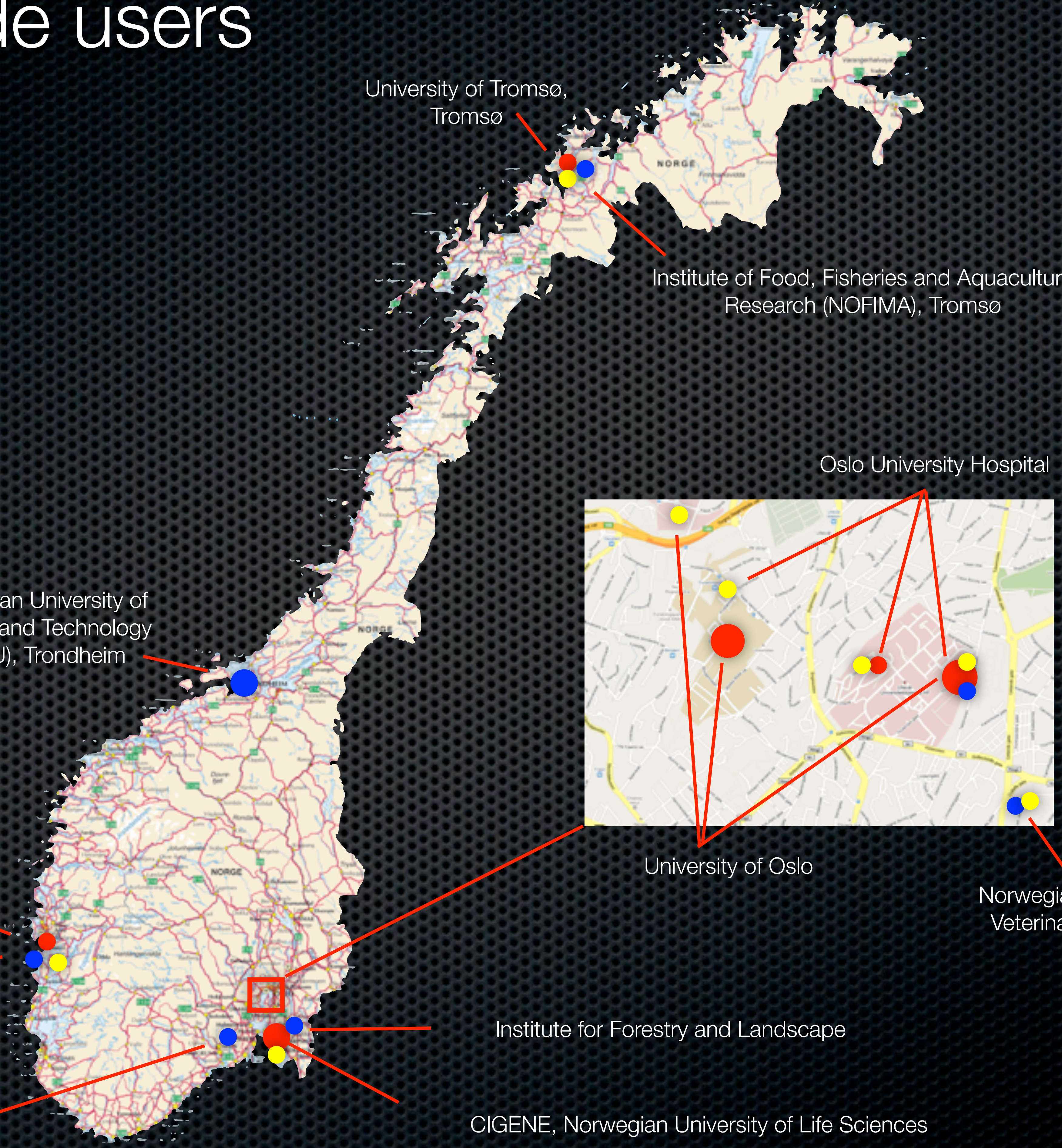






Nationwide users

- 🔴 Run
- 🟡 Scheduled
- 🔵 Contacted



University of Tromsø, Tromsø

Institute of Food, Fisheries and Aquaculture Research (NOFIMA), Tromsø

Oslo University Hospital

Norwegian University of Science and Technology (NTNU), Trondheim

University of Oslo

Norwegian School of Veterinary Science

National Institute of Nutrition and Seafood Research (NIFES), Bergen

University of Bergen, Bergen

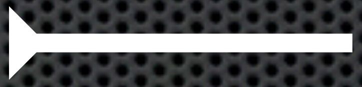
Institute for Forestry and Landscape

Telemark Hospital, Skien

CIGENE, Norwegian University of Life Sciences

Data and analysis

Illumina sequence data

- Random DNA library of short fragments ~300 bp
- ~100-300 million DNA sequences
- 18, 36, 50, 75, 125 bp long
- Single-end reads 
- Paired-end reads

- Run time: 1-10 days
- Data volume: 300 GB.....8 TB

Data issues

- Up to 4 TB/week

- Data storage and backup
- Network speed
- Security (human data)

- Data law - ‘return of results’
- Bioinformatics

Users

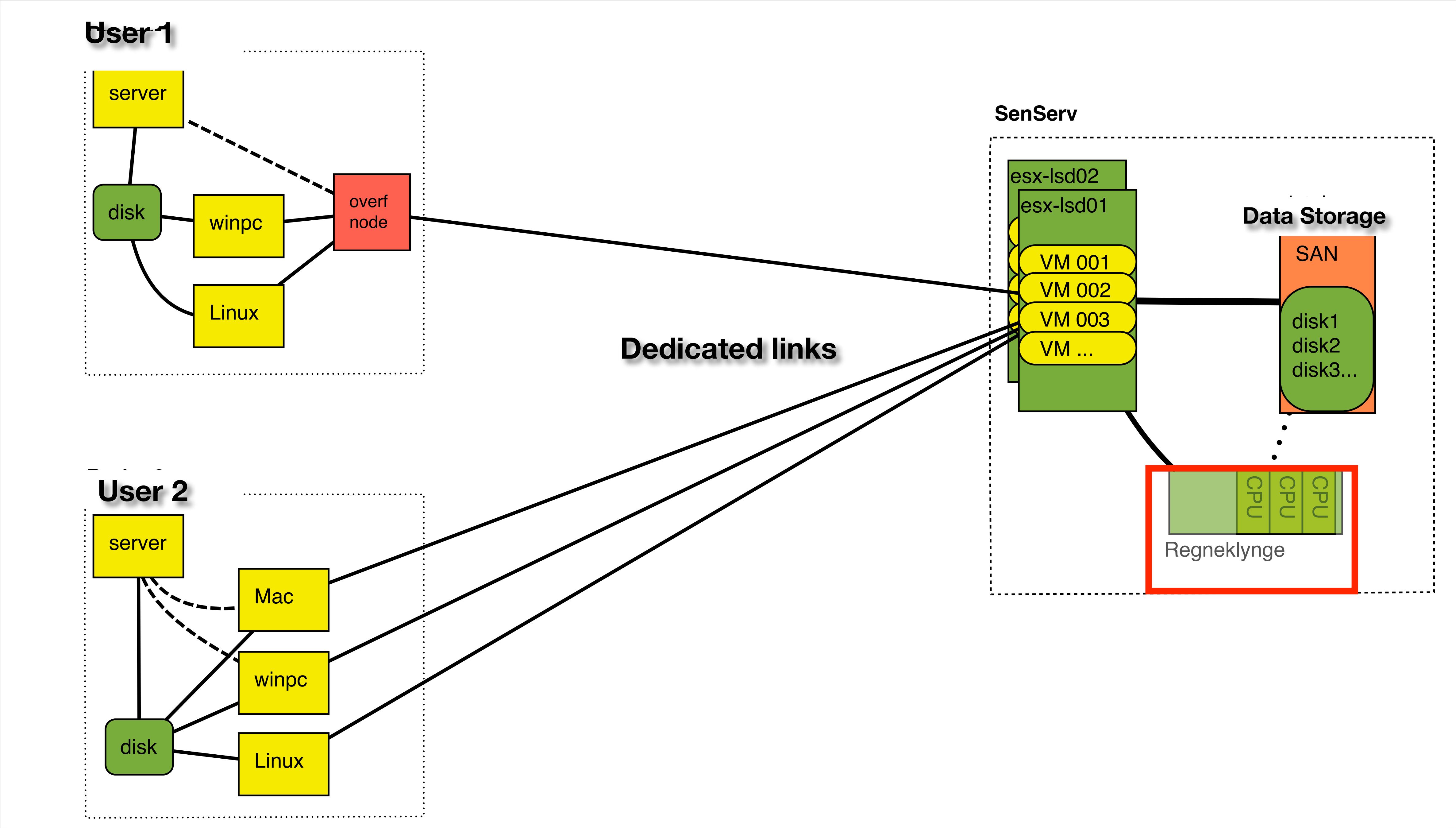
- Many users
- Many institutes
- Many applications

→ Bioinformatic challenge

User data storage - Phases

Phase	Provision	Timeline
0	Storage/backup of non-sensitive data from NSC (NorStore)	Complete 12.2009
1	High capacity secure storage of coded but indirectly identifiable data at local level (OUS, UiO) Establish routines for backup	Complete 6.2010
2	Robust secure solution for HTS data at the national level Exchange of data collaborators Infrastructure for analysis through the Bioinformatics platform	Start 11.2010
3	Clinical usage, secure handling of person-identifiable data being part of patient journals	?

User data storage - Phase 1

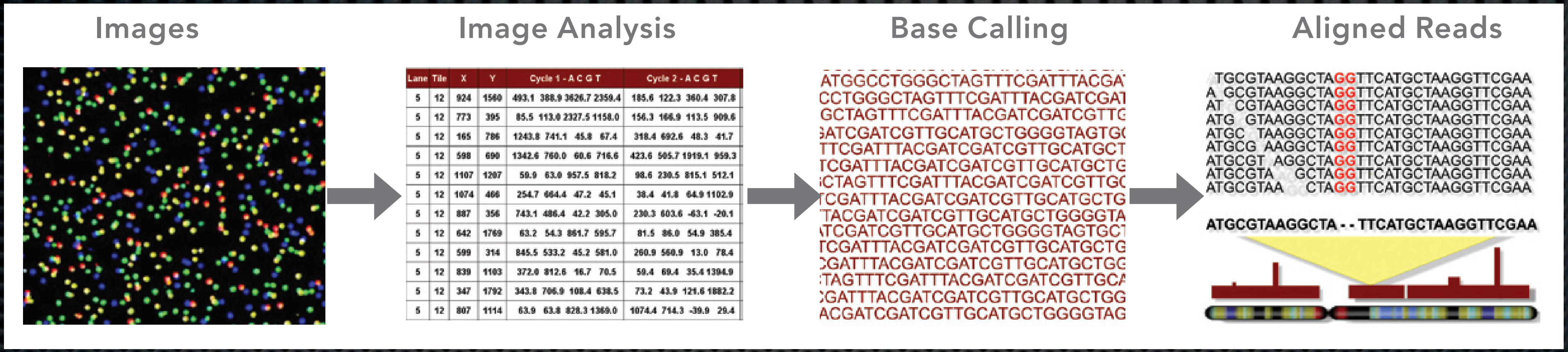


Secure storage/backup for sensitive data

Illumina analysis pipeline

Illumina Pipeline 1.4

SCS | Firecrest | Bustard | GERALD



Lane	Tile	X	Y	Cycle 1 - A C G T				Cycle 2 - A C G T			
5	12	924	1560	493.1	388.9	3626.7	2359.4	185.6	122.3	360.4	307.8
5	12	773	395	85.5	113.0	2327.5	1158.0	156.3	166.9	113.5	909.6
5	12	165	786	1243.8	741.1	45.8	67.4	318.4	692.6	48.3	41.7
5	12	598	690	1342.6	760.0	60.6	716.6	423.6	505.7	1919.1	959.3
5	12	1107	1207	59.9	63.0	957.5	818.2	98.6	230.5	815.1	512.1
5	12	1074	466	254.7	664.4	47.2	45.1	38.4	41.8	64.9	1102.9
5	12	887	356	743.1	486.4	42.2	305.0	230.3	603.6	-63.1	-20.1
5	12	642	1769	63.2	54.3	861.7	595.7	81.5	86.0	54.9	385.4
5	12	599	314	845.5	533.2	45.2	581.0	260.9	560.9	13.0	78.4
5	12	839	1103	372.0	812.6	16.7	70.5	59.4	69.4	35.4	1394.9
5	12	347	1792	343.8	706.9	108.4	638.5	73.2	43.9	121.6	1882.2
5	12	807	1114	63.9	63.8	828.3	1369.0	1074.4	714.3	-39.9	29.4

```
ATGGCCTGGGCTAGTTTCGATTTACGA
CCTGGGCTAGTTTCGATTTACGATCGA
GCTAGTTTCGATTTACGATCGATCGTT
ATCGATCGTTGCATGCTGGGGTAGTG
TTCGATTTACGATCGATCGTTGCATGCT
TCGATTTACGATCGATCGTTGCATGCTG
CTAGTTTCGATTTACGATCGATCGTTG
TCGATTTACGATCGATCGTTGCATGCTG
TACGATCGATCGTTGCATGCTGGGGTA
TCGATCGTTGCATGCTGGGGTAGTGC
TCGATTTACGATCGATCGTTGCATGCTG
GATTTACGATCGATCGTTGCATGCTGG
TAGTTTCGATTTACGATCGATCGTTGC
GATTTACGATCGATCGTTGCATGCTGG
ACGATCGATCGTTGCATGCTGGGGTAG
```

```
 TGCGTAAGGCTAGGTTCATGCTAAGGTTCGAA
A GCGTAAGGCTAGGTTCATGCTAAGGTTCGAA
AT CGTAAGGCTAGGTTCATGCTAAGGTTCGAA
ATG GTAAGGCTAGGTTCATGCTAAGGTTCGAA
ATGC TAAGGCTAGGTTCATGCTAAGGTTCGAA
ATGCG AAGGCTAGGTTCATGCTAAGGTTCGAA
ATGCGT AGGCTAGGTTCATGCTAAGGTTCGAA
ATGCGTA  GCTAGGTTCATGCTAAGGTTCGAA
ATGCGTAA  CTAGGTTCATGCTAAGGTTCGAA

ATGCGTAAGGCTA--TTCATGCTAAGGTTCGAA
```

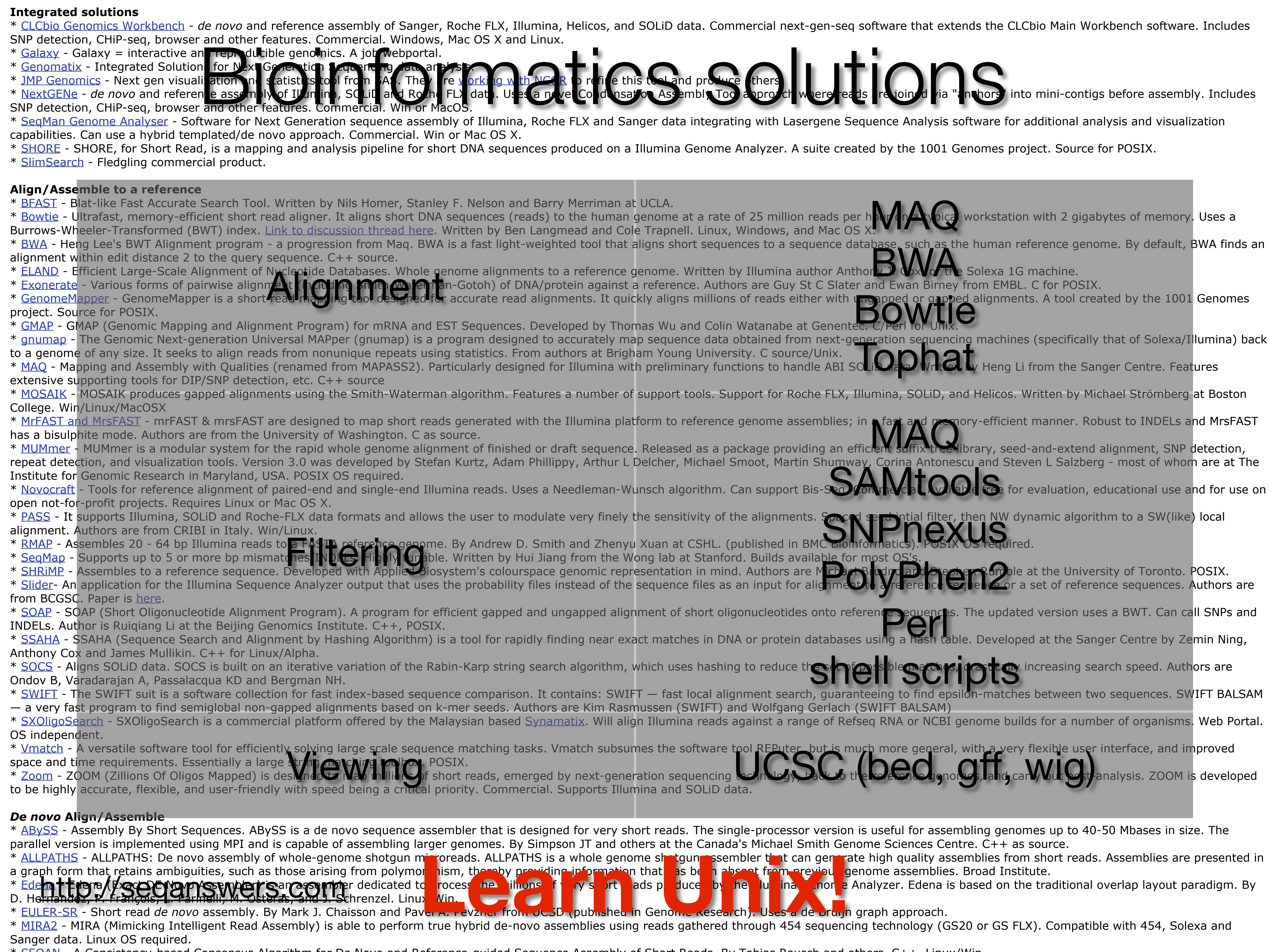
```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;39333
```

FASTQ format

Other software/ analyses

Run size	Images	Intensity files	Sequence files	Analysis output	Total
Minimum	700 GB	100 GB	100 GB	100 GB	**1 TB**
Average	1.4 TB	200 GB	200 GB	200 GB	**2 TB**
Maximum	3.5 TB	500 GB	500 GB	500 GB	**5 TB**

Bioinformatics solutions



Integrated solutions

* CLCbio Genomics Workbench - *de novo* and reference assembly of Sanger, Roche FLX, Illumina, Helicos, and SOLiD data. Commercial next-gen-seq software that extends the CLCbio Main Workbench software. Includes SNP detection, CHiP-seq, browser and other features. Commercial. Windows, Mac OS X and Linux.
* Galaxy - Galaxy = interactive and reproducible genomics. A job webportal.
* Genomatix - Integrated Solutions for Next Generation Sequencing data analysis.
* JMP Genomics - Next gen visualization and statistics tool from SAS. They are working with NCGR to refine this tool and produce others.
* NextGENe - *de novo* and reference assembly of Illumina, SOLiD and Roche FLX data. Uses a novel Condensation Assembly Tool approach where reads are joined via "anchors" into mini-contigs before assembly. Includes SNP detection, CHiP-seq, browser and other features. Commercial. Win or MacOS.
* SeqMan Genome Analyser - Software for Next Generation sequence assembly of Illumina, Roche FLX and Sanger data integrating with Lasergene Sequence Analysis software for additional analysis and visualization capabilities. Can use a hybrid templated/de novo approach. Commercial. Win or Mac OS X.
* SHORE - SHORE, for Short Read, is a mapping and analysis pipeline for short DNA sequences produced on a Illumina Genome Analyzer. A suite created by the 1001 Genomes project. Source for POSIX.
* SlimSearch - Fledgling commercial product.

Align/Assemble to a reference

* BFAST - Blat-like Fast Accurate Search Tool. Written by Nils Homer, Stanley F. Nelson and Barry Merriman at UCLA.
* Bowtie - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hour on a typical workstation with 2 gigabytes of memory. Uses a Burrows-Wheeler-Transformed (BWT) index. Link to discussion thread here. Written by Ben Langmead and Cole Trapnell. Linux, Windows, and Mac OS X.
* BWA - Heng Lee's BWT Alignment program - a progression from Maq. BWA is a fast light-weighted tool that aligns short sequences to a sequence database, such as the human reference genome. By default, BWA finds an alignment within edit distance 2 to the query sequence. C++ source.
* ELAND - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony J. Cox for the Solexa 1G machine.
* Exonerate - Various forms of pairwise alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Birney from EMBL. C for POSIX.
* GenomeMapper - GenomeMapper is a short read mapping tool designed for accurate read alignments. It quickly aligns millions of reads either with ungapped or gapped alignments. A tool created by the 1001 Genomes project. Source for POSIX.
* GMAP - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentec. C/Perl for Unix.
* gnumap - The Genomic Next-generation Universal MAPper (gnumap) is a program designed to accurately map sequence data obtained from next-generation sequencing machines (specifically that of Solexa/Illumina) back to a genome of any size. It seeks to align reads from nonunique repeats using statistics. From authors at Brigham Young University. C source/Unix.
* MAQ - Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina with preliminary functions to handle ABI SOLiD data. Written by Heng Li from the Sanger Centre. Features extensive supporting tools for DIP/SNP detection, etc. C++ source
* MOSAIK - MOSAIK produces gapped alignments using the Smith-Waterman algorithm. Features a number of support tools. Support for Roche FLX, Illumina, SOLiD, and Helicos. Written by Michael Strömberg at Boston College. Win/Linux/MacOSX
* MrFAST and MrsFAST - mrFAST & mrsFAST are designed to map short reads generated with the Illumina platform to reference genome assemblies; in a fast and memory-efficient manner. Robust to INDELs and MrsFAST has a bisulphite mode. Authors are from the University of Washington. C as source.
* MUMmer - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficient suffix tree library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. POSIX OS required.
* Novocraft - Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Can support Bis-Seq. Commercial. Available free for evaluation, educational use and for use on open not-for-profit projects. Requires Linux or Mac OS X.
* PASS - It supports Illumina, SOLiD and Roche-FLX data formats and allows the user to modulate very finely the sensitivity of the alignments. Spaced seed intial filter, then NW dynamic algorithm to a SW(like) local alignment. Authors are from CRIBI in Italy. Win/Linux.
* RMAP - Assembles 20 - 64 bp Illumina reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinformatics). POSIX OS required.
* SeqMap - Supports up to 5 or more bp mismatches/INDELs. Highly tunable. Written by Hui Jiang from the Wong lab at Stanford. Builds available for most OS's.
* SHRiMP - Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Michael Brudno and Stephen Rumble at the University of Toronto. POSIX.
* Slider- An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a reference sequence or a set of reference sequences. Authors are from BCGSC. Paper is here.
* SOAP - SOAP (Short Oligonucleotide Alignment Program). A program for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences. The updated version uses a BWT. Can call SNPs and INDELs. Author is Ruiqiang Li at the Beijing Genomics Institute. C++, POSIX.
* SSAHA - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases using a hash table. Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.
* SOCS - Aligns SOLiD data. SOCS is built on an iterative variation of the Rabin-Karp string search algorithm, which uses hashing to reduce the set of possible matches, drastically increasing search speed. Authors are Ondov B, Varadarajan A, Passalacqua KD and Bergman NH.
* SWIFT - The SWIFT suit is a software collection for fast index-based sequence comparison. It contains: SWIFT — fast local alignment search, guaranteeing to find epsilon-matches between two sequences. SWIFT BALSAM — a very fast program to find semiglobal non-gapped alignments based on k-mer seeds. Authors are Kim Rasmussen (SWIFT) and Wolfgang Gerlach (SWIFT BALSAM)
* SXOligoSearch - SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illumina reads against a range of Refseq RNA or NCBI genome builds for a number of organisms. Web Portal. OS independent.
* Vmatch - A versatile software tool for efficiently solving large scale sequence matching tasks. Vmatch subsumes the software tool REPuter, but is much more general, with a very flexible user interface, and improved space and time requirements. Essentially a large string matching toolbox. POSIX.
* Zoom - ZOOM (Zillions Of Oligos Mapped) is designed to map millions of short reads, emerged by next-generation sequencing technology, back to the reference genomes, and carry out post-analysis. ZOOM is developed to be highly accurate, flexible, and user-friendly with speed being a critical priority. Commercial. Supports Illumina and SOLiD data.

***De novo* Align/Assemble**

* ABySS - Assembly By Short Sequences. ABySS is a de novo sequence assembler that is designed for very short reads. The single-processor version is useful for assembling genomes up to 40-50 Mbases in size. The parallel version is implemented using MPI and is capable of assembling larger genomes. By Simpson JT and others at the Canada's Michael Smith Genome Sciences Centre. C++ as source.
* ALLPATHS - ALLPATHS: De novo assembly of whole-genome shotgun microreads. ALLPATHS is a whole genome shotgun assembler that can generate high quality assemblies from short reads. Assemblies are presented in a graph form that retains ambiguities, such as those arising from polymorphism, thereby providing information that has been absent from previous genome assemblies. Broad Institute.
* Edena - Edena (Exact DE Novo Assembler) is an assembler dedicated to process the millions of very short reads produced by the Illumina Genome Analyzer. Edena is based on the traditional overlap layout paradigm. By D. Hernandez, P. François, L. Farinelli, M. Osteras, and J. Schrenzel. Linux/Win.
* EULER-SR - Short read *de novo* assembly. By Mark J. Chaisson and Pavel A. Pevzner from UCSD (published in Genome Research). Uses a de Bruijn graph approach.
* MIRA2 - MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo assemblies using reads gathered through 454 sequencing technology (GS20 or GS FLX). Compatible with 454, Solexa and Sanger data. Linux OS required.
* SEQAN - A Consistency-based Consensus Algorithm for De Novo and Reference-guided Sequence Assembly of Short Reads. By Tobias Rausch and others. C++, Linux/Win.

Alignment

MAQ
BWA
Bowtie
Tophat

Filtering

MAQ
SAMtools
SNPnexus
PolyPhen2
Perl
shell scripts

Viewing

UCSC (bed, gff, wig)

http://seqanswers.com

Learn Unix!

Analysis hardware

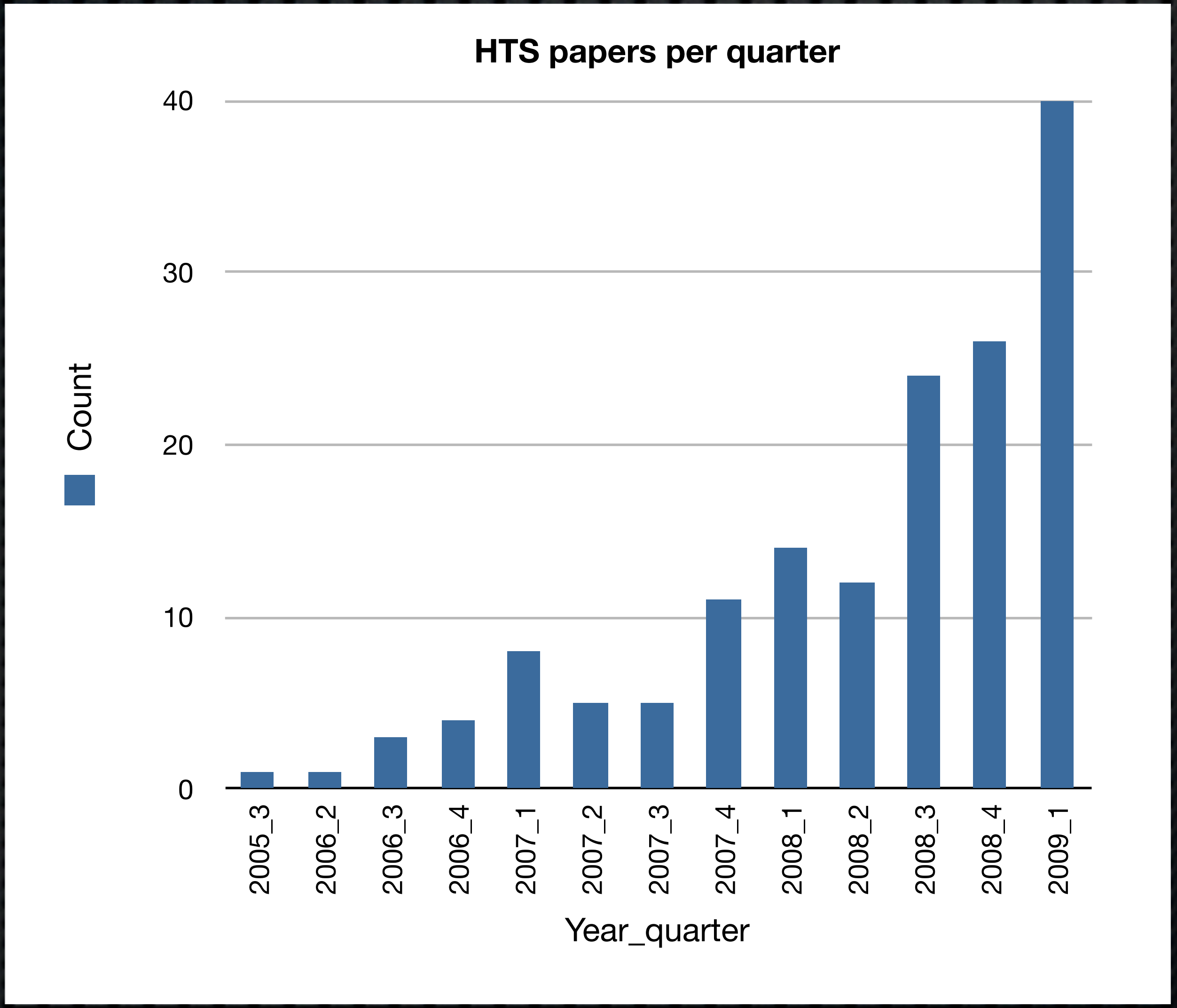
System	Specifications
Pipeline server	
Processor	HP Proliant dl580 g5 rack server (4 quad-core 2.93GHz 64-bit Intel Xeon)
Memory	32 GB
Storage	21 TB (HP 60 MSA)
Operating system	Linux
iPAR	
Processor	HP DL 380 (2 × 5460 3.16 GHz)
Memory	16 GB
Storage	3.2 TB (HP SmartArray P800)
Operating system	Linux/XP
Mac Pro (x2)	
Processor	2 quad-core 2.66 GHz 64-bit Intel Nehalem
Memory	16 GB
Storage	4 TB
Operating system	OS X

NorStore,Titan.....

Break?

Applications

Research publications



Applications

Application	Project
Resequencing	whole genome linkage/association mutation detection
de novo sequencing	metagenomics new species
Expression	transcriptome SAGE miRNA
Epigenetics	DNA methylation ChIP
Variation	SNPs CNVs

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)

Resequencing: mutation detection

Genomic region known

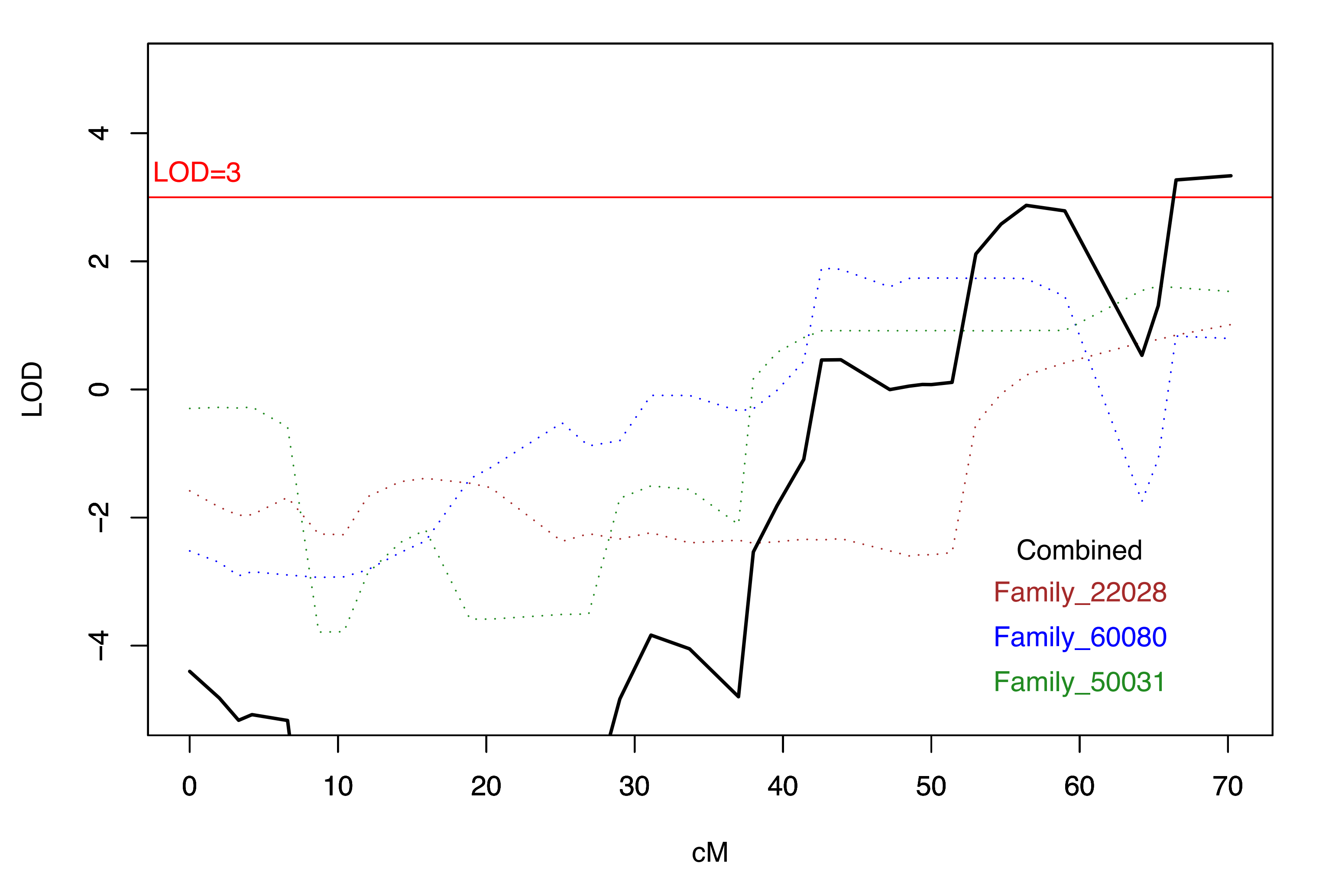
- Linkage peak
- Sequence capture - region of interest

Genomic region unknown

- Rare Mendelian disorders
- Sequence capture - exome
- RNAseq

Region known

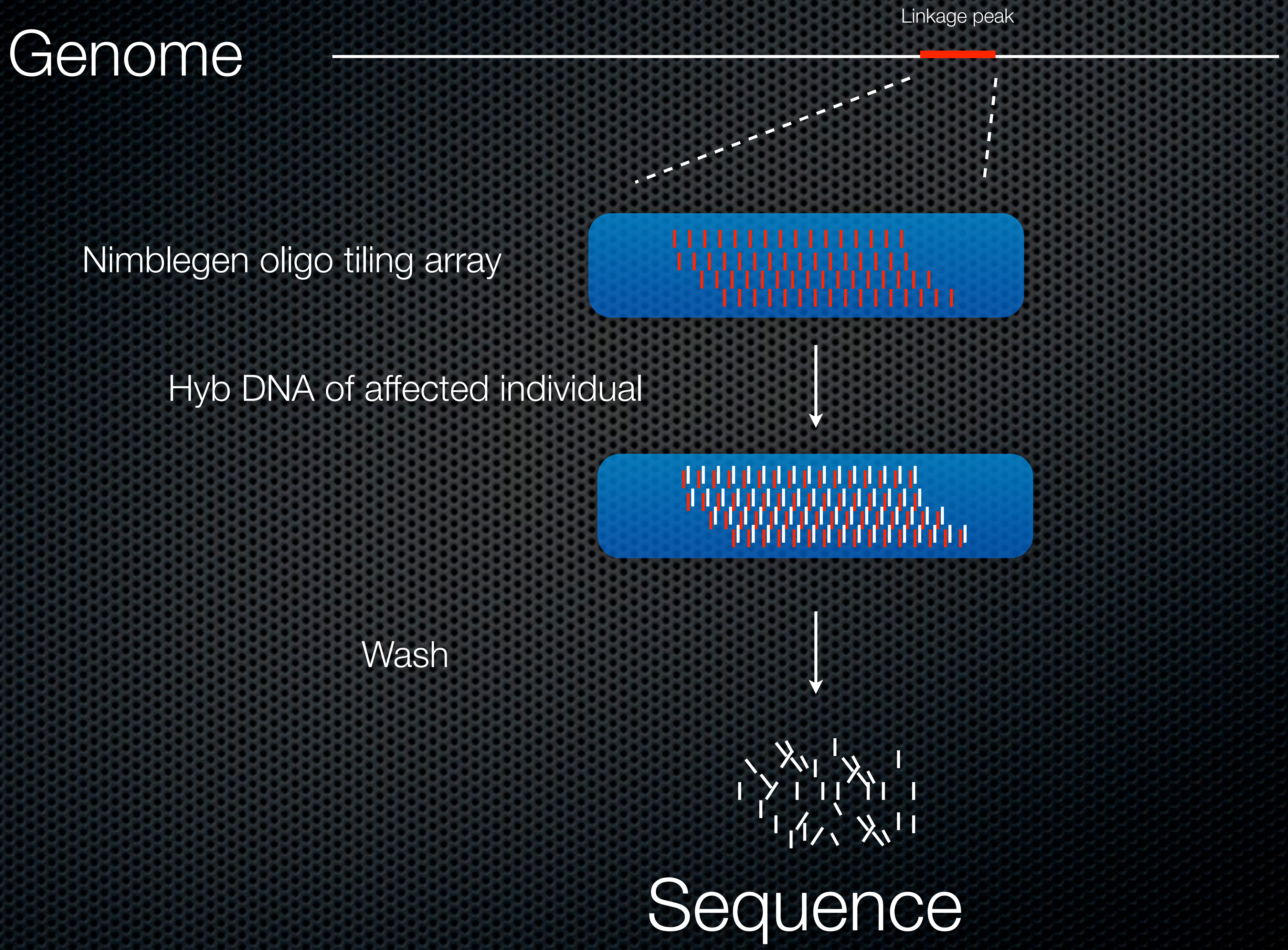
Linkage



1-10 Mb?

How can we capture this region to sequence?

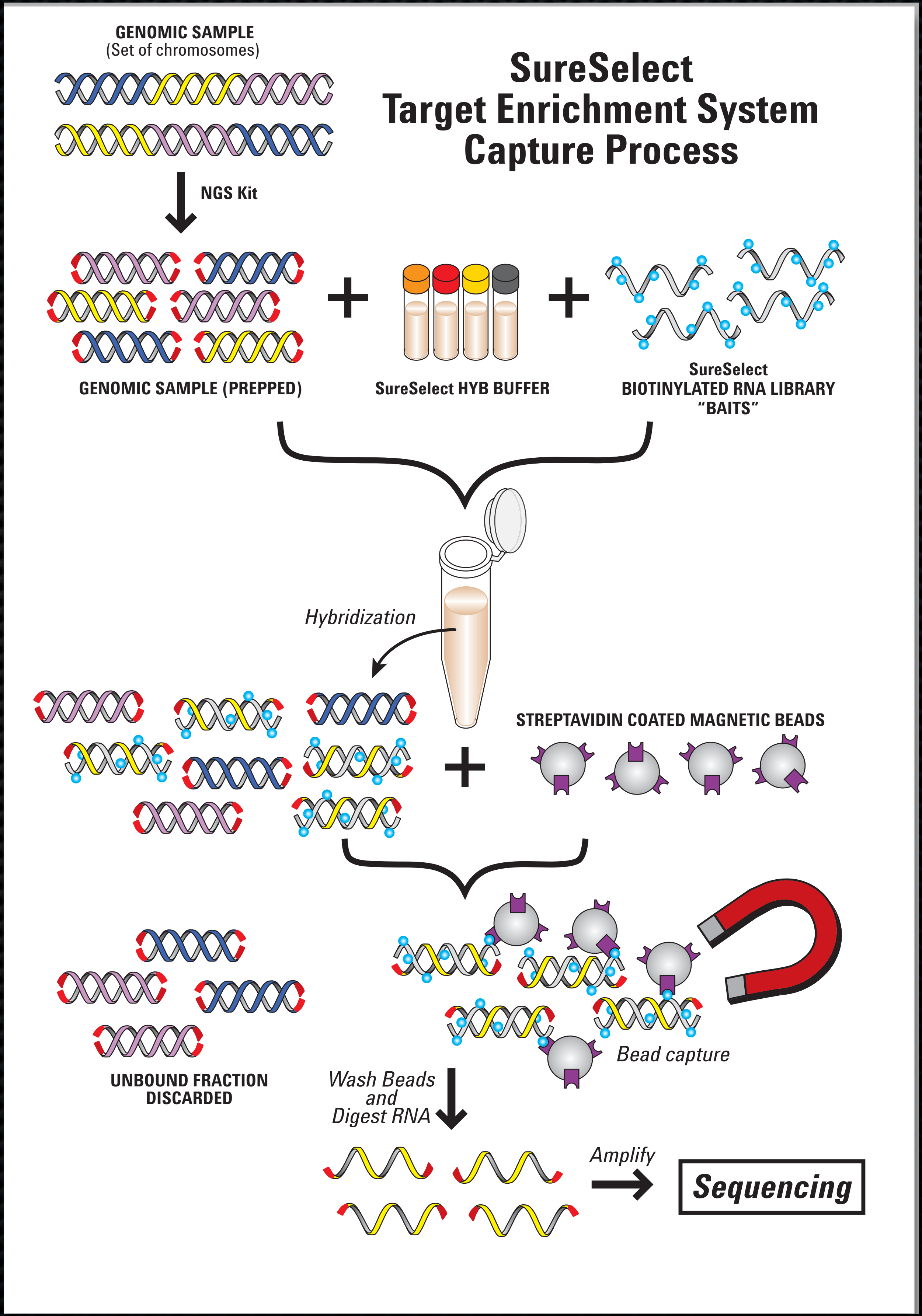
Sequence capture



Agilent SureSelect

- RNA oligonucleotides
- >100 bp
- custom design

www.agilent.com



Analyzing resequencing data

- Capture DNA and sequence
- Prepare sequence files (Perl...)
- Align to reference (MAQ, BWA etc.)
- Format/filter output files (Perl...)
  - .bed, .gtf
  - View on genome browser
  - identify variants

Analysis pipeline

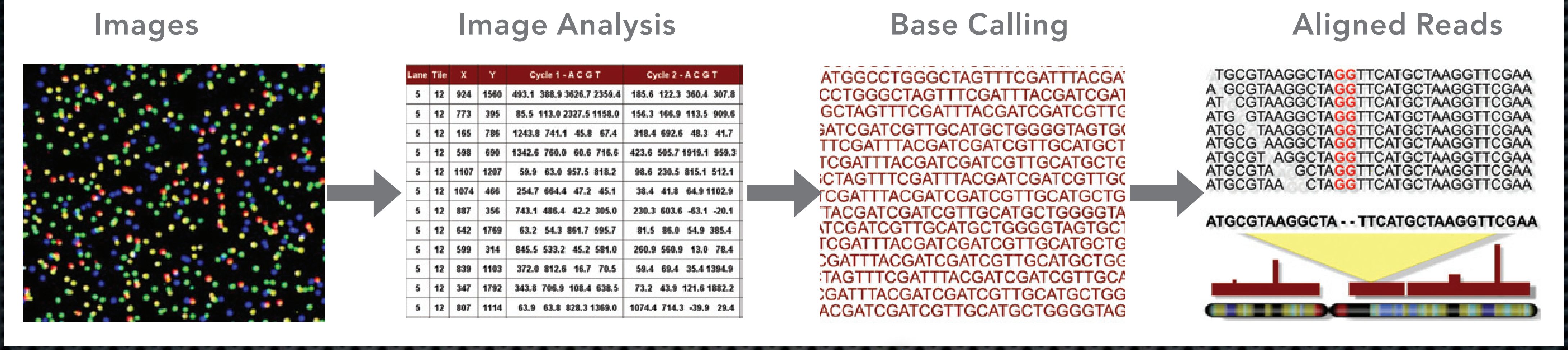
Illumina Pipeline 1.4

SCS

Firecrest

Bustard

GERALD



Images

Image Analysis

Lane	Tile	X	Y	Cycle 1 - A C G T				Cycle 2 - A C G T			
5	12	924	1560	493.1	388.9	3626.7	2359.4	185.6	122.3	360.4	307.8
5	12	773	395	85.5	113.0	2327.5	1158.0	156.3	166.9	113.5	909.6
5	12	165	786	1243.8	741.1	45.8	67.4	318.4	692.6	48.3	41.7
5	12	598	690	1342.6	760.0	60.6	716.6	423.6	505.7	1919.1	959.3
5	12	1107	1207	59.9	63.0	957.5	818.2	98.6	230.5	815.1	512.1
5	12	1074	466	254.7	664.4	47.2	45.1	38.4	41.8	64.9	1102.9
5	12	887	356	743.1	486.4	42.2	305.0	230.3	603.6	-63.1	-20.1
5	12	642	1769	63.2	54.3	861.7	595.7	81.5	86.0	54.9	385.4
5	12	599	314	845.5	533.2	45.2	581.0	260.9	560.9	13.0	78.4
5	12	839	1103	372.0	812.6	16.7	70.5	59.4	69.4	35.4	1394.9
5	12	347	1792	343.8	706.9	108.4	638.5	73.2	43.9	121.6	1882.2
5	12	807	1114	63.9	63.8	828.3	1369.0	1074.4	714.3	-39.9	29.4

Base Calling

```
ATGGCCTGGGCTAGTTTCGATTTACGA
CCTGGGCTAGTTTCGATTTACGATCGA
GCTAGTTTCGATTTACGATCGATCGTT
GATCGATCGTTGCATGCTGGGGTAGTG
TTCGATTTACGATCGATCGTTGCATGCT
TCGATTTACGATCGATCGTTGCATGCTG
GCTAGTTTCGATTTACGATCGATCGTTG
TCGATTTACGATCGATCGTTGCATGCTG
TACGATCGATCGTTGCATGCTGGGGTA
ATCGATCGTTGCATGCTGGGGTAGTGC
TCGATTTACGATCGATCGTTGCATGCTG
CGATTTACGATCGATCGTTGCATGCTGG
TAGTTTCGATTTACGATCGATCGTTGCA
CGATTTACGATCGATCGTTGCATGCTGG
ACGATCGATCGTTGCATGCTGGGGTAG
```

Aligned Reads

```
 TGCGTAAGGCTAGGTTCATGCTAAGGTTCGAA
A GCGTAAGGCTAGGTTCATGCTAAGGTTCGAA
AT  CGTAAGGCTAGGTTCATGCTAAGGTTCGAA
ATG  GTAAGGCTAGGTTCATGCTAAGGTTCGAA
ATGC  TAAGGCTAGGTTCATGCTAAGGTTCGAA
ATGCG AAGGCTAGGTTCATGCTAAGGTTCGAA
ATGCGT AGGCTAGGTTCATGCTAAGGTTCGAA
ATGCGTA   GCTAGGTTCATGCTAAGGTTCGAA
ATGCGTAA   CTAGGTTCATGCTAAGGTTCGAA

ATGCGTAAGGCTA - - TTCATGCTAAGGTTCGAA
```

FASTQ format

```
@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
```

Other software/analyses

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

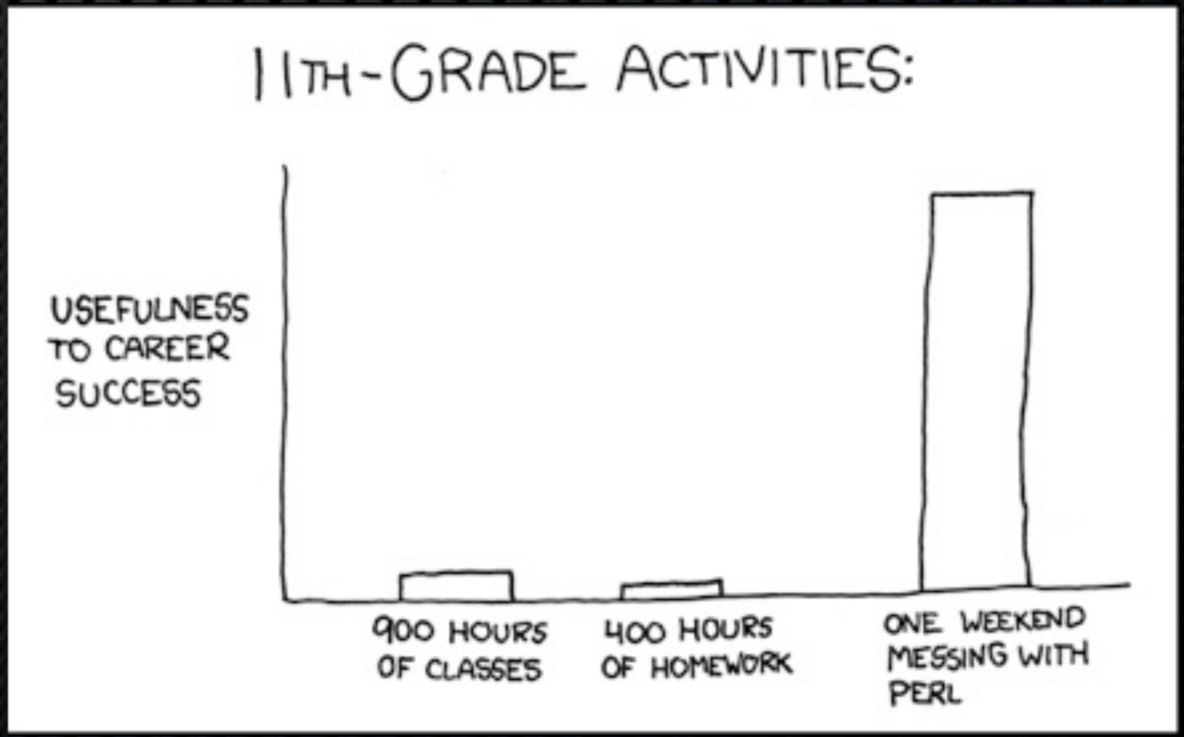
R|G

Sequence

Mutation

Tools for HTS

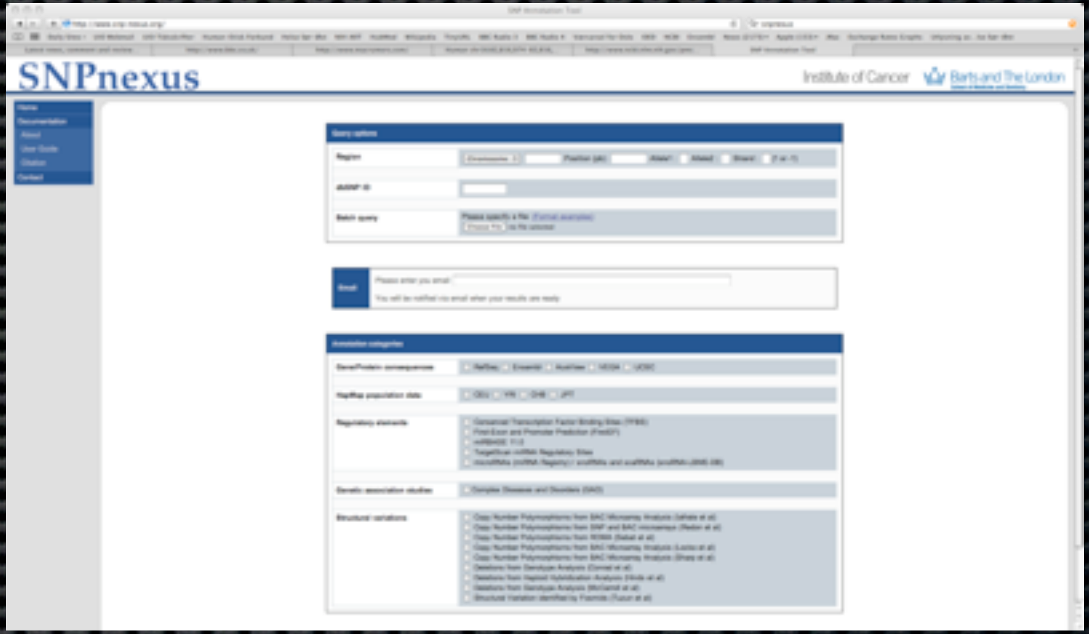
- Alignment - MAQ, BWA
- Filtering, sorting etc. - SAMTools, BEDTools
- Viewing - BED, GFF, UCSC browser
- Perl, unix scripts



xkcd.com

Finding mutations?

- Which variants are deleterious?
- Novel? (dbSNP, 1000genomes, HGMD)
- Synonymous/non-synonymous?
- Conserved?
- Alter protein structure?



PolyPhen2
MutationTaster
ANNOVAR
SeattleSeq Annotation

1000 genomes project

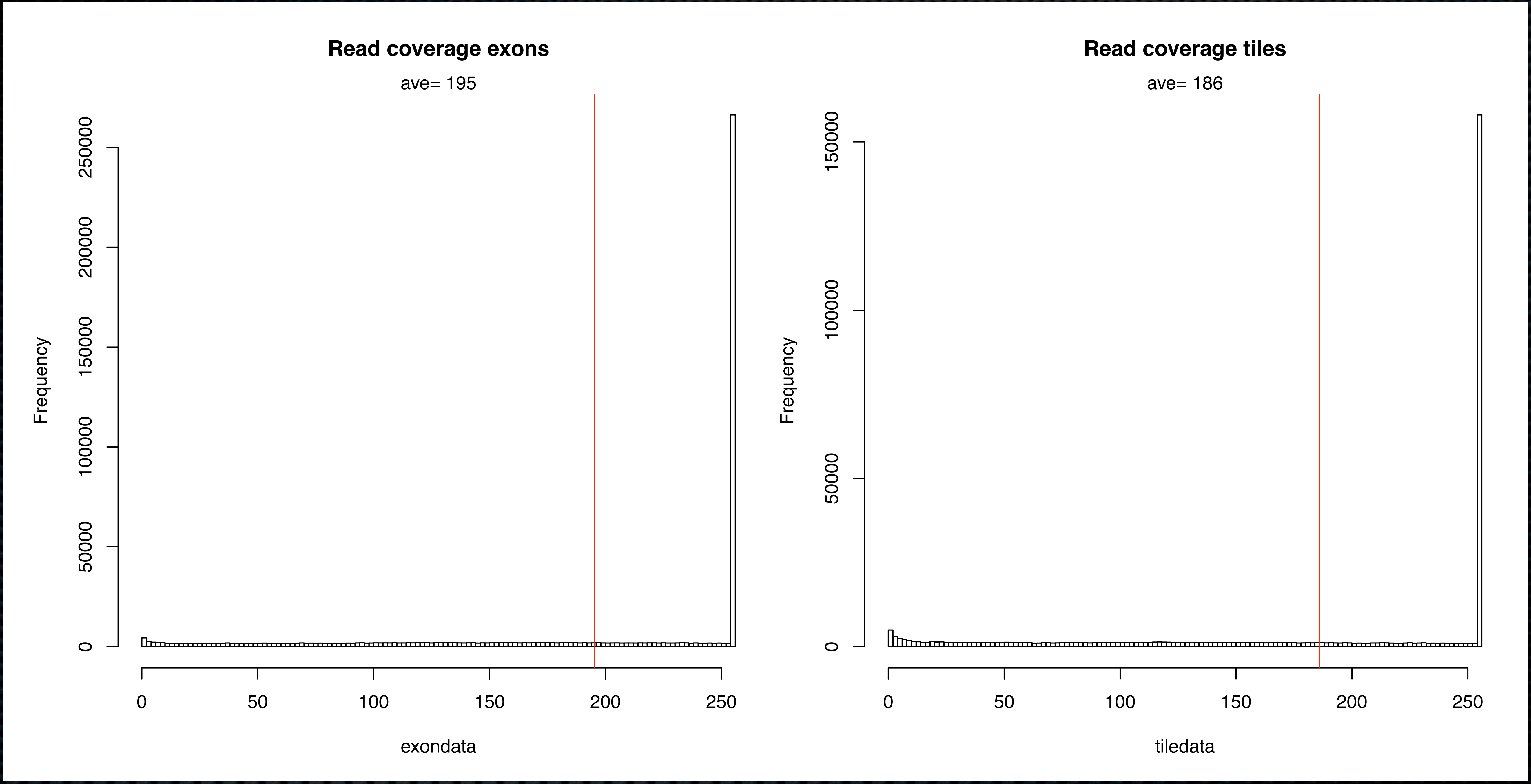
- International consortium
- Sequence 1200 genomes
- Produce a nearly complete catalog of common human genetic variants (defined as frequency 1% or higher; SNPs, CNVs)
  - mutation detection in Mendelian disease
  - accelerate fine-mapping efforts association studies
  - enabling design of next-generation genotyping arrays - improve the power of future genetic association studies
  - improve our ability to “impute” or “predict” untyped genetic variants
- Frequent public data releases

1000genomes.org

Crude analysis pipeline

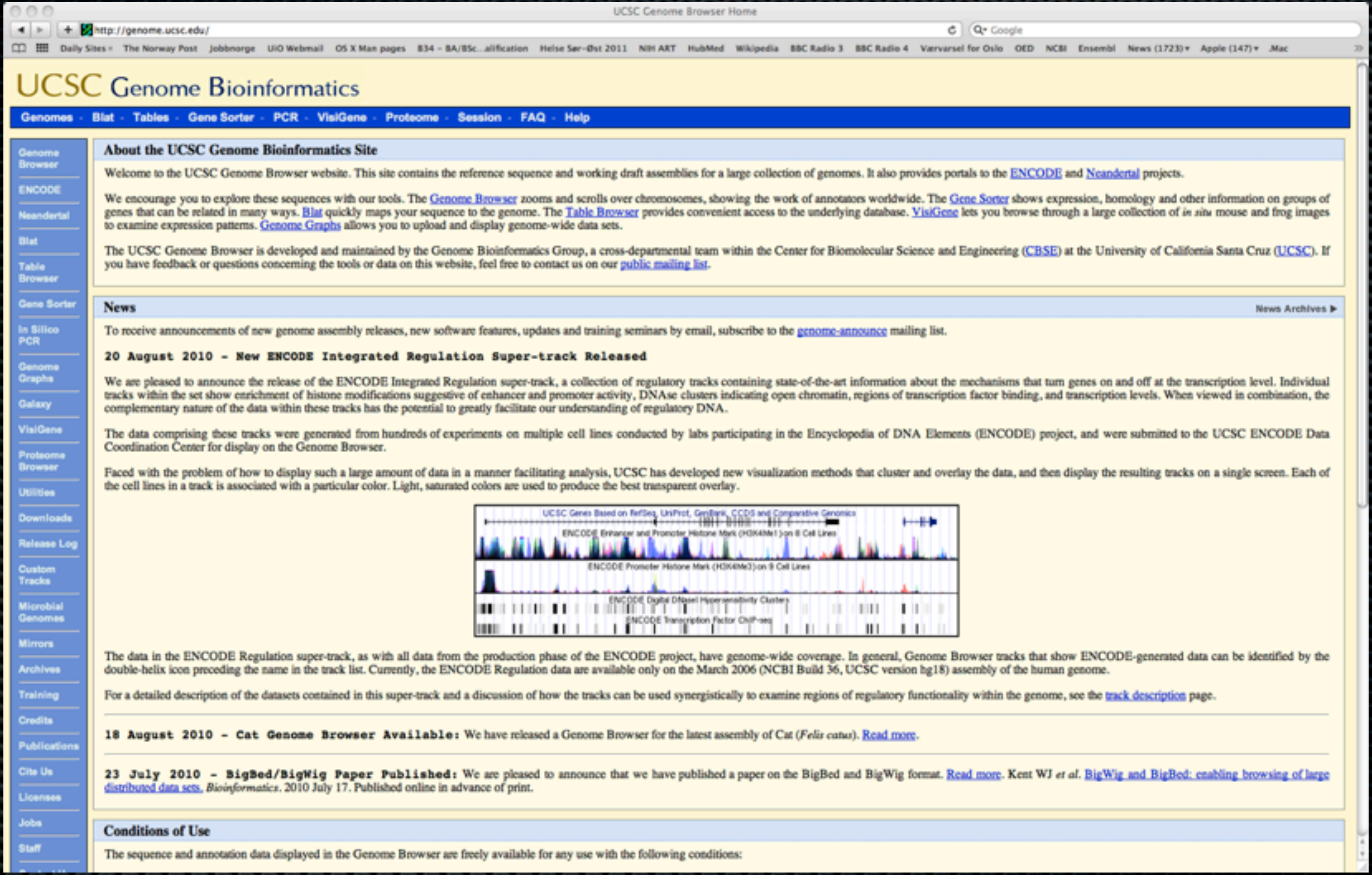
Program/script	Description	Output
maq fasta2bfa	Prepare ref sequences	bfa
maq fastq2bfq	Convert FASTQ reads to BFQ format	bfq
maq map	Align	MAQ aln
maq assemble	Assemble	MAQ cns
maq cns2snp	Call SNPs	MAQ SNP
awk '$2>=29621176 && $2<=39095041'	Filter for ROI	MAQ SNP
maq.pl SNPfilter	Q filter SNPs	MAQ SNP
maq cns2view	MAQ file for ROI	MAQ SNP
maqview2bed.pl	bed file for ROI	bed
maqsnp2bed.pl	bed file for SNPs	bed
maqsnp2snpnexus.pl	Input for SNPnexus	SNPnexus input
parseSNPnexus.pl	Parse SNPnexus output	SNPnexus output
bases2nexus.pl	Variants file	bases file
maqCoverageSummary.pl	Sequence coverage	bed, pdf
coverage_v4.pl	Sequence coverage	bed
lowCoverageSummary.pl	Sequence coverage	bed

Read depth statistics



R

UCSC Genome browser



About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the ENCODE and Neandertal projects.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.

News

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.

20 August 2010 - New ENCODE Integrated Regulation Super-track Released

We are pleased to announce the release of the ENCODE Integrated Regulation super-track, a collection of regulatory tracks containing state-of-the-art information about the mechanisms that turn genes on and off at the transcription level. Individual tracks within the set show enrichment of histone modifications suggestive of enhancer and promoter activity, DNAse clusters indicating open chromatin, regions of transcription factor binding, and transcription levels. When viewed in combination, the complementary nature of the data within these tracks has the potential to greatly facilitate our understanding of regulatory DNA.

The data comprising these tracks were generated from hundreds of experiments on multiple cell lines conducted by labs participating in the Encyclopedia of DNA Elements (ENCODE) project, and were submitted to the UCSC ENCODE Data Coordination Center for display on the Genome Browser.

Faced with the problem of how to display such a large amount of data in a manner facilitating analysis, UCSC has developed new visualization methods that cluster and overlay the data, and then display the resulting tracks on a single screen. Each of the cell lines in a track is associated with a particular color. Light, saturated colors are used to produce the best transparent overlay.

The data in the ENCODE Regulation super-track, as with all data from the production phase of the ENCODE project, have genome-wide coverage. In general, Genome Browser tracks that show ENCODE-generated data can be identified by the double-helix icon preceding the name in the track list. Currently, the ENCODE Regulation data are available only on the March 2006 (NCBI Build 36, UCSC version hg18) assembly of the human genome.

For a detailed description of the datasets contained in this super-track and a discussion of how the tracks can be used synergistically to examine regions of regulatory functionality within the genome, see the track description page.

18 August 2010 - Cat Genome Browser Available: We have released a Genome Browser for the latest assembly of Cat (*Felis catus*). Read more.

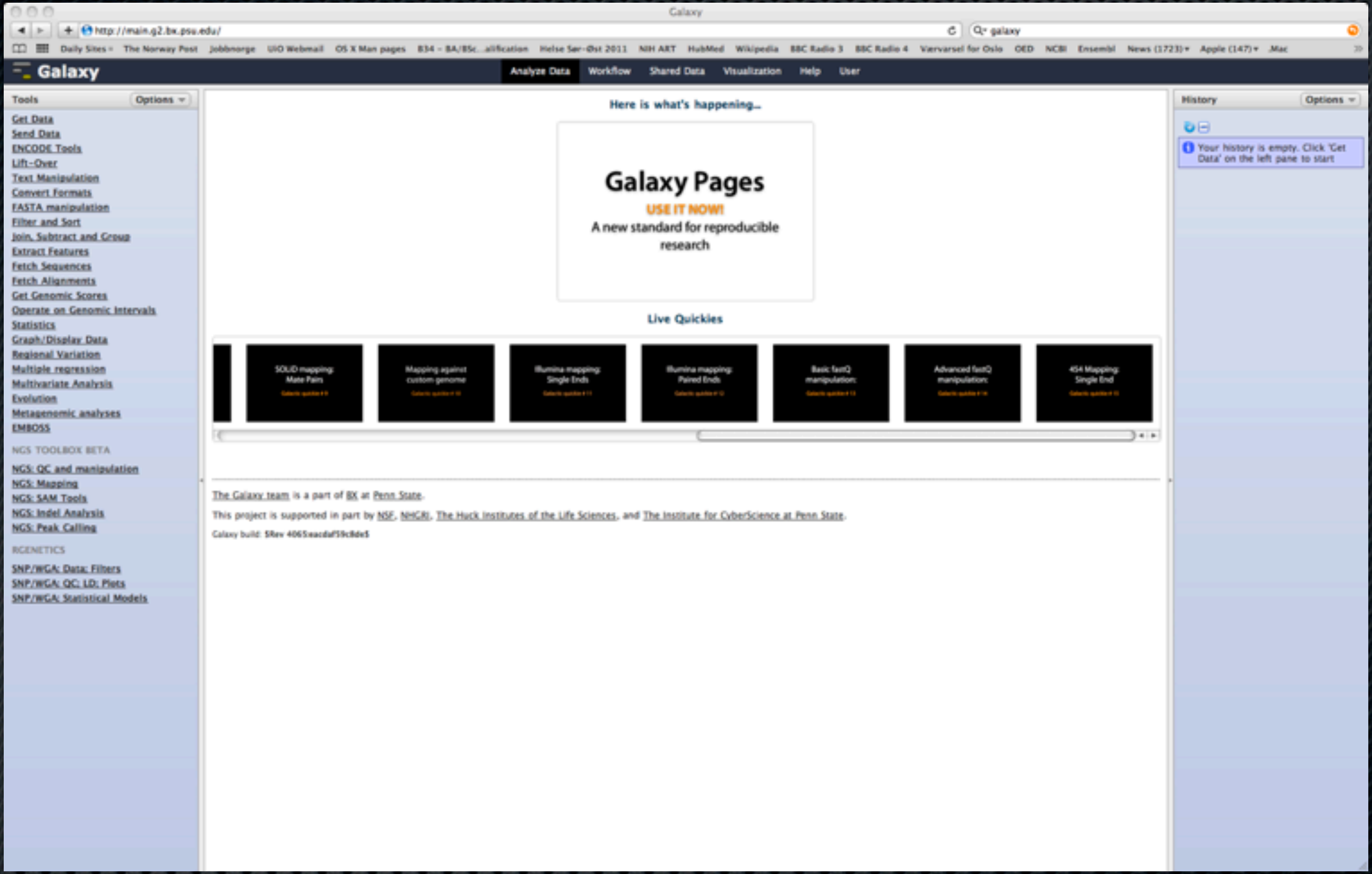
23 July 2010 - BigBed/BigWig Paper Published: We are pleased to announce that we have published a paper on the BigBed and BigWig format. Read more. Kent WJ *et al.* BigWig and BigBed: enabling browsing of large distributed data sets. *Bioinformatics*. 2010 July 17. Published online in advance of print.

Conditions of Use

The sequence and annotation data displayed in the Genome Browser are freely available for any use with the following conditions:

http://genome.ucsc.edu/

Galaxy



Your history is empty. Click 'Get Data' on the left pane to start

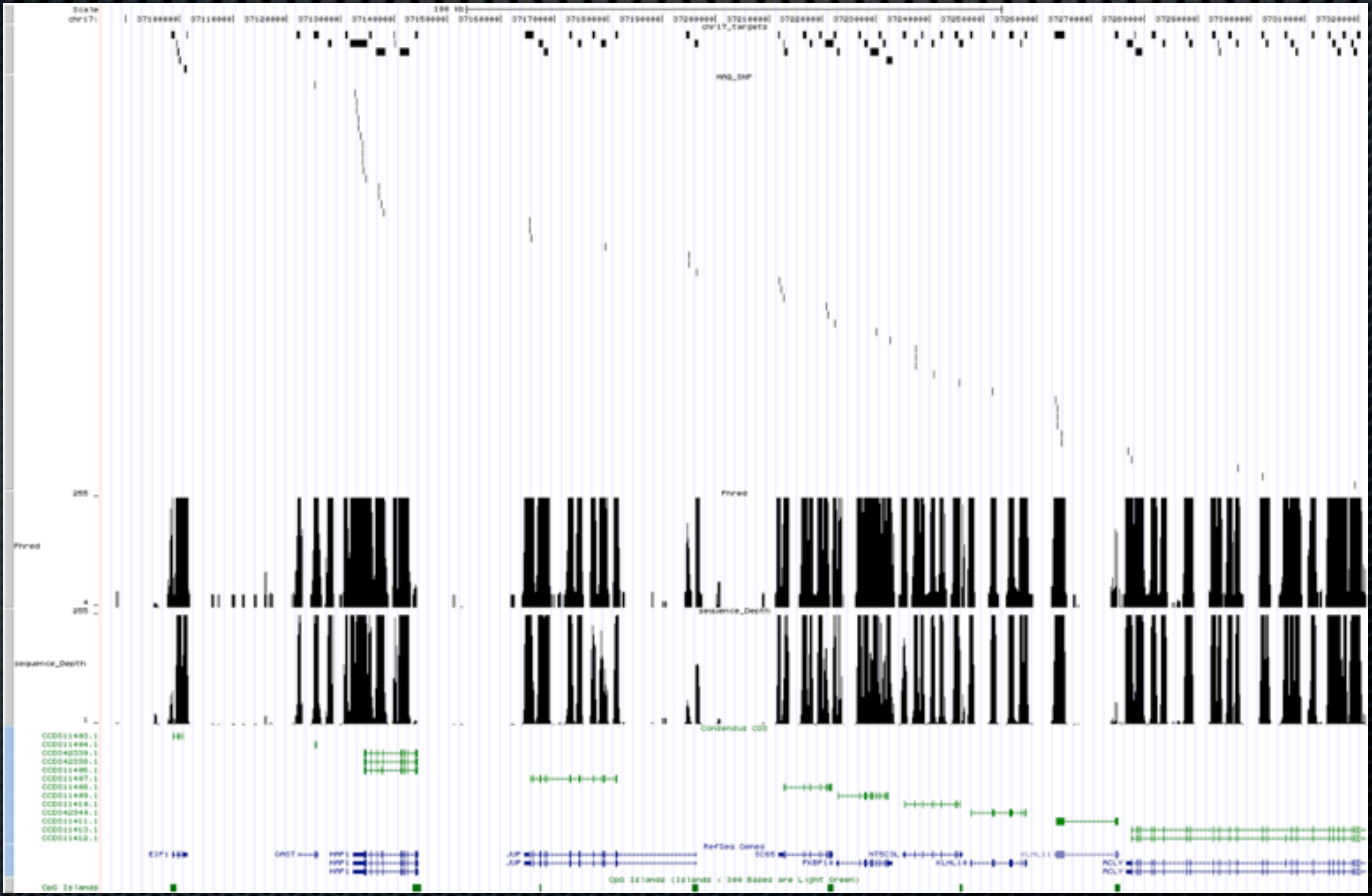
The Galaxy team is a part of BX at Penn State.

This project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, and The Institute for CyberScience at Penn State.

Galaxy build: $Rev 4065:eacdaf59c8de$

http://main.g2.bx.psu.edu/

Viewing data

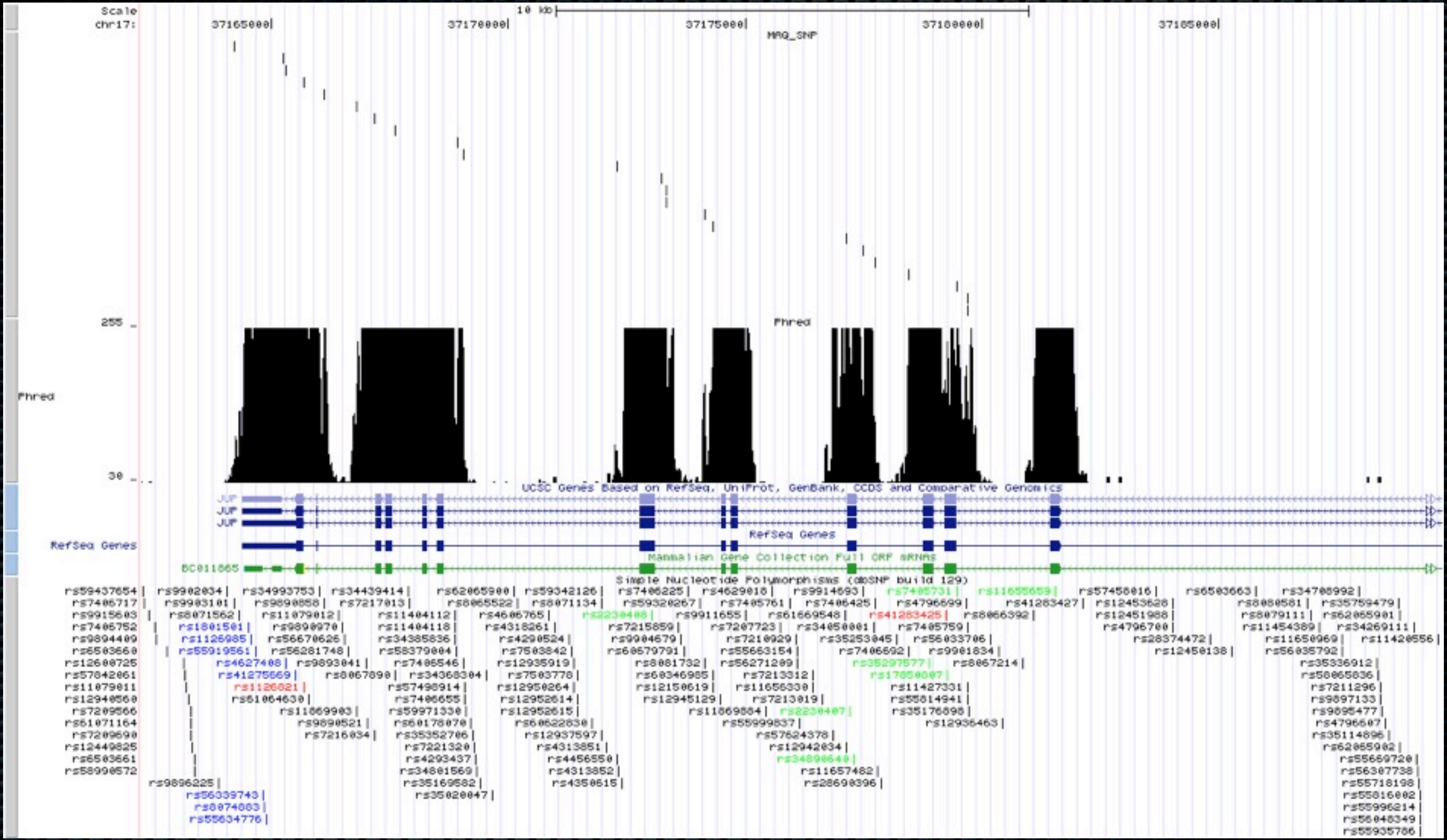


SNP locations

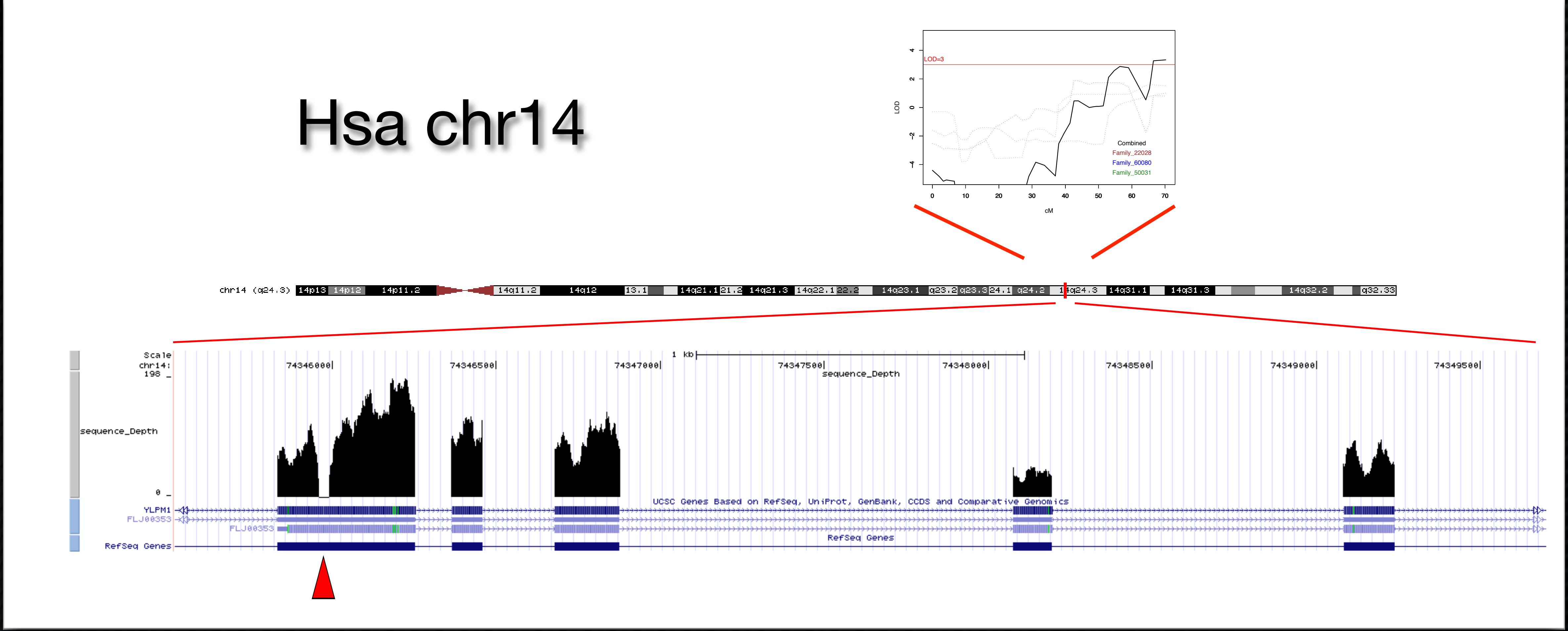
Quality score

Read depth

Genes

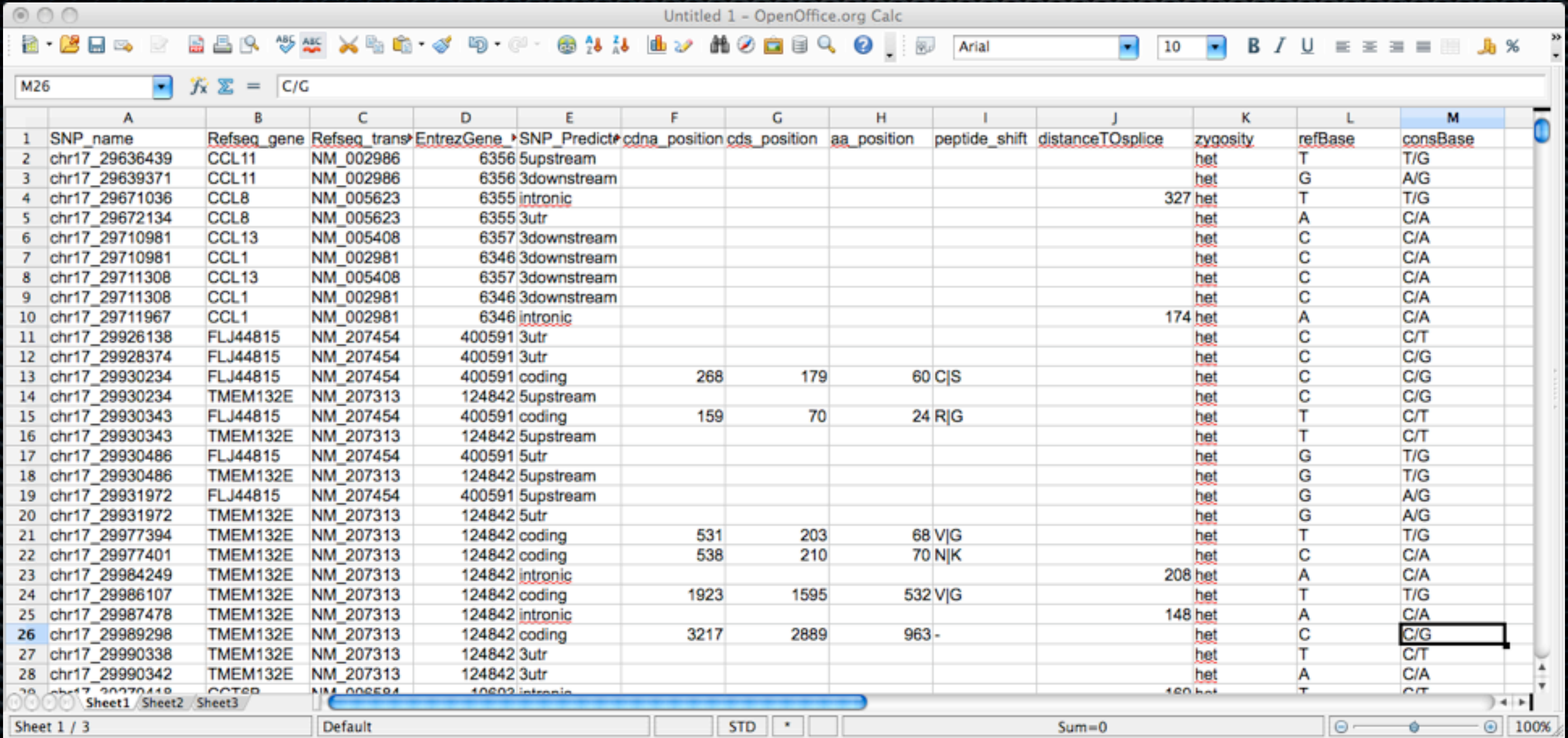

Scale
chr17:
10 kb
37165000
37170000
37175000
37180000
37185000
MAQ_SNP
255
Phred
30
Phred
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
JUP
JUP
JUP
RefSeq Genes
RefSeq Genes
BC011865
Mammalian Gene Collection Full ORF mRNAs
Simple Nucleotide Polymorphisms (dbSNP build 129)

Analyzing resequencing data


Hsa chr14

13 bp insertion

Variants file



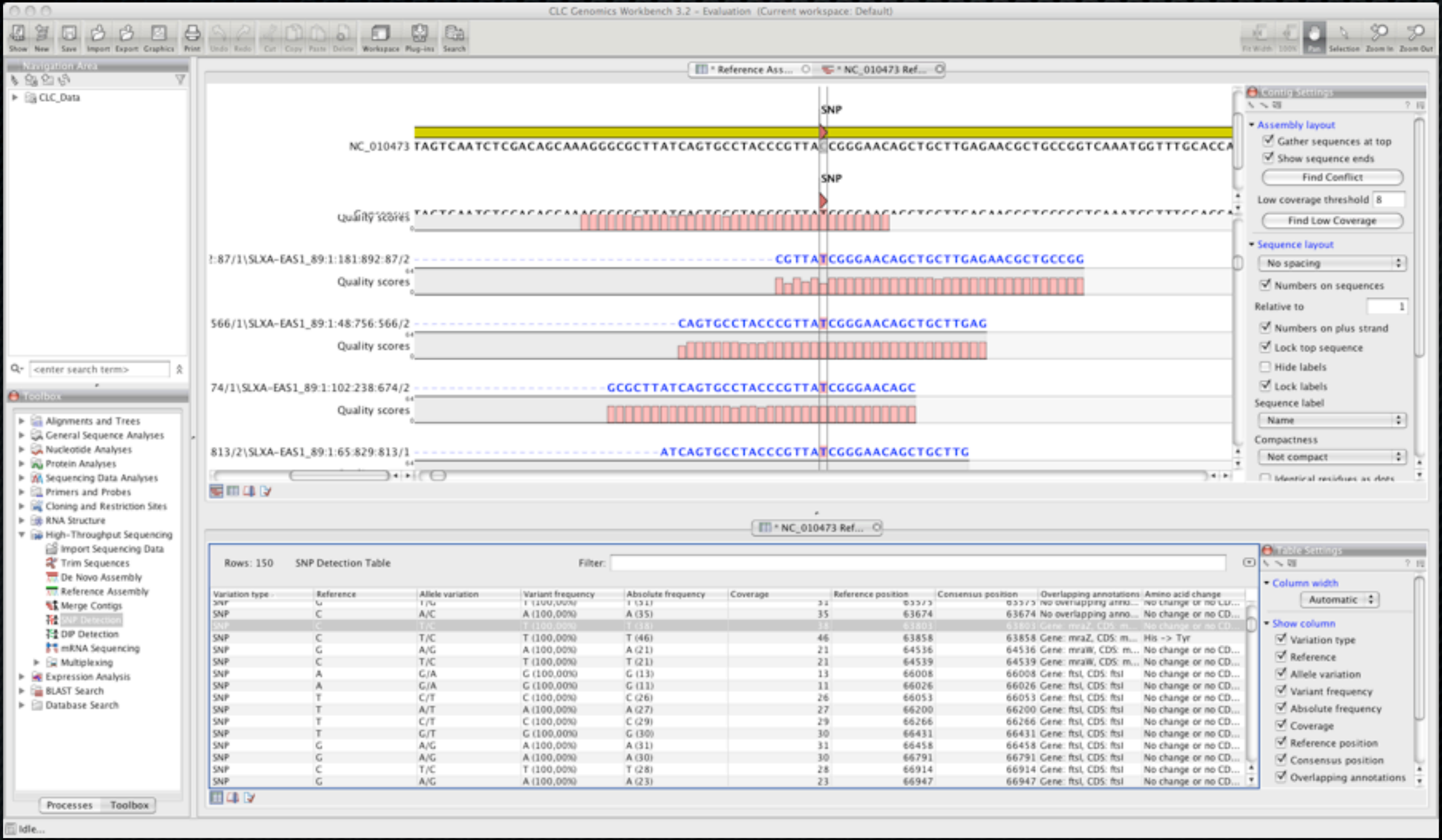
Untitled 1 - OpenOffice.org Calc

M26 C/G

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	SNP_name	Refseq_gene	Refseq_trans	EntrezGene_	SNP_Predict	cdna_position	cds_position	aa_position	peptide_shift	distanceTOsplice	zygosity	refBase	consBase
2	chr17_29636439	CCL11	NM_002986	6356	5upstream						het	T	T/G
3	chr17_29639371	CCL11	NM_002986	6356	3downstream						het	G	A/G
4	chr17_29671036	CCL8	NM_005623	6355	intronic					327	het	T	T/G
5	chr17_29672134	CCL8	NM_005623	6355	3utr						het	A	C/A
6	chr17_29710981	CCL13	NM_005408	6357	3downstream						het	C	C/A
7	chr17_29710981	CCL1	NM_002981	6346	3downstream						het	C	C/A
8	chr17_29711308	CCL13	NM_005408	6357	3downstream						het	C	C/A
9	chr17_29711308	CCL1	NM_002981	6346	3downstream						het	C	C/A
10	chr17_29711967	CCL1	NM_002981	6346	intronic					174	het	A	C/A
11	chr17_29926138	FLJ44815	NM_207454	400591	3utr						het	C	C/T
12	chr17_29928374	FLJ44815	NM_207454	400591	3utr						het	C	C/G
13	chr17_29930234	FLJ44815	NM_207454	400591	coding	268	179	60	C|S		het	C	C/G
14	chr17_29930234	TMEM132E	NM_207313	124842	5upstream						het	C	C/G
15	chr17_29930343	FLJ44815	NM_207454	400591	coding	159	70	24	R|G		het	T	C/T
16	chr17_29930343	TMEM132E	NM_207313	124842	5upstream						het	T	C/T
17	chr17_29930486	FLJ44815	NM_207454	400591	5utr						het	G	T/G
18	chr17_29930486	TMEM132E	NM_207313	124842	5upstream						het	G	T/G
19	chr17_29931972	FLJ44815	NM_207454	400591	5upstream						het	G	A/G
20	chr17_29931972	TMEM132E	NM_207313	124842	5utr						het	G	A/G
21	chr17_29977394	TMEM132E	NM_207313	124842	coding	531	203	68	V|G		het	T	T/G
22	chr17_29977401	TMEM132E	NM_207313	124842	coding	538	210	70	N|K		het	C	C/A
23	chr17_29984249	TMEM132E	NM_207313	124842	intronic					208	het	A	C/A
24	chr17_29986107	TMEM132E	NM_207313	124842	coding	1923	1595	532	V|G		het	T	T/G
25	chr17_29987478	TMEM132E	NM_207313	124842	intronic					148	het	A	C/A
26	chr17_29989298	TMEM132E	NM_207313	124842	coding	3217	2889	963	-		het	C	C/G
27	chr17_29990338	TMEM132E	NM_207313	124842	3utr						het	T	C/T
28	chr17_29990342	TMEM132E	NM_207313	124842	3utr						het	A	C/A

Identifying relevant variants is the hard part

CLC genomics workbench



CLC Genomics Workbench 3.2 – Evaluation (Current workspace: Default)

```
NC_010473 TAGTCAATCTCGACAGCAAAGGGCGCTTATCAGTGCCTACCCGTTACCGGGAACAGCTGCTTGAGAACGCTGCCGGTCAAATGGTTTGCACCA
:87/1\SLXA-EAS1_89:1:181:892:87/2   CGTTATCGGGAACAGCTGCTTGAGAACGCTGCCGG
566/1\SLXA-EAS1_89:1:48:756:566/2   CAGTGCCTACCCGTTATCGGGAACAGCTGCTTGAG
74/1\SLXA-EAS1_89:1:102:238:674/2   GCGCTTATCAGTGCCTACCCGTTATCGGGAACAGC
813/2\SLXA-EAS1_89:1:65:829:813/1   ATCAGTGCCTACCCGTTATCGGGAACAGCTGCTTG
```

Contig Settings

Assembly layout
- ☑ Gather sequences at top
- ☑ Show sequence ends
- Find Conflict
- Low coverage threshold: 8
- Find Low Coverage

Sequence layout
- No spacing
- ☑ Numbers on sequences
- Relative to: 1
- ☑ Numbers on plus strand
- ☑ Lock top sequence
- ☐ Hide labels
- ☑ Lock labels
- Sequence label: Name
- Compactness: Not compact
- ☐ Identical residues as dots

Rows: 150 SNP Detection Table Filter:

Variation type	Reference	Allele variation	Variant frequency	Absolute frequency	Coverage	Reference position	Consensus position	Overlapping annotations	Amino acid change
SNP	G	T/G	T (100,00%)	T (31)	31	63575	63575	No overlapping anno...	No change or no CD...
SNP	C	A/C	A (100,00%)	A (35)	35	63674	63674	No overlapping anno...	No change or no CD...
SNP	C	T/C	T (100,00%)	T (33)	33	63803	63803	Gene: mraZ, CDS: m...	No change or no CD...
SNP	C	T/C	T (100,00%)	T (46)	46	63858	63858	Gene: mraZ, CDS: m...	His -> Tyr
SNP	G	A/G	A (100,00%)	A (21)	21	64536	64536	Gene: mraW, CDS: m...	No change or no CD...
SNP	C	T/C	T (100,00%)	T (21)	21	64539	64539	Gene: mraW, CDS: m...	No change or no CD...
SNP	A	G/A	G (100,00%)	G (13)	13	66008	66008	Gene: ftsI, CDS: ftsI	No change or no CD...
SNP	A	G/A	G (100,00%)	G (11)	11	66026	66026	Gene: ftsI, CDS: ftsI	No change or no CD...
SNP	T	C/T	C (100,00%)	C (26)	26	66053	66053	Gene: ftsI, CDS: ftsI	No change or no CD...
SNP	T	A/T	A (100,00%)	A (27)	27	66200	66200	Gene: ftsI, CDS: ftsI	No change or no CD...
SNP	T	C/T	C (100,00%)	C (29)	29	66266	66266	Gene: ftsI, CDS: ftsI	No change or no CD...
SNP	T	G/T	G (100,00%)	G (30)	30	66431	66431	Gene: ftsI, CDS: ftsI	No change or no CD...
SNP	G	A/G	A (100,00%)	A (31)	31	66458	66458	Gene: ftsI, CDS: ftsI	No change or no CD...
SNP	G	A/G	A (100,00%)	A (30)	30	66791	66791	Gene: ftsI, CDS: ftsI	No change or no CD...
SNP	C	T/C	T (100,00%)	T (28)	28	66914	66914	Gene: ftsI, CDS: ftsI	No change or no CD...
SNP	G	A/G	A (100,00%)	A (23)	23	66947	66947	Gene: ftsI, CDS: ftsI	No change or no CD...

Table Settings

Column width
- Automatic

Show column
- ☑ Variation type
- ☑ Reference
- ☑ Allele variation
- ☑ Variant frequency
- ☑ Absolute frequency
- ☑ Coverage
- ☑ Reference position
- ☑ Consensus position
- ☑ Overlapping annotations

Commercial software

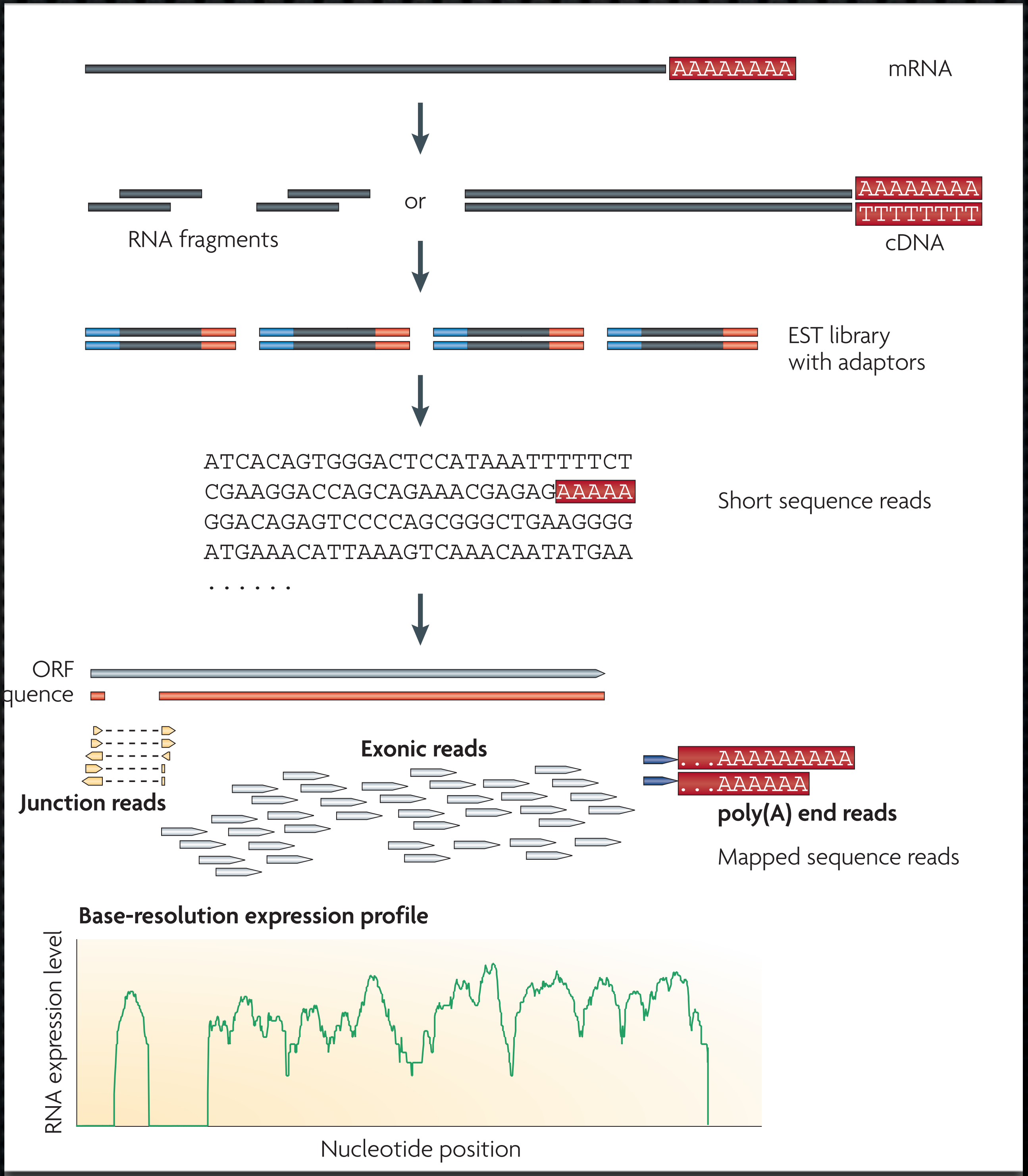
Detecting all variants

VARIANT	SINGLE READ	SHORT INSERT PAIRED-ENDS (200–500 bp)	LONG INSERT MATE PAIRS (2–5 kb)	PAIRED-END AND MATE PAIR COMBINED
SNP	++	++++	++	++++
Small indels	++	++++	++	++++
Insertion	+	+++	+++	++++
Amplification	++	+++	+++	++++
Deletion	+	+++	++	++++
Inversion	+	+++	++	++++
Complex rearrangement	+	+++	++	++++
Large rearrangement	+	++	+++	++++

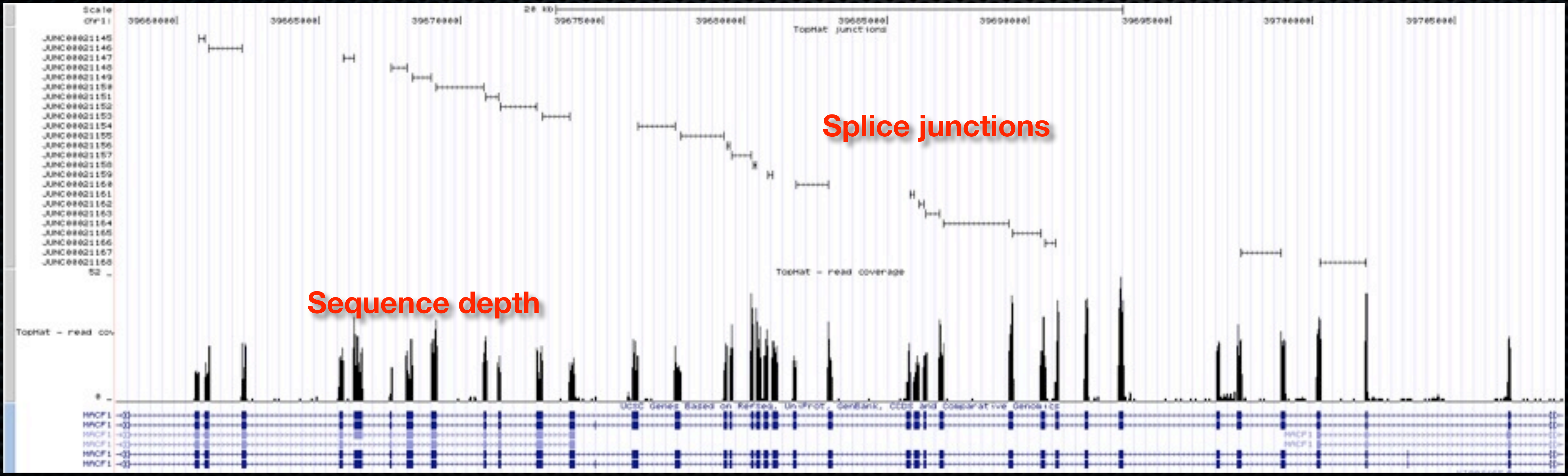
Region unknown

- Sequence capture - exome: sequence all exons

- RNAseq
- Sequence total polyA RNA
- Map reads to reference
- Identify mutations/variants



RNAseq data



Splice junctions

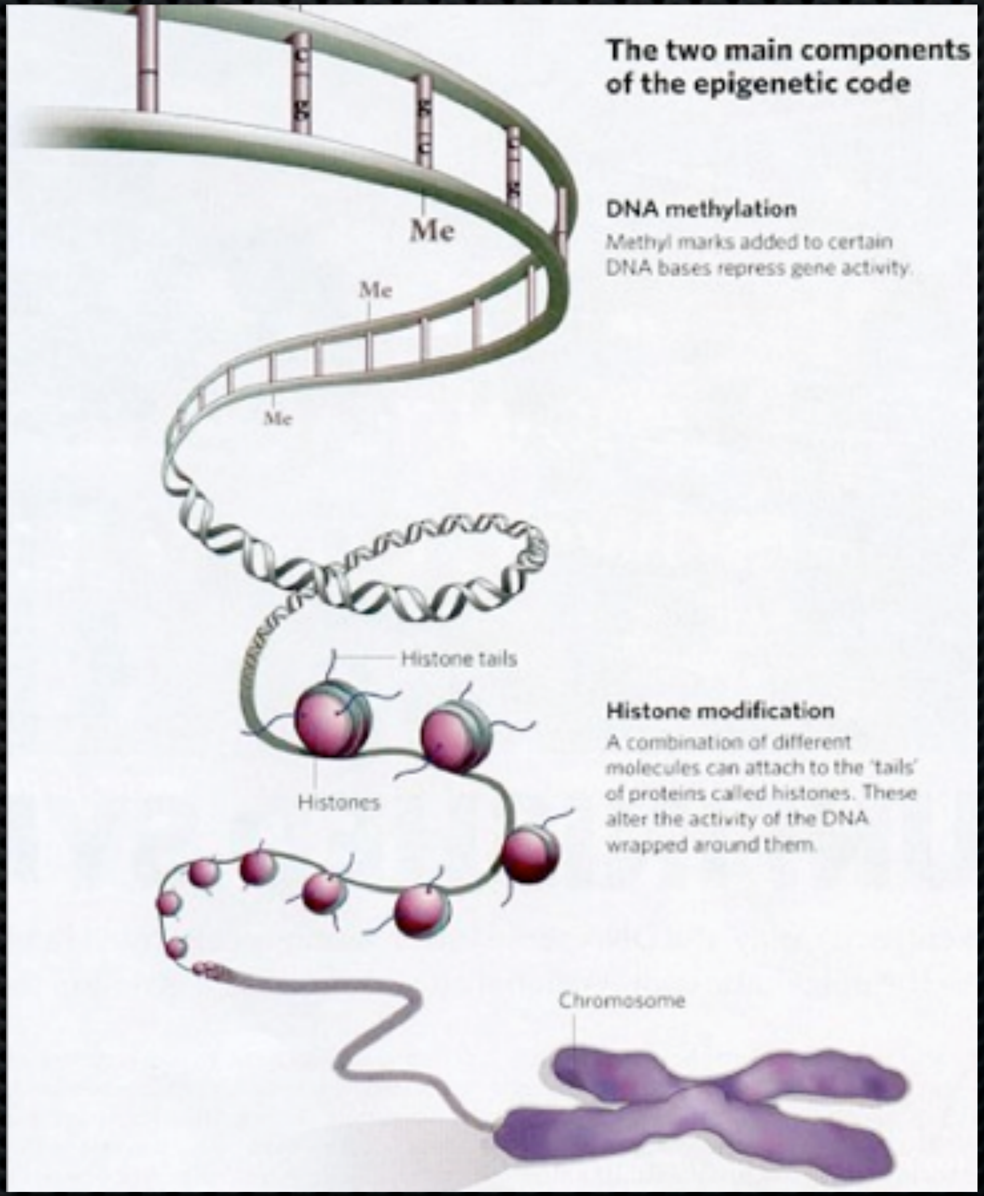
Sequence depth

Position along Hsa chr1

Identifying relevant variants is the hard part

Epigenetics

- DNA methylation
  - CpG dinucleotides
- Histone modifications
  - acetylation
  - phosphorylation
  - methylation
  - ubiquitination



The two main components of the epigenetic code

DNA methylation

Methyl marks added to certain DNA bases repress gene activity.

Histone modification

A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.

→ Control of gene expression

Epigenetics II

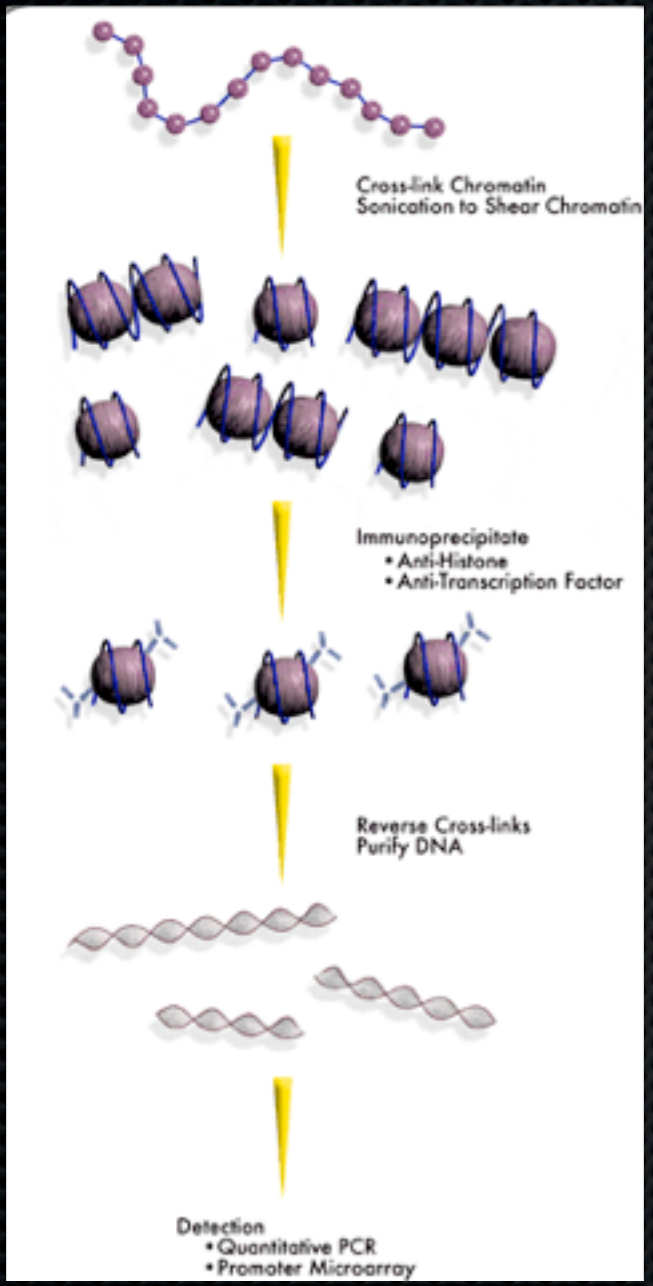
- DNA methylation
  - Long-term epigenetic silencing of specific sequences
  - transposons, imprinted genes, pluripotency genes

- Histone modifications
  - Short term, flexible epigenetic control

→ Control of gene expression

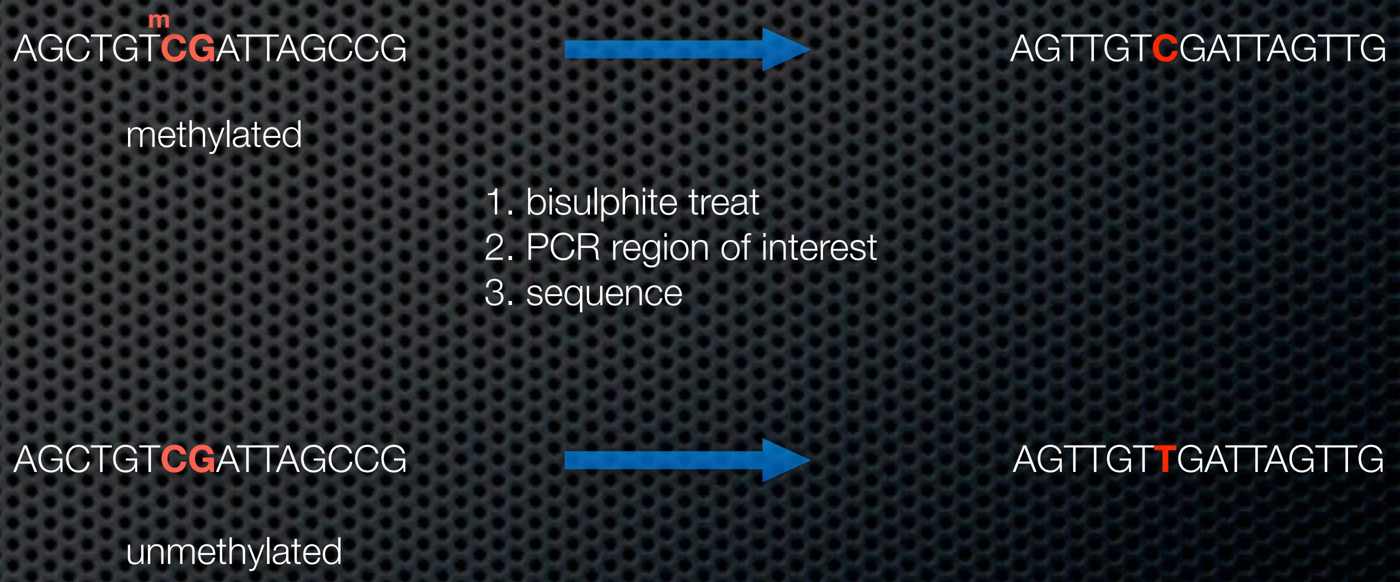
HTS and epigenetics

ChIP
chromatin immunoprecipitation



Quantifying DNA methylation

Bisulphite sequencing (BiS)



AGCTGT^{m}**CG**ATTAGCCG → AGTTGT**C**GATTAGTTG

methylated

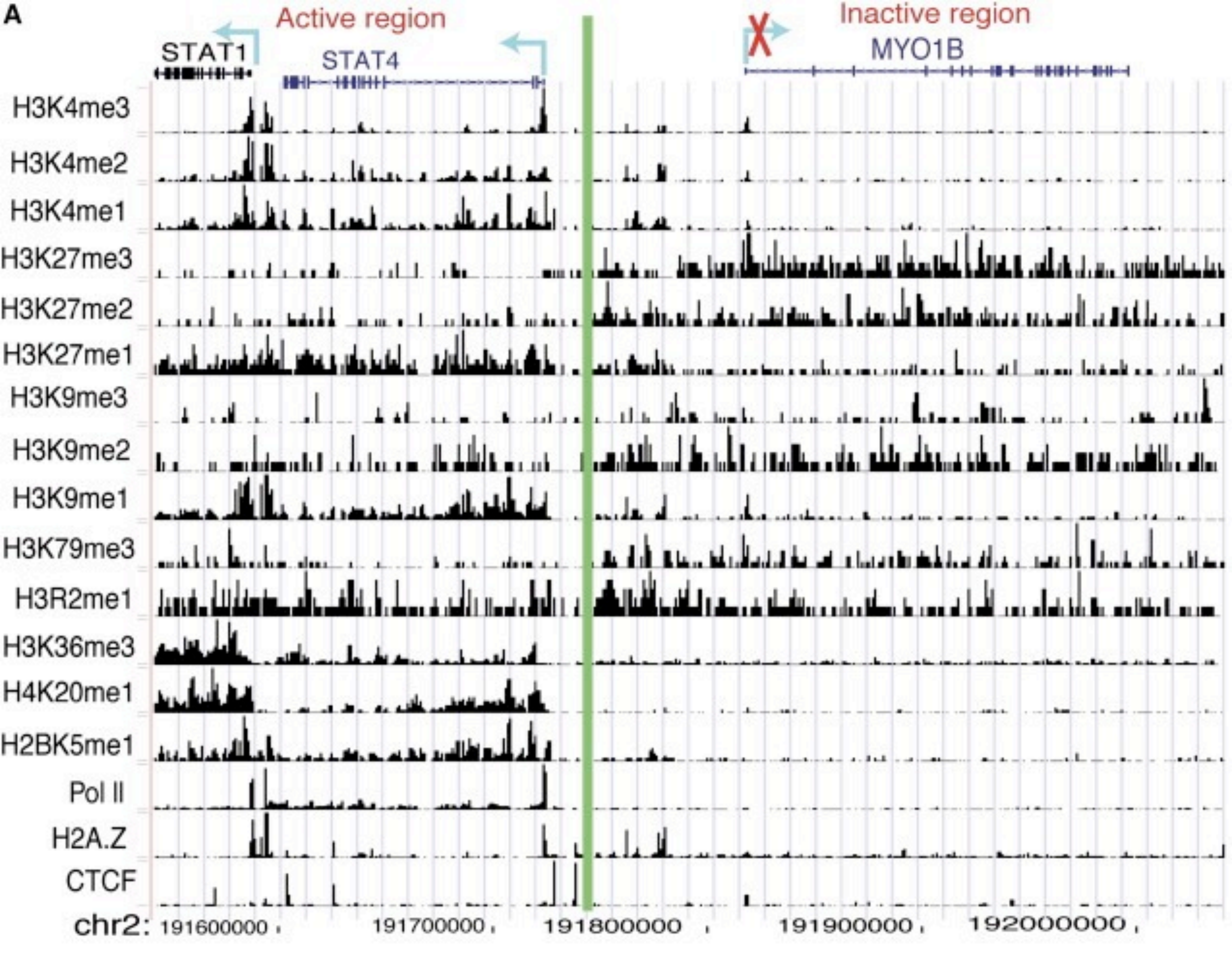
1. bisulphite treat
2. PCR region of interest
3. sequence

AGCTGT**CG**ATTAGCCG → AGTTGT**T**GATTAGTTG

unmethylated

HTS to identify genome-wide status/variation

ChIP-seq example



Barski et al, Cell 129, 823–837, May 18, 2007

Summary

- High-throughput sequencing
  - Dramatic increase in sequence production
  - Many applications on one platform
  - Field new and moving very quickly
- Bioinformatics challenges/opportunities
  - Data storage
  - Data analysis

Visit?

Robert.Lyle@medisin.uio.no