Visualizing NGS data with GenomeView

June 2nd, 2010, Espoo, Finland

Thomas Abeel – thomas@abeel.be
VIB - Ghent University, Gent, Belgium
Broad Institute of MIT and Harvard, Cambridge, MA, USA

Overview

- Introduction to NGS
  - What is NGS? Platforms?
  - Applications of NGS:
    - Assembly
    - Mapping: RNA-seq, chip-seq, resequencing, etc.
- Visualizing NGS and other data
  - Description of GenomeView

NGS INTRODUCTION

Next-gen sequencing

- Next-generations sequencing machine
  - ABI Solid, Illumina (Solexa), and Roche 454
  - Helicos Biosciences, Pacific Biosciences
- Pour DNA in machine and you get reads
- Read = fragment of sequenced DNA

Gen.	**1st**	**next**	**next**	**next**
Platform	Sanger	Roche 454	Illumina	SOLiD
Data/run	100kb	500 Mb	8 Gb	16Gb
Read len.	1000 nt	500 nt	100 nt	50 nt
$/genome	1,000k	25k	25k	25k

1st, next, 3rd?

- Discussion what's 'next'-generation?

Generation	Platforms	Status
1	Sanger	Production
2 (next)	454	Production
2.5 (next)	Illumina, Solid	Production
3	Helicos, Pacbio	First commercial models available
3.x (not available)	Nanopore, ….	Prototype

- For ease of discussion, NGS will be used for anything not Sanger seq.

What to do with short reads

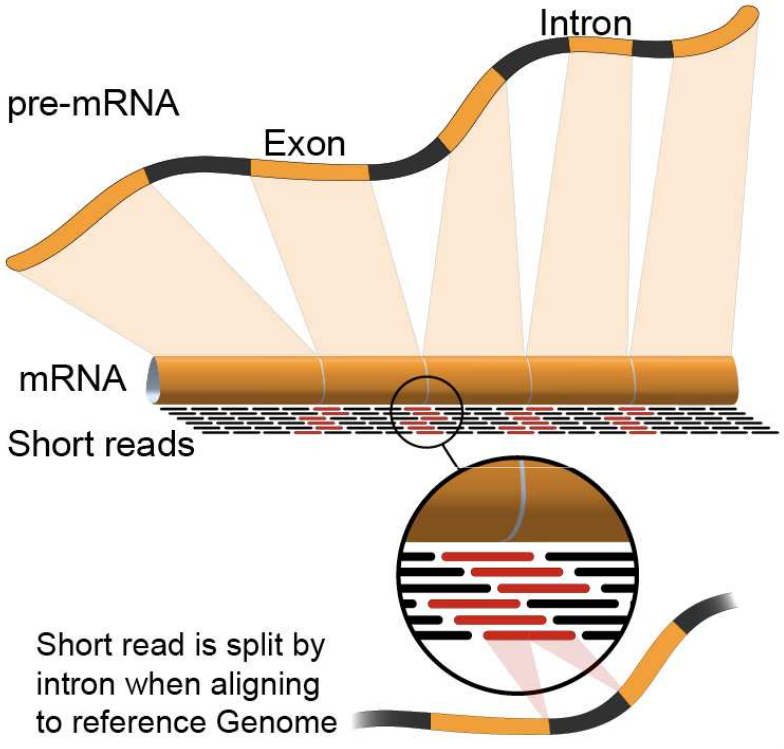
- Assembly
  - At least 40x coverage
  - Result is large number of contigs
- **Map to a reference genome**
  - **Re-sequencing → SNP discovery**
  - **ChIP-seq**
  - **RNA-seq**

Re-sequencing

- Sequence the genome of an organism for which you already have a genomic sequence
  - Genomic diversity: SNPs, indels, but also structural variation

RNA-seq

- Whole genome transcriptome sequencing
- Extract mRNA by picking up the poly-A tail
- Sequence cDNA



RNA-seq applications

- Gene expression
  - If a read maps to a gene, that gene is expressed
- Expression measure
  - If 10 reads map to the same place in a gene, it's 10 fold expressed
- Condition/cell type specific expression
- → Microarray replacement
- Annotation

ChIP-seq

- ChIP produces a library of target DNA to which a protein of interest binds in-vivo
- -seq sequences the library
- Less bias than ChIP-on-chip because not limited to probes
- Application:
  - Find transcription factor binding sites

Short read mapping

- Each of the proposed applications boil down to aligning the reads to a reference sequence
- Characteristics:
  - Many short queries (reads)
  - Single large reference (reference)
  - Mismatches allowed

Alignment challenges

- Efficiency
  - Need to align several billion reads (300 Gb)
  - To a genome of several billion nucleotides (3 Gb)
  - Preferably over the weekend
- Ambiguity
  - Sequencing errors
  - Genetic variation
  - Alignment with mismatches

Assigning reads

- 36 nt → with a couple of mismatches reads may align to multiple locations in the genome
- Unique: where it maps
- Non-unique:
  - all locations
  - random
  - nowhere

Short read aligners

- Bfast
- BioScope
- Bowtie
- BWA
- CLC bio
- CloudBurst
- Eland/Eland2
- GenomeMapper
- GnuMap
- Karma
- MAQ
- MOM
- Mosaik
- MrFAST/MrsFAST
- NovoAlign
- PASS
- PerM
- RazerS
- RMAP
- SSAHA2
- Segemehl
- SeqMap
- SHRiMP
- Slider/SliderII
- SOAP/SOAP2
- Srprism
- Stampy
- vmatch
- ZOOM
- .....

Slide adapted from Heng Li

Typical NGS analysis pipeline

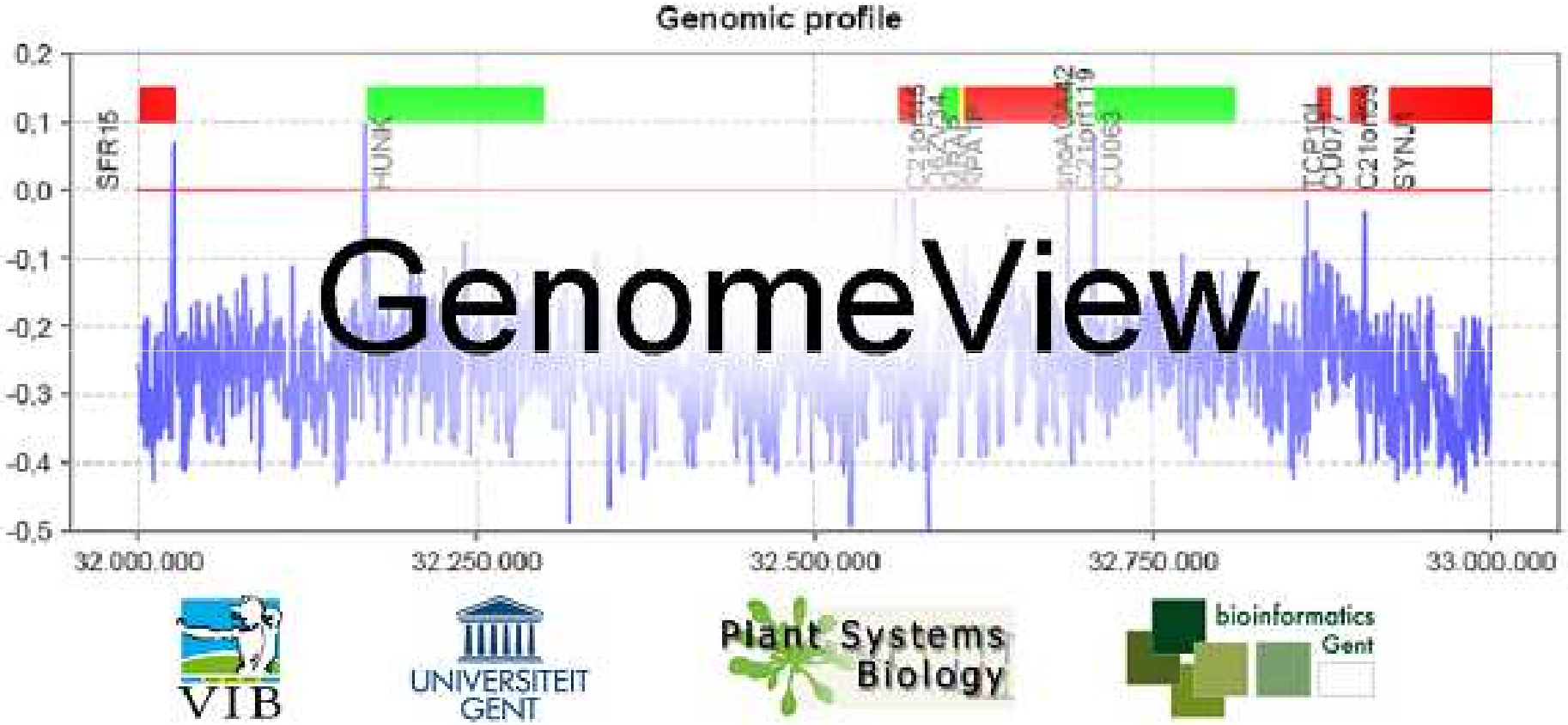
- Extracted DNA (or somebody else did it)
- Put it in machine, got a lot of data in return
- We have aligned it to a reference sequence
- We now have an even bigger file with even more information.

Analyses

- SNP calling
- Peak detection
- Gene prediction
- Transcript identification/quantification
- Visualization

Reasons for visualization

- Taking a look at the data
  - Sanity check on the data
  - Hypothesis generation
- Provide insights in large-scale data sets
  - Augment ability to reason about complex data
  - Make it easier to develop algorithms
  - The appropriate image makes the solution obvious



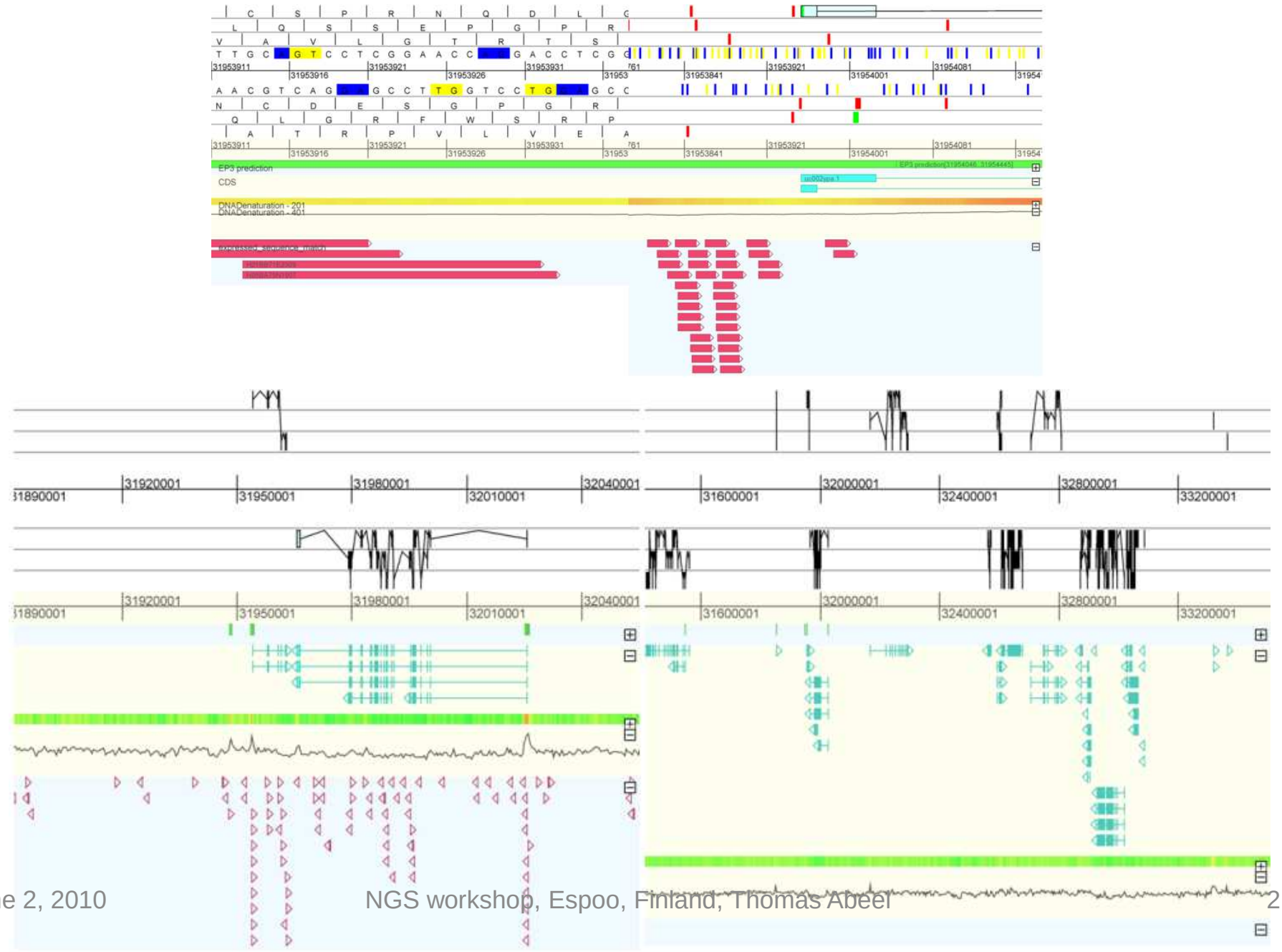
VISUALIZING NGS DATA

GenomeView

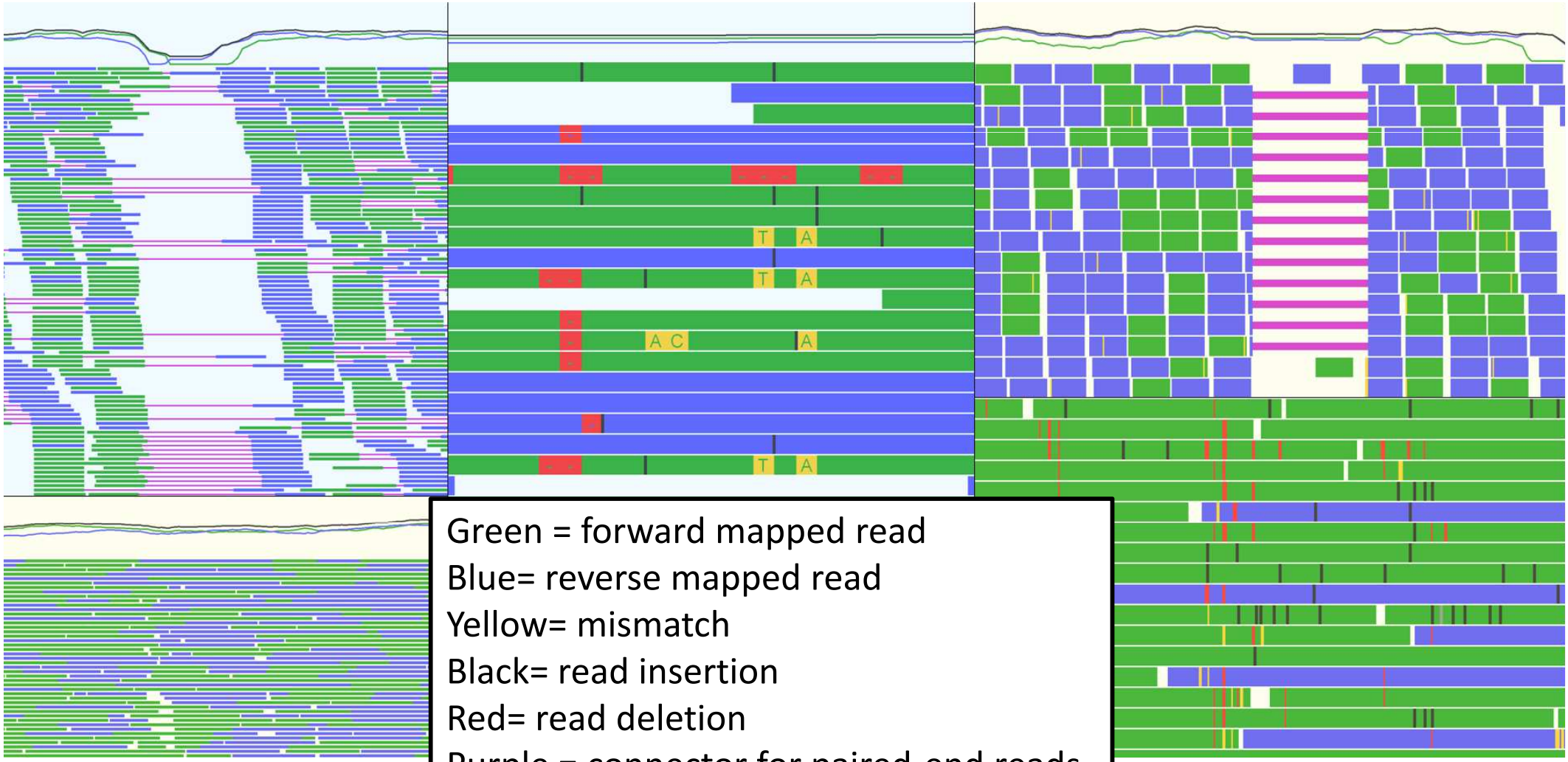
- Interactive genome browser/editor
- Annotations and mapped experimental data
- Short read alignments (*-seq)
- Multiple alignments

GUI Overview

Basic features

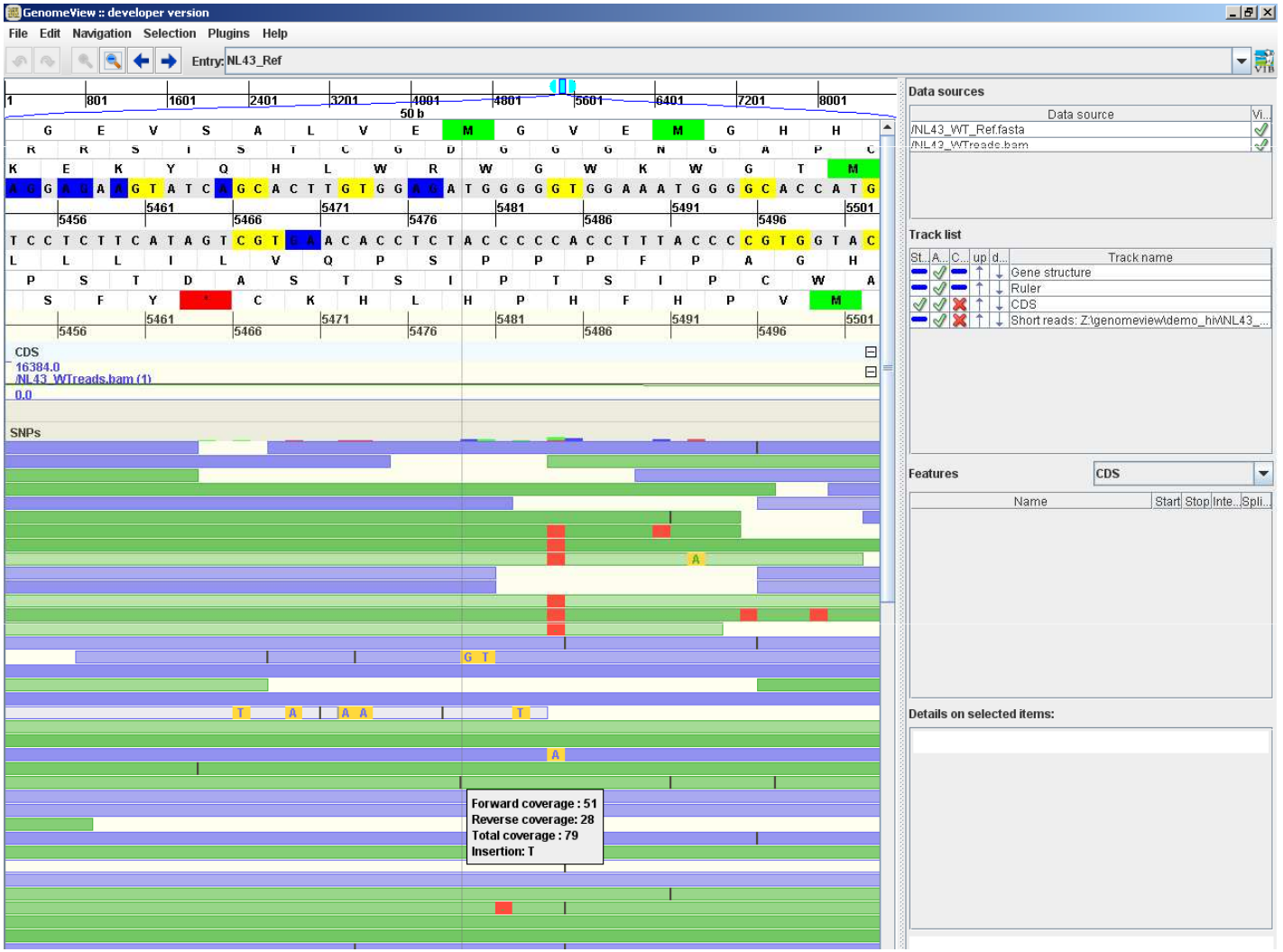


Short read mappings

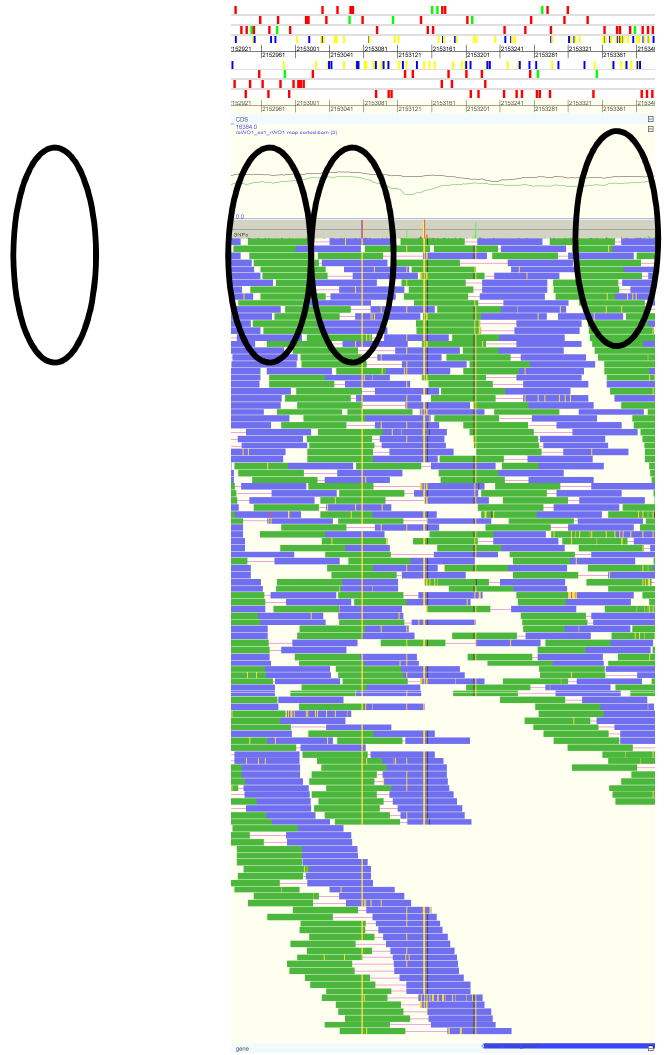


Green = forward mapped read
Blue= reverse mapped read
Yellow= mismatch
Black= read insertion
Red= read deletion
Purple = connector for paired-end reads and spliced reads

Quality, indels and mismatches

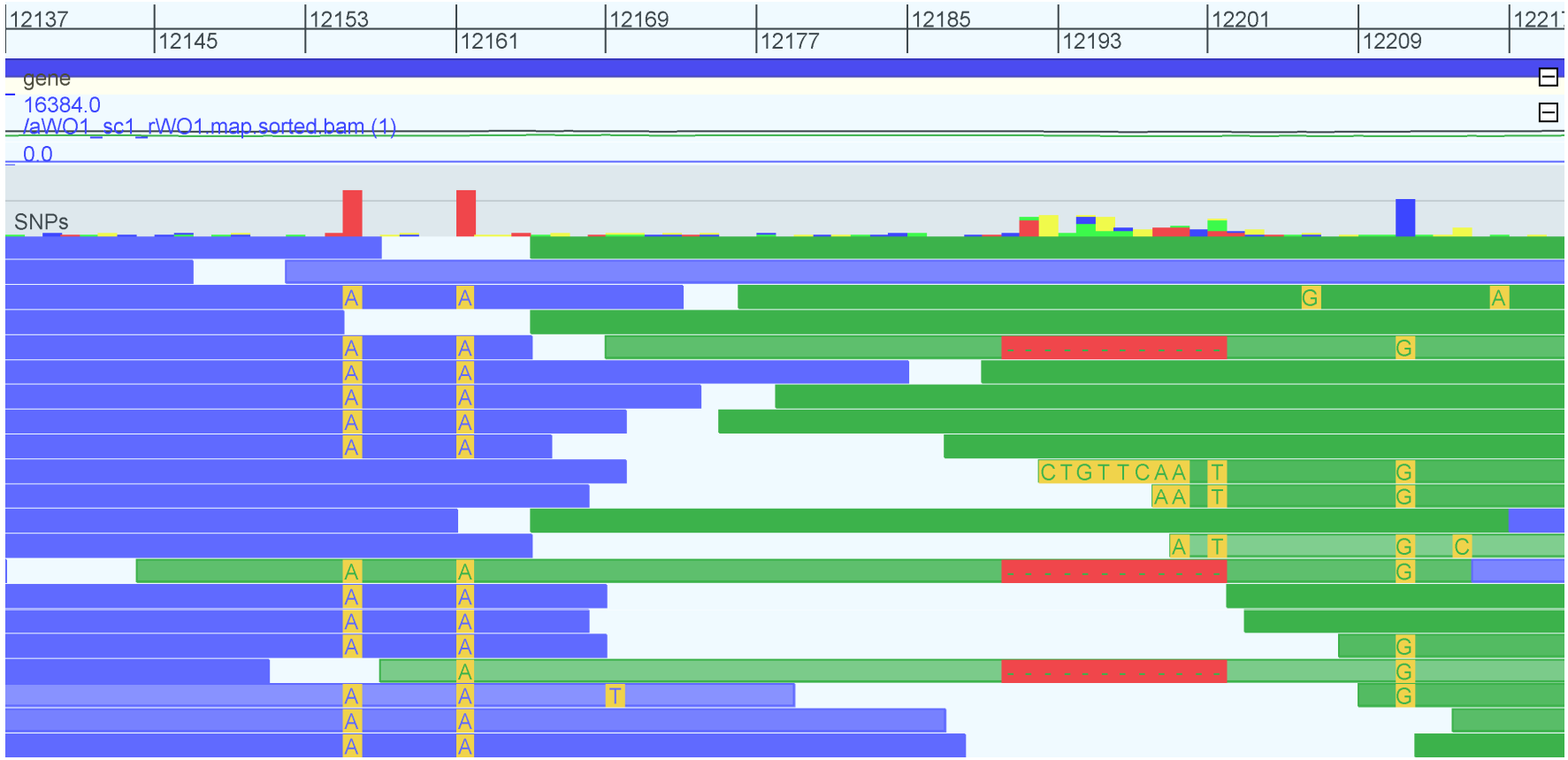


SNP track



Re-sequencing demo

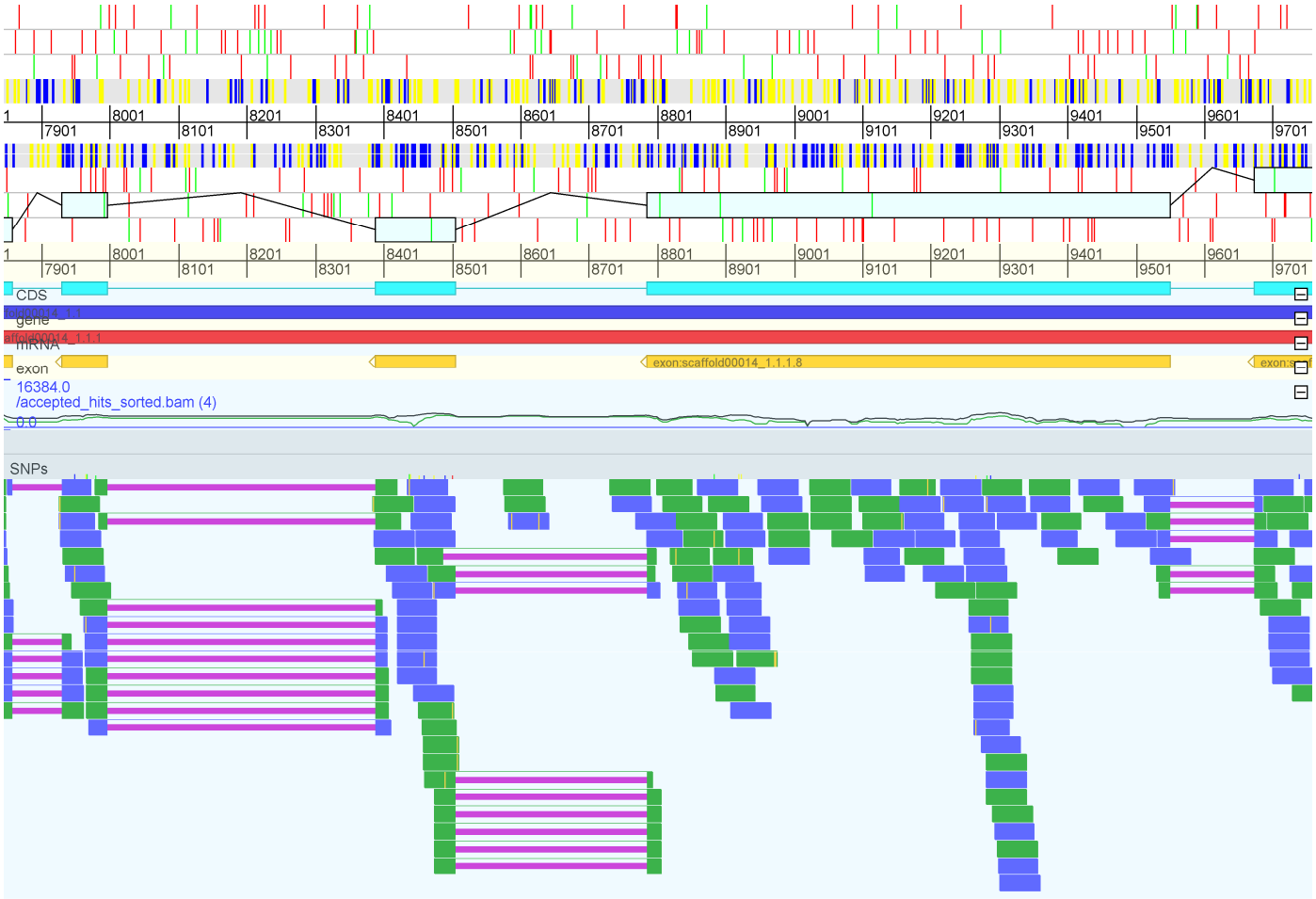
SNPs



http://www.youtube.com/watch?v=KPgARXGbDaM

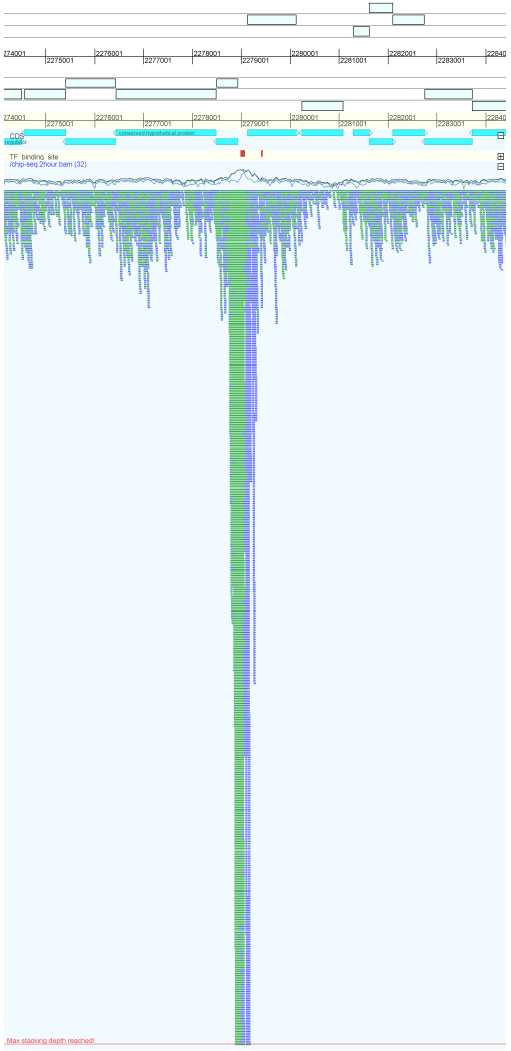
RNA-seq demo

RNA-seq



http://www.youtube.com/watch?v=PgBU1gWCkWU

Chip-seq


2274001
2275001
2276001
2277001
2278001
2279001
2280001
2281001
2282001
2283001
2284
CDS
TF binding site
r/chip-seq.2hour.bam (32)
Max stacking depth reached!

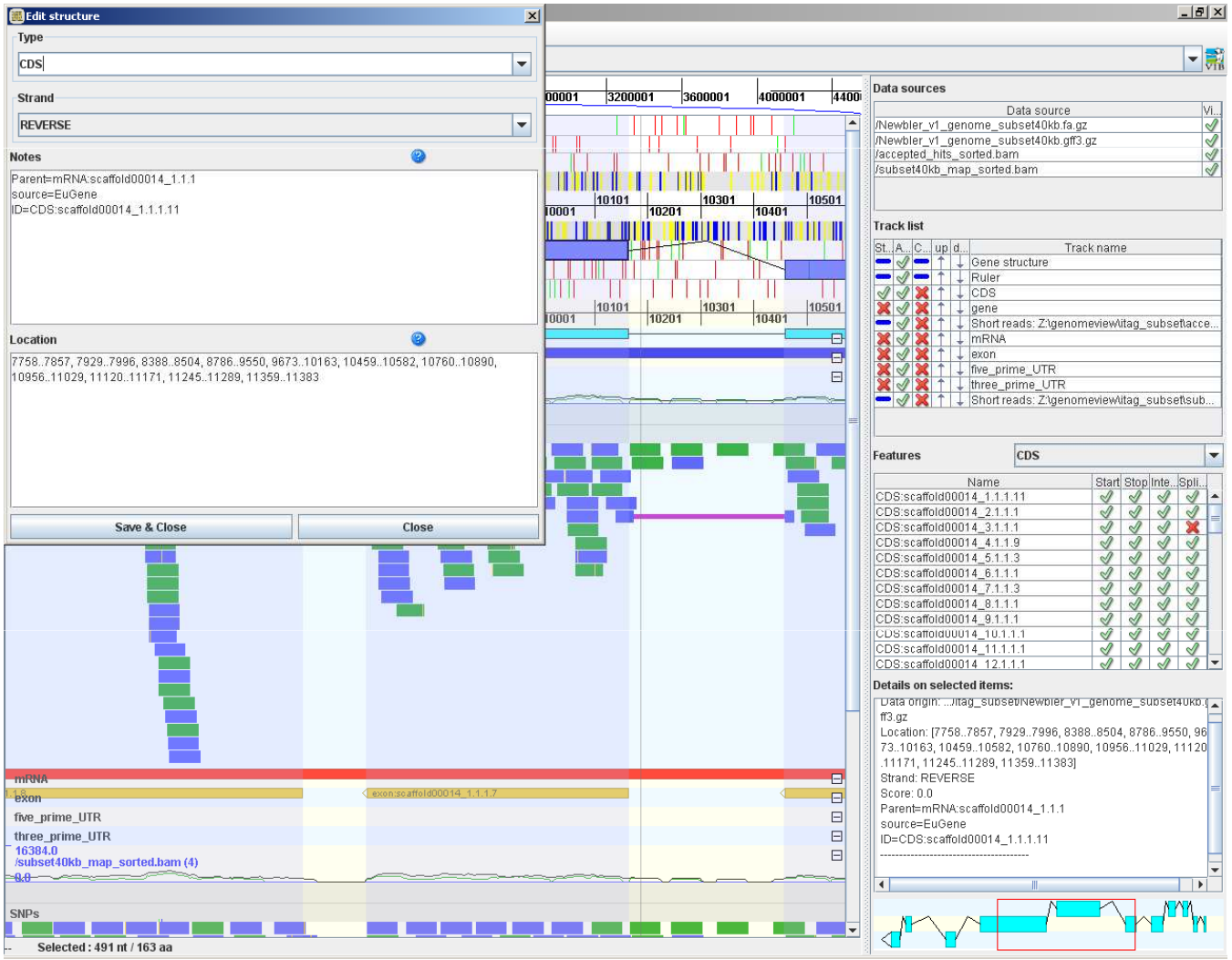
Other features

Plug-ins, editor, integration, multiple alignments

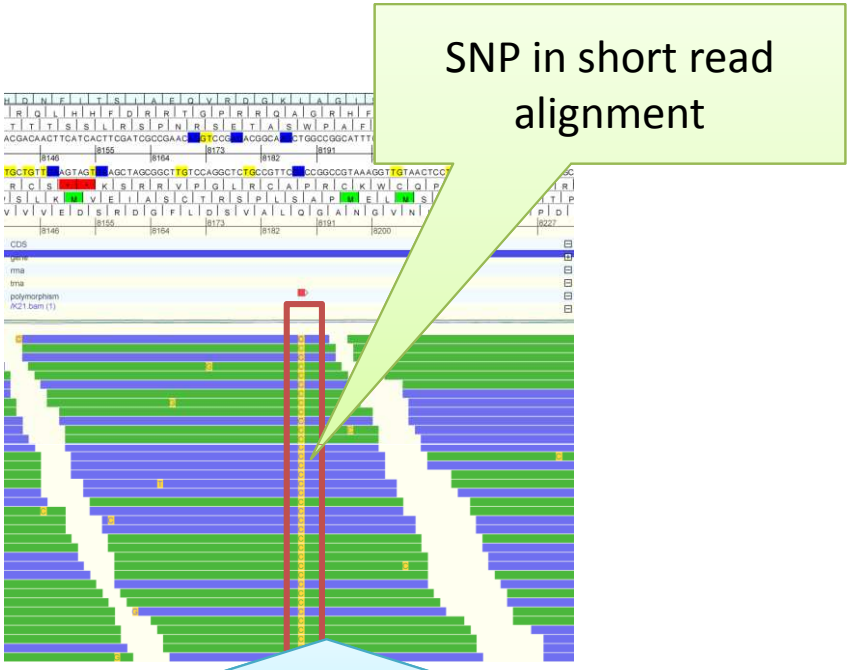
Plug-ins

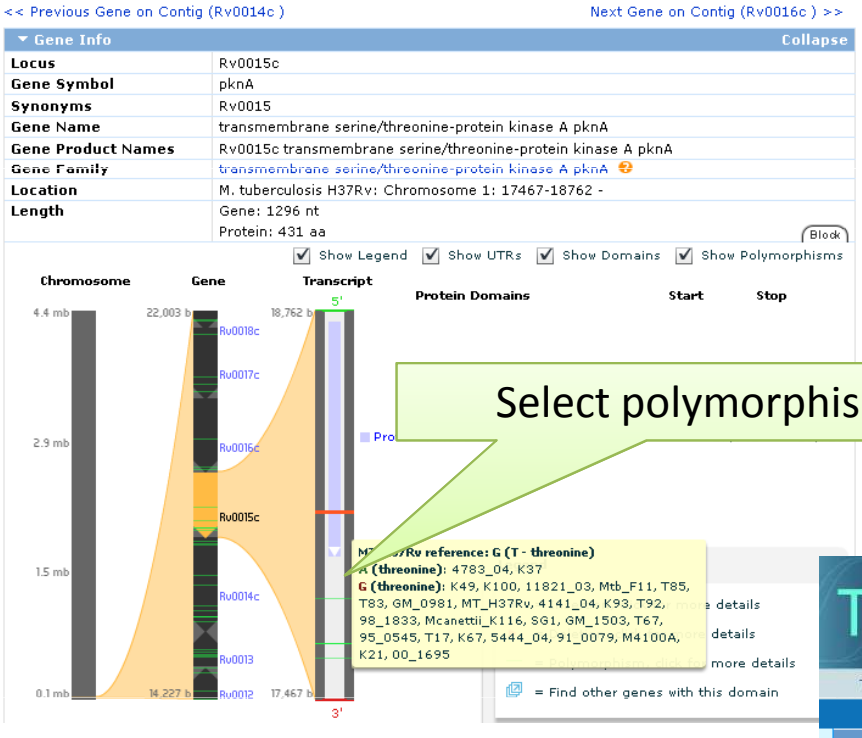
- Computational analyses can be plugged in
  - Gene prediction
    - Core promoter prediction (EP3, Prosom)
    - Splice site prediction (SpliceMachine)
    - Translation start site prediction (StartScan)
    - Coding potential prediction
  - Physical properties of DNA
  - Blast
  - ...

Editor

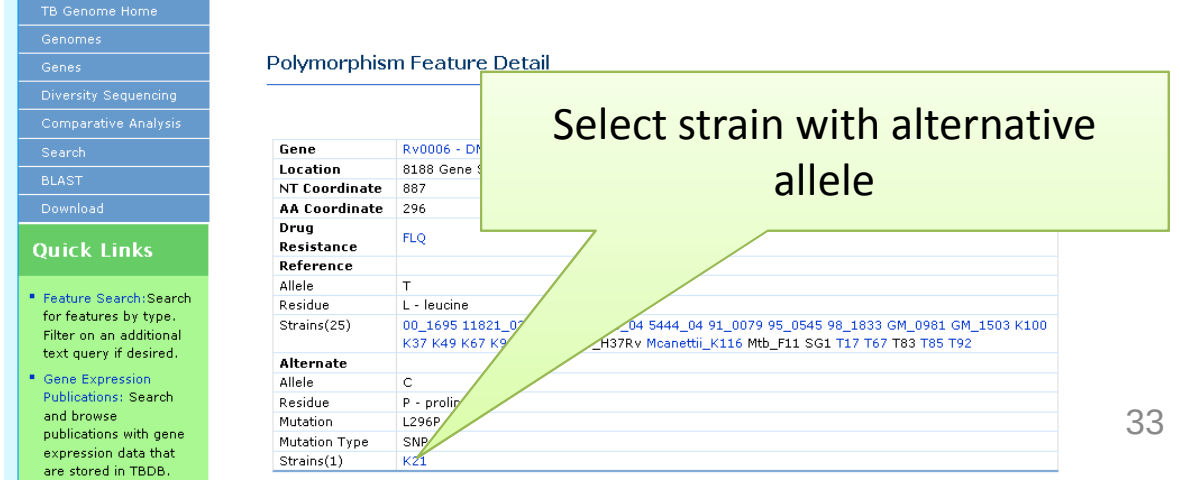

Edit structure
Type
CDS
Strand
REVERSE
Notes
Parent=mRNA:scaffold00014_1.1.1
source=EuGene
ID=CDS:scaffold00014_1.1.1.11
Location
7758..7857, 7929..7996, 8388..8504, 8786..9550, 9673..10163, 10459..10582, 10760..10890, 10956..11029, 11120..11171, 11245..11289, 11359..11383
Save & Close
Close
Data sources
Data source
/Newbler_v1_genome_subset40kb.fa.gz
/Newbler_v1_genome_subset40kb.gff3.gz
/accepted_hits_sorted.bam
/subset40kb_map_sorted.bam
Track list
Track name
Gene structure
Ruler
CDS
gene
mRNA
exon
five_prime_UTR
three_prime_UTR
Features
CDS
Name
Start
Stop
Details on selected items:
Strand: REVERSE
Score: 0.0
Parent=mRNA:scaffold00014_1.1.1
source=EuGene
ID=CDS:scaffold00014_1.1.1.11
mRNA
exon
five_prime_UTR
three_prime_UTR
/subset40kb_map_sorted.bam (4)
SNPs
Selected : 491 nt / 163 aa

Browser integration



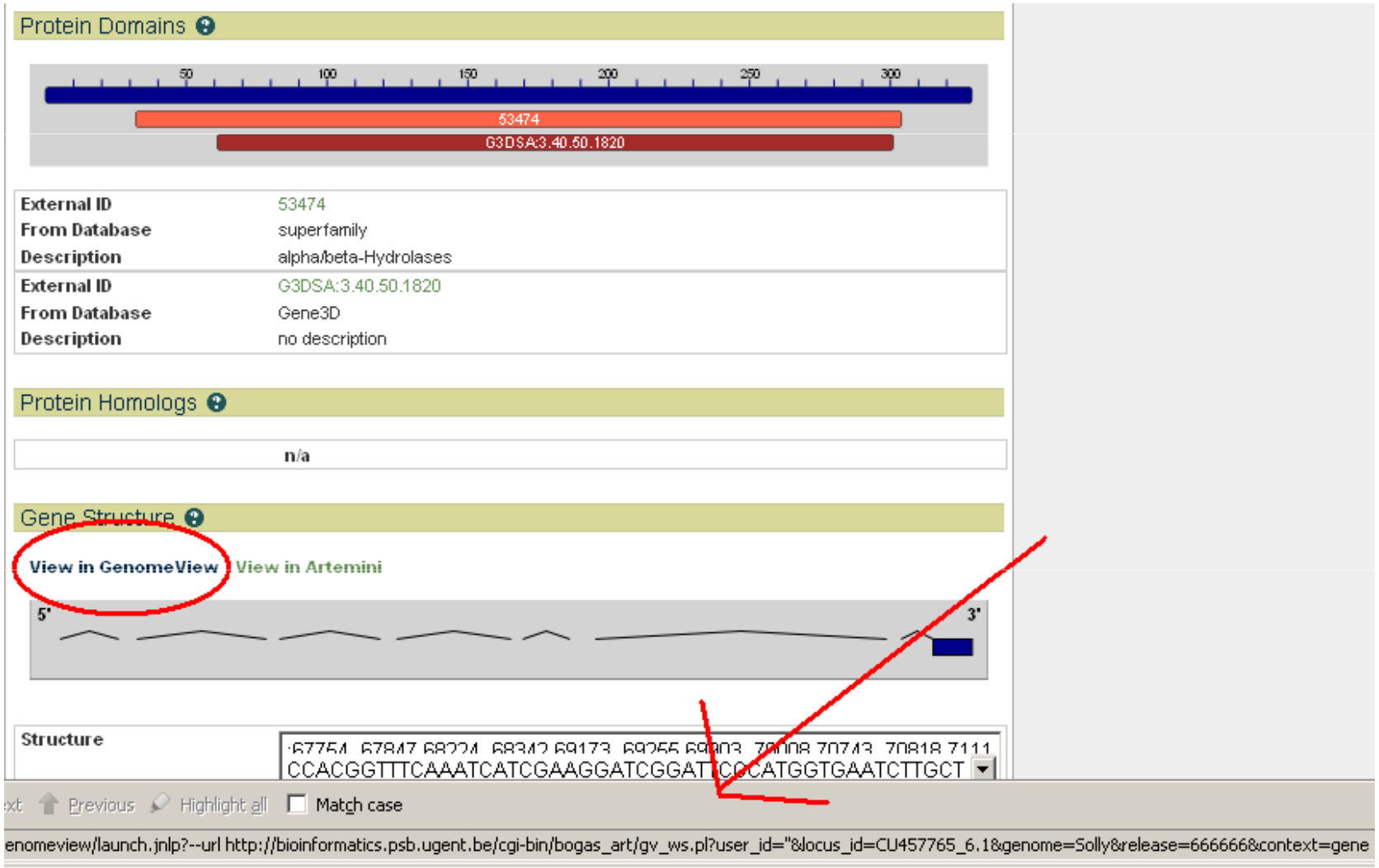


Go to polymorphism detail page

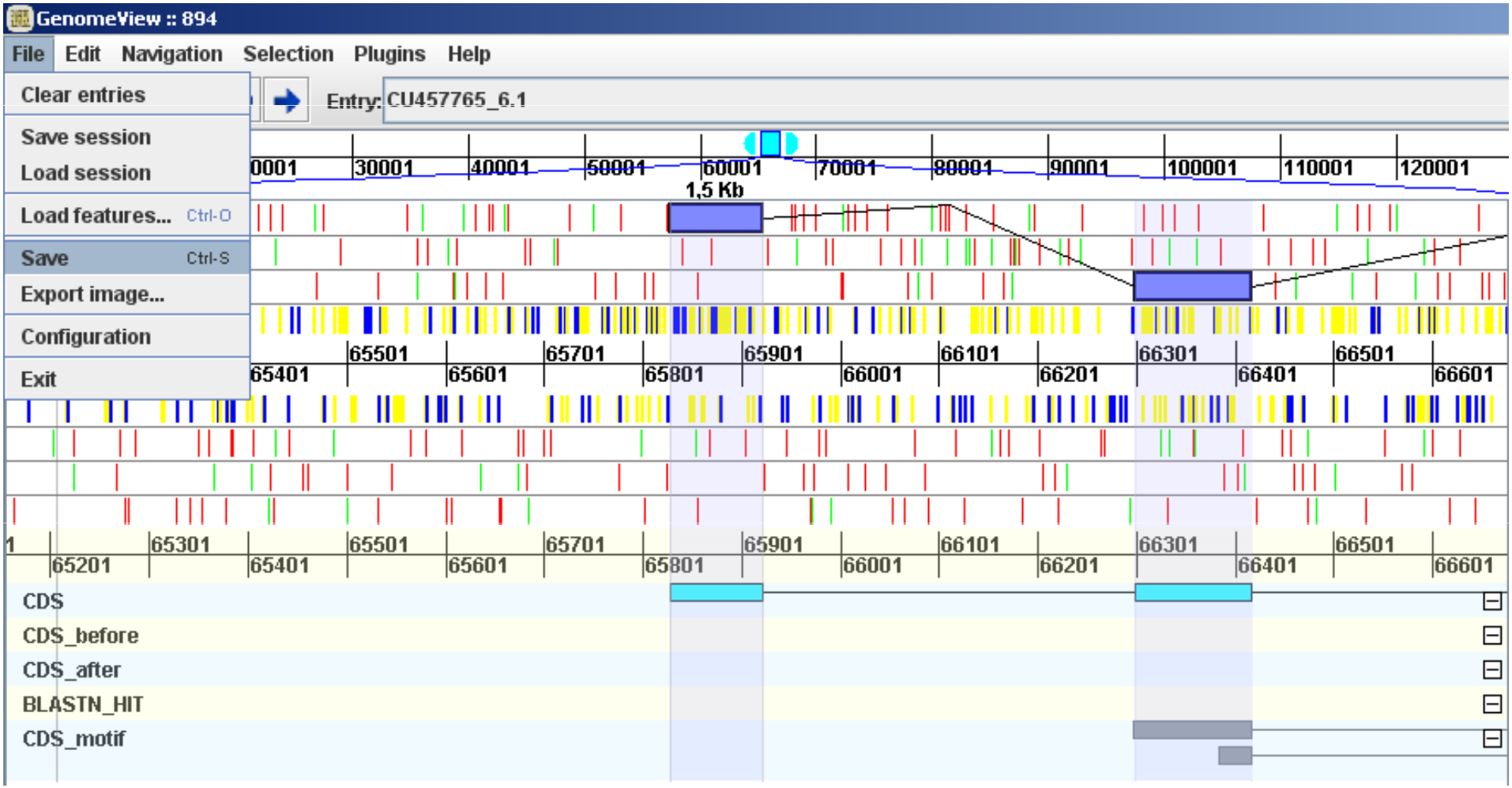


Select strain with alternative allele

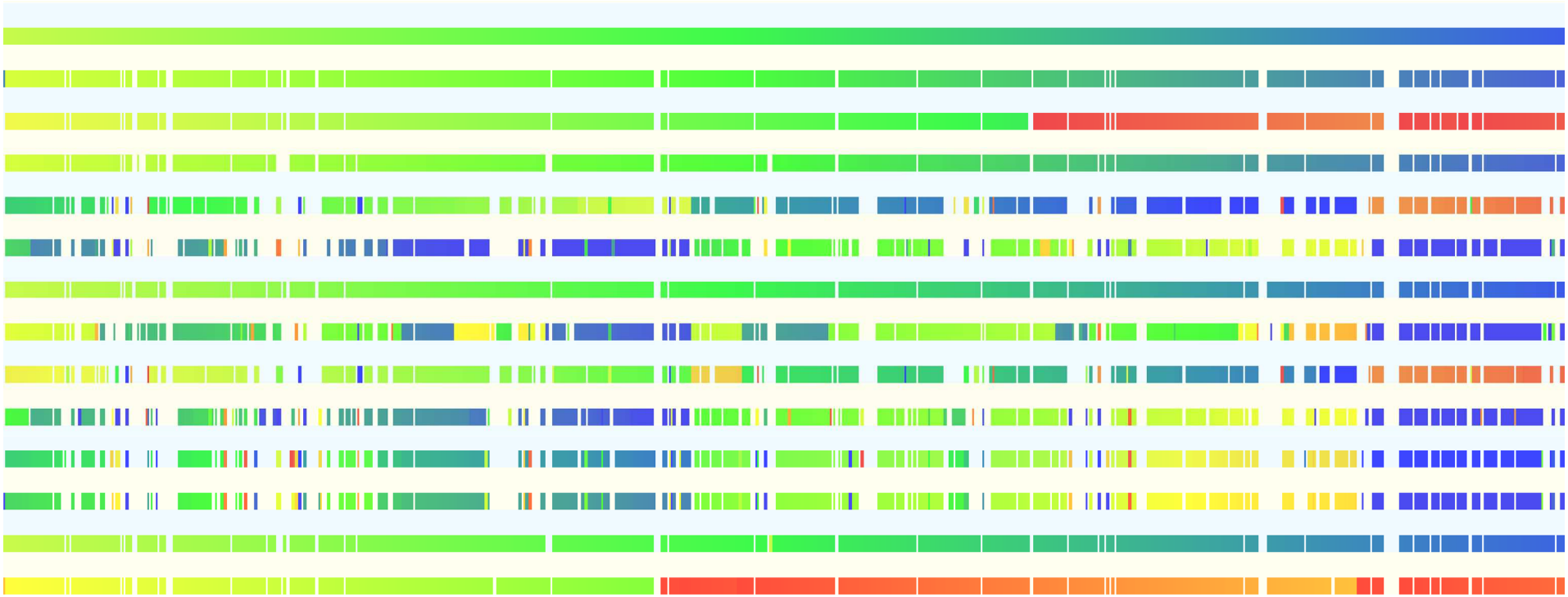
Editor integration



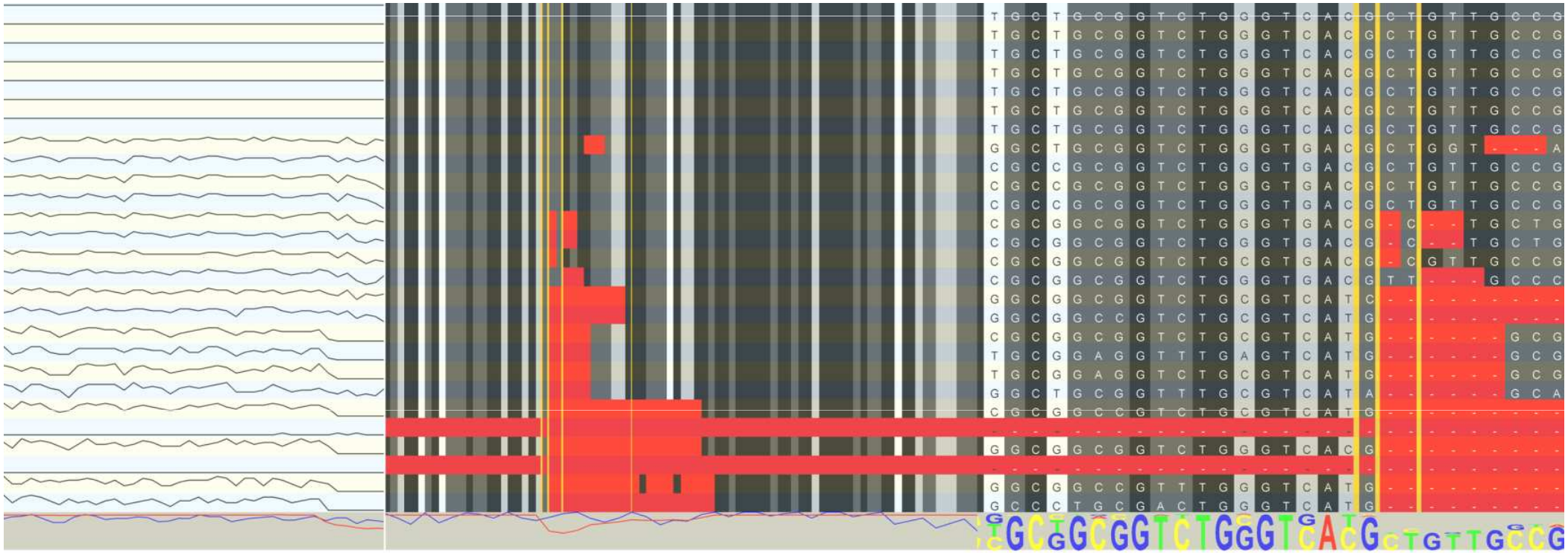
Saving back to the server…



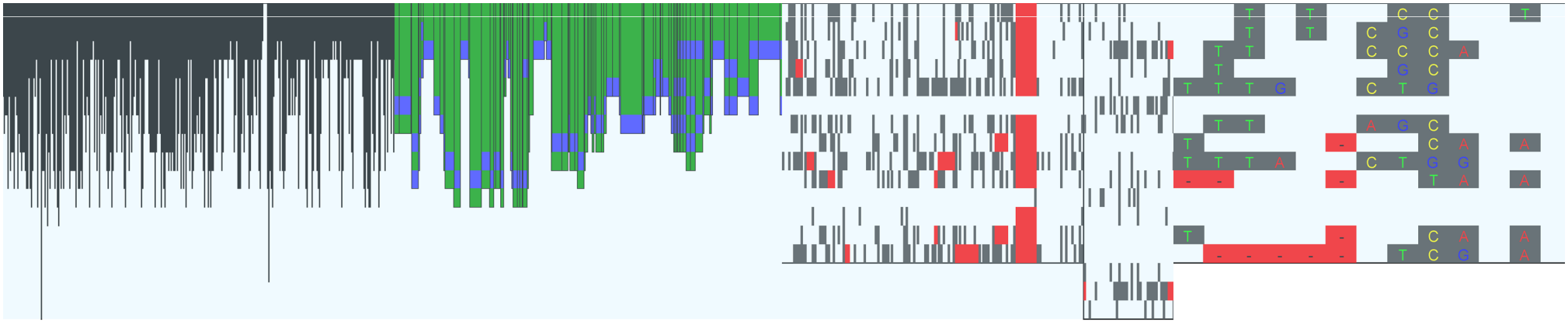
Synteny



Multiple alignments



Next-gen multiple alignments



Things to come

- Indexed sequence data structure
- Indexed feature data structure
- → Faster and more interactive

Summary

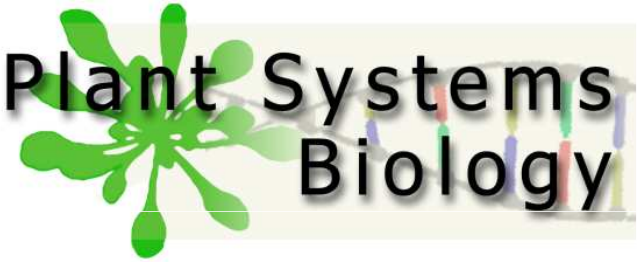
- NGS data sets are large and daunting
- Visualization makes them less daunting
- Visualization is essential
- GenomeView can visualize a broad set of NGS data sets (and other types of data)

- GenomeView is freely available @

http://genomeview.org

Acknowledgements

- Yves Van de Peer
- Yvan Saeys
- James Galagan
- Thomas Van Parys



http://genomeview.org