

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

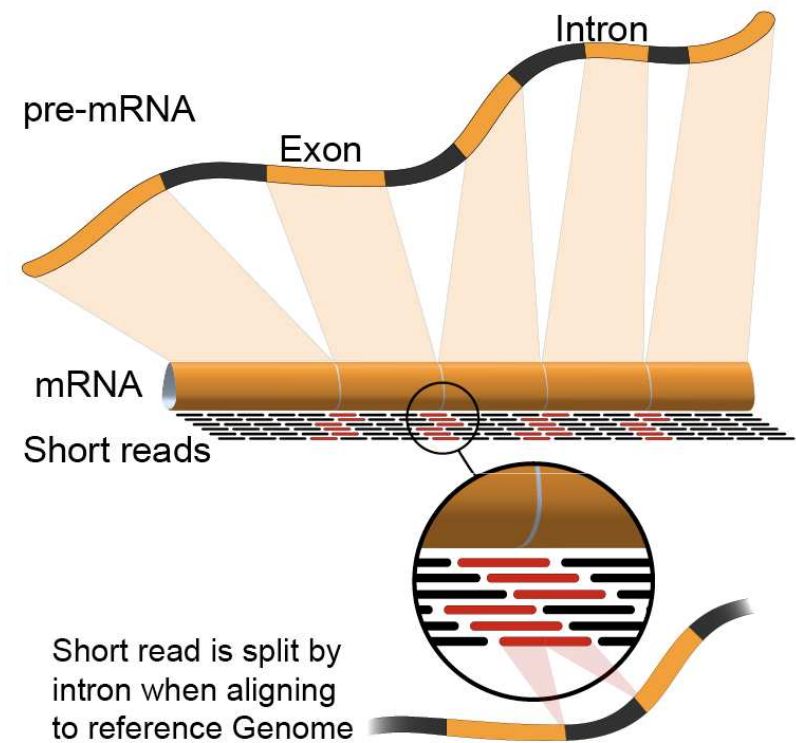
- Assembly
 - At least 40x coverage
 - Result is large number of contigs
- **Map to a reference genome**
 - **Re-sequencing → SNP discovery**
 - **ChIP-seq**
 - **RNA-seq**

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

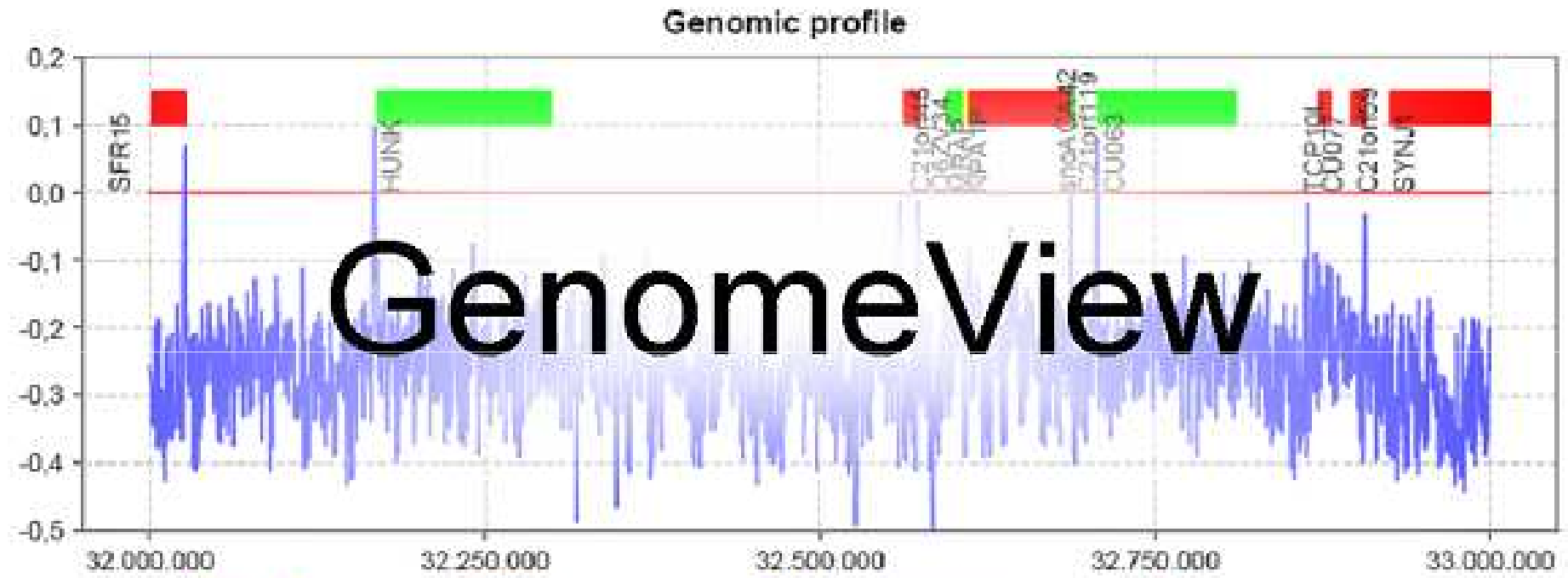
- Extracted DNA (or somebody else did it)
- Put it in machine, got a lot of data in return
- We have aligned it to a reference sequence
- We now have an even bigger file with even more information.

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

The screenshot shows the GenomeView application interface. The main window displays a genomic track with various data sources and visualization tracks. The interface includes a menu and toolbar at the top, a navigator on the left, and a track management panel on the right. The main display area shows multiple tracks, including gene structure, ruler, CDS, and gene tracks. The visualization tracks show read alignments and feature details. The information panels on the right provide details for selected features, including their location, strand, and score.

Labels in the image:

- Navigator
- Menu and Toolbar
- Loaded data
- Track management
- Information panels
- List of features
- Selected feature details
- Gene frame structure
- Visualization tracks

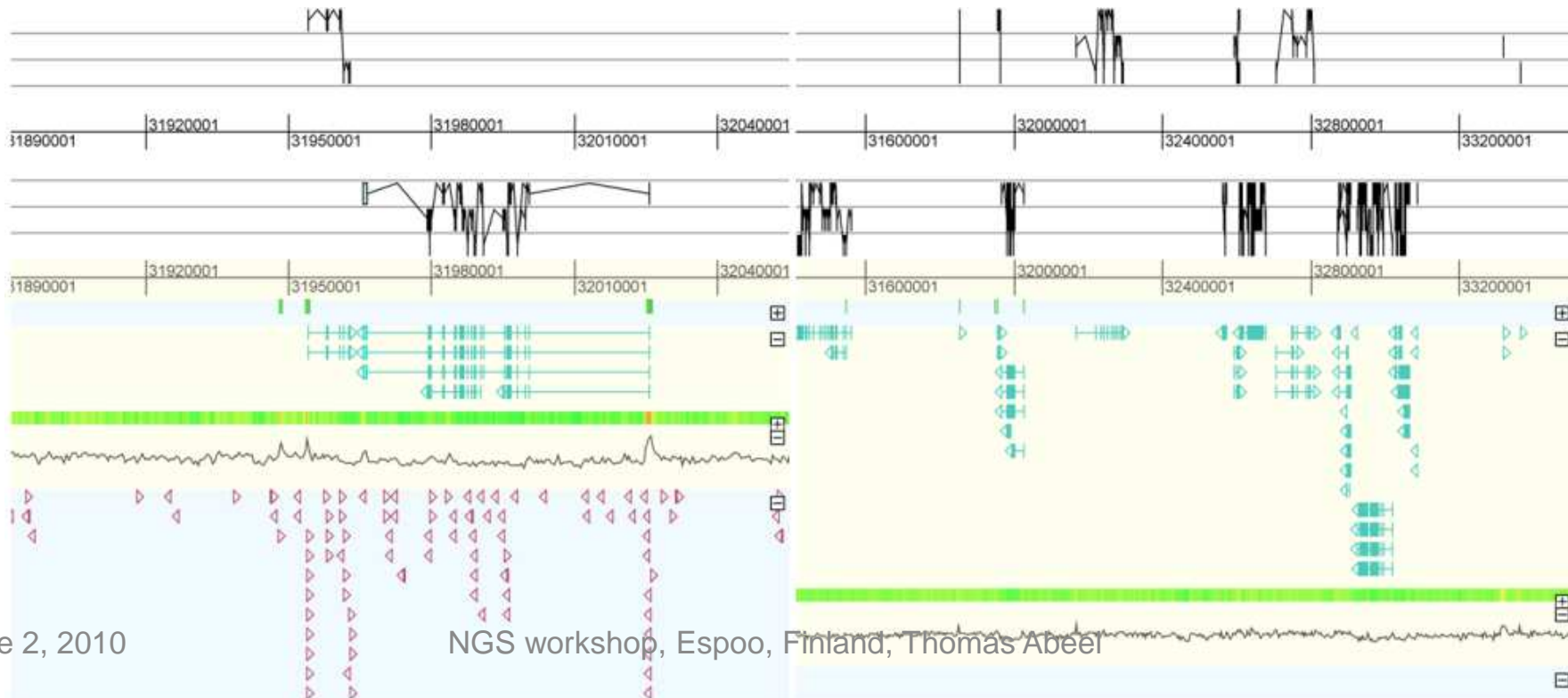
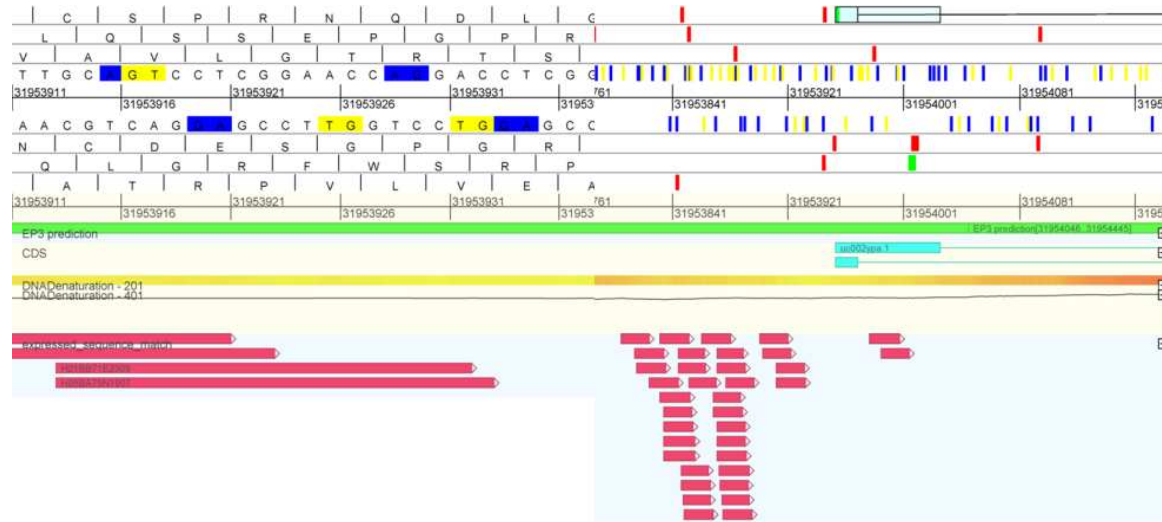
St.	A.	C.	up d.	Track name
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Gene structure
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Ruler
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	CDS
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	gene
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Short reads: http://www.broadinstitute.org/soft...
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Short reads: http://www.broadinstitute.org/soft...
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Short reads: http://www.broadinstitute.org/soft...

Name	Start	Stop	Inte	Spli
CDS[248624..248899..248958..249050...			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
CDS[254709..254802..254976..255079...			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
CDS[257414..257661..258021..258358...			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
CDS[286352..286420..287509..288003...			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
CDS[305901..306089..306291..306609...			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
CDS[311402..311577..311635..311785...			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
CDS[318826..319129..319739..320418...			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
CDS[343544..343652..343897..344036...			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
CDS[343544..343652..343897..344036...			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
CDS[357243..357330..357379..357485...			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
CDS[368778..369249..370427..370764...			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

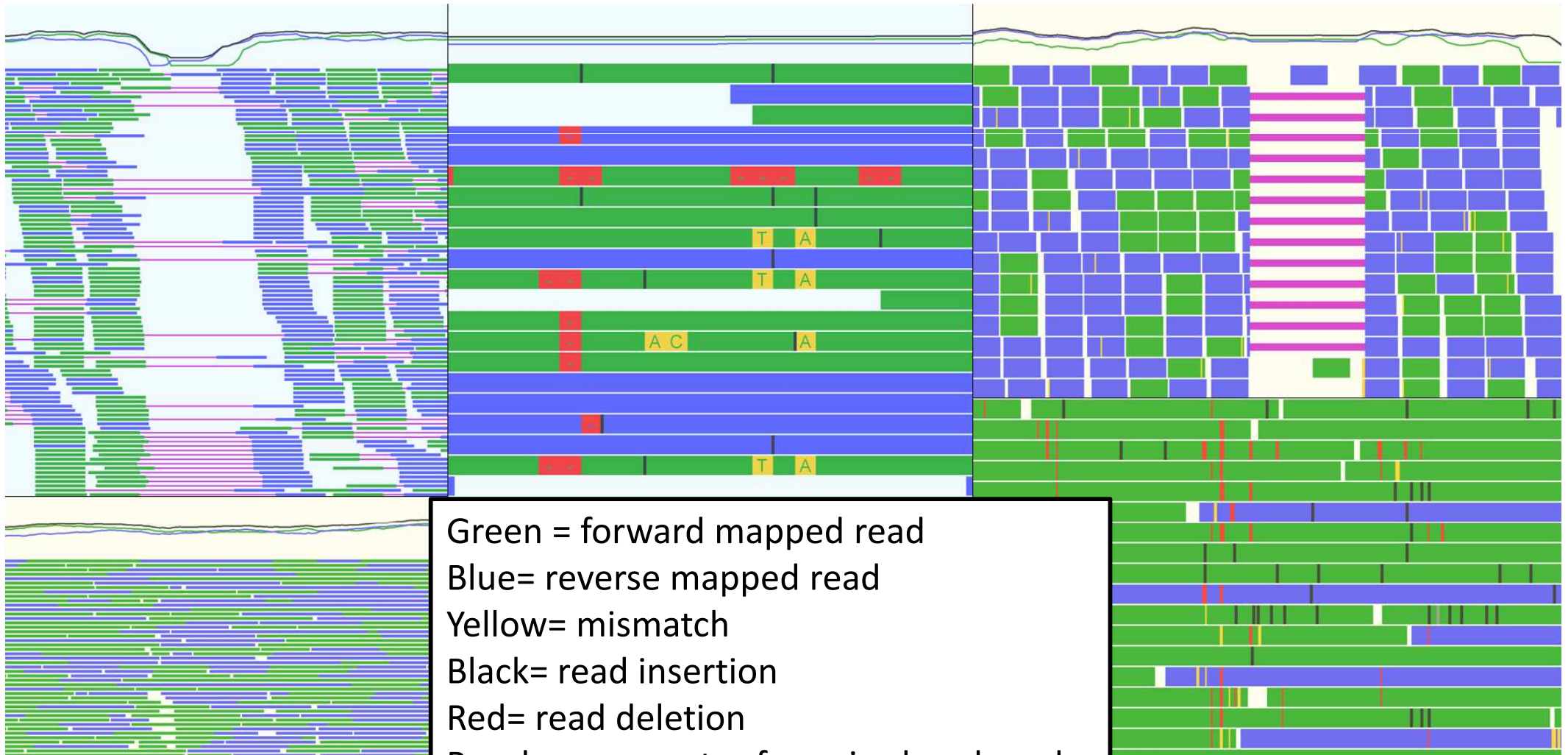
Details on selected items:

Data origin: ..\cache\17ADD90FE865FA7BFCC2930992665FFC...
Location: [311402..311577..311635..311785..311837..312879..31...]
Strand: REVERSE
Score: 0.0
Parent="C53D5.6"
source=curated

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

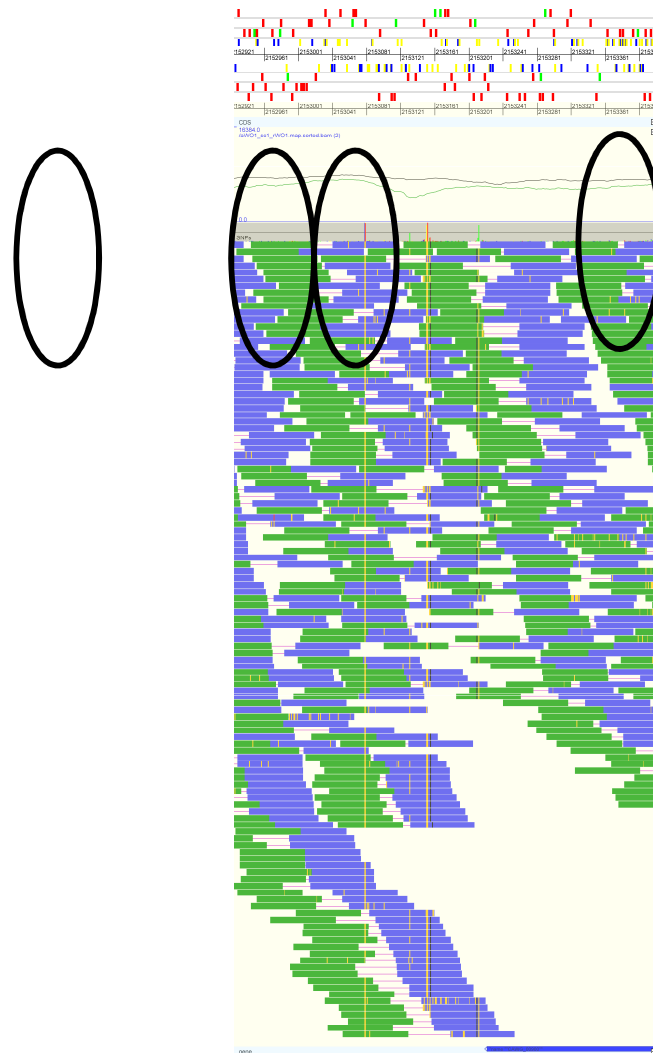
Quality, indels and mismatches

The screenshot shows the GenomeView software interface. The main window displays a genomic track for chromosome 1, with coordinates ranging from 801 to 8001. The track includes a reference sequence (G E V S A L V E G G V E G G H H C), a CDS (Gene structure), and SNPs. A ruler track is also visible. The short reads track shows multiple reads with various colors (green, blue, red) indicating mismatches or indels. A tooltip is displayed over a read, showing the following statistics:

Forward coverage	: 51
Reverse coverage	: 28
Total coverage	: 79
Insertion	: T

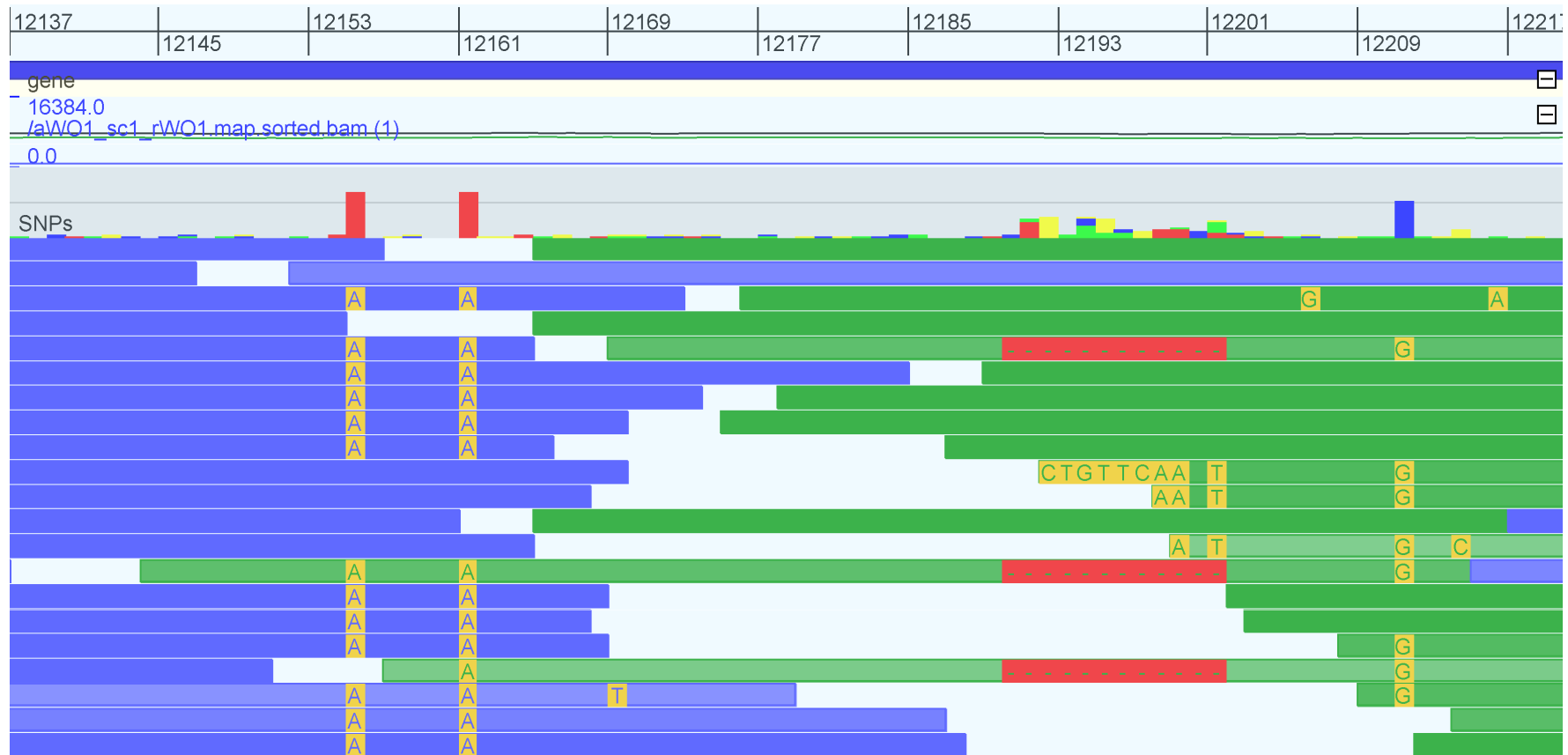
The right-hand side of the interface contains a 'Data sources' panel with two entries: /NL43_WT_Ref.fasta and /NL43_WTreads.bam. Below it is a 'Track list' panel with columns for 'St.', 'A.', 'C.', 'up', and 'd.', and a 'Track name' column. The 'Features' panel is currently set to 'CDS' and shows a table with columns for 'Name', 'Start', 'Stop', 'Inte.', and 'Split.'. The 'Details on selected items:' panel is currently empty.

SNP track



Re-sequencing demo

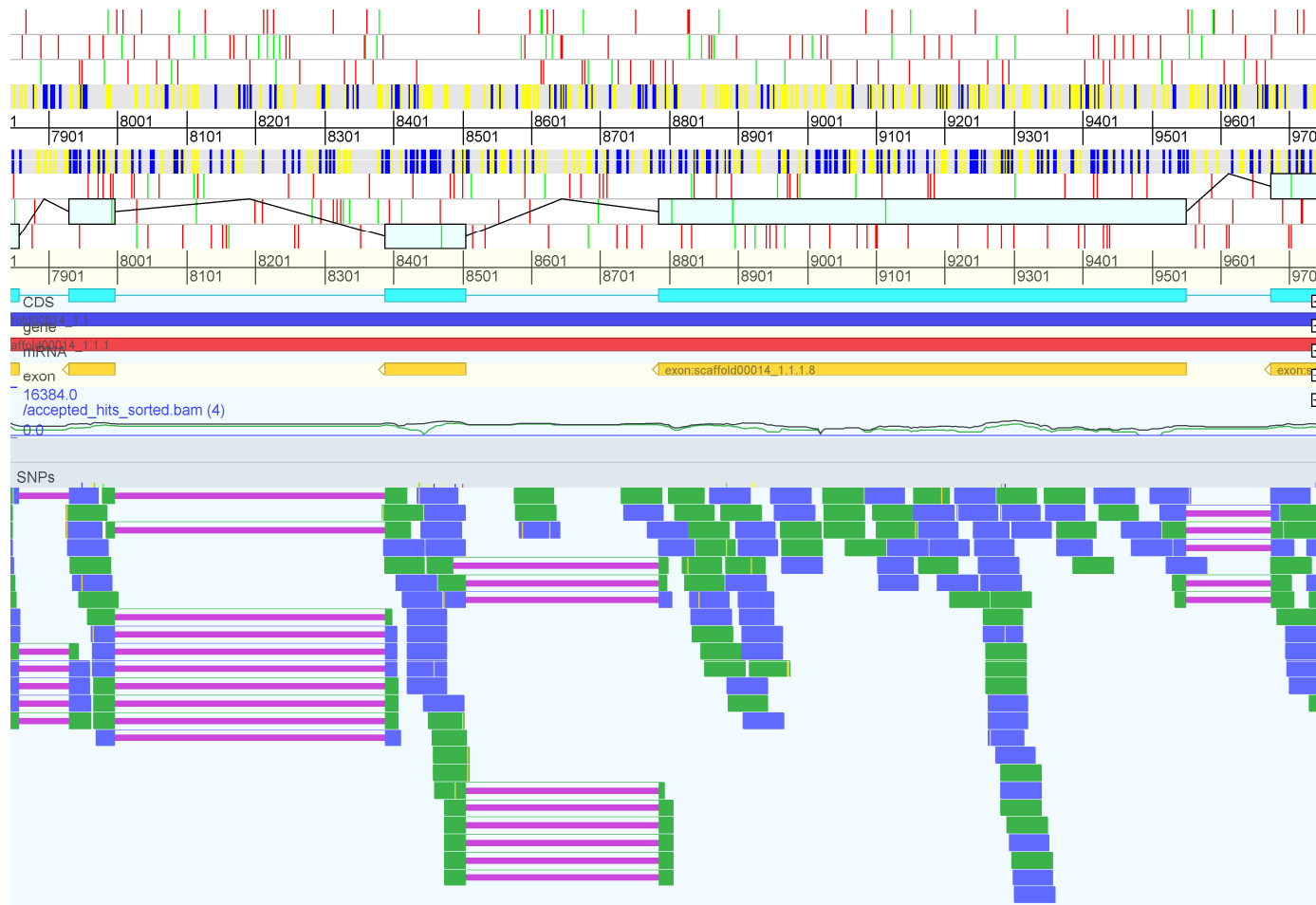
SNPs



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

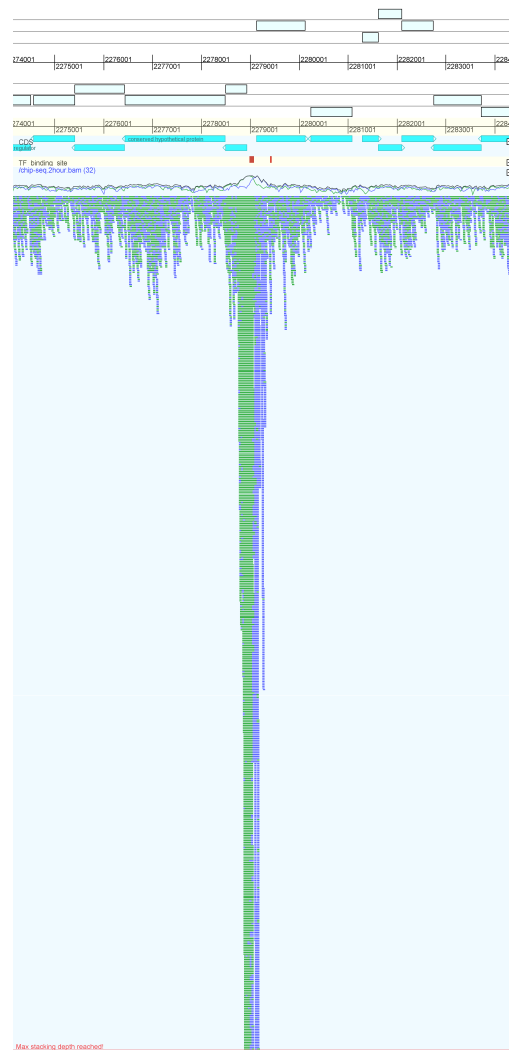
RNA-seq demo

RNA-seq



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

Chip-seq



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

The screenshot displays a genome browser editor interface. On the left, an 'Edit structure' window is open, showing the following details:

- 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

The main window shows a genomic track with coordinates from 000001 to 440001. The track includes a ruler, gene structure, CDS, gene, mRNA, exon, five_prime_UTR, three_prime_UTR, and short reads. A legend on the right lists these tracks with their status (checked/unchecked) and update options.

Below the tracks, a table lists features for the selected CDS:

Name	Start	Stop	Inte	Spli
CDS:scaffold00014_1.1.1.11	✓	✓	✓	✓
CDS:scaffold00014_2.1.1.1	✓	✓	✓	✓
CDS:scaffold00014_3.1.1.1	✓	✓	✓	✗
CDS:scaffold00014_4.1.1.9	✓	✓	✓	✓
CDS:scaffold00014_5.1.1.3	✓	✓	✓	✓
CDS:scaffold00014_6.1.1.1	✓	✓	✓	✓
CDS:scaffold00014_7.1.1.3	✓	✓	✓	✓
CDS:scaffold00014_8.1.1.1	✓	✓	✓	✓
CDS:scaffold00014_9.1.1.1	✓	✓	✓	✓
CDS:scaffold00014_10.1.1.1	✓	✓	✓	✓
CDS:scaffold00014_11.1.1.1	✓	✓	✓	✓
CDS:scaffold00014_12.1.1.1	✓	✓	✓	✓

At the bottom right, a 'Details on selected items' window shows the following information:

- Data origin:** .../itag_subset/newbler_v1_genome_subset40kb.../ff3.gz
- Location:** [7758..7857, 7929..7996, 8388..8504, 8786..9550, 9673..10163, 10459..10582, 10760..10890, 10956..11029, 11120..11171, 11245..11289, 11359..11383]
- Strand:** REVERSE
- Score:** 0.0
- Parent:** mRNA:scaffold00014_1.1.1
- source:** EuGene
- ID:** CDS:scaffold00014_1.1.1.11

The bottom of the interface shows a track for SNPs with a selection of 491 nt / 163 aa.

Browser integration

<< Previous Gene on Contig (Rv0014c) Next Gene on Contig (Rv0016c) >>

Gene Info Collapse

Locus	Rv0015c
Gene Symbol	pknA
Synonyms	Rv0015
Gene Name	transmembrane serine/threonine-protein kinase A pknA
Gene Product Names	Rv0015c transmembrane serine/threonine-protein kinase A pknA
Gene Family	transmembrane serine/threonine-protein kinase A pknA
Location	M. tuberculosis H37Rv: Chromosome 1: 17467-18762 -
Length	Gene: 1296 nt Protein: 431 aa

Show Legend Show UTRs Show Domains Show Polymorphisms

Chromosome **Gene** **Transcript** **Protein Domains** **Start** **Stop**

4.4 mb 22,003 b 18,762 b

2.9 mb

1.5 mb

0.1 mb 14,227 b 17,467 b

Mycobacterium tuberculosis H37Rv reference: G (T - threonine)

A (threonine): 4783_04, K37

G (threonine): K49, K100, 11821_03, Mtb_F11, T85, T83, GM_0981, MT_H37Rv, 4141_04, K93, T92, 98_1833, Mcanettii_K116, SG1, GM_1503, T67, 95_0545, T17, K67, 5444_04, 91_0079, M4100A, K21, 00_1695

[more details](#)

[more details](#)

[more details](#)

[Find other genes with this domain](#)

Select polymorphism

SNP in short read alignment

View polymorphism in GV

Go to polymorphism detail page

TB Database AN INTEGRATED PLATFORM FOR TUBERCULOSIS RESEARCH

QUICK SEARCH Search Search for

TBDB Home Search My TBDB TBExpression TBGenomes Tools and Analysis TB Community Help

Genomes Genes Diversity Sequencing Comparative Analysis Search BLAST Download

Quick Links


- Feature Search: Search for features by type. Filter on an additional text query if desired.
- Gene Expression Publications: Search and browse publications with gene expression data that are stored in TBDB.

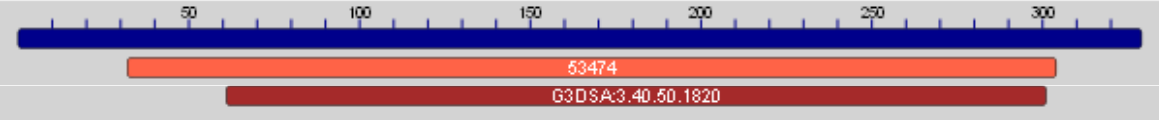
Polymorphism Feature Detail

Gene	Rv0006 - D
Location	8188 Gene S
NT Coordinate	887
AA Coordinate	296
Drug	FLQ
Resistance	
Reference	
Allele	T
Residue	L - leucine
Strains(25)	00_1695 11821_07 04 5444_04 91_0079 95_0545 98_1833 GM_0981 GM_1503 K100 K37 K49 K67 K9
Alternate	
Allele	C
Residue	P - prolin
Mutation	L296P
Mutation Type	SNP
Strains(1)	K21


Select strain with alternative allele

Editor integration


Protein Domains 



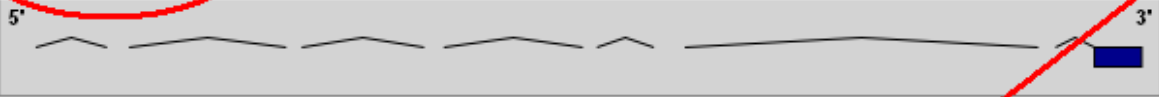
External ID	53474
From Database	superfamily
Description	alpha/beta-Hydrolases
External ID	G3DSA:3.40.50.1820
From Database	Gene3D
Description	no description

Protein Homologs 

n/a



Gene Structure 

[View in GenomeView](#) [View in Artemini](#)



Structure

```
..67754 67817 68024 68317 69173 69255 69303 70108 70713 70818 7111  
CCACGGTTTCAAATCATCGAAGGATCGGATCCCATGGTGAATCTTGCT
```

xt  Previous  Highlight all Match case

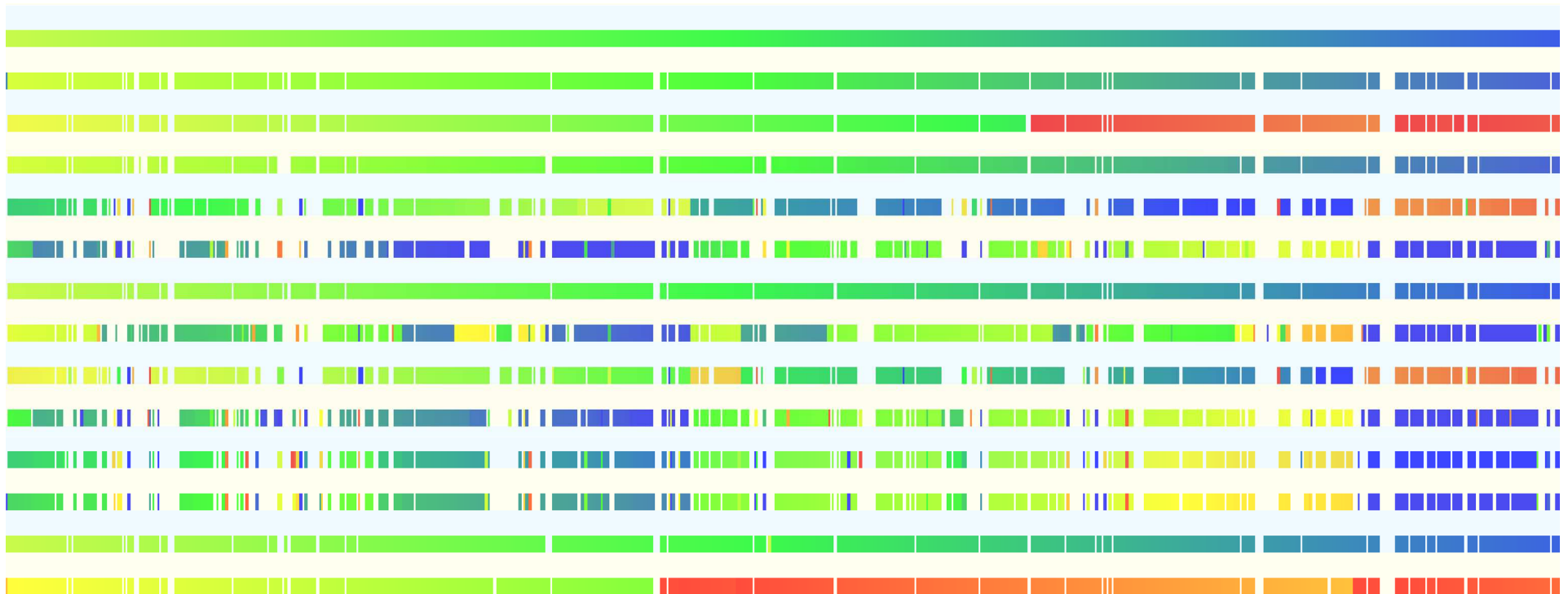
enomeview/launch.jnp?--url http://bioinformatics.psb.ugent.be/cgi-bin/bogas_art/gv_ws.pl?user_id=""&locus_id=CU457765_6.1&genome=Solly&release=666666&context=gene

Saving back to the server...

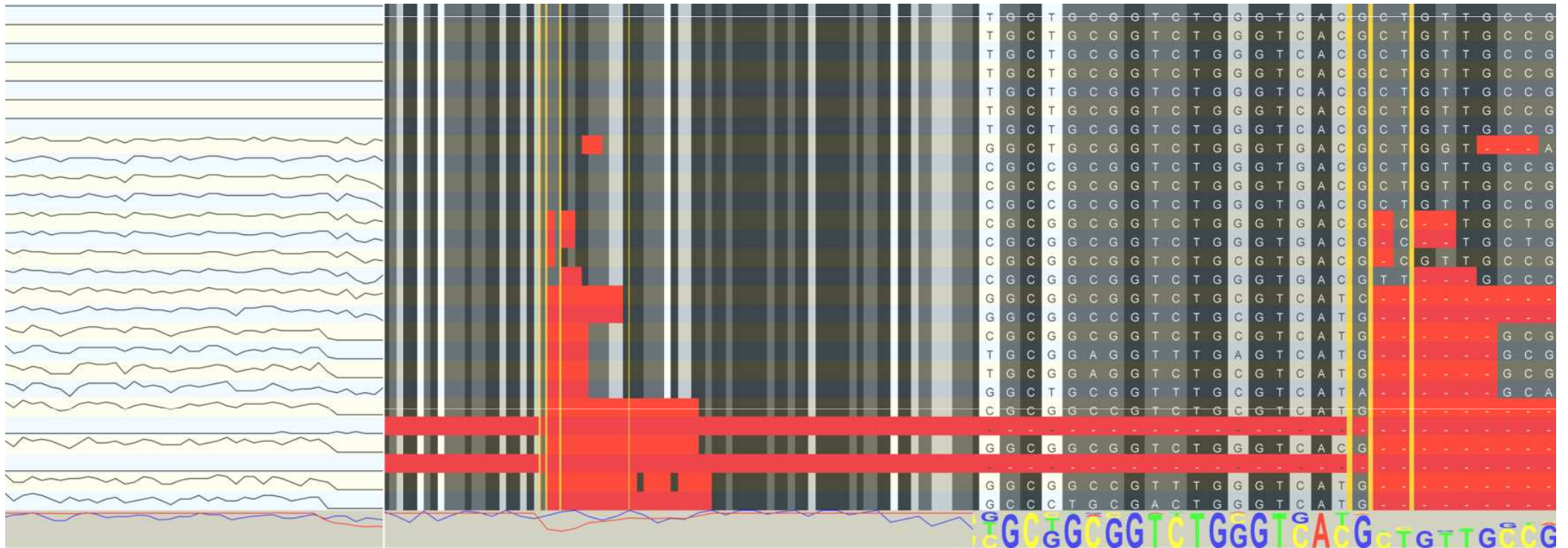
The screenshot displays the GenomeView application window titled "GenomeView :: 894". The menu bar includes "File", "Edit", "Navigation", "Selection", "Plugins", and "Help". A dropdown menu is open under "File", showing options: "Clear entries", "Save session", "Load session", "Load features... Ctrl-O", "Save Ctrl-S", "Export image...", "Configuration", and "Exit". The "Entry" field contains "CU457765_6.1".

The main view shows a genomic track with a scale from 0001 to 120001. A 1.5 Kb region is highlighted in blue, spanning from approximately 65801 to 66301. Below the main track, there are several tracks: "CDS", "CDS_before", "CDS_after", "BLASTN_HIT", and "CDS_motif". The "CDS" track shows a cyan bar within the highlighted region. The "BLASTN_HIT" track shows a yellow bar within the highlighted region. The "CDS_motif" track shows a grey bar within the highlighted region.

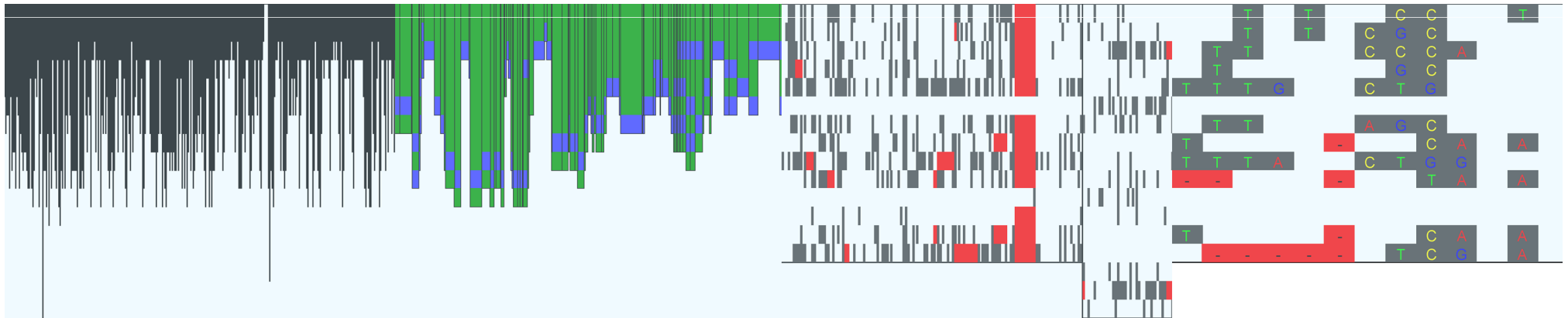
Synteny



Multiple alignments



Next-gen multiple alignments



Things to come

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

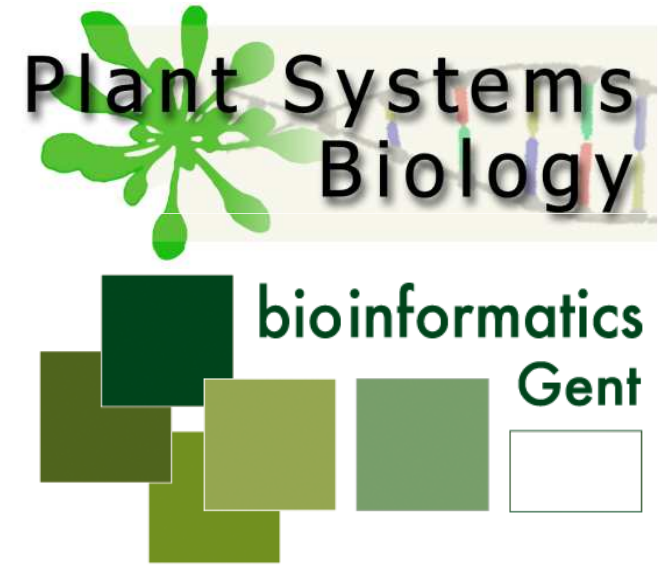
Summary

- 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>