Next-Gen Sequencing and data processing

Li-Jun Ma
Umass Amherst

Fusarium Workshop Asilomar
March 15 2011

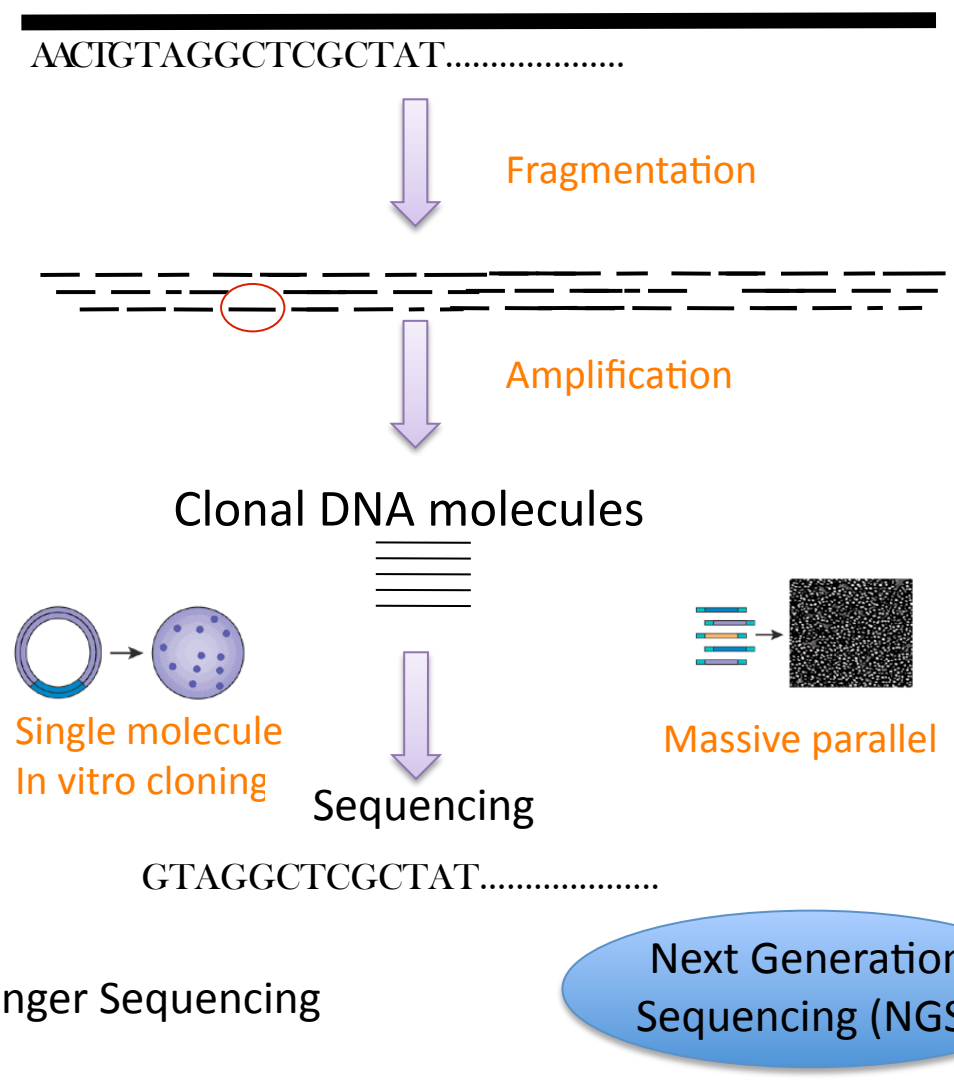
A little ride to the west

University of Massachusetts Amherst MA ⟵ Broad Institute at Cambridge MA

A big change

What can a biologist do facing the explosion of data

Sequencing – the principles



Sanger Sequencing

Next Generation Sequencing (NGS)

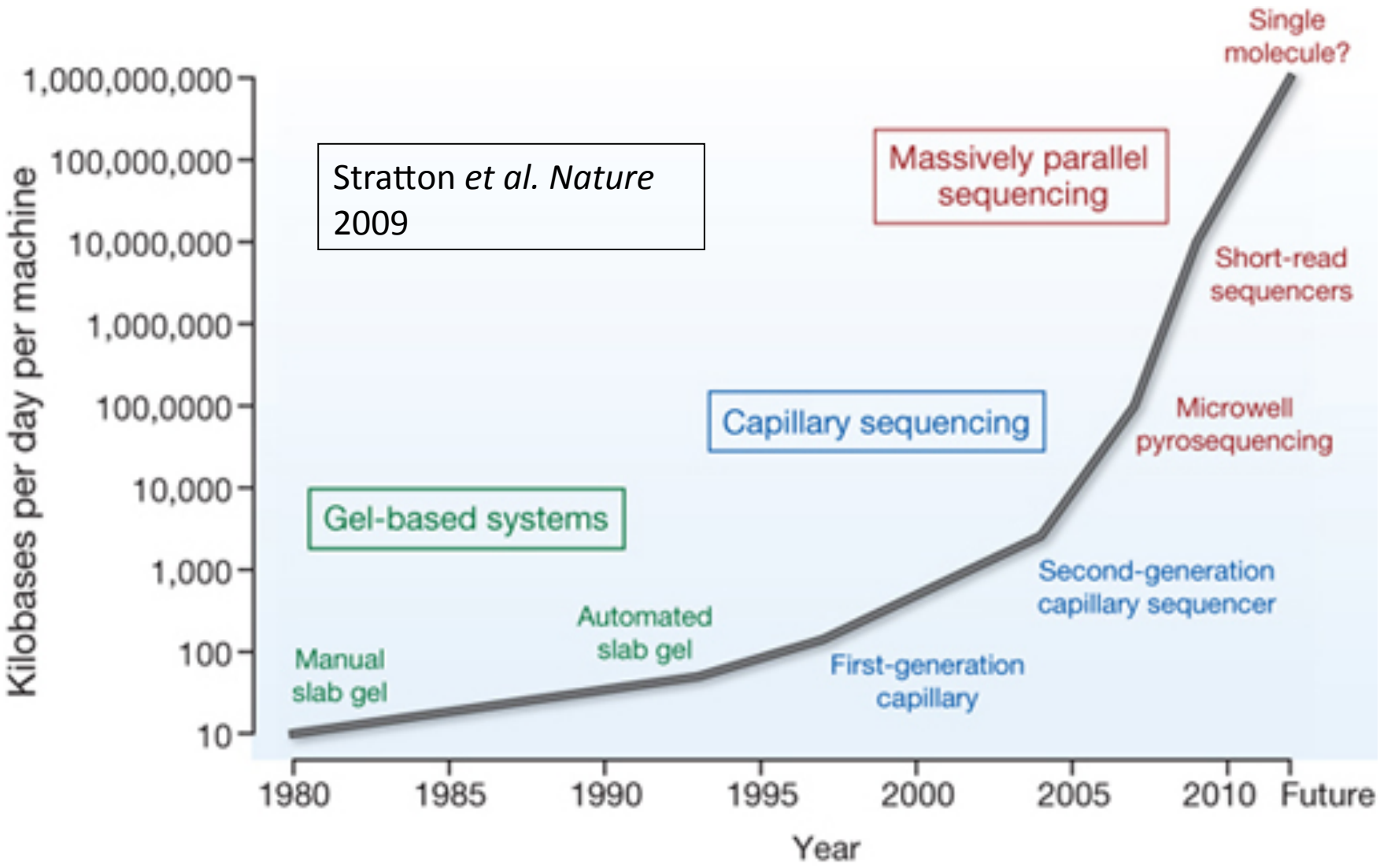
sequencing

Single Molecule Sequencing (SMS)
Third generation

Massive

Cheap

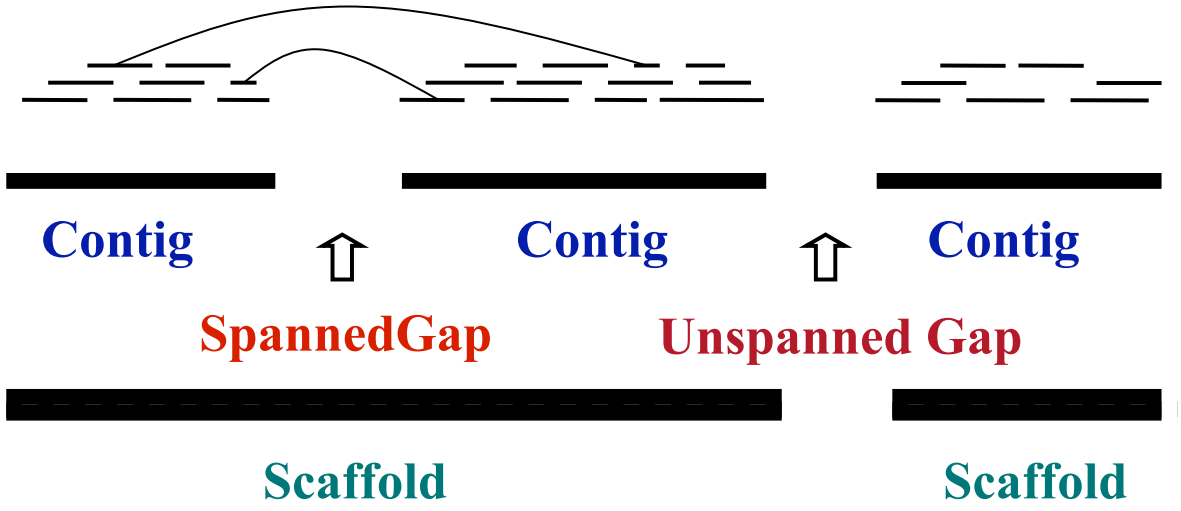
Genome sequencing is getting faster and cheaper



Usage of the NGS data

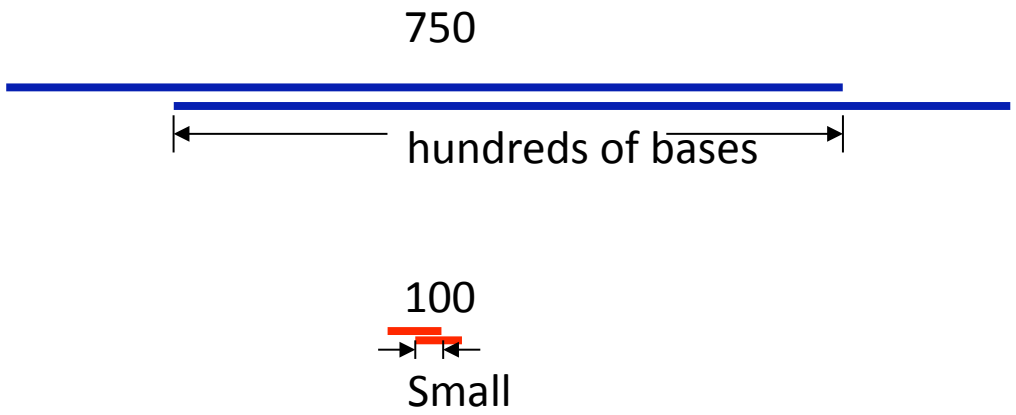
- Genome sequencing – whole assembly
- Population study – SNP and other genetic variations
- Transcriptional profiling – normalized sequence coverage
- Protein-DNA interactions – pick calling

Whole Genome Shotgun Assembly



Assemble NGS

With small reads, false overlaps dominate



Lack large insert libraries

Assembly tools

- ALLPATHs (Broad Institute)
- Velvet (European Bioinformatics Institute)
- Newbler (454 Life Sciences)
- ABySS (Canada Genome Sciences Center)
- Soap *de novo* (Beijing Genome Insititute)

Usage of the NGS data

- Genome sequencing – whole assembly

- Population study – SNP and other genetic variations
  - Mapping
  - SNP calling
  - indel

- Transcriptional profiling – normalized sequence coverage
  - Mapping, or assembly then mapping
  - Coverage
  - normalization

- Protein-DNA interaction – pick calling
  - Mapping
  - Coverage
  - pick calling

Mapping – find homolog

Tool	Algorithm
MAQ	Seed Matching
BWA	Burrows-Wheeler-Transformation
Bowtie	Burrows-Wheeler-Transformation
MOSAIK	Smith-Waterman
Novalign	Iterative Search
SOAP	Burrows-Wheeler-Transformation

Real data

	File name	size
Initial file	62MA3AAXX.1.fq	2,348,282,969
 Mapping		
Detailed mapping info	62MA3AAXX.1.sam	3,304,581,692
 SAM tools		
Binary of the .sam file	62MA3AAXX.1..bam	626,868,823
Summarized mapping results	62MA3AAXX.1.pileup	676,262,764

Files and their formats

Fastq:

```
@62MA3AAXX100922:1:100:10000:11938
ACAGTCGGGGGCATCAGTATTCAATTGTCAGAGGTG
+
EEGGGEDEGGGEGGGGEGGEGFEEGGGGGD?B=B;H
```

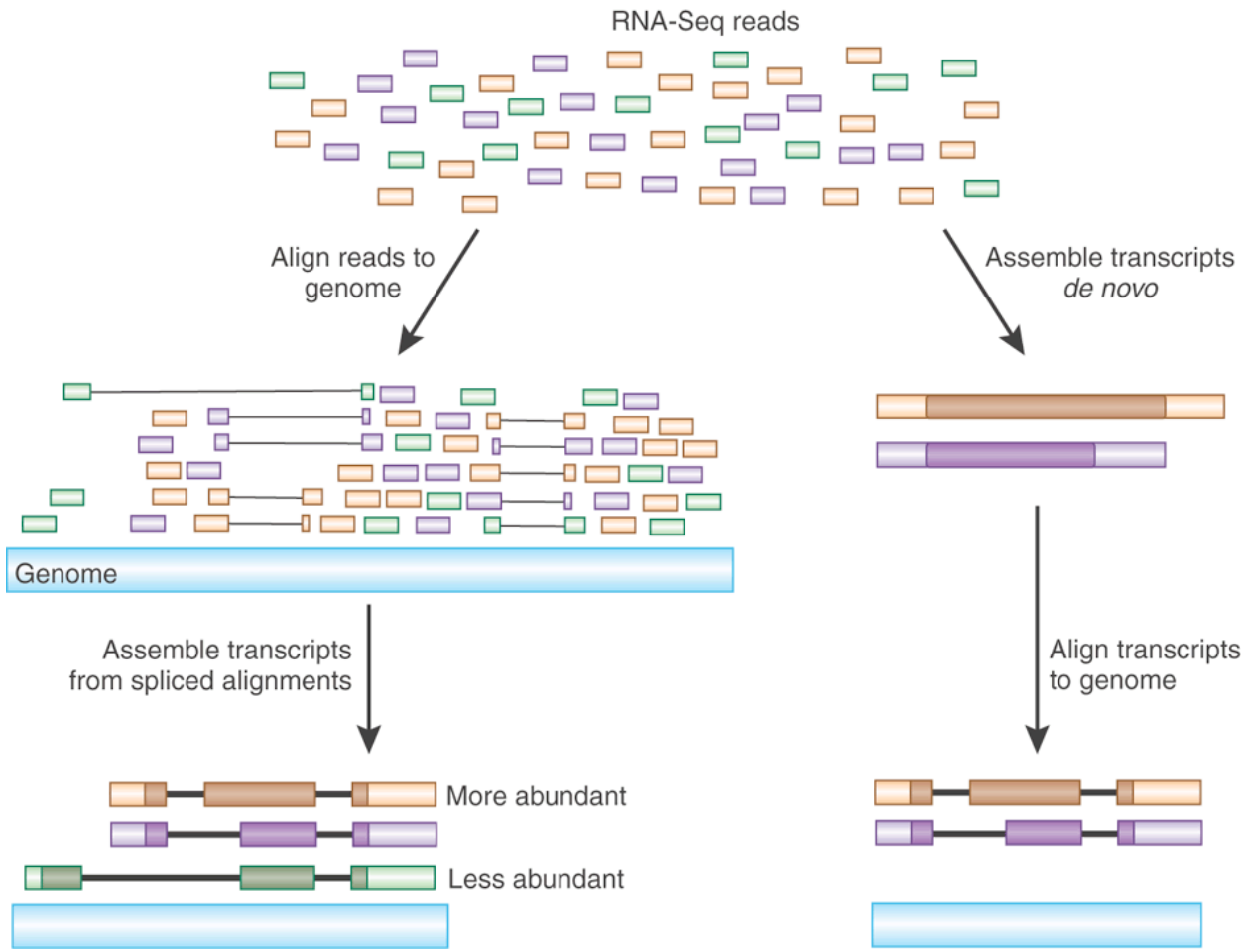
Pileup:

```
Seq1    2554    C    8    ,,,t^!,^!t^!,^!,      9IIHDFIH
Seq1    2555    A    8    ,,,,,,,,      1HGE>EED
Seq1    2556    C    8    ,,,,,,,,      =IIHFIIH
Seq1    2557    C    9    ,,,,,A,,^!,    ;IGHGIBHI
Seq1    2558    C    9    ,,,,,,,,,     ;IIHGIIHG
Seq1    2559    C    9    ,,,,,,,,,     =HIHGIHHI
Seq1    2560    G    9    ,,,,,,,,,     =IIHHHIHI
Seq1    2561    T    9    ,,,,,,,,,     EIIHIIIHI
Seq1    2562    A    10    ,,,,,,,,,^!,   AIIGGHHHIB
Seq1    2563    C    10    ,,,,,,,,,,    EIIFIIIHIG
Seq1    2564    A    10    ,,,$,,,,,,,   EIIHHIIHDG
```

Filters for SNP discover

- Reference re-sequence
  - Assembly errors
- Coverage
  - Confidence level
- Quality (haploid genome)
  - if the second best is the same as the reference (repeats)
  - if the best base quality is less than phrap 20
  - if the minimal base quality in the neighbor 6 bases less than 15
  - if the consensus is ambiguous

RNA-seq and transcriptomics



Haas B.J. and M.C. Zody 2010 Nature Biotechnology

RNA-seq methods

- TopHat http://tophat.cbcb.umd.edu/
  - a fast splice junction mapper.
  - using short read aligner Bowtie,
  - identifying splice junctions between exons.

- Cufflinks http://cufflinks.cbcb.umd.edu/
  - assembles transcripts
  - estimates their abundances,
  - tests for differential expression and regulation in RNA-Seq samples.

- Inchworm http://inchworm.sourceforge.net/
  - employs the Kmer graph method
  - reconstruct transcripts from Illumina RNA-Seq (prefer strand-specific)
  - Both de novo assemble the transcriptome and align to the reference genome

RNA-seq – expression

- Coverage (abundance of alignments) as the measure of expression level

- FPKM (RPKM) normalization
  - Reads Per Kilobase of exon per Million mapped sequence reads

- Different read lengths
  - Average coverage over a feature on 1 Gb mapped bases

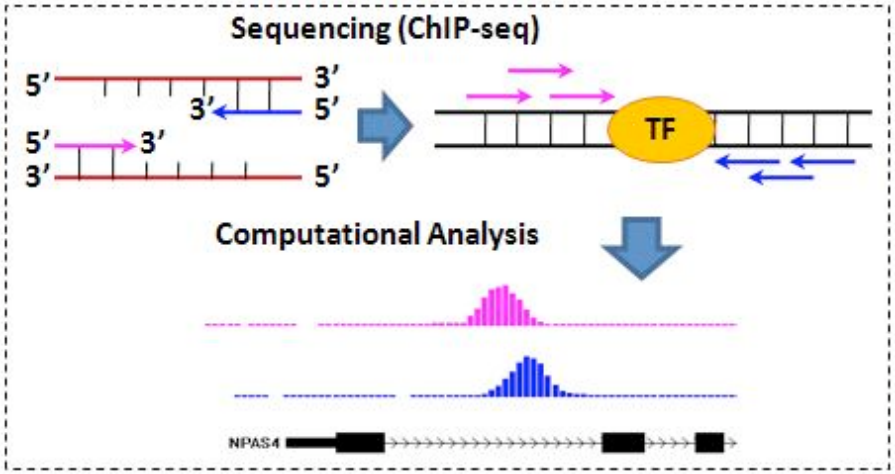
Coverage

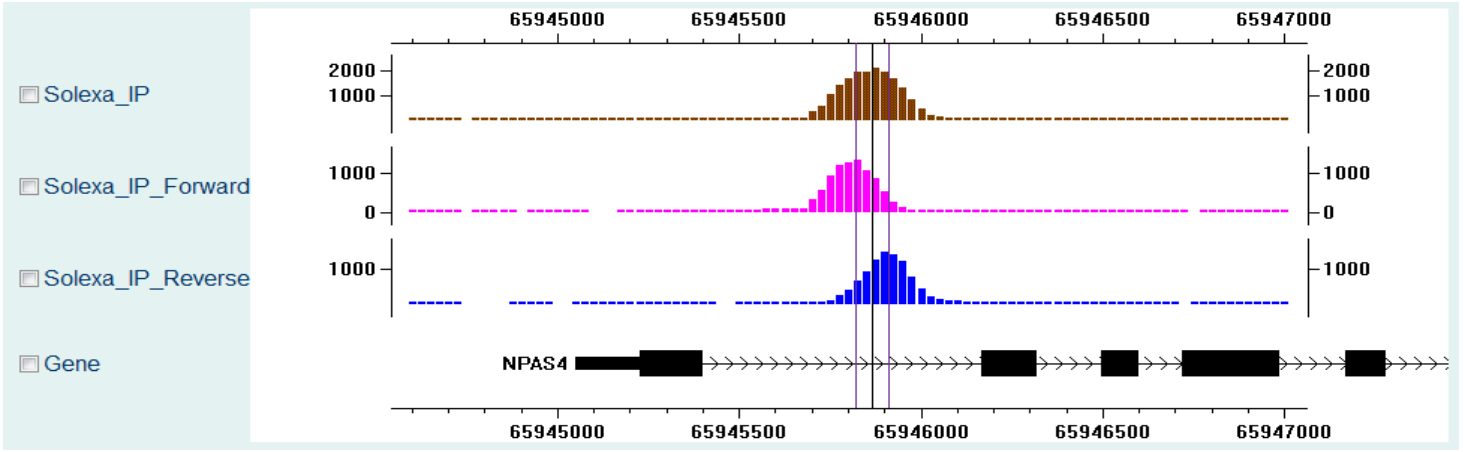
Pileup:

```
Seq1    2554    C    8    ,,,t^!,^!t^!,^!,      9IIHDFIH
Seq1    2555    A    8    ,,,,,,,,      1HGE>EED
Seq1    2556    C    8    ,,,,,,,,      =IIHFIIH
Seq1    2557    C    9    ,,,,,,A,^!,    ;IGHGIBHI
Seq1    2557    C    9    ,,,,,,^!,    ;IGHGIBHI
Seq1    2558    C    9    ,,,,,,,,,      ;IIHGIIHG
Seq1    2559    C    9    ,,,,,,,,,      =HIHGIHHI
Seq1    2560    G    9    ,,,,,,,,,      =IIHHHIHI
Seq1    2561    T    9    ,,,,,,,,,      EIIHIIIHI
Seq1    2562    A    10    ,,,,,,,,,^!,    AIIGGHHHIB
Seq1    2563    C    10    ,,,,,,,,,,     EIIFIIIHIG
Seq1    2564    A    10    ,,,$,,,,,,,    EIIHHIIHDG
```

• -- a match to the reference base on the forward strand
, -- a match on the reverse strand,
‘>’ or ’<’ for a reference skip,
‘\+[0-9]+[ACGTNacgtn]+’ -- an insertion
‘-[0-9]+[ACGTNacgtn]+’ -- a deletion from the reference
‘^’ -- the start of a read
‘$’ -- the end of a read segment

Peak calling





Hongkai Ji et al. Nature Biotechnology 26: 1293-1300. 2008

Peak detection

- FindPeaks (http://www.bcgsc.ca/platform/bioinfo/software/findpeaks)
- BS-Seq (http://epigenomics.mcdb.ucla.edu/BS-Seq/download.html)
- SISSRs (http://www.rajajothi.com/sissrs/)
- QuEST (http://mendel.stanford.edu/SidowLab/downloads/quest/)
- MACS (http://liulab.dfci.harvard.edu/MACS/)
- CisGenome (http://www.biostat.jhsph.edu/~hji/cisgenome/index.htm)

Challenges

- short reads, low quality, Massive
  - Data storage
  - Data processing
  - Data interpretation

Solution in my lab

- Local server
  - Dell R710 (64-bit, 6 cores)
- Storage
  - Dell MD1220 (7X2 Tb)
- Power backup
  - SMT2200

Setup cost and management cost

Local clusters

- Not always available
- Not compile for bioinformatics purposes
- Computing limits

Cloud computing

Ideal cloud world

Storage Computing

Pay as you use

Speed of data transfer

Local infrastructure (Linux environment)

Public data in the cloud

- Amazon hosts many public data sets that can be used in the cloud
  - ✓ (http://bit.ly/amazonpublicdata)

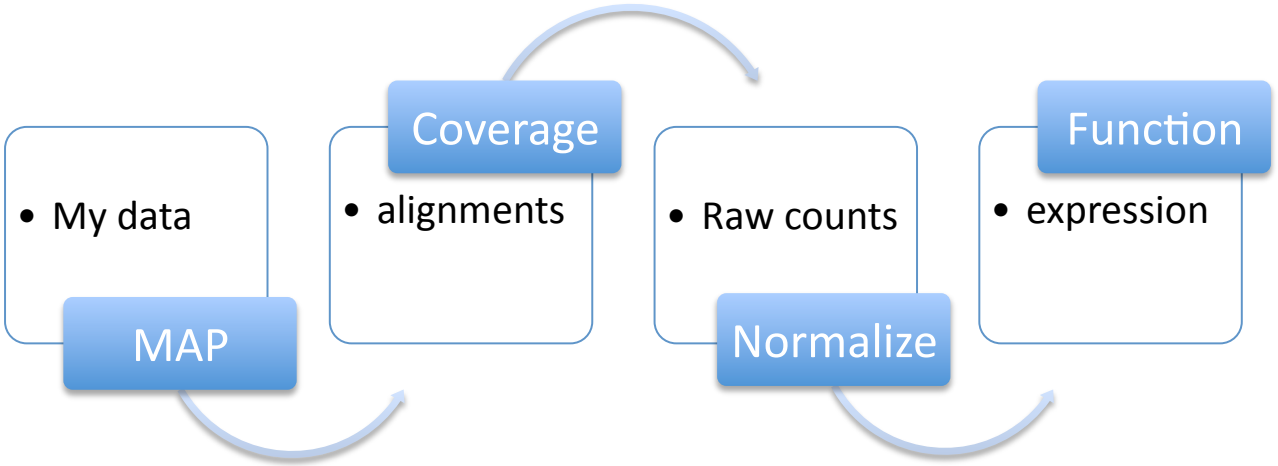
- Public data sets are freely hosted by Amazon
  - ✓ Money and speed
  - ? Give up access control rights to data

http://developer.amazonwebservices.com/connect/entryCreate!default.jspa?categoryID=244&entryTypeID=14

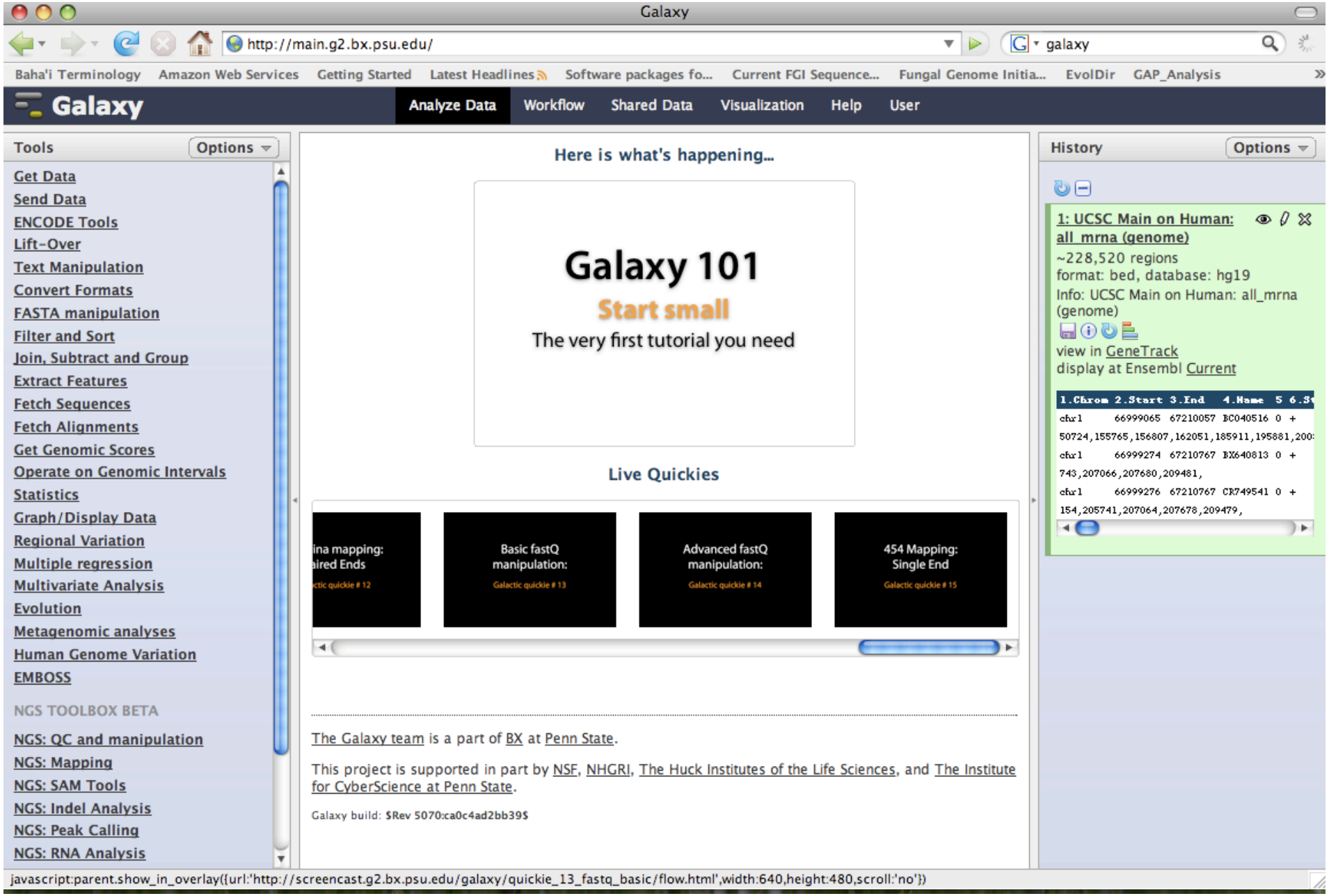
Local machine to access the cloud

- Purchase Hardware & ensure it's all compatible
- Appropriate resources for hardware (power, cooling, rack space, etc)
- Set up & configure hardware
- Install baseline software (OS, packages)

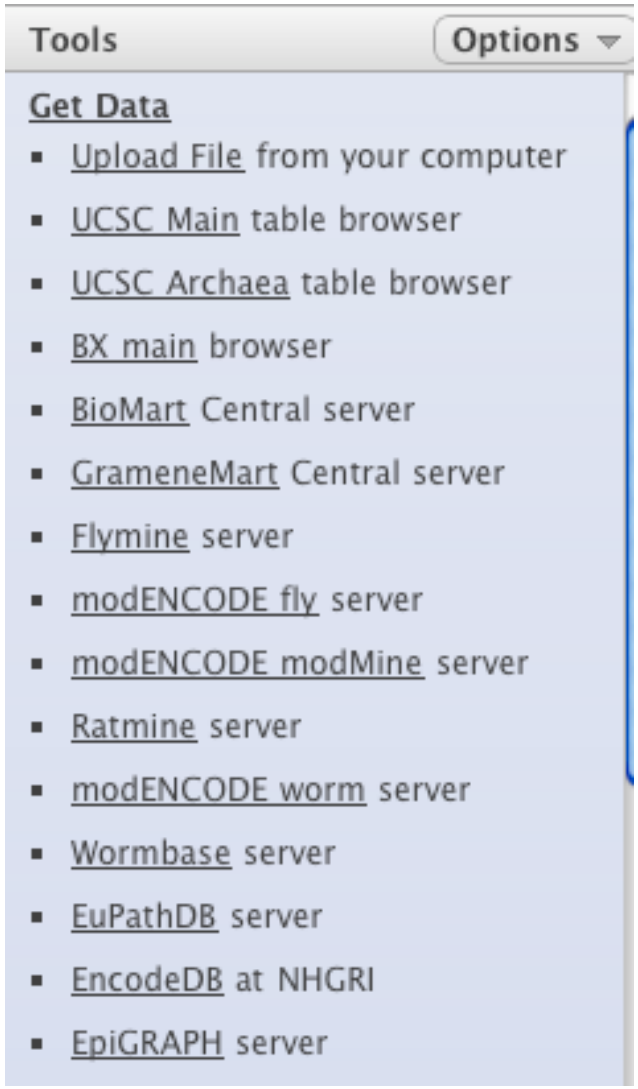
Simplify the complications



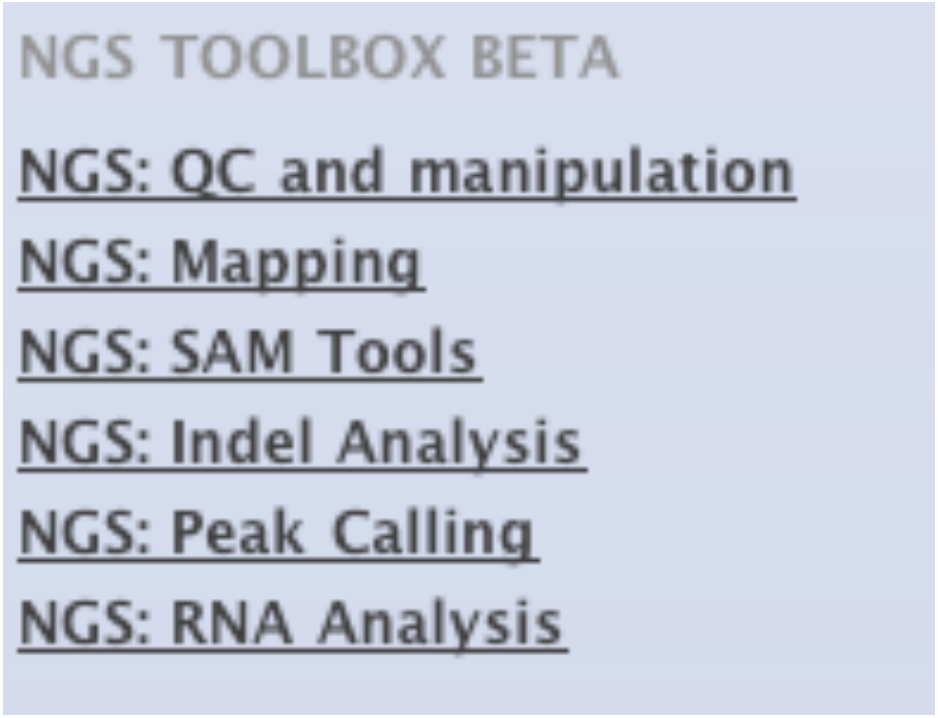
Galaxy


Galaxy
http://main.g2.bx.psu.edu/
galaxy
Baha'i Terminology
Amazon Web Services
Getting Started
Latest Headlines
Software packages fo...
Current FGI Sequence...
Fungal Genome Initia...
EvolDir
GAP_Analysis
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
Options
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
Here is what's happening...
Galaxy 101
Start small
The very first tutorial you need
Live Quickies
Basic fastQ manipulation:
Galactic quickie # 13
Advanced fastQ manipulation:
Galactic quickie # 14
454 Mapping: Single End
Galactic quickie # 15
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: $Rev 5070:ca0c4ad2bb39$
History
Options
1: UCSC Main on Human: all_mrna (genome)
~228,520 regions
format: bed, database: hg19
Info: UCSC Main on Human: all_mrna (genome)
view in GeneTrack
display at Ensembl Current
1.Chrom 2.Start 3.End 4.Name 5 6.S
chr1 66999065 67210057 BC040516 0 +
50724,155765,156807,162051,185911,195881,200
chr1 66999274 67210767 BX640813 0 +
743,207066,207680,209481,
chr1 66999276 67210767 CR749541 0 +
154,205741,207064,207678,209479,
javascript:parent.show_in_overlay({url:'http://screencast.g2.bx.psu.edu/galaxy/quickie_13_fastq_basic/flow.html',width:640,height:480,scroll:'no'})

Get data


Tools
Options
Get Data
Upload File from your computer
UCSC Main table browser
UCSC Archaea table browser
BX main browser
BioMart Central server
GrameneMart Central server
Flymine server
modENCODE fly server
modENCODE modMine server
Ratmine server
modENCODE worm server
Wormbase server
EuPathDB server
EncodeDB at NHGRI
EpiGRAPH server

Next-Gen tools



Great advantages

- Powerful interface
- Create work flow for repeated analysis
- Share work flow among a group
- Create local copy to access your data locally

Data interpretation
-- visualization

Visualization tools

- Genome browser
  - Ensembl Genome Browser (Stalker, Gibbins et al. 2004)
  - NCBI Entrez Map Viewer (Wolfsberg 2007)
  - Golden Path Genome Browser (Kent, Sugnet et al. 2002)
  - Apollo (Ed, Nomi et al. 2009)
  - Argo (www.broad.mit.edu/annotation/argo/).

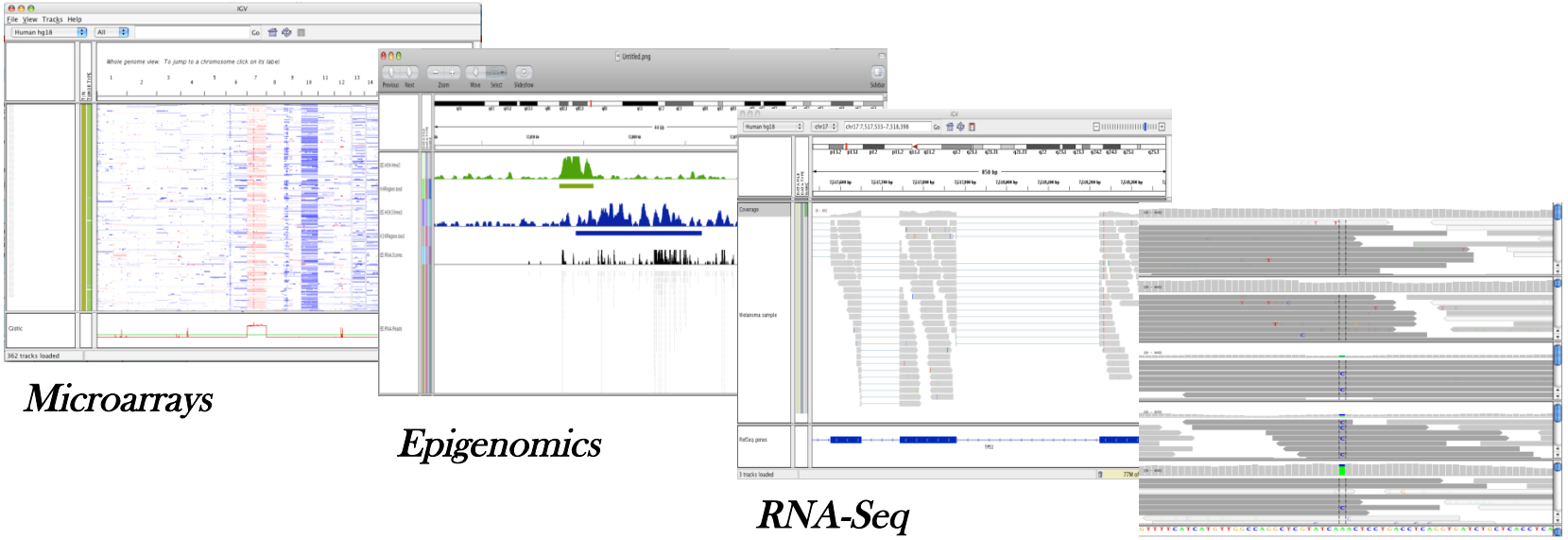
Integrated Genome View (IGV)

A desktop application

for integrated visualization

of multiple data types and annotations

in the context of the genome



Microarrays

Epigenomics

RNA-Seq

NGS alignments

•http://www.broadinstitute.org/igv

Adaptable

Data types

- Any data tied to genomic coordinates
- Genome annotations
- Sample attributes/annotations

File formats

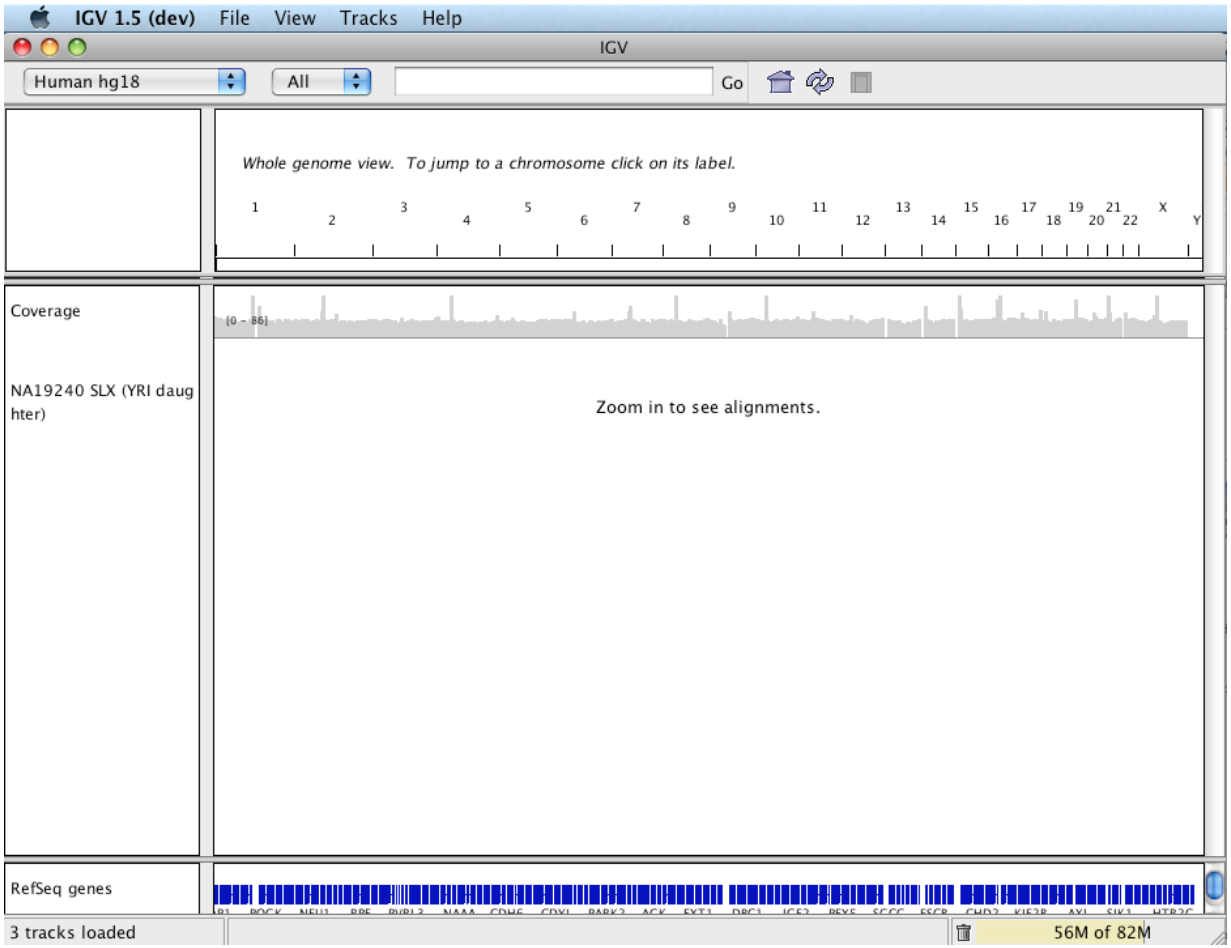
- Many different file formats supported

Data sources

- Local files
- URLs
- IGV data server

Viewing NGS data

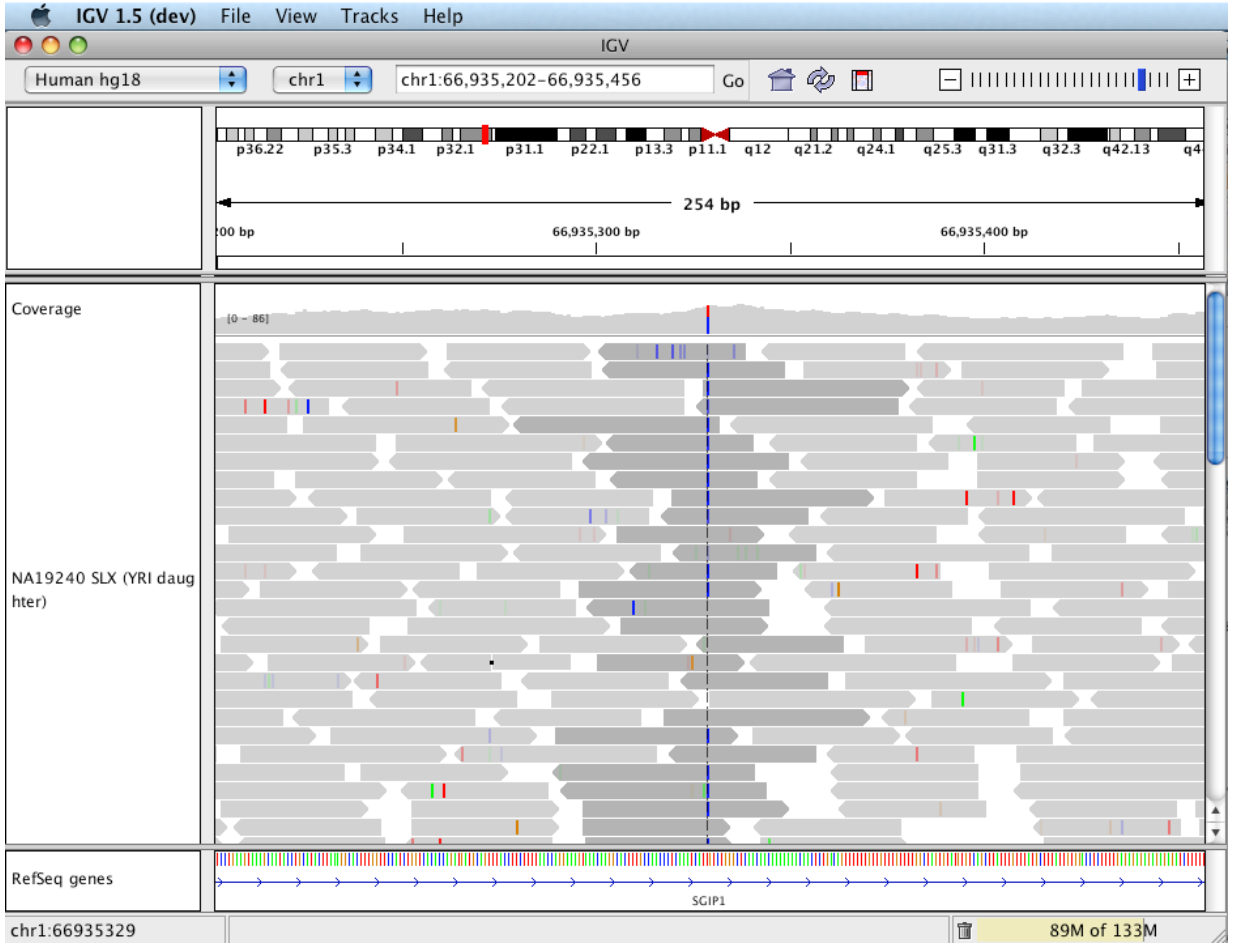
Whole genome view – coverage data only



Viewing NGS data

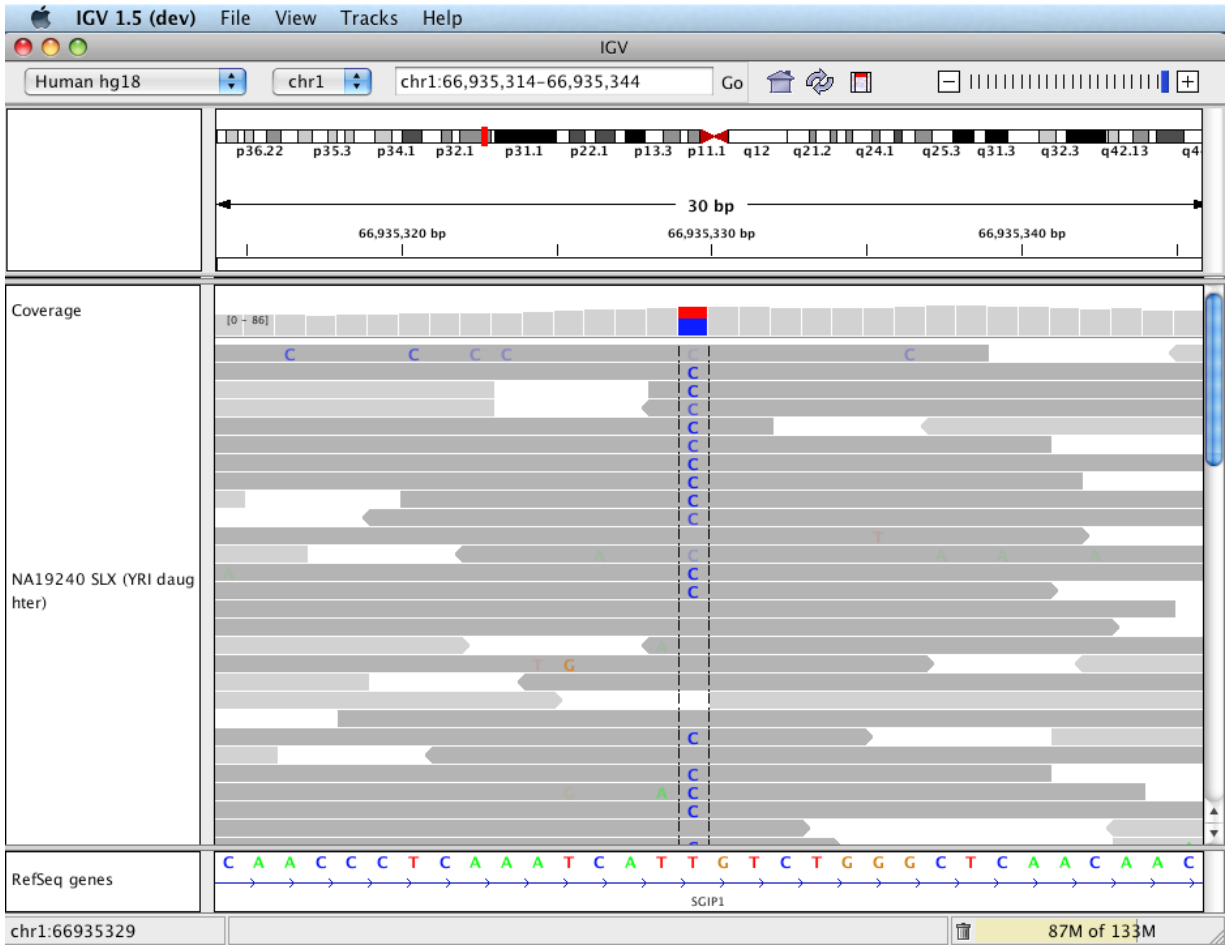
Zoomed in to 254 bp –

see more detail in alignments: base mismatches & read direction



Viewing NGS data

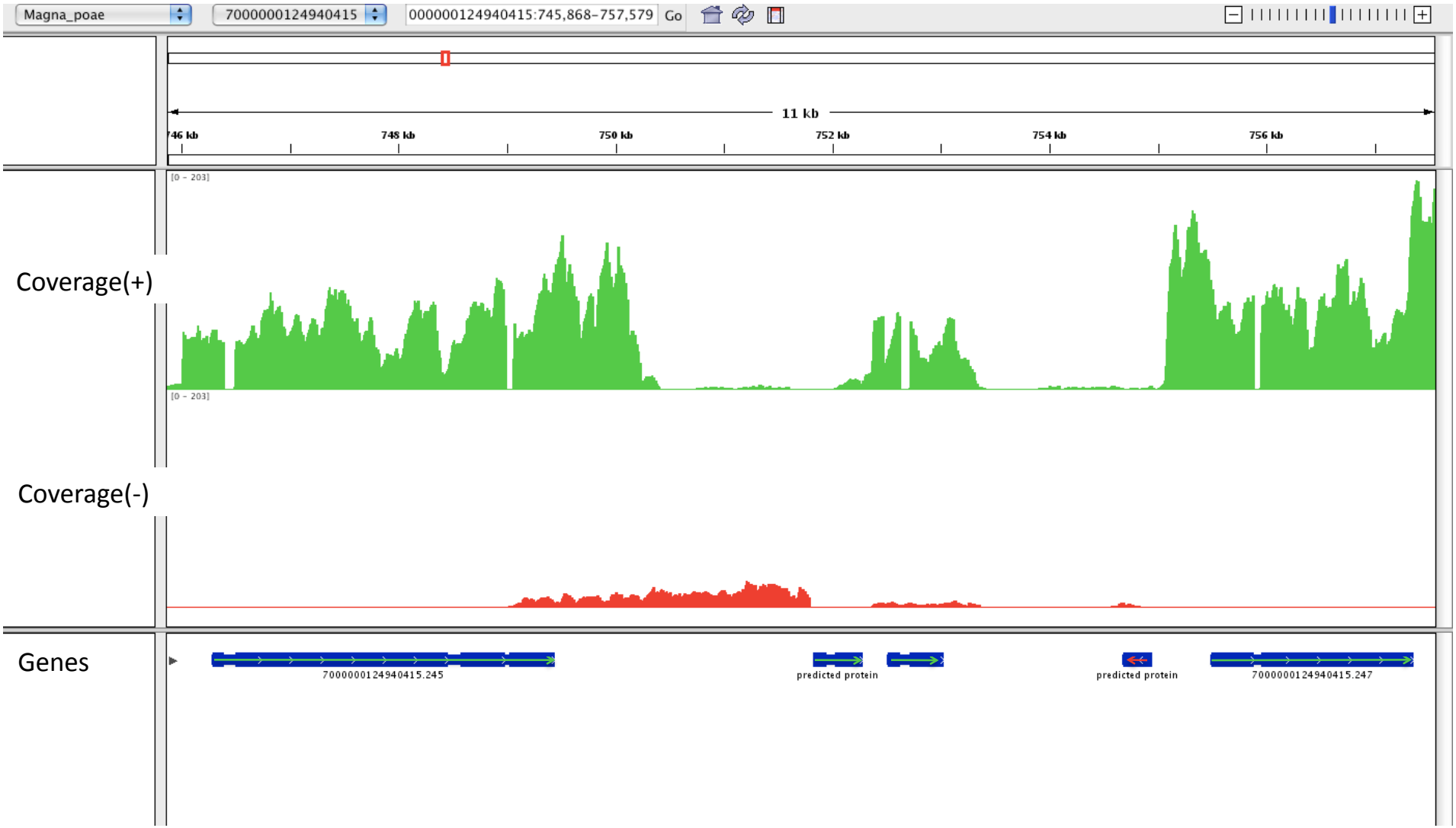
Zoomed in to 30 bp – see bases



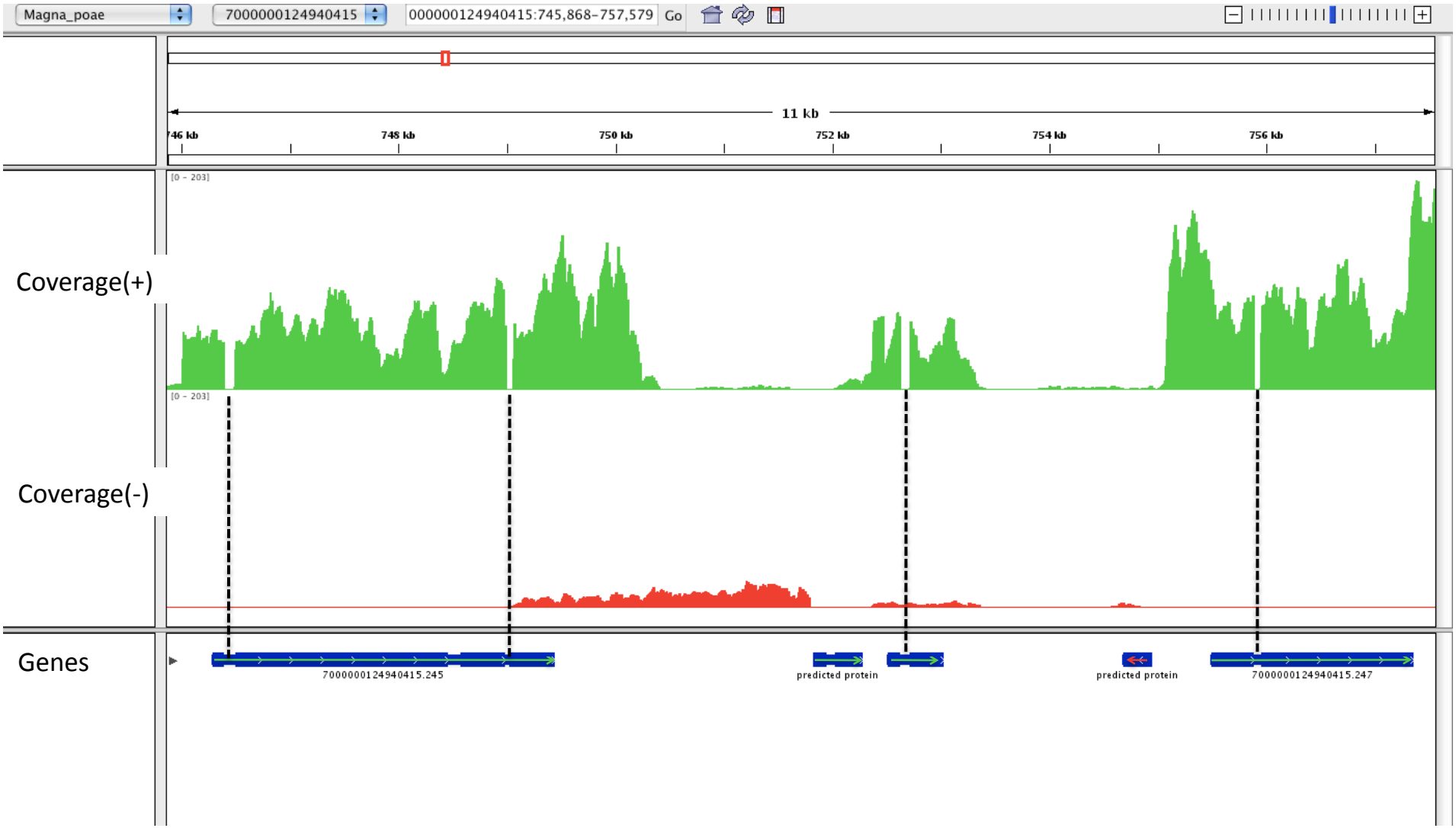
IGV RNA-seq


Magna_poae
All
Go
Whole genome view. To jump to a chromosome click on its label.
NAME
DATA FILE
DATA TYPE
Mpoae.cold_stress.+.bam.tdf
Mpoae.cold_stress.-.bam.tdf
Mpoae.sporulation.-.bam.tdf
Mpoae.sporulation.+.bam.tdf
Mpoae.heat_stress.-.bam.tdf
Mpoae.sporulation.+.bam.tdf
Mpoae.osmotic_stress.-.bam.tdf
Mpoae.control.-.bam.tdf
Mpoae.light_stress.+.bam.tdf
Mpoae.light_stress.-.bam.tdf
Mpoae.heat_stress.+.bam.tdf
Mpoae.osmotic_stress.+.bam.tdf
Mpoae.control.+.bam.tdf
Gene
threonine synthase
None
predicted protein
None
hypothetical protein
hypothetical protein
predicted protein
predicted protein
predicted protein
hypothetical protein
None
hypothetical protein
None
None

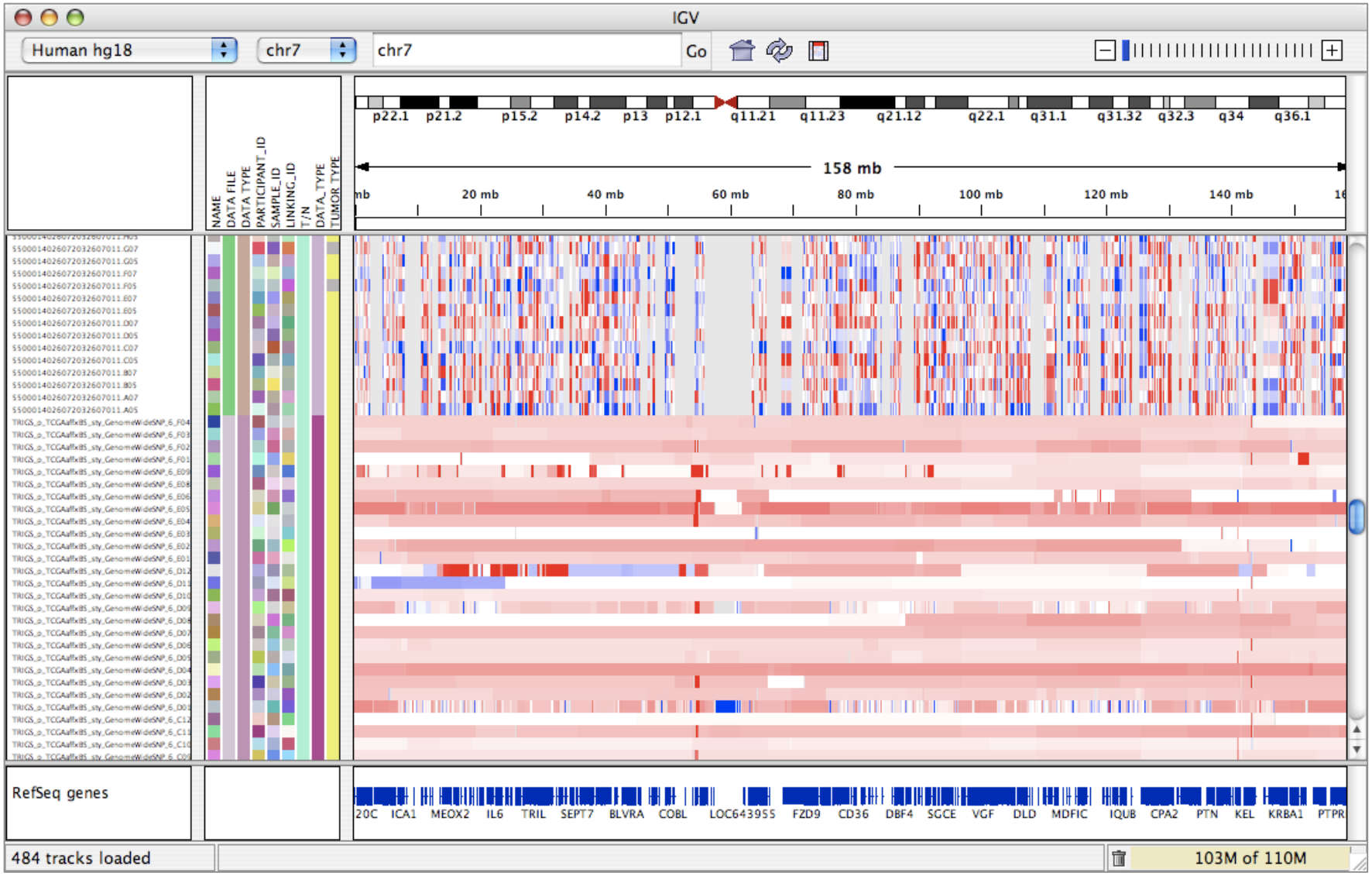
IGV RNA-seq


Magna_poae
7000000124940415
000000124940415:745,868-757,579
Go
11 kb
746 kb
748 kb
750 kb
752 kb
754 kb
756 kb
[0 - 203]
Coverage(+)
[0 - 203]
Coverage(-)
Genes
7000000124940415.245
predicted protein
predicted protein
7000000124940415.247

Annotation


Magna_poae
7000000124940415
000000124940415:745,868-757,579
Go
11 kb
746 kb
748 kb
750 kb
752 kb
754 kb
756 kb
[0 - 203]
Coverage(+)
[0 - 203]
Coverage(-)
Genes
7000000124940415.245
predicted protein
predicted protein
7000000124940415.247

Heat Map


IGV
Human hg18
chr7
chr7
Go
p22.1 p21.2 p15.2 p14.2 p13 p12.1 q11.21 q11.23 q21.12 q22.1 q31.1 q31.32 q32.3 q34 q36.1
158 mb
mb 20 mb 40 mb 60 mb 80 mb 100 mb 120 mb 140 mb
NAME
DATA FILE
DATA TYPE
PARTICIPANT_ID
SAMPLE_ID
LINKING_ID
T/N
DATA_TYPE
TUMOR TYPE
RefSeq genes
20C ICA1 MEOX2 IL6 TRIL SEPT7 BLVRA COBL LOC643955 FZD9 CD36 DBF4 SGCE VGF DLD MDFIC IQUB CPA2 PTN KEL KRBA1 PTPR
484 tracks loaded
103M of 110M