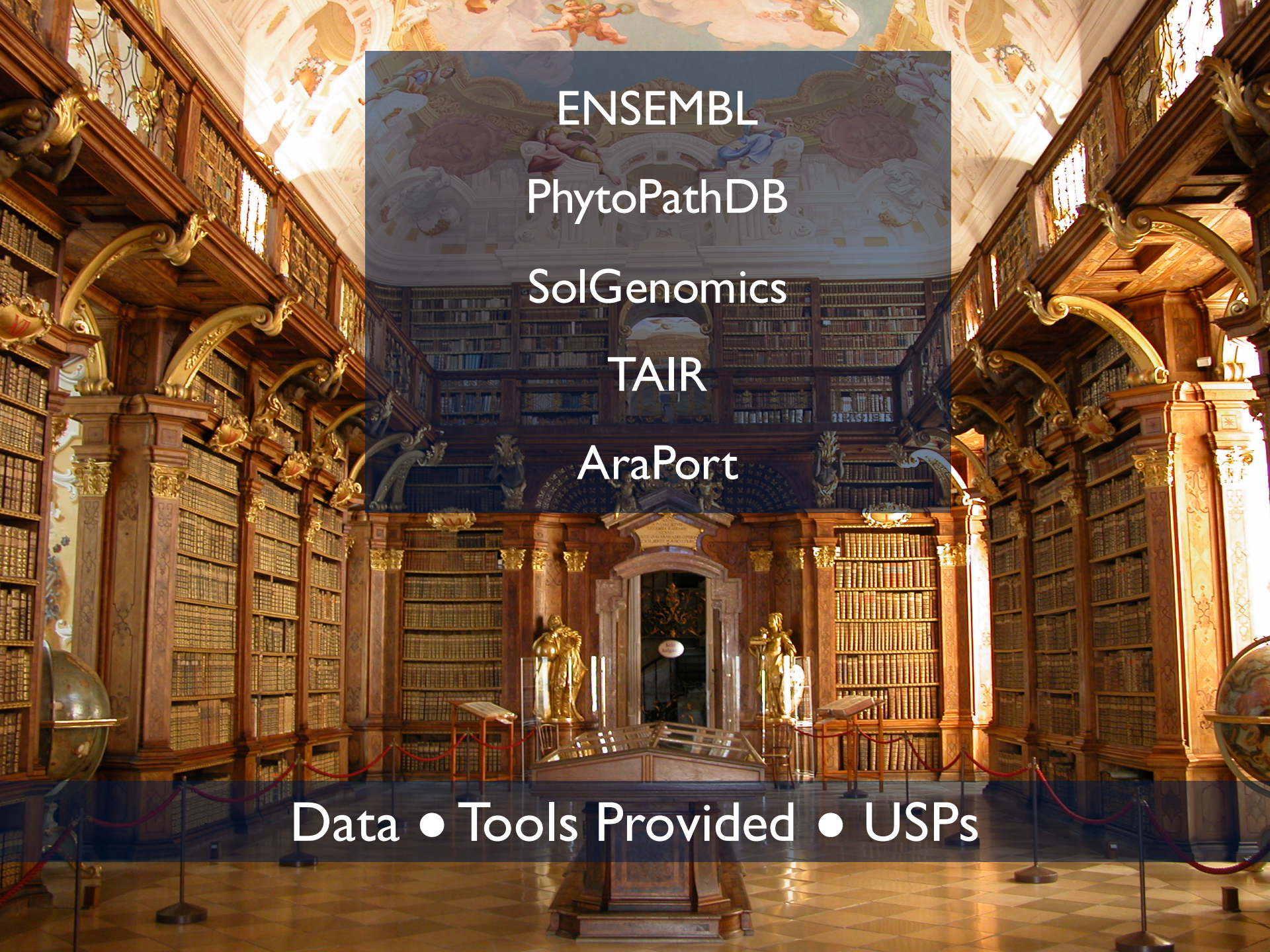

ENSEMBL
PhytoPathDB
SolGenomics
TAIR
AraPort
Data • Tools Provided • USPs

ENSEMBL

url: ensembl.org

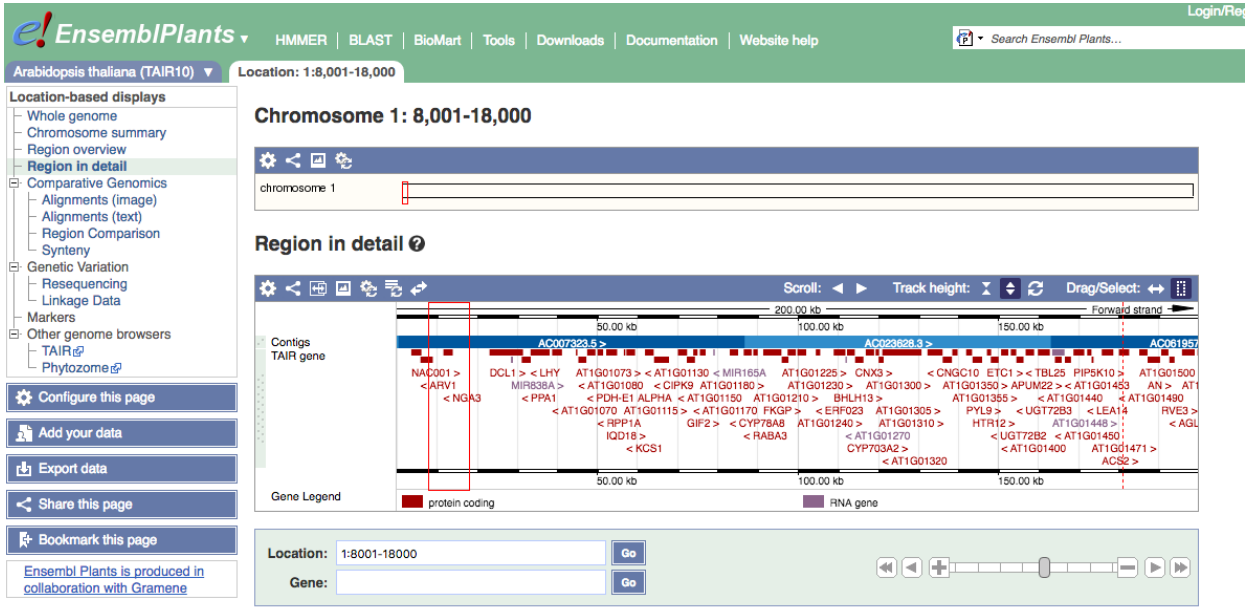
Family of DBs
(Human is the flagship)

Assemblies • Genes • Annotations • Expression • Variants • Comparative Genomics

ENSEMBL PLANTS

url: plants.ensembl.org

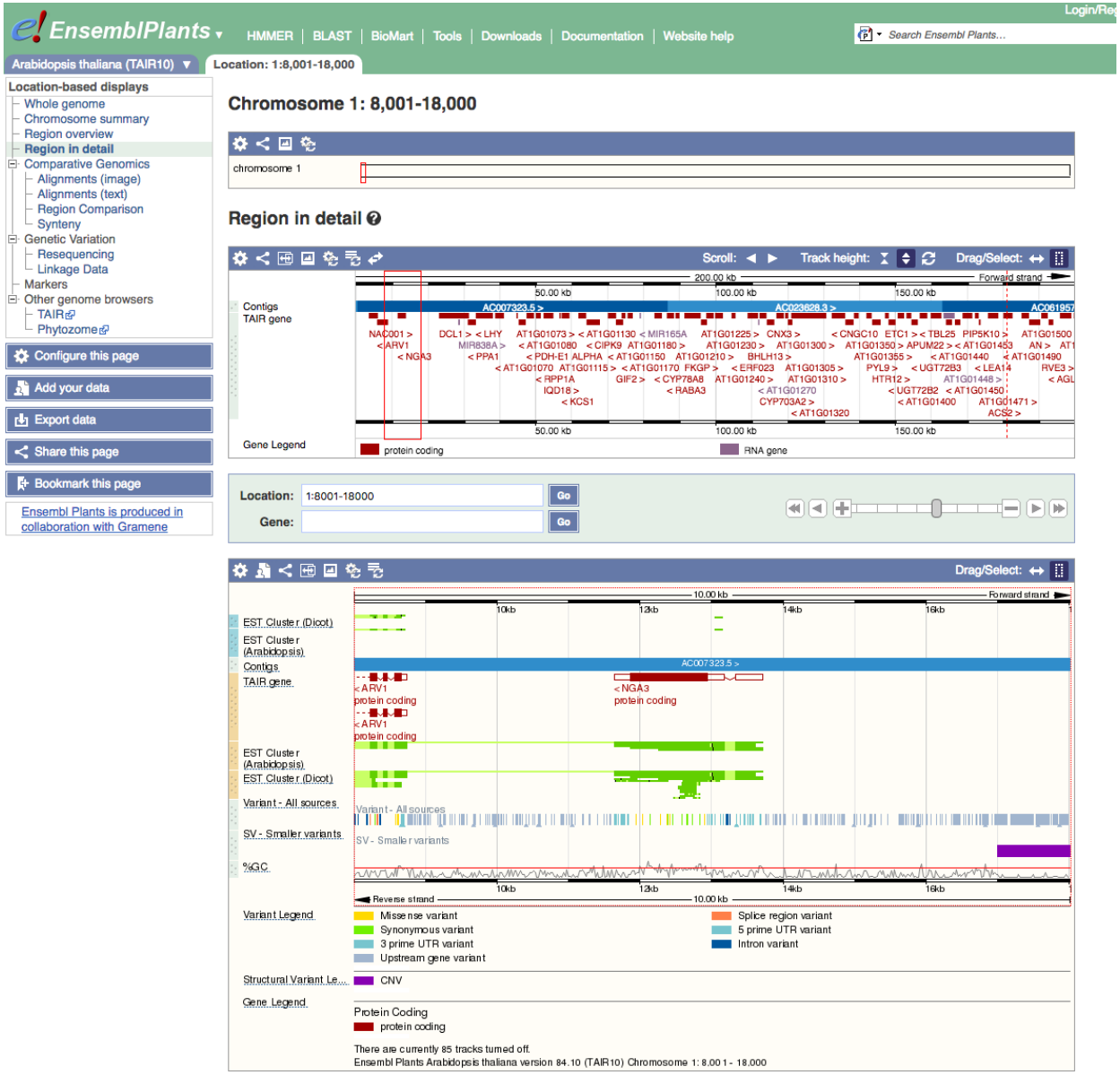
39 Plant Species



ENSEMBL PLANTS

url: plants.ensembl.org

39 Plant Species



ENSEMBL PLANTS

url: plants.ensembl.org

Gene
- trees – gene families, across plant kingdom, wider!
- secondary structure
- orthologs
- GO terms
- PO terms
- genomic alignments
- gene expression (outlinks)
- variants

Transcript
- exons
- splice variants
- cDNA

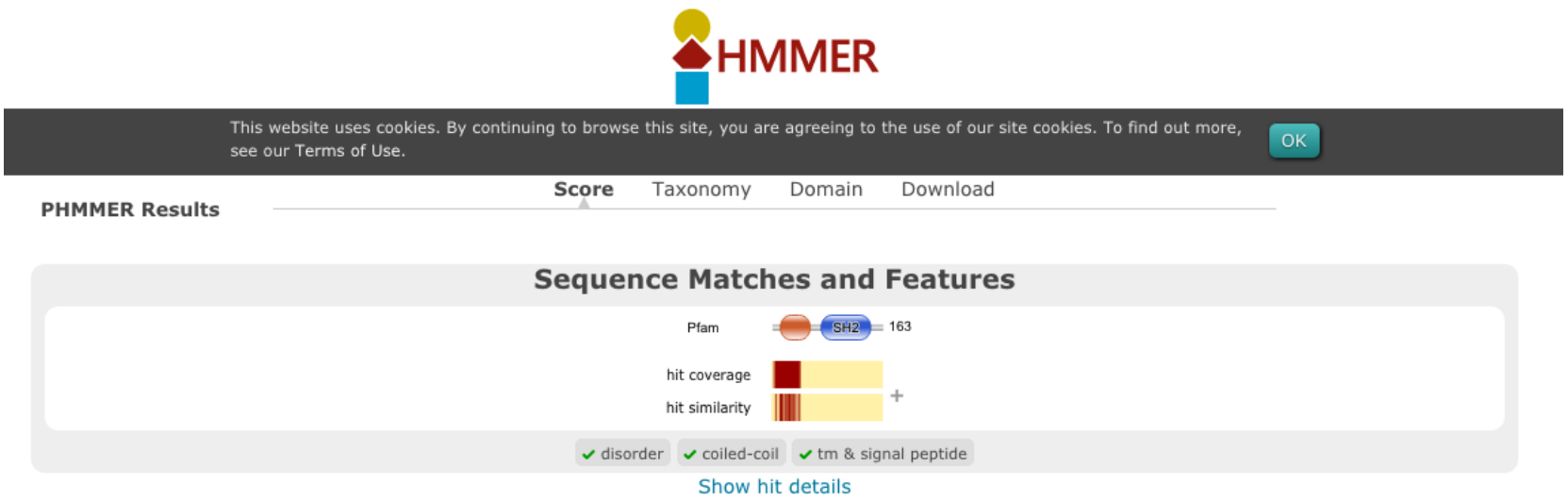
Protein
- domain
  - prints, PFAM, PROSite, Panther, Gene3D
- variants

ENSEMBL PLANTS

url: plants.ensembl.org

Tools

HMMER



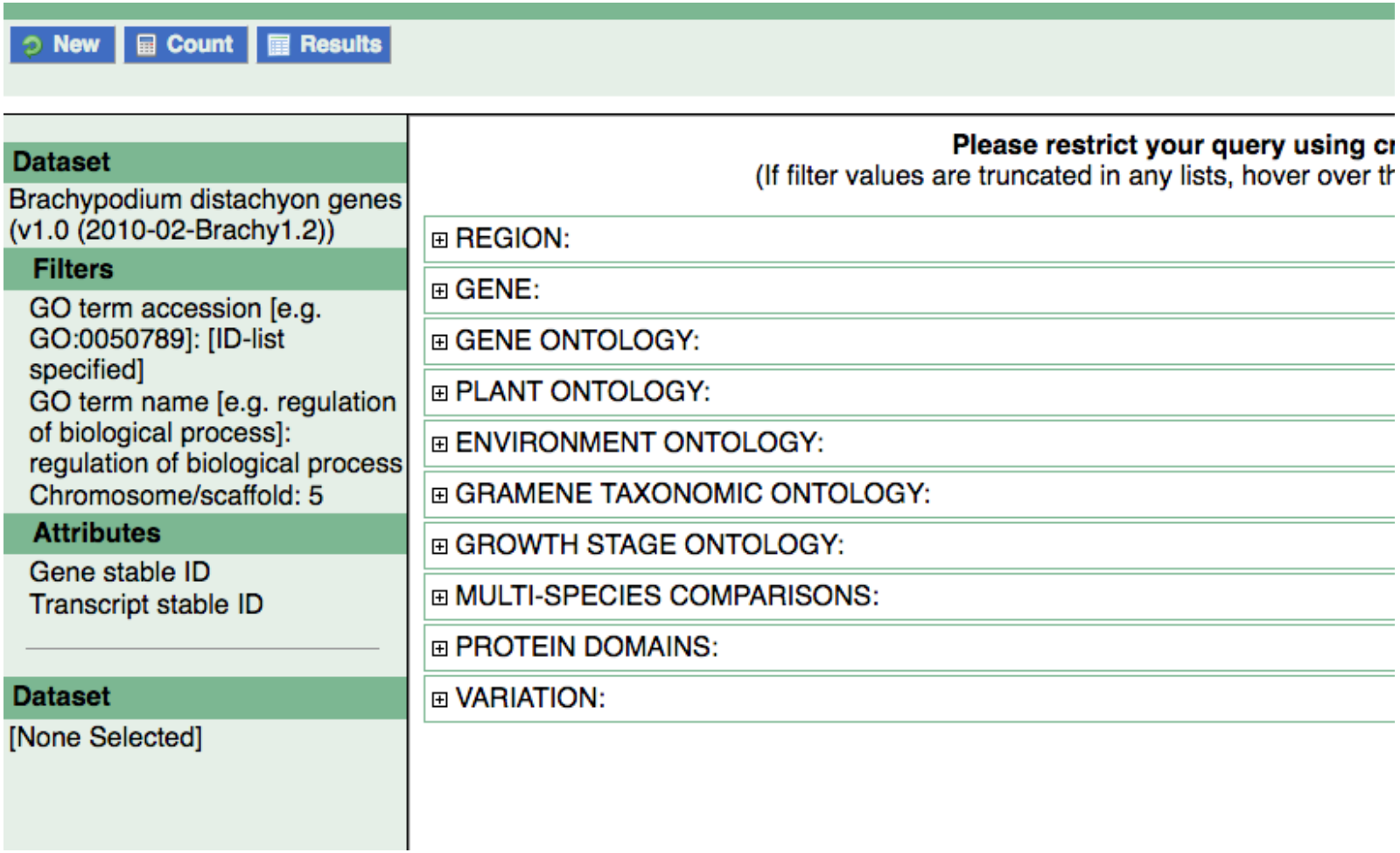
BLAST

ENSEMBL PLANTS

url: plants.ensembl.org

Tools

BIOMART



ENSEMBL FUNGI, BACTERIA, PROTISTS

url: fungi.ensembl.org, bacteria.ensembl.org, protists.ensembl.org

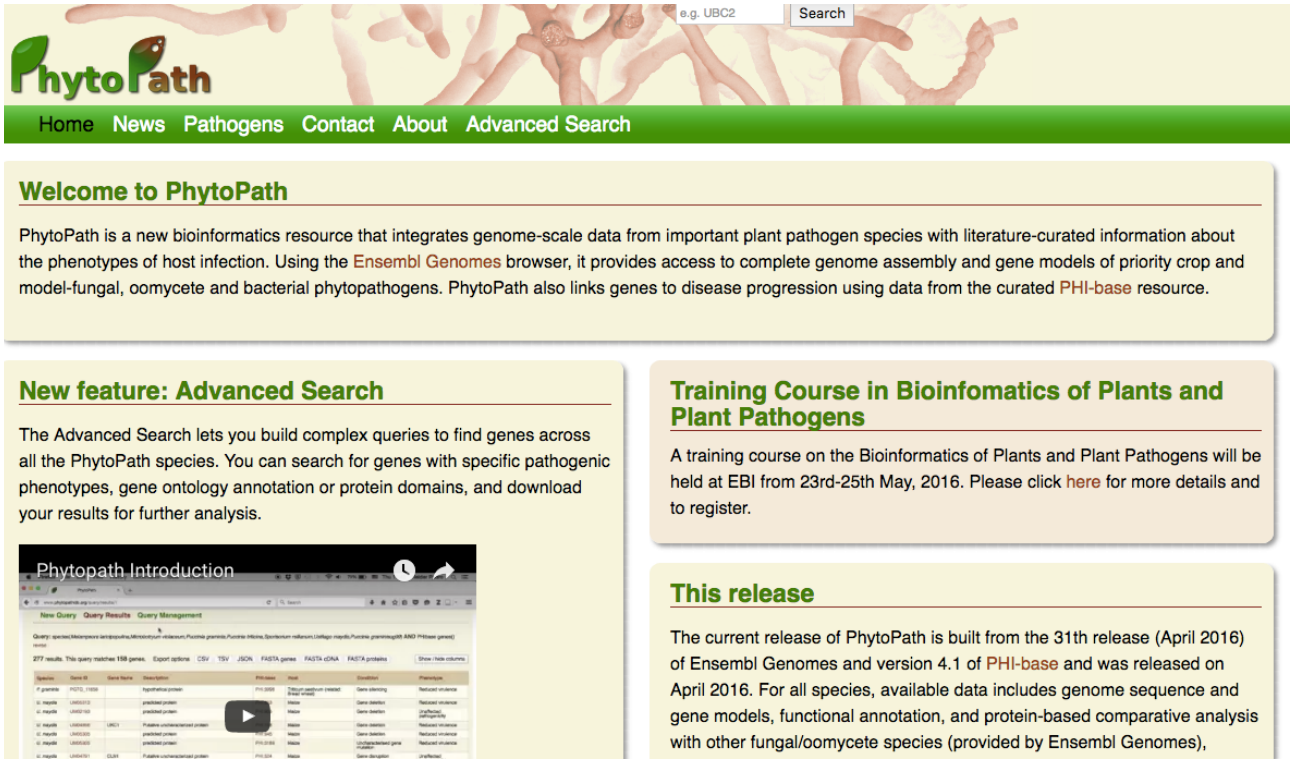
- *Gene*
  - ~~trees – gene families, across plant kingdom, wider!~~
  - secondary structure
  - orthologs
  - ~~GO terms~~
  - ~~PO terms~~
  - ~~genomic alignments~~
  - ~~gene expression (outlinks)~~
  - variants
- *Transcript*
  - exons
  - splice variants
  - cDNA
- *Protein*
  - domain
    - prints, PFAM, PROSite, Panther, Gene3D
  - variants

PhytoPath

url: phythopath.org

Based on ENSEMBL

Portal for Plant Pathogens



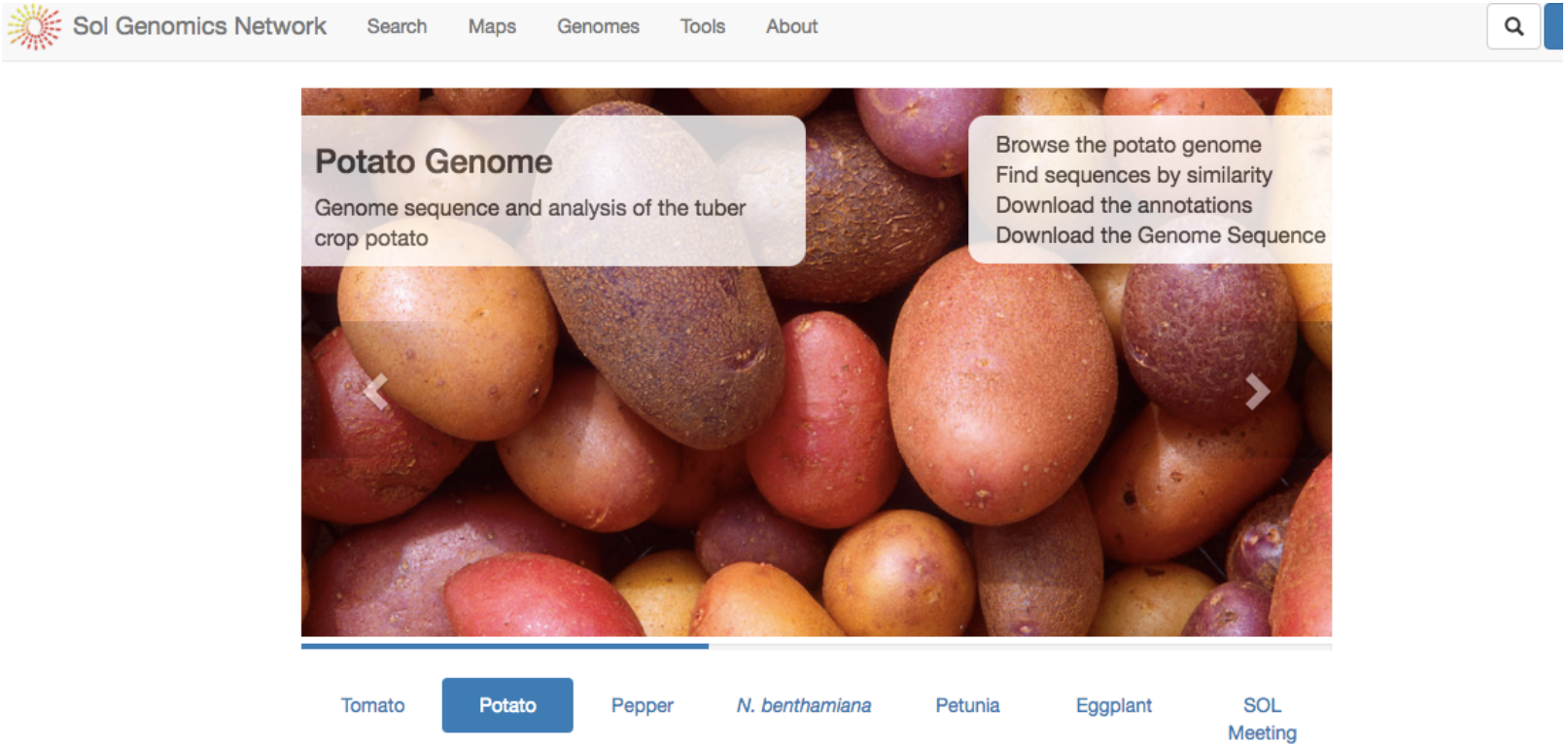
Literature Information,

disease progression info from PHI-BASE

SolGenomics

url: solgenomics.net

Community Site for Solanaceae

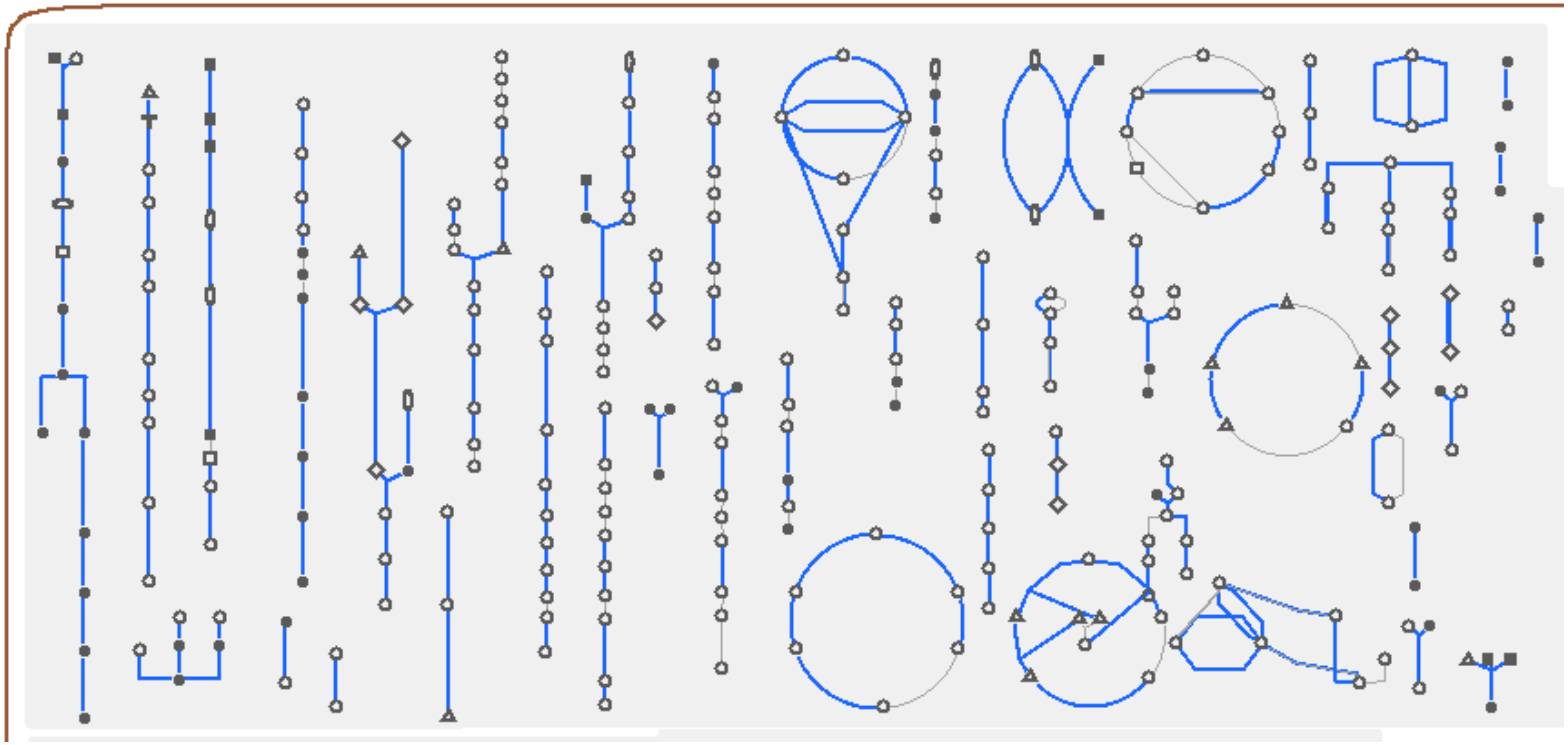


SolGenomics

url: solgenomics.net

Tools

Biochemical pathways

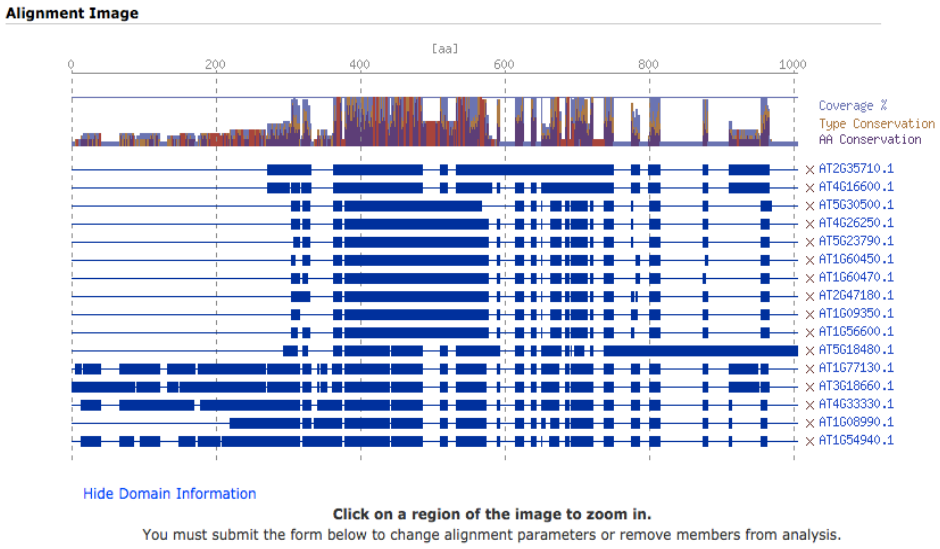


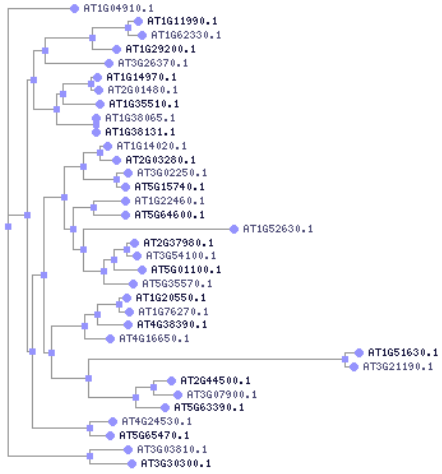
SolGenomics

url: solgenomics.net

Tools

- CAPS experiment designer
- VIGs tool
- Alignment Analyser
- Tree Browser





TAIR

url: arabidopsis.org

Gold Standard DB for Arabidopsis

Data

- Genes - Transcripts – Proteins - Genomes
- GO / PO
- Locus history
- Markers
- Array elements, expression, labs
- Protocols, Publications
- Seeds, Germplasm

Tools

- Gbrowse
- Synteny browser
- Protein-Protein Interaction Browser
- AraCyc Metabolic Pathways
- BLAST & WU-BLAST
- Pattern/Motif Finders
- Expression / Array Clustering

Portals

- Clone stocks
- Education
- Gene Expression
- Gene Annotation
- Metabolome
- Mutant (Genetics)

AraPort

url: araport.org

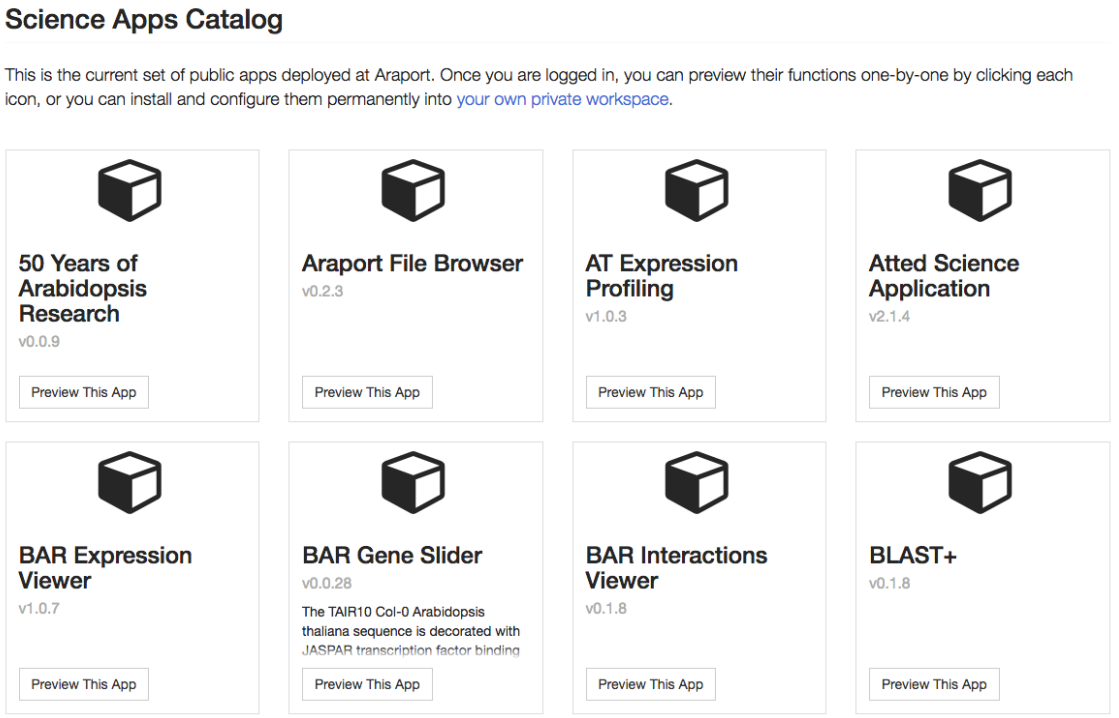
Next-Gen Portal for Arabidopsis

Data

- TAIR 10 Genes - Transcripts – Proteins - Genomes
- Araport 11
- Live community annotation

Tools

- JBrowse
- App Store
  - 20+ apps
  - sequence tree views
  - search
  - phosphorylation
  - interactions
  - BLAST+



Genome Assembly and Annotation

Best Practices, Best Tools

Structure

AIM:

Help you develop a strategy for performing a genome assembly and assessing its quality

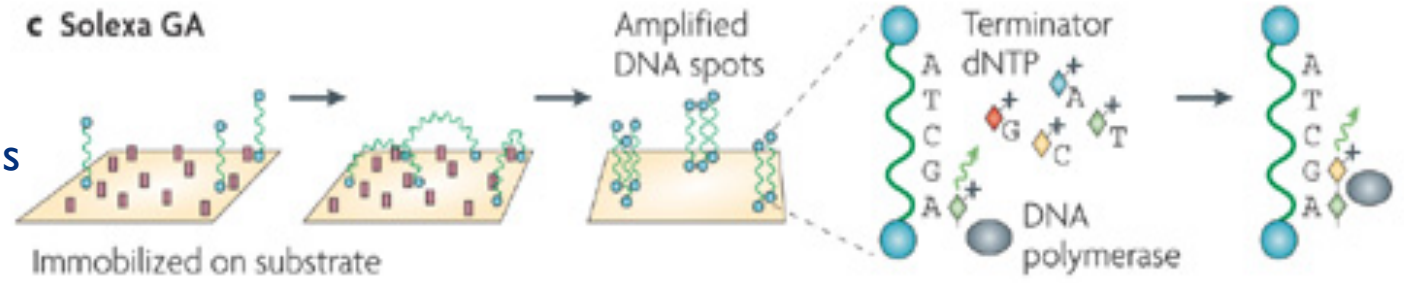
- Theory
  - Sequence Types
  - Algorithm Basics
  - Assessing an Assembly
- Tools
  - For Assembly
  - For Assessment
  - For Annotation

Sequencing

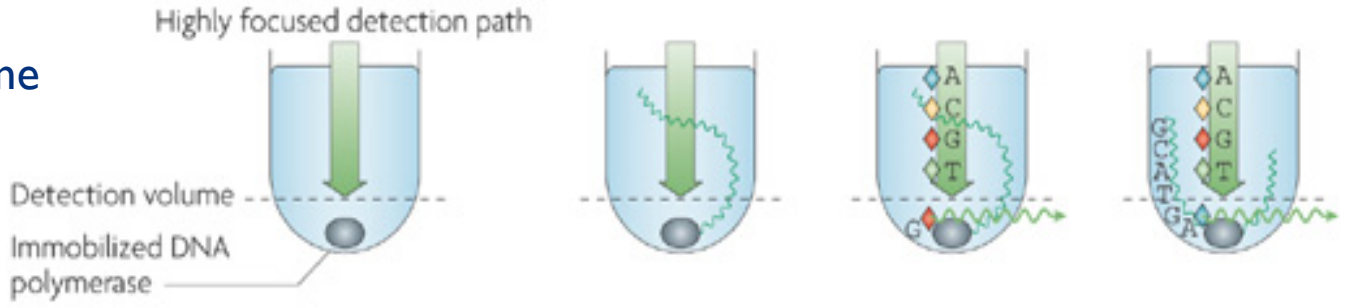
Two major kinds

- Illumina
- Pacific BioSciences

Illumina
Sequence by Synthesis



PacBio
Single Molecule Real Time



Paired-end and Mate Paired Reads

Paired End
Read both ends of the molecule
Separated by ~ 500 bp

Mate Pair
Circularize long molecules (1-10kbp)
Shear into fragments then sequence

Brief sequence comparison

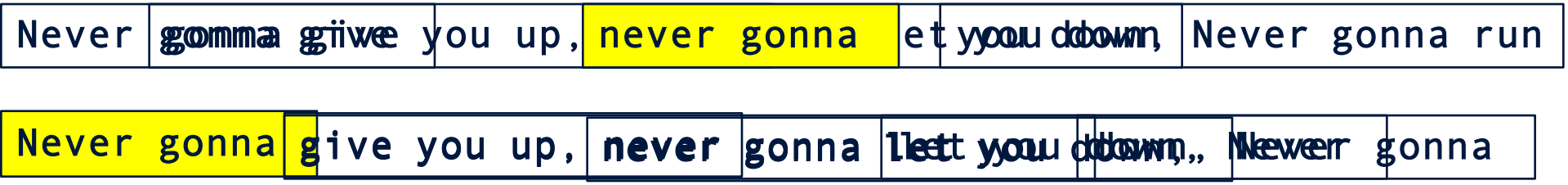
	Illumina	PacBio
Length	<300bp	<10kb
Reads per run	25 – 400 million	~50 thousand
accuracy	>99%	~86%

The Assembly Problem

Rebuilding a shredded text

Rick Astley accidentally shreds the lyrics to his first album…

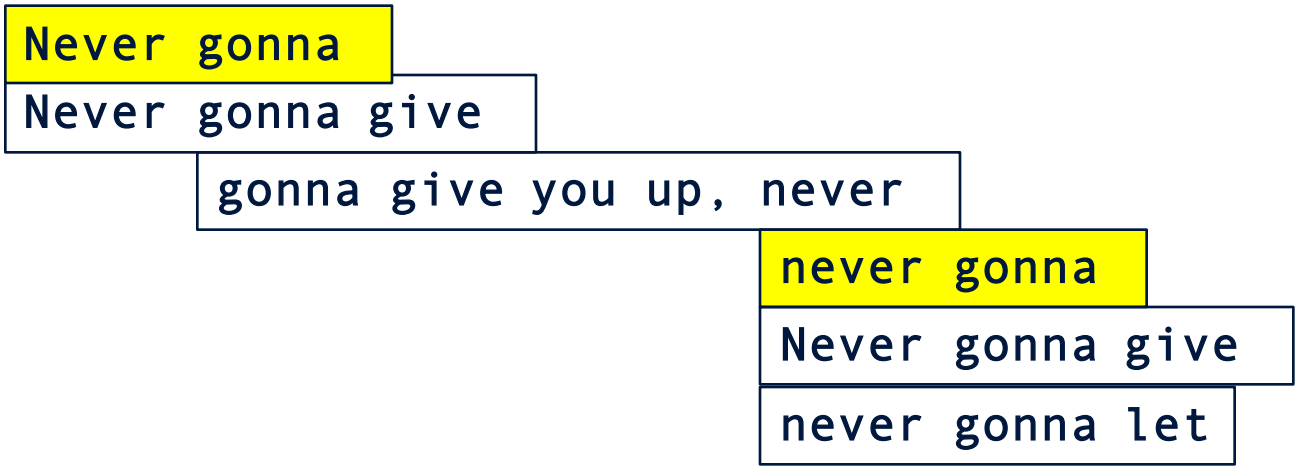
```
Never gonna give you up, never gonna let you down, Never gonna
run around and desert you, Never gonna make you cry, never gonna
say goodbye, Never gonna tell a lie and hurt you
```



How can he put them back together?

- The fragments from all copies are mixed together
- Some are identical

Greedy Algorithm

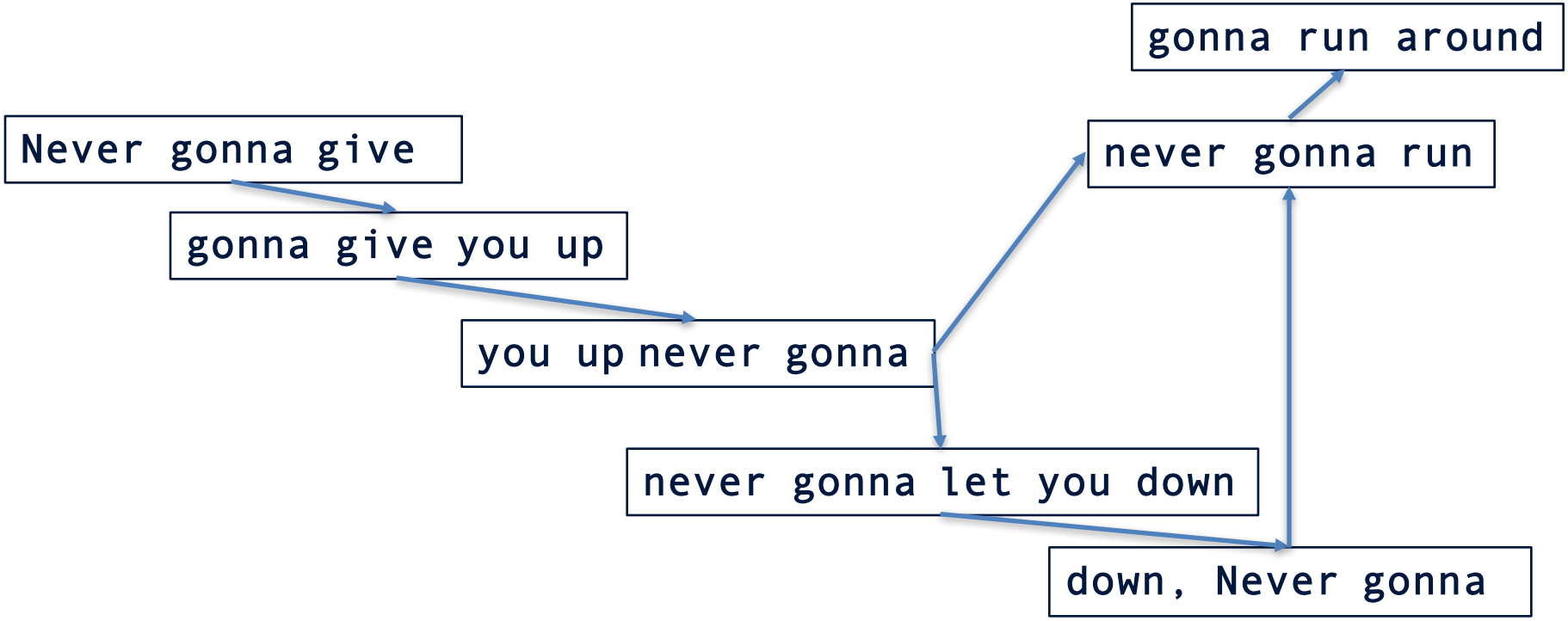


Repeated sections make the proper assembly ambiguous

de Bruijn Graph

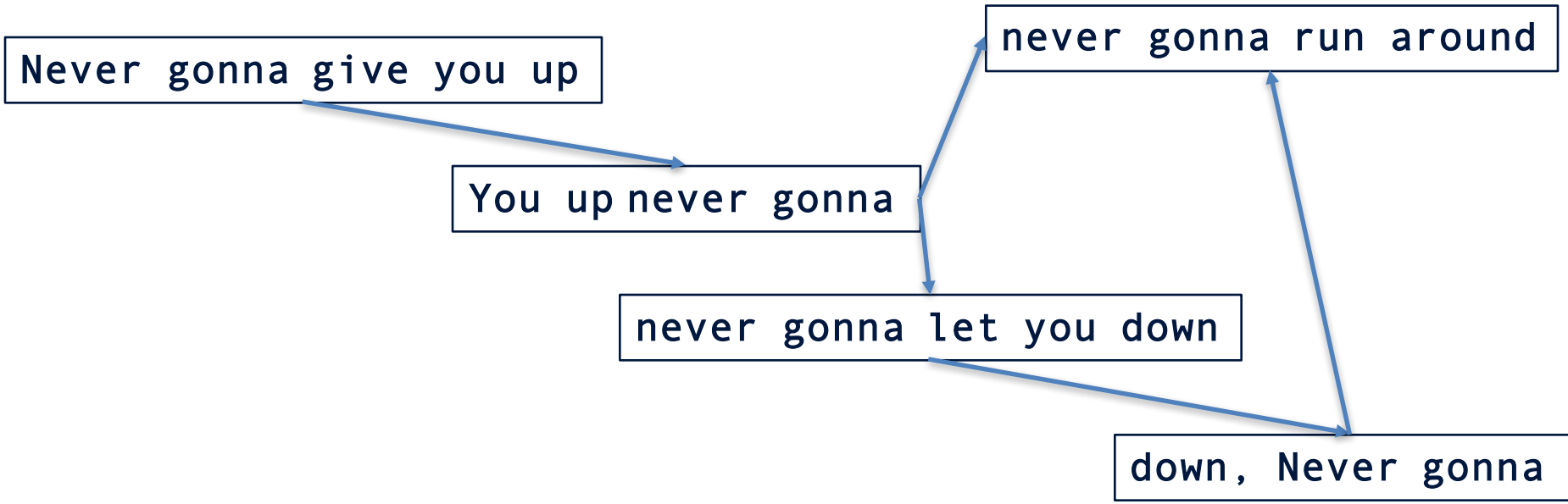
Make all ‘k’ length fragments
Network of possible directional links – from overlaps

de Bruijn Graph Assembly



Build Graph and simplify

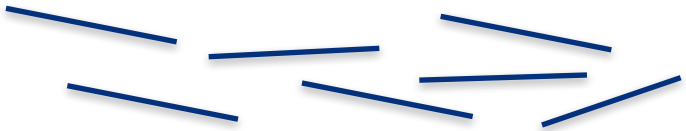
de Bruijn Graph Assembly



Build graph and simplify

Assembling a genome

1. Shear and sequence DNA

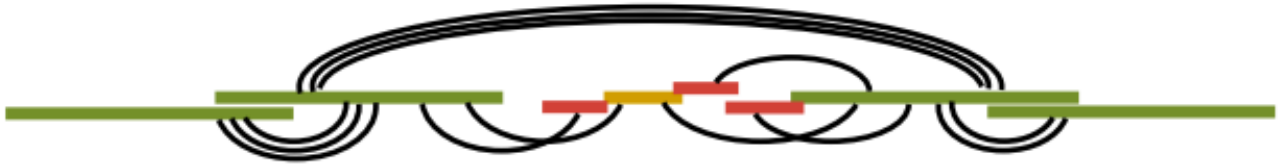


2. Make assembly graph from overlapped reads

AGCCTTAGACCTACAGGATGCGCGCACAGT
GATGCGCGCACAGTAGCCTTAGACT

3. Simplify assembly graph

4. Detangle graph with long reads, mate pairs and other links



Why are genomes hard to assemble?

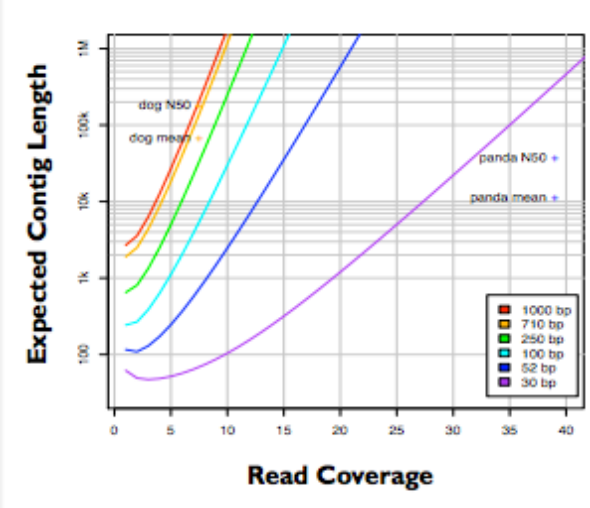
1. Biological Problems
   Polyploidy, heterozygosity
   Repeats, repeats

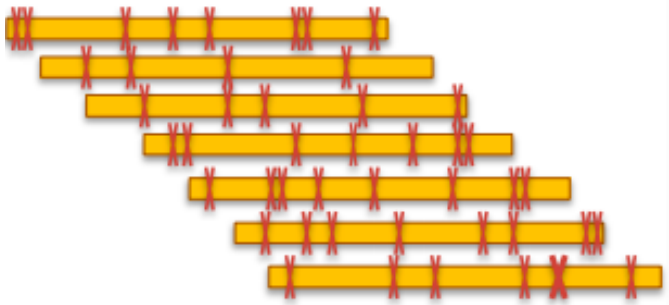
2. Sequencing problems
   Large genomes genome >>> reads
   sequence read errors

3. Computational
   Very large genomes
   Complex repeat structure

4. Accuracy
   What is correct assembly?
   How do we tell?

What you need for a good assembly





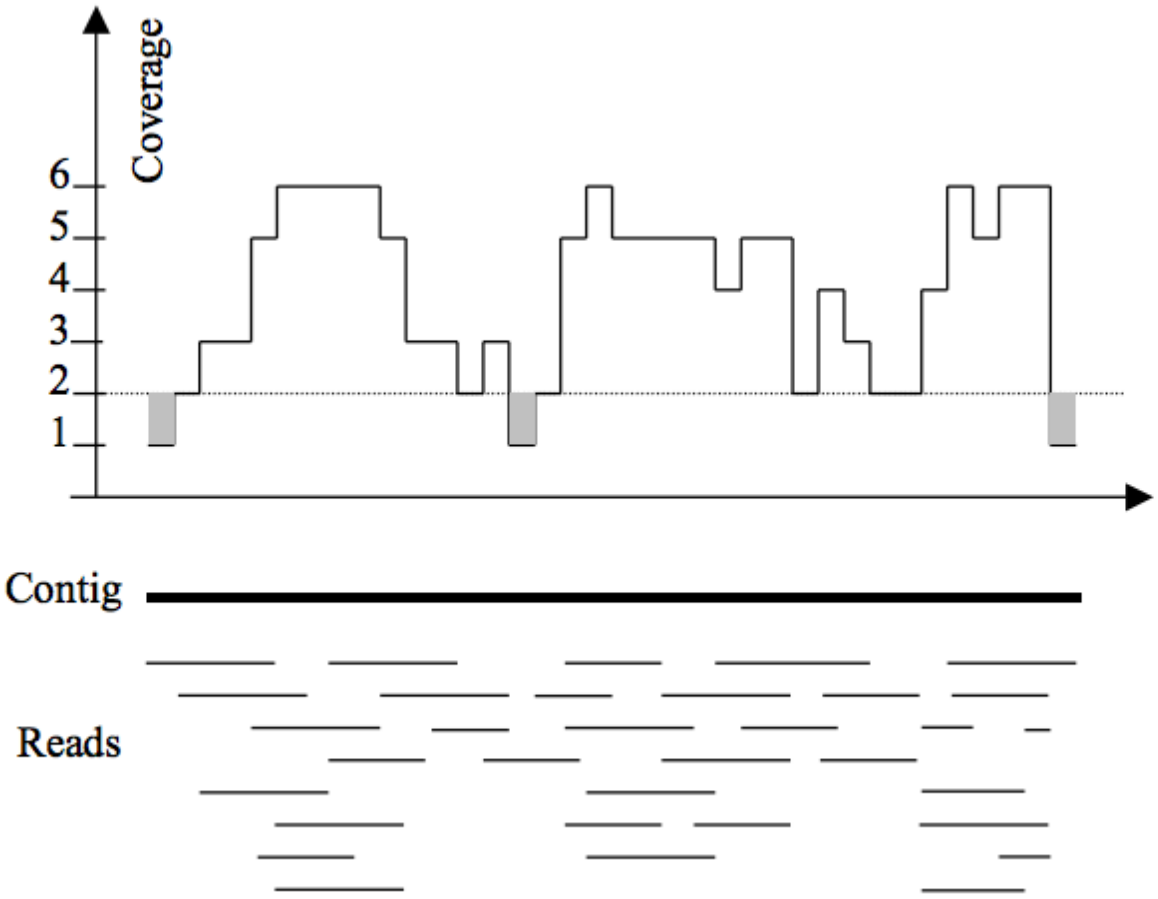
Coverage
- Oversample

Read Length
- Longer than repeats

Quality
- Errors mess up overlaps

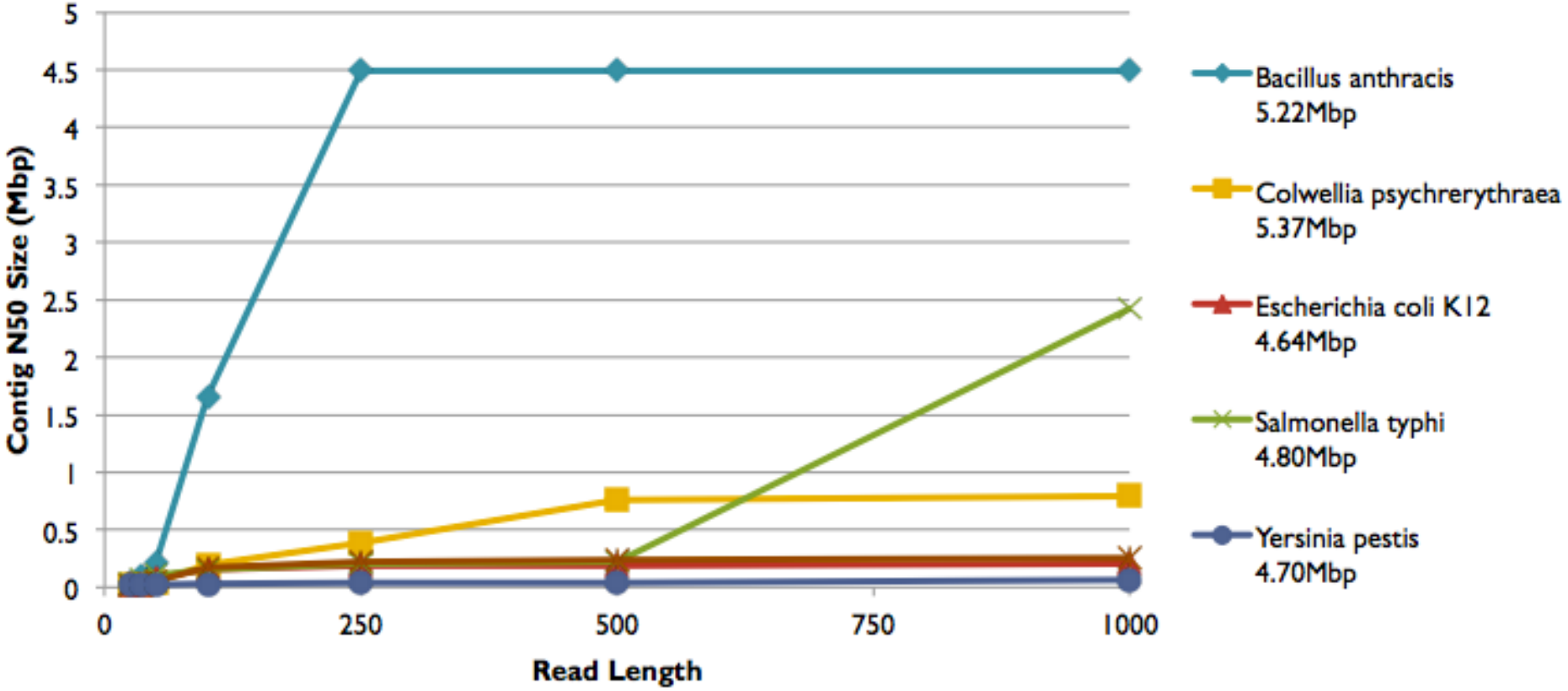
Coverage

Times sampled



Some segments may not even be sequenced!

Read Length and Repeats



Assembly length increases with read length, especially when it passes the repeat length

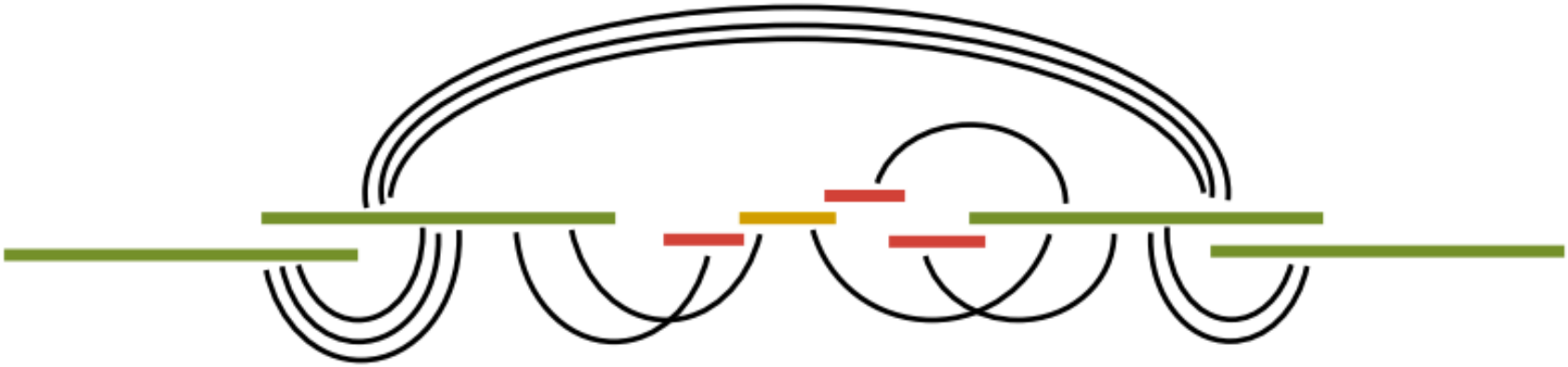
Scaffolding

Contigs end at

- coverage gaps
- errors
- repeat boundaries

Use paired end/mate information to place contigs relative to each other

- has gaps



Assembly Tools

Loads! Including but not limited to ..

SOAPdenovo, HyDA, ABySS, ALLPATHS-LG, Velvet, Celera, PacBio Corrected Reads, Mira, meraculous, Newbler, Ray, SGA, CLC Assembly Cell …

Assembly Tools

SOAPdenovo

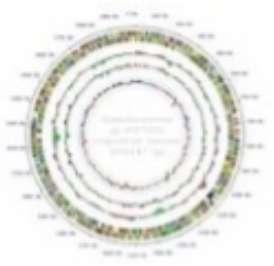
BGI's assembler
(Li et al. 2010)

De bruijn graph
Short reads

Most flexible, but requires a lot of tuning

http://soap.genomics.org.cn/soapdenovo.html

Celera Assembler



JCVI's assembler
(Miller et al. 2008)

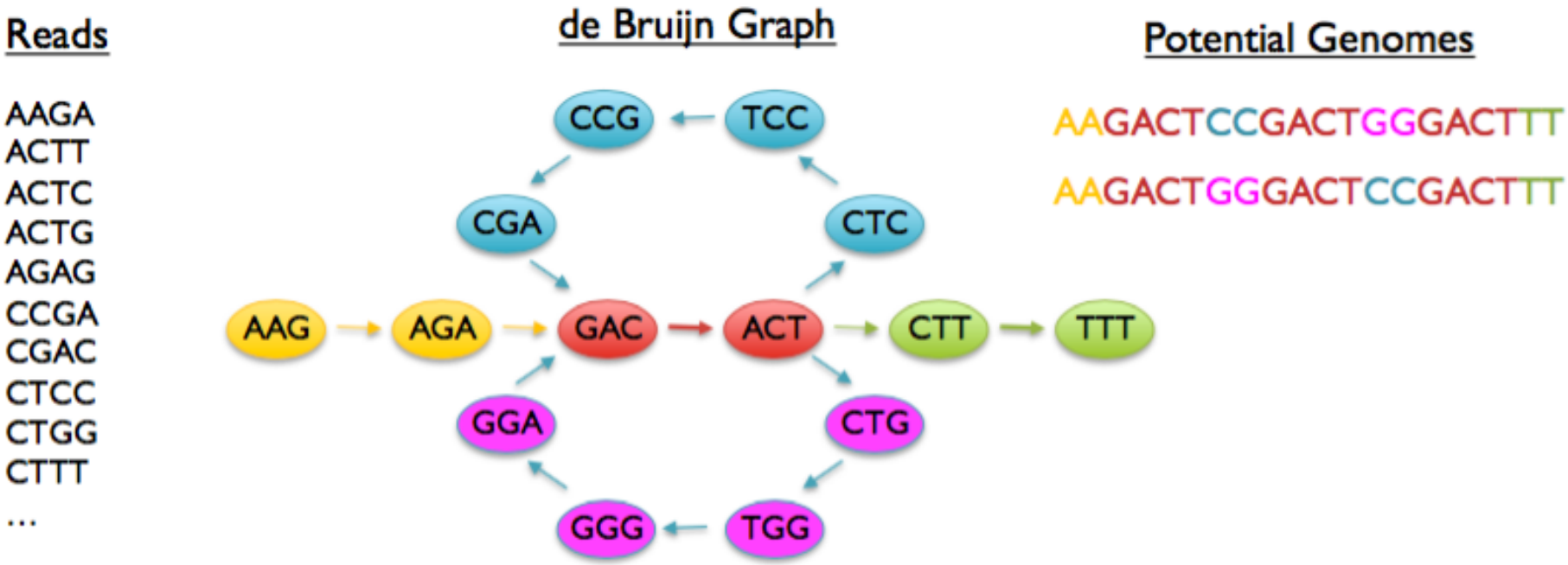
Overlap graph
Medium + Long reads

Supports Illumina/454/PacBio Hybrid assemblies

http://wgs-assembler.sf.net

Short read assembly

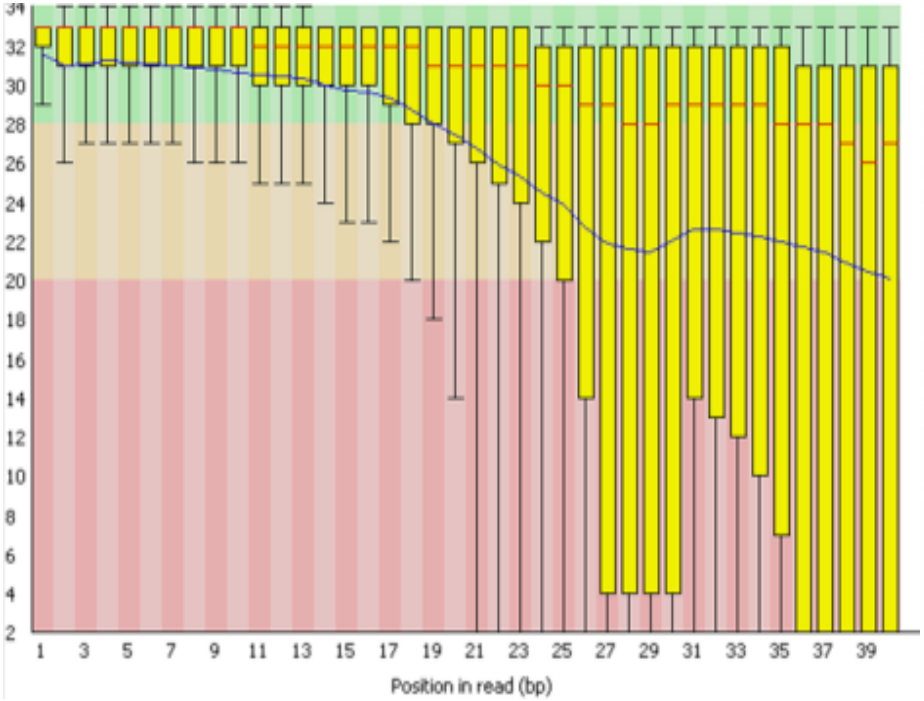
SOAPdenovo



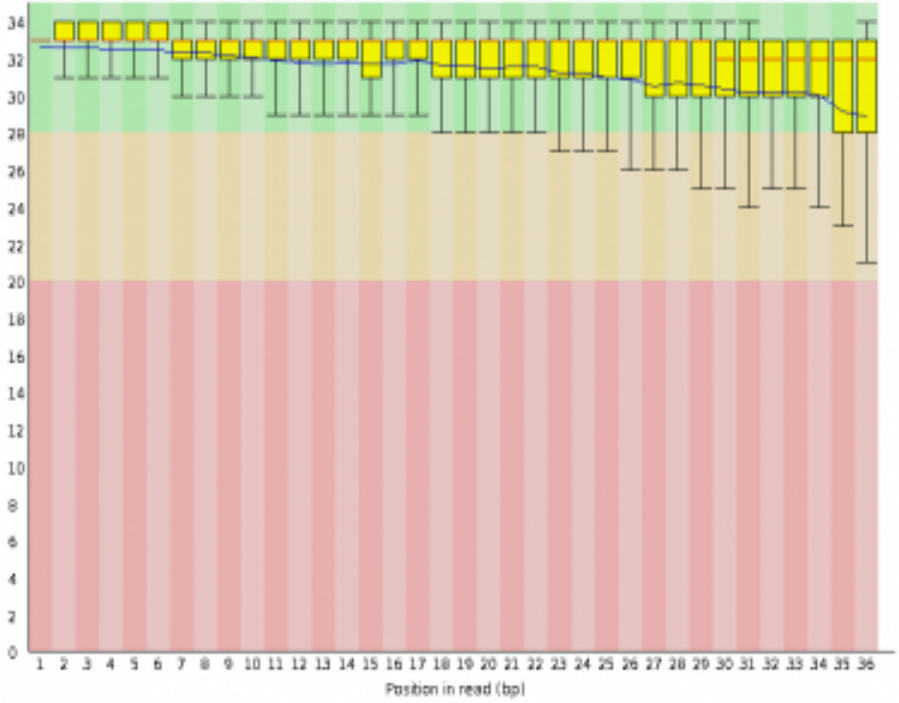
1. Get a big computer
    1. Different specs for every genome and sequence data.
    2. > 128Gb RAM for Pathogen genomes
    3. > 1Tb for Plant genomes

PreProcessing

FASTQC plots quality scores and sequence metrics and allows development of filter thresholds for whole sequence sets



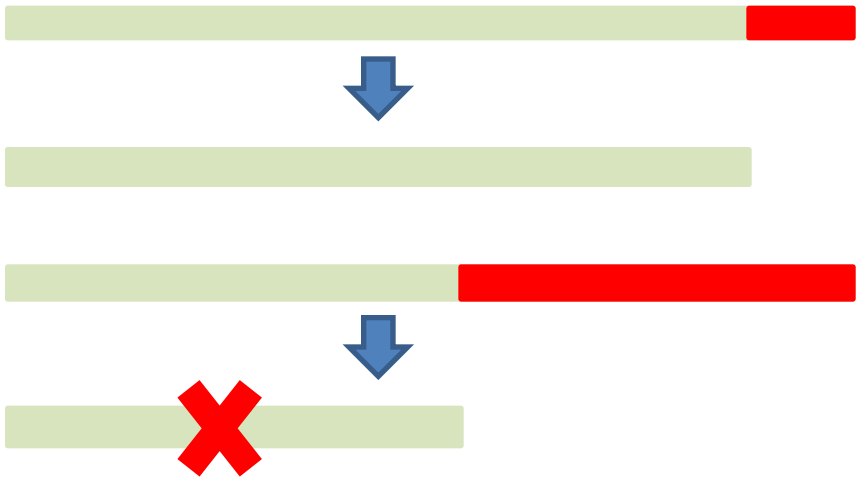
before



after

PreProcessing

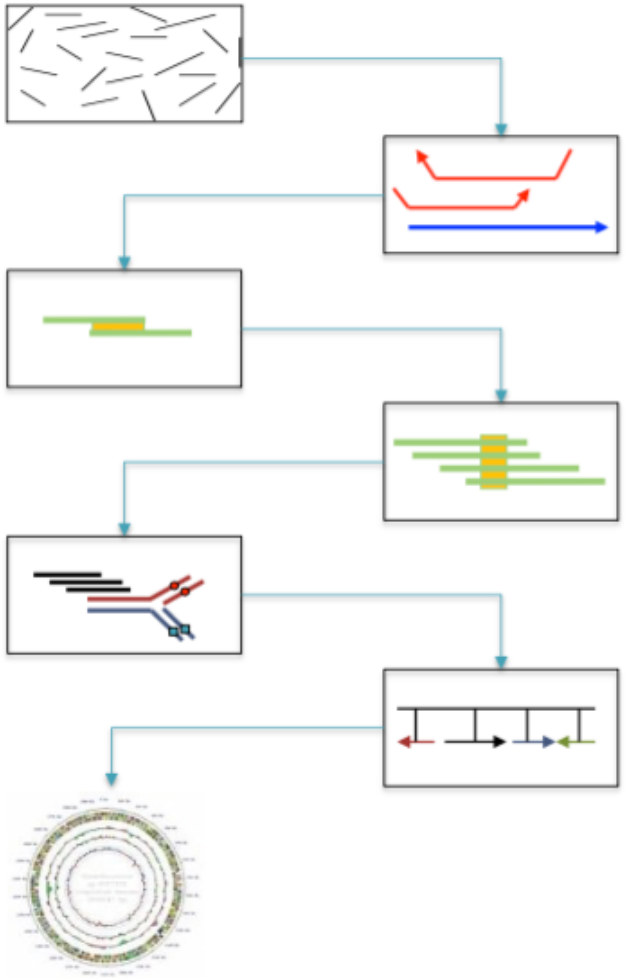
Trimmomatic uses a sliding window approach from the 5` end to identify low quality regions which are then trimmed from the 3` end. Reads < 36 bp are discarded



Long read assembly

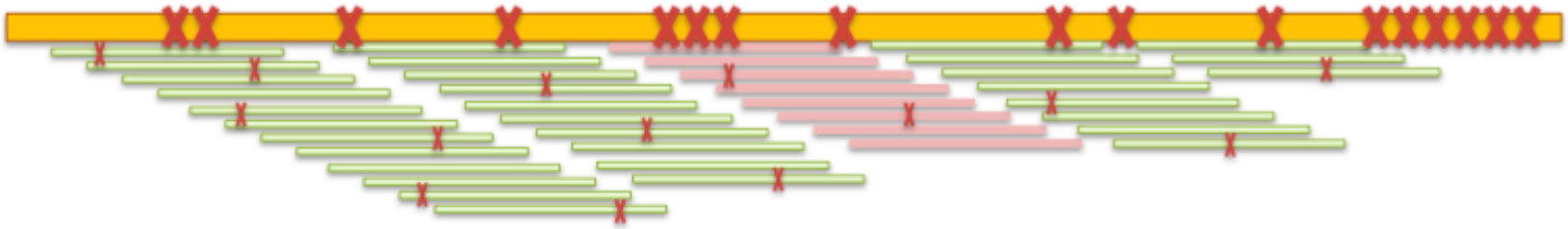
Celera Assembler

1. Pre-overlap
   - Consistency checks
2. Trimming
   - Quality trimming & partial overlaps
3. Compute Overlaps
   - Find high quality overlaps
4. Error Correction
   - Evaluate difference in context of overlapping reads
5. Unitigging
   - Merge consistent reads
6. Scaffolding
   - Bundle mates, Order & Orient
7. Finalize Data
   - Build final consensus sequences



Hybrid assembly

1. Map Illumina reads to PacBio reads to correct errors
2. Assemble



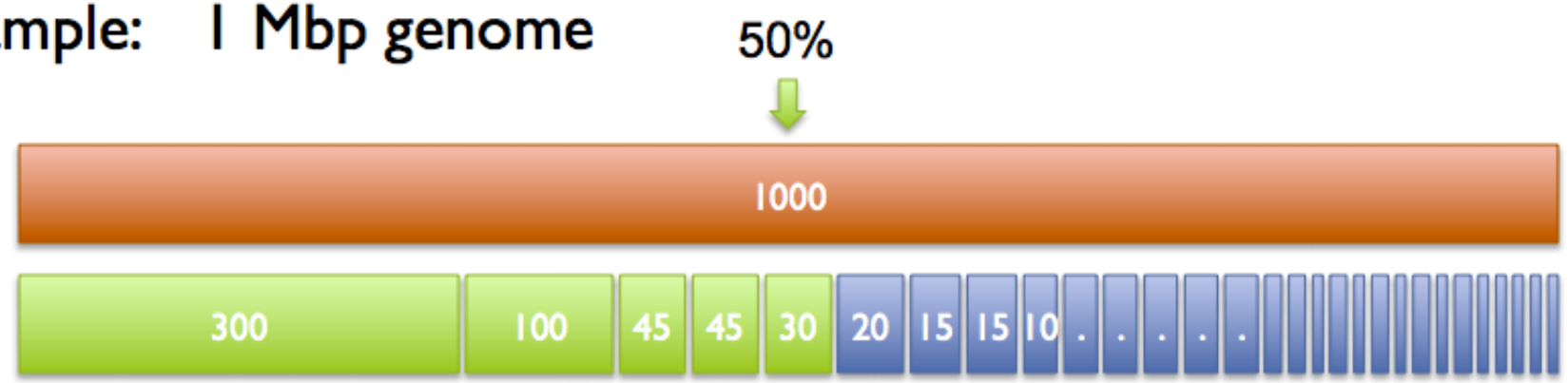
Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, et al. (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

Assessing an assembly

N50

50 % of genome in contigs longer than the N50 contig

Example: 1 Mbp genome



N50 size = 30 kbp

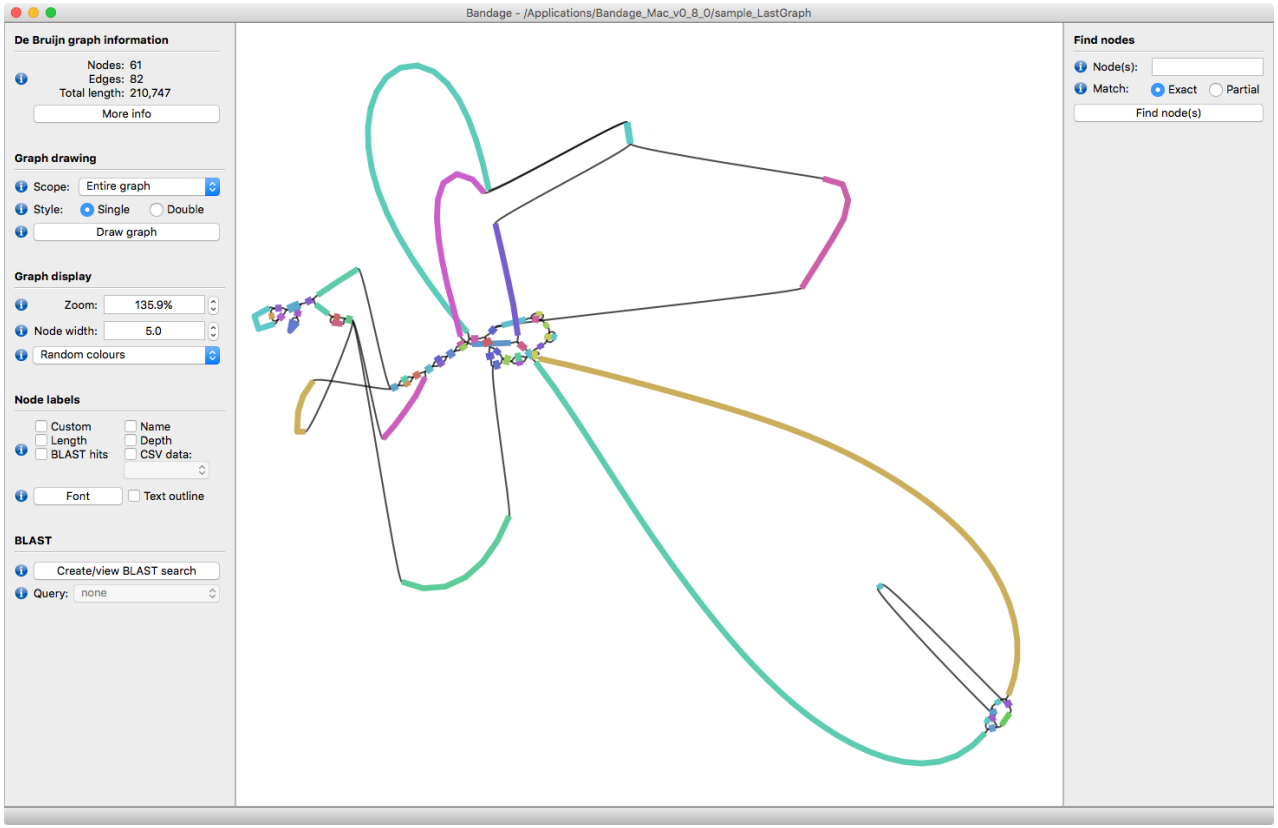
(300k+100k+45k+45k+30k = 520k >= 500kbp)

Massively Abused!!

Assessing an assembly

Visual Graph Inspection - Bandage

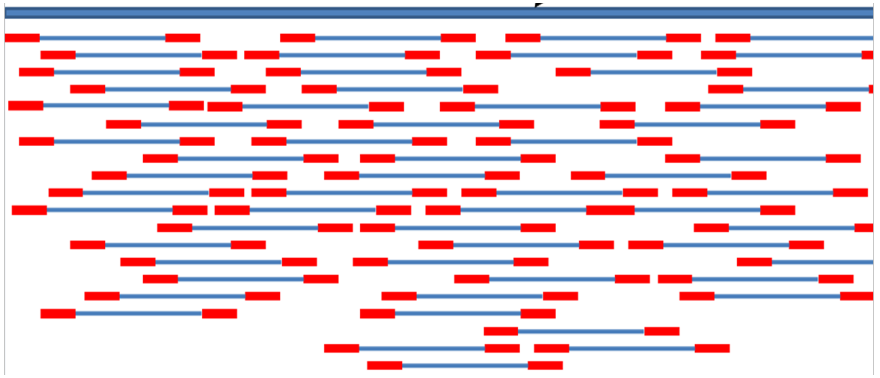
Allows easy visualisation of some graphs. Assess structure (and sanity) of the assembly



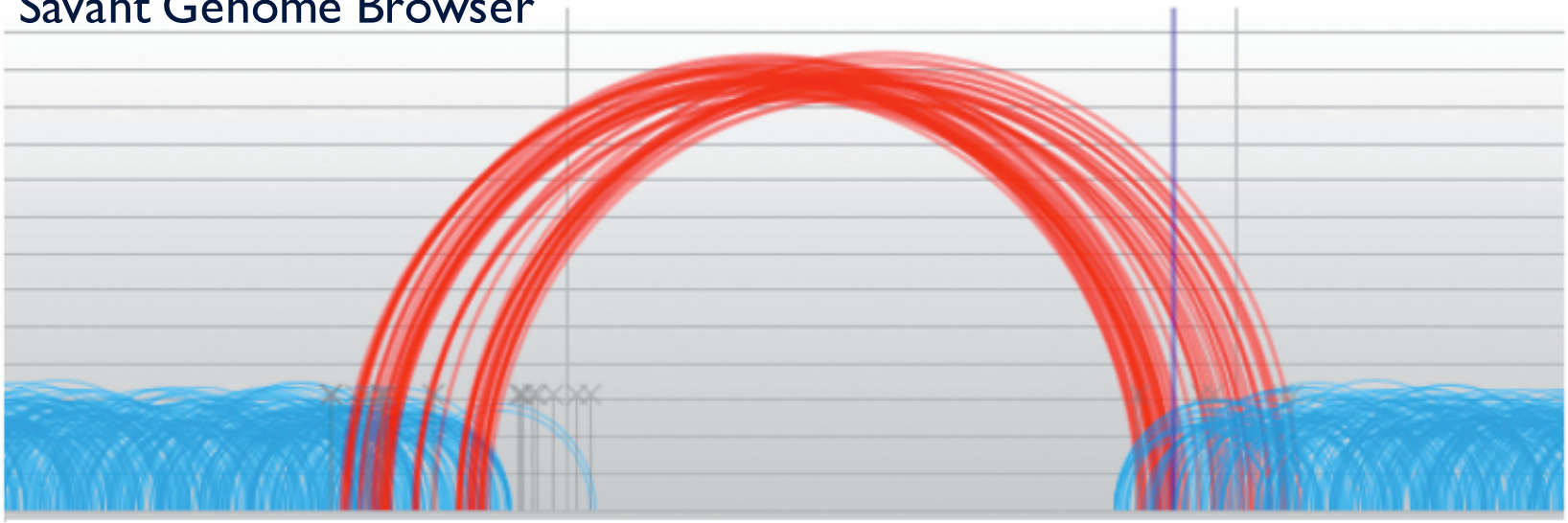
Assessing an assembly

Internal contig consistency

Align original reads to contigs.
Do they all align without breaks?
Is the distance between pairs consistent?
Is coverage as expected?



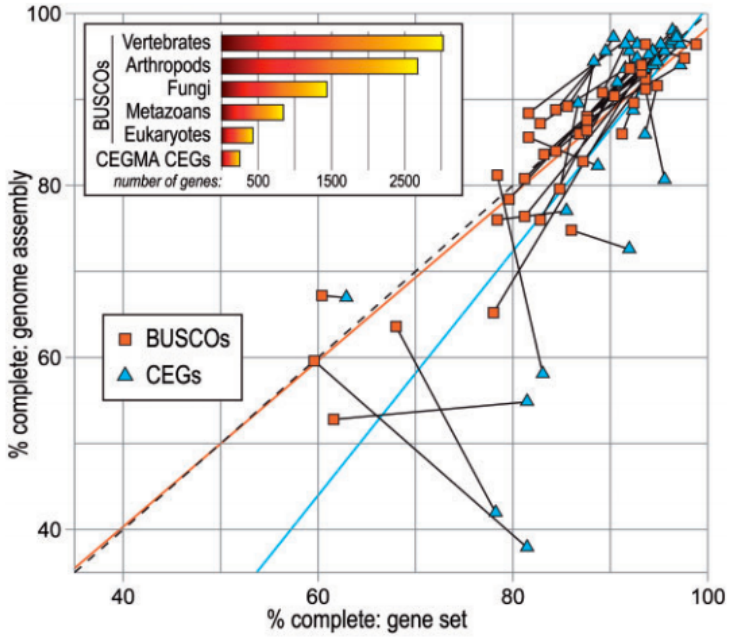
Savant Genome Browser



Assessing an assembly

Gene Content

Core Orthologues – are they all there?



BUSCO – identify and quantify completeness of core orthologues per clade in an assembly

RNA Seq

Best Practices, Best Tools

Structure

AIM:

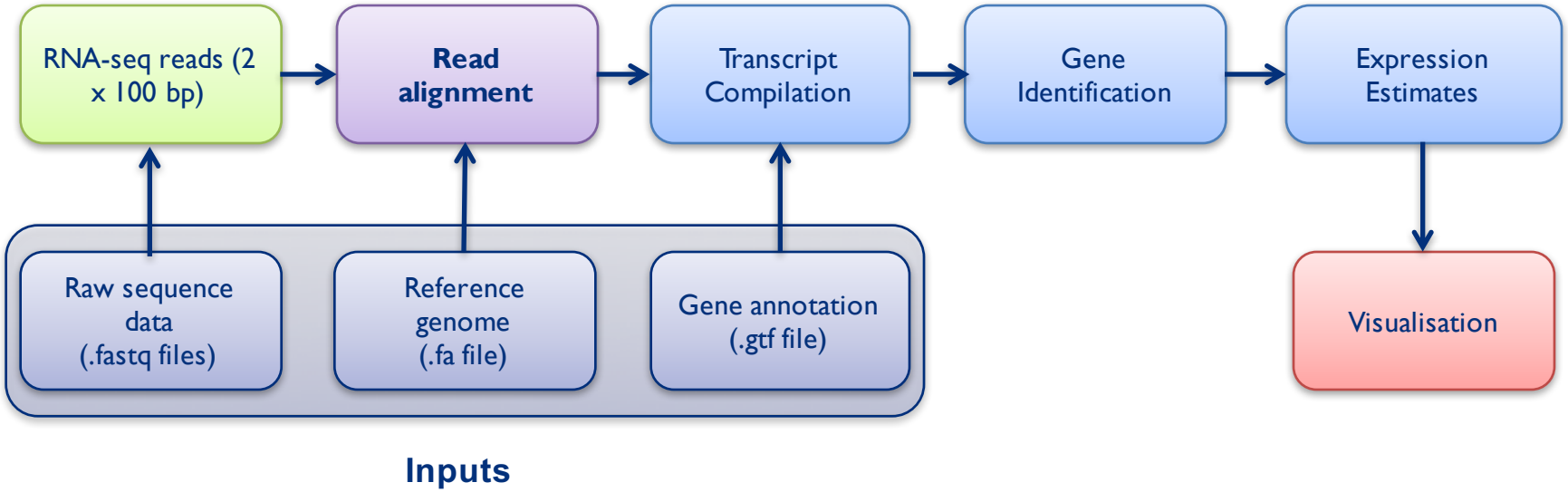
Help you develop a strategy for performing RNAseq and assembly and assessing analysis quality

- Theory
  - Sequence Types
  - Algorithm Basics/ Tool Types
  - Assessing an Alignment/Assembly
- Tools
  - For Alignment/Assembly
  - For Differential Expression

Why RNA Seq?

- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Allele specific expression
- Relating to SNPs or mutations
- Mutation discovery
- Fusion detection
- RNA editing

The RNAseq Pipeline



Alignment with Millions of Sequences

- Computational cost
  - 100's of millions of reads
  - Oldstyle (Smith-Waterman) – Newstyle (Burrows Wheeler Transform, Suffix Trees)
- Introns!
  - Spliced vs. unspliced alignments
- Can I just align my data once using one approach and be done with it?
  - Unfortunately probably not

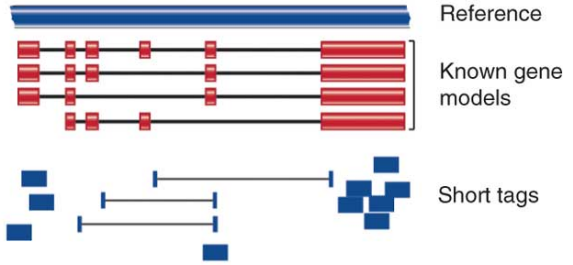
Alignment (mapping) strategies

De novo assembly

Assemble transcripts from overlapping tags

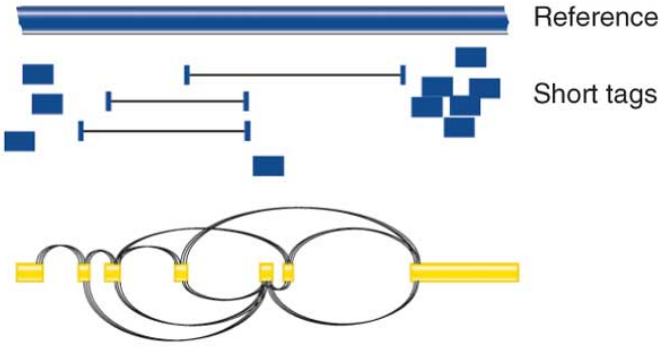
Optional: align to genome to get exon structure

Align to transcriptome



Use known and/or predicted gene models to examine individual features

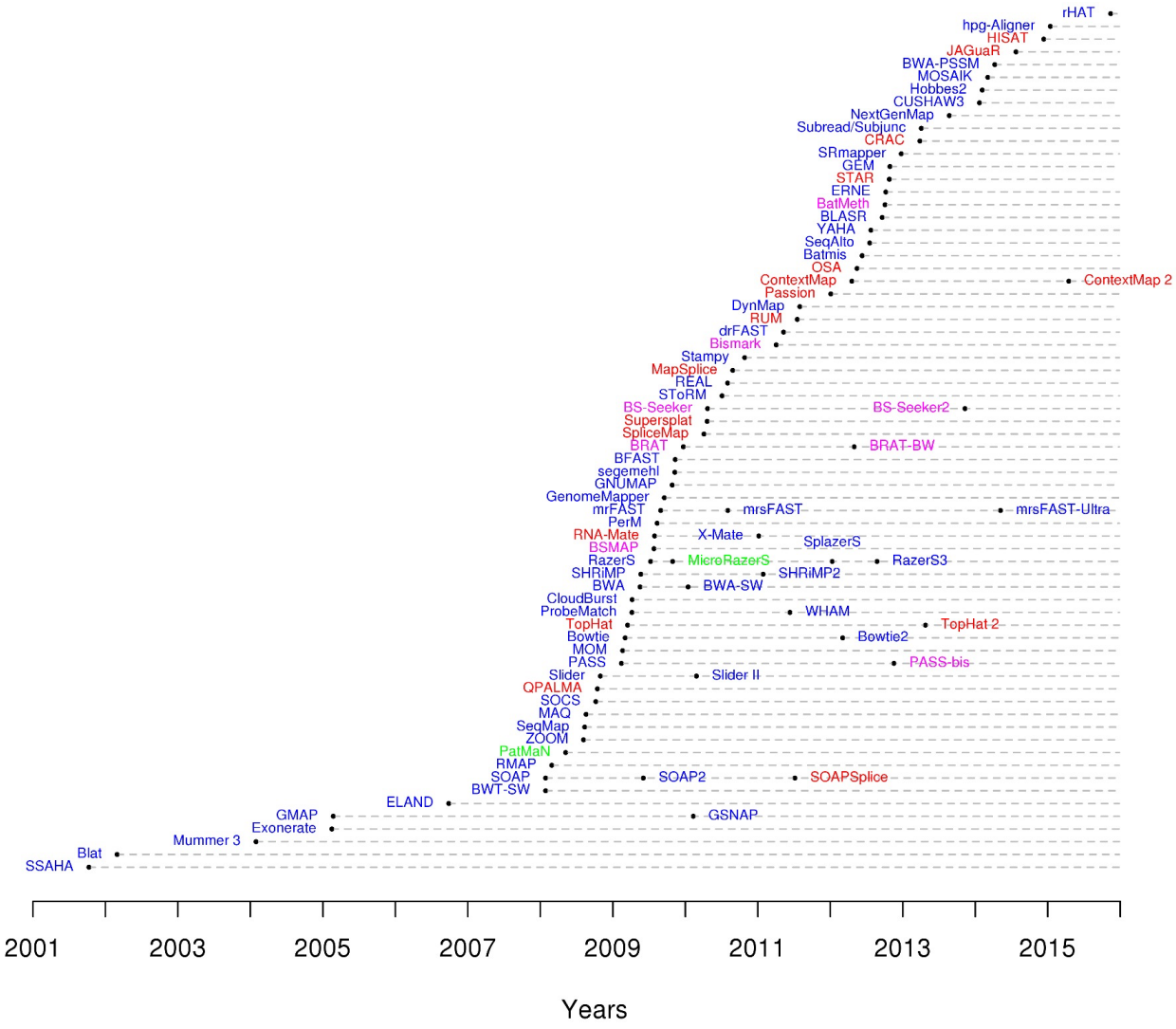
Align to reference genome



Infer possible transcripts and abundance

Diagrams from Cloonan & Grimmond, Nature Methods 2010

Alignment Tools

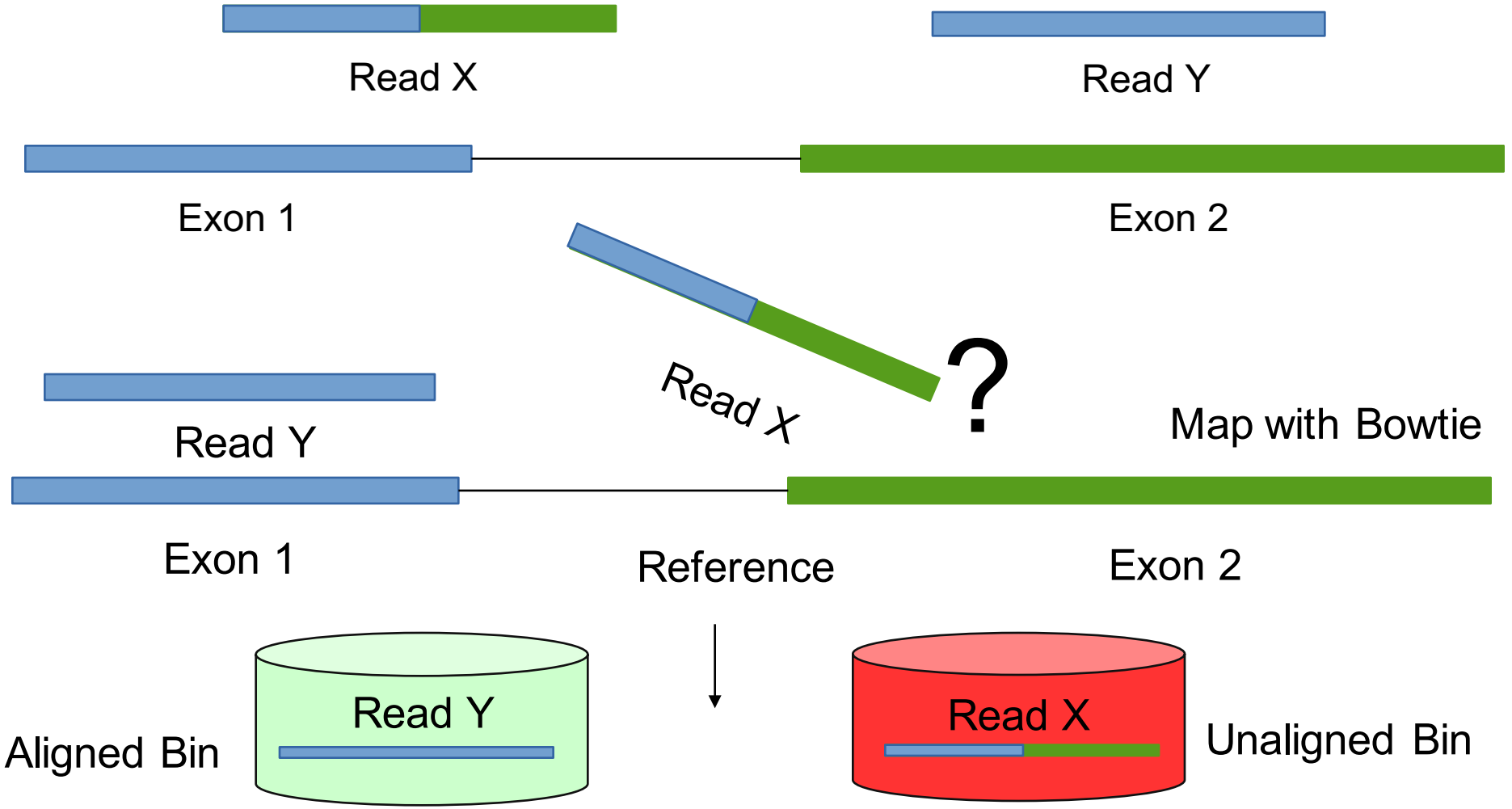


RNA
Bisulfite
DNA
microRNA

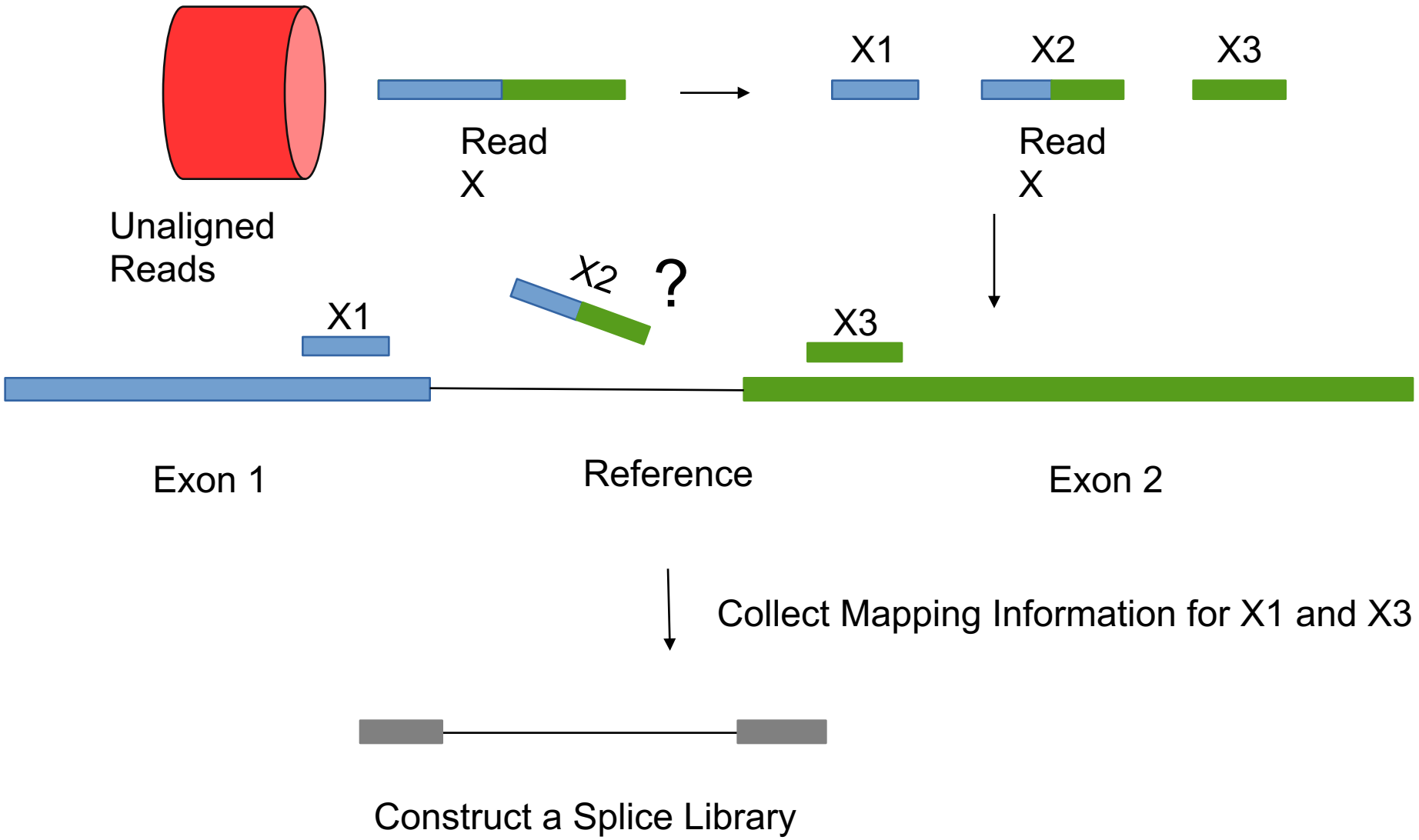
http://wwwdev.ebi.ac.uk/fg/hts_mappers/

Splice Aware Aligners – Tophat/Bowtie

RNA-seq reads may span introns



Splice Aware Aligners - Tophat



SAM/BAM

Example SAM/BAM header section (abbreviated)

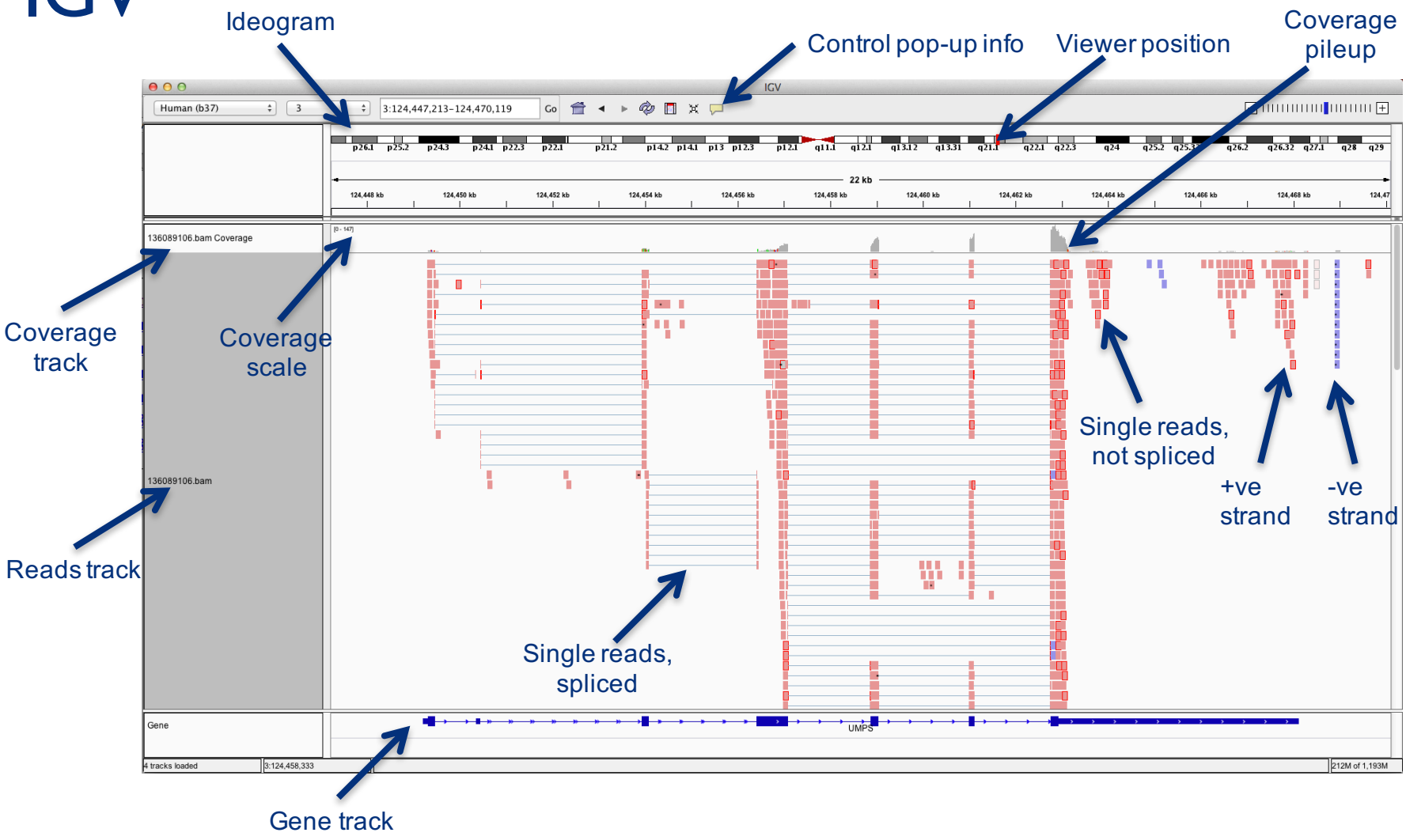
```
mgriffit@linus270 ~> samtools view -H /gscmnt/gc13001/info/model_data/2891632684/build136494552/alignments/136080019.bam  | grep -P "SN\:22|HD|RG|PG"
@HD     VN:1.4  SO:coordinate
@SQ     SN:22   LN:51304566     UR:ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa.gz AS:GRCh37-lite  M5:a718acaa6135fdca8357d5bfe9
4211dd  SP:Homo sapiens
@RG     ID:2888721359   PL:illumina     PU:D1BA4ACXX.3  LB:H_KA-452198-0817007-cDNA-3-lib1      PI:365  DS:paired end   DT:2012-10-03T19:00:00-0500     SM:H_KA-452198-0817007  CN:WUGSC
@PG     ID:2888721359   VN:2.0.8        CL:tophat --library-type fr-secondstrand --bowtie-version=2.1.0
@PG     ID:MarkDuplicates       PN:MarkDuplicates       PP:2888721359   VN:1.85(exported)       CL:net.sf.picard.sam.MarkDuplicates INPUT=[/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blad
e10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300.bam] OUTPUT=/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jw
alker-15434-136080019/scratch-ILg6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300-post_dup.bam METRICS_FILE=/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-1543
4-136080019/staging-1iuJS/H_KA-452198-0817007-cDNA-3-lib1-2888360300.metrics REMOVE_DUPLICATES=false ASSUME_SORTED=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=9500 TMP_DIR=[/gscmnt/gc13001/info/build_merged_al
ignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y] VALIDATION_STRINGENCY=SILENT MAX_RECORDS_IN_RAM=500000    PROGRAM_RECORD_ID=MarkDuplicates PROGRAM_GROUP_NAME=Mark
Duplicates MAX_SEQUENCES_FOR_DISK_READ_ENDS_MAP=50000 SORTING_COLLECTION_SIZE_RATIO=0.25 READ_NAME_REGEX=[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).* OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 VERBOSITY=INFO
QUIET=false COMPRESSION_LEVEL=5 CREATE_INDEX=false CREATE_MD5_FILE=false
mgriffit@linus270 ~>
```

Example SAM/BAM alignment section (only 10 alignments shown)

```
mgriffit@linus270 ~> samtools view -f 3 -F 1804 /gscmnt/gc13001/info/model_data/2891632684/build136494552/alignments/136080019.bam | head
HWI-ST495_129147882:3:2114:15769:38646  99      1       11306   3       100M    =       11508   302     ACTGCGGGGCCCTCTTGCTTACTGTATAGTGGTGGCACGCCGCCTGCTGGCAGCTAGGGACATTGCAGGGTCCTCTTGCTCAAGGTGTAGTGGCAGCACGC
CCFFFFFHHHGHJJJJJJJIJJJHGIJJIJJHIIJJJJJJHFDDDDDDDDDDDDCDDDDDDDDDDDDD<CDDDDDDDDCCCC>@>CDDDDDD?1<B        CC:Z:15 MD:Z:5A94       PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102519765  AS:i:-5 XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:2114:15769:38646  147     1       11508   3       100M    =       11306   -302    ACTCCTAAATATGGGATTCCTGGGTTTAAAAGTATAAAATAAATATGTTTAATTTGTGAACTGATTACCATCAGAATTGTACTGTTCTGTATCCCACCAG5
;5:CDCDCDECEFCD@9E=?7EEIIIIHCEGGIJJJJIIJJIHF@?00IHHFFGG?*JJJIJGHGEIJJIJJJJJJIHHCIEJJJHFHHGHFFEDFCCB        CC:Z:15 MD:Z:34A65      PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102519563  AS:i:-6 XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1210:1257:16203   163     1       11810   3       100M    =       12055   345     CCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTTCTTTGACCTCTTCTTTCTGC
CCFFFFFHFHAFGGIIIJJJEEHGIGGGIJIJJGI?@EHIGIJDGHIHGGIJJJJJJJJIJGHHHGHFFFCDDDDDDCDCCCCCA;>@>@AA@:AA>AA        CC:Z:15 MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:102519261  AS:i:0  XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:1210:1257:16203   83      1       12055   3       100M    =       11810   -345    GAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAGTGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATAC
CC>4C>DCCCACACDCC?BDCEE@ECFFFFHHHHHIJJJIIJJIIIHHEHIIGJIJIJJIGHIIIJJJJJIIJJJJJIIJJIJIJJHGHHHDFEFFCCC        CC:Z:15 MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:102519016  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:2111:3117:78828   163     1       12634   3       100M    =       12746   212     GCCCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTGCACAGGTGAGAGGAGAG<
@@FFFFFDHHHH9FHGIIFGAFDHEGII>GHIIIIIIIIIIIIIIIFHDDFFEEECEECCCACCCCCC:AADCCBCC>CAC<CCCCCC:@CB@@BAB##        CC:Z:15 MD:Z:85G14      PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102518437  AS:i:-5 XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:2111:3117:78828   83      1       12746   3       100M    =       12634   -212    GGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCATCTCCTGTCTCCTGGAGAGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTD
DCABDBDDDDDDDDDDDCDDDDBDB@BDDDB@;CCCCCDEFD@;.?<HIGGEIGEHIGJJJIIGIGIIHEGFEHFJIIIIIGJJJJHHHHHFFFFFC@@        CC:Z:15 MD:Z:37G62      PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102518325  AS:i:-5 XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:1102:4242:26638   99      1       13503   3       100M    =       13779   376     CGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAC
CCFFFFFHHHHHJJJIJJJJJJJJJJJJJJGIIIIJJFHGGIJGIJJJEGIJIJJHHIHHGHFFEFDEEEECCCAACDDACDCDDDDDB?8?<B>A@CDC        CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357414  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1309:15328:74082  99      1       13534   3       100M    =       13780   346     AGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCACAGG@
CCFFFADHHHHFIJJJJJIJIJIHIJJJIHJJIJJIIIJJIJIJJJJJBFHIIJJJJJJIJHH=EEFFFFCEEECEDCDCDDDDDDDDDDDDDDDBDCCD        CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357383  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1308:10126:19636  99      1       13779   3       100M    =       14027   348     CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTGC
CCFFFFFHHGHHJJJJJJJJJJIJJJJJJIJJJJJJJJJIJIJJJJJJJJJHJJJHGHHFFFFFFEEEEEEDDDDCDCDCDFADCACCACDDCCCCCCD        CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357140  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1102:4242:26638   147     1       13779   3       100M    =       13503   -376    CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTG#
##DCCDDDCCBBBABCCDDDCBDDBBDHC?=GIIJIIIIJIGIIIIIJHJJIJJIGCIIJJJJJIGHGJJIJIJJJJJIJIIIGGFGHHHHFFFFFCCC        CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357140  AS:i:0  XS:A:+  YT:Z:UU
mgriffit@linus270 ~>
```

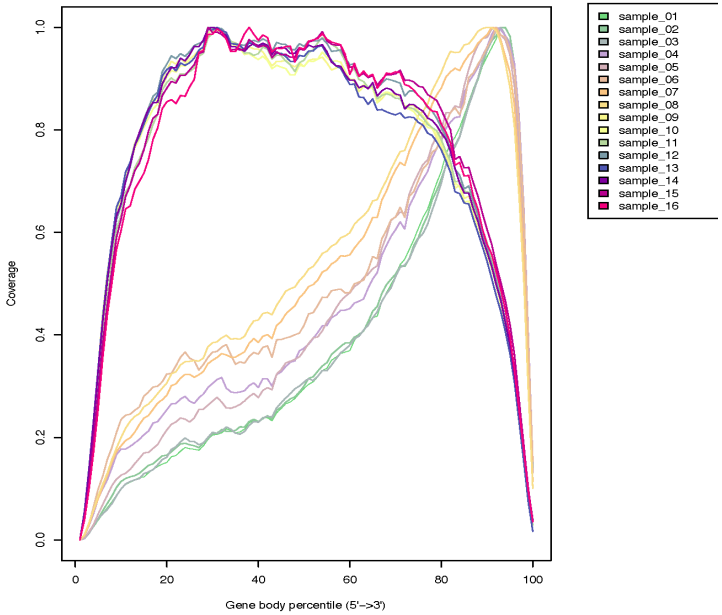
Visualizing Alignments

IGV

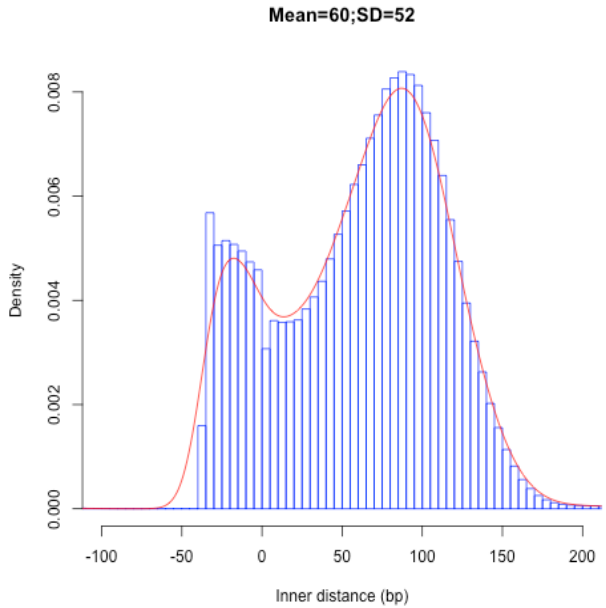


Alignment QC

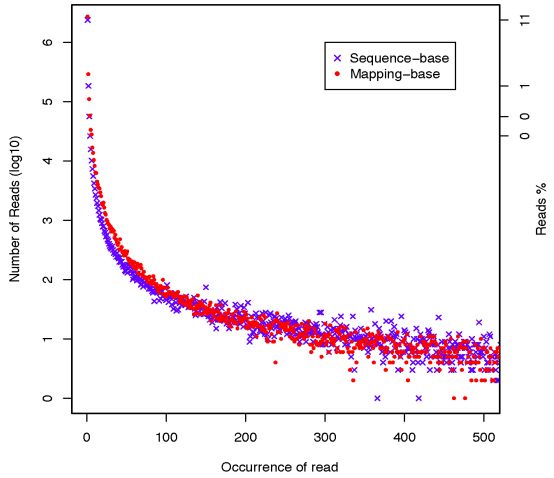
Transcript coverage 3' Bias



Insert Size

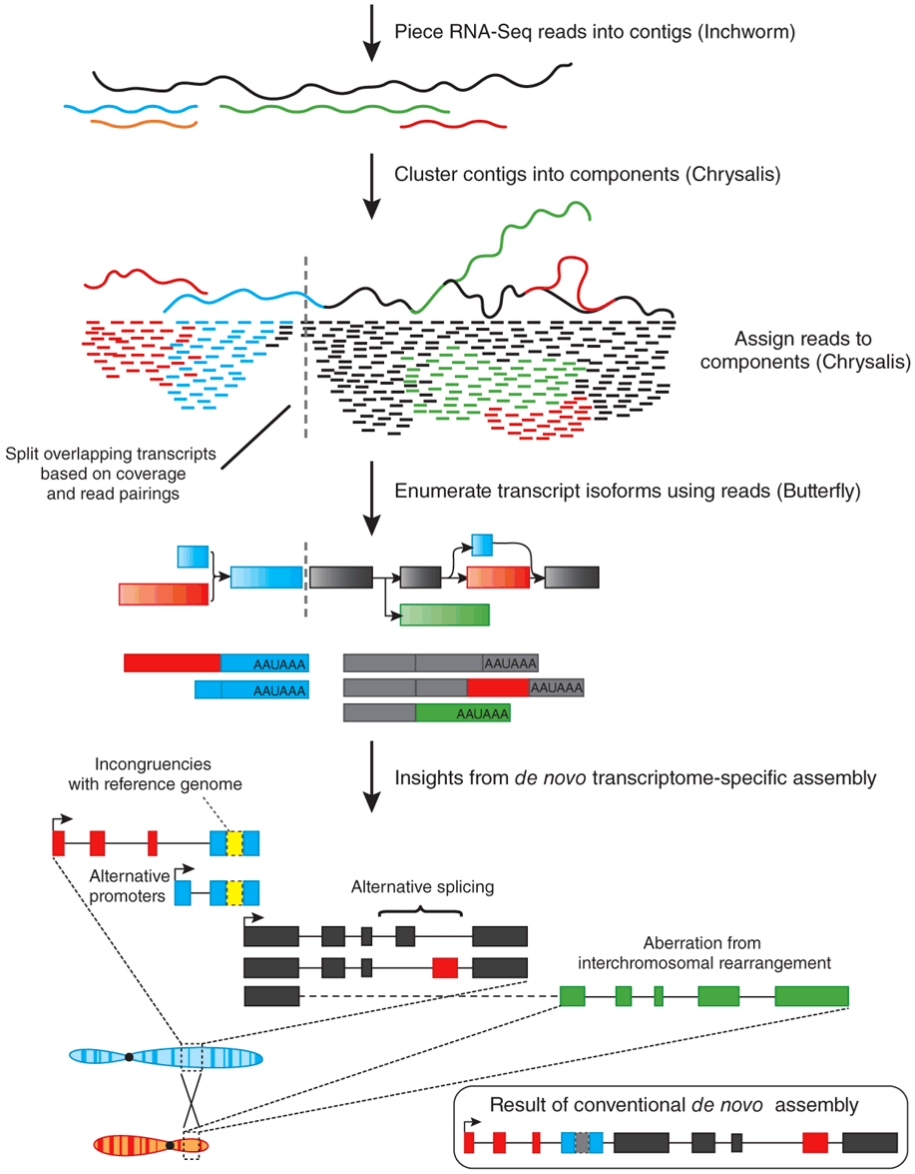


PCR Duplicates



http://rseqc.sourceforge.net

Complete de novo – no reference

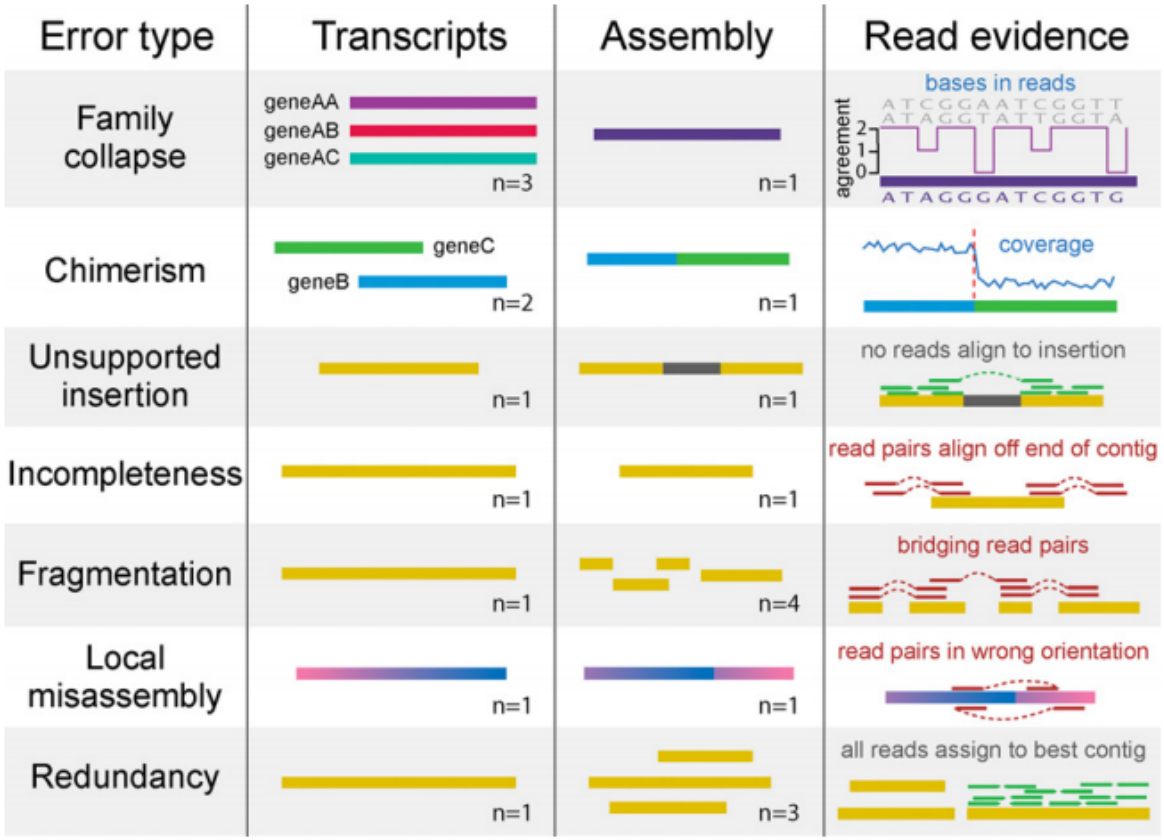


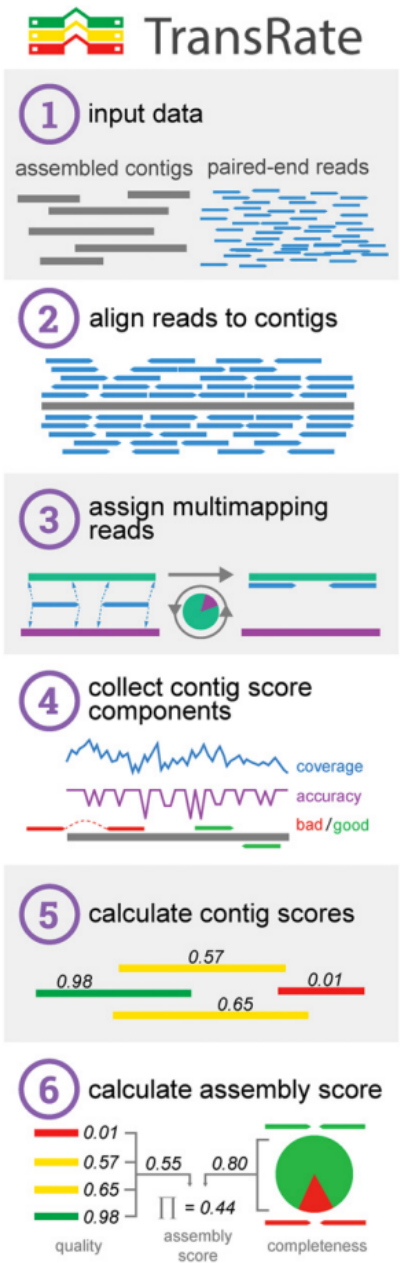
Trinity

https://github.com/trinityrnaseq

Assembly QC

TransRate

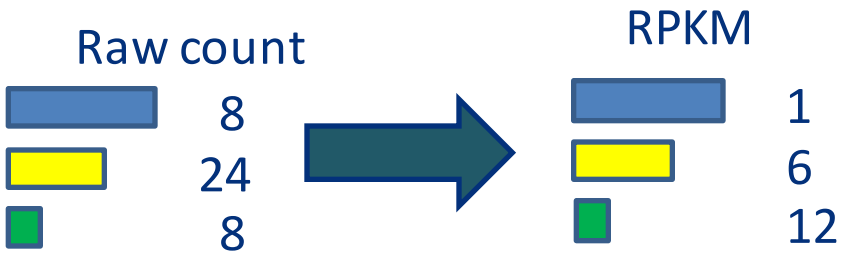




Alignments Mappings to expression

Normalize counts for expression levels –

to account for differences in gene length, read counts are converted to ***Reads per kilobase of transcript mapped (RPKM)***



$$RPKM_{geneA} = 10^9 C_{geneA} / NL$$

C_{geneA} = number of reads mapped to geneA
N = total number of reads
L = length of transcript in units of Kb

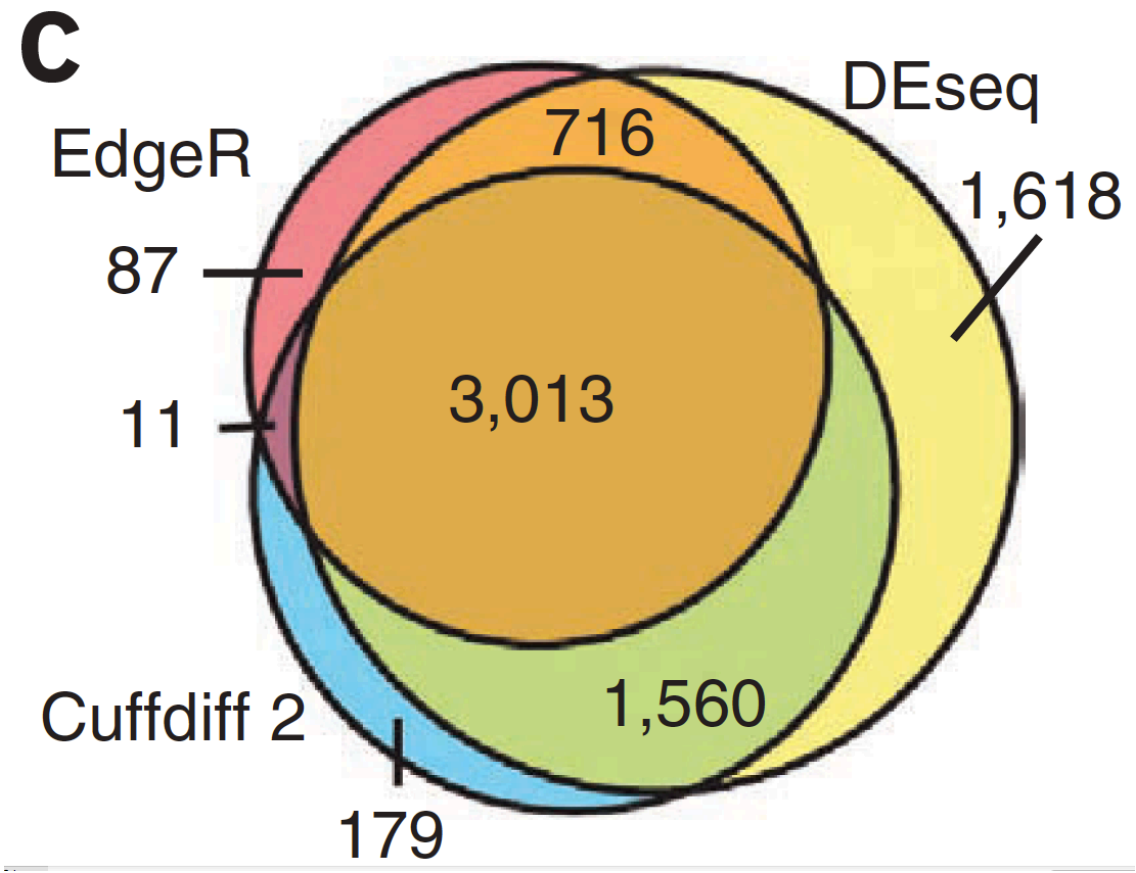
Cufflinks/CuffDiff
Trinity

Alternative differential expression methods

Raw count approaches

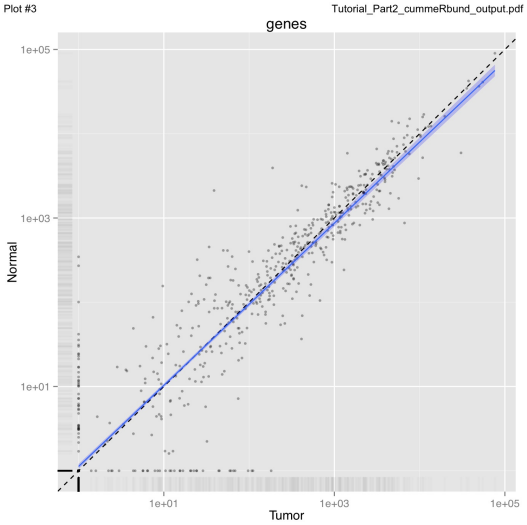
DESeq - http://www-huber.embl.de/users/anders/DESeq/

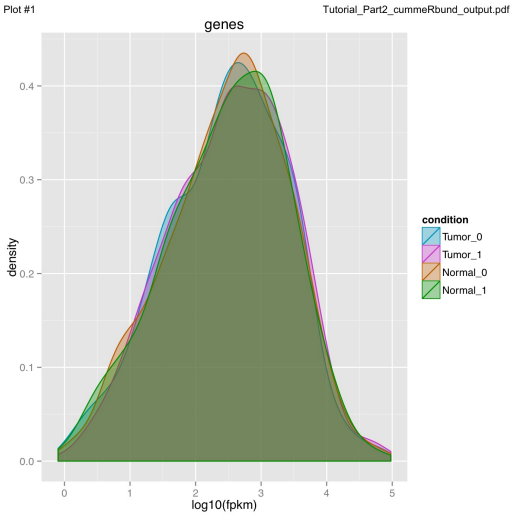
edgeR - http://www.bioconductor.org/packages/release/bioc/html/edgeR.html

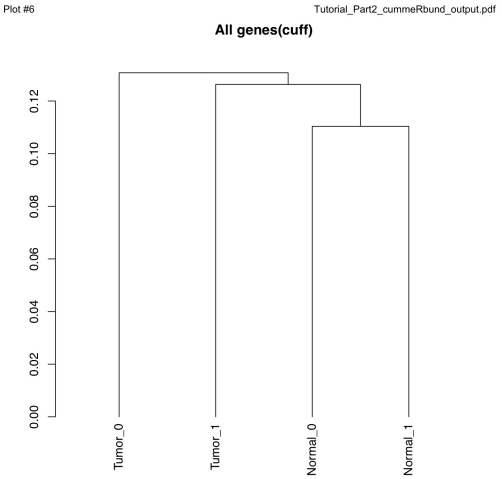


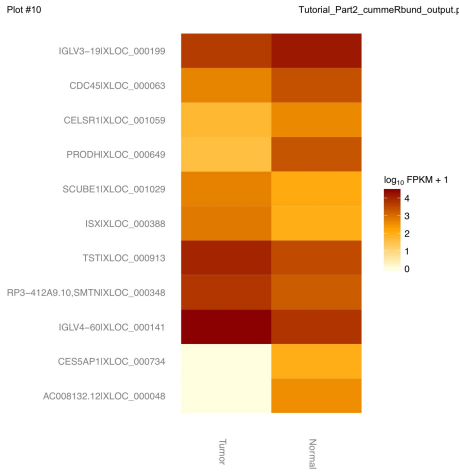
cummeRbund

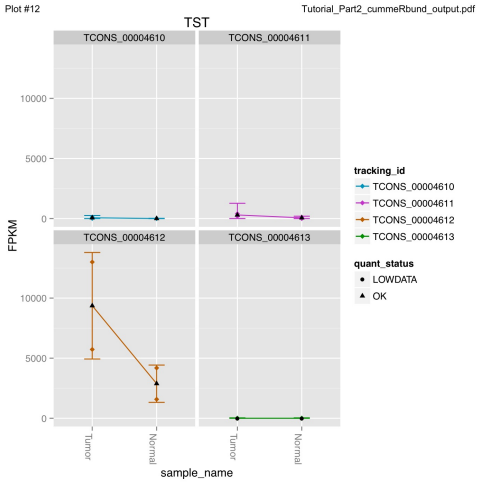
Analysis of expression in R

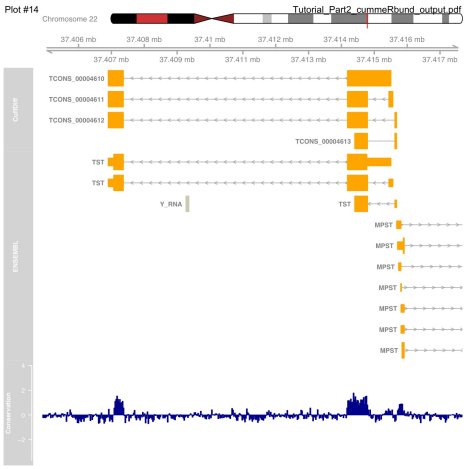










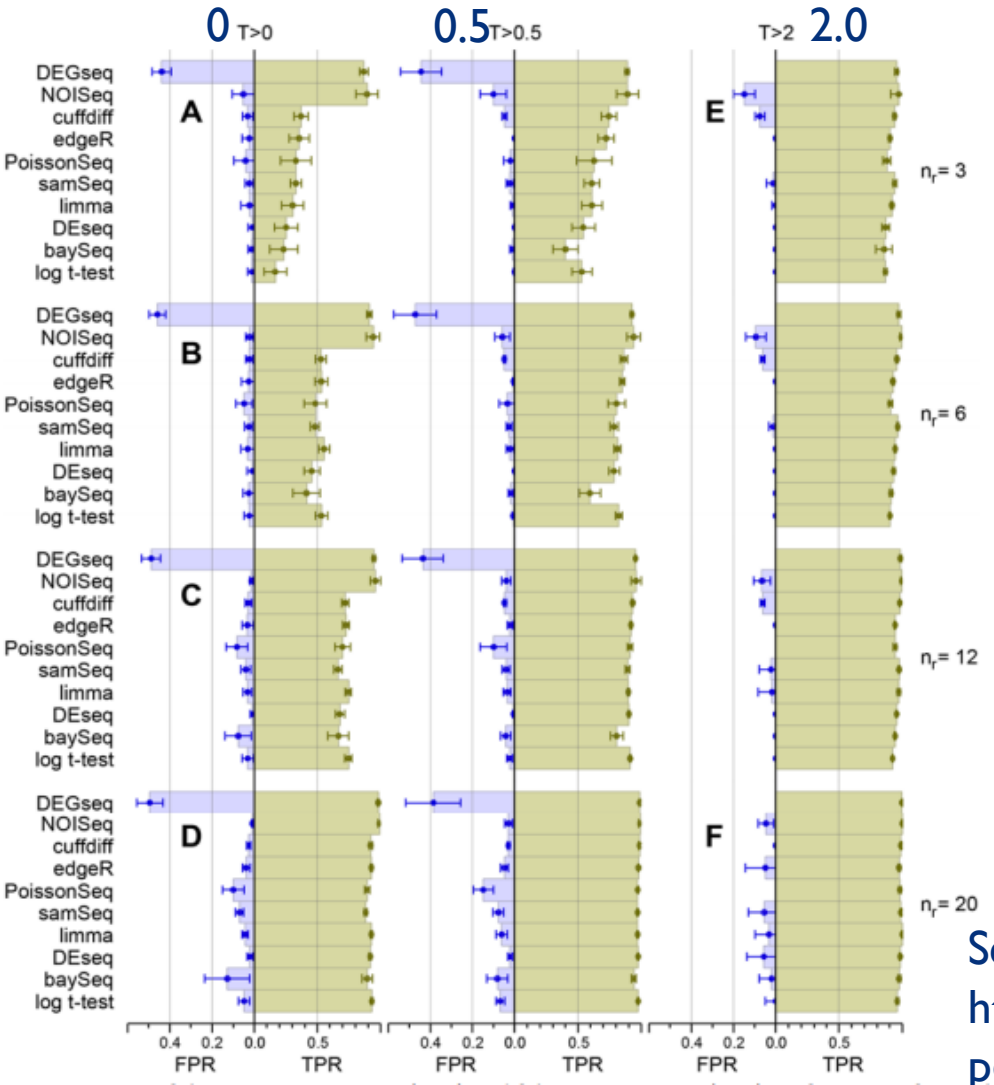


Which tool? Which experimental design?

False/True Positive Rate at different fold changes and replicate numbers

Fold Change Cutoff

Tool



Samples

Schurch et al http://arxiv.org/pdf/1505.02017v2.pdf

Which tool? Which experimental design?

- At least six replicates per condition for all experiments.
- At least 12 replicates per condition for experiments where identifying the majority of all DE genes is important.
- For experiments with <12 replicates per condition; use edgeR(exact).
- For experiments with >12 replicates per condition; useDESeq.
- Apply a fold-change threshold appropriate to the number of replicates per condition between $0.1 \leq T \leq 0.5$.

Schurch et al
http://arxiv.org/pdf/1505.02017v2.pdf

Metagenomics

Best Practices, Best Tools

Structure

AIM:

Help you understand strategies used in metagenomics experiment

Experimental Aims
Pipelines and Tools

Metagenome Taxonomy
Metagenome Function
Metatranscriptomes

What is metagenomics?

Environmental genomics

Samples of mixed genomes, from some place

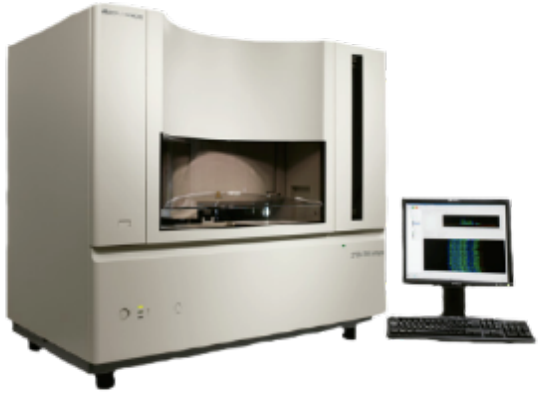
Microbiomes
- e.g genomes of microbes in soil samples

Plant Pathogen Infection
- e.g field pathogenomics

Not so much marker gene surveys
- eg. 16S

Effective way to profile structure and function of communities

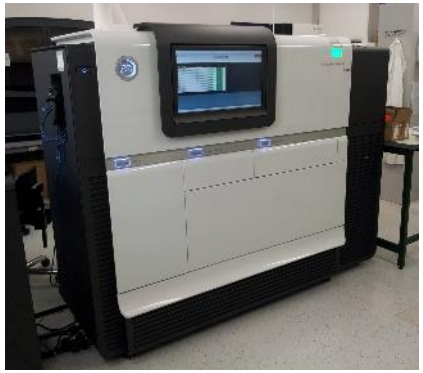
Metagenomics Sequencing



Sanger

Roche 454

Illumina *Seq



Pacific Biosciences

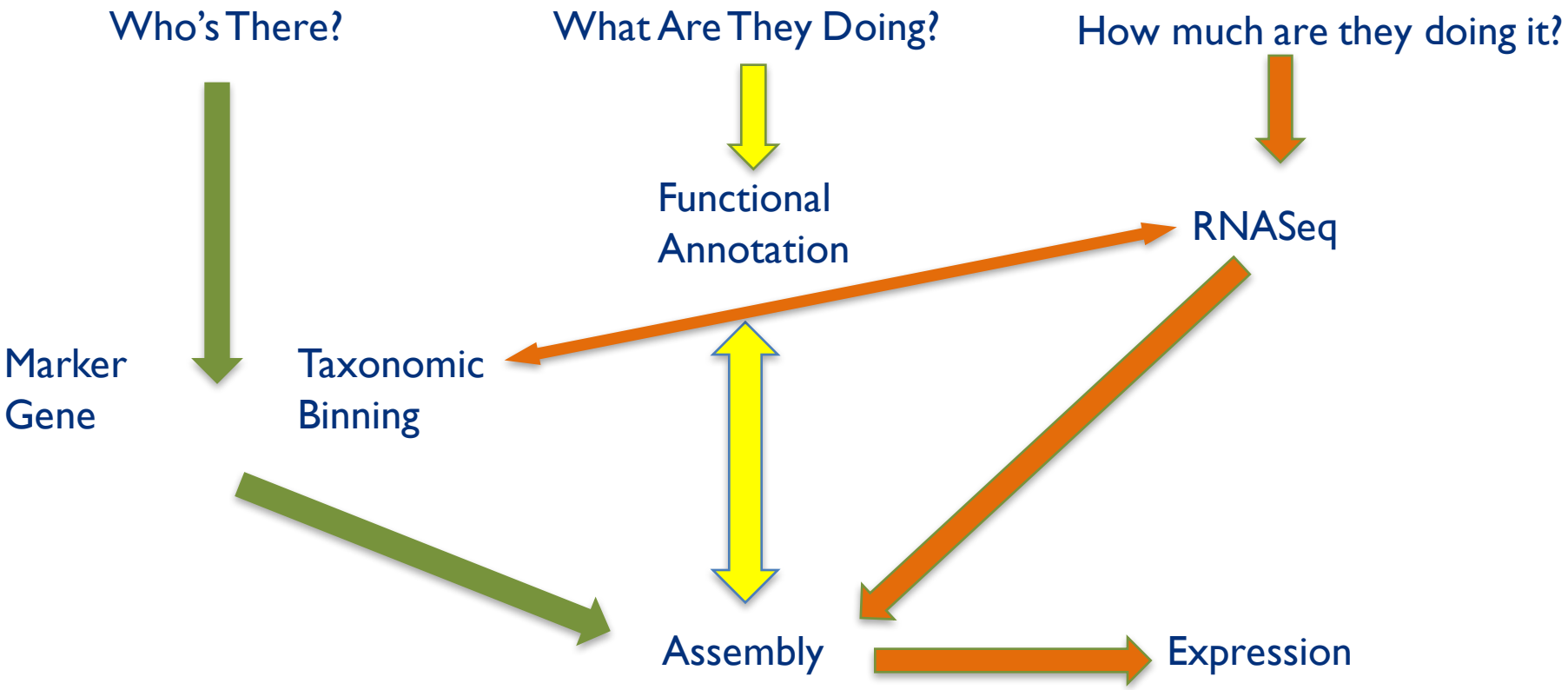
Ion Torrent

Nanopore

Data Quality Issues

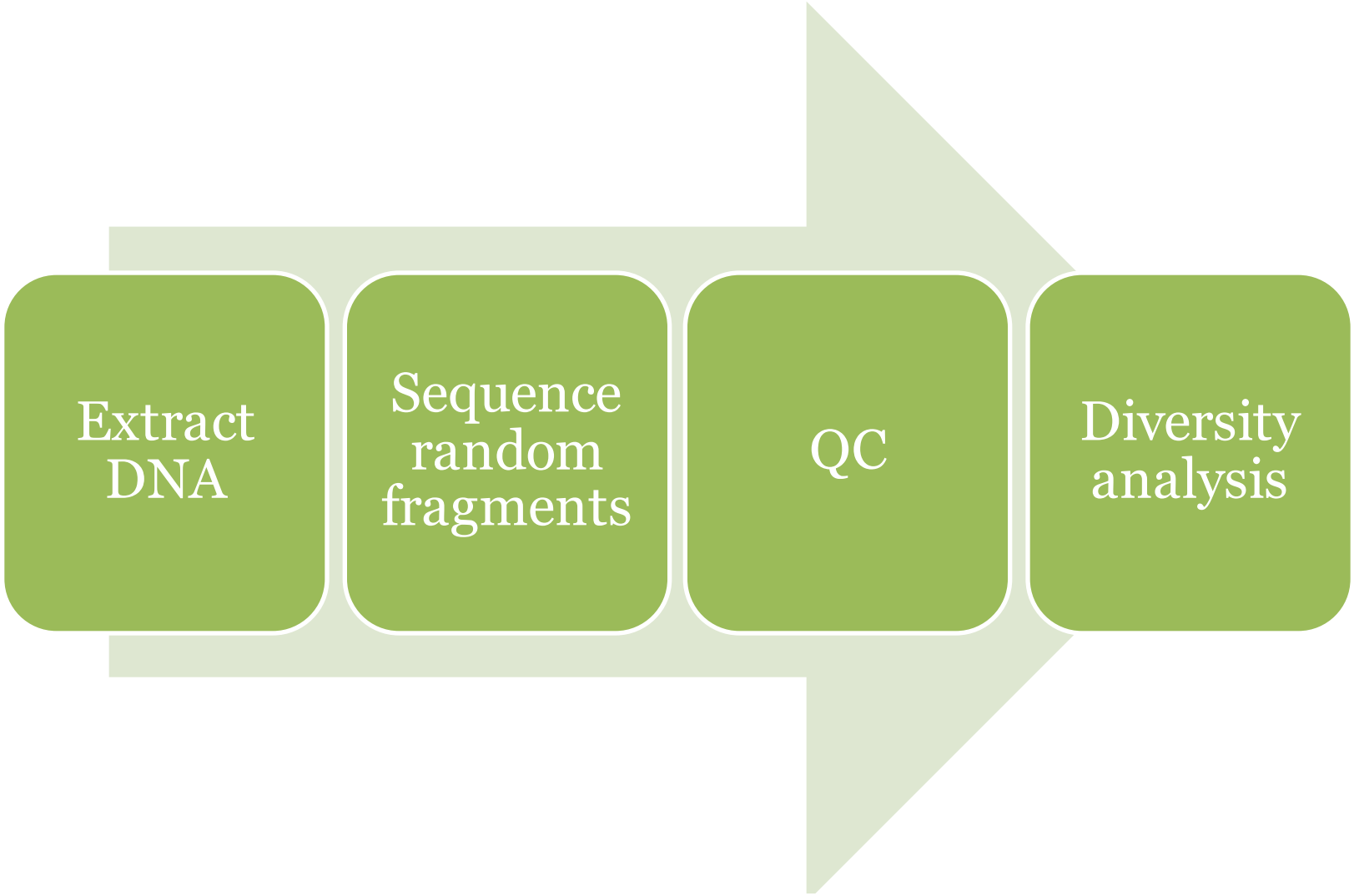
- Sequencing errors
  - Error rates, error *type* (PacBio: 10% random, Illumina – 0.1% substitution)
- Chimeras
  - Amplification artifacts, cloning of restriction fragments

Metagenomics approaches - Summary



Metagenome Taxonomy

Who's there?



Metagenome Taxonomy
Who's there?

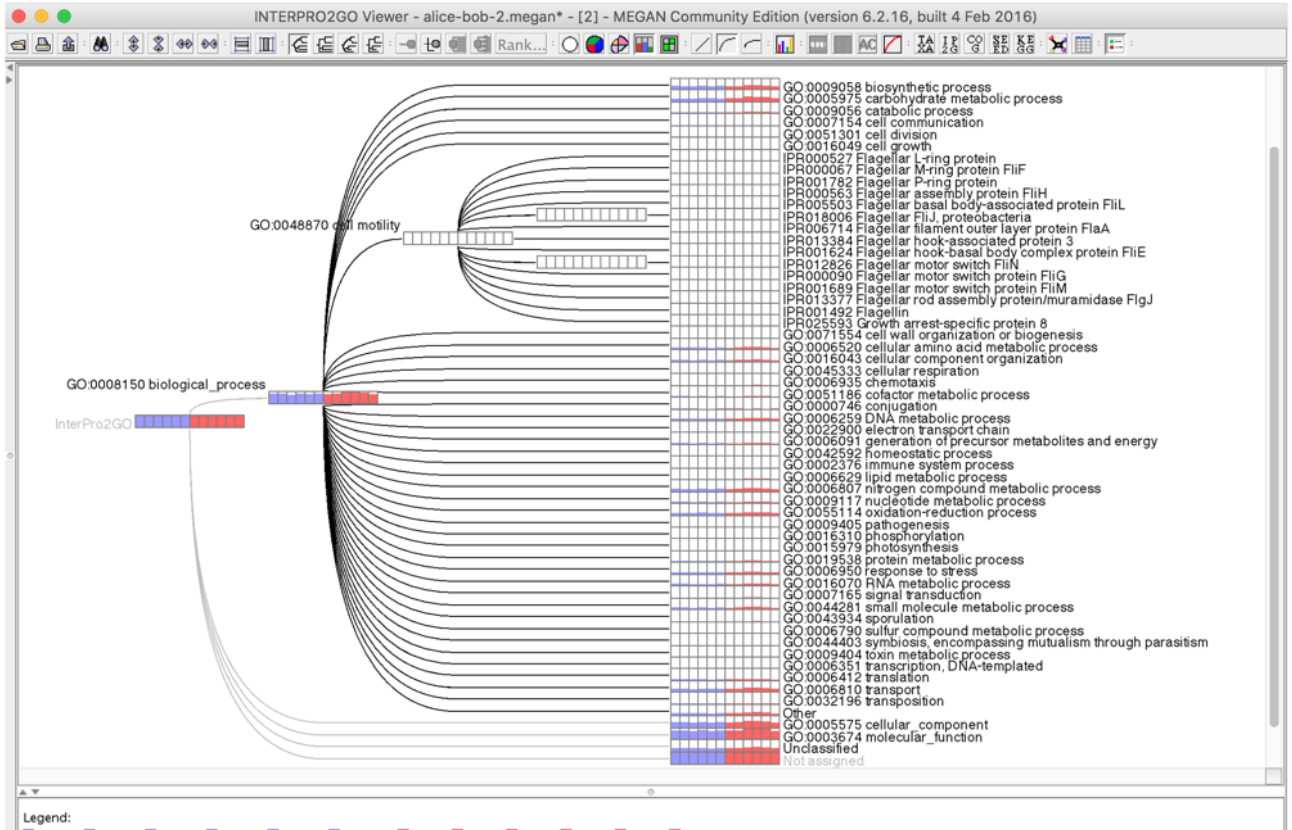
- Goal: Identify the relative abundance of different microbes in a sample
- Problems:
  - Reads are all mixed together
  - Reads can be short (~100bp)
- Two broad approaches
  1. Binning Based
  2. Marker Based

Binning Methods

- Attempts to "bin" reads into the genome from which they originated
- Composition-based
  - Uses GC composition or k-mers (e.g. Naïve Bayes Classifier)
  - Generally not very precise and not recommended
- Sequence-based
  - Compare reads to large reference database using BLAST (or some other similarity search method)
  - Reads are assigned based on "Best-hit" or "Lowest Common Ancestor" approach

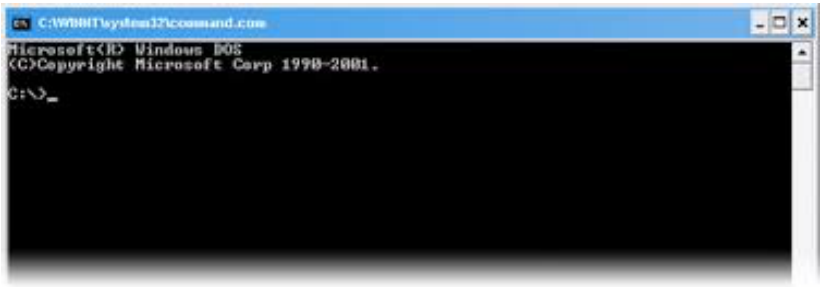
Binning Tools

- Notable Examples:
  - MEGAN: http://ab.inf.uni-tuebingen.de/software/megan6/
    - One of the first metagenomic tools
    - Does functional profiling



Binning Tools

- Notable Examples:
  - Kraken: https://ccb.jhu.edu/software/kraken/
    - Fastest binning approach to date and very accurate.
    - Large computing requirements (e.g. >128GB RAM)



Marker Methods

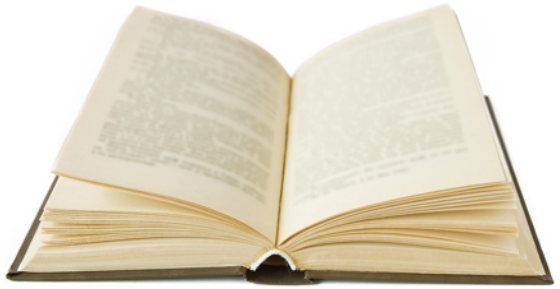
- Single Gene
  - Identify and extract reads hitting a single marker gene (e.g. 16S, cpn60, or other "universal" genes)
  - Use existing bioinformatics pipeline (e.g. QIIME.)
- Multiple Gene
  - Several universal genes
    - PhyloSift Uses 37 universal single-copy genes

‘left-over’ Reads

- What about the reads that can’t be classified?
  - Errors?
  - Novel Sequence?
  - Just not in database?

  - Try assembling the left-overs

Metagenomic assembly



Metagenomic assembly

Specific assemblers

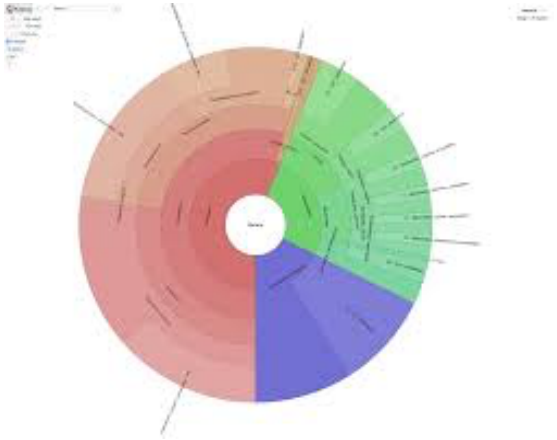
- IDBA-UD,
- Meta-IDBA,
- Meta-Velvet,
- Ray Meta,
- Meta-Cortex

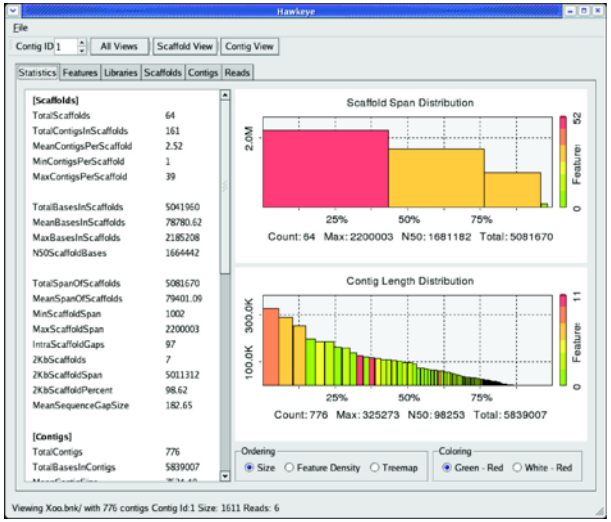
Validation

- LAP,
- ALE,
- QUAST,
- FRCbam,
- REAPR

Analysis Pipelines

- MetAMOS





What do we mean by function?

- General categories
  - Photosynthesis
  - Nitrogen metabolism
  - Glycolysis
- Specific groups of orthologs
  - EC: 1.1.1.1 (alchohol dehydrogenase)
  - K00929 (butyrate kinase)

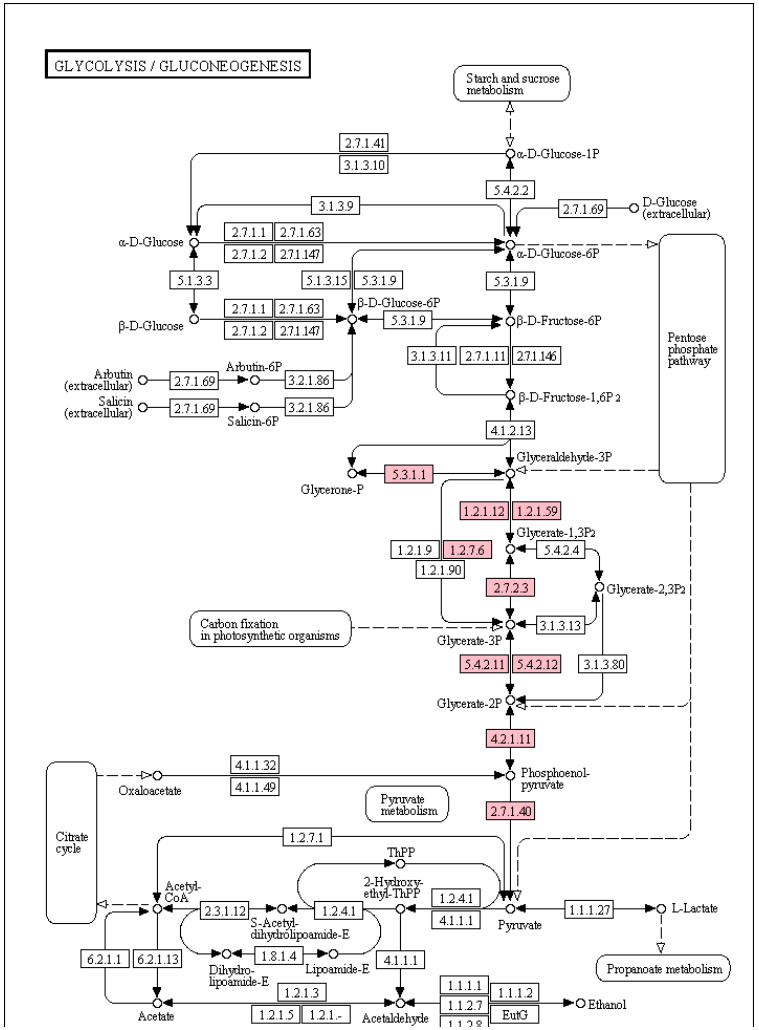
KEGG for functional annotation

- KEGG Orthologs (KOs) http://www.genome.jp/kegg/ko.html
  - Most specific. Thought to be homologs and doing the same exact "function"
  - ~12,000 KOs in the database
  - These can be linked into KEGG Modules and KEGG Pathways,
  - Identifiers: K01803, K00231, etc.

Entry	K01803 KO
Name	TPI, tpiA
Definition	triosephosphate isomerase (TIM) [EC:5.3.1.1]
Pathway	ko00010 Glycolysis / Gluconeogenesis ko00051 Fructose and mannose metabolism ko00562 Inositol phosphate metabolism ko00710 Carbon fixation in photosynthetic organisms ko01200 Carbon metabolism ko01230 Biosynthesis of amino acids
Module	M00001 Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate M00002 Glycolysis, core module involving three-carbon compounds M00003 Gluconeogenesis, oxaloacetate => fructose-6P
Disease	H00664 Anemia due to disorders of glycolytic enzymes
Brite	KEGG Orthology (KO) [BR:ko00001] Metabolism Overview 01200 Carbon metabolism K01803 TPI, tpiA; triosephosphate isomerase (TIM) 01230 Biosynthesis of amino acids K01803 TPI, tpiA; triosephosphate isomerase (TIM) Carbohydrate metabolism 00010 Glycolysis / Gluconeogenesis

KEGG for functional annotation

- KEGG Pathways
  - Groups KOs into large pathways (~230)
  - Each pathway has a graphical map
  - Individual KOs or Modules can be highlighted within these maps
  - Pathways can be collapsed into very general functional terms (e.g. Amino Acid Metabolism, Carbohydrate Metabolism, etc.)



Metagenomic Annotation tools

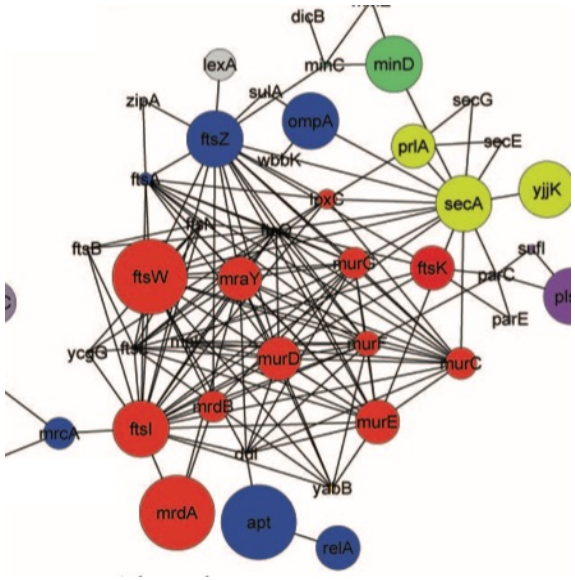
- Web-based
  - EBI Metagenomics Server
  - MG-RAST
  - IMG/M
  - Koala
- GUI-Based
  - MEGAN
    - Allows connection between taxonomy and function
  - ClovR
    - Virtual Machine based - microbial
- Local-based
  - Prokka
  - MetAMOS
    - Built in assembly, highly customizable,
  - DIY
    - Set up your own in-house custom computational pipeline

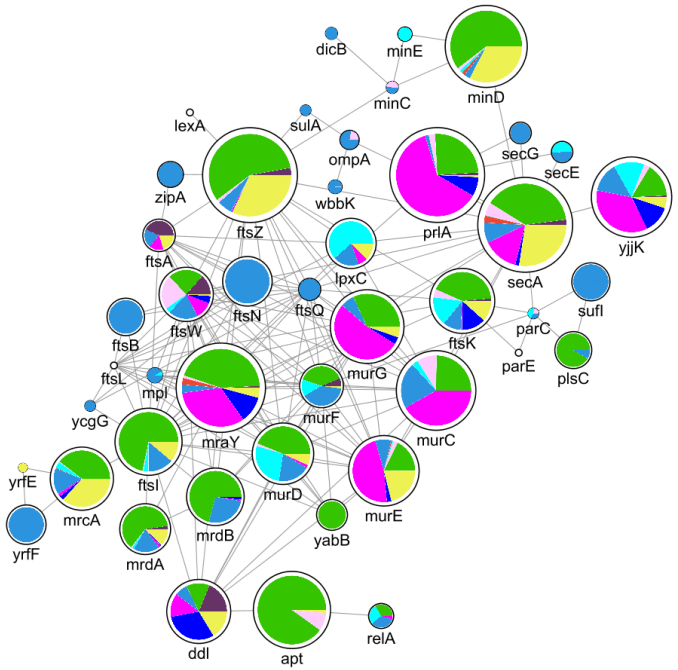
Metatranscriptomics for community activity

Functional Quantification

Use RNA-Seq to determine which genes and pathways are being actively expressed within a community

Genes involved in pathways associated with cell wall biogenesis

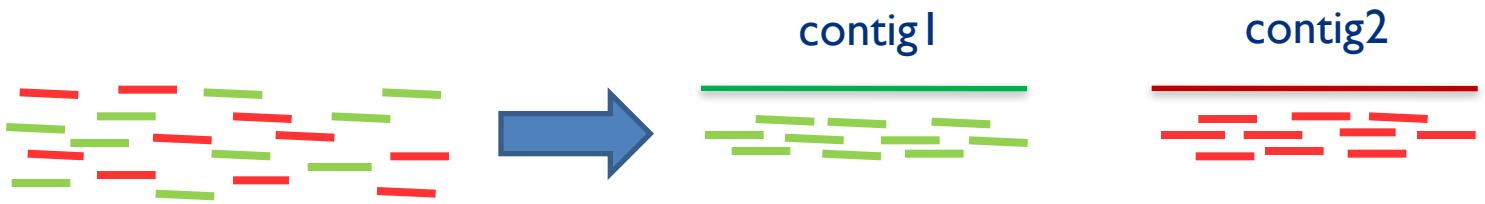


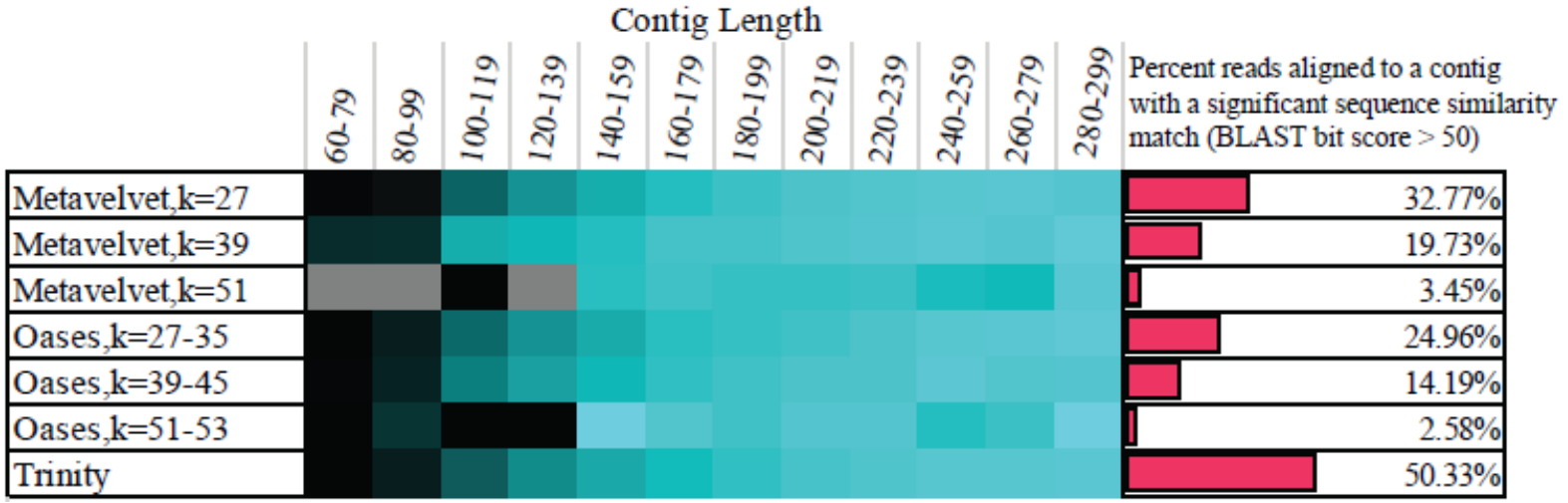


Relative abundance of transcripts

Relative abundance and contributing taxa

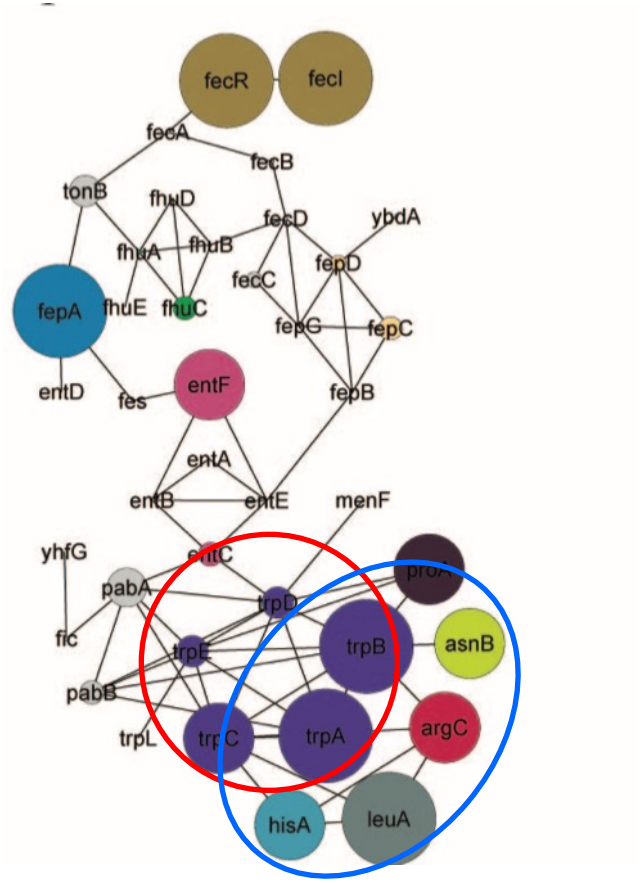
Metatranscriptomics: Assembling to annotate





Metatranscriptomics: Understanding results

How do we identify those systems which are differentially regulated across samples (e.g. systems up-regulated in disease vrs. health)?



Gene Set Enrichment Analysis

- Node Size = number genes with an annotation

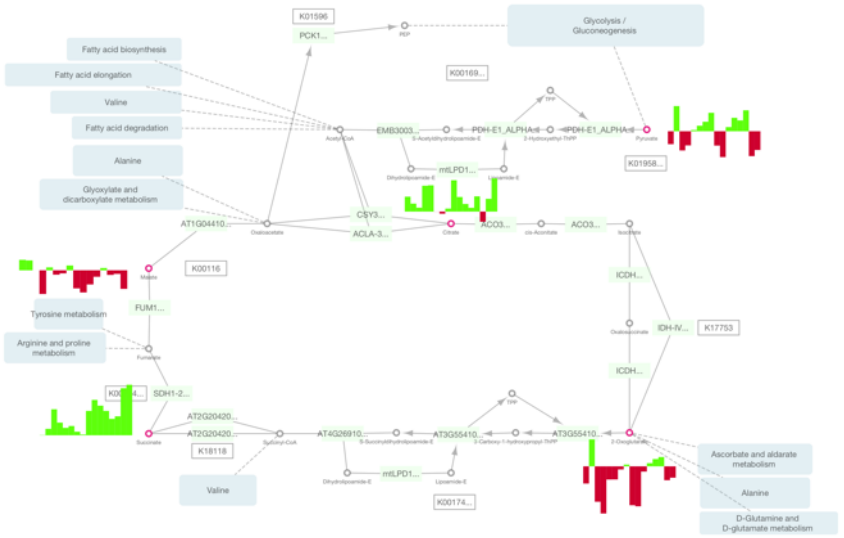
Network analysis

– genes with related function and similar changes in expression

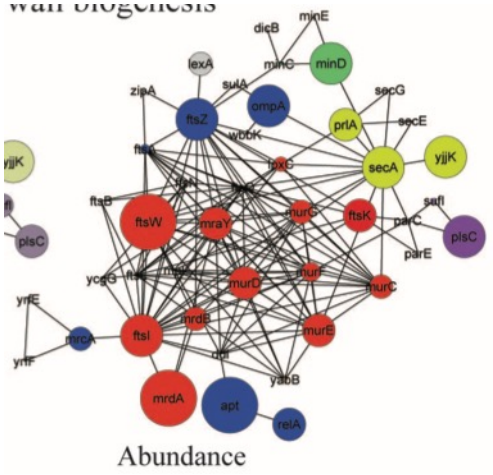
Metatranscriptomics: Understanding results

Cytoscape
www.cytoscape.org
MeV
www.tm4.org/mev.html

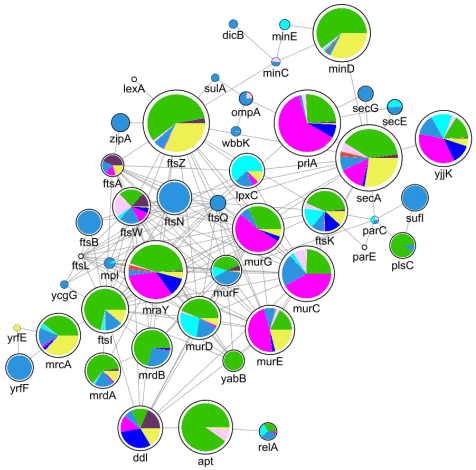
R
D3.js



Genes involved in pathways associated with cell wall biogenesis



Relative abundance



Relative abundance and contributing taxa