Automated annotation

Dan Jones, Bioinformatics Institute

Automated annotation

- Many programs available that vary considerably in focus:
  - Some are taxon or species specific (e.g. specific to microbes, Arabidopsis)
  - Some are very specific in what they predict
    - Genes
    - Promotors / Regulatory regions
    - Repeat elements
    - CpG islands
    - Transposable elements

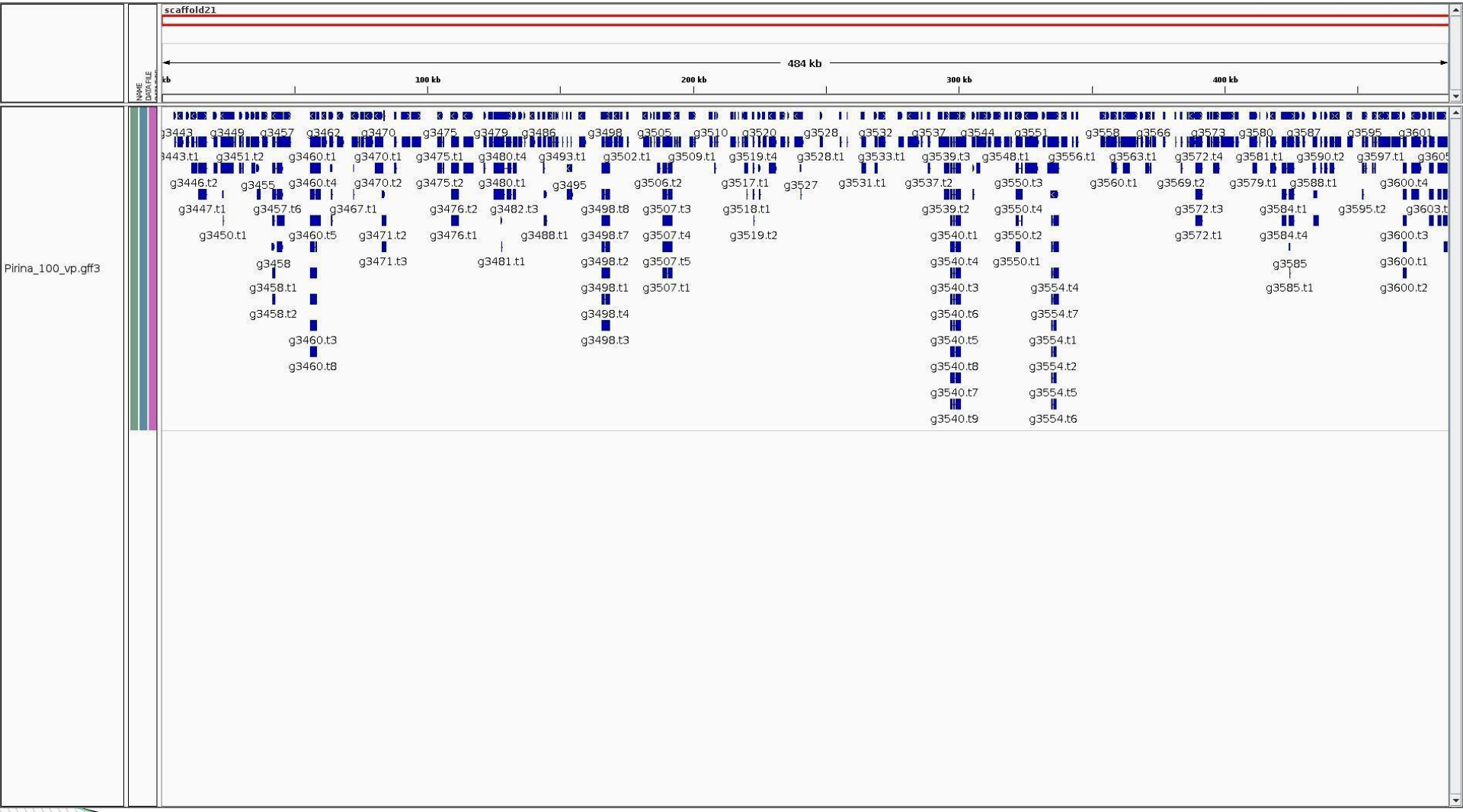
Automated annotation

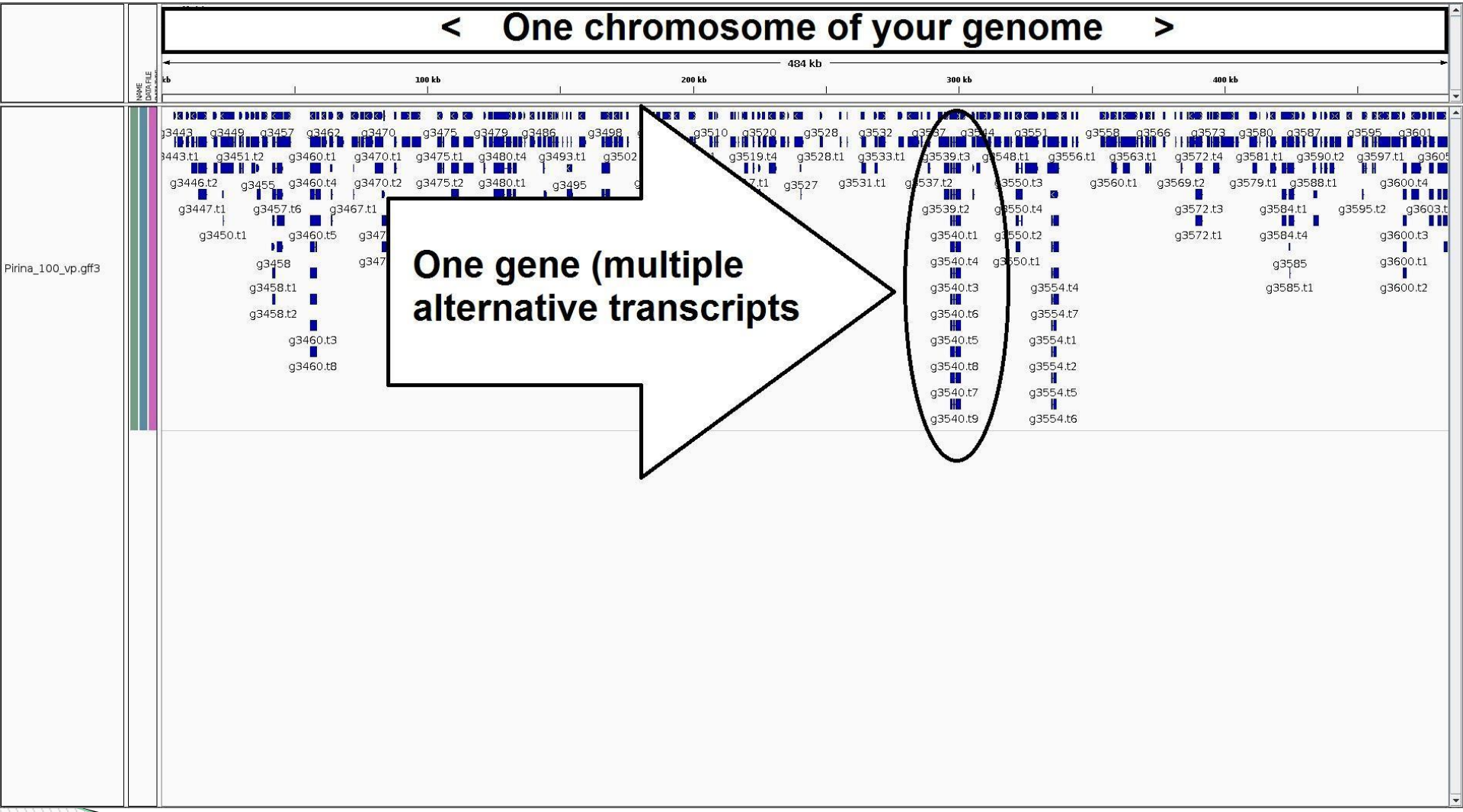
- Generally, **gene prediction** programs require a set of parameters, which tell you:

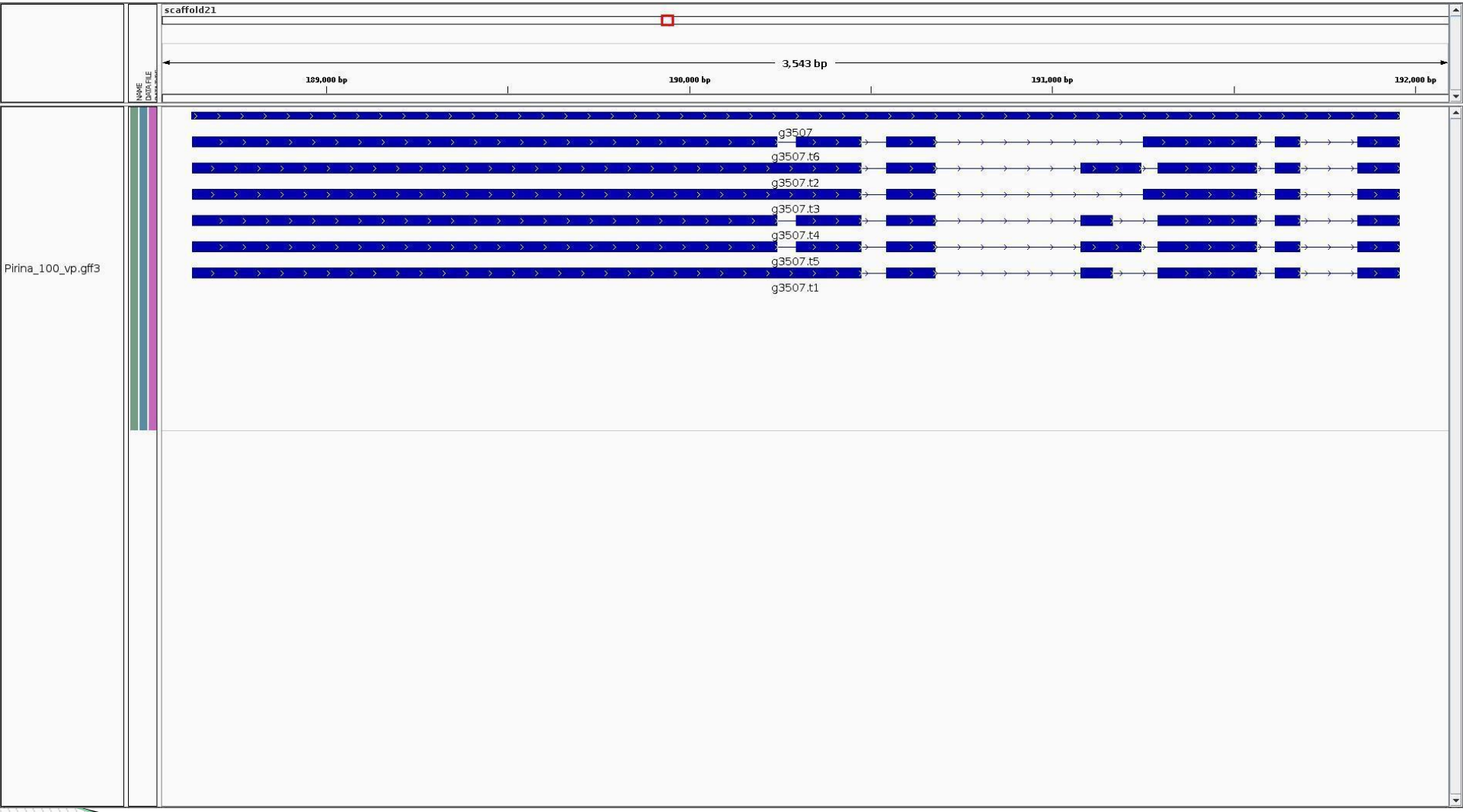
  "This is what a gene tends to look like in this species"

  (Intron / exon size, splice sites, UTRs)

- These are derived from known-to-be-good gene models
  - If you don't have any known-to-be-good gene models, you can use parameters from a related species, but take it with a grain of salt
  - An RNAseq run is a good option to get data to make your own known-to-be-good gene models


scaffold21
484 kb
100 kb
200 kb
300 kb
400 kb
Pirina_100_vp.gff3


< One chromosome of your genome >
484 kb
100 kb
200 kb
300 kb
400 kb
Pirina_100_vp.gff3
One gene (multiple alternative transcripts
g3443
g3449
g3457
g3462
g3470
g3475
g3479
g3486
g3498
g3510
g3520
g3528
g3532
g3544
g3551
g3558
g3566
g3573
g3580
g3587
g3595
g3601
g3443.t1
g3451.t2
g3460.t1
g3470.t1
g3475.t1
g3480.t4
g3493.t1
g3502
g3519.t4
g3528.t1
g3533.t1
g3539.t3
g3548.t1
g3556.t1
g3563.t1
g3572.t4
g3581.t1
g3590.t2
g3597.t1
g3446.t2
g3455
g3460.t4
g3470.t2
g3475.t2
g3480.t1
g3495
g3527
g3531.t1
g3537.t2
g3550.t3
g3560.t1
g3569.t2
g3579.t1
g3588.t1
g3600.t4
g3447.t1
g3457.t6
g3467.t1
g3539.t2
g3550.t4
g3572.t3
g3584.t1
g3595.t2
g3603.t
g3450.t1
g3460.t5
g3540.t1
g3550.t2
g3572.t1
g3584.t4
g3600.t3
g3458
g3540.t4
g3550.t1
g3585
g3600.t1
g3458.t1
g3540.t3
g3554.t4
g3585.t1
g3600.t2
g3458.t2
g3540.t6
g3554.t7
g3460.t3
g3540.t5
g3554.t1
g3460.t8
g3540.t8
g3554.t2
g3540.t7
g3554.t5
g3540.t9
g3554.t6


scaffold21
3,543 bp
189,000 bp
190,000 bp
191,000 bp
192,000 bp
Pirina_100_vp.gff3
g3507
g3507.t6
g3507.t2
g3507.t3
g3507.t4
g3507.t5
g3507.t1

Automated annotation - other genomic features

Repeats

RepeatScout

REPET

Telomeres

TelSeq (predicts using defined telomeric repeats, requires you to map your reads to your genome)

Automated annotation - using RNAseq to improve your gene calls

Augustus can use two sources of information to improve gene calls

- cDNA sequences

ESTs, assembled cDNAs from RNAseq projects, known genes

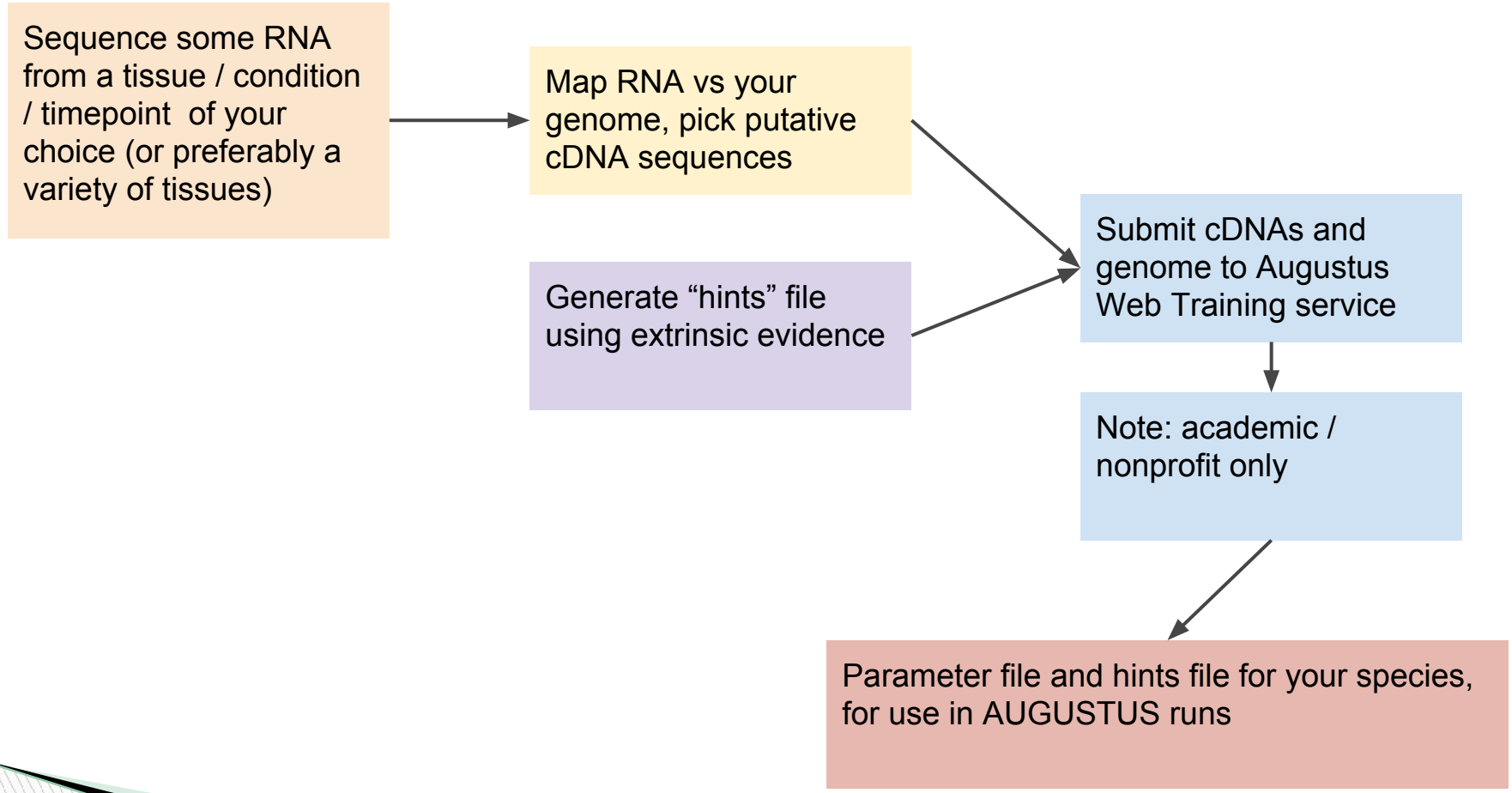
- Raw RNA reads (but not if you're using the web version)

- "Hints"

Hints are extrinsic information on other gene features, e.g. Splice sites, UTRs, non-coding regions, transcription start/stop sites.

Different types of hints can be generated from proteins (using EXONERATE), cDNA (using BLAT), or information on repeat regions (RepeatScout/REPET)

Automated annotation - using RNAseq to improve your gene calls



Viewing and using

Dan Jones

Outputs of a well-designed *de novo* genomics program

- An assembled genome, with an appropriate level of coverage (at least 30x)
- Metrics to assess the assembly quality
- Predicted genes, repeats (low-complexity regions, transposable elements)
- Annotation of the annotations: what are my predicted genes most similar to?
- Mitochondrial and (Chloroplast) genomes
  - Depending on your tissue of origin, organelle genomes will often have a much higher level of coverage than the nuclear genome
- All of the above, in a searchable form, and visualised in a genome browser

Browsers

Free: **IGV**

Tablet (but only for mapped reads)

Gbrowse (not simple to setup, but useful, fast and can be web-facing)

ENSEMBL browser (they do workshops on it!)

ARGO (Broad institute)

Commercial: **Geneious**

CLC Bio

Making your genome searchable

You can (and should) make your genome into a BLAST database to enable searching against this

- Easy to do on command-line (see exercise)
- Many programs like Geneious can do this for you

Collaboration

- Many organisations offering to host your genome in their browser
  - IGV hosts several genomes
  - 1000 fungal genomes

..although microbial genomes + annotations are probably small enough to email around!

- Web-facing browsers like Gbrowse enable you to setup your own website/browser
  - PFR Strawberry genome is a good example

Where to go for help

- SeqAnswers!

www.seqanswers.com

How can we tell if a genome assembly is any good?

N50 (or L50) and L50 (or N50)

N50: Given a set of sequences of varying lengths, the N50 length is defined as the length N for which 50% of all bases in the sequences are in a sequence of length L < N.

L50: The number of contigs that are smaller than the N50 contig

(Why is the N50 a length, and the L50 a number?)
(Contradictory definitions and arguments in SeqAnswers)
(http://seqanswers.com/forums/showthread.php?p=41420)

"Old" methods

Other metrics

Number of contigs
Mean contig length
Average contig length
Average / Mean / Number of contigs of all contigs above a certain size (not always explicitly reported!!)

“Old” methods

Other metrics

Mate-pair analysis: are mate-pairs acting as expected?

Coverage analysis: what's the coverage and how variable is it?

Repeat analysis: do de novo repeat tools (like TRF) match the expected characteristics of the genome?

Breakpoint analysis: where do scaffolds end and why?

"Old" methods

ALE: Assembly Likelihood Evaluator (http://www.ncbi.nlm.nih.gov/pubmed/23303509)

CGAL: Computing Genome Assembly Likelihoods (http://genomebiology.com/2013/14/1/R8)

QUAST: Quality Assessment Tool for Genome Assemblies (http://www.ncbi.nlm.nih.gov/pubmed/23422339)

REAPR (http://www.ncbi.nlm.nih.gov/pubmed/23710727)

LAP (http://www.biomedcentral.com/content/pdf/1756-0500-6-334.pdf)

Mauve (http://www.ncbi.nlm.nih.gov/pubmed/21810901)

AMOSvalidate (http://www.ncbi.nlm.nih.gov/pubmed/18341692)

Plantagora (http://www.plantagora.org/tools_downloads/)

New methods

ALE: Assembly Likelihood Evaluator
(http://www.ncbi.nlm.nih.gov/pubmed/23303509)

Can be used for metagenome or genome assemblies

Does not require reference

Produces a single “ALE” score of likelihood

New methods

CGAL: Computing Genome Assembly Likelihoods
(http://genomebiology.com/2013/14/1/R8)

Can be used for genome assemblies

Does not require reference

Produces a single “score” of likelihood

New methods

QUAST: Quality Assessment Tool for Genome Assemblies (http://www.ncbi.nlm.nih.gov/pubmed/23422339)

Can be used with or without a reference genome

Produces a number of stats in graphical form

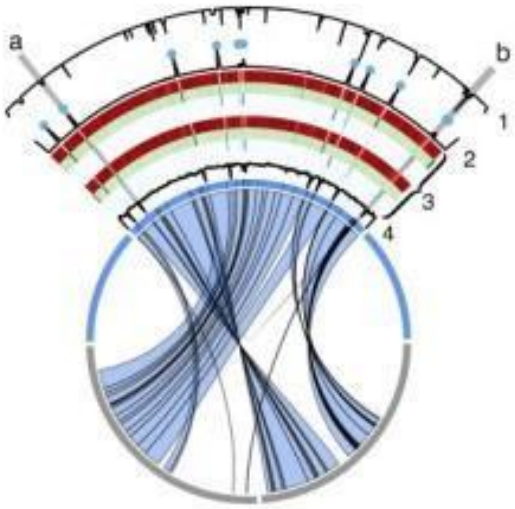
- Trends of stats across the genome (mismatches, indels)
- No. of predicted genes (using GeneMark or Glimmer)

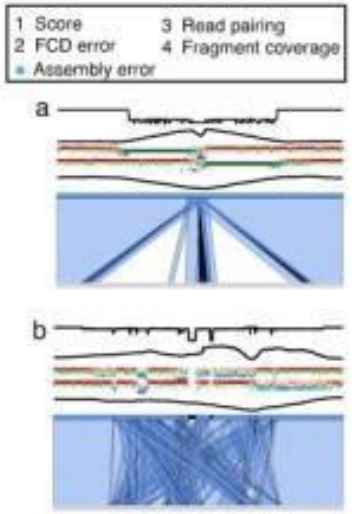
New methods

REAPR (http://www.ncbi.nlm.nih.gov/pubmed/23710727)

Does not require a reference genome

Produces a number of stats in graphical form, some of which are very groovy





New methods

LAP (http://www.biomedcentral.com/content/pdf/1756-0500-6-334.pdf)

Mauve (http://www.ncbi.nlm.nih.gov/pubmed/21810901)

AMOSvalidate (http://www.ncbi.nlm.nih.gov/pubmed/18341692)

Plantagora (http://www.plantagora.org/tools_downloads/)

New methods