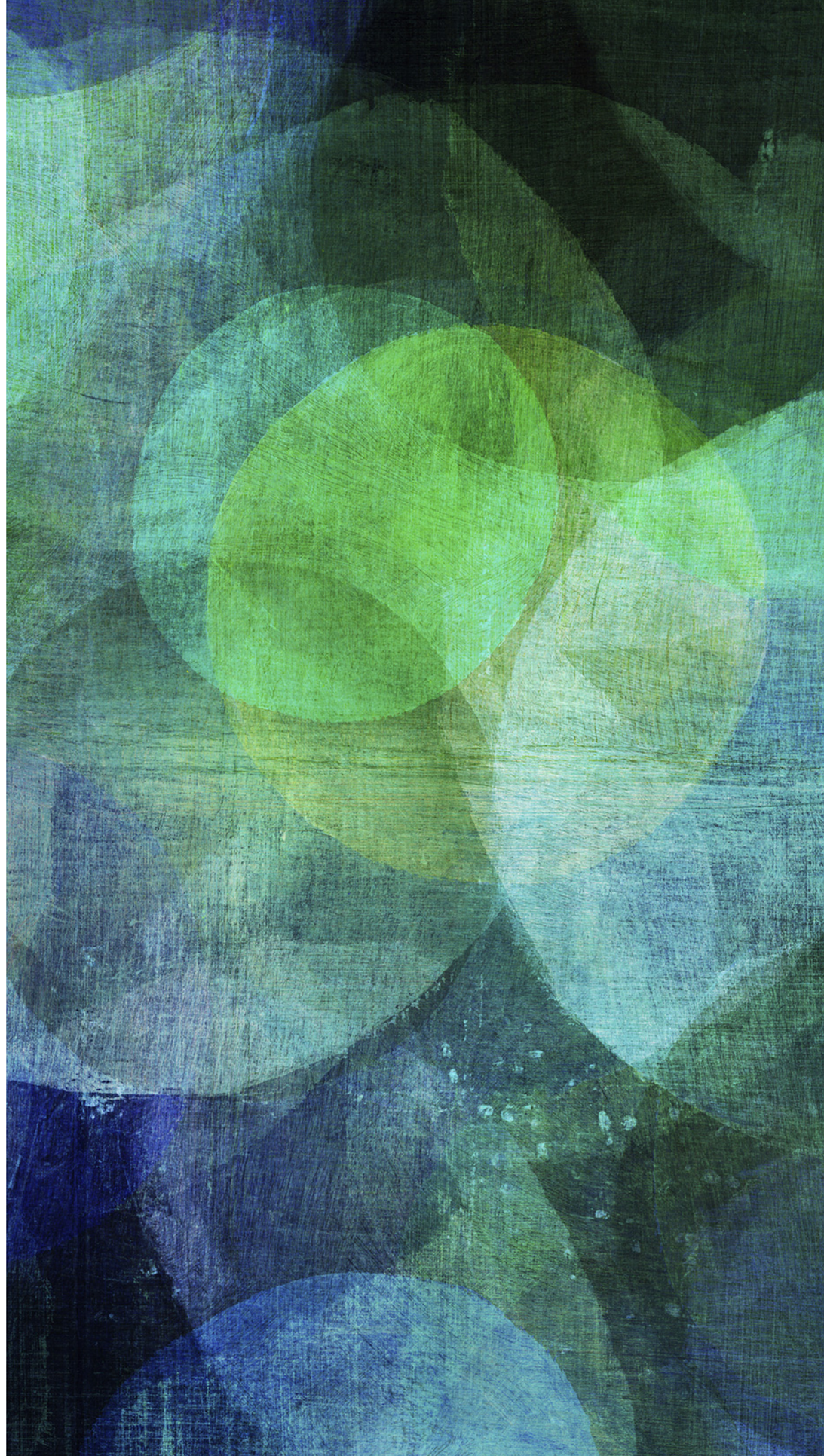
GENE PREDICTION IN AUGUSTUS

Vanessa L. González
Research Bioinformatician
Global Genome Initiative (GGI)

OVERVIEW

- Introduction to Genome Annotation
- General terminology and concepts
- Getting your data ready for gene prediction/annotation.
- Gene prediction in AUGUSTUS
- References/Resources:
  - A beginner's guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342
  - Genome Annotations by Dr. Michael Schatz: http://schatzlab.cshl.edu/teaching/2014/index.shtml#2014AdvSeq
  - Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus
  - Maker Tutorial: http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial

INTRODUCTION TO GENOME ANNOTATION



GENE ANNOTATION

➤ aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgc
atgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatgctaatg
aatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtctt
gggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgcta
atgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaag
ctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcctgcggctatgctaat
gaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtc
ttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctggga
tccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcg
gctatgctaatgcatgcggctatgctaagctcatgcggctatgctaagctgggaatgcatgcggctatgctaagctg
ggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagct
gcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatggtcttgggatttaccttggaatgcta
agctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcat
gcggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaat
gcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagct
catgcgg

Genome Annotations by Dr. Michael Schatz: http://schatzlab.cshl.edu/teaching/2014/index.shtml#2014AdvSeq

GENE ANNOTATION

➤ aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgc
atgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatgctaatg
aatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtctt
gggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgcta
atgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaag
ctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcctgcggctatgctaat
gaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtc
ttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaagctgggatccgatgac
aatgcatgcggctatgctaatgcatgcggctatgctaagctcatgcggctatgctaagctgggaatgcatgcggctat
gctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactat
gctaagctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatggtcttgggatttaccttg
gaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgc
taatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctat
gctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgct
aagctcatgcgg

GENE

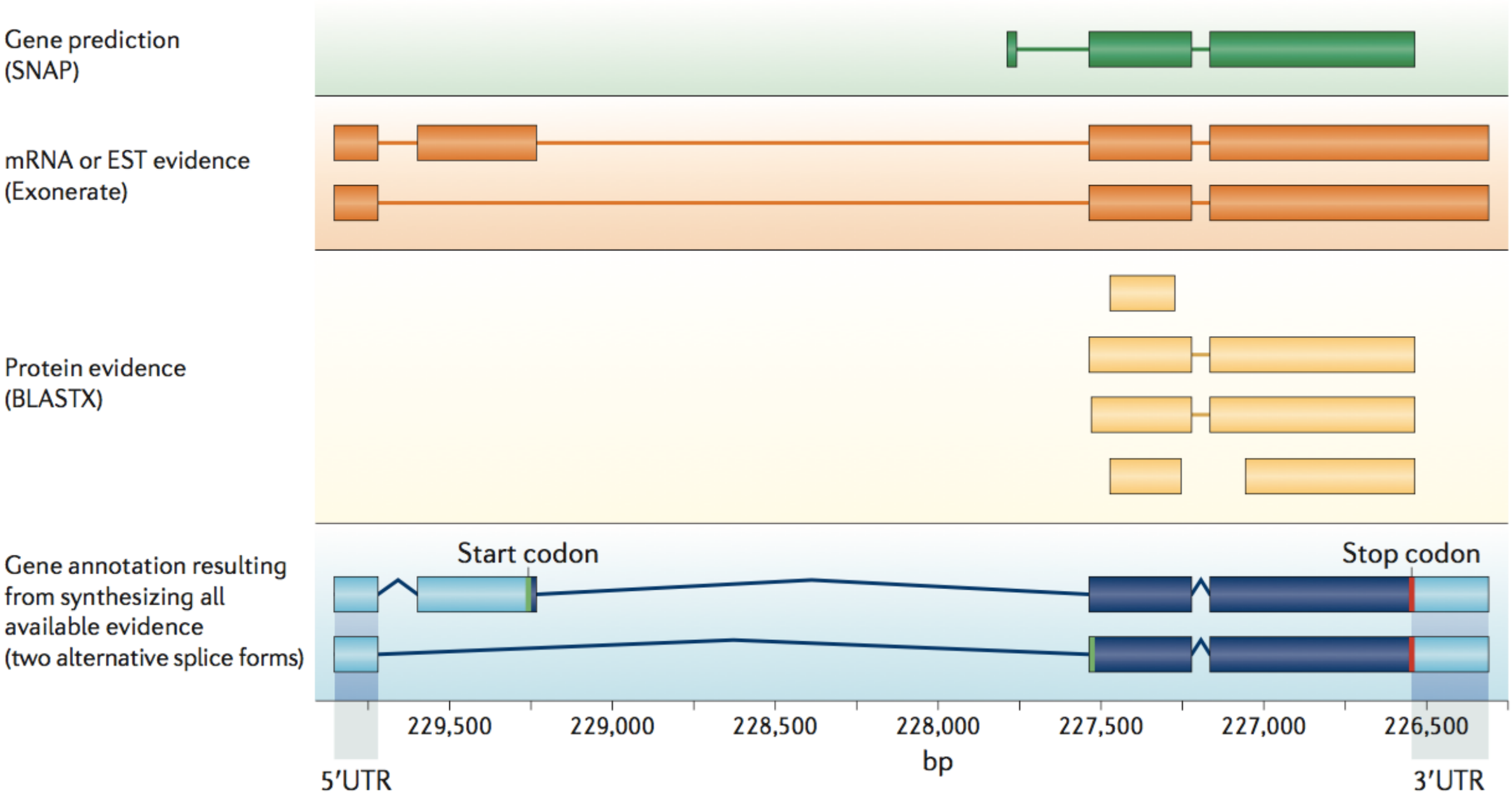
Genome Annotations by Dr. Michael Schatz: http://schatzlab.cshl.edu/teaching/2014/index.shtml#2014AdvSeq

GENE PREDICTION VERSUS GENE ANNOTATION

- **Gene prediction ≠ Gene annotation.**
- Gene predictions are partial gene models.
- Gene annotations are gene models but should include a documented evidence trail supporting the model in addition to quality control metrics.

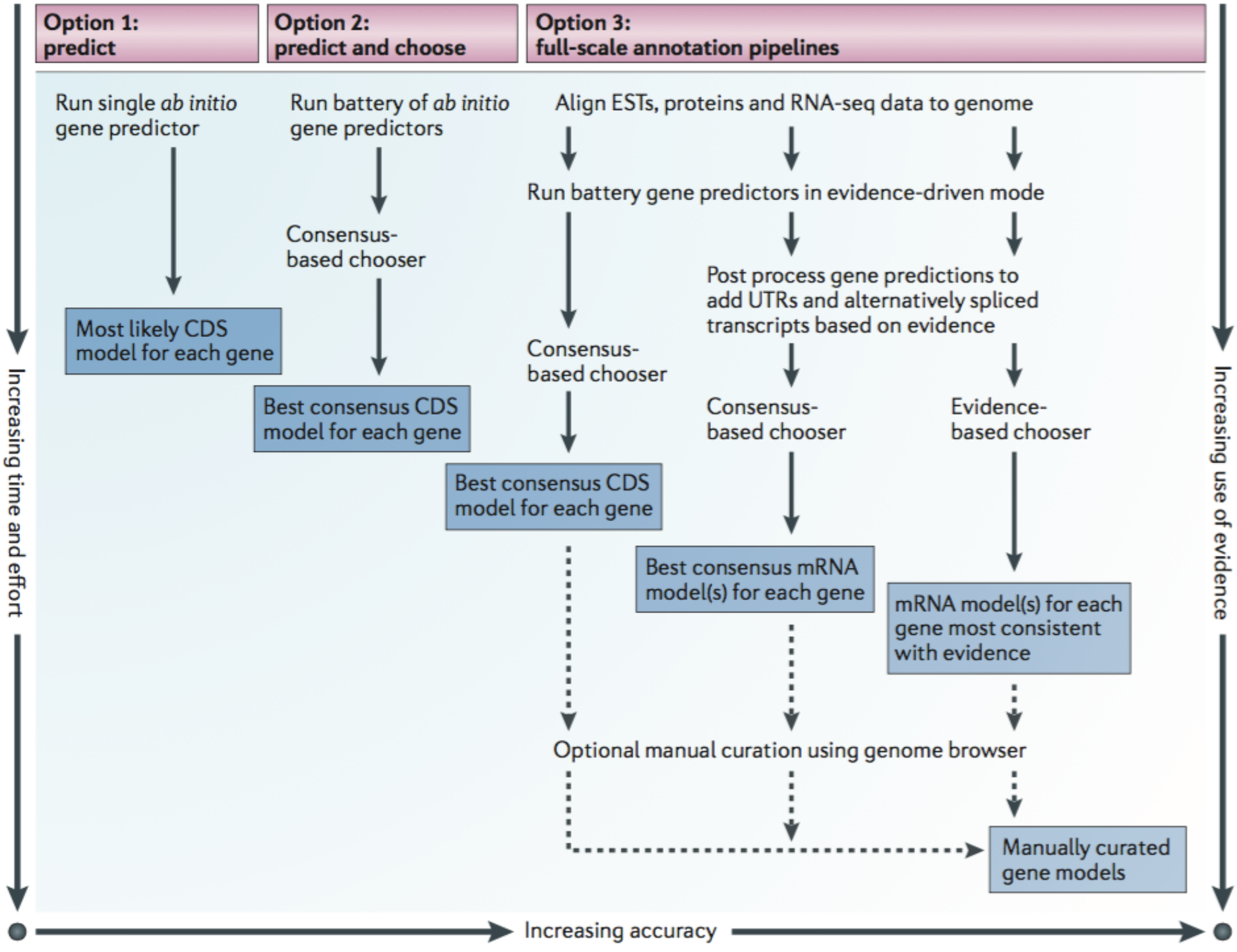
A beginner's guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342

GENE PREDICTION VERSUS GENE ANNOTATION



A beginner's guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342

BASIC APPROACHES TO GENOME ANNOTATION

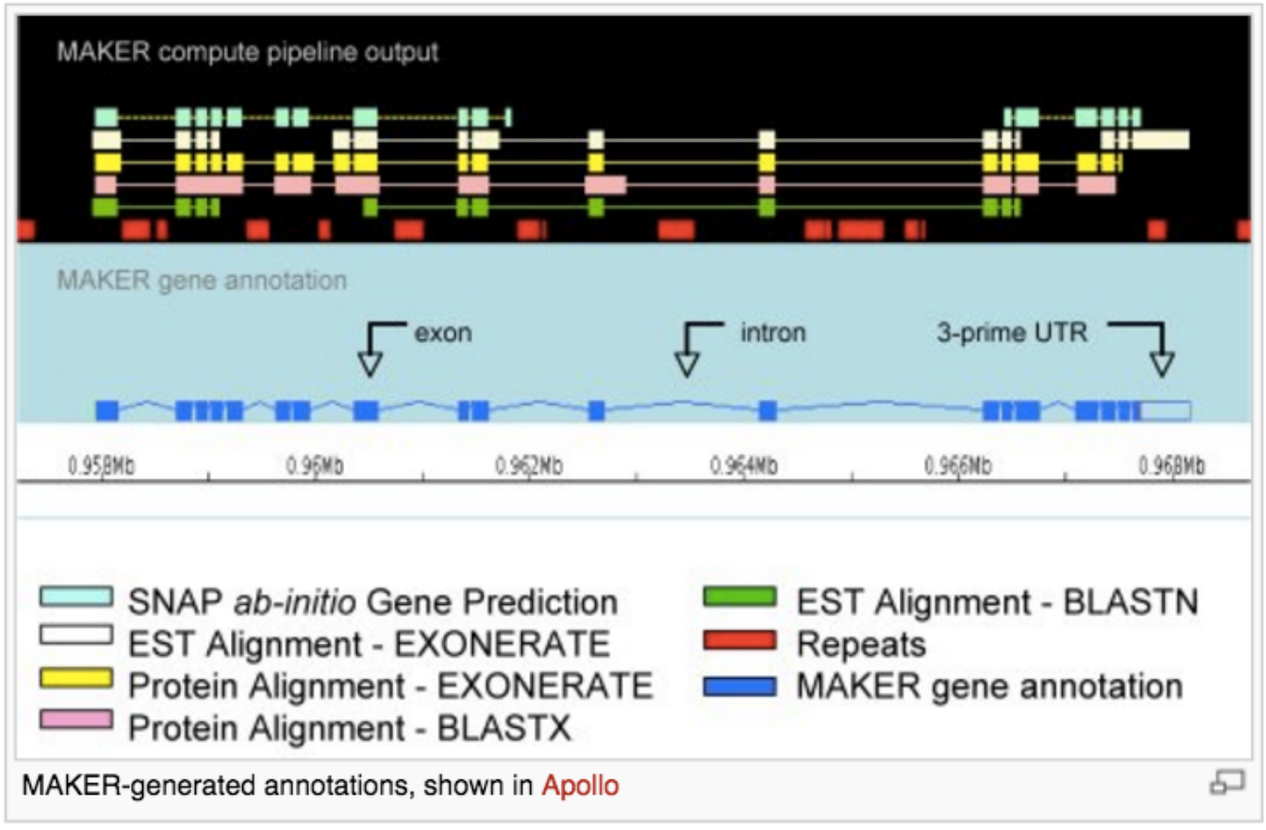


A beginner's guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342

GENE ANNOTATION PIPELINE EXAMPLE: MAKER

What does MAKER do?

- Identifies and masks out repeat elements
- Aligns ESTs to the genome
- Aligns proteins to the genome
- Produces *ab initio* gene predictions
- Synthesizes these data into final annotations
- Produces evidence-based quality values for downstream annotation management



Maker Tutorial: http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial

ANNOTATIONS.

- Annotations are descriptions of different features of the genome, and they can be **structural** or **functional** in nature.
- **Structural genome annotation** is the process of identifying genes in their intron-exon structures.
  - i.e. exons, introns, UTRs, splice forms.
- **Functional genome annotation** is the process of attached meta-data such as gene ontology terms to structural annotations.

A beginner's guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342

SETTING UP AN ANNOTATION PROJECT.

- Quality of genome assembly - important consideration for genome annotation.
- “Gene size N50 scaffold length is needed for a decent target annotation.”
- Consider Scaffold and contig N50, percent gaps, and percent coverage.

A beginner’s guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342

ANNOTATION STEPS.

- Two phases of genome annotation:
- Phase one: Computation phase. ESTs, proteins, and so on are aligned to the genome and ab innate and/or evidence-driven gene predictions are generated.
- Phase two: Annotation phase. These data are synthesized into gene annotations.
- Programs that assemble compute data and use it to create annotations are referred to as “pipelines.” E.g. Maker, JAMg, PASA.
- Current annotation pipelines are focused on the annotation of protein-coding genes.

A beginner’s guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342

COMPUTATION PHASE

- training, optimizing, and configuring gene prediction tools.
- 1. Repeat Identification
  - refers to ‘low complexity’ sequences as well as homopolymers, transposable elements, long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs).
  - Identifying repeats in complicated.
  - Tools exist to identify stretches of a sequence in a target genome that are homologous to known repeats. E.g. RepeatMasker.
  - Masking = transform all indentified repeats to Ns or lower case nucleotides (soft masking) to signal downstream applications that these are repeats.

A beginner’s guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342

COMPUTATION PHASE

- training, optimizing, and configuring gene prediction tools.
- 2. Evidence Alignment - Examples of evidence supporting a structural annotation: Ab initio gene predictions, Transcribed RNA (mRNA-Seq/ESTs/cDNA/transcript), Proteins
- 3. *Ab initio* gene prediction - *Ab initio* prediction means that no other input is used than the target genome itself. They provide a fast and easy way to identify genes in assembled DNA sequences
- 4. Evidence driven gene prediction - Example, using ESTs to identify exon boundaries.

A beginner's guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342

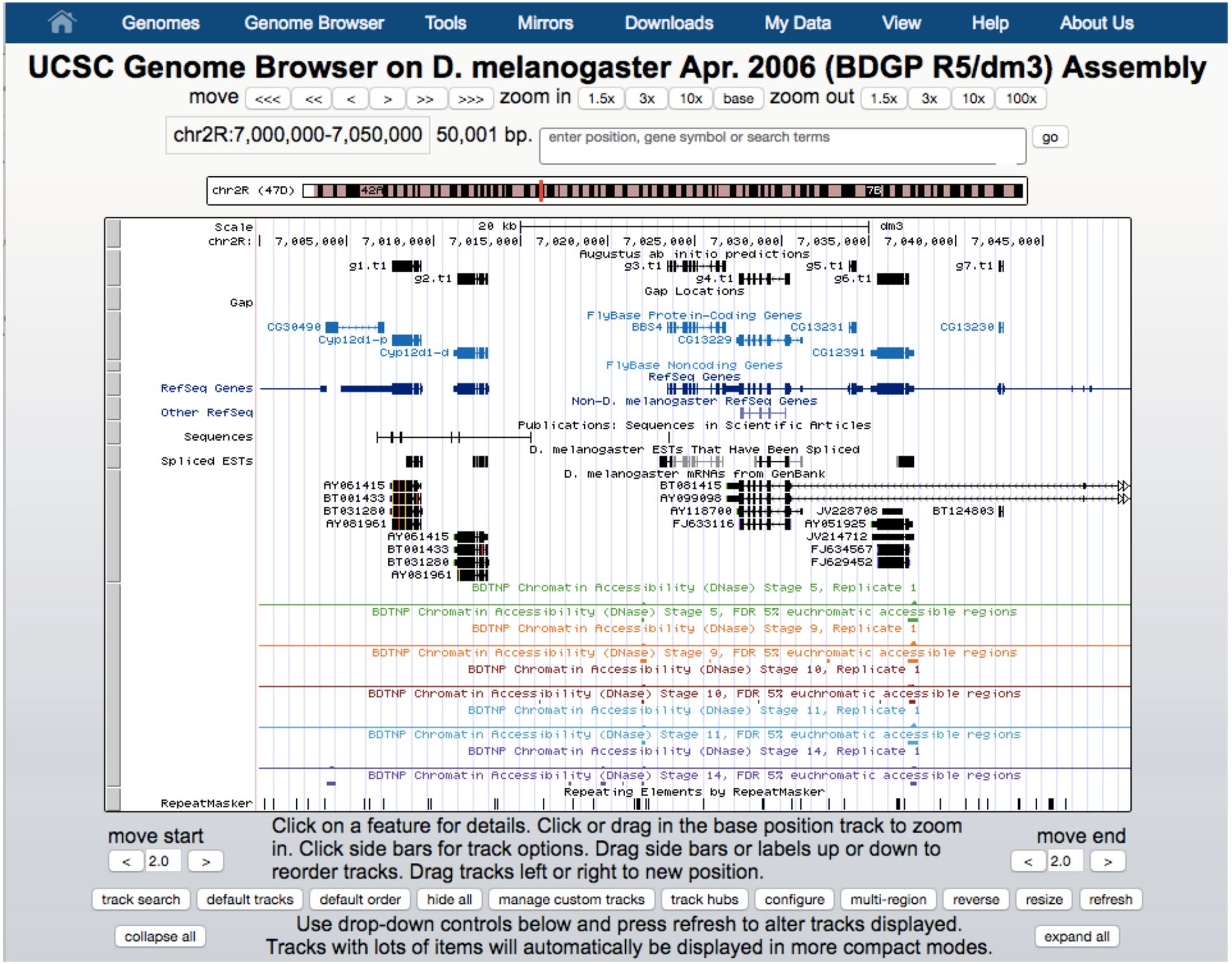
ANNOTATION PHASE

- obtaining a synthesis of alignment-based evidence with *ab initio* gene predictions to obtain a final set of gene annotations.
  - Manual Curation - E.g. Web Apollo
  - Automated annotation - run different gene finders on the genome and then use a ‘combiner’ aka ‘chooser algorithm’ to select single prediction whose intron-exon structure best represents the consensus of the models from among the overlapping predictions that define each putative gene locus. E.g GLEAN, JIGSAW.

A beginner’s guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342

VISUALIZING ANNOTATION DATA

- Apollo
- JBROWSE
- IGV



A beginner's guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342

“

Like parenthood, annotation responsibilities do not end in birth.

-Yandell and Ence (2012)

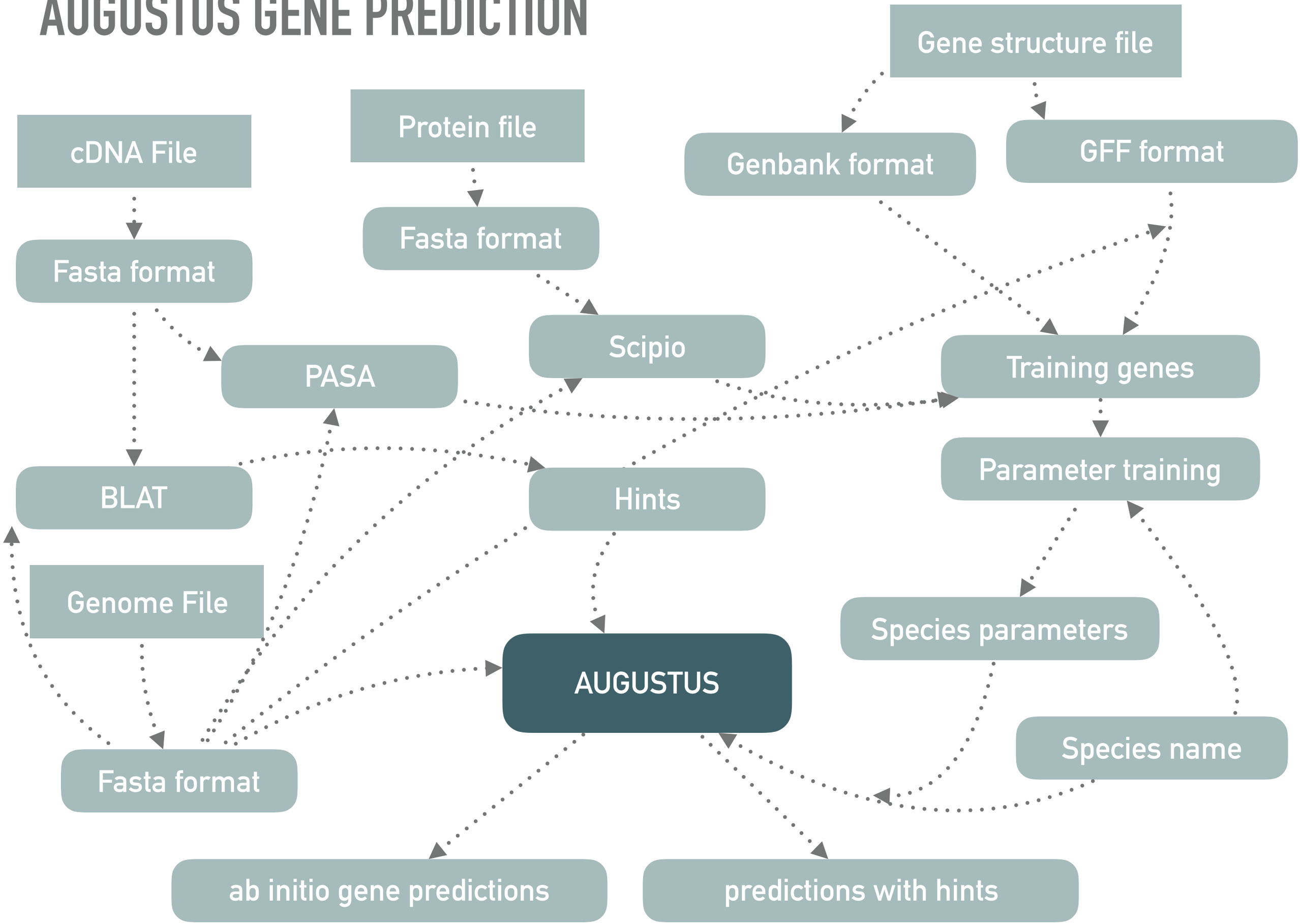
A beginner’s guide to eukaryotic genome annotation. Yandell and Ence (2012) *Nature Reviews*, 13:329-342

GENE PREDICTION IN AUGUSTUS

ab initio and evidence-drivable (hints) gene predictor

Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus

AUGUSTUS GENE PREDICTION



AUGUSTUS — STEPS

- 1. Compile a training set
- 2. Train the coding regions of AUGUSTUS
- 3. *Ab initio* predict the genes in the genome.
- 4. Prepare Hints
- 5. Predict genes using hints
- 6. Identify members of a protein family (optional, see tutorial for more information).

Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus

1. COMPILE A TRAINING SET

- You will need a set of genomic sequences with bona fide gene structures (sequence coordinates of starts and ends of exons and genes).
- In many cases, or as a first step towards modeling complete genes, it is sufficient to have only the coding parts of the gene structure (CDS).
- There are several typical options for creating a training set to estimate the parameters of gene finders.

Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus

1. COMPILE A TRAINING SET

- Augustus uses Scipio and PASA to create a (training) set of gene structures.
- In the case where a genome and protein sequences from the same or very closely related species are given, Scipio can be used to construct the gene structure from the sequences.
- Scipio uses BLAT to align the protein sequences against the genome and then determines the exact boundaries of the exons, and adds small exons that were not found by BLAT.
- PASA is an annotation pipeline that aligns EST and protein sequences to the genome and produces evidence-driven gene models.

Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus

1. COMPILE A TRAINING SET

- Other options for compiling a set of gene structures
  - pre-existing gene structures (e.g. from GenBank)
  - spliced alignments of ESTs against the assembled genomic sequence (e.g. using PASA)
  - spliced alignments of de novo assembled transcriptome short reads (RNA-Seq)
  - spliced alignments of protein sequences of the same or a very closely related species against the assembled genomic sequence, e.g. using Scipio.
  - gene structures from a related species
  - iteration of training with predicted genes, starting with an existing parameter set

Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus

2. TRAIN CODING REGIONS OF AUGUSTUS

- Parameter sets exist for many taxa, however, you can optimize for you taxon.
- To estimate the parameters from the training set, AUGUSTUS will split gene structure sets into a training set and test set, to then test accuracy.
- Once you have made an initial training set. Species specific parameters must be optimized for prediction accuracy.

Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus

3. AB INITIO PREDICT GENES IN THE GENOME

- *Ab initio* prediction means that no other input is used than the target genome itself. In this step AUGUSTUS will predict the protein-coding parts of the genes.
- Example: Predict the genes in the range 7,000,001-7,500,000 of chr2R of D. melanogaster.
- augustus --species=fly --predictionStart=7000001 --predictionEnd=7500000 chr2R.fa > augustus.abinitio.gff

```
# This output was generated with AUGUSTUS (version 2.5).
...
# start gene g1
chr2R AUGUSTUS   gene        7007533 7010935 0.02  - .  g1
chr2R AUGUSTUS   transcript  7007533 7010935 0.02  - .  g1.t1
chr2R AUGUSTUS   tts         7007533 7007533 .     - .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   exon        7007533 7008630 .     - .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   stop_codon  7007610 7007612 .     - 0  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   intron      7008631 7008694 1     - .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   intron      7008812 7008865 0.88  - .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   intron      7009192 7009251 0.95  - .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   CDS         7007610 7008630 1     - 1  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   CDS         7008695 7008811 0.88  - 1  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   exon        7008695 7008811 .     - .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   CDS         7008866 7009191 0.99  - 0  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   exon        7008866 7009191 .     - .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   CDS         7009252 7009353 0.95  - 0  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   exon        7009252 7009429 .     - .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   start_codon 7009351 7009353 .     - 0  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   exon        7010820 7010935 .     - .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   tss         7010935 7010935 .     - .  transcript_id "g1.t1"; gene_id "g1";
# protein sequence = [MNTLSSARSVAIYVGPVRSSRSASVLAHEQAKSSITEEHKTYDEIPRPNKFKFMRAFMPGGEFQNASITEYTSAMRKR
# YGDIYVMPGMFGRKDWVTTFNTKDIEMVFRNEGIWPRRDGLDSIVYFREHVRPDVYGEVQGLVASQNEAWGKLRSAINPIFMQPRGLRMYYEPLSNIN
# NEFIERIKEIRDPKTLEVPEDFTDEISRLVFESLGLVAFDRQMGLIRKNRDNSDALTLFQTSRDIFRLTFKLDIQPSMWKIISTPTYRKMKRTLNDSL
# NVAQKMLKENQDALEKRRQAGEKINSNSMLERLMEIDPKVAVIMSLDILFAGVDATATLLSAVLLCLSKHPDKQAKLREELLSIMPTKDSLLNEENMK
# DMPYLRAVIKETLRYYPNGLGTMRTCQNDVILSGYRVPKGTTVLLGSNVLMKEATYYPRPDEFLPERWLRDPETGKKMQVSPFTFLPFGFGPRMCIGK
# RVVDLEMETTVAKLIRNFHVEFNRDASRPFKTMFVMEPAITFPFKFTDIEQ]
# end gene g1
...
```

Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus

4. PREPARE HINTS

- Hints are extrinsic evidence about the location and structure of genes in a particular GFF format.
- When predicting genes AUGUSTUS can incorporate these hints, which will change the likelihood of gene structures candidates. It will tend to predict gene structures that are in agreement with the hints.
- Sources of Hints:
  - ESTs or mRNAs transcriptome reads, long enough to span several exons (454, Sanger)
  - RNA-Seq high coverage short read transcriptome sequences (Illumina, SOLiD)
  - Genomic conservation

Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus

4. PREPARE HINTS

- Most commonly, AUGUSTUS is used to constructs extrinsic evidence about genes from transcriptome data (ESTs and RNA-Seq).

Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus

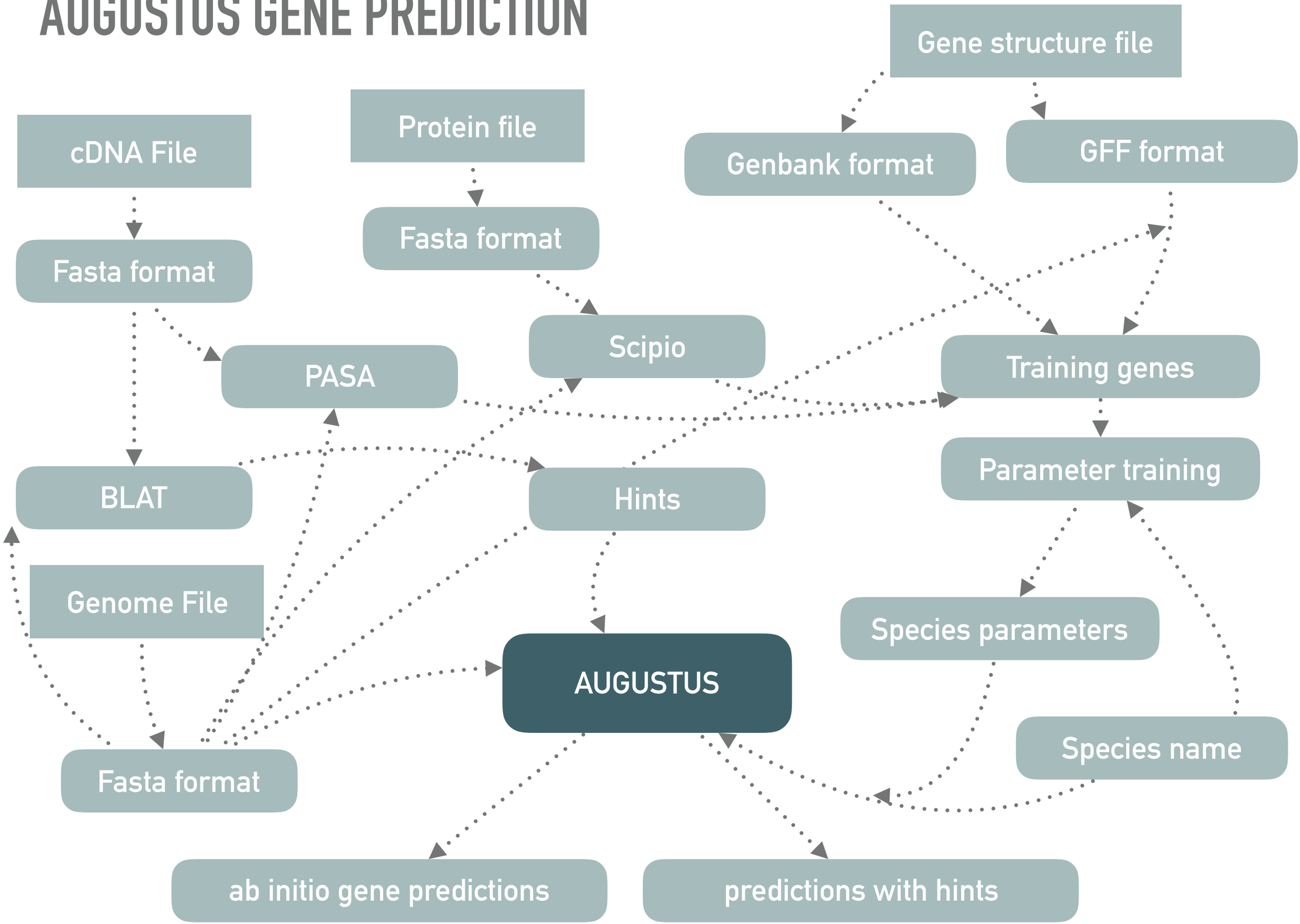
PREDICT GENES USING HINTS

- Structural annotation step: Once you have the hints and parameters trained you can get a more accurate prediction of gene models.
- Example: Predict the genes in the range 7,000,001-7,500,000 of chr2R of D. melanogaster using evidence from hints.

```
# start gene g1
chr2R AUGUSTUS   gene        7007533 7009385 0.2 -  .  g1
chr2R AUGUSTUS   transcript  7007533 7009385 0.2 -  .  g1.t1
chr2R AUGUSTUS   tts         7007533 7007533 .   -  .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   exon        7007533 7008630 .   -  .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   stop_codon  7007610 7007612 .   -  0  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   intron      7008631 7008694 1   -  .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   intron      7008812 7008865 1   -  .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   intron      7009192 7009251 1   -  .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   CDS         7007610 7008630 1   -  1  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   CDS         7008695 7008811 1   -  1  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   exon        7008695 7008811 .   -  .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   CDS         7008866 7009191 1   -  0  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   exon        7008866 7009191 .   -  .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   CDS         7009252 7009353 0.94-  0  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   exon        7009252 7009385 .   -  .  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   start_codon 7009351 7009353 .   -  0  transcript_id "g1.t1"; gene_id "g1";
chr2R AUGUSTUS   tss         7009385 7009385 .   -  .  transcript_id "g1.t1"; gene_id "g1";
# protein sequence = [MNTLSSARSVAIYVGPVRSSRSASVLAHEQAKSSITEEHKTYDEIPRPNKFKFMRAFMPGGEFQNASITEYTSAMRKR
# YGDIYVMPGMFGRKDWVTTFNTKDIEMVFRNEGIWPRRDGLDSIVYFREHVRPDVYGEVQGLVASQNEAWGKLRSAINPIFMQPRGLRMYYEPLSNIN
# NEFIERIKEIRDPKTLEVPEDFTDEISRLVFESLGLVAFDRQMGLIRKNRDNSDALTLFQTSRDIFRLTFKLDIQPSMWKIISTPTYRKMKRTLNDSL
# NVAQKMLKENQDALEKRRQAGEKINSNSMLERLMEIDPKVAVIMSLDILFAGVDATATLLSAVLLCLSKHPDKQAKLREELLSIMPTKDSLLNEENMK
# DMPYLRAVIKETLRYYPNGLGTMRTCQNDVILSGYRVPKGTTVLLGSNVLMKEATYYPRPDEFLPERWLRDPETGKKMQVSPFTFLPFGFGPRMCIGK
# RVVDLEMETTVAKLIRNFHVEFNRDASRPFKTMFVMEPAITFPFKFTDIEQ]
# Evidence for and against this transcript:
# % of transcript supported by hints (any source): 100
# CDS exons: 4/4
#      E:   4
#      W:   4
# CDS introns: 3/3
#      E:   3
# 5'UTR exons and introns: 1/1
#      E:   1
# 3'UTR exons and introns: 1/1
#      W:   1
# hint groups fully obeyed: 137
#      E:   4 (gi|15542574,SRR023546.8642467/1)
#      W: 133
# incompatible hint groups: 18
#      E:  18 (gi|13769068,gi|4203815,gi|15543927,gi|38623822,gi|15539951,gi|14693753,gi|14699170,...)
# end gene g1
...
```

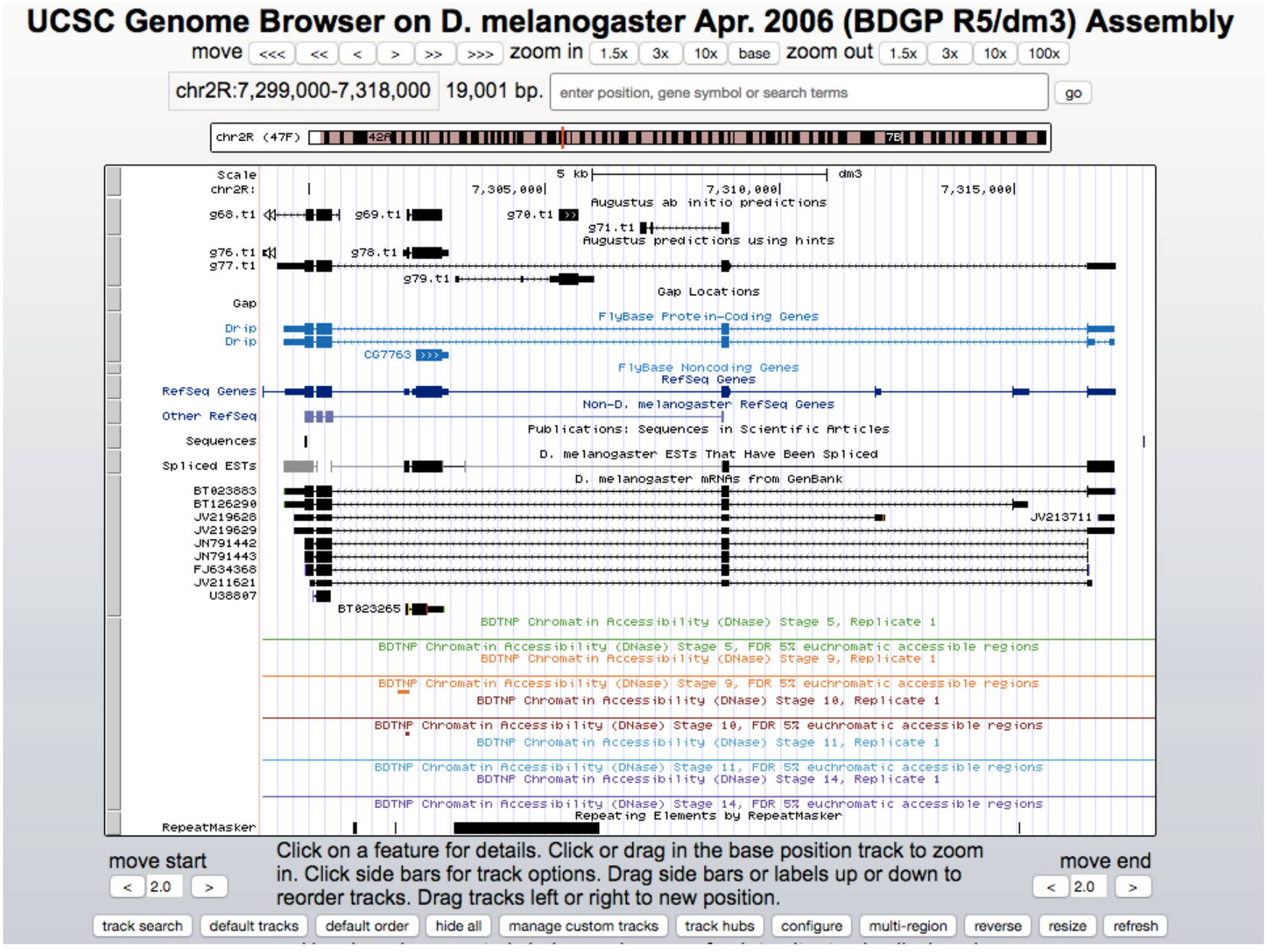
Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus

AUGUSTUS GENE PREDICTION



VISUALIZE AB INITIO AND HINT PREDICTED GENES

- Now that you have your predicted genes you can visualize these in a genome browser.



Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus

AUGUSTUS — WEB SERVER TO TRAIN AUGUSTUS

- http://bioinf.uni-greifswald.de/webaugustus/training/create

Data Input for Training AUGUSTUS

Use this form to submit data for training AUGUSTUS parameters for novel species/new genomic data.

Before submitting a training job for your species of interest, please check whether parameters have already been trained and have been made publicly available for your species at **our species overview table**

Please read the **training tutorial** before submitting a job for the first time. Example data for this form is available **here**. You may also use the button below to insert sample data. Please note that you will always need to enter the verification string at the bottom of the page, yourself, in order to submit a job!

Fill in Sample Data

We strongly recommend that you specify an **E-mail address**! Please read the **Help** page before submitting a job without e-mail address! You have to give a **species name**, and a **genome file**!

Current problem: Regrettably, our server is currently connected to the internet via a rather unreliable connection. This may cause connection timeouts (caused by server side) when uploading big files. Please use the web link upload option, instead, if you experience such problems. We apologize for the inconvenience!

E-mail **Help**

Species name * **Help**

There are two options for sequence file (fasta format) transfer:
You may **either** upload data files from your computer **or** specify web links. **Help**

Please read our **instructions about fasta headers** before using this web service! Most problems with this web service are caused by a wrong fasta header format!

Genome file * (max. 250000 scaffolds) **Help**

Upload a file (max. 100 MB): Choose File No file chosen

or

specify web link to genome file (max. 1 GB):

Augustus Tutorials: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.Augustus