Genome sequencing, assembly and annotation

Erwin Datema

26-01-2010

PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

Overview

- What and why?
- Sequencing methodologies and technologies
- Sequence assembly
- From sequence to biology: extracting information

What?

- Genome sequencing
  - Determining the order of nucleotides in a DNA molecule
- Genome assembly
  - Reconstructing the complete sequence of a DNA molecule from short sequence fragments (“reads”)
- Genome annotation
  - Assigning a (possible) function to a string of nucleotides

Why?

- Genome sequencing
  - Identification of the molecular ‘blueprints’ for traits of interest (disease, agriculture, etc)
- Genome assembly
  - It is currently impossible to sequence a complete chromosome in one go
- Genome annotation
  - Unraveling the (molecular) mystery of life

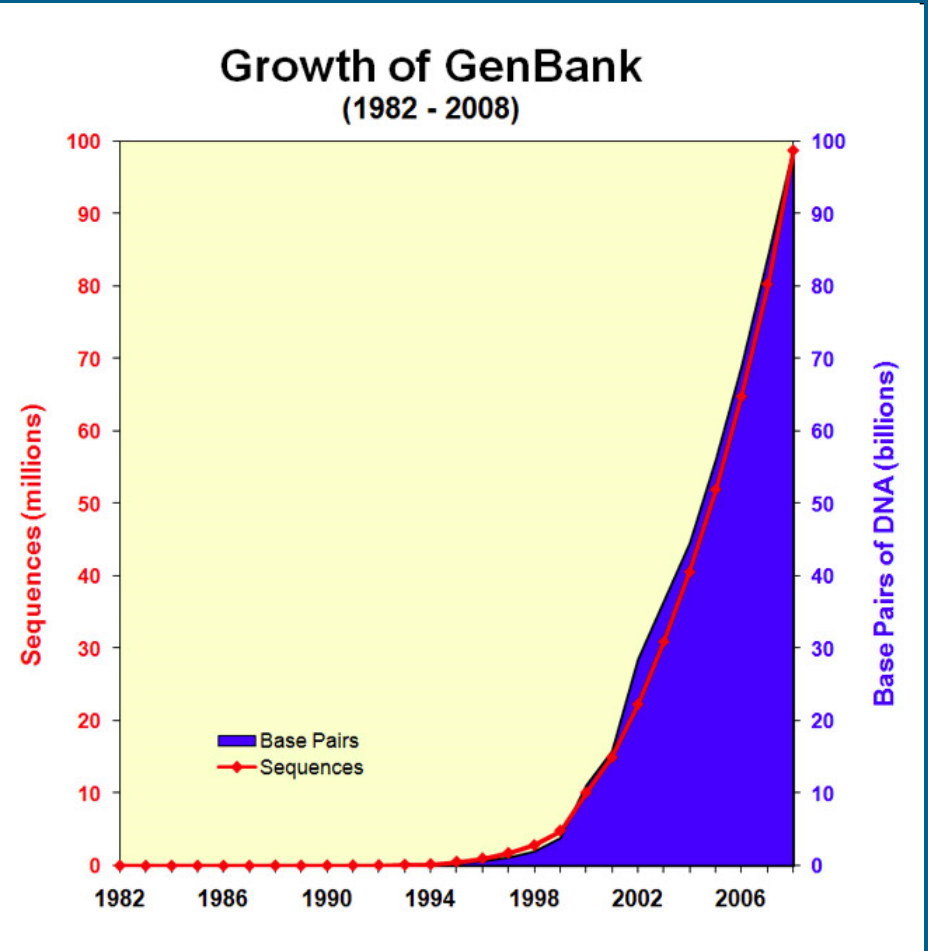
Sequencing and assembly: terms and definitions

- Sequence: linear order of nucleotides as they appear on the DNA molecule
- Read: single observation of the (partial) sequence of a DNA molecule
- Contig: contiguous stretch of sequence, often derived from multiple reads
- Scaffold: linearly ordered and oriented group of contigs

Current advances in genome sequencing

- 2000: 1st Human genome sequence ‘completed’
  - ~3,000,000,000 US dollar
  - Ten years
  - Many large sequencing and bioinformatics centers
- 2007: Watson’s genome sequence ‘completed’
  - ~2,000,000 US dollars
  - Couple of months
  - Small group of companies and research institutes
- Today: Your genome sequence ‘completed’
  - ~20,000 US dollars
  - Couple of weeks
  - One person in the lab, one behind the computer

Sequence data is accumulating…

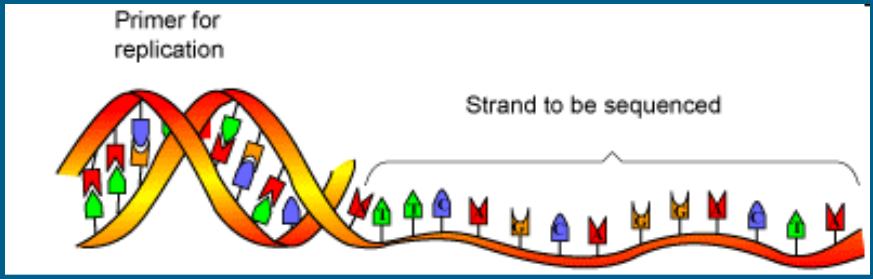


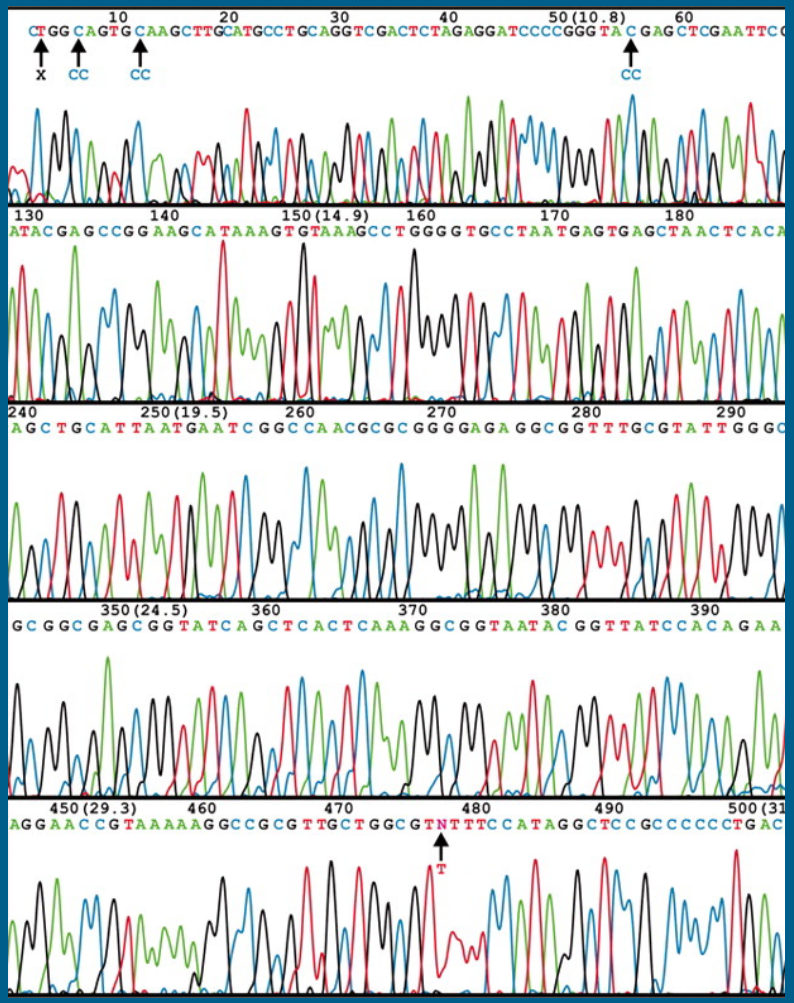
YEAR	# NT	# SEQS
1982	680,338	606
1992	101,008,486	78,608
1997	1,160,300,687	1,765,847
2000	11,101,066,288	10,106,023
2008	99,116,431,942	98,868,465

Sequencing technologies

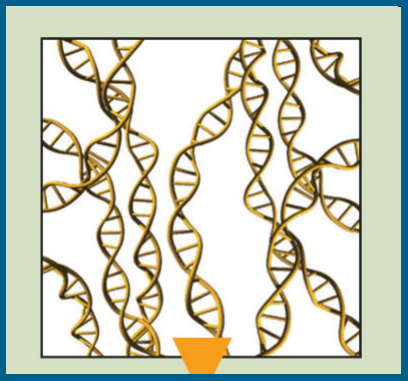
Sanger	800 nt	~80 kb
Roche / 454	450 nt	0.5 Gb
AB / SOLiD	2x 50 nt	60 Gb
Illumina / Solexa	2x100 nt	200 Gb

Sanger sequencing



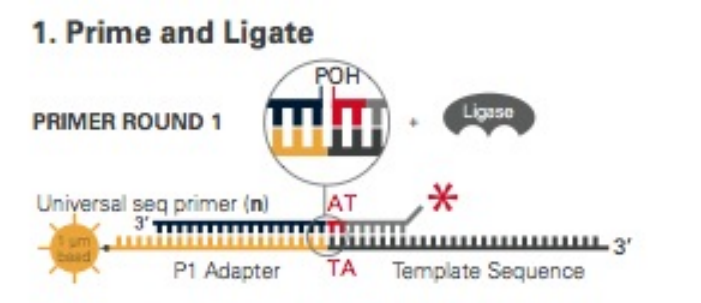


Roche/454 Sequencing

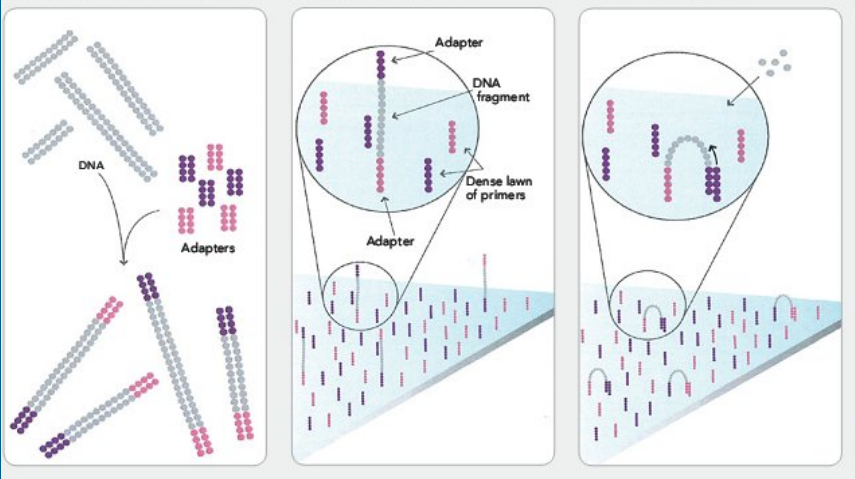


- DNA isolation and fragmentation
- Adapter ligation
- Hybridisation to micro-bead
  - one DNA molecule to one bead
- emPCR amplification
  - Population of identical molecules on each bead
- Micro-beads are loaded onto pico-titerplate and sequenced
  - pyro-sequencing

ABI/SOLiD sequencing



Illumina/Solexa sequencing



Sequencing approaches

- Clone-based sequencing
  - Requires ‘tiling path’ of clones that span the genome
  - Reduces the assembly problem to small fragments
  - ‘Finished’ genome sequence
  - Expensive

- Whole Genome Shotgun sequencing
  - No additional molecular or genetic resources required
  - Higher complexity of sequence assembly
  - Draft genome sequence
  - Less expensive

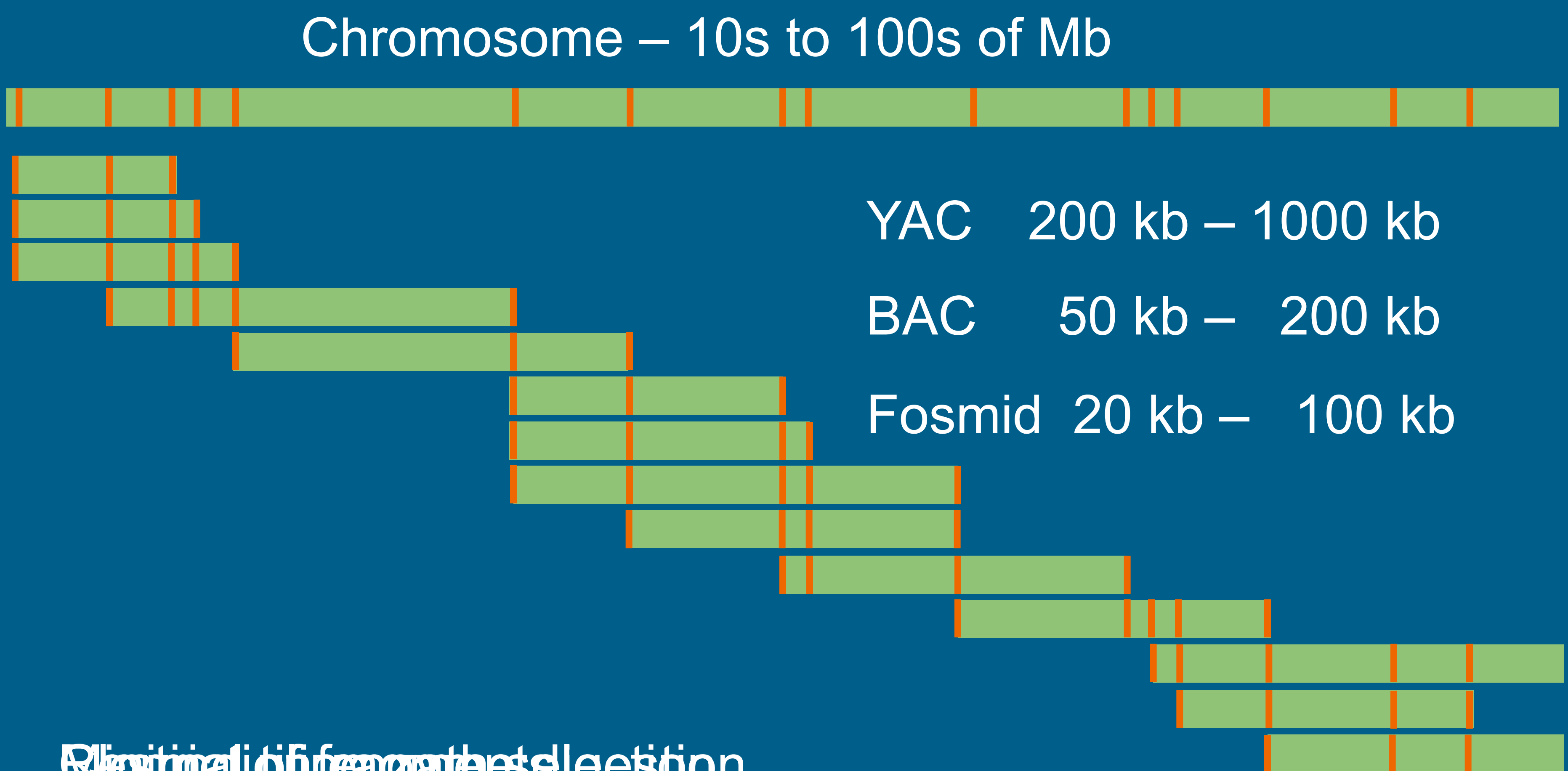
Clone-based sequencing (1)

- Construction of clone libraries
  - Restriction enzyme digestion or random shearing
  - Ligation into cloning vector (BAC, YAC, fosmid)
  - Cloning vector is propagated in host (e.g. *E. coli*)
  - Each host colony contains a unique insert
  - Each library represents several genome equivalents (redundant)
- Some clones can not be propagated in the host
  - Toxicity to the host
  - Secondary sequence structures

Clone-based sequencing (2)

- Anchoring of clones to a genetic map
  - Identification of genetic markers in clones (e.g. through hybridization)
  - Clones are fingerprinted and clustered into groups based on overlapping fingerprints (physical map)
- Clone-by-clone sequencing of a minimal tiling path
  - Reduces the amount of redundant sequence

Clone-based sequencing (3)



Partial digestion
Cloning of fragments
Vector ligation

Shotgun sequencing

- Applied to whole genomes and selected clones
- Shotgun sequencing
  - Several size equivalents of DNA is sheared into small fragments (0.5 – 2 kb)
  - Each fragment is sequenced from one end
- ‘Double-barreled’ shotgun sequencing
  - DNA is sheared into larger fragments (2 kb – 20 kb)
  - Each fragment is sequenced from both ends (‘matepairs’)

Intermezzo

- Go to http://www.google.com/
- Fill in your favorite food species
  - animal, vegetable, fruit, nut
- Add “genome sequence”

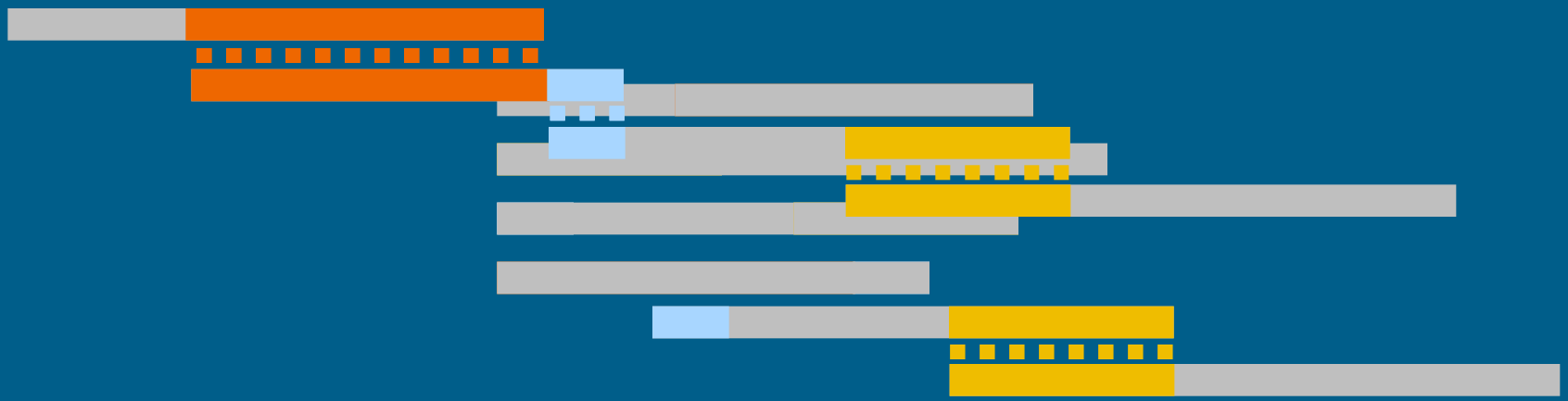
- How many of these genomes are being sequenced?

- http://www.genomesonline.org/gold.cgi

Sequence assembly

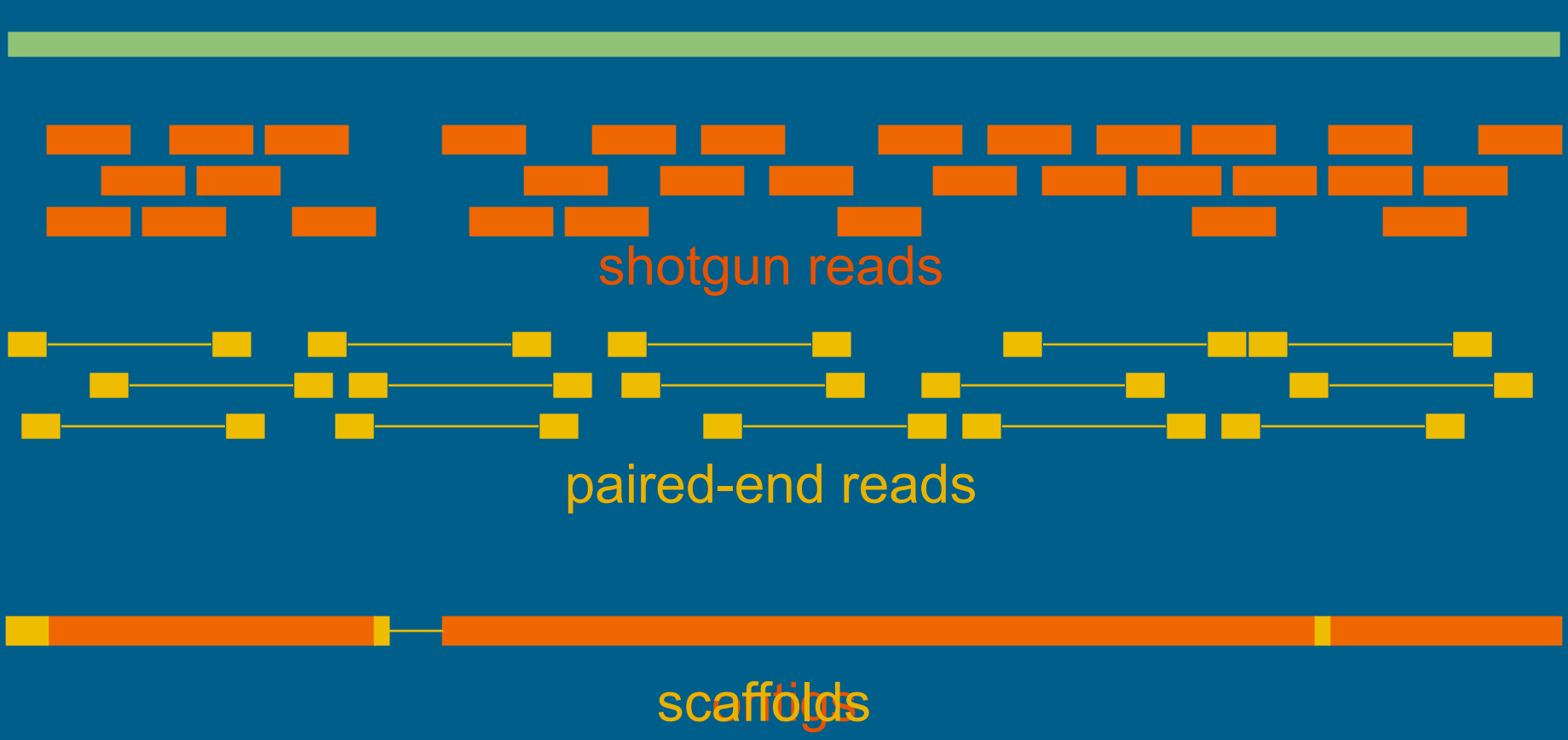
- Reconstructing the complete sequence of a DNA molecule from short sequence fragments ("reads")
- Overlap-layout-extend
  - All-vs-all sequence similarity comparison
  - Alignment of sequences with high sequence identity
- Graph-based assembly
  - Construction of *k*-mer graph
  - Extraction of linear paths from the graph

Sequence assembly: overlap-layout-extend (1)



Align reads to each other with high sequence identity

Sequence assembly: layout-overlap-extend (2)

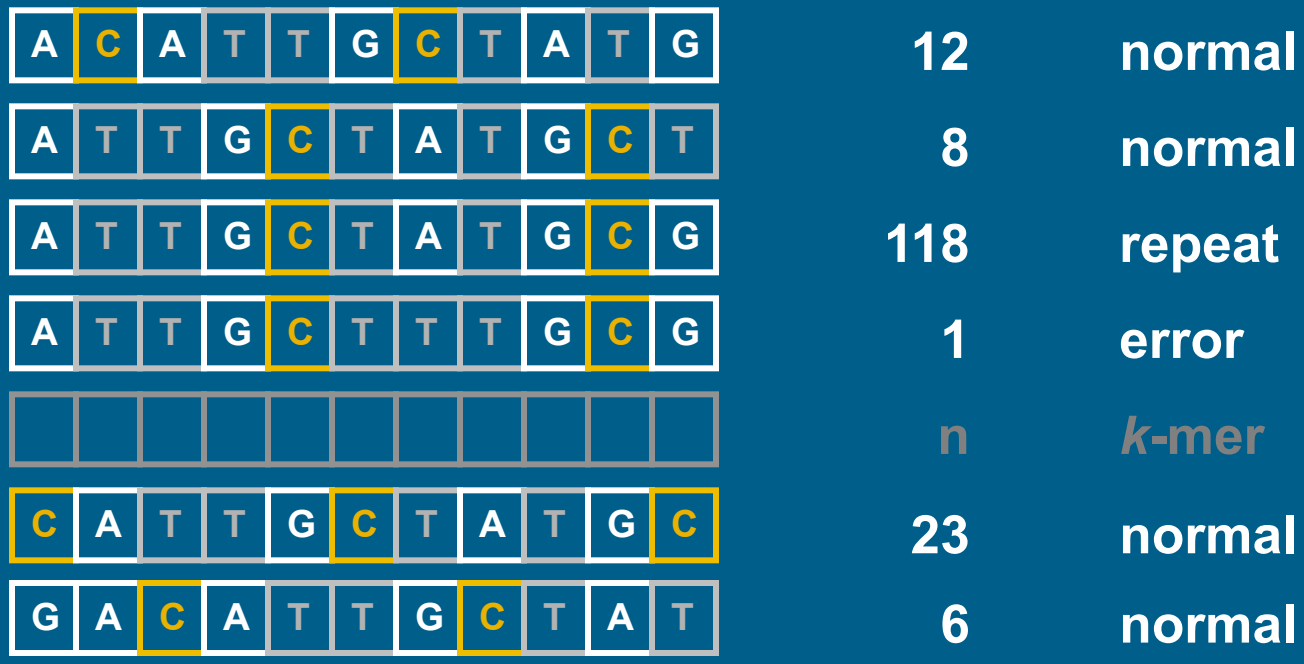


Graph-based sequence assembly

- Short read length and large data volume
  - All-vs-all comparison becomes impractical
  - Many mutually inconsistent overlaps
- Solving the assembly problem with De Bruijn graphs
  - Nodes are *k*-mers
  - Edges are *k*-mers that overlap on *k*-1 positions

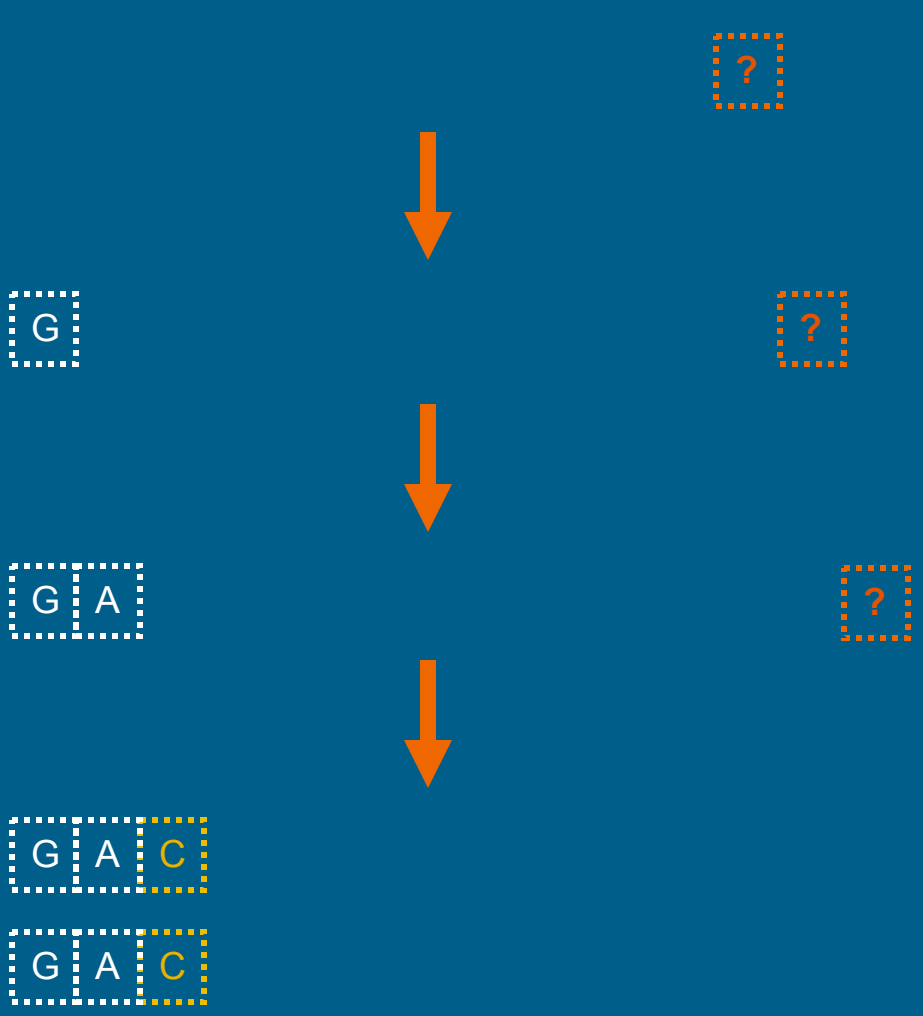
Generating *k*-mers from reads

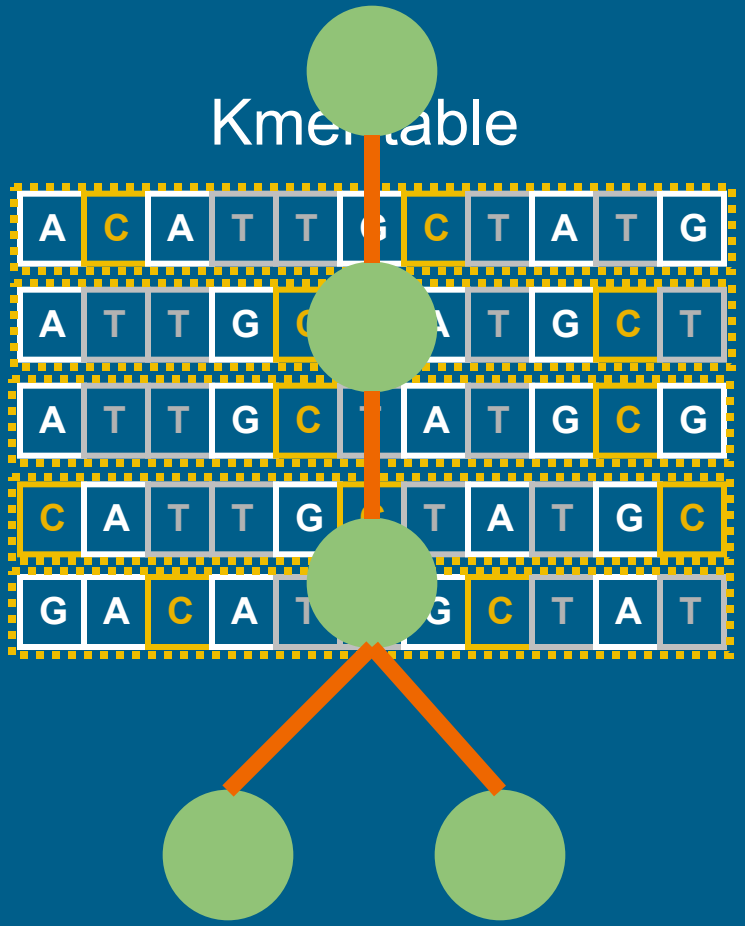
Creating a *k*-mer table



n	*k*-mer
12	normal
8	normal
118	repeat
1	error
23	normal
6	normal

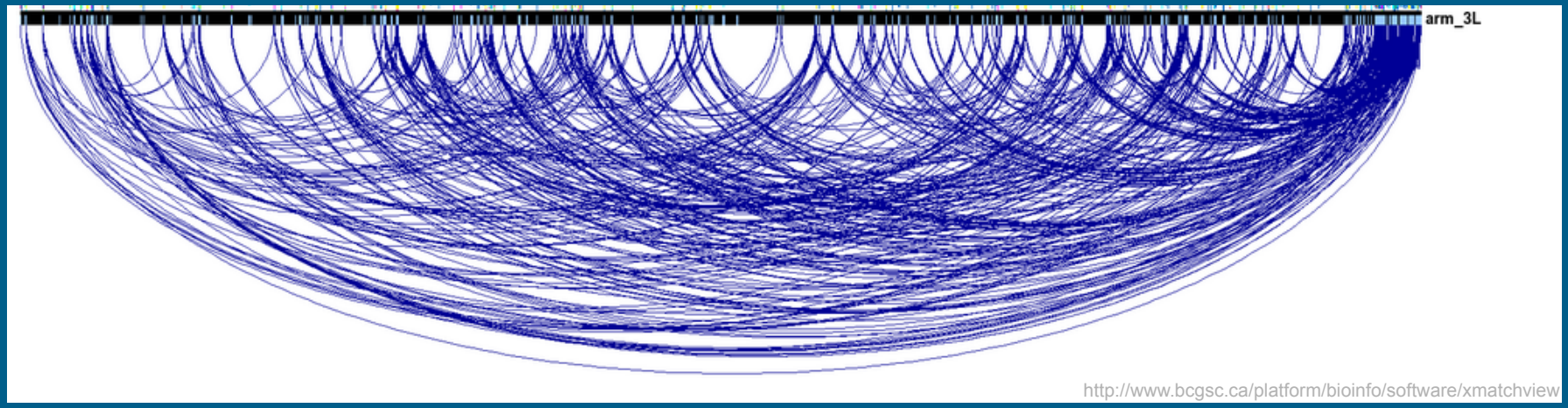
Graph-based assembly of *k*-mers





The largest challenge in genome assembly (1)

- Repeats
  - Multiple overlaps in overlap-layout-extend approach
  - Multiply connected nodes in graph-based approach



Repeats in *D. melanogaster* chromosome 3L

The largest challenge in genome assembly (2)

Table 2. Repeat Content in Eukaryotic Genomes[a]

species	sequencing method	TR-derived repeats (fraction of genome %)				segmental duplications (fraction %)
		LINE	SINE	DNA transposon	LTR retrotransposon	
C. elegans	hierarchical	0.3	0.1	5.3	<0.1	?
C. briggsae	WGS				22.4[b]	
D. discoideum[69]	hierarchical		4.6	1.4	4.4	?
Fruit fly[70]	Hybrid	0.5	0.3	1.5	2.6	?
Silkworm	WGS	6.7	1.4	1.7	11.1	?
Mosquito	WGS	0.2	3.5	1.1	11.2	?
Fugu fish	WGS	1.2	<0.1	0.5	0.8	?
Puffer fish	WGS	<0.1	0	<0.1	<0.1	?
Chicken	WGS	6.5	<0.1	0.8	1.3	~2.8
Dog	WGS	5.4	6.7	0.13	0.17	~0.6
Human	hierarchical	20.4	13.1	2.8	8.3	3.2
Mouse	WGS	19.2	8.2	0.9	9.9	1.2
Rat	Hybrid	23.1	7.1	0.8	9.0	2.9
Chimpanzee	WGS	23.1	7.0	0.8	9.0	2.5
Mustard weed	hierarchical	0.5	0.5	5.1	4.8	58[d]
Rice[c]	WGS	1.19	0.09	2.8	9.3	?
Rice[c]	hierarchical	1.12	0.06	13.0	18.1	60[d]

The largest challenge in genome assembly (3)

- Repeats can be resolved by longer reads
  - If the read is larger than the repeat, it poses no problem
  - Many repeats are several to tens of kbs long …
- Repeats can be circumvented by using matepairs
  - If the matepair is larger than the repeat, we can connect both sides of the repeat to each other
  - If one end of the matepair is unique, it allows us to position repeat types/families inside scaffolds

Intermezzo

- Which sequencing (and corresponding assembly) approach would you take to sequence your favorite genome?
- Why?
- (There are no wrong answers!)

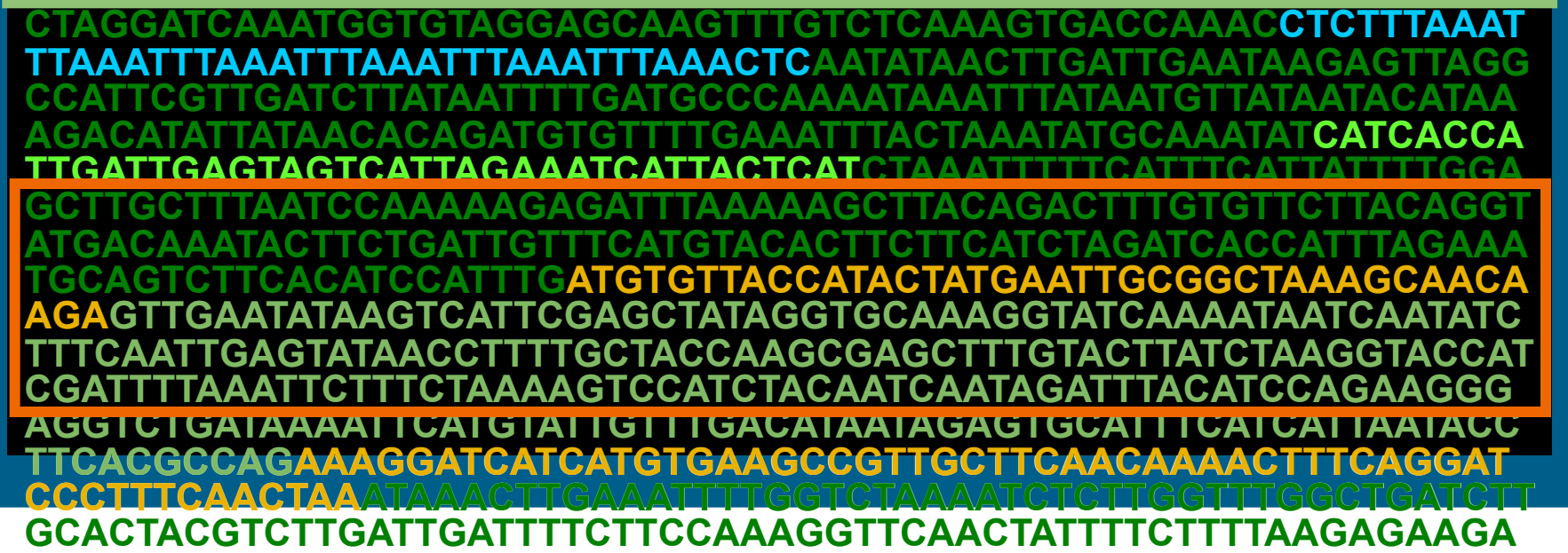
Computational genome annotation

Gene: SL6G63120.1

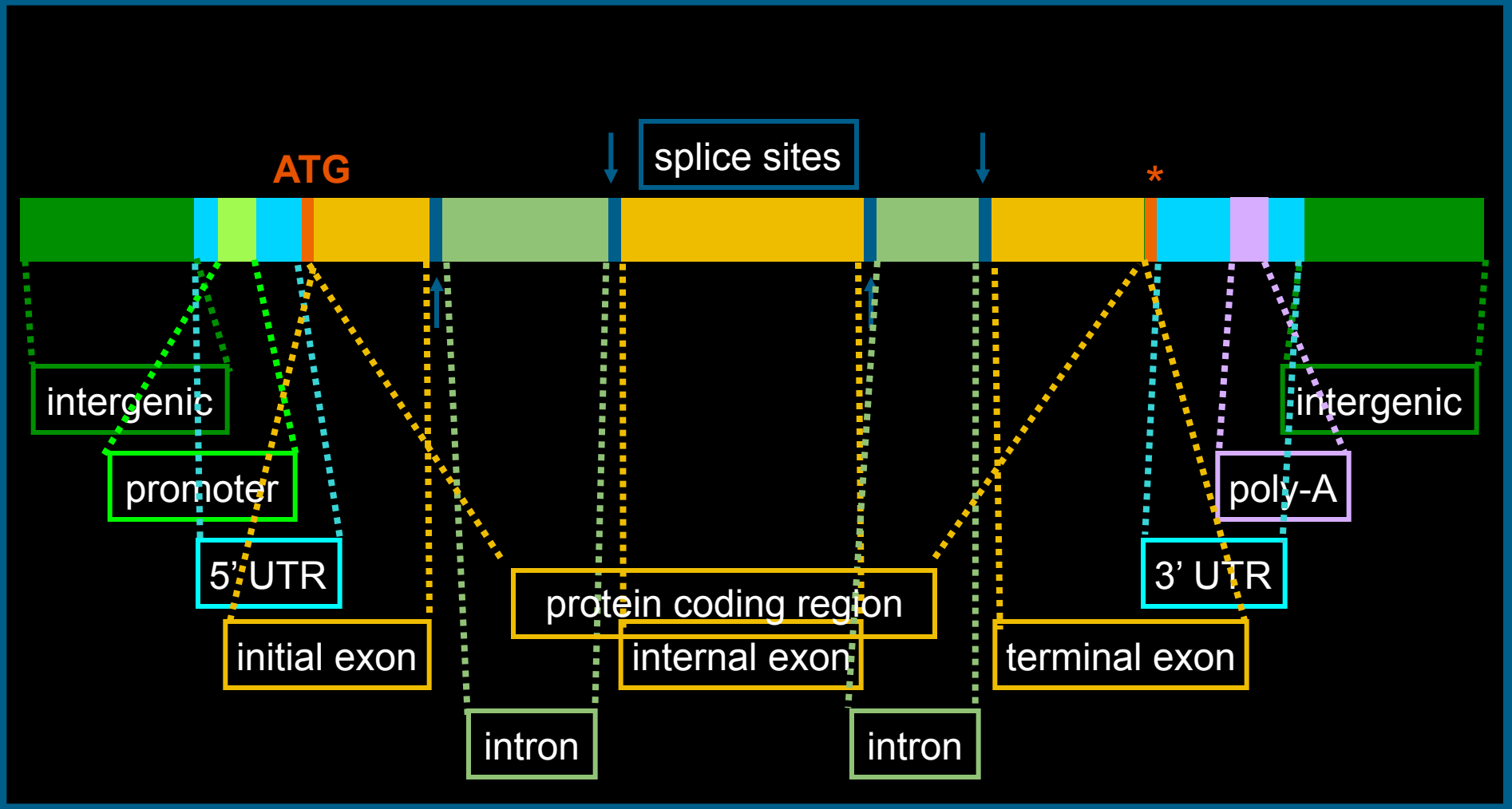
Description: Putative disease resistance gene, Mi-homolog

Domains: NBS, LRR

Best Blast hit: SA004581

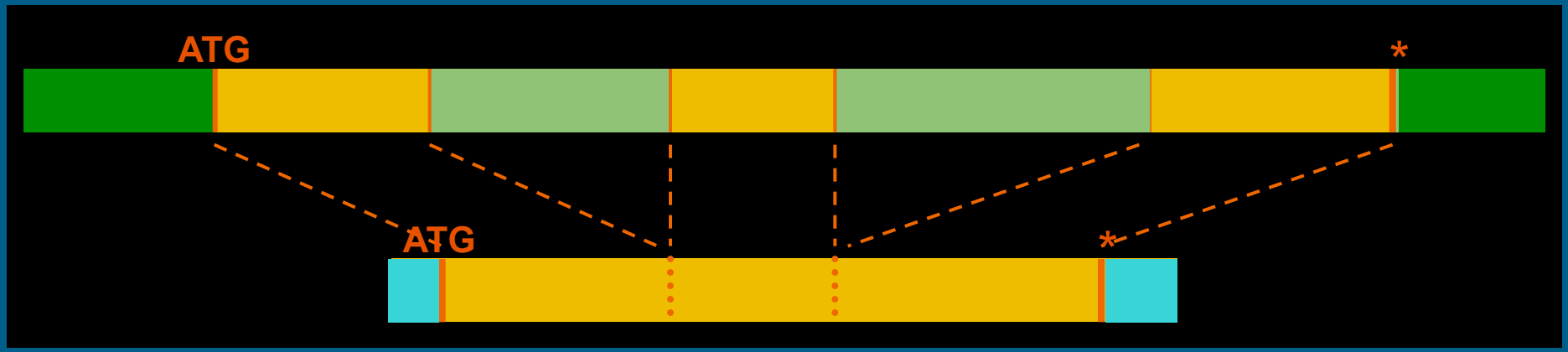


Gene prediction – the eukaryotic gene model



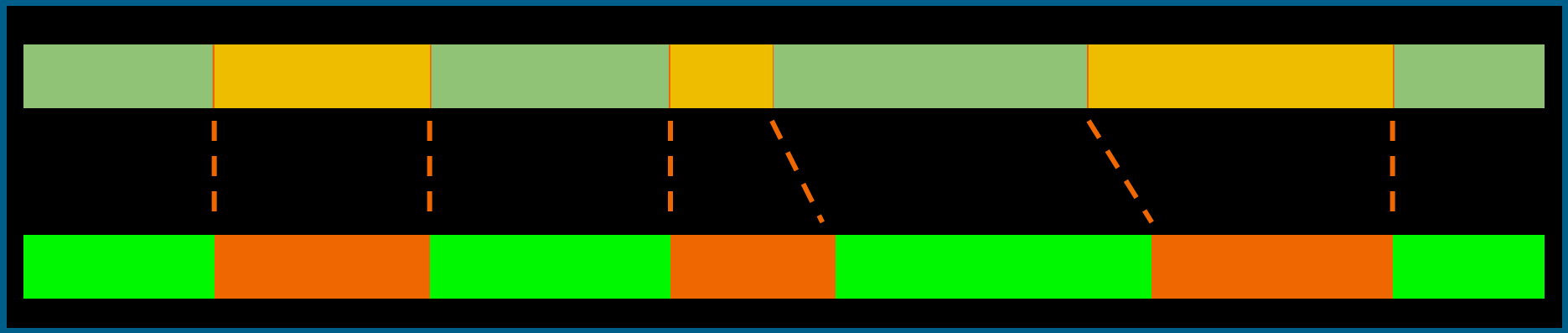
Structural gene annotation – alignment–based (1)

- Prediction of gene structures based on alignments
  - Transcripts and proteins provide direct evidence
  - Requires experimental data for each gene

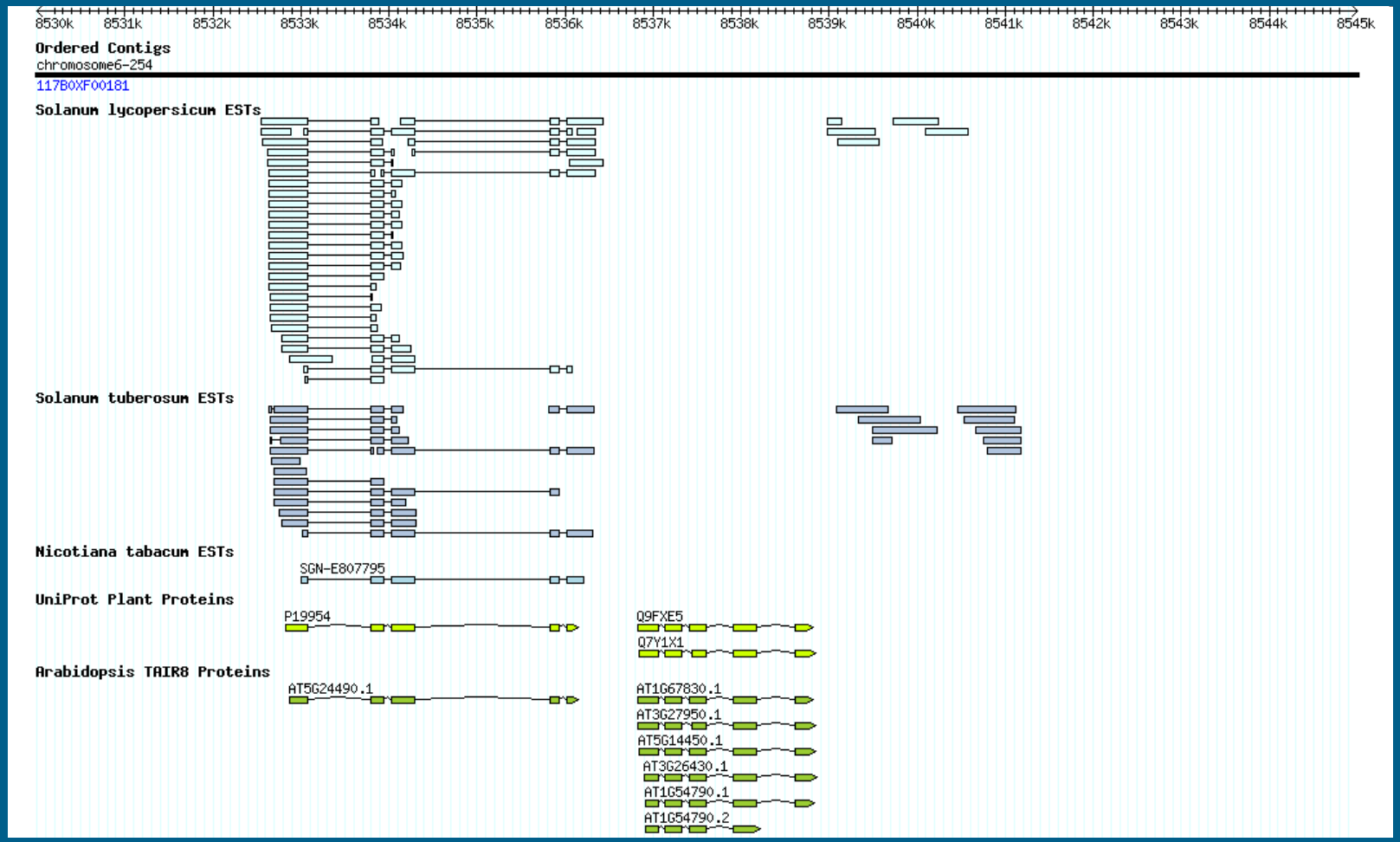


Structural gene annotation – alignment–based (2)

- Genome-to-genome alignment
  - Requires **annotated** genome of closely related species

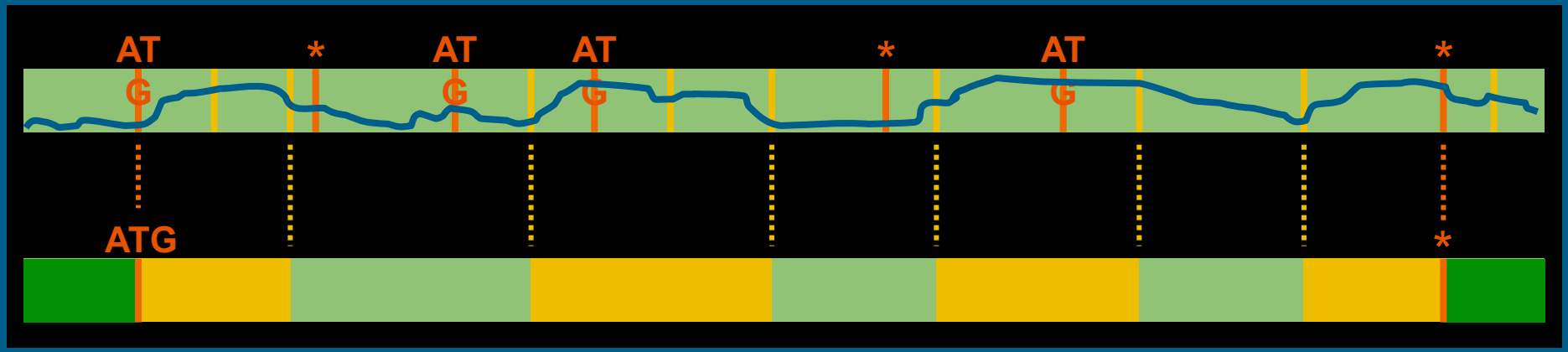


Example – tomato genome browser



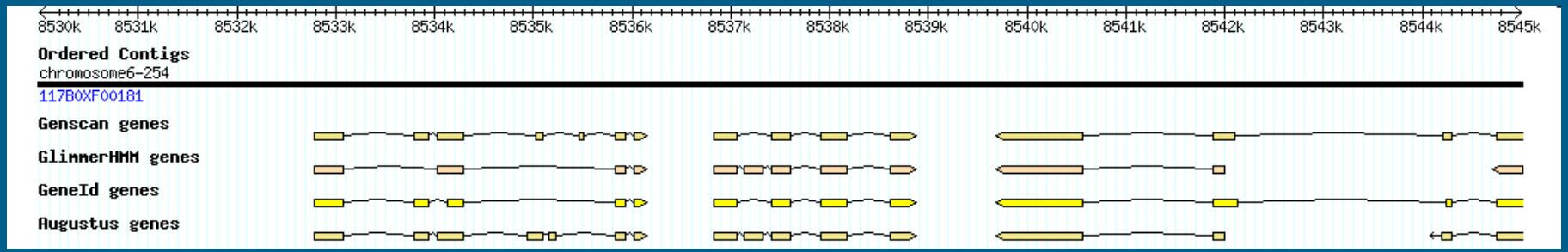
Structural gene annotation – *ab initio* (1)

- Prediction of gene structures based on ‘gene model’
  - Start, stop, splice sites
  - Exon, intron, intergenic length distributions
  - Triplet/hexamer frequencies (coding vs. non-coding)



Structural gene annotation – *ab initio* (2)

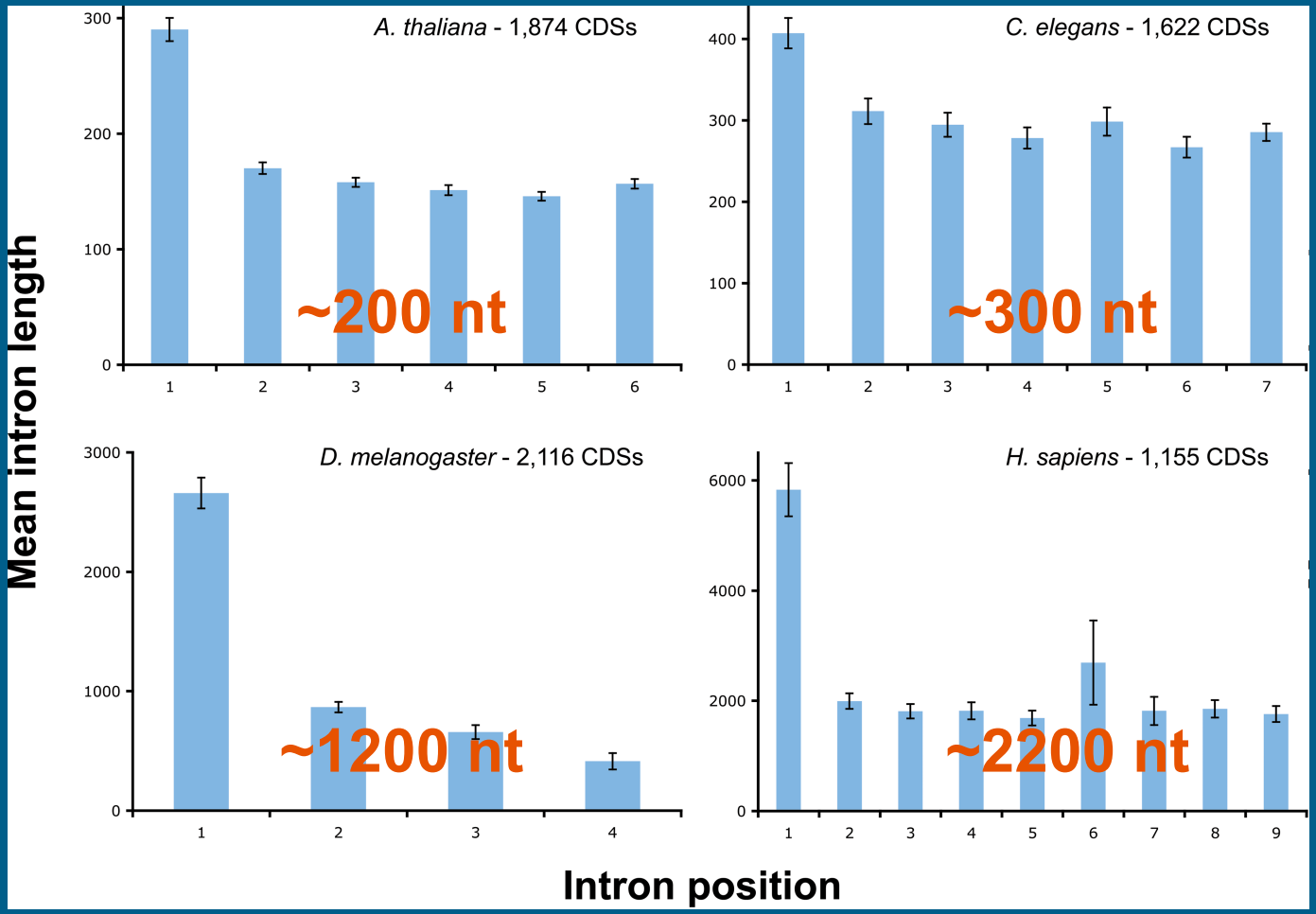
- Different predictors produce different results
  - Underlying models (HMM, SVM, …)
  - Quality of training
  - Lack of understanding of biology



Structural gene annotation – *ab initio* (3)

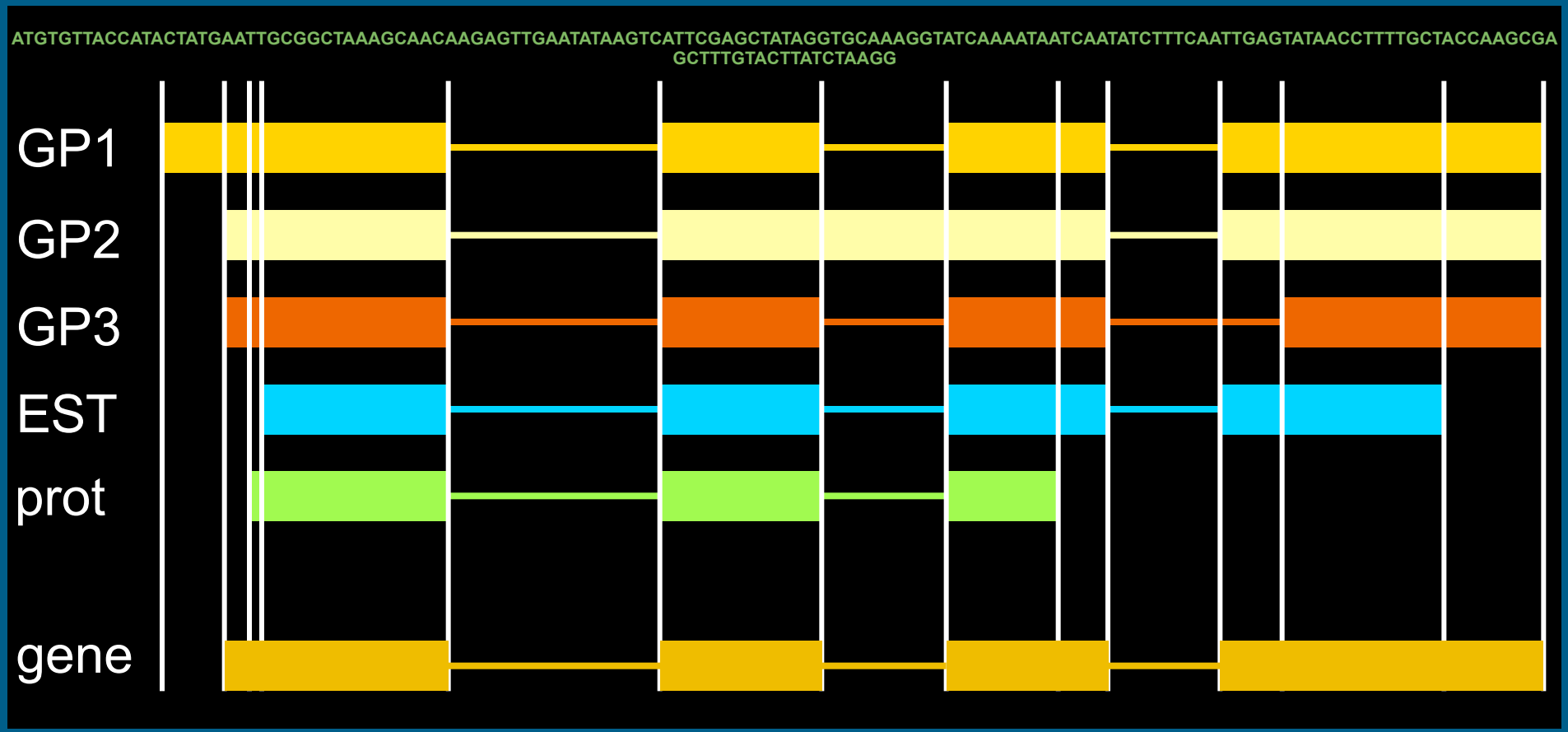
- Requirements for a*b initio* gene predictors
  - Training through verified transcript (and protein) alignments
  - Sufficient sequence context in order to make accurate predictions
- Some properties are common for all eukaryotes
  - Start, stop, splice site consensus
- Many properties differ, even between related species
  - Intron and intergenic length distribution
  - Codon usage
  - Additional splicing signals (e.g. branchpoints)

Example - differences in intron lengths

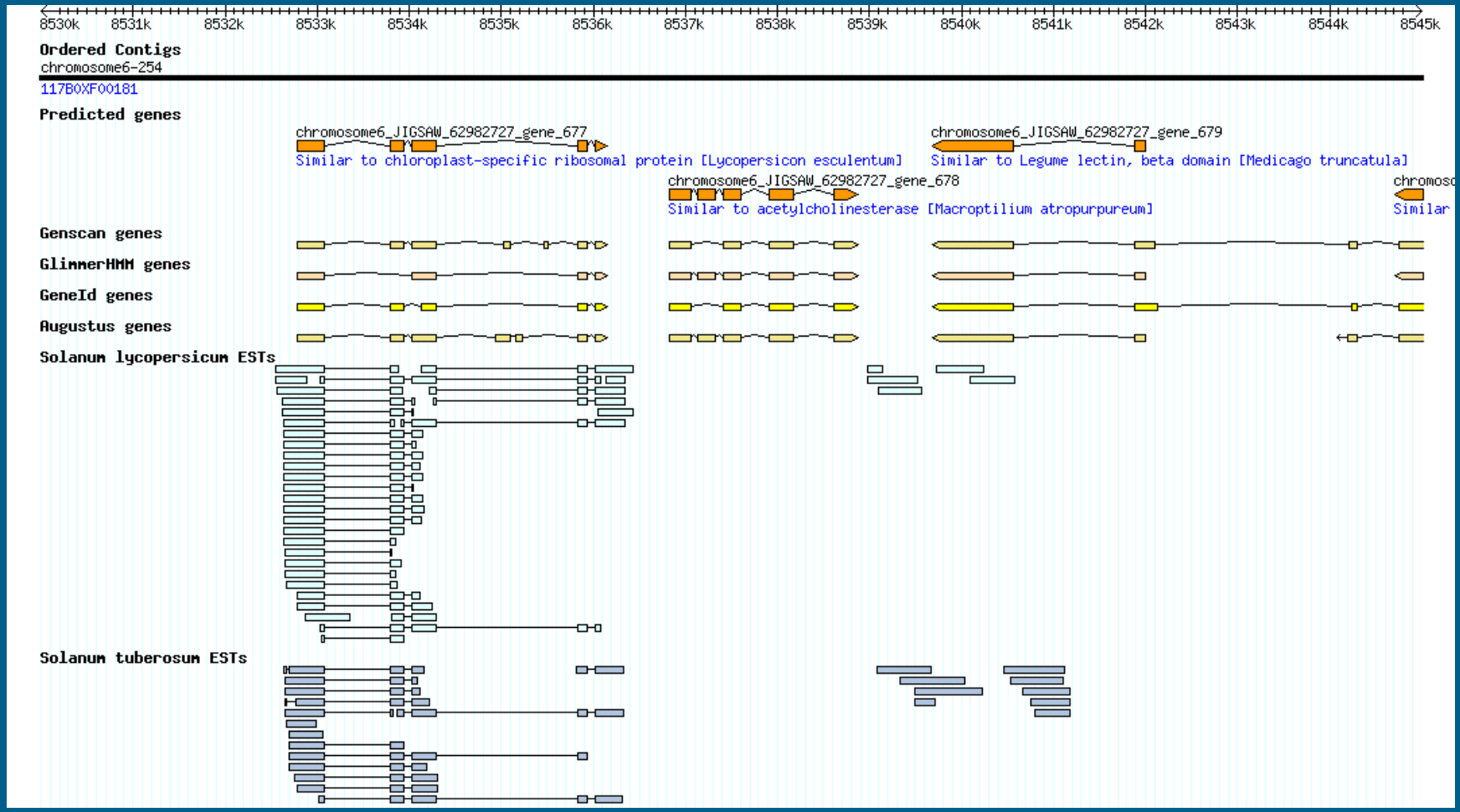


Bradnam and Korf, PLoS One. 2008

Generation of a consensus gene structure



Example – tomato genome browser



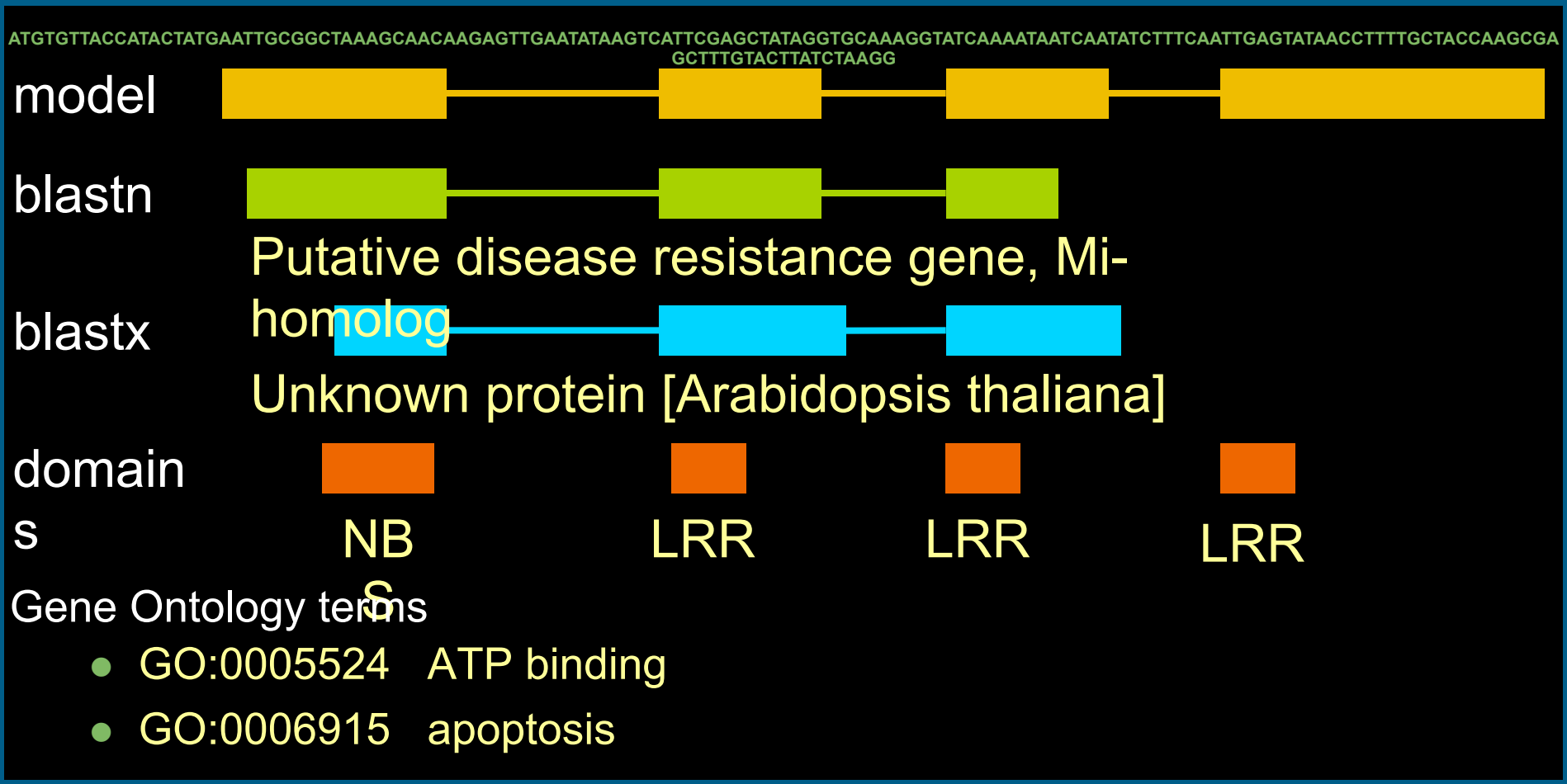
Functional gene annotation – alignment–based

- Inferring function through sequence similarity
  - Proteins with similar sequence often share function
- Annotation quality of database sequences
  - Many proteins with unknown function
  - Propagation of erroneous annotation

Functional gene annotation – domain–based

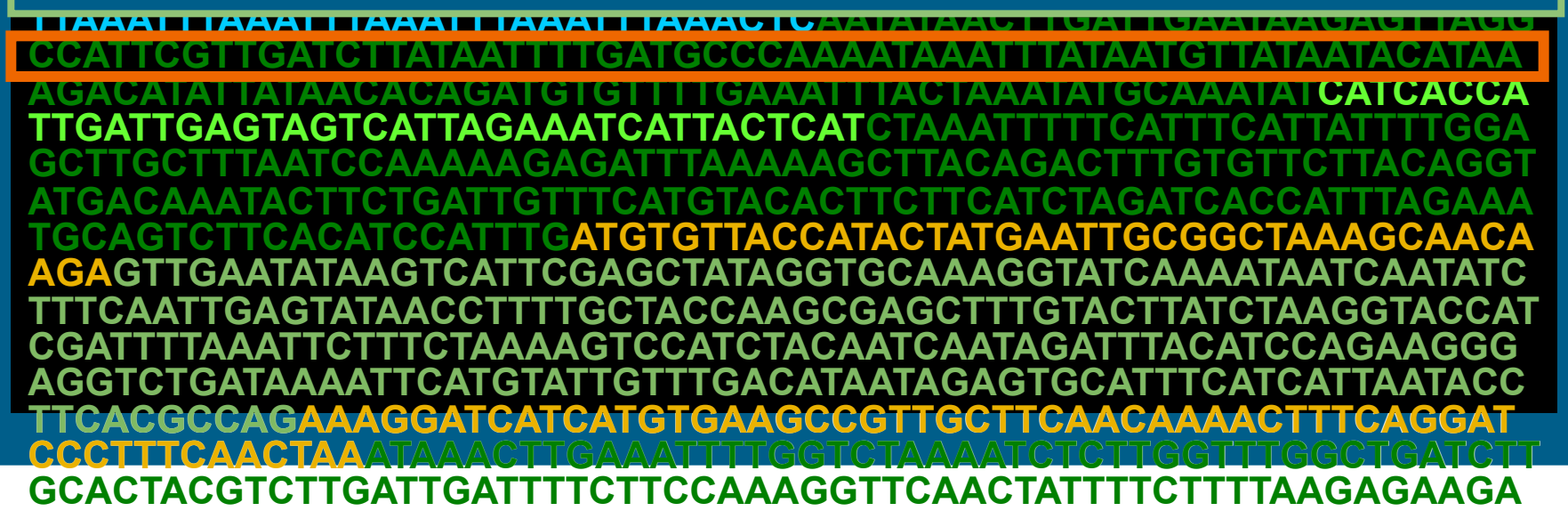
- Inferring function through domain searches
  - Domains are the functional parts of a protein
- Global functional annotation of the protein
  - E.g. kinase, ATP-binding
  - Gene Ontology (GO) terms

The annotated gene



Genome annotation: more than gene finding

- Repetitive sequences
  - Interspersed repeats (e.g. transposons)
  - Tandem repeats, SSRs
  - miRNAs
  - ...

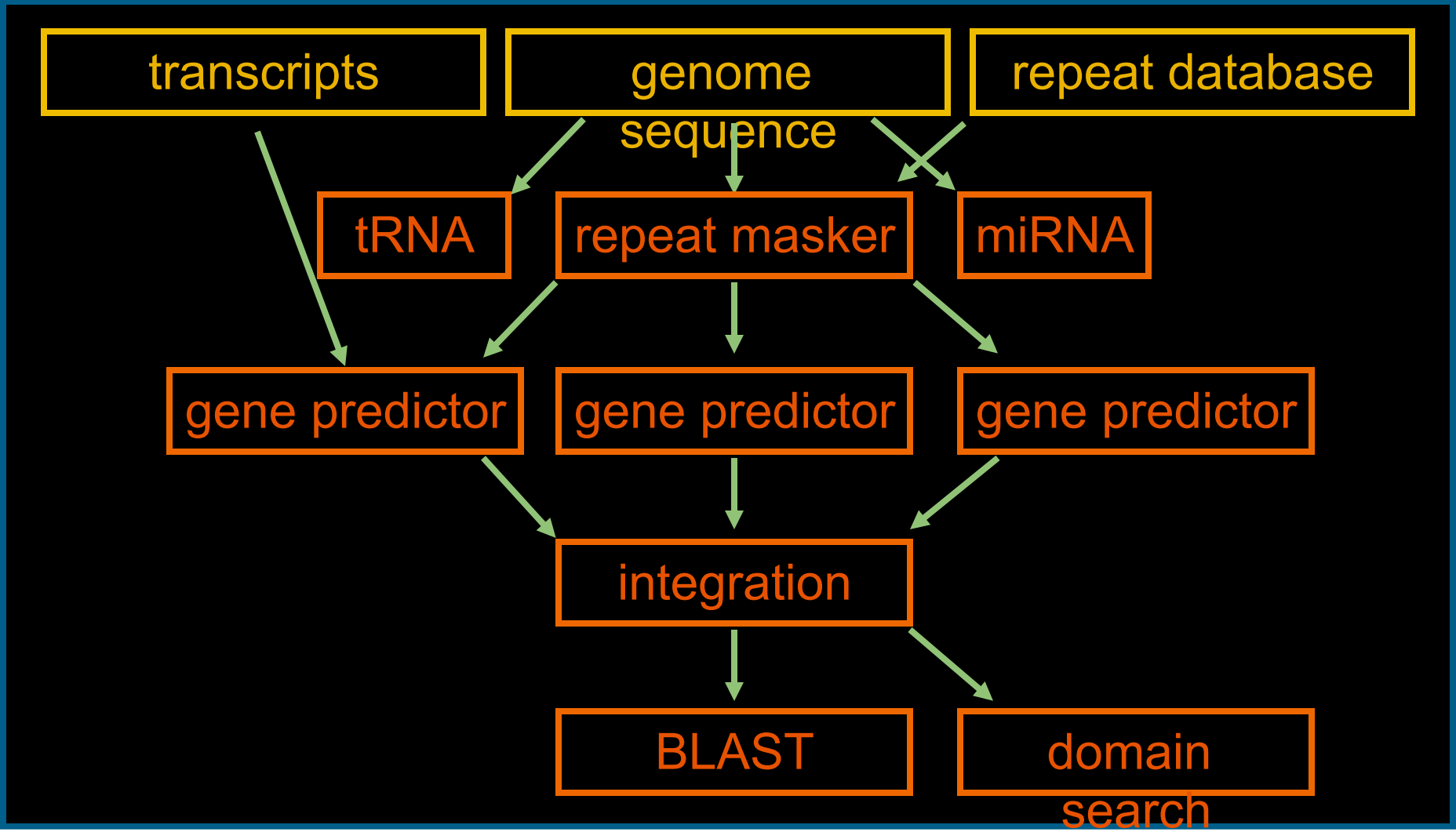


Repeat identification and masking

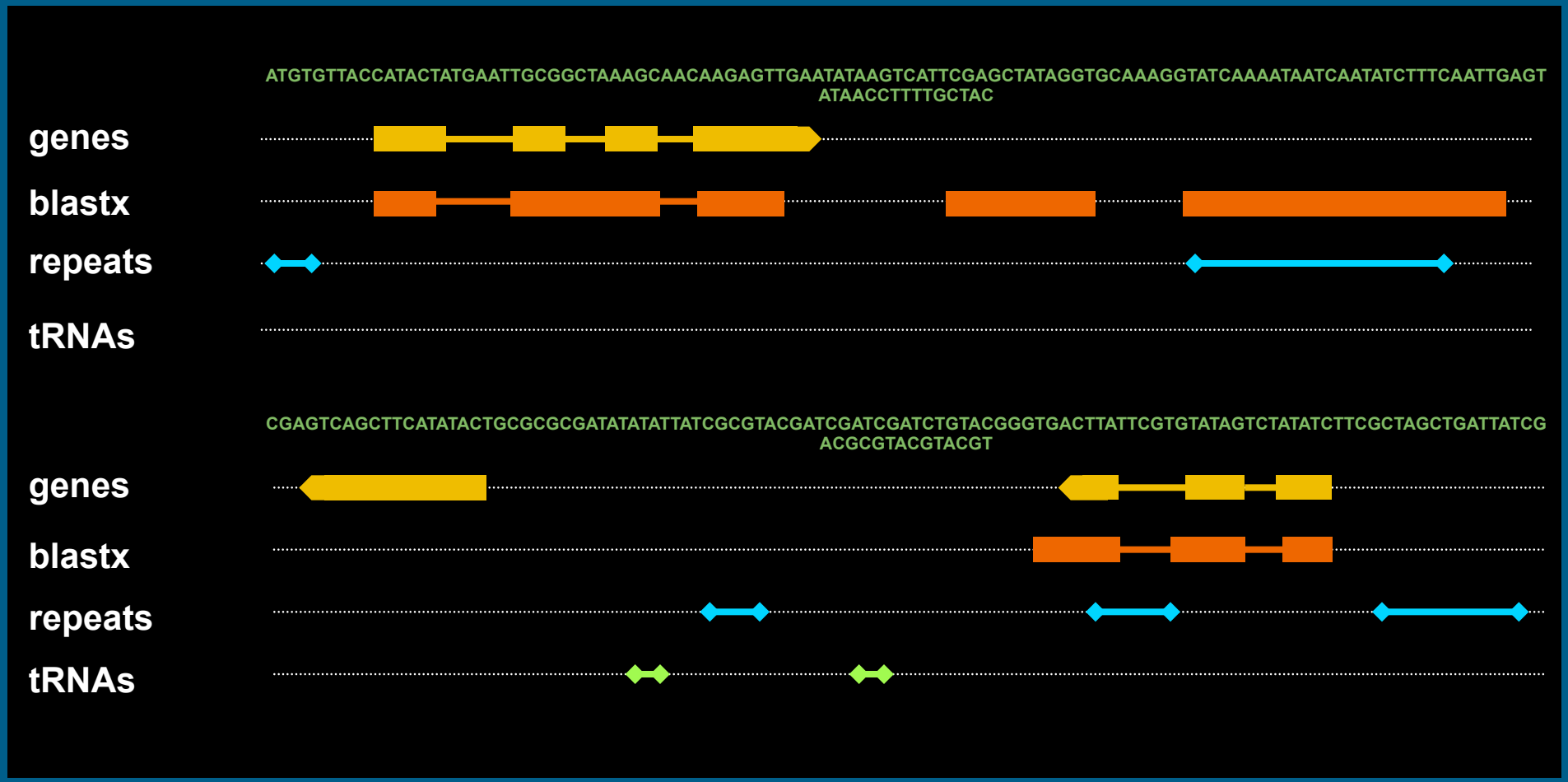
- Repeats containing coding elements (e.g., reverse transcriptase and polymerase in a retrotransposon) result in many ‘false’ gene predictions
  - 56,797 genes predicted in rice
  - 16,220 of these are repeat-related!

- Prior to gene prediction, repeats should be masked
  - Requires database of known repeats
  - Sequence similarity
  - De novo

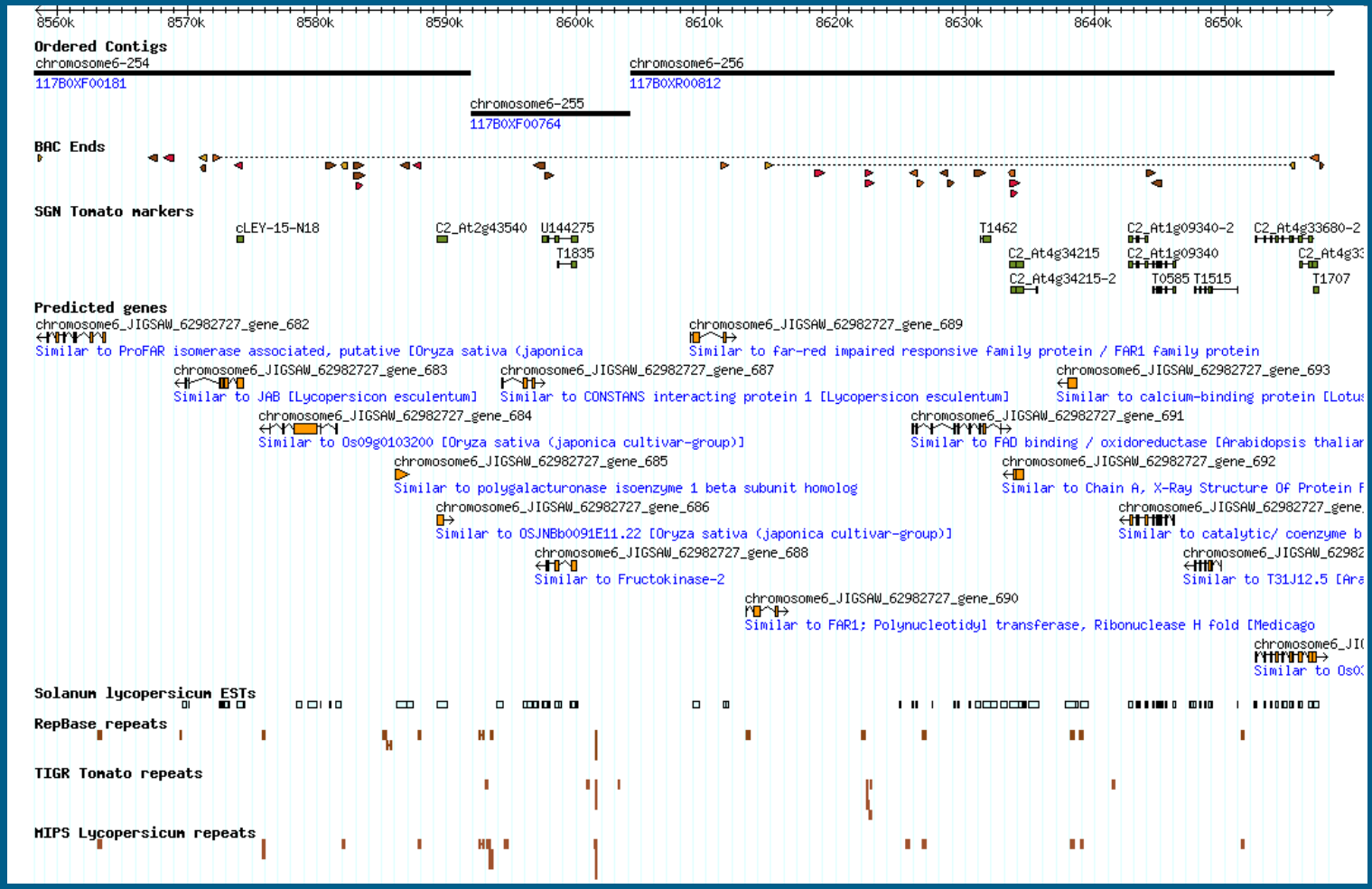
Genome annotation pipeline



The annotated genome



Example – tomato genome browser



Tomato genome structure

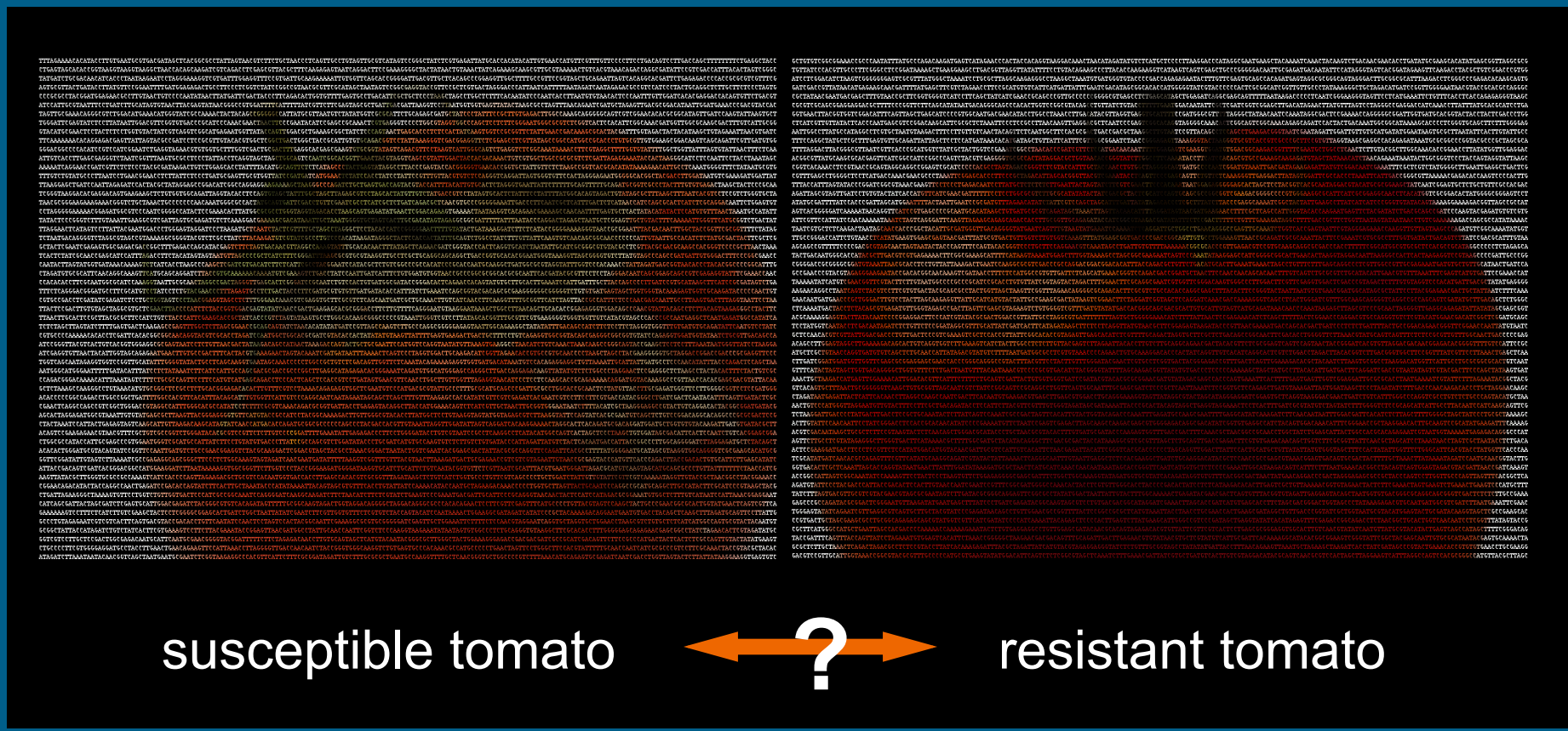
Peters, Datema *et al*, Plant J. 2009

Beyond genome annotation

- Automated annotation can provide candidate genes
  - Similarity to known genes from other species
  - Targets for crop improvement, treatment of (genetic) diseases, etc.

- Comparative genomics
  - Study the evolution of species
  - What makes a species unique?
  - What makes an individual unique?

Back to biology…



susceptible tomato ⟷ ? ⟷ resistant tomato

Introduction to exercise (1)

- We identified a marker in our tomato population that segregates with the resistance phenotype
  - The resistance gene(s) are probably located close this this marker on the genome!
- We identified a BAC clone that hybridizes to this marker
  - This BAC could contain the gene(s) that confer the resistance!
- We sequenced the BAC…
  - … and now we would like to identify the candidate gene(s)!

Introduction to exercise (2)

- The BAC was sequencing using the double-barreled shotgun approach on a Sanger platform
  - Fragments of ~2 kb
  - Matepair reads of ~700 nt each
- For each read there are two types of data
  - Nucleotides
  - Quality values: these provide an estimate of the chance that the nucleotide was called incorrectly

Exercise

- Assemble the BAC using the Celera Assembler
  - How many scaffolds do you end up with, and what is their total length?
  - Optional: how many reads are there originally, and how long are they in total? What is the average read coverage of the BAC?
  - Optional: try some different ‘error rates’ in each of the assembly steps to see how these influence the assembly
  - Optional: try a different assembler, such as CAP3
- Help us find our candidate gene(s) in the BAC
  - How many genes can you find in the BAC sequence?
  - Do any of them have interesting functional annotations?
  - Optional: try and compare three different predictors
  - Optional: try to mask for repeats before predicting the genes

Assemble the BAC using the Celera Assembler

- http://sourceforge.net/apps/mediawiki/wgs-assembler/

- CA wants .frg files, we have .fasta and .qual
  - Convert the files using **convert-fasta-to-v2.pl**
  - Don't forget to remove any vector contamination…

- CA has a LOT of parameters!
  - The defaults should be fine for assembling our data

Find and annotate the genes

- There are many different gene finders for plants
  - genscan, geneid, fgenesh, augustus, …
  - Most predictors have several training sets
  - Tomato (*Solanum lycopersicum*) is a dicot plant!
- There are many ways to annotate genes (proteins)
  - BLAST (sequence similarity searches)
  - InterProScan (domain searches)
- Many of these tools can be run on the web!

F1!!! F1!!!

- Converting the files
  - http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Formatting_Inputs
  - convert-fasta-to-v2.pl -l H019E05 -s H019E05.fasta -q H019E05.qual -v H019E05.vector > H019E05.frg
- Running the assembler
  - http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=RunCA
  - runCA -d assembly -p H019E05 H019E05.frg
- Finding the output
  - Look for the .scf.fasta file in your assembly directory
- Finding the genes
  - genscan http://genes.mit.edu/GENSCAN.html
  - geneid http://genome.crg.es/software/geneid/geneid.html
  - fgenesh http://mendel.cs.rhul.ac.uk/mendel.php?topic=fgen
  - augustus http://augustus.gobics.de/submission

F1!!! F1!!! (optionals)

- Counting the number of reads and the total length
  - grep ">" H019E05.fasta | wc
  - wc H019E05.fasta
- Configuring the allowed error rates in CA
  - utgErrorRate, ovlErrorRate, cnsErrorRate, cgwErrorRate
  - Error rates must obey the relationship utg ≤ ovl ≤ cns ≤ cgw
- CAP3
  - http://seq.cs.iastate.edu/cap3.html
- RepeatMasker
  - http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker

The End

PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR