Module 7
Genome Annotation

Introduction

One of the key goals of producing a draft genome is to define the genes encoded in it. This is the first step in answering many important questions. How many genes does this organism have? What metabolic pathways might this organism have? Are there novel gene families encoded in the genome? Subsequent uses of the genome, such as proteomics and transcriptomics experiments rely on accurate gene models. We concentrate here on protein coding genes, however one would also try to identify non-protein coding RNAs such as tRNAs, rRNAs, miRNAs, transposons etc.

Producing accurate **gene models** is just as hard as producing a good assembly. You have seen that one way of producing a set of gene models is to transfer them from a closely related organism. However, if there is not a closely related genome, or the most closely related genome is not well annotated, this may not be an option. Furthermore, even if there is a closely related, well annotated reference genome as in the case of the malaria parasite - *Plasmodium falciparum* strain 3D7 and *P. falciparum* strain IT, there may be regions of your genome of interest which are not syntenic to the reference. Indeed, this is the case here, as the subtelomeric regions of *Plasmodium* chromosomes are highly variable and cannot be used to transfer gene models, even between strains of the same species!

When there is no reference genome, or for regions which are not syntenic to the reference, we can use ***ab initio* gene finding** methods. These identify genes based on properties of the genome sequence independent of whether they show homology to known genes. They can be trained and it is common to identify the most well conserved genes by homology, then to train an *ab inito* gene finding algorithm using these well conserved genes so that it can learn what a gene looks like in the particular genome you are interested in. In this module we will use the program **Augustus** to predict gene models *ab initio*.

Another approach to identifying gene models is to determine those regions of the genome which are transcribed. Prior to the advent of second generation sequencing technologies Sanger capillary sequencing was used to sequence mRNAs and generate Expressed Sequence Tags (ESTs). These could be used to identify the most highly expressed genes and improve some gene models. With second generation sequencing technologies we are able to sequence mRNA transcripts from essentially all the genes which are expressed. This is known as **RNA sequencing** or RNA-seq (Mortazavi et al., 2008; Wang et al, 2009) and the resulting data provides incredible resolution of gene structure. This method is not biased by which genes are present in previously sequenced genomes (as with RATT) or by how much their structure reflects that of other gene in the genome (as with Augustus), but rather by how highly expressed they are and how extensively we sequence the transcriptome.

In this module you will generate a set of gene models *ab initio* using the gene prediction tool Augustus. It has been trained using the highly accurate, manually curated gene models of *P. falciparum* 3D7. These gene models will be used to fill in those regions of the *P. falciparum* IT assembly which could not be annotated using RATT because they are not syntenic to *P. falciparum* 3D7. You will then map RNA-seq data to your assembly and use this to improve the gene models. There will be inaccuracies from the RATT transfer due to technical error and due to real differences in the gene structures. There will be inaccuracies in the Augustus predictions due to gene models which do not follow the expected pattern of a *Plasmodium* gene and due to inaccuracies in the genome assembly. These can be addressed using the RNA-seq mapping.

Below is a model of the eukaryotic protein-coding gene highlighting features relevant to their annotation. Note that compared to bacteria, eukaryotic genes frequently have multiple exons, separated by un-translated introns and 5` and 3` Un-Translated Regions (UTRs) which do not encode part of the protein sequence.

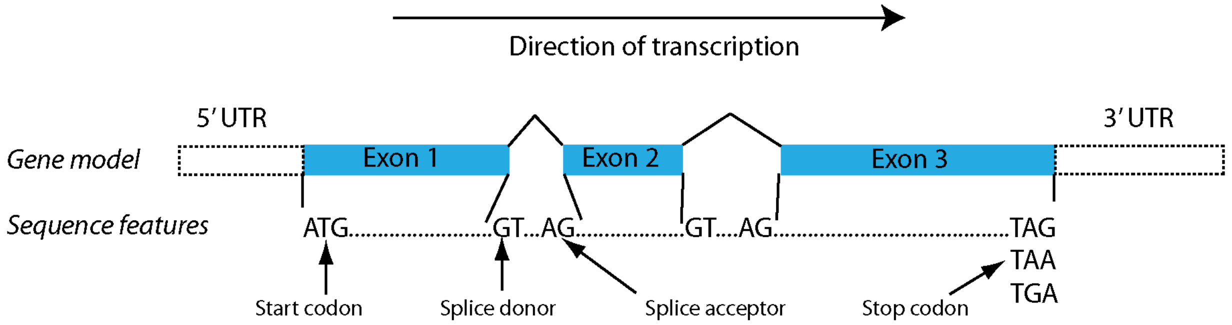

Direction of transcription
5′UTR
3′UTR
Gene model
Exon 1
Exon 2
Exon 3
Sequence features
ATG........................GT...AG..................GT....AG...TAG
TAA
TGA
Start codon
Splice donor
Splice acceptor
Stop codon

Figure showing the key features of a eukaryotic gene. The exact DNA sequence of the splice sites may vary between species.

Module Summary

1. Generating an initial set of gene models (merging RATT and Augustus)
2. Mapping RNA-seq data to a reference
3. Viewing RNA-seq mapping in Artemis
4. Correcting gene models by hand
5. Automatically generating gene models based on RNA-seq data
6. Using RNA-Seq to improve annotation

1. Generating an initial set of gene models

A. Generate gene models *ab initio*

The *ab initio* gene prediction algorithm Augustus has already been trained with gene models from *Plasmodium falciparum* 3D7. You will now run it on your *Plasmodium* IT strain chromosome to predict a set of gene models.

Navigate to the Module 7 data directory. On the command line, type:

```
augustus --species=3D7 IT.genome.fa > augustus.gtf &
```

The file `augustus.gtf` now contains your predicted gene models

Next we are going to convert the Augustus GFF to EMBL format. On the command line, type:

```
cat augustus.gtf | augustus2embl.pl > augustus.embl
```

Although Artemis can display gff files, the visualization is better for embl files (if you want you can also try loading the gff file into Artemis).

It is typically much more challenging to train an *ab initio* gene finder than to run it. Most *ab initio* gene finders require a set of very accurate models to train it, which may be predicted from highly conserved genes, RNA-Seq and typically manual curation of a few hundred gene models. The chosen training set can influence which types of gene models get better or worse predictions, so should be carefully chosen.

Tip: skip to page 9; "Mapping RNA-Seq data to a reference", and start the tophat mapping (the first blue box) before continuing to page 4, because the mapping might take several minutes.

B. Examine gene model predictions

First we will open some RATT-transferred gene models from *P. falciparum* 3D7 in Artemis and compare them to the gene models predicted *de novo* by Augustus.

On the command line, type:

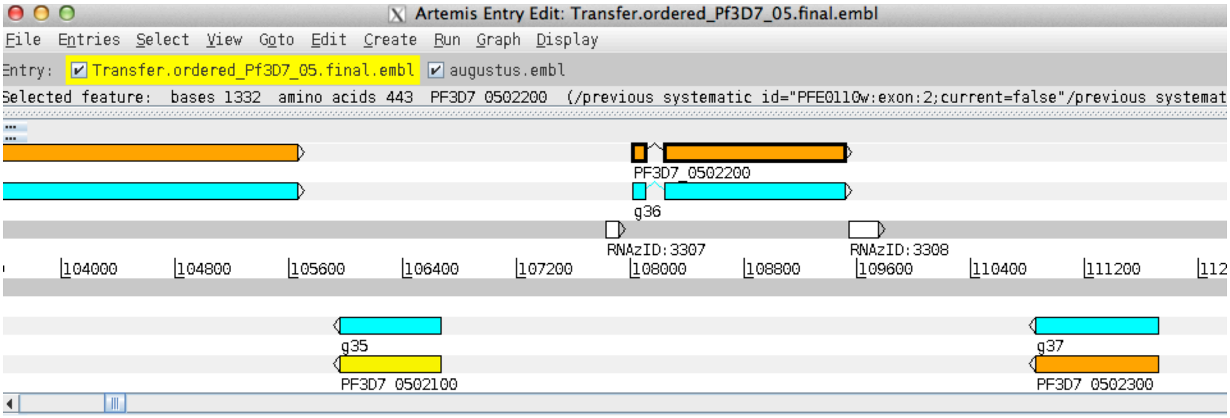
```
art Transfer.ordered_Pf3D7_05.final.embl &
```

Load in Augustus models. In Artemis:

File -> Read An Entry and select the file `augustus.embl`.

Right click in genome window, select "One Line Per Entry".

The transferred models have different colours and annotation, while the newly predicted genes have the default colour (blue), and are named g1, g2, g3…

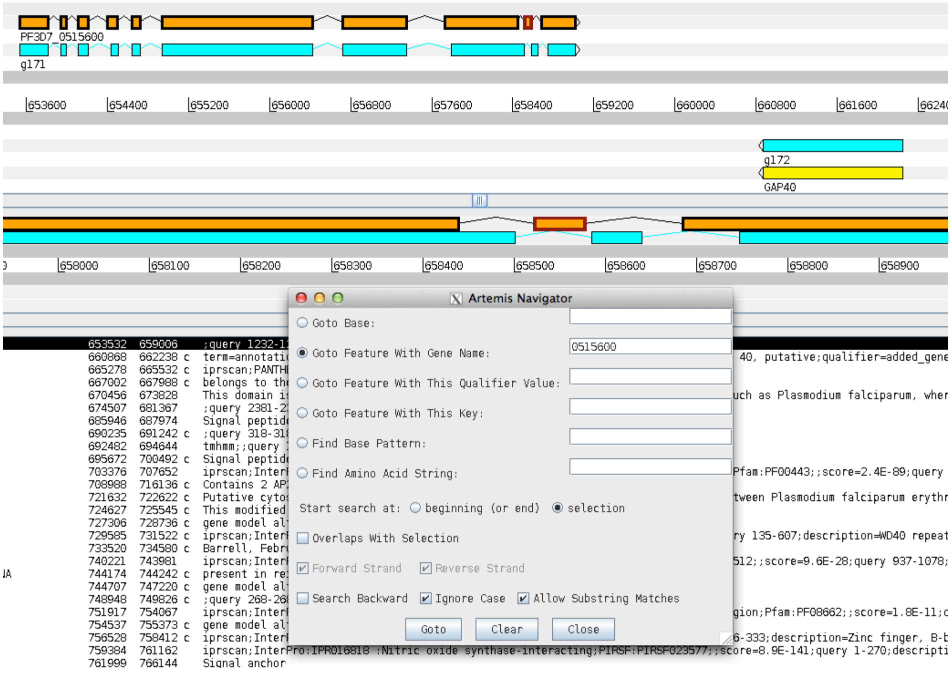


Artemis Entry Edit: Transfer.ordered_Pf3D7_05.final.embl

Entry: ☑ Transfer.ordered_Pf3D7_05.final.embl ☑ augustus.embl

```
Selected feature:  bases 1332  amino acids 443  PF3D7 0502200  (/previous systematic id="PFE0110w:exon:2;current=false"/previous systemat
```

Have a look at the gene PF3D7_0515600. Does Augustus perform better than RATT? What has gone wrong with the prediction? Go to the gene by Goto -> Navigator -> “Goto Feature With Gene Name”. Tip: you don’t have to type in the complete gene name.



○ Goto Base:

● Goto Feature With Gene Name: 0515600

○ Goto Feature With This Qualifier Value:

○ Goto Feature With This Key:

○ Find Base Pattern:

○ Find Amino Acid String:

Start search at: ○ beginning (or end) ● selection

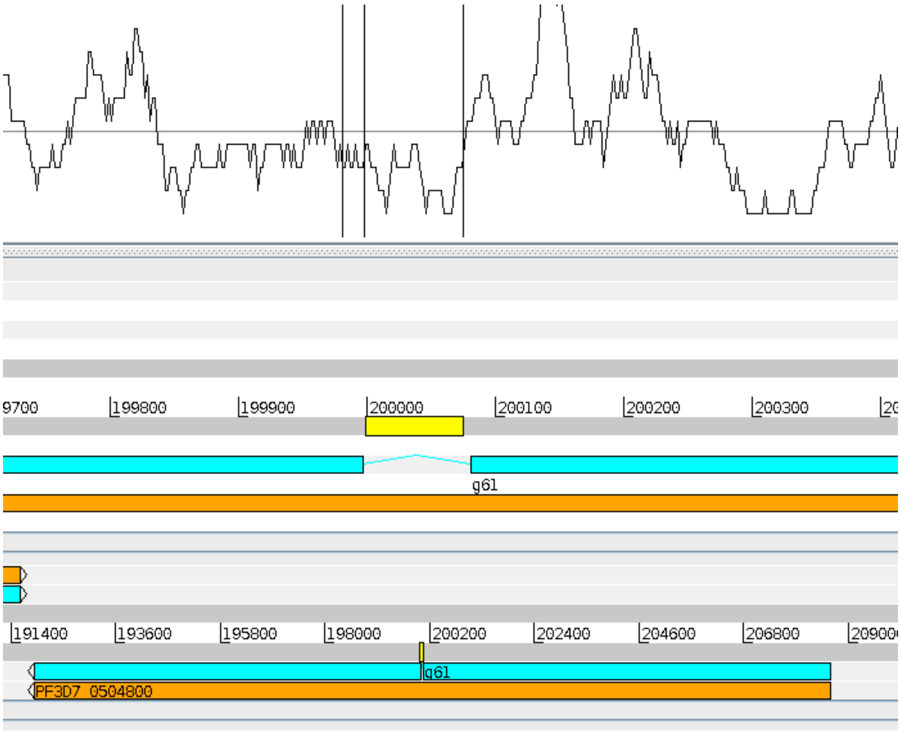
□ Overlaps With Selection

☑ Forward Strand ☑ Reverse Strand

□ Search Backward ☑ Ignore Case ☑ Allow Substring Matches

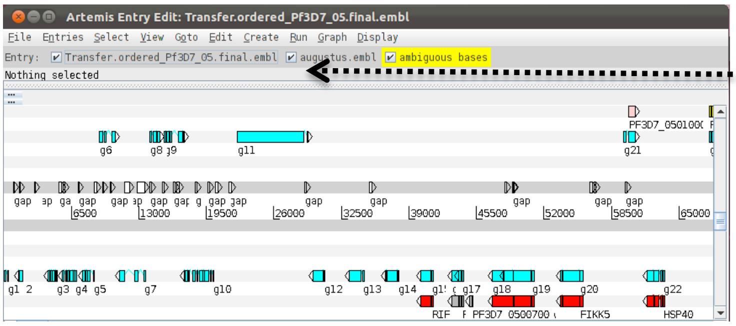
```
653532  659006    ;query 1232-1
660868  662238 c  term=annotati                                               40, putative;qualifier=added_gene
665278  665532 c  iprscan;PANTH
667002  667988 c  belongs to th
670456  673828    This domain i                                               uch as Plasmodium falciparum, wher
674507  681367    ;query 2381-2
685946  687974    Signal peptid
690235  691242 c  ;query 318-31
692482  694644    tmhmm;;query
695672  700492 c  Signal peptid
703376  707652    iprscan;Inter                                               Pfam:PF00443;;score=2.4E-89;query
708988  716136 c  Contains 2 AP
721632  722622 c  Putative cyto                                               tween Plasmodium falciparum erythr
724627  725545 c  This modified
727306  728736 c  gene model al
729585  731522 c  iprscan;Inter                                               ry 135-607;description=WD40 repeat
733520  734580 c  Barrell, Febr
740221  743981    iprscan;Inter                                               512;;score=9.6E-28;query 937-1078;
744174  744242 c  present in re
744707  747220 c  gene model al
748948  749826 c  ;query 268-26
751917  754067    iprscan;Inter                                               gion;Pfam:PF08662;;score=1.8E-11;c
754537  755373 c  gene model al
756528  758412 c  iprscan;Inter                                               6-333;description=Zinc finger, B-b
759384  761162    iprscan;InterPro:IPR016818 :Nitric oxide synthase-interacting;PIRSF:PIRSF023577;;score=8.9E-141;query 1-270;descripti
761999  766144    Signal anchor
```

Look at the gene PF3D7_0504800. Perhaps due to the low GC content of this region Augustus decided that it is intronic (Graph -> GC content (%)). How could you verify the prediction?



C. Combine RATT and Augustus gene models

Augustus is able to predict gene models in regions of the IT genome which have no synteny with the 3D7 genome, principally the subtelomeres. However, on the whole, the RATT-transferred gene models ought to be more accurate because the IT and 3D7 genome are otherwise very similar. Therefore it is perhaps most useful to add in only those Augustus gene models which do not overlap RATT-transferred models. To do this, the Augustus gene models must be converted into EMBL format.


Artemis Entry Edit: Transfer.ordered_Pf3D7_05.final.embl
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ Transfer.ordered_Pf3D7_05.final.embl ☑ augustus.embl ☑ ambiguous bases
Nothing selected
g6
g8 g9
g11
g21
PF3D7_050100
gap
6500
13000
19500
26000
32500
39000
45500
52000
58500
65000
g1 2
g3 g4 g5
g7
g10
g12
g13
g14
g17
g18
g19
g20
g22
RIF
PF3D7_0500700
FIKK5
HSP40

Unselect the “augustus.embl” entry.

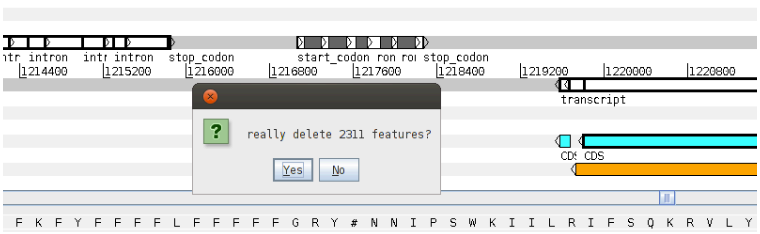
Go to Select -> All CDS features

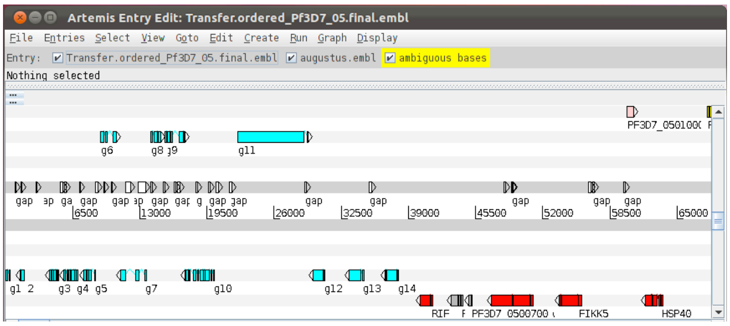
Re-select the “augustus.embl” entry.

Go to Select -> Features Overlapping Selection

n.b. deselect any augustus (blue) gene models which you have edited and want to keep by shift-clicking that model.

The Augustus gene models which overlap RATT models should now be selected. Delete them! Edit -> Selected Features -> Delete


really delete 2311 features?


Artemis Entry Edit: Transfer.ordered_Pf3D7_05.final.embl
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ Transfer.ordered_Pf3D7_05.final.embl ☑ augustus.embl ☑ ambiguous bases
Nothing selected
g6
g8 g9
g11
PF3D7_050100
gap
6500
13000
19500
26000
32500
39000
45500
52000
58500
65000
g1 2
g3 g4 g5
g7
g10
g12
g13
g14
RIF
PF3D7_0500700
FIKK5
HSP40

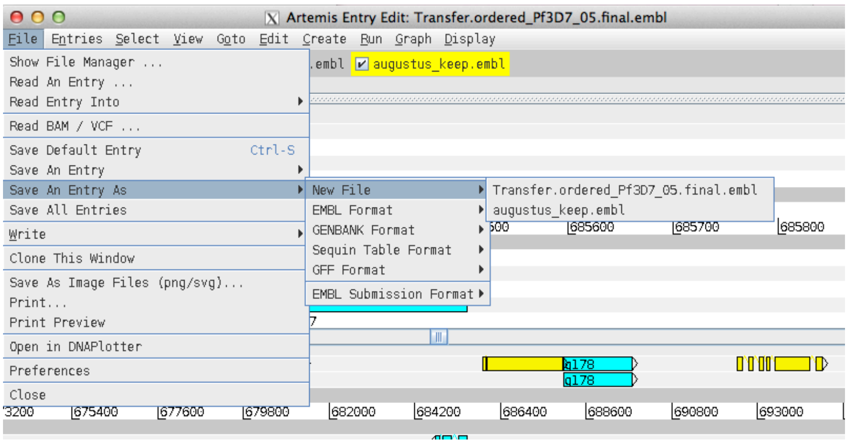
If you want to keep any augustus gene model you have edited, then delete the overlapping RATT model. How does the annotation look? How many extra gene models do we have compared to using the RATT-transferred ones alone? Use the function View -> Overview to see some stats. In the next section we are going to show how RNA-Seq data can be used to correct gene models.

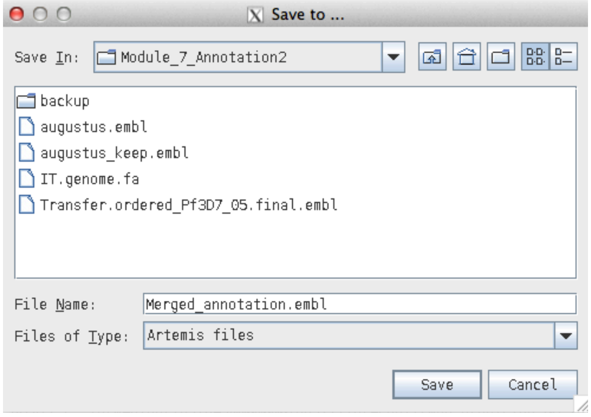
We need to save the merged annotation:
File -> Save An Entry As -> New File -> augustus.embl
Save to … "augustus_keep.embl"

Now merge it into the original file:
File -> Read Entry Into -> Transfer… Transfer.ordered_Pf3D7_05.final.embl
Select a file … "augustus_keep.embl"

And now save the merged file:
File -> Save An Entry As -> New File -> Transfer.ordered_Pf3D7_05.final.embl
Save to … "Merged_annotation.embl"

This file "Merged_annotation.embl" can now be used as a basis for further annotation. Always remember to save your work!





Save In: Module_7_Annotation2

backup
augustus.embl
augustus_keep.embl
IT.genome.fa
Transfer.ordered_Pf3D7_05.final.embl

File Name: Merged_annotation.embl

Files of Type: Artemis files

Gene models can also be merged/overlapped/removed bioinformatically using custom scripts or the free command-line software BEDtools http://bedtools.readthedocs.org/en/latest/ . The most convenient format for manipulating annotation is GFF/GFF3/GTF and BED.

D. Functional annotation

For those gene models transferred by RATT, you will have a range of functional information which will help you identify the types of genes present in your genome. This functional information has been manually curated for *P. falciparum* 3D7 based on the literature. For those genes predicted *de novo* by Augustus there is no such information. It is beyond the scope of this module to present a solution for assigning functional annotation for all these extra genes, but you can annotate a few of them yourself. Product calls for genes are usually defined by looking for orthologues or best BLAST hits in other organisms for which a gene has been annotated with a useful name. Many annotation databases exist for different purposes; Pfam for functional domains pfam.sanger.ac.uk, GOanna for GO-terms agbase.msstate.edu/cgi-bin/tools/GOanna.cgi and KAAS for enzymes www.genome.jp/tools/kaas/. There are often specialized databases for specific classes of genes you might be particularly interested in.

For *Plasmodium* annotation, the best place to look is PlasmoDb, a large resource of comparative genomics data for these species.

Right click on an augustus gene model (blue ones), View -> Amino Acids of Selection as Fasta -> Ctrl-A -> Ctrl-C

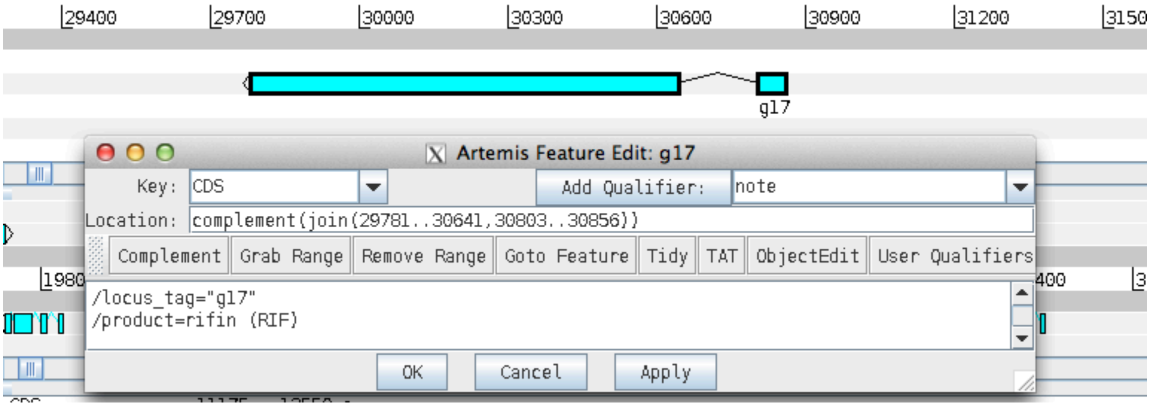
In a web browser, navigate to http://plasmodb.org/plasmo/

Under the Tools menu, select BLAST. In the web form, select “Proteins”, “blastp”, select all Target Organisms, paste in your sequence (Ctrl-V) and “Get Answer”. It is essential that your BLAST-search parameters match the search you are trying to do (protein versus protein).

If the top hit has a good E-value (e.g. less than 1e-20), select it and copy the description. Then select the gene model, press “Ctrl+e”, and add a new line as below. In the example below Augustus prediction g17 is a gene from the rifin (RIF) family.

```
/product=“rifin (RIF)”
```

Click OK to save the annotation.



Key: CDS
Add Qualifier: note
Location: complement(join(29781..30641,30803..30856))

```
/locus_tag="g17"
/product=rifin (RIF)
```

2. Mapping RNA-Seq data to a reference

We will use the program TopHat, part of the Tuxedo suite, to map RNA-Seq reads to our references genome, chromosome 5 of *P. falciparum* IT. In this case we are mapping single-end reads generated from RNA extracted during the blood stage of malaria. N.b. do not close the Artemis window.

TopHat requires an index of the reference. Create the index:

```
bowtie2-build IT.genome.fa IT.genome
```

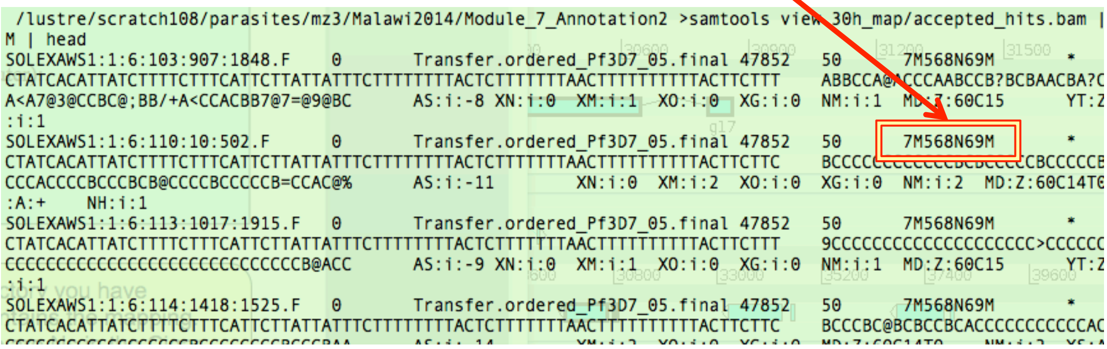
Now run TopHat (n.b. this may take several minutes depending on computer):

```
tophat -o 30h_map -I 10000 IT.genome blood_stage_30h.fastq &
```

To read the BAM file:

```
samtools view 30h_map/accepted_hits.bam | head
```

TopHat will generate several files in the new "30h_map" directory you have specified. The most important is accepted_hits.bam. This contains the mapping. Briefly read through the file if you like. For some reads, there are Ns in the Cigar line (column 6, see below). What do these mean?



```
 /lustre/scratch108/parasites/mz3/Malawi2014/Module_7_Annotation2 >samtools view 30h_map/accepted_hits.bam |
M | head
SOLEXAWS1:1:6:103:907:1848.F    0       Transfer.ordered_Pf3D7_05.final 47852   50      7M568N69M       *
CTATCACATTATCTTTTCTTTCATTCTTATTATTTCTTTTTTTTACTCTTTTTTTAACTTTTTTTTTTTACTTCTTT   ABBCCA@ACCCAABCCB?BCBAACBA?C
A<A7@3@CCBC@;BB/+A<CCACBB7@7=@9@BC      AS:i:-8 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:60C15       YT:Z
:i:1
SOLEXAWS1:1:6:110:10:502.F      0       Transfer.ordered_Pf3D7_05.final 47852   50      7M568N69M       *
CTATCACATTATCTTTTCTTTCATTCTTATTATTTCTTTTTTTTACTCTTTTTTTAACTTTTTTTTTTTACTTCTTC   BCCCC                 CBCCCCCB
CCCACCCCBCCCBCB@CCCCBCCCCCB=CCAC@%      AS:i:-11        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:60C14T0
:A:+    NH:i:1
SOLEXAWS1:1:6:113:1017:1915.F   0       Transfer.ordered_Pf3D7_05.final 47852   50      7M568N69M       *
CTATCACATTATCTTTTCTTTCATTCTTATTATTTCTTTTTTTTACTCTTTTTTTAACTTTTTTTTTTTACTTCTTT   9CCCCCCCCCCCCCCCCCCCC>CCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCB@ACC     AS:i:-9 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:60C15       YT:Z
:i:1
SOLEXAWS1:1:6:114:1418:1525.F   0       Transfer.ordered_Pf3D7_05.final 47852   50      7M568N69M       *
CTATCACATTATCTTTTCTTTCATTCTTATTATTTCTTTTTTTTACTCTTTTTTTAACTTTTTTTTTTTACTTCTTC   BCCCBC@BCBCCBCACCCCCCCCCCCAC
```

We have to index the bam using SAMtools, in order to view the data in Artemis.

```
samtools index 30h_map/accepted_hits.bam
```

The output index file is called `30h_map/accepted_hits.bam.bai`

The reads you just mapped does not cover the whole genome, so that the mapping would go faster. We will instead view the much larger pre-computed file which does cover the whole genome:

```
blood_stage_30h.bam
```

3. Viewing RNAseq mapping in Artemis

We will now examine the read mapping in Artemis using the BAM view feature.

If you have closed the Artemis window, then first type:

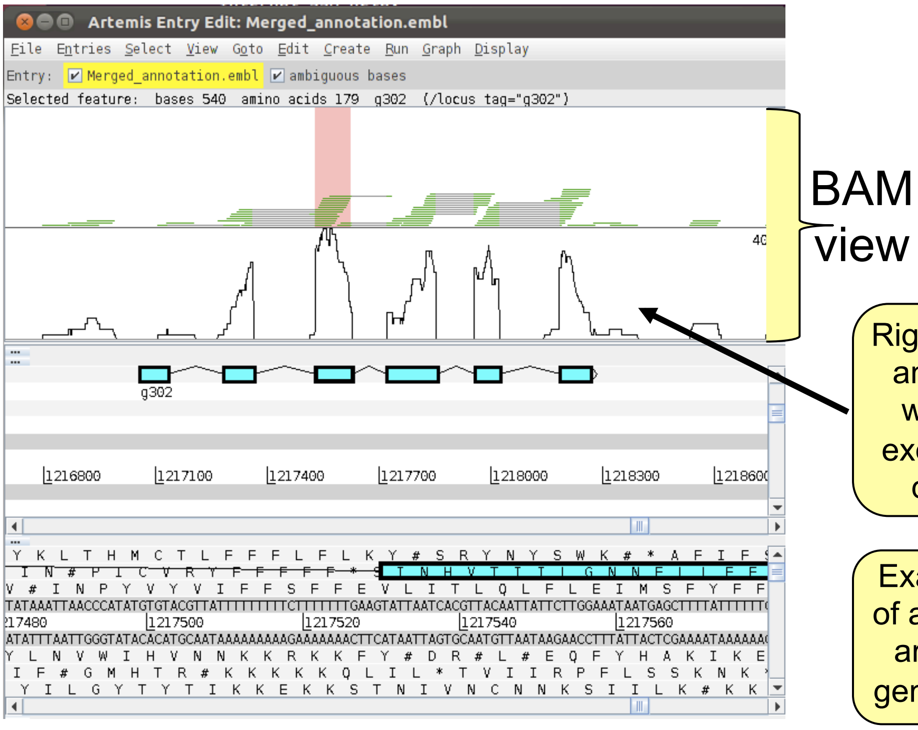
```
art Merged_annotation.embl &
```

In Artemis, highlight gaps in the assembly which might mislead you about the meaning of the RNAseq data:

Create -> Mark Ambiguities

Load the BAM file:
File -> Read BAM / VCF -> Select, "blood_stage_30h.bam" -> "OK"

This opens a new window at the top. Right click on the BAMview window, select Graph -> coverage.



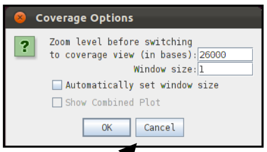
Artemis Entry Edit: Merged_annotation.embl

Entry: ☑ Merged_annotation.embl ☑ ambiguous bases

Selected feature: bases 540 amino acids 179 g302 {/locus tag="g302"}

```
 |1216800    |1217100    |1217400    |1217700    |1218000    |1218300    |1218600
```

```
 Y  K  L  T  H  M  C  T  L  F  F  F  L  F  L  K  Y  #  S  R  Y  N  Y  S  W  K  #  *  A  F  I  F
  I  N  #  P  I  C  V  R  Y  F  F  F  F  F  *  S  T  N  H  V  T  T  T  I  G  N  N  F  I  I  F  F
V  #  I  N  P  Y  V  Y  V  I  F  F  S  F  F  E  V  L  I  T  L  Q  L  F  L  E  I  M  S  F  Y  F  F
TATAAATTAACCCATATGTGTACGTTATTTTTTTTTCTTTTTTTGAAGTATTAATCACGTTACAATTATTCTTGGAAATAATGAGCTTTTATTTTTTC
217480         |1217500           |1217520           |1217540           |1217560
ATATTTAATTGGGTATACACATGCAATAAAAAAAAAGAAAAAAACTTCATAATTAGTGCAATGTTAATAAGAACCTTTATTACTCGAAAATAAAAAAG
Y  L  N  V  W  I  H  V  N  N  K  K  R  K  K  F  Y  #  D  R  #  L  #  E  Q  F  Y  H  A  K  I  K  E
 I  F  #  G  M  H  T  R  #  K  K  K  K  K  Q  L  I  L  *  T  V  I  I  R  P  F  L  S  S  K  N  K
  Y  I  L  G  Y  T  Y  T  I  K  K  E  K  K  S  T  N  I  V  N  C  N  N  K  S  I  I  L  K  #  K  K
```



Coverage Options

Zoom level before switching to coverage view (in bases): 26000

Window size: 1

☐ Automatically set window size

☐ Show Combined Plot

Right click on the coverage plot and select Options... Set the window size to 1 to see the exon boundaries (You have to disable "Automatically...")

Examine the exon boundaries of a couple of genes. How well are splice sites identified by gene predictors vs. RNA-Seq?

Try out the different read views in the BAMview. Right click on the BAMview, select Views then "Strand Stack". Paired views will not work because we are working with single-ended reads.

What do the black lines in the middle of some of the mapped reads mean? Right click on one of these reads, select "Show details of" and examine the cigar string.

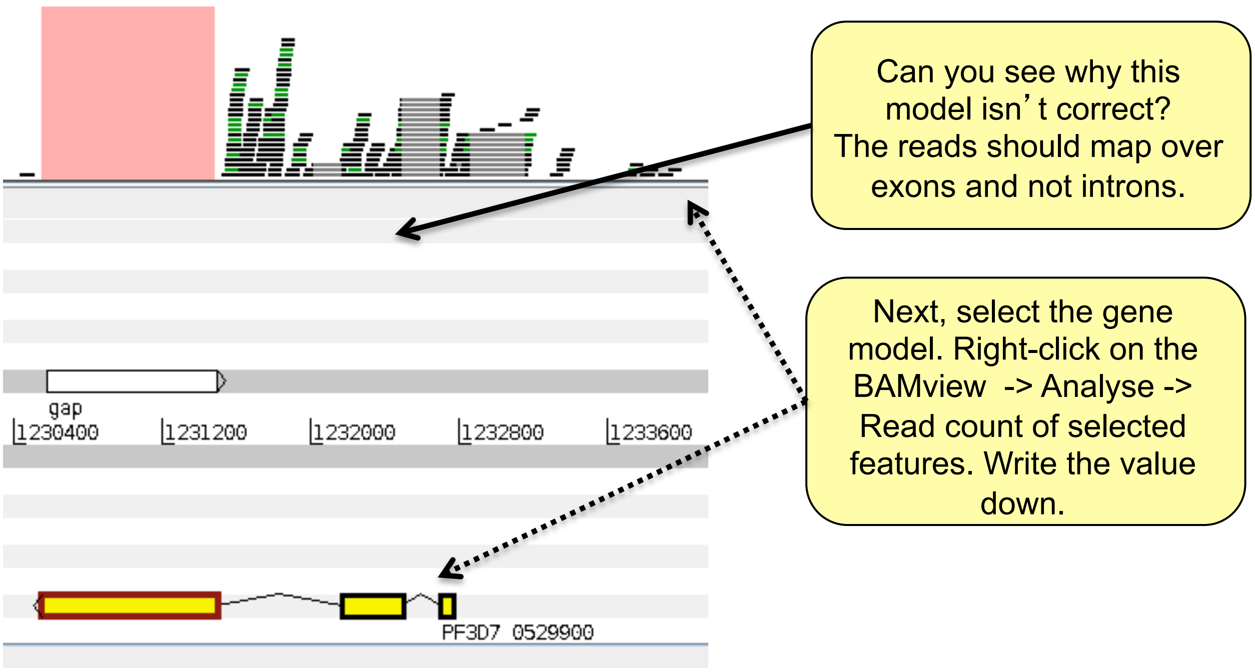
4. Correcting gene models by hand

Scroll along the chromosome and examine the read coverage. How well does it correlate with the gene models? Notice how different genes have different depths of coverage. Why do some genes have little or no coverage? What does this mean for annotation?

Why do some reads map where there are no genes?

Scroll along the chromosome. Can you see any gene models which might be incorrect based on the RNAseq data? Find one and correct it. PF3D7_0529900 is a good example.

Bonus question: Can you figure out why RATT got this gene model so wrong!



Can you see why this model isn’t correct? The reads should map over exons and not introns.

Next, select the gene model. Right-click on the BAMview -> Analyse -> Read count of selected features. Write the value down.

Now correct the last exon. You can move gene-boundaries by left-clicking at the edge of a gene-model and dragging it to the right position. If you are very zoomed out and the model is hard to catch, try shift + left-click instead. You can add exons/introns to a model by selecting the model and then clicking “e” to open the Feature editor. Left-click and highlight the region on the genome where you want to add an exon. Then switch to the Feature editor box again, and click the button “Grab range” or “Remove range”. Click “Apply” and you can straight away see the gene-model change. Under Edit -> Trim../Extend... there are some useful short-cuts for adjusting gene-boundaries. Look again at read counts/RPKM values. Have they changed? **n.b. Don’t forget to save any changes you make!**

Optional: Is the start of the gene model correct? There seem to be spliced reads confirming another exon.

5. Automatically generating gene models based on RNA-Seq data

Cufflinks is another program in the Tuxedo suite. It can be used to generate gene models based on the RNA-Seq mapping. Rather than fix each gene model by hand, we could replace them with RNA-Seq based predictions if these are better.

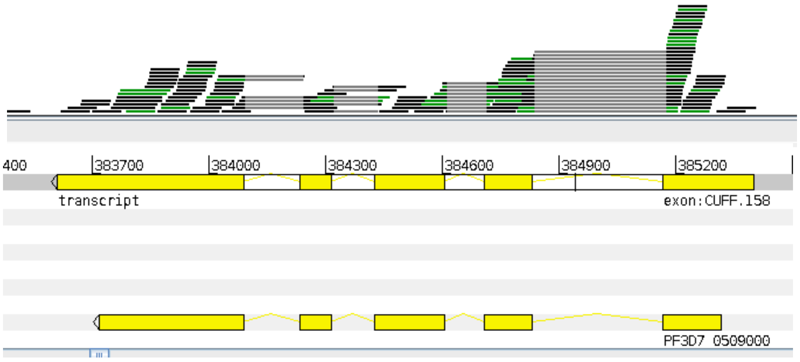
Run Cufflinks

```
cufflinks blood_stage_30h.bam &
```

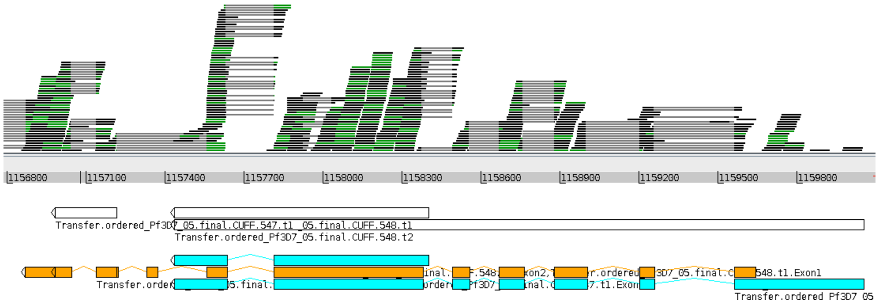
The key results file from Cufflinks is transcripts.gtf. When the output file `transcripts.gtf` has been created, change the format to display better in Artemis:

```
cufflinks2gff.pl transcripts.gtf
```

Read this into Artemis and compare the results to the RNAseq coverage plots and to the existing gene predictions. In Artemis,
File -> Read An Entry, select "Files of type: All files", select "transcripts.gtf"
Right click on the genome window, select "One Line Per Entry". Skim through the annotation. Does cufflinks confirm the Augustus predictions? What can we say where a gene model has no coverage? How can we remedy this?



Why are the Cufflinks predictions often longer than the gene model predictions? PF3D7_050900 is a good example. Remember we are examining a eukaryote. Why does this make it difficult to incorporate these models directly into our gene set? Would this explain the splice reads of the gene PF3D7_0529900 on page 11?



Optional: Can cufflinks help us to find alternative splicing? Maybe check the gene PF3D7_0527600 – Cufflinks has predicted two splice forms for this gene. Right-click in the genome window and choose "Feature Stack View" to see both splice forms.

6. Using RNA-Seq to improve predictions

Perhaps the best option would be to use the RNA-Seq to guide the *ab initio* predictions? Augustus allows you to to create hints-files from RNA-Seq. You can read more about how to create your own hints on augustus help pages: augustus.gobics.de/binaries/readme.rnaseq.html

On the command line, you first have to copy a suitable configuration-file from the augustus folder:

```
cp /usr/local/augustus.2.7/config/extrinsic/
extrinsic.M.RM.E.W.cfg .
```

Make hints from the bam-file. On the command-line, type:

```
bam2augustusHints.pl blood_stage_30h.bam > hints.introns.gff
```

Now predict new models using RNA-Seq hints, you'll notice it takes a bit longer than running without hints, expect a few minutes. On the command-line, type (as one line).

```
augustus --species=3D7
  --extrinsicCfgFile=extrinsic.M.RM.E.W.cfg
  --hintsfile=hints.introns.gff IT.genome.fa >
augustus.hints.gtf
```

Convert the Augustus GFF to EMBL format, so you can look at it in Artemis. On the command line, type:

```
cat augustus.hints.gtf | augustus2embl.pl >
augustus.hints.embl
```

Load the new predictions into Artemis like you did before, and compare this prediction to your earlier prediction without hints, to see which one you think is better. Try for instance to look at model PF3D7_0523600 and PF3D7_0528500.

Tip: if you want to see the difference between the models more clearly, you can in Artemis un-tick all files except for augustus.hints.embl at the bar at the top, choose Select -> "All CDS features". Click Edit -> "Qualifier of selected feature(s)" -> Change. In the drop-down menu in the pop-up box choose "colour", and then click "Insert qualifier:". Change the text in the box to /colour=3 and click "Add". All your models predicted using hints are now green.

Optional 1: Learn more augustus; a) learn how to train augustus with your favorite species from the online manual, b) check out the useful scripts that come with augustus; /usr/local/augustus.2.7/scripts/ , c) What does the Augustus option --alternatives-from-evidence do?

Optional 2: Try to look at all the differences between augustus predictions with and without hints 1. using programming, 2. in Artemis. How would you choose which predictions are the best? If you had to write a script to automatically choose between the models, what would that script contain?

Key aspects of genome annotation

Quality

The better the genome prediction is, the better the gene-models will be. If the genome is miss-assembled that can lead to partial or chimeric gene-models. The gold standard for gene-models is manual gene-model curation, but for draft genomes there often not resources available to do this. So for draft genomes you may have to accept that you will not have a perfect set of genes. Ten years of annotating the malaria genome by hand using all possible lines of evidence has not resulted in a perfect annotation (although it is a very good one)! How do you think the quality of the gene-models affect the analysis you can do, and the conclusions you can draw?

Several lines of independent evidence are best

Predicting gene-models, you will find that one of the hardest things to do is to choose which set of models are the best; all methods are good at some types of genes and bad at others. Augustus has a built-in quality check, in which you can compare your training models with your predicted models. Use a variety of prediction approaches, as you have done here, to capture as many of the genes as possible and to improve their accuracy. It is always a good idea to try different programs for any particular problem in computational biology: if they all produce the same answer you can be more certain it is correct. In the case of gene model predictions they will frequently disagree. If several predictions are of similar high quality, perhaps the best option is to combine different sets of gene using tools such as Jigsaw (Allen & Salzberg, 2005) and EVM (Haas et al., 2008)?

Over-prediction

There is a balance to strike between having almost all the genes and lots of erroneous ones as well, or to miss some genes but have relatively few incorrect ones. It may be important to find as many of the real genes as possible, however once you have published an erroneous model to the public sphere, for instance by submitting it to GenBank, it can be very hard to retract it later, and it may cause problems for other people using that gene for their analyses.

Bacteria are simpler

If you work on bacteria you will encounter fewer problems with accurately predicting gene models as they almost always have single-exon genes. The program Glimmer3 (Delcher et al., 1999) is an alternative to Augustus for *ab initio* gene prediction in bacteria, but since version 2.7 Augustus has improved its prediction methods for bacterial genomes. Another alternative for both gene prediction and functional annotation for bacteria is Prokka www.vicbioinformatics.com/software.prokka.shtml

Alternative splicing

Many genes in more complex organisms have several alternative splice-forms, including/excluding different UTRs and exons. Predicting alternative splicing is much harder than predicting a “canonical”; most complete, gene-model (like the ones you have just predicted). It is currently only possible to predict alternative transcripts reliably for model genomes which already are very well assembled and annotated, and have a wealth of supporting evidence (like human and *C. elegans*).

Key aspects of RNA-seq mapping

Non-unique/repeat regions

A sequence read may map equally well to multiple locations in the reference genome. Different mapping algorithms have different strategies for this problem, so be sure to check the options in the mapper. A low GC content, such as in *Plasmodium falciparum* (81% AT) means that reads are more likely to map to multiple locations in the genome by chance.

Insert size

When mapping paired reads, the mapper (e.g. TopHat) takes the expected insert size into account. If the fragments are expected to on average be 200bp, and the reads are 50bp, then the insert between the paired reads should be ~100bp. If the paired reads are significantly further apart than expected, we can suspect that the reads have not mapped properly and discard them. Removing poorly mapping reads can produce a more reliable mapping.

Spliced mapping

Eukaryotic mRNAs are processed; after transcription introns are spliced out. Therefore some reads (those crossing exon boundaries) should be split when mapped to the reference genome sequence in which intron sequences are still present. TopHat is one of few mappers which can split reads while mapping them, making it very suitable for mapping RNA-Seq. Beware that TopHat cannot recognize donor and acceptor splice-sites so it will split reads only based on optimizing the mapping, and you will occasionally see a couple of bases of the read having ended up on the wrong side of the intron.

Alternative mappers

Alternative short read mappers which do not split reads include SOAP (Li et al., 2008b), SSAHA (Ning et al., 2001), BWA (Li et al., 2009) and Bowtie2 (Langmead B, Salzberg S. 2012), SMALT (Ponstingl, unpublished). All of these may be appropriate for bacterial RNA-Seq. Where introns are an issue RUM (Grant et al., 2011) is one alternative to TopHat (Trapnell et al., 2009).

New tools for mapping sequence reads are continually being developed. This reflects improvements in mapping technology, but it is also due to changes in the sequence data to be mapped. The sequencing machines we are using now (e.g. Illumina HiSeq, 454 GS FLX etc) will perhaps not be the ones we are using in a few years time, and the data the new machines produce may not be best mapped with current tools.

Beware of the genes!

In spite of our very best efforts, it is not always possible to predict genes accurately. There are many phenomena which can throw both automatic and manual predictions off track. For instance (but not limited to): seleno-proteins containing “stop-codons” as part of the coding sequence, polycistronic genes, splice-leader trans-splicing, long non-coding RNAs, repetitive genomic regions and pseudogenes. Before publishing, it is always a good idea to try to estimate how correct the models are, and doing some sanity-checking of the gene-models.