Group task Version 3.1701

The idea of the group task is for you to apply the learnt from the week and apply it yourself. The groups, best three students, should be assembled Thursday lunch. Thursday late afternoon you might meet and discuss how to proceed.
Friday morning you will have to work on the project, which will be presented in the afternoon. The presentation should be short 3-4 minutes and EVERY member of the group should present something.

It is really informal, so no panic!

If you have a nice dataset you want to work on, please let one of the instructors know in advance if it is a good project. For example, mapping 24 RNA-Seq conditions might be too much, but if you already have the read counts or SNP data, and the other members of the group are happy to do the project, it would work.

Another option for group tasks would be to download data from known publication, especially if read counts are available. Below we have some projects you could do.

Overview
De novo assembly with longreads (de novo assembly)
Analysis of HIV (virus genomics)
Analysis genes important for Drug resistance in Malaria (mapping)
Find gene responsible for gametocyte production (mapping and variant calling)
Expression of type I interferon in different species
Differential expression in P. falciparum
Differential expression in P. chabaudi
scRNA-Seq in mammary epithelial cells
scRNA-Seq in malaria

Groups Task: Assembly of virus
Assemble and compare HIV samples.
- Use IVA to assemble the genome.
- Compare it to the reference (act). Can you find any clear differences?
- Map a further sample against the new assembly and the reference. Which regions of the genome are the most diverse?

Data:

Group Task de novo Assembly.

Covers: De novo assembly, Annotation, RNA-Seq, Artemis, act, canu, PlasmoDB

Question/Aim: How different is are chromosomes of *P. falciparum* in terms of genome structure, genes and expression.

The aim for this task is to assemble chromosome four of the *Plasmodium falciparum* HB3 clone, annotate it, use RNA-Seq to look at the expression and do analysis of genes or regions in the genome that looks different to the Chr4 of the *P. falciparum* 3D7.
As data you have Pacific Bioscience of HB3 in a fastq files and Illumina RNA-Seq data from one time points of PfHB3 and also Pf3D7.
We also have the read counts of 7 timepoints of PfHB3 and Pf3D7 over the 48h malaria blood cycle.

Work tips:

- Assemble the PacBio reads of the direction ~/GroupTasks/Task1
- Annotate the new chromosome with Companion
- Map the RNA-Seq against and compare the RPKM with the BAM file of Pf3D7 Chr4. (Ask Uli how to down load the Chr4 of P. falciparum from geneDB)
- Generate an act comparison file and search for differences between the genomes
- For interesting genes, go to webpages like plasmoDB or Panoptis to analyse those genes further.

Characterization of genes responsible to drug resistance in Malaria

P. falciparum is the causative agent of **the most dangerous form of malaria in humans**. The reference genome for *P. falciparum* strain 3D7 was determined and published about 10 years ago (Gardener et al., 2002). Since then the genomes of several other species of *Plasmodium* that infect humans or animals have been elucidated. Malaria is widespread in tropical and subtropical regions, including parts of Asia, Africa, and the Americas. Each year, there are approximately 350–500 million cases of malaria killing more than one million people, the majority of whom are young children in sub-Saharan Africa.
To date, the genomes of several strains of *P. falciparum* have been sequenced completely. For this exercise we will examine 76bp paired-end sequence read data from the malaria strains Dd2 and IT. In particular the *P. falciparum* Dd2 strain is well known for its **resistance to commonly used antimalarial drugs such as chloroquine**. Working with the mapped sequence data and Artemis we will have a closer look at some SNPs and CNVs that contribute directly to the drug-resistance phenotype of this deadly parasite.

Data:
ftp://ftp.sanger.ac.uk/pub/pathogens/tdo/Exercise/GrouptaskMappingMalaria.zip

Task
1. Download the zip file, unzip it. Do a QC of the reads. Map the reads (**DD2.Chr5_?.fastq**) against the reference **Pf3D7_05.fasta** and look if the quality of the experiment is good (**Artemis**).
2. Do the SNP calling.
3. The gene of interest is MDR1. Can you characterize the locus? (SNPs, indels, CNV)? (btw, Plasmodium is haploid)

Group Task: Which gene is the important one for gametocyte production?
Malaria parasites in culture loose the ability to produce gametocyte, which gives them a growth advantage in culture. In a study of Andy Waters, Abhiny (PhD student) grow severel P. berghei parasites, which over time (like after month) lost the ability to grow gametocytes.

- Map the reads of the parent and clones against the reference
- Call variants
- Find the genes that is responsible for gametocyte production.

Group Task: differential expression in P. falciparum

Covers: Different statistical methods to analyse RNA-Seq data and comparison in PlasmoDB

Question/Aim: Which genes are differential expressed between two lab strains over the red blood cycle; RNA-Seq, DESeq, Cluster methods, PlasmoDB, Functional analysis like GO enrichments

The core genes of *Plasmodium falciparum* are very conserved, but are they also expressed the same way? In this project you will get the reads count of three *P. falciparum* parasites (3D7, IT, HB3), over 7 time points. The aim is to work in R to try to do some differential expression. This could be done through DESeq2, pairwise, or through clustering methods. The aim is to find genes that behave different. Next you could perform analysis on genes that are differentially expressed (GO enrichment, pathways analysis in PlasmoDB) or look at genes that are very similar expressed to see how conserved they are (orthoDB).

Work tips:
- Load the read count into R (~/GroupTasks/Task2)
- Look at some BAM files in artemis to get an idea of the quality (~/GroupTasks/Task2)
- Run DEseq2 and maybe MBcluster.seq
- Look for genes differential expressed in plasmoDB
- Look for conserved genes. Do they also have less mutations (PlasmoDB, PanOptis)

Expression of type I interferon in different species

Following a viral infection, mammalian species express type I interferon (IFN) which brings about the host innate immune response and the upregulation of hundreds of interferon-stimulated genes (ISGs). To determine the interferome of the pig, *Sus scrofa*, primary cells were stimulated with type I IFN or left as controls. Four replicates were sequenced on the Ion Proton and the data is available here: https://www.ebi.ac.uk/ena/data/view/PRJEB21332. The accession numbers of the samples you will need are available in Table 1.

Table 1. Accession numbers for the *Sus scrofa* experiment.

Run	Sample name	Control/IFN
ERR2012489	Pig_1	Control
ERR2012490	Pig_2	Control
ERR2012491	Pig_3	Control
ERR2012492	Pig_4	Control
ERR2012493	Pig_ifn1	IFN
ERR2012494	Pig_ifn2	IFN
ERR2012495	Pig_ifn3	IFN
ERR2012496	Pig_ifn4	IFN

The aim of your group task is the determine the upregulated and downregulated genes following type I IFN stimulation and to describe the types of genes that are affected. It will involve carrying out the following analyses:

1) Downloading the data from the public repository.
2) Assessing the presence of cell culture contaminants (optional).
3) Downloading and indexing the pig genome (http://www.ensembl.org/Sus_scrofa/Info/Index).
4) Aligning the Ion Proton reads to the host genome (e.g., Hisat2 + bowtie2).
5) Counting the reads mapping to the annotated genome and performing differential expression.
6) Producing a heatmap of the results.
7) Determining the GO of the significantly upregulated and significantly downregulated genes.

Group Task: Detect genes important between serial transmitted and naturally transmitted infections

In this group task you will address the effect of vector transmission on gene expression of the malaria parasite. Is the transcriptome of a mosquito-transmitted parasite different from one which has not passed through a mosquito? The key reason for asking this question is that parasites which are transmitted by mosquito are less virulent than those which are serially blood passaged in the laboratory. Figure 1A shows the malaria life cycle, the red part highlighting the mosquito stage. Figure 1B shows the difference in virulence, measured by blood parasitemia, between mosquito-transmitted and serially blood passaged parasites. The data in this exercise, as well as figures 1B and 1C are taken from Spence et al. (2013).

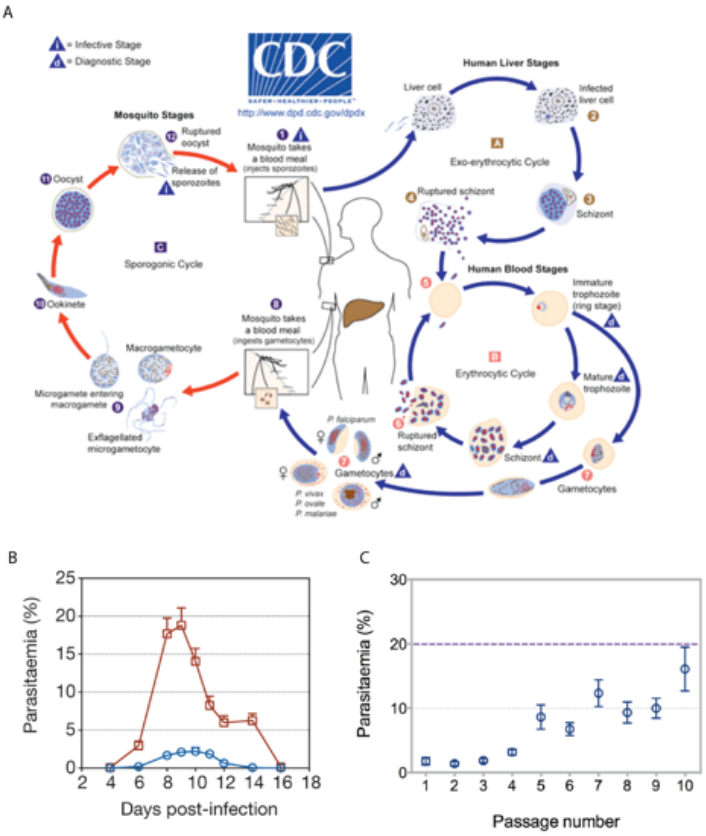

Figure 1. Serial blood passage increases virulence of malaria parasites. (A) The lifecycle of plasmodium parasites involves mammalian and mosquito stages. Experiments in the lab often exclude the mosquito stage (red) and instead remove parasites from the blood of a mouse to infect another mouse (serial blood passage). (B) Serially blood passaged parasites (red) are more virulent than mosquito-transmitted parasites (blue) as shown by their higher parasitemia over the course of infection. (C) As mosquito transmitted parasites are serially blood passaged an increasing number of times, they return to a higher level of parasitemia.

Are there any differences between the transcriptomes of serially blood passaged parasites and mosquito-transmitted parasites which might explain how they are able to do this?

To perform this exercise you can follow the first RNA-Seq exercise.

Data: /nfs/pathogen003/tdo/workshop/GroupTasks/Pc.RNA-Seq

scRNA-Seq in mammary epithelial cells

An example of scRNA-Seq with 10X genomics. Old colleagues of mine published that, looking at the dynamics of breast cells of mice.

Download the read counts of the below paper.

- Perform the clustering
- Perform the pseudo time analysis

Which genes are interesting?

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-02001-5 OPEN

Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing

Karsten Bach[1,2,3], Sara Pensa[1,3], Marta Grzelak[2,3], James Hadfield[2,3], David J. Adams[3,4], John C. Marioni[2,3,4,5] & Walid T. Khaled[1,3]

Characterising the hierarchy of mammary epithelial cells (MECs) and how they are regulated during adult development is important for understanding how breast cancer arises. Here we report the use of single-cell RNA sequencing to determine the gene expression profile of MECs across four developmental stages; nulliparous, mid gestation, lactation and post involution. Our analysis of 23,184 cells identifies 15 clusters, few of which could be fully characterised by a single marker gene. We argue instead that the epithelial cells—especially in the luminal compartment—should rather be conceptualised as being part of a continuous spectrum of differentiation. Furthermore, our data support the existence of a common luminal progenitor cell giving rise to intermediate, restricted alveolar and hormone-sensing progenitors. This luminal progenitor compartment undergoes transcriptional changes in response to a full pregnancy, lactation and involution. In summary, our results provide a global, unbiased view of adult mammary gland development.

Groups Task: scRNA-Seq in Malaria

Get the dataset https://www.biorxiv.org/highwire/filestream/30918/field_highwire_adjunct_files/0/105015-1.gz, (https://www.biorxiv.org/content/early/2017/02/10/105015.figures-only) table S6.
Can you also load the read counts in a different, more efficient way (look for the EBI array express).

Which genes are the different differentially expressed within each group (asexual / gametocyte), and between the groups?

Can you do a GO enrichment (via topGO or PlasmoDB)? What are the conclusions?

Optional: Can you replicate the heatmap from the paper (Figure 2d)?
Data: Online, BioRxiv and ArrayExpress