**Sensurveiledning skriftlig eksaemn INF-BIO5121 og 9121, høst 2015**

**1 Assembly**

1a) What is the conceptual difference between contigs and scaffolds?

5 points for getting that contigs are gapless and build from overlapping reads && scaffolds are ordered and oriented contigs && with gaps in-between

4 points  
• order and orientation missed  
• gaps not mentioned for scaffolds

1b) Given even, high coverage error-free reads, describe what can cause assemblies to become fragmented

5 points for mentioning repeats longer than the read length 3 when repeats, but not 'longer than read length' is mentioned

'bonus' when faulty assembly program is mentioned

1c) Explain why it is difficult to say, given a set of assemblies, which one is “best”

10 points if explicitly saying depends on research question/biological question, and gives an explanation/example  
8 if it is clear that is what is meant, but it is not said explicitly  
6 as above, but without example/explanation  
'bonus' for adding time and money considerations

1d) What is the conceptual difference between how the de Bruijn and Overlap (overlap layout consensus) graphs are used in the assembly process

OLC based on overlap between reads, usually for longer reads, requires a smart way to do all-against-all overlaps. Contigs are built from consensus of alignments in non-bifurcating paths through the graph  
DBG based on kmer graph, overlaps are edges between adjacent kmers. Contigs are constructed from the kmers in non-bifurcating paths through the graph. Usually for shorter reads. More sensitive to read errors, need more memory.

**2 Variant calling**

2a) Describe at least 2 factors that introduce uncertainty into the variant calling process and explain how and why they introduce uncertainty.

>> covered in lectures, there is a slide that summarises this

Here are the bare minimum explanations. Description (2 points) + how and why (2 points), good detail for (example a graphic) or more than 2 factors provided (1 point)

\* Base quality:   
\* Base quality is the quality of the base in the sequence read and relates directly to probability that the base call is incorrect. This quality varies from base to base in a sequence read and typically deteriorates towards the end of a read.  
 \* Low quality base calls introduce uncertainty as to what the true underlying base is and can affect both the correct mapping of reads and variant calling. If many bases at a particular site are affected by poor quality (this is most likely to happen at sites of low coverage), the quality of the variant call at that site will also be poor.

\* Mapping quality:  
 \* Mapping quality is Phred-scaled probability that a read is incorrectly mapped. This probability is computed by the mapping algorithm.   
 \* Poor mapping quality introduces uncertainty as to whether the read actually originates from the location that it is mapped to. A typical example of reads which are difficult to map are those coming from regions that have homologous regions in the genome. If reads are wrongly mapped, then there may be mismatches between the read and the reference due to the mismapping rather than to a true difference between the sample and the reference.

\* Alignment:  
 \* Alignment is carried out following mapping of reads as the mapping algorithm does not necessarily correctly align all the bases in the read to the corresponding base in the reference.  
 \* It is important that each base in the read is correctly aligned to its corresponding base in the reference as any misalignment will appear as a mismatch between the sample and the reference and may thus result in the site being erroneously called variant. Misalignment is common around indels.

\* PCR duplicates:  
 \* PCR duplicates occur because Polymerase Chain Reaction is used in the sample preparation. Typically only one fragment of each PCR product is sequenced, but sometimes more than one is sequenced. If this occurs we call these PCR duplicates.  
 \* The PCR process is not error free and thus some PCR duplicates will not be exact copies of the original fragment. If several such fragments are sequenced and mapped to the genome, they will make the site to which they are mapped appear variant when in fact the sample is the same as the reference. This uncertainty can be overcome by removing all PCR duplicates prior to variant calling.

\* Sequencing depth:  
 \* Sequencing depth is the number of reads from a sample covering a site in the genome. One typically aims for a coverage of 30X in order to get good quality variant calls.  
 \* Low sequencing depth introduces uncertainty into the variant calling process because 1. the base calling is not perfect and thus one needs several reads of each base in the reference genome to have high confidence that a variant is true and not just the result of a base call error, and 2. in a diploid genome where a sample may be heterozygous one needs good depth in order to ensure that both alleles have been sampled multiple times.

2b) What is re-alignment, and why is it done?

>> covered at length in the lectures  
1. What is re-alignment (2 points)  
Re-alignment is a step in the variant calling pipeline performed after mapping. The mapping algorithm's purpose is to map individual reads to the region that they originate from but does not necessarily produce a high quality alignment. A much better alignment of reads to the reference can be achieved by aligning jointly all reads mapping to a given region of the genome.

2. Why is it done  
Basic (2 points): A correct alignment of reads is key to variant calling as misaligned reads can cause the variant calling to miss true variants and to call false positive variants.  
Extra (1 point): This is particularly a problem when the sample carries insertions or deletions.

2c) Describe briefly and in general terms what the Burrows-Wheeler transform is, where it is used in the field of high throughput sequencing, and why it is needed.

WHAT THE BURROWS-WHEELER TRANSFORM IS (3 points)  
The Burrows-Wheeler transform of a String T is the last column of the Burrows Wheeler matrix of T. The rows of the BW Matrix are the consecutive rotations of T. The BWT(T) has properties that facilitate rapid search of T for any substring of T.

WHERE IT IS USED IN THE FIELD OF HIGH THROUGHPUT SEQUENCING (3 points)  
The application of the BWT to the reference sequence enables a rapid search of the reference for the location that a read originated from in other words it is used to map a read to the reference.

WHY IT IS NEEDED (3 points)  
It is needed because mapping millions of reads to a large genome would be extremely time consuming if one were to use a brute force approach of evaluating every read at every possible position in the genome.

Extra point for good explanation beyond the bare minimum presented here.

**3 RNA-SEQ**

3a) What is the definition of a transcriptome and how would you design an RNAseq appriach to make a reference transcriptome for a non-model organism? 5p

**Answer:**  A transcriptome consists of all RNAs present in a given organism, tissue, cell or isolate at a given time.

Given a non-model organism a *de novo* approach or a reference based approach using a closely related species are the options. Using long-read technology is beneficial to ensure high-quality transcripts and resolution of rare splice variants. Medium length technology is also usable (less expensive) especially if differential expression analysis is intended down stream. Also, if differential expression is intended biological replicates are needed. A *de novo* approach requires greater sequencing depth to resolve transcripts and decrease the number of assembly artefacts. Distribution of libraries across different tissues and/ or time points is beneficial when making a reference transcriptome to cover any time/tissue specific expression.

* Selection of approach: de novo or reference based = 1 p.
* Selection of read length (no need to mention name of technology) = 1p
* Coverage/depth = 1 p
* Distribution of libraries = ½ p

More sequencing/biological replicates = ½ p.

**#** Second part of the question could be interpreted two ways. Teacher’s intention: experimental design. Alternative impretation: step by step procedure as done during the course. If students have discussed the elements listed above (even when doing the alternative interpretation) they have been given points otherwise not.

On the definition of a transcriptome I deducted ¼ p if the time element was skipped and if the student did not specify where the isolate came from (whole organism, tissue, cell…).

I further deducted ½ p if reasons were not stated. Examples: I would choose a medium length sequencing technology = ½ p. I would choose a medium length sequencing technology because I would like to do differential expression / it’s cheaper = 1p. I would use a de novo approach = ½ p. I would use a de novo approach because this is a non-model organism presumably without a genome = 1 p.

3b) Briefly describe the pros and cons of *ab initio* (reference based) and *de novo* transcriptome strategies

**Answer:**  Artefacts: ab initio will not generate any other artefacts than those present in the reference genome whereas de novo will, due to the assembly process, likely generate faulty transcripts. This is especially true for rare transcripts/splice variants. Non-reference transcripts: any transcripts not present in the reference genome will be lost when performing ab initio transcriptomics; however, de novo transcriptomics will detect all transcripts if enough sequencing is performed. Resolution: Ab initio is more sensitive and provides higher resolution of rare/novel transcripts compared to de novo. This is due to using mapping instead of assembly - easier to obtain good quality transcripts will few mapped reads that few reads used as input for assembly. Fragmentation: This is also the reason for de novo transcriptomes being more fragmented – single exons are assembled but not merged together as whole transcripts. Time/computing: Ab initio requires much less time and computing power due to the lacking assembly step which is present in de novo. Coverage: Ab intio requires less coverage compared to de novo. Reference: Ab intio is heavily dependent on reference quality as errors in the genome will be transferred to the transcriptome. Splice junctions: Ab initio requires a certain coverage of splice junctions to resolve splice variants not needed for de novo. Complex/gene dense organisms: De novo can be difficult for complex eukaryotes and gene dense species as the assembly algorithm will struggle to resolve the transcripts properly.

* Artefacts 1p
* Non-reference transcripts 1p
* Resolution 2p – I weighted this more as this is one of the central differences. Gave ½ p if the student stated that ab initio had better resolution without giving more information.
* Degree of fragmentation 1p
* Time/computation power consumption 1p
* Coverage/sequencing depth 1p
* Reference quality 1p
* Splice junction coverage 1p
* Gene dense/complex organisms 1p

3c) Briefly describe the main differences between DESeq and EdgeR and the assumptions you make during DE analysis when performing a comparison of two groups with replicates (PhD only)

**Answer:**

**Comparisons:**

Variance estimation: DESeq estimates dispersion (variance) per gene using the relationship between the mean and variance and then fits a curve through the estimates. It then assigns the per gene fit for all genes above the curve (large variation) and the curve fit for all genes below (little variation). EgdeR estimates the dispersion using a Cox-reid profile adjusted maximum likelihood method per gene. A trended curve is fitted to these estimates. Genes above the curve is assigned the per gene fit and those below are assigned the trended curve fit.

DESeq test vs GLM: DESeq tests for differential expression using a regression test based on negative binomial distribution where as edgeR uses a general linearized model approach.

Filtering: edgeR filters away all genes with low counts regardless wheres DESeq uses all genes in the matrix unless otherwise stated.

Type I error profile: This profile is connected to the dispersion estimate. DESeq becomes very conservative on lowly expressed genes due a skewed error I profile connected to the assignment of curved fit to genes with per gene fit below the curved fit. This profile is more even in edgeR as it uses a maximum likelihood dispersion estimate/trended curve fit where the assignmet of trended curved fit to genes with less variation does not affect the error I profile in the same degree as in DESeq.

**Assumptions:**

Negative binominal error distribution: Both packages assume that errors are distributed according to the negative binomial distribution as several papers in litterature show this. However; this is not the case for all genes and thus you enforce a less fitting distribution on some of your data.

Independent samples: Both packages assume that your samples are independent. This is not always the case. Sampling the same individual over time would require a different approach. You also assume that the expression level of gene A at time 1 does not affect the expression of gene A at time 2.

Transcript length: Both packages assumes that transcript length does not affect the outcome by stating that the length of differential expressed transcripts in condition A is similar to that of condition B. This assumption is not always right especially when treatment affects different gene families.

* Variance estimation: 2p – I’ve give 1p per R package
* DEseq negative binomial test vs GLM 2p – I’ve give 1p per R package
* Filtering 1p
* Type 1 error 2p – 1 point for the skewed error profile and 1 point for DESeq being more conservative
* Negtive binomial distribution 1p
* Independent samples 1p
* Transcript length 1p

**4 - Statistical genomics and reproducibility**

4A. Briefly describe why reproducibility is important for computer-based analyses? Explain some of the main problems and how they may be overcome?

**Note**: it was not assumed (because of time constraints) that the student would not be able answer with all the details give below. The points were given based upon whether the aspects were more or less covered, with less point given if the aspect was mention too briefly. One extra point was given to students that were felt to be scoring too low based upon a general overview of their answer.

Answer:

Using computer-based analyses in science should be considered in the same way as with other more traditional experiments. It is important to make the research reproducible, for several reasons.  
- Other researchers should be able to reproduce your experiments to check your findings. This is integral to the system of peer-reviewed journals. Specifically, people reviewing your article should also be able to find the details of what you have done, so that they can assess the scientific value of ones findings, and be able to report errors if any. This is a moral obligation as researchers. (3p)  
- Often, the person making use of a computer-based analysis is the one that need to (possibly years later) reproduce the findings. Also, one may want to reuse the methodology/software for other projects later. This is a more practical argument, making life easier for the researcher. (1p)

Main problems:

Which data were used?  
- Note the exact place (URL) the software was downloaded, which time, and which changes were done to it. Also provide the datasets in the raw and preprocessed forms.  
- Which software was used? In what version? Be sure to store intermediate data. (2p)

Which software was used?  
- Provide the name of the software with version numbers. Provide archive of software for distribution. Also, for custom scripts, provide public access to the scripts. (2p)

Which analysis parameters where used?  
- Provide the exact parameters used (for monte carlo analyses, this would include null model, test statistic and random number seed). Avoid manual steps without documentation. Even better is to providing the user with the possibility of rerunning the analyses with same or different parameters. One may use analysis frameworks such as Galaxy to support reproducibility in a simple way. (2p)

(This question was based upon the "Ten simple rules for reproducible computational resarch" article is the major one here, as well as slides from the Galaxy and Statistical genomics modules)

4B. Briefly explain the key components of Hypothesis testing.

Note: The points were given based upon whether the aspects were more or less covered, with less point given if the aspect was mention too briefly. Just giving a list of the components without explanation would only give a few points in total. One extra point was given to students that were felt to be scoring too low based upon a general overview of their answer.

Alternative hypothesis (H1): (2p)  
- What you really want to show (e.g. relation between tracks A and B)

Null hypothesis/model (H0): (2p)  
- A neutral baseline (e.g. no relation exists between tracks A and B). A Null model is a model from which the null hypothesis arises.

Test statistic: (2p)  
- A measure of the aspect of interest (e.g. base pair overlap)

P-value: (2p)  
- How likely is the observation (or more extreme), given H0

Significance level: (2p)  
- The p-value “cut-off” (0.05) is called significance level (usually denoted by α)  
- Observation is unlikely (e.g. p-value less than α) -> reject H0, data supports H1 (=significant results)

(This question was based upon slides and exercises from the Statistical genomics module)