microRNA analysis

Merete Molton Worren
Ståle Nygård

Help personnel:
Daniel Vodak (BCF)
Morten Johansen (BCF)

Background

- Dysregulation of miRNA expression has been connected to progression and development of atherosclerosis
- The hypothesis:
  - miRNA expression profiles differ between baboons with high and low serum LDL-C
  - A subset of these miRNAs regulate genes are relevant to dyslipidemia and risk of atherosclerosis

Experiment

- Baboons divided into two groups based on their LDL-C response levels
- Two diets
  - Chow (baseline diet)
  - HCHF (High Cholesterol High Fat diet)
- Compared miRNA expression profiles from liver biopsies collected before and after the HCHF challenge diet

Paper

- Karere GM, Glenn JP, Vandeberg JL, Cox LA. 2012 **Differential microRNA response to a high-cholesterol, high-fat diet in livers of low and high LDL-C baboons**. BMC Genomics 13(1):32

- http://www.ncbi.nlm.nih.gov/pubmed/22809019

Experiment setup

- 12 samples available from baboon liver biopsies
  - 3 replicates from high-LDL, baseline (chow) diet
  - 3 replicates from high-LDL, HCHF diet
  - 3 replicates from low-LDL, baseline (chow) diet
  - 3 replicates from low-LDL, HCHF diet

- We will work with the high-LDL samples only

QUESTIONS?

miRNA analysis workflow

- Preparations
- Alignment
- Annotation
- Read count per miRNA
- Normalization
- Differential expression analysis

Preparations, read files

- Starting point: sequence read files

- Practical info
  - File format (.fastq, .fasta, .sff)
  - Base quality scale used (Phred+33, Phred+64)
  - Quality control
  - Adapters

Preparations, reference

- Find sequences you will align to
- Resources:
  - UCSC: http://hgdownload.cse.ucsc.edu/downloads.html
  - NCBI: http://www.ncbi.nlm.nih.gov/guide/all/#downloads
  - ENSEMBL: http://www.ensembl.org/info/data/ftp/index.html

- We will use sequences from miRBase that we will customize to our needs
  - http://www.mirbase.org/

Preparations, miRNA

- Collapse files for increased speed
- Check that reads/reference are both RNA or both DNA

QUESTIONS?

Next:
Alignment

Alignment

- The sequence read is not very informative in itself.
- Alignment to a reference -> location of origin

TAGCTTATCAGACTGATGTTGA

Alignment

- The sequence read is not very informative in itself.
- Alignment to a reference -> location of origin

TAGCTTATCAGACTGATGTTGA

TACACTGAGTCGACTGAGTGGCCTAGCTTATCAGACTGATGTTGAATTTATATAGGCGTGACGTA

Alignment

- The sequence read is not very informative in itself.
- Alignment to a reference -> location of origin

TAGCTTATCAGACTGATGTTGA

TACACTGAGTCGACTGAGTGGCCTAGCTTATCAGACTGATGTTGAATTTATATAGGCGTGACGTA

mir-21-5p

What to align to

- Often several possible references
  - The genome of the organism or a closely related organism
  - Specific sequences like mRNA or miRNAs
  - Databases such as Rfam
    - filter out reads
    - check what the unaligned reads are

Aligners

- Many different aligners
  - Novoalign
  - Bowtie
  - BWA (Burrows-Wheeler Aligner)
  - SOAPaligner/soap2 (Short Oligonucleotide Analysis Package)

Alignment challenges

- Sequencing errors
- True differences from reference
- Repeat regions
- Similar coding sequences

Alignment process

- Indexing
  - An index is made of the genome
- Mapping based on index
  - Finds several possible alignment location
- Final alignment
  - Thorough search of possible mapping locations

Simplified example, k-mer=2

- ”Genome”: TTATATTTTATTAATTTTA

Possible 2-mers	Location positions
TT	1,6,7,8,11,15,16,17
TA	2,4,9,12,18
AT	3,5,10,14
AA	13

- Our read: ATT
- Start 2-mer AT

Options

- Different options for different aligners

- Reading reference manuals is a big part of doing bioinformatics

Some common options

- Level of similarity, genome/reads
- Output formats
- Options to control multiple mappers

QUESTIONS?

Next:
Annotation

Annotation

- Aligned to sequence of interest
  - The annotation is already in place
- Aligned to genome
  - Either: coordinates for the features of interest are already known
  - Or: part of project is to annotate the genome

Annotation

- Some genomes are better annotated than others
- Coordinates for specific genomic features can be found through several sites
  - ENSEMBL http://www.ensembl.org/index.html
  - UCSC Table browser http://genome.ucsc.edu/cgi-bin/hgTables
  - miRBase http://www.mirbase.org/ftp.shtml

QUESTIONS?

Next:
Read counts

Read counts

- How much do we have?
- Read counts is the starting point for finding expression and differential expression
- When counting, multiple mappers must be considered

Read count miRNA

- Multiple mappers are common
  - Size of reads
  - Similarity of different miRNAs

- This leads to some additional challenges when counting the reads of miRNAs

Aligning miRNAs to miRNAs

Reads\Reference	hsa miRs	all mirs
hsa miRs (2578 reads)	2578/2494 **84 miRs lost**	2578/1669 **909 miRs lost**
all miRs (30424 reads)	10111/9653 **458 miRs lost**	30424/15732 **14692 miRs lost**

Counting multiple mappers

- Disregard all the reads that maps to multiple locations
- Keeping one or more randomly
- Keeping them all

QUESTIONS?

Next:
Normalization

Normalization

- The purpose of normalization is to remove technical bias while maintaining true biological signal.

Normalization

- Normalization is necessary because the true expression is not the only factor influencing the read counts
- Other factors may be
  - Sample size
  - Differences in extraction
  - Differences during the sequencing procedure

Normalization methods for miRNA

- No method for best practice yet
  - Normalization by read counts
  - Several other methods tried with better results
  - TMM

QUESTIONS?

Next:
Differential expression analysis

Differential expression analysis

- Differential expression analysis aims at finding the miRNAs that are truly differentially expressed under the relevant conditions
- The challenge is to be able to find as many as possible of the true ones, while avoiding to call a miRNA differentially expressed when the changes are random or due to other factors

Differential expression analysis

FLU (subject1)	HEALTHY (subject2)

- Is this a good way of checking what miRNAs are differentially expressed when having the flu?

Differential expression analysis

FLU (subject1)	HEALTHY (subject2)
Male	Female
Arthritis	Allergy
Eats vegetables	Eats burgers and fries

- Is this a good way of checking what miRNAs are differentially expressed when having the flu?

Biological replicates

- There will be a lot of miRNAs that have different expression in two samples
  - Some due to the condition you are investigating
  - Some due to other factors
- The solution: biological replicates
  - Average out unknown/unwanted differences
  - Call differential expression only where it is due to the factor you are investigating

QUESTIONS?

Next:
A little statistics with Ståle Nygård (Eivind Hovig's group)

QUESTIONS?

Next:
Discussion of the results

Comparison

Our analysis

miRNA	FDR
• miR-206	2.7-23
• miR-1	2.8-19
• miR-133a-3p	4.2-19
• miR-133b	3.1-18
• miR-95-3p	8.7-08
• miR-499a-5p	5.7-05
• miR-4454	0.016
• miR-452-5p	0.021

Paper

miRNA	FDR
• miR-29b	0.01
• miR-222	0.01
• miR-221	0.02
• miR-1	0.35

Possible sources of difference

- From fastq files to read counts
  - Differences in aligner/reference
  - Different way of counting

- Statistics
  - Different normalization
  - Different statistical tests used

Learning points

- Different pipelines will give different results
- Very important to give good descriptions of exactly what you have done
- PCR validation is important