High-Throughput Sequencing Course
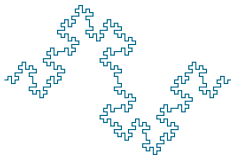## Introduction

Biostatistics and Bioinformatics

Summer 2017

FROM RAW UNALIGNED READS


```
@SRR546799.1 HWI-1KL120:92:C0F56ACXX:1:1101:1203:2232 length=50
CATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAAC
+SRR546799.1 HWI-1KL120:92:C0F56ACXX:1:1101:1203:2232 length=50
B@CFFFEFHHHHHFGHIGHHIJIHIIIJJIHBFHG=FFCEIIEAACECDE
@SRR546799.2 HWI-1KL120:92:C0F56ACXX:1:1101:1152:2242 length=50
CTCGTGAACTCATCTCCGGGGGTAGAGCACTGTTTCGGCAAGGGGGTCAT
+SRR546799.2 HWI-1KL120:92:C0F56ACXX:1:1101:1152:2242 length=50
@=?DBDBDFHHDDGHGHGIGGG77BFHIFIHIIIIGHGHHFF=?ADDB<A
@SRR546799.3 HWI-1KL120:92:C0F56ACXX:1:1101:1429:2119 length=50
ACCACGTGTCCCGCCCTACTCATCGAGCTCACAGCATGTGCATTTTTGTG
+SRR546799.3 HWI-1KL120:92:C0F56ACXX:1:1101:1429:2119 length=50
@@@FFFDFHHHH:EGIHGIEHEGHHHEHFHCFGCGGFHGHHIIHIIIIII
@SRR546799.4 HWI-1KL120:92:C0F56ACXX:1:1101:1376:2136 length=50
GTTAATCGGGGCAGGGTGAGTCGACCCCTAAGGCGAGGCCGAAAGGCGTA
+SRR546799.4 HWI-1KL120:92:C0F56ACXX:1:1101:1376:2136 length=50
@?@FFFFFHGHHHHJJ9CBGGHIIIJJJFGIGIIIBGHHFFDDDDDDD;?
@SRR546799.5 HWI-1KL120:92:C0F56ACXX:1:1101:1417:2140 length=50
CTGGGTTGTTTCCCTCTTCACGACGGACGTTAGCACCCGCCGTGTGTCTC
+SRR546799.5 HWI-1KL120:92:C0F56ACXX:1:1101:1417:2140 length=50
B?BFFFFDFHHHGJFHHGIGJFGHJGBGHIIJIGFHGIGGIHHEDDECEE
@SRR546799.6 HWI-1KL120:92:C0F56ACXX:1:1101:1320:2224 length=50
CCCAGAGCCTGAATCAGTGTGTGTGTTAGTGGAAGCGTCTGGAAAGGCGC
+SRR546799.6 HWI-1KL120:92:C0F56ACXX:1:1101:1320:2224 length=50
:
```

TO ALIGNED READS

TO COUNTS


```
> head(counts(htseq),20)[,1:15]
          7A_E 7A_G 7A_K 7A_N 7A_P 7B_E 7B_G 7B_K 7B_N 7B_P 7C_E 7C_G 7C_K 7C_N 7C_P
gene0        9   17   11   17   11   12   22   20    6    9   19   20   17    5   20
gene1      108  170   97   88  173  119  241  103   51  162  155  149  124   88  128
gene10       3    0    7    3    3    2    1    1    2    2    2    2    2    7    5
gene100     24   27   15   16   23   11   24   28    5   30   24   20   22   15   25
gene1000    11    5    8    2   13   10    8    7    2   13    8    2    5   13    9
gene1001     1    3    2    5    2    3    1    1    3    5    3    4    4    1    2
gene1002    32   11   19   12   23   31   29   19   11   34   22   20   19   12   27
gene1003    80   60  109   58   68  100   57   74   36   74   76   75   85   55   58
gene1004     1    2    1    1    3    0    5    0    0    1    1    3    1    2    0
gene1005   873  499  713  356  662 1259  575  585  236  820  937  521  486  317  809
gene1006    24   14   33   17   28   25   20   20   10   21   21   15   17   27   12
gene1007    64   29   86   46   49   79   52   57   28   65   67   22   75   38   54
gene1008    16    6   23   14   11   21   21   26   10   15   25   12   23   14   20
gene1009     9    8   17    5   14   17   13    9    2   12   18    6    5    9    7
gene101     29   39   29   42   47   46   68   40   16   41   48   80   46   28   41
gene1010     0    1    2    0    1    4    0    0    0    2    0    0    1    0    1
gene1011     0    1    0    0    0    0    0    1    0    0    2    0    0    0    1
gene1012     2    0    1    0    1    2    1    0    1    0    0    1    0    1    0
gene1013     0    0    2    0    2    0    0    0    1    1    0    0    0    0    1
gene1014     2    0    1    0    1    2    0    0    0    0    1    1    0    0    0
>
```

Differential Expression


```
fitting model and testing
-- replacing outliers and refitting for 46 genes
-- DESeq argument 'minReplicatesForReplace' = 7
-- original counts are preserved in counts(dds)
estimating dispersions
fitting model and testing
log2 fold change (MAP): trt 8 vs 7
Wald test p-value: trt 8 vs 7
DataFrame with 4444 rows and 6 columns
            baseMean log2FoldChange      lfcSE        stat      pvalue       padj
           <numeric>      <numeric>  <numeric>   <numeric>   <numeric>  <numeric>
gene0      15.274431     0.28920009  0.2167382   1.3343292 0.1820959756 0.334270077
gene1     145.603062     0.43095114  0.1292386   3.3345378 0.0008544128 0.004147663
gene10      2.605083    -0.28595073  0.3674671  -0.7781668 0.4364706803 0.614286779
gene100    20.323396     0.08658647  0.1486582   0.5824532 0.5602614320 0.723906417
gene1000    6.582580    -0.43057986  0.2612653  -1.6480558 0.0993412243 0.214598998
...              ...            ...        ...         ...         ...         ...
gene995     1.6041044     0.6238433  0.4006699   1.5570009   0.1194703  0.2450365
gene996    10.3271263    -0.2176632  0.1992665  -1.0923221   0.2746915  0.4504187
gene997     6.8183976    -0.2618863  0.2651733  -0.9876041   0.3233466  0.5039471
gene998    29.3582205    -0.2004418  0.1752968  -1.1434424   0.2528549  0.4264820
gene999     0.6089341    -0.1343551  0.5377144  -0.2498632   0.8026931  0.8962573
>
```

DIFFERENTIAL EXPRESSION
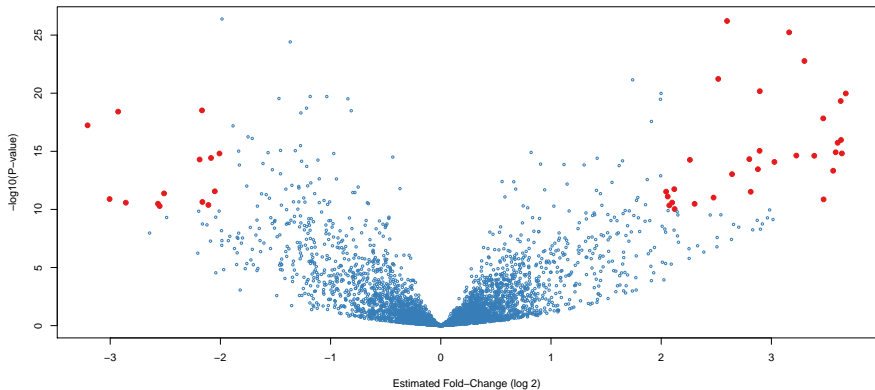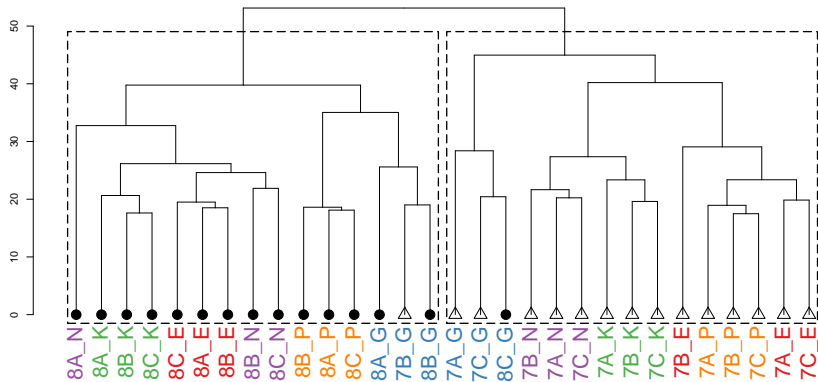
PCR/Microarray versus RNA-seq: Common objectives and challenges

- ▶ Hypothesis testing: Is the mRNA abundance related to a phenotype, or changed in response to treatment or over time

- ▶ Effect size estimation: How to quantify the effect size and then how to estimate it from data

- ▶ Classification: Predict an outcome on the basis of baseline RNA levels from multiple genes

- ▶ Class Discovery: Discover subsets on the basis of baseline levels or changes in the levels of multiple genes

- ▶ Multiplicity: how to deal with testing not a single marker but thousands if not millions of markers ($P < 0.05$ makes no sense here or anywhere)

RNA-Seq: A tool for measuring abundance of RNA from cells

- The data observed are not gene expressions (quantified on a continuum)
- We observe the number of reads mapped to each gene
- These are counts
- Microarrays: consider distributions and regression models for quantitative traits (often assume that these are normally distributed)
- RNA-Seq: consider distributions and regression models for counts

MRNA ABUNDANCE, GENE EXPRESSIONS AND READ COUNTS

- ▶ Suppose that Y is the true abundance for a gene of interest
- ▶ $\hat{Y}$: the "expression" measured by microarray transcript (e.g., oligo nucleotide)
- ▶ K: The number of RNA-Seq reads mapped to the gene
- ▶ Questions:
  - ▶ Is $\hat{Y}$ close to Y (the truth)?
  - ▶ Is K close to Y (the truth)?
  - ▶ Should K even be compared with Y?

RNA-Seq: Two Approaches

- Two-stage method:
  - Convert counts to "Expression" (e.g., RPKM, FPKM, TPM)
  - plug these into a standard tests or regressions models
  - In essence: Force things into a microarray problem
- One-stage method:
  - Relate the counts directly to the phenotype
  - Use distributions and regression models for counts

Emphasis, Focus, Approach and Topics

- ▶ Concepts rather than on mechanics (e.g., which software or method to use to fit a regression model)
- ▶ How statistical concepts are misunderstood or misinterpreted
- ▶ How and why things could go wrong
- ▶ Use simulation as a tool to illustrate these issues
- ▶ Topics:
    - ▶ Statistical Inference (testing and estimation)
    - ▶ Supervised learning (classification and regression)
    - ▶ Unsupervised learning (class discovery)
    - ▶ Multiple testing
    - ▶ Pathway/Gene-Set Analysis
    - ▶ Meta-Analysis
    - ▶ Distributions and regression models for counts

DECISION VERSUS TRUTH

- Any statistical method will yield a decision
- Whether that conclusion of the decision is close to the truth or even reasonable will remain unknown
- We have to accept that the decision may be wrong
- Goal: Bound the probability of a wrong decision through the use of proper statistical design and methods
- and *proper* and *measured* interpretation of the results

The Simulation Method

- Simulate data from the "truth" *in silico* using computers
- Apply your proposed statistical method to the simulated (synthetic) data
- Repeatedly compare the decision at which you arrive (by virtue of the chosen statistical method) to the truth (under your control)

THE SIMULATION METHOD: NOISE DISCOVERY

- ▶ Simulate data from noise
- ▶ Example: simulate treated and untreated samples from the same distribution
- ▶ Assess the proportion of times you arrive at the wrong conclusion
- ▶ Wrong conclusion: Conclude that there a treatment effect
- ▶ Important tool for identifying self-fulfilling prophecies.

On Statistics, Conclusions and Solutions

"No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the 'one chance in a million' will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us."

Ronald Aylmer Fisher (The Design of Experiments (1935), 16)

"Doing statistics is like doing crosswords except that one cannot know for sure whether one has found the solution."

John Wilder Tukey (Annals of Statistics, 2002:30(6))