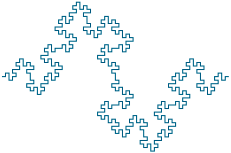
High-Throughput Sequencing Course

Welcome

Biostatistics and Bioinformatics



Summer 2018

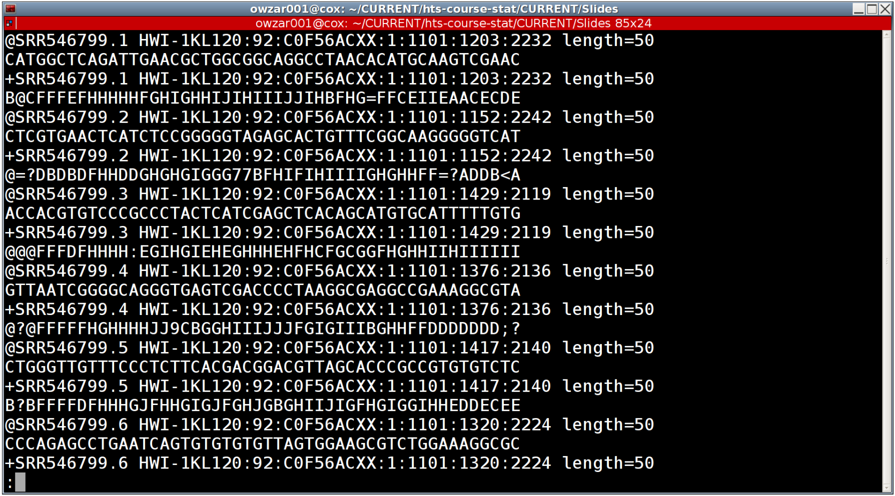
Duke University
School of Medicine

DUKE BIOSTATISTICS
& BIOINFORMATICS

Welcome from HTS Course Faculty and Staff

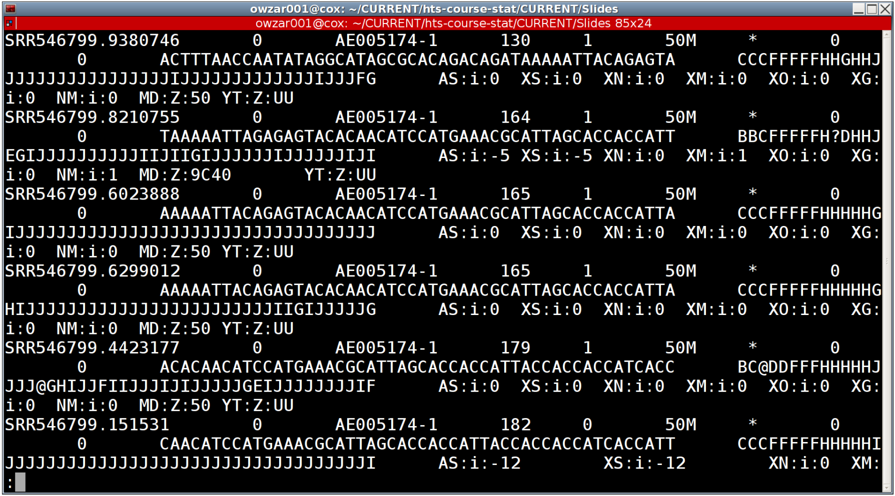
- *Biology and Computational Biology*
  - David Corcoran
  - Holly Dressman
  - Raluca Gordân
  - Josh Granek
  - Kathleen Miglia
- *Computing*
  - Cliburn Chan
  - Janice McCarthy
- *Statistics*
  - Andrew Allen
  - Yi-Ju Li
  - Kouros Owzar
  - Jichun Xie
- *Program Evaluation*
  - Ed Neal
- *Translational Bioinformatics*
  - Anna-Maria Masci
  - Jessica Tenenbaum
- *Teaching Assistants*
  - Jeremy Gresham
  - Kuei (Clint) Yueh Ko
  - Benji Wagner
  - Paul Zweck
  - C?
- *Resource specialist*
  - Sharon Updike
- Administration
  - Tasha Allison
  - Tim Durning
  - Dawn Hails
  - James Thomas
- Special Thanks: Liz Delong, Tim Reddy

Raw Unaligned Reads


owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides
owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides 85x24

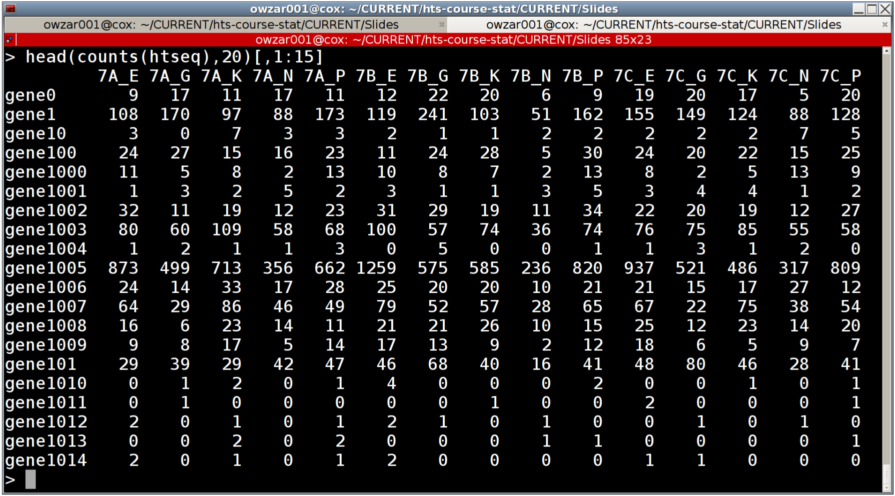
```
@SRR546799.1 HWI-1KL120:92:C0F56ACXX:1:1101:1203:2232 length=50
CATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAAC
+SRR546799.1 HWI-1KL120:92:C0F56ACXX:1:1101:1203:2232 length=50
B@CFFFEFHHHHHFGHIGHHIJIHIIIJJIHBFHG=FFCEIIEAACECDE
@SRR546799.2 HWI-1KL120:92:C0F56ACXX:1:1101:1152:2242 length=50
CTCGTGAACTCATCTCCGGGGGTAGAGCACTGTTTCGGCAAGGGGGTCAT
+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
:
```

Aligned Reads


owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides
owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides 85x24

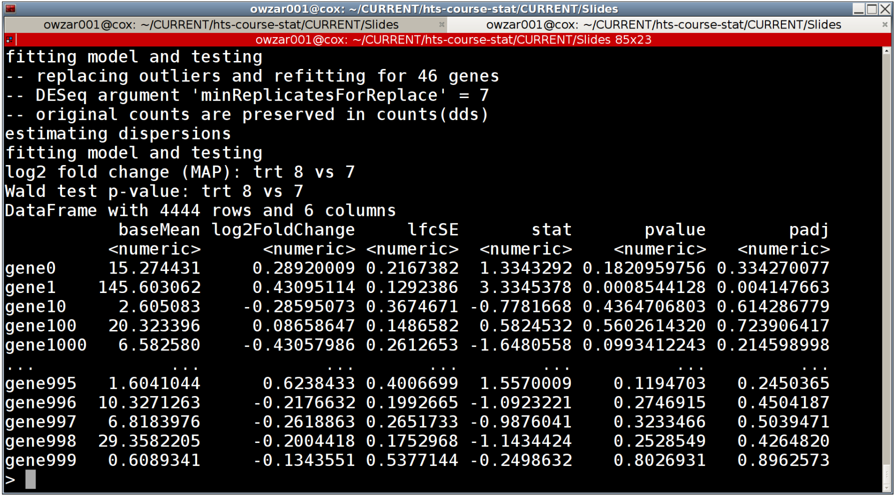
```
SRR546799.9380746       0       AE005174-1      130     1       50M     *       0
       0       ACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTA      CCCFFFFFHHGHHJ
JJJJJJJJJJJJJJJJIJJJJJJJJJJJJJJIJJJFG      AS:i:0  XS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:
i:0  NM:i:0  MD:Z:50 YT:Z:UU
SRR546799.8210755       0       AE005174-1      164     1       50M     *       0
       0       TAAAAATTAGAGAGTACACAACATCCATGAAACGCATTAGCACCACCATT      BBCFFFFFH?DHHJ
EGIJJJJJJJJJJJIIJIIGIJJJJJJJIJJJJJJJIJI      AS:i:-5 XS:i:-5 XN:i:0  XM:i:1  XO:i:0  XG:
i:0  NM:i:1  MD:Z:9C40       YT:Z:UU
SRR546799.6023888       0       AE005174-1      165     1       50M     *       0
       0       AAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTA      CCCFFFFFHHHHHG
IJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ      AS:i:0  XS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:
i:0  NM:i:0  MD:Z:50 YT:Z:UU
SRR546799.6299012       0       AE005174-1      165     1       50M     *       0
       0       AAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTA      CCCFFFFFHHHHHG
HIJJJJJJJJJJJJJJJJJJJJJJJJJJJIIGIJJJJJG      AS:i:0  XS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:
i:0  NM:i:0  MD:Z:50 YT:Z:UU
SRR546799.4423177       0       AE005174-1      179     1       50M     *       0
       0       ACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACC      BC@DDFFFHHHHHJ
JJJ@GHIJJFIIJJJIJIJJJJJGEIJJJJJJJJJIF      AS:i:0  XS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:
i:0  NM:i:0  MD:Z:50 YT:Z:UU
SRR546799.151531        0       AE005174-1      182     0       50M     *       0
       0       CAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATT      CCCFFFFFHHHHHI
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJI      AS:i:-12        XS:i:-12        XN:i:0  XM:
:
```

Counts


owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides
owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides
owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides
owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides 85x23

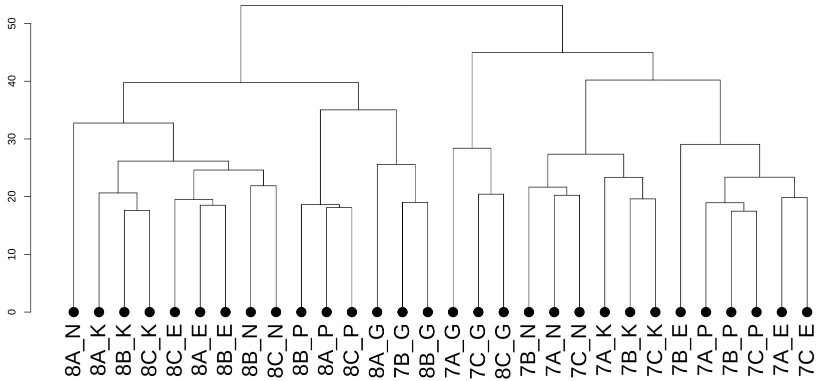
```
> head(counts(htseq),20)[,1:15]
         7A_E 7A_G 7A_K 7A_N 7A_P 7B_E 7B_G 7B_K 7B_N 7B_P 7C_E 7C_G 7C_K 7C_N 7C_P
gene0       9   17   11   17   11   12   22   20    6    9   19   20   17    5   20
gene1     108  170   97   88  173  119  241  103   51  162  155  149  124   88  128
gene10      3    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
> 
```

Downstream Statistical Analysis

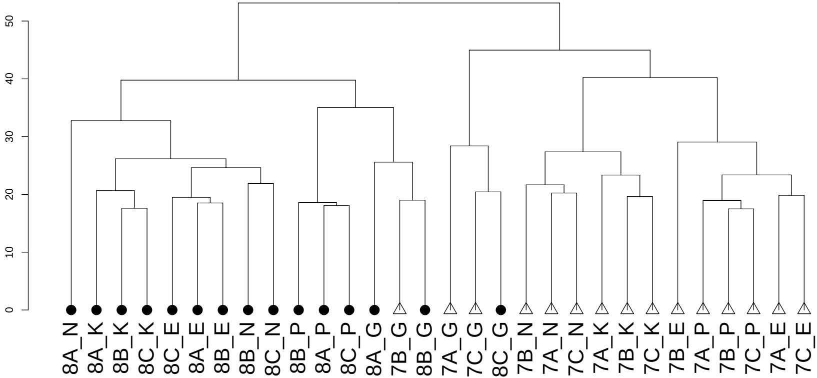

owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides
owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides
owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides
owzar001@cox: ~/CURRENT/hts-course-stat/CURRENT/Slides 85x23

```
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
> 
```

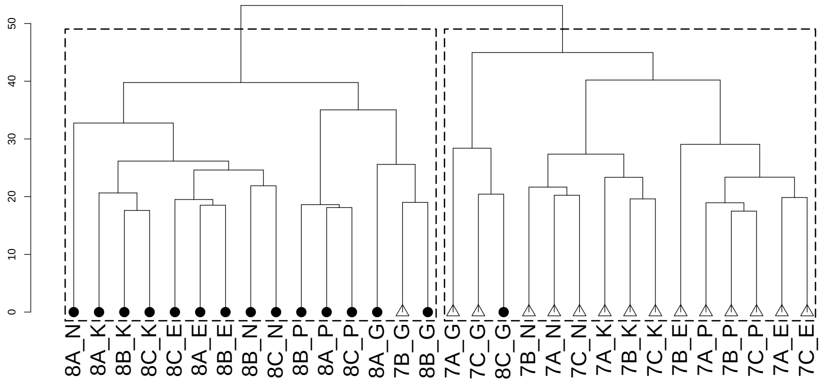
Summer 2015 Student Data



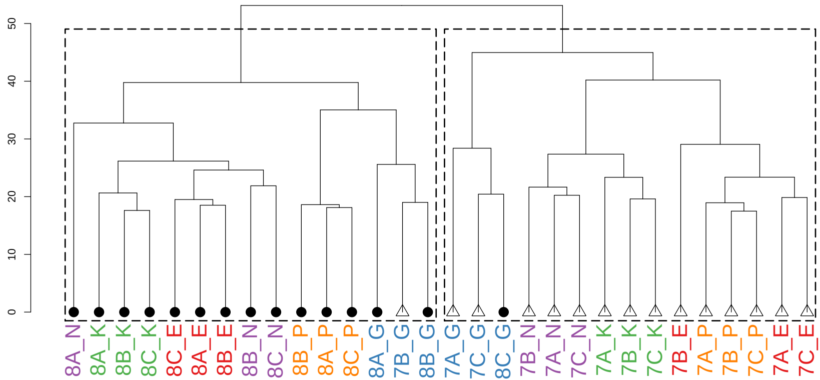
Summer 2015 Student Data



Summer 2015 Student Data



Summer 2015 Student Data



Beyond the mechanics of data analysis

- Proper lab practices for building sequencing libraries
- Computational Biology concepts and algorithms
- Introduction to ”tidy” programming
- Pre-processing and QC of raw sequencing data
- Statistics: Concepts, limitations, abuse
- Simulation and noise discovery
- Distributions for counts
- Reproducible analysis and literate programming
- Virtual computing
- Translational bioinformatics

The tidyverse approach (Data analysis task)

Task: Summarize the mean expression levels for genes 1 and 2 by mutation status (WT vs MT)

```
## # A tibble: 20 x 3
##    mutation   gene1    gene2
##    <fct>      <dbl>    <dbl>
##  1 MT       -0.381  -0.722
##  2 MT        0.202  -1.37
##  3 MT       -0.124  -0.773
##  4 WT       -0.0492 -1.06
##  5 WT       -0.227  -0.192
##  6 WT       -0.0440  0.00387
##  7 MT        1.72   -0.108
##  8 MT       -1.10   -0.288
##  9 WT        0.696   1.81
## 10 WT        2.22    0.103
## 11 MT        1.95   -0.226
## 12 MT       -1.18   -1.18
## 13 WT       -1.18   -0.281
## 14 WT       -0.874   1.12
## 15 WT        0.865   0.0713
## 16 MT       -0.268   0.277
## 17 WT        0.341  -0.00142
## 18 MT       -0.452  -0.430
## 19 MT        0.102   0.0960
## 20 WT        1.11    0.975
```

The tidyverse approach (Messy programming)

```r
x0 <- mydat[mydat$mutation == "WT", ]
x1 <- mydat[mydat$mutation == "MT", ]
# Mean expression of gene 1 in WT
mean(x0$gene1)
```

```
## [1] 0.2865462
```

```r
# Mean expression of gene 1 in MT
mean(x1$gene1)
```

```
## [1] 0.04775764
```

```r
# Mean expression of gene 2 in WT
mean(x0$gene2)
```

```
## [1] 0.2550259
```

```r
# Mean expression of gene 2 in MT
mean(x1$gene1)
```

```
## [1] 0.04775764
```

Find the error!

The tidyverse approach (Tidy programming)

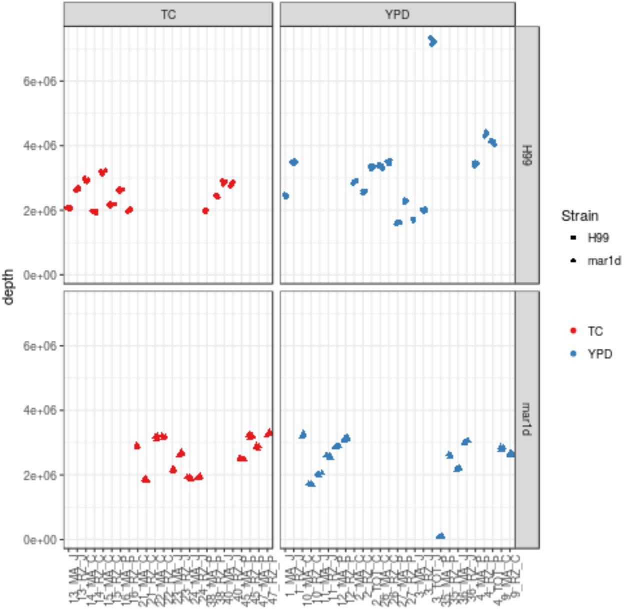
```r
mydat %>% group_by(mutation) %>% summarize_at(vars(gene1, gene2), mean)
```

```
## # A tibble: 2 x 3
##   mutation  gene1  gene2
##   <fct>     <dbl>  <dbl>
## 1 MT       0.0478 -0.472
## 2 WT       0.287   0.255
```

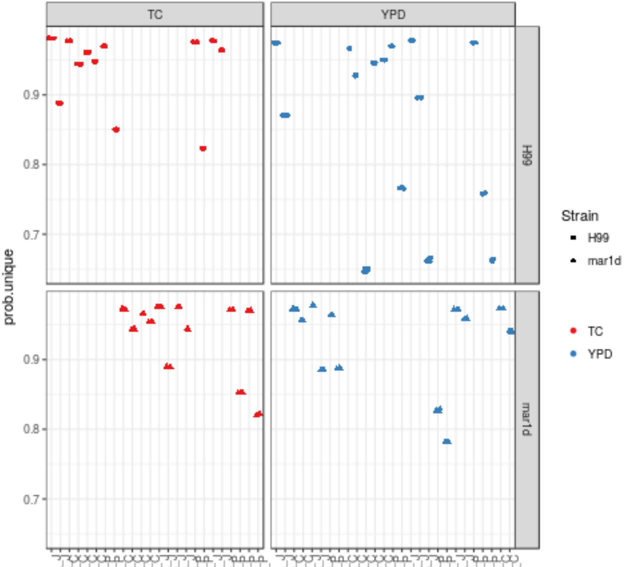
2018 Pilot Data

- *Cryptococcus neoformans*
- Experimental Design
  - Two by two factorial design
  - Factor 1: Treatment
  - Factor 2: Strain
- The experimental design will enable us to address a number of scientific questions
- The experimental design will also enable us consider methods for assessment of batch effects

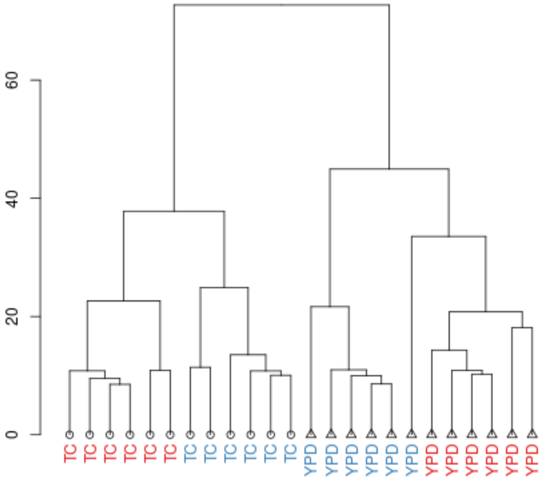
2018 Pilot Data: Sequencing Depth



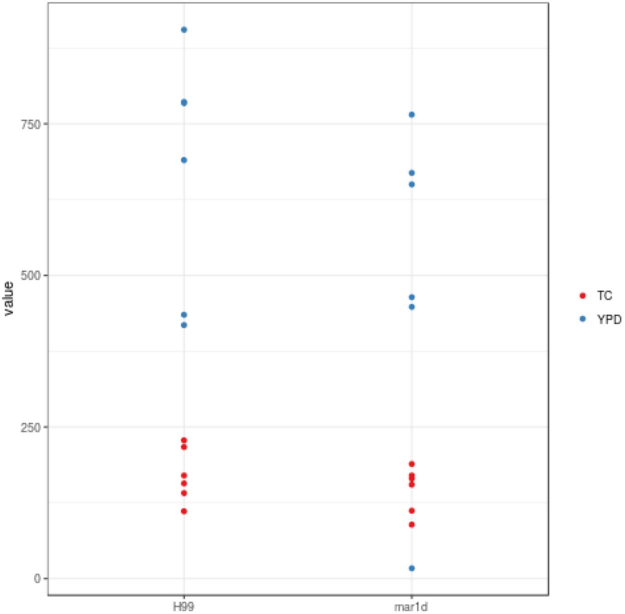
2018 Pilot Data: Proportion of unique mapped reads



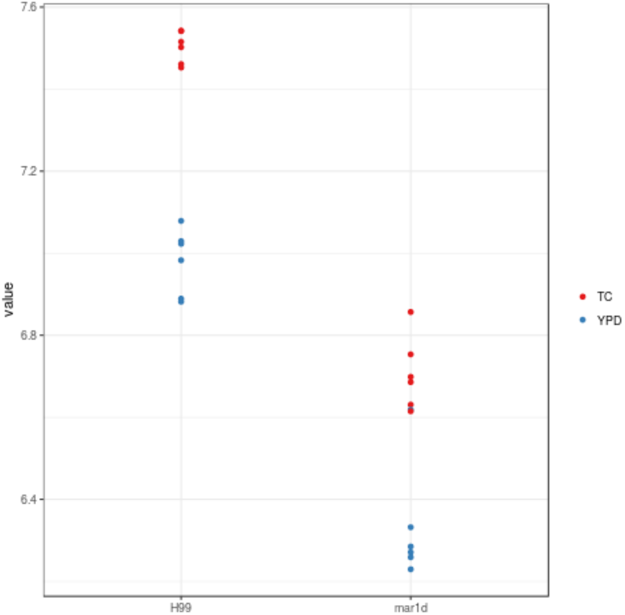
2018 Pilot Data: Dendrogram



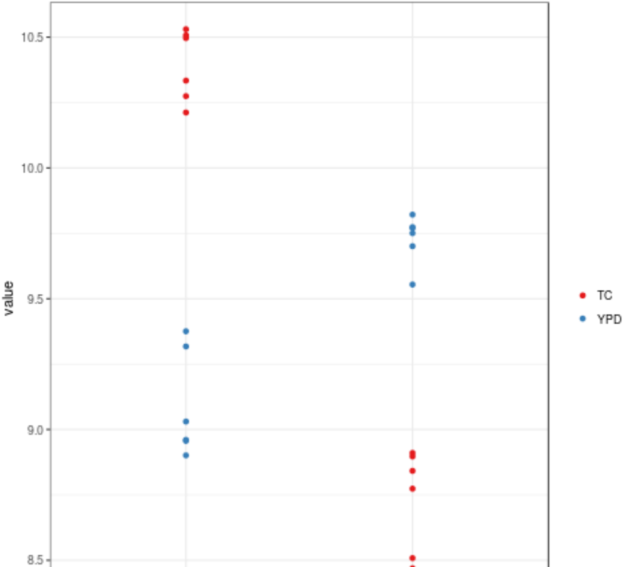
2018 Pilot Data: DE with respect to Media



2018 Pilot Data: DE with respect to Strain



2018 Pilot Data: Strain by Treatment interaction



Overview: Format

- Weeks 1, 3-5: Lectures and Workshops (Statistics, Computing, Bioinformatics, Translational Biomedical Informatics)
- Week 2: Wet lab work (build RNA library)
- Week 6: Group work/poster: data analysis, preparation and presentation
- Most statistical lectures (taught in the morning) are followed by a computing workshop (in the afternoon)
- Weekly assessments (Weeks 1-5)

Locations

1. CRTP Classroom (Hock 2nd floor; present location)
2. B&B Classroom (Hock 11025; 11th floor)
3. B&B Breakroom
4. BioSci Lab 0032/0066 (Directions have been provided)

B&B: Department of Biostatistics and Bioinformatics

Overview: Schedule

- Week 1: Thu-Fri (two days)
- Week 2: Mon-Fri (five days)
- Weeks 3-6: Mon-Thu (four days per week)
- Four sessions per day (0900-1015; 1030-1145; 1315-1430; 1445-1600)
- Lunch 1145-1315
- Locations:
  - Lectures and computing workshops: Hock CRTP Classroom
  - Wet lab work: 0032/0066 Biosci Lab
- Exceptions: 07/07 (this Friday) and 07/27 (Thursday week 4) will be moved to Hock 11025

Weekly Assessment

1. Format: 10 multiple choice or True/False questions
2. Administered during last 35 minutes of the last day of the week
3. 20 minutes for completion + 15 minutes for group feedback
4. Purpose: To help instructors *and* students identify topics and issues that need clarification
5. Improve course content and delivery for this *and* next year
6. A formal assessment is a requirement of the grant funding this course

Changes from 2017

- The course structure has been substantially revised in response to comments from student evaluations
- A two session workshop on microbiome sequencing studies has been added to the curriculum
- The data analysis practicum has been substantially expanded and revised:
    - The data analysis component will start earlier (in Week 2)
    - A workshop on pathway analysis (in addition to lectures on the topic) will be held
    - A four session guided analysis worskhop of the 2018 pilot data
    - An advanced bioinformatics workshop will be held (using packages from the Bioconductor project)

Week 1

- Virtual computing environment setup
- Introduction to the `R` statistical environment, Jupyter (iPython) notebooks and UNIX (the main computing framework for the course)
- Introduction to statistical consideration of Design of Experiments (DOE)
- Introduction to sequencing technologies
- Wetlab reproducibility
- Location: CRTP classroom

Week 2

- Lab work (RNA, library prep)
- Libraries sent to sequencing core
- Day 1:
  - Option 1: Lab: basics (0032/0066)
  - Option 2: Computing Lab (CRTP classroom)

Week 3

- Design of experiments
- Elements of statistical inference
- Unsupervised learning
- Supervised learning (aka machine learning)
- `R` graphics

Week 4

- Models for counts
- Generalized linear model for RNA-Seq
- Multiple testing
- Gene expression networks
- Reproducible analysis
- Bioinformatics computing/Computational biology
- Big Data and distributed computing

Week 5

- Translational bioinformatics
- Microbiome case study
- Human Genetics: Resources and Examples
- HTS pre-processing
- HTS pipeline
- Downstream analysis using the `DESeq2` package

Week 6

- Analysis of team data
- If time allows: Analysis of course data and pilot data
- Poster preparation
- Final presentation

Course Certificate

- There are 19 days of lectures or workshops in weeks 1 through 5
- The criteria for earning a course certificate
  - Complete all online quizzes on or before 08/09
  - The passing score for each quiz is 80% and you can retake each quiz as many times as needed.
  - Attend at least 17 out of 19 days in weeks 1 through 5
  - Fully attend the last day (08/09)
  - Actively participate in the team presentation

Dinner

- Optional group dinner on Wednesday (08/08)
- Location to be determined

Additional Resources (Hock 11th floor)

- Coffee and filtered water
- Kitchen sink
- Refigerator, microwave, toaster oven

Questions

- Ask us (don't be shy)
- Email: `htscourse@duke.edu`

Plan for today

- Quick Introduction (all)
- Questions
- Review of 2015 and 2016 experiments (Josh Granek)
- Preview of 2017 experiment (Josh Granek)
- Setup of virtual computing environment (Cliburn Chan and Janice McCarthy)
- Introduction to `R` and `UNIX` (Cliburn Chan and Janice McCarthy)
- Pizza lunch 1145-1315 (in CRTP classroom)

Acknowledgement

- A hands-on, integrative next-generation sequencing course: design, experiment, and analysis
- National Institute of Biomedical Imaging and Bioengineering (NIBIB)
- Education Projects (R25)
- 1R25EB023928-01

From all of us: Welcome!