Summer Institutes of Statistical Genetics, 2022

Module 6: GENE EXPRESSION PROFILING

Greg Gibson and Peng Qiu

Georgia Institute of Technology

Lecture 1: EXPERIMENTAL DESIGN

greg.gibson@biology.gatech.edu

http://www.cig.gatech.edu

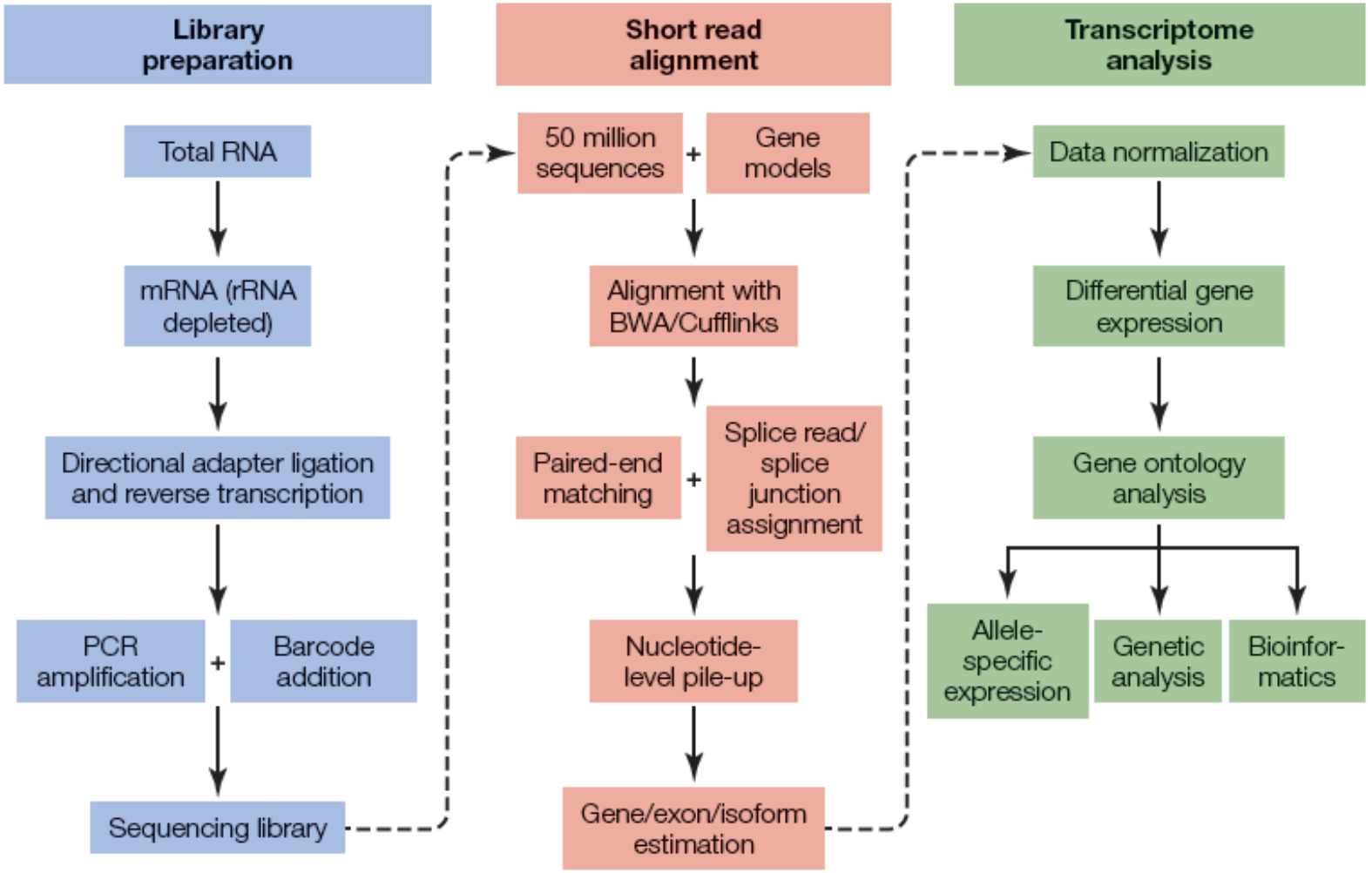
SISG Module 6 Schedule

Date	Time (PST)	Time (EST)	Topic	Instructor
Wednesday, July 13	11:30 – 12:00	2:30 – 3:00	Introductions	
	12:00 – 1:00	3:00 – 4:00	Experimental Design for Gene Expression Profiling	GG
	1:00 – 1:20	4:00 – 4:20	Break	
	1:20 – 2:20	4:20 – 5:20	Hypothesis Testing, Significance and Power	GG
	2:20 – 3:00	5:20 – 6:00	Q&A Discussions	
Thursday, July 14	8:00 – 9:00	11:00 – 12:00	Foundations of Clustering and Dimension Reduction	PQ
	9:00 – 9:20	12:00 – 12:20	Break	
	9:20 – 10:20	12:20 – 1:20	Normalization of Transcriptome Datasets	GG
	10:20 – 11:00	1:20 – 2:00	Q&A Discussions	
Thursday, July 14	12:00 – 1:00	3:00 – 4:00	Epigenomics	GG
	1:00 – 1:20	4:00 – 4:20	Break	
	1:20 – 2:20	4:20 – 5:20	Computational Challenges in scRNAseq Analysis	PQ
	2:20 – 3:00	5:20 – 6:00	Q&A Discussions	
Friday, July 15	8:00 – 9:00	11:00 – 12:00	Clustering, Trajectory Inference and Geometric Properties in scRNAseq	PQ
	9:00 – 9:20	12:00 – 12:20	Break	
	9:20 – 10:20	12:20 – 1:20	Handling the sparsity of scRNAseq Analysis	PQ
	10:20 – 11:00	1:20 – 2:00	Q&A Discussions	
Friday, July 15	12:00 – 1:00	3:00 – 4:00	eQTL and Genetics of Gene Expression	GG
	1:00 – 1:20	4:00 – 4:20	Break	
	1:20 – 2:20	4:20 – 5:20	Automated Cell Type Annotation and Interpretation	PQ
	2:20 – 3:00	5:20 – 6:00	Q&A Discussions	

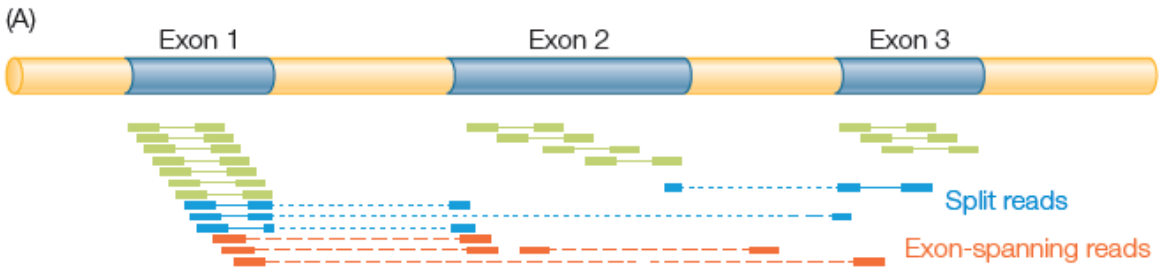
Steps in a Gene Expression Profiling Study

1. Experimental Design	(this session)
2. RNA Sequencing	(next)
3. Short read alignment	
4. Normalization	(tomorrow morning)
5. Hypothesis testing	(after the break today)
6. Downstream analyses	(Module 16)
7. Genetic analysis	(Friday afternoon)

RNAseq Workflow



Read Alignment



(B)

Reference:

```
ACGGCATTCATCCTACGCGCCATCCACTACGGCTGCTAAGCCACACCCATATACCGGC
   GGCATTCATCCTACGCGCCATCCACTACGACTGCTAAG
        TTCATCCTACGCGCCATCCACTACGGCTGCTAAGC
          CATCCTACGCGCCATCCACTACGACTGCTAAGCCA
               CTACGCGCCATCCACTACGACTGCTAAGCCACAC
                TACGCGCCATCCACTACGGCTGCTAAGCCACACCCAT
                    GCGCCATCCACTACGGCTGCTAAGCCACACCCAT
                    GCGCCATCCACTACGACTGCTAAGCCACACCCAGAT
                     CGCCATCCACTACGGCTGCTAAGCCACACCCATATAC
                     CGCCATCCACTACGGCTGCTAAGCCACACCCATATAC
                           ACTACGACTGCTAAGCCACACCCAGATACCG
                             TACGACTCCTAAGCCACACCCAGATACCGG
                                GGCTGCTAAGCCACACCCATATACCGGC
```

(C)

```
seq1  272  G  17   , . $ . , , , , , , , , , . .    < < < + ; < < < < < < ; < & & <
seq2  273  G  18   . , A , . a . a . a , T . , a T . T    < < ; ; < < < < < < & 3 ; < = & < <
seq3  274  C  18   . . , , , . , , , , , , , . $ . ,    < < ; : < < & < < & < ; < < & < = 6
seq4  275  T  16   . . , , , , , , , , , , , , , ,    < < < + < < = < < & 7 ; < < < ;
seq5  276  G  17   . , , , , , , , C , , , , , , .    < 6 < < < < < < ; < < < < & 4 <
seq6  277  C  15   . , , , , , , , , , , , , , .    < ; ; < < < < < < < < : < < <
```

Modes of Bulk RNA sequencing

RNA is prepared, mRNA is captured on polyT beads, fragmented, and converted to cDNA using either a stranded or unstranded protocol, usually with 12-24X multiplexing

1. Single-end reads
   - Maximizes the total number of independent reads (50M optimal)
   - When RNA is degraded, eg FFPE specimens

2. Paired-end reads
   - Slightly more accurate alignment
   - But typically lower coverage (25M reads)
   - Better for estimation of alternate splicing and ASE

3. 3’ targeted
   - Lexogen protocol is one fifth the cost ($70 vs $350 per sample)
   - Ideal for large sample studies when funds are a concern
   - Single Cell drop digital dd-scRNASeq is also 3’ targeted

RNAseq Software

1. Short Read Alignment
   - STAR https://github.com/alexdobin/STAR/releases
   - HISAT2 https://ccb.jhu.edu/software/hisat2/index.shtml

2. Read counting
   - HTseq http://www-huber.embl.de/HTSeq/doc/overview.html
   - SAMtools http://www.htslib.org/

3. Differential Expression
   - DESeq https://bioconductor.org/packages/release/bioc/html/DESeq2.html
   - DExSeq https://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html
   - edgeR https://bioconductor.org/packages/release/bioc/html/edgeR.html
   - Voom http://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/limma/html/voom.html

4. Data Normalization
   - SVASeq https://www.bioconductor.org/packages/release/bioc/html/sva.html
   - Combat https://www.rdocumentation.org/packages/sva/versions/3.20.0/topics/ComBat
   - PEER http://www.sanger.ac.uk/science/tools/peer
   - SNM https://www.bioconductor.org/packages/release/bioc/html/snm.html

Another option is the Tuxedo protocol (Bowtie, Tophat, Cufflinks, Cuffdiff, https://ugene.net/wiki/display/WDD31/RNA-seq+Analysis+with+Tuxedo+Tools

Notes on Alignment Biases

1. Adapter trimming can alter read alignment and counts – pay attention to all steps in workflow
2. Very short reads make it more difficult to align across introns: software biases exist in old datasets
3. Some regions of the genome are known not to align well, often due to repeats or odd GC content
4. There is a global tendency for reads to align preferentially to the reference allele
    - Consequently, average allelic expression % is closer to 60-40 (or 55-45) rather than 50-50
    - This may affect pile-ups of homologous genes that differ by just a SNP or two
5. Alignment algorithms treat things like whether the library is stranded differently

    (stranded libraries allow you to evaluate whether the Watson or Crick strand was transcribed)

Basics of Experimental Design: Levels of Replication

Often you will have a fixed budget that constrains how many arrays can be processed. So your first task is to determine what levels of replication you can afford, and how they will impact statistical power.

Technical Replication:

- RNA preparation (eg. from adjacent biopsies)
- cDNA synthesis (pooling minimizes outlier effects)
- library preparation
- sequencing lane or array hybridization (usually a minimal effect)

Biological Replication:

Fixed effects:

- sex
- treatment (drug, growth regimen, tissue)
- time of sampling (repeated measures in some cases)
- genotype (IF specifically chosen and resampled)

Random effects

- individual from a population
- field plot

Basics of Experimental Design: Specifying Contrasts of Interest

At the same time, you need to be aware of the contrasts you wish to make since by tweaking the design you may gain a lot in terms of what you can infer.

Suppose you want to compare B cells and T cells from Healthy controls and COVID-19 patients, and you have the funds to generate 24 RNASeq profiles

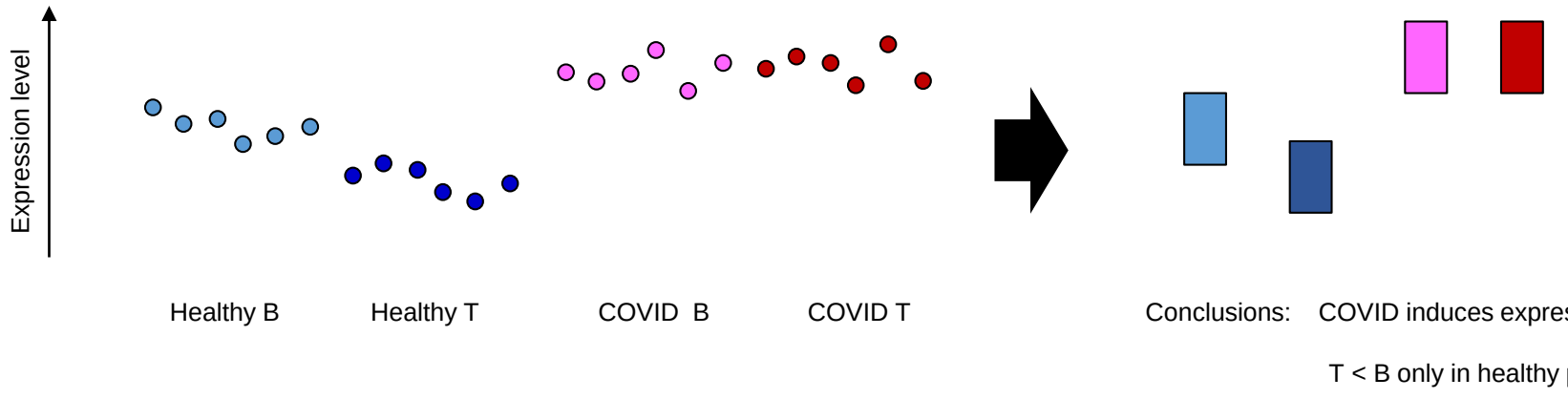
What is the best design?

- 6 controls and 6 patients, each donating both a B and a T cell sample
- 12 controls and 12 patients, each donating either a B or a T cell sample
- 3 controls and 3 patients, each donating a B and a T cell sample, processed twice
- 3 controls and 3 patients, each donating 2 B and 2 T cell samples, on separate days
- same as above, but only men or only women
- 12 controls and 12 patients, each donating either a B or a T cell sample, but pooling two visits

Main effects can only be contrasted if you have biological replicates:
reducing the number of individuals may allow you to address intra-individual variability

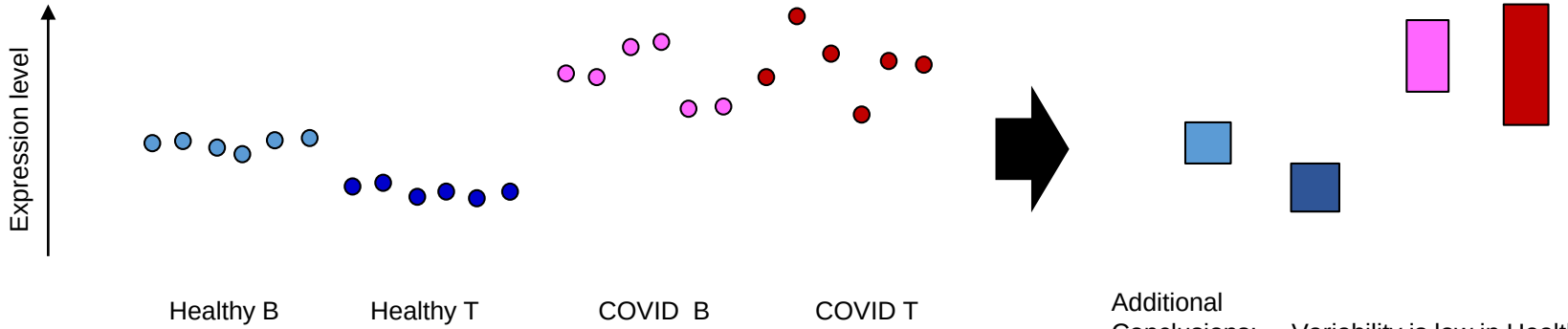
Interaction effects allow you to ask questions like whether B cells and T cells differ more between healthy volunteers or patients

Two Hypothetical Sets of Results Illustrating Design Principles



Conclusions: COVID induces expression

T < B only in healthy people

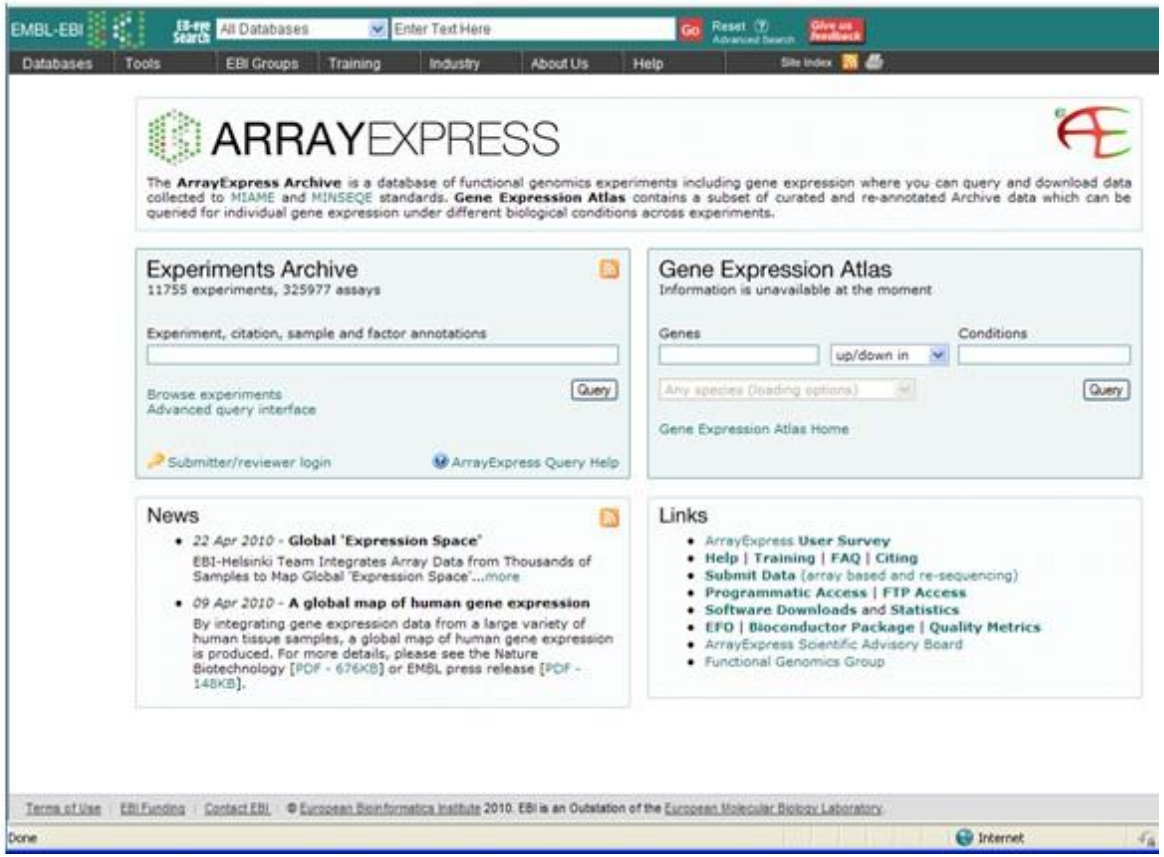


Additional Conclusions: Variability is low in Healthy controls

COVID B cell response is individualized

COVID T cell response is hypervariable

Reporting Results to Public Databases



GEOquery is R code for retrieving datasets from GEO:

https://www.bioconductor.org/packages/release/bioc/vignettes/GEOquery/inst/doc/GEOquery.html