

RNASEQ PROJECT DESIGN

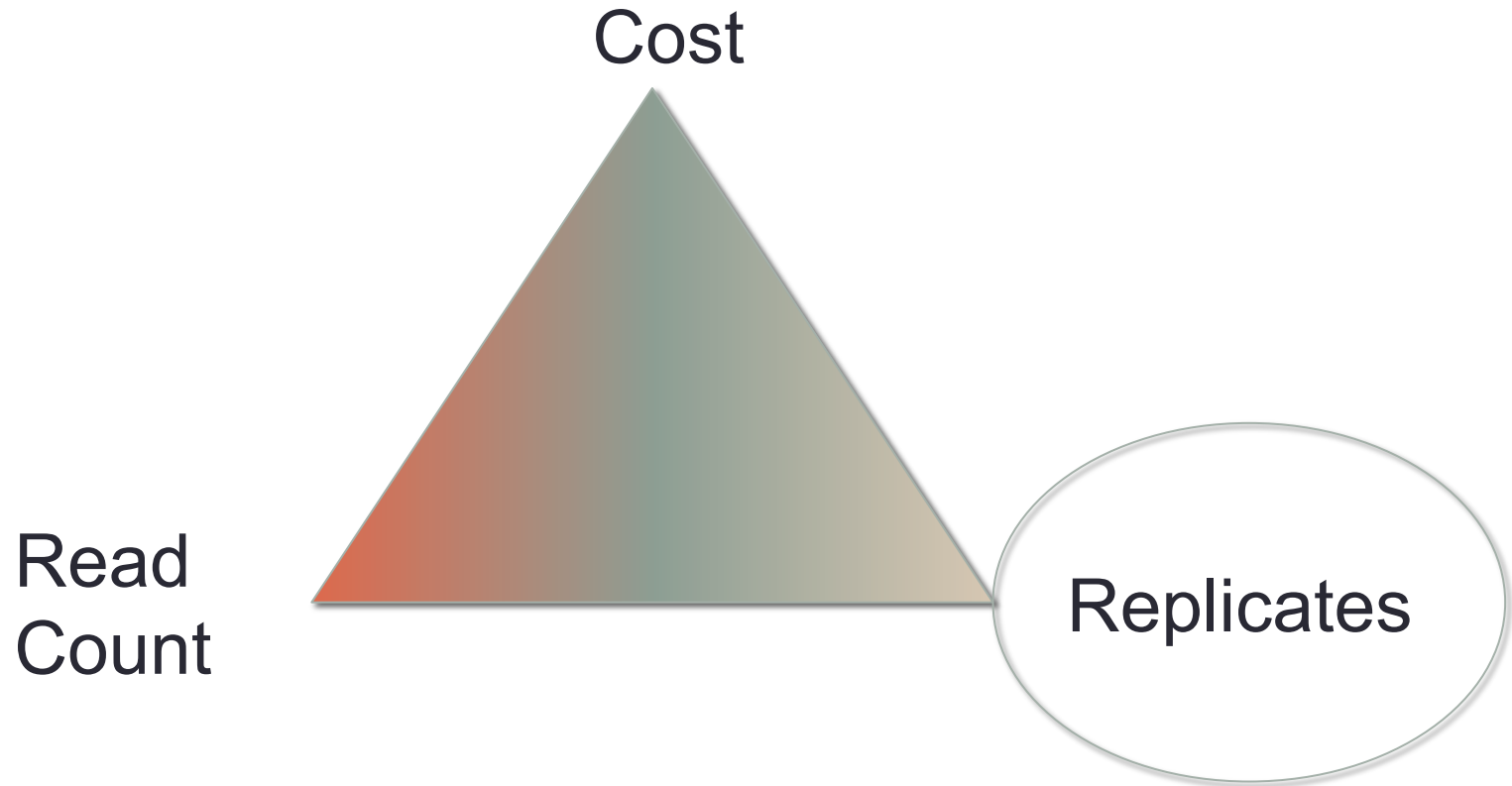
Experimental Design

Assembly in Non-Model Organisms

And other (hopefully useful) Stuff

Meg Staton
mstaton1@utk.edu
University of Tennessee
Knoxville, TN

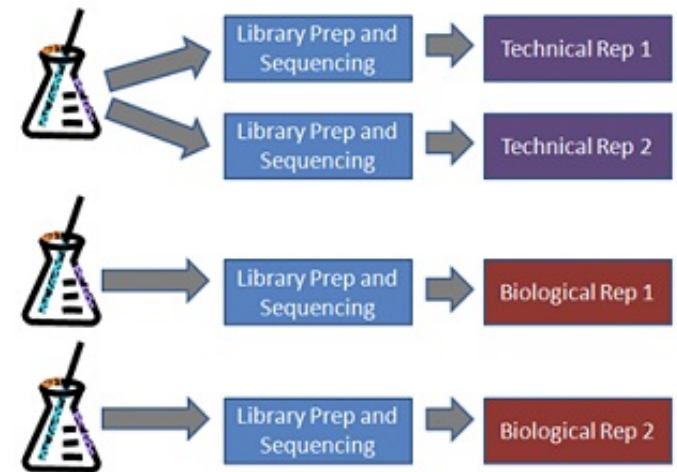
Major Considerations for Project Design



Who is your resident statistician and/or bioinformatician? Buy them a coffee and make friends. **Preferably before starting the experiment!**

Replicates

- Biological Replicates – independent biological sample, processed separately and barcoded
 - Technical Replicates – independent library construction or sequencing of the same biological sample
-
- Technical reproducibility is very good for RNASeq
 - Biological variation is much greater!



“Thinking About RNA Seq Experimental Design for Measuring Differential Gene Expression: The Basics”
<http://gkno2.tumblr.com/post/24629975632/thinking-about-rna-seq-experimental-design-for>

Replicates – How many?

- beyond a depth of 10 million reads, replicates provide more statistical power than depth for detecting differential gene expression

Liu Y, Zhou J, White KP. **RNA-seq differential expression studies: more sequence or more replication?** Bioinformatics. 2014;30(3):301-304. doi:10.1093/bioinformatics/btt688.

- Very difficult to publish with 1 rep
- Publications still coming out with 3 reps

Replicates – How many?

- The ultimate test – 48 replicates. What were the results?

How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

[Nicholas J. Schurch](#),^{1,6} [Pietá Schofield](#),^{1,2,6} [Marek Gierliński](#),^{1,2,6} [Christian Cole](#),^{1,6} [Alexander Sherstnev](#),^{1,6} [Vijender Singh](#),² [Nicola Wrobel](#),³ [Karim Gharbi](#),³ [Gordon G. Simpson](#),⁴ [Tom Owen-Hughes](#),² [Mark Blaxter](#),³ and [Geoffrey J. Barton](#)^{1,2,5}

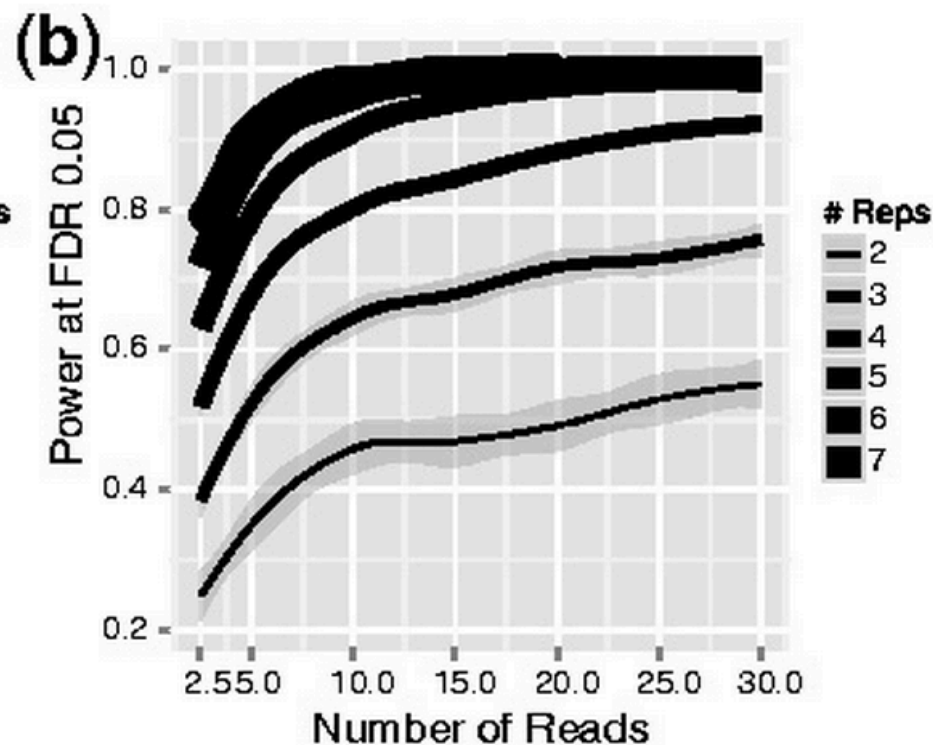
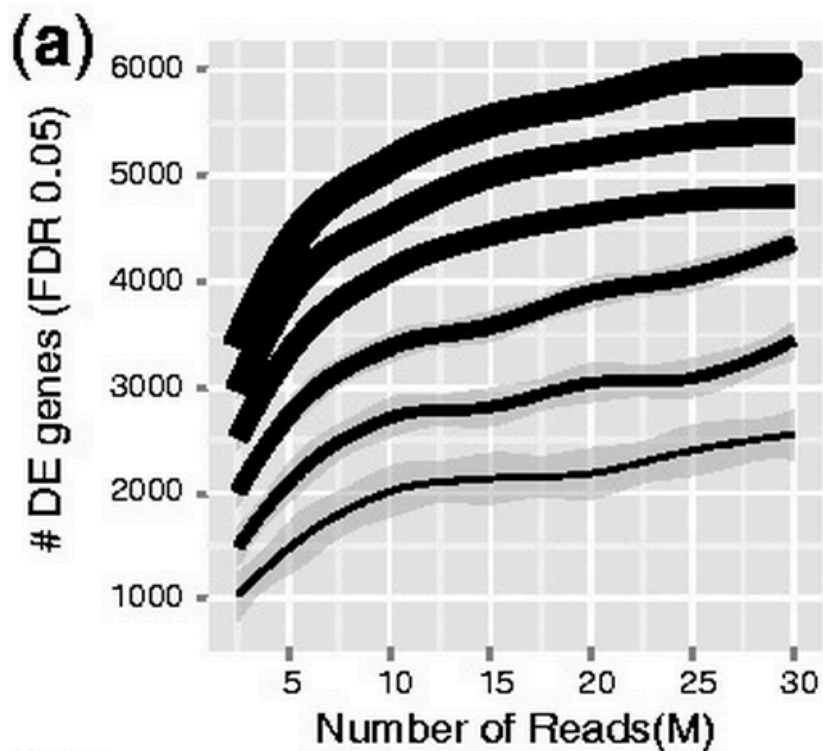
“With three biological replicates, nine of the 11 tools evaluated found only 20%–40% of the significantly differentially expressed (SDE) genes identified with the full set of 42 clean replicates.”

“these results suggest that at least six biological replicates should be used, rising to at least 12 when it is important to identify SDE genes for all fold changes”

“If fewer than 12 replicates are used, a superior combination of true positive and false positive performances makes edgeR and DESeq2 the leading tools.”

Replicates – How many?

Liu Y, Zhou J, White KP. **RNA-seq differential expression studies: more sequence or more replication?** Bioinformatics. 2014;30(3):301-304. doi:10.1093/bioinformatics/btt688.



Replicates – Software?

- Both EdgeR and DeSeq will calculate variance from replicates
- Which to use?
- From the horse's mouth:
- “Of course, we like to claim that DESeq is better than edgeR, and for only two or three replicates, I do think so, but for five or more replicates, edgeR's ‘moderation’ feature really pays off.”
-Simon Anders on SeqAnswers

From Schurch et al 2014:

“For experiments with <12 replicates per condition; use edgeR (exact) or DESeq2.

For experiments with >12 replicates per condition; use DESeq.”

Pooling

Does pooling my samples count as biological replicates?

No. With pooling, you will get an accurate mean, but not an accurate measure of variability across individuals.

Literature is mixed on this issue. But it doesn't make solid statistical sense and the downsides are significant:

“the DEGs identified in pooled samples suffered low positive predictive values” - Rajkumar et al, 2015

Blocking

- Randomized Block Design
- Divide samples (individuals) into blocks in order to control variation between blocks
- Randomize - assigning individuals at random to treatments in an experiment

	West Virginia	South Carolina
Early flowering cultivar	20	20
Late flowering cultivar	20	20

Blocking

Lane effects

- systematically bad sequencing cycles and errors in base calling

A cautionary tale

- Original paper:

Comparison of the transcriptional landscapes between human and mouse tissues

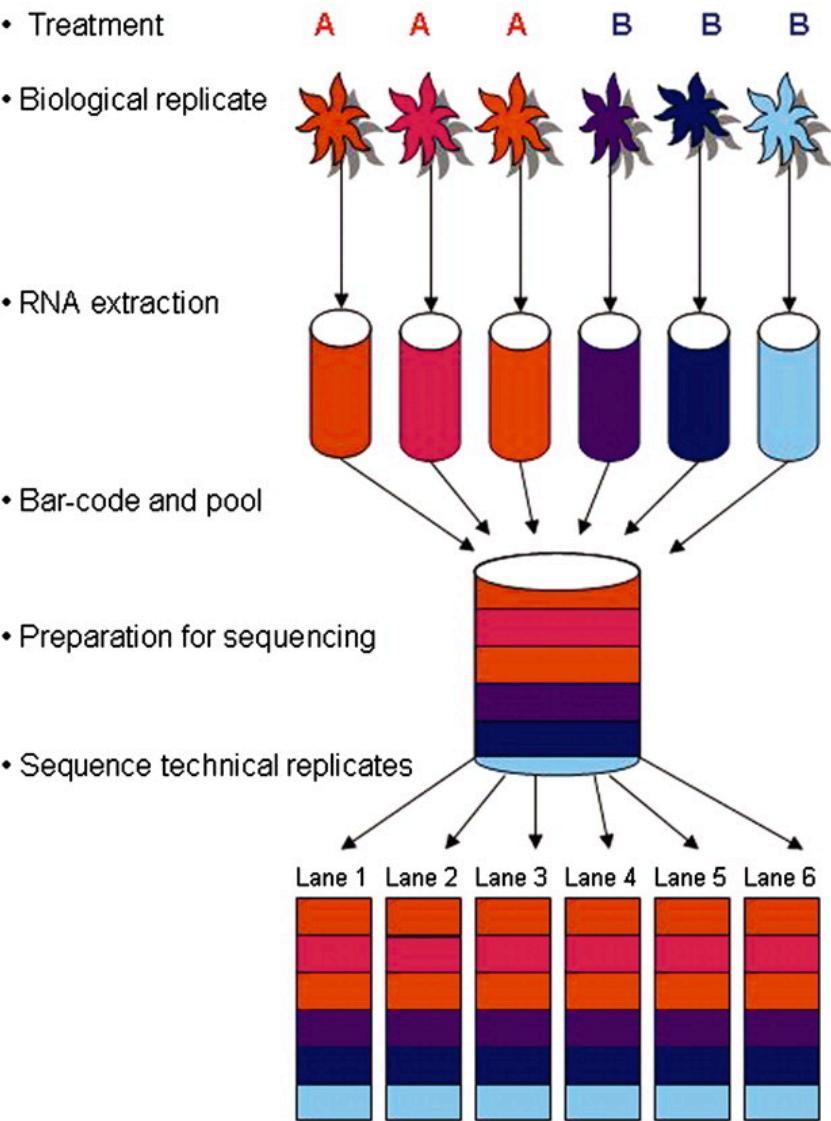
Shin Lin^{a,b,1}, Yiling Lin^{c,1}, Joseph R. Nery^d, Mark A. Urich^d, Alessandra Breschi^{e,f}, Carrie A. Davis^g,

- Reanalysis pointing out flawed statistical design and questioning results

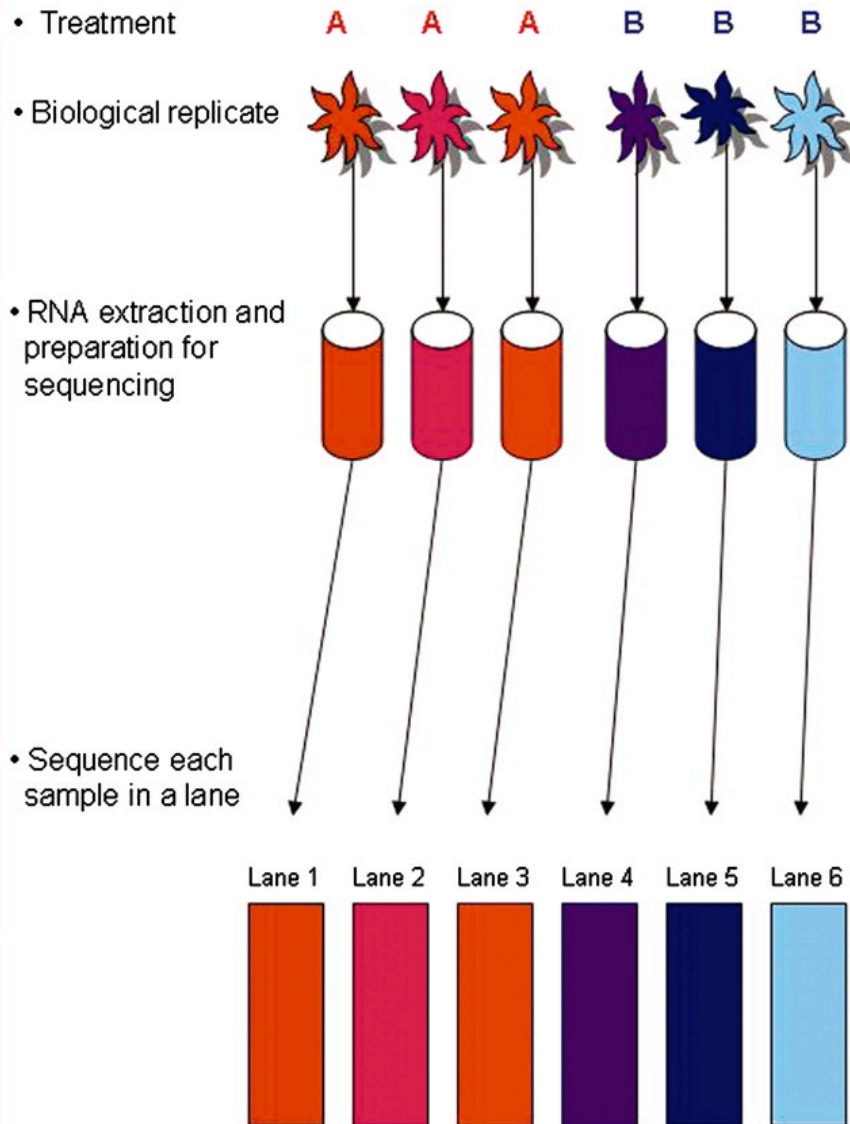
A reanalysis of mouse ENCODE comparative gene expression data [version 1; referees: 3 approved, 1 approved with reservations]

 Yoav Gilad, Orna Mizrahi-Man

Balanced Blocked Design



Confounded Design



Major Considerations for Project Design

Cost



Read
Count

Replicates

Read Count - How to Decide?

- Standards, Guidelines and Best Practices for RNA-Seq
- V1.0 (June 2011)
- The ENCODE Consortium
- What are you trying to do?
 - Compare two mRNA samples for differential expression (30M PE per sample)
 - Discover novel elements, perform more precise quantification, especially of lowly expressed transcripts (100-200M PE per sample)

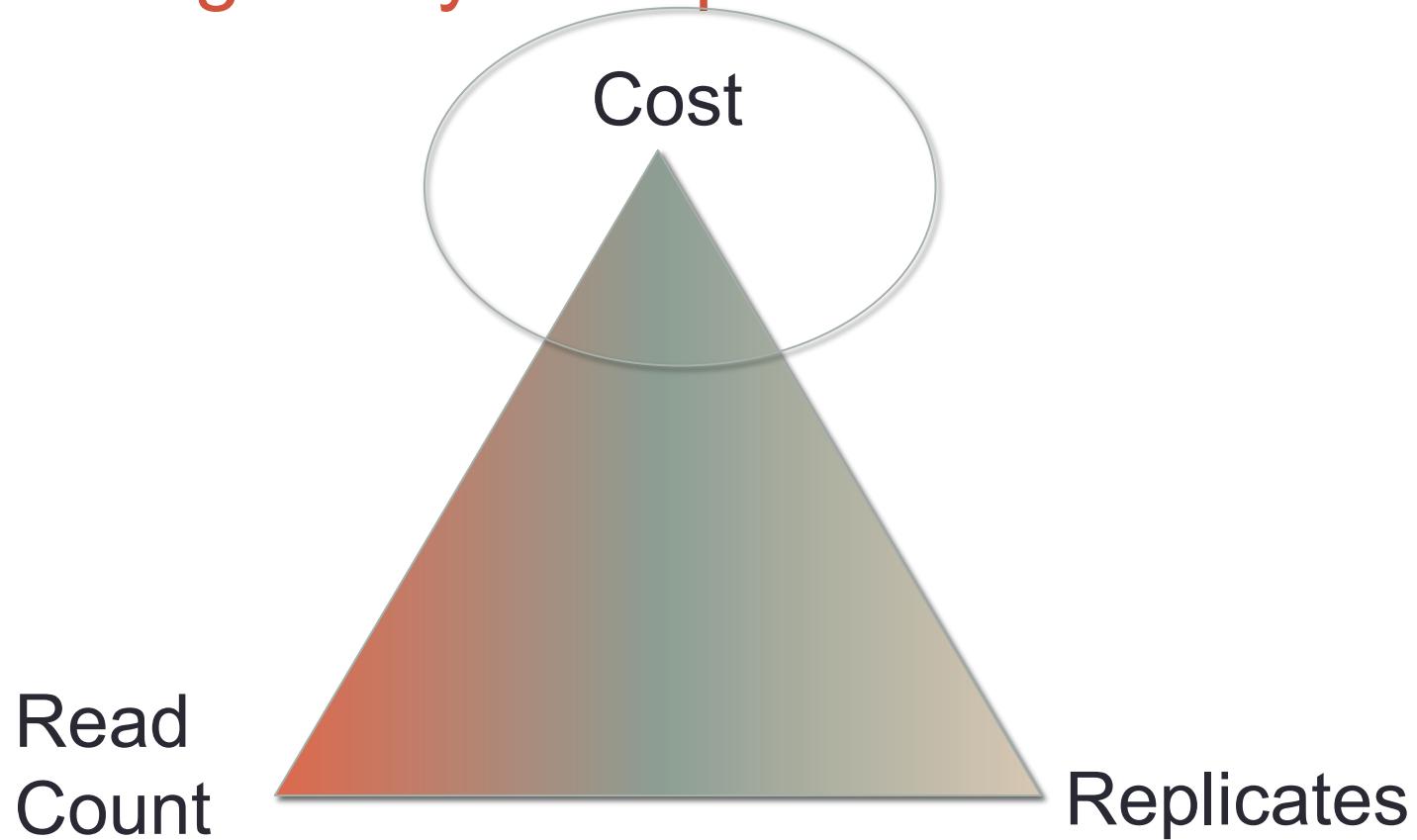
Read Count - How to Decide?

- “As low as one million reads can provide the same sequencing accuracy in transcript abundance ($r=0.99$) as >30 million reads for highly-expressed genes in all six species”
- Caveat: This only applies to the 50% most highly expressed genes
 - Lei R, Ye K, Gu Z, Sun X. (2014) Diminishing returns in next-generation sequencing (NGS) transcriptome data. *Gene* S0378-1119(14)01386-9.
- Beyond a depth of 10 million reads, replicates provide more statistical power than depth for detecting differential gene expression
 - Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*. 2014;30(3):301-304. doi:10.1093/bioinformatics/btt688.

Read Count - How to Decide?

- General recommendations:
- If you have to choose between depth and replicates, choose more replicates
- Look at what is being published in your community
- What resources do you already have?
 - Well assembled and annotated genomes – save money by using single ends, shorter reads
 - De novo transcriptome assembly – longer reads, paired ends

What's right for your experiment?



Publicly Posted Pricing

UT Genomics Core

http://mbrf.utk.edu/miseq_cost.php

UTexas Austin Genomic Sequencing and Analysis Facility

<https://wikis.utexas.edu/display/GSAF/Pricing+and+Service+Descriptions>

Science Exchange

<https://www.scienceexchange.com/services/illumina-ngs>

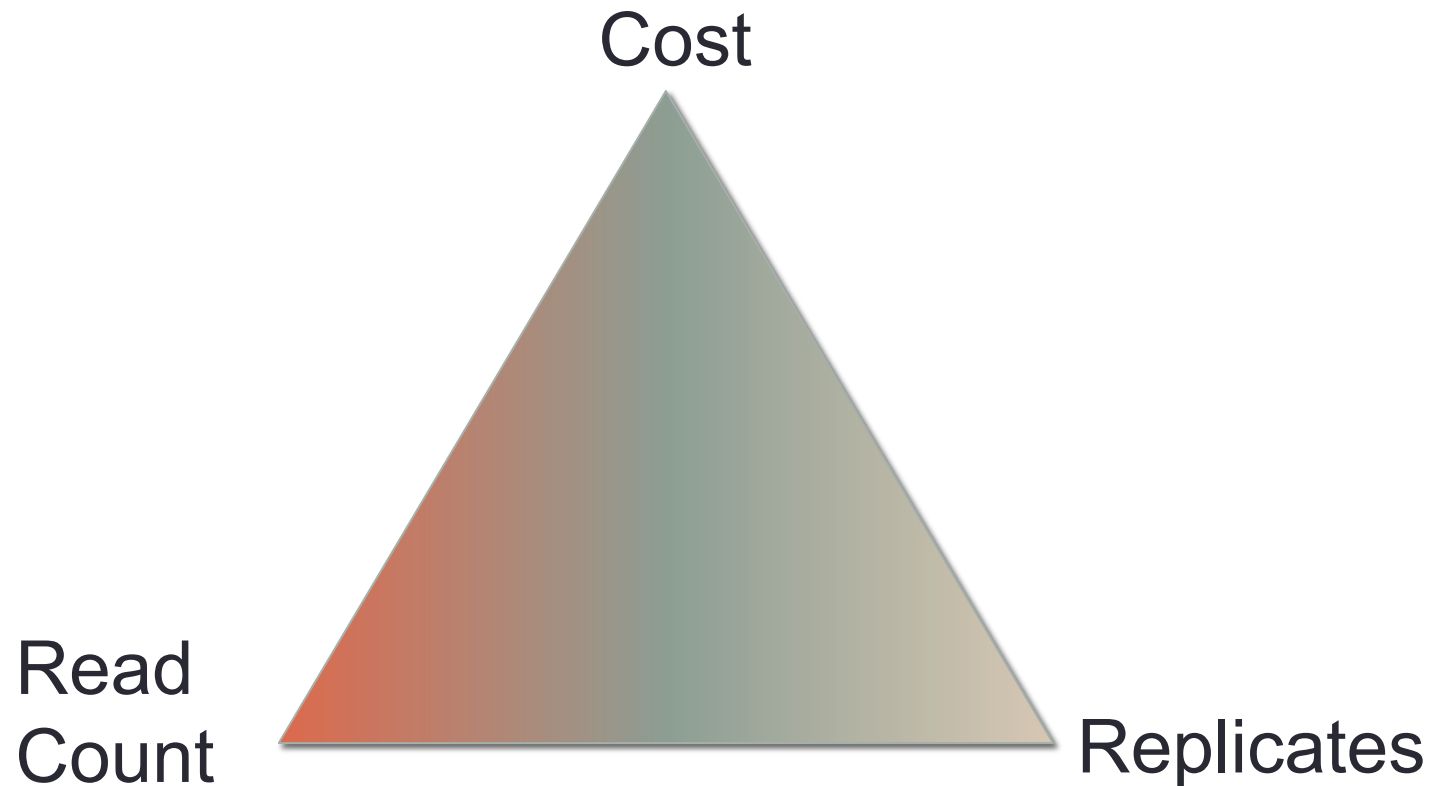
Cornell University Institute of Biotechnology

<http://www.biotech.cornell.edu/brc/genomics/services/price-list>

UC Davis Genome Center

<http://dnatech.genomecenter.ucdavis.edu/prices/>

Major Considerations for Project Design



Pro Tip: Who is your resident statistician? Buy them a coffee and make friends.

Sam, Bam and Cram format

SAM Format

- SAM = Sequence Alignment/Map format
- Tab delimited plain text
- Store large nucleotide sequence alignments
 - Alignment of every read
 - Including gaps, SNPs and structural variants
 - Pairing of reads
 - Can record more than one alignment location in the genome
 - Stores quality values
 - Stores information about duplication

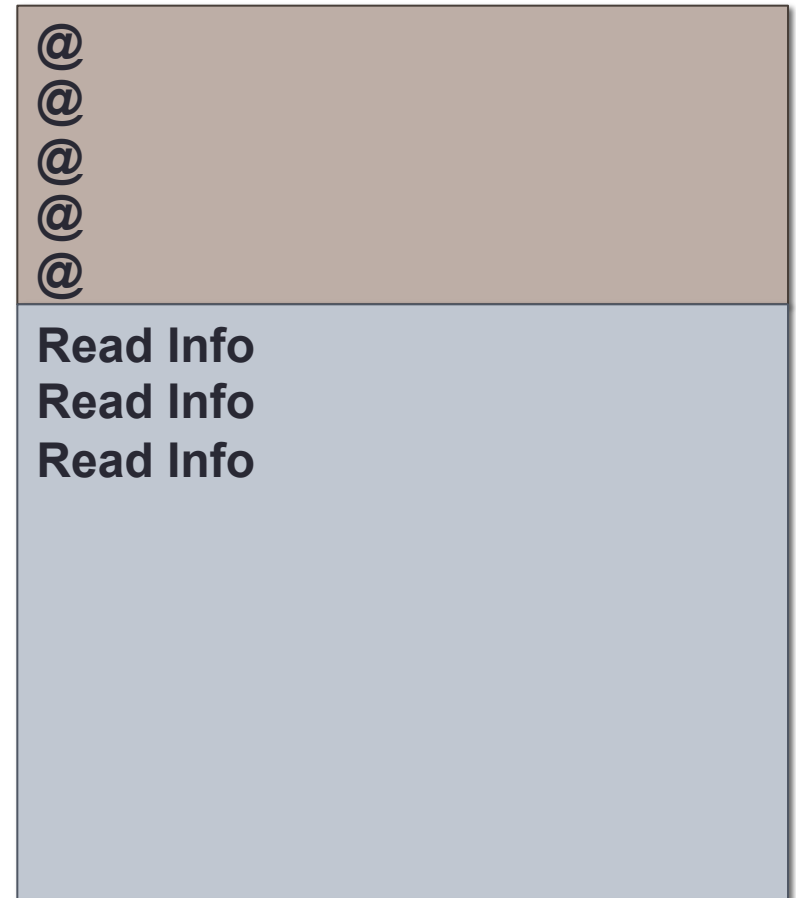
SAM Format

Strengths

- Flexible
- Useful for operations on very large sequences
- Extremely detailed documentation
 - <https://samtools.github.io/hts-specs/SAMv1.pdf>
- Manipulations can be done with the software samtools

SAM - Header

- Structure
 - Optional Header at top of file
 - Alignment information



SAM - Header

- Header lines start with @ symbol
- Always at top of file
- Contain lots of information about what was mapped, what it was mapped to, and how (**metadata**)
 - the version information for the SAM/BAM file
 - whether or not and how the file is sorted
 - information about the reference sequences
 - any processing that was used to generate the various reads in the file
 - software version

Alignment Line

- Below the headers are the alignment records
 - Tab-delimited fields
-
- | | | |
|-------|-------|--|
| • 1 | QNAME | Query template/pair NAME |
| • 2 | FLAG | bitwise FLAG |
| • 3 | RNAME | Reference sequence NAME |
| • 4 | POS | 1-based leftmost POSition/coordinate of clipped sequence |
| • 5 | MAPQ | MAPping Quality (Phred-scaled) |
| • 6 | CIGAR | extended CIGAR string |
| • 7 | MRNM | Mate Reference sequence NaMe ('=' if same as RNAME) |
| • 8 | MPOS | 1-based Mate POSition |
| • 9 | TLEN | inferred Template LENgth (insert size) |
| • 10 | SEQ | query SEQUENCE on the same strand as the reference |
| • 11 | QUAL | query QUALity (ASCII-33 gives the Phred base quality) |
| • 12+ | OPT | variable OPTional fields in the format TAG:VTYPE:VALUE |

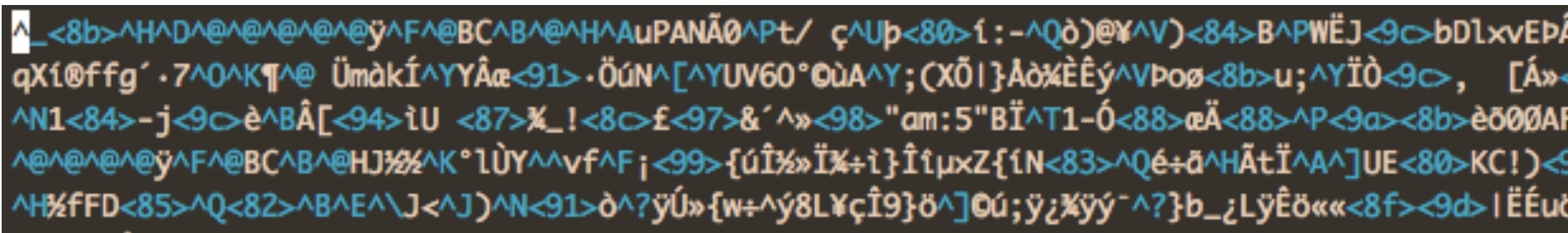
SAM

Example sam with one read:

```
@SQ SN:gi|254160123|ref|NC_012967.1| LN:4629812
@PG ID:bwa PN:bwa VN:0.7.12-r1039 CL:/lustre/
projects/rnaseq_ws/apps/bwa-0.7.12/bwa sampe ../raw_data/
NC_012967.1.fasta aln_SRR030257_1.sai
aln_SRR030257_2.sai ../raw_data/SRR030257_1.fastq ../
raw_data/SRR030257_2.fastq
SRR030257.1 99 gi|254160123|ref|NC_012967.1|
950180 60 36M = 950295 151
TTACACTCCTGTTAATCCATACAGCAACAGTATTGG
AAA;A;AA?A?AAAAA?;?A?1A;;????566)=*1
XT:A:U NM:i:1 SM:i:37 AM:i:25 X0:i:1 X1:i:0
XM:i:1 XO:i:0 XG:i:0 MD:Z:32C3
```

BAM Format

- Sister format to SAM
- BAM – Binary version of SAM
- compressed **BGZF** (Blocked GNU Zip Format) - a variant of GZIP (GNU ZIP),
- files are bigger than GZIP files, but they are much faster for random access
- Can index and then look up information embedded in the file with decompressing the whole file
- up to 75% smaller in size
- Not readable by people

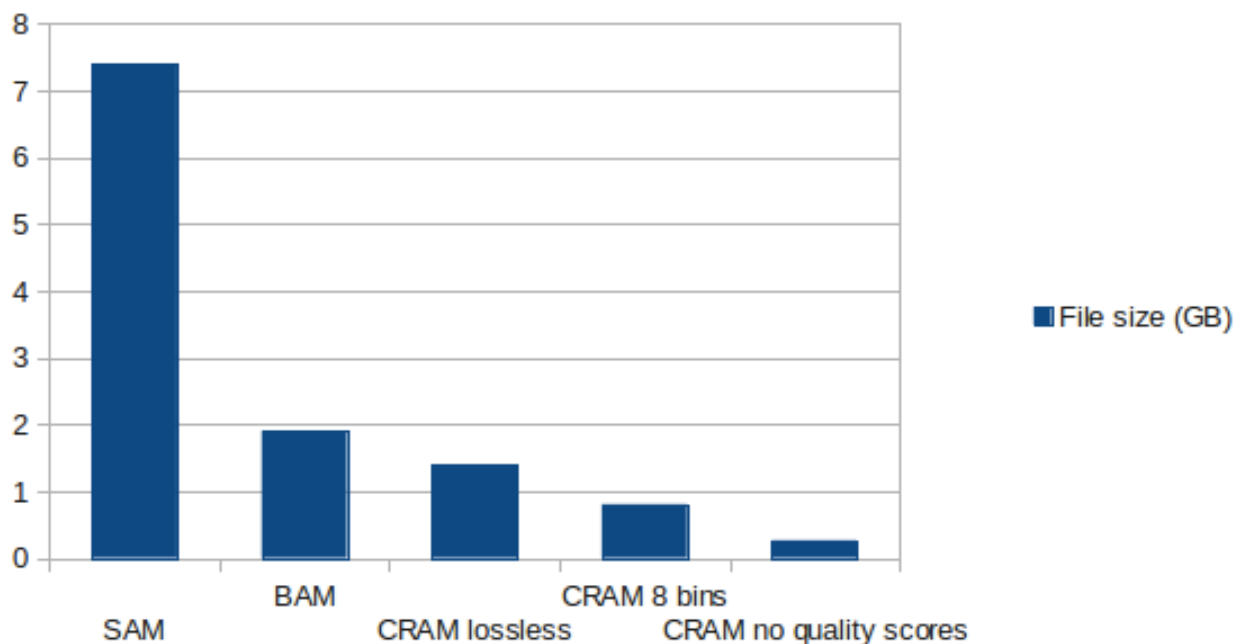


^_<8b>^H^D^@^@^@^@^@y^F^@BC^B^@^H^A^uPAN^0^P^t/ ç^Up<80>í:-^Qð)^@Y^V)^<84>B^PWËJ<9C>bDl^xvE^P^
qXí@ffg'.7^0^K^T^@ ÜmàkÍ^YY^Â^æ<91>·ÖúN^[^YUV60°0ùA^Y;(XÖI}Àð%ÈÈý^V^pø<8b>u;^YÏÒ<9C>, [Á»
^N1<84>-j<9C>è^B^Â[<94>iU <87>%_!<8C>f<97>&'^»<98>"am:5"BÏ^T1-Ó<88>æÄ<88>^P<9a><8b>èð00AF
^@^@^@^@y^F^@BC^B^@HJ}%^K°lÙY^^vf^F; <99>{úÎ%»Î%+i}ÎîµxZ{îN<83>^Qé+ð^H^tÎ^A^]UE<80>KC!)<9
^H%FFD<85>^Q<82>^B^E^^\J<^J)^N<91>ð^?ýÚ»{w+^ý8L^çÎ9}ð^]0ú;ÿ¿%ÿý^-^?}b_¿LÿÊö««<8f><9d>|ËÉuð

CRAM

- Introduced in 2011 by EMBL/EBI
- Even smaller and more efficient than BAM files
- Rare

EBI has a cram toolkit
<https://www.ebi.ac.uk/ena/software/cram-toolkit>



Fritz, Markus Hsi-Yang, et al. "Efficient storage of high throughput DNA sequencing data using reference-based compression." *Genome research* 21.5 (2011): 734-740.

<http://www.uppmax.uu.se/using-cram-to-compress-bam-files-on-uppnex>

Software - samtools

Samtools Home Download ▾ Workflows ▾ Documentation ▾ Support ▾

Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

Samtools	Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
BCFtools	Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
HTSlib	A C library for reading/writing high-throughput sequencing data

Samtools and BCFtools both use HTSlib internally, but these source packages contain their own copies of htslib so they can be built independently.

<http://www.htslib.org/>