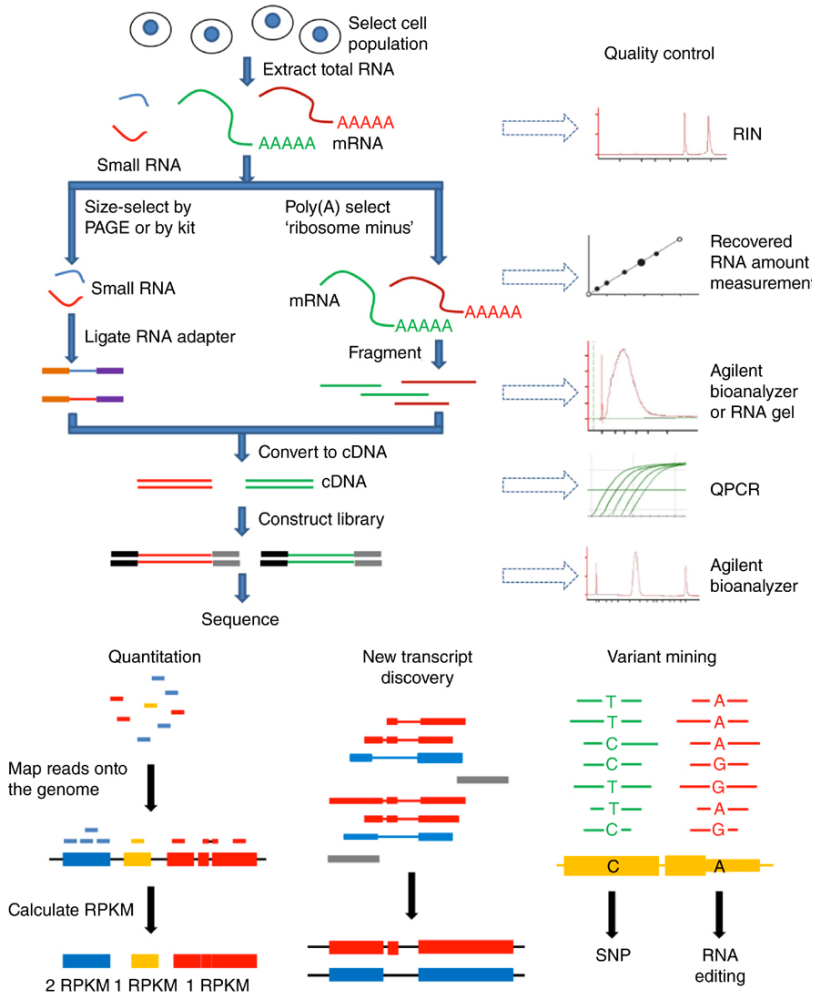
Intro to Gene Annotations and RNA-seq

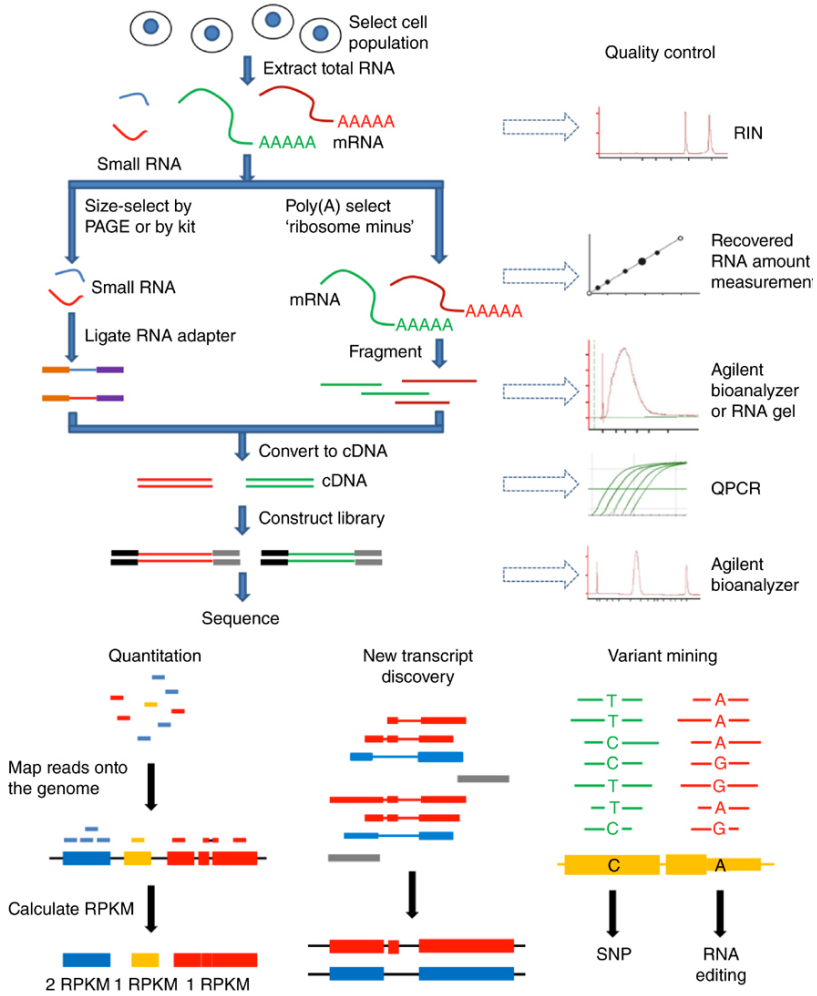
Overview

- Overview of RNA-seq
  - Alignment
  - Visualization
- Genome Annotations
  - Coordinate Systems
  - Introduction to File Formats
  - Explore GTF File
- RNA-seq analysis
  - Quantification
  - Differential Expression
- Astrocyte Workflows

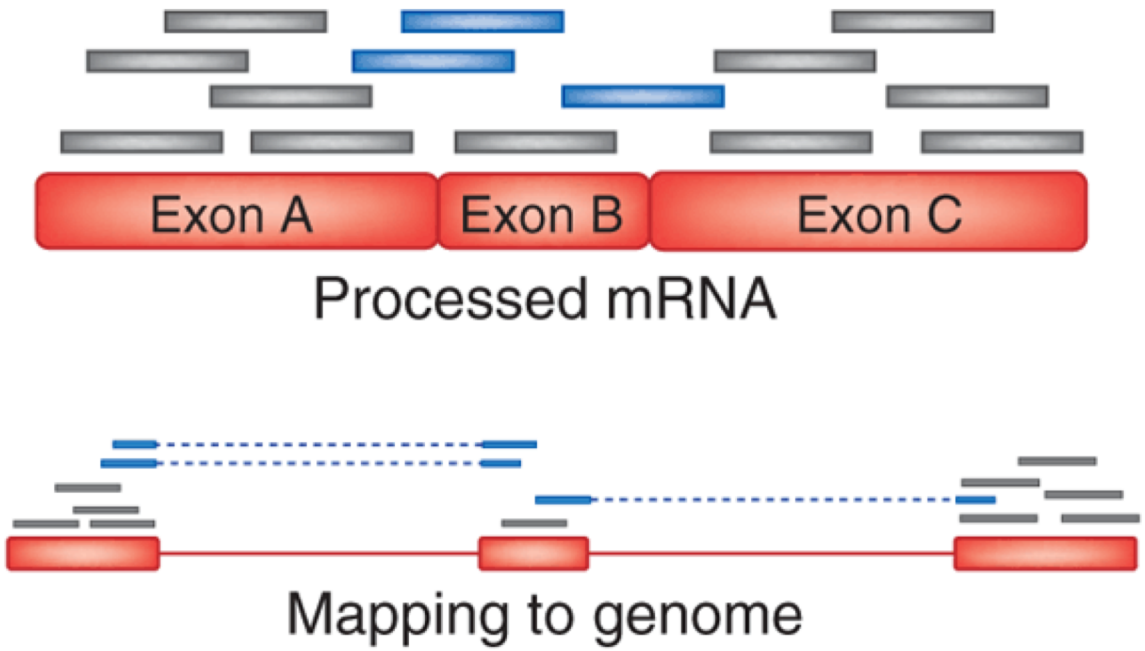


Intro to RNA-seq and review

Overview of RNA-seq

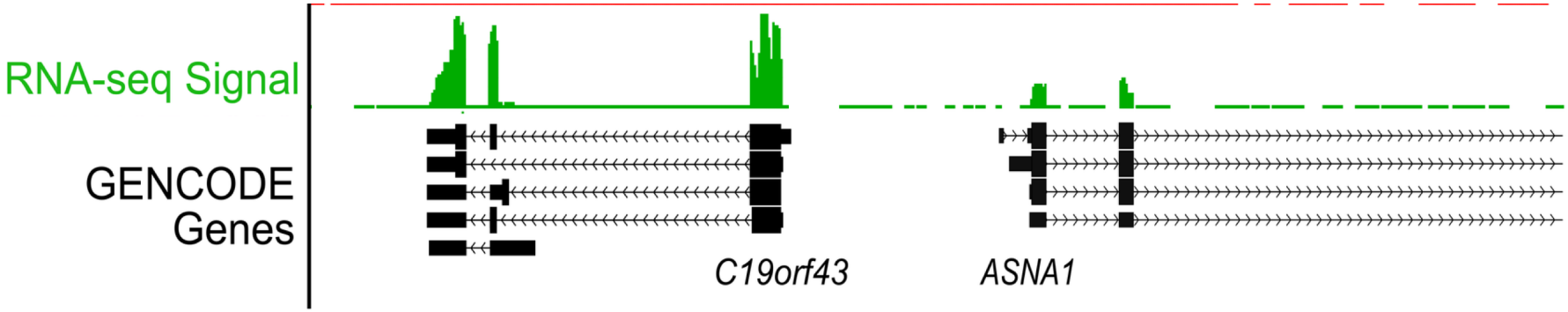


Alignment to Genome and Reference Annotations



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Visualization of Alignment



Review of Alignment and Intro to RNA-seq

What is a genome annotation?

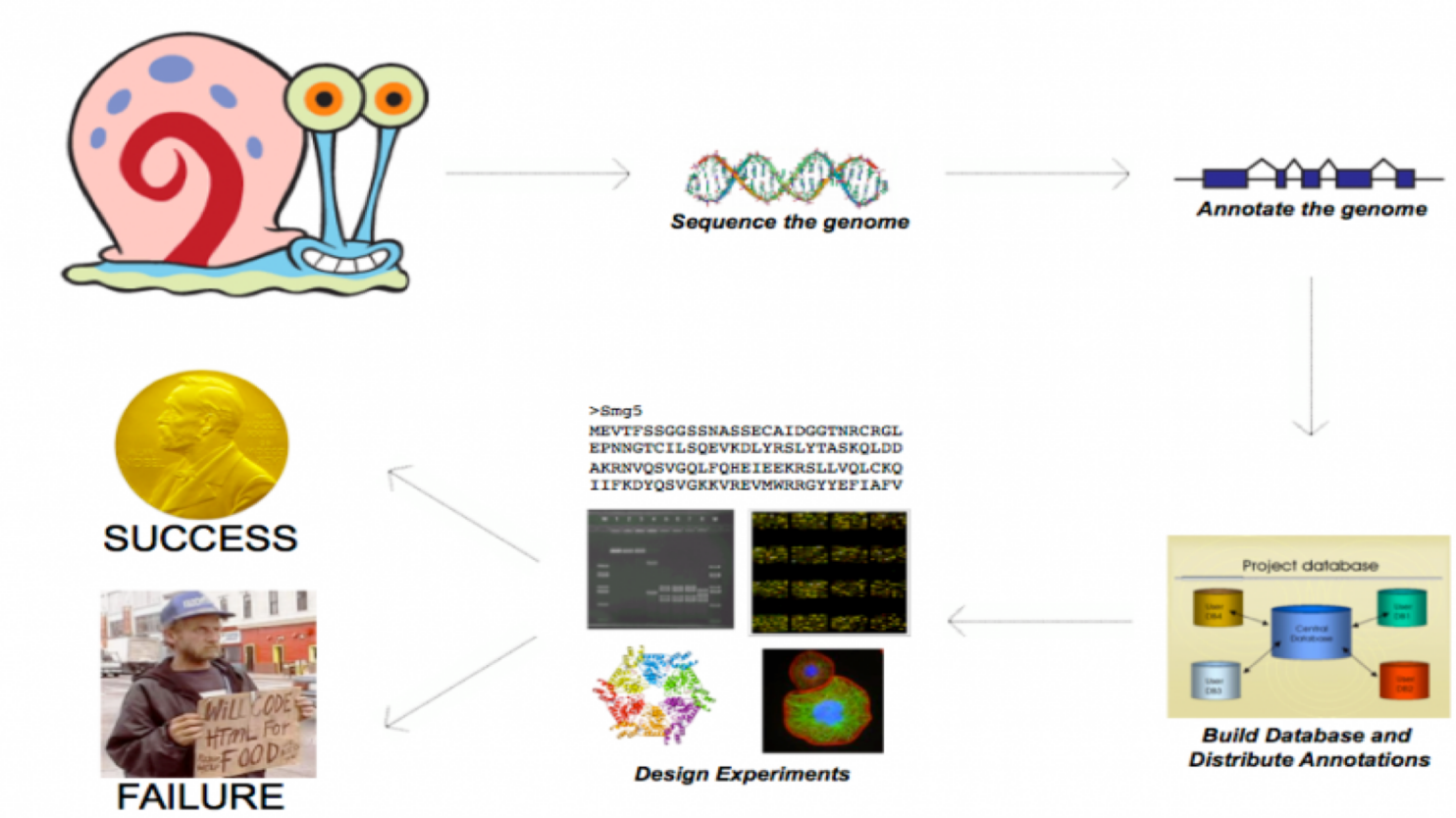
CTGCCGTCTGCTGCCATCGGAGCCCAAAGCCGGGCTGTGACTGCTCAGAC
CAGCCGGCTGGAGGGAGGGGCTCAGCAGGTCTGGCTTTGGCCCTGGGAGA
GCAGGTGGAAGATCAGGCAGGCCATCGCTGCCACAGAACCCAGTGGATTG
GCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCTCTGGGTGGTAGGTGC
AGAGACGGGAGGGGCAGAGCCGCAGGCACAGCCAAGAGGGCTGAAGAAAT
GGTAGAACGGAGCAGCTGGTGATGTGTGGGCCCACCGGCCCCAGGCTCCT
GTCTCCCCCCAGGTGTGTGGTGATGCCAGGCATGCCCTTCCCCAGCATCA
GGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTG
AGTGTCCCCAGTGTTGCAGAGGTGAGAGGAGAGTAGACAGTGAGTGGGAG
TGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAG
AGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCATCT
GGAGCCCTGCTGCTTGCGGTGGCCTATAAAGCCTCCTAGTCTGGCTCCAA
GGCCTGGCAGAGTCTTTCCCAGGGAAAGCTACAAGCAGCAAACAGTCTGC
ATGGGTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGGCCCCCA
AGAAAGGCTCTGGTGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCAT
AGGCAAGCCTGGCTGCCTCCAGCTGGGTCGACAGACAGGGGCTGGAGAAG
GGGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGA
GGAAGGAGAAGGGGATGCACTGTTGGGGAGGCAGCTGTAACTCAAAGCCT
TAGCCTCTGTTCCCACGAAGGCAGGGCCATCAGGCACCAAAGGGATTCTG
CCAGCATAGTGCTCCTGGACCAGTGATACACCCGGCACCCTGTCCTGGAC
ACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGG
TTCTGCCATTGCTGCTGTGTGGAAGTTCACTCCTGCCTTTTCCTTTCCCT
AGAGCCTCCACCACCCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCC
AGTTTCACCAGAAGTAGGCCTCTTCCTGACAGGCAGCTGCACCACTGCCT
GGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATG

What is an annotation?

- Gene
- Repeat
- Disease Locus
- Something interesting

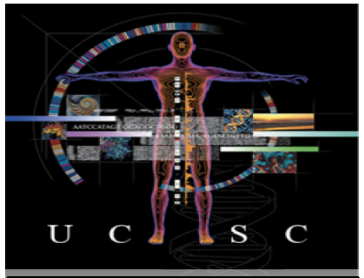
CTGCCGTCTGCTGCCATCGGAGCCCAAAGCCGGGCTGTGACTGCTCAGAC
CAGCCGGCTGGAGGGAGGGGCTCAGCAGGTCTGGCTTTGGCCCTGGGAGA
GCAGGTGGAAGATCAGGCAGGCCATCGCTGCCACAGAACCCAGTGGATTG
GCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCTCTGGGTGGTAGGTGC
AGAGACGGGAGGGGCAGAGCCGCAGGCACAGCCAAGAGGGCTGAAGAAAT
GGTAGAACGGAGCAGCTGGTGATGTGTGGGCCCACCGGCCCCAGGCTCCT
GTCTCCCCCCAGGTGTGTGGTGATGCCAGGCATGCCCTTCCCCAGCATCA
GGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTG
AGTGTCCCCAGTGTTGCAGAGGTGAGAGGAGAGTAGACAGTGAGTGGGAG
TGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAG
AGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCATCT
GGAGCCCTGCTGCTTG**CGGTGGCCTATAAAGCCTCCTAGTCTGGCTCCAA**
GGCCTGGCAGAGTCTTTCCCAGGGAAAGCTACAAGCAGCAAACAGTCTGC
ATGGGTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGGCCCCCA
AGAAAGGCTCTGGTGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCAT
AGGCAAGCCTGGCTGCCTCCAGCTGGGTCGACAGACAGGGGCTGGAGAAG
GGGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGA
GGAAGGAGAAGGGGATGCACTGTTGGGGAGGCAGCTGTAACTCAAAGCCT
TAGCCTCTGTTCCCACGAAGGCAGGGCCATCAGGCACCAAAGGGATTCTG
CCAGCATAGTGCTCCTGGACCAGTGATACACCCGGCACCCTGTCCTGGAC
ACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGG
TTCTGCCATTGCTGCTGTGTGGAAGTTCACTCCTGCCTTTTCCTTTCCCT
AGAGCCTCCACCACCCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCC
AGTTTCACCAGAAGTAGGCCTCTTCCTGACAGGCAGCTGCACCACTGCCT
GGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATG

Why are genome annotations important?



Slide from Mark Yandell

Annotations Provide Context



Conservation
Repeat elements
Genome Gaps
Cytobands
Gene annotations
"Mappability"
DeCIPHER
ISGA

dbSNP
Short Genetic Variations

gnomAD browser

ClinVar

OMIM
Online Mendelian Inheritance in Man

1000 Genomes
A Deep Catalog of Human Genetic Variation

Genetic variation

...CCTCATGCATGGAAA...
...CCTCATGTATGGAAA...
...CCTCATGCATGGAAA...
...CCTCATGCATGGAAA...
...CCTCATGTATGGAAA...
...CCTCATGCATGGAAA...
...CCTCATGTATGGAAA...

ENCODE

Chromatin marks
DNA methylation
RNA expression
TF binding

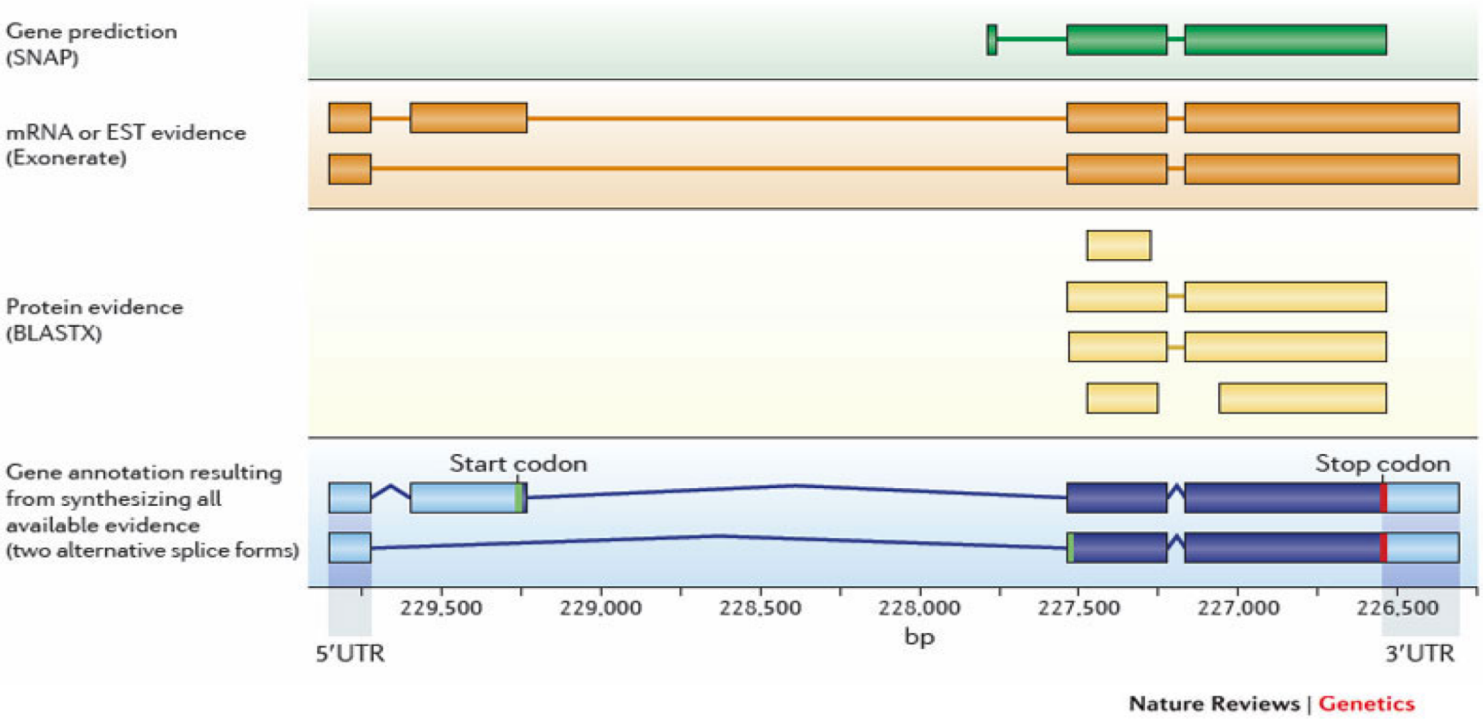
Pfam

KEGG
Kyoto Encyclopedia of Genes and Genomes

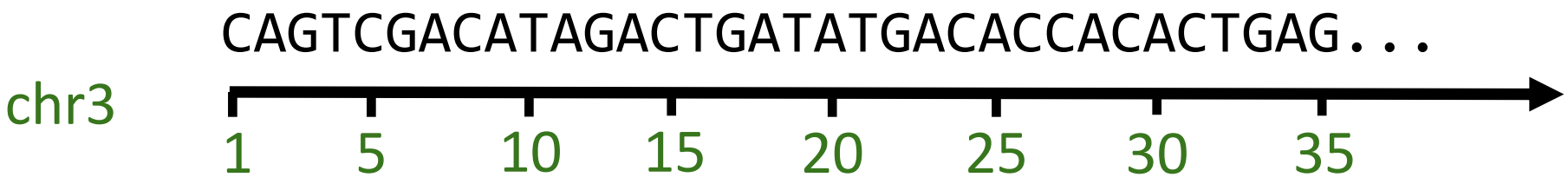
HGMD®

Human Protein Reference Database

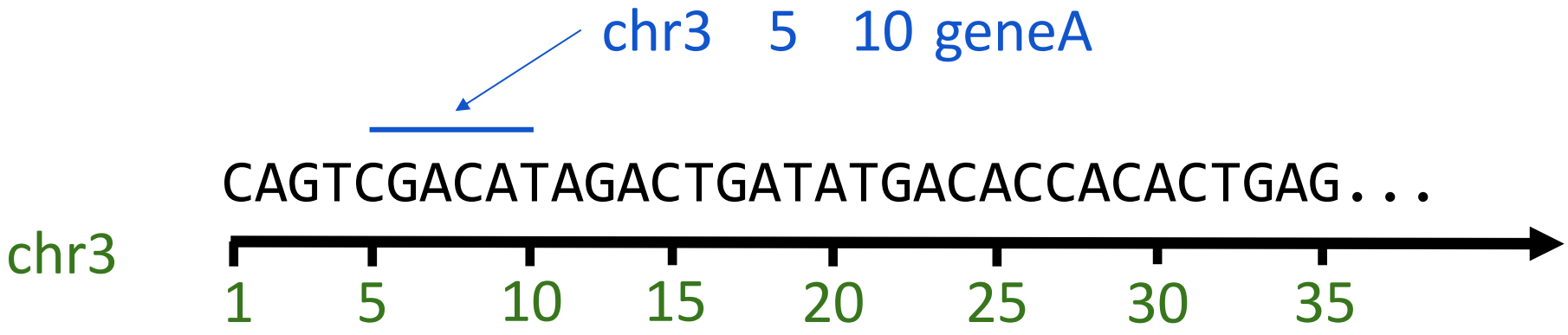
How do we annotate the genome?



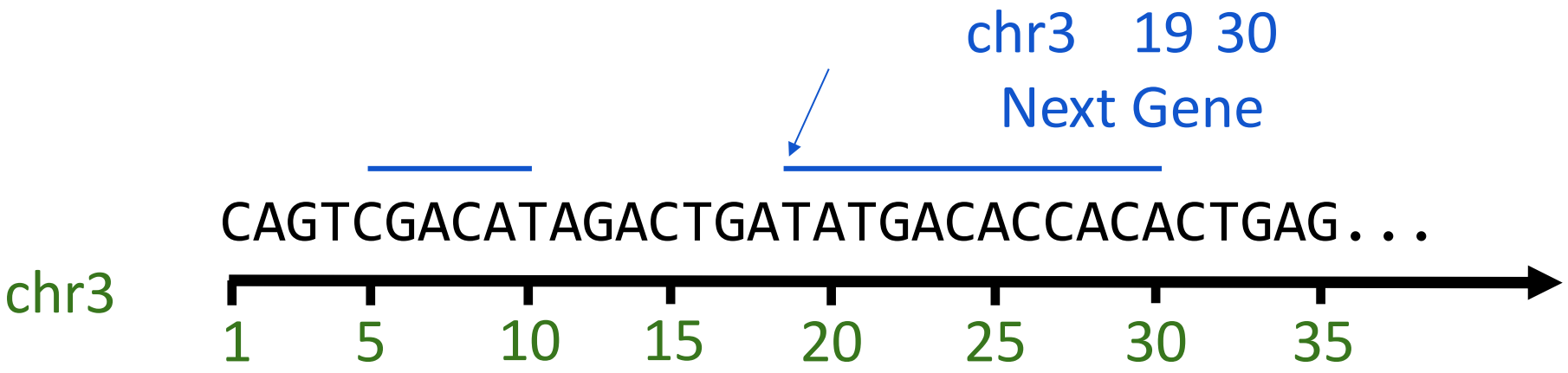
The genome as coordiantes



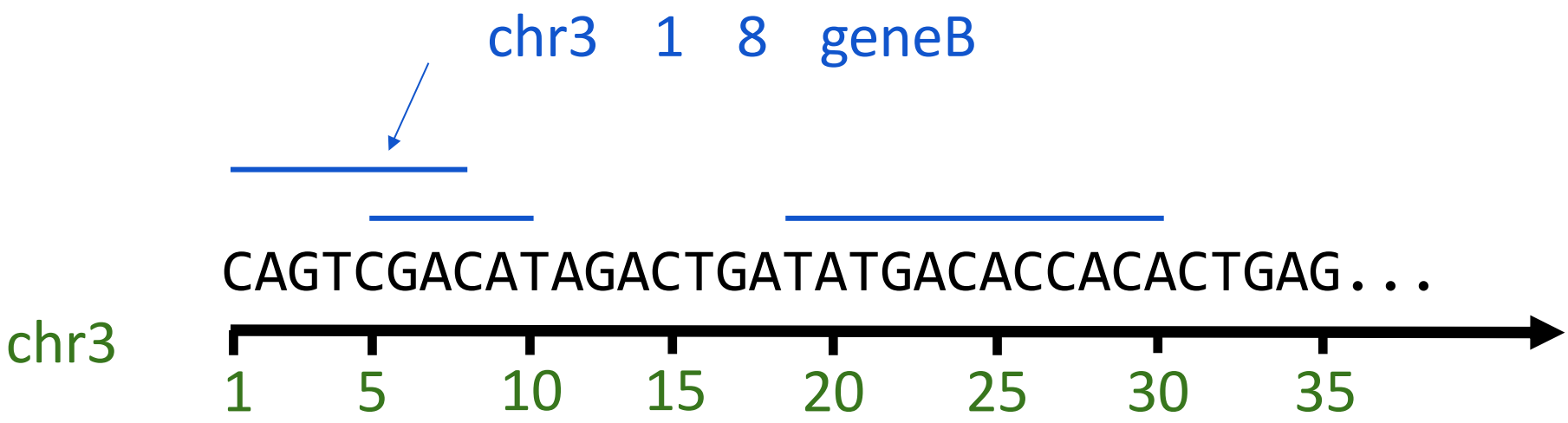
How do you describe a gene?



What about 2 consecutive genes?



What about overlapping genes?



File Formats

- BED (Browser Extensible Format)
- GFF (General Feature Format)
- GTF (Gene Transfer Format, GTF2.2)

BED File Format

BED format Index ▷

BED format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.

If your data set is BED-like, but it is very large (over 50MB) and you would like to keep it on your own server, you should use the bigBed data format.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.

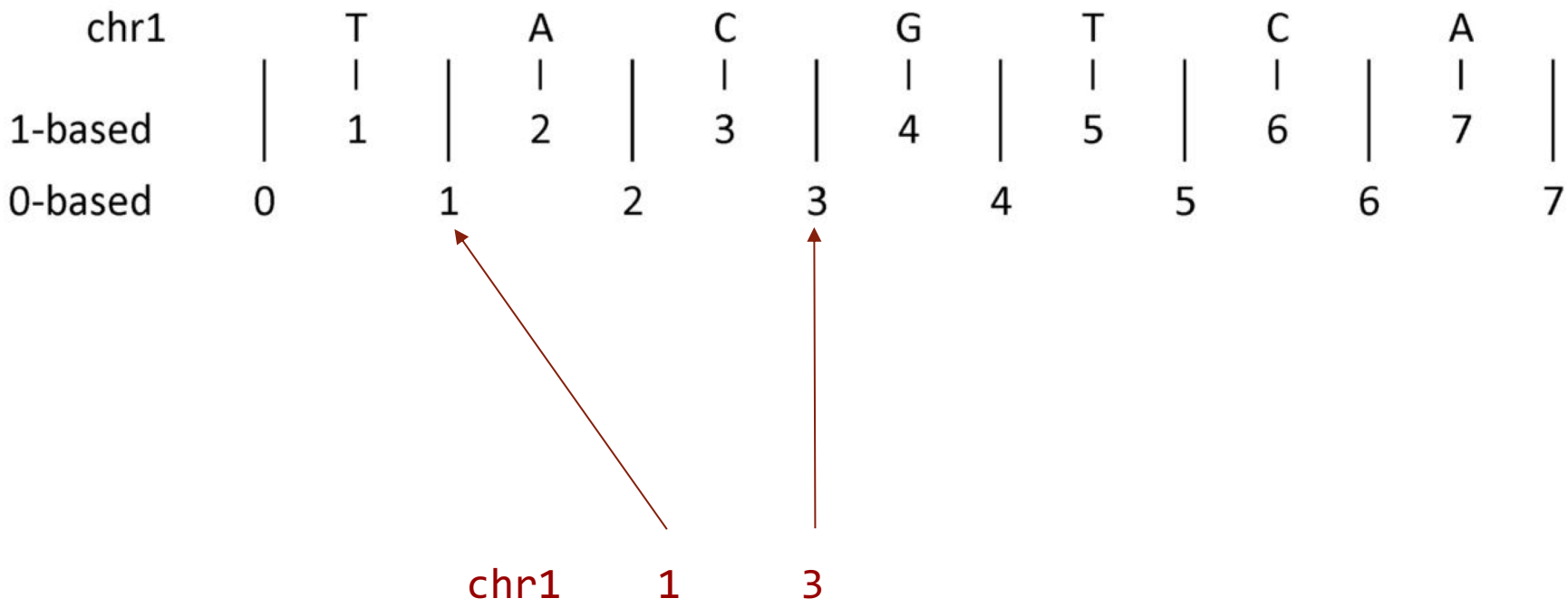
The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

shade								
score in range	≤ 166	167-277	278-388	389-499	500-611	612-722	723-833	834-944	≥ 945

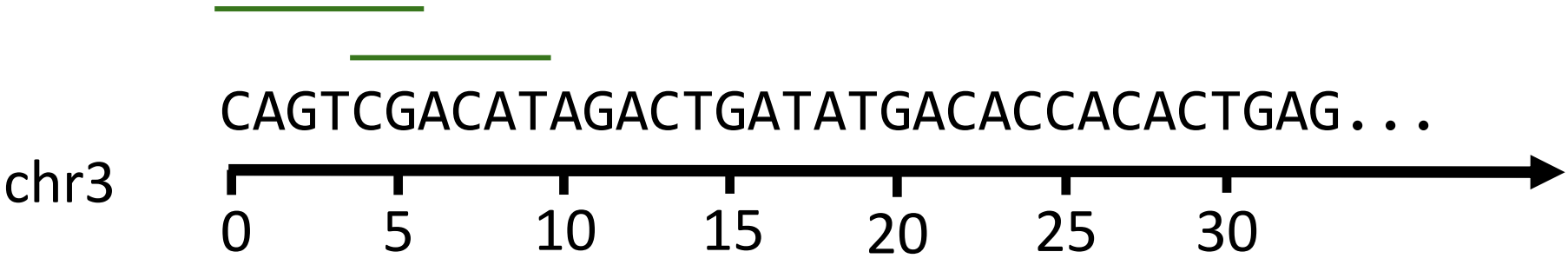
6. **strand** - Defines the strand - either '+' or '-'.
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, thickStart and thickEnd are usually set to the chromStart position.
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

BED File use 0-based, half open intervals



https://www.biostars.org/p/84686/

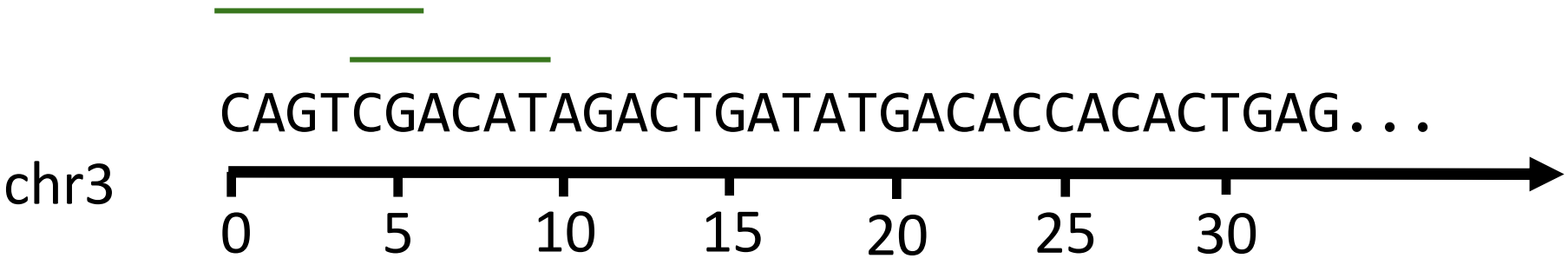
Example BED3




```
chr3    0    7
chr3    4    10
```


https://www.biostars.org/p/84686/

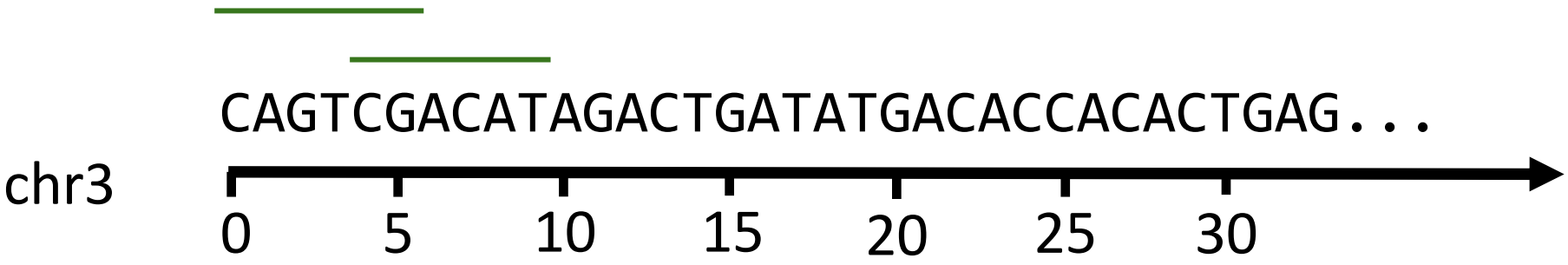
Add Name in 4th column to BED format




```
chr3    0    7    first
chr3    4    10   second
```


https://www.biostars.org/p/84686/

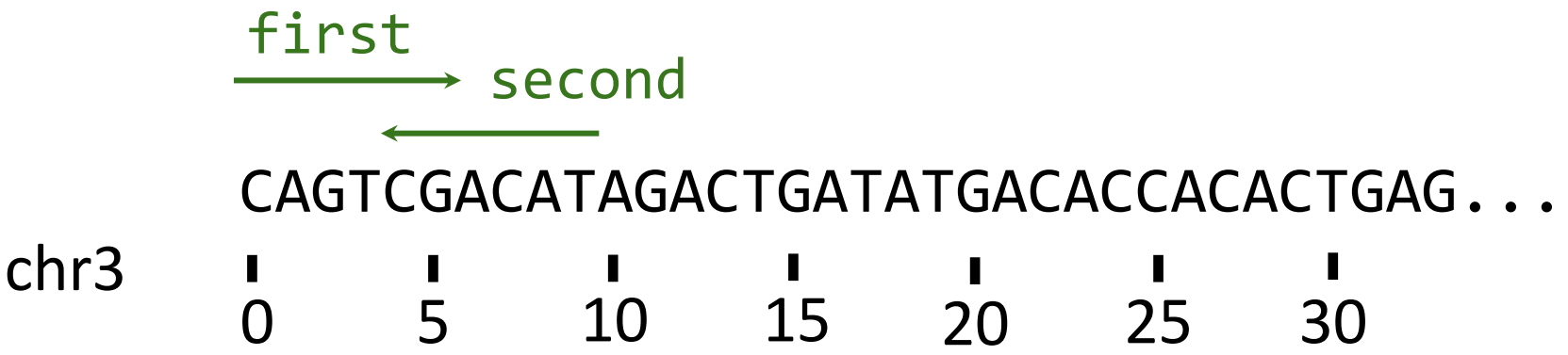
Add Intensity in 5th column to BED format




```
chr3    0     7    first    7.39
chr3    4     10   second   1e-8
```


https://www.biostars.org/p/84686/

Add orientation/strand to the 6th field




```
chr3    0     7    first    7.39  +
chr3    4     10   second   1e-8  -
```


https://www.biostars.org/p/84686/

BED can also track Genome Annotations

BED format Index ▷

BED format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.

If your data set is BED-like, but it is very large (over 50MB) and you would like to keep it on your own server, you should use the bigBed data format.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.

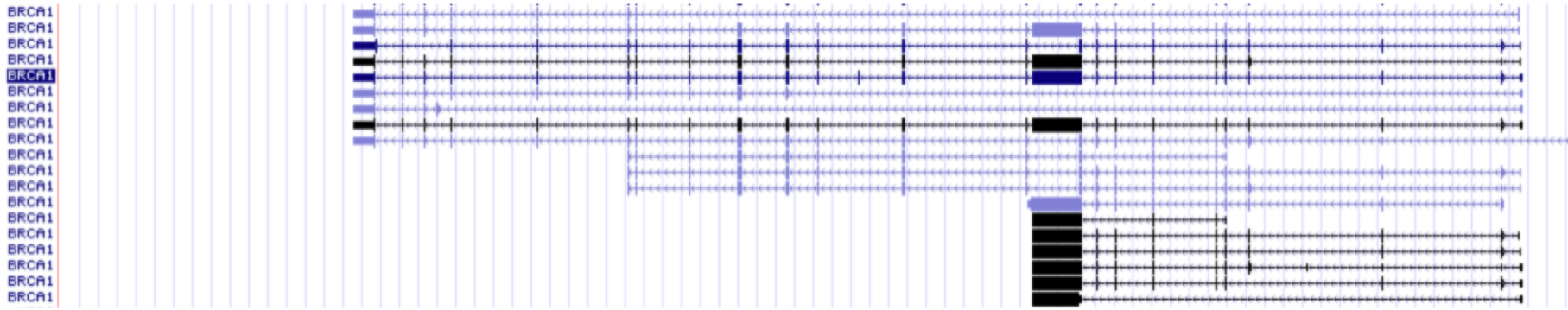
The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

shade								
score in range	≤ 166	167-277	278-388	389-499	500-611	612-722	723-833	834-944	≥ 945

6. **strand** - Defines the strand - either '+' or '-'.
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, thickStart and thickEnd are usually set to the chromStart position.
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

BED12 Example




```
chr17   41196311        41277340        uc010whm.2      0       -       41197694        41277202        0       8       1508,61,74,55,84,41,78,142,     0,3348,4826,6768,12757,19038,19579,80887,
chr17   41196311        41277340        uc002icp.4      0       -       41197694        41258496        0       23      1508,61,74,55,84,41,78,88,311,191,127,172,89,3426,77,46,103,140,89,56,54,99,142,
0,3348,4826,6768,12757,19038,19579,23313,26633,30036,32193,38109,46649,47140,51551,52949,55480,59827,60573,62183,71431,79722,80887,
chr17   41196311        41277468        uc002icu.3      0       -       41197800        41276113        0       22      1508,61,55,84,41,78,88,311,191,124,172,89,117,77,46,106,140,89,78,54,99,175,
0,3348,6768,12757,19038,19579,23313,26633,30036,32193,38109,46649,50449,51551,52949,55480,59827,60573,62161,71431,79722,80982,
chr17   41196311        41277468        uc010cyx.3      0       -       41197694        41258543        0       22      1508,61,74,55,84,41,78,88,311,191,127,172,89,3426,77,46,106,140,89,78,99,175,
0,3348,4826,6768,12757,19038,19579,23313,26633,30036,32193,38109,46649,47140,51551,52949,55480,59827,60573,62161,79722,80982,
chr17   41196311        41277500        uc002ict.3      0       -       41197694        41276113        0       24      1508,61,74,55,84,41,78,88,311,191,124,66,172,89,3426,77,46,106,140,89,78,54,99,213,
0,3348,4826,6768,12757,19038,19579,23313,26633,30036,32193,35039,38109,46649,47140,51551,52949,55480,59827,60573,62161,71431,79722,80976,
chr17   41196311        41277500        uc010whn.2      0       -       41197694        41226495        0       11      1508,61,74,55,84,41,78,88,311,191,213,  0,3348,4826,6768,12757,19038,19579,23313,26633,30036,80976,
chr17   41196311        41277500        uc010who.3      0       -       41197694        41202109        0       5       1508,61,74,129,213,     0,3348,4826,5767,80976,
chr17   41196311        41277500        uc002icq.3      0       -       41197694        41276113        0       23      1508,61,74,55,84,41,78,88,311,191,127,172,89,3426,77,46,106,140,89,78,54,99,213,
0,3348,4826,6768,12757,19038,19579,23313,26633,30036,32193,38109,46649,47140,51551,52949,55480,59827,60573,62161,71431,79722,80976,
chr17   41196311        41322420        uc010whp.2      0       -       41197694        41258543        0       22      1508,61,74,55,84,41,78,88,311,191,124,172,89,117,77,46,106,140,89,78,54,278,
0,3348,4826,6768,12757,19038,19579,23313,26633,30036,32193,38109,46649,50449,51551,52949,55480,59827,60573,62161,71431,125831,
chr17   41215349        41256973        uc010whq.1      0       -       41215349        41256198        0       12      41,78,88,311,191,127,172,89,117,106,140,89,
0,541,4275,7595,10998,13155,19071,27611,31411,36442,40789,41535,
chr17   41215349        41277468        uc002idc.1      0       -       41215349        41276113        0       18      41,78,88,311,191,127,172,89,117,77,46,103,140,89,78,54,99,175,
0,541,4275,7595,10998,13155,19071,27611,31411,32513,33911,36442,40789,41535,43123,52393,60684,61944,
chr17   41215349        41277468        uc010whr.1      0       -       41215349        41258543        0       17      41,78,88,311,191,127,172,89,117,77,46,106,140,89,78,99,175,
0,541,4275,7595,10998,13155,19071,27611,31411,32513,33911,36442,40789,41535,43123,60684,61944,
chr17   41243116        41276132        uc002idd.3      0       -       41243347        41276113        0       9       3761,77,46,106,140,89,78,54,99, 0,4746,6144,8675,13022,13768,15356,24626,32917,
chr17   41243451        41256973        uc002ide.1      0       -       41243452        41256198        0       4       3426,103,140,89,        0,8340,12687,13433,
chr17   41243451        41277340        uc010cyy.1      0       -       41243452        41276113        0       10      3426,77,46,106,140,89,78,54,99,142,     0,4411,5809,8340,12687,13433,15021,24291,32582,33747,
chr17   41243451        41277468        uc010whs.1      0       -       41243452        41276113        0       10      3426,77,46,106,140,89,78,54,99,175,     0,4411,5809,8340,12687,13433,15021,24291,32582,33842,
chr17   41243451        41277500        uc010cyz.2      0       -       41243452        41258543        0       11      3426,77,46,106,140,89,78,116,54,99,213, 0,4411,5809,8340,12687,13433,15021,19030,24291,32582,33836,
chr17   41243451        41277500        uc010cza.2      0       -       41243452        41276113        0       9       3426,77,46,106,140,89,54,99,213,        0,4411,5809,8340,12687,13433,24291,32582,33836,
chr17   41243451        41277500        uc010wht.1      0       -       41243452        41246659        0       2       3426,213,       0,33836,
chr17   41277599        41292342        uc002idf.3      0       +       41277599        41277599        0       4       188,63,182,1669,        0,5625,7373,13074,
chr17   41277599        41292342        uc010czb.2      0       +       41277599        41277599        0       2       188,1669,       0,13074,
chr17   41277599        41297125        uc002idg.3      0       +       41277599        41277599        0       5       188,63,266,468,381,     0,5625,13074,14233,19145,
chr17   41277599        41305688        uc002idh.3      0       +       41277599        41277599        0       8       188,63,125,182,205,266,120,70,  0,5625,7016,7373,12573,13074,16874,28019,
```

GTF/GFF

GFF (General Feature Format)

GTF (General Transfer Format)

Note: GFFv2 == GTF

Fields:

1. Seqname – name of chrom
2. Source – name of data source
3. Feature - type
4. Start – Start position
5. End – End position
6. Score
7. Strand
8. Frame – Indicates that base of codon
9. Attribute – Semicolon list

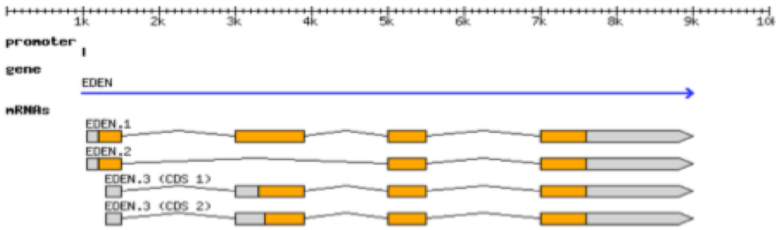
```
chr22  TeleGene enhancer  10000000  10001000  500 +  .  touch1
chr22  TeleGene promoter  10010000  10010100  900 +  .  touch1
chr22  TeleGene promoter  10020000  10025000  800 -  .  touch2
```

Note that the start and end coordinates are 1-based versus 0-based BED format

GTF/GFF Example

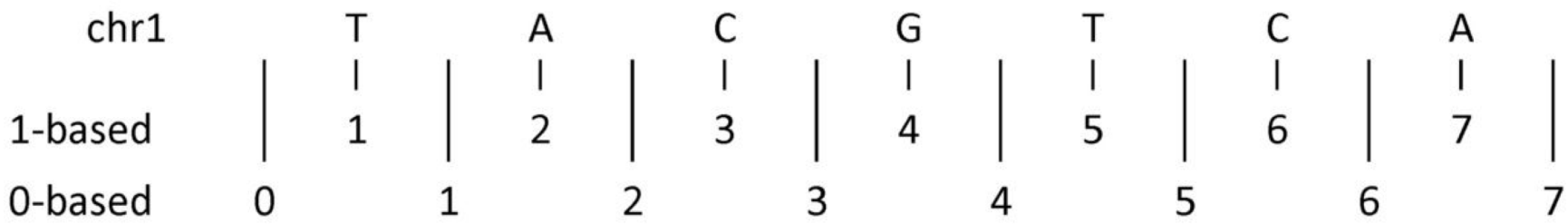
GFF Example

Gene "EDEN" with 3 alternatively spliced transcripts, isoform 3 has two alternative translation start sites




```
##gff-version 3
##sequence-region   ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN

ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001

ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3

ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003

ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1

ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2

ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3

ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

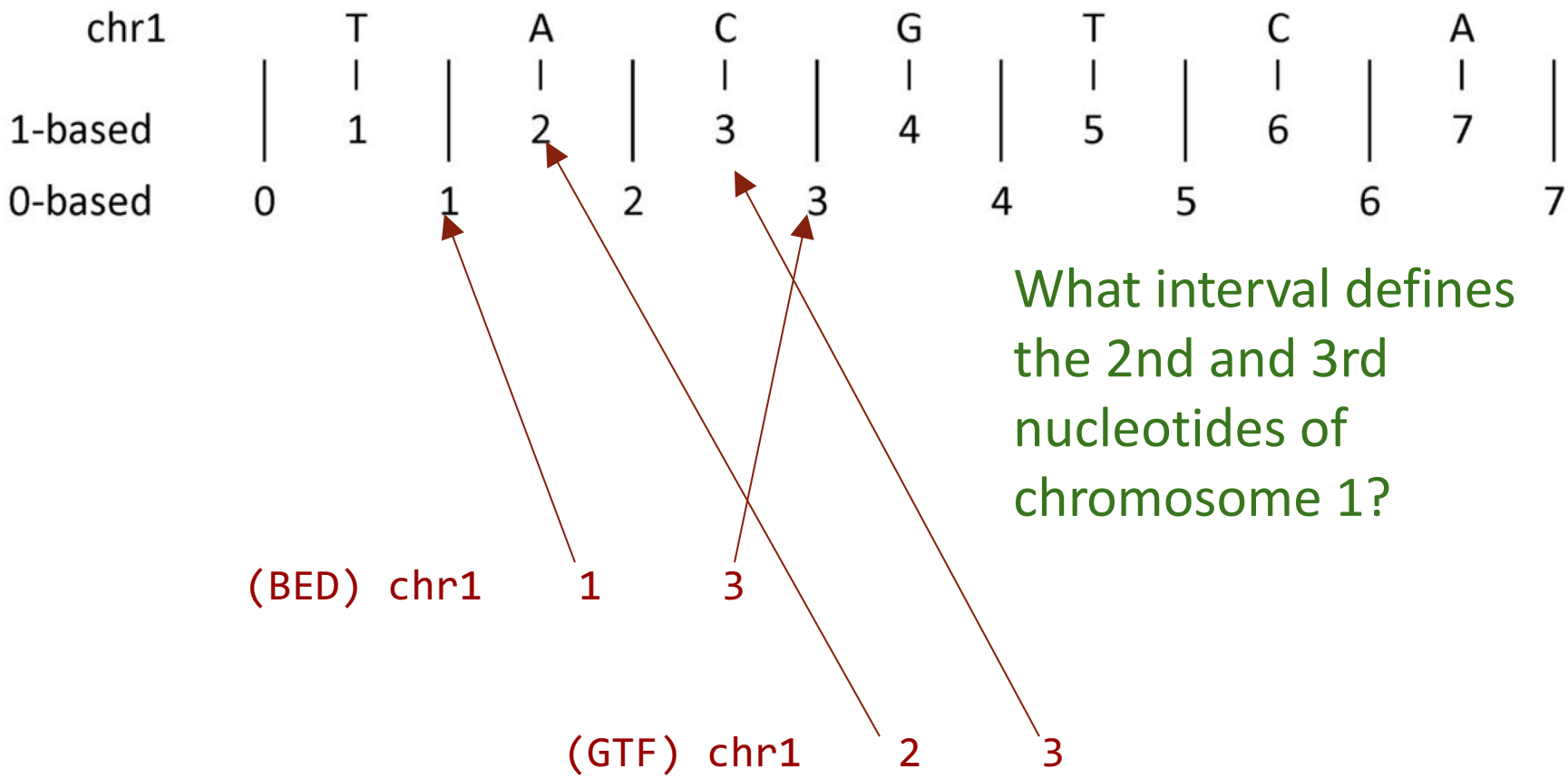
How does BED and GTF define a genomic range?



What interval defines the 2nd and 3rd nucleotides of chromosome 1?

https://www.biostars.org/p/84686/

How does BED and GTF define a genomic range?



https://www.biostars.org/p/84686/

Not all file formats are created Equal

BED: 0-based, half-open

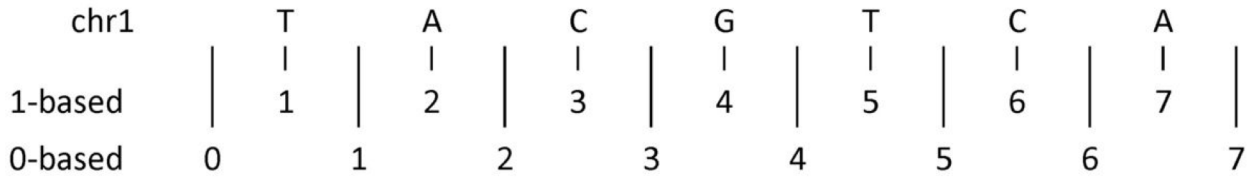
GFF: 1-based, closed

SAM: 1-based, closed

BAM: 0-based, half-open.

VCF: 1-based, closed

…



Workshop

GENCODE Annotation

How many transcripts are there for Tnnt2?

```
grep 'Tnnt2' gencode.gtf | cut -f9 | grep ENSMUST | cut -f2,8 -d ';' | sort | uniq |wc
```

How many transcripts are there for Tnnt2?

```
grep 'Tnnt2' gencode.gtf | cut -f9 | grep ENSMUST | grep  'transcript_type "protein_coding"' | cut -f2,8 -d ';' | sort | uniq | wc
```

GENCODE Annotation

How many lines in the GTF file?

```
wc -l gencode.gtf
```

How would you view the first 5 lines of the file?

```
head –n 5 gencode.gtf
```

What are the first 5 lines of the file?

##description: evidence-based annotation of the mouse genome (GRCm38), version M10 (Ensembl 85)
##provider: GENCODE
##contact: gencode-help@sanger.ac.uk
##format: gtf
##date: 2016-07-19

Why might this information be important?

Indicates the version of the annotation use and when it was generated.

RNA-seq Analysis

Sequence RNA transcripts (usually cDNA) to understand how gene regulation varies in different cell types, genetic backgrounds, or conditions.

Gene/Transcript Quantification

The number of reads that align to each gene is an estimate of how many RNA transcripts were present in the sample for that gene.

Problems:

- Paralogs (reads don't always map uniquely)
- Reads that span 2 or more exons
- Sampling error

http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Raw Read Counts per Gene/Transcript

Gene	Control	Experimental Condition
Gene A	27	27
Gene B	90	270
Gene C	280	640
Gene D	1003	3021
Gene E	3100	3342

Wow! Gene E is expressed at a 3X higher rate than Gene D!

Oooh! The experimental condition caused genes B,C,D, and E to be overexpressed!

FeatureCounts is a useful tool for generating counts:
https://www.ncbi.nlm.nih.gov/pubmed/24227677

Raw Read Counts per Gene/Transcript

Gene	Control	Experimental Condition
Gene A	27	27
Gene B	90	270
Gene C	280	640
Gene D	1003	3021
Gene E	3100	3342

Oooh! The experimental condition caused genes B,C, and D to be overexpressed!

Would you reach the same conclusion if you knew that the control experiment used 10 million reads while the experimental condition used 30 million reads?

Possible explanations:

1. 3 times as many transcripts are expressed for Gene E than Gene D.
2. Gene E is 3X as long as Gene D.
3. Gene D and E are the same length and produce the same number of transcripts, but Gene D has a close paralog that has acted as a "sponge" for ⅔ of its alignments.

Gene	Control	Experimental Condition
Gene A	27	27
Gene B	90	270
Gene C	280	640
Gene D	**1003**	3021
Gene E	**3100**	3342

Why we Normalize

- To smooth out technical variations amoung samples:
  - Sequencing depth: genes have more reads in a deeper sequenced library
  - Gene length: longer genes are likely to have more reads that the shorter reads

- **CPM (counts per million):** counts scaled by total number of reads. This method accounts for sequencing depth only.

- **TPM (transcripts per kilobase million):** counts per length of transcript (kb) per million reads mapped. This method accounts for both sequencing depth and gene length.

- **RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped):** similar to TPM, as this method also accounts for both sequencing depth and gene length as well

RPKM (or FPKM): reads per kilobase of exon model per million reads

Gene	Gene Size	Control	Experimental Condition
Gene A	1 kb	27	27
Gene B	2 kb	90	270
Gene C	6 kb	280	640
Gene D	30 kb	1003	3021
Gene E	40 kb	3100	3342

RPKM (or FPKM): reads per kilobase of exon model per million reads

Gene	Gene Size	Control	Experimental Condition
Gene A	1 kb	27	27
Gene B	2 kb	90	270
Gene C	6 kb	280	640
Gene D	30 kb	1003	3021
Gene E	40 kb	3100	3342
Total (w/ counts for ~19000 other genes)		**20,000,000**	**40,000,000**

Step 1: Normalize (i.e., adjust) gene counts by the total amount of sequences in the experiment

RPKM (or FPKM): reads per kilobase of exon model per million reads

Gene	Gene Size	Control (RPM)	Experimental Condition (RPM)
Gene A	1 kb	1.34	0.675
Gene B	2 kb	4.5	6.75
Gene C	6 kb	14.0	16.0
Gene D	30 kb	50.15	75.525
Gene E	40 kb	155.0	83.55
Total		**20,000,000**	**40,000,000**
Millions of reads		**20**	**40**

Step 1: Normalize gene counts the total amount of sequences in the experiment

RPKM (or FPKM): reads per kilobase of exon model per million reads

Gene	Gene Size	Control (RPM)	Experimental Condition (RPM)
Gene A	1 kb	1.34	0.675
Gene B	2 kb	4.5	6.75
Gene C	6 kb	14.0	16.0
Gene D	30 kb	50.15	75.525
Gene E	40 kb	155.0	83.55
Total		**20,000,000**	**40,000,000**
Millions of reads		**20**	**40**

Step 2: Normalize gene counts RPM by gene length

RPKM (or FPKM): reads per kilobase of exon model per million reads

Gene	Gene Size	Control (RPKM)	Experimental Condition (RPKM)
Gene A	1 kb	1.34	0.675
Gene B	2 kb	2.25	3.375
Gene C	6 kb	2.33	2.67
Gene D	30 kb	1.67	2.52
Gene E	40 kb	3.875	2.09
Total		**20,000,000**	**40,000,000**
Millions of reads		**20**	**40**

Step 2: Normalize gene counts RPM by gene length

Which Genes is most expressed in each condition

Gene	Gene Size	Control (RPKM)	Experimental Condition (RPKM)
Gene A	1 kb	1.34	0.675
Gene B	2 kb	2.25	3.375
Gene C	6 kb	2.33	2.67
Gene D	30 kb	1.67	2.52
Gene E	40 kb	**3.875**	2.09
Total		**20,000,000**	**40,000,000**
Millions of reads		**20**	**40**

RPKM is best for *within-sample* comparisons of gene expression. TPM is best for inter-sample comparisons

Differential Expression Between Samples

METHOD | OPEN ACCESS

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love, Wolfgang Huber and Simon Anders

Genome Biology 2014 15:550 | DOI: 10.1186/s13059-014-0550-8 | © Love et al.; licensee BioMed Central. 2014
Received: 27 May 2014 | Accepted: 19 November 2014 | Published: 5 December 2014

Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]
[1]Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010 and [2]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia

Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments*

Gordon K. Smyth
Walter and Eliza Hall Institute of Medical Research
Melbourne, Vic 3050, Australia

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728800/pdf/13059_2016_Article_881.pdf

Why you need replicates?

A true technical replicate repeats sample and library prep.

Library prep → sequencing → Technical replicate #1

Library prep → sequencing → Technical replicate #2

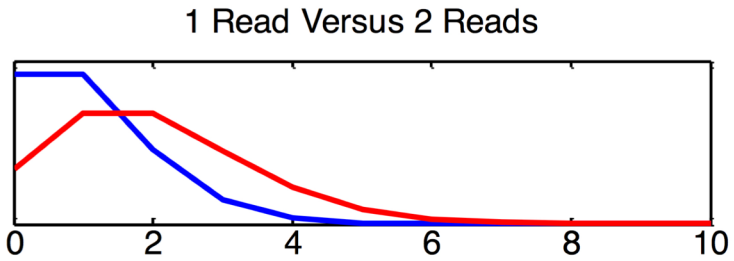
Library prep → sequencing → Biological replicate #1

Library prep → sequencing → Biological replicate #1

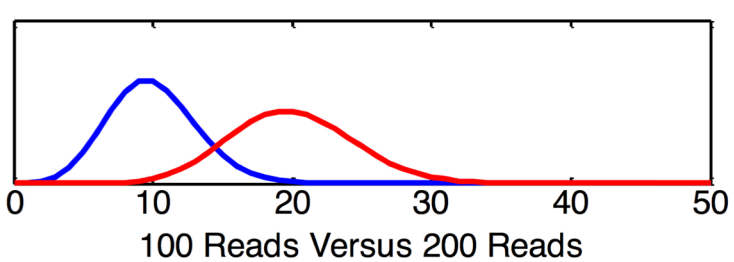
http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Source of Variation. Poisson (counting) noise

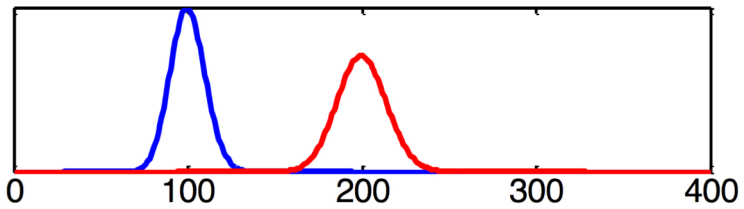
Gene A



Gene B



Gene C

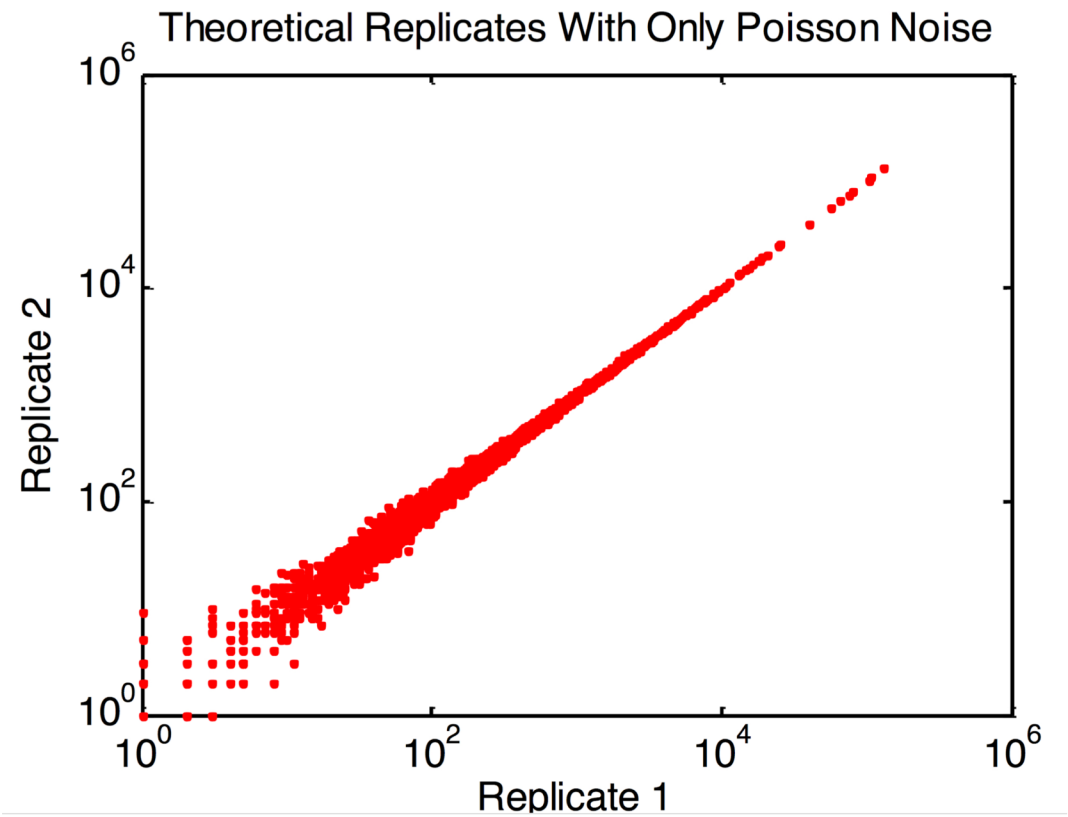


In each case, there is an apparent 2X difference in the mean read counts between control and experimental.

Yet poisson variance is higher relative to the total count when counts are low versus when they are high. For example, the difference in expression of a gene measured with one read versus two reads is inherently less certain than the differences in expression of a gene measured with 100 reads versus 200 reads, even though both differences are nominally a 2X fold change.

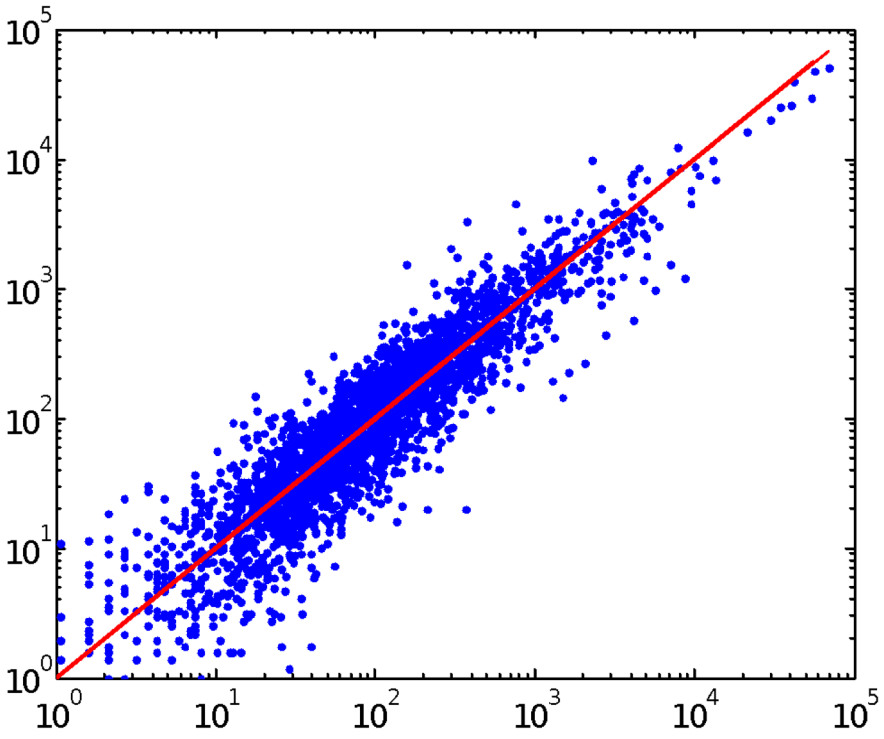
http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Source of Variation. Poisson (counting) noise



http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Source of Variation. Non-Poisson technical noise



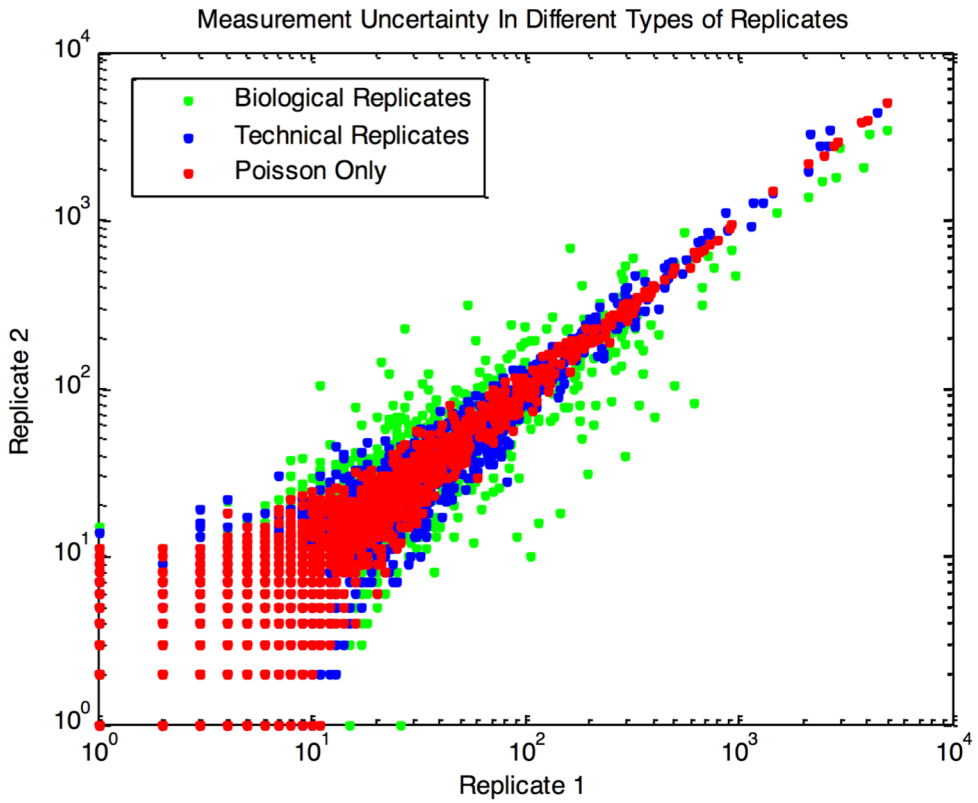
"Non-Poisson Technical Variance is measurement imprecision that stems from the inability of RNA-Seq measurements to measure expression perfectly.

This imprecision is seen when expression from the same sample is measured twice. The expression measurements will not match exactly, and have error greater than what is expected from Poisson noise alone.

Sources of measurement imprecision may include PCR amplification errors during library preparation or machine errors."
-- Michele Busby

http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Sources of variance. Biological Variance



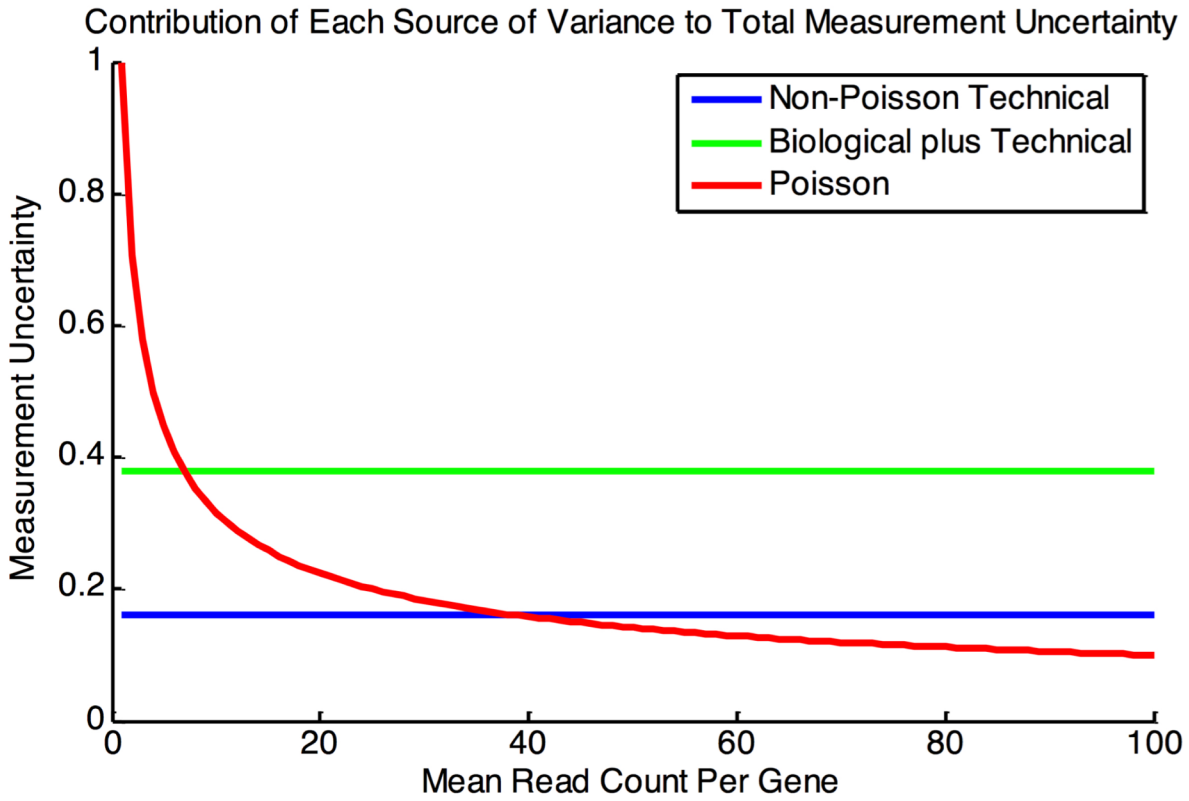
Biological variance is variance that naturally occurs within the samples under investigation. This variance stems from the fact that the expression of any given gene is likely to naturally fluctuate within the cells themselves, and between samples of the same condition. Sources of biological variance include genetic differences among samples and gene expression responses to the environment.

Green dots: biological replicates from S. cerevisiae

Blue dots: technical replicates from S. cerevisiae

Red dots: simulated replicates with only Poisson noise

Relative contribution of sources of variance



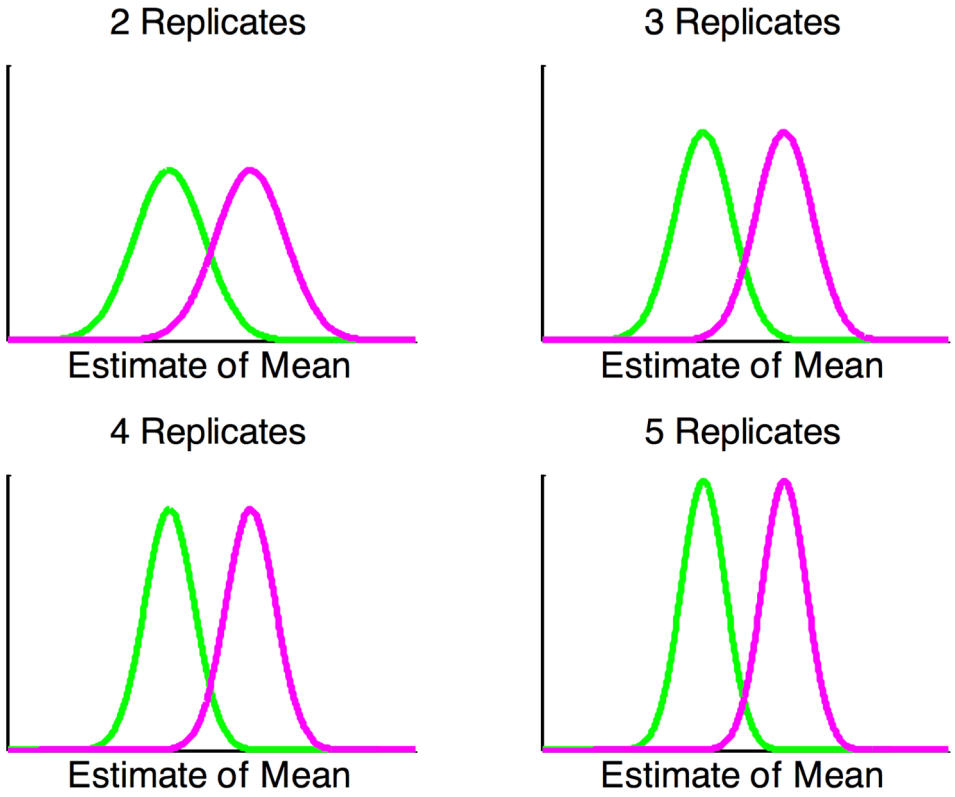
How to decrease uncertainty? Increase Reads

	Replicate 1	Replicate 2	Replicate 3	Mean
Control	12	14	19	15
Test	12	41	7	20

	Replicate 1	Replicate 2	Replicate 3	Mean
Control	22	26	38	30
Test	23	82	15	41

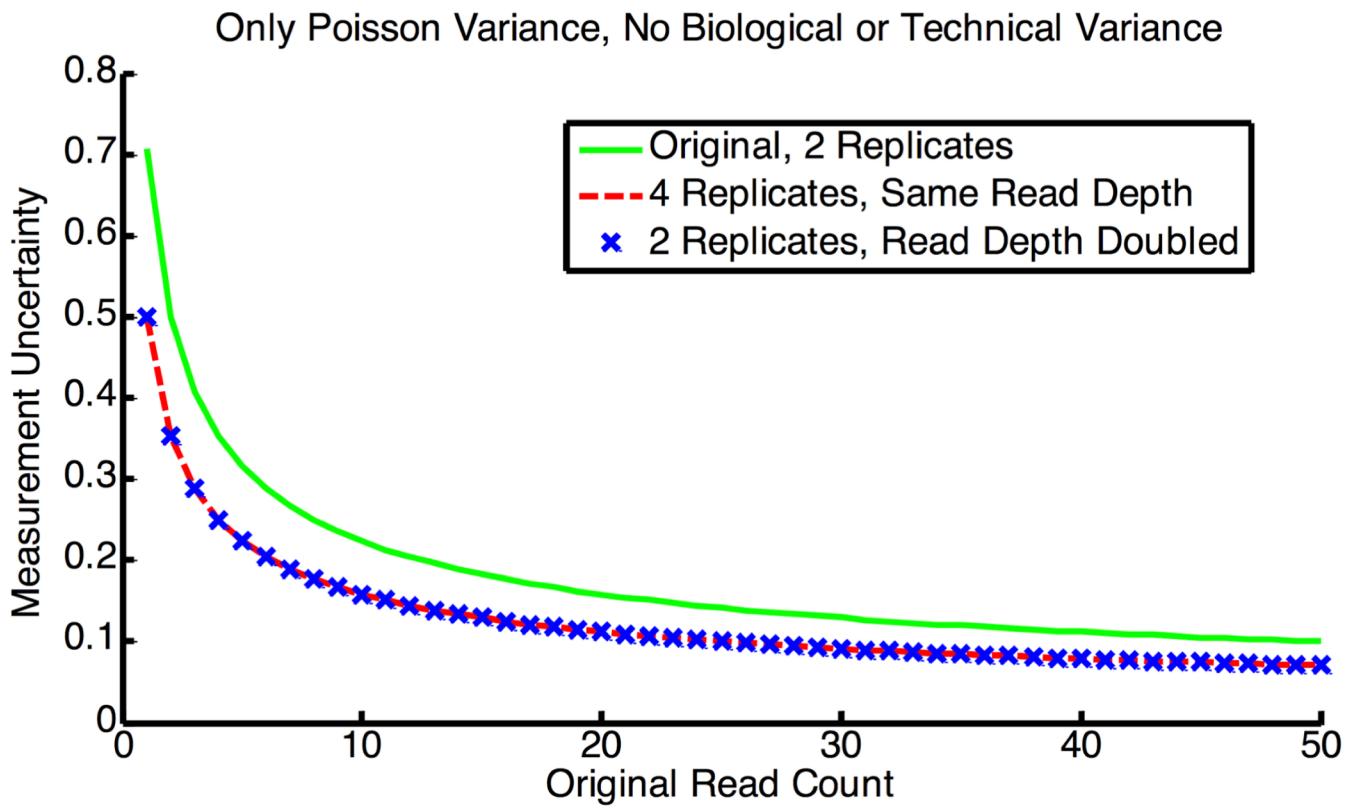
http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

How to decrease uncertainty? Increase Replicates



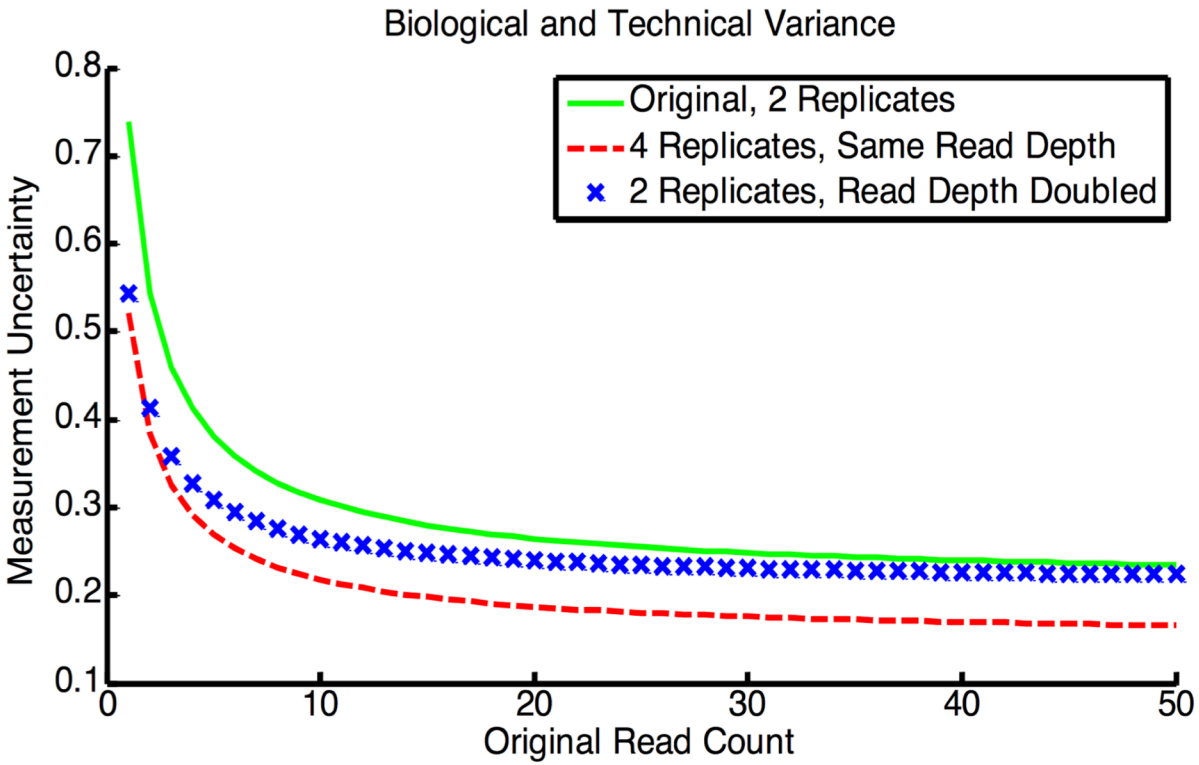
http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Which is better (Depth or Replicates)?



http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Which is better (Depth or Replicates)?



http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Correcting for Multiple Testing

1. If you test 30,000 genes for differential gene expression, and you use a significance cut off of $p<0.05$, then you should expect to call approximately 1500 (i.e., 5% of 30000) genes to exhibit differential expressed solely random chance.
2. Thus, if your list of differentially expressed genes at $p<0.05$ is about 1500 genes long, then either there are no genes differentially expressed between the two conditions, or your experiment is underpowered.

False Discovery Rate

It is informative to report a False Discovery Rate (FDR) as well as a p-value. This value is the expected proportion of false positives among all of the significant results (100% in the previous example).

Usually we strive for a FDR at 5% or lower. Importantly, the FDR can only be calculated after an experiment is run, because it requires knowing how many genes were called differentially expressed.

Workshop