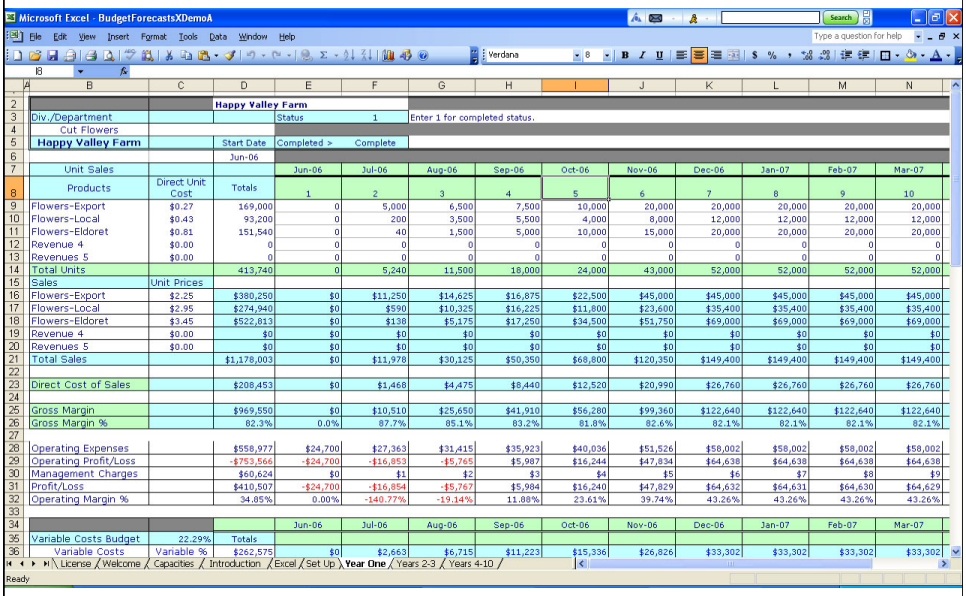
File formats

Wrapping your data in the right package

Carol Bult, Ph.D.
Professor
Deputy Director, The Jackson Laboratory Cancer Center

The Jackson Laboratory
Leading the search for tomorrow's cures

Short Course in Medical Genetics 2014



http://www.downloadsoftfree.com/windows/Business/Business-Finance/Budgeting-Spreadsheets-for-Excel-1-2-13-4345-1-0-0.html

```
<Styles>
 <Style ss:ID="Default" ss:Name="Normal">
  <Alignment ss:Vertical="Bottom"/>
  <Borders/>
  <Font ss:FontName="Arial"/>
  <Interior/>
  <NumberFormat/>
  <Protection/>
 </Style>
 <Style ss:ID="s16">
  <Alignment ss:Vertical="Bottom" ss:WrapText="1"/>
  <Font ss:FontName="Arial" ss:Bold="1"/>
 </Style>
 <Style ss:ID="s17">
  <Alignment ss:Horizontal="Right" ss:Vertical="Bottom" ss:WrapText="1"/>
  <Font ss:FontName="Arial" ss:Bold="1"/>
 </Style>
 <Style ss:ID="s18">
  <NumberFormat ss:Format="#,##0"/>
 </Style>
 <Style ss:ID="s19">
  <Alignment ss:Horizontal="Left" ss:Vertical="Bottom"/>
  <NumberFormat ss:Format="#,##0"/>
 </Style>
 <Style ss:ID="s20">
  <Alignment ss:Horizontal="Right" ss:Vertical="Bottom"/>
 </Style>
 <Style ss:ID="s21">
  <Alignment ss:Horizontal="Left" ss:Vertical="Bottom"/>
 </Style>
</Styles>
```

```
track name="tb_gap" description="table browser query
chr1 167280   217280
chr1 257582   307582
chr1 461231   511231
chr1 2624080  2674080
chr1 3835128  3895128
```


```
track name="tb_gap" descript
chr1^I167280^I217280$
chr1^I257582^I307582$
chr1^I461231^I511231$
chr1^I2624080^I2674080$
chr1^I3835128^I3895128$
```

```
track name="tb_gap" description="table browser query
chr1    167280  217280
chr1    257582  307582
chr1    461231  511231
chr1    2624080     2674080
```


```
track name="tb_gap" description
chr1  167280  217280$
chr1  257582  307582$
chr1  461231  511231$
chr1  2624080     2674080$
chr1  3835128     3895128$
```

Control Characters: invisible to you but not to software

Carriage return (CR): \r or ^M

Line feed (LF): \n or ^J

Unix/Linux: uses LF character

Macs: uses CR character

Windows: uses CR followed by LF

http://danielmiessler.com/study/crlf/

Most bioinformatics packages expect:

- A plain text file
  - Not a word or excel document
- A particular field delimiter
  - often tab or comma, sometimes pipe
  - Unix style line terminators

Read file specifications!*

* Even though they may not be complete

Vince Buffalo
@vsbuffalo

Following

If I had one thing to tell biologists learning bioinformatics, it would be "write code for humans, write data for computers".

NCBI data representation:

- Uses ASN.1
- Not easily human readable
- Limited flexibility
- Robust validation tools
- Not easily parsed by Perl/Python

```
Seq-entry ::= seq {
  id {
    general {
      db "WGS:AMYH01" ,
      tag
        str "chr1_315417" } ,
    genbank {
      accession "JH976292" ,
      version 1 } ,
    gi 409188728 } ,
  descr {
    title "Homo sapiens chromosome 1 genomic scaffold" ,
    source {
      genome genomic ,
      org {
        taxname "Homo sapiens" ,
        common "human" ,
        db {
          {
            db "taxon" ,
            tag
              id 9606 } } ,
        orgname {
          name
            binomial {
              genus "Homo" ,
              species "sapiens" } ,
            lineage "Eukaryota; Metazoa; Chordata; Craniat
 Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Pri
 Catarrhini; Hominidae; Homo" ,
            gcode 1 ,
            mgcode 2 ,
            div "PRI" } } ,
      subtype {
        {
          subtype chromosome ,
          name "1" } ,
        {
          subtype cell-line ,
          name "CHM1htert" } ,
        {
          subtype tissue-type ,
          name "hydatidiform mole" } ,
```

Typical bioinformatics data representation:

Tab delimited file

```
track name="tb_snp135Common" description="table browser query on snp135Common" visibility=3 url=
chr1  10491 10492 rs55998931 0    +
chr1  10582 10583 rs58108140 0    +
chr1  14929 14930 rs75454623 0    +
chr1  20244 20245 rs71262674 0    -
chr1  20303 20304 rs71262673 0    -
chr1  33494 33495 rs75468675 0    +
```

- Flexible
  - Good: with rapidly changing data/tech (but don't change/add columns!)
  - Poor: validation
- Human Readable
  - Convenient for de-bugging
  - Computer doesn't care!

Putting the data in the right package

- Sequences
  - FASTA
  - FASTQ
  - SAM/BAM
- Alignments
  - SAM/BAM
  - MAF
- Annotations
  - Genes
    - GFF3
    - GTF
  - Variation
    - VCF
    - GVF
    - HGVS
  - General
    - GFF3
    - BED

http://deannachurch.github.io/BHSC_Bioinformatics/formats.html

FASTA

```
>OTTMUST00000058732 cdna:all chromosome:VEGA44:6:17256335:17291324:1 Gene:OTTMUSG00000024023
GCCCCCCGCCAGCGCCTTTCCCCCCTCTATACAATACAAGATCTTCCTTCCTCAGTTCTC
TTAAATCACAGCCCAGGGAAACCTCCTCAGAGCCTGCAGCCAGCCACGCGCCAGCATGTC
TGGGGGCAAATACGTAGACTCCGAGGGACATCTCTACACTGTTCCCATCCGGGAACAGGG
CAACATCTACAAGCCCAACAACAAGGCCATGGCAGACGAGGTGACTGAGAAGCAAGTGTA
TGACGCGCACACCAAGGAGATTGACCTGGTCAACCGCGACCCCAAGCATCTCAACGACGA
CGTGGTCAAGATTGACTTTGAAGATGTGATTGCAGAACCAGAAGGGACACACAGTTTCGA
CGGCATCTGGAAGGCCAGCTTCACCACCTTCACTGTGACAAAATATTGGTTTTACCGCTT
```

FASTQ

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

FASTQ Details

Sequence data format

- Text based
  - Encodes sequence calls and quality scores with ASCII characters
  - Stores minimal information about the sequence read
  - 4 lines per sequence
    - Line 1: begins with @; followed by sequence identifier and optional description
    - Line 2: the sequence
    - Line 3: begins with the "+" and is followed by sequence identifiers and description (both are optional)
    - Line 4: encoding of quality scores for the sequence in line 2

References
http://maq.sourceforge.net/fastq.shtml
Cock et al. (2009) Nuc Acids Res 38: 1767-1771

FASTQ Example

For analysis, it may be necessary to convert to the Sanger form of FASTQ.

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""7F071,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D02EA5':>5?:%A;A8A;?9B;D0
/=<?7=9<2A8==
```

@title and optional description
sequence line(s)
+optional repeat of title line
quality line(s)

FASTQ example from Cock et al., 2009

Quality Scores

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

Q = Phred Quality Scores
P = Base-calling error probabilities

Quality Scores

Not always directly comparable between to programs/pipelines

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
```

Format/Platform	QualityScoreType	ASCII encoding	
Sanger	Phred: 0-93	33-126	
Solexa	Solexa:-5-62	64-126	
Illumina 1.3	Phred: 0-62	64-126	
Illumina 1.5	Phred: 0-62	64-126	
Illumina 1.8	Phred: 0-62	33-126	*** Sanger format!

Need to know what your program is expecting
Likely to change again (to improve compressing data)

SAM (Sequence Alignment/Map)

Alignment data format

- Standard output of aligners that map reads to a reference genome
  - Tab delimited w/ header section and alignment section
    - Header sections begin with @ (are optional)
    - Alignment section has 11 mandatory fields
  - BAM is the binary format of SAM

http://samtools.sourceforge.net/

Mandatory Alignment Fields

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next fragment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next fragment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	fragment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

http://samtools.sourceforge.net/SAM1.pdf

Alignments example

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                       CAGCGCCAT
```

CIGAR string -> 8M2I4M1D3M

Alignments in SAM format

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

http://samtools.sourceforge.net/SAM1.pdf

Annotation Formats

- Mostly tab delimited files that describe the location of genome features (i.e., genes, etc.)
- Also used for displaying annotations on standard genome browsers
- Important for associating alignments with specific genome features
- Descriptions
- Knowing format details can be important to translating results!
  - BED is zero based/exclusive
  - GTF/GFF are one based/inclusive

BED: zero based, start inclusive, stop exclusive

```
chr1 10491     10492     rs55998931     0     +
chr1 10582     10583     rs58108140     0     +
```

- First base on the chromosome is 0
- Length = stop - start

GTF/GFF: one based, inclusive

```
chr1 snp135Com exon 10492     10492     0.000
chr1 snp135Com exon 10583     10583     0.000
```

- First base on the chromosome is 1
- Length = stop – start+1

BED format

Annotation data format

Required (1-3) Optional (4-12) →

```
chr1   86114265    86116346    nsv433165
chr2   1841774     1846089     nsv433166
chr16  2950446     2955264     nsv433167
chr17  14350387    14351933    nsv433168
chr17  32831694    32832761    nsv433169
chr17  32831694    32832761    nsv433170
chr18  61880550    61881930    nsv433171
```

```
chr1   16759829    16778548    chr1:21667704 270866 -
chr1   16763194    16784844    chr1:146691804      407277 +
chr1   16763194    16784844    chr1:144004664      408925 -
chr1   16763194    16779513    chr1:142857141      291416 -
chr1   16763194    16779513    chr1:143522082      293473 -
chr1   16763194    16778548    chr1:146844175      284555 -
chr1   16763194    16778548    chr1:147006260      284948 -
chr1   16763411    16784844    chr1:144747517      405362 +
```

GFF3

Annotation data format

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

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

 4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDE
 5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDE
 6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDE

 7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
 8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA0000
 9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA0000
10  ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA0000
    ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA0000

    ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=eden
```

Fixed columns:
Column 1: Sequence Id
Column 2: Source
Column 3: Feature type
Column 4: Start (1-based)
Column 5: End
Column 6: Score
Column 7: Strand
Column 8: Phase (0,1,2)

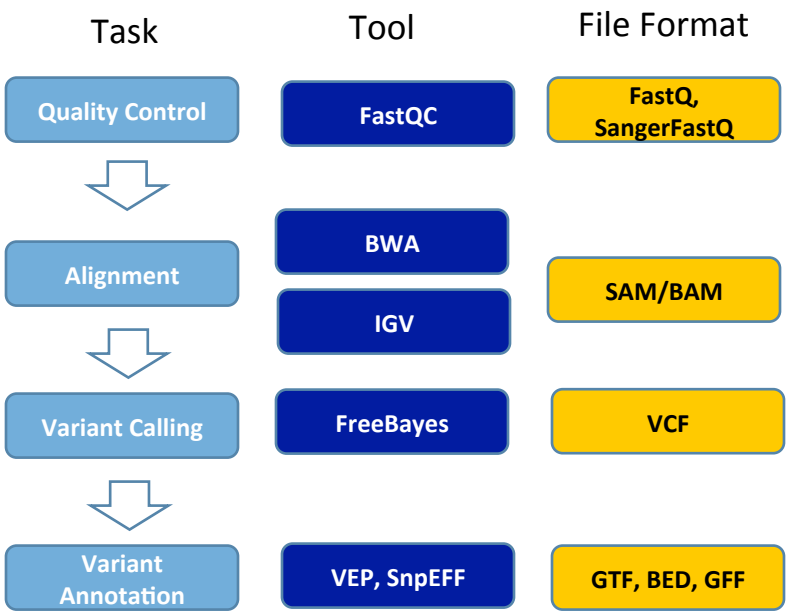
Flexible column:
Column 9: attributes

Semi-colon delimited tag=value pairs. Some tags are reserved (ID, Name, etc).

http://www.sequenceontology.org/resources/gff3.html

DNASeq Tasks, Tools and File Formats

We'll re-visit this on Friday during the Galaxy tutorial.



Take home messages

- Understand how your tools work
  - What is the tool expecting?
  - What type of data am I representing?
  - What type of data will it produce
- Output of programs/pipelines are not always comparable
  - Score values
- Know how to count (starting at 0 or 1)
- Just because 2 files are of the same type (BED, GFF3) it does not mean they are identical or 'standard'.

What to do next

- Work on the file format exercises on the workshop web site
- Explore the links on the File Formats section of the course web site
- The file formats that will be most relevant to you this week:
  - Fastq
  - SAM/BAM
  - BED
  - VCF