File formats

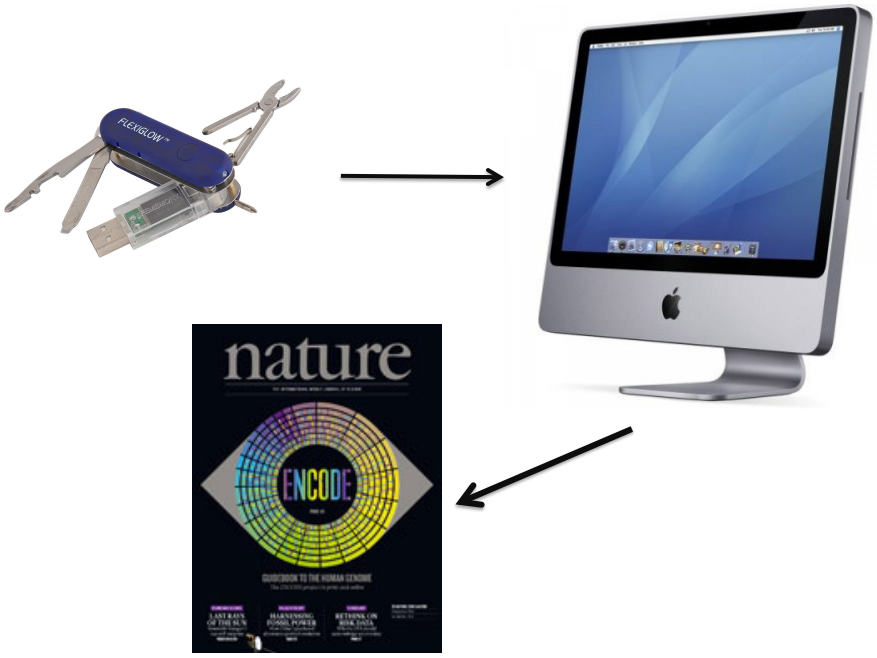
Wrapping your data in the right package

Deanna M. Church
Senior Director of Genomics and Content
Personalis, Inc

Personalis

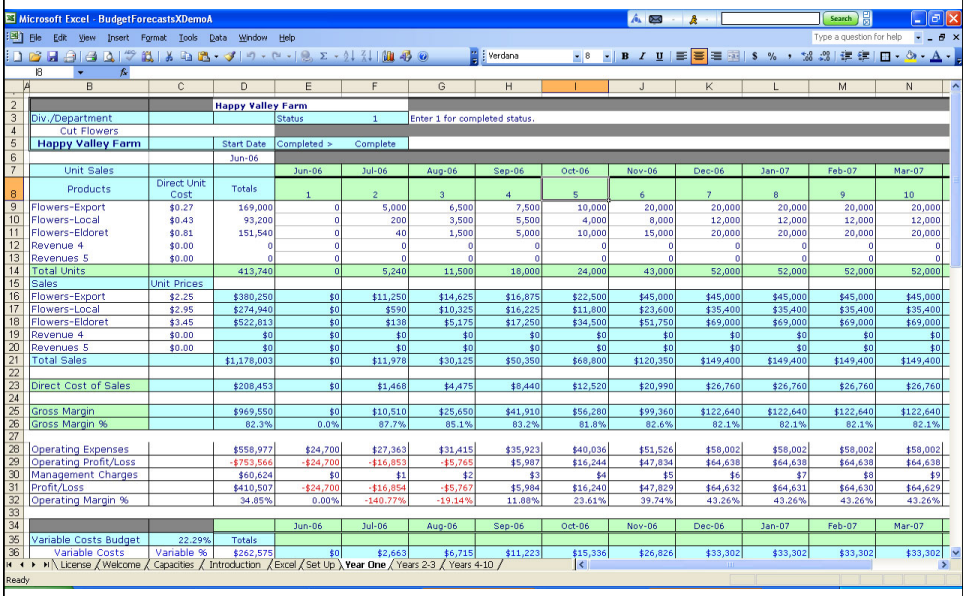
@deannachurch

Short Course in Medical Genetics 2013



http://www.dansdata.com

http://www.nature.com/encode



Microsoft Excel - BudgetForecastsXDemoA

2			Happy Valley Farm										
3	Div./Department			Status	1	Enter 1 for completed status.							
4	Cut Flowers												
5	Happy Valley Farm		Start Date	Completed >	Complete								
6			Jun-06										
7	Unit Sales			Jun-06	Jul-06	Aug-06	Sep-06	Oct-06	Nov-06	Dec-06	Jan-07	Feb-07	Mar-07
8	Products	Direct Unit Cost	Totals	1	2	3	4	5	6	7	8	9	10
9	Flowers-Export	$0.27	169,000	0	5,000	6,500	7,500	10,000	20,000	20,000	20,000	20,000	20,000
10	Flowers-Local	$0.43	93,200	0	200	3,500	5,500	4,000	8,000	12,000	12,000	12,000	12,000
11	Flowers-Eldoret	$0.81	151,540	0	40	1,500	5,000	10,000	15,000	20,000	20,000	20,000	20,000
12	Revenue 4	$0.00	0	0	0	0	0	0	0	0	0	0	0
13	Revenues 5	$0.00	0	0	0	0	0	0	0	0	0	0	0
14	Total Units		413,740	0	5,240	11,500	18,000	24,000	43,000	52,000	52,000	52,000	52,000
15	Sales	Unit Prices											
16	Flowers-Export	$2.25	$380,250	$0	$11,250	$14,625	$16,875	$22,500	$45,000	$45,000	$45,000	$45,000	$45,000
17	Flowers-Local	$2.95	$274,940	$0	$590	$10,325	$16,225	$11,800	$23,600	$35,400	$35,400	$35,400	$35,400
18	Flowers-Eldoret	$3.45	$522,813	$0	$138	$5,175	$17,250	$34,500	$51,750	$69,000	$69,000	$69,000	$69,000
19	Revenue 4	$0.00	$0	$0	$0	$0	$0	$0	$0	$0	$0	$0	$0
20	Revenues 5	$0.00	$0	$0	$0	$0	$0	$0	$0	$0	$0	$0	$0
21	Total Sales		$1,178,003	$0	$11,978	$30,125	$50,350	$68,800	$120,350	$149,400	$149,400	$149,400	$149,400
23	Direct Cost of Sales		$208,453	$0	$1,468	$4,475	$8,440	$12,520	$20,990	$26,760	$26,760	$26,760	$26,760
25	Gross Margin		$969,550	$0	$10,510	$25,650	$41,910	$56,280	$99,360	$122,640	$122,640	$122,640	$122,640
26	Gross Margin %		82.3%	0.0%	87.7%	85.1%	83.2%	81.8%	82.6%	82.1%	82.1%	82.1%	82.1%
28	Operating Expenses		$558,977	$24,700	$27,363	$31,415	$35,923	$40,036	$51,526	$58,002	$58,002	$58,002	$58,002
29	Operating Profit/Loss		-$753,566	-$24,700	-$16,853	-$5,765	$5,987	$16,244	$47,834	$64,638	$64,638	$64,638	$64,638
30	Management Charges		$60,624	$0	$1	$2	$3	$4	$5	$6	$7	$8	$9
31	Profit/Loss		$410,507	-$24,700	-$16,854	-$5,767	$5,984	$16,240	$47,829	$64,632	$64,631	$64,630	$64,629
32	Operating Margin %		34.85%	0.00%	-140.77%	-19.14%	11.88%	23.61%	39.74%	43.26%	43.26%	43.26%	43.26%
34				Jun-06	Jul-06	Aug-06	Sep-06	Oct-06	Nov-06	Dec-06	Jan-07	Feb-07	Mar-07
35	Variable Costs Budget	22.29%	Totals										
36	Variable Costs	Variable %	$262,575	$0	$2,663	$6,715	$11,223	$15,336	$26,826	$33,302	$33,302	$33,302	$33,302

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://www.ncbi.nlm.nih.gov/staff/church/GenomeAnalysis/index.shtml#FILES

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+&$#""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?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

```
##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=EDE
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDE
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDE

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

                         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

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'.