MCB 5472
Lecture #1 – NCBI and public data

Jan 27/14

Pretest question #1

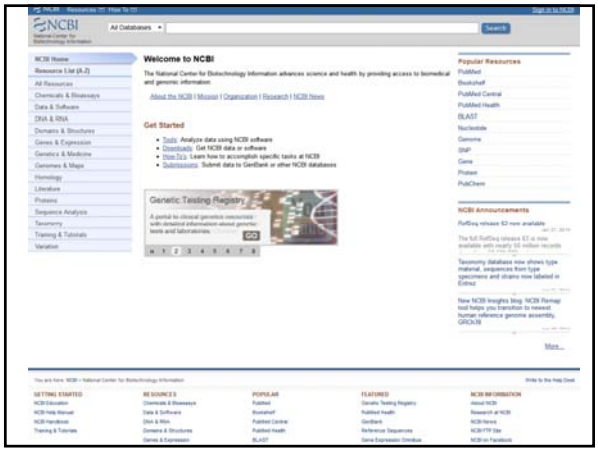
- Define “Evolution by Natural Selection” in one or two sentences
  - Variation exists in a population
  - Variation is inherited
  - Individuals reproduce more than the environment can support, therefore competition occurs
  - Differential reproduction is driven by hereditable variation allowing more efficient resource use

Pretest question #2

- Can two sequences that are aligned reasonably well over their whole lengths be 66% homologous?
  - Sequences either are or are not homologous.
  - Homologous sequences can vary in their similarity

NCBI

- www.ncbi.nlm.nih.gov
- National Center for Biotechnology Information
- Run by the US National Institute of Health National Library of Medicine
- Since 1988 has been the designated US repository for data relating to biological research
- Also runs its own research program in computational biology



Comprised of 42 databases containing together >900 million records

Table 1. The Entrez databases (as of 3 September 2013)

Database	Records	Section within this article	Data source
NCBI Web Site	31,978	Introduction	N
PubMed	23,052,796	Literature	C
PMC	2,436,592	Literature	D, C
NLM Catalog	1,463,089	Literature	C, N
MeSH	243,770	Literature	N
Books	372,232	Literature	C, N
Taxonomy*	1,153,795	Taxonomy	C, N
Nucleotide*	181,999,766	DNA and RNA	D (GenBank), C, N
EST*	74,911,096	DNA and RNA	D (GenBank)
GSS*	36,999,049	DNA and RNA	D (GenBank)
BioSample	2,180,817	DNA and RNA	N
SRA*	475,684	DNA and RNA	D
PDB*	102,118	DNA and RNA	D (GeoBank)
Protein*	94,162,424	Proteins	C, N
Protein Clusters*	202,691	Proteins	N
GEO Profiles*	91,392,791	Genes and expression	D
Probe	31,367,498	Genes and expression	D
Gene*	14,167,800	Genes and expression	C, N
UniGene*	6,867,083	Genes and expression	N
GEO Datasets*	1,044,344	Genes and expression	N
Biosystems*	522,277	Genes and expression	C
Homologene*	133,546	Genes and expression	N
Clone*	33,133,797	Genomes	D, N
LinkSNP*	343,812	Genomes	D (dbSNP)
BioProject*	98,538	Genomes	D
Assembly	17,707	Genomes	C, N
Genome*	10,929	Genomes	C, N
MapViewer/genomes*	10,811	Genomes	D
SNP*	300,238,943	Genetics and medicine	D (dbSNP), N
dbVar*	3,584,019	Genetics and medicine	D
MedGen*	169,431	Genetics and medicine	C, N
dbGaP	134,971	Genetics and medicine	D
ClinVar*	49,040	Genetics and medicine	D, N
PubMed Health	41,262	Genetics and medicine	C
GTR*	29,212	Genetics and medicine	D
OMIA	2844	Genetics and medicine	C
PubChem Substance*	119,813,046	Chemicals and bioassays	D
PubChem Compound*	47,737,896	Chemicals and bioassays	N
PubChem Bioassay*	717,420	Chemicals and bioassays	D
Structure*	92,993	Domains and structures	C, N
CDD*	48,034	Domains and structures	C, N

*Indicates that the data in this resource are available by FTP.
D, direct submission; C, collaboration/agreement; N, internal NCBI/NLM curation.

NCBI Resource Coordinators Nucl. Acids Res. 2014 42: D7-D17

GenBank

- Central database for nucleotide sequences
- Synced daily with alternative independent databases
  - European Nucleotide Archive (ENA)
  - DNA Data Bank of Japan (DDBJ)
- Linked to other NCBI resources via Entrez system

GenBank sequence types

- WGS (whole genome sequencing)
- TSA (transcriptome shotgun assembly)
- 12 standard nucleotide types
  - Split by taxon + ENV for environmental sequences
- PAT (patent sequences)
- 6 high-throughput types
  - EST (expressed sequence tag)
  - GSS (genome survey sequence)
  - HTC (high-throughput cDNA)
  - HTG (high-throughput genomic)
  - STS (sequence tagged sites)

NCBI raw nucleotide data

- SRA – Sequence Read Archive: raw next-generation sequencing data, typically underpins a GenBank genome assembly
- Trace Archive: raw Sanger sequencing data

Table 1. Growth of GenBank divisions (nucleotide base pairs)

Division	Description	Release 197 (8/2013)	Annual increase (%)[a]
WGS	Whole-genome shotgun data	500 420 412 665	62.4
TSA	Transcriptome shotgun data	8 633 123 935	49.9
PHG	Phages	119 812 712	42.5
VRL	Viruses	1 757 202 472	22.9
BCT	Bacteria	10 281 048 518	21.8
ENV	Environmental samples	3 743 277 434	10.9
INV	Invertebrates	2 737 140 646	9.8
PAT	Patented sequences	13 290 161 247	9.7
PLN	Plants	5 963 882 822	8.8
GSS	Genome survey sequences	23 726 384 753	8.1
VRT	Other vertebrates	3 068 956 026	6.3
MAM	Other mammals	911 342 025	5.6
HTG	High-throughput genomic	25 184 819 955	3.4
HTC	High-throughput cDNA	656 196 063	2.7
UNA	Unannotated	130 510	2.1
EST	Expressed sequence tags	41 665 629 009	1.9
PRI	Primates	6 425 093 034	1.7
SYN	Synthetic	941 078 074	1.4
ROD	Rodents	4 451 315 297	0.4
STS	Sequence tagged sites	636 326 479	0.0
TOTAL	All GenBank sequences	654 613 333 676	45.1

[a]Measured relative to Release 191 (8/2012).

Benson et al. Nucl. Acids Res. 2014 42: D32-D37

RefSeq

- RefSeq is a curated collection of annotated sequences
- Typically, raw nucleotide sequences come from GenBank
- Annotations are typically from NCBI, but sometimes from collaborations with the broader community
- Some quality restrictions for RefSeq annotating GenBank data

Discuss

1. What makes a database? Why use a database, e.g., vs. a spreadsheet
2. GenBank is a repository, but RefSeq is not. Why is this significant?

Discuss

3. What data types of genomic data are there?

4. How are these data types related

.gbk files

- NCBI's standard format of annotating nucleotide sequences

```
LOCUS       NR_102804               1542 bp    rRNA    linear   BCT 24-MAY-2013
DEFINITION  Escherichia coli str. K-12 substr. MG1655 strain K-12 16S ribosomal
            RNA, complete sequence.
ACCESSION   NR_102804
VERSION     NR_102804.1  GI:507147997
DBLINK      Project: 33175
            BioProject: PRJNA33175
KEYWORDS    RefSeq.
SOURCE      Escherichia coli str. K-12 substr. MG1655
  ORGANISM  Escherichia coli str. K-12 substr. MG1655
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 1542)
  AUTHORS   Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
            Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,
            Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H.,
            Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.
  TITLE     Escherichia coli K-12: a cooperatively developed annotation
            snapshot--2005
  JOURNAL   Nucleic Acids Res. 34 (1), 1-9 (2006)
   PUBMED   16397293
  REMARK    Publication Status: Online-Only
REFERENCE   2  (bases 1 to 1542)
  AUTHORS   Blattner,F.R. and Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (14-JAN-1997) Laboratory of Genetics, University of
            Wisconsin, 4250 Henry Hall, Madison, WI 53706-1580, USA
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence is identical to U00096:223771-225312.
            This record has been curated by collaboration with an international
            consortium of ribosomal RNA databases.
            COMPLETENESS: full length.
PRIMARY     REFSEQ_SPAN         PRIMARY_IDENTIFIER PRIMARY_SPAN        COMP
            1-1542              U00096.2           223771-225312
FEATURES             Location/Qualifiers
     source          1..1542
                     /organism="Escherichia coli str. K-12 substr. MG1655"
                     /mol_type="rRNA"
                     /strain="K-12"
                     /sub_strain="MG1655"
                     /db_xref="taxon:511145"
     rRNA            1..1542
                     /product="16S ribosomal RNA"
ORIGIN
        1 aaattgaaga gtttgatcat ggctcagatt gaacgctggc ggcaggcctc acacatgcaa
       61 gtcgaacggt aacaggaaga agcttgcttc tttgctgacg agtggcggac gggtgagtaa
                                   ...
     1441 agcttaacct tcgggagggc gcttcccacc ttgtgattcc tgactggggt gaagtcgtaa
     1501 caaggtaacc gtaggggaac ctgcggttgg atcacctcct ta
//
```

Accession: NCBI's identifier for this sequence record

Version: NCBI's identifier for this sequence record version

GI: NCBI's identifier for this unique document

These features are the only ones that are "stable", i.e., absolute identifiers

Locus: A name for this stretch of DNA

Length of DNA

Molecule type

GenBank division

Date of last modification

Dblink: other NCBI databases related to this record

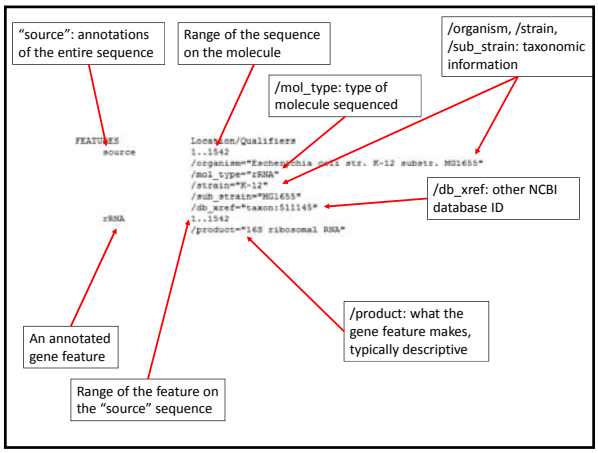
Source: the organism from which this sequence was determined

Organism: Taxonomic breakdown for the SOURCE organism, according to NCBI's Taxonomy DB

Reference: reference(s) relating to the sequence

Origin: The nucleotide sequence

"//": always marks the end of the file




```
LOCUS       NT_077402             257719 bp    DNA     linear   CON 13-AUG-2013
DEFINITION  Homo sapiens chromosome 1 genomic contig, GRCh37.p13 Primary
            Assembly.
ACCESSION   NT_077402 GPS_000125215 NT_077911
VERSION     NT_077402.2  GI:224514618
DBLINK      BioProject: PRJNA168
            Assembly: GCF_000001405.25
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 257719)
  AUTHORS   Gregory,S.G., Barlow,K.F., McLay,K.E., Kaul,R., Swarbreck,D.,
            Dunham,A., Scott,C.E., Howe,K.L., Woodfine,K., Spencer,C.C.,
            Jones,M.C., Gillson,C., Searle,S., Zhou,Y., Kokocinski,F.,
            McDonald,L., Evans,R., Phillips,K., Atkinson,A., Cooper,R.,
            Jones,C., Hall,R.E., Andrews,T.D., Lloyd,C., Ainscough,R.,
            Almeida,J.P., Ambrose,K.D., Anderson,F., Andrew,R.W., Ashwell,R.I.,
            Aubin,K., Babbage,A.K., Bagguley,C.L., Bailey,J., Beasley,H.,
            Bethel,G., Bird,C.P., Bray-Allen,S., Brown,J.Y., Brown,A.J.,
            Buckley,D., Burton,J., Bye,J., Carder,C., Chapman,J.C., Clark,S.Y.,
            Clarke,G., Clee,C., Cobley,V., Collier,R.E., Corby,N.,
            Coville,G.J., Davies,J., Deadman,R., Dunn,M., Earthrowl,M.,
            Ellington,A.G., Errington,H., Frankish,A., Frankland,J., French,L.,
            Garner,P., Garnett,J., Gay,L., Ghori,M.R., Gibson,R., Gilby,L.M.,
            Gillett,W., Glithero,R.J., Grafham,D.V., Griffiths,C.,
            Griffiths-Jones,S., Grocock,R., Hammond,S., Harrison,E.S., Hart,E.,
            Haugen,E., Heath,P.D., Holmes,S., Holt,K., Howden,P.J., Hunt,A.R.,
            Hunt,S.E., Hunter,G., Isherwood,J., James,R., Johnson,C.,
            Johnson,D., Joy,A., Kay,M., Kershaw,J.K., Kibukawa,M.,
            Kimberley,A.M., King,A., Knights,A.J., Lad,H., Laird,G., Lawlor,S.,
            Leongamornlert,D.A., Lloyd,D.M., Loveland,J., Lovell,J., Lush,M.J.,
            Lyne,R., Martin,S., Mashreghi-Mohammadi,M., Matthews,L.,
            Matthews,N.S., McLaren,S., Milne,S., Mistry,S., Moore,M.J.,
            Nickerson,T., O'Dell,C.N., Oliver,K., Palmeiri,A., Palmer,S.A.,
            Parker,A., Patel,D., Pearce,A.V., Peck,A.I., Pelan,S., Phelps,K.,
            Phillimore,B.J., Plumb,R., Rajan,J., Raymond,C., Rouse,G.,
            Saenphimmachak,C., Sehra,H.K., Sheridan,E., Shownkeen,R., Sims,S.,
```

A scaffold from the human genome project

```
            Skuce,C.D., Smith,M., Steward,C., Subramanian,S., Sycamore,N.,
            Tracey,A., Tromans,A., Van Helmond,Z., Wall,M., Wallis,J.M.,
            White,S., Whitehead,S.L., Wilkinson,J.E., Willey,D.L., Williams,H.,
            Wilming,L., Wray,P.W., Wu,Z., Coulson,A., Vaudin,M., Sulston,J.E.,
            Durbin,R., Hubbard,T., Wooster,R., Dunham,I., Carter,N.P.,
            McVean,G., Ross,M.T., Harrow,J., Olson,M.V., Beck,S., Rogers,J.,
            Bentley,D.R., Banerjee,R., Bryant,S.P., Burford,D.C., Burrill,W.D.,
            Clegg,S.M., Dhami,P., Dovey,O., Faulkner,L.M., Gribble,S.M.,
            Langford,C.F., Pandian,R.D., Porter,K.M. and Prigmore,E.
  TITLE     The DNA sequence and biological annotation of human chromosome 1
  JOURNAL   Nature 441 (7091), 315-321 (2006)
   PUBMED   16710414
  REMARK    Erratum:[Nature. 2006 Oct 26;443(7114):1013. Banerjee, R [added];
            Bryant, SP [added]; Burford, DC [added]; Burrill, WDH [added];
            Clegg, SM [added]; Dhami, P [added]; Dovey, O [added]; Faulkner, LM
            [added]; Gribble, SM [added]; Langford, CF [added]; Pandian, RD
            [added]; Porter, KM [added]; Prigmore, E [added]]
REFERENCE   2  (bases 1 to 257719)
  CONSRTM   International Human Genome Sequencing Consortium
  TITLE     Finishing the euchromatic sequence of the human genome
  JOURNAL   Nature 431 (7011), 931-945 (2004)
   PUBMED   15496913
REFERENCE   3  (bases 1 to 257719)
  AUTHORS   Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C.,
            Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., Funke,R.,
            Gage,D., Harris,K., Heaford,A., Howland,J., Kann,L., Lehoczky,J.,
            LeVine,R., McEwan,P., McKernan,K., Meldrim,J., Mesirov,J.P.,
            Miranda,C., Morris,W., Naylor,J., Raymond,C., Rosetti,M.,
            Santos,R., Sheridan,A., Sougnez,C., Stange-Thomann,N.,
            Stojanovic,N., Subramanian,A., Wyman,D., Rogers,J., Sulston,J.,
            Ainscough,R., Beck,S., Bentley,D., Burton,J., Clee,C., Carter,N.,
            Coulson,A., Deadman,R., Deloukas,P., Dunham,A., Dunham,I.,
            Durbin,R., French,L., Grafham,D., Gregory,S., Hubbard,T.,
            Humphray,S., Hunt,A., Jones,M., Lloyd,C., McMurray,A., Matthews,L.,
            Mercer,S., Milne,S., Mullikin,J.C., Mungall,A., Plumb,R., Ross,M.,
            Shownkeen,R., Sims,S., Waterston,R.H., Wilson,R.K., Hillier,L.W.,
            McPherson,J.D., Marra,M.A., Mardis,E.R., Fulton,L.A.,
            Chinwalla,A.T., Pepin,K.H., Gish,W.R., Chissoe,S.L., Wendl,M.C.,
            Delehaunty,K.D., Miner,T.L., Delehaunty,A., Kramer,J.B., Cook,L.L.,
            Fulton,R.S., Johnson,D.L., Minx,P.J., Clifton,S.W., Hawkins,T.,
```

The human genome project had a few authors!

```
            Branscomb,E., Predki,P., Richardson,P., Wenning,S., Slezak,T.,
            Doggett,N., Cheng,J.F., Olsen,A., Lucas,S., Elkin,C.,
            Uberbacher,E., Frazier,M., Gibbs,R.A., Muzny,D.M., Scherer,S.E.,
            Bouck,J.B., Sodergren,E.J., Worley,K.C., Rives,C.M., Gorrell,J.H.,
            Metzker,M.L., Naylor,S.L., Kucherlapati,R.S., Nelson,D.L.,
            Weinstock,G.M., Sakaki,Y., Fujiyama,A., Hattori,M., Yada,T.,
            Toyoda,A., Itoh,T., Kawagoe,C., Watanabe,H., Totoki,Y., Taylor,T.,
            Weissenbach,J., Heilig,R., Saurin,W., Artiguenave,F., Brottier,P.,
            Bruls,T., Pelletier,E., Robert,C., Wincker,P., Smith,D.R.,
            Doucette-Stamm,L., Rubenfield,M., Weinstock,K., Lee,H.M.,
            Dubois,J., Rosenthal,A., Platzer,M., Nyakatura,G., Taudien,S.,
            Rump,A., Yang,H., Yu,J., Wang,J., Huang,G., Gu,J., Hood,L.,
            Rowen,L., Madan,A., Qin,S., Davis,R.W., Federspiel,N.A.,
            Abola,A.P., Proctor,M.J., Myers,R.M., Schmutz,J., Dickson,M.,
            Grimwood,J., Cox,D.R., Olson,M.V., Kaul,R., Raymond,C., Shimizu,N.,
            Kawasaki,K., Minoshima,S., Evans,G.A., Athanasiou,M., Schultz,R.,
            Roe,B.A., Chen,F., Pan,H., Ramser,J., Lehrach,H., Reinhardt,R.,
            McCombie,W.R., de la Bastide,M., Dedhia,N., Blocker,H.,
            Hornischer,K., Nordsiek,G., Agarwala,R., Aravind,L., Bailey,J.A.,
            Bateman,A., Batzoglou,S., Birney,E., Bork,P., Brown,D.G.,
            Burge,C.B., Cerutti,L., Chen,H.C., Church,D., Clamp,M.,
            Copley,R.R., Doerks,T., Eddy,S.R., Eichler,E.E., Furey,T.S.,
            Galagan,J., Gilbert,J.G., Harmon,C., Hayashizaki,Y., Haussler,D.,
            Hermjakob,H., Hokamp,K., Jang,W., Johnson,L.S., Jones,T.A.,
            Kasif,S., Kaspryzk,A., Kennedy,S., Kent,W.J., Kitts,P.,
            Koonin,E.V., Korf,I., Kulp,D., Lancet,D., Lowe,T.M., McLysaght,A.,
            Mikkelsen,T., Moran,J.V., Mulder,N., Pollara,V.J., Ponting,C.P.,
            Schuler,G., Schultz,J., Slater,G., Smit,A.F., Stupka,E.,
            Szustakowski,J., Thierry-Mieg,D., Thierry-Mieg,J., Wagner,L.,
            Wallis,J., Wheeler,R., Williams,A., Wolf,Y.I., Wolfe,K.H.,
            Yang,S.P., Yeh,R.F., Collins,F., Guyer,M.S., Peterson,J.,
            Felsenfeld,A., Wetterstrand,K.A., Patrinos,A., Morgan,M.J., de
            Jong,P., Catanese,J.J., Osoegawa,K., Shizuya,H., Choi,S. and
            Chen,Y.J.
  CONSRTM   International Human Genome Sequencing Consortium
  TITLE     Initial sequencing and analysis of the human genome
  JOURNAL   Nature 409 (6822), 860-921 (2001)
   PUBMED   11237011
  REMARK    Erratum:[Nature 2001 Aug 2;412(6846):565]
```

```
COMMENT     REFSEQ INFORMATION: The reference sequence is identical to
            GL000001.1.
            On or before Jun 10, 2009 this sequence version replaced
            gi:29794400, gi:29794392.
            Assembly Name: GRCh37.p13 Primary Assembly
            The DNA sequence is composed of genomic sequence, primarily
            finished clones that were sequenced as part of the Human Genome
            Project. PCR products and WGS shotgun sequence have been added
            where necessary to fill gaps or correct errors. All such additions
            are manually curated by GRC staff. For more information see:
            http://genomereference.org.

            ##Genome-Annotation-Data-START##
            Annotation Provider         :: NCBI
            Annotation Status           :: Full annotation
            Annotation Version          :: Homo sapiens Annotation Release 105
            Annotation Pipeline         :: NCBI eukaryotic genome annotation
                                           pipeline
            Annotation Software Version :: 5.1
            Annotation Method           :: Best-placed RefSeq; Gnomon
            Features Annotated          :: Gene; mRNA; CDS; ncRNA
            ##Genome-Annotation-Data-END##
FEATURES             Location/Qualifiers
     source          1..257719
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="1"
```

Annotation information
division-specific methods

Source field (sequence overview)

```
     gene            1874..4409
                     /gene="DDX11L1"
                     /note="DEAD/H (Asp-Glu-Ala-Asp/His) box helicase 11 like
                     1; Derived by automated computational analysis using gene
                     prediction method: BestRefSeq."
                     /pseudo
                     /db_xref="GeneID:100287102"
                     /db_xref="HGNC:37102"
     misc_RNA        join(1874..2227,2613..2721,3221..4409)
                     /gene="DDX11L1"
                     /product="DEAD/H (Asp-Glu-Ala-Asp/His) box helicase 11
                     like 1"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /pseudo
                     /transcript_id="NR_046018.2"
                     /db_xref="GI:389886562"
                     /db_xref="GeneID:100287102"
                     /db_xref="HGNC:37102"
     gene            complement(<4362..19370)
                     /gene="WASH7P"
                     /gene_synonym="FAM39F; WASH5P"
                     /note="WAS protein family homolog 7 pseudogene; Derived by
                     automated computational analysis using gene prediction
                     method: BestRefSeq."
                     /pseudo
                     /db_xref="GeneID:653635"
                     /db_xref="HGNC:38034"
     misc_RNA        complement(join(<4362..4829,4970..5038,5796..5947,
                     6607..6765,6858..7055,7233..7368,7606..7742,7915..8061,
                     8268..8366,14738..14891,19321..19370))
                     /gene="WASH7P"
                     /gene_synonym="FAM39F; WASH5P"
                     /product="WAS protein family homolog 7 pseudogene"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /pseudo
                     /transcript_id="NR_024540.1"
                     /db_xref="GI:215277009"
                     /db_xref="GeneID:653635"
```

A couple of pseudogenes

Note strand information, splice junctions

```
     STS             4596..4719
                     /standard_name="G54113"
                     /db_xref="UniSTS:109350"
     gene            20366..20503
                     /gene="MIR1302-2"
                     /gene_synonym="hsa-mir-1302-2; MIRN1302-2"
                     /note="microRNA 1302-2; Derived by automated computational
                     analysis using gene prediction method: BestRefSeq."
                     /db_xref="GeneID:100302278"
                     /db_xref="HGNC:35294"
                     /db_xref="miRBase:MI0006363"
     precursor_RNA   20366..20503
                     /gene="MIR1302-2"
                     /gene_synonym="hsa-mir-1302-2; MIRN1302-2"
                     /product="microRNA 1302-2"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /transcript_id="NR_036051.1"
                     /db_xref="GI:301173067"
                     /db_xref="GeneID:100302278"
                     /db_xref="HGNC:35294"
                     /db_xref="miRBase:MI0006363"
     ncRNA           20438..20458
                     /gene="MIR1302-2"
                     /gene_synonym="hsa-mir-1302-2; MIRN1302-2"
                     /ncRNA_class="miRNA"
                     /product="hsa-miR-1302"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /db_xref="miRBase:MIMAT0005890"
                     /db_xref="GeneID:100302278"
                     /db_xref="HGNC:35294"
                     /db_xref="miRBase:MI0006363"
```

Sequence tagged site

ncRNA gene that undergoes processing

```
     gene            59091..60008
                     /gene="OR4F5"
                     /note="olfactory receptor, family 4, subfamily F, member
                     5; Derived by automated computational analysis using gene
                     prediction method: BestRefSeq."
                     /db_xref="GeneID:79501"
                     /db_xref="HGNC:14825"
                     /db_xref="HPRD:14974"
     mRNA            59091..60008
                     /gene="OR4F5"
                     /product="olfactory receptor, family 4, subfamily F,
                     member 5"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /transcript_id="NM_001005484.1"
                     /db_xref="GI:53828739"
                     /db_xref="GeneID:79501"
                     /db_xref="HGNC:14825"
                     /db_xref="HPRD:14974"
     CDS             59091..60008
                     /gene="OR4F5"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /codon_start=1
                     /product="olfactory receptor 4F5"
                     /protein_id="NP_001005484.1"
                     /db_xref="GI:53828740"
                     /db_xref="CCDS:CCDS30547.1"
                     /db_xref="GeneID:79501"
                     /db_xref="HGNC:14825"
                     /db_xref="HPRD:14974"
                     /translation="MVTEFIFLGLSDSQELQTFLFMLFFVFYGGIVFGNLLIVITVVS
                     DSHLHSPMYFLLANLSLIDLSLSSVTAPKMITDFFSQRKVISFKGCLVQIFLLHFFGG
                     SEMVILIAMGFDRYIAICKPLHYTTIMCGNACVGIMAVTWGIGFLHSVSQLAFAVHLL
                     FCGPNEVDSFYCDLPRVIKLACTDTYRLDIMVIANSGVLTVCSFVLLIISYTIILMTI
                     QHRPLDKSSKALSTLTAHITVVLLFFGPCVFIYAWPFPIKSLDKFLAVFYSVITPLLN
                     PIIYTLRNKDMKTAIRQLRKWDAHSSVKF"
```

A protein-coding gene

Note translation

```
     gene            complement(132447..164392)
                     /gene="LOC100996442"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Gnomon."
                     /db_xref="GeneID:100996442"
     misc_RNA        complement(join(132447..133011,145767..145831,
                     154263..154791,155884..155942,158100..158165,
                     163753..164392))
                     /gene="LOC100996442"
                     /product="uncharacterized LOC100996442, transcript variant
                     X2"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Gnomon. Supporting evidence
                     includes similarity to: 1 EST, and 88% coverage of the
                     annotated genomic feature by RNAseq alignments, including
                     3 samples with support for all annotated introns"
                     /transcript_id="XR_246629.1"
                     /db_xref="GI:530360230"
                     /db_xref="GeneID:100996442"
     misc_RNA        complement(join(148340..148674,154263..154791,
                     155884..155942,158100..158165,163753..164392))
                     /gene="LOC100996442"
                     /product="uncharacterized LOC100996442, transcript variant
                     X1"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Gnomon. Supporting evidence
                     includes similarity to: 4 ESTs, and 86% coverage of the
                     annotated genomic feature by RNAseq alignments, including
                     4 samples with support for all annotated introns"
                     /transcript_id="XR_159064.3"
                     /db_xref="GI:530360229"
                     /db_xref="GeneID:100996442"
```

A gene with 2 splice varients

```
     assembly_gap    167418..217417
                     /estimated_length=50000
                     /gap_type="within scaffold"
                     /linkage_evidence="unspecified"
     gene            complement(217770..218778)
                     /gene="RPL23AP21"
                     /gene_synonym="RPL23A_1_1"
                     /note="ribosomal protein L23a pseudogene 21; Derived by
                     automated computational analysis using gene prediction
                     method: Curated Genomic."
                     /pseudo
                     /db_xref="GeneID:728481"
                     /db_xref="HGNC:35827"
CONTIG      join(complement(AP006221.1:36117..36731),AL627309.15:103..166904,
            gap(50000),AP006222.1:1..40302)
//
```

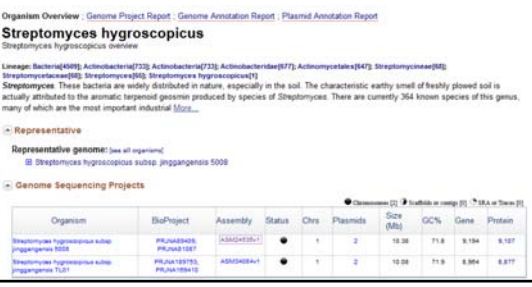
Gap in the sequence

Another pseudogene

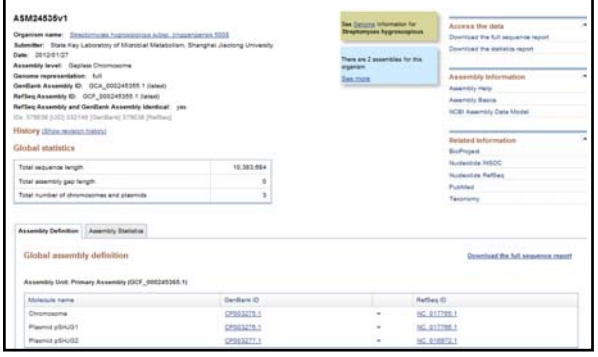
- Note that instead of the nucleotide sequences there is instead a reference to several sequences joined together and a gap
  - Reflects that the “completed genome” is anything but!

NCBI genomes: Bioprojects

- Allow multiple related data to be linked together



NCBI genomes: Assembly

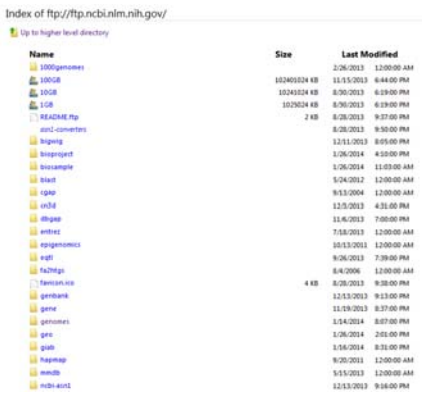


NCBI genomes

- Frankly, this is a bit of a mess on the website
  - Often, records are not in their complete form, i.e., instead consisting of links to subordinate data types, e.g., assembly->nucleotide->protein
  - But sometimes these are from GenBank, other times from Refseq
- Genbank, Refseq, Entrez data structures to not overlap in a simple manner
- If you want to download genomic data, this is not the way to do it!

NCBI FTP portal

- ftp://ftp.ncbi.nlm.nih.gov/
- Allows access both from the web and from Terminal (using the program `ftp`)
- Includes all of NCBI's data structures as separate database folders



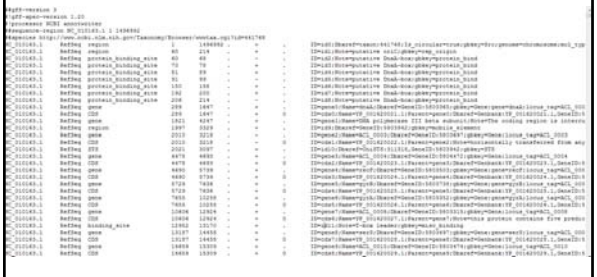
.fna: chromosome nucleotides
.ffn: genes
.faa: proteins
.gbk: annotated chromosomes
.gbs: scaffolds
.gff: all annotations
.ptt: protein annotations
.rnt: RNA annotations (not mRNA)

gff (Genome Feature Format)

- A compact tabular annotation format, especially common for larger eukaryotic genomes
- Does not include sequences
- An input file type for many genomics programs (particularly sequence viewers)

Columns:

1: Seq ID	6: Score
2: Source	7: Strand
3: Type	8: Phase
4: Start	9: Attribues (itself subdivided)
5: End	

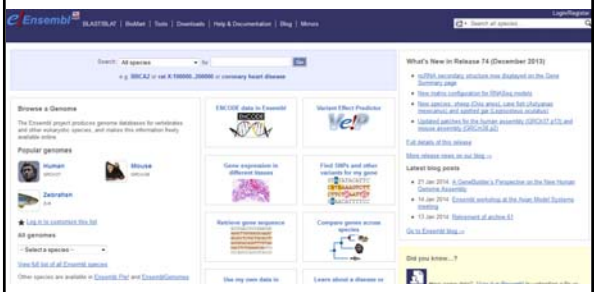


NCBI FTP portal

- Fungi have a similar download structure as bacteria
- Other Eukaryotic genomes unfortunately do not have a standard format in this folder, meaning you have to dig through the Refseq and Genbank folders
- Genome reports have helpful tables linking all this information together: ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/
- Unfortunately, @$*%# NCBI did not make them tab-delimited so you can't easily use them in Excel, or even Perl!

IMG: Integrated Microbial Genomes http://img.jgi.doe.gov/

Ensembl: www.ensembl.org



- Next: Dr. Jill Wegrzyn, Biotechnology Center Cluster
- Reminder: Assignments due Wed before class
  - Note: I will not be very available to answer questions Tues evening
- This Wed: Introduction to Perl part 2