

Genome Annotations

Michael Schatz

Nov 15, 2014
Adv Sequencing Course



Goal: Genome Annotations

aatgcatgcggctatgcta atgcatgcggctatgcta agctggatccgat gaca atgc atgcggctatgcta
tgc atgcggctatgca agctggatccgat gactatgcta agctggatccgat gaca atgc atgcggctatg
taatgaatggtcttggatt tacctt ggaatgcta agctggatccgat gaca atgc atgcggctatgcta
atggtcttggatt tacctt ggaatatgcta atgc atgcggctatgcta agctggatccgat gaca atgc
ggctatgcta atgc atgcggctatgca agctggatccgat gactatgcta agctgcggctatgcta
cggtatgcta agctggatccgat gaca atgc atgcggctatgcta atgc atgcggctatgca
agctggatccgat gaca atgc atgcggctatgcta atgc atgcggctatgca
ctgcggctatgcta atga atggtcttggatt tacctt ggaatgcta agctggatccgat gaca atgc
atgcggctatgcta atga atggtcttggatt tacctt ggaatatgcta atgc atgcggctatgcta
cggtatgcta agctggatccgat gaca atgc atgcggctatgcta atgc atgcggctatgca
cgatgactatgcta agctgcggctatgcta atgc atgcggctatgcta agctcatgc
ggctatgcta agctggatccgat gaca atgc atgcggctatgcta atgc atgcggctatgca
aatgc atgcggctatgcta agctggatccgat gaca atgc atgcggctatgcta atgc
atgcggctatgcta agctggatccgat gaca atgc atgcggctatgcta agctcg
ggctatgcta agctggatccgat gaca atgc atgcggctatgcta agctggatccgat
atggtcttggatt tacctt ggaatgcta agctggatccgat gaca atgc atgcggctatgcta
atga atggtcttggatt tacctt ggaatatgcta atgc atgcggctatgcta
ttggatccgat gactatgcta agctgcggctatgcta atgc atgcggctatgcta
agctggatccgat gaca atgc atgcggctatgcta atgc atgcggctatgcta
tgatccgat gaca atgc atgcggctatgcta atgc atgcggctatgca
tgatccgat gaca atgc atgcggctatgcta agctggatccgat gactatgcta
tgatccgat gaca atgc atgcggctatgcta atgc atgcggctatgca

Goal: Genome Annotations

aatgcatcggtatgctaattgcattgcggctatgctaagctggatccgatgacaatgcattgcggctatgctaa
tgcatcggtatgcaagctggatccgatgactatgctaagctggatccgatgacaatgcattgcggctatgc
taatgaatggtcttggatttaccttggaatatgctaagctggatccgatgacaatgcattgcggctatgctaatga
atggtcttggatttaccttggaatatgctaattgcattgcggctatgctaagctggatccgatgacaatgcattgc
ggctatgctaattgcattgcggctatgcaagctggatccgatgactatgctaagctggatccgatgacaatgcattgc
ggctatgctaagctggatccgatgacaatgcattgcggctatgctaattgcattgcggctatgctaagctggatc
ctggatccgatgacaatgcattgcggctatgctaattgcattgcggctatgctaagctggatccgatgacaatgcattgc
ctatgctaattgcattgcggctatgctaattgcattgcggctatgctaattgcattgcggctatgctaagctggatc
cggtatgctaagctggatccgatgacaatgcattgcggctatgctaattgcattgcggctatgctaagctggatc
ctggatccgatgacaatgcattgcggctatgctaattgcattgcggctatgctaattgcattgcggctatgctaagctgg
aatgcatcggtatgctaagctggatccgatgacaatgcattgcggctatgctaattgcattgcggctatgctaag
ctggatccgatgactatgctaagctggatccgatgacaatgcattgcggctatgctaagctggatccgatgacaatga
atggtcttggatttaccttggaatatgctaattgcattgcggctatgctaagctggatccgatgacaatgcattgcggctatg
ttggatccgatgacaatgcattgcggctatgctaagctggatccgatgacaatgcattgcggctatgctaagctgg
gatccgatgacaatgcattgcggctatgctaattgcattgcggctatgctaagctggatccgatgacaatgcattgcgg
tgccatgctaattgcattgcggctatgctaattgcattgcggctatgctaagctggatccgatgacaatgcattgcgg

Gene!



Outline

1. Alignment to other genomes
2. Prediction aka “Gene Finding”
3. Experimental & Functional Assays
4. Online Resources



Outline

1. Alignment to other genomes
2. Prediction aka “Gene Finding”
3. Experimental & Functional Assays
4. Online Resources

Basic Local Alignment Search Tool

- Rapidly compare a sequence Q to a database to find all sequences in the database with a score above some cutoff S .
 - Which protein is most similar to a newly sequenced one?
 - Where does this sequence of DNA originate?
- Speed achieved by using a procedure that typically finds “most” matches with scores $> S$.
 - Tradeoff between sensitivity and specificity/speed
 - Sensitivity – ability to find all related sequences
 - Specificity – ability to reject unrelated sequences

(Altschul et al. 1990)

Seed and Extend

FAKDFLAGGVAAAISKTAVAPIERVKLLLQVQHASKQITADKQYKGIIDCVVRIPKEQGV
F D +GG AAA+SKTAVAPIERVKLLLQVQ ASK I DK+YKGI+D ++R+PKEQGV
FLIDLASGGTAAAVSKTAVAPIERVKLLLQVQDASKAIAVDKRYKGIMDVLIRVPKEQGV

- Homologous sequences are likely to contain a **short high scoring word pair**, a seed.
 - Smaller seed sizes make the sense more sensitive, but also (much) slower
 - Typically do a fast search for prototypes, but then most sensitive for final result
- BLAST then tries to extend high scoring word pairs to compute **high scoring segment pairs** (HSPs).
 - Significance of the alignment reported via an e-value

BLAST E-values

E-value = the number of HSPs having alignment score **S** (or higher) expected to occur **by chance**.

- Smaller E-value, more significant in statistics
- Bigger E-value, less significant
- Over 1 means expect this totally by chance
(not significant at all!)

The expected number of HSPs with the score at least **S** is :

$$E = K * n * m * e^{-\lambda S}$$

K, λ are constant depending on model

n, m are the length of query and sequence

E-values quickly drop off for better alignment bits scores

Very Similar Sequences

Query: HBA_HUMAN Hemoglobin alpha subunit

Sbjct: HBB_HUMAN Hemoglobin beta subunit

Score = 114 bits (285), Expect = 1e-26

Identities = 61/145 (42%), Positives = 86/145 (59%), Gaps = 8/145 (5%)

Query 2 LSPADKTNVKAAGKVGAAHAGEYGAEALERMFLSFPTTKTYFPHF-----DLSHGSAQV 55
L+P +K+ V A WGKV + E G EAL R+ + +P T+ +F F D G+ +V

Sbjct 3 LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKV 60

Query 56 KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRDPVNFKLLSHCLLVTLAHLPA 115
K HGKKV A ++ +AH+D++ + LS+LH KL VDP NF+LL + L+ LA H

Sbjct 61 KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLGNVLVCVLAHHFGK 120

Query 116 EFTP AVHASLDKFLASVSTVLTSKY 140

EFTP V A+ K +A V+ L KY

Sbjct 121 EFTPPVQAAYQKV VAGVANALAHKY 145

Quite Similar Sequences

Query: HBA_HUMAN Hemoglobin alpha subunit

Sbjct: MYG_HUMAN Myoglobin

Score = 51.2 bits (121), Expect = 1e-07,

Identities = 38/146 (26%), Positives = 58/146 (39%), Gaps = 6/146 (4%)

Query 2 LSPADKTNVKAAWGKVGAHAGEYGAELERMFLSFPTTKTYFPF-----DLSHGSAQV 55
LS + V WGKV A +G E L R+F P T F F D S +

Sbjct 3 LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFHKHLKSEDEMKAEDL 62

Query 56 KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRDPVNFKLLSHCLLVTAAHLPA 115
K HG V AL + + L+ HA K ++ + +S C++ L + P

Sbjct 63 KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLOSKHPG 122

Query 116 EFTPRAVHASLDKFLASVSTVLTSKYR 141
+F +++K L + S Y+

Sbjct 123 DFGADAQGAMNKALELFRKDMASNYK 148

Not similar sequences

Query: HBA_HUMAN Hemoglobin alpha subunit

Sbjct: SPAC869.02c [Schizosaccharomyces pombe]

Score = 33.1 bits (74), Expect = 0.24

Identities = 27/95 (28%), Positives = 50/95 (52%), Gaps = 10/95 (10%)

Query 30 ERMFLSFPTTKTYFPHFDSLHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAH 89
++M ++P P+F+ +H + +A AL N ++DD+ +LSA D

Sbjct 59 QKMLGNYPEV---LPYFNKAHQISL--SQPRILAFALLNYAKNIDDL-TSLSAFMDQIVV 112

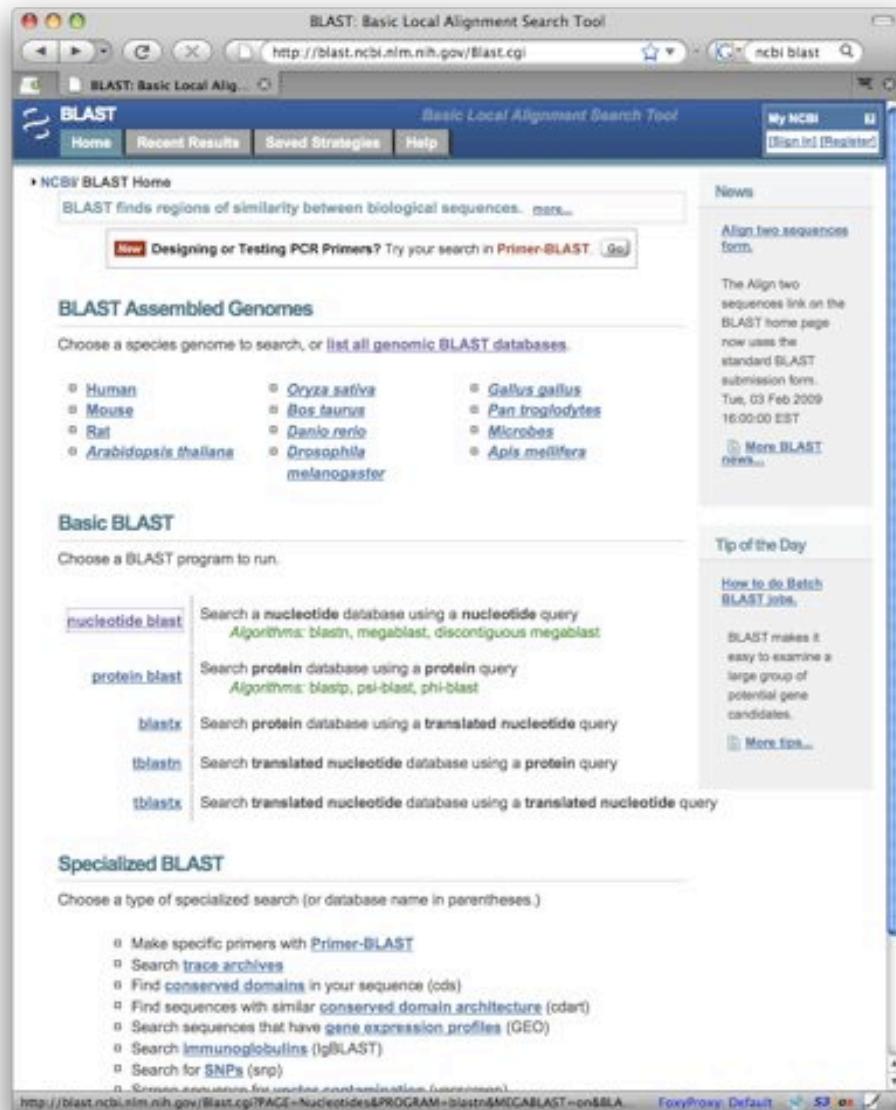
Query 90 K---LRVDPVNFKLLSHCLLVTAAHLPAEF-TPA 120
K L++ ++ ++ HCLL T+ LP++ TPA

Sbjct 113 KHVGQLQIKAEHYPIVGHCLLSTMQELLPSDVATPA 147

Blast Versions

Program	Database	Query
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Protein	Nucleotide translated in to protein
TBLASTN	Nucleotide translated in to protein	Protein
TBLASTX	Nucleotide translated in to protein	Nucleotide translated in to protein

NCBI Blast



- Nucleotide Databases
 - nr:All Genbank
 - refseq: Reference organisms
 - wgs:All reads

- Protein Databases
 - nr:All non-redundant sequences
 - Refseq: Reference proteins

Genomic Coordinates

What are coordinates of “TAC”
in GATTACA?

1-based coordinates

- Base 4 through 6: [4,6] “closed”
- Base 4 through 7: [4,7) “half-open”
- 3 bases starting at base 4: [4, +3]

GAT**TAC**A
1234567

0-based coordinates

- Position 3 through 5: [3,5] “closed”
- Position 3 through 6: [3,6) “half-open”
- 3 bases starting at position 3: [3, +3]

GAT**TAC**A
0123456

Genomic Conventions

1-based coordinates

- BLAST/MUMmer alignments
- Ensembl Genome Browser
- SAM,VCF, GFF and Wiggle

GATTYACA
1234567

0-based coordinates

- BAM, BCFv2, BED, and PSL
- UCSC Genome Browser
- C/C++, Perl, Python, Java

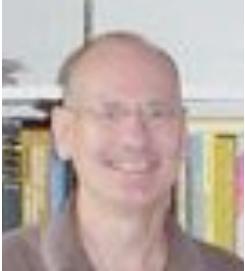
GATTYACA
0123456

Always double check the manual!
You will get this wrong someday 😞



Outline

1. Alignment to other genomes
2. Prediction aka “Gene Finding”
3. Experimental & Functional Assays
4. Online Resources



Bacterial Gene Finding and Glimmer

(also Archaeal and viral gene finding)

Arthur L. Delcher and Steven Salzberg
Center for Bioinformatics and Computational Biology
Johns Hopkins University School of Medicine

Step One

- Find open reading frames (ORFs).

The diagram illustrates a DNA sequence with a green rectangular box highlighting a specific segment. Within this segment, a red arrow points to the first codon, labeled "Start codon". At the far right end of the highlighted segment, two red arrows point to consecutive codons, both labeled "Stop codon". The sequence itself is composed of various nucleotides, some of which are highlighted in yellow or orange, indicating specific base pairs.

Start codon

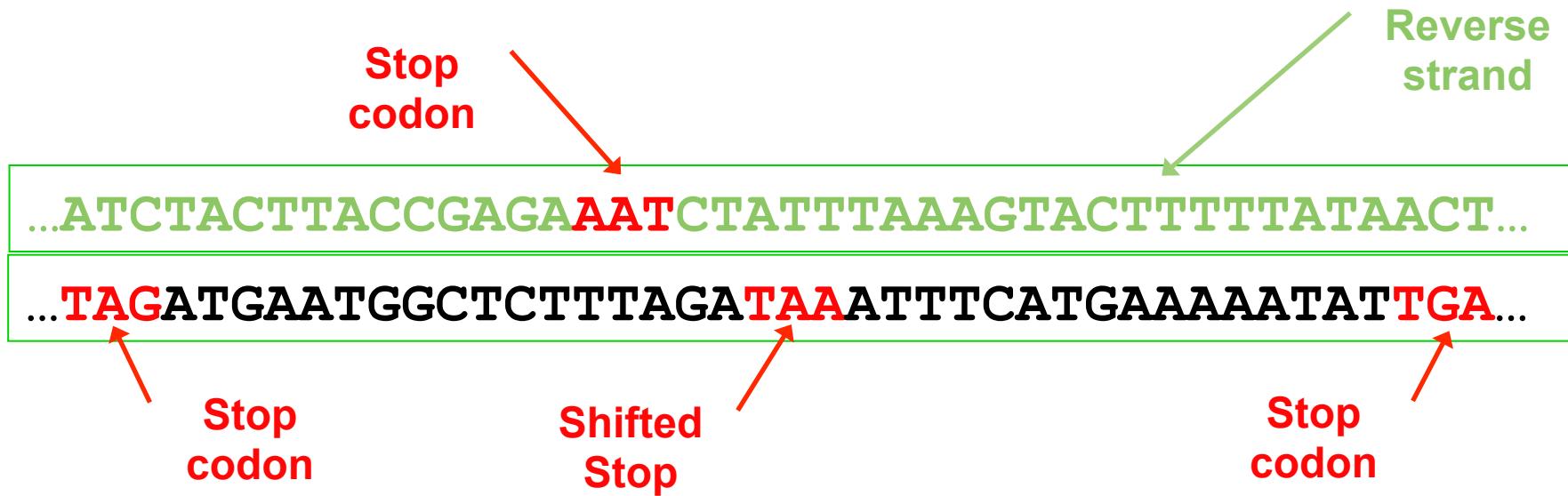
Stop codon

Stop codon

...TAGATGAATGGCTCTTTAGATAAATTTCATGAAAAATATTGA...

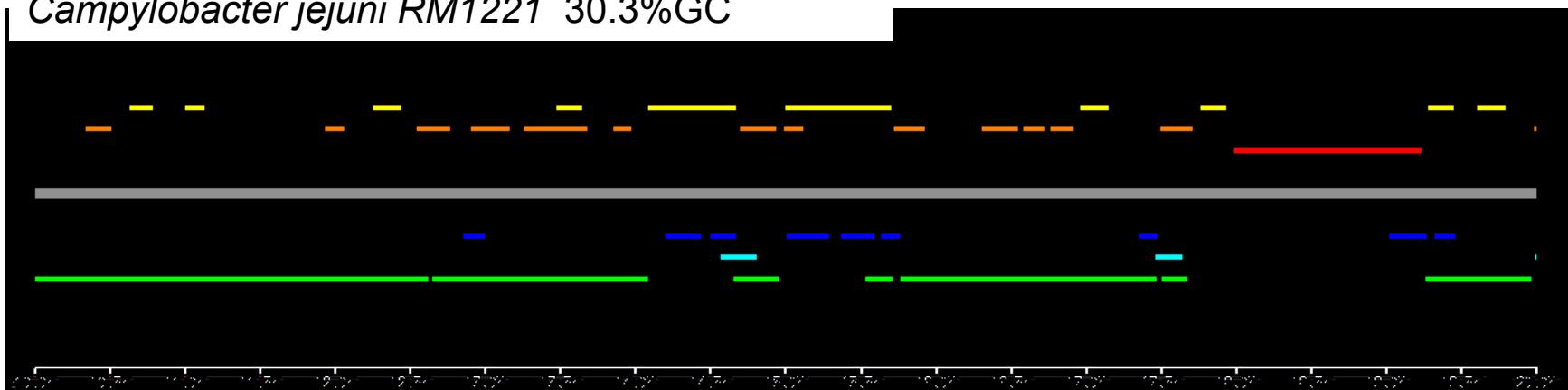
Step One

- Find open reading frames (ORFs).



- But ORFs generally overlap ...

Campylobacter jejuni RM1221 30.3%GC

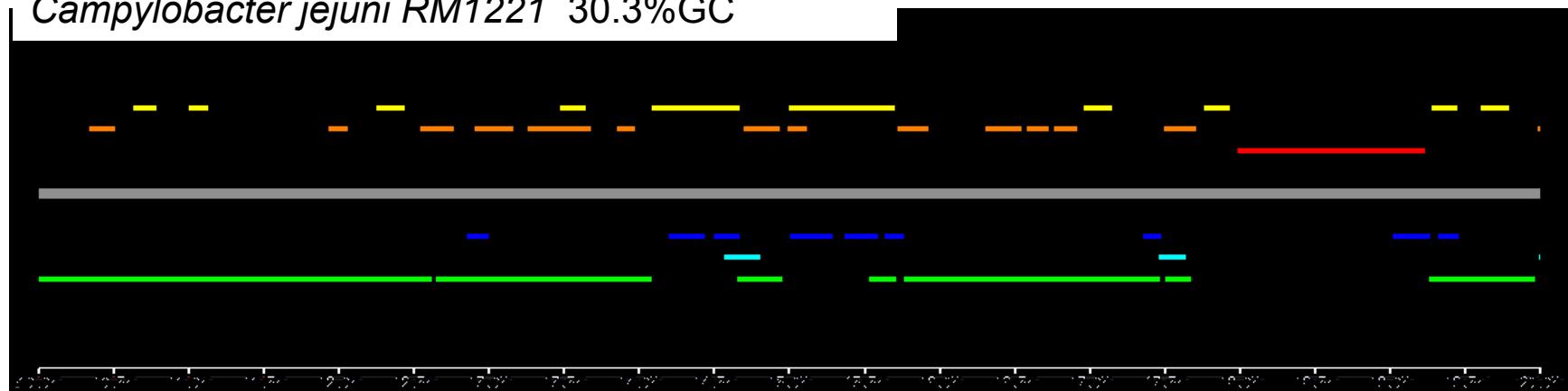


All ORFs longer than 100bp on both strands shown
- color indicates reading frame
Longest ORFs likely to be protein-coding genes

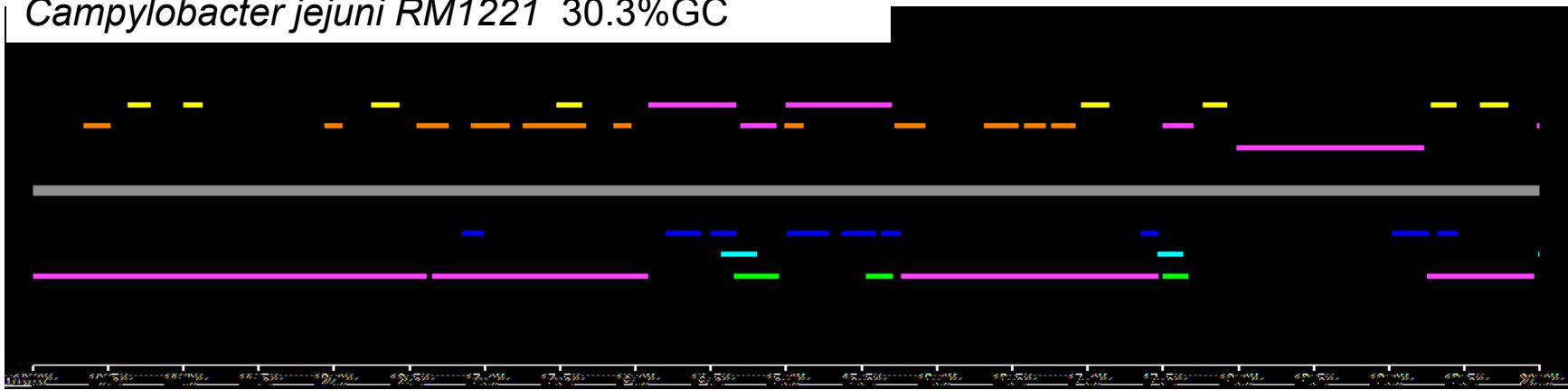
Note the low GC content

All genes are ORFs but not all ORFs are genes

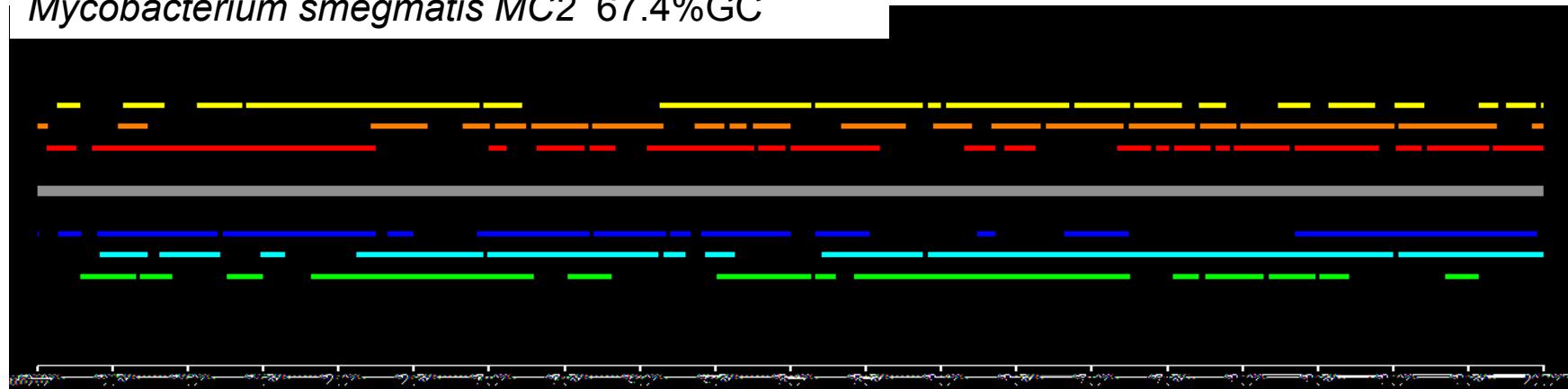
Campylobacter jejuni RM1221 30.3%GC



Campylobacter jejuni RM1221 30.3%GC

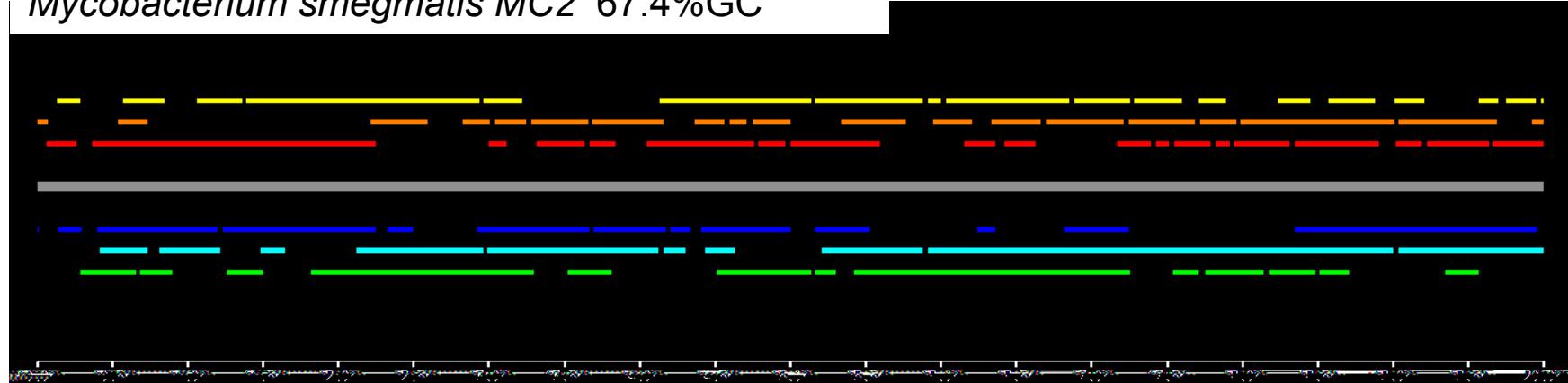


Mycobacterium smegmatis MC2 67.4%GC

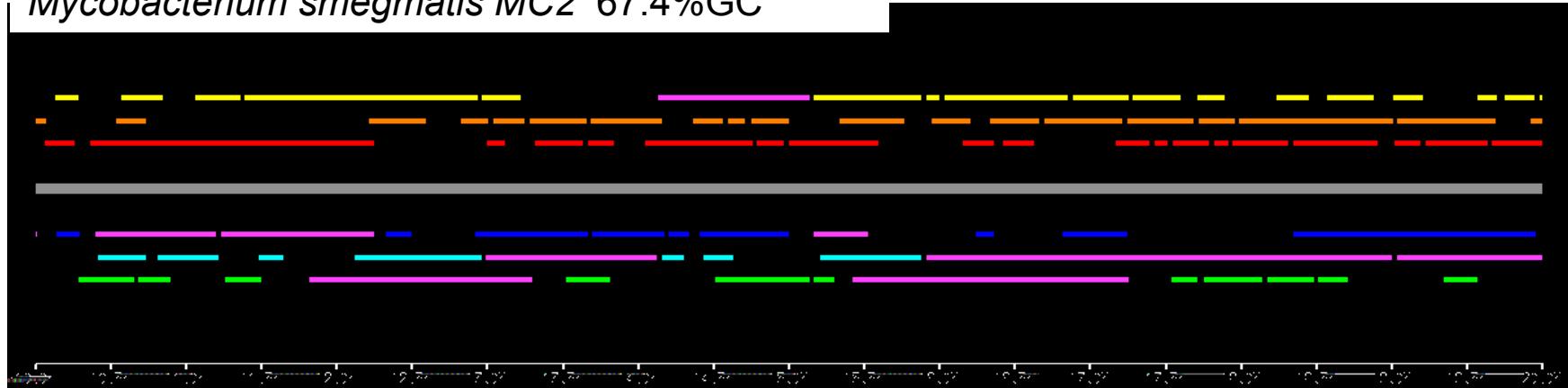


Note what happens in a high-GC genome

Mycobacterium smegmatis MC2 67.4%GC



Mycobacterium smegmatis MC2 67.4%GC



Probabilistic Methods

- Create models that have a probability of generating any given sequence.
 - Evaluate gene/non-genome models against a sequence
- Train the models using examples of the types of sequences to generate.
 - Use RNA sequencing, homology, or “obvious” genes
- The “score” of an orf is the probability of the model generating it.
 - Most basic technique is to count how kmers occur in known genes versus intergenic sequences
 - More sophisticated methods consider variable length contexts, “wobble” bases, other statistical clues

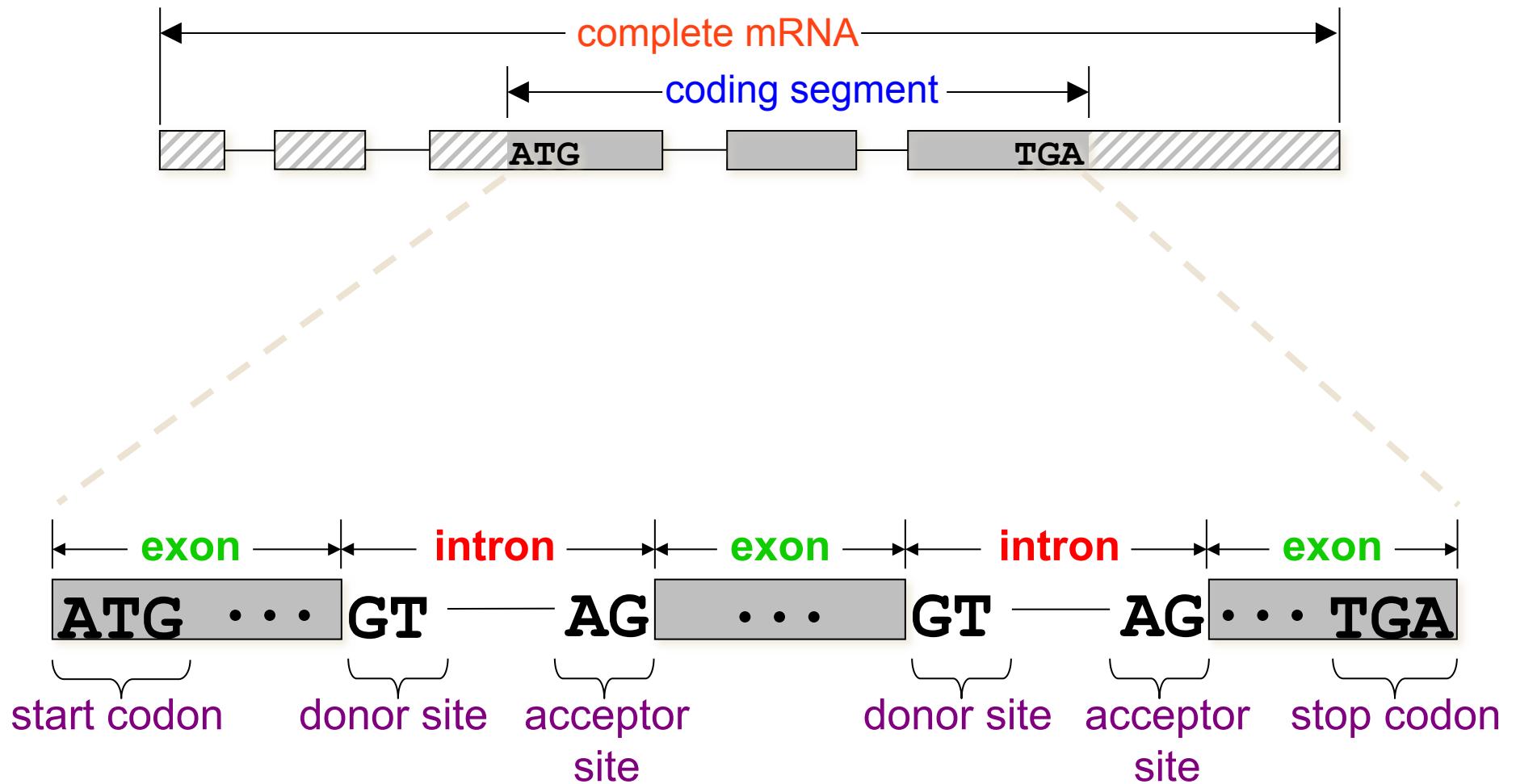


Overview of Eukaryotic Gene Prediction

CBB 231 / COMPSCI 261

W.H. Majoros

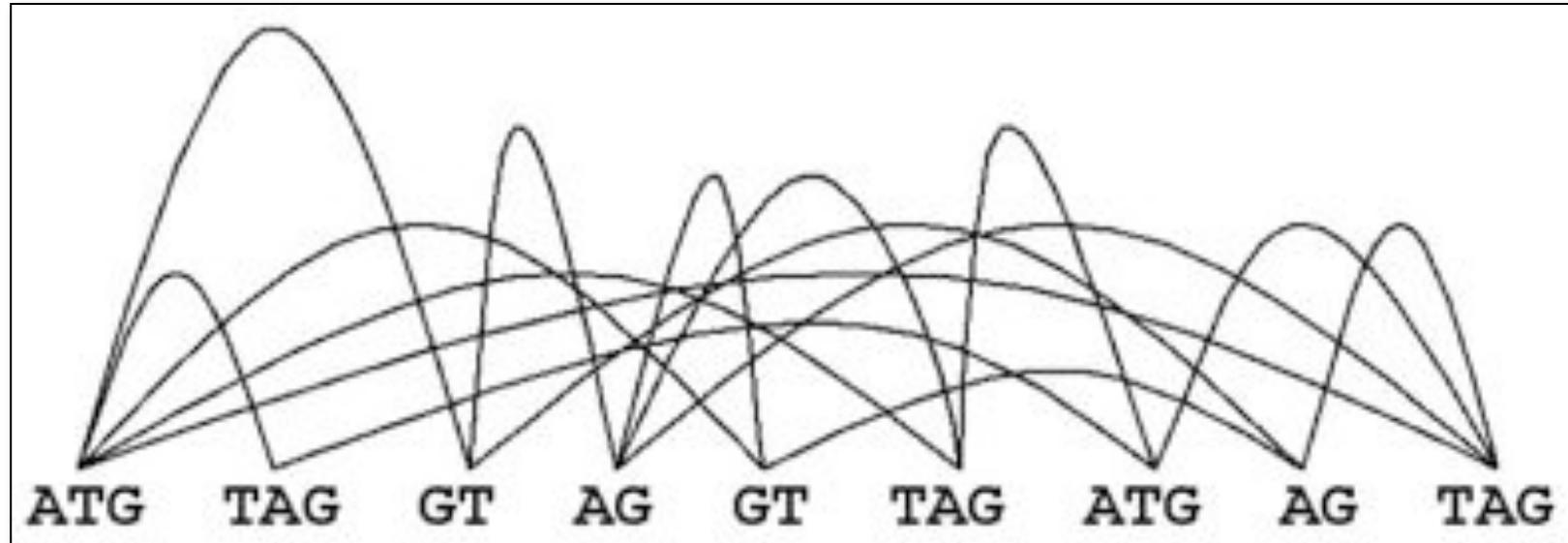
Eukaryotic Gene Syntax



Regions of the gene outside of the CDS are called **UTR**'s (*untranslated regions*), and are mostly ignored by gene finders, though they are important for regulatory functions.

Representing Gene Syntax with ORF Graphs

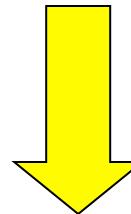
After identifying the most promising (i.e., highest-scoring) signals in an input sequence, we can apply the gene syntax rules to connect these into an *ORF graph*:



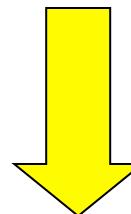
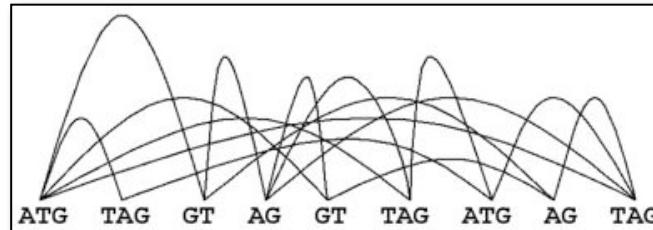
An ORF graph represents all possible *gene parses* (and their scores) for a given set of putative signals. A *path* through the graph represents a single gene parse.

Conceptual Gene-finding Framework

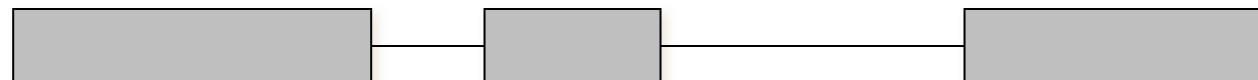
TATTCCGATCGATCGATCTCTCTAGCGTCTACG
CTATCATCGCTCTATTATCGCGCGATCGTCG
ATCGCGCGAGAGTATGCTACGTGATCGAATTG



identify most promising signals, score signals and content regions between them; induce an ORF graph on the signals



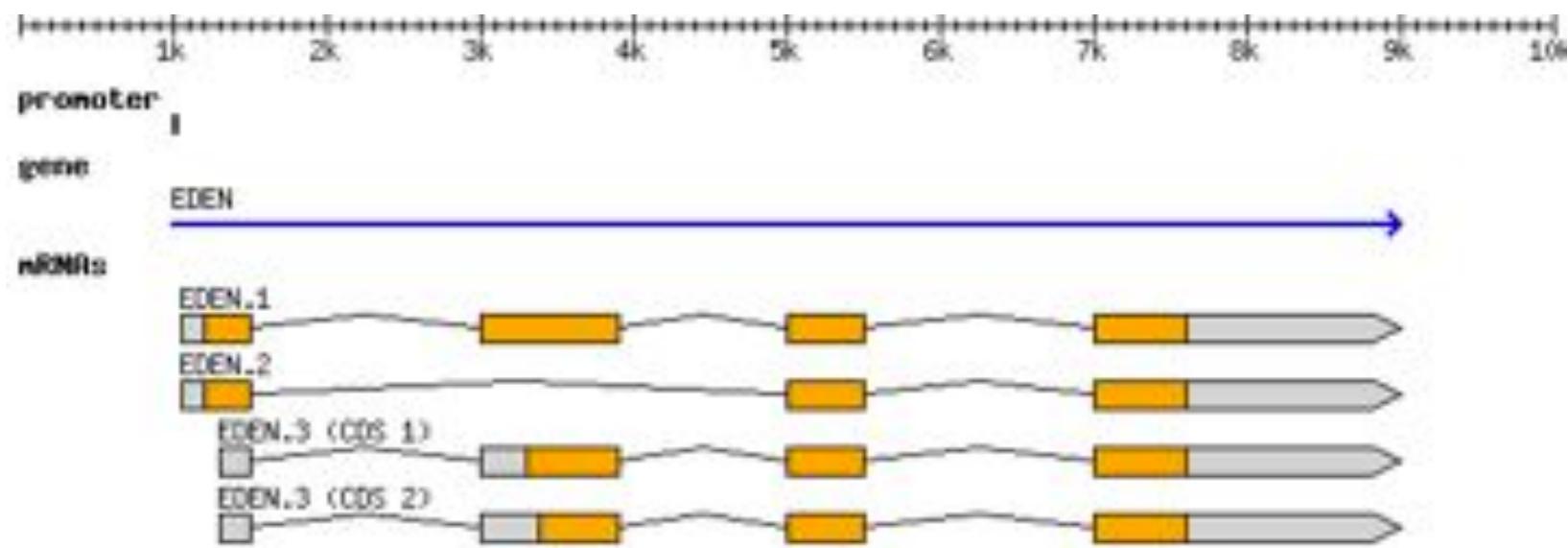
find highest-scoring path through ORF graph;
interpret path as a gene parse = gene structure



Gene Finding Overview

- Prokaryotic gene finding distinguishes real genes and random ORFs
 - Prokaryotic genes have simple structure and are largely homogenous, making it relatively easy to recognize their sequence composition
- Eukaryotic gene finding identifies the genome-wide most probable gene models (set of exons)
 - “Probabilistic Graphical Model” to enforce overall gene structure, separate models to score splicing/transcription signals
 - Accuracy depends to a large extent on the quality of the training data

Gene Models



- “Generic Feature Format” (GFF) records genomic features
 - Coordinates of each exon
 - Coordinates of UTRs
 - Link together exons into transcripts
 - Link together transcripts into gene models

<http://www.sequenceontology.org/gff3.shtml>

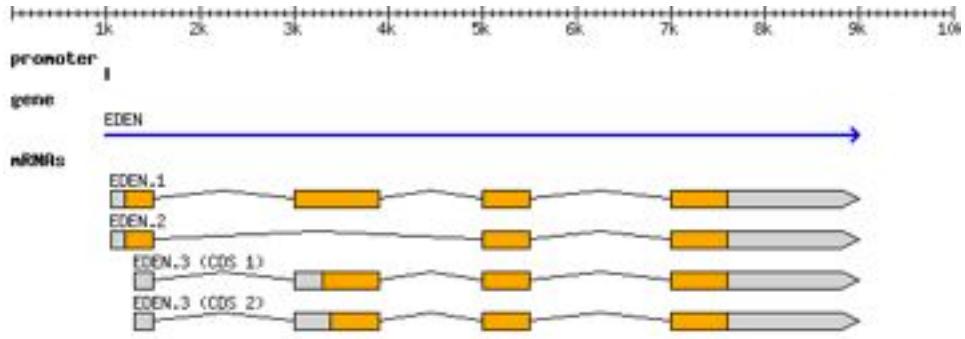
GFF File format

GFF3 files are nine-column, tab-delimited, plain text files

- 1. *seqid*:** The ID of the sequence
- 2. *source*:** Algorithm or database that generated this feature
- 3. *type*:** gene/exon/CDS/etc...
- 4. *start*:** 1-based coordinate
- 5. *end*:** 1-based coordinate
- 6. *score*:** E-values/p-values/index/colors/...
- 7. *strand*:** “+” for positive “-” for minus, “.” not stranded
- 8. *phase*:** For “CDS”, where the feature begins with reference to the reading frame (0,1,2)
- 9. *attributes*:** A list of tag=value features
 - Parent: Indicates the parent of the feature (group exons into transcripts, transcripts into genes, ...)

GFF Example

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

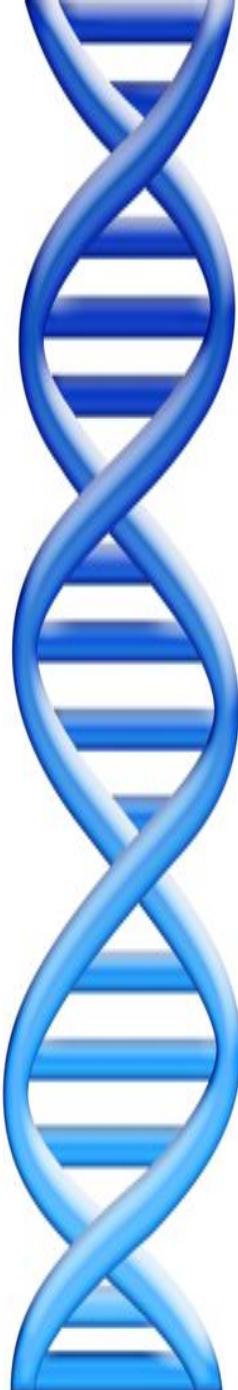


```
##gff-version 3
##sequence-region ctg123 1 1497228
ctg123 . gene      1000 9000 . + . ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000 1012 . + . ID=tfbs00001;Parent=gene00001
ctg123 . mRNA      1050 9000 . + . ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA      1050 9000 . + . ID=mRNA00002;Parent=gene00001;Name=EDEN.2
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
```



THE G-NOME PROJECT

Break

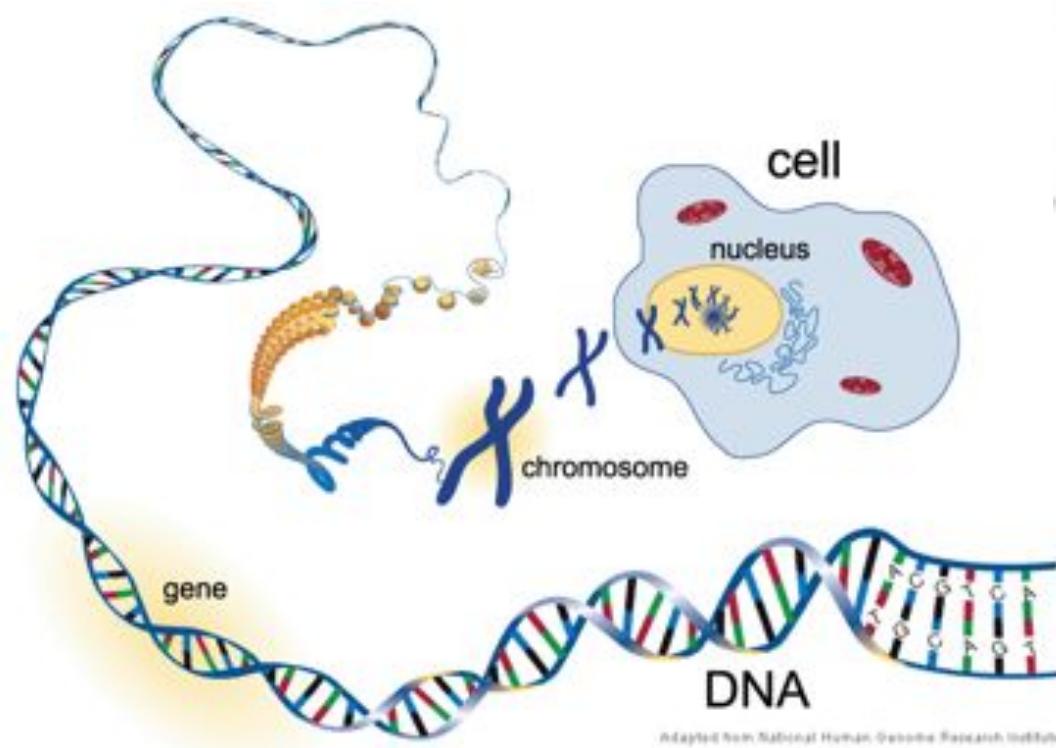


Outline

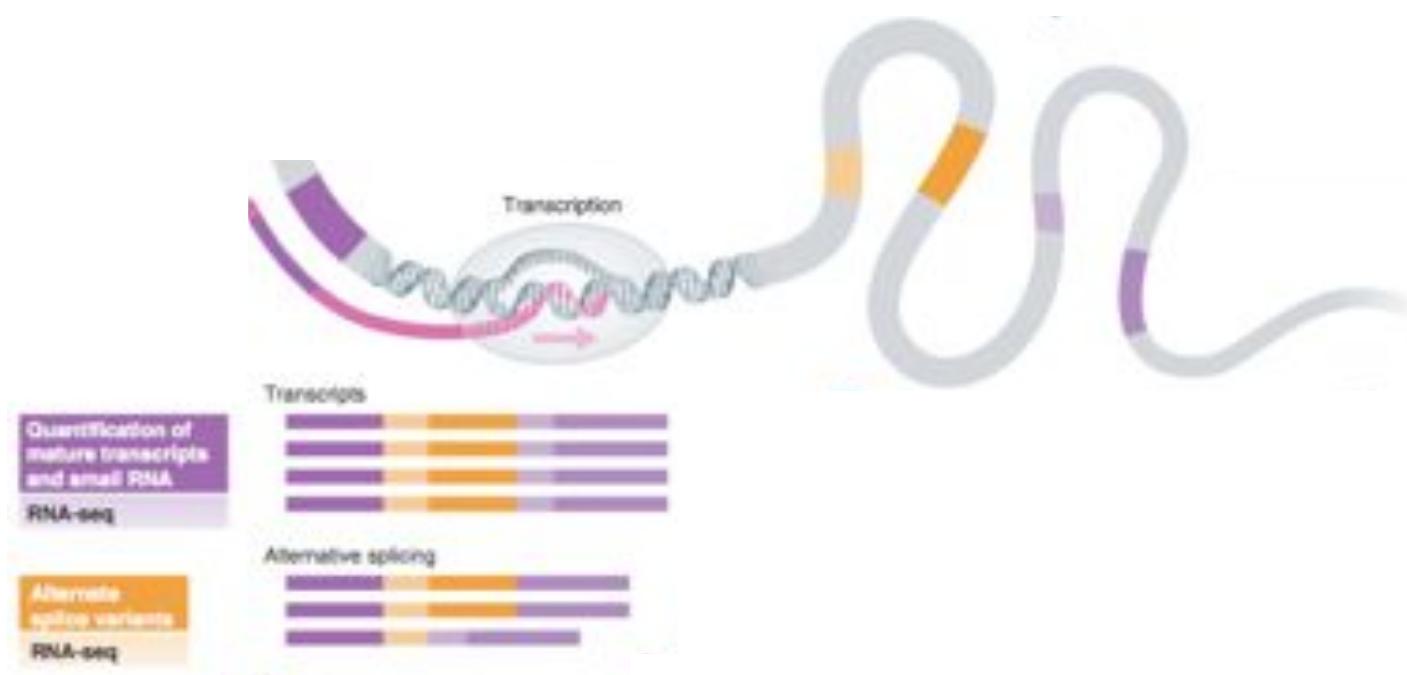
1. Alignment to other genomes
2. Prediction aka “Gene Finding”
3. **Experimental & Functional Assays**
4. Online Resources

Sequencing techniques

Much of the capacity is used to sequence genomes (or exomes) of individuals...



... but biology is much more than just genomes...



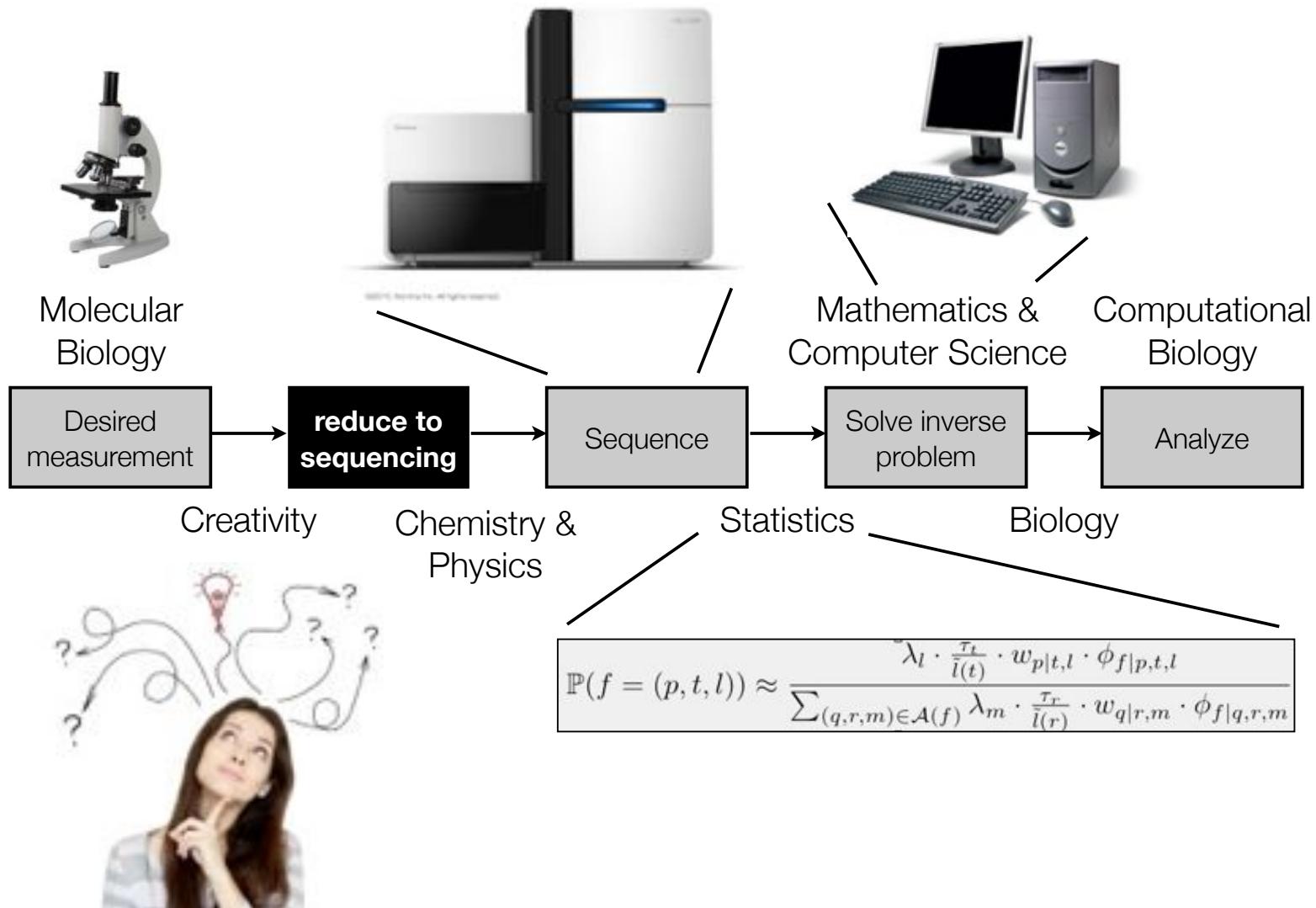
Soon et al., Molecular Systems Biology, 2013

Sequencing Assays

The *Seq List (in chronological order)

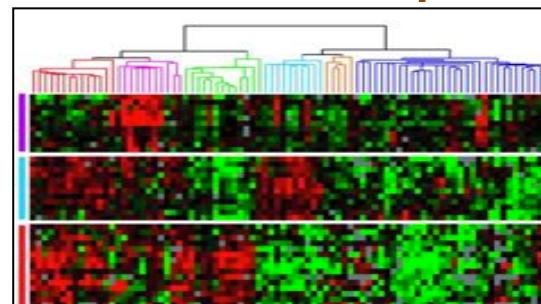
1. Gregory E. Crawford et al., "Genome-wide Mapping of DNase Hypersensitive Sites Using Massively Parallel Signature Sequencing (MPSS)," *Genome Research* 16, no. 1 (January 1, 2006): 123–131, doi:10.1101/gr.4074106.
2. David S. Johnson et al., "Genome-Wide Mapping of in Vivo Protein-DNA Interactions," *Science* 316, no. 5830 (June 8, 2007): 1497–1502, doi:10.1126/science.1141319.
3. Tarjei S. Mikkelsen et al., "Genome-wide Maps of Chromatin State in Pluripotent and Lineage-committed Cells," *Nature* 448, no. 7153 (August 2, 2007): 553–560, doi:10.1038/nature06008.
4. Thomas A. Down et al., "A Bayesian Deconvolution Strategy for Immunoprecipitation-based DNA Methylome Analysis," *Nature Biotechnology* 26, no. 7 (July 2008): 779–785, doi:10.1038/nbt1414.
5. Ali Mortazavi et al., "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq," *Nature Methods* 5, no. 7 (July 2008): 621–628, doi:10.1038/nmeth.1226.
6. Nathan A. Baird et al., "Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers," *PLoS ONE* 3, no. 10 (October 13, 2008): e3376, doi:10.1371/journal.pone.0003376.
7. Leighton J. Core, Joshua J. Waterfall, and John T. Lis, "Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters," *Science* 322, no. 5909 (December 19, 2008): 1845–1848, doi:10.1126/science.1162228.
8. Chao Xie and Martti T. Tammi, "CNV-seq, a New Method to Detect Copy Number Variation Using High-throughput Sequencing," *BMC Bioinformatics* 10, no. 1 (March 6, 2009): 80, doi:10.1186/1471-2105-10-80.
9. Jay R. Hesselberth et al., "Global Mapping of protein-DNA Interactions in Vivo by Digital Genomic Footprinting," *Nature Methods* 6, no. 4 (April 2009): 283–289, doi:10.1038/nmeth.1313.
10. Nicholas T. Ingolia et al., "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling," *Science* 324, no. 5924 (April 10, 2009): 218–223, doi:10.1126/science.1168978.
11. Alayne L. Brunner et al., "Distinct DNA Methylation Patterns Characterize Differentiated Human Embryonic Stem Cells and Developing Human Fetal Liver," *Genome Research* 19, no. 6 (June 1, 2009): 1044–1056, doi:10.1101/gr.088773.108.
12. Mayumi Oda et al., "High-resolution Genome-wide Cytosine Methylation Profiling with Simultaneous Copy Number Analysis and Optimization for Limited Cell Numbers," *Nucleic Acids Research* 37, no. 12 (July 1, 2009): 3829–3839, doi:10.1093/nar/gkp260.
13. Zachary D. Smith et al., "High-throughput Bisulfite Sequencing in Mammalian Genomes," *Methods* 48, no. 3 (July 2009): 226–232, doi:10.1016/j.ymeth.2009.05.003.
14. Andrew M. Smith et al., "Quantitative Phenotyping via Deep Barcode Sequencing," *Genome Research* (July 21, 2009), doi:10.1101/gr.

What is a *Seq assay?

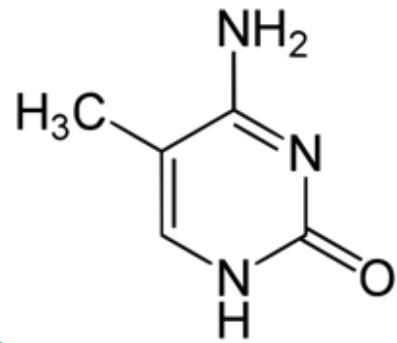


*-seq in 3 short vignettes

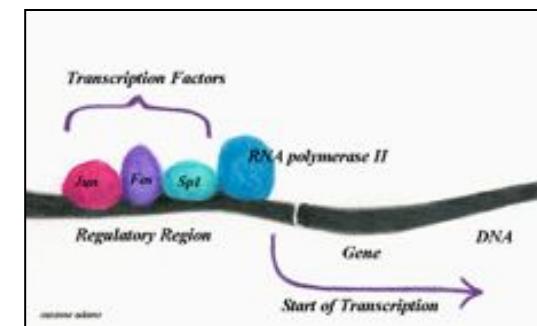
RNA-seq



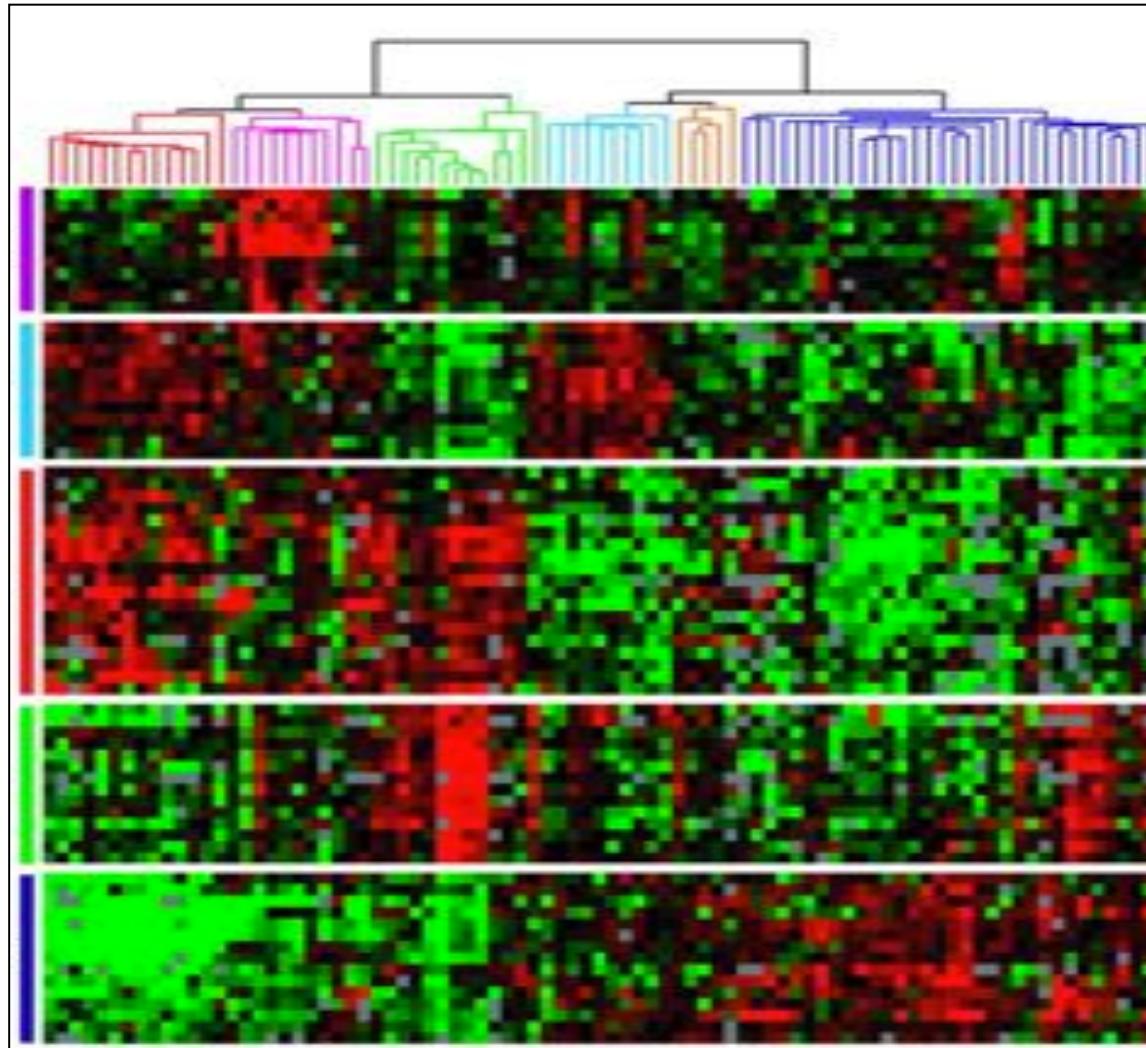
Methyl-seq



ChIP-seq

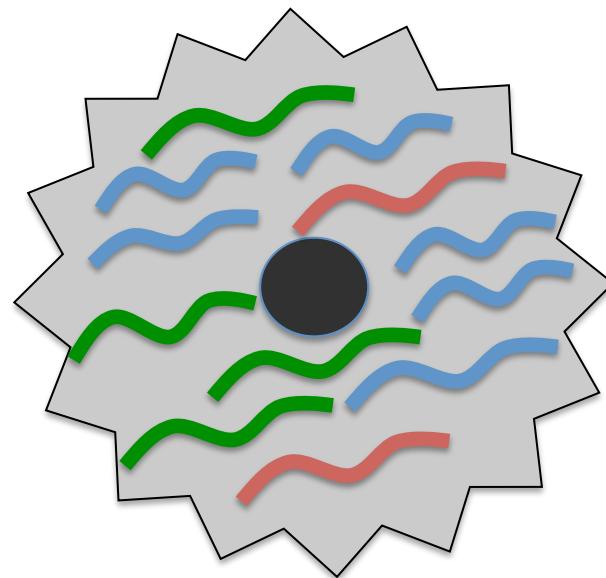


RNA-seq

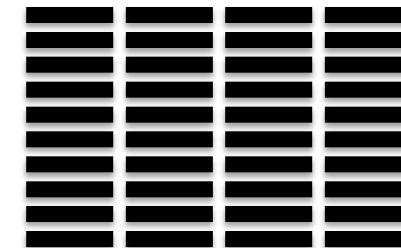


Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.
Sørlie et al (2001) PNAS. 98(19):10869-74.

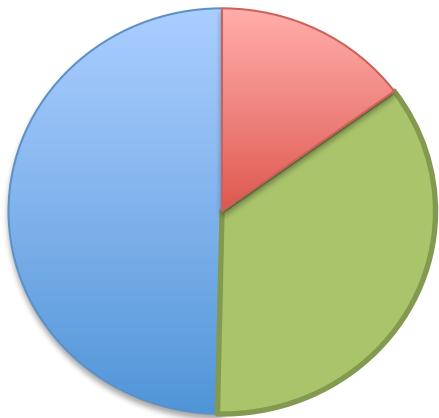
RNA-seq Overview



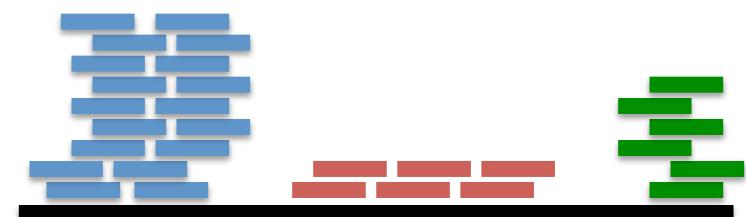
Sequencing



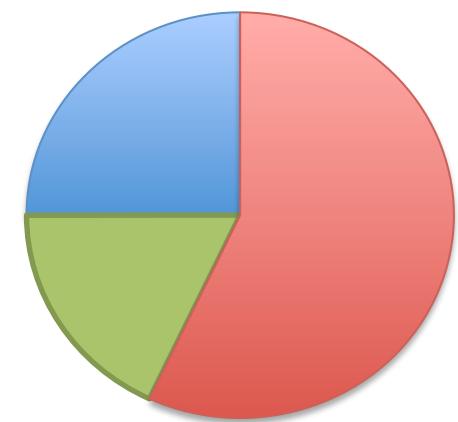
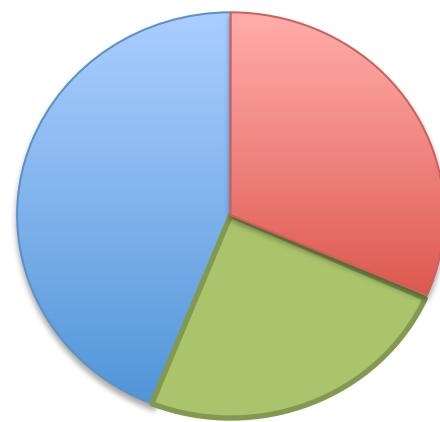
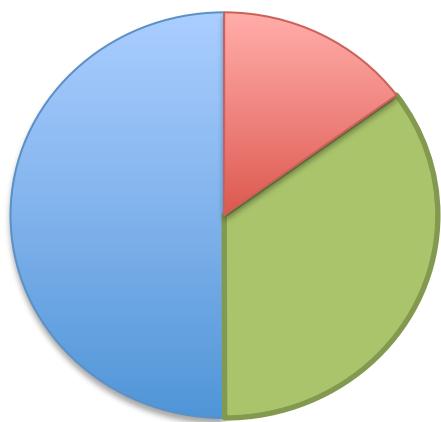
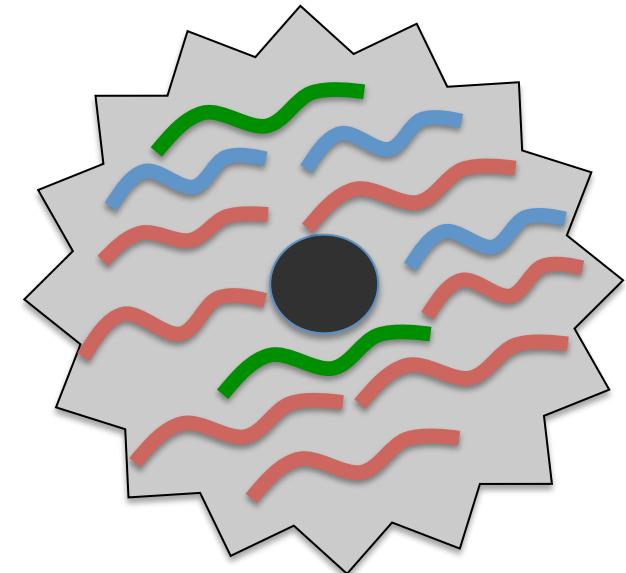
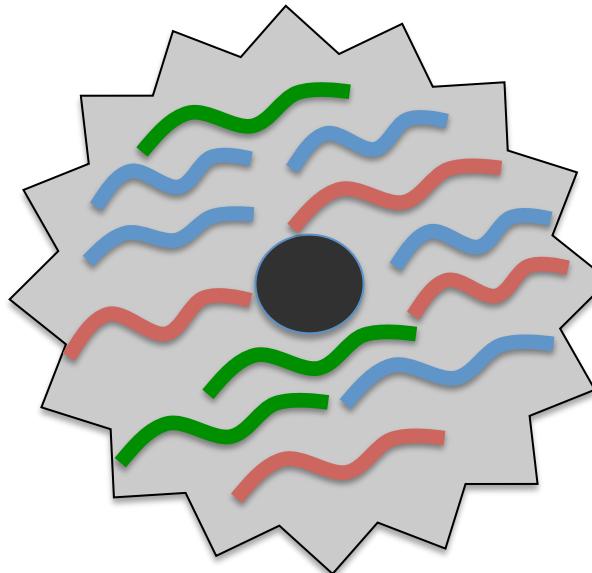
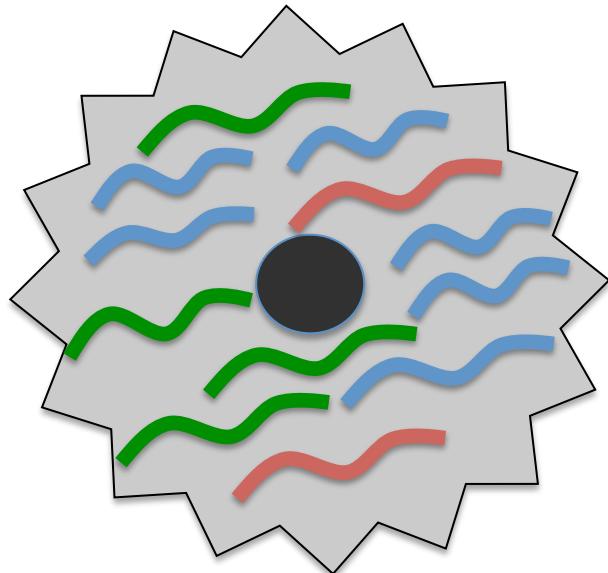
Mapping
& Assembly



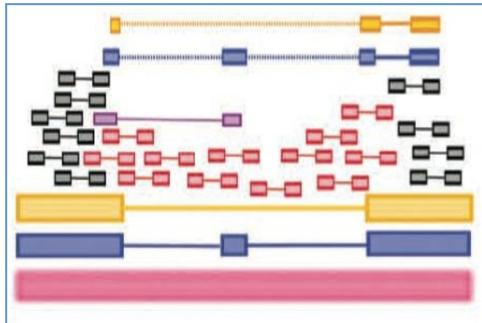
Quantification



RNA-seq Overview



RNA-seq Challenges

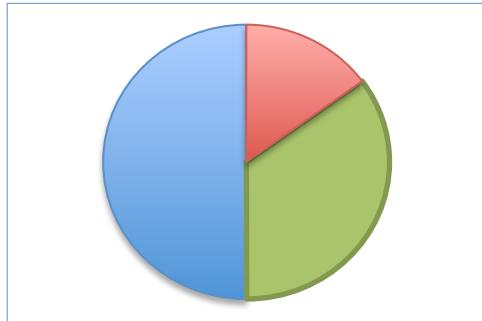


Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

TopHat: discovering spliced junctions with RNA-Seq.

Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111

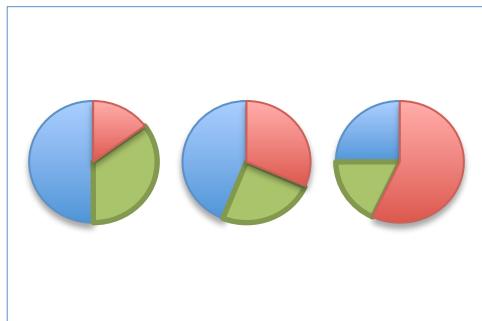


Challenge 2: Read Count != Transcript abundance

Solution: Infer underlying abundances (e.g. FPKM)

Transcript assembly and quantification by RNA-seq

Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



Challenge 3: Transcript abundances are stochastic

Solution: Replicates, replicates, and more replicates

RNA-seq differential expression studies: more sequence or more replication?

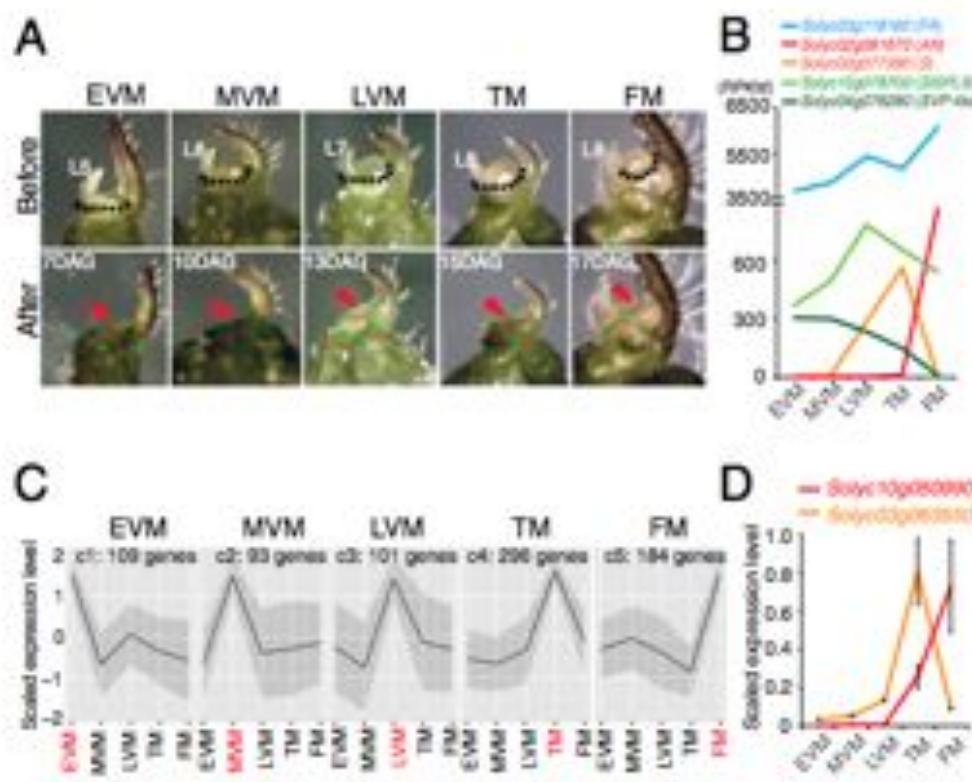
Liu et al (2013) *Bioinformatics*. doi:10.1093/bioinformatics/btt688

Rate of meristem maturation determines inflorescence architecture in tomato

Soon Ju Park¹, Ke Jiang¹, Michael C. Schatz, and Zachary B. Lippman²

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

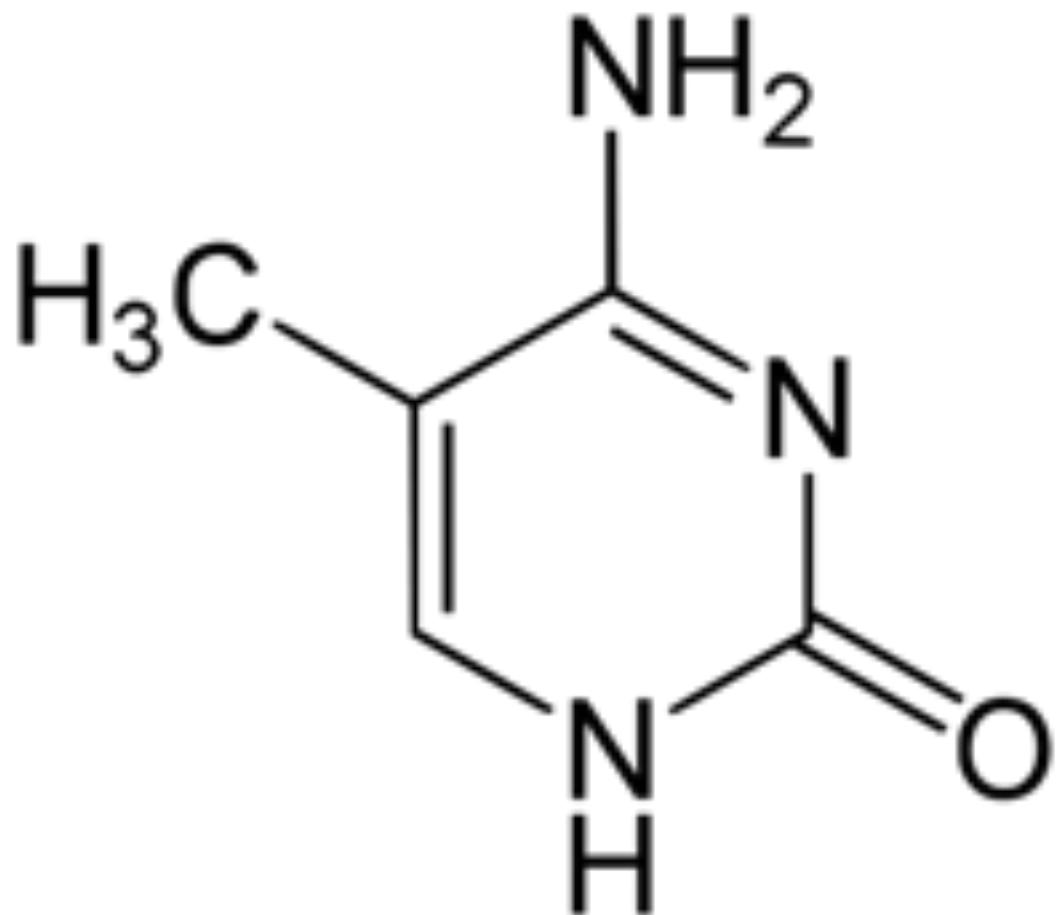
Edited by Maarten Koornneef, Wageningen University and Research Centre, Cologne, Germany, and approved November 28, 2011 (received for review September 12, 2011)



RNA-seq to determine the expression dynamics during development

- Laser microdissection to precisely extract tissue from developing organs
- Use RNA-seq to watch different classes of genes become activated at different stages of development
- When those genes are delayed or interrupted, tomato mutants take on very different branching patterns.

Methyl-seq



Finding the fifth base: Genome-wide sequencing of cytosine methylation
Lister and Ecker (2009) *Genome Research.* 19: 959-966

The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers

Frank Lyko^{1,2*}, Sylvain Foret^{2,3}, Robert Kucharski³, Stephan Wolf⁴, Cassandra Falckenhayn¹, Ryszard Maleszka^{3*}

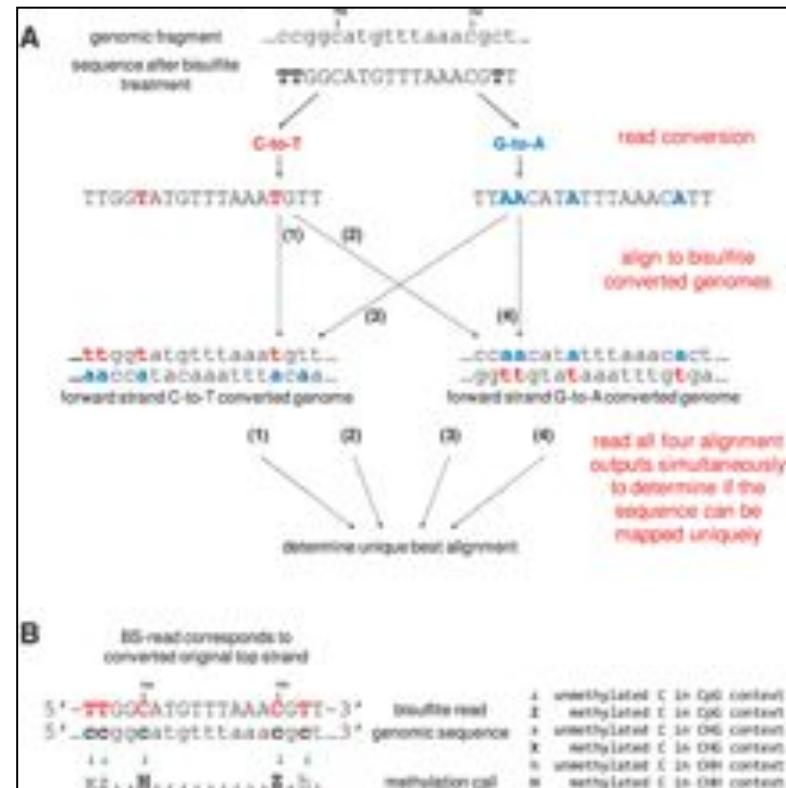
1 Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany, 2 ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Australia, 3 Research School of Biology, the Australian National University, Canberra, Australia, 4 Genomics and Proteomics Core Facility, German Cancer Research Center, Heidelberg, Germany



Bisulfite Conversion

Treating DNA with sodium bisulfite will convert unmethylated C to T

- 5-MethylC will be protected and not change, so can look for differences when mapping
- Requires great care when analyzing reads, since the complementary strand will also be converted (G to A)
- Typically analyzed by mapping to a “reduced alphabet” where we assume all Cs are converted to Ts once on the forward strand and once on the reverse



Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

Bisulfite Conversion

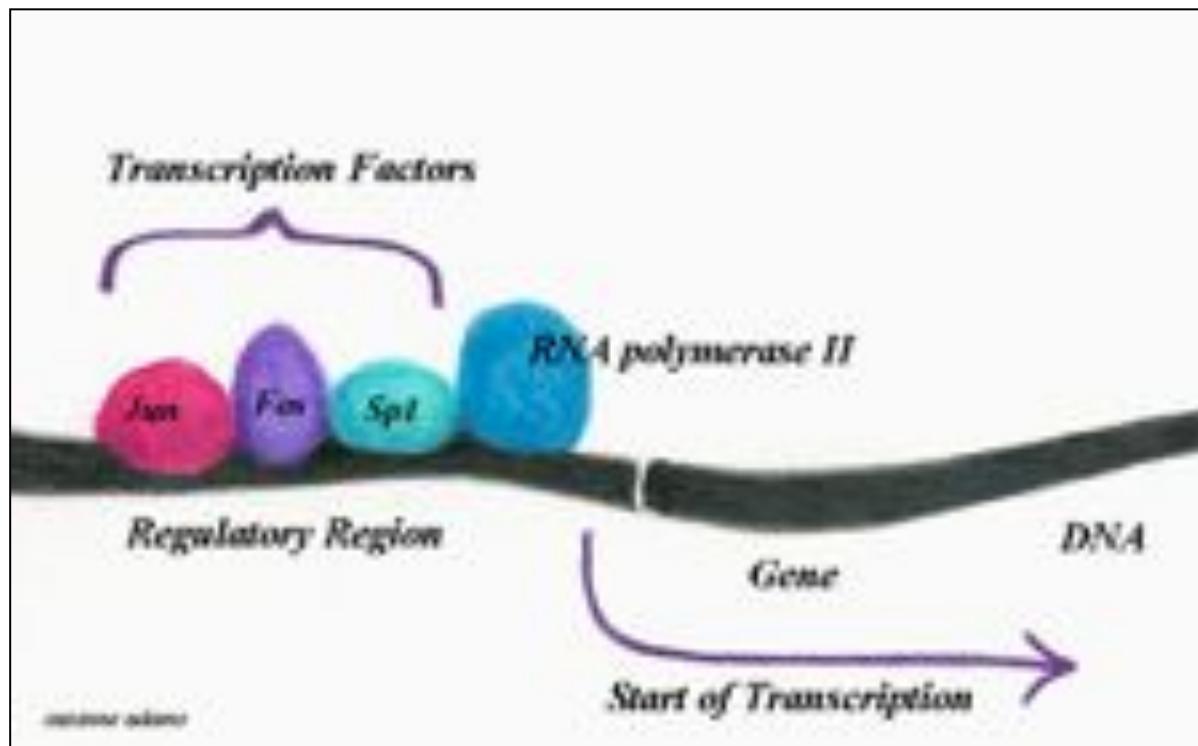
T
W

-
-
-
-
-



Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

ChIP-seq

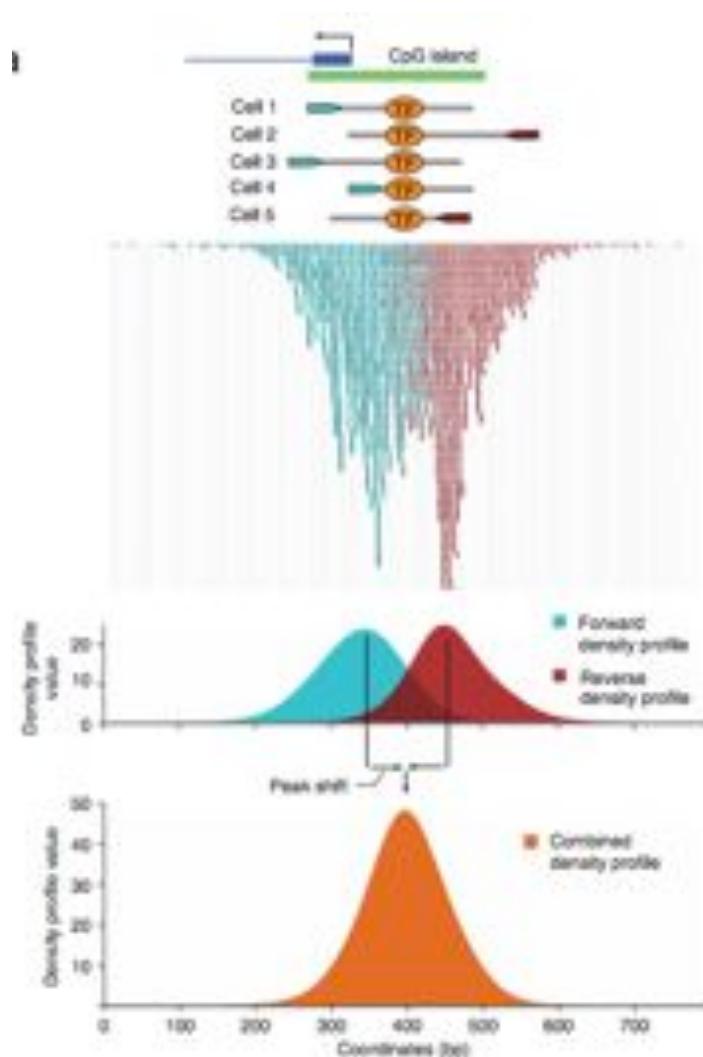
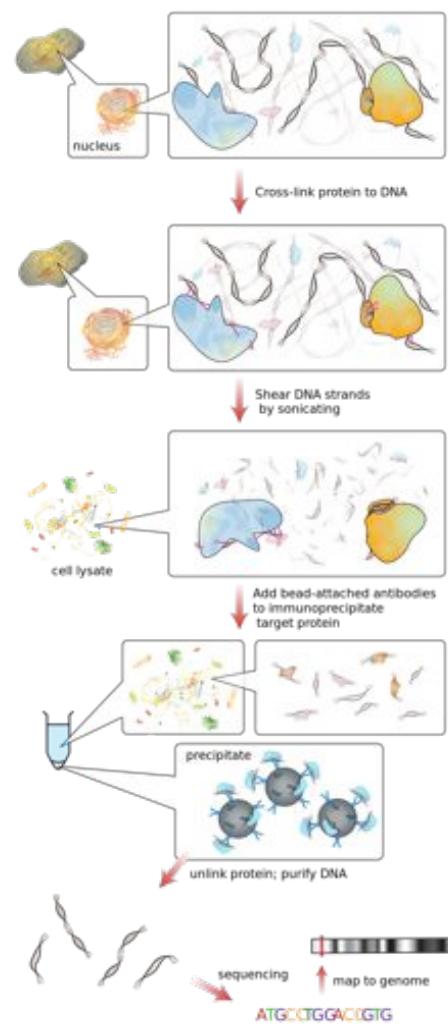


Genome-wide mapping of in vivo protein-DNA interactions.
Johnson et al (2007) Science. 316(5830):1497-502

ChIP-seq

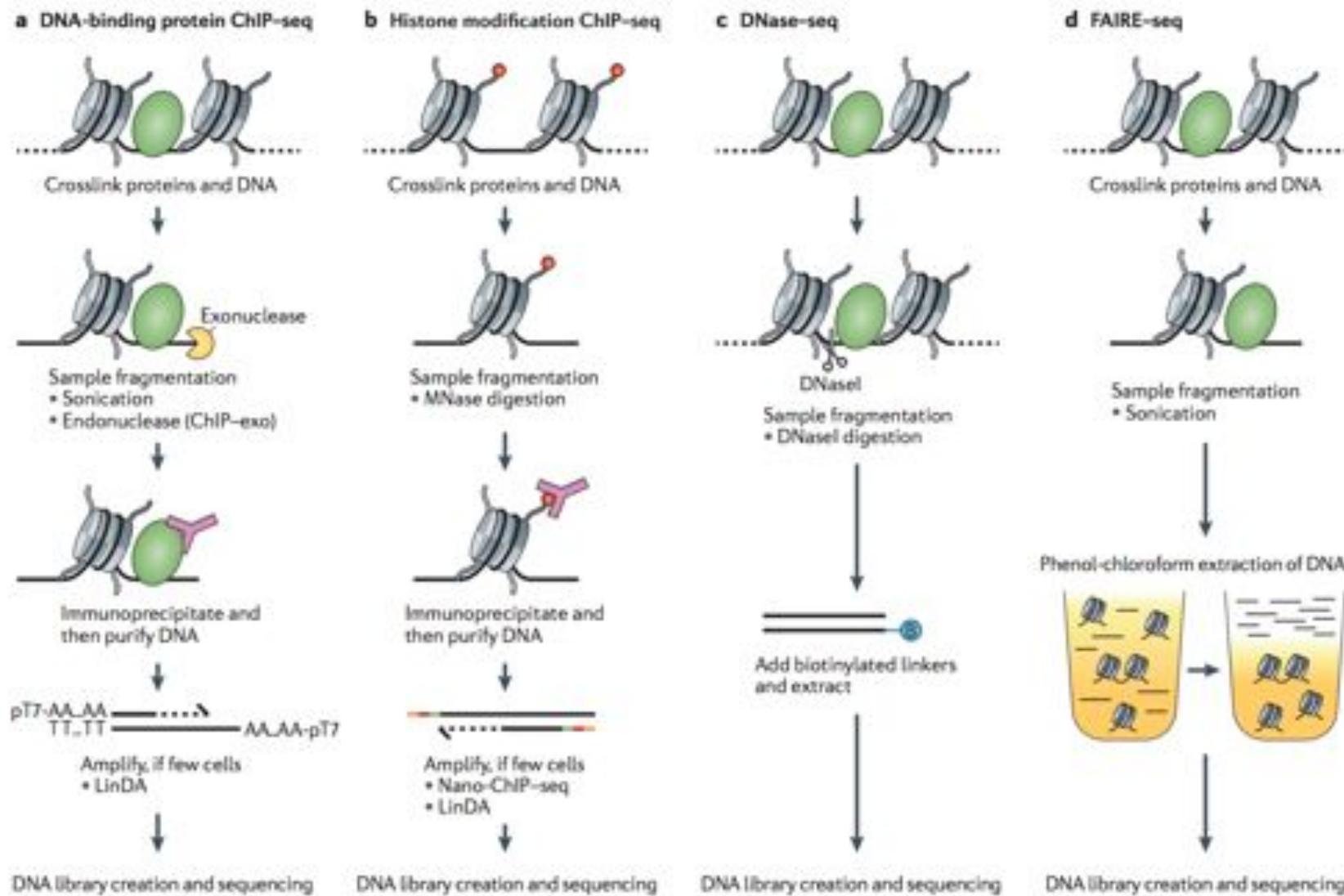
Goals:

- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?



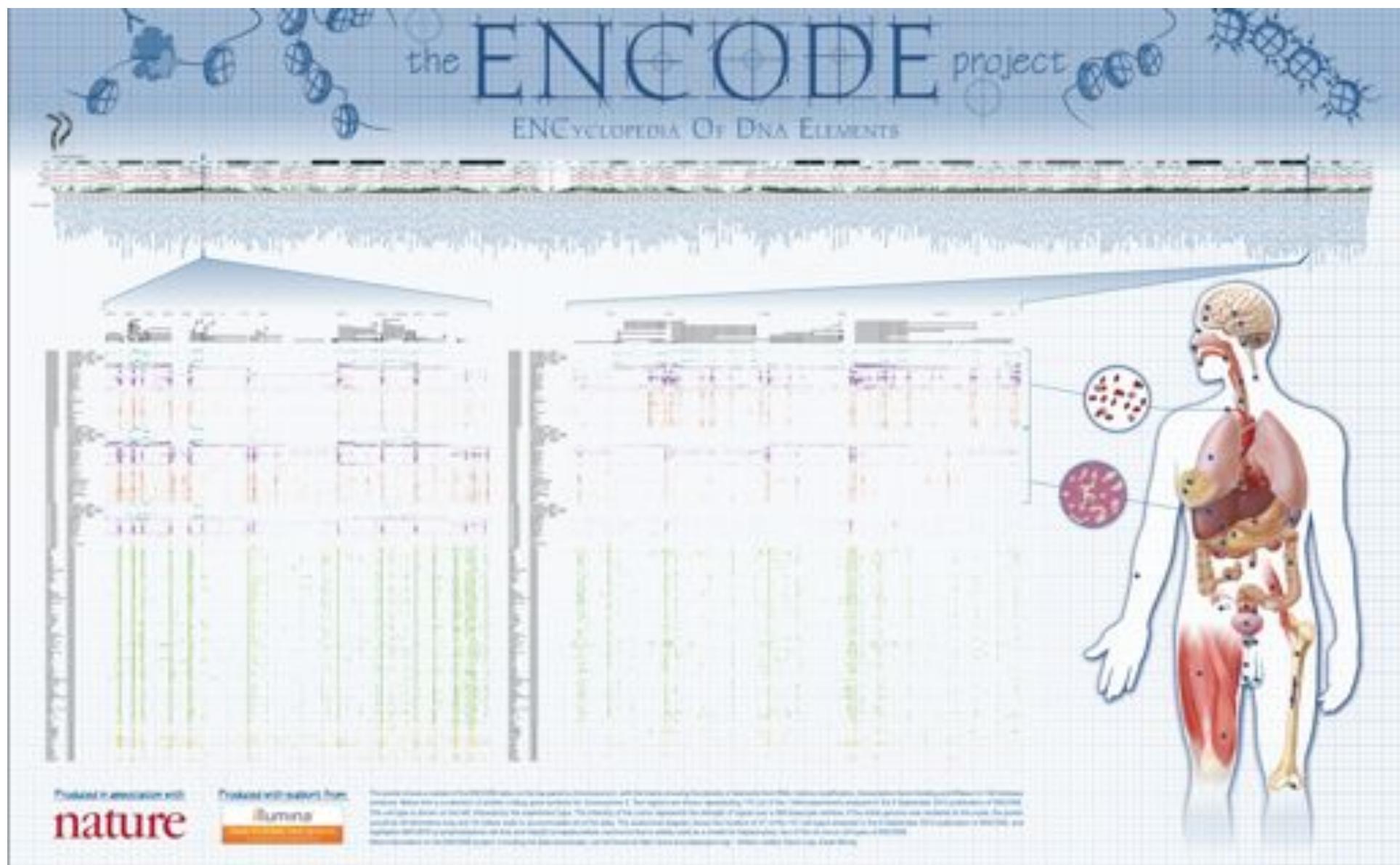
Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data
Valouev et al (2008) *Nature Methods*. 5, 829 - 834

Related Assays



ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions
Furey (2012) *Nature Reviews Genetics.* 13, 840-852

ENCODE Data Sets



1,640 data sets total over 147 different cell types

Summary of ENCODE elements

“Accounting for all these elements, a surprisingly large amount of the human genome, 80.4%, is covered by at least one ENCODE-identified element”

- 62% transcribed
- 56% enriched for histone marks
- 15% open chromatin
- 8% TF binding
- 19% At least one DHS or TF Chip-seq peak
- 4% TF binding site motif
- (Note protein coding genes comprise ~2.94% of the genome)

*“Given that the ENCODE project did not assay all cell types, or all transcription factors, and in particular has sampled few specialized or developmentally restricted cell lineages, **these proportions must be underestimates of the total amount of functional bases.**”*

ChromHMM: Signal Integration

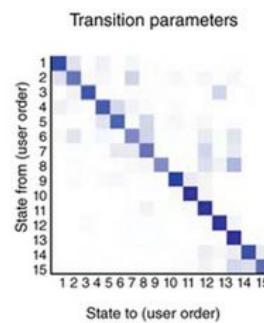
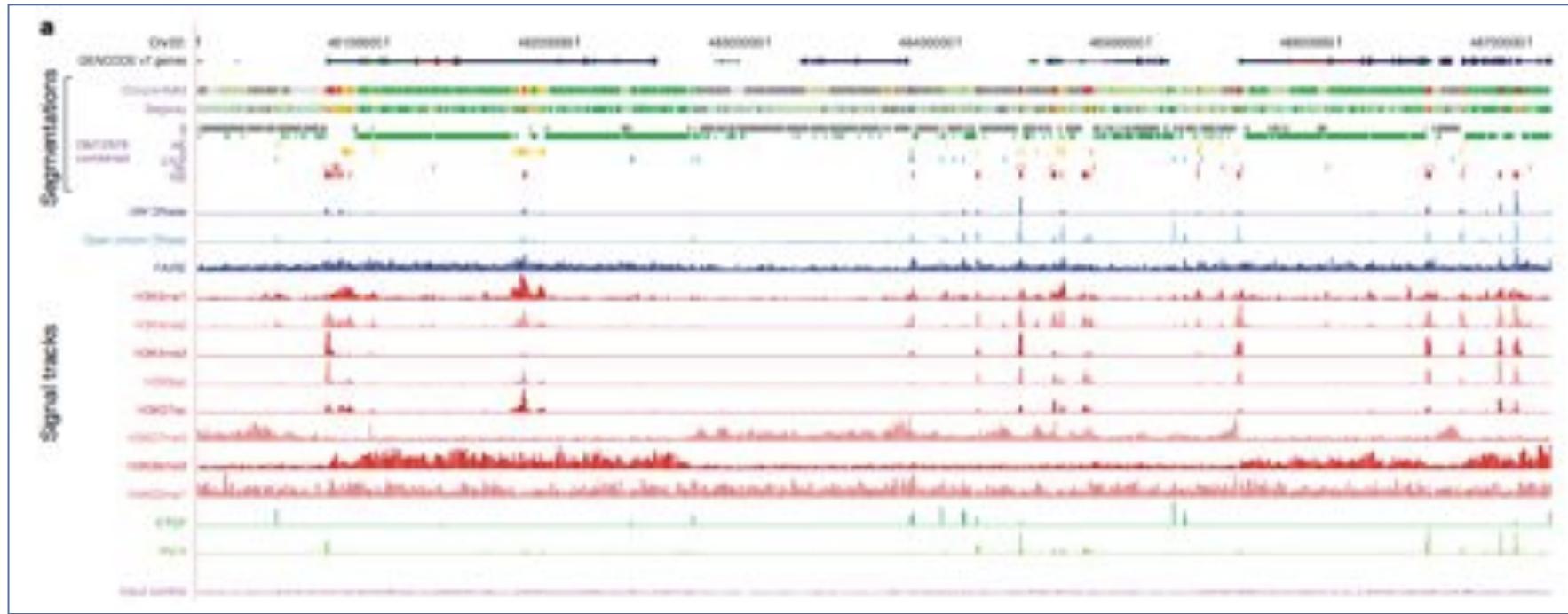
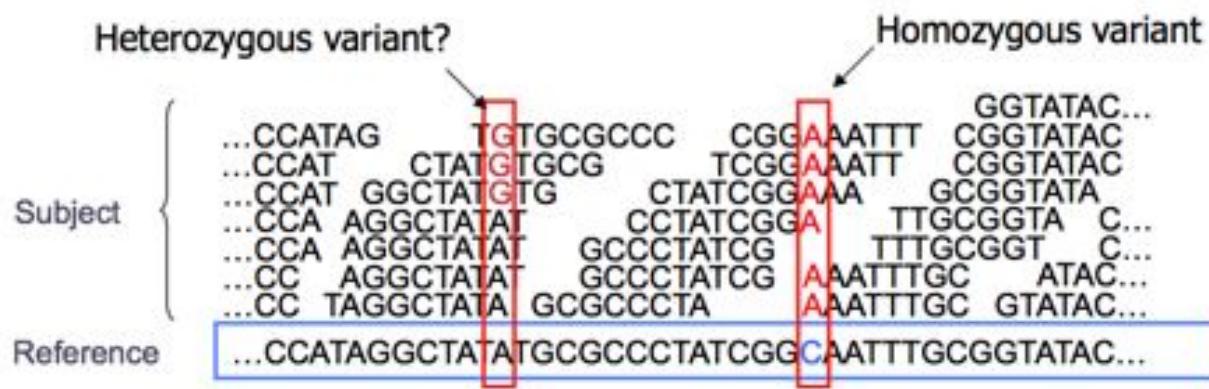


Table 3 | Summary of the combined state types

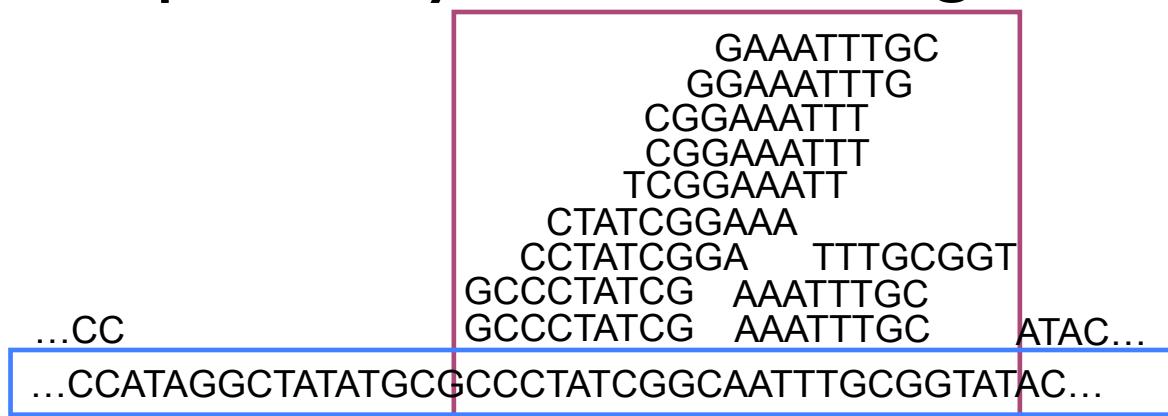
- Summarize the individual assays into 7 functional/regulatory states using an HMM across the genome

Genotyping vs *-seq

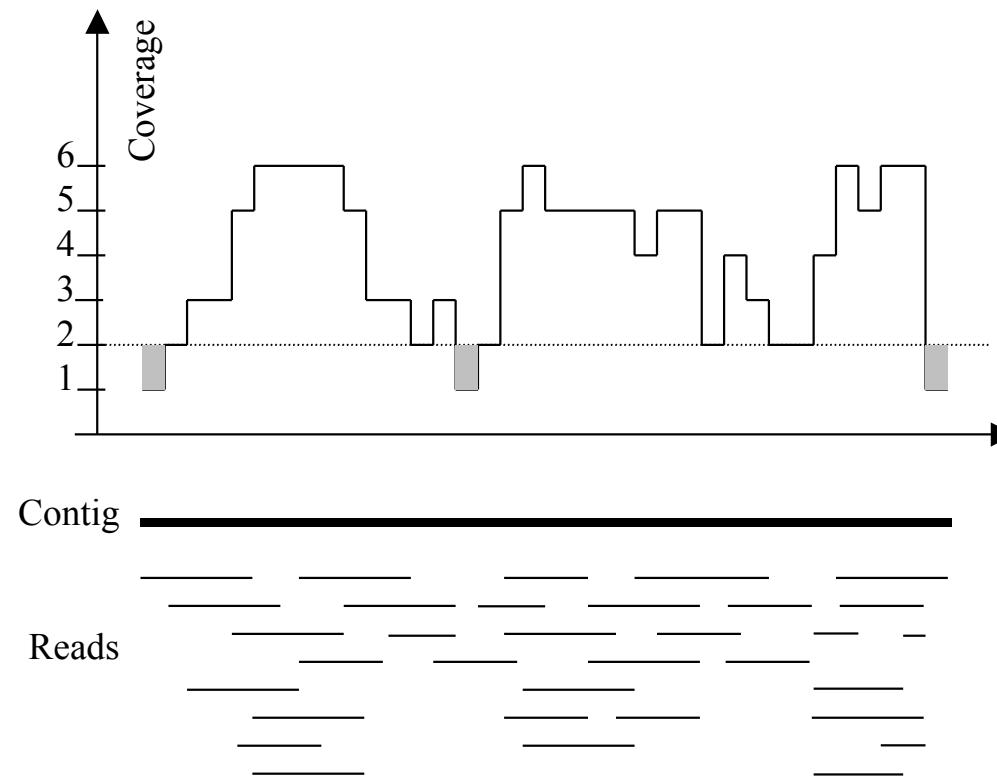
- Genotyping: Identify Variations



- *-seq: Classify & measure significant peaks



WIG/bigWIG Format



- Coverage can change at every single position (3B integers)
- But we often want to summarize to every 100th or every 1000th
- WIG format to the rescue!

WIG/bigWIG Format

Wiggle format is line-oriented, 1st line must be a track definition, followed by declaration lines and data lines

fixedStep is for data with regular intervals between new data values

```
fixedStep chrom=chrN start=position step=stepInterval [span>windowSize]  
dataValues
```

```
fixedStep chrom=chr3 start=400601 step=100  
11  
22  
33
```

variableStep is for data with irregular intervals

```
variableStep chrom=chrN [span>windowSize]  
chromStartA dataValueA
```

```
variableStep chrom=chr2  
300701 12.5  
300702 12.5  
300703 12.5  
300704 12.5  
300705 12.5
```

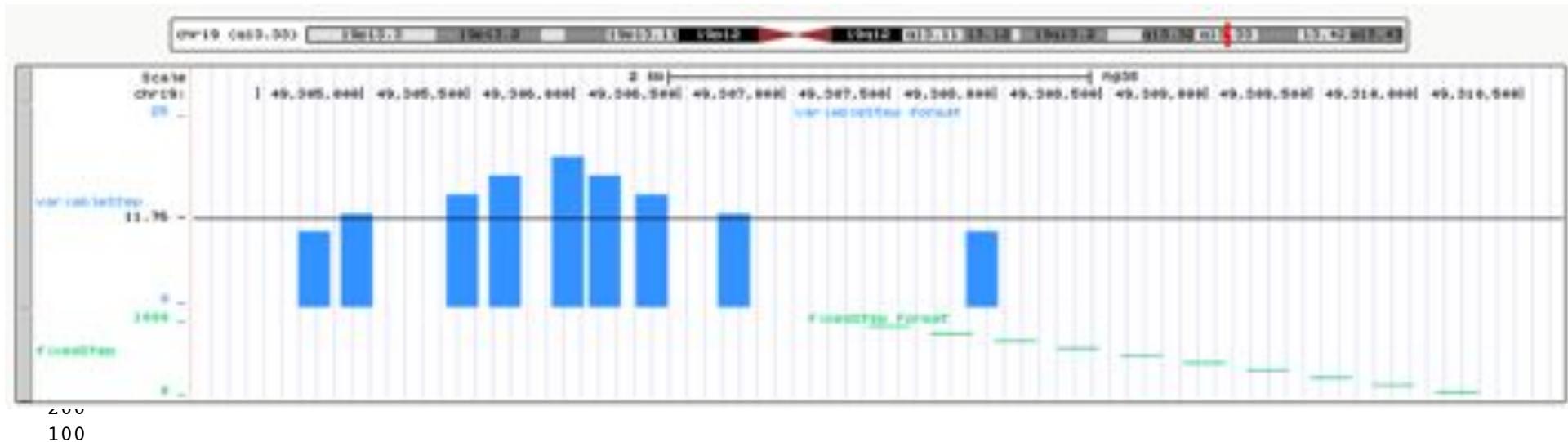
WIG Example

```
browser position chr19:49304200-49310700
browser hide all
#      150 base wide bar graph at arbitrarily spaced positions,
#      threshold line drawn at y=11.76
#      autoScale off viewing range set to [0:25]
#      priority = 10 positions this as the first graph
#      Note, one-relative coordinate system in use for this format
track type=wiggle_0 name="variableStep" description="variableStep format" visibility=full autoScale=off
viewLimits=0.0:25.0 color=50,150,255 yLineMark=11.76 yLineOnOff=on priority=10
variableStep chrom=chr19 span=150
49304701 10.0
49304901 12.5
49305401 15.0
49305601 17.5
49305901 20.0
49306081 17.5
49306301 15.0
49306691 12.5
49307871 10.0

#      200 base wide points graph at every 300 bases, 50 pixel high graph
#      autoScale off and viewing range set to [0:1000]
#      priority = 20 positions this as the second graph
#      Note, one-relative coordinate system in use for this format
track type=wiggle_0 name="fixedStep" description="fixedStep format" visibility=full autoScale=off
viewLimits=0:1000 color=0,200,100 maxHeightPixels=100:50:20 graphType=points priority=20
fixedStep chrom=chr19 start=49307401 step=300 span=200
1000
900
800
700
600
500
400
300
200
100
```

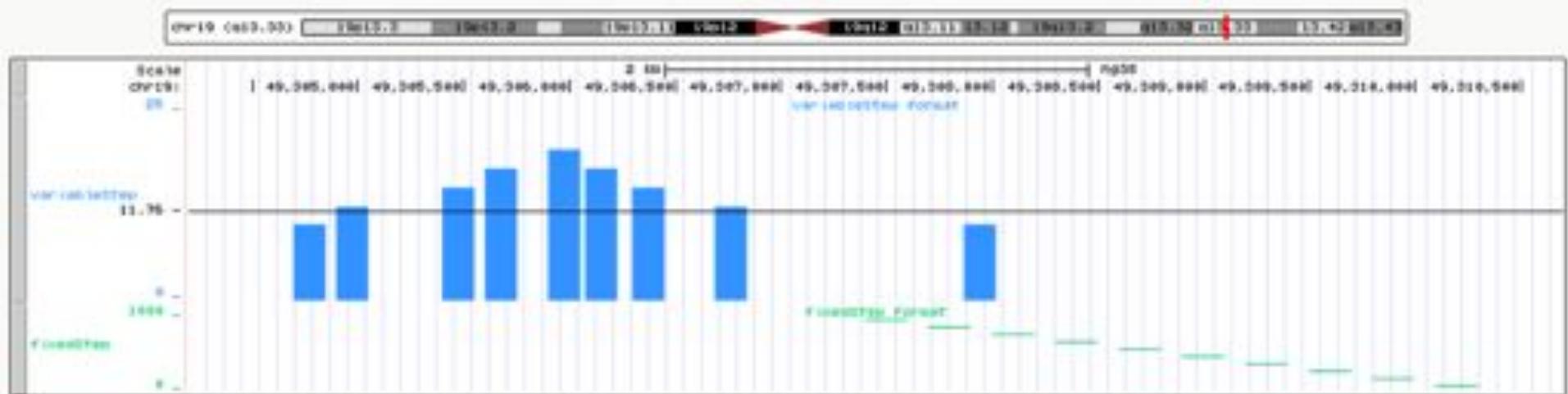
WIG Example

```
browser position chr19:49304200-49310700
browser hide all
#      150 base wide bar graph at arbitrarily spaced positions,
#      threshold line drawn at y=11.76
#      autoScale off viewing range set to [0:25]
#      priority = 10 positions this as the first graph
#      Note, one-relative coordinate system in use for this format
track type=wiggle_0 name="variableStep" description="variableStep format" visibility=full autoScale=off
viewLimits=0.0:25.0 color=50,150,255 yLineMark=11.76 yLineOnOff=on priority=10
variableStep chrom=chr19 span=150
49304701 10.0
49304901 12.5
49305401 15.0
49305601 17.5
49305901 20.0
49306081 17.5
49306301 15.0
49306691 12.5
49307871 10.0
```



WIG Example

```
browser position chr19:49304200-49310700
browser hide all
#      150 base wide bar graph at arbitrarily spaced positions,
#      threshold line drawn at y=11.76
```



```
#      200 base wide points graph at every 300 bases, 50 pixel high graph
#      autoScale off and viewing range set to [0:1000]
#      priority = 20 positions this as the second graph
#      Note, one-relative coordinate system in use for this format
track type=wiggle_0 name="fixedStep" description="fixedStep format" visibility=full autoScale=off
viewLimits=0:1000 color=0,200,100 maxHeightPixels=100:50:20 graphType=points priority=20
fixedStep chrom=chr19 start=49307401 step=300 span=200
1000
900
800
700
600
500
400
300
200
100
```

BED Format

Simple tab-delimited general format for recording “intervals”

Required fields:

1. chrom: The name of the sequence
2. chromStart: The 0-based starting position
3. chromEnd: The 0-based half-open ending position

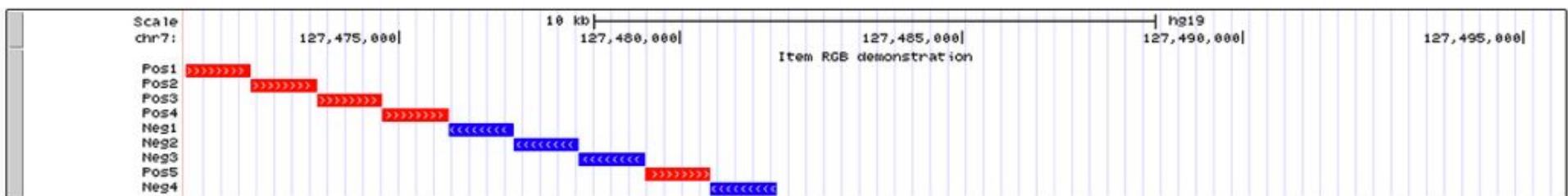
The first 100 bases of a sequence are defined as chromStart=0, chromEnd=100, and span the bases numbered 0-99.

The 9 additional optional BED fields are:

4. name: Defines the name of the BED line
5. score: A score value (typically between 0 and 1000)
6. strand: Defines the strand - either '+' or '-'.
7. thickStart: The starting position at which the feature is drawn thickly
8. thickEnd: The ending position at which the feature is drawn thickly
9. itemRgb: An RGB value of the form R,G,B (e.g. 255,0,0).
10. blockCount: The number of blocks (exons) in the BED line.
11. blockSizes: A comma-separated list of the block sizes.
12. blockStarts: A comma-separated list of block starts.

BED Example

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7 127471196 127472363 Pos1 0      +    127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0      +    127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0      +    127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0      +    127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0      -    127475864 127477031 0,0,255
chr7 127477031 127478198 Neg2 0      -    127477031 127478198 0,0,255
chr7 127478198 127479365 Neg3 0      -    127478198 127479365 0,0,255
chr7 127479365 127480532 Pos5 0      +    127479365 127480532 255,0,0
chr7 127480532 127481699 Neg4 0      -    127480532 127481699 0,0,255
```





Outline

1. Alignment to other genomes
2. Prediction aka “Gene Finding”
3. Experimental & Functional Assays
4. Online Resources

Common Genomics Questions

- What is the closest gene to this ChIP-seq peak?
- Is my latest discovery novel?
- Is there strand bias in my data?
- How many genes does this mutation affect?
- Where did I fail to collect sequence coverage?
- Is this feature significantly correlated with some other feature?

Solution is to integrate (many) online resources
with your own data!

NCBI

<http://www.ncbi.nlm.nih.gov/>

NCBI Resources How To

mchatch@osu.edu My NCBI Sign Out

NCBI National Center for Biotechnology Information

All Databases Search

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Genomic Structural Variation

dbVar archives large scale genomic variation data and associates defined variants with phenotypic information.



1 2 3 4 5 6 7 8

Popular Resources

PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI Announcements

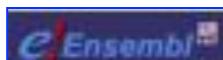
NCBI's July Newsletter is on the Bookshelf

13 Aug 2012

Introduction to the 1000 Genomes Browser, PubMed's Citation Manager and

Ensembl

<http://www.ensembl.org>

Ensembl  BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Miroir

Search: 

BRCA2 or rat 5-62797363-63627969 or coronary heart disease

Browse a Genome
The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

 Human Homo sapiens	 Mouse Mus musculus
 Zebrafish Danio rerio	

 Log in to customize this list.
[All genomes](#)
[Select a species...>](#)
[View full list of all Ensembl species](#)
Other species are available in [Ensembl.Pan](#) and [Ensembl.Genomes](#)

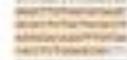
Ensembl supports data from external projects through [DataTables](#) 

ENCODE data in Ensembl 

Variant Effect Predictor 

Gene expression in different tissues 

Find SNPs and other variants for my gene 

Retrieve gene sequence 

Compare genes across species 

Use my own data in Ensembl 

Learn about a disease or phenotype 

Did you know...? 
Last... by our [SocialMedia](#)

 **What's New in Release 77 (October 2014)**

- [GENCODE 21 on human GRCh38](#)
- [First Ensembl-Havana cat merge](#)
- [New species: Vervet African green monkey](#)
- [Transcript supporting levels for human](#)

[Full details](#) | [All web updates](#) | [By release](#) | [More news on our blog](#)

 **Latest blog posts**

- 12 Nov 2014: [ensembl67.ensembl.org brief downtime - 17th Nov 2014](#)
- 31 Oct 2014: [Catra extra extra: Ensembl wodaboo at the Flat Genetics and monoids conference!](#)
- 29 Oct 2014: [What's coming in Ensembl release 78](#)

[Go to Ensembl blog...](#)

Tweets 

 **Ensembl** [@ensembl](#)
SyntenyMapper analysis of micro-rearrangements, including #ENSEMBLcompars #OxfordEnsembl @PLOSOne [buff.ly/1pD3o](#)
[See Summary](#)

 **Paul Flicek** [@pflcok](#)
Our own Bronwen Allen from @ensembl talking about the cat genome on the @ NakedScientists [theresatthesciences.com/n/174/c/podcast/...](#)
[#Reposted by Ensembl](#)
[Expand](#)

 **Ensembl** [@ensembl](#)
Hear Ensembl's @BronwenAllen on the @ NakedScientists talking about the effects of domestication on the cat genome [buff.ly/1yF03](#)
[Tweet to Bronwen](#)

sanger Ensembl is a joint project between EMBL-EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.
Ensembl receives major funding from the Wellcome Trust. Our [acknowledgements](#) page includes a list of additional current and previous funding bodies. [How to cite Ensembl](#) in your own publications.

EMBL-EBI 

Ensembl release 77 - October 2014 © EBI 2014 - 128
Permanent link - <http://www.ensembl.org>

   [About Ensembl](#) | [Privacy Policy](#) | [Contact Us](#)

Biomart

<http://www.biomart.org>

The screenshot shows the Biomart website homepage. On the left is a vertical navigation bar with links: HOME, TOOLS, COMMUNITY, PUBLICATIONS, NEWS, CREDITS, DOCUMENTATION, VERSION 0.7, and CONTACT. Below this is a section for 'ENTERPRISE' with a 'READ MORE' link. The main content area features a 'BioMart' header with a sub-header: 'is a community-driven project to provide unified access to distributed research data to facilitate the scientific discovery process.' Below this is a paragraph about the project's mission and open source philosophy. To the right is a 'JOIN OUR COMMUNITY' section with three bullet points: 'Set up your own data source with a click of a button', 'Expose your data to a world wide scientific community through BioMart Portal', and 'Federate your local data with data from other community members'. At the bottom of this section is a 'BROWSE 2000+ DATA SOURCES' link. A large yellow callout box contains text about the number of servers and data sources, with a 'READ MORE' link. To the right is a world map showing the distribution of data sources across continents.

bio.mart

[30th October 2014] BioMart Portals (0.7 and 0.8) updates; ensembl; ensembl genomes and uniprot; ... ➔

HOME

TOOLS

COMMUNITY

PUBLICATIONS

NEWS

CREDITS

DOCUMENTATION

VERSION 0.7

CONTACT

[READ MORE](#)

BioMart

is a community-driven project to provide unified access to distributed research data to facilitate the scientific discovery process.

The BioMart project provides free software and data services to the international scientific community in order to foster scientific collaboration and facilitate the scientific discovery process. The project adheres to the open source philosophy that promotes collaboration and code reuse.

JOIN OUR COMMUNITY

- Set up your own data source with a click of a button
- Expose your data to a world wide scientific community through BioMart Portal
- Federate your local data with data from other community members

[BROWSE 2000+ DATA SOURCES](#) ➔ [ENTERPRISE](#)

A large number of servers that provide access to a wide range of research data have been set up by the BioMart community. Using BioMart's unique data federation technology, a Central Portal was established to provide a convenient single point of access to all of these data, which is distributed worldwide.

[READ MORE](#)

40 DATABASES, 4 CONTINENTS AND GROWING

A world map with a grid overlay. Numerous small yellow dots are scattered across the map, primarily concentrated in North America, Europe, and Asia, representing the locations of data sources. A larger, more prominent yellow dot is located over the United States. A legend in the top right corner indicates that the size of the dots corresponds to the number of data sources in that region, with a scale from 1 to 20.

20

UCSC Genome Browser

<http://genome.ucsc.edu/>

move <<< << < > >> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr21:33,031,597-33,041,570 9,974 bp. enter position, gene symbol or search terms go

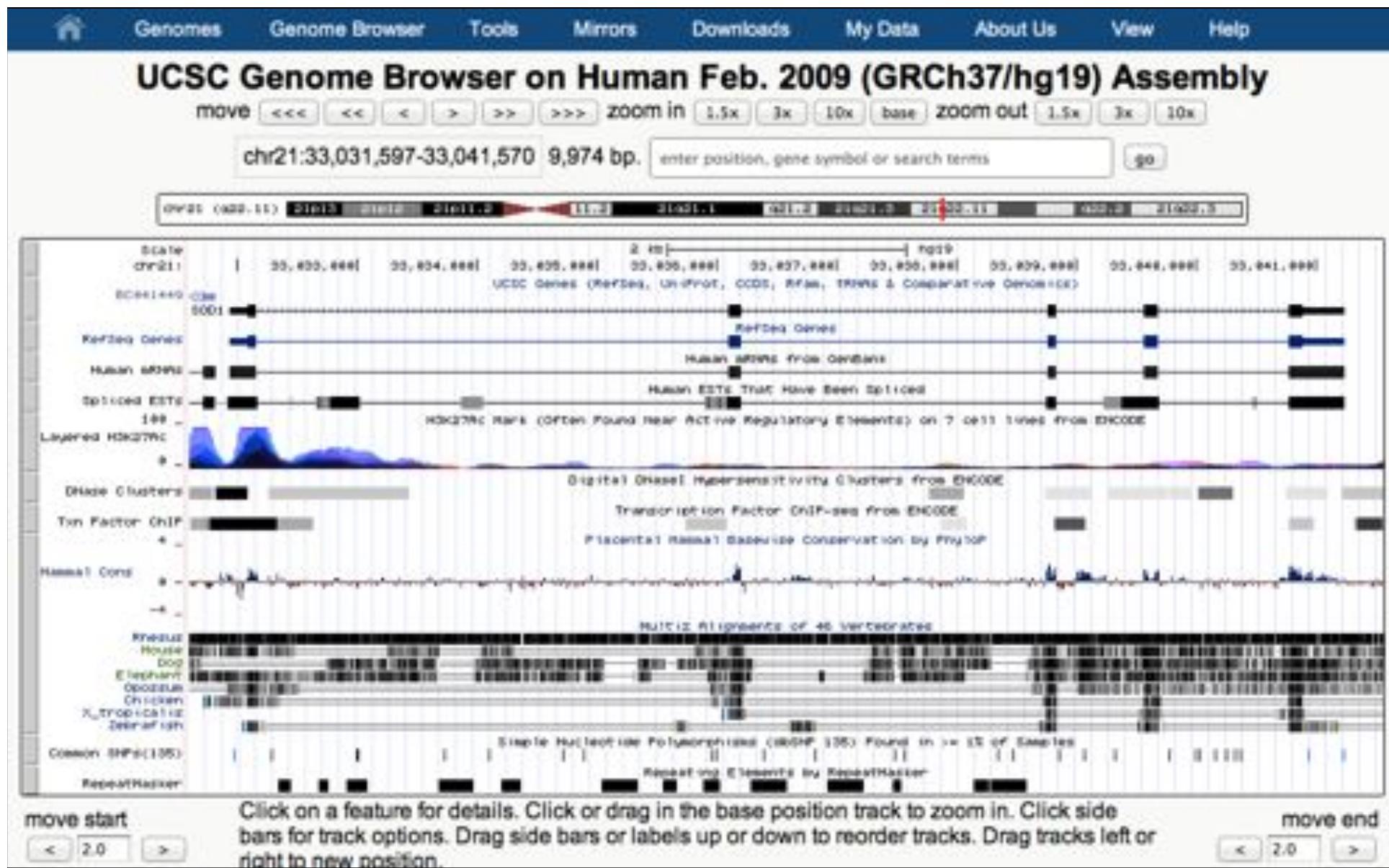
chr21 (GRCh37/hg19) 21612 21613 21614 21615 21616 21617 21618 21619 21620 21621 21622 21623

UCSC Genome Browser (RefSeq, UniProt, CCDS, Rfam, TRNAs & Comparative Genomics)

5001 RefSeq Genes Human mRNA Human ESTs Human ESTs That Have Been Spliced Human ChIP-seq ChIP-seq (Often Found Near Active Regulatory Elements) on 7 Cell Lines From ENCODE Layered H3K27me3 Digital ChIP-seq Hyperacetylated Clusters from ENCODE Transcription Factor ChIP-seq from ENCODE PIGMENTATION MARKER Sequence Conservation by PhyloP Human Conservation Human Conservation Rat Conservation MUSCLE Alignments of 46 Vertebrates Common SNPs (105) RepeatMasker

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

move start < 2.0 > move end < 2.0 >



UCSC Genome Browser / Table Browser

<http://genome.ucsc.edu/cgi-bin/hgTables?command=start>

The screenshot shows a web browser window titled "Table Browser". The address bar displays the URL: "http://genome.ucsc.edu/cgi-bin/hgTables?command=start". The main content area is titled "Table Browser" and contains several input fields and buttons. At the top left, there is a navigation menu with links to "Home", "Genomes", "Genome Browser", "Stat", "Tables", "Gene Sorter", "PCR", "Sessions", "FAQ", and "Help". Below the menu, a sub-menu titled "Table Browser" is shown. The main body of the page contains the following controls:

- clade:** dropdown menu set to "Mammal" (highlighted in blue), with options "genomes", "human", and "assembly".
- assembly:** dropdown menu set to "Feb 2008 UCSC Genome" (highlighted in blue).
- group:** dropdown menu set to "Comparative Genomics" (highlighted in blue), with options "Annotation", "All tracks", and "Custom tracks".
- track:** dropdown menu set to "Annotation" (highlighted in blue), with options "All tracks", "Annotation", "Custom tracks", and "Regions".
- table:** dropdown menu set to "Primate Core (phylogenetic)" (highlighted in blue), with options "Describe table schema" and "Regions".
- region:** dropdown menu set to "position chr1:33011821-33011844" (highlighted in blue), with options "Position" and "Regions".
- filters:** dropdown menu set to "None" (highlighted in blue), with options "None", "Intersection", "Correlation", and "Output format".
- output format:** dropdown menu set to "Send output to Galaxy" (highlighted in blue), with options "Send output to Galaxy", "Send output to GREAT", and "Output file".
- output file:** text input field containing "Leave blank to keep output in browser".
- file type reference:** radio buttons for "plain text" (selected) and "gzip compressed".

Below these controls, there is a note: "Note: to review more than 100,000 lines, change the filter setting (above). The entire data set may be available for download as a very large file that contains the original data values (not compressed into the wiggle format) -- see the Download page." It also includes two small buttons: "get output" and "summary statistics".

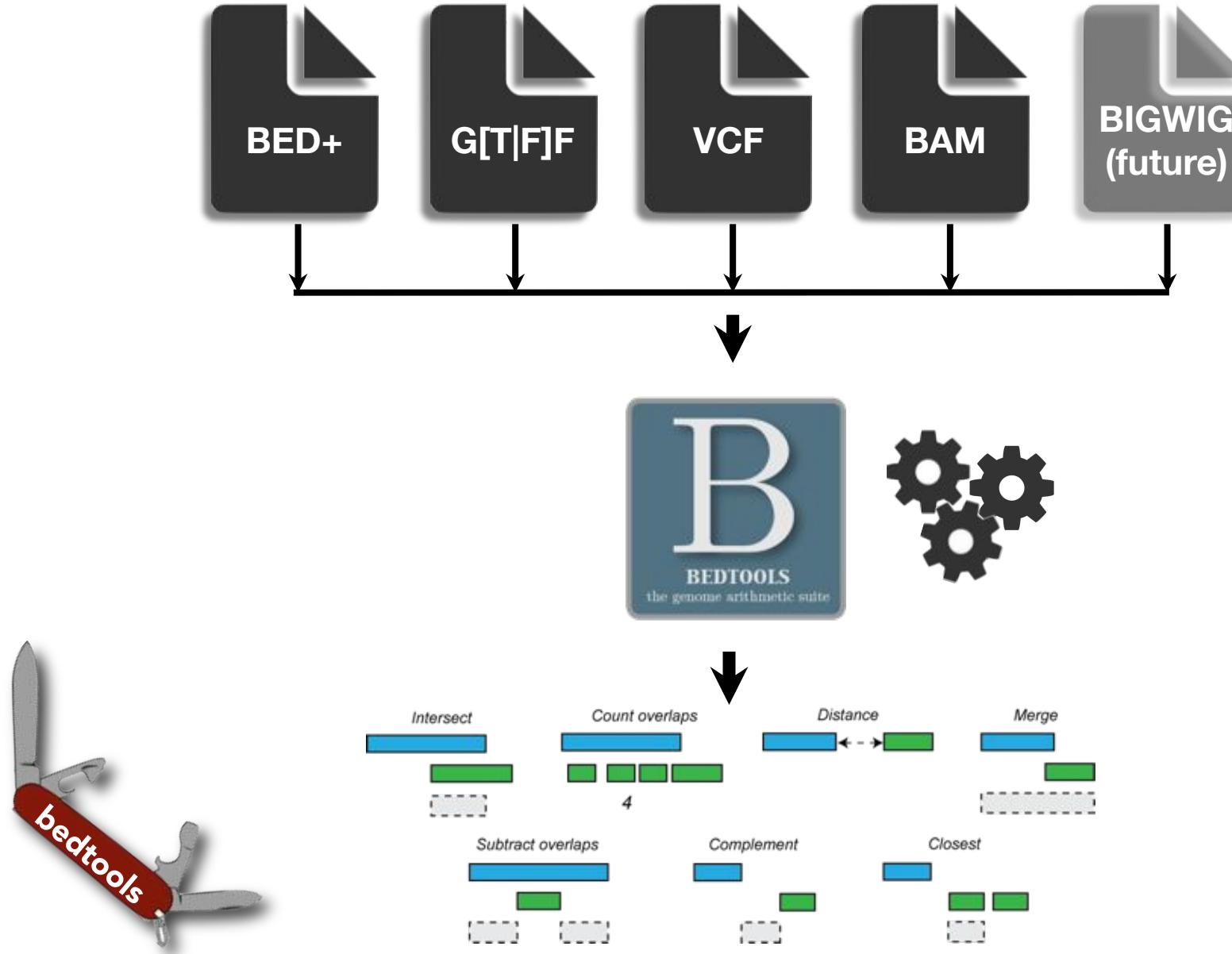
At the bottom of the main content area, there is a link: "To reset all user cart settings (including custom tracks), click here."

Using the Table Browser

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the [Table Browser User's Guide](#).

- **clade:** Specifies which clade the organism is in.
- **genome:** Specifies which organism data to use.
- **assembly:** Specifies which version of the organism's genome sequence to use.
- **group:** Selects the type of tracks to be displayed in the track list. The options correspond to the track groupings shown in the Genome Browser. Select 'All Tracks' for an alphabetical list of all available tracks in all groups. Select 'All Tables' to see all tables including those not associated with a track.
- **database:** (with 'All Tables' group option) Determines which database should be used for options in table menu.
- **track:** Selects the annotation track data to work with. This list displays all tracks belonging to the group specified in the group list.
- **table:** Selects the SQL table data to use. This list shows all tables associated with the track specified in the track list.
- **describe table schema:** Displays schema information for the tables associated with the selected track.
- **region:** Restricts the query to a particular chromosome or region. Select "position" to apply the query to the entire genome or ENCODE regions. To limit the query to a specific position, type a

BEDTools to the rescue!



Find SNPs that have the potential to alter gene expression regulation by affecting methylation at CpG islands.

Wednesday @ 1pm

Annotation Summary

- Three major approaches to annotate a genome

1. Alignment:

- Does this sequence align to any other sequences of known function?
- Great for projecting knowledge from one species to another

2. Prediction:

- Does this sequence statistically resemble other known sequences?
- Potentially most flexible but dependent on good training data

3. Experimental:

- Lets test to see if it is transcribed/methylated/bound/etc
- Strongest but expensive and context dependent

- Many great resources available

- Learn to love the literature and the databases
- Standard formats let you rapidly query and cross reference
- Google is your number one resource ☺

- Coming up:

- IGV, QC, Variant Analysis, De novo assembly, Transcriptome, etc...





Thank you!

<http://schatzlab.cshl.edu>

@mike_schatz