

# Genome Annotations

Michael Schatz

Nov 15, 2014

Adv Sequencing Course



# Goal: Genome Annotations

aatgcatgicggctatgctaagcatgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaa  
tgcatgicggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgicggctatgc  
taatgaatggtcttgggatttaccttgggaatgctaagctgggatccgatgacaatgcatgicggctatgctaatga  
atggtcttgggatttaccttgggaatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgic  
ggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatc  
cggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatc  
ctgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicgg  
ctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatc  
cggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatc  
cgatgactatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctggg  
aatgcatgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaag  
ctgggatccgatgactatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaag  
atggtcttgggatttaccttgggaatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicgg  
ttgggatttaccttgggaatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgg  
gatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicgg  
tgicggctatgctaagctgggatccgatgacaatgcatgicggctatgctaagctgggatccgatgacaatgcatgicgg

# Goal: Genome Annotations

aatgcatgicggctatgctaagctgicggctatgctaagctgicggatccgatgacaatgcatgicggctatgctaa  
tgcatgicggctatgcaagctgicggatccgatgactatgctaagctgicggatccgatgacaatgcatgicggctatgic  
taatgaatggtcttgggatttaccttgggaatgctaagctgicggatccgatgacaatgcatgicggctatgctaatga  
atggtcttgggatttaccttgggaatgctaagctgicggctatgctaagctgicggatccgatgacaatgcatgic  
ggctatgctaagctgicggctatgcaagctgicggatccgatgactatgctaagctgicggctatgctaagctgic  
cggctatgctaagctgicggatccgatgacaatgcatgicggctatgctaagctgicggctatgcaagctgicggatc  
ctgicggctatgctaagctgicggatccgatgacaatgcatgicggctatgctaagctgicggatccgatgacaatgcatgicgg  
ctatgctaagctgicggatccgatgacaatgcatgicggctatgctaagctgicggatccgatgacaatgcatgicgg  
cggctatgctaagctgicggatccgatgacaatgcatgicggctatgctaagctgicggatccgatgacaatgcatgicgg  
cggctatgctaagctgicggatccgatgacaatgcatgicggctatgctaagctgicggatccgatgacaatgcatgicgg  
cggctatgctaagctgicggatccgatgacaatgcatgicggctatgctaagctgicggatccgatgacaatgcatgicgg  
aatgcatgicggctatgctaagctgicggatccgatgacaatgcatgicggctatgctaagctgicggatccgatgacaatgcatgicgg  
ctgicggatccgatgactatgctaagctgicggctatgctaagctgicggctatgctaagctgicggctatgctaagctgicgg  
atggtcttgggatttaccttgggaatgctaagctgicggatccgatgacaatgcatgicggctatgctaagctgicgg  
ttgggatttaccttgggaatgctaagctgicggatccgatgacaatgcatgicggctatgctaagctgicggatccgatgacaatgcatgicgg  
gatccgatgacaatgcatgicggctatgctaagctgicggatccgatgactatgctaagctgicggatccgatgacaatgcatgicgg  
tgicggctatgctaagctgicggatccgatgacaatgcatgicggatccgatgacaatgcatgicgg

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 an 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

# Seed and Extend

```
FAKDFLAGGVAAAI SKTAVAPIERVKLLLQVQHASKQITADKQYKGIIDCVVRIPKEQGV  
F D +GG AAA+ SKTAVAPIERVKLLLQVQ ASK I DK+YKGI+D ++R+PKEQGV  
FLIDLASGGTAAAV SKTAVAPIERVKLLLQVQDASKAIAVDKRYKGYMDVLI RVPKEQGV
```

- 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 :

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

$K, \lambda$  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  LSPADKTNVKAAWGKVG AHAGEYGA EALERMFLSFPTTKTYFPHF-----DL SHGSAQV 55
      L+P +K+ V A WGKV + E G EAL R+ + +P T+ +F F D G+ +V
Sbjct 3  LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV 60

Query 56 KGHGKKVADALTNAVAHVDDMPNALSALS DLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA 115
      K HGKKV A ++ +AH+D++ + LS+LH KL VDP NF+LL + L+ LA H
Sbjct 61  KAHGKKVLGAFSDGLAHL DNLKGT FATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK 120

Query 116 EFTP AVHASLDKFLASVSTVLTSKY 140
      EFTP V A+ K +A V+ L KY
Sbjct 121 EFTPPVQAAYQKVVAGVANALAHKY 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  LSPADKTNVKAAWGKVGHAHAGEYGAELERMFLSFPTTKTYFPHF-----DLSHGSAQV 55
      LS +  V  WGKV A    +G E L R+F  P T  F  F      D  S  +
Sbjct 3  LSDGEWQLVLNVWVGKVEADIPGHGQEV LIRLFKGH PETLEKFDKFKHLKSEDEM KASEDL 62

Query 56 KGHGKKVADALTNVAHVDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA 115
      K HG V AL  +          + L+ HA K ++      + +S C++ L + P
Sbjct 63 KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG 122

Query 116 EFTP AVHASLDKFLASVSTVLT SKYR 141
      +F      +++K L      + S Y+
Sbjct 123 DFGADAQGAMNKALELFRKDMASNYK 148
```

# Not similar sequences

Query: HBA\_HUMAN Hemoglobin alpha subunit  
Sbjct: SPAC869.02c [Schizosaccharomyces pombe]

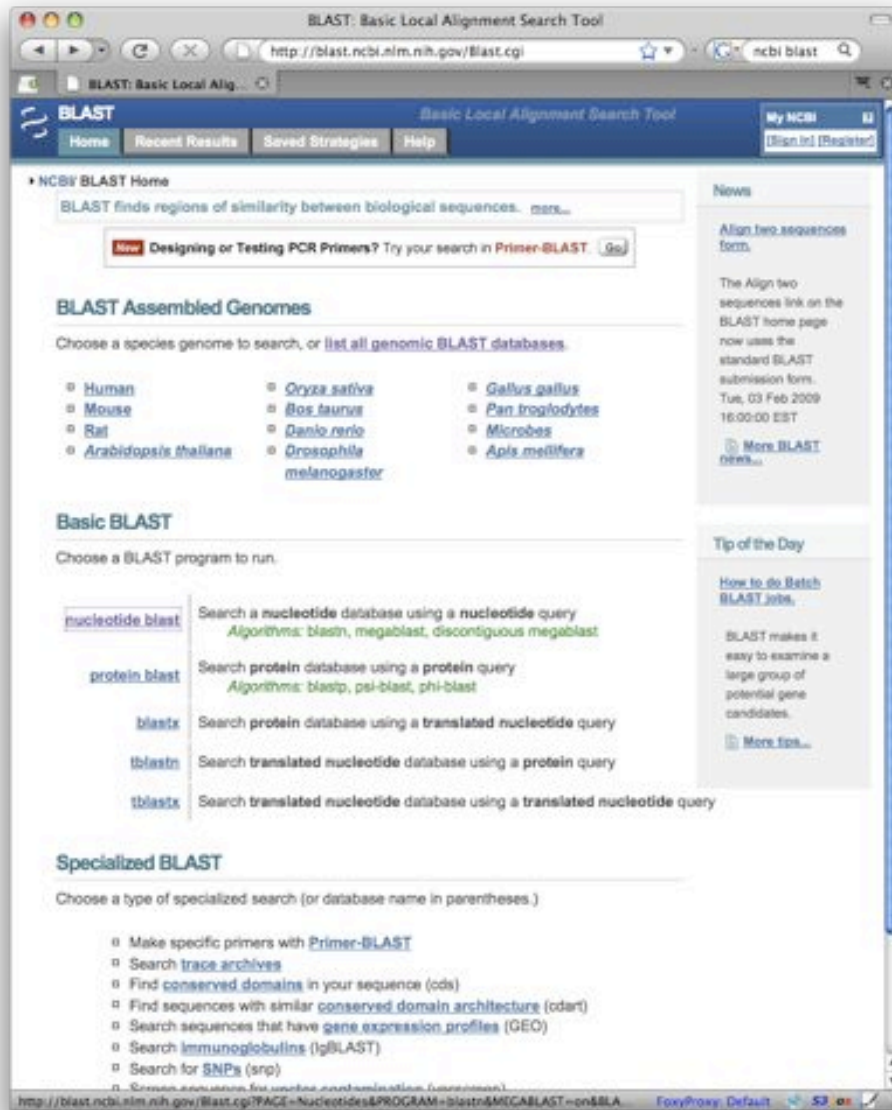
Score = 33.1 bits (74), Expect = 0.24  
Identities = 27/95 (28%), Positives = 50/95 (52%), Gaps = 10/95 (10%)

```
Query 30  ERMFLSFPTTKTYFPHFDSLHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALS  
      ++M  ++P      P+F+ +H  +      + +A AL N  ++DD+  +LSA  D  89  
Sbjct 59  QKMLGNYPEV---LPYFNKAHQISL--SQPRILAFALLNYAKNIDDL-TLSAFMDQIVV 112  
  
Query 90  K---LRVDPVNFKLLSHCLLVTLAAHLPAEF-TPA 120  
      K   L++   ++ ++ HCLL T+   LP++  TPA  
Sbjct 113  KHVGLQIKAEHYPIVGHCLLSTMQELLPSDVATPA 147
```

# Blast Versions

<b>Program</b>	<b>Database</b>	<b>Query</b>
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]

GATTACA  
1 2 3 4 5 6 7

## *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]

GATTACA  
0 1 2 3 4 5 6

# Genomic Conventions

## ***1-based coordinates***

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

GATTACA

1234567

## ***0-based coordinates***

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

GATTACA

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

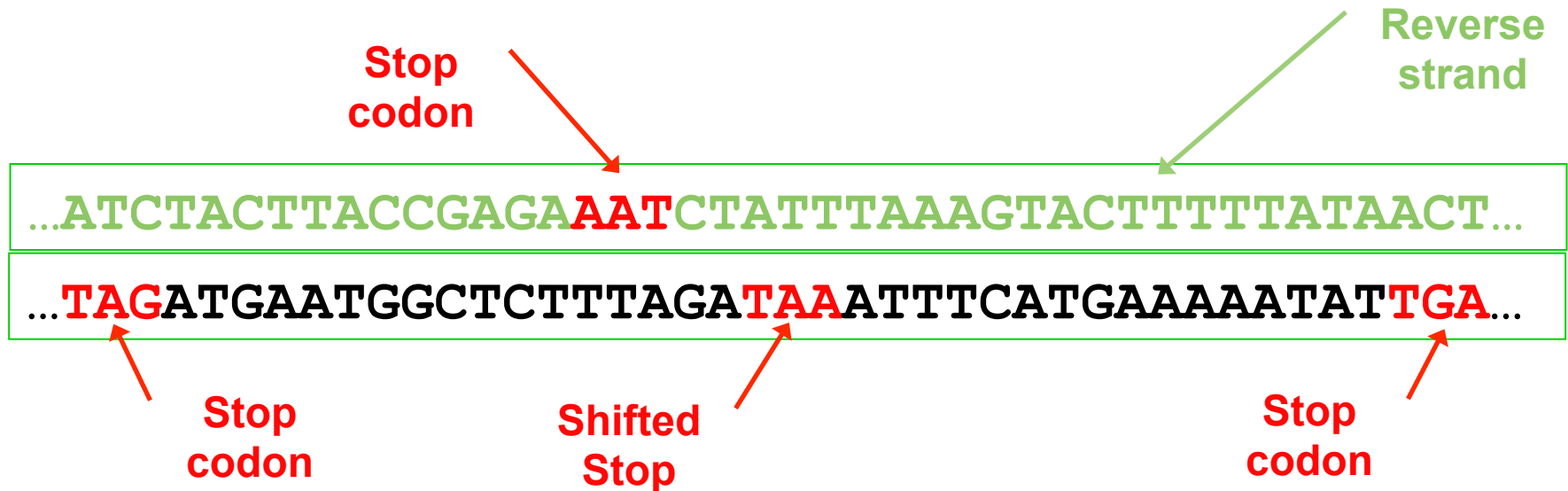
# Step One

- Find open reading frames (ORFs).



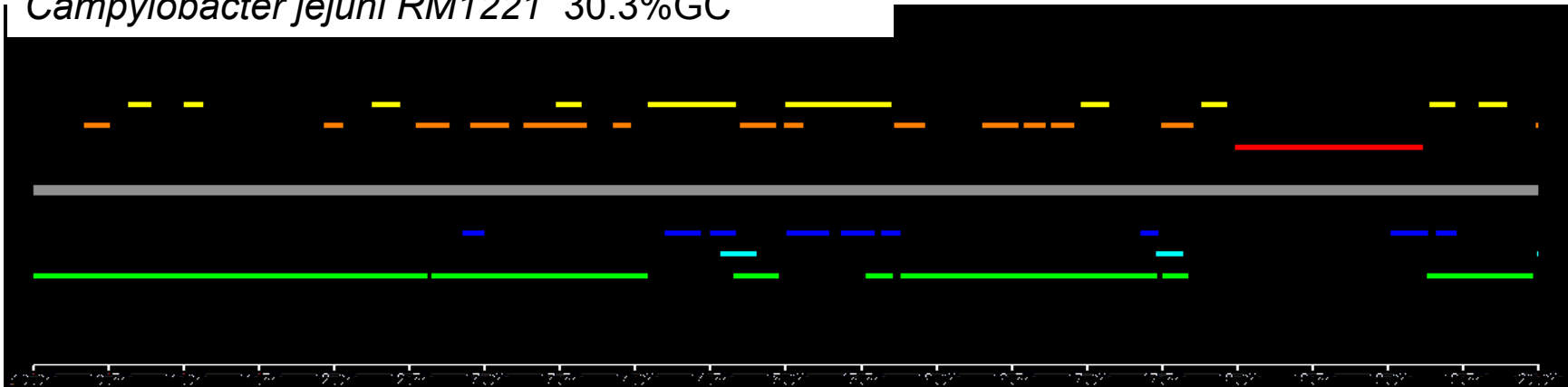
# Step One

- Find open reading frames (ORFs).



- But ORFs generally overlap ...

*Campylobacter jejuni* RM1221 30.3%GC

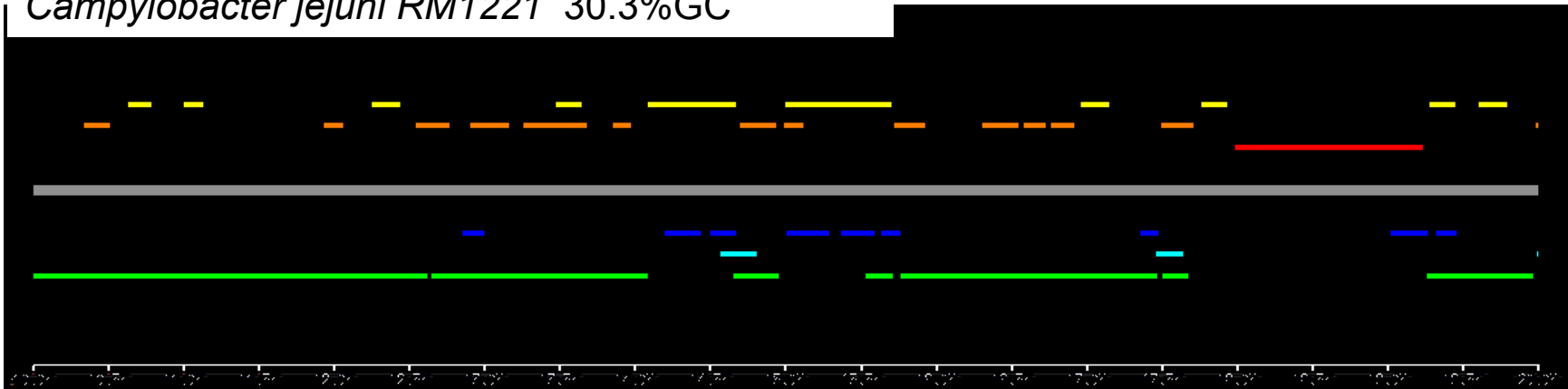


All ORFs longer than 100bp on both strands shown  
- color indicates reading frame  
Longest ORFs likely to be protein-coding genes

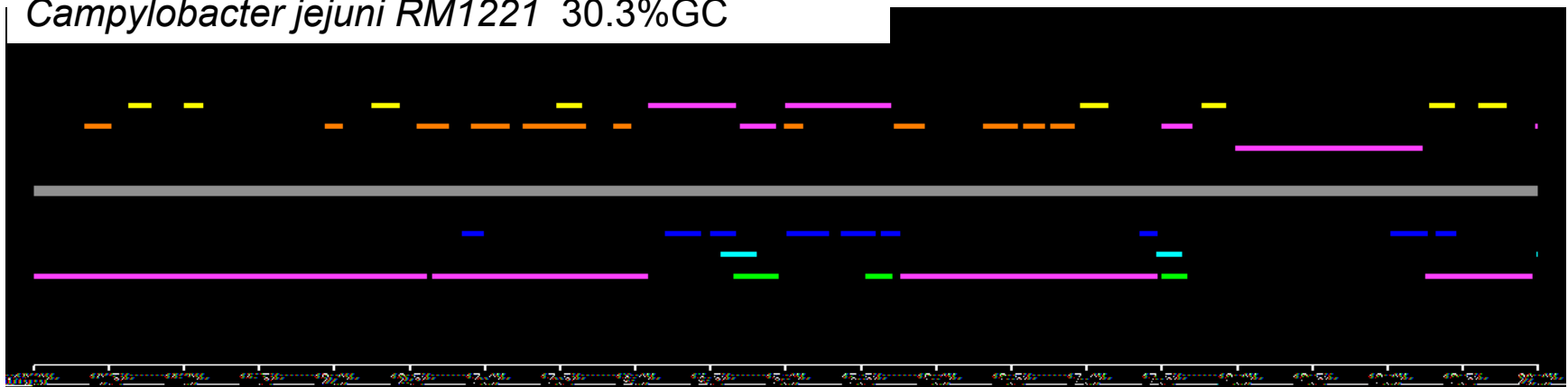
Note the low GC content

All genes are ORFs but not all ORFs are genes

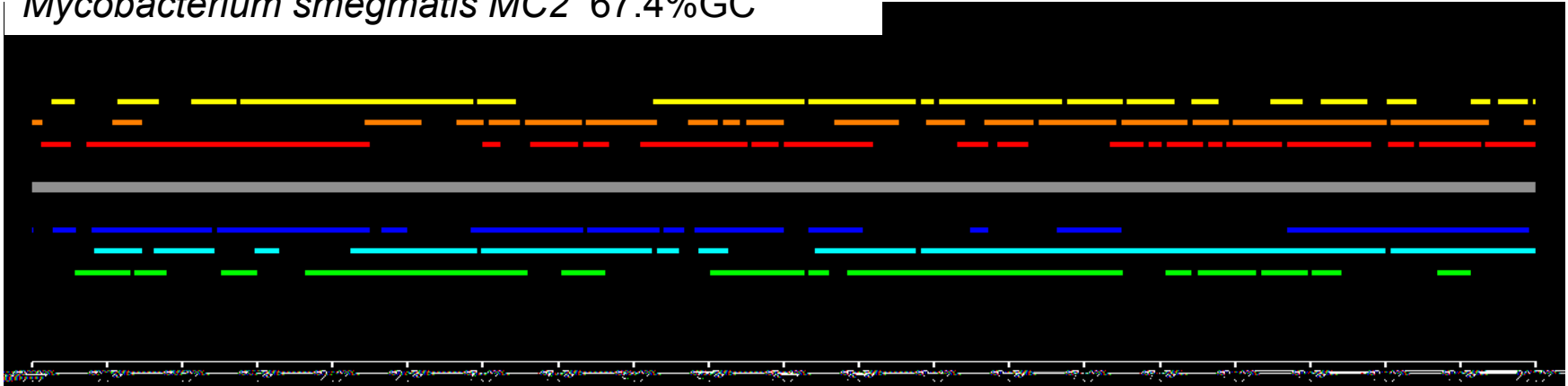
*Campylobacter jejuni* RM1221 30.3%GC



*Campylobacter jejuni* RM1221 30.3%GC

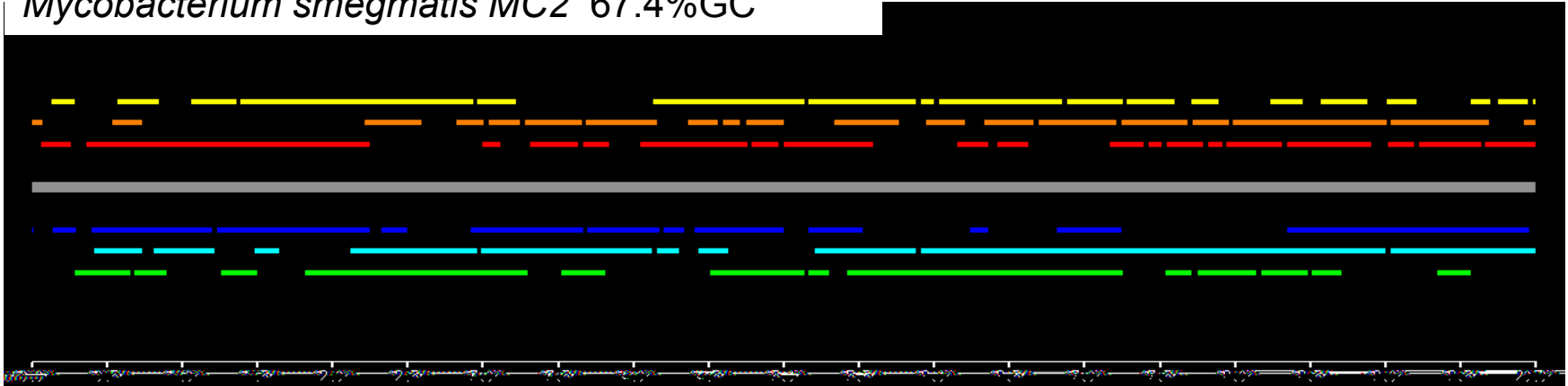


*Mycobacterium smegmatis* MC2 67.4%GC

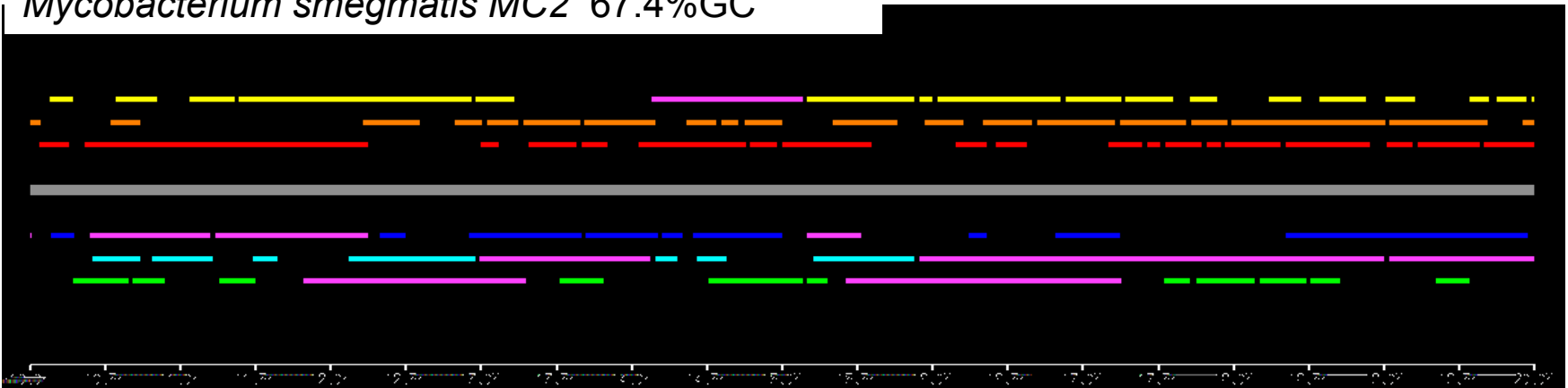


Note what happens in a high-GC genome

*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



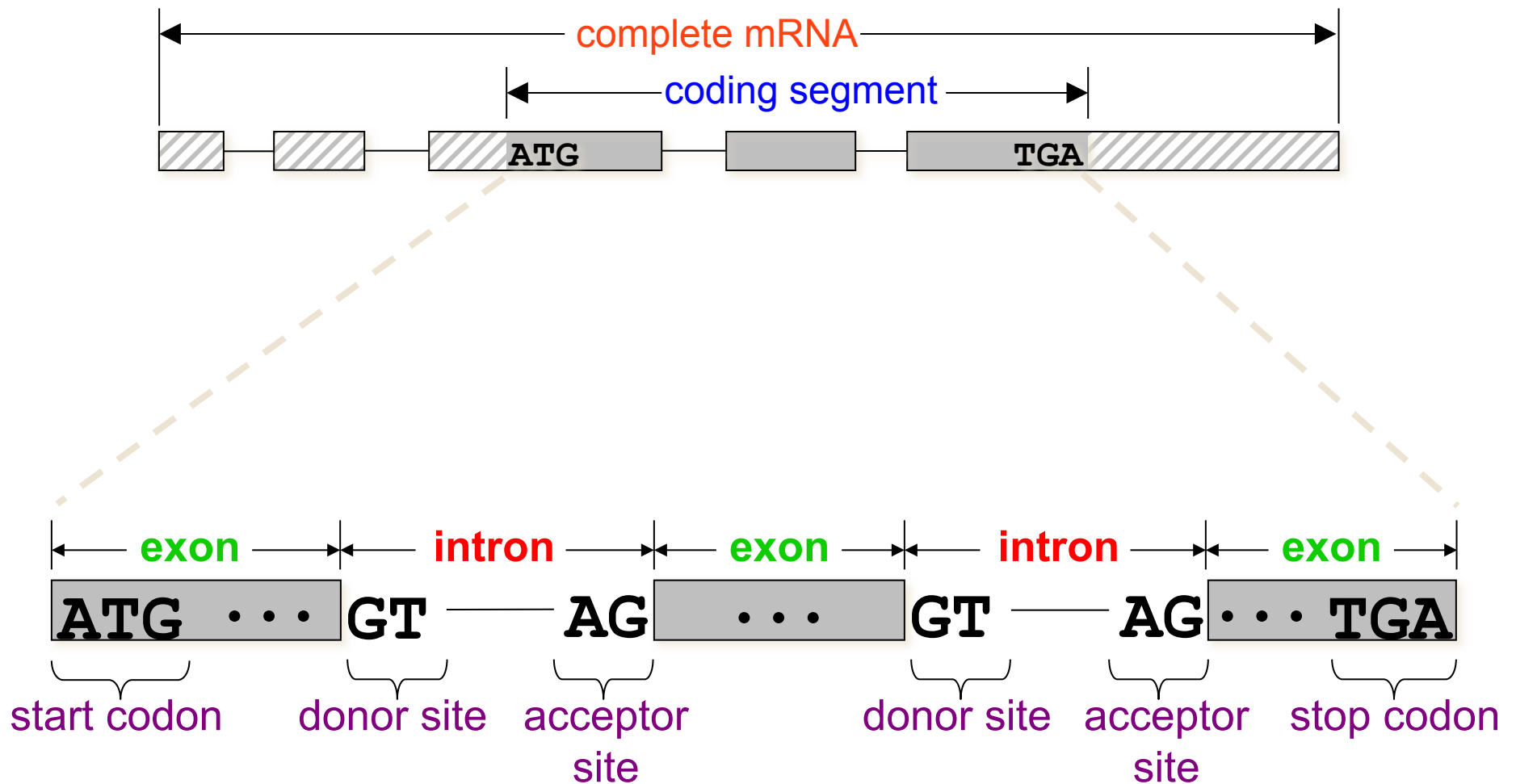


# Overview of Eukaryotic Gene Prediction

CBB 231 / COMPSCI 261

*W.H. Majoros*

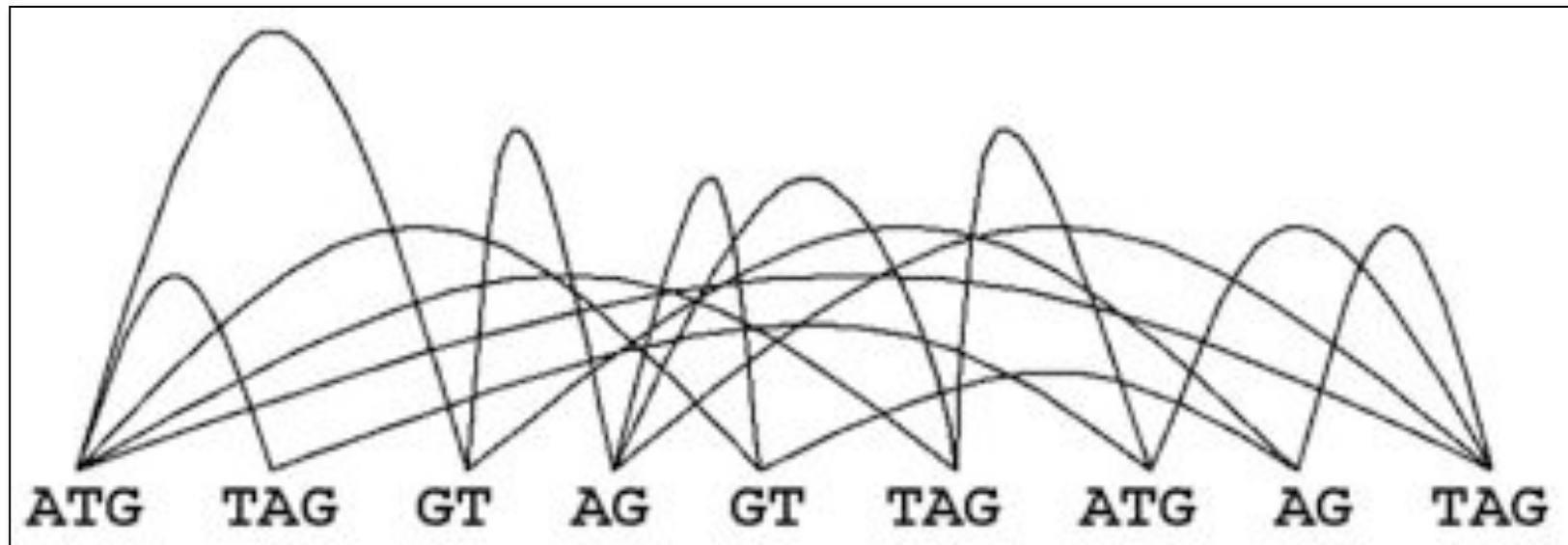
# Eukaryotic Gene Syntax



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

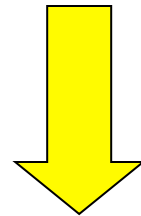
After identifying the most promising (i.e., highest-scoring) signals in an input sequence, we can apply the gene syntax rules to connect these into an *ORF graph*:



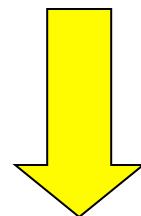
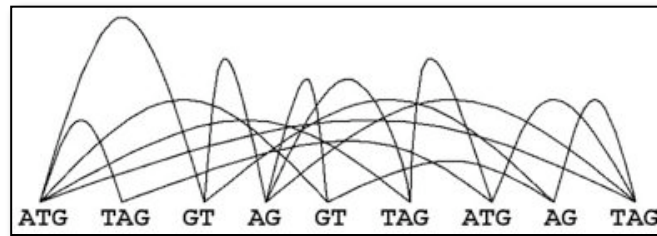
An ORF graph represents all possible *gene parses* (and their scores) for a given set of putative signals. A *path* through the graph represents a single gene parse.

# Conceptual Gene-finding Framework

TATTCCGATCGATCGATCTCTCTAGCGTCTACG  
CTATCATCGCTCTCTATTATCGCGCGATCGTCG  
ATCGCGGAGAGTATGCTACGTCGATCGAATTG



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



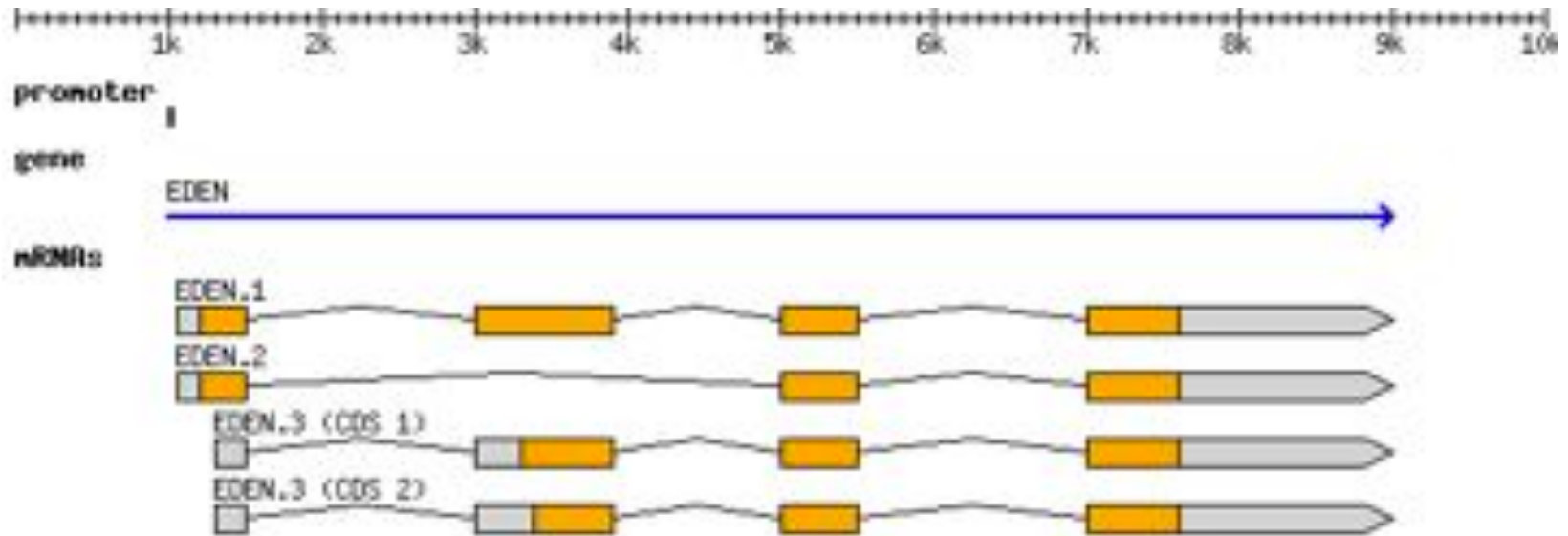
find highest-scoring path through ORF graph; interpret path as a gene parse = gene structure



# 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>

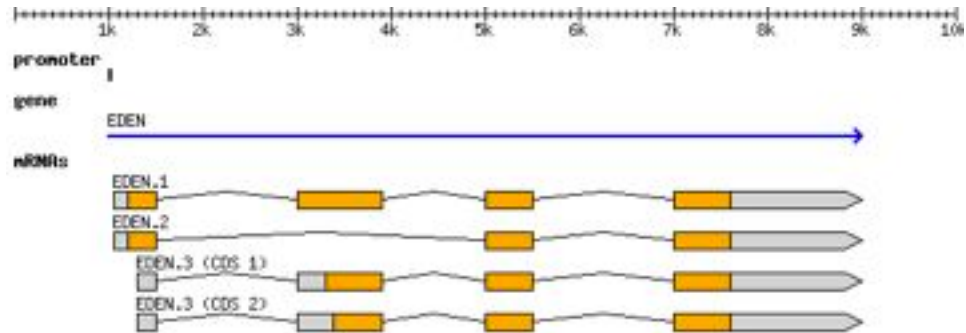
# GFF File format

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
```





Break

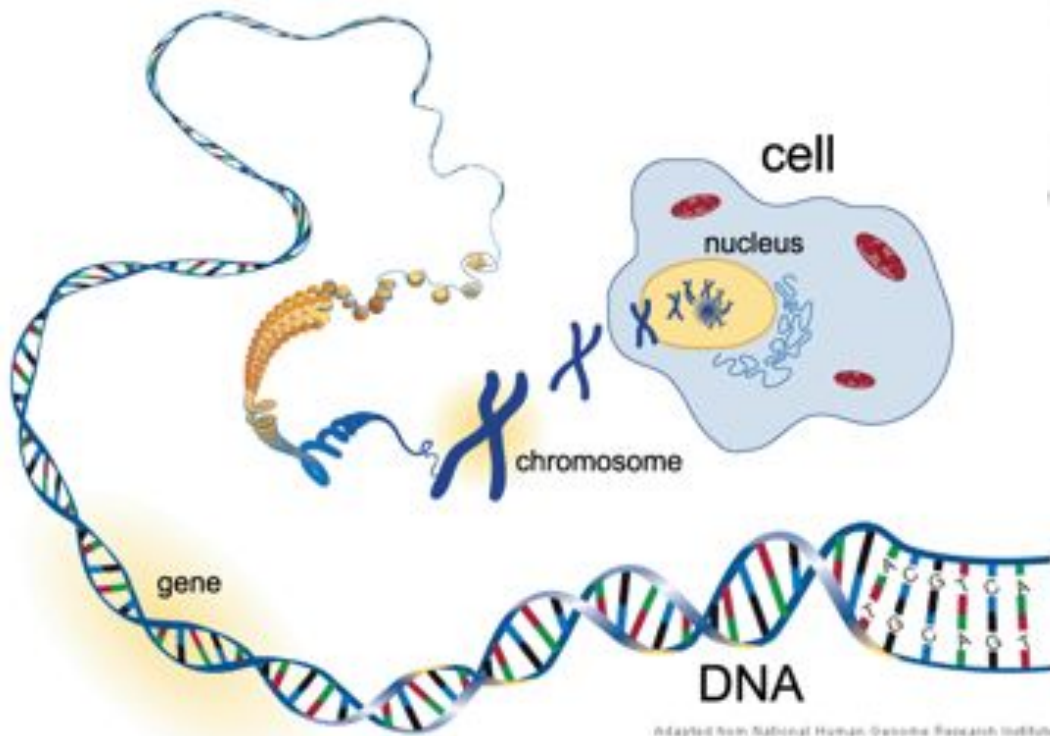
# Outline

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

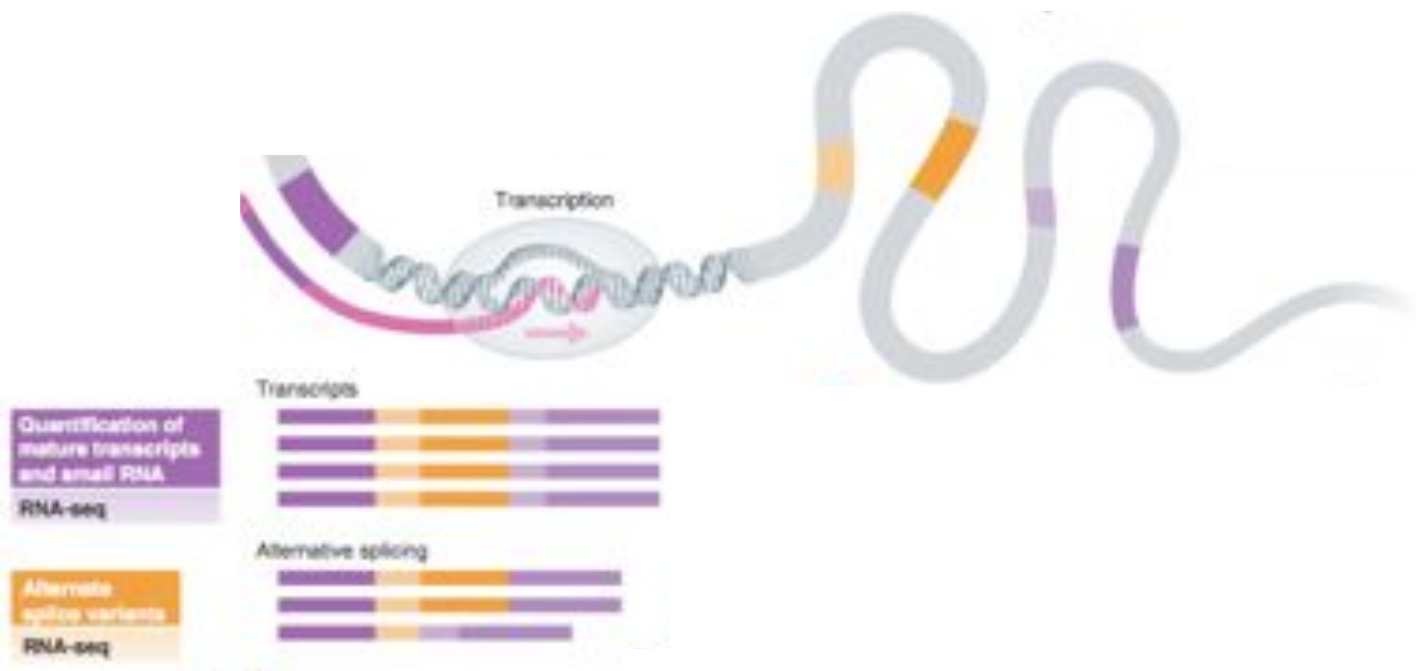


# Sequencing techniques

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



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

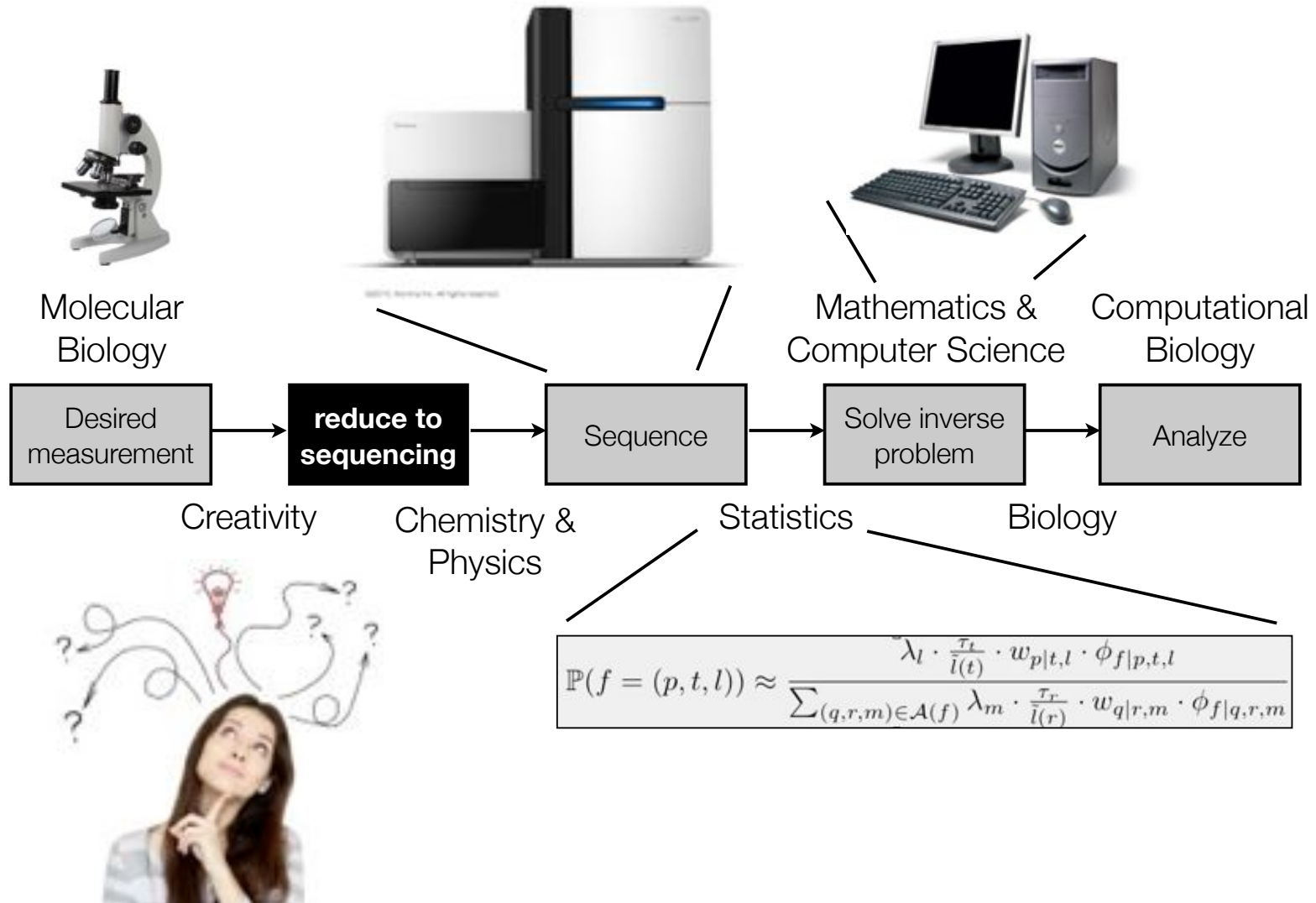


# Sequencing Assays

## The \*Seq List (in chronological order)

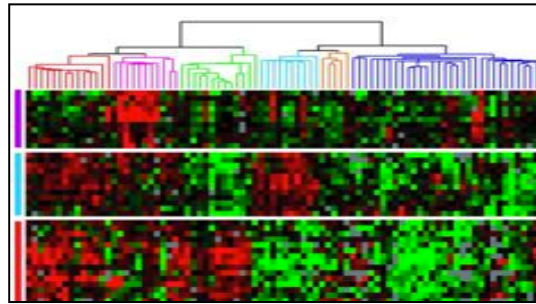
1. Gregory E. Crawford et al., "Genome-wide Mapping of DNase Hypersensitive Sites Using Massively Parallel Signature Sequencing (MPSS)," *Genome Research* 16, no. 1 (January 1, 2006): 123–131, doi:10.1101/gr.4074106.
2. David S. Johnson et al., "Genome-Wide Mapping of in Vivo Protein-DNA Interactions," *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?

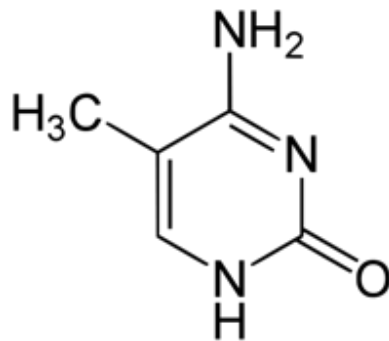


# \*-seq in 3 short vignettes

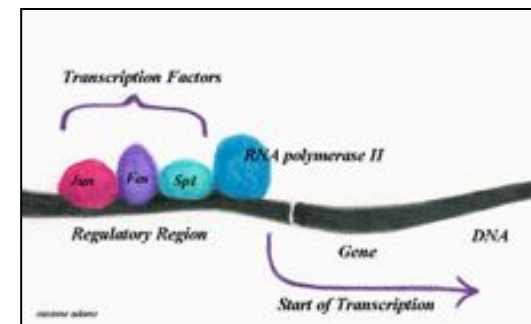
## RNA-seq



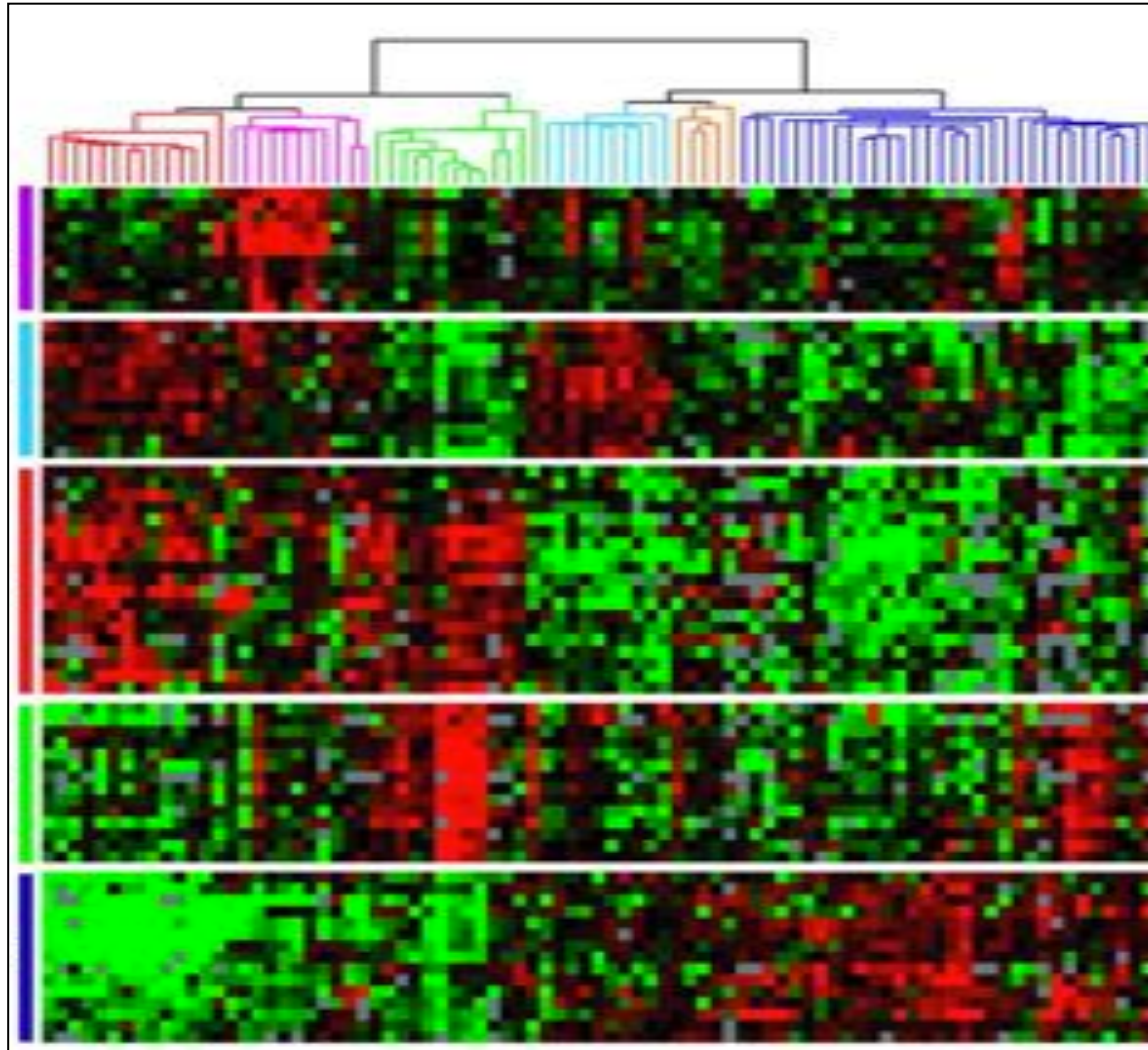
## Methyl-seq



## ChIP-seq



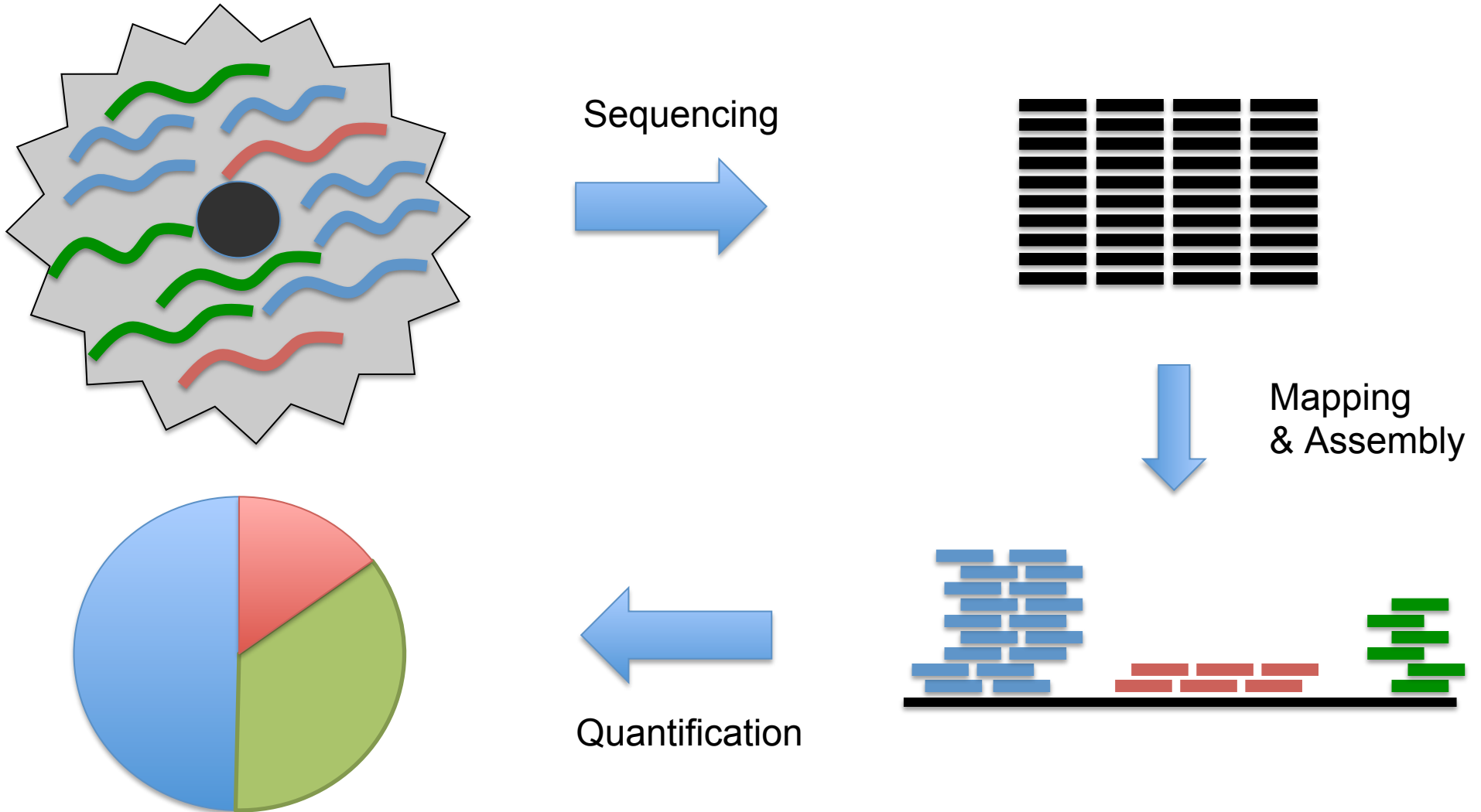
# RNA-seq



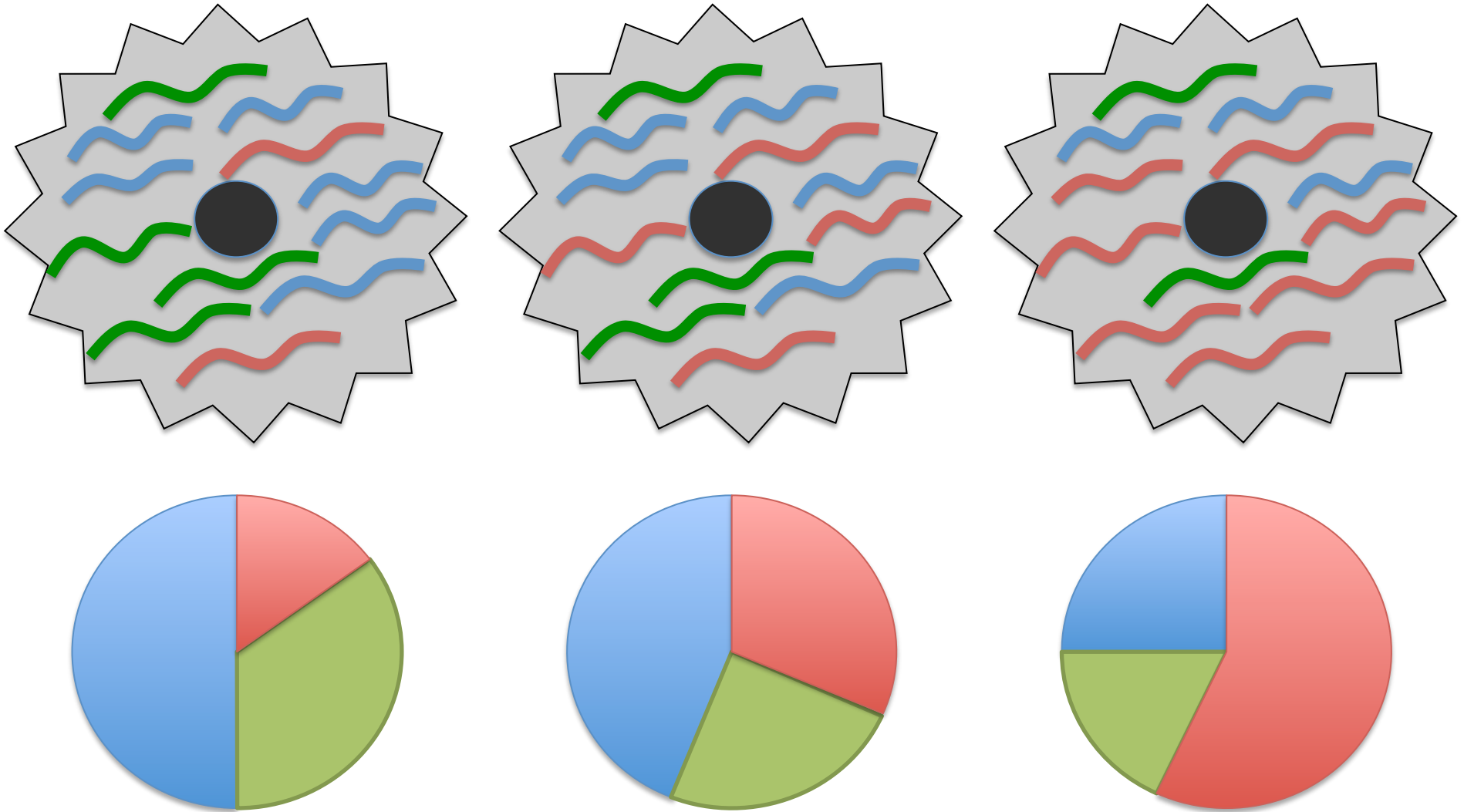
**Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.**  
Sørlie et al (2001) *PNAS*. 98(19):10869-74.



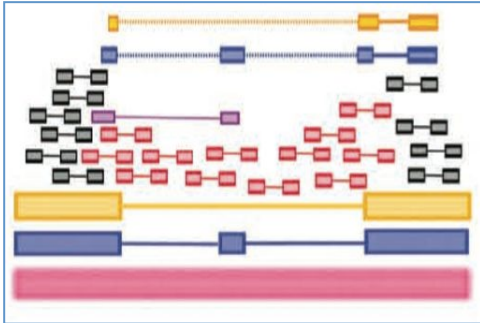
# RNA-seq Overview



# RNA-seq Overview



# RNA-seq Challenges

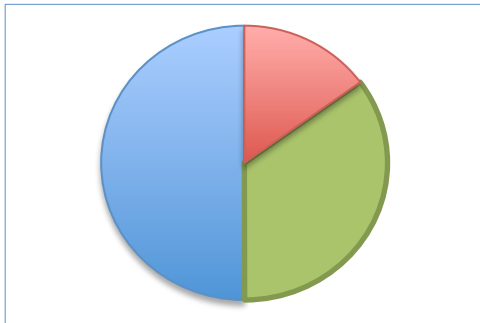


## Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

### TopHat: discovering spliced junctions with RNA-Seq.

Trapnell et al (2009) *Bioinformatics*. 25:0 | 105-111

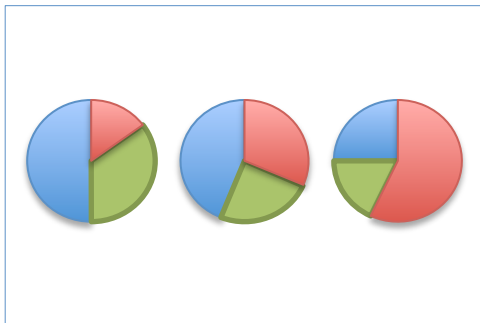


## Challenge 2: Read Count $\neq$ Transcript abundance

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?

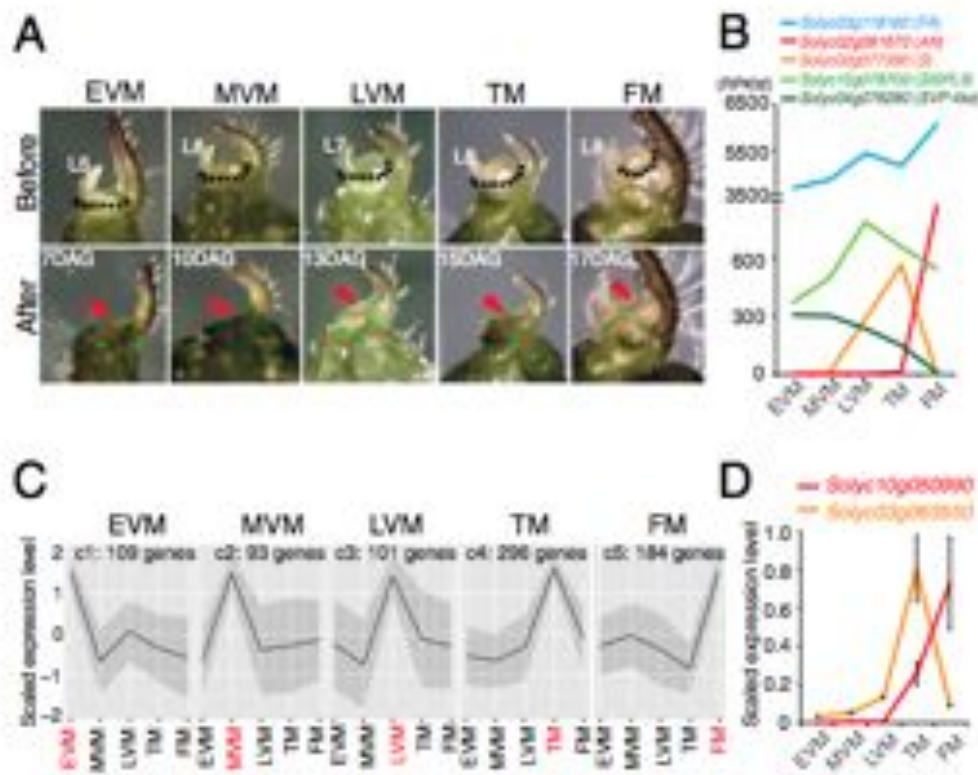
Liu et al (2013) *Bioinformatics*. doi:10.1093/bioinformatics/btt688

# Rate of meristem maturation determines inflorescence architecture in tomato

Soon Ju Park<sup>1</sup>, Ke Jiang<sup>1</sup>, Michael C. Schatz, and Zachary B. Lippman<sup>2</sup>

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

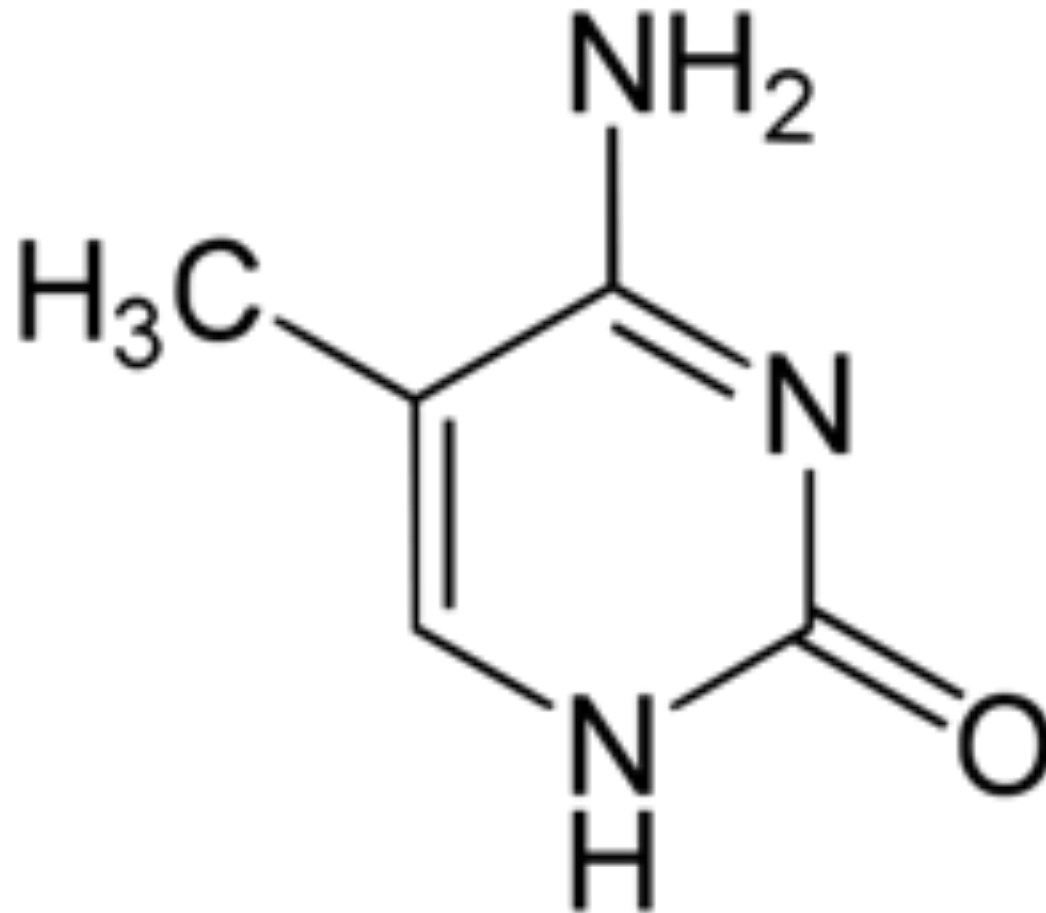
Edited by Maarten Koornneef, Wageningen University and Research Centre, Cologne, Germany, and approved November 28, 2011 (received for review September 12, 2011)



## 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<sup>1\*</sup>, Sylvain Foret<sup>2,3</sup>, Robert Kucharski<sup>3</sup>, Stephan Wolf<sup>4</sup>, Cassandra Falckenhayn<sup>1</sup>, Ryszard Maleszka<sup>3\*</sup>

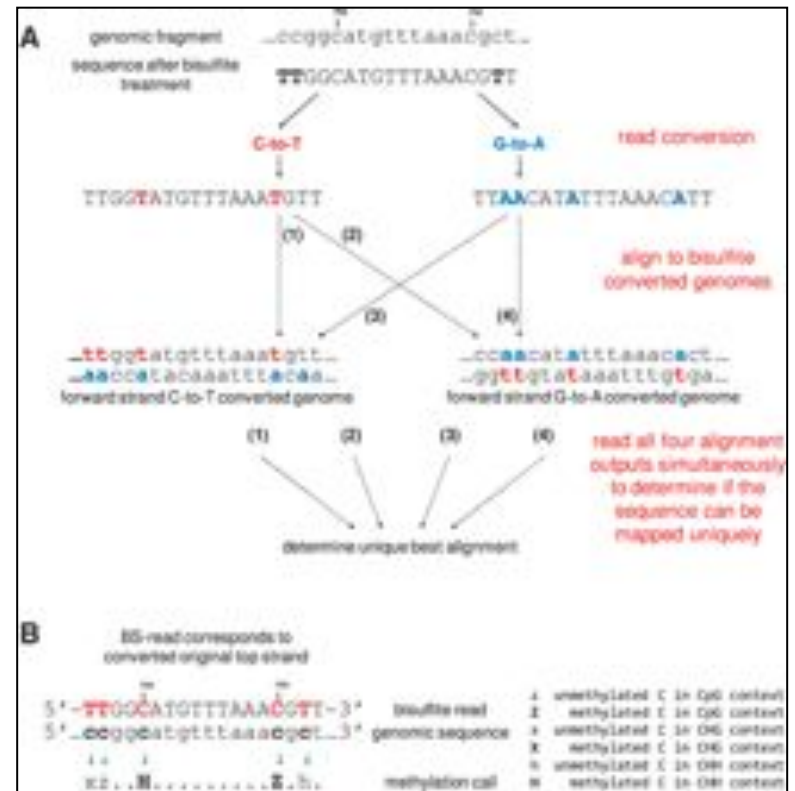
**1** Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany, **2** ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Australia, **3** Research School of Biology, the Australian National University, Canberra, Australia, **4**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

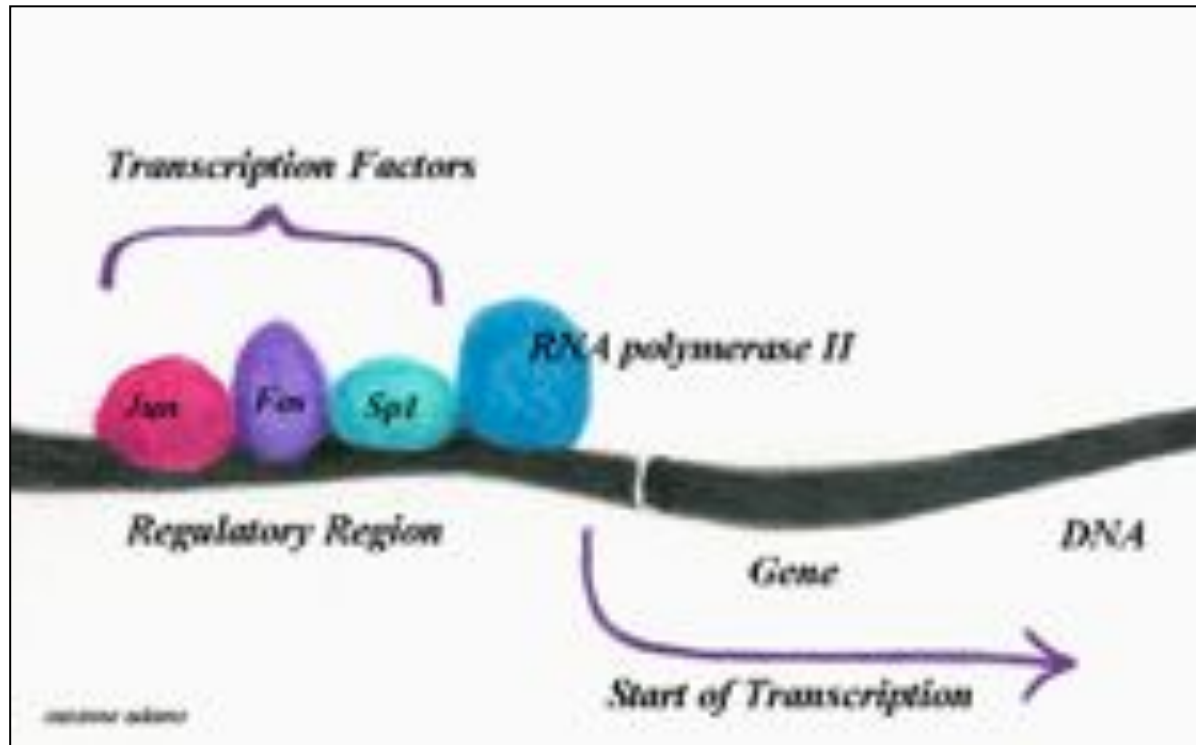
T  
W  
•  
•  
•



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



# ChIP-seq



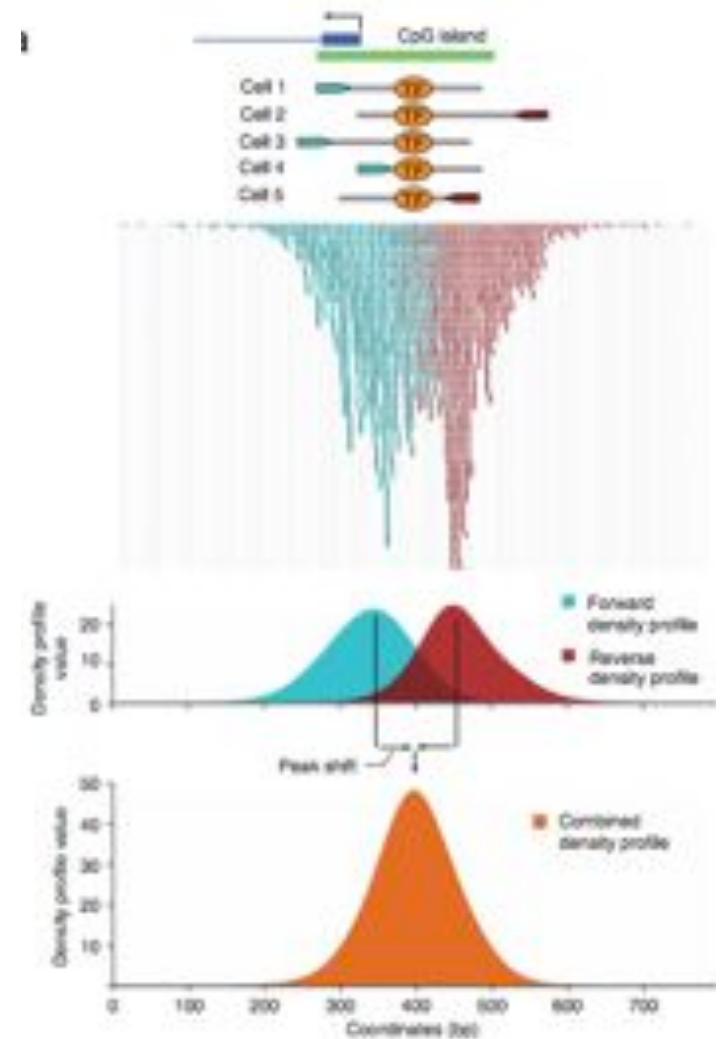
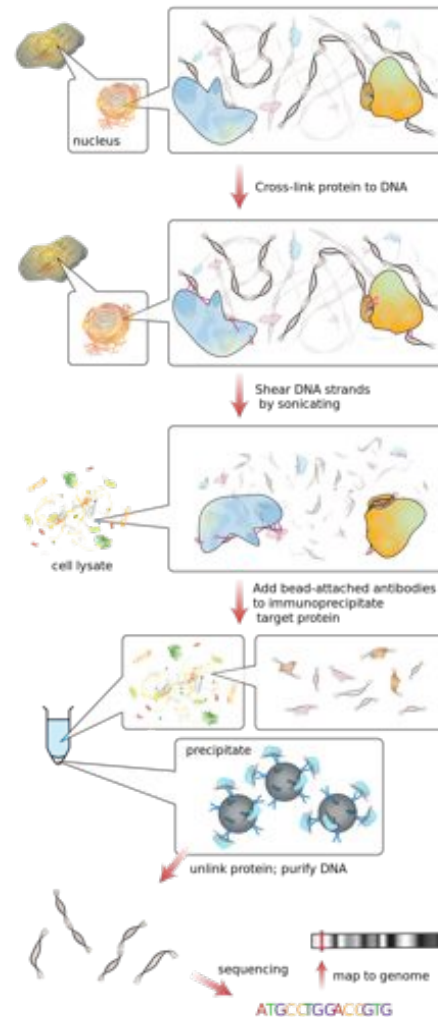
**Genome-wide mapping of in vivo protein-DNA interactions.**

Johnson et al (2007) *Science*. 316(5830):1497-502

# ChIP-seq

## Goals:

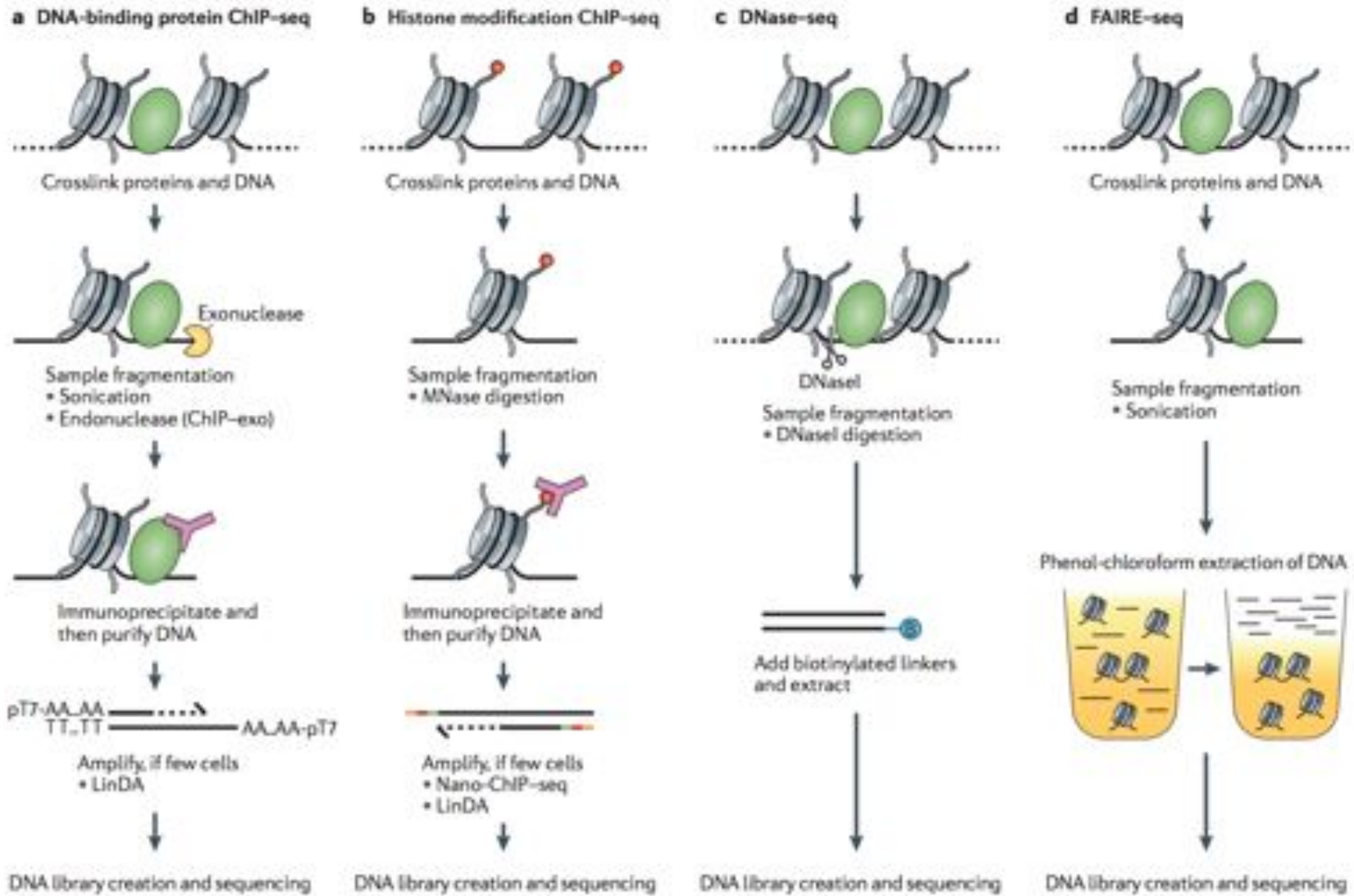
- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?



## Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data

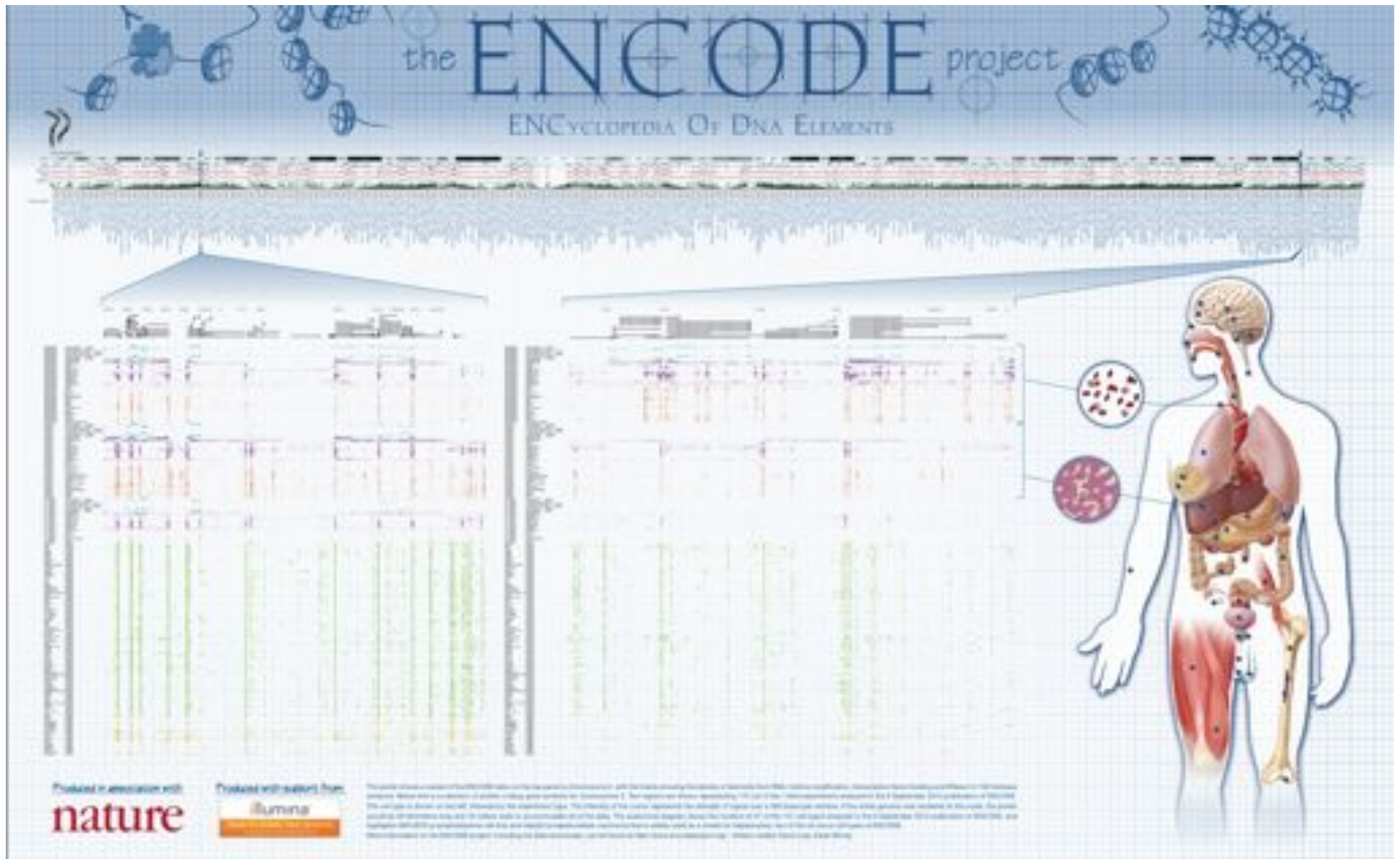
Valouev et al (2008) *Nature Methods*. 5, 829 - 834

# 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***

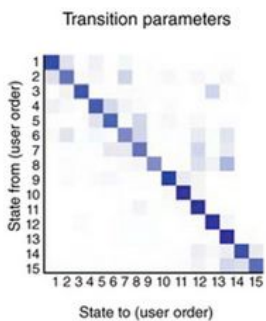
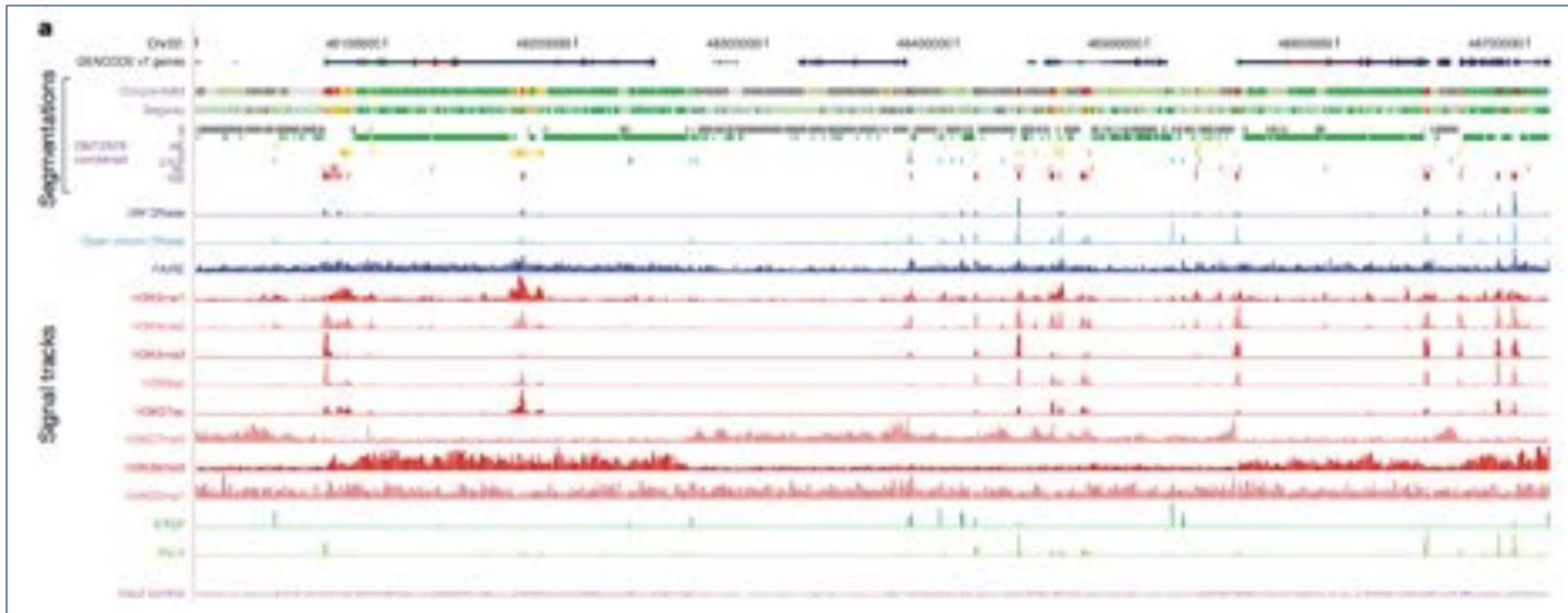
# Summary of ENCODE elements

*“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



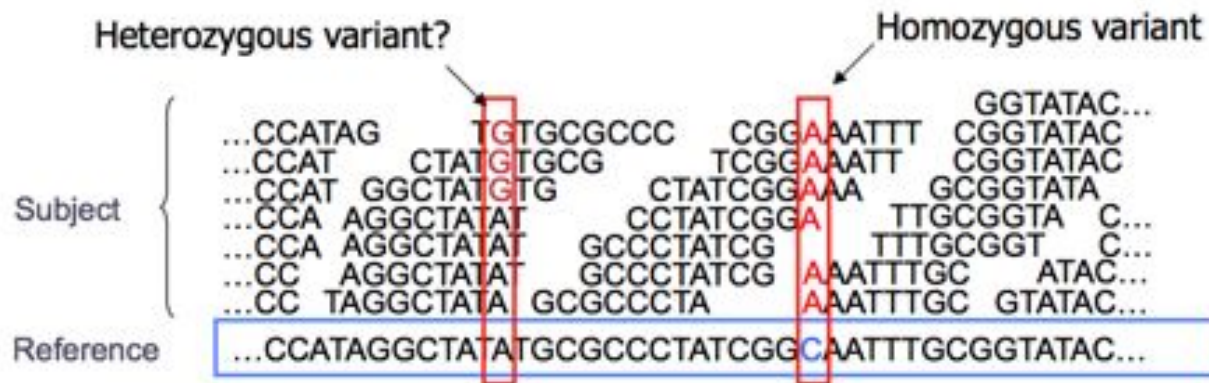
Label	Description	Details*	Colour
CTCF	CTCF-enriched element	Sites of CTCF signal lacking histone modifications, often associated with open chromatin. Many probably have a function in insulator assays, but because of the multifunctional nature of CTCF, we are conservative in our description. Also enriched for the cohesin components RAD21 and SMC3; CTCF is known to recruit the cohesin complex.	Turquoise
E	Predicted enhancer	Regions of open chromatin associated with H3K4me1 signal. Enriched for other enhancer-associated marks, including transcription factors known to act as enhancers. In enhancer assays, many of these (>50%) function as enhancers. A more conservative alternative would be cis-regulatory regions. Enriched for sites for the proteins encoded by EP300, FOS, FOSL1, GATA2, HDACB, JUNB, JUN, NFE2, SMARCA4, SMARCB1, SIRT5 and TAL1 genes in K562 cells. Have nuclear and whole-cell RNA signal, particularly poly(A)- fraction.	Orange
PF	Predicted promoter flanking region	Regions that generally surround TSS segments (see below).	Light red
R	Predicted repressed or low-activity region	This is a merged state that includes H3K27me3 polycomb-enriched regions, along with regions that are silent in terms of observed signal for the input assays to the segmentations (low or no signal). They may have other signals (for example, RNA, not in the segmentation input data). Enriched for sites for the proteins encoded by BRP2, CIZ1B, MARK, TRIM28, ZNF274 and SETDB1 genes in K562 cells.	Grey
TSS	Predicted promoter region including TSS	Found close to or overlapping GENCODE TSS sites. High precision/recall for TSSs. Enriched for H3K4me3. Sites of open chromatin. Enriched for transcription factors known to act close to promoters and polymerases Pol II and Pol III. Short RNAs are most enriched in these segments.	Bright red
T	Predicted transcribed region	Overlap gene bodies with H3K36me3 transcriptional elongation signal. Enriched for phosphorylated form of Pol II signal (elongating polymerase) and poly(A) <sup>+</sup> RNA, especially cytoplasmic.	Dark green
WE	Predicted weak enhancer or open chromatin cis-regulatory element	Similar to the E state, but weaker signals and weaker enrichments.	Yellow

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

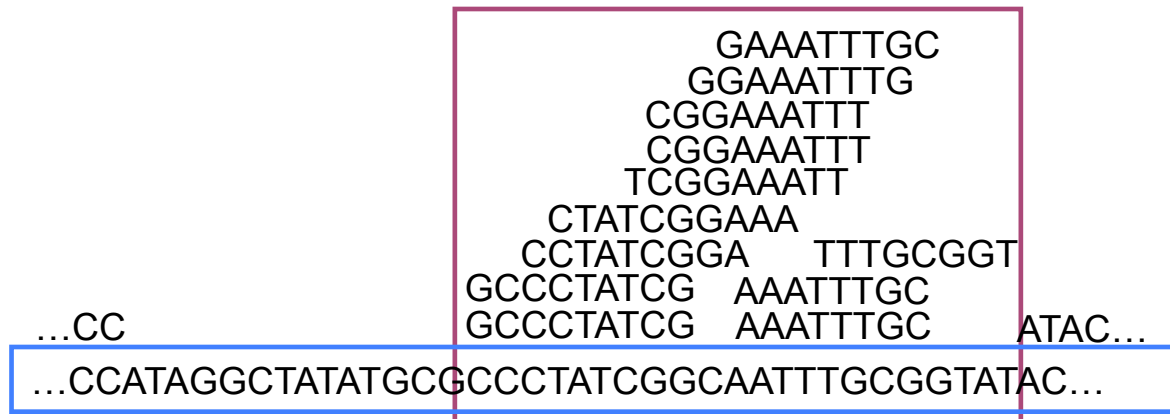
**ChromHMM: automating chromatin-state discovery and characterization**  
 Ernst & Kellis (2012) *Nature Methods*. doi:10.1038/nmeth.1906

# Genotyping vs \*-seq

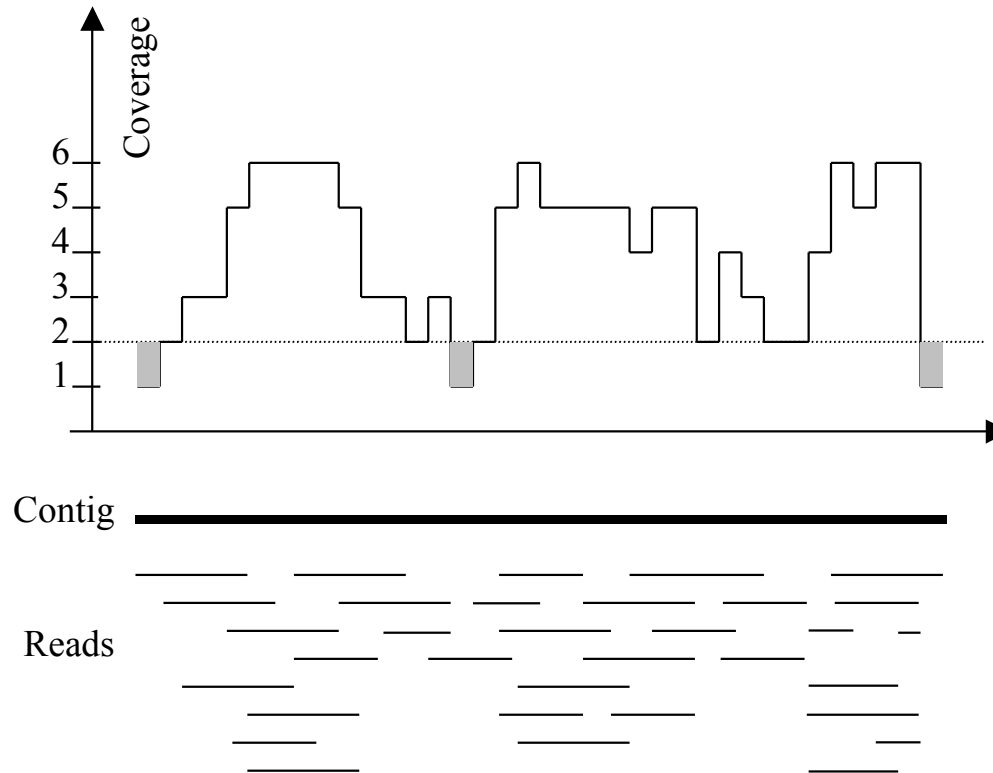
- Genotyping: Identify Variations



- \*-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 100<sup>th</sup> or every 1000<sup>th</sup>
- 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
```

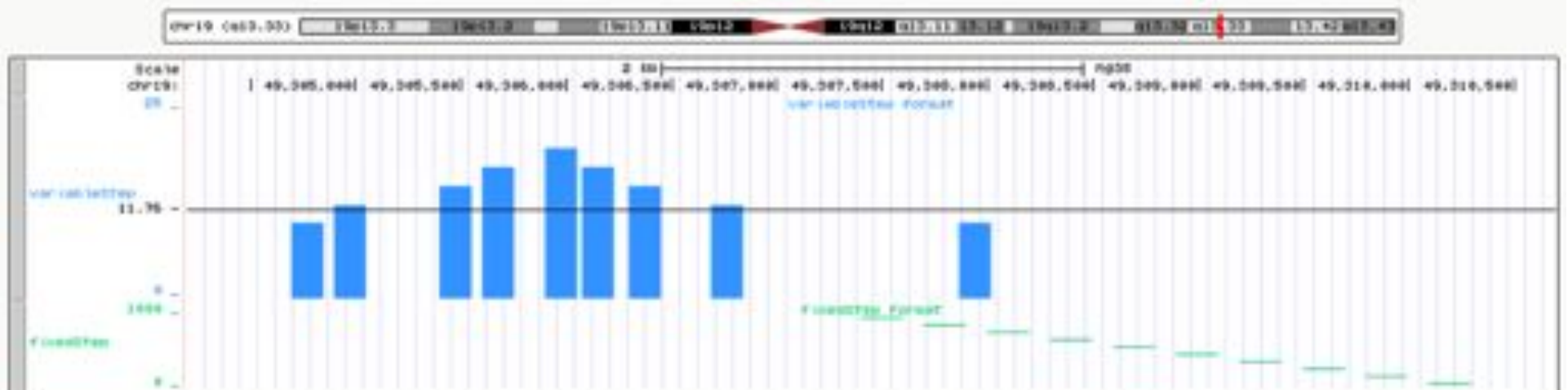
# WIG Example

```
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
```

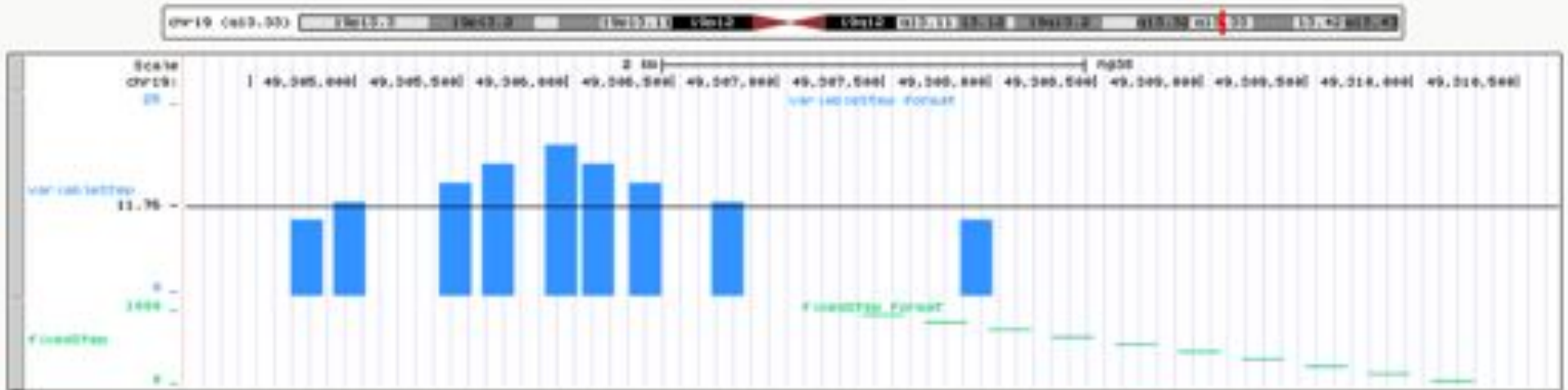
# WIG Example

```
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.***

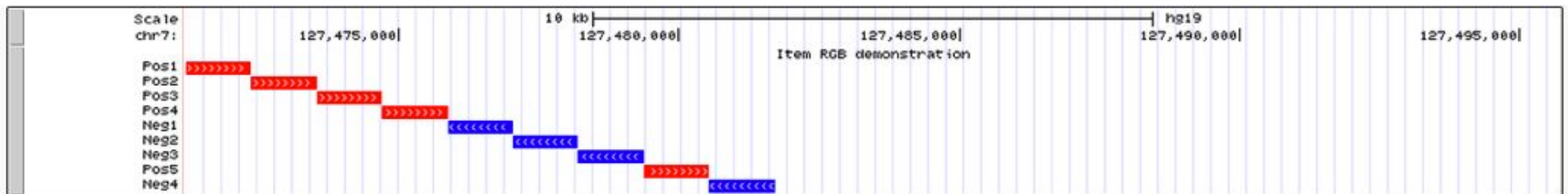
The 9 additional optional BED fields are:

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.

<http://genome.ucsc.edu/FAQ/FAQformat.html>

# 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
```



<http://genome.ucsc.edu/FAQ/FAQformat.html>

# 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/>

The screenshot shows the NCBI website homepage. At the top, there is a navigation bar with "NCBI Resources" and "How To" menus, and a user profile for "machatn@ohio.edu" with "My NCBI" and "Sign Out" options. Below this is a search bar with a dropdown menu set to "All Databases" and a "Search" button. On the left side, there is a vertical navigation menu with various categories like "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", and "Variation". The main content area features a "Welcome to NCBI" section with a brief description of the center's mission and a list of links for "About the NCBI", "Mission", "Organization", "Research", and "RSS Feeds". Below this is a "Get Started" section with a bulleted list of links for "Tools", "Downloads", "How-To's", and "Submissions". A featured article titled "Genomic Structural Variation" is displayed, with a text box stating "dbVar archives large scale genomic variation data and associates defined variants with phenotypic information." and an image of a corn cob. At the bottom of the featured article is a pagination bar with numbers 1 through 8. On the right side, there is a "Popular Resources" section with links to "PubMed", "Bookshelf", "PubMed Central", "PubMed Health", "BLAST", "Nucleotide", "Genome", "SNP", "Gene", "Protein", and "PubChem". Below this is an "NCBI Announcements" section with a link to "NCBI's July Newsletter is on the Bookshelf" and a date of "13 Aug 2012".

NCBI Resources How To machatn@ohio.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>

The screenshot shows the Ensembl website interface. At the top, there is a navigation bar with the Ensembl logo and links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. A search bar is located on the right side of the header. Below the header, there is a main search area with a dropdown menu for species and a search input field. The main content area is divided into several sections:

- Browse a Genome:** A section with a description of the Ensembl project and a list of popular genomes including Human, Mouse, and Zebrafish. It also includes a search for all genomes and a link to view the full list of species.
- Functional Genomics:** A grid of tiles for ENCODE data, Variant Effect Predictor (Vep), Gene expression in different tissues, and Find SNPs and other variants for my gene.
- Genomic Tools:** Tiles for Retrieve gene sequence, Compare genes across species, Use my own data in Ensembl, and Learn about a disease or phenotype.
- What's New in Release 77 (October 2014):** A section with a list of updates, including GENCODE 21, First Ensembl Havana rat merge, and New species: Yarrow Akodon green monkey.
- Latest blog posts:** A list of recent blog entries with dates and titles.
- Tweets:** A section displaying tweets from Ensembl and other users, including a tweet from Paul Flicek about the cat genome.
- Did you know...?** A section with a globe icon and a link to the Ensembl blog.

At the bottom of the page, there is a footer with the Sanger logo, a description of the Ensembl project as a joint effort between EMBL-EBI and the Wellcome Trust, and a link to the acknowledgements page. There are also social media icons for Twitter, Facebook, and YouTube, and a link to the Ensembl GitHub repository.

# Biomart

<http://www.biomart.org>

The screenshot shows the BioMart website homepage. At the top left is the BioMart logo. A navigation menu on the left lists: HOME, TOOLS, COMMUNITY, PUBLICATIONS, NEWS, CREDITS, DOCUMENTATION, VERSION 0.7, and CONTACT. A 'Download v 0.8' button is in the top right. A news banner at the top center reads: '(20th October 2014) BioMart Portals (0.7 and 0.8) updates, ensemble, ensemble genomes and uniprot.' The main content area features a 'BioMart' section with a definition: 'is a community-driven project to provide unified access to distributed research data to facilitate the scientific discovery process.' Below this is a paragraph: '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.' 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'. Below the text are four icons: 'SEARCH DATA', 'BI CONVERSION', 'GENOMES ACTUAL', and 'LINK LINKED ANALYSIS'. A 'READ MORE' link is below the icons. At the bottom left, there are two featured articles: 'ENTREPRISE' and 'Ensemble Genomes'. At the bottom right, a yellow box contains the text: '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.' To the right of this box is a world map titled '46 DATABASES, 4 CONTINENTS AND GROWING' with a '21' badge and a 'View map' link.

bio·mart

(20th October 2014) BioMart Portals (0.7 and 0.8) updates, ensemble, ensemble genomes and uniprot.

Download v 0.8

## 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

WORKING TOGETHER TO GET THE MOST FROM OUR SOFTWARE TOGETHER

SEARCH DATA BI CONVERSION GENOMES ACTUAL LINK LINKED ANALYSIS

READ MORE

## ENTREPRISE

FROM THE SOURCE: FROM A CLIP TO A PROGRAM

## Ensemble Genomes

FORWARD BY MULTIPLE PATHWAYS

### 46 DATABASES, 4 CONTINENTS AND GROWING

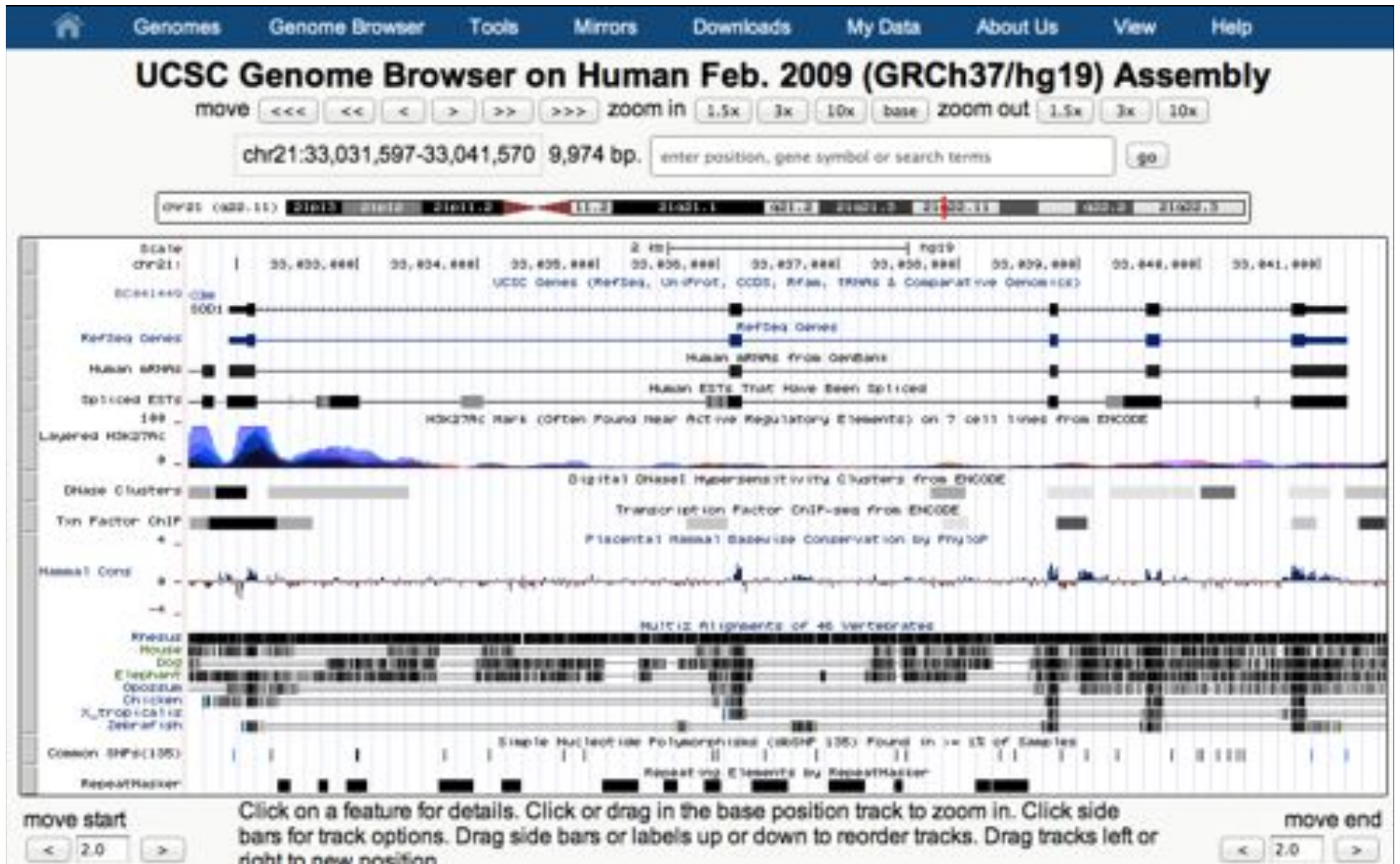
21

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.

View map

# UCSC Genome Browser

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



# UCSC Genome Browser / Table Browser

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



The screenshot shows the UCSC Table Browser web interface. At the top, there is a navigation menu with links for Home, Genomes, Genome Browser, Blast, Tables, Gene Sorter, PCR, Session, FAQ, and Help. The main content area is titled "Table Browser" and contains a detailed form for querying genomic data. The form includes fields for clade (set to "human"), genome (set to "human"), assembly (set to "Feb. 2009 GRCh37/hg19"), group (set to "Comparative Genomics"), track (set to "Conservation"), and table (set to "Primate Cons. (phyloP/cons/Primate)"). There are also buttons for "describe table schema", "define regions", "filter", "subtrack merge", "intersection", "correlation", "output format" (set to "text format"), "Send output to" (with options for Galaxy and GREAT), "output file", and "file type returned" (with options for plain text and gzip-compressed). A note at the bottom of the form states: "Now, to return 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." Below the form are buttons for "get output" and "summary/statistics". A link is provided to "reset all user cart settings (including custom tracks), click here".

**Table Browser**

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix [Table Browser tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data.

clade:  genome:  assembly:

group:  track:

table:

region:  genome  position:

filter:

subtrack merge:

intersection:

correlation:

output format:  Send output to  Galaxy  GREAT

output file:  (leave blank to keep output in browser)

file type returned:  plain text  gzip-compressed

*Now, to return 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.*

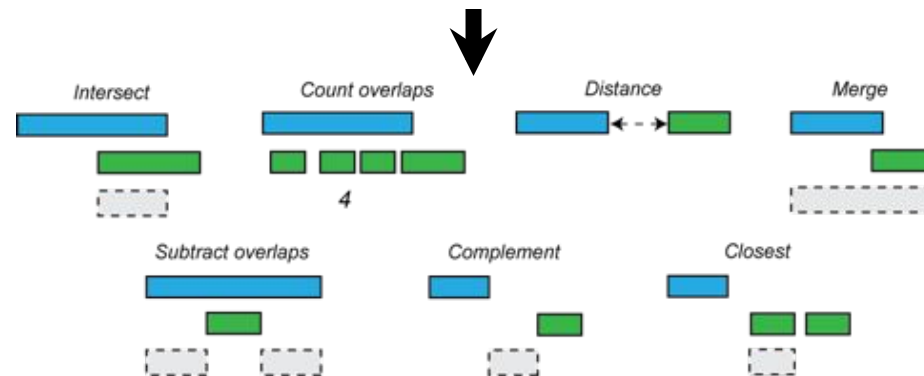
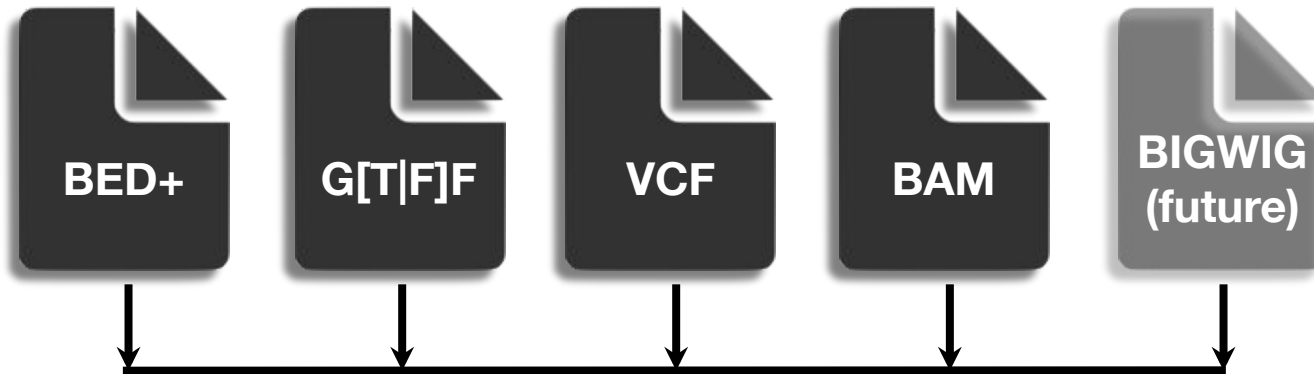
To reset all user cart settings (including custom tracks), [click here](#).

**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 *genome* to apply the query to the entire genome or *ENCODE* to examine only the 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