Genomics

- Genome sequencing & fragment assembly
- Gene prediction

**What is genomics?**
- The study of all DNA of an organism (genome)

**Issues**
- Sequencing the genome
- Identifying genes (transcribed DNA) on the genome
- Determining gene function
- Discover how each gene is regulated
- Study natural variations in gene among / between species
- Study variations between healthy / diseased genes
Genomics

- Genome sequencing
  - Made possible by automated methods
  - Scientists → less data generation, more data analysis

Genomics Overview

- Outline
  - Molecular biology techniques
    - Restriction enzyme digests
    - Cloning
    - Sequence tagged sites (STS)
  - Sequencing
  - Assembly
  - Gene structure
  - Gene prediction
Technique – Restriction Enzymes

- **Restriction enzyme**
  - Proteins found in bacteria
  - Cuts DNA at specific pattern (usually palindrome)

- **Example**
  - EcoRI
    
    5’ – G A A T T C – 3’
    3’ – C T T A A G – 5’
    
    cut

    5’ – G C T T A A A T A C – 3’
    3’ – A T A C G – 5’

Sticky ends

Technique – Restriction Enzymes

- **Restriction enzymes**
  - Over 300 restriction enzymes found in bacteria
  - Cut DNA at different **recognition sites** (patterns)
    - Smaller pattern → more frequent cuts → many small fragments
    - Larger pattern → less frequent cuts → few large fragments

- **Restriction mapping**
  - Cutting DNA with multiple restriction enzymes
  - Analyzing # of fragments, order of breaks
  - Determines relative positions of recognition sites in DNA

- **Clone library**
  - Collection of DNA fragments from genome
  - Contains redundancy, overlaps
Clone Library Example

Molecular Biology Technique – Cloning

🔹 Cloning
  - Creates large amounts of target DNA
    1. Insert DNA fragments into vectors
    2. Grow vector in laboratory
    3. Recover DNA from vector

🔹 Cloning vectors
  - Chromosome-like carriers for target DNA
  - Plasmids (small extra-chromosomal pieces of DNA)
    - Up to 25K base pairs
  - Bacteria BACs (Bacterial Artificial Chromosome) / yeast DNA
    - For 100K to 1 million base pairs
Molecular Biology Technique – Cloning

- **Cloning algorithm**
  1. Cut DNA & vector with restriction enzymes
  2. Use complementary **sticky ends** to join DNA to vector
  3. Grow vector
  4. Extract DNA from vector

![Cloning algorithm diagram](image)

Technique – Sequence Tagged Sites (STS)

- **Sequence tagged site (STS)**
  - A short DNA sequence (about 200-300 bases)
  - Unique position in the genome
  - Probe for STS
    - Short strand of labeled DNA
    - Attaches (hybridizes) to STS

- Use STS probe to provide rough map of clones
Genomics Overview

- **Outline**
  - Molecular biology techniques
  - Sequencing
    - Physical mapping
    - Ordered cloning
    - Primer walking
    - Shotgun sequencing
  - Assembly
  - Gene structure
  - Gene prediction

Sequencing an Entire Genome

- **Physical mapping**
  - Break genome into clones (large contiguous fragments)
  - Find markers along the genome
  - Find unique overlapping clones covering the genome
    - Find which STS probes attach to which clone
    - Find order & orientation of clones

- **Sequencing clones**
  - Break clone into several short fragments (< 700 bps)
  - Automatically sequence fragments
  - Assemble fragments together
Sequencing – Using STS Probes

- **Physical mapping**
  - Find STS probes present in each contig
  - Results form STS matrix
  - Find permutation of columns in STS matrix [Booth 76]
    - Where 1’s in each row are consecutive
    - Yields order & overlap of contigs

- **Complications**
  - False positive – clone does not actually contain STS
  - False negative – clone contains unreported STS
  - Chimera – multiple DNA fragments combine and act as clone

Sequencing – Clones and Probes
### Sequencing – STS Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

### Sequencing – STS Matrix (Reordered)

<table>
<thead>
<tr>
<th></th>
<th>D</th>
<th>B</th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Sequencing – Automatic Sequencing

- **Automatic sequencers**
  - Alternative / complement to physical mapping
  - Limited to ~700-800 chunks known as “reads” due to
    - Biochemistry of DNA polymerase enzyme
    - Resolution of gel / capillary electrophoresis

- **Sequencing projects must (at some point)**
  1. Divide DNA into overlapping 700 bp fragments
  2. Assemble fragments into contiguous sequences (contigs)

- **Assembly is a computational problem**

Sequencing – Sequencing Strategy

- **Approaches to genome sequencing**
  - Ordered sub-cloning
  - Primer walking
  - Shotgun sequencing

- **Selecting an approach based on faith in**
  - Speed of sequence analysis
  - Reliability of assembly software
Sequencing – Ordered Cloning

◆ **Approach**
  - Divide large clones into small ordered overlapping fragments
  - Applying more detailed physical mapping to each clone

◆ **Observations**
  - Requires much more initial cloning work in the laboratory
  - Reduces # of actual sequencing reads required
  - Much easier to assemble the reads
  - Used by researchers who don't trust assembly software

Sequencing – Primer Walking

◆ **Approach**
  - Make new primer from the end of each new sequence read
  - Apply PCR to isolate next section of DNA
  - Sequence new section of DNA, repeat

◆ **Observations**
  - Each sequencing step uses information from previous read
  - Requires fast & accurate analysis of sequence reads
  - Skips sub-cloning of clones
  - Both order and overlap of reads are known
  - Very easy to assembly reads
  - Expensive to make a lot of PCR primers
Sequencing – Primer Walking Example

CMSC 838T – Lecture 10

Sequencing – Shotgun Sequencing

- **Approach**
  - Fragment complete genome into small DNA fragments
  - Assemble all fragments at once

- **Observation**
  - Exploits speed & low cost of automated sequencing
  - Relies on robust assembly software
  - Works well on small bacteria / virus genomes

- **Problem**
  - May not result in single contig for larger genomes
  - Rely on ordered cloning / primer walking to connect contigs
Sequencing – The Human Genome

- **Race to sequence the human genome**
  - Human Genome Project (academic consortium)
  - Celera (private company)

- **Human Genome Project used ordered cloning**
  - Breaking the genome into mapped BAC clones
  - Shotgun sequence the BAC clones

- **Celera used a modified shotgun method**
  - Random clones of various sizes (size selected libraries)
  - Plus relative mapping of clone ends (they must be located in the assembly at the correct distance and orientations)
  - Created custom assembly software
  - Made use of the “scaffold” built by the HGP

---

Genomics Overview

- **Outline**
  - Molecular biology techniques
  - Sequencing
  - Assembly
    - Issues
    - Algorithms
  - Gene structure
  - Gene prediction
Fragment Assembly

◆ Given
   - A collection of DNA fragments
   - Assemble fragments into maximal length contiguous sequences (contigs) using overlap information

Fragment Assembly – Approach

◆ Approach
   - Look for ungapped overlaps at end of fragments
   - High degree of identity over a short region
   - Exclude chance matches, but tolerate sequencing errors

◆ Match must be at ends of sequence
Fragment Assembly – Issues

- **Reads have errors**
  - Incorrectly determined bases, insertions & deletions
  - Error rate highest at beginning / end of reads
    - Precisely regions that need to be overlapped
- **Lack of sufficient coverage**
  - Fragments do not cover entire clone → separate contigs
- **Different fragments may combine (chimeras)**
- **Unknown orientation (which strand of DNA?)**
- **Repeats in the genome**
- **Vector contamination**
  - Sequences from cloning vectors included in read
  - Often at beginning / end of reads

Fragment Assembly – Possible Overlaps

CMSC 838T – Lecture 10
Fragment Assembly – Approach

◆ **Orientation**
  - For each fragment and partial contig formed
  - Consider both the sequence and its reverse complement

◆ **Optimal solution (in the absence of errors)**
  - Find shortest common superstring
  - Problem is NP-hard

◆ **Greedy algorithm**
  - Find two fragments with maximum overlap, combine
  - Repeat by treating contigs as fragments
  - Refinements reduce approximation factor to 2.2

Fragment Assembly – Refinements

◆ **Preprocessing**
  - Eliminate pairs of fragments w/o significant overlap
  - Compute optimal overlap between promising pairs
    - Using dynamic programming
  - If fragment is completely contained in another
    - Discard shorter fragment

◆ **Generating consensus sequence**
  - Find all overlapping fragments
  - Perform multiple sequence alignment
  - Results in better contigs
Genomics Overview

◆ **Outline**
  - Molecular biology techniques
  - Sequencing
  - Assembly
  - Gene structure
    - Promoter elements
    - Regulatory proteins
    - Open reading frames (ORF)
    - Alternative splicing
  - Gene prediction

Gene Structure

◆ **Views of a gene**
  - Portion of genomic DNA transcribed / expressed in mRNA
  - Active / useful portion of genome
  - DNA processed by RNA-polymerase enzyme

◆ **Real story is more complicated**
  - In prokaryotes, RNA-polymerase translates most ORFs
  - In eukaryotes, RNA-polymerase looks for many signals
  - mRNA undergoes processing before translation to protein
**Gene Structure**

- **RNA requires post-transcription modifications**
  - Capping – chemical alterations to 5’ end of RNA
  - Splicing – wholesale removal of sections of RNA
  - Polyadenylation – adding ~250 A’s to 3’ end of RNA
  - Produces many (heterogeneous) hnRNA intermediates

```
DNA  5'--------[Promoter]--------3'  
      |                        |  
      | Transcription           |  
      | 3'--[Exon:Intron:Exon]--3' |  
      |                      |  
      | RNA Splicing          |  
hnRNA 5'----[5' Cap]----------3'  
      |                        |  
      | mRNA                  |  
      | 3'--[5' Cap]----------3' |  
```

- **Promoter elements**
  - Portions of genomic DNA that act as transcription signals

- **Regulatory proteins**
  - Proteins bind to promoter elements
  - Positively or negatively regulates transcription
    - Enhances / inhibits RNA-polymerase

- **Open reading frames (ORF)**
  - Portion of DNA translated by ribosome to protein

- **Pseudo-genes**
  - Originally active gene
  - Rendered inactive due to mutations
Gene Structure – Prokaryotes

- **Features of prokaryotic genes**
  - High gene density (85% coding), no introns
  - Start with ATG, finish with TAA, TGA, TAG
  - Long open reading frames (ORF)
    - Usually > 180+ amino acids in length
  - Different composition (AT / GC ratio) in coding regions
  - Single RNA polymerase (from multiple proteins)
  - Promoter sequences in 5’ flanking region
    - E. Coli has 7 promoters located at –35 and –10 bases
      - $\sigma^{70} – TTGACA (-35) \& TATAAT (-10)$
      - $\sigma^{32} – TCTC?CCCTGAA (-35) \& CCCCAT?TA (-10)$
  - Shine-Dalgarno sequence (AGGAGGU) in 5’ UTR
    - Ribosome loading site
Gene Structure – Eukaryotes

- **Features of eukaryotic genes**
  - Low gene density (3% coding, 27% promoters / introns)
  - High variability in size / composition of genes
  - Three kinds of RNA polymerase (from 8-12 proteins)
  - Promoter sequences in 5' flanking region (may be distant)
    - RNA polymerase I −45 to +20 bases
    - RNA polymerase II far upstream to -25 bases
    - RNA polymerase III +50 to +100 bases
  - Different promoter sequence(s) for each gene
    - Example – TATA box (-25) in 70% of genes
  - Many regulatory proteins (12+ basal transcription factors)
    - Bind to transaction factor binding sites in specific order
    - Facilitate transcription by RNA polymerase

- **Intron Structure**
  - **Intron begins with GT**
  - **Intron ends with AG**

---

CMSC 838T – Lecture 10
Gene Structure – Eukaryotes

- **GC content**
  - CG dinucleotides *(CpG islands)*
    - Underrepresented by 80% in DNA
    - Generally found upstream of 5’ ends of genes
      - From −1500 to +500
    - Rarely found in non-coding regions
  - **Isochores** (long regions of DNA with uniform GC ratio)
    - 5 types of isochores in humans (39, 42, 46, 49, 54%)
    - Most genes in high GC isochores (20x ratio genes)
  - **Codon usage bias**
    - Organisms prefer certain triplet codons for amino acid
    - Distinguishes genes from random DNA sequences

- **Splicing**
  - Eight types of introns found
  - Most protein-coding gene introns conform to GU–AG rule
    - Additional splicing signals within intron (minimum 60 bps)
    - Average intron ~450 bp, most between 100 and 2000+ bps
    - 95% human genes with 1+ introns, some with 100+
    - **Alternative splicing** creates multiple proteins
Gene Structure – Alternative Splicing
Gene Structure – Alternative Splicing

Genomics Overview

- **Outline**
  - Molecular biology techniques
  - Sequencing
  - Assembly
  - Gene structure
  - Gene prediction
    - cDNA sequencing
    - EST clustering
    - Microarrays covering entire genome
    - Genetics in model organisms
    - Mutation rate comparisons (across & within species)
    - Computational gene finding
Gene Prediction – Motivation

- **Identifying genes is important**
  - Targets for expression microarrays
  - Producing proteins
- **The full gene (including 5’ and 3’ UTRs) is needed**
  - Avoiding misleading fragmentation / fusion artifacts
  - Understanding mRNA targeting and stability
  - Finding transcription factor binding sites
  - Understanding regulatory networks
- **Unreal (incorrectly labeled) genes can**
  - Mislead analysis of multiple sequence alignments
  - Distort protein classification systems and phylogenies
  - Misclassify other genes (since genes annotated by homology)

Gene Prediction – Computational Gene Finding

- **Annotation of Celera human genome assembly**
  - Small section of chromosome 3
  - Requires expert curation, very labor intensive
Gene Prediction – Methods

- **Methods for identifying genes**
  - cDNA sequencing
  - EST clustering
  - Microarrays covering entire genome
  - Genetics in model organisms
  - Mutation rate comparisons (across & within species)
  - Computational gene finding

Gene Prediction – cDNA Sequencing

- **Collecting cDNAs**
  - Extract mRNA from cells
  - Apply reverse transcriptase and a poly-U primer
  - Convert to cDNA starting at poly-A tail
  - Insert cDNA into vectors
  - Sequence read insert using primers on vector
  - If sequence looks to be new, sequence full cDNA

- **Artifacts and limitations are possible at each stage**
Gene Prediction – cDNA Problems & Solutions

- For rarely expressed genes little RNA is available
  - Normalize libraries
  - Use embryonic and exotic tissues as mRNA source

- Reverse transcriptase problems
  - Falls off before finishing (produces fragments)
    - Preferentially taking larger cDNAs
    - Normalizing only on 5’ ends (Soares)
  - High error rate, prone to small deletions
    - Compare cDNA to genomic DNA
    - Sequence multiple cDNA clones

Gene Prediction – cDNA / EST Problems

- cDNA includes introns, UTRs

![Diagram showing genomic DNA, hnRNA, problematic cDNA, and mature mRNA with exons and introns.](image)
**Gene Prediction – EST Clustering**

- **Simplified version of cDNA analysis**
  - Extract mRNA from cells
  - Apply reverse transcriptase
  - Convert to cDNA
  - Sequence fragment of cDNA from either 5' or 3' end

- **Result**
  - Sequence for only parts of cDNA
  - Called “expressed sequence tag” (EST)
  - Disadvantage
    - High error rates, partial cDNA
  - Advantage
    - Automated, high volume!

**EST example**

![Diagram of genomic DNA, mRNA, and ESTs](image)
Gene Prediction – EST Clustering

- **Clustering**
  - Build clusters of ESTs from the same gene
  - Can help identify gene

- **Simple approach**
  - Use pairwise comparisons to put ESTs into clusters
    - Compare all pairs of ESTs
    - Use fragment assembly software
  - Problems
    - ESTs from different individuals / strains of one species
    - Distinguishing between mutations and sequencing errors
    - Genomic & protein databases provide additional clues

Gene Prediction – EST Clustering Problem

- **Lack of Coverage**

```markdown
mRNA

ESTs

gaps
```
Gene Prediction – EST Clustering Problem

- **Duplicated genes**

  mRNA (from gene)        mRNA (from duplicated gene)

  ESTs

  high degree of similarity

Gene Prediction – EST Clustering

- **More efficient approach**
  - Initially, treat each EST as a cluster by itself
  - If two ESTs from two different clusters show significant overlap, merge the clusters
  - Use union-find data structure

  pass alignment test → merge → single cluster
Gene Prediction – EST Clustering

- **Quality of overlap**
  - Length of maximal common substring

- **Promising pairs**
  - Pairs with maximal common substring length \( \geq \psi \)
  - Find promising pairs on demand
  - Find promising pairs in decreasing order of quality
  - Can use generalized suffix tree

Gene Prediction – Whole Genome Microarrays

- **Microarray**
  - Technique for directly detecting cDNA
  - Based on hybridization to thousands of oligomers (short DNA sequences) at once
  - Can now cover non-repetitive portions of entire chromosomes

- **Observations**
  - Brute force, no homology required
  - Detect lower concentrations of mRNA than randomly sequencing EST
  - Rarely expressed genes may not stand out above background
  - Have to cope with cross-hybridization, other issues
Gene Prediction – Genetics in Model Organisms

♦ Approach
  - Zap yeast, plants, flies, mice with x-rays
  - Inbreed offspring and look for genetic defects

♦ Advantages
  - Works at DNA level, so expression level doesn't matter
  - Immediate hints of gene function
  - Discover gene interactions by breeding mutants

♦ Disadvantages
  - Finding mutated DNA may be slow & difficult
  - Essential genes can be hard to find
    - Reduced fertility in the inbreeding stage
  - Genes only needed in certain environments
    - Unable to detect all gene mutations

Gene Prediction – Mutation Rate Comparisons

♦ Approach
  - Compare mutation rates in genome
  - Compare across species & individuals
  - Look for highly conserved regions

♦ Motivation
  - Mutations occur randomly across genome, but...
  - Mutations in functional areas reduced by natural selection
  - Comparing DNA across species / individuals
  - Functional areas (genes, promoters) are more conserved
Gene Prediction – Mutation Rates Across Species

- **Mutation rates in beta-like globin genes**
  - Comparing human, mouse, rabbit, cow
  - Nucleotide substitution rate/site/billion years

![Graph showing mutation rates across different regions of the gene](image)

Prediction – Mutation Rate Across Individuals

- **% conserved positions in human genes**
  - 3165 mappings of human RefSeq mRNAs to the genome
  - Sampling 200 evenly spaced bases in different gene regions
  - Peaks of conservation at transitions between regions
    - Start/end codons, GU-AG splicing signals, etc...

![Graph showing conservation rates across different regions of the gene](image)
Gene Prediction – Computational Gene Finding

- **Computational gene finding**
  - Identify genes in DNA sequences using computer analysis
  - Look for gene features & compare with EST / protein databases
  - Discover exons, introns, promoters, etc...
  - Simple for prokaryotes (bacteria), difficult for eukaryotes

Gene Prediction – Computational Gene Finding

- **Approaches**
  - Homology based
    - Search against translated protein sequences
  - Direct analysis methods (content-based & site-based)
    - Grammar based
    - Neural networks
    - Hidden markov models (HMMs)
  - Composite methods (combines direct analysis & homology)
    - EST data
    - Gene homology
    - Multiple specie genomes
Gene Prediction – Homology Based

- **Approach (Procrustes 1996)**
  - Takes protein sequence as input
  - Uses dynamic programming spliced alignment algorithm
  - Coding regions must be fairly well conserved
- **Result**
  - Find best exons matching protein
  - Incomplete gene structure
    - No promoters, etc…
- Or just use BLAST…

Gene Prediction – Direct Analysis

- **Content based**
  - Relies on overall (bulk) properties of sequence
    - Codon frequency
    - Periodicity of repeats
    - Compositional complexity
- **Site based**
  - Focus on presence / absence of specific patterns
    - Binding sites for transcription factors (promoters)
    - Donor & acceptor splice sites
    - Start & stop codons
Gene Prediction – Grammar Based

**Approach** *(GeneLang, 1994)*
- Encode rules for gene features as context-free grammar
- Generate parser for grammar
- Attempt to syntactically recognize target sequences

Gene Prediction – Neural Networks

**Approach** *(GRAIL, 1991)*
- Fragment sequence into 4096 6-base hexamers
- Compute probability of hexamers at each gene location
- Recognize gene factors
  - Codon usage
  - Base composition
  - Splice site characteristics
  - PolyA signals
  - Di-, tri-, hexa-nucleotide frequencies
  - Translation signals
  - Transcription signals
  - Size distributions
Gene Prediction – Hidden Markov Models

**Approach (GenScan, 1997)**
- Multiple probabilistic models for different gene structures
- Analyze sequence to assign probabilities for exons, etc…
- Recognize gene factors using 5th-order Markov model
  - GC content
  - # of genes
  - Exon / intron
  - Mean length of
    - Exon / intron
    - Transcript
    - Inter-gene region
  - Signal models for
    - Coding
    - Splicing
    - Etc…

Gene Prediction – Problems

**Homology based methods**
- Can only find genes we already know
  - By searching comparisons to known protein
- Does not detect promoters, UTRs

**Direct analysis methods**
- Tend to overpredict genes
  - Many false positives
- Introns are vast, GT/AG splice signals are small
  - Coding signal is stronger than start / stop signal
  - Difficult to predict splice sites
  - Gene fragmentation / fusion often result

CMSC 838T – Lecture 10
Gene Prediction – Composite Methods

◆ Composite methods
  - Combines homology & direct analysis methods
  - Use bioinformatic databases to correct / enhance predictions

◆ Approaches
  - Use EST info to constrain prediction
    • Genie (Generalized HMM + EST alignment)
  - Use protein homology info to constrain prediction
    • GenomeScan
  - Use cross-species genomic alignment to improve prediction
    • Twinscan, SLAM, SGP

Gene Prediction – Accuracy

◆ GASP1
  - Genome Annotation Assessment Project, 1999
  - Experimentally compare computational gene finding

◆ Evaluation measures
  - True Positive (TP), False Positive (FP), False Negative (FN)
  - Sensitivity – % found = TP / (TP + FN)
  - Specificity – % correct = TP / (TP + FP)
  - Missed / wrong exons
  - Missed / wrong genes
  - Split / joined genes
    \[\text{Incorrect boundaries}\]
Gene Prediction – GASP1 Results

- **Results**
  - Genie (constrained w/ EST database) a top performer
  - Incorrectly split genes more problematic than joined genes
  - Including homology does not always yield improvement
  - HMM seems to be best approach
  - Poor prediction of promoters
  - Computational gene finding not sufficiently accurate

<table>
<thead>
<tr>
<th></th>
<th>Bases</th>
<th>Exons</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensitivity (% found)</strong></td>
<td>97%</td>
<td>77%</td>
<td>65%</td>
</tr>
<tr>
<td><strong>Specificity (% correct)</strong></td>
<td>91%</td>
<td>55%</td>
<td>38%</td>
</tr>
<tr>
<td>Missed</td>
<td></td>
<td>5%</td>
<td>11%</td>
</tr>
<tr>
<td>Wrong</td>
<td></td>
<td>20%</td>
<td>42%</td>
</tr>
</tbody>
</table>

- Genes
- Exons
- Bases

Genomics – Summary

- **Sequencing & assembly**
  - Reasonably well understood, quality solutions available
  - Assembly computationally intensive for large sequences

- **Gene prediction**
  - Many laboratory & computational techniques
  - Major effort for bioinformatics researchers
  - Computational techniques insufficiently precise