What is BLAST?

- Basic BLAST search
  - What is BLAST?
  - The framework of BLAST
  - Different BLAST programs
  - BLAST databases you can search
  - Where can I run BLAST?

- BLAST stands for
  - Basic Local Alignment Search Tool

- Why BLAST is popular?
  - Good balance of sensitivity and speed
  - Reliable
  - Flexible

- Produce local alignments: short significant stretches of similarity, irrespective of where they are in the sequence

BLAST Programs

The most common BLAST search include five programs:

<table>
<thead>
<tr>
<th>Program</th>
<th>Database (Subject)</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTN</td>
<td>Nucleotide</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>BLASTP</td>
<td>Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>BLASTX</td>
<td>Protein</td>
<td>Nt. → Protein</td>
</tr>
<tr>
<td>TBLASTN</td>
<td>Nt. → Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>TBLASTX</td>
<td>Nt. → Protein</td>
<td>Nt. → Protein</td>
</tr>
</tbody>
</table>

BLASTN

- BLASTN
  - The query is a nucleotide sequence
  - The database is a nucleotide database
  - No conversion is done on the query or database

- DNA :: DNA homology
  - Mapping oligos to a genome
  - Annotating genomic DNA with transcriptome data from ESTs and RNA-Seq
  - Annotating untranslated regions

BLASTP

- BLASTP
  - The query is an amino acid sequence
  - The database is an amino acid database
  - No conversion is done on the query or database

- Protein :: Protein homology
  - Protein function exploration
  - Novel gene → make parameters more sensitive

BLASTX

- BLASTX
  - The query is a nucleotide sequence
  - The database is an amino acid database
  - All six reading frames are translated on the query and used to search the database

- Coding nucleotide seq :: Protein homology
  - Gene finding in genomic DNA
  - Annotating ESTs and transcripts assembled from RNA-Seq data
**TBLASTN**

- **TBLASTN**
  - The query is an amino sequence
  - The database is a nucleotide database
  - All six frames are translated in the database and searched with the protein sequence
- **Protein :: Coding nucleotide DB homology**
  - Mapping a protein to a genome
  - Mining ESTs and RNA-Seq data for protein similarities

**TBLASTX**

- **TBLASTX**
  - The query is a nucleotide sequence
  - The database is a nucleotide database
  - All six frames are translated on the query and on the database
- **Coding :: Coding homology**
  - Searching distantly related species
  - Sensitive but computationally expensive

**BLAST output**

1. List of sequences with scores
   - Raw score
     - Higher is better
     - Depends on aligned length
   - Expect Value (E-value)
     - Smaller is better
     - Independent of length and database size
2. List of alignments

**The Databases (1)**

- GenBank NR (protein and nucleotide versions)
  - Non-redundant large databases (compile and remove duplicates)
  - Anyone can submit, you can call your sequence anything
  - Low quality; names can be meaningless
- Transcriptome Shotgun Assembly (TSA) Database
  - Transcripts assembled from overlapping ESTs and RNA-Seq reads
  - Most of the sequences have no annotations

**The Databases (2)**

- UniProt/Swiss-Prot
  - Curated from literature
  - **REAL** proteins; **REAL** functions; small;
- Genomic databases
  - Human, Mouse, *Drosophila*, Arabidopsis, etc.
  - NCBI, species-specific web pages

**Where Can I run BLAST?**

1. NCBI **BLAST** web service
2. EBI **BLAST** web service
   - [https://www.ebi.ac.uk/Tools/sss/ncbiblast/](https://www.ebi.ac.uk/Tools/sss/ncbiblast/)
3. FlyBase **BLAST**
   - [https://flybase.org/blast/](https://flybase.org/blast/)
   - *Drosophila melanogaster*