

Beyond The Genome 2012  
Informatics Challenge  
*One Possible Approach*

Michael Schatz

James Taylor

David Dooling

# This Approach

- Uses the command line on a GNU/Linux x64 system
- Uses `fixed-width` font for things you should type (do not type the prompt, \$)
- Does not use full path for all input files

# Download Challenge

```
$ mkdir btg
```

```
$ cd btg
```

```
$ wget http://goo.gl/3Zwkk
```

```
$ tar -x -z -f BeyondTheGenome2012InformaticsChallenge.tar.gz
```

# Assemble Reads

- Download velvet
  - \$ wget [http://www.ebi.ac.uk/~zerbino/velvet/velvet\\_1.2.07.tgz](http://www.ebi.ac.uk/~zerbino/velvet/velvet_1.2.07.tgz)
  - \$ tar -x -z -f velvet\_1.2.07.tar.gz
  - \$ cd velvet\_1.2.07
- Compile velvet allowing larger k-mers
  - \$ make MAXKMERLENGTH=51
- Run velveth
  - Try a k-mer length of 45 (must be odd!)
  - Use MiSeq-like interleaved reads
  - \$ ./velveth velvet-45 45 -fastq -shortPaired \  
../BeyondTheGenome2012InformaticsChallenge/i2x250f700.fq
- Run velvetg
  - Discard error-derived low-coverage k-mers (-cov\_cutoff)
  - Discard repeat-derived high-coverage k-mers (-max\_coverage)
  - Expected k-mer coverage is  $C \cdot (L - k + 1) / L$ , C is base coverage, L is read length, and k is k-mer size (HINT: C is 100 for this data set)
  - \$ ./velvetg velvet-45 -ins\_length 700 -cov\_cutoff 20 \  
-exp\_cov 82.4 -max\_coverage 150

# BLAST Contigs

- Use NCBI BLAST to blast contigs.fa from Velvet output directory against refseq\_genomic database
- [http://blast.ncbi.nlm.nih.gov/Blast.cgi?  
PROGRAM=blastn&BLAST\\_SPEC=WGS&BLAST\\_PRO  
GRAMS=megaBlast&PAGE\\_TYPE=BlastSearch](http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_SPEC=WGS&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch)
- Use megablast

# NCBI BLAST

**BLAST®** Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI BLAST/blastn suite **Standard Nucleotide BLAST**

blastn blastx blastz blastn blastx

Enter Query Sequence BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter accession number(s), gi(s), or FASTA sequence(s)  Clear Query subrange

From

To

Or, upload file  config.txt

Job Title  Enter a descriptive title for your BLAST search.

Align two or more sequences

Choose Search Set

Database  Human genomic + transcript  Mouse genomic + transcript  Others (nr etc.):

+

Organism Optional   Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional  Models (XM/XP)  Uncultured/environmental sample sequences

Entrez Query Optional  Enter an Entrez query to limit search.

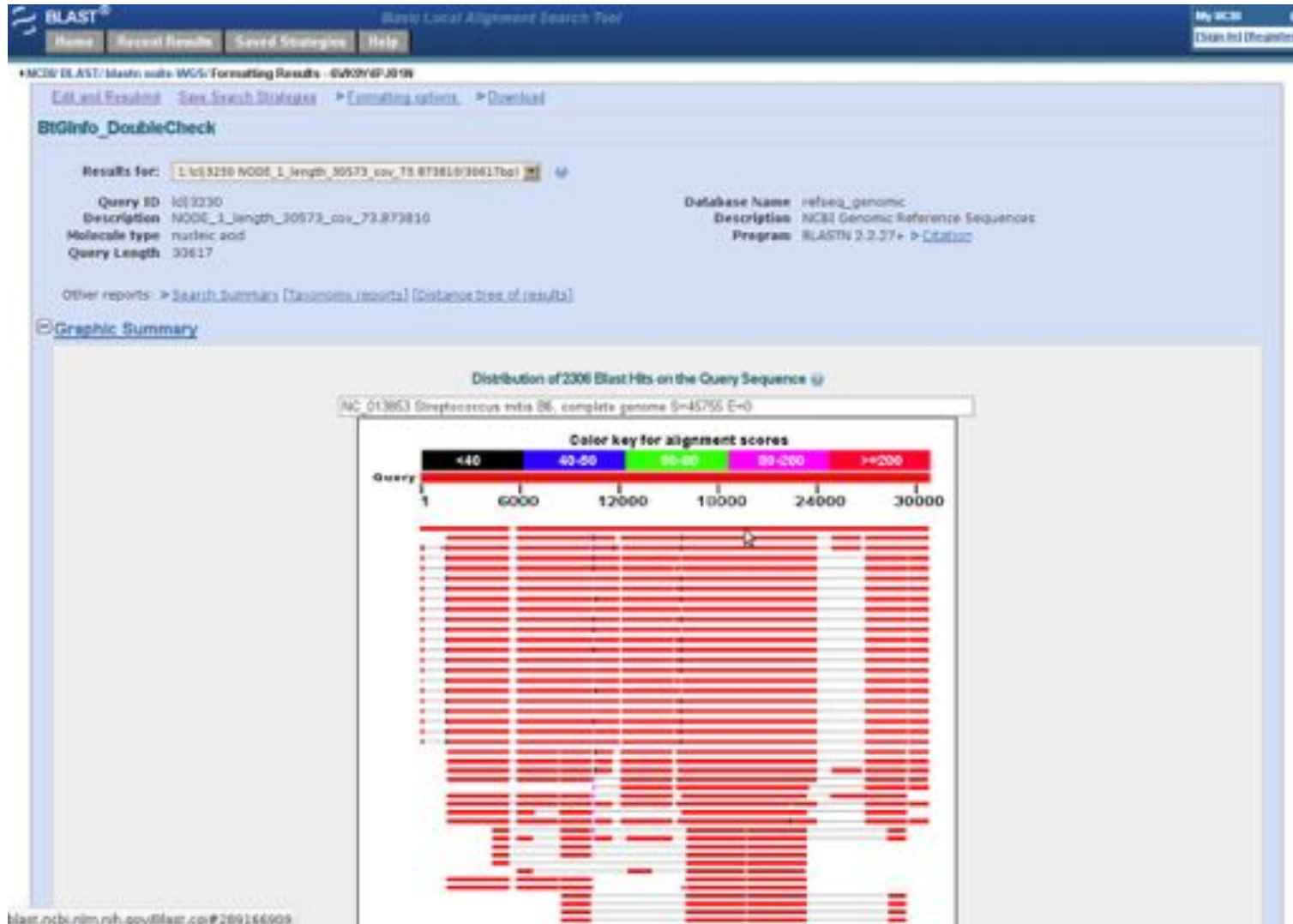
Program Selection

Optimize for  Highly similar sequences (megablast)

More dissimilar sequences (discontiguous megablast)

Somewhat similar sequences (blastn)

# BLAST Results



Click on top hit

# Top Hit

```
> refseq\_013853.11  Streptococcus mitis B6, complete genome
Length=2146611
Show report for NC_013853
Sort alignments for this subject sequence by:
E value Score Percent identity
Query start position Subject start position

Features in this part of subject sequence:
gasga inducible DnaF protein
DcaA protein

Score = 45755 bits (24777), Expect = 0.0
Identities = 24785/24789 (99%), Gaps = 0/24789 (0%)
Strand=Plus/Minus

Query 5829      ACCCTGCTGTCCCTGAGARGAAGTGACTCGGGTTGTAACACCAGATGTTAATCAATTG 5888
                ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 1924958    ACCCTGTTGTCCCTGAGARGAAGTGACTCGGGTTGTAACACCAGATGTTAATCAATTG 1924899

Query 5889      ATGAGTGGCAGCCTTCTAARCGTCCGAGARGAACAAACAAACCAAGACCCCTCTTCTACA 5948
                ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 1924898    ATGAGTGGCAGCCTTCTAARCGTCCGAGARGAACAAACAAACCAAGACCCCTCTTCTACA 1924839

Query 5949      TGGCAGATGGTGAGCTTGTTCAGCCCTTCTAGTTGGAAATGACCAGCTCAATGARGTTA 6008
                ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 1924898    TGGCAGATGGTGAGCTTGTTCAGCCCTTCTAGTTGGAAATGACCAGCTCAATGARGTTA 1924779

Query 6009      AGTTGAAGAACCCTTGGGAGCAGATTTTTTTTGGCGTTGCGAGCGAAGAAGAGTGGCA 6068
                ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 1924778    AGTTGAAGAACCCTTGGGAGCAGATTTTTTTTGGCGTTGCGAGCGAAGAAGAGTGGCA 1924719

Query 6069      GTGTTGTTCAAGCAGGATTTGCTTCACTTGGACAGTTGCTTTGCCGGAGARTGTTAAAA 6128
                ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 1924718    GTGTTGTTCAAGCAGGATTTGCTTCACTTGGACAGTTGCTTTGCCGGAGARTGTTAAAA 1924659

Query 6129      TCATTGCCGACCGTARGGTGCAAGATGTCCATAATGCAGTTGTCGGTGCTAOCGARGATG 6188
                ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 1924658    TCATTGCCGACCGTARGGTGCAAGATGTCCATAATGCAGTTGTCGGTGCTAOCGARGATG 1924599

Query 6189      GCTACCCTTGACTGGTGTGAATCCAGGTCGTGACTTTACTGCAGAATATGTGGATATCC 6248
                ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
Sbjct 1924598    GCTACCCTTGACTGGTGTGAATCCAGGTCGTGACTTTACTGCAGAATATGTGGATATCC 1924539

Query 6249      GTGAAGTTGTTGAGGGTGAATTTCTCCAGACCGAAGGTTGCTCTTAACCTTGGCGGTG 6308
                ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
```

Click on reference name



# Streptococcus mitis

[http://en.wikipedia.org/wiki/Streptococcus\\_mitis](http://en.wikipedia.org/wiki/Streptococcus_mitis)

Streptococcus mitis is a mesophilic alpha-hemolytic species of Streptococcus that inhabits the human mouth. It is a Gram positive, coccus, facultative anaerobe and catalase negative. It can cause endocarditis. It has been widely reported that this organism survived for over two years on the Surveyor 3 probe on the moon; but some NASA scientists suggest this may be a result of contamination during or after return of Surveyor parts to Earth, as the person assembling the camera may have sneezed.

# Download Reference

The screenshot shows the NCBI GenBank interface for the Streptococcus mitis B6 complete genome (NC\_013853.1). A 'Send' dropdown menu is open, displaying various download options. The 'Format' section is expanded to show 'FASTA' as the selected option. The background shows the sequence details, including the locus information, definition, accession number, version, and keywords.

**Streptococcus mitis B6, complete genome**  
NCBI Reference Sequence: NC\_013853.1  
[FASTA](#) [Graphics](#)

**LOCUS** NC\_013853 2146611 bp DNA circular BCT 25-JAN-2012  
**DEFINITION** Streptococcus mitis B6, complete genome.  
**ACCESSION** NC\_013853  
**VERSION** NC\_013853.1 GI:289166909  
**DBLINK** Project: [ds002](#)  
BioProject: [BXJN450097](#)

**KEYWORDS** complete genome.  
**SOURCE** Streptococcus mitis B6  
**ORGANISM** [Streptococcus\\_mitis\\_B6](#)  
Bacteria; Firmicutes; Lactobacillales; Streptococcaceae; Streptococcus.

**REFERENCE** 1 (bases 1 to 2146611)  
**AUTHORS** Denapate,D., Bruckner,R., Nuhn,M., Reichmann,P., Henrich,B., Maurer,P., Schahle,Y., Selbmann,P., Zimmermann,W., Warbut,R. and Hakenbeck,R.  
**TITLE** The genome of Streptococcus mitis B6 -what is a commensal?  
**JOURNAL** PLoS ONE 5 (2), E9426 (2010)  
**PUBMED** [20185536](#)  
**REMARK** Publication Status: Online-Only

**REFERENCE** 2 (bases 1 to 2146611)  
**CONSTRM** NCBI Genome Project  
**TITLE** Direct Submission  
**JOURNAL** Submitted (17-FEB-2010) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA

**REFERENCE** 3 (bases 1 to 2146611)

**Send**  Complete Record  
 Coding Sequences  
 Gene Features

**Choose Destination**  
 File  Clipboard  
 Collections  Analysis Tool

Download 1 items  
**Format**  
FASTA  
Summary  
GenBank  
GenBank (Full)  
FASTA  
ASN.1  
XML  
FASDSeq XML  
TinySeq XML  
Feature Table  
Accession List  
GI List

# Find Insert

- Download and compile MUMmer
  - <http://sourceforge.net/projects/mummer/files/latest/download?source=files>
  - \$ tar -x -z -f MUMmer3.23.tar.gz
  - \$ cd MUMmer3.23
  - \$ make
- Align contigs against reference
  - \$ ./nucmer -maxmatch Streptococcus\_mitis\_B6.fasta \contigs.fa
- Identify insert
  - \$ delta-filter -q out.delta > out.delta.q
  - \$ show-coords -qclo out.delta.q
  - Look for a single contig that has two non-overlapping mappings to the reference with a span on the contig between the two parts of the contig that map
  - The bases in the span of the contig that do not map are the insert

# Extract the Sequence

- Download and compile samtools
  - <http://sourceforge.net/projects/samtools/files/samtools/0.1.18/samtools-0.1.18.tar.bz2/download>
  - \$ tar -x -j -f samtools-0.1.18.tar.bz2
  - \$ cd samtools-0.1.18
  - \$ make
- Extract sequence and decode it
  - Insert is bases 5413-5836 (inclusive) on NODE\_1
  - Assembly reverse complemented the DNA
  - \$ ./samtools faidx contigs.fa \  
NODE\_1\_length\_30573\_cov\_73.873810:5413-5836 |  
perl dna-encode --decode --reverse-complement