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
$ 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
```

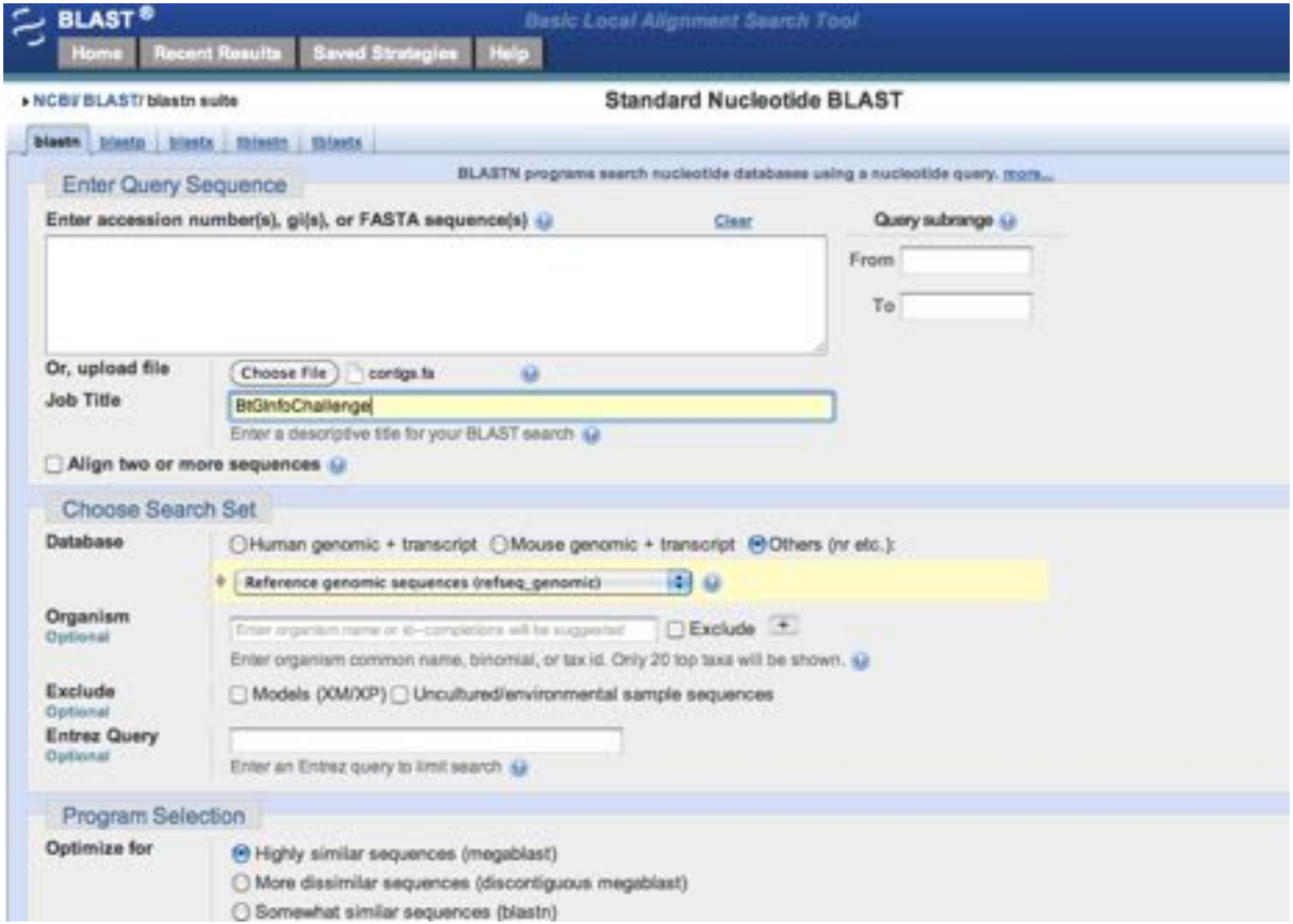
- Run velvetg
  - Discard error-derived low-coverage k-mers (-cov_cutoff)
  - Discard repeat-derived high-coverage k-mers (-max_coverage)
  - Expected k-mer coverage is C*(L-k+1)/L, C is base coverage, L is read length, and k is k-mer size (HINT: C is 100 for this data set)

```
$ ./velvetg velvet–45 –ins_length 700 –cov_cutoff 20 \
    –exp_cov 82.4 –max_coverage 150
```

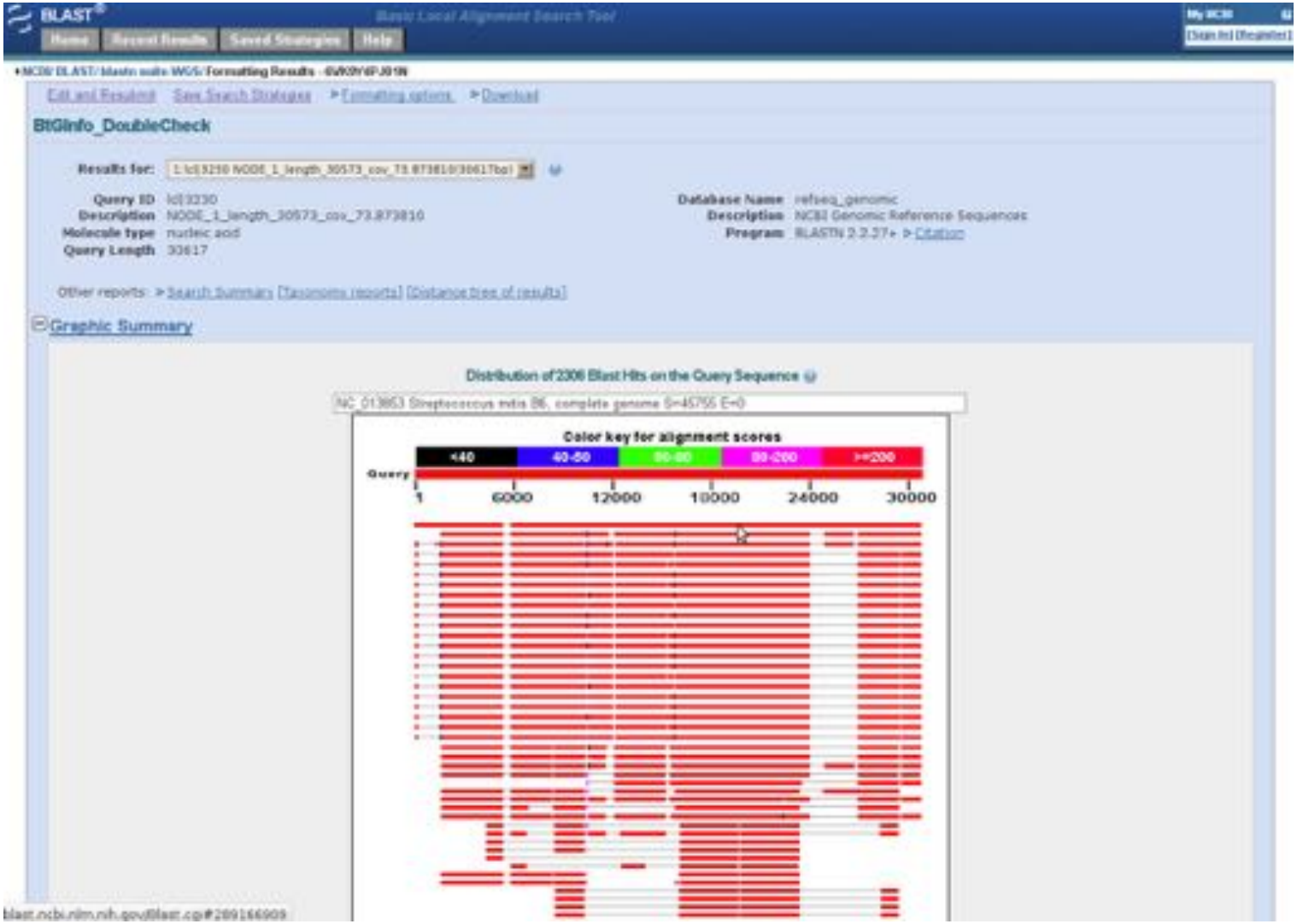
BLAST Contigs

- 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_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
BLASTN programs search nucleotide databases using a nucleotide query. more...
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Choose File
contigs.fa
Job Title
BtGInfoChallenge
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.)
Reference genomic sequences (refseq_genomic)
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

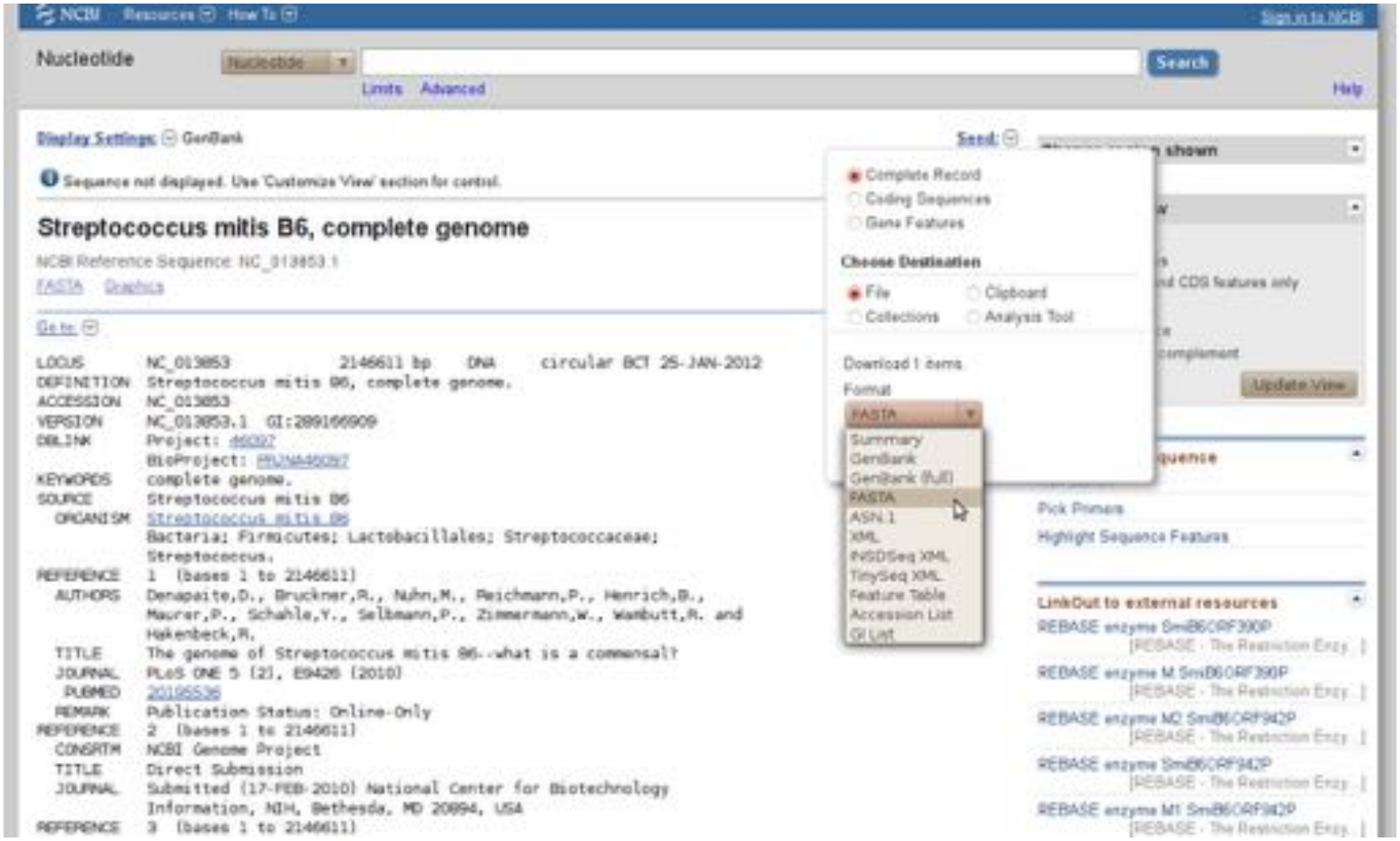
Click on reference name

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



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
```