

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 $C \cdot (L - k + 1) / L$, 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_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch
- Use megablast

NCBI BLAST

 **BLAST®** *Basic Local Alignment Search Tool*

HomeRecent ResultsSaved StrategiesHelp

NCBI/ BLAST/ blastn suiteStandard Nucleotide BLAST

blastnblastpblastxtblastntblastx

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

Choose File

 contigs.fa 

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.):

◆ 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

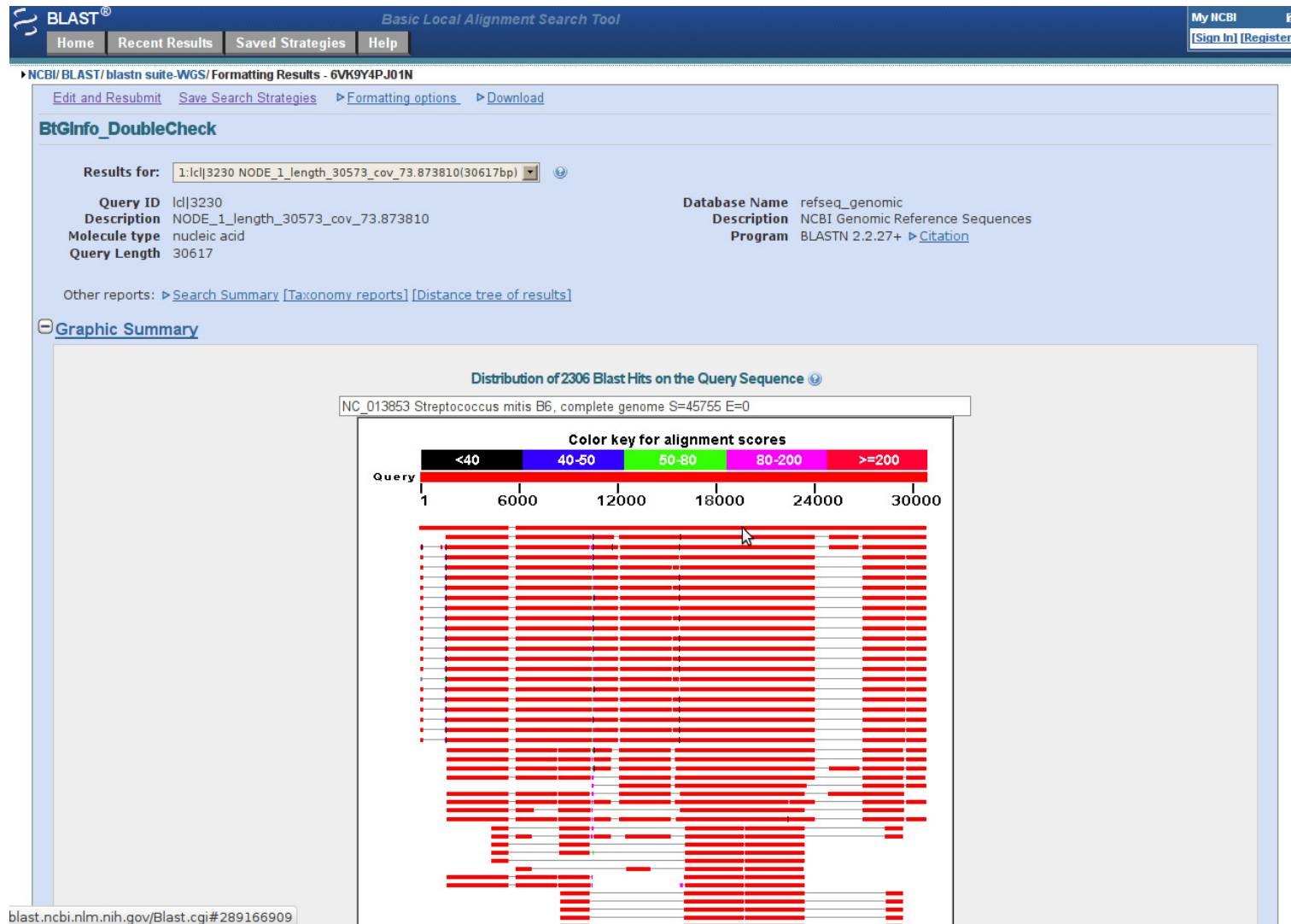
Optimize for

☒ Highly similar sequences (megablast)

☐ More dissimilar sequences (discontiguous megablast)

☐ Somewhat similar sequences (blastn)

BLAST Results



Click on top hit

Top Hit

> [ref|NC_013853.1|](#)  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:
[damage inducible DinF protein](#)
[RecA protein](#)

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

Query	5829	ACCCTGCTGTCGCTGAAGAAGAAGTGA	CGGGTTGTAACACCAGATGTTAAATCAATTG	5888
Sbjct	1924958	ACCGTGTTGTCGCTGAAGAAGAAGTGA	CGGGTTGTAACACCAGATGTTAAATCAATTG	1924899
Query	5889	ATGAAGTGGCAGCCTTCTAAACGTCCCAGAAGAACAACAATCAAGACCCTCTTCTACA		5948
Sbjct	1924898	ATGAAGTGGCAGCCTTCTAAACGTCCCAGAAGAACAACAATCAAGACCCTCTTCTACA		1924839
Query	5949	TGGCAGATGGTGAGCTTGTTCAGCCCTTCTAGTTGGAAATGACCAGCTCAATGAAGTTA		6008
Sbjct	1924838	TGGCAGATGGTGAGCTTGTTCAGCCCTTCTAGTTGGAAATGACCAGCTCAATGAAGTTA		1924779
Query	6009	AGTTGAAGAACCACTTGGGAGCAGATttttttGACGTTGCGAGCGAAGAAGAAGTGGCAA		6068
Sbjct	1924778	AGTTGAAGAACCACTTGGGAGCAGATttttttGACGTTGCGAGCGAAGAAGAAGTGGCAA		1924719
Query	6069	GTGTTGTTCAAGCAGGATTTGGTTCACCTGGACCAGTTGGTTTGCCGGAGAATGTTAAAA		6128
Sbjct	1924718	GTGTTGTTCAAGCAGGATTTGGTTCACCTGGACCAGTTGGTTTGCCGGAGAATGTTAAAA		1924659
Query	6129	TCATTGCCGACCGTAAGGTGCAAGATGTCCATAATGCAGTTGTCGGTGCTAACGAAGATG		6188
Sbjct	1924658	TCATTGCCGACCGTAAGGTGCAAGATGTCCATAATGCAGTTGTCGGTGCTAACGAAGATG		1924599
Query	6189	GCTACCACTTGACTGGTGTGAATCCAGGTCGTGACTTTACTGCAGAATATGTGGATATCC		6248
Sbjct	1924598	GCTACCACTTGACTGGTGTGAATCCAGGTCGTGACTTTACTGCAGAATATGTGGATATCC		1924539
Query	6249	GTGAAGTTCGTGAGGGTGAAATTTCTCCAGACGGACAAGGTGTCCTTAACCTTGCGCGTG		6308

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

The screenshot shows the NCBI GenBank interface for the Streptococcus mitis B6 complete genome (NC_013853.1). A dropdown menu is open, showing options for downloading the complete record, coding sequences, or gene features. The 'File' option is selected under 'Choose Destination'. The 'Format' dropdown is set to 'FASTA'. The 'Download 1 items.' section shows the 'FASTA' format selected. The 'LinkOut to external resources' section lists various external links, including REBASE enzyme links.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Limits Advanced Help

Display Settings: GenBank

Sequence not displayed. Use 'Customize View' section for control.

Streptococcus mitis B6, complete genome

NCBI Reference Sequence: NC_013853.1

[FASTA](#) [Graphics](#)

Go to:

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: [46097](#)
BioProject: [PRJNA46097](#)

KEYWORDS complete genome.

SOURCE Streptococcus mitis B6

ORGANISM [Streptococcus mitis B6](#)
Bacteria; Firmicutes; Lactobacillales; Streptococcaceae;
Streptococcus.

REFERENCE 1 (bases 1 to 2146611)

AUTHORS Denapaita,D., Bruckner,R., Nuhn,M., Reichmann,P., Henrich,B.,
Maurer,P., Schahle,Y., Selbmann,P., Zimmermann,W., Wambutt,R. and
Hakenbeck,R.

TITLE The genome of Streptococcus mitis B6--what is a commensal?

JOURNAL PLoS ONE 5 (2), E9426 (2010)

PUBMED [20195536](#)

REMARK Publication Status: Online-Only

REFERENCE 2 (bases 1 to 2146611)

CONSRM 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)

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
INSDSeq XML
TinySeq XML
Feature Table
Accession List
GI List

LinkOut to external resources
REBASE enzyme SmiB6ORF390P
[REBASE - The Restriction Enzy...]
REBASE enzyme M.SmiB6ORF390P
[REBASE - The Restriction Enzy...]
REBASE enzyme M2.SmiB6ORF942P
[REBASE - The Restriction Enzy...]
REBASE enzyme SmiB6ORF942P
[REBASE - The Restriction Enzy...]
REBASE enzyme M1.SmiB6ORF942P
[REBASE - The Restriction Enzy...]

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
```