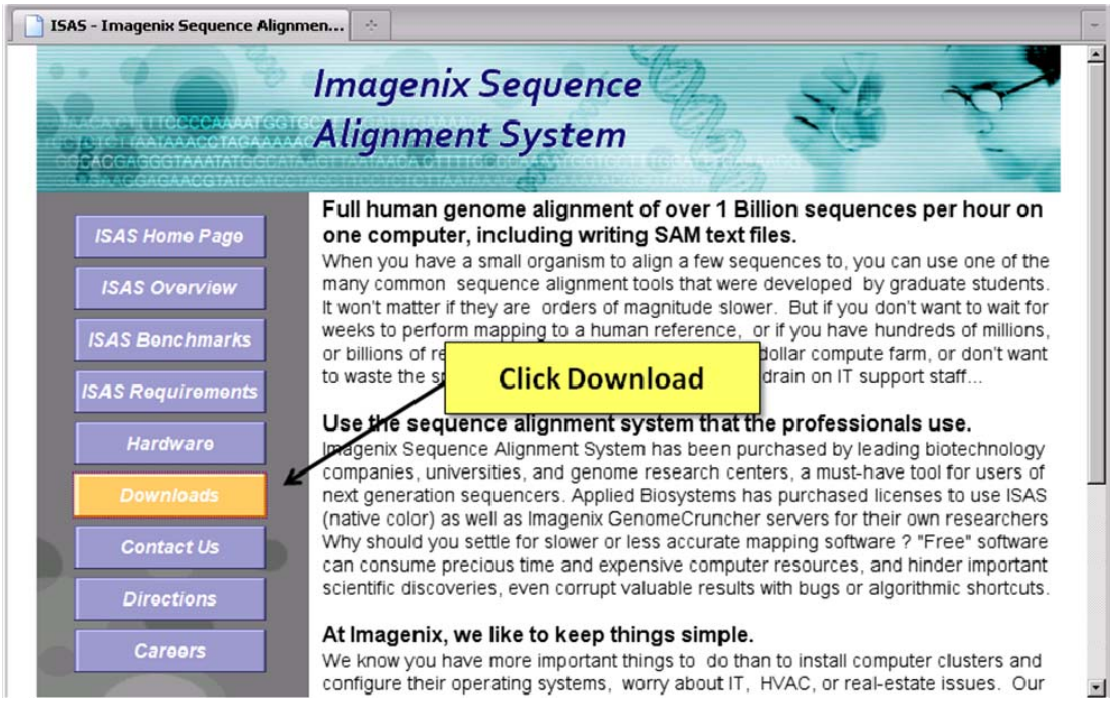
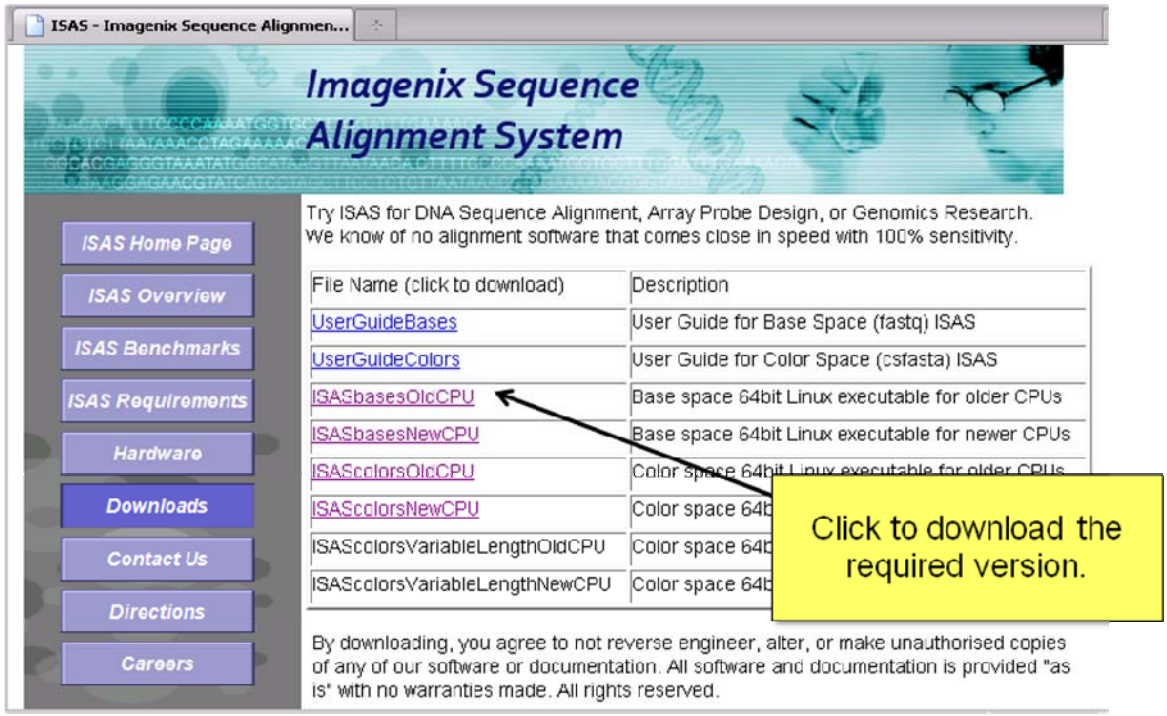
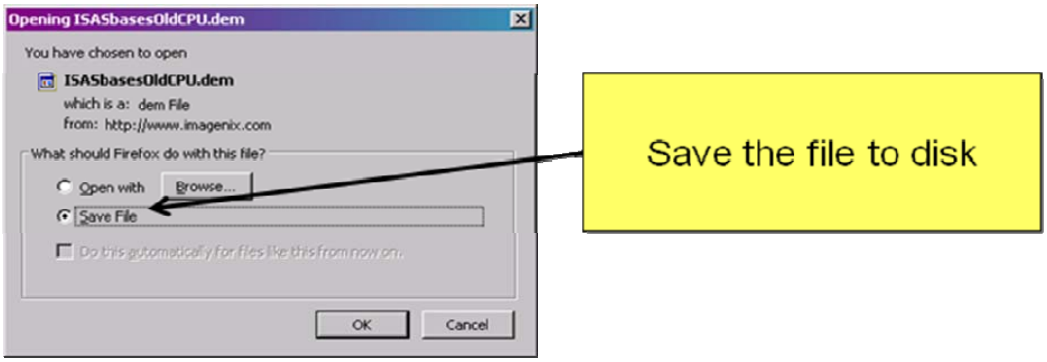
Supplementary File 2

This document provides a tutorial on how to use the ISAS demonstration software to generate custom uniqueomes for genomes of interest, using the example of ce6 (for *C. elegans*). The ISAS demonstration software can be accessed from the Imagenix website at http://www.imagenix.com/. Please see the website for details on requirements, specifications, and the full user manual.

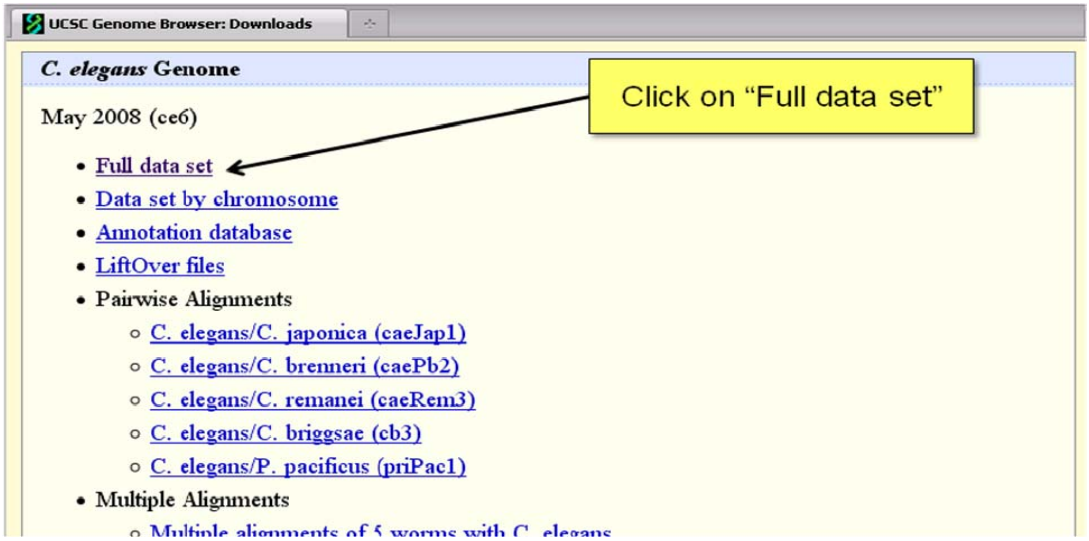
Downloading the Required Files

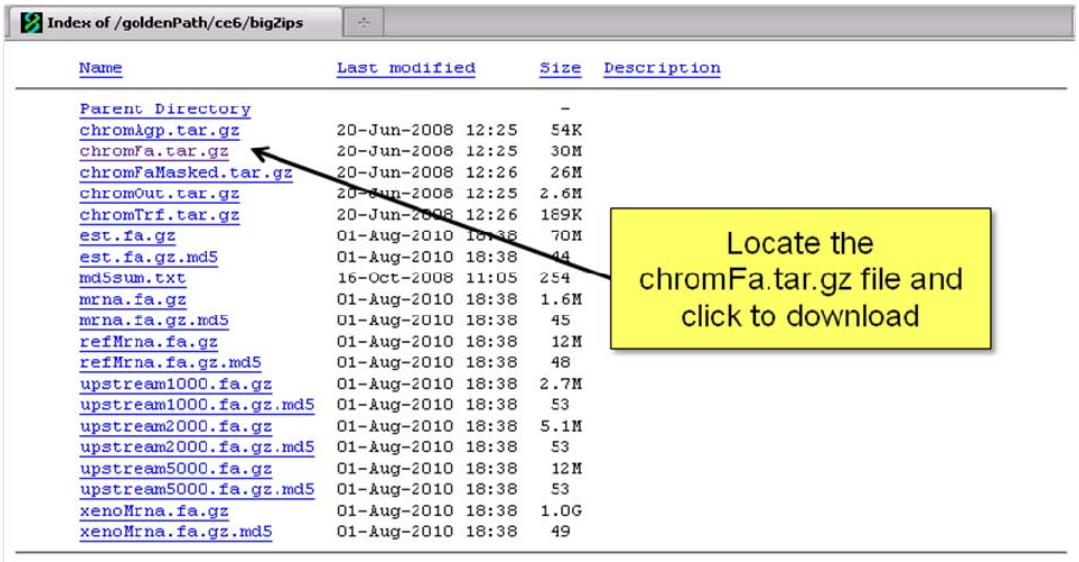






You will also need to download the genome of interest from a site like the UCSC Genome Browser (http:// genome.ucsc.edu/downloads.html). Click on the genome of interest. eg. *C. elegans*





Finally, you will also need to download the perl scripts for processing and converting the ISAS output to bigWig or bigBED files. You can access these scripts from http://grimmond.imb.uq.edu.au/unqiueome/.

Installing the ISAS demonstration software and associated files.

1. Create a directory to run the ISAS software from. Change to that directory. Commands have been highlighted in green for visualization purposes.

```
[a.user@foo]$ mkdir /data/ncloonan/ISASdemo/
[a.user@foo /]$ cd /data/ncloonan/ISASdemo/
[a.user@foo ISASdemo]$
```

2. Move all of the required files (the demonstration software, the scripts, and the genome fasta files) to that directory.

```
[a.user@foo ISASdemo]$ mv /data/downloads/ISAScolorsOldCPU.dem .
[a.user@foo ISASdemo]$ mv /data/downloads/uniqueome_scripts.tar.gz .
[a.user@foo ISASdemo]$ mv /data/downloads/chromFa.tar.gz .
```

3. Unpack the scripts from their archive

```
[a.user@foo ISASdemo]$ tar -xzf uniqueome_scripts.tar.gz
```

4. Create a directory for your genome (eg. ce6 for *C. elegans*). Within that directory create a subfolder called “reference”. Move and unpack the fasta files into that directory.

```
[a.user@foo ISASdemo]$ mkdir ce6
[a.user@foo ISASdemo]$ mkdir ce6/reference
[a.user@foo ISASdemo]$ mv chromFa.tar.gz ce6/reference/
[a.user@foo ISASdemo]$ cd ce6/reference/
[a.user@foo reference]$ tar -xzf chromFa.tar.gz
```

5. ISAS requires write permissions enabled for the directory the demo sits in, and the genome directory as well. Use chmod to alter the permissions.

```
[a.user@foo ISASdemo]$ chmod 777 /data/ncloonan/ISASdemo/
[a.user@foo ISASdemo]$ chmod 777 /data/ncloonan/ISASdemo/ce6/
```

6. ISAS only accepts fasta files that are named chr[integer].fa (eg. chr10.fa). If necessary, rename the files to conform to this requirement, then run the script called **document_filename_changes.pl** to generate a renamed_chromosomes.txt file for deconvolving these chromosome names later.

```
[a.user@foo reference]$ ls –lha
total 128M
drwxr-xr-x 2 a.user A-Group  4.0K Aug  9 16:50 .
drwxr-xr-x 3 a.user A-Group  4.0K Aug  9 16:49 ..
-rw-r--r-- 1 a.user A-Group  15M Jun 21  2008 chrI.fa
-rw-r--r-- 1 a.user A-Group  15M Jun 21  2008 chrII.fa
-rw-r--r-- 1 a.user A-Group  14M Jun 21  2008 chrIII.fa
-rw-r--r-- 1 a.user A-Group  18M Jun 21  2008 chrIV.fa
-rw-r--r-- 1 a.user A-Group  14K Jun 21  2008 chrM.fa
-rw-r--r-- 1 a.user A-Group  21M Jun 21  2008 chrV.fa
-rw-r--r-- 1 a.user A-Group  18M Jun 21  2008 chrX.fa
[a.user@foo reference]$ mv chrI.fa chr1.fa
[a.user@foo reference]$ mv chrII.fa chr2.fa
[a.user@foo reference]$ mv chrIII.fa chr3.fa
[a.user@foo reference]$ mv chrIV.fa chr4.fa
[a.user@foo reference]$ mv chrV.fa chr5.fa
[a.user@foo reference]$ mv chrM.fa chr6.fa
[a.user@foo reference]$ mv chrX.fa chr7.fa
[a.user@foo reference]$ ls -lha
total 128M
drwxr-xr-x 2 a.user A-Group  4.0K Aug  9 21:16 .
drwxr-xr-x 3 a.user A-Group  4.0K Aug  9 16:49 ..
-rw-r--r-- 1 a.user A-Group  15M Jun 21  2008 chr1.fa
-rw-r--r-- 1 a.user A-Group  15M Jun 21  2008 chr2.fa
-rw-r--r-- 1 a.user A-Group  14M Jun 21  2008 chr3.fa
-rw-r--r-- 1 a.user A-Group  18M Jun 21  2008 chr4.fa
-rw-r--r-- 1 a.user A-Group  21M Jun 21  2008 chr5.fa
-rw-r--r-- 1 a.user A-Group  14K Jun 21  2008 chr6.fa
-rw-r--r-- 1 a.user A-Group  18M Jun 21  2008 chr7.fa
[a.user@foo reference]$ cd /data/ncloonan/ISASdemo/
[a.user@foo reference]$ ./document_filename_changes.pl –p ce6/reference/
[a.user@foo reference]$ ls –lha ce6/reference/
total 128M
drwxr-xr-x 2 a.user A-Group  4.0K Aug  9 21:22 .
drwxr-xr-x 3 a.user A-Group  4.0K Aug  9 16:49 ..
-rw-r--r-- 1 a.user A-Group  15M Jun 21  2008 chr1.fa
-rw-r--r-- 1 a.user A-Group  15M Jun 21  2008 chr2.fa
-rw-r--r-- 1 a.user A-Group  14M Jun 21  2008 chr3.fa
-rw-r--r-- 1 a.user A-Group  18M Jun 21  2008 chr4.fa
-rw-r--r-- 1 a.user A-Group  21M Jun 21  2008 chr5.fa
-rw-r--r-- 1 a.user A-Group  14K Jun 21  2008 chr6.fa
-rw-r--r-- 1 a.user A-Group  18M Jun 21  2008 chr7.fa
-rw-r--r-- 1 a.user A-Group  116 Aug  9 21:22 renamed_chromosomes.txt
[a.user@foo reference]$ more ce6/reference/renamed_chromosomes.txt
chrI.fa        chr1.fa
chrII.fa       chr2.fa
chrIII.fa      chr3.fa
chrIV.fa       chr4.fa
chrV.fa        chr5.fa
chrM.fa        chr6.fa
chrX.fa        chr7.fa
```

Running the ISAS demonstration for the first time

1. When running the ISAS demonstration for the first time, you will need to create the binary files of the reference genome. Firstly, start ISAS.

```
[a.user@foo /]$ cd /data/ncloonan/ISASdemo/
[a.user@foo ISASdemo]$ ./ISAScolorsOldCPU.dem
```

ISAS will now enter interactive mode, and will display many errors as the correct binary files have not been created in the appropriate directories.

```
ISAS Version 3.2.47 for colors
Copyright (c) 2008-2010 Imagenix Technologies Corporation.  All rights reserved.
Imagenix Sequence Alignment System (colors) is Initializing.
Hardware detected: 8 logical processors, 67,665,047,552 bytes RAM.
Could not open file "ReferenceDirectory.txt". Defaulting reference directory to "hg19".
Loading settings for hg19
Could not read settings file hg19/settings-colors.txt
```

```
The settings file will be ignored. Using default settings. The settings file will be
overwritten when ISAS exits.

Error: cannot open file "hg19/reference/chr1bases.bin" for reading



**** Database could not be fully loaded. Check error messages above ******


Error: cannot open file "hg19/1-25-0a.bin" for reading


Loading took 0.0 Seconds. Database size 0MB min 0MB max. Buffer size 0 (0MB)

Current reference directory: hg19
Chr. 1 through Chr. 25 total 0 bases.

 *** REFERENCE DATA NOT LOADED CORRECTLY. Use "CHR=" command to set chromosome range.
 *** If the chromosome range is correct, then use the FASTA command to create reference
files.

Mode=2: ReadLength=25 Mismatches=2 (up to 2 mismatches).
Any bases beyond 25 will be ignored.
Regular Output.
Maximum no. of matches (LIMIT) reported=2.
Sequence Filtering level is 3.

Enter command, or type "?" (and ENTER) for list of commands.
```

2. Tell ISAS what the directory of the genome is using the “database” command. **IMPORTANT:** do not append a trailing ‘/’ to the name of the directory, and do not specify the full path.

```
database=ce6
Saving settings for database=hg19


*** Error: cannot open file "hg19/settings-colors.txt" for saving settings ***

Loading settings for ce6
Could not read settings file ce6/settings-colors.txt
The settings file will be ignored. Using default settings. The settings file will be
overwritten when ISAS exits.
Loading reference for ce6

Error: cannot open file "ce6/reference/chr1bases.bin" for reading



**** Database could not be fully loaded. Check error messages above ******


Error: cannot open file "ce6/1-25-0a.bin" for reading


Current reference directory: ce6
Chr. 1 through Chr. 25 total 0 bases.

 *** REFERENCE DATA NOT LOADED CORRECTLY. Use "CHR=" command to set chromosome range.
 *** If the chromosome range is correct, then use the FASTA command to create reference
files.

Mode=2: ReadLength=25 Mismatches=2 (up to 2 mismatches).
Any bases beyond 25 will be ignored.
Regular Output.
Maximum no. of matches (LIMIT) reported=2.
Sequence Filtering level is 3.

Enter next command, or type "?" (and ENTER) for list of commands.
```

3. Set the chromosome range using the “chr” command.

```
chr=1,7
Reloading reference data...
```

```
Error: cannot open file "ce6/reference/chr1bases.bin" for reading


**** Database could not be fully loaded. Check error messages above ******


Error: cannot open file "ce6/1-7-0a.bin" for reading


Enter next command, or type "?" (and ENTER) for list of commands.
```

4. Use the “fasta” command to make the reference files.

```
fasta
 ce6/reference/chr1.fa to ce6/reference/chr1colors.bin...15072421 bases (0 Ns)
 ce6/reference/chr2.fa to ce6/reference/chr2colors.bin...15279323 bases (0 Ns)
 ce6/reference/chr3.fa to ce6/reference/chr3colors.bin...13783681 bases (0 Ns)
 ce6/reference/chr4.fa to ce6/reference/chr4colors.bin...17493785 bases (0 Ns)
 ce6/reference/chr5.fa to ce6/reference/chr5colors.bin...20919568 bases (0 Ns)
 ce6/reference/chr6.fa to ce6/reference/chr6colors.bin...13794 bases (0 Ns)
 ce6/reference/chr7.fa to ce6/reference/chr7colors.bin...17718854 bases (0 Ns)

Total 100281426 bases (0 Ns).


Error: cannot open file "ce6/1-7-0a.bin" for reading


Enter next command, or type "?" (and ENTER) for list of commands.
```

5. Use the “makebin” command to make the binary files. You will need to type “yes” to confirm that you wish to do this. This is a quick process for a small genome like ce6, however for large mammalian genomes this process can take around 90 minutes. This and the previous four commands will only need to be run **once** per genome.

```
makebin
Are you sure you want to make bin file for chr. 1 through 7 (type "yes" to continue) ?yes
Evaluating 0a
Creating 0a
Sorting 0a
Saving ce6/1-7-0a.bin. writing 1024MB...writing 382MB...Done 1 of 7
Evaluating 1a
Creating 1a
Sorting 1a
Saving ce6/1-7-1a.bin. writing 1024MB...writing 382MB...Done 2 of 7
Evaluating 2a
Creating 2a
Sorting 2a
Saving ce6/1-7-2a.bin. writing 1024MB...writing 382MB...Done 3 of 7
Evaluating 3a
Creating 3a
Sorting 3a
Saving ce6/1-7-3a.bin. writing 256MB...writing 382MB...Done 4 of 7
Evaluating 4a
Creating 4a
Sorting 4a
Saving ce6/1-7-4a.bin. writing 256MB...writing 382MB...Done 5 of 7
Evaluating 5a
Creating 5a
Sorting 5a
Saving ce6/1-7-5a.bin. writing 256MB...writing 382MB...Done 6 of 7
Evaluating 6a
Creating 6a
Sorting 6a
Saving ce6/1-7-6a.bin. writing 256MB...writing 382MB...Done 7 of 7
The current configuration aligns 1000 million sequences at a time.
Makebin took 297 Seconds
```

Creating uniqueomes using the ISAS demonstration software

1. Turn filtering off (when filtering is enabled, this means that ISAS does not perform an exhaustive search).

```
filter=0
Filter turned off.
```

2. Set the mode of the alignment. For an explanation of the modes available, type “mode=?”.

```
mode=?
The following modes are supported:

MODE=0     First 25 bases must have 0 mismatches (perfect match).
MODE=1     First 25 bases can have up to 1 mismatch (i.e. 0 or 1).
MODE=1VA   First 25 bases can have up to 1 mismatch, where two valid
           adjacent mismatches (VA) count as 1 mismatch.
MODE=2     First 25 bases can have up to 2 mismatches (0, 1, or 2).
MODE=2VA   First 25 bases can have up to 2 mismatches, where two valid
           adjacent mismatches (VA) count as 1 mismatch.
MODE=02    First find all perfect matches for first 25, then add all
           locations with up to 2 mismatches for first 25.
MODE=012   First find all perfect matches for first 25, then add all
           locations with 1 mismatch for first 25, then add all locations
           with 2 mismatches for first 25.
MODE=02VA  First find all perfect matches for first 25, then add all
           locations with up to 2 mismatches for first 25 with VA pairs
           counting as singles.
MODE=012VA First find all perfect matches for first 25, then add all 1
           mismatches for first 25, then add all 2VA mismatches for
           first 25.
MODE=3     First 50 bases can have up to 3 mismatches (0, 1, 2 or 3).
MODE=3VA   First 50 bases can have up to 3 mismatches, where two valid
           adjacent mismatches (VA) count as 1 mismatch.
MODE=4     First 50 bases can have up to 4 mismatches (0, 1, 2, 3 or 4).
MODE=4VA   First 50 bases can have up to 4 mismatches, where two valid
           adjacent mismatches (VA) count as 1 mismatch.
MODE=5     First 50 bases can have up to 5 mismatches (0, 1, 2, 3, 4 or 5).

In any "first 25" mode, ReadLength can be extended beyond 25 by using the command
"GLOBAL=ReadLength,TotalMismatches".

In any mode, searching for any sequence stops after a total of LIMIT (currently set to 2)
matches are found for that sequence.


Enter next command, or type "?" (and ENTER) for list of commands.
mode=5
The current configuration aligns 1000 million sequences at a time.
```

3. To confirm that all the settings for the uniqueome are as desired, use the “status” command.

```
status

Current reference directory: ce6
Chr. 1 through Chr. 7 total 100,281,426 bases.
Mode=5: ReadLength=50 Mismatches=5 (up to 5 mismatches).
Any bases beyond 50 will be ignored.
Regular Output.
Maximum no. of matches (LIMIT) reported=2.
Sequence Filtering is disabled (lossless alignment).
The current configuration can align up to 1000 million sequences at a time.

Enter next command, or type "?" (and ENTER) for list of commands.
```

4. Run the command “makestats”.

```
makestats
The current configuration aligns 821 million sequences at a time.
Stats will be created in 1 loops.
Loop no. 1 (of 1)
L0a 3.0 Sec.
R5 75.4 Sec.
```

```
L1a 3.0 Sec.
R5 72.0 Sec.
L2a 2.4 Sec.
R5 64.5 Sec.
L3a 1.3 Sec.
R5 112.0 Sec.
L4a 1.4 Sec.
R5 117.5 Sec.
L5a 1.3 Sec.
R5 110.6 Sec.
L6a 1.4 Sec.
R5 99.7 Sec.

Wrote UniqueChr1Length50Mismatches5ce6
Wrote UniqueChr2Length50Mismatches5ce6
Wrote UniqueChr3Length50Mismatches5ce6
Wrote UniqueChr4Length50Mismatches5ce6
Wrote UniqueChr5Length50Mismatches5ce6
Wrote UniqueChr6Length50Mismatches5ce6
Wrote UniqueChr7Length50Mismatches5ce6
The current configuration aligns 1000 million sequences at a time.
Stats took 680 Seconds

Enter next command, or type "?" (and ENTER) for list of commands.
quit

Bye !
Saving settings for database=ce6
```

Alternatively, the “makestats” command can be run in non-interactive mode (recommended for larger genomes.

```
[a.user@foo ISASdemo]$ nohup ./ISAScolorsOldCPU.dem makestats &
```

Working with ISAS uniqueome output

The output from ISAS is a compact text file that mirrors the chromosomes used as the reference, with one file per chromosome and a one-to-one coordinate-to-score correspondence. A score of 1 indicates a single (self) match for the given word size and stringency criteria. Scores of 2-9 indicate the corresponding number of genome matches and an “X” is used to denote starting positions for N-mers with 10 or more matches. Starting positions that do not correspond to unambiguous words, i.e. those containing “N” and those at the end of the sequence, are denoted with “N”. How to convert these files to BED plots and wiggle plots is described below.

1. Use the script **make_and_compress_BED.pl** to convert the ISAS output into a single BED file containing the positive strand unique start sites.

```
[a.user@foo ISASdemo]$ ./make_and_compress_BED.pl -p /data/ncloonan/ISASdemo/ -f
/data/ncloonan/ISASdemo/ce6/reference/renamed_chromosomes.txt -o /data/ncloonan/ISASdemo/
-n ce6_uniqueome.unique_starts.color-space.50.5.positive.BED
Processing file name: /data/ncloonan/ISASdemo//UniqueChr1Length50Mismatches5ce6...
chromosome: chrI
Processing file name: /data/ncloonan/ISASdemo//UniqueChr2Length50Mismatches5ce6...
chromosome: chrII
Processing file name: /data/ncloonan/ISASdemo//UniqueChr3Length50Mismatches5ce6...
chromosome: chrIII
Processing file name: /data/ncloonan/ISASdemo//UniqueChr4Length50Mismatches5ce6...
chromosome: chrIV
Processing file name: /data/ncloonan/ISASdemo//UniqueChr5Length50Mismatches5ce6...
chromosome: chrV
Processing file name: /data/ncloonan/ISASdemo//UniqueChr6Length50Mismatches5ce6...
chromosome: chrM
Processing file name: /data/ncloonan/ISASdemo//UniqueChr7Length50Mismatches5ce6...
chromosome: chrX

[a.user@foo ISASdemo]$ ls -lha
total 102M
```

```
drwxrwxrwx 6 a.user A-Group 4.0K Aug 12 10:57 .
drwxr-xr-x 9 a.user A-Group  68K Aug 11 12:35 ..
drwxrwxrwx 3 a.user A-Group 4.0K Aug 11 12:16 ce6
-rw-r--r-- 1 a.user A-Group 3.4M Aug 12 10:57 ce6_uniqueome.unique_starts.color-
space.50.5.positive.BED
-rwxr-xr-x 1 a.user A-Group 1.4M Aug  2 15:40 ISAScolorsOldCPU.dem
-rwxr-xr-x 1 a.user A-Group 3.4K Aug 12 10:52 make_and_compress_BED.pl
-rwxr-xr-x 1 a.user A-Group 3.7K Aug 12 10:52 make_COV_plot.pl
-rwxr-xr-x 1 a.user A-Group 1.5K Aug 12 10:52 make_negative_BED.pl
-rw-r--r-- 1 a.user A-Group  314 Aug 11 12:23 ReferenceDirectory.txt
-rw-r--r-- 1 a.user A-Group  15M Aug 12 10:51 UniqueChr1Length50Mismatches5ce6
-rw-r--r-- 1 a.user A-Group  15M Aug 12 10:51 UniqueChr2Length50Mismatches5ce6
-rw-r--r-- 1 a.user A-Group  14M Aug 12 10:51 UniqueChr3Length50Mismatches5ce6
-rw-r--r-- 1 a.user A-Group  17M Aug 12 10:51 UniqueChr4Length50Mismatches5ce6
-rw-r--r-- 1 a.user A-Group  21M Aug 12 10:51 UniqueChr5Length50Mismatches5ce6
-rw-r--r-- 1 a.user A-Group  14K Aug 12 10:51 UniqueChr6Length50Mismatches5ce6
-rw-r--r-- 1 a.user A-Group  18M Aug 12 10:51 UniqueChr7Length50Mismatches5ce6
[a.user@foo ISASdemo]$ tail --lines=5 ce6_uniqueome.unique_starts.color-
space.50.5.positive.BED
chrX    17715934        17716276        1
chrX    17716278        17716282        1
chrX    17716283        17716381        1
chrX    17716387        17717840        1
chrX    17718013        17718703        1
```

2. Use the script **make_negative_BED.pl** to create a BED file containing the negative strand unique start sites. This script requires a chrom.sizes file that can be generated with a tool (**fetchChromSizes**) downloadable from the UCSC genome at http://hgdownload.cse.ucsc.edu/downloads.html. This is simply a tab delimited text file with the chromosome names, and the number of nucleotides in each chromosome as follows:

```
[a.user@foo ISASdemo]$ more ce6.chrom.sizes
chrI    15072421
chrII   15279323
chrIII  13783681
chrIV   17493785
chrM    13794
chrV    20919568
chrX    17718854
[a.user@foo ISASdemo]$ ./make_negative_BED.pl -f
/data/ncloonan/ISASdemo/ce6_uniqueome.unique_starts.color-space.50.5.positive.BED -s /d
ata/ncloonan/ISASdemo/ce6.chrom.sizes -l 50
[a.user@foo ISASdemo]$ tail --lines=5 ce6_uniqueome.unique_starts.color-
space.50.5.negative.BED
chrX    17715984        17716326        1
chrX    17716328        17716332        1
chrX    17716333        17716431        1
chrX    17716437        17717890        1
chrX    17718063        17718753        1
```

3. Use the script **make_COV_plot.pl** to create an unstranded coverage plot of uniqueness. As described above, this tool also needs a chrom.sizes text file.

```
[a.user@foo ISASdemo]$ ./make_negative_BED.pl -f
/data/ncloonan/ISASdemo/ce6_uniqueome.unique_starts.color-space.50.5.positive.BED -s /d
ata/ncloonan/ISASdemo/ce6.chrom.sizes -l 50
[a.user@foo ISASdemo]$ tail --lines=5 ce6_uniqueome.unique_starts.color-space.50.5.Wig
chrX    17718747        17718748        10
chrX    17718748        17718749        8
chrX    17718749        17718750        6
chrX    17718750        17718751        4
chrX    17718751        17718752        2
```

4. Tools for converting BED and WIG files to bigBed (**bedToBigBed**) and bigWig (**wigToBigWig**) are available to download from the UCSC genome browser page at http://hgdownload.cse.ucsc.edu/downloads.html. These files convert the text files generated here, to binary files that can be viewed easily through the UCSC genome browser.