Installation

MSG currently runs on any linux platform.

Place `msg-version.tar.gz` in any directory:

```
$ tar zxf msg-version.tar.gz
$ cd msg_version/
$ make
```

Please ensure that the following dependencies are installed before running MSG (working versions are indicated in parentheses):

- Python (2.6)
- bwa (0.5.7)
- samtools (0.1.9-3)
- biopython-1.53
- Pyrex-0.9.9
- pysam-0.1.2 (apply fix*)
- R packages (HiddenMarkov 1.3-1, zoo 1.6-2, R.methodsS3 1.2.0 and R.oo 1.7.3)

*http://code.google.com/p/pysam/issues/detail?id=22&can=1&q=dandavison0

Setting up the MSG analysis directory

Note that all data files must be located within your MSG analysis directory (links to files are acceptable).

1. Create a text file called “msg.cfg”. This file will specify the location of your data files, and a few other details. You can find an example of an msg.cfg file here: (http://genomics.princeton.edu/AndolfattoLab/MSG_files/msg.cfg).

2. Create a barcode file. You can find an example of a barcode file here: (http://genomics.princeton.edu/AndolfattoLab/MSG_files/barcodes_file.txt).

3. Create (or download) two parental reference genomes in fasta format (links to examples are given at the end of this document).

4. Download read data from an MSG library for a backcross experiment and/or parental genomes (links to examples are given at the end of this document).

5. Create a link to the msg software within your MSG analysis directory:
   ```
   $ ln -s <path_to_msg> msg
   ```

6. To run MSG, simply type the following from within your MSG analysis directory:
   ```
   $ perl msg/msgCluster.pl
   ```

Sample data

Short-read Illumina data from manuscript

F1-parental backcross data
ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX029/SRX029935/SRR071201/

Parental data for Dsim_w501
ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX032/SRX032362/SRR074287/

Parental data for Dsec_w1
ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX032/SRX032363/SRR074288/

To convert these to fastq format, download the SRA Toolkit (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software) and type something like:
./<path to fastq-dump> -A <sra accession number> -D <path to sra file> -O <output directory> &

e.g.
./sratoolkit.2.0rc4-mac64/fastq-dump -A SRR071201 -D sra/SRR071201.sra -O fastq &

Reference genomes
D. simulans reference genome
ftp://ftp.flybase.net/genomes/Drosophila_simulans/current/fasta/dsim-all-chromosome-r1.3.fasta.gz

D. sechellia reference genome
ftp://ftp.flybase.net/genomes/Drosophila_sechellia/current/fasta/dsec-all-chromosome-r1.3.fasta.gz

Barcode file
Link.