Short ibd_stitch tutorial

Based on information provided by Fiona Grimson
Edited and updated by Elizabeth Thompson

August 16, 2016

1 The executable and required files

This document contains a minimal example for running ibd_stitch and interpreting the output. The method is described in [1] and software can be downloaded at [2]. See the web page http://www.stat.washington.du/thompson/Genepi/Stitch.shtml for installation help.
The program is run by calling the installed executable with a single parameter file. For example:

```
~/ibd_stitch/cabal-dev/bin/ibd_stitch test.stitch.par
```

The parameter file "test.stitch.par" specifies the marker data file:

```
input marker data file "test.stitch.markers"
```

The parameter file and marker data file are the only required inputs.

The program will generate an output file, and also a log file. The output file should also be specified in the parameter file:

```
output extra file "test.stitch.out"
```

On some systems it may also be necessary specify the log file explicitly:

```
set log file "test.stitch.log"
```

2 The marker data file

The data input to the program is in a marker data file in a MORGAN format. The MORGAN Online Tutorial [3] may be consulted for more information, but note that `ibd_stitch` does not have the flexibility of MORGAN in specifying data inputs.

An example marker data file for 5 individuals and 20 SNP markers is given in the file `stitch.test.markers`. For `ibd_stitch`, the marker data file should provide the map locations of markers (in centiMorgans), the marker allele frequencies, and the marker genotypes of individuals, in the format given in the example file, and printed below:

```
map marker positions  0.01 0.08 0.12 0.16 0.73 0.84 0.95 0.99 1.03 1.06 1.11
  1.18 1.22 1.26 1.83 1.94 2.05 2.19 2.23 2.26
set marker 1 allele freq 0.824275 0.175725
set marker 2 allele freq 0.2181 0.7819
set marker 3 allele freq 0.172025 0.827975
set marker 4 allele freq 0.1836 0.8164
set marker 5 allele freq 0.139025 0.860975
```

```
set marker 6 allele freq 0.154075 0.845925
set marker 7 allele freq 0.7522 0.2478
set marker 8 allele freq 0.24485 0.75515
set marker 9 allele freq 0.232075 0.767925
set marker 10 allele freq 0.206125 0.793875
set marker 11 allele freq 0.824275 0.175725
set marker 12 allele freq 0.2181 0.7819
set marker 13 allele freq 0.172025 0.827975
set marker 14 allele freq 0.1836 0.8164
set marker 15 allele freq 0.139025 0.860975
set marker 16 allele freq 0.154075 0.845925
set marker 17 allele freq 0.7522 0.2478
set marker 18 allele freq 0.24485 0.75515
set marker 19 allele freq 0.232075 0.767925
set marker 20 allele freq 0.206125 0.793875
set marker data 20
996677 1 1 2 2 2 2 2 2 2 2 2 2 1 2 1 2 1 2 1 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 1 2 1 2 1 2
986212  1 1 2 2 1 2 2 2 2 2  2 2 1 1 2 1 2 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 1 2 1 2 1 2 1 2
996322 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 1 2 1 2 2 2 2 2 1 2 1 2 2 2 2 2 2
997644 1 1 1 2 1 2 2 2 2 1 1 2 1 1 2 2 2 2 2 2 1 1 2 2 2 1 2 2 2 2 2 2 1 2 1 2 1 2 1 2
1009106 1 1 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 1 2 2 1 1 1 2 1 2 1
```

3 Parameter File

An example parameter file is given in `test.stitch.par`, and is printed below. The parameters are described in the ibd_stitch paper [1] and the download page [2]. Note that although many of the statements are in MORGAN format, the formats are far more fragile. For example: Be careful with white space. On some systems, it is necessary to specfy log files explicitly.

```
input marker data file "test.stitch.markers"          # location of marker data file
output extra file "test.stitch.out"                   # location of output
set iterations 2                                      # number of iterations
select unphased data                                  # marker data phased/unphased
set printlevel 3                                      # output print level
set seed 12345                                        # random seed

set population kinship 0.05                           # model parameters
set kinship change rate 0.5
set transition matrix null fraction 0.05
set genotyping error rate 0.01
set logfile "test.stitch.log"                         # default, but may be needed
select 2011 state transition matrix
set number haplotype sets 4
set haplotype set size 4
```

4 Interpretation of the Program Output

4.1 First example

The output is directed to the file specified by the "output extra file" in the parameter file. In this example this is the file "test.stitch.out" which is printed below, with some comments added:

```
# --- first iteration starts here ---- #
996677 0 1 0  # one gamete for id 996677 -- all loci have fgl 1
996677 0 2 0  # other gamete for id 996677 -- all loci have fgl 2
986212 0 3 0
986212 0 4 0
996322 0 5 0
996322 0 6 0
997644 0 5 3 4 7 14 3 16 7   # this gamete has some switches in fgl
997644 0 4 2 4 8 11 3
1009106 0 9 3 4 6 11 9 17 1
1009106 0 10 0
# --- second iteration starts here ----  #
996677 0 1 0
996677 0 2 0
986212 0 3 0
986212 0 4 2 4 2 11 4
996322 0 5 1 4 1
996322 0 6 0
997644 0 7 2 14 2 15 4
997644 0 3 3 4 8 15 2 17 8
1009106 0 6 1 3 9
1009106 0 5 3 4 1 8 10 17 3
```

Each iteratiom consists of a realization of an ibd graph on the k individuals of the input data; in this example the five individuals 996677, 986212, 996322, 997644, and 1009106 ($k = 5$). Each individual has two gametes; each of the two rows of an individual corresponds to one gamete.

The graph is written in terms of fgl labels at the supplied markers, where fgl refers to an arbitrary founder genome label. The format of the ibd graph is as for the The formatting of the output is the same as MORGAN/gl_auto, described in section 9.5 of [3]. Each line of output has the format:

`[name] [ 0 ] [ first fgl ] [ number switches ]  *[ switch point ] [ new fgl ]*`

where switch points and new fgls within * are provided for each switch. For example, line 7 of the output, the first gamete of 997644, reads `997644 0 5 3 4 7 14 3 16 7` . This indicates that this gamete starts with fgl 5 at marker 1, and will have 3 switch points. At marker 4 it will switch to fgl 7, at marker 14 it will switch to fgl 3, at marker 16 it will switch to fgl 7.

If two gametes have the same fgl at a locus then they are ibd at that locus. The ibd graph contains the information needed to determine the ibd state among the $2k$ gametes of the k individuals at any marker locus. For example, in the first realization, individual 996322 has fgl 5 over all the markers. from the above example the first gamete of individual 997644 also has fgl 5 at markers 1, 2 and 3. So at markers 1, 2 and 3, there is ibd between one gamete of individuals 996322 and one gamete of 997644.

4.2 An alternate example

In the zip file of examples, there is an additional example for a trio of individuals. The parameter file is `trio_test.par` and the specified marker data file is `trio_thin1.markers`. This example has only three individuals, but uses 334 markers and produces 1000 realizations.

For both examples, in addition to the parameter and marker data files, the output and log files are included in the zip file for reference.

References

[1] Glazner, C., Thompson, EA. (2015) Pedigree-free descent-based gene mapping from population samples. *Hum Hered.* 80: 21-35

http://www.ncbi.nlm.nih.gov/pubmed/26228693

[2] ibd_stitch 1.0 [Software] (2016) *https://github.com/cglazner/ibd_stitch*

[3] MORGAN 3.3 Online Tutorial (2015)

http://www.stat.washington.edu/thompson/Genepi/MORGAN/morgan-tut_33_html/morgan-tut.html

If this link no longer works, please check the software page at

http://www.stat.washington.edu/thompson/Genepi/pangaea.shtml

for updated information.