Perl for human linkage analysis

Karl W. Broman

Department of Biostatistics, Johns Hopkins University

http://www.biostat.jhsph.edu/~kbroman

Data

- A set of pedigrees
  - Family, individual, mom, dad, sex
- Phenotypes (traits) and covariates on many individuals
  - Binary, ordinal, continuous, etc.
- Genotypes at a set of markers
  - A pair of numbers at each marker
- Genetic map for the markers
  - Chromosome, position

Goal: Identify regions of the genome harboring genes that influence the phenotype.

Perl for human linkage analysis

- Manipulation of text files
  - Convert data files from one format to another.
- Automation of analysis
  - Run a program many times and combine/summarize the output.
- Computer simulations
  - Simulate data.
  - Apply a program/method.
  - Extract the interesting bits from the output.
  - Repeat many times.

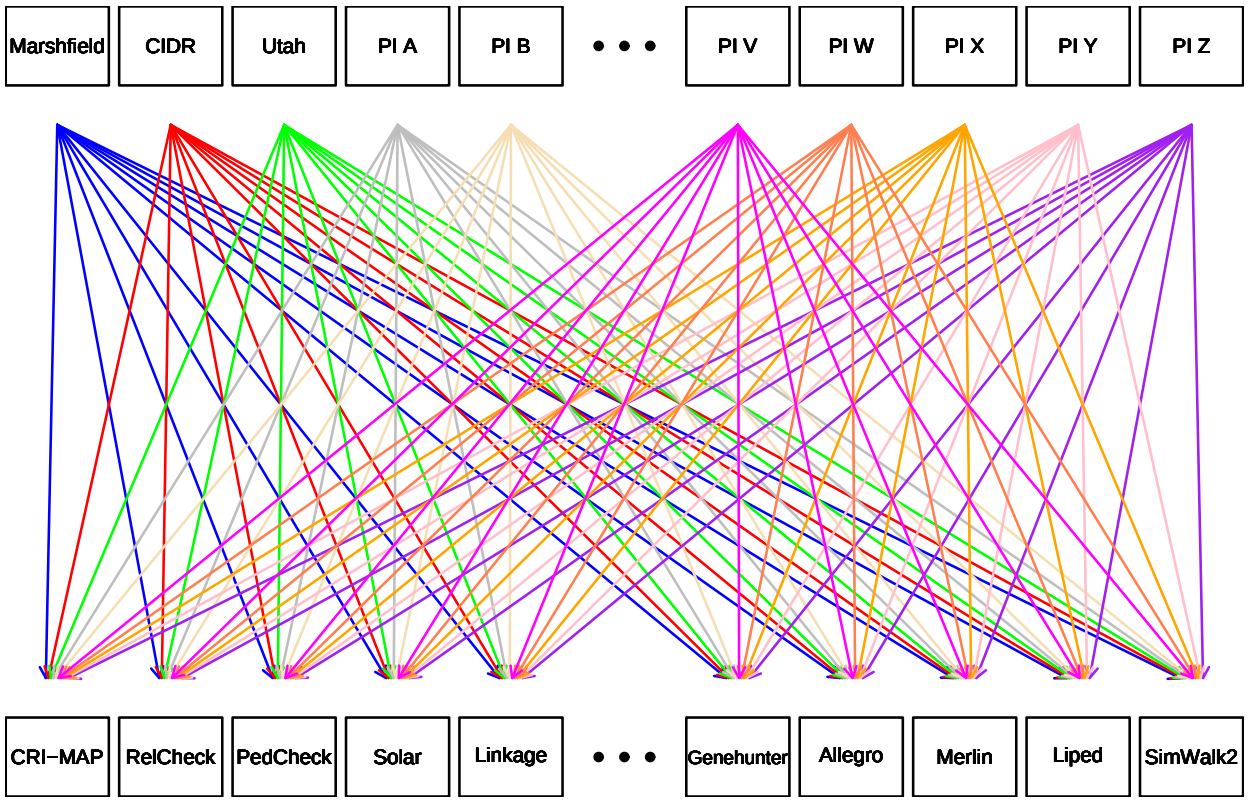
Steps in the analysis

- Verify/correct relationships.
- Identify/resolve genotyping errors.
- Identify/resolve strange phenotypes/covariates.
- Perform the actual analysis.
- Conduct simulations to assess performance or to obtain P-values with proper adjustment for test multiplicity.

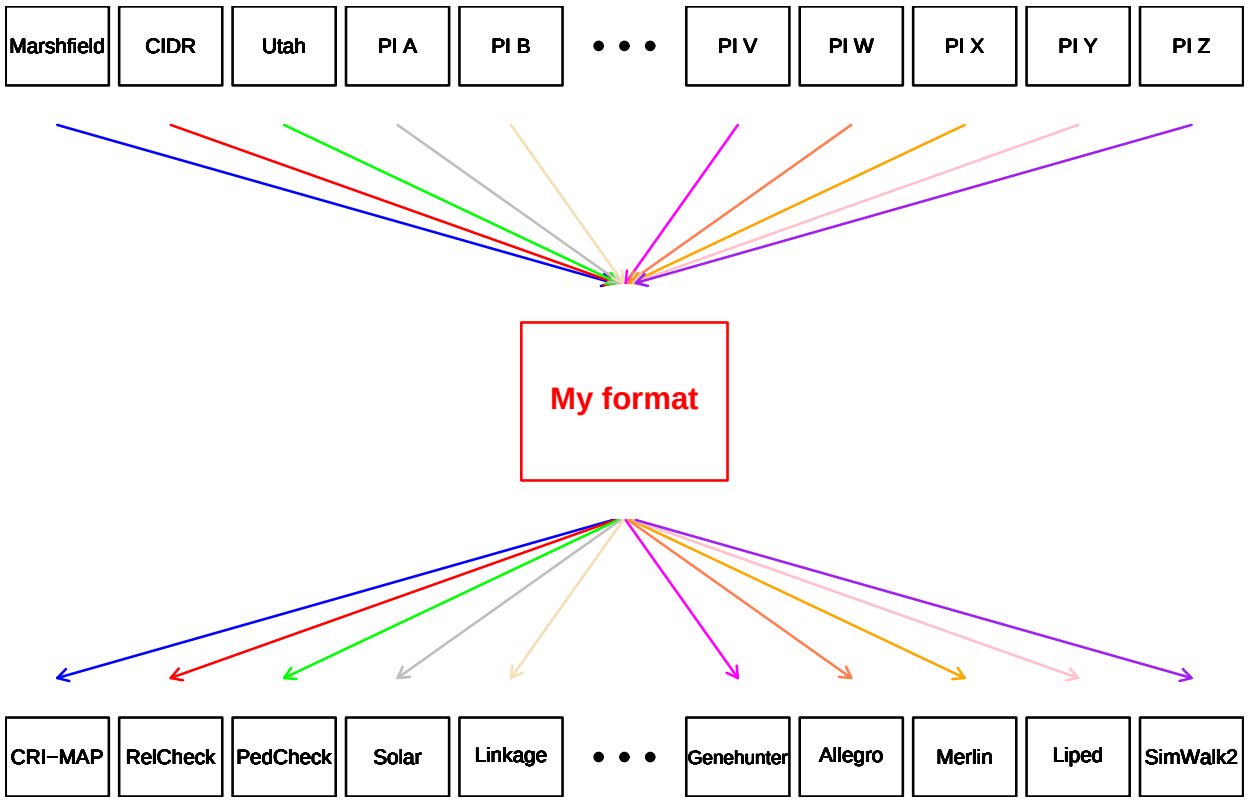
Guiding principles

- Never modify the data “by hand”.
  - Exact record of what you’ve done.
  - Avoid errors.
  - Automation (if the data should change. . .)
- Manipulation of data files is best performed by the analyst.
- Be organized, keep notes.
- Re-use code
  - Plan for the future.
  - Comment your code.
  - Code in a modular and reasonably general form.
- There’s probably an easier way, but. . .

Manipulation of data files



Manipulation of data files



Input: Genotype data

~/Projects/PIn/Rawdata/ [402 files]

```
002vd5.out      224zg11.out     ATA1B07.out
002zf1.out      242zg5.out      ATA20G07Z.out
029xg1.out      248vc5.out      ATA21A03.out
035xb9.out      254ve1.out      ATA22D02.out
044xg3.out      280we5.out      ATA22E01.out
049xd2.out      2QTEL47.out     ATA22G07.out
059xa9.out      303zg9.out      ATA23A10.out
063xf4.out      306yg5.out      ATA23G05.out
077xd3.out      309va9.out      ATA24A08.out
150xf10.out     310wd9.out      ATA24F10.out
157xg3.out      337zh9.out      ATA25E07.out
164xb8.out      350VD1.out      ATA25F04.out
165xc11.out     350zc9.out      ATA25F09.out
16PTEL06.out    4PTEL04.out     ATA26B08.out
178xc3.out      6QTEL54.out     ATA26D07.out
18QTEL11.out    ACT1A01.out     ATA27A03.out
198zb4.out      ATA10F11.out    ATA27A06.out
1QTEL19.out     ATA10H11.out    ATA27C07.out
203yg9.out      ATA11D10.out    ATA27C09.out
218yb5.out      ATA18A07.out    ATA27D04.out
                                    ⋮
                                Mfd259.out
                                Mfd265.out
                                Mfd313.out
                                PAH.out
                                SDF1.out
                                SE30.out
                                SRA.out
                                UT1307.out
                                UT1355.out
                                UT1772.out
                                UT2095.out
                                UT254.out
                                UT5029.out
                                UT532.out
                                UT7129.out
                                UT7136.out
                                UT721.out
                                UT7544.out
```

Input: Genotype data

~/Projects/PIn/Rawdata/GATA8F07.out

Marker: GATA8F07 Project: [masked] Gel: none User: none

[family]	[ind'l]	[dad]	[mom]	[sex]	[genotype]	[quality]	
1	1	105	106	1	245	245	0.990
1	2	105	106	2	245	245	0.990
1	3	105	106	1	253	245	0.990
1	9	105	106	2	245	245	0.990
1	10	107	108	1	253	249	0.990
1	11	0	0	2	253	253	0.990
1	12	10	11	1	253	253	0.990
1	13	10	11	1	253	249	0.990
1	15	13	14	2	253	245	0.990
1	16	0	0	2	249	249	0.990
1	17	13	16	1	249	249	0.990
1	18	13	16	2	253	249	0.990
1	19	10	11	1	253	253	0.990
1	20	0	0	2	249	245	0.990
1	21	19	20	2	253	245	0.990
1	22	19	20	1	253	249	0.990
1	23	19	20	2	253	249	0.990

Input: Phenotype data

~/Projects/PIn/Rawdata/pheno.csv

```
Pedigree#,ID#,Father,Mother,Sex,Phe1,Phe2,Phe3,Phe4,Liability
1,1,105,106,1,2,2,2,2,4
1,2,105,106,2,2,2,2,2,4
1,3,105,106,1,1,2,2,1,4
1,9,105,106,2,1,1,1,1,4
1,10,107,108,1,2,2,2,2,4
1,11,0,0,2,1,1,1,1,4
1,12,10,11,1,0,0,0,0,3
1,13,10,11,1,2,1,2,1,3
1,14,0,0,2,1,1,1,1,2
1,15,13,14,2,1,1,1,1,1
1,16,0,0,2,1,1,1,1,2
1,17,13,16,2,1,1,1,1,1
1,18,13,16,2,1,1,1,1,1
1,19,10,11,1,1,1,1,1,3
1,20,0,0,2,1,1,1,1,3
1,21,19,20,2,1,1,1,1,1
1,22,19,20,1,1,1,1,1,1
1,23,19,20,2,1,1,1,1,1
```

Input: Genetic maps

`~/Projects/GenMaps/FullInfo/full??.csv`

```
marker,dnum,genbank,pos,ave,female,male,onmap,dup?,het
12QTEL36,D12S2341,Unknown,391,168.79,210.59,127.98,12QTEL36,,0.56
12QTEL82,D12S2342,Unknown,390,168.79,210.59,127.98,12QTEL82,,0.76
12QTEL87,D12S2343,Unknown,392,170.66,214.45,127.98,12QTEL87,,0.50
1GF1,Unknown,Unknown,287,109.47,142.29,77.33,1GF1,,0.42
224yf10,D12S323,Unknown,240,95.03,118.64,72.00,224yf10,,0.17
238yb10,D12S102,Unknown,179,75.17,96.20,54.93,238yb10,,0.77
249vf9,D12S103,Unknown,156,69.23,87.21,51.72,249vf9,,0.38
273zc9,D12S336,Unknown,43,19.68,12.85,26.63,273zc9,,0.82
AFM010th7,D12S76,Z16423,345,136.82,181.20,92.87,AFM010th7,,0.70
AFM026tb5,D12S77,Z16443,46,20.27,14.05,26.63,AFM026tb5,,0.89
AFM026tf3,D12S78,Z16444,291,111.87,146.85,77.33,AFM026tf3,,0.91
AFM026wh7,D12S1595,Z50892,76,36.06,33.30,38.92,AFM026wh7,,0.60
AFM065ye9,D12S317,Z23320,295,114.28,149.14,80.07,AFM065ye9,,0.71
AFM067yc5,D12S79,Z16516,334,125.31,161.73,89.79,AFM067yc5,,0.88
AFM073wh7,D12S320,Z23326,61,30.60,30.08,31.08,AFM073wh7,,0.82
AFM077yc1,D12S1345,Z50928,248,96.09,120.77,72.00,AFM077yc1,,0.78
AFM086xd7,D12S330,Z23349,300,116.08,152.54,80.07,AFM086xd7,,0.75
AFM092wd11,D12S331,Z23351,113,54.46,62.74,46.39,AFM092wd11,,0.61
```

My format

`stem.csv`	Pedigree info + phenotypes + covariates
`stem??.gen`	Genotype data [CRI-MAP format]
`stem??.map`	Genetic map
`stem??.frq`	Allele frequencies

?? = `01`, `02`, `...`, `23`

Output: `stem.csv`

Simple, comma-delimited file

`fam,ind,mom,dad,sex,phe1,phe2,...`

Output: `stem??.gen`

```
2                                   [no. families]
31                                  [no.markers]
D1S468
D1S214
D1S450
[...]
D1S2836
1                                   [family ID]
32                                  [no. individuals]
1 202 201 1                         [individual, mom, dad, sex]
206 208 145 145 [...] 243 250       [genotypes]
2 202 201 1
206 208 122 145 [...] 243 243
274 0 0 1
0 0 0 0         [...] 0 0
[...]
2                                   [the next family]
34
101 151 150 0
200 202 0 0     [...] 243 243
```

Output: stem??.map

```
D1S468 4.22 4.46 3.54              [marker, sex-ave, female, male]
D1S214 14.04 15.57 12.39
D1S450 20.61 20.00 21.04
D1S2667 24.68 26.18 23.19
D1S2697 37.05 40.41 34.04
D1S199 45.33 50.41 40.52
D1S234 55.10 65.26 45.36
D1S255 65.47 82.41 48.95
D1S2797 75.66 100.52 51.36
D1S2890 85.68 115.09 56.70
D1S230 95.31 127.72 63.10
D1S2841 106.45 141.34 72.16
D1S207 113.69 151.07 76.96
D1S2868 126.16 167.43 85.53
D1S206 134.20 176.02 93.04
D1S2726 144.38 188.92 100.54
D1S252 150.27 197.51 103.74
[...]
D1S2836 285.75 363.91 208.50
```

Output: stem??.frq

```
D1S468 7                   [marker, no. alleles]
190 0.02105263             [allele, frequency]
196 0.02105263
200 0.24210526
202 0.11578947
204 0.04210526
206 0.44210526
208 0.11578947
D1S214 10
122 0.15306122
134 0.06122449
137 0.18367347
138 0.03061224
139 0.07142857
140 0.06122449
142 0.08163265
143 0.17346939
144 0.02040816
145 0.16326531
D1S450 8
[...]
```

The ubiquitous line endings problem

Text files in Unix, MacOS, and DOS (ie Windows) have different conventions regarding the character(s) at the ends of lines in text files.

This causes problems when moving files from Windows to Unix.

I use the perl script `clean` (written by guys at the Berkeley Statistical Computing Facility) to convert Windows-type text files to the Unix format.

http://www.biostat.jhsph.edu/~kbroman/perlintro/clean.html

Code snippet 1

```
$dir = "Rawdata";
opendir(DIR, $dir);
while(defined($file = readdir(DIR))) {
  unless($file =~ /(\w*)\.out/) {
    next;
  }
  $mar = $1;
  push(@markers, $mar);

  $file = $dir . "/" . $file;
  open(IN, $file) or die("Cannot open $file.\n");
  $line = <IN>;
  while($line = <IN>) {
    chomp($line);
    @v = split(/\s+/, $line);
    if($v[0] eq "") { shift @v; }
    ($fam, $ind, $dad, $mom, $sex, $g1, $g2, $qual) = @v;
    @{$gen{$fam}{$ind}{$mar}} = ($g1, $g2);
  }
  close(IN);
}
```

Code snippet 2

```
while($line = <IN>) {
  chomp($line);
  ($fam,$ind,$dad,$mom,$sex,@v) = split(/,/, $line);

  $mom{$fam}{$ind} = $mom;
  $dad{$fam}{$ind} = $dad;
  $sex{$fam}{$ind} = $sex;
  @{$phe{$fam}{$ind}} = @v;
}

foreach $fam (sort numerically keys %mom) {
  foreach $ind (sort numerically keys %{$mom{$fam}}) {

    $phe{$fam}{$ind}[2] = [...];

    @{$phe{$fam}{$ind}}[3..4] = [...];
  }
}

sub numerically { $a <=> $b; }
```

Code snippet 3a

```
foreach $i (1..23) {

  while($line = <IN>) {
    chomp($line);
    ($mar,$sexave,$female,$male) = split(/,/, $line);

    $chr{$mar} = $i;
    $pos{$mar} = $sexave;
  }

}


@markers = sort bypos keys %chr;

sub bypos {
  if($chr{$a} == $chr{$b}) {
    return($pos{$a} <=> $pos{$b});
  }
  $chr{$a} <=> $chr{$b};
}
```

Code snippet 3b

```
foreach $i (1..23) {

  while($line = <IN>) {
    chomp($line);
    ($mar,$sexave,$female,$male) = split(/,/, $line);

    push(@{$markers{$i}}, $mar);
    $pos{$mar} = $sexave;
  }

}


foreach $mar (sort bypos2 @{$markers{$i}}) {
  [...]
}

sub bypos2 { $pos{$a} <=> $pos{$b}; }
```

Code snippet 4

```
foreach $mar (@markers) {
  foreach $fam (keys %mom) {
    foreach $ind (keys %{$mom{$fam}}) {

      ($g1, $g2) = @{$gen{$fam}{$ind}{$mar}};

      unless($g1==0 or $g2==0) {
        ($freq{$mar}{$g1})++;
        ($freq{$mar}{$g2})++;
        ($ng{$mar}) += 2;
      }
    }
  }

  foreach $allele (keys %{$freq{$mar}}) {
    $freq{$mar}{$allele} /= $ng{$mar};
  }

}
```

~/Projects/Db

```
db2gh.pl
db2mcmc.pl
db2msrelcheck.pl
db2pedcheck.pl
db2relcheck.pl
db2simwalk2.pl
db2solar.pl
checkSex.pl
checkXinher.pl
countAlleles.pl
countMissing.pl
findFamilies.pl
findMarker.pl
findNoData.pl
gendisplay.pl
ind2drop.pl
removeFams.pl
removeInd.pl
summarize.pl
makePar.pl
runChrompics.pl
runGH.pl
runGHP.pl
runIlink.pl
runMlink.pl
runPedCheck.pl
runSinglepoint.pl
simulateData.pl
```

Resources

- http://www.biostat.jhsph.edu/~kbroman/perlintro
  - My “Intro to perl” page.
  - Goes through an example of manipulating a text file for human linkage analysis, in extremely gory detail.
- http://stein.cshl.org/genome_informatics/#schedule
  - Cold Spring Harbor genome informatics course.
  - Some really good bits I refer to regularly.
- Perl Cookbook
  - The book I refer to most often.
- Perl CD Bookshelf
  - A very good deal; contains Programming Perl, Perl Cookbook, others.
  - I prefer to also have hardcopies.
- http://www.cpan.org

Summary

- Be organized.
- Think long term.
- Write code that works.
- Dump `vi`; use `emacs`