TetraOrigin: Haplotype reconstruction in outcrossing tetraploids

Chaozhi Zheng
Biometris, Wageningen University and Research centre
May 7, 2015

Helps on "TetraOrigin`"

In[1]:=
```
SetDirectory[
  ParentDirectory[NotebookDirectory[]] <> "\\TetraOrigin_Packages"];
Needs["TetraOrigin`"]
SetDirectory[NotebookDirectory[]];
? "TetraOrigin`*"
```

▼ TetraOrigin`

bivalentDecoding	getSummaryITO	inferTetraPhase	maxIteration	maxStuck	relabelHaplo	toGridProb
bivalentPhasing	inferTetraOrigin	loglTetraOrigin	maxPhasingRun	minRepeatRun	saveAsSummaryITO	

In[5]:=
```
Options[inferTetraPhase]
```

Out[5]=
```
{maxStuck → 10, maxIteration → 100, minRepeatRun → 3,
 maxPhasingRun → 20, bivalentPhasing → True}
```

In[6]:=
```
? bivalentPhasing
? inferTetraPhase
```

bivalentPhasing is an option for phasing algorithm. If bivalentPhasing = True, phasing algorithm accounts for only bivalent chromosome pairing but not qudrivalent pairing. If bivalentPhasing = False, phasing algorithm accounts for both bivalent chromosome pairing and qudrivalent pairing.

inferTetraPhase[SNPDose, chrsubset, snpsubset, epsF, eps, ploidy] returns {founderhaplo, loglhistory} where founderhaplo is the estimated parental haplotypes and loglhistory is the records of log likelihood for each proposed parental haplotypes. The ploidy = 4 for tetraploid species. epsF and eps are the dosage error probabilities for parents and siblings, respectively. SNPDose is the input marker data include the genetic map and the dosages for two parents and their full sibs. chrsubset is a list of indices for linkage groups, e.g. chrsubset={1,3} the first and the third linkage groups will be analyzed, and chrsubset= "All" all the likage groups will be analyzed. snpsubset is a list of indices for SNP markers of each linkage group, e.g. snpsubset={2, 5, 10} the second, the fifth, and the tenth markers of each linkage groups will be analyzed, and snpsubset= "All" all the markers will be analyzed.

Example data

- Simulated example data consisting: (50 full sibs + 2 parents) x 75 SNP markers, with true dososage error eps =0. Genetic map is required with SNP locations in cM.

In[8]:=
```
SNPDose = Import["TetraOrigin_Input_SNPDose_ExampleData.csv"];
Dimensions[SNPDose]
SNPDose[[ ;; 10,
   ;; ;; Round[Last[Dimensions[Rest[SNPDose]]] / 9]]] // TableForm
```

Out[9]=
```
{55, 76}
```

Out[10]//TableForm=

marker	A0113	A0241	A0369	A0497	A0625	A0753
chromosome	A	A	A	A	A	A
position	11.9003	23.5561	37.4988	48.2708	59.6022	74.8123
P1	3	0	1	0	2	2
P2	1	1	1	2	0	2
F1_001	2	0	1	1	0	1
F1_009	2	1	0	1	0	2
F1_017	3	1	1	1	0	3
F1_025	3	1	1	1	1	1
F1_033	2	0	2	0	1	2

- Input CSV formatted data
  - Rows 1-3 show the genetic map.
  - The rest shows SNP dosage data for two parents and sampled full sibs.
  - Dosage ranges from 0 to 4 for a tetraploid species.

Haplotype phasing

- Estimating parental likage phases, assume only bivalent chromosome pairing.

In[11]:=
```
{chrsubset, snpsubset} = {"All", "All"};  epsF = 0; eps = 10^(-3.);
ploidy = 4; {esthaplo, loglhistory} = inferTetraPhase[SNPDose,
  chrsubset, snpsubset, epsF, eps, ploidy, bivalentPhasing → True];
```

```
Start Date =Thu 7 May 2015 11:47:10. Chromosome  = A

Finish Phasing Date =Thu 7 May 2015 11:47:37. Time used in Phasing = 27.3
  Seconds. log posterior of phasing runs = {-1813.36, -1813.36, -1813.36}
```

In[12]:=
```
{esthaplo[[All, ;; ;; 10]] // TableForm, loglhistory // TableForm}
```

Out[12]=

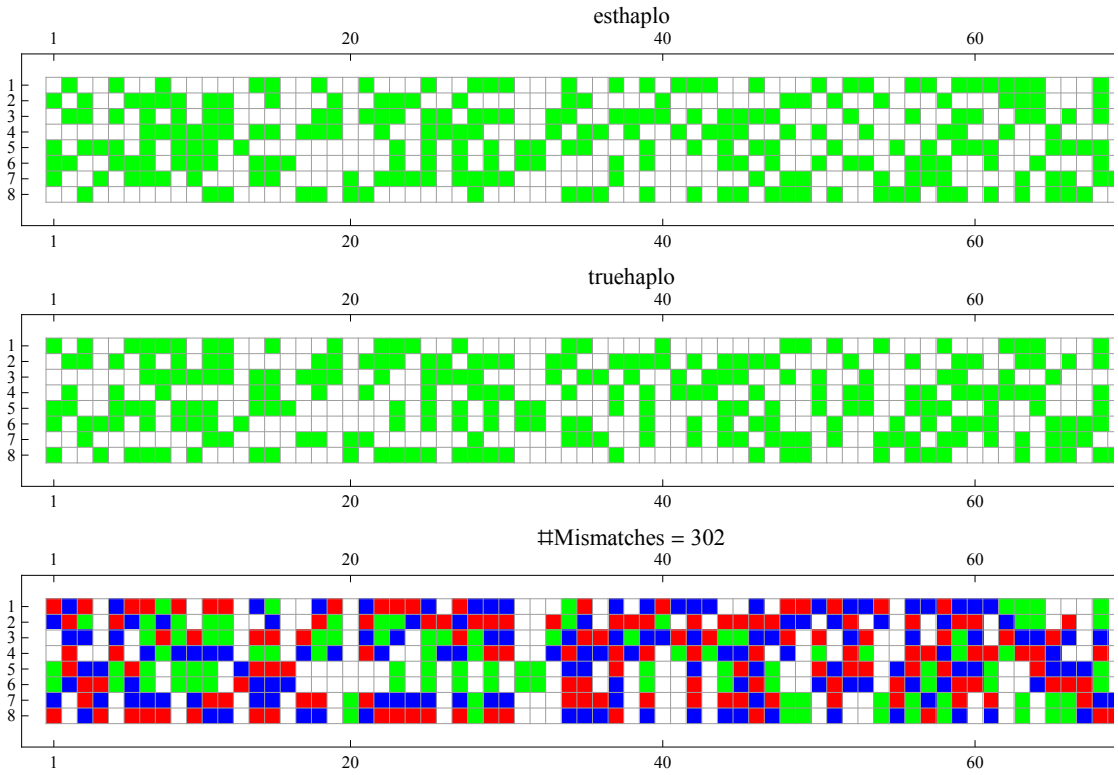
SNP	A0145	A0305	A0465	A0625	A0785	A0945	A1105
Chromosome	A	A	A	A	A	A	A
P1_1	1	1	2	1	2	2	2
P1_2	1	1	1	2	1	1	1
P1_3	1	1	2	2	1	2	1
P1_4	2	1	1	1	2	1	2
P2_1	2	1	1	1	1	2	1
P2_2	2	1	1	1	2	1	1
P2_3	2	2	2	1	1	1	2
P2_4	1	2	1	1	1	1	2

,

Haplotype phasing (cont.)

- The estimated haplotypes vs the true haplotypes

Out[20]=



- Relabel the estimated haplotypes withe respect to the true haplotypes.

In[21]:=
```
founderhaplo = relabelHaplo[esthaplo, refhaplo];
founderhaplo[[3 ;;, 2 ;;]] == refhaplo[[3 ;;, 2 ;;]]
```

Out[22]=
```
True
```

Estimating dosage error probability

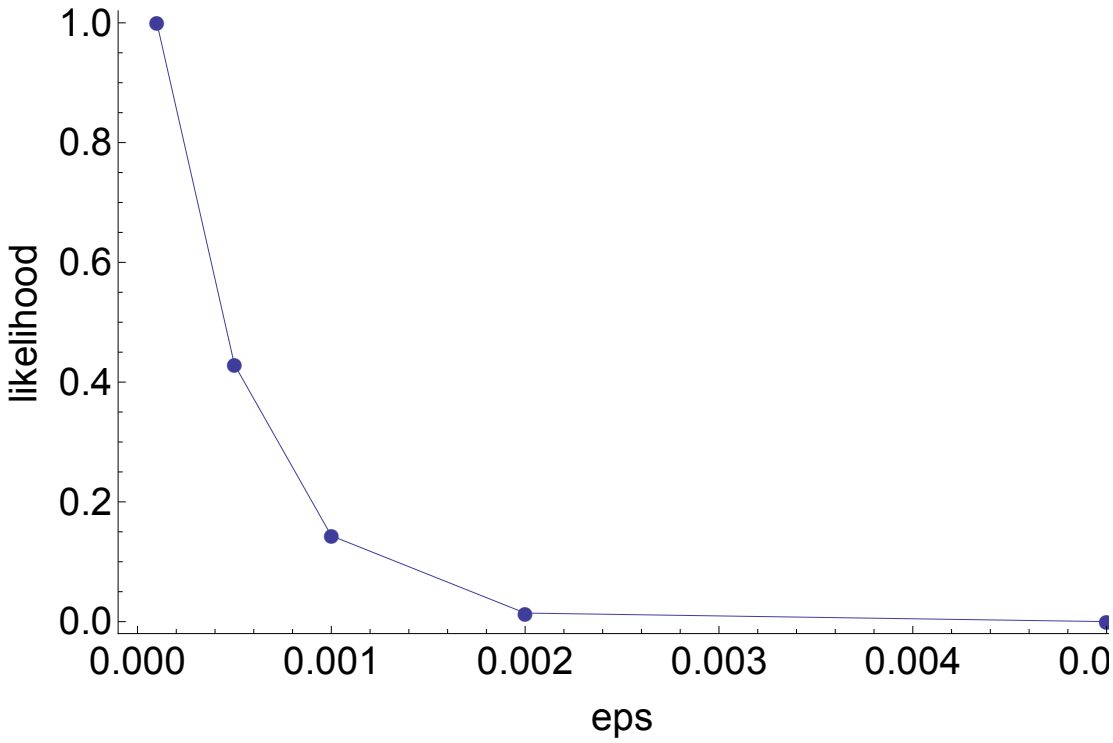
- Calculate the log likelihood given estimated parental hapotypes.

In[23]:=
```
{chrsubset, snpsubset} = {"All", "All"}; eps = 10^(-3.); ploidy = 4;
loglTetraOrigin[SNPDose, chrsubset, snpsubset, eps,
  ploidy, esthaplo, bivalentDecoding → True] // AbsoluteTiming
loglTetraOrigin[SNPDose, chrsubset, snpsubset, eps, ploidy,
  esthaplo, bivalentDecoding → False] // AbsoluteTiming
```

Out[24]= {0.880026, -1813.36}

Out[25]= {11.720352, -1144.66}

Out[28]=



Posterior decoding

In[29]:=
```
Options[inferTetraOrigin]
```

Out[29]=
```
{maxStuck → 10, maxIteration → 100, minRepeatRun → 3,
 maxPhasingRun → 20, bivalentPhasing → True, bivalentDecoding → False}
```

In[30]:=
```
? inferTetraOrigin
```

inferTetraOrigin[SNPDose, chrsubset, snpsubset, eps, ploidy, founderhaplo, outputid] calculates the posteriorprobabilitiesfor each sib at each SNP marker given the the parentalhaplotype founderhaplo, and the results are saved in the file "TetraOrigin_Output_outputid_LinkageGroup1.txt" for the first linkage group, and so on for the rest linkage groups Refer to inferTetraPhasefor estimating parentalhaplotype and descriptionsof other paremters

inferTetraOrigin[SNPDose, chrsubset, snpsubset, eps, ploidy, outputid] is a combinationof {founderhaplo, loglhistory} = inferTetraPhase[SNPDose, chrsubset, snpsubset, epsF, eps, ploidy] and inferTetraOrigin[SNPDose, chrsubset, snpsubset, eps, ploidy, founderhaplo, outputid].

In[31]:=
```
outputid = "ExampleData";
inferTetraOrigin[SNPDose, chrsubset,
  snpsubset, eps, ploidy, founderhaplo, outputid,
  bivalentPhasing → True, bivalentDecoding → False];
```

```
Start Date =Thu 7 May 2015 11:48:49. Outputfile=
 TetraOrigin_Output_ExampleData_LinkageGroup1.txt

Finish Date =Thu 7 May 2015 11:49:09
 . Time used in Posteriordecoding = 20.3 Seconds.

--------------------------------------------------------------------------------
  --------
```

Summary the TetraOrigin output

- Re-save the output of inferTetraOrigin as summary.

In[33]:=
```
tetraResultFile = "TetraOrigin_Output_" <> outputid <> "_LinkageGroup1.txt"
summaryFile = StringReplace[tetraResultFile, {".txt" → "_Summary.csv"}]
saveAsSummaryITO[tetraResultFile, summaryFile];
```

Out[33]= TetraOrigin_Output_ExampleData_LinkageGroup1.txt

Out[34]= TetraOrigin_Output_ExampleData_LinkageGroup1_Summary.csv

- Import the summary for further analysis in mathematica

In[36]:=
```
res = getSummaryITO[summaryFile];
First[res] // MatrixForm
```

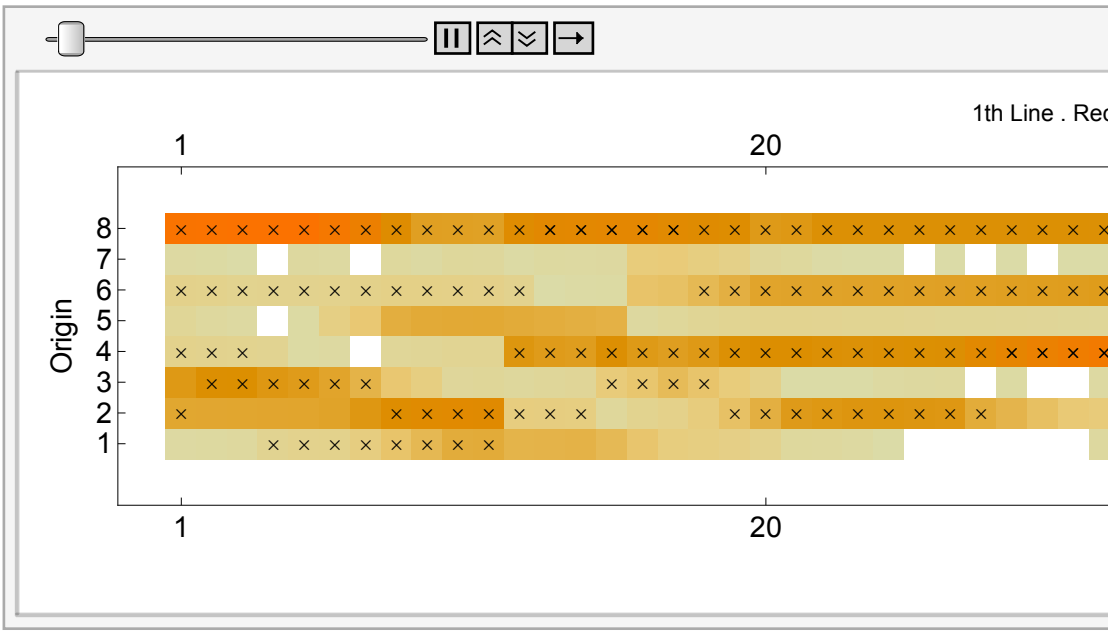
Out[37]//MatrixForm=

inferTetraOrigin-Summary	Genetic map of biallelic markers
inferTetraOrigin-Summary	MAP of parental haplotypes
inferTetraOrigin-Summary	Reference haplotypes
inferTetraOrigin-Summary	ln marginal likelihood of each valent of each
inferTetraOrigin-Summary	ln marginal likelihood given the LG type of eac
inferTetraOrigin-Summary	Genotypes in order
inferTetraOrigin-Summary	Conditonal genotype probability
inferTetraOrigin-Summary	Conditonal haplotype probability

Conidtional posterior probability

- Posterior haplotype porbabilities at markers of each sampled lines.

Out[41]=



Put together

```
{esthaplo, loglhistory} = inferTetraPhase[SNPDose, chrsubset,
    snpsubset, epsF, eps, ploidy, bivalentPhasing → True];
founderhaplo = relabelHaplo[esthaplo, refhaplo];
inferTetraOrigin[SNPDose, chrsubset,
  snpsubset, eps, ploidy, founderhaplo, outputid,
  bivalentPhasing → True, bivalentDecoding → False];
(*Analyze results of Linkage groups one by one*)
tetraResultFile =
  "TetraOrigin_Output_" <> outputid <> "_LinkageGroup1.txt";
summaryFile = StringReplace[tetraResultFile,
    {".txt" → "_Summary.csv"}];
saveAsSummaryITO[tetraResultFile, summaryFile];
```

Altenatively

```
inferTetraOrigin[SNPDose, chrsubset, snpsubset, epsF, eps, ploidy,
  outputid, bivalentPhasing → True, bivalentDecoding → False];
tetraResultFile = "TetraOrigin_Output_" <>
    outputid <> "_LinkageGroup1.txt";
summaryFile = StringReplace[tetraResultFile,
    {".txt" → "_Summary.csv"}];
refhaploFile =
  "TetraOrigin_Truevalues_founderhaplo_ExampleData.csv";
saveAsSummaryITO[tetraResultFile, summaryFile, refhaploFile];
```

Referece

- Zheng, C., et al. 2015. Multilocus haplotype reconstruction in outcrossing tetraploids. Manucript.