STARRInIGHTS tutorial administrivia

★Please self-assemble into groups with at least one command line/R-capable laptop (Linux or Mac with Terminal + R)

★Each group will analyze **one** of the **three** biological scenarios:

`/scenario1/STARRI/`

`/scenario2/STARRI/`

`/scenario3/STARRI/`

★Save the files locally in a directory with the **full** path:

`/STARRI/`

(otherwise you'll have to edit the key.txt file)

Human Microbiome DPWG Bioinformatics Tutorial: Introducing STARRInIGHTS

Jesse Shapiro
jesse1@mit.edu
Alm Lab, MIT Computational & Systems Biology
11 March 2011

Goals & Questions in Human Microbiomics

1. Identifying Microbe-Disease associations

   - *moving from 16S to genome-wide studies*
   - *importance of within-species diversity (identical 16S, distinct ecology)*
   - *can we achieve gene/allele specific resolution?*

Goals & Questions in Human Microbiomics

1. Identifying Microbe-Disease associations
    - *moving from 16S to genome-wide studies*
    - *importance of within-species diversity (identical 16S, distinct ecology)*
    - *can we achieve gene/allele specific resolution?*
2. Thinking about microbial populations and gene pools
    - *do disease/healthy microbiomes constitute separate gene pools?*
    - *are microbial populations mostly clonal or sexual?*

Goals & Questions in Human Microbiomics

1. Identifying Microbe-Disease associations
    - *moving from 16S to genome-wide studies*
    - *importance of within-species diversity (identical 16S, distinct ecology)*
    - *can we achieve gene/allele specific resolution?*
2. Thinking about microbial populations and gene pools
    - *do disease/healthy microbiomes constitute separate gene pools?*
    - *are microbial populations mostly clonal or sexual?*
3. Understanding mechanisms of niche adaptation and speciation
    - *which genes are under distinct selective pressures in different environments/niches?*
    - *when is recombination an adaptive event?*

STARRInIGHTS

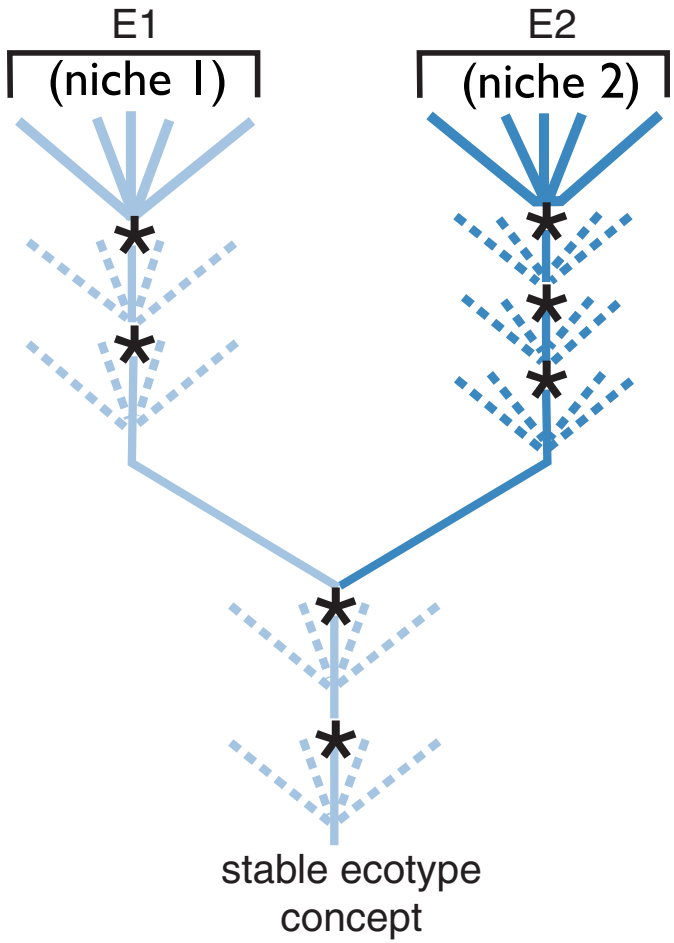
- is a software package for inferring homologous recombination in populations of microbial genomes
- identifies locations of recombination breakpoints in a maximum likelihood, phylogenetic framework
- pinpoints parts of the genome that are strongly associated with ecology of interest (e.g. healthy/sick)
- works best for closely-related, well-aligned genomes

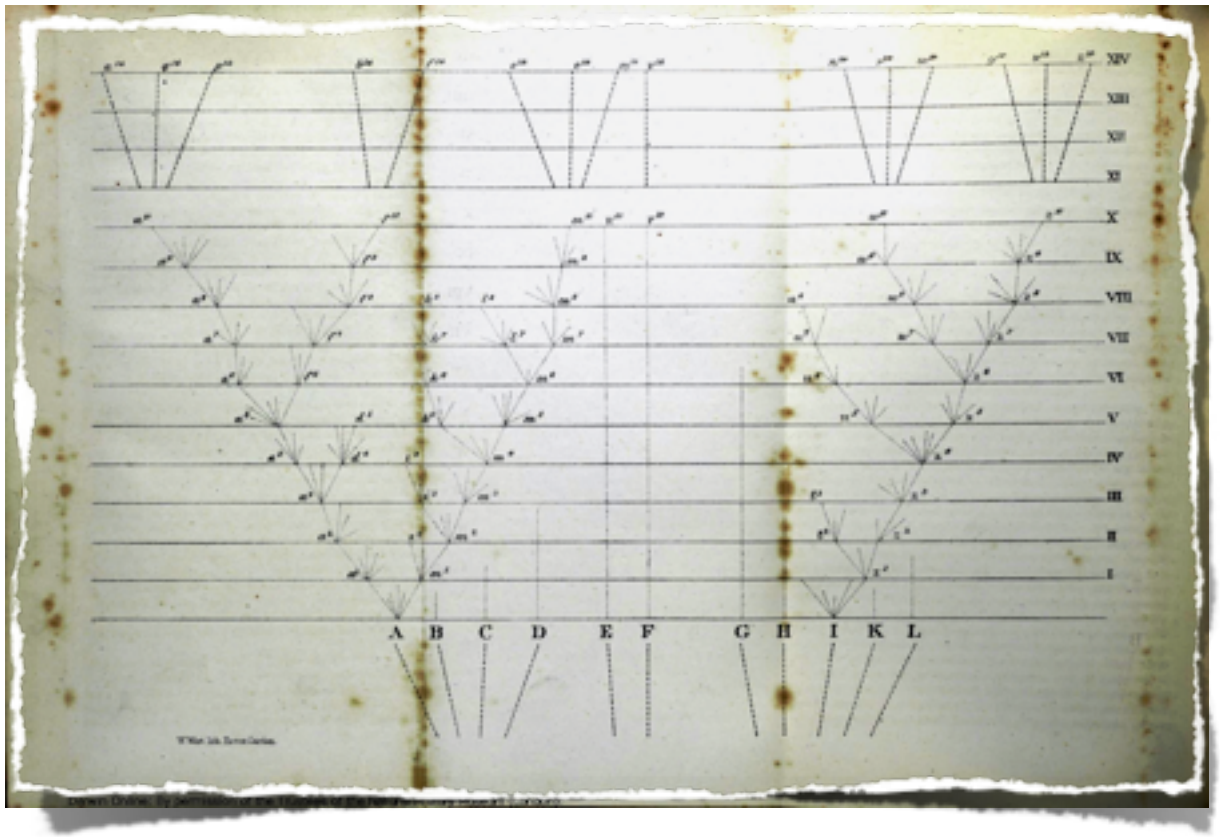
Workshop outline

- Microbial evolution & the importance of recombination
- STARRInIGHTS method description
- Hands-on example
- Discussion
  - Possible uses
  - Limitations
  - Downstream analysis

Microbes: Mechanisms of ecological differentiation

- Clonal expansion ("ecotype" model)





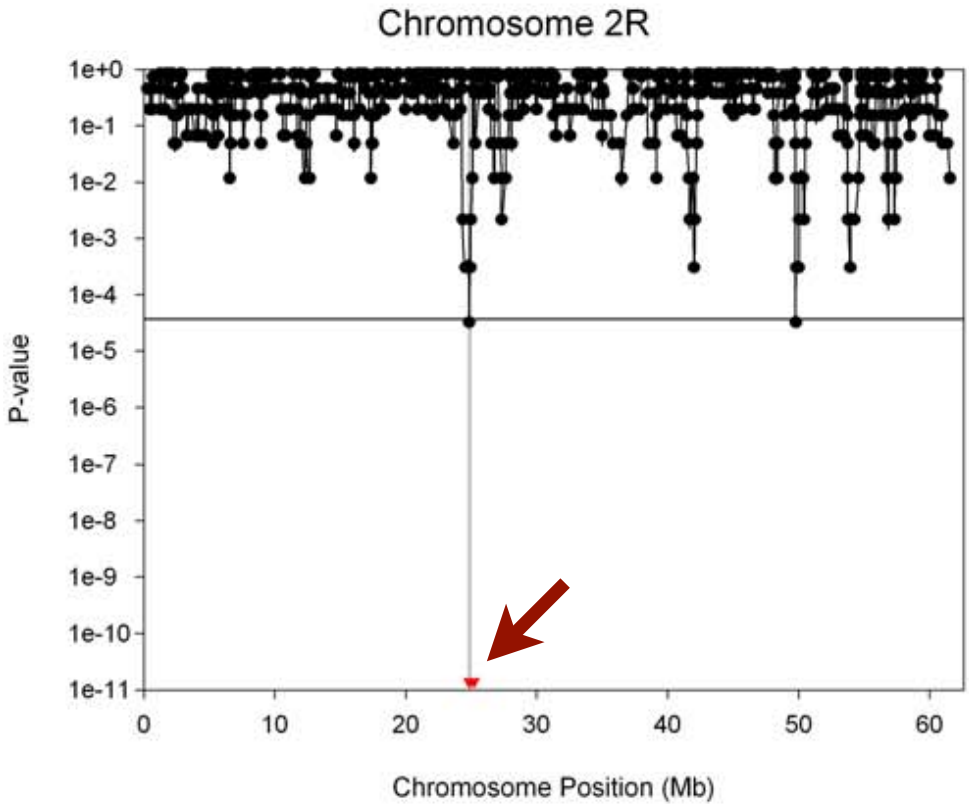
Darwin 1859

Cohan 2001;
Gevers et al. 2005
Fraser et al. 2009

Animals: Mechanisms of ecological differentiation

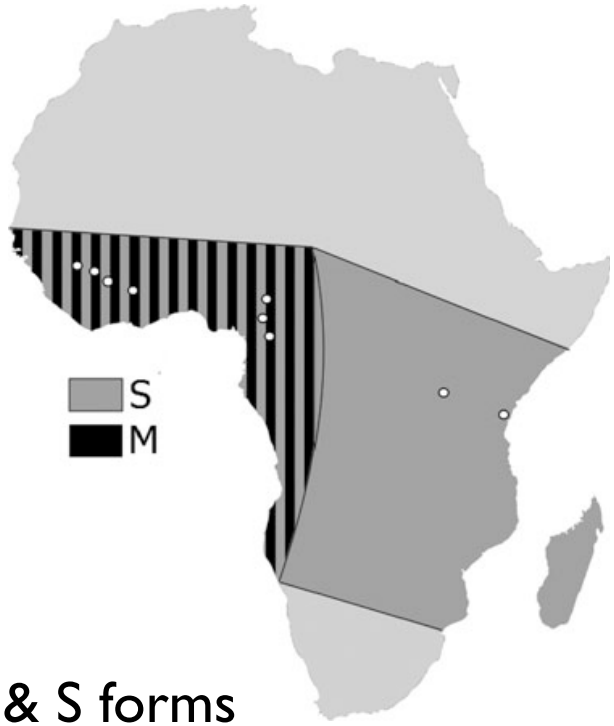
- M and S forms of *Anopheles gambiae*
  - no apparent barriers to gene flow...

... except in 'genomic islands of speciation'



intermixed

divergent between M & S forms



Turner et al. 2005, 2010

White et al. 2010

The importance of recombination

- in bacteria, sex (recombination) is optional: some bacterial recombine a lot, others are clonal
- difficult to pinpoint ecologically important loci or perform tests for natural selection if bacteria don't recombine enough
- recombination may itself be an adaptive event
- (further reading):

Review

Cell PRESS

Evolutionary Microbiology

Looking for Darwin's footprints in the microbial world

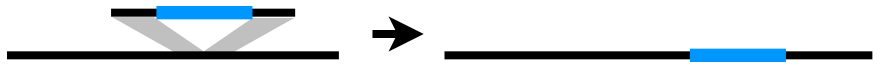
B. Jesse Shapiro[1], Lawrence A. David[1], Jonathan Friedman[1] and Eric J. Alm[1,2,3,4,5]

[1] Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3] Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4] The Virtual Institute of Microbial Stress and Survival, Berkeley, CA 94720, USA
[5] The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

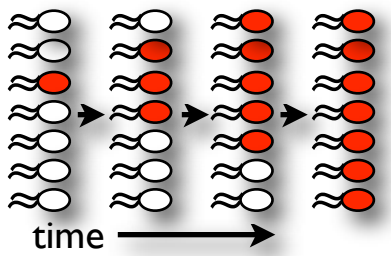
Microbial population genetics terminology

Recombination/sex: *The exchange of a stretch of homologous DNA. Applies to the* ***core*** *genome.*

Illegitimate Recombination (horizontal transfer): *The acquisition of entirely novel genes or operons. Applies to the* ***flexible*** *genome.*



Positive selection: *The evolutionary force favoring adaptive alleles, allowing them to increase in frequency in a population. May lead to differentiation/speciation between populations.*



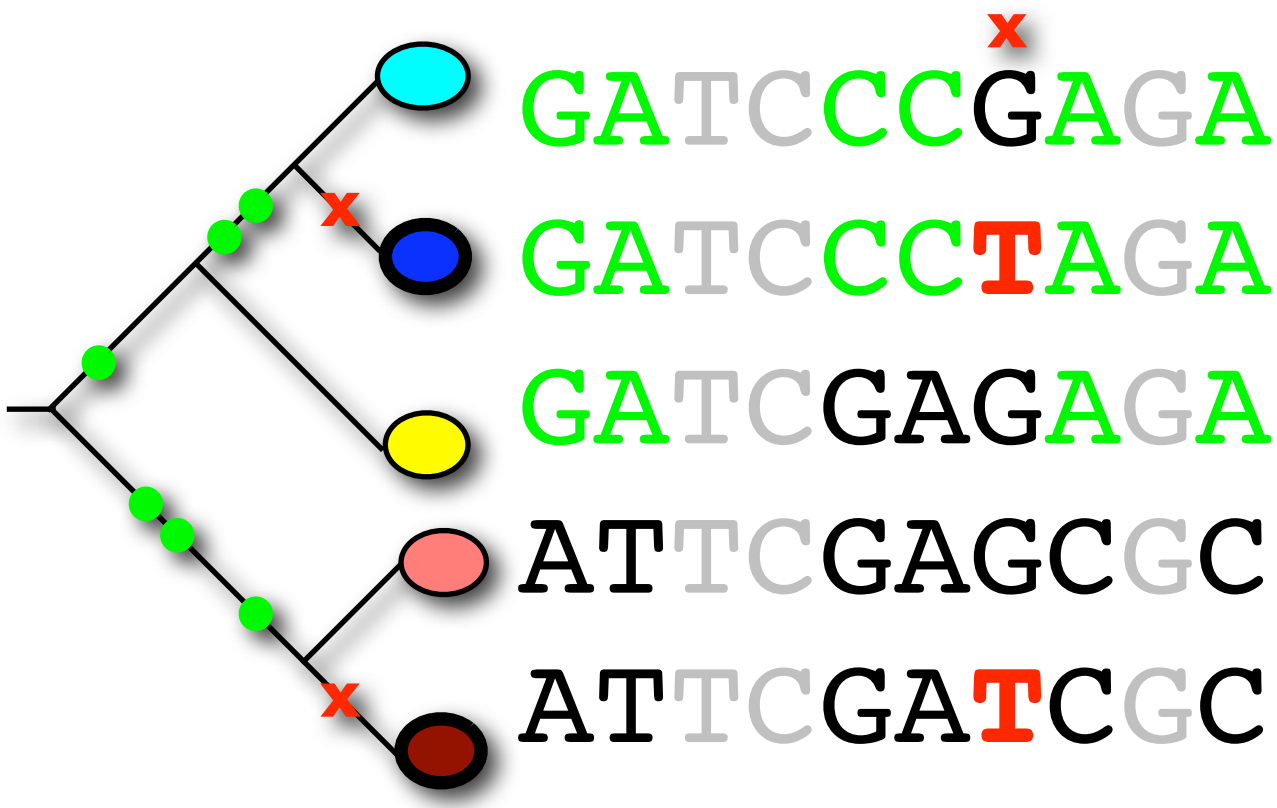
How to detect recombination events?

- assume the genome can be divided into one or more recombinant blocks
- each block can have its own evolutionary history of recombination (its own phylogenetic tree)
- most mutations within a block will support the block's tree, some may reject it (homoplasies)

Finding recombination events in the genome

e.g. 1: Genome consists of 1 block / 1 tree topology

cost = 0 break + 17 homoplasies



x = homoplasic / unparsimonious site

● = SNP (single nucleotide polymorphism) supporting topology **A**

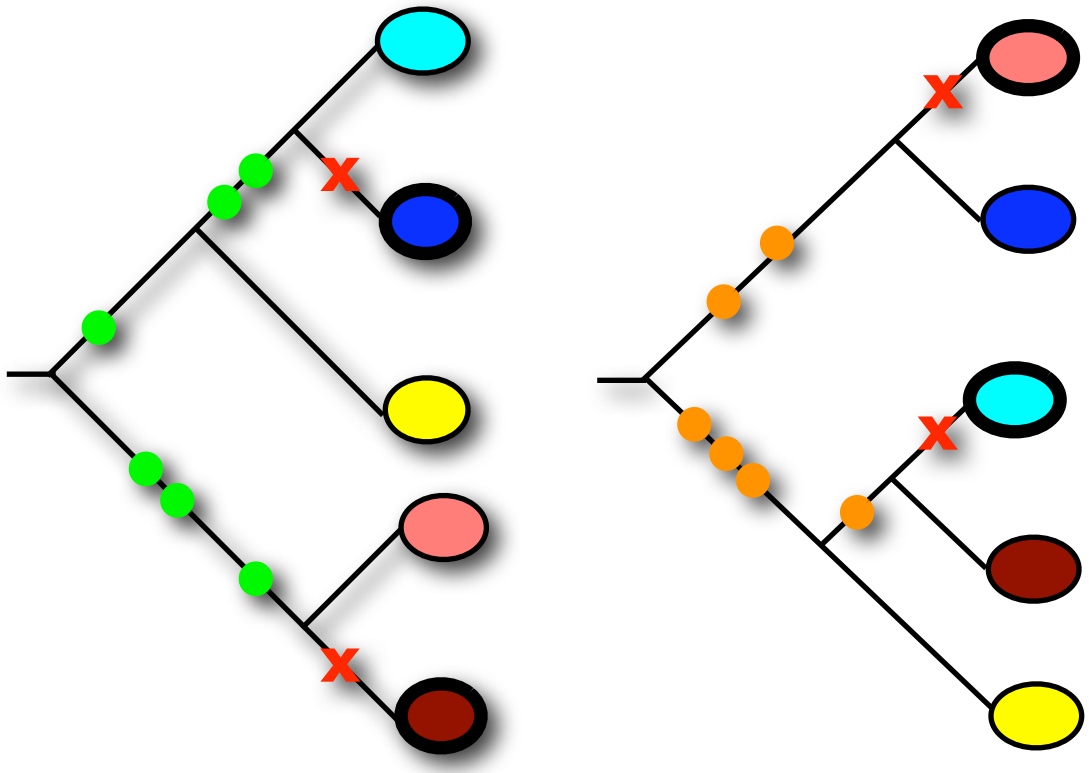
Finding recombination events in the genome

e.g. 1: Genome consists of 1 block / 1 tree topology

cost = 0 break + 17 homoplasies

e.g. 2: Genome consists of 2 blocks / 2 trees

cost = 1 break + 6 homoplasies



x = homoplasic / unparsimonious site

= SNP (single nucleotide polymorphism) supporting topology **A**

= SNP supporting topology **B**

Finding recombination events in the genome

e.g. 1: Genome consists of 1 block / 1 tree topology

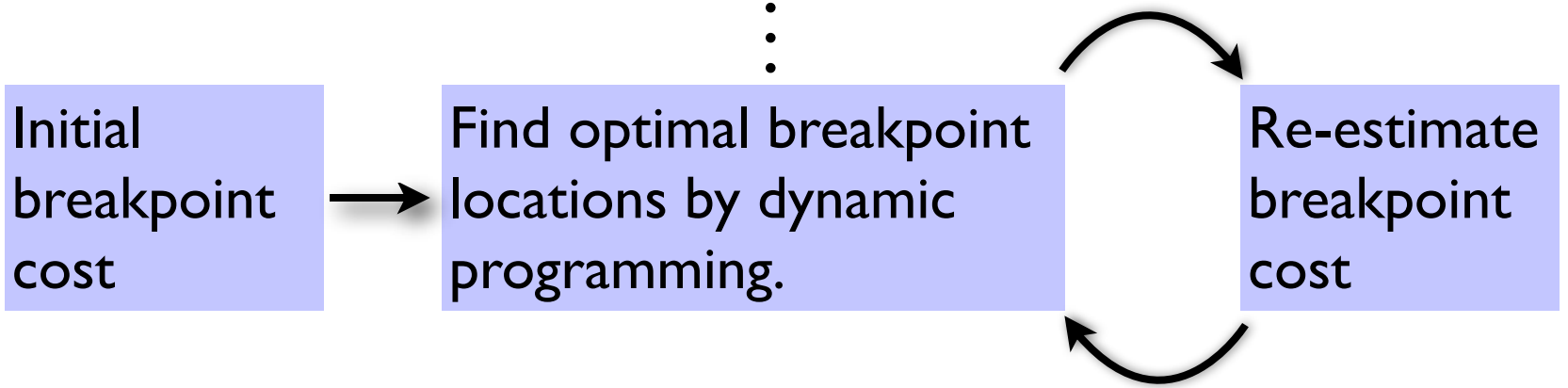
cost = 0 break + 17 homoplasies

e.g. 2: Genome consists of 2 blocks / 2 trees

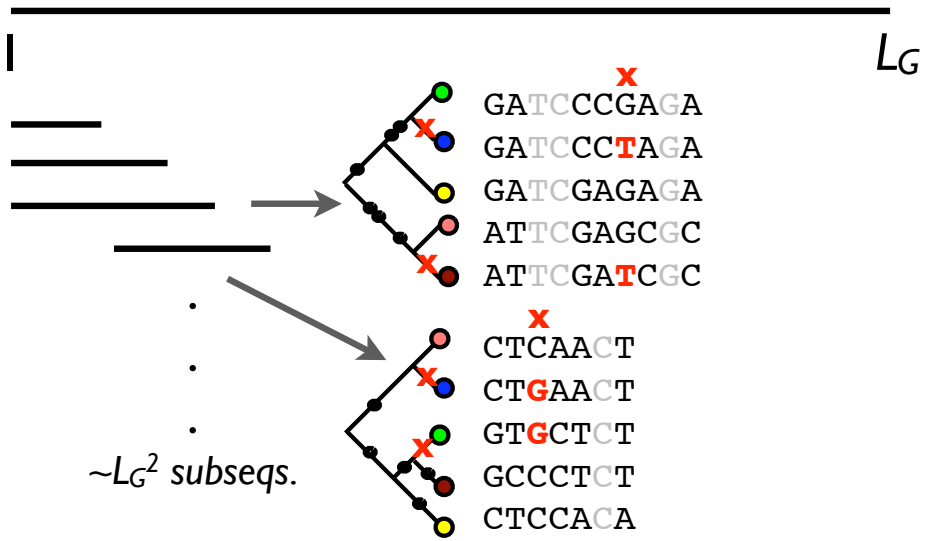
cost = 1 break + 6 homoplasies

e.g. 3: Genome consists of 4 blocks / 4 trees

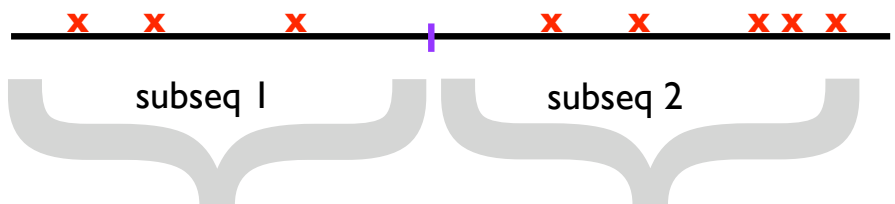
cost = 3 breaks + 2 homoplasies



Finding recombination events in the genome



e.g. *1*. combine 2 subseqs with a breakpoint in between:



$C(1,L_G) = c_b(1) + c_{nb}(L_G\text{-}1) + c_{Tree\ 1} + c_{Tree\ 2}$

1. Consider all subsequences of the core genome (~L^2).

2. Each subsequence gets an ML tree.

x = homoplasic / unparsimonious site

3. Define a cost function for recomb. breakpoints (*b*) and trees in intervening sequences.

$$C(i,j) = c_b \cdot b_{ij} + c_{nb} \cdot (l_{ij} - b_{ij}) + c_{Tree(i,j)}$$

-log probabilities

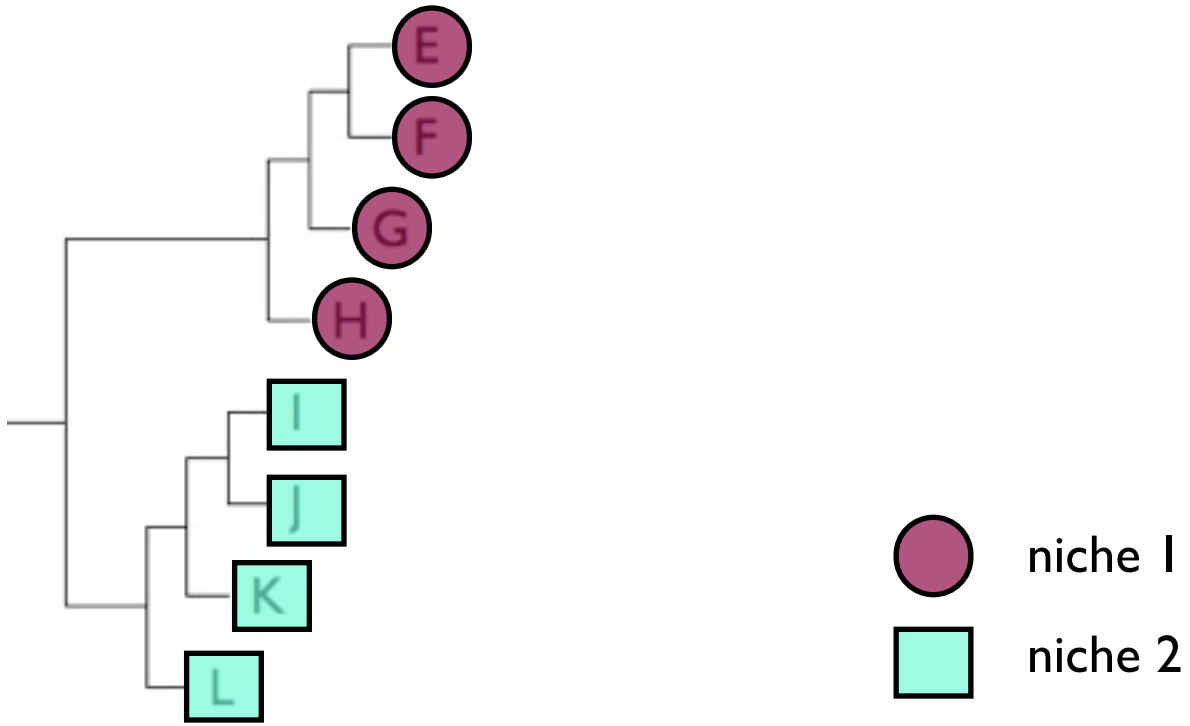
events (breakpoint or not)

4. Find optimal breakpoint locations by dynamic programming (DP). Estimate c_b by Expectation-Maximization.

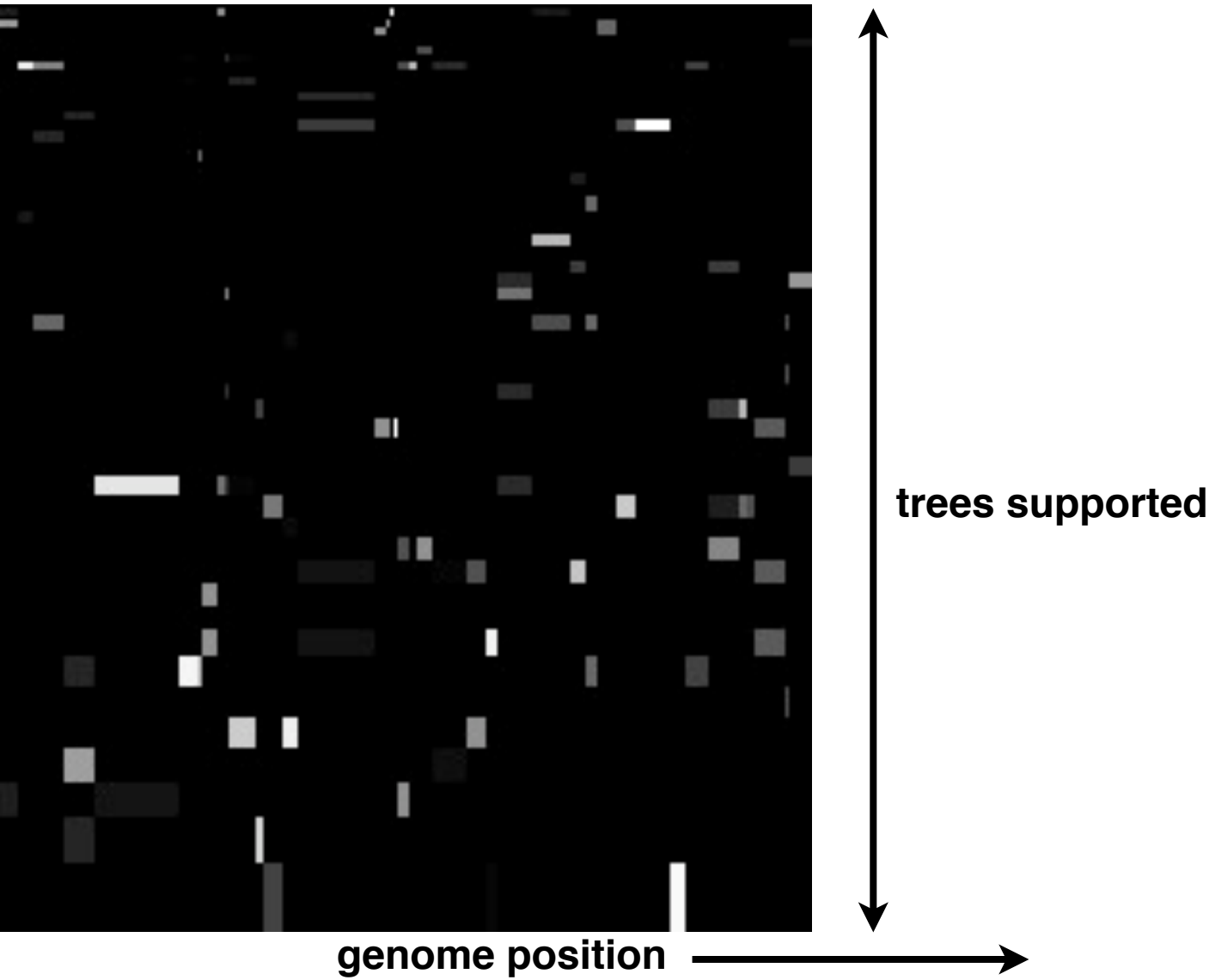
output

- locations of recombination breakpoints, if any, separating recombinant blocks
- each block has its own phylogenetic tree

★ trees may support/reject ecological/biological hypotheses:



output



STARRInIGHTS

Strain-based **T**ree **A**nalysis & **R**ecombinant **R**egion **In**ference **I**n **G**enomes from **H**igh-**T**hroughput **S**equencing-projects

Example STARRInIGHTS analysis

- You have isolated 4 strains from healthy individuals:
  - H1, H2, H3, H4
- 4 from sick patients affected by a disease of interest
  - S1, S2, S3, S4
- Strains are closely-related and genomes are easily aligned
- Questions:
  - How much recombination occurs among these strains?
  - Are H and S gene pools separate or mixed?
  - Can we pinpoint disease-associated loci?

Example STARRInIGHTS analysis

2. Maximum-likelihood (ML) trees have been pre-computed based on every possible subsequence (i,j) of informative SNPs in the core genome. You can view these trees and their associated log-likelihoods in the file **lk.txt**.

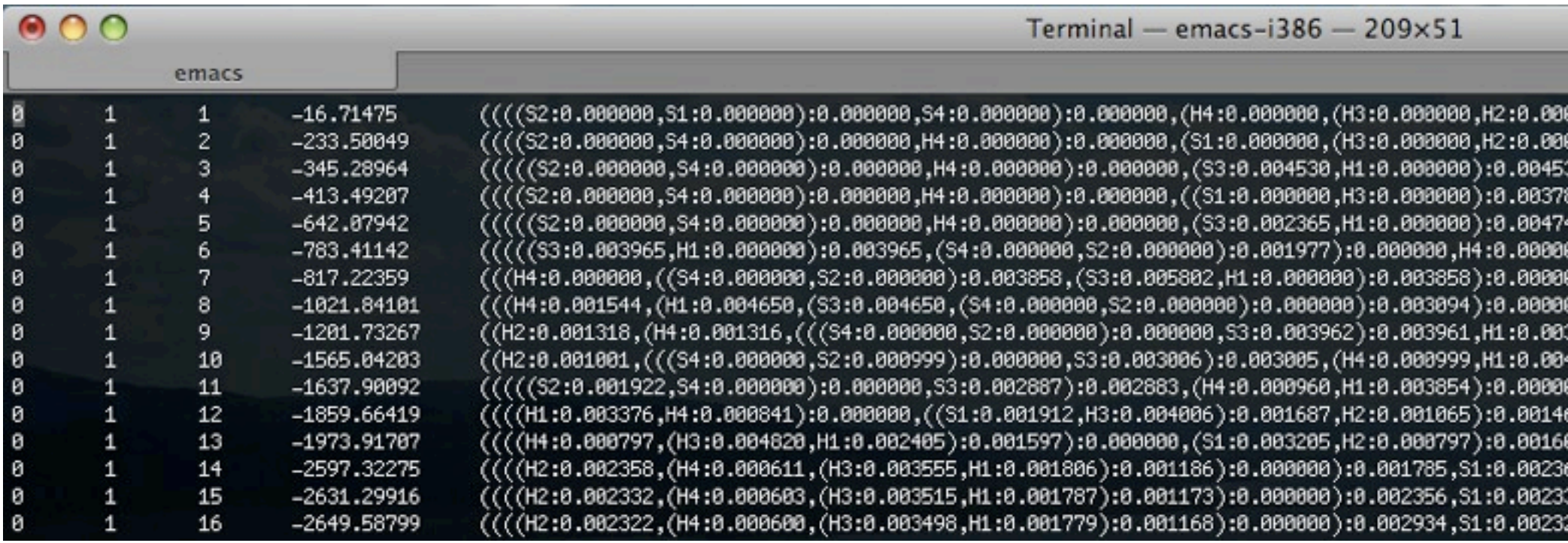
*Note: for real genomes (~5 Mbp, ~50,000 informative SNPs), would have to build ~millions of trees using parallel computing

Example STARRInIGHTS analysis

2. Maximum-likelihood (ML) trees have been pre-computed based on every possible subsequence (i,j) of informative SNPs in the core genome. You can view these trees and their associated log-likelihoods in the file **lk.txt**.

Do the trees for different subsequences appear to be the same or different?

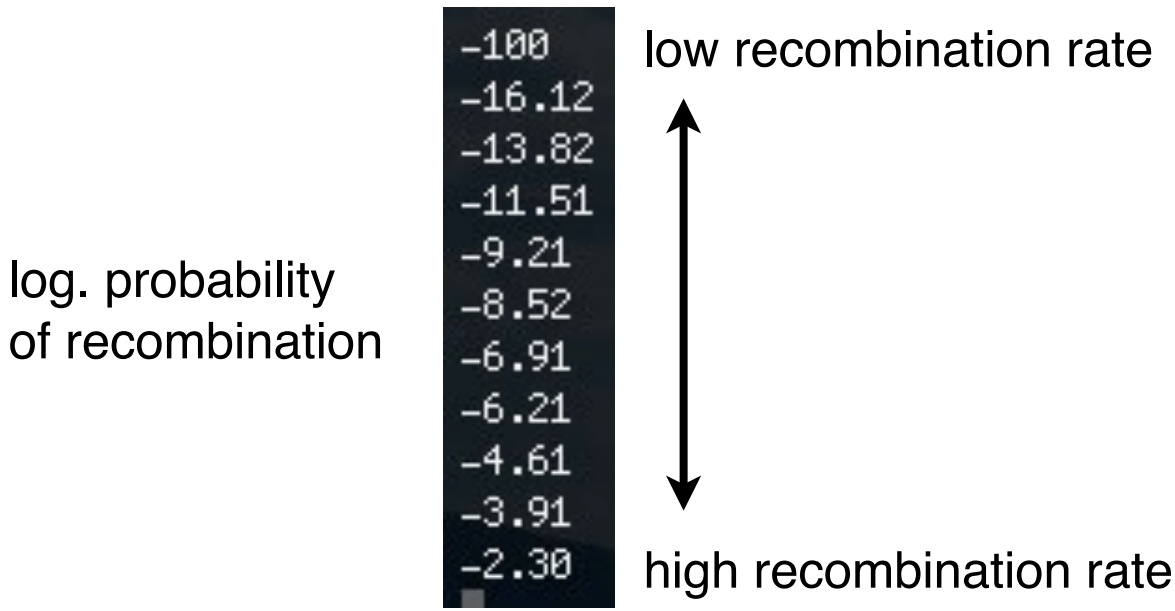
Tree viewer:
http://itol.embl.de/



Example STARRInIGHTS analysis

Combine the subsequences using dynamic programming and a range of recombination breakpoint penalties (found in the file pB.list) by running:

```
perl dp.pl /STARRI/ pB.list > dp.screenout
```



Example STARRInIGHTS analysis

Combine the subsequences using dynamic programming and a range of recombination breakpoint penalties (found in the file pB.list) by running:

```
perl dp.pl /STARRI/ pB.list > dp.screenout
```

You can print a summary of the results using different breakpoint penalties by doing:

```
perl findML_noEM.pl starri.init.pB_-
```

How do different pB settings affect the results?
What is the most likely number of recombination breakpoints? number of events?

Example STARRInIGHTS analysis

How do different pB settings affect the results?
What is the most likely number of recombination breakpoints? number of events?

Scenario 1:

pBinit	pB	like	nb
-100	na	-3835.62158	0
-11.51	na	-3733.79988325205	4
-13.82	na	-3743.01733482071	4
-16.12	na	-3752.21509912514	4
-2.30	na	-3975.89919761141	8
-3.91	na	-3760.29254717745	8
-4.61	na	-3731.13901498	4
-6.21	na	-3717.59495168907	4
-6.91	na	-3717.86649669679	4
-8.52	na	-3722.31270045286	4
-9.21	na	-3724.82454746064	4

Scenario 2:

like	nb
-3643.61762	0
-3643.64269336944	0
-3643.62010880279	0
-3643.61786952438	0
-3896.77155393373	6
-3693.43366524902	6
-3666.14762498	4
-3647.75420546514	1
-3645.92271146446	1
-3644.11626829013	0
-3643.86771761682	0

Scenario 3:

like	nb
-3638.01	0
-3638.03	0
-3638.01	0
-3638.01	0
-3890.00	5
-3683.44	4
-3660.68	4
-3641.44	1
-3639.60	1
-3638.51	0
-3638.26	0

Intuitively, higher breakpoint probabilities result in more breakpoints!

Best just to choose a conservative value of pB to get high-confidence breakpoints only?

Example STARRInIGHTS analysis

Repeat, but letting E-M converge on an optimal (maximum-likelihood) value of pB from different starting values:

```
perl dp+em.pl /STARRI/ pB.list > dp+em.screenout

perl findML.pl starriEM.init.pB_
```

Does the E-M converge?

What is the likeliest number of breakpoints in the genome?

Which value of pB provides the best (maximum likelihood) inferences of recombination breakpoints?

Example STARRInIGHTS analysis

Repeat, but letting E-M converge on an optimal (maximum-likelihood) value of pB from different starting values:

```
perl dp+em.pl /STARRI/ pB.list > dp+em.screenout

perl findML.pl starriEM.init.pB_
```

Does the E-M converge?

What is the likeliest number of breakpoints in the genome?

Scenario 1: nb = 4
Scenario 2: nb = 0
Scenario 3: nb = 0 (but 1 is nearly as likely!)

Which value of pB provides the best (maximum likelihood) inferences of recombination breakpoints?

Scenario 1: pB = -6.44
Scenario 2: pB = -16.12
Scenario 3: pB = -16.12

Example STARRInIGHTS analysis

3. Using the value of pBinit that converges on the best pB (call this `[best_pBinit]`, let's visualize the data and see which parts of the genome, if any, are associated with disease.

```
perl parseBlockParsTrees.pl /STARRI/key.txt /STARRI/ [best pBinit]
STARRI.2.0.EM
```

Now we will use this output to make a final summary of our results and plot it visually:

```
perl parseBlockIngroupStats.pl /STARRI/key.txt /STARRI/ [best
pBinit] STARRI.2.0.EM pars.snp.pB_[best pBinit].txt healthy.txt
sick.txt
```

This produces our final results file:

```
trees.pB_[best_pBinit]+stats.ingr.healthy.txt.txt
```

Example STARRInIGHTS analysis

This produces our final results file:

`trees.pB_[best_pBinit]+stats.ingr.healthy.txt.txt`

Scenario 1:

blk	start	end	parsSNP	homoSNP	domGrp	domFrac	parsTree	Nintree
1	1	424	5	0	H1.S3.	0.400	H1.S3._H2.H3.S1._H3.S1._	2_1_2_
2	425	766	4	0	S1.S2.S3.S4.	0.500	H3.H4._S1.S2.S3.S4._S2.S3.S4._	1_2_1_
3	767	1266	4	0	H1.H3.S2.	0.500	H1.H3.S2._H2.S1._H3.S2._	2_1_1_
4	1267	1935	4	0	S1.S2.S3.S4.	1.000	S1.S2.S3.S4._	4_
5	1936	2275	4	0	H1.H4.S1.S4.	1.000	H1.H4.S1.S4._	4_

Scenario 2:

blk	start	end	parsSNP	homoSNP	domGrp	domFrac	parsTree	Nintree
1	1	2224	18	0	H1.S3.	0.389	H1.S3._H2.H3.S1._H3.S1._H4.S2._	7_1_7_3_

Scenario 3:

blk	start	end	parsSNP	homoSNP	domGrp	domFrac	parsTree	Nintree
1	1	2407	12	0	S1.S2.S3.S4.	0.333	H1.H2._H3.H4._S1.S2.S3.S4._S2.S3.S4._S3.S4._	2_2_4_3_1_

Example STARRInIGHTS analysis

This produces our final results file:

`trees.pB_[best_pBinit]+stats.ingr.healthy.txt.txt`

Scenario 1: **5 recombinant blocks; 2 disease-associated regions**

blk	start	end	parsSNP	homoSNP	domGrp	domFrac	parsTree	Nintree
1	1	424	5	0	H1.S3.	0.400	H1.S3._H2.H3.S1._H3.S1._	2_1_2_
2	425	766	4	0	S1.S2.S3.S4.	0.500	H3.H4._S1.S2.S3.S4._S2.S3.S4._	1_2_1_
3	767	1266	4	0	H1.H3.S2.	0.500	H1.H3.S2._H2.S1._H3.S2._	2_1_1_
4	1267	1935	4	0	S1.S2.S3.S4.	1.000	S1.S2.S3.S4._	4_
5	1936	2275	4	0	H1.H4.S1.S4.	1.000	H1.H4.S1.S4._	4_

Scenario 2: **clonal evolution; no disease association**

blk	start	end	parsSNP	homoSNP	domGrp	domFrac	parsTree	Nintree
1	1	2224	18	0	H1.S3.	0.389	H1.S3._H2.H3.S1._H3.S1._H4.S2._	7_1_7_3_

Scenario 3: **clonal evolution, with disease association**

blk	start	end	parsSNP	homoSNP	domGrp	domFrac	parsTree	Nintree
1	1	2407	12	0	S1.S2.S3.S4.	0.333	H1.H2._H3.H4._S1.S2.S3.S4._S2.S3.S4._S3.S4._	2_2_4_3_1_

Example STARRInIGHTS analysis

directory containing R code snippets to make plots:

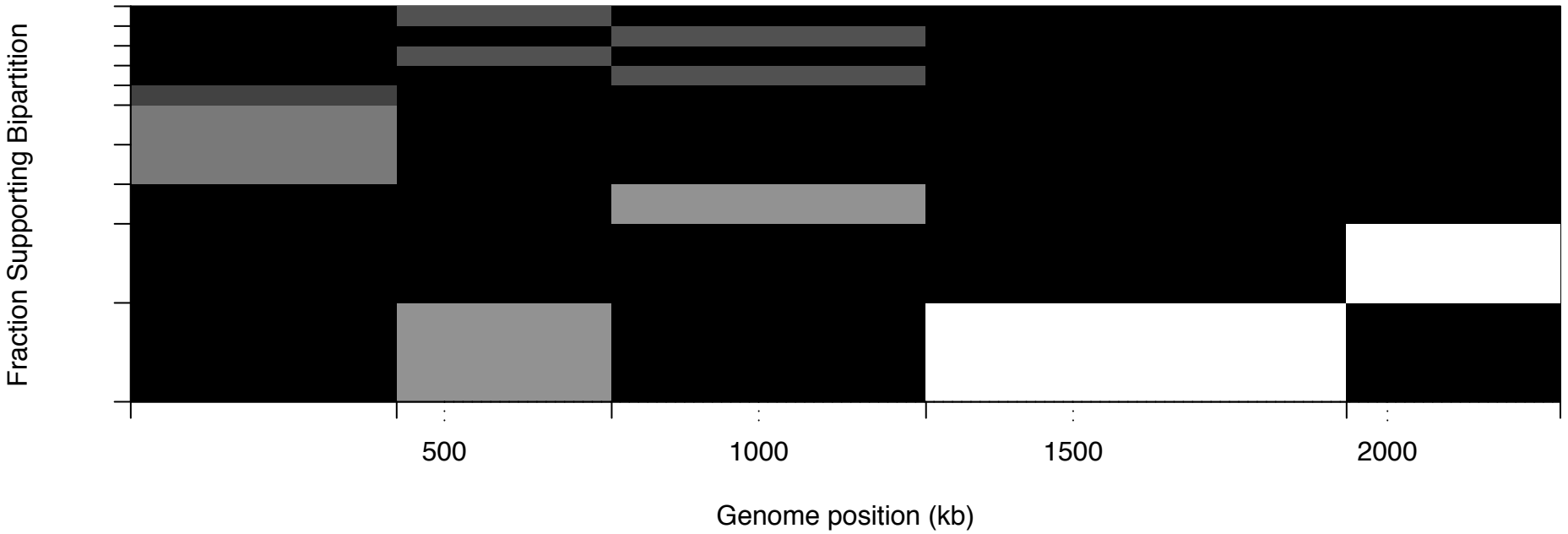
`output.pB_[best pBinit]/`

You can plot a heatmap of support for different tree partitions across the genome by pasting `[#].support.txt` into an R command window.

Brighter shades or white/gray indicate stretches of the genome supporting different phylogenetic partitioning of strains (shown on the Y-axis; see `[#].Ykey.txt`).

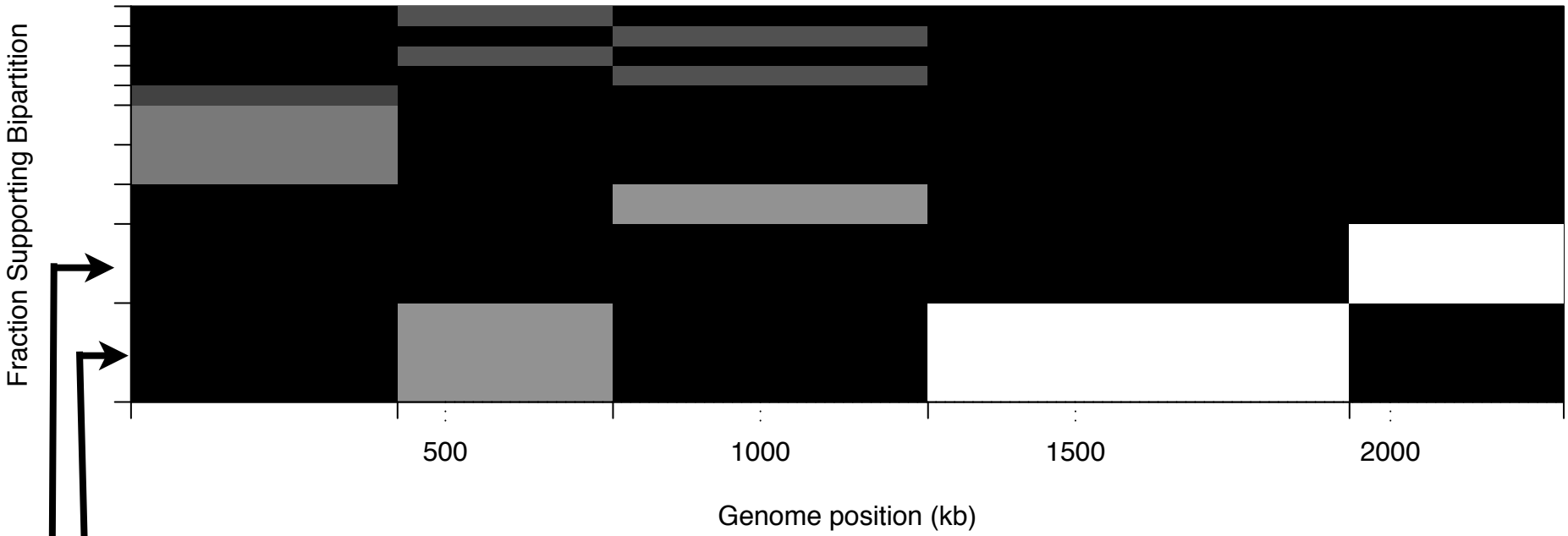
Example STARRInIGHTS analysis

Scenario 1: **5 recombinant blocks; 2 disease-associated regions**



Example STARRInIGHTS analysis

Scenario 1: **5 recombinant blocks; 2 disease-associated regions**

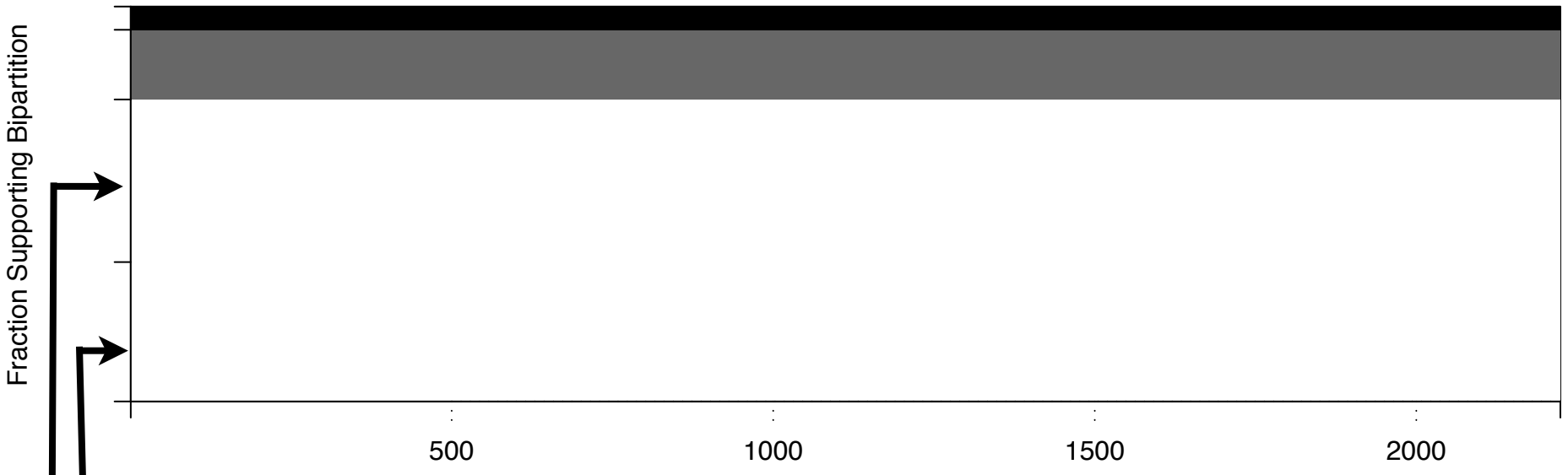


# SNPS	y1	y2	tree partition	
6	0	6	S1.S2.S3.S4.	H1.H2.H3.H4.
4	6	10	H1.H4.S1.S4.	H2.H3.S2.S3.
2	10	12	H1.H3.S2.	H2.H4.S1.S3.S4.
2	12	14	H1.S3.	H2.H3.H4.S1.S2.S4.
2	14	16	H3.S1.	H1.H2.H4.S2.S3.S4.
1	16	17	H2.H3.S1.	H1.H4.S2.S3.S4.
1	17	18	H2.S1.	H1.H3.H4.S2.S3.S4.
1	18	19	H3.H4.	H1.H2.S1.S2.S3.S4.
1	19	20	H3.S2.	H1.H2.H4.S1.S3.S4.
1	20	21	S2.S3.S4.	H1.H2.H3.H4.S1.

1.Ykey.txt

Example STARRInIGHTS analysis

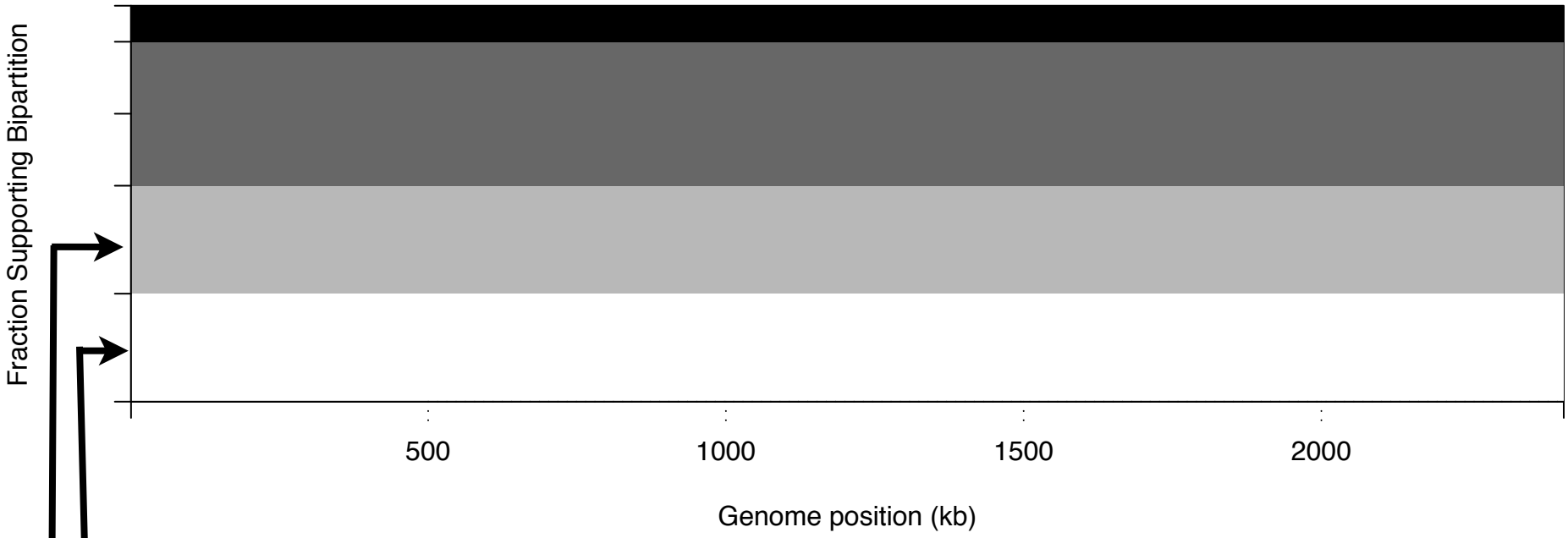
Scenario 2: **clonal, no disease association**



# SNPS	y1	y2	tree partition	
7	0	7	H1.S3.	H2.H3.H4.S1.S2.S4.
7	7	14	H3.S1.	H1.H2.H4.S2.S3.S4.
3	14	17	H4.S2.	H1.H2.H3.S1.S3.S4.
1	17	18	H2.H3.S1.	H1.H4.S2.S3.S4.

Example STARRInIGHTS analysis

Scenario 3: **clonal, with disease association**



# SNPS	y1	y2	tree partition	
4	0	4	S1.S2.S3.S4.	H1.H2.H3.H4.
3	4	7	S2.S3.S4.	H1.H2.H3.H4.S1.
2	7	9	H1.H2.	H3.H4.S1.S2.S3.S4.
2	9	11	H3.H4.	H1.H2.S1.S2.S3.S4.
1	11	12	S3.S4.	H1.H2.H3.H4.S1.S2.

Example STARRInIGHTS analysis

Is there any support for association between microbes and disease state (healthy/sick)?

Is association localized to certain genomic regions?

If so, what downstream analyses would you perform on these regions?

If not, can you suggest another strategy to pinpoint genes/mutations associated with disease?

Example STARRInIGHTS analysis

Is there any support for association between microbes and disease state (healthy/sick)?

For scenarios 1 and 3, yes.

Is association localized to certain genomic regions?

In scenario 1, two regions show disease association.

If so, what downstream analyses would you perform on these regions?

Gene-finding, tests for positive selection (*e.g.* McDonald-Kreitman test, dN/dS)

If not, can you suggest another strategy to pinpoint genes/mutations associated with disease?

Convergence tests: Are certain mutations repeatably associated with disease? (for example, in another sample taken at a different time or geographic location).

Discussion

1. Biological limitations

    - *breakpoints ≠ events*

2. Technical limitations

    - *runtime scales with genetic diversity*
    - *will usually require parallel computing to pre-compute ML trees*
    - *need to correct for model complexity (each new breakpoints adds more parameters to the model). Do so empirically, or just choose a conservative pB.*

Discussion

1. Biological problems addressed:

   - how common is recombination?
   - does recombination cross niche boundaries?

     - implications for species concepts

   - which recombinant loci are associated with meta-data of interest

     (meta-data could include disease information, environmental variables, geography)

2. Many possible uses:

   - exploratory analysis of closely related genomes (what are the dominant phylogenetic groupings?)
   - hypothesis-driven analysis (*e.g.* disease vs. healthy associations)

Ongoing work

1. Projects:
   - *Vibrio*: associated with different marine particles
   - *S. pneumoniae:* virulent/avirulent strains
   - *E. faecalies*: antibiotic resistant/sensitive strain

2. Future work:
   - web interface?
   - cloud?

Software & documentation

http://almlab.mit.edu/ALM/star.html

Acknowledgments

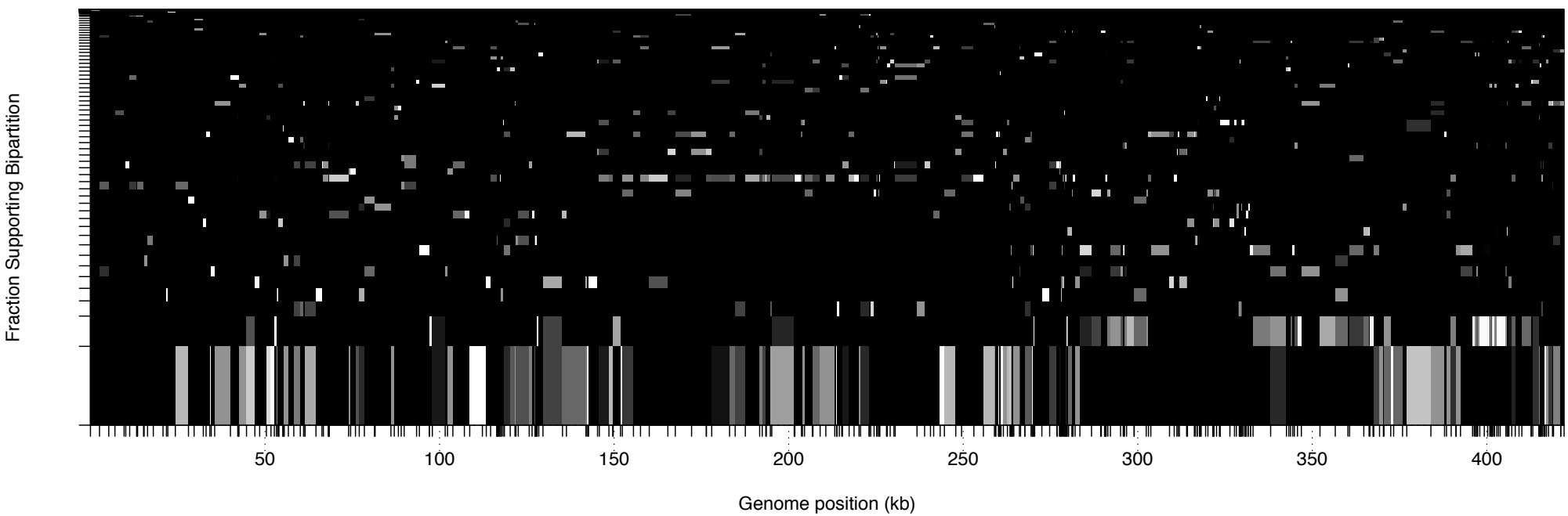
Eric Alm
Jonathan Friedman
Otto Cordero

Martin Polz
John Wakeley
Pardis Sabeti

Funding

Feedback/Questions:
jesse1@mit.edu

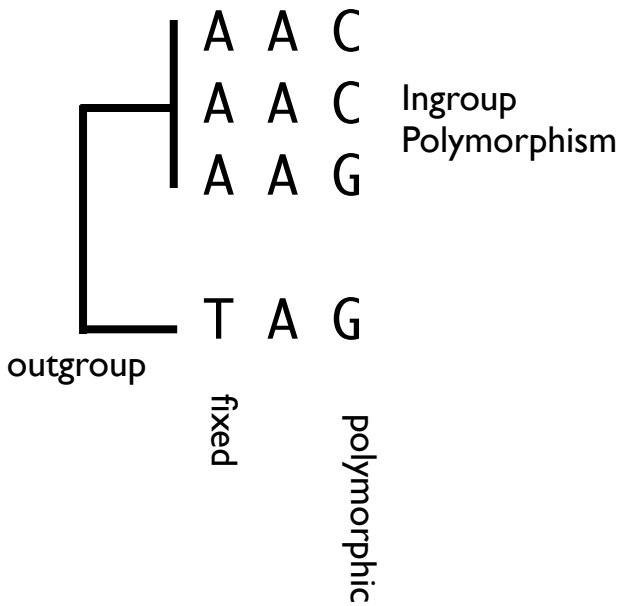
extra


Fraction Supporting Bipartition
50
100
150
200
250
300
350
400
Genome position (kb)

McDonald-Kreitman test for positive selection

- Has there been an excess of fixed amino acid changes between the 2 ecological groups?

- Compare *fixed A/S* to *polymorphic A/S*:



SNPs
Fixed between ingroup & outgroup
Polymorphic in ingroup

change **A**A	**S**ilent	**A**/**S**
4	2	2
1	2	0.5

'fixation index' = 2 / 0.5 = 4

Convergence test