Outcrossing

Yu Huang

October 10, 2008

Abstract

Contents

1 Introduction

Arabidopsis thaliana is a highly selfing species, with a very small portion resulting from outcrossing. An early electrophoretic study based on polymorphic enzymes[1] put the outcrossing rate to be less than 0.3%.

Given the polymorphism data of 149 SNPs in 5720 strains, we set out to improve the estimate of the outcrossing rate in terms of accuracy and resolution.

2 Data

The strains were collected worldwide by numerous scientists including Eric Holub, Diane Myers, Justin Borevitz, Jon Agen, Megan Dunning (top 5 collectors) etc. The polymorphism data was generated by Sequenom, Inc. The 149 SNPs were picked based on an earlier work involving PCR sequencing[10]. Figure 1 shows the spacing of 149 SNPs. Total 6389 runs of genotyping. 669 of them are technical duplications of 640 strains. So end up with 5720 strains based on ecotypeid. The 5720 strains were collected from around the world (Figure 2).

Further, Remove 27 strains with all-NA data. Remove 14 strains with no GPS info.

So ends up with 5679 strains.

2.1 tables showing intricate relationship between these strains

Table 1: Cross (nativename,stkparent) to ecotypeid duplicated times, ecotypeid-with-all-NA, ecotypeid-with-no-gps

nativename	stkparent	ecotypeid duplicate					ecotypeid-with-all-NA					ecotypeid-with-no-gps				
1	4925	4925	0	0	0	0	1	0	0	0	0	14	0	0	0	0
2	693	158	614	0	0	0	23	23	0	0	0	0	0	0	0	0
3	32	4	4	28	0	0	1	2	0	0	0	0	0	0	0	0
5	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0

Table 2: (nativename,stkparent) duplicates after ecotypeid-with-all-NA removal

nativename	stkparent
0	24
1	4949
2	648
3	29
4	1

Table 3: (nativename,stkparent) duplicates after ecotypeid-with-no-gps removal

nativename stkparent	
0	14
1	4911
2	693
3	32
5	1

Table 4: (nativename,stkparent) duplicates after ecotypeid-with-all-NA removal, cross it to ecotypeid duplicated times, ecotypeid-with-no-gps

nativename stkparent all-NA removal		ecotypeid duplicate				ecotypeid-with-no-gps			
0	24	0	0	0	0	0	0	0	0
1	4949	4949	0	0	0	14	0	0	0
2	648	158	569	0	0	0	0	0	0
3	29	4	4	25	0	0	0	0	0
4	1	1	0	1	0	0	0	0	0

Table 5: (nativename,stkparent) duplicates after all-NA and no-gps removal

nativename stkparent	
0	38
1	4935
2	648
3	29
4	1

Table 6: overlapping between all weirdos

all-NA runs	all-NA ecotypeid	no-genotyping	no-gps	all-NA and no-genotyping	all-NA and no-gps	no-genotyping a
76	51	899	195	0	0	181

3 inconsistency in duplicated calls

The duplication arises in two situations. 1. For one strain, several runs of genotyping was carried out due to either failure of earlier experiments or lack of communication between different labs. 2. Several strains were recorded as different strains in the database.

Mentioned earlier, 640 strains with same ecotypeid has duplicated calls on one snp locus (1st kind). 89 strains out of 5679 different ecotypeid strains have same nativename and stockparent (2nd kind), which adds more to duplicated calls.

Among $(640 + 89) * 149 = 108621$ genome positions, 749 (ratio=0.00689553585402) show inconsistency (more than 1 non-NA calls). A simple voting scheme is used to resolve these duplicated inconsistent call.

Final data matrix is 5590 strains by 149 snps. 10.6% of this matrix is NA. Figure 3 is what matrix looks like after coding different calls into integers.

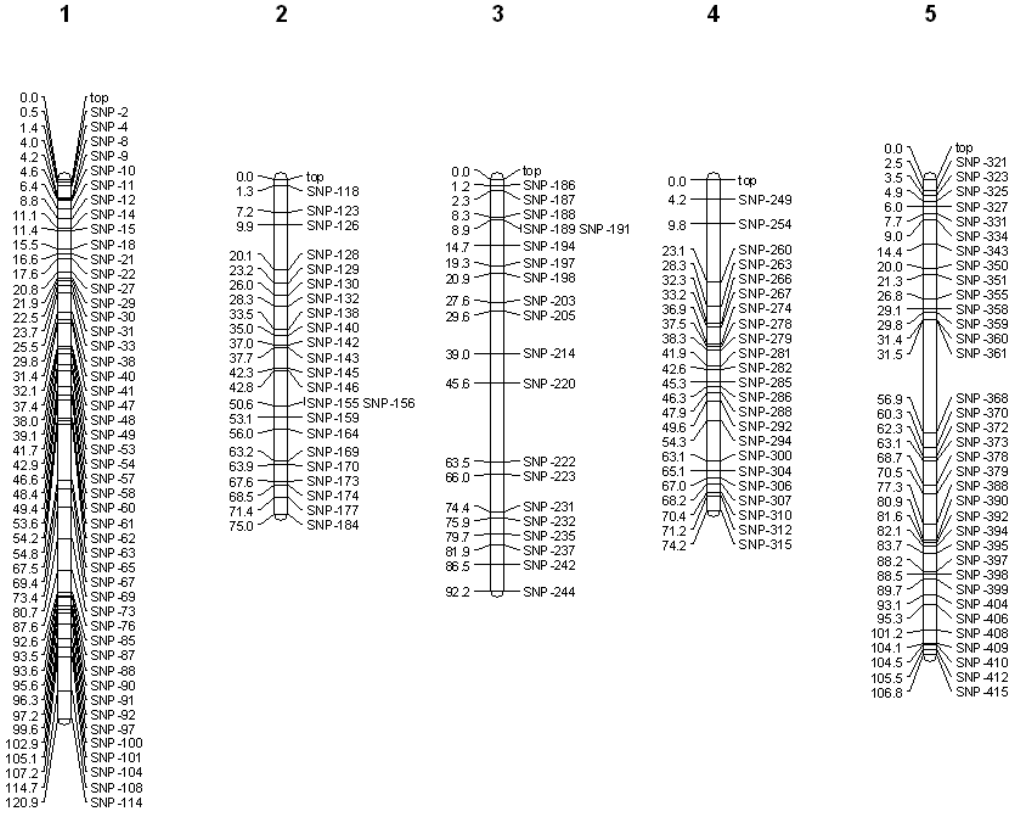

Figure 1: location of 149 snps on chromosomes. the number on the left of chromosome is centimorgan distance?. the label on the right is the snp id.

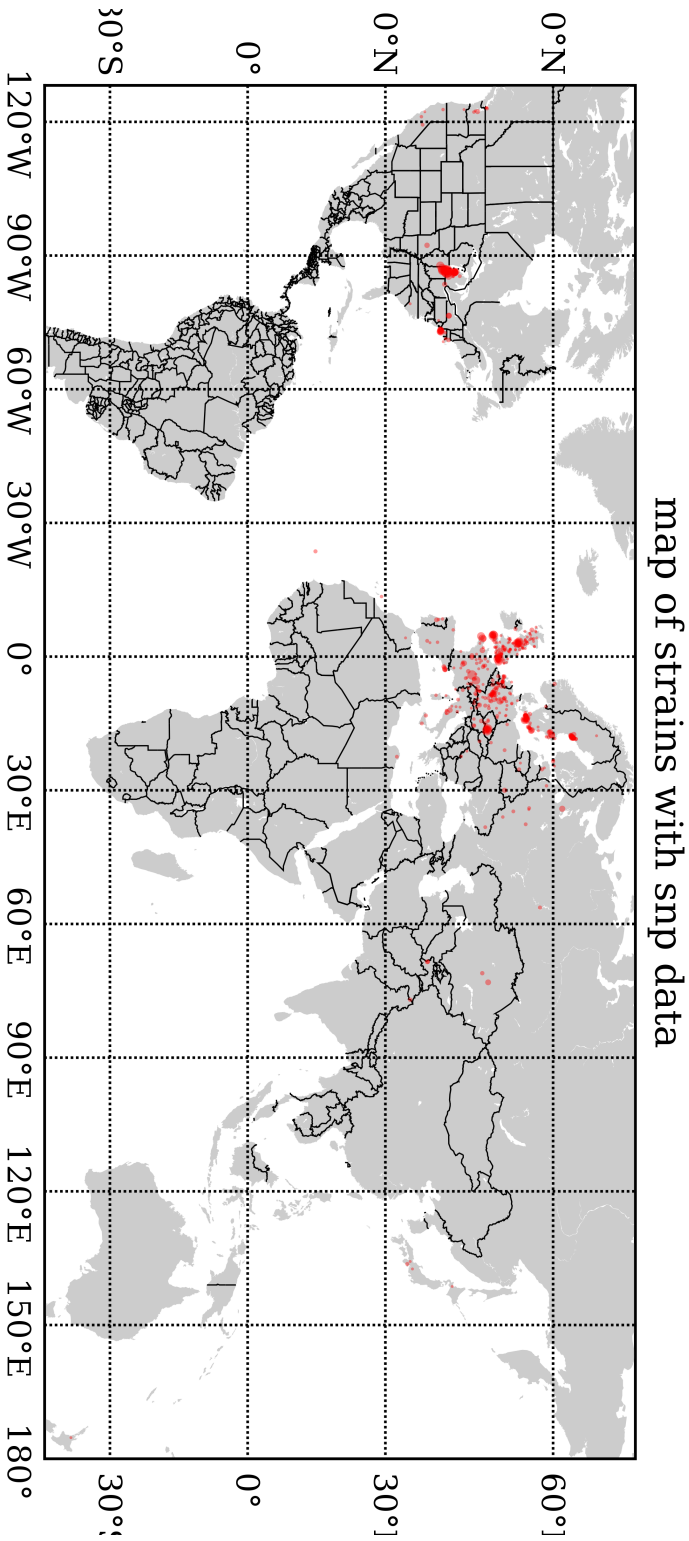

Figure 2: worldwide distribution of strains. the size of the red circle corresponds to the number of strains from that region.

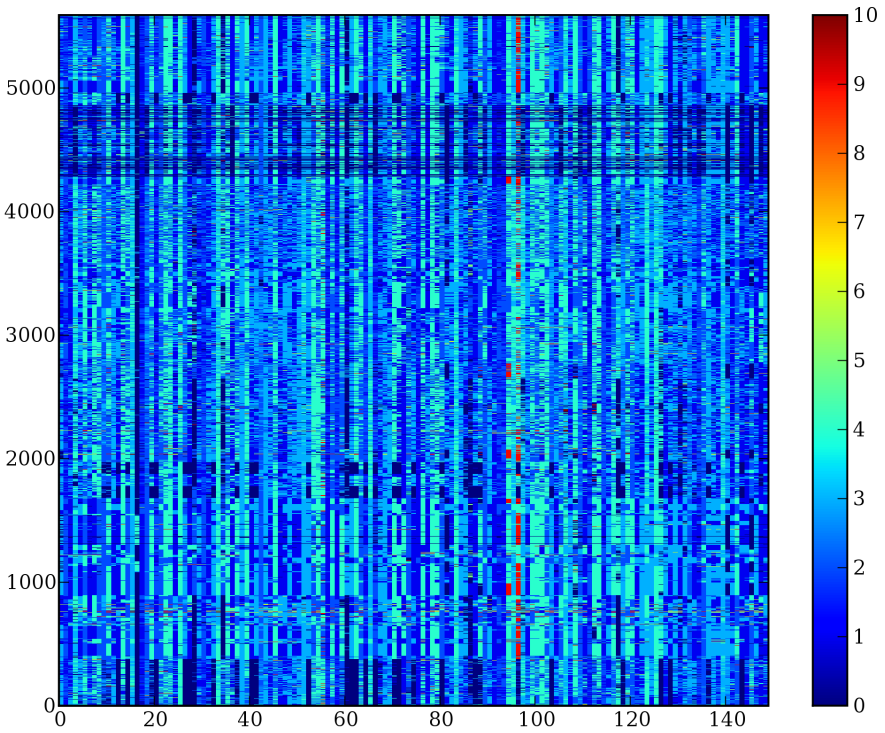

Figure 3: 0 is NA. 1-4 are ACGT. 5-10 are heterozygous calls.

4 Genotyping Error Rate

The 96 strains from [10] were genotyped again, which offered an opportunity to have a rough estimate of genotyping error rate by comparing the data from two different experiments. Figure 4 is an overview of all the differences.

The optimistic error rate (not counting the NAs) is 121/(12408+121) = 0.96%. Taking the NAs into account, the error rate would be (121+893+804)/(12408.0+121+893+804) = 12.78%.

5 NA Filtering

Figure 5 is the Strain by SNP matrix after being sorted. A bunch of NA-rich strains are shown on top of the figure.

Figure 6 is a histogram of NA percentage in all SNPs. 5 SNPs with $\geq 40\%$ NAs were removed. Figure 7 is a histogram of NA percentage in all strains. 194 strains with $\geq 40\%$ NAs were removed.

6 Bogus Heterozygous Calls(column-wise)

From figure 3 or figure 5, there're columns with an unreasonable whopping number of heterozygous calls.

Check figure 8 for the histogram of heterozygosity of each strain. Among 5602-194=5408 strains, 2407 (?) strains have no heterozygous calls. 1961 (?) strains have only one heterozygous call. None of the strains' heterozygosity exceeds 0.5.

A simple statistical model is used to quantitively tell how unreasonable each column is.

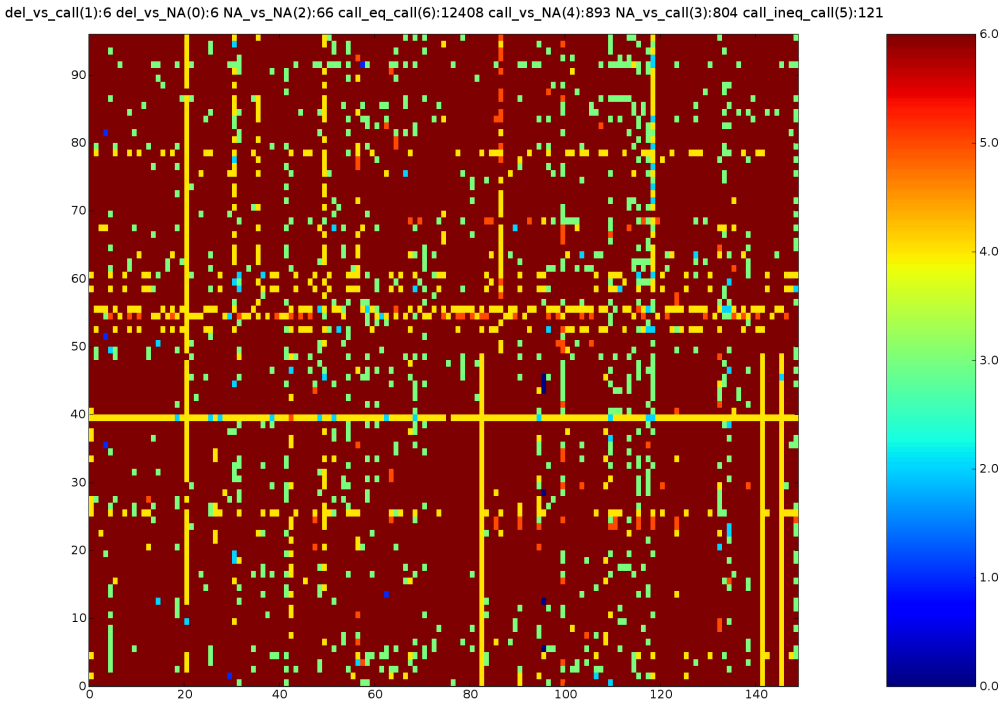

Figure 4: 2010 versus Justin

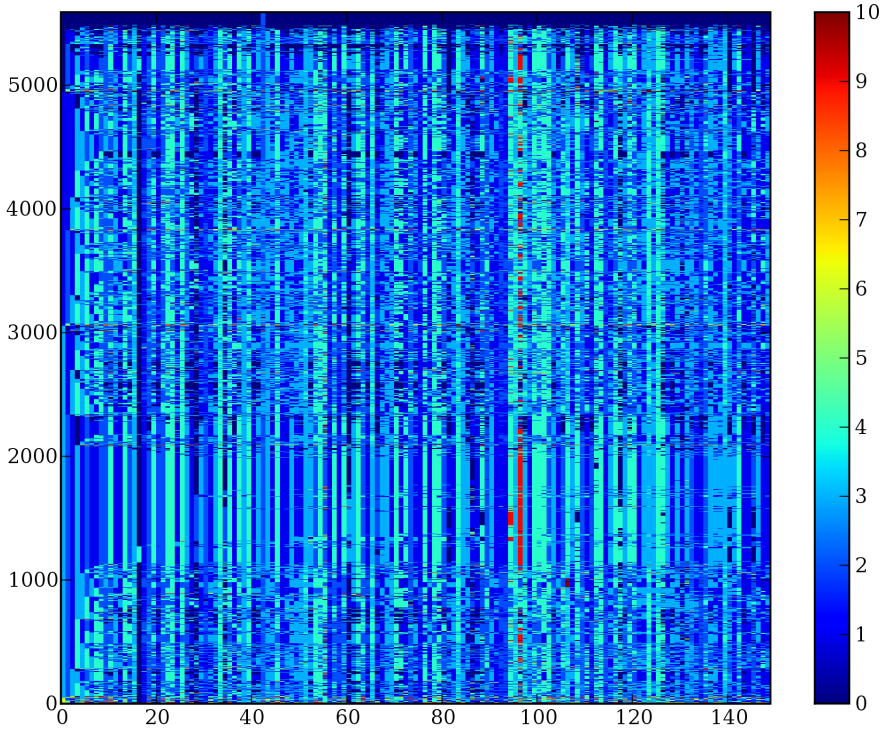

Figure 5: 0 is NA. 1-4 are ACGT. 5-10 are heterozygous call

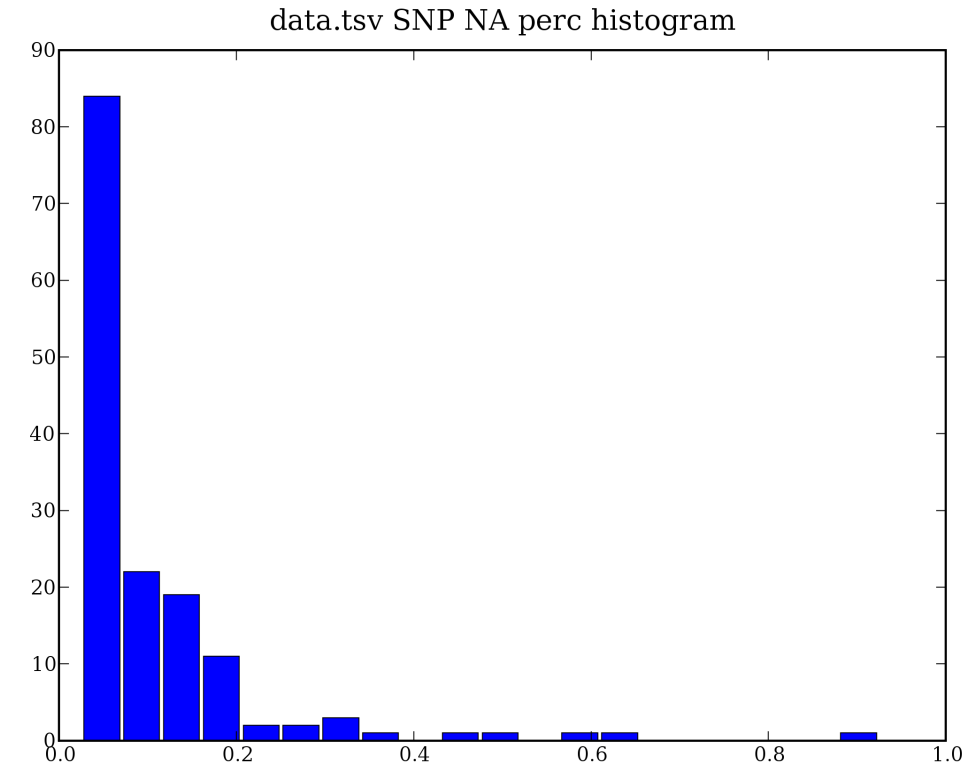

Figure 6:

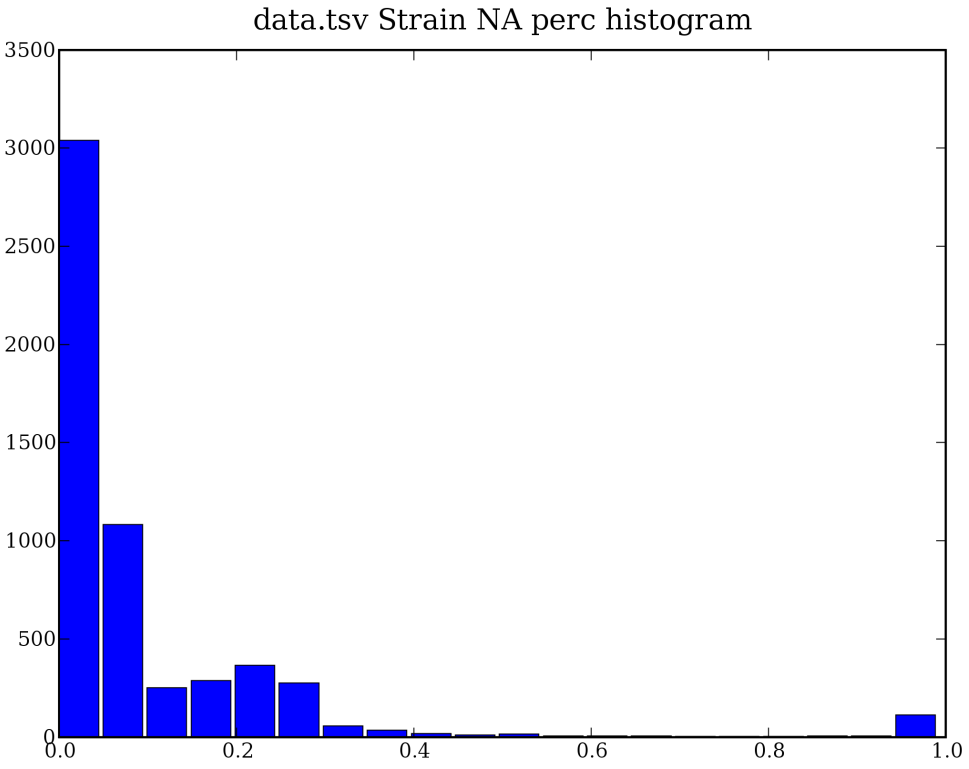

Figure 7:

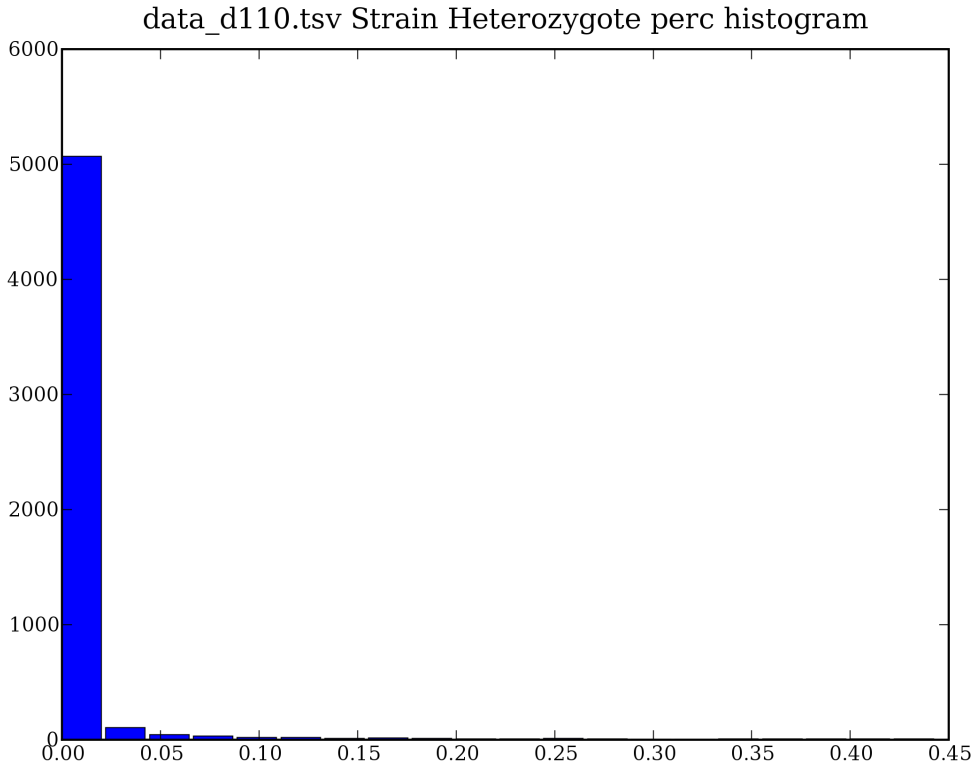

Figure 8:

6.1 model for snp locus

$$P(SNP_j|Strain \quad heterozygous \quad info) = \prod_{i=1}^{N} p_i^{a_i^j}(1-p_i)^{1-a_i^j} \tag{1}$$

p_i is probability that one strain has heterozygous call.

a_i^j is indicator whether SNP_j is homozygous(= 1) or not(= 0) for $Strain_i$.

Figure 9 is the histogram of the probability for each SNP locus after taking logarithm.

4 SNP loci (AtMSQTsnp 267, AtMSQTsnp 232, AtMSQTsnp 138, AtMSQTsnp 263) with log probability ≤ -0.5 were removed. These four loci show long stretch of heterozygous calls among lots of strains in Figure 8.

These long stretch of heterozygous calls could trace to technical error, segmental duplication in all strains, segmental duplication in some strains or real heterozygotes caused by population structure.

7 Linkage Disequilibrium

we calculate LD among all pairwise loci on the same chromosome. In the calculation, all pairs with one or two missing values are discarded. all pairs with both heterozygous calls are discarded as the phase is unknown.

The only with considerable high LD, figure 10-17 (r^2=0.9114) is 'AtMSQTsnp 22' versus 'AtMSQTsnp 27'. These two snp loci is only 1713bp apart. The next closest pair is 4520bp apart with very low LD (r^2=0.0351). The next top 9 highest LDs (their r^2 ranges from 0.2423 to 0.2882) are all from chromosome 1 although their distance ranges from 1.1MB to 25MB. Other chromosomes have shorter distance pairs.

In figures about LD, the curve-fitting method is spline.

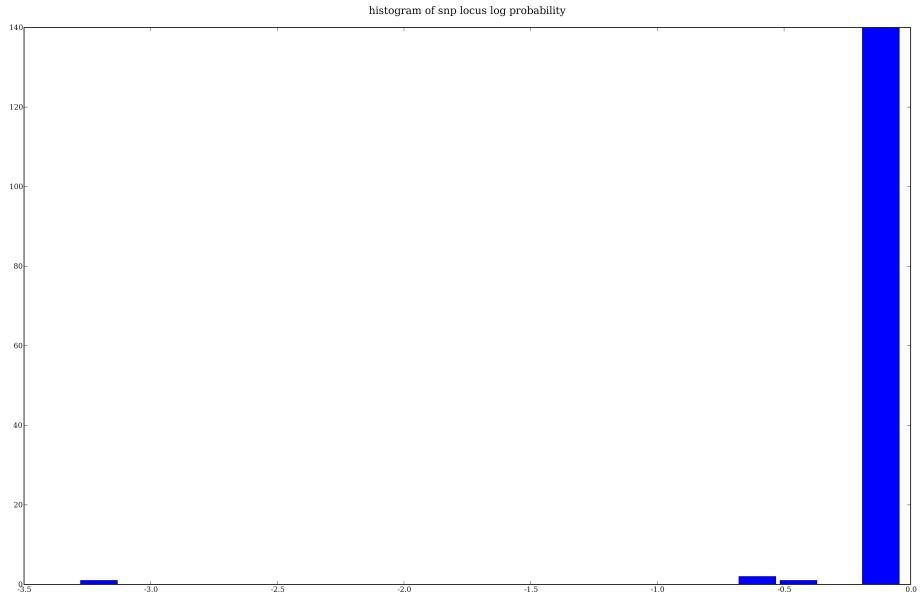

Figure 9:

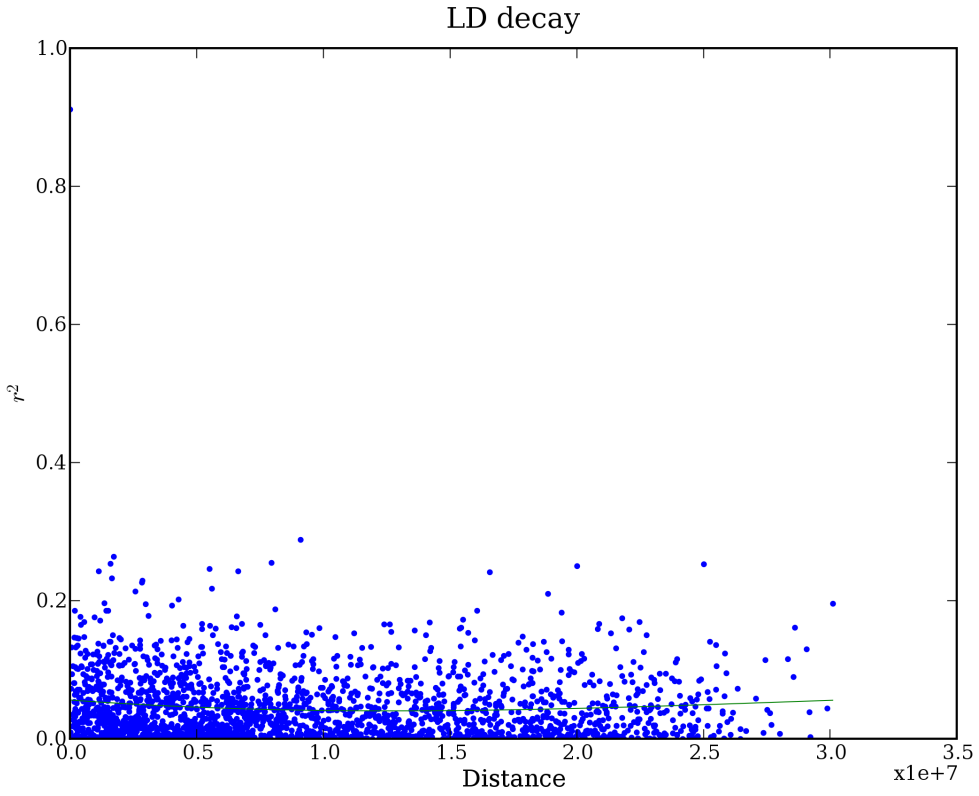

Figure 10: LD wih r^2

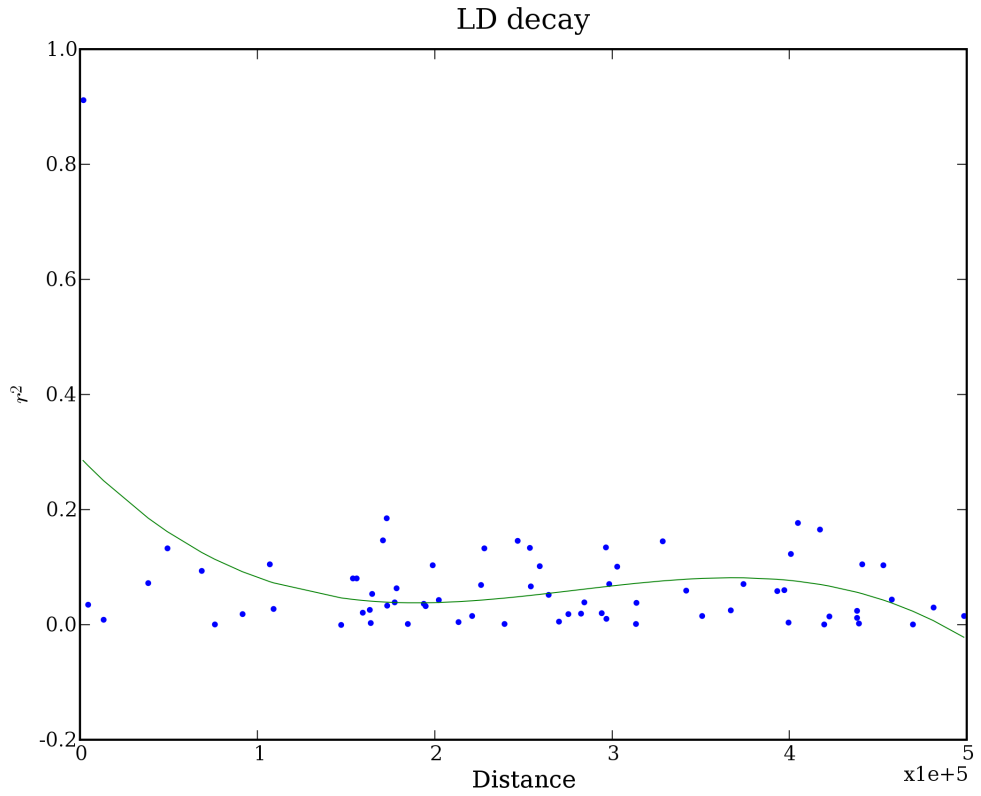

Figure 11: LD with r^2 of all pairs within 500KB

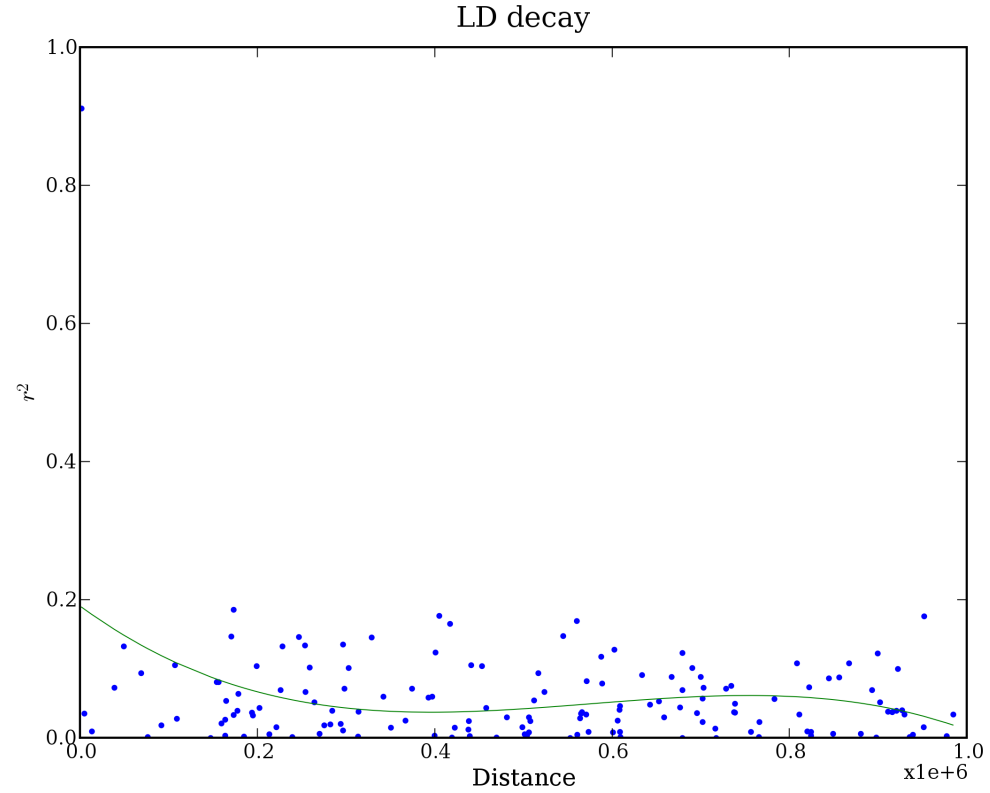

Figure 12: LD with r^2 of all pairs within 1MB

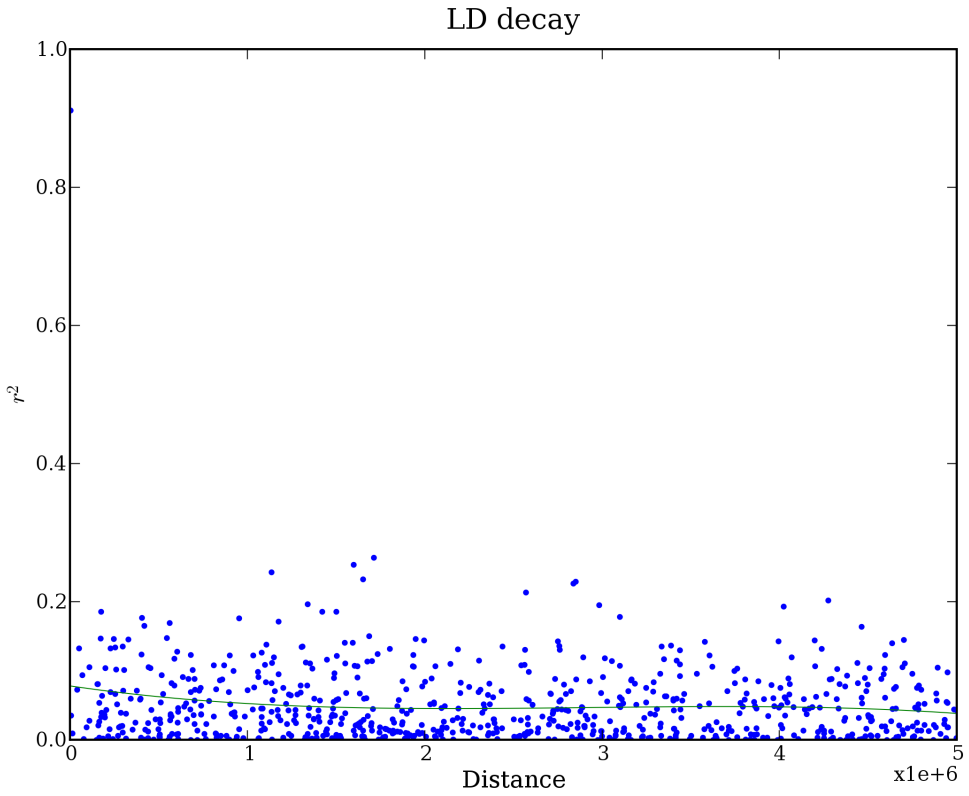

Figure 13: LD with r^2 of all pairs within 5MB

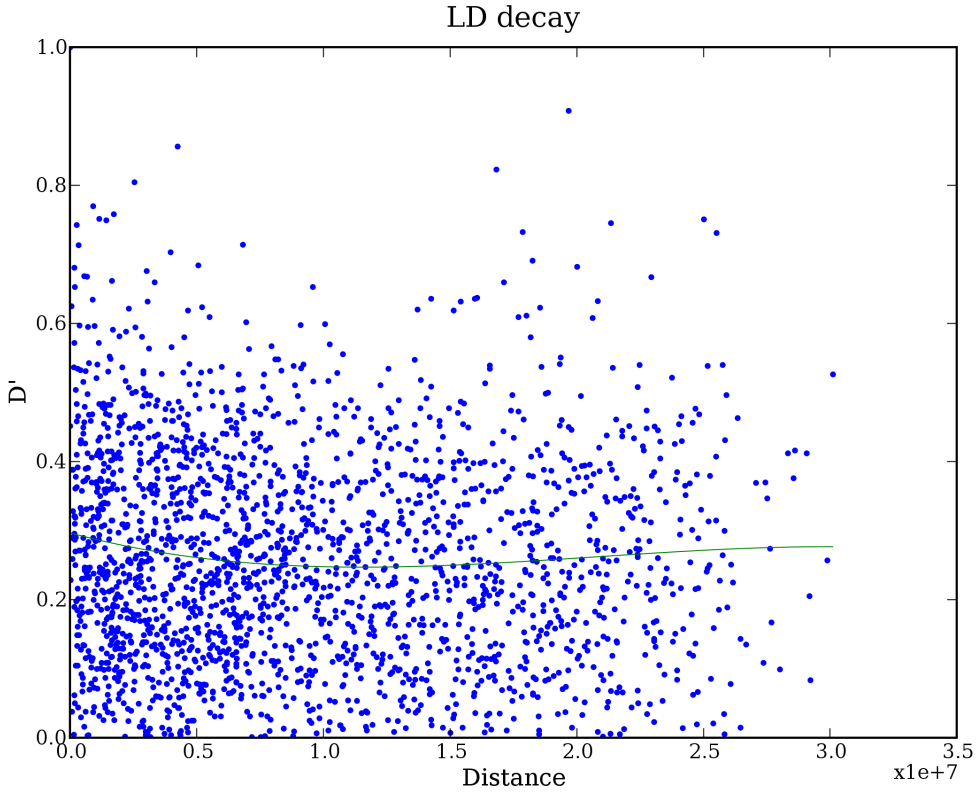

Figure 14: LD wih D'

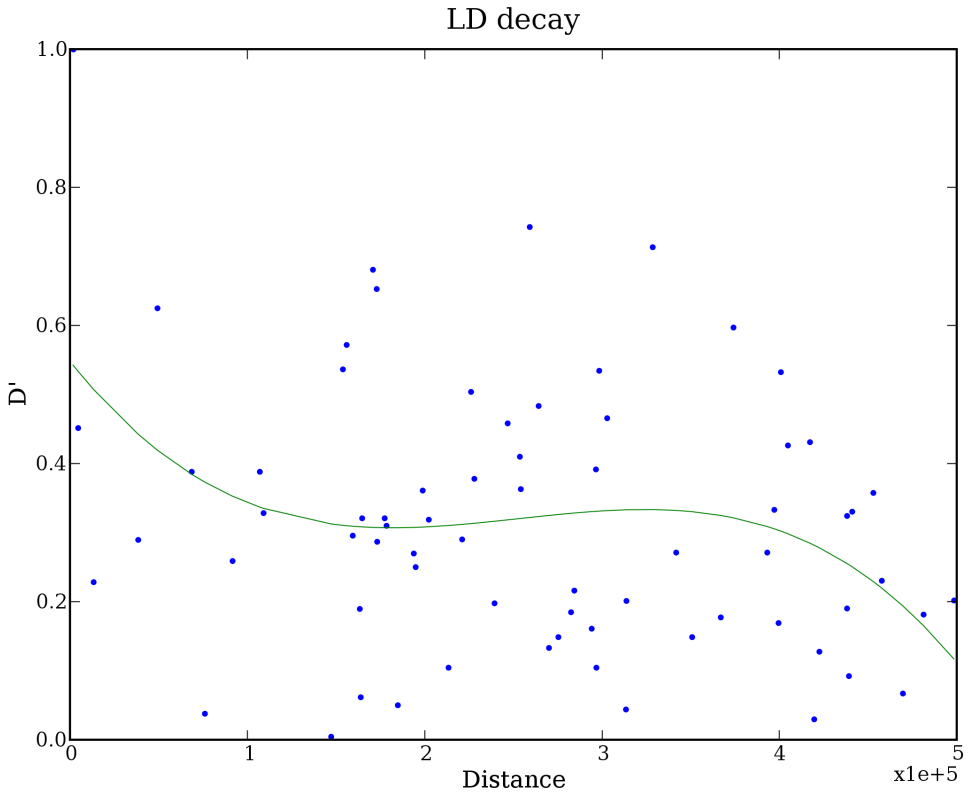

Figure 15: LD with D' of all pairs within 500KB

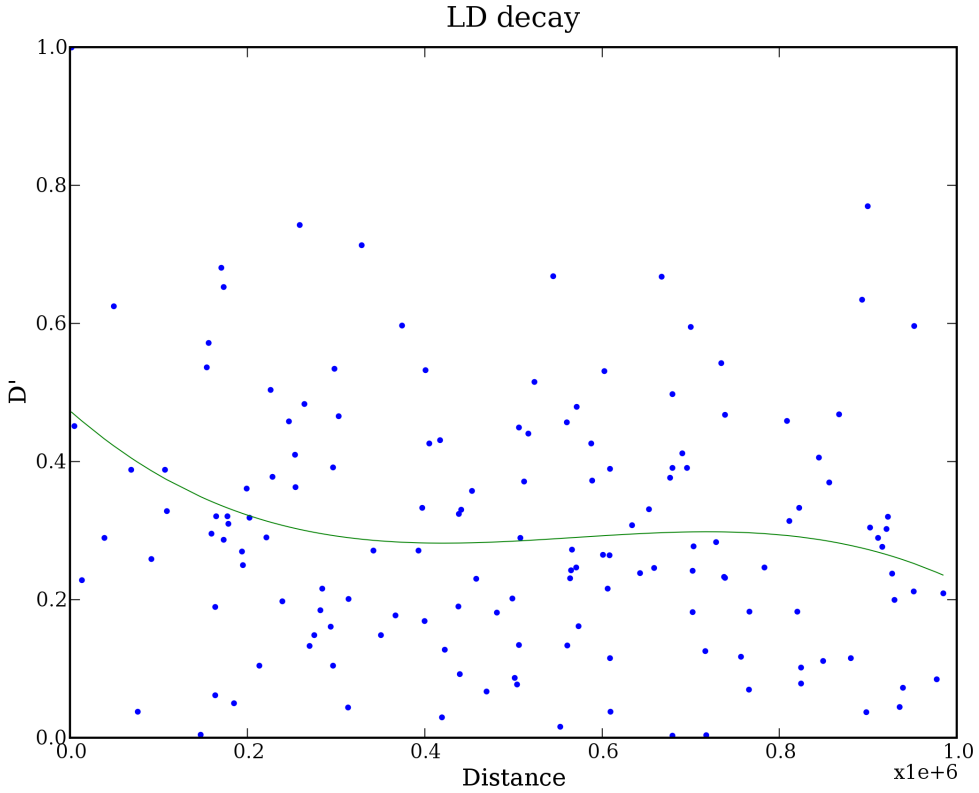

Figure 16: LD with D' of all pairs within 1MB

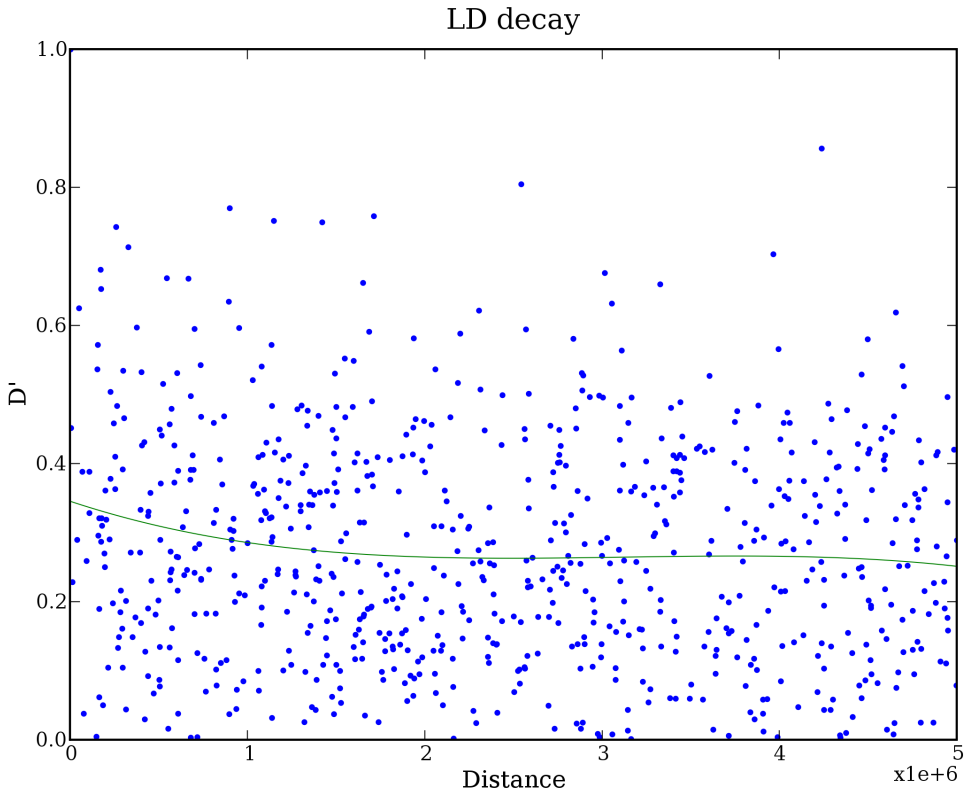

Figure 17: LD with D' of all pairs within 5MB

7.1 LD in North American Samples

the LD in North America is considerably higher than in the globe (Figure 18, 19, 20, 21).

8 Genetic and Geographic Structure

8.1 Correlation between genotype and geographic distance.

As in previous section, most unique haplotypes spread in local populations. So we suspected that the pairwise distance based on genotype might be correlated to geographic distance.

Figure 22 is a histogram of distance showing a bi-modal shape, which reflects that most strains come from either north america or europe.

Figure 23 shows the correlation between genotype and geographic distance. Due to the whopping number of pairwise comparisons, 14 million, 10000 samples were taken to plot the correlation. The inter-continent comparisons cause a distant cloud of data. So we separate the data into 3 groups, europe, north america and inter-continent.

Europe, figure 24, shows a positive correlation. While in north america, figure 25, the correlation is botched up by two very similar populations 1000km apart. Between the two continents, figure 26, the correlation is noisier.

8.2 correlation between the pairwise genotype distance using 149snp and full 2010 data

There're 247 arabidopsis strains in 2010 data. 244 of them have their 149 snps genotyped. The sequencing coverage of the 247 strains varies. 96 of them, used in [10], have about 1500s loci, from which the 149 snps are picked. The other 96 have about 112 loci.

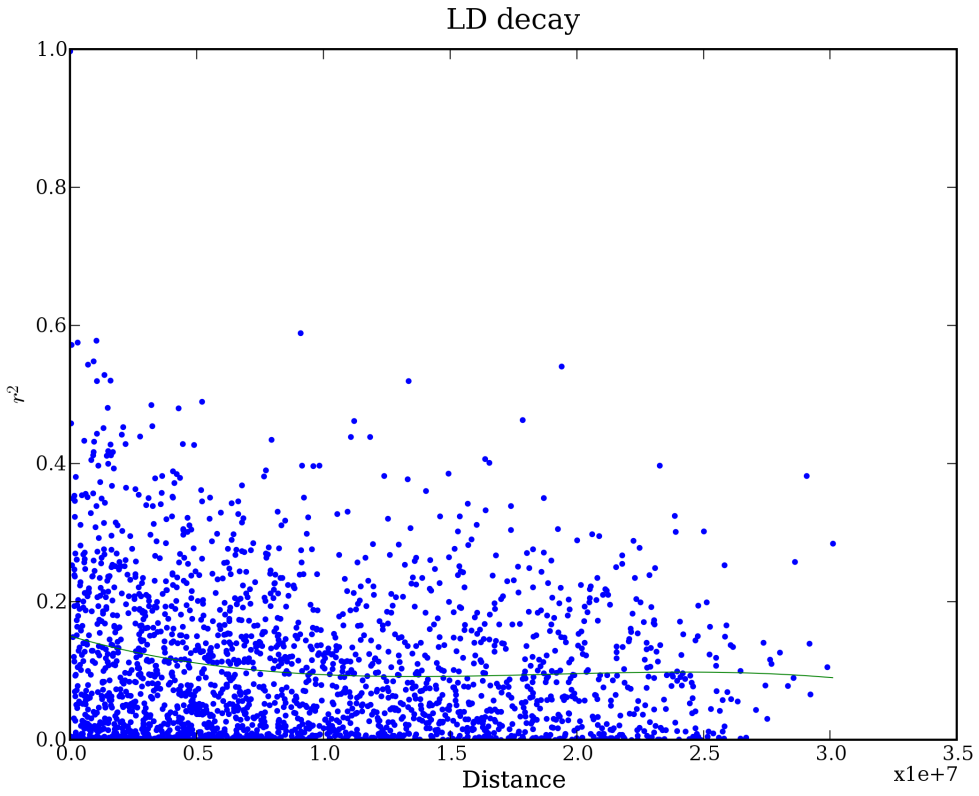

Figure 18: LD wih r^2 in North America

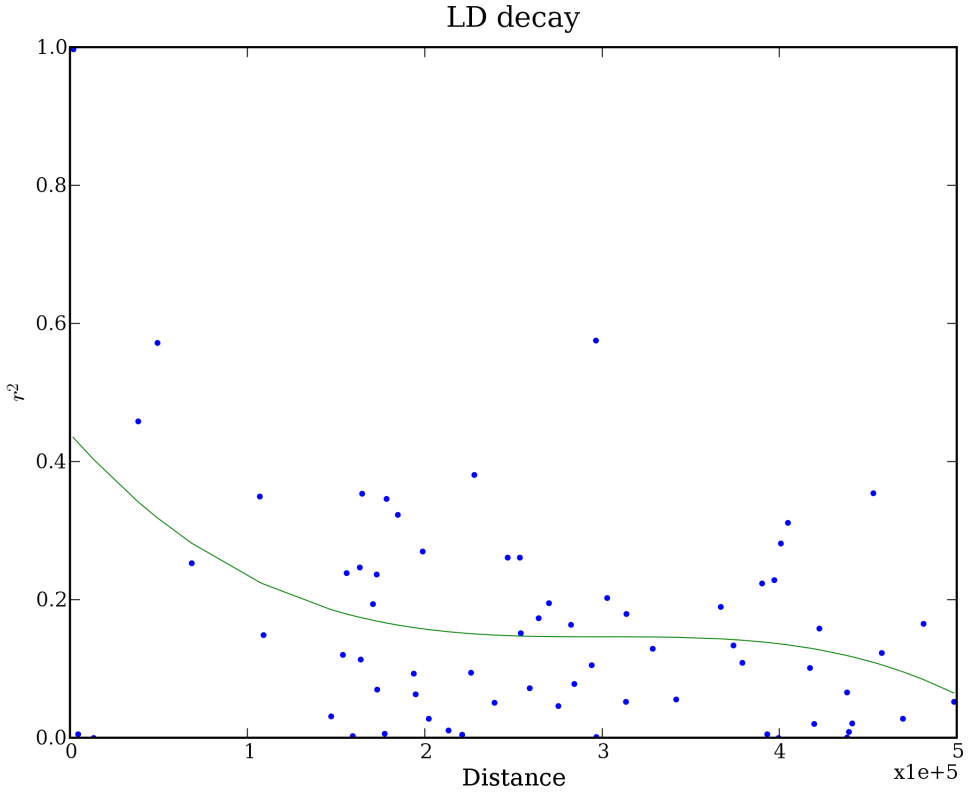

Figure 19: LD with r^2 of all pairs within 500KB in North America

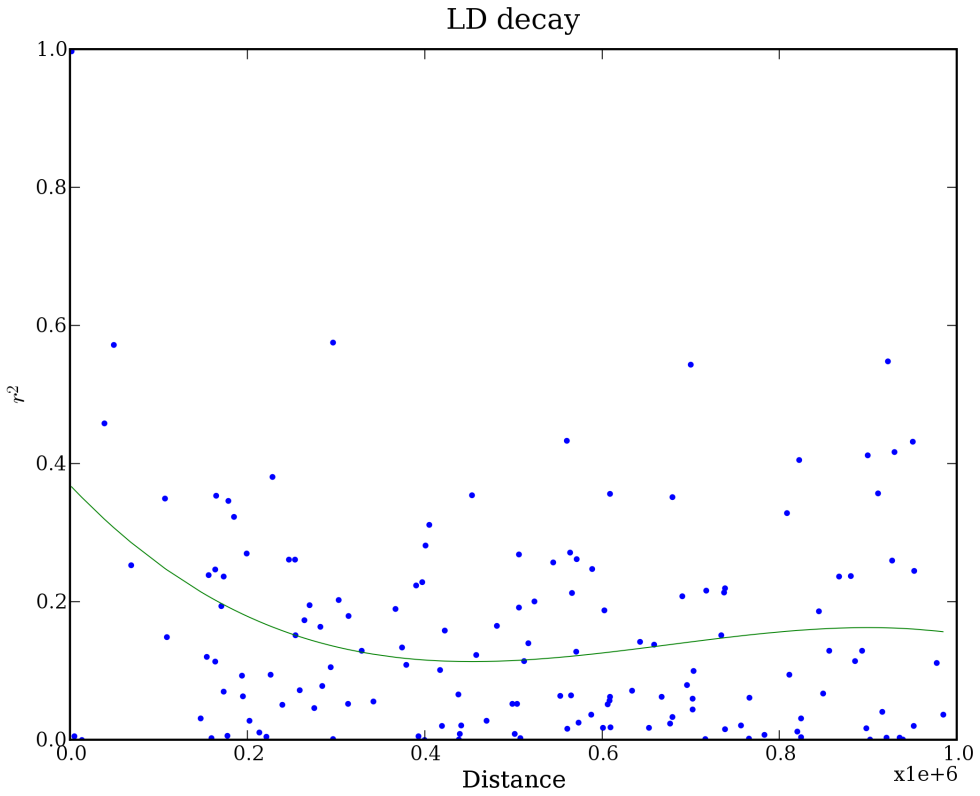

Figure 20: LD with r^2 of all pairs within 1MB in North America

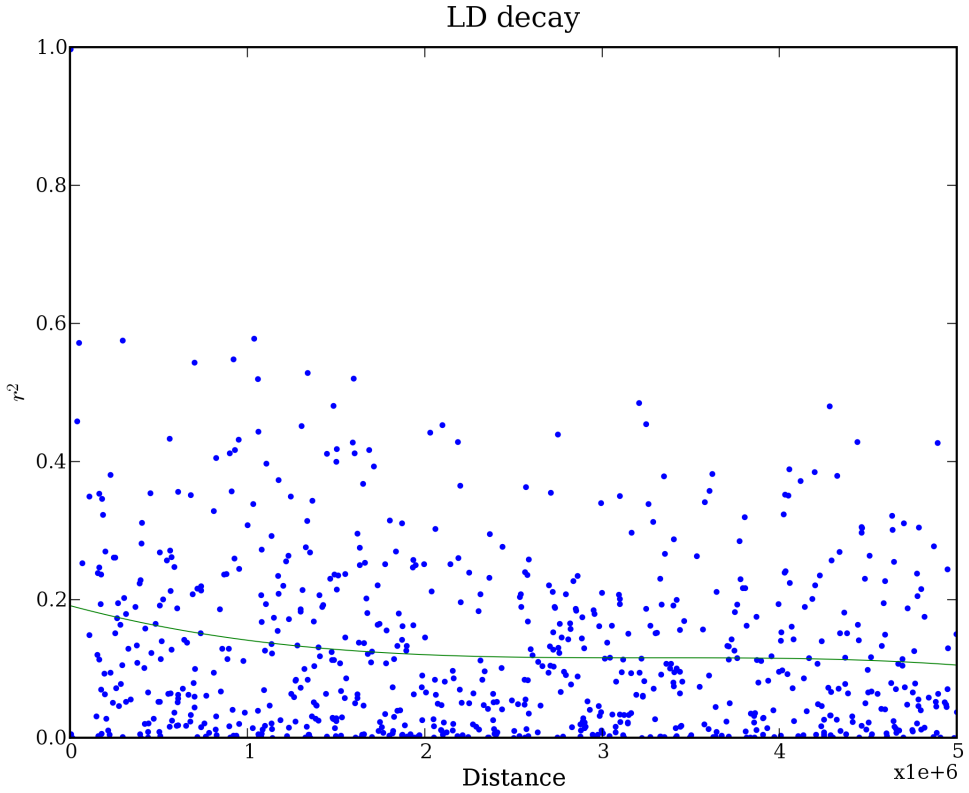

Figure 21: LD with r^2 of all pairs within 5MB in North America

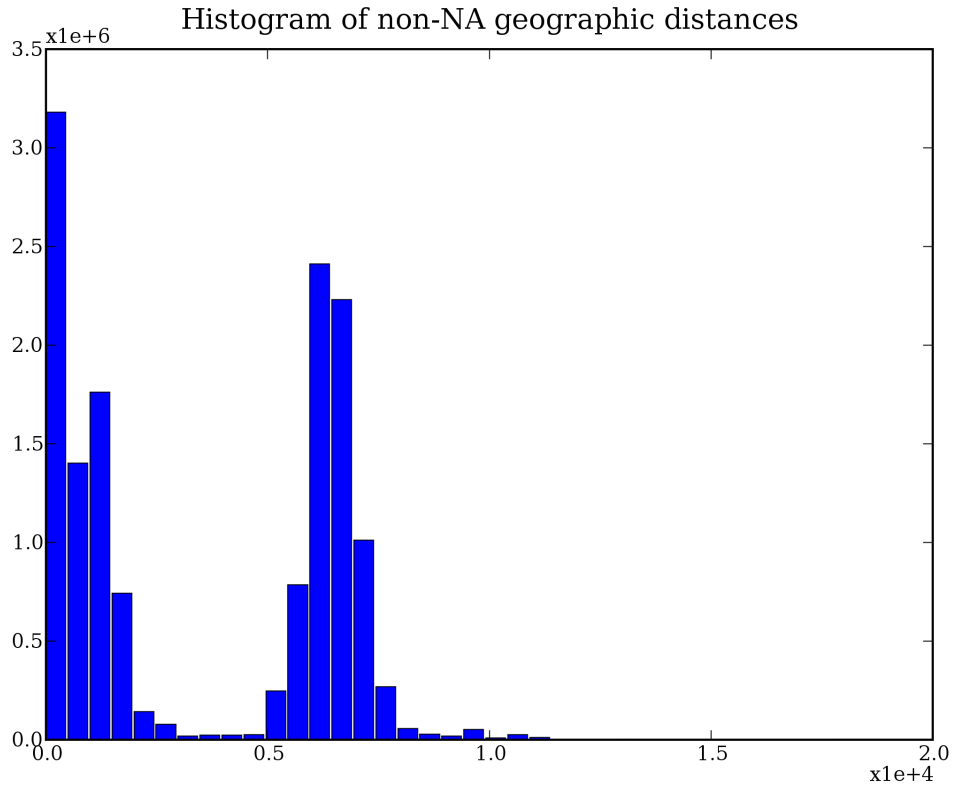

Figure 22: Histogram of geographic distance

9 identity strains across globe

Here is to see how identity strain pairs are distributed across globe. We first tried a loose criteria to define identity. If one of the two calls is NA, it's deemed as identical too. Hence, the transivity of identity relationship is not guaranteed. For example, given strain 1 and 2 are identical and strain 1 and 3 are also identical, strain 1 and 3 might not be identical.

Figure 27 is a histogram of pairwise distances (binary/hamming). 433,295 identity pairs is a major portion of the leftest spike.

In total, there're 433,295 identity pairs.

9.1 connected components in the identity strain graph

Think of it as a graph, which can be partitioned into connected components. Within each connected component, any two strains could be reached via some intermediate strains based on identity relationship. Each connected component could loosely correspond to a distinct haplotype ('loosely' because its transivity is not guaranteed and there're still some strains who are not identical.).

There're not many long-distance connected components. Most of them are within one country. So far two components were spotted as cross ocean.(still need systematically determine how many cross-ocean components there are.)

Figure 28 shows the cross-atlantic component. Figure 29 is the data matrix corresponding to the cross-atlantic component. Notice this component is actually a clique (transivity reached, any two strains are identical to each other.). Magnus and I just sat down and found three Col-0 strains are in this component, which raises lab-contamination suspicion over this component. This component is probably Col-0 haplotype.

There's another component linking america and japan, figure 30. the picture was cut improperly. Check figure 33 to see it's actually connected all the way to japan.

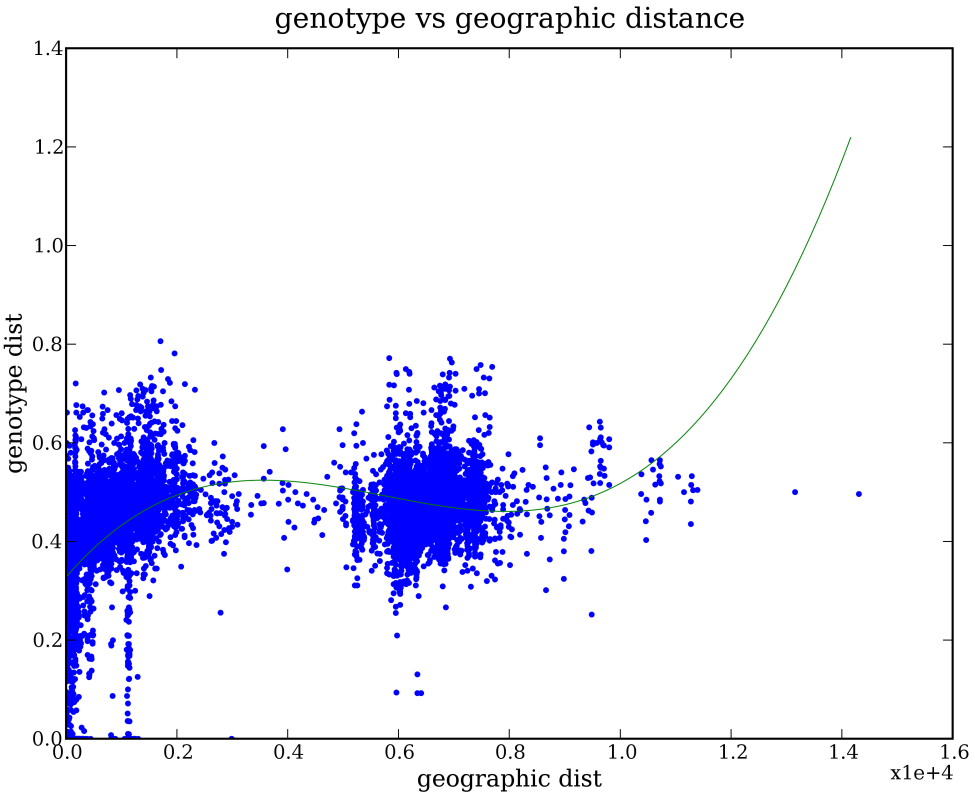

Figure 23: Correlation between genotype and geographic distance (10000 sampling from all data). the green curve is spline-fitting.

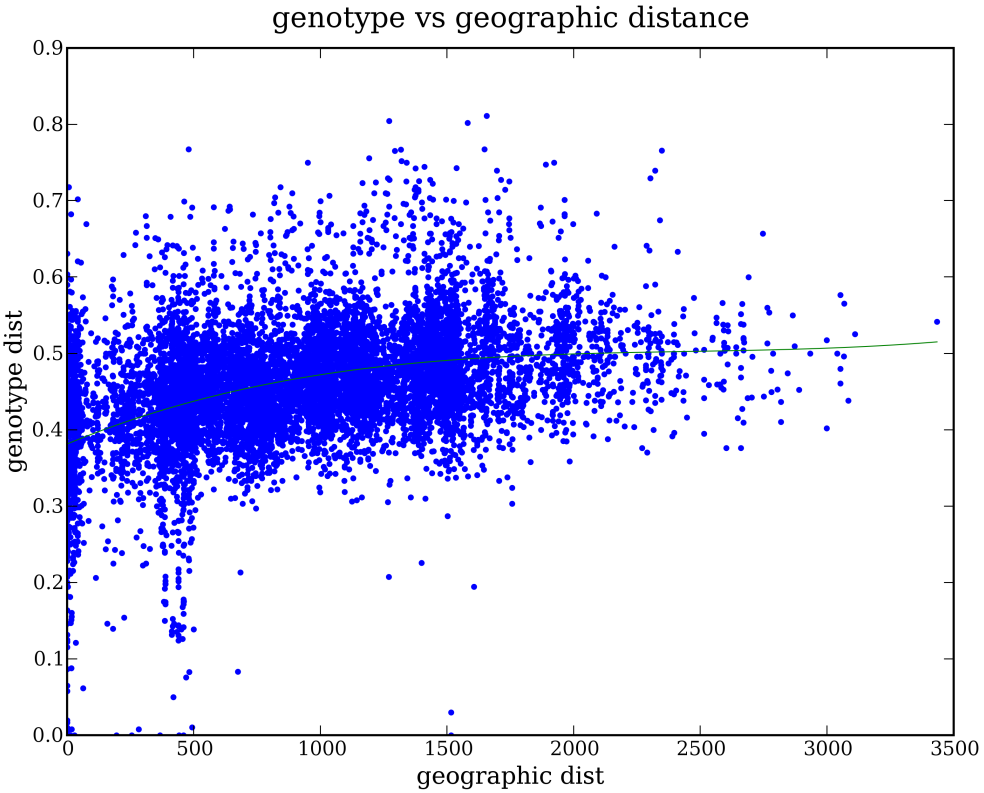

Figure 24: Correlation between genotype and geographic distance (10000 sampling from comparisons within europe). the green curve is spline-fitting.

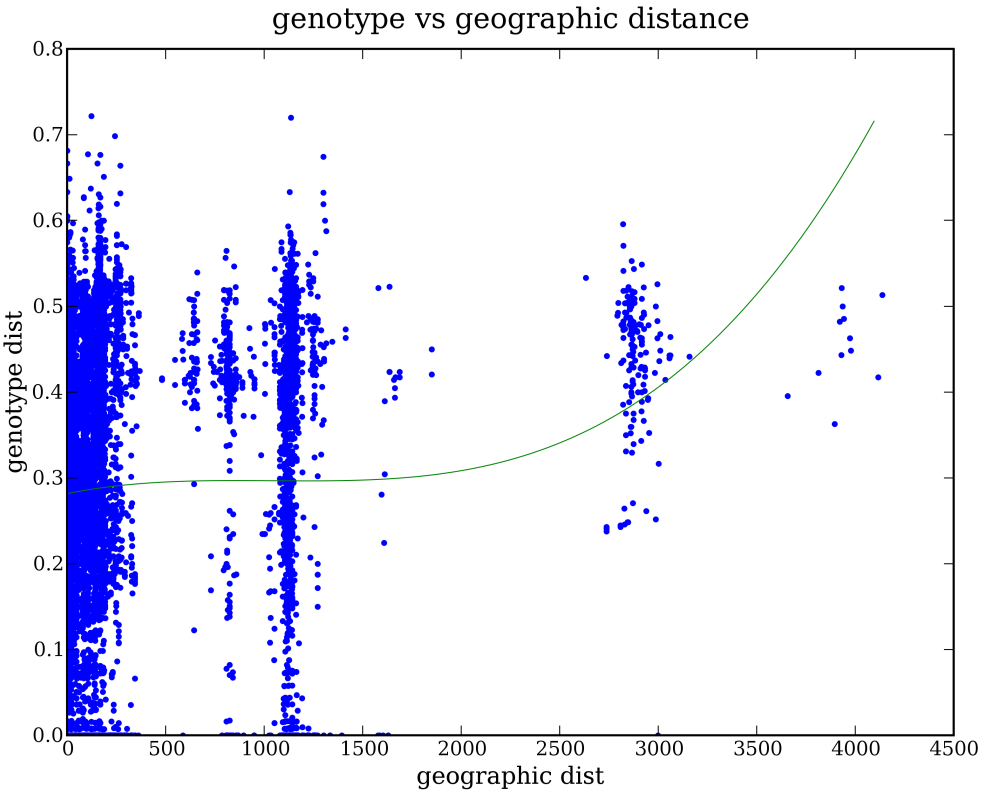

Figure 25: Correlation between genotype and geographic distance (10000 sampling from comparisons within north america). the green curve is spline-fitting. around 1000km apart, there's a long stretch of everything

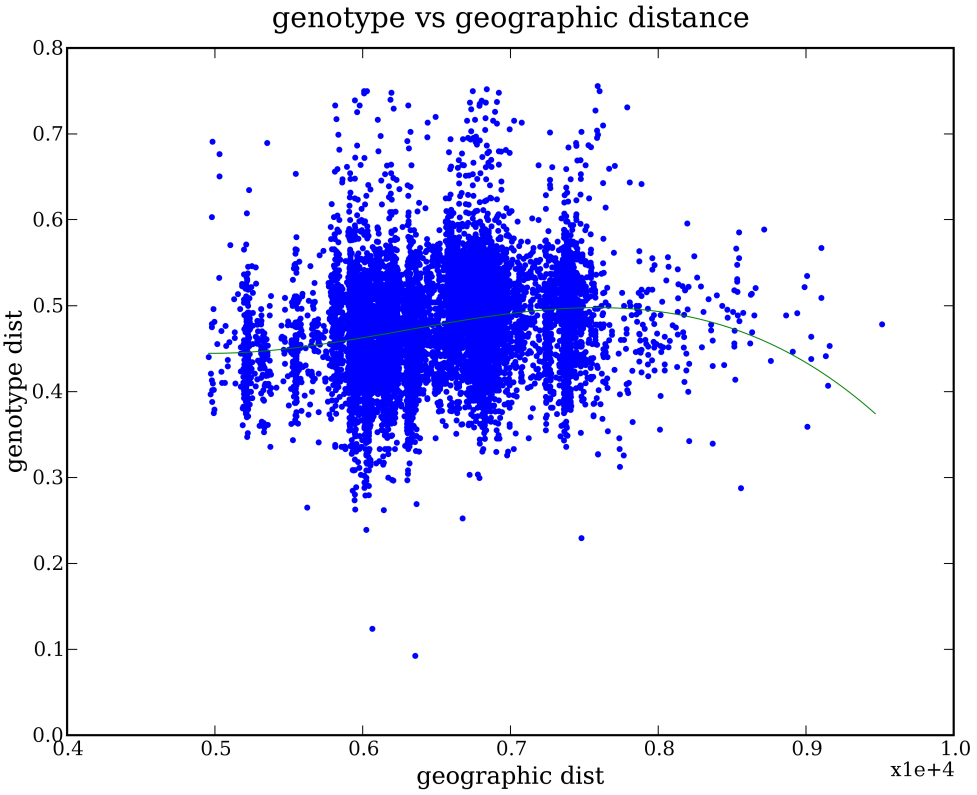

Figure 26: Correlation between genotype and geographic distance (10000 sampling from comparisons between europe and north america). the green curve is spline-fitting.

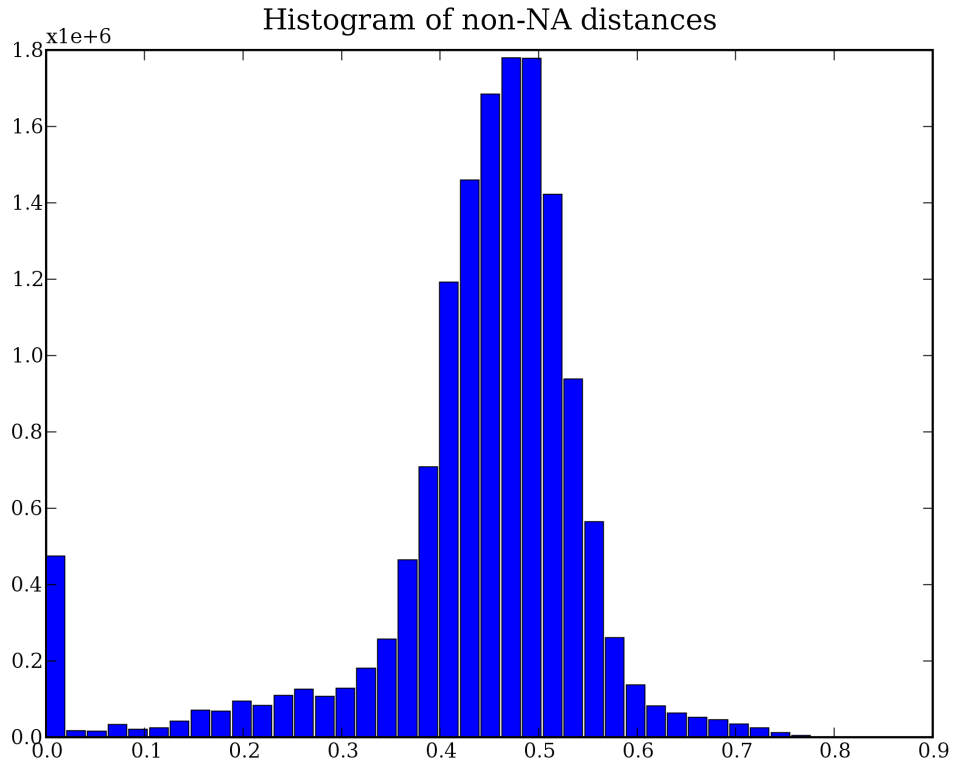

Figure 27: histogram of pairwise distance

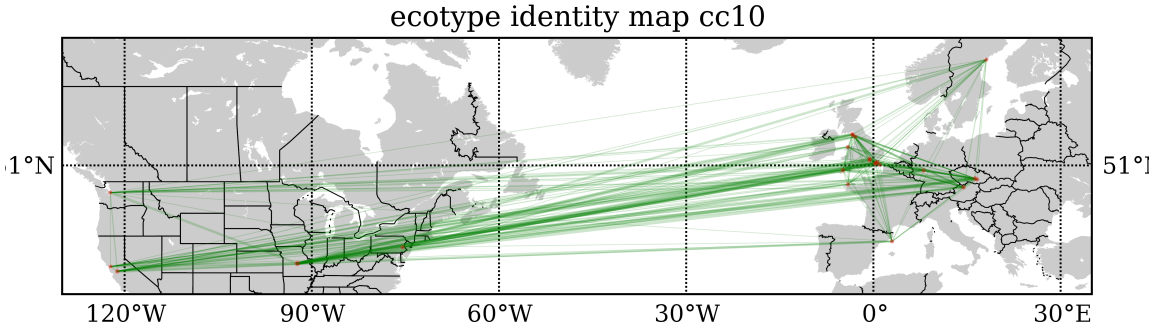

Figure 28: the cross-atlantic component

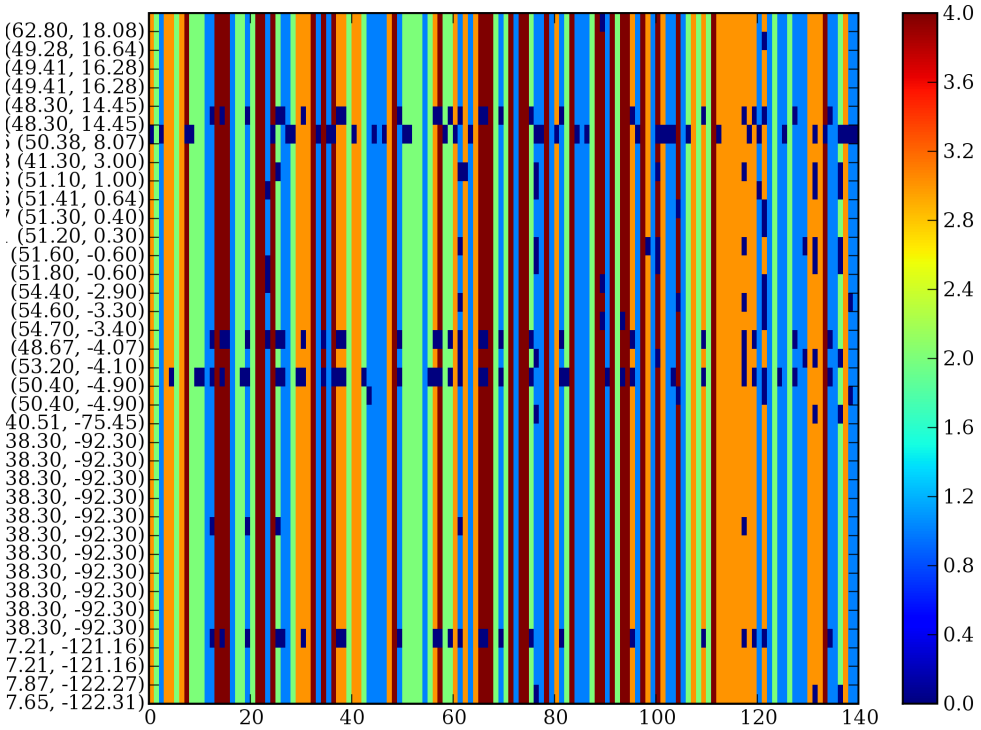

Figure 29: the data corresponding to the cross-atlantic component in figure 28

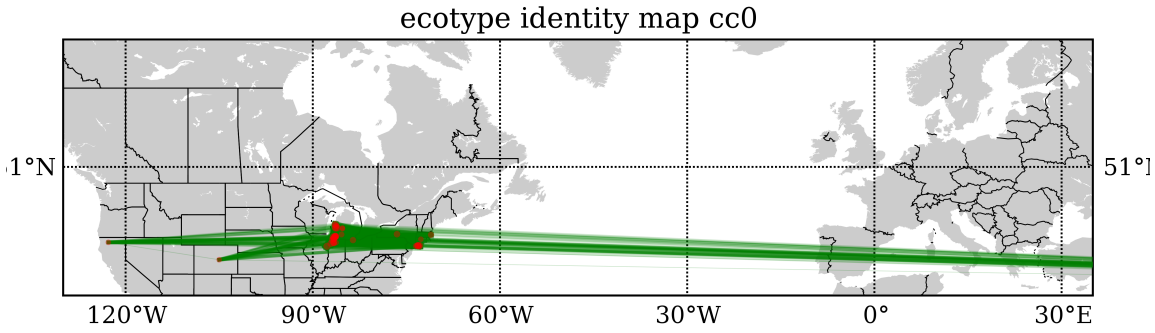

Figure 30: the america-japan component

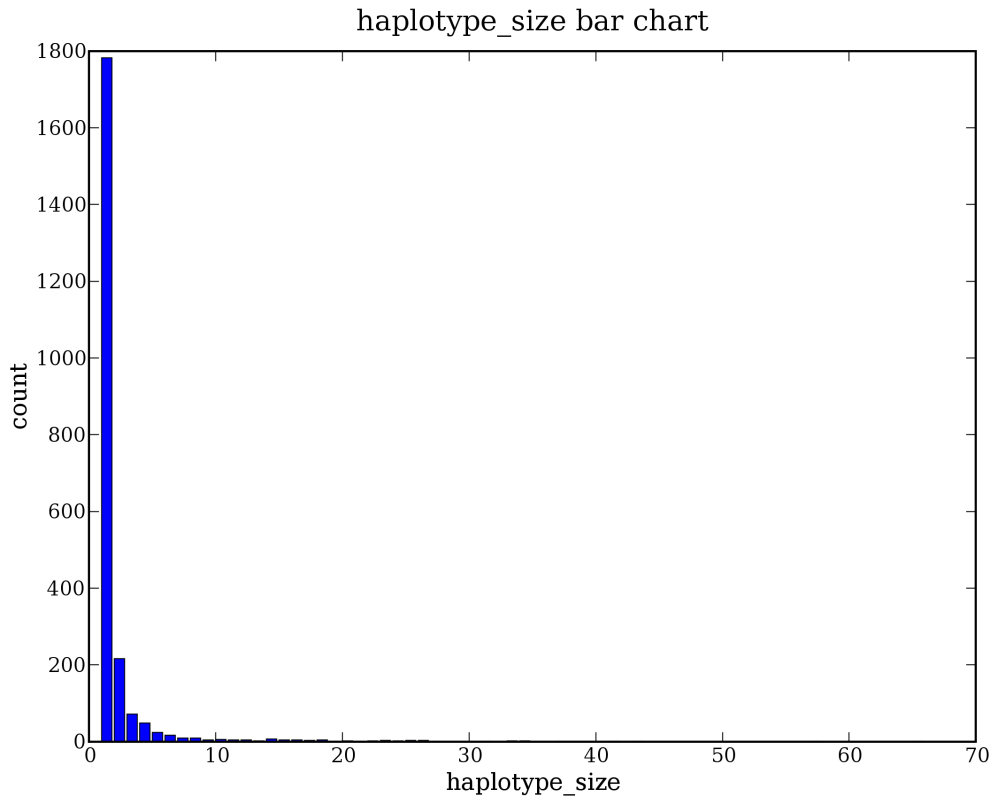

Figure 31: x axis is how many times one unique haplotype occur in the population, which is labeled haplotype size. y axis is the number of unique haplotypes with that number of occurrence. for example, 1st bar means there're 1782 unique haplotypes that occur only once in the population. An outlier whose haplotype size is 900 is excluded in this figure and it only occurs once, figure 32.

9.2 unique haplotypes

With transivity not guaranteed in the connected component, we further break the connected components into cliques. A clique is a graph in which any two vertices are neighbors. Every clique could represent one unique haplotype.

Figure 31 shows the distribution of the number of times unique haplotypes (magnus, i changed 'genotypes' to 'haplotypes') were sampled. There're 2256 unique haplotypes in total. The distribution is extremely skewed: 1782(78.99%) of all haplotypes are seen once, and of those seen more than once, 217(9.62%) are only seen twice. Only 257 haplotypes are seen more than twice. The most frequent one is seen 900 times. Figure 32 shows the geographic distribution of all 900 instances. only one of them appear outside north america, which is suspected to be the result of academia interaction.

For 257 haplotypes that appear in more than three times, most of them appear in local populations, or at least within one continent. Apart from the one, figure 32 crossing pacific, there're six cross-atalantic, table 8, 9, 10, 11, 12, 13 and one cross-eurasia, table 7.

9.3 Identity Map on the scale of population

In order to gain a global look. we grouped strains into population. Each population is formed via connecting close sites. There's a distance (great circle distance) threshold to determine how close sites within a population should be.

289,867(66.9%) pairs are across-population for population model thresholded by 25km. In figure 33, Each population is denoted as a red dot with its diameter proportional to the number of identity pairs within that population. An edge is connected if there's identity relationship between two populations. The thickness of an edge corresponds to the number identity pairs between them.

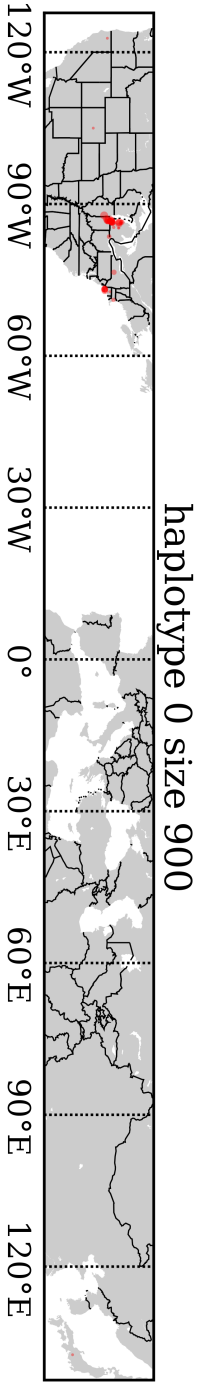

Figure 32: the most frequent haplotype, also the only one cross-pacific. most of them, 899 are in north america (US). only 1 (nativename is Gifu-2) is in japan.

Table 7: Cross Eurasia Haplotype

ecotypeid	nativename	stockparent	latitude	longitude	site	country
7374	CS28781	CS6926	34.43	136.31	Tsu	JPN
6972	CS28782	CS22641	34.43	136.31	Tsu	JPN
7373	Tsu-0	CS6874	34.43	136.31	Tsu	JPN
8394	Tsu-1	CS22641	34.43	136.31	Tsu	JPN
7375	Tu-0	CS6875	45	7.5	Tu	ITA
8395	Tu-0	N1567	45	7.5	Tu	ITA

Table 8: Cross Atlantic 1

ecotypeid	nativename	stockparent	latitude	longitude	site	country
6714	BG-7	CS22347	47.6479	-122.305	BG	USA
7012	Berkeley	CS8068	37.8695	-122.271	Berkeley	USA
8377	Santa Clara	N8069	37.21	-121.16	Santa Clara	USA
7329	Santa Clara	CS8069	37.21	-121.16	Santa Clara	USA
7090	Col-2	CS907	38.3	-92.3	Col	USA
7089	Col(gl1)	CS3879	38.3	-92.3	Col	USA
7088	Col-0	CS60000	38.3	-92.3	Col	USA
6909	Col-0	CS22625	38.3	-92.3	Col	USA
7086	CS28175	CS6930	38.3	-92.3	Col	USA
7087	Col-1	CS3176	38.3	-92.3	Col	USA
7082	Col-0	CS6673	38.3	-92.3	Col	USA
7083	Col-4	CS933	38.3	-92.3	Col	USA
7084	Col-7	CS3731	38.3	-92.3	Col	USA
7085	Col-3	CS908	38.3	-92.3	Col	USA
7091	Col-6(gl1)	CS8155	38.3	-92.3	Col	USA
7233	Limeport	CS8070	40.5088	-75.4472	Limeport	USA
4778	UKSW06-178		50.4	-4.9	St Columb	UK
4855	UKSW06-255		50.4	-4.9	St Dennis	UK
5712	UKID5		53.2	-4.1	Bangor	UK
241	MOG-36		48.6667	-4.06667	MOG	FRA
5639	UKNW06-476		54.7	-3.4	Cockermouth	UK
5485	UKNW06-232		54.6	-3.3	High Lorton	UK
5656	UKNW06-493		54.4	-2.9	Windemere	UK
5714	UKID7		51.6	-0.6	Becconsfield	UK
5818	UKID114		51.8	-0.6	Tring (N.History Museum)	UK
5221	UKSE06-450		51.2	0.3	Tonbridge castle	UK
5177	UKSE06-375		51.3	0.4	Wateringbury	UK
6896	Sq-4	CS22243	51.4083	0.6383	SQ	UK
5766	UKID61		51.1	1	Port Lympne	UK
7018	Bla-12	CS6624	41.3	3	Bla	ESP
7226	Li-5	CS6775	50.3833	8.0666	Li	GER
6414	Uod-3		48.3	14.45	Uod	AUT
8428	Uod-2		48.3	14.45	Uod	AUT
5880	DraIII-7		49.4112	16.2815	DraIII	CZE
5879	DraIII-5		49.4112	16.2815	DraIII	CZE
6294	UduI 1-8		49.281	16.6353	UduI 1	CZE
1441	SL-3		62.8	18.0833	SL	SWE

Table 9: Cross Atlantic 2

ecotypeid	nativename	stockparent	latitude	longitude	site	country
7805	ME3.09		42.093	-86.74	ME (Benton Harbor)	USA
7821	ME3.25		42.093	-86.74	ME (Benton Harbor)	USA
7831	ME3.35		42.093	-86.74	ME (Benton Harbor)	USA
7844	ME3.48		42.093	-86.74	ME (Benton Harbor)	USA
8623	1130ME1.52		42.093	-86.359	ME (Benton Harbor)	USA
8624	1130ME1.53		42.093	-86.359	ME (Benton Harbor)	USA
299	TOU-A1-138		46.6667	4.11667	TOU-A1	FRA

Table 10: Cross Atlantic 3

ecotypeid	nativename	stockparent	latitude	longitude	site	country
7524	Rmx-A02	CS22568	42.036	-86.511	RMX	USA
5764	UKID59		54.7	-2.8	Penrith	UK
5768	UKID63		54.1	-1.5	Ripon	UK
5778	UKID73		52.2	1.5	Snape Malting	UK

Table 11: Cross Atlantic 4

ecotypeid	nativename	stockparent	latitude	longitude	site	country
6708	BG-1	CS22341	47.6479	-122.305	BG	USA
6709	BG-2	CS22342	47.6479	-122.305	BG	USA
6711	BG-4	CS22344	47.6479	-122.305	BG	USA
6908	CIBC-5	CS22602	51.4083	0.6383	CIBC	UK

Table 12: Cross Atlantic 5

ecotypeid	nativename	stockparent	latitude	longitude	site	country
7350	Tac-0	CS3885	47.2413	-122.459	Tac	USA
7349	Ta-0	CS6867	49.5	14.5	Ta	CZE
8389	Ta-0	N1549	49.5	14.5	Ta	CZE

Table 13: Cross Atlantic 6

ecotypeid	nativename	stockparent	latitude	longitude	site	country
7523	Pna-17	CS22570	42.0945	-86.3253	PNA	USA
5708	UKID1		56.8	-3.9	Aberfeldy	UK
5730	UKID23		55.3	-1.8	Cragside	UK

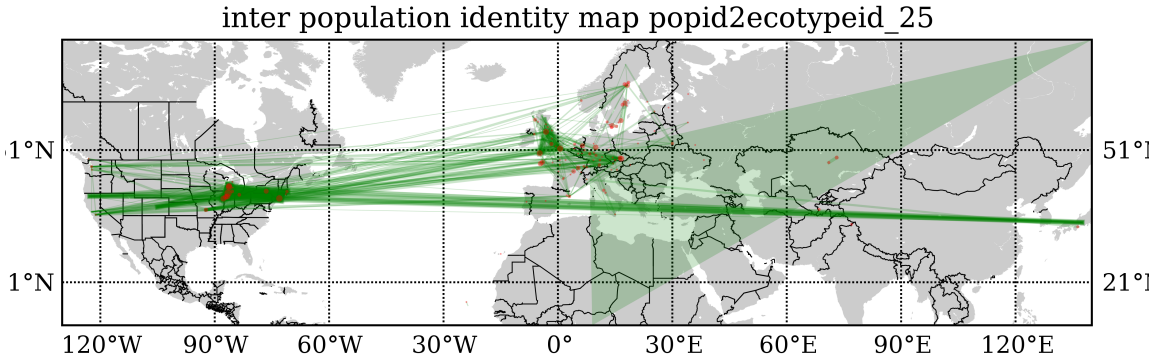

Figure 33: Population Identity Map

10 Remove Identity Strains

Due to the highly inbreeding nature of Arabidopsis, there're lots of identity strains in the samples collected from nature. In Figure 5, blocks of identity strains are very obvious. In the detection of identity strains, NA is regarded to be identical to any genotyping call. Because different strains have NAs in different SNP loci, the identity relation is not transitive. For example, A is identical to B and B is identical to C but A might not be identical to C. A graph is constructed with strains as nodes and two nodes are connected if these two strains are deemed as identical. An example graph is shown in Figure (?). A greedy algorithm is used to remove nodes with highest degree until no edges left.

In the end, 2575 strains were removed.

11 Detecting Recombinant Inbred Line(RIL)

Figure 34 is a diagram showing the production of recombinant imbred lines by selfing.

use Dynamic programming to identify RILs with minimum number of recombinations. Regardless of how many selfing generations, the algorithm tells you whether one strain results from the outcrossing of two strains.

This step gives out results flooded with false trios. The 1st kind of false trio is that one parent is very close to the child, the other is a little farther (but still with substantial similarity) and just help to save some few incompatibilites. The 2nd kind is that one parent and child are close. The other parent is kind of irrelevant and saves the incompatibilites by being 'NA'.

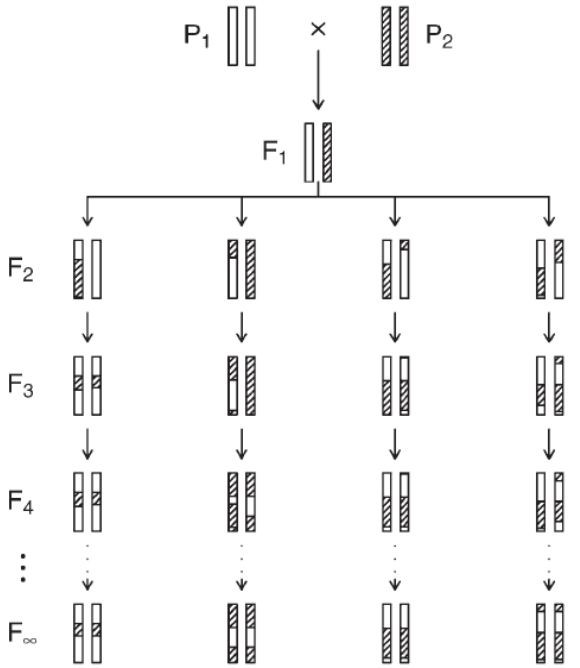

Figure 34: from [2]

12 Estimate the number of selfing generations since the last outcrossing event

Figure 34 is a diagram showing the production of recombinant imbred lines by selfing. F1 would be 0 selfing generation away from the last outcrossing event as it's the direct result of outcrossing. F2 is 1 selfing generation away from the last outcrossing event.

The following model is from [4].

12.1 model to estimate the number of selfing generations since the last outcrossing event

$$D_i = 1 - \sum_{j=1}^{k} p_{ij}^2 \tag{2}$$

p_{ij} is probability of allele j of SNP i. k is the number of alleles.
D_i is probability of SNP i being heterozygous.

$$likelihood(S_n) = \prod_{i=1}^{L} (\frac{D_i}{2^n})^{a_i} (1 - \frac{D_i}{2^n})^{1-a_i} \tag{3}$$

S_n is the strain which is n selfing generations away from the last outcrossing event.
L is the number of loci.
a_i is indicator whether SNP i is heterozygous(= 1) or not(= 0).
$\hat{n}$ is the MLE.

This model assumes the independence among the SNP loci, which is not true in general due to LD. According the recent Kim et al.(2007) paper (in preparation), LD decays within 10kb on average. In our data, the spacing between neighboring SNP loci is generally huge (see Figure 35). If sorted, the list of gaps between loci looks like 1713, 4520, 13250, 49053, ..., 4482219, 6334701. So only the first two gaps are within 10kb frame. So statistical independence could be well assumed.

Figure 36 is the result. The spike in 6 is probably spurious estimates due to only 1 or 2 hets in one strain which could be caused by genotyping error. If the whole population is in inbreeding equilibrium. The data should be geometric distribution with p=outcrossing rate. With large data, it's supposed to look like exponential decay. Our data is not true in this sense which hints inbreeding equilibrium hasn't been reached.

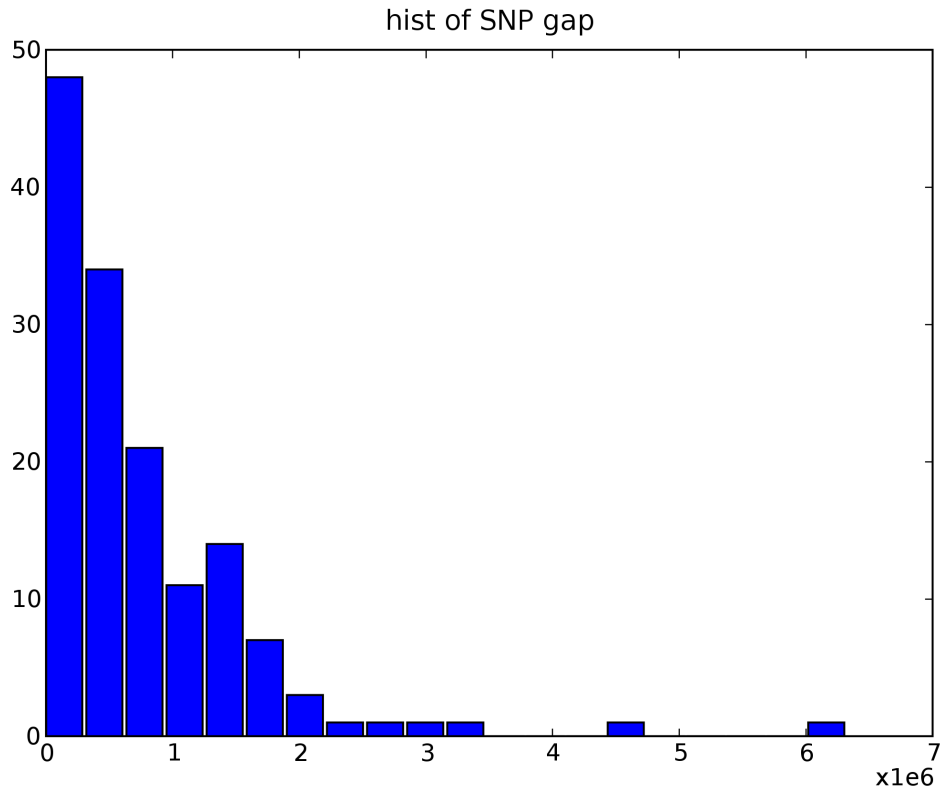

Figure 35:

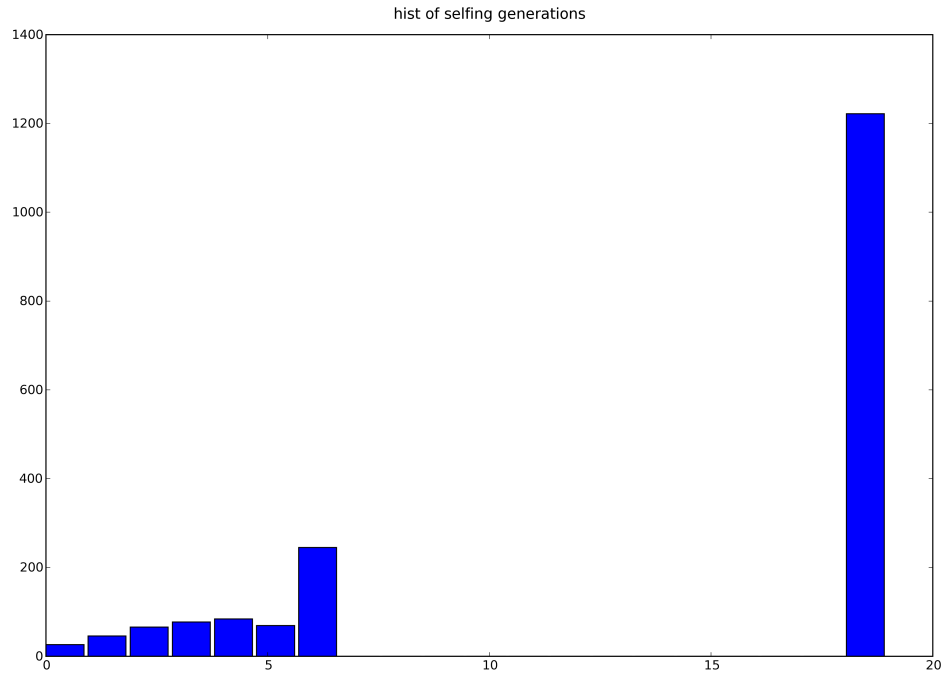

Figure 36:

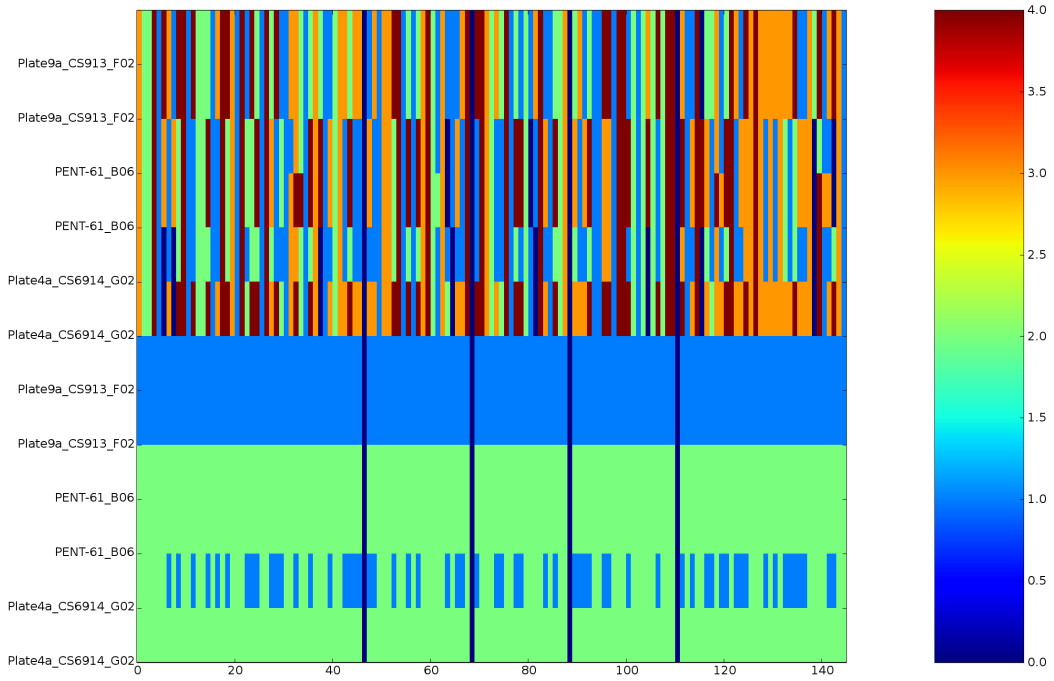

Figure 37:

Figure 37 is an example showing one strain with 0 selfing generations from the last outcrossing event.

13 Estimate selfing rate

Here gives a summary of the methods tried so far. the goal is to see whether there're major difference among different methods. the quick answer is no.

13.1 Jarne2006 [8]

simplest, single-locus, F_{IS}-based. $F_{IS} = 1 - H_{obs}/H_e$, where H_e is the expected heterozygosity, $H_e = 2pq$ and H_{obs} is the observed heterozygosity. $s = 2F_{IS}/(1 + F_{IS})$.

13.2 Robertson1984 [12]

single-locus, F_{IS}-based. not significantly different from Jarne2006. estimate is slightly higher than Jarne2006's. in several cases, it gives estimate slightly over 1.

13.3 Weir1984 [13]

single-locus, multi-loci, F_{IS}, F_{ST}-based. The single locus estimate gives similar result to Jarne2006 or Nordborg1997. But the multi-locus gives significantly lower estimate.

13.4 Nordborg1997 [9]

single-locus, colascent-based. This estimate gives value equal to Weir1984's single-locus one.

13.5 David2007 [3]

it's based on identity disequilibrium, which is an excess of heterozygotes or homozygotes across multiple loci. this method is most computing intensive and shows greatest difference.

13.6 todo

1. Enjalbert2000 [4] method
2. Simulate data with different selfing rate and apply different algorithms to see how effective different methods are.

How to define a population remains a question. Gao2007 [6], Pritchard2000 [11], Guillot2005 [7] and Francois2006 [5] have interesting solutions. But before figuring out how to define populations, we just tried a simple method.

13.7 Selfing rates across globe

Each population is formed via connecting close sites. There's a distance (great circle distance) threshold to determine how close sites within a population should be. For each population, its data undergoes a few preprocessing procedures. 1. remove strains with >= 40% NA. 2. remove snps with >= 40% NA. 3. remove snps with too many heterozygous inconsistent with the strains' homozygous state (detailed in a previous section).

so far 5km (120 populations), 10km (96 populations), 25km (66 populations) have been tried. The figures below are shown in this order for each particular region (England, Europe Continent, North America, Sweden).

In figures below, selfing rates are estimated by Jarne2006. Other methods (except David2007's) show minor differences most of the time. we'll come to the differing parts later. Each population is labeled by selfing rate X 1000. 0 denotes not enough data to make inference (at least 5 samples/strains). The diameter of the circle corresponds to the size of the population.

13.7.1 Standard deviation of these selfing rates

Due to the fact that Jarne2006's estimate is single-locus based, the selfing rate that's put on the map is the average of all loci. For those low average selfing rates, the std (standard deviation) is also bigger, which means for most loci of that population, selfing rate estimate is pretty close to 1 and just a few with a lot of heterozygous calls which pulls down the average. However remember, bad snps with too many heterozygous inconsistent with the strains' homozygous state have already been eliminated. These ones are probably not technical errors.

13.7.2 Negative selfing rate estimates

In short, negative estimates are caused by the excess of heterozygosity compared to the Hardy Weinberg prediction ($2pq$). Supposedly, inbreeding would cause a deficiency of heterozygosity. Possibly due to sampling and extremely minor allele frequency, excess is observed. The identity-disequilibrium-based David2007's method doesn't suffer this problem.

It's hard to see those selfing rates without significantly magnification. I have high-resolution png figures.

14 2008-02-14 Try Kruskal-Wallis

14.1 data

- 96X250k matrix. Each strain of the 96 is picked as the one with lowest error rates among replicates.
- phenotype data is LD(long-day) flowering time from the supplemental table in Keyan's plos genetic paper. No normalization.
- another 'good' data: 96-strain X 193149 SNPs matrix is got after chopping off erroneous (relative to Perlegen data) SNPs.

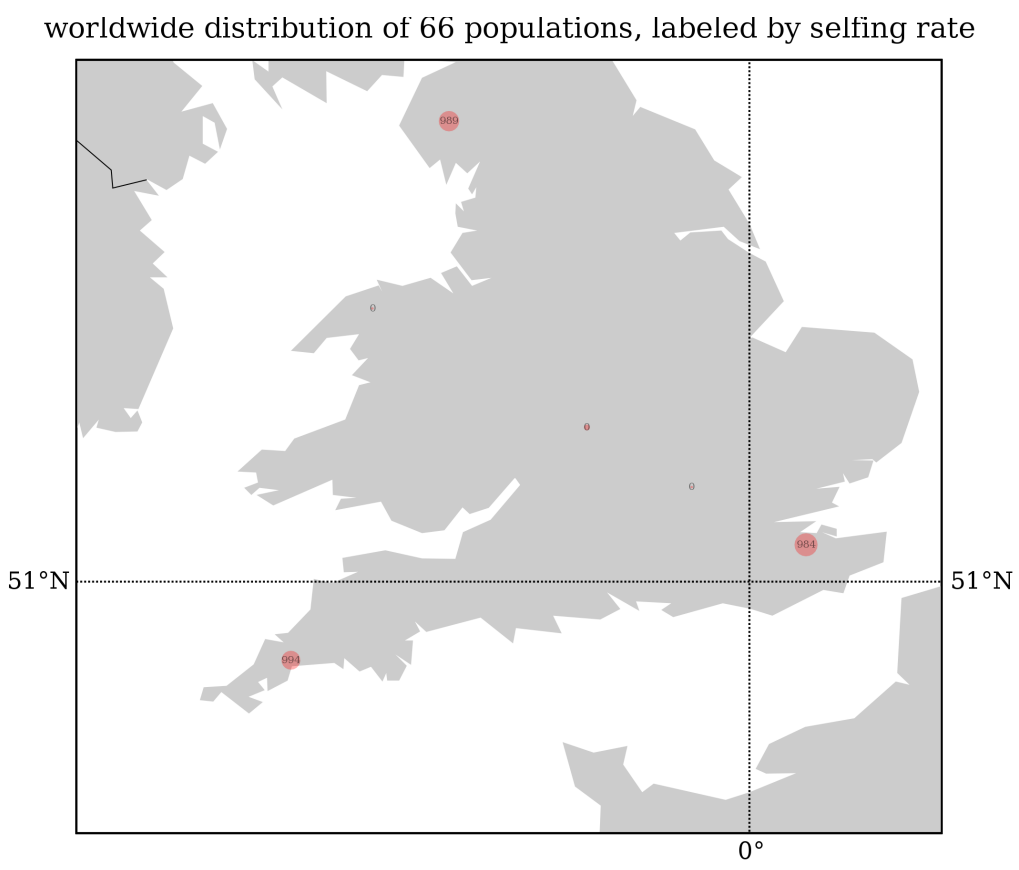

Figure 38: England

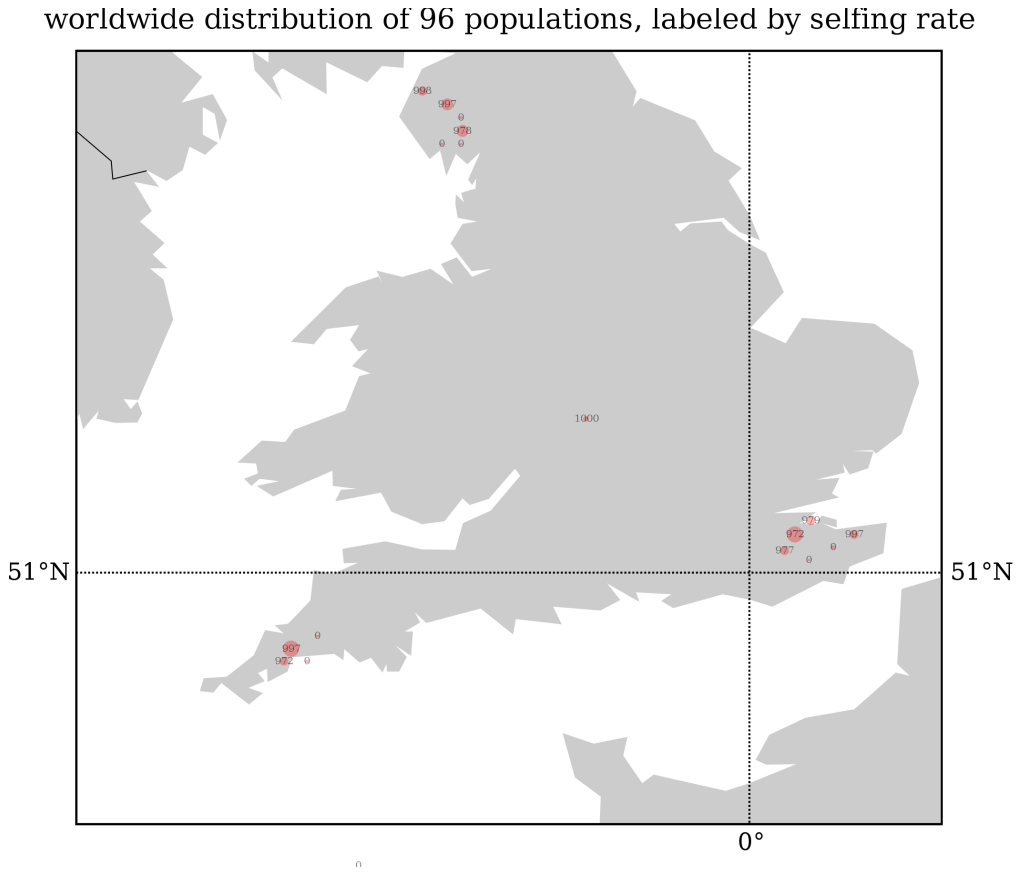

Figure 39: England

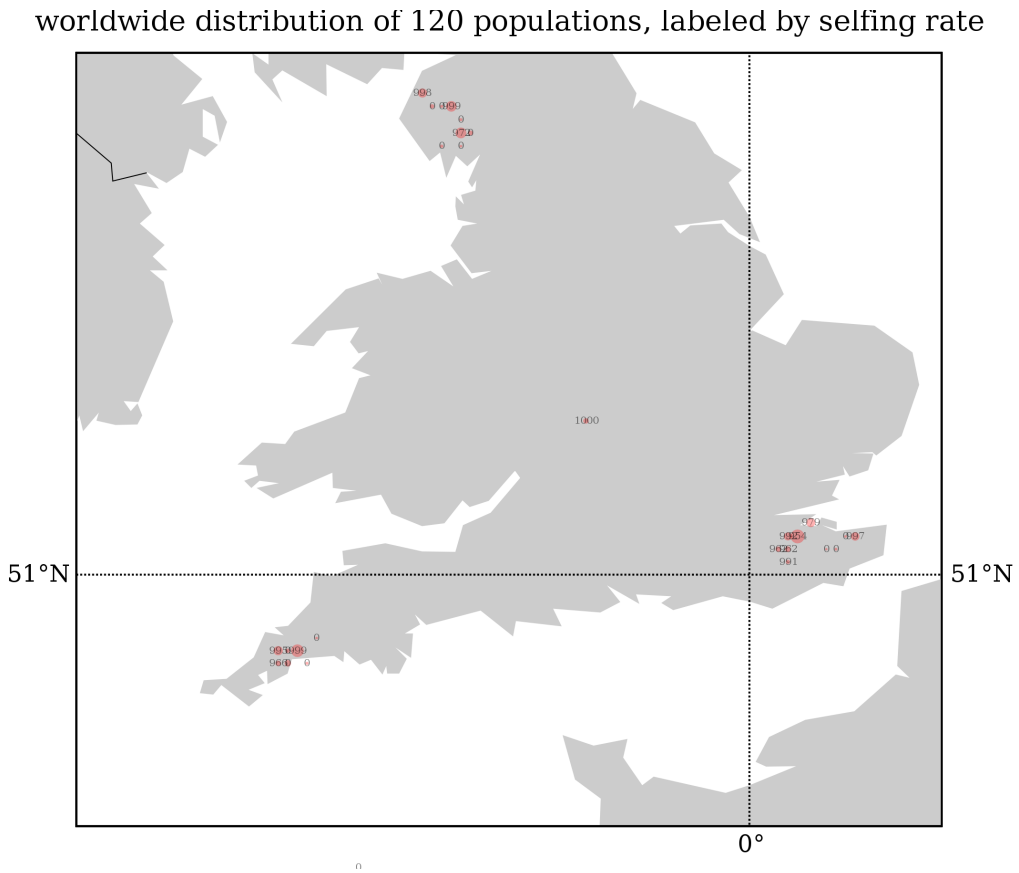

Figure 40: England

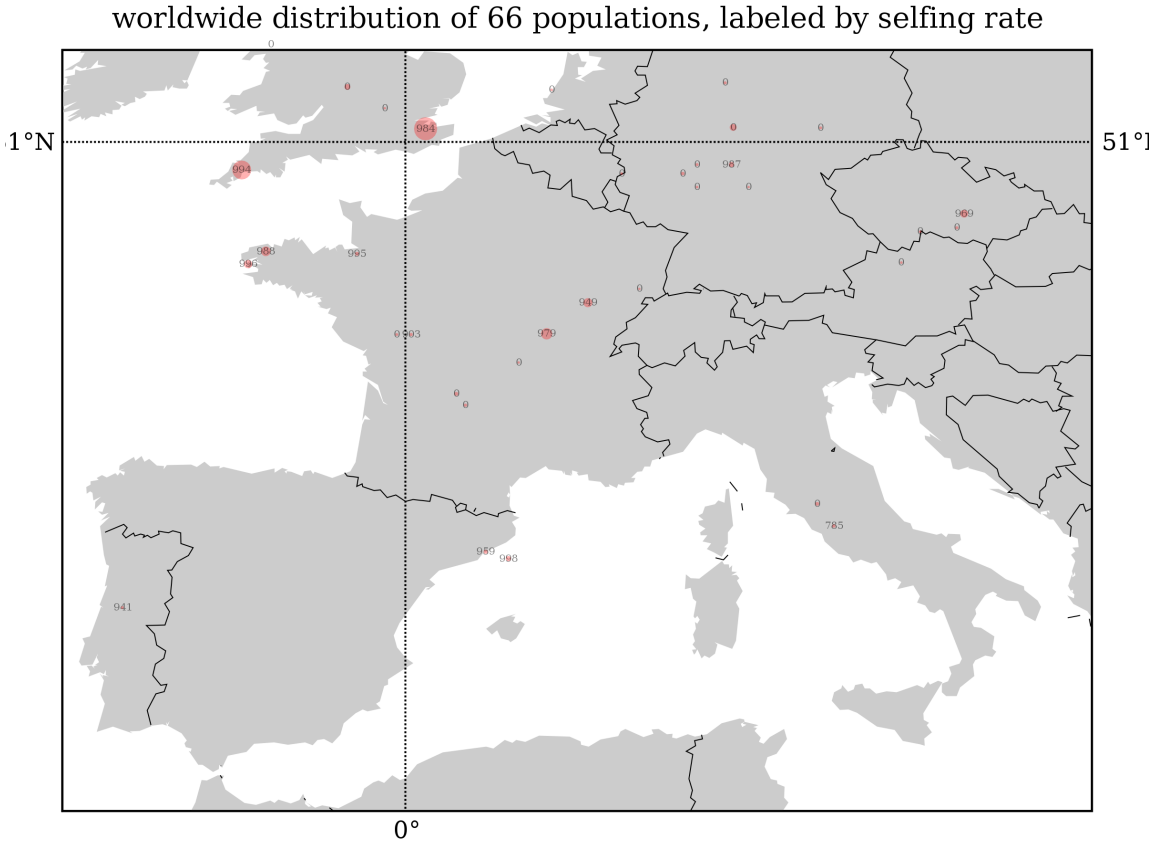

Figure 41: Europe Continent

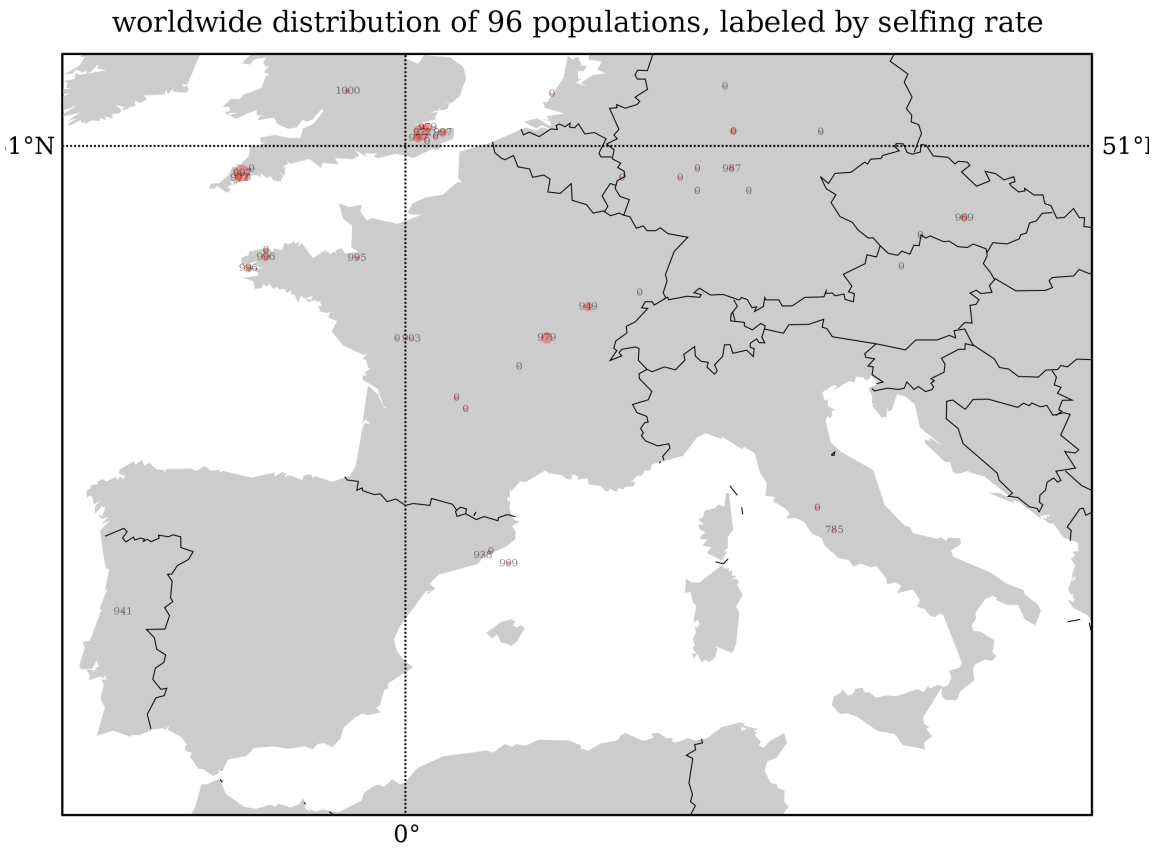

Figure 42: Europe Continent

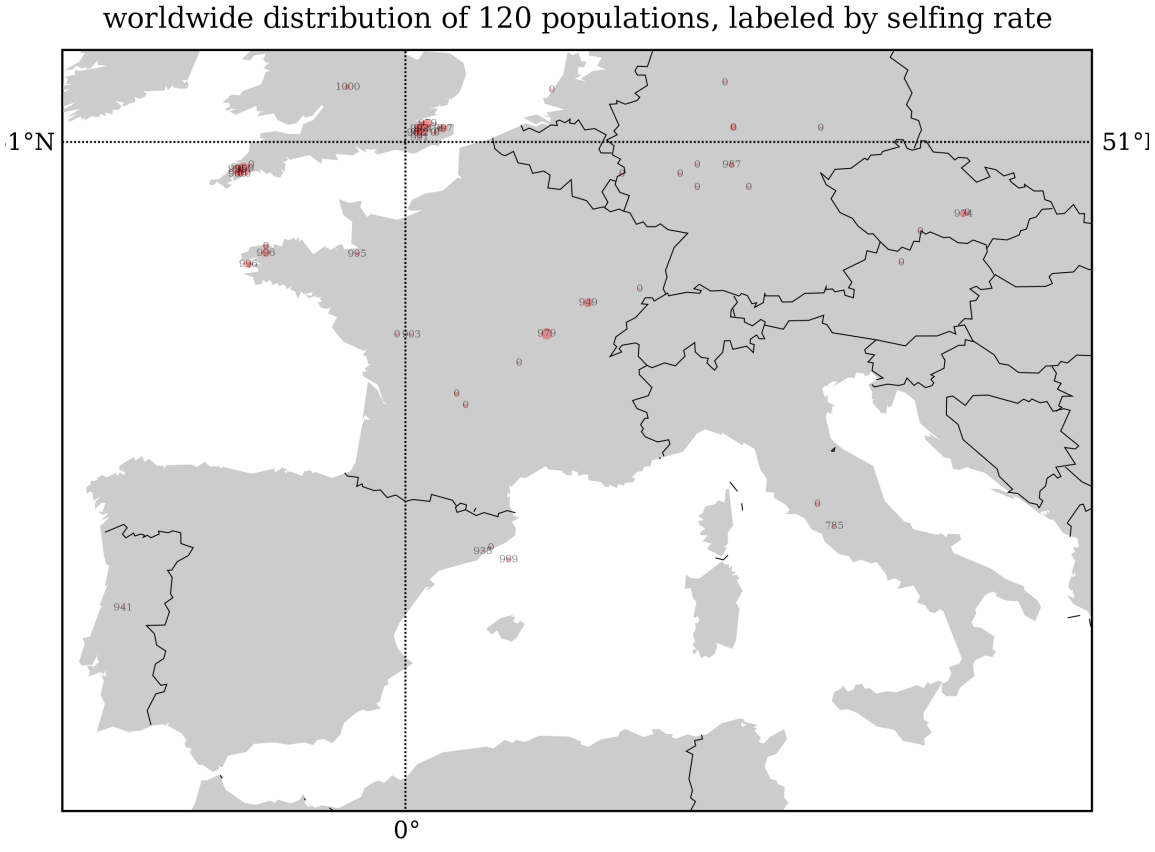

Figure 43: Europe Continent

worldwide distribution of 66 populations, labeled by selfing rate

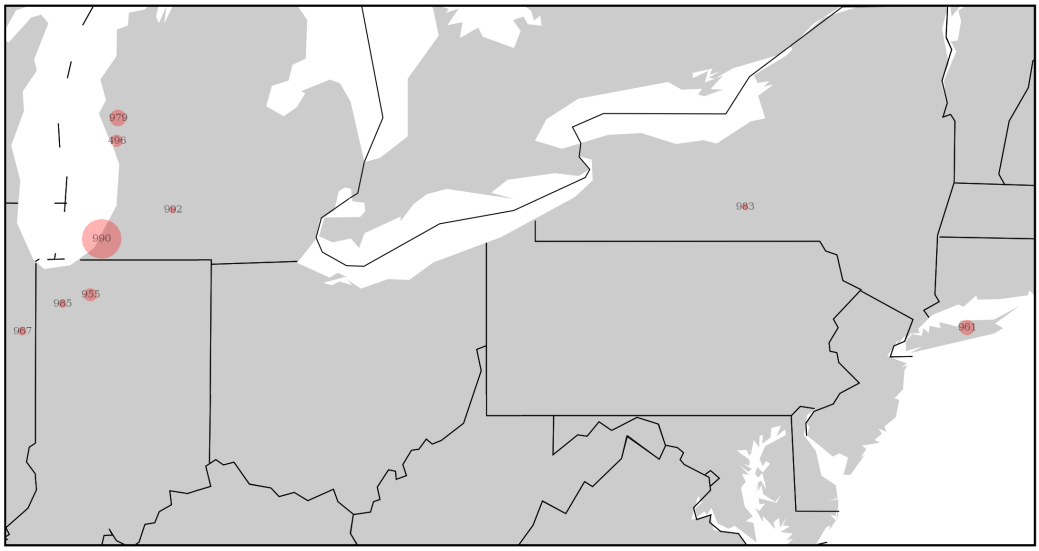

Figure 44: North America

worldwide distribution of 96 populations, labeled by selfing rate

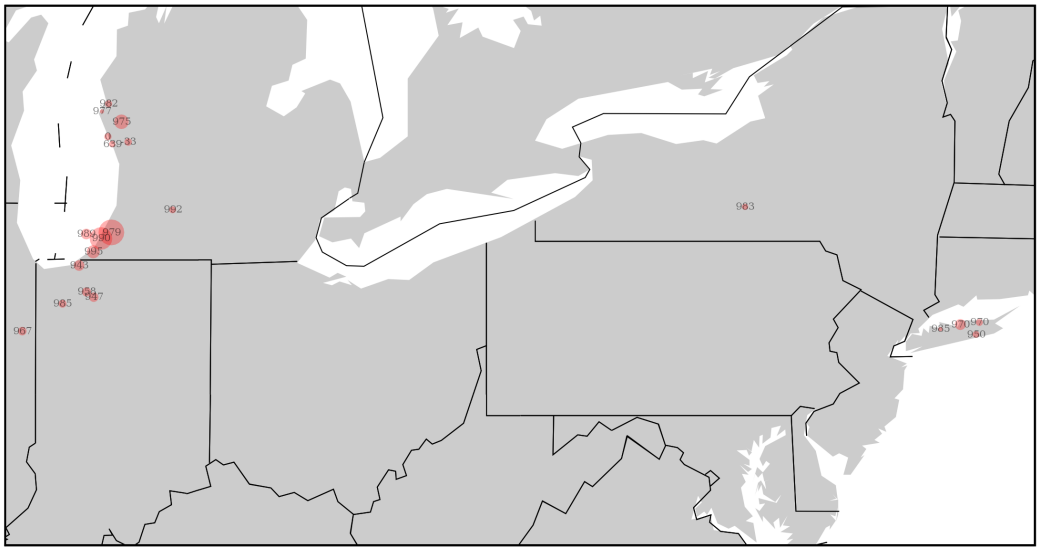

Figure 45: North America

worldwide distribution of 120 populations, labeled by selfing rate

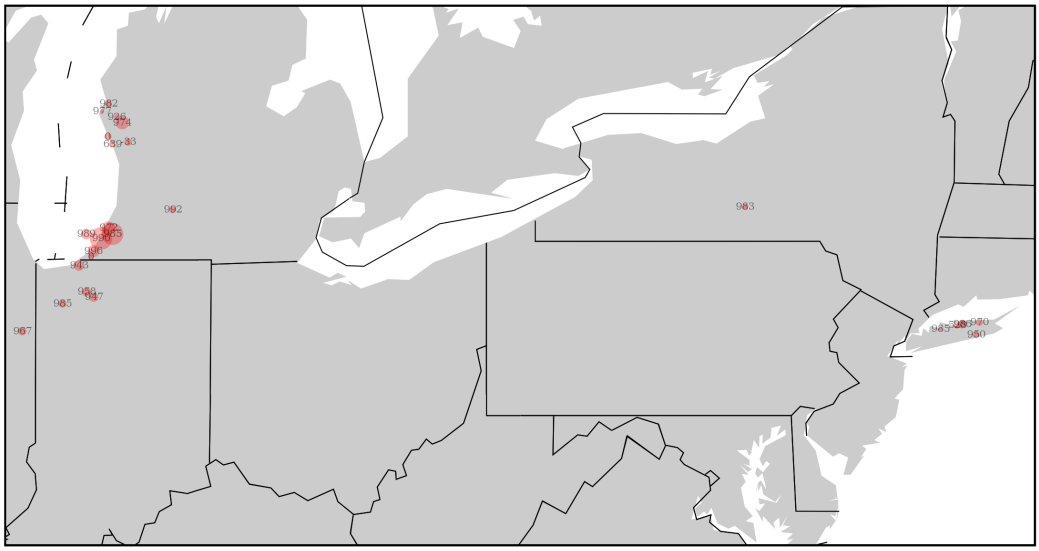

Figure 46: North America

worldwide distribution of 66 populations, labeled by selfing rate

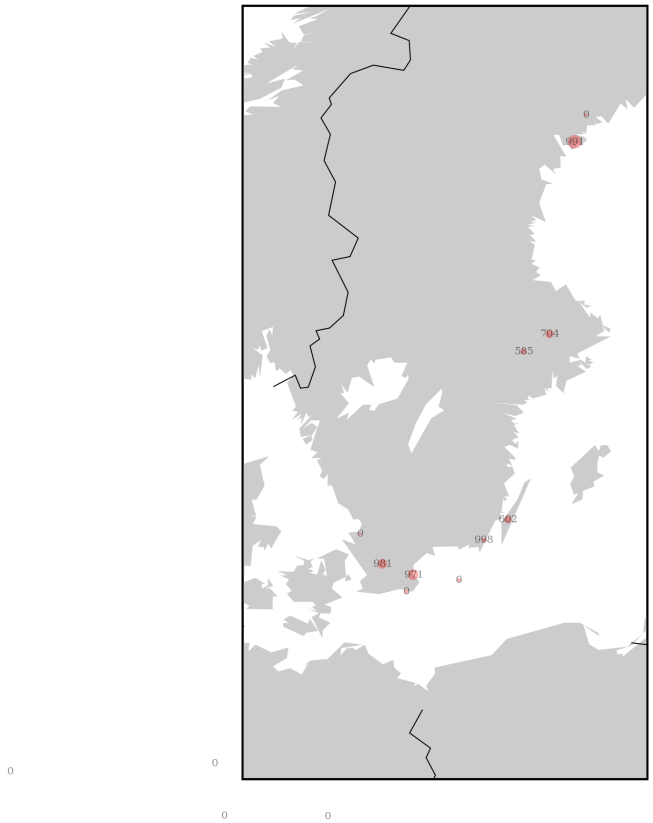

Figure 47: Sweden

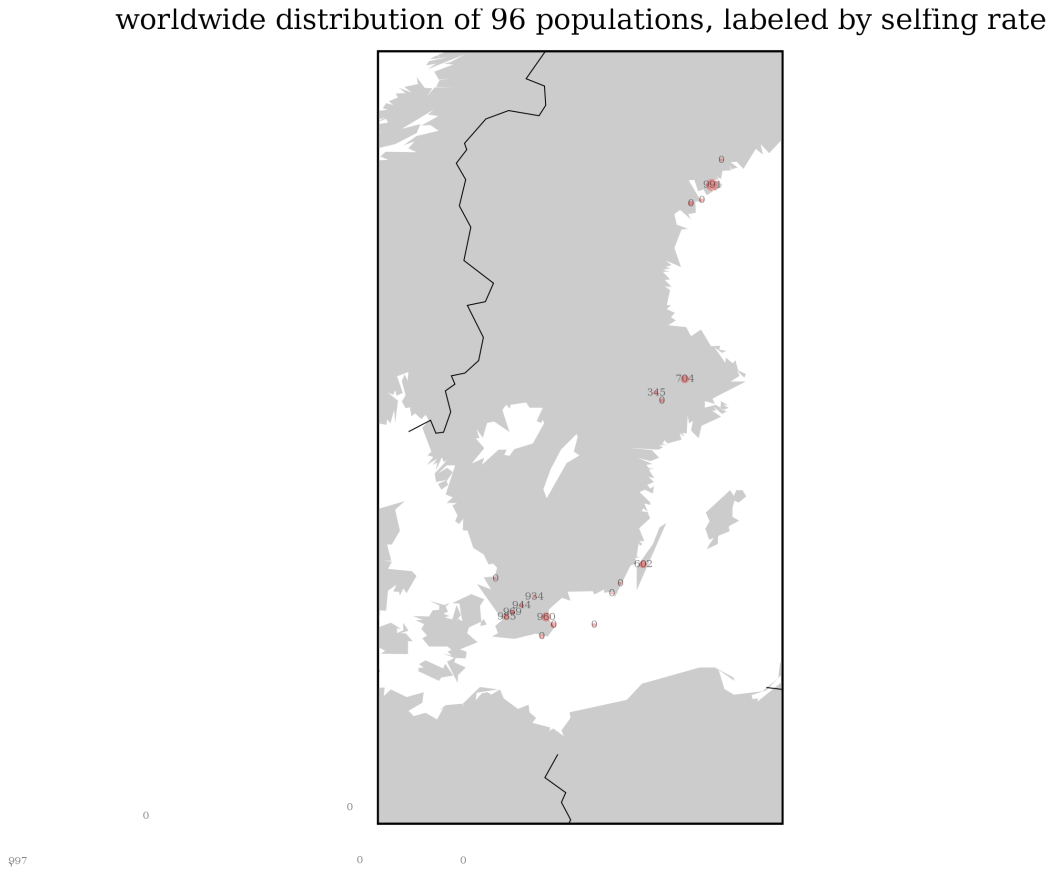

Figure 48: Sweden

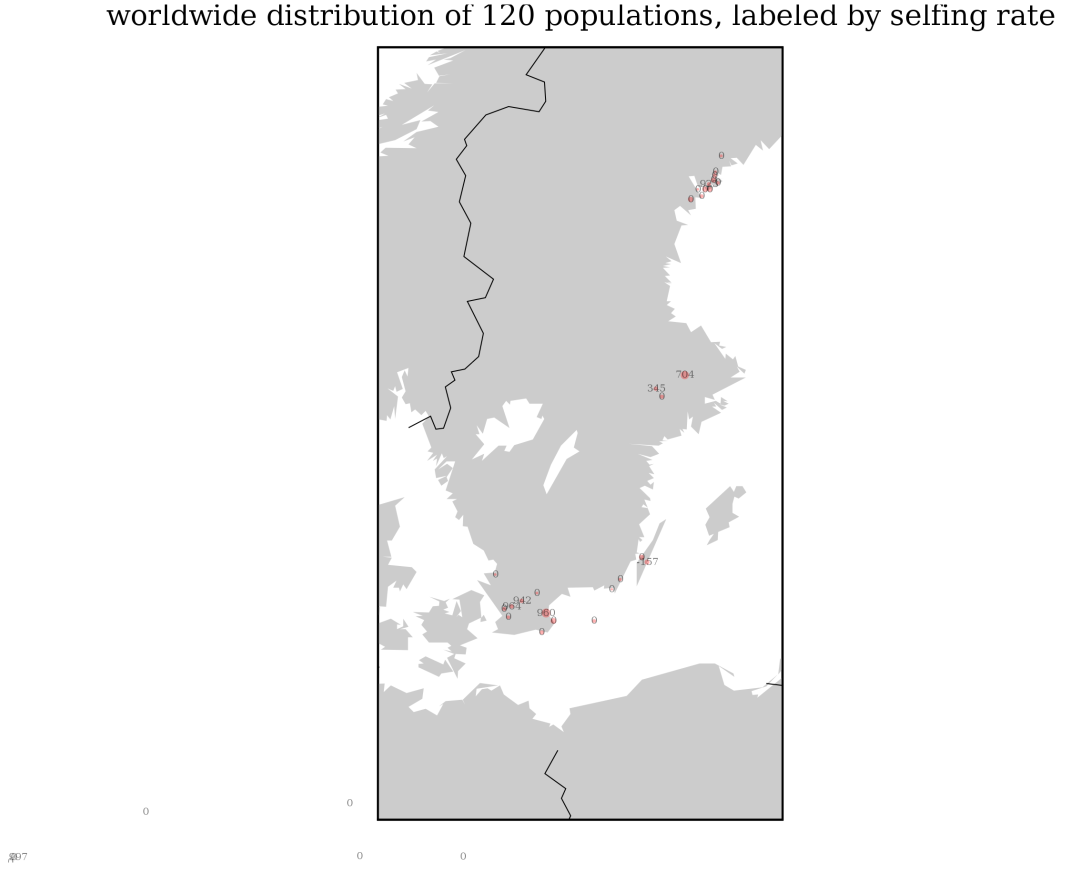

Figure 49: Sweden

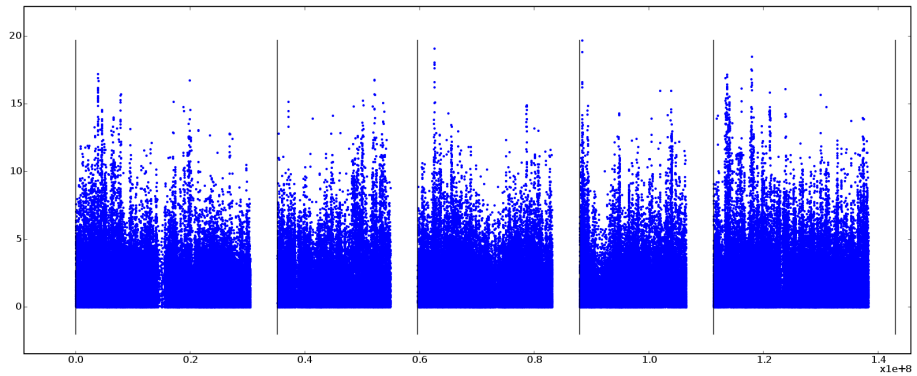

Figure 50: Genome-wide view of the -log(kruskal-wallis p-value) with original data

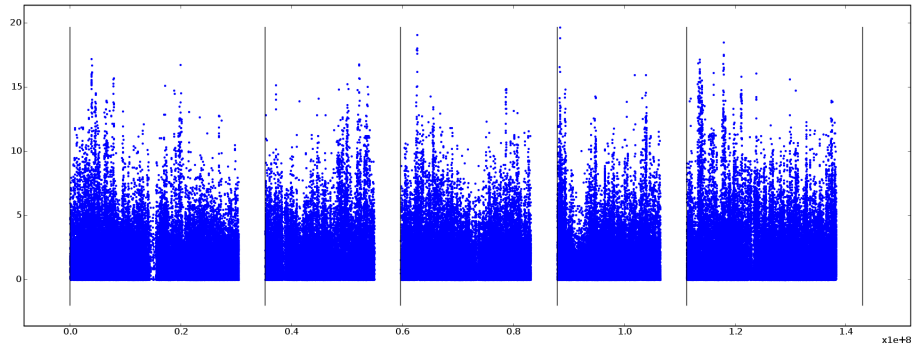

Figure 51: Genome-wide view of the -log(kruskal-wallis p-value) with good data

14.2 genome-wide pattern of SNP pvalue same with good or original data.

14.3 FRI locus

The SNP with most significant p-value (snp chromosome: 4, position: 454422, pvalue: 19.6820786624) is nearly 200kb further down the chromosome from the FRI locus(gene id: 828044. symbol: At4g00650. description: FRIGIDA protein. type of gene: protein-coding. chromosome: 4. start: 269026. stop: 271503. strand: 1).

The most significant SNP is near these genes.

```
gene id: 827921. symbol: At4g01050. description: hydroxyproline-rich glycoprotein family protein. type_of_g
snp chromosome: 4, position: 454422, pvalue: 19.6820786624
gene id: 827924. symbol: At4g01040. description: glycosyl hydrolase family 18 protein. type_of_gene: prote
gene id: 827926. symbol: At4g01030. description: pentatricopeptide (PPR) repeat-containing protein. type_o
gene id: 827917. symbol: At4g01060. description: myb family transcription factor. type_of_gene: protein-co
gene id: 826439. symbol: At4g01020. description: helicase domain-containing protein / IBR domain-containing
gene id: 826427. symbol: At4g01010. description: cyclic nucleotide-regulated ion channel, putative (CNGC13)
gene id: 826450. symbol: At4g01023. description: zinc finger (C3HC4-type RING finger) family protein. type
```

14.4 FLC locus

FLC locus (gene id: 830878. symbol: At5g10140. description: MADS-box protein flowering locus F (FLF). type of gene: protein-coding. chromosome: 5. start: 3173498. stop: 3179449. strand: -1) is 750kb close to a very significant SNP (snp chromosome: 5, position: 2355408, pvalue: 17.1743250098).

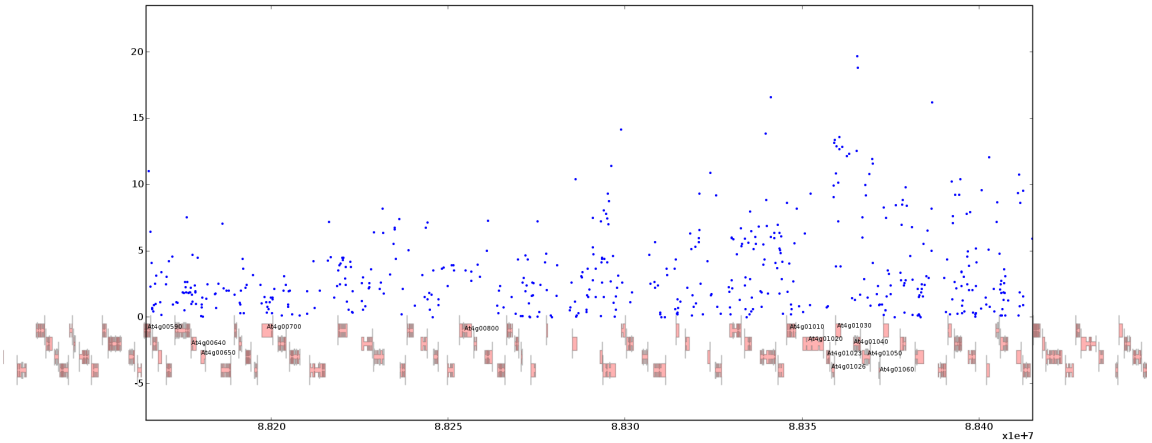

Figure 52: the -log(kruskal-wallis p-value) at FRI locus/At4g00650

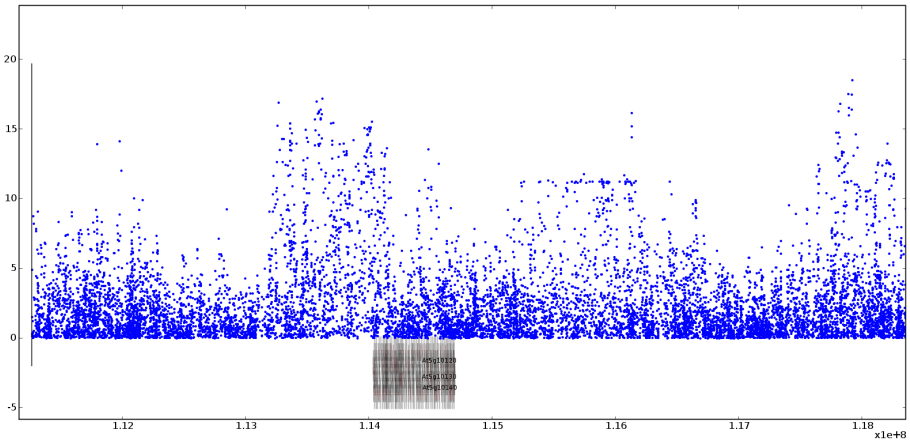

Figure 53: the -log(kruskal-wallis p-value) at FLC locus/At5g10140

15 plant terms

vernalization: during rosette. >= 2 weeks after germination. keep same daylight. set temperature from 18C to 4C. could last a month.

stratification: seed cold treatment, 3 days, 4C. pre-germination. encourage plants to germinate.

flowering time usually refers to the time from germination to first flower open.

16

References

[1] RJ Abbott and MF Gomes. Population genetic structure and outcrossing rate of arabidopsis thaliana(l.) heynh. *Heredity*, 62:411–418, 1989.

[2] Karl W Broman. The genomes of recombinant inbred lines. *Genetics*, 169(2):1133–46, 2005.

[3] Patrice David, Benoit Pujol, Frederique Viard, Vincent Castella, and Jerome Goudet. Reliable selfing rate estimates from imperfect population genetic data. *Molecular Ecology*, 16:2474–2487, 2007. doi:10.1111/j.1365-294X.2007.03330.x.

[4] J Enjalbert and J L David. Inferring recent outcrossing rates using multilocus individual heterozygosity: application to evolving wheat populations. *Genetics*, 156(4):1973–82, 2000.

[5] Olivier Francois, Sophie Ancelet, and Gilles Guillot. Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics*, 174:805–816, 2006. 10.1534/genetics.106.059923.

[6] Hong Gao, Scott Williamson, and Carlos D. Bustamante. A markov chain monte carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics*, 176:1635–1651, 2007. 10.1534/genetics.107.072371.

[7] Gilles Guillot, Arnaud Estoup, Frederic Mortier, and Jean Francois Cosson. A spatial statistical model for landscape genetics. *Genetics*, 170:1261–1280, 2005. 10.1534/genetics.104.033803.

[8] Philippe Jarne and Josh R Auld. Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals. *Evolution Int J Org Evolution*, 60(9):1816–24, 2006.

[9] M Nordborg and P Donnelly. The coalescent process with selfing. *Genetics*, 146(3):1185–95, 1997.

[10] Magnus Nordborg, Tina T. Hu, Yoko Ishino, Jinal Jhaveri, Christopher Toomajian, Honggang Zheng, Erica Bakker, Peter Calabrese, Jean Gladstone, Rana Goyal, Mattias Jakobsson, Sung Kim, Yuri Morozov, Badri Padhukasahasram, Vincent Plagnol, Noah A. Rosenberg, Chitiksha Shah, Jeffrey D. Wall, Jue Wang, Keyan Zhao, Theodore Kalbfleisch, Vincent Schulz, Martin Kreitman, and Joy Bergelson. The pattern of polymorphism in arabidopsis thaliana. *PLoS Biology*, 3:e196 EP –, 2005.

[11] Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

[12] Alan Robertson and William G. Hill. Deviations from hardy-weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics*, 107:703–718, 1984.

[13] B. S. Weir and C. Clark Cockerham. Estimating f-statistics for the analysis of population structure. *Evolution*, 38:1358–1370, 1984.