The Scale of Population structure in partial selfer Arabidopsis thaliana

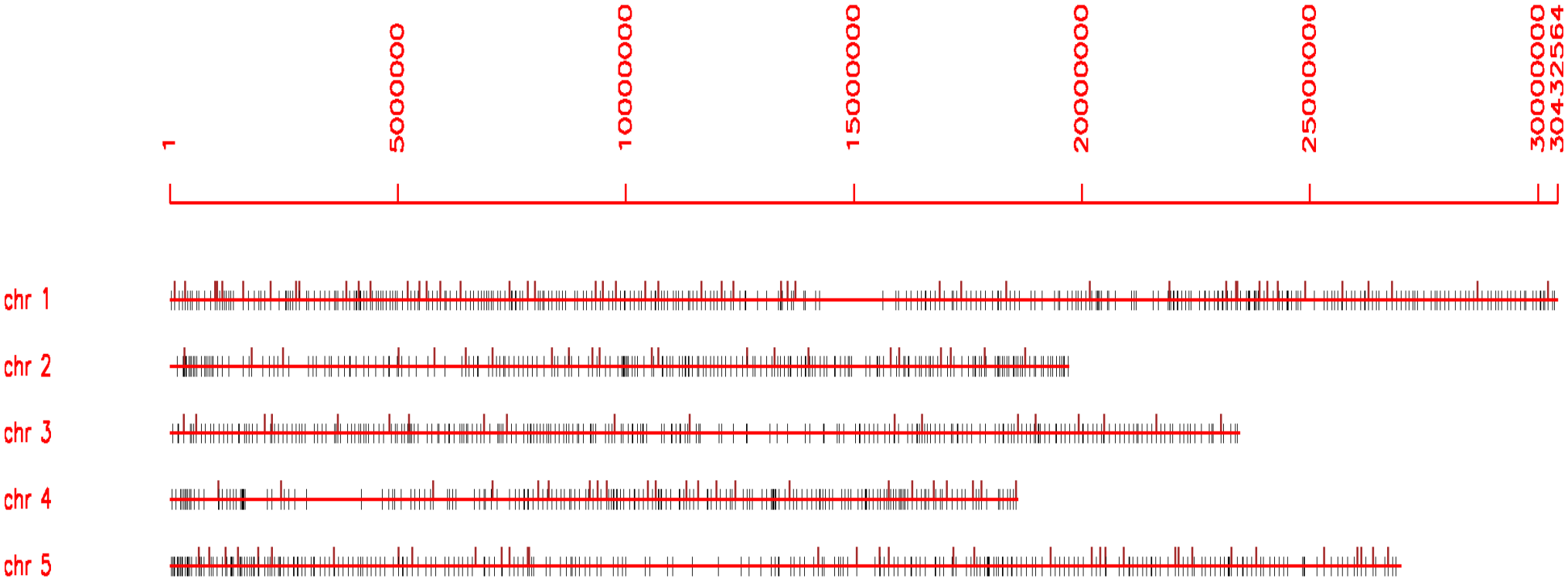
Lab Meeting

Yu Huang

2007-10-23

Data

- 6634 strains from around the world
- 149 SNPs



red tick: 149 SNPs

black crossing tick/block: 2010 fragments

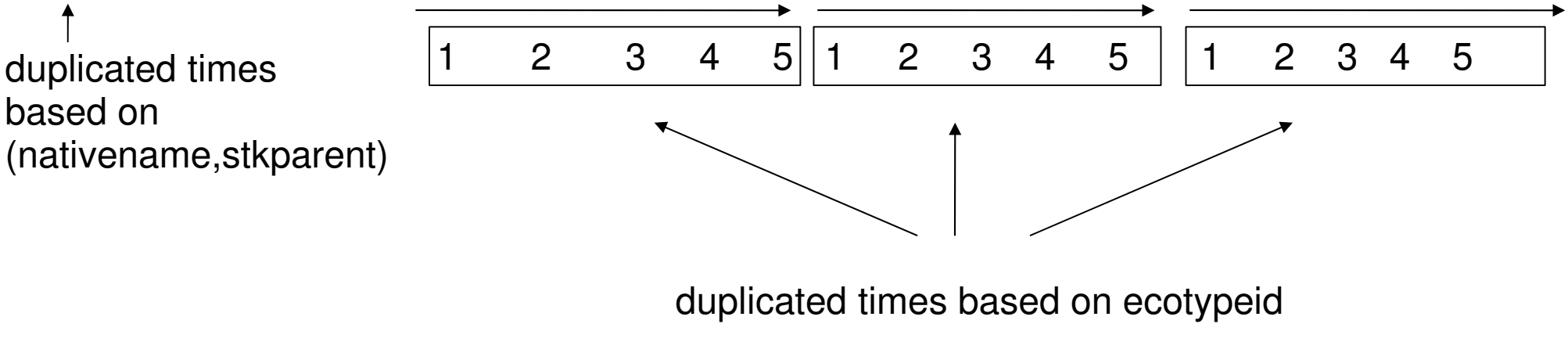
Questions

- What's the genotyping error rate (duplicate, compare to 2010, etc)?
- bogus heterozygous calls?
- Population Structure (by identity strains, etc)
- Which strains are recent products of outcrossing?
- How many selfing generations are they away from the last out-crossing?
- how's selfing rate in different populations?

the extent of duplication

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



inconsistent rate among duplicated calls ~ 0.69%

2010 versus 149 SNP

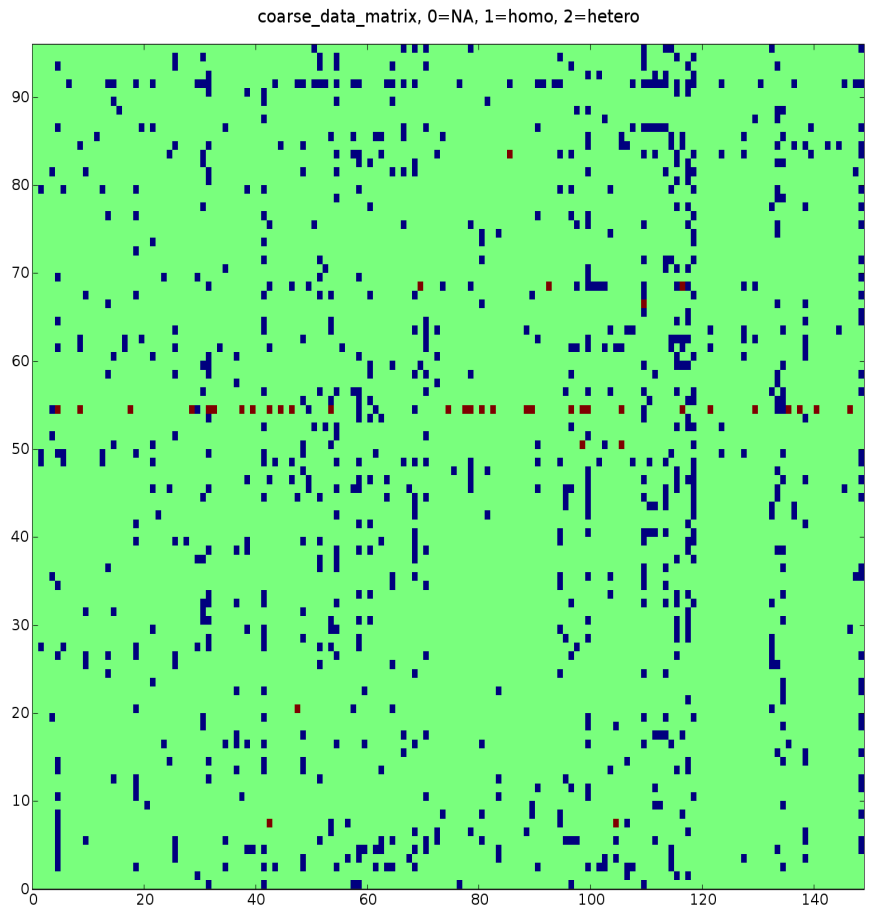
NA

Homozygous

Heterozygous

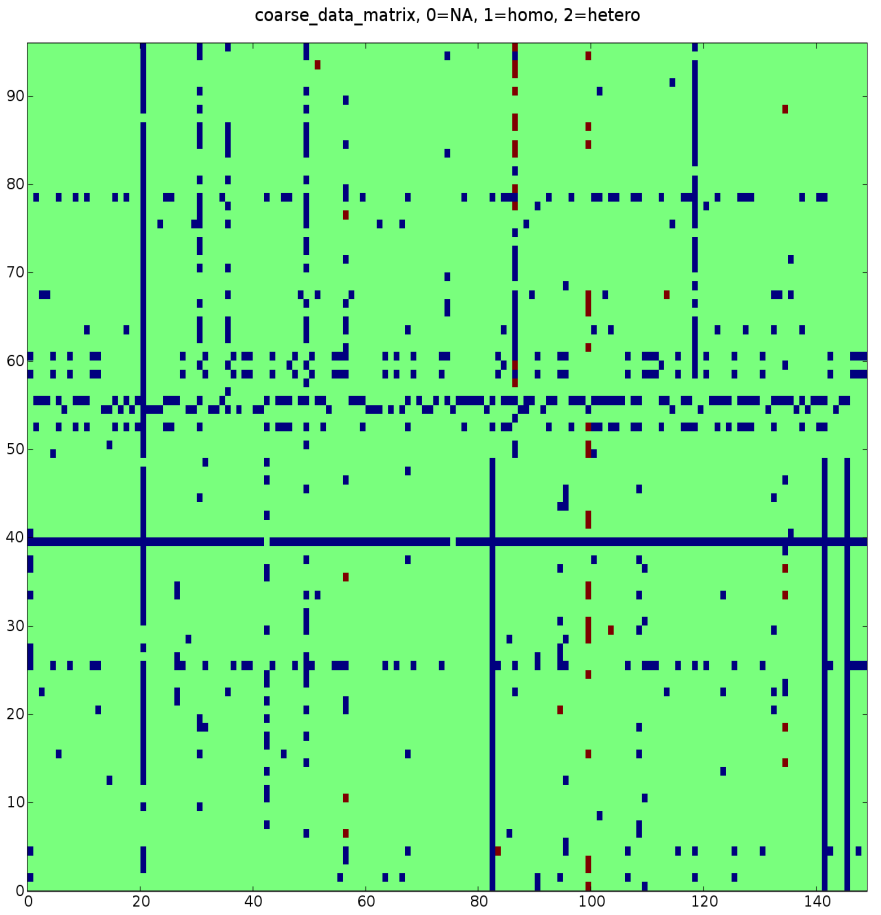
2010 data

149-SNP data



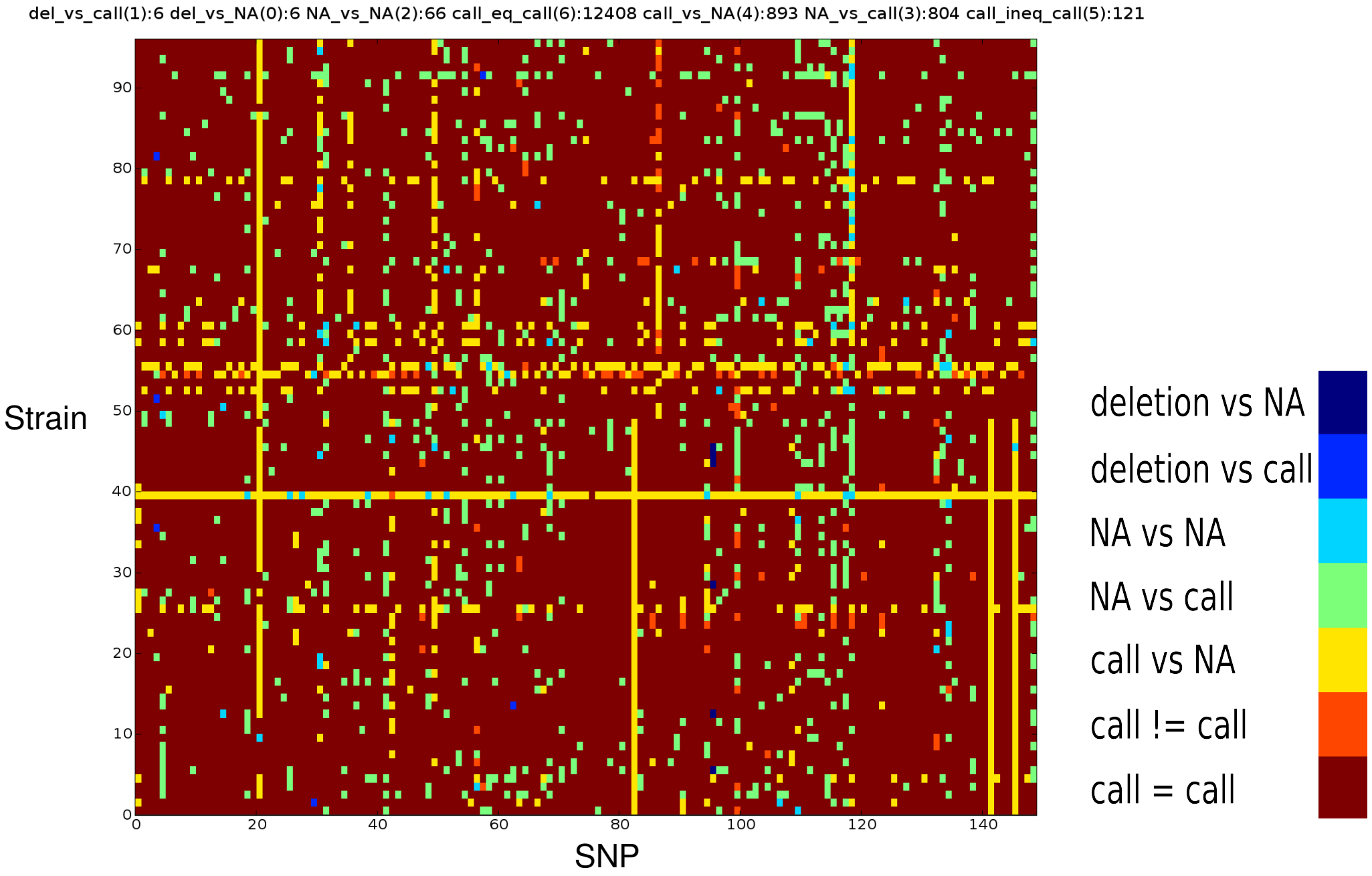
Strain

SNP



SNP

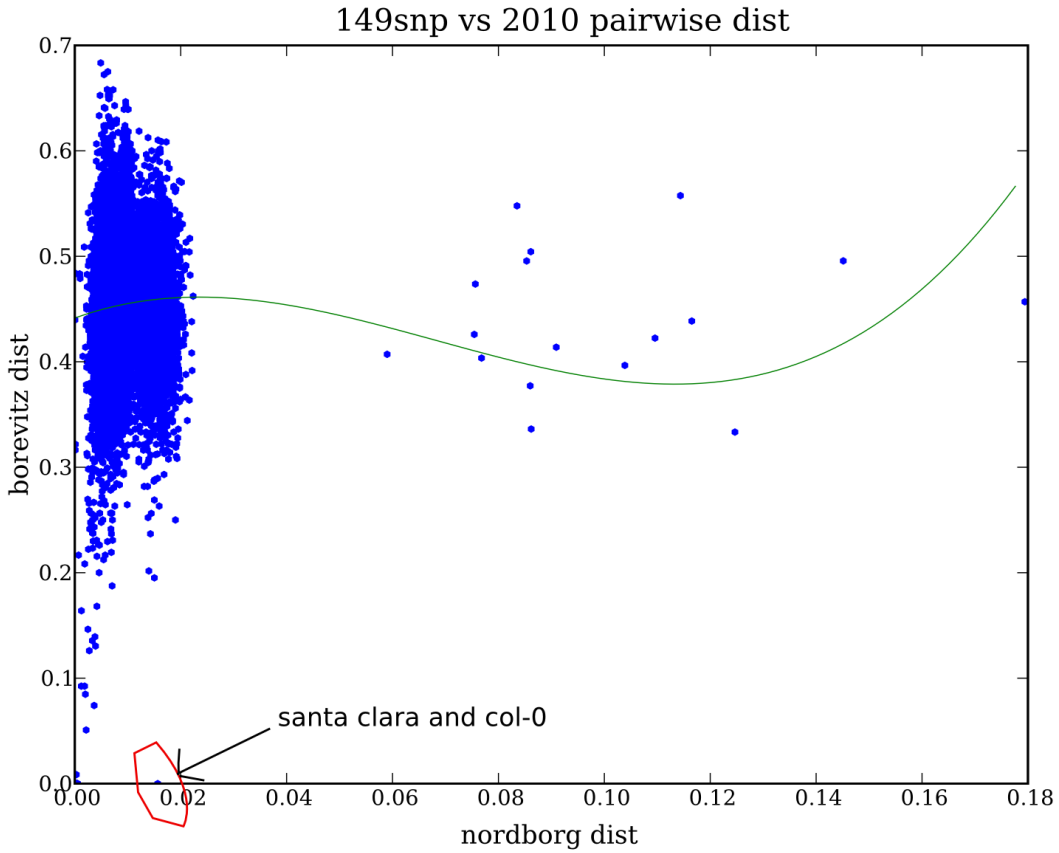
2010 versus 149 SNP



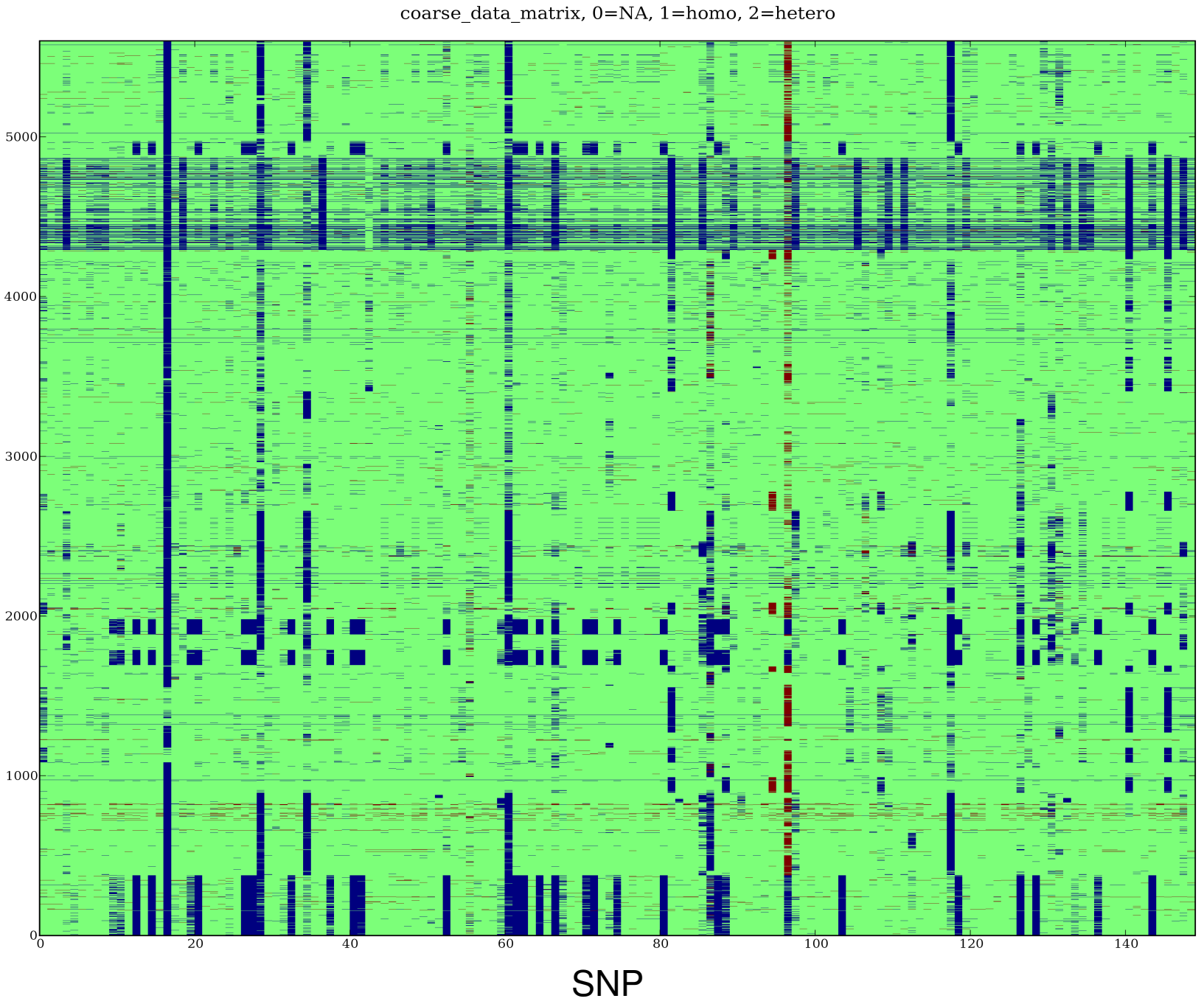
Error rate estimate

- optimistic error rate is 121/ (12408+121) ~0.96%
- including call-vs-NA and NA-vs-call (121+893+804)/(12408.0+121+893+804) ~ 12.78%

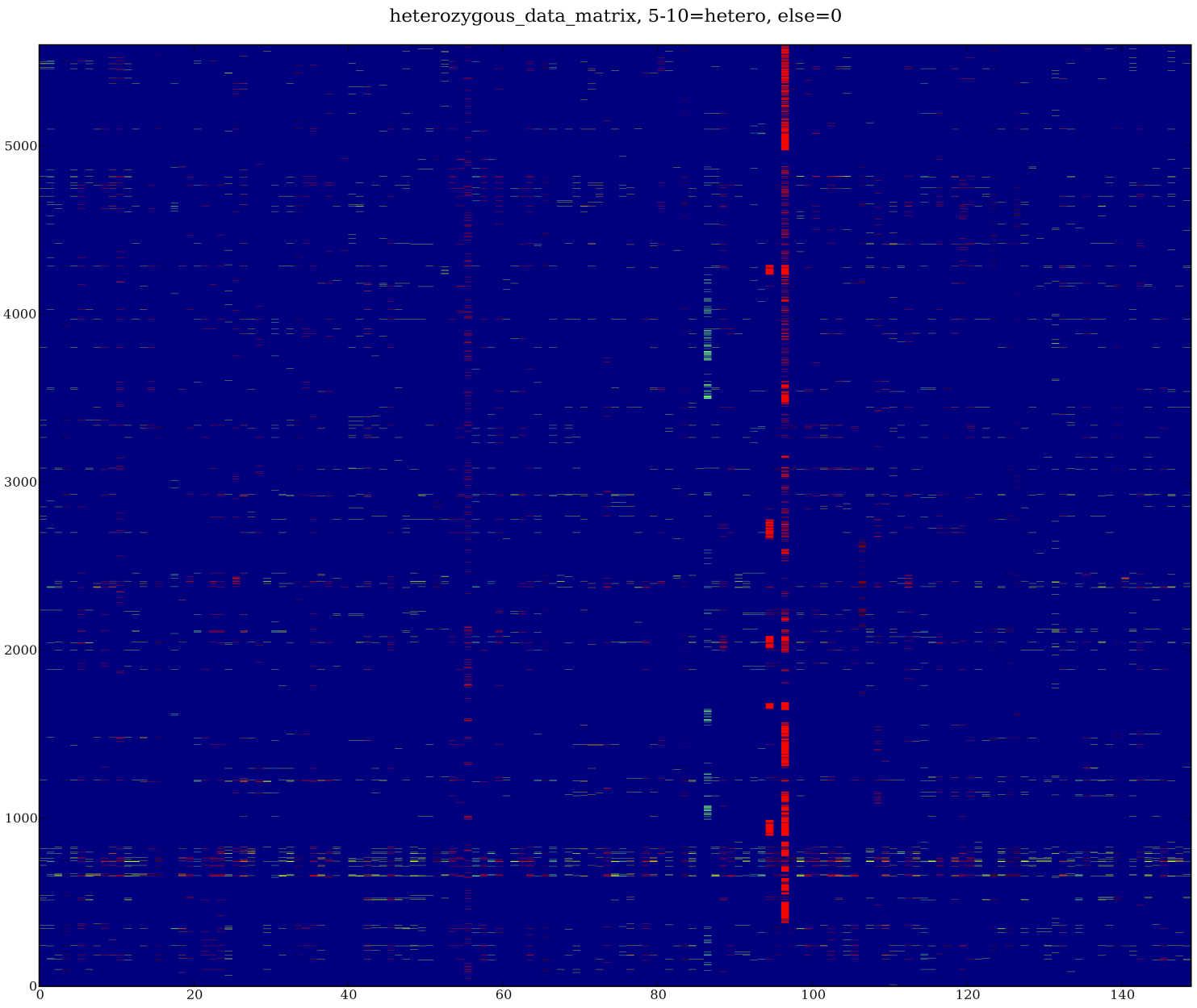
pairwise distance comparison of **192** strains



Bogus Hetero-calls



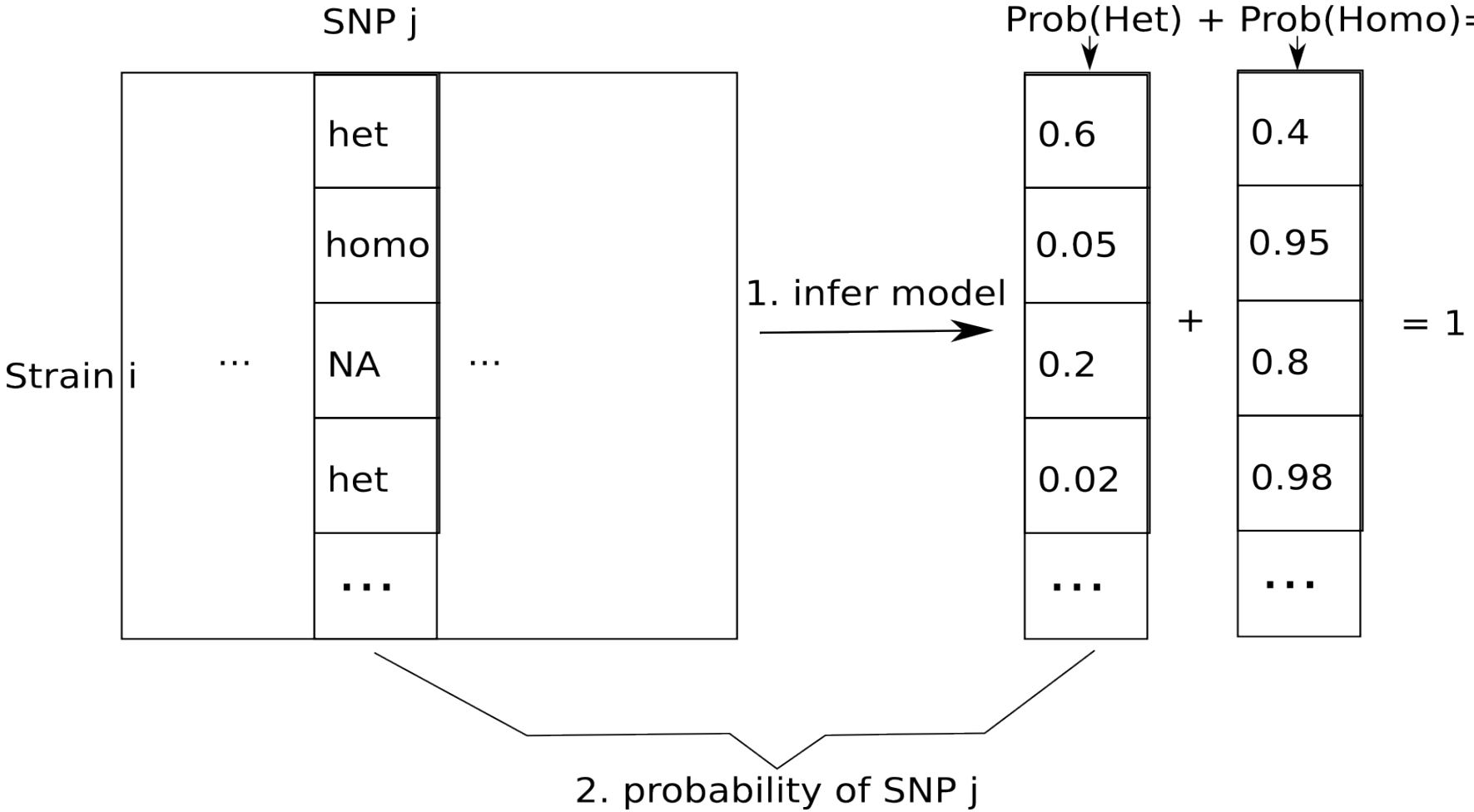
Non-background -> Heterozygous

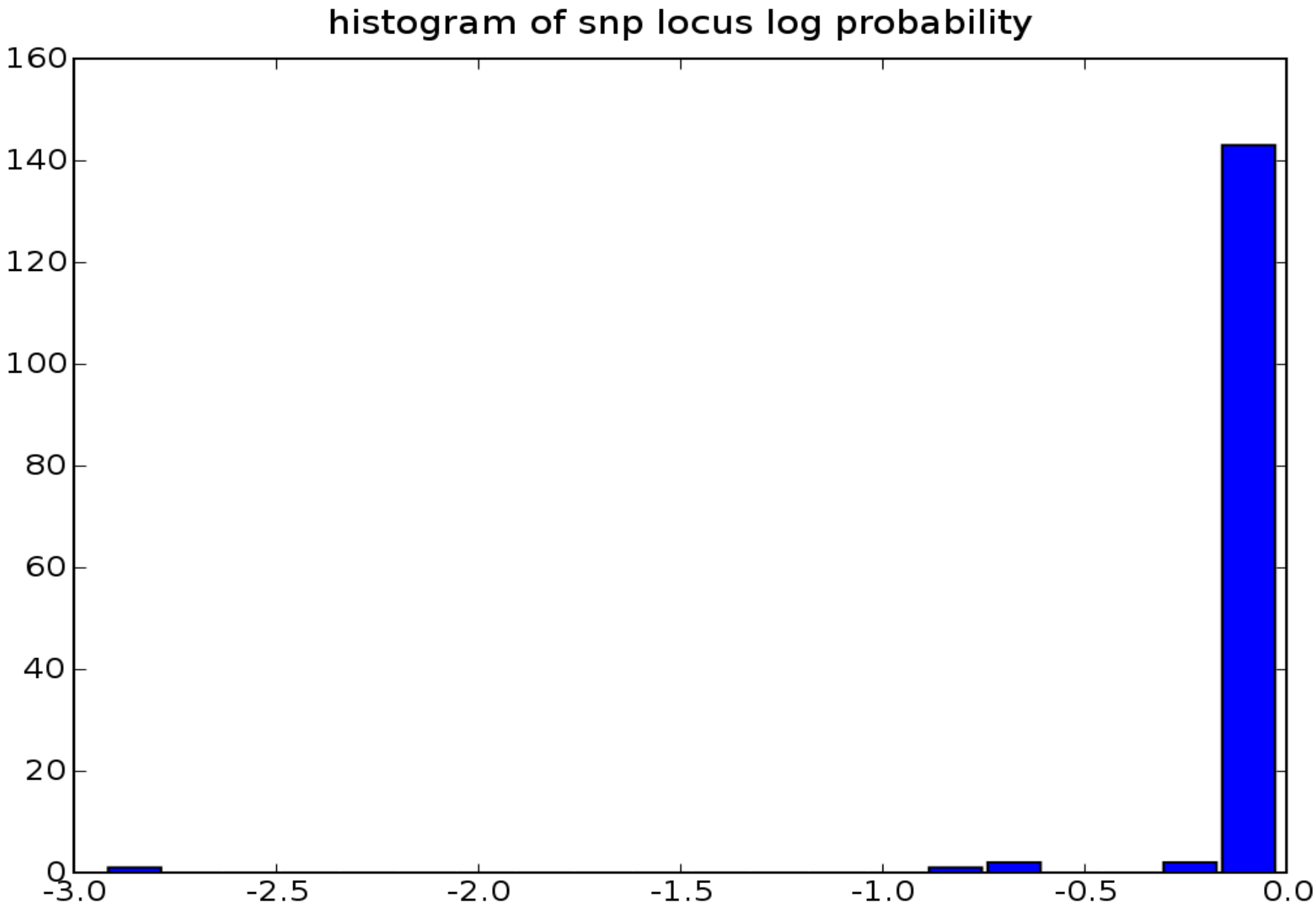


probability model to identify bogus het

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

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 $Strain_i$.

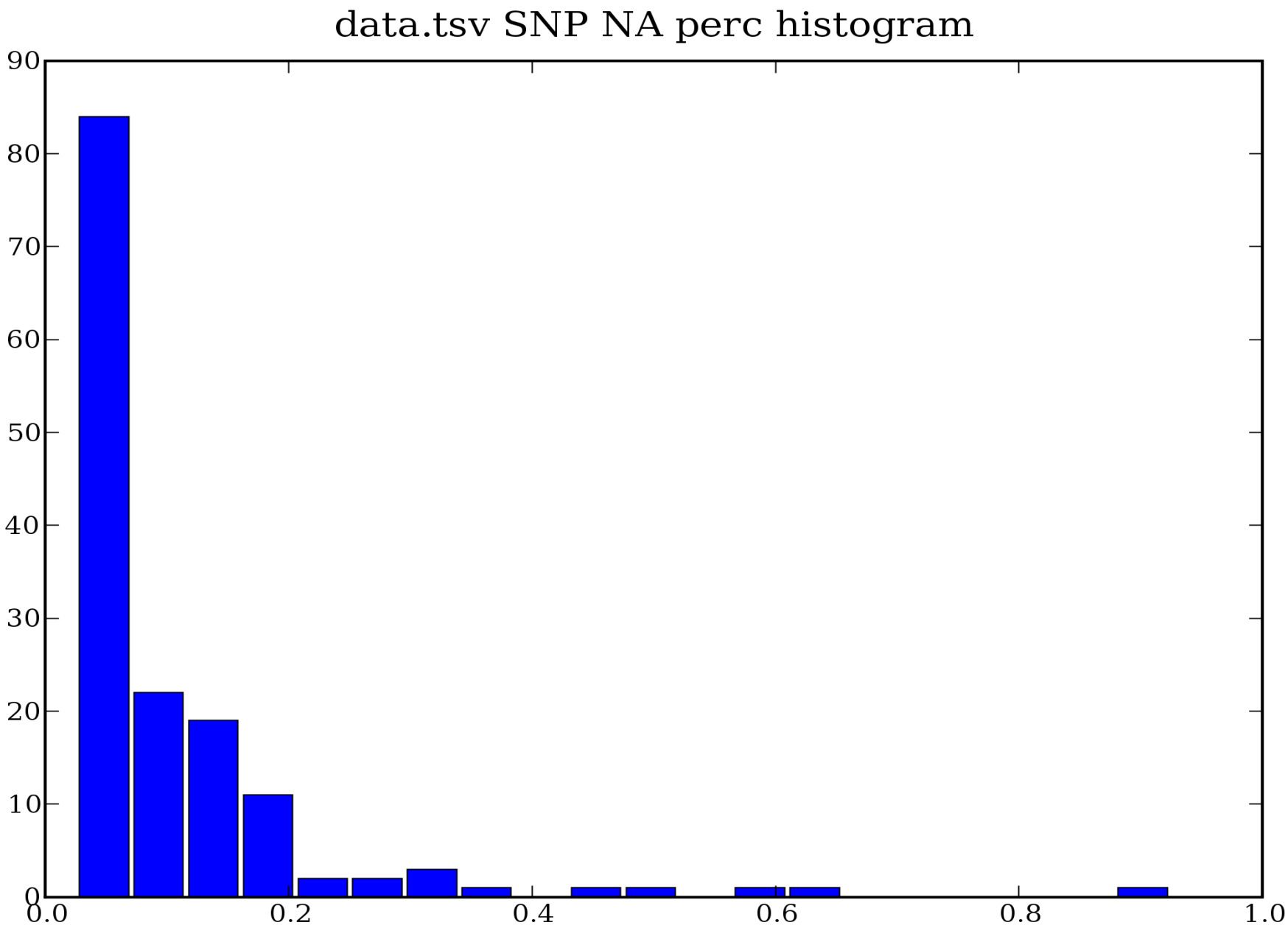


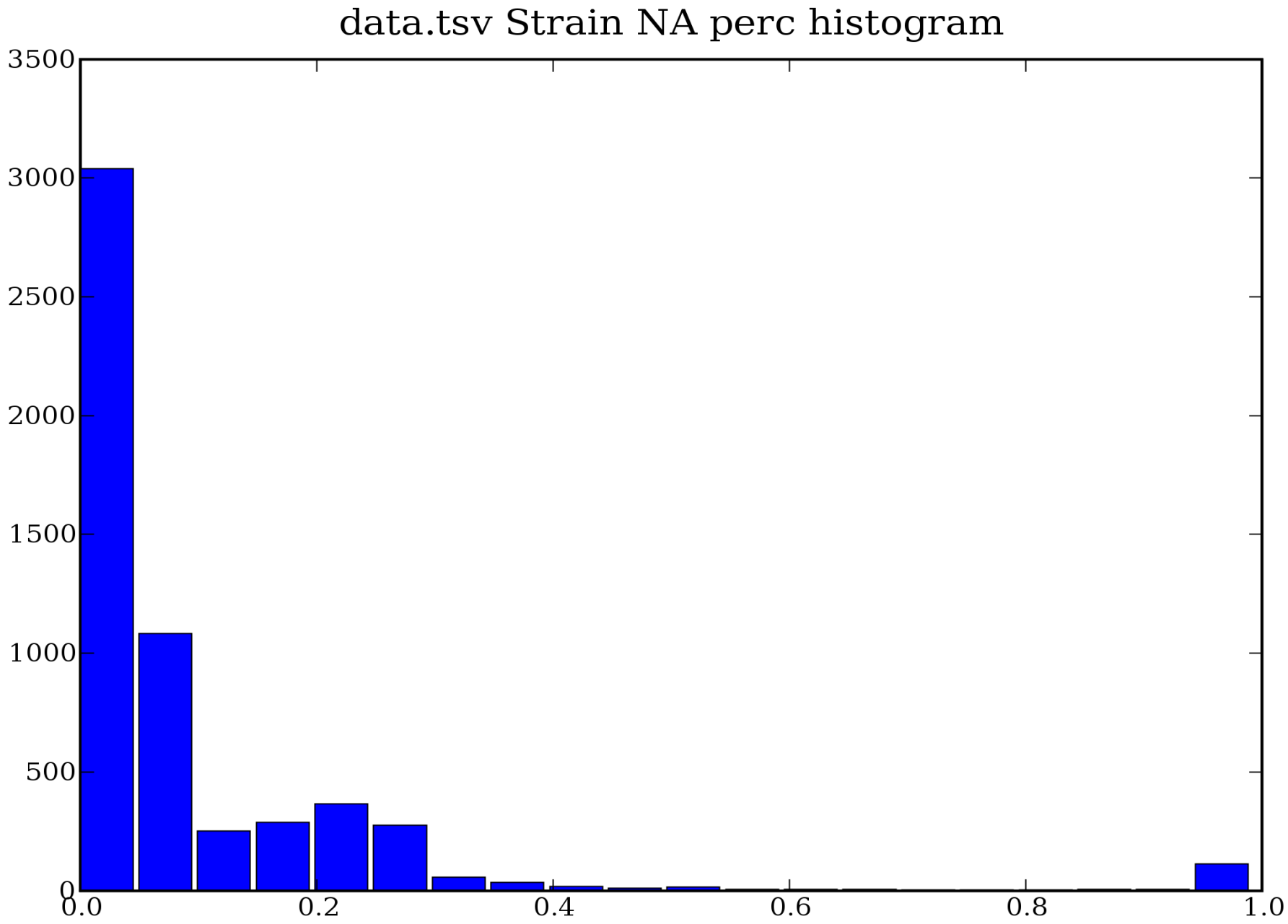

histogram of snp locus log probability
160
140
120
100
80
60
40
20
0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0

cause of heterozygous calls

- technical error
- segmental duplication in all strains
- segmental duplication in some strains
- real heterozygous

NA

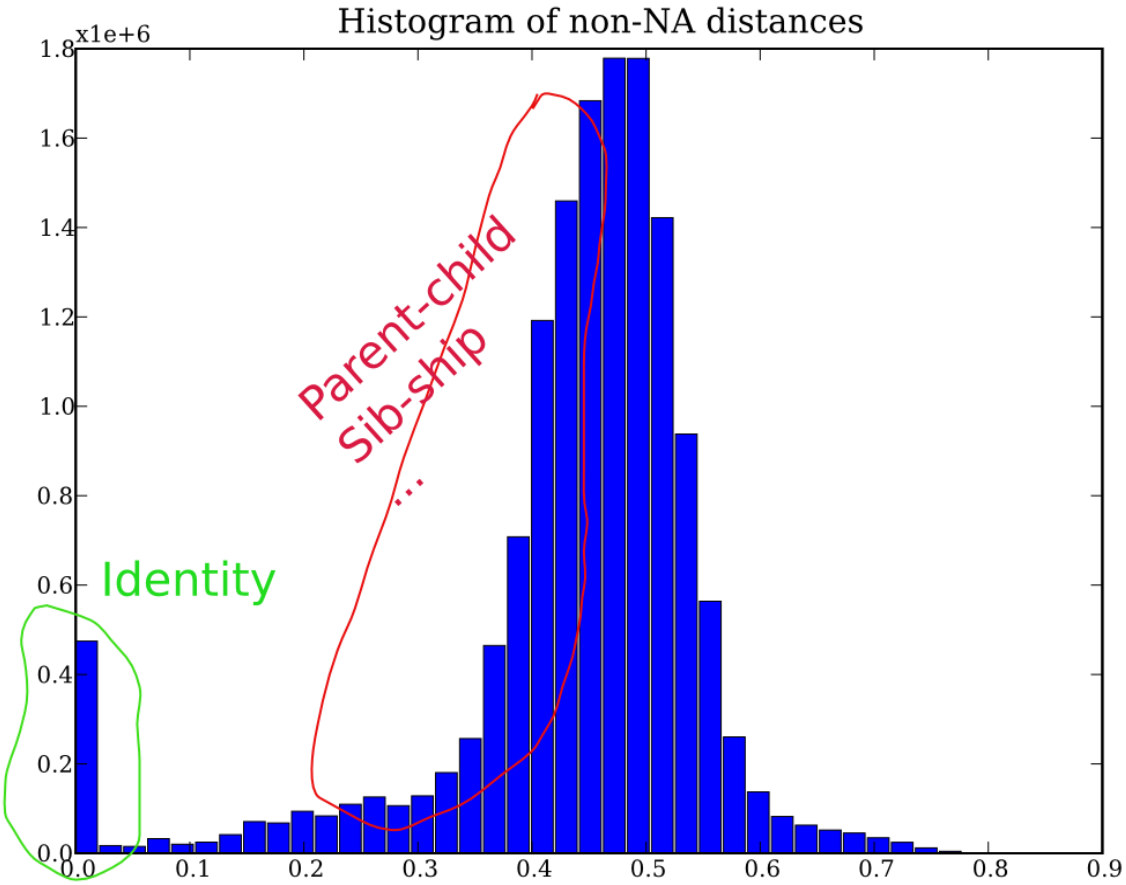

data.tsv SNP NA perc histogram
90
80
70
60
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0


data.tsv Strain NA perc histogram
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0

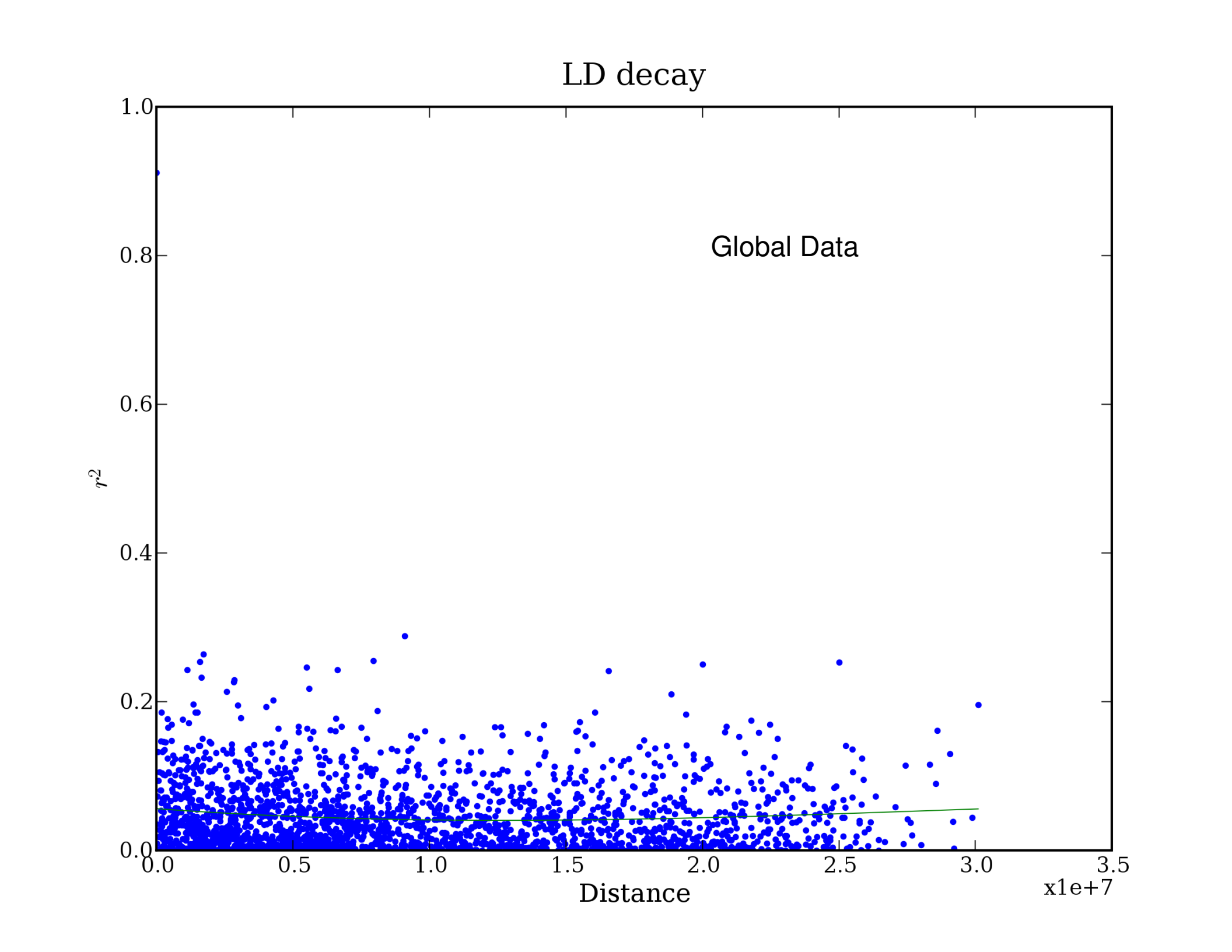
NA filtering result

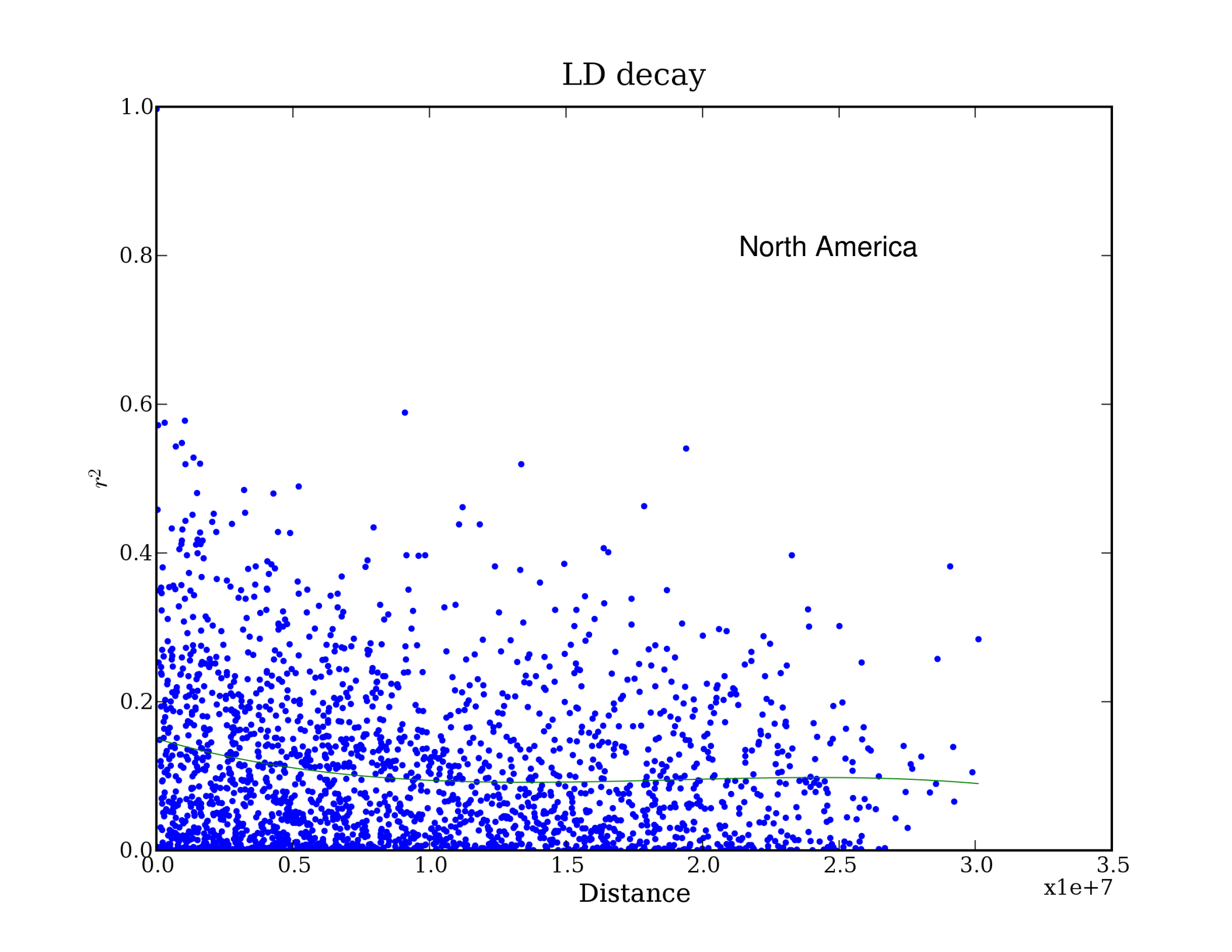
- 194 strains with >=40 % NA, removed
- 5 SNPs with >=40% NA, removed

Population Structure

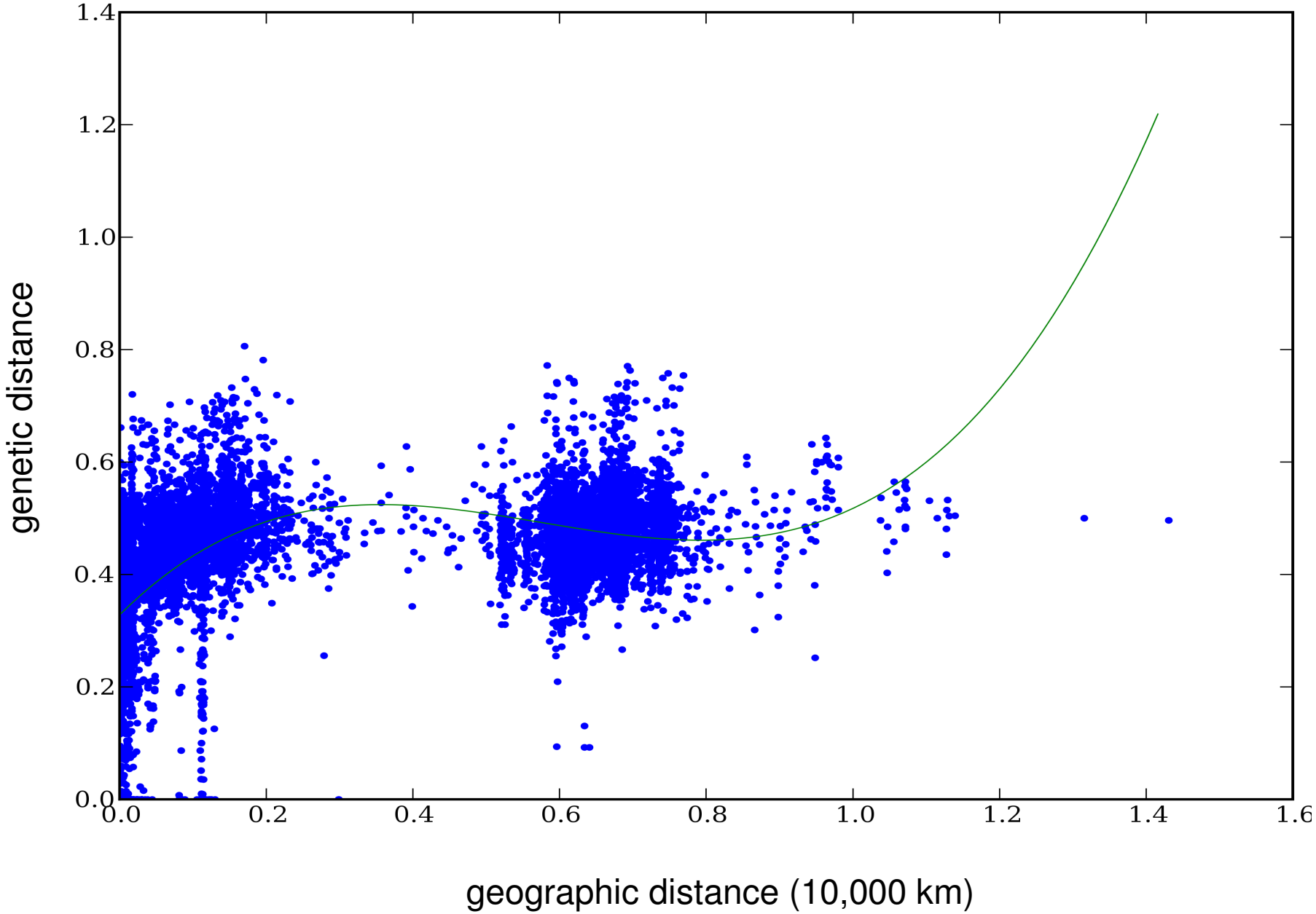


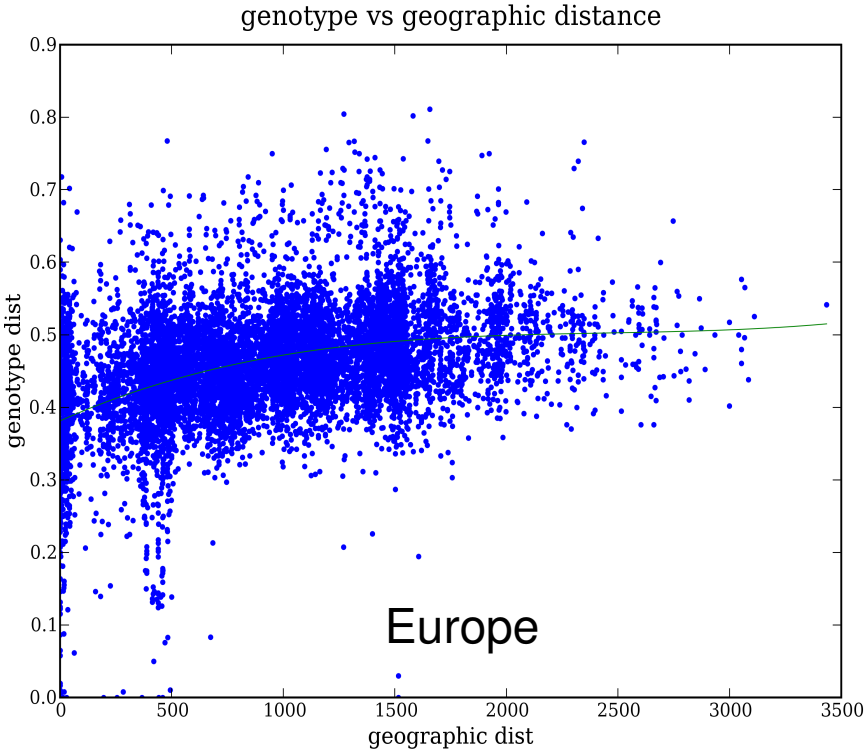
Hamming(Binary) Distance between two strains

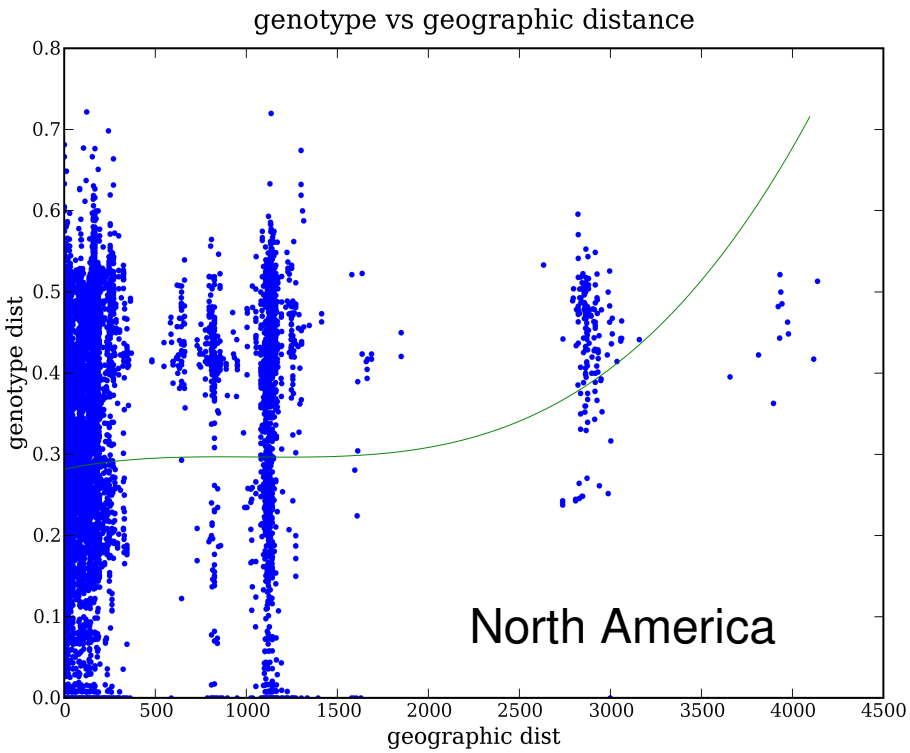

LD decay
Global Data
r^2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Distance
x1e+7

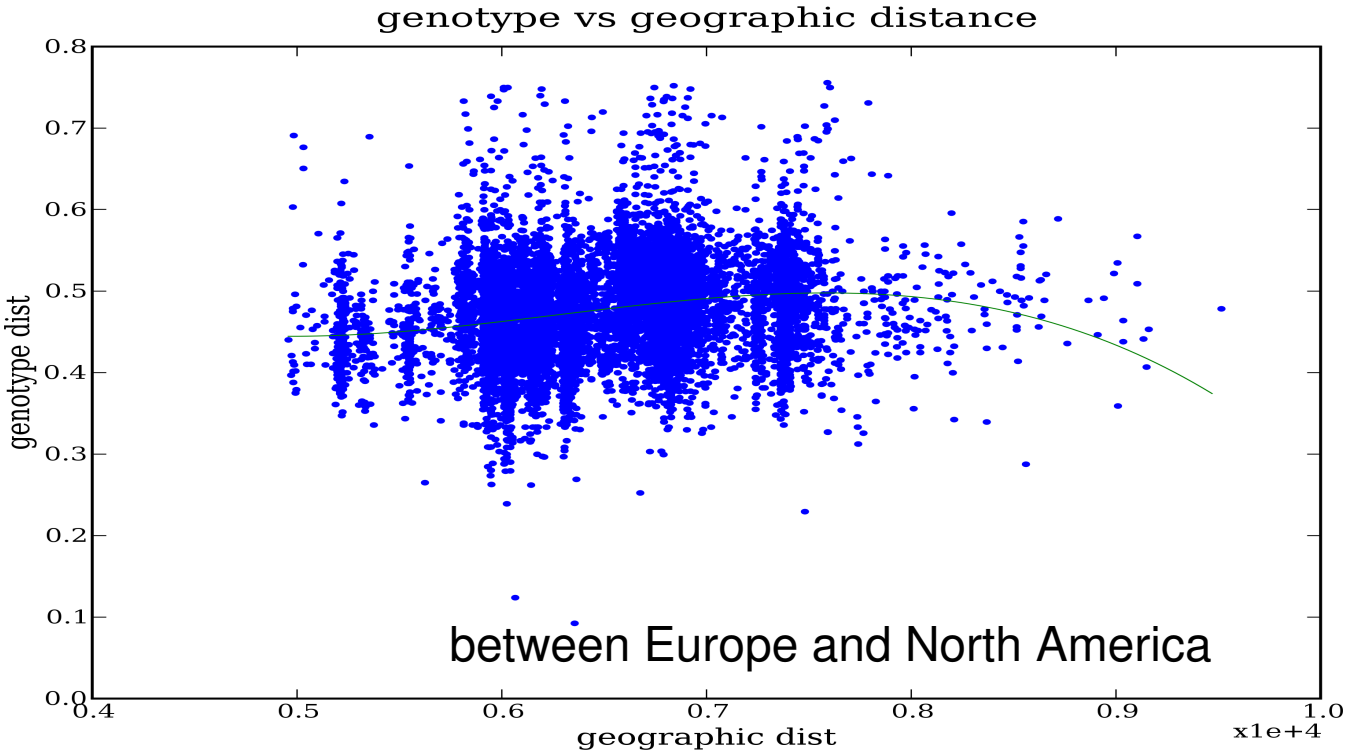

LD decay
North America
r^2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Distance
x1e+7

linear relationship between genetic and geographic distance?



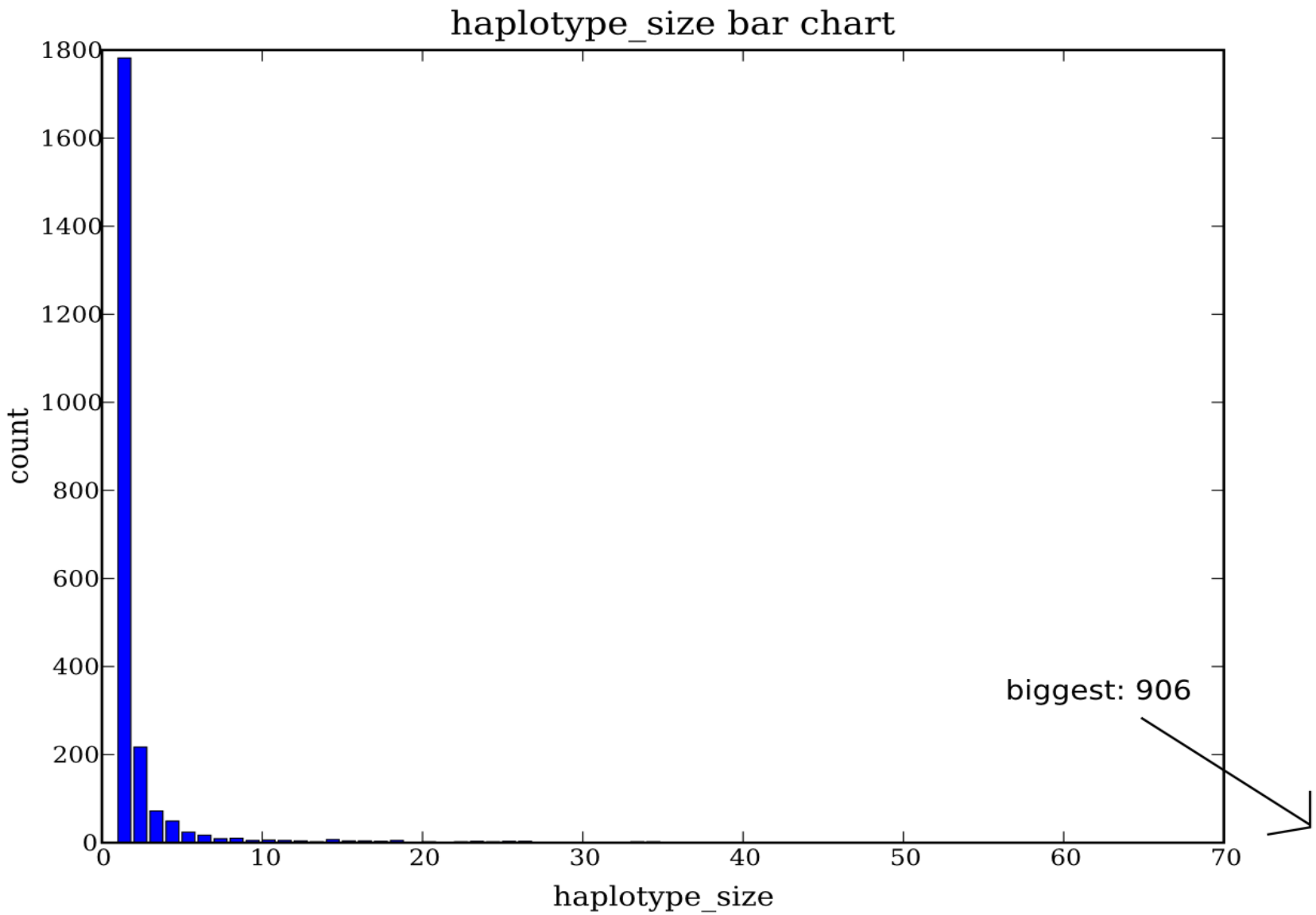

genotype vs geographic distance
genotype dist
geographic dist
Europe


genotype vs geographic distance
genotype dist
geographic dist
North America

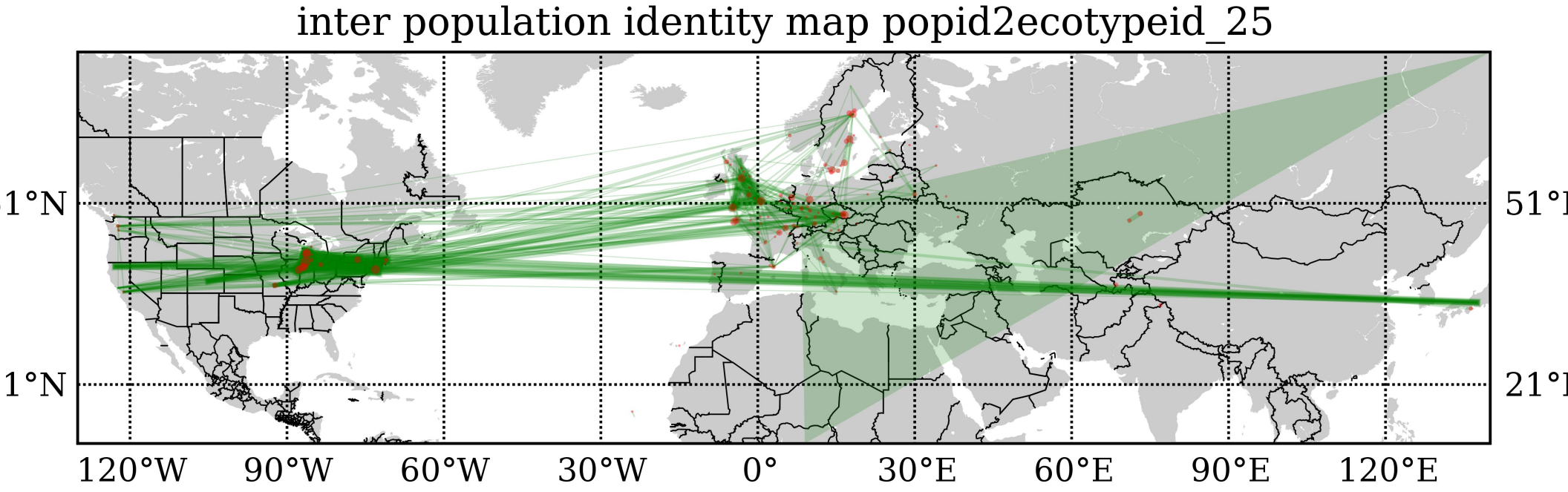

genotype vs geographic distance
genotype dist
geographic dist
x1e+4
between Europe and North America

one group of identical strains => unique haplotype

group size => haplotype size



how populations are connected by identity

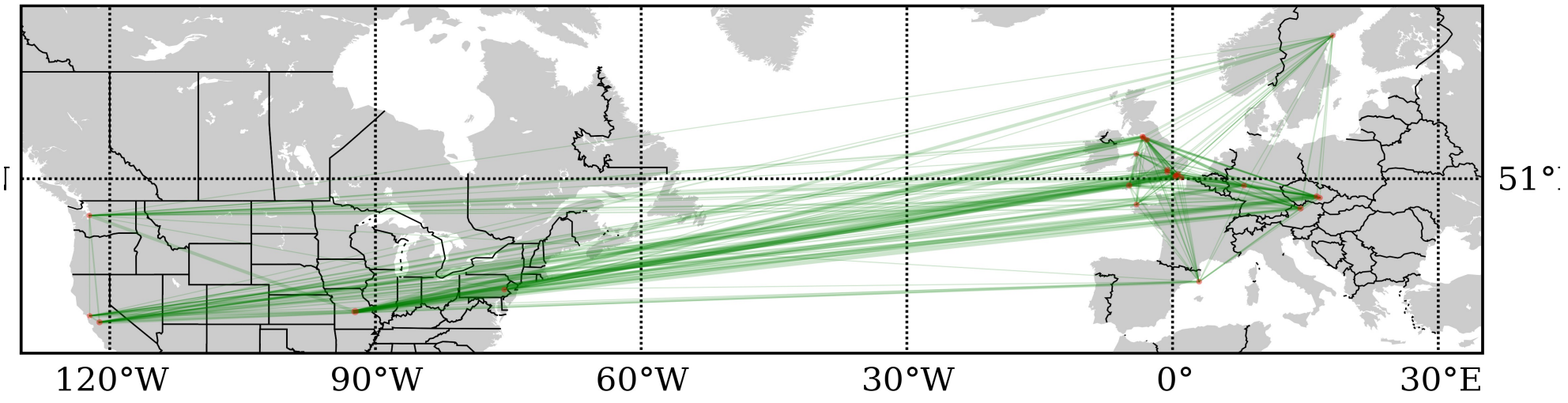


contamination by famous lab strains

Landsberg

id	name	nativename	stkparent	lat	lon	collector	country	site
7099	CS28209	Di-2	CS6682	47.0	5.0	Albert Kranz	FRA	Di
7096	CS28205	Di-G	CS910	47.3239	5.04278	Maarten Koornneef	FRA	Di
446	EM-048	EM-048		51.3	0.5	Eric Holub	UK	East Malling Research
7213	CS28445	Ler-0	CS20	47.984	10.8719	Maarten Koornneef	GER	Ler
7212	CS28446	Ler-0	CS24238	47.984	10.8719	Jorge Casal	GER	Ler
7215	CS28447	Ler-0	CS24596	47.984	10.8719	Maarten Koornneef	GER	Ler
6932	CS28449	Ler-1	CS22618	47.984	10.8719	Eric Holub	GER	Ler
7214	CS28450	Ler-2	CS8581	47.984	10.8719	Maarten Koornneef	GER	Ler
5534	UKNW06-352	UKNW06-352		54.6	-3.1	Eric Holub	UK	Keswick
5205	UKSE06-427	UKSE06-427		51.3	0.4	Eric Holub	UK	Hadlow
5254	UKSE06-501	UKSE06-501		51.2	0.8	Eric Holub	UK	Pluckley
4781	UKSW06-181	UKSW06-181		50.4	-4.9	Eric Holub	UK	St Columb
6434	ZdrI 2-9	ZdrI 2-9		49.3853	16.2544	Jirina Relichov	CZE	ZdrI 2

Columbia



Real Cross-Continent?

Table 249: haplotype clique 315 has 4 ecotypes from Asia and Europe. check Figure **??**.

id	name	nativename	stkparent	lat	lon	collector	country	site
7093	CS28194	Condara	CS6175	38.48	68.49	Igor Vizir	TJK	Kondara
9059	Hg-2	Hg-2		62.79	17.9	Magnus Nordborg	SWE	Hg
6929	CS28418	Kondara	CS22651	38.48	68.49	Igor Vizir	TJK	Kondara
7200	CS28417	Kondara	CS916	38.48	68.49	Igor Vizir	TJK	Kondara

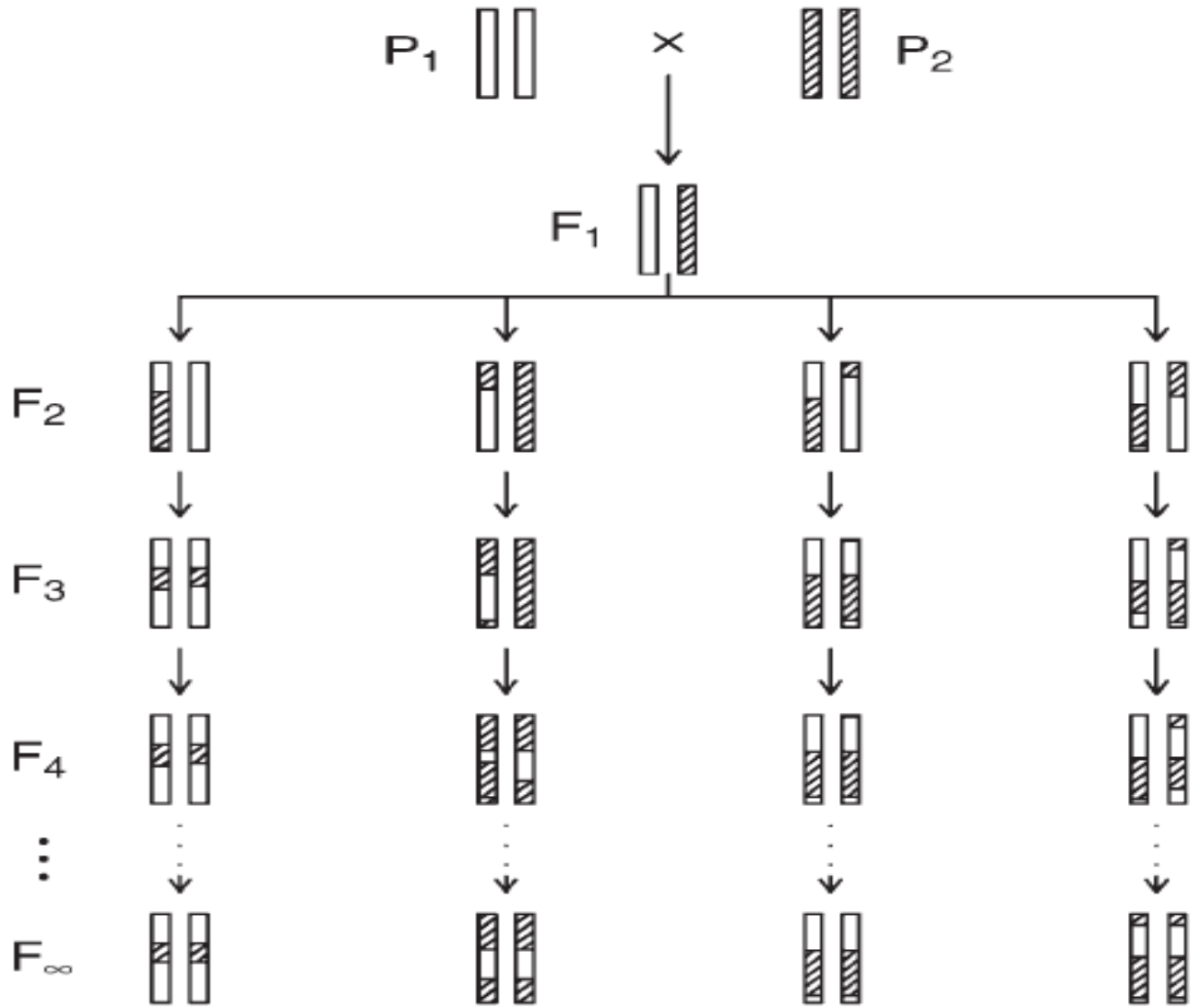
Table 256: haplotype clique 282 has 4 ecotypes from Europe and North America. check Figure **??**.

id	name	nativename	stkparent	lat	lon	collector	country	site
6708	CS28068	BG-1	CS22341	47.6479	-122.305	Juliette Winterer	USA	BG
6709	CS28069	BG-2	CS22342	47.6479	-122.305	Juliette Winterer	USA	BG
6711	CS28071	BG-4	CS22344	47.6479	-122.305	Juliette Winterer	USA	BG
6908	CS28143	CIBC-5	CS22602	51.4083	-0.6383	Mick Crawley	UK	CIBC

Table 257: haplotype clique 331 has 4 ecotypes from Europe and North America. check Figure **??**.

id	name	nativename	stkparent	lat	lon	collector	country	site
7524	CS28689	Rmx-A02	CS22568	42.036	-86.511	Joy Bergelson	USA	RMX
5764	UKID59	UKID59		54.7	-2.8	Eric Holub	UK	Penrith
5768	UKID63	UKID63		54.1	-1.5	Eric Holub	UK	Ripon
5778	UKID73	UKID73		52.2	1.5	Eric Holub	UK	Snape Malting

#selfing generations, RIL

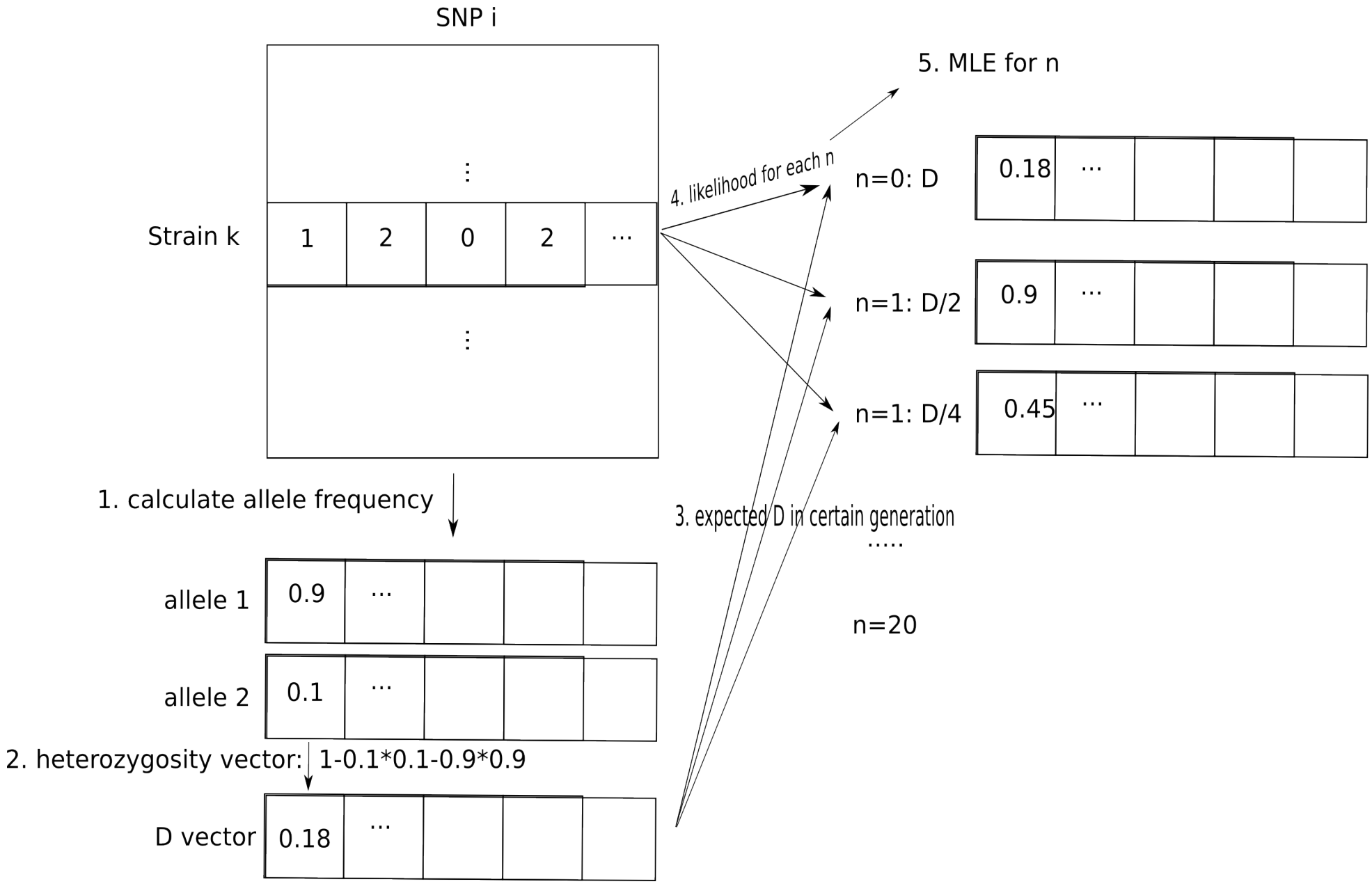


model to estimate #selfing generations

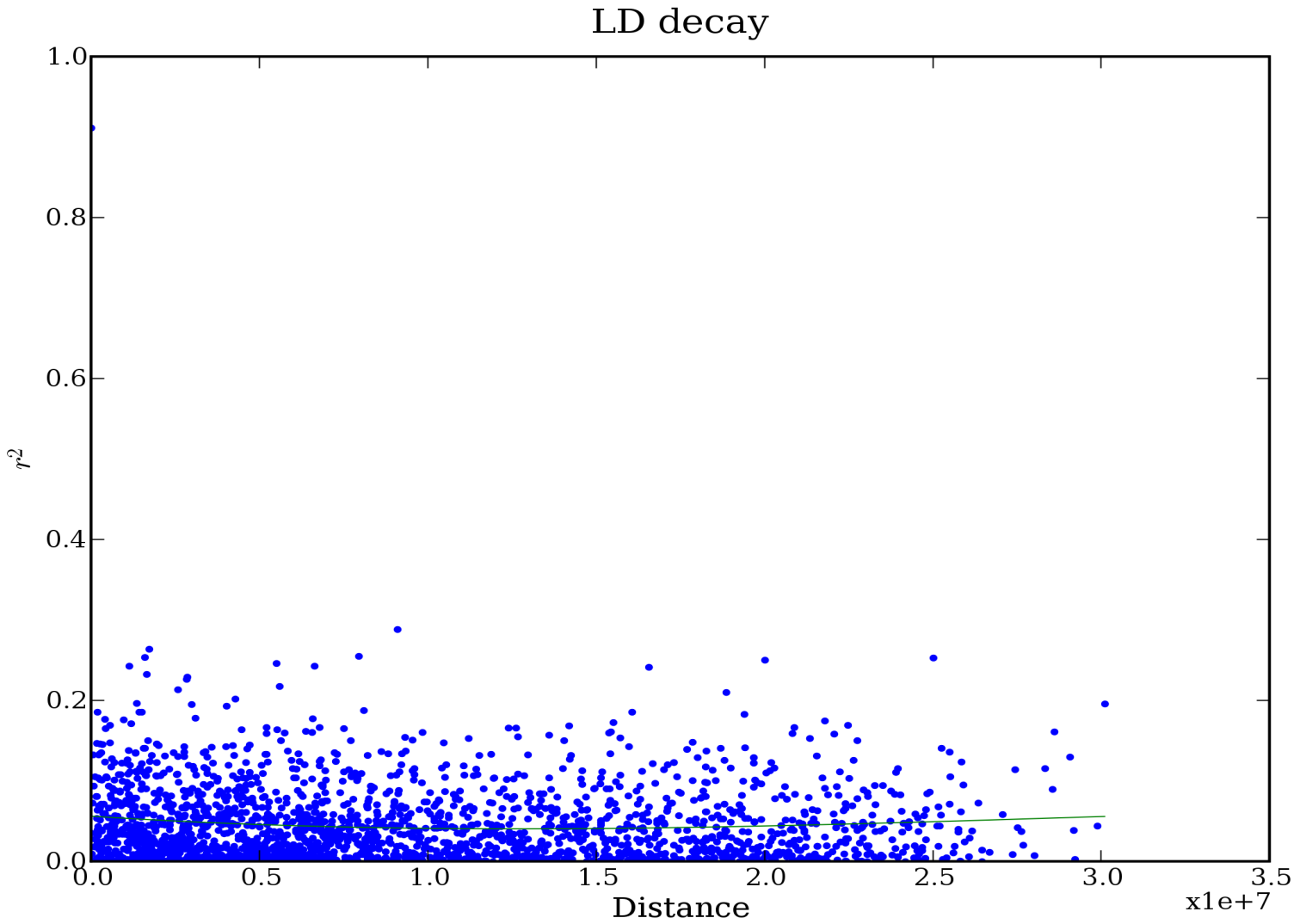
$$D_i = 1 - \sum_{j=1}^{k} p_{ij}^2$$

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.

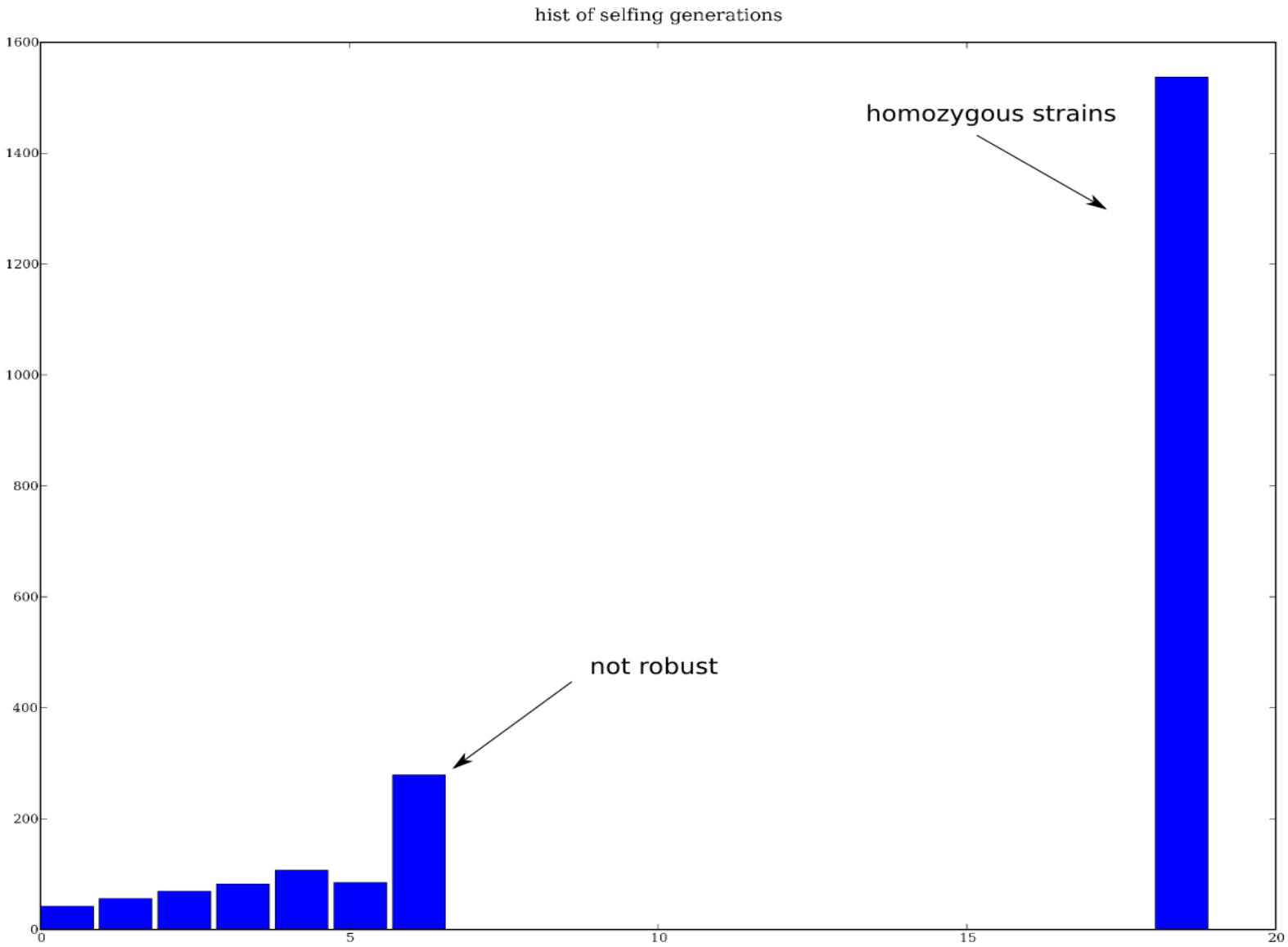
$$likelihood(S_n) = \prod_{i=1}^{L} (\frac{D_i}{2^n})^{a_i} (1 - \frac{D_i}{2^n})^{1-a_i}$$


SNP i
Strain k
1
2
0
2
...
1. calculate allele frequency
allele 1
0.9
...
allele 2
0.1
...
2. heterozygosity vector: 1-0.1*0.1-0.9*0.9
D vector
0.18
...
3. expected D in certain generation
.....
n=20
4. likelihood for each n
n=0: D
n=1: D/2
n=1: D/4
5. MLE for n
0.18
...
0.9
...
0.45
...

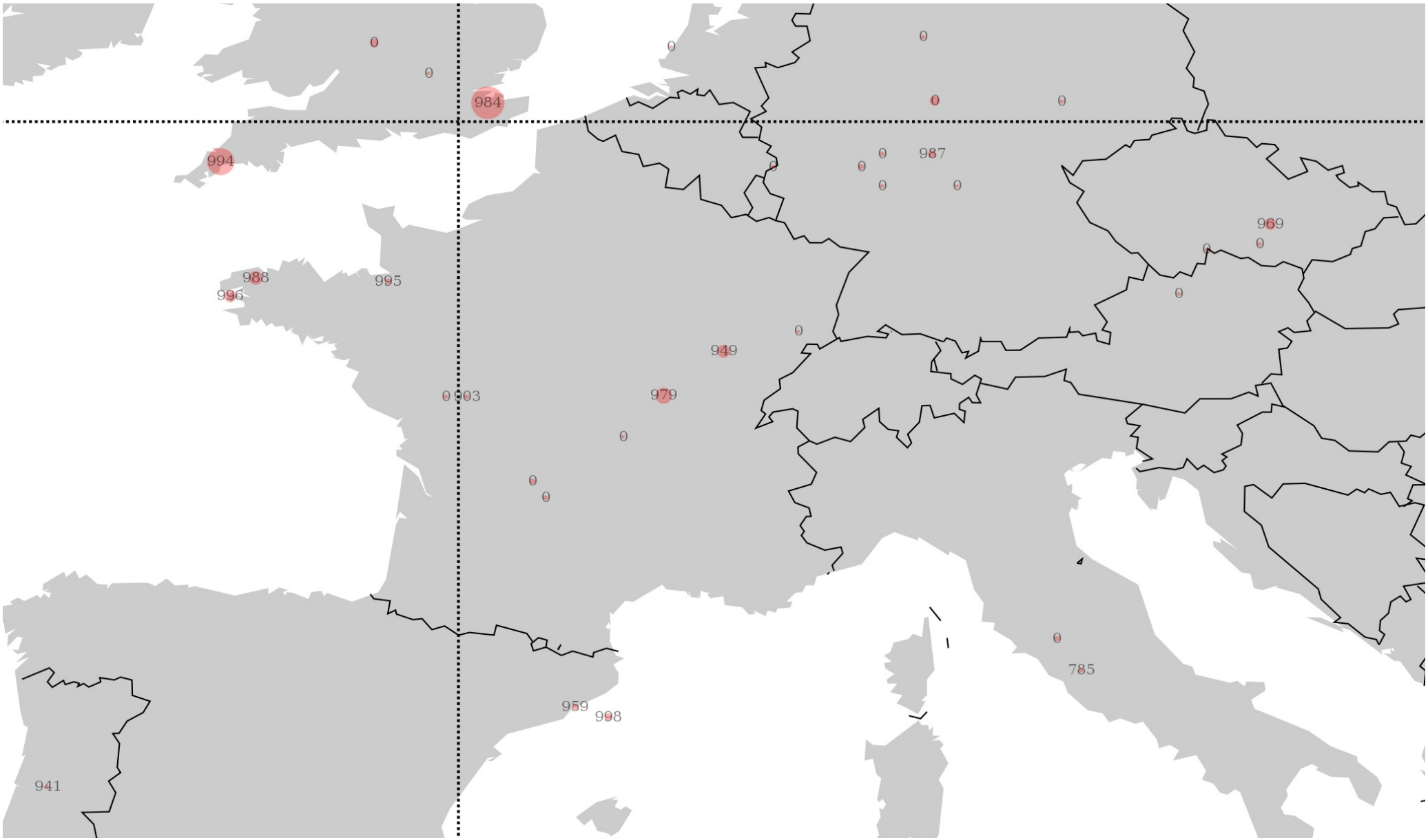
loci independence could be assumed



histogram of #selfing generations



Selfing Rate



ToDo

- Recombinant Inbred Line
- Parent-child pairs, shared block
- Sib-ship
- ...