Exploratory Allele Specific Expression Analysis on X chromosome

Vasyl Zhabotynsky

Introduction

- We will use the Pickerel at al. dataset of RNA-seq data for Yoruba HapMap cell lines sequenced using the Illumina GA2 platform to search for a possible highly expressed genes in chromosome X.
- Potential interest will be to find expressed genes and to check whether we can see evidence that X chromosome inactivation takes place.

Setup

- Original dataset includes 65 subjects, out of which 38 are females
- (select females to look into X chromosome)
- Aggregate allele specific counts by gene
- For each individual that has at least 10 counts on both alleles perform Fisher's exact test.

Fisher's exact test

Suppose we have an individual (NA18499) we have such counts for her:

Gene	Allele A	Allele B
ENSG0000014715	8	49
ENSG00000196924	77	14
…	…	…
ENSG00000182872	0	0
Totals:	354	392

Note, many of genes for each individual still are (0,0), since we are choosing such genes where at least one individual have more than 10 counts.
After all, for each gene many individuals are homozygotic.

Fisher's exact test 2

	Allele A	Allele B	Totals
ENSG0000014715	a=8	b=49	a+b=57
Totals- gene count	c=354-8=346	d=392-49=343	c+d=689
Totals	a+c=354	b+d=392	n=354+392=746

Where n=a+b+c+d.
Note, that in such case probability of getting such values under independence assumption is:

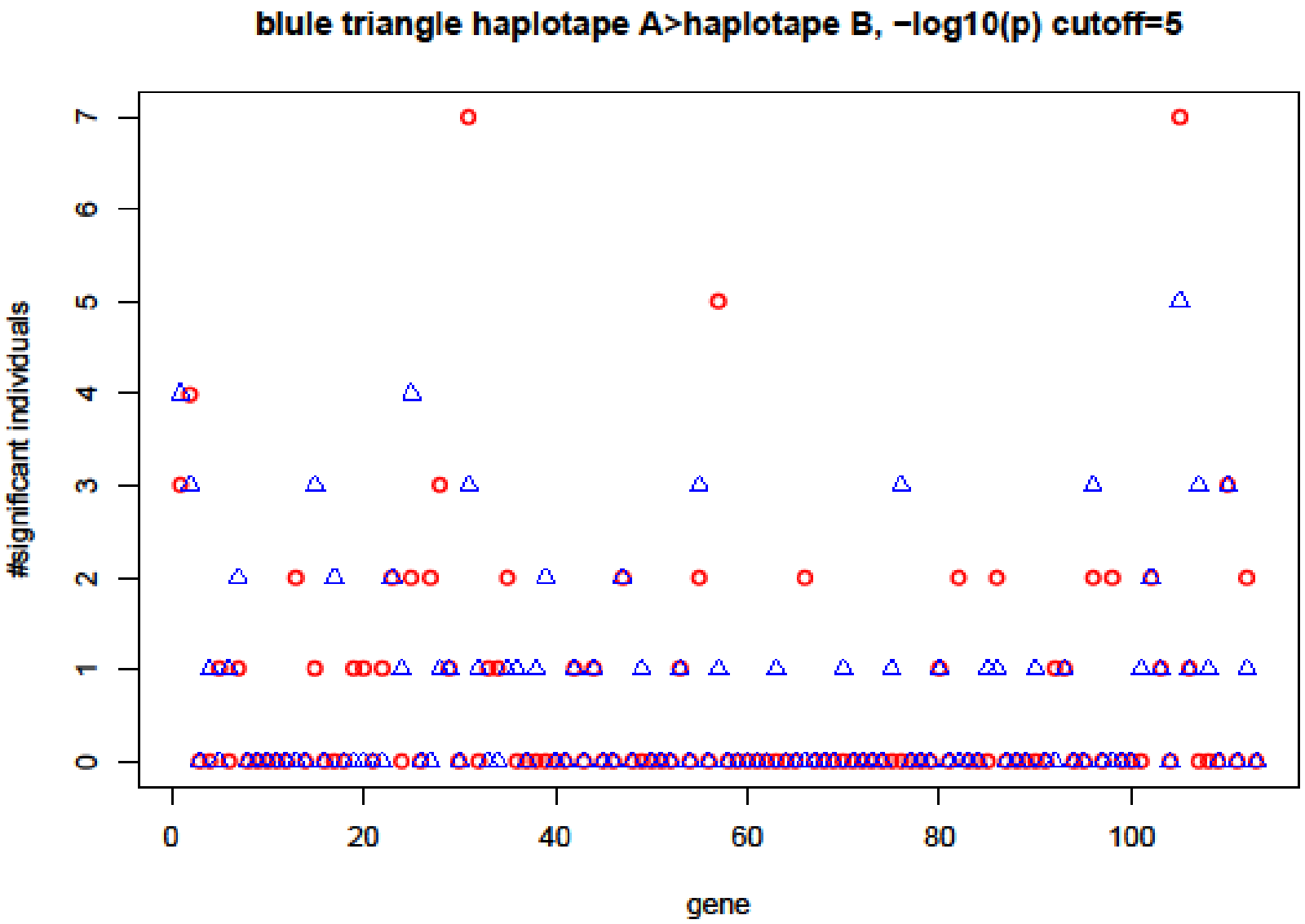
$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

We can calculate as extreme or more extreme values of the table to calculate a p-value.
Well, we don't really need to write our own function – there is an appropriate R function fisher.test.
Note, if you wish to write your own function in such case I would strongly recommend to use logarithms, which is very handy here since we move to sums of logs and avoid big-big overflow and rounding problems.

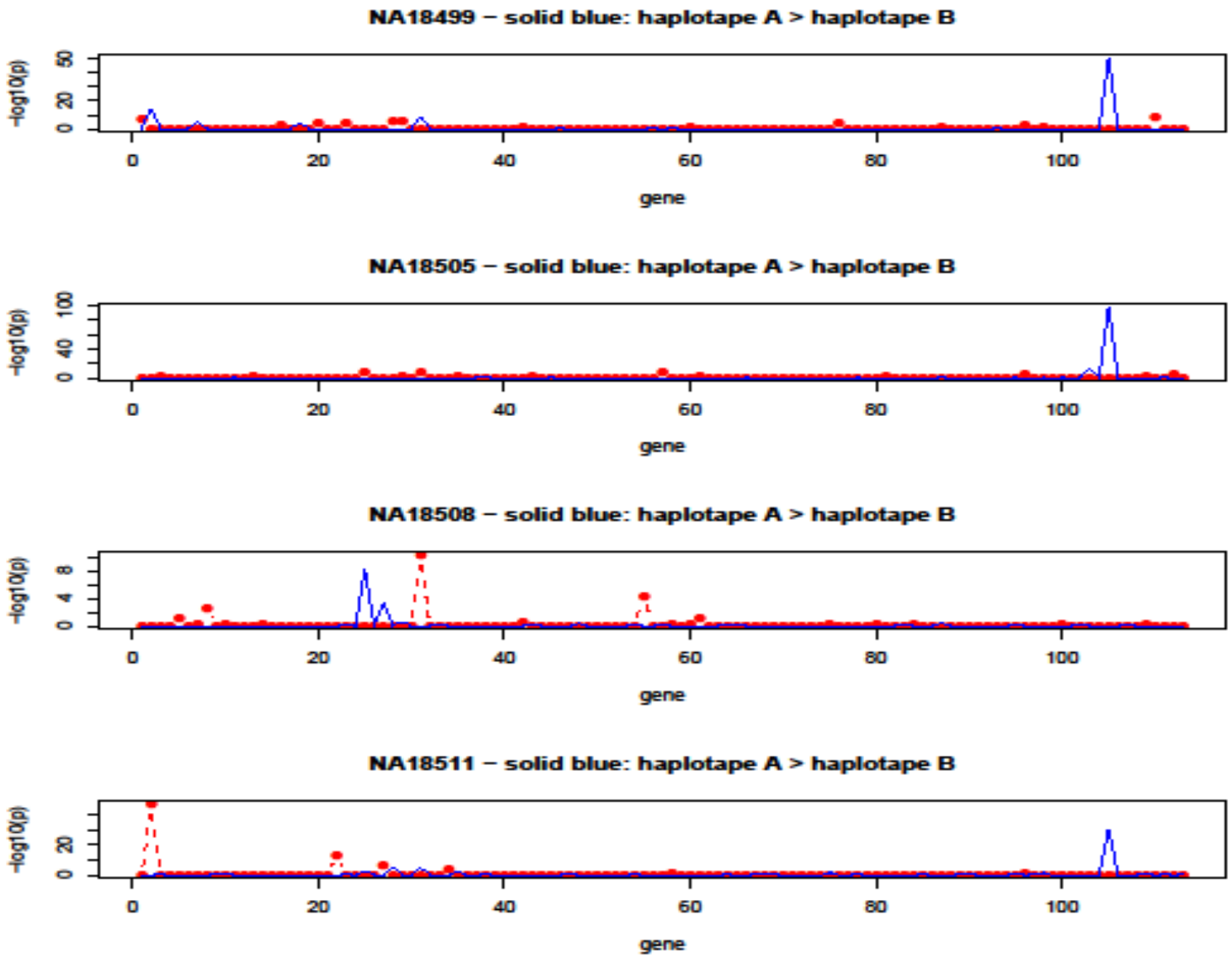
Results 1

- Note, first that we may run into a situation when allele A counts are bigger than allele B counts and to the opposite situation. We will distinguish such situations (and study them a bit more closely).
- Having cutoffs as stated couple slides before (at least 10 counts for at least one individual) we find that there are 113 such genes.

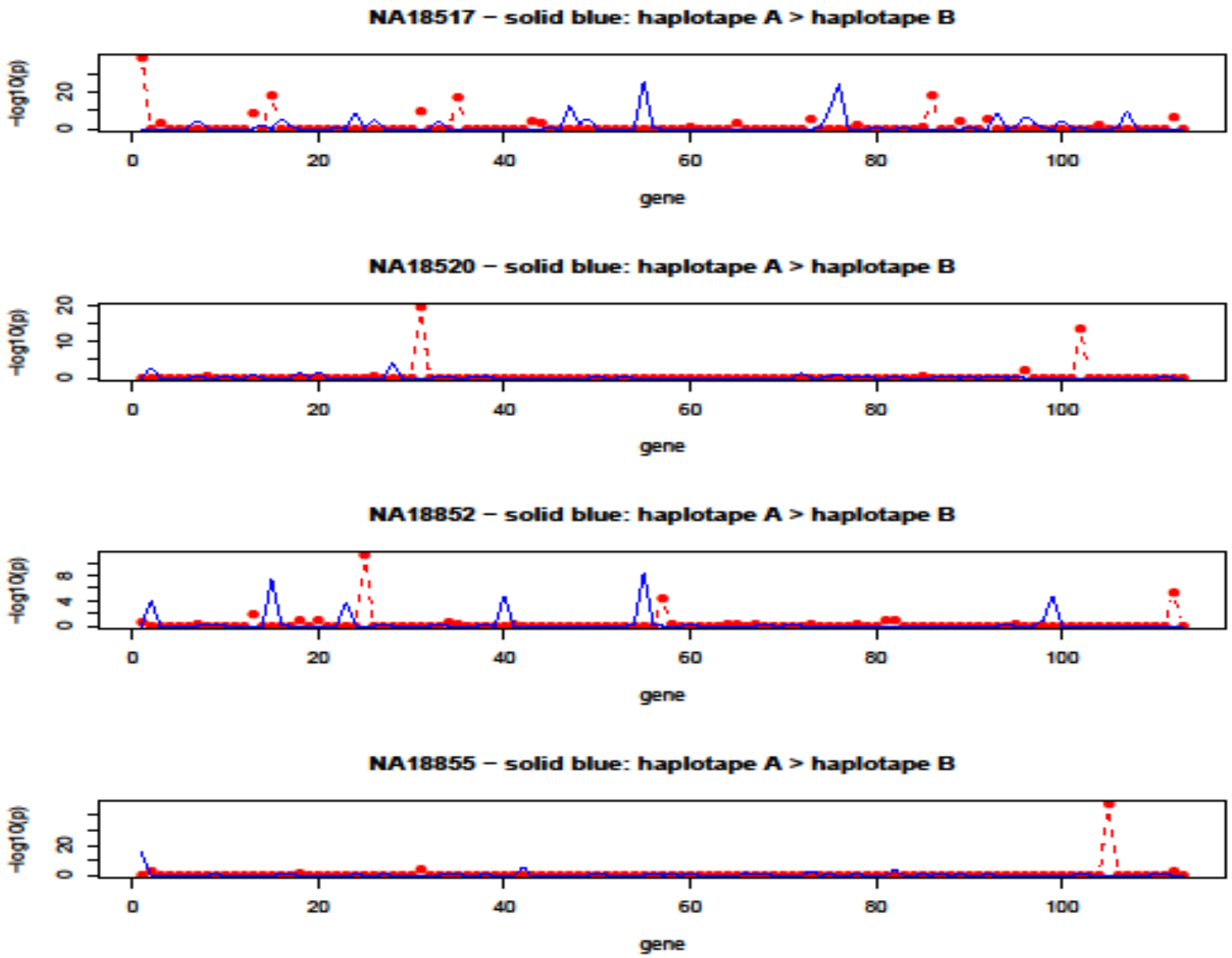
Results 2


blule triangle haplotape A>haplotape B, −log10(p) cutoff=5
#significant individuals
0
1
2
3
4
5
6
7
0
20
40
60
80
100
gene

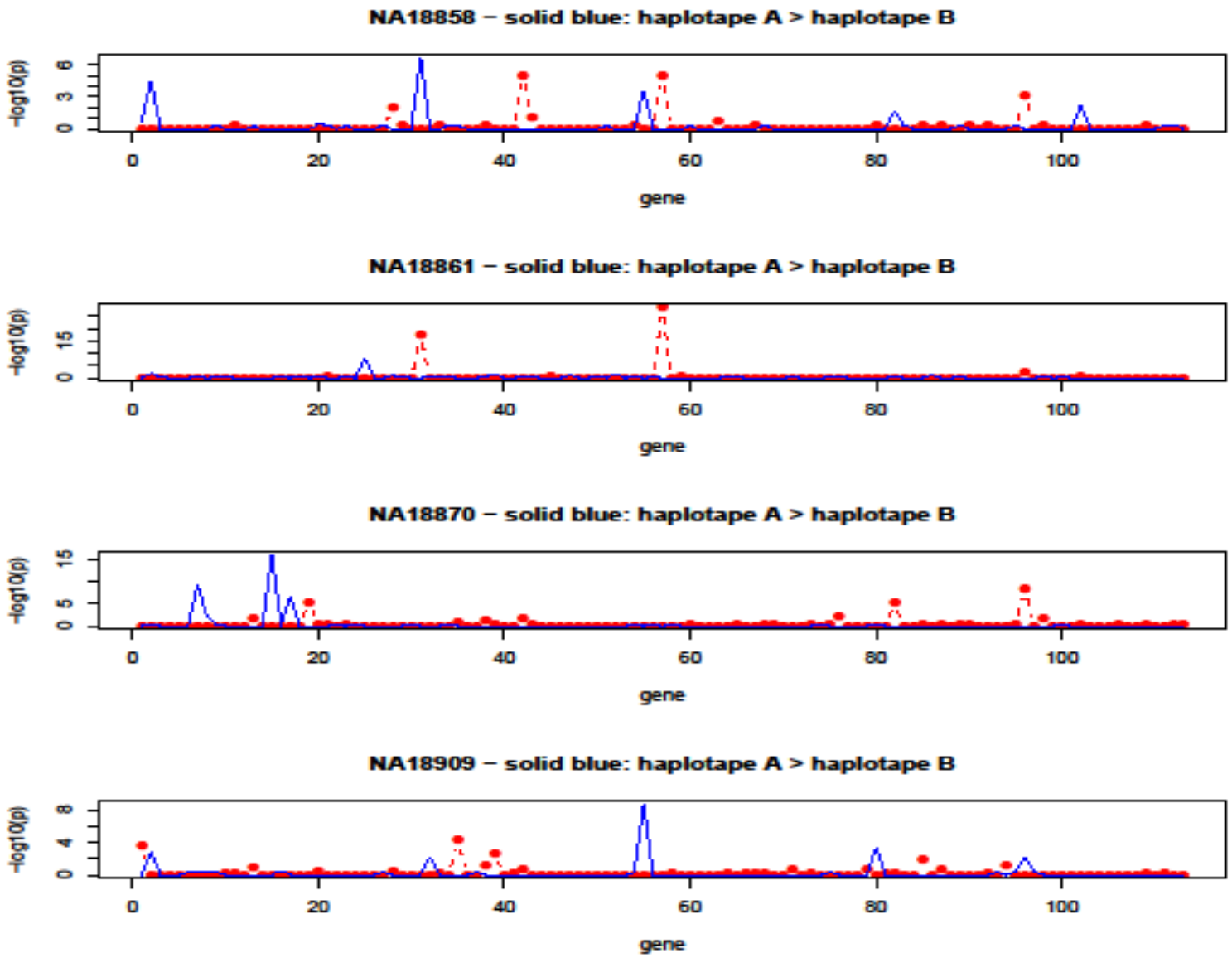
Results 2-1


NA18499 – solid blue: haplotape A > haplotape B
-log10(p)
0
20
50
0
20
40
60
80
100
gene
NA18505 – solid blue: haplotape A > haplotape B
-log10(p)
0
40
100
0
20
40
60
80
100
gene
NA18508 – solid blue: haplotape A > haplotape B
-log10(p)
0
4
8
0
20
40
60
80
100
gene
NA18511 – solid blue: haplotape A > haplotape B
-log10(p)
0
20
0
20
40
60
80
100
gene

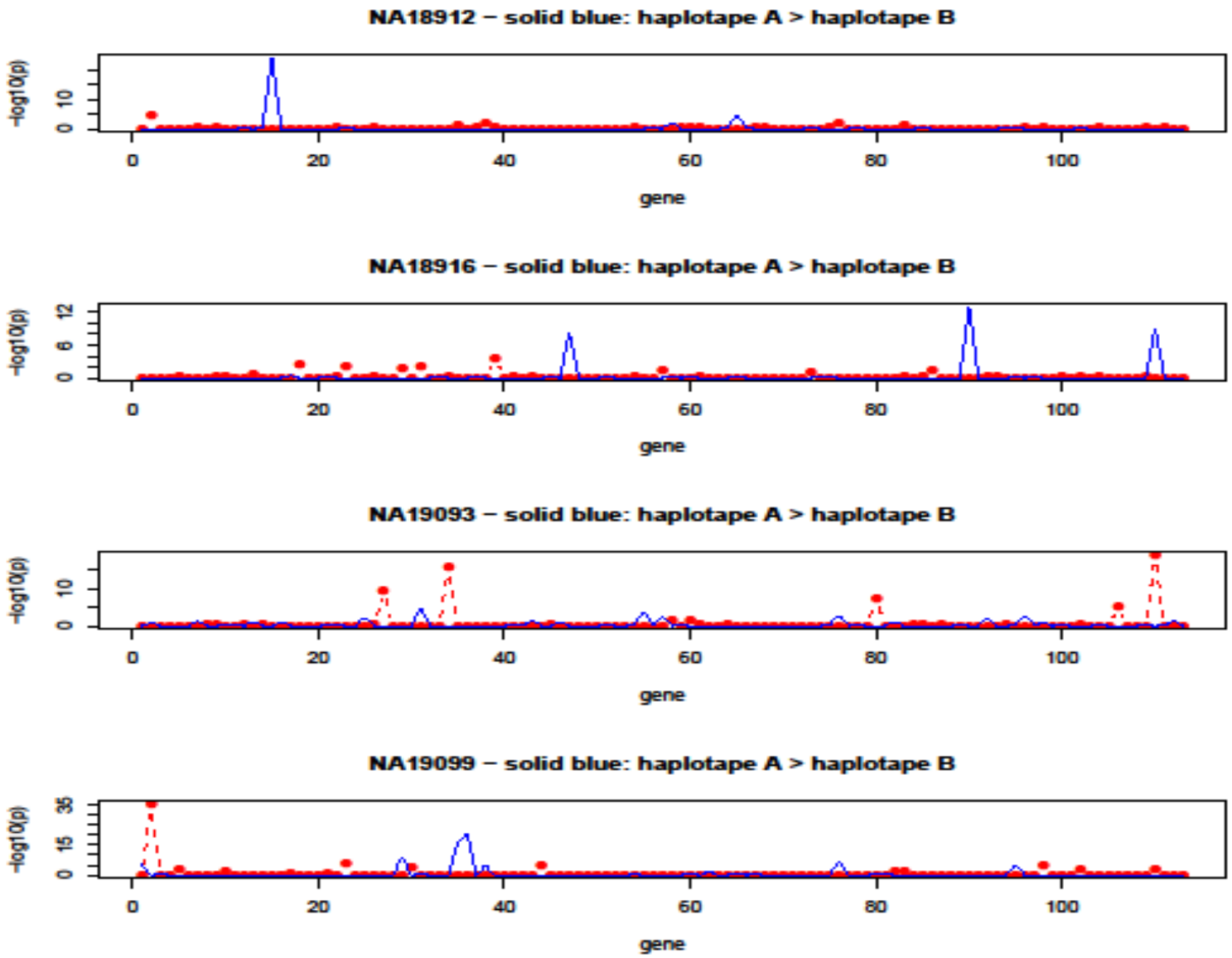
Results 2-2


NA18517 – solid blue: haplotape A > haplotape B
-log10(p)
0
20
0
20
40
60
80
100
gene
NA18520 – solid blue: haplotape A > haplotape B
-log10(p)
0
10
20
0
20
40
60
80
100
gene
NA18852 – solid blue: haplotape A > haplotape B
-log10(p)
0
4
8
0
20
40
60
80
100
gene
NA18855 – solid blue: haplotape A > haplotape B
-log10(p)
0
20
0
20
40
60
80
100
gene

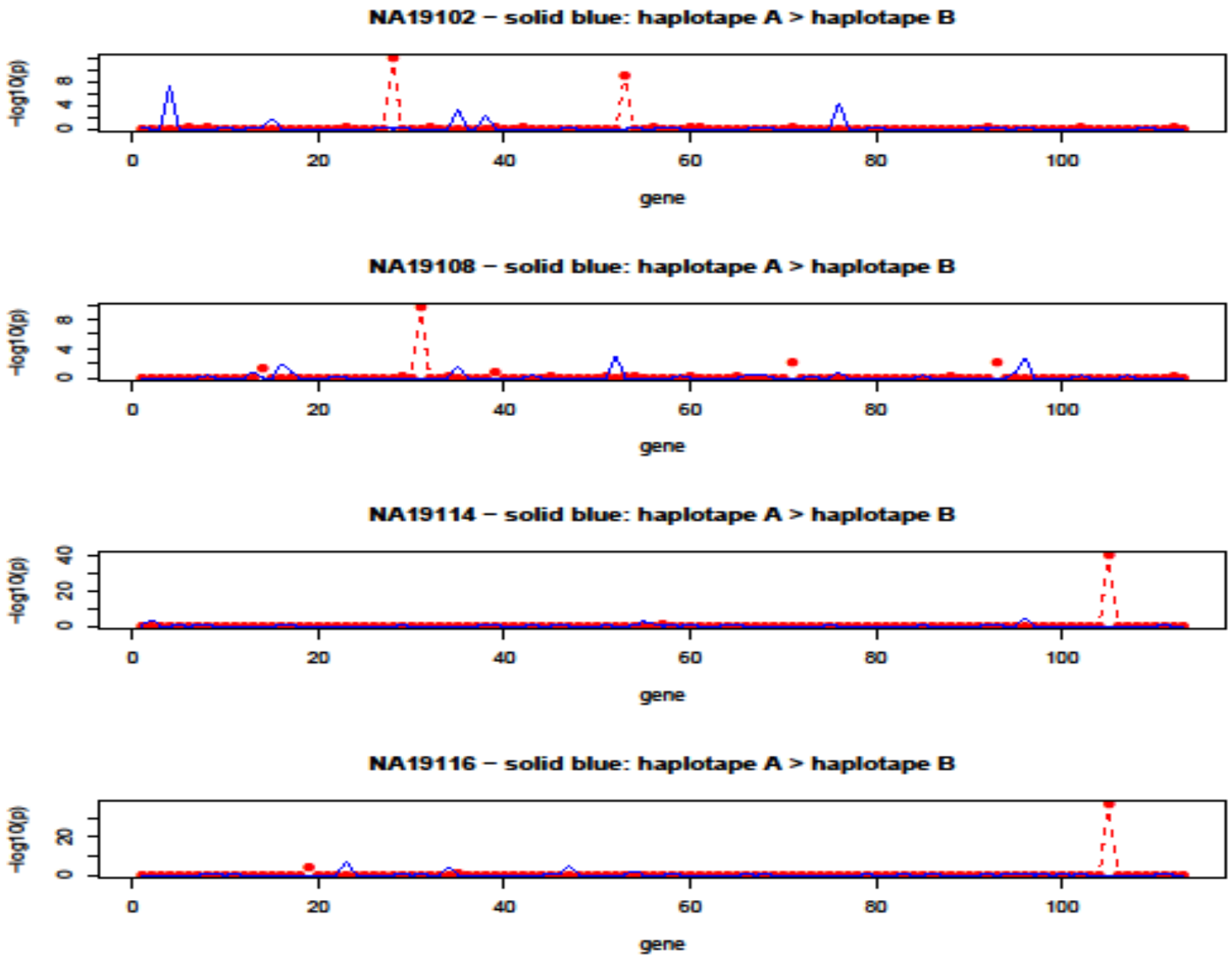
Results 2-3


NA18858 – solid blue: haplotape A > haplotape B
-log10(p)
0
3
6
0
20
40
60
80
100
gene
NA18861 – solid blue: haplotape A > haplotape B
-log10(p)
0
15
0
20
40
60
80
100
gene
NA18870 – solid blue: haplotape A > haplotape B
-log10(p)
0
5
15
0
20
40
60
80
100
gene
NA18909 – solid blue: haplotape A > haplotape B
-log10(p)
0
4
8
0
20
40
60
80
100
gene

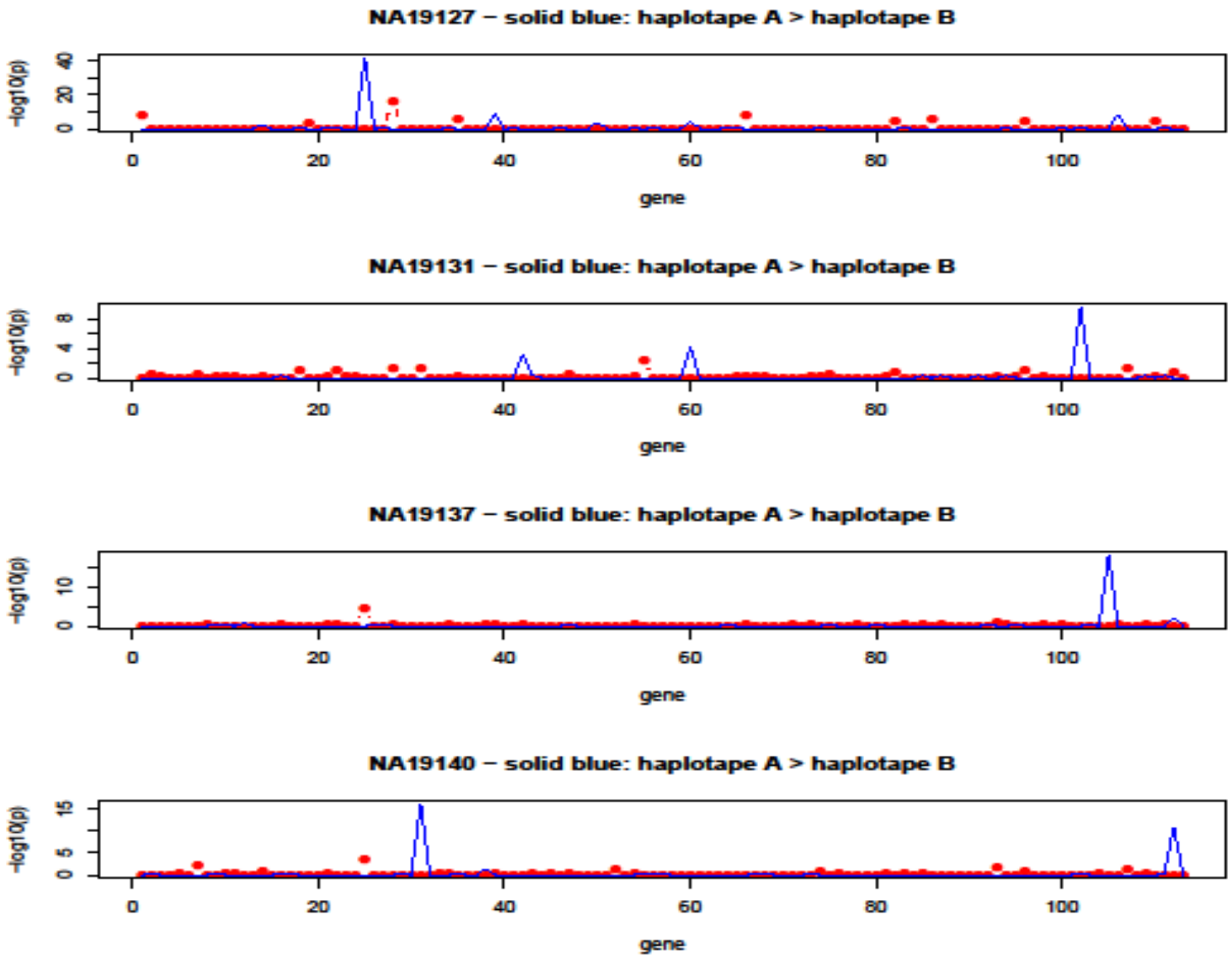
Results 2-4


NA18912 – solid blue: haplotape A > haplotape B
-log10(p)
0
10
0
20
40
60
80
100
gene
NA18916 – solid blue: haplotape A > haplotape B
-log10(p)
0
6
12
0
20
40
60
80
100
gene
NA19093 – solid blue: haplotape A > haplotape B
-log10(p)
0
10
0
20
40
60
80
100
gene
NA19099 – solid blue: haplotape A > haplotape B
-log10(p)
0
15
35
0
20
40
60
80
100
gene

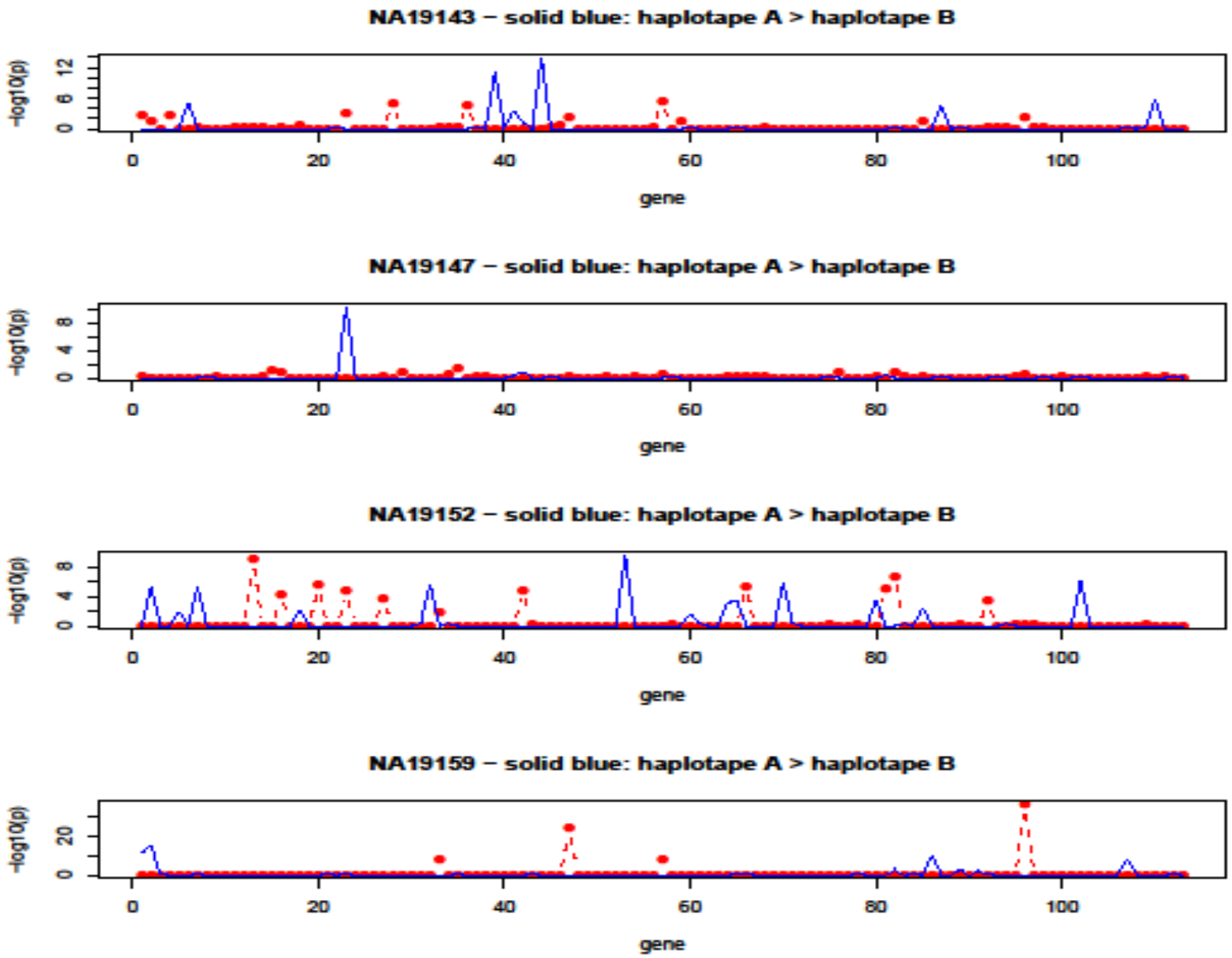
Results 2-5


NA19102 – solid blue: haplotape A > haplotape B
-log10(p)
0
4
8
0
20
40
60
80
100
gene
NA19108 – solid blue: haplotape A > haplotape B
-log10(p)
0
4
8
0
20
40
60
80
100
gene
NA19114 – solid blue: haplotape A > haplotape B
-log10(p)
0
20
40
0
20
40
60
80
100
gene
NA19116 – solid blue: haplotape A > haplotape B
-log10(p)
0
20
0
20
40
60
80
100
gene

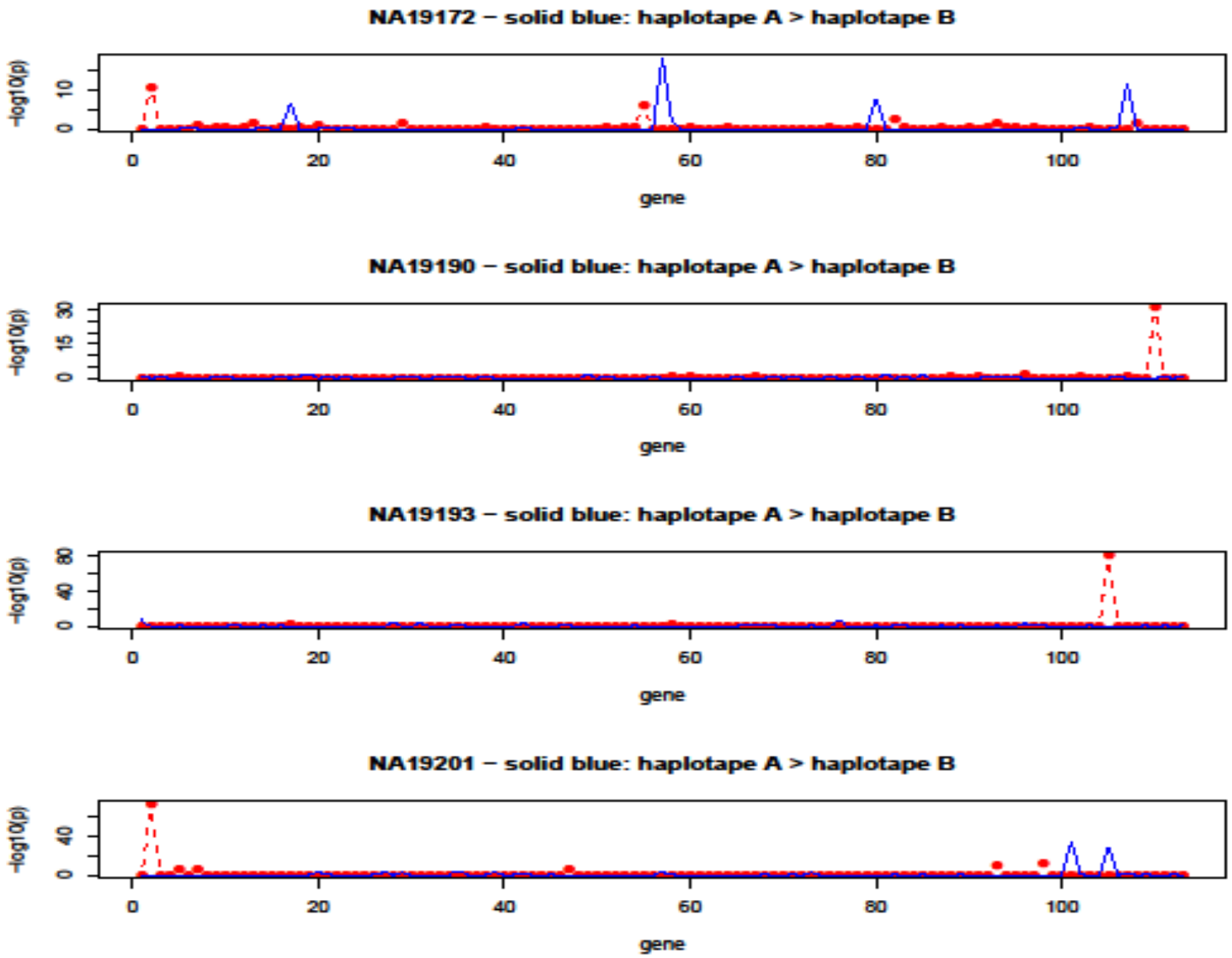
Results 2-6


NA19127 – solid blue: haplotape A > haplotape B
-log10(p)
0
20
40
0
20
40
60
80
100
gene
NA19131 – solid blue: haplotape A > haplotape B
-log10(p)
0
4
8
0
20
40
60
80
100
gene
NA19137 – solid blue: haplotape A > haplotape B
-log10(p)
0
10
0
20
40
60
80
100
gene
NA19140 – solid blue: haplotape A > haplotape B
-log10(p)
0
5
15
0
20
40
60
80
100
gene

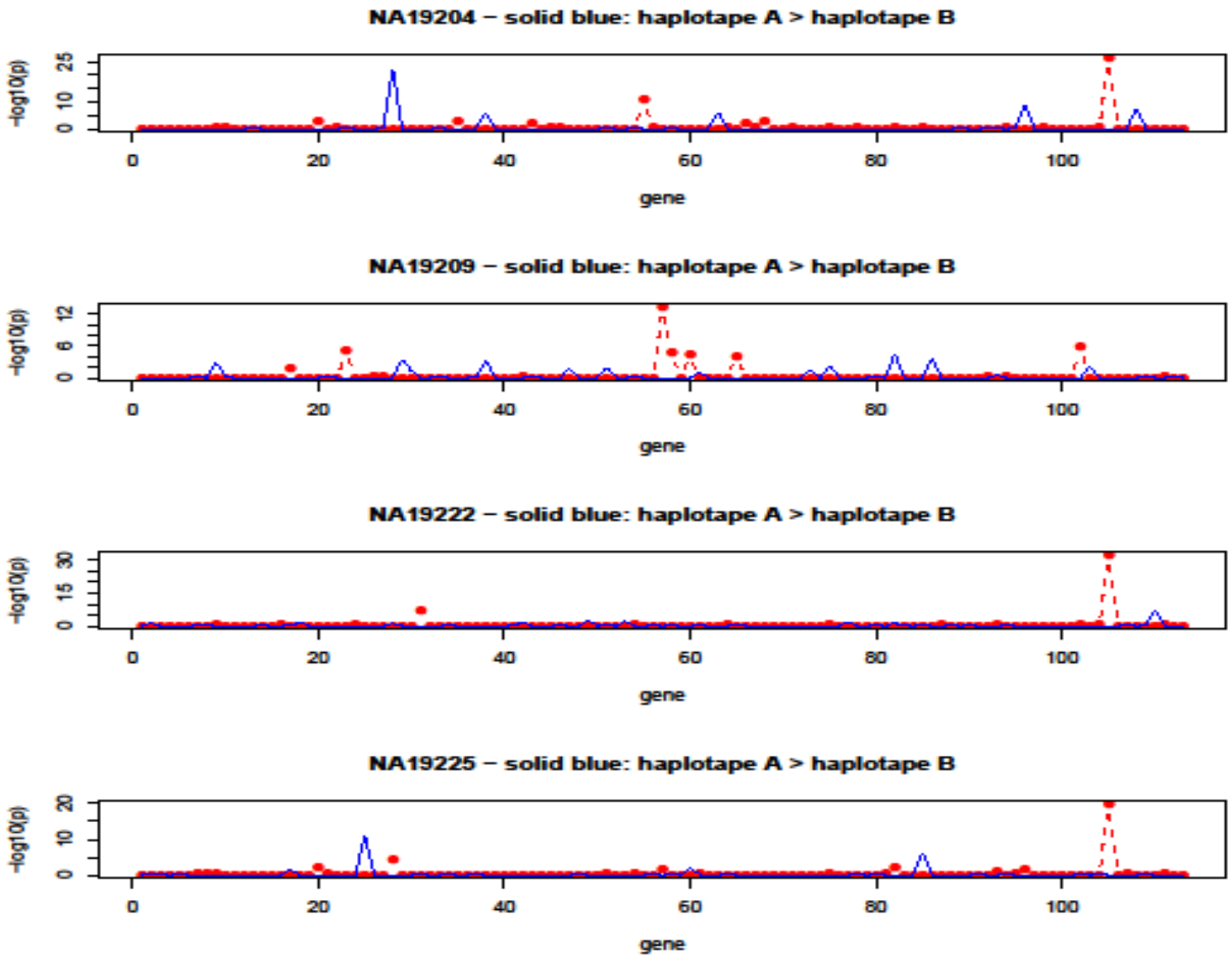
Results 2-7


NA19143 – solid blue: haplotape A > haplotape B
-log10(p)
0
6
12
0
20
40
60
80
100
gene
NA19147 – solid blue: haplotape A > haplotape B
-log10(p)
0
4
8
0
20
40
60
80
100
gene
NA19152 – solid blue: haplotape A > haplotape B
-log10(p)
0
4
8
0
20
40
60
80
100
gene
NA19159 – solid blue: haplotape A > haplotape B
-log10(p)
0
20
0
20
40
60
80
100
gene

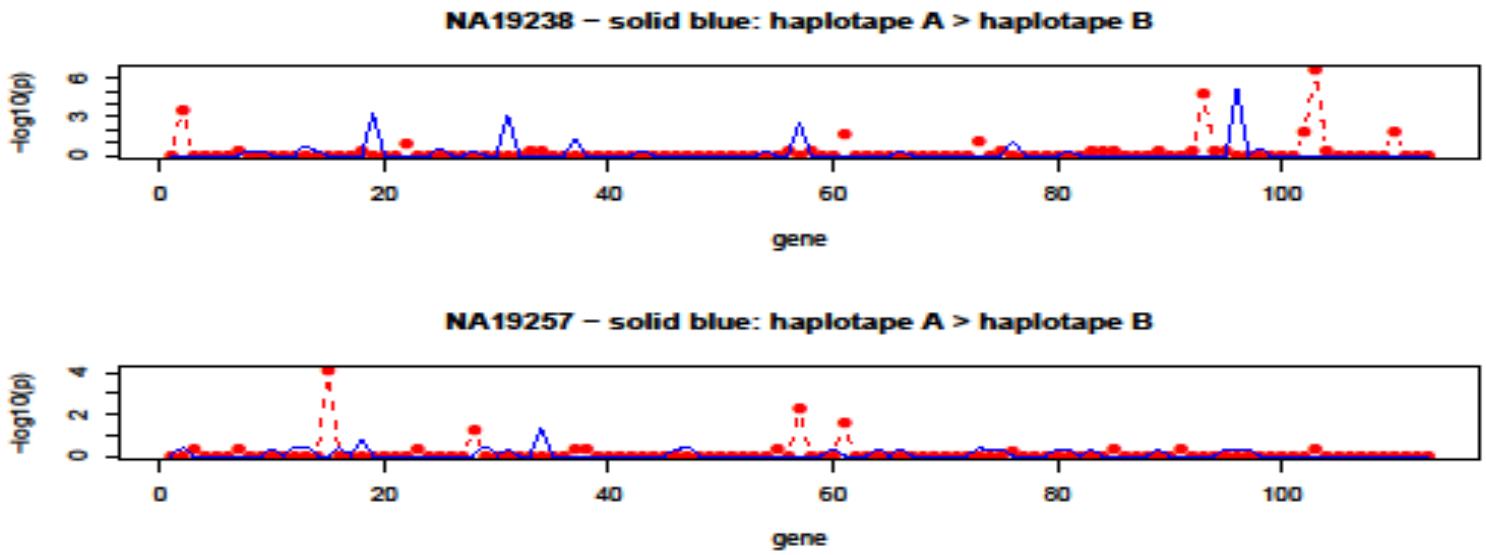
Results 2-8


NA19172 – solid blue: haplotape A > haplotape B
-log10(p)
0
10
0
20
40
60
80
100
gene
NA19190 – solid blue: haplotape A > haplotape B
-log10(p)
0
15
30
0
20
40
60
80
100
gene
NA19193 – solid blue: haplotape A > haplotape B
-log10(p)
0
40
80
0
20
40
60
80
100
gene
NA19201 – solid blue: haplotape A > haplotape B
-log10(p)
0
40
0
20
40
60
80
100
gene

Results 2-9

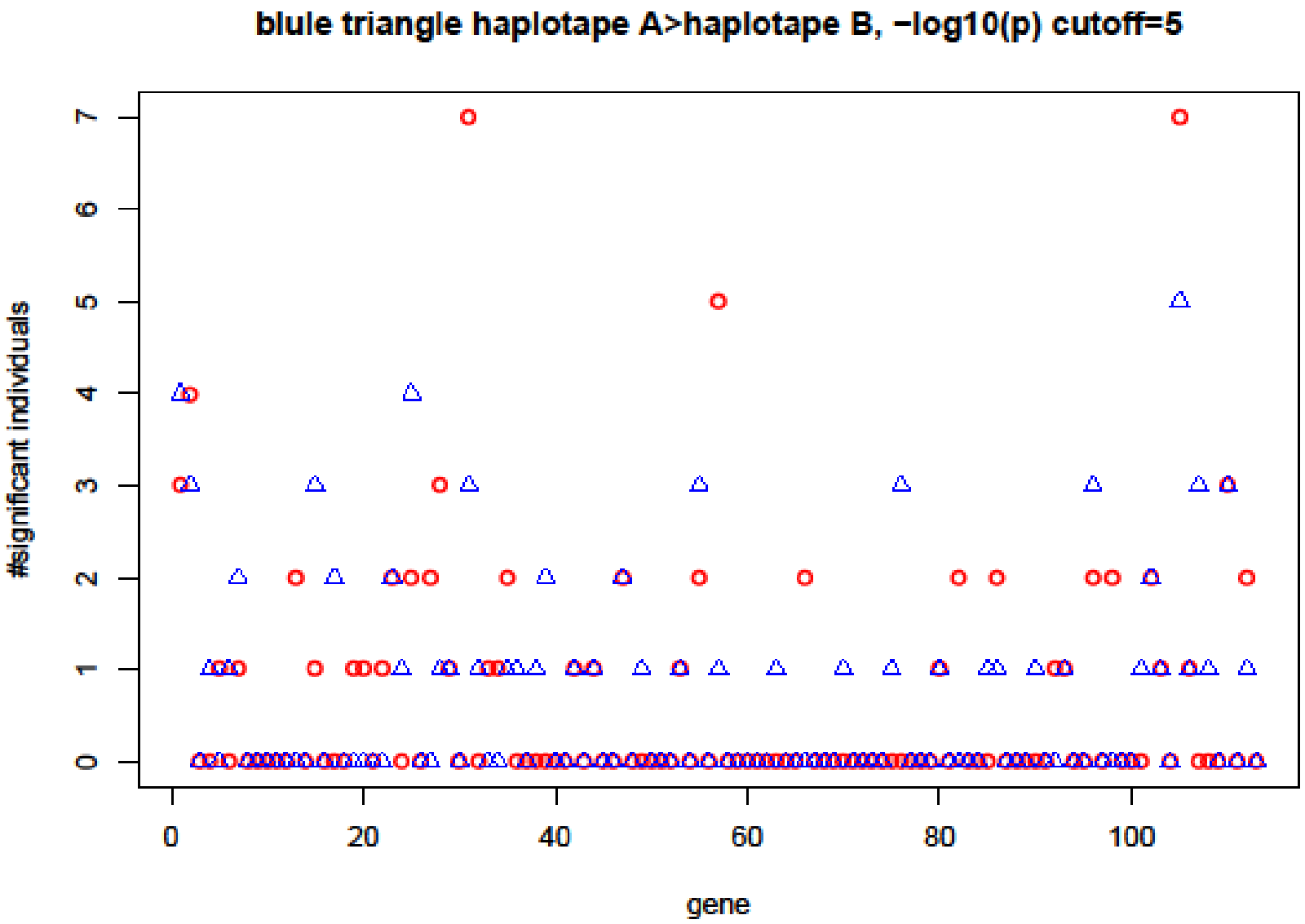

NA19204 – solid blue: haplotape A > haplotape B
-log10(p)
0
10
25
0
20
40
60
80
100
gene
NA19209 – solid blue: haplotape A > haplotape B
-log10(p)
0
6
12
0
20
40
60
80
100
gene
NA19222 – solid blue: haplotape A > haplotape B
-log10(p)
0
15
30
0
20
40
60
80
100
gene
NA19225 – solid blue: haplotape A > haplotape B
-log10(p)
0
10
20
0
20
40
60
80
100
gene

Results 2-10



For each individual many genes are not expressed at all (practically all the scores just at 0 reflect this picture)
Also there are several places when we see allele specific expression for quite several individuals (for example position 105)
Also, note that this position has several cases when allele A is more expressed than allele B, and also several cases when the picture is opposite. We can see a summary on the next slide.

Results 3


blule triangle haplotape A>haplotape B, −log10(p) cutoff=5
#significant individuals
0
1
2
3
4
5
6
7
0
20
40
60
80
100
gene

Results 4

- One interesting thing is to see what genes were significant for several individuals.

ENSG00000147155	emopamil binding protein (sterol isomerase)	
	[Source:HGNC Symbol;Acc:3133]	[Type: protein coding Ensembl/Havana merge gene
ENSG00000196924	filamin A, alpha	
	[Source:HGNC Symbol;Acc:3754]	[Type: protein coding Ensembl/Havana merge gene
ENSG00000147010	SH3-domain kinase binding protein 1	
	[Source:HGNC Symbol;Acc:13867]	[Type: protein coding Ensembl/Havana merge gene
ENSG00000102265	TIMP metallopeptidase inhibitor 1	
	[Source:HGNC Symbol;Acc:11820]	[Type: protein coding Ensembl/Havana merge gene
ENSG00000198359	Retired (see below for possible successors) Database: homo_sapiens_core_38_36 pseudogene	
ENSG00000102317	RNA binding motif (RNP1, RRM) protein 3	
	[Source:HGNC Symbol;Acc:9900]	[Type: protein coding Ensembl/Havana merge gene
ENSG00000130826	dyskeratosis congenita 1, dyskerin	
	[Source:HGNC Symbol;Acc:2890]	[Type: protein coding Ensembl/Havana merge gene
ENSG00000125354	septin 6	
	[Source:HGNC Symbol;Acc:15848]	[Type: protein coding Ensembl/Havana merge gene
ENSG00000102316	melanoma antigen family D, 2	
	[Source:HGNC Symbol;Acc:16353]	[Type: protein coding Ensembl/Havana merge gene
ENSG00000213684	lactate dehydrogenase B pseudogene 2	
	[Source:HGNC Symbol;Acc:6543]	[Type: pseudogene Havana gene]
ENSG00000179031	is no longer in the database	[Type: pseudogene Havana gene]

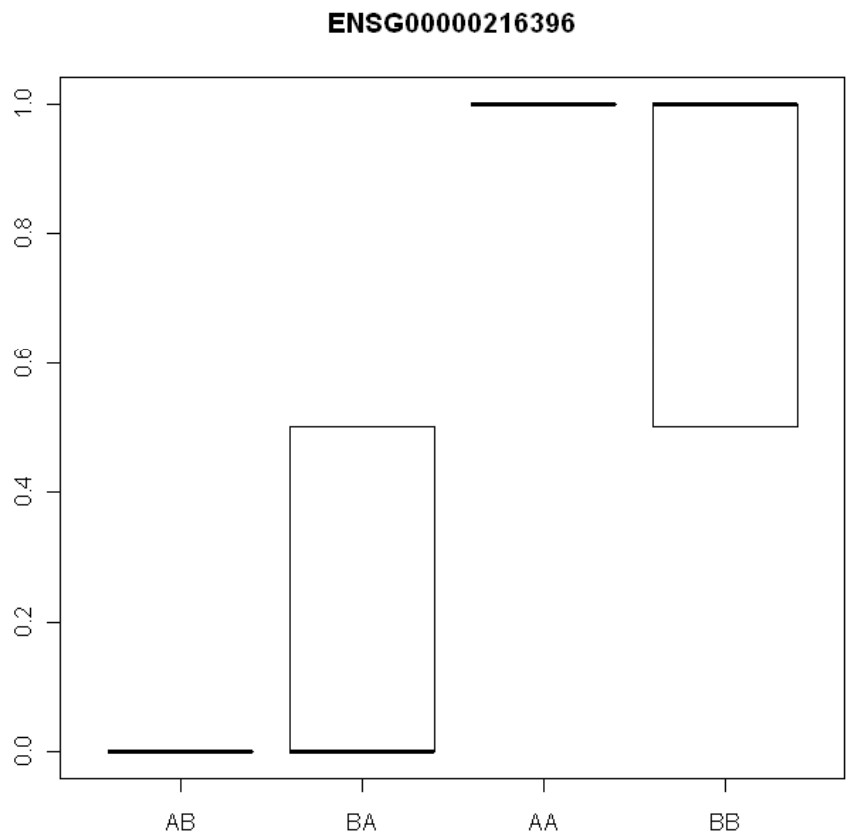
Closer look at ENSG00000213684

- In a graphs above we have seen that this entity had an interesting situation when we have 5 cases when allele A is more expressed and 7 cases when allele B is more expressed with p-values 10^-5.
- If we look for actual snp for this gene, we will find out that when A is more expressed we have AB=CG alleles, and when B is more expressed we have AB=GC, i.e. we just have different labels, and in all cases C is more expressed than G
- (also, these individuals happen to be the only heterozygotous at this snp, all the rest are homozygotous)

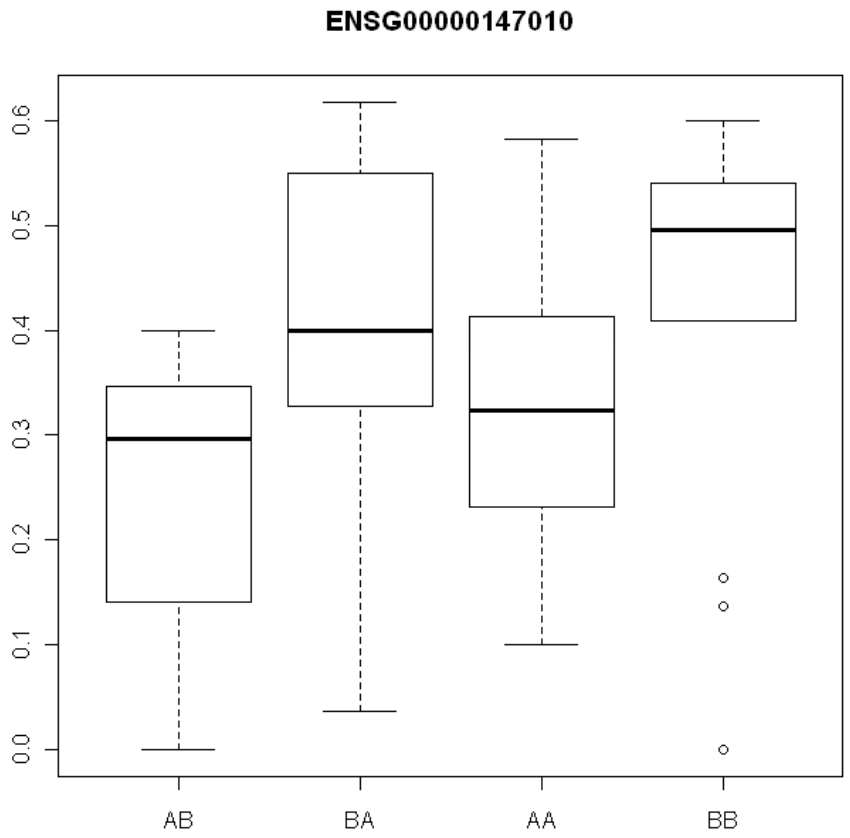
Other genes

- It is not as easy to compare other genes, since we have more than one snp for each gene. However we can try to calculate distances between haplotypes using some sort of distance.
- (For this case I used simple number of mismatches, other distances could be of interest too)
- Doing so, I could not see any significant evidence towards A=B for such genes: ENSG00000147155, ENSG00000196924, ENSG00000102265, ENSG00000102317, ENSG00000130826, ENSG00000102316, ENSG00000125354. Possibly we could see something with different metric.
- However for genes ENSG00000216396, ENSG00000147010, ENSG00000198359 and ENSG00000179031 one can notice some evidence:

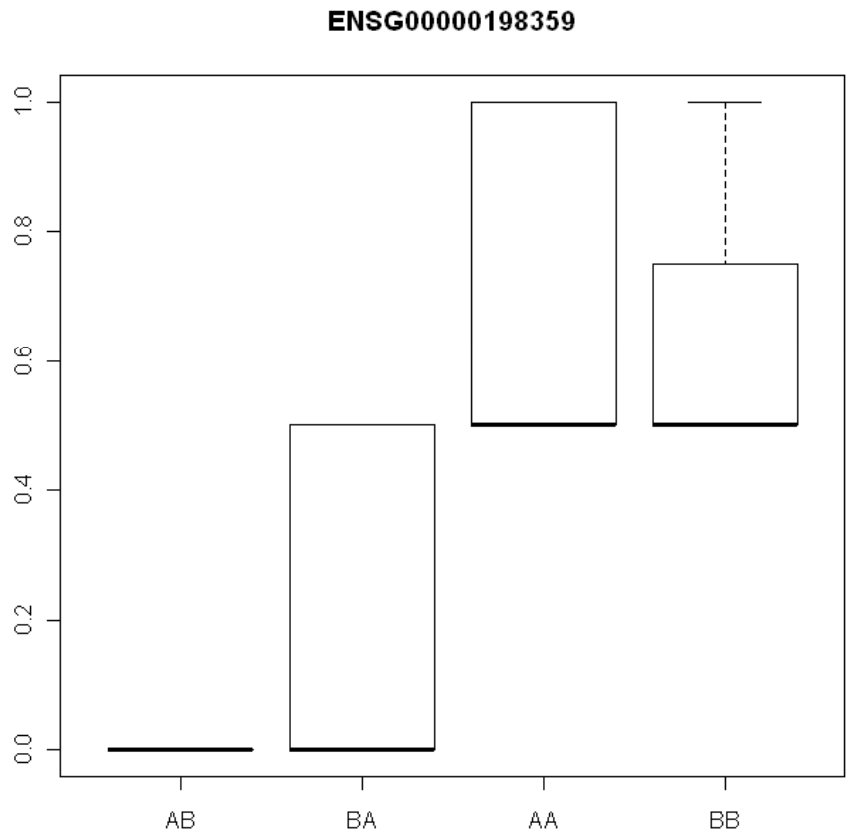
ENSG00000216396



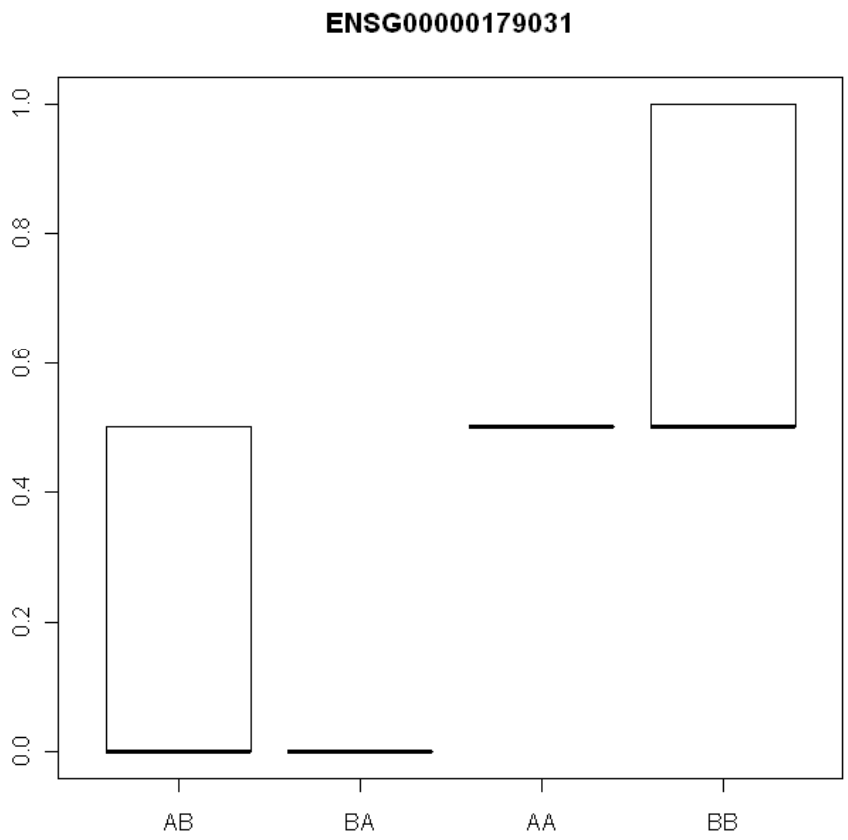
ENSG00000147010



ENSG00000198359



ENSG00000179031



Summary

- This project illustrates the behavior of genes on X chromosome.
- It has quite several highly expressed genes, whenever we have enough allele specific counts we typically see one allele much more expressed than the other.
- Also, we have seen several cases when we have allele A quite highly expressed and allele B quite highly expressed for different individuals. In one case we could totally attribute it to relabeling and in 4 more there are some evidence towards relabeling.
- In 7 more genes we could not see such switch. Possibly we observed this due to the effect of X-inactivation.