A table side chat on thinking about your NGS data statistically

Goals

I am not planning on trying to provide any sort of overview of statistical methods for genomic data. Instead I am going to provide a few short ideas to think about.

Statistics (like bioinformatics) is a rapidly developing area, in particular with respect to genomics. Rarely is it clear what the "right way" to analyze your data is.

Instead I hope to aid you in using some common sense when thinking about your experiments for using high throughput sequencing.

Useful reference

Paul L. Auer and R.W. Doerge 2010. Statistical Design and Analysis of RNA-Seq Data. Genetics. 10.1534/genetics.110.114983
PMID: 20439781

Designing your experiment before you start.

Sampling

Replication

Blocking

Randomization

Over all we are going to be thinking about how to **avoid Confounding** sources of variation in the data.

All of these are larger topics that are part of **Experimental Design**.

Sampling

Sampling

Replication

Blocking

Randomization

Sampling design is all about making sure that when you "pick" (sample) observations, you do so in a **random** and **unbiased** manner.

Proper sampling aims to control for unknown sources of variation that influence the outcome of your experiments.

This seems reasonable, and often intuitive to most experimental biologists, but it can be very insidious.
Whiteboard…

Sampling

Sampling

Replication

Blocking

Randomization

Replication

Sampling

R**eplication**

Blocking

Randomization

Imagine you have an experiment with one factor (sex), with two treatment levels (males and females).

You want to look for sex specific differences in the brains of your critters based on transcriptional profiling, so you decide to use RNA-seq.

Perhaps you have a limited budget so you decide to run one sample of male brains, and one sample of female brains, each in one lane of a flow cell.

What (useful) information can you get out of this?

Not much (but there may be some). Why?

Replication

Sampling

R**eplication**

Blocking

Randomization

Why?

No replication. How will you know if the differences you observe are due to differences in males and females, random (biological) differences between individuals, or technical variation due to RNA extraction, processing or running the samples on different lanes.

All of these sources of variation are confounded, and there are no particularly good ways of separating them out.

But there are lots of sources of variation, so how do we account for these?

Replication

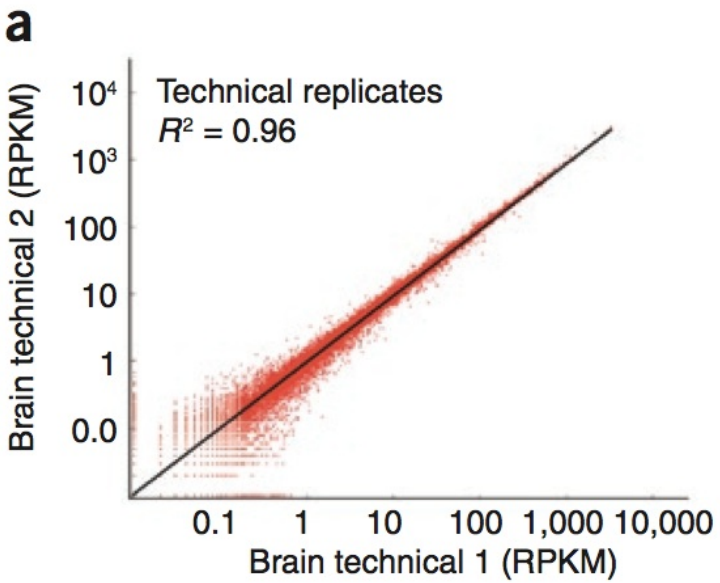
To date, several studies have suggested that “technical” replicates for RNA-seq show very little variation/ high correlation.

Sampling

R**eplication**

Blocking

Randomization



Mortazavi et al. 2008

How might such a statement be misleading about variation?

Replication

Sampling

Replication

Blocking

Randomization

This study looked at a single source of technical variation.

Running exactly the same sample on two different lanes on a flow cell.

This completely ignores other sources of "technical variation"
variation due to RNA purification
variation due to fragmentation, labeling, etc..
lane to lane variation
flow cell to flow cell variation

All of these may be important (although unlikely interesting) sources of variation…

However…..

Replication

Sampling

Replication

Blocking

Randomization

Many studies have ignored the BIOLOGICAL SOURCES of VARIATION between replicates. In most cases biological variation between samples (from the same treatment) are generally far more variable than technical sources of variation.

While it would be nice to be able to partition various sources of technical variation (such as labeling, RNA extraction), it often too expensive to perform such a design (see white board).

IF you have limited resources, it is generally far better to have biological replication (independent biological samples for a given treatment) than technical replication.

Does these lead to confounded sources of variation?

Blocking

Sampling

Replication

Blocking

Randomization

Blocks in experimental design represent some factor (usually something not of major interest) that can strongly influence your outcomes. More importantly it is a factor which you can use to group other factors that you are interested in.

For instance in agriculture there is often plot to plot variation. You may not be interested in the plot themselves but in the variety of crops you are growing.

But what would happen if you grew all of strain 1 on plot 1 and all of strain 2 on plot 2?

Whiteboard.

These plots would represent blocking levels

Blocking

Sampling

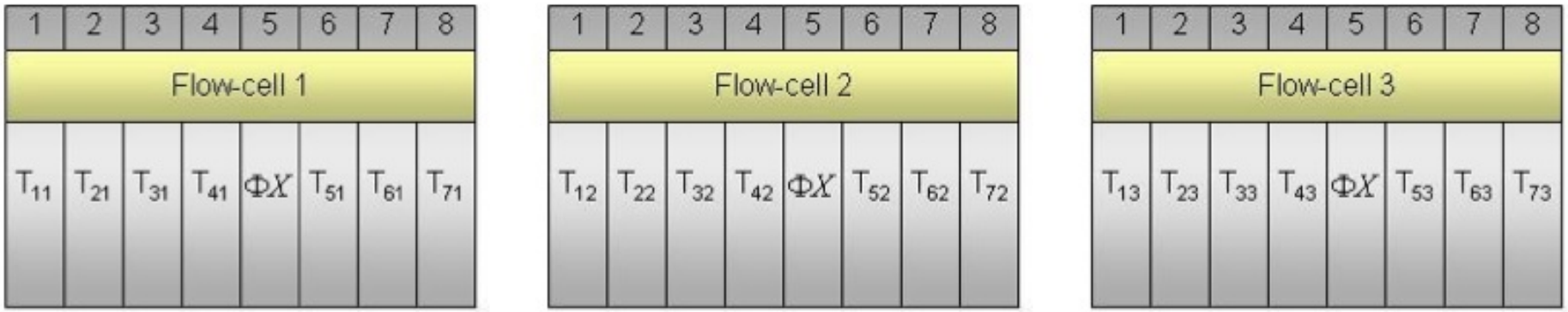
Replication

Blocking

Randomization

In genomic studies the major blocking levels are often the slide/chip for microarrays (i.e. two samples /slide for 2 color arrays, 16 arrays/slide for Illumina arrays).

For GAII RNA-seq data the major blocking effect is the flow cell. Soon the lanes within the flow cell will also be often used as a blocking factor.



Blocking

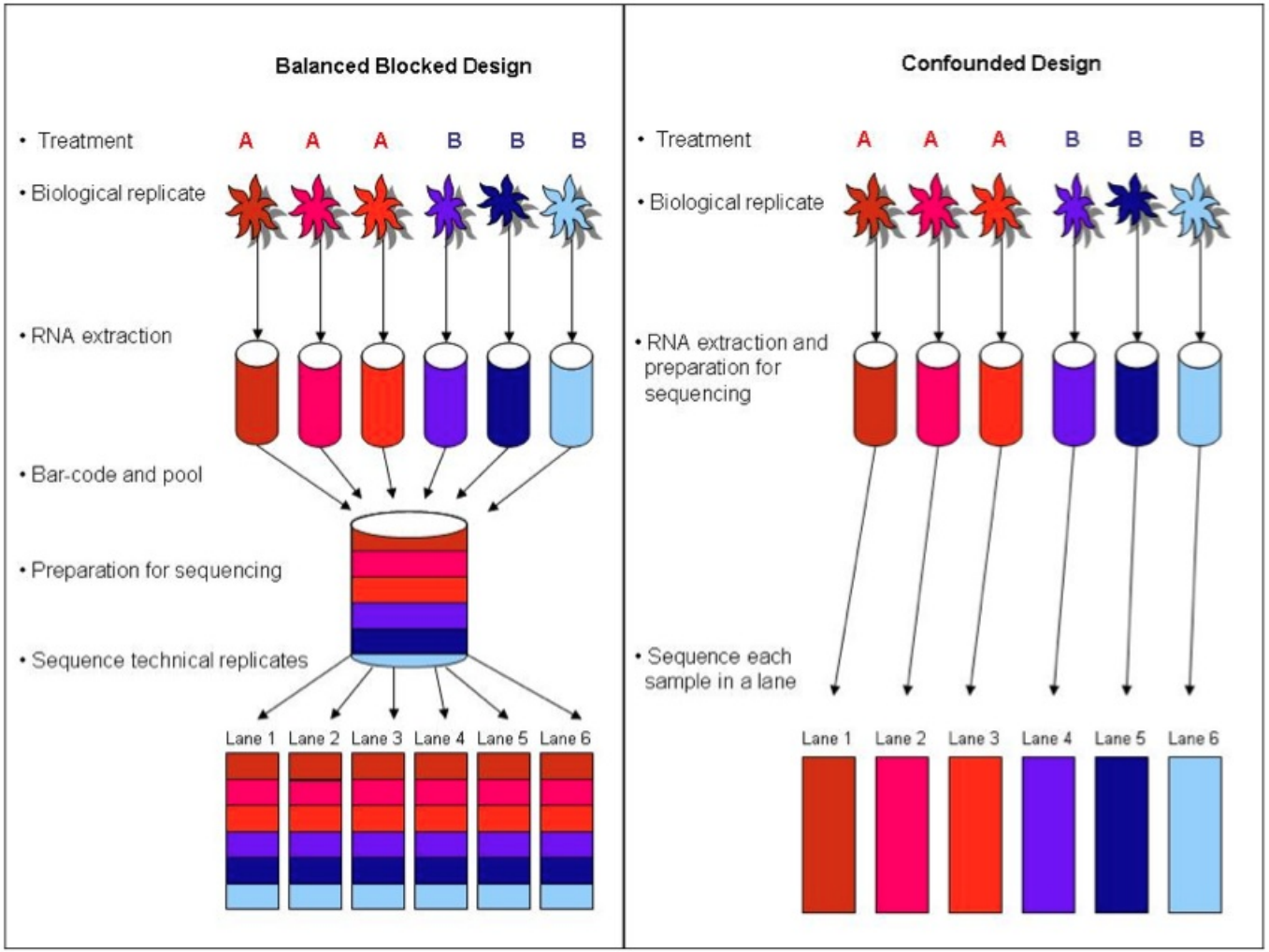
Sampling

Replication

Blocking

Randomization

Incorporating lanes as a blocking effect



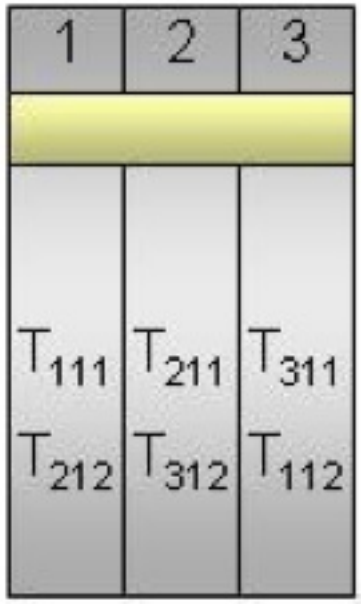
Blocking designs

Sampling

Replication

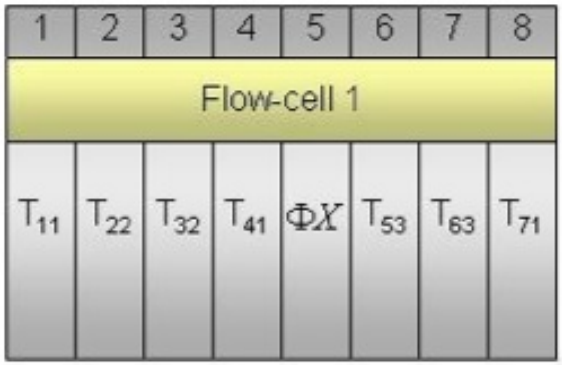
Blocking

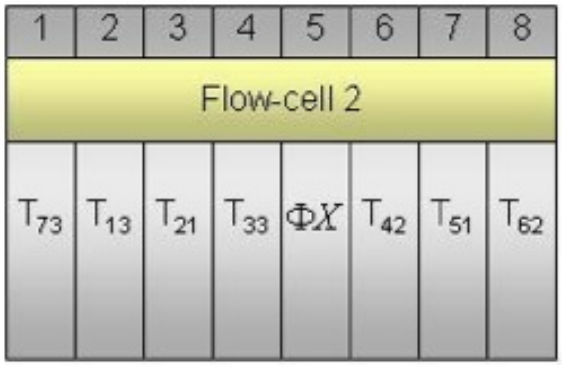
Randomization

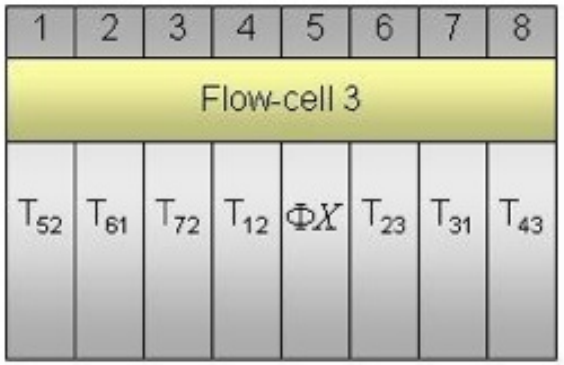


Balanced **I**ncomplete **B**locking **D**esign (BIBD)

Let's dissect these subscripts.







Balanced for treatments across flow cells.. Randomized for location

You have designed and run the experiment… now what?

First a couple of quotes from a great statistician:

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.
John Tukey

Numerical quantities focus on expected values, graphical summaries on unexpected values.
John Tukey

Graphical examination of your data.

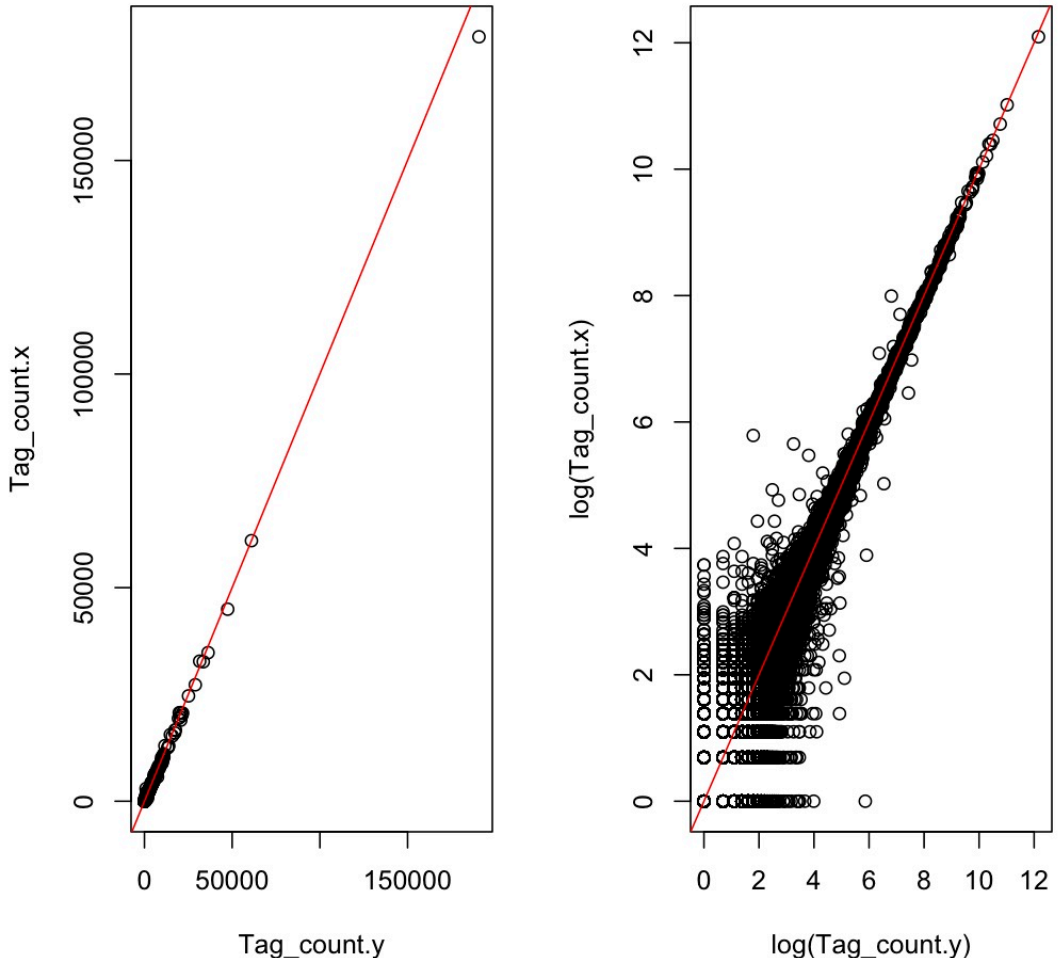
By far the single most important thing you should be thinking about with your data, at every stage of the analysis (raw, filtered, normalized, modeled) is how to present the data graphically.

Plots are a very good way to pick out if something wacky is going on with your data.

Some examples: Digital Gene Expression (Sequence tags) for RNA quantification

A comparison of two lanes of DGE sequence tags.

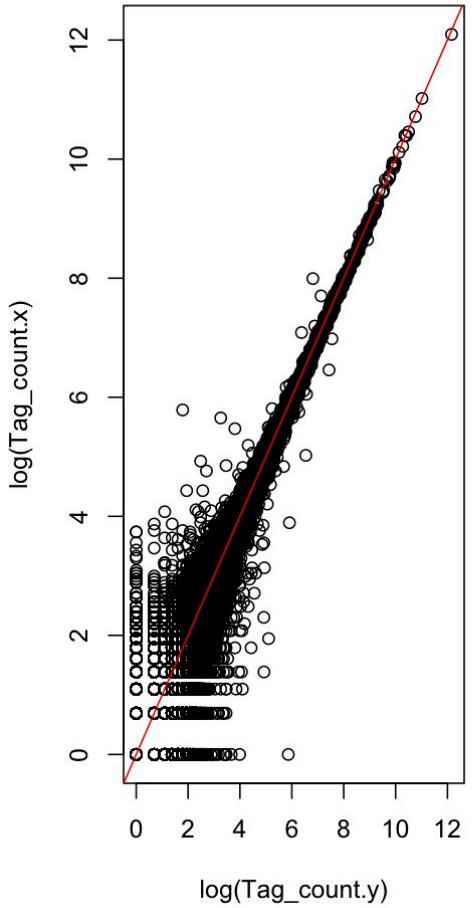
What's the difference between these plots?

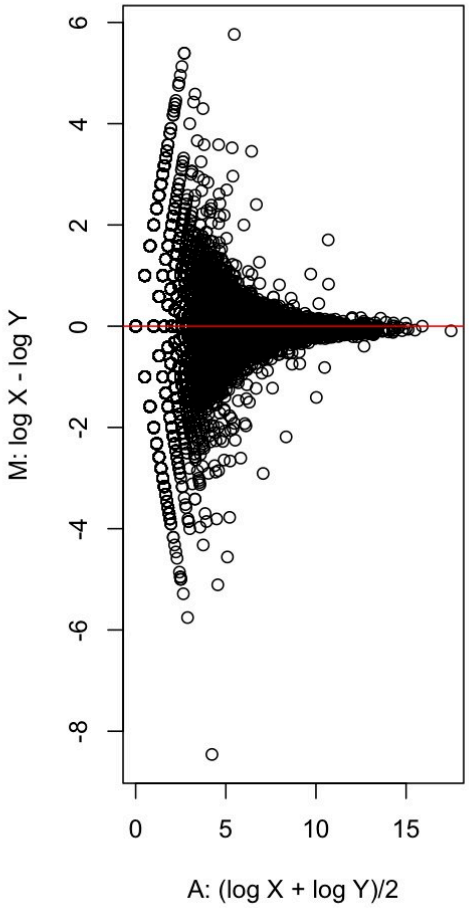


Some examples: Digital Gene Expression (Sequence tags) for RNA quantification

A comparison of two lanes of DGE sequence tags.

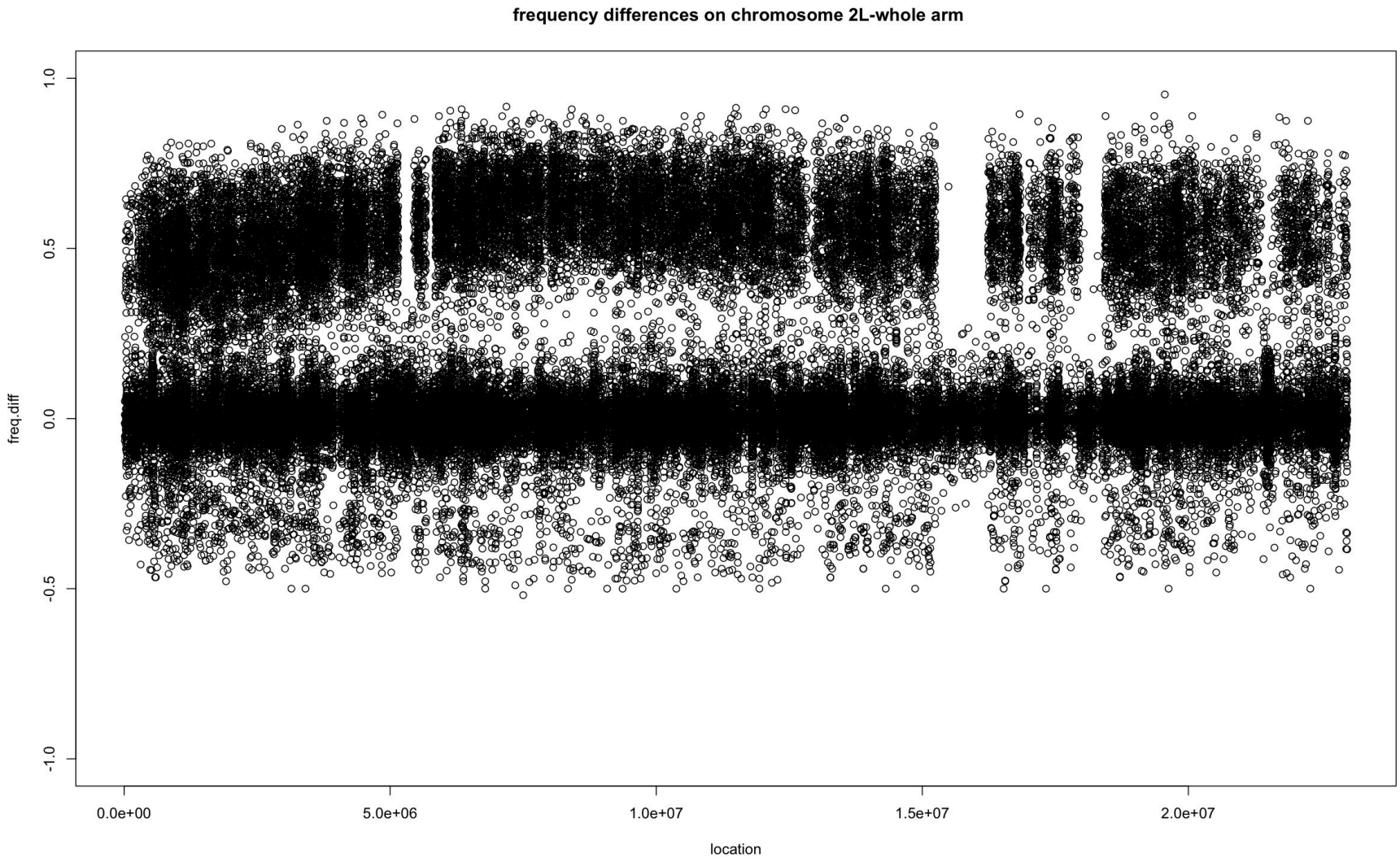
What's the difference between these plots?



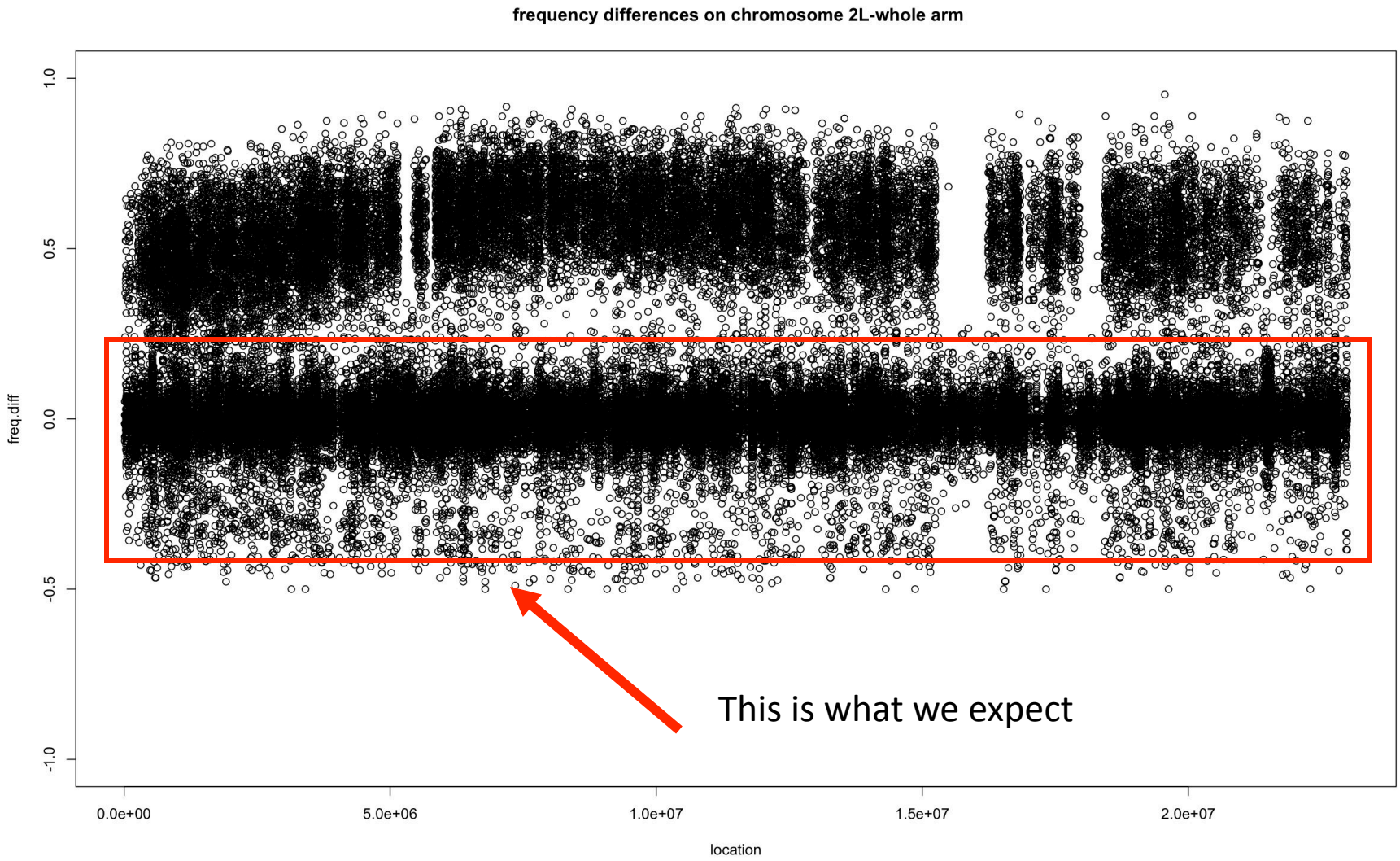


MA plot

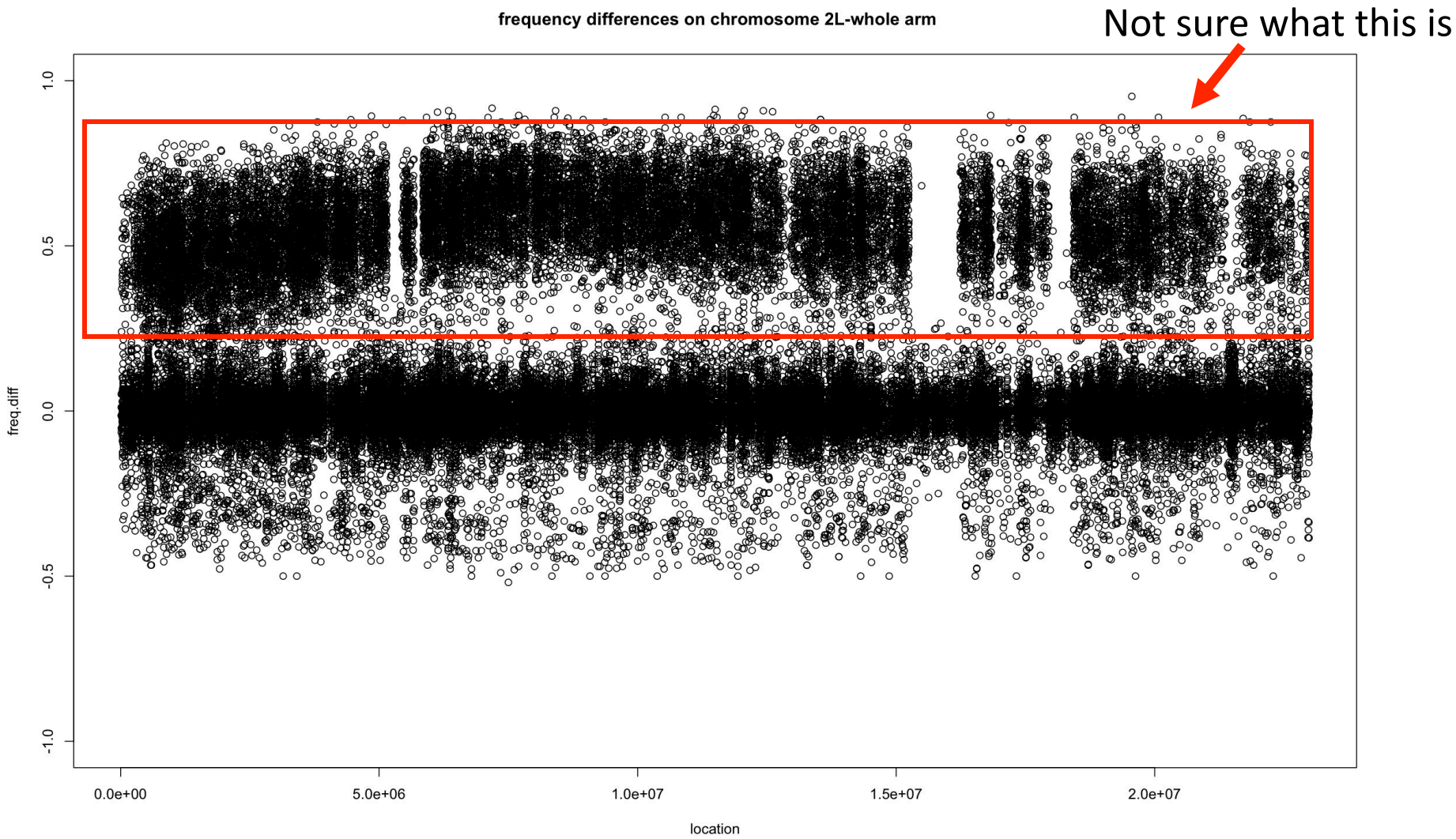
Some examples: resequencing pools of individuals for mapping…



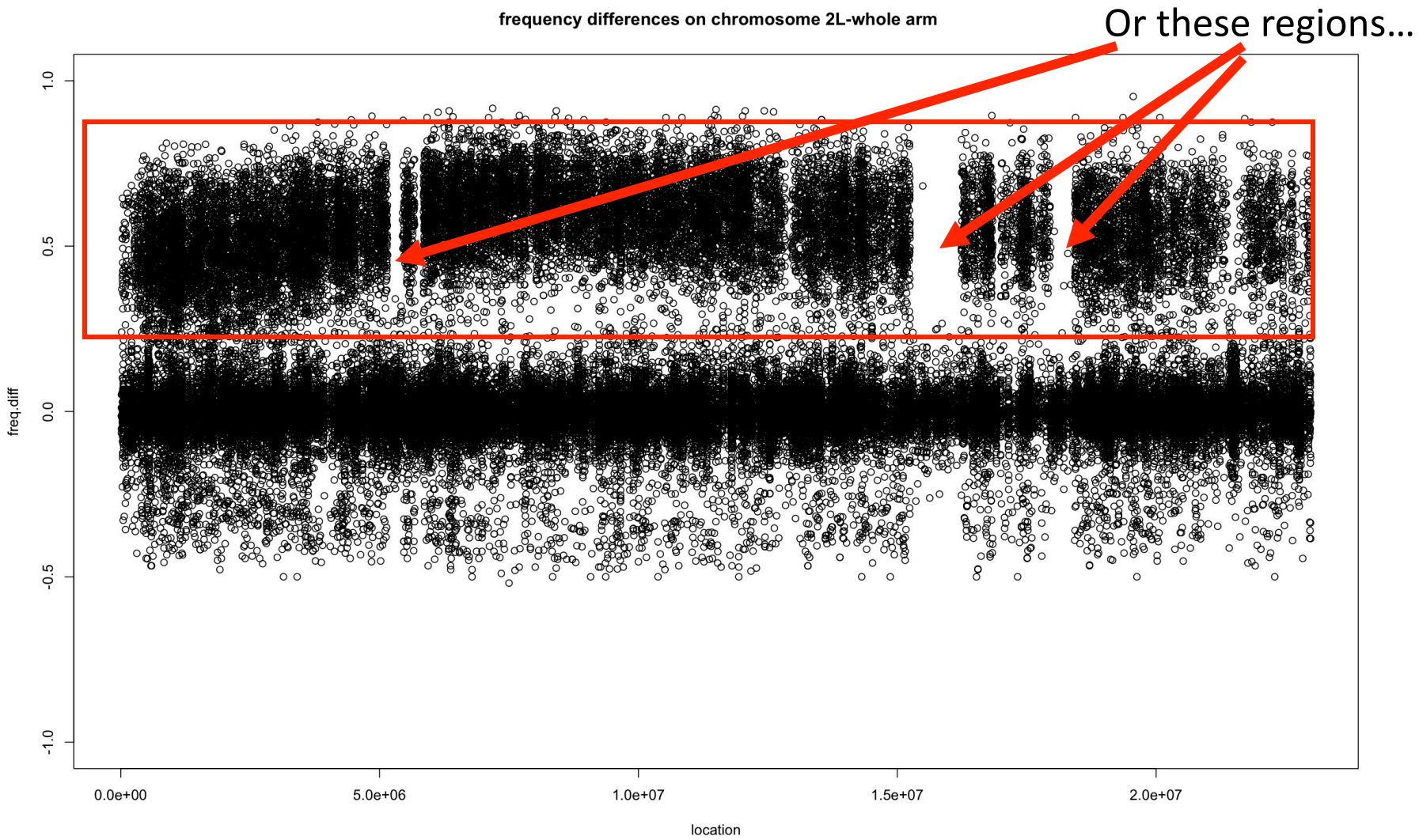
WTF….


frequency differences on chromosome 2L-whole arm
freq.diff
1.0
0.5
0.0
-0.5
-1.0
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
location
This is what we expect

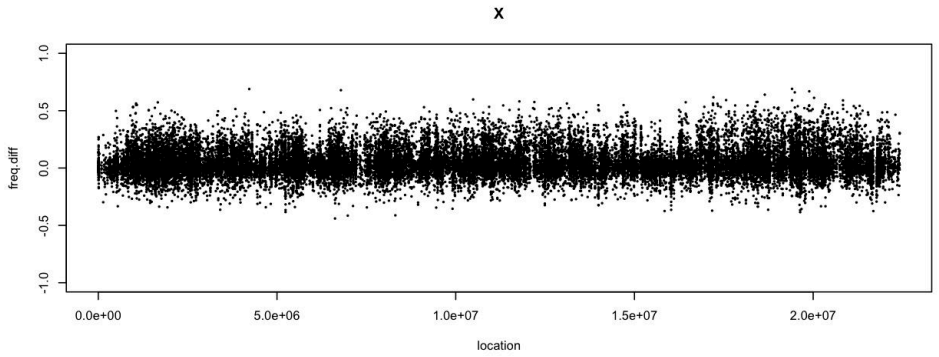
WTF….

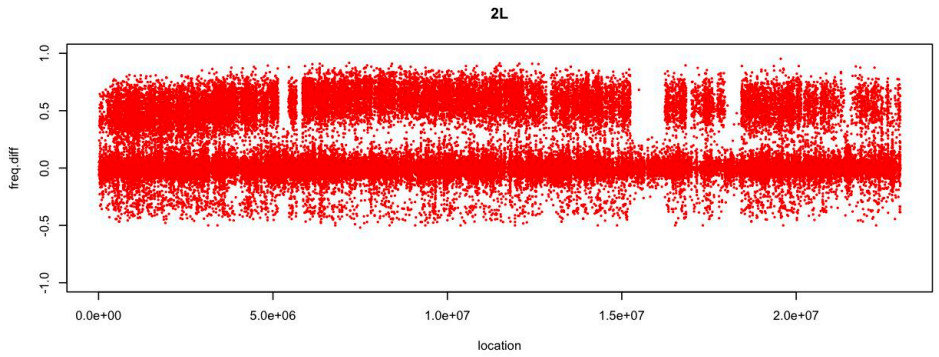

frequency differences on chromosome 2L-whole arm
Not sure what this is
freq.diff
1.0
0.5
0.0
-0.5
-1.0
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
location

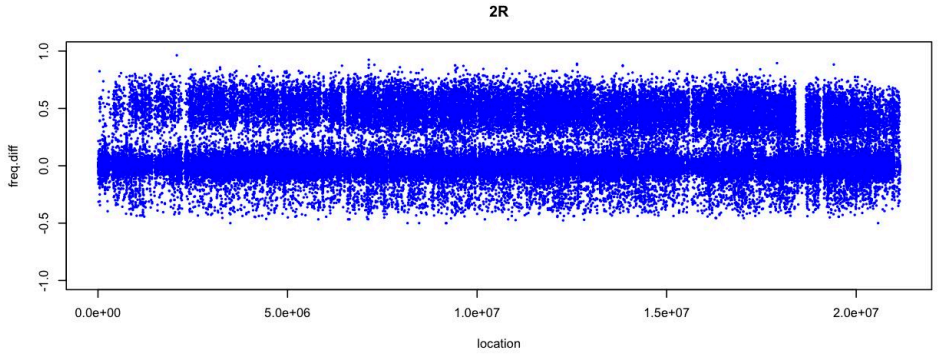
WTF….

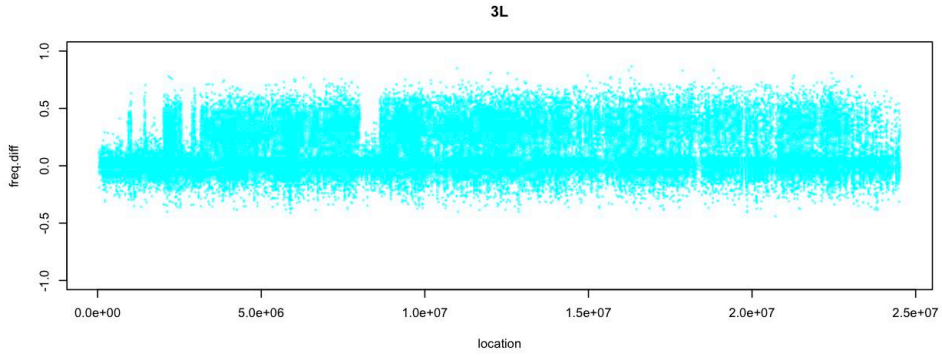

frequency differences on chromosome 2L-whole arm
Or these regions…
freq.diff
1.0
0.5
0.0
-0.5
-1.0
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
location

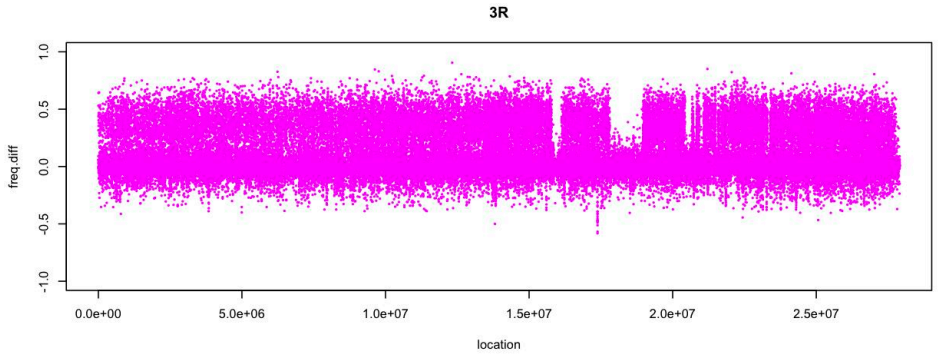
Whole genome







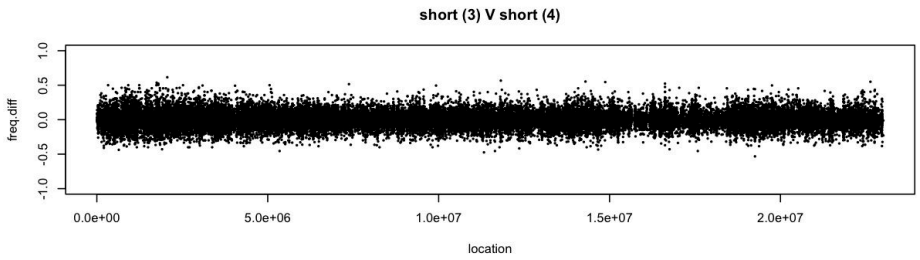


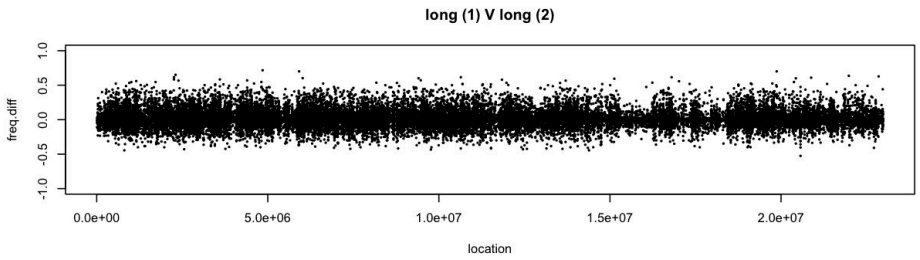


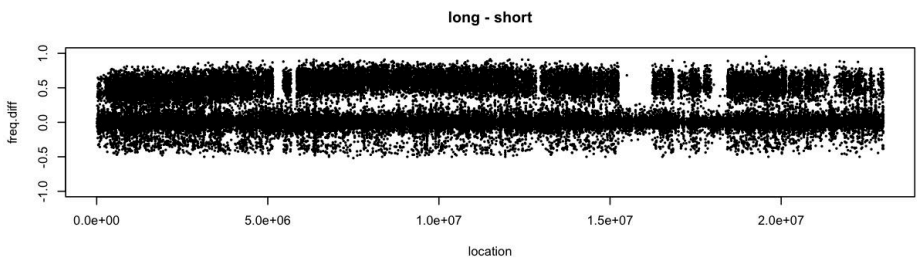
Why are there two bands

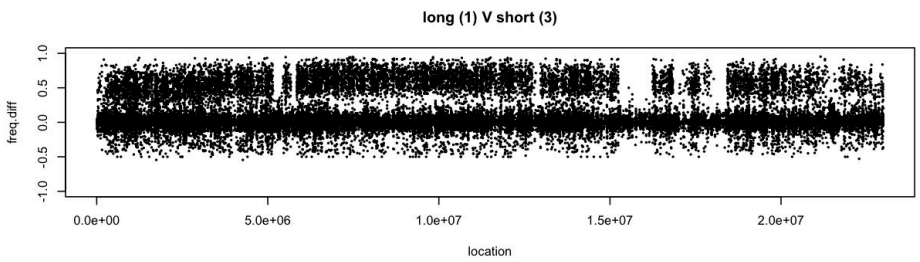
- You could get the upper band if there was highly unequal genomic contributions in the F2's. However you would not get the lower band in this case.
- Trying to figure it out.
- First thought, one of the lanes got really messed up.

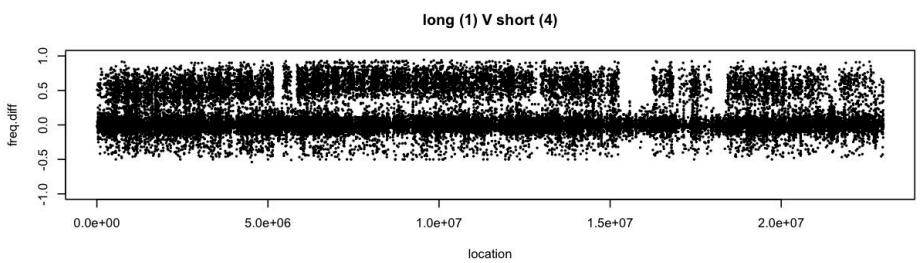
Troubleshooting….

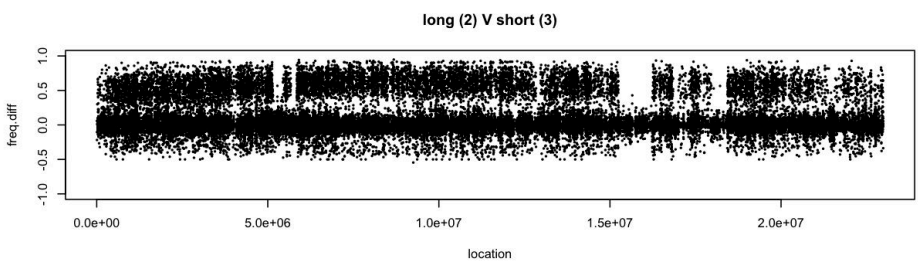


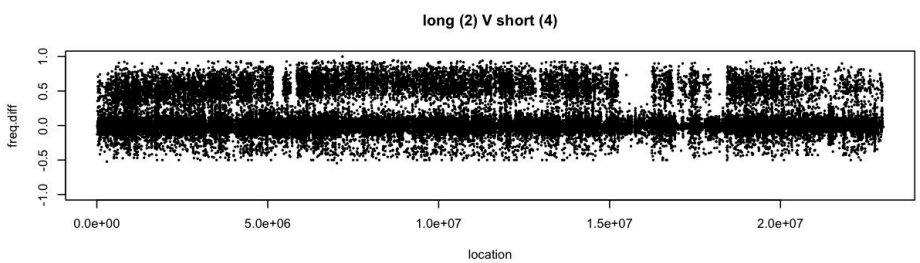


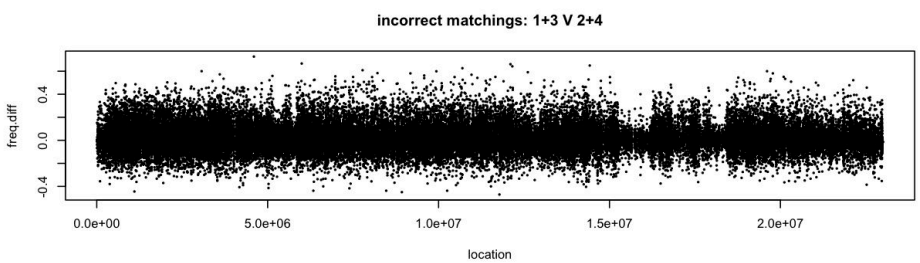












Plotting

- People are developing new approaches to plotting the data all of the time. Titus will be showing you some approaches, in particular to spot potential regions of reads (3' end) with errors.

- It is also important to map the reads back onto the genome to make sure this is sensible (Likit, Jason and Rose).

Final thoughts

- One BIG mistake people make with BIG genomics data sets is treat each gene (or genomic interval) as an independent data point. In particular for correlation analysis.

- This is almost never the case…

chalkboard.