UPPSALA
UNIVERSITET

Normalization of Human Y Chromosome SNP-Array Data

Using Female Population Data

Jon Jakobsson

Degree project in biology, Bachelor of science, 2013
Examensarbete i biologi 15 hp till kandidatexamen, 2013
Biology Education Centre and Department of Organismal Biology, Uppsala University
Supervisors: Martin Johansson and Elena Jazin

Abstract

A common way to examine chromosomes and genes is by the use of Single Nucleotide Polymorphism arrays (SNP-arrays). These arrays can detect genes and report what alleles the person have. This is done by hybridization of specific DNA sequences. There is also the possibility to detect at what copy number a specific sequence is in the DNA of the subject. Copy number differences can then be linked to phenotypic traits. The X and Y chromosome have evolved from a regular pair of autosomes. Even to this day, some parts of the X and Y chromosomes have high sequence similarity. SNP-arrays uses hybridization to find specific sequences, and this becomes a problem when examining the X and Y chromosomes, since their sequence similarity can make it hard for the SNP-array to distinguish between X and Y sequences. When examining males Y chromosomes with an SNP-array, there is always the possibility that some of the sequences originated from the X chromosome. Females examined in the same way should also show some sequences on the Y chromosome, even though they have no Y chromosome. This study will try to use the female data (XX) and the male data (XY) to find only the information originating from Y. The subjects come from a control group of a Norwegian schizophrenia study. 49 females was used to find a signal profile on the Y chromosome and 240 males was corrected using this information. Many novel copy number variations was found by comparing the copy number variations found before and after the correction. These novel detections where mostly located at the areas with high X and Y sequence similarity. Four of these novel putative deletions have been tested with PCR, and none of them was confirmed.

Contents

Introduction

The Y chromosome

The evolution of sex chromosomes is not something unique to mammals. Animals such as birds, snakes, amphibians and fish are all vertebrates with a separate sex chromosome evolution. Furthermore, invertebrates have also evolved sex chromosomes many times. All of this convergent evolution gives a clue of the benefits of sex chromosomes. The mammalian sex chromosomes emerged 200-300 million years ago with the addition of a sex-regulating gene on one of two regular autosomes. This new "proto-Y" chromosome is identical to the "proto-X" chromosome except for the new sex-determining region on the Y chromosome (SRY). This acquired difference has a great impact on the evolution of these two chromosomes. The fact that females have two X chromosomes allows for purging of harmful genes by recombination. Males on the other hand, have one Y chromosome and one X chromosome, and these have large areas that cannot recombine. This will lead to accumulation of mutations and deletions in this area. Today, the Y chromosome only has 78 protein-coding genes compared to approximately 800 in the X chromosome. When the SRY had just emerged, the X and Y chromosomes were still recombining all over the chromosome, except over the new SRY region. With the increase in mutations on the Y chromosome, the area without recombination grew steadily. That area does now cover 95% of the chromosome, leaving 5% that still recombine with the X chromosome. This 95% big "non-recombinant region" was earlier called the NRY. It was then discovered that a large segment of the X chromosome has been transposed to the Y chromosome around 3-4 million years ago. It is now called the "male specific region" (MSY), since recombination has obviously occurred. (Bachtrog, 2013).

There are long stretches of repetitive sequences all over the Y chromosome, which makes it very hard to sequence the chromosome. These sequences are called heterochromatin and they cover large part of the Q-arm and the centromere. The remaining sequences are called euchromatin and they can be categorized into X-transposed, X-degenerate and ampliconic sequences. What characterize these categories will be covered in subtopics below. Figure 1 shows the layout of these sequences. (Skaletsky et al., 2003)

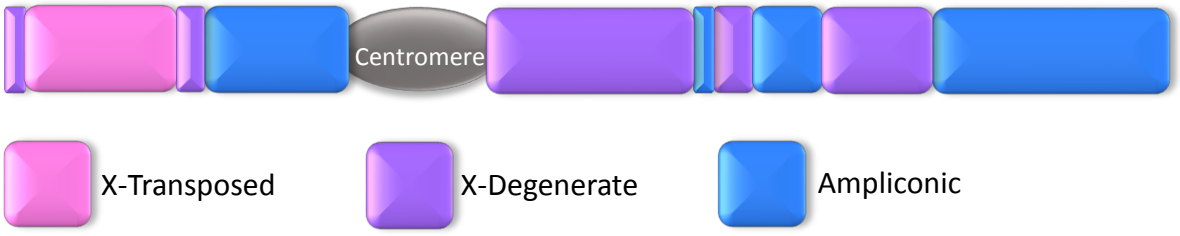

Figure 1: The different types of sequences on Y chromosome euchromatin. The X-transposed sequence (pink) is on the P-arm (left of the centromere). X-degenerate (purple) sequences are on both the P-arm and the Q-arm (right of the centromere). Ampliconic (blue) sequences in the P-arm are tandem repeats and ampliconic sequences on the Q-arm are palindromes. Modified from (Skaletsky et al., 2003)

X-degenerate regions

Sequences that show a sequence similarity of 60-96% to the existing X chromosome are called X-degenerate. These regions have been conserved throughout evolution for their vital functions and there are around 16 genes in these sequences, most of them are genes with housekeeping functions. (Bachtrog, 2013) (Skaletsky et al., 2003)

X-transposed regions
Around 3-4 million years ago, a large chunk of DNA was transposed to the P-arm from an X chromosome. This is unique to humans, and is hence not found in the chimpanzee Y chromosome. These sequences have 99% sequence similarity to sequences of the X chromosome. (Bachtrog, 2013) (Skaletsky et al., 2003)

Ampliconic regions
There are two types of ampliconic sequences. On the P-arm of the Y chromosome, there are tandem repeats, and on the Q-arm, there are eight palindromes. A palindrome is two sequences (arms) that are inverted copies of each other. These are found next to each other on the chromosome. The ampliconic sequences have 60 protein-coding genes from nine families, and they are mostly expressed in the testis. There are multiple copies of genes on different arms of a palindrome. This facilitates for gene conversion (recombination between sequences on the same chromosome), which can help protect important genes form deleterious mutations. One could argue that this is an adaptation to the Y-chromosome's haploid nature, since it provides opportunity for the chromosome to self-recombine, and hence remove deleterious mutations. As an effect of this, the palindrome arm sequences show up to 99,9% identity to each other. There are many different origins to the sequences in the ampliconic regions. Some genes originate from autosomes, and some are remnants from the old X chromosome. (Bachtrog, 2013) (Skaletsky et al., 2003)

SNP arrays
With the completion of the human genome project, scientists now have a crude reference of what the human genome looks like. The fact that every person has a slightly different genome does however complicate things. One such difference can be a single nucleotide polymorphism (SNP), where one person has e.g. an A base and another has a C/T/G base at a specific location. Other differences can be more prominent, where a person has duplications or deletions of whole genes or chromosomes. (Scharpf et al., 2008)

SNP-arrays employ the DNA-molecule's hybridization property. Hybridization means that it will form bonds with complementary DNA strands. Probes (complementary DNA strands) are produced on a microchip, and every probe can hybridize with a different part of the human genome. Many identical probes make up a patch on the macro-chip. These patches are called markers. To know where hybridization has occurred between the sample DNA and the probes, florescent molecules are attached to the sample's DNA. A machine can then laser-scan the microchip and report with what intensity every marker is illuminating (Siggberg et al., 2012) (Figure 2). Some markers are designed to hybridize to the exact same position on the genome, but they will prefer slightly different versions of that position. These are called SNP markers, and they will report information about one SNP. If one of these markers has intensity, its sequence is the one present in the sample DNA. If none of the SNP markers has any intensity, that sequence is not present in the sample DNA. See SNP 1 and SNP 2 in figure 2. The person can be either homozygote or heterozygote for a given position in the genome, and this will be determined by an algorithm comparing all SNP markers for that SNP. Modern microchips also contain markers that only detect presence of a given sequence, and will not report any SNPs. These are called copy number markers (CN markers) or non-polymorphic (NP) markers (Wang & Bucan, 2008). If a CN marker has very low intensity values, that sequence is not present in the person's genome. With high intensities, there are many copies of that sequence in the person's genome.

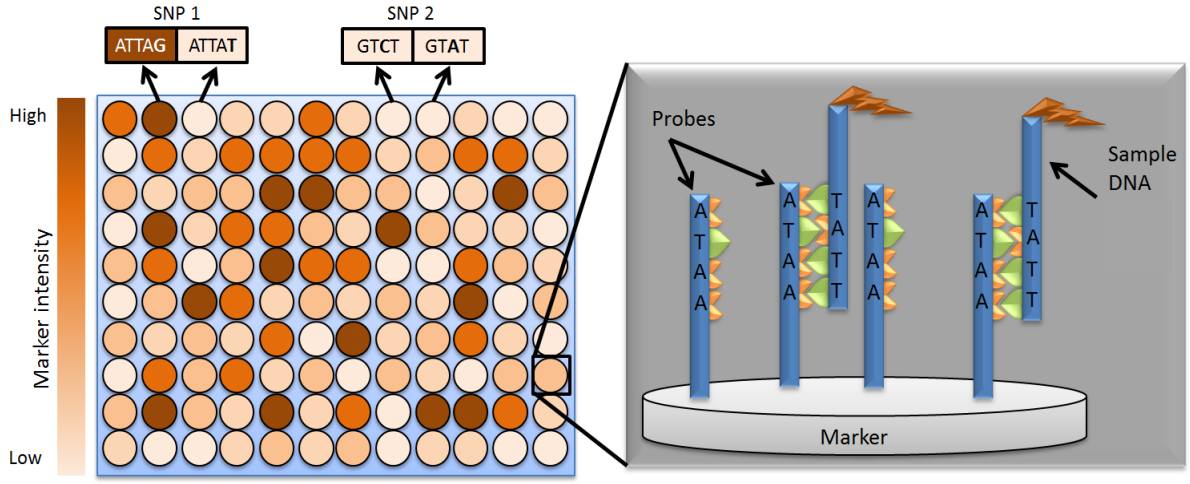

Figure 2: Microchips are filled with markers**.** Markers are patches filled with short DNA sequences called probes. These probes can hybridize with the specific sequences of the sample's DNA and make it stick on the marker after washing the microchip. The sample DNA have a florescent molecule adhered to it. The molecule will illuminate in one wavelength if illuminated by another, and this is utilized to determine the amount of molecules and hence the amount of sample DNA that got stuck to every marker. SNP 1 has the sequence of ATTA**G** in this sample, not ATTA**T**. SNP 2 is not present at all in this sample.

Problems

When doing copy number (CN) analysis of a genome, there is no guaranty a marker only hybridizes to sequences actually representing the sequence that marker was designed to represent. E.g. if parts of chromosome 1 are copied to chromosome 2, the SNP array will report a duplication on chromosome 1, and show a normal chromosome 2, even though it is chromosome 2 that has the aberration. A similar effect could occur on the sex chromosomes. As mentioned before, the X-transposed region on the Y chromosome shares 99% similarity to the X chromosome. This could in theory mask true deletions of this area in males, since the male's X chromosome could be the source of the signals.

Goals

The first hypothesis is that females should increase marker intensity on the X-transposed areas on the Y chromosome SNP-array. As an extension to this, the X-degenerate areas should have higher marker intensity than the ampliconic areas. This will be tested by analysing many females from the same population and examining the average intensity of all markers to the map of the sequence type regions in figure 1.

Since females produce a signal on the Y chromosome markers, that data will be used to correct marker intensities on males. The female markers with a high average intensity over all females are considered erroneous since females do not have a Y chromosome. These markers will be used to subtract intensities from the corresponding male markers' intensities. Female markers with low average intensity are not erroneous since they should be close to zero. These markers will not subtract anything from their corresponding male markers. The SD of one marker between the females indicates the confidence that can be put on that markers behaviour. Low SD will therefore increase the modification and high SD decrease the modification made with that marker's average intensity. This will in theory remove some internal artefacts of the SNP-array and reduce the intensities that the X and autosomal chromosomes induce on the Y chromosome markers.

Using an algorithm that reports copy number variations (CNVs), the effect of the modification can be examined by comparing the original data to the modified data. The aim is to find novel deletions in the X-transposed area. These deletions will then have to be confirmed or discarded by conventional PCR.

Since the examination is made on males from one population, there is an expectations to find similar CNV (the Y chromosome is paternally inherited), and therefore some CNV pattern should be detectible. The same argument can be made for SNP-array artefacts, which will be due to the design of the array.

Method

Preparing the data

All raw files (CEL-files) were analysed using Affymetrix's Genotyping Console™ Software (GC). All files were imported to GC and a quality control was performed. All samples that were inbound (passed the control) were used to perform a Copy Number/LOH analysis. The samples with a MAPD lower than or equal to 0,35 (default threshold in GC) were selected for use in the study. With the export Copy Number/LOH results function, chromosomal position and intensity values (Log2) for the Y chromosome were exported to a separate text file for every sample. Male and female sample text files were placed in different folders.

Modulation of data

All calculations were performed using Microsoft's Excel software.

The mean values and standard deviation (SD) of the Log2 values for markers between the females were calculated, and saved in new columns.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = mean\ value\ (mean\ value\ of\ one\ marker)$$

$$n = Sample\ size\ (number\ of\ females)$$

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The modulation of the male Log2 value was based upon the female mean and standard deviation for that particular marker. In order to use simple subtraction as modulation, a linear relation was needed between the Log2 values and the copy number state (CN state). To calculate this, a data list with Log2 values corresponding to CN state were taken from the GC's Hidden Markov Model (HMM) default settings. By using an iterative function, the Log2 values could be exponentiated with the base that gave the highest R^2-value for a linear fitting. This base is annotated as Best Base (BB) in the following equations.

To determine the amount of influence (gain) a specific marker had when modifying the corresponding male marker, the females SD of that marker was used:

$$Gain = e^{-SD*x} * y$$

Where x was a user setting modulating the effect of the SD, and y was a user setting amplifying the whole gain. The amount to be subtracted from the $maleBB^{Log2}$ was determined with this equation:

$$Correction = Gain * (femaleBB^{Log2} - y\ interception)$$

The y interception made sure that a female marker representing CN state 0 (no Y chromosome present) gave zero in value to the correction. The final modulation of the male Log2 was done by this equation:

$$Modified\ male\ Log2 = Log_{BB}(Normal\ male\ BB^{Log2} - Correction)$$

Copy Number Analysis

To fit a normal curve to the male sample, the SD and the mean of that sample were needed. With this information, probabilities could be calculated, and significant CNVs could be found. This, however, assumed that the whole sample had the same mean. When analysing copy number variations, there is an assumption that the mean should differ at the deleted or duplicated areas. Therefore, for a sample with many CNVs the SD should be high, which would have made it hard to use probabilities to detect these CNVs. There was however a method to calculate an alternative to SD, that would allow the sample mean to shift. The Median of the Absolute Values of all Pairwise Differences (MAPD) measured the sample's variation and allowed for local shifts. By pairing every marker with its neighbour and measure the absolute differences between them, the median of these values would represent MAPD. The relation between MAPD and SD is $MAPD/0{,}96 = SD$ when Log2 values are distributed normally. (Affymetrix®, 2008)

Copy Number variation criteria

To report a CNV, five consecutive markers had to be beyond a threshold (upper or lower limits for duplication or deletion respectively). The threshold was designed to give a less than five percent probability for the event to occur by chance per sample. Using the excel function "*NORM.INV*" a value could be returned that covered x percentage of the sample, starting from low to high using a normal distribution.

$$Upper\ Limit = NORM.INV((1 - x), mean, MAPD)$$

$$Lower\ Limit = NORM.INV(x, mean, MAPD)$$

The probability of finding one randomly chosen marker outside of the upper/lower limit was x. The probability to find five in a row was x^5

The number of tries to find x markers in a row was:

$$Tries = total\ number\ of\ markers - x\ markers\ needed\ in\ row + 1$$

The total chance to find a false positive was then:

$$Chance\ to\ find\ 5\ markers\ in\ a\ row = Tries * x^5$$

X was selected to give a false positive under 5% for five markers in a row.

When a CNV was detected, the start and stop positions of that variation were stored in a list together with the settings used and the sample number/name.

Graphical representation

The individual males were plotted using their markers' Log2 values against the markers' chromosomal positions. A running Log2 mean with a window of 31 markers was plotted for both the modified and original male data in the same graph. Two other graphs plotted the modified respectively original data including the males' Log2 values and their 31 running window Log2 mean. Opaque orange squares represented a five running window Log2 mean that went over or under the upper or lower limits. Red squares represented a 31 running window Log2 mean that passed the

limits. Black triangles represented markers that had passed the limits at least five markers in a row (the statistically significant sensor).

Accompanying the graphs were information about the current settings, sample information (average, MAPD, confidence calculations) in a tabular form and a graph showing the amount of gain and SD for every marker (Appendix, Figure 1). This information was saved as one XPS file for every scanned male.

User prompted data

It was not clear what would be a good way to use the SD to influence the gain for a given marker. To allow for easy experimentation with this value, it was implemented as a setting. Other settings were the upper, lower limits and the gain amplification.

Saving data

Working with the data in excel could be very repetitive. Excel did however allow for macros, which could be used to automate different tasks. The excel file developed during this project had one customized ribbon with one button. This button would do everything from importing new male data, updating the calculations, saving a graph as an XPS document, copy-pasting information from 4 tables to one table and finally selecting all the new data rows on the "sample name" column allowing for easy naming of the sample just added. This new table could be exported to R for further analysis, or to another excel document.

R statistical software

R is a statistical software package used by many scientists to make models and perform statistical analysis. It also has powerful graphing features. Two R functions were designed during this project. The first function would collect all the information from the female files outputted from GC and output a table with all the females' Log2 values. This file could then be imported into the excel document as the female data (instead of adding them one by one by hand).

The second function would give a CNV summary plot where every CNV found in the analysis was plotted in one graph. The sample number determined the plotting position for that sample on the y-axis. In this plot, the modified data were masked by the normal data to allow for detection of novel CNV's. This function would also make a CNV density plot, which show how many times a CNV, was found at a chromosomal position. These plots were saved to separate files. See figure 6 and appendix, figure 2.

Results

The data

The initial data was acquired from a Norwegian research group lead by prof. Ingrid Agartz and Srdjan Djurovic. Some of the data were out of bounds after quality control (QC), and were hence left out from the study. Table 1 shows where the samples failed in the GC analysis.

Table 1: Showing the initial sample size and in what quality control step samples were out of bounds. The sample size left is the samples used in the modification analysis.

Gender	Females	Males
Initial sample size	63	271
Out of bounds QC	-1	-18
Out of bounds MAPD	-13	-13
Sample size left	49	240

The Hidden Markov Model data from GC was used to calculate the best base (Table 2 and Figure 3).

Table 2: Log2 values and their corresponding CN state. Default settings used by the official GC software HMM.

CN state	Log2 values
0	-2
1	-0,533
2	0
3	0,363
4	0,567

The iterative function in Excel found the base 3,18 to be the best base (BB) for a linear fitting of this dataset.

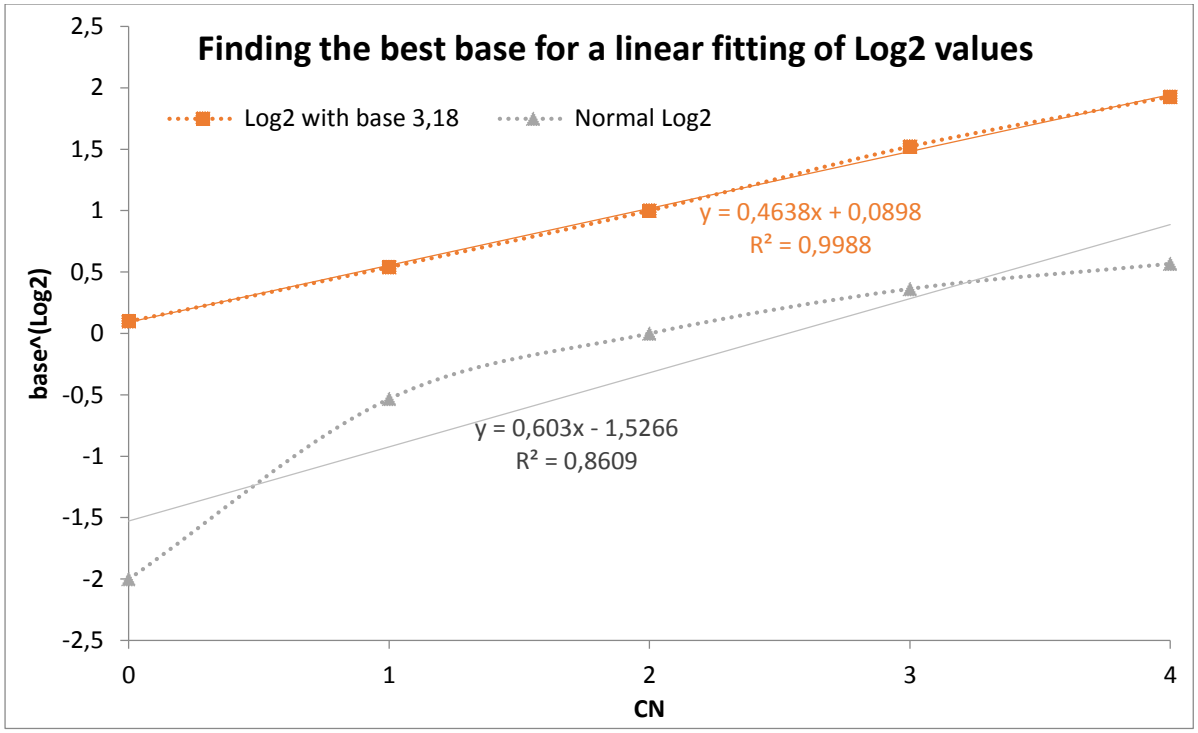

Figure 3: Log2 values with the base of 3,18 is shown with orange squares connected with a dotted orange line and with an orange linier trend line and the trend line's equation. The normal Log2 values are shown with grey triangles connected with a grey dotted line and with a grey trend line and corresponding trend line equation. Log2 values with a base of 3,18 gives the best linear fit (highest R^2).

The y intercept was 0,0898 as shown in the trend line equation for the 3,18 base data in figure 3.

Female statistics

The female data was analysed using the average from 49 females. The average Log2 values were calculated. They were illustrated using a running average of 50 markers. The average Log2 values correlated to the three different sequence types on the Y chromosome. The X-transposed had the highest Log2 values. The X-degenerate had medium Log2 values, and the ampliconic had the lowest Log2 values. See figure 4.

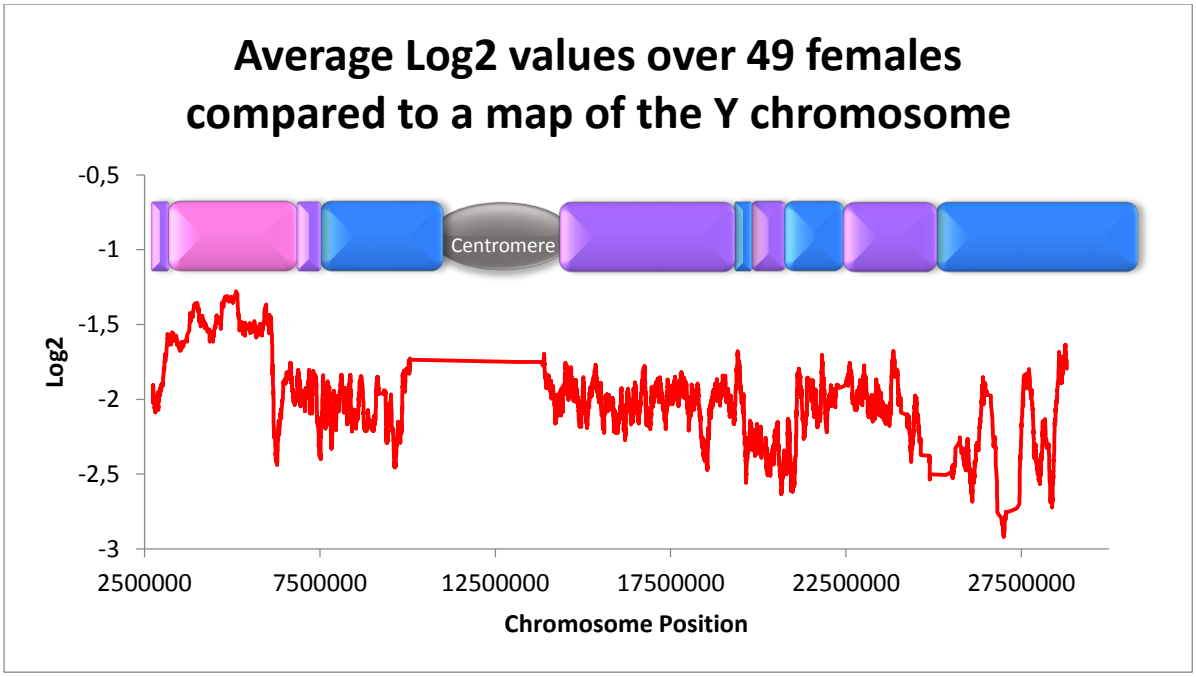

Figure 4: Comparing the map of X-transposed (pink), X-degenerate (purple) and ampliconic (Blue) sequence areas to average female data. The red line is the running average with a window size of 50 markers over the average Log2 values from 49 females. The highest Log2 values are on the X-transposed area. The X-degenerate area on the Q-arm has higher Log2 values than the ampliconic palindrome P4-6.

The SD of the females' Log2 values was calculated from the 49 females. There was a small decrease in SD in the beginning of the P-arm (Figure 5).

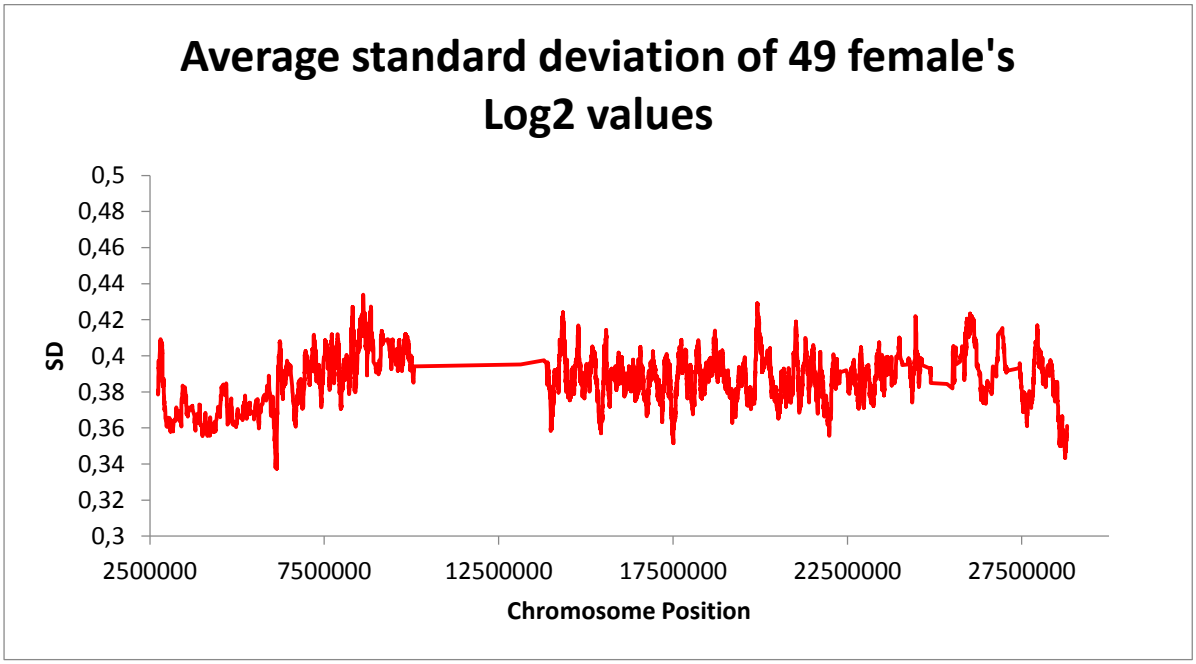

Figure 5: Showing a running average of 50 markers average standard deviation of 49 females Log2 values. There is a slight decrease in SD in the beginning of the P-arm

Detection of CNVs

The Genome-Wide Human SNP array 6.0 has 8179 Y chromosome specific markers; the total number of tries with a sliding window of five markers in a row was then 8175. The limit used in the study was nine percent (chosen approximately to give a final probability under five percent). This gave the probability of $0{,}09^5 = 0{,}0000059$ to find five markers in a row. The total chance over the whole sample was then $8175 * 0{,}0000059 = 0{,}0483 = 4{,}83\%$. Probabilities for other marker counts decreased dramatically, promoting further examinations of these detections (Table 3).

Table 3: Shows the probability to get x number of markers in a row using the limit used in this study (9%).

Probability to get # of markers in a row by chance in one sample with the limit of 9%						
4	5	6	7	8	9	10
53,64%	4,83%	0,43%	0,04%	0,00%	0,00%	0,00%

All samples where scanned for CNVs using the original data and the modified data. The setting used in this study was SD of five, Amplification of 100% and an upper and lower limit of nine percent. The CNVs found were then plotted in a graph where every sample had its own row. The modified data's CNVs were masked by the normal data's CNVs, highlighting only novel CNVs (only found after modification). The bright red and bright blue boxes in appendix, figure 2 were deletions and duplications respectively and were found only after modifying the data. The light blue and red boxes were CNVs found with the normal data. CNVs were found on 178 out of 240 males (Appendix, Figure 2).

Every CNV detected was added to a trend graph. A duplication added 1 to a blue line on its start to stop position. A deletion subtracted 1 to a red line in the same fashion. After all CNVs had been processed, a trend started to appear (Figure 6). There were three large peaks of deletions on the later part of the P-arm. There was also some frequent duplications one third in on the visible Q-arm. The "Deleted in Azoospermia" region (DAZ) which is positioned on the far right of the chromosome was a frequent place for both duplications and deletions.

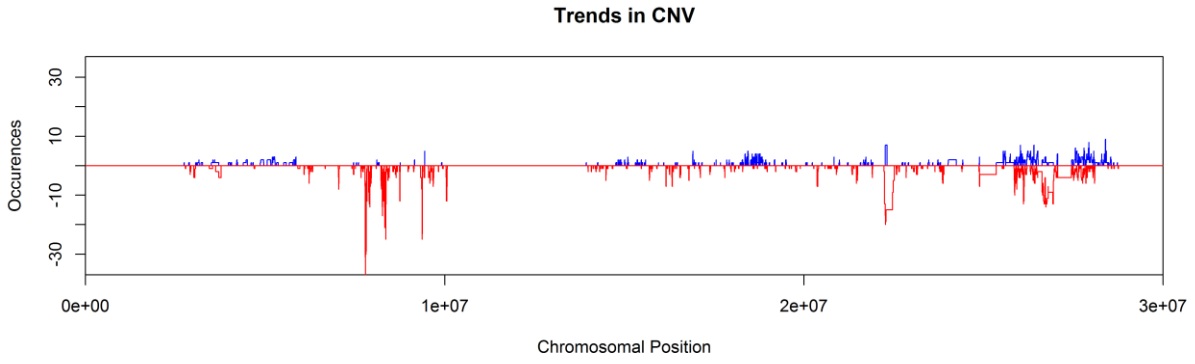

Figure 6: Showing the trend of CNVs in the population. The red line represents deletions and the blue line represents duplications. There is a trend of deletions on the P-arm with 3-5 deletion points with 10-35 occurrences. On the Q-arm, there is a duplication trend with about five occurrences. The DAZ area is a common place for both deletions (10-20 occurrences) and duplications 5-10 occurrences).

The novel deletions on the P-arm were examined for presence of STS primers using UCSC. The STS primers found in the deletions are shown in table 4. Only one deletion covered a protein-coding gene (TGIF2LY). The full gene name was TGF (beta)-induced transcription factor 2-like Y, and it had a homologous gene on the X chromosome (TGIF2LX). The gene was expressed in the testis and it had shown correlation to the size of the testis. (Aarabi et al., 2008) (Skaletsky et al., 2003)

Table 4: Eight small novel deletions in the X-transposed area. Six of them have five markers in a row and two of them have eight markers in a row. The STS IDs in red are Y chromosome specific STSs and the black are specific to both the X and Y chromosome.

Sample No.	Sample	CNV Type	Start	Stop	kbp	Markers	Covering genes	STSs
21	k5128	Deletion	3016489	3044113	27,624	5	N/A	Sy3033, Sy3032
51	u151	Deletion	3016489	3065666	49,177	8	N/A	Sy3033, Sy3032
92	k5519	Deletion	3016489	3044113	27,624	5	N/A	Sy3033, Sy3032
146	k5581	Deletion	3016489	3044113	27,624	5	N/A	Sy3033, Sy3032
26	k5278	Deletion	3443385	3539834	96,449	8	TGIF2LY	Sy1254, Sy3022, Sy872
150	k5524	Deletion	3622708	3777756	155,048	5	N/A	DXYS109, D1S57, Sy3024
46	u537	Deletion	3712712	3778385	65,673	5	N/A	D1S57, Sy3024
70	u904	Deletion	3712712	3778385	65,673	5	N/A	D1S57, Sy3024

Confirmation of deletions

Sample number 21, 92, 146 where used in PCR using the Sy3033 STS primer. They did all produce PCR product and all negative and positive controls were ok. The same with sample number 26 using the Sy3022 STS primer (Table 4).

Discussion

Female data

This study was established on the hypothesis that the X chromosomes and to some extent also the autosomes influence the Y chromosome signals in SNP-array data. When examining the results in figure 4, there is clear evidence of consistent influence by females on the Y chromosome markers. The running average alone show segments of different Log2 values, but when matched with earlier studies of the Y chromosome's relation to the X chromosome, these segments are coherent with that data. They align well with the locations of the different sequence types proposed by Skaletsky et al.. The X-transposed sequences have a sequence similarity of 99% to the X chromosome, and it has the highest Log2 values. The X-degenerate sequences have sequence similarity of about 60-96% but the ampliconic have even less similarity. The effect of this can be seen on the Q-arm where the palindromes correlates to sharp drops of the Log2 values within a block of X-degenerate sequences. It is therefore highly probable that the X-transposed region on the Y chromosome SNP array suffers from inability to detect CNVs.

Modulation

By exponentiating the Log2 values to a linear relation, it was possible to subtract the female signal from the male signal. It is of importance to subtract what the Log2 values are representing (CN states), not the Log2 values them self. It is therefore important to have the same value between every CN state. The following example uses the base of two and illustrates the effect of subtracting data

point four from data point five, and then data point two from data point three, which should give the same result (one), with and without the logarithmized value. See figure 7.

1. Subtracting 3-2 with the logarithmized value: $1{,}58 - 1 = 0{,}58 \rightarrow 2^{0{,}58} = \mathbf{1{,}49}$
2. Subtracting 3-2 without the logarithmized value: $2^{1{,}58} - 2^1 = 3 - 2 = \mathbf{1}$
3. Subtracting 5-4 with the logarithmized value: $2{,}32 - 2 = 0{,}32 \rightarrow 2^{0{,}32} = \mathbf{1{,}25}$
4. Subtracting 5-4 without the logarithmized value: $2^{2{,}32} - 2^2 = 5 - 4 = \mathbf{1}$

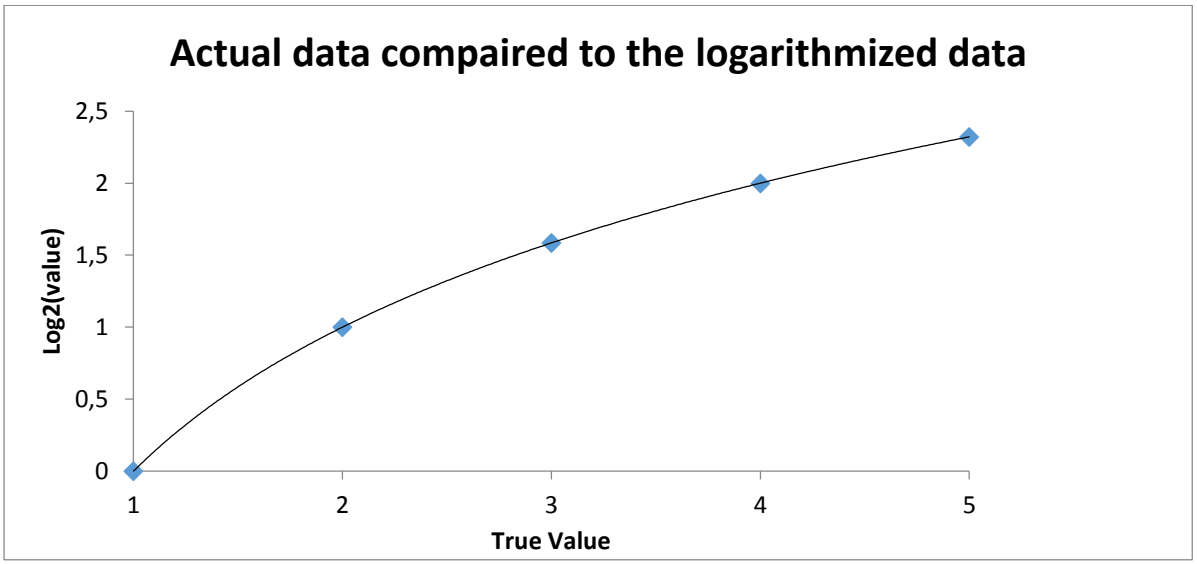

Figure 7: The true value plotted against its logarithmized value using the base of two. The difference between the logarithmized values decreases with a higher true value.

The Log2 values are called so because they are logarithmized in the base of two. So why did the base of two not give the best linear fit? The data were collected in a machine that detected intensity from small patches of probes. One could be tempted to assume that the intensity should be twice as high, if there is twice as much DNA in the sample. However, when one thinks about the probability of finding empty probes when the probes start to fill up in that marker, the linear relation falls apart. We get a saturation of that marker.

Another aspect of the modulation was the use of SD to determine how much female's markers should modulate the male markers. If a marker had a high SD, not much was known about that marker, and to use it to modulate the male marker would make no sense. If a marker had low SD, it could be predicted, and use of this marker is supported by statistics. It is nevertheless, hard to set a number on what is a good and a bad SD. In this study, this factor was part of the user settings. In order to keep down the false positives as much as possible, a strict SD modulation was used. No markers had more than 50% gain, and most of them had around 20%, see lower right corner of figure 1 in appendix. These setting are speculative and encouraged to either play around with or find a mathematical backing for them.

One assumption made in this study was that both the X chromosome and the autosomes contribute to the Y chromosome markers' signals equally throughout the chromosome. There is strong evidence of the influence from the X chromosome in the X-transposed area and the X-degenerate area, but it is unknown how much of this influence is induced by the X chromosomes and not the autosomes. Since females have two X chromosomes and males only one, the signals induced by the females X chromosomes would be twice than what the males' single X chromosome could produce. On the other hand, signals on the ampliconic regions should be affected from autosomes mainly, since there are

autosomal gene homologs in this area, making the males able to produce the same signal as the females. Since the main focus of this study were the X-transposed area, which had a high similarity to the X chromosome, the assumption were made that the signals were mostly from X chromosomes. Because of the males' inability to produce such high signal with their single X chromosome, the maximum gain was set to 50%. If one were to study the ampliconic areas, a gain of 100% could theoretically be applied.

Novel detections

During this study, a number of novel putative deletions where found (Appendix, figure 2). They are spread out all over the chromosome, but with a little clustering in the beginning of the P-arm. This is not so strange, since this is the area with the highest female Log2 values (see figure 4), and the lowest SD values (see figure 5), which lead to considerable modification. There are also, what looks like two big novel deletions and a multitude of large deletions in the DAZ area. These are however artefacts caused by a gap of markers in that location. There is also an artefact covering the whole centromere, which has no markers, on row 127. Only one of these novel deletions was found over a protein-coding gene (Table 4). This gene has a homolog on the X chromosome, and according to studies, an abnormality of the X-version of this gene contribute to azoospermia where as an abnormality of the Y-version gives the phenotype of smaller testis. There are many more novel CNVs than those listed in table 4, but these were not examined any further because of time restrictions.

Four of the deletions have been examined using PCR in an attempt to confirm them. They were unfortunately disproven. It would be interesting to examine more of the deletions, and not only from the X-transposed region, to see if false positives are produced by the hard modification of that area. The next step would be to examine CNV found before any modulation to see if the detection algorithm is the source of the false positives.

Trends in the population

As hypothesized, some trends in the CNV emerged (Figure 6). There are three deletion peaks on the P-arm and one duplication peak on the Q-arm. There are also both duplications and deletions in the DAZ area. The intention was to look further in to these trends, but due to time issues, the data was only summarized. It is hard to tell if this trend originates from a population, or from the array design itself. This could be figured out by examining other populations. If it were a population trend, the two populations would have somewhat different trends. It is not necessary for the CNVs to have different positions, the frequency distributions of both the deletions and the duplications can be used to differentiate between populations. If two populations' trends look the same, the trends are most likely caused by an array error.

Sensitivity

The algorithm that detects CNVs used in this study uses statistical significance. It allows for findings of CNVs far smaller than the method used by GC (Hidden Markov Modell). If this comes with the cost of a higher false positive rate is unknown, but it seems likely. The method only reports duplications or deletions were as HMM reports the CN state (total loss, haploid, diploid triploid etc.)

Errors/improvements

In this study, the MAPD was used as SD instead of using the MAPD/0,96. The effect on the outcome of this study should be that more CNVs where detected than otherwise would have been. So implementing MAPD/0,96 should remove some false positives.

Using excel files to make all the calculations is not the best way to make the experiment reproducible for other scientists. All these calculations could have been integrated in an R package. This would have made it much easier for any scientist to replicate the results and to see all the calculations and statistical analysis done, simply by reading the code. Excel on the other hand is something that many

people are used to work with, and with a graphical user interface and macros, the calculations is only one button away. It is unfortunately as mentioned before much harder to get a grip of what is happening with the data and hence harder to find errors.

Unfinished project goal

Ten weeks was dedicated to this project. The results published in this report are the results from six of these weeks. The remaining four weeks was used to investigate RNA-sequencing data to see expression of genes from both adult and fetal brain tissue in humans. The goal was to automatically get a graph of the RNA-sequencing data covering the CNVs found in the first part of the project. This attempt did however not work out in time.

What's next

Confirming or discarding more of the putative deletions could give useful data. A false positive rate could be estimated with this information. Since there are large amounts of SNP-array data out in the public, many more population studies equivalent to this one could be done.

Acknowledgements

I would like to thank Elena Jazin, who gave me the opportunity to do this study, and Martin Johansson who has supported me with guidance and knowledge throughout the whole process. I would also like to thank Ingrid Agartz and Srdjan Djurovic for the opportunity to use their SNP-array data. Last but not least a big thanks to my organizer David van der Spoel, for valuable feedback on my report.

References

Aarabi, M. et al., 2008. Association of TGIFLX/Y mRNA expression with azoospermia in infertile men. *Molecular reproduction and development*, 75(12), pp.1761–1766.

Affymetrix®, 2008. Median of the Absolute Values of all Pairwise Differences and Quality Control on Affymetrix Genome-Wide Human SNP Array 6.0.

Bachtrog, D., 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature Reviews Genetics*, 14(2), pp.113–124.

Scharpf, R.B. et al., 2008. Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *The annals of applied statistics*, 2(2), p.687.

Siggberg, L. et al., 2012. High-resolution SNP array analysis of patients with developmental disorder and normal array CGH results. *BMC medical genetics*, 13(1), p.84.

Skaletsky, H. et al., 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423(6942), pp.825–837.

Wang, K. & Bucan, M., 2008. Copy number variation detection via high-density SNP genotyping. *Cold Spring Harbor Protocols*, 2008(6), p.pdb–top46.

Appendix

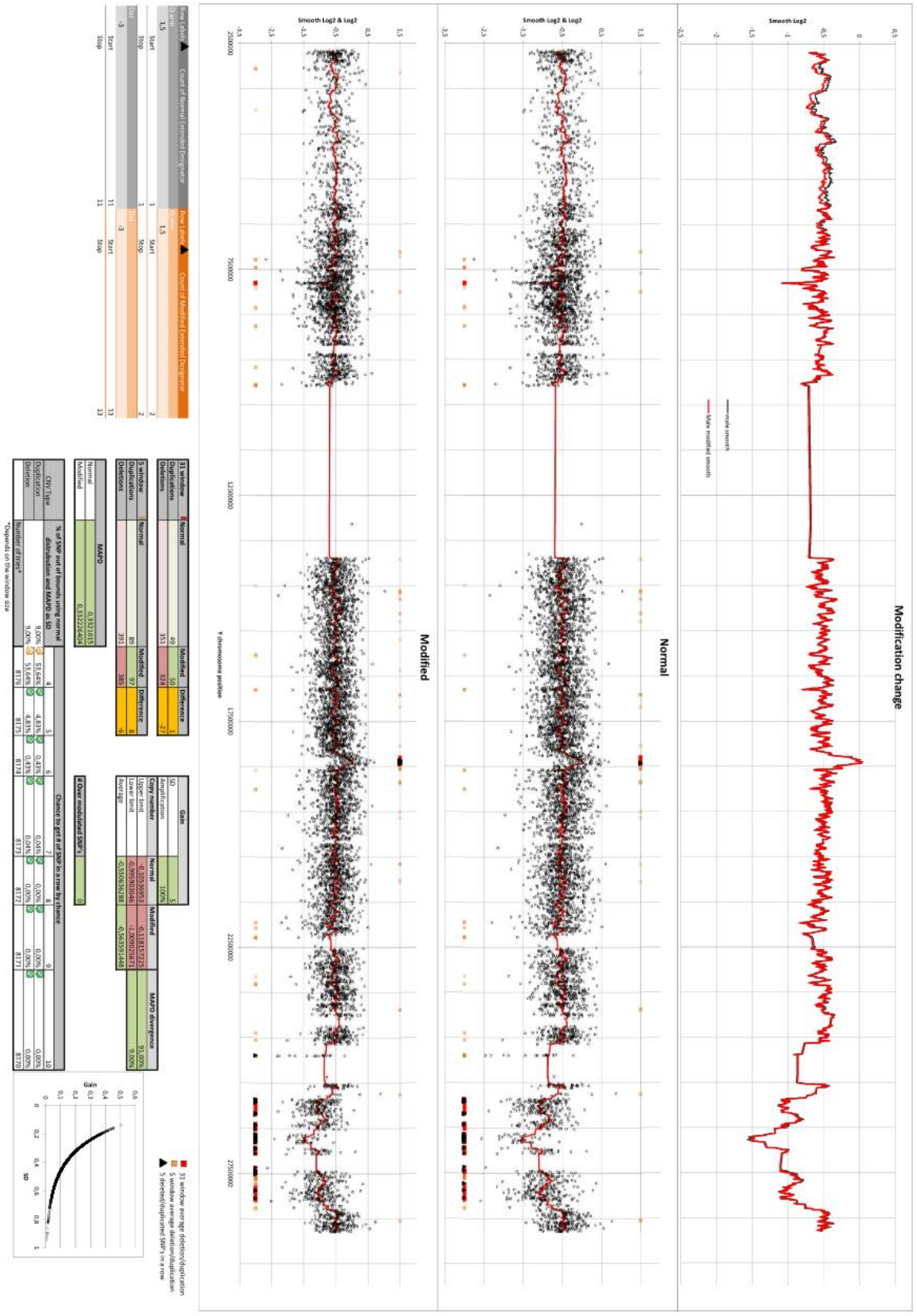

Figure 1: The graphical output showing from top to bottom; a graph with the 31 running mean value of Log2 values, normal data in black and modified data in red. Two graphs, illustrating the actual Log2 values of the markers for the normal and

modified data respectively. The black triangles are markers that are past the upper or lower limit for more than five markers in a row (statistically significant CNVs). Tables at the bottom display information about the sample statistics, and a graph showing how the SD influences the gain.

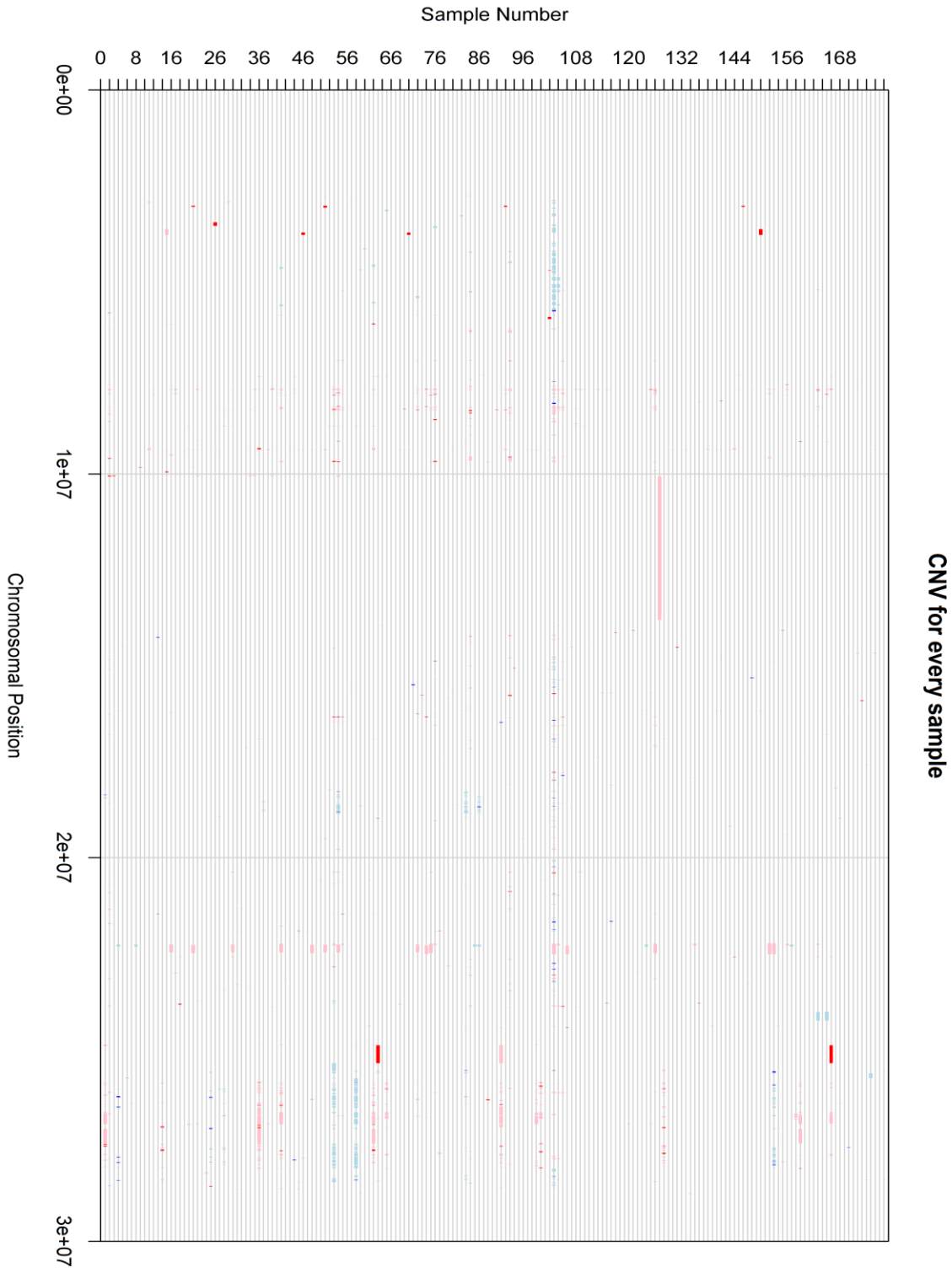

Figure 2: Every CNV found in the study are plotted in this graph. The sample numbers are on the y-axis, and the chromosomal positions of the CNVs are on the x-axis. The modified data's CNVs are plotted in bright red and blue colours (red is deletion and blue is duplication). The modified data's CNVs are masked by the normal data's CNVs, which are plotted in light red and light blue colours.