1 Results

In this section we examine the output of seven linkage projects; five small straightforward of which three are non-problematic, and two with larger more complex pedigrees. Due to the focus of the Nephrology group at the Royal Free Hospital, all linkage projects follow rare disease models, with a primary aim in determining a single disease locus indicative of a causative variant.

Bits	**No. of Genotyped Individuals**	**No. of Markers**	**Model**
3	5	35,155	Autosomal Recessive
5	6	41,479	Autosomal Recessive
7	9	43,421	Autosomal Dominant
9	8	43,421	Autosomal Dominant
15	11	41,480	X-Linked Dominant
18	7	50,110	Autosomal Recessive
21	10	15,914	Autosomal Dominant
23	11	15,914	Autosomal Dominant
29	14	47,595	Autosomal Recessive

Table 1: Summary table of the pedigrees evaluated in ascending order of bit-size

Timings of each pipeline run are considered later in the section, and haplotypes are compared using HaploPainter and HaploHTML5.

1.1 Linkage Projects

The small pedigrees underwent two types of runs: single-core and multi-core. Due to the speed at which both types operate on small pedigrees, each type performed 10 trials each. The larger pedigrees were for the most part single-core due to resource limitations, but parallelizations were undertaken where possible.

We will examine the pipeline by looking at the four main visual components of each analysis: the pedigree, the GRR relationship charts, the Mendelian error plots, and finally the linkage plots. Run times will be displayed at the end of the section.

1.1.1 Small Pedigrees

Here we look at a variety of pedigrees, each with an increasingly larger bit-size within the 19-bit limit that defines the threshold for large "big data" pedigrees. Small pedigrees are run through the pipeline without any runtime modifications required, resulting in a complete set of plots from all linkage programs.

3-Bit Autosomal Recessive

Figure 1 shows a very simple pedigree; depicting one affected male (2053) amongst unaffected male (2052) and female (2054) siblings from two parents (2056 and 2055). All individuals are genotyped; the parents and the affected offspring being the main targets of informative meiosis, and the siblings acting as controls.

The GRR relationship chart shows normal bunching of sib-pairs (red), two close clusters of parent-offspring relations (yellow), and one single unrelated connection (cyan) between the two parents. For small non-consanguineous pedigrees, parents typically exhibit separate clusters of relation to their offspring since allele inheritance is never perfectly even and some offspring will share more alleles with one parent than another.

In larger (usually inbred) pedigrees, these two parental clusters either bunch together due to the parents being more related, or the gap between two outlying parent clusters are filled with other more-related parental clusters to form one large group that is still distinct from other sib-pair and half-sib groups.

The number of markers used for this analysis was 35,155 of which only 9 exhibited Mendelian errors and were filtered out from subsequent analysis. This constitutes 0.025% of the number of markers and is of no cause for concern, allowing the rest of the analysis to progress.

Due to the low bit-size, the estimated lod score for the pedigree gave a maximum score of 0.2499 (Allegro and Genehunter), which is reflected in the genome-wide plot. Unfortunately most regions of the genome reached this score (approx >~ 80%) likely due to there not being enough lack of homology between the affected and unaffected siblings. The score is also far below the minimum LOD score for informative linkage (LOD > 2) meaning that the analysis was likely not very useful in identifying or excluding regions of interest for a causative disease locus.

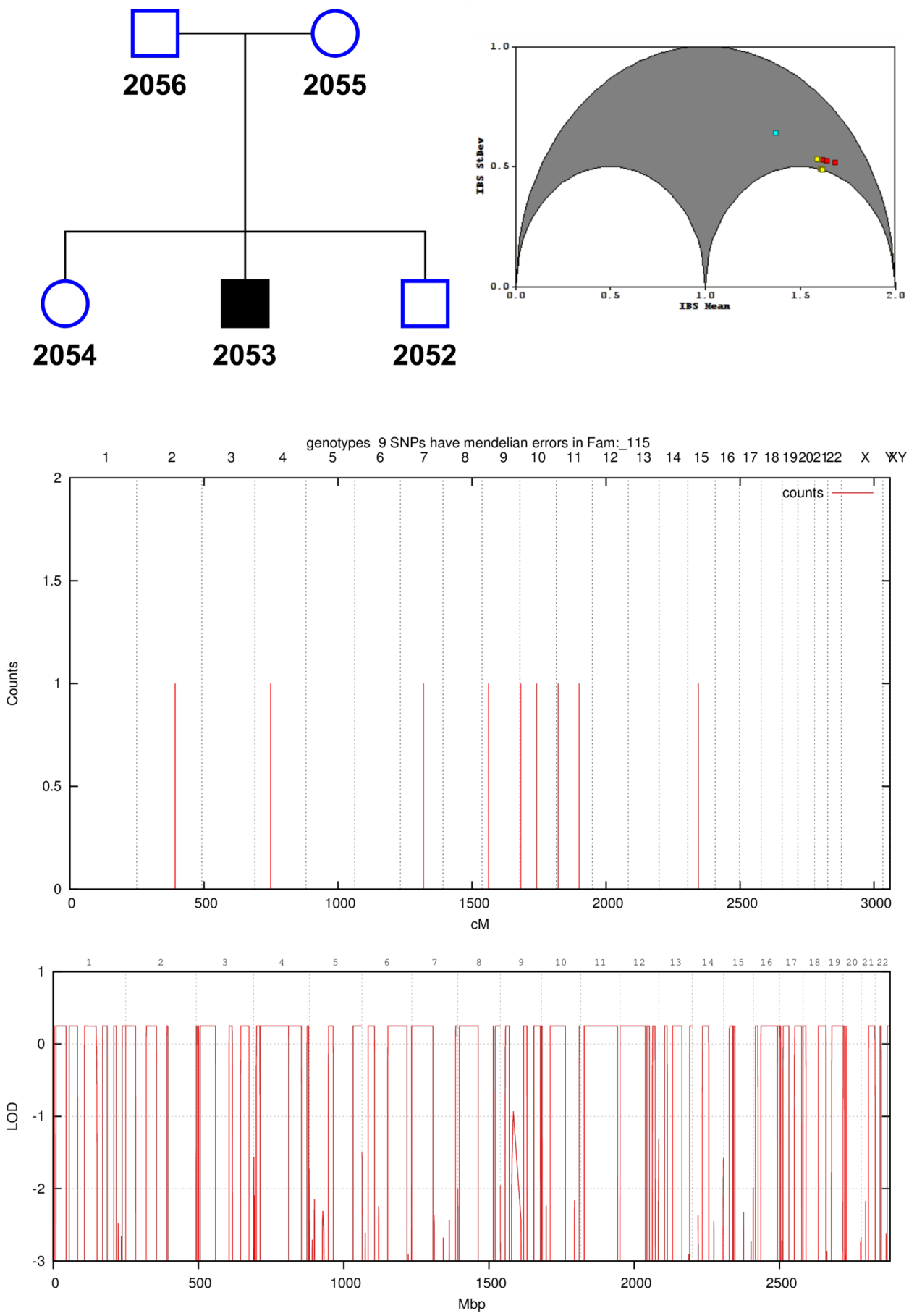


Figure 1: 3-bit Autosomal Recessive Pedigree. (Top) Family Tree and GRR, (Middle) Mendelian errors, (Bottom) Genomewide linkage plot.

5-Bit Autosomal Recessive

This pedigree is very similar to the 3-bit pedigree above, but the bit-size increases by 2-bits for each non-founder added, and here we have one extra affected offspring. The GRR relationship diagram in Figure 2 also shows a similar pattern of groups as before, but this time there are more parental clusters which exhibit the faint beginnings of a larger parental cluster.

55 SNPs (of 41,479 markers) were identified having Mendelian errors, with a cluster of 6 towards the telomeric p-arm of chromosome 8. Removing these from the analysis also had negligible impact (0.133%).

The estimated LOD score was 0.8519 (Allegro and Genehunter) which was once again reached by the linkage with a smaller number of peaks to indicate locus of interest. Though a single causative peak could not be determined, wide regions of exclusion defined chromosomes 2, 3, 6, and 7, as well much narrower peaks genome-wide.

We can also see a few distinct types of peaks: the nicely squared “flat” plateau peaks, the “rounded” bullet-like peaks, and the sharp “shard” peaks. Flat peaks tend to be of more worth to the analysis because they indicate that there are *at minimum* three markers exhibiting the same LOD score to constitute a left-wall, a right-wall and a point in-between.

Rounded peaks show discord or a lack of resolution within the linkage region. The small peak at chromosome 7 (LOD=-0.1) shows that *at maximum* there there are three markers constituting the peak with the left and right walls in rough agreement LOD-score wise, but that a single marker in-between them is significantly higher, forcing the plotting program (GNUplot) to smooth the peak into a rounded shape. These types of peaks are extremely common in low-marker analyses (< 5000 SNPs) and are rare in high-marker analyses (> 50,000 SNPs). Shards convey the same level of information, with the further disadvantage of only being constituted by two markers, both with unequal LOD scores.

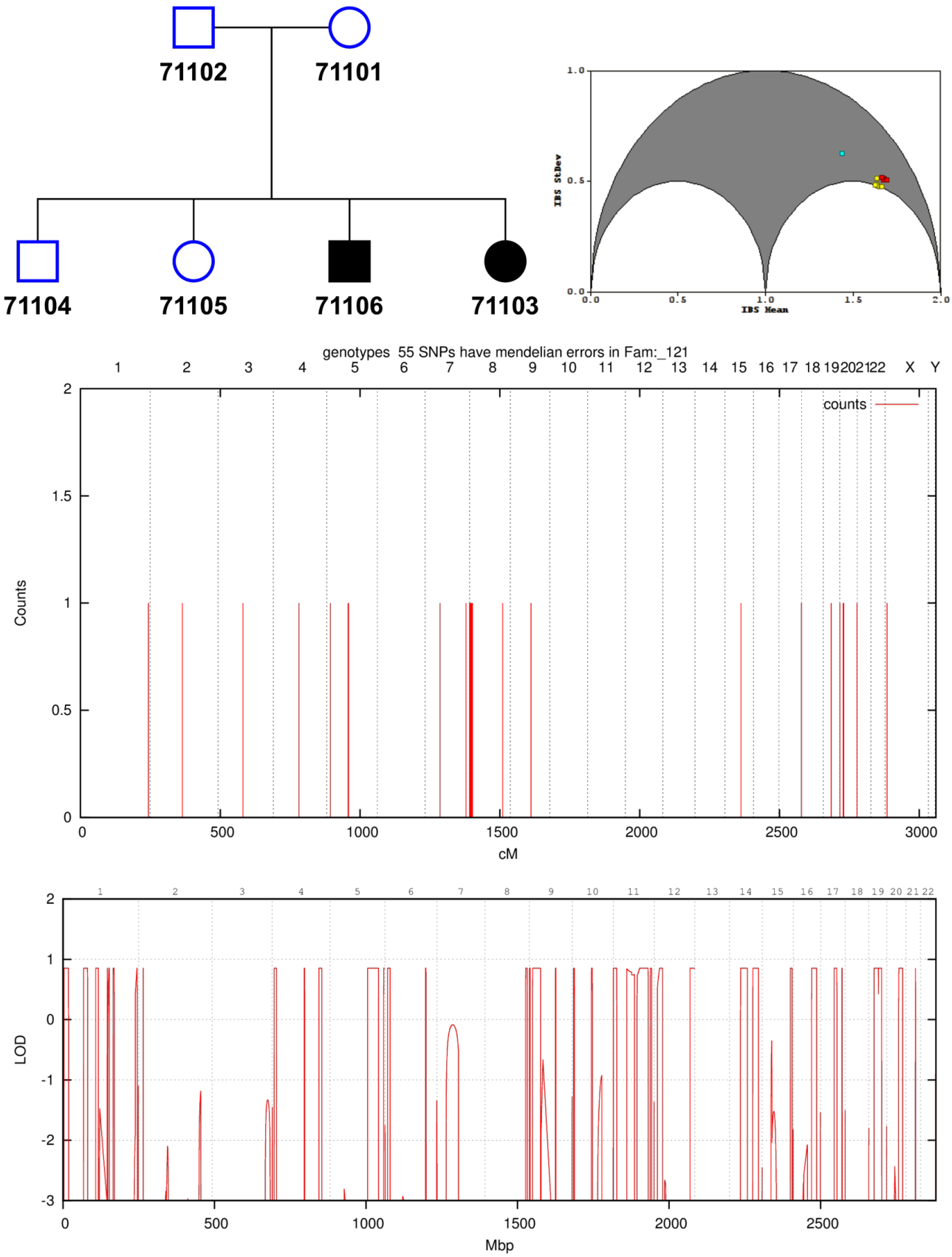


Figure 2: 5-bit Autosomal Recessive Pedigree (119)

7-bit and 9-bit Autosomal Dominant

In this analysis we see examples of excessive Mendelian errors, as well as the trial-and-error process of trying to determine the 'true' model of a given data set through successive run iterations.

Figure 3 shows ten pedigree variants from the same initial base, each constructed to pinpoint the source of the high Mendelian errors shown in Figure 4 for the base scenario; 8000 out of 43,421 SNPs (18.4%) . The GRR plot for the run (Figure 5, 01) shows inconsistent bunching between sib-pair and parent-offspring groups with some mixing, clearly showing that something was indeed wrong with the pedigree.

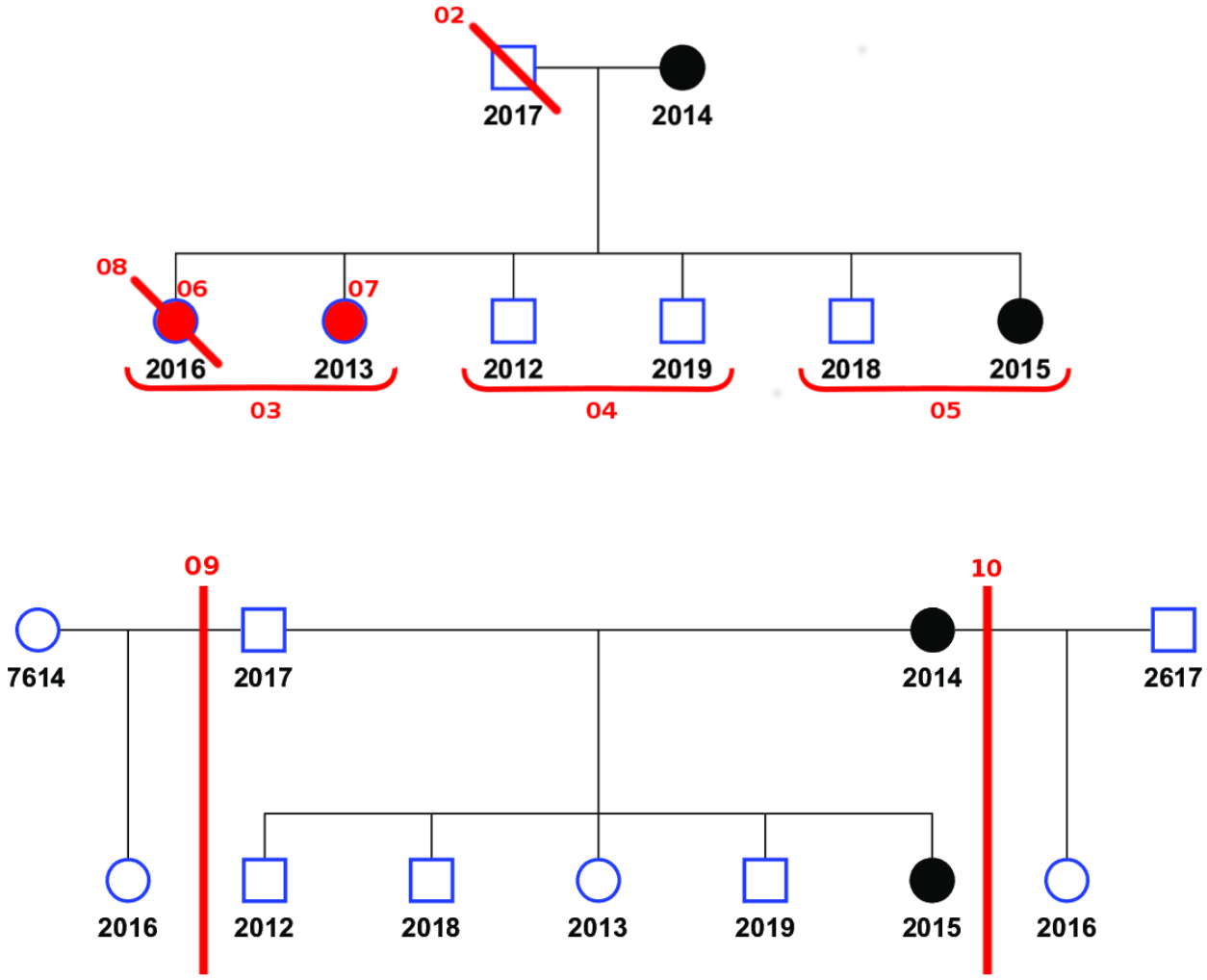


Figure 3: Autosomal Dominant Pedigree split into 10 scenarios. The base scenario (Top) exists without any red-line modifications. Individuals with a diagonal red line crossing through them denote a scenario (numbered in red) where that individual was removed from the pedigree. Individuals filled in red denote analyses where only that individual was considered in relation to their parents with all other offspring omitted. Red underlined groups denote parent-sibling analyses where siblings are grouped and evaluated separately. The bottom pedigree is the same as the top base scenario, but with alternate mother and father for individual 2016 making it half-sib with other offspring.

The second scenario (run 02) assumed that the majority of errors stemmed from the father (2017) since genetic testing often reveals bad paternity 10% of the time (XXXREF). The father was truncated from the analysis to be replaced by an ungenotyped placeholder parent, which halved the number of errors to 3433 SNPs but also proved that the father was not the main source of the errors,

hinting that one of the offspring was responsible. The GRR plot for the scenario (02) reflects this, for there is no longer any mixing between sib-pair and parent-offspring clusters, though the two extremely disparate groups for each relation still remains.

To pinpoint the offspring individual at fault with a reasonably small number of re-runs, offspring were split into pairs and run with their original parents in independent analyses: (run 03) 16 and 13, (run 04) 12 and 19, (run 05) 18 and 15. The analyses were terminated prematurely before the linkage stage in order to compare their relative errors without bias.

Run 03 reproduced over 99% of the original base errors with 7991 SNPs with Mendelian errors, with run 04 and run 05 producing 55 erroneous SNPs between them. The GRR plots also represented this grouping, with the runs 04 and runs 05 showing the classic close groupings of parent-offspring relations orbiting around the main sib-pair cluster, compared to run 03 which spread the parent-offspring relations in a more diverse fashion[1].

This clearly implicated 16 and 13 as the source of the errors, and the next two runs attempted to resolve this by splitting the pair and examining them separately: (run 06) 16 only, and (run 07) 13 only. Figure 4 shows that the errors between runs 06 and 07 clearly implicate individual 16 as the source of the errors, since the Mendelian errors once again rises to a high number of 7980 SNPs. GRR confirmed this by producing a cleaner parent-offspring grouping cluster for individual 13.

The next analysis (run 08) omitted 16 from the base analysis, preserving the original parents and offspring. This resulted in only 37 SNPs (0.0852%) with Mendelian errors across the entire analysis. GRR also showed better group clustering for sib-pair and parent-offspring relations, with much less disparity within each group, and no mixing.

Due to the more complete nature of this run, a genomewide linkage plot was also produced (Figure 6 (top)) showing distinct peaks in chromosomes 4, 7, 8, 9, 12, 13, and 17, 20, and 22 each with a LOD score of 1.25. The estimated LOD gave an expected maximum of 1.52 (Allegro and Genehunter) which was not reached, and even upon success would not have been informative for linkage.

1 Indeed, the relation between mother (14) and child (16) was scored to be the same as the unrelated relationship between the two parents (14 and 17)

Reconfirmation upon the sample data prompted a two more genotype sets to be introduced for a mother and a father of individual 16. Runs 09 and 10 denote alternate testing of each new parent in conjunction with an existing parent in order keep a half-sib relation between 16 and the other offspring.

This created some disarray in the GRR plots for run 09 and 10, and even some relation mixing for run 10. The number of Mendelian errors also rose to significant levels (1849 and 2978 SNPs respectively) though the linkage peaks remained relatively unchanged.

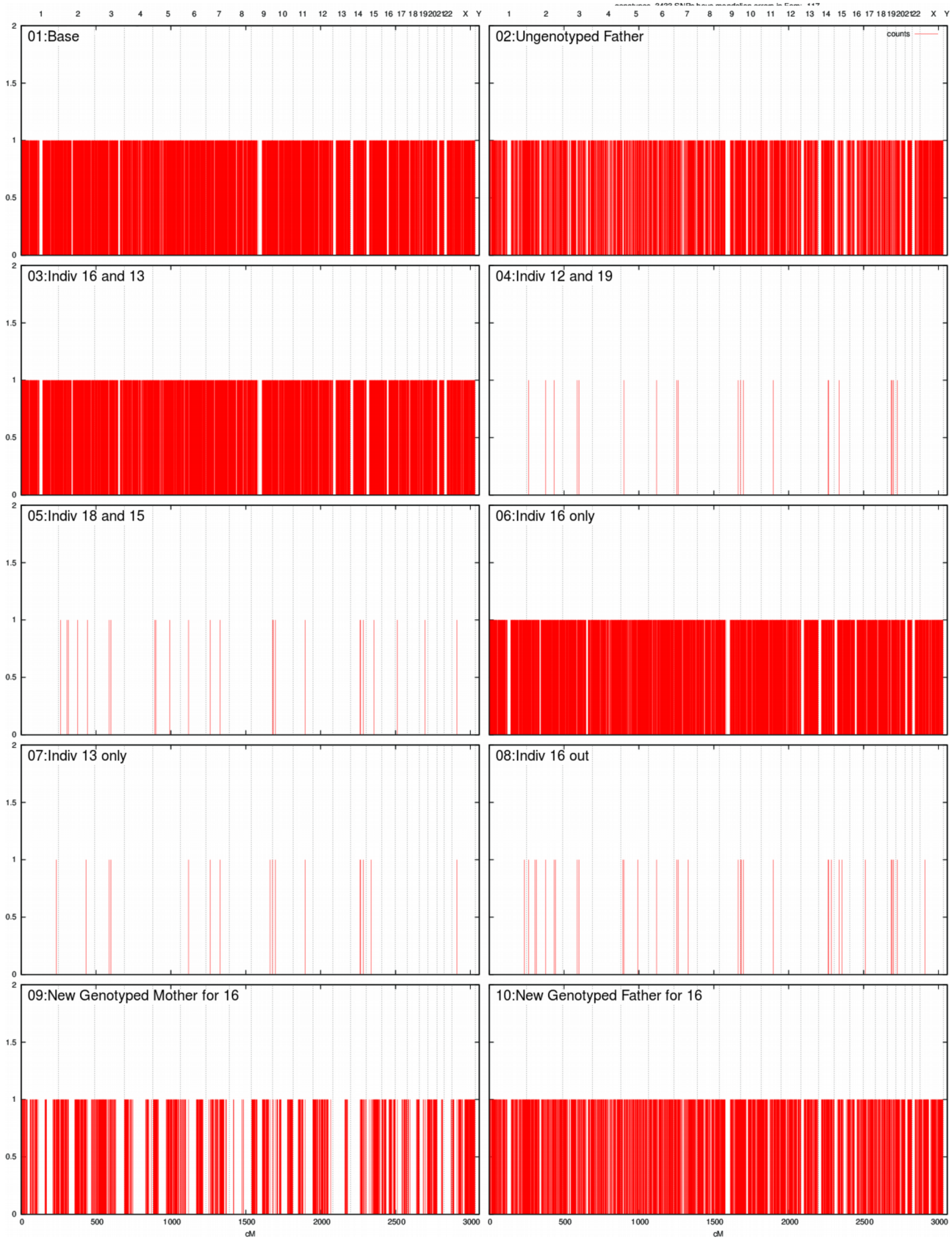


Figure 4: Mendelian errors for the ten scenarios depicted in Figure 3. Each of the scenarios represent the following number of errors: 01 (all base) 8000 SNPs, 02 (father excluded) 3433 SNPs, 03 (16+13 only) 7991 SNPs, 04 (12 +19 only) 24 SNPs, 05 (18+15 only) 26 SNPs, 06 (16 only) 7980 SNPs, 07 (13 only) 18 SNPs, 08 (16 out) 37 SNPs, 09 (new mother for 16) 1849 SNPs, 10 (new father for 16) 2978 SNPs

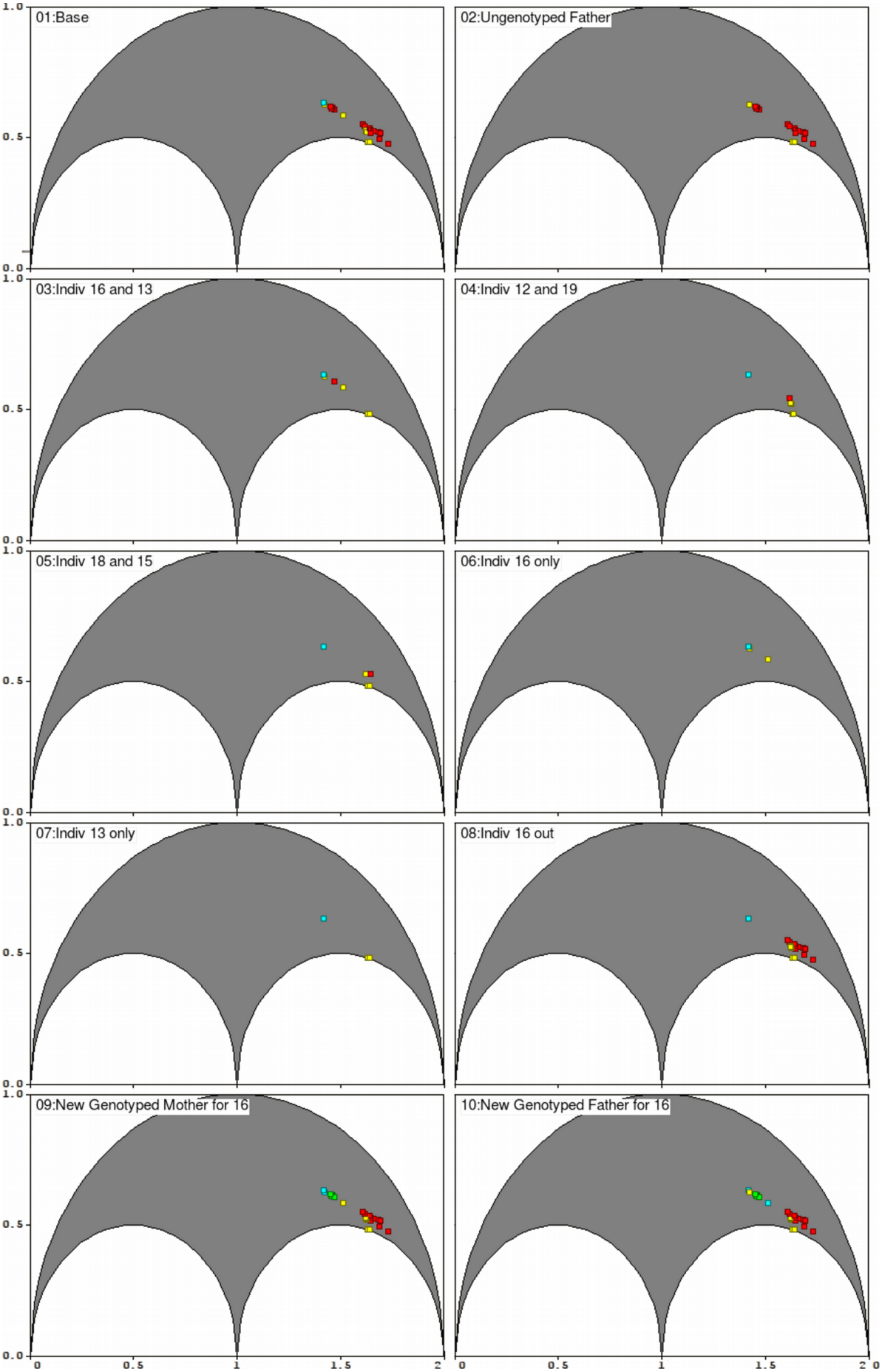


Figure 5: GRR plots for the ten scenarios. Identity-by-State (IBS) is plotted as a function of Mean against Standard Deviation and bound within an arced range of possible values.

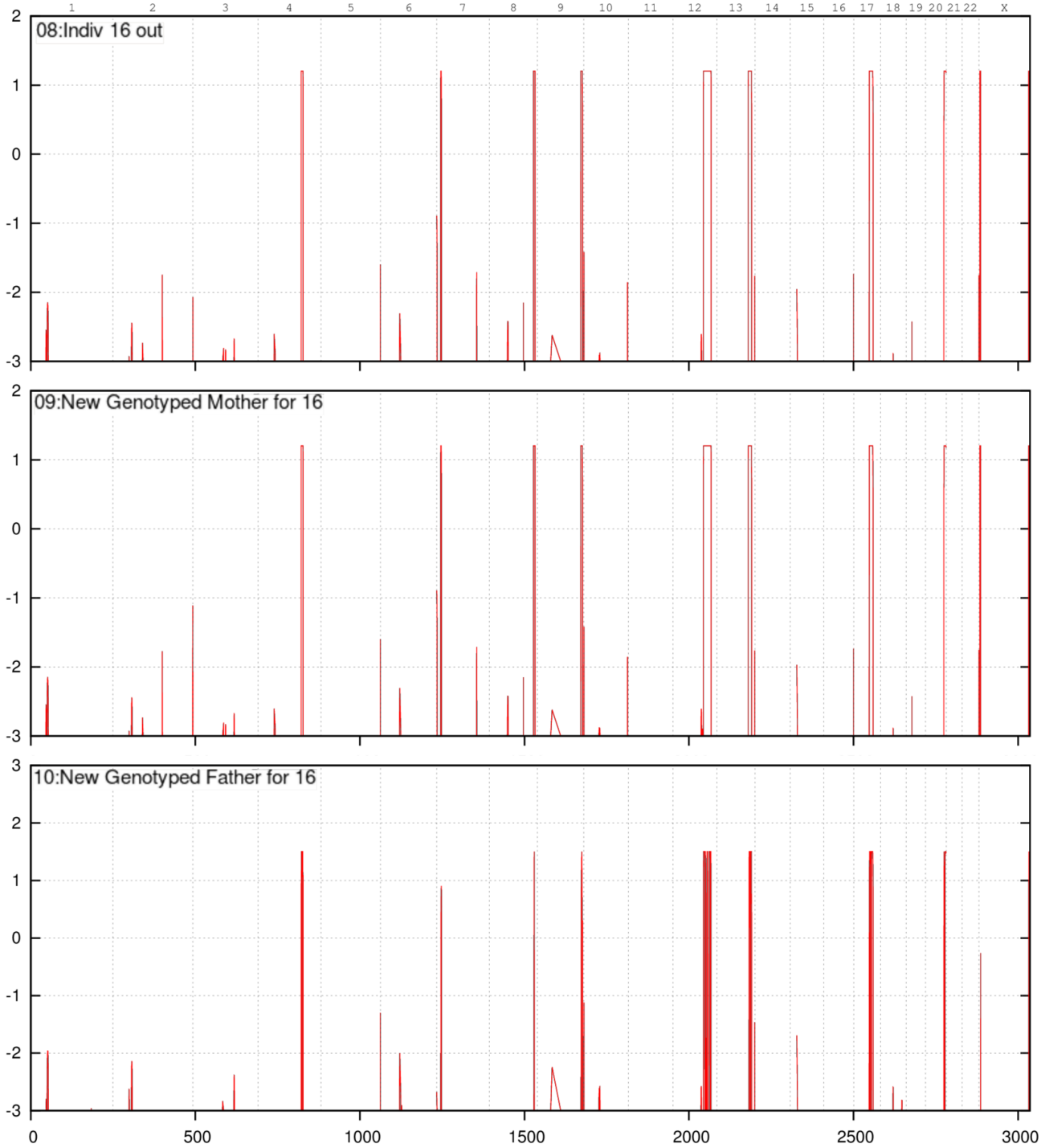


Figure 6: Genomewide linkage plots for three of the ten scenarios that did not produce Mendelian errors: (08) with individual 16 omitted, (09) with a new genotyped mother for individual 16, and (10) with a new genotyped father for individual 16

15-bit X-linked Dominant

Figure 7 below shows a larger pedigree, with three individuals of unknown affectation status. The penetrance model is dominant because an affected individual resides at each generation, and it is suspected to be X-linked because of the absence of male-male transmission though that is not to say that it is not autosomal.

Normally when a penetrance model is fully described in a pedigree, then individuals of unknown affectation are pre-emptively set to either affected or unaffected in order to better accommodate the model. Here the phenotype is not fully penetrant, so ambiguity was purposefully left for the linkage pipeline to interpret.

Despite autosomal dominant and X-linked dominant not being biologically compatible models, computationally they are processed in the same run because each chromosome is treated as fully independent of each other. The linkage pipeline treats X-linked models as "special cases" in conjunction with the standard dominant and recessive runs. It is for this reason that the X chromosome is included with autosomal chromosomes in the genome-wide plots.

The low 162 Mendelian errors are usually negligible for marker set of 41,480 SNPs (0.391%), but the close bunching of 52 SNPs in the q-arm of chromosome X raises some concern on possible genotyping errors. This is later reflected in the genome-wide linkage plot which consists of extremely "noisy" peaks: sharp and intermittent, consisting of no more than two markers before the signal is lost and drops below significance, only to resurrect again within close proximity of the previous signal.

The size and frequency of the peaks indicate at least one of four issues:

- Large numbers of recombinations exist within the three generations of the pedigree.
- The family was badly genotyped with errors stemming from the genotyping process.
- The pedigree structure is not correct and the family model is more complex
- The penetrance model is incorrect

Many of the recombinations occur within extremely small proximity of one another, and since physical distance scales linearly with centiMorgans, a large number of meioses in those regions are extremely unlikely. Similarly, the family could not have been badly genotyped nor could the pedigree structure be incorrect, since the number of Mendelian errors are relatively low.

This defers us to the conclusion that the penetrance model is incorrect, with the affectation of certain individuals (specifically those not already set to ‘unknown’) not being correctly determined by the physician, possibly due to a late-onset phenotype.

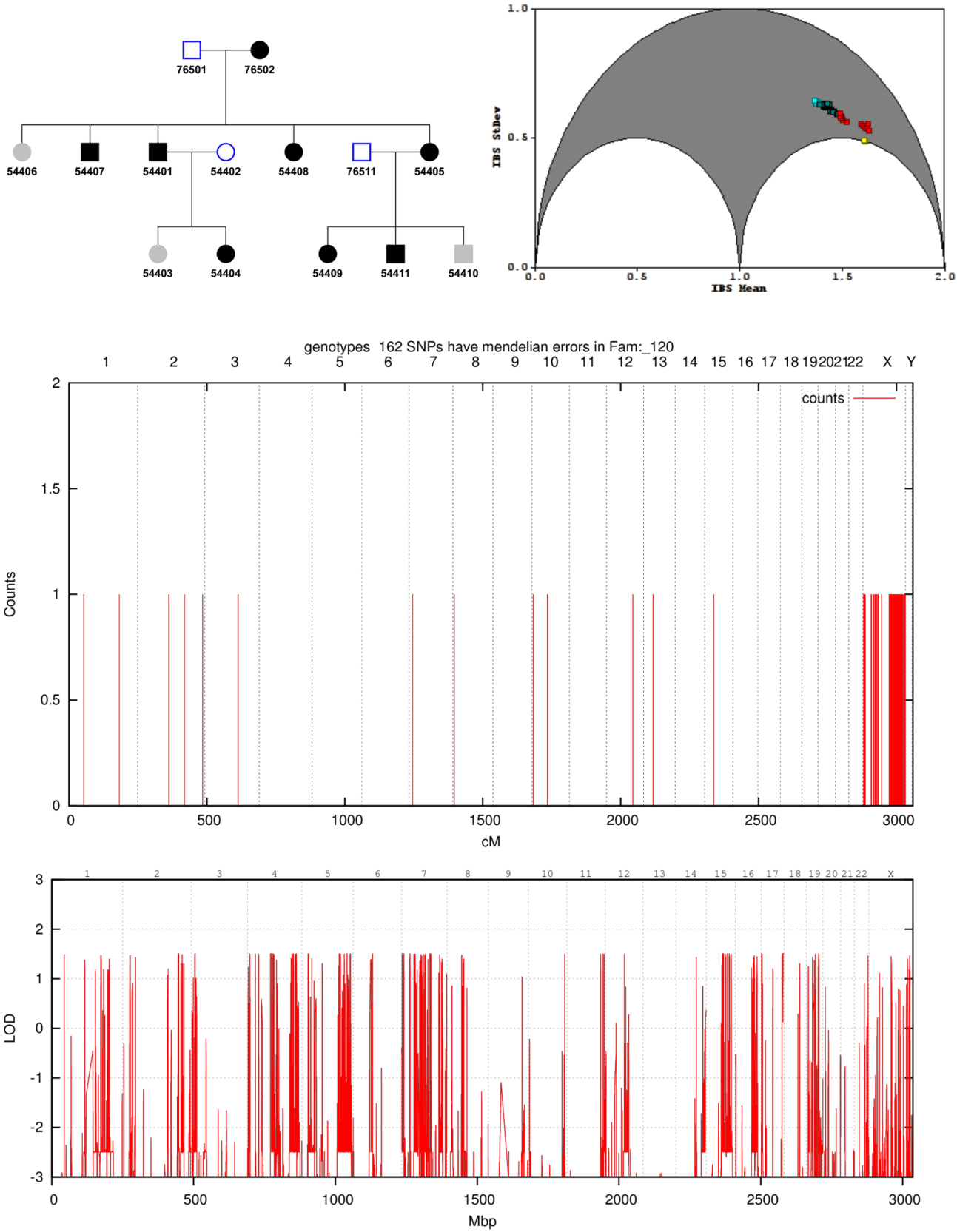


Figure 7: X-linked Dominant pedigree, with three individuals of unknown affectation marked in grey (54406, 54403, 54410)

Further probing of the pedigree under different models (autosomal/X-linked recessive) would be required under various combinations of individual affectation in order to stabilize the linkage peaks and determine the true model behind the data.

18-bit Autosomal Recessive Pedigree

Figure 8 shows a family with an inbreeding loop stemming primarily from the founders on the right (101 and 102) that results in a second-cousin consanguineous pairing (401 and 402).

Occasionally members of consanguineous families have a tendency of obscuring such relations, but the members of this family are fully described with their GRR relationship plot clustering in a good distinct groupings. The lack of extra relations in the plot is due to the only 7 of the individuals being genotyped (401, 402, 404, 403, 501, 502, 503).

Of the 50,110 SNPs in the input marker set, 22 negligible Mendelian errors were reported (0.439%). The preliminary LOD-score estimates provided a score of 2.9, which was met accordingly in the actual linkage output with peaks at chromosome 2, 5, and 7. Figure 9 below shows the zoomed in plot of chromosome 7, with chromosomal bands (and sub-bands) overlayed with a centromere. The darkness of the band is indicative of the density of the chromatin, hinting at regions of heterochromatin (dark) and euchromatin (light) which provide cues about gene expression. A single characteristic flat peak spans a region of 30 Mbp along the q-arm, followed by sharp recombination peak artefacts.

The peaks at the other chromosomes also display valid linkage, and it is up to the researcher to examine all valid regions through further experimental sequencing analysis in order to truly pinpoint the causative gene/mutation.

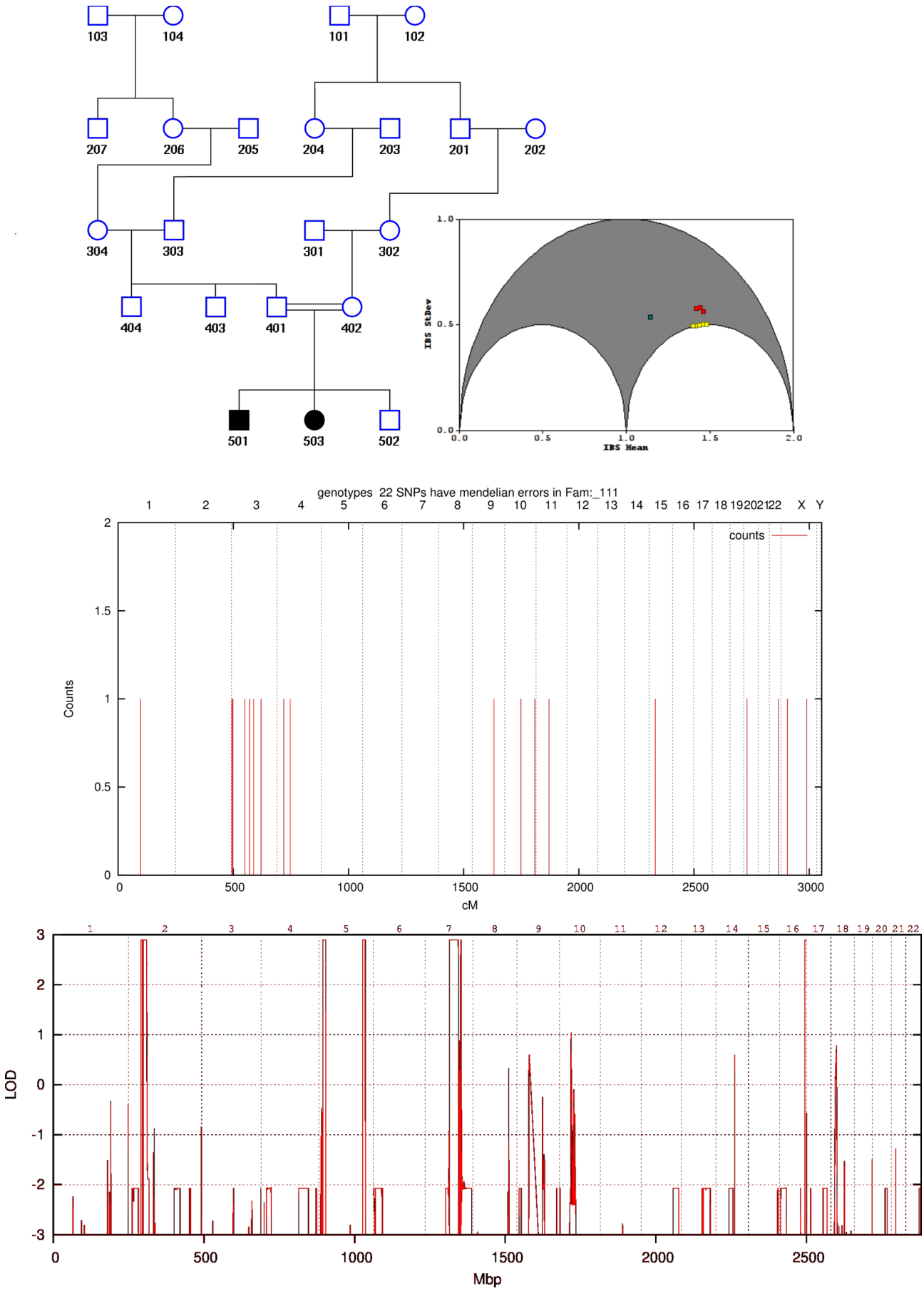


Figure 8: 18-bit Autosomal Recessive pedigree

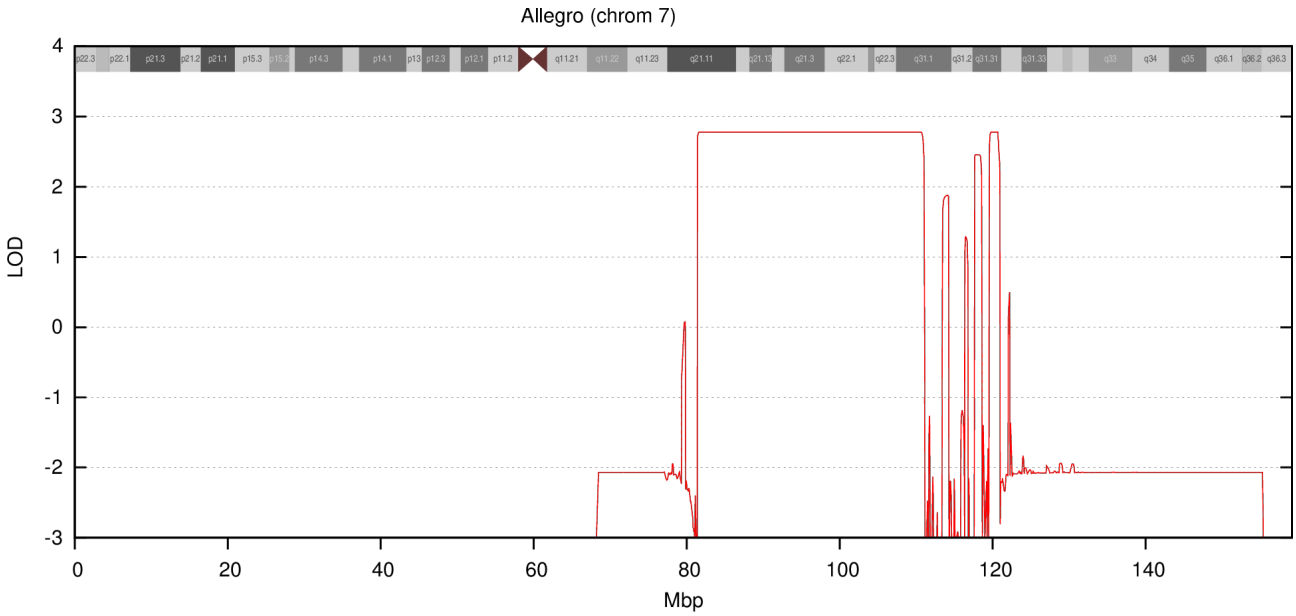


Figure 9: 18-bit Autosomal Recessive Pedigree, chromosome 7 (allegro)

1.1.2 Large Pedigrees

The pedigrees evaluated here required the big data patching scripts mentioned previously in the Methods to operate; the default Allegro and GeneHunter binaries failed due to the combined size of the marker set and the pedigree complexity, requiring the modified Allegro binary to operate as well as some trials with Simwalk.

21-bit and 23-bit Autosomal Recessive Pedigree

Here we examine another large consanguineous pedigree with two inbreeding loops occurring both within the second generation (7621 and 7622, 7623 and 7624) from the two founder couples (7611 and 7612, 7613 and 7614) with seven variations upon the inclusion of unaffected individuals spawned from a base analysis where only affected individuals were included. The base analysis has a 21-bit pedigree complexity, and as per equation <XXXREF> on page <XXXREF>, each non-founder contributes 2-bits which places all subsequent analyses at 23-bits.

Each of the 8 total scenarios contributed negligible Mendelian errors (in total less than 10 SNPs of the 15,914 in the marker set), with an example chart shown in Figure 11 for completion[§]. A total of 18 genotyped individuals contributed to the pedigree (25 individuals) providing a sizeable cluster of

§ Full scores and charts for all 8 scenarios are provided in the Appendix (page <XXXREF>)

'other' relationships in the GRR plot (Figure 12) for scenario 02 (individual 19[2] omitted). No significant changes in the GRR plots for the other scenarios were observed.

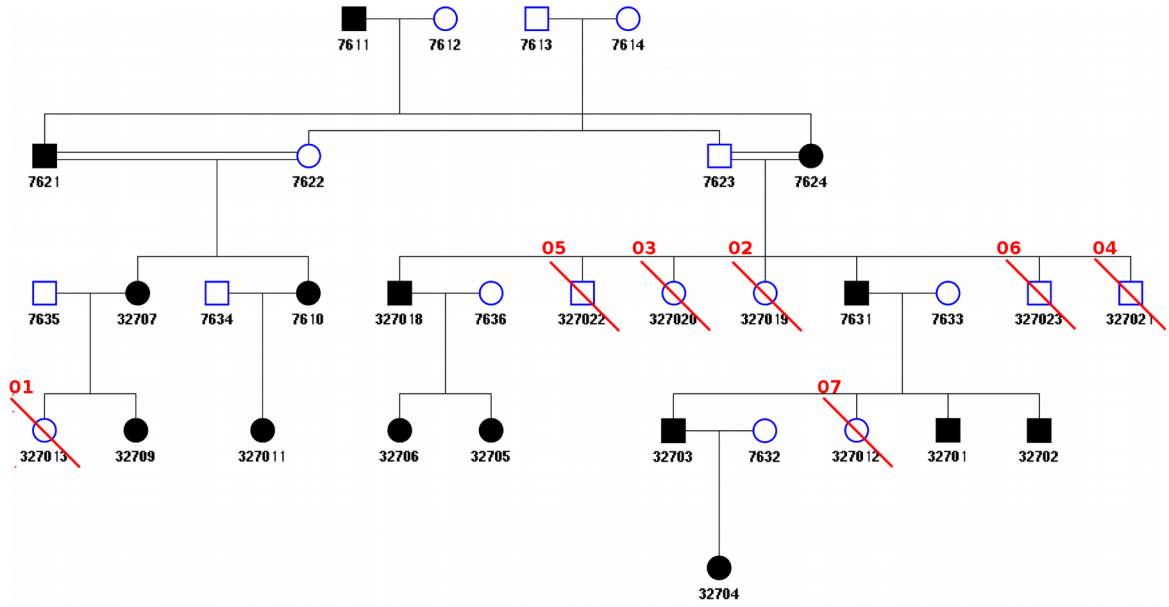


Figure 10: 23 bit pedigree with red-lines indicating the individuals omitted from a given analysis.

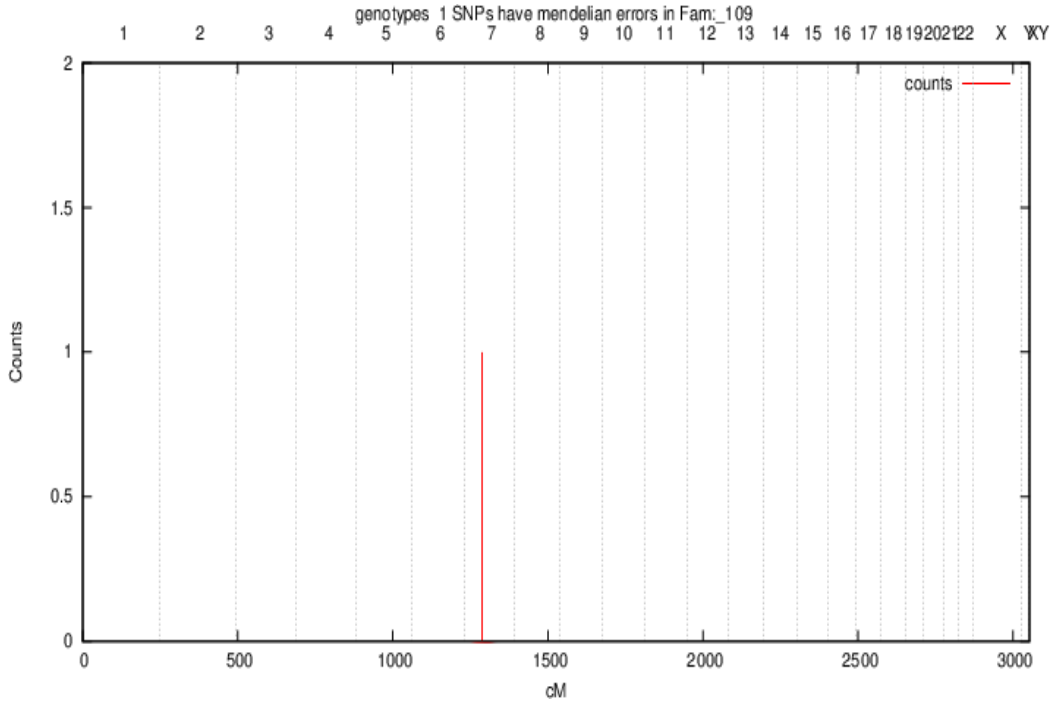


Figure 11: Mendelian error for the base scenario (21-bit). All other analyses gave either 0 or at maximum 3 erroneous SNPs

2 Individual 19 = 327019, but we have removed the '3270' prefix for readability.

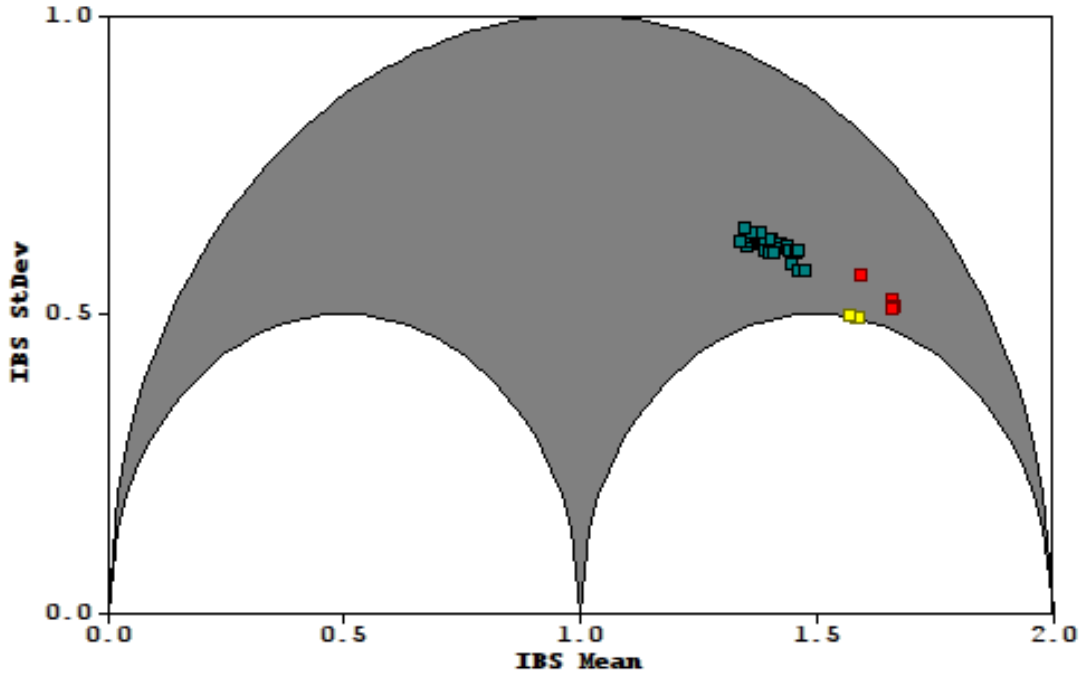


Figure 12: GRR relationship chart of a 23-bit pedigree for scenario 02 (individual 327019 omitted).

The maximum LOD score for the base scenario (21-bit) was estimated to 3.31, and the estimate LODs for the other (23-bit) analyses gave an average of 3.62[§].

The independent addition of each of the unaffected individuals for the seven non-base scenarios (individuals: 13, 19, 20, 21, 22, 23, and 12) created small but noticeable differences in linkage plots (Figure 13). The base analysis (scenario 0) shows peaks at chromosomes 3, 4 and 11 all reaching the maximum estimated LOD score (3.35) with the peak at chromosome 4 being the broadest. A zoomed in plot of chromosome 4 for that analysis reveals a slight drop in the peak at 45 Mbp (Figure 14, top-left). The peaks at chromosomes 3 and 11 are equally viable linkage peaks at this point[3].

By looking at each linkage result evaluated, we can groups the plots into three result types:

3 Plots for all chromosomes can be found in the Appendix (page <XXXREF>)

1. With the independent additions of individuals 13 (scenario 1), 22 (scenario 5), and 12 (scenario 7), the peak at chromosome 4 increases to the new maximum (3.62) whilst the slight drop disappears, and the peaks at chromosomes 3 and 11 decrease.
2. The independent addition of individuals 19 (scenario 2), 20 (scenario 3) cause the chromosome 4 peak to drop as well as the chromosome 11 peak, leaving a fractured peak at chromosome 3 that raises doubt on the informativeness of the peak.
3. Individuals 21 (scenario 4) and 23 (scenario 6) preserve the peaks at chromosomes 3 and 4, but greatly fractures them both too, hinting at an incompatibility.

The width of the chromosome 4 peak (26.8 Mbp) does not narrow with the addition of individuals 13, 22, and 22 as in the first result type, suggesting that the disease locus is within a reasonably conserved locus across the pedigree. A higher resolution trial would likely be required to narrow the region.

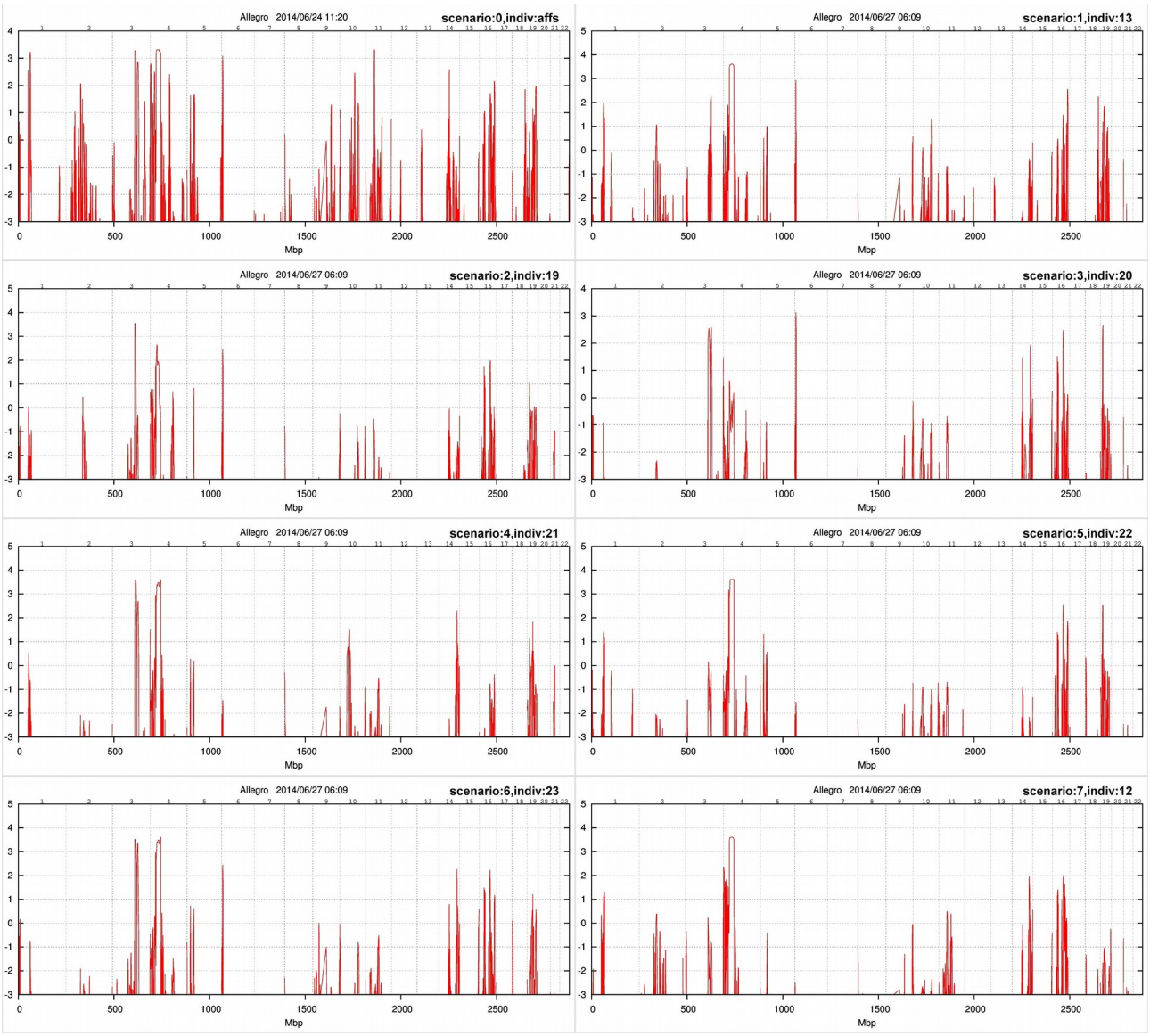


Figure 13: Genomewide linkage plots for each of the eight scenarios for the 21 to 23-bit autosomal dominant pedigree

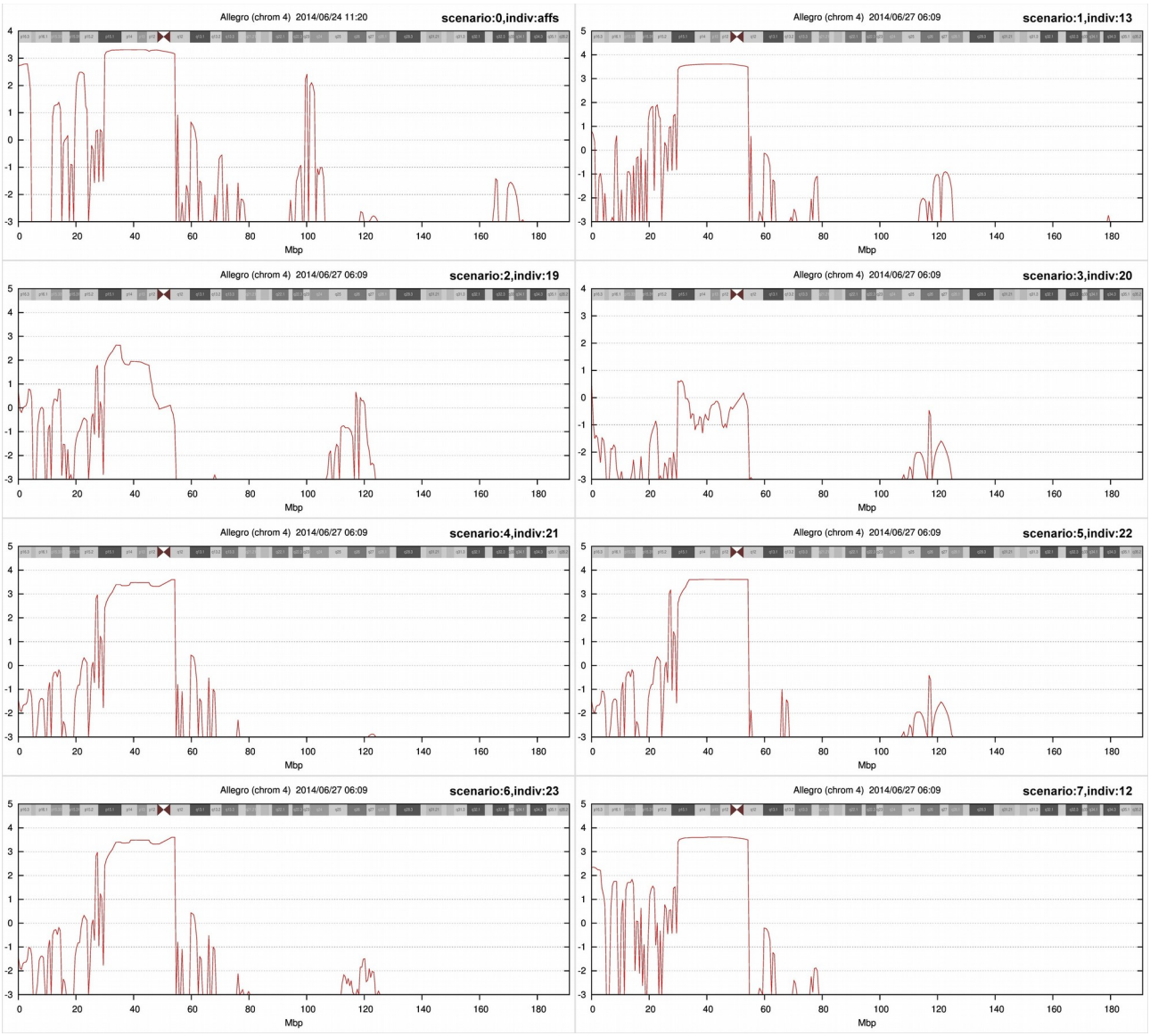


Figure 14: Linkage plots of chromosome 4 for each of the eight analyses

29-bit Autosomal Recessive

This is the largest family ever considered by the pipeline. Figure 15 shows Consisting of three inbreeding loops that span four generations stemming from the same founder couple, where all individuals in the last generation were genotyped along with their parents. Parent-offspring genotyping is always more favoured than grandparent-offspring genotyping because it leaves little ambiguity in tracing the path of inheritance of the disease locus, and makes detecting Mendelian errors much more effective.

Three pedigrees were used in this analysis, the first being the 29-bit pedigree used to lend power to the linkage study and the latter two (Figure 16) not being informative for linkage at all but aided in

the haplotype reconstruction process to reconstitute the genotypes of grandparents and great-grandparents.

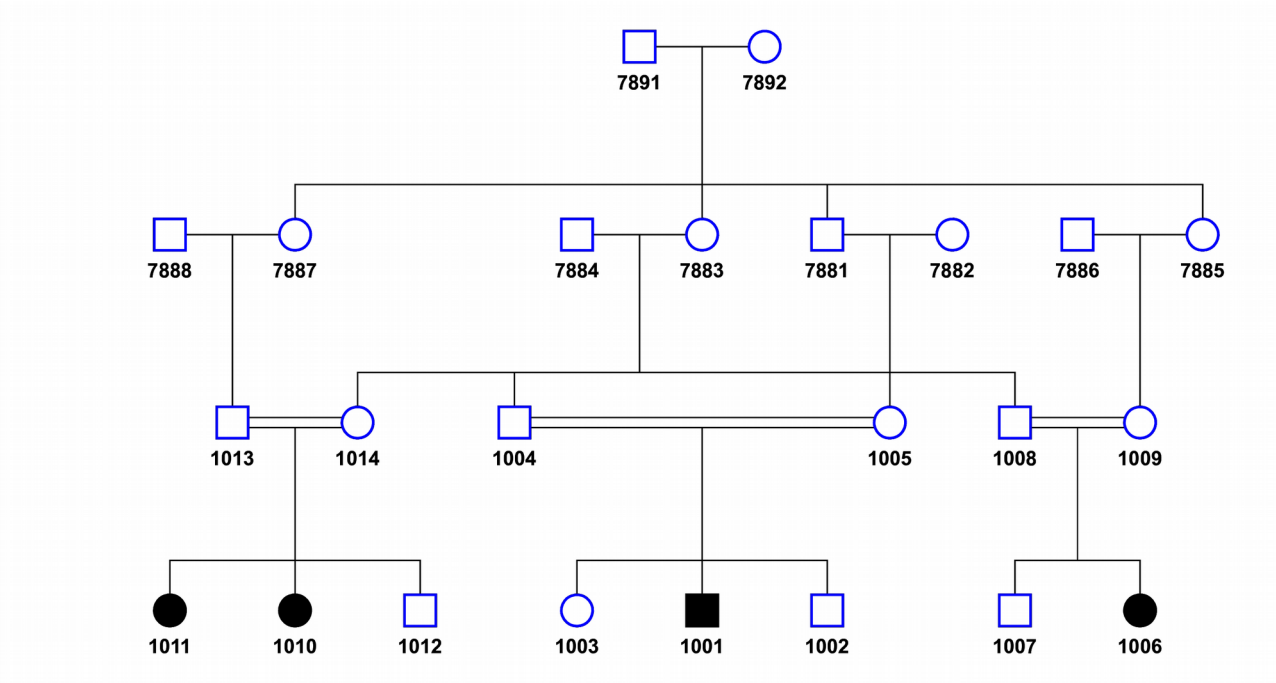


Figure 15: 29-bit Autosomal Recessive Pedigree, 4 affecteds in the last generation.

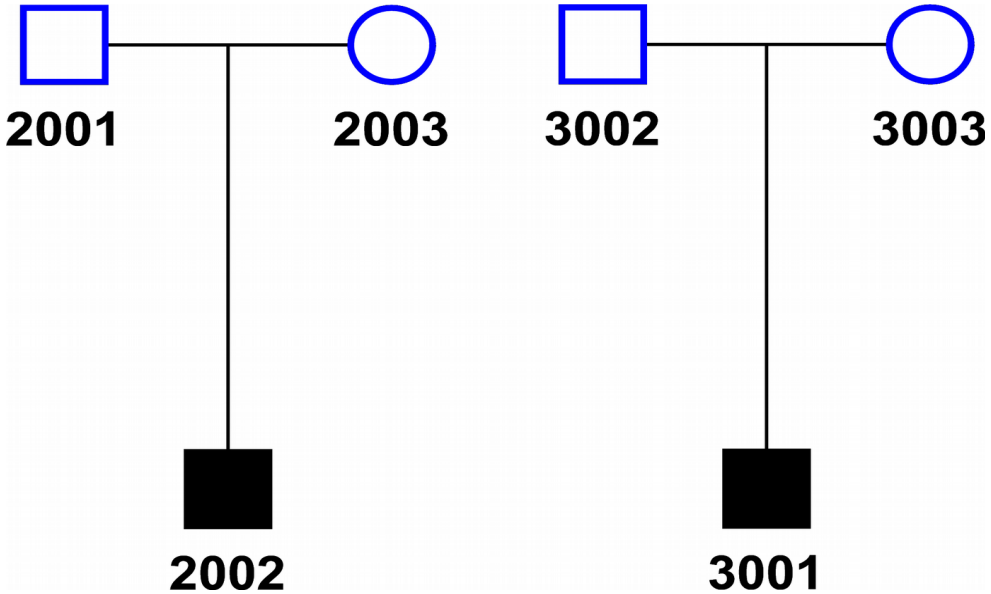


Figure 16: Two completely uninformative pedigrees for linkage (bit-size of -1)

The combined Mendelian errors across the three pedigrees amounted to no more than 175 SNPs out of the total 47,595 selected for linkage (0.368%) , with 3 SNPs encountered more than once as shown in Figure 17.

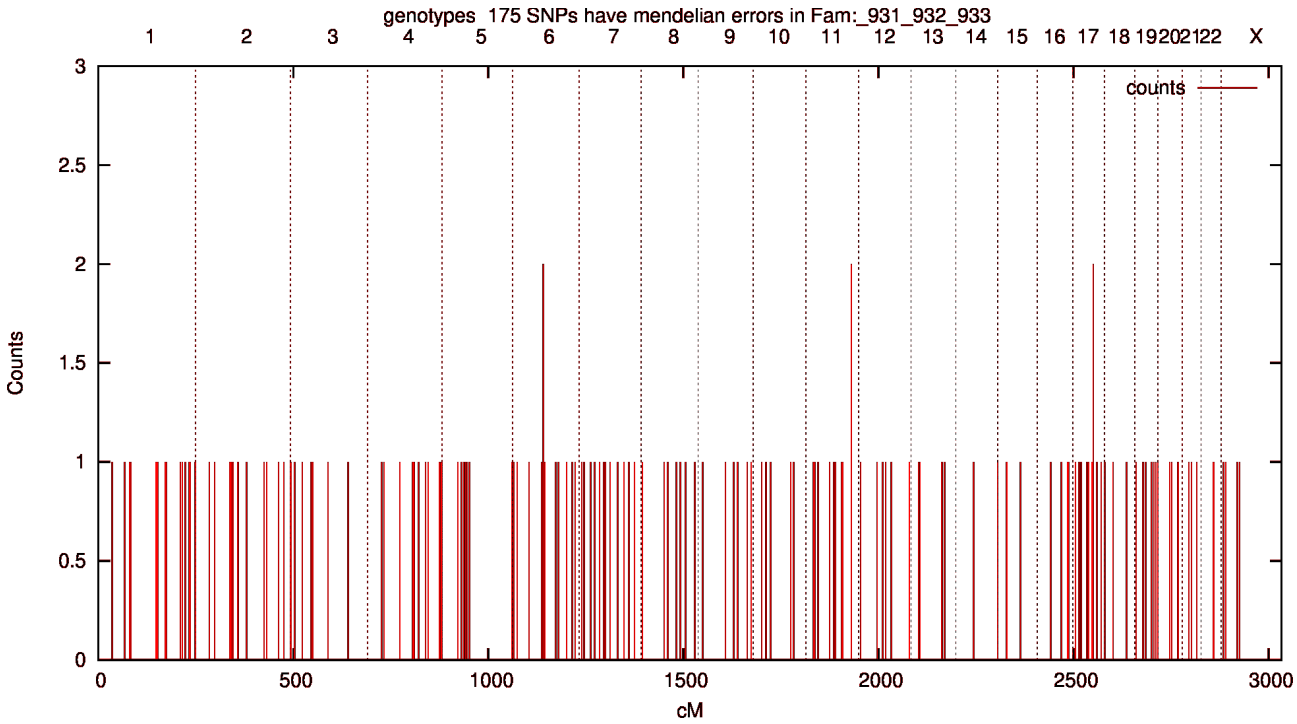


Figure 17: Mendelian errors for the three families considered under the 21-bit Autosomal Recessive analysis. The three peaks relate to an overlap of an erroneous marker in two of the three families.

The GRR plot of the first family (Figure 18) showed good distinct relational grouping with some potential border overlap between the sib-pair and parent-offspring relations, but no actual outliers in either group. The GRR plots of the other two families were normal (Figure 19).

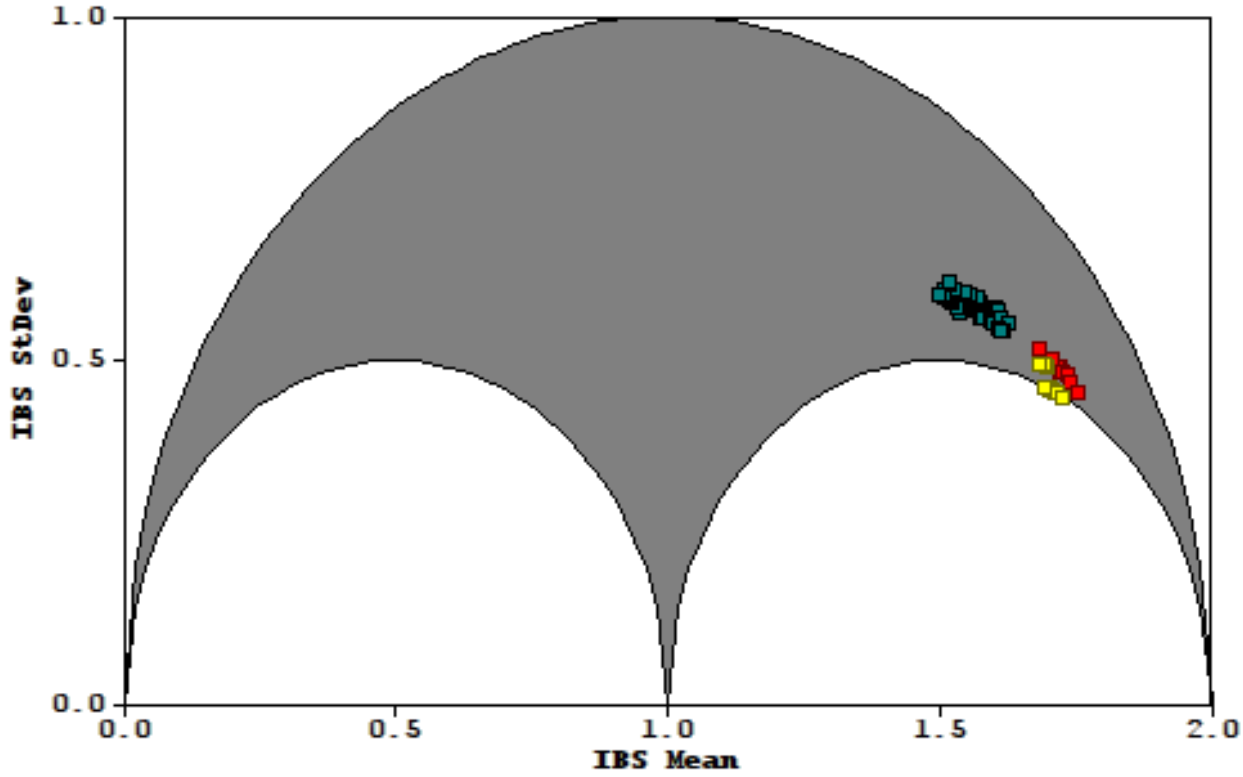


Figure 18: GRR plot of 29-bit family

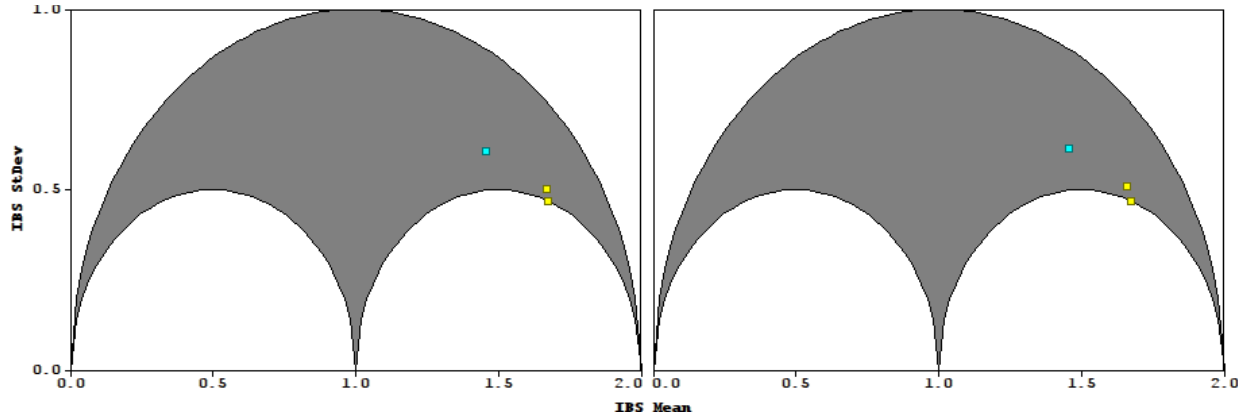


Figure 19: GRR plots of the smaller two families that contributed to the 29-bit analysis

Due to the complexity of the analysis, genome-wide analysis had to be performed via Simwalk due to the big data errors mentioned previously (see page <XXXREF> of Methods). An estimated maximum LOD score could also not be computed because of this reason since the calculation relied

upon GeneHunter and Allegro, both of which were not able to process the analysis in any reasonable timeframe.

The sliding-window approach of Simwalk allowed the parallelization techniques employed by the *simwalk_multicore.sh* script to fully utilize all threads available on platform without throttling the RAM. Timings of these runs are discussed later in the next section.

The Simwalk run produced a single peak on the p-arm (p13.3 to p13.2) of chromosome 16 with a LOD score of 5.21 (Figures <> and <>) . In order to confirm the Simwalk result and to gain haplotype reconstruction, Allegro was required. A single chromosomal analysis was performed on chromosome 16, reproducing the peak at the same locus with a LOD score of 5.23 (Figure 21).

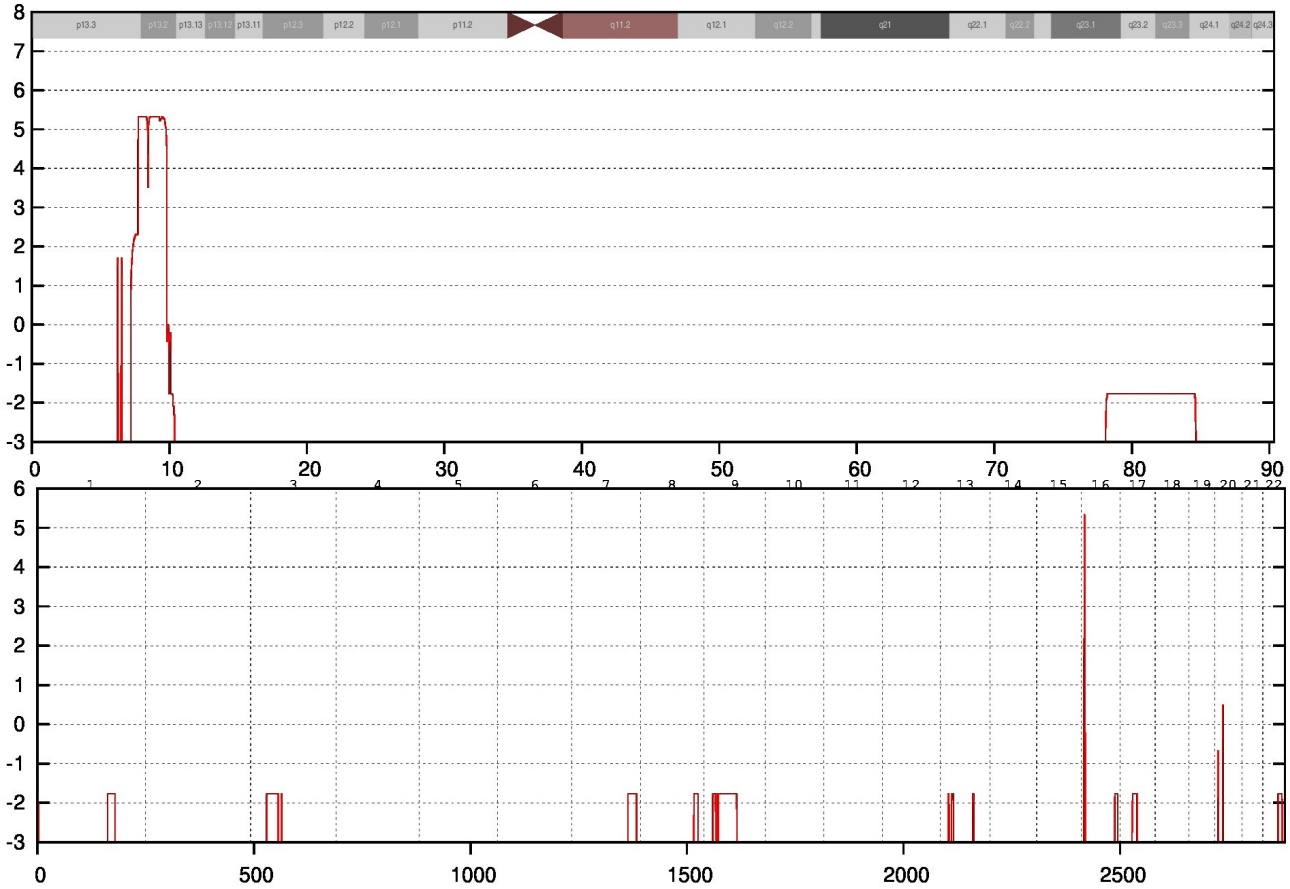


Figure 20: Simwalk plots of 29-bit family. Genomewide (Top), and Chromosome 16 (Bottom)

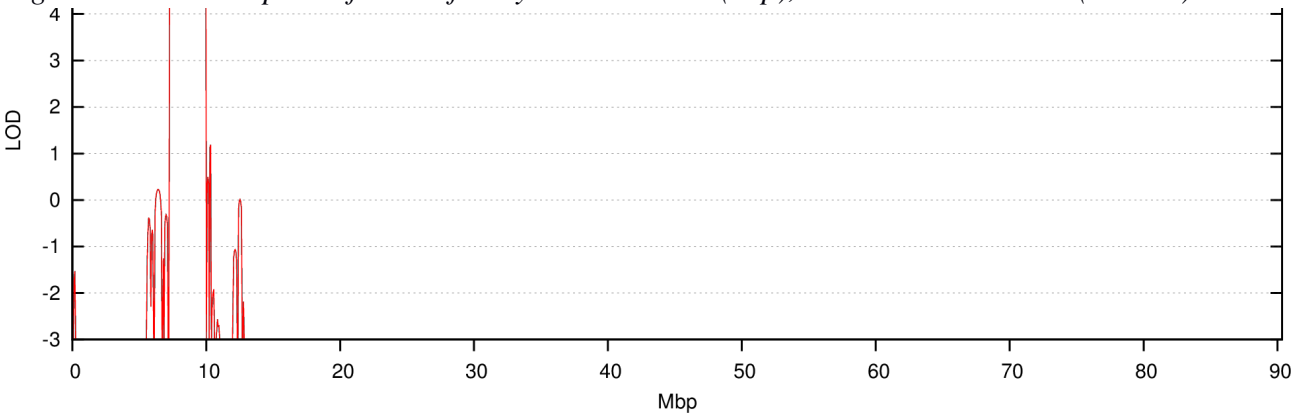


Figure 21: Allegro chromosome 16 of the 29-bit pedigree for the three families. Only the first family contributes to the peak score.

1.1.3 Pedigree Run Times

All analyses were timed using the GNU **time** utility in order to get the "wall clock" runtimes from the calling (bash) script.

For all projects with pedigree complexities less than 19-bits, the main limiting factor in their runtimes was the graphical GRR Xautomation script. Table 2 below shows how the mean runtime remains at a reasonably constant 122 seconds up to the 19-bit limit, where GRR window polling then begins to take precedence and longer wait times are required for the genotypes to fully load.

Bit-size	**No. of Genotyped Individuals**	**No. of Markers**	**Total Genotypes**	**1**	**2**	**3**	**4**	**Mean**
3	5	35,155	75,775	121.91	122.08	121.87	121.62	121.87
5	6	41,479	248,874	121.85	121.52	121.62	121.81	121.70
7	9	43,421	390,789	121.96	121.50	122.29	122.34	122.02
9	8	43,421	347,368	122.18	121.98	122.28	122.33	122.19
15	11	41,480	456,280	122.14	121.51	122.31	122.03	121.99
18	7	50,110	350,770	123.21	122.75	122.59	122.23	122.69
21	10	15,914	159,140	129.82	130.11	130.20	129.15	129.82
23	11	15,914	175,054	152.81	151.11	151.64	152.37	151.99
29	14	47,595	666,330	178.66	180.42	179.31	183.75	180.54

Table 2: GRR run times over 4 trials, ordered by ascending bit-size. The number of genotyped individuals and the number of genotypes per individuals are also listed.

A plot of Table 2 shows GRR mean run times against normalized bit-size, number of genotyped individuals, number of markers and total genotypes. The total genotypes, bit-size, and number of genotypes appear to follow an upwards sloping power trend, wheareas the run time appears to be more independent of the number of markers.

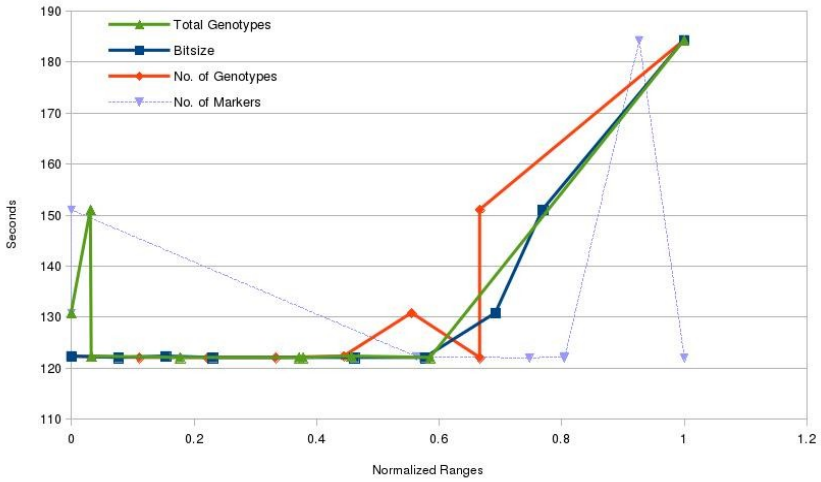


Figure 22: GRR run times. (Normalized) Genotypes, Markers, Total Genotypes, and Bitsize against Run Time.

The GRR runtimes appear to be constant at approximately 122 seconds within the established 19 bit limit with a small increase to 123 seconds at 18-bits. Larger pedigrees begin to scale more linearly, rising to 180 by the 29-bit stage, at which stage the contribution of the GRR window polling script runtime becomes negligible in contrast to the Allegro runtimes outlined in Table 3 below.

Bit-Size	**Total Runtime (seconds)**									
	SingleCore					**Multicore**				
	#1	***#2***	***#3***	***#4***	***Mean***	***#1***	***#2***	***#3***	***#4***	***Mean***
3	81.63	80.12	82.66	84.10	82.13	55.79	55.88	54.00	50.37	54.01
5	86.84	83.55	83.57	87.07	85.26	57.53	60.07	57.24	60.24	58.77
7	146.72	148.15	147.51	150.77	148.29	62.13	56.32	58.51	60.09	59.26
9	378.72	379.96	378.67	377.17	378.63	81.70	80.74	86.99	81.97	82.85
15	775.54	735.08	741.71	768.11	755.11	116.62	108.27	103.63	112.82	110.34
18	2779.57	2809.55	2795.13	2758.11	2785.59	389.55	415.02	400.36	404.79	402.43
21	943891.22	940984.64	948114.72	946515.60	944876.54	-	-	-	-	-
23	6531019.41	6780937.40			6655978.41	-	-	-	-	-

Table 3: Total Genome-wide Allegro Run Times for MPT and Haplotypes; single-core and multi-core. Singlecore timings for 23-bit pedigree underwent two trials only. Multicore timings for 21-bit and 23-bit analyses were not processed.

The following table contains the average chromosome-specific[4] runtimes that comprise the single-core components of Table 3.

Chrom	Bit-size								
	3	**5**	**7**	**9**	**15**	**18**	**21**	**23***	**29***
1	3.97	4.02	12.88	50.67	138.89	595.14	87849.27	730875.58	-
2	4.51	4.02	10.27	40.16	99.07	433.59	83949.84	635076.98	-
3	4.14	3.63	8.84	29.94	68.51	322.90	76173.28	665699.06	-
4	3.69	3.78	7.93	24.66	48.83	232.71	70044.33	600316.88	-
5	3.39	3.75	7.31	19.69	35.43	177.24	64033.08	493473.21	-
6	3.94	4.72	6.63	16.78	28.98	128.61	59448.95	461203.36	-
7	4.01	3.91	6.19	14.36	23.52	99.59	53240.12	409079.04	-
8	3.68	3.62	5.83	12.47	21.52	74.21	48531.97	378953.42	-
9	2.86	3.39	5.64	11.52	18.19	59.10	44332.31	280859.56	-
10	3.91	3.45	5.51	10.68	16.42	47.62	39859.10	275189.27	-
11	3.63	4.02	5.36	10.30	14.96	38.11	35716.50	259160.94	-
12	3.60	4.56	5.29	9.82	15.22	30.90	32045.80	217073.37	-
13	3.49	3.37	5.17	9.60	15.14	26.25	28290.74	167115.23	-
14	3.09	3.21	5.16	9.57	14.12	23.50	24899.32	148455.32	-
15	2.38	3.09	5.13	9.43	15.16	19.08	21915.39	110163.35	-
16	3.19	3.42	5.06	9.17	13.22	17.93	19353.51	76464.53	315615.20
17	3.77	4.32	5.09	9.07	14.46	17.95	16933.29	81678.43	-
18	3.73	3.69	5.02	9.04	14.43	15.91	14746.18	66359.72	-
19	3.07	3.44	5.08	9.26	13.77	16.14	12955.06	33812.14	-
20	3.31	3.23	5.00	9.01	14.27	15.33	11266.49	19439.38	-
21	3.71	2.91	5.06	9.11	14.18	16.13	10049.32	15547.82	-
22	3.54	3.51	4.99	9.00	13.71	14.22	9065.98	35267.98	
X	3.52	4.20	9.87	35.33	83.09	363.44	80176.73	494713.86	-

Table 4: Average Single-Core Allegro Run Times for Multi-point Parametric Linkage and Haplotype Reconstruction; Bit-Size (columns) against Chromosome (rows). All analyses are the result of 4 trials, except the 23-bit analysis which was only able to run for 2 trials and the 29-bit analysis which only run on chromosome 16, both analyses without haplotypes.

4 For a full disclosure of all trials run, please see *Allegro Single-core Runtimes* Section in the Appendix.

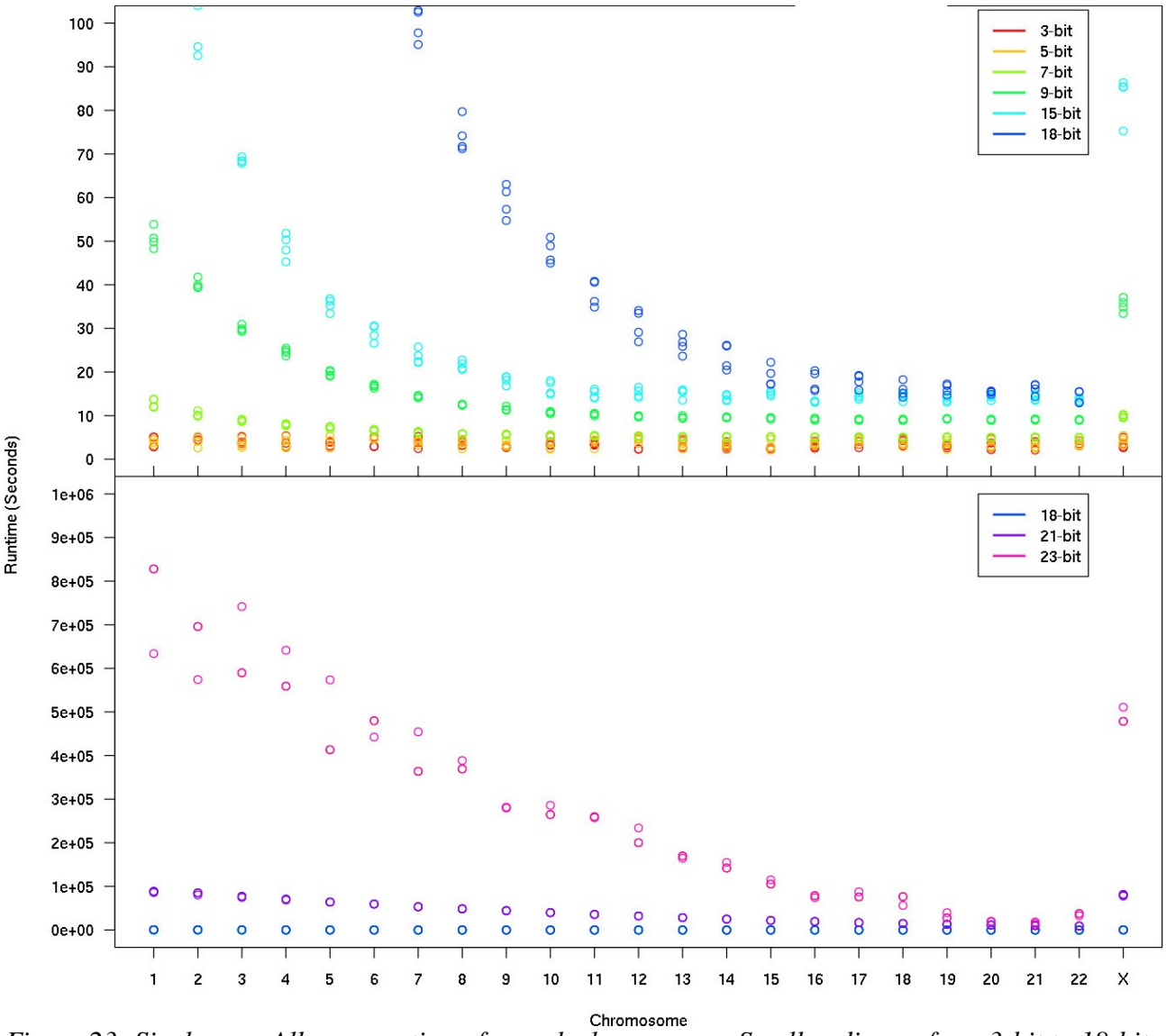


Figure 23: Single-core Allegro run times for each chromosome. Small pedigrees from 3-bit to 18-bit with runtimes within [0,400] range (Top), larger pedigrees from 18-bit to 23-bit shown with runtimes as powers of 10 (Bottom).

Figure 23 shows the plot of Table 3 at different scale; the top image showing the general logarithmic decrease in runtime with increasing chromosome number (1 to 22). The sharp rise at the end is due to the inclusion of chromosome X which in terms of size is more similar to chromosomes 1 and 2 than chromosome 22. For 3-bit, 5-bit, and 7-bit pedigrees, the time required to process each chromosome is almost constant; the main contributing factor to the timing being the overhead in the system calls related to the loading and unloading the Allegro binary. Should the Allegro binary remain in memory throughout the duration of a genomewide run, the plots may reflect the size of the chromosomes they process more accurately. 9-bit, 15-bit, and 18-bit pedigrees begin to take on

the logarithmic shape that better exemplifies the size of the chromosome, where the overhead in system calls becomes more negligible.

The bottom image represents a zoomed out Y-axis scale of the same data by 5 orders of magnitude. The smaller ({3,5,7,15,18}-bit) pedigrees appear grouped as a flat plot towards the bottom, whilst the 21-bit and 23-bit rise up above it. The difference in runtimes between 21-bit and 23-bit greatly deviates for the lower-number chromosomes, with chromosome 1 taking approximately 8 times longer to process for the latter analysis.

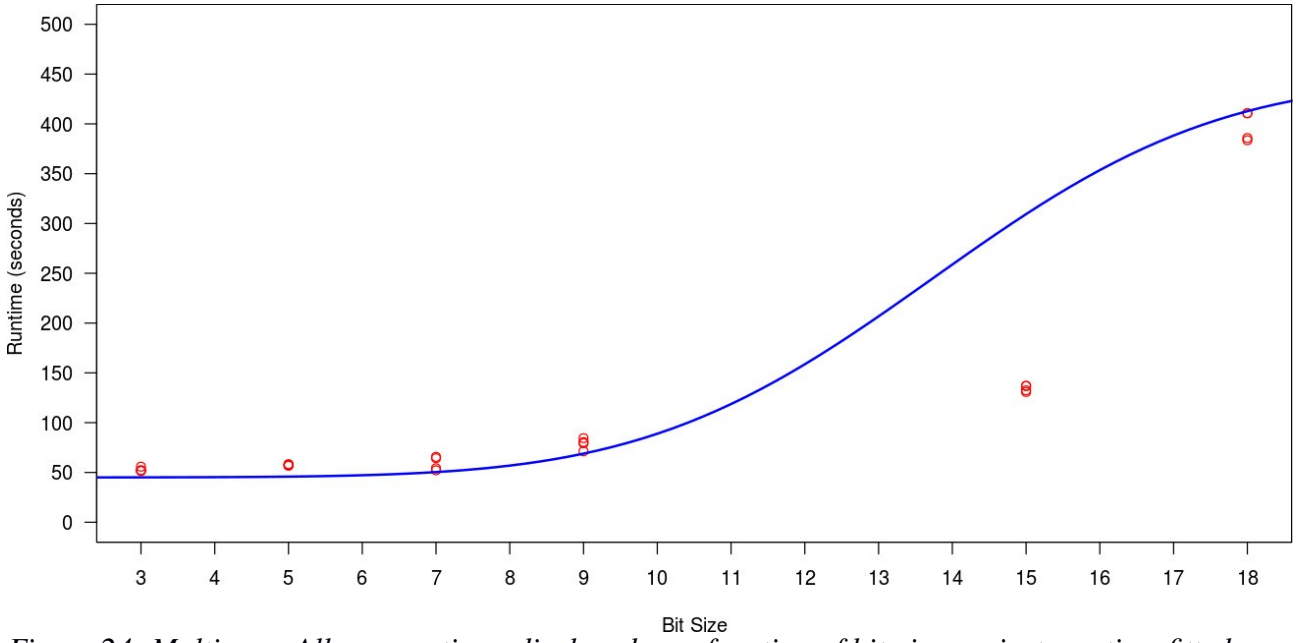


Figure 24: Multi-core Allegro runtimes displayed as a function of bit-size against run time fitted with a power curve. Individual chromosome runtimes were not recorded due to parallel evaluation.

The multi-core run-times shown in second portion of Table 3 are plotted above in Figure 24, where individual chromosome-runtimes were ommitted for simplicity. Runtimes remain reasonably linear from 3-bit to 15-bits where a sharp rise in runtime occurs 18-bits where the system resources requested by active Allegro instances begin to compete for memory.
<XXX change this to exp curve, and backdrop max RAM usage for both single and multicore>
Due to RAM constraints, the pedigrees larger than 18-bits were not able utilize parallelization effectively, only being able to process a single chromosome at a time (i.e. single-core) and so they were removed from the analysis.

For the 29-bit pedigree, Simwalk performed the genomewide analysis and then Allegro reconfirmed the chromosome 16 peak. Allegro took 315615.20 seconds (3.7 days) to process chromosome 16 alone. Table <XXXREF not invented> shows how all 256GB of physical memory was used within the first 4 hours of processing, and the 16TB swap memory was then used. A large portion of the processing at that point was memory paging in order to preserve all the LOD calculations being made. Simwalk writes straight to disk and has an extremely small RAM usage globally, so despite being the traditionally slower program (due to the sliding-window approach creating numerous overlaps), it did not suffer from paging bottlenecks. Due to the Simwalk parallelization script not being chromosome-specific (i.e. it executes any window-processing job it is dispatched), only the genome-wide runtime was recorded at 682117.03 seconds (~ 7.9 days).

1.2 Haplotype Resolution Results

We will look at the haplotypes of three large pedigrees with the following penetrance models: Autosomal Dominant, Autosomal Recessive, X-Dominant.

All the pedigrees viewed in the Results so far were produced by HaploPainter, and here we will examine some of the same pedigrees now rendered by HaploHTML5. Haplotype rendering and resolution between the two programs will also be compared.

23-bit Autosomal Dominant Pedigree and Haplotypes
Family 109 in the figure below consists of many members and so it was prudent to compare only the affecteds at a given locus in order to prioritize the haplotypes on screen. The colouring of the haploblocks is always semi-random so it is the position and lengths of the blocks themselves that we are most concerned with.

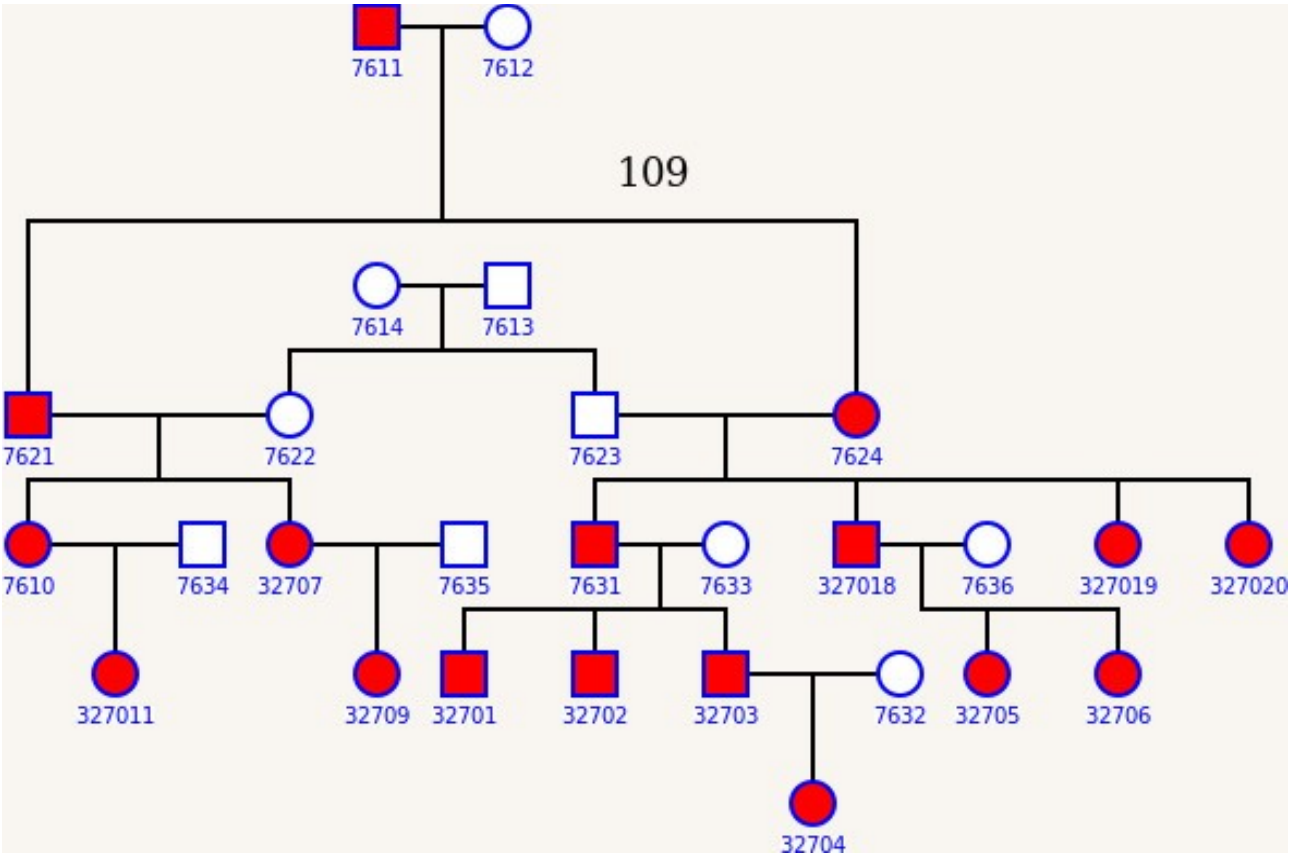


Figure 25: HaploHTML5 default rendering of pedigree 109 from Allegro haplotype input file. This is the 23-bit pedigree shown earlier inFigure 10. Since both founder couples (7611 and 7612, 7613 and 7614) were at the same grid 'level', a line overlap was inevitable and so 7611 and 7612 were raised to better distinguish childlines.

The first red arrow marks a point between rs7693827 and rs11735648 which refers to a recombination in both haploblock renditions where the founder allele switches to a red block HaploPainter which is equivalent to the green block in HaploHTML5. Both blocks persist fully in the right allele of 32707, 32709, 7610, and 327011. The red arrow at point 3 shows agreability too between the applications, though HaploHTML5 appears to have selected two very similar looking colors for the pink block just above point 3, which looks identical to the right allele of 32704 but has a different set of genotypes. Debugging the allele shows that HaploHTML5 does assign them different color groups, but the groups themselves have high colour homology due to an over abundance of founder alleles (9 founders, 18 alleles) and a limited selection of colors.

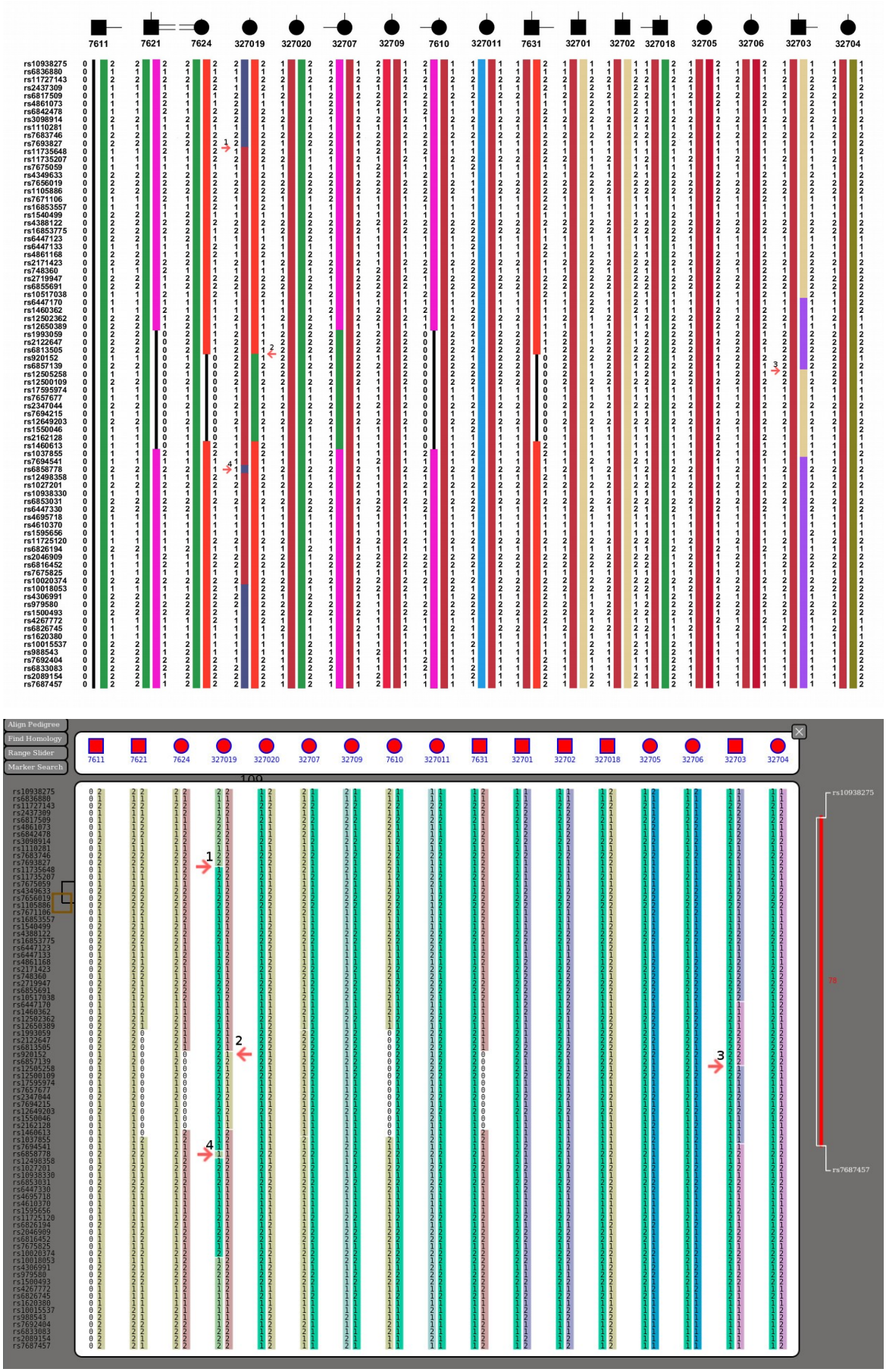


Figure 26: Haplotypes of family 109. HaploPainter modified to display affecteds inline (Top). HaploHTML5 default comparison view showing affecteds only (Bottom). Four noteable points of comparison points are highlighted with red arrows.

The red arrow at point 4 shows a genotyping error that is correctly picked up upon by both programs and assigned the first available group outside of the block it resides in.

In terms of design and presentation, HaploPainter prefers to present its haplotypes in a family-tree structure resorting in more manual methods of aligning the individuals of the pedigree for haplotype inspection. HaploPainter also places the text of the haplotype outside of the coloured block. HaploHTML5 presents the text within the block itself in order to preserve horizontal space, and as a result must use lighter shades of colours to improve the readability of the black text.

29-bit Autosomal Recessive Pedigree and Haplotypes

Here we have the three inbreeding loops that we encountered previously in the section, with consanguinuity correctly detected for all the couples in the second to last generation. There is only one founder couple, but there is inter-generational marrying that makes it impossible for lines to not overlap (7612 and 7611 for example).

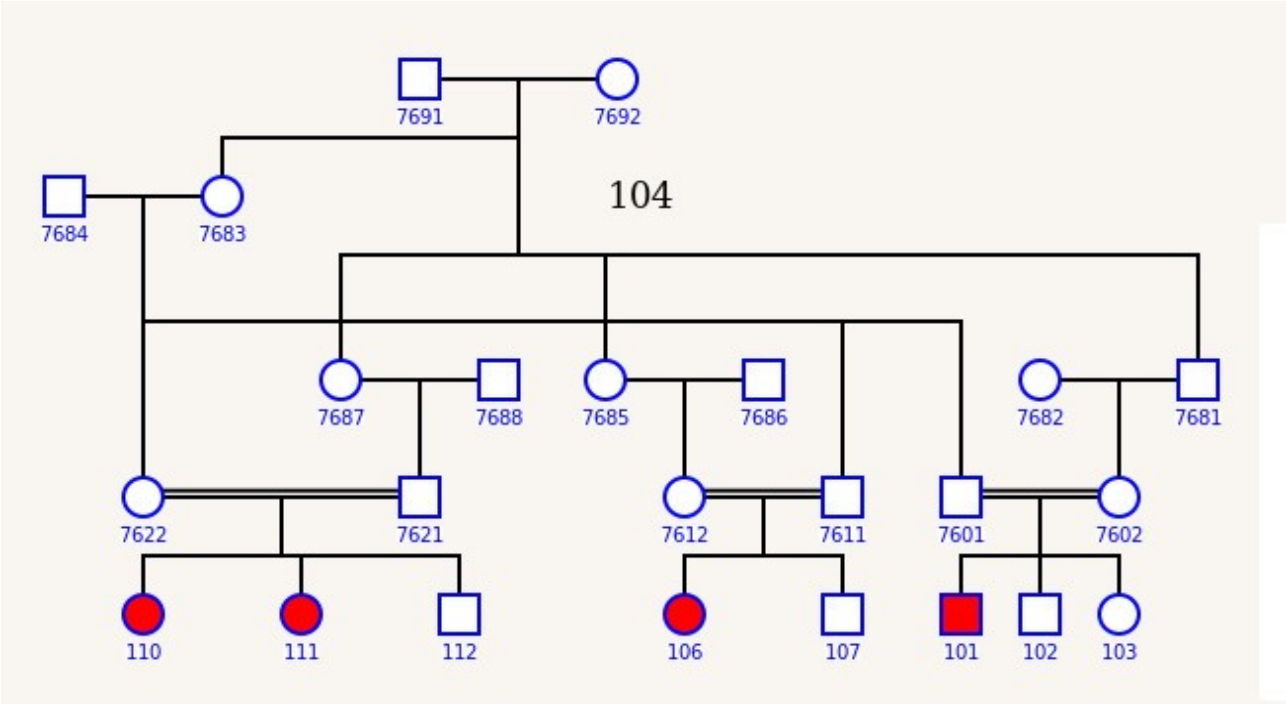


Figure 27: HaploHTML5 rendering of family 104, of the 29-bit pedigree originally shown in Figure 15 on page 25.

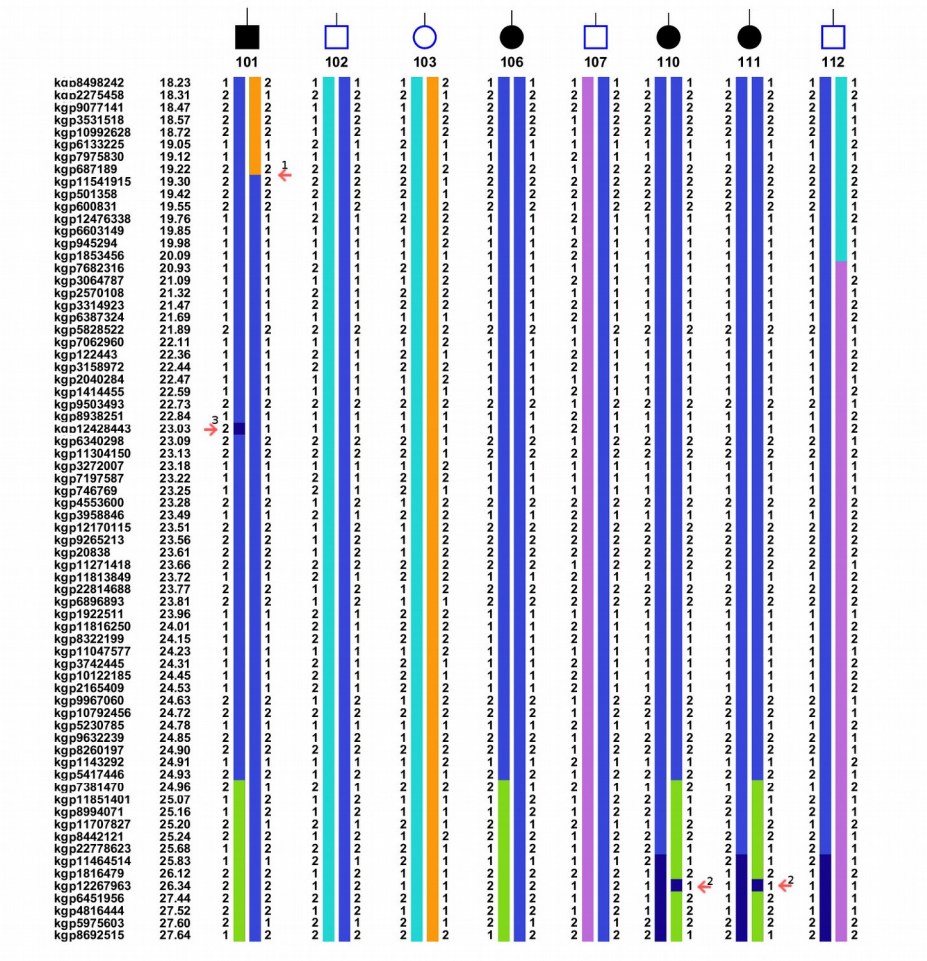
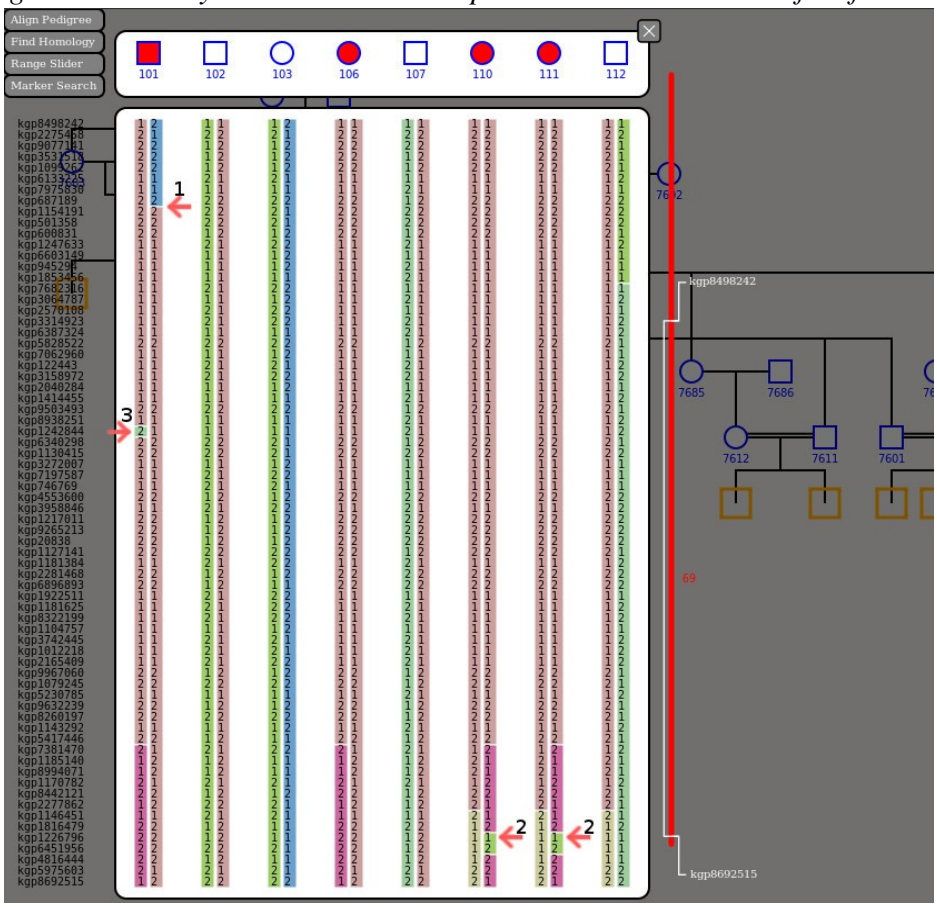


Figure 28: Family 104 rendered in HaploPainter and then modified for



horizontal alignment (Top) and HaploHTML (Bottom).

There are only 4 affected individuals in this pedigree, so we can now include some unaffecteds too (the last generation in this case). Here we once again see good accordance between HaploPainter and HaploHTML5 for the genotyping error at point 3, and the recombination at point 1.

However there is a slight difference at the point 2 location(s) where HaploHTML5 extends the small block at kgp12267963 to two marker loci and not the single locus given by HaploPainter. Both configurations are valid because the purple and green block (in the case of HaploHTML5) both share a '2' allele index at that locus. However, HaploPainter does not maximise the block and makes it appear more as a genotyping error or artefact instead of the small haploblock as HaploHTML5 does. This can be easily rectified by adjusting the *minimum stretch* parameter for the A* path finding algorithm such that it does not consider blocks less spanning less than or equal to two markers to be actual blocks (and thus does not try to maximise it), but this places an unfair constraint on the data because single-locus haploblocks may be valid for many-generational pedigrees with sparse markers.

A better compromise would be to allow the inclusion of marker map data (such as is read in by HaploPainter) so that the algorithm can determine whether the inter-marker distances surrounding a marker is great enough to warrant a haploblock on its own.
Figure 29 below compares the homology tools used by the two applications. HaploPainter was adapatedto take in messner files and marker pairs, where regions of homology could be detected using our **haploregion.py** script. Messner regions are boxed in a blue regions, and regions of homology are surrounded by red boxes where in this case the affecteds are all homozygous for the regions shown, with the extra condition that the unaffecteds are not also similarly homozygous.
HaploHTML5 performs all homology detection within the browser, but uses a different approach where it assumes that homology detection is not so binary and assigns a "homology score" at each locus which is overlayed against the haplotypes and marker slider. Haplopainter is much clearer in showing homology in this instance, but HaploHTML5 allows for some leeway for erroneous genotypes that might otherwise obstruct a binary homology detection.

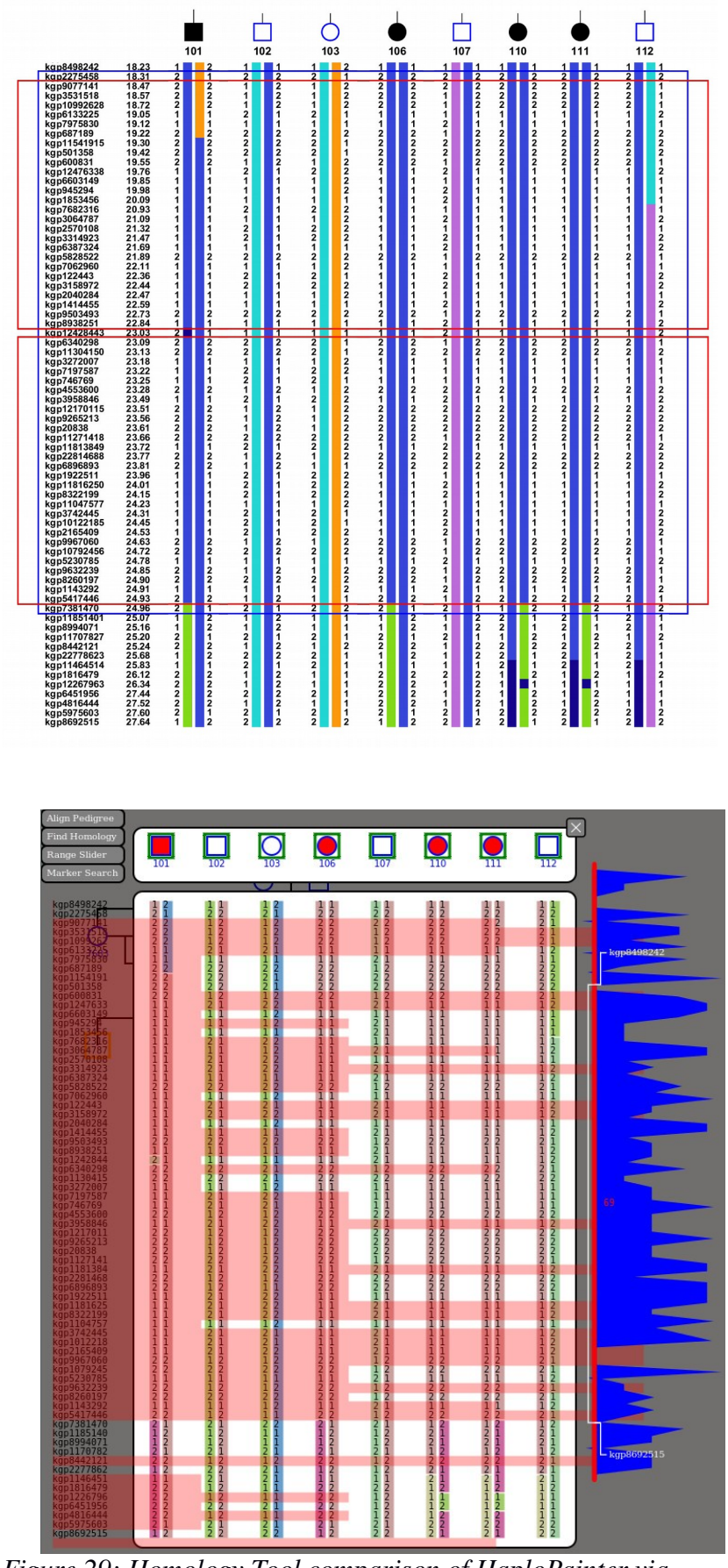


Figure 29: Homology Tool comparison of HaploPainter via HaploPainterRFH modification script (Top), and HaploHTML5 Homology Mode (Bottom)

15-Bit X-Linked Dominant Pedigree and Haplotypes

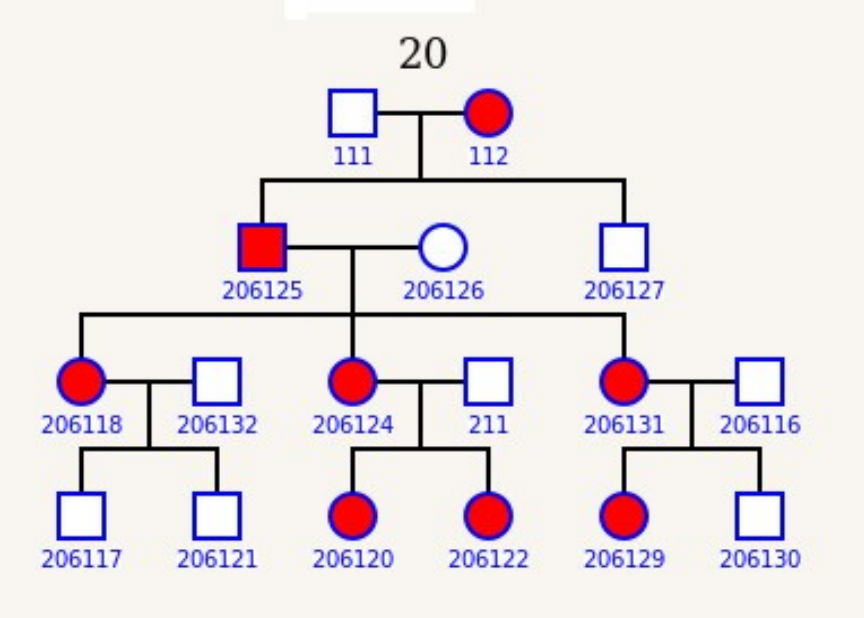


Figure 30: X-Linked Pedigree, 17 members of which 8 are affected individuals

We once again take a look at our X-linked Dominant pedigree mentioned previously on page 14, where we can examine the haplotypes of all individuals in the pedigree.

The figures below provide three ranges of which to examine the X chromosome where recombination errors were observed within HaploPainter. HaploHTML5 sets the correct penetrance model and provides no such recombinations, often providing complete non-recombinants for the loci considered.

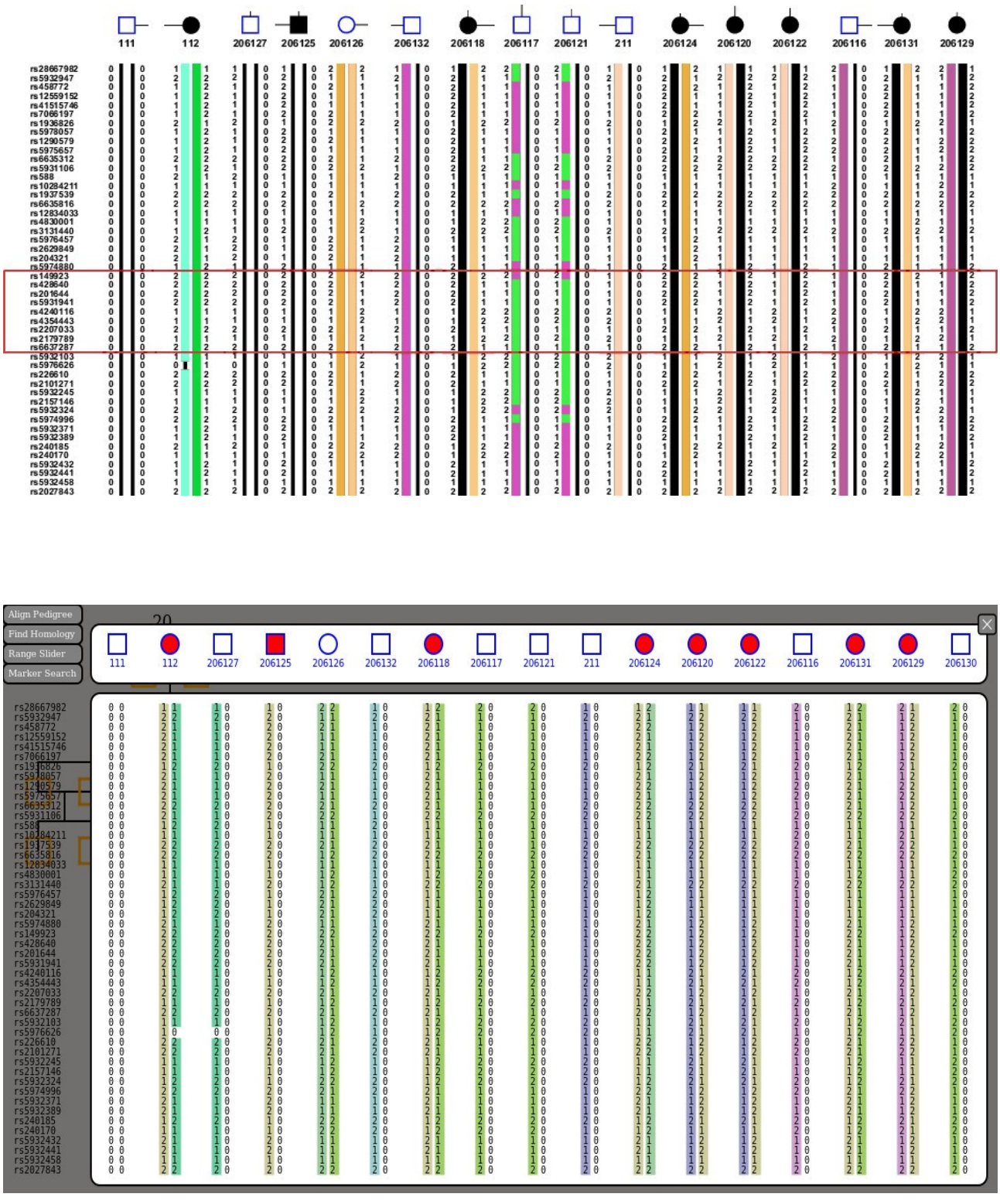


Figure 31: Starting telomeric P-arm region of chrX spanning 60 markers. (Top) HaploPainter recombination errors caused by bad X-inheritance. (Bottom) HaploHTML5 complete haploblocks due to correctly parsing the genotypes.

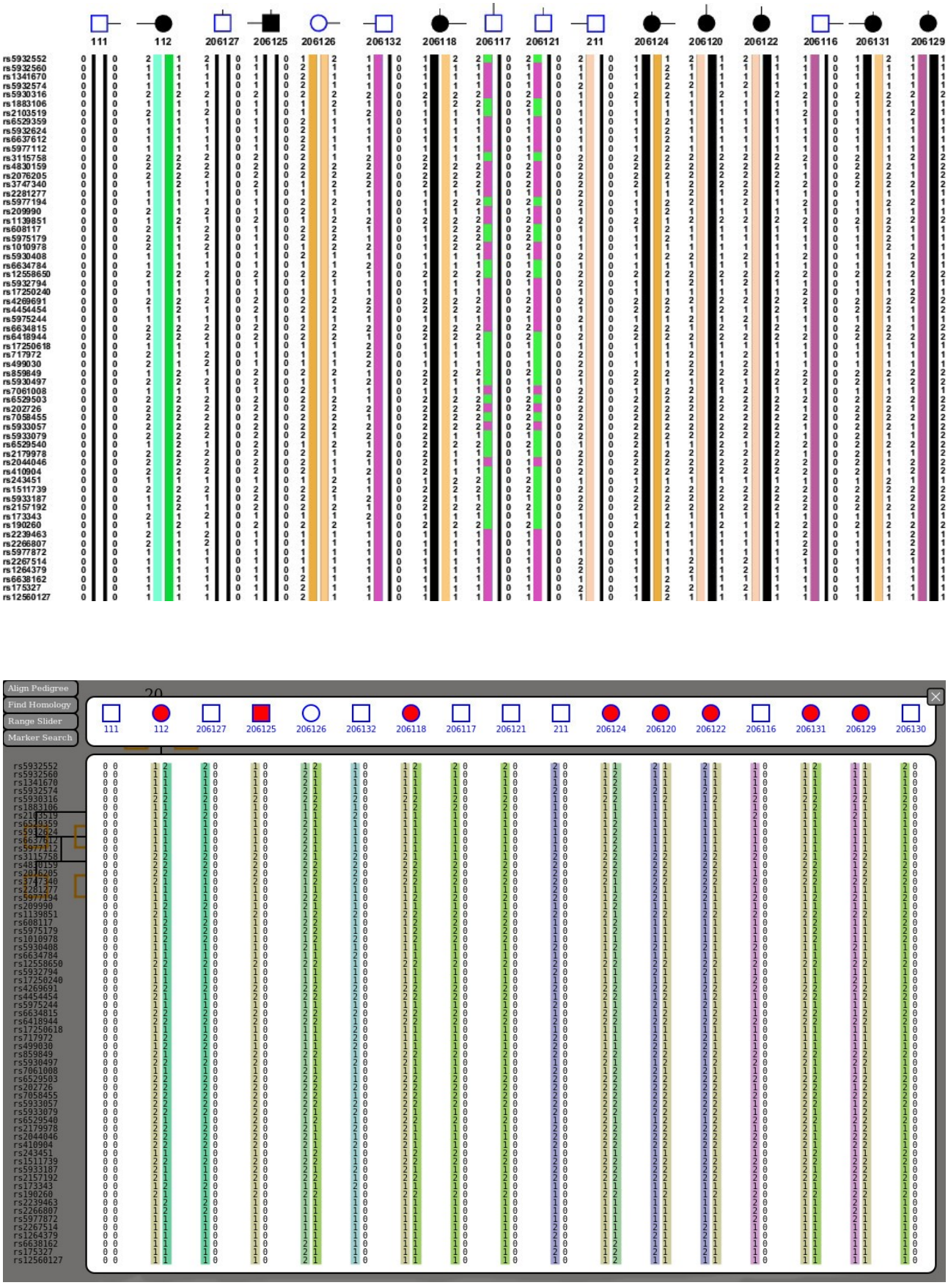


Figure 32: Middle region of chrX spanning 72 markers. HaploPainter (Top) and HaploHTML5 (Bottom).

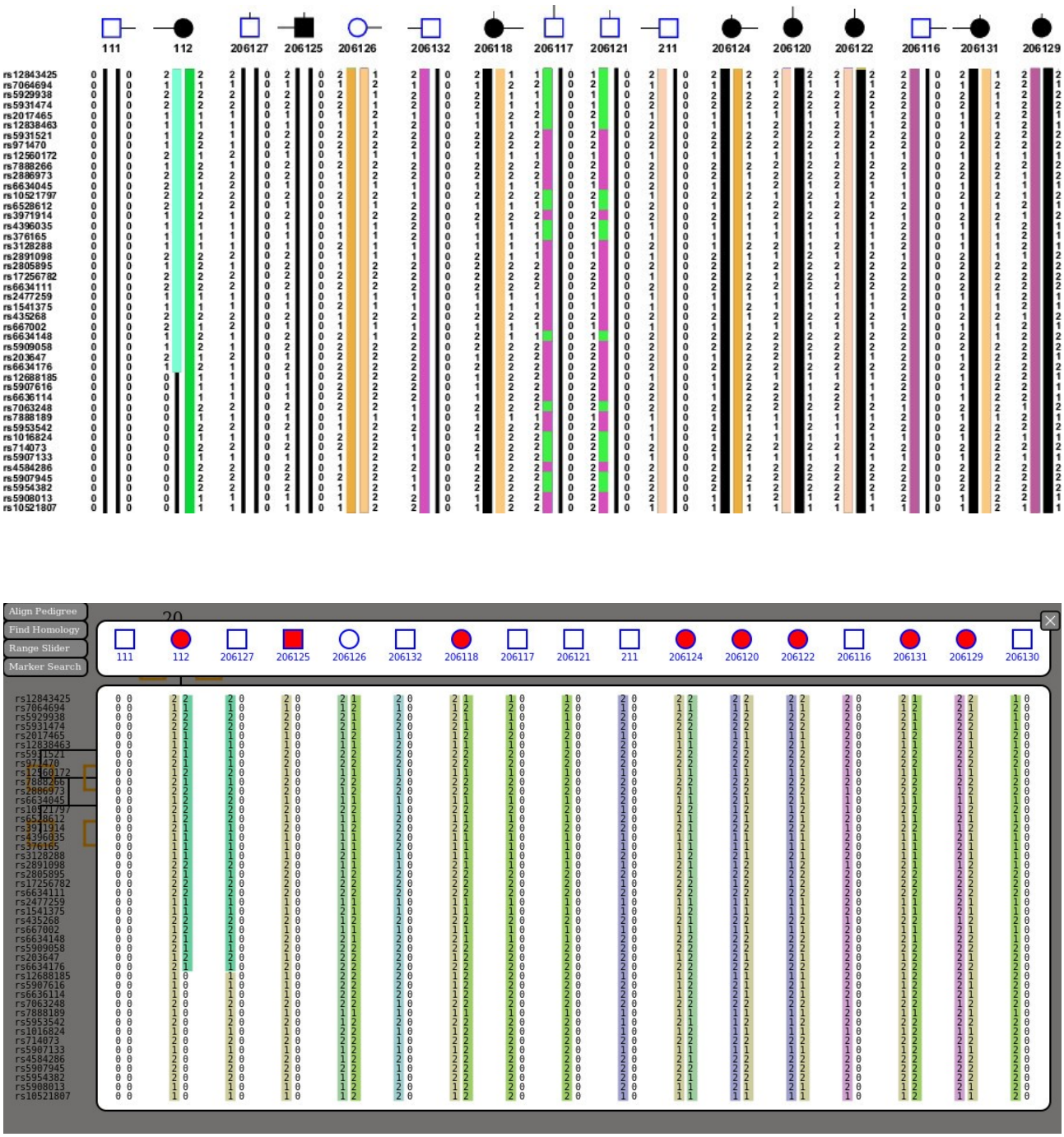


Figure 33: Finishing telomeric Q-arm region of chrX spanning 56 markers. HaploPainter (Top) and HaploHTML5 (Bottom).