### **Appendix V: Bioinformatic pipelines/scripts**

### **C. Illumina HisSeq 2000 dataset - data treatment**

Raw List

Reads from bam files mapped to genome reference were extracted and remapped to 46 (45 on Genebank and 1 novel allele in P5 sample set) reference PRDM9 alleles using BWA aligner with seed length increased from 19 to 100bp to effect virtually 100% stringency (zero mismatch). Coverage data was then obtained via VCF files for each sample-reference pair (sam/ref). This was the raw data used for analysis.

IF minbymean

DP\_min / DP\_mean gives a measure of how low the coverage could go across the ZnF array. This was used to normalise the data across samples so that they could be compared. For this measure to be truly comparable and diagnostic for different alleles, the measure was fitted into thresholds levels 0.9, 0.8, 0.75, 0.7, 0.65, 0.6, 0.5 and 0.4. For example, if min=112 and mean=140, then min/mean=0.8. If min/mean was equal to or more than 0.8, a score of ‘0’ was returned for that sam/ref. For this sam/ref the score at 0.9 threshold level would be ‘1’.

I do realise most people would want to see a 1 when min/mean is equal or higher than the threshold level and 0 when it is lower. Yet the reasoning behind using the opposite is that what we are looking for really is the situation when min/mean is lower than the threshold level and ‘1’ represents the ‘flag’ indicating that is it lower. Visually its more intuitive to ignore the zeros and look at the 1s as they are what we should be concerned about. Hope that makes sense.

IF minbymean VALUES COPY

Values from previous sheet were copied to preserve formula values during consequent sorting and other processing. min/mean was less than 0.9 for all samples and therefore not informative, so that category was removed from subsequent sheets.

Colouredmin/mean threshold levels were colour coded. Additionally sam/refs in the Sample column were coloured according to the highest DP\_mean/DP\_mean category for which they got a ‘0’ score. Here are two example sam/refs:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | DP\_min | DP\_max | DP\_mean | 0.80 | 0.75 | 0.70 | 0.65 | 0.60 | 0.50 | 0.40 | DP\_stdev |
| 001\_bas\_\_bas1\_srt-flalleleA | 112 | 172 | 140 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |
| 480\_ser\_\_SE30\_srt-flalleleL2 | 156 | 264 | 203 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 21 |

Measures of central tendency was also obtained to get a general idea for read depth:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | DP\_min | DP\_max | DP\_mean | 0.80 | 0.75 | 0.70 | 0.65 | 0.60 | 0.50 | 0.40 | DP\_stdev |
| mean | **24.1** | **219.5** | **129.1** |  |  |  |  |  |  |  | **46.5** |
| median | 0 | 212 | 125 |  |  |  |  |  |  |  | 44 |
| mode | 0 | 195 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| stdev | 36 | 57 | 39 |  |  |  |  |  |  |  | 18 |

Mean and median values across data set were similar. However, using these values as cutoffs (to delete sam/ref entries below these levels) will definitely delete true PRDM9 alleles (see P5 Sanger Confirmed Alleles). Hence, this summary is not useful going forward.

Also, samples which carry the same PRDM9 alleles have been shown (eg. compared 001 which was ‘confirmed from previous mismatch allowed mapping results viewed from IGV and 013 which was found to be A/A via Sanger) to have differences in DP values:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | DP\_min | DP\_max | DP\_mean | 0.80 | 0.75 | 0.70 | 0.65 | 0.60 | 0.50 | 0.40 | DP\_stdev |
| 001\_bas\_\_bas1\_srt-flalleleA | 112 | 172 | 140 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |
| 013\_aboaus\_C\_AMD\_srt-flalleleA | 106 | 207 | 144 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 20 |

For DPmin, I have highlighted all sam/refs which returned a score of ‘0’ since this means there is atleast one base within the ZnF array for which no segment of any read mapped. In fact, it has been shown in IGV that these ‘gaps’ are longer than a mere base as they correspond to samples which do not contain a ZnF that is in the reference given.

For DPstdev, I have highlighted all sam/refs from 47 and above with reference to the mean DPstdev of 46.5 across the data set. This is not meant as a cutoff. As seen in the A/L4 Sanger sequenced sample (see P5 Sanger Confirmed Alleles sheet), DPstdev can be as high as 80 when the sample is heterozygous for the allele length.

In summary, using any one of these values by itself cannot help deduce the allele for any particular sam/ref but the min/mean value and the threshold levels I have setup can be used to compare and understand between sam/refs in order to deduced the alleles carried by the samples in the data set.

DPstdev 1to8 removed

I reordered according to DPstdev and found that DPstdev 1-8 contained all 46 sam/refs for samples 449 and 458. Additionally, the coverage for these samrefs were considerably low compared to the next set of samples:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | DP\_min | DP\_max | DP\_mean | 0.80 | 0.75 | 0.70 | 0.65 | 0.60 | 0.50 | 0.40 | DP\_stdev |
| 449\_pal\_J1\_5370\_srt-flalleleL24 | 0 | 31 | 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 |
| 449\_pal\_J1\_5370\_srt-flalleleL9 | 0 | 31 | 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 |
| 449\_pal\_J1\_5370\_srt-flalleleL20 | 0 | 31 | 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8 |
| 449\_pal\_J1\_5370\_srt-flalleleL3 | 2 | 35 | 13 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8 |
| 151\_ork\_\_ORK585\_srt-flalleleA | 54 | 101 | 77 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 9 |
| 151\_ork\_\_ORK585\_srt-flalleleE | 54 | 101 | 77 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 9 |
| 151\_ork\_\_ORK585\_srt-flalleleL2 | 49 | 101 | 76 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 10 |

So it was decided to mark these samples on PRDM9 Assignion sheet as -/- i.e. their PRDM9 alleles connoted be determined via DP data.

Dpmin0 removed

Removed all Dpmin=0 sam/refs since they have been demonstrated on IGV to show clear vertical gaps where no reads could align. Typically these gaps were atleast 6-10 bases long. This indicated that the sample did not carry a specific ZNF type in the given reference. Hence, 10584 sam/refs were removed from consideration.

0.8 Score

It became apparent that sam/refs for A and E (a truncated version of A with the middle 5 ZnFs missing) give ‘perfect’ mapping when samples are A/A.

|  |  |  |
| --- | --- | --- |
| Allele Name | ZnF Repeat No. | ZnF Array |
| A | 13 | ABCDDECFGHFIJ |
| E | 8 | ABCD HFIJ |

Looking at the adjacent ZnFs, DH in E and DD in A allele, it is not quite clear why the mapping looks so perfect in both. D to H ZnFs have only a 51base stretch (17 on 5’prime end and 34 on 3’ end) where the sequence is identical.

D/H ZnF:

TGTGGGCGGGGCTTTAGCCGGCAGTCAGTCCTCCTCACTCACCAGAGGAGACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG

TGTGGGCGGGGCTTTAGAGATAAGTCAAACCTCCTCAGTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG

On the one hand, mapping with a seed length of 100bp would have eliminated any reads that did not have a clear D to H ZnF sequence. So the mapping result must be accurate. But looking retrospectively at some 215 samples that I have been assigned A/A or A/(structurally related N allele) and the fact that the frequency of E allele is 0.019 in European populations (according to Berg papers which are mostly northern European) and 0 in Africans, I cannot say that all of these samples are A/E. Yet, the 4 samples that gave DPmin/mean=0.8 are from different populations including Greek, Himalayan and Basque. I will need to get some information from that Oxford paper on haplogroups to see how different these populations are but given the high coverage in these 4 samples I am inclined to assign them as A/E individuals. I think Sanger sequencing should be on some of these samples to disprove this. For the rest of the samples where sam/refs for E allele has DPmin/mean=0.75 and lower threshold levels I have been more careful with assigning A/N alleles, in most cases preferring the longer alleles as theoretically the longer the allele reference the better the mapping and coverage.

There is a way to confirm that all these samples are not A/E or A/N but A/A without Sanger sequencing. This is by using the mapping results obtained earlier which allowed for mistmatches. I have demonstrated before that B allele which has ABCDDCCFGHFIJ allows me to detect A/A samples due to the drop in coverage due to the 1base difference in E and C ZnFs.

E/C ZnF:

TGTGGGCGGGGCTTTAGCTGGCAGTCAGTCCTCCTCAGTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG

TGTGGGCGGGGCTTTAGCTGGCAGTCAGTCCTCCTCACTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG

So for the moment, for the min/mean>=0.8, I decided to score as A/E. But I will resolve the E ZnF issue after my holidays. I have also removed all remaining sam/refs for these 4 samples from subsequent lists.

DPmin/mean-based ID

Based on the Sanger confirmed alleles DP data, a DPstdev High Stringency approach is not appropriate. For example:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | PRDM9 1 | PRDM9 2 | DP\_min | DP\_max | DP\_mean | DPmax/mean | 0.80 | 0.75 | 0.70 | 0.65 | 0.60 | 0.50 | 0.40 | DP\_stdev |
| 409\_pal\_G2a\_4943\_srt-flalleleA | A |  | 67 | 413 | 206 | 2.0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 80 |
| 409\_pal\_G2a\_4943\_srt-flalleleL4 |  | L4 | 58 | 241 | 153 | 1.6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 42 |

Continued with DPmin/mean High Stringency approach i.e. for each sam/ref I chose the alleles with the highest threshold levels (0 score).

While going down list of 0.75 level, the first two alleles for any samples were noted in the PRDM9 Assignment sheet. If both alleles do not occur in the 0.75 level, then I went down the threshold until the two alleles were found.

If more than 2 sam/refs appear together at same level when selecting the second allele, then I selected the one allele which had the highest DPmin.

Sometimes, the DPstdev were the same for more than one sam/ref. So I introduced a new measure DPmax/mean where anything equal or more than 1.5 is highlighted and not used for assignion. The rationale for this was that Dpmax/mean gives a normalised measure for when reads that do not really belong over a particular region of the reference causes stacked reads. This measure would be an indicator of stacked reads. However, as seen above with the A/L4 sample, it is more of an indicator of length heterozygosity. This somewhat helped me when deciding between sam/refs which gave similar DPmin/mean and DPstdev values.

The colour coding helped to visualise this pattern when looking for the right alleles to assigne a sample. The A/A sample here which was Sanger sequenced and all the A/A alleles I came across had a similar profile.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | DP\_min | DP\_max | DP\_mean | DPmax/mean | 0.80 | 0.75 | 0.70 | 0.65 | 0.60 | 0.50 | 0.40 | DP\_stdev |
| **013\_aboaus\_C\_AMD\_srt-flalleleA** | 106 | 207 | 144 | 1.4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 20 |
| 013\_aboaus\_C\_AMD\_srt-flalleleE | 106 | 207 | 144 | 1.4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 20 |
| 013\_aboaus\_C\_AMD\_srt-flalleleL3 | 106 | 207 | 151 | 1.4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 24 |
| 013\_aboaus\_C\_AMD\_srt-flalleleL5 | 112 | 221 | 156 | 1.4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 27 |
| 013\_aboaus\_C\_AMD\_srt-flalleleL2 | 94 | 207 | 142 | 1.5 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 24 |
| 013\_aboaus\_C\_AMD\_srt-flalleleL39 | 106 | 214 | 157 | 1.4 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 27 |
| 013\_aboaus\_C\_AMD\_srt-flalleleL11 | 86 | 207 | 146 | 1.4 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 29 |
| 013\_aboaus\_C\_AMD\_srt-flalleleL44 | 70 | 207 | 138 | 1.5 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 26 |
| 013\_aboaus\_C\_AMD\_srt-flalleleL33 | 60 | 207 | 134 | 1.5 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 31 |

Yet, from the 0.70 level and downwards, I had to look at the structure of the ZnF arrays that consistently come up in the same levels. I decided that samples would be tentatively assigned A/A with a note of the other alleles that came up as candidate so that I can look into it further in mapping results which allowed mismatch (VCF and IGV). Below are some of the most common alleles that came up:

|  |  |  |
| --- | --- | --- |
| A | 13 | ABCDDECFGHFIJ |
| E | 8 | ABCDHFIJ |
| L3 | 12 | ABCDECFGHFIJ |
| L39 | 9 | ABCFGHFIJ |
| L5 | 10 | ABCDDECFIJ |

Decide not to eliminate sam/refs with DPmin/mean=0.4 level where the score was ‘1’ since the Sanger confirmed samples also had alleles that gave a ‘1’ score i.e. fell below 0.4.

A/L33 has been assigned in many cases but there are arguments for an against this allele. In some of cases, the Dmin/mean score was too high to ignore. Yet, it could be that the mapping algorithm is trying to fit these reads (100bp even though its paired end reads), spreading them over the L33 reference.

|  |  |  |
| --- | --- | --- |
| Allele Name | ZnF Repeat No. | ZnF Array |
| A | 13 | ABCD DECFGHFIJ |
| L33 | 14 | ABCDDDECFGHFIJ |

A portion of these samples with the higher DPmin/mean scores should also be Sanger sequenced if necessary (even though we are only interested in very rare and novel alleles). L33 was reported in the Hussin 2013 paper so there is no frequency data (?) for it.

L5 (missing FGH motif compared to A but begins again with F ZnF) and L3 (one D missing compared to A) also came up a lot in the higher DPmin/mean scores. When either of these alleles came up before or after A (in terms of score) then I have assigned A/L3 or A/L5. When both L3 and L5 gave high scores, I considered the DPmeanscore and assign for the allele with higher coverage. However, they could all be just A/A but I have assigned them so at this stage so that I can cross-check with the allele frequencies in populations (L3 Eur 0.003 Afr 0 and L5 0 Eur Afr 0.0007).

|  |  |  |
| --- | --- | --- |
| Allele Name | ZnF Repeat No. | ZnF Array |
| A | 13 | ABCDDECFGHFIJ |
| L3 | 12 | ABCD ECFGHFIJ |
| L5 | 10 | ABCDDEC FIJ |

When the DPmin/mean scores for all sam/refs for any one sample were below the 0.4 threshold I have left them unassigned. When only one sam/ref gave a positive score within one of the thresholds and the rest of the sam/refs were below all thresholds I have left the second allele unassigned. In total, there were 24 samples with ?/? and 33 samples with N/?.

PRDM9 Assignment

This is the list of samples and assigned alleles.

P5 All Alleles

For reference, all the sam/refs for the 5 samples that were Sanger sequenced.

P5 Sanger Confirmed

For reference, all the Sanger-confirmed sam/refs for the 5 samples.

Alleles by Length

For reference, ZnF array structures of all known PRDM9 alleles.

Znfs compared to K ZnF

Hussin 2013 reported an overrepresentation of K-finger containing alleles. This includes all Ct alleles plus D and L20 allele. As part of the B-ALL study, we have been leveraging the use of a PRDM9 associated SNP haplotype network. So far we have found that there is no excess of Ct alleles in patients compared to controls. Similarly, we found no excess of D and L20 alleles in patients compared to controls. So why did they report an excess of K-finger containing alleles? According to the supplementary data, the sequencing as done using SOLID platform with 50bp read length. Since we are struggling to make 100bp reads give true results by simple mapping, then surely their mapping to known references will ‘yield the results they expect’ i.e. if you map such short reads to C allele reference then the sample will appear to have a C allele.

However, we do think that their report of K-fingers is worth looking at. To supplement this work, it would help to do a K-finger search for the NGS data set. We could determine if non-Ct, D and L20 alleles could contain K-ZnFs in this more diverse data set. If we find such samples and fully characterise using Sanger and our mapping results, and find this to be a novel allele, then it is possible that the leukaemia cohort might be carrying rare alleles containing K-ZnFs.

The first thing I needed to do was to compared K-ZnF to all of the known ZnFs. As you can see, a larger proportion of the 67 ZnFs I compared can be removed due to the 18th base from the 5’ end. Further filtering shows 4 ZnFs with 1base difference to the K-ZnF. A combination of base18-22 and those single base differences (38, 39, 63, 72) can be used for this search. Firstly, mapping will be done with 55bp seed length with K-finger allele reference. So the seeds will be identical to K-ZnFs from base18 to 72. All other reads will be eliminated. Samples which do no contain any K-ZnFs will also be evident.

Next Steps

1. Use mismatch allowed mapping results for reference B allele to confirm A/A allele signature (all A/N samples)
2. Conduct K-finger search
3. Resolve the A/E issue (get haplogroup information, use reference B mapping to confirm A/A allele signature)
4. Sanger sequence ?/?, N/? samples
5. Sanger sequence a selection of A/L3, A/L5, A/L33, etc samples

Sanger Sequencing Confirmation of PRDM9 Allele Assignment by Remapping and Analysis Pipeline

As part of determining whether the three (03) individuals homozygous A for SNP35 in the NGS set carried at least one (01) D allele, additional DNA samples from the NGS dataset were handpicked for their read depth scores (min/mean, standard deviation across ZnF array, minimum depth=0 removed) as follows:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Original file name | P1 | P2 | DP\_min | DP\_max | DP\_mean | Score | Notes |
| 001\_bas\_\_bas1\_po | A |  | 112 | 172 | 140 | 0.80 | E ZnF issue |
|  |  | E | 112 | 172 | 140 | 0.80 |  |
| 479\_ser\_\_SE26\_po | A |  | 152 | 247 | 195 | 0.75 | E 0.75 L39>1.5 |
|  |  | L3 | 152 | 287 | 199 | 0.75 |  |
| 335\_spa\_\_SP51\_po | L2 |  | 116 | 224 | 161 | 0.70 |  |
|  |  | A | 116 | 224 | 166 | 0.65 |  |
| 194\_fri\_\_fri1722\_po | L5 |  | 88 | 186 | 127 | 0.65 | A/A? |
|  |  | A | 71 | 186 | 115 | 0.60 |  |
| 280\_mbuti\_B2b2\_AFP13\_ph | C |  | 68 | 200 | 126 | 0.50 |  |
|  |  | GMO15 | 51 | 200 | 120 | 0.40 |  |
| 092\_gre\_E1b1b1a2\_GR99-78\_po | L5 |  | 56 | 181 | 139 | 0.40 | L40 was shorter |
|  |  | L9 | 63 | 181 | 126 | 0.40 |  |
| Potential D Carriers |  |  |  |  |  |  |  |
| 320\_nor\_\_N27\_po | L5 |  | 151 | 291 | 207 | 0.70 |  |
|  |  | A | 128 | 260 | 191 | 0.65 |  |
| 168\_tur\_\_TK10\_po | L39 |  | 91 | 159 | 121 | 0.75 |  |
|  |  | L3 | 58 | 159 | 109 | 0.50 |  |
| 334\_tur\_\_TK7\_po | L5 |  | 126 | 247 | 172 | 0.70 |  |
|  |  | A | 103 | 216 | 156 | 0.65 |  |

Samples where PRDM9 allele assignments were based on a range of min/mean depth read scores to assess the reliability of using depth read parameters for allele assignment.

This was done for the five (05) samples that were used for Sanger, Ion Torrent and MinION sequencing:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sanger |  |  |  |  |  |  |  |  |  |  |  |  |  | Remapping Assignment |  |
| Sample | PRDM9 1 | PRDM9 2 | DP\_min | DP\_max | DP\_mean | DPmax/mean | 0.80 | 0.75 | 0.70 | 0.65 | 0.60 | 0.50 | 0.40 | DP\_stdev | P1 | P2 |
| 013\_aboaus\_C\_AMD\_srt-flalleleA | A | A | 106 | 207 | 144 | 1.4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 20 | A | A |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 383\_aboaus\_C\_AMB\_srt-flalleleA | A |  | 48 | 193 | 125 | 1.5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 36 | ? |  |
| 383\_aboaus\_C\_AMB\_srt-flalleleL7 |  | L7 | 81 | 193 | 132 | 1.5 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 23 |  | L7 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 391\_kung\_A2\_GMO3043\_srt-flalleleGMO15 | GMO15 |  | 48 | 148 | 115 | 1.3 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 21 |  | L4 |
| 391\_kung\_A2\_GMO3043\_srt-flalleleC |  | C | 72 | 148 | 120 | 1.2 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 14 | C |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 394\_biaka\_\_K736\_srt-flalleleC | C |  | 31 | 150 | 92 | 1.6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 25 | ? |  |
| 394\_biaka\_\_K736\_srt-flalleleA |  | A | 45 | 137 | 97 | 1.4 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 22 |  | ? |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 409\_pal\_G2a\_4943\_srt-flalleleA | A |  | 67 | 413 | 206 | 2.0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 80 | ? |  |
| 409\_pal\_G2a\_4943\_srt-flalleleL4 |  | L4 | 58 | 241 | 153 | 1.6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 42 |  | ? |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

According to these results, PRDM9 assignment was not reliable in all cases. The loss of valuable reads from initial mapping to B allele in the reference genome may be partly responsible for this. But since some assignments proved true, a small survey is warranted. It may help in provide inside to what additional parameters could be combined to improve allele assignment.