SI Appendix

Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome

Alexandre Kuhn, Yao Min Ong, Ching-Yu Cheng, Tien Yin Wong, Stephen R. Quake, William F. Burkholder

Proc Natl Acad Sci U S A. 2014

Table of contents

1. Methods

Samples and construction of L1-seq libraries

We obtained 20 DNA samples (SI Appendix, Dataset S1) from individuals of multiethnic Asian origin (1, 2) and constructed L1-seq libraries based on the protocol of Ewing et al. (3). Briefly, we linearly amplified 3' flanking regions of L1Hs elements using an L1Hs-specific primer followed by a hemi-specific PCR using a set of eight degenerate primers (3) (see SI Appendix, Table S1 for PCR conditions). We made dilutions of the hemi-specific PCR reactions and used them in a second round of PCR which amplified the library and incorporated the appropriate Illumina sequencing adapters. After the PCR, we ran the products on a 2% TAE agarose gel and excised the fragments between 200 and 500 bp. We gel-extracted the DNA and pooled the eluted DNA from the eight reactions and further column-purified it. We next removed adenine overhangs by incubating with Pfu Turbo DNA polymerase (Agilent). Finally, we purified the blunt-ended DNA and analyzed it using the High Sensitivity Chip on the Agilent 2100 BioAnalyzer to determine the DNA concentration and quality. L1-seq libraries were sequenced on the Illumina HiSeq 2000 and the data were deposited as controlled-access data at dbGAP under accession number phs000732.v1.p1. This study was reviewed and approved by the National University of Singapore IRB, the Singapore Eye Research Institute IRB, and the SingHealth Centralised IRB. In addition, subjects gave informed consent prior to enrolment.

Computational pipeline for L1 calling

Reads with an average per base quality score of less than 15 or with more than 6 bases with a quality score less than or equal to 2 were filtered out. We trimmed the first 10 bases of each read because they represented the degenerate anchor part of the primers. Potential polyT sequences (corresponding to the polyA at the 3' end of the L1 sequence) were identified as 5 or more consecutive thymines and were trimmed. Processed reads that were less than 20 nucleotides long were eliminated. We then aligned the processed reads against the reference human genome (hg19) using bwa (4) and a maximum edit distance (option -n) of 4. We finally discarded non-uniquely aligned reads (i.e. reads that did not have the type XT:A:U) and eliminated reads with MAPQ score of 0 (using samtools (5) and the option -q 1).

Read pile-ups indicated L1 3' flanks and, for each sample, we identified all read pile-ups along with their position and coverage. Read pile-ups that were separated by less than 1000 bp were considered to mark the same L1 element and were merged into a single peak. For each peak, we recorded 4 parameters: width, mean and maximal coverage and number of unique reads (i.e. reads with the same start and end coordinates were considered as 1 read). L1s were called by setting thresholds on these 4 parameters. The sensitivity and specificity of L1 was adjusted by comparing L1 calls with the results of PCR validation experiments (see "PCR validation of L1 elements"). Read quality control, trimming and L1 calling were implemented in R (6) using Bioconductor (7) packages including ShortRead (8) and GenomicRanges (9).

PCR validation of L1 elements in our samples

We performed extensive PCR validation of L1-seq results. We developed a custom computational pipeline to design PCR primers to test for the presence of an L1 element at the specific genomic location predicted by L1-seq: for a given potential insertion, the PCR primer "G" specifically hybridized to the genomic region located upstream of the read pile up and was oriented so as amplify the targeted flanking region in conjunction with the L1Hs-specific primer originally used for library construction. As control and to test for the absence of the insertion, we also designed a primer "E" hybridizing to the genomic region downstream of the potential insertion. To decrease the total number of PCR reactions, we adopted the following 2-stage validation process: for each L1, we first ensured that primers worked as expected by testing both PCR reactions (i.e. detecting the presence and the absence of the particular L1) on a subset of the 20 samples. Using gel electrophoresis analysis, we checked that the PCR reactions yielded clear bands of the expected size. In particular, primers yielding smears or multiple bands were discarded. We successfully designed primers for 91 L1 elements absent from the reference genome and corresponding to L1 peaks with varying width and coverage, with varying numbers of unique reads and with varying levels of polymorphism across the 20 samples (Dataset S3). In a second stage, we tested for the presence of each insertion in each of the 20 samples based on the PCR reaction using the "G" primer and the L1Hs-specific primer. Dataset S4 in the SI Appendix provides PCR-based genotyping results (presence/absence) for 91 L1s across our 20 samples. SI Appendix contains gel electrophoresis analyses for all PCR validation reactions performed in the first and second stage of the validation process (see "3. Gel electrophoresis analyses of site-specific PCR validations for 91 L1s identified in our data"). We used the same PCR conditions for all reactions: 95 °C for 2 min, followed by 30 cycles of 95 °C for 30 sec, 57 °C for 30 sec, 73 °C for 1 min, followed by 73 °C for 2 min.

These PCR results allowed us to fine-tune the parameter thresholds used for peak calling in our L1-seq computational pipeline. Specifically, we performed single-parameters ROC analyses (using the R package pROC (10)) for peak width, mean peak coverage, maximal peak coverage and number of unique reads per peak (SI Appendix, Figure S2a-d) and found that peak coverage had the highest classifying performance, followed by number of unique peaks per read. We thus performed two-parameter ROC analyses based on these 2 parameters (SI Appendix, Figure S2e-j). Manual inspection of the relation between peak characteristics and PCR validation success, however, revealed that setting thresholds on peak width efficiently eliminated rare, but artifactual peaks of excessively long or short width. Setting fixed thresholds for peak width (comprised between 100 and 1000 bp), we performed a heuristic parameter search in the remaining other 2 dimensions (coverage and number of unique read per peak). Specifically, and in light of the downstream population genetic analyses, we favored a high specificity (around 95%) and varied the 2 parameters so as to maximize sensitivity. Final peak calling parameters were set as follows: maximal peak coverage > 80 and number of unique reads per peak > 4, resulting in a sensitivity of 76% for a specificity of 95%. To verify L1-SNP association results using a set of higher confidence L1s, we raised the number of unique reads per peak > 20 and identified a subset of 564 L1s with specificity of 98% (and sensitivity of 39%).

Despite the systematic and stringent validation approach followed here, sensitivity and specificity estimated from PCR validation are approximations only. Note that for particular candidate insertions, PCR primer design could fail altogether. This occurred for instance when the candidate insertion flanked a repetitive or a low-complexity region (and potentially resulted in candidate sampling bias). Moreover, in the case of (apparently) successful primer design, a single round of primer design was not always able to produce a primer that efficiently generated the targeted

amplicon. In these specific cases, additional rounds of primer design insertions sometimes allowed us to eventually validate the insertions. Such an optimization, however, could not be conducted for all candidate insertions. In summary, several intrinsic features of the PCR validation method impinged on the estimation of L1-seq performance.

SNP genotyping and quality control

SNP genotype data was previously obtained for 17 samples (CS60827, CS61814, CS57398, CS56899, CS58948, CS57848, CS50957, CS52543, CS52982, MS11015, MS11078, MS16259, MS15212, MS14984, MS24031, MS11355, MS13212) using Illumina Human610-Quad BeadChips. Quality control was performed as previously described (11).

Integrated L1-SNP analysis of our L1-seq data

L1-seq allowed us to call the presence (or absence) of L1 but it did not differentiate between individuals that were homozygous for L1 presence (i.e. with a particular insertion on the 2 chromosomes) and heterozygous. We tested the association between each detected L1 element and surrounding SNPs using Fisher's exact test. For each L1, we recorded the maximal $-\log_{10}$(p-value) as tagging score. We used a permutation approach to assess the significance of these scores and address the multiple testing issue. Specifically, we repeatedly performed the same analysis but permuted L1 presence/absence labels (100 permutations). The false discovery rate at a given score value was estimated by the ratio of the number of L1 that obtained an identical or higher score using the permuted data over the number of L1 that obtained an identical or higher score with the original data.

To compare the observed L1-SNP association with SNP-SNP association genome-wide, we calculated the association between randomly selected SNP and their surrounding SNPs. Specifically, we recoded SNP genotypes to mimic L1 detection by L1-seq. Thereby homozygotes for the ancestral allele were coded as 0 (absent) and heterozygotes and homozygotes for the derived allele were coded as 1 (present). We then randomly selected identically-sized sets of (recoded) SNPs presenting the same population frequency as L1s and tested association with their surrounding (non-recoded) SNPs in the same way as for L1-SNP association. Since L1 containing regions presented, on average less SNPs compared to randomly selected SNPs, we subsampled surrounding SNPs so as to have the same number as for L1s. Similarly as for the L1-SNP analysis, the significance of the SNP-SNP association for each SNP set (100 sets) was assessed using a permutation approach in which we randomly permuted the (recoded) SNP genotypes before calculating the association with surrounding SNPs (10 permutations were sufficient to obtain a stable FDR estimate). This analysis was implemented in R/Bioconductor (including packages snpStats and GenomicRanges(9)).

Integrated L1-SNP analysis of the 1000GP data

We obtained L1 genotype data for the 3 populations studied in the pilot phase of the 1000GP. Shortly, Stewart et al. (12) detected L1s as deletions ("deletion set", i.e. L1s present in the reference genome) or as insertions ("insertion set", i.e. L1s absent from the reference genome). For L1s detected as deletions, we used the genotype information provided in the VCF files CHBJPT.low_coverage.2010_10.deletions.genotypes.vcf, CEU.low_coverage.2010_10.deletions.genotypes.vcf, YRI.low_coverage.2010_10.deletions.genotypes.vcf. Following Stewart et al. (12), we identified L1 elements as deletions that matched L1 (L1Hs) sequences in the reference genome, within a window of 200 bp for both the start and end coordinates (a list of L1HS elements of the human genome in hg18 coordinate was obtained via the RepeatMasker Table of the UCSC genome browser). For L1 detected as insertions, we used genotype information provided in the file union.2010_06.MobileElementInsertions.genotypes.vcf and extracted mobile elements flagged as L1 (based on the tag <INS:ME:L1>). SNP genotypes were obtained from the files CHBJPT.low_coverage.2010_07.genotypes.vcf, CEU.low_coverage.2010_07.genotypes.vcf and YRI.low_coverage.2010_07.genotypes.vcf. The L1 information obtained from the 1000GP data provided us with the full genotype (and not only presence/absence information as for L1-seq) and LD between L1 elements and surrounding SNPs was assessed using the R^2 metric. The quality of L1 genotyping in the "deletion" and "insertion" sets was not directly comparable. Following Stewart et al, we focused on L1s with a high genotyping quality score (GQ ≥ 10 and GQ ≥ 7 for the "deletion" and "insertion" sets, respectively, see (12)). We eliminated L1s that had did not have at least 16 individuals with high quality genotypes. For each L1 set, we compared with the overall level of LD in the genome and calculated R^2 for sets of randomly selected SNPs (1000 sets). Each SNP set was jointly matched for (derived) allele frequency and sample size (SI Appendix, Figure S7). Thereby, each set contained SNPs of the same (derived) allele frequency spectrum as L1s and we also subsampled individuals for each SNP to match the number of samples used in the L1 analysis. For control SNPs corresponding to L1s in the "deletion" set, we left a gap of the corresponding L1 length around SNPs before selecting surrounding SNPs but it did not qualitatively change the results of the analysis. We verified that the number of SNPs around L1s and control SNPs was not significantly different. All analyses were implemented in R/Bioconductor (including packages VariantAnnotation (13), GenomicRanges (9), snpStats).

PCR validation of L1s in the 1000GP data

We obtained 40 CEU and 29 CHB HapMap (DNA) samples used by the 1000GP. We selected a set of L1s with max. R^2 values (obtained using 1000GP genotypes) approximately evenly distributed across the whole range from 0 to 1 and from both the "insertion" and "deletion" sets. We used our custom computational pipeline to design PCR primers for L1s of the "insertion" set. For L1s of the "deletion" set, we designed PCR primers by selecting 300 bp upstream and downstream of the L1 element and running the sequence through the Primer3 primer design program (14). Since the L1s of the "insertion" set were not present in the reference genome, their orientations were not known. Hence, in order to ensure that the targeted flanking region was amplified regardless of orientation, both designed primers upstream (primer "E") and downstream (primer "G") of the L1 element were used in conjunction with the L1Hs-specific primer in separate PCR reactions. The list of primers, gel electrophoresis analyses of PCR reactions and PCR-based genotypes are provided in SI Appendix (see

Dataset S5, "4. Gel electrophoresis analyses of site-specific PCR validations for 47 L1s from the 1000GP", and Dataset S6, respectively). PCR conditions were identical to those used for the validation of L1s in our samples.

Association of L1 LD with genomic features

The location of L1s was compared against gene boundaries obtained from the May 2009 archive of Ensembl (queried via biomaRt (15)). GC content around L1 elements (20 kb) and chromomsome length were obtained using the hg18 version of human genome available in the Bioconductor data package BSgenome.Hsapiens.UCSC.hg18. Centromere position was obtained from the gap Table of the UCSC genome browser. The number of elements belonging to the L1 or Alu repeat families and surrounding (20 kb) L1s genotyped by the 1000GP was extracted from the RepeatMasker table of the UCSC genome browser.

Extended haplotype homozygosity (EHH) around L1s

We tested for differences in the haplotypic structure around L1 elements based on the EHH statistic originally proposed by Sabeti et al. (16). For a SNP at distance x of the candidate variant under selection (i.e. in our case L1), the EHH statistic is the probability of drawing 2 alleles that are homozygous between the variant and the SNP (included) at distance x, from the population of alleles carrying the candidate variant. For comparison, the same EHH statistic is calculated for the ancestral allele, i.e. in our case for alleles that do not carry the specific L1 under investigation. Calculation of the EHH run statistic around a candidate variant requires that it is phased with surrounding markers which is not the case for the L1 genotype data provided by the 1000 GP. To circumvent this problem, we focused our analysis on L1s that had 5 or more homozygous individuals of both types (i.e. ≥ 5 homozygotes for the presence and ≥ 5 homozygotes for the absence of the particular L1), thereby avoiding the phasing problem. We thus used haplotypes obtained from L1 homozygotes to calculate the EHH run statistic for the L1-bearing allele and used haplotypes obtained from homozygotes for the absence of L1 to calculate the EHH run statistic for the allele without L1. Following Voight et al. (17), we summarized the difference between the two EHH run signals using the (log of the) ratio of the integrated EHH signals, i.e. log(iHS L1 / iHS no L1) where iHS is the integrated EHH signal (or integrated Haplotype homozygosity Score (17)). If log(iHS L1 / iHS no L1) = 0, integrated EHH signals in both groups are equal. Positive values indicate extended haplotype homozygosity (i.e. possibly positive selection) around the L1-bearing allele whereas negative values indicate the reverse.

We checked that the use of homozygotes to calculate EHH statistics would not result in systematic differences as compared to using all individuals. We thus selected a large number of random, frequency-matched SNPs and performed EHH calculations based on homozygous individuals only or including heterozygous individuals as well. This was possible in the case of SNPs as the phase between the test SNP and surrounding SNPs was known. We found that EHH scores (i.e log(iHS L1 / iHS no L1)) calculated using the 2 methods were highly correlated (Pearson's correlation coefficient = 0.95) and concluded that estimates based on alleles from homozygous individuals only did not present significant bias or noise as compared to estimates based on all alleles. We also asked if L1

length influenced the results of our analysis: For control SNPs corresponding to L1s in the "deletion" set, we left a gap corresponding to the matching L1 length around SNPs before selecting surrounding SNPs. EHH scores calculated with or without a gap were highly correlated (e.g. Pearson's correlation coefficient = 0.97 for 10 sets of 8 control SNPS in the CEU analysis) and their difference was not systematically different from 0.

To assess the significance of EHH scores obtained with L1s, we compared with scores obtained from SNPs genome-wide. We thus repeatedly selected identically-sized sets of SNPs that were matched for (derived) allele frequency. Based on 1000 sets of random SNPs, we calculated for each L1 the average number of SNPs (over sets) that obtained an (absolute) EHH score identical or higher than the (absolute) EHH score observed with the particular L1. The ratio of this number over the rank of the L1 in the set can be interpreted as a false discovery rate and estimates how unusual the observed score is compared to SNPs genome-wide. Since negative EHH score can be observed, we also calculated FDR based on signed EHH scores (in a two-sided manner). We obtained about the same significance levels as with absolute EHH scores.

We performed EHH analysis on all 9 L1s that had enough homozygotes in the CEU population. We could validate 8 of them (P1_MEI_1280&P2_MEI_1388, P1_M_061510_1_185, P1_M_061510_1_239, P1_M_061510_1_391, P1_M_061510_4_203, P1_M_061510_4_354, P1_M_061510_9_218, P1_M_061510_10_203) and used PCR-based genotypes to calculate EHH scores. Due to the smaller number of samples in the CHB population, only 3 L1s had enough homozygous individuals to be tested. We aimed to increase the number of L1s considered and used unfiltered genotypes (i.e. without excluding calls with low GQ scores) to screen for L1s with unusual EHH scores. We calculated EHH scores and significance of 18 L1s in the "insertion" set (i.e. all L1s with ≥ 6 homozygotes for L1 presence and ≥ 6 homozygotes for L1 absence) based on their (unfiltered) 1000GP genotypes. For significance calculations, random SNP sets were matched for allele frequency and sample size (i.e. for each frequency-matched SNP, we subsampled homozygous individuals to match the number of homozygous individuals actually used for the L1 analysis.). The top 6 L1s obtained unusual EHH scores and were validated using site-specific PCR (P1_MEI_2516&P2_MEI_1951, P1_MEI_539&P2_MEI_776, P1_MEI_190&P2_MEI_1442, P1_MEI_3095&P2_MEI_1256, P1_MEI_414&P2_MEI_691, P1_MEI_1381). We then used the PCR-based genotypes to obtain final EHH scores. The significance of these 6 L1s was established by comparison against sets of 18 (instead of 6) frequency-matched, random SNPs (to prevent overestimation of significance resulting from pre-screening 18 L1s).

L1 frequency stratification in the 1000GP

We used the fixation index Fst to compare L1 allele frequencies between the CEU and YRI populations. Allele frequencies were obtained based on the (GQ-filtered) 1000GP genotypes. We did not consider the CHB population as the smaller sample size and the smaller number of genotyped L1s did not allow for a comparable analysis. We first assessed the significance of Fst values by permutation. We repeatedly permuted the population labels of individuals and recalculated Fst (1000 permutations). We corrected for multiple testing of L1s using an FDR approach (18). We also compared the distribution of Fst values obtained with L1s against the distribution of Fst values obtained with SNPs genome-wide and generated 1000 sets of random SNPs where each set was

jointly matched for L1 (derived) allele frequency (averaged over both populations) and sample size. All analyses were implemented in R/Bioconductor (including packages VariantAnnotation (13), GenomicRanges (9), snpStats).

References

1. Foong AWP et al. (2007) Rationale and methodology for a population-based study of eye diseases in Malay people: The Singapore Malay eye study (SiMES). *Ophthalmic Epidemiol* 14:25–35.

2. Lavanya R et al. (2009) Methodology of the Singapore Indian Chinese Cohort (SICC) eye study: quantifying ethnic variations in the epidemiology of eye diseases in Asians. *Ophthalmic Epidemiol* 16:325–336.

3. Ewing AD, Kazazian HH Jr (2010) High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* 20:1262–1270.

4. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

5. Li H et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

6. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria) Available at: http://www.R-project.org.

7. Gentleman RC et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5:R80.

8. Morgan M et al. (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* 25:2607–2608.

9. Lawrence M et al. (2013) Software for computing and annotating genomic ranges. *PLoS Comput Biol* 9:e1003118.

10. Robin X et al. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77.

11. Cornes BK et al. (2012) Identification of four novel variants that influence central corneal thickness in multi-ethnic Asian populations. *Hum Mol Genet* 21:437–445.

12. Stewart C et al. (2011) A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* 7:e1002236.

13. Obenchain V et al. (2014) VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*.

14. Untergasser A et al. (2012) Primer3--new capabilities and interfaces. *Nucleic Acids Res* 40:e115.

15. Durinck S et al. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21:3439–3440.

16. Sabeti PC et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

17. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4:e72.

18. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J Roy Statist Soc Ser B* 57:289–300.

2. Supplementary Tables, Datasets and Figures

Supplementary table captions

Table S1: PCR conditions for L1-seq library construction. a) Linear amplification and first-round PCR. b) Second-round PCR.

Table S2: List of L1 elements with significant (FDR < 0.05) Fst between CEU and YRI samples and present in genes. AF is allele frequency, p-value is the nominal p-value obtained using permutation testing (1000 permutations). L1 entries with identical start and end coordinates belong to the "insertion" set (i.e. L1 absent from the reference genome), for which the insertion lengths were not determined by the 1000GP.

Supplementary dataset captions

Dataset S1: Sample names and phenotypic information. For each sample, the table provides unique sample identifier (SERI identifier), study-specific sample name (sample name) and corresponding L1-seq library name (sequencing identifier), as well age, gender and race information.

Dataset S2: Location and peak statistics for L1s identified in our L1-seq data. "Chromosome", "start", "end" and "strand" correspond to features of (read) peaks obtained by L1-seq. Since L1-seq reads tag L1 3' ends, L1s are predicted to be located downstream of peaks (up to a few 100 bp) and on the opposite strand. "Mean coverage" and "Number of unique reads" are (per sample) averages over all samples in which the particular L1 was detected. "Frequency" indicates the number of samples (between 1 and 20) in which the L1 was detected. "KR_KNR" indicates if the L1 matches a known reference (KR) or a known nonreference (KNR) L1 (according to the lists of KR and KNR L1s in Table 5 of Evrony et al.). We defined a match as a strand-specific overlap between a region 100 bp upstream and 300 bp downstream of the KR or KNR L1 and an L1 peak. "High specificity set" indicates which L1s belong to the set of higher-specificity L1s used as control in Figure S4b.

Dataset S3: Primers used for PCR validation of L1s identified in our data. "EG" refers to the location of the primer. "E" identifies the primer on the L1 5' flank (on the same strand as L1). "G" identifies the primer on the L1 3' flank (on the opposite strand). "L1 name" indicate the L1 targeted for validation.

Dataset S4: PCR-based genotypes obtained for 91 L1s in our 20 samples. The first row (column header) provides sample names and the first column provides L1 names. "0" indicates absence and "1" indicates presence of the L1. L1s and corresponding PCR primers used for validation are provided in Dataset S2 and Dataset S3, respectively. Analyses by gel electrophoresis for all 91 L1s are shown in SI Appendix (see "3. Gel electrophoresis analyses of site-specific PCR validations for 91 L1s identified in our data").

Dataset S5: Primers used for PCR validation of L1s of the 1000GP. "EG" refers to the location of the primer. "E" identifies the primer on the L1 5' flank (on the same strand as L1). "G" identifies the primer on the L1 3' flank (on the opposite strand). "L1 name" indicate the L1 targeted for validation.

Dataset S6: PCR-based genotypes obtained for 47 L1s of the 1000GP. The first row (column header) provides L1 names and the first column provides sample names. The first 40 samples are CEU individuals and the next 29 samples are CHB individuals. "0" indicates homozygotes for the absence of L1, "1" indicates heterozygotes and "2" indicates homzygotes for the presence of L1. Corresponding PCR are listed in Dataset S5 and analyses by gel electrophoresis for all 47 L1s are shown in SI Appendix (see "4. Gel electrophoresis analyses of site-specific PCR validations for 47 L1s from the 1000GP").

Dataset S7: List of perfect tagging SNPs (max. $R^2 = 1$) identified from 40 CEU samples and 29 CHB samples in the 1000GP data.

Supplementary figure captions

Figure S1: Schematic representation of study methodology. **a**: We performed genome-wide profiling of L1 elements (L1-seq) in 20 Asian individuals and parallel SNP genotyping (Illumina BeadChips) in 17 of the 20. Blue and black ticks on chromosomes represent L1 and SNP locations, respectively. **b**: We characterized the performance of our L1 profiling method by combined analysis of L1-seq data and PCR validation results. L1-seq library reads are shown for a novel, polymorphic L1 insertion on chromosome 2. Reads indicate that the L1 insertion is present in samples 1 and 2 but not in sample 3 (top). However, accurately calling the presence or absence of an L1 could be affected by the greatly variable read coverage across samples (in this example, maximal coverage was 18,691 and 889 reads in samples 1 and 2, respectively). We therefore used PCR validation to verify the presence of the polymorphic L1 in samples 1 and 2 and its absence in sample 3 (bottom left). PCR validation across all 20 Asian samples of 91 L1 insertions predicted by L1-seq confirmed the high sensitivity and specificity of the method (example ROC analysis, bottom right).

Figure S2: ROC analysis of L1 detection by L1-seq using 108 L1s validated by PCR. **a-d**: We performed single-parameter ROC analyses for each of the 4 peak parameters, i.e. peak width (a), mean (b) and maximal coverage (c) and number of unique reads (d). For each parameter, the area under the curve (AUC), parameter threshold, specificity and sensitivity for the point with the maximum sum of specificity and sensitivity are indicated. **e-g**: 2-D ROC analyses of mean coverage (x-axis) and number of unique reads (y-axis). The plots show the simultaneous effect of these two peak parameters on sensitivity (e), specificity (f) and their sum (g). **h-j**: 2-D ROC analyses of maximal coverage and number of unique reads. The plots show the simultaneous effect of the two peak parameters on sensitivity (h), specificity (i) and their sum (j).

Figure S3: Specificity and sensitivity of L1 detection by L1-seq for 108 L1s stratified by frequency. **a**: Sensitivity stratified by frequency. Open circles represent L1s. Sensitivity and frequency values are slightly jittered to visualize data point density (standard deviation of frequency and sensitivity jitters are 0.05 and 0.005, respectively). Green ticks indicate the mean of L1s at each frequency. **b**: same as (a) for specificity.

Figure S4: L1-SNP and SNP-SNP taggability in our data. **a**: Distribution of SNP-SNP taggability measured using R^2 for identically-sized sets of frequency-matched SNPs. More than 50% of SNPs were perfectly tagged (max. $R^2 = 1$) by neighbouring SNPs, in line with the high significance of

tagging scores obtained using Fisher's exact test. Error bars indicate 10th and 90th percentile (10 sets). **b**: L1-SNP association (red) of a set of 564 high confidence L1s with at least 1 non-monoallelic SNP in the surrounding 100 kb (FDR estimation is identical to Figure 1d). The significance of L1-SNP association was much smaller compared to frequency-matched SNPs (black). **c**: L1-SNP association for the subset of 681 L1s present in 2 to 15 individuals with at least 1 non-monoallelic SNP in the surrounding 100 kb (red). The significance of L1-SNP association was much smaller compared to frequency-matched SNPs (black). **d**: A small subset of L1s expected to be efficiently tagged on this array showed highly significant FDR (red). We identified 43 perfectly tagged L1s (max. $R^2 = 1$) and their tagging SNPs in individuals of Asian descent (CHB samples) of the 1000GP (SI Appendix, Dataset S7). We translated the coordinates of these L1s to hg19 (using the Bioconductor package rtracklayer) and matched them to L1s identified from our L1-seq data. A subset of 12 L1s that was detected in our L1-seq data, was polymorphic across the 17 samples and had at least one tagging SNP on the array was used as positive control and tested for taggability. L1-SNP association (red) was highly significant (10^{th} and 90^{th} percentile of FDR were 0 for the top 8 L1s) ad was similar to the SNP-SNP association (black).

Figure S5: Errors in L1 calls did not critically affect L1-SNP association in our data. **a**: L1-SNP association obtained using PCR-based genotypes (x-axis) and L1-seq-based genotypes (y-axis) for 44 polymorphic L1s validated by site-specific PCR and with at least 1 non-monoallelic SNP in the surrounding 100 kb. Gray level encodes the number of overlapping points as indicated in the legend (top-left corner of the plot). **b**: FDR for L1-SNP association obtained using PCR-based genotypes (red) and L1-seq-based genotypes (black).

Figure S6: L1 taggability for 3 human populations (CEU, CHB and YRI samples in the 1000 Genome Project, pilot phase data). **a**: Inverse cumulative distribution of the maximal R^2 observed between 67 L1s ("deletion" set) and SNPs in their surrounding 20 kb region (containing on average 70 SNPs) (red line), or random SNP sets (1000 sets of 67 SNPs) and SNPs in their surrounding 20 kb region (containing on average 65 SNPs) (black line). Each SNP set used to calculate SNP-SNP association comprised randomly selected SNP sets that had a similar joint derived allele frequency and sample size distribution as for L1s. Error bars show 10^{th} and 90^{th} percentile obtained from 100 SNP sets. **b**: same as a for 102 L1s of the "insertion" set in the CEU population. **c**: same as a for 60 L1s ("deletion" set) in the CHB population. **d**: same as a for 24 L1s ("insertion" set) in the CHB population. **e**: same as a for 78 L1s ("deletion"n set) in the YRI population. **f**: same as a for 110 L1s ("insertion" set) in the YRI population.

Figure S7: Allele frequency and sample size for L1s and comparison SNPs used for LD analysis of the CEU population in the 1000 GP. **a**: Allele frequency versus sample size for 67 L1s of the "deletion" set. L1s are preferentially of high allele frequency and have high quality genotype in most samples. **b**: Derived allele frequency versus sample size for an random SNP set used to compare against L1s of the deletion set (LD analysis). The joint distribution of allele frequency and sample size matched the distribution observed for L1s (a). **c**: Same as (b) but for 102 L1s of the "insertion" set. L1 have lower allele frequency and lower genotyping quality score resulting in smaller sample size, compared to the deletion set (a). **d**: Derived allele frequency versus sample size for a random SNP set used to compare against L1s of the insertion set (LD analysis). The joint distribution of allele frequency and sample size matched the distribution observed for L1s (c). Sample size was slightly jittered (standard deviation = 0.05) in all panels to prevent overlaps of data points.

Figure S8: L1 taggability assessed using 1000GP genotypes is underestimated due to L1 genotyping errors. **a**: Maximal R^2 calculated using 1000GP genotypes (x-axis) and PCR-based genotypes (y-axis) for 41 L1s validated using site-specific PCR in 40 CEU samples. Validated L1s were selected to be approximately evenly distributed across the whole range of maximal R^2 values obtained using 1000GP genotypes. The gray level of dots encodes the number of overlapping points (3 data points overlap at (1,1)). **b**: same as (a) for 8 L1 validated in 29 CHB samples (5 data point overlap at (1,1)).

Figure S9: Genotype quality (GQ) scores for correct and wrong genotype calls are similarly distributed (42 PCR-validated L1s in the CEU samples). A further increase of the GQ score threshold is thus unlikely to yield a subset of L1s with less erroneous genotyping calls. **a**: Distribution of GQ scores for correct genotypes for 22 validated L1s from the "insertion" set in the 1000GP. **b**: Same as (a) for 20 validated L1s from the "deletion" set in the 1000GP. **c**: Distribution of GQ scores for wrong genotypes for the 22 validated L1s from the "insertion" set. **d**: same as (c) for the 20 validated L1s from the "insertion" set. The comparison of (a) and (c) for the "deletion" set and (b) and (d) for the "insertion" set shows that, in both sets, distributions of GQ scores are very similar for correct and wrong calls.

Table S1

a

Step	Temperature (°C)	Duration (min)	Notes
1	95	2:30	
2	95	0:30	
3	65	0:30	
4	72	0:10	
5			Go to Step 2, 19x (20 cycles)
6	-60	(hold)	Add degenerate primers
7	95	0:30	
8	43	0:30	
9	72	0:30	
10			Go to Step 7, 9x (10 cycles)
11	95	0:30	
12	55	0:30	
13	72	0:30	
14			Go to Step 11, 9x (10 cycles)
15	72	5:00	
16	4	hold	

b

Step	Temperature (°C)	Duration (min)	Notes
1	95	2:30	
2	95	0:30	
3	60	0:30	
4	72	1:00	
5			Go to Step 2, 29x (30 cycles)
6	72	5:00	
7	4	hold	

Table S2

Chromosome	L1 start	L1 end	AF (CEU)	AF (YRI)	Fst	p-value	Gene symbol	Entrez gene ID
3	131830256	131835798	0.8	0.279412	0.267964	<0.001	131873	COL6A6
10	52733103	52733103	0	0.157895	0.229257	0.012	5592	PRKG1
15	27198766	27199188	0.229167	0.672414	0.202772	<0.001	321	APBA2
18	46165017	46165282	0.987179	0.685714	0.185244	<0.001	220134	SKA1
3	126372743	126373518	0.1	0.410714	0.138629	<0.001	84561	SLC12A8
8	17903777	17904702	0.9625	0.708333	0.131408	<0.001	5108	PCM1
12	116671308	116671308	0.166667	0.5	0.12983	0.003	283455	KSR2
17	26683717	26684067	0.455882	0.803571	0.122843	<0.001	4763	NF1
7	30445373	30451439	1	0.864865	0.106637	<0.001	10392	NOD1

Figure S1

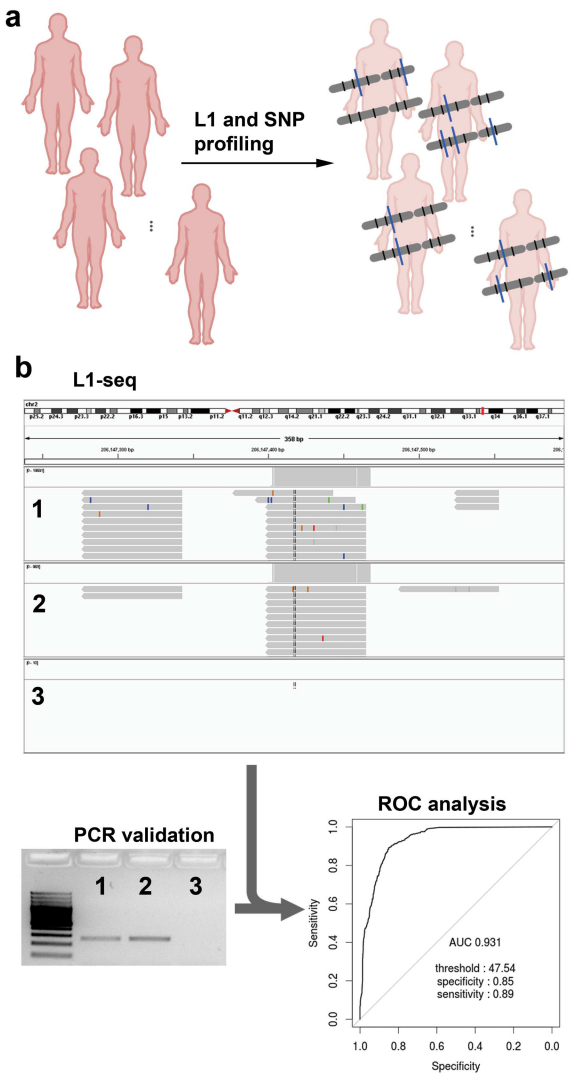

a
L1 and SNP profiling
b
L1-seq
1
2
3
PCR validation
1 2 3
ROC analysis
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC 0.931
threshold : 47.54
specificity : 0.85
sensitivity : 0.89

Figure S2

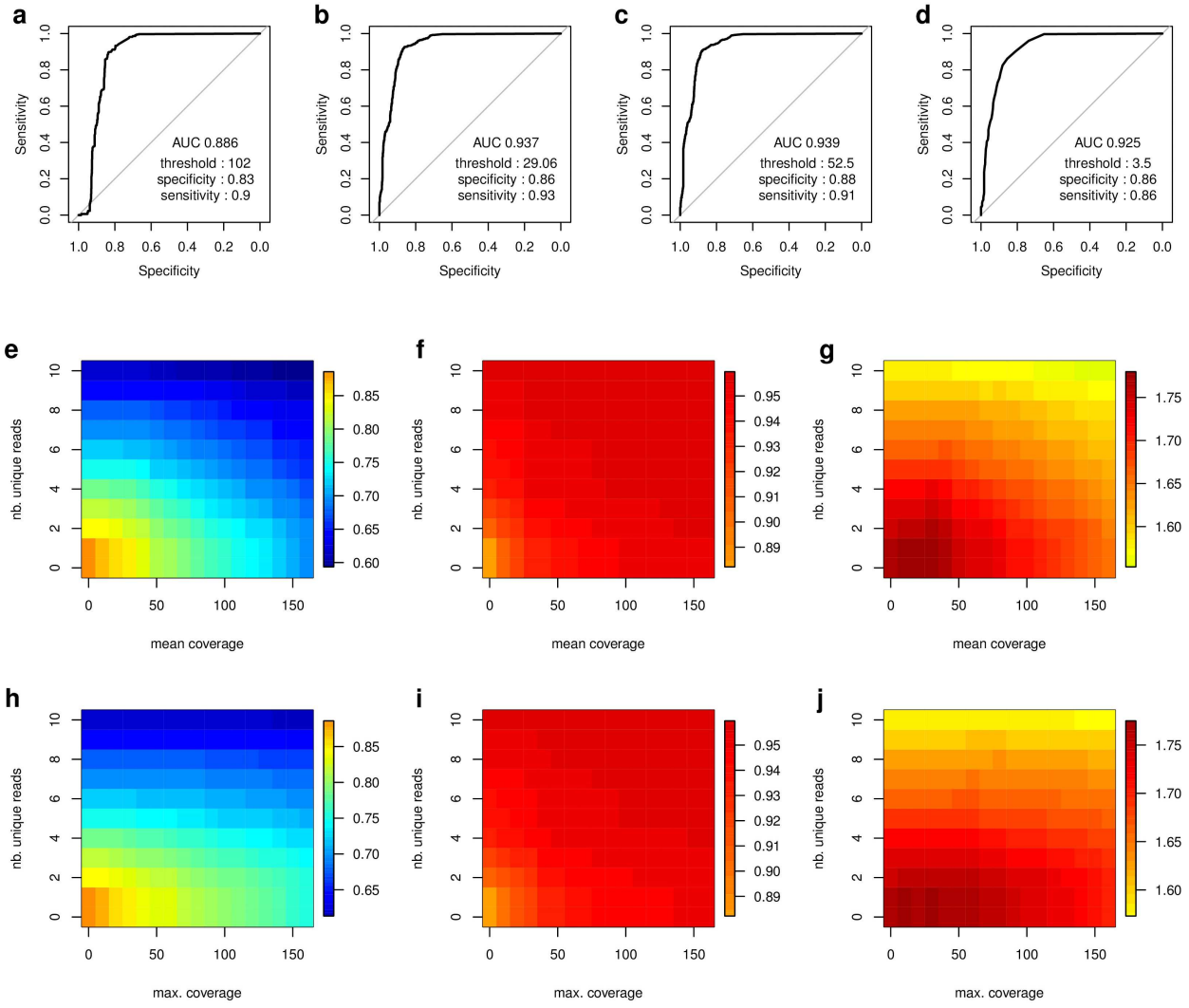

a
Sensitivity
Specificity
AUC 0.886
threshold : 102
specificity : 0.83
sensitivity : 0.9
b
Sensitivity
Specificity
AUC 0.937
threshold : 29.06
specificity : 0.86
sensitivity : 0.93
c
Sensitivity
Specificity
AUC 0.939
threshold : 52.5
specificity : 0.88
sensitivity : 0.91
d
Sensitivity
Specificity
AUC 0.925
threshold : 3.5
specificity : 0.86
sensitivity : 0.86
e
nb. unique reads
mean coverage
f
nb. unique reads
mean coverage
g
nb. unique reads
mean coverage
h
nb. unique reads
max. coverage
i
nb. unique reads
max. coverage
j
nb. unique reads
max. coverage

Figure S3

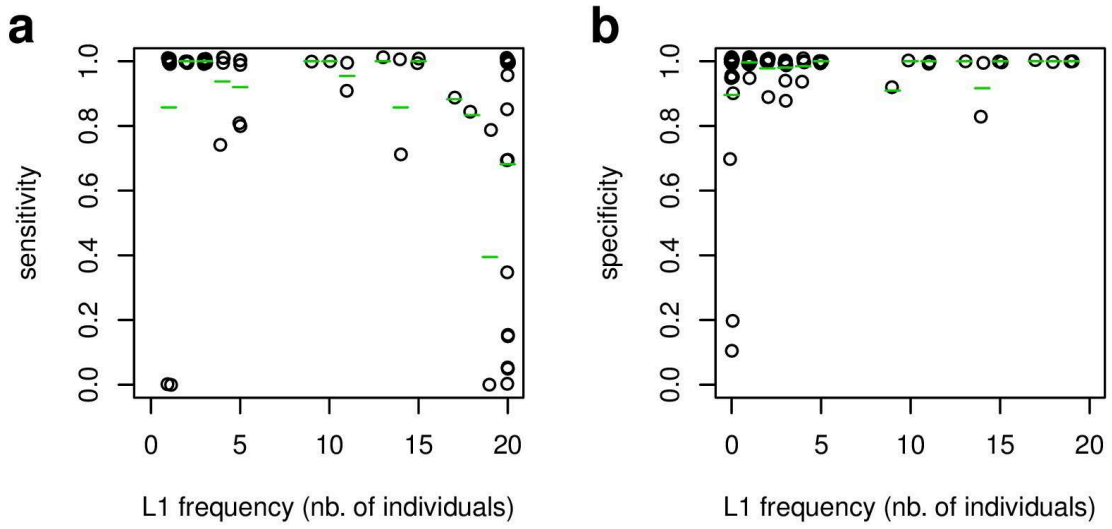

a
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
L1 frequency (nb. of individuals)
b
specificity
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
L1 frequency (nb. of individuals)

Figure S4

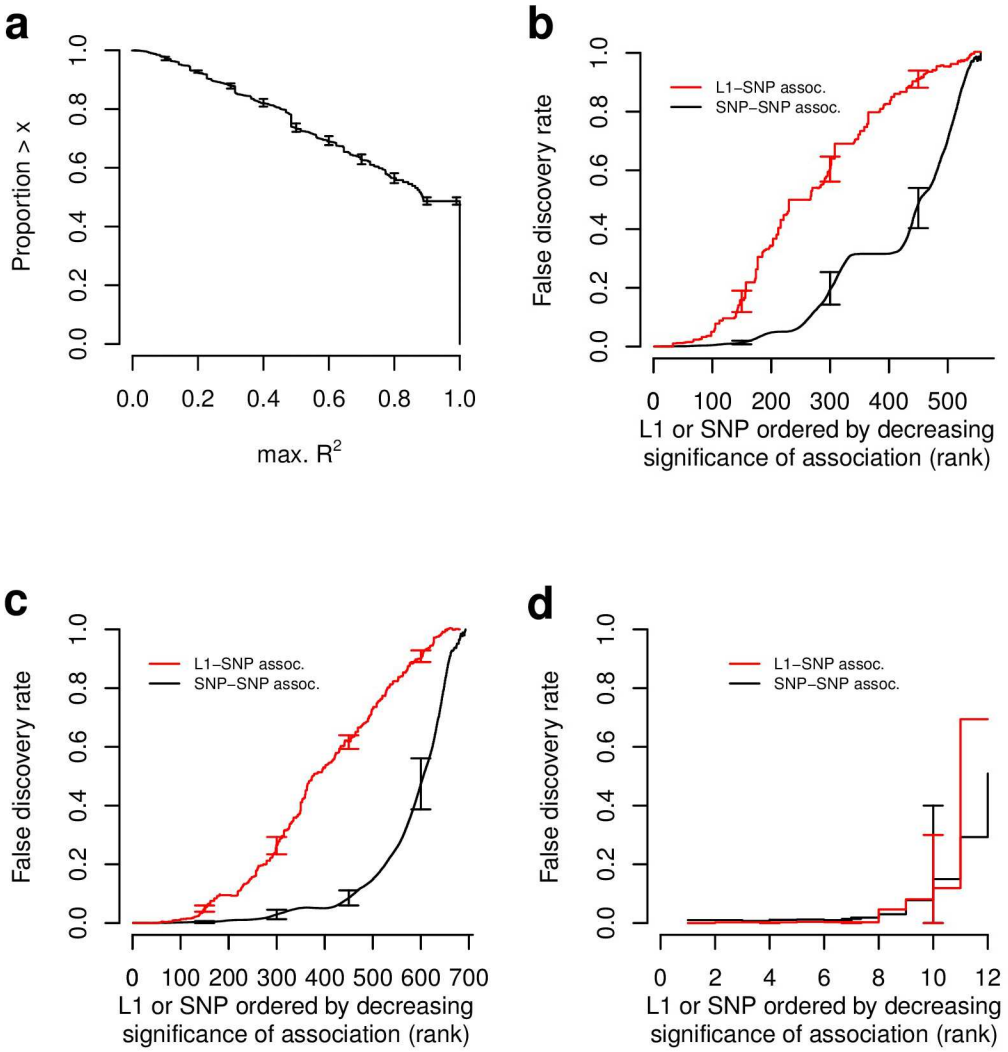

a
Proportion > x
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
max. R^2
b
False discovery rate
L1–SNP assoc.
SNP–SNP assoc.
0
100
200
300
400
500
L1 or SNP ordered by decreasing significance of association (rank)
c
False discovery rate
L1–SNP assoc.
SNP–SNP assoc.
0
100
200
300
400
500
600
700
L1 or SNP ordered by decreasing significance of association (rank)
d
False discovery rate
L1–SNP assoc.
SNP–SNP assoc.
0
2
4
6
8
10
12
L1 or SNP ordered by decreasing significance of association (rank)

Figure S5

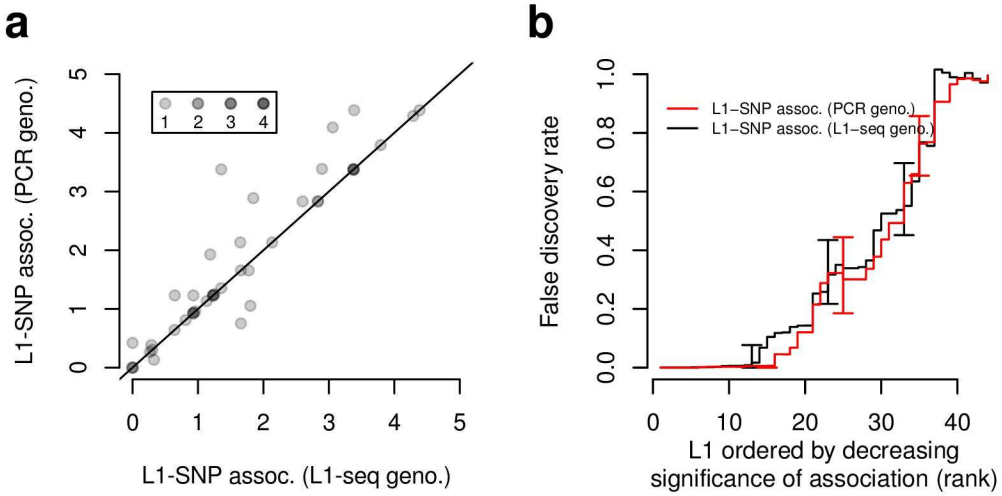

a
L1-SNP assoc. (PCR geno.)
L1-SNP assoc. (L1-seq geno.)
1 2 3 4
0 1 2 3 4 5
b
False discovery rate
L1−SNP assoc. (PCR geno.)
L1−SNP assoc. (L1−seq geno.)
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40
L1 ordered by decreasing
significance of association (rank)

Figure S6

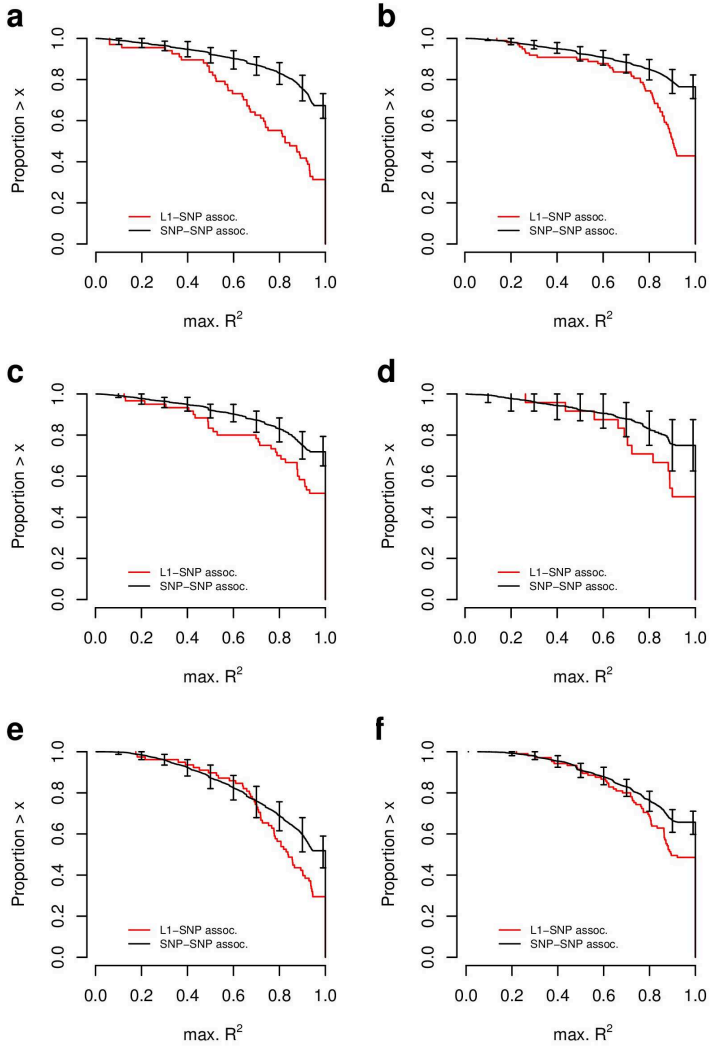

a
b
c
d
e
f
Proportion > x
max. R²
0.0
0.2
0.4
0.6
0.8
1.0
L1–SNP assoc.
SNP–SNP assoc.

Figure S7

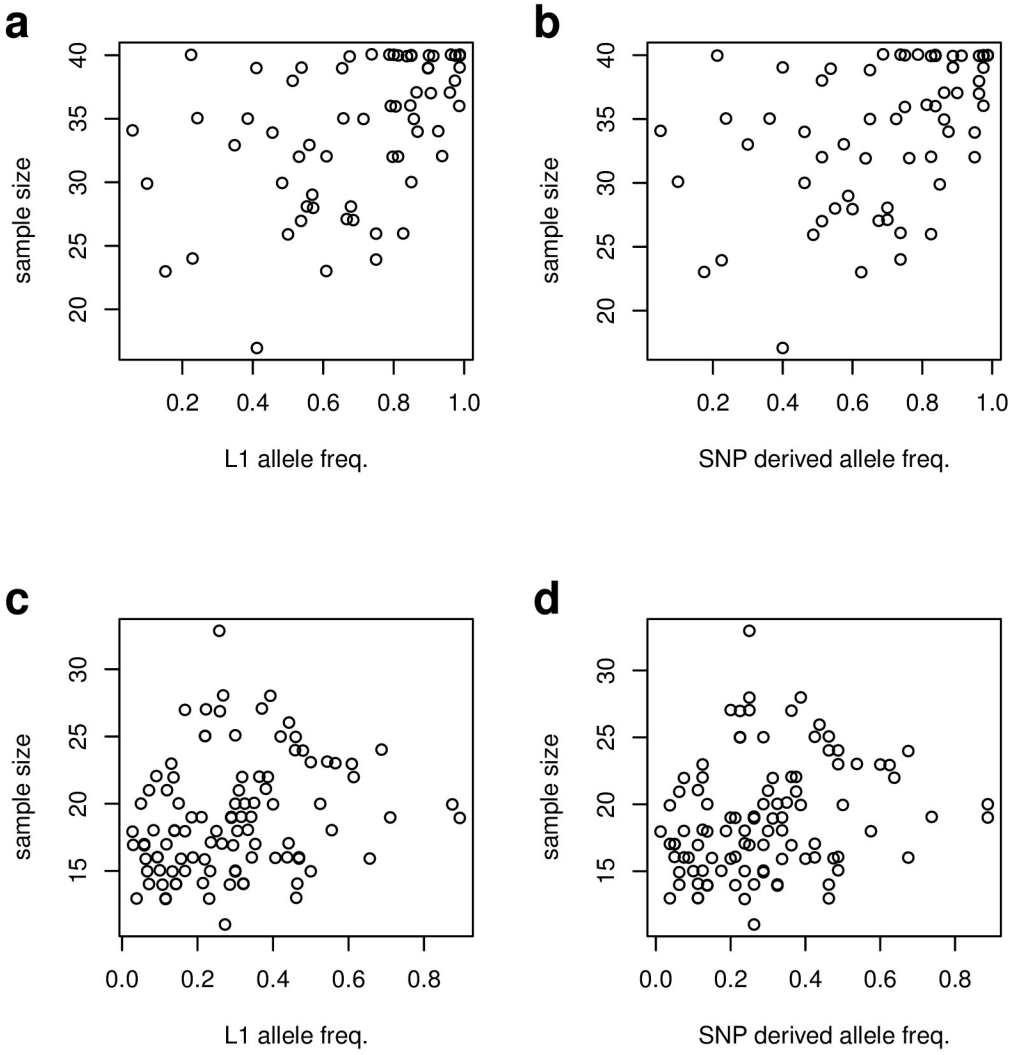

a
sample size
20
25
30
35
40
0.2
0.4
0.6
0.8
1.0
L1 allele freq.
b
sample size
20
25
30
35
40
0.2
0.4
0.6
0.8
1.0
SNP derived allele freq.
c
sample size
15
20
25
30
0.0
0.2
0.4
0.6
0.8
L1 allele freq.
d
sample size
15
20
25
30
0.0
0.2
0.4
0.6
0.8
SNP derived allele freq.

Figure S8

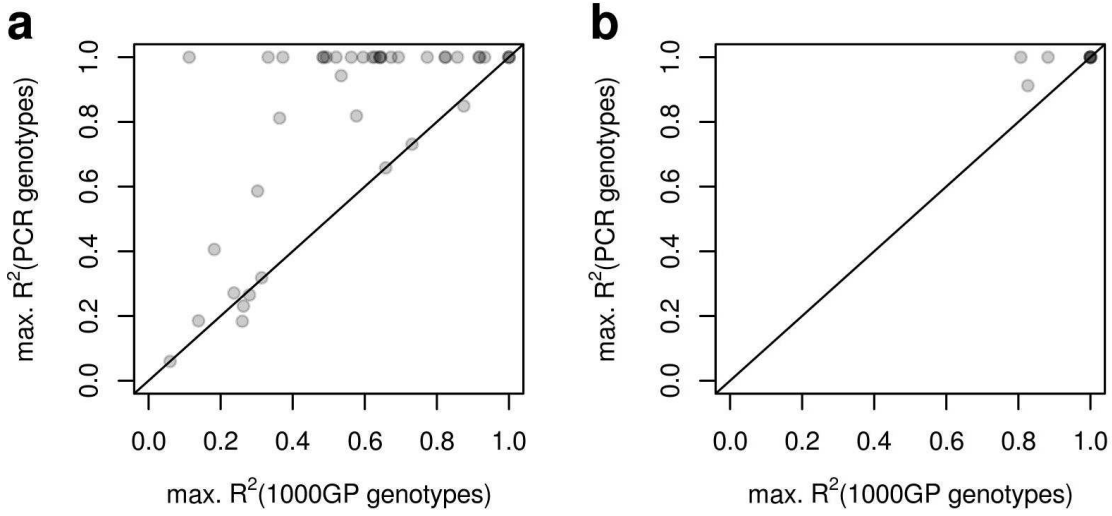
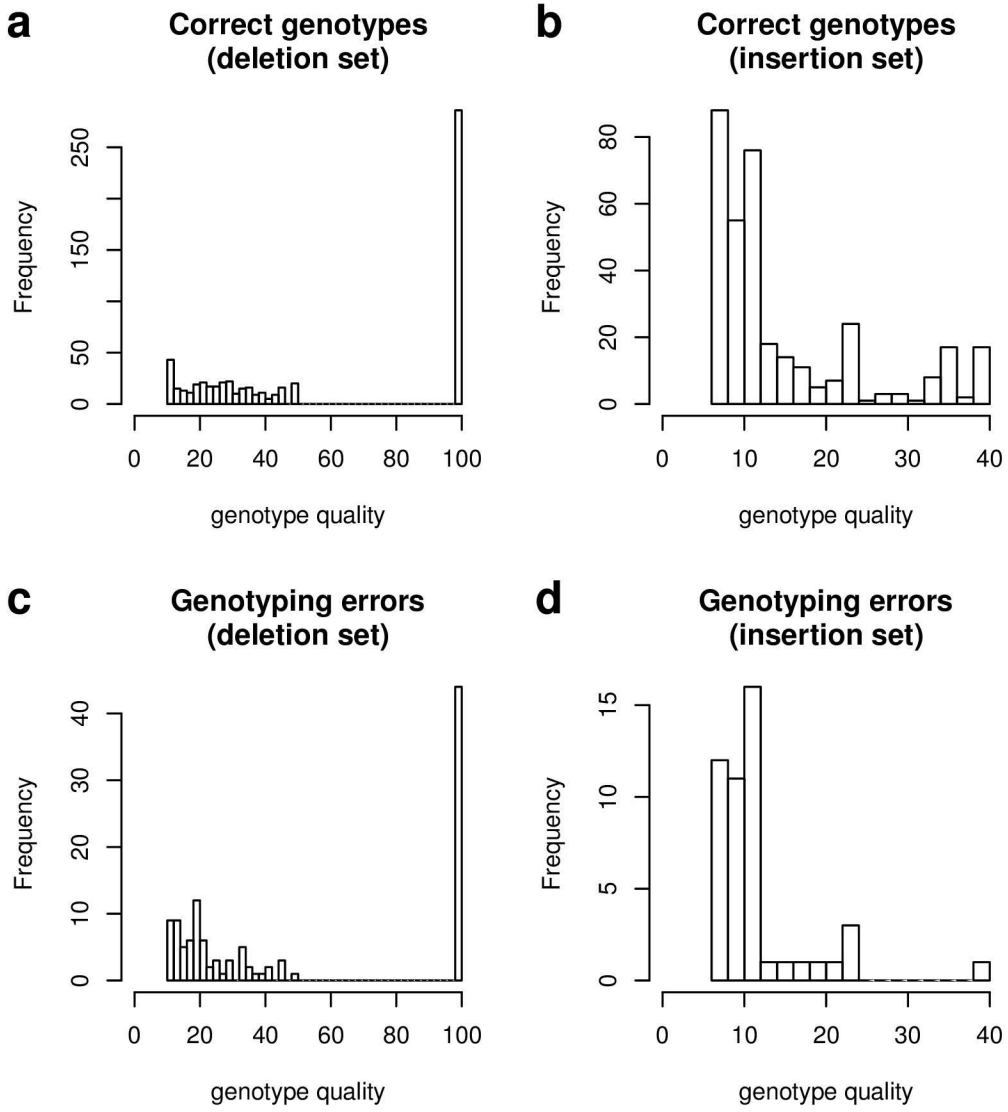

a
b
max. R2(PCR genotypes)
max. R2(1000GP genotypes)
0.0
0.2
0.4
0.6
0.8
1.0

Figure S9


a
Correct genotypes
(deletion set)
Frequency
0
50
150
250
0
20
40
60
80
100
genotype quality
b
Correct genotypes
(insertion set)
Frequency
0
20
40
60
80
0
10
20
30
40
genotype quality
c
Genotyping errors
(deletion set)
Frequency
0
10
20
30
40
0
20
40
60
80
100
genotype quality
d
Genotyping errors
(insertion set)
Frequency
0
5
10
15
0
10
20
30
40
genotype quality

3. Gel electrophoresis analyses of site-specific PCR validations for 91 L1s identified in our data

Section A: First validation stage (pages 2-17)

This section presents gel electrophoresis analyses of PCR reactions used to verify the specificity of primers used to validate 91 L1s identified in our L1-seq data for the Asian cohort samples. For each L1, we tested 2 PCR reactions in a subset of our 20 samples (generally including samples predicted to contain the tested L1 insertion):

1) Reactions labelled "E" test for the absence of the L1 insertion (primers E and G).
2) Reactions labelled "G" test for the presence of the L1 insertion (L1Hs-specific primer and primer G).

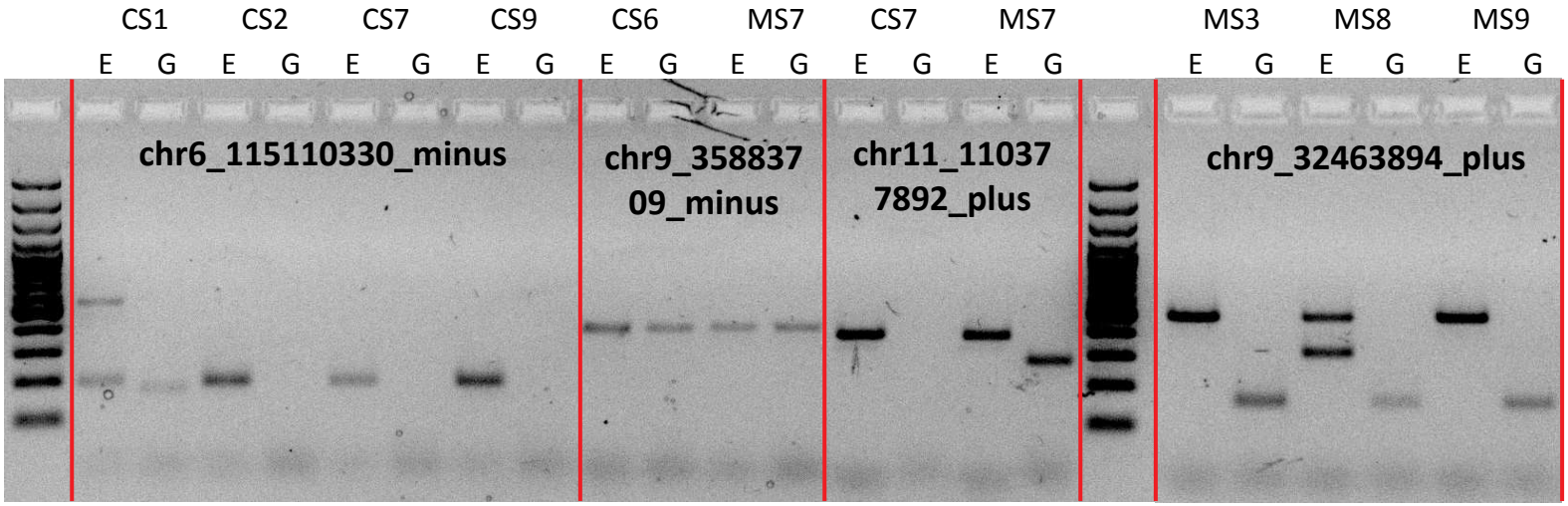
For each L1 insertion, the subset of samples used to test primers is indicated in the corresponding table (tables on pages 3, 5, 9, 14, 16).

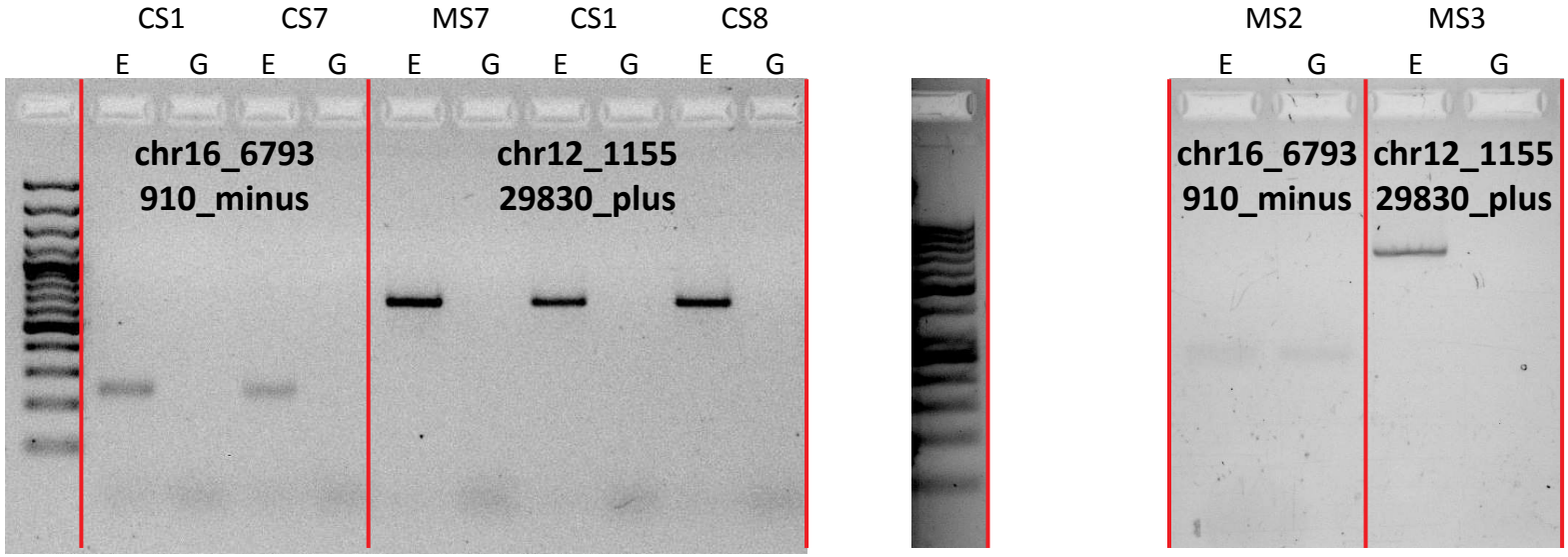
Section B: Second validation stage (pages 18-47)

This section contains gel electrophoresis analysis of PCR validation reactions used to test for the presence of L1 (L1Hs-specific primer and primer G) in each of our 20 samples.

Section A

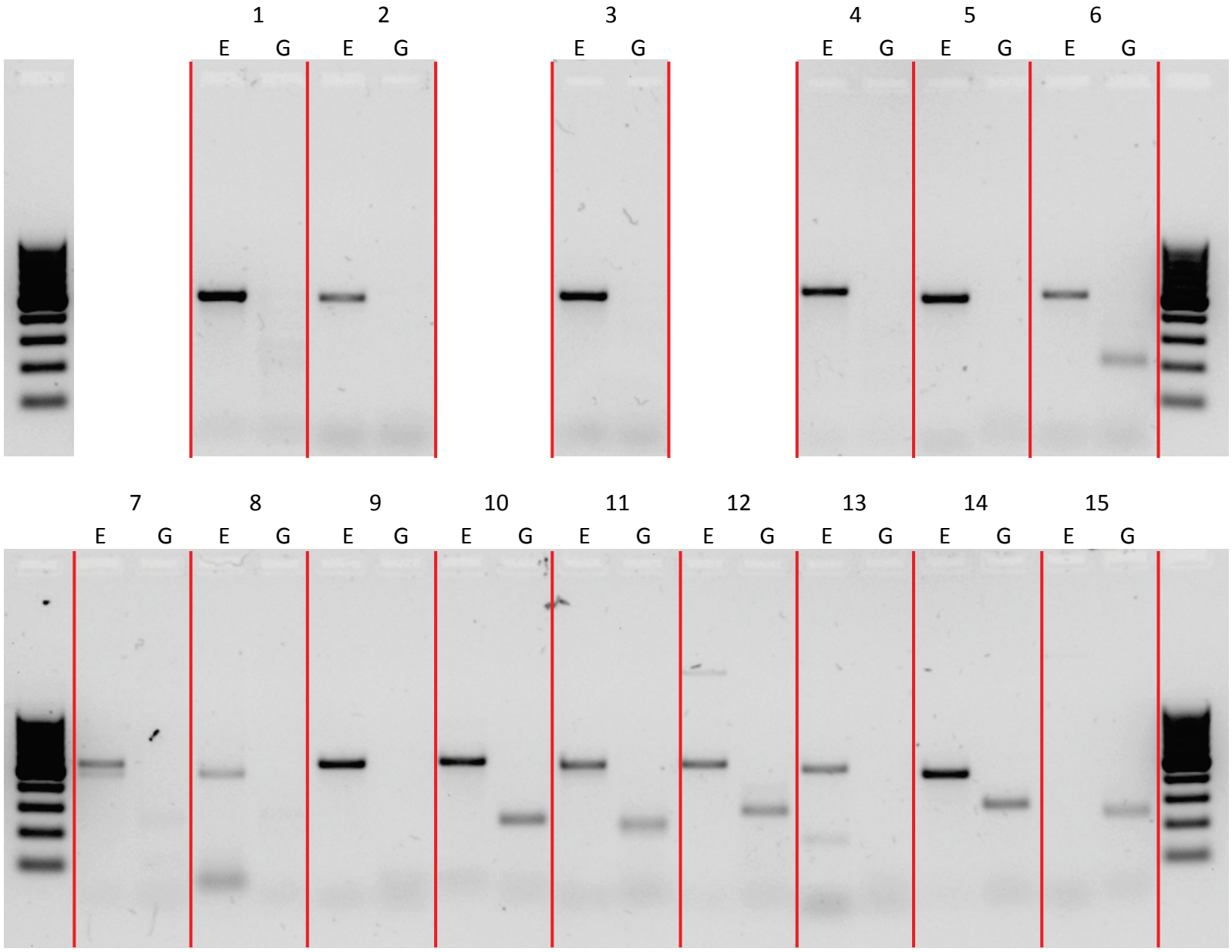
No.	Site	Samples			
1	chr6_115110330_minus	CS1	CS2	CS7	CS9
2	chr9_35883709_minus	CS6	MS7		
3	chr11_110377892_plus	CS7	MS7		
4	chr9_32463894_plus	MS3	MS8	MS9	
5	chr16_6793910_minus	CS1	CS7	MS7	
6	chr12_115529830_plus	CS1	CS8	MS2	MS3


CS1 CS2 CS7 CS9 CS6 MS7 CS7 MS7 MS3 MS8 MS9
E G E G E G E G E G E G E G E G E G E G E G
chr6_115110330_minus
chr9_35883709_minus
chr11_110377892_plus
chr9_32463894_plus

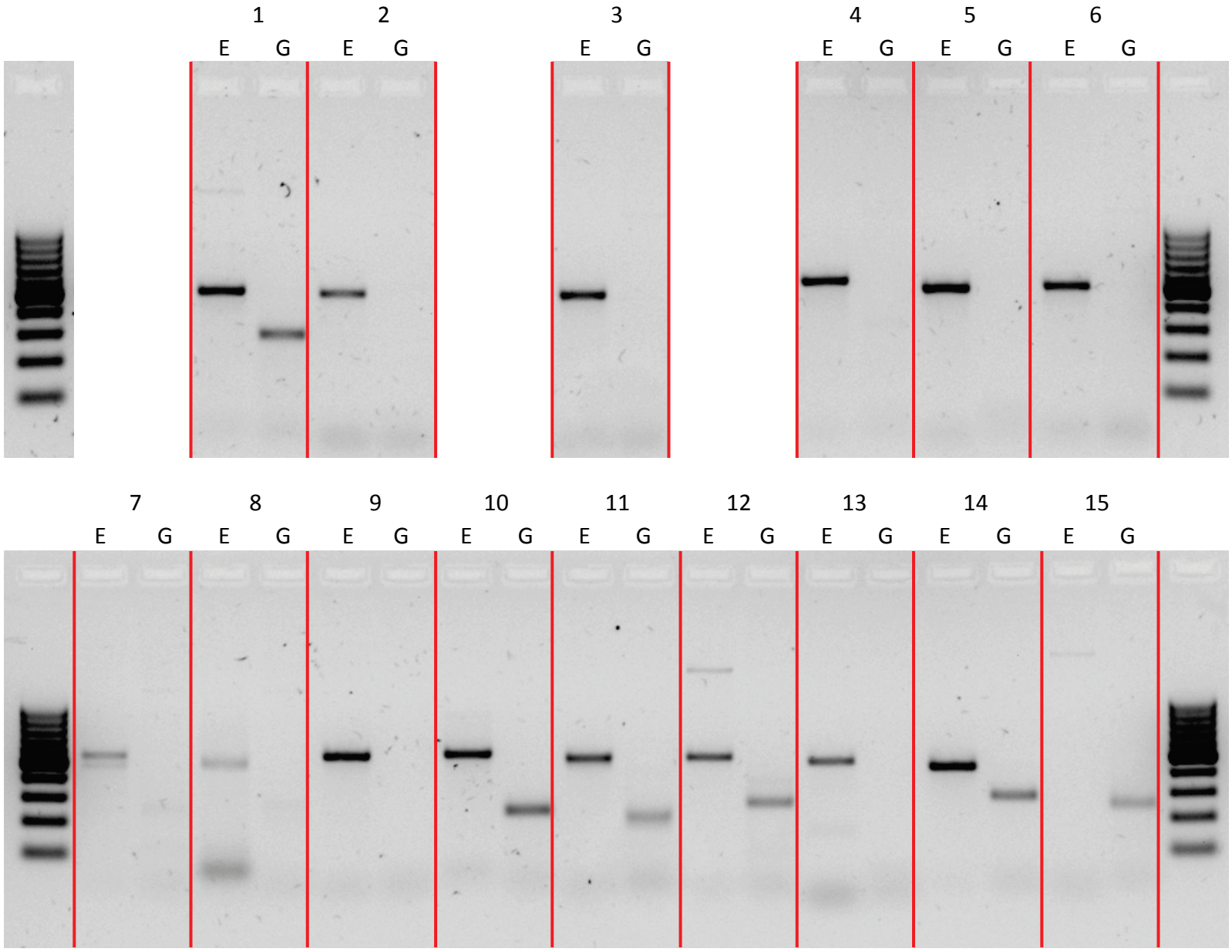

CS1 CS7 MS7 CS1 CS8 MS2 MS3
E G E G E G E G E G E G E G E G
chr16_6793910_minus
chr12_115529830_plus
chr16_6793910_minus
chr12_115529830_plus

No.	Site	Samples		
1	chr20_43880572_plus	CS1	CS2	MS1
2	chr8_59528830_plus	CS1	CS2	MS1
3	chr2_239270468_minus	CS1	CS2	MS1
4	chr4_18032493_minus	CS1	CS2	MS1
5	chr5_155487136_minus	CS1	CS2	MS1
6	chr10_2704044_plus	CS1	CS2	MS1
7	chr2_9444519_minus	CS1	CS2	MS1
8	chr19_7908394_minus	CS1	CS2	MS1
9	chr1_176722954_plus	CS1	CS2	MS1
10	chr3_656447_plus	CS1	CS2	MS1
11	chr3_67793591_plus	CS1	CS2	MS1
12	chr3_84593287_plus	CS1	CS2	MS1
13	chr1_121352055_minus	CS1	CS2	MS1
14	chr1_242496549_minus	CS1	CS2	MS1
15	chr2_147020918_minus	CS1	CS2	MS1

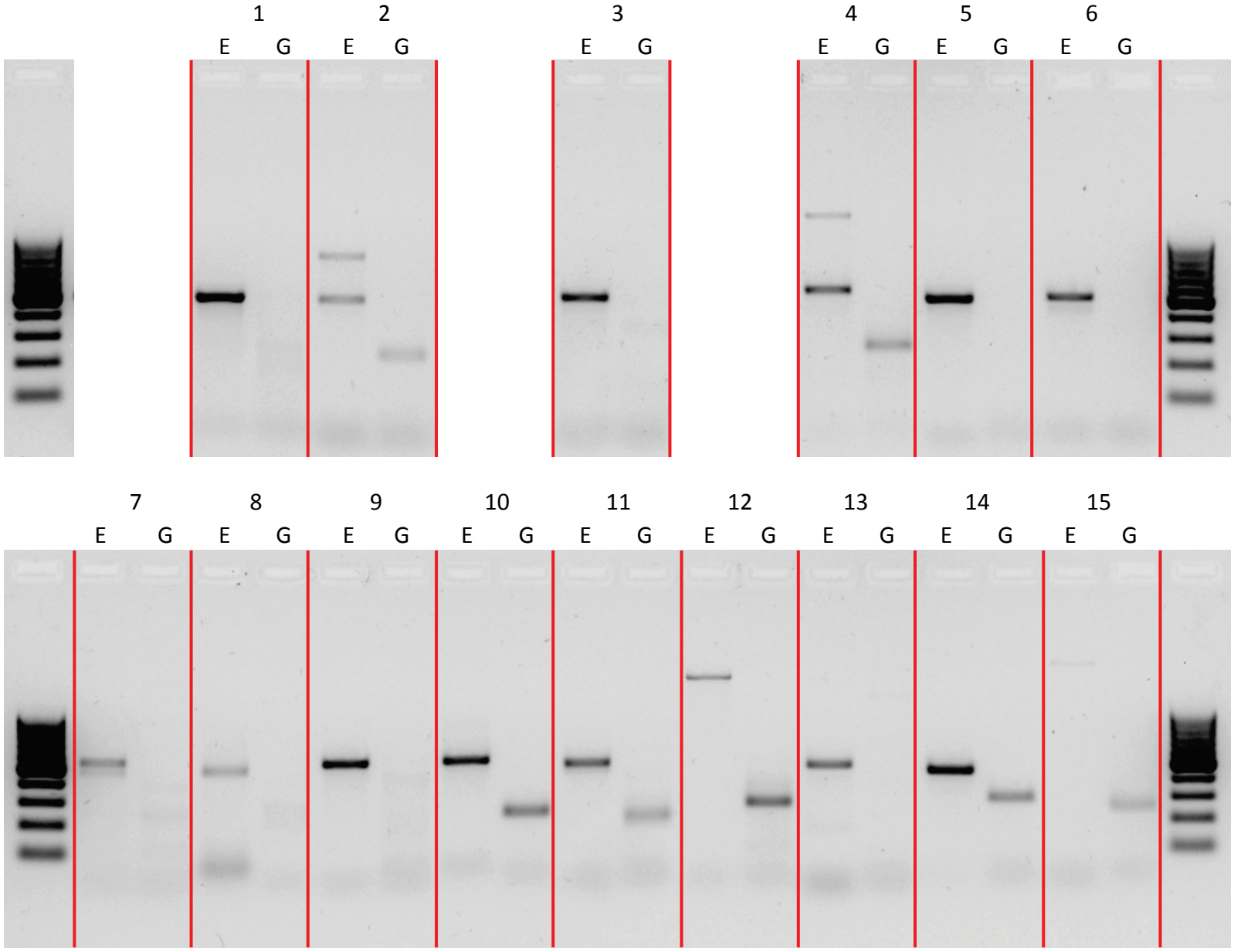
CS1


1
2
3
4
5
6
E G E G E G E G E G E G
7
8
9
10
11
12
13
14
15
E G E G E G E G E G E G E G E G E G

CS2

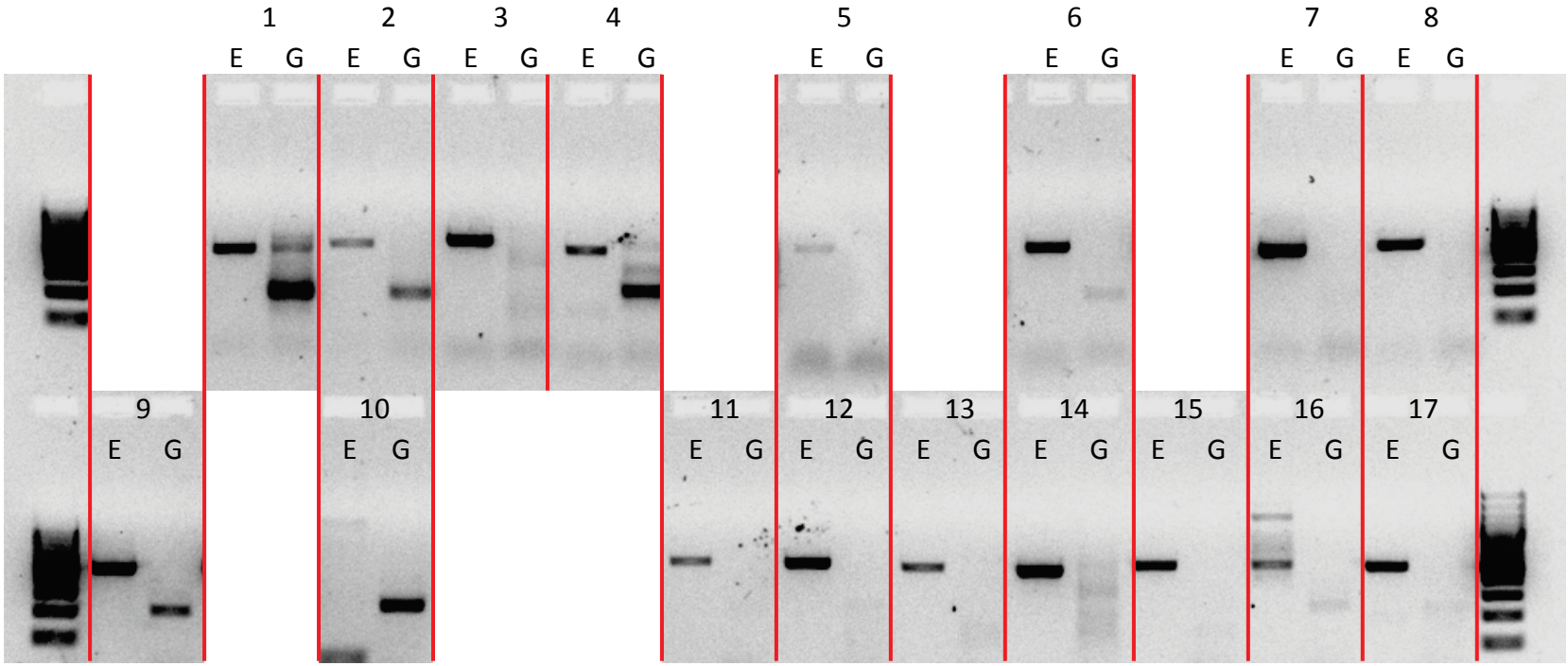

1
2
3
4
5
6
E
G
E
G
E
G
E
G
E
G
E
G
7
8
9
10
11
12
13
14
15
E
G
E
G
E
G
E
G
E
G
E
G
E
G
E
G
E
G

MS1

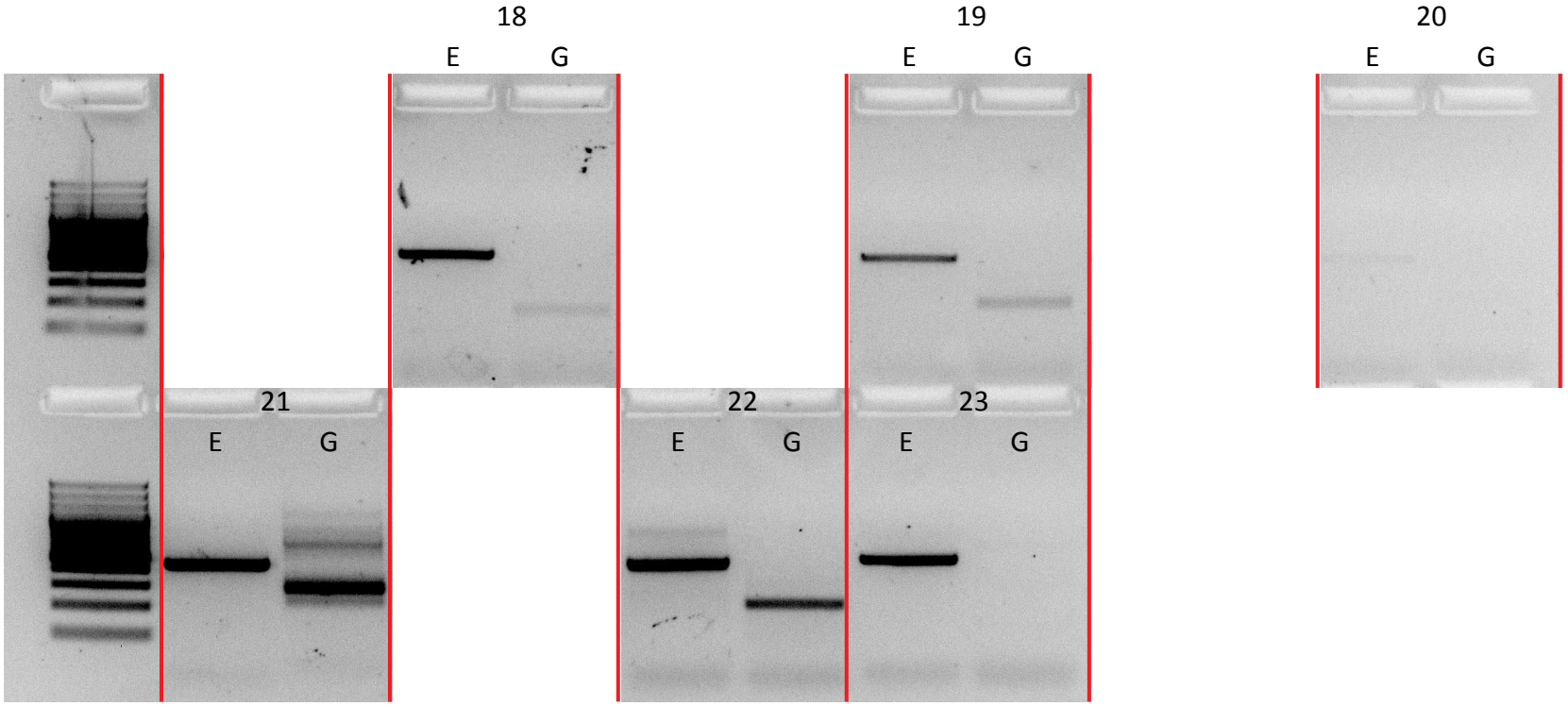

1
2
3
4
5
6
E G E G E G E G E G E G
7
8
9
10
11
12
13
14
15
E G E G E G E G E G E G E G E G E G

No.	Site	Sample				
1	chrX_51211114_minus	CS1	CS2	CS6	CS7	MS1
2	chrX_75949612_minus	CS1	CS2	CS6	CS7	MS1
3	chr21_9495872_minus	CS1				
4	chr21_10760899_minus	CS1	CS2	CS6	CS7	MS1
5	chr21_21834907_minus	CS1				
6	chr6_152840265_minus	CS1				
7	chr8_5786746_minus	CS1				
8	chr8_73206437_minus	CS1				
9	chr8_79944646_minus	CS1	CS2	CS6	CS7	MS1
10	chr9_6290490_minus	CS1				
11	chr10_96542886_minus	CS1				
12	chr11_87976718_minus	CS1				
13	chr11_99693360_minus	CS1				
14	chr12_17908608_minus	CS1				
15	chr12_23788622_minus	CS1				
16	chr12_74082832_minus	CS1				
17	chr14_47643264_minus	CS1				
18	chr15_60482742_minus	CS1	CS2	CS6	CS7	MS1
19	chr16_73741680_minus	CS1	CS2	CS6	CS7	MS1
20	chr18_18510934_minus	CS1				
21	chr18_67598691_minus	CS1				
22	chr20_29591543_minus	CS1	CS2	CS6	CS7	MS1
23	chr20_51777689_minus	CS1				

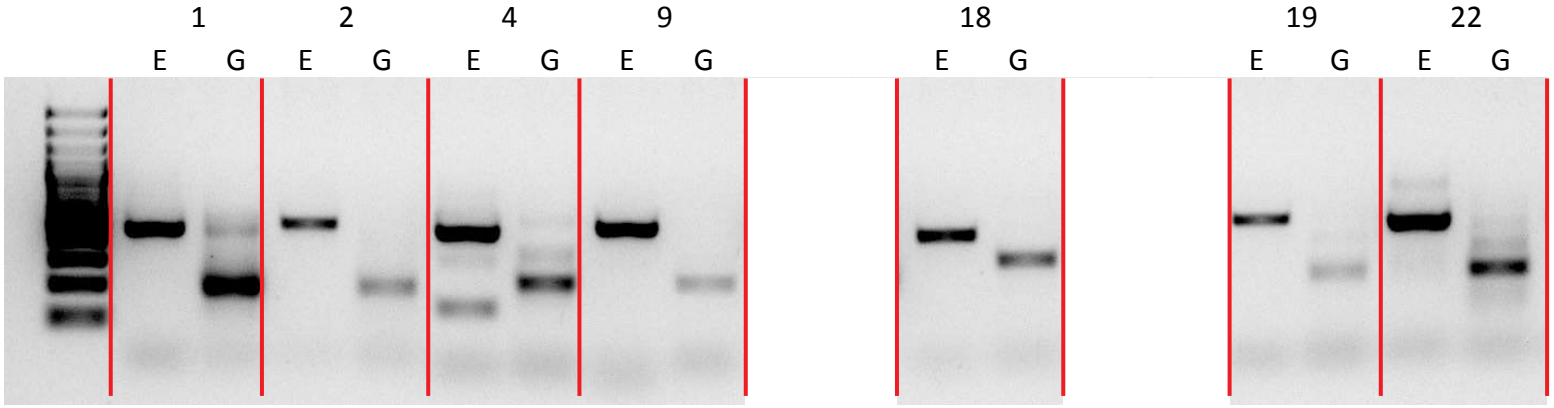
CS1


1
E G
2
E G
3
E G
4
E G
5
E G
6
E G
7
E G
8
E G
9
E G
10
E G
11
E G
12
E G
13
E G
14
E G
15
E G
16
E G
17
E G

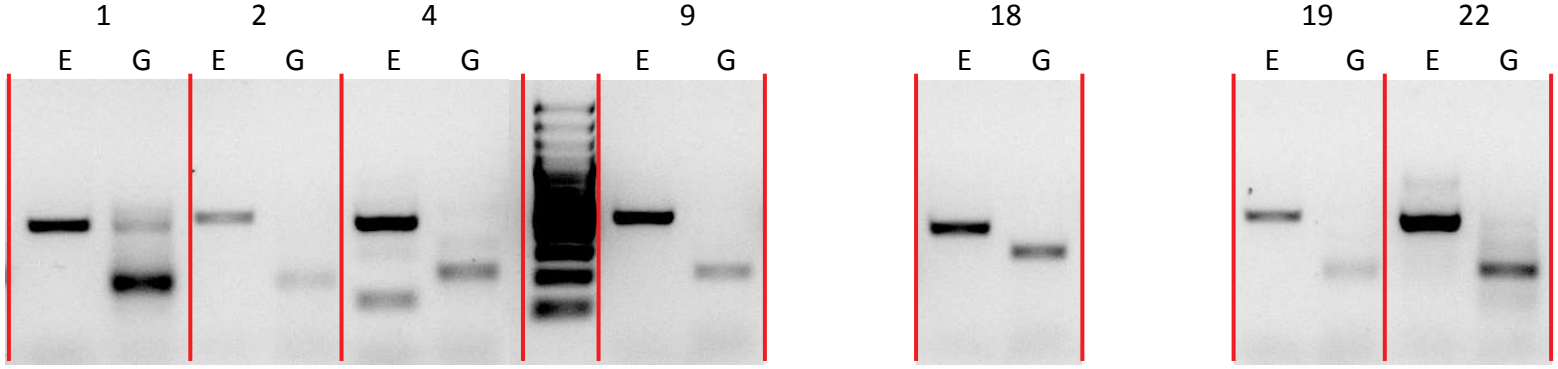
CS1



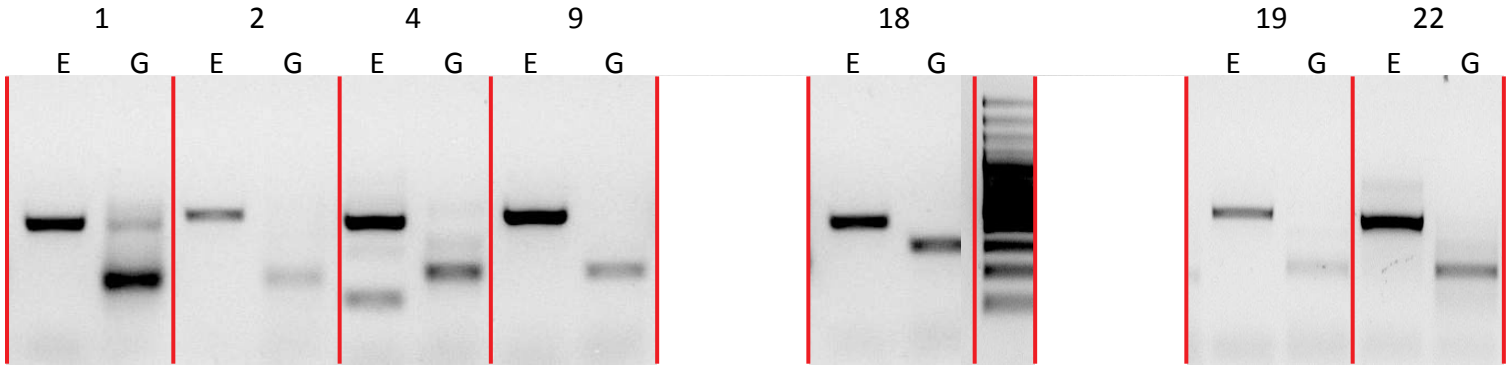
CS2



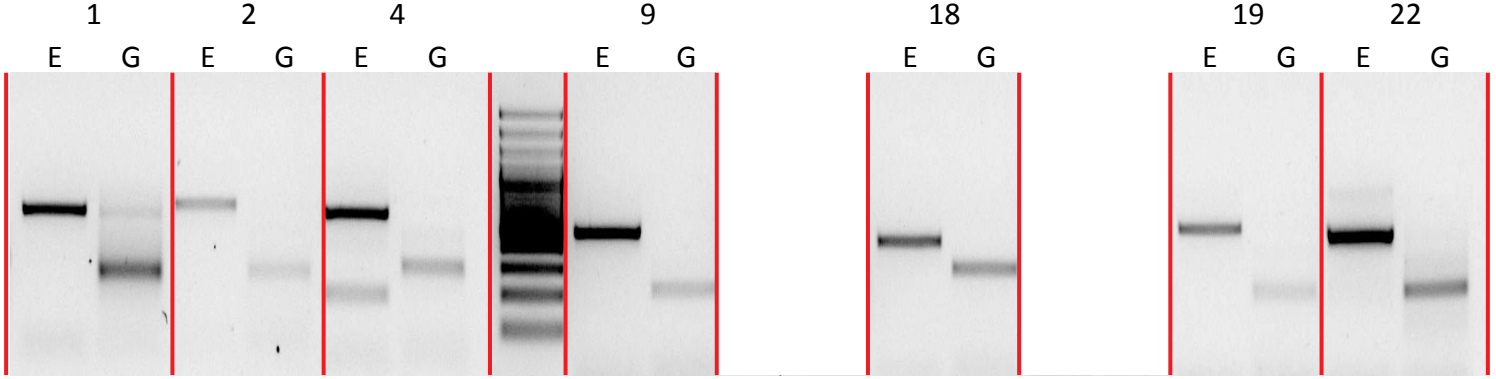
CS6



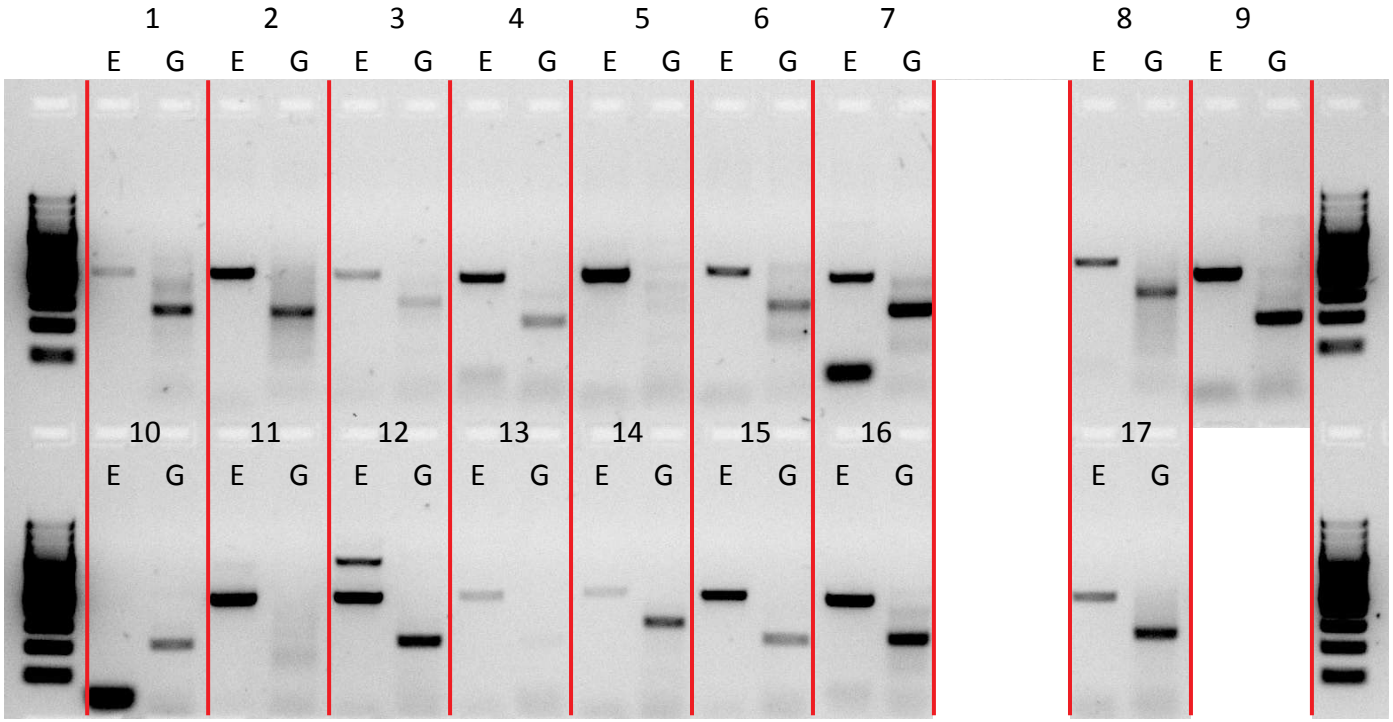
CS7



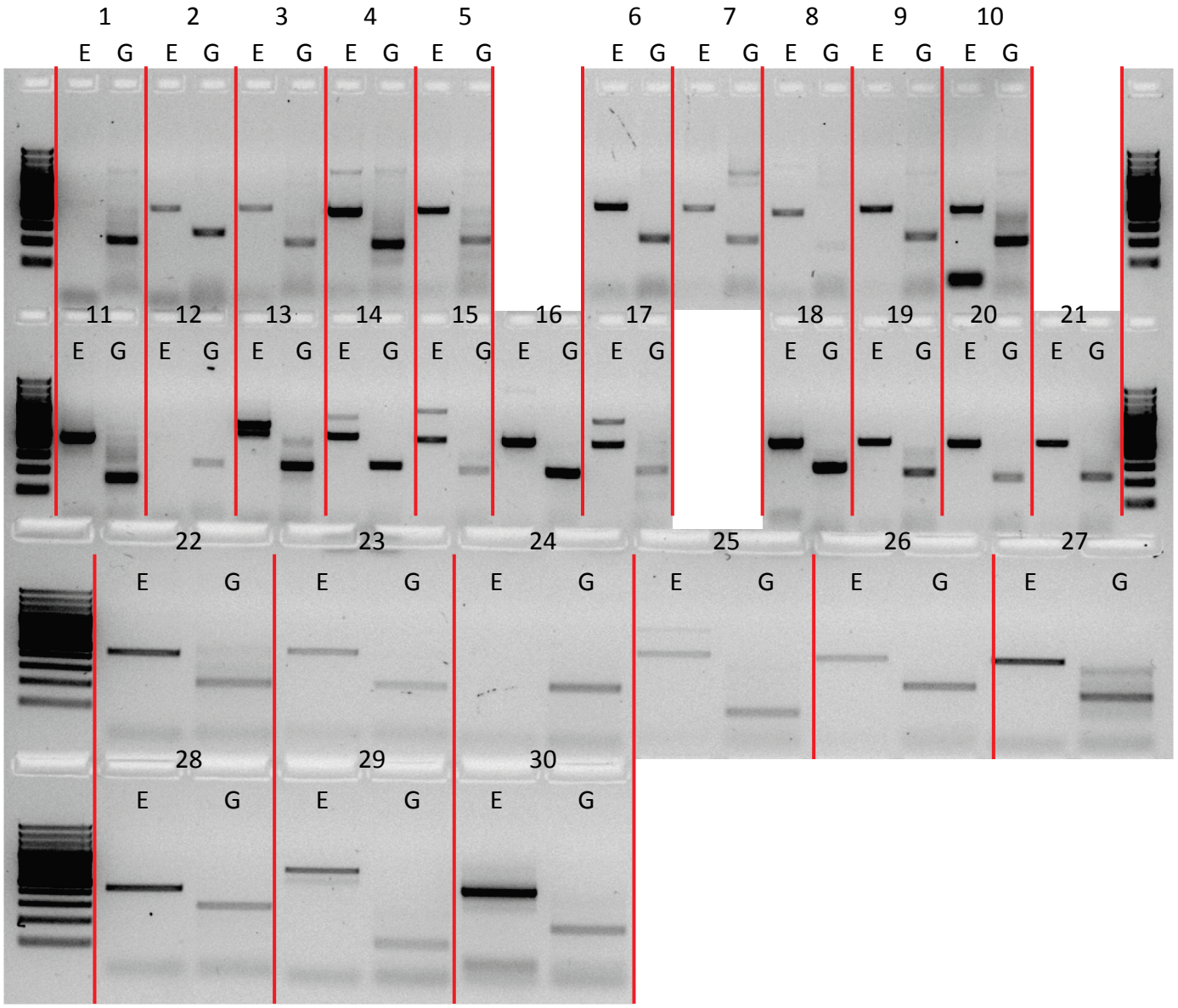
MS1



No.	Site	Sample
1	chr1_106396431_plus	MS1
2	chr1_177792483_plus	MS3
3	chr1_216097436_minus	CS2
4	chr1_230241576_minus	MS1
5	chr2_6142873_plus	CS5
6	chr2_75918908_plus	MS4
7	chr2_177704492_plus	MS6
8	chr3_88467483_minus	CS1
9	chr3_139044424_minus	MS6
10	chr2_53460045_minus	MS6
11	chr2_73768551_minus	CS3
12	chr2_106163155_minus	CS6
13	chr2_201638065_minus	CS8
14	chr4_118765302_minus	CS7
15	chr4_190659692_minus	CS1
16	chr5_102747623_minus	MS7
17	chr6_77222572_minus	CS3

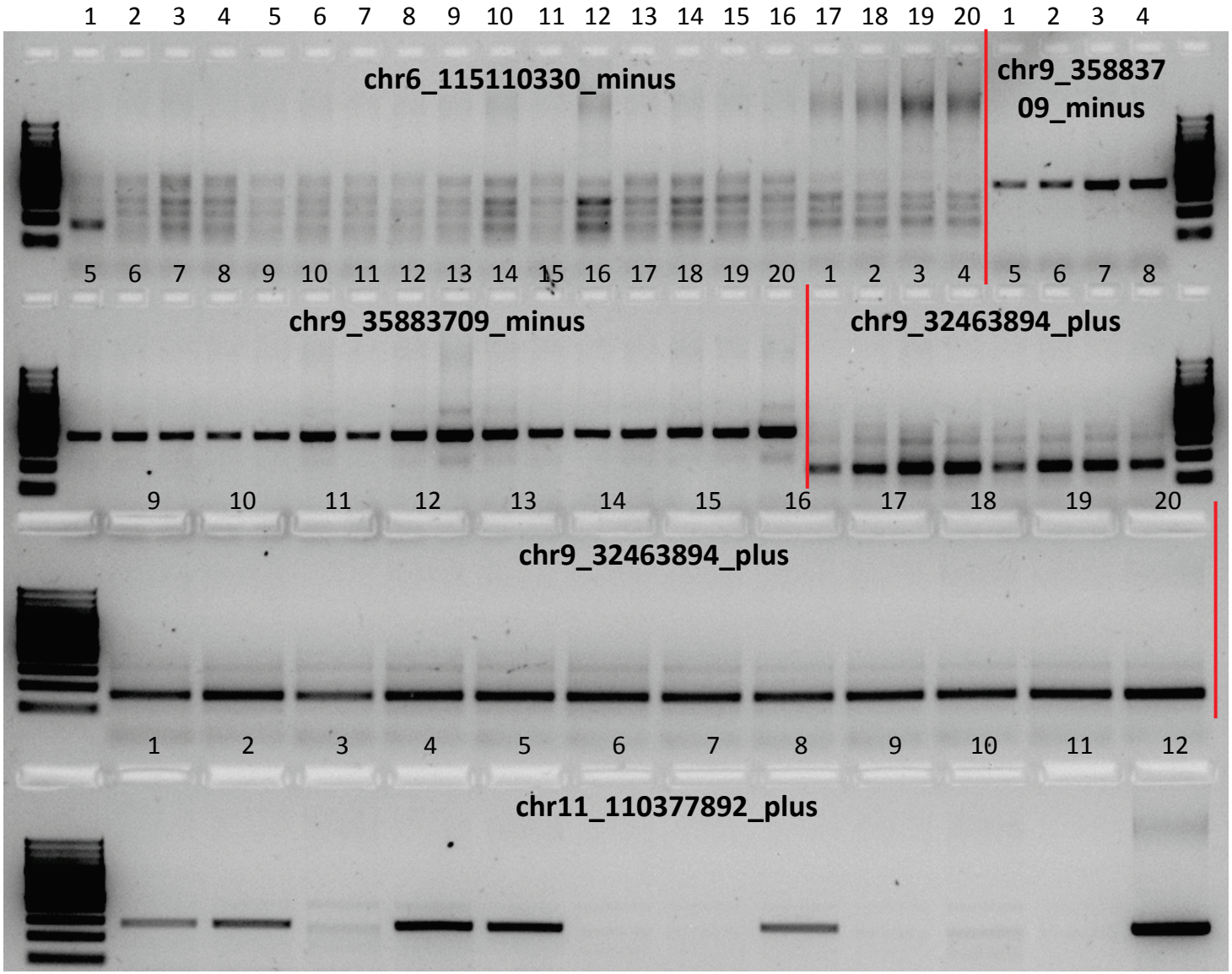

1
2
3
4
5
6
7
8
9
E G E G E G E G E G E G E G E G E G
10
11
12
13
14
15
16
17
E G E G E G E G E G E G E G E G

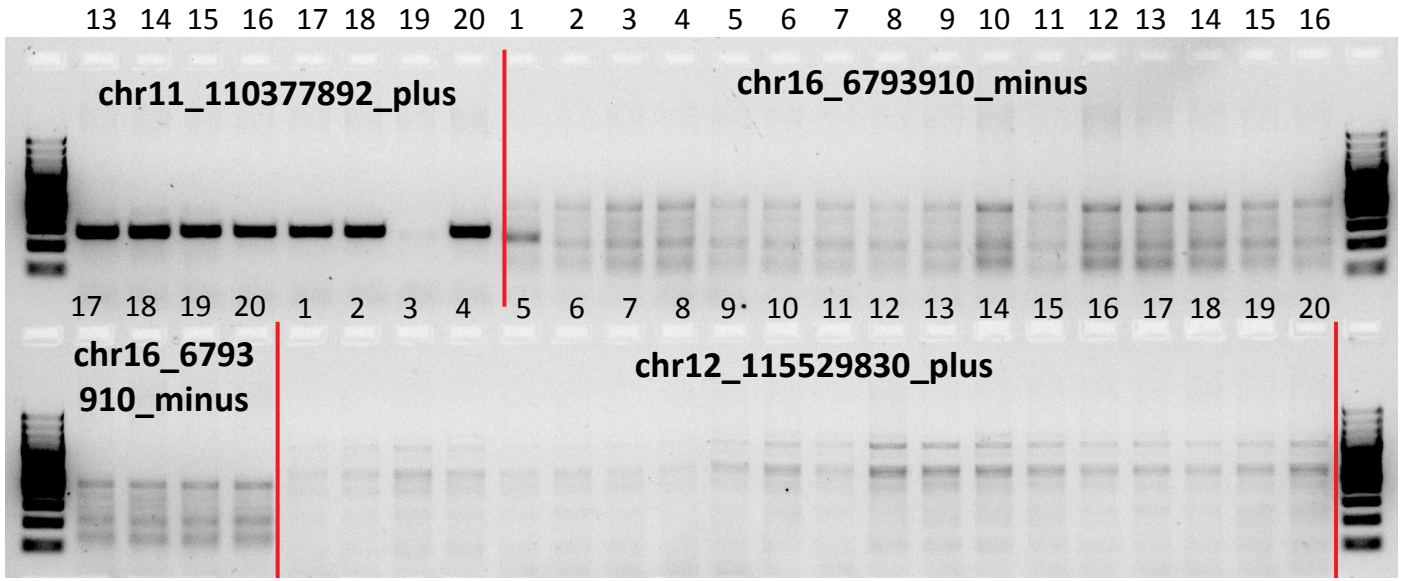
No.	Site	Sample
1	chr2_192493518_plus	MS7
2	chr3_82144860_plus	CS1
3	chr4_18194359_plus	CS1
4	chr4_130361674_plus	CS6
5	chr4_75917687_minus	CS2
6	chr6_74935597_plus	CS5
7	chr7_118993490_minus	CS6
8	chr7_153509409_minus	CS1
9	chr11_97434312_plus	CS6
10	chr12_33325810_plus	CS6
11	chr1_75192913_plus	MS2
12	chr1_98363699_plus	CS2
13	chr2_1482203_plus	CS4
14	chr2_206147278_minus	CS1
15	chr3_108138687_minus	CS3
16	chr1_165531920_plus	CS7
17	chr1_192874933_plus	CS9
18	chr1_58691704_minus	CS3
19	chr1_165553152_plus	CS1
20	chr2_42051402_minus	CS2
21	chr2_83188323_minus	CS2
22	chr4_80768538_minus	CS1
23	chr6_114091139_minus	CS2
24	chr3_85576556_minus	CS1
25	chr4_47007900_plus	CS2
26	chr4_58126678_plus	CS1
27	chr4_147225279_minus	CS1
28	chr12_11191387_plus	CS2
29	chr3_176091206_minus	CS1
30	chr4_170281928_plus	CS1

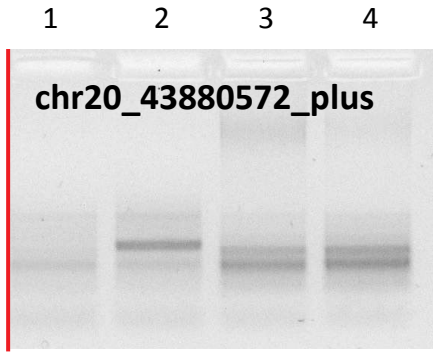

1 2 3 4 5 6 7 8 9 10
E G E G E G E G E G E G E G E G E G E G
11 12 13 14 15 16 17 18 19 20 21
E G E G E G E G E G E G E G E G E G E G E G
22 23 24 25 26 27
E G E G E G E G E G E G
28 29 30
E G E G E G

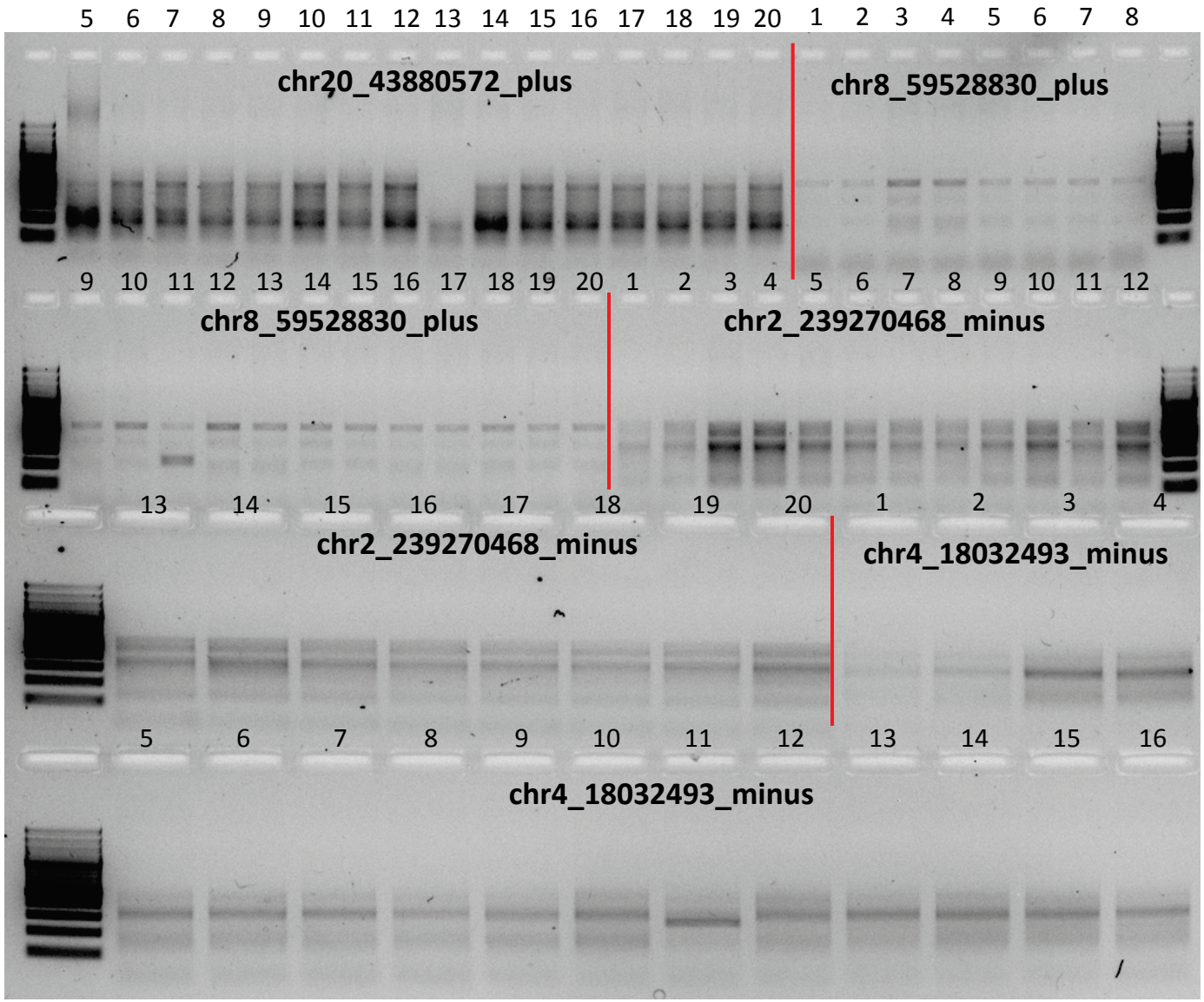
Section B

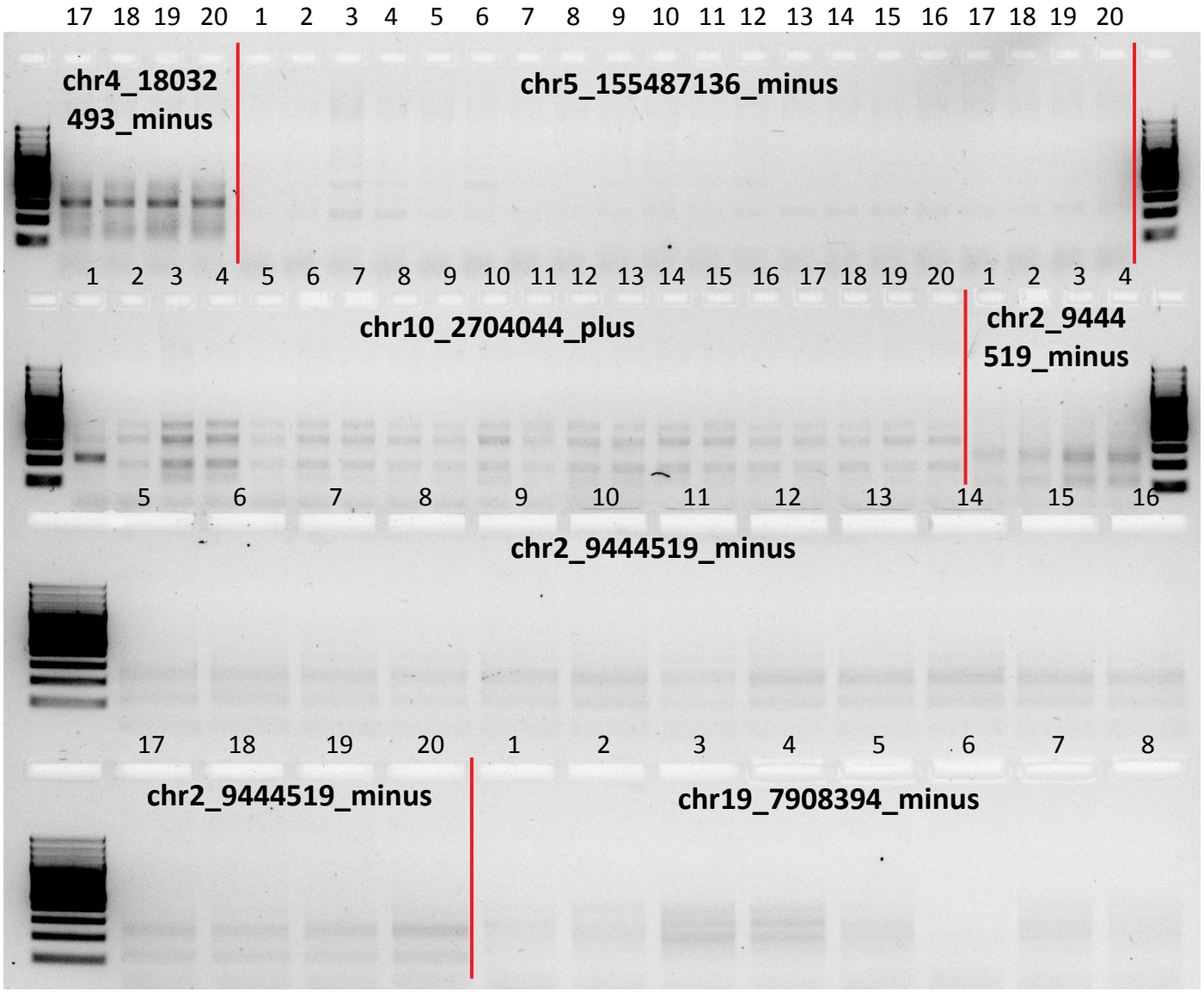
Lane No.	Sample Name
1	CS1
2	CS2
3	CS3
4	CS4
5	CS5
6	CS6
7	CS7
8	CS8
9	CS9
10	CS10
11	MS1
12	MS2
13	MS3
14	MS4
15	MS5
16	MS6
17	MS7
18	MS8
19	MS9
20	MS10

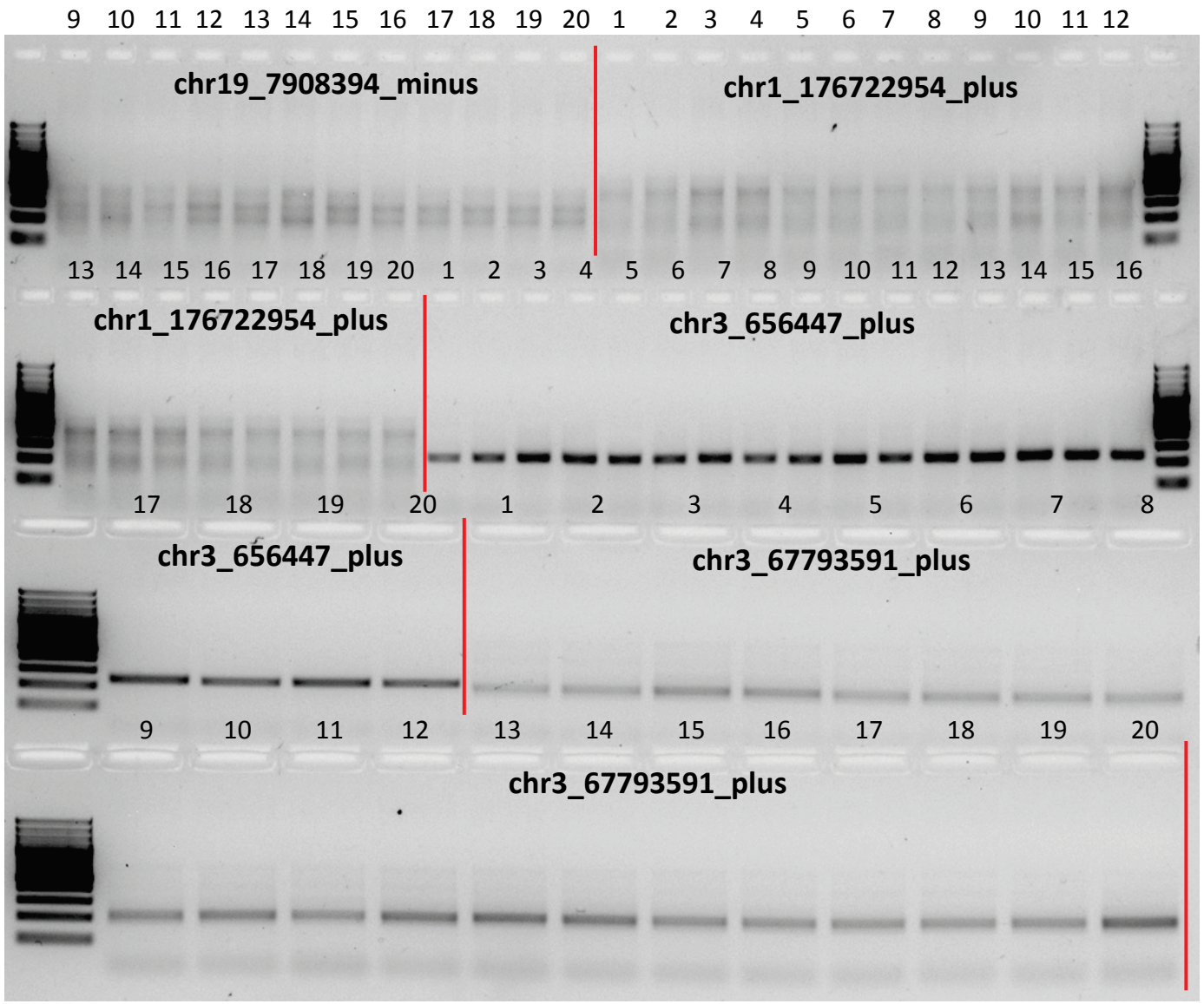

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4
chr6_115110330_minus
chr9_358837
09_minus
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8
chr9_35883709_minus
chr9_32463894_plus
9 10 11 12 13 14 15 16 17 18 19 20
chr9_32463894_plus
1 2 3 4 5 6 7 8 9 10 11 12
chr11_110377892_plus

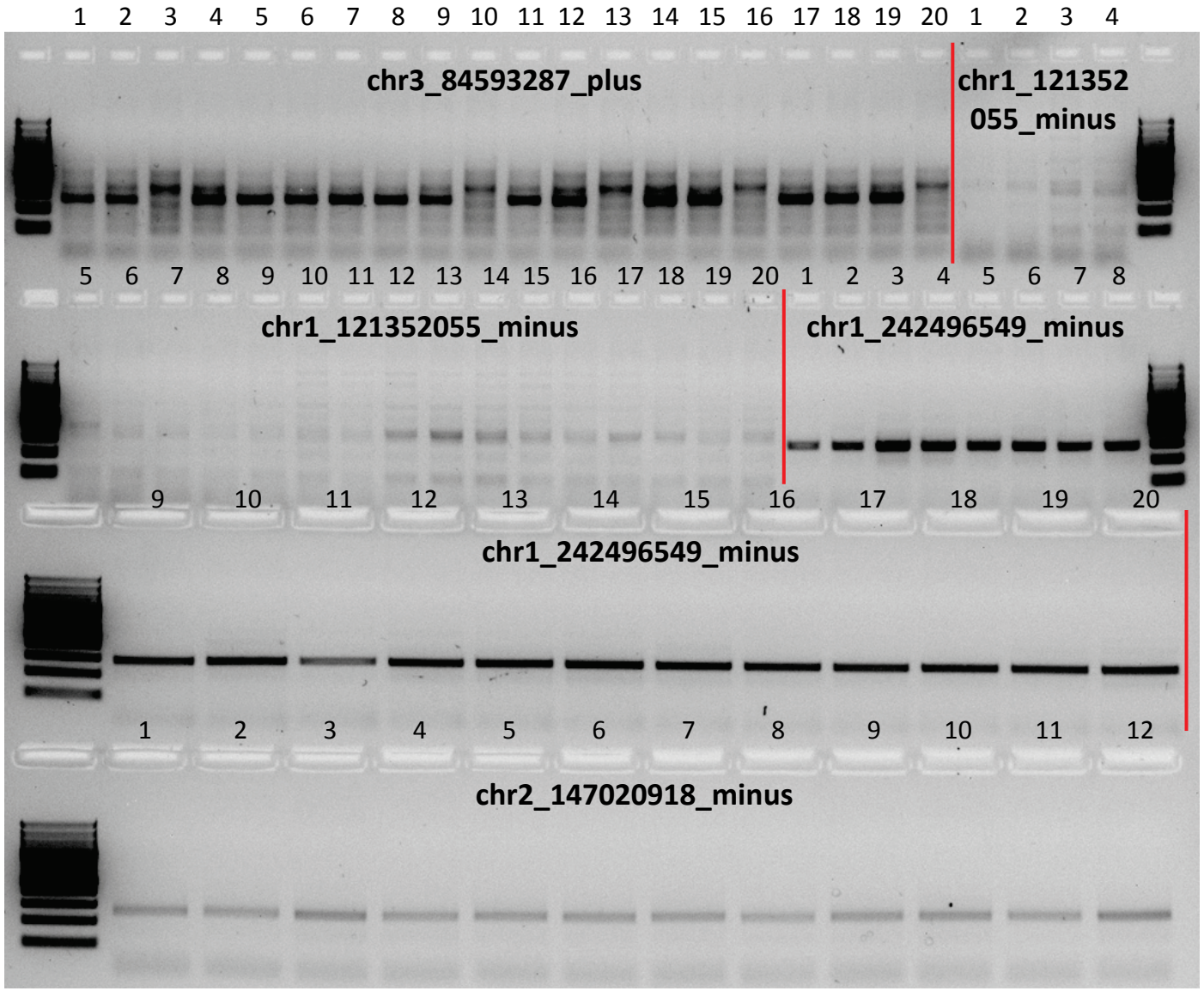

13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
chr11_110377892_plus
chr16_6793910_minus
17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
chr16_6793
910_minus
chr12_115529830_plus

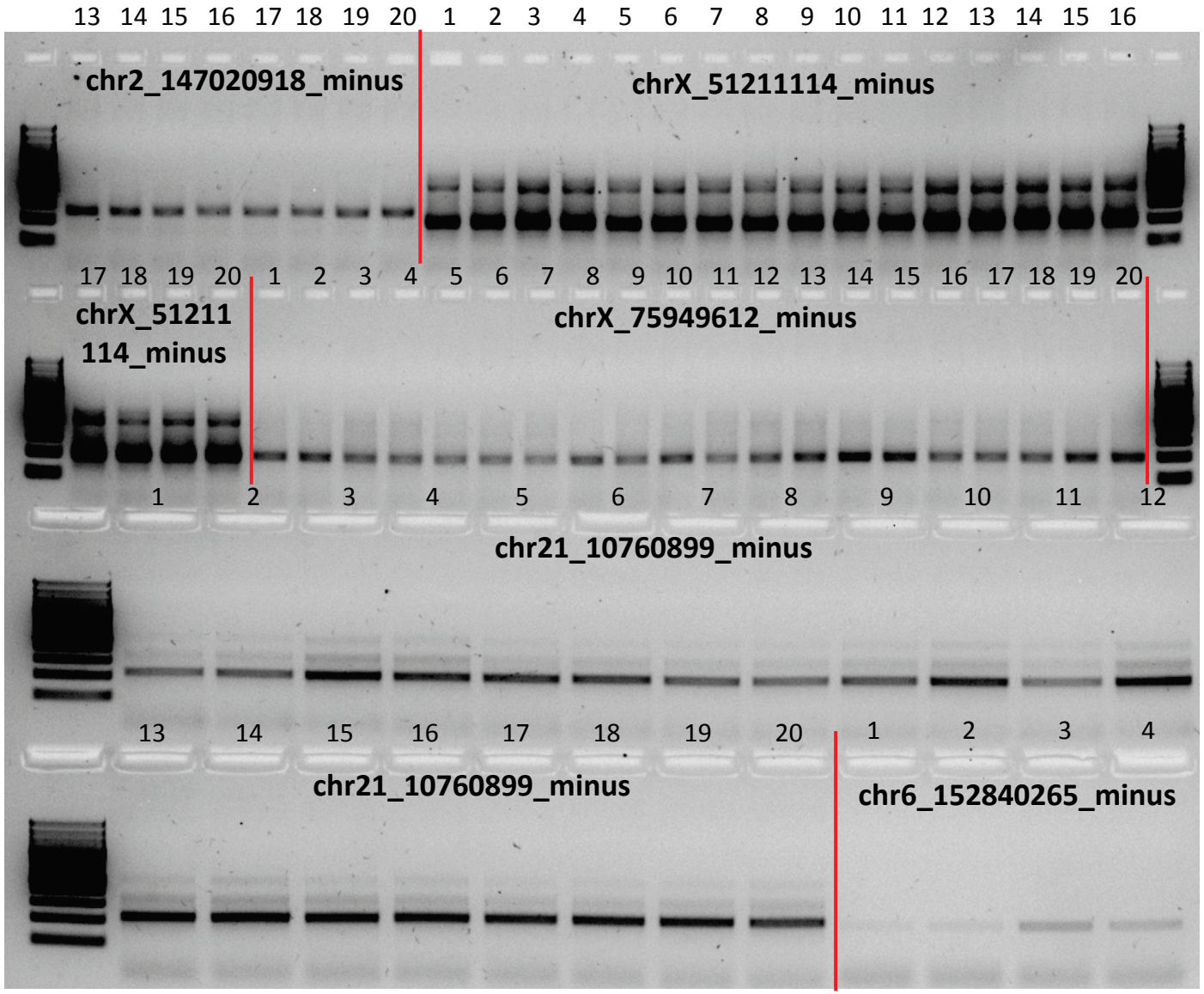

1 2 3 4
chr20_43880572_plus

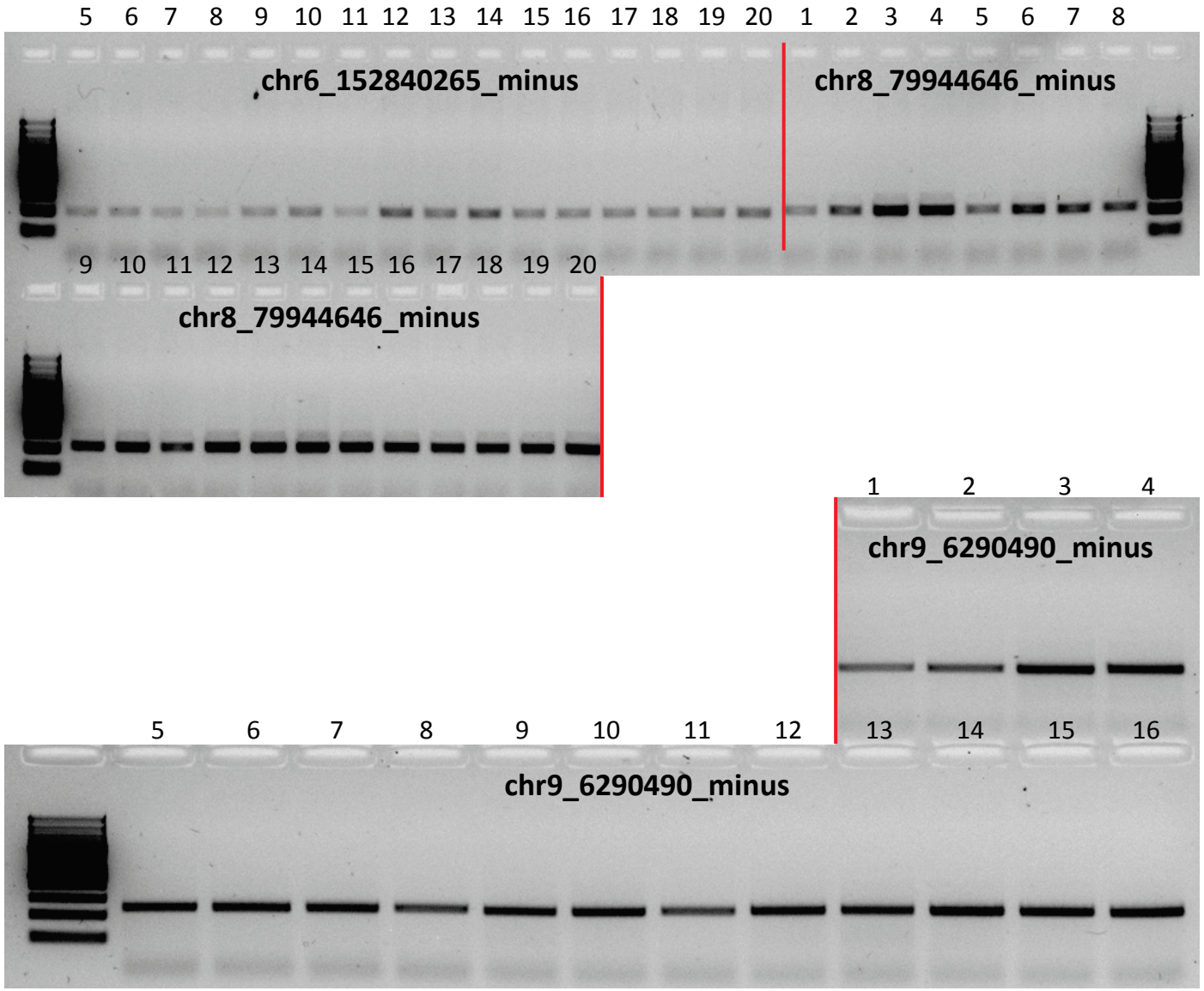

5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8
chr20_43880572_plus
chr8_59528830_plus
9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12
chr8_59528830_plus
chr2_239270468_minus
13 14 15 16 17 18 19 20 1 2 3 4
chr2_239270468_minus
chr4_18032493_minus
5 6 7 8 9 10 11 12 13 14 15 16
chr4_18032493_minus

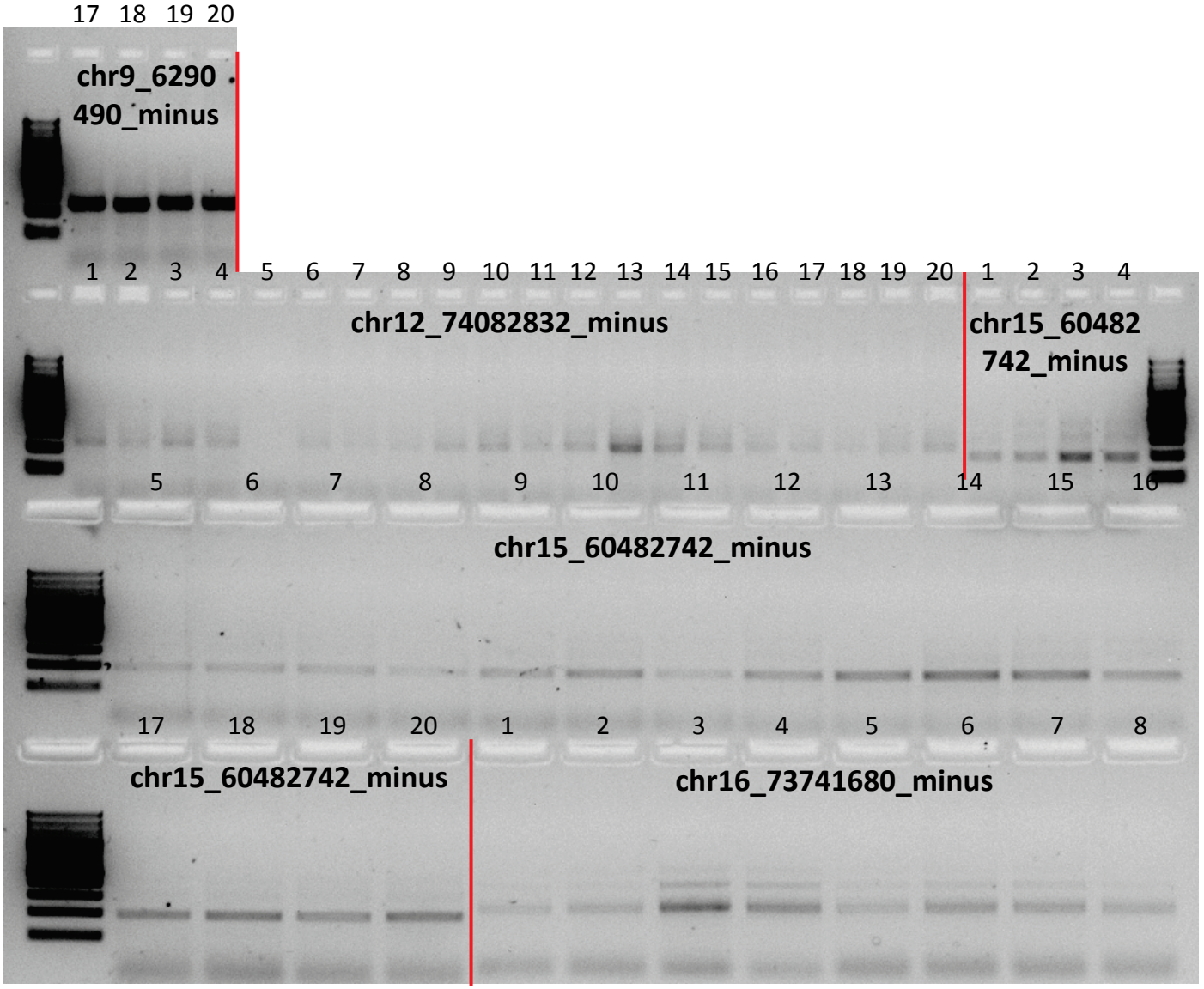

17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
chr4_18032493_minus
chr5_155487136_minus
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4
chr10_2704044_plus
chr2_9444519_minus
5 6 7 8 9 10 11 12 13 14 15 16
chr2_9444519_minus
17 18 19 20 1 2 3 4 5 6 7 8
chr2_9444519_minus
chr19_7908394_minus

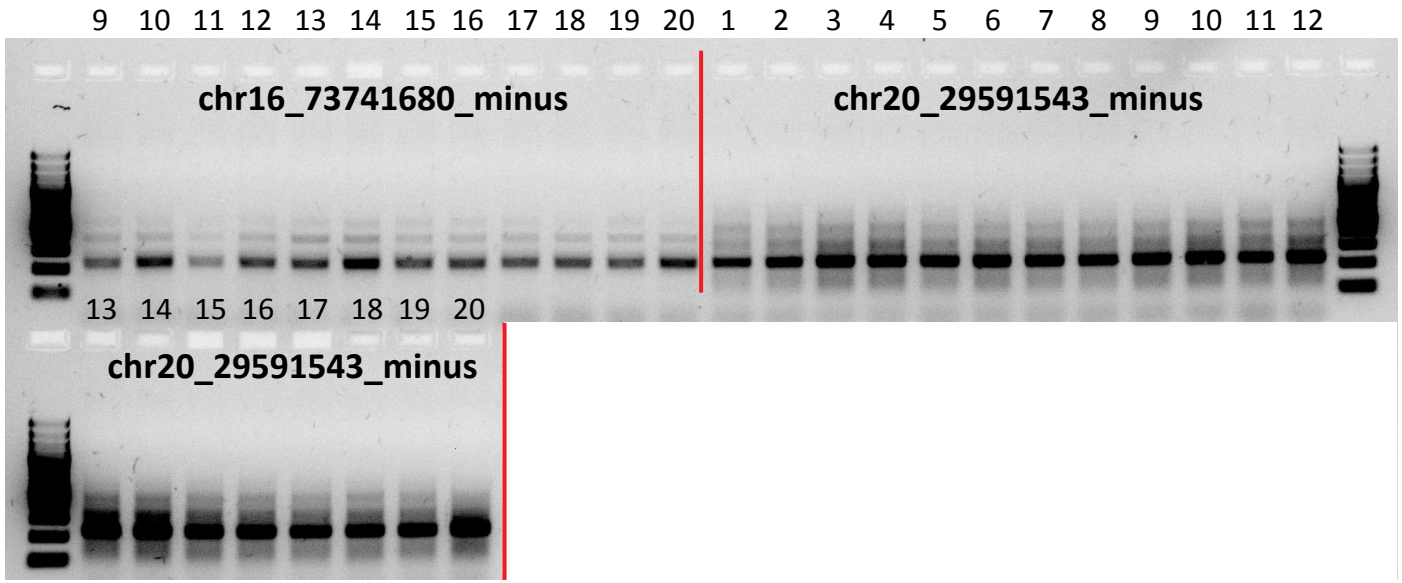

9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12
chr19_7908394_minus
chr1_176722954_plus
13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
chr1_176722954_plus
chr3_656447_plus
17 18 19 20 1 2 3 4 5 6 7 8
chr3_656447_plus
chr3_67793591_plus
9 10 11 12 13 14 15 16 17 18 19 20
chr3_67793591_plus

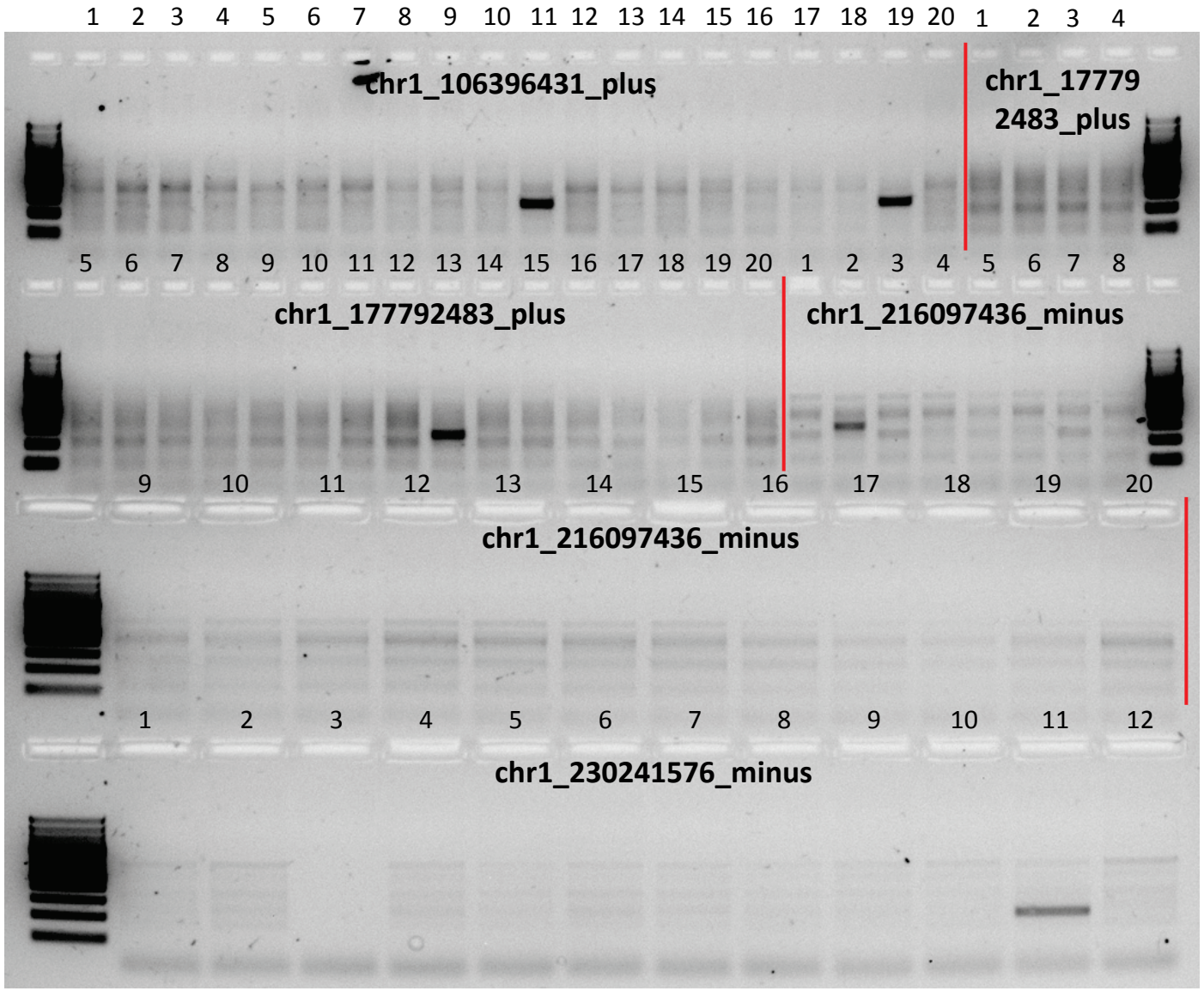

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4
chr3_84593287_plus
chr1_121352055_minus
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8
chr1_121352055_minus
chr1_242496549_minus
9 10 11 12 13 14 15 16 17 18 19 20
chr1_242496549_minus
1 2 3 4 5 6 7 8 9 10 11 12
chr2_147020918_minus

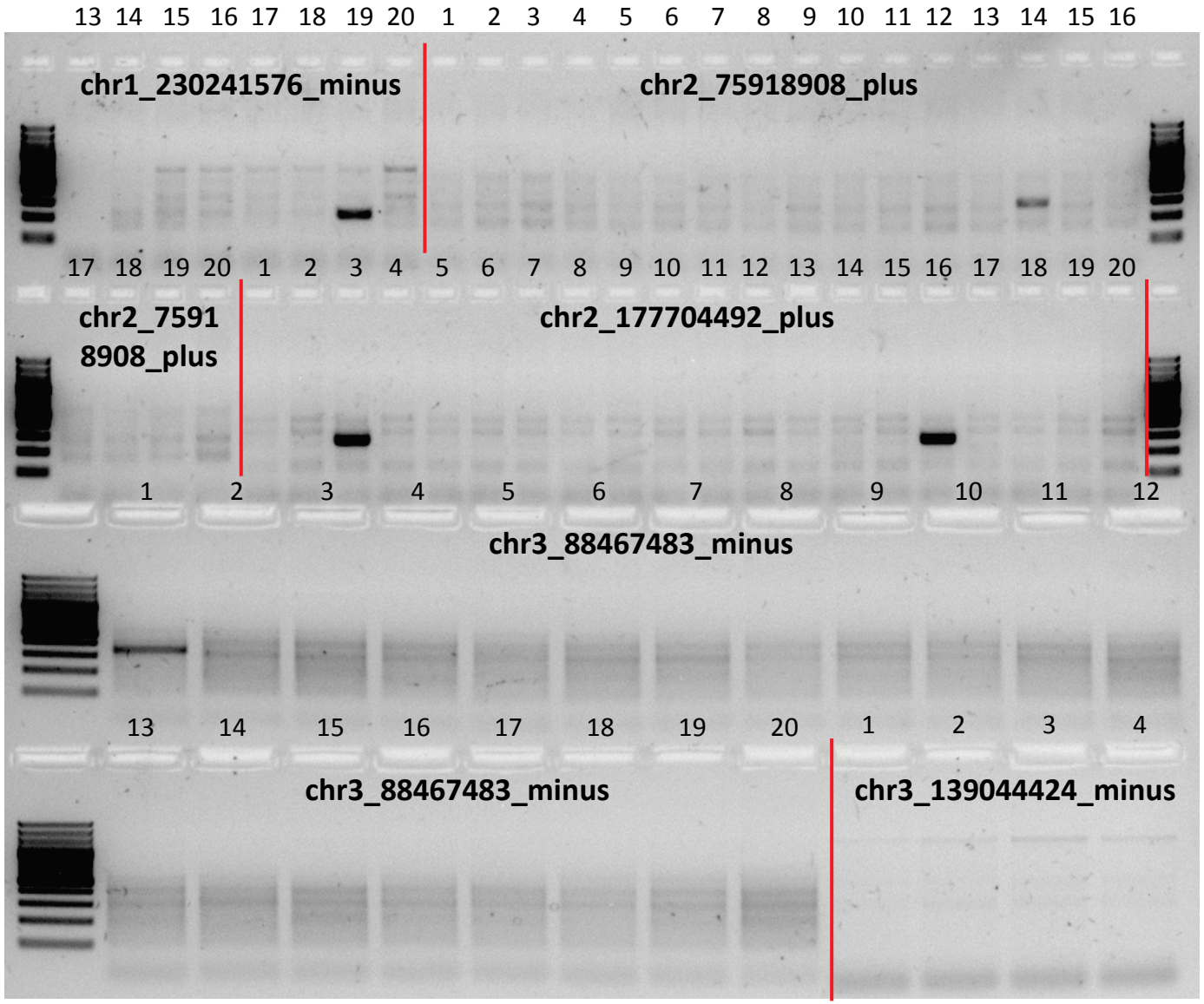

13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
chr2_147020918_minus
chrX_51211114_minus
17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
chrX_51211
114_minus
chrX_75949612_minus
1 2 3 4 5 6 7 8 9 10 11 12
chr21_10760899_minus
13 14 15 16 17 18 19 20 1 2 3 4
chr21_10760899_minus
chr6_152840265_minus

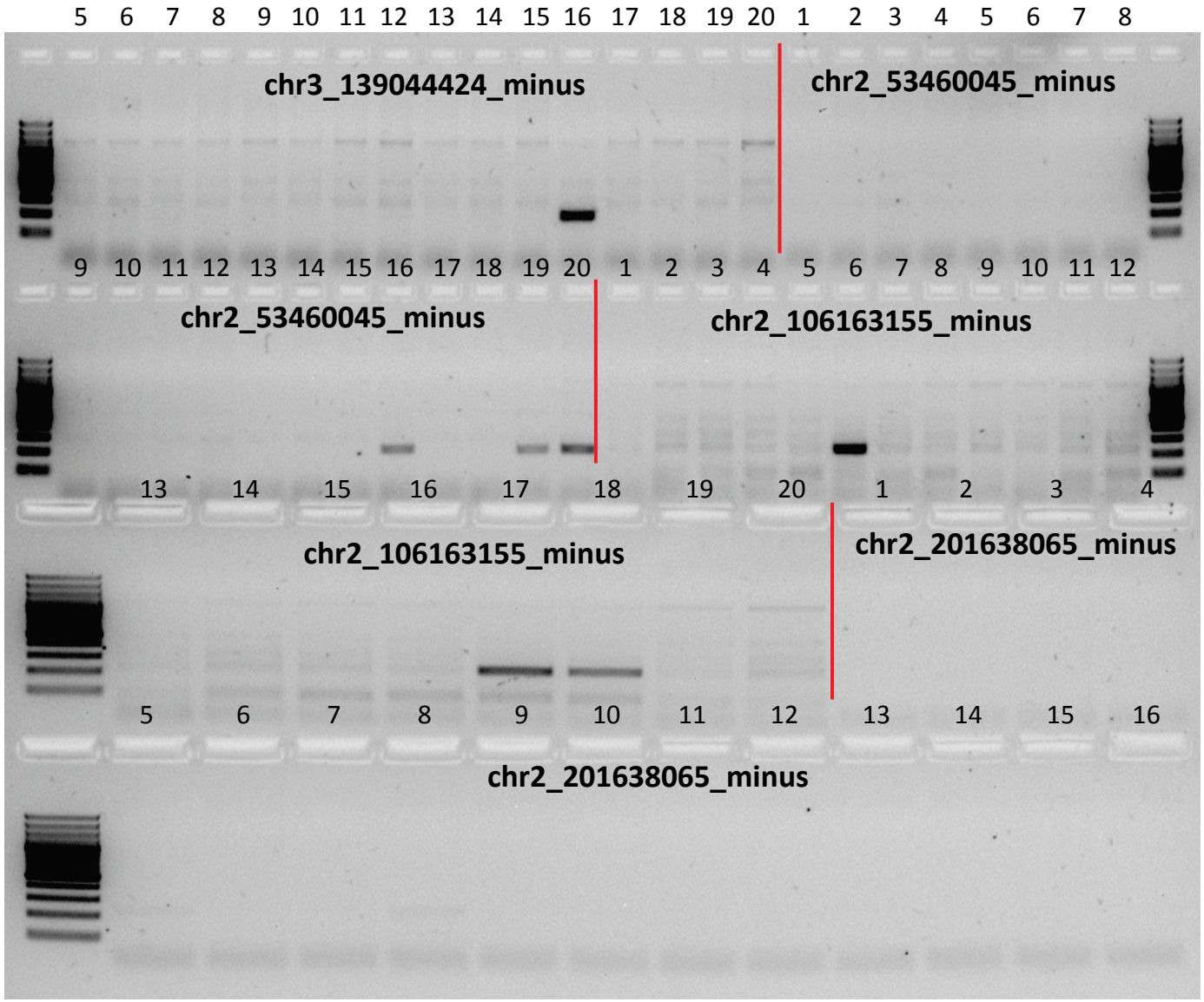

5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8
chr6_152840265_minus
chr8_79944646_minus
9 10 11 12 13 14 15 16 17 18 19 20
chr8_79944646_minus
1 2 3 4
chr9_6290490_minus
5 6 7 8 9 10 11 12 13 14 15 16
chr9_6290490_minus

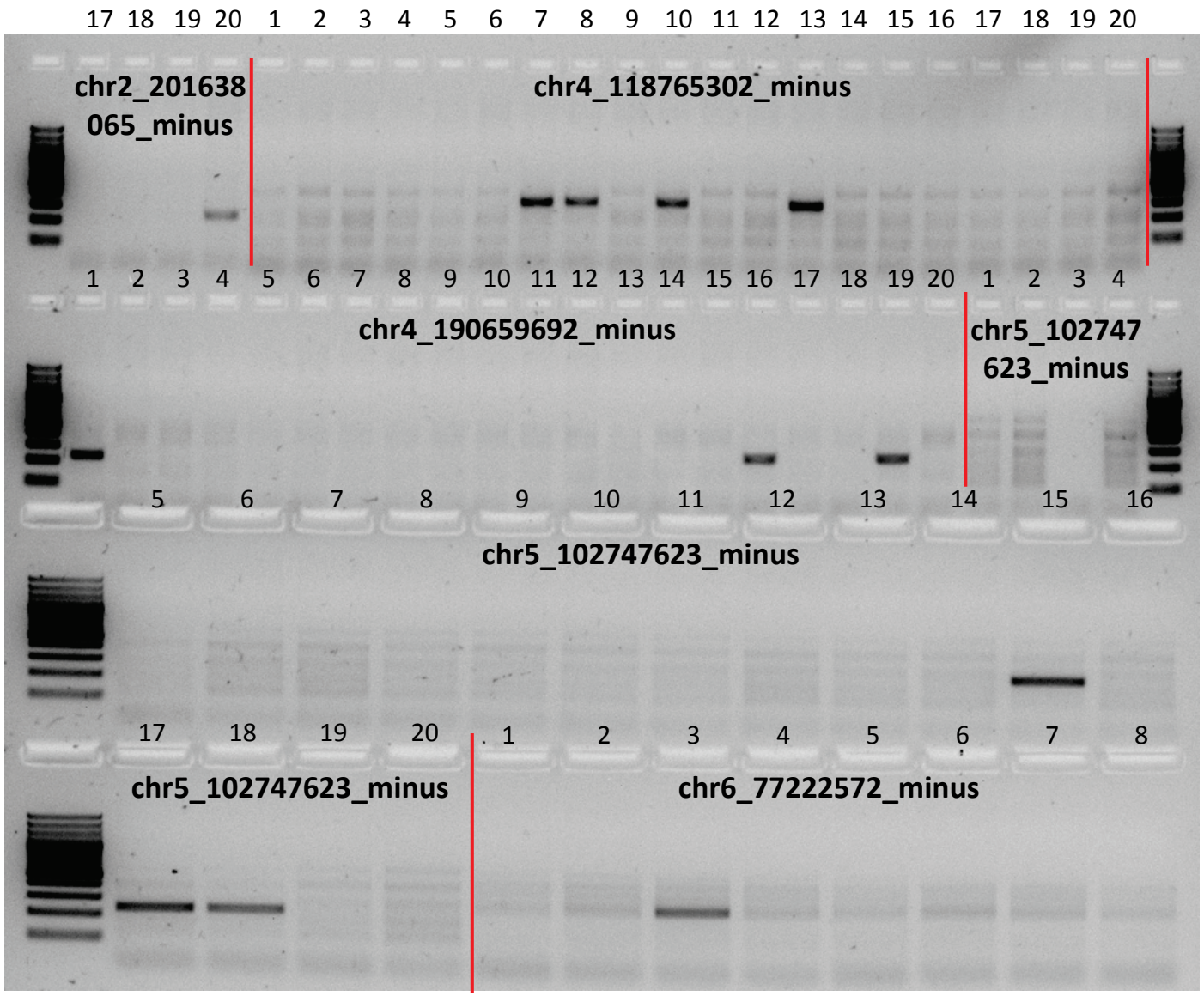

17 18 19 20
chr9_6290
490_minus
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4
chr12_74082832_minus
chr15_60482
742_minus
5 6 7 8 9 10 11 12 13 14 15 16
chr15_60482742_minus
17 18 19 20 1 2 3 4 5 6 7 8
chr15_60482742_minus
chr16_73741680_minus

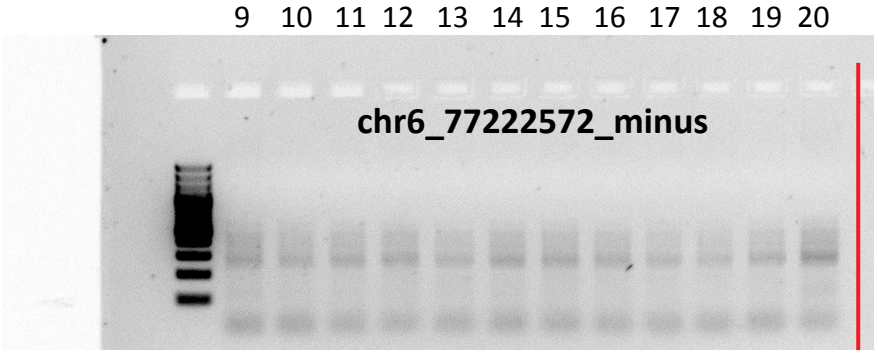

9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12
chr16_73741680_minus
chr20_29591543_minus
13 14 15 16 17 18 19 20
chr20_29591543_minus

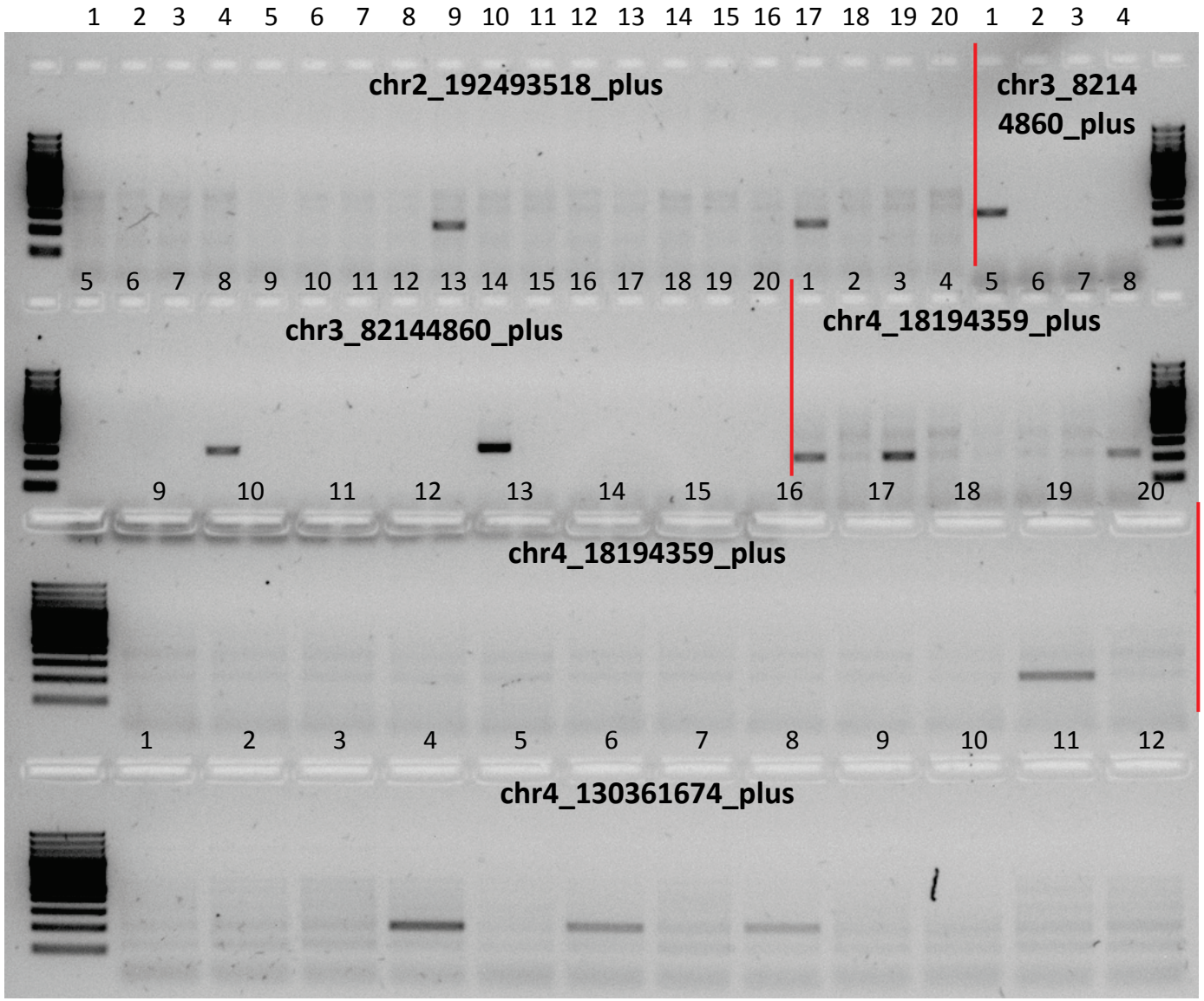

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4
chr1_106396431_plus
chr1_17779
2483_plus
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8
chr1_177792483_plus
chr1_216097436_minus
9 10 11 12 13 14 15 16 17 18 19 20
chr1_216097436_minus
1 2 3 4 5 6 7 8 9 10 11 12
chr1_230241576_minus

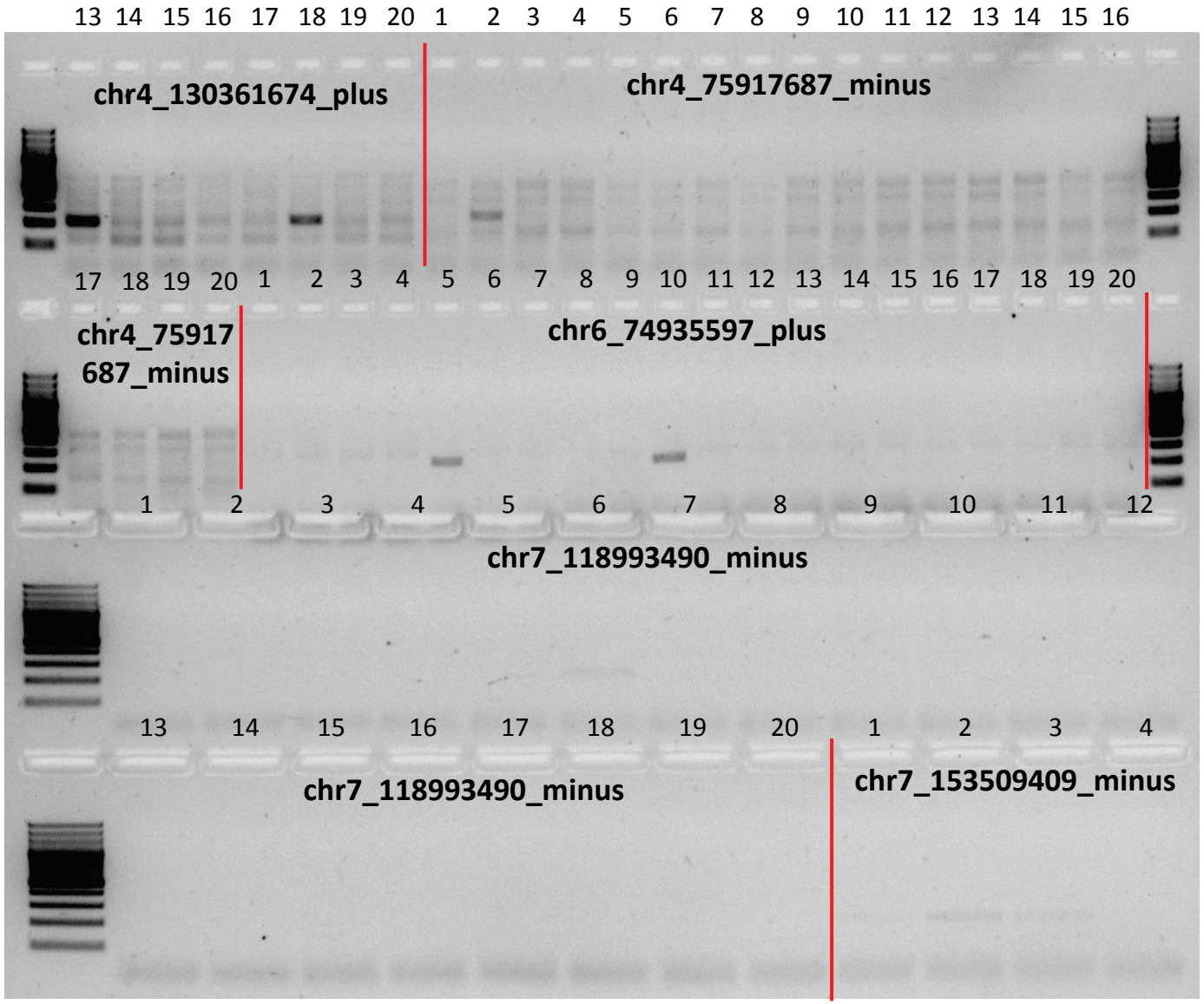

13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
chr1_230241576_minus
chr2_75918908_plus
17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
chr2_7591
8908_plus
chr2_177704492_plus
1 2 3 4 5 6 7 8 9 10 11 12
chr3_88467483_minus
13 14 15 16 17 18 19 20 1 2 3 4
chr3_88467483_minus
chr3_139044424_minus

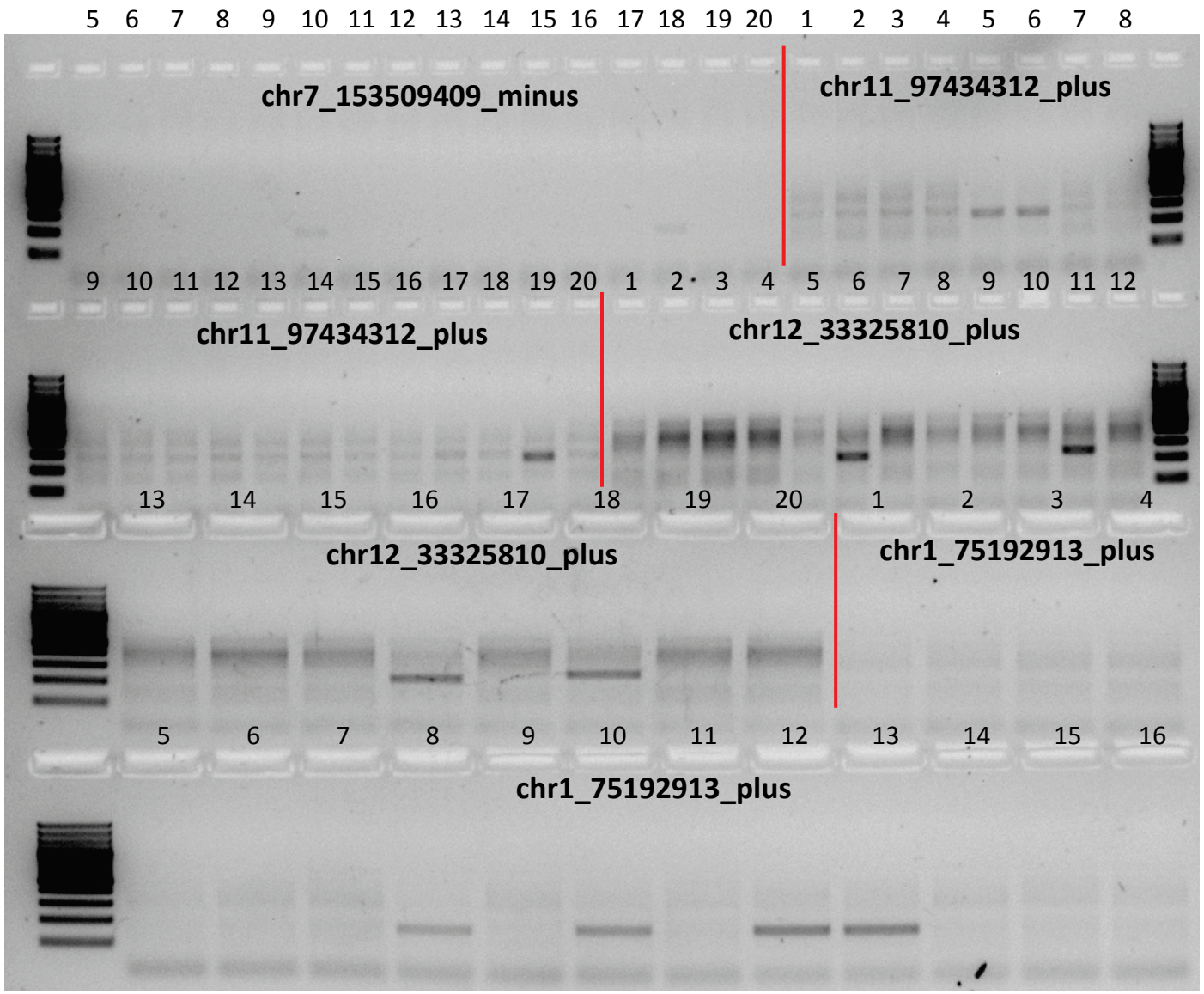

5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8
chr3_139044424_minus
chr2_53460045_minus
9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12
chr2_53460045_minus
chr2_106163155_minus
13 14 15 16 17 18 19 20 1 2 3 4
chr2_106163155_minus
chr2_201638065_minus
5 6 7 8 9 10 11 12 13 14 15 16
chr2_201638065_minus

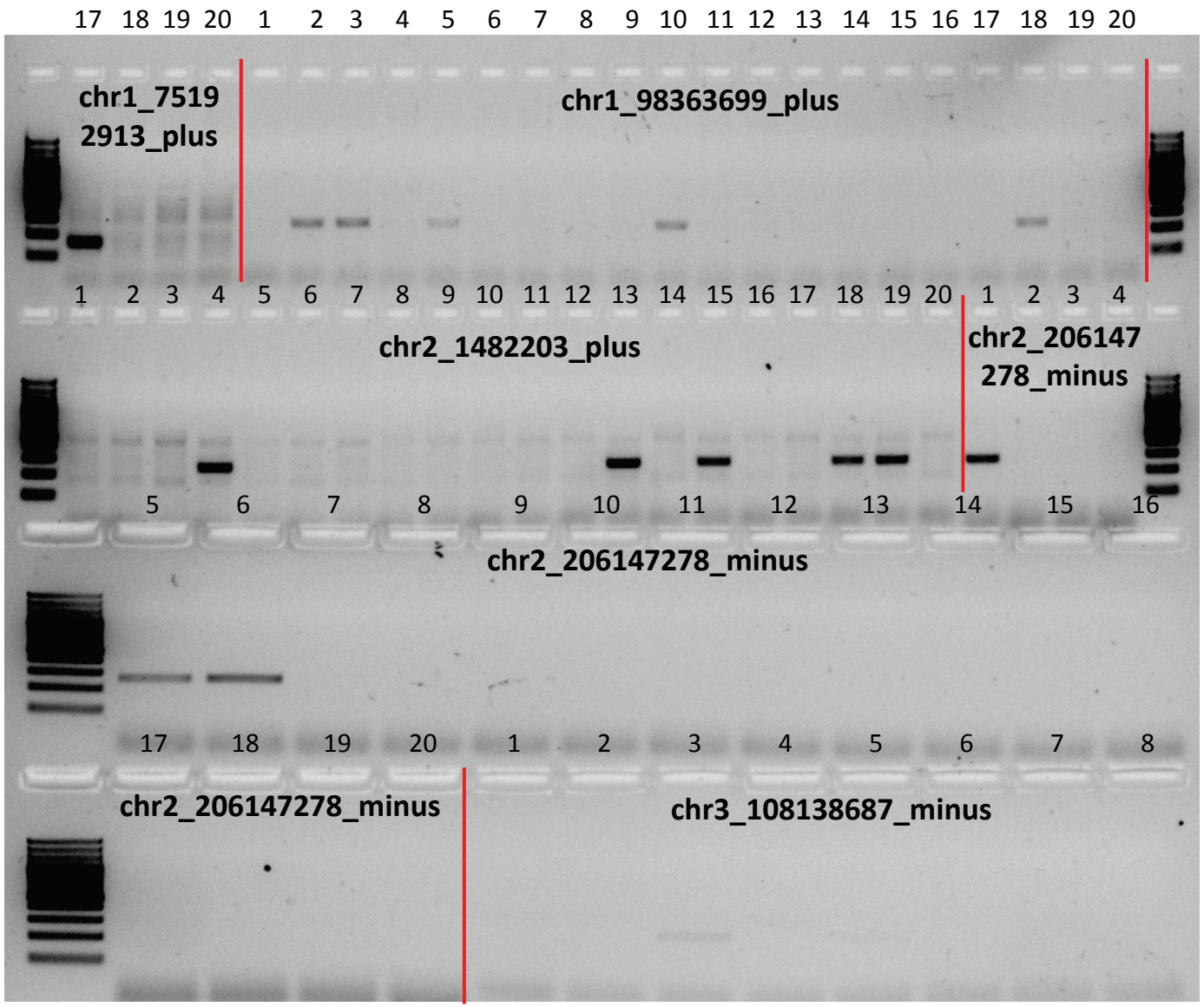

17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
chr2_201638065_minus
chr4_118765302_minus
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4
chr4_190659692_minus
chr5_102747623_minus
5 6 7 8 9 10 11 12 13 14 15 16
chr5_102747623_minus
17 18 19 20 1 2 3 4 5 6 7 8
chr5_102747623_minus
chr6_77222572_minus

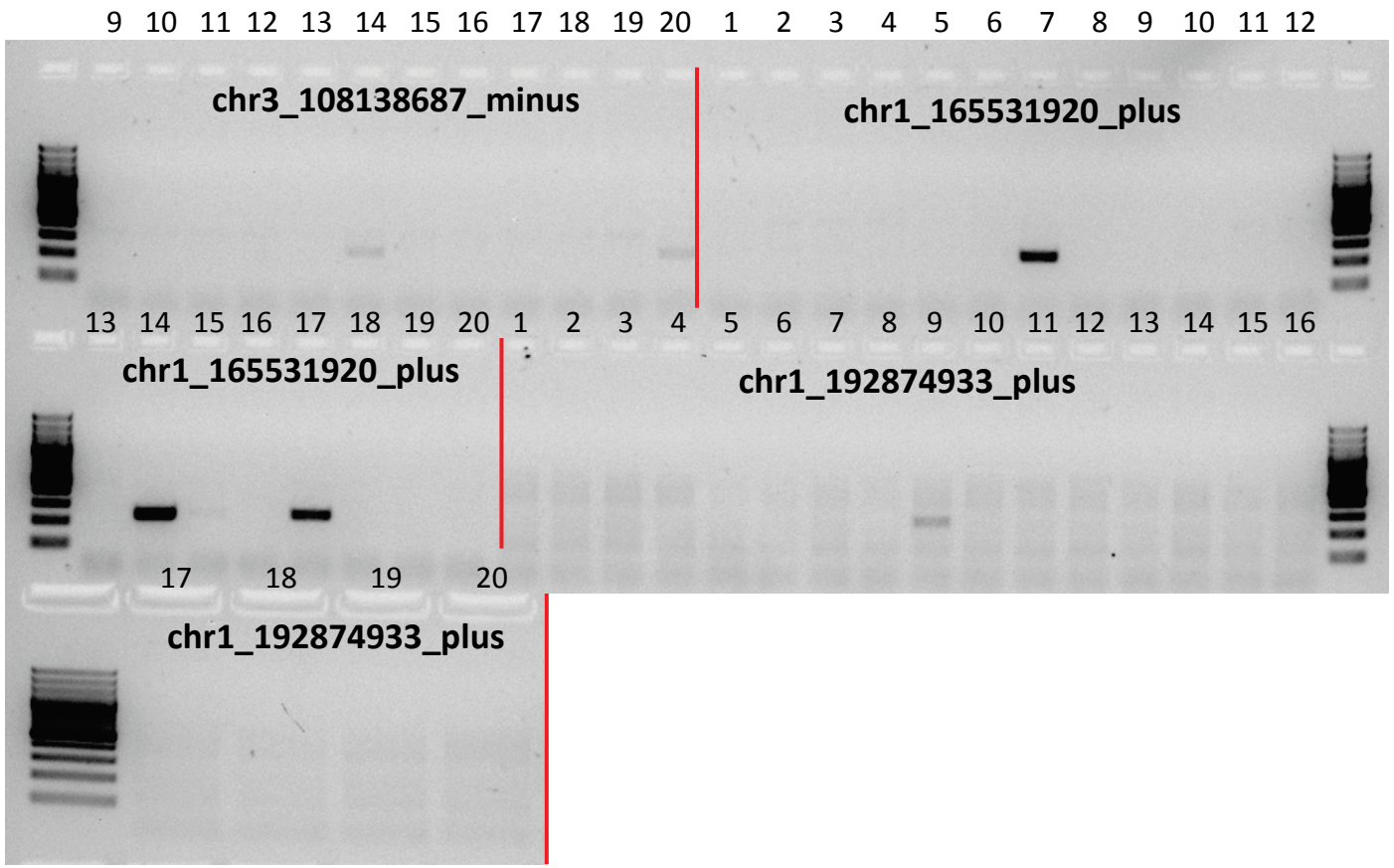

9 10 11 12 13 14 15 16 17 18 19 20
chr6_77222572_minus

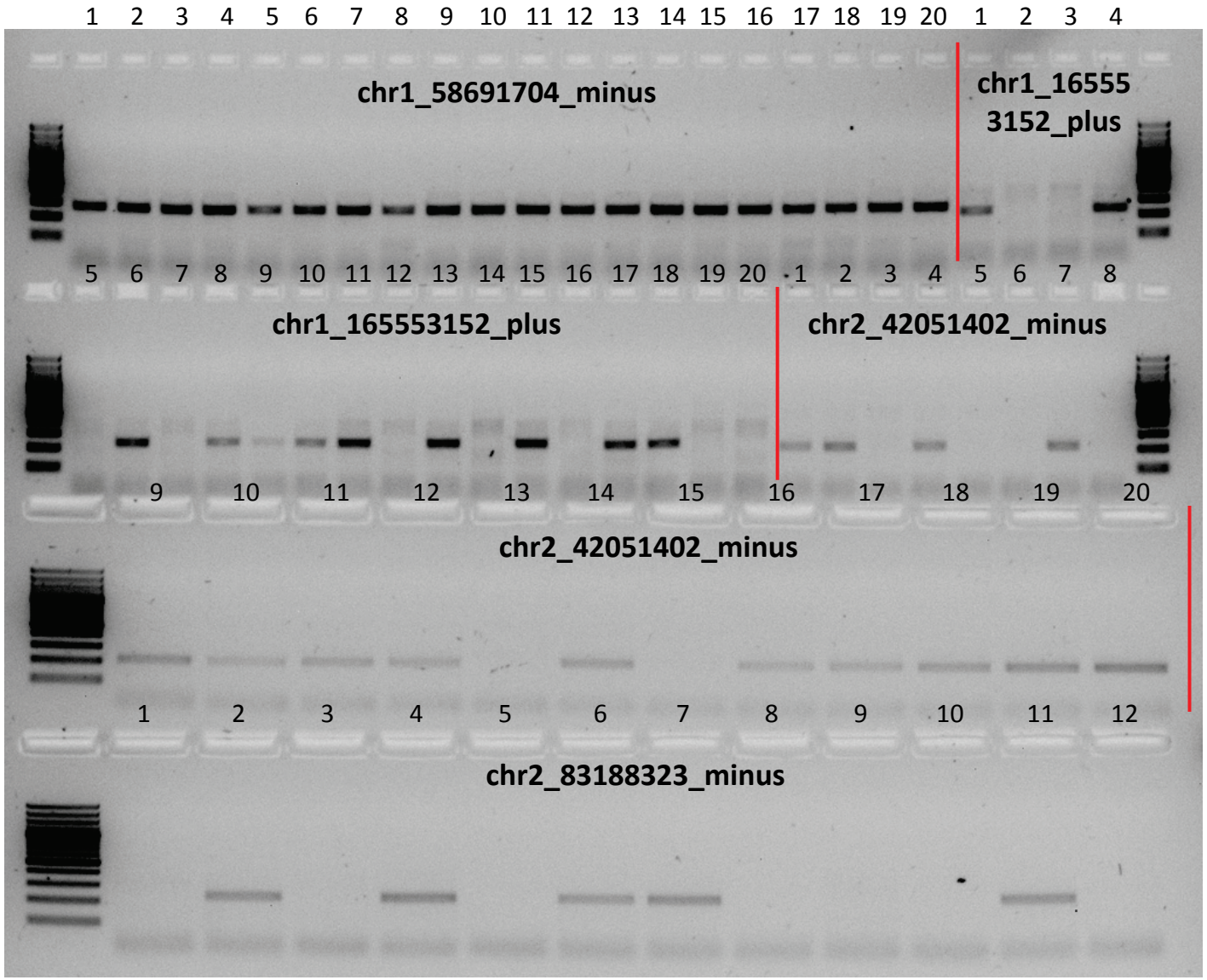

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4
chr2_192493518_plus
chr3_8214
4860_plus
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8
chr3_82144860_plus
chr4_18194359_plus
9 10 11 12 13 14 15 16 17 18 19 20
chr4_18194359_plus
1 2 3 4 5 6 7 8 9 10 11 12
chr4_130361674_plus

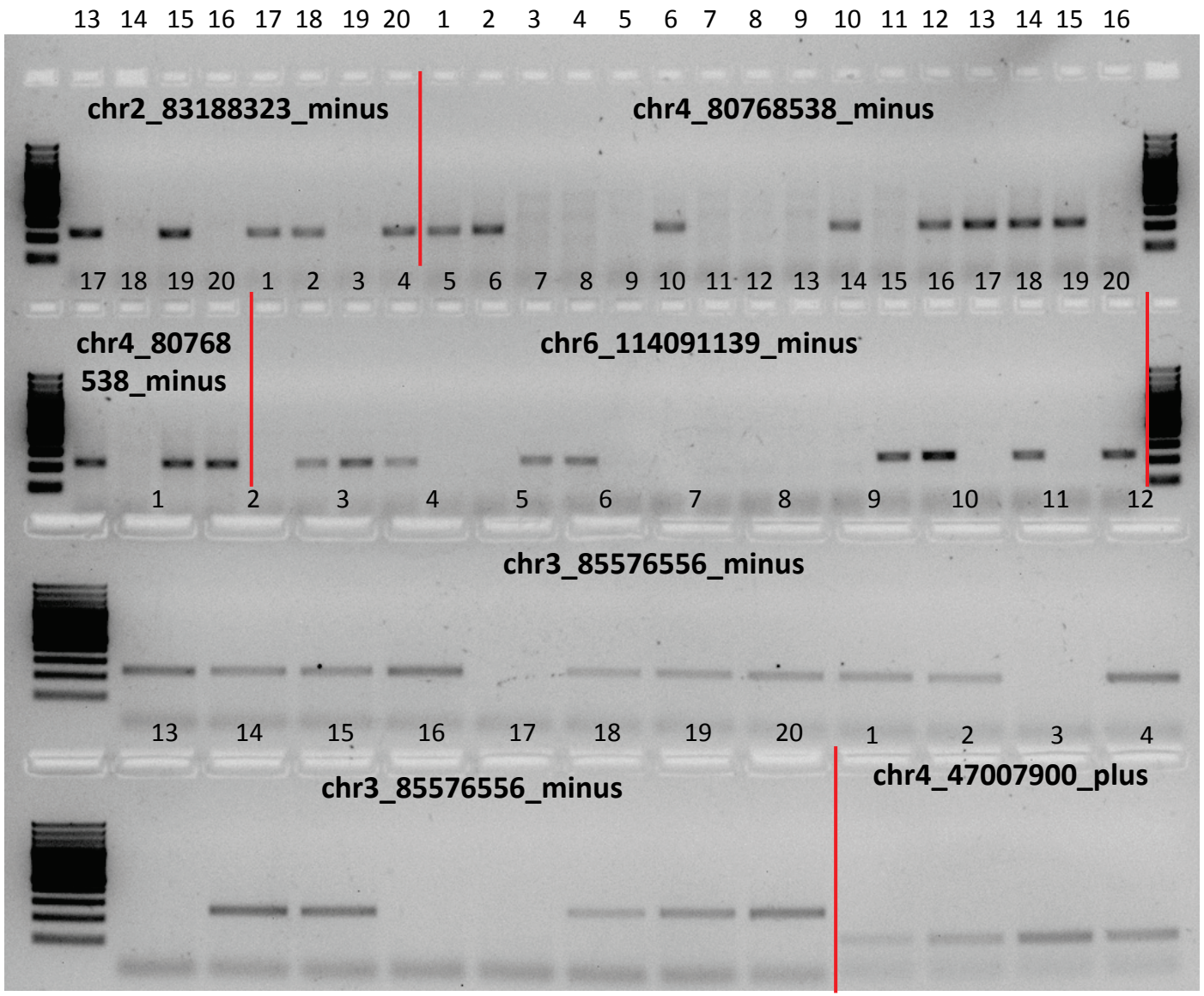

13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
chr4_130361674_plus
chr4_75917687_minus
17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
chr4_75917
687_minus
chr6_74935597_plus
1 2 3 4 5 6 7 8 9 10 11 12
chr7_118993490_minus
13 14 15 16 17 18 19 20 1 2 3 4
chr7_118993490_minus
chr7_153509409_minus

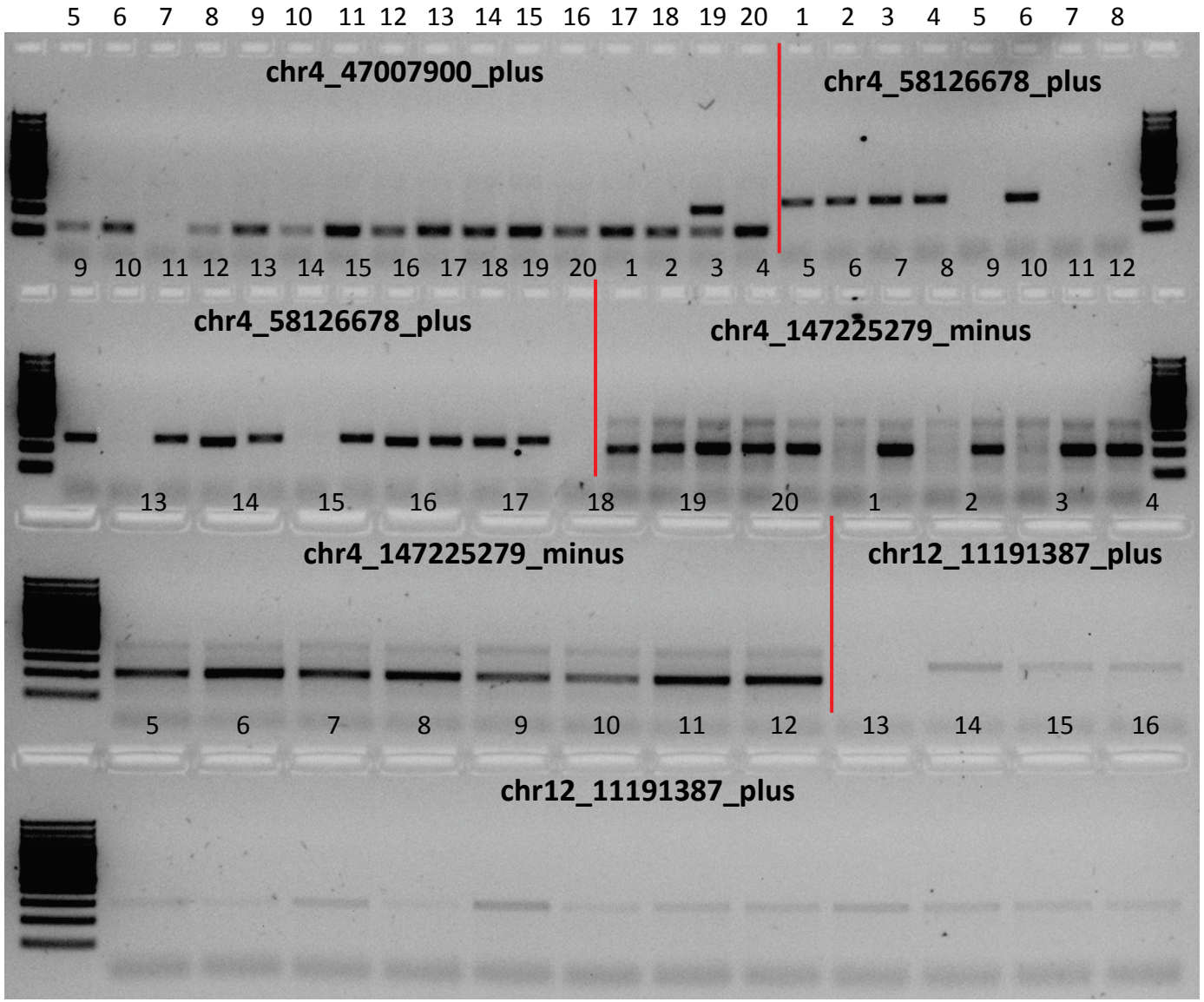

5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8
chr7_153509409_minus
chr11_97434312_plus
9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12
chr11_97434312_plus
chr12_33325810_plus
13 14 15 16 17 18 19 20 1 2 3 4
chr12_33325810_plus
chr1_75192913_plus
5 6 7 8 9 10 11 12 13 14 15 16
chr1_75192913_plus

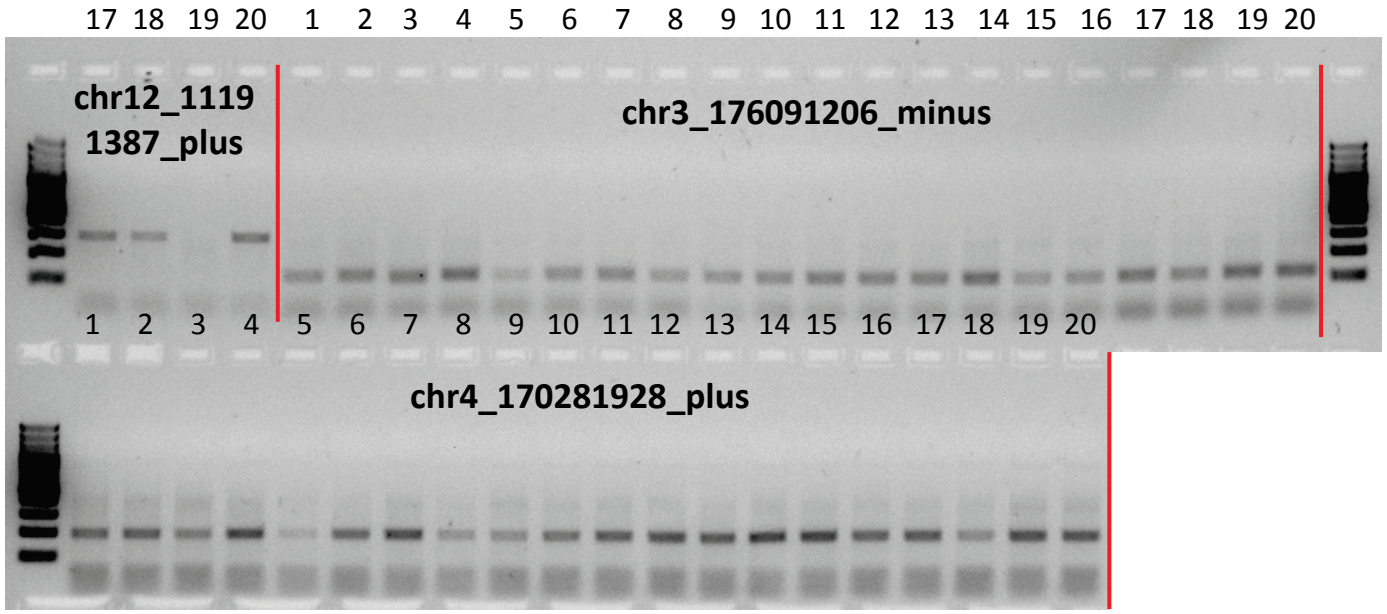

17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
chr1_7519
2913_plus
chr1_98363699_plus
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4
chr2_1482203_plus
chr2_206147
278_minus
5 6 7 8 9 10 11 12 13 14 15 16
chr2_206147278_minus
17 18 19 20 1 2 3 4 5 6 7 8
chr2_206147278_minus
chr3_108138687_minus

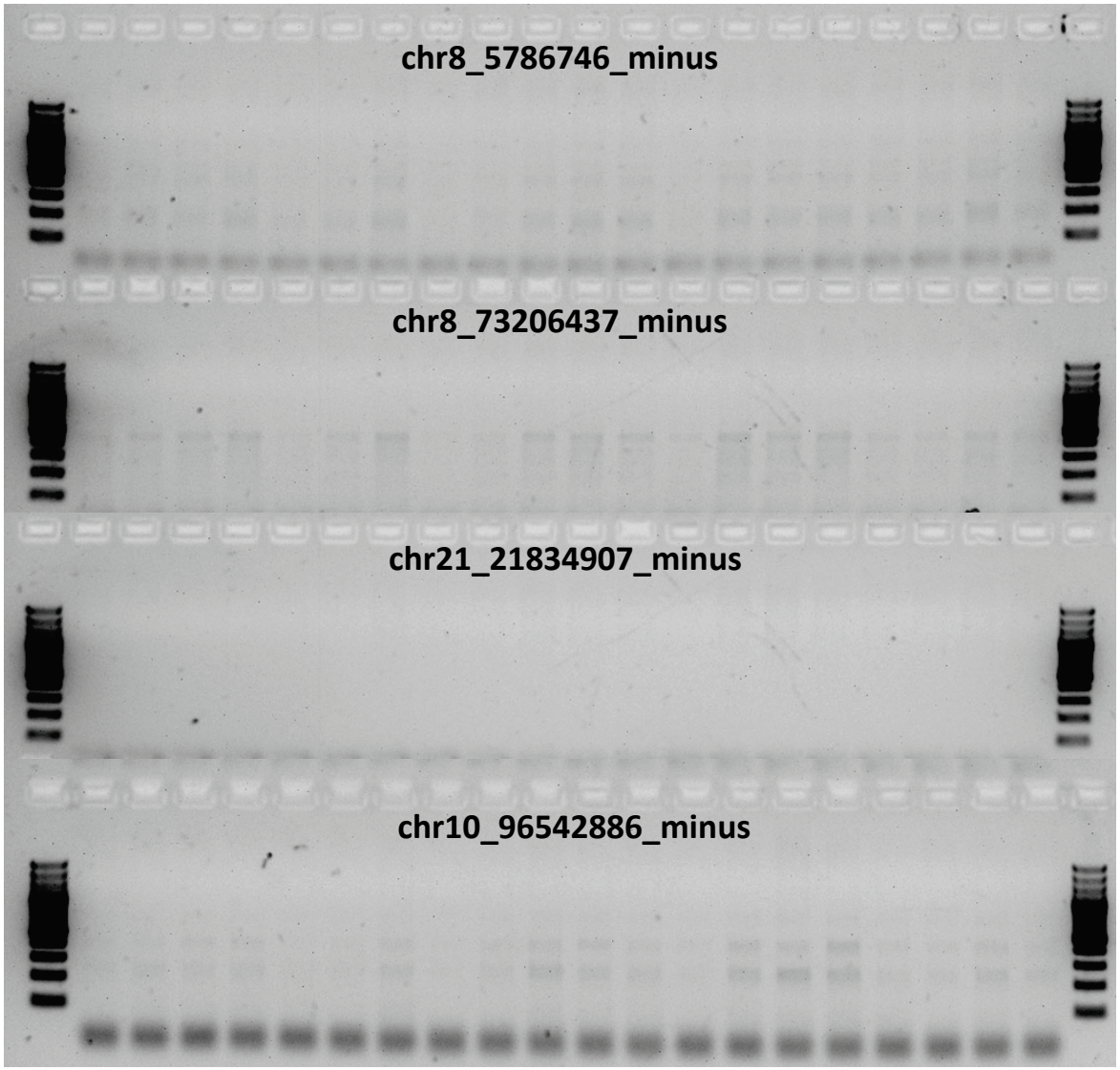

9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12
chr3_108138687_minus
chr1_165531920_plus
13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
chr1_165531920_plus
chr1_192874933_plus
17 18 19 20
chr1_192874933_plus

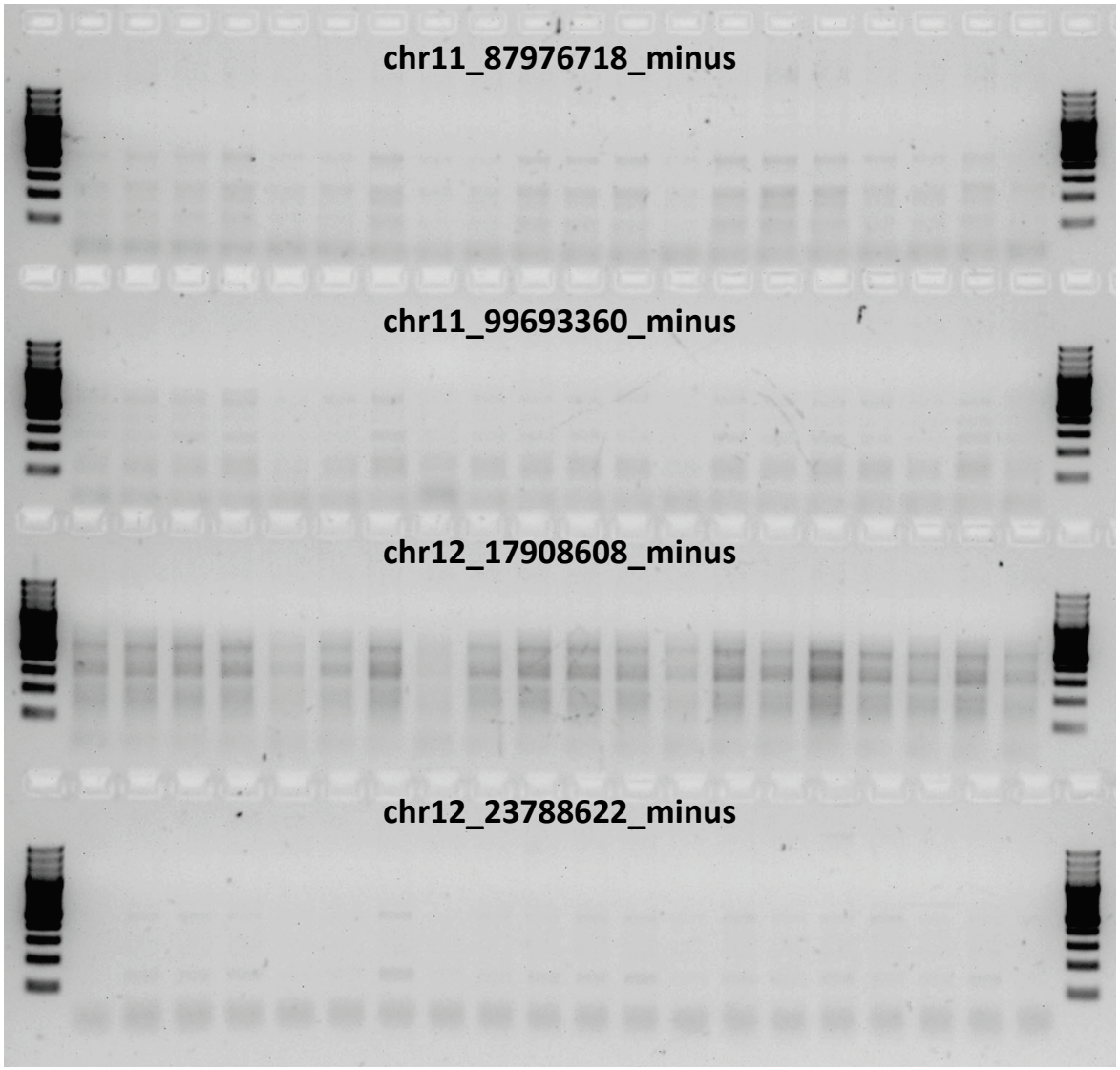

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4
chr1_58691704_minus
chr1_16555
3152_plus
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8
chr1_165553152_plus
chr2_42051402_minus
9 10 11 12 13 14 15 16 17 18 19 20
chr2_42051402_minus
1 2 3 4 5 6 7 8 9 10 11 12
chr2_83188323_minus

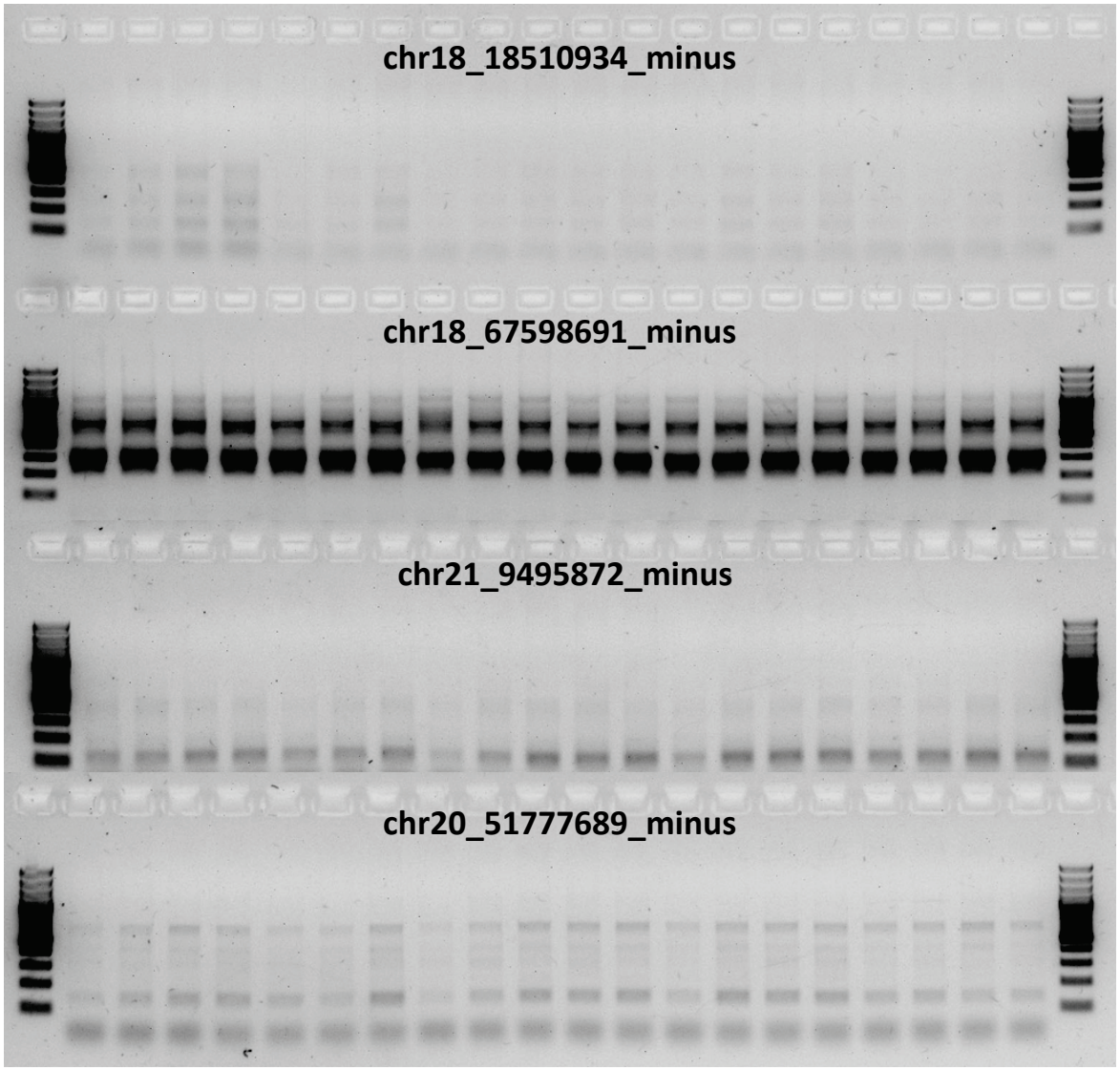

13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
chr2_83188323_minus
chr4_80768538_minus
17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
chr4_80768538_minus
chr6_114091139_minus
1 2 3 4 5 6 7 8 9 10 11 12
chr3_85576556_minus
13 14 15 16 17 18 19 20 1 2 3 4
chr3_85576556_minus
chr4_47007900_plus

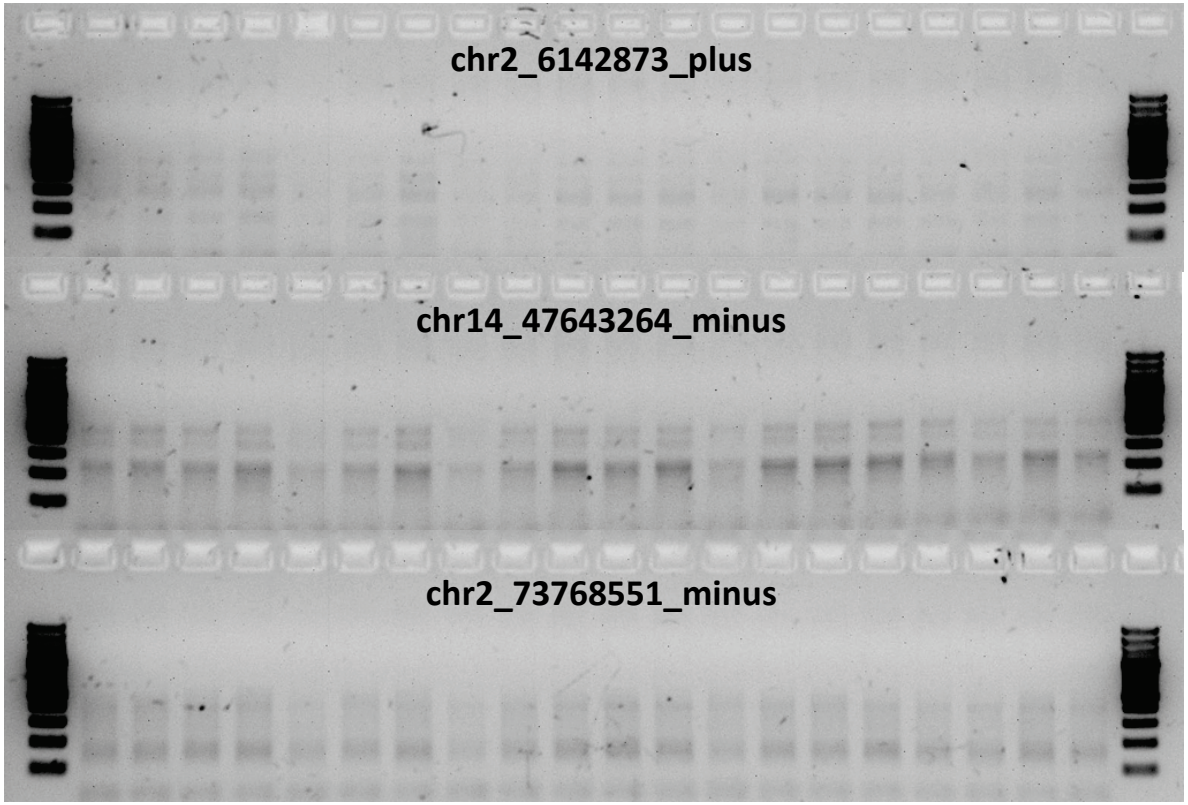

5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8
chr4_47007900_plus
chr4_58126678_plus
9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12
chr4_58126678_plus
chr4_147225279_minus
13 14 15 16 17 18 19 20 1 2 3 4
chr4_147225279_minus
chr12_11191387_plus
5 6 7 8 9 10 11 12 13 14 15 16
chr12_11191387_plus


17 18 19 20 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
chr12_1119
1387_plus
chr3_176091206_minus
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
chr4_170281928_plus


chr8_5786746_minus
chr8_73206437_minus
chr21_21834907_minus
chr10_96542886_minus


chr11_87976718_minus
chr11_99693360_minus
chr12_17908608_minus
chr12_23788622_minus


chr18_18510934_minus
chr18_67598691_minus
chr21_9495872_minus
chr20_51777689_minus


chr2_6142873_plus
chr14_47643264_minus
chr2_73768551_minus

4. Gel electrophoresis analyses of site-specific PCR validations for 47 L1s from the 1000GP

This part contains gel electrophoresis analyses of PCR reactions used to validate and genotype 47 L1s identified by the 1000GP. For each L1, each sample was tested with two or three pairs of primers depending on whether it was from the "deletion" or "insertion" set. This allowed us to obtain allelic genotypes for all L1s in all tested samples (40 CEU and/or 29 CHB samples). L1s of the "deletion" set were assayed with 2 PCR reactions testing for the absence (primers E and G) and the presence (L1Hs-specific and primer G) of the L1. The orientation of L1s in the "insertion" set was unknown and we tested for the presence of the L1 in either orientation (i.e. using either of the E and G primers in conjunction with the L1Hs-specific primer).

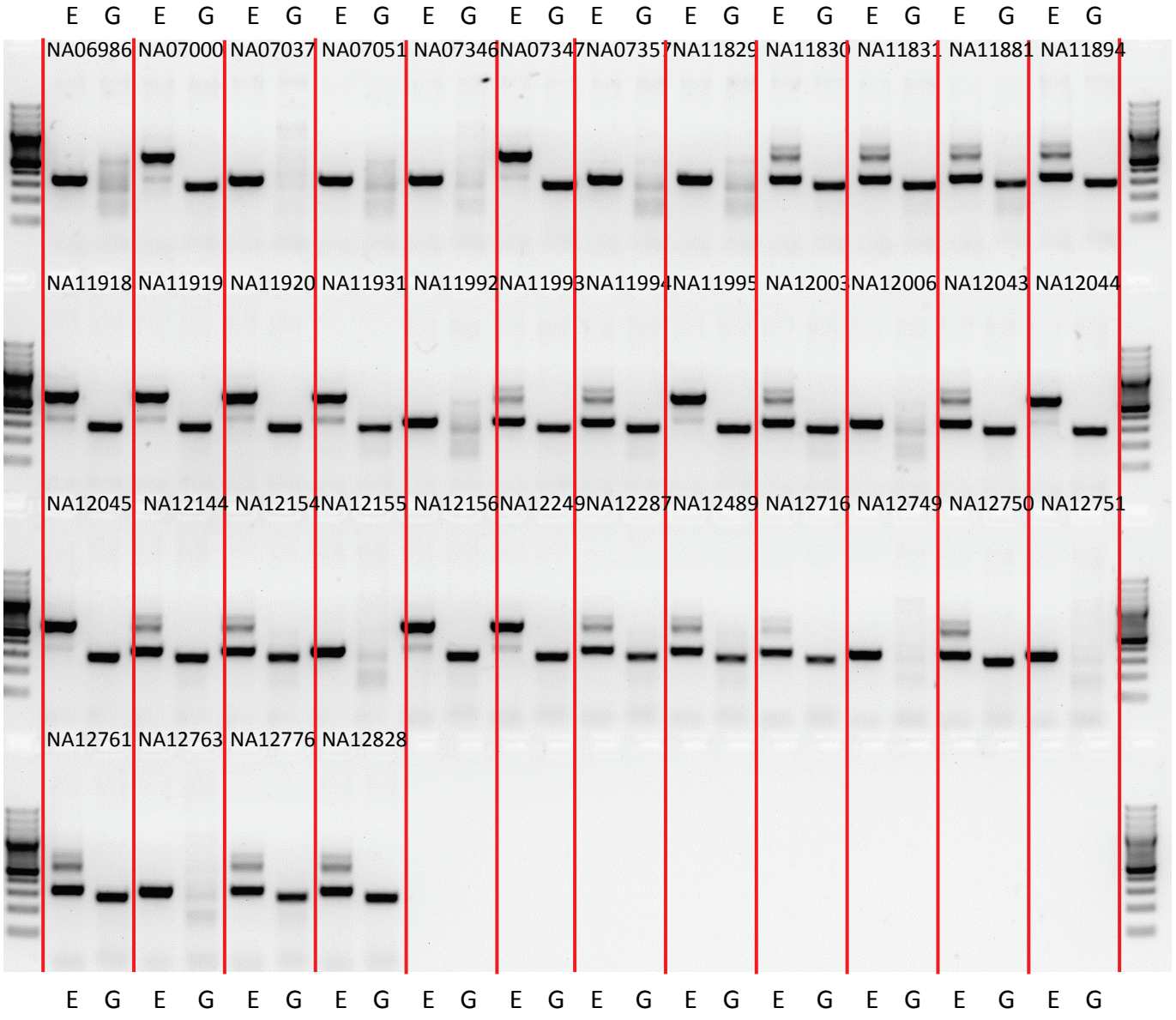
1) Reactions labelled “E” test for the absence of the L1 insertion (primers E and G), for L1s of both the "deletion" and "insertion" sets.
2) Reactions labelled “G1” test for the presence of the L1 insertion in one orientation (L1Hs-specific primer and primer E, reaction used for L1s of the "insertion" set only).
3) Reactions labelled “G” (L1s of the "deletion" set) or “G2” (L1s of the "insertion" set) test for the presence of the L1 insertion (L1Hs-specific primer and primer G) in the known orientation (L1s of the “deletion” set) or the other orientation (L1s of the “insertion” set).

CEU samples: Pages 2-65

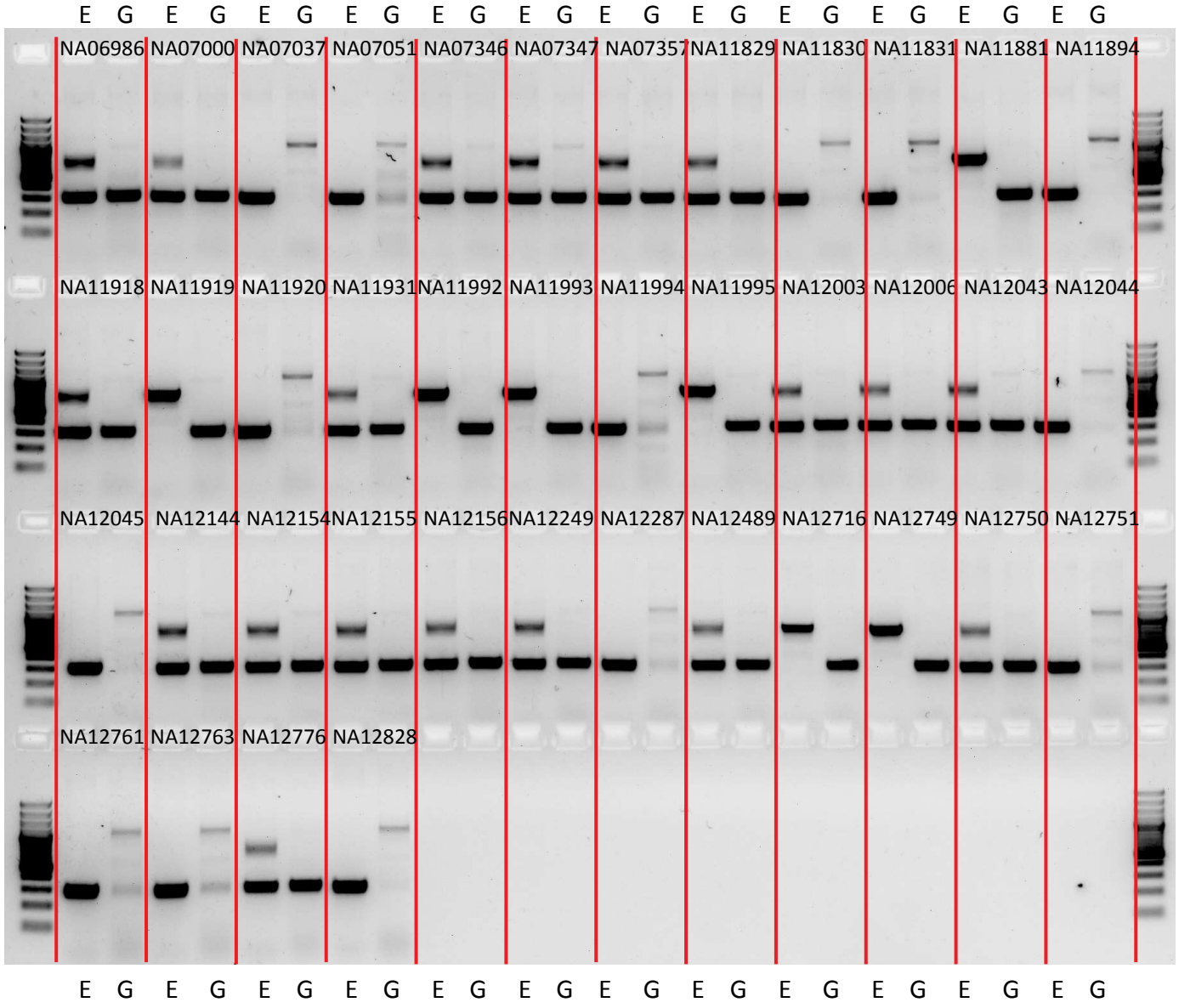
CHB samples: Pages 66-80

CEU

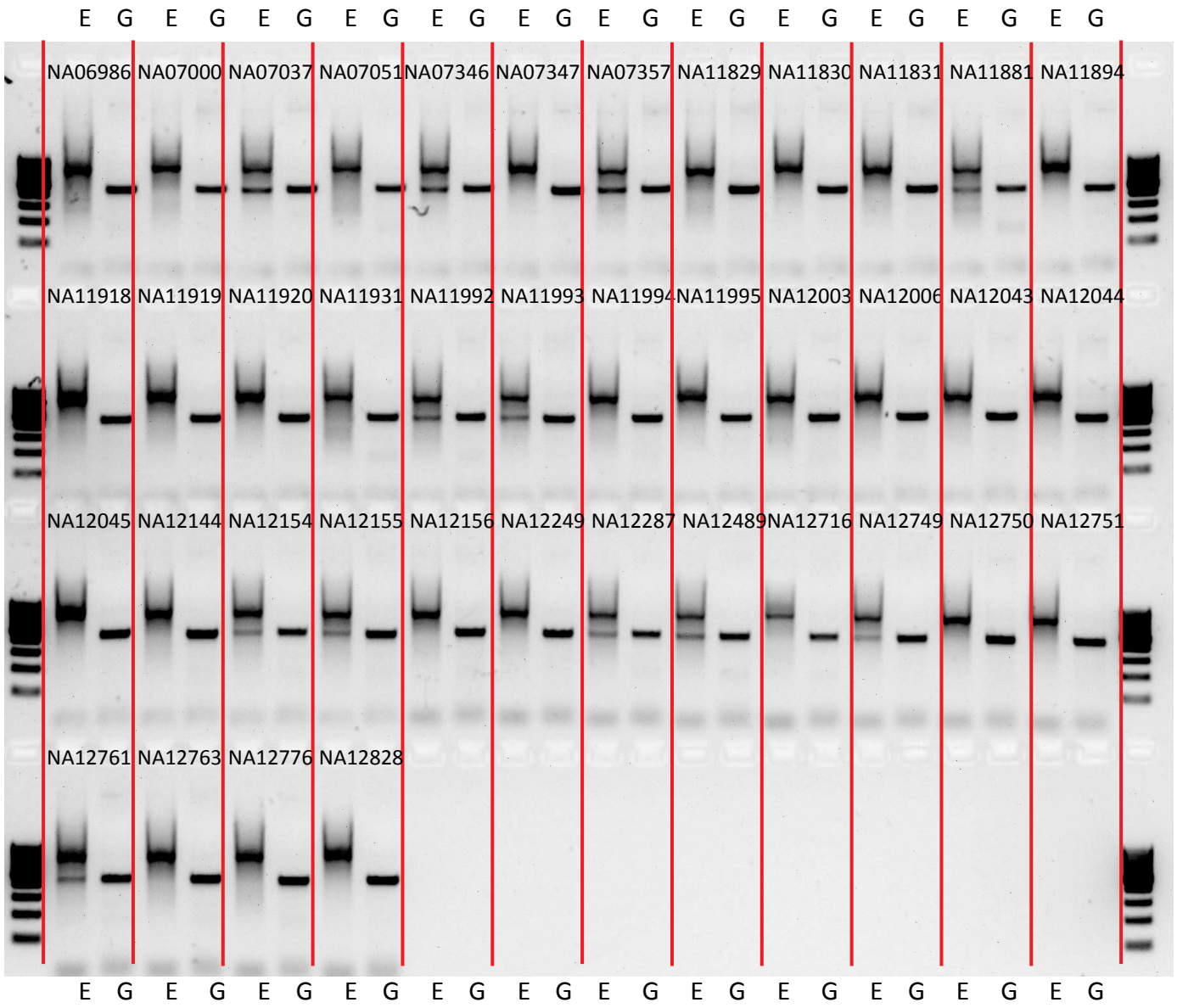
P1_M_061510_1_185



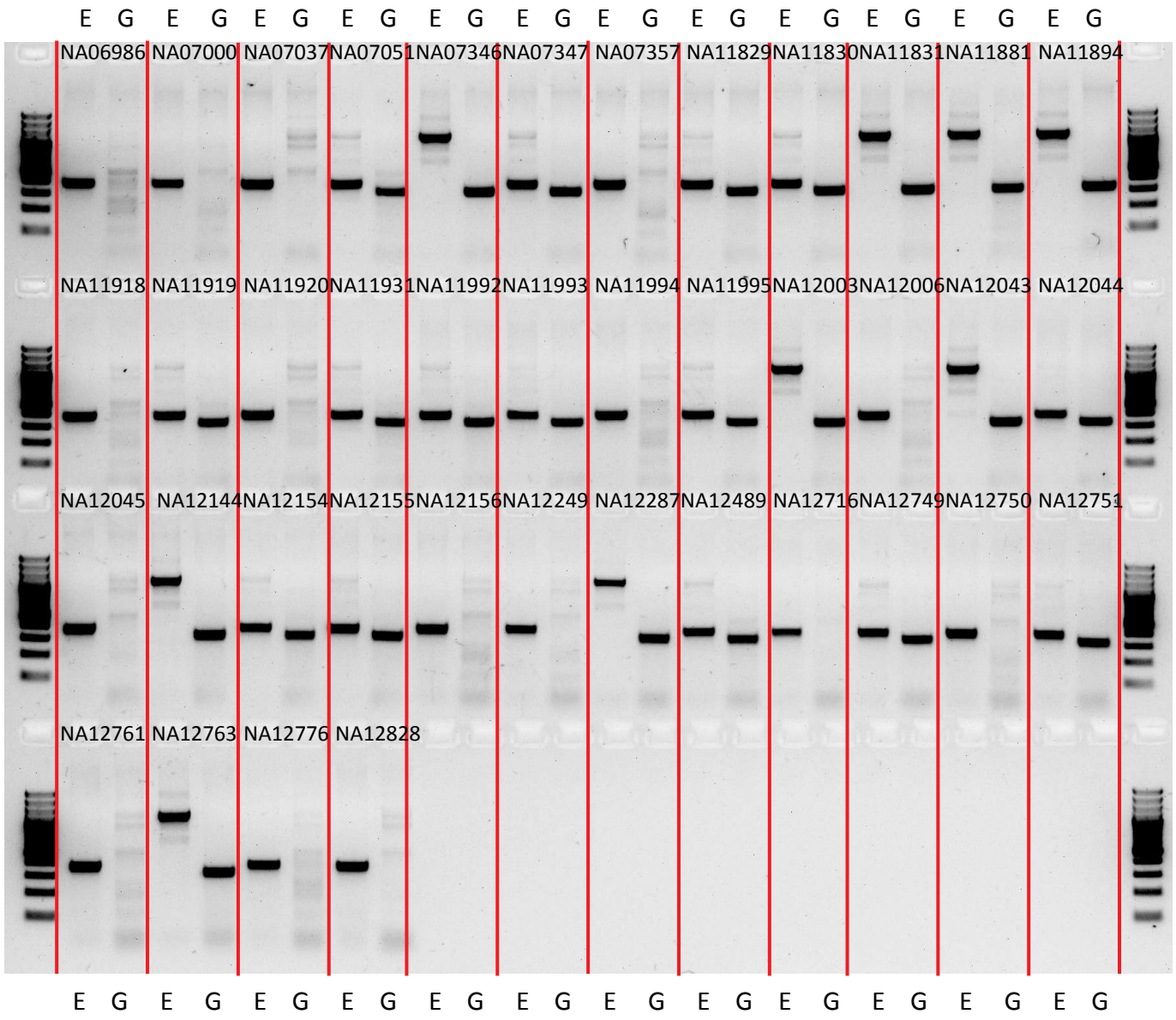
P1_M_061510_1_239



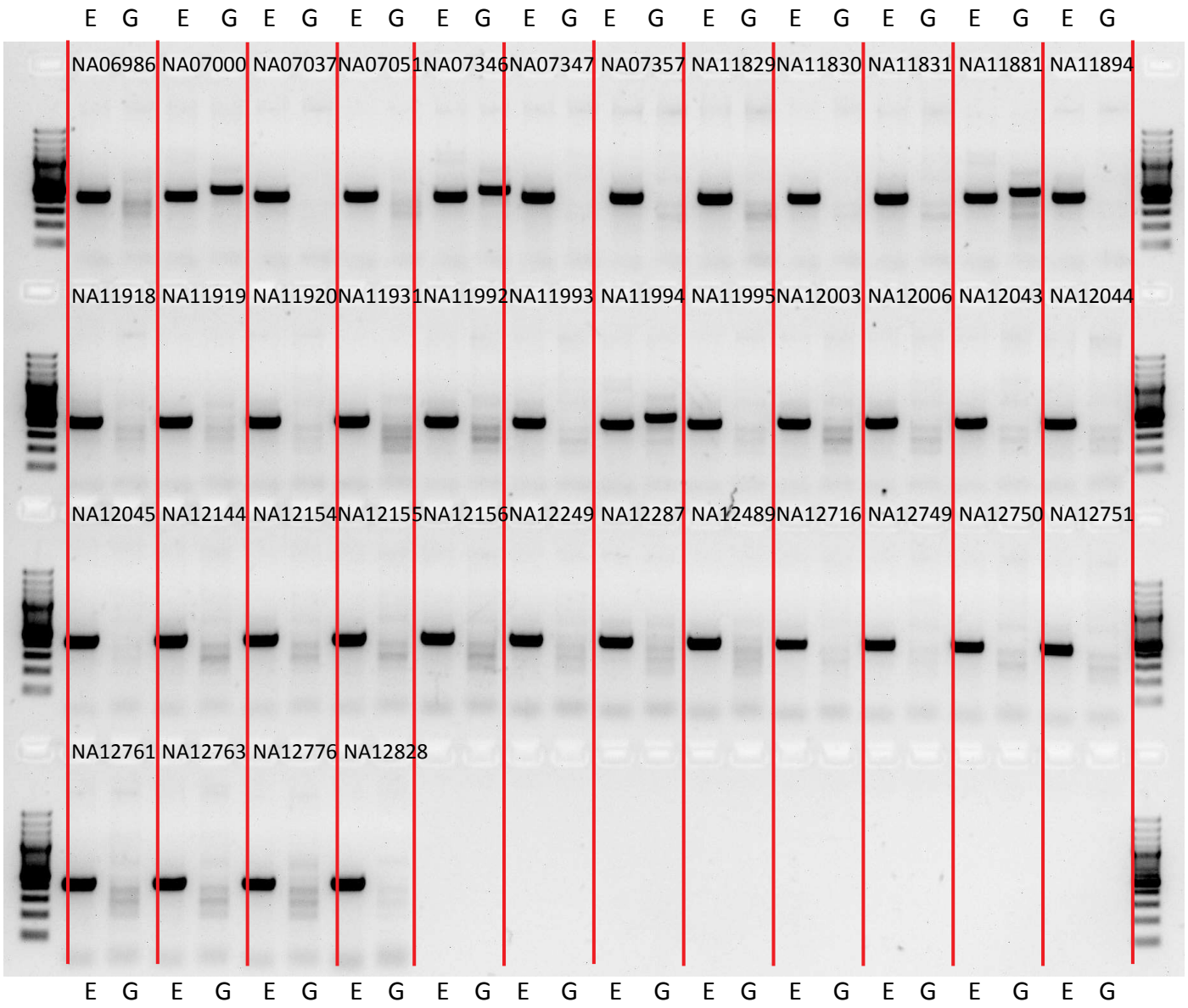
P1_M_061510_1_282



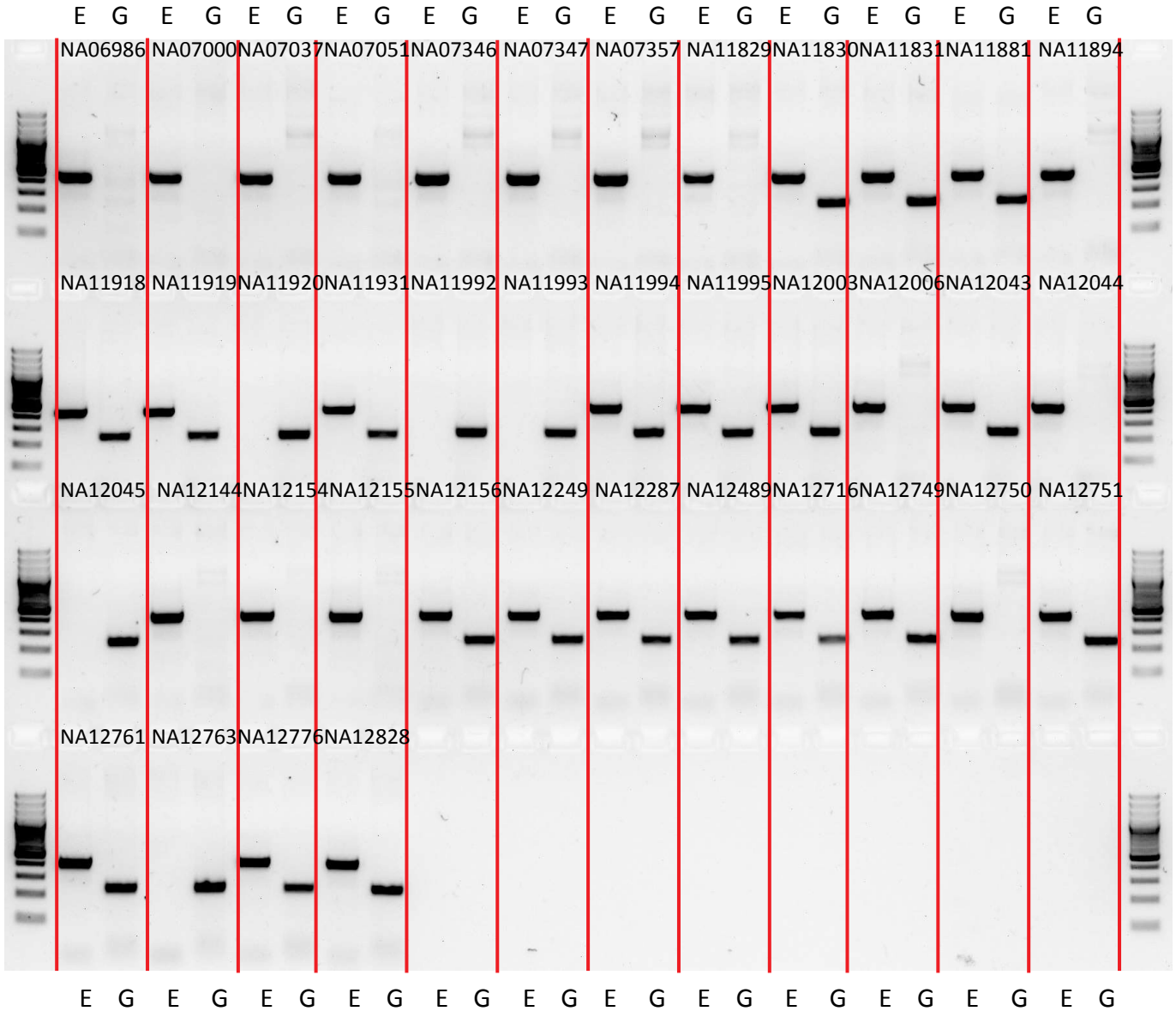
P1_M_061510_1_391



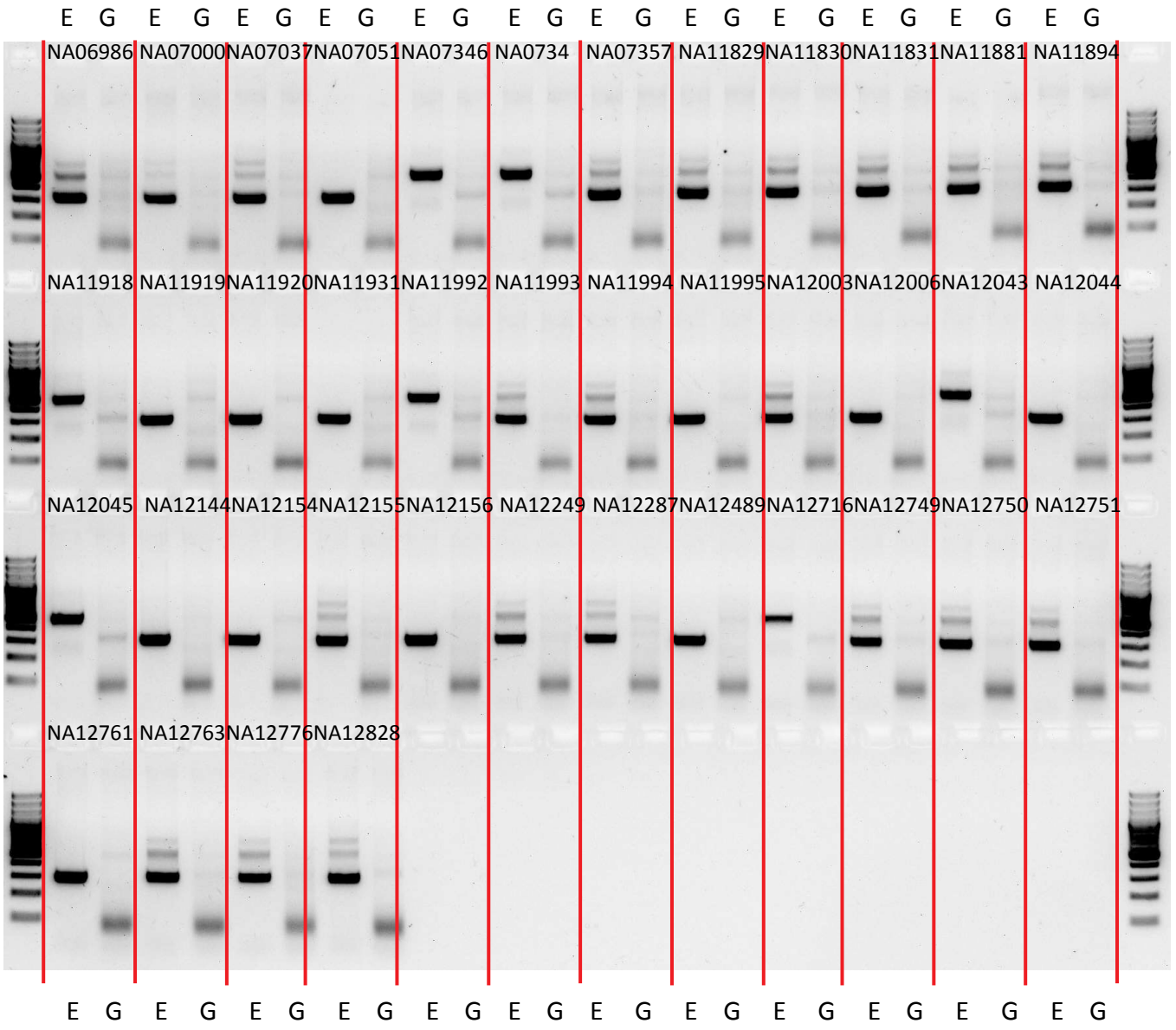
P1_M_061510_3_265



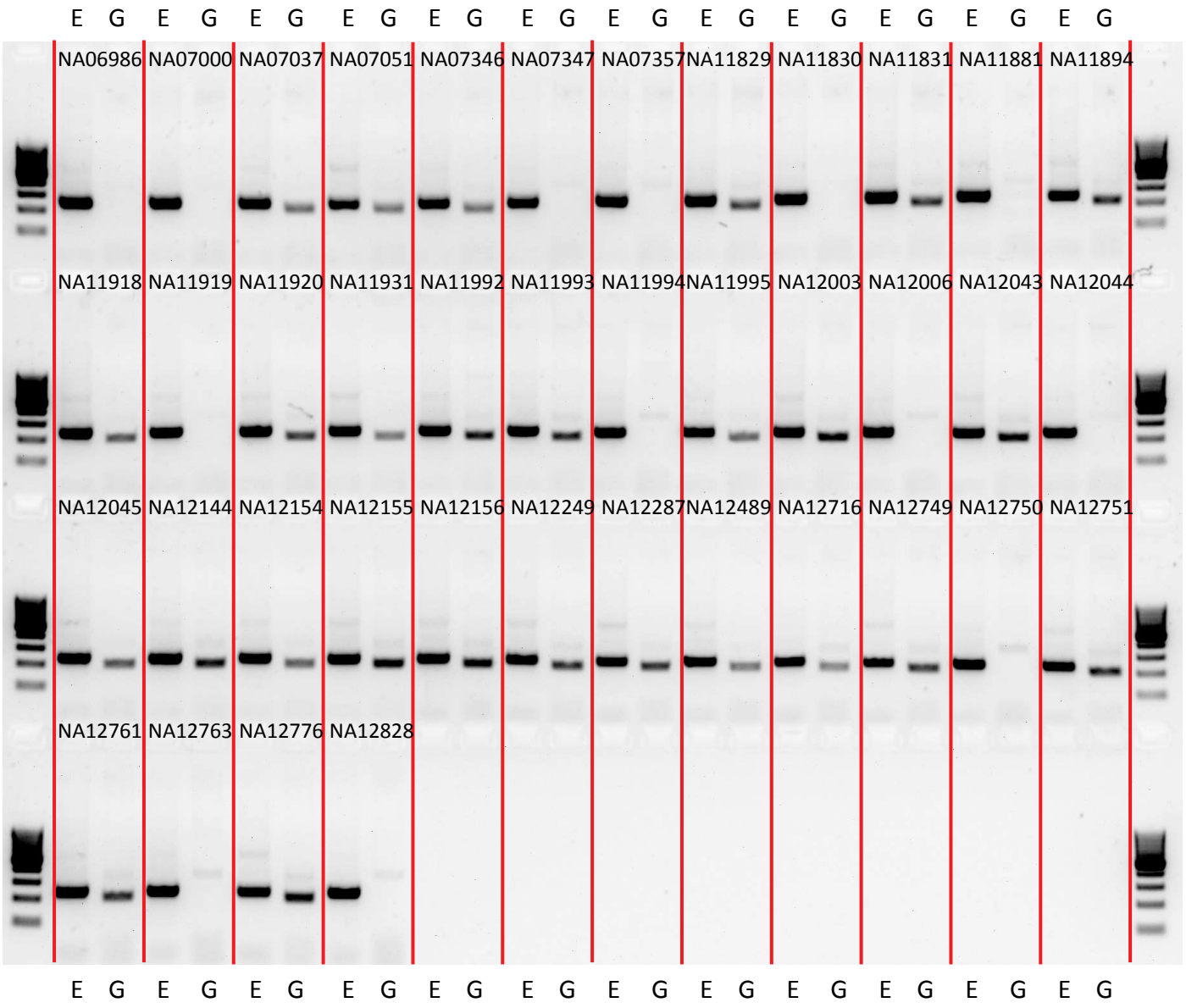
P1_M_061510_4_203



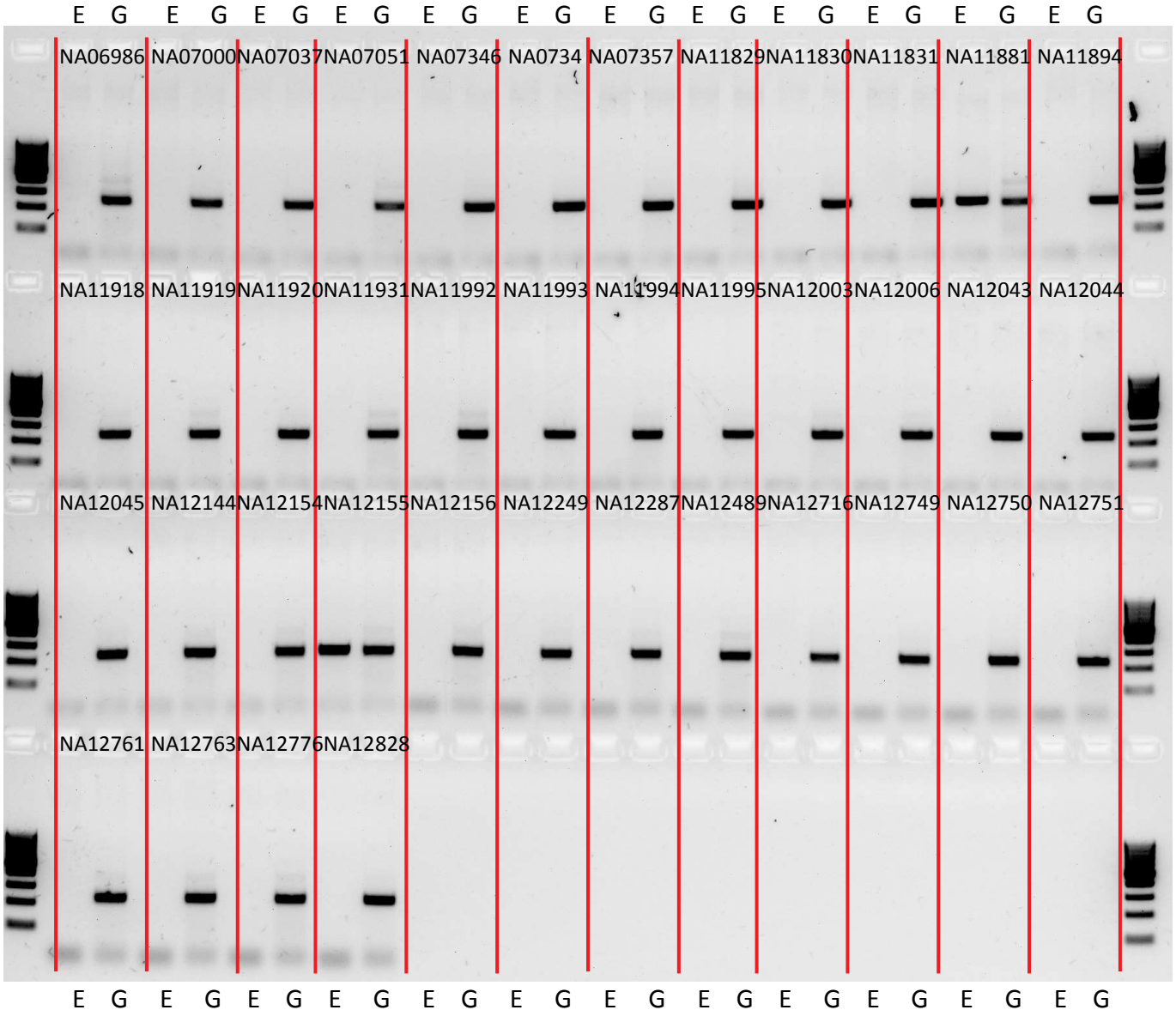
P1_M_061510_4_354



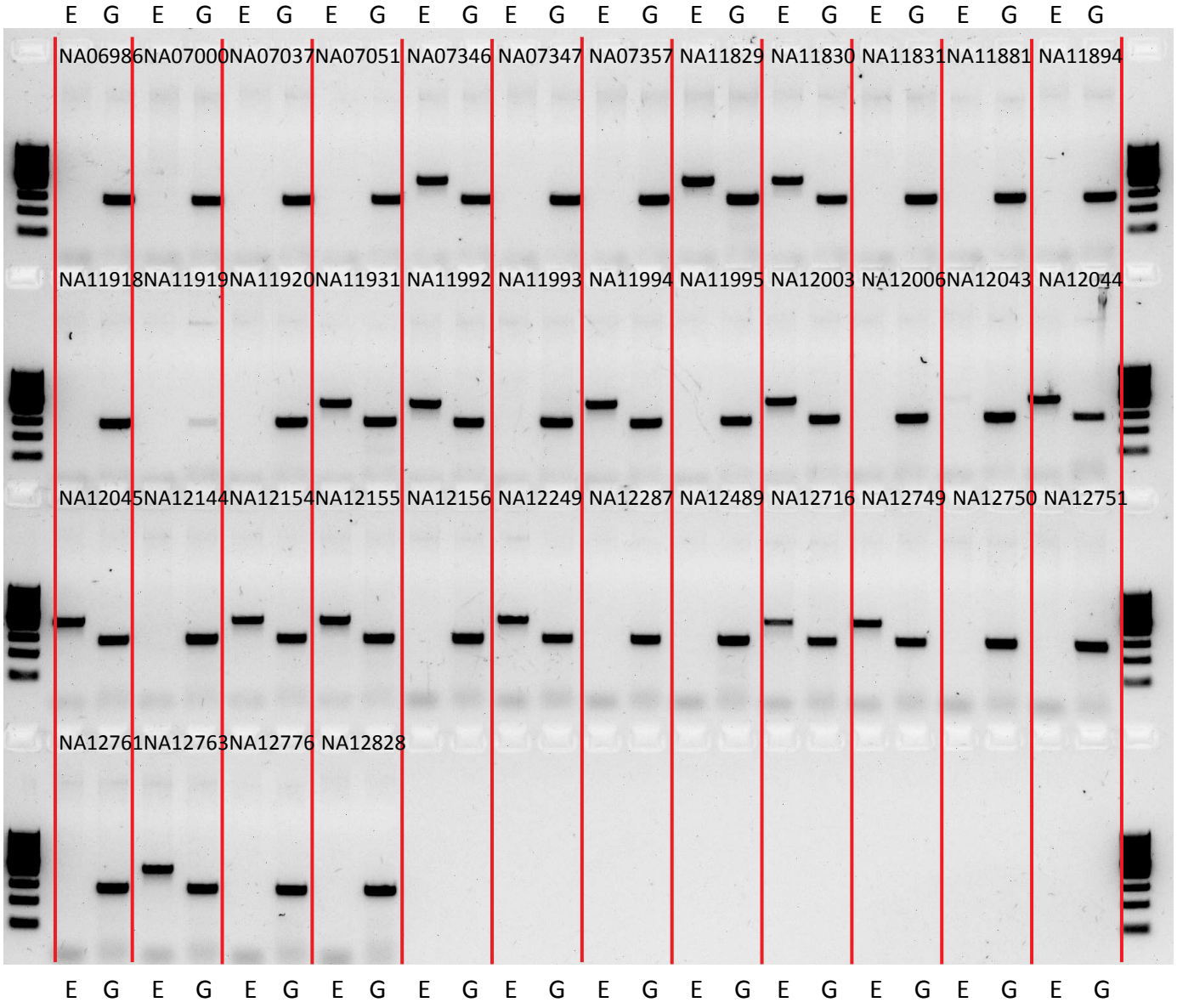
P1_M_061510_6_114



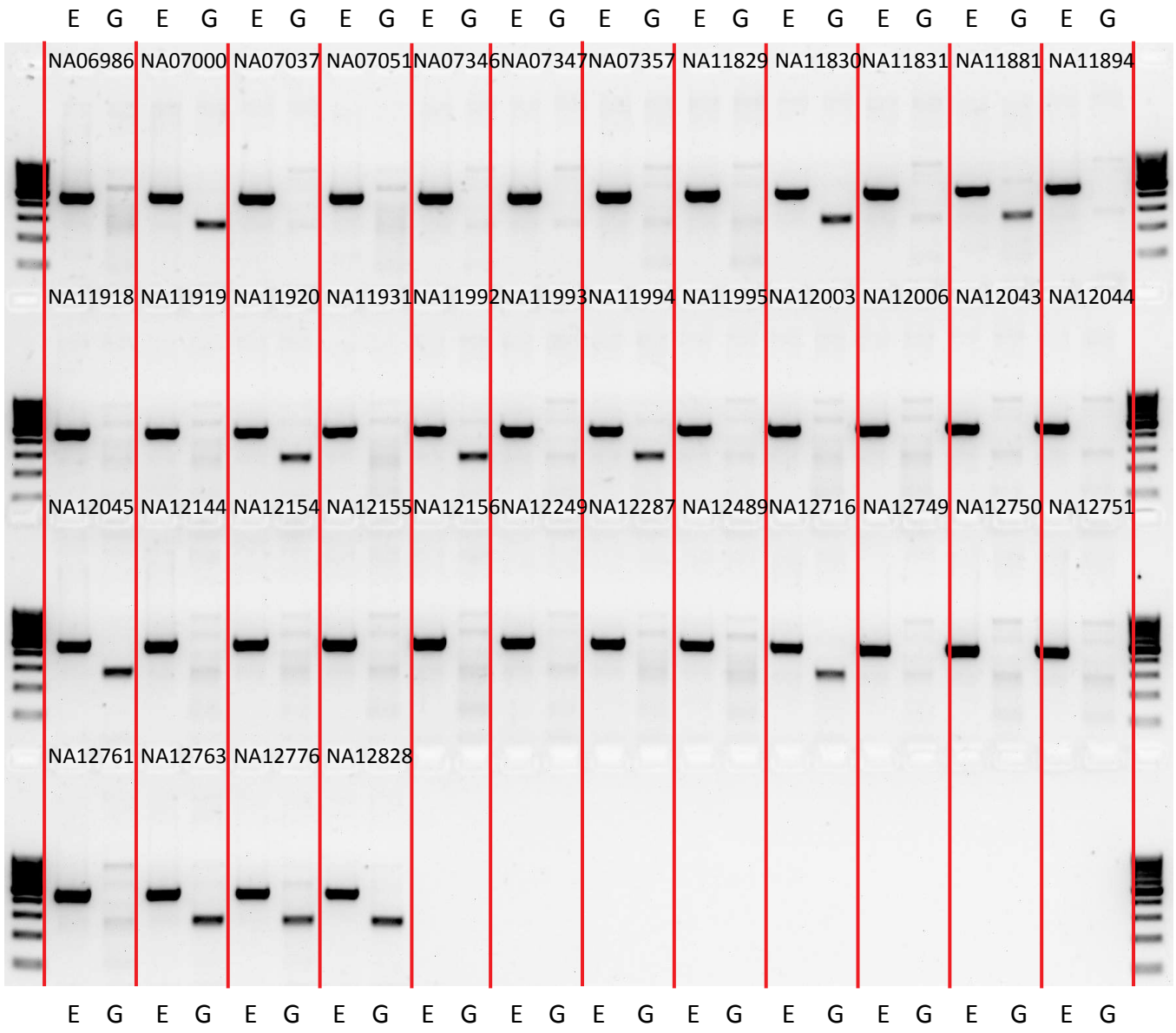
P1_M_061510_6_306



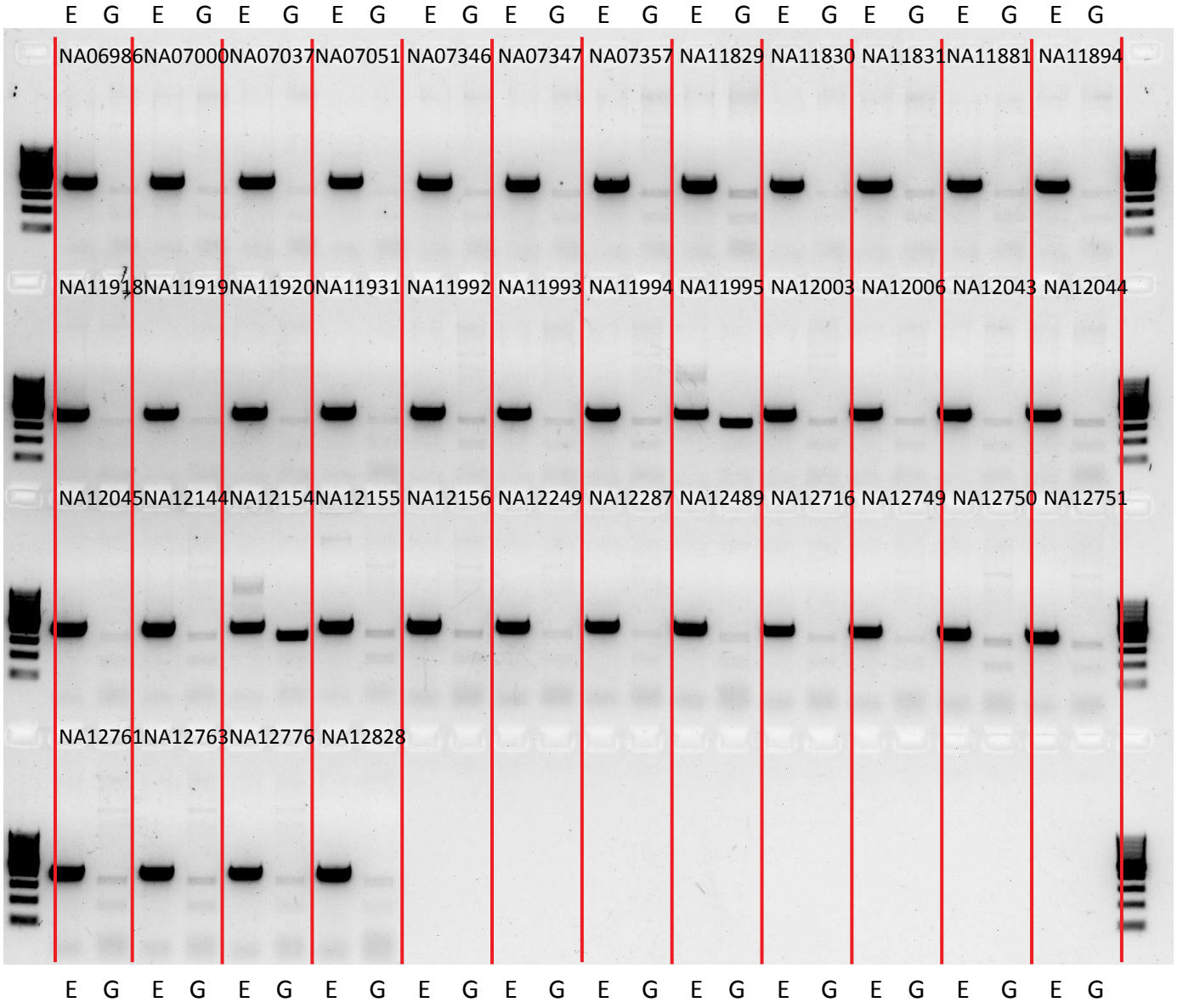
P1_M_061510_6_823



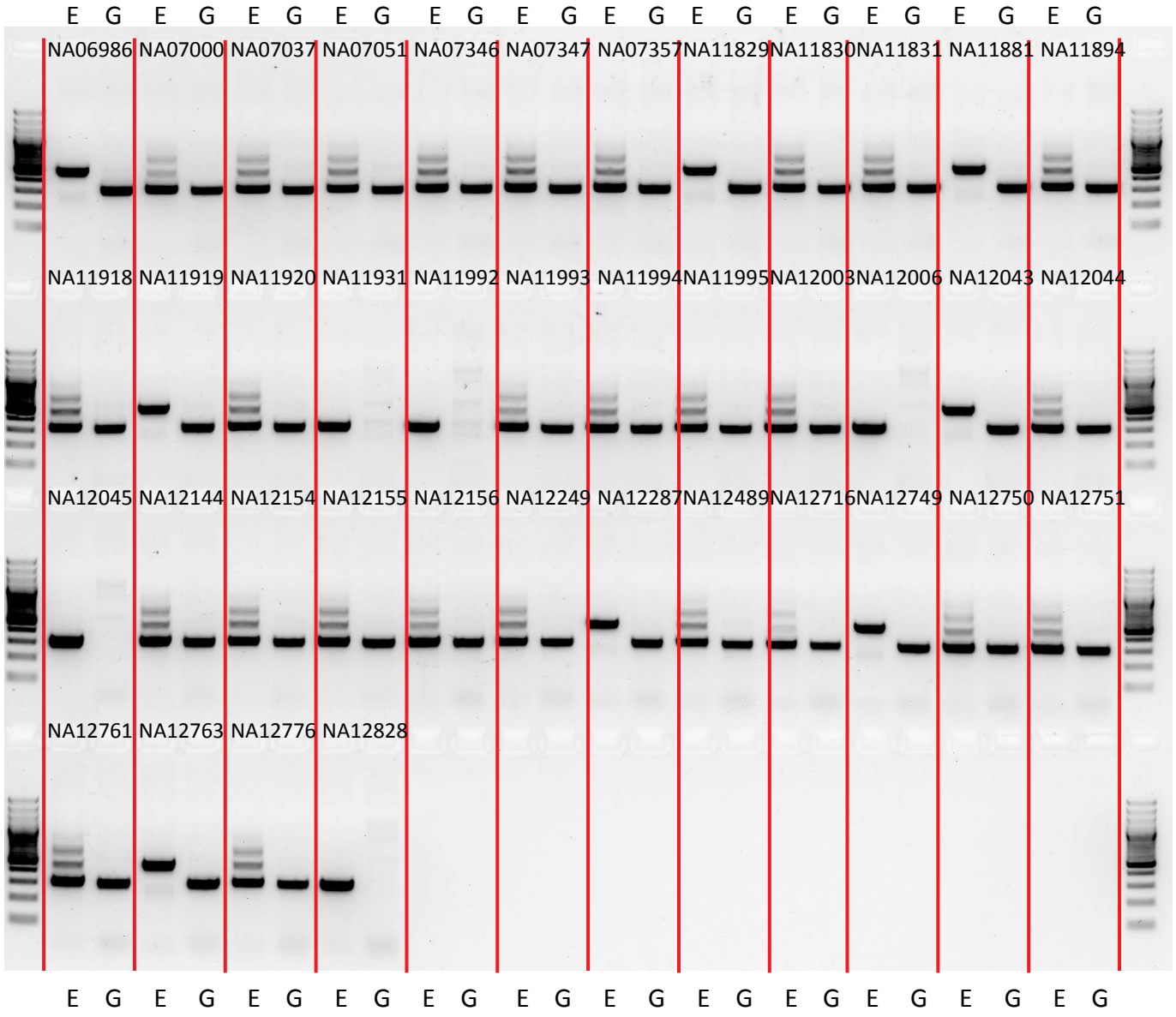
P1_M_061510_8_216



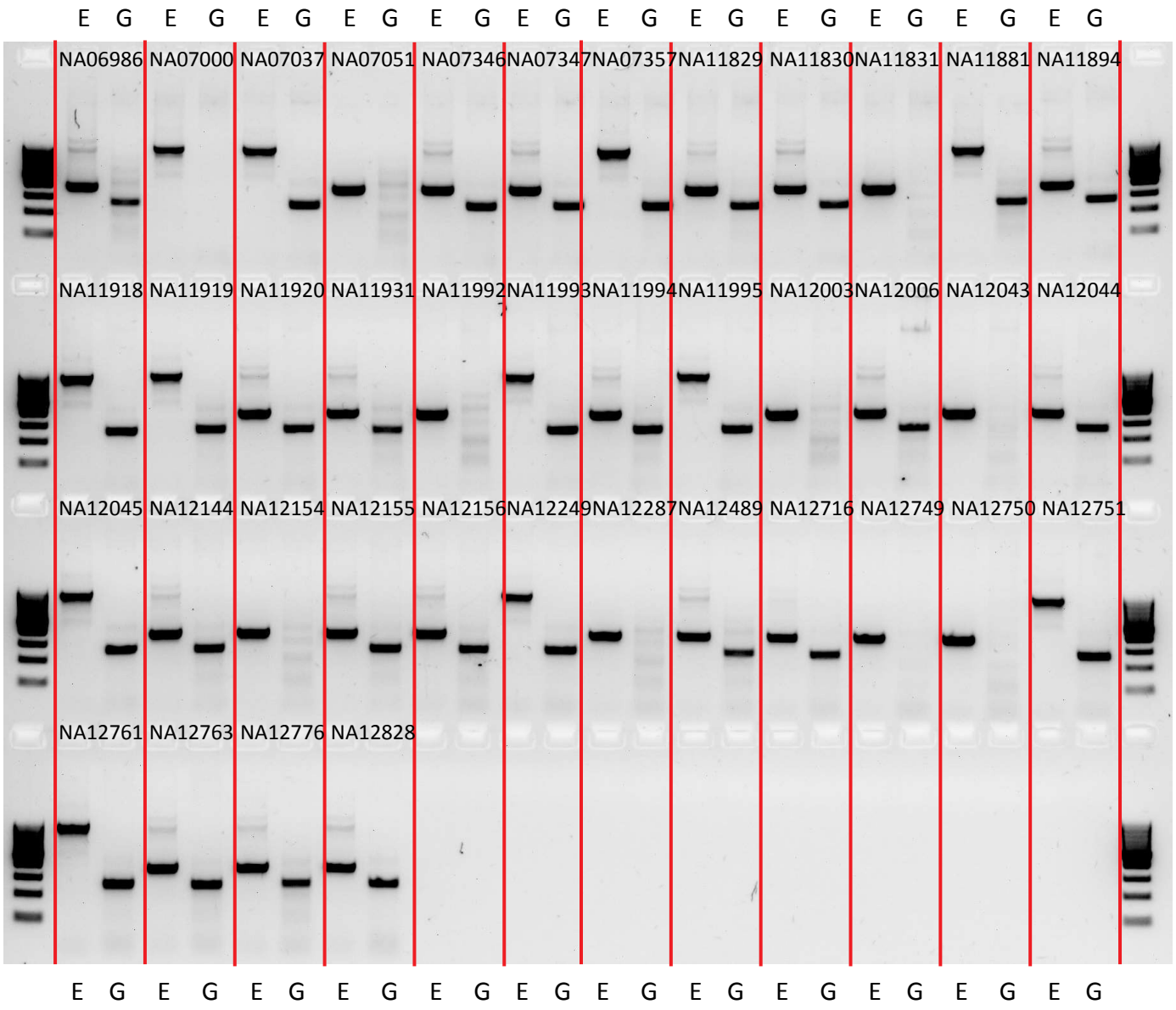
P1_M_061510_9_215



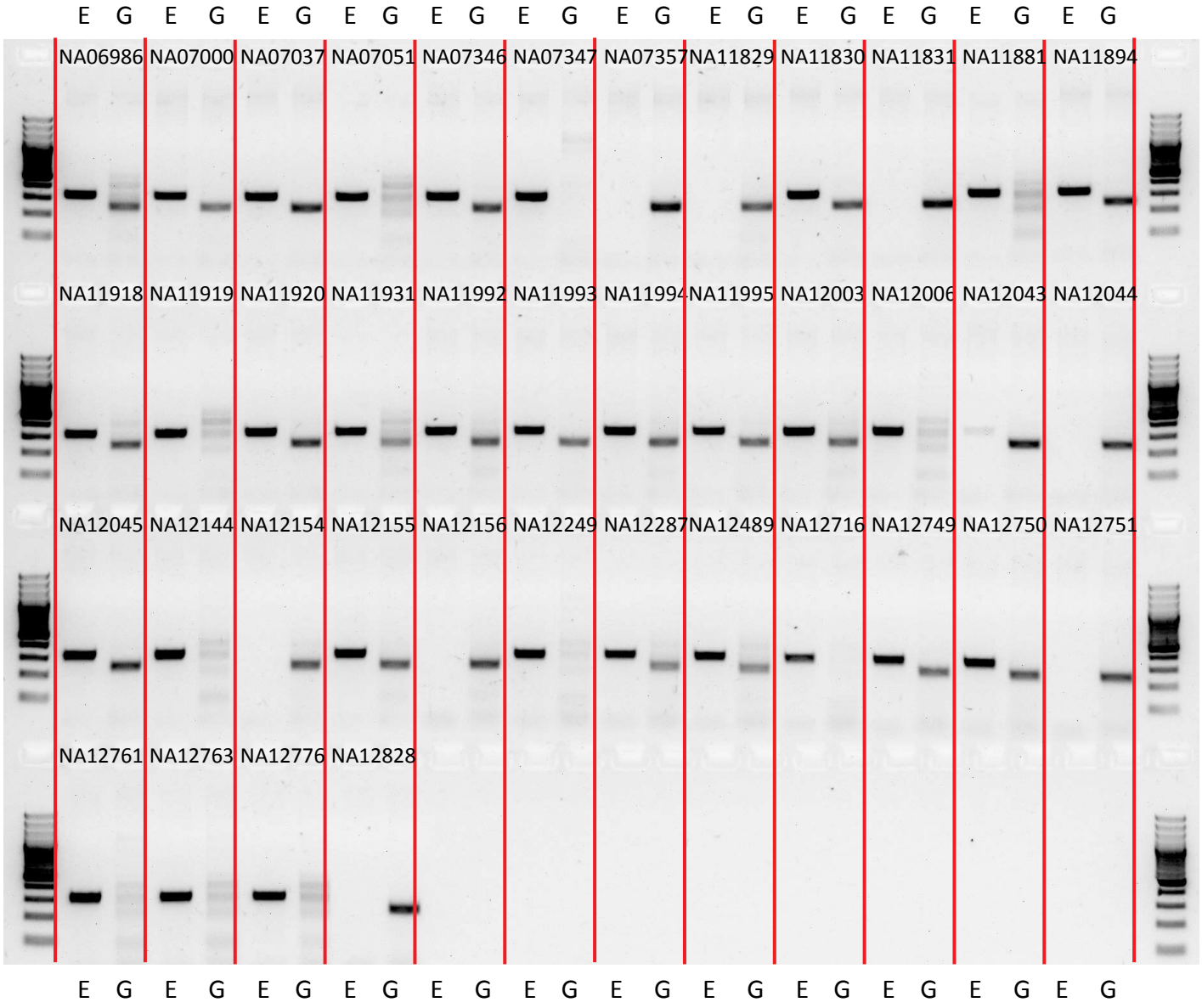
P1_M_061510_9_218



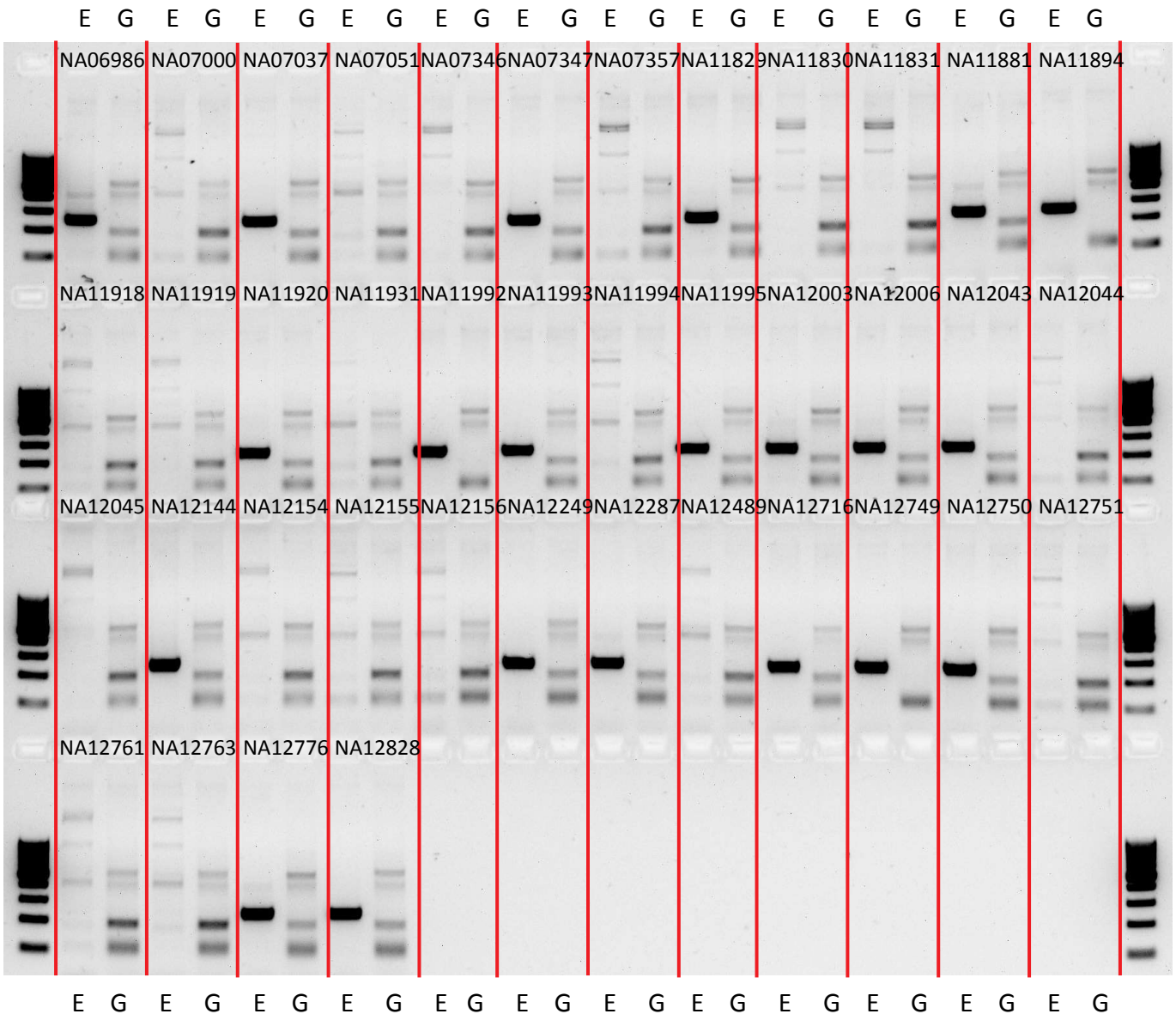
P1_M_061510_10_118



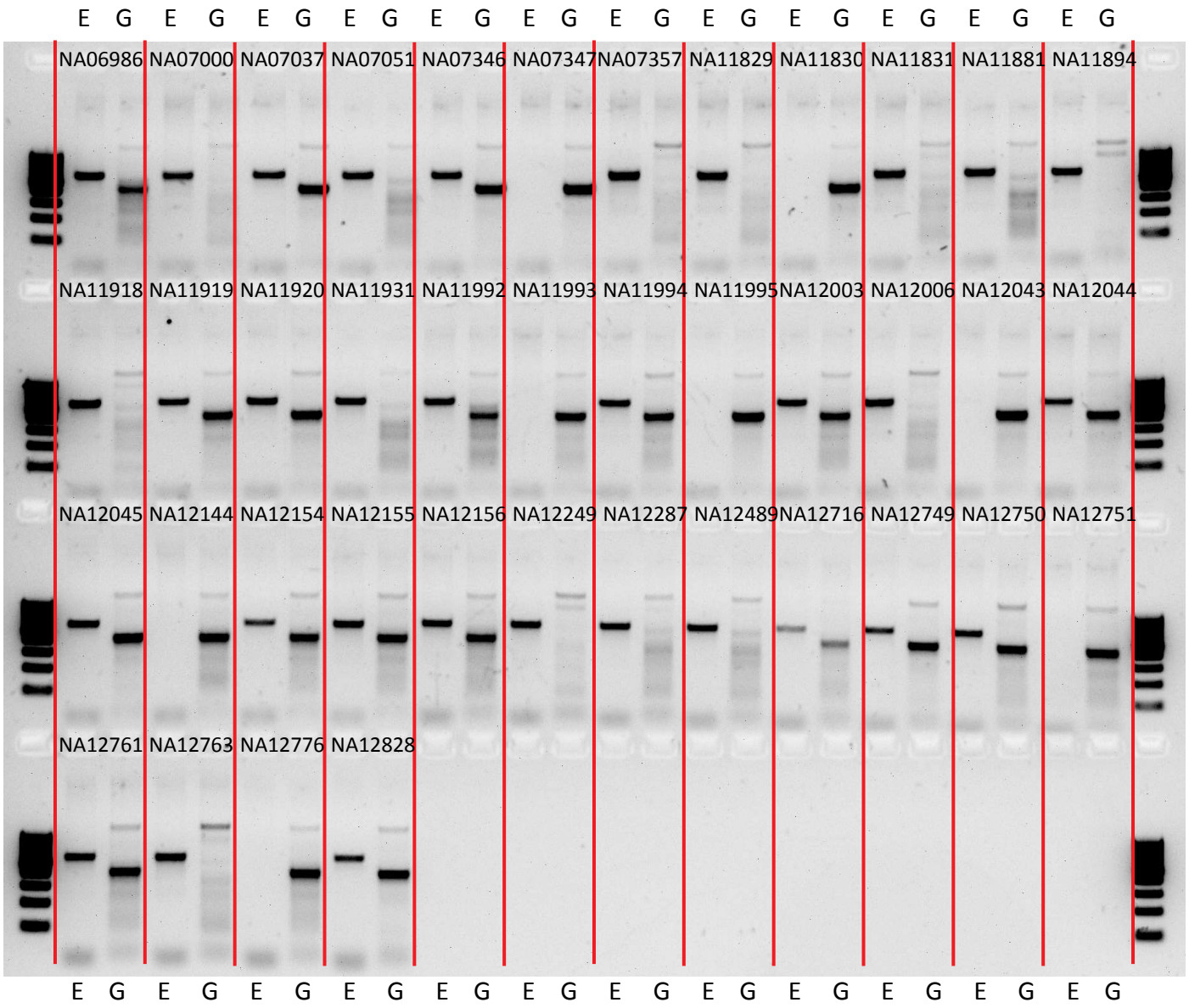
P1_M_061510_10_203



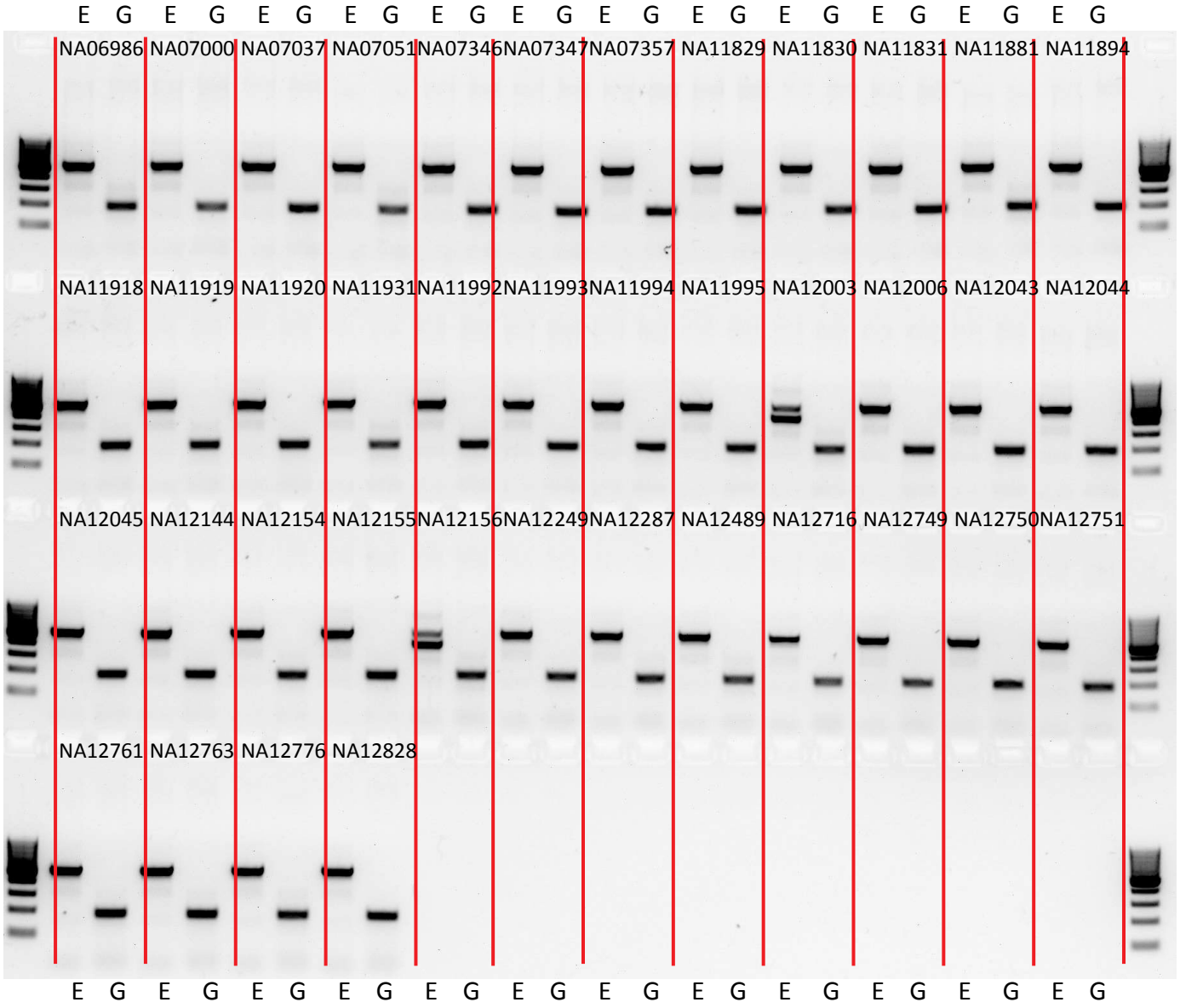
P1_M_061510_12_282



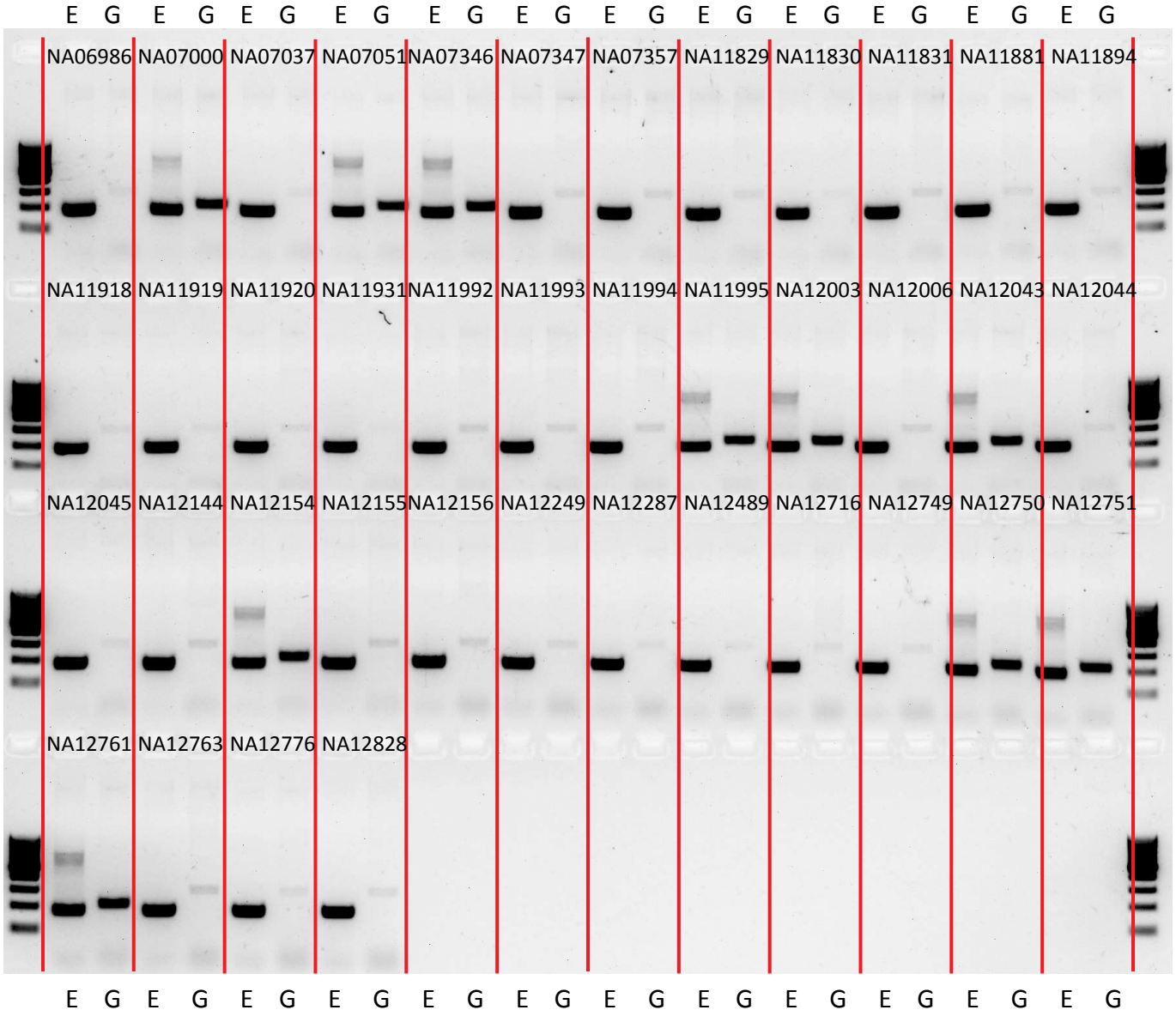
P1_M_061510_13_47



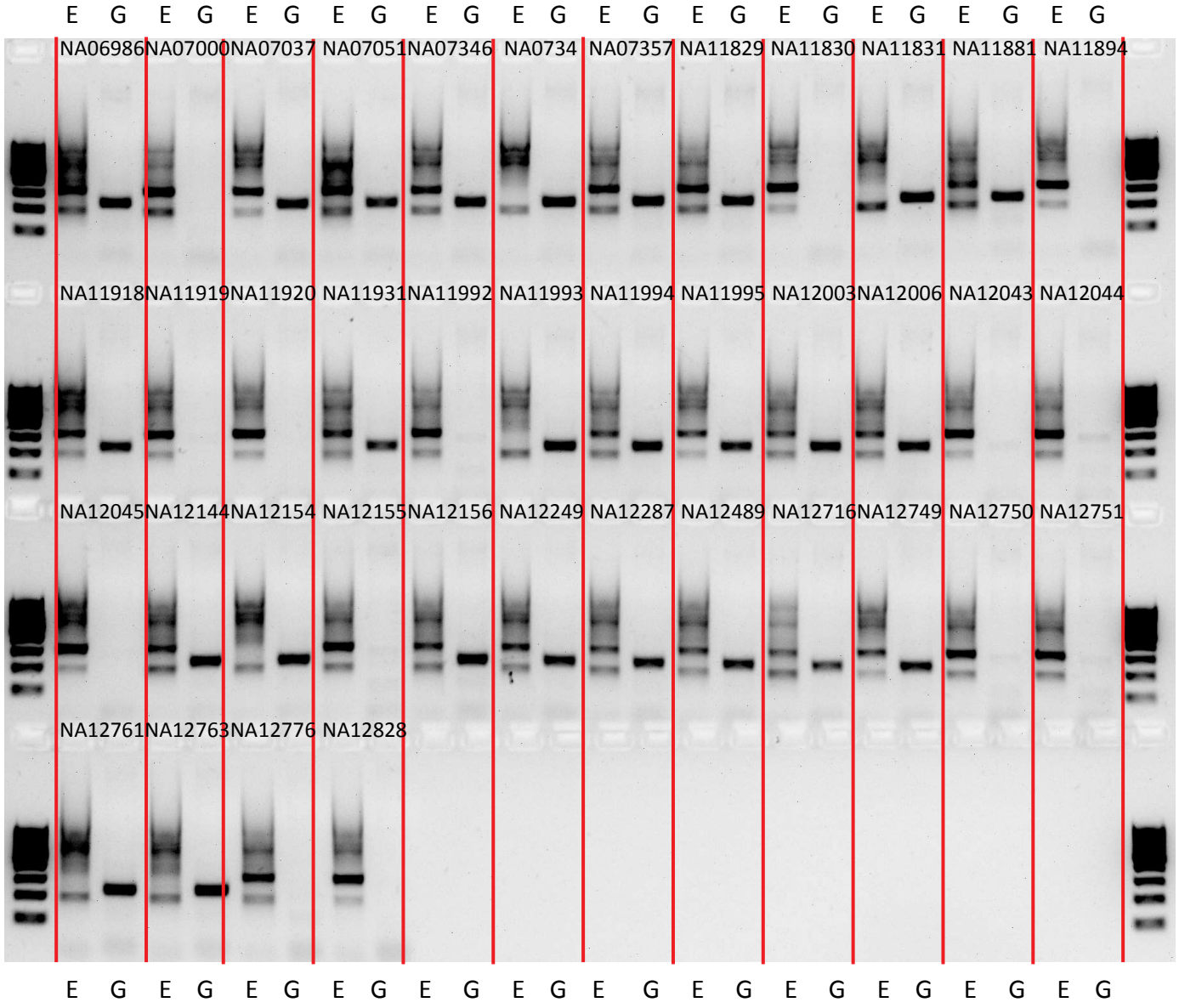
P1_M_061510_14_175



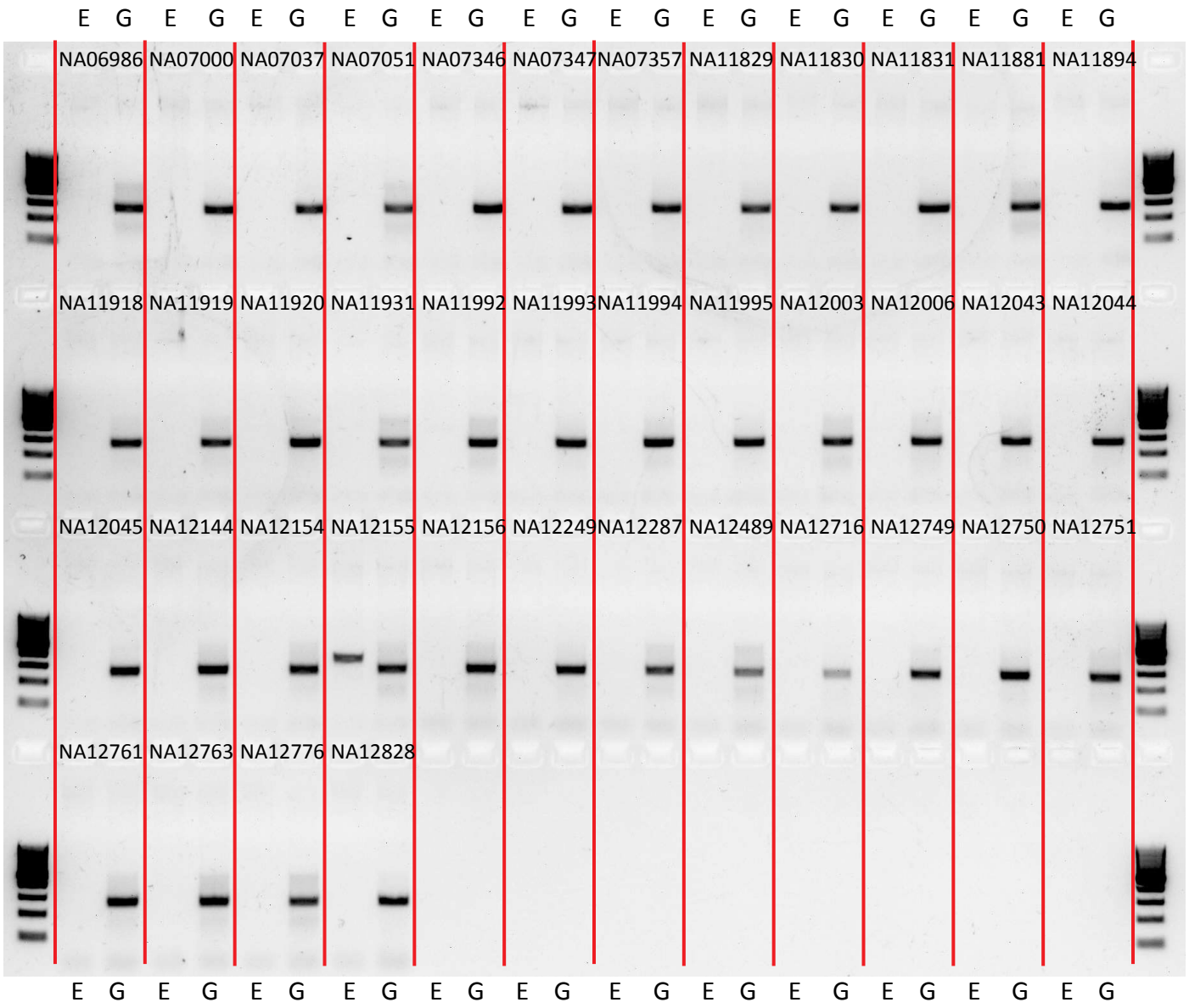
P1_M_061510_15_41



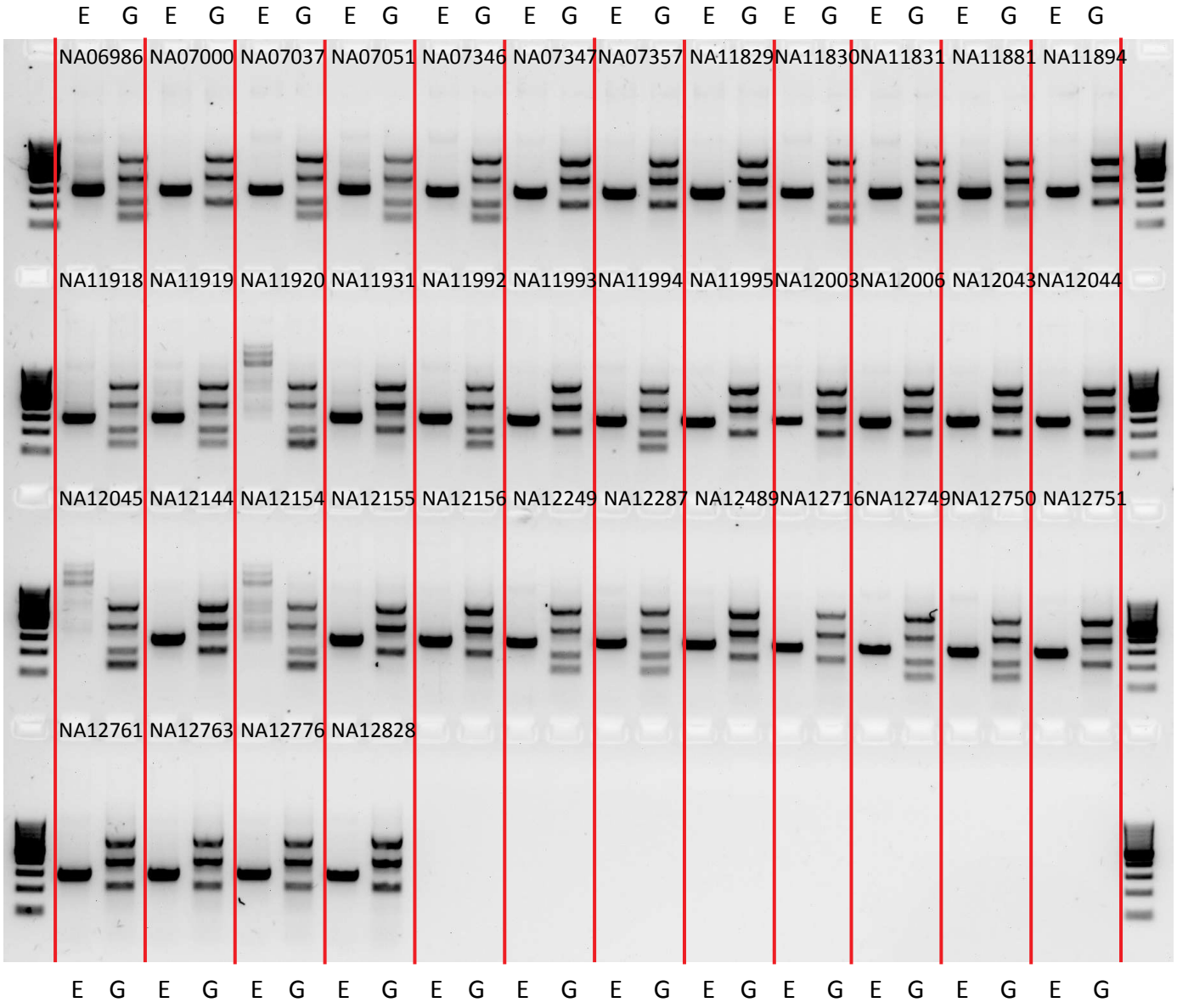
P1_M_061510_16_54



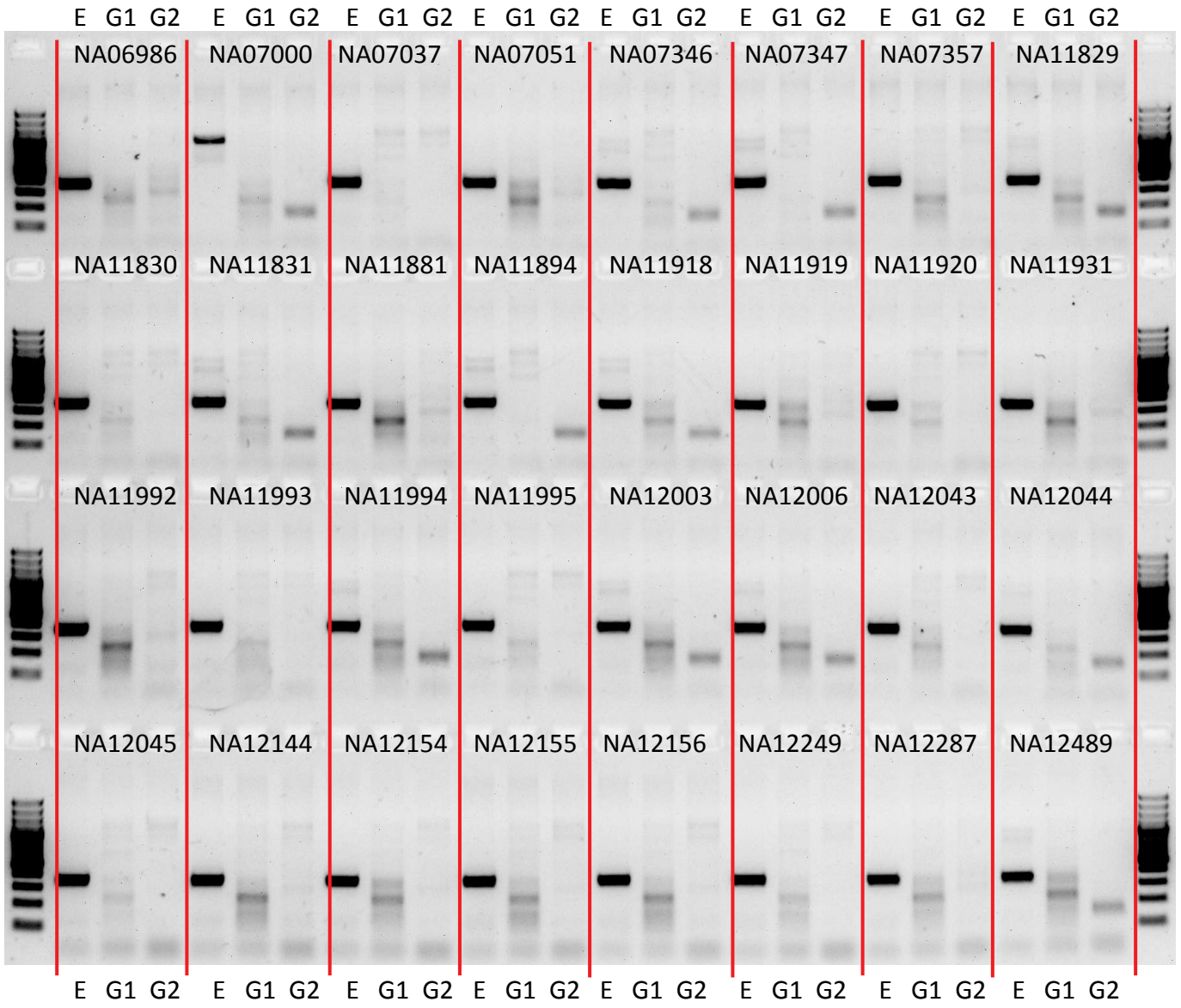
P1_M_061510_18_386



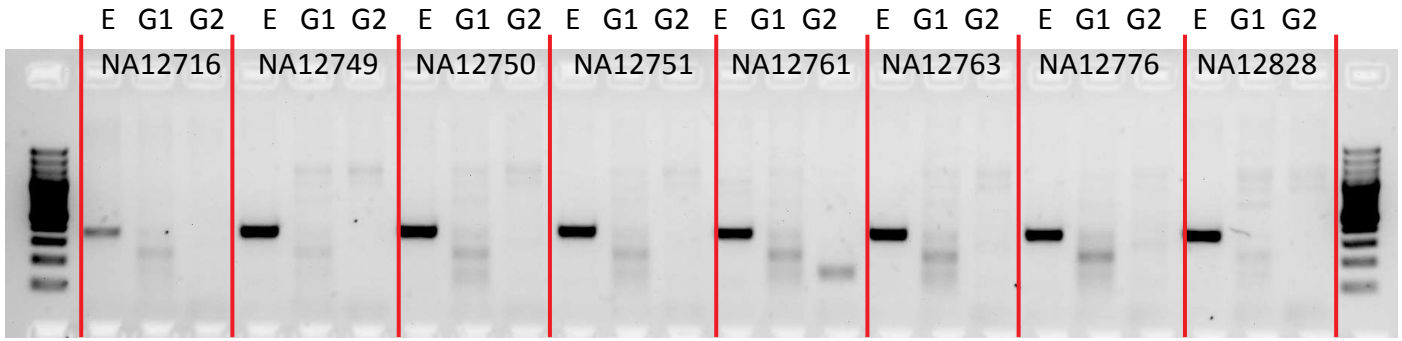
P1_M_061510_20_89



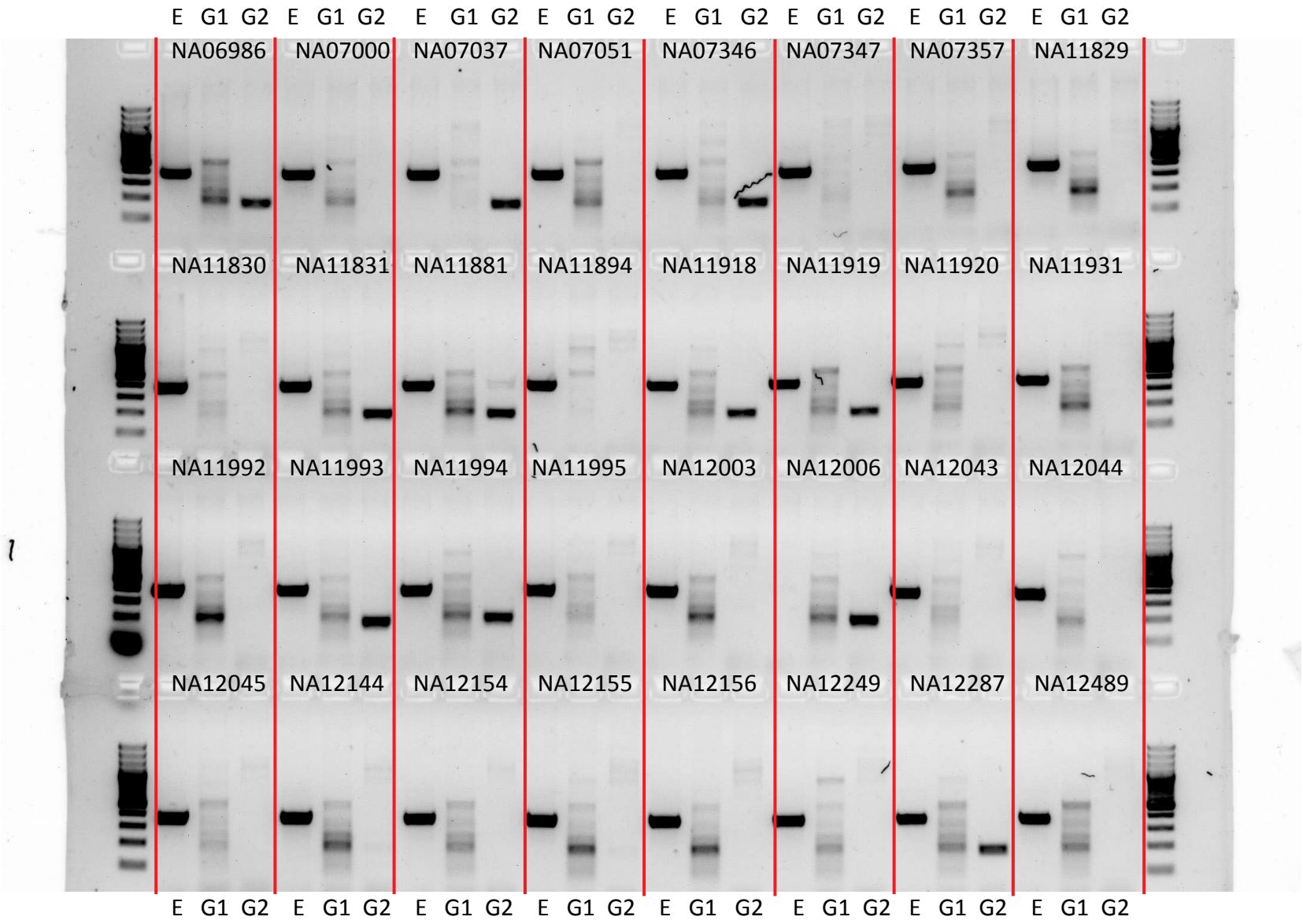
P1_MEI_58&P2_MEI_1302



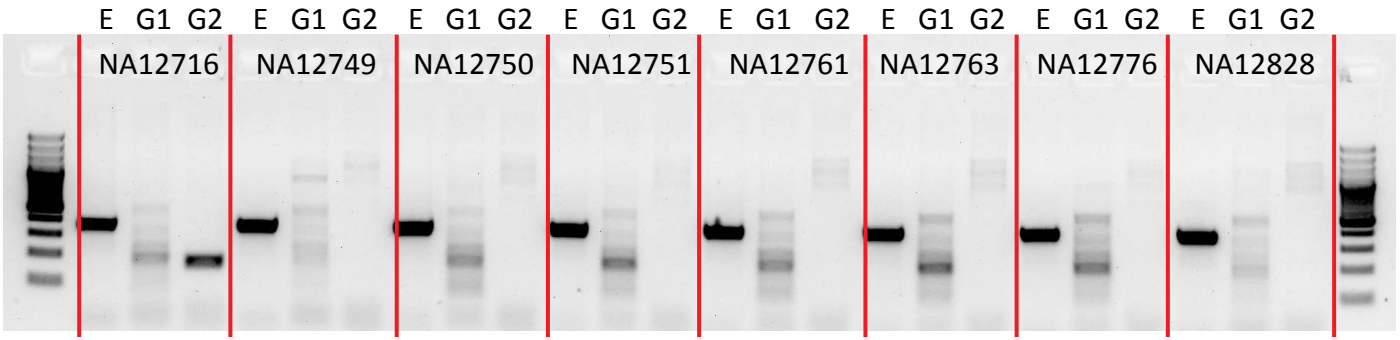
P1_MEI_58&P2_MEI_1302



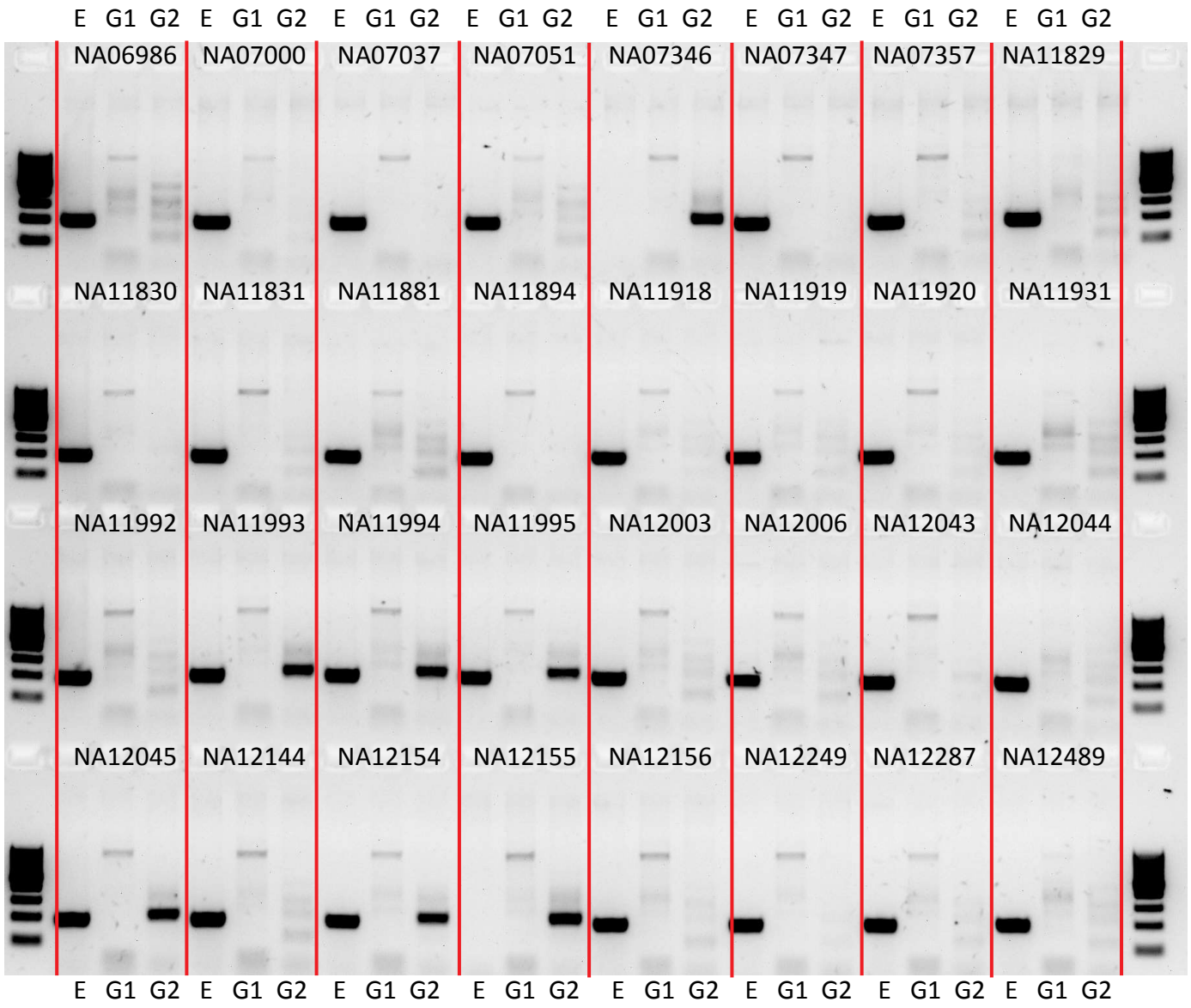
P1_MEI_170


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA06986 NA07000 NA07037 NA07051 NA07346 NA07347 NA07357 NA11829
NA11830 NA11831 NA11881 NA11894 NA11918 NA11919 NA11920 NA11931
NA11992 NA11993 NA11994 NA11995 NA12003 NA12006 NA12043 NA12044
NA12045 NA12144 NA12154 NA12155 NA12156 NA12249 NA12287 NA12489
E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2

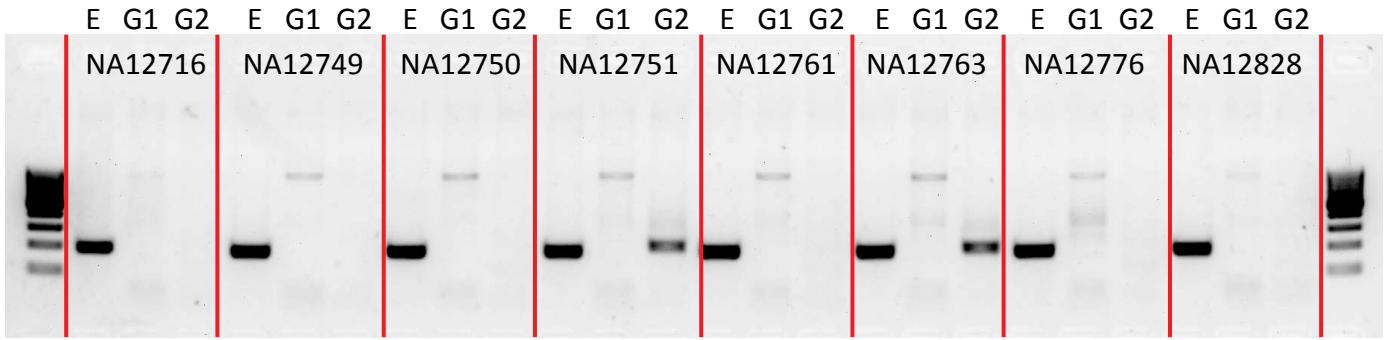
P1_MEI_170


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA12716
NA12749
NA12750
NA12751
NA12761
NA12763
NA12776
NA12828

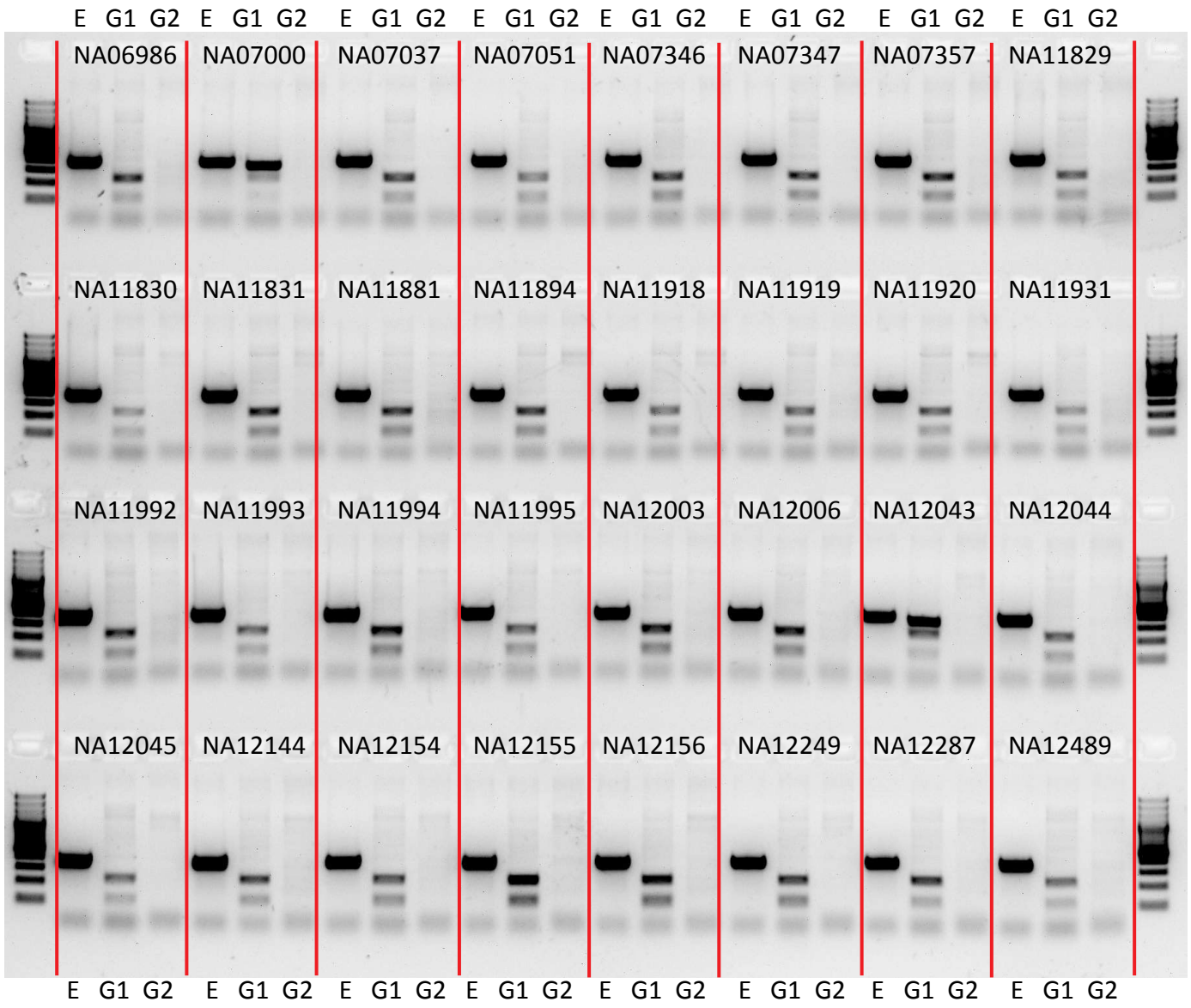
P1_MEI_201


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA06986 NA07000 NA07037 NA07051 NA07346 NA07347 NA07357 NA11829
NA11830 NA11831 NA11881 NA11894 NA11918 NA11919 NA11920 NA11931
NA11992 NA11993 NA11994 NA11995 NA12003 NA12006 NA12043 NA12044
NA12045 NA12144 NA12154 NA12155 NA12156 NA12249 NA12287 NA12489
E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2

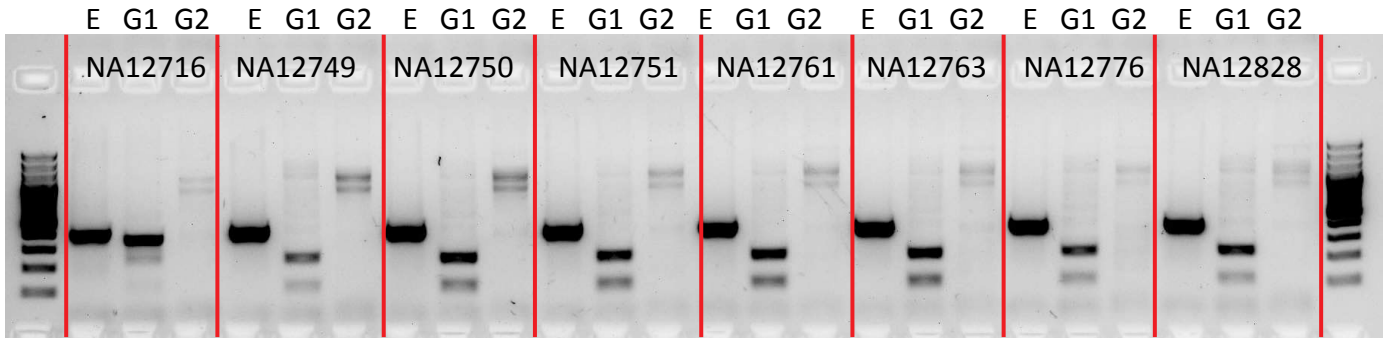
P1_MEI_201


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA12716
NA12749
NA12750
NA12751
NA12761
NA12763
NA12776
NA12828

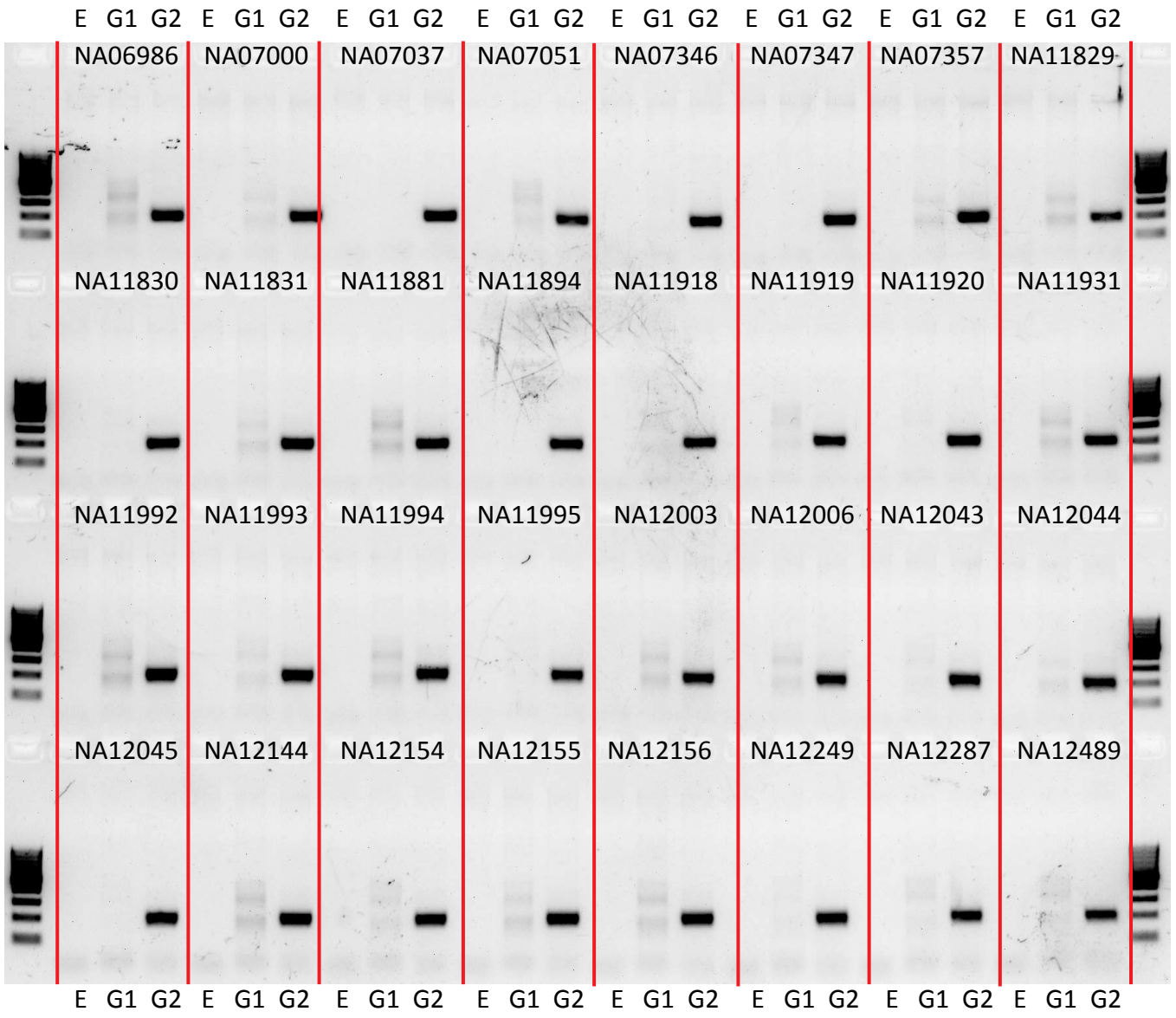
P1_MEI_322


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA06986 NA07000 NA07037 NA07051 NA07346 NA07347 NA07357 NA11829
NA11830 NA11831 NA11881 NA11894 NA11918 NA11919 NA11920 NA11931
NA11992 NA11993 NA11994 NA11995 NA12003 NA12006 NA12043 NA12044
NA12045 NA12144 NA12154 NA12155 NA12156 NA12249 NA12287 NA12489
E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2

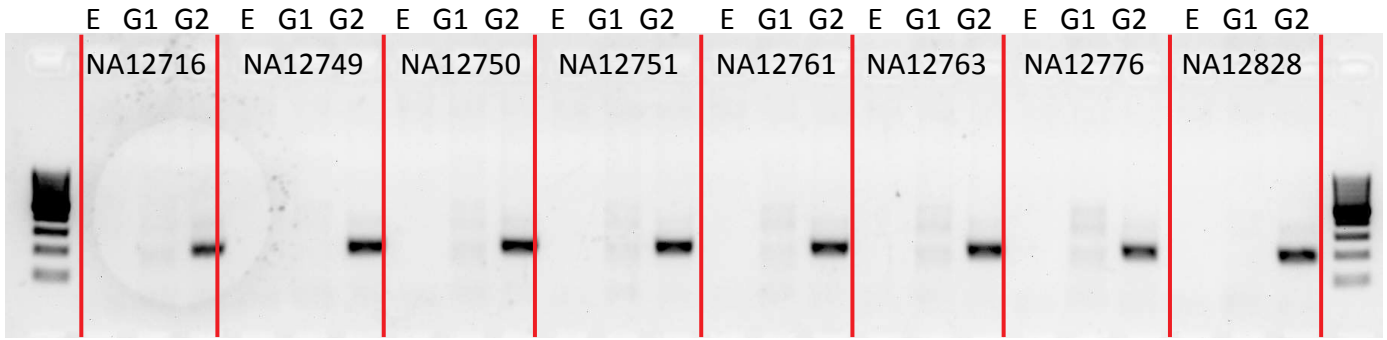
P1_MEI_322


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA12716
NA12749
NA12750
NA12751
NA12761
NA12763
NA12776
NA12828

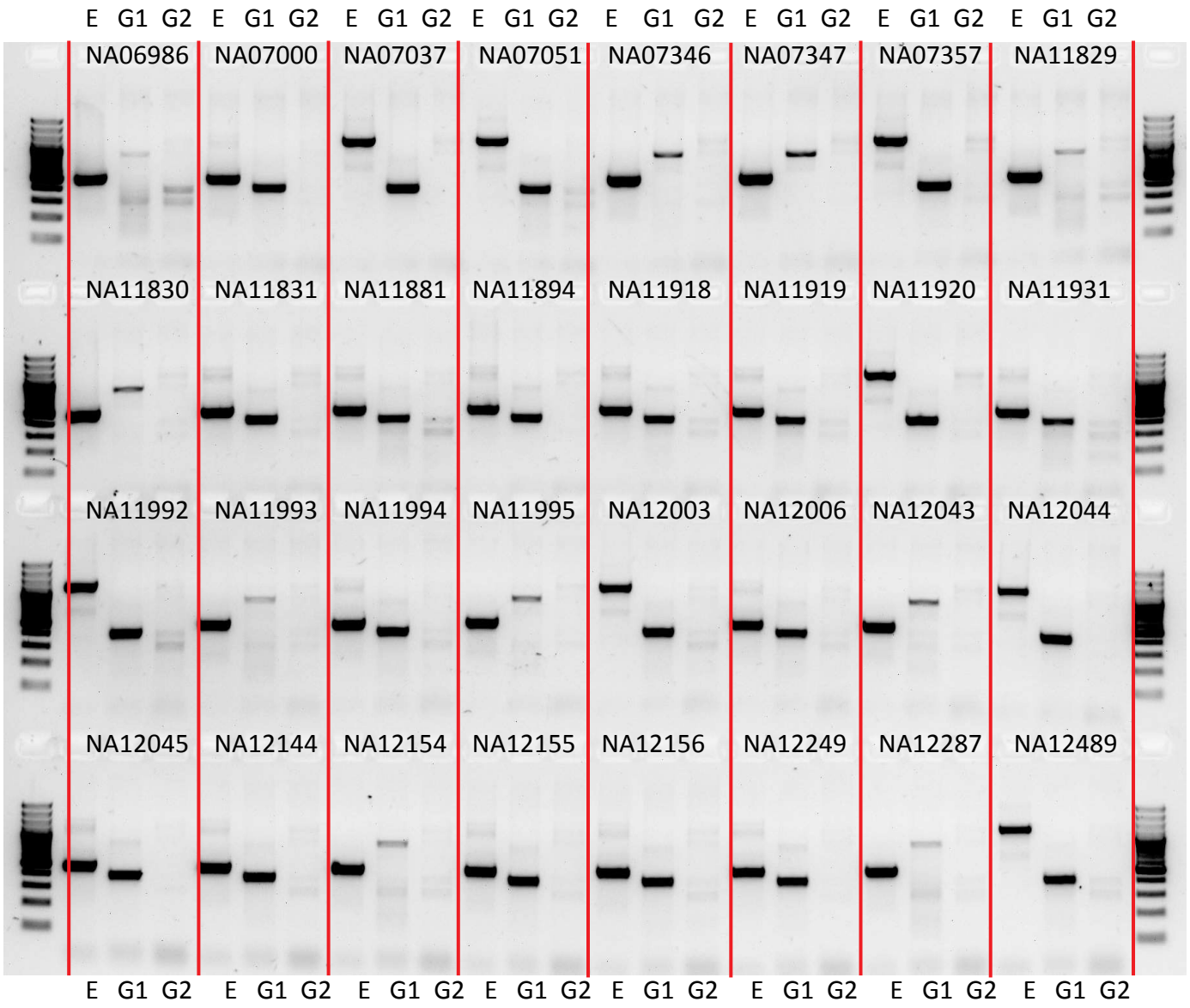
P1_MEI_460&P2_MEI_693



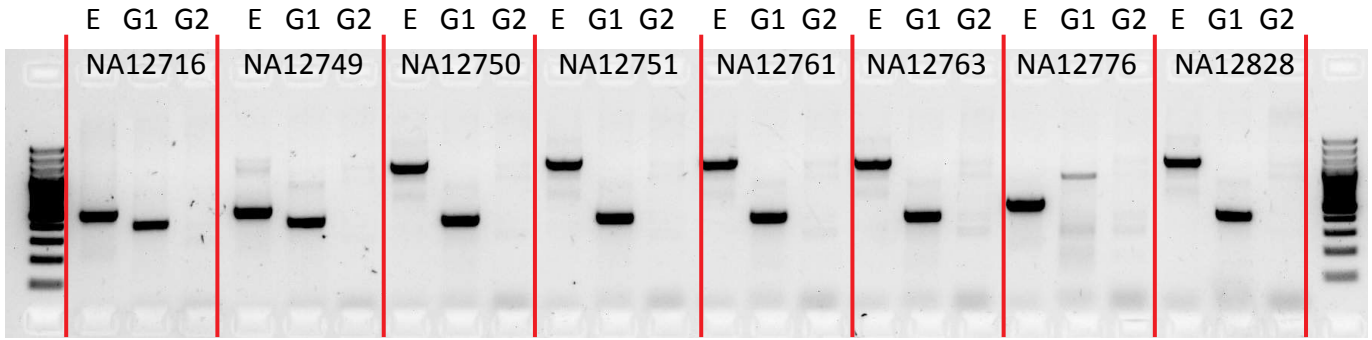
P1_MEI_460&P2_MEI_693



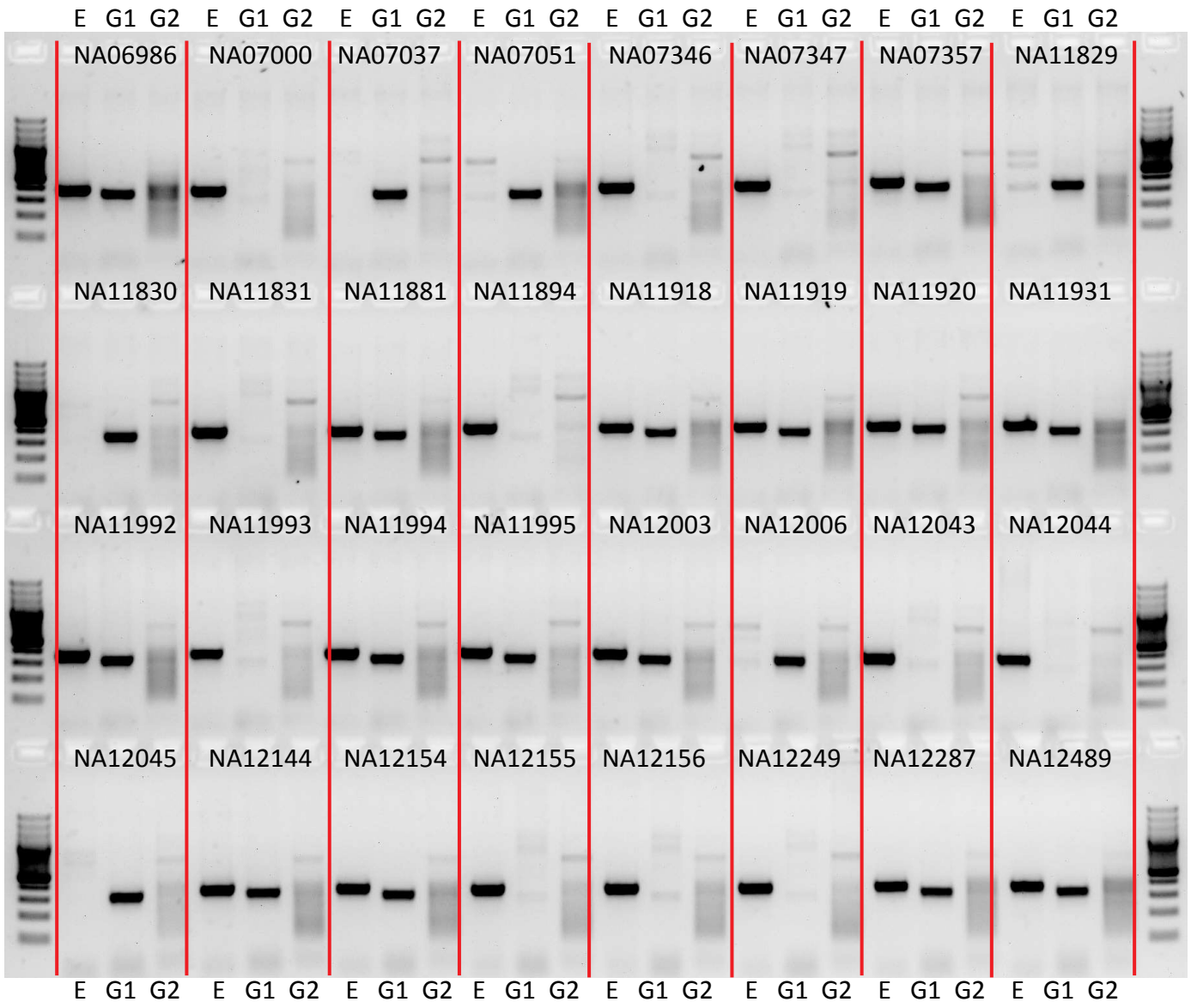
P1_MEI_539&P2_MEI_776



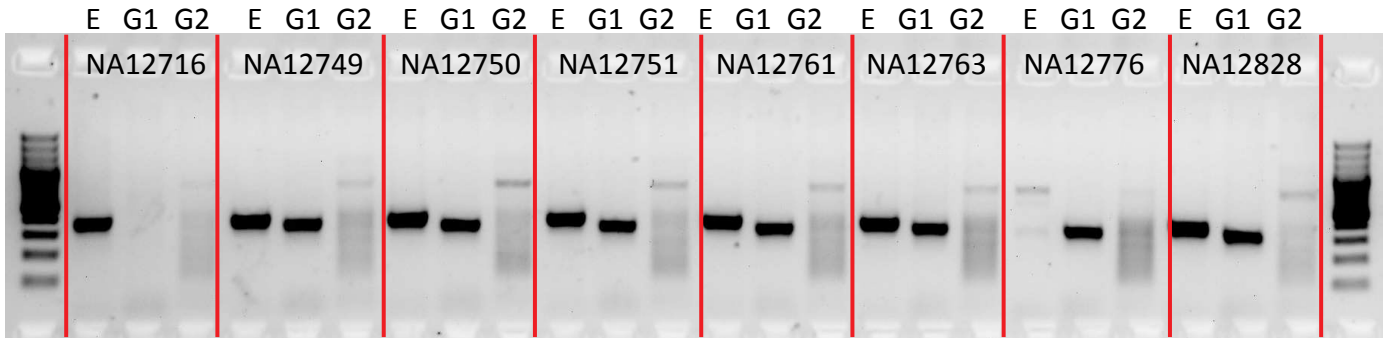
P1_MEI_539&P2_MEI_776



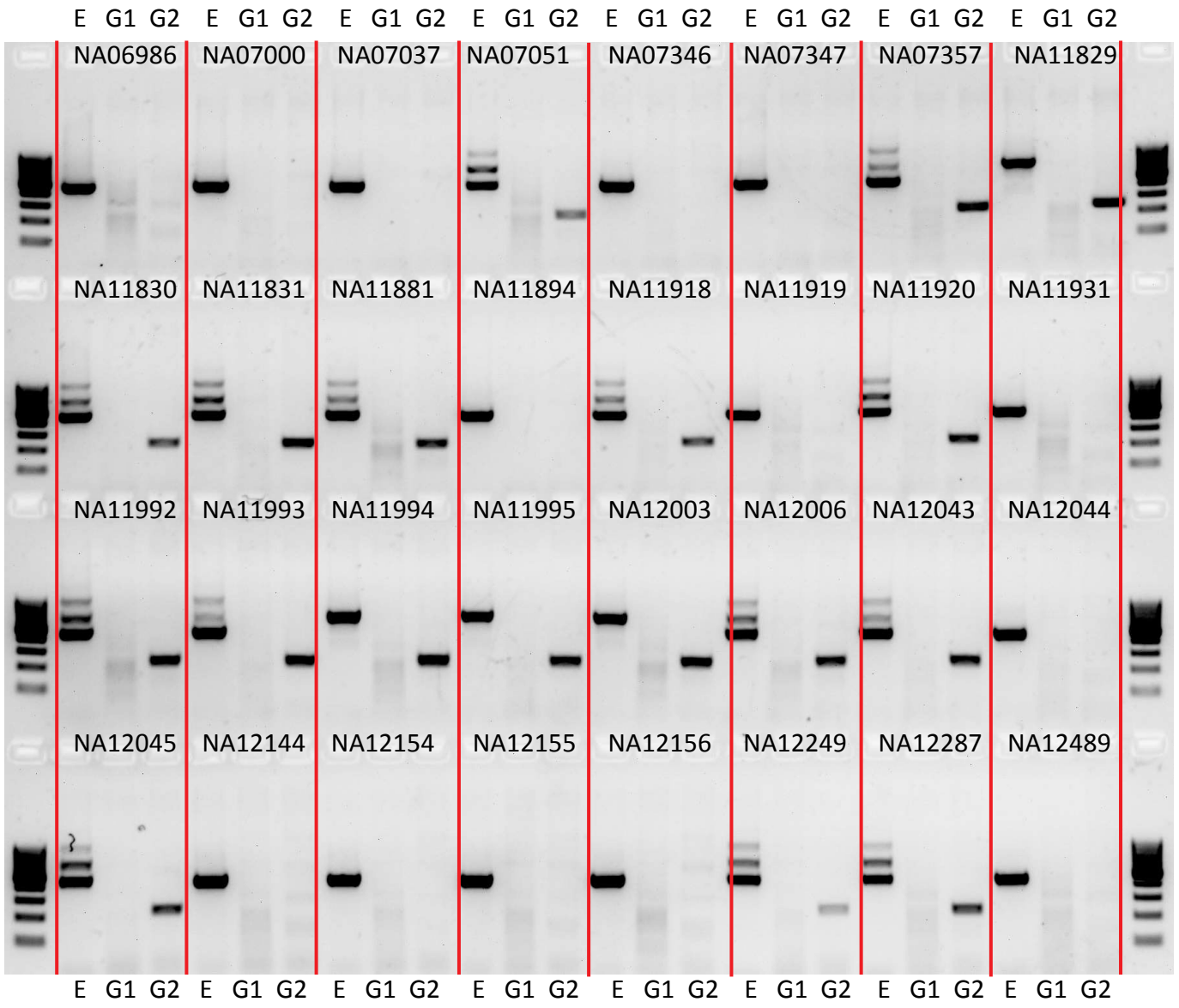
P1_MEI_707&P2_MEI_865



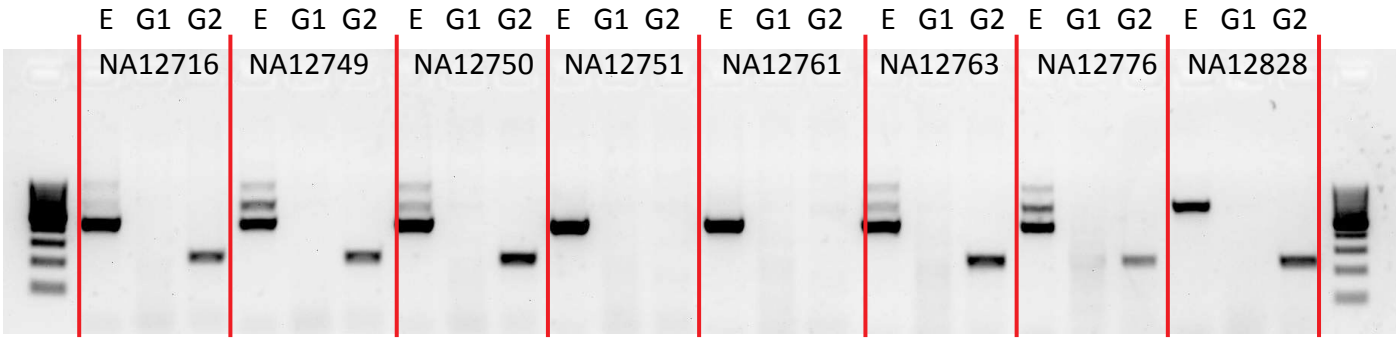
P1_MEI_707&P2_MEI_865



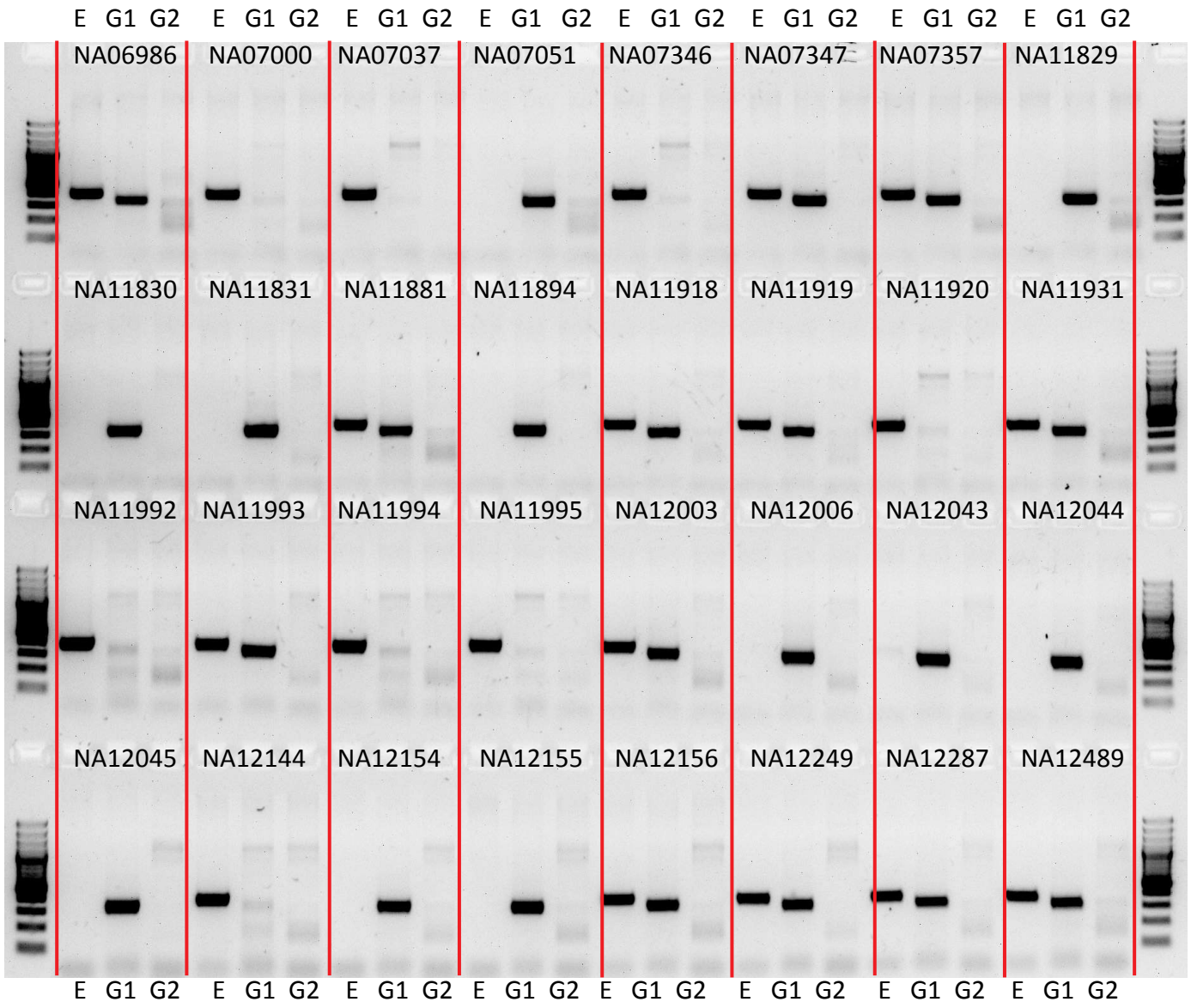
P1_MEI_1023&P2_MEI_147



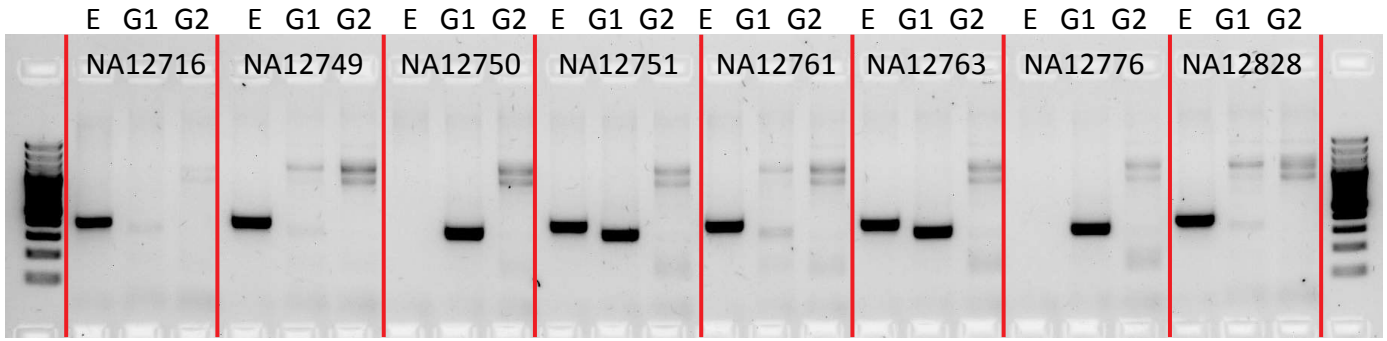
P1_MEI_1023&P2_MEI_147



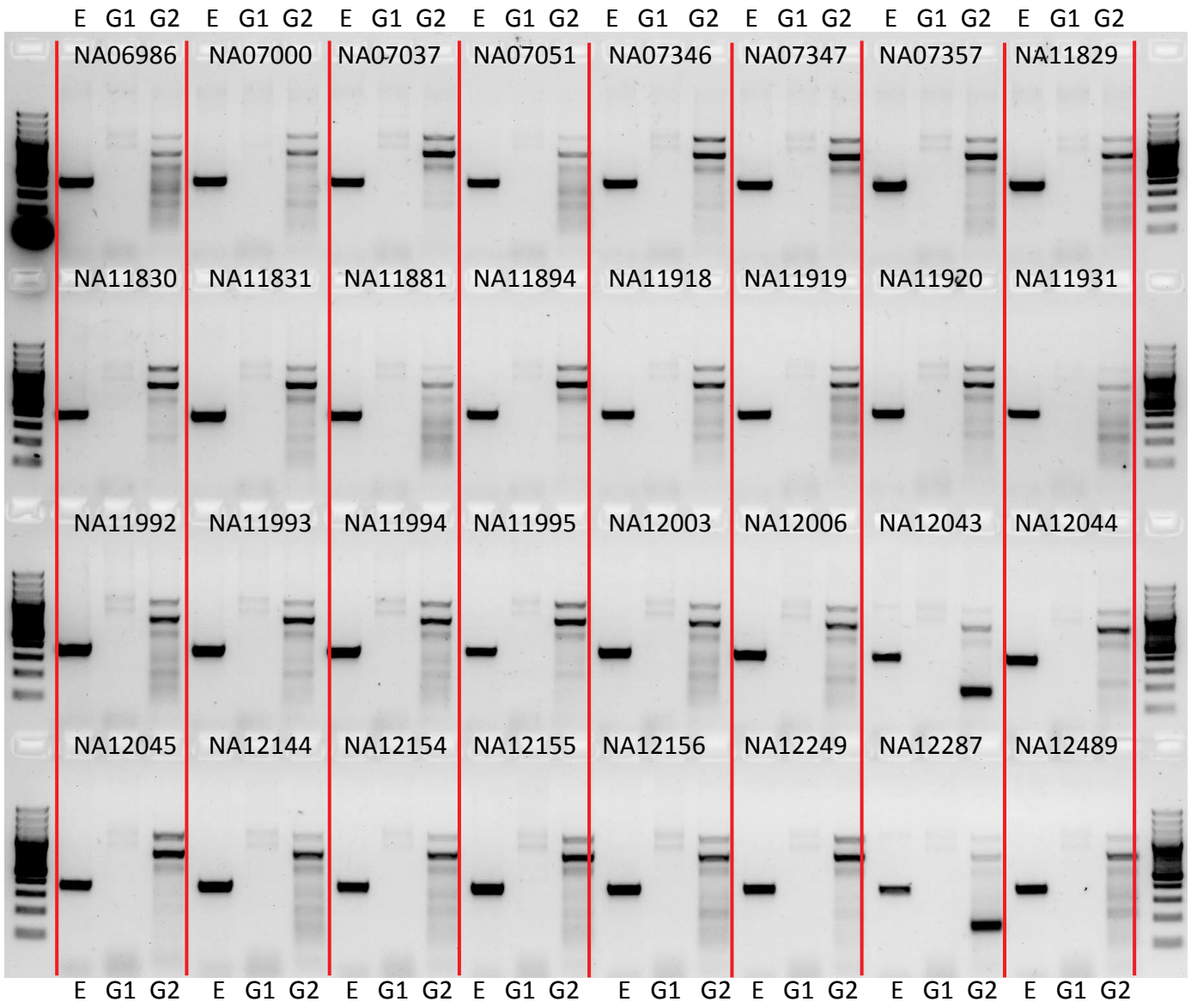
P1_MEI_1051&P2_MEI_170



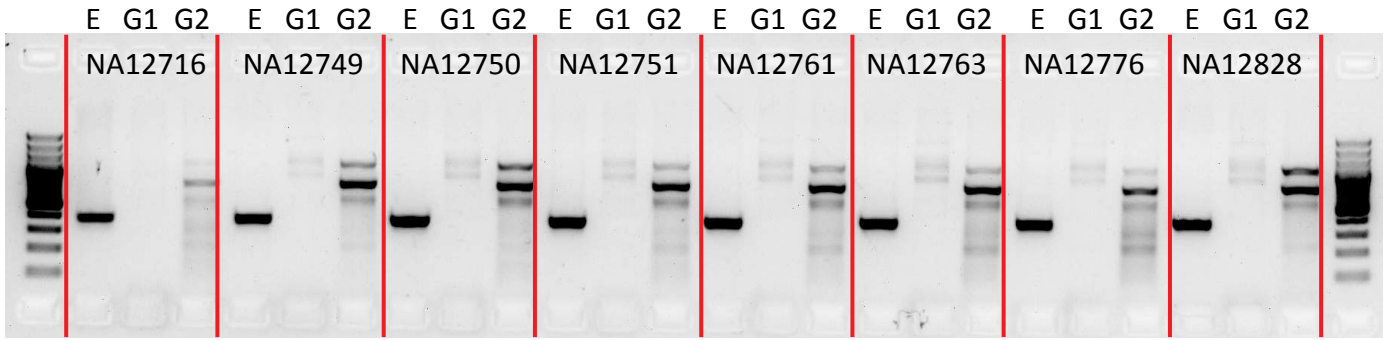
P1_MEI_1051&P2_MEI_170



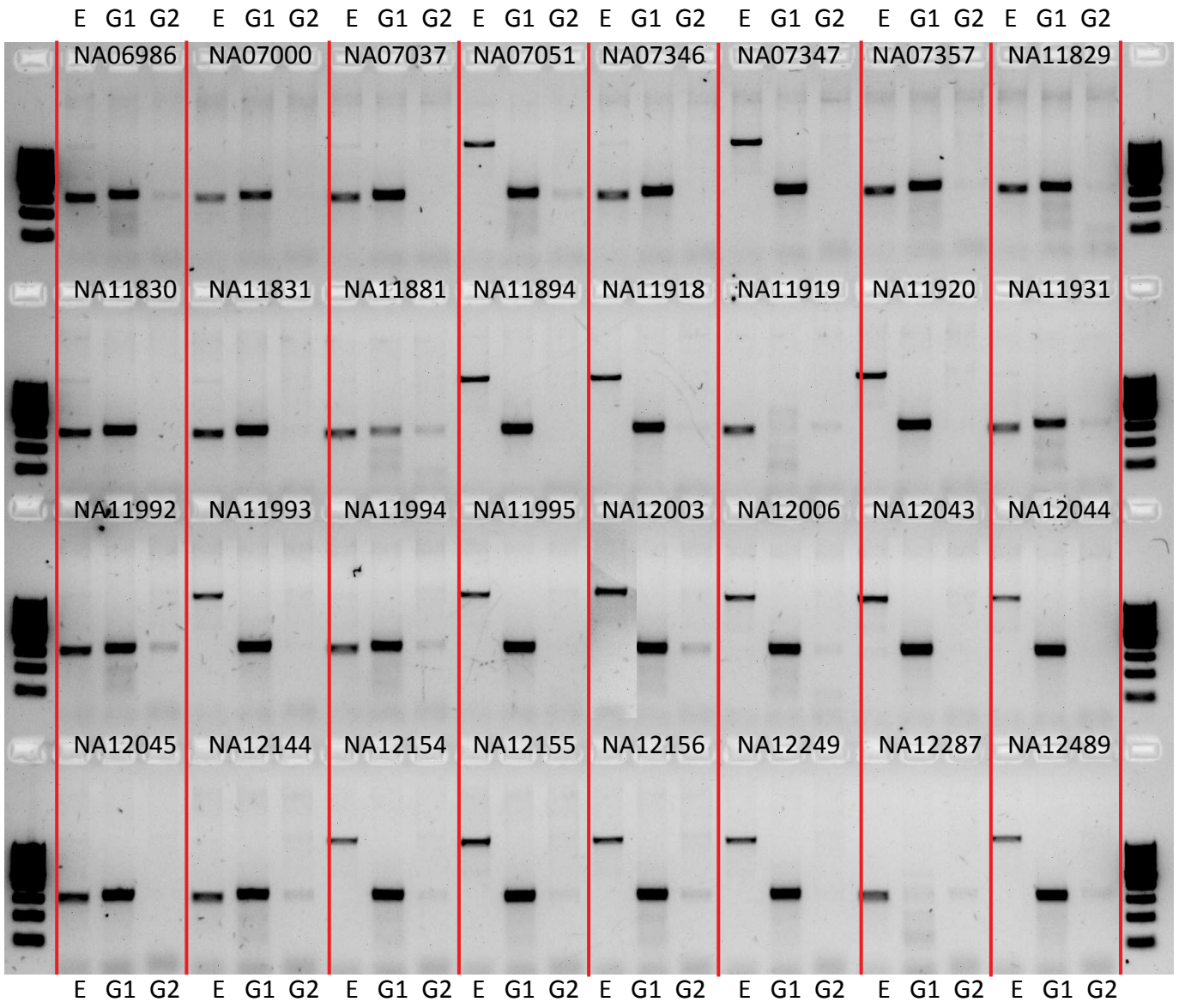
P1_MEI_1062&P2_MEI_1128



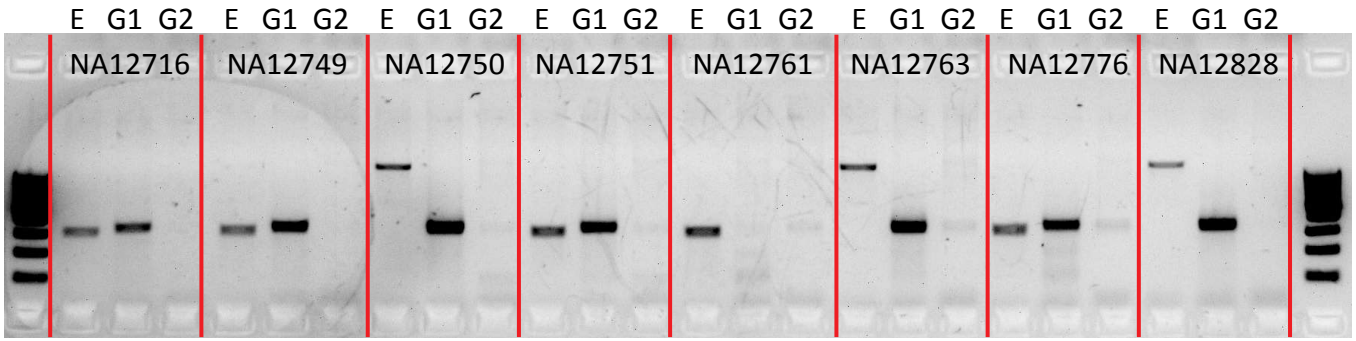
P1_MEI_1062&P2_MEI_1128



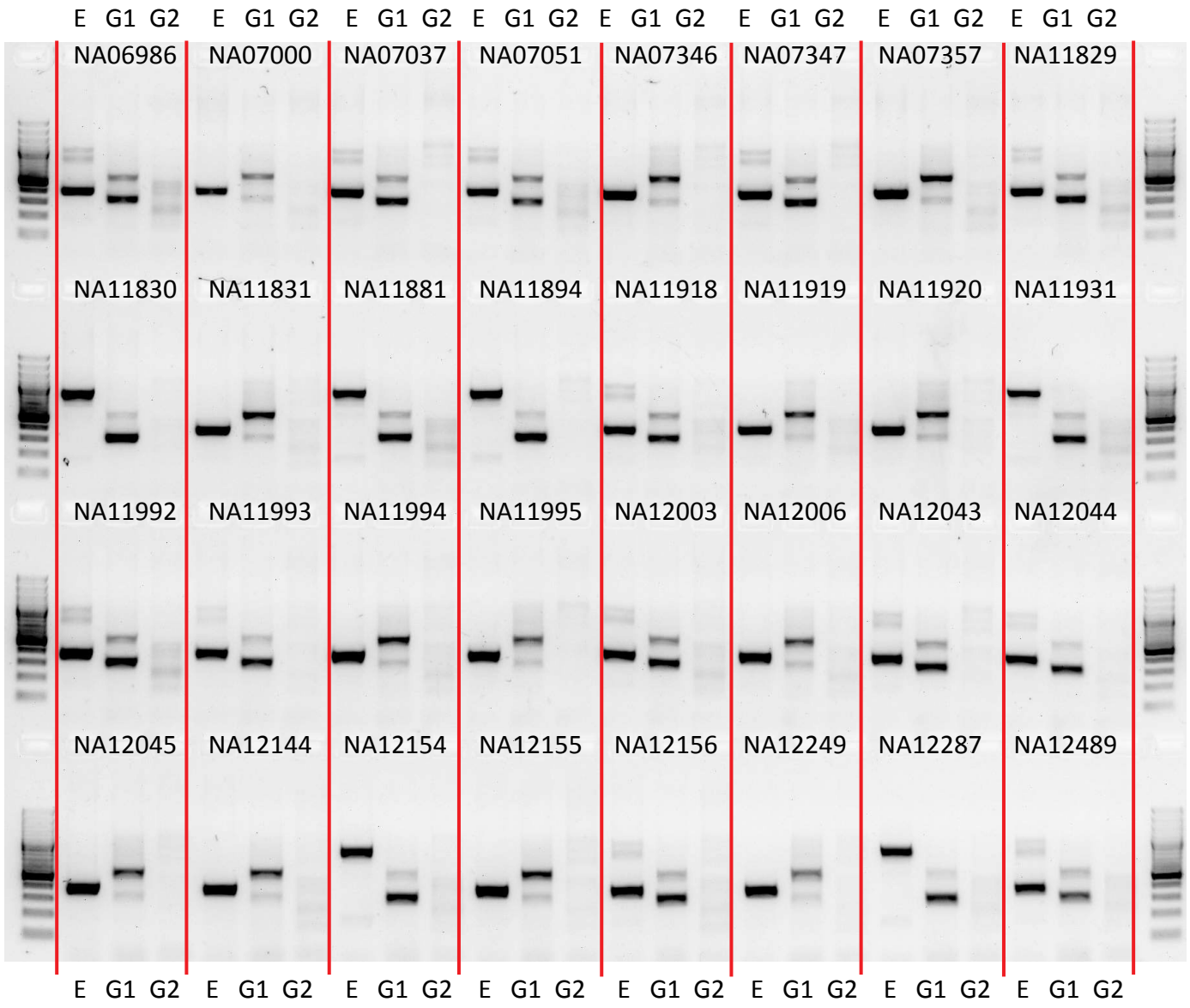
P1_MEI_1067&P2_MEI_1133



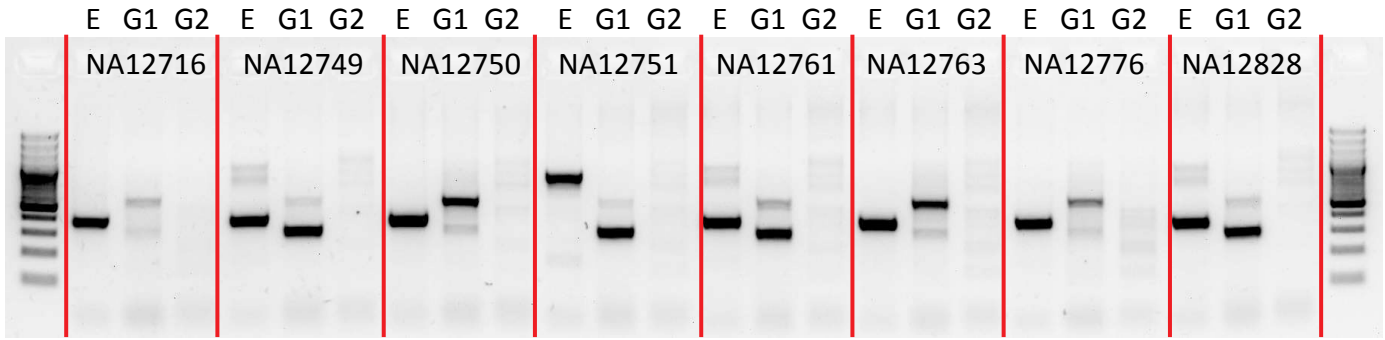
P1_MEI_1067&P2_MEI_1133



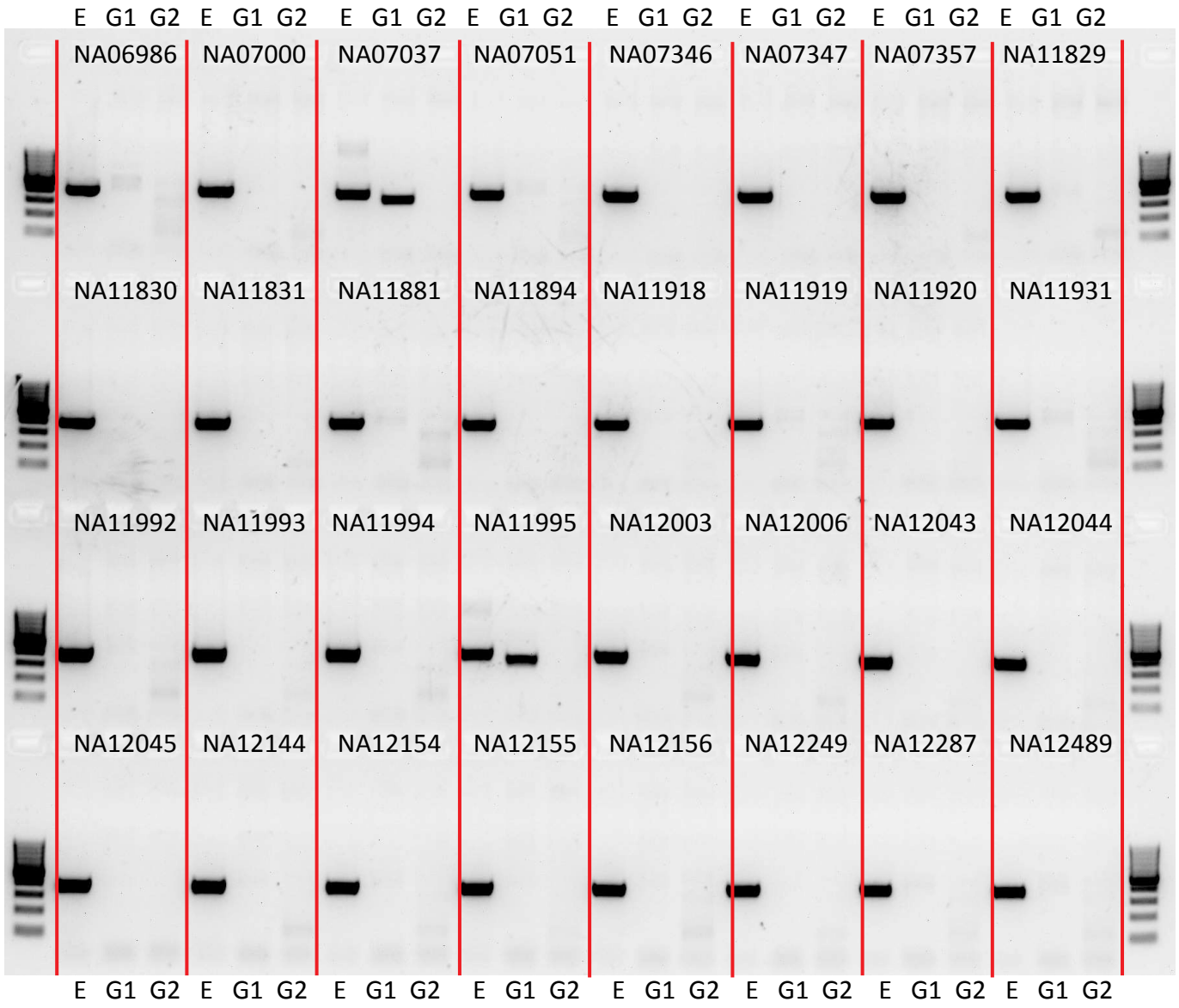
P1_MEI_1280&P2_MEI_1388



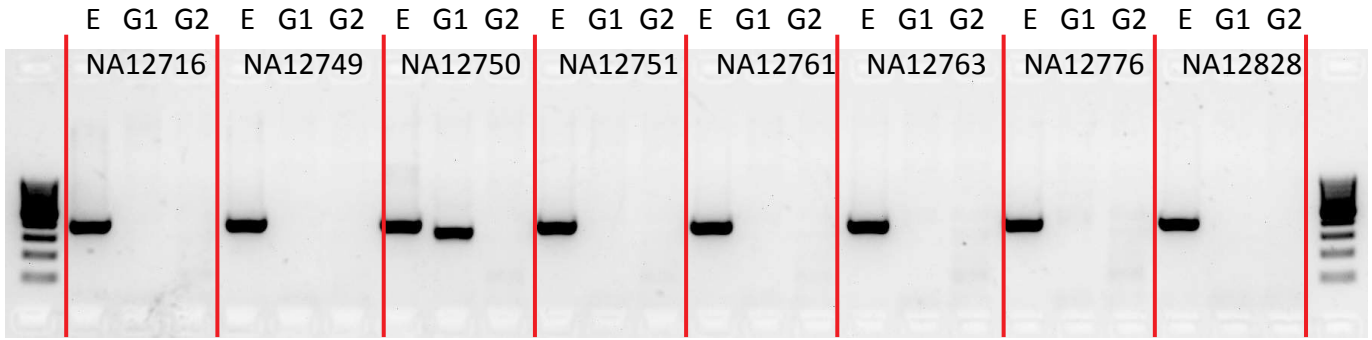
P1_MEI_1280&P2_MEI_1388



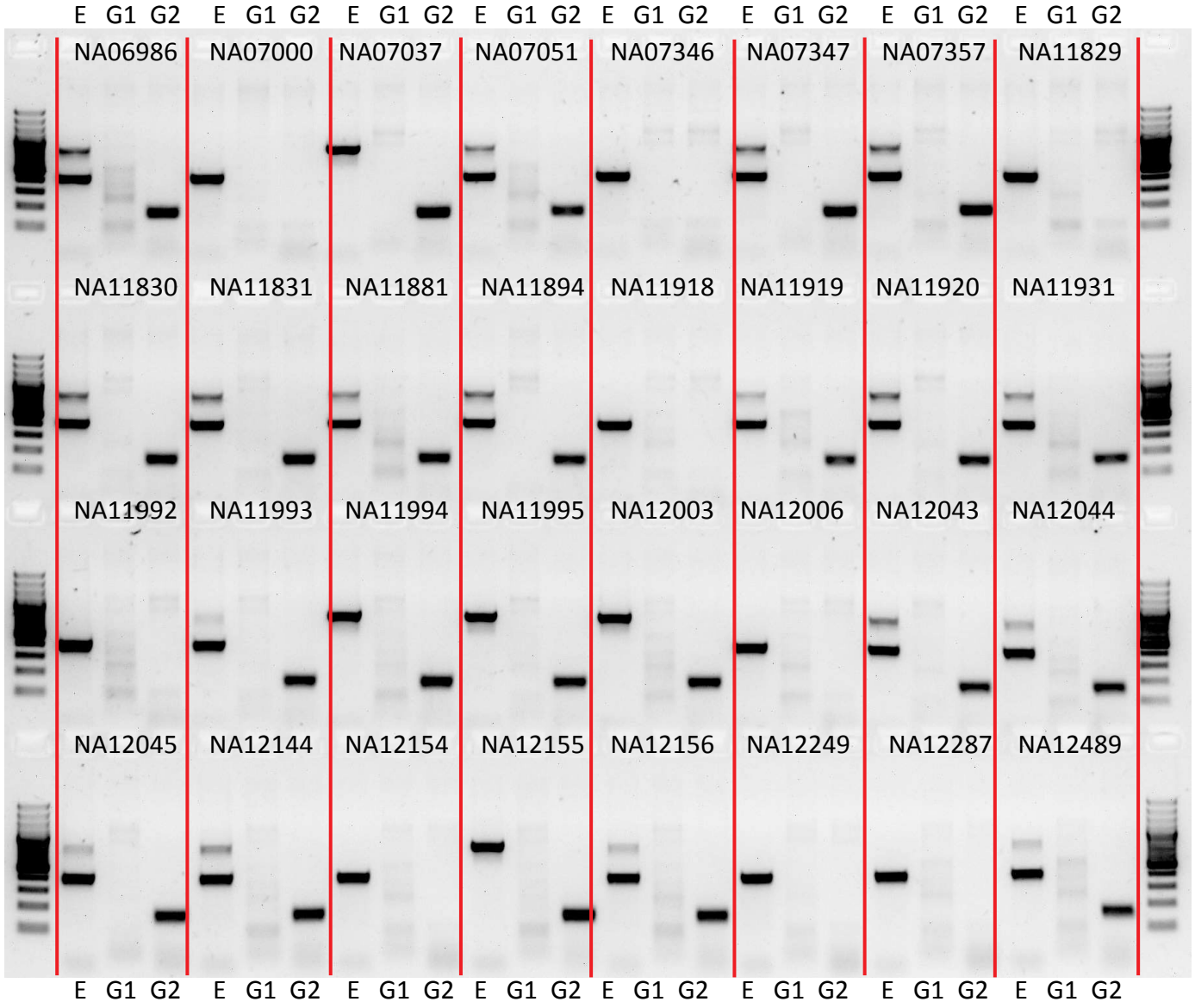
P1_MEI_1569


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA06986 NA07000 NA07037 NA07051 NA07346 NA07347 NA07357 NA11829
NA11830 NA11831 NA11881 NA11894 NA11918 NA11919 NA11920 NA11931
NA11992 NA11993 NA11994 NA11995 NA12003 NA12006 NA12043 NA12044
NA12045 NA12144 NA12154 NA12155 NA12156 NA12249 NA12287 NA12489
E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2

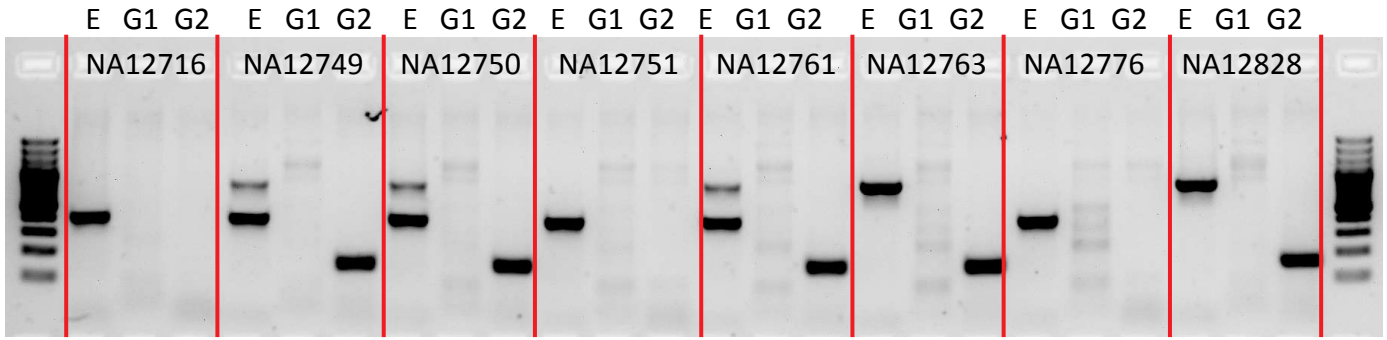
P1_MEI_1569


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA12716
NA12749
NA12750
NA12751
NA12761
NA12763
NA12776
NA12828

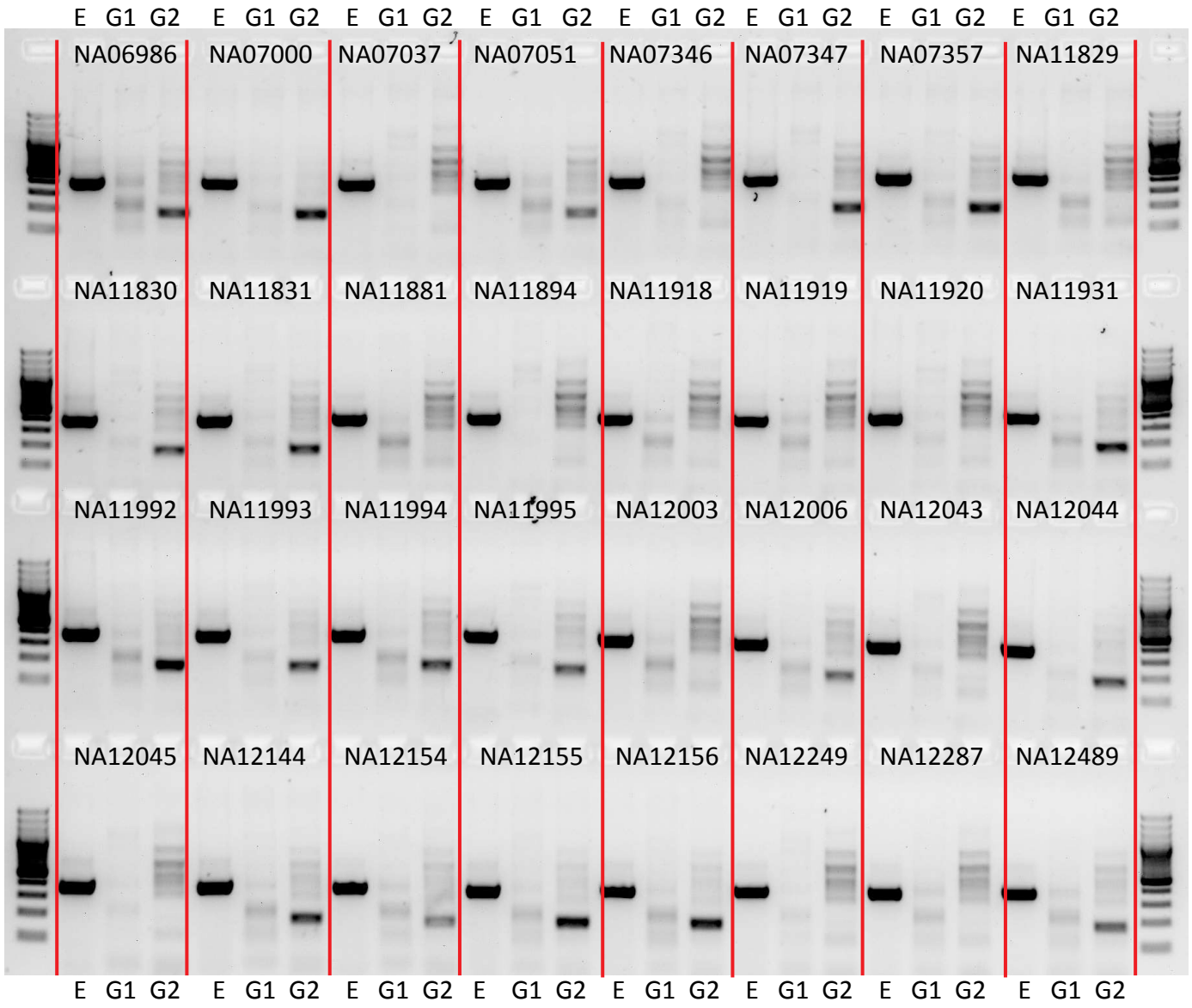
P1_MEI_1894&P2_MEI_1613



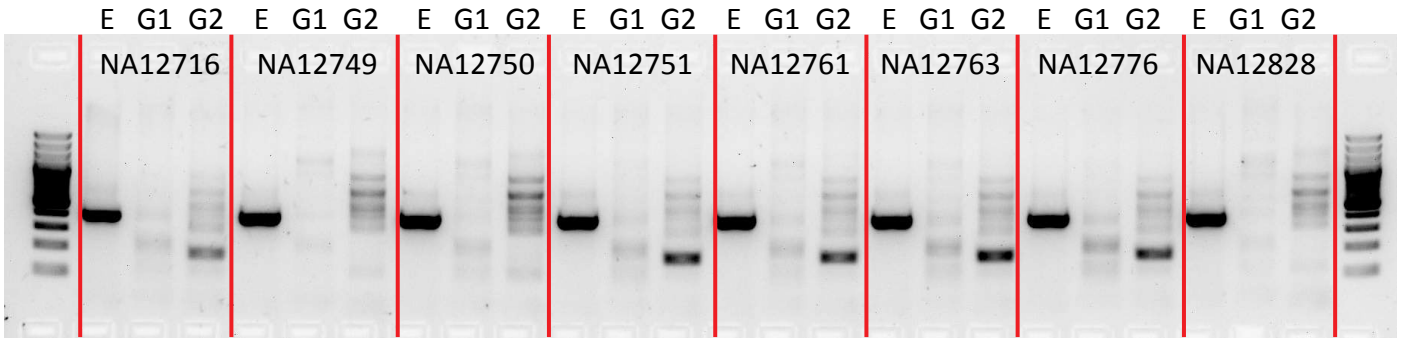
P1_MEI_1894&P2_MEI_1613



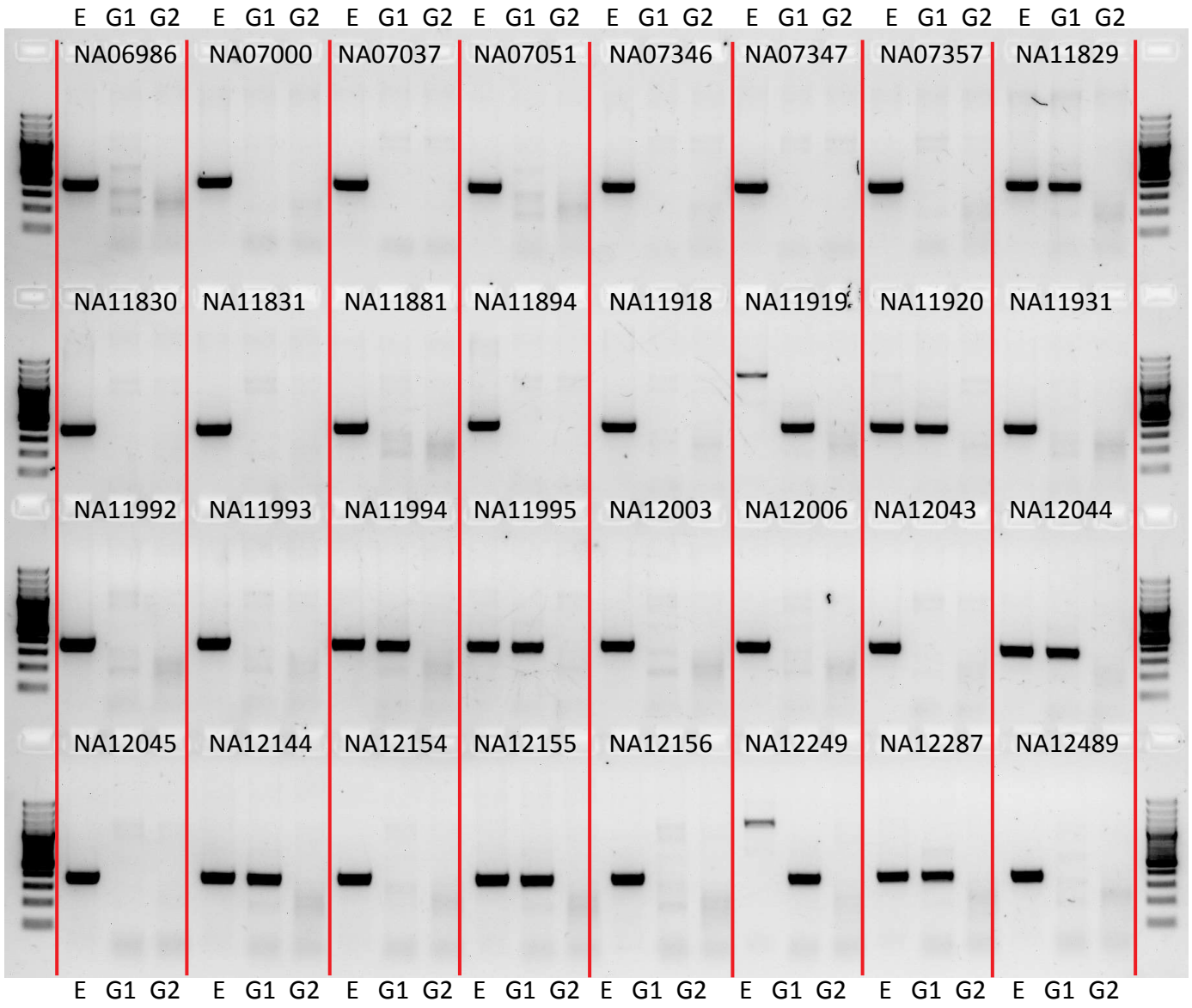
P1_MEI_1949&P2_MEI_1636



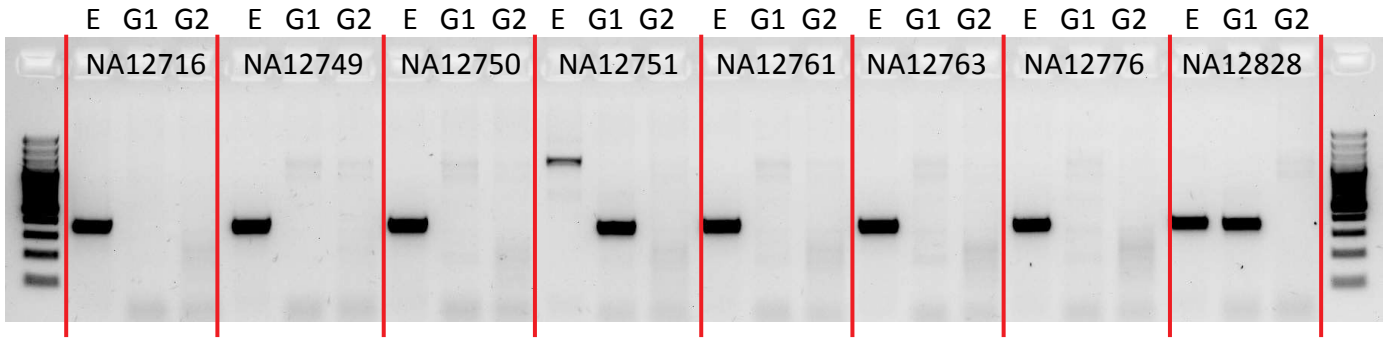
P1_MEI_1949&P2_MEI_1636



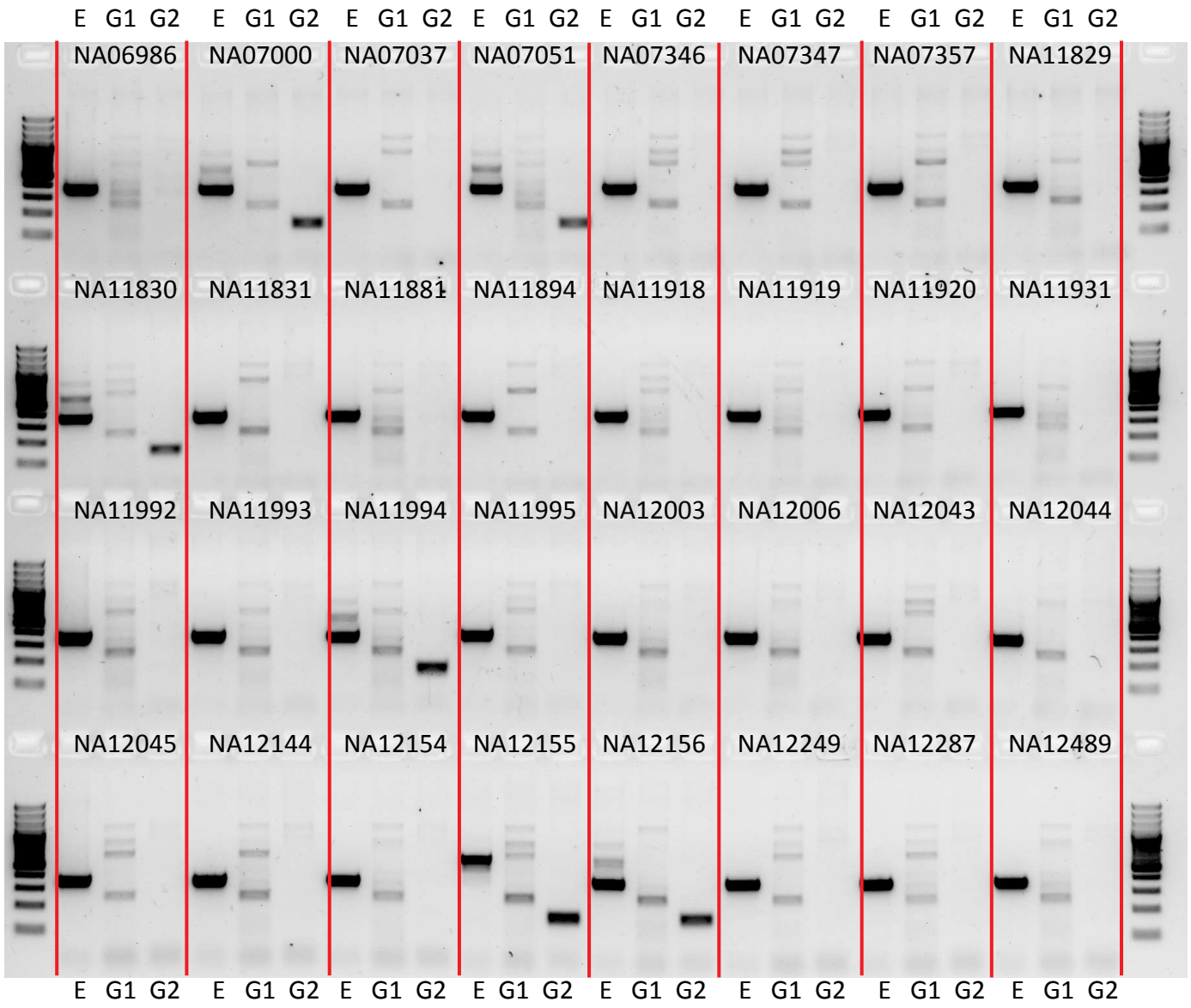
P1_MEI_2433


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA06986 NA07000 NA07037 NA07051 NA07346 NA07347 NA07357 NA11829
NA11830 NA11831 NA11881 NA11894 NA11918 NA11919 NA11920 NA11931
NA11992 NA11993 NA11994 NA11995 NA12003 NA12006 NA12043 NA12044
NA12045 NA12144 NA12154 NA12155 NA12156 NA12249 NA12287 NA12489
E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2

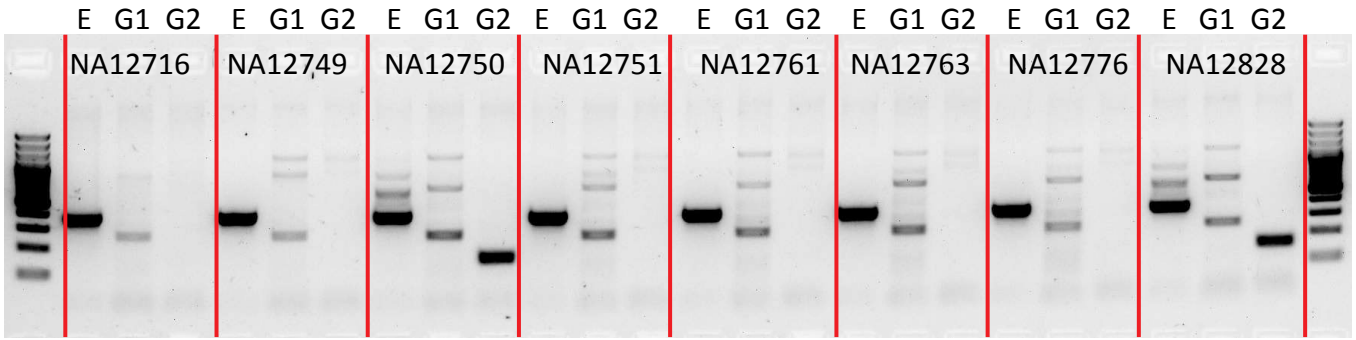
P1_MEI_2433


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA12716
NA12749
NA12750
NA12751
NA12761
NA12763
NA12776
NA12828

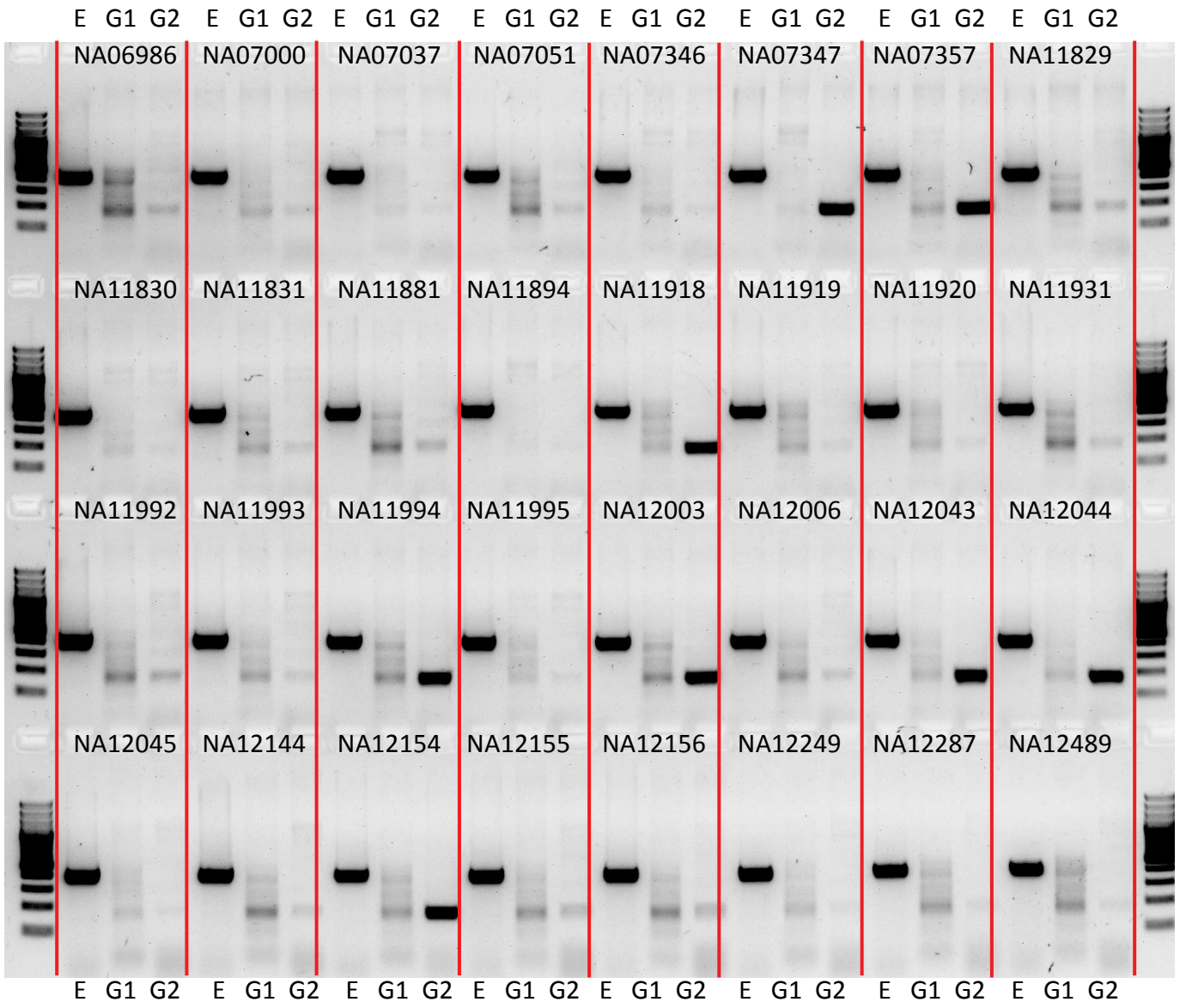
P1_MEI_2730


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA06986 NA07000 NA07037 NA07051 NA07346 NA07347 NA07357 NA11829
NA11830 NA11831 NA11881 NA11894 NA11918 NA11919 NA11920 NA11931
NA11992 NA11993 NA11994 NA11995 NA12003 NA12006 NA12043 NA12044
NA12045 NA12144 NA12154 NA12155 NA12156 NA12249 NA12287 NA12489
E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2

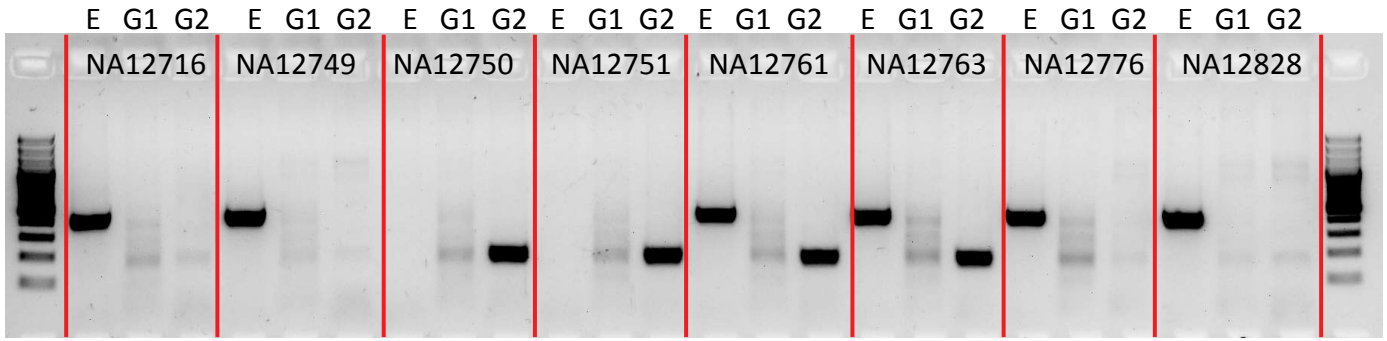
P1_MEI_2730


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA12716
NA12749
NA12750
NA12751
NA12761
NA12763
NA12776
NA12828

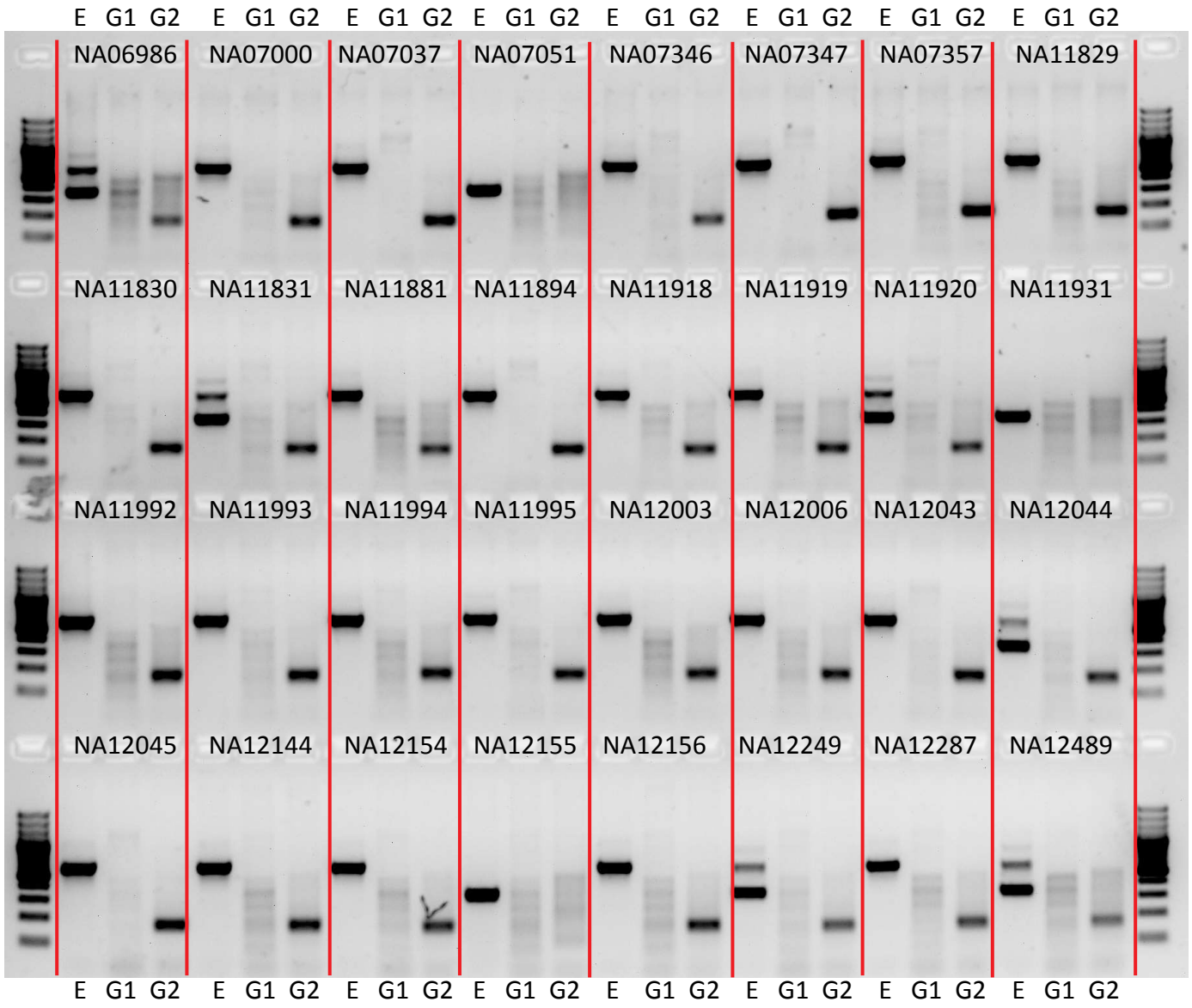
P1_MEI_2893&P2_MEI_1141



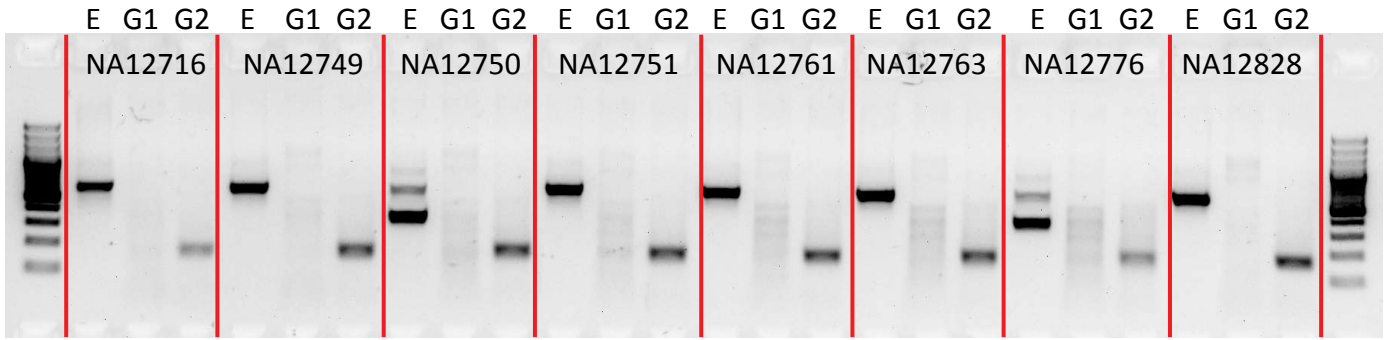
P1_MEI_2893&P2_MEI_1141



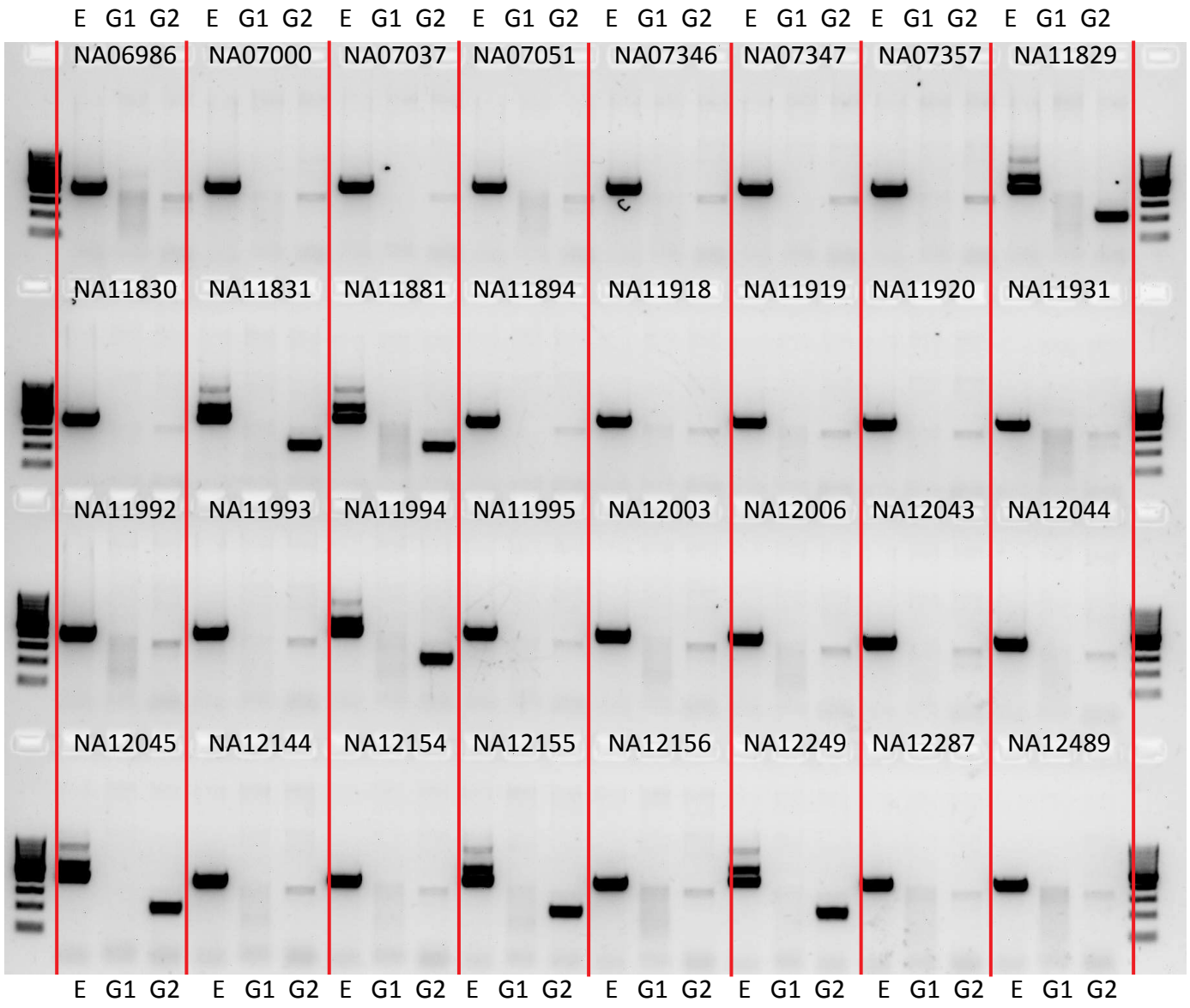
P1_MEI_2925&P2_MEI_1154



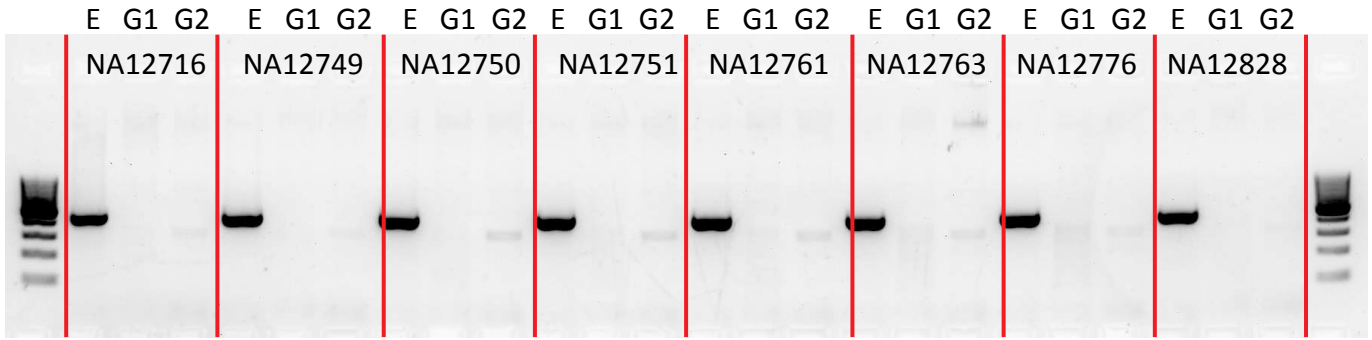
P1_MEI_2925&P2_MEI_1154



P1_MEI_3120


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA06986 NA07000 NA07037 NA07051 NA07346 NA07347 NA07357 NA11829
NA11830 NA11831 NA11881 NA11894 NA11918 NA11919 NA11920 NA11931
NA11992 NA11993 NA11994 NA11995 NA12003 NA12006 NA12043 NA12044
NA12045 NA12144 NA12154 NA12155 NA12156 NA12249 NA12287 NA12489
E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2

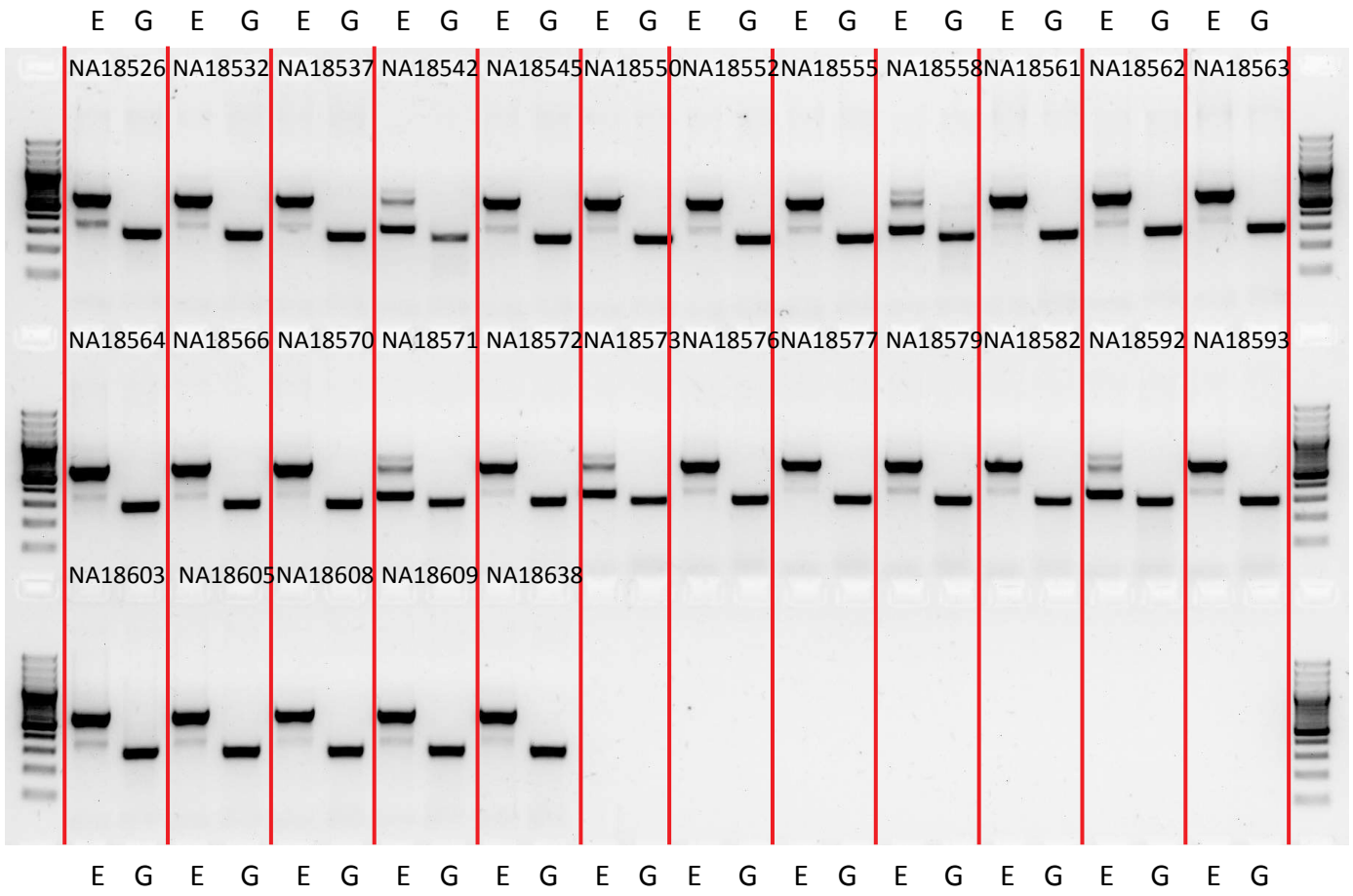
P1_MEI_3120


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA12716
NA12749
NA12750
NA12751
NA12761
NA12763
NA12776
NA12828

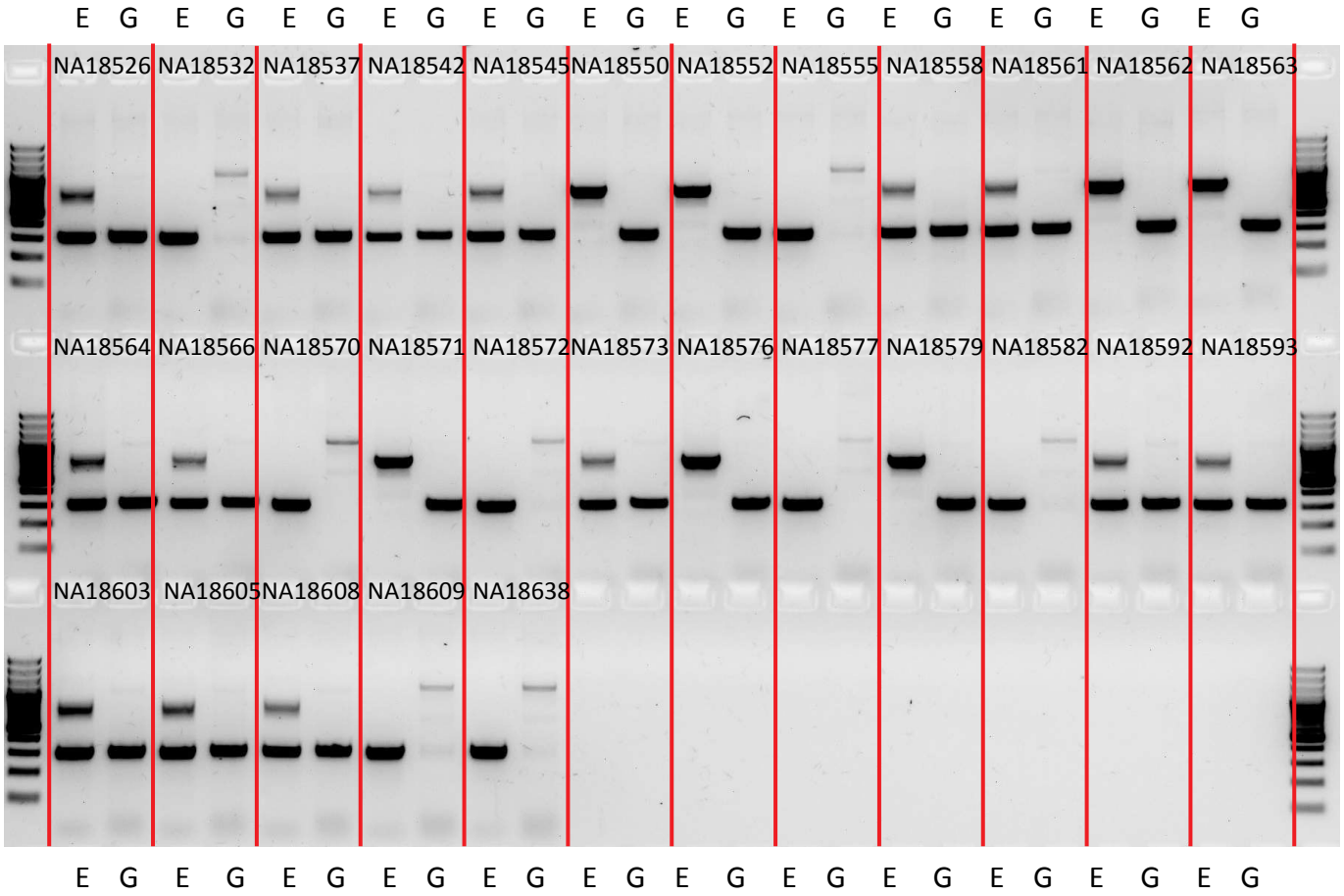
CHB

Site	Page
P1_M_061510_1_185	68
P1_M_061510_1_239	69
P1_M_061510_1_391	70
P1_M_061510_4_203	71
P1_M_061510_4_354	72
P1_M_061510_9_218	73
P1_M_061510_10_203	74
P1_MEI_190&P2_MEI_1442	75
P1_MEI_414&P2_MEI_691	76
P1_MEI_539&P2_MEI_776	77
P1_MEI_1381	78
P1_MEI_2516&P2_MEI_1951	79
P1_MEI_3095&P2_MEI_1256	80

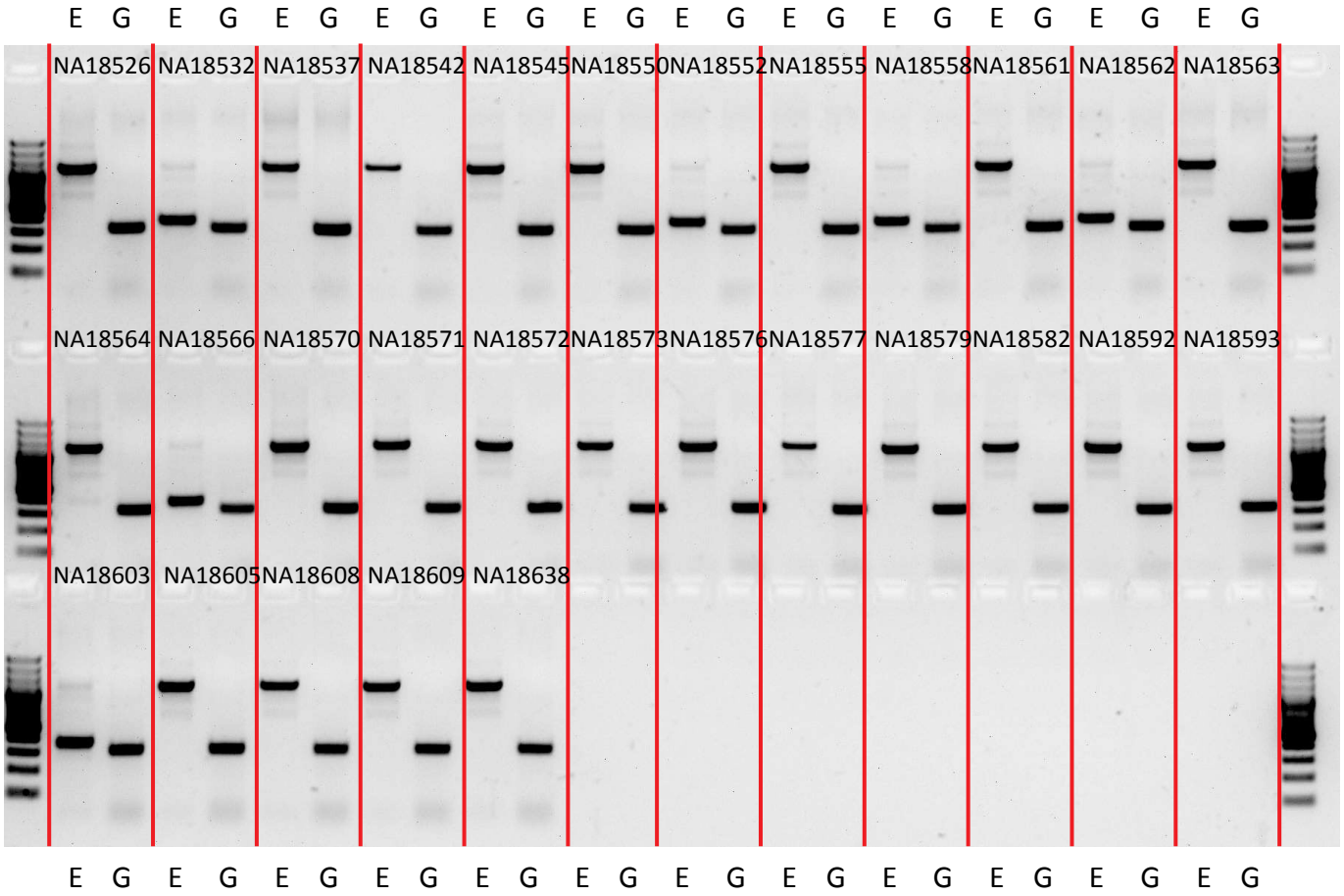
P1_M_061510_1_185



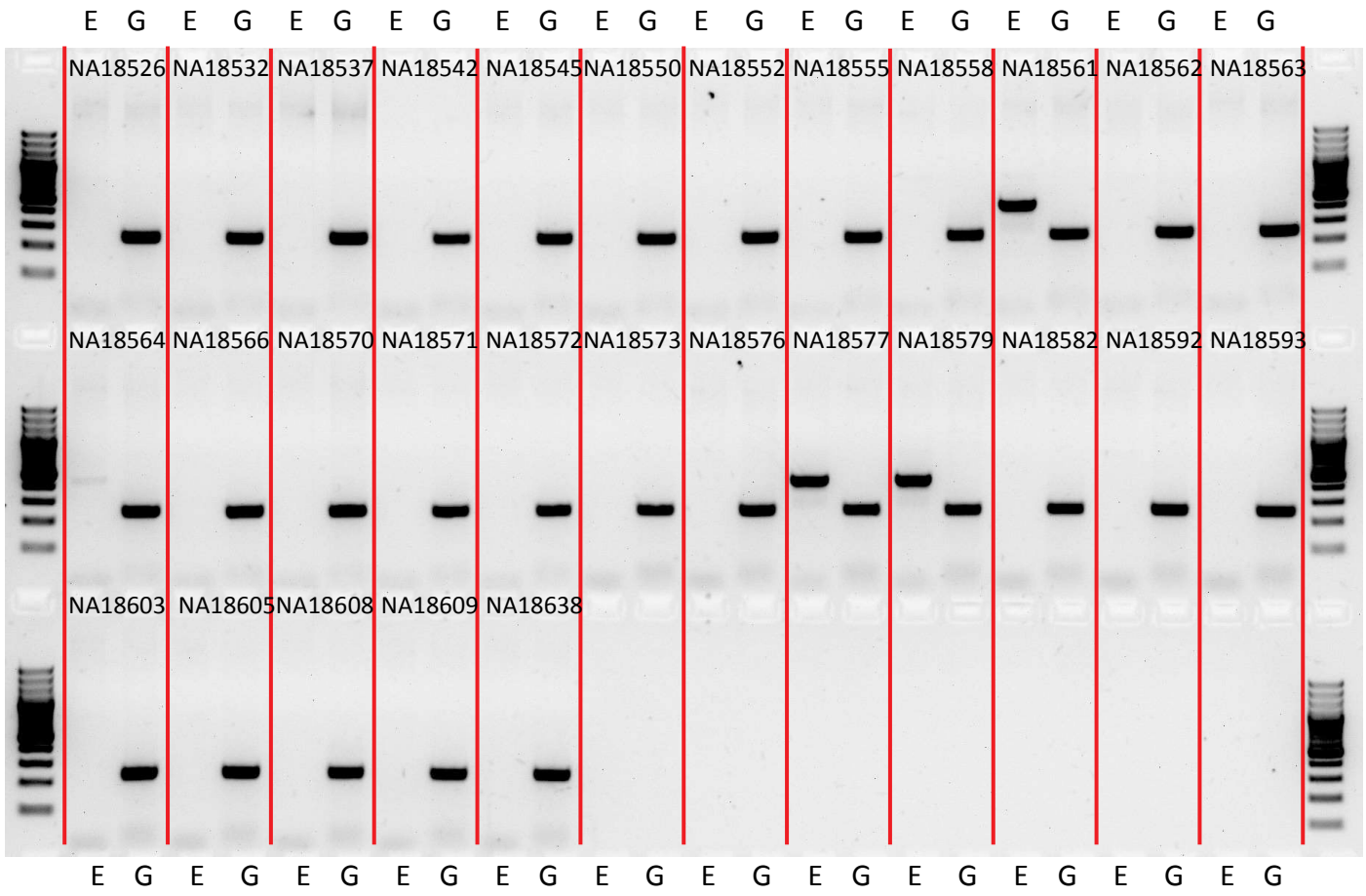
P1_M_061510_1_239



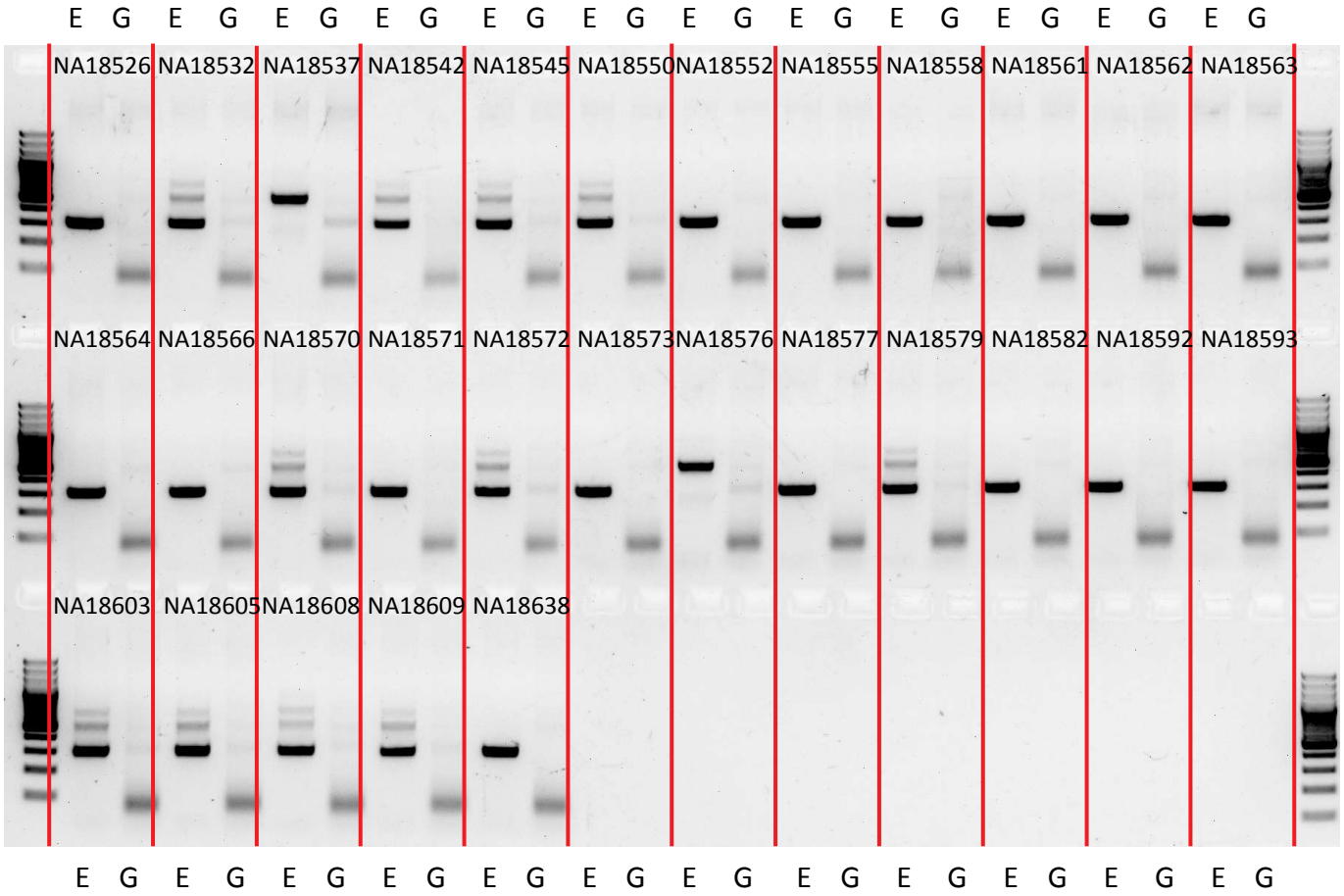
P1_M_061510_1_391



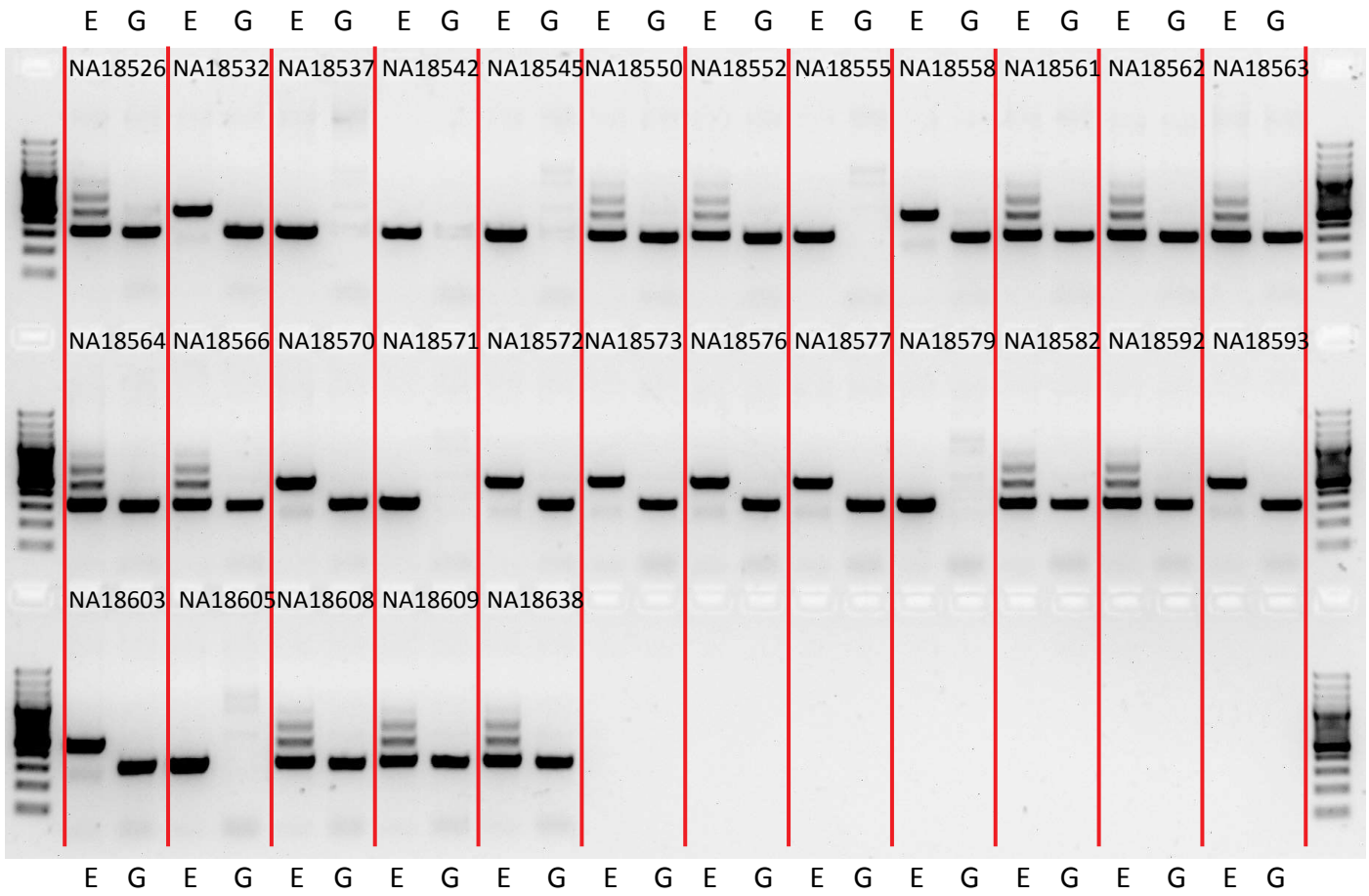
P1_M_061510_4_203



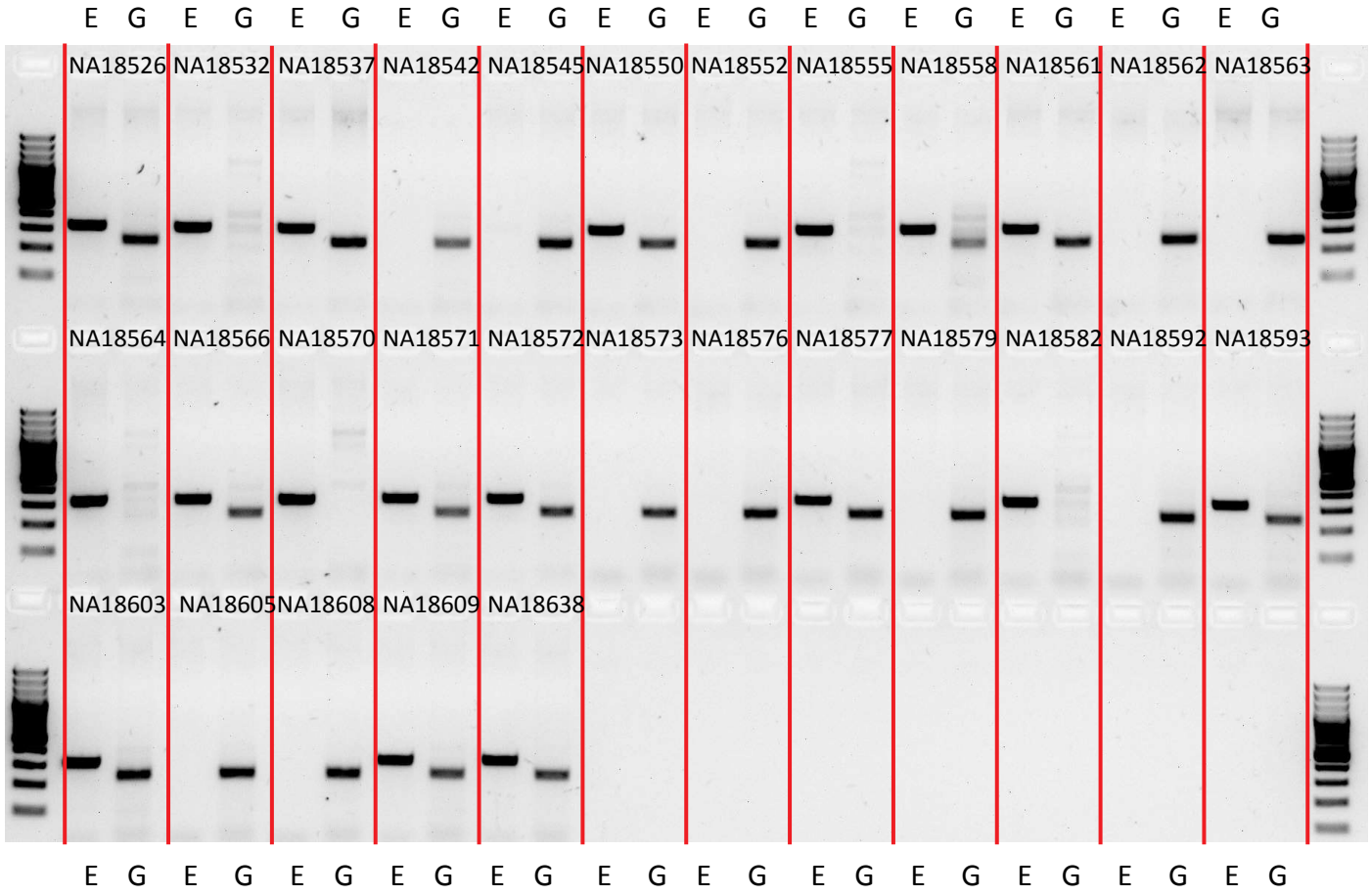
P1_M_061510_4_354



P1_M_061510_9_218

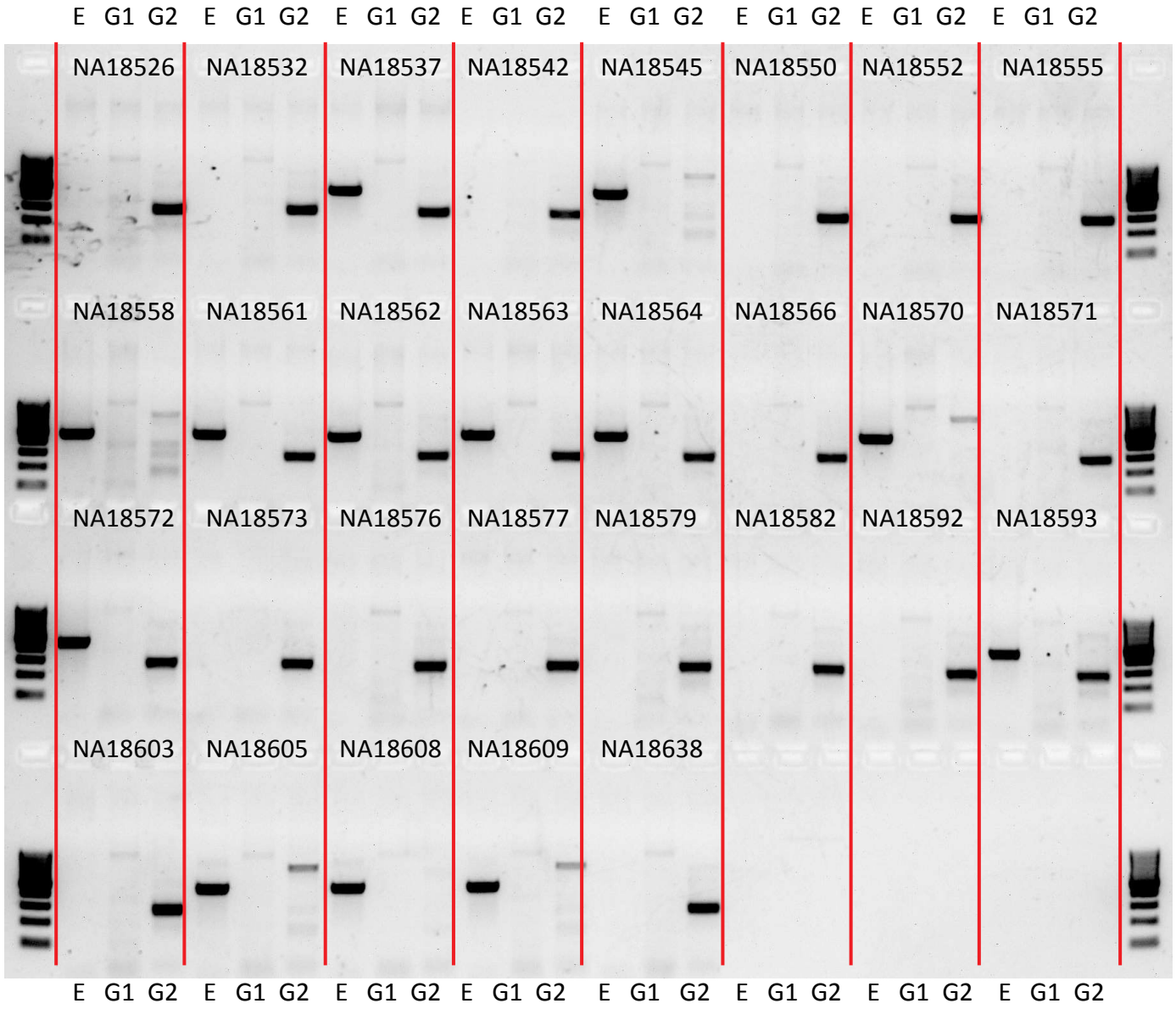


P1_M_061510_10_203

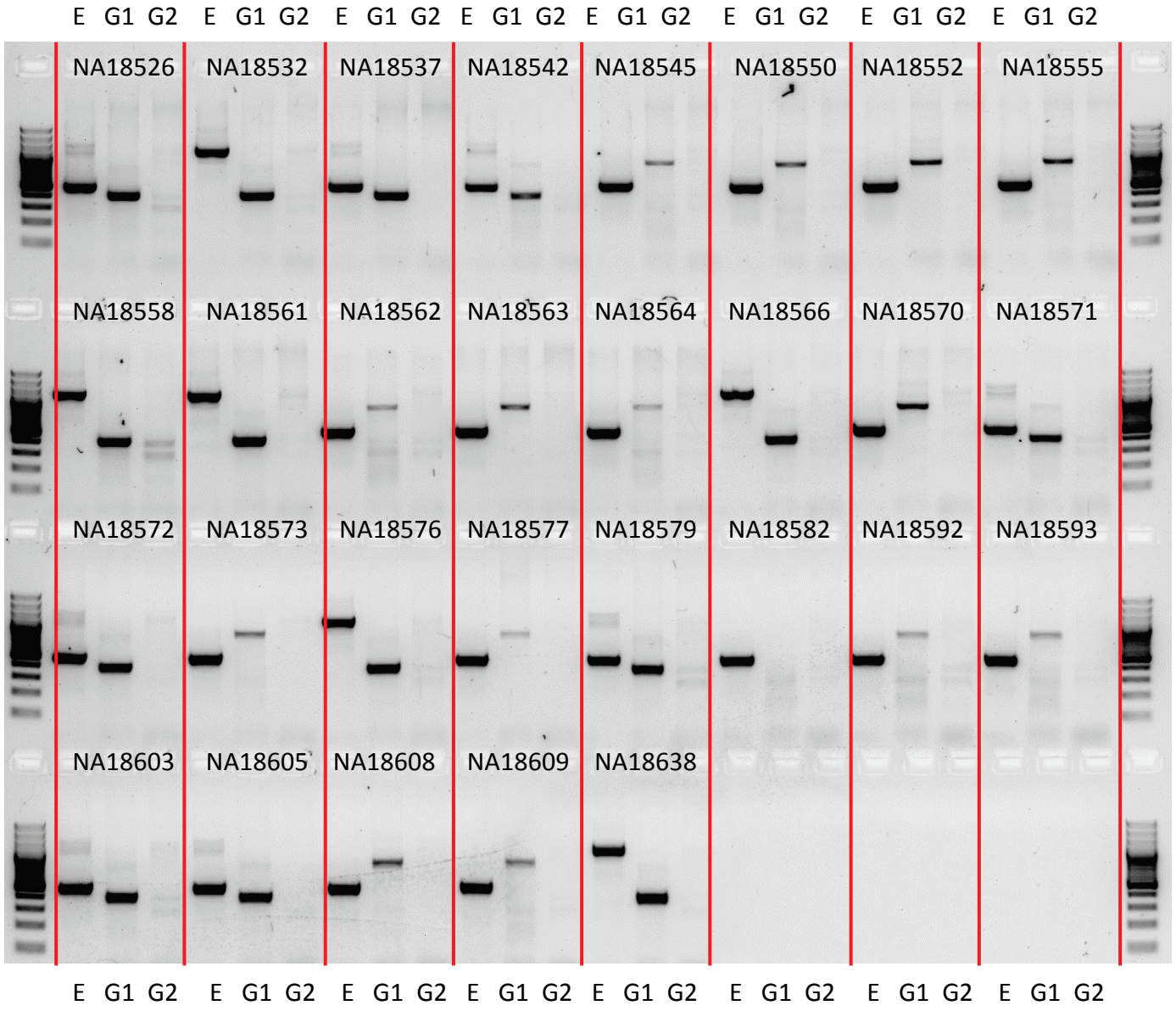


P1_MEI_190&P2_MEI_1442

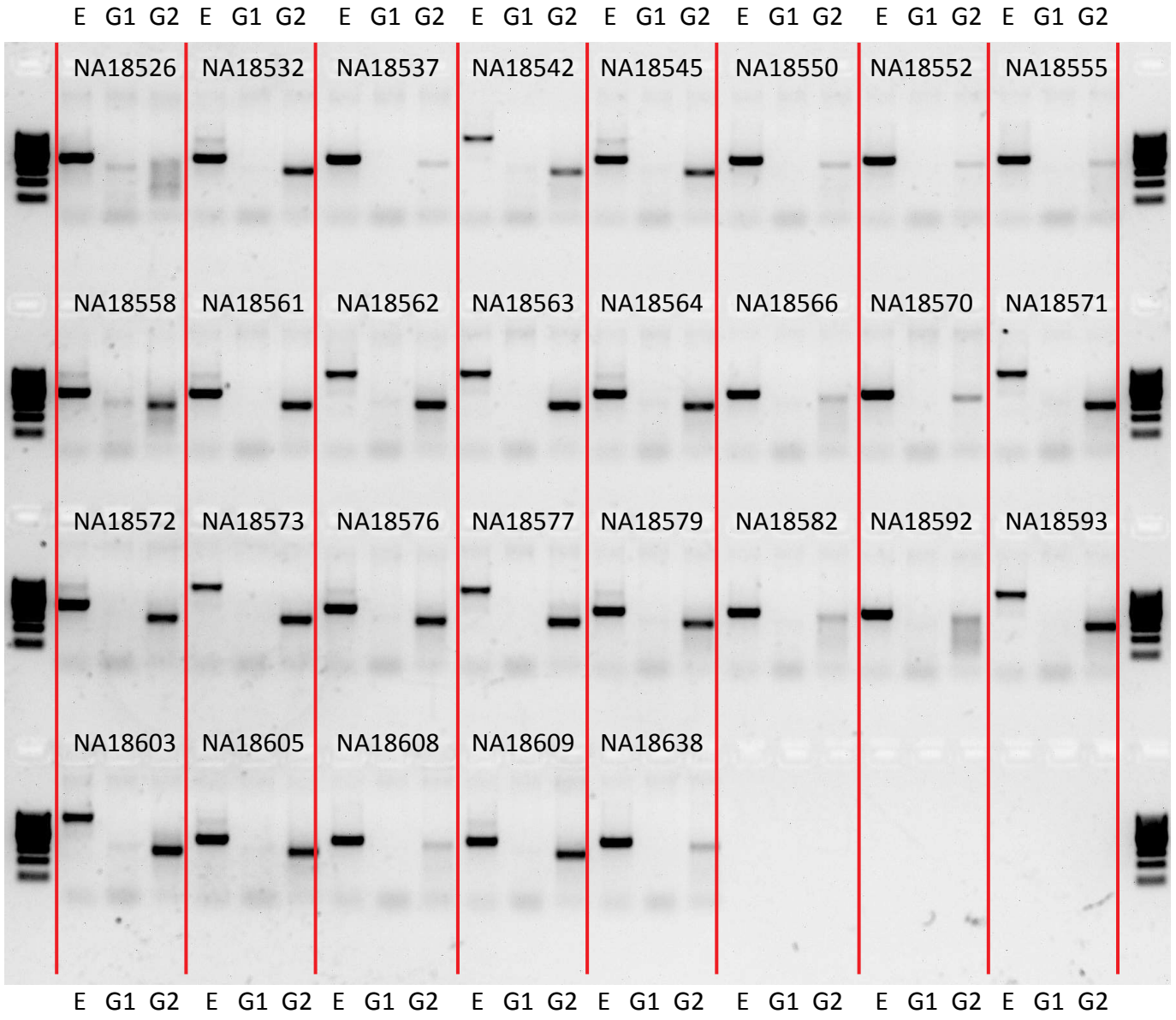
P1_MEI_414&P2_MEI_691



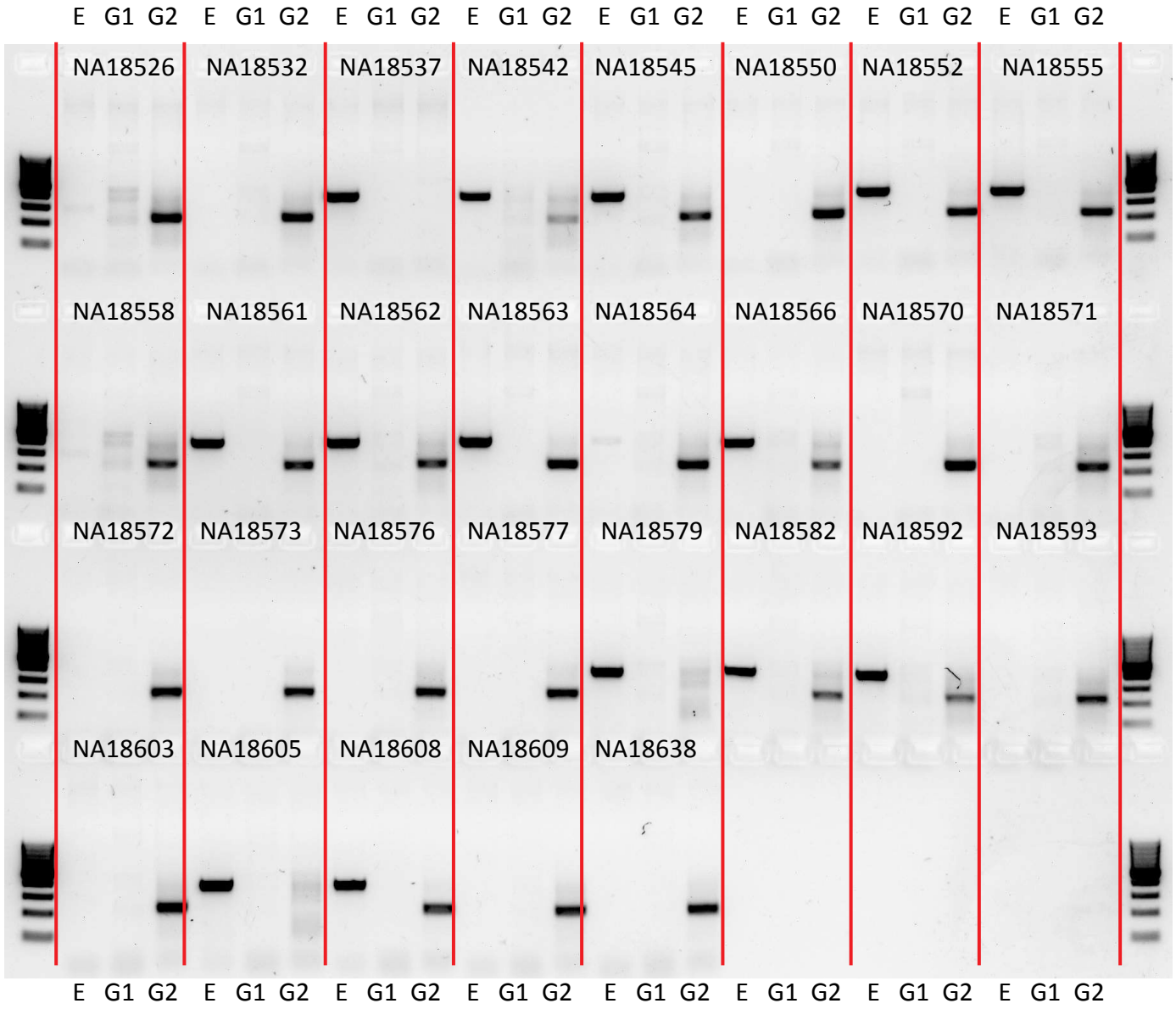
P1_MEI_539&P2_MEI_776



P1_MEI_1381


E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2
NA18526 NA18532 NA18537 NA18542 NA18545 NA18550 NA18552 NA18555
NA18558 NA18561 NA18562 NA18563 NA18564 NA18566 NA18570 NA18571
NA18572 NA18573 NA18576 NA18577 NA18579 NA18582 NA18592 NA18593
NA18603 NA18605 NA18608 NA18609 NA18638
E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2 E G1 G2

P1_MEI_2516&P2_MEI_1951



P1_MEI_3095&P2_MEI_1256