Variant Phasing and Haplotypic Expression from Single-molecule

Long-read Sequencing in Maize

Bo Wang[1], Elizabeth Tseng[2], Primo Baybayan[2], Kevin Eng[2], Michael Regulski[1], Yinping Jiao[1],

Liya Wang[1], Andrew Olson[1], Kapeel Chougule[1], Peter Van Buren[1], Doreen Ware[1,3*]

[1] Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America

[2] Pacific Biosciences, 1380 Willow Road, Menlo Park, California, United States of America

[3] USDA ARS NEA Robert W. Holley Center for Agriculture and Health Cornell University, Ithaca, New York, United States of America

*Corresponding author:

E-mail: ware@cshl.edu

Abstract

Haplotype phasing of genetic variants is important for interpretation of the maize genome, population genetic analysis, and functional genomic analysis of allelic activity. Accordingly, accurate methods for phasing full-length isoforms are essential for functional genomics study. In this study, we performed an isoform-level phasing study in maize, using two inbred lines and their reciprocal crosses, based on single-molecule full-length cDNA sequencing. To phase and analyze full-length transcripts between hybrids and parents, we developed a tool called IsoPhase. Using this tool, we validated the majority of SNPs called against matching short read data and identified cases of allele-specific, gene-level, and isoform-level expression. Our results revealed that maize parental and hybrid lines exhibit different splicing activities. After phasing 6,847 genes in two reciprocal hybrids using embryo, endosperm and root tissues, we annotated the SNPs and identified large-effect genes. In addition, based on single-molecule sequencing, we identified parent-of-origin isoforms in maize hybrids, different novel isoforms between maize parent and hybrid lines, and imprinted genes from different tissues. Finally, we characterized variation in cis- and trans-regulatory effects. Our study provides measures of haplotypic expression that could increase power and accuracy in studies of allelic expression.

Introduction

Phasing of genetic variants is crucial for identifying putative causal variants and characterizing the relationship between genetic variation and phenotype [1]. Maize is a diploid organism that not only has high genetic diversity, but also exhibits allele-specific expression (ASE) [2,3,4], i.e., unequal transcription of parental alleles. The effects of ASE vary by cell/tissue type, developmental stage, and conditions [5]. Because alleles from the same gene can generate heterozygous transcripts with distinct sequences, a comprehensive analysis of ASE is necessary in order to achieve a thorough understanding of transcriptome profiles [6,7]. Previously, ASEs were studied using short-read RNA-seq, which could quantify alleles at the SNP level but was unable to provide full-length haplotype information [8,9]. Third-generation sequencing technologies such as PacBio and Oxford Nanopore offer full-length transcript sequencing that eliminates the need for transcript reconstruction, and these methods have been widely adopted for genome annotations [10–14]. However, only a handful of studies used long reads for isoform-level haplotyping [6].

In this study, we used the Pacific Biosciences Sequel platform to produce a single-molecule full-length cDNA dataset for two maize parental lines, the temperate line B73 and the tropical line Ki11, as well as their reciprocal hybrid lines (B73 × Ki11;

3

58   Ki11 × B73). We also sequenced RNAs from the same tissues using 150-bp

59   paired-end (PE) sequencing on the Illumina platform. We developed a tool called

60   IsoPhase to phase the allelic isoforms in the hybrids based on the single-molecule

61   transcriptome datasets. This is the first full-length isoform phasing study in maize,

62   or in any plant, and thus provides important information for haplotype phasing in

63   other species, including polyploid species. Using IsoPhase, we demonstrated that

64   haplotype phasing by full-length transcript sequencing can reveal allele-specific

65   expression in maize reciprocal hybrids. Our approach does not require parental

66   information (although parental data could be used to assign maternal and paternal

67   alleles) and can be applied to exclusively long-read data. Moreover, we show that

68   single molecules can be attributed to the alleles from which they were transcribed,

69   yielding accurate allelic-specific transcriptomes. This technique allows the

70   assessment of biased allelic expression and isoform expression.

71

72   **Results**

73   **Full-Length Transcript Sequencing and Bioinformatics Pipeline**

74   Reciprocal hybrids of the maize inbred lines B73 and Ki11 exhibited dramatic

75   heterosis in plant height, primary and lateral root number, biomass and 100-kernel

76   Weight (S1A–S1H Fig), making this group well suited to the allelic study of

4

77   heterosis. To identify as many transcripts as possible, we extracted high-quality

78   RNA from root, embryo, and endosperm of B73, Ki11, as well as their reciprocal

79   crosses (B73 × Ki11, Ki11 × B73), and subjected the RNA to reverse transcription.

80   Tissue-specific barcodes were added before pooling for subsequent amplification.

81   Barcoded SMRTBell libraries were sequenced on a PacBio Sequel 1M platform

82   with 15 SMRT Cells using 2.1 chemistry, yielding 4,898,979 circular consensus

83   sequences (CCS, also called reads of insert). We pooled the reads from all four

84   lines and processed them using the IsoSeq3 workflow (Fig 1); 76.3% (3,739,812)

85   of reads were classified as full-length based on the presence of barcoded primers

86   and polyA tails (S1 Table). IsoSeq3 processing yielded 250,168 full-length, high-

87   quality (HQ) consensus transcript sequences.

88   The HQ transcript sequences were then mapped to the maize RefGen_v4

89   genome assembly [14] using minimap2 [15]. Of the 250,168 HQ sequences,

90   248,424 (99.3%) were mapped to the genome, among which 229,757 (91.8%) were

91   selected according to two criteria: min-coverage 99% and min-identity 95%. These

92   sequences were further collapsed into 90,419 non-redundant transcripts (S2 Table).

93   Using the QC tool SQANTI [16], we discarded 15,301 (16.9%) of the non-

94   redundant transcripts due to intra-priming, RT switching, or non-canonical

95   junctions unsupported by matching short-read data (S3 Table). Genome-wide

96   BLASTN to the NCBI RefSeq NR database revealed that 523 (30%) of the 1,744

97   unmapped sequences fell into gaps in the assembly, whereas the remaining

98   sequences could be mapped to other organisms, implying that they represented

99   biological contamination from endophytes or other sources (S4 Table). Our final

100 dataset consists of 75,118 transcripts covering 23,412 gene loci, with lengths

101 ranging from 80 to 11,495 bp and an average transcript length of 2,492 bp (S2 Fig).

102 **Isoform Characterization in Maize B73, Ki11, and Reciprocal Lines**

103 We used SQANTI to compare reference transcripts against the maize B73

104 RefGen_v4 annotation [14], matching 20,068 of the 23,412 loci to a reference gene

105 locus; the remaining 3,344 transcripts, with an average gene length of 2,350 bp,

106 were deemed novel. Among these novel transcripts 3,193 (95.5%) are intergenic,

107 and 319 (4.5%) are antisense. The novel transcripts are not annotated in the

108 existing maize B73 genome.

109   SQANTI classified the transcripts into six groups (Fig 2A): a] 36,005

110 (47.9%) isoforms were FSM (full splice matches), matching a reference perfectly

111 for all exonic junctions; b] 8,910 (11.9%) isoforms were ISM (incomplete splice

112 matches), matching a reference perfectly but with fewer exons; c] 13,521 (18.0%)

113 were NIC (novel in catalog), novel isoforms using combinations of known

114 donor/acceptor sites; d] 13,170 (17.5%) were NNC (novel not in catalog), novel

115 isoform using at least one novel donor or acceptor site; e] 319 (0.4%) were

6

116   Antisense; and f] 3,193 (4.3%) were Intergenic. The Iso-Seq data recovered

117   proportionately just as many full-length reference transcripts at longer read lengths

118   (Fig 2B), whereas the shortest transcripts (< 1 kb) and novel genes were enriched

119   for mono-exonic transcripts (Fig 2C–2D), of which 44.3% (4,702) were non-

120   coding. In addition, we identified alternative start and end sites for these transcripts

121   (Fig 2E–2F).

122

123   We then demultiplexed the pooled transcripts by looking at the number of reads

124   associated with each strain-tissue (ex: B73-root) in each transcript. If a sample had

125   at least one full-length read associated with a transcript, it was considered to be

126   expressed. Each sample contained between 20,000 and 30,000 expressed

127   transcripts, with relatively similar length distributions (S5 Table and S3 Fig). To

128   determine the degree of saturation of the data, we subsampled the full-length reads

129   associated with the transcripts by strain and by tissue; the results of this analysis

130   revealed that the Iso-Seq data was saturated at the gene level, but was still

131   revealing additional diversity at the transcript level (S4A–S4D Fig). Note, however,

132   that this saturation analysis was limited by the input library size and subject to

133   sequencing bias on the PacBio platform.

134

7

135 Comparison of genes/isoforms between parents and hybrids revealed that

136 reciprocal F1s exhibited maternal dominance (S5 Fig). We found a number of

137 shared and genotype-specific genes and isoforms between the parental line and two

138 hybrids (S6 Fig). Investigation of the main splicing patterns revealed that intron

139 retention was the dominant pattern across these four genotypes, and alternative last

140 exon the least common pattern (S7A–S7C Fig). Quantification using Illumina short

141 reads showed that most genes exhibited additive expression patterns in three

142 tissues. We also observed differences between the two reciprocal hybrids:

143 specifically, B73 × Ki11 had more non-additive genes than Ki11 × B73 (S8A–S8C

144 Fig), which could contribute to phenotypic differences between these two hybrid

145 lines. However, there were no significant differences in the number of isoforms

146 between genes with additive expression and those with non-additive expression,

147 nor were there differences in genes with non-additive expression between the two

148 reciprocal hybrids (S9A–S9C Fig).

149

150 **Full-Length Transcripts Enable Accurate Haplotyping**

151 To phase Iso-Seq transcripts, we developed a new tool called IsoPhase. For each

152 gene, we aligned all full-length reads to the gene region, and then called SNPs

153 individually (currently, IsoPhase only calls substitution SNPs). We then used the

154 full-length read information to reconstruct the haplotypes and used a simple error

155   correction scheme to obtain the two alleles (Fig 3A). To determine which allele

156   belongs to B73 or Ki11, we took advantage of the fact that all B73 reads must only

157   express one allele, and all Ki11 reads must only express the other. Once the

158   parental alleles were identified, we obtained the allelic counts for the F1 hybrids

159   (Fig 3B). We applied IsoPhase to the 9,463 genes that had at least 40 FL read

160   coverage, of which 6,907 had at least one SNP and could be readily classified as

161   the B73 or Ki11 alleles.

162

163   We validated the SNPs called from IsoPhase using short-read data. Considering

164   only substitution SNPs at positions for which there was at least 40 FL read

165   coverage, 96% (74280 of 77540) of IsoPhase SNPs were validated by short-read

166   data. The remaining 4% (3260) of SNPs that were PacBio-specific were mostly

167   due to insufficient coverage of the UTR regions by short-read data. Conversely,

168   short reads identified an additional 26,774 SNPs in the regions that were not

169   confirmed by IsoPhase. There were several reasons for this: (1) low or dropped

170   coverage of Iso-Seq data; (2) alignment artifacts; and (3) indels masquerading as a

171   series of consecutive SNPs. Both short read and Iso-Seq data showed variable

172   coverage at the 5' ends. In some cases, short-read data called additional 5' SNPs

173   (Fig 4A). At positions where short reads called a SNP but Iso-Seq had sufficient

174   coverage, however, read mis-alignment was a common issue (S10A–S10B Fig). In

9

175   summary, we have confidence in the joint SNP calls, and attribute the unique SNP

176   calls to either false negatives (due to low coverage) or false positives (due to mis-

177   alignment).

178   **Full-Length Transcript Reads Reveal Allelic Specific Expression**

179   An advantage of full-length transcript sequencing is the ability to characterize

180   isoforms with haplotype information. Among the highly expressed genes, we

181   observed cases of allelic specific expression. For example, only the maternal allele

182   of the gene Zm00001d037529 (PB.16588) was expressed in the F1 hybrids, and

183   both Iso-Seq and short-read data supported maternal-only expression (Fig 4B). The

184   Zm00001d040612 (PB.8517) gene provides a remarkable example of allelic-

185   specific isoform expression. Its two most abundant isoforms were PB.8517.1 and

186   PB.8517.4, which differ only in the last exon: PB.8517.4 has a single 3' exon,

187   whereas in PB.8517.1 the last exon is spliced into two exons. In B73, PB.8517.4

188   was expressed but PB.8517.1 was not, whereas in Ki11 the opposite pattern was

189   observed. In both F1 hybrids, both isoforms were expressed, but only the Ki11

190   allele was detected for the PB.8517.1 isoform, whereas only the B73 allele was

191   detected for the PB.8517.4 isoform (Fig 5). Short-read junction data supported this

192   observation. In total, we identified 221 monoallelic genes in embryo, 527 in

193   endosperm, and 271 in root (S11A Fig). Comparison of the number of isoforms in

10

194   these monoallelic genes revealed differences among the three tissues: in embryo

195   and root, the two reciprocal hybrids were inclined to exhibit the maternal effect

196   pattern, whereas in endosperm, reciprocal hybrid B73 × Ki11 had more isoforms

197   (S11B Fig).

198

199   We conclude that while short-read data achieves higher sequencing depth and can

200   call more SNPs, full-length transcripts deliver accurate haplotype information with

201   high specificity and can be used to study allele-specific expression. In future work,

202   combining the deep coverage of short read data with full-length long-read data

203   should drastically improve both the sensitivity and specificity of transcript

204   haplotyping.

205

206   **Functional Annotation of the SNPs**

207   We performed functional annotation of SNPs called by IsoPhase using the maize

208   reference genome annotation v4. Among all SNPs, 24% were synonymous variants.

209   Of the non-synonymous variants, 22,093 had potential large effects on the function

210   of 5,140 genes, including 21,685 missense, 287 splice donor/acceptor, and 243

211   stop gained/retained variations. In addition, 10% and 17% variants were in 3'

212   UTRs and 5' UTRs, respectively (Fig 6A–6B). Among those, 2,556 genes had

213   SIFT (sort intolerant from tolerant) scores < 0.05 and were therefore predicted to

11

214   be deleterious mutations. Gene Ontology analysis revealed that most of the large-

215   effect genes were associated with 'molecular function' terms related to catalytic

216   activity and binding, and 'biological process' terms related to metabolic and

217   cellular processes (Fig 6C). The resultant differences in these processes could

218   contribute to the phenotypic differences between Ki11 and B73, as well as the

219   differences between the hybrids and their parents. We also found that these large-

220   effect genes had more isoforms in root and embryo tissues in B73, but more

221   isoforms in endosperm tissue in Ki11. In addition, the number of isoforms in the

222   two reciprocal hybrids exhibited the high-parent value in root and embryo, but not

223   in endosperm, suggesting that genes in endosperm play a role in determining the

224   developmental differences between B73 and Ki11 plants (Fig 6D–6F).

225

226   **Imprinted genes and cis-/trans-regulatory effects**

227   We identified a number of imprinted genes based on the phasing results from

228   IsoPhase: 172 paternally inherited genes in embryo, 221 in endosperm, and 200 in

229   root; and 193 maternally inherited genes in embryo, 326 in endosperm, and 196 in

230   root. The expression of each allele confirmed the paternal and maternal expression

231   pattern of these imprinted genes (Fig 7A–7F). Comparison of splicing patterns

232   between parents and hybrids revealed that overall intron retention was the

233   predominant pattern across the four genotypes, regardless of tissue and paternal or

234   material imprinting; however, the proportion of each splicing pattern varied among

235   genotypes in specific tissues (Fig 7G–7L).

236

237   Variation in cis- and trans-regulation can be distinguished by allelic expression

238   ratios between parents relative to the F1 interspecific hybrid or allotetraploid. In

239   order to see cis- and trans-effects on gene expression divergence, we calculated the

240   allelic ratios in parents and reciprocal hybrids from the phasing results, and used

241   this information to identify genes with cis or trans effects. The results revealed that

242   cis + trans effects were predominant in the two hybrids among all the three tissues,

243   following by cis × trans effects; conserved genes were the least common (S12A

244   Fig). In addition, the number of isoforms was slightly higher in conserved genes

245   than in genes with other effects (S12B Fig).

246

247   **Discussion**

248

249   Maize is a diploid important genetic model for elucidating transcriptional networks.

250   Recently, full-length transcript sequencing using long-read technology has enabled

251   us to characterize alternative splicing events and improve the maize genome

252   annotation [14,17]. However, the general Iso-Seq algorithm ignores SNP-level

253   information, focusing instead on identifying alternative splicing differences.

254   Heterosis has been extensively studied in plants using transcriptome sequencing

255   approaches, revealing differentially expressed genes and biased expressions

256   between parents and hybrid lines [18–20]. The results of these studies have

257   provided clues about the molecular mechanisms underlying plant heterosis [21]. To

258   date, however, no study has characterized allelic expression at single-molecule

259   full-length transcript level. Maize, with its high diversity, provides an excellent

260   model for studying heterosis.

261

262   On the other hand, transcriptome sequencing has a wide variety of applications,

263   including evaluation of differences in gene expression between tissues, conditions,

264   etc. By contrast, allele-specific expression analysis evaluates expression

265   differences between two parental alleles in their hybrids. Both differential and

266   allele-specific expression analyses have been employed to study heterosis using

267   traditional RNA-seq [22,23], but the short read lengths of this technique make it

268   impossible to construct the parental origins at the full-length transcript level. Full-

269   length, single-molecule sequencing provides an unprecedented allele-specific view

270   of the haploid transcriptome. Haplotype phasing using long reads allowed us to

271   accurately calculate allele-specific transcript and gene expression, as well as

272   identify imprinted genes and investigate the cis/trans regulatory effects.

273   Sequencing of full-length haplotype-specific isoforms enabled accurate assessment

274 of allelic imbalance, which could be used to study the molecular mechanisms

275 underlying genetic or epigenetic causative variants and associate expression

276 polymorphisms with plant heterosis.

277

278 Maize is an excellent model for haplotype phasing study, as it is very polymorphic

279 as well as both inbreds and hybrids are easily obtainable and reference genomic

280 sequence is available. In this study, we sequenced full-length cDNAs in two

281 reciprocal hybrids and their parental lines using the PacBio Iso-Seq method. To

282 phase isoforms in the hybrids, we developed IsoPhase, an accurate method for

283 reconstructing haplotype-specific isoforms. Based on the assumption that the SNPs

284 in each inbred line are homozygous, our method uses splice mapping to partition

285 the reads into parental haplotypes. Using IsoPhase, we successfully phased 6,847

286 genes using single-molecule sequencing data from maize reciprocal F1 hybrids and

287 their parents. It is important to note that we only phased genes supported by more

288 than 40 FL reads; the rest of the genes were not phased due to the sequencing

289 depth cutoff. In addition, we used short-read data from the same tissues to confirm

290 the SNPs called from long-reads sequencing, giving us high confidence in the

291 resultant phased genes. This is the first full-length isoform phasing study in maize,

292 or in any plant, and thus provides important information for haplotype phasing to

293 other organisms, including polyploid species. IsoPhase can also be used for self-

15

294   incompatible species for haplotype phasing, although a high level of

295   heterozygosity in such species makes it very challenging and would require deeper

296   sequencing.

297

298   We noticed that the allelic expression of phased genes varied among tissues.

299   Among the phased genes, we also identified imprinted genes in endosperm,

300   embryo and root, and we found that genes expressed in the embryo had more

301   isoforms than imprinted genes expressed in endosperm and root. However, due to

302   the large number of full-length reads required for high-confidence phasing by

303   IsoPhase, future studies involving large-scale deep sequencing will be required to

304   phase all genes in maize.

305

306   Methods

307   **Plant materials**

308   Maize inbred lines B73 and Ki11 were grown at CSHL Uplands Farm, and reciprocal crosses

309   were made between the two lines. For tissue collection, embryo and endosperm at 20 DAP (days

310   after pollination) were collected for each genotype in two biological replicates, and root tissues

311   were collected at 14 DAG (days after germination). All tissues were immediately frozen in liquid

312   N_2. For each tissue, at least 10 plants were pooled in each biological replicate.

313

314 **RNA preparation**

315 Total RNA was prepared by grinding tissue in TRIzol reagent (Invitrogen 15596026) on dry ice

316 and processing as recommended by the manufacturer. To remove DNA, an aliquot of total RNA

317 was treated with RQ1 DNase (Promega M6101), followed by phenol/chloroform/isoamyl alcohol

318 extraction and chloroform/isoamyl alcohol extraction using Phase Lock Gel Light tubes (5

319 PRIME 2302800), and ethanol precipitation. Precipitated RNA was stored at -20°C.

320

321 **Illumina RNA-Seq library construction**

322 Total RNA (20 μg) was used for poly(A)$^+$ selection using oligo(dT) magnetic beads (Invitrogen

323 610-02), eluted in water, and subjected to RNA-seq library construction using the ScriptSeq™

324 kit (Epicentre SS10906). Libraries were amplified by 15 cycles of PCR, and then sequenced in

325 two lanes on the HiSeq 2500 PE150 platform at Woodbury Genome Center, Cold Spring Harbor

326 Laboratory.

327

328 **PacBio library construction and single-molecule sequencing**

329 cDNA was generated from 1 μg of total RNA per sample using the Clontech SMARTer PCR

330 cDNA Synthesis Kit (catalog# 634925 or 634926) according to PacBio's Iso-Seq Template

331 Preparation for the Sequel System (https://www.pacb.com/wp-content/uploads/Procedure-

332 Checklist-Iso-Seq-Template-Preparation-for-Sequel-Systems.pdf). cDNAs were barcoded using

333 barcoded oligo-dT during first-strand synthesis. The 16-bp barcode sequences used for this

334 study are shown in Additional file 1: S6 Table. The embryo, root, and endosperm cDNAs were

335 enriched by PCR using PrimeSTAR GXL DNA Polymerase (Clontech, catalog# R050A or

336 R050B). Amplification conditions used were as follows: initial denaturation at 98°C for 30

337 seconds, followed by 9–12 cycles of 98°C for 10 seconds, 65°C for 15 seconds, and 68°C for 10

338 minutes. A final extension was performed at 68°C for 5 minutes. After amplification, tissues

339 (embryo, root and endosperm) from the same strain were pooled in equimolar quantities,

340 yielding four pools, which were subsequently used to construct SMRTbell libraries using the

341 SMRTbell Template Prep Kit 1.0 (Pacific Biosciences, Part No. 100-259-100). After library

342 construction, each SMRTbell library was size-fractionated using SageELF (Sage Science). The

343 final SMRTbell libraries were annealed with Sequencing Primer v4 (Pacific Biosciences Part No.

344 101-359-000) and bound with Sequel Binding Kit 2.1(Pacific Biosciences Part No. 101-429-300).

345 The polymerase-bound SMRTbell libraries were loaded at 3–10 pM on-plate concentrations and

346 sequenced using the Sequel Sequencing Kit 2.1 (Pacific Biosciences Part No. 101-312-100) and

347 Instrument Software v5.1.

348 **Illumina data analysis**

349 Raw reads were aligned to the B73 reference genome (RefGen_v4) [14] using STAR 2.4.2a [24]

350 with minimum intron length of 20 bp and maximum intron length of 50 kb; default settings were

351 used for the other parameters. Quantification of genes and isoforms was performed using

352 cufflinks version 2.2.1 [25] using the GTF annotation file generated by PacBio sequencing.

353

354 **PacBio data analysis**

355 PacBio data were analyzed by running the IsoSeq3 application in PacBio SMRT Analysis v6.0 to

356 obtain high-quality, full-length transcript sequences, followed by downstream analysis.

357

358 **Full-Length reads classification**

18

359   Full-length reads were identified as CCS reads that contained both the 5' and 3' primer and the

360   polyA tail preceding the 3' primer. The 5' primer consists of the Clontech SMARTer cDNA

361   primer with an ATGGG overhang. The 3' primer consists of a tissue-specific 16-bp PacBio

362   barcode followed by the Clontech SMARTer cDNA primer (S6 Table).

363

364   **Isoform-level clustering analysis to obtain high-quality transcript sequences**

365   To increase detection of rare isoforms, the demultiplexed FL reads were pooled to perform

366   isoform-level clustering analysis [26]. After clustering, consensus sequences were called using

367   the Arrow algorithm, and only polished sequences with predicted consensus accuracy $\geq$ 99%

368   were considered high-quality (HQ) and retained for the next step.

369

370   **Mapping to B73 genome and filtering**

371   The HQ transcript sequences were mapped to B73 RefGen_v4 genome using minimap2 (version

372   2.11-r797) [15] using parameters '-ax splice -t 30 -uf --secondary=no -C5'. We then filtered for

373   alignments with $\geq$ 99% coverage and $\geq$ 95% identity and removed redundancy using scripts from

374   cDNA_Cupcake (http://github.com/Magdoll/cDNA_Cupcake). The full list of commands used at

375   this step is as follows:

```
376   minimap2 -ax splice -uf -C5 --secondary=no B73_RefV4.fa hq.fastq > hq.fastq.sam
377   sort -k 3,3 -k 4,4n hq.fastq.sam > hq.fastq.sorted.sam
378   collapse_isoforms_by_sam.py --input hq.fastq --fq -s hq.fastq.sorted.sam \
379       -c 0.99 -i 0.95 --dun-merge-5-shorter -o hq.no5merge
380   get_abundance_post_collapse.py hq.no5merge.collapsed cluster_report.csv
381   filter_away_subset.py hq.no5merge.collapsed
```

382

383   **Removing potential artifacts using SQANTI**

19

384   We applied further filtering criteria to remove potential genomic contamination and rare PCR

385   artifacts. We run a modified version of SQANTI [16] that categorizes each isoform according to

386   existing B73 RefGen_v4 annotations and a list of short-read junctions from the same samples.

387   An isoform is retained in the dataset if: 1) it is FSM/ISM/NIC and does not have intra-priming; 2)

388   It is NNC, does not have intra-priming, is not RT-switching, and all junctions are either all

389   canonical or supported by short reads; or 3) it is antisense, intergenic, genic, does not have intra-

390   priming, is not RT-switching, and all junctions are either all canonical or supported by short

391   reads. The rationale behind the filtering is to eliminate artifacts that come from intra-priming (dT

392   priming off genomic 'A' stretches), potential RT-switching, and other library or sequencing

393   errors that could introduce erroneous splice junctions. Isoforms that are categorized as

394   FSM/ISM/NIC are isoforms that use all known splice junctions, and are therefore trusted. For all

395   other categories, we only retain the isoform if the junctions are either canonical or supported by

396   short reads. This approach yields a high-confidence dataset that is well-supported by existing

397   annotation and matching short-read data.

398

399   **De-multiplex final isoforms by sample and rarefaction analysis by subsampling**

400   We recovered the relative abundance of each the final isoforms in each sample by extracting the

401   fraction of full-length reads supporting each isoform from each sample. To draw rarefaction

402   curves, we used the unique Iso-Seq IDs (format: PB.X.Y) to indicate unique transcripts and the

403   matching reference gene name (ex: Zm00001d027230) from B73 annotation. If a gene was novel,

404   we created a novel gene ID (ex: novelGene_124) for each non-overlapping strand-specific locus.

405   Subsampling was performed at 10,000 FL read intervals for 100 iterations, taking the average

406   number of unique transcript/genes observed to plot the rarefaction curves.

407

SNP calling and phasing using Iso-Seq data

409 All full-length reads from all 12 samples were aligned to the B73 RefGen_v4 genome using

410 minimap2 (v2.11-r797) [15] to create a pileup. Then, at each position with at least 10-base

411 coverage, Fisher's exact test with the Bonferroni correction was applied with a p-value cutoff of

412 0.01. Only substitution SNPs were called. Then, sample-specific haplotype information was

413 obtained by looking at the number of FL reads associated with each allele. To account for

414 residual errors in the FL reads, we error-corrected the haplotypes down to the two dominant

415 haplotypes (maternal and parental haplotype) that minimizes the edit distance. Functional

416 annotation of SNPs was performed using SnpEff [27] and Gramene [28] based on the maize B73

417 genome annotation RefGen_v4. SIFT 4G [29] was used to predict deleterious mutations through

418 SciApps platform [30].

419

Identification of paternal and maternal imprinted genes

421 Bm represents the percentage of expression of maternal alleles out of all expressed genes,

422 including the maternal and paternal alleles. The Bm ratio is calculated for each cross maternal

423 depth/(maternal + paternal depth) and averaged. The overall mean and standard deviation are

424 calculated and used to define cutoffs for MEG and PEG variants. A variant is marked MEG if its

425 Bm is at least two standard deviations above the mean. Conversely, it is marked PEG if its Bm is

426 at least two standard deviations below the mean.

427

Contributions

429 B.W., E.T., and D.W. conceived the idea for the study; B.W. collected the tissues; M.R.

430 generated RNA and Illumina libraries; P.B. and K.E. generated the PacBio libraries and data;

431 B.W., E.T., Y.J., L.W., A.O., and K.C. analyzed the data; and B.W. and E.T. wrote the

432 manuscript.

Accession codes

434 The data generated in this study, including PacBio Iso-Seq reads and Illumina short reads, have

435 been submitted to ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) under accession numbers

436 E-MTAB-7837 and E-MTAB-7394. The IsoPhase tool developed in this study is available in the

437 GitHub repository: https://github.com/magdoll/cdna_cupcake.

Acknowledgements

Competing financial interests

443 E.T., P.B., and K.E. are full-time employees of Pacific Biosciences. All other authors declare no

444 competing financial interests.

445

References

1. Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. Nat Commun. 2016; 7:12817.

2. Springer NM, Stupar RM. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. Plant Cell. 2007; 19(8):2391-402.

3. Stupar RM, Hermanson PJ, Springer NM. Nonadditive expression and parent-of-origin effects identified by microarray and allele-specific expression profiling of maize endosperm. Plant Physiol. 2007;145(2):411-25.

4. Waters AJ, Bilinski P, Eichten SR, Vaughn MW, Ross-Ibarra J, Gehring M, Springer NM. Comprehensive analysis of imprinted genes in maize reveals allelic variation for imprinting and limited conservation with other species. Proc Natl Acad Sci U S A. 2013; 110(48):19639-44.

5. Gregg C. Known unknowns for allele-specific expression and genomic imprinting effects. F1000Prime Rep. 2014; 6:75.

6. Deonovic B, Wang Y, Weirather J, Wang XJ, Au KF. IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. Nucleic Acids Res. 2017; 45(5):e32.

7. Waters A, Makarevitch I, Eichten SR, Swanson-Wagner RA, Yeh CT, Xu W, et al. Parent-of-origin effects on gene expression and DNA methylation in the maize endosperm. Plant Cell. 2011; 23(12):4221-33.

467   8. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best

468   practices for data processing in allelic expression analysis. Genome Biol. 2015 Sep

469   17;16:195. doi: 10.1186/s13059-015-0762-6.

470   9. Shao L, Xing F, Xu C, Zhang Q, Che J, Wang X, Song J, Li X, Xiao J, Chen LL, Ouyang Y,

471   Zhang Q. Patterns of genome-wide allele-specific expression in hybrid rice and the

472   implications on the genetic basis of heterosis. Proc Natl Acad Sci U S A.

473   2019;116(12):5653-5658.

474   10. Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, et al. Unveiling the complexity of

475   the maize transcriptome by single-molecule long-read sequencing. Nat Commun. 2016; 7:

476   11708.

477   11. Abdel-Ghany S, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, et al. A survey of

478   the sorghum transcriptome using single-molecule long reads. Nat Commun. 2016; 7: 11706.

479   12. Wang B. Regulski M, Tseng E, Olson A, Goodwin S, McCombie WR, et al. A comparative

480   transcriptional landscape of maize and sorghum obtained by single-molecule sequencing.

481   Genome Res. 2018; 28(6):921-932.

482   13. Wang M, Wang P, Liang F, Ye Z, Li J, Shen C, Pei L, et al. A global survey of alternative

483   splicing in allopolyploid cotton: landscape, complexity and regulation. New Phytol. 2018;

484   217(1):163-178.

485   14. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference

486   genome with single-molecule technologies. Nature. 2017; 546(7659):524-527.

487   15. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;

488   34(18):3094-3100.

489   16. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, et al.

490   SQANTI: extensive characterization of long-read transcript sequences for quality control in

491   full-length transcriptome identification and quantification. Genome Res. 2018; 28(7):1096.

492   17. Wang B, Kumar V, Olson A, Ware D. Reviving the Transcriptome Studies: An Insight Into

493   the Emergence of Single-Molecule Transcriptome Sequencing. Front Genet. 2019; 10:384.

494   doi: 10.3389/fgene.2019.00384. eCollection 2019.

495   18. Hochholdinger F, Hoecker N. Towards the molecular basis of heterosis. Trends Plant Sci.

496   2007; 12(9):427-32.

497   19. Flint-Garcia SA, Buckler ES, Tiffin P, Ersoz E, Springer NM. Heterosis is prevalent for

498   multiple traits in diverse maize germplasm. PLoS One. 2009; 4(10):e7433.

499   20. Paschold A, Jia Y, Marcon C, Lund S, Larson NB, Yeh CT, et al. Complementation

500   contributes to transcriptome complexity in maize (Zea mays L.) hybrids relative to their

501   inbred parents. Genome Res. 2012; 22(12):2445-54.

502   21. Chen Z. Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci. 2010;

503   15(2):57-71.

504   22. Song G, Guo Z, Liu Z, Cheng Q, Qu X, Chen R, et al. Global RNA sequencing reveals that

505   genotype-dependent allele-specific expression contributes to differential expression in rice

506   F1 hybrids. BMC Plant Biol. 2013; 13:221.

507   23. Bell G, Kane N, Rieseberg L, Adams K. RNA-seq analysis of allele-specific expression,

508   hybrid effects, and regulatory divergence in hybrids compared with their parents from

509   natural populations. Genome Biol Evol. 2013; 5(7):1309-23.

510   24. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast

511   universal RNA-seq aligner. Bioinformatics. 2013; 29, 15-21.

512 25. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and

513   transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat

514   Protoc. 2012; 7(3):562-78.

515 26. Gordon S, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, et al. Widespread Polycistronic

516   Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. PLoS One. 2015; 10,

517   e0132628.

518 27. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for

519   annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in

520   the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;

521   6(2):80-92.

522 28. Tello-Ruiz M, Naithani S, Stein JC, Gupta P, Campbell M, Olson A, et al. Gramene 2018:

523   unifying comparative genomics and pathway resources for plant research. Nucleic Acids

524   Res. 2018; 46(D1):D1181-D1189.

525 29. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes.

526   Nat Protoc. 2016; 11(1):1-9.

527 30. Wang L, Lu Z, Van Buren P, Ware D. SciApps: a cloud-based platform for reproducible

528   bioinformatics workflows. Bioinformatics. 2018; 34(22):3917-3920.

529

530

531 # Figure Legends

532 **Fig 1. Full-length Transcript Sequencing Workflow.**

533 Full-length transcript data are generated on the PacBio Sequel platform following the

534 recommended procedure for Iso-Seq multiplexed library preparation. Iso-Seq analysis

535 demultiplexes each read and removes the barcodes, cDNA primers, and polyA tail. The

536 demultiplexed reads are pooled and run through an isoform-level clustering algorithm that

537 produces high-quality transcript sequences, which is then aligned to the reference genome and

538 removed of artifacts. The genome-mapped genes are then used to run IsoPhase analysis to call

539 SNPs and haplotypes.

540 **Fig 2. Iso-Seq transcript categorization against maize B73 RefGen_v4 annotation using the**

541 **SQANTI software.**

542 **(A)** Isoform distribution across structural categories. FSM=Full Splice Match: matches a

543 reference transcript exon by exon. ISM=Incomplete Splice Match: matches a reference transcript

544 exon by exon, but is missing one or more 5' exons. NIC=Novel In Catalog: novel isoform using

545 known splice sites. NNC=Novel In Catalog: novel isoform using at least one novel splice site.

546 Because this analysis is performed after SQANTI filtering, all junctions must be in the

547 annotation, canonical, or supported by matching short-read data. **(B)** Classification by transcript

548 length, normalized. **(C)** Transcript lengths, mono vs multi-exon. **(D)** Distribution of mono- vs

549 multi-exon transcripts. **(E)** Distance to annotated polyadenylation site. **(F)** Distance to annotated

550 transcription start site.

551   **Fig 3. IsoPhase workflow.**

552   **(A)** For each gene, full-length reads from all 12 samples are aligned to a gene region. SNPs are

553   called individually for each position using Fisher's exact test with the Bonferroni correction,

554   applied with a p-value cutoff of 0.01. Only substitution SNPs are called. The full-length reads are

555   then used to reconstruct the haplotypes, and a simple error-correction scheme is applied to obtain

556   the two alleles. **(B)** To determine which allele is derived from B73 vs Ki11, we use the FL count

557   information associate with the homozygous parents: B73 would only express the B73 allele,

558   whereas Ki11 would only express the Ki11 allele.

559

560   **Fig 4. Example of phasing using IsoPhase.**

561   **(A)** The gene PB.12426 (Zm00001d053356) phased by IsoPhase. The top two tracks show the

562   B73 and Ki11 FL reads, and the bottom two tracks show the FL reads from the two F1 hybrids,

563   with the reads segregated by parental origin. SNPs are depicted between the B73 and Ki11 track;

564   six SNPs are shown. SNPs #2–6 were called based on both long- and short-read data (purple);

565   SNP #1 was missed by the long-read data due to reduced coverage (blue). **(B)** IsoPhase can

566   identify maternal-specific gene expression. The gene PB.16588 (Zm00001d037529) is expressed

567   in both B73 and Ki11. However, in the F1 hybrids, only the maternal allele is expressed. In Ki11

568   × B73, the maternal allele is Ki11. In B73 × Ki11, the maternal allele is B73. Short-read data

569   confirmed this maternal-specific gene expression.

570

571   **Fig 5. IsoPhase identifies allelic-specific isoform expression.**

572   The gene PB.8517 (Zm00001d040612) shows allelic-specific expression. B73 dominantly

573   expresses the PB.8517.4 isoform, which has an unspliced 3' exon. Ki11 dominantly expresses

28

574   the PB.8517.1 isoform that has the last 3' exon spliced. The two F1 hybrids express both

575   isoforms, but each isoform is associated with the parental allele.

576

577   **Fig 6. Annotations of the SNPs called from IsoPhase.**

578   **(A)** Distribution of different categories of SNPs from variant effect predictor. **(B)** Proportions of

579   different categories of SNPs from coding sequences. **(C)** Gene Ontology analysis of large-effect

580   genes. **(D–F)** Number of isoforms of large-effect genes in two parents and reciprocal hybrids

581   from root (D), endosperm (E) and embryo (F).

582

583   **Fig 7. Allelic expression and alternative splicing patterns of paternal expressed genes (PEG)**

584   **and maternal expressed (MEG) genes from phasing.**

585   **(A–C)** Allelic expression of maternal imprinted genes in reciprocal hybrids in embryo (A),

586   endosperm (B), and root (C). **(D–F)** Allelic expression of paternal imprinted genes in reciprocal

587   hybrids in embryo (D), endosperm (E) and root (F). **(G–L)** Alternative splicing patterns of PEG

588   and MEG genes between parents and two reciprocal hybrids in embryo (G–H), endosperm (I–J),

589   and root (K–L). p0 represents the B73 allele; p1 represents the Ki11 allele.

590

591

592   # Supporting information

593
594   **S1 Table. 16-mer Barcodes corresponding to the sample.**

595

596   **S2 Table. Number of full-length, non-concatemer (FLNC) reads from each of the 12**

597   **samples after demultiplexing.**

598

599   **S3 Table. Mapping high-quality (HQ) transcript sequences to the genome and filtering**

600   **criteria**.

601   HQ sequences were mapped to B73 v4 genome and filtered for 99% coverage and 95% identity.

602   Redundant transcripts are collapsed.

603

604   **S4 Table. Removal of library artifacts using a modified version of the SQANTI software.**

605

606   **S5 Table. Top BLASTN hit counts of the unmapped high-quality (HQ) transcript**

607   **sequences to the NR database.**

608   BLASTN was run with a report best hit with E-value cutoff 0.1, 1669 of 1744 HQ sequences had

609   a BLASTN hit.

610

611   **S6 Table. Sample-specific transcript counts.**

612   Using the demultiplexed full-length reads, we assigned Iso-Seq transcripts back to each sample.

613   If a transcript contained at least one full-length read from a sample, it was considered to be

614   expressed.

S1 Fig. Phenotype of maize B73, Ki11, and two reciprocal hybrids (B73 × Ki11, Ki11 × B73) at 14DAG and measure of different traits between parents and the hybrids.

(A) Shoot and root phenotype of B73, Ki11, and the two reciprocal hybrids. (B) Seed phenotype of B73, Ki11, and the two hybrids. (C) Plant height, (D) primary root length, (E) lateral root number, (F–G) biomass, and (H) 100-kernel weight of the parents and hybrids. ** $p < 0.01$.

S2 Fig. Length distribution of the final transcript set.

After mapping the high-quality (HQ) sequences to the B73 v4 genome and filtering for coverage, identity, and running the SQANTI software to remove library artifacts, we obtained 75,118 transcripts. Min: 80 bp, Max: 11,495 bp, Mean: 2,482 bp, 5th–95th percentile, 363–4,975 bp.

S3 Fig. Sample-specific transcript length distribution.

Using the demultiplexed full-length reads, we assigned Iso-Seq transcripts back to each sample. If a transcript contains at least one full-length read from a sample, it is considered expressed.

S4 Fig. Rarefaction curves against known genes and transcripts (A−B) by strain, (C−D) by tissue.

For each subpanel, the X-axis shows the number of subsampled full-length reads and the Y-axis shows the number of observed unique genes or transcripts. Both known and novel genes/transcripts are considered.

S5 Fig. Number of genes and isoforms between parents and two reciprocal hybrids in embryo, endosperm, and root.

S6 Fig. Overlap of genes and isoforms among parents and two hybrids in embryo, endosperm, and root.

S7 Fig. Alternative splicing pattern between parents and two reciprocal hybrids in embryo (A), endosperm (B), and root (C).

S8 Fig. Number of additive and non-additive expression genes in embryo (A), endosperm (B), and root (C).

S9 Fig. Number of isoforms of additive and non-additive expression genes in embryo (A), endosperm (B) and root (C).

S10 Fig. IsoPhase phasing example. (A) The gene PB.21897 (Zm00001d045657) phased by IsoPhase. SNPs are depicted between the B73 and Ki11 track; seven SNPs are shown. SNP #2–#7 were called based on both long and short read data (purple), SNP #1 was missed by long read due to reduced coverage (blue); suspicious short read-only SNPs are marked in red. (B) Zoomed-in region between SNP #6 and SNP #7 showing suspicious SNPs (red) called based on short read data only. Top track shows Ki11 Iso-Seq FL reads; bottom track shows Ki11 short reads.

S11 Fig. Expression and number of isoforms of mono-allelic genes in different tissues.
(A) Allelic expression of mono-allelic genes in reciprocal hybrids, and (B) number of isoforms between parents and hybrids of mono-allelic genes. p0 represents the B73 allele; p1 represents the Ki11 allele.

S12 Fig. Distribution of cis-, trans-regulated genes, and number of isoforms of each category.

(A) Distribution of different cis-, trans-regulated genes. (B) Number of isoforms in each category of cis-, trans-regulated genes.
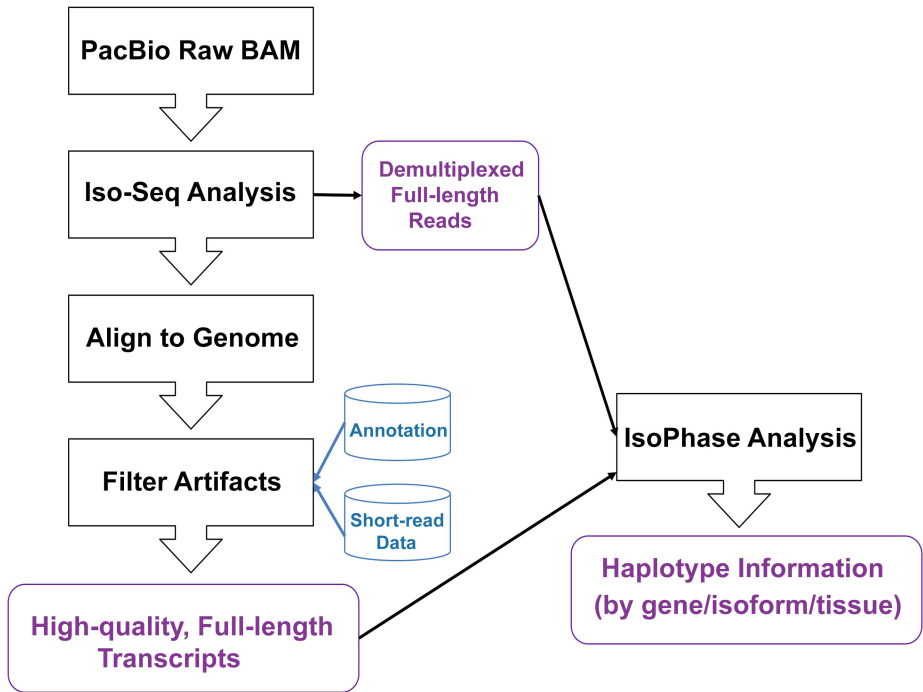
A Isoform distribution across structural categories
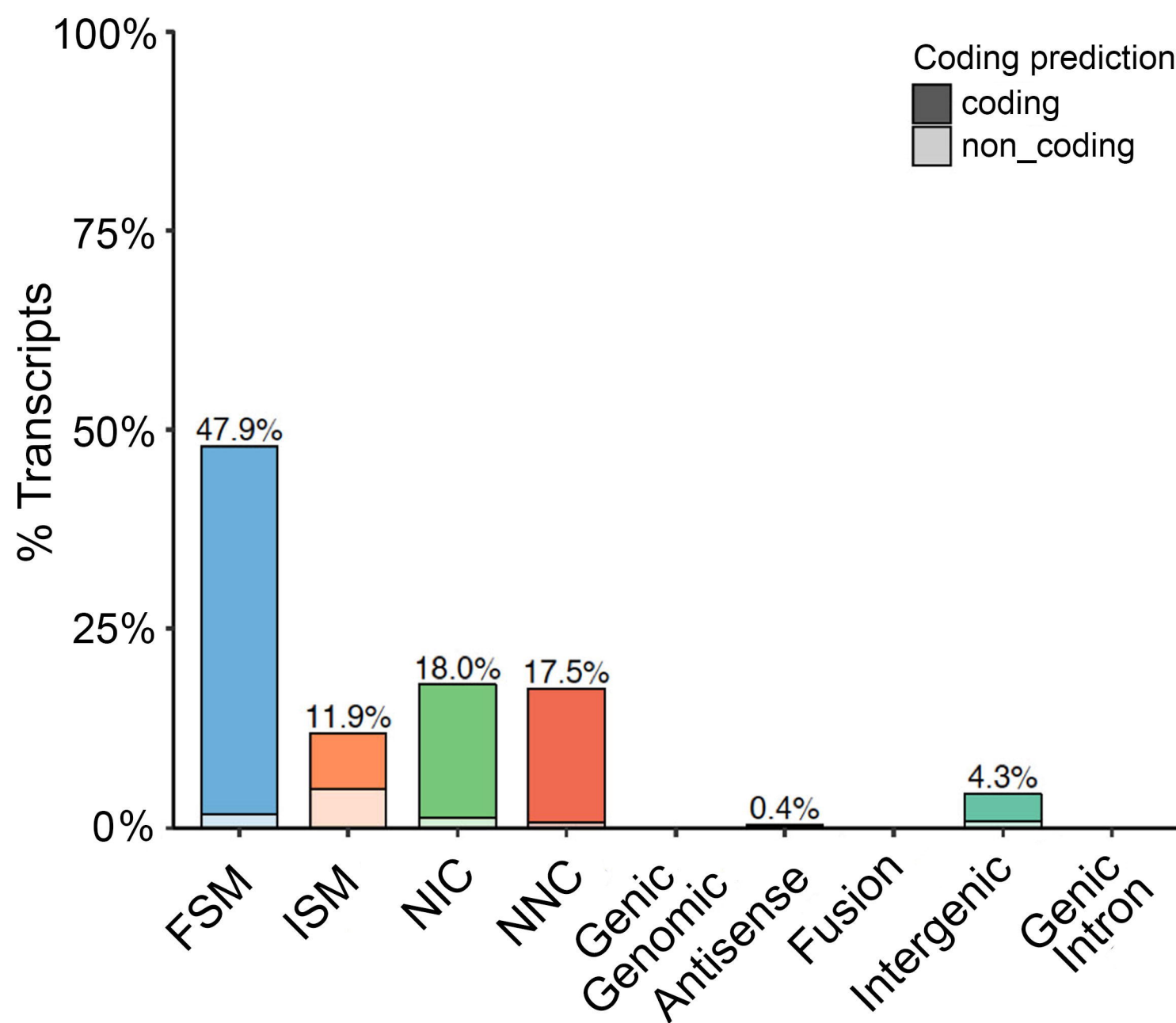
B Classification by transcript length, normalized
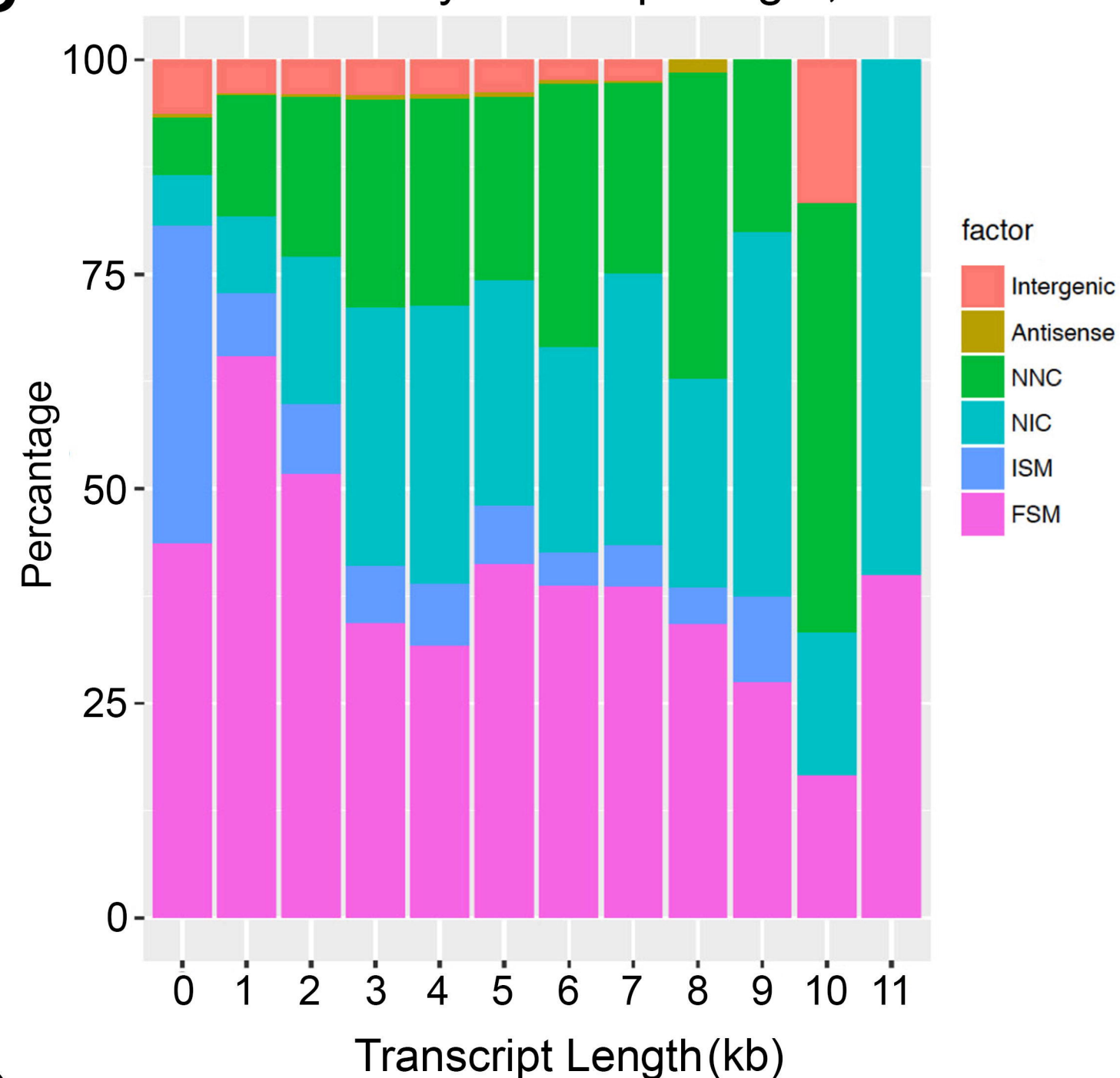
C Transcript Lengths, Mono- vs Multi-Exons
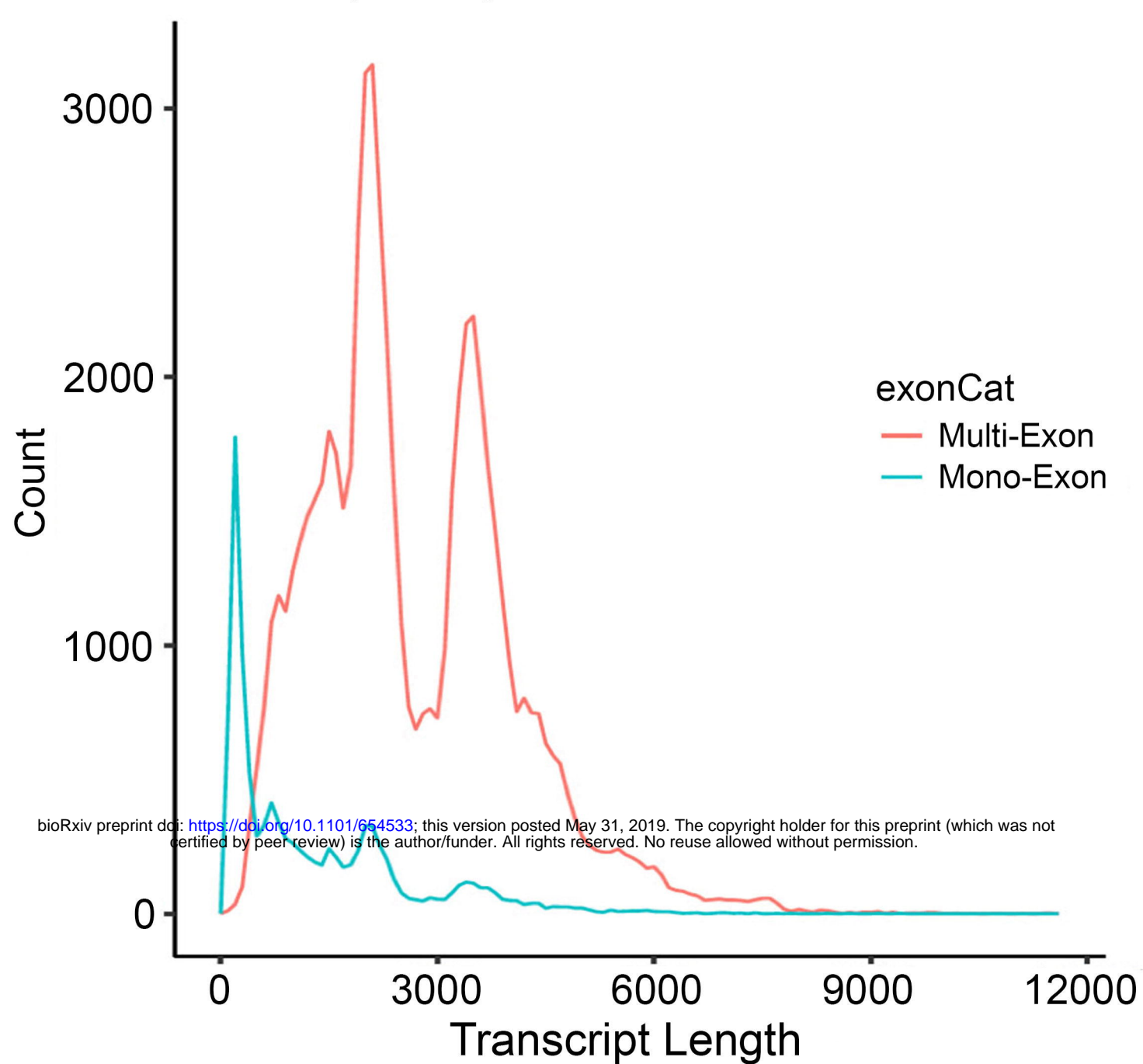
D Distribution of Mono- vs Multi-Exon Transcripts
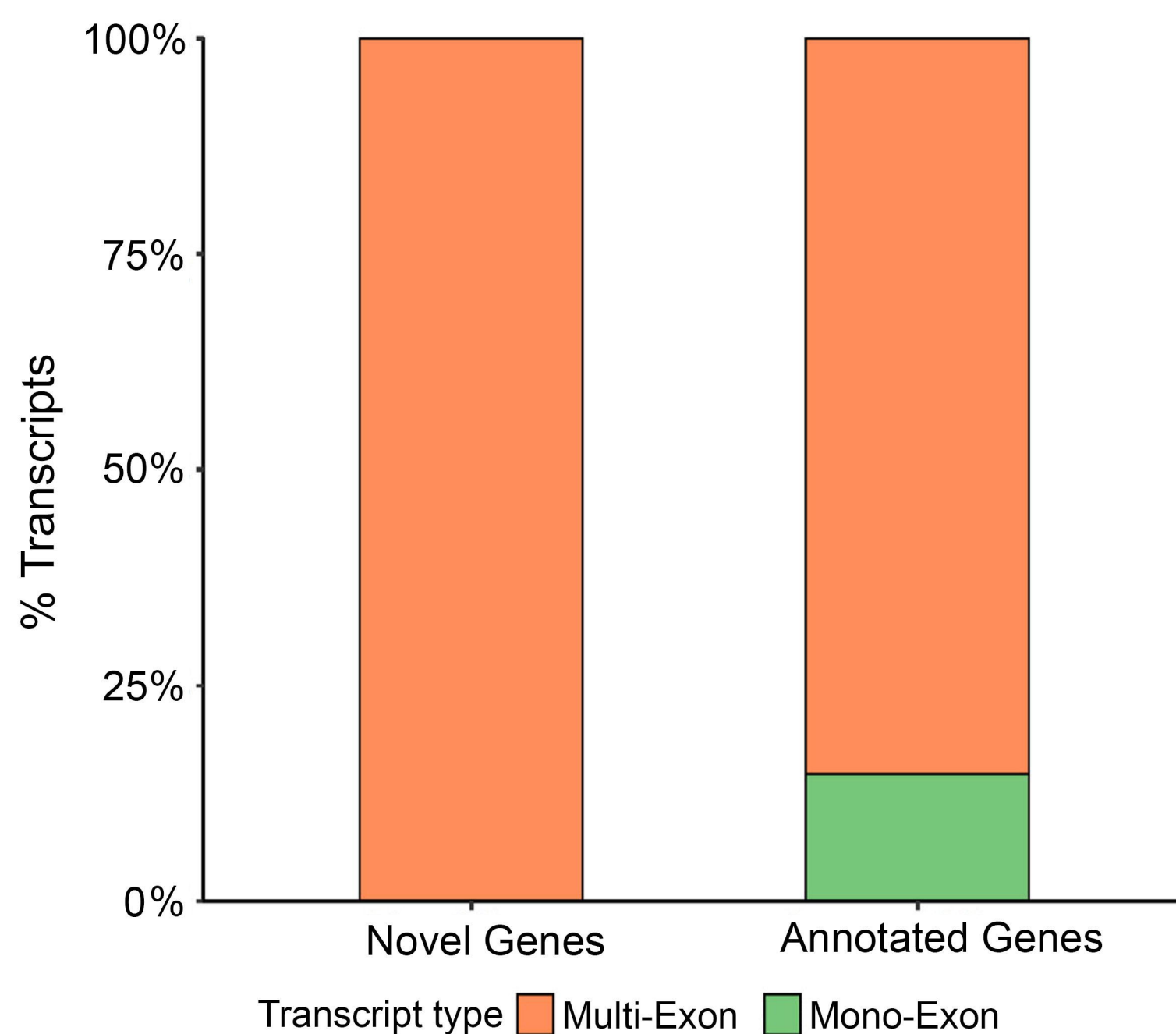
E
- distance: upstream of polyA site
+ distance: downstream of polyA site
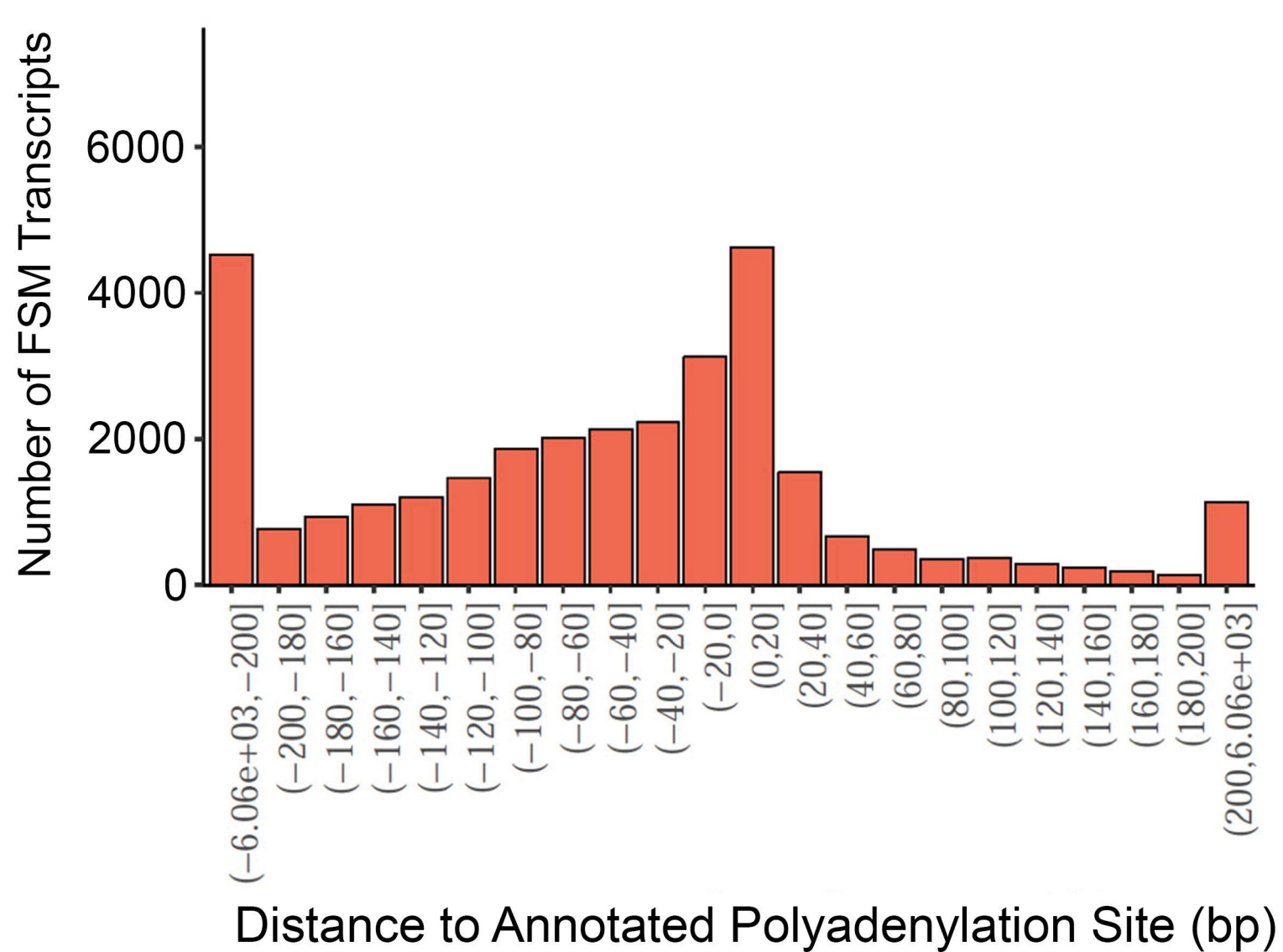
F
- distance: downstream of TSS
+ distance: upstream of TSS
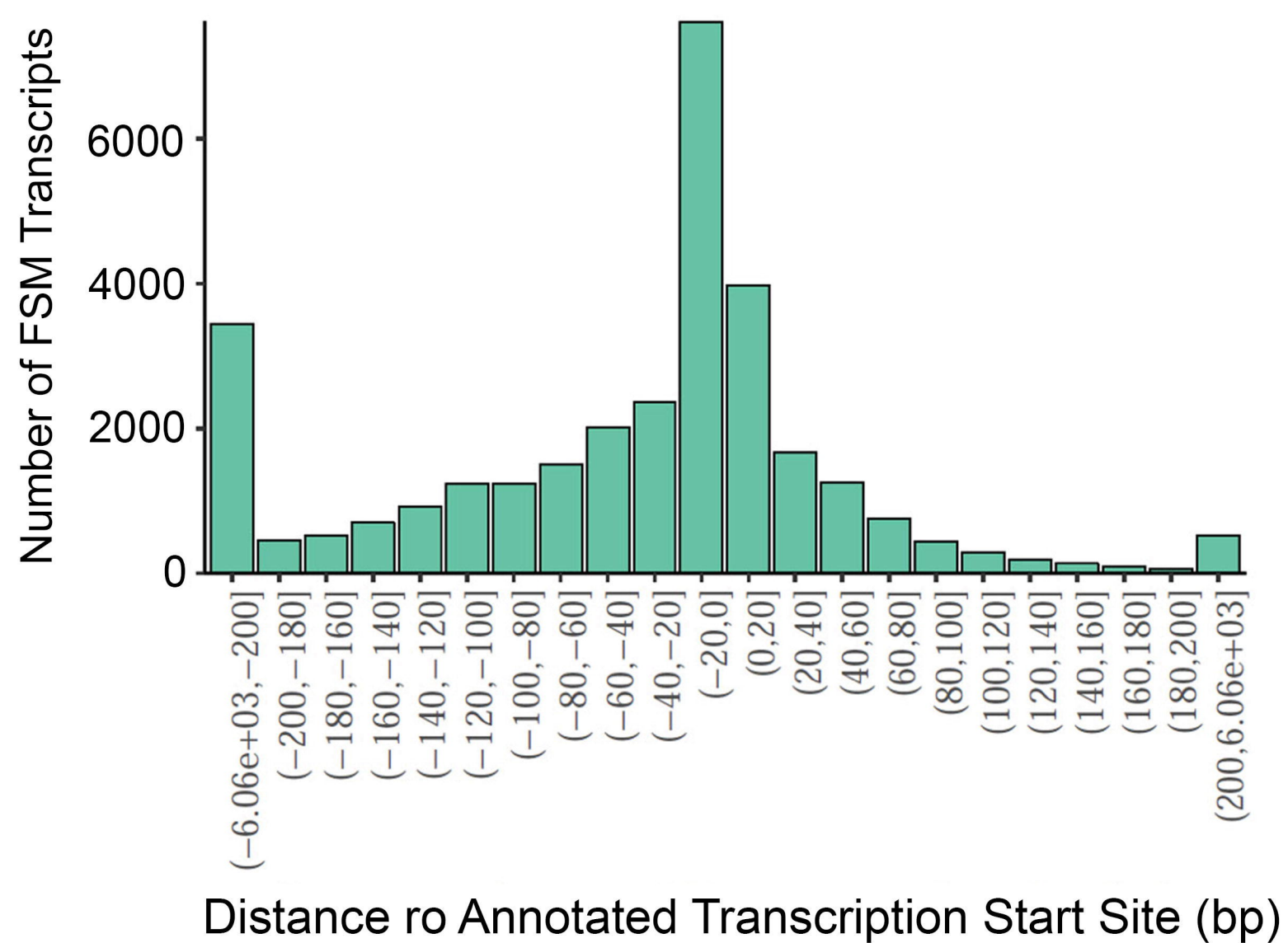
A

ALIGNMENT
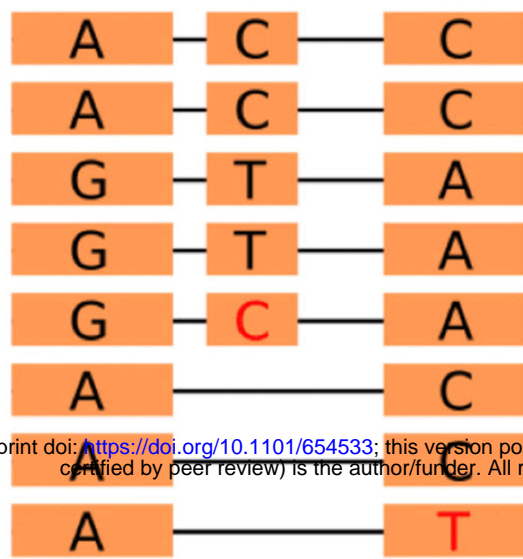
SNP CALLING

Position	SNPs
POS1	A, G
POS2	C, T
POS3	C, A

PHASING
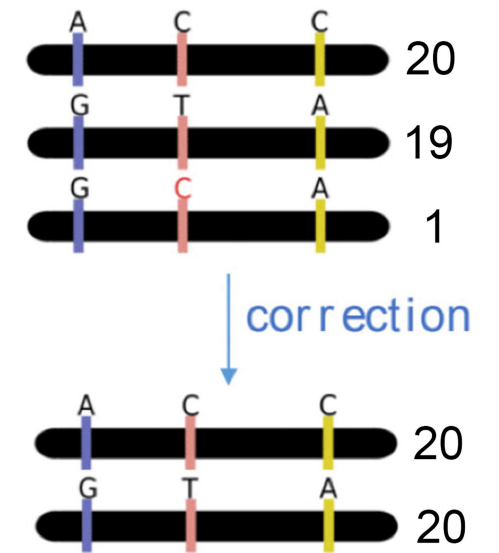
correction

VCF OUTPUT

```
##fileformat=VCFv4.2
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO            FORMAT  ISOFORM1    ISOFORM2
chr1    105 .   A   G   .       PASS    DP=40;AF=0.50   GT:HQ   0|1:20,20   0:15
chr1    190 .   C   T   .       PASS    DP=40;AF=0.50   GT:HQ   0|1:20,20   0:15
chr1    336 .   C   A   .       PASS    DP=40;AF=0.50   GT:HQ   0|1:20,20   0:15
```

B

Pooled Reads

B73 FL reads
Ki11 FL reads
B73 x Ki11 FL reads
Ki11 x B73 FL reads

IsoPhase

Allele and Per-sample Read Counts

	B73	Ki11	B73xKi11	Ki11xB73
A C C	30	0	15	15
G T A	0	20	15	3

Identify B73-only allele as phase0
Identify Ki11-only allele as phase1

Final Results

Phase 0 (B73) : A-C-C

Phase 1 (Ki11) : G-T-A

B73 x Ki11 : 15, 15

Ki11 x B73 : 15, 3

IsoPhase is run individually for each gene
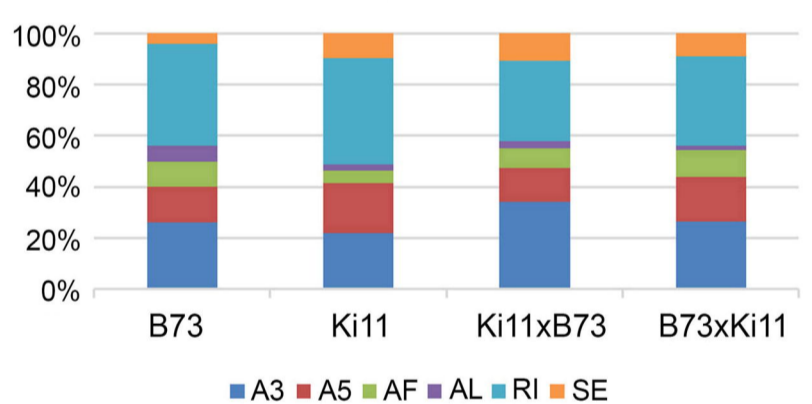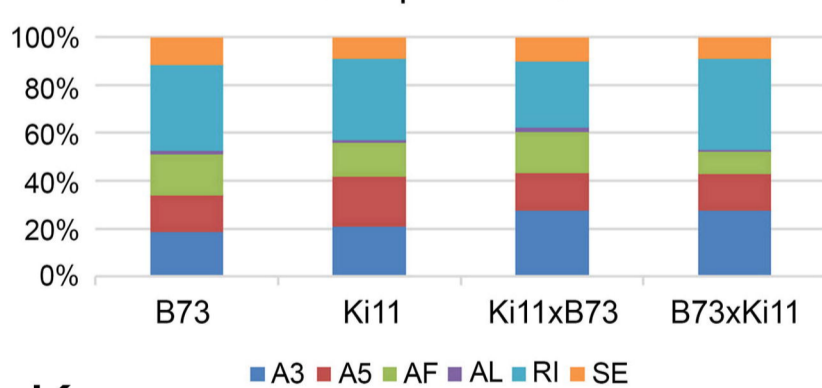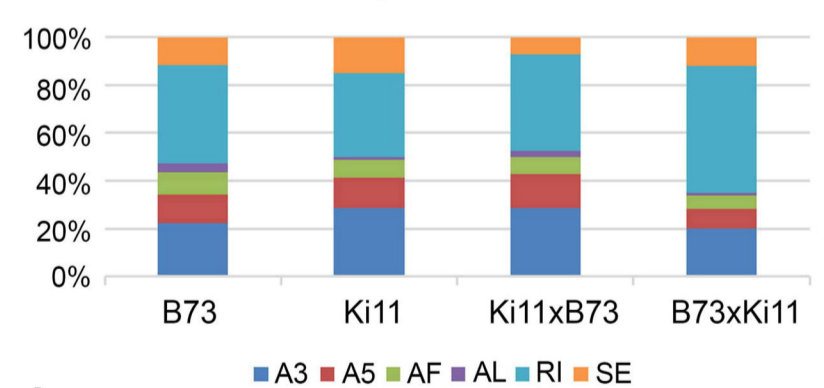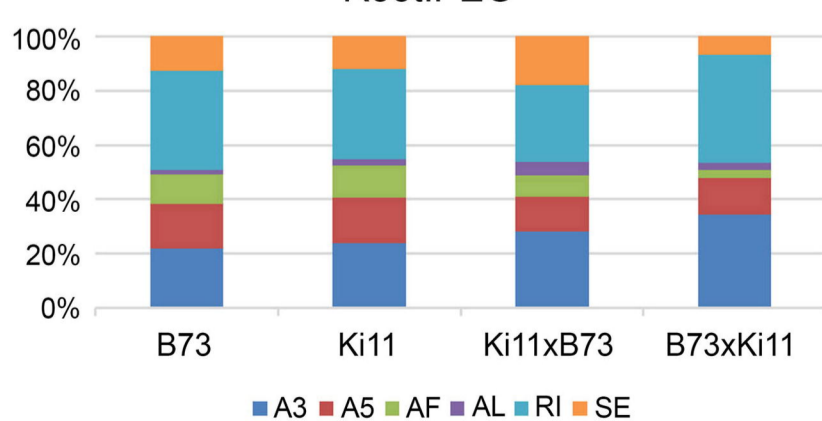
A chr4
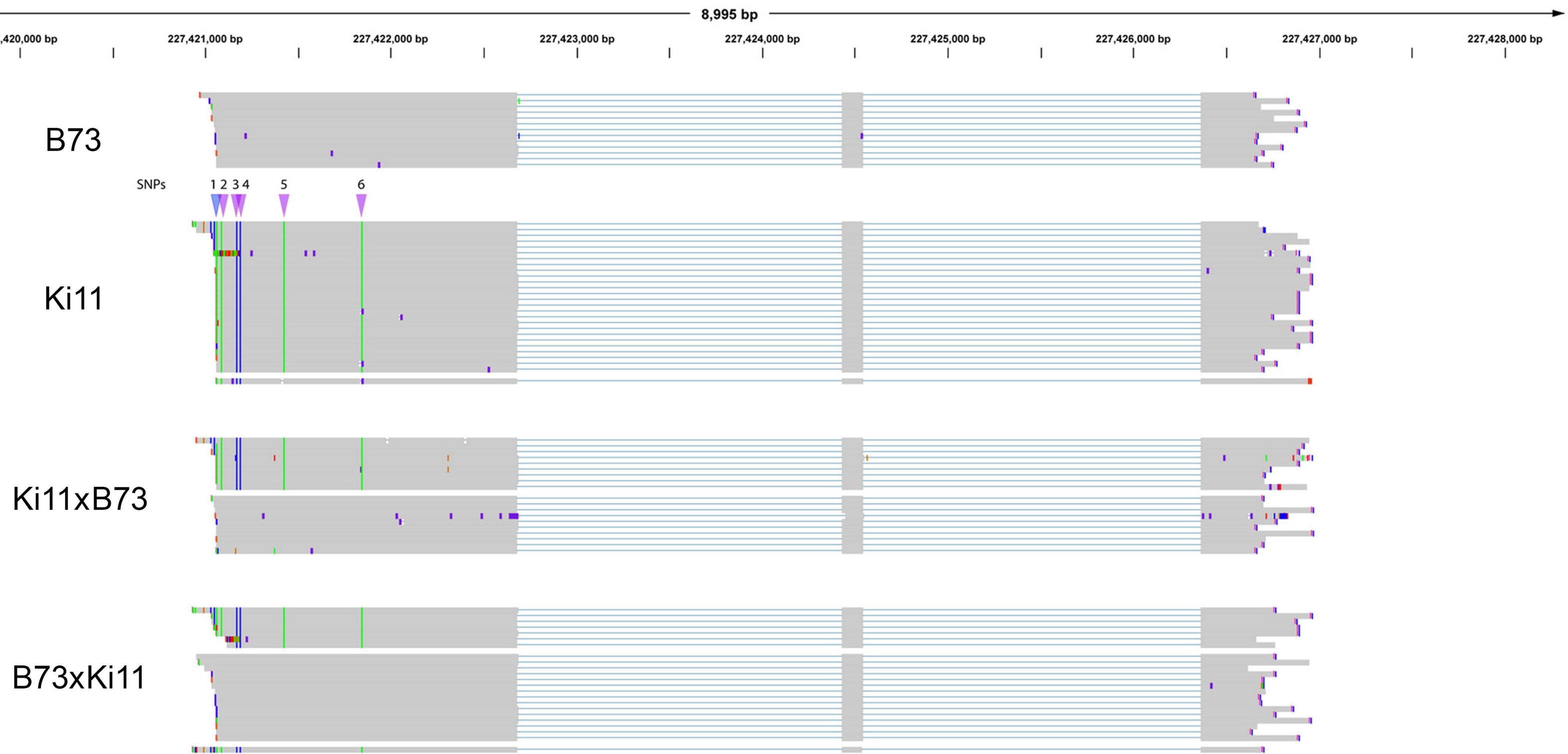
B73

Ki11

Ki11xB73

B73xKi11

B chr6

B73
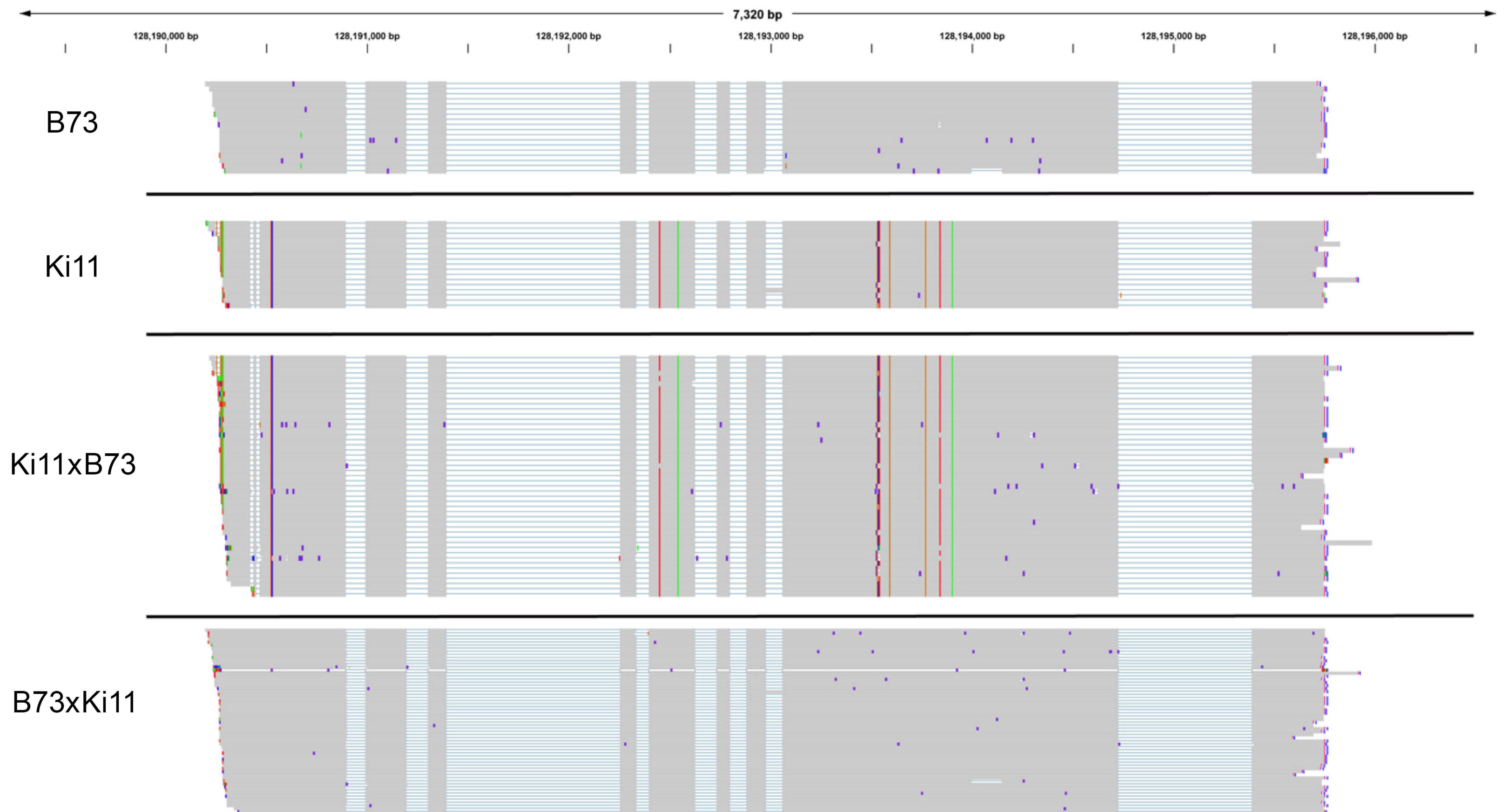
Ki11

Ki11xB73

B73xKi11
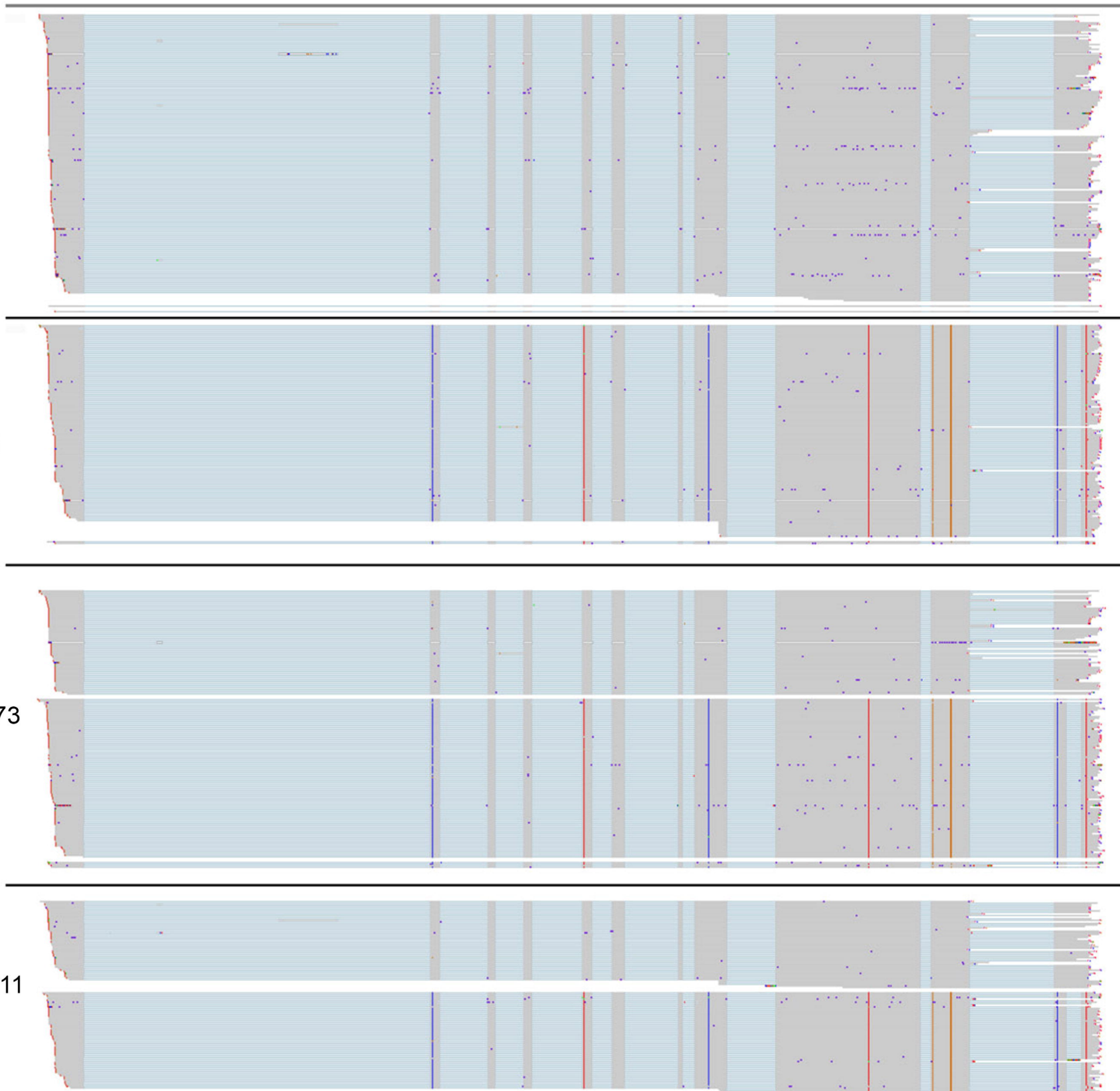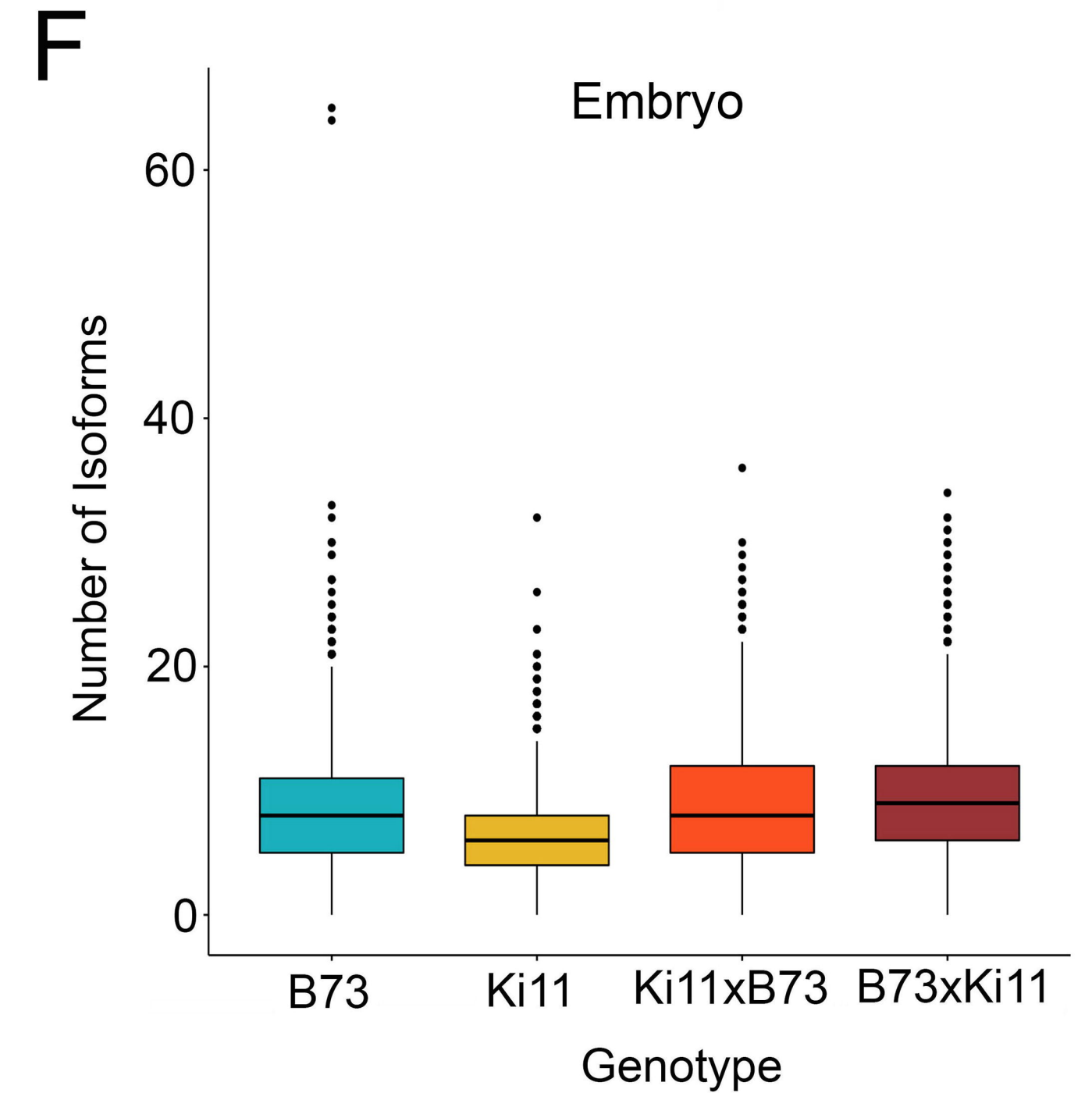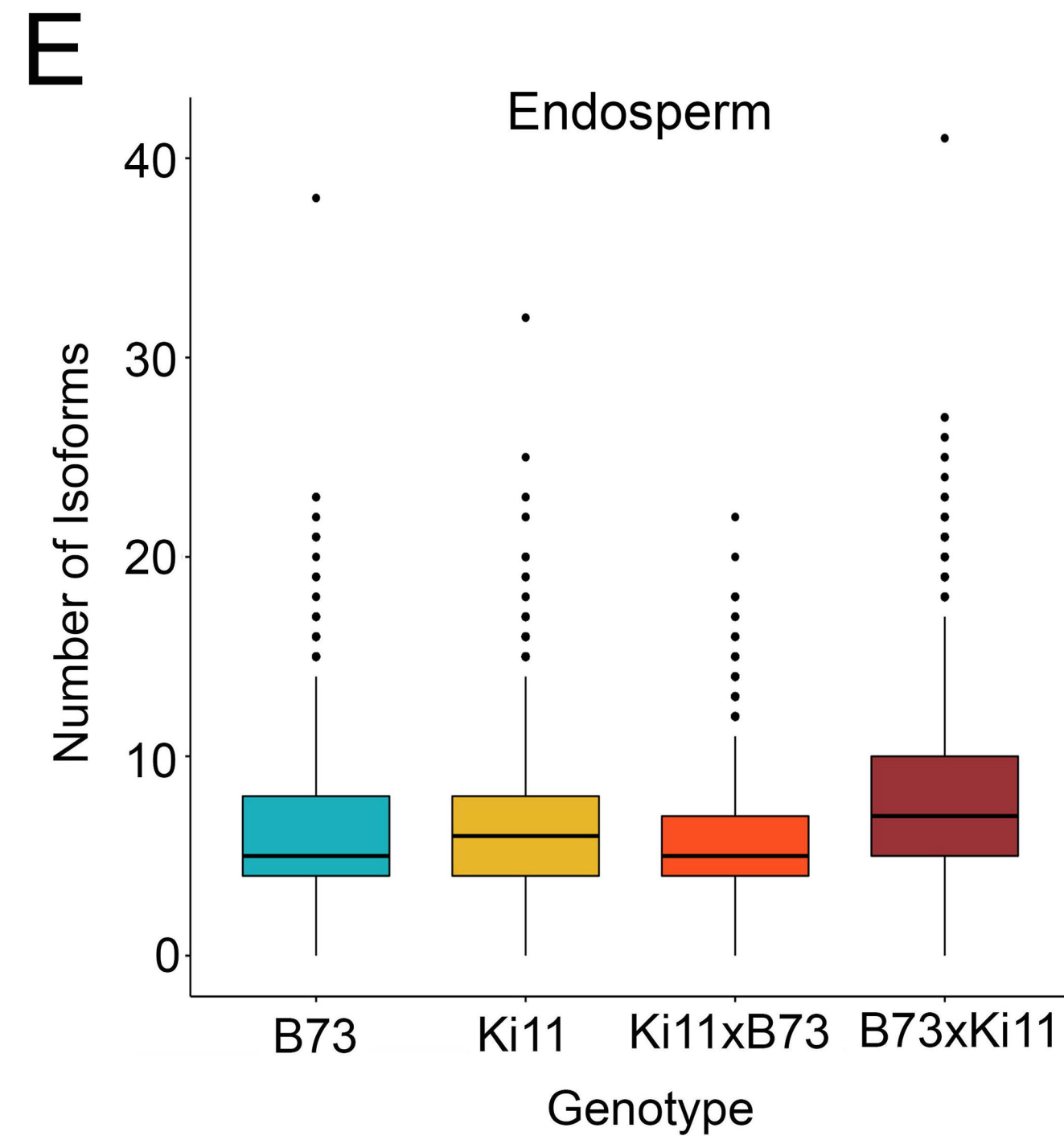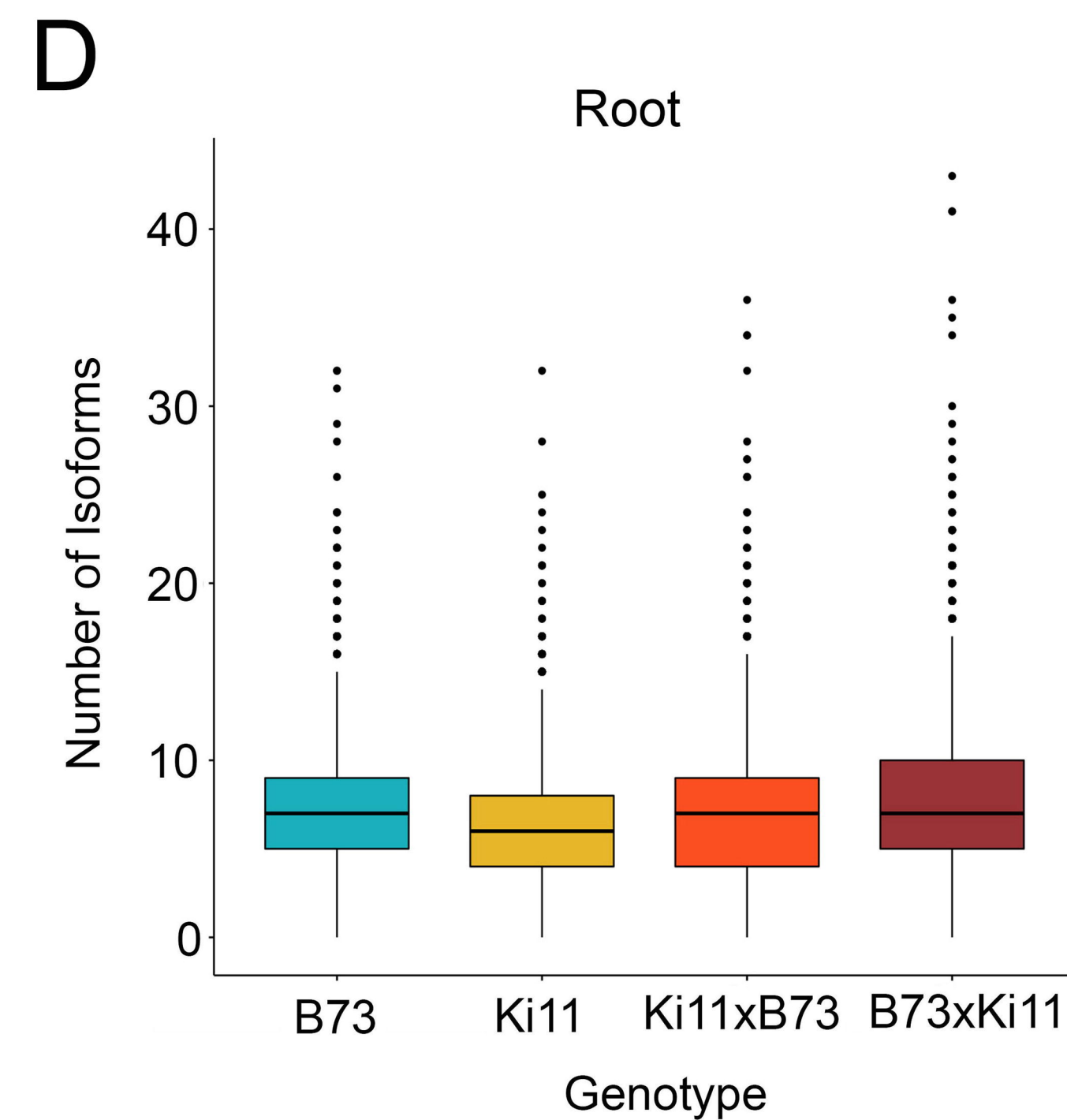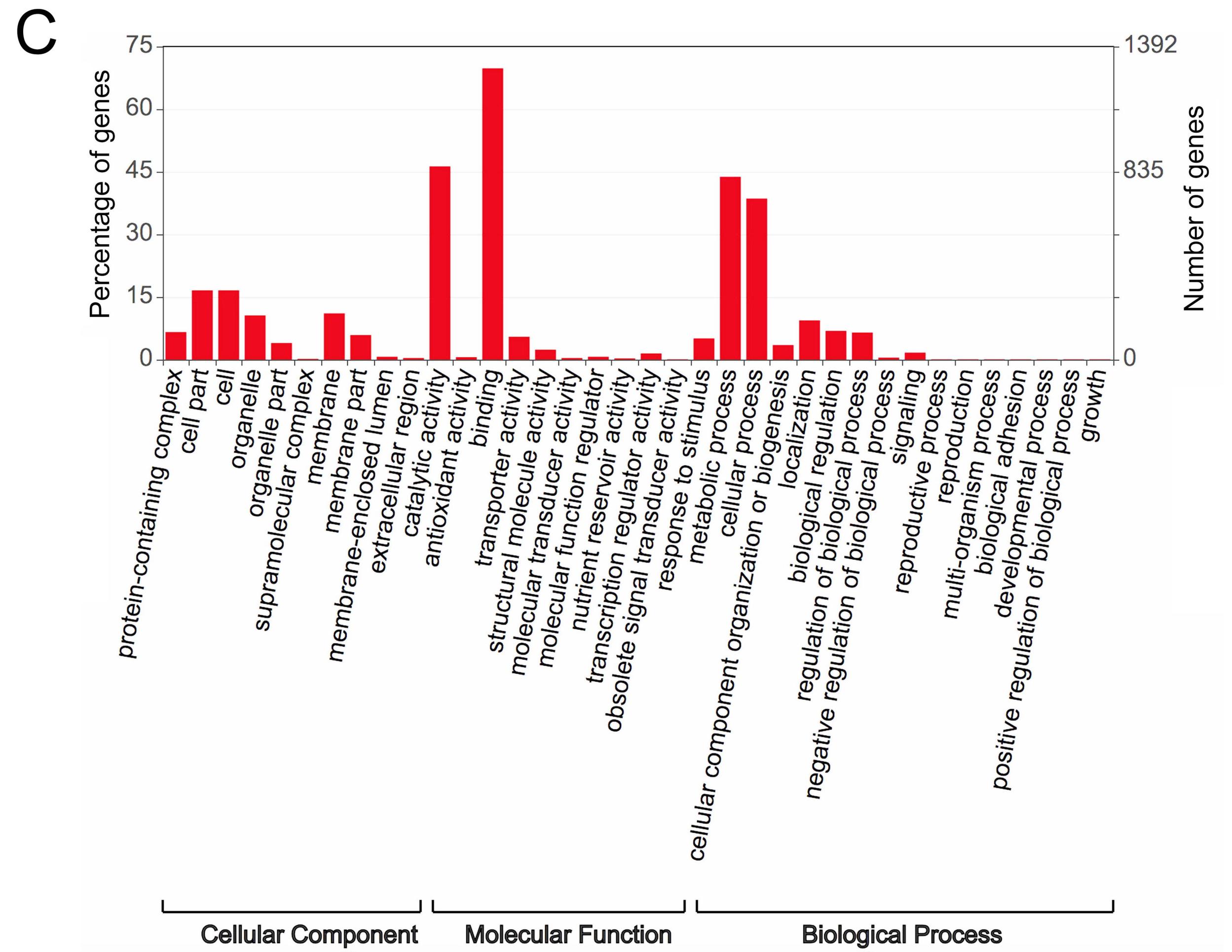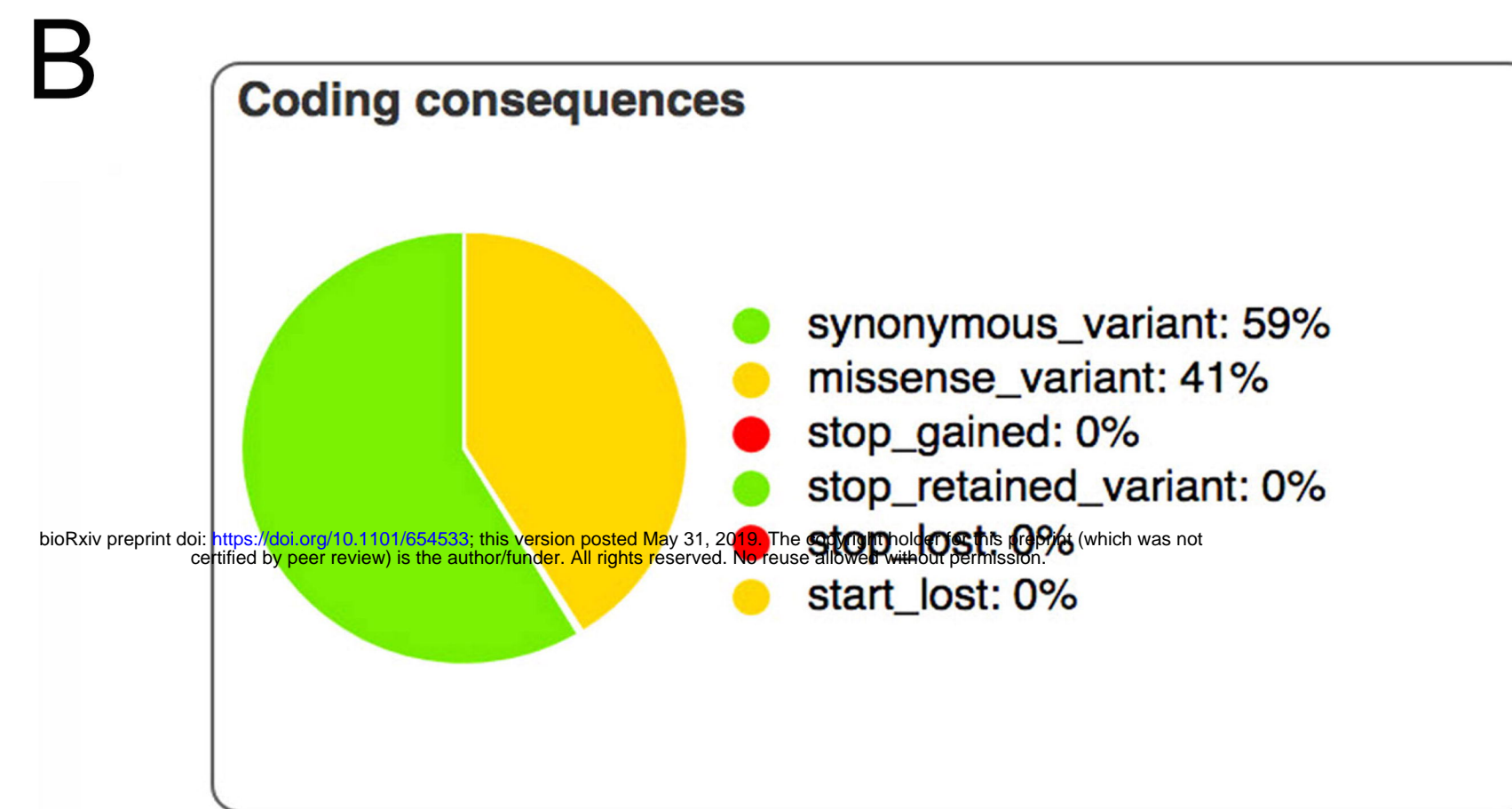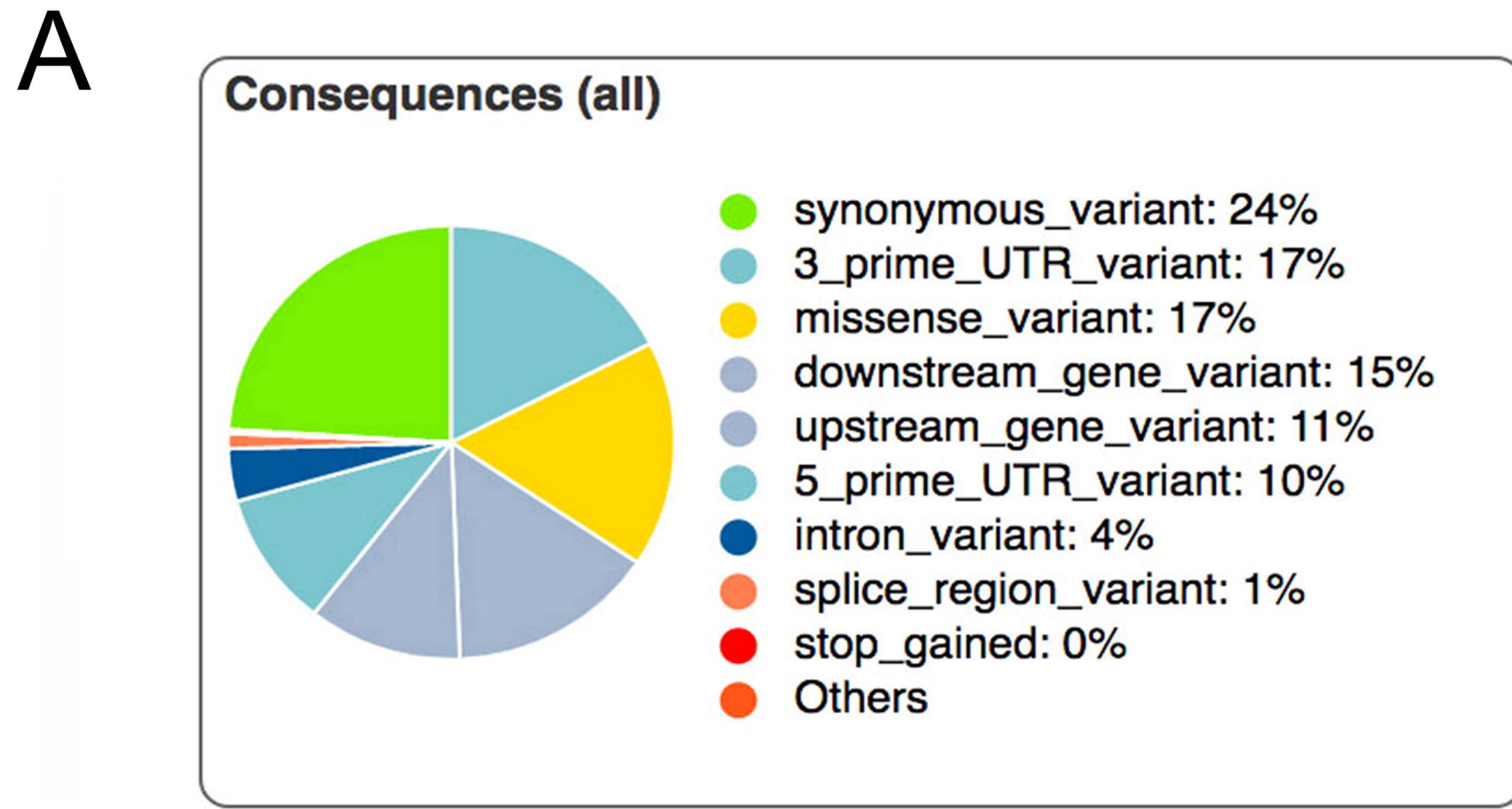
A Imprinted.embryoMEG.allele
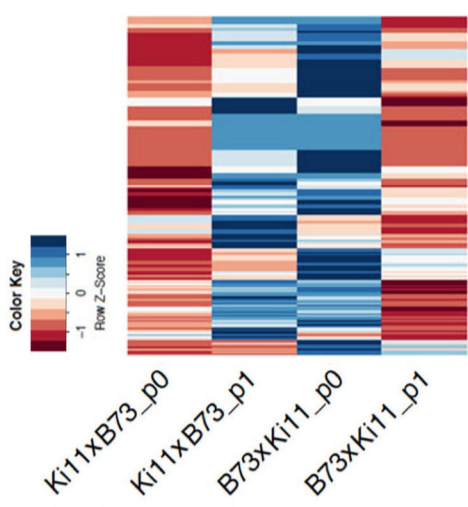
B Imprinted.endospermMEG.allele
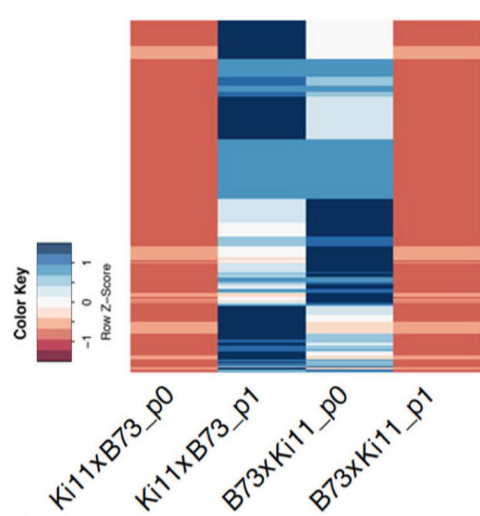
C Imprinted.rootMEG.allele
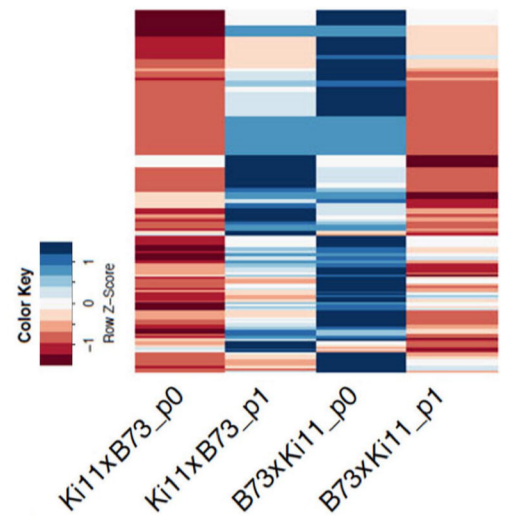
D Imprinted.embryoPEG.allele
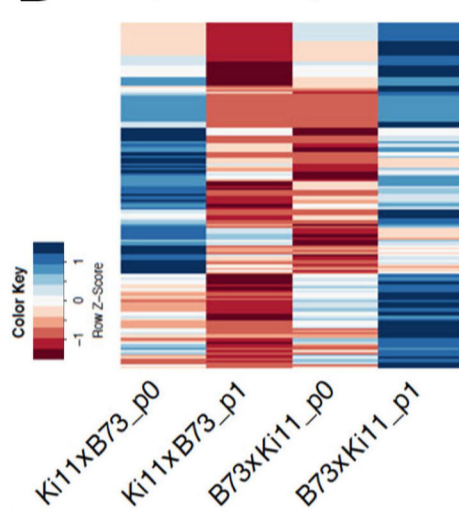
E Imprinted.endospermPEG.allele
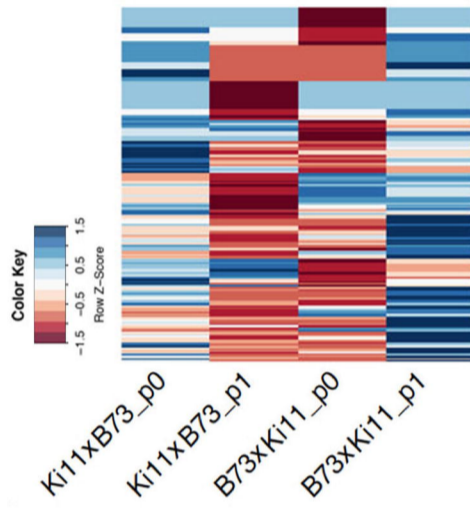
F Imprinted.rootPEG.allele
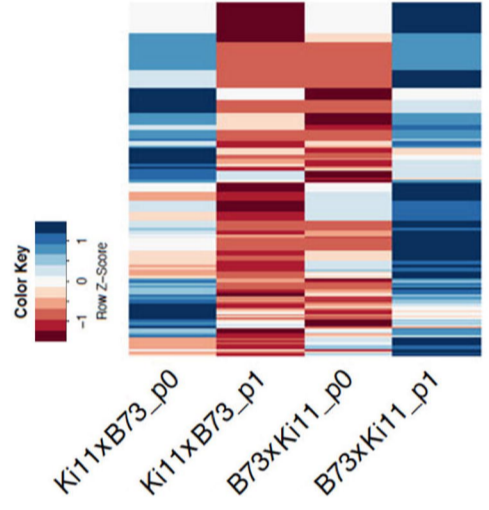
G Embryo.PEG
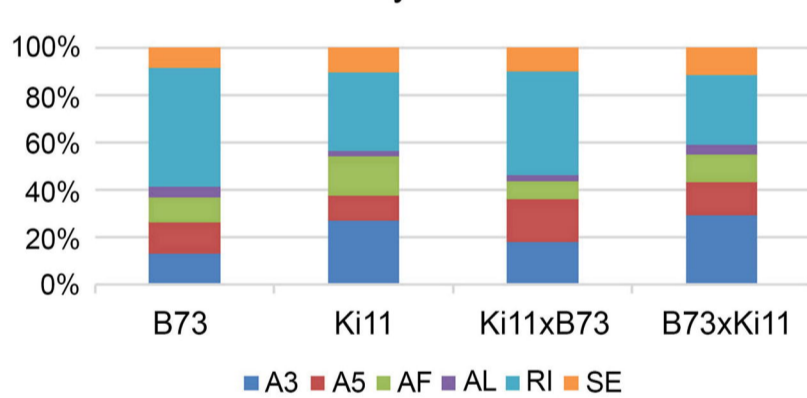
H Embryo.MEG

I Endosperm.PEG

J Endosperm.MEG

K Root.PEG

L Root.MEG

A3 A5 AF AL RI SE