Improving isoform-level eQTL and integrative genetic analyses of breast cancer risk with long-read RNA transcript assemblies

S. Taylor Head[1]*, Aryun Nemani[2], Yung-Han Chang[3], Tabitha A. Harrison[4], Sean T. Bresnahan[1], Joseph H. Rothstein[1], Weiva Sieh[1], Sara Lindström[4,5], Arjun Bhattacharya[1,6]

[1]Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

[2]Bioscience Program, Rice University, Houston, TX, USA[1]

[3]Quantitative Sciences Program, University of Texas MD Anderson Cancer Center, UTHealth Houston Graduate School of Biomedical Sciences, Houston, TX, USA

[4]Department of Epidemiology, University of Washington, Seattle, WA, USA

[5]Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA

[6]Institute for Data Science in Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

*Corresponding author: sthead@mdanderson.org

ABSTRACT

Most eQTL and TWAS analyses quantify expression using aggregate, tissue-agnostic transcript annotations and ignore isoform-level regulation, potentially obscuring or misattributing regulatory mechanisms. Here, we developed a framework leveraging publicly available long-read RNA-seq data to perform tissue-informed inference of genetic regulation and prioritize candidate causal isoforms for breast cancer risk. We quantified gene- and isoform-level expression in breast tumor (TCGA), non-cancerous mammary tissue, and cultured fibroblasts (GTEx) using three transcriptome annotations: standard GENCODE, tissue-specific long-read-derived assemblies, and combined annotations incorporating transcript-isoforms from both. While GENCODE cataloged over 250,000 pan-tissue isoforms, the tissue-specific long-read assemblies captured reduced sets of 74,717 isoforms in tumor, 48,057 in fibroblasts, and 22,941 in healthy breast. We performed eQTL mapping and fine-mapping, followed by colocalization with overall and subtype-specific breast cancer GWAS and isoform-level TWAS. While most eGenes were concordant across annotations, approximately 1/3 of lead cis-eQTLs for shared eGenes differed between long-read assemblies and GENCODE. Further, eIsoform discovery was highly annotation-specific. In healthy breast tissue, the gold standard tissue for building gene expression prediction models for TWAS of breast cancer, 46% of eIsoforms identified by the long-read annotation were unique to that annotation even though 93.7% of them are present in GENCODE. Despite combined annotations expanding the GENCODE catalog by only 0.6-7.6% depending on tissue source, 69% of unique significant isoform-trait associations were specific to a single annotation. Long-read-informed annotations uncovered regulatory associations entirely missed by GENCODE, including a candidate regulatory isoform at the *MARK1* locus captured only in fibroblasts and a previously unannotated splice variant prioritized as the likely effector transcript at *NUP107*. These findings demonstrate that transcript annotation is not merely a technical consideration but critically defines the biological hypothesis space for

41 regulatory mechanisms and shapes discovery. Incorporating tissue-resolved isoform
42 annotations from long-read RNA-seq improves the specificity of regulatory inference and
43 enhances identification of candidate causal isoforms at GWAS loci.

INTRODUCTION

Identifying the likely transcriptomic mechanisms affected by the >200 genetic variants linked to breast cancer risk remains an open challenge[1–7]. While integrative approaches such as colocalization and transcriptome-wide association studies (TWAS) help identify causal genes at breast cancer genome-wide association study (GWAS) loci, these approaches often rely on total gene expression aggregated across transcript isoforms, potentially obscuring isoform-specific regulatory effects. Indeed, modeling isoforms rather than genes has been found to increase the discovery of transcriptome-mediated genetic associations with complex traits by as much as 60%[8–10]. Ignoring isoforms is a notable limitation in breast cancer, where isoform-specific antagonistic effects are well-documented in genes such as *BRCA1*, *ERBB2*, *KLF6*, *AIB1*, *TNC37-42*, and *ESR1*[11–16].

Both gene- and isoform-level integrative analyses depend upon accurate quantification of expression. Aggregate transcriptome references like GENCODE, which are often used for short-read RNA-seq quantification, can obscure the tissue-specific regulatory mechanisms underlying tumorigenesis. They can contain hundreds of thousands of annotated isoforms, many of which are only weakly or not at all expressed in breast tissue[17,18]. Isoform quantification is further complicated by the limitations of short-read RNA-seq approaches that can struggle to accurately resolve transcript isoforms because of read-to-transcript mapping ambiguity[19]. When sequenced fragments align to exons shared by multiple isoforms, which is a frequent occurrence in genes with complex splicing patterns, probabilistic assignment of reads generates substantial uncertainty in isoform-level expression estimates[20], leading to reduced statistical power in downstream integrated analyses[19]. However, recent technological advances in long-read RNA-seq have enabled more comprehensive detection of full-length transcript isoforms. While short-read sequencing typically produces reads under 300 bp in length, long-read RNA sequencing

technologies can generate reads up to 15-20 kilobases that span entire transcripts. This offers direct insight into exon structure and a more comprehensive view of transcript isoform diversity within a tissue or cell type[21–25].

In this study, we leveraged long-read RNA-seq-derived transcriptome references for breast-relevant tissues to deliver an improved, tissue-specific view of genetic regulation at the isoform level across non-cancerous mammary tissue, malignant breast tumors, and cultured fibroblasts. We compared gene- and isoform-level expression quantitative trait loci (eQTL) mapping, colocalization, and TWAS for overall and subtype-specific breast cancer across multiple annotations and tissues. Our analyses yielded several major findings. First, although the long-read transcriptome assemblies are highly specific and contain >100-200K fewer transcripts than GENCODE, they retained comparable power for mapping GWAS loci with detected eQTLs, indicating robust coverage of risk-associated loci. Second, the long-read-informed analyses revealed multiple novel isoform-level regulatory mechanisms that are undetectable with standard reference transcriptomes. Third, we demonstrated that colocalization and TWAS results are highly sensitive to even minor changes in the transcriptome assembly used to quantify short-read RNA-seq samples, underscoring the importance of reference selection for downstream analyses. Finally, we provide an initial catalog of germline genetic effects on isoform-level expression across tissue states that can serve as a resource for future studies of breast cancer.

RESULTS

In this study, we re-quantified short-read RNA-seq data from non-cancerous breast tissue, breast tumors, and fibroblasts using long-read derived transcriptome assemblies. We then systematically compared *cis*-eQTL mapping, colocalization of eQTL and breast cancer GWAS signals, and isoform-level TWAS results against those obtained using GENCODEv45 and a combined annotation (Fig. 1a). Despite containing 3-11 times fewer transcripts than

92 GENCODE, long-read assemblies recovered comparable numbers of eQTL-linked independent
93 GWAS loci. However, they yielded markedly fewer colocalized loci and TWAS associations,
94 potentially reflecting tissue specificity of causal regulatory mechanisms and reduced spurious
95 findings rather than diminished power. Aggregate annotations like GENCODE can dilute signal
96 across structurally similar transcripts and misattribute reads to transcripts that are lowly or not
97 expressed in the target tissue, elevating quantification uncertainty and producing unstable
98 colocalization or TWAS results (Fig. 1b). Here, we show how long-read assemblies can help
99 resolve isoform-level ambiguity at known GWAS loci, uncover novel isoforms mediating genetic
100 risk, and inform biological interpretation of potential disease mechanisms in the target tissue
101 context.

102 **Long-read transcriptome assemblies greatly refine gene and isoform space**

103 We applied a robust transcriptome assembly and quantification framework[26,27] to jointly
104 discover and quantify expression of transcript isoforms in three sets of breast cancer-relevant
105 long-read RNA-seq libraries (Fig. 1a). Compared to the GENCODE reference transcriptome
106 containing 252,931 transcript isoforms across 63,187 genes[28], these *de novo* long-read
107 assemblies yielded markedly lower numbers of annotated isoforms and genes, both before and
108 following comprehensive quality control protocols. Across the three tissues, isoform counts from
109 long-read assemblies decreased over 70% relative to GENCODE, and this corresponded to a
110 decrease in gene counts by over 63% (Fig. 2a). We resolved the greatest number of high-
111 confidence isoforms (74,717) in our malignant breast tumor samples (N=26; primary biopsies,
112 patient-derived xenograft tumors, and cell lines) mapping to 18,100 genes. The healthy breast
113 samples (N=5; primary mammary tissue and cell lines) demonstrated the most pronounced
114 transcriptome refinement, with only 22,941 high-confidence isoforms assembled across 11,015
115 genes.

116 Given the limited number of healthy breast libraries, cell-cultured fibroblasts provided an
117 additional non-malignant and non-epithelial reference tissue with a long-read sample size
118 (N=22) comparable to breast tumor. We observed filtering of approximately 91% of isoforms in
119 both healthy breast and tumor. However, in fibroblasts, we observed the most stringent quality
120 control filtering of long-read isoform calls, with 48% (48,057 isoforms of 100,190 assembled)
121 mapped to a total of 12,990 genes retained as high-confidence. 92.3% of all filtered isoforms
122 lacked sufficient splice junction coverage in short-reads or full-length reads (Fig. S1, Table S1).
123 While we consistently detected 8,469 shared genes and 12,775 shared isoforms across
124 all three long-read assemblies (Fig. 2b), 58% of genes and 87% of long-read isoforms displayed
125 some degree of tissue-specificity. Differences in detected and retained tissue-specific
126 transcriptome sizes likely reflect a combination of tissue-specific alternative splicing programs,
127 long-read sample size, and heterogeneity of long-read sample type. In particular, there are 5,048
128 genes and 45,293 isoforms unique to the breast tumor assembly, comprising 25% and 44% of
129 all assembled features (Fig. 2b). For ubiquitous genes, the number of isoforms per gene was
130 substantially lower than in GENCODE (median 9) across all assemblies, but the tumor assembly
131 again retained the highest number (median 4; Fig. 2c). Overall, isoform counts per gene
132 decreased relative to the reference for 78.3%, 81.4%, and 91.1% of shared genes in tumor,
133 fibroblast, and healthy breast long-read assemblies, respectively (Fig. S2). These observations
134 of increased isoform diversity in the tumor assembly are consistent with abnormal regulation of
135 alternative splicing in malignant tumors relative to healthy tissue[29–31].

136 **Long-read assemblies reveal novel isoform structures in breast tumor and fibroblasts**

137 The majority of long-read-derived isoforms are full-splice matches (FSM) with splice
138 junction structure perfectly matching an annotated isoform in the reference GENCODE
139 transcriptome: 68.3%, 72%, and 93.7% in fibroblasts, tumor, and healthy breast, respectively

140 (Fig. 2d). However, there was a non-negligible contribution of novel transcript isoform structures
141 to the tumor and fibroblast assemblies (Table S2). While there were only 1,455 novel isoforms
142 in healthy breast tissue, we assembled 20,933 novel isoform structures in tumor tissue and
143 15,258 in fibroblasts. The vast majority (66.2% and 61.5%) of these were “novel in catalog” (NIC)
144 isoforms, reflecting a new combination of previously annotated splice junctions, and 28-29%
145 were “novel not in catalog” (NNIC) isoforms containing one or more previously unannotated
146 splice donor or acceptor sites[27]. Novel tumor isoforms also contained more exons (median 9
147 [IQR 7]) than their corresponding FSM counterparts (median 6 [IQR 8]; Fig. 2e) and were
148 significantly longer (one-sided t-test P<2.2E-16). While these findings highlight the potential of
149 long-read transcriptome assemblies to capture complex tissue-specific isoform structures
150 currently missing from aggregate references, we note that the higher number of novel isoforms
151 in tumor and fibroblasts may reflect a combination of biological differences and technical factors
152 including varied sample sizes and sample composition (tissue vs. cultured cells).

153 Importantly, using these assemblies as refined transcriptome annotations directly
154 impacted expression quantification in our short-read RNA-seq datasets (N=363 healthy breast
155 mammary tissue, N=483 cell-cultured fibroblasts, N=1,085 breast tumors). We used GENCODE
156 and the *de novo* long-read assemblies to construct three annotation sets per tissue (Fig. 2f):
157 GENCODE isoforms alone, tissue-matched long-read assemblies alone (“LR”), and all
158 GENCODE isoforms plus novel high-confidence non-FSM isoforms identified in the
159 corresponding long-read samples (“combined”). Per-sample correlations of expression
160 estimates across annotation pairs revealed that gene-level quantification is relatively robust to
161 annotation choice, with consistently high correlations across all tissues (Fig. 2g). Isoform-level
162 correlations, however, were markedly more variable. They exhibited heavy-tailed distributions
163 even between minimally divergent catalogs like GENCODE and the combined annotations,
164 despite the latter containing only 0.6-7.6% more isoforms. In practical terms, a low correlation

165 here means that the relative expression of an isoform across samples is not preserved between
166 annotations and suggests isoform-level quantification is highly sensitive to the transcriptome
167 annotation choice, consistent with prior studies[32,33].

168 **Long-read transcriptomes reveal alternative genetic architectures for shared regulatory**
169 **features**

170 To identify regulatory genetic variants that may underlie causal breast cancer pathways,
171 we first performed eQTL mapping in each tissue. We identified 5,827, 2,305, and 1,177 genes
172 in fibroblasts, breast, and breast tumor, respectively, as harboring one or more significant eQTL
173 (eGenes) that replicated across all three annotations (Fig. 3a, Online Data). These common
174 eGenes represented over half (52.9%) of all fibroblast eGenes (identified by any annotation),
175 31.3% of all breast eGenes, but less than 18% of tumor eGenes identified by at least one
176 annotation.

177 As expected, GENCODE and combined annotations showed particularly strong
178 concordance in eGenes, with over 89% of eGenes overlapping between the two in each tissue.
179 More notably, long-read assemblies consistently identified a substantial number of unique
180 features not detected by either reference-inclusive annotation: 1,289 unique eGenes in
181 fibroblasts, 954 in healthy breast, and 1,320 in tumor, accounting for 16.9%, 28.3% and 50.2%
182 of all eGenes identified by long-read in each tissue. Fine-mapped eQTLs yielded similar patterns
183 of cross-tissue and -annotation eGene discovery (Fig. S3).

184 While the proportion of long-read unique eIsoforms identified in tumors and in fibroblasts
185 did not differ considerably from that observed for eGenes (differing only by 4.5-10.8%, Fig. 3a),
186 healthy breast tissue showed a striking departure. Approximately 46.0% of all eIsoforms
187 identified in healthy breast tissue using the long-read annotation were unique hits not similarly
188 identified with GENCODE or combined annotations. This represents a 1.6-fold increase relative

189 to the gene-level rate of long-read-specific signals. Notably, this occurred despite 93.7% of all
190 isoforms in the healthy breast long-read assembly being FSM to previously annotated
191 transcripts.

192 We also observed substantial annotation-specific eIsoform signals for reference-informed
193 annotations. For example, in breast tumor, 34.6% of GENCODE eIsoforms were unique despite
194 all being present in the combined tumor transcriptome annotation file, and this jumped to 43.7%
195 following fine-mapping. These findings further suggest that isoform-level regulatory architecture
196 is far more sensitive to transcriptome annotation choice than gene-level analyses, and while the
197 degree of sensitivity differs across tissues, long-read-specific signals persist across all three
198 tissue contexts. We provide full lists of eGenes and eIsoforms from conditional linear modeling
199 and fine-mapping online (see Online Data).

200 We next compared the underlying regulatory variants driving shared eGenes and
201 eIsoforms across annotation sets. Lead cis-eQTLs for shared genes were highly concordant
202 between the GENCODE and combined annotations (>82.1%) but substantially less-so for
203 comparisons involving the long-read assemblies (49.0-65.3%, Fig. 3b). This pattern persisted at
204 the isoform level in fibroblast and healthy breast tissue, yet concordance dropped markedly in
205 tumor where fewer than 41.5% of shared eIsoforms had matching lead isoQTLs across any
206 annotation pair, including GENCODE-informed.

207 Among the shared regulatory features (eGenes and eIsoforms) with discordant lead
208 eQTLs, linkage disequilibrium (LD) R^2 was often moderate to high (median R^2=0.83) but highly
209 variable (IQR=0.59). This suggests that eQTL discordance was not solely driven by the selection
210 of near-identical proxy variants (Fig. S4). Expanding beyond the single top variant to all
211 significant cis-eQTLs per feature, we saw similar patterns of discordance across annotation sets.
212 (Fig. S5).

While the eGenes and eIsoforms identified above through linear regression modeling have strong genetic signals, standard eQTL mapping does not confidently resolve the underlying causal variants. To address this, we performed Bayesian eQTL fine-mapping to more rigorously compare the sets of regulatory features identified across annotations while more directly incorporating LD structure. These fine-mapping results further reinforced observed discordance. Although the number of credible sets per feature and their sizes were comparable across annotations, the composition of these 95% coverage credible sets differed substantially. Among shared eGenes and eIsoforms identified by both GENCODE and long-read annotations, 20.0-31.2% showed <50% overlap in credible set variants (Fig. S6). Even at the stage of eQTL mapping, annotation choice can lead to materially different inferences about putative causal regulatory variants.

eQTLs in long-read assemblies tag GWAS loci at similar rates to GENCODE

Strikingly, despite the reduced gene and isoform space cataloged by each long-read assembly, long-read-derived eQTLs tagged 332 independent breast cancer GWAS risk loci in tumor and fibroblasts at rates comparable to those observed for the larger GENCODE-inclusive annotations (Fig. 3c). Across annotations, between 41.5-56.6% of all tagged loci were observed in one tissue only, suggesting substantial tissue-specific regulatory architecture.

When aggregating across all tissues, long-read-defined eQTLs tagged >89.2% as many loci as GENCODE or combined annotations. Only in healthy breast tissue did the long-read annotation tag fewer loci overall compared to reference-inclusive annotations. However, the number of loci tagged by long-read eQTLs still exceeded 72.4% of the number tagged using GENCODE for both eGenes and eIsoforms. This is notable, given the breast long-read assembly contains 80% fewer genes and 91% fewer transcripts than GENCODE. These patterns remain

236 largely consistent across different LD thresholds used to define locus tagging and restriction to
237 fine-mapped eQTLs (Fig. S7).

238 With the exception of gene-level eQTLs in breast tumor, where 46.7% of loci tagged by
239 long-read were unique, fewer than 20.0% of loci tagged by long-read were unique across
240 remaining tissues. This indicates that long-read annotations largely implicated the same GWAS
241 loci as reference transcriptomes. However, the specific eGenes and eIsoforms underlying these
242 shared loci differed substantially. Fibroblasts provide an example, as this tissue showed the
243 greatest number of tagged GWAS loci and similar tagging rates across all annotations. Among
244 shared GWAS loci in fibroblasts, eGenes were nearly perfectly concordant between GENCODE
245 and combined annotations, whereas concordance dropped to an average of 72.9% at the
246 isoform level (Fig. S8). Concordance was lower still when comparing long-read to GENCODE,
247 with only 42.1% average overlap in eIsoforms tagging the same shared locus.

248 Together, these results suggest that expansive reference annotations do not necessarily
249 yield expanded insight into trait-relevant regulation. Long-read assembly-derived eQTLs tagged
250 GWAS loci at rates largely similar to those observed with GENCODE. While most of these loci
251 were consistent across annotations, this did not imply concordance in the underlying regulatory
252 features. Instead, the specific eGenes and eIsoforms implicated at shared loci differed
253 depending on the transcriptome annotation used for expression quantification. Signals relevant
254 to breast cancer risk thus appear to be captured well by a subset of tissue-relevant isoforms
255 reliably observed with long-read RNA-seq. These findings suggest that long-read assemblies
256 provide a more parsimonious and tissue-informed representation of the regulatory landscape
257 while retaining the capacity to implicate genetic loci relevant to breast cancer risk.

258 **Transcriptome annotation drives colocalization and TWAS hits**

259 To further evaluate how transcript annotation shapes inference at GWAS loci, we
260 compared isoform-trait associations prioritized by both colocalization and TWAS for risk of breast
261 cancer (overall and five intrinsic-like subtypes). Across annotations, we identified 556 unique
262 significant associations, each defined by an isoform, trait, and tissue (Table S3). Despite similar
263 rates of GWAS locus tagging by eQTLs, GENCODE-informed annotations identified
264 approximately 2.5 times as many associations as assembly-derived annotations (Fig. 4a). This
265 ratio was highest in breast tissue (5.04) and fibroblasts (2.30) but was approximately equal in
266 tumor. Importantly, this difference in the number of prioritized hits does not reflect weaker
267 association signals with long-read-informed analyses. Among significant associations, the
268 strength of colocalization signals was comparable across annotations (median posterior
269 probability of shared causal variant: 0.98 for long-read, 0.98 for combined, 0.97 for GENCODE).
270 Similarly, the magnitude of significant TWAS effect sizes did not meaningfully differ (median 4.90
271 for GENCODE, 4.79 for long-read, and 4.87 for combined). Thus, fewer prioritized isoforms from
272 long-read annotations do not reflect weaker statistical support.

273 Just as many of the eIsoforms tagging the same GWAS locus were largely discordant,
274 annotation-dependent variability in isoform prioritization was substantial. Of the 556
275 colocalization and TWAS-prioritized associations, 381 (69%) were unique to a single annotation
276 within a tissue. Additionally, for each phenotype-tissue pair harboring at least one signal, the
277 number of annotation-unique associations consistently exceeded the number identified by two
278 or more annotations (Fig. S9). Notably, this variability occurred despite the high similarity
279 between GENCODE and combined annotations, indicating that even a modest contribution of
280 novel isoforms can substantially alter isoform-level inference. These patterns were even more
281 pronounced for the less common breast cancer subtypes (luminal B, HER2-enriched, luminal
282 B/HER2-, triple negative). Of these 57 tissue-trait-isoform associations, 47 (82.5%) of prioritized
283 hits were annotation-unique.

284 Many isoform-trait associations identified only with long-read annotations might be
285 expected to also appear with GENCODE, though assigned to a different but structurally similar
286 isoform from the same gene. However, we found that many long-read-prioritized associations
287 lacked any corresponding signal with GENCODE. Specifically, in fibroblasts, healthy breast
288 tissue, and tumor, 38.8%, 50.0%, and 62.5% of long-read-prioritized isoform-trait associations,
289 respectively, showed neither a significant trait association for the same isoform in GENCODE
290 nor any significant association for another isoform of the corresponding gene (Fig. 4b).

291 Conversely, when examining associations detected only with GENCODE, only 13.8%
292 (healthy breast), 56.2% (breast tumor) and 43.0% (fibroblasts) of unique isoforms prioritized by
293 GENCODE were present in the corresponding long-read annotations. This suggests that up to
294 half of regulatory hypotheses derived from tissue-agnostic annotations may involve transcripts
295 that are weakly expressed or possibly absent from the relevant tissue.

296 Next, we restricted analyses to prioritized associations within 1MB of independent breast
297 cancer risk GWAS loci. While long-read annotations yielded significantly fewer associations
298 overall, this also reduced the number of candidate isoforms per GWAS locus (Fig. 4c). Of note,
299 GENCODE prioritized up to 19 candidate isoforms (median 3) and 13 candidate genes (median
300 2) at a GWAS locus, while long-read annotations prioritized at most 7 isoforms (median 2) and
301 4 genes (median 1) at those same loci. Furthermore, at these shared loci, we observed
302 considerable variation in the top isoform prioritized by colocalization at each locus across all
303 annotation pairs (Fig. S10). For example, in fibroblasts for our two most well-powered
304 phenotypes (overall breast cancer and luminal A-like cancer), GENCODE and combined
305 annotations prioritized the same isoform approximately half of the time while GENCODE- and
306 long-read-prioritized isoforms agreed only approximately 30% of the time. Variability was even
307 greater in breast tumor analyses, although interpretation was limited by the much smaller
308 number of overall association signals in that tissue.

309 Long-read assemblies also have the distinct advantage of enabling direct evaluation of
310 novel, previously unannotated transcript structures. While both combined and long-read-only
311 annotations identified significant associations for novel isoforms (23 and 16, respectively), those
312 from the long-read annotations exhibited greater structural diversity (Fig. 4d). These included
313 five hits for transcripts categorized as incomplete-splice matches, while isoforms identified by
314 the combined annotations were largely novel-in-catalog transcripts.

315 **Long-read annotations prioritize new regulatory features**

316 *MARK1* (1q41) illustrates a high-confidence isoform-trait association detected only with
317 long-read annotations, with no corresponding TWAS or colocalization signal identified under
318 GENCODE. Only the fibroblast long-read annotation identified a putative driver isoform
319 associated with both overall breast cancer risk and luminal A-like cancer. *MARK1* encodes a
320 serine/threonine protein kinase involved in the regulation of microtubule dynamics and cell
321 polarity through the phosphorylation of microtubule-associated proteins[34]. This gene has been
322 recently implicated in the prediction of treatment response in HER2-positive breast cancer
323 tumors to targeted therapy[35] and has also been studied as a target of dysregulated miRNAs in
324 cervical cancer[36].

325 GENCODE version 45 catalogs 18 *MARK1* transcript-isoforms (Fig. 5a). However, when
326 we quantified the short-read fibroblast samples using GENCODE, only two of these isoforms
327 passed the minimum expression threshold (TPM>0.1 in ≥25% of samples) and were retained
328 for analysis (Fig. 5b). By contrast, when quantifying the same samples using the long-read
329 annotation, two isoforms also passed the expression threshold, but these were completely
330 distinct from the two retained under GENCODE. This gene precisely illustrates how two
331 underlying annotated isoform spaces used for expression quantification may partially overlap,

332 but the isoform sets surviving standard post-quantification quality control can be completely
333 distinct.

334 Across GENCODE, long-read, and combined annotations, we observed only one
335 significant colocalization between GWAS and isoQTL signals at this locus (Fig. 5c). This signal
336 corresponded to ENST00000366917 and was driven by lead variant rs11118564 in the long-
337 read annotation. Gene-level eQTLs were significant only under the long-read annotation and
338 attenuated when quantified using GENCODE or the combined references. Mapping short-read
339 alignments from five randomly selected GTEx fibroblast samples using STAR showed that reads
340 supporting this driver isoform were dispersed across multiple structurally similar transcripts when
341 quantified against GENCODE (Fig. 5d), thereby diluting isoform-specific regulatory signal.

342 While *MARK1* illustrates how long-read annotations can refine regulatory signals within a
343 fully annotated transcript space, long-read assemblies also uniquely enable the discovery and
344 functional evaluation of novel, previously unannotated transcript structures. One example is
345 *NUP107* (12q15, Fig. 6). *NUP107* encodes a core component of the nuclear pore complex, and
346 its associated subcomplex is essential for mRNA export and cell differentiation[37]. While not a
347 canonical cancer susceptibility gene, *NUP107* transcripts were prioritized by both colocalization
348 and TWAS in fibroblasts under GENCODE (ENST00000378905.6) and long-read annotations
349 (lead SNP rs7976687). However, the long-read prioritized isoform is a novel-in-catalog isoform
350 distinguished from reference transcripts by an exon composed of previously annotated splice
351 junctions at chr12:68713730-68715740 (Fig. 6a). Integrated 18-state chromatin learning data
352 from fibroblasts in the ROADMAP Epigenomics Project[38] showed that the start of this exon
353 coincided with a predicted enhancer (Fig. 6b). This observation supports a model in which exon
354 inclusion may be regulated by chromatin looping with the promoter or through effects on
355 transcriptional elongation and splicing dynamics.

356

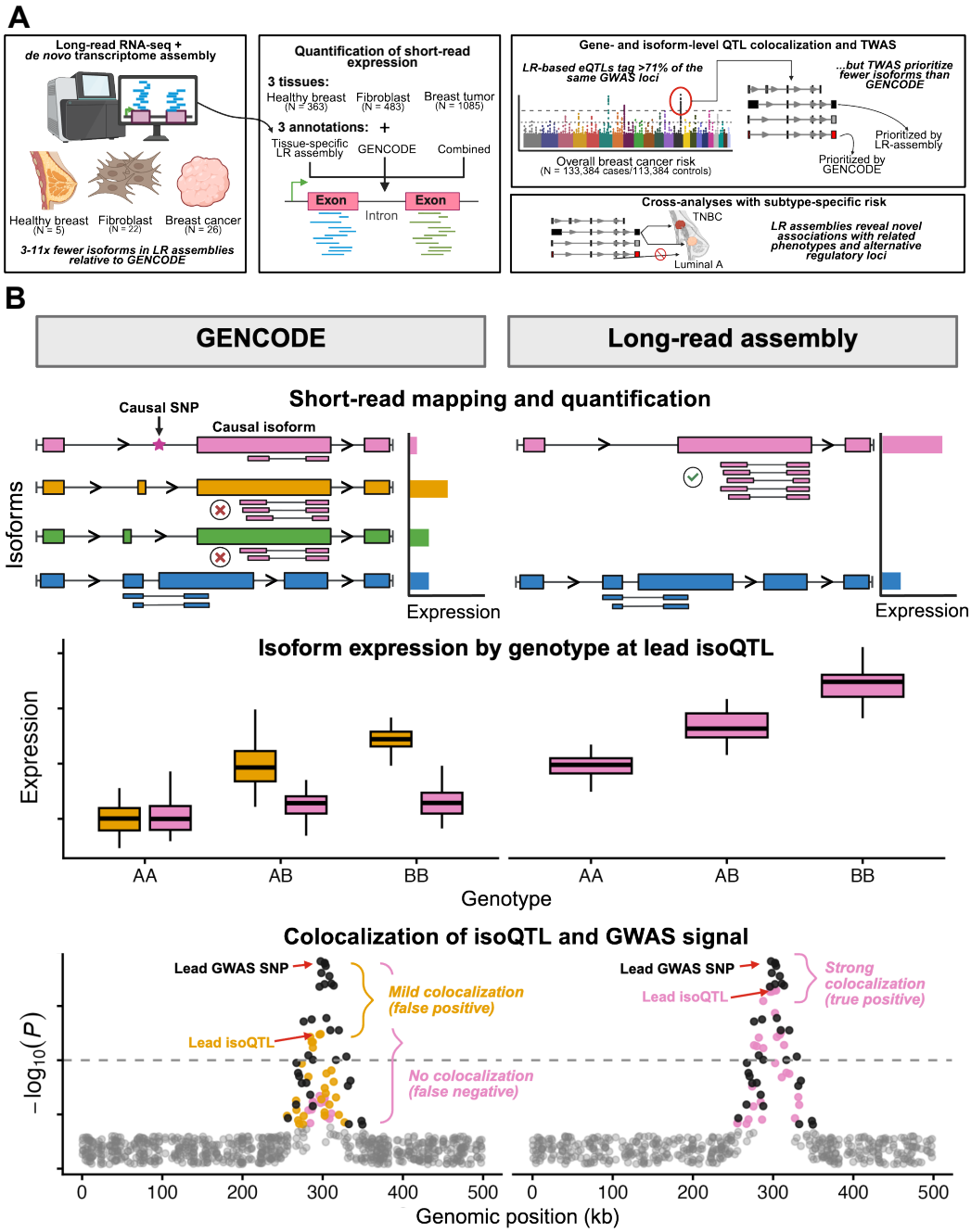


Figure 1. (A) In this project, we developed three de novo transcriptome assemblies in healthy breast tissue, fibroblasts, and breast tumor tissue derived from publicly available LR RNA-seq data (left panel). Next, we used these tissue-specific LR assemblies, the GENCODEv45 reference transcriptome, and combined annotations each to re-quantify short-read RNA-seq samples in healthy breast tissue, fibroblasts, and breast cancer tumor tissue (middle panel). We performed isoform- and gene-level eQTL mapping and fine-mapping, as well as colocalization and TWAS analyses with overall breast cancer GWAS risk loci and subtype-specific breast cancer loci (right panel). (**B**) In this example, the GENCODE transcript reference annotation contains four structurally similar isoforms for a gene. Only two are relevant to the tissue of interest and only one is causal for a trait, leading to misattribution of reads to irrelevant isoforms in GENCODE. This can propagate downstream and manifest as inaccurate isoform expression quantification, spurious eQTL mapping, and false positives and negatives in colocalization with GWAS loci. Abbreviations: LR, long-read; eQTL, expression quantitative trait loci; TWAS, transcriptome-wide association study; isoQTL, isoform-level eQTL; GWAS, genome-wide association study.

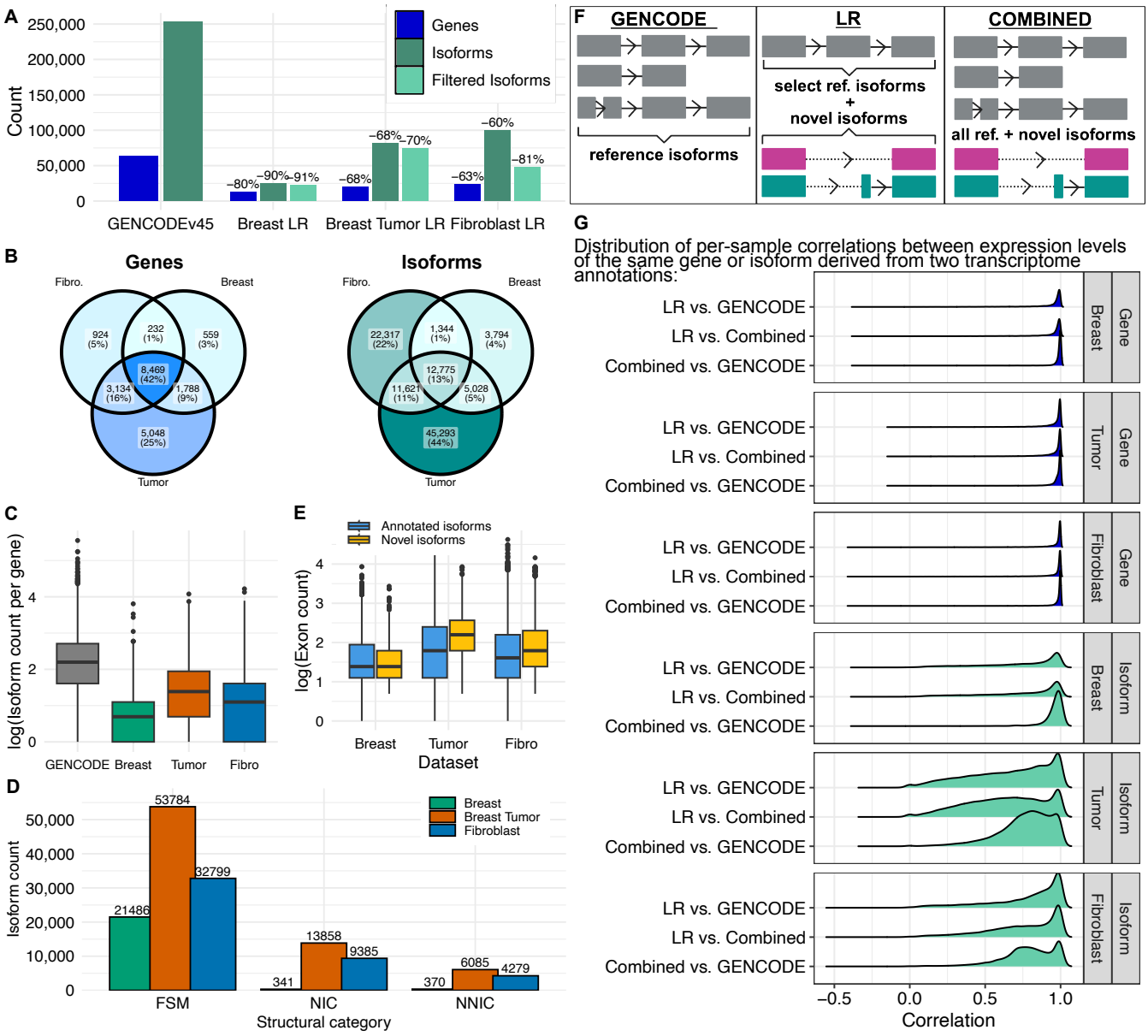


Figure 2. (**A**) Number of genes and isoform transcripts in the GENCODEv45 reference transcriptome compared to LR RNA-seq-derived healthy breast, breast tumor, and fibroblast transcriptome assemblies. Percentages shown represent percent decrease in these counts relative to GENCODE. (**B**) Overlap in post-filtering genes and isoforms across the three *de novo* long-read assemblies. (**C**) Log isoform counts per gene in LR derived assemblies and GENCODE for genes present in both long-read assemblies and GENCODE reference. Whiskers of boxplots extend to maximum/minimum point that is less than 1.5 × interquartile range from the third/first quartiles. **(D)** For each LR assembly, distribution of log exon counts for novel, unannotated isoforms versus transcripts present in GENCODE. (**E**) High confidence transcript isoforms assembled in the LR RNA-seq derived stratified by transcript structural category. (**F**) Graphical representation of how the three transcriptome annotations were constructed in each of our three short-read tissue contexts. Combined annotations incorporate all isoforms present in GENCODEv45 appended with all novel, non-FSM isoforms assembled in the long-read read samples of a given tissue. (**G**) Correlation of expression levels in re-quantified short-read samples across annotations for each gene (blue) and isoform (green) present in each pairwise annotation set. Abbreviations: LR, long-read; FSM, full splice match isoforms to annotated transcript; NIC, novel in catalog isoform; NNC, novel not in catalog isoform.

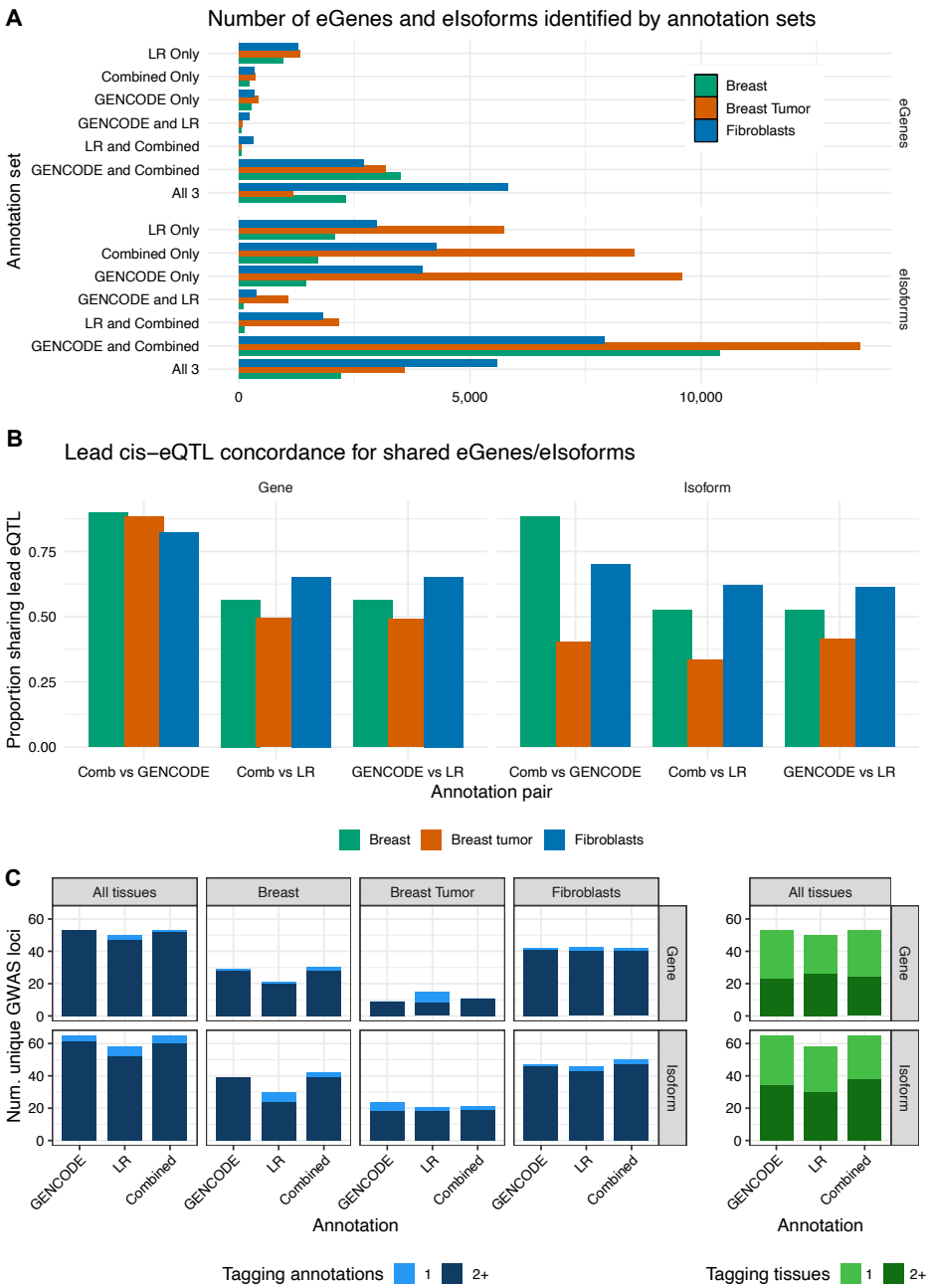


Figure 3. (**A**) Bars indicate the overlap of eIsoforms (isoforms with at least one significant eQTL) and eGenes (genes with at least one significant eQTL) identified in short-read healthy breast, breast cancer tumors, and fibroblasts across the three annotations. Sets representing intersections of two annotations (e.g., GENCODE and LR) indicate eGenes/eIsoforms identified only in those two annotations and not the third. (**B**) Bar heights reflect the proportion of eGenes (left) and eIsoforms (right) sharing the same lead cis-eQTL for features shared across annotations. **(C)** Gene- and isoform-level eQTL tagging of independent breast cancer GWAS loci. A tagged locus indicates an eQTL is within 1Mb and LD R^2>0.9 with the GWAS risk variant. The bars are colored to reflect the proportion of GWAS loci that were uniquely tagged by that single annotation (left) or in a single tissue (right). For all panels (**A**)-(**C**), eQTLs were identified from the conditional linear model analysis scheme. Abbreviations: LR, long-read; eQTL, expression quantitative trait loci; LD, linkage disequilibrium; GWAS, genome-wide association study; comb, combined.

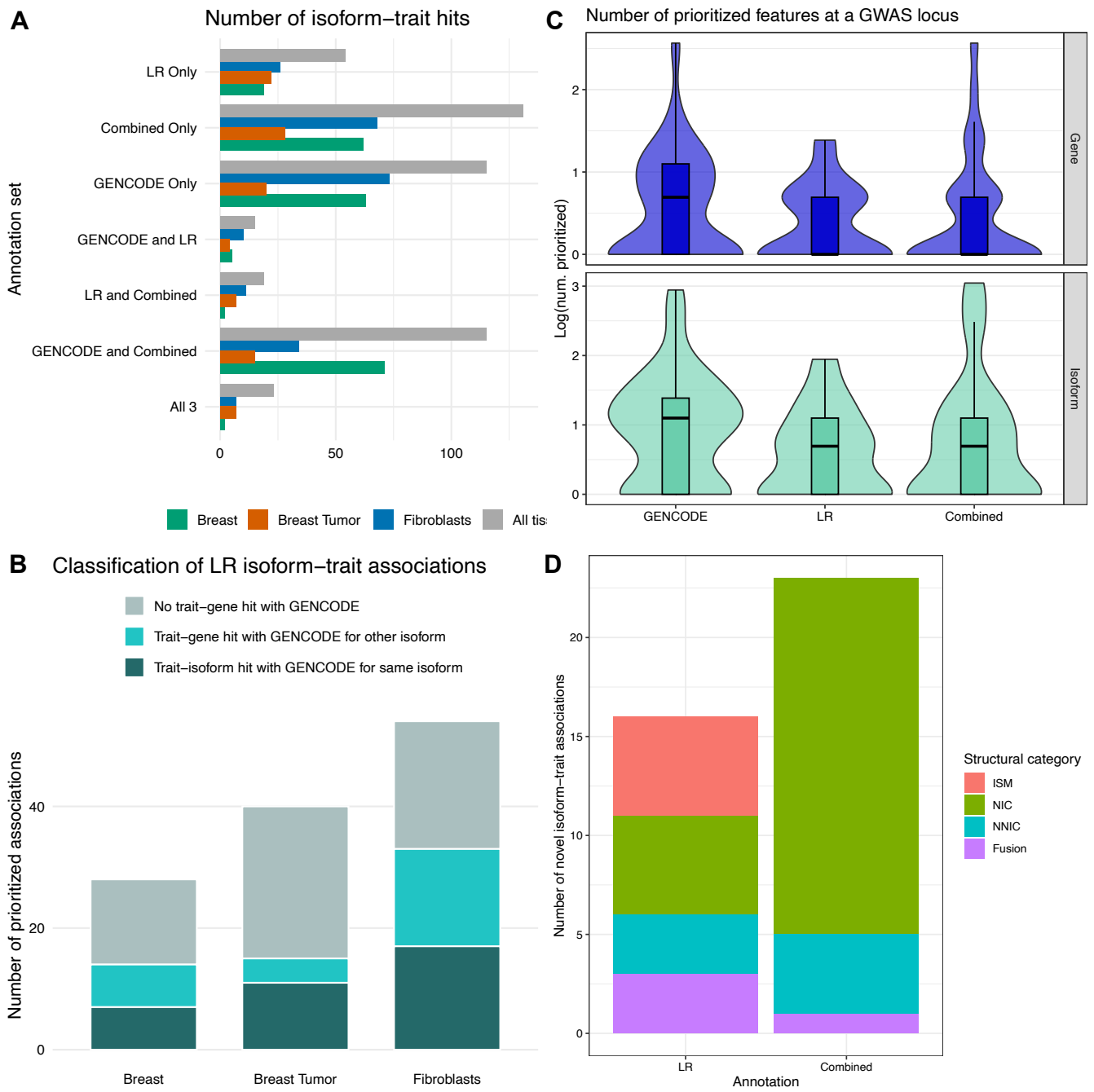


Figure 4. (**A**) Plots show the number of isoform-trait associations prioritized by both TWAS and colocalization in each annotation set for each tissue type. Unique hits aggregated across all tissues are shown in gray. Sets representing intersections of two annotations (e.g., GENCODE and LR) indicate associations identified only in those two annotations and not the third. (**B**) Bar height indicates the number of isoform-trait associations identified using long-read-derived transcriptome annotations by tissue. Bar fill indicates the proportion of associations with a corresponding significant isoform-level association for the same isoform and trait in GENCODE (dark blue), with a corresponding isoform-trait hit in GENCODE but for a different isoform of the gene (turquoise), and with no hits in GENCODE for that trait for any isoforms of the gene (pale blue). (**C**) Boxplots show the distribution of the number of prioritized features (top: genes, bottom: isoforms) within 1Mb of a GWAS risk locus by annotation aggregated across tissues. (**D**) Bar heights mark the number of prioritized isoform-trait associations aggregated across tissues for novel, unannotated isoforms only in the LR (left) and the combined annotations (right). Bar fill shows the distribution of the structural category of these novel isoforms. Abbreviations: TWAS, transcriptome-wide association study; LR, long-read; ISM, incomplete-splice match; NIC, novel in catalog; NNIC, novel not in catalog; GWAS, genome-wide association study.

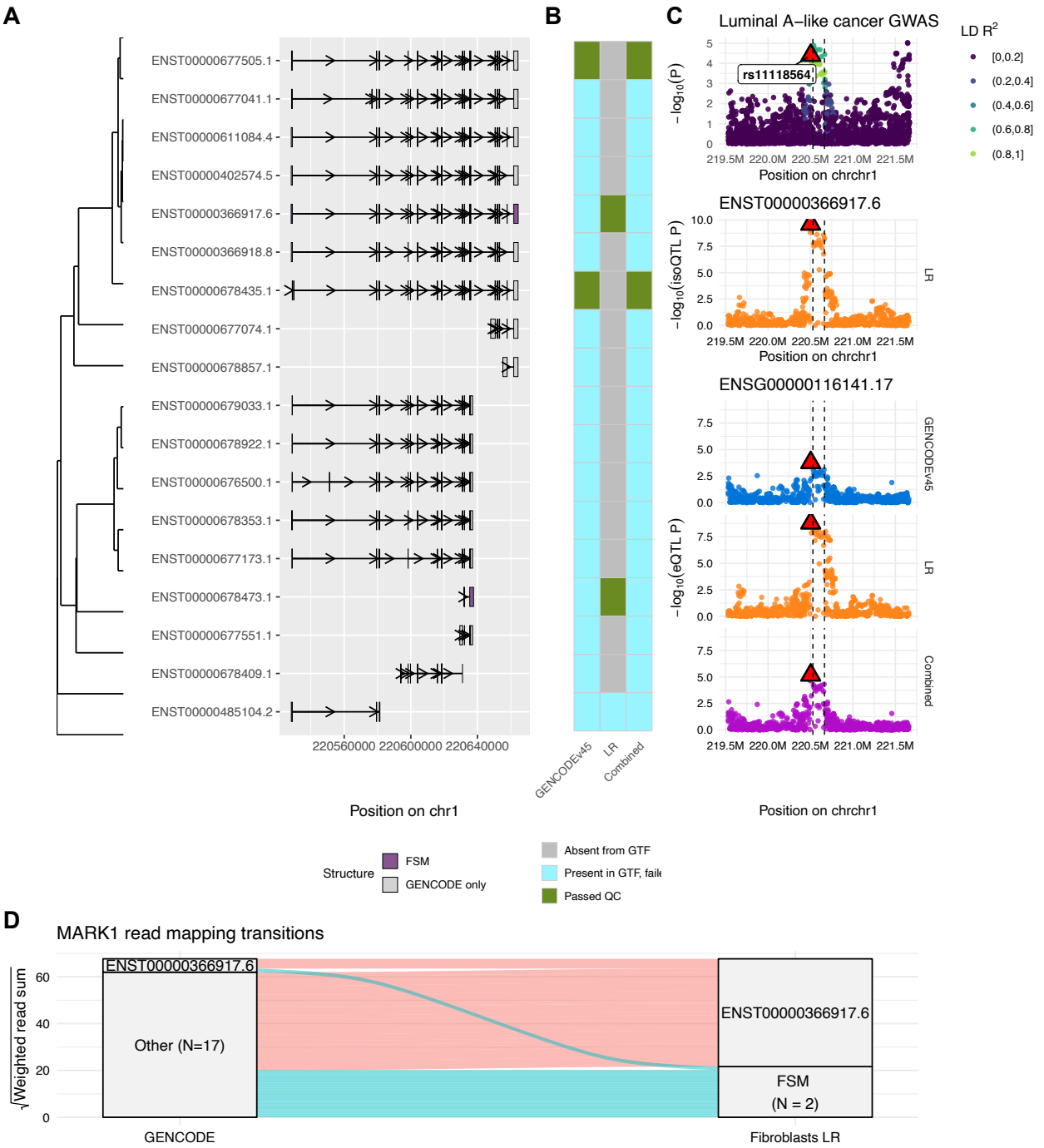


Figure 5. (**A**) Exon-intron structures of *MARK1* transcript isoforms on chromosome 1. Exons are colored by isoform presence in the fibroblast long-read assembly (FSM, purple) or GENCODE only (gray). (**B**) Isoform detection and QC in fibroblasts: gray indicates isoforms not detected in the long-read assembly, blue indicates detected isoforms that failed quality control filtering, and green indicates isoforms that passed. (**C**) Regulatory signals in the *MARK1* locus. Top panel: -log10(GWAS P-value) for variants within 1 Mb of the gene. The lead eQTL variant for the LR-prioritized isoform (rs11118564) is marked by a red triangle. Variant colors reflect LD R² with this lead variant in LR fibroblasts. Middle panel: eQTL signal for the prioritized isoform (ENST00000366917.6) in the long-read annotation. Bottom three panels: gene-level eQTL signals across all three transcript annotations in fibroblasts. (**D**) Short-read alignment transitions to *MARK1* transcripts in five randomly selected GTEx fibroblast samples. Left: alignments to GENCODE annotation; right: alignments to LR fibroblast annotation. All transcripts except the LR-prioritized isoform are collapsed for ease of visualization. Abbreviations: FSM, full-splice match; LR, long-read; LD, linkage disequilibrium; GWAS, genome-wide association study; eQTL, expression quantitative trait loci; isoQTL, isoform eQTL; GTF, gene transfer format file; QC, quality control.

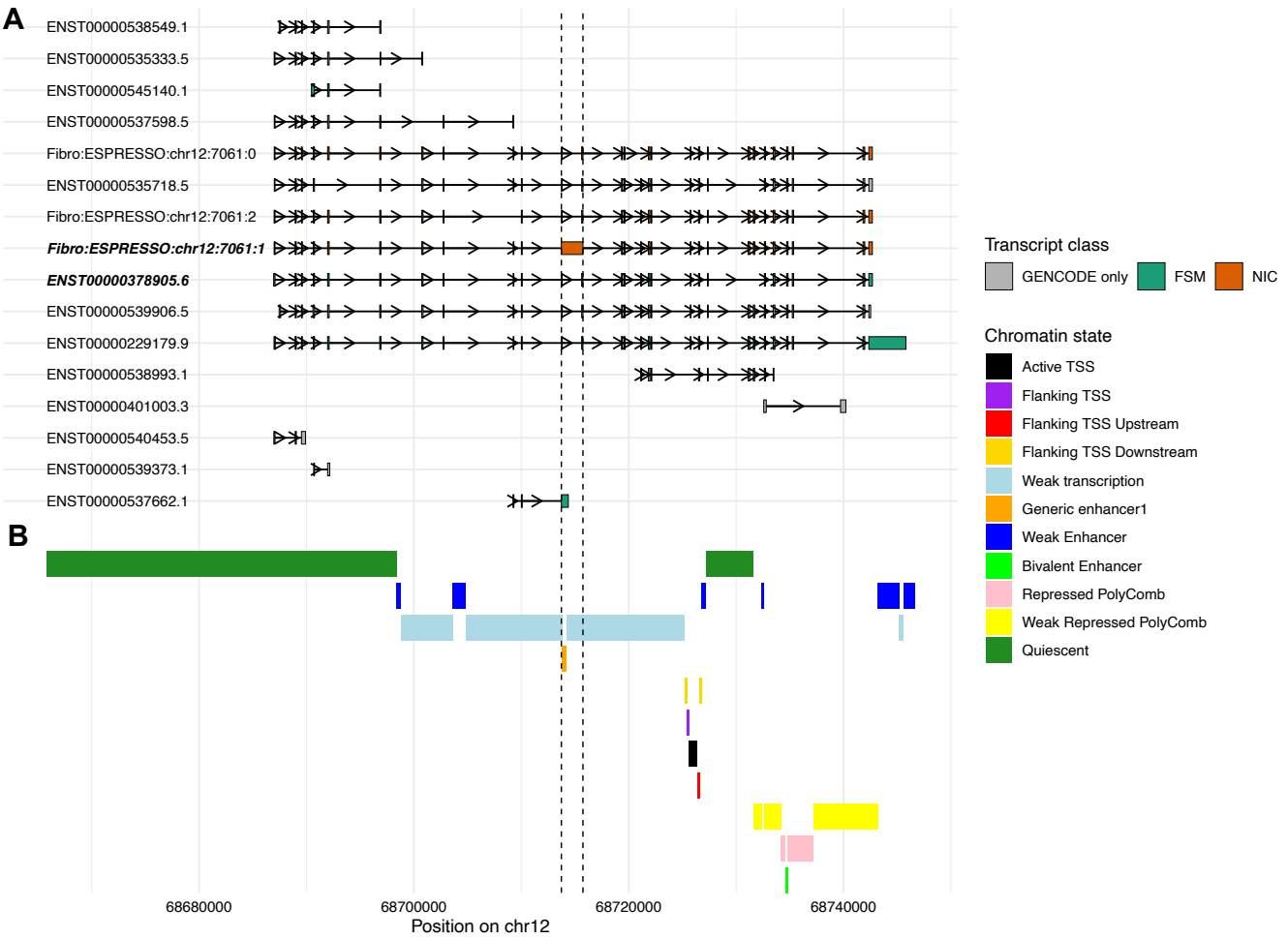


Figure 6. (**A**) Exon-intron structures of *NUP107* transcript isoforms on chromosome 12. Exons are colored by isoform presence and structural classification in the fibroblast long-read assembly: FSM (green), GENCODE only (gray), and NIC (orange). The prioritized novel isoform assembled with ESPRESSO and the GENCODE-prioritized isoform of this gene are bolded. (**B**) Chromatin state landscape across the *NUP107* locus. Genomic regions are annotated using the 18-state ChromHMM model from ROADMAP Epigenomics lung fibroblasts. Colors indicate predicted functional states, including active TSS, weak enhancers, generic enhancers, quiescent regions, and other functional categories. This track highlights the local regulatory context overlapping the LR-prioritized isoform. Abbreviations: FSM, full splice match; NIC, novel in catalog; TSS, transcription start site; LR, long-read.

438 **DISCUSSION**

439 In this study, we identified 556 unique isoform-level associations with breast cancer risk using
440 three short-read RNA-seq datasets from malignant and normal breast tissues and cultured
441 fibroblasts. We found striking differences in the association patterns depending on the transcript
442 annotations used for expression quantification. These findings suggest that transcriptome
443 annotations should be treated not merely as default inputs in integrative analyses of gene
444 regulation, but rather as biologically meaningful choices that shape downstream inference.
445 Tissue-specific long-read assemblies (1) substantially reduce the isoform space, focusing the
446 hypothesis space for causal isoforms in the relevant tissue context, (2) avoid spurious
447 associations with tissue-irrelevant transcripts, and (3) improve the specificity of inferred
448 regulatory mechanisms at GWAS loci. By prioritizing empirically supported isoforms, long-read-
449 informed annotations can help resolve ambiguity at the isoform level, the most fundamental unit
450 of transcription that remains unexplored in most integrative analyses of complex traits.

451 TWAS have come under recent scrutiny for high false positive rates, high false sign rates,
452 and limited replicability[33,39–41]. While the polygenicity of complex traits has been proposed as
453 one major driver of these challenges, our current and previous findings suggest that variability
454 in transcriptome annotations and quantification methods may also contribute substantially[33].
455 Standard transcriptomic references such as GENCODE and Ensembl are largely tissue-agnostic
456 and catalog hundreds of thousands of transcripts spanning protein-coding genes, long
457 noncoding RNAs, pseudogenes, and other biotypes. This expansive isoform space continues to
458 grow with each release and may increase the likelihood of read misassignment to lowly
459 expressed or tissue-irrelevant transcripts. In turn, this can dilute true regulatory signals and
460 ultimately obscure biologically relevant mechanisms. Additional critical strides must be made to
461 better understand how annotation choice shapes downstream genetic inference.

One such critical next step is to accurately quantify how uncertainty arising from overpopulated transcript references impacts statistical power and type I error of eQTL discovery, TWAS, and colocalization studies. Our results directly motivate the development of improved simulation frameworks to address this question. Existing tools for QTL simulation rely on simplified distributional assumptions that do not capture the complexity of isoform-level RNA-seq data. More realistic approaches that generate raw sequencing reads, model multiple modes of genetic regulation, and incorporate modern library preparation protocols will be needed to systematically evaluate these annotation-dependent biases in downstream analyses.

While our results suggest that re-quantification using tissue-matched long-read-informed annotations provides the most precise framework for regulatory inference, this approach may not always be feasible for large legacy datasets. These constraints motivate the development of methods that preserve the benefits of annotation-aware inference while reducing the need for full RNA-seq reprocessing. Moving beyond simple associations toward causal resolution of driver isoforms will likely require isoform-level fine-mapping frameworks that jointly model the effects of individual transcripts, total gene expression, and nearby regulatory variants. We believe incorporating biologically informed priors derived from long-read evidence (e.g., structural support, splice junction validation) may further improve causal inference while reducing the need for large-scale resequencing efforts.

Long-read panels remain limited by sample size, high sequencing costs, and scattered representation across tissue contexts. In this study, healthy breast tissue posed a particular challenge despite its importance for modeling germline susceptibility to breast cancer and its status as the gold-standard context in existing breast cancer TWAS. Limited long-read sample sizes can lead to incomplete transcript capture and reduced sensitivity for low-abundance isoforms, which in turn, can severely limit discovery. As publicly available long-read resources expand, combining samples across independent and diverse short-read cohorts for a given

487 tissue may help improve power while preserving tissue specificity. Despite these challenges, we
488 are optimistic that ongoing technological advances and growing long-read catalogs will likely
489 substantially mitigate these concerns.

490 **METHODS**

491 **Long-read transcriptome assembly construction**

492 We obtained raw long-read RNA sequencing FASTQ files via the European Genome
493 Archive database (accession number EGAS00001004819) from Veiga et al.[18], who generated
494 PacBio single-molecule reads from a total of 30 breast tissue samples. These included 4 healthy
495 breast samples (two primary tissues and two from cell lines HMEC and HS578BST) and 26
496 breast cancer samples comprising 13 primary breast tumor biopsies (covering hormone
497 receptor-positive, HER2+, and triple-negative subtypes), 9 patient-derived xenograft tumors, and
498 4 breast cancer cell lines (HS578T, MCF-7, MDA-MB-231, MDA-MB-468). Detailed sample
499 characteristics (subtype, origin, receptor status, patient demographics) are available in the
500 supplement of the original publication, where applicable[18]. We also obtained 22 long-read RNA
501 sequencing samples from cell-cultured fibroblasts and one non-malignant breast mammary
502 tissue sample from GTEx V9[42]. Fibroblast samples were derived from skin biopsies from the
503 lower leg, and cell lines were cultured under standard conditions with a subset undergoing
504 *PTBP1* siRNA knockdown, high-quality RNA extraction, poly(A) selection, cDNA library
505 preparation, and Oxford Nanopore long-read sequencing according to manufacturer protocols
506 described elsewhere[43]. These raw long-read data served as input for the transcriptome
507 assembly and quantification procedures described below.

508 We converted raw FASTQ files to unmapped BAM files with samtools[44] import and
509 processed as in Veiga et al[18] with the IsoSeq workflow[45]. Briefly, we used lima to remove New
510 England Biolabs and Clontech adapters and to orient the reads. We used IsoSeq refine to filter

511 out reads without poly(A) tails and then trim poly(A) tails and concatemers. We then aligned the
512 full-length non-concatemer poly(A) reads to the human genome assembly analysis set hg38
513 (GenBank assembly 000001405.15, no ALT contigs) using minimap2[46] with a k-mer length of
514 14, minimizer window size of 4, and in splice-aware mode with –splice-flank=no to relax
515 assumptions about bases adjacent to splice junction donors and acceptors. We filtered
516 secondary alignments and coordinate sorted primary alignments with samtools[44].

517 For the healthy breast, tumor, and fibroblast samples separately, we merged the filtered
518 and sorted alignments from chromosomes 1-22, X, Y, and M across all samples into one BAM
519 file. This served as input for *de novo* transcript assembly using ESPRESSO[26]. This method has
520 performed well in maximizing discovery while minimizing false positives in error-prone long-read
521 RNA-seq data[26]. We then used SQANTI[27] to characterize the *de novo* assemblies and classify
522 isoforms into the following structural categories relative to GENCODE v45 transcripts: full splice
523 matches (FSM) with a perfect structural match to an annotated isoform in GENCODE,
524 incomplete splice matches (ISM) with some splice junctions matching reference transcripts,
525 novel in catalog (NIC) with new combinations of annotated splice junctions in reference
526 transcripts, novel not in catalog (NNC) with novel donors and/or receptors, and other classes
527 such as fusion products and antisense transcripts[47].

528 We implemented custom filtering protocols to define high-confidence ESPRESSO
529 transcripts with integrated orthogonal evidence including short-read splicing junction support,
530 proximity to external transcription start and termination site (TSS, TTS) annotations, and
531 sequence-based artifact filtering. Orthogonal sources included DNase-seq peaks from
532 ENCODE[48], 3' and sequencing alternative polyadenylation site data from PolyASite[49–51], and
533 CAGE-seq peaks from PHANTOM5[52]. We provide full accessions in Table S4. We applied filters
534 to novel isoforms to remove likely technical artifacts such as RT switching, intrapriming, and

535 spurious nonsense-mediated decay predictions. We provide descriptions of full quality control
536 criteria elsewhere[53].

537 We generated "combined" gene and transcript isoform annotations by integrating high-
538 confidence novel isoforms from our long-read healthy breast, breast tumor, and fibroblast
539 assemblies with the GENCODE reference. We constructed a transcript-to-gene mapping by
540 assigning transcripts to annotated genes or, when absent, to novel gene IDs. We then appended
541 novel genes and transcript isoforms from ESPRESSO absent in GENCODE to the reference.
542 We harmonized merged annotations, resulting in two combined gene transfer format (GTF) files
543 containing both GENCODE genes and high-confidence long-read transcripts for downstream
544 analyses per tissue. We then used MashMap[54] to construct decoy-aware transcriptome indices
545 for Salmon[55] for the long-read-derived healthy, tumor, and fibroblast assemblies, the reference
546 GENCODE annotation, and the three combined annotations of GENCODE concatenated with
547 unannotated non-FSM transcripts from the long-read assemblies.

548 **Genotype processing**

549 *GTEx V8 dataset.* We describe the full protocols for processing GTEx genotypes elsewhere[8].
550 Briefly, we obtained genotypes from GTEx V8[56] from whole-genome sequencing data of 838
551 postmortem donors. We retained only single nucleotide polymorphisms (SNPs) and next filtered
552 out variants with elevated genotype missingness (>5%), variants with minor allele frequency
553 (MAF) <1%, and those showing significant deviation from Hardy-Weinberg equilibrium (P < 10E-
554 6). We additionally removed SNPs with a MAF<1% in the European ancestry subset of the 1000
555 Genomes Project[57]. We were left with a total of 1,131,422 autosomal variants (GRCh38) for
556 analysis.

557 *TCGA dataset.* We converted Affymetrix SNP 6.0 CEL files for TCGA breast carcinoma (BRCA)
558 blood-derived normal samples to variant call format (VCF) using a custom R script. We first

559 prepared a probe annotation file mapping SNP probe identifiers to GRCh38 coordinates by
560 obtaining hg19 probe positions from the UCSC snpArrayAffy6 table and lifting them to GRCh38
561 using the UCSC liftOver chain via the rtracklayer R package, retaining probe identifiers,
562 reference and alternative alleles, and strand information. We then performed genotype calling
563 on each CEL file individually using the Corrected Robust Linear Model with Maximum Likelihood
564 Distance (CRLMM) algorithm, implemented in the crlmm R/Bioconductor package[58]. We
565 recorded CRLMM genotype calls to VCF genotype notation and mapped to the lifted-over
566 GRCh38 probe coordinates. We converted genotype confidence scores returned by CRLMM to
567 Phred-scaled genotype quality (GQ) values using the formula $GQ = -10 \times \log_{10}(1 - p)$, where p
568 is the CRLMM confidence score, with values capped at 99. Variants with GQ < 20 were flagged
569 as FAIL in the VCF FILTER field, and we wrote per-sample VCFs with GT and GQ FORMAT
570 fields. Probes that failed to return a genotype call were excluded. We filtered VCFs with bcftools
571 to exclude variants with GQ < 20 or FILTER status of FAIL. We sorted filtered VCFs, compressed
572 them with bgzip, and indexed with tabix. We validated variant calls using GATK ValidateVariants.

573 We phased filtered VCFs using Eagle2[59] with the 1000 Genomes Project high-coverage
574 phased reference panel. This panel was prepared by filtering for sites with minor allele count ≥
575 2, normalizing multiallelic variants with bcftools norm, removing duplicate sites, and converting
576 to BCF format. We used a GRCh38 genetic map for recombination rate estimation.

577 We performed genotype imputation using Minimac4 v4.1.6 with the 1kGP high-coverage
578 phased reference panel in MSAV format[60]. We concatenated imputed per-chromosome BCFs
579 into a final genome-wide imputed BCF and indexed. Following imputation, we filtered out
580 variants with imputation R^2<0.8, those with high missingness (>5%), variants with MAF<1%, and
581 those failing Hardy-Weinberg equilibrium (P < 10E-6). Among individuals with two genotyped
582 samples derived from both blood and solid tissue, we preferentially kept those from blood. We

583 were left with a total of 2,679,319 autosomal variants (GRCh38) for analysis across 1,096
584 unrelated individuals.

585 **RNA-seq processing**

586 We first obtained 459 aligned BAM files for breast mammary tissue and 504 files for
587 cultured fibroblast samples from the GTEx project V8 (dbGaP: phs000424.v8.p2) that passed
588 GTEx Analysis Freeze quality control. We used SAMtools[44] to convert the BAM files to paired-
589 end FASTQ format for quantification via Salmon[55] (described below) and discarded unpaired
590 and orphaned reads. Next, we used trimmomatic[61] to quality trim the raw paired-end FASTQ
591 files. For quality-control alignment metrics, we built a STAR genome index from the GRCh38
592 reference FASTA and GENCODEv45 gene annotation and re-aligned the trimmed reads to
593 generate BAM files. For TCGA, we extracted RNA-seq reads from GDC-aligned BRCA BAM
594 files by converting to paired-end FASTQ format using samtools. We performed adapter trimming
595 and quality filtering with fastp using default parameters. We subsequently retained only primary
596 solid tumor breast cancer samples, excluding metastatic and tumor-adjacent samples.

597 For both GTEx and TCGA datasets, we used Picard[62] to collect GC bias, RNA-seq, and
598 alignment summary metrics. We aggregated all results with MultiQC[63] to produce inputs for
599 downstream hidden covariate estimation.

600 **Expression quantification**

601 We quantified the GTEx trimmed reads using Salmon with three transcript annotations:
602 GENCODEv45, the long-read *de novo* breast and fibroblast assemblies, and the combined
603 GTFs concatenating unannotated non-FSM transcripts to the reference annotation. We similarly
604 quantified trimmed reads with the GENCODE reference transcriptome, as well as the long-read
605 tumor assembly and the combined annotation of those two. We built Salmon v1.10.2 indices for
606 each reference using a decoy-aware strategy. We used the generateDecoyTranscriptome.sh

607 script (provided with Salmon) to construct a gentrome FASTA by appending the GRCh38
608 reference genome (GCA_000001405.15, no-alt analysis set) as decoy sequence to the
609 transcript FASTA. We used MashMap for decoy identification and BEDtools for coordinate
610 extraction. We generated Salmon indices from the resulting gentrome with a k-mer size of 31
611 and decoy sequence identifiers provided via a decoys list. We performed transcript-level
612 quantification with Salmon in mapping-based mode[55] using automatic library type detection,
613 mapping validation, sequence-specific bias correction, and 50 bootstrap replicates for
614 uncertainty estimation.

615 We imported isoform-level quantifications into R using tximeta[64] and removed non-
616 autosomal transcripts. We aggregated isoform-level quantifications to the gene-level using
617 summarizeToGene() from tximeta. We removed lowly expressed transcripts and genes by
618 requiring each to have TPM>0.1 in at least 25% of samples. To address overdispersion arising
619 from read-to-transcript ambiguity, we applied catchSalmon() from the edgeR[65] package and
620 divided TPM of each isoform by its estimated overdispersion factor. We normalized gene- and
621 transcript-level quantifications using trimmed-method of means (TMM, edgeR[65]) and
622 subsequently transformed them to log scale using the variance-stabilizing transformation
623 implemented in DESeq2[66]. We then removed outlier samples with gene-level WGCNA network
624 connectivity scores below -3[67].

625 To visually assess read mapping transitions between GENCODE and the long-read-
626 derived fibroblast annotation, we also performed alignment of reads with STAR in five randomly
627 selected short-read samples.

628 **eQTL mapping and fine-mapping**

629 Prior to eQTL mapping, we first estimated hidden covariates to control for latent technical
630 variation in expression matrices. We estimated these components using a regularized latent

631 factor model incorporating RNA-seq, GC content, and genome-level quality metrics from Picard
632 as prior covariates[62,68]. We applied this procedure to normalized gene- and isoform-level
633 expression matrices and inferred k latent components for downstream analyses. For each
634 annotation and tissue, we ran this procedure four times assuming k=(15,50,100,200) hidden
635 technical covariates.

636 We next performed gene- and isoform-level *cis*-eQTL mapping using QTLtools[69] with the
637 conditional linear regression framework. We first ran permutation-based analysis with 1,000
638 permutations to inform the nominal p-value thresholds at FDR 5%. This permutation procedure
639 adjusts nominal p-values for the number of variants in the *cis* region and LD within this 1MB
640 window. We corrected expression phenotypes for covariates (age, sex, top 5 genetic principal
641 components to control for population structure, and hidden covariates). We determined the
642 optimal number of hidden covariates for use in eQTL mapping, as well as colocalization and
643 TWAS, as those that maximized eGenes and eIsoform discovery, respectively. Following
644 WGCNA-based outlier removal and restriction to samples with complete genotype, covariate,
645 and sequencing data, we retained 483/483/476 fibroblasts (GENCODE/combined/LR),
646 360/360/362 healthy breast samples, and 1,083/1,085/1,085 tumors. Differences in sample
647 counts reflect minor annotation-specific variation in WGCNA-based outlier detection. After
648 covariate adjustment, we rank normal transformed each adjusted expression vector. Following
649 this, we performed eQTL fine-mapping using the SuSiE Bayesian variable selection
650 framework[70]. We again used a 1MB window around the gene or isoform to extract genotype data
651 and centered and scaled genotypes. With residualized expression matrices, we fit SuSiE with
652 L=10 single-effect components to estimate posterior inclusion probabilities (PIP) and identify
653 credible sets with 95% minimum coverage.

654 **BCAC GWAS summary statistics**

655 We obtained summary-level breast cancer GWAS data from the Breast Cancer
656 Association Consortium (BCAC, see Web Resources). The summary statistics for variant-level
657 associations with breast cancer risk and risk of specific breast cancer subtypes were the result
658 of a large multi-study GWAS of women of European genetic ancestry[5]. The overall breast cancer
659 analysis used genotype data from cases (invasive, *in situ*, unknown invasiveness) and controls
660 across 82 BCAC studies that were genotyped using either the iCOGS or OncoArray Illumina
661 genome-wide custom arrays. For this overall analysis, data from 11 other breast cancer GWASs
662 were incorporated. This yielded a total sample size of 133,384 cases and 113,789 controls. The
663 authors estimated SNP-disease associations using standard logistic regression, adjusting for
664 country of origin and top principal components. Results were obtained for iCOGS subjects,
665 OncoArray subjects, and additional GWASs separately and then combined via fixed-effects
666 meta-analysis.

667 In addition to the summary statistics for the outcome of overall breast cancer risk, this
668 study also published summary statistics for the association of variants with risk of specific
669 intrinsic-like subtypes of breast cancer: luminal A-like cancer, luminal B-like cancer, luminal
670 B/HER2-negative-like cancer, HER2-enriched-like cancer, and triple-negative cancer. These
671 subtype analyses were performed by fitting two-stage polytomous logistic regression models.
672 Full details on these models are described elsewhere[71,72]. Only invasive cases were considered
673 for this analysis, and samples from the 11 additional GWASs were not included due to missing
674 tumor marker information. The final sample for the GWAS subtype analyses included 106,278
675 cases and 91,477 controls.

676 Lastly, a meta-analysis was performed combining results from the analysis of triple-
677 negative breast cancer cases from the BCAC and the separate analysis of cases and controls
678 with a pathogenic *BRCA1* variant from the Consortium of Investigators of Modifiers of BRCA1/2
679 (CIMBA)[73]. The CIMBA participants were also of European ancestry. Authors performed a fixed-

effects meta-analysis, combining odds ratio estimates from the BCAC study and hazard ratio estimates from the CIMBA study. As approximately 69% of breast cancer cases in individuals with *BRCA1* mutations are triple negative[74], our "triple-negative" TWAS results utilized summary statistics from this meta-analysis for greatest power. Quality control, imputation, and ancestry classification protocols for all sets of genotype data used in the study are described separately[1,6,75,76], and additional details on the statistical methods used in the breast cancer GWAS are provided by Zhang et al.[5] The full list of all contributing studies to each analysis (overall breast cancer, subtype specific, and CIMBA) can be found in Tables S1-S3 of Zhang et al.[5] Once we obtained all summary data from these GWASs, we performed liftover to map GWAS variants to GRCh38. Subsequently, we removed any variants with imputation R^2 below 0.3 when available, strand-ambiguous alleles, and non-biallelic SNPs.

To define independent risk loci across all 6 sets of breast cancer GWAS summary statistics, we first pruned GWAS SNPs to those with $P < 5E-8$. We then performed LD-based clumping in PLINK[77] using the 1000 Genomes European reference panel[57] ($R^2 < 0.05$ within 1 MB and secondary threshold $P < 1E-4$) and retained the lead SNP per locus. This procedure yielded 332 independent breast cancer risk loci.

Colocalization and isoform-level TWAS

We performed Bayesian colocalization at the gene- and isoform-resolution to test for a single shared genetic variant driver of both expression signals from marginal variant-level linear regression and GWAS summary statistics using the coloc.abf() functionality within coloc with default priors[78]. Next, we utilized the isoTWAS framework to train predictive models of isoform expression using all cis-SNPs within 1MB of each gene body across the six breast cancer phenotypes. Details on these statistical models are provided elsewhere[8]. We selected the optimal model for each isoform with 10-fold cross-validation, and we retained only those features

704 with cross-validated $R^2 > 0.01$. These models were then applied in a weighted burden framework
705 to evaluate associations between the genetically-predicted component of isoform expression
706 and cancer risk. To address the multiple testing burden for isoform-level analyses, we
707 implemented a two-stage testing procedure as described previously[8]. For prioritized isoforms,
708 we further evaluated whether SNP-expression weights contributed signal beyond SNP-disease
709 associations observed in the GWAS using a conservative permutation test. We permuted SNP-
710 isoform weights 10,000 times to generate a null distribution of the test statistic and derived
711 empirical p-values. We restricted permutation testing to features with FDR-adjusted $P < 0.05$
712 and isoTWAS hits meeting both false discover rate-adjusted $P < 0.05$ and family-wide error rate-
713 adjusted $P < 0.05$.

714 **Data availability**

715 Long-read RNA-seq-derived transcriptome annotations, scripts, and all code used to generate
716 the data and figures presented in this paper are available upon publication at
717 https://github.com/staylorhead/Long_Read_Informed_Breast_Cancer. The GWAS summary
718 statistic data for risk of overall and subtype-specific breast cancers are publicly available and
719 hosted at: https://www.ccge.medschl.cam.ac.uk/breast-cancer-association-consortium-
720 bcac/data-data-access/summary-results. Individual-level datasets from TCGA were obtained via
721 dbGaP accession number phs000178.v11[79]. Individual-level GTEx data were accessed through
722 dbGaP under accession phs000424.v8.p2[56]. LD reference data from the 1000 Genomes
723 Project[57] were obtained from the International Genome Sample Resource data portal at
724 https://www.internationalgenome.org/data-portal/sample. The GENCODE human reference
725 transcriptome data (version 45) were downloaded from
726 https://www.gencodegenes.org/human/release_45.html.

727 **Acknowledgements**

728 This work was supported by the National Institutes of Health (R21CA293419, R01CA237541,
729 R01CA264987).

730 **References**

731 1. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature*
732 **551**, 92–94 (2017).
733 2. Adedokun, B. *et al.* Cross-ancestry GWAS meta-analysis identifies six breast cancer loci in
734 African and European ancestry women. *Nat. Commun.* **12**, 4198 (2021).
735 3. Shu, X. *et al.* Identification of novel breast cancer susceptibility loci in meta-analyses
736 conducted among Asian and European descendants. *Nat. Commun.* **11**, 1217 (2020).
737 4. Ahearn, T. U. *et al.* Common variants in breast cancer risk loci predispose to distinct tumor
738 subtypes. *Breast Cancer Res.* **24**, 2 (2022).
739 5. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer
740 susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* **52**, 572–581
741 (2020).
742 6. Milne, R. L. *et al.* Identification of ten variants associated with risk of estrogen-receptor-
743 negative breast cancer. *Nat. Genet.* **49**, 1767–1778 (2017).
744 7. Garcia-Closas, M. *et al.* Genome-wide association studies identify four ER negative–
745 specific breast cancer risk loci. *Nat. Genet.* **45**, 392–398 (2013).
746 8. Bhattacharya, A. *et al.* Isoform-level transcriptome-wide association uncovers genetic risk
747 mechanisms for neuropsychiatric disorders in the human brain. *Nat. Genet.* **55**, 2117–2128
748 (2023).
749 9. Chang, Y.-H. *et al.* Isoform-level analyses of 6 cancers uncover extensive genetic risk
750 mechanisms undetected at the gene-level. *Br J Cancer* **133**, 874–885 (2025).
751 10. Wen, C. *et al.* Cross-ancestry atlas of gene, isoform, and splicing regulation in the
752 developing human brain. *Science* **384**, eadh0829 (2024).

753 11. Hatami, R. *et al.* KLF6-SV1 drives breast cancer metastasis and is associated with poor
754 survival. *Sci. Transl. Med.* **5**, 169ra12 (2013).
755 12. Nielsen, F. C., van Overeem Hansen, T. & Sørensen, C. S. Hereditary breast and ovarian
756 cancer: new genes in confined pathways. *Nat. Rev. Cancer* **16**, 599–612 (2016).
757 13. Wang, Y. *et al.* The BRCA1-Δ11q Alternative Splice Isoform Bypasses Germline Mutations
758 and Promotes Therapeutic Resistance to PARP Inhibition and Cisplatin. *Cancer Res.* **76**,
759 2778–2790 (2016).
760 14. Weigelt, B. & Reis-Filho, J. S. Activating Mutations in HER2: Neu Opportunities and Neu
761 Challenges. *Cancer Discov.* **3**, 145–147 (2013).
762 15. Sharif, G. M. *et al.* An AIB1 Isoform Alters Enhancer Access and Enables Progression of
763 Early-Stage Triple-Negative Breast Cancer. *Cancer Res.* **81**, 4230–4241 (2021).
764 16. Borsi, L. *et al.* Expression of different tenascin isoforms in normal, hyperplastic and
765 neoplastic human breast tissues. *Int. J. Cancer* **52**, 688–692 (1992).
766 17. Mudge, J. M. *et al.* GENCODE 2025: reference gene annotation for human and mouse.
767 *Nucleic Acids Res.* **53**, D966–D975 (2025).
768 18. Veiga, D. F. T. *et al.* A comprehensive long-read isoform analysis platform and sequencing
769 resource for breast cancer. *Sci. Adv.* **8**, eabg6711 (2022).
770 19. Baldoni, P. L. *et al.* Dividing out quantification uncertainty allows efficient assessment of
771 differential transcript expression with edgeR. *Nucleic Acids Res.* **52**, e13 (2024).
772 20. Singh, N. P., Wu, E. Y., Fan, J., Love, M. I. & Patro, R. Tree-based differential testing using
773 inferential uncertainty for RNA-Seq. *bioRxiv* 2023.12.25.573288 (2025)
774 doi:10.1101/2023.12.25.573288.
775 21. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of
776 nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).

22. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).
23. Leung, S. K. *et al.* Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Rep.* **37**, (2021).
24. Paniagua, A. *et al.* Evaluation of strategies for evidence-driven genome annotation using long-read RNA-seq. *Genome Res.* **35**, 1053–1064 (2025).
25. Glinos, D. A. *et al.* Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**, 353–359 (2022).
26. Gao, Y. *et al.* ESPRESSO: Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. *Sci. Adv.* **9**, eabq5072 (2023).
27. Pardo-Palacios, F. J. *et al.* SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat. Methods* **21**, 793–797 (2024).
28. Mudge, J. M. *et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res.* **53**, D966–D975 (2025).
29. Peng, Q. *et al.* Impacts and mechanisms of alternative mRNA splicing in cancer metabolism, immune response, and therapeutics. *Mol. Ther.* **30**, 1018–1035 (2022).
30. Urbanski, L., Leclair, N. & Anczuków, O. Alternative-splicing defects in cancer: splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. *Wiley Interdiscip. Rev. RNA* **9**, e1476 (2018).
31. Zhang, Y., Qian, J., Gu, C. & Yang, Y. Alternative splicing and cancer: a systematic review. *Signal Transduct. Target. Ther.* **6**, 78 (2021).
32. Chisanga, D., Liao, Y. & Shi, W. Impact of gene annotation choice on the quantification of RNA-seq data. *BMC Bioinformatics* **23**, 107 (2022).

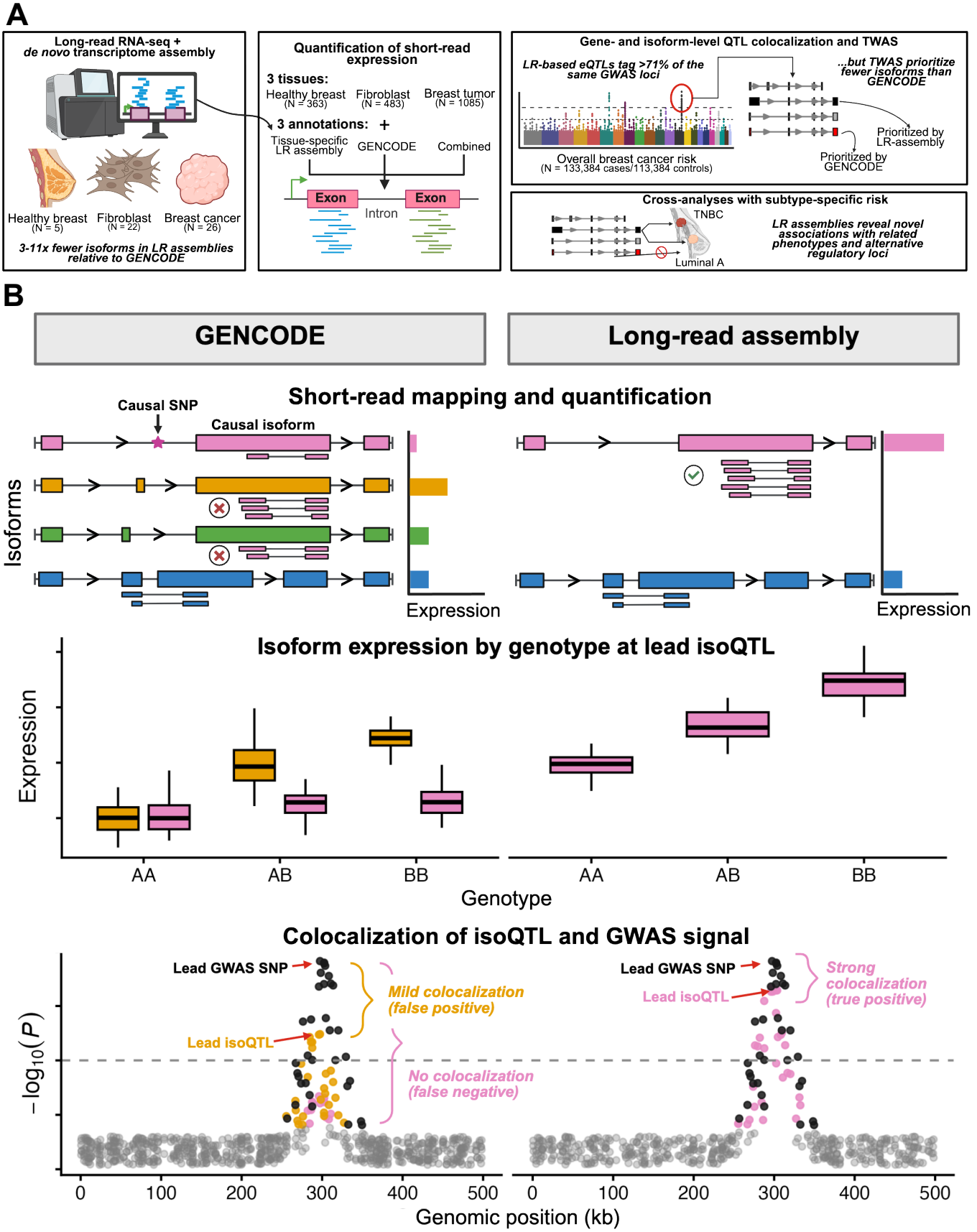
800 33. Head, S. T., Bresnahan, S. T., Cole, N., Wu, W. & Bhattacharya, A. Quantification method
801 affects replicability of eQTL analysis, colocalization, and TWAS. 2025.08.20.671303
802 Preprint at https://doi.org/10.1101/2025.08.20.671303 (2025).
803 34. Gu, G. J. *et al.* Role of Individual MARK Isoforms in Phosphorylation of Tau at Ser262 in
804 Alzheimer's Disease. *NeuroMolecular Med.* **15**, 458–469 (2013).
805 35. Debets, D. O. *et al.* Predicting treatment outcome using kinome activity profiling in HER2+
806 breast cancer biopsies. *iScience* **27**, 109858 (2024).
807 36. Natalia, M.-A., Alejandro, G.-T., Virginia, T.-V. J. & Alvarez-Salas, L. M. MARK1 is a Novel
808 Target for miR-125a-5p: Implications for Cell Migration in Cervical Tumor Cells. *MicroRNA*
809 **7**, 54–61.
810 37. González-Aguilera, C. & Askjaer, P. Dissecting the NUP107 complex: Multiple components
811 and even more functions. *Nucleus* **3**, 340–348 (2012).
812 38. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat.*
813 *Biotechnol.* **28**, 1045–1048 (2010).
814 39. Liang, Y., Nyasimi, F. & Im, H. K. Pervasive polygenicity of complex traits inflates false
815 positive rates in transcriptome-wide association studies. *bioRxiv* 2023.10.17.562831 (2024)
816 doi:10.1101/2023.10.17.562831.
817 40. Leeuw, C. de, Werme, J., Savage, J. E., Peyrot, W. J. & Posthuma, D. On the interpretation
818 of transcriptome-wide association studies. *PLOS Genet.* **19**, e1010921 (2023).
819 41. Gerlach, P. A., Milind, N., Spence, J. P. & Pritchard, J. K. High false sign rates in
820 transcriptome-wide association studies. 2025.12.19.695550 Preprint at
821 https://doi.org/10.64898/2025.12.19.695550 (2025).
822 42. Glinos, D. A. *et al.* Transcriptome variation in human tissues revealed by long-read
823 sequencing. *Nature* **608**, 353–359 (2022).

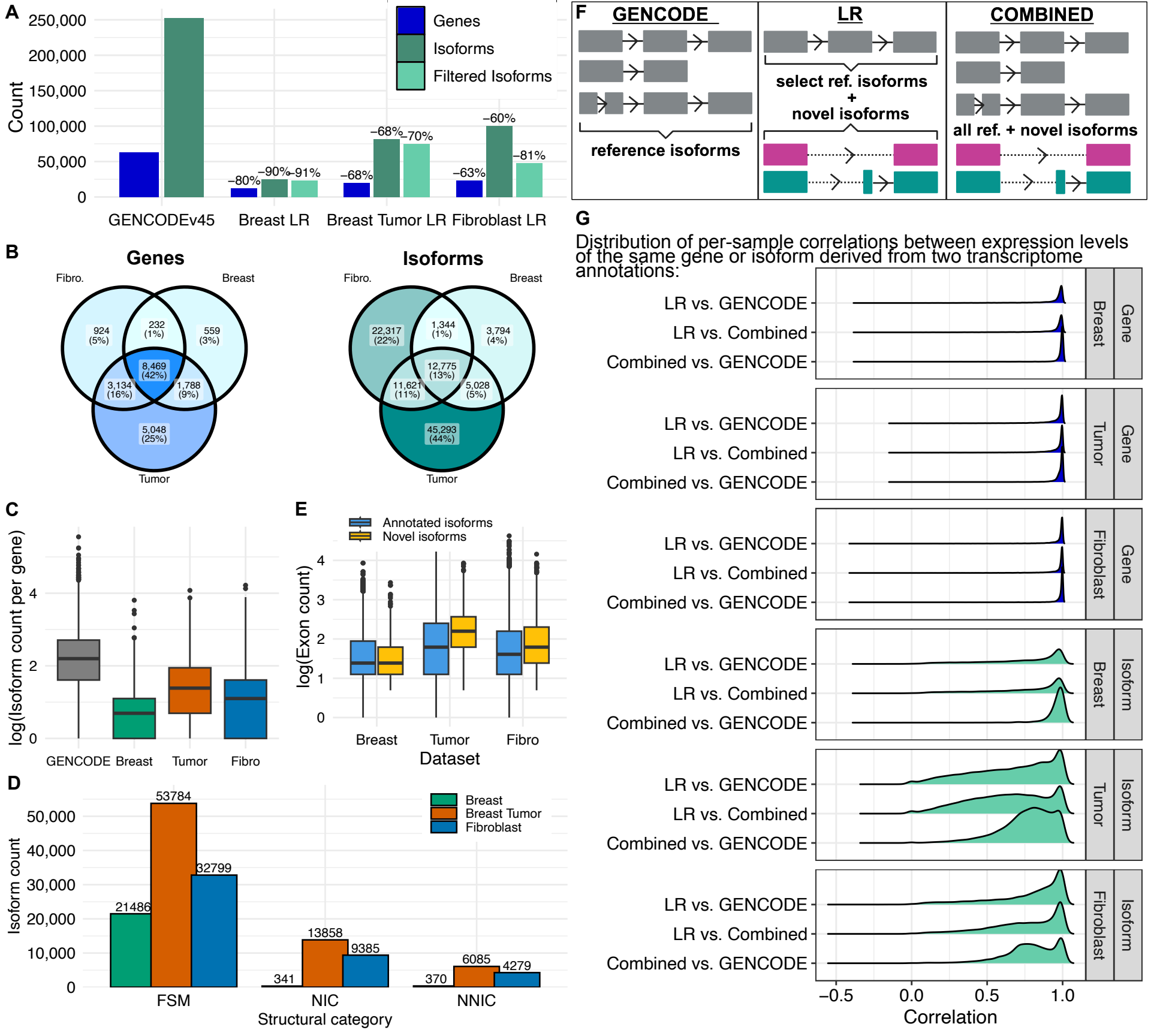
824 43. Transcriptome variation in human tissues revealed by long-read sequencing | Nature.
825 https://www.nature.com/articles/s41586-022-05035-y.
826 44. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
827 2079 (2009).
828 45. Guizard, S. *et al.* nf-core/isoseq: simple gene and isoform annotation with PacBio Iso-Seq
829 long-read sequencing. *Bioinformatics* **39**, btad150 (2023).
830 46. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
831 3100 (2018).
832 47. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences
833 for quality control in full-length transcriptome identification and quantification. *Genome Res.*
834 **28**, 396–411 (2018).
835 48. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human
836 genome. *Nature* **489**, 57–74 (2012).
837 49. Gruber, A. J. *et al.* A comprehensive analysis of 3' end sequencing data sets reveals novel
838 polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on
839 cleavage and polyadenylation. *Genome Res.* **26**, 1145–1159 (2016).
840 50. Herrmann, C. J. *et al.* PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3'
841 end sequencing. *Nucleic Acids Res.* **48**, D174–D179 (2020).
842 51. Moon, Y., Herrmann, C. J., Mironov, A. & Zavolan, M. PolyASite v3.0: a multi-species atlas
843 of polyadenylation sites inferred from single-cell RNA-sequencing data. *Nucleic Acids Res.*
844 **53**, D197–D204 (2025).
845 52. Noguchi, S. *et al.* FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4**,
846 170112 (2017).

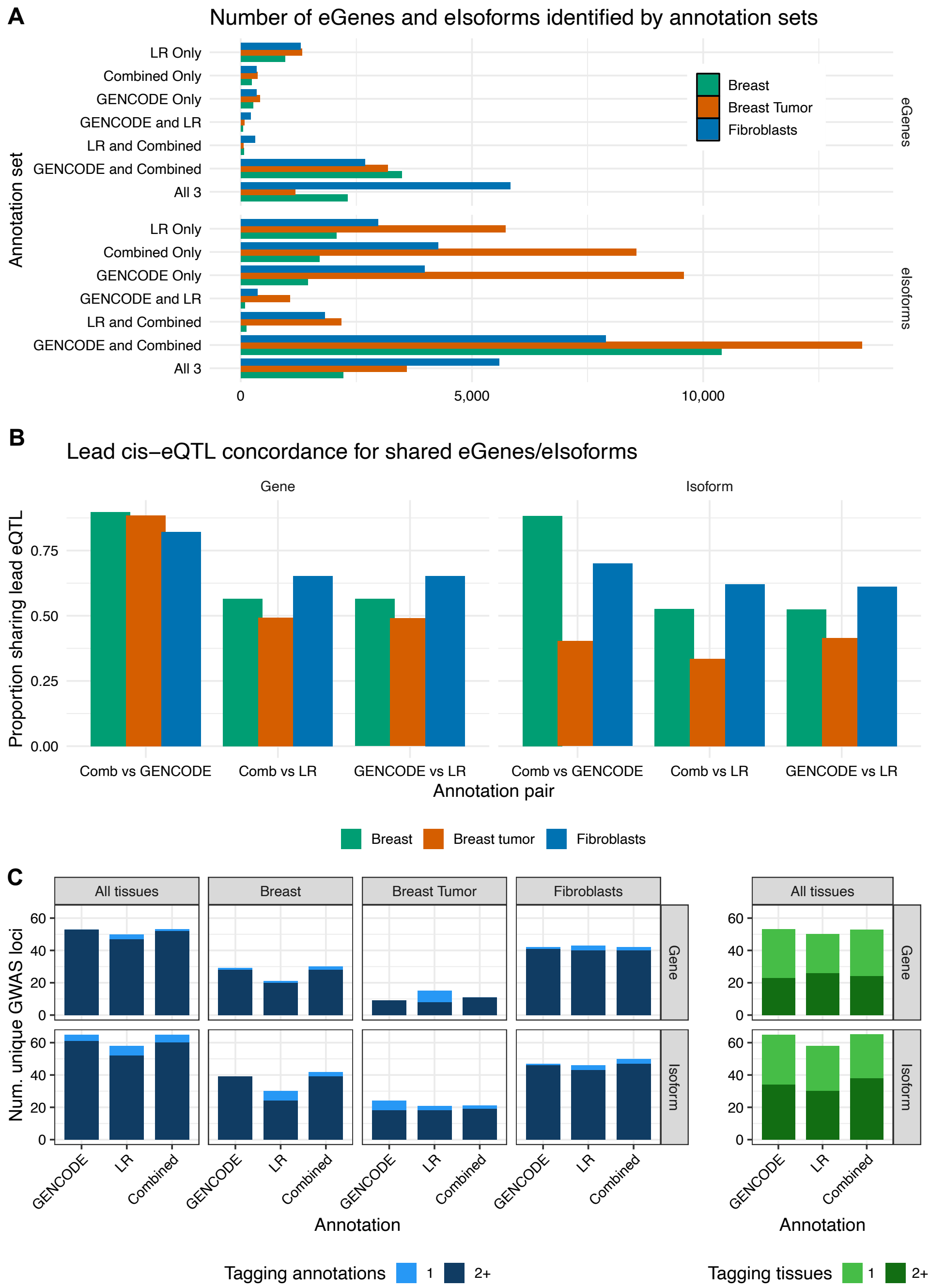
847 53. Bresnahan, S. T. *et al.* Long-read transcriptome assembly reveals vast transcriptional
848 complexity in the placenta associated with metabolic and endocrine function.
849 2025.06.26.661362 Preprint at https://doi.org/10.1101/2025.06.26.661362 (2025).
850 54. Kille, B., Garrison, E., Treangen, T. J. & Phillippy, A. M. Minmers are a generalization of
851 minimizers that enable unbiased local Jaccard estimation. *Bioinforma. Oxf. Engl.* **39**,
852 btad512 (2023).
853 55. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and
854 bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
855 56. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human
856 tissues. *Science* **369**, 1318–1330 (2020).
857 57. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
858 58. Scharpf, R. B., Irizarry, R. A., Ritchie, M. E., Carvalho, B. & Ruczinski, I. Using the R
859 Package crlmm for Genotyping and Copy Number Estimation. *J. Stat. Softw.* **40**, 1–32
860 (2011).
861 59. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium
862 panel. *Nat. Genet.* **48**, 1443–1448 (2016).
863 60. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation.
864 *Bioinformatics* **31**, 782–784 (2015).
865 61. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
866 data. *Bioinformatics* **30**, 2114–2120 (2014).
867 62. Picard Toolkit. *Broad Inst. GitHub Repos.* https://broadinstitute.github.io/picard/ (2019).
868 63. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for
869 multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
870 64. Love, M. I. *et al.* Tximeta: Reference sequence checksums for provenance identification in
871 RNA-seq. *PLoS Comput. Biol.* **16**, e1007664 (2020).

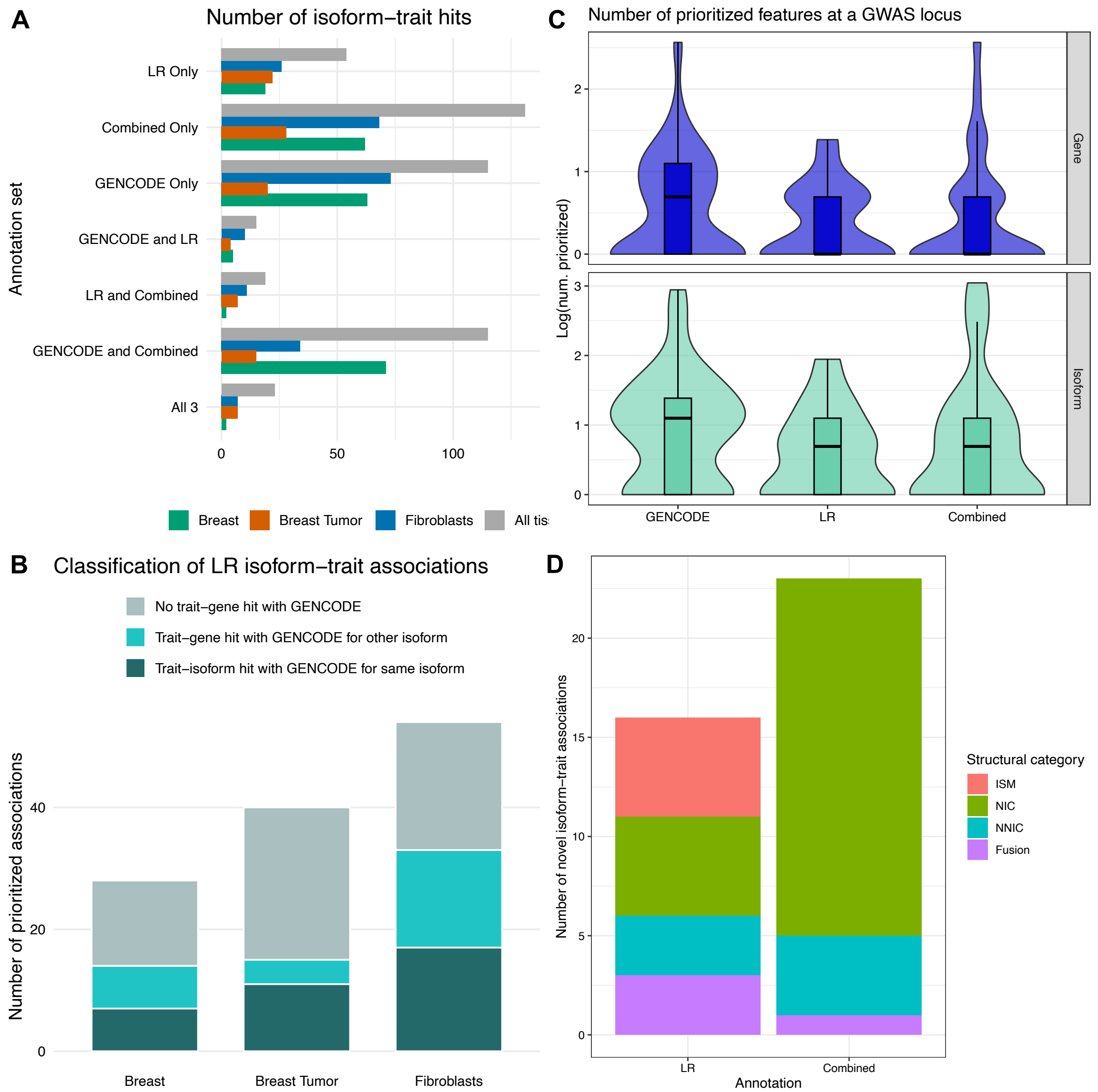
872 65. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for
873 differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140
874 (2010).
875 66. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
876 RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
877 67. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
878 analysis. *BMC Bioinformatics* **9**, 559 (2008).
879 68. Mostafavi, S. *et al.* Normalizing RNA-Sequencing Data by Modeling Hidden Covariates with
880 Prior Knowledge. *PLOS ONE* **8**, e68141 (2013).
881 69. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat.*
882 *Commun.* **8**, 15452 (2017).
883 70. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable
884 Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B*
885 *Stat. Methodol.* **82**, 1273–1300 (2020).
886 71. Zhang, H. *et al.* A mixed-model approach for powerful testing of genetic associations with
887 cancer risk incorporating tumor characteristics. *Biostatistics* **22**, 772–788 (2021).
888 72. Chatterjee, N. A Two-Stage Regression Model for Epidemiological Studies With Multivariate
889 Disease Classification Data. *J. Am. Stat. Assoc.* **99**, 127–138 (2004).
890 73. Chenevix-Trench, G. *et al.* An international initiative to identify genetic modifiers of cancer
891 risk in BRCA1 and BRCA2 mutation carriers: the Consortium of Investigators of Modifiers
892 of BRCA1 and BRCA2 (CIMBA). *Breast Cancer Res.* **9**, 104 (2007).
893 74. Mavaddat, N. *et al.* Pathology of breast and ovarian cancers among BRCA1 and BRCA2
894 mutation carriers: results from the Consortium of Investigators of Modifiers of BRCA1/2
895 (CIMBA). *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored*
896 *Am. Soc. Prev. Oncol.* **21**, 134–147 (2012).

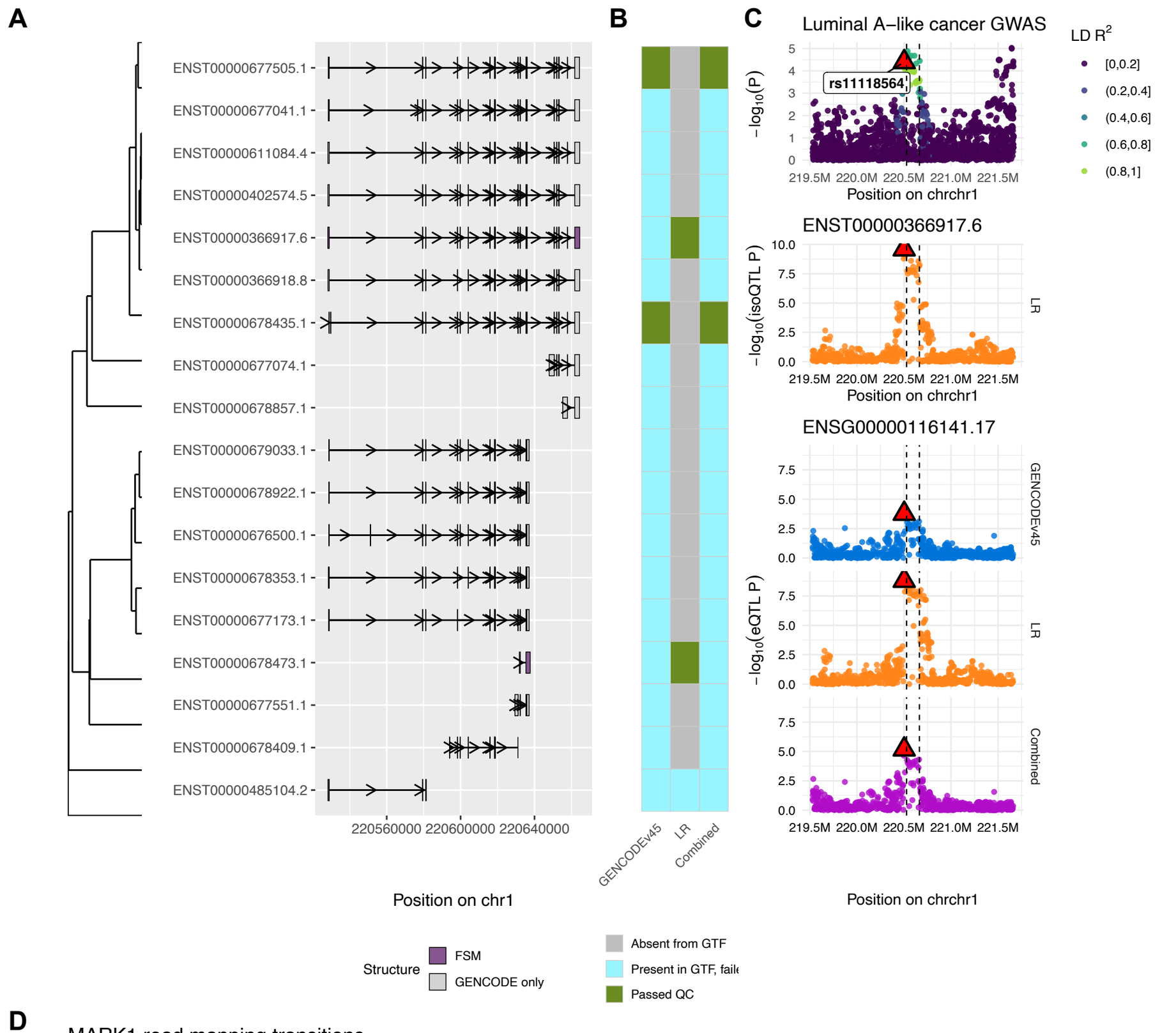
897 75. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast
898 cancer risk. *Nat. Genet.* **45**, 353–361 (2013).
899 76. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals
900 identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47**, 373–380 (2015).
901 77. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based
902 Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
903 78. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic
904 Association Studies Using Summary Statistics. *PLOS Genet.* **10**, e1004383 (2014).
905 79. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*
906 **45**, 1113–1120 (2013).
907

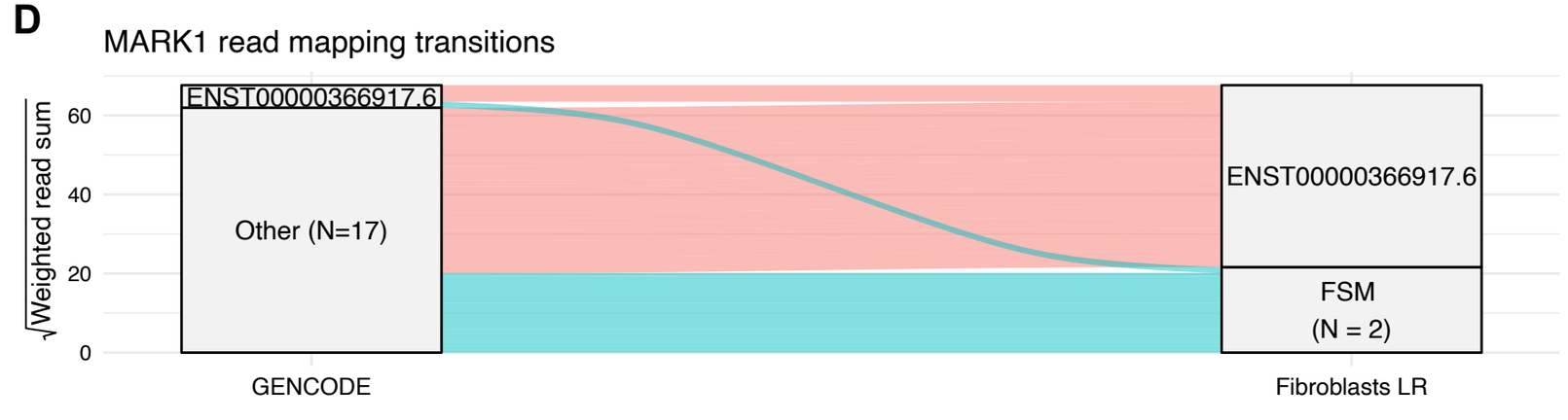

A
Long-read RNA-seq +
de novo transcriptome assembly
Healthy breast
(N = 5)
Fibroblast
(N = 22)
Breast cancer
(N = 26)
3-11x fewer isoforms in LR assemblies relative to GENCODE
Quantification of short-read expression
3 tissues:
Healthy breast
(N = 363)
Fibroblast
(N = 483)
Breast tumor
(N = 1085)
3 annotations: +
Tissue-specific LR assembly
GENCODE
Combined
Exon
Intron
Exon
Gene- and isoform-level QTL colocalization and TWAS
LR-based eQTLs tag >71% of the same GWAS loci
...but TWAS prioritize fewer isoforms than GENCODE
Overall breast cancer risk
(N = 133,384 cases/113,384 controls)
Prioritized by LR-assembly
Prioritized by GENCODE
Cross-analyses with subtype-specific risk
TNBC
Luminal A
LR assemblies reveal novel associations with related phenotypes and alternative regulatory loci
B
GENCODE
Long-read assembly
Short-read mapping and quantification
Causal SNP
Causal isoform
Isoforms
Expression
Expression
Isoform expression by genotype at lead isoQTL
Expression
AA
AB
BB
AA
AB
BB
Genotype
Colocalization of isoQTL and GWAS signal
Lead GWAS SNP
Mild colocalization (false positive)
Lead isoQTL
No colocalization (false negative)
Lead GWAS SNP
Lead isoQTL
Strong colocalization (true positive)
$-\log_{10}(P)$
0
100
200
300
400
500
0
100
200
300
400
500
Genomic position (kb)

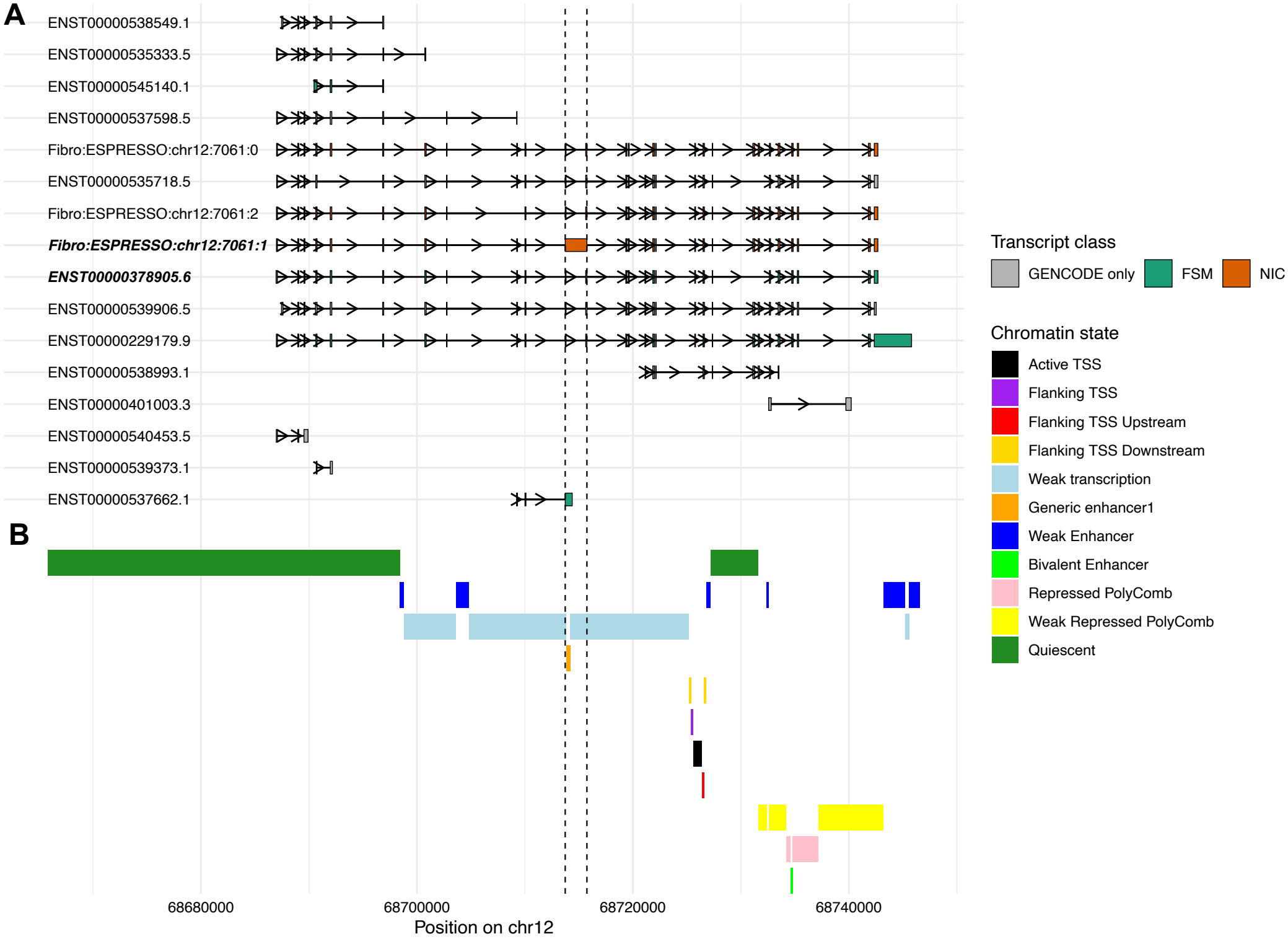

A
Count
250,000
200,000
150,000
100,000
50,000
0
Genes
Isoforms
Filtered Isoforms
−80%
−90%
−91%
−68%
−68%
−70%
−63%
−60%
−81%
GENCODEv45
Breast LR
Breast Tumor LR
Fibroblast LR
B
Genes
Fibro.
Breast
Tumor
924 (5%)
232 (1%)
559 (3%)
8,469 (42%)
3,134 (16%)
1,788 (9%)
5,048 (25%)
Isoforms
22,317 (22%)
1,344 (1%)
3,794 (4%)
12,775 (13%)
11,621 (11%)
5,028 (5%)
45,293 (44%)
C
log(Isoform count per gene)
GENCODE
Breast
Tumor
Fibro
D
Isoform count
Breast
Breast Tumor
Fibroblast
53784
32799
21486
13858
9385
341
6085
4279
370
FSM
NIC
NNIC
Structural category
E
Annotated isoforms
Novel isoforms
log(Exon count)
Breast
Tumor
Fibro
Dataset
F
GENCODE
LR
COMBINED
reference isoforms
select ref. isoforms
+
novel isoforms
all ref. + novel isoforms
G
Distribution of per-sample correlations between expression levels of the same gene or isoform derived from two transcriptome annotations:
LR vs. GENCODE
LR vs. Combined
Combined vs. GENCODE
Breast
Tumor
Fibroblast
Gene
Isoform
−0.5
0.0
0.5
1.0
Correlation


A
Number of eGenes and eIsoforms identified by annotation sets
Annotation set
LR Only
Combined Only
GENCODE Only
GENCODE and LR
LR and Combined
GENCODE and Combined
All 3
eGenes
eIsoforms
Breast
Breast Tumor
Fibroblasts
0
5,000
10,000
B
Lead cis−eQTL concordance for shared eGenes/eIsoforms
Gene
Isoform
Proportion sharing lead eQTL
0.00
0.25
0.50
0.75
Comb vs GENCODE
Comb vs LR
GENCODE vs LR
Annotation pair
Breast
Breast tumor
Fibroblasts
C
All tissues
Breast
Breast Tumor
Fibroblasts
Num. unique GWAS loci
0
20
40
60
Gene
Isoform
GENCODE
LR
Combined
Annotation
Tagging annotations
1
2+
Tagging tissues
1
2+


A
Number of isoform−trait hits
Annotation set
LR Only
Combined Only
GENCODE Only
GENCODE and LR
LR and Combined
GENCODE and Combined
All 3
0
50
100
Breast
Breast Tumor
Fibroblasts
All tis
C
Number of prioritized features at a GWAS locus
Log(num. prioritized)
Gene
Isoform
GENCODE
LR
Combined
B
Classification of LR isoform−trait associations
No trait−gene hit with GENCODE
Trait−gene hit with GENCODE for other isoform
Trait−isoform hit with GENCODE for same isoform
Number of prioritized associations
Breast
Breast Tumor
Fibroblasts
D
Number of novel isoform−trait associations
Annotation
LR
Combined
Structural category
ISM
NIC
NNIC
Fusion


A
ENST00000677505.1
ENST00000677041.1
ENST00000611084.4
ENST00000402574.5
ENST00000366917.6
ENST00000366918.8
ENST00000678435.1
ENST00000677074.1
ENST00000678857.1
ENST00000679033.1
ENST00000678922.1
ENST00000676500.1
ENST00000678353.1
ENST00000677173.1
ENST00000678473.1
ENST00000677551.1
ENST00000678409.1
ENST00000485104.2
220560000 220600000 220640000
Position on chr1
B
GENCODEv45
LR
Combined
C
Luminal A−like cancer GWAS
rs11118564
−log10(P)
Position on chrchr1
LD R2
[0,0.2]
(0.2,0.4]
(0.4,0.6]
(0.6,0.8]
(0.8,1]
ENST00000366917.6
−log10(isoQTL P)
ENSG00000116141.17
−log10(eQTL P)
GENCODEv45
LR
Combined
Position on chrchr1
Structure
FSM
GENCODE only
Absent from GTF
Present in GTF, faile
Passed QC


D
MARK1 read mapping transitions
√Weighted read sum
ENST00000366917.6
Other (N=17)
ENST00000366917.6
FSM
(N = 2)
GENCODE
Fibroblasts LR


A
ENST00000538549.1
ENST00000535333.5
ENST00000545140.1
ENST00000537598.5
Fibro:ESPRESSO:chr12:7061:0
ENST00000535718.5
Fibro:ESPRESSO:chr12:7061:2
Fibro:ESPRESSO:chr12:7061:1
ENST00000378905.6
ENST00000539906.5
ENST00000229179.9
ENST00000538993.1
ENST00000401003.3
ENST00000540453.5
ENST00000539373.1
ENST00000537662.1
B
68680000
68700000
68720000
68740000
Position on chr12
Transcript class
GENCODE only
FSM
NIC
Chromatin state
Active TSS
Flanking TSS
Flanking TSS Upstream
Flanking TSS Downstream
Weak transcription
Generic enhancer1
Weak Enhancer
Bivalent Enhancer
Repressed PolyComb
Weak Repressed PolyComb
Quiescent