Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer.

Stephanie N. Dorman[1], Coby Viner[2], Peter K. Rogan[1,2]*

Departments of Biochemistry[1] and Computer Science[2], Western University
*progan@uwo.ca

Supplemental Materials

SomaticSniper Supplementary Materials

Supplementary Methods – Variant Calling Methods

Two independent variant callers, Strelka[1] and SomaticSniper[2], were evaluated. The main analysis performed using results from Strelka, which has greater sensitivity and ability to detect subclonal mutations, by minimizing reporting of spurious variants and germline polymorphisms[3]. Additionally, the SomaticSniper methods and results are reported below.

Before running SomaticSniper, all DNA sequencing BAM files were realigned using the Genome Analysis Toolkit (GATK) Indel Realigner program [4]. In addition to default parameters, the `knownAlleles` parameter was used with the well-documented insertions/deletions (indels) files: `Mills_and_1000G_gold_standard.indels.b37.sites.vcf` [5] and `1000G_phase1.indels.b37.vcf` [6], available through the bioinformatic resource Galaxy [7, 8]. SomaticSniper data was then post-processed to only include variants with both mapping and somatic qualities of at least 40 (equivalent to running it with `-Q 40 -q 40`).

Supplementary Results – SomaticSniper Variant Calling Results

SomaticSniper variant predictions are summarized in Table 1. Notably, there were 1,208 variants from SomaticSniper that are predicted to affect both protein coding and splicing 594 genes. In the SomaticSniper data, mutations classified as both protein coding and splicing variants were found in 383 tumours, with 63 of these variants in *PASD1,* 61 in *PRSS3,* 52 in *NF1.* The variants in these genes, as well as others that were highly mutated, are the exact same genomic location and nucleotide change, suggesting that SomaticSniper reported higher numbers of SNPs[3] that were not annotated with dbSNP135 in >1% of the population, which was used to filter out common SNPs. There were 248 variants in 186 tumours from the SomaticSniper set that were classified as silent amino acid changes from ANNOVAR, but were revealed to affect splicing from the Shannon Pipeline predictions.

There was relatively low concordance between the two variant callers, which reported variant lists with less than 50% similarity. There were 21,112 protein coding and 1,811 splicing variants common to both Strelka and SomaticSniper. The predicted variants were compared to the previously reported TCGA Level 2 somatic mutations (Table 2). Strelka showed the highest concordance with TCGA mutations, reporting 82.1% of protein coding mutations, and 86.5% of the splicing variants. Conversely, SomaticSniper predicted 73.4% protein coding and 75.3% splicing variants reported by TCGA.

Both of the somatic variant callers we employed utilize Baysian methods to elucidate somatic event probabilities. Strelka and SomaticSniper were found to be the two best variant callers in a comparison by Roberts et al 2013. Additionally, these two are a valuable combination, in that SomaticSniper is useful to generate "a variety of candidate SNV sites without any particular drawbacks", although with a fair amount of false positives, while Strelka is least prone to returning germ-line polymorphisms. The relative stringency of Strelka was our main reason for performing most of our analyses with it, along with the fact that many of its candidates (at probability 0.2) were also returned by other callers. It is worth mentioning that different callers have been found to have poor correlations at the same sites; in particular, Strelka and SomaticSniper were found to have a 0.21 Pearson correlation coefficient in the abovementioned study. Our use of Veridical to validate splicing variants with functional evidence of the mutation significantly resolves the inconsistency between somatic variant callers (for this type of mutation).

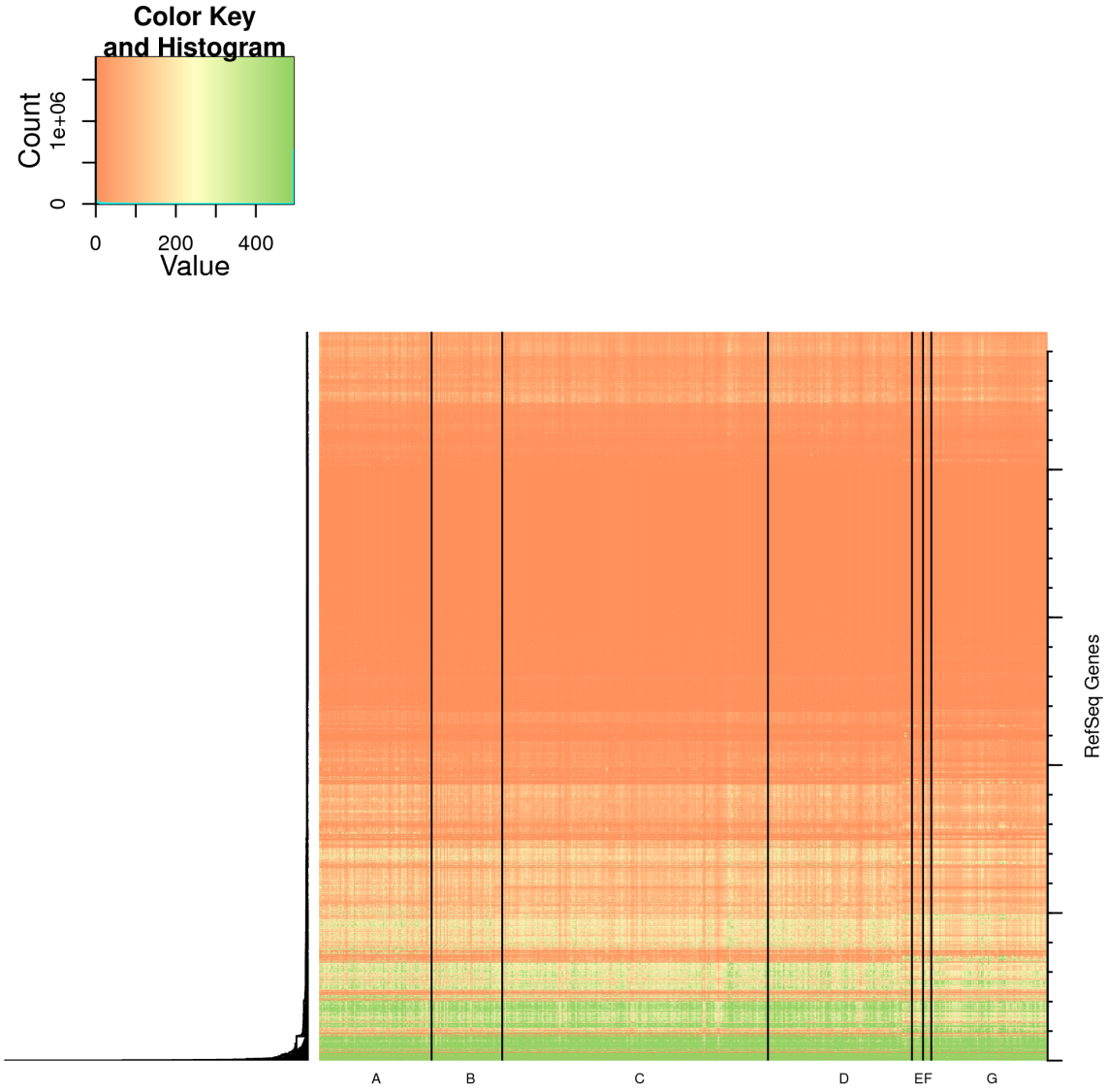
Supplementary Data 1 | Variant Summaries by Mutation Type

	SomaticSniper software
ANNOVAR protein coding variants	
Synonymous	23,458
Nonsynonymous	52,634
Stop gain or loss	2,127
Total protein coding variants	**78,219**
Shannon Pipeline splicing variants	
Cryptic	6,441
Inactivating	2,685
Leaky	10,648
Total splicing variants	**19,774**
Synonymous	248
Nonsynonymous	905
Stop gain or loss	55
Total	**1,208**
% Synonymous also splicing	1.0572%
% Nonsynonymous also splicing	1.7194%
% Stop gain or loss also splicing	2.5858%

Supplementary Data 2 | SomaticSniper Variants compared to TCGA Findings

TCGA Protein Coding Variants	**Total TCGA**	**TCGA predicted by SomaticSniper**
SNVs Validated	5,557	4,365 (77.3%)
SNVs Not Validated	18,197	13,380 (72.2%)
Indels Validated	125	N/A
Indels Not Validated	1,758	N/A
Total	**25,637**	**17,745 (73.4%)**
TCGA Splicing Variants		
SNVs Validated	87	70 (80.5%)
SNVs Not Validated	342	253 (74.0%)
Total	**429**	**323 (75.3%)**

Supplementary Figure S1. RNA-Seq Coverage Heat Map by Subtype. Heatmap depicting coverage per exonic base of TCGA RNA-Seq tumour and normal data. Expression based on RNA-Seq datasets is shown along the x-axis, with tumours first, ordered by subtype, followed by matched normal breast tissues. These categories are demarcated within the heatmap by black vertical lines, which correspond to the sample types: (A) basal-like; (B) HER2-enriched; (C) luminal A; (D) luminal B; (E) tumour, subtype not available; (F) normal-like tumor; and (G) normal control samples. The y-axis consists of all RefSeq genes (with major and minor tick marks every 5,000 and 1,000 genes, respectively), clustered to form a dendrogram, which is visible on the left side of the graph. Genes will low nominal expression levels were below minimum threshold read counts for analysis by Veridical.



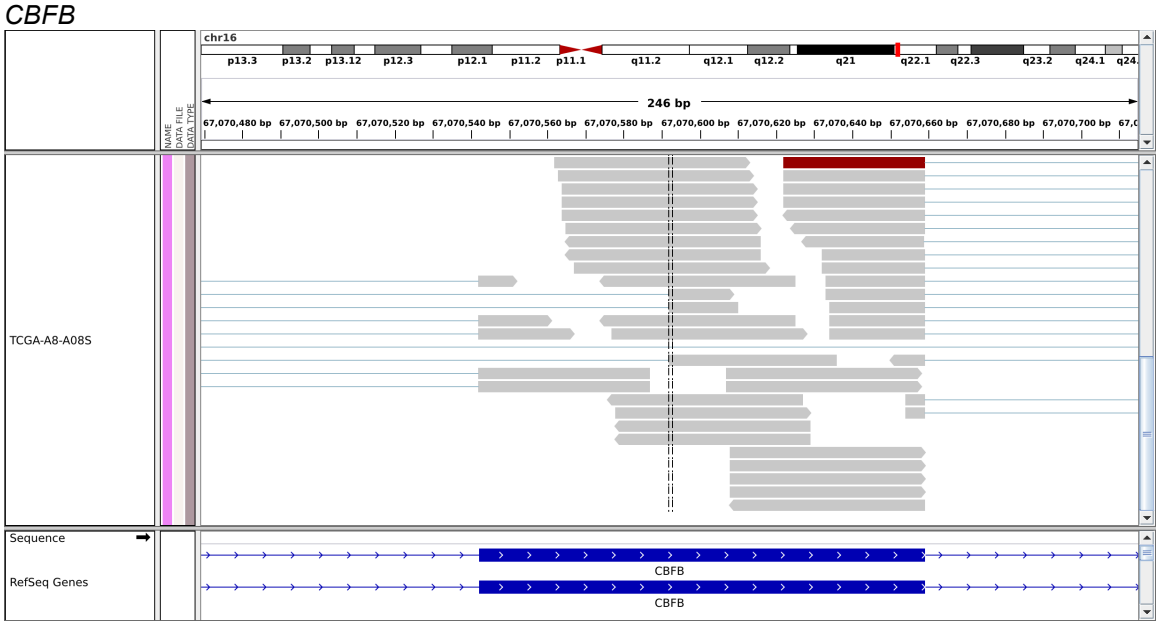
Supplementary Figure S2. Information theory based analysis and corresponding evidence demonstrating abnormal mRNA splicing in predicted mRNA splicing mutations. (A) Table indicates the TCGA sample identifier, variant, information analysis and statistical support for the mutation. (B) Screenshots from the Integrative Genomics Viewer (IGV) displaying junction-spanning reads that demonstrate cryptic splicing for mutations predicted by the Shannon Pipeline in the genes *CBFB, GATA3, PALB2,* and *ABL1*. The normal exonic structure is indicated by blue, with the thick bars representing exons, and the thin lines introns. RNA-Seq reads are shown in grey with the vertical dotted black lines demarcate the location of the cryptic splice site.

(A)

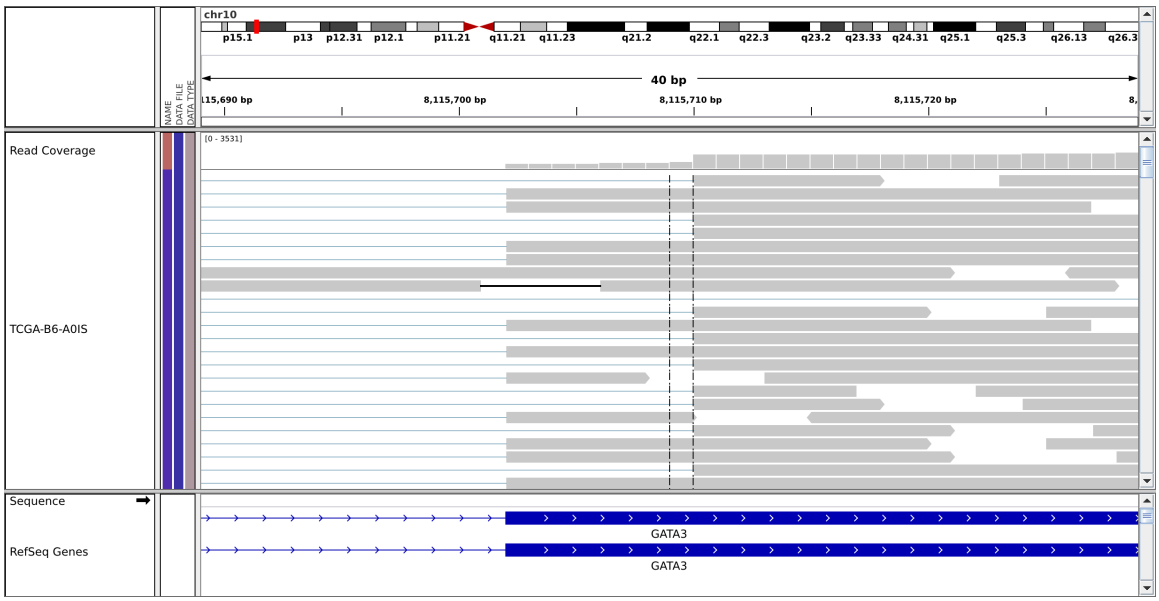
Sample	Gene	Splice Site Coordinate	Variant Coordinate	Ref/ Var	R_i - initial	R_i - final	ΔR_i	Cryptic Site Use P-Value	Exon Skipping P-Value
TCGA-A8-A08S	CBFB	chr16:67070591	chr16:67070577	G/T	5.6	7.5	1.9	< 0.005	0.12
TCGA-B6-A0I5	GATA3	chr10:8115709	chr10:8115702	A/C	4.2	5.9	1.7	< 0.005	NA
TCGA-B6-A0RT	PALB2	chr16:23637694	chr16:23637710	T/A	5.3	7.0	1.7	< 0.005	0.05
TCGA-B6-A0RV	ABL1	chr9:133750256	chr9:133750254	G/C	0.8	9.6	8.8	< 0.005	NA

(B)

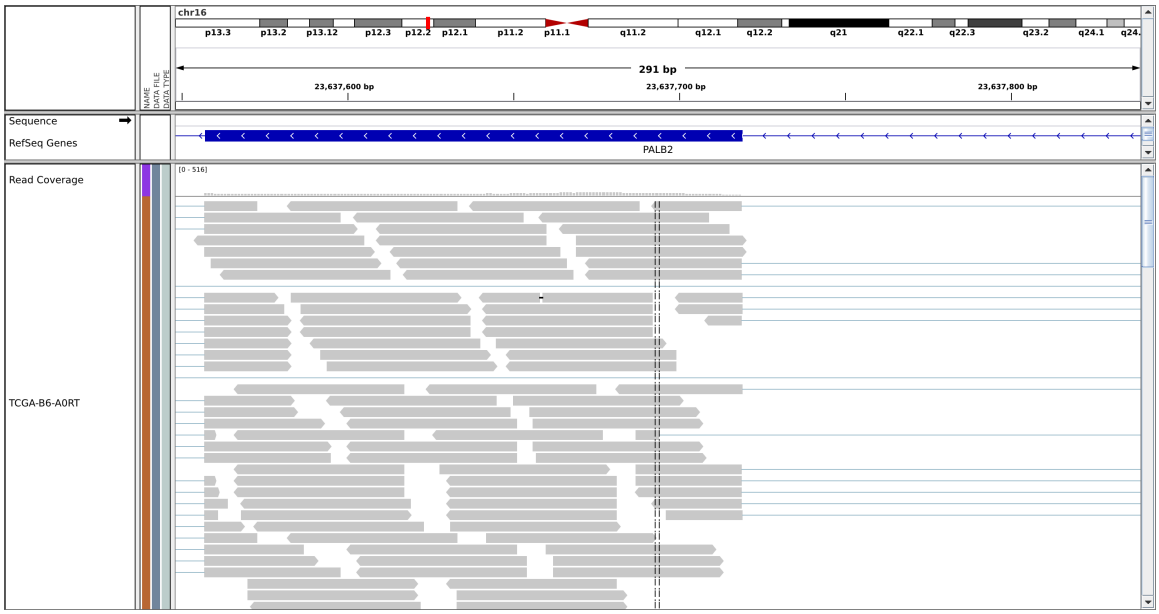
CBFB



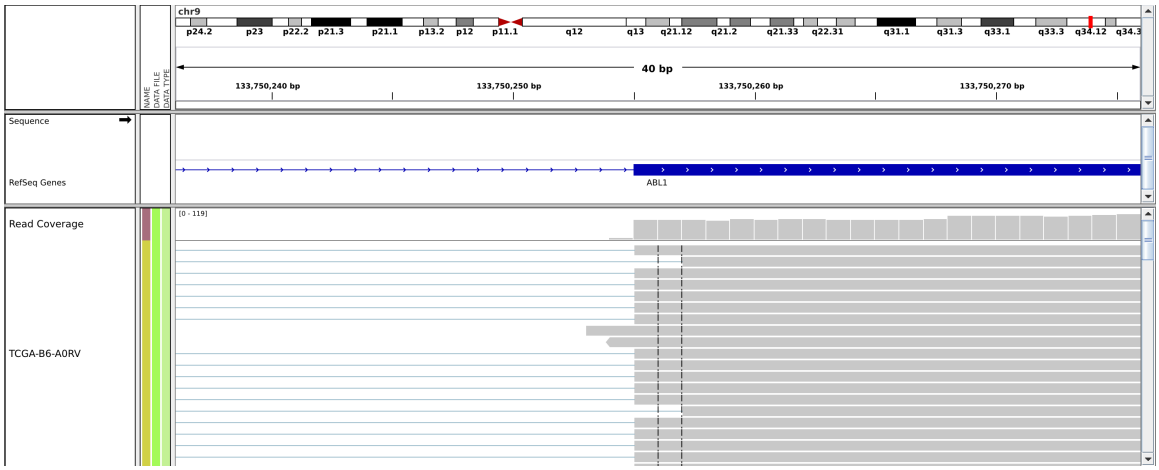
GATA3



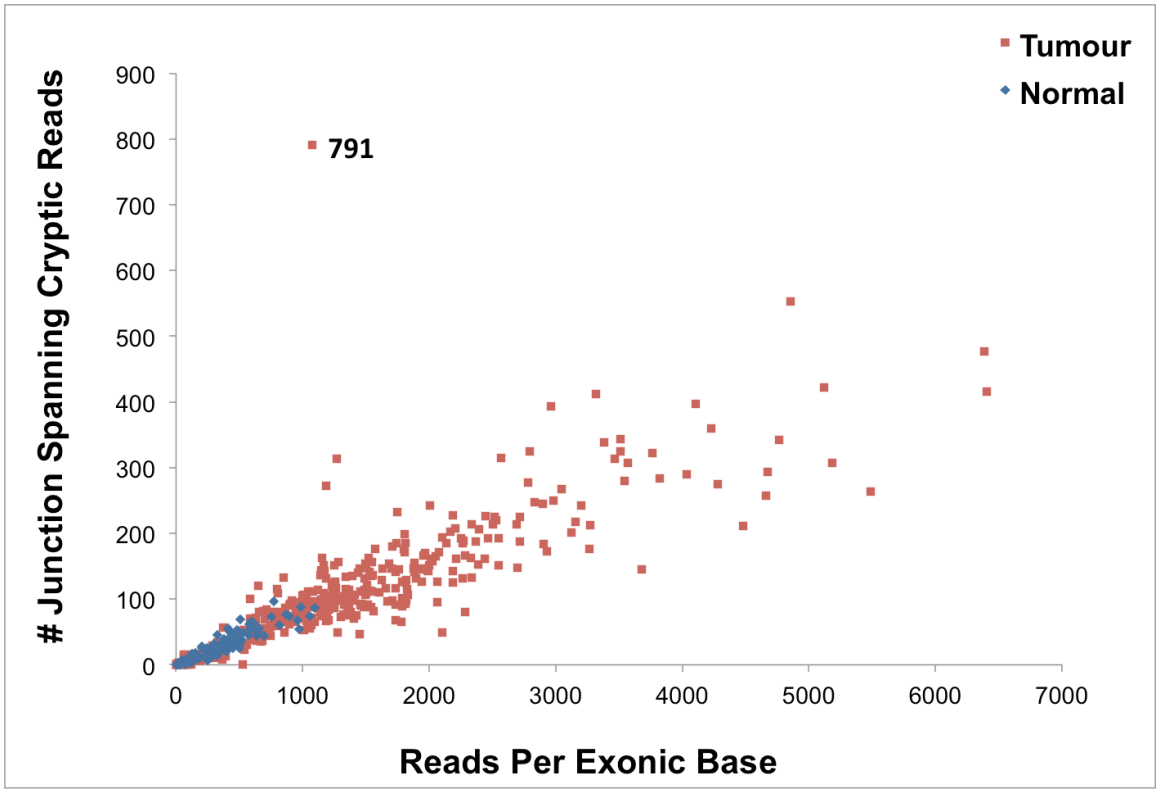
PALB2


chr16
p13.3 p13.2 p13.12 p12.3 p12.2 p12.1 p11.2 p11.1 q11.2 q12.1 q12.2 q21 q22.1 q22.3 q23.2 q24.1 q24.
291 bp
23,637,600 bp
23,637,700 bp
23,637,800 bp
NAME
DATA FILE
DATA TYPE
Sequence
RefSeq Genes
PALB2
Read Coverage
[0 - 516]
TCGA-B6-A0RT

ABL1

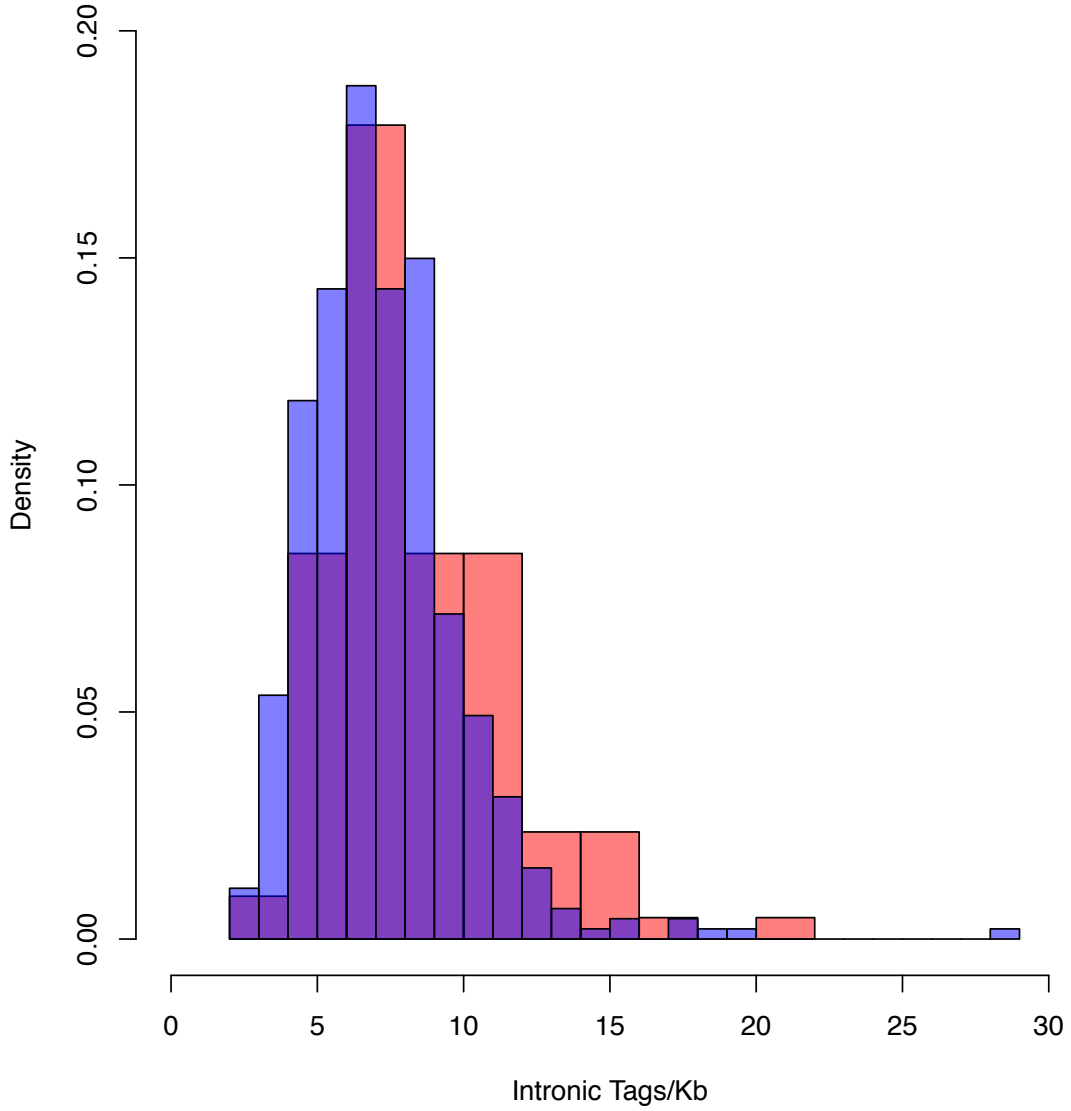

chr9
p24.2 p23 p22.2 p21.3 p21.1 p13.2 p12 p11.1 q12 q13 q21.12 q21.2 q21.33 q22.31 q31.1 q31.3 q33.1 q33.3 q34.12 q34.3
40 bp
133,750,240 bp
133,750,250 bp
133,750,260 bp
133,750,270 bp
NAME
DATA FILE
DATA TYPE
Sequence
RefSeq Genes
ABL1
Read Coverage
[0 - 119]
TCGA-B6-A0RV

Supplementary Figure S3. Junction-spanning, cryptic splicing read counts for GATA3 mutation (chr10: g.8115702A>C). The number of RNA-Seq reads per exonic base were plotted against the number of reads demonstrating *GATA3* cryptic splicing in the variant-containing tumours and controls. The variant containing tumour is indicated by the number of cryptic splicing reads (n = 791), tumours that do not contain this variant are in red, and normal controls are in blue. Cryptic splicing in the control samples likely occurs because the cryptic splice site (Ri = 4.2 bits) exceeds the strength of the natural splice site (Ri = 0.9 bits). However, the mutation further weakens the natural splice site (final Ri = 0.0 bits), while simultaneously strengthening the cryptic splice site (final Ri = 5.8 bits), which consistent with the RNA-Seq analysis.



Supplementary Figure S4. Intron Inclusion in tumour and normal breast genomes, based on RNA-Seq evidence. Histogram of the density of intronic sequence reads for normal (blue) and tumour (red) RNA-Seq samples. Purple shading represents overlapping components of the two density distributions. Intron inclusion was calculated with RSeQC's ReadDist script and RefSeq's gene annotation.

High levels of unspliced isoforms with intron inclusion were the most frequent outcome of mutations with significant effects on mRNA splicing. Nevertheless, when considering non-specific aberrant splicing across the transcriptome, the numbers of junction-spanning, intron inclusion reads present in normal and tumour samples did not significantly differ ($p > 0.1$). In fact, non-junction-spanning, intronic read-abundance reads of normal controls exceeded those of the tumour samples ($p < 0.01$; Supplementary Figure 4). This suggests that validation events in these tumour samples are not due solely to intron inclusion and aberrant mRNA splicing known to be present in breast tumours[9]. It is notable, however, that the levels of intronic inclusion for validated mutations significantly exceeded the read counts for all controls that did not contain these variants.



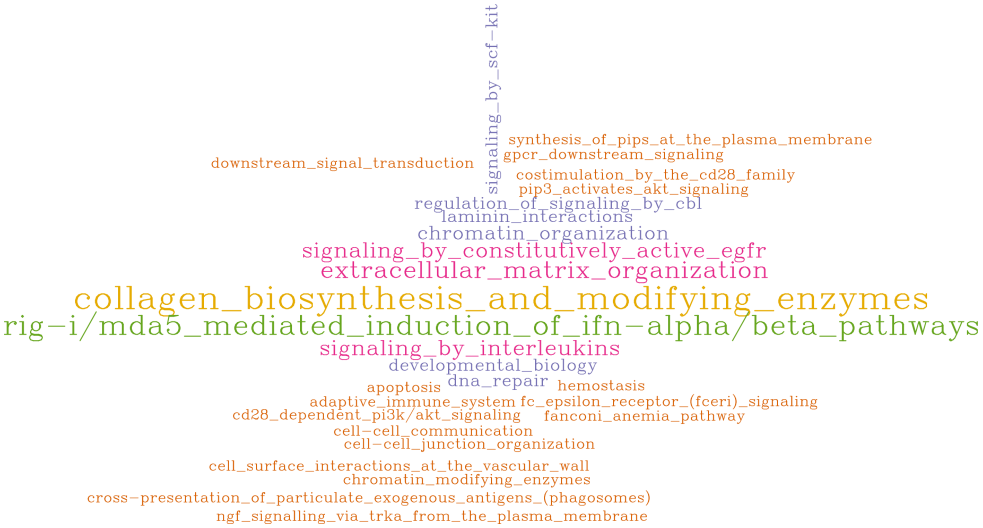
Supplementary Figure S5. Word Clouds of Overrepresented Pathways by Subtype. Word Clouds of overrepresented Reactome pathways for mutations in breast tumours, stratified by lymph node status (positive or negative) and by breast cancer subtype (basal-like, *HER2*-enriched, Luminal A, or Luminal B). The size of each word is proportional to its frequency in the abstracted list of overrepresented pathways.

Basal-like lymph node positive

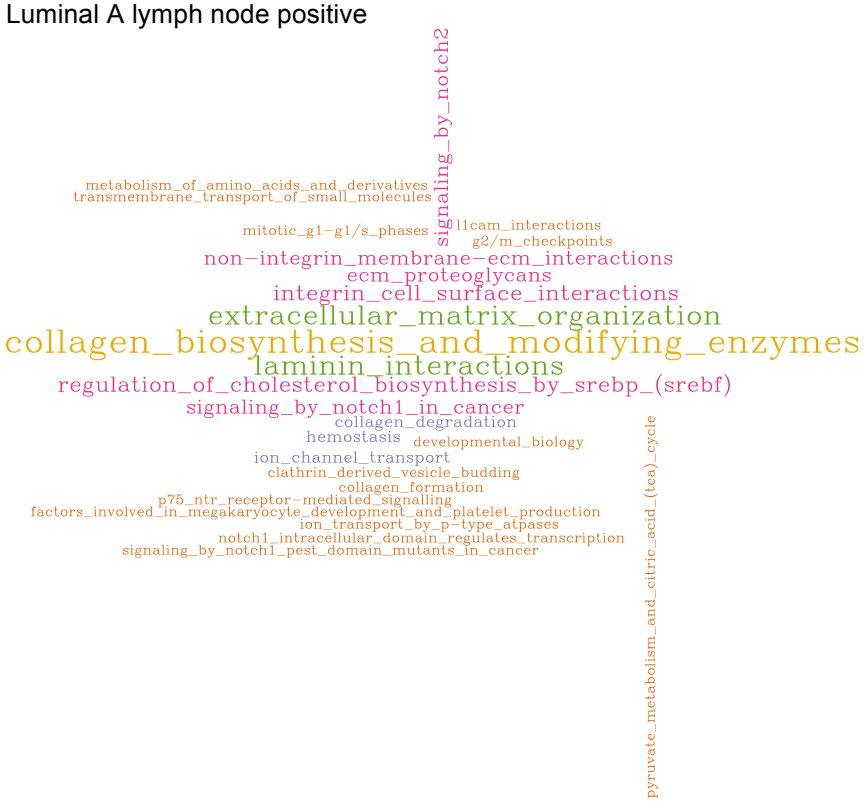
Basal-like lymph node negative

HER2-enriched lymph node positive

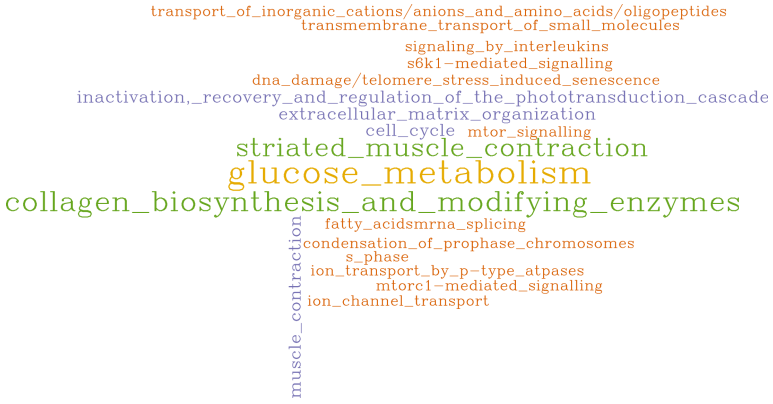
HER2-enriched lymph node negative



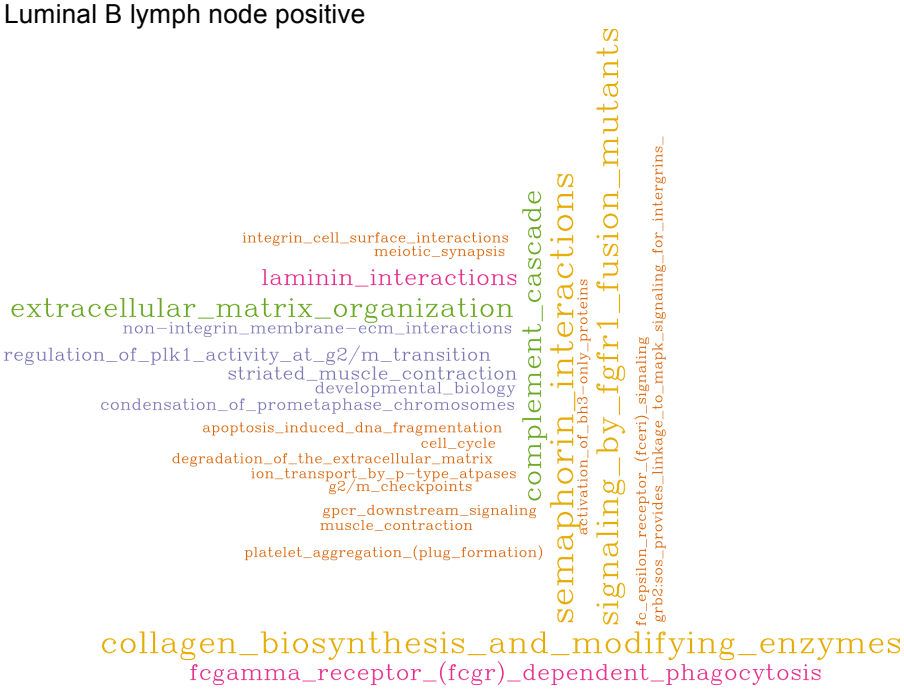
Luminal A lymph node positive



Luminal A lymph node negative

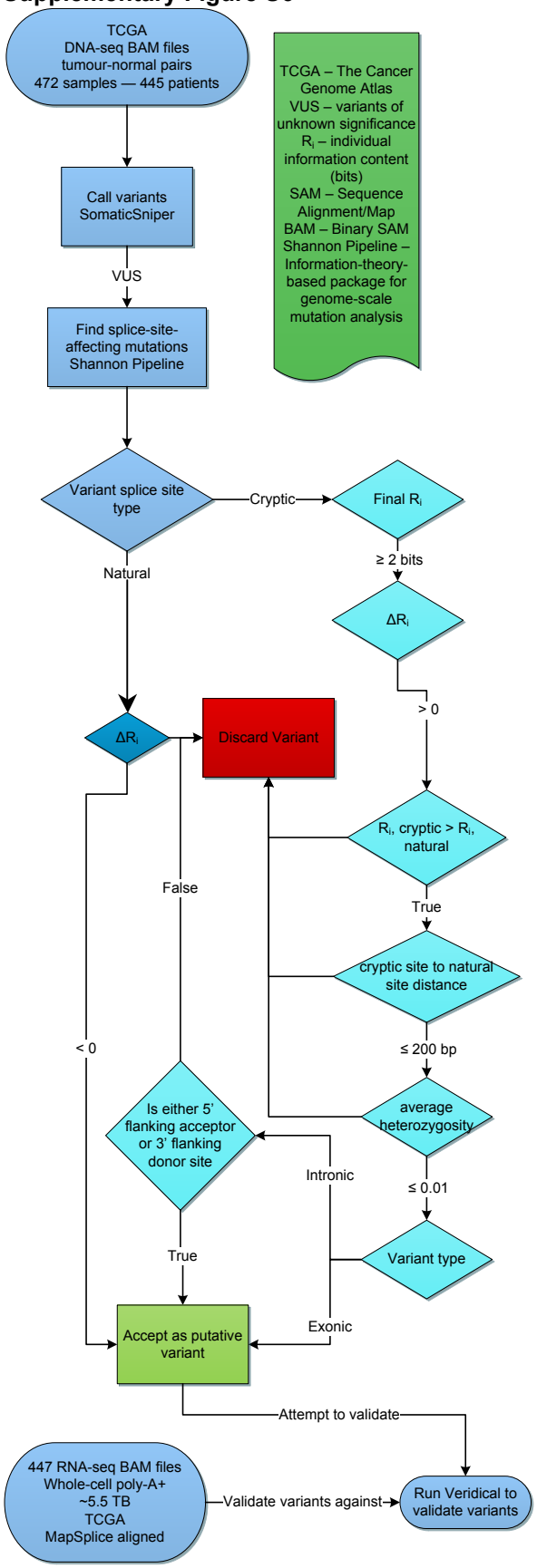


Luminal B lymph node positive



Luminal B lymph node negative

Supplementary Figure S6



Supplementary Figure S6. Flowchart indicating procedure for filtering splicing mutation variants. Shannon pipeline splicing variants output was filtered using the steps shown in this flowchart to identify those variants that are likely to cause aberrant splicing. Upon identifying variants with Strelka (or Somatic Sniper), the VCF files were submitted to the Shannon splicing mutation pipeline, then categorized as either mutations affecting natural splice sites (3' acceptor, or 5' donor) or cryptic splice site strengths. In a smal number of cases, both natural and cryptic splice sites were simultaneously altered. Natural sites that were predicted to be abolished were further considered. Predicted leaky splicing mutations were excluded from the present analysis, since the validation methods for such mutations has not yet been assessed. Aside from standard information theory-based mutation criteria, cryptic splicing mutation candidates were also filtered for proximity to the nearest neighboring natural splice site and population frequency. The filtered variant subset (n = 5,206) was used for all subsequent analyses.

References

1. Saunders, C. T. *et al*. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811-1817 (2012).

2. Larson, D. E. *et al*. Somaticsniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311-317 (2012).

3. Roberts, N. D. *et al*. A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics* **29**, 2223-2230 (2013).

4. McKenna, A. *et al*. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303 (2010).

5. https://usegalaxy.org/library_common/ldda_info?library_id=f9ba60baa2e6ba6d&show_deleted=False&cntrller=library&folder_id=709338dd4741c085&use_panels=False&id=a78cd737488dc2c1.

6. https://main.g2.bx.psu.edu/library_common/ldda_info?library_id=f9ba60baa2e6ba6d&show_deleted=False&cntrller=library&folder_id=709338dd4741c085&use_panels=False&id=93fcddc692e0986f#.

7. Giardine, B. *et al*. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15**, 1451-1455 (2005).

8. Blankenberg, D. *et al*. Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* **Chapter 19**, Unit 19.10.1-21 (2010).

9. Eswaran, J. *et al*. RNA sequencing of cancer reveals novel splicing alterations. *Scientific Reports* **3** (2013).