Splicing imbalances in basal-like breast cancer underpin perturbation of cell surface and oncogenic pathways and are associated with patients' survival

Filipe Gracio, Brian Burford, Patrycja Gazinska, Anca Mera, Aisyah Mohd Noor, Pierfrancesco Marra, Cheryl Gillett, Anita Grigoriadis, Sarah Pinder, Andrew Tutt^{*} and Emanuele de Rinaldis^{*}

*corresponding authors





IHC-based Classification	ER-	N=148 (84%)
	ER+	N=28 (16%)
	HER2+	N=30 (17%)
	ER- PR- HER2- (Triple-Negative)	N=93 (53%)
	Basal-Like	N=104 (59%)
	HER2	N=35 (20%)
Molecular Subtype	Luminal A	N=20 (11%)
	Luminal B	N=7 (4%)
	Normal-like	N=10 (6%)

Supplemental File 2



Supplemental File 3



Supplemental File 4

Classification: Basal / NBT



Number of Variables available to the model





Experimental protocol for generation of amplified full length cDNAs for all RNA isoforms of a given gene Supp

AURKA





Gel-like pseudo-image obtained from Bioanalyzer analysis of amplified isoforms. Green arrows highlight isoforms differential expressed between basal-like tumours and normal samples





Gel-like pseudo-image obtained from Bioanalyzer analysis of amplified isoforms. Green arrows highlight isoforms differential expressed between basal-like tumours and normal samples

BCL2-α

200 Kb				
60,800,000	60,850,000	60,900,000 RefSeq Genes	60,950,000	61,000,
	******************			BCL2



Gel-like pseudo-image obtained from Bioanalyzer analysis of amplified isoforms. Green arrows highlight isoforms differential expressed between basal-like tumours and normal samples







Gel-like pseudo-image obtained from Bioanalyzer analysis of amplified isoforms. Green arrows highlight isoforms differential expressed between basal-like tumours and normal samples

RRM2





Gel-like pseudo-image obtained from Bioanalyzer analysis of amplified isoforms. Green arrows highlight isoforms differential expressed between basal-like tumours and normal samples

TGFBR1



Gel-like pseudo-image obtained from Bioanalyzer analysis of amplified isoforms. Green arrows highlight isoforms differential expressed between basal-like tumours and normal samples







Gel-like pseudo-image obtained from Bioanalyzer analysis of amplified isoforms. Green arrows highlight isoforms differential expressed between basal-like tumours and normal samples

ZBTB16





Gel-like pseudo-image obtained from Bioanalyzer analysis of amplified isoforms. Green arrows highlight isoforms differential expressed between basal-like tumours and normal samples

CCR7





Gel-like pseudo-image obtained from Bioanalyzer analysis of amplified isoforms. Green arrows highlight isoforms differential expressed between basal-like tumours with respectively high and low splicing index

Paxillin Signaling



© 2000-2013 Ingenuity Systems, Inc. All rights reserved.



MORI_IMMATURE_B_LYMPHOCYTE_UP

			GNF2_CBFB MODULE_430 CELL_CORTEX GSE22886_UNSTIM_VS_IL15_STIM_NKCELL_UP YAGUE PRETUMOR DRUG RESISTANCE DN
			SCHAEFFER_PROSTATE_DEVELOPMENT_AND_CANCER_BOX1_UP V\$SP1_Q6_01 MORF_RAD23B MORF_TPR
			NUCLEOBASENUCLEOSIDENUCLEOTIDE_KINASE_ACTIVITY GSE13493_CD4INTCD8POS_VS_CD8POS_THYMOCYTE_UP GSE360_T_GONDII_VS_B_MALAYI_LOW_DOSE_MAC_DN KIM_ALL_DISORDERS_CALBI_CORR_UP
			BARIS_INTROID_CANCER_DN GTATGAT,MIR-154,MIR-487 GNF2_FGR MODULE_275 GSE1771_POLYIC_VS_CPG_24H_BMDM_DN
			GSE11864_CSF1_VS_CSF1_IFNG_IN_MAC_UP ELVIDGE_HYPOXIA_BY_DMOG_DN ACAACTT,MIR-382 LASTOWSKA_COAMPLIFIED_WITH_MYCN
			GSE37416_CTRL_VS_3H_F_TULARENSIS_LVS_NEUTROPHIL_UP NAKAYAMA_FGR2_TARGETS CATION_BINDING PERINUCLEAR_REGION_OF_CYTOPLASM CORECORY_SYNTLETIC_LETHAL_WITH_IMATINIP
			GSE11057_PBMC_VS_MEM_CD4_TCELL_UP GSE1721_0.5H_VS_24H_PAM3CSK4_BMDM_UP GSE392_BASOPHIL_VS_CENT_MEMORY_CD4_TCELL_DN GSE14308_TH17_VS_NAIVE_CD4_TCELL_DN
			PID_ARF6_TRAFFICKINGPATHWAY SCHLOSSER_SERUM_RESPONSE_UP PID_RHOA_REG_PATHWAY MIKKELSEN_MCV6_LCP_WITH_H3K4ME3
			COLLAGEN BURTON_ADIPOGENESIS_7 GSE13306_TREG_VS_TCONV_SPLEEN_UP LIEN_BREAST_CARCINOMA_METAPLASTIC
			GSE360_DC_VS_MAC_DN MANTOVANI_NFKB_TARGETS_UP DNA_PACKAGING LINDGREN_BLADDER_CANCER_HIGH_RECURRENCE
			EXTRACELLULAR_MATRIX GSE13493_DP_VS_CD4INTCD8POS_THYMOCYTE_UP GSE27786_LSK_VS_NEUTROPHIL_UP GSE27786_NKCELL_VS_NEUTROPHIL_UP
			TGAATGT,MIR-181A,MIR-181B,MIR-181C,MIR-181D ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE KOYAMA_SEMA3B_TARGETS_UP GSE22886_NAIVE_TCELL_VS_NKCELL_DN
			CATGTAA,MIR-496 GSE22886_CD4_TCELL_VS_BCELL_NAIVE_UP GSE27786_LIN_NEG_VS_BCELL_DN GSE3982_MAC_VS_TH2_UP GSE3982_MAC_VS_TH2_UP
			GSE27786_INKTCELL_VS_NEUTROPHIL_UP REACTOME_GPVI_MEDIATED_ACTIVATION_CASCADE V\$PEA3_Q6 GNF2_CDH11
			KESHELAVA_MULTIPLE_DRUG_RESISTANCE GNF2_LYN PID_INTEGRIN1_PATHWAY KEGG_LYSOSOME
			GSE339_EX_VIVO_VS_IN_CULTURE_CD8POS_DC_UP GSE339_CD8POS_VS_CD4CD8DN_DC_UP GNF2_PTX3 GSE3337_CTRL_VS_4H_IFNG_IN_CD8POS_DC_UP GSE29618_PBE_VS_DAYZ_ELU_VACCINE_PDC_DN
			GSE22886_NAIVE_TCELL_VS_MONOCYTE_UP GSE17721_CPG_VS_GARDIQUIMOD_16H_BMDM_DN GSE360_CTRL_VS_L_DONOVANI_MAC_DN GSE7460_CTRL_VS_TGFB_TREATED_ACT_FOXP3_MUT_TCONV_DN
			ATPASE_ACTIVITY CHANDRAN_METASTASIS_TOP50_DN UDAYAKUMAR_MED1_TARGETS_UP VANHARANTA_UTERINE_FIBROID_UP
			GSE3982_BCELL_VS_TH2_UP REACTOME_NCAM_SIGNALING_FOR_NEURITE_OUT_GROWTH GSE17721_POLYIC_VS_CPG_16H_BMDM_UP GSE26669_CD4_VS_CD8_TCELL_IN_MLR_COSTIM_BLOCK_DN PID_SVIDECON_1_REATUMAY
			REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION REACTOME_COLLAGEN_FORMATION V\$PITX2_02 PROTEINACEOUS_EXTRACELLULAR_MATRIX
			PID_TCR_PATHWAY SHAFFER_IRF4_TARGETS_IN_ACTIVATED_B_LYMPHOCYTE TMTCGCGANR_UNKNOWN MORF_CDK2
			GSE22886_NAIVE_CD8_TCELL_VS_DC_UP GSE380_T_GONDII_VS_B_MALAYI_HIGH_DOSE_DC_DN GSE380_HIGH_DOSE_B_MALAYI_VS_M_TUBERCULOSIS_DC_UP GSE39820_IL1B_IL6_VS_IL1B_IL6_IL23A_TREATED_CD4_TCELL_DN
			GSE9006_HEALTHY_VS_TYPE_2_DIABETES_PBMC_AT_DX_UP ACTGAAA,MIR-30A-3P,MIR-30E-3P KEGG_EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION GSE17721_LPS_VS_POLYIC_1H_BMDM_DN
			GSE3982_NEUTROPHIL_VS_CENT_MEMORY_CD4_TCELL_DN GSE17721_0.5H_VS_24H_GARDIQUIMOD_BMDM_DN TIEN_INTESTINE_PROBIOTICS_2HR_DN GSE27786_CD8_TCELL_VS_ERYTHROBLAST_UP
			GSE339_UD8PUS_VS_CD4CD8DN_DC_IN_CULTURE_DN CAGCACT,MIR-512-3P TOOKER_GEMCITABINE_RESISTANCE_UP GSE17721_PAM3CSK4_VS_GADIQUIMOD_8H_BMDM_UP GSE27786_LSK_VS_ERYTHRORI &ST_1IP
			V\$E2F_Q2 V\$USF_Q6 PROTEIN_LOCALIZATION NUCLEOCYTOPLASMIC_TRANSPORT
			GSE11924_TFH_VS_TH1_CD4_TCELL_UP GSE6269_HEALTHY_VS_STREP_AUREUS_INF_PBMC_UP GSE17721_POLYIC_VS_GARDIQUIMOD_24H_BMDM_DN SCHLOSSER_SERUM_RESPONSE_AUGMENTED_BY_MYC
			GTGTTAC,MIR-194 GGTGTGT,MIR-329 V\$CETS1P54_01 MORF_TERF1
			MORF_UBE2I GCM_DDX5 GNF2_INPP5D GSE15930_STIM_VS_STIM_AND_TRICHOSTATINA_72H_CD8_T_CELL_UP
			GSE17721_12H_VS_24H_POLYIC_BMDM_DN GSE27766_LSK_VS_CD4_TCELL_DN GSE29618_PRE_VS_DAY7_POST_TIV_FLU_VACCINE_PDC_DN GINESTIER_BREAST_CANCER_ZNF217_AMPLIFIED_DN
			WATANABE_RECTAL_CANCER_RADIOTHERAPY_RESPONSIVE_UP TTGGAGA,MIR-515-5P,MIR-519E MORF_CTBP1 MORF_ACP1 MORF_PP2CA
			GNF2_CASP4 PROTEIN_TARGETING GSE22886_NAIVE_CD4_TCELL_VS_DC_UP GSE34205_HEALTHY_VS_RSV_INF_INFANT_PBMC_UP
			GSE3982_NEUTROPHIL_VS_EFF_MEMORY_CD4_TCELL_DN GNF2_VAV1 BIOCARTA_MAPK_PATHWAY BIOCARTA_INTEGRIN_PATHWAY
			REGG_GLYCEROLIPID_METABOLISM REACTOME_TRANS_GOLGI_NETWORK_VESICLE_BUDDING GSE15324_NAIVE_VS_ACTIVATED_ELF4_KO_CD8_TCELL_DN GSE14308_TH1_VS_TH17_DN BIOCARTA_MTOR_PATHWAY
			IRITANI_MAD1_TARGETS_DN KEGG_VIBRIO_CHOLERAE_INFECTION REACTOME_GLUCOSE_METABOLISM PID_NFAT_3PATHWAY
			GTGGTGA,MIR-197 AGTCTTA,MIR-499 CAGCAGG,MIR-370 YGCGYRCGC_UNKNOWN
			MORF_RPA1 MORF_SP3 GCM_RAD21 GCM_MAX
			GNF2_PAK2 PROTEIN_OLIGOMERIZATION CELLULAR_PROTEIN_COMPLEX_DISASSEMBLY NUCLEAR_MEMBRANE PROTEIN_DOMAIN_SPECIEIC_BINDING
			GSE17721_LPS_VS_CPG_1H_BMDM_UP GSE2886_NAIVE_CD8_TCELL_VS_NEUTROPHIL_UP GSE26928_CENTR_MEMORY_VS_CXCR5_POS_CD4_TCELL_DN GSE2766_BCELL_VS_ERVTHROBLAST_UP
			GSE3982_MAST_CELL_VS_BCELL_DN GSE3982_NEUTROPHIL_VS_BCELL_DN OUYANG_PROSTATE_CANCER_PROGRESSION_UP BLALOCK_ALZHEIMERS_DISEASE_INCIPIENT_DN
			MELLMAN_TUT1_TARGETS_DN BIOCARTA_RHO_PATHWAY MODULE_248 ATGCAGT,MIR-217 PID_CDRTCRPATHWAY
			MODULE_32 GSE360_L_DONOVANI_VS_M_TUBERCULOSIS_DC_UP GSE17721_CTRL_VS_GARDIQUIMOD_24H_BMDM_UP REACTOME_TRIF_MEDIATED_TLR3_SIGNALING
			GSE27786_LSK_VS_CD8_TCELL_DN V\$HNF4_01_B MOREIRA_RESPONSE_TO_TSA_UP V\$STAT3_02
			V\$AF2_U6_U1 NIKOLSKY_BREAST_CANCER_17Q21_Q25_AMPLICON CELL_FRACTION GSE20366_CD103_POS_VS_NEG_TREG_KLRG1NEG_DN CCTGTGA,MIR-513
			GSE36392_TYPE_2_MYELOID_VS_MAC_IL25_TREATED_LUNG_DN ACACTGG,MIR-199A,MIR-199B V\$CACBINDINGPROTEIN_O6 GAVIN_FOXP3_TARGETS_CLUSTER_P3
			GSE13306_TREG_VS_TCONV_SPLEEN_DN VANLOO_SP3_TARGETS_DN ATATGCA,MIR-448 IVANOVA_HEMATOPOIESIS_INTERMEDIATE_PROGENITOR GSE1721_LPS_VS_CPG_4H_BMDM_DN
			PID_VEGFR1_2_PATHWAY FIGUEROA_AML_METHYLATION_CLUSTER_7_UP GSE360_DC_VS_MAC_UP GSE22886_NAIVE_CD4_TCELL_VS_48H_ACT_TH1_UP
			GSE/460_FOXP3_MUT_VS_WT_ACT_WITH_TGFB_TCONV_UP REACTOME_SIGNALING_BY_INSULIN_RECEPTOR BIOCARTA_AT1R_PATHWAY MORF_FDXR MORF_UBE2N
			GCM_ACTG1 GCM_DFFA GNF2_MSN NUCLEAR_IMPORT
			PROTEIN_IMPORT_INTO_NUCLEUS DNA_HELICASE_ACTIVITY HELICASE_ACTIVITY ATP_DEPENDENT_DNA_HELICASE_ACTIVITY ATPASE_ACTIVITY_COURTED
			GSE360_DC_VS_MAC_B_MALAYI_HIGH_DOSE_UP FAELT_B_CLL_WITH_VH3_21_UP MORF_G22P1 WATANABE_ULCERATIVE_COLITIS_WITH_CANCER_DN
			ATAGGAA,MIR-202 OHM_EMBRYONIC_CARCINOMA_UP GSE16522_ANTI_CD3CD28_STIM_VS_UNSTIM_MEMORY_CD8_TCELL_UP WATANABE_RECTAL_CANCER_RADIOTHERAPY_RESPONSIVE_DN GSE22965_NAME_COR_TCTL_NEX_NOVEMENT
			GSE22886_NAIVE_CD8_TCELL_VS_MONOCYTE_UP TGCTTTG,MIR-330 GSE17580_UNINFECTED_VS_S_MANSONI_INF_TEFF_UP CEBALLOS_TARGETS_OF_TP53_AND_MYC_DN V\$DR3_Q4
			GSE7460_CD8_TCELL_VS_CD4_TCELL_ACT_UP GSE3982_MAST_CELL_VS_DC_UP GSE7460_WT_VS_FOXP3_HET_ACT_WITH_TGFB_TCONV_DN GTPASE_REGULATOR_ACTIVITY
			REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLECULES GSE1460_DP_THYMOCYTE_VS_NAIVE_CD4_TCELL_ADULT_BLOOD_DN GUANYL_NUCLEOTIDE_EXCHANGE_FACTOR_ACTIVITY MEMBRANE_FRACTION GSE22886_NAIVE_CD4_TCELL_VS_NONCOURS_VT
			UEDA_PERIFERAL_CLOCK GSE13738_RESTING_VS_BYSTANDER_ACTIVATED_CD4_TCELL_UP SHEN_SMARCA2_TARGETS_UP JOHNSTONE_PARVB_TARGETS_1_DN
			COULOUARN_TEMPORAL_TGFB1_SIGNATURE_UP GILDEA_METASTASIS CACCAGC,MIR-138 TIEN_INTESTINE_PROBIOTICS_24HR_DN
			GOLGI_APPARATUS_PART GSE39820_CTRL_VS_TGFBETA3_IL6_CD4_TCELL_DN ASTON_MAJOR_DEPRESSIVE_DISORDER_DN OUILLETTE_CLL_13014_DELETION_DN YOSHIMURA_MAPK8_TARGETS_DN
			GSE3982_DC_VS_EFF_MEMORY_CD4_TCELL_DN DACOSTA_UV_RESPONSE_VIA_ERCC3_UP GSE16522_MEMORY_VS_NAIVE_ANTI_CD3CD28_STIM_CD8_TCELL_DN GGGACCA,MIR-133A,MIR-133B
			SHIPP_DLBCL_VS_FOLLICULAR_LYMPHOMA_UP GTGCAAT,MIR-25,MIR-32,MIR-32,MIR-363,MIR-367 CACTTTG,MIR-520G,MIR-520H GSE14350_IL2RB_KO_VS_WT_TREG_UP
			GSE20366_1 HEG_VS_NAIVE_CD4_TCELL_HOMEOSTATIC_CONVERSION_UP GABRIELY_MIR21_TARGETS SESTO_RESPONSE_TO_UV_C8 GSE31082_DN_VS_CD4_SP_THYMOCYTE_DN TGACAGNY_V\$MEIS1 01
			LI_INDUCED_T_TO_NATURAL_KILLER_UP MODULE_324 GSE17580_TREG_VS_TEFF_DN GSE10325_CD4_TCELL_VS_BCELL_UP
			CAMP_UP.V1_UP V\$MAZ_Q6 GSE360_DC_VS_MAC_B_MALAYI_LOW_DOSE_UP GSE24634_IL4_VS_CTRL_TREATED_NAIVE_CD4_TCELL_DAY7_DN
			OHM_METHYLATED_IN_ADULT_CANCERS AGCATTA,MIR-155 KENNY_CTNNB1_TARGETS_DN GSE22886_NAIVE_VS_MEMORY_TCELL_DN
			GSE13738_TCR_VS_BYSTANDER_ACTIVATED_CD4_TCELL_UP INTRACELLULAR_TRANSPORT GSE17721_CTRL_VS_LPS_12H_BMDM_UP ASTIER_INTEGRIN_SIGNALING
			GSE360_CTRL_VS_B_MALAYI_HIGH_DOSE_MAC_DN ATGAAGG,MIR-205 WELCSH_BRCA1_TARGETS_DN GSE27786_BCELL_VS_NKCELL_DN GSE31082_CD4_VS_CD8_SP_THYMOCYTE_UD
			GSE31002_CD4_VS_CD8_SP_IHYMOCYTE_UP GSE32423_MEMORY_VS_NAIVE_CD8_TCELL_DN MARTINEZ_RB1_AND_TP53_TARGETS_DN ACTGCCT,MIR-34B MARTINEZ_TP53_TARGETS_DN
			GTATTAT,MIR-369-3P RASHI_RESPONSE_TO_IONIZING_RADIATION_5 BILD_MYC_ONCOGENIC_SIGNATURE V\$AML1_06
			v\$AML1_01 GSE29618_PRE_VS_DAY7_POST_LAIV_FLU_VACCINE_PDC_DN ATGTACA,MIR-493 ZHAN_MULTIPLE_MYELOMA_HP_DN MORF_RAD23A
			MORF_GNB1 GSE10856_CTRL_VS_TNFRSF6B_IN_MACROPHAGE_UP GSE15930_STIM_VS_STIM_AND_TRICHOSTATINA_48H_CD8_T_CELL_UP GSE22886_NAIVE_CD8_TCELL_VS_DC_DN
			GSE27786_BCELL_VS_CD4_TCELL_UP GSE17721_LPS_VS_POLYIC_12H_BMDM_UP GINESTIER_BREAST_CANCER_20013_AMPLIFICATION_DN GSE17721_CPG_VS_GARDIQUIMOD_1H_BMDM_DN GSE380_L_MA_IOR_VG_R_MM_USY_STATEST
			MORF_DEK MORF_HAT1 CHROMOSOME_ORGANIZATION_AND_BIOGENESIS MARTENS_TRETINOIN_RESPONSE_DN
			MORF_HDAC2 V\$USF_01 FLECHNER_BIOPSY_KIDNEY_TRANSPLANT_OK_VS_DONOR_UP PENG_GLUTAMINE_DEPRIVATION_DN
			GSE27786_NEUTROPHIL_VS_MONO_MAC_DN GSE37416_0H_VS_12H_F_TULARENSIS_LVS_NEUTROPHIL_UP REACTOME_TRANSPORT_OF_MATURE_MRNA_DERIVED_FROM_AN_INTRONLESS_TRANSCRIP



Supplemental File 11

Gene Expression

Exon Expression



Splicing Index

Supplemental File 13

Validation of survival analysis results using external data sets

To our knowledge, the only external dataset publically available with exon level results along with clinical outcome is the RNA-Seq dataset deposited in Tumor Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/). This dataset encompasses 921 breast cancer samples annotated according with PAM50 subtype classification, of which 140 are annotated as basal-like,

Our plan was to first check concordance at gene-level, and then to proceed with the further validation comparison at exon- and splicing-index level.

We tried to validate our gene-level results by comparing genes whose overall expression was associated with prognosis in basal-like breast cancer in both data sets. We took prognostic genes in basal-like tumours from our data set (204 genes having q-val < 0.1) and compared them with an equivalent number of genes with lowest p-values for association to prognosis from the TCGA data set (to be noticed that none of the TCGA genes passed the threshold of q-value < 0.1). Surprisingly, none of our prognostic genes was found in the list of TCGA genes.

To understand these results, we then explored further the TCGA dataset and run survival analyses across all PAM50 subtypes. As a result, we found that in this data set the basal-like subtype does not show up as the most aggressive subtype, in overt contrast with all previous results published on the same subject [1-3]. Even more puzzling was the observation that he subtype with the worst prognosis is the Normal-like one (see Figure below).



Survival analysis of TCGA dataset. Kaplan-Meier survival curves of breast cancer samples from the TCGA breast cancer RNA-Seq based database (UCSC browser was used for the analysis)

An additional observation was that if we take the list of genes with lowest p-values for association with prognosis from TCGA, this list is not enriched for immune related genes, as we observed in our data set and as it should be expected from recent literature, showing that immune infiltration is the among the most important prognostic factors in basal-like and triple-negative breast cancers. On these bases we came to the conclusion that the survival data of the TCGA dataset present remarkable divergences from the majority of other data sets in the field [9], and that this probably explains the reasons for lack of convergence with our results. Based on these observations, it came by no surprise that also the exon-level results (EE and SI associated to prognosis) when compared between our and TCGA-based analyses present no overlap.

We therefore moved to the comparison with another public resource, accessible on-line (<u>http://kmplot.com/analysis/</u>). [10]. This database, here referred to as the KMP DB, is the largest public resource of standard Affymetrix-based gene expression data and clinical information from breast cancer, extracted from a large number of publically available studies (from GEO, TCGA and EGA). It encompasses data from 4,142 breast tumours, 54,675 Affymetrix probe set IDs and 70,632 gene symbols.

Out of the 204 genes associated with basal-like prognosis in our dataset (q-val < 0.1), 168 (83%) had q-val <0.1 in the KMP database (Fisher-test p-value of the overlap < 10^{-20}). As a negative control, when we took the 204 genes with lowest association with prognosis from our dataset, only 25% had a q-value < 0.1 in the KMP database.

We went on with the comparison of the list of highest and lowest genes associated with prognosis from the TCGA data set and those associated with prognosis in the KMP database, and we found a very limited overlap (30%). This was not significantly different from the overlap between the 204 genes with lowest association with prognosis from TCGA, used as a negative control (19%) (see Figure below).

In summary:

- 1. Gene-level comparison with the only data set publically available comprising exon-level expression and clinical level information (TCGA RNA-Seq breast cancer) showed no concordance. This data set presented remarkable deviations with respect to survival analysis from what expected from other published literature and data sets.
- 2. Conversely, gene-level comparison with the largest existing resource of standard Affymetrixbased gene expression data and clinical information – KMP - provided striking confirmation of our gene-level results of survival associated genes.
- 3. TCGA gene-level prognostic results were not overlapping with KMP-based ones, pointing again to the fact that this data presents peculiar survival features when compared to other published data sets.



Comparison between basal-like breast cancer prognostic genes resulting from the analysis of our data set (Guy's), and two independent datasets (TCGA and KMP data sets). Each bar represents the fraction of genes identified as associated with survival in one dataset and checked in the second data set. From each database, q-value < 0.1 was set as a threshold for association with survival (Benjamini-Hochberg corrected p-value for multiple testing). Green and red colours respectively indicate presence confirmed and not-confirmed. From left to right: i) basal-like prognostic genes identified in Guy's and checked in KMPDB ii) basal-like non-prognostic genes identified in Guy's and checked in KMPDB iii) basal-like prognostic genes identified in TCGA and checked in KMPDB iv) basal-like non-prognostic genes identified in TCGA and checked in KMPDB v) basal-like prognostic genes identified in Guy's and checked in TCGA vi) basal-like non-prognostic genes identified in Guy's and checked in TCGA.

1. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S *et al*: **Repeated observation of breast tumor subtypes in independent gene**

expression data sets. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(14):8418-8423.

- 2. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L *et al*: **The molecular portraits of breast tumors are conserved across microarray platforms**. *BMC Genomics* 2006, **7**:96.
- 3. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z *et al*: **Supervised risk predictor of breast cancer based on intrinsic subtypes**. *J Clin Oncol* 2009, **27**(8):1160-1167.
- 4. Ignatiadis M, Singhal SK, Desmedt C, Haibe-Kains B, Criscitiello C, Andre F, Loi S, Piccart M, Michiels S, Sotiriou C: Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis. *J Clin Oncol* 2012, **30**(16):1996-2004.
- 5. Szasz AM, Lanczky A, Nagy A, Forster S, Hark K, Green JE, Boussioutas A, Busuttil R, Szabo A, Gyorffy B: Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget* 2016.

CYFIP2



Breast Cancer Specific Survival

WIPF1



SLAMF1





Gene Expression

Supplemental File 18

Exon Expression

Splicing Index

Exon Array Data Processing Workflow

RLE plot

Relative Log Expression (RLE) values are computed by calculating for each probe-set the ratio between the expression of a probe-set and the median expression of this probe-set across all arrays of the experiment. The outlier normal samples highlighted by a red star were removed from the analysis

Hierarchical Clustering

Hierarchical clustering of all samples was computed was computed after normalization. Two normal samples not clustering with the other normal samples (highlighted by red stars) were removed from the analysis

Description of additional data files 1 2

Description of additional data files 3

4

5	Additional File 1
6	Title: Sample phenotypic information.
7	Description: Sample_ID and Patient_ID: anonymized sample and patient IDs;
8	SampleType: tumour or normal breast tissue; HER2, ER, PR: receptor status (P:
9	positive, N: negative, NA: not available); Final_Grade: Tumour Grade;
10	Percentage_of_Lymphocitic_cells: percentage of lymphocytic cells infiltrated in the
11	Tumour; Diag_To_LastObs:number of days between first and last observation;
12	AGE_at_DIAG: the patient's age at diagnosis. Surv_Status: the patient's status at last
13	observation (1 deceased, 2 alive); PAM50: the sample PAM50 classification.
14	
15	Additional File 2
16	Title: Intrinsic molecular subtypes and clinical markers.
17	Description: Characterization of different tumour types: PR, ER, HER2 receptor
18	status according to IHC; histological grade; molecular intrinsic subtype assigned
19	using transcriptional data and the PAM50 algorithm.
20	
21	Additional File 3
22	Title: Coefficient of determination across sub types
23	Description: Distribution of R^2 values of each sample of a given group with all other
24	samples of the same subtype (blue) or of different subtypes (red). From left to right,
25	the three panels represent the results of the analysis using gene expression (GE), exon
26	expression (EE) and splicing index (SI) values.
27	
28	Additional File 4
29	Title: q-q plots of p-values from pairwise comparisons.

1

Description: Each of the panels shows the quantile-quantile plot for the distribution of the logarithm of p-values of a particular pairwise comparison against the logarithm of the uniform distribution. Each of the coloured lines represents different data types as indicated in the figure caption.

34

35 Additional File 5

36 Title: Classification Models.

37 Description: Sensitivity and Specificity plots for a classification model built to

classify basal-like tumours and NBT. The horizontal axis represents the number of

39 variables used in the model, the vertical axis represents sensitivity (in solid) and

40 specificity (in dashed) measures. Different coloured lines refer to the different

41 variables used. Variables were chosen at random from all genes/exons. More

- 42 variables, naturally improve model quality.
- 43

44 Additional File 6

45 Title: Samples clustering using Principal Components Analysis (PCA).

46 Description: Principal components analysis on three types of data: gene expression

47 (leftmost panel), exon expression (center panel) and splicing index (rightmost panel).

48 Normal breast tissue samples are represented in red, and basal-like tumour samples in

49 black. On the horizontal axis the second principal component, and on the vertical axis

50 is the third principal component.

51

52 Additional File 7

53 Title: Comparison with independent studies

54 Description: Results of comparison between our results and three independent studies.

55

56 Additional File 8

57 Title: Experimental validation of differential splicing on 9 genes

58 Description: Description of the experimental protocol used and the results obtained

for 9 genes differentially spliced in basal-like tumours vs normal samples or in basal-

60 like tumours with good vs bad prognosis

61

2

62

63	Additional File 9)
00		

64 Title: Pathway Analysis Results

65 Description: 1st worksheet: Results from Ingenuity Pathway Analysis (IPA) for the

set of genes only differentially spliced between basal-like and Normal Breast Tissue

67 Ingenuity Canonical Pathways: the name of the pathway in the Ingenuity database. –

log(p-value): the statistical significance of the enrichment as reported by Ingenuity.

69 Ratio: the fraction of genes in the IPA pathway that overlapped the input list.

70 Molecules: the names of the genes in the input gene list that are part of the IPA

canonical pathway. 2nd and 3rd worksheets: p-values for gene set enrichment of

72 differentially expressed or spliced genes against the gene sets of Molecular Signatures

73 Database (MSigDb)

74 Additional File 10

75 Title: Paxillin Signalling Pathway

76 Description: The paxillin signalling pathway as represented by Ingenuity (Ingenuity[®]

77 Systems). In purple are genes affected by differential splicing between basal-like

tumours and normal breast tissues, with no evidence of whole-gene differential

79 expression.

80

81 Additional File 11

Title: Gene sets perturbed at Splicing Index but not Gene Expression level – complete
 results.

84 Description: In this heatmap, each column is a pairwise comparison (either at GE or

SI level), each row is a gene set and colour coded is the $log_{10p value}$ for significant

86 enrichment of that gene set for the list of differentially spliced genes for the respective

87 pairwise comparison.

88

89 Additional File 12

90 Title: q-q plots of the p-values for association with Survival.

- 91 Description: Panels represent, from left to right, Gene Expression, Exon Expression,
- 92 Splicing Index. Deviations from the diagonal (in red) indicate that the p-values differ
- from the theoretical uniform distribution expected for no association with survival.
- 94

95 Additional File 13

96 Title: Validation of survival analysis results using external data sets

97 Description: Document describing the rationale and the results obtained on gene

- 98 expression and survival association from external datasets
- 99

100 Additional File 14.

101 Title: Results of Survival Analysis.

102 Description: This table contains all genes where either total expression or splicing

index could be associated with survival in basal-like breast cancer. The 1^{st} and 2^{nd}

104 worksheets contain respectively gene and exon level information. The 1st contains all

the genes where at least one exon that could be associated with survival in basal-like

¹⁰⁶ breast cancer, either by their expression level, or by their splicing index (SI). It

107 includes the Ensemble ID, the gene symbol, the coordinates of the probeset, the q-

value and hazard ratios for association with survival in the one-factor model with EE,

109 the q-value and hazard ratios for association with survival in the one-factor model

110 with SI, the q value and hazard ratios for association with survival in the two-factor

111 models (i.e. EE + lymphocytic infiltration, and SI + lymphocytic infiltration).

112 Additional File 15

113 Title: Kaplan–Meier curves for CYFIP2

114 Description: On the top: a schematic representation of the exon level gene model for

115 CYFIP2 (taken from UCSC genome browser). On the bottom: the three panels, from

- 116 left to right, show Kaplan-Meier survival curves for Gene Expression, Exon
- 117 Expression, and Splicing Index. In each plot, the three lines represent the top tercile

(red), middle tercile (blue), and lower tercile (green) for the value of the variable. In

insert are the q-value for association with survival, and the hazard ratio with 95%

120 confidence intervals.

121

4

- 122 Additional File 16
- 123 Title: Kaplan–Meier curves for WIPF1.

124 Description: On the top: a schematic representation of the exon level gene model for

125 WIPF1 (not in scale). Dark grey represent protein coding isoforms, and light grey

non-coding. On the bottom: the three panels, from left to right, show Kaplan-Meier

127 survival curves for Gene Expression, Exon Expression, and Splicing Index. In each

128 plot, the three lines represent the top tercile (red), middle tercile (blue), and lower

129 tercile (green) for the value of the variable. In insert are the q-value for association

130 with survival, and the hazard ratio with 95% confidence intervals.

131 Additional File 17

132 Title: Kaplan–Meier curves for SLAMF1

Description: On the top: a schematic representation of the exon level gene model for SLAMF1 (not in scale). Dark grey represent protein coding isoforms, and light grey non-coding. On the bottom: the three panels, from left to right, show Kaplan-Meier survival curves for Gene Expression, Exon Expression, and Splicing Index. In each plot, the three lines represent the top tercile (red), middle tercile (blue), and lower tercile (green) for the value of the variable. In insert are the q-value for association with survival, and the hazard ratio with 95% confidence intervals.

140

141 Additional File 18

142 Title: Effect of lymphocytic infiltration in survival multivariate model.

143 Description: Panels, from left to right, are for Gene Expression, Exon Expression, and Splicing Index. Each point in the plot is a gene or probe; the horizontal axis shows the 144 log_{10pval} for association with survival in the univariate Cox Hazard model (GE, EE, or 145 SI), and the vertical axis shows the difference in $log_{10ppval}$ for that gene/exon when 146 lymphocytic infiltration is introduced as an additional covariate in the multivariate 147 model. Negative values on this axis indicate loss of significance. Highlighted in red 148 are genes related to lymphocytic function, as assessed either by gene expression or by 149 database functional annotations (see Materials and Methods). Highlighted in blue are 150 151 genes retaining statistical significance in the multivariate model. 152 Additional File 19 153 Title: Exon Array data processing analytical workflow and QCs 154 Description: Overview of the Exon Array data processing analytical workflow and 155 results from microarray QC 156

157

158 Additional File 20

159 Title: Excel file with gene- and exon-level results of comparative analyses.

160 Description: q-values, and log fold change for statistically significant differences

between sample groups (genes and exons showing no significance in any of the tests,

162 were omitted from the table).

163