Supplemental Materials and Methods for:

Essential gene profiles in breast, pancreas and ovarian cancer cells

Richard Marcotte[4,*], Kevin R. Brown[1,6,*], Fernando Suarez[4*], Azin Sayad[1], Konstantina Karamboulas[1], Paul M. Krzyzanowski[4], Fabrice Sircoulomb[4], Mauricio Medrano[3,4], Judice L.Y. Koh[1,6], Yaroslav Fedyshyn[1], Dewald van Dyk[1], Bodhana Fedyshyn[1], Marianna Luhova[1], Glauber C. Brito[6], Franco J. Vizeacoumar[1], Frederick S. Vizeacoumar[5], Alessandro Datti[5], Dahlia Kasimer[1], Alla Buzina[1], Patricia Mero[1], Christine Misquitta[1], Josee Normand[4], Maliha Haider[4], Troy Ketela[1,6], Jeffrey L. Wrana[2,5], Robert Rottapel[3,4,6,ψ], Benjamin G. Neel[3,4ψ], and Jason Moffat[1,2,ψ]

[1]Donnelly Centre and Banting & Best Department of Medical Research, [2]Department of Molecular Genetics, and [3]Department of Medical Biophysics, University of Toronto; [4]Ontario Cancer Institute, University Health Network, Toronto, Canada; [5]Samuel Lunenfeld Research Institute, Toronto, Canada; [6]Ontario Institute for Cancer Research, Toronto, Canada

*These authors contributed equally to this manuscript.

[ψ]Correspondence should be addressed to R.R. (rottapel@uhnresearch.ca), B.G.N. (bneel@uhnresearch.ca), or J.M. (j.moffat@utoronto.ca)

Supplemental Methods

Fast/Slow Dropout Validation by siRNA

To differentiate between fast and slow dropouts, live cells were monitored by time-lapse imaging over a period of seven days following siRNA transfection. Briefly, cells were stably infected with RFP-expressing lentiviruses derived from pLJM2, and selected using puromycin. RFP-expressing cells were seeded in 384 well plates such that untransfected cells would reach 90-100% confluence within 7 days. Transfection was performed by liquid handling robots (Biomek FX) in duplicate using Lipofectamine RNAiMax and Dharmacon SMARTpool siRNA reagents, at a concentration of 40nM. Twenty-four hours post-transfection, plates were imaged using the Evotech Opera automated imaging platform (Perkin Elmer) with a 4X objective to capture the entire well. This was taken as day one, and plates were imaged repeatedly for seven days. Images were segmented using Acapella (v2.0), and the number of red cells was estimated using the spot detection algorithm.

Essentiality was determined by comparing duplicate cell counts for each gene knock-down against the cell counts obtained for 60 replicate mock transfections. Statistical significance was established using the non-parametric Wilcoxon rank-sum test followed by Benjamini-Hochberg multiple testing correction; cell counts falling below the mean of the mock transfections and with a significant adjusted p-value ($\alpha = 0.05$) were considered validated. Similar results were obtained using the lower limit of the 95% confidence interval of the mock transfections.

Hierarchical Clustering

General essential genes were defined as genes with GARP p-values ≤ 0.05 in at least half (n = 36) of the cell lines, resulting in a list of 297 genes. To represent these genes, the rank product (1) was calculated using TMeV (2) and plotted against the number of cell lines in which the gene was called essential (GARP p-value ≤ 0.05) (Figure 3A). Word clouds were generated using Wordle Advanced (http://www.wordle.net/advanced), and the number of genes belonging to each term was determined using DAVID (http://david.abcc.ncifcrf.gov/).

To identify tumor-specific genes, the non-parametric Wilcoxon rank-sum test was applied on genes that were called essential in two or more cell lines (GARP p-value ≤

0.05). Each tumor type was compared with the other two tumor types, and genes with a significant difference between groups ($p \leq 0.01$) were selected. This resulted in 66 ovarian-, 155 breast-, and 187 pancreatic-specific genes. The analysis was performed using using R (v.2.12.2).

For breast cancer subtype identification, normalized GARP scores were processed as follows. First, only genes called essential (GARP p-value ≤ 0.05) in at least two breast cancer cell lines were included (n = 3472). Second, only the top 10 percent of genes by variance were used for unsupervised hierarchical analysis (n = 348). The resulting dendrogram was cut into two clusters, which were used to define the minimal number of genes required to classify the cell lines using a Student's t-test, after applying the Benjamini-Hochberg multiple testing correction (FDR ≤ 0.1). This identified 41 genes that were able to classify the cell lines into luminal and basal subtypes. A second analysis was performed as above, except using a supervised approach where the clusters were pre-determined by grouping luminal/HER2+ cell lines and basal cell lines. This resulted in 26 genes capable of recapitulating the known subtypes.

Stability of the unsupervised subtype identification was examined by performing the unsupervised clustering approach, as above, while adjusting the variance threshold from 10% to 20%, 30%, 40% and 50%.

All heatmaps, visual and graphical representations were done using R (v.2.12.2).

Transcriptome processing and analysis

Transcriptome sequencing was performed on the AB SOLiD platform at the Ontario Institute for Cancer Research (Toronto, Canada). Using a starting input of 2μg of total RNA, two rounds of poly-A RNA selection was carried out using Thermo Sera-Mag Magnetic Oligo(dT) Beads. Multiplexed SOLiD whole transcriptome libraries were constructed using 50-100ng of poly-A selected material according to SOLiD Total RNA-Seq kit protocols. Pooled libraries were amplified and enriched following standard SOLiD EZ Bead bulk sample preparation methods. Multiplexed sequencing (1x50bp) was performed on three independent flowcells using SOLiD 4 sequencers. Data was obtained for the following ovarian cell lines: OV1946, OVCAR3, OVCAR5, OVCAR8, SK-OV3, TOV1369TR, and TOV1946.

Colourspace reads were aligned using SHRiMP (2.1.1d) against human RefSeq RNA reference sequences dated November 24, 2010. Results were filtered with samtools to exclude unmapped reads. To compute Reads Per Kilobase per Million mapped reads (RPKM) values, the number of reads mapping to each transcript sequence was normalized by the template length in kilobases and divided by the number of reads mapping to the transcriptome.

Calculation of Non-Expression Rates

To compute Non-Expression Rates (NERs) for shRNA score thresholds, scores were sorted in ascending order, and for each range of top N genes, the proportion of genes without detectable expression (RPKM = 0) was determined. This proportion was considered the NER for each list of top N genes. The mean NER for each N and scoring metric was calculated using all ovarian cell lines to report an overall NER range for each scoring metric examined.

Analysis of Gene Expression and Breast Tumor Subtyping

Breast gene expression and sample-associated clinical data was obtained from The Cancer Genome Atlas (TCGA), and samples were assigned to five subtypes (Basal, ERBB2, Luminal A/B, Normal-Like) as follows. Level 3 normalized and summarized expression data was downloaded from the TCGA data portal (http://tcga-data.nci.nih.gov/tcga/) for 345 primary tumor and 25 normal samples, consisting of sample-level median-centered expression values for 17814 genes, indexed by gene symbol. The data was further median-centered for each gene (3).

Tumor subtypes were established using three publicly available single sample predictors (SSPs) (4-7), applied to each sample using the sweave/R code of Weigelt *et al*. (8). The published gene signatures were matched to the TCGA samples by gene symbols, using the updated Entrez Gene annotation provided by Weigelt *et al*.

Each SSP assigned a subtype classification to each TCGA sample, determined by the highest Spearman rank correlation, while correlation values below 0.1 led to a

classification of 'indeterminate subtype'. The final subtype label was determined by a majority vote among the three SSPs, with ambiguous cases labeled as indeterminate.

In addition to expression-based subtyping, TCGA provided ERBB2 immunohistochemistry (IHC) and in-situ hybridization (ISH) results for a subset of the TCGA tumor samples. Where available, a positive ERBB2 IHC or ISH result superseded the expression-based subtype classification, assigning ERBB2-positive samples as ERBB2 subtype. The following table lists the number of TCGA samples assigned to each subtype.

Table 1: Results of breast tumor subtype classification by gene expression and IHC/ISH.

	SSP-based subtypes	SSP subtypes + ERBB2 IHC/ISH status
Basal	70	70
ERBB2+	43	74
Luminal A	96	87
Luminal B	74	63
Normal-like	27	26
Indeterminate	35	25
Total tumor samples	345	345

Copy Number Variation – Cell Lines

Cell line SNP data (downloaded April 29, 2011) were provided by the Cancer Genome Project group (CGP) at the Wellcome Trust Sanger Institute and can be obtained from http://www.sanger.ac.uk/genetics/CGP/CellLines/. This data, processed by the CGP using the PICNIC algorithm, included 20 breast, 5 ovarian and 10 pancreatic lines that were also profiled in our shRNA screens. PICNIC files for the relevant lines contained a copy number status for each profiled SNP. Genomically contiguous SNPs with matching copy numbers were collapsed into contiguous segments, and then mapped to Entrez Gene using hg18 genomic coordinates. Genomic coordinates used were those included in the JISTIC software bundle (9).

Several cell lines were manifestly triploid or tetraploid, as the majority of the genes had predicted copy numbers of three or four. As such, we defined a gain as a predicted copy number at least one higher than the median copy number of all genes in a given sample.

Copy Number Variation – Tumor samples

TCGA was used for both breast and ovarian primary tumor copy number data. The breast samples were profiled on Affymetrix human SNP array 6.0, and segmented via Circular Binary Segmentation (CBS) (10). For ovarian samples, CBS-processed Agilent 1M aCGH data for 564 primary serious ovarian carcinomas was obtained.

For pancreatic tumor data, we used the data from 28 microdissected pancreatic ductal adenocarcinomas profiled on Affymetrix 100K SNP arrays (11). Data was obtained from the GEO database (GEO accession GSE7130), and processed using the Aroma.CN package, including SNP-RMA normalization followed by CBS segmentation.

For all tumor samples, a Log R-Ratio threshold of 0.33 was used to call gains. For both cell line and tumor data, segmented samples were mapped to genes using Entrez Gene hg18 coordinates.

Enrichment of Functional Annotations

Enrichment for functional annotation was derived using the cumulative hypergeometric probability distribution for finding at least *k* genes from a particular category within a cluster of size *n* using Equation 2:

$$P = 1 - \sum_{k} \frac{\binom{f}{g}\binom{g-f}{n-f}}{\binom{g}{n}} \qquad \text{Eq. 2}$$

where *f* is the number of genes within a category and *g* is the total number of genes studied. P-values were adjusted for multiple testing using the false discovery rate (FDR). For enrichment heatmaps, annotations are filtered to remove those with adjusted P-value > 0.05, and the $-\log_{10}$(P-value) is shown such that a darker blue indicates a

more significant enrichment. Biological process categories were derived from a standard set of Gene Ontology (GO) Biological Process terms downloaded from Entrez Gene (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA, ca. Aug 11, 2010). Biological Process pathways were derived from 880 canonical pathways included in the Molecular Signatures Database (MSigDB) (http://www.broadinstitute.org/gsea/msigdb, version 3.0).

Validation assay - Alamar Blue

Cells were plated in 96-well plates (2 x 10^3 cells/well), incubated for 18 – 24 hours at 37°C, after which the media was replaced with fresh media containing polybrene (6μg/ml). Lentiviruses (10μl) expressing individual shRNA were added to each well. Twenty-four hours after infection, cells were selected using 3μg/ml of puromycin for 48 hours, after which the media was exchanged with fresh media. Seventy-two hours later, fresh media containing 10% Alamar blue (Invitrogen) was added, cultures were incubated for 2 hours, and fluorescence was read using a SpectraMax M2 fluorescent plate reader (excitation wavelength = 544nm, emission wavelength = 590nm). Each experiment was performed in triplicate, and individual shRNA readings were normalized to GFP, LacZ, and luciferase shRNA control.

Validation Assay – Hoechst/Opera

Cells were plated in 96-well plates (~ 1.5 x 10^3 cells/well) in media containing polybrene (8μg/mL). Two hours after plating, cells were infected with lentiviruses expressing individual shRNAs. Twenty-four hours post-infection, virus was removed and cells were selected with puromycin (~2μg/mL) for 48 hours, after which puromycin-containing media was exchanged with fresh media. Ninety-two hours later, Hoechst (33342) was added to live cells. Cells were then fixed with 4% PFA and imaged using the Evotec Opera automated imaging platform (Perkin Elmer). Nuclei were counted using Acapella (v2.0). Each experiment was performed in triplicate.

Rescue of ITGAV Essentiality

Mouse Itgav (BC167182) was cloned into the pMAL-RFCA destination plasmid (kind gift of Luigi Naldini). PL45 cells were infected with lentivirus expressing pMAL-mITGAV two days prior to infecting with shITGAV_39, shLuciferase (shLUC) or shPSMD1. Twenty-four hours post-infection with the shRNA-containing viruses, virus was removed and cells were selected with puromycin (~4 μg/mL) for 48 hours, after which puromycin-containing media was exchanged with fresh media. Five days post selection, cells were

harvested by trypsinization, stained with 7-AAD (BioLegend) and counted using a BD FacsCalibur analyzer with CellQuest Pro software and analyzed using FlowJo (v.7.6.5). Each sample was run for two minutes at a constant flow rate of 35ul/min and the number of live cells in each of the GFP negative (uninfected) and GFP positive sub-populations was determined.

qPCR Knock-down Validation

Cells were plated into 24-well plates (5 x 10^4 cells/well) and infected the following day with the appropriate shRNA. Twenty-four hours post-infection, cells were selected with puromycin and allowed to grow for an additional 48 hours. RNA was extracted using the RNeasy mini kit (Qiagen) following the manufacturer's protocol. RNA was then reverse-transcribed using the SuperScript III First-Strand Synthesis kit (Invitrogen) following the manufacturer's protocol, and 20 ng of cDNA was mixed with iQ SYBR green supermix (Bio-Rad) along with primers specific to the targeted genes. Reactions were run using the CFX96 real-time PCR machine (Bio-Rad). $\Delta\Delta$CT was calculated using TBP as a loading control and knock-down was calculated relative to a GFP shRNA control.

Western blot

Cells were plated in 6-well plates (~1 x 10^5 cells/well) and infected with the appropriate shRNA. Twenty-four hours post-infection, virus was removed and cells were selected with puromycin for 48 hours. Cell lysates were harvested in RIPA buffer 4-5 days after infection. Western blot analysis was performed using antibodies directed against EPS8 (BD Biosciences), RAB31 (Santa Cruz Biotechnology), and ITGAV (Cell Signaling Technology) at a dilution of 1:1000. Blots were also probed with anti-β-actin (Abcam) as a loading control.

Overlap of Essential Genes With Previous Datasets

Essential genes from Cheung *et al.* (12) were identified as the genes appearing in the top 5% of at least two of the three scoring methods mentioned in their paper (Top hairpin, Second best hairpin, and RIGER Kolmogorov-Smirnov). Common essential genes were genes identified as essential in 50% or more of the screened cell lines. The common essential set was then intersected with the 205 common essential genes identified in the current study and found in the 54k shRNA library used by Cheung *et al*.

In order to ensure that GARP scoring did not bias the overlap between the two screens, we recomputed our common essential gene set by applying the above scoring methodology to our screening data prior to intersecting the genes. Overlaps were only performed for genes common to both shRNA libraries. The significance of the overlap was determined using Fisher's exact test (R, v2.12.0).

Cell line-by-cell line overlaps were also performed for all cell lines common to the two studies. In this case, hairpin-level scores from both datasets were scored by GARP, and genes were considered essential in a cell line if they appeared in the top 5% of ranked GARP scores. The essential gene lists for all common cell lines were extracted, intersected, and a p-value was determined for the overlap using Fisher's exact test. The comparisons were restricted to genes common to both shRNA libraries.

Supplementary Tables

Note: All supplementary tables are included as a separate Excel file.

Supplemental Table 1. 80K library contents
Supplemental Table 2. Cell lines, growth conditions and other info.
Supplemental Table 3. shRNA class rules.
Supplemental Table 4. Enumeration of the shRNA classes.
Supplemental Table 5. Genes selected for the siRNA followup experiment and the results.
Supplemental Table 6. General essential genes.
Supplemental Table 7. Overlap of essential genes with Cheung *et al*.
Supplemental Table 8. Tissue-specific essentials with enrichment analyses.
Supplemental Table 9. Breast subtype-specific essentials with enrichment analyses.

Supplementary Figure Legends

Supplementary Figure 1: Full set of (A) enriched GO slim 'Biological Process' terms and (B) MSigDB pathways within classified hairpins shown in Figure 1C.

Supplementary Figure 2: Complete growth curves for fast and slow dropouts in the siRNA validation assay described in Figure 1D. Each SMARTPool siRNA was normalized to mock transfected controls. Each point is the mean ± standard deviation of the replicate cell counts. Growth curves denoted by asterisks correspond to genes significantly different than the control at day 7 by Wilcoxon rank-sum test (adjusted p-value < 0.05).

Supplementary Figure 3: Properties of hairpins and genes classified as fast, continuous or slow dropouts. A) Boxplots of the shARP scores for fast, continuous and slow dropouts in the three cell lines used for analysis in Figure 2. B) Overlap of 'General Essential' genes with hairpin classes derived in each cell line. All points are mean ± standard deviation; 'Random' consists of 1000 selections of 297 random genes.

Supplementary Figure 4: Overlap of the common essential genes determined in this study with common essential genes extracted from a previously published pooled shRNA screen (12). (A) Marcotte common essentials (this study) determined as described in the main text. (B) Marcotte essential genes defined using the same rules as for Cheung *et al*., and common essentials as genes called essential in >50% of the cell lines.

Supplementary Figure 5: Unsupervised clustering dendrogram of all 72 cancer cell lines. Genes were filtered first to those deemed essential in at least two cell lines (GARP p-value ≤ 0.05; n = 5508), and then to the top 10% most variable genes (n = 551). Complete linkage clustering was performed using (1 - Pearson correlation) of zGARP values as the distance metric.

Supplementary Figure 6: Heatmap showing Molecular Signatures Database (MSigDB; http://www.broadinstitute.org/gsea/msigdb/index.jsp) pathways enriched in tumor type-specific genes.

Supplementary Figure 7: Unsupervised clustering of GARP scores for breast subtype data with altered variance cut-offs. (A) 10%. (B) 20%. (C) 30%. (D) 40%. (E) 50%.

Supplementary Figure 8: Chromosome ideograms indicate known regions of amplification in tumors. Barplots above the X-axis depict frequency of cell lines in which indicated gene is essential by GARP score, while bars below the X-axis depict the frequency of observed amplifications in cell lines (orange) or tumors (red). For each plot, the ten genes upstream and downstream of the suspected oncogenic driver gene (yellow) are shown. Plots were generated in R (v2.12.1) using the ggplot2 package (v.0.8.9, http://cran.r-project.org/web/packages/ggplot2/index.html). (‡ - genes present in the chromosomal region but not present in the 80k shRNA library)

Supplemental References

1. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett. 2004;573:83-92.

2. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 2003;34:374-8.

3. Sorlie T, Borgan E, Myhre S, Vollan HK, Russnes H, Zhao X, et al. The importance of gene-centring microarray data. Lancet Oncol. 2010;11:719-20; author reply 20-1.

4. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics. 2006;7:96.

5. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160-7.

6. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America. 2001;98:10869-74.

7. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, 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;100:8418-23.

8. Weigelt B, Mackay A, A'Hern R, Natrajan R, Tan DS, Dowsett M, et al. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. Lancet Oncol. 2010;11:339-49.

9. Sanchez-Garcia F, Akavia UD, Mozes E, Pe'er D. JISTIC: identification of significant targets in cancer. BMC Bioinformatics. 2010;11:189.

10. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5:557-72.

11. Harada T, Chelala C, Bhakta V, Chaplin T, Caulee K, Baril P, et al. Genome-wide DNA copy number analysis in pancreatic cancer using high-density single nucleotide polymorphism arrays. Oncogene. 2008;27:1951-60.

12. Cheung HW, Cowley GS, Weir BA, Boehm JS, Rusin S, Scott JA, et al. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proceedings of the National Academy of Sciences of the United States of America. 2011.