Integrative Genomic Approaches to Dissecting Host-Tumor and Host-Pathogen Immune Processes

by
Michael Steven Rooney

B.S. Engineering
Harvard College, 2007

SUBMITTED TO THE HARVARD-MIT PROGRAM IN HEALTH SCIENCES AND TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTORATE OF PHILOSOPHY IN MEDICAL ENGINEERING AND MEDICAL PHYSICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2015

Signature of Author: Signature Redacted
Harvard-MIT Program in Health Sciences and Technology
May 15, 2015

Certified by: Signature Redacted
Nir Hacohen
Associate Professor of Department of Medicine, Harvard Medical School
Thesis Supervisor

Accepted by: Signature Redacted
Emery Brown
Professor of Computational Neuroscience and Health Sciences and Technology
Chairman, Committee for Graduate Student

Copyright Permissions Notice

Integrative Genomic Approaches to Dissecting Host-Tumor and Host-Pathogen Immune Processes

by
Michael Steven Rooney

B.S. Engineering
Harvard College, 2007

SUBMITTED TO THE HARVARD-MIT PROGRAM IN HEALTH SCIENCES AND TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTORATE OF PHILOSOPHY IN MEDICAL ENGINEERING AND MEDICAL PHYSICS

Abstract

Two parallel research efforts were pursued.

First, we conducted a systematic exploration of how the genomic landscape of cancer shapes and is shaped by anti-tumor immunity. Using large-scale genomic data sets of solid tissue tumor biopsies, we quantified the cytolytic activity of the local immune infiltrate and identified associated properties across 18 tumor types. The number of predicted MHC Class I-associated neoantigens was correlated with cytolytic activity and was lower than expected in colorectal and other tumors, suggesting immune-mediated elimination. We identified recurrently mutated genes that showed positive association with cytolytic activity, including beta-2-microglobulin (B2M), HLA-A, -B and -C and Caspase 8 (CASP8), highlighting loss of antigen presentation and blockade of extrinsic apoptosis as key strategies of resistance to cytolytic activity. Genetic amplifications were also associated with high cytolytic activity, including immunosuppressive factors such as PDL1/2 and ALOX12B/15B. Our genetic findings thus provide evidence for immunoediting in tumors and uncover mechanisms of tumor-intrinsic resistance to cytolytic activity.

Second, we combined measurements of protein production and degradation and mRNA dynamics so as to build a quantitative genomic model of the differential regulation of gene expression in lipopolysaccharide-stimulated mouse dendritic cells. Changes in mRNA abundance play a dominant role in determining most dynamic fold changes in protein levels. Conversely, the preexisting proteome of proteins performing basic cellular functions is remodeled primarily through changes in protein production or degradation, accounting for more than half of the absolute change in protein molecules in the cell. Thus, the proteome is regulated by transcriptional induction for newly activated cellular functions and by protein life-cycle changes for remodeling of preexisting functions.

Thesis Supervisor: Nir Hacohen
Title: Associate Professor of Department of Medicine, Harvard Medical School

Author Biography

Michael Rooney graduated summa cum laude from Harvard College in 2007 with a Bachelor of Science degree in engineering. Before graduate school, he worked as a research fellow at the Harvard Initiative for Global Health in Cambridge, MA, and as an analyst for the Cadmus Group in Arlington, VA. He joined the Harvard-MIT Health Science and Technology Program in 2010 adopting coursework in Bioinformatics and Integrative Genomics and joining the Hacohen lab at the Broad Institute. He lives with his partner Vishnu in Cambridge, MA.

Author Acknowledgments

I would like to thank my family – Linda (Mom), Steven (Dad), Kristen, Rich, Mark, Marion, Jean, and Vishnu – for their care and support through the years. I am especially grateful to my cousin Stan Jurga for encouraging me to apply to MIT and advising me in writing my application essay. I also want to thank my many helpful advisors since college: Sumeeta Srinivasan, Majid Ezzati, Frank Letkiewicz, and most of all Nir Hacohen, who was a fantastic coach for the last five years. Finally, I would like to express my appreciation to my thesis committee, Ernest Fraenkel and Mikael Pittet, for their valuable feedback and advice and generous commitment of time.

Table of Contents

Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity

Authors and Affiliations

Michael S. Rooney[1,2], Sachet A. Shukla[1,3], Catherine J. Wu[1,3,4], Gad Getz[1,5] and Nir Hacohen[1,4,6]

[1]*The Broad Institute, Cambridge, MA*
[2]*Harvard/MIT Division of Health Sciences and Technology, Cambridge, MA*
[3]*Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA*
[4]*Department of Medicine, Harvard Medical School, Boston, MA*
[5]*Massachusetts General Hospital Cancer Center and Department of Pathology, Charlestown, MA*
[6]*Center for Immunology and Inflammatory Diseases & Department of Medicine, Massachusetts General Hospital, Charlestown, MA*

Author Contributions

M.S.R. and N.H. conceived the project, designed analysis strategies, and wrote the manuscript. M.S.R. developed and performed the computational analyses. S.S., C.J.W., and G.G. developed the tools for HLA genotyping, mutation calling, and neo-epitope prediction.

Manuscript Acknowledgments

We are grateful to the TCGA Research Network (http://cancergenome.nih.gov/), the Genotype-Tissue Expression (GTEx) Project (dbGaP phs000424.v3.p1), the FANTOM research consortium (http://fantom.gsc.riken.jp/), and the Broad-Novartis Cancer Cell Line Encyclopedia (http://www.broadinstitute.org/ccle/home) for providing the data analyzed in this manuscript. We thank Mara Rosenberg, Amaro Taylor-Weiner, Chloe Villani, Arnon Arazi, Pavan Bachireddy, and Dan-Avi Landau for their feedback and help accessing data and analysis tools, and Leslie Gaffney for assistance with artwork. Funding was provided by the Blavatinik Family Foundation (NH), the MGH Research Scholars Program (NH), and the NIH Training Program in Bioinformatics and Integrative Genomics training grant (MSR).

Introduction

Recent impressive achievements in cancer immunotherapy are redefining cancer survivability and reshaping the landscape of cancer research. Key advances include the "checkpoint" blockade drugs, like Ipilimumab (targeting CTLA-4) and Nivolimab (targeting PD-1), as well as chimeric antigen receptor T cells (CAR-T cells). The checkpoint blockade drugs have produced 15-50% response rates in metastatic melanoma and metastatic lung cancer, malignancies that have traditionally been resistant to nearly all forms of treatment (Hamid et al., 2013; Hodi et al., 2010; Pazdur). Given these successes, it is likely that tumor-immune axis will become an intensely studied research topic during the next decade. To move forward efficiently, we need new unbiased approaches to determine what drives anti-tumor immune responses and what tumors do to avoid them.

Our understanding of tumor immunity today is the result of decades of research centered mostly on mouse models, cell lines, serology, and immunostaining. These studies have uncovered processes that drive immune responses, such as MHC-mediated presentation of abnormal antigens to T cells (as through viral infection (Saiki et al., 1996), ectopic gene expression (Andersen et al., 2012), or mutational antigens (Linnemann et al., 2015)), damage and stress signaling (as from necrosis (Vakkila and Lotze, 2004) or expression of stress ligands (Groh et al., 1999; Textor et al., 2011)), and paracrine signaling networks (Lin and Karin, 2007). Furthermore, they have revealed a wide array of mechanisms by which tumors can evade or suppress anti-tumor immunity including "repolarization" of the tumor microenvironment with suppressive immune cell types (*e.g.* regulatory T cells (Liyanage et al., 2002) and myeloid-

derived cells (Biswas et al., 2006; Mantovani et al., 2008)), the loss of normal antigen presentation (Khong and Restifo, 2002), and the manipulation of immune-modulatory receptors on leukocytes (most famously, PD1 (Hirano et al., 2005) and CTLA-4 (Leach et al., 1996)). Despite this rich literature, it is still not clear which of these processes are most important in human cancer and whether any important components are being missed.

One important complementary approach will likely be the genomic analysis of tumor biopsies, which can yield unbiased and unanticipated insights. The traditional cancer community has effectively capitalized on biopsy genomics, building a successful genomics-to-bedside pipeline. For instance, recurrent mutations were observed in the kinase BRAF in 2002 (Davies et al., 2002). Just ten years later, small molecules targeting BRAF were clinically available and providing objective responses in 65% of melanoma patients (Robert et al., 2015). In the case of the gene ALK, recurrent alterations were observed in lung cancer in 2007 (Soda et al., 2007), and just six years later, a drug targeting ALK-positive lung tumors was FDA-approved (DiGiulio, 2013). With the arrival of The Cancer Genome Atlas, a collection of biopsy data of unprecedented breadth and depth, the field is poised to continue their success in illuminating the broken circuitry of cancer. The question is whether researchers studying tumor immunity can similarly benefit from these data.

Gene expression as a tool for inferring the roles of immune infiltrates

So far, tumor immunologists applying genome-wide approaches have mostly focused on gene expression data, using this data as a window into which immune cell types infiltrate the tumor microenvironment and how their presence might affect patient survival. One of the most

influential researchers in this arena is Jerome Galon. In a 2006 paper, Galon and colleagues profiled the expression of 18 immune genes across 75 colorectal patients (Galon et al., 2006). While not quite yet "genome-wide", this analysis identified a signature of "T_H1" genes (T_H1 is a class of helper T cells that guard against intracellular pathogens) strongly associated with improved survival (though the top genes, *CD8A*, *GNLY*, and *GZMB*, are more strongly associated with CD8+ T cell in our own analyses). Following up with immunostaining in 415 individuals, the T cell signature out-performed standard histological staging. The authors argued that this suggested immune-mediated rejection of persistent tumor cells after resection. Following this, other groups focused on how gene expression in the tumor stroma might contribute to patient survival. Park and colleagues tackled the question directly by using laser capture micro-dissection to physically separate tumor and stromal cells in breast cancer (Finak et al., 2008). Their genome-wide expression analysis of the stromal tissues revealed clusters of immune genes as being most strongly associated with survival, leading to a simple signature for positive prognosis dominated by marker genes for effector lymphocytes, such as *CD8A* and *GZMA* (though they, like Galon, called this a T_H1 signature).

Others sought to streamline the analysis of stromal signatures in unfractionated heterogeneous tissue biopsies, such as those available from TCGA. Yoshihara *et al.* mined the Gene Expression Omnibus for stroma-specific and immune-specific genes (Yoshihara et al., 2013), building these into indices of stromal and immune abundance using ssGSEA enrichment tool (Barbie et al., 2009), and validating them on the TCGA cohort by showing anti-correlation (r=-0.74) with DNA copy number-based purity estimates determined by the Absolute algorithm (Carter et al., 2012). While this analysis indicated surprisingly high levels of immune infiltration in kidney clear

cell tumors, the notion of an "immune" signature was probably not precise enough to be highly useful – notably, no associations with survival were presented. Fortunately, around this time, Galon and colleagues worked to define gene expression signatures that could be used to support inference on the relevant infiltrating cell types. First curating gene expression profiles for diverse immune cell types as well as other stromal and epithelial cell types present in the colorectal microenvironment, they found marker genes and analyzed their expression in 105 colon cancer biopsies profiled by whole-genome microarray (Bindea et al., 2013). As previous, a cluster of T cell genes (broadly associated with cytotoxic T cells, γδ T cells, and T_H1 cells) predicted improved survival; follow-up qPCR on 81 representative genes in a larger cohort of 153 samples showed genes *GZMA*, *PRF1*, *CD8A*, *GNLY*, *GZMH*, *CXCL13*, *IFNG*, *CXCR6*, *LTK*, and *CCR2* to be the best individual predictors. Since many of these genes encode the cytolytic effector molecules of T cells, the results further vouched for productive protective immunity in colorectal cancer.

More recently, it has been possible to query gene expression signatures specifically for what they predict about immunotherapy response. In 2013, Ji *et al.* analyzed microarray gene expression data from Ipilimumab (CTLA-4 blockade) responders and non-responders and found "Cytotoxic T lymphocyte-mediated apoptosis of target cells" to be the most significantly enriched biological process in responders pre-treatment (Ji et al., 2012). Comparing pre- and post-treatment biopsies to determine which genes were most up-regulated in responders, T cell effector processes were again implicated with *CD8A*, *GZMA*, *GZMK*, and *GZMH* among the top 15 genes. These results suggested that the therapy was most effective in patients with pre-existing immune responses and confirmed that the drug worked by up-regulating these

responses. Results from the new PD1 and PDL1 trials in 2014 provided another opportunity to examine the gene expression correlates of immunotherapy, though among the major papers publishing on these data (Herbst et al., 2014; Powles et al., 2014; Tumeh et al., 2014), only Hodi and colleagues employed gene expression profiling (albeit on a limited set of 100 immune genes). Hodi found that PD-L1 expression, along with a signature of "T_H1" genes (*GZMB*, *CD8A*, *CD27*, *CXCR3*, *CTLA4*, *CD45RO*; again, probably more suggestive of CD8+ T cells than T_H1) were predictive of therapy response to the PDL-1 inhibitor MPDL3280A. Meanwhile, the genes most upregulated in productive response were also associated with effector T cells (*GZMA*, *PRF1*, and *TNF*). Thus, the gene expression correlates of PD-L1 blockade appeared to directly mirror those of CTLA-4.

The need to look beyond gene expression data

One drawback to these analyses was that the high level of correlation among immune genes made it difficult to pinpoint exactly which process was contributing to the observed protective effect. For instance, even with a carefully curated collection of immune gene expression profiles, Galon could not clearly determine whether it was CD8+ T cells, γδ T cells, or T_H1 cells driving his signal. Thus, while useful for biomarker discovery, these gene expression analyses were not well-equipped to reveal novel mechanistic insights. Furthermore, meaning could only be derived by showing associations with external clinical data, greatly limiting the number of hypotheses that could be explored.

Cancer genetic alterations, such as somatic point mutations and copy number alterations, represent another important dimension to biopsy data that had not been extensively analyzed in the context of immunity. Mutations, particularly point mutations, can be extremely informative because frequencies and patterns can imply importance even in the absence of clinical data and because they can implicate causative roles for specific genes and pathways. Furthermore, we know that many of the most important phenotypic changes that allow tumors to grow (Hanahan and Weinberg, 2011) and to resist therapy (Lo, 2012) are effected by mutations. Thus, careful analysis of cancer genetic alterations can provide an unbiased glimpse of cancer's strategy playbook.

To harness this power, the cancer community has invested considerable effort in building informatics tools that leverage mutational profiles. Importantly, significance of a gene can be captured by determining how much more frequently it is mutated than expected given the neutral mutation rate (a measure of the selective force favoring the variant). Advanced algorithms have been developed to extract hits from point mutation data (Hodis et al., 2012; Lawrence et al., 2013), other data types (Landau et al., 2014; Zack et al., 2013), and combinations of data types (Akavia et al., 2010) leading to therapeutic advances (Davies et al., 2002; Levine et al., 2005; Parsons et al., 2008); however, these approaches have not been optimized to find alterations that relate to immunity.

Immuno-editing model bridges host immunity and the cancer genome

The "immuno-editing" model, first proposed by Robert Schreiber in 2002 (Dunn et al., 2002), suggests that host immunity should leave an identifiable footprint on the cancer genome. This

model envisions three principal phases to cancer's relationship with the immune system – elimination, equilibrium, and escape. During the *elimination* phase, the host immune system first becomes aware of the tumor, and active killing of tumor cells may occur, possibly resulting in complete elimination of the tumor. At *equilibrium*, the tumor develops countermeasures that partially neutralize the anti-tumor response. Finally, during *escape*, the tumor outstrips any remaining potency left in the immune response (or learns to manipulate immune signaling to support its own agenda), resulting in uncontrolled tumor outgrowth. Notably, both the elimination and escape phases imply that tumoral alterations influence the immune micro-environment. During the elimination phase, alterations might initially alert the immune system that the cancer cells are abnormal (as through the expression of stress ligands, ectopic gene expression, mutational epitopes, *etc.*). During the escape phase, these alterations might provide the mechanism(s) by which the tumor avoids immune destruction (*e.g.,* disruption of normal antigen presentation).

The immune-editing concept has been around since the 1970s (then known as "immune-surveillance") (Burnet, 1971), but it was not until the mid-1990s that experimental results began to consistently indicate the existence of immune-mediated tumor elimination. The first major study was one that showed that interferon gamma knockout mice were significantly inferior to controls in their ability to control tumors produced by the mutagen methylcholanthrene (MCA). Following this result, other knockout mice were developed. Since perforin was known to be one of the key proteins used by effector lymphocytes to kill target cells, a $PRF1^{-/-}$ mouse was developed and shown to be similarly deficient in its ability to control the MCA-induced tumors. Most critical, however, was the development RAG1 and RAG2

knockout mice. RAG1 and RAG2 are genes necessary for B and T cells to develop adaptive responses (via VDJ recombination), so when it was found that $RAG2^{-/-}$ mice are also poor controllers of MCA tumors (Shankaran et al., 2001), it was the most compelling evidence yet for immune control of cancer and further suggested that immune control was at least partially mediated through the recognition of cancer-specific epitopes.

Evidence that elimination also occurs in humans came from epidemiological studies showing that immunocompromised patients were at greater risk for developing (non-viral) cancers (one well-controlled analysis of 5,692 renal transplant recipients under long-term immunosuppressive regimens showed >2-fold increases in the incidences of multiple tumor types (Birkeland et al., 1995)) and that patients with abundant T cell infiltrates (as measured by histopathology) had improved prognosis (Fridman et al., 2012). A medical case in 2010 showed a concrete case in which a kidney transplant recipient developed a melanoma derived from the donor (Strauss and Thomas, 2010). Alive and well, the donor presumably had trace melanoma cells remaining his body, even though he had been clinically free of disease for 32 years. The outgrowth of the tumor in an immunosuppressed host suggested that the extended "equilibrium" phase in the donor was immune-mediated. This case, along with several others documenting spontaneous remissions from melanoma (Savarrio et al., 1999) provided evidence for immune-editing in humans.

The existence of productive endogenous anti-tumor immunity in humans suggests that there may be genetic signatures associated with immunity arising because they 1) drive the response or 2) provide resistance to it. Robert Schreiber further motivated these ideas through

experiments detailing how tumors change when passaged through immune-competent host mice. First, he showed that tumors grown in $RAG2^{-/-}$ were rejected when transplanted into wild type mice (Shankaran et al., 2001). Conversely, if the tumors were grown in wild type mice and transplanted to other wild type mice, they flourished. The implication was that tumors grown in immune-competent mice adapted to have reduced immunogenicity (or directly immune-suppressive), whereas such selection did not occur in the absence of productive adaptive immunity. Schreiber then explored whether these changes were evident at the genomic level (Matsushita et al., 2012). According to well-established immuno-biology, when proteins are digested in the proteasome they are converted to short peptides (give range), a small fraction of which (~0.5% (Yewdell and Bennink, 1999)) avidly bind the MHC antigen presentation machinery; T cells scan for peptides that are non-self. Hypothesizing that mutational epitopes were driving the T cell response in MCA mouse tumors, Schreiber and colleagues collected whole-exome sequencing for a tumor from a $RAG2^{-/-}$ mouse and ranked mutations by their ability to produce MHC-binding non-self peptides (using an established prediction algorithm (Nielsen et al., 2007a)). When they transplanted this tumor line into immune-competent mice, they were able to observe specific T cell responses against their top-predicted epitope as well as the selective outgrowth of subclones lacking the mutation responsible for the epitope (a non-silent point mutation in spectrin-β2). This study gave direct evidence for how genetic alterations might drive immune responses and how the corresponding immune-mediated selection pressure can influence which alterations are accepted or rejected in an evolving tumor.

Efforts to leverage genetic alterations in the context of tumor immunity

Motivated by the immune-editing concept, some tumor immunologists have explored how genetic changes in human tumor biopsies might inform mechanism. In early, targeted (not genome-wide) studies, researchers identified mutations in the invariant chain of MHC Class I (β2-microglobulin) (Hicklin et al., 1998), in the Fas death receptor (Grønbæk et al., 1998; Landowski et al., 1997; Shin et al., 1999), and in other genes in the extrinsic apoptosis pathway (Shin et al., 2002a; Shin et al., 2002b). These alterations appeared to be immune escape variants that enabled tumors to avoid detection of T cells (via deficient antigen presentation) or avoid FasL-mediated cytotoxicity by killer lymphocytes. In one notable study from by Restifo and colleagues in 2005 (Chang et al., 2005), researchers collected biopsies from five melanoma patients with initial responses to a T cell-based immunotherapy. Amazingly, all five patients had mutations in β2M, the invariant chain of the MHC Class I antigen presentation molecule (and interestingly, 3 of the five had the same CT dinucleotide deletion frameshift). Nonetheless, most of these reports were anecdotal in nature, not reaching the sample sizes necessary to implicate positive selection statistically. Furthermore, their targeted nature meant that they were not well-equipped to discover new biology. A genome-wide context was missing.

Galon and colleagues brought such analysis to the next level, systematically characterizing copy number alterations affecting cytokine loci in colorectal cancer (Mlecnik et al., 2014). While not quite genome-wide (only copy loci containing known cytokines were examined), this study identified amplifications and deletions associated with lymph node metastasis, non-metastatic disease, and risk of relapse. Unfortunately, these results were difficult to evaluate as they were presented without *p*-values or commentary on whether the observed events were broad

(involving large chromosomal regions containing many genes, thus nominating many potential drivers other than the cytokine in question) or focal (involving just the cytokine and its immediate neighboring genes). Around this time, Rutledge *et al.* published a cleaner glioblastoma-focused analysis (Rutledge et al., 2013), which asked whether point mutations and copy number alterations in known significantly mutated genes were associated with the level of T cell infiltration (as assessed by histology). This analysis was motivated by the notion that these mutational features defined molecular classes of glioblastoma that may have different immunological properties. Among the 171 samples (from the TCGA collection), the authors found positive T cell associations for RB1 and NF1 point mutations as well as for EGFR and PTEN copy number alterations, though the associations were barely significant, even without correction for multiple hypotheses. Nonetheless, the approach was straight-forward and novel and could easily be applied to much larger TCGA sample sets to achieve greater power. Given the recent identification of HLA genes as being significantly mutated (according to MutSigCV) in lung squamous cell carcinoma (Cancer Genome Atlas Research Network, 2012) and gastric cell carcinoma (Cancer Genome Atlas Research Network, 2014), we noted that an expanded analysis would be poised to uncover other putative escape variants.

More recently, Holt and colleagues investigated how the cancer genome might be contributing to these immune responses (rather than providing a mechanism of escape), exploring Schreiber's idea that cancer immunogenicity may be driven in part by mutational epitopes (Brown et al., 2014). Holt profiled tumors from six TCGA tumor types, determining which non-silent mutations in each tumor were likely to bind the corresponding patient's HLA, thereby yielding accessible peptide epitopes (neoAgs). They found a strong association between the

epitope count and the level of CD8A expression (used to mark CD8+ T cells; $p=2.0\times10^{-6}$) as well as a modest protective effect on survival ($p=0.02$). Very recently (after the publication of the work presented in this thesis document), Chan and colleagues showed that the count of predicted neoAgs was predictive of PD1 response in lung cancer arguing that the re-activation of neoAg-specific T cells may be the basis of therapy response (Rizvi et al., 2015). In the vein of Rutledge *et al.*, Chan looked for genes that were more frequently mutated in responders or non-responders, but did not uncover significant hits (likely due to the small sample size; $N_R=14$, $N_{NR}=17$).

The need for a more comprehensive genomic analysis of tumor immunity

We decided that there was a need for a more systematic analysis of the connection between immune activity and the cancer genome. Even with a number of papers already published using TCGA data, publications from the consortium had mostly overlooked tumor immunity. In 2012, two years after dramatic melanoma responses were observed to CTLA-4 blockade (Hodi et al., 2010), the official TCGA publication on melanoma did not make any comment regarding immunity (Hodis et al., 2012). After the even more impressive responses to PD1 blockade in 2012 (Topalian et al., 2012), TCGA released a 2013 perspective paper omitting immunity from its list of 12 proposed focus areas for future pan-cancer research (Weinstein et al., 2013). Thus, we perceived a gap between what current cancer genomics analyses were providing and the current trends in cancer therapeutics.

We envisioned that there were several different modes of interaction that could potentially yield observable statistically significant associations between the cancer genome and host

immunity. First, like Holt and colleagues (Brown et al., 2014), we hypothesized that immunogenic factors, including (but not limited to) the count of mutational epitopes, would be positively associated with the level of anti-tumor immunity across a cohort of histologically similar tumors. Second, we predicted that tumors with high levels of anti-tumor immunity would be enriched with compensatory escape mutations in specific genes (such as those uncovered by Restifo and colleagues (Chang et al., 2005)) enabling them to persist in the otherwise hostile immune environment. Third, we recognized that some escape mutations would have strong extrinsic effects on the tumor microenvironment, thereby reversing the conditions promoting their emergence. Thus, with sufficient effect size, this last class of lesions would be enriched in tumors with the lowest level of anti-tumor immune activity.

To enable this line of investigation, we needed a measure for the level of anti-tumor immunity in a tumor. Though our understanding of anti-tumor immune responses is still evolving, it has become evident that these responses are typically ultimately mediated by effector CD8+ lymphocytes or other closely related effector lymphocytes (NK cells, NKT cells, γδ T cells, or possibly T_H1 cells). In a recent meta-analysis of 62 published articles covering 14 different tumor types, immunohistochemical detection of CD8+ T cells was positively associated with prognosis 95% of the time (Fridman et al., 2012).

Cytolytic lymphocytes share an expression program notable for the activation of genes encoding effector molecules, including granzymes, perforin, and Fas ligand that directly mediate target cell death (Russell and Ley, 2002). This gene expression signature has repeatedly emerged as the best immunological prognostic marker in multiple tumor types (Bindea et al.,

2013; Donson et al., 2012; Finak et al., 2008; Galon et al., 2006), consistent with the notion that their expression reflects protective anti-tumor immune processes. Furthermore, these genes are among those most strongly induced in anti-CTLA-4 and anti-PDL1-responsive tumors and are among best for predicting response for both drugs (Herbst et al., 2014; Ji et al., 2012). Therefore, we developed a simple metric based on the two highly expressed and well-documented effector genes, *GZMA* and *PRF1* (expressed in transcripts per million , combined by geometric mean, and referred to as "Cytolytic Activity" or "CYT").

Employing the analysis approach described above, we systematically surveyed what alterations were most strongly associated with immunity in 18 TCGA tumor types, focusing first on those that might drive immune responses and then on those that might enable immune evasion (published (Rooney et al., 2015) and discussed (Burgess, 2015)). We found that neoAgs were significantly positively associated with the level of cytolytic activity in nearly half the tumors types we analyzed. Importantly, mutational signatures suggest that these antigenic mutations are depleted from the cancer genome consistent with T cell-mediated elimination. Meanwhile, viral infection associated with increased cytolytic activity in only several tumor types, and cancer-testis antigens (long considered the most likely instigators of anti-tumor immune responses (Scanlan et al., 2002)) never showed an association. Analysis of point-mutated genes showed that highly infiltrated tumors have increased rates of alteration in the antigen presentation machinery and in genes supporting the extrinsic apoptosis pathway. In addition to these proof-of-principle hits, we nominated novel escape variants. Analysis of copy number alterations similarly revealed expected regulators, such as PD-L1, as well as several other less anticipated hits, like the ALOX gene family.

While these results require further experimental investigation and validation, they provide additional compelling evidence for the immune-editing of tumors, implicate a set of highly immune-reactive tumor types (including colorectal, uterine, and head/neck), and highlight a set of escape variants that could potentially serve as therapeutic targets or as markers for monitoring immunotherapy response. Most importantly, the analysis provides proof of concept for an integrative immuno-genomics cancer research approach, which we hope will influence the design of future cancer studies.

Results

A metric for immune cytolytic activity based on gene expression in TCGA tumors

To study immune effector activity in solid tumors, we focused on cytotoxic T cells (CTL) and natural killer cells (NK) because of their potent ability to kill tumor cells and numerous studies showing that effector T cells at the tumor site predict favorable outcome across many cancers (Pages et al., 2005; Sato et al., 2005; Schumacher et al., 2001). Using RNA-Seq data from thousands of TCGA solid tumor biopsies, we devised a simple and quantitative measure of immune cytolytic activity ('CYT') based on transcript levels of two key cytolytic effectors, granzyme A (GZMA) and perforin (PRF1), which are dramatically upregulated upon CD8+ T cell activation (Johnson et al., 2003) and during productive clinical responses to anti-CTLA-4 and anti-PD-L1 immunotherapies (Ji et al., 2012) (Herbst et al., 2014). Consistent with their coordinated roles, GZMA and PRF1 were tightly co-expressed in TCGA samples (**Figure S1A, Figure S1B**) and showed CTL-specific expression in panels of human cell types (**Figure S1C, Figure S1D**), thus serving as highly specific markers in heterogeneous tumor samples.

We found that the levels of cytolytic activity were highest in kidney clear cell carcinomas and cervical cancers, lowest in glioma and prostate cancers, and average (albeit skewed to high levels) in melanoma (**Figure 1A; Table S1A, B**). Most normal tissues (from TCGA or the Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2013a)) showed definitively lower (6 tissues) or equal (7 tissues) cytolytic activity compared to their corresponding tumors, but two showed definitively higher activity (lung and colon).

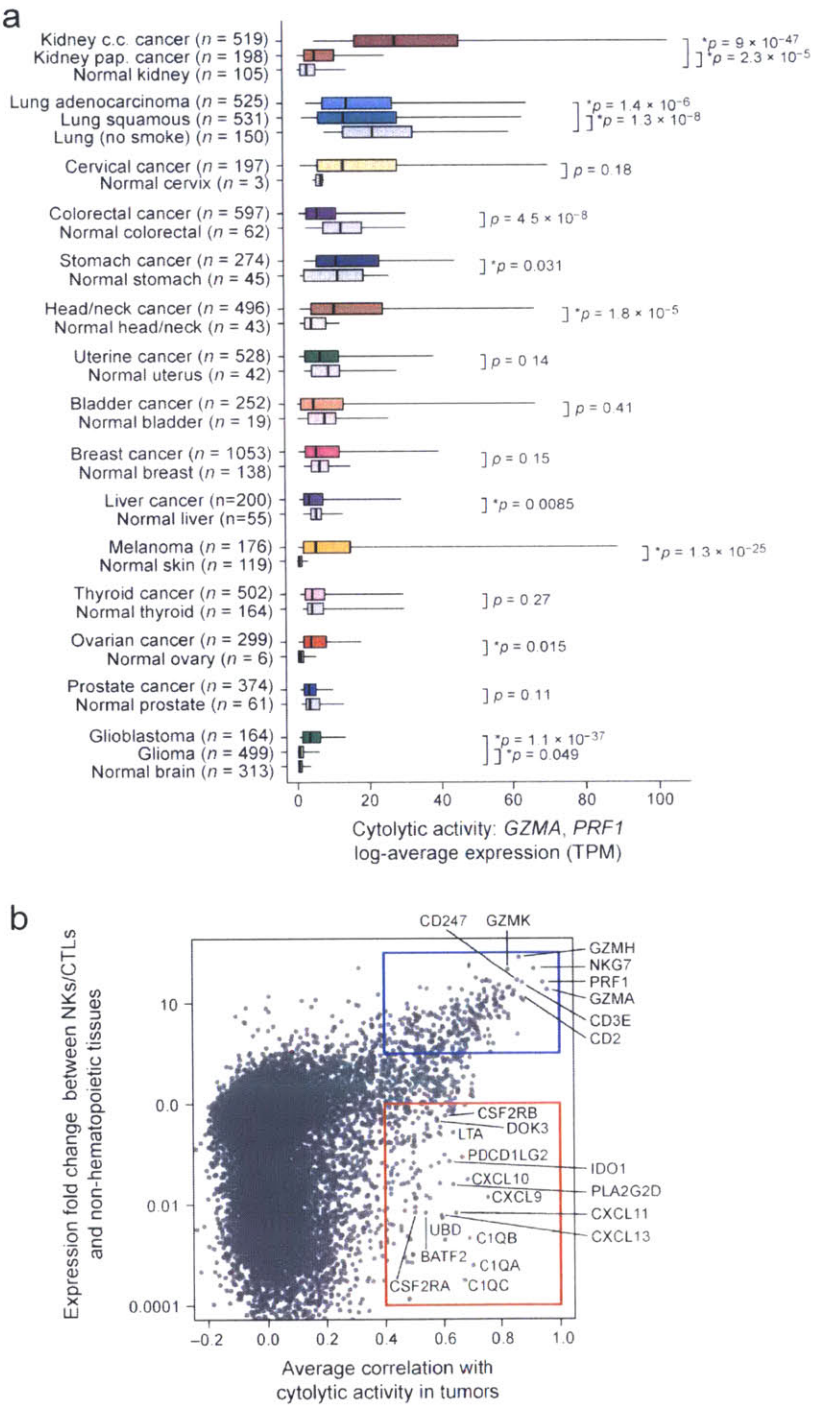


Figure 1. Immune cytolytic activity (CYT) varies across tumor types and is associated with suppressive factors

(A) Cytolytic activity (CYT), defined as the log-average (geometric mean) of GZMA and PRF1 expression in transcripts per million (TPM), is shown for each of 18 TCGA tumor types and normal tissues. Normal tissue samples include TCGA controls and GTEx samples, excluding smokers for lung tissues. Boxes in box plot represent interquartile ranges and vertical lines represent 5th-95th percentile ranges, with a notch for the median. P-values are unadjusted and calculated by Wilcoxon rank-sum test (comparison to relevant normal), and asterisks denote events significant at 10% FDR. **(B)** The correlation of a gene with CYT across all tumor types is shown (X-axis) relative to its relative expression in CTL/NK cells. Top right, genes expressed in CTL/NK cells that are associated with CYT. Bottom right, non-CTL/NK genes associated with CYT. Average Spearman correlation of expression with CYT was calculated across 18 tumor types. Y-axis: for each gene, median expression in NKs and CTLs divided by median expression in non-hematopoietic cells using CAGE data from Fantom5. See also **Figure S1** and **Table S1**.

Of note, CYT in colorectal tumors increased considerably given high microsatellite instability (MSI) (**Figure S1E**) (Schwitalle et al., 2008). The differences in cytolytic activities across tumor types and compared to normal tissues are likely to reflect a combination of tissue- and tumor-specific mechanisms that regulate local immunity.

Cytolytic activity is associated with counter-regulatory immune responses and improved prognosis

To determine whether cytolytic activity is associated with other immune cell types and functions, we calculated the enrichment of 15 immune cell type and function gene sets in the same samples (**Table S1C**; expression data from Fantom5 project (Fantom Consortium et al., 2014)). While CYT showed moderate correlation with B cells and weak correlation with macrophages, it showed strong correlation with: (i) CTL markers, as expected; (ii) plasmacytoid dendritic cells; (iii) counter-regulatory Tregs and known T-cell co-inhibitory receptors, as seen in chronic inflammatory conditions (**Figure S1F**) (Lund et al., 2008). We note that expression of the pre-defined gene sets was similarly enriched in most tumor and normal tissues, with some notable differences (**Figure S1G**), and not typically connected to tumor stage (**Figure S1H**, **Figure S1I**). Finally, when we looked for CYT correlations with any transcript (filtering out CTL and NK genes), we found that CYT was best correlated with immunosuppressive factors (Spranger et al., 2013), such as PDCD1LG2 (PDL2), IDO1/2, DOK3 (Lemay et al., 2000), GMCSF receptor (CSF2RA, CSF2RB) and the C1Q complex (**Figure 1B**). In addition, it was also associated with interferon-stimulated chemokines (CXCL9, CLCL10, and CXCL11) that attract T cells, as observed previously (Bindea et al., 2013). We conclude that tumors can differ dramatically in

their infiltrate levels and composition, and that cytolytic activity is associated with counter-regulatory activities that limit the immune response.

When we used CYT and these other metrics to identify predictors of survival (controlling for tumor histology and stage), we found that high-CYT (and other T cell markers) is associated with a modest but significant pan-cancer survival benefit (**Figure S1J**). While no individual immune cell type metrics were associated with poorer prognosis, higher expression of macrophage markers relative to other markers was consistently linked with poor prognosis, while higher expression of CYT or CTL markers was correlated with improved prognosis (**Figure S1J**).

Tumor cytolytic activity is associated with oncogenic viruses in some tumors

Viruses account for a subset of malignancies and are also known to activate high affinity antigen-specific CTLs against non-self viral antigens. Thus, we tested for correlation of cytolytic activity levels with transcripts from oncogenic viruses – including Epstein Barr virus (EBV), hepatitis B and C (HBV and HCV), human papilloma virus (HPV), Kaposi sarcoma virus (KSV), and polyoma viruses (**Table S2A**). Consistent with previous analysis of TCGA data (Tang et al., 2013), HPV infection was most abundant in cervical cancer (91%), but also frequent in head and neck cancer (12%; with more men than women, OR=4.9; p=8.5e-4) and bladder cancer (2%). We also observed occasional cases in colorectal, kidney clear cell, glioma, lung squamous cell carcinoma, and uterine cancer (**Figure 2A**).

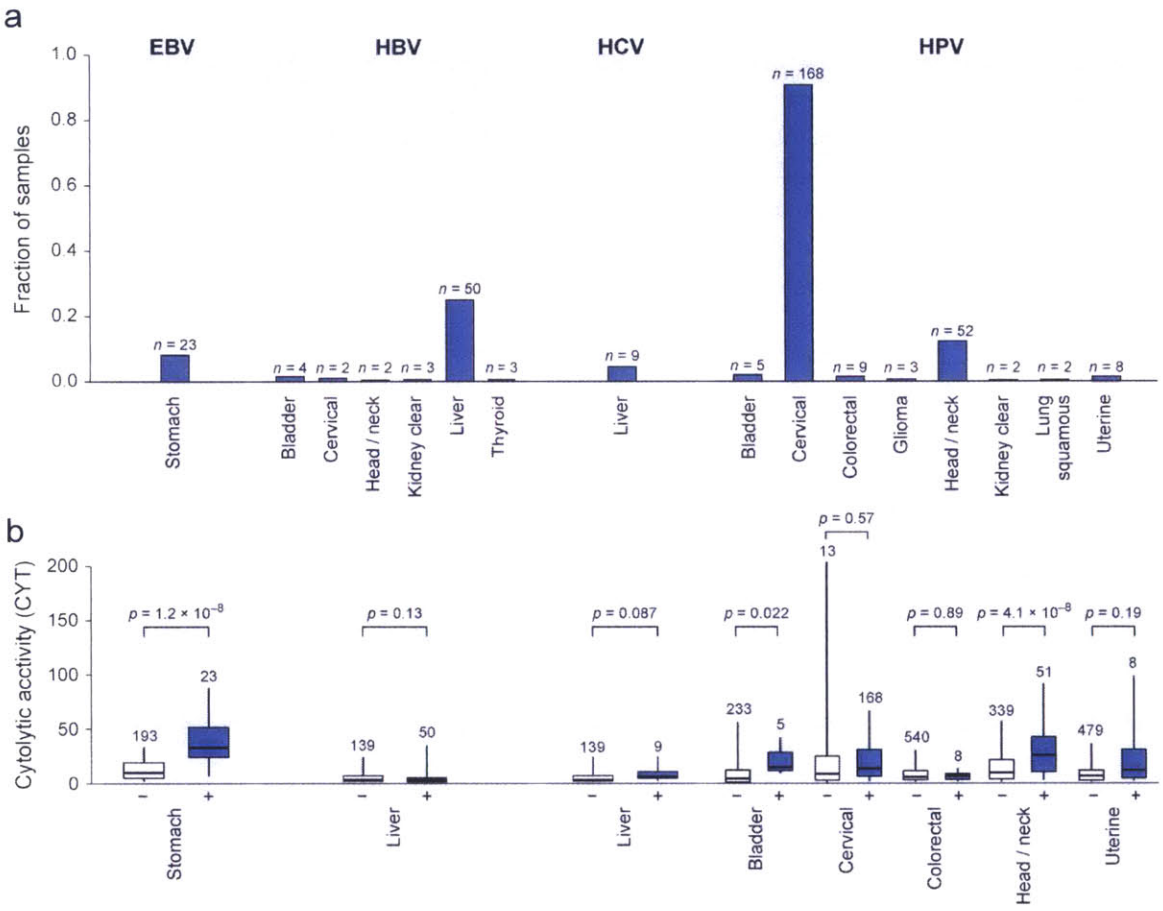


Figure 2. Viral infection is tumor-specific and associated with higher CYT in a subset of tumor types
(A) Rates of viral infection, as defined by viral RNA-Seq read counts exceeding those observed in GTEx, for tumor types exhibiting at least one case. Isolated cases of several other viruses were also observed. (B) Distribution of CYT in tumor samples with (+) or without (-) viral infection. In tumor types affected by multiple viruses, "negative" samples include only those negative for all viruses. Box plots as in Figure 1. P-values are according to Wilcoxon rank-sum test. See also **Figure S2** and **Table S2**.

Only stomach cancer demonstrated definitive instances of EBV infection (8%; **Table S2A**), which was associated with high expression of specific EBV genes EBER-1 and RPMS1 (**Figure S2A**). Asian patients, known to exhibit increased rates of stomach cancer (Jemal et al., 2007), were not more likely than other stomach cancer patients to harbor EBV (p=0.63). Consistent with a role for viral infection in the induction of CTLs, >2-fold increases in cytolytic activity were observed in EBV+ vs. EBV- stomach cancers and HPV+ vs. HPV- head and neck cancers, bladder cancers, uterine cancers and possibly cervical cancers (**Figure 2B**). Strikingly, all the gene sets that were most tightly associated with EBV infection in stomach cancer related to T cell activation (**Table S2B**).

HBV and HCV were primarily observed in liver cancer (25% and 5%, respectively), as expected, with occasional instances of HBV infection in diverse tumor types. The extra hepatic cases do not exhibit hepatic gene expression signatures, suggesting that these are not the result of metastases (**Figure S2B**). We also observed singleton cases of Kaposi sarcoma virus (lung squamous cell carcinoma and stomach cancer), BK polyoma (bladder cancer), and Merkel cell polyoma (ovarian cancer). While we did observe type I interferon activation and B cell infiltration for HCV+ liver cancer (**Figure S2C**), these viruses did not show an identifiable association with cytolytic activity.

To probe indirectly for the presence of viruses, we looked for associations between CYT and two other correlates of viral infection, HLA genotype and APOBEC activity. While association with HLA genotype was not observed for a single tumor type (although there was a pan-cancer

association with HLA-A31; **Figure S2D**), we did detect association with high APOBEC activity in tumors with viral involvement (head and neck, cervical) and those without known viral involvement (breast, bladder) (**Figure S2E**), suggesting potential for unknown virus infections in some tumors.

Cytolytic cells are likely to be targeting tumor neoantigens

With recent studies from our group and others showing the presence of neoepitope-specific T cells in patients (Fritsch et al., 2014), we tested for CYT association with the overall rate of mutation and the rate of mutations predicted to yield a neoepitopes (*i.e.*, an expressed peptide capable of binding each patient's imputed HLA alleles) (**Figure S3A, S3B, Table S3**). On average, 50% of non-silent mutations yielded ≥1 predicted neoepitope, and 39% of these impacted a substantially expressed gene (median expression ≥10 TPM in the given tissue type). Despite considerable inter-tumoral heterogeneity (**Table S4A**), both metrics exhibited significant positive association with CYT in multiple tumor types, most notably uterine cancer, breast cancer, stomach cancer, cervical cancer, and lung adenocarcinoma (**Figure 3A, 3B**). Consistent with a smoking etiology, lung adenocarcinomas from ever-smokers demonstrated significantly higher CYT than those from never-smokers (p=0.003) (**Figure S3C**). Melanoma mutations exhibited a likely association with CYT. Associations of mutations or neoepitopes with CYT were matched by correlations for other T cell markers, but less so with interferon-responsive genes (**Figure S3D, S3E**). These data are consistent with neoepitopes driving CYT for many tumor types.

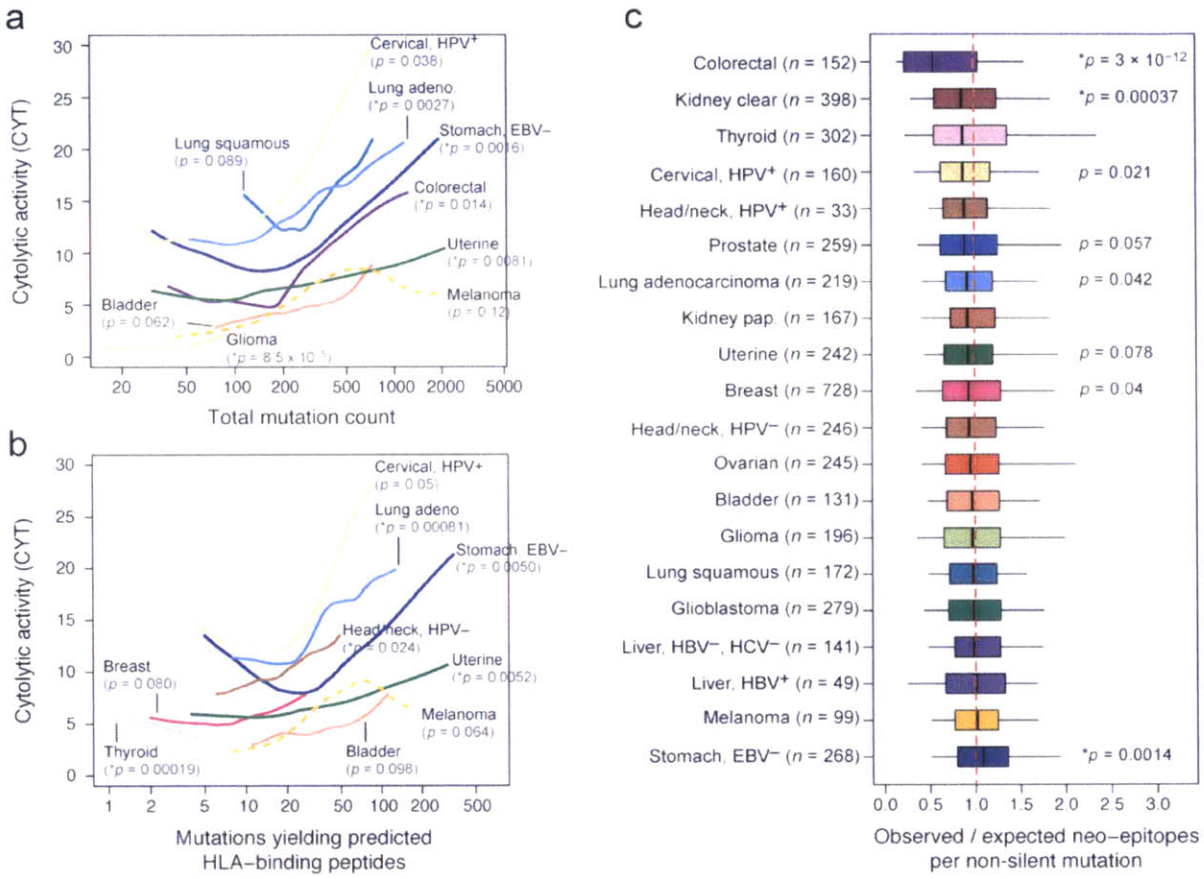


Figure 3. Count of predicted antigenic mutations per sample is linked with cytolytic activity and selectively depleted in certain tumor types

(A) Local regression curves showing significant relationships between CYT and total mutation count in eight tumor types ($p<0.1$, Spearman rank correlation), plus melanoma (dotted line). Curves span the 5th to 95th percentile of the mutation count variable. Colors correspond to tumor type and are the same as appear in Figure 1. **(B)** Analogous to (**A**), but based on the count of point mutations predicted to yield an antigenic neo-epitope. Potential for antigenicity was defined based on gene expression and potential to bind the corresponding patient's imputed HLA with high affinity. **(C)** For each tumor, the count of point mutations predicted to generate neo-epitopes was divided by the total count of non-silent point mutations to yield B_{obs}/N_{obs}. This observed ratio was compared to an expected ratio, B_{pred}/N_{pred}, estimated from the mutational spectra of the silent point mutations in the given sample using an empirical model (Methods). The ratio of the observed and predicted ratios represents the relative deviation of the neo-epitope rate from expectation. P-values reflect Wilcoxon rank-sum tests for deviation from 1. Asterisks denote trends significant at 10% FDR for all panels. See also **Figure S3, Table S3**, and **Table S4**.

However, since the per-sample rate of neoepitope yielding mutations closely tracks with the overall rate of mutation (Spearman rho=0.91; **Figure S3F**), CYT may be driven by mutation rate rather than neoepitopes. To test a role for neoepitopes, we reasoned that T cell-mediated immune surveillance would lead to elimination of immunogenic sub-clones expressing neoepitopes. To quantify neoepitope depletion, we determined how the rate of predicted neoepitopes generated per non-silent point mutation deviated from a null model based on the observed mutation rate of silent point mutations. We found that colorectal cancer and kidney clear cell cancer demonstrated dramatic depletions of neoepitopes (**Figure 3C;** associated gene expression changes, **Table S4B**). Because neoepitope predictions are dependent on HLA genotypes, we reasoned that random shuffling of HLA genotypes would abrogate the depletion signal (**Figure S3G**). As expected, depletion was eliminated for colorectal cancer and kidney clear cell cancer (and we note that the residual enrichment for other tumor types may reflect degeneracy of peptide binding across HLA alleles). These findings are consistent with a model in which immune surveillance activities cull subclones expressing immunogenic antigens.
We conclude that neoepitopes are likely to be driving cytolytic activity in a number of tumors, and that the resulting antigen-specific CTLs can eliminate tumor clones harboring these neoepitopes.

Ectopic gene expression, endogenous retroviruses and necrosis associated with CYT

Another potential source of tumor antigens is a unique set of genes, known as cancer testis (CT) antigens, which are not expressed in healthy tissues, except germ cells, but are aberrantly

expressed in tumors and associated with antigen-specific responses in patients harboring these tumors. Ectopic expression is likely due to disturbances in genomic methylation and reactivation of stem-like expression programs that may contribute to tumorigenicity (Simpson et al., 2005). Using a set of 276 known CT genes (Almeida et al., 2009), we used GTEx to identify a subset of 60 that are transcriptionally silent in normal non-germline tissues. Ectopic expression was observed for most tumor types, especially melanoma, head and neck, lung, liver, stomach, and ovarian cancer (**Figure S4A; Table S4A**). In no tumor type was there a clear positive association between the CYT and the count of expressed CT antigens (**Figure S4B**). We queried individual CT antigens for correlation with CYT (**Table S5A**), and observed positive associations for CSAG2 in breast cancer (p=1.2e-15), head and neck cancer (p=1.9e-7), kidney clear cell cancer (p=9.9e-5), and other tumor types. Associations for canonical antigens, such as NY-ESO-1 (*CTAG1*), were less consistent. We hypothesized that T cell surveillance would lead to CT antigen silencing through chromosomal deletions, but compelling evidence for this was not observed (**Figure S4C**).

Endogenous retroviruses (ERVs) are another class of germline-encoded elements that may be re-activated in tumors, and we considered whether these might also contribute to anti-tumor immunity. TLR7 or RAG knockouts in mice develop uncontrolled ERV expression, ERV infectivity, and ERV insertion-driven tumors (Young et al., 2012; Yu et al., 2012) yet little is known about ERV-immune and ERV-cancer interactions in humans. Given reports that these elements are transcriptionally and sometimes even translationally active in humans (Boller et al., 1997; Schmitt et al., 2013), we considered the possibility that they trigger immune sensing in tumors.

Therefore, we mapped TCGA RNA-Seq data to a recently published annotation of 66 expressed ERV family members (**Table S5B, Figure S4D**) and assessed associations with cytolytic activity (Mayer et al., 2011). By comparing GTEx and TCGA tissue controls to TCGA tumor samples, we observed numerous instances of ERVs demonstrating re-activation in tumors, including one instance of an ERVH-2 element exceeding 2,700 reads per million in a stomach adenocarcinoma (**Figure S4E**). From these data we surprisingly discovered a conservative set of three tumor-specific endogenous retroviruses ('TSERVs') all with minimal to undetectable expression in normal tissues and elevated expression in tumor tissues (**Figure 4A**). Assessing the gene expression correlates of each TSERV in the tumor type exhibiting highest expression, we observed that immune pathways were typically the most significantly enriched (**Table S5C**). Many ERVs, in addition to the TSERVs, demonstrated association with CYT in multiple tumor types (**Figure 4B**). While we cannot determine whether ERVs activate immunity or inflammation triggers ERVs (Manghera and Douville, 2013), we conclude that ERVs are highly dysregulated in tumors and speculate that they may yield tumor-specific peptide epitopes (Boller et al., 1997) or act as immunological adjuvants to activate local immunity (Yu et al., 2012).

Another potential source of antigens and immunostimulatory ligands is dying cells. Thus, we explored the potential role for necrosis in driving CYT and immune infiltration in general. Rates of necrosis were highest in glioblastoma (**Figure S4F**) and showed modest positive association ($p<0.05$) with CYT in glioblastoma, bladder, and ovarian cancer; but notably, association with macrophage markers was consistently stronger (**Figure S4G**).

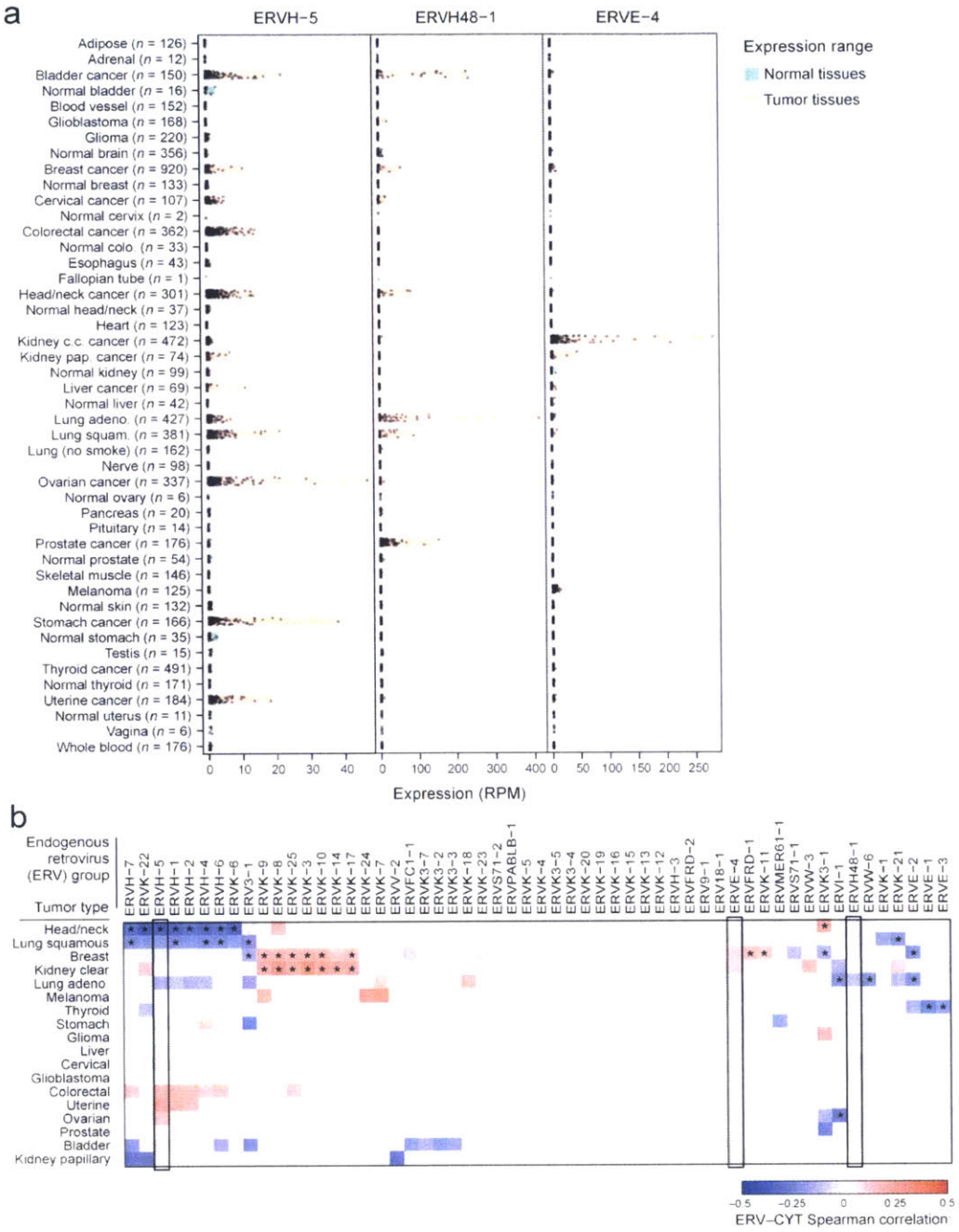


Figure 4. Endogenous retroviruses tied to local immunity

(A) RNA-Seq-derived ERV expression in reads per million (RPM) across 18 TCGA tumor types and 27 non-tumor tissue types (from TCGA and GTEX) for three elements found to be tumor-specific. The expression ranges (minimum value to maximum value) are highlighted in orange (for tumor tissues) or green (for non-tumor tissues). **(B)** Spearman-rank correlations between CYT and ERV expression. Gray squares indicate non-significant association (unadjusted p>0.05) and blank squares indicate no over-expression of the given ERV in the given tumor type (expression strictly below the normal tissue maximum). Asterisks (*) denote Bonferroni-significant associations (adj. p<0.05). See also Figure S4 and Table S5.

Mutations in specific driver genes were enriched in tumors with higher cytolytic activity

We hypothesized that high cytolytic activity could select for tumors with somatic mutations that render them resistant to immune attack. We therefore asked whether CYT is associated with mutations in 373 'driver' genes that are frequently mutated in cancer based on analysis of TCGA exome sequencing data (q<0.1 by MutSigCV (Lawrence et al., 2013); **Table S6A**). Using a regression-based approach to look for pan-cancer association of these mutated genes with CYT, controlling for tumor type and background mutation rate, we found 35 genes (adjusted p<0.1; **Figure 5A**, **Figure S5A**, **Table S6B**). In contrast, synonymous somatic mutations were not associated with CYT (adj. p_{min}=0.09). Of the top 10 CYT-associated mutations, 8 were also associated with an independent marker of CTLs (CD8a; 10% FDR; **Figure S5B**), demonstrating the robustness of our CYT metric. Of the individual tumor types, uterine, stomach and colorectal had the most associations (15, 11, 6 respectively) while kidney clear cell and ovarian, which showed markedly higher CYT compared to normal tissue, had just one each, and lung adenocarcinoma had none. Strikingly, somatic mutations, except *TP53*, were all positively associated with CYT, consistent with a model in which tumors develop resistance mutations under selection pressure.

We note that while we predicted that cytolytic activity would have the strongest impact on the mutation landscape, we also identified gene mutations strongly associated with other immune cell types/functions (adj. *p*<.01; **Figure S5B**), including STK11 and VHL with reduced macrophage signature, BRAF with increased expression of costimulatory genes, and AXIN2, SNX25 and others with the differential enrichment score of CD8+ T compared to Treg.

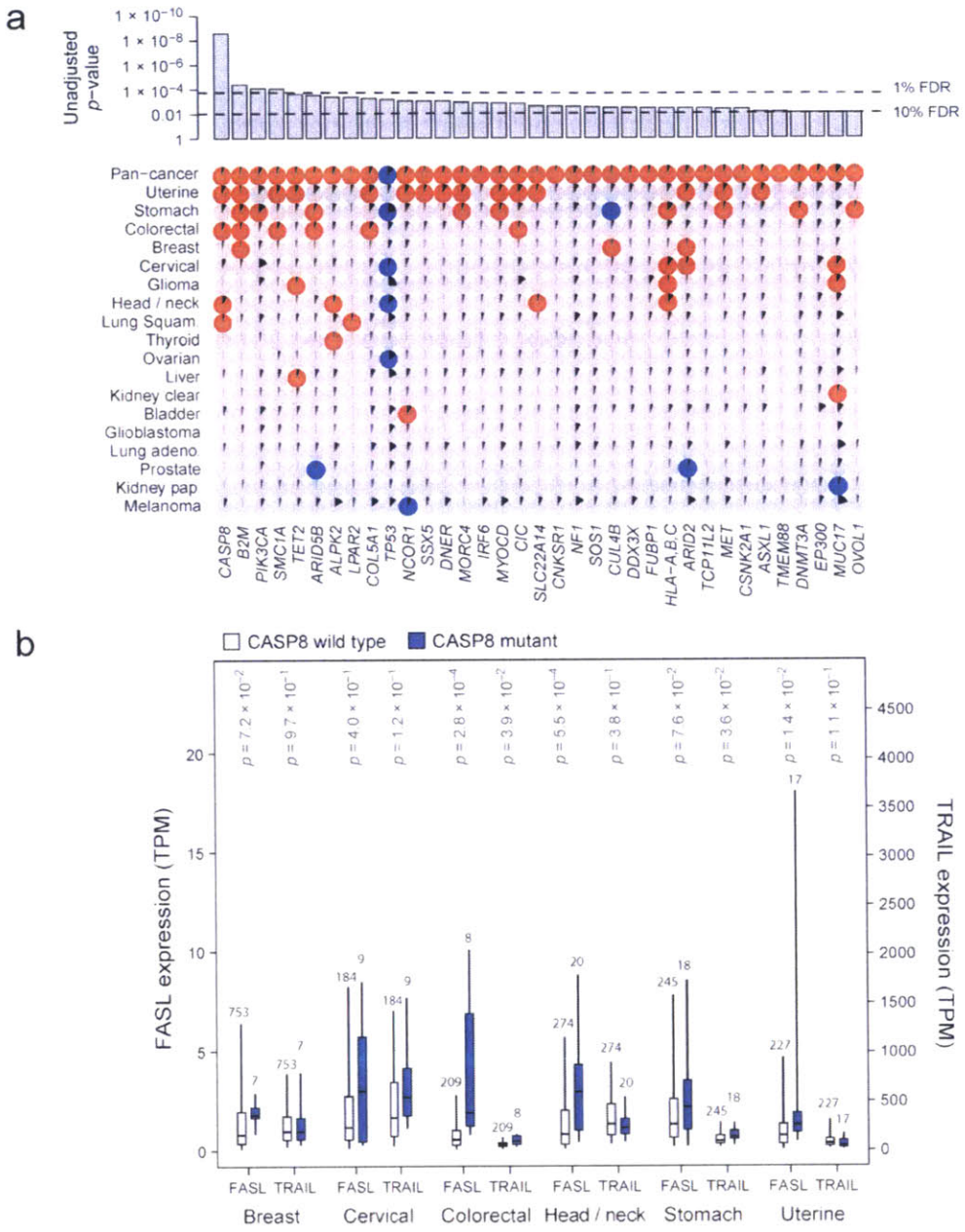


Figure 5. Gene mutations associated with high or low immune cytolytic activity

(A) Only genes showing pan-cancer significance (adj. p<0.1, red for positive, blue for negative and grey for non-significant association) for non-silent mutation association with CYT are shown in top row. Additional rows, clustered by similarity, show independent significant (unadjusted p<0.05) enrichment upon sub-analysis. The black wedges represent the share of samples exhibiting mutation. Bar plot indicates unadjusted pan-cancer p-values for mutational association with CYT, dashed lines indicating thresholds yielding 1% and 10% FDRs. **(B)** Association between CASP8 mutational status and FASLG (left axis) and TRAIL (right axis) gene expression (TPM) for tumor types demonstrating at least 5 instances of nonsynonymous CASP8 mutation. Light and dark bars correspond to wild type and (nonsynonymous) mutant samples, respectively. Box plots as in **Figure 1**. P-values are calculated by Wilcoxon rank-sum test. See also **Figure S5** and **Table S6**.

Higher CYT was associated with mutations in genes involved in antigen-presentation, extrinsic apoptosis and innate immune sensing

Several themes emerged when we considered the known functions of the identified genes. First, the most enriched gene, *CASP8* (adj. p=8.8e-7), is a critical player in the extrinsic apoptosis pathway and was enriched in head and neck cancer, colorectal cancer, lung squamous cell carcinoma, and uterine cancer (where it showed a maximal mutation frequency of 7.0%). The pattern of mutation was diffuse and suggested loss of function (**Figure S5C**), a potential mechanism by which a tumor cell could evade FasL- or TRAIL-induced apoptosis. Between FasL and TRAIL, FasL is most correlated with CASP8 mutations and thus more consistent with such a hypothesis (**Figure 5B**). A study in mice indeed demonstrated that blockade of CASP8 results in tumor escape from CTLs (Medema et al., 1999), and our result indicates that this may be a common mechanism in human tumors (that may evade CTLs or NK cells). Interestingly, four additional genes with significant but less definitive statistical enrichment also had well-established roles in regulating extrinsic apoptosis. These include, CNKSR1 (Garimella et al., 2014), MET (Fan et al., 2001; Garofalo et al., 2009), CSNK2A1 (Ravi and Bedi, 2002) (Izeradjene et al., 2005) (Llobet et al., 2008; Wang et al., 2006), and PIK3CA (Saturno et al., 2013; Song et al., 2010). PIK3CA mutations, which were often the well-known activating alterations E545K and H1047R (Samuels and Ericson, 2006), showed their strongest enrichment in stomach cancer, demonstrating a 20% mutation rate and a strong positive association with EBV infection (p=2.9e-10). As in the case of CASP8, mutations in each of these genes were more closely associated with FASL expression than TRAIL expression. We conclude that loss of the extrinsic

apoptosis pathway may represent a general mechanism for tumors to escape immune cytolytic activity.

Second, the invariant chain of MHC Class I, B2M, was the next most strongly enriched gene (adj. p=7.1e-3), showing independently significant association in uterine, breast, colorectal cancer, and stomach cancer, which exhibited the highest rate, 5.7%. The most frequent event was the same CT dinucleotide deletion observed previously in melanoma patients relapsing from T cell-based immunotherapy (Chang et al., 2005). The MHC Class I locus itself was also significant (**Table S6C**; HLA-A, -B, -C mutations considered jointly, adj. p=5.3e-2). HLA-A and HLA-B alleles were mutated about 3 times as frequently as HLA-C alleles. No specific alleles showed strong evidence for being especially frequently mutated. The tumor types with the highest rates of HLA mutation, stomach cancer (14%), cervical cancer (12%), and head and neck cancer (11%), were also among those with frequent viral involvement. However, viral infection was not significantly associated with HLA mutation in any of them (**Table S6D**). Given the requirement of MHC Class I and B2M in presenting tumor antigens to cytotoxic CD8 T cells, we consider the enrichment of MHC Class I and B2M mutations in high-CYT tumors (Khong and Restifo, 2002) as an independent and strong validation of CYT as a measure of cytolytic activity. While MHC Class II genes were not significantly mutated pan-cancer, class II gene mutations, considered collectively, were positively associated with CYT (unadj. p=0.017) with independent significance in bladder cancer (unadj. p=0.0084).

Other hits included the CT antigens MORC4 (Liggins et al., 2007) and SSX5 (Ayyoub et al., 2004) and genes with roles in innate immune sensing, including DDX3X (Oshiumi et al., 2010) and ARID2 (Yan et al., 2005). We also note that mutant TP53 is negatively correlated with CYT, which may be explained either by a role for p53 in regulating immunity (*e.g.*, loss of p53-regulated stress ligands that induce cytoxicity, (Textor et al., 2011) or from absence of viral infection (consistent with p53 mutations being anti-correlated with viral infection in stomach (p=2.3e-5) and head and neck cancer (p=2.6e-4); **Table S6D**).

Because MSI-high colorectal tumors are known to be immunogenic (Kloor et al., 2010), we also considered whether MSI-high tumors were enriched for mutations in particular genes with respect to MSI-low and microsatellite stable (MSS) tumors. Mirroring the CYT analysis, *CASP8* and MHC Class I mutations were the most enriched mutations in MSI-high tumors (p adj. = 1.5e-5 and 1.4e-12, respectively), with *COL5A1*, *SMC1A*, *CIC*, *ARID2*, *CNKSR1*, and *DNMT3A* also significant (adj. $p < 0.05$) (**Table S6E**).

Finally, we note that some candidate genes with well-known immune function (**Table S6A**) did not show association with CYT. However, enrichment in the expression of immune-related genes were observed in tumors with mutations in some of these genes (*TNFRSF14*, *CLEC4E*, *CD1D*, *IL32*; **Table S6F**).

Loci containing known immune regulators show copy number alterations associated with CYT

We also considered the possibility that specific regions of the genome may be preferentially focally amplified or deleted (based on a dataset of TCGA samples profiled with SNP6.0 arrays) in high- or low- CYT tumors. As with the point mutation analysis, we looked for pan-cancer CYT association with copy number alterations (CNAs) using regression and controlling for cancer subtype and background mutation rate (of amplifications and deletions). This approach yielded 13 significantly amplified regions (with 3 adjacent to each other on 6q) and 1 significantly deleted region (FDR=10%) (**Figure 6A**, **Table S7**). Although CNAs include variable segments of a chromosomal region and do not typically identify causative genes, many of the identified regions harbored plausible candidates.

On chromosomes 9 and 8, we found two well-known targets of cancer immunotherapy. First, amplification of 9p23-p24.2 (**Figure 6B**), a region including PDL1 (CD274) and PDL2 (PDCD1LG2),was positively associated with CYT in lung squamous cell carcinoma, head and neck cancer, cervical cancer, stomach cancer, and colorectal cancer (**Figure 6E**). While tumor cells and tumor infiltrating leukocytes are known to express these ligands, our results suggest that tumor-expressed ligands affect tumor fitness in the presence of cytolytic activity. Second, 8p11.21-8p11.23 (**Figure S6A**) showed increased probability of amplification in low-CYT tumors (pan-cancer and breast) and is adjacent to IDO1 and IDO2, enzymes that degrade extracellular tryptophan and create a potent immunosuppressive microenvironment, which may explain the associated reduction in CYT (Uyttenhove et al., 2003).

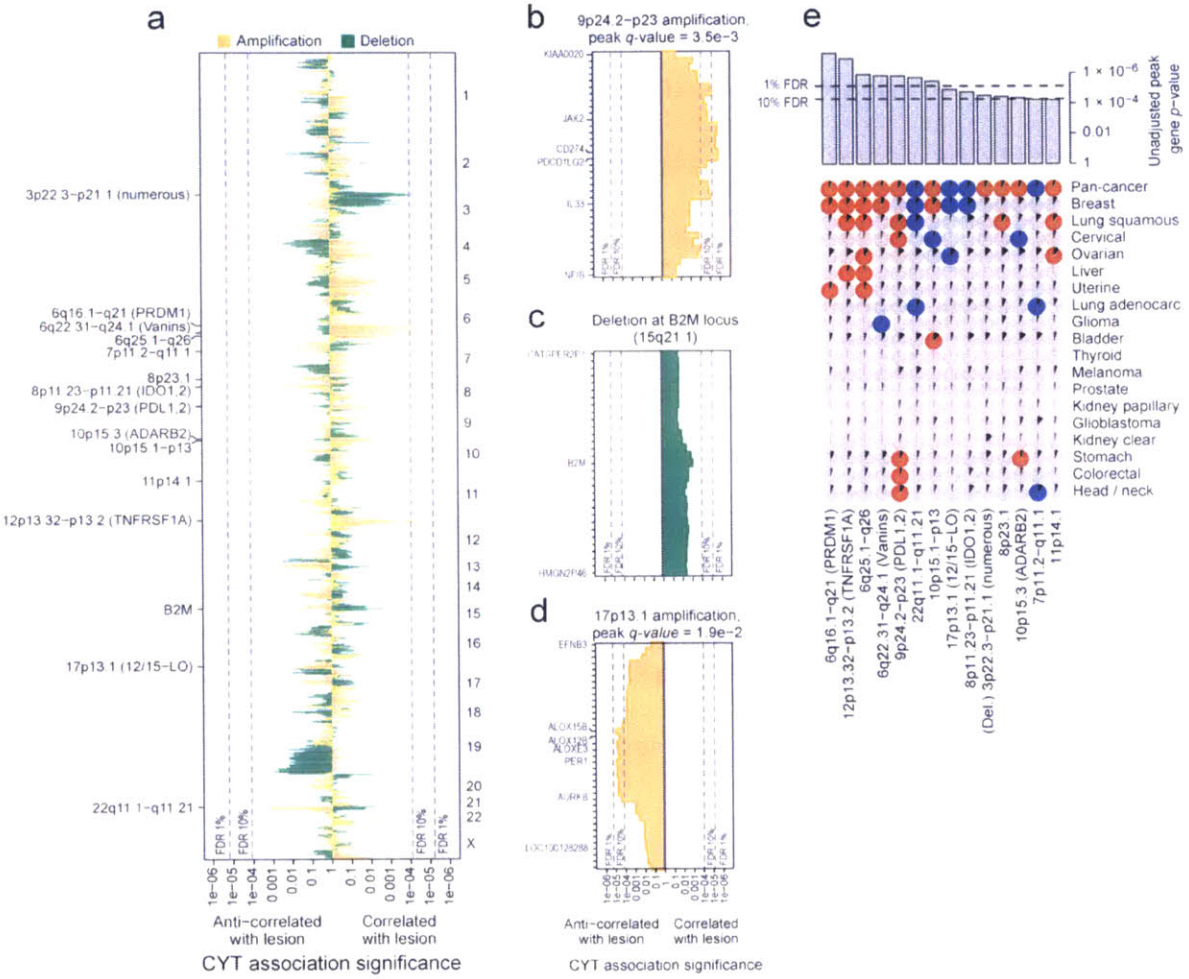


Figure 6. Amplifications and deletions are associated with cytolytic activity in tumors

(A) The significance of association between CYT and amplification (orange) and between CYT and deletion (green) for all genic loci. Upward lines show unadjusted p-values for instances in which the lesion was positively associated with CYT, and downward lines show unadjusted p-values for instances in which the lesion was negatively associated with CYT. Dotted lines represent the significance cutoff yielding 1% and 10% FDRs (and also appear in parts **B-E**). Labels on the right side mark events significant at the 10% FDR, plus B2M. Potential driver genes appear in parentheses. **(B)** Locus zoom on the 9p24.2-p23 amplification, each bar corresponding to a single gene. Labeled genes include those with driver potential or those on the locus boundary. **(C)** Locus zoom on the region containing B2M, which was not genome-wide significant. **(D)** Locus zoom on the 17p13.1 amplification. **(E)** Significant associations between CNAs and CYT on the pan-cancer and cancer-specific level (as in Figure 5). Pan-cancer significance was defined at a 10% FDR, and significance for individual tumor types was defined at unadjusted $p<0.05$. Positive association is indicated with red circles, negative with blue circles, and non-association with gray circles. Black wedges indicate the share of samples exhibiting the event (*ie.* non-zero GISTIC score at the locus). Bar plot indicates unadjusted pan-cancer p-values for CNAs, sorted by significance, with dashed lines indicating thresholds yielding 1% and 10% FDRs. See also **Figure S6** and **Table S7**.

In addition, potential new targets were identified. These included 17p13.1, which was preferentially amplified in low-CYT tumors (**Figure 6D**), including breast and ovarian. The peak genes, ALOX12B/ALOX15B (12/15-LO) regulate immunity in many ways, including blocking the uptake of apoptotic cells by inflammatory monocytes in a manner that decreases antigen presentation to T cells (Uderhardt et al., 2012), which may explain the observed decrease in CYT. Further supporting this model, the amplification was associated with higher necrosis in breast ($p=0.002$) and kidney clear cell cancer ($p=0.0002$), though not ovarian cancer. Other peaks included ones near TNFRSF1A and PRDM1 (**Figure S6A**) as well as a suggestive, but not genome-wide significant enrichment at B2M (**Figure 6C**). In considering how other enrichment signatures might associate with CNAs (**Figure S6B**), we observed a dramatic positive association between increased MHC Class I expression and amplification of the MHC Class II complex (adj. $p<5e-4$).

Discussion

Based on the notion that effective natural anti-tumor immunity requires a cytolytic immune response (**Figure 7A**), we quantified cytolytic activity using a simple expression metric of effector molecules that mediate cytolysis. Our analysis was designed to address which genetic and environmental factors drive tumor-associated cytolytic activity, and how this cytolytic activity selects for genetic resistance in tumors. Our results suggest that neoantigens and viruses are likely to drive cytolytic activity, and reveal known and novel mutations that enable tumors to resist immune attack.

We considered several explanations for the elevated immune cytolytic activity observed in some tumors (**Figure 7A**). First, we asked whether neoantigens play a role. These are a compelling set of antigens because of their absence from the thymus and thus lack of central tolerance that would normally delete cognate high-affinity T cells. Indeed, we found that neoantigen load positively associated with cytolytic activity across multiple tumor types, and that neoantigens appear to be depleted in tumors relative to their expected numbers based on the silent mutation rate, consistent with the notion of immunoediting (Schreiber et al., 2011). Second, when we analyzed CT antigens that are expressed selectively in tumors, we could not detect a positive correlation between the number of expressed CT genes and cytolytic activity. In addition, CT antigen genes were not contained within deletions associated with CYT, contrary to what would be expected if there were immune pressure on CT antigens. Although we did not uncover a role for CT antigens in spontaneous immunity (perhaps because our methods were not optimized to detect CT depletion), we did highlight a subset of 60 CT antigens that are

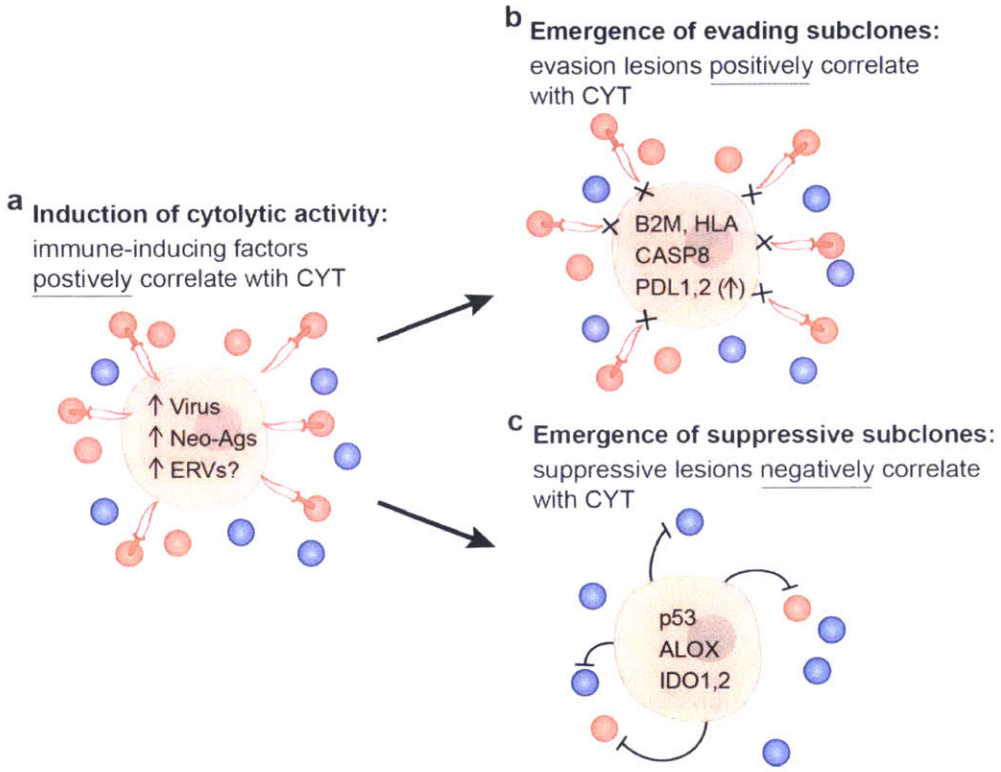


Figure 7. Proposed model for evolution of tumor-immune associations
(A) As the tumor develops, we propose that intrinsic tumor factors – such as mutated neoantigens or viruses – induce local immune infiltrates (blue circles) that include cytolytic effector cells (expressing GZMA/PRF1; red circles) that kill tumors (daggers). These factors are expected to be positively correlated with CYT across tumors. **(B,C)** Under pressure from cytolytic immune cells, subclones with resistance mutations will grow out over time. **(B)** One subset of these mutations would enable tumors to evade killing, but does not impact the infiltrate, and are positively correlated with CYT (i.e. higher infiltrate samples are enriched for these mutations). **(C)** Another subset suppresses the immune infiltrate (i.e. lower infiltrate samples are enriched for these mutations), and is negatively correlated with CYT. Notably, p53 mutations and ALOX amplifications were also significantly negatively associated with CD8A, suggesting a reduction in cell numbers and not just activity. See also **Figure S7** and **Table S8**.

highly tumor-specific and may be (or are already) excellent targets for immunotherapy, including vaccines, adoptive T cell transfer or CAR-T therapy. Third, we asked whether viruses could be inducers of immune responses. In some tumors, we observed that cytolytic activity does indeed associate with the presence of exogenous or endogenous viruses, and we expect that some viruses would trigger immunity through RNA and DNA sensors and generate immunogenic antigens for the adaptive immune response.

To learn more about how tumors adapt to attack by cytolytic immune cells, we also searched for enrichment of somatic genetic alterations in tumors with high versus low cytolytic activity. As expected, we observed enrichment of mutations in antigen presentation machinery (thus validating our cytolytic metric), including HLA and B2M, as well as extrinsic apoptosis genes, such as CASP8, that would prevent cytolytic cells from killing tumors via FasL-Fas interactions. In addition, we found cytolytic activity correlating with amplifications in regions containing genes that function in immunosuppression, such as PDL1/2. Most of the identified mutations – including HLA, B2M and CASP8 – were positively correlated with CYT and are likely to represent autonomous escape mechanisms (**Figure 7B**). In addition, we identified a smaller number of mutations that correlated negatively with cytolytic activity -- including IDO1 and IDO2, p53, and the ALOX locus – and may represent non-autonomous mechanisms of suppressing immunity (**Figure 7C**). Finally, we were surprised that CYT-associated genetic lesions represent ~10% of drivers, and these genes had largely not been studied in the context of immunity. However, given the importance of immune responses in controlling tumor progression (Pages et al., 2005), tumors may have evolved several mechanisms of evasion.

Our approach has allowed us to positively identify the subset of tumor types that are sensitive to spontaneous cytolytic activity (**Figure S7**, **Table S8**). If we consider positive correlation of HLA, B2M or CASP8 mutations with CYT as a 'signature' of selection pressure by the immune system, we find that colorectal, uterine, stomach, head and neck, cervical, lung squamous and breast tumors are most susceptible to immune elimination. If we further consider depletion of neoepitopes as an independent signature of selection, we identify colorectal as well as kidney clear cell cancer as immune-susceptible tumors. For these tumor types, we thus suggest that spontaneous tumor immunity can delete tumor cells.

For several tumor types, we did not find evidence for immunoediting. This could be due to: insufficient power to detect associations in tumors with low rates of spontaneous immunity, non-genetic evasion mechanisms that we cannot detect, or true absence of immune cytolytic activity (perhaps for thyroid and prostate cancers, for example).

Finally, the mutations associated with cytolytic activity reveal potential genetic biomarkers for predicting outcome and candidate targets for immunotherapy. To assess the utility of these markers, one would need to genotype tumors for the 35 identified genes at clonal or subclonal levels, and test if pre-treatment or post-treatment mutations predict refractoriness or relapse in response to cytolytic immunotherapy. We predict that the presence of these mutations (assuming they do not lead to complete loss of susceptibility) indicates that re-activation of CD8 T cells would be therapeutically effective. In addition, we identified new candidates for therapeutic development, including the ALOX enzymes and their products, the PIK3CA protein that is enriched in activating mutations in high-CYT stomach cancers, and FASL which may be

useful to upregulate in T cells to enhance the anti-tumor activity of adoptively transferred T cells.

Analysis of TCGA samples has revealed environmental and genetic mechanisms that impact tumor-immune interactions. While we chose to focus on cytolytic activity because of its central role in tumor elimination and the feasibility of monitoring its activity, we did not consider other tumoricidal activities (such as antibody-dependent cell-mediated cytotoxicity) because we are not aware of transcript-based markers for these activities. In addition, the CYT metric we used is transcript-based and thus may not reflect changes in cytolytic activity due to post-transcriptional regulation, and is a snapshot in time that may miss previous activity that impacted tumor growth. We anticipate that improved experimental measurements of anti-tumor immune activity will further reveal the genetic and epigenetic changes that underlie co-evolution of tumor cells and immune cells.

Methods

Selection of tumor types

Tumor types were selected for analysis based on publication availability as determined by The Cancer Genome Atlas (TCGA) embargo dates in September 2014, excluding non-solid tumor types. The analyzed tumor types and their corresponding project codes were urothelial bladder cancer (BLCA), breast cancer (BRCA), cervical cancer (CESC), colon and rectal adenocarcinoma (COAD and READ, a.k.a. CRC), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), clear cell kidney carcinoma (KIRC), papillary kidney carcinoma (KIRP), lower grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), papillary thyroid carcinoma (THCA), and uterine corpus endometrial carcinoma (UCEC). Clinical data for these samples were accessed through the TCGA Portal (National Institute of Health) ftp on March 26, 2014 (https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/<*tumortype*>/bcr/nationwidechildrens.org/bio/clin/nationwidechildrens.org_<*tumortype*>.bio.Level_2.X.X.X/). These data included assessment of histological subtype, tumor stage, specimen characteristics (such as percent necrosis), patient survival, and smoking history. Analyzed samples represent untreated primary tumors, except for melanoma, which includes untreated metastases. Melanoma metastases to lymph nodes were excluded from all analyses. Patients that received some form of neo-adjuvant therapy were excluded.

Data types, sources, and initial processing

Different types of data were considered: gene expression (RNA-Seq and array-based), viral expression, endogenous retrovirus expression, HLA type (and mutational status), point mutation (as identified by whole exome sequencing (WES)), neopeptide HLA-binder predictions, copy number alteration (CNA) data, and reference gene expression profiles. These data were obtained from TCGA (http://cancergenome.nih.gov/), the Genotype-Tissue Expression project (GTEx) (GTEx Consortium, 2013a), Fantom5 (Fantom Consortium et al., 2014), and/or the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012). Some data were accessed and used directly, and some required post-processing to generate, as described below.

Clinical data

Clinical data for each tumor type were accessed from the TCGA public access web portal on March 26, 2014. These data included assessment of histological subtype, microsattelite instability status, tumor stage, specimen characteristics (such as percent necrosis), patient survival, and smoking history.

RNA-Seq-based gene expression data

TCGA gene expression data (90% tumor biopsies, 10% solid tissue controls) were obtained through GDAC Firehose (Broad Institute TCGA Genome Data Analysis Center, 2014) and included all available "Level_3" gene-level data (a mix of Illumina HiSeq and Illumina GA data). Samples from Genotype-Tissue Expression project (GTEx) were accessed through the GTEx web portal in November 2013 (GTEx Consortium, 2013b). For both data sets, raw read counts were tallied per gene symbol and divided by the gene symbol's maximum transcript length to

represent coverage depth. Transcript lengths and mappings between gene symbols and transcript IDs were obtained from UCSC Genome Browser's table "knownIsoforms" (hg19 version) (Karolchik et al., 2004). For each sample, the corresponding coverage estimates across all genes were rescaled to sum to a total depth of 1e6, such that expression estimates may be interpreted as Transcripts Per Million transcripts (TPM). A listing of patient IDs for TCGA tumor samples can be found in **Table S1A**, which also lists the histological subtype of each tumor, if defined. A listing of IDs for TCGA normal and GTEx samples can be found in **Table S1B**, which also indicates tissue type/lineage designations for these samples.

RNA-Seq-based sequence data

TCGA data were accessed from CGHub and included TCGA .bam files with "RNA-Seq" indicated in the library strategy field. Because this data set was too large to store locally, analyses were conducted "on-the-fly." Therefore, analyses based on the TCGA RNA-Seq sequence data do not always comprise the same samples set (reflecting the ongoing additional of samples to CGHub). RNA-Seq .bams for GTEx were downloaded from Short Read Archive (SRP012682, corresponding to dbGap phs000424) on July 1, 2014. **Table S1A** and **S1B** list the samples used in each analysis.

Microarray-based gene expression data

While RNA-Seq-based gene expression data were used for most analyses, microarray-based data were used for assessing the baseline expression of GZMA and PRF1 in non-TCGA cancer cell lines. Data were obtained through the CCLE web portal (http://www.broadinstitute.org/ccle/home; file: CCLE_Expression_2012-2009-2029.res), and

probes 205488_at and 1553681_a_at were used to represent GZMA and PRF1, respectively (Affymetrix U133+2 platform; RMA processing).

CAGE cell type expression profiles

Human cell type gene expression profiles were downloaded from the Fantom5 website on October 8, 2014: http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/hg19.cage_peak_tpm_ann.osc.txt.gz

Point mutation data

When possible, data was obtained from TumorPortal (Lawrence et al., 2014), which supplies the following ".maf" (Mutation Annotation Format) file: http://cancergenome.broadinstitute.org/data/per_ttype_mafs/PanCan.maf. When a patient was not present in this .maf, mutation data were obtained from Cyriac Kandoth's Synapse workspace syn1729383 (https://www.synapse.org/#!Synapse:syn1729383; corresponding paper (Kandoth et al., 2013)). When a patient was not available in the previous sources, data were obtained from the TCGA Data Portal (National Institute of Health) from the files:

genome.wustl.edu_CESC.IlluminaGA_DNASeq_curated.Level_2.1.0.0
broad.mit.edu_BLCA.IlluminaGA_DNASeq_curated.Level_2.1.4.0
broad.mit.edu_PRAD.IlluminaGA_DNASeq_curated.Level_2.1.4.0
broad.mit.edu_KIRP.IlluminaGA_DNASeq_curated.Level_2.1.1.0
broad.mit.edu_LGG.IlluminaGA_DNASeq_curated.Level_2.1.2.0

When not available in the previous sources, data for liver cancer patients was obtained from the GDAC Firehose standard analysis pipeline (Broad Institute TCGA Genome Data Analysis Center, 2014) (accessed August 8, 2014). Finally, data from the recent TCGA stomach adenocarcinoma analysis (Cancer Genome Atlas Research Network, 2014) (https://tcga-data.nci.nih.gov/docs/publications/stad_2014/; file: "Public Mutations") were included and used preferentially when a patient was already in one of the above sets.

Several possible mutation-calling artifacts were identified for the genes ZNF43, XYLT2, PAX6, PAFAH1B1, ALPK2. In each case, the gene had a reported indel appearing in multiple subjects at the edge of a homopolymer stretch. These events were manually excised from the maf.

HLA type, HLA mutations, and predicted neo-antigen binders

Whole exome sequencing data (.bam's) were downloaded from CGHub (University of California, 2012) for all samples for which a tumor and normal sample were available for a given patient (hg19-mapped .bam's were used when available; all files included unmapped reads). The 4-digit HLA type for each sample was inferred using the POLYSOLVER (POLYmorphic loci reSOLVER) tool which uses a normal tissue .bam file as input and employs a Bayesian classifier to determine genotype (Shukla et al, manuscript in review, *Nature Biotechnology*). The algorithm selects and aligns putative HLA reads to an imputed library of full-length genomic HLA allele sequences. The alignments then serve as a basis for the inference step that incorporates the number and base qualities of aligned reads, the empirical library insert size distribution and population-based allele frequencies.

Because standard mutation calling algorithms are not well-equipped to deal with highly variant regions such as the MHC loci, mutations in class I HLA genes were determined using the POLYSOLVER-based mutation detection pipeline (Shukla et al, manuscript in review, *Nature Biotechnology*) that takes a tumor/germline exome pair as input, and first characterizes the HLA alleles in the individual by applying POLYSOLVER on the germline data. Putative HLA reads from both the tumor and germline exomes are then aligned to the inferred alleles separately and likely erroneous alignments are filtered out. Somatic changes are subsequently identified by comparative evaluation of the aligned tumor and germline files using the Mutect (Cibulskis et al., 2013) and Strelka (Saunders et al., 2012) tools. Since CGHub contains .bam files that have not yet been processed into .maf's by the TCGA Data Coordinating Center, there are more patients with HLA mutation calls than patients appearing in the general mutation .maf.

Individual-specific HLA-binding peptides were identified by a neo-antigen prediction pipeline (Rajasagi et al., 2014) that uses all detected somatic mutations for the individual (obtained from the general mutation .maf). Binding affinities of all possible 9 and 10-mer mutant peptides to the corresponding POLYSOLVER-inferred HLA alleles were predicted using NetMHCpan (v2.4) (Nielsen et al., 2007b).

Viral expression

Variant sequences for ten putative oncoviruses were accessed from NCBI Nucleotide (http://www.ncbi.nlm.nih.gov/nucleotide/) using the following search terms:

Virus	Query	Retrieved Count
JC polyomavirus	JC Polyomavirus[Organism] AND "complete genome" AND 5000:5400[Sequence Length]	564

BK polyomavirus	BK Polyomavirus[Organism] AND "complete genome" AND 5000:5400[Sequence Length]	282
KI polyomavirus	KI Polyomavirus[Organism] AND "complete genome" AND 5000:5400[Sequence Length]	10
WU polyomavirus	WU Polyomavirus[Organism] AND "complete genome" AND 5000:5400[Sequence Length]	80
Merkel cell polyomavirus	Merkel cell polyomavirus[Organism] AND "complete genome" AND 5200:5700[Sequence Length]	42
Human papillomavirus	Human papillomavirus AND "complete genome" AND 5000:10000[Sequence Length]	741
Epstein-Barr virus	Human herpesvirus 4[Organism] AND "complete genome" AND 150000:200000[Sequence Length]	9
Kaposi sarcoma virus	Human herpesvirus 8[Organism] AND "complete genome" AND 130000:140000[Sequence Length]	4
Hepatitis B virus	Hepatitis B virus[Organism] AND "complete genome" AND 2800:3500[Sequence Length]	4834
Hepatitis C virus	Hepatitis C virus[Organism] AND "complete genome" AND 9000:10000[Sequence Length]	912

These sequences were then filtered using Tandem Repeat Finder (Benson, 1999) using options "2 7 7 80 10 24 50 -m -h" to mask low-complexity sequences. These sequences, as well as a decoy fasta of homopolymer repeats, were concatenated into a single fasta and converted into a bowtie2 (Langmead and Salzberg, 2012) search index. TCGA RNA-Seq .bam files were downloaded from CGHub, and the unmapped reads were mapped using bowtie2 and search parameters "-q --end-to-end -k 1 --no-unal". Because the read mapping pipeline used to generate the .bam's hosted at CGHub does not produced files fully consistent with SAM format (Li et al., 2009), it was not possible to revert the .bam's to paired mate-1 and mate-2 .fastq files (UNC Lineberger Comprehensive Cancer Center, 2013). Therefore, reads were mapped in

single-end mode, and it was considered after-the-fact whether reads with the same name suffix had both aligned to the same virus. For consistency, the same approach was used for mapping the GTEx RNA-Seq data. For the TCGA data, fragment ends had been sequenced to 48 or 76 bases, and each .bam contained 114,000,000-181,000,000 mapped read ends (IQR). For GTEx, fragment ends had been sequenced to 76 bases, and each .bam contained 63,000,000-110,000,000 mapped read ends (IQR). Due to lags in sample processing in TCGA and GTEx, the counts of processed RNA-Seq .bam's do not exactly match the corresponding counts of samples with full-transcriptome expression estimates (described above). **TableS1A** and **TableS1B** list the samples for which viral expression was quantified.

Upon pilot analysis, the requirement that both read ends successfully map appeared to be inadequately sensitive; therefore, the paired-end nature of the data was ignored for the purposes of the viral analysis. To guard against false positives caused by spurious read mapping, non-zero viral expression calls were contingent on reads mapping to multiple loci in the viral reference sequence. For a given virus, the specific operation was to 1) identify all viral reference sequence covered by at least one read and 2) determine whether the count of unique 20mers within that sequence (as calculated using Jellyfish (Marcais and Kingsford, 2011)) is greater than 300 (it was not possible to simply measure the length of covered sequence because multiple fastas were used to represent each virus). Given successful clearance of this robustness test, viral expression was quantified by taking the count of read ends mapping to each virus (summing across the variant sequences for the virus) and dividing by the total count of human genome-mapped read ends in the original .bam. This number was multiplied by 1e6 in order to express viral titer as viral Reads Per Million reads mapped to the

human genome (RPM). "Positivity" for viral infection was determined based on whether the expression exceeded the maximum observed in the GTEx normal, with the assumption that very low levels may simply reflect the trace presence of previously exposed leukocytes in the tissue sample. We acknowledge that these are conservative assessments and may miss some cases of viral infection that are transcriptionally silent. A table of expression values from both studies can be found in **Table S2A**.

Endogenous retrovirus (ERV) expression

A list of GenBank accessions corresponding to transcriptionally active endogenous retroviruses was obtained (Mayer et al., 2011) and contained sequences representing 66 ERV species. These were converted into a bowtie2 index, and .bams from TCGA and GTEx were remapped to this index (preserving both mapped and unmapped reads) using the same bowtie2 search parameters used in the viral analysis. Paired read ends were assigned to the same ERV sequence if there was one that provided an acceptable alignment for both; otherwise, the reads ends were aligned individually. When multiple ERV sequences provided an equally good match, ties were broken at random. ERV expression was quantified in RPM in the same manner used in the viral analysis (**Table S5B**).

Copy number events

GISTIC2 (Mermel et al., 2011) "Level 4" copy number calls for each patient were accessed from GDAC Firehose in March 2014. GISTIC2 uses data from copy number arrays (in this instance, Affymetrix Genome-Wide Human SNP 6.0 arrays) to identify regions of copy number variation across the genome. The files accessed through GDAC Firehose contained a score for each gene

representing whether that gene was in a region that was focally amplified or deleted in the given tumor (larger events, such as whole genome amplifications, were ignored). Values of zero indicate that there is no evidence of copy number alteration for the given gene, whereas positive and negative values represent amplification and deletion, respectively. Even though each tumor subclone contains an integral copy number for each locus in the genome, biopsies potentially contain multiple tumor subpopulations as well as stromal tissues; therefore, the reported values in the GISTIC2 output are continuous rather than integral. Since stromal contamination (which would presumably correlate with CYT) tends to regress the signal toward zero, each sample was rescaled so that the median non-zero event amplitude was 1.

Data Analysis

Definition of cytolytic activity metric "CYT"

Cytolytic Activity Metric (CYT) was obtained by calculating the geometric mean of GZMA and PRF1 expression (as measured in TPM) per sample. The geometric mean was preferred to the arithmetic mean because it is not arbitrarily affected by the expression scales of the two genes being averaged. Because the geometric mean function requires a log transformation, 0.01 was added to each expression value before transformation in order to avoid logging zero. In order to assure robust statistical results, some analyses (association with genetic alterations) included an additional rank-transformation (across samples) to the CYT values, which was rescaled such that the values were uniformly distributed between 0 and 100. GZMA and PRF1 were selected based on their known roles in target cell lysis in addition to corroborating expression profile-

based evidence that they were specific to killer lymphocytes (a point on which GZMB and FASL, other well-known effector genes, failed).

Definition and analysis of cell type expression markers and immune meta-genes

Cap analysis of gene expression (CAGE) data from Fantom5 were used to define a set of transcriptomic markers for immune cell subtypes of interest. We note that "regulatory T cells" in Fantom5 were represented by CD4+CD25^{hi}CD45ra- cells. We further note that "myeloid dendritic cells" (mDCs) in Fantom5 were monocyte-derived rather than primary and were therefore not used. Data were collapsed to the gene symbol level using summation. For each gene in each cell type, a median expression level was calculated over the given replicates (an offset of 5 TPM was added to all expression values). To determine a specificity ratio for markers, the expression of each gene was compared to maximum expression of the gene in the other immune cell types (listed in **Table S1C; B, Treg, NK, CD8 T cell, neutrophil, macrophage, pDC**) as well as all non-hematopoietic, non-cancer cell types in FANTOM5. This specificity ratio had to be at least 2 to consider a gene as a marker for a cell type, with up to 10 markers per cell type. Because the Fantom5 project did not include activated/effector CD8+ T cells, many of the genes initially identified as NK-specific within Fantom5 were actually shared between NK cells and activated CD8+ T cells when we considered data from the DMAP human blood profiling project (http://www.broadinstitute.org/dmap) (Novershtern et al., 2011). Therefore, we used data from the DMAP project to find the genes most highly expressed in NK cells (median of types "A1", "A2", "A3", and "A4") vs. activated/effector CD8 T cells (median of types "T cell 1", and "T cell3"), identified the top 20, and then obtained a revised NK marker gene list by intersecting the FANTOM NK markers with the DMAP NK markers. We note, however, that even

these genes exhibited a substantial degree of expression in activated CTLs, consistent with the lack of known highly-specific NK markers.

Several other meta-genes were defined. Sets of co-inhibitory and co-stimulatory receptors expressed on T cells and antigen presenting cells (APCs) were defined based on a recent review (Chen and Flies, 2013). Type I and Type II-specific interferon response genes were defined based on recent study comparing responses of macrophages to these two stimuli (Liu et al., 2012) (Supplementary Table 1 of referenced). HLA Class I genes were defined as HLA-A, B2M, and TAP1. The final set of selected markers can be found in **Table S1C**.

The enrichment of a cell type meta-gene in a given sample was then calculated using single sample gene set enrichment analysis (ssGSEA) (Barbie et al., 2009) as used before to analyze TCGA samples for immune/stromal infiltrates and implemented in the 'GSVA' R package (Hänzelmann et al., 2013), with subsequent z-scoring across samples. Note that these enrichments should not be interpreted as deconvolutions of actual cell type proportions. In several instances in which CYT was directly compared to the ssGSEA enrichments, CYT was also calculated according to the ssGSEA approach (rather than geometric mean) in order to make a fair comparison. These include **Figure S1F** (cell type enrichments *vs.* CYT), **Figure S1G** (tumor-normal comparison of enrichments), **Figure S1J** (survival analysis of enrichments). We note, however, that the two CYT calculations are nearly identical (Spearman correlation 0.96). Values for (log-average) CYT and the ssGSEA enrichment scores can be found in **Table S1A** (tumor samples) and **Table S1B** (normal samples).

Relationships with tumor stage

To test for an overall association between CYT and tumor stage, the Pearson correlation was calculated between log-CYT and stage (stage was converted to a numeric variable: "stage 1"=1, "stage 2"=2, *etc.*) Z-scored ssGSEA enrichments of marker genes were also compared to stage in this manner. Since gliomas are not staged, grade (G2 or G3) was analyzed in place of stage for this tumor type.

Survival analysis

Patient samples grouped according to histological subtype (samples were excluded when histological subtype was not available) and tumor stage. Groups with fewer than 8 samples were excluded. To assess the survival effect of a continuous variable *x*, each group was split equally into high-*x* and low-*x* patients. High-*x* patients were pooled pan-cancer and low-*x* patients were pooled pan-cancer and analyzed as two distinct cohorts using Cox proportional hazards modeling. Note that the two cohorts have identical admixtures of tumor type and stage. In some analyses, the variable *x* represented the z-scored ssGSEA enrichment of a metagene (*e.g.* macrophage marker genes); in other analyses it was the arithmetic difference between the z-scored ssGSEA enrichments of two meta-genes (*e.g.* Treg marker gene enrichment minus CTL marker gene enrichment). In contrast to most other analyses, CYT was calculated according to the ssGSEA approach to enable the analysis of differential enrichment with respect to the meta-genes.

Transcriptomic assessment of viral infection

Association between CYT and viral infection status was characterized using Wilcoxon rank sum tests for tumor types exhibiting at least five cases of infection with the given virus.

To further characterize viral transcription, representative variants (one for which a large number of reads mapped and for which there exists good gene annotation) were selected for each of virus that was detected, and remapped reads (pooled from all TCGA cancer samples) to these variants. These read depths are presented in **Figure S2A**.

In order to assess the general gene expression correlates of viral infection in a given tumor type, a Wilcoxon rank-sum test was performed for each gene to test differential expression between infected and non-infected, and a score was assigned by multiplying the sign of the association by the negative log p-value. Genes were ranked by this differential expression score and submitted in forward and reverse order to "GOrilla" gene ontology enrichment analysis and visualization tool (Eden et al., 2007; Eden et al., 2009), to assess for gene set enrichment (results reported in **Table S2B**).

To assess whether extra-hepatic cases of HBV infection were metastases originating from the liver, samples from tumor types with at least 1 HBV+ cases were plotted according to the first two principal components of their global log-transformed gene expression. The clustering of HBV+ samples (with liver or with the uninfected samples of the corresponding tumor type) was assessed visually.

Association between HLA type and cytolytic activity

HLA types (at two-digit granularity) were assessed for association for cytolytic activity in each tumor type using Wilcoxon rank sum tests. The overall significance of an HLA type pan-cancer was assessed using Fisher's method to combine the p-values of the individual tumor type Wilcoxon rank sum test p-values. The overall significance of a tumor type for HLA-CYT association was assessed using an F test of a linear regression modeling rank-scaled CYT in that tumor type as a function of HLA type.

Characterization of mutational spectra

Using the general-analysis .maf (which contains only coding region mutations), single-nucleotide variants were identified and characterized as C→A, C→G, C→T, A→C, A→G, or A→T (if the reference allele was T or G, the event was analyzed from the perspective of the opposite strand). In addition, the identities of the upstream and downstream reference bases were used to further categorize the mutational events. **Figure S2E** depicts the rate of each mutation type, per sample, for high-CYT tumors and low-CYT tumors as well the difference in the rates (high minus low). High-CYT tumors were defined as those with CYT in the top quartile for the given tumor type. Low-CYT tumors were defined as those with CYT in the bottom quartile for the given tumor type.

To test whether the rate of Apobec-characteristic mutations (reference allele C with upstream T) was differential between high-CYT and low-CYT tumors, the count of Apobec-characteristic mutations in each tumor samples was divided by the count of all other mutations and this ratio was assessed for Spearman rank correlation with CYT. The ratio was tested for association with

viral infection status and with ERV expression using Wilcoxon rank-sum test and Spearman rank correlation, respectively. For the ERV analysis, p-values were corrected by B-H method across the 18-cancer × 66-ERV matrix of p-values. While several ERVs association with Apobec-characteristic mutations narrowly reached significance (p adj.<.05) in stomach cancer (ERVH-2, ERVE-2) and breast cancer (ERVI-1), the directions of association were not consistent amongst tumor types leaving no definitive result.

Neo-antigen analysis

If the mutation was predicted to produce a "binder" neopeptide with affinity <500 nM and if the corresponding gene was expressed greater than 10 TPM (evaluated based on median expression in the given tumor type rather than the specific sample, as mutations may affect transcript quantification), the mutation was designated as putatively antigenic. For each tumor type, the count of total mutations and the count putatively antigenic mutations per sample was compared to the CYT. Tumor types displaying a spearman rank correlation p-value less than 0.1 (only significant positive associations were observed) are presented in **Figures 3A** and **3B**. For each cancer type, a local regression curve (as calculated by the R lowess() implementation (Cleveland, 1981), default parameters) is drawn over inner 90th percentile range of the independent variable. The raw data for these curves can be found in **Table S4A**.

To determine whether the number of neo-antigens predicted for a tumor was more or less than expected given its mutation rate, a null model for mutation was developed to control for the differing rates of mutational "spectra" observed in different tumors (a result of differing mutagenic processes). Indels and mutations in genes significantly mutated in cancer (**Table S6A**;

described in a later section) were excluded from the analysis. 192 mutational spectra were defined based on the old base, the new base, and the identities of the nucleotides 1 base upstream and 1 base downstream (from the perspective of the coding strand). For each spectrum *s*, two rates were estimated empirically pan-cancer: the expected number of non-silent mutations per silent mutation, $\bar{N}_s$, and the expected number of high-affinity neo-peptide binders (not considering gene expression) per non-silent mutation, $\bar{B}_s$,. Using these rates, we used the silent mutational events in each tumor sample to predict the number of non-silent mutations, N_{pred}, and the number of neo-peptide binders, B_{pred}, expected for that tumor under null model in which there is no selection against mutations that yield HLA binders:

$$N_{pred} = \sum_{m}^{Silent\ SNVs} \bar{N}_{s(m)}$$

$$B_{pred} = \sum_{m}^{Silent\ SNVs} \bar{N}_{s(m)}\bar{B}_{s(m)}$$

where *s(m)* represents the spectrum of the given mutation. Having calculated N_{pred} and B_{pred} for a sample, these were compared to the actual counts in the sample, N_{obs} and B_{obs}, to define the ratio between the observed and expected rate of neo-peptides, *R*:

$$R = \frac{B_{obs}/N_{obs}}{B_{pred}/N_{pred}}$$

R was characterized for the samples corresponding to each tumor type, and Wilcoxon rank sum tests were used to determine whether tumor types were significantly different from *R*=1. Note that since $\bar{N}_s$, and $\bar{B}_s$, were estimated empirically, they are under-estimates if strong selection against binder-yielding mutations is occurring. However, since these values are estimated pan-cancer, *R* can still be interpreted in a relative sense.

As a control, we randomly scrambled HLA genotypes across patients and re-ran the analysis using the resulting new set of predicted neo-epitopes (but still using $\bar{N}_s$ and $\bar{B}_s$ as estimated previously).

Smoking

For the lung cancers, clinical data included the smoking history of the patients. For lung adenocarcinoma and lung squamous cell carcinoma, ever-smokers (excluding those reformed at least 15 years prior and those with an unknown number of years of reform) were compared to never-smokers in terms of CYT using the Wilcoxon rank sum test to assess significance.

Assessing ectopic transcription

In order to define a set of genes whose expression could be considered ectopic and thereby potentially immunogenic, a candidate list was first created using the list of cancer testis (CT) antigens maintained at CTdatabase (http://www.cta.lncc.br/) (Almeida et al., 2009). Using RNA-Seq data from normal samples in GTEx, the 95th percentile expression value was calculated for each tissue type, including blood, as an estimate of the upper bound of the expression of the gene in that tissue type. If no tissue exceeded a threshold of 1 TPM, then the gene was included in our ectopic gene set. This filtering step was applied in order to avoid CT antigens identified genes that may be expressed stromally, which would confound association analyses. The degree of ectopic expression in a given tumor sample was determined by counting the number of ectopic genes expressed greater than 1 TPM. For each tumor type, association with CYT was determined by binning samples by the count of ectopic genes expressed greater than 1 TPM (bins: 0 genes, 1-5 genes, 6-10 genes, and >10 genes); for adjacent bins containing ≥10

samples, CYT was compared by Wilcoxon rank-sum test. (We note that an alternative approach, in which the association was assessed by Spearman rank correlation, did not show any compelling cases of positive association.) Separately, the expression levels of individual ectopic genes were assessed for association with CYT using Spearman rank correlation.

To explore the hypothesis that CT antigens might be chromosomally deleted as a mode of immune evasion, we explored several properties of their copy number alteration status. First, we determined whether deletion reduced the expression of each gene in each tumor type by assessing for whether there was a significant negative Pearson correlation between the gene's GISTIC deletion signal (0-censoring values in the direction of amplification) and its log expression (using a log offset of +1 TPM). Second, we determined whether high CYT was associated with deletion by looking for significant positive Pearson correlation between each gene's GISTIC deletion signal (0-censoring values in the direction of amplification) and log-scale CYT. Finally, for each tumor type, we calculated the count of instances in which each gene was deleted (GISTIC score < 0), divided it by the count of total alterations (GISTIC score $\neq 0$), and calculated the average across all genes. Using this deletion:alteration ratio, we calculated whether each CT gene was significantly more frequently deleted than amplified in comparison to genes in that tumor type in general (according to a binomial distribution). Even with loose thresholds, there were no instances in which the three tests agreed for a given gene-cancer combination (**Figure S4C**).

Analysis of ERV expression

To define a set of tumor-specific ERVs (TSERVs), the 95th percentile expression value was calculated for each ERV in each tumor tissue type and each normal tissue type. This value was considered to represent a robust estimate of the upper limit of the expression range. If this value did not exceed 10 RPM in any normal tissue type, did exceed 10 RPM in at least one tumor tissue type, and if there existed at least one tumor tissue type with a value 5-fold greater than any normal tissue type, then the corresponding ERV was considered to be a TSERV.

Functional motifs within ERV sequences were obtained by determining the consensus sequence for each ERV (among aligning reads), translating all ORFs greater than length 75 and processing using InterProScan (Jones et al., 2014).

For each TSERV, gene set enrichment analysis was performed for the tumor type demonstrating maximum expression. This was done in the same fashion as for the viral gene set enrichment analysis, but using Spearman rank correlation to determine sign and p-value rather than Wilcoxon rank-sum test.

For all ERVs that exhibited overexpression in a given tumor type (as defined by expression exceeding that observed in normal tissues), ERV-CYT expression was assessed using Spearman rank correlation.

Correlates of necrosis

Association between percent necrosis (based on TCGA clinical data) and various meta-genes (including CYT) was assessed using Spearman rank correlation.

Identifying genes significantly point-mutated in high-/low-CYT tumor biopsies

A set of candidate genes was defined by running MutSigCV on each individual tumor type and on the entire pan-cancer .maf. MutSigCV is a tool designed to identify genes that are mutated in a non-random manner and considers variables such as the ratio of nonsynonymous to synonymous events (Lawrence et al., 2013). As described previously, the .maf contains "point mutations" (SNV, DNVs, indels and other variants that can be identified using whole exome sequencing) but excludes larger chromosomal derangements. In line with previous application of MutSigCV (Lawrence et al., 2014), genes significant at a 10% false discovery rate in any of these MutSgCV runs were deemed to be significantly mutated. Genes that were not identified in this analysis but were identified in previous pan-cancer MutSigCV application (Lawrence et al., 2014), were added to the candidate list. **Table S6A** presents the full list of candidate genes and the analysis(-es) supporting the inclusion of each (373 genes total).

To assess whether a gene's mutational status was significantly associated with CYT, rank-transformed CYT was modeled (using linear regression) as a function of the gene's mutational status (ignoring synonymous events), cancer type (encoded as dummy variables), and the rank-transformed count of total non-synonymous mutations. The latter two variables were included to diminish confounding effects. Cancer type was defined based on the histological subtype of the tumor (indicated in the clinical data; 40 types total), and samples were excluded when the histological subtype was not defined (**Table S1A** lists the histological subtype and missingness status for each sample). The p-value of the mutation status coefficient ("beta") and its sign were used a measure of enrichment for the given gene. As previously described, rank-scaling transformed data such that values were uniformly distributed between 0 and 100. It was

employed as a conservative measure to avoid results driven by outliers. Thus, beta values (reported in **Table S6B**) should be interpreted as the expected change in CYT percentile given nonsynonymous mutation, and a positive beta value implies a positive relationship between CYT and mutational status. The p-values across the 373 genes tested were corrected for multiple hypothesis testing using "method=BH" (Benjamini & Hochberg) in R's p.adjust() function. A set of "hits" was defined by setting an adjusted p-value cutoff of 0.1. (The pan-cancer association analysis was also conducted using synonymous mutations only and ignoring nonsynonymous events. This was to determine whether any "hits" would be discovered in a scenario in which none were expected.)

Since HLA mutations were called through a separate pipeline that could be applied to all available CGHub .bam files, the calls were available for a much larger number of samples than appeared in our pan-can .maf. Analyzed on the larger set, HLA mutation reached an adjusted p-value of 3.0e-13 for CYT association, with 5 tumor types (colorectal, head and neck, uterine, stomach, and cervical cancer) independently showing this association; however, this analysis could not apply the correction for background mutation rate given the absence of mutation calls for other genes.

To further characterize each hit, the data was parsed into 18 subsets corresponding to each tumor type, CYT was re-rank-transformed per subset, and the linear regression (using the same covariates, excluding cancer type) was repeated on each subset (no control for histological subtype). An uncorrected p-value less than 0.05 for the (nonsynonymous) mutation status

variable was considered evidence for association. The beta values can be interpreted as the expected change in CYT percentile (for the given tumor type) given mutation.

In exploring the relationship between CASP8 mutation and FASL and TRAIL expression (**Figure 5B**), tumor types were analyzed if they had at least 5 instances of CASP8 mutation. For those that did, association p-values were assigned using the Wilcoxon rank sum test.

Associations between the hits and viral infection status were characterized using Fisher's exact test. For testing a given virus, uninfected samples were excluded if demonstrating non-zero transcriptional titer for any virus.

In order to visualize the mutations affecting each significantly CYT-associated gene, a representation was modeled after a popular cancer genomics tool (Gao et al., 2013) (**Figure S5C**). To define the functional subdomains of each gene, the amino acid sequence was processed by InterProScan (Jones et al., 2014) which identified known motifs. When enriched domains overlapped, the smallest was selected for visual representation. In order to depict clusters of mutation, the local density of mutations was depicted using the density() function in R, specifying a smoothing bandwidth of 30 nucleotides.

Though CYT was the primary focus, we also explored whether other cell type signatures (quantified by ssGSEA) would have mutational associations. For this, we used the same "hit" identification pipeline as described above for CYT.

To identify genes specifically mutated in MSI-high vs. MSI-low/MSS tumors, Fisher's exact test was used to test for enrichment of non-silent mutation status in each of the 373 candidate genes. P-values were adjusted using the Benjamini Hochberg (BH) method.

A set of additional immune genes, which were frequently mutated in cancer but did not show mutational associations with CYT or the cell type expression markers, were assessed in terms of their gene expression correlates using an unbiased approach. These genes included *CARD11*, *CD1D*, *CLEC4E*, *CXCL9*, *IFITM1*, *IL32*, *IL7R*, *IRF4*, *MYD88*, *PRDM1*, *TAP1*, *TNF*, and *TNFRSF14*. To characterize the gene expression correlates of a given gene's mutation, Wilcoxon rank-sum tests were applied to all genes' expression profiles within the tumor type exhibiting the strongest MutSigCV p-value for the gene in question. Association scores were defined by multiplying the association sign by the negative log p-value, and genes were sorted by score and submitted to GOrilla (in forward and reverse order).

Identifying copy number alterations (CNAs) significantly enriched in high-/low-CYT tumor biopsies

To test for CNA association, a regression approach was utilized similar to that used for the point mutation analysis. To test a given gene, rank-scaled CYT across all TCGA tumor samples was modeled as a function of the gene's copy number, cancer type (at the histological subtype level, as described previously), and three variables representing the overall copy number disruption of each tumor. These latter three variables were meant as additional controls for stromal biopsy fraction (which may negatively impact the ability to make focal amplification/deletion calls) and included 1) a rank-scaled count of genes with positive copy number signal 2) a rank-scaled count of genes with negative copy number signal and 3) a rank-scaled estimate of the

number of chromosomal "events" (obtained by placing the genes in genomic order and counting the number of times the copy number signal switched between positive/zero/negative). This linear regression approach was applied twice. The first run was amplification-centric, so the copy number variable was adjusted such that negative values were set to zero (such that neutral and deleted regions are zero, and amplified regions are positive). The second run was deletion-centric, so the copy number variable was adjusted such that positive values were set to zero and the sign flipped (such that neutral and amplified regions are zero and deleted regions are positive). Thus, in both regressions, a positive copy number coefficient represented a positive association between CYT and lesion, and a negative copy number coefficient represented a negative association between copy number and lesion. The p-value of the coefficient was considered a measure of the strength of the evidence for association.

Because copy number alterations rarely affect a single gene, association signals were highly auto-correlated, meaning that genomic neighbors likely had a similar enrichment score. Because gene scores do not truly represent independent tests, standard multiple hypothesis correction procedures could not be employed at the per-gene level. Instead, an alternative approach based on permutation testing was used to assign adjusted p-values to each "peak." A "peak" was defined as a continuous stretch of genes (arranged in genomic order) with a nominal p-value less than 0.01, and the peak score was defined as the minimum p-value in the peak. To obtain the null distribution of peak scores, the CYT variable was randomly re-permuted and the entire process repeated (testing individual genes, defining peaks, and obtaining peak scores). This was repeated 500 times each for the amplification analysis and the

deletion analysis yielding a peak score null distribution. The quantile of each true peak score within the peak score null distribution was taken as a peak p-value. The set of peak p-values were then subjected to standard B-H correction.

For each amplification hit, the copy number of the peak gene was then tested for association with CYT in each individual cancer type (following the same approach taken in the point mutation analysis). Cancer-specific association was defined when the uncorrected p-value was less than 0.05.

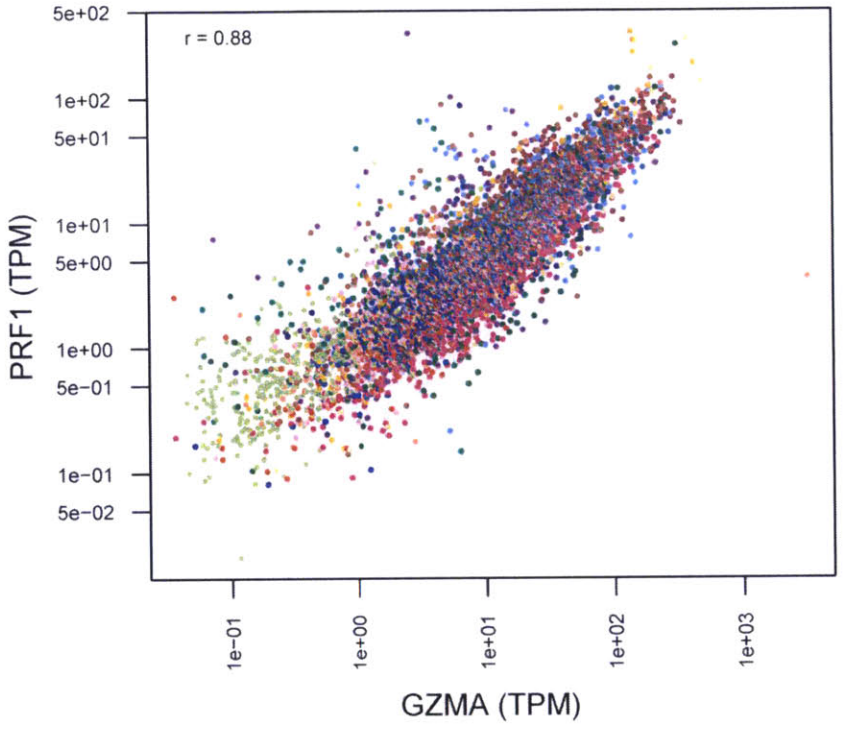
As in the point mutation analysis, the pipeline was repeated exploring for CNA associations with other cell type signatures.

Necrosis and ALOX amplifications

The amplification of ALOX15B was tested for association with necrosis in each tumor type by using a linear regression that modeled percent necrosis as a function of ALOX15B amplification and three additional background mutation rate variables added to avoid confounding (rank-transformed count of amplified genes, rank-transformed count of deleted genes, and rank-transformed count of events, as described previously).

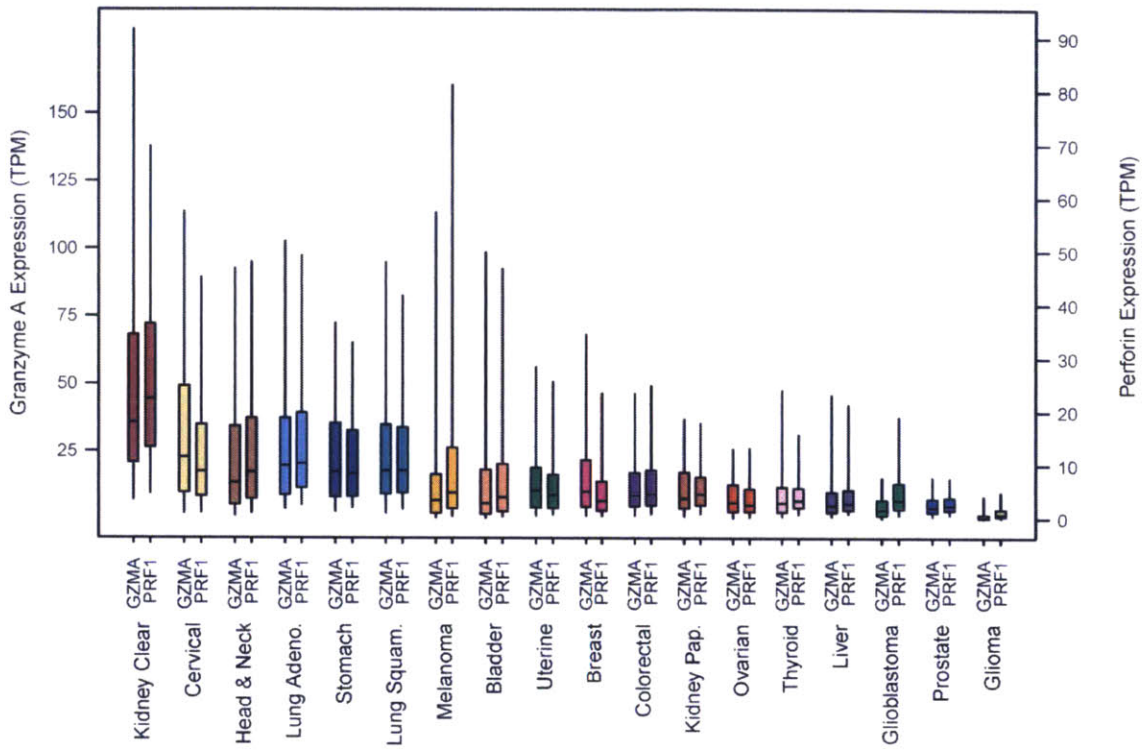
Supplemental Figures

Figure S1. Cytolytic activity and its expression correlates, related to Figure 1



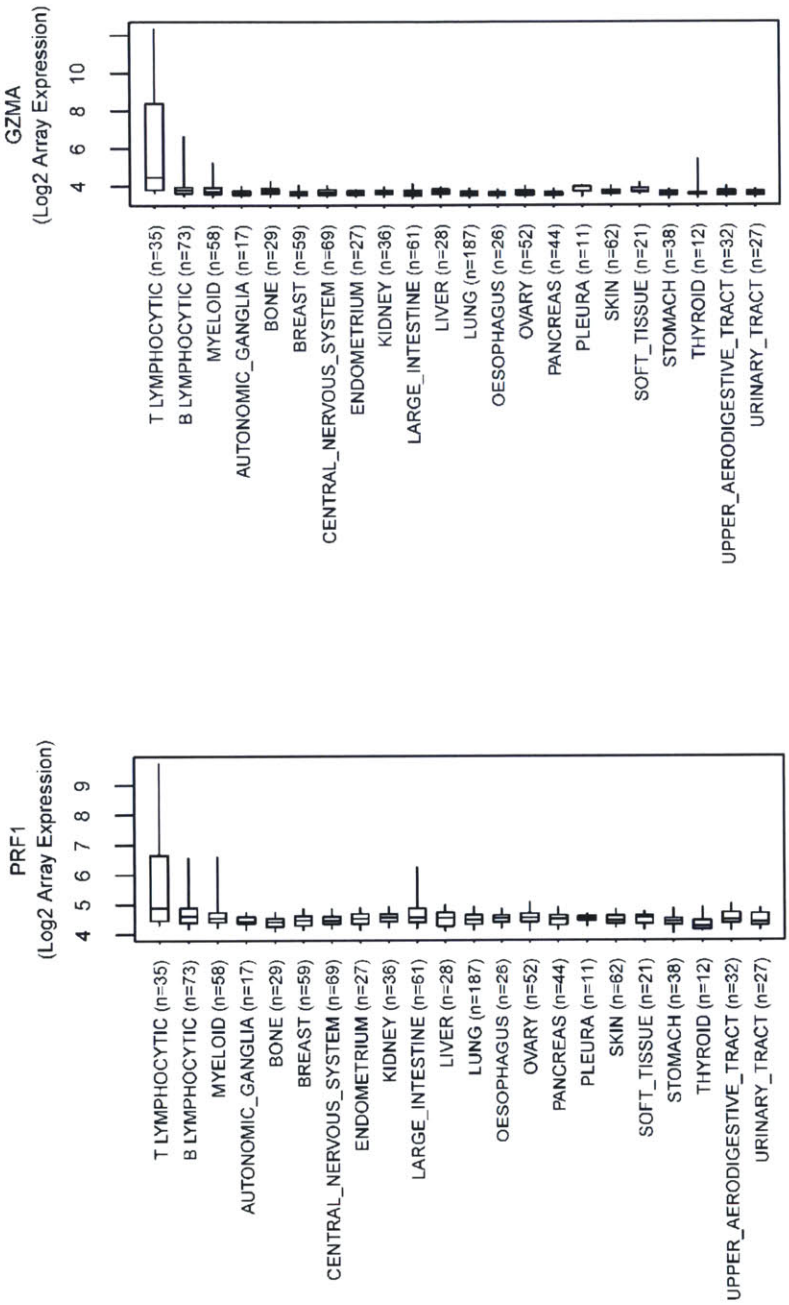
(a) GZMA *vs.* PRF1 expression across TCGA tumor biopsies. Points are colored according to cancer type using the same color-coding employed in **Figure 1**. Pan-cancer, a spearman rank correlation (r) of 0.88 was observed.

Figure S1, continued



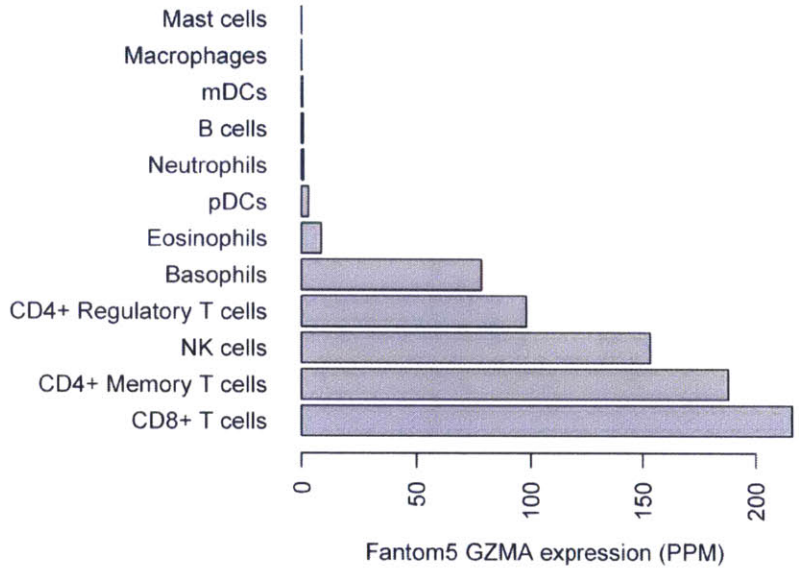
(b) GZMA and PRF1 expression across TCGA tumor biopsies. Solid bodies represent interquartile ranges and are notched by the median; vertical lines demarcate the 5th to 95th percentile range.

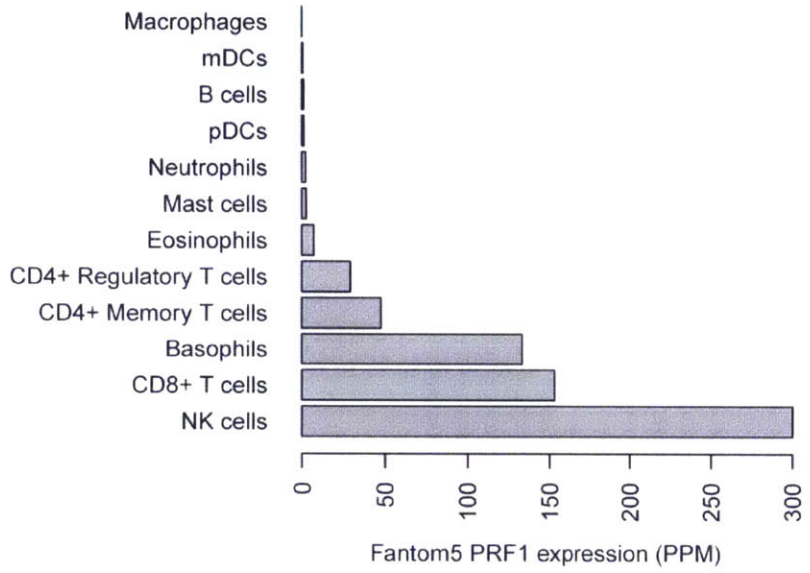
Figure S1, continued



(c) GZMA and PRF1 expression in Cancer Cell Line Encyclopedia (CCLE) cancer cell lines. Log2 (Affymetrix U133) array expression for ~1000 cancer cell lines grouped by cell lineage. Probes 205488_at and 1553681_a_at were used to represent GZMA and PRF1, respectively. Hematopoietic cell lines were further subdivided as T lymphocytic, B lymphocytic, or myeloid. Solid bodies represent interquartile ranges and are notched by the median; vertical lines demarcate the 5^{th} to 95^{th} percentile range.

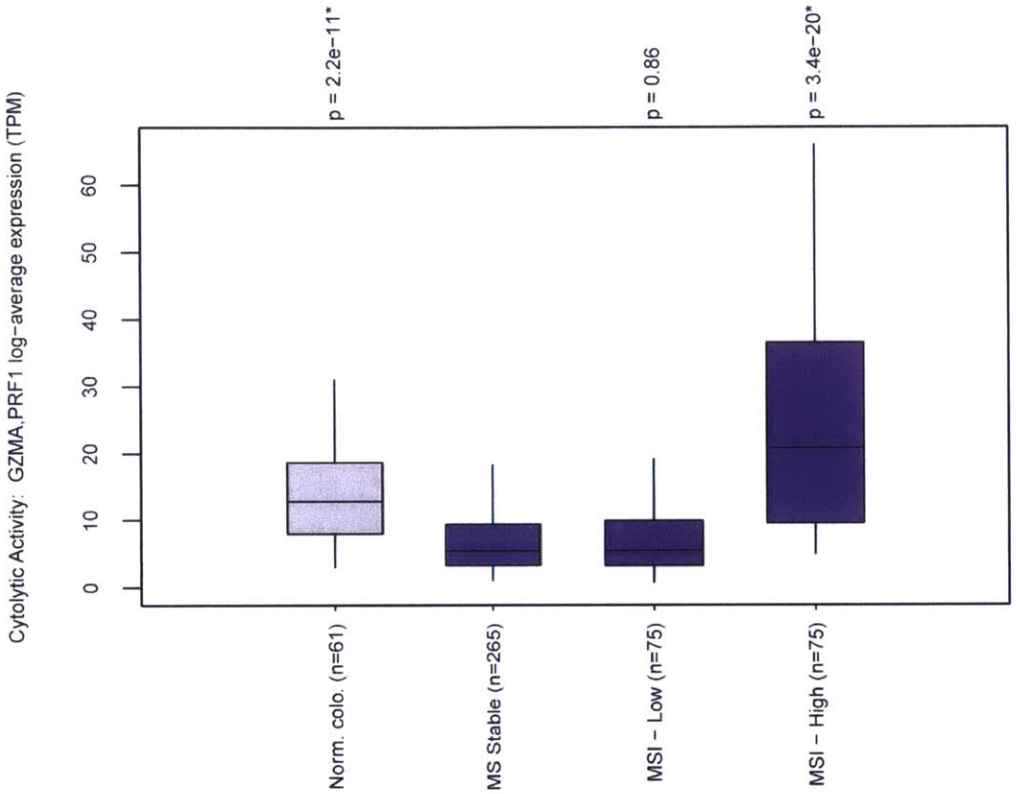
Figure S1, continued





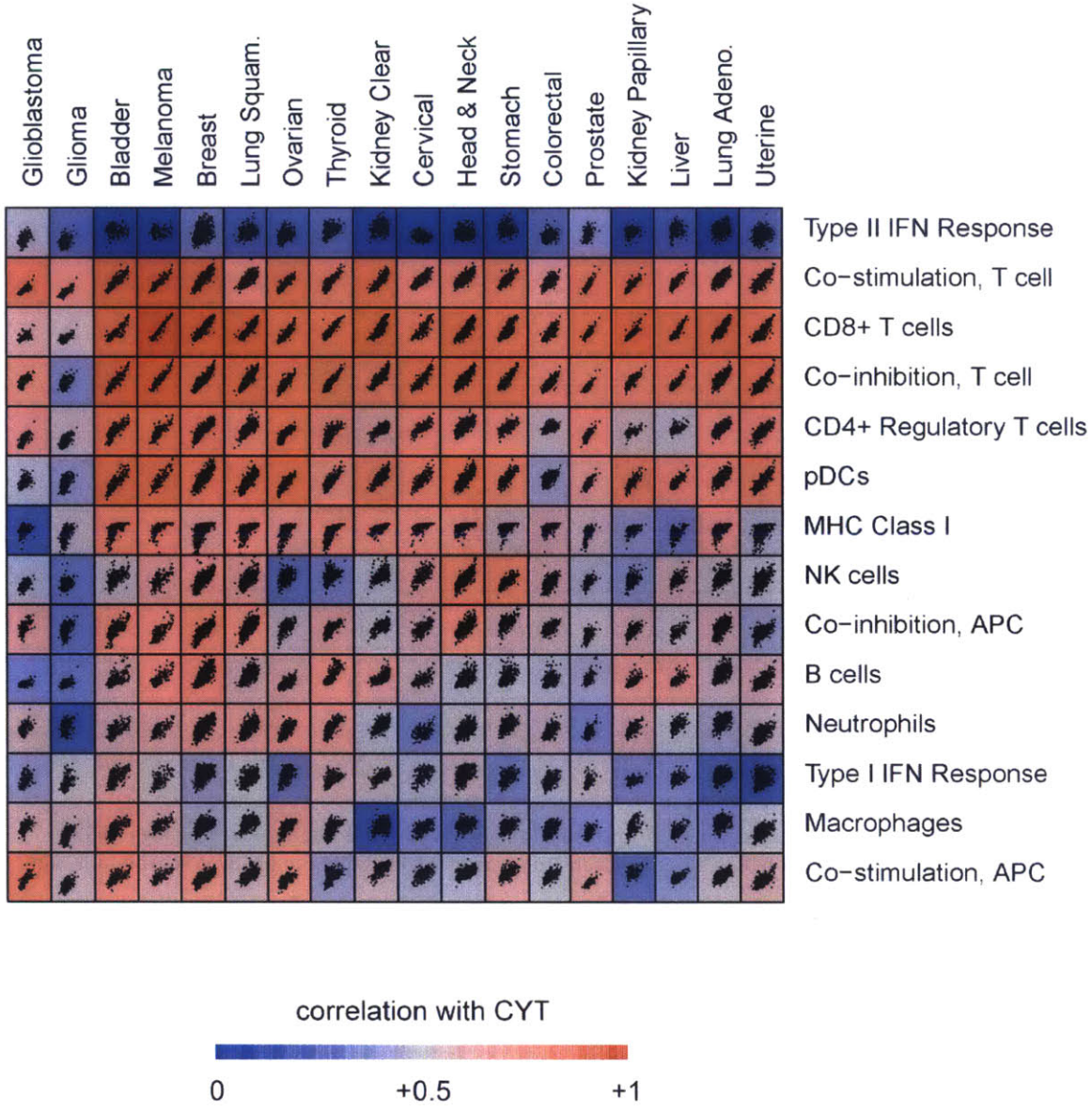
(d) Fantom5 CAGE expression (parts per million) of GZMA and PRF1 in 12 immune cell types.

Figure S1, continued



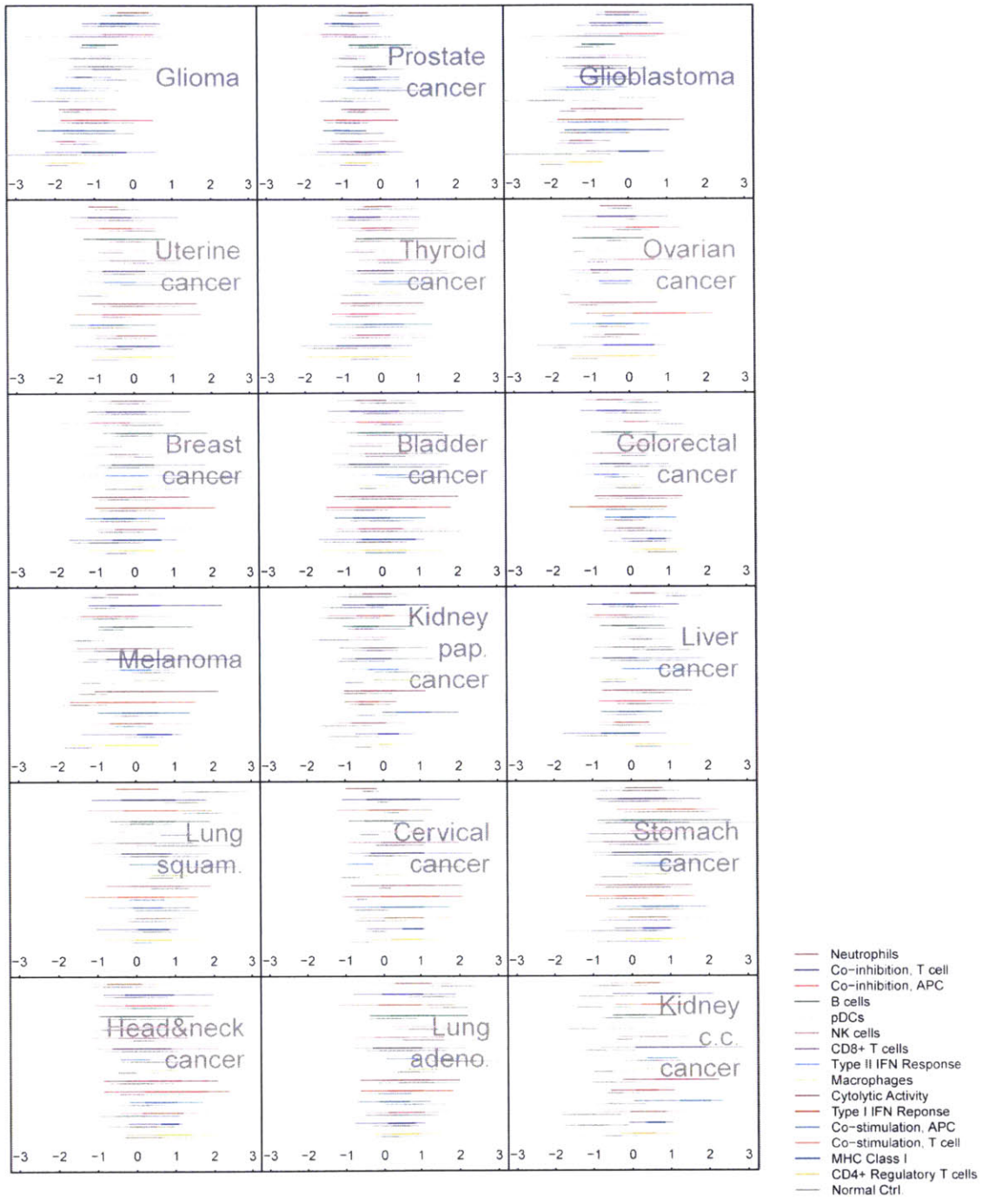
(e) CYT in normal colon and in colorectal cancer by microsatellite instability status (stable, low MSI, high MIS). Quantiles are represented as in part **b**. P-values correspond to comparison to stable tumors by Wilcoxon rank-sum test.

Figure S1, continued



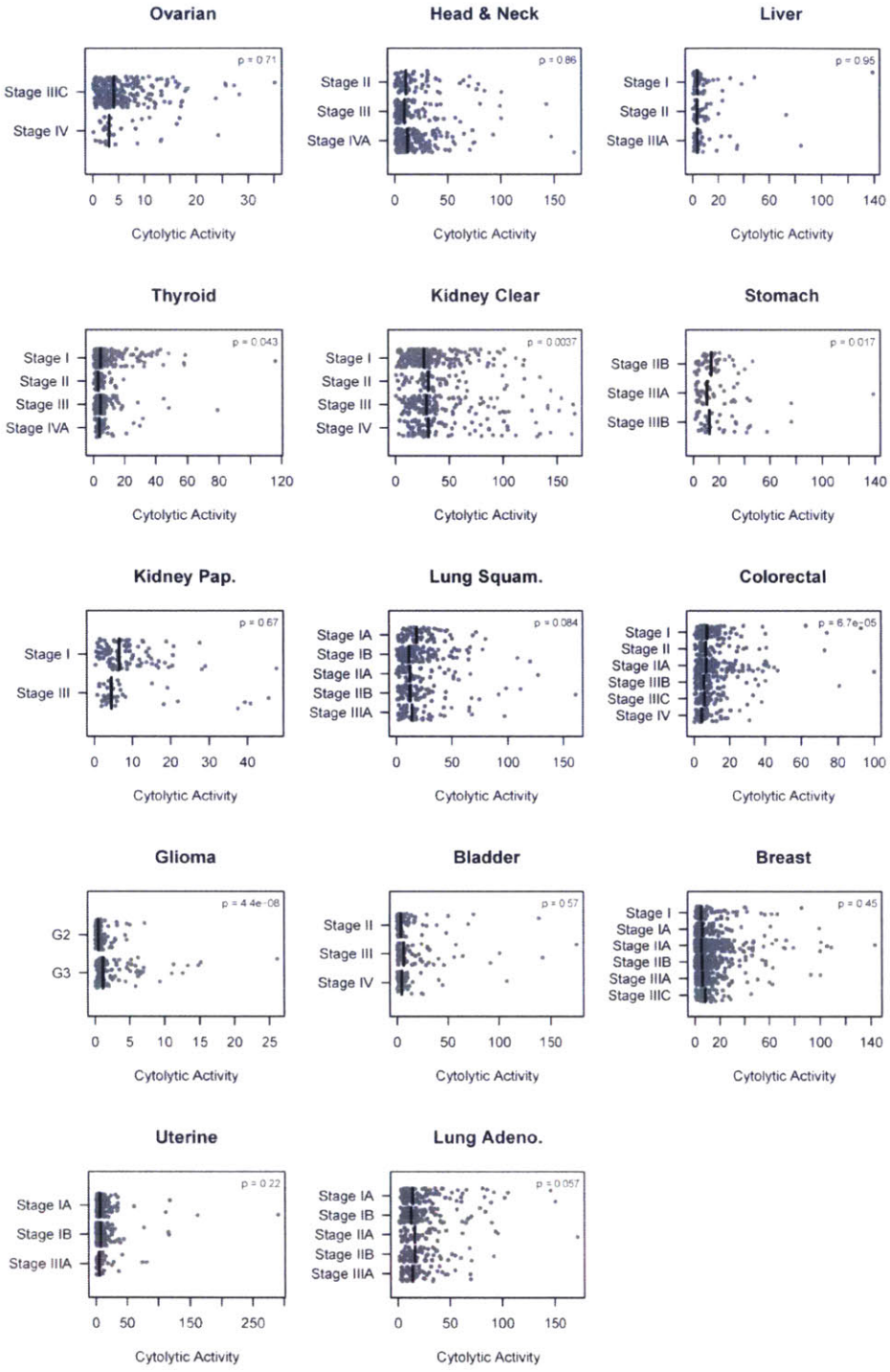
(f) Cell type marker enrichments *vs.* Cytolytic Activity (all calculated by ssGSEA). Each panel represents a scatter plot of z-scored enrichment scores with CYT on the x-axis and the relevant cell type on the y-axis. Background color of each scatter corresponds to the Spearman rank correlation, the color mapping indicated in the legend. We note that there are limitations to the precision of our markers genes; for example, we could not identify markers for NK cells that are not expressed (to some level) in activated CTLs.

Figure S1, continued



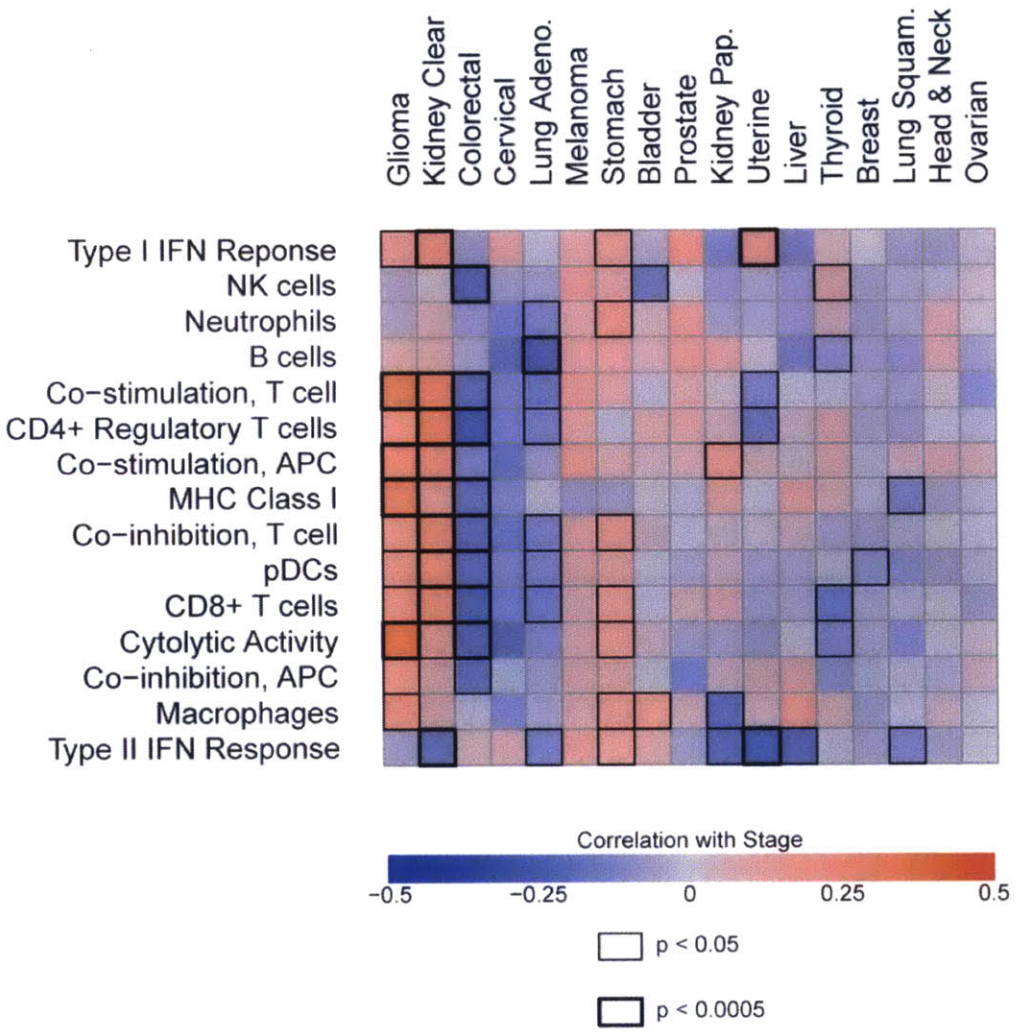
(g) Tumor-normal expression differences of z-scored cell type marker enrichments (all by ssGSEA, including CYT). Thin lines span the 5th to 95th percentile range and thick lines span the interquartile range. Colors correspond to cell type as indicated in the figure; gray bars represent the enrichment of the adjacent cell type in normal control tissues (from TCGA and GTEx).

Figure S1, continued



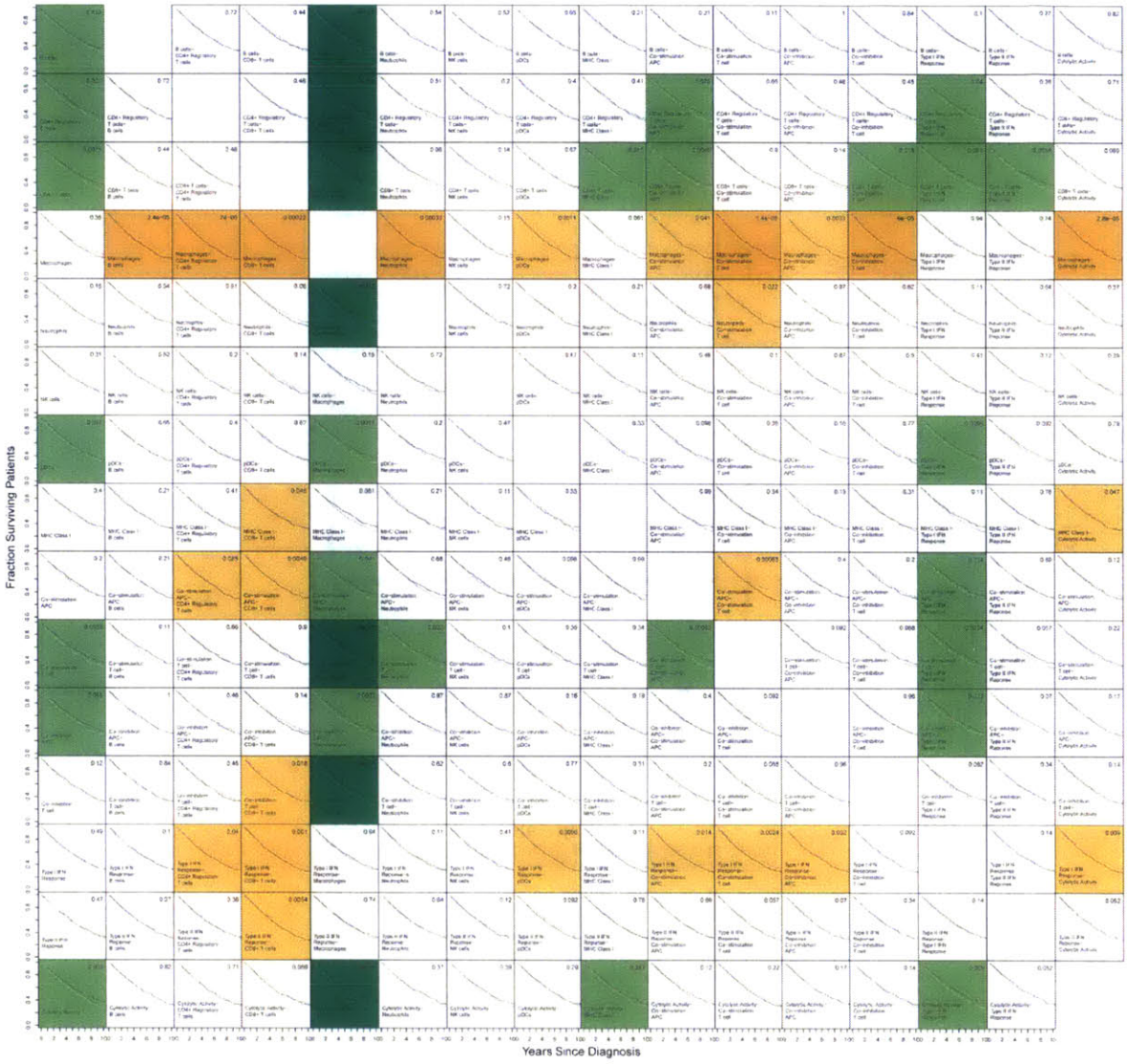
(h) CYT (geometric mean) by tumor stage. Stages are shown with at least 30 corresponding samples. Each gray dot represents a sample, and black lines mark the medians. P-values (upper right of each plot) correspond to Pearson correlation between log CYT and rank stage (*i.e.*. "stage 1A"=1, "stage 1B"=2, *etc.*).

Figure S1, continued



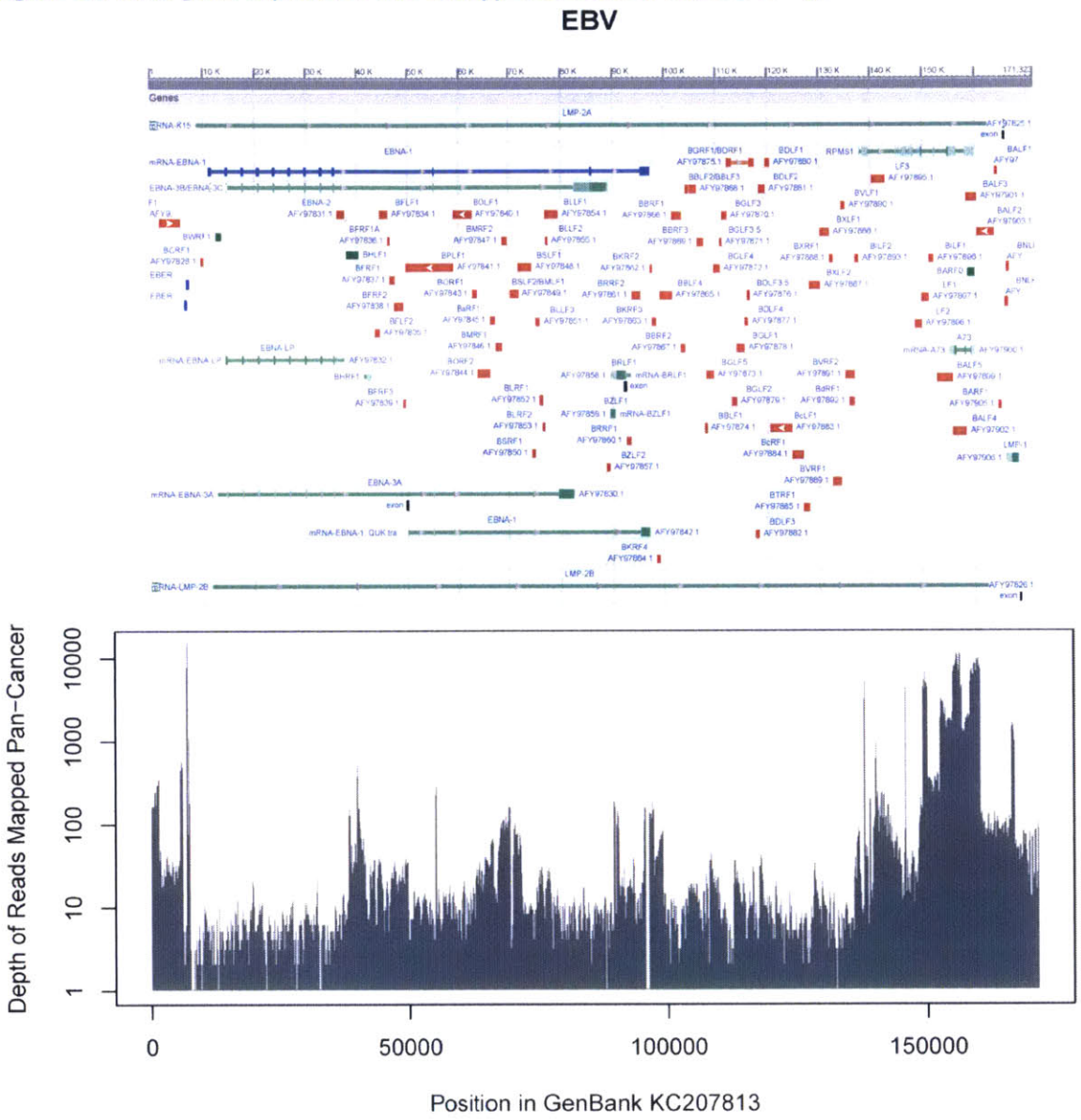
(i) Heatmap indicating the association between rank stage and z-scored marker gene enrichment in each tumor type. Colors represent the magnitude and direction of the correlation as indicated in the legend. Cell borders indicate significance levels (thin black lines, $p<0.05$; thick black lines, $p<0.0005$).

Figure S1, continued



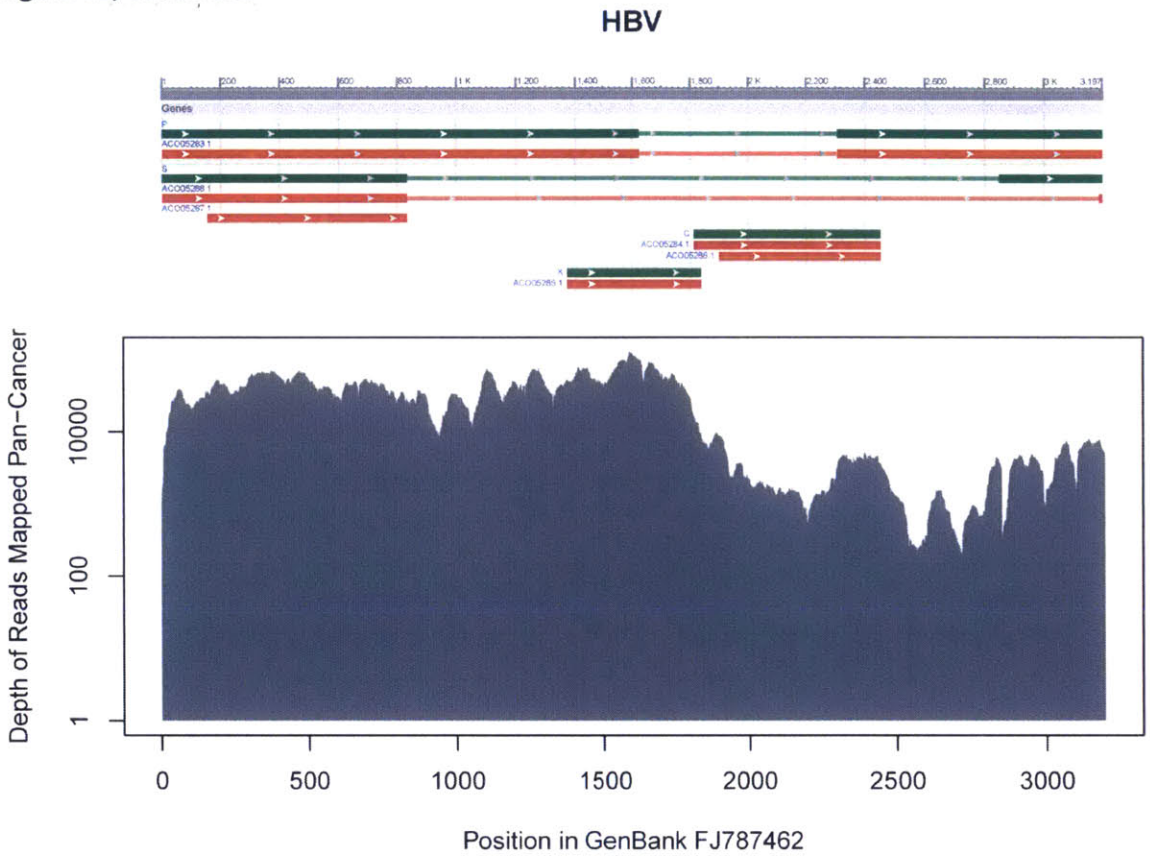
(j) Survival curves based on cytolytic activity and other cell type markers. In each survival analysis, patients were segregated into "high" (black line) and "low" (gray line) cohorts, each with an identical admixture of tumor histological type and stage (Methods). In the leftmost column, "high" and "low" were based on metagene expression. In other panels, "high" and "low" were based on expression ratios, as indicated. P-values were assigned based on Cox proportional hazards models. Panels are highlighted in green when the "high" group had a advantage and in orange when the "low" group had a survival advantage (using a nominal significance cutoff of $p<0.05$). Darker orange and green correspond to stronger unadjusted p-values ($p<0.0005$).

Figure S2. Viral gene expression and cell type correlates, related to Figure 2



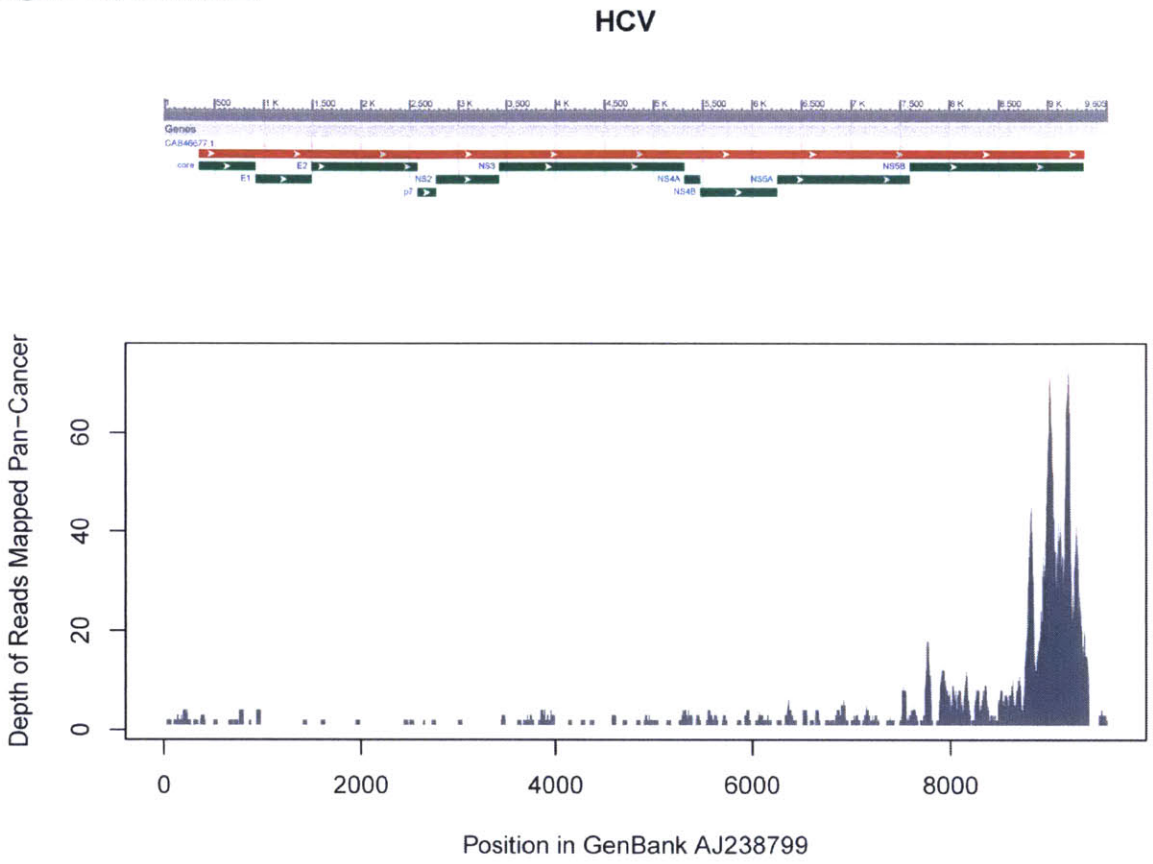
(a) Part 1 of 7. Read depths are presented on log-scale for viruses with depths exceeding 100 and on linear scale otherwise. GenBank annotations of known viral elements are presented above.

Figure S2, continued



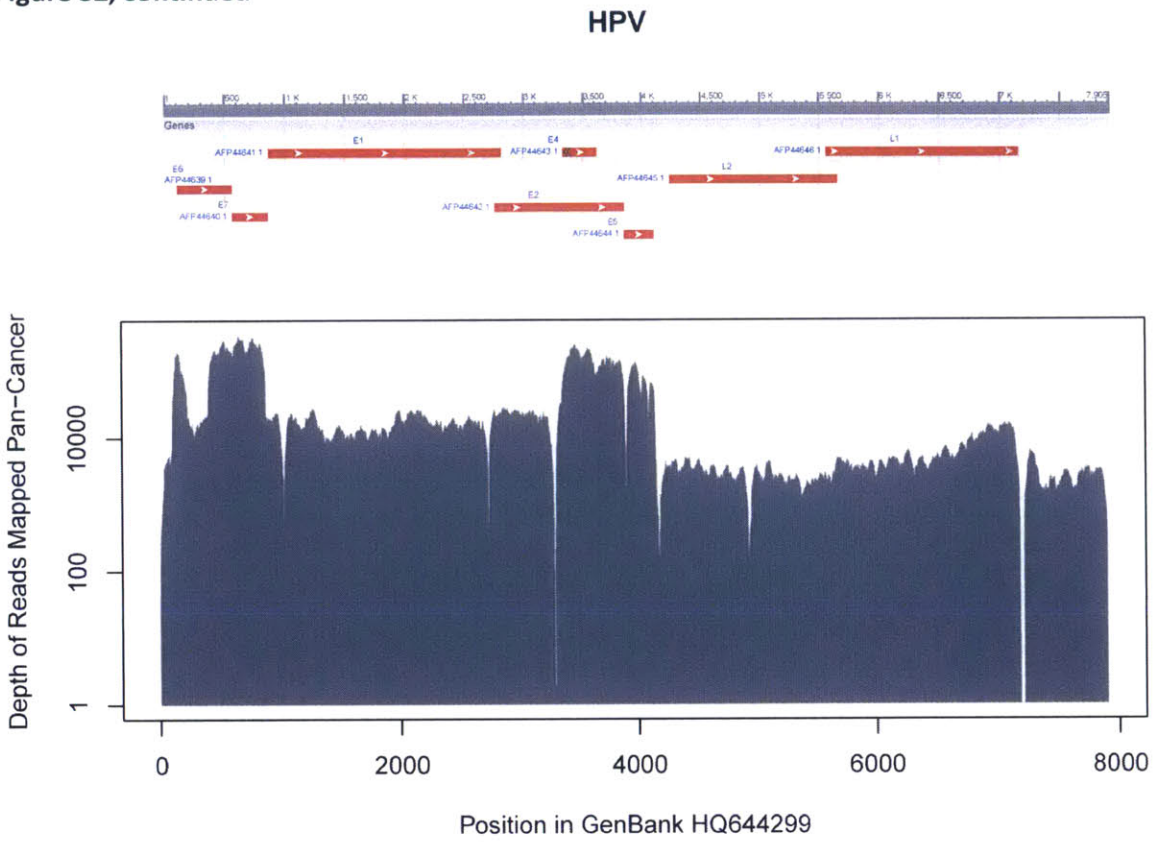
(a) Part 2 of 7. Read depths are presented on log-scale for viruses with depths exceeding 100 and on linear scale otherwise. GenBank annotations of known viral elements are presented above.

Figure S2, continued



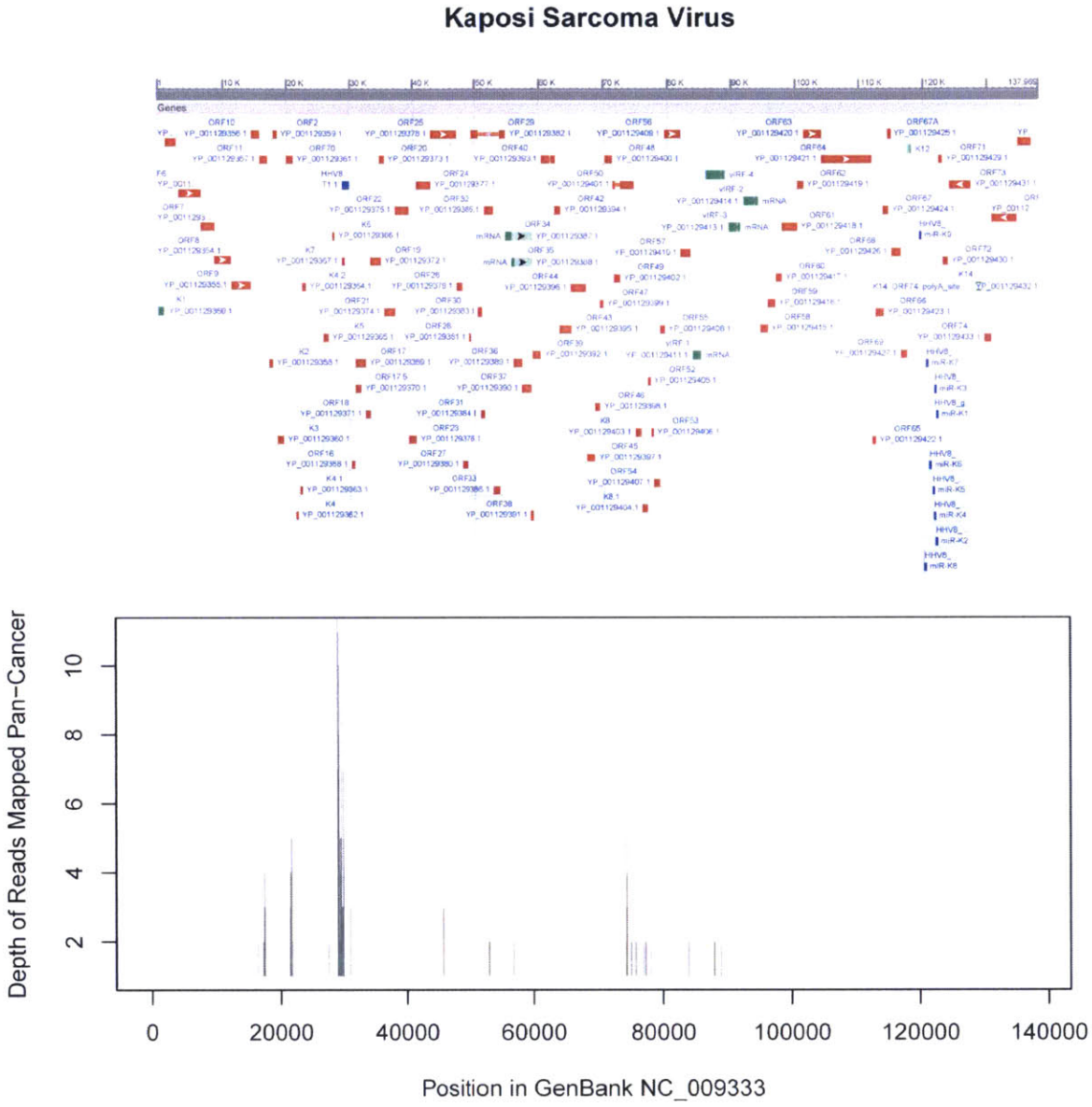
(a) Part 3 of 7. Read depths are presented on log-scale for viruses with depths exceeding 100 and on linear scale otherwise. GenBank annotations of known viral elements are presented above.

Figure S2, continued



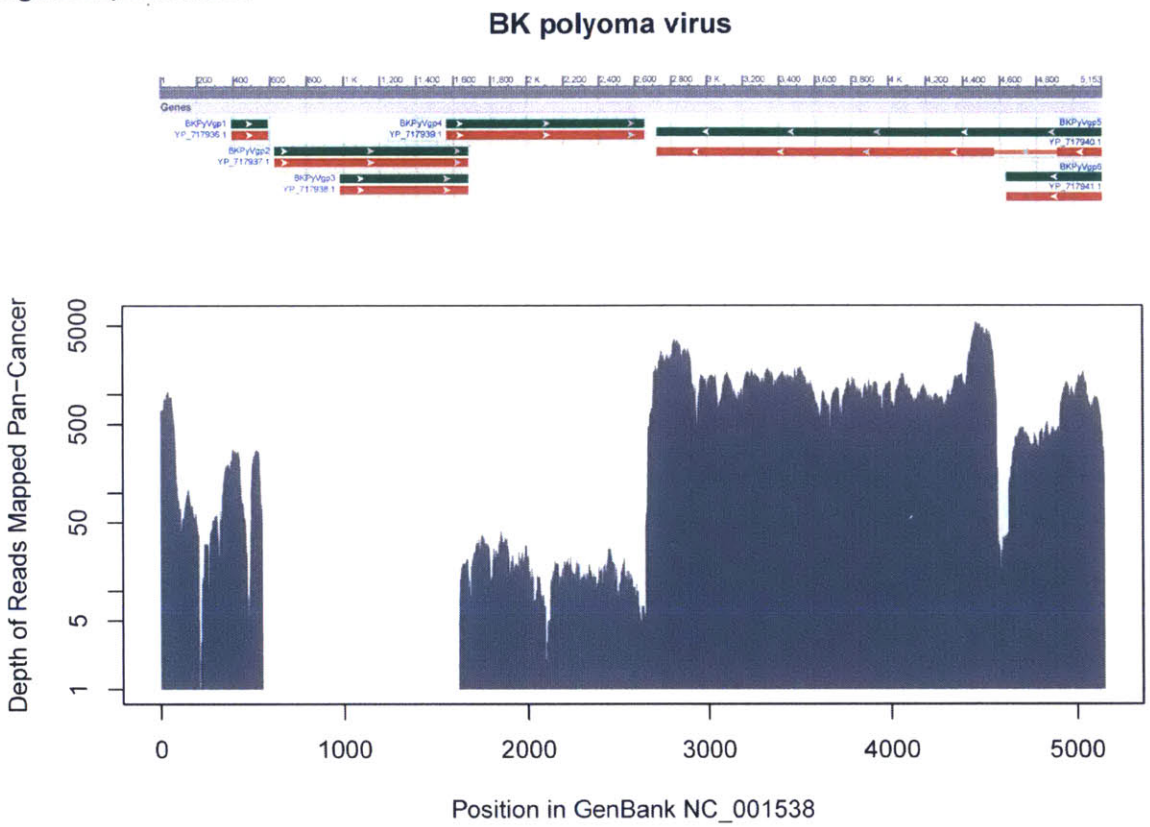
(a) Part 4 of 7. Read depths are presented on log-scale for viruses with depths exceeding 100 and on linear scale otherwise. GenBank annotations of known viral elements are presented above.

Figure S2, continued



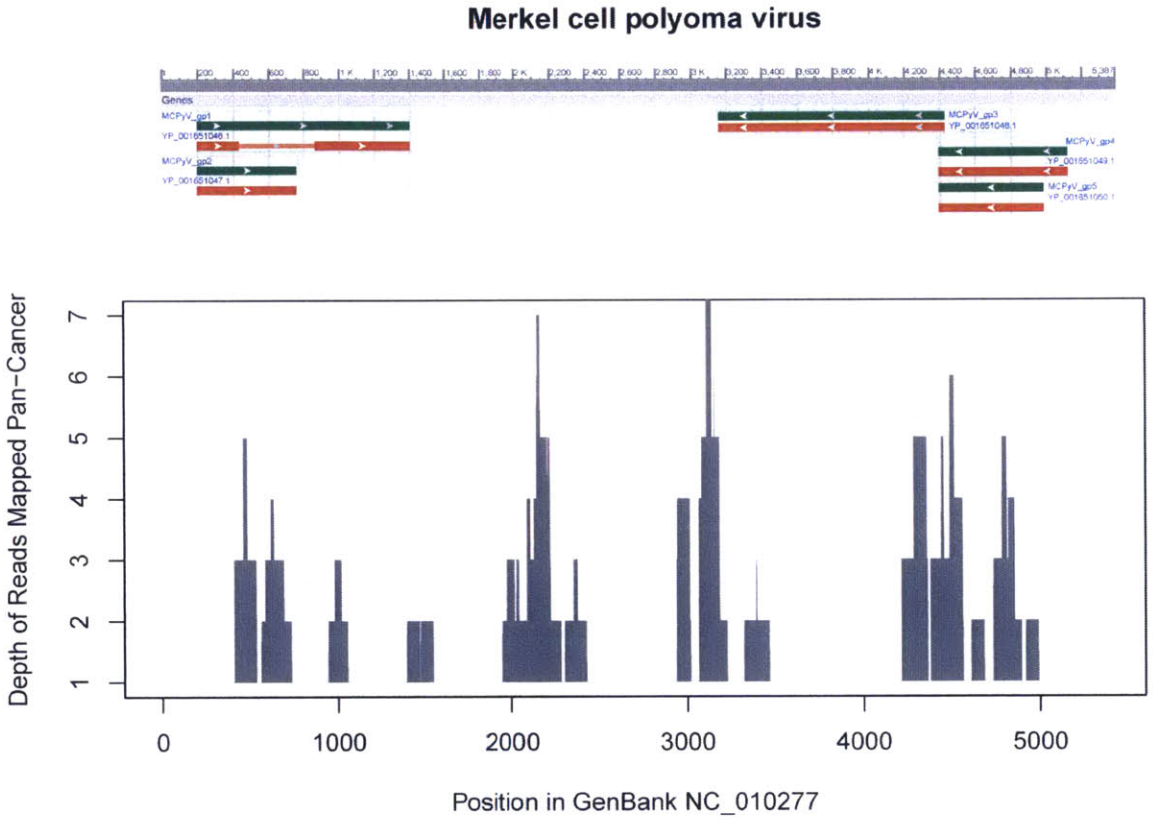
(a) Part 5 of 7. Read depths are presented on log-scale for viruses with depths exceeding 100 and on linear scale otherwise. GenBank annotations of known viral elements are presented above.

Figure S2, continued



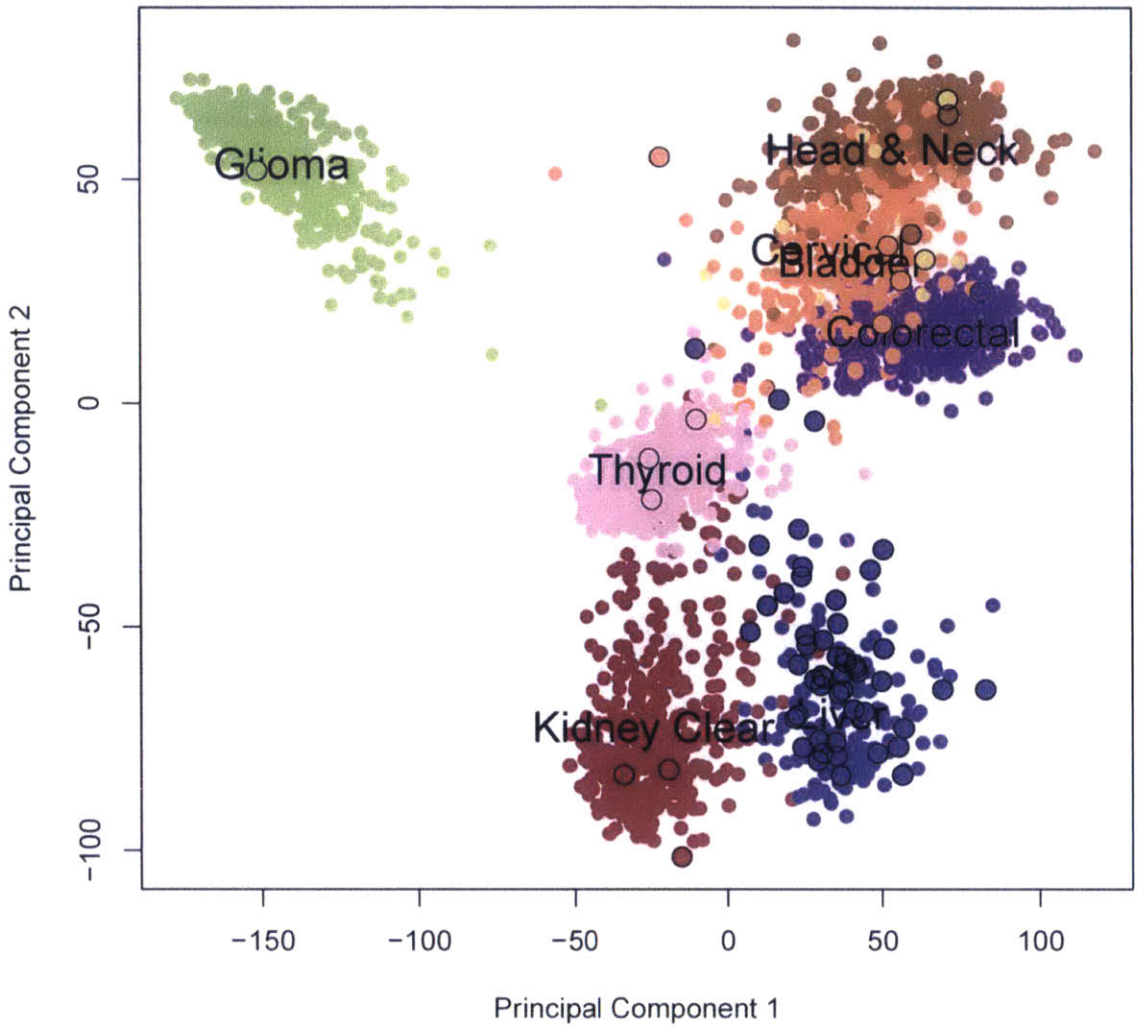
(a) Part 6 of 7. Read depths are presented on log-scale for viruses with depths exceeding 100 and on linear scale otherwise. GenBank annotations of known viral elements are presented above.

Figure S2, continued



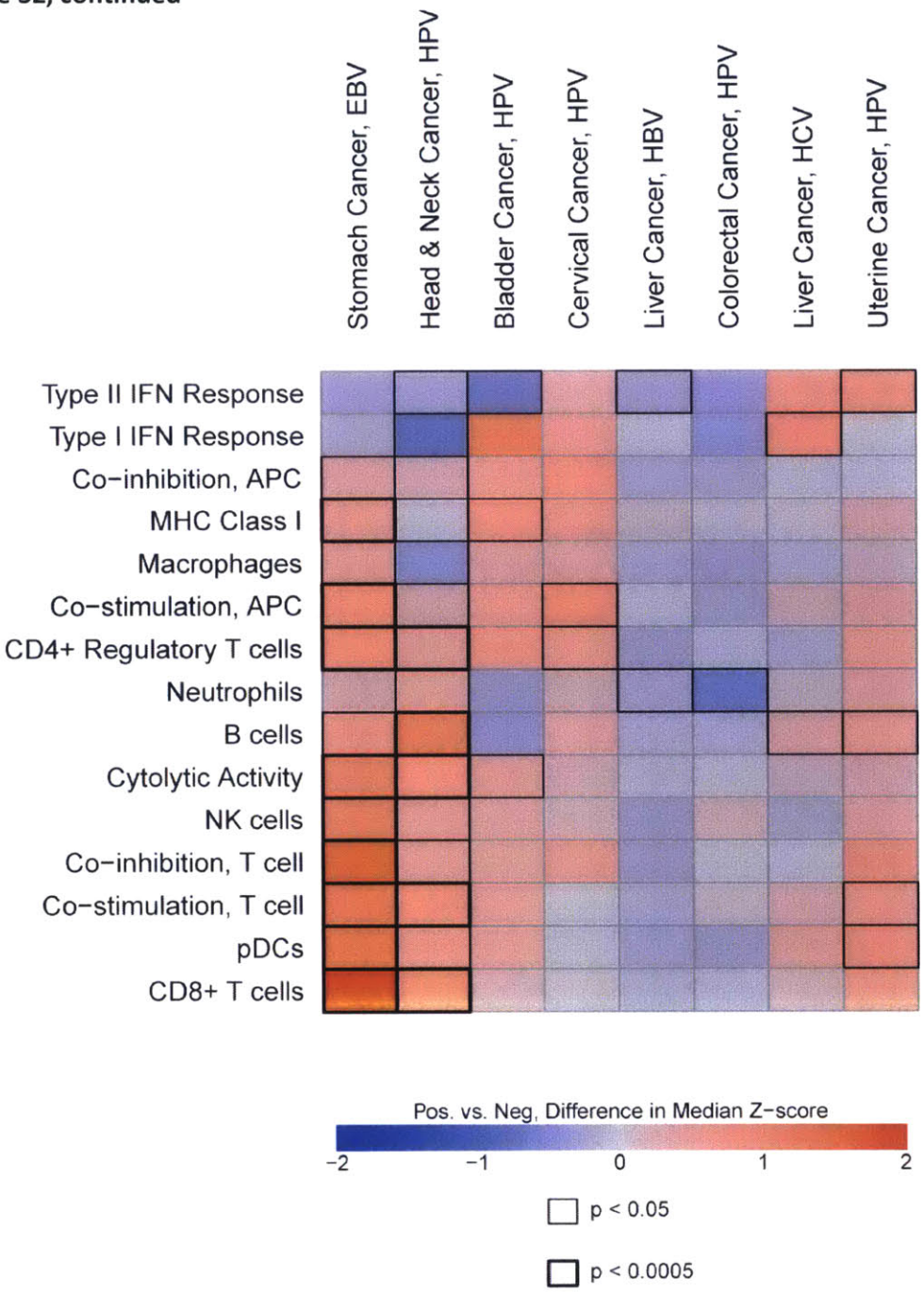
(a) Part 7 of 7. Read depths are presented on log-scale for viruses with depths exceeding 100 and on linear scale otherwise. GenBank annotations of known viral elements are presented above.

Figure S2, continued



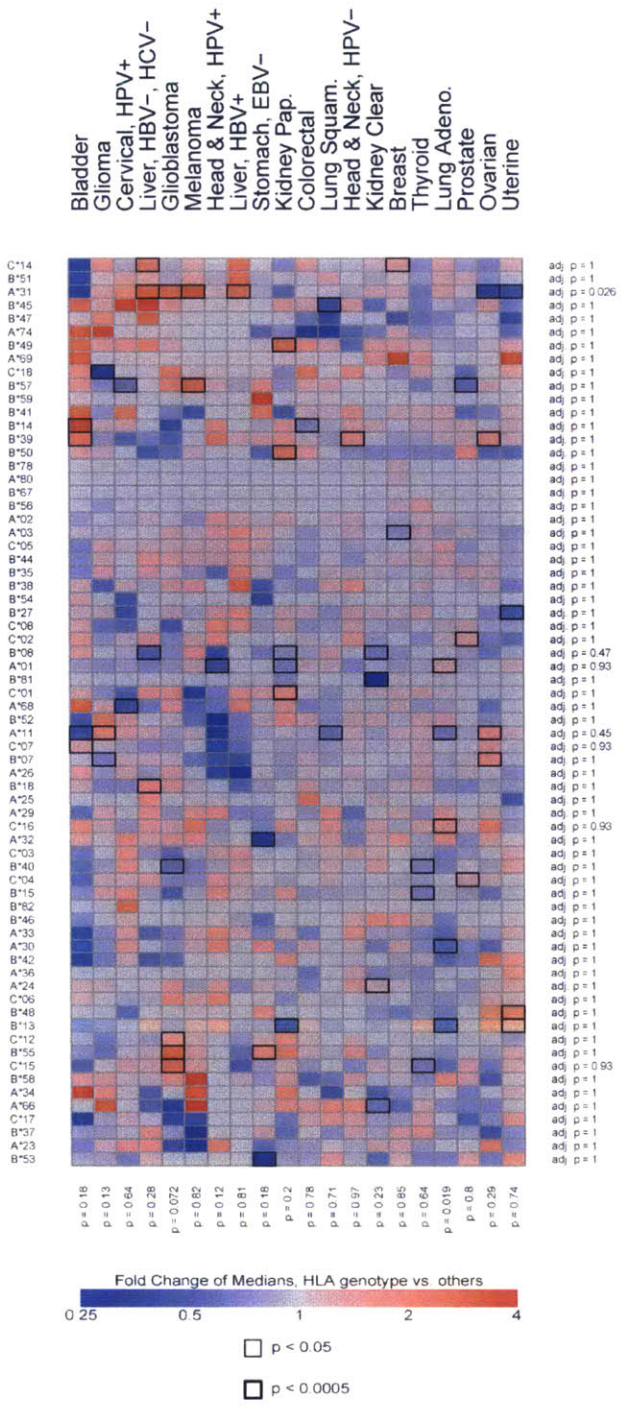
(b) Tumor samples plotted according to the first two principal components log-transformed gene expression (for tumor types with ≥1 HBV+ case). Color coding corresponds to that used in **Figure 1**. HBV-infected samples are represented by larger, black-outlined points.

Figure S2, continued



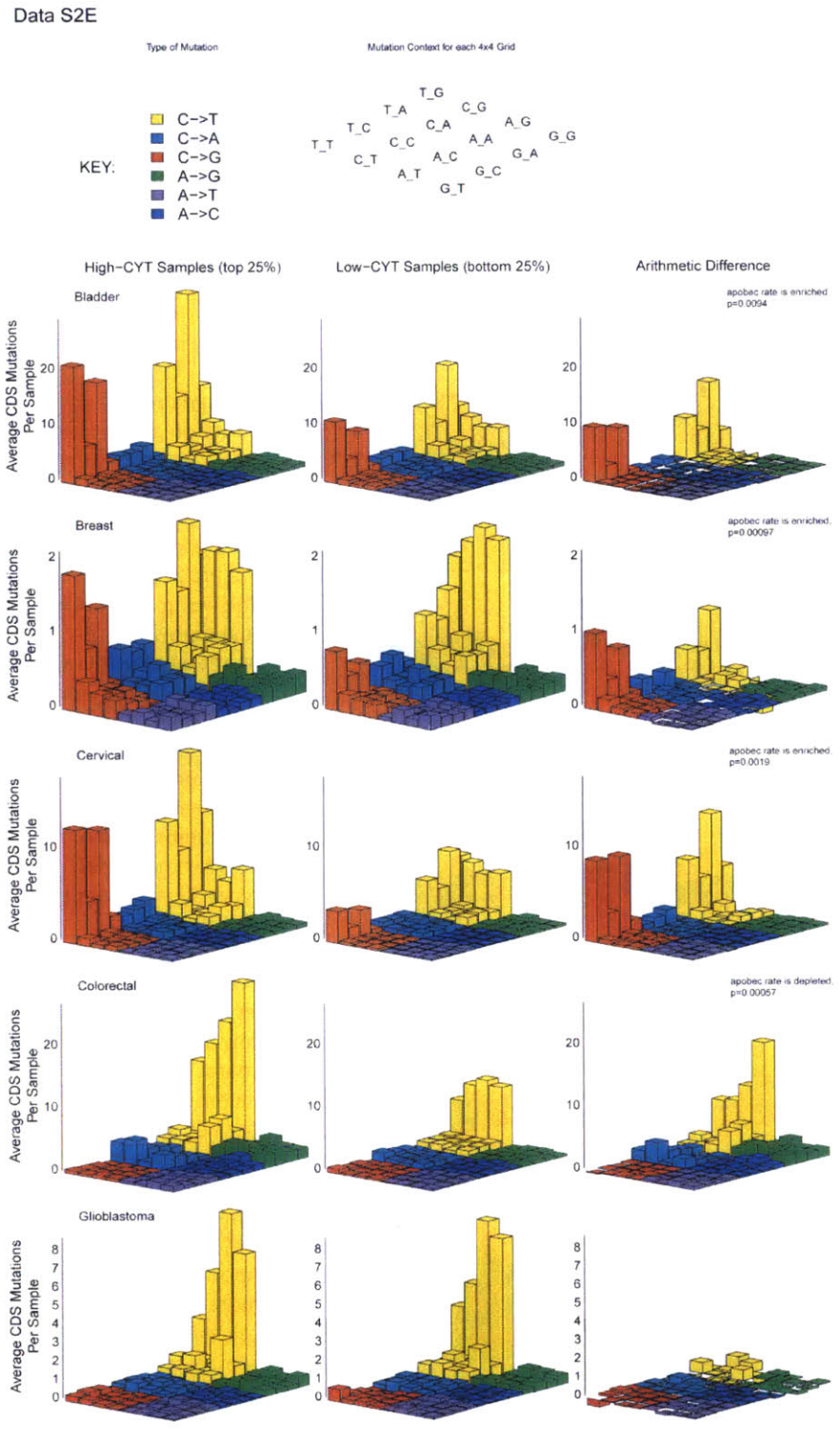
(c) Heatmap showing association between viral infection status and the enrichment of cell type markers. Colors correspond to the difference in z-scored enrichment between infected and non-infected samples. Cell borders indicate the unadjusted significance of the association according to Wilcoxon rank-sum test.

Figure S2, continued



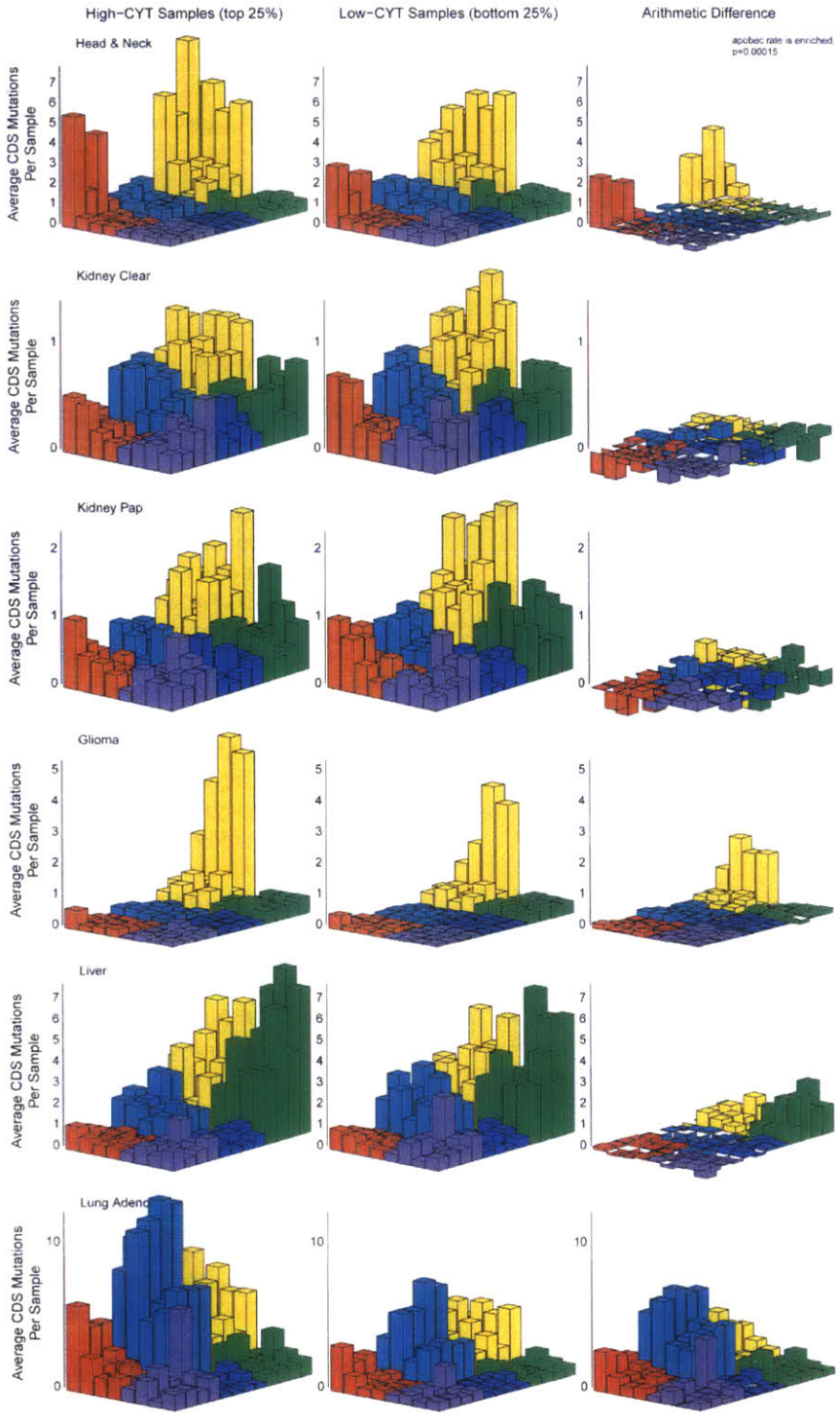
(d) Heatmap showing associations between HLA type and CYT. Colors correspond to the fold change between median expression in infected and non-infected samples. Cell borders indicate the significance of the association according to Wilcoxon rank sum test. Marginal HLA type significances are based on combination of each row's p-values by Fisher's method and are adjusted by BH method. Marginal tumor type significances are based on rank-CYT ANOVA and are presented without multiple comparisons correction.

Figure S2, continued



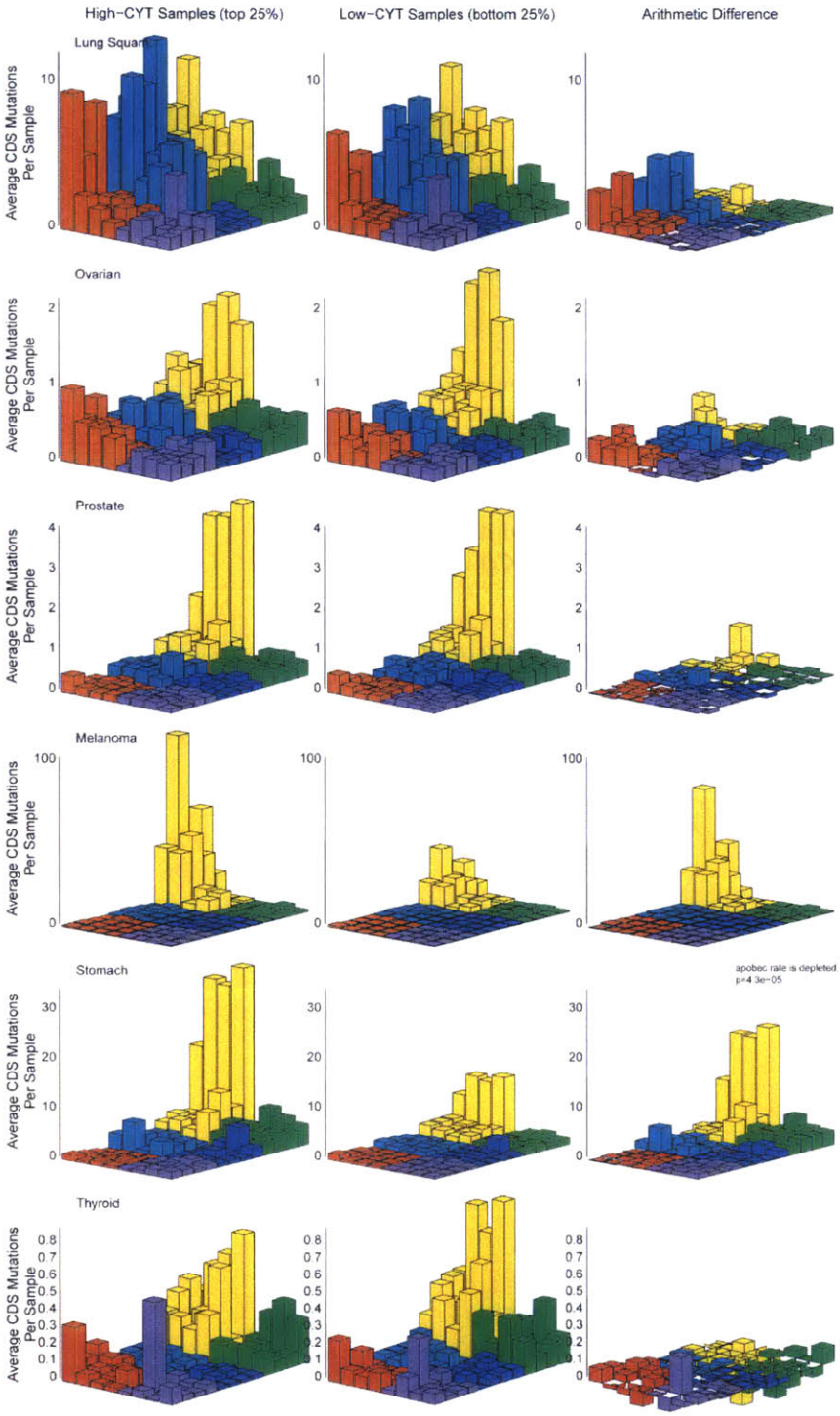
(e) Part 1 of 4. Single-nucleotide variant spectra for high- vs. low-CYT tumors. Mutational spectra are defined by the base change and the sequence context one base upstream and one base downstream. T→X and G→X mutations are considered from the perspective of the opposite strand such that all mutations are A→X or C→X. The average rate of each mutation per sample (counting mutations in coding sequence only) is represented in an 8x12 grid according to the provided legends. The first plot in each row represents mutation rate averages for high-CYT tumors (top 25% for that tumor type). The middle plot represents mutation rate averages for low-CYT tumors (bottom 25% for that tumor type). The third plot represents the arithmetic difference. In each plot, the back left row of bars corresponds to Apobec-characteristic tCx→tXx mutations. To assess Apobec enrichment for a tumor type, the Spearman rank correlation between CYT and the Apobec/non-Apobec mutation ratio was calculated across all samples.

Figure S2, continued



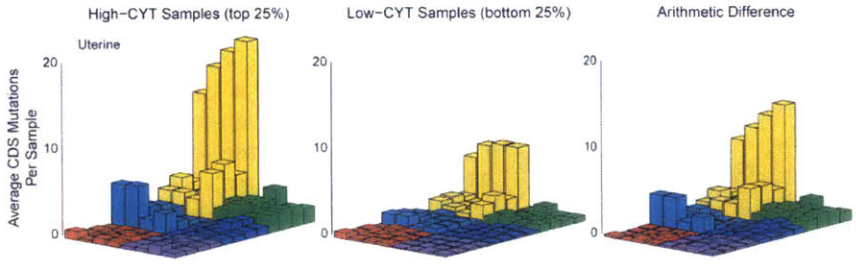
(e) Part 2 of 4. Single-nucleotide variant spectra for high- vs. low-CYT tumors. Mutational spectra are defined by the base change and the sequence context one base upstream and one base downstream. T→X and G→X mutations are considered from the perspective of the opposite strand such that all mutations are A→X or C→X. The average rate of each mutation per sample (counting mutations in coding sequence only) is represented in an 8x12 grid according to the provided legends. The first plot in each row represents mutation rate averages for high-CYT tumors (top 25% for that tumor type). The middle plot represents mutation rate averages for low-CYT tumors (bottom 25% for that tumor type). The third plot represents the arithmetic difference. In each plot, the back left row of bars corresponds to Apobec-characteristic tCx→tXx mutations. To assess Apobec enrichment for a tumor type, the Spearman rank correlation between CYT and the Apobec/non-Apobec mutation ratio was calculated across all samples.

Figure S2, continued



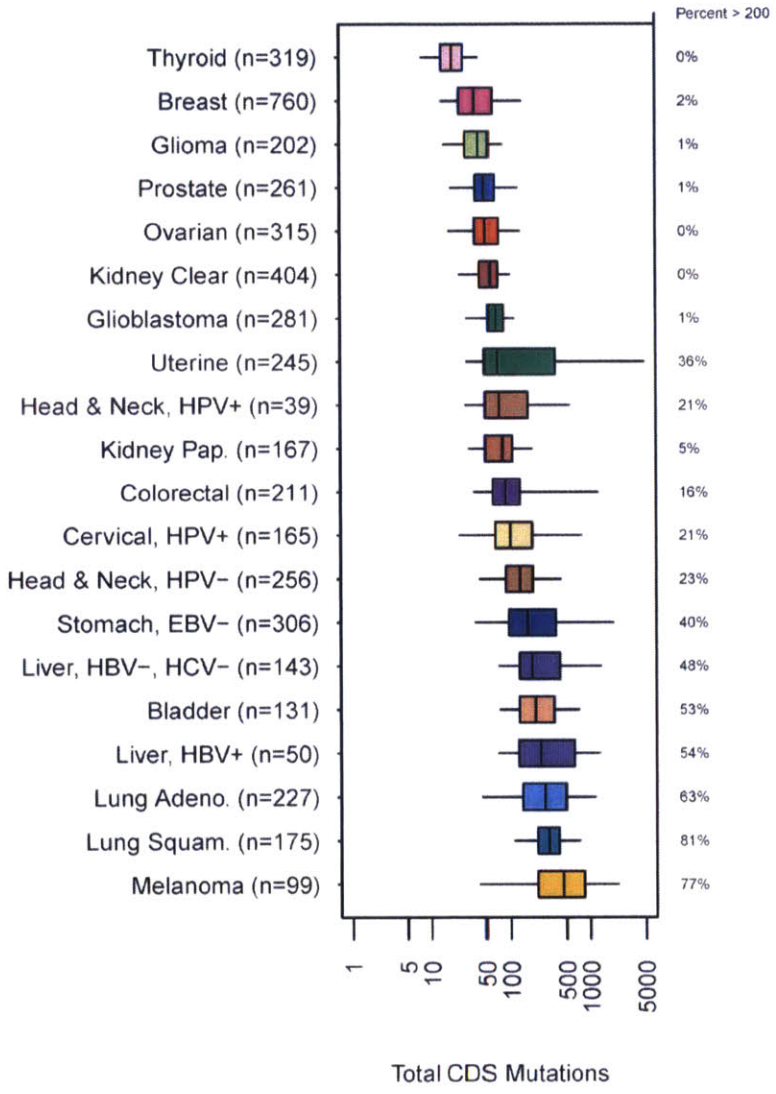
(e) Part 3 of 4. Single-nucleotide variant spectra for high- vs. low-CYT tumors. Mutational spectra are defined by the base change and the sequence context one base upstream and one base downstream. T→X and G→X mutations are considered from the perspective of the opposite strand such that all mutations are A→X or C→X. The average rate of each mutation per sample (counting mutations in coding sequence only) is represented in an 8x12 grid according to the provided legends. The first plot in each row represents mutation rate averages for high-CYT tumors (top 25% for that tumor type). The middle plot represents mutation rate averages for low-CYT tumors (bottom 25% for that tumor type). The third plot represents the arithmetic difference. In each plot, the back left row of bars corresponds to Apobec-characteristic tCx→tXx mutations. To assess Apobec enrichment for a tumor type, the Spearman rank correlation between CYT and the Apobec/non-Apobec mutation ratio was calculated across all samples.

Figure S2, continued



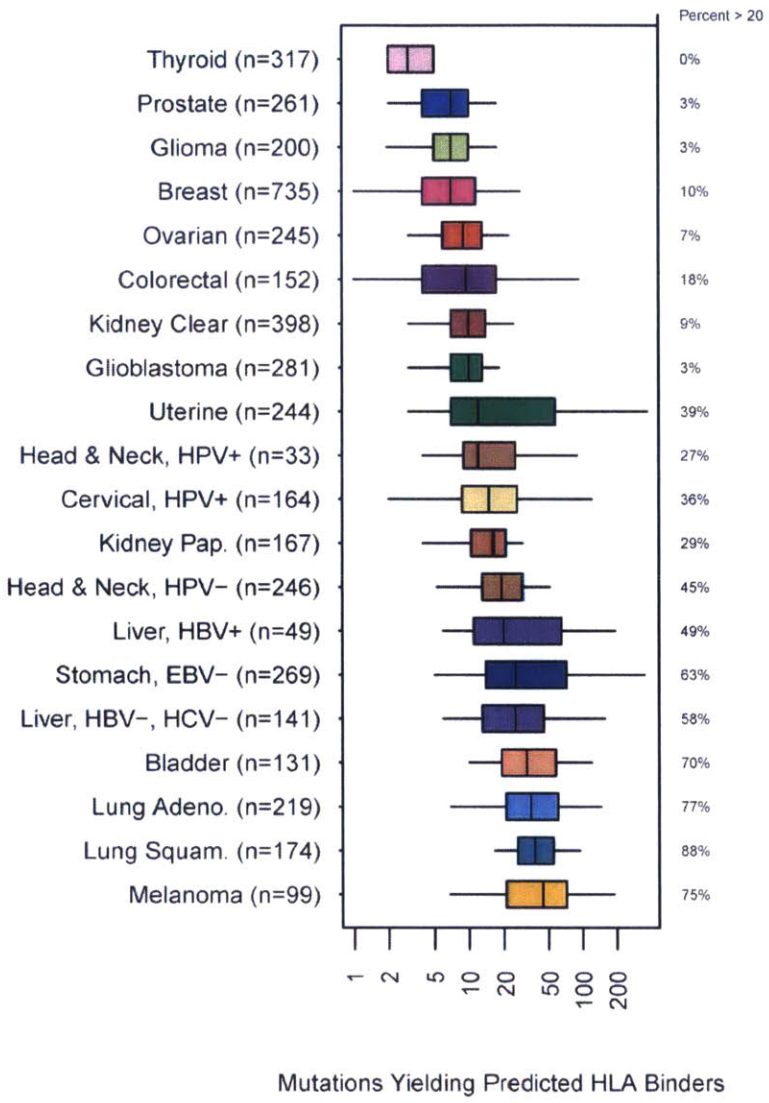
(e) Part 4 of 4. Single-nucleotide variant spectra for high- vs. low-CYT tumors. Mutational spectra are defined by the base change and the sequence context one base upstream and one base downstream. T→X and G→X mutations are considered from the perspective of the opposite strand such that all mutations are A→X or C→X. The average rate of each mutation per sample (counting mutations in coding sequence only) is represented in an 8x12 grid according to the provided legends. The first plot in each row represents mutation rate averages for high-CYT tumors (top 25% for that tumor type). The middle plot represents mutation rate averages for low-CYT tumors (bottom 25% for that tumor type). The third plot represents the arithmetic difference. In each plot, the back left row of bars corresponds to Apobec-characteristic tCx→tXx mutations. To assess Apobec enrichment for a tumor type, the Spearman rank correlation between CYT and the Apobec/non-Apobec mutation ratio was calculated across all samples.

Figure S3. Mutations, Neo-epitopes and their correlates, related to Figure 3



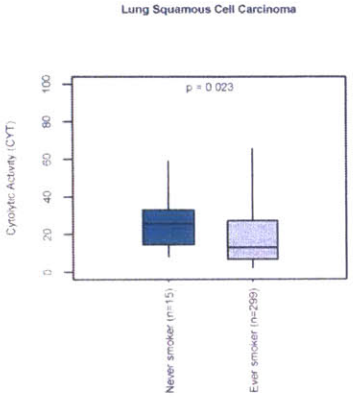
(a) Boxplots indicate typical rates of non-silent mutation (coding sequence events only) in each tumor type. Solid bodies represent interquartile ranges and are notched by the median; lines demarcate the 5th to 95th percentile range. The right axis indicates the fraction of samples with >200.

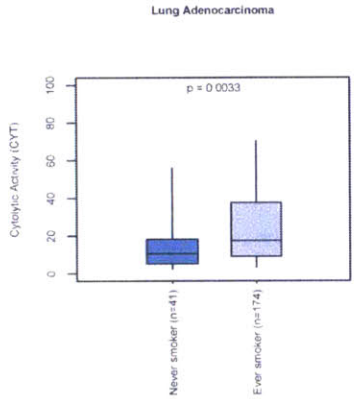
Figure S3, continued

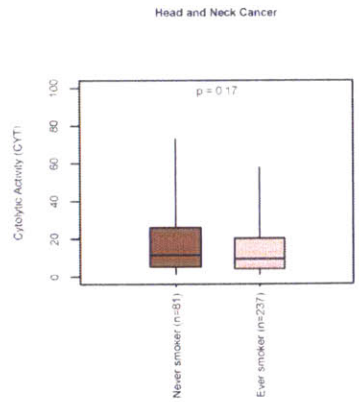


(b) Analogous to (a) but presenting the rate of mutations expected to yield an epitope with strong predicted binding to patient-matched HLA and moderate-to-high expression (median expression ≥10 TPM within the given tumor type). The right axis indicates the fraction of samples with >20.

Figure S3, continued

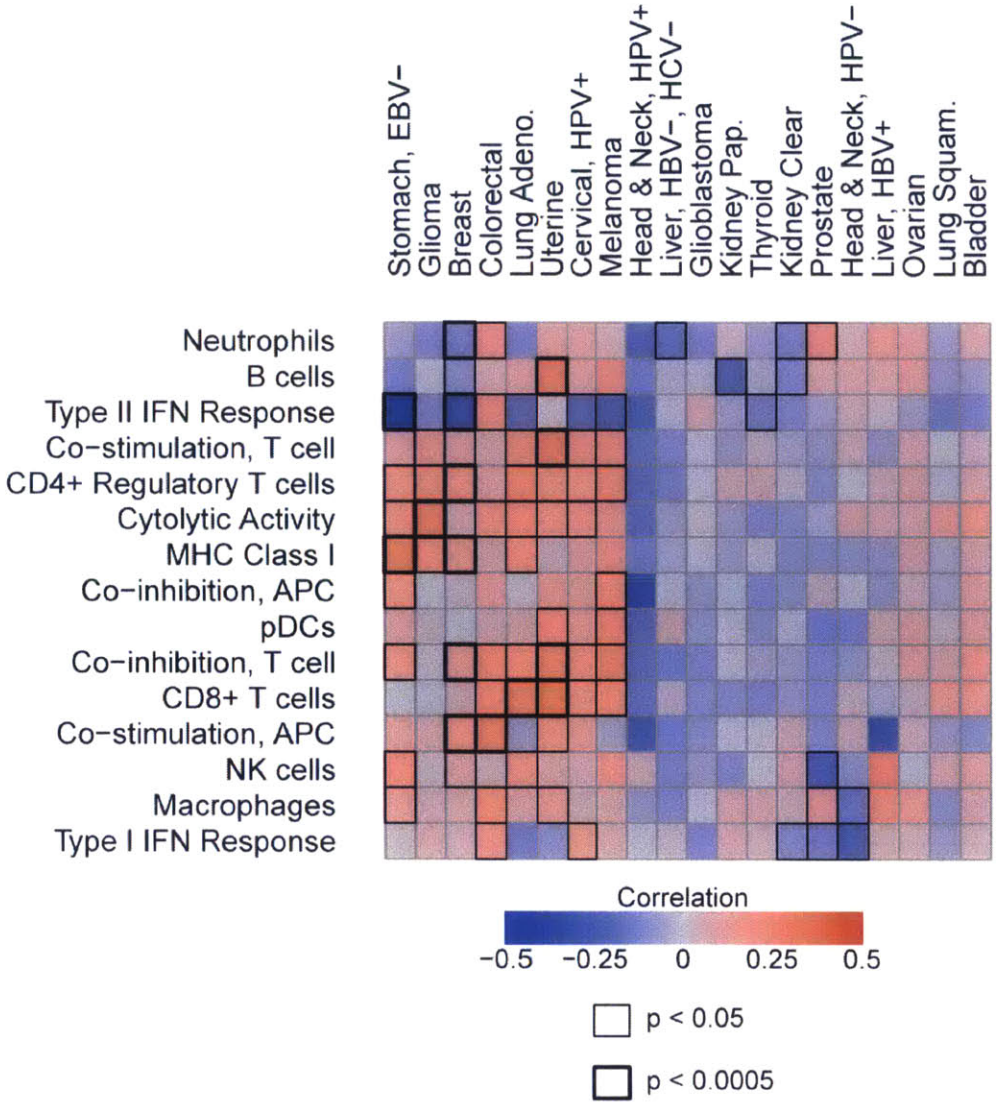






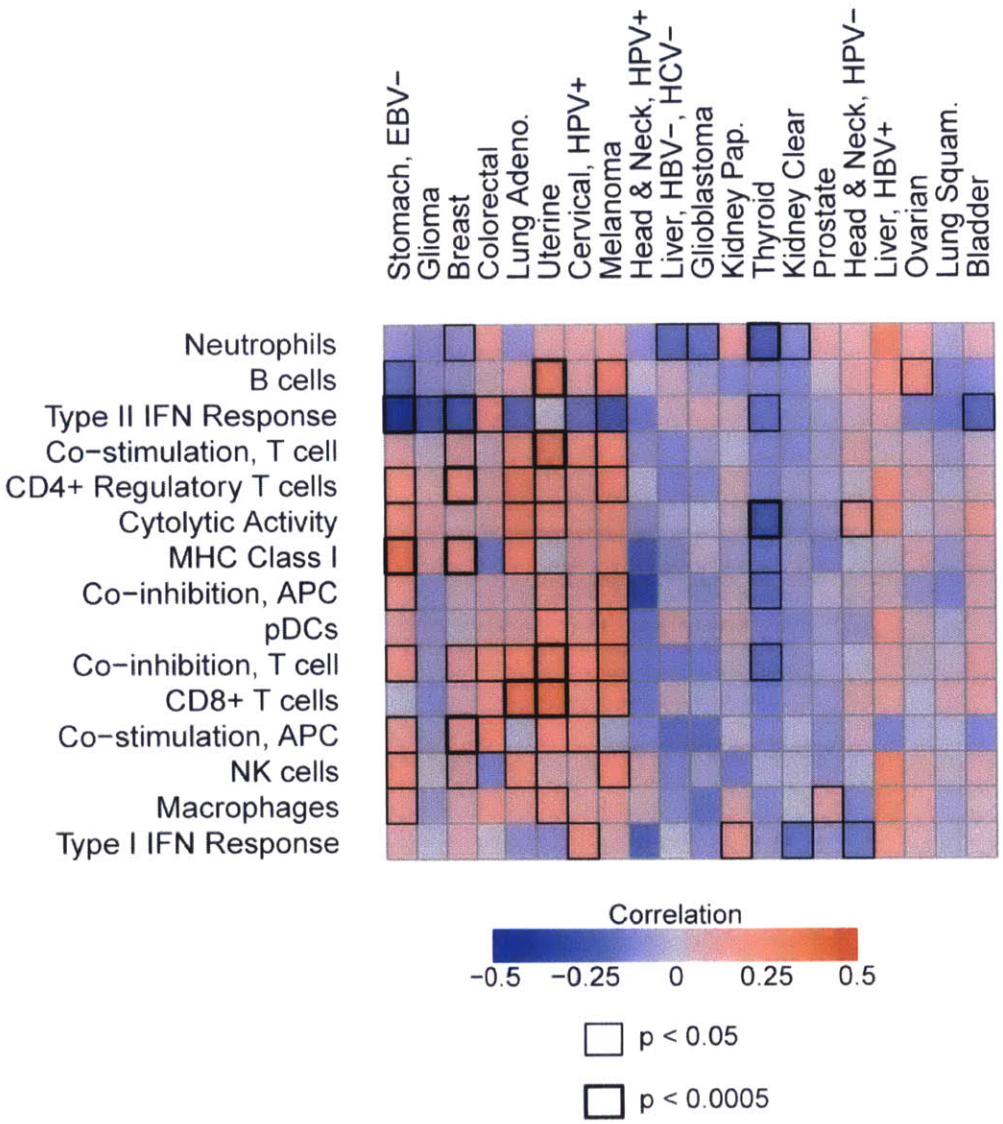
(c) Relationship between smoking and cytolytic activity in lung and head and neck tumors. Cytolytic activity for smokers and those reformed for less than 15 years verses never-smokers in lung squamous cell carcinoma, lung adenocarcinoma, and head and neck cancer. Solid bodies represent interquartile ranges and are notched by the median; vertical lines demarcate the 5th to 95th percentile range. P-values reflect Wilcoxon rank-sum tests.

Figure S3, continued



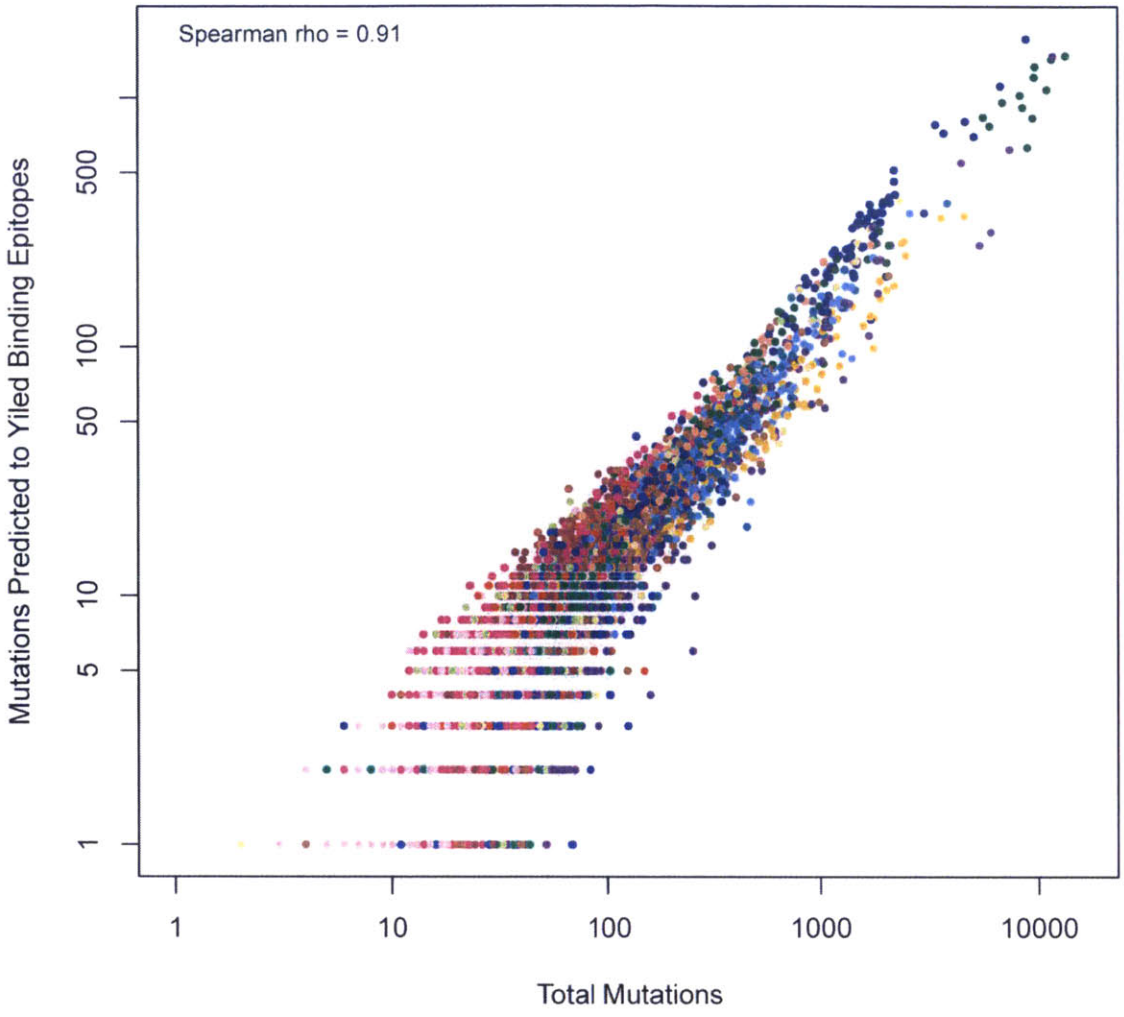
(d) Heatmap showing association between total count of mutations and cell type marker gene enrichment. Colors correspond to Spearman correlation, and borders indicate unadjusted p-value.

Figure S3, continued



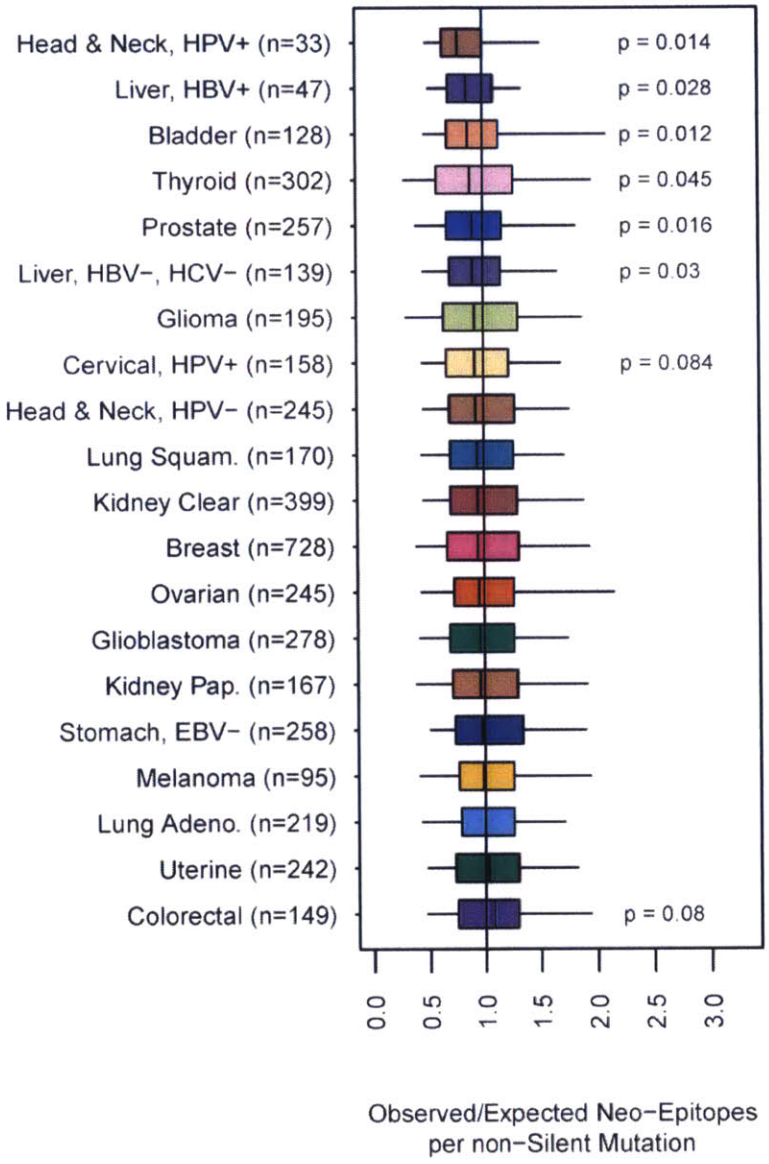
(e) Analogous to **(c)** but presenting association for neo-epitope counts shown in **(b)**

Figure S3, continued



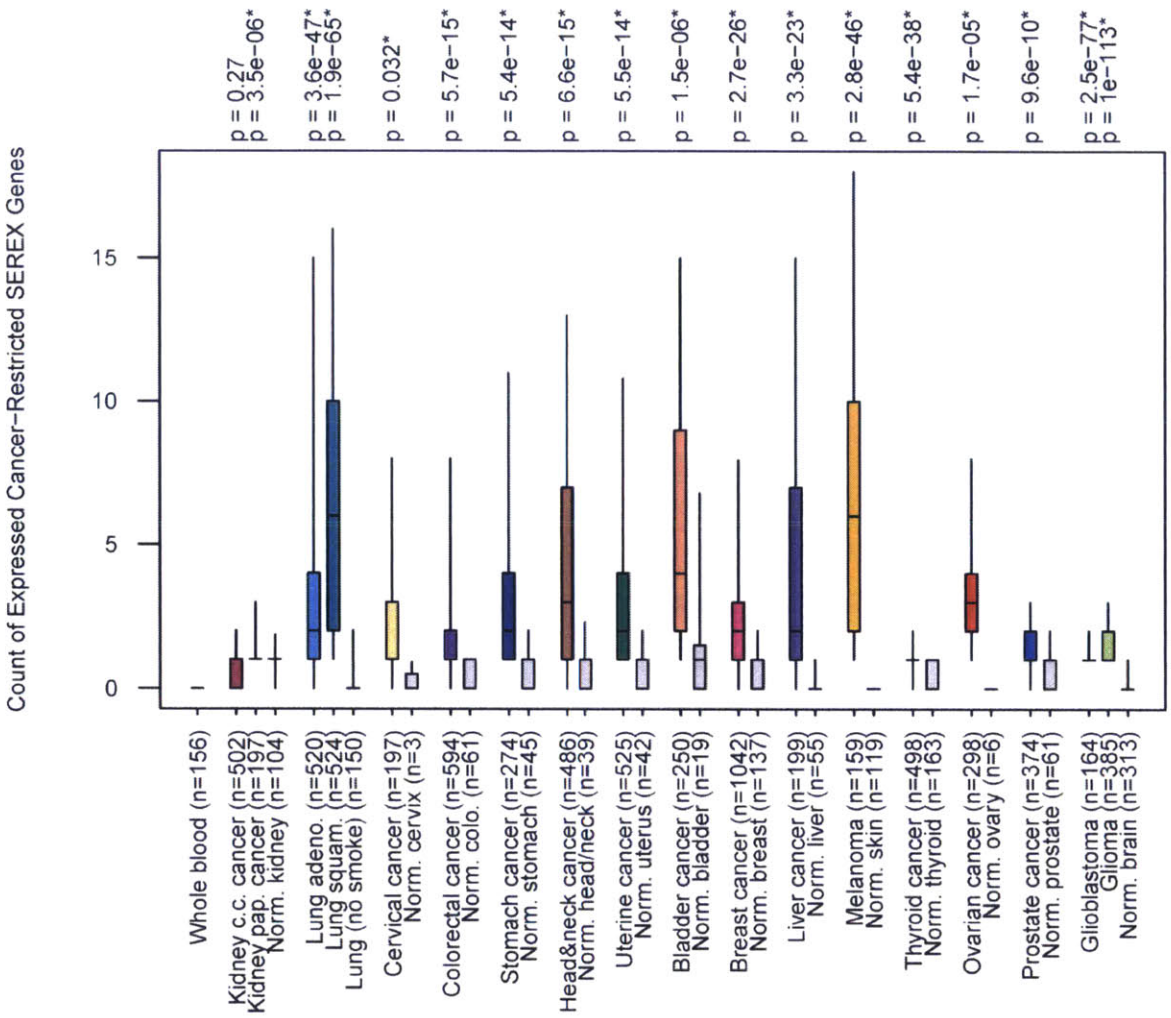
(f) Scatter plot showing correlation of total mutations and the count of predicted expressed neo-epitopes.

Figure S3, continued



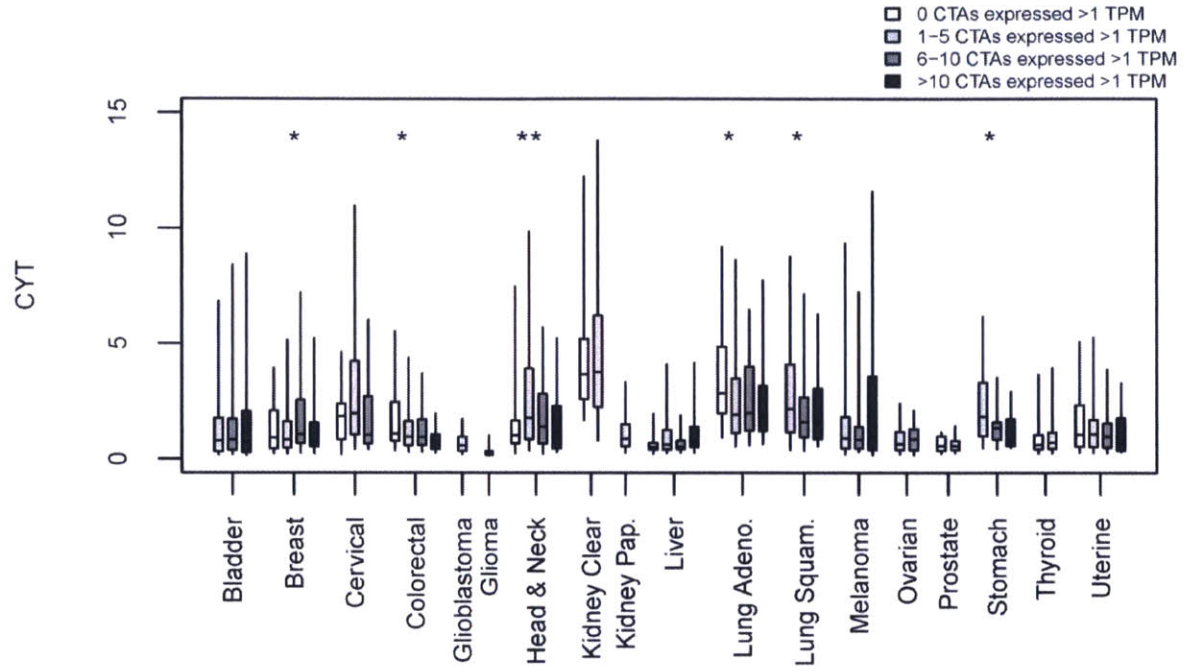
(**g**) Analogous to **Figure 3C**, but using neo-epitope prediction based on randomly re-permuted HLA genotype assignments (across patients).

Figure S4. Ectopic Gene expression and its correlates; Necrosis, related to Figure 4



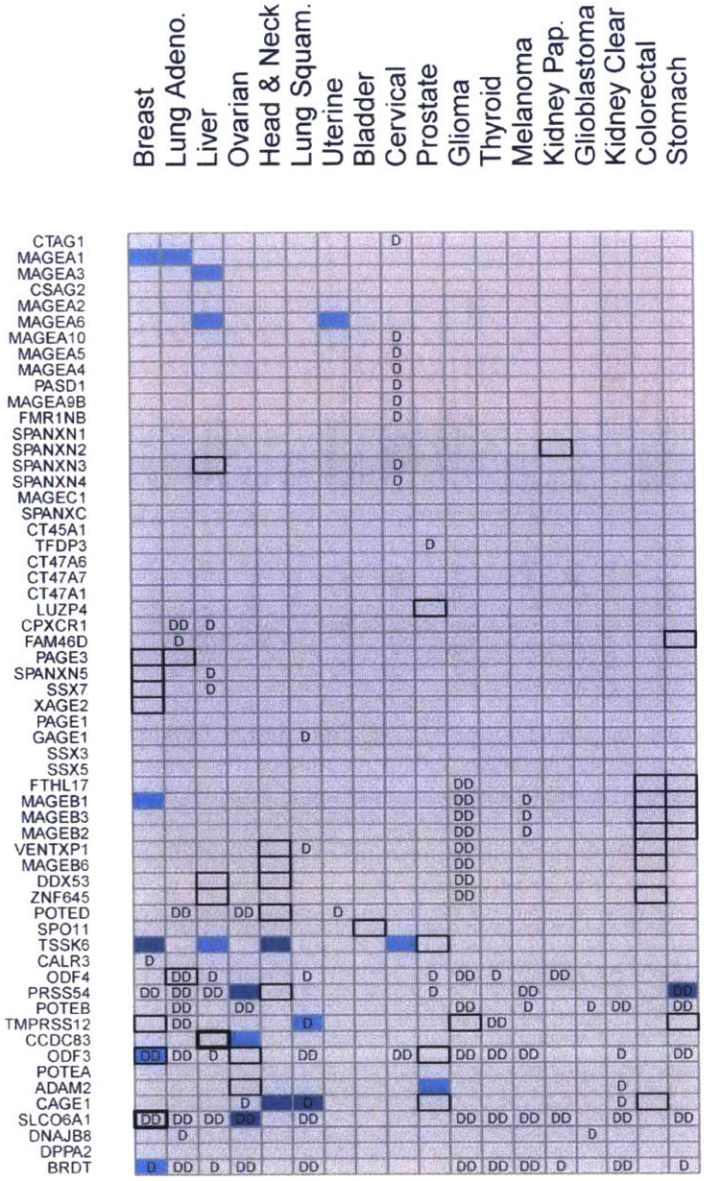
(a) Tumor-normal differences in the expression of cancer testis genes not expressed in GTEx normals. The count of select cancer testis genes (60 total; see Methods) expressed >1 TPM was calculated per sample and the distributions characterized in TCGA tumor samples verses normal samples from TCGA and GTEx. Solid bodies represent interquartile ranges and are notched by the median; vertical lines demarcate the 5th to 95th percentile range

Figure S4, continued



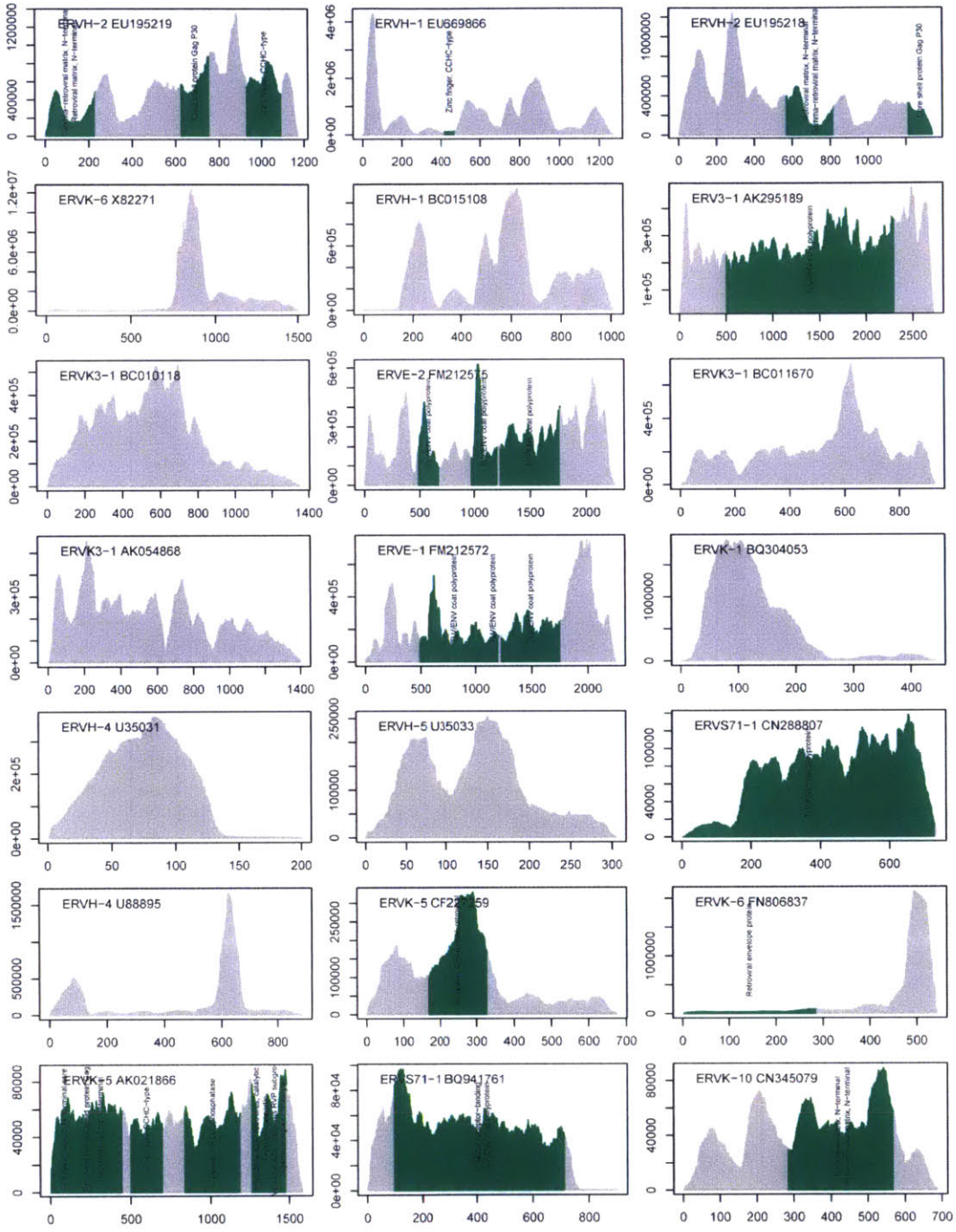
(b) Distributions of cytolytic activity in according to the count of CT antigens (CTAs) expressed >1 TPM (only bins with ≥10 samples are shown). Solid bodies represent interquartile ranges and are notched by the median; vertical lines demarcate the 5th to 95th percentile range. Asterisks denote $p<0.05$ for comparison of adjacent distributions (Wilcoxon rank-sum test).

Figure S4, continued



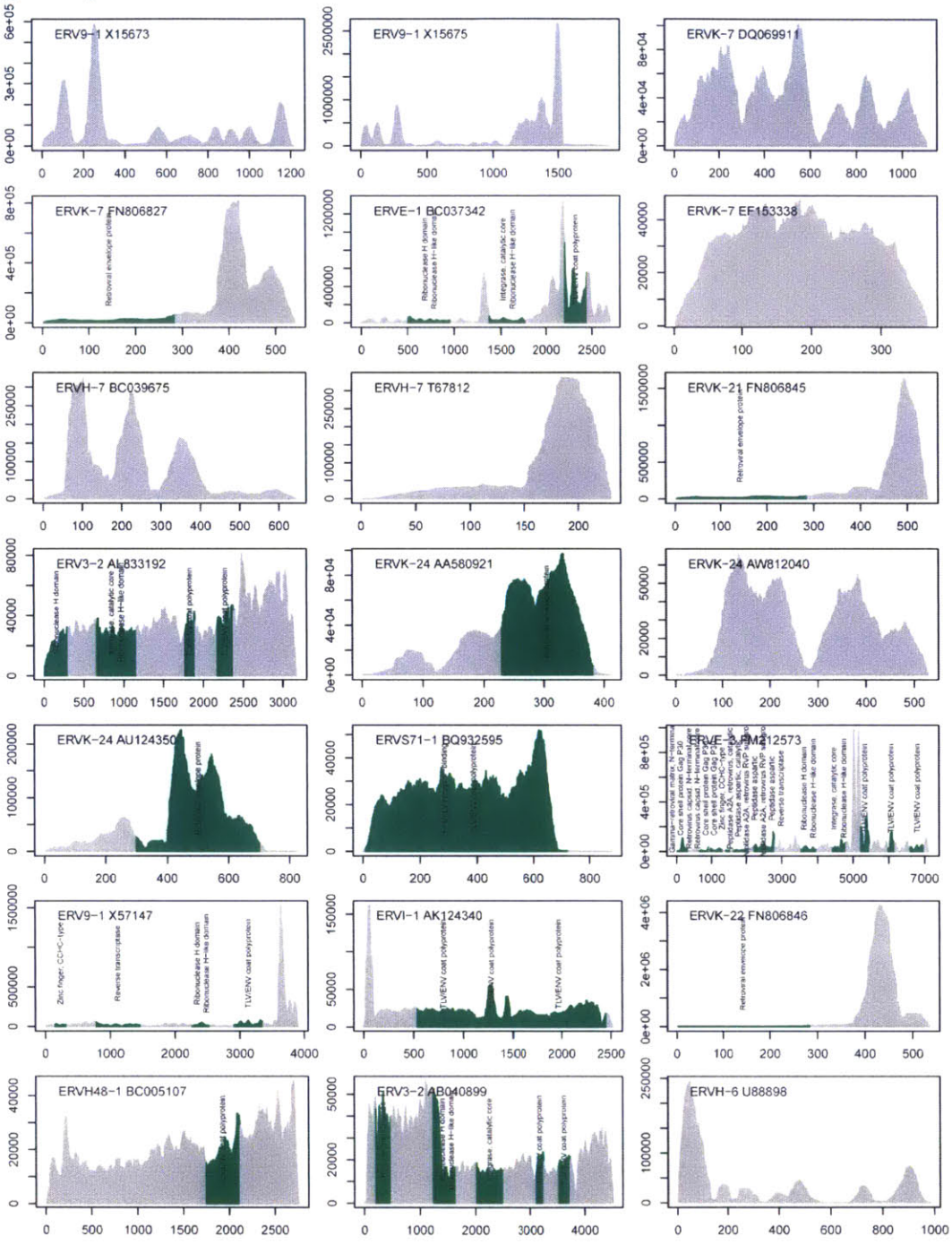
(**c**) Heatmap exploring chromosomal deletions targeting cancer testis genes. The color of each box indicates whether the given gene had lower expression when its locus was deleted. The outline indicates whether high CYT was positively associated with deletion status. The text (*blank* / "D" / "DD") indicates whether the locus was more likely to be deleted than amplified (with respect to average rate across genes in the tumor type). Thresholds reflect liberal nominal p-values, $p<0.1$ and $p<0.01$.

Figure S4, continued



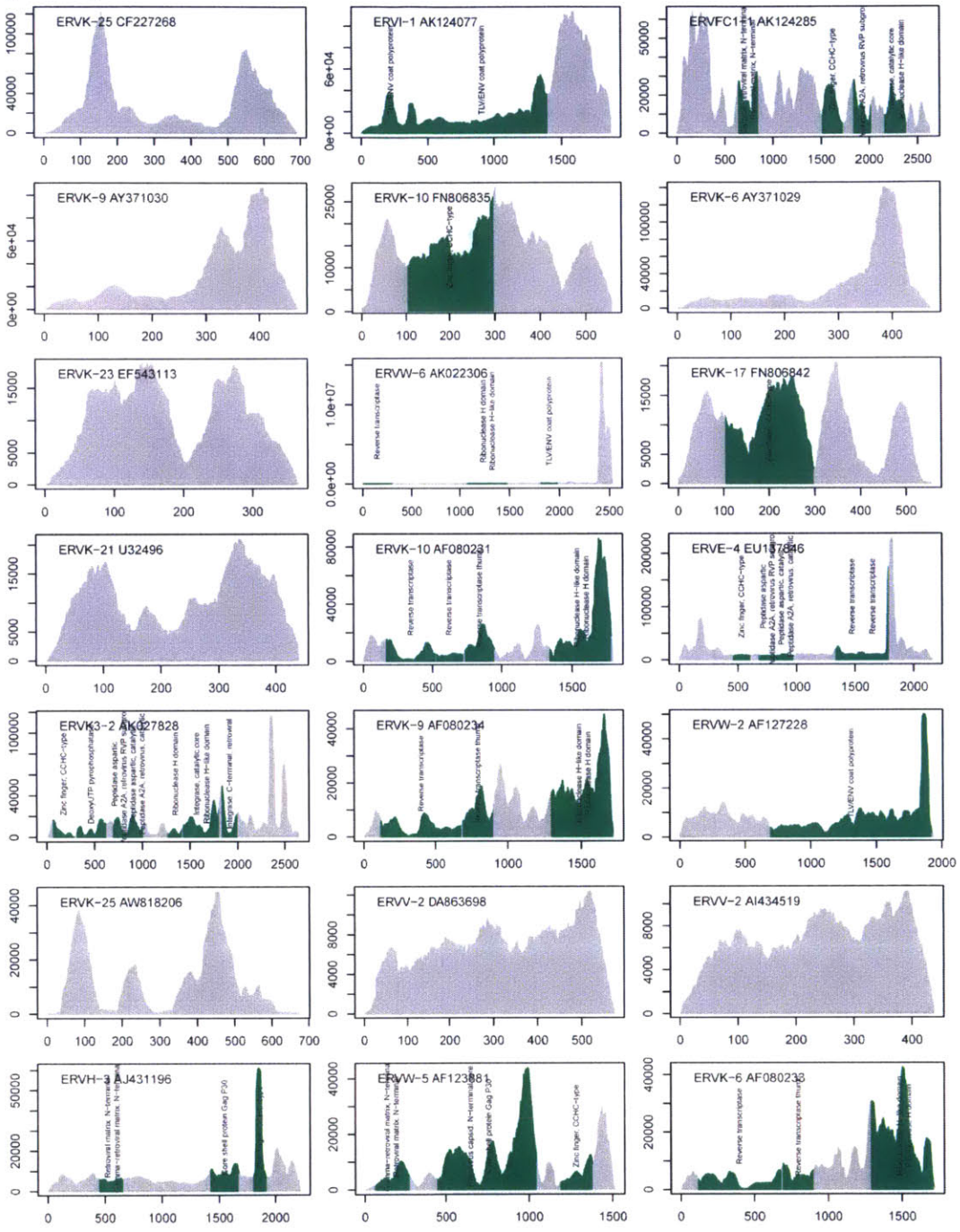
(d) Part 1 of 6. Coverage depth of ERV of reference sequence with reads from TCGA tumor samples. Each plot represents the depth of reads mapping to a given ERV reference sequence. Some ERVs are represented by multiple sequences. ORFs of length greater than 75nt that scored for InterProScan motifs are highlighted in green along with the name of the motif for which they scored.

Figure S4, continued



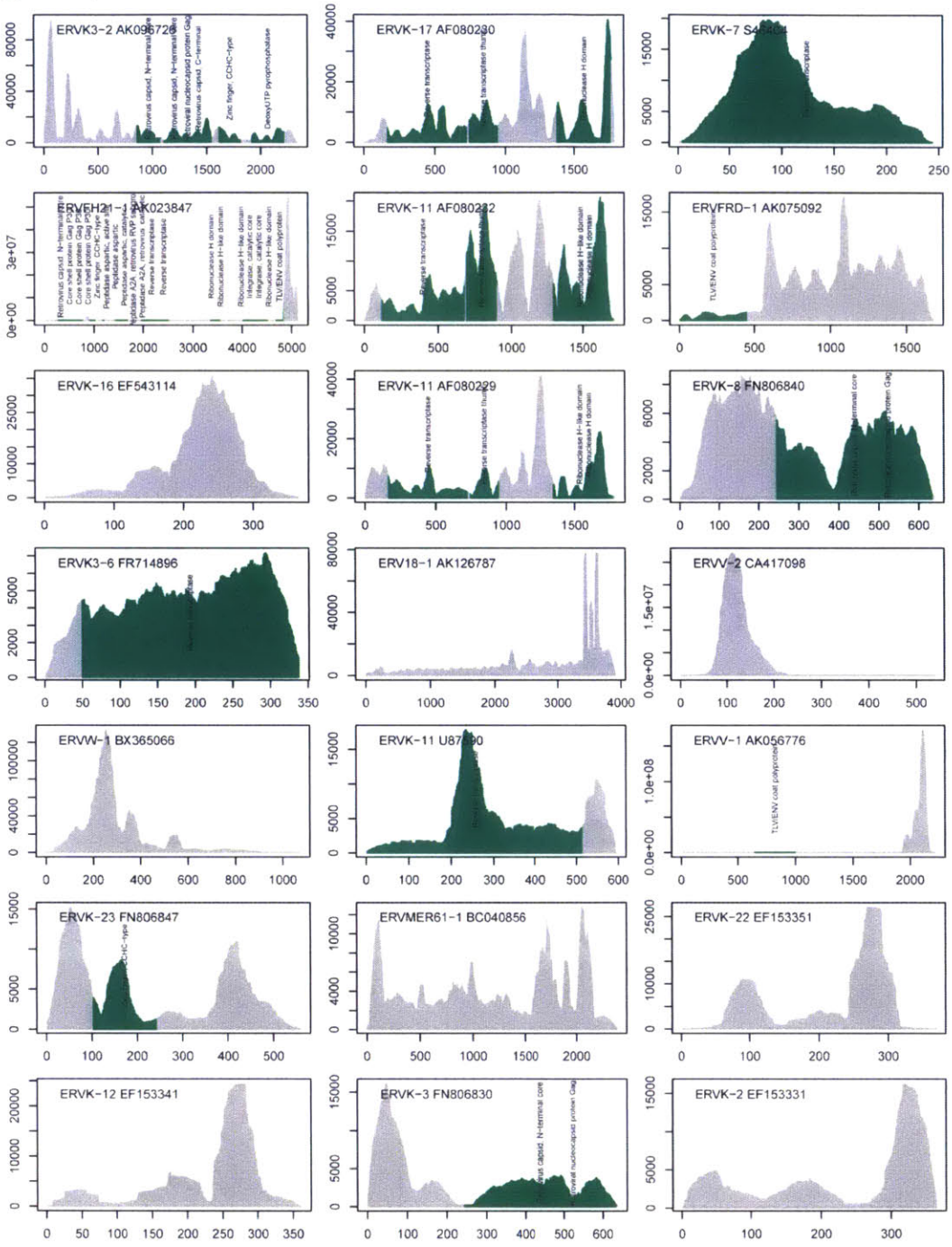
(d) Part 2 of 6. Coverage depth of ERV of reference sequence with reads from TCGA tumor samples. Each plot represents the depth of reads mapping to a given ERV reference sequence. Some ERVs are represented by multiple sequences. ORFs of length greater than 75nt that scored for InterProScan motifs are highlighted in green along with the name of the motif for which they scored.

Figure S4, continued



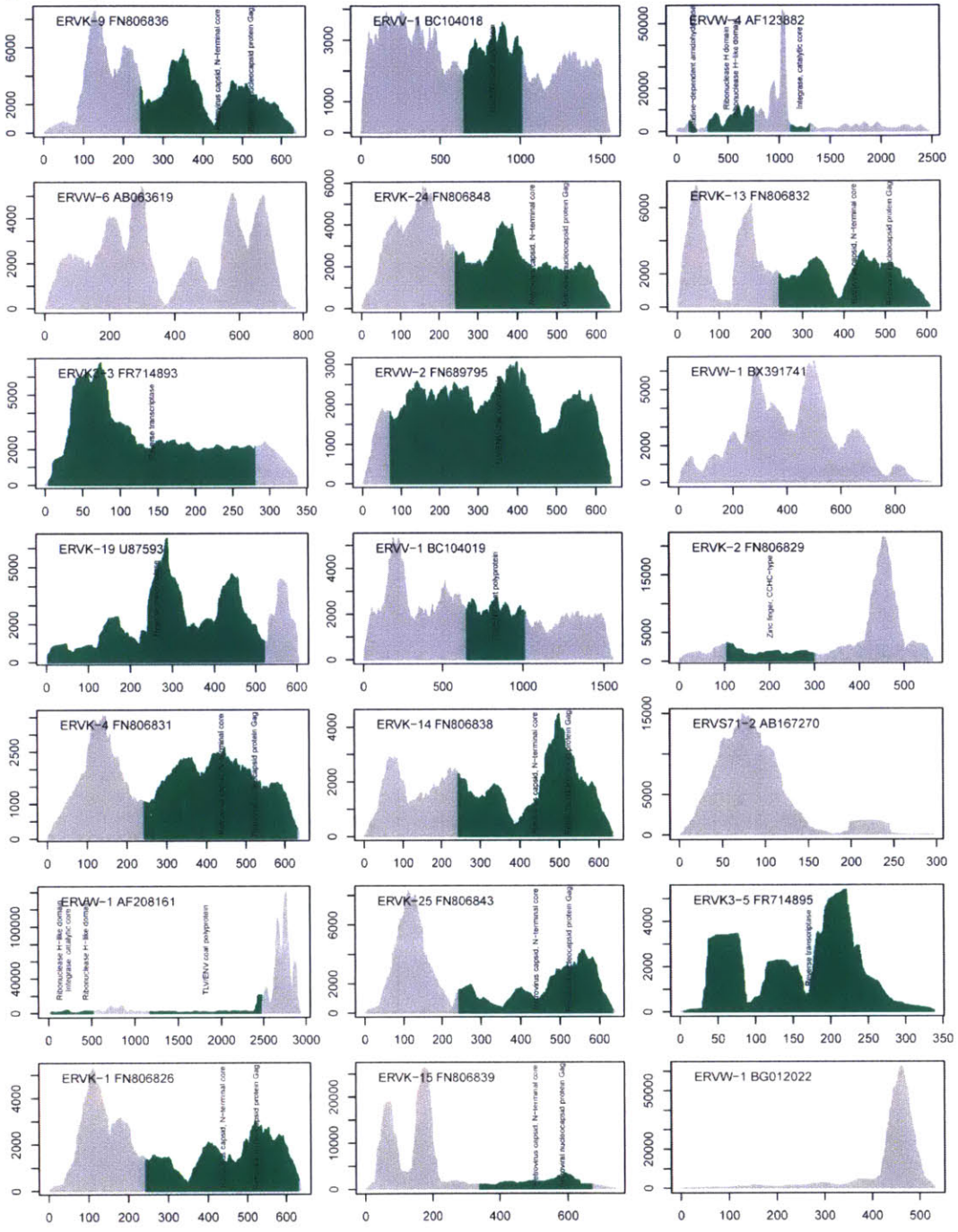
(d) Part 3 of 6. Coverage depth of ERV of reference sequence with reads from TCGA tumor samples. Each plot represents the depth of reads mapping to a given ERV reference sequence. Some ERVs are represented by multiple sequences. ORFs of length greater than 75nt that scored for InterProScan motifs are highlighted in green along with the name of the motif for which they scored.

Figure S4, continued



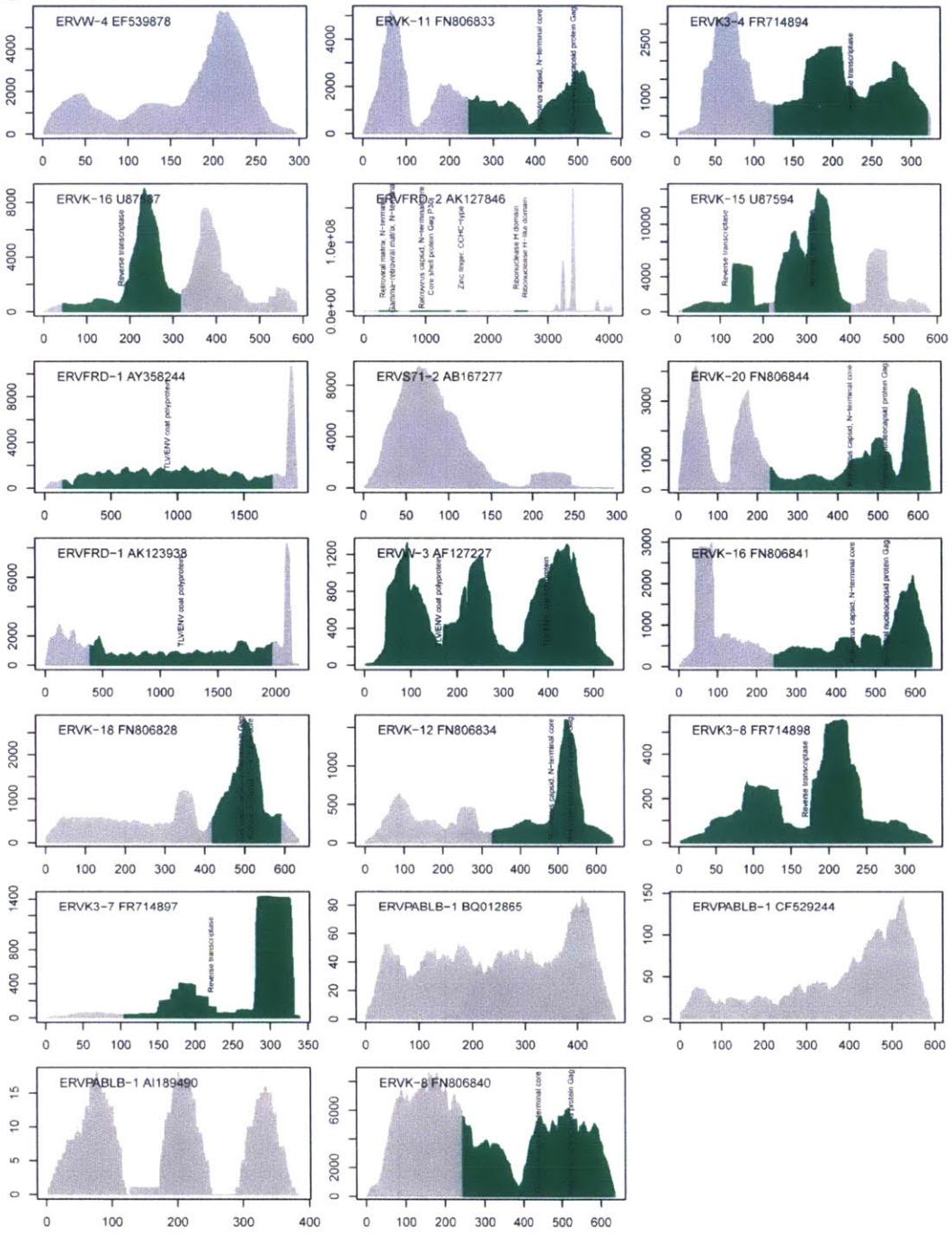
(d) Part 4 of 6. Coverage depth of ERV of reference sequence with reads from TCGA tumor samples. Each plot represents the depth of reads mapping to a given ERV reference sequence. Some ERVs are represented by multiple sequences. ORFs of length greater than 75nt that scored for InterProScan motifs are highlighted in green along with the name of the motif for which they scored.

Figure S4, continued



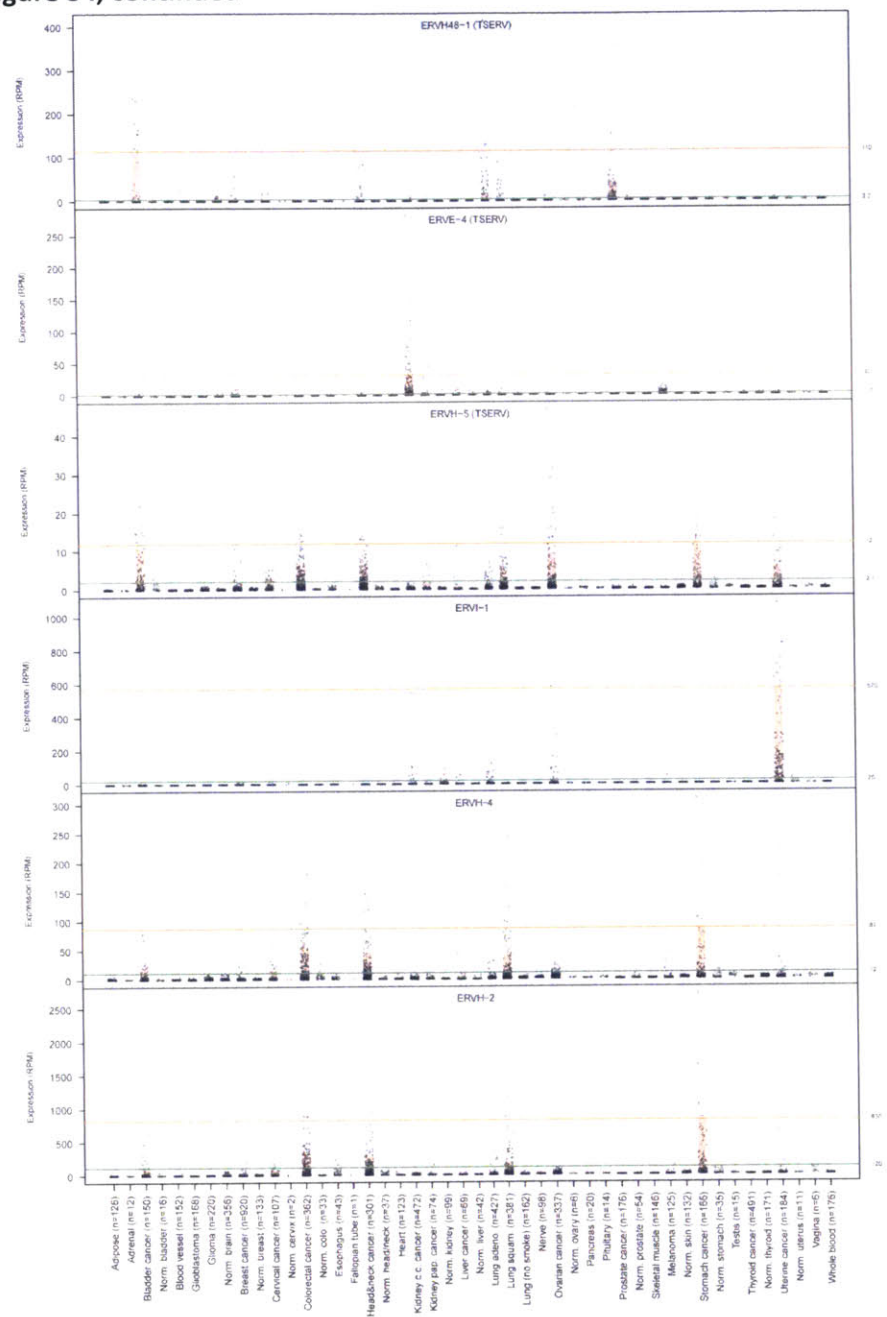
(d) Part 5 of 6. Coverage depth of ERV of reference sequence with reads from TCGA tumor samples. Each plot represents the depth of reads mapping to a given ERV reference sequence. Some ERVs are represented by multiple sequences. ORFs of length greater than 75nt that scored for InterProScan motifs are highlighted in green along with the name of the motif for which they scored.

Figure S4, continued



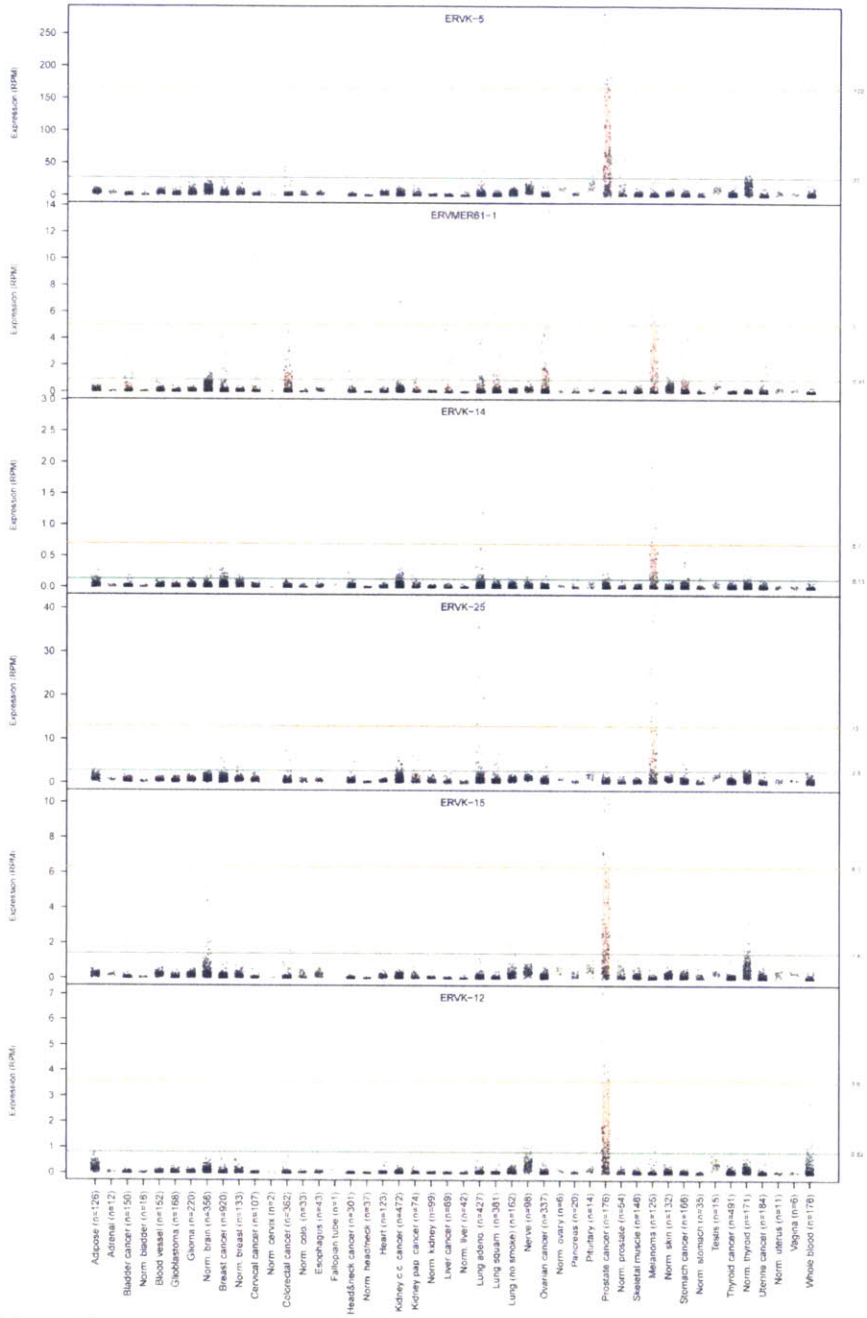
(d) Part 6 of 6. Coverage depth of ERV of reference sequence with reads from TCGA tumor samples. Each plot represents the depth of reads mapping to a given ERV reference sequence. Some ERVs are represented by multiple sequences. ORFs of length greater than 75nt that scored for InterProScan motifs are highlighted in green along with the name of the motif for which they scored.

Figure S4, continued



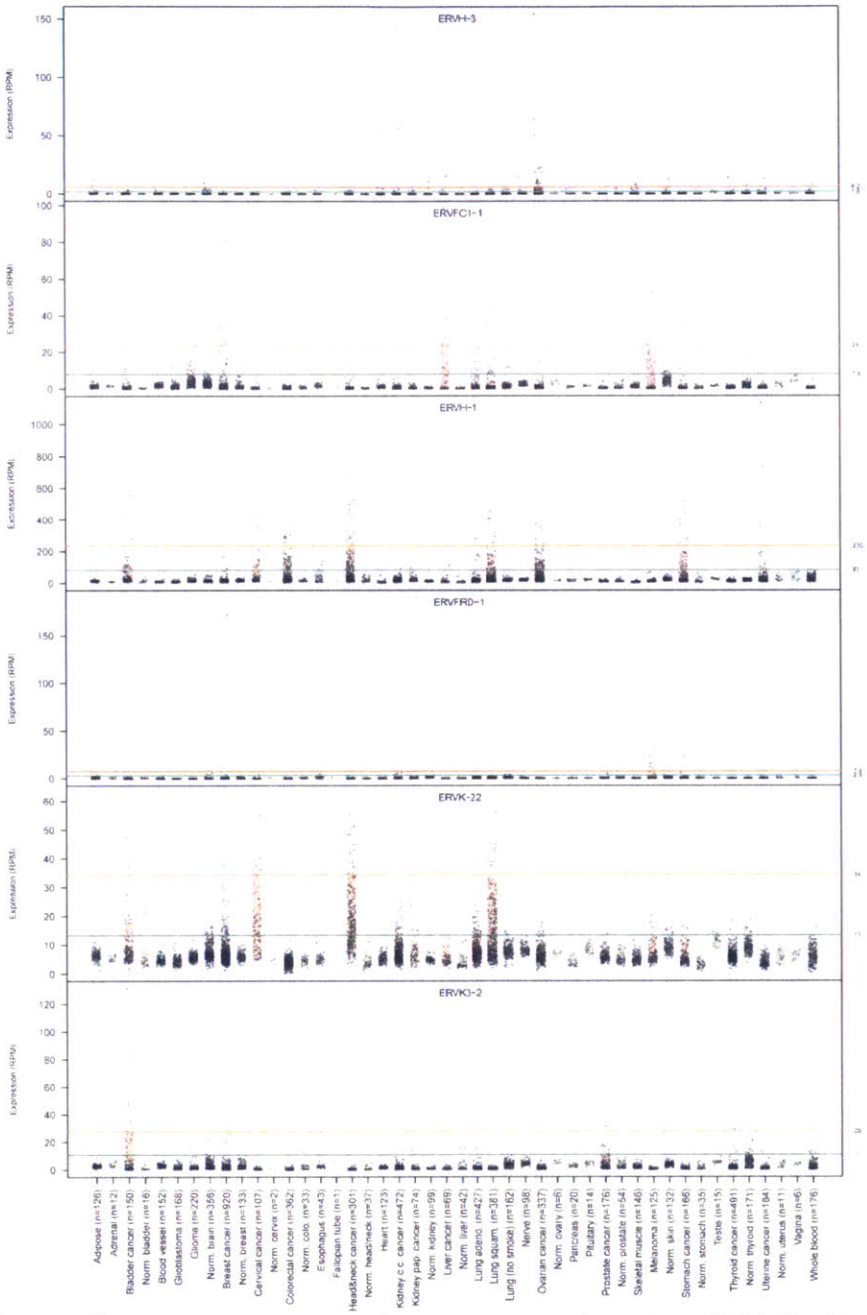
(e) Part 1 of 11. ERV expression ranges for tumor vs. normal samples, all ERVs. Semi-transparent left-right jittered points represent the expression values observed in a compendium of tissues. The 5th to 95th percentile range is highlighted in orange for tumor tissues and in green for normal tissues. The maximum 95th percentile value observed in tumors is marked with an orange horizontal line, and the corresponding maximum for normal tissues is marked with a green horizontal line. These values (marked on the right axis) were the basis for determining tumor-specific expression. ERVs designated as TSERVs are marked as such. Many ERVs, while not specific to tumors, were elevated in tumors.

Figure S4, continued



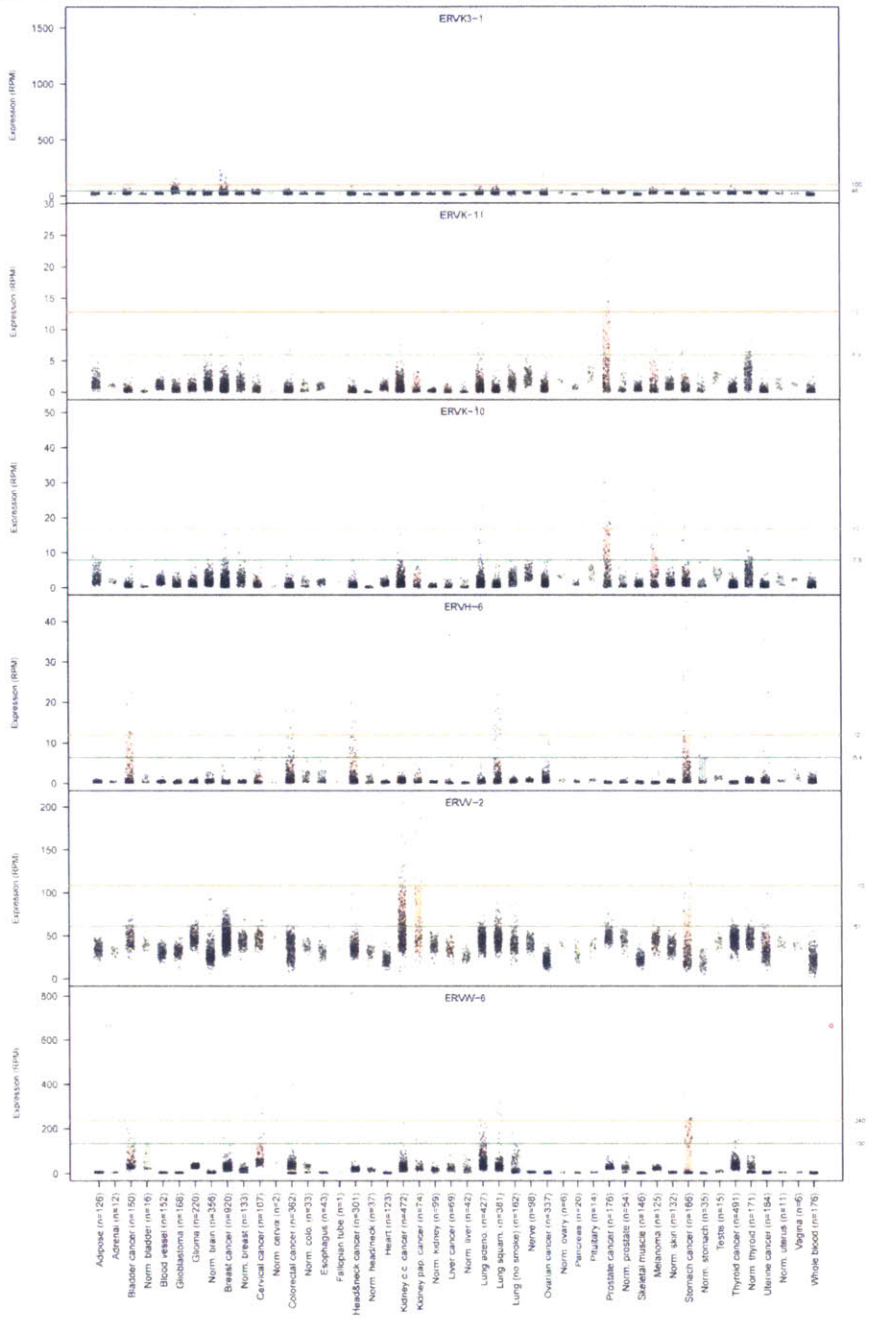
(e) Part 2 of 11. ERV expression ranges for tumor vs. normal samples, all ERVs. Semi-transparent left-right jittered points represent the expression values observed in a compendium of tissues. The 5th to 95th percentile range is highlighted in orange for tumor tissues and in green for normal tissues. The maximum 95th percentile value observed in tumors is marked with an orange horizontal line, and the corresponding maximum for normal tissues is marked with a green horizontal line. These values (marked on the right axis) were the basis for determining tumor-specific expression. ERVs designated as TSERVs are marked as such. Many ERVs, while not specific to tumors, were elevated in tumors.

Figure S4, continued



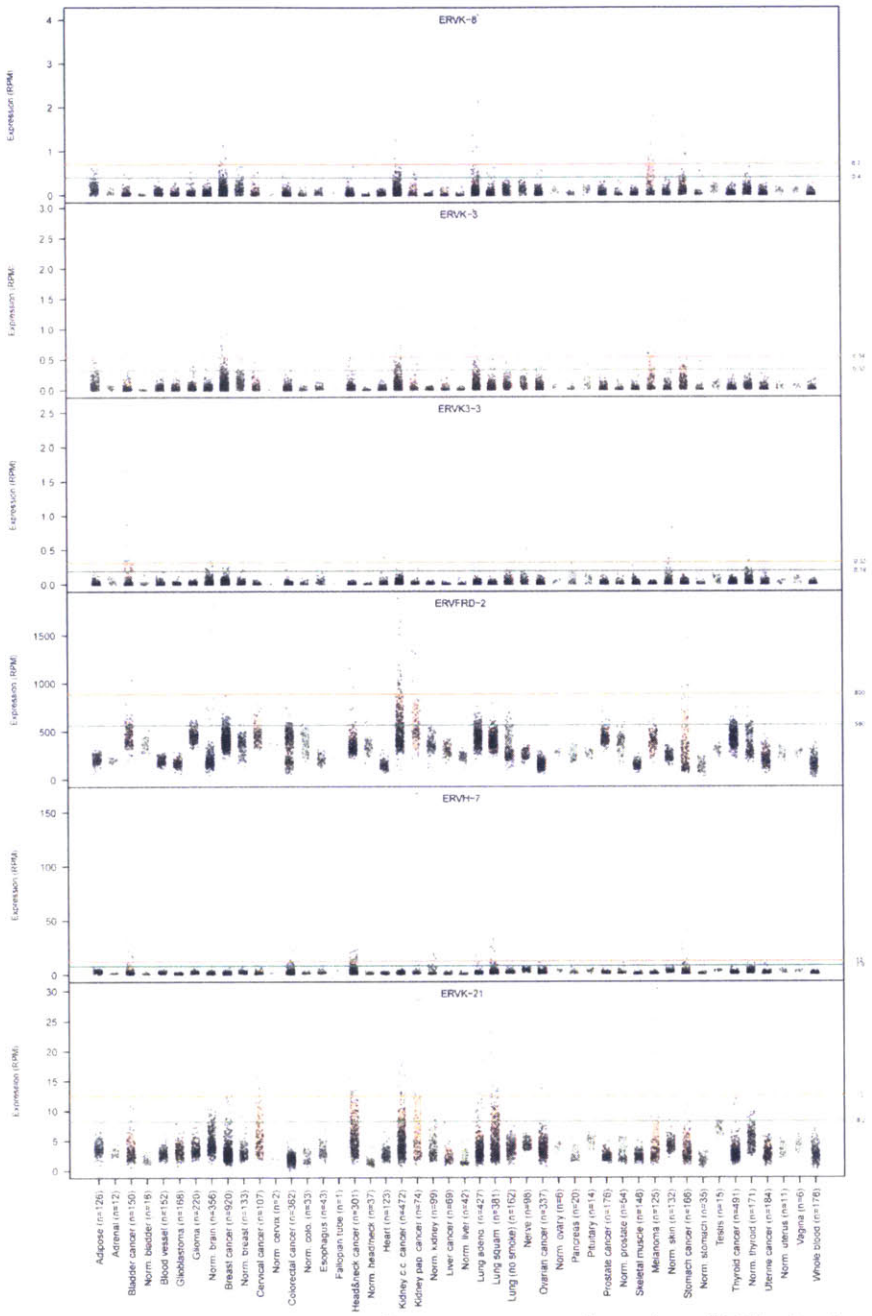
(e) Part 3 of 11. ERV expression ranges for tumor vs. normal samples, all ERVs. Semi-transparent left-right jittered points represent the expression values observed in a compendium of tissues. The 5th to 95th percentile range is highlighted in orange for tumor tissues and in green for normal tissues. The maximum 95th percentile value observed in tumors is marked with an orange horizontal line, and the corresponding maximum for normal tissues is marked with a green horizontal line. These values (marked on the right axis) were the basis for determining tumor-specific expression. ERVs designated as TSERVs are marked as such. Many ERVs, while not specific to tumors, were elevated in tumors.

Figure S4, continued



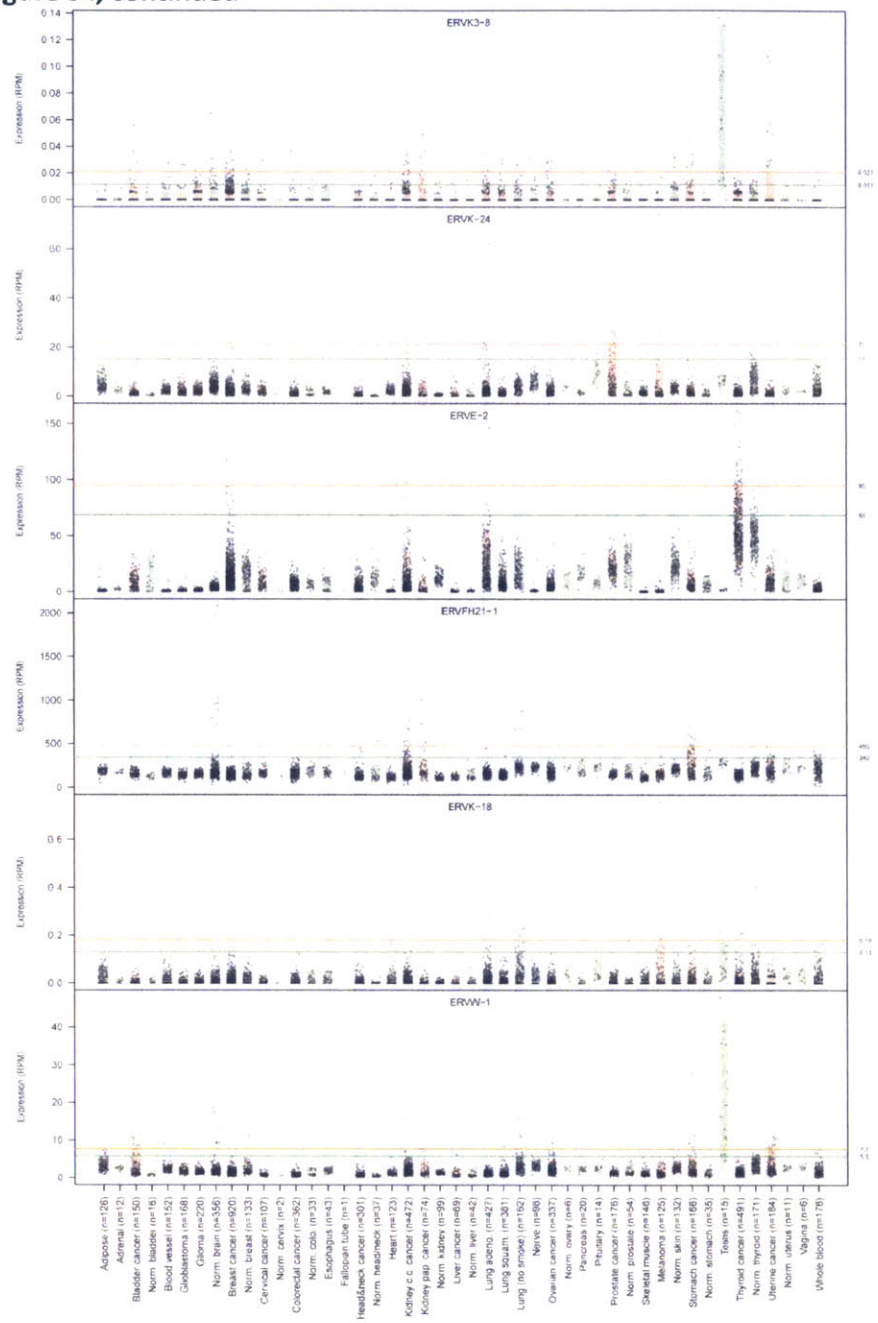
(e) Part 4 of 11. ERV expression ranges for tumor vs. normal samples, all ERVs. Semi-transparent left-right jittered points represent the expression values observed in a compendium of tissues. The 5th to 95th percentile range is highlighted in orange for tumor tissues and in green for normal tissues. The maximum 95th percentile value observed in tumors is marked with an orange horizontal line, and the corresponding maximum for normal tissues is marked with a green horizontal line. These values (marked on the right axis) were the basis for determining tumor-specific expression. ERVs designated as TSERVs are marked as such. Many ERVs, while not specific to tumors, were elevated in tumors.

Figure S4, continued



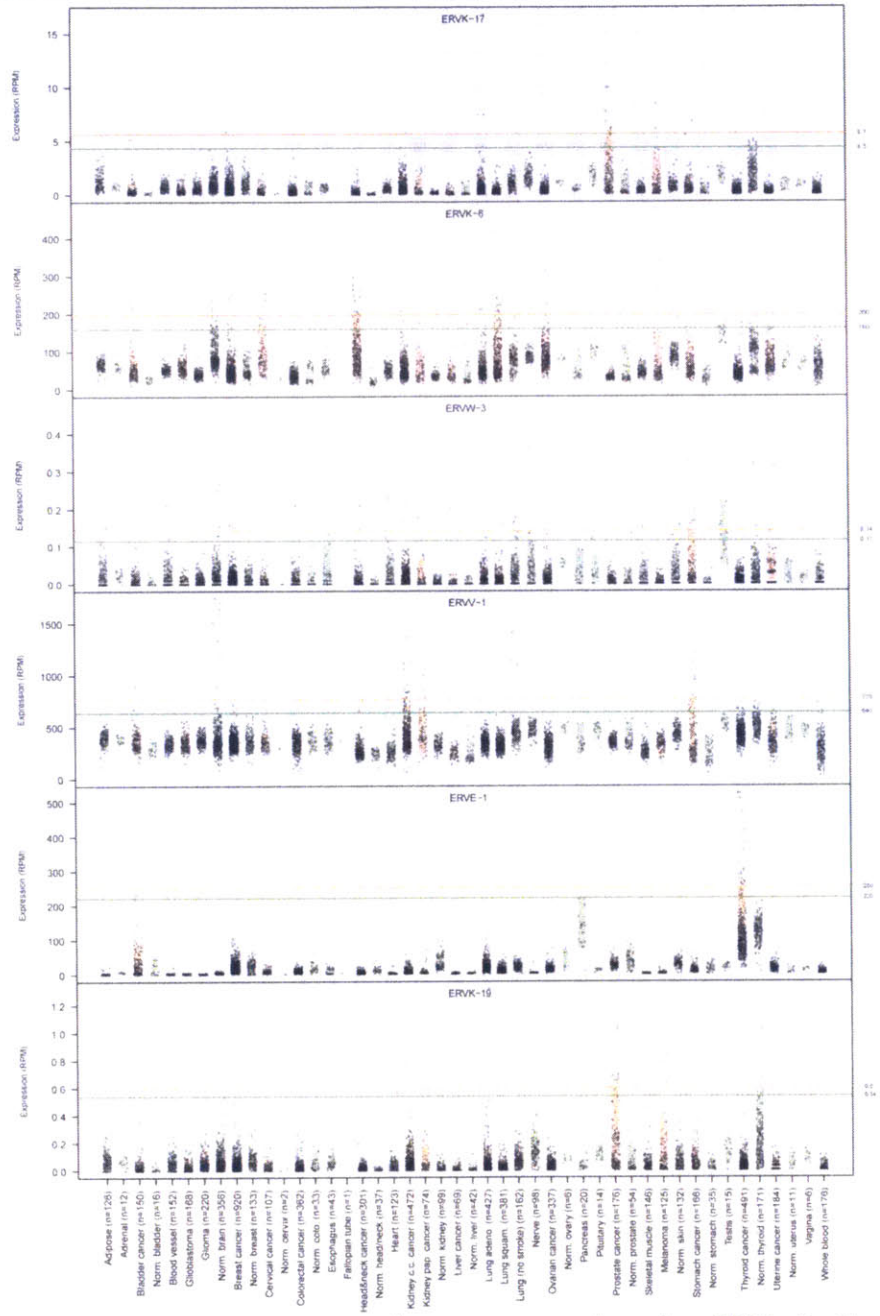
(e) Part 5 of 11. ERV expression ranges for tumor vs. normal samples, all ERVs. Semi-transparent left-right jittered points represent the expression values observed in a compendium of tissues. The 5th to 95th percentile range is highlighted in orange for tumor tissues and in green for normal tissues. The maximum 95th percentile value observed in tumors is marked with an orange horizontal line, and the corresponding maximum for normal tissues is marked with a green horizontal line. These values (marked on the right axis) were the basis for determining tumor-specific expression. ERVs designated as TSERVs are marked as such. Many ERVs, while not specific to tumors, were elevated in tumors.

Figure S4, continued



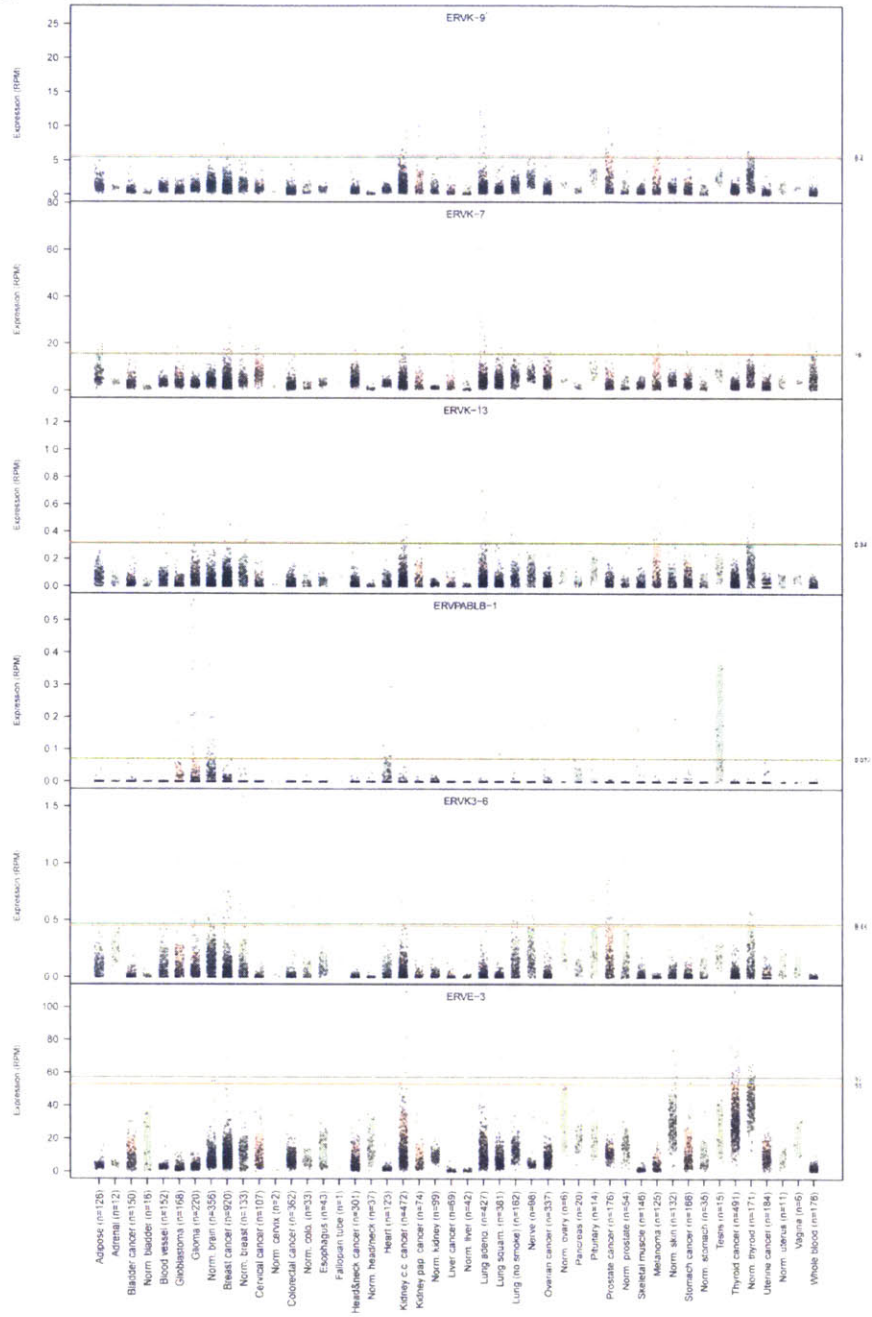
(e) Part 6 of 11. ERV expression ranges for tumor vs. normal samples, all ERVs. Semi-transparent left-right jittered points represent the expression values observed in a compendium of tissues. The 5th to 95th percentile range is highlighted in orange for tumor tissues and in green for normal tissues. The maximum 95th percentile value observed in tumors is marked with an orange horizontal line, and the corresponding maximum for normal tissues is marked with a green horizontal line. These values (marked on the right axis) were the basis for determining tumor-specific expression. ERVs designated as TSERVs are marked as such. Many ERVs, while not specific to tumors, were elevated in tumors.

Figure S4, continued



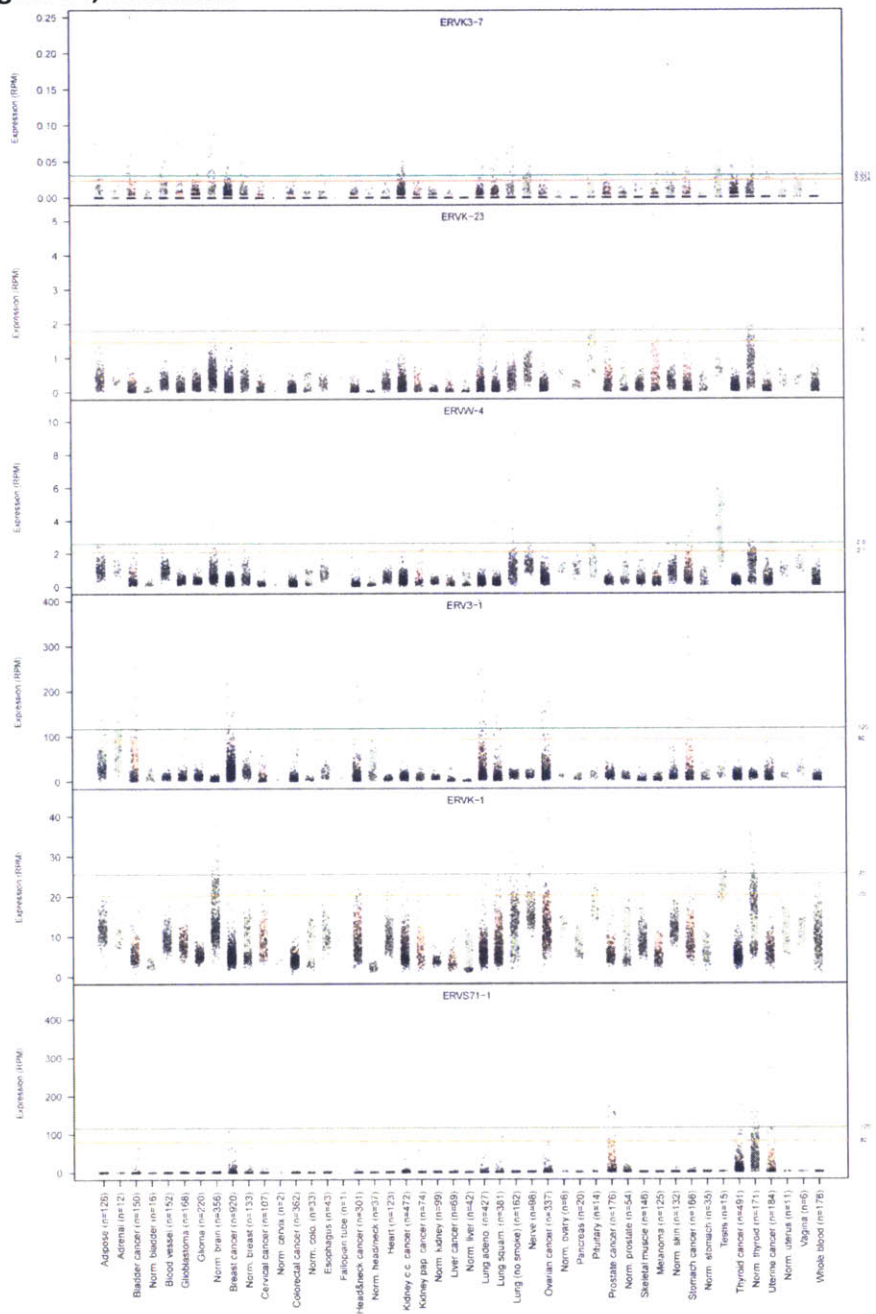
(e) Part 7 of 11. ERV expression ranges for tumor vs. normal samples, all ERVs. Semi-transparent left-right jittered points represent the expression values observed in a compendium of tissues. The 5th to 95th percentile range is highlighted in orange for tumor tissues and in green for normal tissues. The maximum 95th percentile value observed in tumors is marked with an orange horizontal line, and the corresponding maximum for normal tissues is marked with a green horizontal line. These values (marked on the right axis) were the basis for determining tumor-specific expression. ERVs designated as TSERVs are marked as such. Many ERVs, while not specific to tumors, were elevated in tumors.

Figure S4, continued



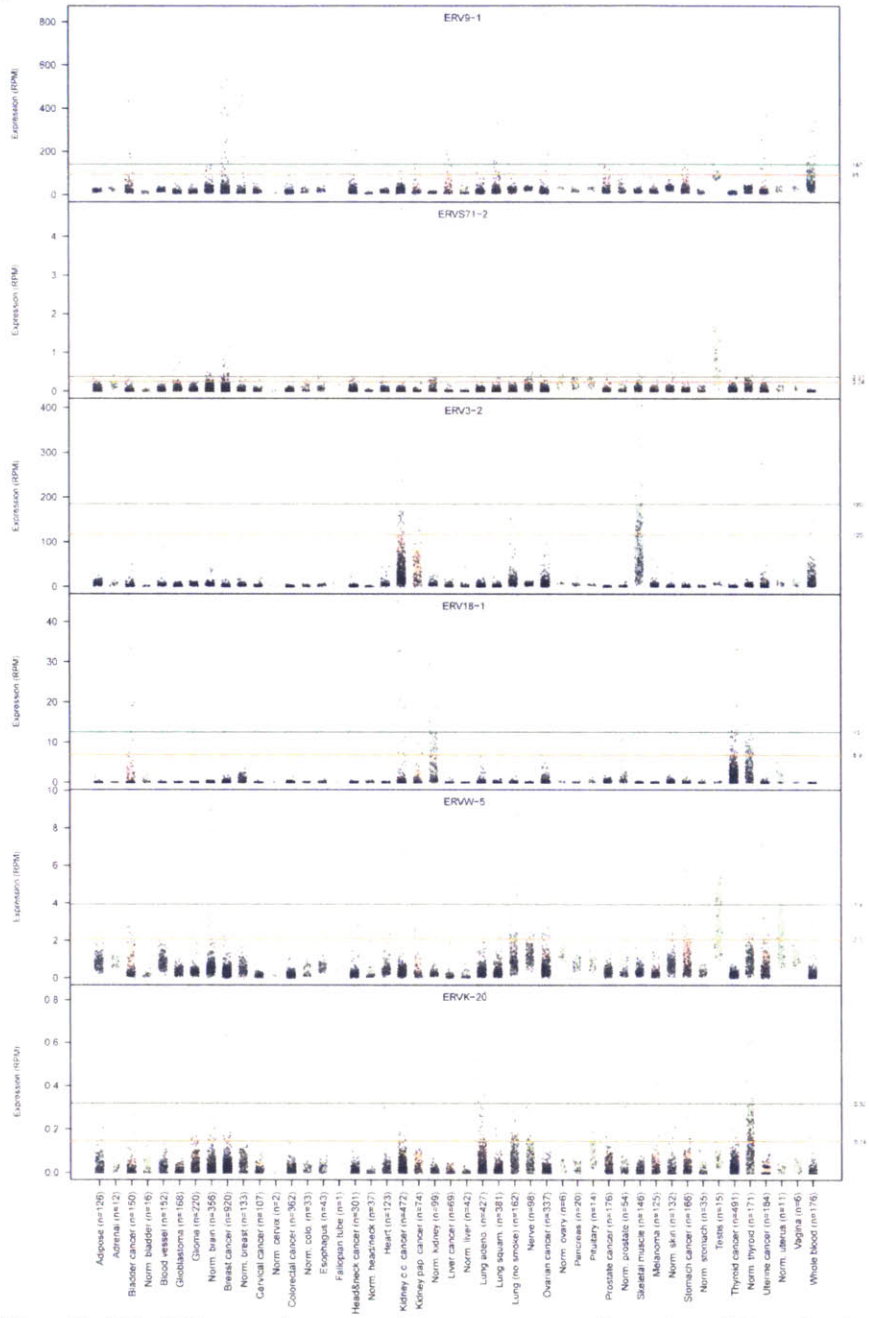
(e) Part 8 of 11. ERV expression ranges for tumor vs. normal samples, all ERVs. Semi-transparent left-right jittered points represent the expression values observed in a compendium of tissues. The 5th to 95th percentile range is highlighted in orange for tumor tissues and in green for normal tissues. The maximum 95th percentile value observed in tumors is marked with an orange horizontal line, and the corresponding maximum for normal tissues is marked with a green horizontal line. These values (marked on the right axis) were the basis for determining tumor-specific expression. ERVs designated as TSERVs are marked as such. Many ERVs, while not specific to tumors, were elevated in tumors.

Figure S4, continued



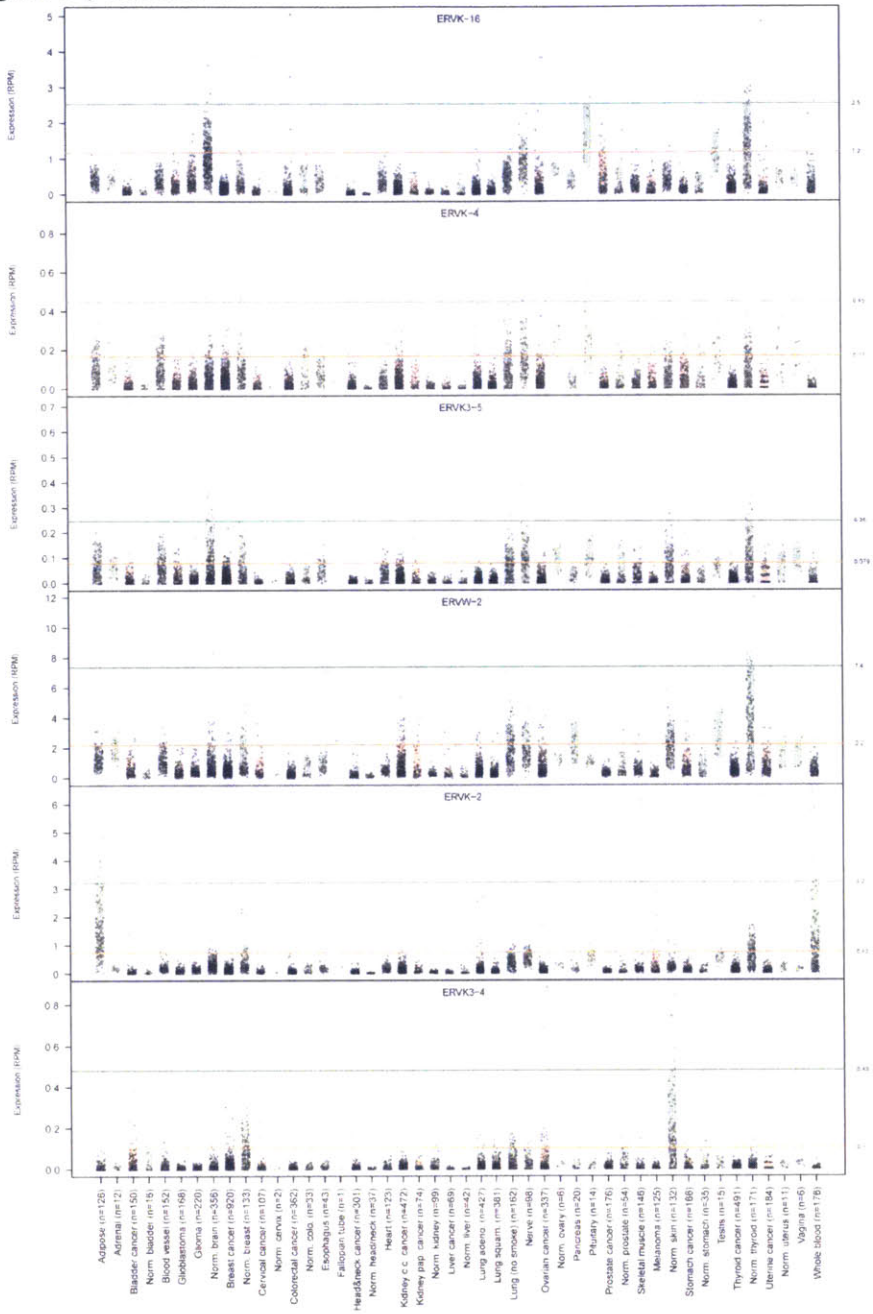
(e) Part 9 of 11. ERV expression ranges for tumor vs. normal samples, all ERVs. Semi-transparent left-right jittered points represent the expression values observed in a compendium of tissues. The 5th to 95th percentile range is highlighted in orange for tumor tissues and in green for normal tissues. The maximum 95th percentile value observed in tumors is marked with an orange horizontal line, and the corresponding maximum for normal tissues is marked with a green horizontal line. These values (marked on the right axis) were the basis for determining tumor-specific expression. ERVs designated as TSERVs are marked as such. Many ERVs, while not specific to tumors, were elevated in tumors.

Figure S4, continued

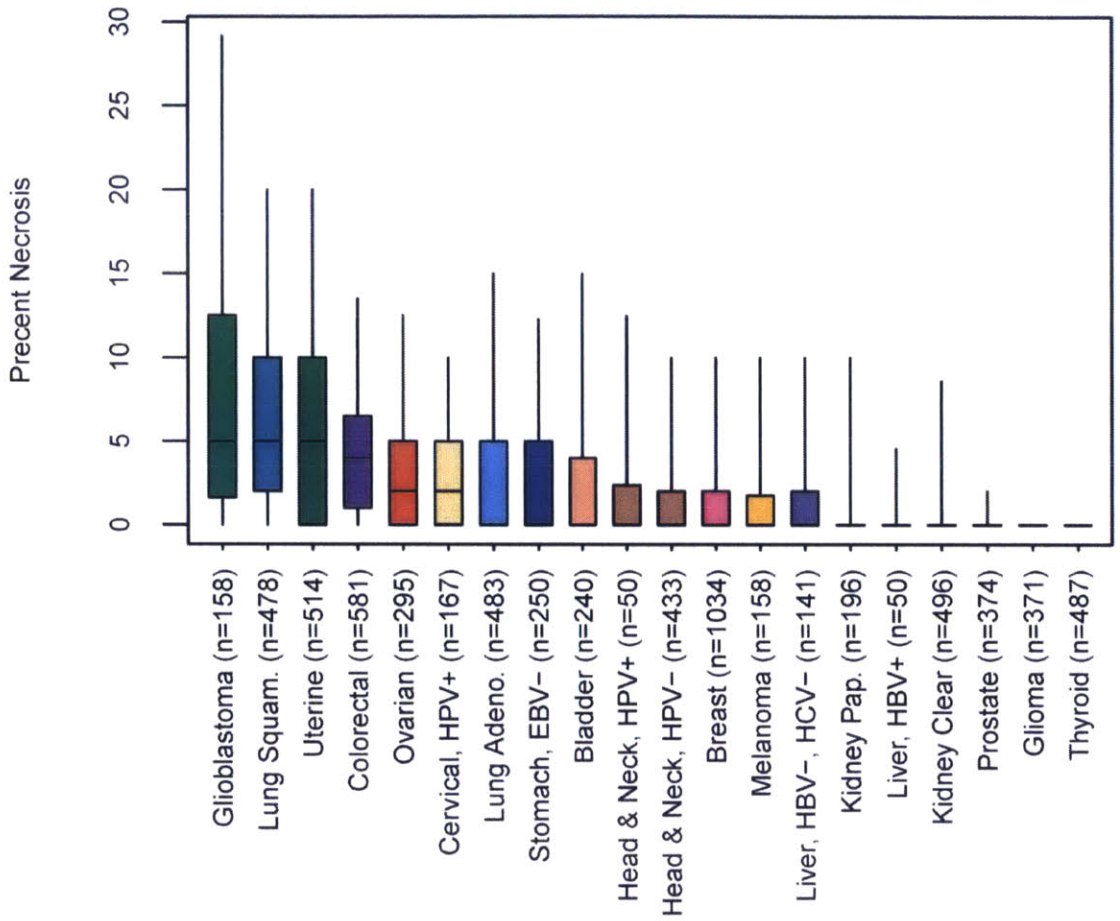


(e) Part 10 of 11. ERV expression ranges for tumor vs. normal samples, all ERVs. Semi-transparent left-right jittered points represent the expression values observed in a compendium of tissues. The 5th to 95th percentile range is highlighted in orange for tumor tissues and in green for normal tissues. The maximum 95th percentile value observed in tumors is marked with an orange horizontal line, and the corresponding maximum for normal tissues is marked with a green horizontal line. These values (marked on the right axis) were the basis for determining tumor-specific expression. ERVs designated as TSERVs are marked as such. Many ERVs, while not specific to tumors, were elevated in tumors.

Figure S4, continued

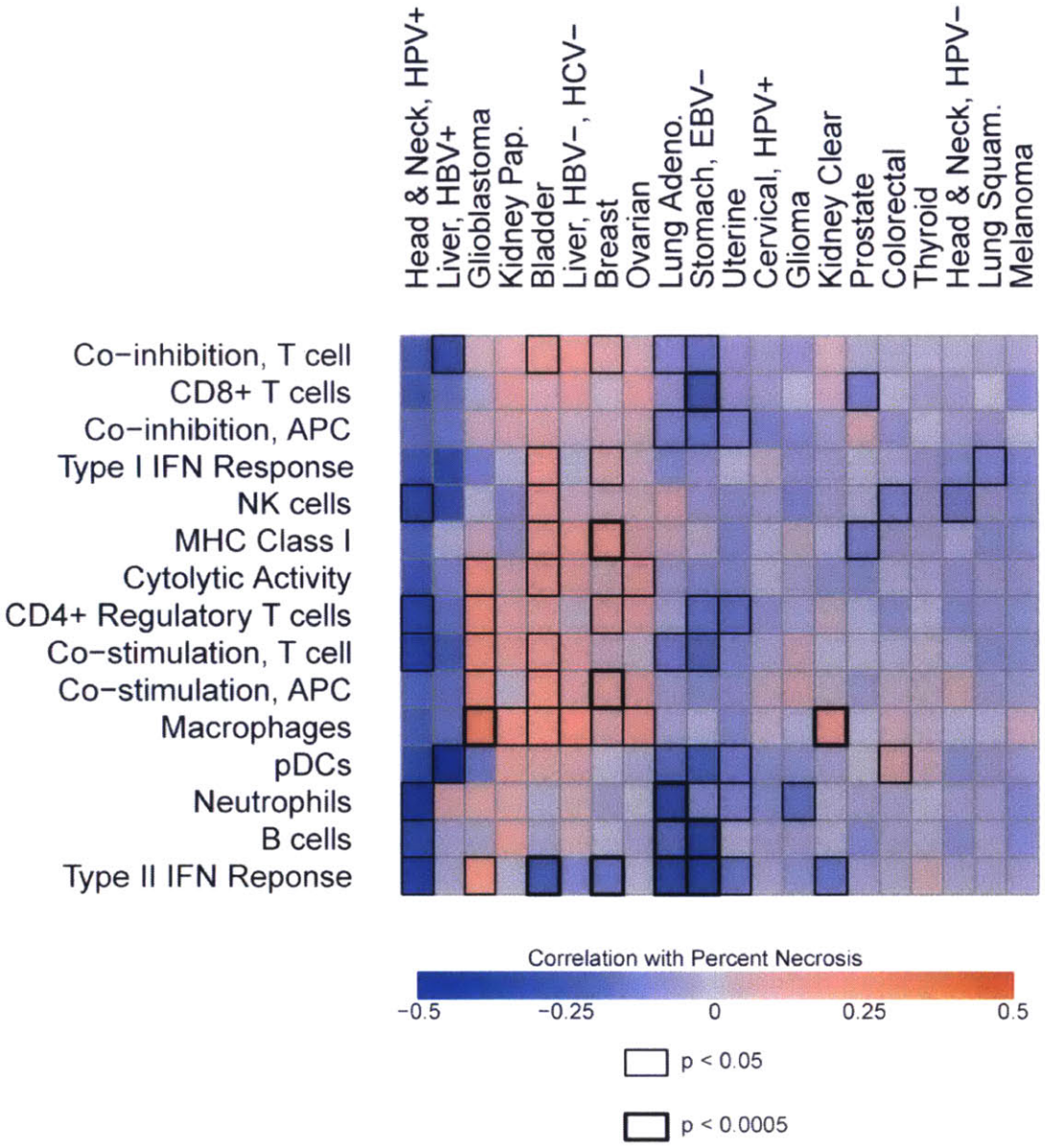


(e) Part 11 of 11. ERV expression ranges for tumor vs. normal samples, all ERVs. Semi-transparent left-right jittered points represent the expression values observed in a compendium of tissues. The 5th to 95th percentile range is highlighted in orange for tumor tissues and in green for normal tissues. The maximum 95th percentile value observed in tumors is marked with an orange horizontal line, and the corresponding maximum for normal tissues is marked with a green horizontal line. These values (marked on the right axis) were the basis for determining tumor-specific expression. ERVs designated as TSERVs are marked as such. Many ERVs, while not specific to tumors, were elevated in tumors.



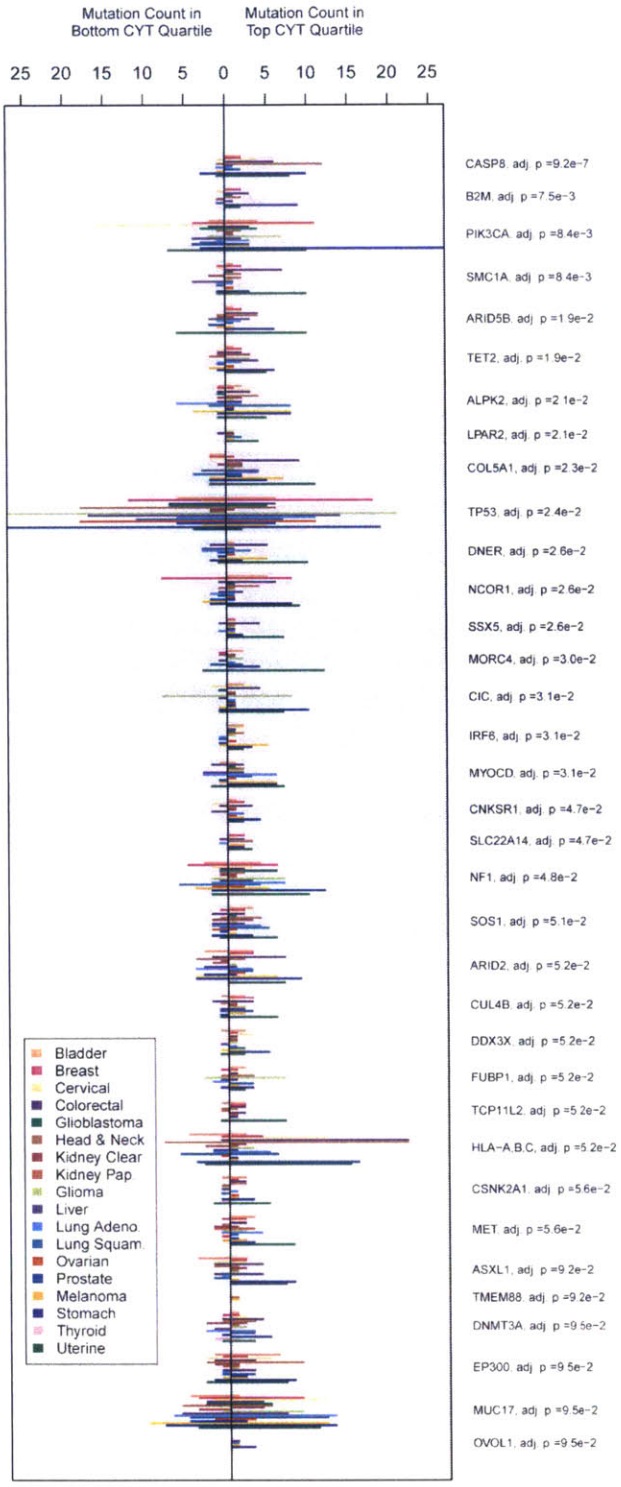
(f) Percent necrosis by tumor type. Solid bodies represent interquartile ranges and are notched by the median; lines demarcate the 5th to 95th percentile range.

Figure S4, continued



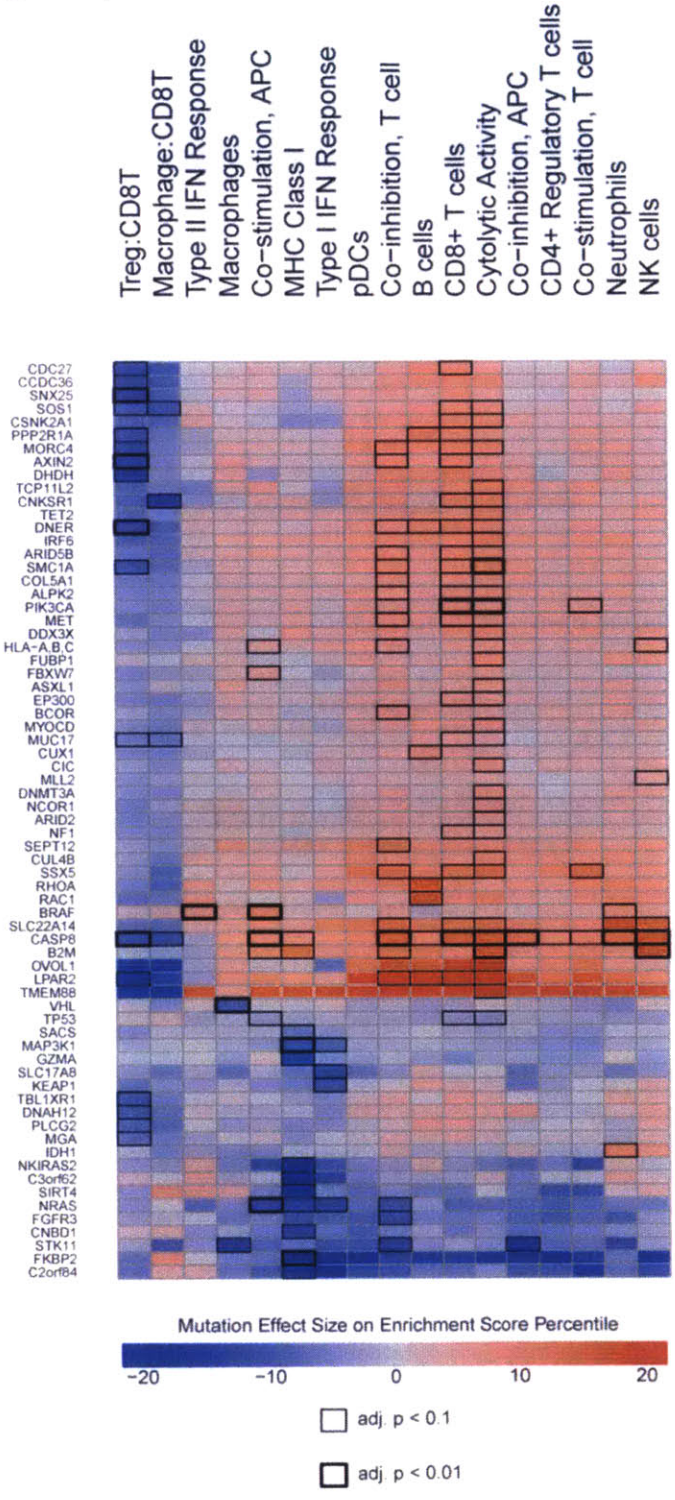
(g) Heatmap indicating association between percent necrosis and ssGSEA enrichments for markers for various immune cell types, by cancer. Colors correspond to Spearman correlations, and cell borders correspond to association p-values, as indicated in the legend.

Figure S5. Genes with enriched point mutation in high- or low-CYT tumors, related to Figure 5



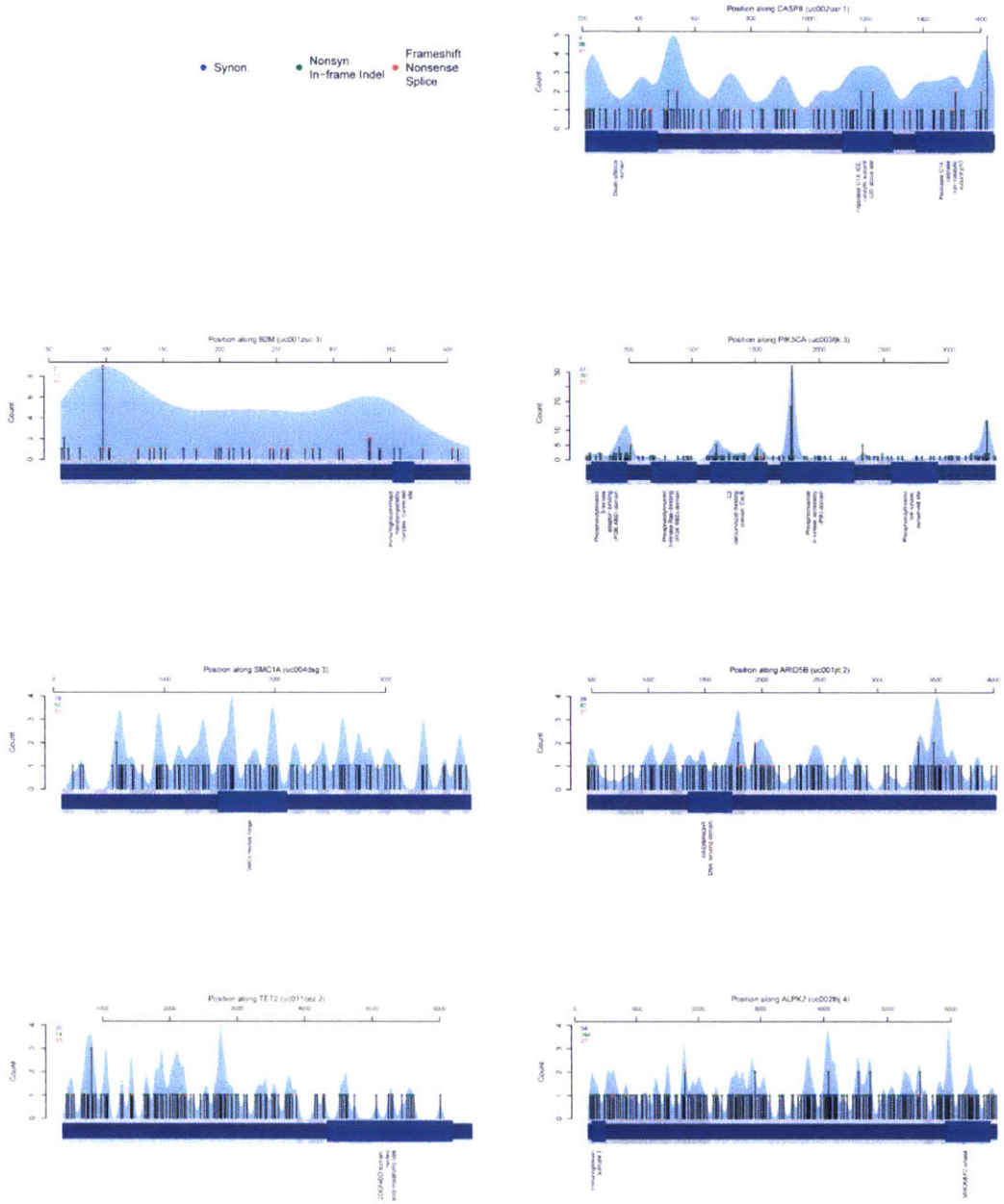
(a) Non-silent mutation counts for significant genes in high- and low-CYT tumors. High- and low-CYT tumors were defined as the top and bottom CYT quartile, respectively, per tumor type. Mutation counts in high-CYT samples point upward from the x-axis, mutation counts in low-CYT samples point downward from the x-axis. Bars are color-coded according to tumor type using the color code indicated in the legend and used elsewhere. For a given gene, tumor types exhibiting no mutations among the high-CYT or low-CYT samples are not depicted. Gene names and pan-cancer adjusted p-values (BH method) appear at the top of the figure.

Figure S5, continued



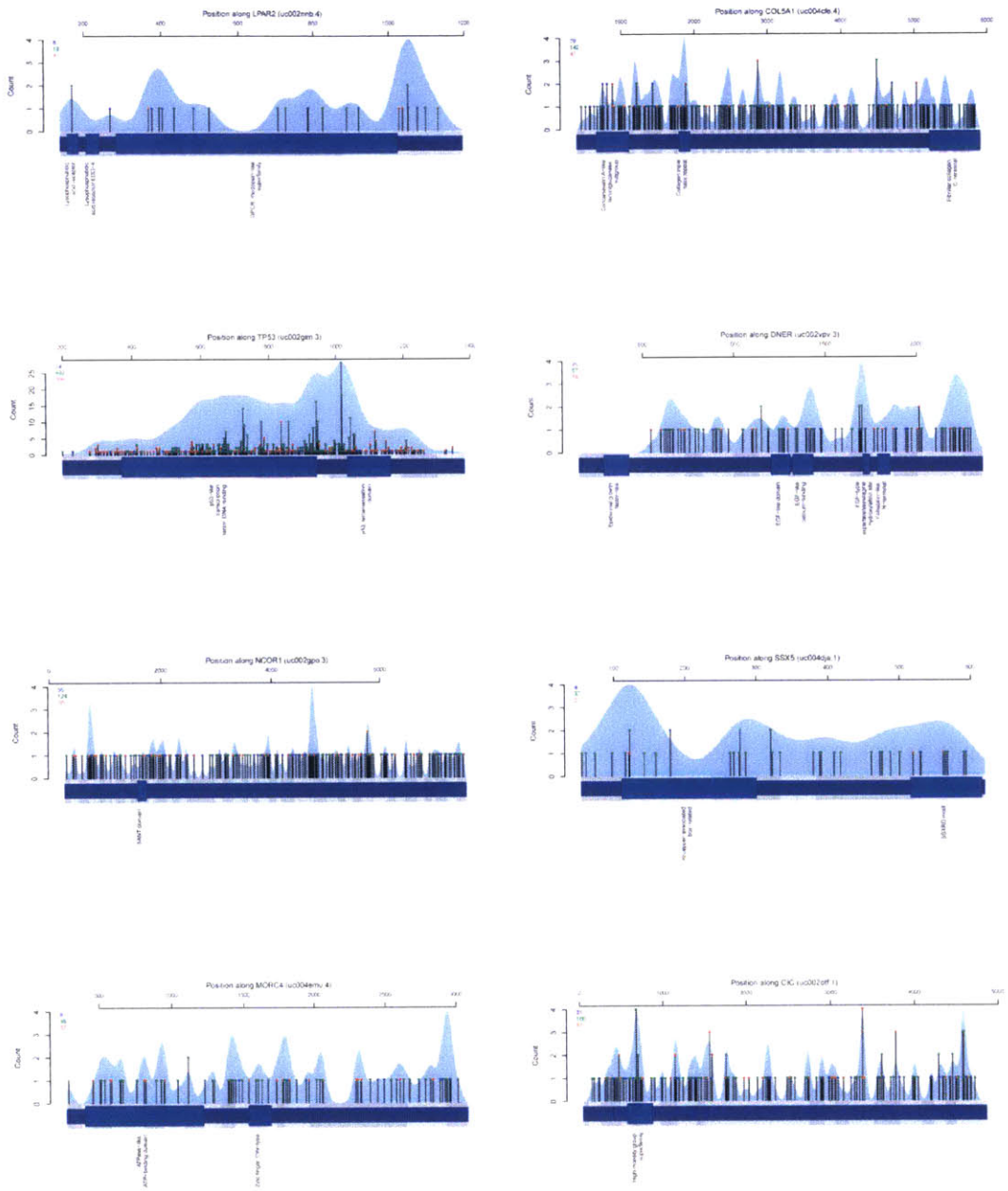
(b) Heatmap representing pan-cancer enrichments identified for other cell type signatures. Color corresponds to the effect size of non-silent mutation on rank-transformed signature expression; cell borders represent the adjusted p-value for association.

Figure S5, continued



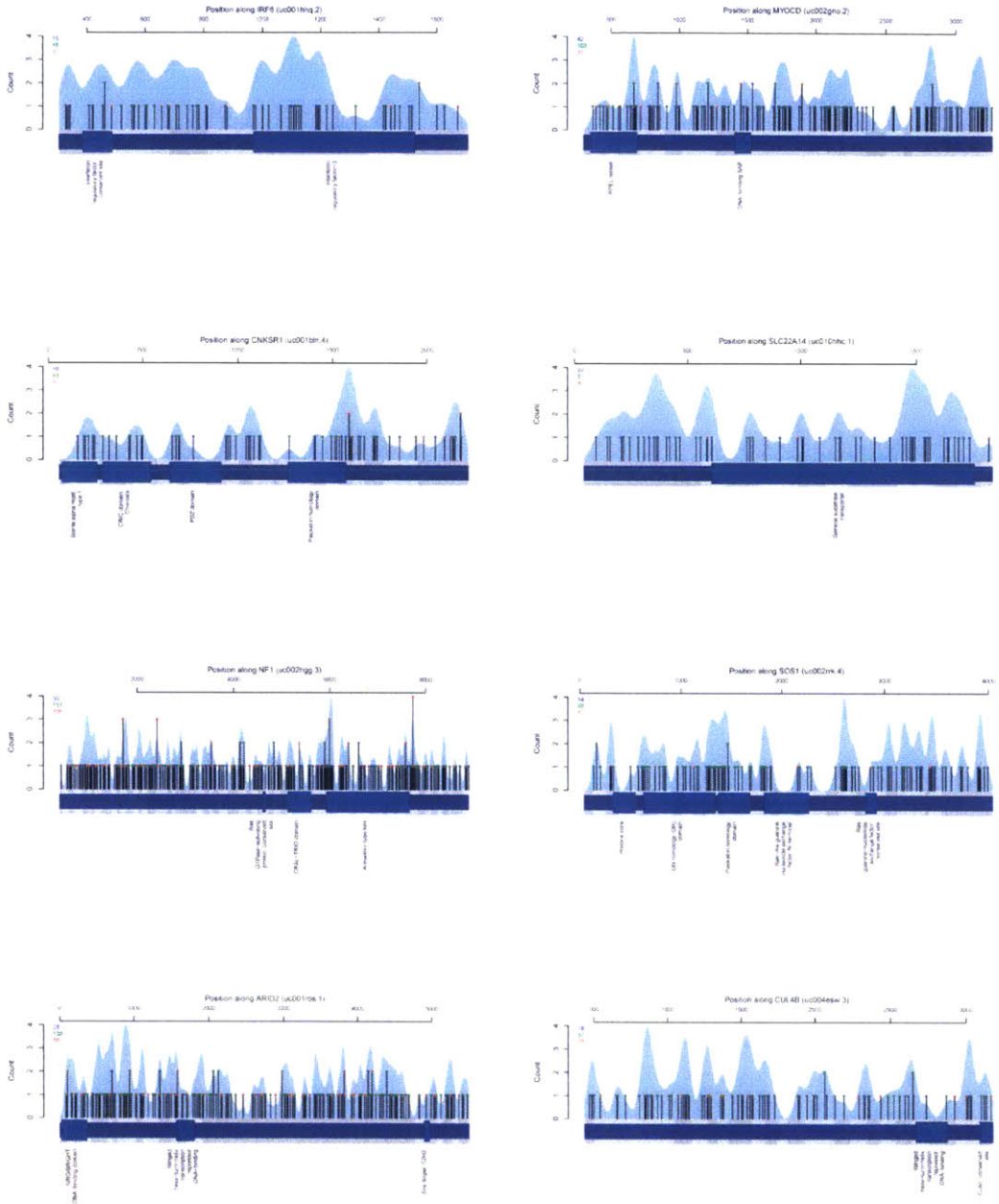
(c) Part 1 of 5. Positions of point mutation in CYT-associated genes. Mutated positions and their functional classifications are reported for each gene. Colors indicate mutation severity (synonymous, nonsynonymous, or probable loss of function), and vertical height represents event frequency. Total counts of each class of mutation appear in the upper left corner. Light blue peaks represent the relative local density of events as estimated using a smoothing bandwidth of 30 nucleotides. Sequence domains, as scored by Interpro, are represented by widened regions. Distinct exons are demarcated by alternating domains of gray and light gray.

Figure S5, continued



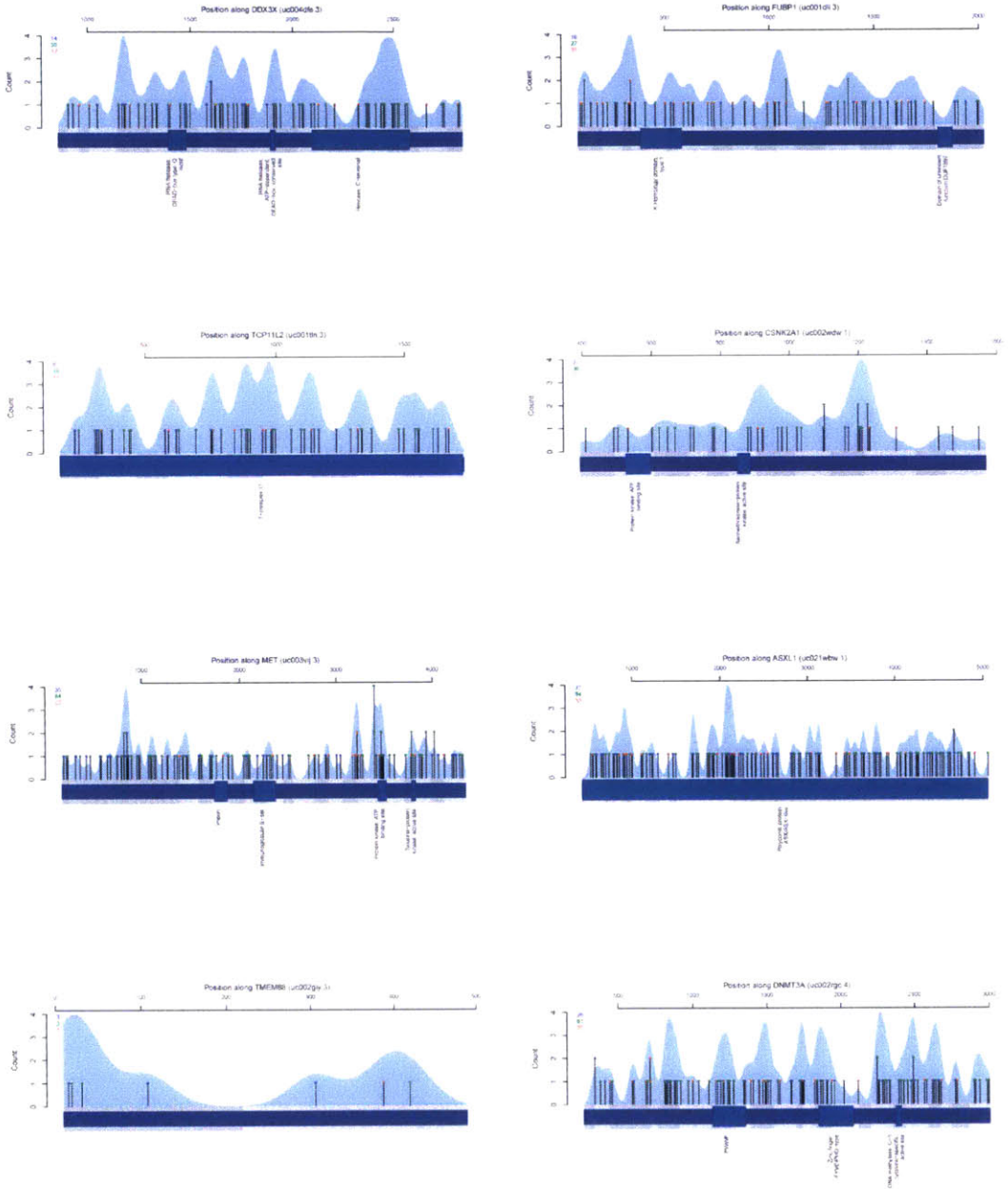
(c) Part 2 of 5. Positions of point mutation in CYT-associated genes. Mutated positions and their functional classifications are reported for each gene. Colors indicate mutation severity (synonymous, nonsynonymous, or probable loss of function), and vertical height represents event frequency. Total counts of each class of mutation appear in the upper left corner. Light blue peaks represent the relative local density of events as estimated using a smoothing bandwidth of 30 nucleotides. Sequence domains, as scored by Interpro, are represented by widened regions. Distinct exons are demarcated by alternating domains of gray and light gray.

Figure S5, continued



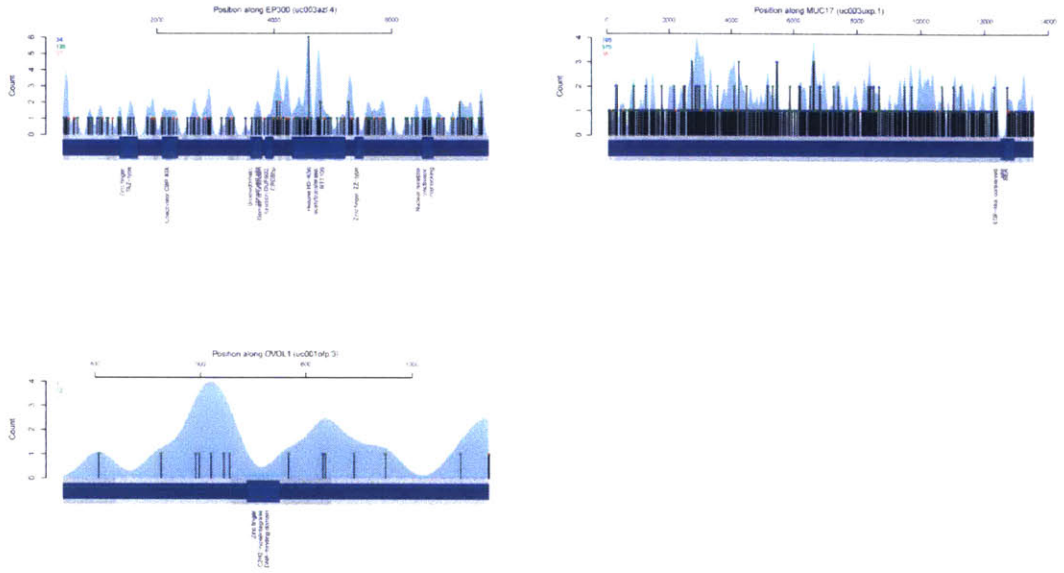
(c) Part 3 of 5. Positions of point mutation in CYT-associated genes. Mutated positions and their functional classifications are reported for each gene. Colors indicate mutation severity (synonymous, nonsynonymous, or probable loss of function), and vertical height represents event frequency. Total counts of each class of mutation appear in the upper left corner. Light blue peaks represent the relative local density of events as estimated using a smoothing bandwidth of 30 nucleotides. Sequence domains, as scored by Interpro, are represented by widened regions. Distinct exons are demarcated by alternating domains of gray and light gray.

Figure S5, continued



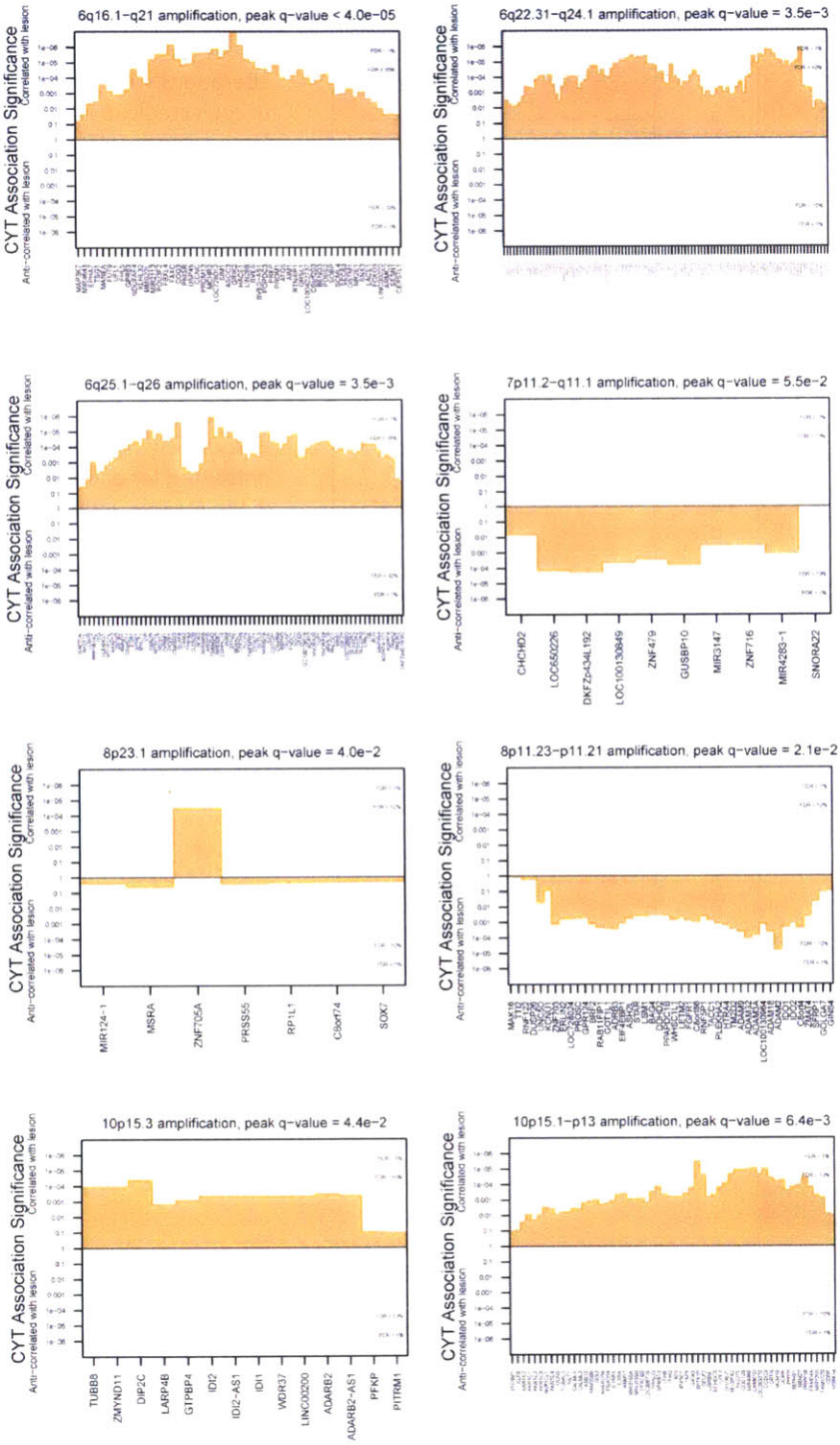
(c) Part 4 of 5. Positions of point mutation in CYT-associated genes. Mutated positions and their functional classifications are reported for each gene. Colors indicate mutation severity (synonymous, nonsynonymous, or probable loss of function), and vertical height represents event frequency. Total counts of each class of mutation appear in the upper left corner. Light blue peaks represent the relative local density of events as estimated using a smoothing bandwidth of 30 nucleotides. Sequence domains, as scored by Interpro, are represented by widened regions. Distinct exons are demarcated by alternating domains of gray and light gray.

Figure S5, continued



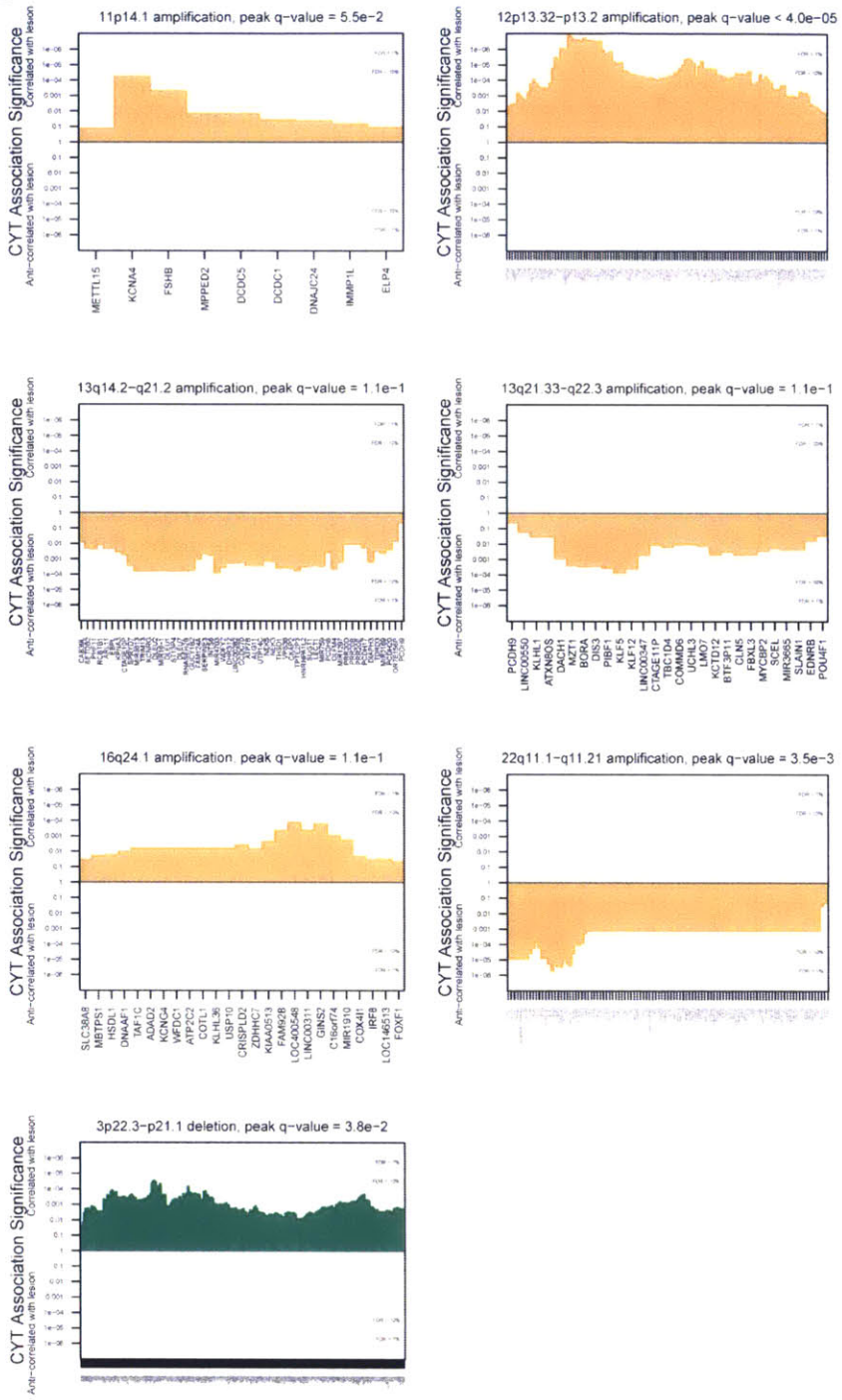
(c) Part 5 of 5. Positions of point mutation in CYT-associated genes. Mutated positions and their functional classifications are reported for each gene. Colors indicate mutation severity (synonymous, nonsynonymous, or probable loss of function), and vertical height represents event frequency. Total counts of each class of mutation appear in the upper left corner. Light blue peaks represent the relative local density of events as estimated using a smoothing bandwidth of 30 nucleotides. Sequence domains, as scored by Interpro, are represented by widened regions. Distinct exons are demarcated by alternating domains of gray and light gray.

Figure S6. Significant copy number alterations, related to Figure 6



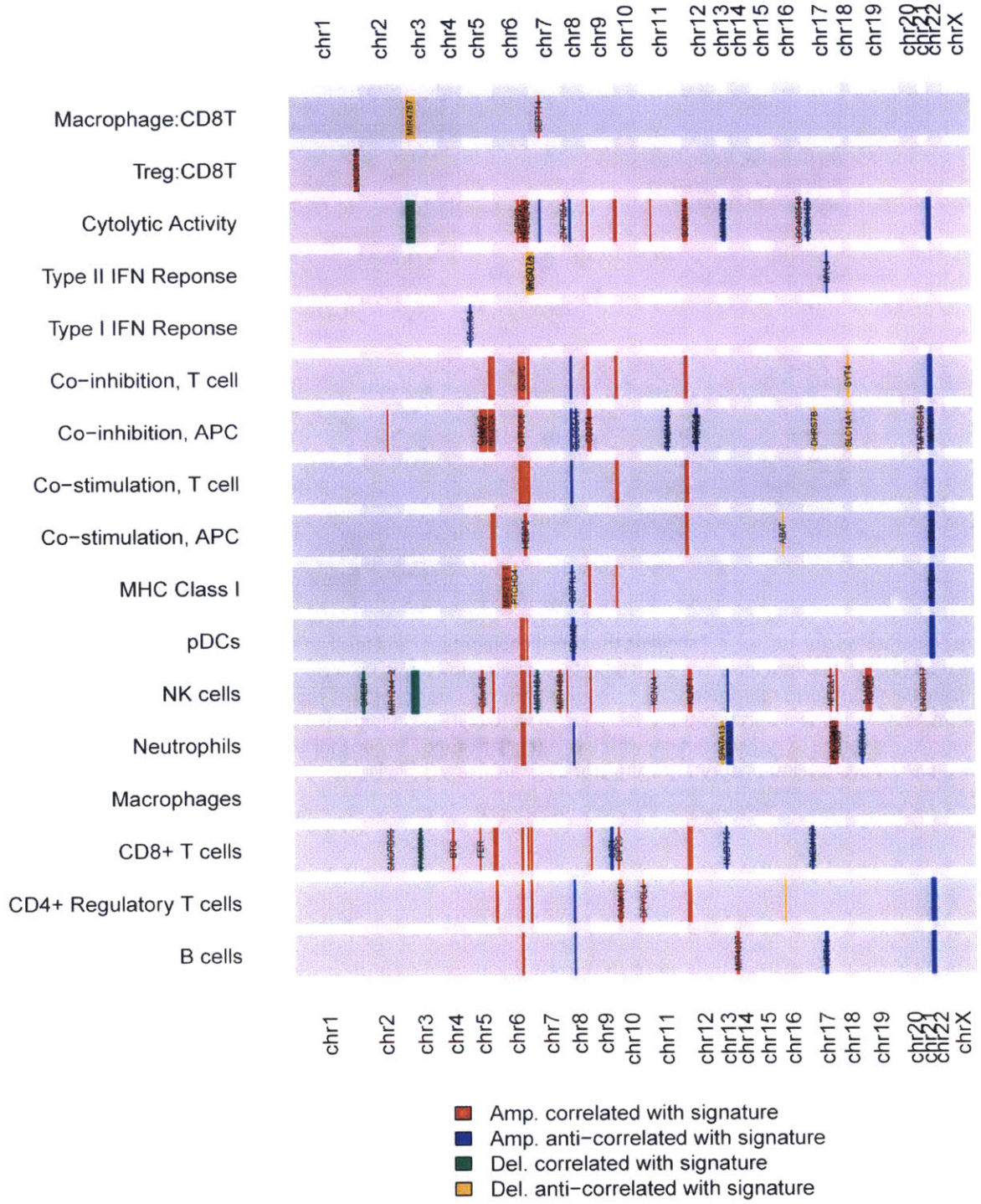
(a) Part 1 of 2. Locus zooms for copy number alterations with uncorrected p<0.05. Plots indicate CYT association for amplifications (orange) and deletions (green) in significant and near-significant regions. Upward/downward direction indicates positive/negative association of lesion with CYT. One bar is presented for each gene in the region. Dotted lines indicate the uncorrected p-values at which a 1% and 10% FDRs are established.

Figure S6, continued



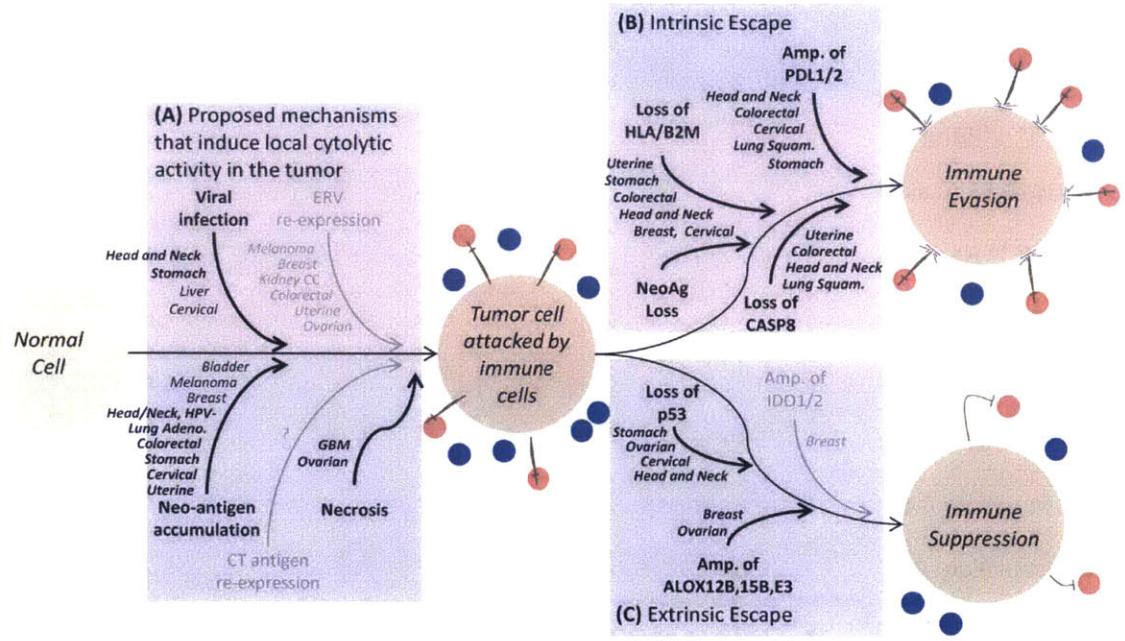
(a) Part 2 of 2. Locus zooms for copy number alterations with uncorrected p<0.05. Plots indicate CYT association for amplifications (orange) and deletions (green) in significant and near-significant regions. Upward/downward direction indicates positive/negative association of lesion with CYT. One bar is presented for each gene in the region. Dotted lines indicate the uncorrected p-values at which a 1% and 10% FDRs are established.

Figure S6, continued



(b) CNA associations with enrichments of other cell type markers. Significant regions are highlighted according to the class of lesion (amplification or deletion) and the direction of the association. Many loci were significantly associated with multiple cell type markers. For the analysis yielding the strongest signal for a locus, the unbiased peak gene is labeled.

Figure S7. Classifying tumors by their immunological properties, related to Figure 7



A more elaborate version of Figure 7 showing the specific tumor types and mechanisms implicated in **(a)** immune provocation **(b)** intrinsic immune escape and **(c)** extrinsic immune escape. Red cells represent cytolytic effectors (with spears) and blue cells represent other immune infiltrates. Large green cell represents tissue pre-transformation, and brown cells represent the tumor in various stages of immune interaction. In **(b)**, "ricochet" lines indicate resistance to cytolytic action; whereas in **(c)**, flat-capped arrows indicate suppression/removal/exclusion of cytolytic effector cells. Relates to main Figure 7.

Supplemental Tables

Since many of the supplemental tables are large, they are available online only:
http://www.sciencedirect.com/science/article/pii/S0092867414016390
(DOI:10.1016/j.cell.2014.12.033)

Table S1. Data inventory and cell type gene expression markers, related to Figure 1
(a) Inventory of TCGA tumor samples. Columns indicate patient ID, available data types, tumor type (at the TCGA project level), histological subtype, and a flag indicating whether histological subtype was considered to be known or unknown. Additional columns present the CYT calculations (log-average GZMA and PRF1 expression in TPM) and cell type/process enrichment scores (z-scored ssGSEA) for the samples. **(b)** Inventory of TCGA and GTEX samples used as normal controls. Columns indicate the patient ID, study origin, tissue of origin, and the available data (gene expression, viral expression, and ERV expression) for each sample. Additional columns present CYT and enrichment calculations, as in **(a)**. Note that z-scoring was applied to a joint data set of tumor and normal samples, so values in **(a)** and **(b)** can be directly compared. Data availability discrepancies reflect lags between the holdings of GDAC Firehose (gene expression data) and CGHub (viral and ERV quantification) and the different dates on which CGHub was accessed. **(c)** Gene symbols for best transcriptomic cell type markers according to analysis of Fantom5 CAGE data.

Table S2. Viral expression and gene set enrichment, related to Figure 2
(a) Viral titers quantified in viral reads per million mapped to the human genome (RPM). Viral titers are quantified in samples from TCGA and GTEx. Non-zero calls required at least 300nt of the viral reference genome to be covered at a read depth of at least 1. **(b)** Gene sets enriched and depleted in virally infected tumors. Gene sets enrichments are presented for tumor types exhibiting at least five cases of infection with the given virus. Enrichments are GOrilla interpretation of ranked lists of differentially expressed genes.

Table S3. Neopeptide binding predictions, related to Figure 3
Mutation-introduced novel peptides predicted to load imputed HLA alleles (tab-delimited text; HLA allele calls are masked for patient privacy). Data includes mutated genes, the corresponding wild type and mutant peptides, and affinity scores for the alleles they were predicted to bind.

Table S4. Mutation/Neopeptide counts and gene set enrichment, related to Figure 3
(a) Table presents counts of total mutations, counts of mutations predicted to yield HLA-binding neopeptides, the ratios of observed vs. expected binders per non-silent mutation, and viral infection statuses for all samples. Adjacent scatter plots show the raw data used to generate **Figure 3A** and **Figure 3B**. **(b)** Gene sets up-regulated given lower than expected rate of neo-epitopes per non-silent mutation in colorectal cancer.

Table S5. Ectopic gene expression, related to Figure 4
(a) Genes with testis-specific expression and their respective re-expressing tumor types. Genes with testis-specific expression, including both SEREX-identified and novel genes, are presented along with the rates at which they were observed at >1 TPM in each tumor type, omitting tumor types in which they were never expressed >1 TPM. Spearman correlations with CYT are also presented. **(b)** Expression values for 66 ERVs in samples from TCGA and GTEx. Expression values are expressed in reads per million mapped to the human genome (RPM). **(c)** Genes sets significantly up or down-regulated in TSERV-high tumors. For each TSERV, gene set enrichment analysis was conducted on the genes most significantly correlated and anti-correlated with the expression of the element in the tissue that demonstrated maximal expression.

Table S6. Analysis of significantly point-mutated genes, related to Figure 5
(a) Candidate genes significantly mutated in cancer. Genes identified in previous pan-cancer MutSigCV analysis at an FDR of 10% or in the current study at an FDR of 10% were included and listed noting the analyses supporting them. **(b)** Enrichment statistics for pan-cancer significant genes. Statistics are presented for the overall pan-cancer analysis and for tumor type-specific sub-analyses. Beta values reflect that the dependent variable (CYT) was transformed to rank values scaled from 0 to 100. The table also presents counts of mutated and total samples per tumor type / gene. Note that HLA mutations were called for a larger number of samples than general mutations. **(c)** HLA mutational status for available TCGA tumor samples. Coding mutations for HLA-A, B, and C are presented. These were called through application of the POLYSOLVER algorithm rather than by the Firehose/Synapse mutation calling pipelines. **(d)** Mutations associated with viral infection. Table indicates odds ratios, Fisher exact test p-values and counts for virally infected *vs.* mutant samples. **(e)** Mutations associated with microsatellite instability (MSI) in colorectal cancer (MSI-high vs. MSI-low and microsatellite stable) at 10% FDR. Table indicates odds ratios, Fisher exact test p-values and counts for virally infected *vs.* mutant samples, and BH-correctef p-values. **(f)** Gene set enrichments associated with other significantly mutated immune genes. Some with definitive roles in immunity were identified as significantly mutated in cancer but did not show mutational association with CYT. For each of these genes, the gene expression correlates of its mutation were analyzed in the tumor type exhibiting the strongest MutSigCV p-value. GOrilla results returning immunologically relevant enrichments are presented.

Table S7. CYT association statistics for significant copy number alterations (CNAs), related to Figure 6

Statistics are presented for the overall pancancer analysis and for tumor type-specific sub-analyses. Beta values reflect that the dependent variable (CYT) was transformed to rank values scaled from 0 to 100 and the scaling of the CNA events per sample to have a median amplitude of 1. The table also presents counts of mutated and total samples per tumor type / gene; mutant condition was based on nonzero GISTIC score.

Table S8. A summary of immunological properties per tumor type, related to Figure 7

Immune attributes and associations (rows) and the corresponding tumor types in which they manifest (columns) with a "1" (highlighted in red) to mark positive instances of the trend or attribute and a "-1" (highlighted in blue) to mark cases in which the opposite trend was observed.

Dynamic profiling of the protein life cycle in response to pathogens

Authors and Affiliations

Marko Jovanovic[1,*], Michael S. Rooney[1,2,*], Philipp Mertins[1], Dariusz Przybylski[1], Nicolas Chevrier[1,3], Rahul Satija[1], Edwin H. Rodriguez[4], Alexander P. Fields[4], Schraga Schwartz[1], Raktima Raychowdhury[1], Maxwell R. Mumbach[1], Thomas Eisenhaure[1,5], Michal Rabani[1], Dave Gennert[1], Diana Lu[1], Toni Delorey[1], Jonathan S. Weissman[4,6], Steven A. Carr[1], Nir Hacohen[1,5,7], Aviv Regev[1,6,8]

[1]Broad Institute of MIT and Harvard, Cambridge, MA 02142
[2]Harvard/MIT Division of Health Sciences and Technology, Cambridge, MA 02141
[3]Harvard FAS Center for Systems Biology
[4]Department of Cellular and Molecular Pharmacology, California Institute for Quantitative Biomedical Research, University of California, San Francisco, San Francisco, CA 94158
[5]Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital
[6]Howard Hughes Medical Institute
[7]Harvard Medical School
[8]Department of Biology, Massachusetts Institute of Technology
*contributed equally

Author Contributions

M.J., M.S.R, A.R, N.H., and S.A.C. conceived of the project, designed the experiments, and interpreted the results. M.J. conducted the experiments with assistance from P.M., N.C., R.S., S.S., R.R., M.R.M., T.E., D.G., D.L., and T.D. M.S.R. developed statistical models for analysis with assistance from D.P., R.S., and M.R. E.H.R. and A.P.F. contributed ribosomal profiling measurements under the guidance of J.S.W.

Manuscript Acknowledgments

We thank members of the Regev, Hacohen, and Carr groups, as well as G. Brar, N. Slavov, and E. Airoldi for constant input and discussions. We thank L. Gaffney for help with the figures and K. Lage and A. Kashani for help with some of the analyses. This work was supported by National Human Genome Research Institute Centers of Excellence in Genomics Science P50 HG006193 (A.R, N.H., S.A.C.) and Broad Institute Funds. A.R. was supported by an NIH Pioneer Award, the Klarman Cell Observatory, and HHMI. M.J. was supported by fellowships of the Swiss National Science Foundation for advanced researchers (SNF) and the Marie Sklodowska-Curie International Outgoing Fellowships. M.S.R. was supported by the NIH Training Program in Bioinformatics and Integrative Genomics training grant. S.S. was supported by a Rothschild Fellowship, a European Molecular Biology Organization fellowship, and Human Frontier Science Program fellowships. E.H.R. was supported by the Howard Hughes Medical Institute Gilliam Fellowship for Advanced Study.

Introduction

With greater than 55,000 expression profiling studies hosted on the Gene Expression Omnibus (GEO) web server (http://www.ncbi.nlm.nih.gov/geo), it is clear that expression profiling of mRNA has become a ubiquitous scientific tool for understanding the multidimensional processes of health and disease. Using mRNA levels as a readout on cellular state, these studies have presumed that mRNA levels highly correlate with protein abundance – the actual effector molecules that drive cellular processes. However, for greater than a decade, a series of scientific reports have argued that transcriptional data is in fact highly insufficient for inferring protein levels and that the post-transcriptional processes of protein translation and degradation determine a larger share of the variance in protein levels. If this is true, it argues for greatly expanding research on post-transcriptional rather than transcriptional processes that influence protein expression.

Post-transcriptional regulation mainly operates at the level of changing per-mRNA translation rates or changing the rates at which proteins are degraded. Part of my graduate thesis work was devoted to quantifying the role of transcription versus post-transcriptional processes in determining the abundance of proteins. A rigorous solution to this problem would address a basic cell biology question and have implications for how researchers should interpret mRNA levels as reflecting protein abundance.

Mechanisms of protein synthesis

In eukaryotes, the process of translation begins with the ribosome recognizing the 5' cap structure (a prepended, 5'-to-5' linked guanine) of the incipient mRNA. The first form of regulation that occurs is the recognition of the first AUG, which is enhanced if it falls within a Kozak motif (aAaAaAATGTCt in yeast, gcc[a/g]ccATGG in vertebrates) (Hamilton et al., 1987; Kozak, 1987). Transcripts with stronger Kozak motifs are more rapidly translated (Kozak, 1986). Interestingly, initiation from the first AUG is not a firm requirement, with some translation beginning at CUGs (Ingolia et al., 2011) or at internal downstream ribosomal entry cites (IRESs) (Filbin and Kieft, 2009; Graber and Holcik, 2007). Additionally, the main open reading frame of the transcript may come after one or more short upstream open reading frames (uORFs), which appear to compete with the main open reading frame (Calvo et al., 2009). These alternative forms of translation initiation are thought to contribute additional regulation to the rate at which translation occurs. As the ribosome elongates the nascent polypeptide, additional forms of regulation come into play. For one, tRNAs species are not all equally abundant meaning that codon usage in the transcript can affect the rate at which cognate tRNAs are recruited. This appears to be an especially important constraint in yeast, in which highly expressed genes have evolved to have a much higher level of codon usage "efficiency" (Bennetzen and Hall, 1982). RNA secondary structures and protein chaperone elongation factors (notably, eEF1A, eEF1B and eEF2) contribute to elongation rates by causing and resolving translational stalls, respectively (Browne and Proud, 2002). The total length of the polyA tail as well as the presence of RNA-binding proteins known to interact with the polyA tail (possibly by promoting more optimal secondary structure) are also known to affect translation rate (Lackner et al.,

2007; Preiss and Hentze, 1998; Schwartz and Parker, 1999). One especially well-studied mechanism of post-transcriptional control is microRNAs. These small RNAs bind the 3' UTRs of mRNAs hosting cognate seed sites, resulting in a moderate decrease in the translation of the target mRNA. However, it is not known how much of this effect is mediated by reduced stability of the mRNA target (an effect present in measured mRNA levels) or by direct interaction with the translational machinery (an effect measured RNA levels would not account for) (Huntzinger and Izaurralde, 2011).

Mechanisms of protein degradation

The degradation of proteins is mediated by two major systems: lysosomal degradation and ubiquitin mediated proteolysis. Lysosomal degradation involves proteins being engulfed into a lysosomal vacuole where they are exposed to degradative enzymes. Cytosolic proteins assist in guiding proteins to the lysosome and may recognize specific protein motifs (*e.g.* KFERQ; (Dice, 1990)). Ubiquitination is a process by which a ubiquitin molecule (a 76AA protein) is bound to a lysine residue on a substrate protein through the action of one of three ubiquitin ligases (E1, E2, and E3) (Joazeiro and Weissman, 2000). Ubiquitins can form chains, with each new ubiquitin binding to one of 7 possible lysines in the previous ubiquitin. Depending on which lysine is chained on, the protein may be targeted for degradation in the proteosome (as with chaining on K48; (Hicke, 2001)) or re-routed to some other process (as with chaining on K63; (Miranda and Sorkin, 2007)). In addition to lysosomal degradation and ubiquitination, several other mechanisms are known to affect protein degradation rates. A special class of proteins called N-

recognins recognize N-terminal patterns on proteins (in mammals, for instance, arginylated N-terminal Asn and Glu residues) and tag them for degradation, a process which appears to be largely dependent on whether the N-end is buried or solute exposed (Tasaki et al., 2012). In addition, PEST sequences (proline-glutamic acid-serine-threonine) have been found to be enriched in proteins with a rapid half-life (Rogers et al., 1986), though it is unclear whether the effect is mediated through the proteasome or the non-lysosomal cysteine protease calpain (Reverte et al., 2001; Shumway et al., 1999). Finally, proteins with large unstructured domains are also known to exhibit reduced half-lives (Gsponer et al., 2008), though the exact mechanism of this is not fully understood either.

Quantifying contributions of transcriptional and post-transcriptional processes in determining protein abundance

Given all these mechanisms, it is prudent to systematically assess their overall contribution to homeostatic protein levels as well as to protein level fold changes induced by stimulus or differentiation. The answers to these questions bear heavily on both the appropriateness of RNA expression profiling as a window into biological state and on the relative effort future research should invest in understanding transcriptional regulation *vs.* the many post-transcriptional processes listed above. While many groups have tackled this question and declared a major contribution from post-transcriptional regulation, the specific estimates have varied widely by methodology and experimental platform leading to a level of uncertainty (Maier et al., 2009).

For the most part, quantification of post-transcriptional contribution has been approached using coefficients of determination, *i.e.* R^2. R^2 measures the fraction of the variance in a dependent variable (in this case, protein levels across the genome) that can be explained using a model based on one or more predictor variables. Alternatively, it can be thought of the square of the correlation (r) between the dependent variable and the model predictions. Those arguing for the importance of post-transcriptional processes have pointed to the low R^2 of models in which RNA level is the only predictor variable or to the increase in R^2 that can be achieved by adding per-gene translation and degradation rates into such models (Gygi et al., 1999; Schwanhäusser et al., 2011). (Others have also used this framework for determining the respective contributions of RNA transcription and degradation rates to total RNA levels (Rabani et al., 2014), but this is not discussed here.) Though this formulation seems reasonably concise, there are certain operating assumptions that are important to consider from the outset. First of all, the consensus of the field has settled on log-protein expression (rather than absolute protein expression) as the dependent variable of interest, probably motivated by the desire to have a result that is broadly representative of the genome and not driven by several abundantly expressed genes (Futcher et al., 1999). Second, analyses have limited their scope to genes with non-zero protein expression. Naturally, including transcriptionally silent genes (necessarily absent from the proteome) would greatly increase the variance explained by mRNA levels (Li et al., 2014); however, the field has generally adopted the former approach, which is probably more representative of the entire RNA-to-protein life cycle. Third, there is the question of generalizability across organisms. Many authors in this field have framed their results as being broadly extendable across eukaryotes; however, in actuality, the results in the most studied

system, yeast, have been variable enough that organism-level granularity is likely not yet within reach (Futcher et al., 1999; Gygi et al., 1999; Lee et al., 2011; Vogel et al., 2011). Finally and most importantly, there is role of error. The measure of interest relates *true* RNA levels, *true* translation rates, and *true* degradation rates to *true* protein levels. Once stochastic error (error that varies across replicates) or systematic error (platform-dependent error) is introduced into the estimation of these variables, the derived R^2 will necessarily be lower unless addition statistical corrections are made (Li et al., 2014). Conversely, if model predictions and measured protein levels share correlated errors (as might occur if translation rates and protein levels are measured using the same experimental platform), the derived R^2 may be inflated (Schrimpf et al., 2009).

Efforts to derive R^2 statistics for post-transcriptional processes have been developed for both homeostatic (or "resting", or "baseline") protein levels as well as for protein level changes (or "dynamics") in the course of changes in cellular state. Both are described here with an initial focus on baseline protein levels.

The determinants of homeostatic protein levels

In 1998, Gygi *et al.* published the first large-scale analysis addressing baseline protein expression, measuring over 150 yeast proteins by 2D radio-labeled gel electrophoresis and corresponding mRNAs by serial analysis of gene expression (SAGE) lookup tables (Gygi et al., 1999). The research group reported an R^2 of just 0.36 (data not log-transformed). They interpreted this to mean that post-transcriptional processes were more important in determining protein levels (explaining the remaining 64% of protein variance), declaring that

there was "no predictive correlation" between steady-state mRNA levels and protein levels. Shortly thereafter, Futcher *et al.*, published a study using a nearly identical experimental approach but arriving at a much higher R^2 (0.57) and a dramatically different conclusion (Futcher et al., 1999). This second group pointed out that much of the discrepancy derived from differences in statistical approaches. First of all, Futcher *et al.* used improved mRNA data by integrating SAGE data and microarray data and adjusting for systematic errors. Second, they log-transformed before computing the correlation (the first group to do so), reducing sensitivity to outliers (in fact, Gygi's figure of 0.36 resulted only after dropping the top 30% (!) of proteins; with these genes included, a drastically different R^2 of 0.87 was obtained). Finally, they interpreted the missing variance more carefully, acknowledging that much of it may relate to measurement error in the either the mRNA or protein levels rather than post-transcriptional processes. Even though the study by Futcher *et al.* was more carefully conducted, Gygi's work became the standard of the field, with four times the number of citations during the following decade (2270 *vs.* 567 (scholar.google.com)).

In 2009, a new generation of studies were published that capitalized on improvements in non-targeted "shotgun" proteomics (Nesvizhskii et al., 2007). First, Schrimpf *et al.* explored the relationship between mRNA levels and protein levels in higher order eukaryotes employing liquid chromatography-electrospray ionization-tandem mass spectrometry (on a tryptic digest) to measure protein levels for >5000 genes (and relying on previously published microarray and SAGE estimates for mRNA levels) (Schrimpf et al., 2009). Results yielded poor correspondence between measured transcriptomic and proteomic values, R^2=0.35 and R^2=0.44 for worm and fly, respectively. Meanwhile, there was a reasonably strong correlation between the measured

protein levels of the two species (r=0.79) and a comparatively weaker correlation in the inter-species mRNA levels (r=0.47, as measured by microarray). Schrimpf *et al.* interpreted this to mean that protein levels are more tightly conserved between species with post-transcriptional processes compensating for drift in transcript levels. However, the intra-species correlations of the microarray data and SAGE data were quite poor (r=0.53 and r=0.46 for worm and fly, respectively), hinting at a high level of measurement error in one or both of the platforms. Unfortunately, this paper did not clarify how much of the unexplained variance in the mRNA-protein relationships could be attributed to experimental error. Thus, it fell in the footsteps of Gygi, reaching bold conclusions regarding the role of post-transcriptional processes without firm evidence the finding was more than the result of measurement error.

Shortly thereafter, Vogel *et al.* tackled the same question in a human meduloblastoma cell line, observing an even weaker mRNA-protein relationship, R^2=0.29 (Vogel et al., 2010). Improving on previous work, Vogel attempted to account for measurement error using a method known as Spearman correction (not related to Spearman rank correlation). Spearman correction is a statistical technique to measure the true correlation between two variables x and y that are measured with error as x_e and y_e (Spearman, 1904). Given replicate measurements of x_e and y_e, the true correlation of x and y is estimated as:

$$\frac{G(Corr(x_{e1}, y_{e1}) \cdot Corr(x_{e1}, y_{e2}) \cdot Corr(x_{e2}, y_{e1}) \cdot Corr(x_{e2}, y_{e2}))}{G\big(Corr(x_{e1}, x_{e2}) \cdot Corr(y_{e1}, y_{e2})\big)}$$

where x_{e1}, x_{e2}, y_{e1}, and y_{e2} represent replicate measurements and where $G()$ is the geometric mean function. Thus, the denominator reflects a correction for data reliability. The approach

only slightly increased Vogel's estimate, revised from R^2=0.29 to R^2=0.32, from which Vogel concluded that the role of experimental error was small. However, the Spearman correction assumes that errors between replicates measures are statistically independent (Spearman, 1904). Therefore, Vogel's approach helped clarify the contribution of stochastic error but left the role of systematic error unaddressed. This represented a significant omission given the known biases of microarray estimates (Yang and Speed, 2002) and the wide variance in the ionization and detection efficiencies of tryptic peptides in mass spectrometers (Steen and Mann, 2004).

A recurring problem with these analyses was the difficulty of obtaining direct measurements of post-transcriptional processes, which had forced authors to draw inference from the variance unexplained by the transcriptome. In lieu of direct measurements, some researchers sought to leverage sequence features for additional insight into post-transcriptional processes. Drawing on the established relationship between gene expression level and codon usage in yeast, Wu *et al.* hypothesized that codon efficiency would be associated with higher protein levels even after correction for mRNA expression (Wu et al., 2008). Indeed, codon bias could explain an additional 9% of protein level variance (though Wu did not rule out residual confounding with true mRNA levels as a possible explanation of this result). Soon thereafter, Vogel adopted a similar approach in analyzing the human meduloblastoma cell line (Vogel et al., 2010). Controlling for microarray-derived mRNA levels (using a technique called partial Spearman correlation), she tested ~200 mRNA sequence feature variables for their association with

protein level and found that they could collectively explain over 67% percent of protein level variance. It was unclear how much of this might reflect over-fitting (just 1000 genes were analyzed, and the explanatory power dropped to "30-60%" upon cross-validation). Another unaddressed concern was whether this approach might inadvertently mine for systematic errors in microarray expression estimates, which are also likely to be related to sequence features.

Ultimately, the field was interested in making the analyses more direct. A significant step forward was Selbach and colleagues' 2011 study of mouse fibroblasts (Schwanhäusser et al., 2011). This study employed a relatively new quantitation technique called SILAC (stable isotope labeling by amino acids in cell culture (Ong et al., 2002)), which could be employed in a pulse-chase strategy, wherein cells were grown in Light-isotope (L) media and switched to Heavy-isotope (H) at t=0. At subsequent time points, absolute protein level (as estimate by the sum of peptide intensities divided by the count of theoretically observable peptides, known as intensity based absolute quantification, or iBAQ), in conjunction with the ratio of H- to L-isotope peptides from each protein, allowed estimation of the amount of pre-existing protein (produced before t=0) and newly produced protein (produced after t=0) for each gene at each time point. By assuming steady state $\frac{dP}{dt} = 0 = k_{tra} \cdot R - k_{deg} \cdot P$, Selbach and colleagues could experimentally estimate the production term $k_{tra} \cdot R$, RNA level R, and protein level P, thereby yielding both k_{tra} and k_{deg}. Improving upon Vogel, Selbach took steps to directly estimate stochastic and systematic (platform-dependent) error in RNA expression data by measuring the correlation between RNA-Seq and Nanostring expression estimates for 79 genes (the latter platform considered a "gold standard" for RNA level). This yielded a correlation of

0.79, which was described as "high" but not further considered in the analysis (as through Spearman correction). Likewise, observed discrepancies between mass spectrometry-based quantitations and the levels of protein spike-ins were not mathematically included in the analysis approach. Most crucially, Selbach and colleagues did not address the fact that their translation rate estimates and total protein abundance estimates were obtained from the same experimental platform and thus subject to correlated errors. Ultimately, Selbach ascribed 40%, 55%, and 5% to RNA levels, per-mRNA translation rates, and protein degradation rates, respectively, for their percent contribution to protein level, declaring translational control to be the dominant regulatory factor.

In response to this analysis, Biggin and colleagues wrote a paper charging that Selbach had significantly underestimated the contribution of RNA to protein levels because of deficient statistical techniques (Li et al., 2014). In many ways this mirrored the earlier squabble between Gygi and Futcher. Biggin first pointed out that Selbach and colleagues had not incorporated the results of the RNA-Seq *vs.* Nanostring comparison into his estimates. By making this addition, Biggin was able to boost the explanatory contribution of RNA from 40% to 56%. While this left 44% to potentially be explained by translation and degradation, Biggin predicted that the derived contribution of RNA would be much higher if systematic errors in the protein measurements could also be accounted for (though he ultimately decided that the data from Selbach's protein standards was not sufficient for making such a correction). Therefore, he attempted to make direct measurements of translation rate using "ribosomal profiling." Developed by Ingola *et al.*, this modified form of RNA-Seq uses a nuclease to digest RNA that is not shielded by ribosome engagement, yielding a reduced library of ~28mer protected

fragments (Ingolia et al., 2009). Dividing ribosomal profiling gene expression estimates by standard RNA-Seq expression values yields per-gene "translational efficiencies" predicted to scale with per-mRNA translation rates. Incorporating these values produced a drastically different result from Selbach and colleagues, ascribing just 8% to translation rates, another 8% to degradation rates and a whopping 84% to RNA levels. Given the 10-years of publications suggesting dominance for translation rates, these results put the field into a state of confusion. This was exacerbated by the relatively complex statistical approach Biggin employed as well as disagreement about how faithfully the new ribosomal profiling method actually captured per-mRNA translation rates.

The determinants of protein level dynamics

With the question of protein regulation at baseline not resolved, another pressing question is how much each of these mechanisms (RNA level, translation rate, and degradation rate) contributes to changes in protein level when cells transition to a new cellular state. This is important because transcriptomics has become a preferred descriptor of cellular state as it relates to differentiation, response to stress, and health *vs.* disease. If RNA levels only play a minor role in determining protein levels, say ~35% as some have argued for the baseline condition, then the transcriptome is a poor proxy for cellular state and transcriptomic analysis is missing important post-transcriptional processes that regulate a larger share that state.

One of the first groups to tackle this question systematically was Mann and colleagues (de Godoy et al., 2008), who compared haploid vs. diploid yeast, measuring transcript level ratios using microarrays and protein level ratios using SILAC (rather than pulse-chase set up, as

employed by Selbach (Schwanhäusser et al., 2011), L-labeled and H-labeled yeast were raised in parallel and their proteins mixed just before quantitation). This analysis produced a shockingly low concordance between the RNA and protein fold changes – just R^2=0.05. Though, notably, when the authors filtered out low expression genes, this jumped to 0.21, and when they focused on just one highly induced pathway comprising 18 genes, the number further climbed to 0.46. The authors acknowledged that stochastic and/or systematic errors in the microarray data may be driving these results though they did not attempt to quantify this statistically. A good statistical adjustment was within reach because fold changes should largely normalize out platform-dependent errors and application of Spearman correction to replicates could have adjusted for stochastic errors.

The next year, Lu *et al.* asked the RNA dependence question again, this time in the context of mouse embryonic stem cells (ESCs) differentiating over a five-day period (Lu et al., 2009). While they did not employ an R^2 approach, the authors noted that of the genes with significant changes in protein levels, only 43-52% showed concordant changes in RNA levels, concluding that translational and post-translational regulatory mechanisms have important roles in ESC fate decisions. However, shortly thereafter, Munoz *et al.* explored a similar model system, the transition of human fibroblasts into induced pluripotent stem cells (iPSCs), and derived a much higher coefficient of variation, R^2=0.49 (Munoz et al., 2011). These discrepancies were mirrored in a series of papers analyzing dynamic processes in yeast. The first, by Fournier *et al.,* profiled the yeast response to rapamycin over a 6-hour time course (Fournier et al., 2010). Considering the fold change (*vs.* baseline) at different time points, Fournier *et al.* found the best correspondence between the 2-hour mRNA change and the 6-hour protein change, R^2=0.36.

Later, Vogel *et al.* explored the yeast response to oxidative stress over a 2-hour time course (Vogel et al., 2011). Though Vogel did not formalize the explanatory capacity of mRNA fold change in terms of a coefficient of variation, she showed recurrent qualitative differences in the protein and mRNA expression level trajectories for many genes.

The third paper in the series (Lee et al., 2011), which considered the yeast response to osmotic stress, improved upon previous work significantly by making the important realization that a mere fold change *vs.* fold change analysis was not sufficient for understanding the contribution of mRNA to the system. For steady-state systems, such as the two resting yeast populations studied by Mann and colleagues, a comparison of fold changes is sufficient: translation rate and degradation rate remaining constant, a doubling of mRNA level should drive a doubling of protein level ($P_{steady} = \frac{R \cdot k_{tra}}{k_{deg}}$). However, this is not the case in a system with changing protein levels, for which constant translation and degradation rates do not guarantee identical RNA and protein level time course trajectories (for instance, if an mRNA drops precipitously, a very stable protein may not decrease appreciably during the monitoring period, and this discrepancy does not implicate a change in protein translation or degradation rate). Realizing this, Lee *et al.* developed a dynamic model for protein production and degradation using their observed RNA level changes and previously published (stationary) translation and degradation rates (Belle et al., 2006; Ghaemmaghami et al., 2003). Using this approach, they were able to explain 77% of the variance in protein up-regulation with mRNA changes – much greater than the previous studies reported. This allowed the authors to focus on a more narrow range of proteins for

which the data suggested additional, non-transcriptional forms of regulation, though they did not directly quantify the contribution of transcriptional regulation overall.

The need for a more systematic experimental and analytic approach

In considering these studies of baseline and dynamic, it is clear that the results have varied widely without the emergence of a clear consensus. One possible explanation is diversity in how difference states are regulated in different organisms. While this certainly contributes some variation in the results, we feel that most of the variation results from inconsistencies in the experimental and analytical procedures, particularly as these relate to error handling. The biggest recurrent problem has been the difficulty of obtaining direct measurements for post-transcriptional processes. As a result, an explanation-by-subtraction approach was used again and again, in which transcript levels were shown to explain a surprisingly small share of the total variance in protein levels. While fine in principle, these explanation-by-subtraction strategies require careful treatment of error – both stochastic and systematic. For the analysis of baseline protein levels, control for stochastic error was occasionally implemented and systematic errors were addressed only by Biggin (Li et al., 2014). For the analysis of protein dynamics, no studies have yet accounted for experimental error (though we do believe that the role of systematic errors may be more minor in this context since values are normalized against baseline). Thus, we believed that a project plan with more direct experimental measurements and improved computational treatment could provide significant clarity in what had become a rather murky field.

Using the response of primary mouse monocyte-derived dendritic cells to the immune stimulant lipopolysaccharide (LPS) as a model system, we built on the previous SILAC approaches by using a triple label (also used recently (Kristensen et al., 2013)) that provided separate channels for measuring protein production and degradation processes. The cells were raised in Medium (M) SILAC media such that all proteins had incorporated the M-labeled amino acids. At *t*=0, cells were stimulated with LPS (or mock stimulus) with concurrent media switch to Heavy (H), such that newly translated proteins would be H-labeled. Before mass spectrometry quantitation, each aliquot was finally spiked with an equal volume of "master mix" cells grown in Light (L) SILAC media (providing the basis for cross-sample normalization). Thus, the M:L ratio was indicative of protein degradation and the H:L ration was indicative of protein production (*i.e.* $k_{tra} \cdot R$). Importantly, we addressed the matter of systematic errors in baseline protein expression by collecting several orthogonal peptide libraries (via different enzymatic digestions). This enabled a quantitative assessment of the reliability of the protein level estimates and provided the opportunity to correlate imputed translation rates with protein levels without the worry that hidden error terms would inflate the estimate (as previously (Schwanhäusser et al., 2011)). Also critical, we additionally acquired ribosomal profiling data to provide an independent estimate of translation rates. This enabled a head-to-head comparison of what proteomics *vs.* ribosomal profiling imply regarding the importance of translation rates to protein levels.

Computationally, we used a system of differential equations to model protein levels. While similar in concept to Lee (Lee et al., 2011), we added in additional complexity by permitting dynamic changes in protein translation and degradation rates. Furthermore, we augmented the

analytic approach to account for label cross-talk, namely the incorporation of M-labeled amino acids (released from newly degraded M-labeled proteins) into new proteins and the degradation of new H-labeled proteins (affecting the perceived rate of production). We fit these dynamic models with a statistical approach that accounted for the signal-to-noise ratio in each protein's measurements and avoided over-fitting by applying Empirical Bayes (software packaged as "DogmaQuant" and available online http://www.sciencemag.org/content/suppl/2015/02/11/science.1259038.DC1/Jovanovic_Rpackage.zip). Most importantly, we accounted for all known sources of error (stochastic and systematic), adjusting for them in our calculations using the Spearman correction.

As we demonstrate, this approach significantly upwardly revises the contribution of RNA to the protein levels at baseline, amounting to roughly twice translation rates and degradation rates combined (published (Jovanovic et al., 2015) and discussed (Li and Biggin, 2015)). Importantly, we show that ribosomal profiling yields a numerical result nearly identical to that obtained from the proteomic modeling. In the dynamic response, we show that RNA is fully dominant, explaining 90% of protein level changes. That being said, while RNA levels drive most of the change, protein degradation rates and baseline protein levels play a considerable role in determining the ultimate trajectories of these responses by tempering the amplitude of change and/or modulating the rate of return to baseline. Furthermore we identify specific, highly-expressed metabolic processes that appear to rely substantially on post-transcriptional regulation. Nonetheless, these results suggest RNA levels and their fold changes contain most of the information necessary to understand dynamics at the protein level (consistent with the

large number of transcription factors and their combinatorial interactions) and reinforce the importance of further deciphering the regulatory code that determines gene transcription.

Results

A pulsed-SILAC strategy to measure protein dynamics

We assessed how protein levels are maintained in the context of the model response of mouse immune bone marrow-derived dendritic cells (DCs) (Steinman and Banchereau, 2007) to stimulation with lipopolysaccharide (LPS) (Amit et al., 2009; Chevrier et al., 2011; Garber et al., 2012; Mellman and Steinman, 2001; Rabani et al., 2011). This is a compelling system, as DCs are mostly post-mitotic, and LPS synchronizes them (Shalek et al., 2013) and causes dramatic regulatory changes from the expression of thousands of transcripts (Amit et al., 2009; Garber et al., 2012; Rabani et al., 2011) to protein phosphorylation (Chevrier et al., 2011). To monitor protein production and degradation during a dynamic response, we used a modified pulsed-SILAC approach (Boisvert et al., 2012) (**Figure 1**, **Methods**) to track newly synthesized and previously labeled proteins over time.

We cultured DCs for 9 days in medium-heavy labeled (M) SILAC medium, then substituted the M SILAC medium with heavy-labeled (H) SILAC medium and immediately stimulated them with LPS or medium (MOCK). Newly-synthesized proteins were thus labeled with heavy (H) amino acids, serving as a proxy for protein synthesis, while proteins with medium-heavy (M) amino acids decayed over time, reflecting cellular half-lives. For normalization, we spiked in a reference sample, extracted from a mix of unstimulated and stimulated DCs grown in light (L) SILAC media. We collected biological replicate samples at 10 time points over 12 hours (0h, 0.5h, 1h, 2h, 3h, 4h, 5h, 6h, 9h, 12h) after LPS or mock stimulation. We quantified 6,079 proteins by LC-MS/MS in at least one sample and 2,288 proteins in all samples (time points,

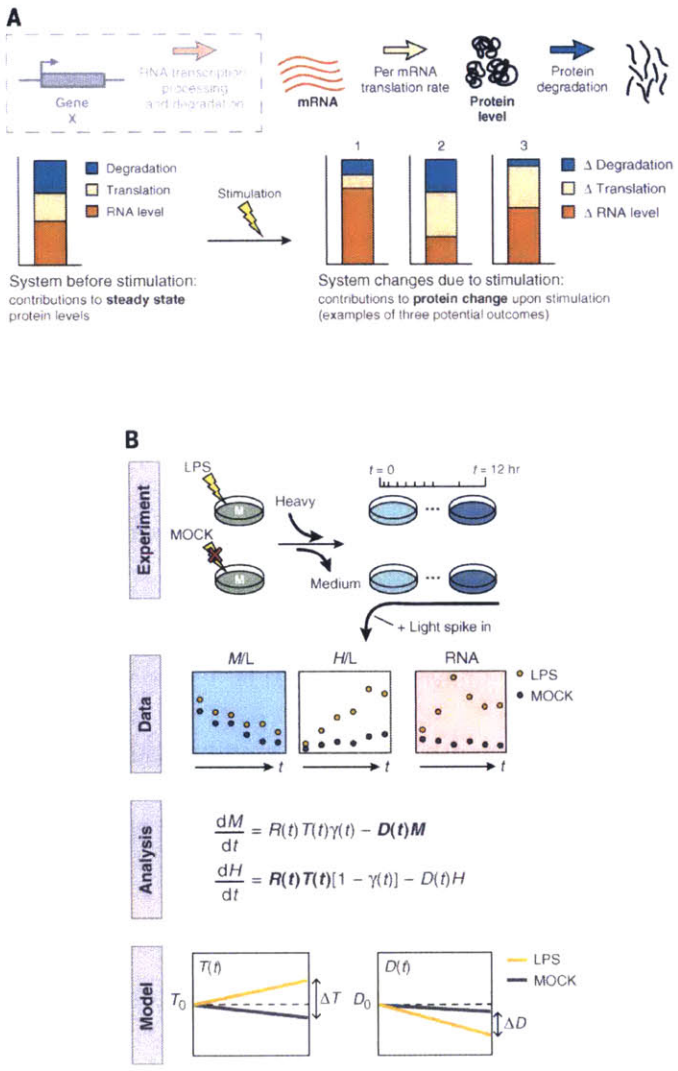


Figure 1. Framework to study the dynamic protein life cycle

(**A**) The dynamic protein life cycle. Top: RNA transcription, processing and degradation (dashed grey box) determine mRNA levels (red), which together with per-mRNA translation (tan) and protein degradation/removal (blue) determine final protein levels. Bottom: Hypothetical contribution of each process (stacked chart) to protein levels at steady-state (left) or to fold changes (right, three hypothetical scenarios). (**B**) Experimental and analysis workflow. From top to bottom: experimental system ("Experiment") consistent of DCs grown in medium-heavy SILAC (M) medium until LPS (top) or MOCK (bottom) stimulation, when heavy (H) SILAC is substituted. A "standard", light (L) SILAC labeled sample is spiked in. The resulting measurements ("Data") include M/L and H/L ratios (proxies for protein degradation/removal and production, respectively), as well as RNA-Seq data at each time point. These are used to fit the parameters of an ODE model ("Analysis"), where R(t) = modeled mRNA change over time; T(t) and D(t) = per-mRNA translation and protein degradation rate constants over time, respectively; γ(t) = recycling ('impurity') rate; H(t) and M(t) = modeled change in heavy (H/L) and medium (M/L) channels, respectively. The results ("Model") are the estimated per-mRNA translation and protein degradation rates over time. See text and **Methods** for details.

conditions and replicates; **Figure 2A**, **Table S1**). We independently measured replicate RNA-Seq profiles under the same conditions (**Figure 2A**, **Table S2**, **Methods**).

A model-based estimation of protein synthesis and degradation rates

We devised a computational strategy to infer per-mRNA translation rates ($T(t)$) and protein degradation rates ($D(t)$) at each time point from the temporal transcriptional profiles ($R(t)$) and H/L and M/L protein ratios ($H(t)$ and $M(t)$, respectively) (**Figures 1B**, **S1**, **Methods**). We defined a model that describes the relevant processes and associated rates (*e.g.*, translation rate, protein degradation rate), and then fitted the parameters (*e.g.* rates) in the model with our mRNA and protein data. Specifically, we used an ordinary differential equations model describing, for each gene i, the changes in $M_i(t)$ and $H_i(t)$ ($dM_i(t)/dt$ and $dH_i(t)/dt$, respectively) as a function of (1) a production term, governed by mRNA abundance $R_i(t)$ and a per-mRNA molecule translation rate constant, $T_i(t)$; and (2) a degradation term, modeled as an exponential decay function, governed by a protein degradation rate constant, $D_i(t)$. Both terms are also affected by $\gamma(t)$, the global M SILAC label recycling rate constant (**Figures 1B**, **S1**, **S2**, **Methods**). All rate constants are dynamic, and the mRNA levels, per-mRNA translation rate constant and protein degradation rate constant are also gene-specific. We modeled the change over time in the per-mRNA translation rate constant ($T_i(t)$) and in the degradation rate constant ($D_i(t)$) as linear functions. This assumption reduces the number of free parameters, thus providing robustness while retaining the capacity to detect the effect of sustained changes, even if these changes do not manifest linearly *in vivo* (as in the case of step functions).

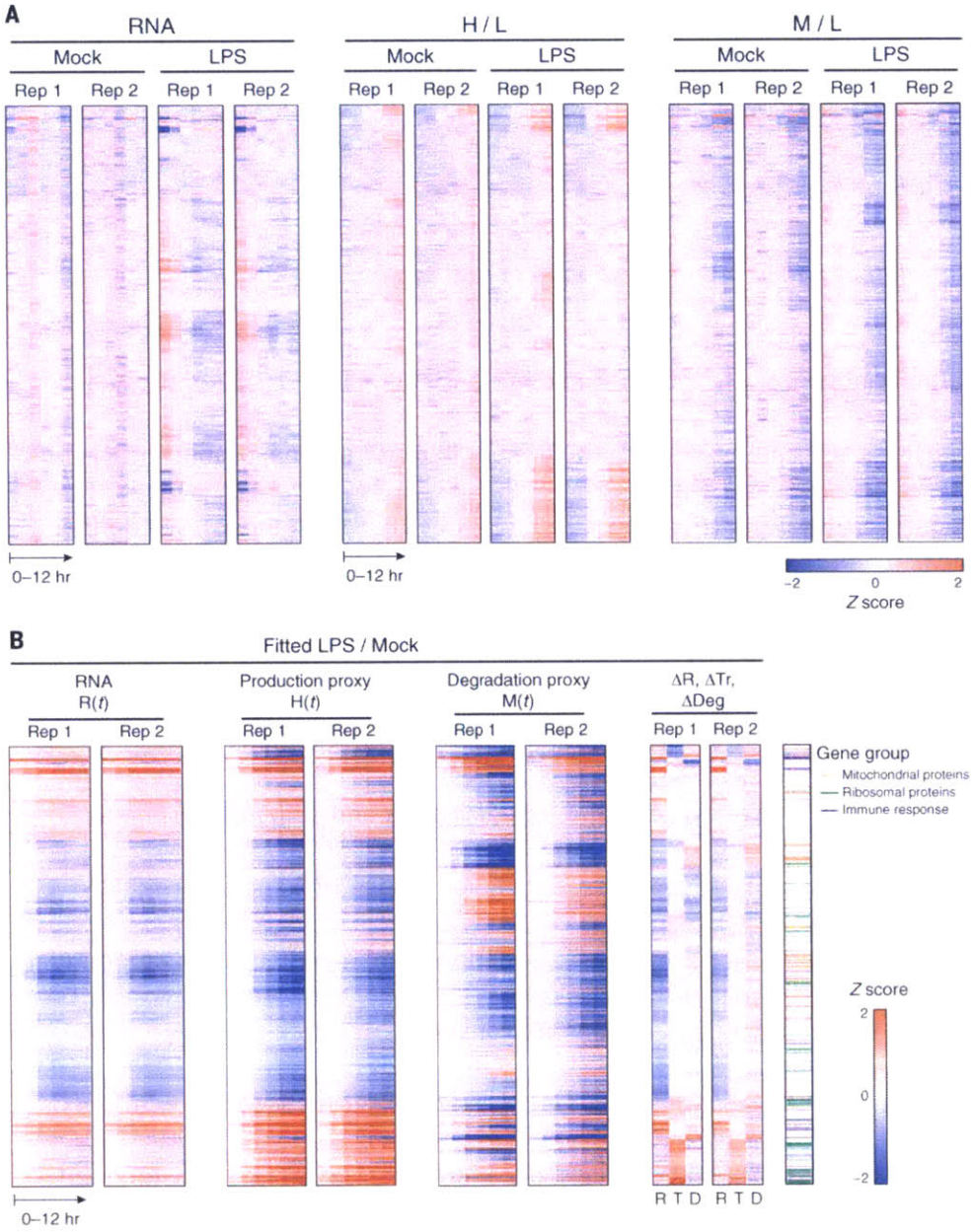


Figure 2. The protein life cycle in LPS stimulated DCs

(**A**) Shown are (left to right) for all 2,288 genes (rows) that were quantified in all samples, mRNA expression, H/L protein expression and M/L protein expression in LPS and MOCK stimulated DCs from each replicate (columns). Gene order is the same across all heatmaps, and determined by hierarchical clustering of fitted fold changes in mRNA level, translation rate, and degradation rate. Values are median normalized by row, logged, and robust z-transformed per map (color scale). (**B**) Fitted differential expression of the same 2,288 genes (rows). Left to right: Robust z-score fitted differential expression ratios (LPS/MOCK; red/blue color scale) for R(*t*), H(*t*) and M(*t*) in LPS *vs.* MOCK stimulated DCs from each replicate (columns) with the log2 fold changes between LPS and MOCK stimulated DCs at 12h post stimulation for mRNA (ΔR), per-mRNA translation rate (ΔTr), and protein degradation rate (ΔDeg) (also z-scored). Right most column: immune response (purple), ribosomal (green) and mitochondrial (orange) proteins.

We fitted the different parameters in the model (**Figure S1**) with the RNA-Seq and mass spectrometry (MS) data, using an empirical Bayes strategy, which prevents over-fitting of noisy MS data by sharing information across genes (**Methods**). In this approach, our most differential and reliable parameter estimates correspond to the well-quantified genes, whereas proteins with less reliable measurements are not associated with reliable changes. This ensures a low rate of false positives (calling a change where none exists), but may result in false negatives, and hence in some underestimation of the contribution of protein synthesis and degradation.

Fitting the parameters for 3,147 genes that passed our filtering criteria, separately for each of our replicates (**Figures 2B**, **S3**, **Table S3**), we found good reproducibility of the LPS/MOCK ratios of key fitted values (**Figures 2B**, **S4**) and of the relative differences in per-mRNA translation rates (*e.g.*, $\Delta T_i(12h) = T_i(12h)_{LPS}/T_i(12h)_{MOCK}$, Pearson $r = 0.68$, **Figure S5A**) or degradation rates (*e.g.*, $\Delta D_i(12h) = D_i(12h)_{LPS}/D_i(12h)_{MOCK}$, $r = 0.62$, **Figure S5B**). The robustness of these results was further supported by: (1) the fair correlation of our translation and protein degradation rate estimates in resting cells (**Table S3, Figure S6**) with previous estimates in mouse fibroblasts (NIH3T3) based on a similar pulsed-SILAC approach (Schwanhäusser et al., 2011) ($r(T_i(0))= 0.35$; $r(D_i(0)) = 0.58$; **Figure S7A**, **S7B**) or on estimates of translation rate efficiency (TE) values based on ribosome profiling in mouse fibroblasts (NIH3T3) (Subtelny et al., 2014) ($r(T_i(0)) = 0.37$; **Figure S7C**); (2) a good correlation between our per-mRNA translation rates and our independent measurement of TE values in DCs using ribosome profiling at t=0h ($r = 0.5$, **Table S4**, **Figure S8A**), comparable to the correlation between TE values in mouse DCs and mouse fibroblasts ($r = 0.54$, **Figure S8B**); (3) the fact that strong early changes are all in immune response proteins (**Figure S4A**); (4) the global increase upon LPS stimulation in protein

production rates ($T_i(12h)_{LPS}$ *vs.* $T_i(12h)_{MOCK}$, $P < 10^{-10}$, Wilcoxon rank-sum Test, **Figure S9A**) and protein degradation rates ($D_i(12h)_{LPS}$ *vs.* $D_i(12h)_{MOCK}$, $P < 10^{-10}$, Wilcoxon rank-sum Test, **Figure S9B**), consistent with other reports (Lelouard et al., 2007; Schmidt et al., 2009); and (5) the increase in the calculated 'degradation rate' – likely reflecting depletion by secretion, or "decreased cellular half life" – of proteins from the recently-characterized secretome of LPS stimulated mouse macrophages (Eichelbaum et al., 2012) ($P < 10^{-10}$, LPS *vs.* MOCK; Wilcoxon rank-sum Test, **Figure S9C**).

mRNA levels contribute the most to protein expression levels before stimulation

To determine the relative contribution of each step to steady state protein levels in unstimulated, post-mitotic DCs, we first estimated absolute protein levels from four additional MS data sets in resting DCs (0h) that rely on distinct peptides (**Methods**): two biological replicate samples, which were each digested in two technical replicates with LysN and AspN, respectively, rather than by trypsin, used for the pulsed SILAC samples.

Next, we assessed the contribution of each regulatory step to gene-to-gene differences in overall protein levels by comparing (with Spearman-corrected coefficients of determination) the independently-measured absolute protein levels to steady state protein levels predicted by our model when setting one or more of the three regulatory steps (mRNA level, per-mRNA translation rate constant, or protein degradation rate constant) to its per-gene inferred value (at time 0h) and setting the remaining steps to their pan-genome median value (**Methods**). By sequentially adding to the model further per-gene values rather than pan-genome medians

(say: mRNA level, translation rate, and finally degradation rate) and assessing the corresponding change in the correlation measure, we can assign additive regulatory contributions to the three steps. Because these three steps are not statistically independent from each other and may interact in a nonlinear manner, we explored every possible ordering of consideration.

Considering all 3 variables together, we account for nearly 79% of the variance of the independently measured protein levels (**Figures S10, S11A**). Of these 79%, mRNA levels explained 59-68%, per-mRNA translation rates 18-26%, and degradation rates 8-22%. (**Figures 3A**, **S11A**). We believe the unexplained variance is due to systematic errors in the measurements and modeling that could not be accounted for. In addition, we have separately estimated the variance in translation rates in the same cells under identical conditions using ribosome profiling to measure TE values. Using TE values instead of our pulsed-SILAC derived translation rates, we estimate a comparable contribution of protein synthesis (**Figures 3B**, **S10**). Thus, in postmitotic DCs, mRNA levels are contributing more to protein-to-protein variation in total protein levels than the protein life cycle (synthesis and degradation rates) combined.

mRNA abundance dynamics dominate protein changes post stimulation

Next, we determined the contribution of each regulatory step to protein fold changes at 12 hours. We used the model fit from a given replicate to predict the protein fold change at 12h, when using either MOCK-estimated parameters or one or more LPS-estimated parameters for mRNA level, translation rate, and degradation rate. We then compared these predictions to the

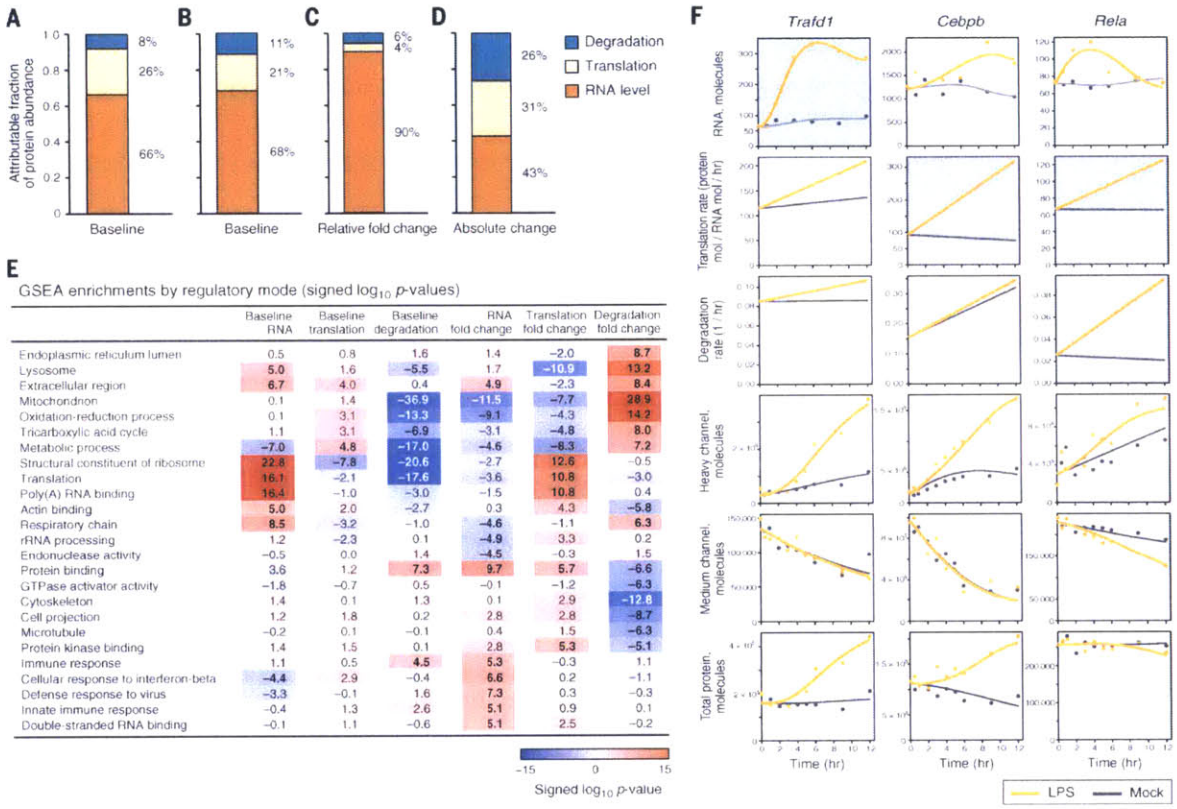


Figure 3. Contributions of mRNA levels and the protein life cycle to steady state and dynamic protein levels
(**A**-**D**) Global contributions of mRNA levels (orange); translation rates (tan); and protein degradation rates (blue) to protein levels. Translation rates were either derived from pSILAC data (**A**, **C** and **D**) or from TE values from ribosome profiling data (**B**). Contributions to steady state protein levels prior to LPS induction (**A** and **B**) or to the change in protein abundance between LPS induced and mock treated cells (**C** and **D**) are shown. The contributions to the fold change (**C**) and to the absolute change in protein abundances (**D**) after LPS stimulation are given. Note that the contributions for steady state presented exclude the percent of the variance in measured protein levels that is not explained by the variance in mRNA, translation or protein degradation (**Figure S10**). Per-gene parameter values were in the order: 1. mRNA; 2. translation; 3. degradation (**Methods**). For all possible orderings see **Figure S11**. (**E**) Functional processes controlled by distinct regulatory steps. For each process (rows) and regulatory step (columns) shown are the magnitudes of the $\log_{10}$(P-values) for the values or differential fold changes (LPS/MOCK at 12h) of mRNA levels, protein synthesis or degradation rates of genes annotated to this process *vs*. the background of all genes fit by the model. Values are signed according to directionality of the enrichment (Wilcoxon Rank Sum test). Shown are the 5 gene sets most enriched for increased or decreased rates for the three 'fold change' columns, along with their scores in all six regulatory modes. Nearly-redundant gene sets were removed (see table S6 for all gene sets). (**F**) Examples of regulation of expression dynamics. For each of three genes in each of LPS (orange) and MOCK (black) condition shown are the measured values (dots) and fits (curves) for (top to bottom) mRNA levels (in mRNA molecules), per-mRNA translation rates (protein molecules / mRNA molecule / hour (hr)), degradation rates (1/hr), H(t), M(t) and total protein ((M+H)(t)). X-axis: time; Y-axis: intensity or rate. Light blue: key regulatory mode. mRNA and protein molecules are only proxies for transcripts per million (TPM) and IBAQ microshares, respectively, in order to help interpretation (**Methods**).

fitted fold changes from the other replicate. Starting with all parameters set to MOCK-estimated rates, we sequentially used LPS-estimated parameters for mRNA, translation rate, and degradation rate (in every possible order), and thus assessed the contribution of each step as the increase in the Spearman-corrected coefficients of determination (**Methods**).

We found that mRNA levels explain ~87-92%, per-mRNA translation rates ~4-7%, and degradation rates ~3-6% of protein fold changes after 12 hours (**Figures 3C**, **S11B**). mRNA fold changes contributed at least eight times as much as the protein life cycle combined for both induced and repressed proteins (**Figure S12**, **Table S5**). However, changes in per-mRNA translation rates contributed more substantially to protein level induction, whereas changes in protein degradation rates mostly contributed to protein level repression (**Figure S12**, **Table S5**).

Fold changes in induced immune response proteins were particularly dominated by mRNA level changes (**Figures 2B**, **3E**, **Table S6**). For example, transient up-regulation of the mRNA encoding the negative immune regulator Trafd1 (Sanada et al., 2008) (**Figure 3F**) is the main cause of a strong increase in its protein. In Trafd1 and hundreds of other genes, a transient, strong, spiked change in mRNA, combined with a time-constant protein half life, much longer than the 12h time course, result in a monotonous increase in protein levels, such that global protein fold changes at 12h post LPS correlate best to mRNA changes at 5h (**Figure S13**). Only a handful of proteins (*e.g.*, Tnfaip2 (Burton et al., 2007; Chevrier et al., 2011; Kuan et al., 1999)), show peaked, transient protein expression within our time scale; all have very high basal degradation rates, which typically do not increase further. Finally, a few key regulators of DCs and the LPS response (*e.g.*, Cebpb, a pioneer transcription factor whose mRNA is already very highly

expressed pre-stimulation, and Rela, **Figure 3F**) are considerably dynamically regulated at the protein level, such that increased protein degradation rates (Rela) and/or increased per-mRNA translation rates (Rela and Cebpb) are main drivers for protein change. These dynamic changes cannot be observed solely from *total* protein and transcript levels, but the corresponding *rate* changes are readily apparent (**Figure 3F**).

Notably, although our global model incorporates the data of only 3,147 genes, several lines of evidence suggest that this did not bias our global conclusions. First, while the 3,147 modeled genes are somewhat enriched for higher expressed genes (**Figure S14**), we do model a substantial number of lowly expressed mRNAs (**Figure S14**). Second, computationally correcting for this bias by recalculating the contributions of mRNA, per-mRNA translation and protein degradation rates while proportionally up-weighting the impact of under-represented expression bins (**Methods**), does not significantly affect our conclusions (**Figure S15**). Third, the correlation between our protein translation at baseline (t=0h), as estimated by pulsed-SILAC data or by TE values, is comparable when considering only the lowest expressed 25% (Pearson r~0.52), the highest expressed 25% (r~0.58), or all modeled proteins (r ~ 0.5) (**Figures S16A, S16B**). Finally, there is no significant difference in the distribution of TE values in the (under-represented) lowly expressed mRNA bins between those proteins we detect (in the 3,147 proteins) versus those we could not include in our model (P=0.069, t-test, **Figures S16C**); thus, it is unlikely that the lowly expressed genes that we could not model have unique regulatory modes.

Protein life cycle changes primarily affect proteins performing basic cellular functions

While mRNA fold changes contributed most to relative changes in protein expression (ratios of LPS to MOCK-simulated protein levels), protein synthesis and degradation rates do change significantly for 357 proteins (~11% of consistently detected proteins, **Tables S7**, **S8**), and in particular for proteins performing essential cellular functions ('housekeeping proteins', **Figures 2B**, **3E**, **Table S6**), including cytoskeletal, metabolic, ribosomal (**Figure 4A**) and mitochondrial proteins (**Figure 4B**). Since these are among the most abundant in the cell (Geiger et al., 2013; Kim et al., 2014; Schwanhäusser et al., 2011; Wilhelm et al., 2014), we reasoned that while mRNA changes may dominate the relative (fold) changes in protein levels following LPS stimulation, changes in the protein life cycle could contribute substantially more to differences in absolute cellular protein abundance than to relative changes. For example, consider two genes: gene 1 is induced 10-fold from 10,000 to 100,000 proteins (a substantial change in relative protein abundance), while gene 2 is induced 1.2-fold from 1,000,000 to 1,200,000 proteins (a substantial change in absolute protein abundance). We asked whether relative and absolute changes are associated with different regulatory mechanisms. Indeed, we found that changes in translation and degradation rates together explain more of absolute protein changes than changes in mRNA levels (mRNA: ~32% to 43% of the fit value; per-mRNA protein production rate: ~22-41%; degradation rates: ~19-36%, **Figures 3D**, **S11C**). Thus, post-transcriptional regulation contributes substantially more to absolute protein level changes than to relative protein level changes.

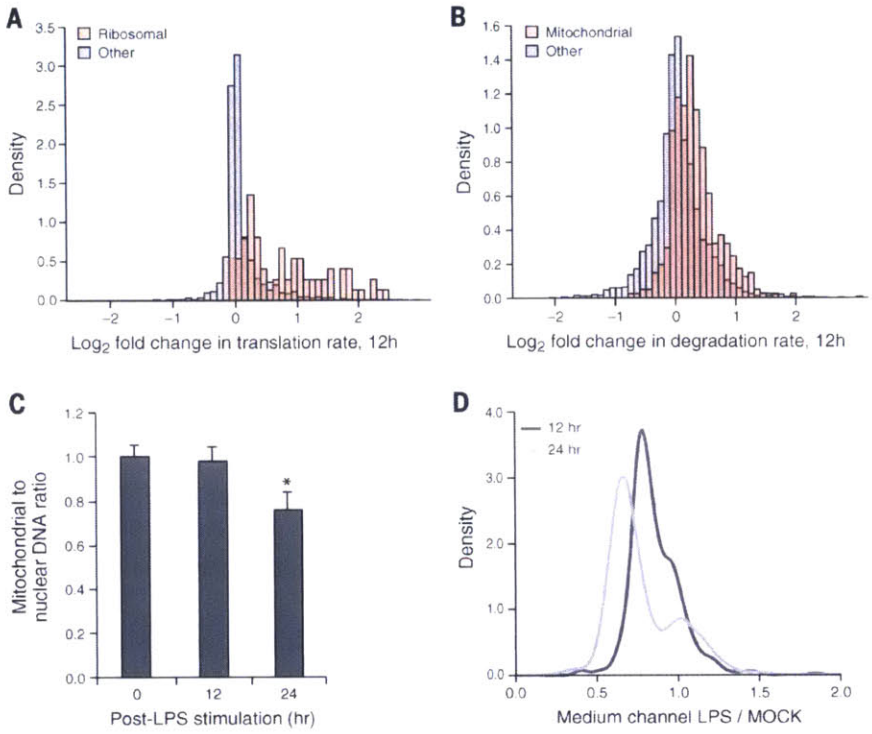


Figure 4. Degradation of mitochondrial proteins following LPS stimulation is associated with mitophagy
(**A**, **B**) Increased translation rates of some ribosomal proteins (**A**) and increased degradation rates of mitochondrial proteins (**B**). Shown are the distributions of log2 fold changes of translation rates (ΔT_i, **A**) or degradation rates (ΔD_i, **B**) between LPS and MOCK stimulated cells of all measured ribosomal proteins (A, red) or mitochondrial proteins (**B**, red; from Mitocarta annotations (Pagliarini et al., 2008), and all measured proteins (grey). (**C**) Evidence of mitophagy in LPS stimulated DCs. Shown is the mitochondrial to nuclear DNA ratio (Y axis) in DCs at 0h, 12h and 24h post LPS stimulation (X axis). Values are normalized to the average mitochondrial to nuclear DNA ratio at 0h. Asterisk: a significant change relative to 0h (*P*-value = 0.016, *t*-test, n=3). (**D**) Distribution of raw $\log_2$ LPS/MOCK M/L ratios (a proxy for protein decay) for all measured mitochondrial proteins (in Mitocarta (Pagliarini et al., 2008)) at 12h (black) and 24h (grey) post stimulation.

An increase in degradation rates of mitochondrial proteins is associated with mitophagy

Upon LPS stimulation, a substantial decrease in the level of mitochondrial proteins is associated with increased degradation rates, although these proteins are among the most stable in unstimulated DCs (**Figures 2B**, **3E**, **4B**, **Table S6**). This increase in protein degradation is accompanied by a significant decrease in mRNA levels (**Figure 3E**, **Table S6**, $P<10^{-10}$,Wilcoxon Rank Sum Test) and in per-mRNA translation rates (**Figure 3E**, **Table S6**, $P<10^{-7}$, Wilcoxon Rank Sum Test), suggesting decreased production of new mitochondrial proteins and increased destruction of old ones. Both structural mitochondrial proteins and enzymes in key mitochondrial metabolic pathways have increased degradation. The increased degradation of key enzymes, such as SUCLA2, ALDH2 and ACO2, is consistent with a reported shift in LPS-stimulated DCs from oxidative phosphorylation and oxygen consumption to glycolysis, glucose consumption and lactate production (Everts et al., 2014; Everts et al., 2012; Krawczyk et al., 2010; Pearce and Pearce, 2013).

The increased loss of structural proteins and enzymes in the mitochondria may be due to either a targeted metabolic shift in carbon and energy metabolism through a reduction of a specific subset of the mitochondrial proteome, or a more global loss of entire mitochondria through mitophagy. To experimentally distinguish between the two hypotheses, we measured the mitochondrial-to-nuclear DNA ratio in unstimulated DCs at 12h and 24h post LPS stimulation (the latter time point was chosen to account for any delay in complete mitochondrial DNA degradation) (**Figure 4C**). There was no significant change in the ratio of mitochondrial-to-nuclear DNA at 12h post LPS stimulation, but a significant (~25%; P=0.016, t-test) reduction at

24h post stimulation (**Figure 4C**). Indeed, analyzing pulsed SILAC data collected at 24h post LPS and mock stimulation, we saw a decrease in the M/L ratios (a proxy for increased degradation) of ~80% of annotated mitochondrial proteins in LPS *vs*. MOCK samples (**Figure 4D**), and in nearly all mitochondrial proteins with a higher mitochondrial localization prediction score (from MitoCarta (Pagliarini et al., 2008) – over 95% of the 156 proteins with a score > 20 of the 472 measured mitochondrial proteins) (**Figure S17**). These results suggest that mitophagy is a driver of LPS induced mitochondrial protein degradation in DCs, consistent with previous observations of mitophagy in virus- or bacteria-infected DCs (Lupfer et al., 2013) and might also contribute to epitope presentation, as previously proposed (Bell et al., 2013).

Discussion

We determined the contribution of changes in mRNA levels, protein synthesis and protein degradation rates during a dynamic response and found that changes in mRNA levels dominate relative fold changes. When considering also absolute changes in protein molecules (abundance), our data suggests a model where the cellular proteome is dynamically regulated through two strategies.

In the first strategy, mRNA regulation acts primarily to ensure that specific functions – here, immune response proteins – are only expressed when needed and thus explains most of the fold-change differences in protein levels, contributing to LPS induced protein fold changes at least 8 times as much as the combined protein life cycle within the 12h time scale of our measurements. It is possible that protein life cycle changes are important to turn over key regulatory and signaling proteins at later phases of the response. While our study does not directly address which steps in mRNA regulation account for this, our related work on the RNA life cycle during the first 3 hours in LPS stimulated DCs suggests that transcriptional changes may in turn dominate differential mRNA expression, whereas dynamic changes in RNA processing or degradation affect only a minority of genes, albeit with important functions (Rabani et al., 2014). Furthermore, in contrast to previous reports where degradation rates contributed only marginally (Kristensen et al., 2013; Schwanhäusser et al., 2011), but consistent with Li et al. (Li et al., 2014), we see that within the protein life cycle, changes in protein degradation rates play an equal role to changes in per-mRNA translation rates. Although some of this is due to turnover from increased secretion of some proteins (**Figures S9C, S18**),

excluding the secretome (Eichelbaum et al., 2012) from our analysis did not strongly alter this global trend (**Figure S18**). Finally, while mRNA changes dominate changes in protein levels, it may be difficult to discern this relationship in the absence of a model-driven analysis. Thus, while mRNA induction is readily reflected in protein level induction, albeit somewhat dampened, few of the 912 repressed mRNAs (>2 fold), show matching protein changes (**Figure S19**, **Table S9**). This could be naively interpreted as substantial posttranscriptional control, but pre-existing proteins, the long protein half-life and the delay of protein changes relative to mRNA changes (**Figure S13**) complicate such an intuitive interpretation, and our analysis shows that mRNA changes drive protein down-regulation as well (**Figure S12**, **Table S5**).

In the second strategy, regulation at the protein level primarily readjusts the pre-existing proteome, especially 'housekeeping' proteins, in order to meet the requirements of a new cellular state, such as change in shape, metabolism, *etc*. Thus, when we consider the contribution of a change in each rate to the change in the *number* of proteins (rather than the relative fold change), the contribution of changes in the protein life cycle is substantially increased (**Figure 3D**). We find similar patterns of contributions when we use the Spearman rank correlation rather than Pearson correlation (**Figure S20**), suggesting that our conclusions are robust to outliers with particularly strong changes.

The extent to which this two-part strategy applies in other dynamic settings remains to be determined. Interestingly, recent studies comparing protein and translation rate differences between different states (*e.g.*, differentiated *vs.* non-differentiated cells or between different yeast strains) suggested that translation rate differences affect differential protein expression

only modestly (Albert et al., 2014; Baek et al., 2008; Hsieh et al., 2012; Ingolia et al., 2011; Kristensen et al., 2013; Selbach et al., 2008), but do impact some highly expressed proteins, including ribosomal proteins (Hsieh et al., 2012; Ingolia et al., 2011), also translationally regulated in our system.

Our analysis of unstimulated (resting) postmitotic DCs refines and extends previous models of protein level regulation in steady state. In our cells, nearly two thirds of the gene-to-gene variation in total protein levels is explained by regulation of mRNA levels, a higher contribution than previously reported in dividing mammalian cells (Schwanhäusser et al., 2011), possibly due to the regulatory mechanisms active in primary post-mitotic, homeostatic resting cells. For example, the increased role we observed for protein degradation, in contrast to prior studies (Kristensen et al., 2013; Schwanhäusser et al., 2011), may be needed by postmitotic cells that cannot simply renew their protein pool by division-coupled passive dilution. Furthermore, our analysis corrected for RNA-Seq expression reproducibility, intra-library protein expression reproducibility, and library-dependent protein expression biases (**Figure S21**), all essential to avoid inadvertent attribution of measurement errors to modeled translation and protein degradation rates. Indeed, whereas from raw data mRNA explains 27% of the gene-to-gene variation in protein levels at baseline (t=0), using modeled expression values it explains 42%, and, once correcting for data reproducibility (**Methods**), it explains 52%. This compares well to a recent study (Li et al., 2014) that found that mRNA levels explain at least 56% of the differences in protein abundance (when estimating the variances of errors with control measurements (Schwanhäusser et al., 2011)), and possibly as much as ~84% (using TE values to estimate the systematic error in translation rates in (Schwanhäusser et al., 2011). Each of these

strategies highlights the importance of determining and correcting for stochastic and systematic errors in the data. Notably, even with our conservative estimates, the protein life cycle is estimated to contribute, at minimum, about a third of the final steady state protein expression level. Since protein expression levels span around 4 to 5 orders of magnitude (Geiger et al., 2013; Kim et al., 2014; Schwanhäusser et al., 2011; Wilhelm et al., 2014), differences between genes in the protein life cycle can easily cause a ten to a hundred fold change in protein expression.

Our experimental and analytical design should be broadly applicable to study similar events in diverse dynamical cell systems. Our analytical model distinguishes per-mRNA protein production and protein degradation rates that were confounded in previous, model-free analyses of raw H/L and M/L ratios from dynamic pulsed-SILAC data (Kristensen et al., 2013), due to *e.g.*, the contribution of mRNA and protein degradation to the H/L signal and of recycled labeled amino acids to the M/L signal (**Methods**). Our empirical Bayes strategy also handles noise in proteomics data in a principled and conservative way. Nevertheless, we make some simplifying assumptions in our model (*e.g.*, linear changes in per-mRNA translation rates and degradation rates) that may be refined in the future (*e.g.*, with sigmoidal functions (Chechik et al., 2008; Rabani et al., 2011; Yosef and Regev, 2011)), allowing us to estimate additional valuable parameters (*e.g.*, time point of rate change). This would require finer-resolution data, such as from ribosome profiling (Ingolia et al., 2009; Ingolia et al., 2011; Stern-Ginossar et al., 2012), puromycin-associated nascent chain proteomics (Aviner et al., 2013), or combining pulsed-SILAC labeling with pulse-labeling using the methionine analogue azidohomoalanine (Eichelbaum and Krijgsveld, 2014; Eichelbaum et al., 2012). Such enhanced methods will

provide a framework to study the contributions of the protein life cycle in diverse dynamic systems and help identify new key regulators of these responses.

Methods

Bone marrow derived dendritic cells (BMDCs) growth conditions

All animal protocols were reviewed and approved by the MIT / Whitehead Institute / Broad Institute Committee on Animal Care (CAC protocol 0609-058-12). To obtain sufficient number of cells, we implemented a modified version of the DCs isolation protocol as described previously (Amit et al., 2009; Chevrier et al., 2011; Garber et al., 2012; Lutz et al., 1999; Rabani et al., 2011). Briefly, 6-8 week old female C57BL/6J mice were obtained from the Jackson Laboratories. RPMI medium (Invitrogen) supplemented with 10% heat inactivated FBS (Invitrogen), β -mercaptoethanol (50uM, Invitrogen), L-glutamine (2mM, VWR), penicillin/streptomycin (100U/ml, VWR), MEM non-essential amino acids (1X, VWR), HEPES (10mM, VWR), sodium pyruvate (1mM, VWR), and GM-CSF (20 ng/ml; Peprotech) was used throughout the study. At day 0, bone marrow-derived dendritic cells (BMDCs) were collected from femora and tibiae and plated on twenty (per mouse), 100mm non tissue culture treated plastic dishes using 10ml medium per plate. At day 2, cells were fed with another 10ml medium per dish. At day 5, cells were harvested from 15ml of the supernatant by spinning at 1,400 rpm for 5 minutes; pellets were resuspended with 5ml medium and added back to the original dish. Cells were fed with another 5ml medium at day 7. At day 8, all non-adherent and loosely bound cells were collected and harvested by centrifugation. Cells were then resuspended with medium, plated at a concentration of $10x10^6$ cells in 10ml medium per 100mm dish. At day 9, cells were stimulated for various time points with LPS (100ng/ml, rough, ultrapure *E. coli* K12 strain, Invitrogen) or MOCK (= no stimulation).

For SILAC experiments, GM-CSF-derived BMDCs were grown and selected as described above, but in media containing medium-heavy L-arginine 13C6 (Arg6) and L-lysine 2H4 (Lys4) (Sigma). Just prior to LPS or MOCK stimulation, all cells were pooled, pelleted and re-suspended in media containing heavy L-arginine 13C6-15N4 (Arg10) and heavy L-lysine 13C6-15N2 (Lys8) (Sigma) and plated at 10^6 cells/ml on non-tissue culture treated Petri dishes. In parallel, GM-CSF derived BMDCs were grown in light L-arginine (Arg) and L-lysine (Lys) (Sigma) containing media. Concentrations for L-arginine and L-lysine were 42 mg/l and 40 mg/l, respectively. The cell culture media, RPMI-1640 deficient in L-arginine and L-lysine, was a custom media preparation from Caisson Laboratories (North Logan, UT) and dialyzed serum was obtained from SAFC-Sigma. We followed all standard SILAC media preparation and labeling steps as previously described (Ong and Mann, 2006).

DC sample collection

Two independent replicates were acquired for both RNA-seq and pulsed SILAC experiment time courses. Another two independent replicates were obtained for measuring protein levels at time 0h (see also below). MOCK and LPS stimulated DCs per replicate were always collected in parallel and started from the same cells (= time point 0h), which were split upon stimulation. For RNA-seq the following time points were collected: 0h (LPS/MOCK together), 1h, 2h, 4h, 6h, 9h, 12h. For the pulsed SILAC experiment the following time points were collected: 0h (LPS/MOCK together – right after medium-heavy to heavy SILAC media switch), 0.5h, 1h, 2h, 3h, 4h, 5h, 6h, 9h, 12h and also 24h (not used for the time course analysis, but only in **Figures 4D,**

S17B). For the light SILAC labeled standard spike-in, DCs were collected at three time points (0h, 6h, 24h), pooled and further processed as a pooled sample.

RNA isolation, library preparation and sequencing

RNA was extracted from cells using RNeasy Mini Kit (Quiagen), according to the manufacturer's protocol. Enrichment of polyadenylated RNA (polyA+ RNA) from total RNA was performed using Oligo(dT) dynabeads (Invitrogen) according to the manufacturer's protocol. The mRNA was chemically fragmented into ~80-nt-long fragments using RNA fragmentation reagent (Ambion), followed by Turbo DNAse treatment (Ambion). Strand-specific RNA-seq libraries were generated as previously described (Engreitz et al., 2013). Briefly, RNA was first subjected to FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific), followed by a 3' ligation of an RNA adapter using T4 ligase (New England Biolabs). Ligated RNA was reverse transcribed using AffinityScript Multiple Temperature Reverse Transcriptase (Agilent), and the cDNA was subjected to a 3' ligation with a second adapter using T4 ligase. The single-stranded cDNA product was then amplified for 9 to 14 cycles in a PCR reaction. Libraries were sequenced on an Illumina HiSeq 2500 generating 30bp paired-end reads.

DC protein isolation and processing for subsequent mass spectrometry

After stimulation (LPS or MOCK) and the appropriate time points, cells were washed twice with PBS and lysed for 30 min in ice-cold lysis urea buffer (8 M urea; 75 mM NaCl, 50 mM Tris HCl pH 8.0, 1 mM EDTA, 2 μg/mL aprotinin (Sigma, A6103), 10 μg/mL leupeptin (Roche, #11017101001), 1 mM PMSF (Sigma, 78830)). Lysates were centrifuged at 20,000*g* for 10 min, and protein concentrations of the clarified lysates were measured via BCA assay (Pierce). From

this procedure, DC lysates produced ~100 μ g of protein per 1 million cells. Light- standard and heavy/medium-labeled sample lysates were then combined in a 1:1 protein ratio (20 μ g each and therefore 40 μ g total). Protein disulfide bonds of the combined lysates were reduced for 45 min with 5 mM dithiothreitol (Thermo Scientific, 20291) and alkylated for 45 min with 10 mM iodoacetamide. Samples were then diluted 1:4 with 50 mM Tris HCl, pH 8.0, to reduce the urea concentration to <2 M. Lysates were digested overnight at room temperature with trypsin in a 1:50 enzyme-to-substrate ratio (Promega, V511X) on a shaker. Peptide mixtures were acidified to a final volumetric concentration of 1% formic acid (Fluka, 56302) and centrifuged at 10,000*g* for 5 min to pellet urea that had precipitated out of solution. The peptide mixtures were fractionated by Strong Cation Exchange (SCX) using StageTips as previously described (Rappsilber et al., 2007) with slight modifications. Briefly, one StageTip was prepared per sample by 3 SCX discs (3M, #2251) topped with 2 C18 discs (3M, #2215). The packed StageTips were first washed with 100 μ l methanol and then with 100 μ l 80% acetonitrile and 0.5% acetic acid. Afterwards they were equilibrated by 100µl 0.5% acetic acid and the sample was loaded onto the discs. The sample was transeluted from the C18 discs to the SCX discs by applying 100 μ l 80% acetonitrile; 0.5% acetic acid, which was followed by 6 stepwise elutions and collections of the peptide mix from the SCX discs. The first fraction was eluted with 50 μ l 50mM NH_4AcO; 20% MeCN (pH 4.1, adjusted with acetic acid), the second with 50 μ l 50mM NH_4AcO; 20% MeCN (pH 4.8, adjusted with acetic acid), the third with 50 μ l 50mM NH_4AcO; 20% MeCN (pH 6.2, adjusted with acetic acid), the fourth with 50 μ l 50mM NH_4AcO; 20% MeCN (pH 7.2), the fifth with 50 μ l 50mM NH_4HCO_3; 20% MeCN (pH 8.5) and the sixth with 50 μ l

0.1% NH_4OH; 20% MeCN (pH 9.5). 200µl of 0.5% acetic acid was added to each of the 6 fractions and they were subsequently desalted on C18 StageTips as previously described (Rappsilber et al., 2007) and evaporated to dryness in a vacuum concentrator. Peptides were reconstituted in 7µl 3% MeCN/0.1% formic acid (at an estimated concentration of 1µg/µl).

LC-MS/MS measurements

All peptide samples were separated on an online nanoflow EASY-nLC 1000 UHPLC system (Thermo Fisher Scientific) and analyzed on a benchtop Orbitrap Q Exactive mass spectrometer (Thermo Fisher Scientific) as previously described (Mertins et al., 2013). Briefly, approximately 1µg of peptides per sample was injected onto a capillary column (Picofrit with 10µm tip opening / 75µm diameter, New Objective, PF360-75-10-N-5) packed in-house with 20cm C18 silica material (1.9µm ReproSil-Pur C18-AQ medium, Dr. Maisch GmbH, r119.aq). The UHPLC setup was connected with a custom-fit microadapting tee (360µm, IDEX Health & Science, UH-753), and capillary columns were heated to 50°C in column heater sleeves (Phoenix-ST) to reduce backpressure during UHPLC separation. Injected peptides were separated at a flow rate of 200 nL/min with a linear 80 min gradient from 100% solvent A (3% acetonitrile, 0.1% formic acid) to 30% solvent B (90% acetonitrile, 0.1% formic acid), followed by a linear 6 min gradient from 30% solvent B to 90% solvent B. Each sample was run for 150 min, including sample loading and column equilibration times. Data-dependent acquisition was performed using Xcalibur 2.2 software in positive ion mode at a spray voltage of 2.00 kV. MS1 Spectra were measured with a resolution of 70,000, an AGC target of $3e^6$ and a mass range from 300 to 1800 *m/z*. Up to 12 MS2 spectra per duty cycle were triggered at a resolution of 17,500, an AGC

target of $5e^4$, an isolation window of 2.5 *m/z* and a normalized collision energy of 25. Peptides that triggered MS2 scans were dynamically excluded from further MS2 scans for 20 s.

Preparation of alternatively digested peptide libraries

BMDCs were grown in conventional medium and under conditions as described above. Two independent replicates were grown and BMDCs were collected as described above at day 9 with no LPS stimulation (i.e., corresponds to time point 0h). Protein isolation and processing for subsequent mass spectrometry was performed as described above, but instead of trypsin, the replicates were digested at a ratio of 1:50 with either LysN (cleaves peptide bonds N-terminal to Lysine residues; U-Protein Express BV) or AspN (cleaves peptide bonds N-terminal to aspartic acid residues; Roche), generating 4 samples (2 replicates and 2 different digests each) from time point 0h with a peptide composition that is entirely different than the time course samples, which have been digested with trypsin (see above). Unfractionated peptide samples were analyzed by LC-MS/MS as described above and identified and quantified by MaxQuant (Cox and Mann, 2008) with the IBAQ feature enabled as described below, with the exception that either LysN or AspN was selected as the digestion enzyme. IBAQ enables relative protein quantification and IBAQ values are proportional to absolute protein values (Schwanhäusser et al., 2011). However, IBAQ values have been reported to estimate protein levels with an accuracy of 2 to 5 fold of the real value (Ahrne et al., 2013) and it has also been reported that this estimation depends partly on the peptide composition of the sample (Peng et al., 2012). Using these 4 independent samples to quantify the relative protein levels at baseline (time point 0h) avoids systematic errors, which otherwise could lead to an overestimation of the

contribution of per-mRNA translation rates to baseline protein levels (for details see below: "**Determining contributions to total protein expression at baseline**").

Identification and quantification of proteins

All mass spectra were analyzed with MaxQuant software version 1.3.5 (Cox and Mann, 2008) using the mouse UniProt database (March 2013). MS/MS searches for the proteome data sets were performed with the following parameters: Oxidation of methionine and protein N-terminal acetylation as variable modifications; carbamidomethylation as fixed modification. Trypsin/P was selected as the digestion enzyme for the pulsed-SILAC experiments and either LysN or AspN was selected as the digestion enzyme for the additional protein quantification experiments for time 0h, and a maximum of 3 labeled amino acids and 2 missed cleavages per peptide were allowed. The mass tolerance for precursor ions was set to 20 p.p.m. for the first search (used for nonlinear mass re-calibration) and 6 p.p.m. for the main search. Fragment ion mass tolerance was set to 20 p.p.m. The IBAQ feature was enabled in order to estimate relative proteins levels (Schwanhäusser et al., 2011). For identification we applied a maximum FDR of 1% separately on the protein and peptide level. We required 2 or more unique/razor peptides for protein identification and a ratio count of 2 or more for protein quantification per replicate measurement.

Integration of proteomic and transcriptomic data

Several transcripts may be encoded by the same locus (as alternative isoforms) but encode the same (or highly similar) proteins. Furthermore, several distinct proteins may not be confidently distinguished by mass spectrometry data. Since our analysis relies on combining information

from the RNA and protein measurements within one model, we first had to map proteins and transcripts despite these possible many-to-many relationships. To this end, we constructed "analysis groups" using the following rules:

1. All transcripts in the same transcript group are in the same analysis group, where transcript groups are defined by the UCSC Table Browser (query settings: assembly="July 2007 (NCBI37/mm9)", group="Gene and Gene Predictions", track="UCSC Genes", table="knownIsoforms"; see the fields "clusterID" and "transcript") and represent groups of transcripts (isoforms) that are derived from the same genic locus and cannot easily be resolved.

2. All proteins in the same MaxQuant protein group are in the same analysis group, where MaxQuant determines protein groups on the basis of the peptide library supplied. The groups represent the level of ambiguity MaxQuant can confidently resolve.

3. If a protein and a transcript are associated with one another, then they are in the same analysis group, where we associate (one or more) Uniprot protein ID with (one or more) UCSC transcript ID based on ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/MOUSE_10090_idmapping.dat

These grouping rules were executed using the clusters() function in the R package igraph. All analysis groups had at least one associated transcript. Some analysis groups had no associated proteins (non-coding RNAs).

RNA-seq read mapping of LPS and MOCK time course

We created a Bowtie15 index based on annotated transcripts from UCSC mm9 and aligned paired-end reads directly to this index using Bowtie v 0.12.7 with command line options: -q --phred33-quals -n 2 -e 99999999 -l 25 -I 1 -X 1000 -a –m 200. Next, we ran RSEM v1.1117 (with default parameters) and used the analysis groups (described above) to define which isoforms belong to the same gene. Analysis groups with no associated proteins (*i.e.* noncoding RNAs) were still used as input to avoid incorrect mappings. Expression quantifications were made for each analysis group, replicate, and condition.

Processing of protein dynamics data

Expression estimates were based on MaxQuant's un-normalized M/L and H/L ratios per protein group. We did not use normalized ratios because at early time points the H/L ratio is expected to be very low and normalized MaxQuant ratios would not represent this low ratio, as it assumes the global median H/L ratio (or M/L) to be 1. This is the case for classical SILAC experiments, but not for *pulsed* SILAC approaches. For each protein and replicate, we scaled the ratio by the mean IBAQ L channel intensity observed for that protein across all conditions and time points (treating missing values as zeroes). These scaled ratios represent M channel and H channel intensities and are the basis of all downstream quantitative analysis.

Data were merged (by addition) to the analysis group level. Analysis groups were named according to the concatenation of all unique gene symbols associated with the UCSC transcripts in the analysis group. For the vast majority (3,270, 88%) of analysis groups, this was only a single gene symbol.

Merging and filtering of RNA and protein data

Analysis groups (henceforth referred to as "genes" for simplicity) were excluded if they were not MS2-identified in at least 6 of the 10 time points in all conditions (time courses for LPS and MOCK stimulation), channels, and replicates (2,609 genes (45%) excluded). After applying this filter, we also excluded 24 genes (0.8%) that did not have positive RNA-Seq values for at least 6 out of the 10 time points in all conditions and channels. 5/24 were histones, whose transcripts are not polyadenylated and hence cannot be analyzed with our data.

These procedures reduced the dataset to 3,147 genes, each linked to an RNA expression time course and protein expression time course. We renormalized such that at each condition/time point the RNA-Seq expression values added up to exactly 1,000,000. Meanwhile, we normalized the protein data such that at each condition/time point protein (M+H) IBAQ values added up to exactly 1,000,000 (missing values were treated as zeroes). We refer to these as transcripts per million (TPM) and IBAQ microshares (IMS), respectively. Except where otherwise noted all analyses and modeling used these units. Figures and tables presenting protein abundances, RNA abundances, or translation rates apply linear scaling factors (derived in a later section "**Rescaling to molecules per cell**") such that these values can be interpreted, approximately, in terms of protein and RNA molecules.

RNA smoothing

To address zero-values in the RNA-Seq time-series, we estimated a detection limit for the experiment, and added this detection limit to every RNA-Seq measurement. To estimate the detection limit, we identified all genes with a mix of zero and positive expression values, and for

each of these genes we identified the smallest non-zero expression value. The detection limit was estimated as the median of all these values, which was 0.1. After the addition of the detection limit, the RNA data was again rescaled such that it summed to 1,000,000 at each condition/time point.

Each gene's time course profile was fit with a 3-point spline (defined using the R function *spline()*, method "natural"). This was necessary so that subsequent optimizations would not encounter numerous local minima. The spline was fit by minimizing squared errors in linear space (not log space). The first and third spline points were constrained to have zero slope and to occur at t=0 and t=12 respectively. The second spline point could occur anywhere between t=0 and t=12. For a given gene, the value of the first spline point was constrained to be the same for LPS and MOCK.

Estimating label switch purity

Ideally, all newly-synthesized proteins would incorporate H label; however, due to incomplete label switch and the degradation of M-labeled proteins there is always a nonzero contamination of M-labeled amino acids in the free amino acid (AA) pool.

We estimated this contamination at each time point by considering the ratio of doubly labeled MH to HH peptides (*e.g.* peptides with one missed cleavage site) among the MaxQuant identifications. Since H label is only present at t>0, all MH and HH peptides can be assumed to have been produced at t>0. In addition, it was assumed that the impurity rate among the newly-synthesized proteins at time *t* would approximate the impurity rate in the free amino acid pool at time *t*.

From the MH/HH ratio we derived the contamination rate as follows:

$MH/HH=2\cdot \gamma\cdot(1-\gamma)/(1-\gamma)^2 = 2\cdot \gamma/h = 2\cdot\gamma/(1-\gamma)$

where γ is the fraction of M-labeled AAs (so that $1-\gamma$ is the fraction of H-labeled AAs), MH is the count of doubly labeled MH peptides, and HH is the count of doubly labeled HH peptides

We derive:

$\gamma=(MH/HH)/(MH/HH + 2)$

We empirically chose a functional form for $\gamma(t)$ that fit the data with few parameters:

$\gamma_t = e^{C_1\cdot t+C_2\cdot\sqrt{t}}$.

We optimized C1 and C2 by maximizing the likelihood of a binomial model of the MH and HH counts:

$$\prod_{t=0\ldots12} \mathrm{P}(\mathrm{MH}_t = \mathrm{x} \mid \mathrm{x} \sim \mathrm{Binom}(\mathrm{n} = \mathrm{MH}_t + \mathrm{HH}_t, \mathrm{p} = 2\cdot\gamma_t/(1-\gamma_t)))$$

Because MH/HH is derived from the set of newly translated proteins rather than the free AA pool, our estimates probably lag (and therefore slightly overestimate) the true impurity rates in the free AA pool.

Dynamic model

We formulate the dynamic model per gene as a system of differential equations:

$$\frac{dM_{ij}}{dt} = R_{ij}(t)\cdot T_{ij}(t)\cdot\gamma_j(t) - D_{ij}(t)\cdot M_{ij}(t)$$

$$\frac{dH_{ij}}{dt} = R_{ij}(t) \cdot T_{ij}(t) \cdot (1 - \gamma_j(t)) - D_{ij}(t) \cdot H_{ij}(t)$$

Where:

$M_{ij}(t)$ is the M channel signal for gene *i* in condition *j* at time t

$H_{ij}(t)$ is the H channel signal for gene *i* in condition *j* at time t

$R_{ij}(t)$ is the RNA expression for gene *i* in condition *j* at time t

$T_{ij}(t)$ is the translation rate of gene *i* in condition *j* at time t

$D_{ij}(t)$ is the degradation rate of gene *i* in condition *j* at time t

$\gamma_j(t)$ is the contamination in the free AA pool in condition *j* at time *t* defined above

The system can be solved as

$$M_{ij}(t) = e^{-\tilde{D}(t)} \cdot \left(M_{0i} + \int_0^t R_{ij}(x) \cdot \gamma_j(x) \cdot T_{ij}(x) \cdot e^{\tilde{D}(x)} \cdot dx \right)$$

$$H_{ij}(t) = e^{-\tilde{D}(t)} \cdot \left(H_{0i} + \int_0^t R_{ij}(x) \cdot (1 - \gamma_j(x)) \cdot T_{ij}(x) \cdot e^{\tilde{D}(x)} \cdot dx \right)$$

where

$$\tilde{D}(x) = \int_0^x D(z) \cdot dz$$

and where

M_{0i} is the initial M channel signal from protein *i* (same in both conditions)

H_{0i} is the initial H channel signal from protein *i* (always assumed to equal zero)

However, background signal in MS1 leads to positive estimates of isotope abundance even when the particular isotope is absent (one effect of this is the presence of positive H channel readings at *t*=0), because the MaxQuant algorithm integrates the random background signal that happens to fall within the expected m/z interval. Accordingly, we also account for a background term B_i for each gene:

$$M_{ij}(t) = e^{-\tilde{D}(t)} \cdot \left(M_{0i} + \int_0^t R_{ij}(x) \cdot \gamma_j(x) \cdot \tilde{T}_{ij}(x) \cdot e^{\tilde{D}(x)} \cdot dx \right) + B_i$$

$$H_{ij}(t) = e^{-\tilde{D}(t)} \cdot \left(H_{0i} + \int_0^t R_{ij}(x) \cdot \left(1 - \gamma_j(x)\right) \cdot \tilde{T}_{ij}(x) \cdot e^{\tilde{D}(x)} \cdot dx \right) + B_i$$

$T(t)$ is defined by three parameters

T_0 is the translation rate at baseline [IMS/(TPM·hr)] and is constrained to be the same for both LPS and MOCK

T_C is the fold change in translation rate at 12h in the MOCK (control) condition

T_L is the fold change in translation rate at 12h in the LPS condition

In a given experimental condition (LPS or MOCK), it is assumed that *T(t)* changes linearly over time from T_0 to its 12h value.

$D(t)$ is likewise defined by three parameters

D_0 is the degradation rate at baseline [1/hr] and is constrained to be the same for both LPS and MOCK

D_C is the fold change in degradation rate at 12h in the MOCK (control) condition

D_L is the fold change in degradation rate at 12h in the LPS condition

In a given experimental condition (LPS or MOCK), it is assumed the *D(t)* changes linearly over time from D_0 to its 12-hour value.

Fitting of parameters by empirical Bayes

All 8 free parameters (T_0, T_C, T_L, D_0, D_C, D_L, M_0, and B) were determined through iterative empirical Bayes fitting. During fitting, all parameters (and their uncertainty distributions) were handled in log-space (*i.e.* we modeled the log of the degradation rate rather than the degradation rate itself) because the log transformation aids optimization and because all the parameters appeared to be approximately log-normally distributed across genes. Meanwhile, model error terms were always considered on a linear scale.

In a given round of fitting, a maximum *a posteriori* (MAP) estimate was determined for the parameters describing each gene. Posterior estimates combine information from the prior and the likelihood density curves.

$$Posterior(\boldsymbol{p_i}) = Prior(\boldsymbol{p_i}) \cdot L(\boldsymbol{p_i})$$

The MAP estimate is the set of values $\boldsymbol{p_i}$ for a gene that maximizes the above expression.

The likelihood $L(\boldsymbol{p_i})$ is defined by the error in the model fits and the estimated noise in the data:

$$L(\boldsymbol{p_i}) = \prod_{j=LPS,MOCK} \prod_{t=0\ldots12} (P\left(M_{i,j,t} = x \mid x \sim N(\mu = \widehat{M}_{i,j,t}, \sigma^2 = E_{i,m})\right) \cdot P\left(H_{ijt} = x \mid x \sim N(\mu = \widehat{H}_{ijt}, \sigma^2 = E_{ih})\right))$$

where

$\boldsymbol{p_i}$ is the vector of model parameters describing gene *i*

$M_{i,j,t}$ is the observed medium channel intensity of gene *i* in condition *j* at time *t*

$\widehat{M}_{i,j,t}$ is the medium channel intensity predicted given the model parameters

$E_{i,j,m}$ is the variance in the M channel signal due to experimental noise

$H_{i,j,t}$ is the observed heavy channel intensity of gene *i* in condition *j* at time *t*

$\widehat{H}_{i,j,t}$ is the heavy channel intensity predicted given the model parameters

$E_{i,j,h}$ is the variance in the H channel signal due to experimental noise

The noise term $E_{i,m}$ is determined by fitting quadratic functions to the LPS and MOCK M channel time series (via linear regression) and averaging the mean squared error (MSE) of the two fits. The noise term $E_{i,h}$ is determined analogously from the H channel.

Missing values were dropped and did not contribute to the likelihood calculations.

The prior for a given gene's parameter set $\boldsymbol{p_i}$ is obtained by multiplying the probabilities of each parameter (T_{0i}, T_{Ci}, T_{Li}, D_{0i}, D_{Ci}, D_{Li}, M_{0i}, and B_i) as defined by their respective prior distributions:

$$Prior(\boldsymbol{p_i}) = \prod_{p_i=T_{0i}\ldots B_i} P(p_i = x \mid x \sim N(\mu = mean(\boldsymbol{p_{-1}}), \sigma^2 = var(\boldsymbol{p_{-1}})))$$

where

$\boldsymbol{p_{-1}}$ is the vector of MAP values (across all genes) for the given parameter in the previous fitting iteration.

Thus, the prior represents our expectations (in the form of a parameterized Normal distribution) for a given gene's parameters considering what has been observed for all other genes. This is the "empirical Bayes" component.

Since we wanted to be able to draw inference on T_L *vs.* T_C and on D_L *vs.* D_C, we shared priors for these parameters, such that there was a single prior for translation rate fold change and a single prior for degradation rate fold changes. For example, the prior for T_C was:

$$N(\mu = \frac{mean(\boldsymbol{T_{c_{-1}}}) + mean(\boldsymbol{T_{L_{-1}}})}{2}, \sigma^2 = \frac{var(\boldsymbol{T_{c_{-1}}}) + var(\boldsymbol{T_{L_{-1}}})}{2})$$

which was also used as the prior for T_L.

In the first fitting iteration, the prior for each parameter was based on a distribution of heuristic estimates for the parameter values. Three fitting iterations were applied, which was sufficient for near-convergence. All posterior maximizations were conducted using the optim() function in R using method BFGS.

Posterior distributions were characterized through inversion of the Hessian of the posterior probability function at its MAP value. This approach was used for establishing credible intervals for individual parameters and for determining $P(D_L>D_C)$ and $P(T_L>T_C)$ for individual genes.

Rescaling to molecules per cell

To aid interpretation, we calculate scaling factors so that model estimates could be interpreted in terms of protein and mRNA molecules rather than microshares and TPM. Comparing the model estimates for M_{0i} to previously curated (*14*) absolute per-cell abundances of 37 overlapping genes, we estimated a scaling factor that minimized the mean log fold discrepancy between our estimates and the curated values. Based on this approach, we determined that replicate 1 and replicate 2 protein microshares should be multiplied by 8,641 and 8,391, respectively, to express protein levels in terms of molecules. For RNA, we used the previous estimate of 2.6×10^{-15} moles nucleotides per cell (mouse fibroblasts; (*11*)), and an average transcript length of 1,500-2,000 nucleotides, to determine that each cell should contain approximately 0.75-1.1 million mRNA transcripts. Therefore, we consider 1 TPM as roughly equivalent to 1 transcript per cell. We emphasize that these are rough estimates only, and rely on general assessments in the literature.

Determining contributions to total protein expression at baseline

Estimates of percent contribution to baseline protein expression were derived by comparing different model-based estimates of protein level to independent proteomic measurements, measured in separate experiments (not those used to fit the model) and with two distinct digestions (see below). Overall, to generate a predicted protein level, we used our model where

the parameters for one or more of the three regulatory step – RNA levels, translation rate, and degradation rate – were set based on their per-gene fit and the others were set for their *median* value across the fits for all genes. In each case we calculated the correlation between the model's prediction and the measurements (see below on how correlation is calculated). By considering the improvement in this correlation as more parameters are set to their per-gene fit, we can estimate their contribution to protein level. For example, comparing the model's ability to predict using only per-gene fits of RNA level (but with translation and degradation rates set to their pan-genome median) *vs*. its quality of prediction based on per-gene fits of RNA and translation rates (with degradation set at pan-genome median), we can determine the contribution of translation rates.

Because RNA levels, translation rates, and degradation rates are not fully orthogonal to each other and because these variables interact in a non-linear manner, explanatory contribution is dependent on the order in which per-gene parameters are allowed into the model. The order we present in **Figure 3** (RNA, translation, degradation) was chosen because it follows the temporal ordering of the protein life cycle and, roughly, accessibility to measurement. It is also consistent with previous computational approaches to this question (Li et al., 2014; Schwanhäusser et al., 2011). Nonetheless, we have explored the results in the context of the other five possible orderings and present these in **Figure S11**. Below we provide full details on this analysis.

To quantify the explanatory capacity of each process we relied on the framework of squared Pearson correlation coefficient, which we treat analogously to the "percent variance explained" or "R^2" term commonly used in the context of linear regression.

Because of stochastic and protocol-dependent errors, we would not expect a perfect correlation between RNA and protein measurements even within the context of a system in which protein level is 100% RNA-dependent. To correct for this, we employed a statistical adjustment, known as Spearman correction (not related to "Spearman" rank correlation), which can estimate the true correlation of two variables *X* and *Y* that are observed as error-prone (but independently distributed) replicate estimates $\tilde{X}_1$, $\tilde{X}_2$, and $\tilde{Y}_1$, $\tilde{Y}_2$. To achieve this estimate, the nominal correlation between X and Y is estimated as the geometric mean of the pairwise correlations $cor(\tilde{X}_1, \tilde{Y}_1)$, $cor(\tilde{X}_1, \tilde{Y}_2)$, $cor(\tilde{X}_2, \tilde{Y}_1)$, and $cor(\tilde{X}_2, \tilde{Y}_2)$ corrected by the geometric mean of the estimated reliabilities of the X and Y measurements, which are $cor(\tilde{X}_1, \tilde{X}_2)$ and $cor(\tilde{Y}_1, \tilde{Y}_2)$, respectively. The Spearman-corrected Pearson correlation may be expressed as:

$$\frac{g(r(\tilde{X}_1, \tilde{Y}_1), r(\tilde{X}_1, \tilde{Y}_2), r(\tilde{X}_2, \tilde{Y}_1), r(\tilde{X}_2, \tilde{Y}_2))}{g(r(\tilde{X}_1, \tilde{X}_2), r(\tilde{Y}_1, \tilde{Y}_2))}$$

where $r(\cdot)$ represents the Pearson correlation of the (log-transformed) input variables and where $g(\cdot)$ represents a geometric mean function. Squaring this correlation expression yields an error-corrected measure of fraction variance explained.

In our case, the underlying variables are model-estimated protein level (based on RNA alone, based on RNA and translation, or based on all three channels) and actual protein level. Estimates of actual protein level were made from *independent experiments* that used LysN or

AspN peptide library preparation protocol (otherwise experimental procedures were identical to those described above for the trypsin-prepared data used to fit the original model). This approach is necessary to satisfy the requirement that error terms be independent. Using trypsin estimates as the measure of actual protein level would be inappropriate because the model-based translation rates are also based on trypsin preparation, which would cause library protocol-dependent errors to artificially inflate the contribution of translation (as both translation rate estimates and total protein level estimates would contain the same trypsin-dependent error per gene). As a quality control measure, we also compared the protein levels estimated by the different digestions and our model for 61 previously determined 'standards' (Li et al., 2014) and show a good correspondence in all cases (**Figure S21**). Note that transcription rates and degradation rates are not sensitive to this issue as RNA is quantitated by an orthogonal system and degradation is inherently on a relative scale (so library-dependent degradation rate error terms should be small). A limitation of this approach is that it does not address systematic errors in RNA-Seq quantitation, *e.g.* biases toward highly sequenceable genes. This may result in an underestimation of the transcriptional contribution to total protein level.

More formally, we define the following variables:

R_1, R_2: Smoothed estimates for baseline (t_0) RNA abundance (two replicates)

T_1, T_2: Model-based estimates of baseline (t_0) translation rates (two replicates)

D_1, D_2: Model-based estimates of baseline (t_0) degradation rates (two replicates)

L_1, L_2: Measured IBAQ values from baseline sample prepared with LysN digest (two replicates)

A_1, A_2: Measured IBAQ values from baseline sample prepared with AspN digest (two replicates)

At steady state, baseline protein expression can be estimated as the product of RNA abundance and per-mRNA translation rate divided by degradation rate. (Of the different possible order of consideration, we describe the case where RNA is considered first, then RNA and translation and then all three steps. We assessed contributions in all other possible orderings in an analogous manner.) We defined three model-based estimates of protein abundance for replicate 1. The first (P_{R1}) utilized per-gene RNA abundance estimates but substituted in pan-genome medians for translation and degradation. The second (P_{RT1}) utilized per-gene RNA abundances and translation rates but held degradation at its median value. The third (P_{RTD1}) used per-gene estimates for all three variables.

$$P_{R1} = R_1 \cdot median(T_1)/median(D_1)$$

$$P_{RT1} = R_1 \cdot T_1/median(D_1)$$

$$P_{RTD1} = R_1 \cdot T_1/D_1$$

Analogous formulas were used to generate P_{R2}, P_{RT2}, and P_{RTD2} for replicate 2.

To obtain the fraction of protein level variance explained by RNA, we calculated a Spearman-corrected correlation and squared it:

$$X_R = \left(\frac{g(r(P_{R1}, L_1), r(P_{R1}, L_2), r(P_{R1}, A_1), r(P_{R1}, A_2), r(P_{R2}, L_1), r(P_{R2}, L_2), r(P_{R2}, A_1), r(P_{R2}, A_2))}{g(r(P_{R1}, P_{R2}), g(r(L_1, A_1), r(L_1, A_2), r(L_2, A_1), r(L_2, A_2)))}\right)^2$$

Correlations were assessed based on the set of overlapping genes between the two experiments, which was ~75% of genes when data originated from different peptide libraries.

To obtain the additional contribution of translation to protein level variance, we calculated:

$$X_T = \left(\frac{g(r(P_{RT1}, L_1), r(P_{RT1}, L_2), r(P_{RT1}, A_1), r(P_{RT1}, A_2), r(P_{RT2}, L_1), r(P_{RT2}, L_2), r(P_{RT2}, A_1), r(P_{RT2}, A_2))}{g(r(P_{RT1}, P_{RT2}), g(r(L_1, A_1), r(L_1, A_2), r(L_2, A_1), r(L_2, A_2)))}\right)^2 - X_R$$

To obtain the additional contribution of degradation to protein level variance, we calculated:

$$X_D = \left(\frac{g(r(P_{RTD1}, L_1), r(P_{RTD1}, L_2), r(P_{RTD1}, A_1), r(P_{RTD1}, A_2), r(P_{RTD2}, L_1), r(P_{RTD2}, L_2), r(P_{RTD2}, A_1), r(P_{RTD2}, A_2))}{g(r(P_{RTD1}, P_{RTD2}), g(r(L_1, A_1), r(L_1, A_2), r(L_2, A_1), r(L_2, A_2)))}\right)^2 - X_R - X_T$$

In our analysis, X_R, X_T, and X_D summed to less than 1. This is due to under-estimation of correlations in the numerators (a possible result of model misspecification) and/or over-estimation of reliabilities in the denominators (a possible result of correlated error terms – these may persist due to aforementioned biases in the RNA-Seq).

In addition to estimating the contribution of each regulatory step across all genes, we also estimated the contribution of RNA level, translation rate, and degradation rate to protein levels on a per-gene basis (**Table S5**). These were defined as $\frac{P_{RTD}}{P_{TD}} = R/median(R)$, $\frac{P_{RTD}}{P_{RD}} = T/median(T)$, and $\frac{P_{RTD}}{P_{RT}} = median(D)/D$, respectively, which may be interpreted as the relative expression difference due to the non-centrality of the given regulatory step.

Determining contributions to dynamic fold changes in protein expression

We used an analogous approach to calculate the contribution of changes in RNA level, translation rate, and degradation rate to LPS-induced changes in total protein level. As before, we considered this in all possible orderings, and present below the ordering of (RNA, translation, degradation).

The following variables were calculated based on data from replicate *i* (*i=1,2*):

P_{Ri}: fold change between two model-based *estimates* of the 12-hour protein level: the first estimate uses the RNA level trajectory fit for LPS conditions and translation/degradation rates estimated under MOCK conditions; the second estimate uses only MOCK-derived parameters.

P_{RTi}: fold change between two model-based *estimates* of 12-hour protein level: the first estimate uses LPS-derived RNA levels and translation rates but control-derived degradation rates; the second estimate uses only control-derived parameters.

P_{RTDi}: fold change between two model-based *estimates* of 12-hour protein level: the first estimate uses LPS-derived parameters exclusively; the second estimate uses only control-derived parameters.

Since the dynamic response is quantified on the basis of ratios (between LPS and control at 12 hours), there is less concern that correlated error terms will arise from systematic biases. Therefore, we dispensed with collecting non-trypsin fold change estimates and simply calculated correlations and reliabilities by comparing (sub-)model estimates from one replicate *i* full model estimates from the other replicate *j*:

$$X_R = \left(\frac{g(r(P_{Ri}, P_{RTDj}), r(P_{Rj}, P_{RTDi}))}{g(r(P_{Ri}, P_{Rj}), r(P_{RTDi}, P_{RTDj}))}\right)^2$$

$$X_T = \left(\frac{g(r(P_{RTi}, P_{RTDj}), r(P_{RTj}, P_{RTDi}))}{g(r(P_{RTi}, P_{RTj}), r(P_{RTDi}, P_{RTDj}))}\right)^2 - X_R$$

$$X_D = \left(\frac{g(r(P_{RTDi}, P_{RTDj}), r(P_{RTDj}, P_{RTDi}))}{g(r(P_{RTDi}, P_{RTDj}), r(P_{RTDi}, P_{RTDj}))}\right)^2 - X_R - X_T = 1 - X_R - X_T$$

Note that a side effect of this approach is that explanatory contributions will necessarily add to 100%.

We also estimated the contribution of changes in RNA level, translation rate, and degradation rate on a per-gene basis (**Table S5**). These were defined as $\frac{P_{RTD}}{P_{TD}}$, $\frac{P_{RTD}}{P_{RD}}$, and $\frac{P_{RTD}}{P_{RT}}$ (dropping replicate subscripts), where variables are defined as above with the addition of:

P_{TD}: fold change between two model-based estimates of 12-hour protein level: the first estimate uses LPS-derived fits of translation rates and degradation rates but MOCK-derived RNA levels; the second estimate uses MOCK-derived parameters exclusively.

P_{RD}: fold change between two model-based estimates of 12-hour protein level: the first estimate uses LPS-derived fits of RNA levels and degradation rates but MOCK-derived translation rates; the second uses MOCK-derived parameters exclusively.

To determine whether there was a relationship between the magnitude of a gene's fold change and the regulatory process driving it, genes were divided into 10 equally sized bins according to their total fitted 12-hour protein fold change (M+H), and the median contribution of each regulatory channel was calculated per bin (**Figure S12**).

Determining contributions to absolute changes in protein expression

To estimate the contributions of changes in RNA levels, translation rates, and degradation rates to absolute changes in protein expression (*ie.* as measured in terms of IBAQ units rather than relative fold change), we re-defined P_{R1}, P_{R2}, P_{RT1}, P_{RT2}, P_{RTD1}, and P_{RTD2} in terms of subtraction rather than division, and X_R, X_T, and X_D were re-calculated using the formulas above. Given that P_{R1}, P_{R2}, P_{RT1}, P_{RT2}, P_{RTD1}, and P_{RTD2} could be positive or negative, they were not log-transformed prior to the correlation calculation (as in the previous analyses). Per gene contributions were defined as $P_{RTD} - P_{TD}$, $P_{RTD} - P_{RD}$, and $P_{RTD} - P_{RT}$ (dropping the replicate subscripts; **Table S5**).

Assessing for robustness to outliers

The above contributions were all calculated based on Pearson correlations. To examine whether any of our conclusions might be substantially influenced by outliers, we repeated the analyses using Spearman rank correlation and observed comparable results (**Figure S20**).

Assessing the impact of ascertainment bias

Since our proteomic measurements are biased toward highly expressed proteins (**Figure S14**), we asked whether our derived explanatory contributions were representative of the full proteome. Since our RNA-seq data is comparatively much more sensitive (**Figure S14**), we defined an "expected proteome" at baseline based on all protein coding genes expressed greater than 2 TPM at t=0h. Logarithmically spaced RNA expression bins were defined (5 per order of magnitude; see **Figure S14**), and the number of modeled proteins was compared to the

number of expressed transcripts in each bin. We then repeated the explanatory contribution analyses using the observed protein *vs.* observed RNA count ratios as weights in the correlation calculations using the weighted correlation formula:

$$corr(x, y; w) = \frac{cov(x, y; w)}{\sqrt{cov(x, x; w)cov(y, y; w)}}$$

where,

$$cov(x, y; w) = \frac{\sum_i w_i(x_i - m(x; w))(y_i - m(y; w))}{\sum_i w_i}$$

where,

$$m(x; w) = \frac{\sum_i w_i x_i}{\sum_i w_i}$$

For example, there were 1,500 distinct genes in the 5-8 TPM RNA expression bin, but only 250 of the corresponding proteins modeled in our analysis. Thus, the correlation calculation (which considers modeled genes only) would give each of these 250 proteins a weight of 1500/250=6. In this manner, low-expression proteins were more highly weighted such that all RNA expression bins contributed equally to the calculations of contributions. It is indeed possible that all, or nearly all, bins will receive a weight >=1. But this is not a problem because the weighted correlation formula automatically normalizes by the sum of the weights, ie the weights only matter in proportion to each other (Pozzi et al., 2012). The result of the weighting is that correlation will now reflect the full expression distribution (in so far as it can be estimated from RNA levels) equally.

Ribosome profiling for baseline translation efficiency estimates in mouse DCs

Unstimulated BMDCs were incubated with 100μg/mL Cycloheximide for 1 minute at 37°C. Cells were collected by centrifugation at 300 ×g for 5 minutes at 4°C and washed twice with Cycloheximide (100μg/mL) in ice-cold PBS. Cell pellets were covered with 400μL of lysis buffer (4.75 mL polysome buffer + 250 μL 20% Triton X-100 (Sigma-Aldrich) + 60 units Turbo DNAse (Ambion)). Polysome buffer is 20mM Tris-HCl pH 7.56, 150mM NaCl, 5mM MgCl2, 100μg/mL CHX, 1mM DTT, and 8% glycerol. Lysis was carried out by triturating 10 times. Whole-cell lysates were clarified by centrifuging at 20,000 ×g for 10 minutes at 4°C. Clarified lysate was collected and flash frozen in liquid nitrogen.

Ribosome-protected fragments were isolated and cloned from lysates via RNAse I (Ambion) treatment as described previously (Ingolia et al., 2012). Samples were depleted of rRNA contamination with the Ribo-Zero rRNA Removal Kit (Epicentre) and sequenced on the Illumina HiSeq2500.

Sequencing reads were stripped of linker sequences using fastx_clipper from the FASTX-Toolkit with the following settings: -a CTGTAGGCACCATCAAT –n -Q 33. Reads mapping to miRNAs, rRNAs, snRNAs, snoRNAs, and tRNAs were filtered using Bowtie2 in --local mode (Langmead and Salzberg, 2012). Remaining reads were aligned using Tophat2 to the UCSC mm9 Known Gene transcriptome. Tophat2 (Kim et al., 2013) was run with the following settings --b2-very-sensitive --transcriptome-only --no-novel-juncs -max-multihits=64.

Footprints were filtered for read lengths 26-33 (inclusive). Footprints were mapped to their approximate P-site position by applying a 12 nucleotide offset from the 5' end. Footprint counts

were averaged over the union of genomic positions corresponding to coding sequences of the genes in each "analysis group" (defined previously in the "**Integration of proteomic and transcriptomic data**" section), excluding positions within 3 nucleotides upstream or 15 nucleotides downstream of start codons or within 9 nucleotides upstream of stop codons.

To estimate translational efficiency (TE) for each gene, the count of aligning nucleotides was divided by the length of the included region. This measure of ribosomal occupancy was renormalized to sum to 1 million across all genes. It was then divided by the RNA-seq-derived abundance (expressed in TPM; **Table S4**).

We then revisited our model for baseline contributions for protein levels, replacing our modeled per-mRNA translation rates with the footprinting-derived translational efficiencies and proceeding as previously (**Figures 3B, S10**).

Generation of heatmaps

Unless otherwise noted, all heatmaps display the 2,288 genes for which we had measurements at all time points in both replicates of the pulsed-SILAC and RNA-Seq experiment. Color intensities are determined by calculating robust z-scores ((x – median of x)/(median absolute deviation of x)) on the data in each map. The order of the genes was determined by hierarchical clustering of robust z-scored fold changes in RNA level, translation rate, and degradation (as estimated LPS/MOCK at 12 hours) using the seriate() function in the R package "seriation" (Buchta et al., 2008; Hahsler et al.).

Secretome enrichment analyses

The "Secretome" was defined as the set of genes appearing in Table S4 "Compendium of Human Secreted Proteins" in (Eichelbaum et al., 2012). The enrichment score with respect to degradation rate fold change was calculated using a Wilcoxon rank sum test.

Identification of functional gene sets with significantly high or low rates

We tested individual functional gene sets for significantly high or low values of RNA baseline, baseline translation, baseline degradation, fold change in RNA, fold change in translation, or fold change in degradation by comparing the rates for the measured genes in the gene set with those of all other measured genes using a Wilcoxon rank sum test. Enrichment level was presented using the absolute $\log_{10}$(p-value) multiplied by the sign of the association.

The gene sets queried were the collection of "Gene Ontology Annotations" from Mouse Genome Informatics (ftp://ftp.informatics.jax.org/pub/reports/gene_association.mgi) (Blake et al., 2013). Baseline RNA was assessed from the fitted values at t=0. Fold change in RNA was assessed as the average RNA expression during the 12-hour LPS experiment divided by the average RNA expression in the 12-hour MOCK experiment.

We used the following definitions for the genes in the "ribosomal," "mitochondrial," and "immune" genes in **Figure 2B** and elsewhere. Ribosomal proteins were defined strictly as the small and large subunits (gene names starting in Rps and and Rpl, excluding several kinases). Mitochondrial genes were defined as all genes appearing in Table S5 of cited (Pagliarini et al., 2008). "Immune genes" were defined as all genes appearing in any of the following gene sets:

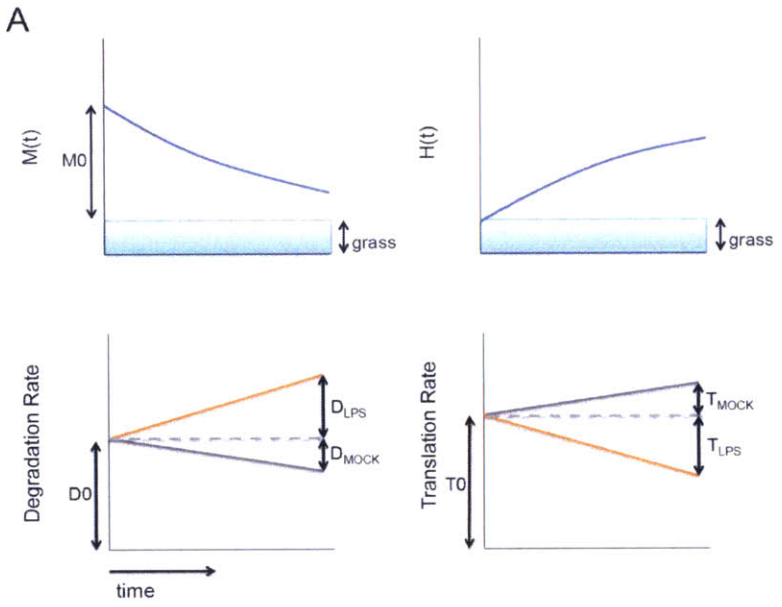
innate immune response in mucosa
activation of immune response
myeloid dendritic cell activation involved in immune response
macrophage activation involved in immune response
leukocyte activation involved in immune response
cytokine production involved in immune response
cytokine secretion involved in immune response
leukocyte migration involved in immune response
regulation of cytokine production involved in immune response
regulation of cytokine secretion involved in immune response
negative regulation of cytokine secretion involved in immune response
positive regulation of cytokine secretion involved in immune response
innate immune response-activating signal transduction
immune response-regulating signaling pathway
immune response-inhibiting signal transduction
immune response-inhibiting cell surface receptor signaling pathway
immune response-regulating cell surface receptor signaling pathway
immune response
innate immune response
regulation of innate immune response
positive regulation of innate immune response
negative regulation of innate immune response
regulation of immune response
negative regulation of immune response
positive regulation of immune response
positive regulation of myeloid leukocyte cytokine production involved in immune response

Mitochondrial to nuclear DNA ratio

We determined the mitochondrial DNA (mtDNA) / nuclear DNA (nDNA) ratio as previously described (Guo et al., 2009). Briefly, genomic DNA was extracted using the phenol-chloroform method. Primers for CO1 (forward: 5'-TGCTAGCCGCAGGCATTAC-3' and reverse: 5'-GGGTGCCCAAAGAATCAGAAC-3') and NDUFV1 (forward: 5'-CTTCCCCACTGGCCTCAAG-3' and reverse: 5'-CCAAAACCCAGTGATCCAGC-3') were used to quantify mtDNA and nDNA, respectively. Real-time PCR was performed based on SYBR Green (Roche) using a Roche LightCycler 480 sequence detection system.

Supplemental Figures

Figure S1. Dynamic Model



We fit 8 free parameters for each gene in our dynamic model. (**1**) *B:* background mass spectrometry signal; (**2**) M_0*:* initial M signal (not counting background); (**3**) D_0*:* initial degradation rate; (**4**) D_M*:* fold change in degradation rate in MOCK at 12h; (**5**) D_L*:* fold change in degradation rate in LPS at 12h; (**6**) T_0*:* initial translation rate; (**7***)* T_M*:* fold change in translation rate in MOCK at 12h; (**8**) T_L*:* fold change in translation rate in LPS at 12h.

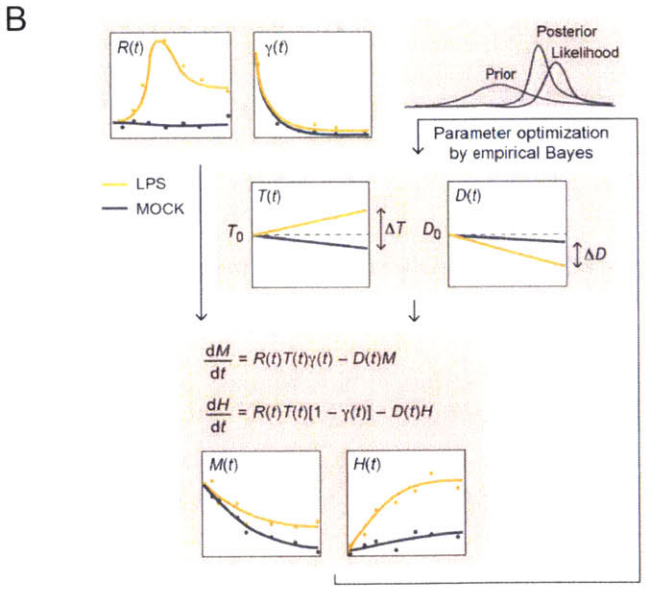
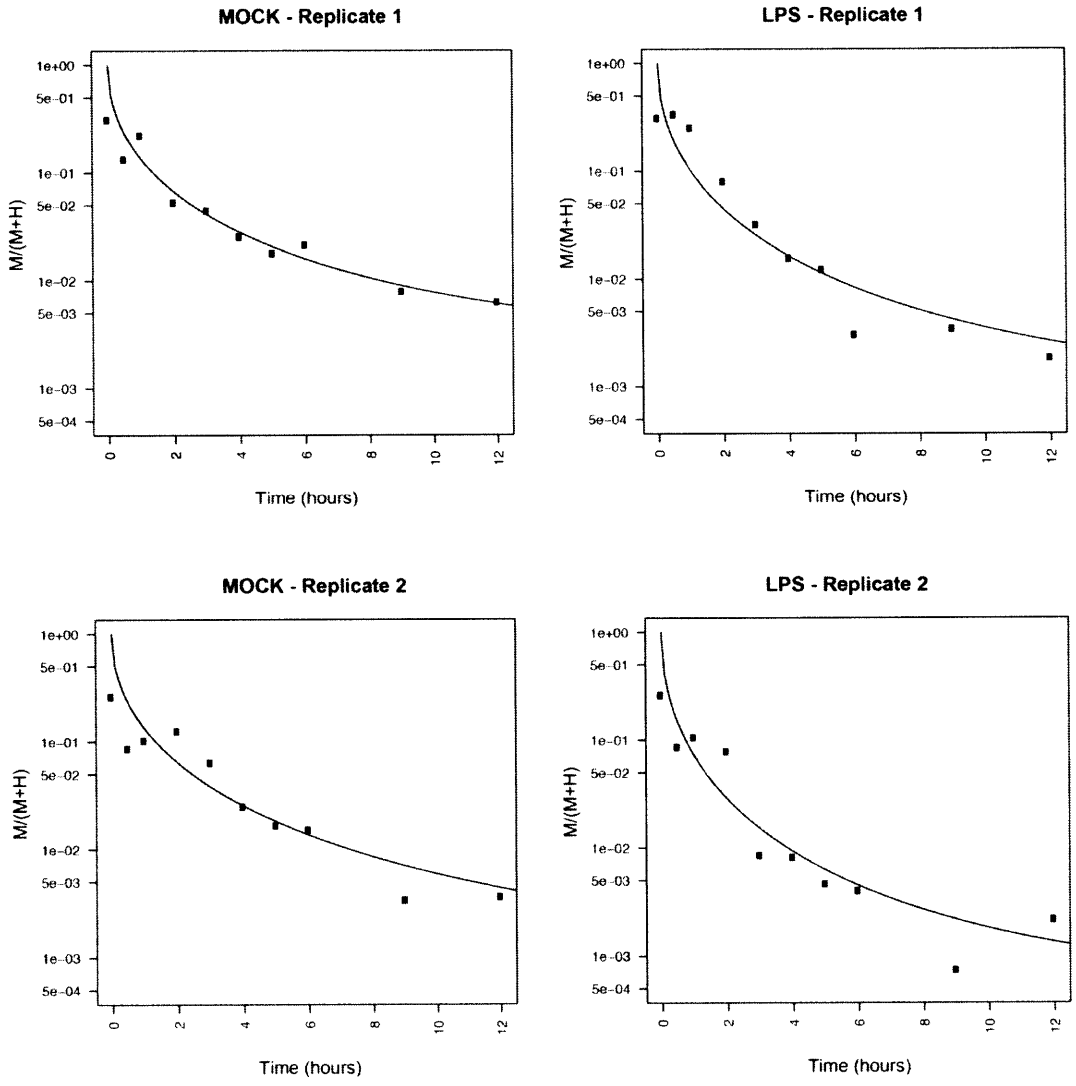
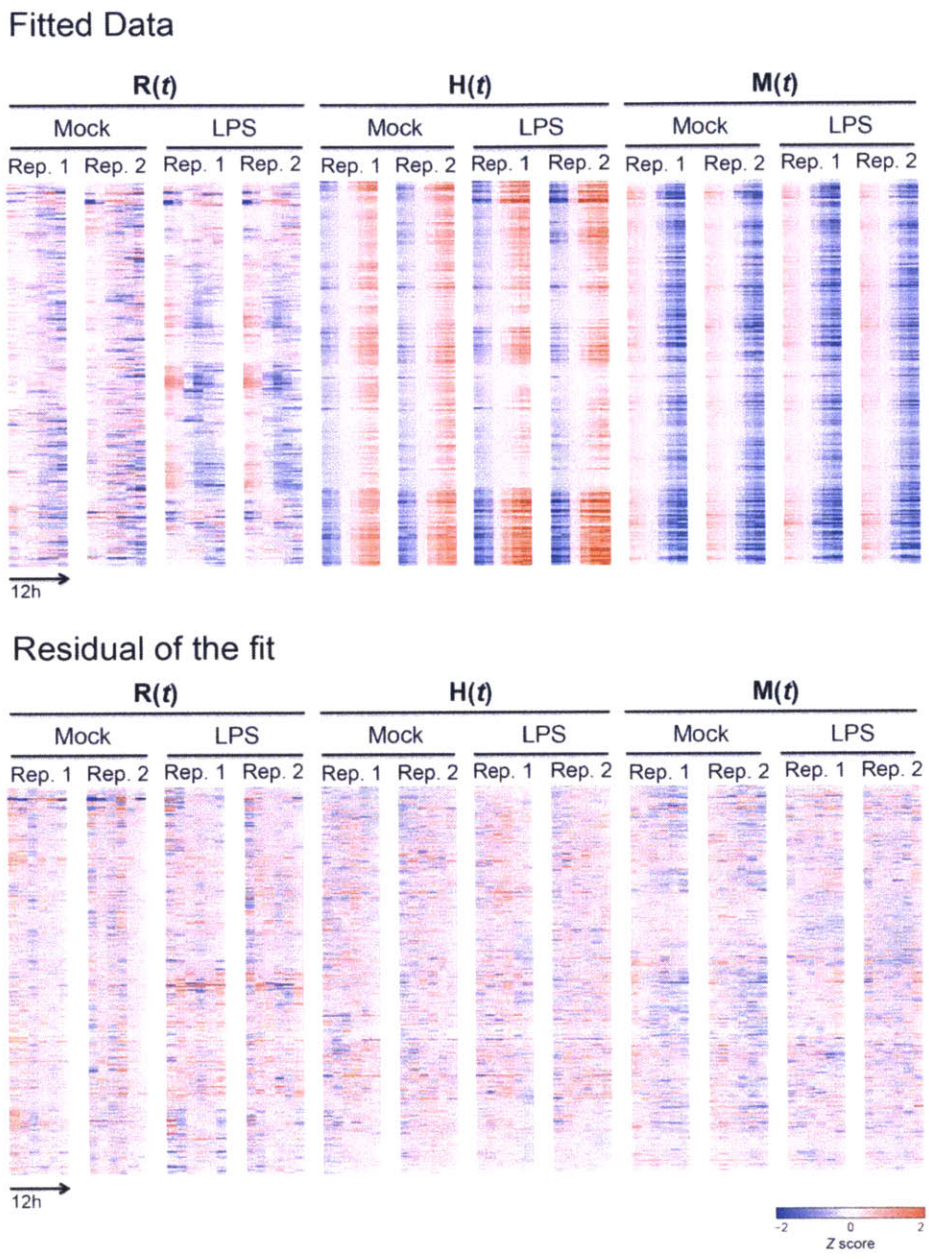


Figure S2. The recycling rate, γ(t)



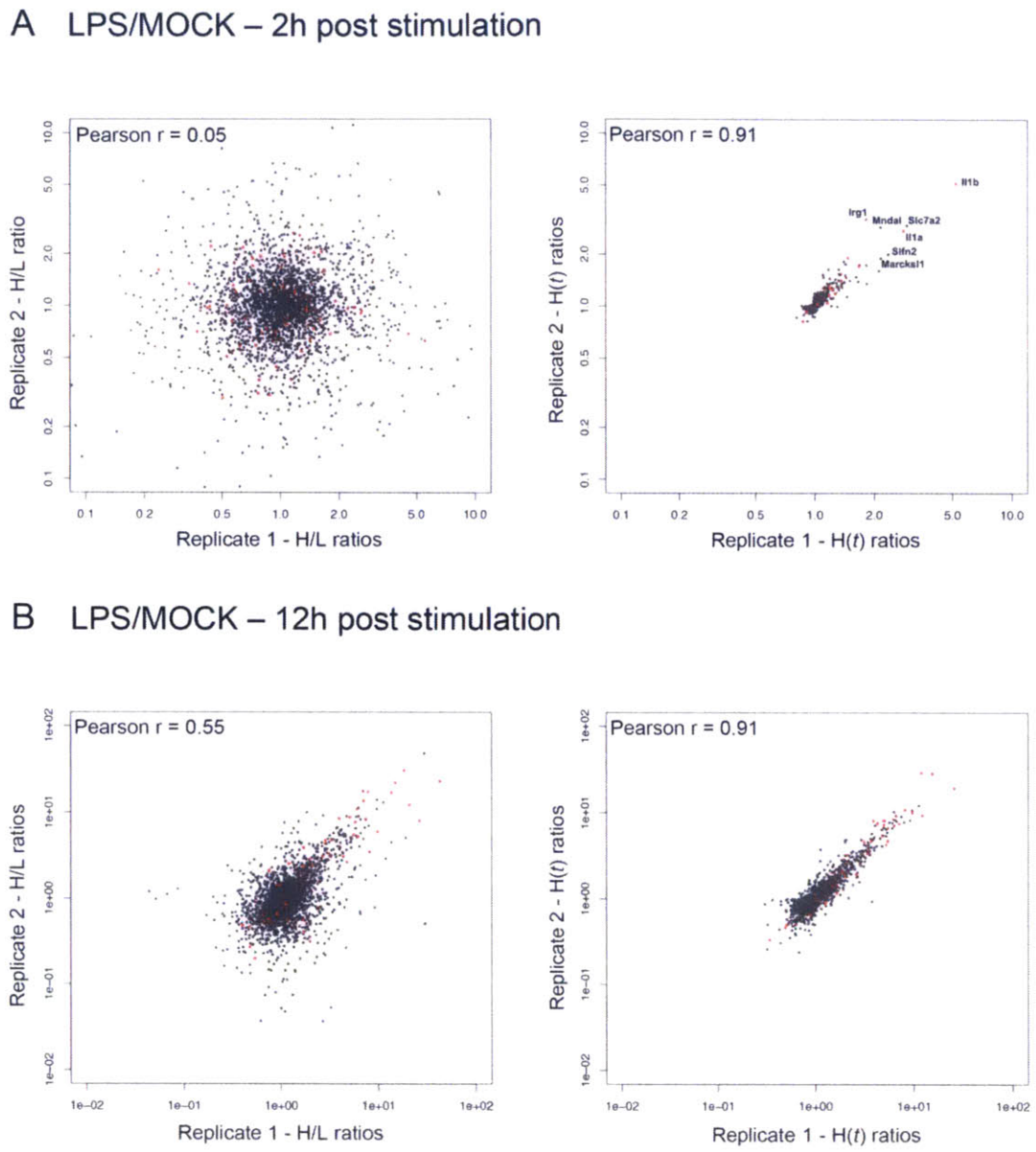
For each stimulation and replicate, shown are plots of the M(M+H) (Y axis) measured values (dots) at each time point (X axis), and the associated fit from which γ(t) was determined.

Figure S3. Fitted Data



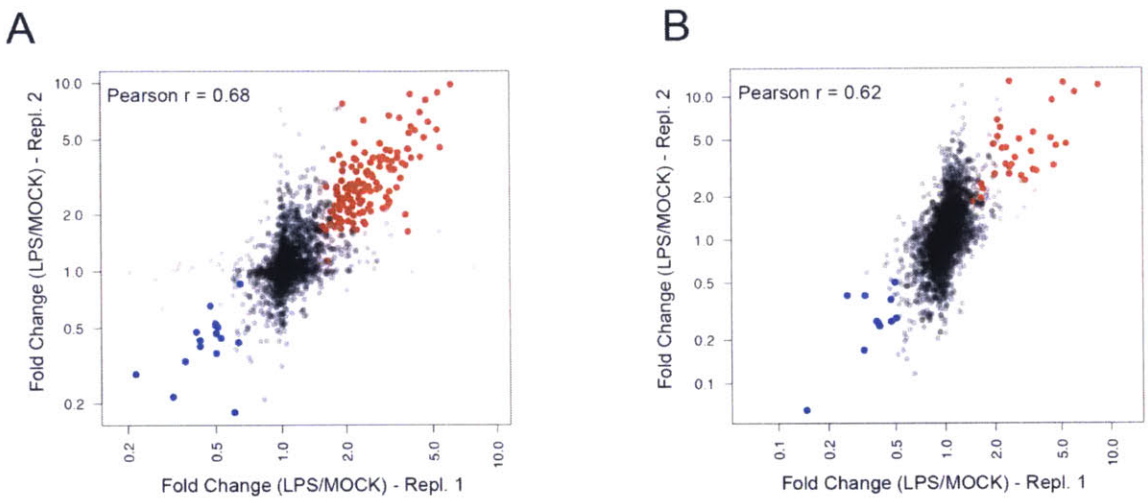
Shown are heat maps for 2,288 genes (rows) at each time point (columns) of (left to right) (**A**) fitted R(*t*), H(*t*) and M(*t*) values and (**B**) corresponding residuals (fitted/raw) for each of two replicates from MOCK- and LPS-stimulated DCs. Gene order is the same across all heatmaps, and determined by hierarchical clustering of fitted fold changes in RNA level, translation rate, and degradation rate. In (**A**) values are median-normalized by row, logged, and subjected to robust z-transformation per map; in (**B**) values are logged, converted to absolute value, and subjected to z-transformation per map (red: high, white: moderate; blue: low; see color scale).

Figure S4. Using empirical Bayes to robustly distinguish signal and noise



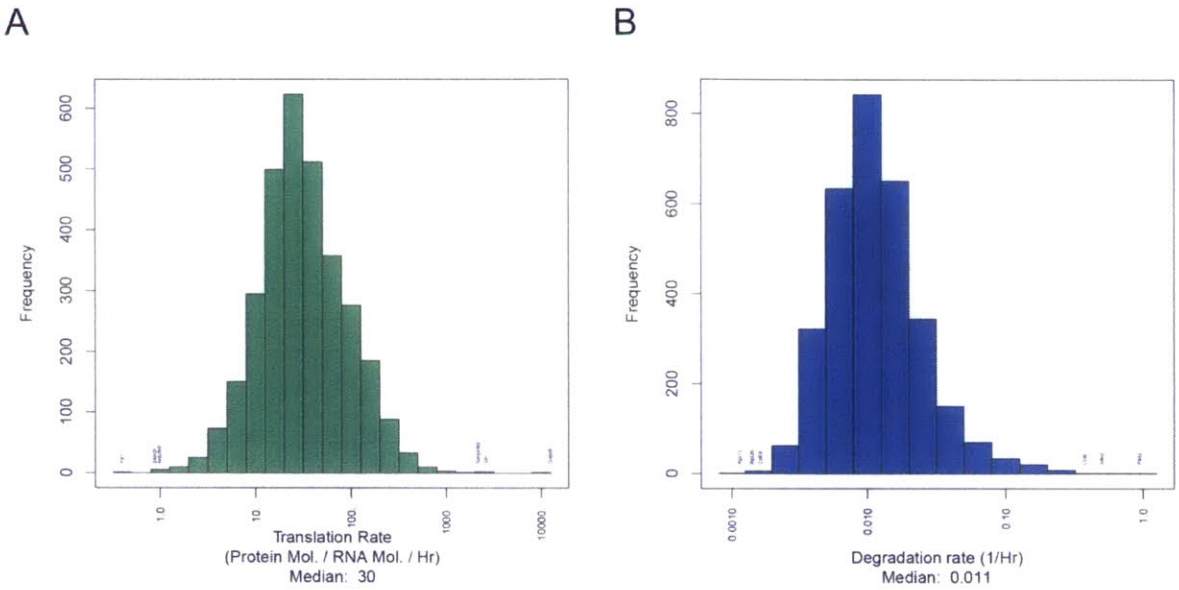
Scatter plots comparing the two replicates by either the raw H/L ratios (left) or the fitted H(t) values (right) of LPS vs. MOCK at (**A**) 2 hours and (**B**) 12 hours. The Pearson correlation coefficient (*r*) of log fold change is depicted in the top left corner of each plot. Proteins annotated with immune response-related GO terms are depicted in red. At the 2h time point, when the amount of *newly produced* protein is typically very low, protein production (H/L) has a low signal-to-noise ratio and reproducibility is much lower than at 12h. The reproducibility of fitted values is substantially better than the raw ones, due to the empirical Bayes approach, which "shrinks" unreliable estimates based on noisy data toward the population mean (while assigning them wide credible intervals). Genes with >2 fold change (in at least one replicate) at 2h post stimulation are named in (**A**). All of these genes have been implicated in immune function.

Figure S5. Reproducibility of estimated changes in production or degradation rates



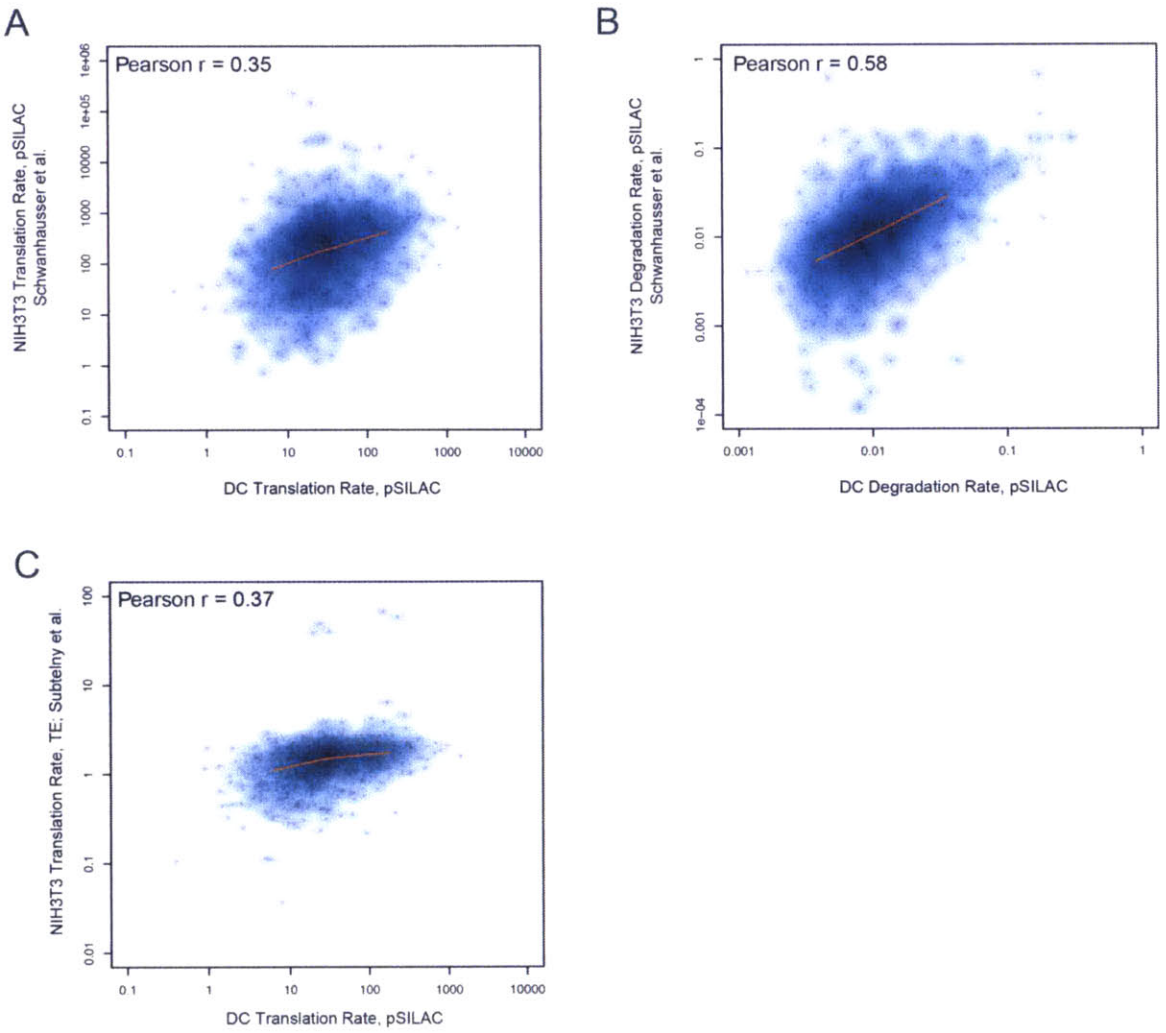
Shown are scatter plots comparing the two replicates for (**A**) per-mRNA translation rate differences ($\Delta T_i = T_i(12h)_{LPS}/T_i(12h)_{MOCK}$) and (**B**) degradation rate differences ($\Delta D_i = D_i(12h)_{LPS}/D_i(12h)_{MOCK}$) . The Pearson correlation coefficient (*r*) of log fold change is depicted in the top left corner of each plot. Red and blue dots: rates significantly up- or down-regulated upon LPS stimulation (as defined by the posterior odds of a rate increase *vs.* a rate decrease being greater than 100 or less than 0.01, respectively). Reproducibility is particularly good for significantly changing rates due to our Empirical Bayes approach.

Figure S6. Basal per-mRNA translation rates and protein degradation rates in DCs



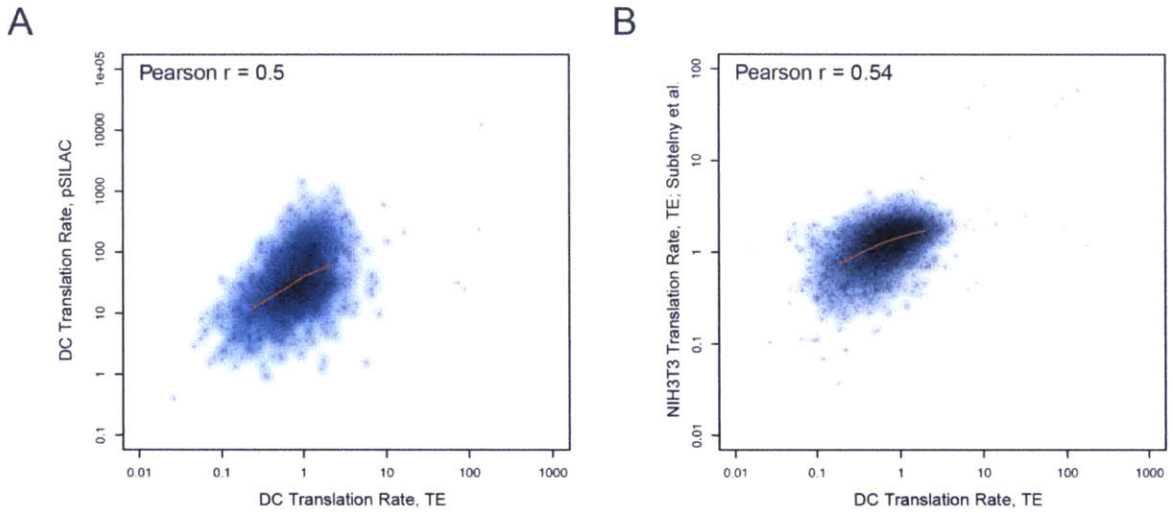
Distributions of (**A**) basal per-mRNA translation rates ($T_i(0h)$) and (**B**) basal degradation rates ($D_i(0h)$) in unstimulated (t=0h) DCs. The per-mRNA translation rate is represented as protein molecules / mRNA transcript / hour (hr).

Figure S7. Fair agreement between estimates of rates in the protein life cycle rates at steady state in mouse DCs and in previous studies in proliferating mouse cells



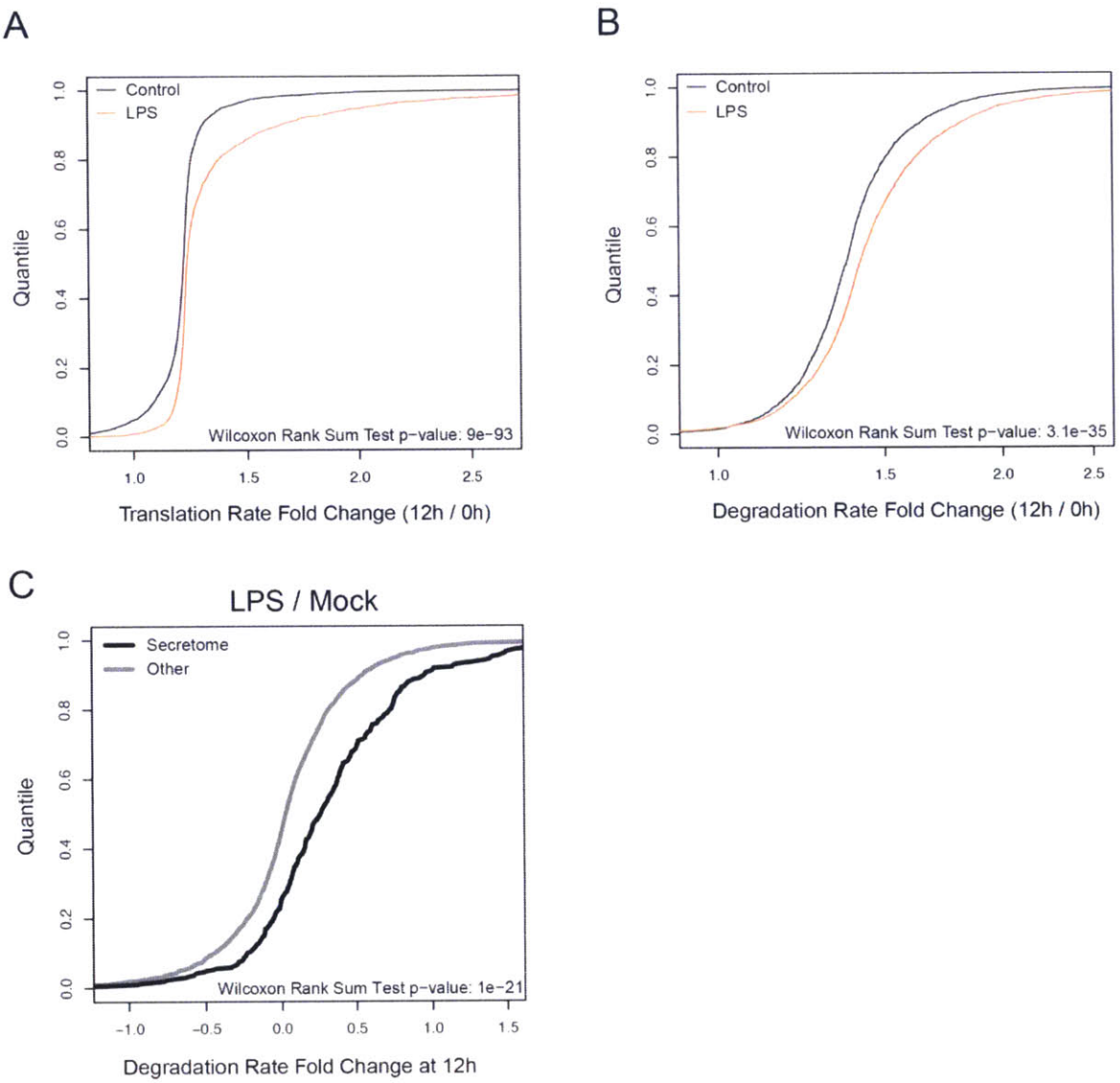
Scatter plots comparing our model-based rate estimates in resting DCs (X axis) to previous measurements in mammalian cells (Y axis). Each plot indicates the Pearson correlation (*r*) of the log-transformed rates in the upper left corner. Each gene is represented with a semi-transparent gray point. Blue shading in the background reflects local point densities (darker = more points), and the red curve represents a local regression fit spanning from the 5th to 95th percentile. In (**A**) and (**B**), per-mRNA translation rates ($T_i(0h)$) and degradation rates ($D_i(0h)$) (X axis), respectively, are compared to previously published rates for cycling mouse fibroblasts (NIH3T3 cells, Y axis), which were also based on a pulsed-SILAC approach (Schwanhäusser et al., 2011). In (**C**), per-mRNA translation rates ($T_i(0h)$, X axis) are compared to translational efficiency (TE, Y axis) estimates obtained by ribosome profiling from mouse NIH 3T3 fibroblasts (Subtelny et al., 2014).

Figure S8. Fair agreement between estimates of translation rates obtained by two alternative approaches at steady state in mouse DC



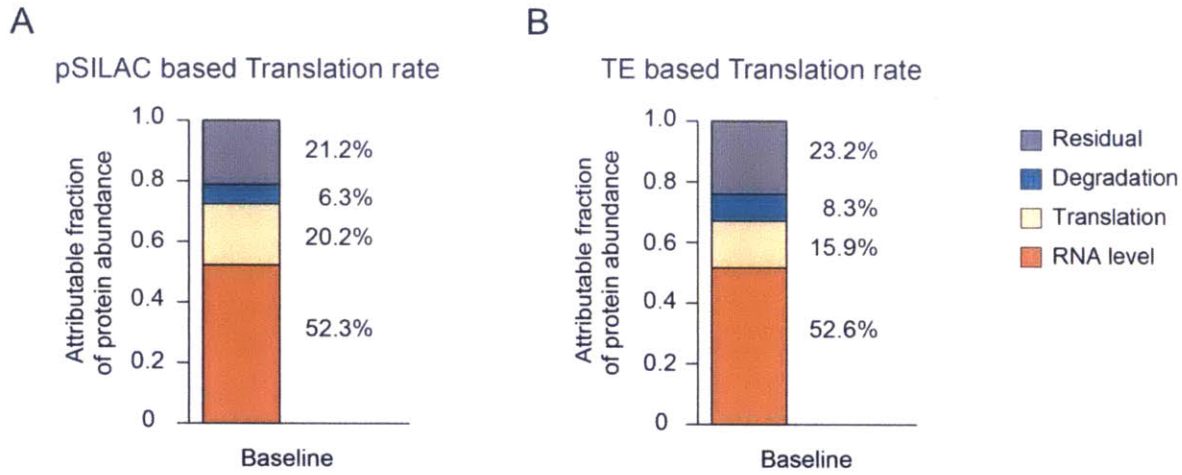
Scatter plots comparing our TE based translation rate estimates based on ribosome profiling in our resting DCs (X axis) to translation rate estimates obtained by us with pulsed-SILAC based in resting DCs (**A**, Y axis) or to TE values measured by others (Subtelny et al., 2014) using ribosome profiling in cycling mouse fibroblasts (NIH3T3 cells) (**B**, Y axis). Each plot indicates the Pearson correlation (*r*) of the log-transformed rates in the upper left corner. Each gene is represented with a semi-transparent gray point. Blue shading in the background reflects local point densities (darker = more points), and the red curve represents a local regression fit spanning from the 5th to 95th percentile.

Figure S9. Increased protein turnover and increased "degradation" of the secretome upon LPS stimulation



Shown are cumulative distribution functions (CDF) plots of (**A**) per-mRNA translation rates in LPS ($TD_{(LPS)i}$ = $T_i(12h)_{LPS}/T_i(0h)$; orange) and MOCK ($TD_{(Mock)i}$ = $T_i(12h)_{Mock}/T_i(0h)$; black) conditions; (**B**) degradation rates in LPS ($DD_{(LPS)i}$ = $D_i(12h)_{LPS}/D_i(0h)$; orange) and MOCK ($DD_{(Mock)i}$ = $D_i(12h)_{Mock}/D_i(0h)$; black) conditions; and (**C**) changes in degradation rates ($\Delta D_i = D_i(12h)_{LPS}/D_i(12h)_{MOCK}$) for all modeled proteins (grey) and for the secretome (as annotated in (Eichelbaum et al., 2012); black). Both translation and degradation rates are overall increased upon LPS stimulation and "degradation" is increased (here: decreased cellular half life) for the secretome. In all three cases: $P < 1*10^{-10}$, Wilcoxon rank sum test.

Figure S10. Contributions of RNA levels and the protein life cycle to steady state protein levels.



As in **Figure 3A** and **3B**, but the unexplained component of the variance is included (grey). Global contributions to steady state protein level of RNA (orange), protein degradation rates (blue), and either per-mRNA translation rates (**A**, from pulsed SILAC) or translation efficiency (**B**, TE, from ribosome profiling) (tan).

Figure S11. Contributions of RNA levels and the protein life cycle to steady state protein levels and to protein expression changes following LPS

A

Baseline (with residual)

Order of parameter addition	*RNA %*	*Translation %*	*Degradation %*	*Residual %*
RNA - Translation - Degradation	52.3	20.2	6.3	21.2
RNA - Degradation - Translation	52.3	15.4	11.1	21.2
Translation - RNA - Degradation	53.8	18.8	6.3	21.2
Translation - Degradation - RNA	48.2	18.8	11.9	21.2
Degradation - RNA - Translation	46.7	15.4	16.7	21.2
Degradation - Translation - RNA	48.2	13.9	16.7	21.2

Baseline (% total variance explained)

Order of parameter addition	*RNA %*	*Translation %*	*Degradation %*
RNA - Translation - Degradation	66.3	25.7	8.0
RNA - Degradation - Translation	66.3	19.6	14.1
Translation - RNA - Degradation	68.2	23.8	8.0
Translation - Degradation - RNA	61.2	23.8	15.0
Degradation - RNA - Translation	59.3	19.6	21.2
Degradation - Translation - RNA	61.2	17.7	21.2

B

Relative Fold Change

Order of parameter addition	*RNA %*	*Translation %*	*Degradation %*
RNA - Translation - Degradation	89.9	4.5	5.7
RNA - Degradation - Translation	89.9	3.8	6.4
Translation - RNA - Degradation	87.4	6.9	5.7
Translation - Degradation - RNA	90.0	6.9	3.0
Degradation - RNA - Translation	91.7	3.8	4.5
Degradation - Translation - RNA	90.0	5.5	4.5

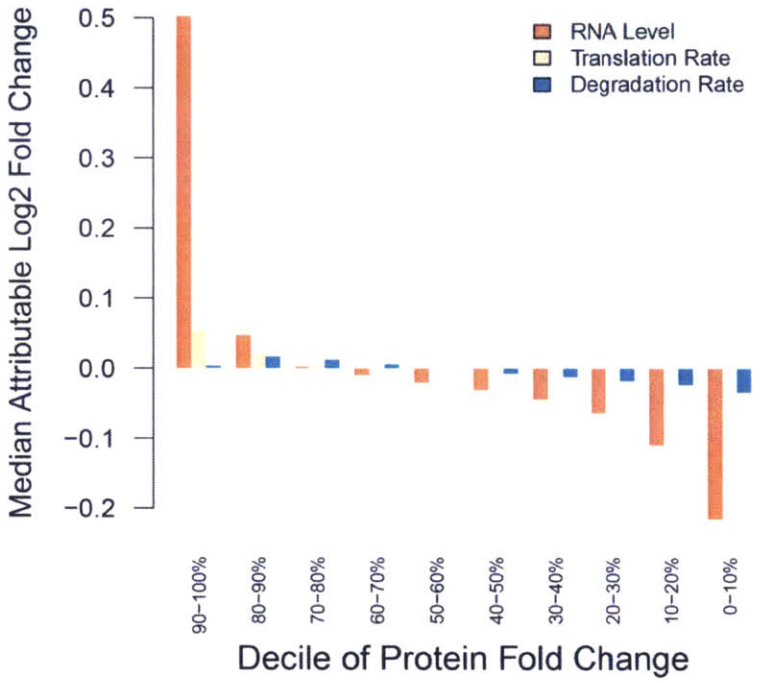
C

Absolute Change

Order of parameter addition	RNA %	Translation %	Degradation %
RNA - Translation - Degradation	42.7	30.8	26.5
RNA - Degradation - Translation	42.7	21.7	35.6
Translation - RNA - Degradation	32.2	41.3	26.5
Translation - Degradation - RNA	39.8	41.3	18.8
Degradation - RNA - Translation	42.8	21.7	35.5
Degradation - Translation - RNA	39.8	24.6	35.5

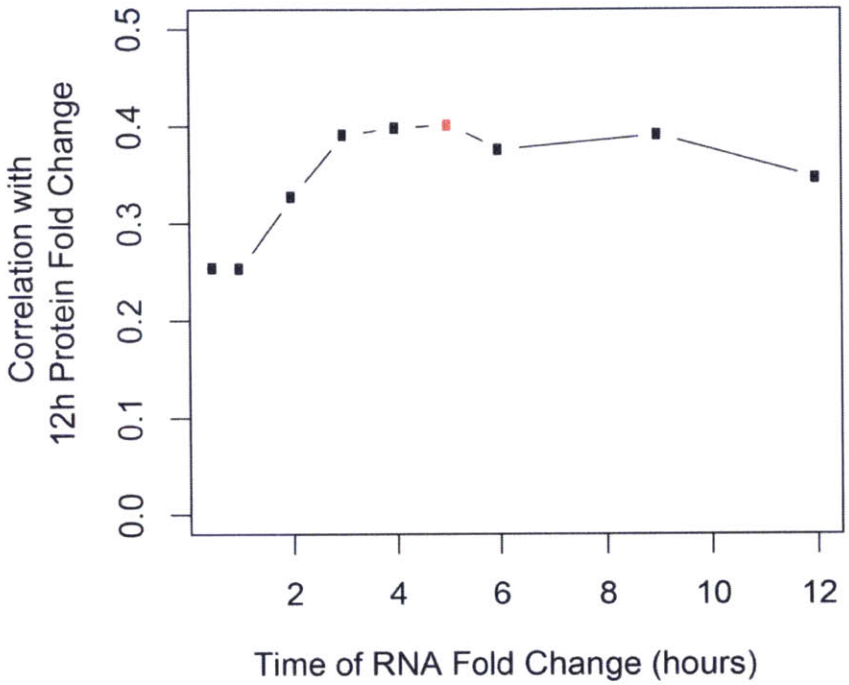
Each table shows the contributions of (left to right) RNA, per-mRNA translation rates, and protein degradation rates to (**A**) steady state protein levels, (**B**) fit fold change in protein expression in LPS *vs.* MOCK and (**C**) fit absolute difference in protein molecules in LPS *vs.* MOCK. Contributions are calculated by the Spearman-corrected Pearson correlation coefficient between model-predicted values and either independently measured protein levels (**A**) or the full models' fit from the other replicate (**B, C**). In (**A**), values were obtained by either including the unexplained component (top table), or by ignoring the unexplained component and rescaling the contributions to sum to 1 (lower table). For each analysis, values are shown for each possible ordering of information (per-gene parameter) addition to the model.

Figure S12. Regulatory contributions per protein expression decile



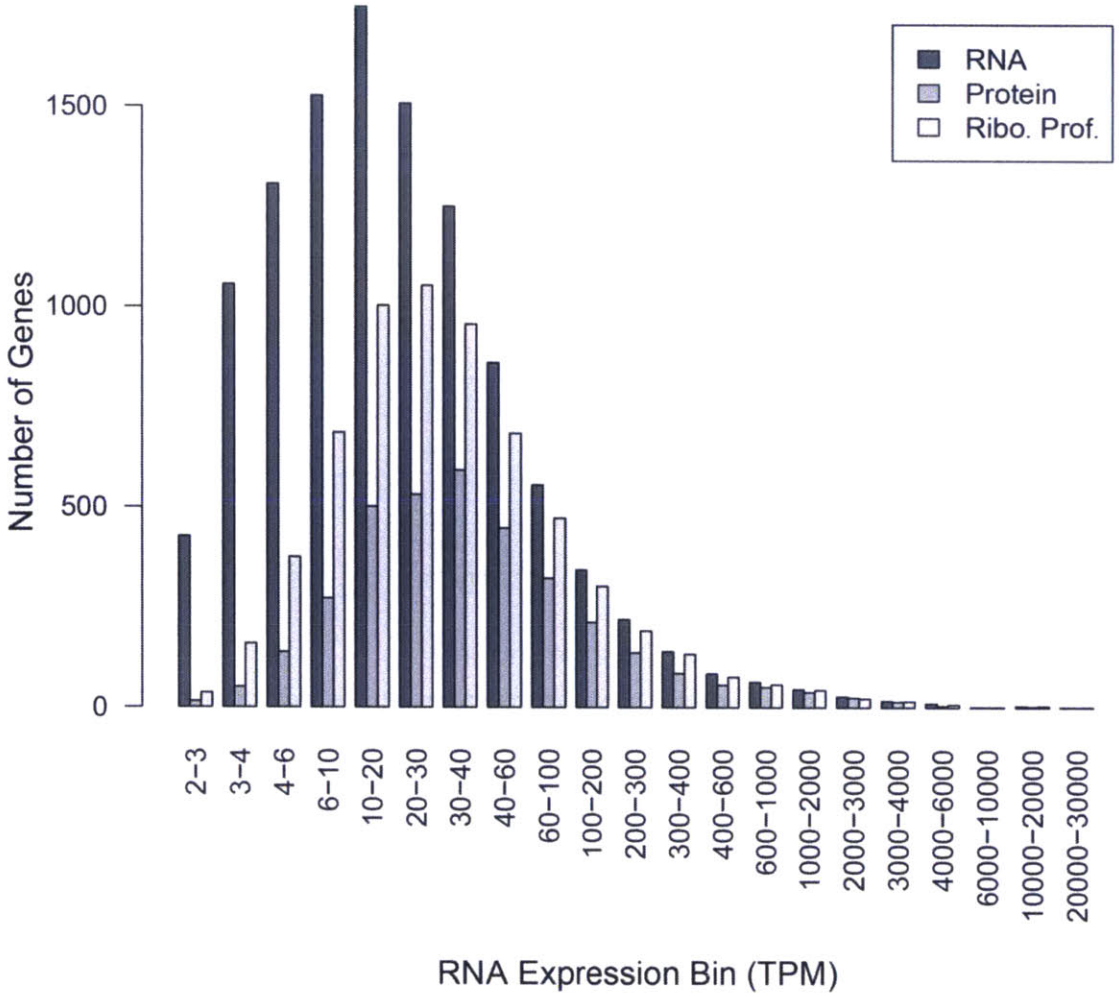
Shown is the proportion of contribution of each process to 12-hour total protein fold change (median log2 contribution to fold change; Y-axis) for genes in each decile of 12-hour total protein fold change (X-axis), ordered from largest increase (leftmost bin) to largest decrease (rightmost bin). These values are available in tabular form per-gene in **Table S5**.

Figure S13. Protein fold changes at 12h post LPS correlate best to mRNA changes at 5h



Shown is the Spearman rank correlation coefficient (ρ, Y axis) between the raw protein fold changes in LPS *vs.* MOCK at 12h and the corresponding raw RNA fold changes at each measured time point (X axis). The time point with the strongest correlation (RNA at 5h) is marked in red.

Figure S14. Proteomic and ribosome profiling detection rate as a function of RNA abundance



Shown are the distribution of number of distinct gene transcripts (black, from RNA-Seq), ribosome-associated transcripts (light grey, from ribosome profiling), and modeled proteins (grey, from proteomics) detected (Y axis) at each of 21 logarithmically-spaced bins (X axis) of RNA expression, for those genes whose mean expression across the two replicate MOCK series is greater than 2 TPM. At lower expression bins, while RNA-Seq detects the most expressed genes, only a portion of those are detectable as ribosome associated (for TE calculations) from ribosome profiling, and only some of those are further possible to model for translation and degradation rates from our proteomics data.

Figure S15. Protein level contributions corrected for detection bias

A

Baseline (with residual)

Order of parameter addition	*RNA %*	*Translation %*	*Degradation %*	*Residual %*
RNA - Translation - Degradation	51.8	20.6	5.2	22.3
RNA - Degradation - Translation	51.8	15.6	10.2	22.3
Translation - RNA - Degradation	53.8	18.6	5.2	22.3
Translation - Degradation - RNA	50.6	18.6	8.5	22.3
Degradation - RNA - Translation	49.4	15.6	12.6	22.3
Degradation - Translation - RNA	50.6	14.5	12.6	22.3

Baseline (% total variance explained)

Order of parameter addition	*RNA %*	*Translation %*	*Degradation %*
RNA - Translation - Degradation	66.7	26.5	6.7
RNA - Degradation - Translation	66.7	20.1	13.1
Translation - RNA - Degradation	69.3	23.9	6.7
Translation - Degradation - RNA	65.1	23.9	10.9
Degradation - RNA - Translation	63.7	20.1	16.2
Degradation - Translation - RNA	65.1	18.7	16.2

B

Relative Fold Change

Order of parameter addition	*RNA %*	*Translation %*	*Degradation %*
RNA - Translation - Degradation	88.3	4.7	7.0
RNA - Degradation - Translation	88.3	3.7	8.0
Translation - RNA - Degradation	84.4	8.6	7.0
Translation - Degradation - RNA	87.9	8.6	3.5
Degradation - RNA - Translation	90.6	3.7	5.8
Degradation - Translation - RNA	87.9	6.3	5.8

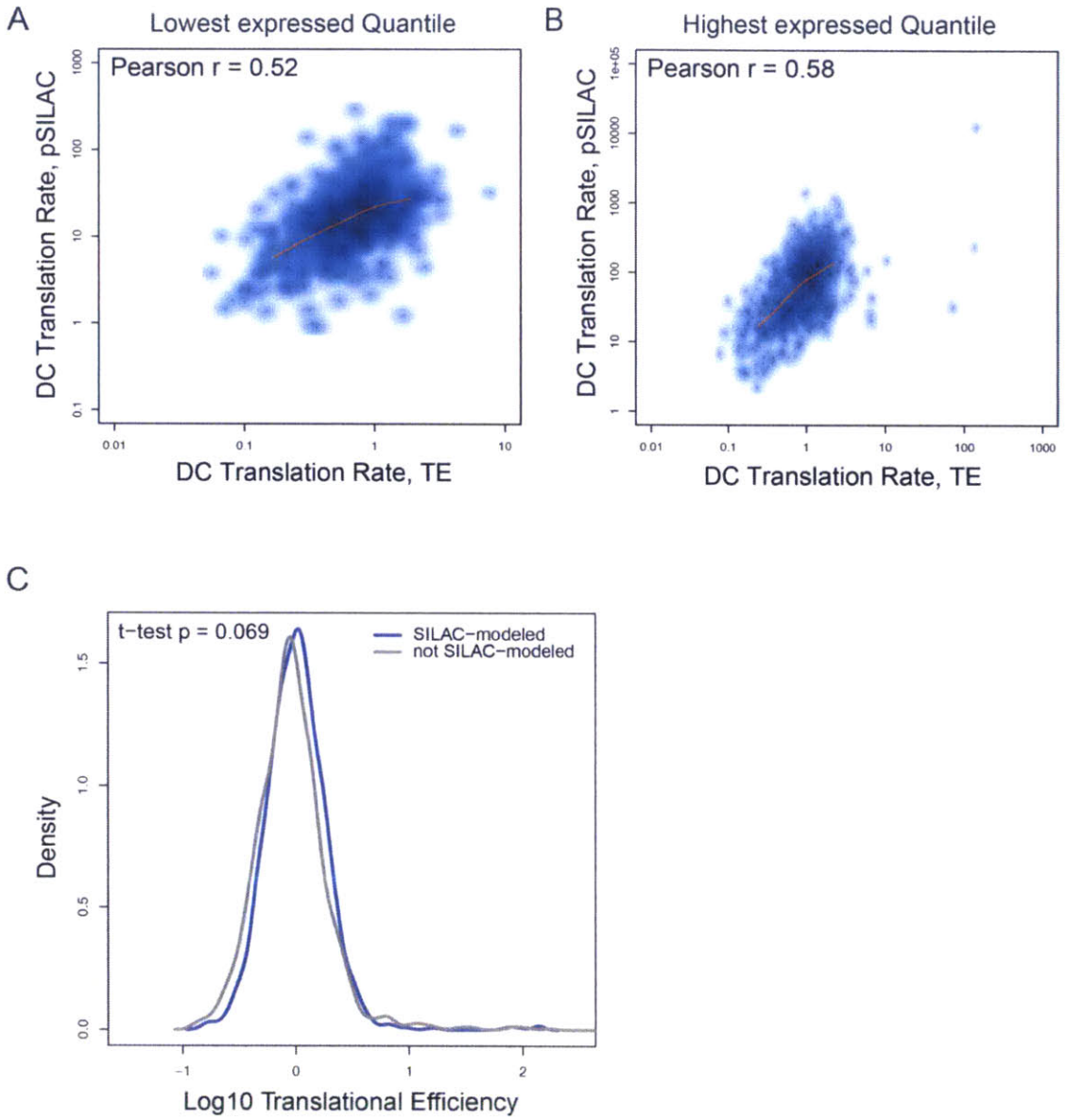
C

Absolute Change

Order of parameter addition	RNA %	Translation %	Degradation %
RNA - Translation - Degradation	43.3	30.6	26.0
RNA - Degradation - Translation	43.3	21.8	34.8
Translation - RNA - Degradation	32.1	41.9	26.0
Translation - Degradation - RNA	40.2	41.9	17.9
Degradation - RNA - Translation	44.1	21.8	34.1
Degradation - Translation - RNA	40.2	25.7	34.1

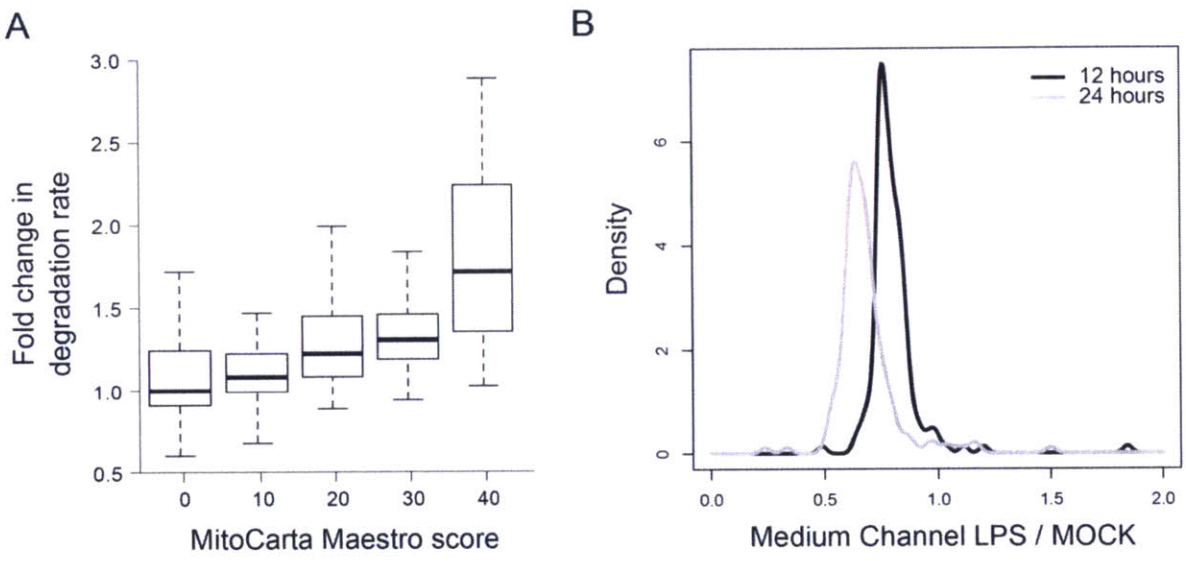
Explanatory contributions were calculated for (**A**) baseline proteins levels, (**B**) LPS-induced protein fold changes, and (**C**) absolute changes in protein abundance, as in **fig. S11,** but after applying weighting each gene proportionally to the sampling bias observed in **fig. S14**.

Figure S16. Correlation of pSILAC based translation rates and DC translational efficiencies at different expression levels



(**A**, **B**) Scatter plots comparing our model-based translation rate estimates from pulsed SILAC data (Y axis) to translational efficiencies (TE) from ribosome profiling (X axis) both in resting DCs but when considering only genes in the bottom (**A**) or top (**B**) quartile of protein expression (as assessed from modeled M_0 values, mean of replicates). The Pearson correlation (*r*) of the log-transformed rates appears in the upper left corner. (**C**) density distributions for $\log_{10}$ TE for low expression genes (mean expression 2-20 TPM in raw, MOCK-stimulated time series data) whose encoded proteins were either included in proteomic modeling (blue) or not (grey). P-value of a t-test on the log-TE values is shown in the upper left corner.

Figure S17. The vast majority of high confidence mitochondrial proteins have decreased half-lives upon LPS stimulation



(**A**) Box plot of the fold change in degradation rate (ΔD_i, Y axis) in LPS *vs.* MOCK for mitochondrial proteins annotated in Mitocarta (Pagliarini et al., 2008) binned by their Maestro scores (reflecting confidence that the protein is indeed localized to the mitochondria). (**B**) Distribution of $\log_2$ LPS/MOCK raw M/L ratios (a proxy for protein decay) for Mitocarta's "high confidence" mitochondrial proteins (Mitocarta Maestro score > 20 (Pagliarini et al., 2008)) measured in our dataset at 12h (black) and 24h (grey) post stimulation.

Figure S18. The strong contribution of protein degradation to the protein life cycle in DCs is independent of the secretome

A

Relative Fold Change – All modeled

Order of parameter addition	*RNA %*	*Translation %*	*Degradation %*
RNA - Translation - Degradation	89.9	4.5	5.7
RNA - Degradation - Translation	89.9	3.8	6.4
Translation - RNA - Degradation	87.4	6.9	5.7
Translation - Degradation - RNA	90.0	6.9	3.0
Degradation - RNA - Translation	91.7	3.8	4.5
Degradation - Translation - RNA	90.0	5.5	4.5

Relative Fold Change – No Secretome

Order of parameter addition	RNA %	Translation %	Degradation %
RNA - Translation - Degradation	92.4	3.2	4.4
RNA - Degradation - Translation	92.4	2.3	5.3
Translation - RNA - Degradation	86.4	9.3	4.4
Translation - Degradation - RNA	88.9	9.3	1.8
Degradation - RNA - Translation	92.6	2.3	5.1
Degradation - Translation - RNA	88.9	6.0	5.1

B

Absolute Change – All modeled

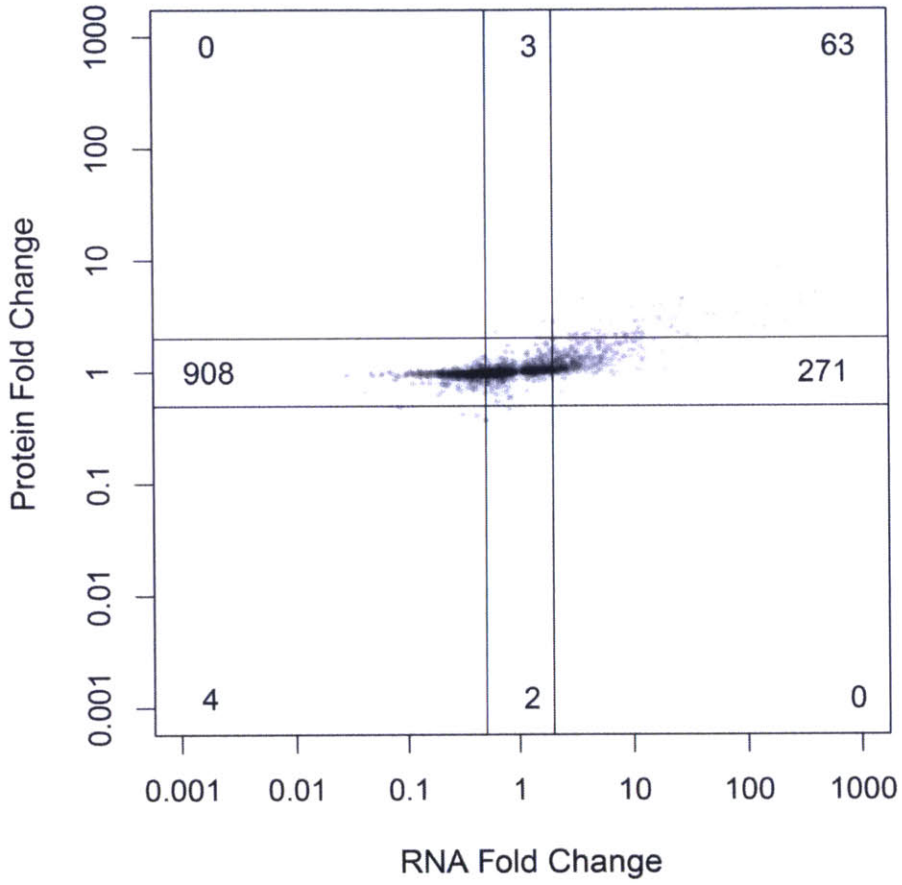
Order of parameter addition	*RNA %*	*Translation %*	*Degradation %*
RNA - Translation - Degradation	42.7	30.8	26.5
RNA - Degradation - Translation	42.7	21.7	35.6
Translation - RNA - Degradation	32.2	41.3	26.5
Translation - Degradation - RNA	39.8	41.3	18.8
Degradation - RNA - Translation	42.8	21.7	35.5
Degradation - Translation - RNA	39.8	24.6	35.5

Absolute Change – No Secretome

Order of parameter addition	RNA %	Translation %	Degradation %
RNA - Translation - Degradation	54.4	34.1	11.5
RNA - Degradation - Translation	54.4	23.7	21.9
Translation - RNA - Degradation	42.6	45.9	11.5
Translation - Degradation - RNA	42.6	45.9	11.5
Degradation - RNA - Translation	42.4	23.7	33.9
Degradation - Translation - RNA	42.6	23.5	33.9

Each table shows the contributions of (left to right) RNA, per-mRNA translation rates, and protein degradation rates to the (**A**) fit fold change in protein expression in LPS *vs.* MOCK and (**B**) fit absolute difference in protein molecules in LPS *vs.* MOCK. Contributions are calculated by the Spearman-corrected Pearson correlation coefficient between model-predicted for one replicate and the full models' fit from the other replicate. Values are shown for each possible ordering of per-gene parameters addition to the model. Results are shown when proteins reported to belong to the secretome (Blake et al., 2013) were removed from the analysis (**top table**) or for all proteins (**lower table,** identical to **the tables in Figs. S11B, S11C**).

Figure S19. Comparison of relative magnitude of RNA and protein fold changes induced by LPS



Shown is the comparison between the maximum fold change for each gene (based on absolute log fold change between modeled LPS and MOCK values, averaged between replicates, over the time course) of RNA (x-axis) and protein (y-axis). Each gene is represented with a semi-transparent gray point. Blue shading in the background reflects local point densities (darker = more points), and the red curve represents a local regression fit spanning from the 5th to 95th percentile. Protein level changes were considered in terms of total protein (M+H).

Figure S20. Contributions, based on Spearman rank, of RNA levels and the protein life cycle to steady state protein levels and to protein expression changes following LPS

A

Baseline (with residual)

Order of parameter addition	*RNA %*	*Translation %*	*Degradation %*	*Residual %*
RNA - Translation - Degradation	49.2	21.4	7.1	22.3
RNA - Degradation - Translation	49.2	14.7	13.8	22.3
Translation - RNA - Degradation	51.1	19.5	7.1	22.3
Translation - Degradation - RNA	44.9	19.5	13.4	22.3
Degradation - RNA - Translation	42.0	14.7	21.1	22.3
Degradation - Translation - RNA	44.9	11.8	21.1	22.3

Baseline (% total variance explained)

Order of parameter addition	*RNA %*	*Translation %*	*Degradation %*
RNA - Translation - Degradation	63.3	27.5	9.2
RNA - Degradation - Translation	63.3	18.9	17.8
Translation - RNA - Degradation	65.7	25.1	9.2
Translation - Degradation - RNA	57.7	25.1	17.2
Degradation - RNA - Translation	54.0	18.9	27.1
Degradation - Translation - RNA	57.7	15.2	27.1

B

Relative Fold Change

Order of parameter addition	*RNA %*	*Translation %*	*Degradation %*
RNA - Translation - Degradation	71.4	10.9	17.6
RNA - Degradation - Translation	71.4	8.0	20.5
Translation - RNA - Degradation	65.1	17.3	17.6
Translation - Degradation - RNA	71.1	17.3	11.6
Degradation - RNA - Translation	76.3	8.0	15.7
Degradation - Translation - RNA	71.1	13.2	15.7

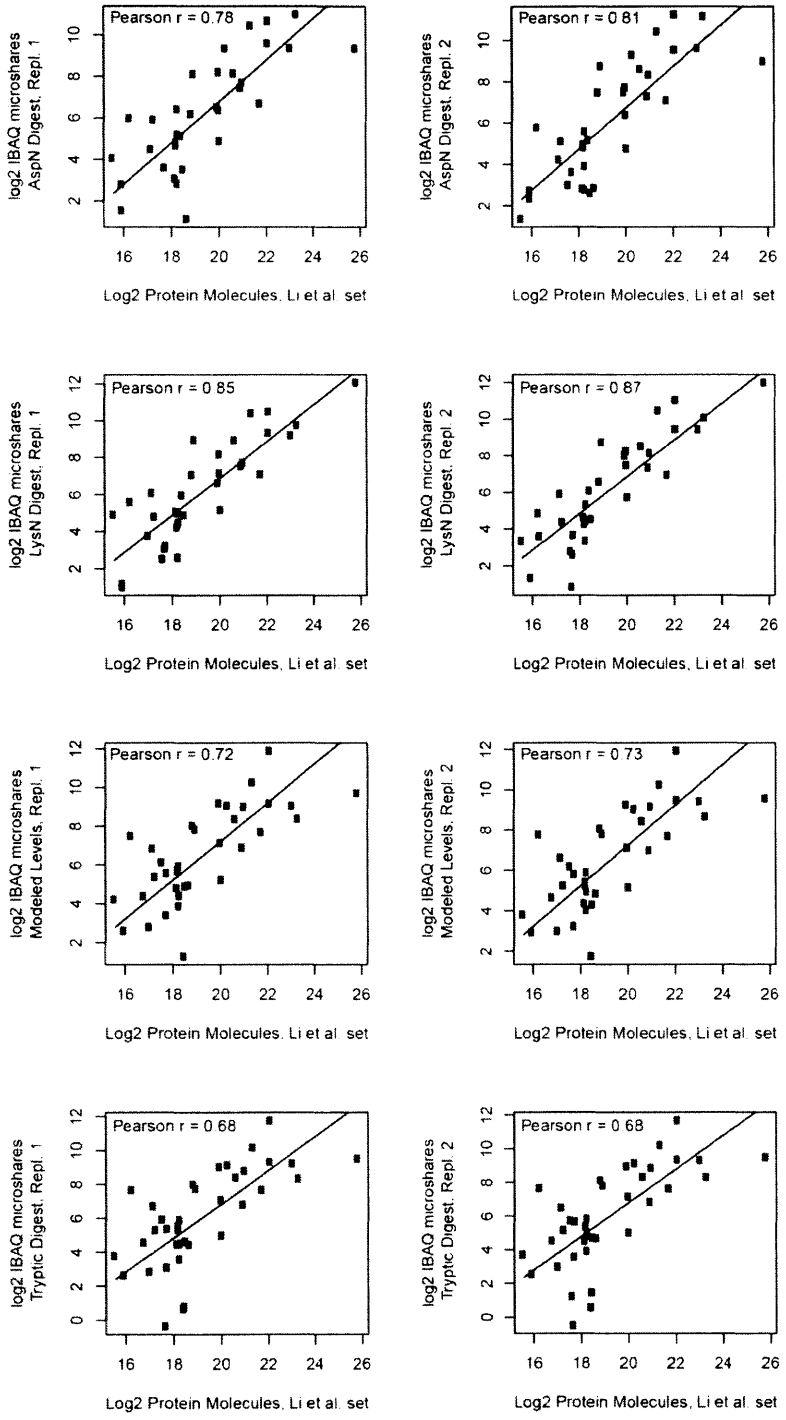
C

Absolute Change

Order of parameter addition	RNA %	Translation %	Degradation %
RNA - Translation - Degradation	51.7	26.6	21.7
RNA - Degradation - Translation	51.7	20.9	27.4
Translation - RNA - Degradation	57.2	21.1	21.7
Translation - Degradation - RNA	55.4	21.1	23.5
Degradation - RNA - Translation	48.1	20.9	31.0
Degradation - Translation - RNA	55.4	13.5	31.0

Each table shows the contributions of (left to right) RNA, per-mRNA translation rates, and protein degradation rates to (**A**) steady state protein levels, (**B**) fit fold change in protein expression in LPS *vs.* MOCK and (**C**) fit absolute difference in protein molecules in LPS *vs.* MOCK. Contributions are calculated by the Spearman-corrected Spearman rank correlation coefficient between model-predicted values and either independently measured protein levels (**A**) or the full models' fit from the other replicate (**B, C**). Values are shown for each possible ordering of information (per-gene parameter) addition to the model. In (**A**), values were obtained by either including the unexplained component (top table), or by ignoring the unexplained component and rescaling the contributions to sum to 1 (lower table). For each analysis, values are shown for each possible ordering of information (per-gene parameter) addition to the model.

Figure S21. Comparison of estimated baseline protein levels from different protein digestions and our model to the expression level of previously determined 'standards'



Scatter plots of estimated baseline protein levels from different protein digest and our model (log_2 IMS values) to previously reported and handpicked 'standards' (log_2 absolute protein levels) (14). Each plot depicts 37 out of all of the 'standard' 61 proteins that could be quantified (or modeled) in our data. Each plot indicates the Pearson correlation coefficient in the top left corner and depicts a fit line based on a linear scale factor.

Supplemental Tables

Since many of the supplemental tables are large, they are available online only:
http://www.sciencemag.org/content/347/6226/1259038.short
(DOI: 10.1126/science.1259038)

Table S1. Protein expression data
Normalized raw protein expression data (IBAQ shares per million (IMS)) for the 3,147 modeled genes by channel (M/L or H/L), experimental condition (LPS or MOCK), replicate (R1 or R2), and time point.

Table S2. RNA expression data
Normalized mRNA expression data (TPM) for the 3,147 modeled genes by experimental condition (LPS or MOCK), replicate (R1 or R2), and time point.

Table S3. Per-gene parameter estimates.
(A) Parameter estimates with the original modeled units IMS and TPM. Listed are parameter estimates for D0: baseline degradation rate [1/hr]; D(MOCK): fold change in degradation rate at 12 hours under MOCK conditions; D(LPS): fold change in degradation rate at 12 hours under LPS conditions; T0: baseline per-mRNA translation rate [IMS / TPM / hour]; T(MOCK): fold change in per-mRNA translation rate at 12 hours under MOCK conditions; T(LPS): fold change in per-mRNA translation rate at 12 hours under LPS conditions; M0: protein expression at baseline [IMS]; grass: background signal [IMS]. Values are presented for each replicate (R1 or R2) in addition to the low and high boundaries of their respective 95% credible intervals (CI.HI and CI.LO). **(B)** Parameter estimates with the units protein and RNA molecules. As in A, but after scaling our model estimates so that they could be interpreted in terms of protein and mRNA molecules rather than microshares. Listed are parameter estimates for D0: baseline degradation rate [1/hr]; D(MOCK): fold change in degradation rate at 12 hours under MOCK conditions; D(LPS): fold change in degradation rate at 12 hours under LPS conditions; T0: baseline per-mRNA translation rate [Protein molecules / RNA transcript / hour]; T(MOCK): fold change in per-mRNA translation rate at 12 hours under MOCK conditions; T(LPS): fold change in per-mRNA translation rate at 12 hours under LPS conditions; M0: protein expression at baseline [Protein molecules]; grass: background signal [Protein molecule equivalents]. Values are presented for each replicate (R1 or R2) in addition to the low and high boundaries of their respective 95% credible intervals (CI.HI and CI.LO).

Table S4. Ribosome profiling and translational efficiencies
Ribosomal footprinting (FP) density was calculated in two replicates (R1 and R2) with transcript length corrections and global normalization applied. These FP values were divided by corresponding mRNA abundances of the MOCK-stimulated time series (averaged, per replicate, across the 10 time points) to obtain translational efficiencies (TE.R1, TE.R2) at time point 0h.

Table S5. Per-gene contributions of RNA, per-mRNA translation rate and protein degradation rate to the final protein level (change)
Per-gene, per-replicate (R1 or R2) estimates of the contribution of each regulatory channel to baseline protein level (sheet: contributions.BASE.csv), 12-hour LPS-induced fold change in protein level (sheet: contributions.FC.csv), and 12-hour LPS-induced absolute changes in protein level (sheet: contributions.ABS.csv). The first two sheets present fold-change data (FC due RNA: contribution of RNA; FC due Translation: contribution of translation; FC due Degradation: contribution of degradation), whereas the third presents expression units [Protein Molecules] (Diff due RNA: contribution of RNA; Diff due Translation: contribution of translation; Diff due Degradation: contribution of degradation).

Table S6. GO term analysis of RNA level, per-mRNA translation rate and protein degradation rate (baseline and fold changes)
GO term $\log_{10}$ p-values (Wilcoxon Rank Sum test, signed such that positive values indicate enrichment and negative values indicated depletion) calculated for baseline mRNA level, translation rate, and degradation rate (R0, T0, and D0, respectively) and 12-hour LPS-induced fold changes in mRNA level, translation rate, and degradation rate (RD, TD, and DD, respectively) along with the count of modeled genes in the gene set.

Table S7. Proteins with significantly increased synthesis rates and/or degradation rates upon LPS stimulation
Gene names and estimated rate fold changes for genes that had significantly higher translation or degradation (posterior odds off rate change increase greater than 100) upon LPS stimulation.

Table S8. Proteins with significantly decreased synthesis rates and/or degradation rates upon LPS stimulation
Gene names and estimated rate fold changes for genes that had significantly slower translation or degradation (posterior odds off rate change decrease greater than 100) upon LPS stimulation.

Table S9. LPS-induced fold changes in RNA and protein
The maximum fold change for each gene (based on absolute log fold change between modeled LPS and MOCK values, averaged between replicates, over the time course) was determined for RNA (**A**) and protein (**B**). Protein level changes were considered in terms of total protein (M+H).

References

1. Ahrne, E., Molzahn, L., Glatter, T., and Schmidt, A. (2013). Critical assessment of proteome-wide label-free absolute abundance estimation strategies. Proteomics *13*, 2567-2578.

2. Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard, P., Mozes, E., Garraway, L.A., and Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. Cell *143*, 1005-1017.

3. Albert, F.W., Muzzey, D., Weissman, J.S., and Kruglyak, L. (2014). Genetic influences on translation in yeast. PLoS genetics *10*, e1004692.

4. Almeida, L.G., Sakabe, N.J., deOliveira, A.R., Silva, M.C., Mundstein, A.S., Cohen, T., Chen, Y.T., Chua, R., Gurung, S., Gnjatic, S., *et al.* (2009). CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. Nucleic acids research *37*, D816-819.

5. Amit, I., Garber, M., Chevrier, N., Leite, A.P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J.K., Li, W., and Zuk, O. (2009). Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science *326*, 257-263.

6. Andersen, R.S., Thrue, C.A., Junker, N., Lyngaa, R., Donia, M., Ellebæk, E., Svane, I.M., Schumacher, T.N., Thor Straten, P., and Hadrup, S.R. (2012). Dissection of T-cell antigen specificity in human melanoma. Cancer research *72*, 1642-1650.

7. Aviner, R., Geiger, T., and Elroy-Stein, O. (2013). Novel proteomic approach (PUNCH-P) reveals cell cycle-specific fluctuations in mRNA translation. Genes & development *27*, 1834-1844.

8. Ayyoub, M., Hesdorffer, C.S., Montes, M., Merlo, A., Speiser, D., Rimoldi, D., Cerottini, J.-C., Ritter, G., Scanlan, M., Old, L.J., *et al.* (2004). An immunodominant SSX-2–derived epitope recognized by CD4+ T cells in association with HLA-DR. Journal of Clinical Investigation *113*, 1225-1233.

9. Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64-71.

10. Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., and Scholl, C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108-112.

11. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., *et al.* (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603-607.

12. Bell, C., English, L., Boulais, J., Chemali, M., Caron-Lizotte, O., Desjardins, M., and Thibault, P. (2013). Quantitative Proteomics Reveals the induction of mitophagy in tumor necrosis factor-α-activated (TNFα) macrophages. Molecular & Cellular Proteomics *12*, 2394-2407.

13. Belle, A., Tanay, A., Bitincka, L., Shamir, R., and O'Shea, E.K. (2006). Quantification of protein half-lives in the budding yeast proteome. Proceedings of the National Academy of Sciences *103*, 13004-13009.

14. Bennetzen, J.L., and Hall, B. (1982). Codon selection in yeast. Journal of Biological Chemistry *257*, 3026-3031.

15. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research *27*, 573-580.

16. Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A.C., Angell, H., Fredriksen, T., Lafontaine, L., Berger, A., *et al.* (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity *39*, 782-795.

17. Birkeland, S.A., Storm, H.H., Lamm, L.U., Barlow, L., Blohmé, I., Forsberg, B., Eklund, B., Fjeldborg, O., Friedberg, M., and Frödin, L. (1995). Cancer risk after renal transplantation in the Nordic countries, 1964–1986. International journal of cancer *60*, 183-189.

18. Biswas, S.K., Gangi, L., Paul, S., Schioppa, T., Saccani, A., Sironi, M., Bottazzi, B., Doni, A., Vincenzo, B., and Pasqualini, F. (2006). A distinct and unique transcriptional program expressed by tumor-associated macrophages (defective NF-κB and enhanced IRF-3/STAT1 activation). Blood *107*, 2112-2122.

19. Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A., and Richardson, J.E. (2013). The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic acids research, gkt1225.

20. Boisvert, F.-M., Ahmad, Y., Gierliński, M., Charrière, F., Lamont, D., Scott, M., Barton, G., and Lamond, A.I. (2012). A quantitative spatial proteomics analysis of proteome turnover in human cells. Molecular & Cellular Proteomics *11*, M111. 011429.

21. Boller, K., Janssen, O., Schuldes, H., Tonjes, R.R., and Kurth, R. (1997). Characterization of the antibody response specific for the human endogenous retrovirus HTDV/HERV-K. J Virol *71*, 4581-4588.

22. Broad Institute TCGA Genome Data Analysis Center (2014). Firehose. In Clinical Data (Broad Institute of MIT and Harvard).

23. Brown, S.D., Warren, R.L., Gibb, E.A., Martin, S.D., Spinelli, J.J., Nelson, B.H., and Holt, R.A. (2014). Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. Genome research *24*, 743-750.

24. Browne, G.J., and Proud, C.G. (2002). Regulation of peptide-chain elongation in mammalian cells. European Journal of Biochemistry *269*, 5360-5368.

25. Buchta, C., Hornik, K., and Hahsler, M. (2008). Getting things in order: an introduction to the R package seriation. Journal of Statistical Software *25*, 1-34.

26. Burgess, D.J. (2015). Cancer genetics: Omics analyses of tumour immunity. Nature Reviews Genetics *16*, 130-131.

27. Burnet, F. (1971). Immunological surveillance in neoplasia. Immunological reviews *7*, 3-25.

28. Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., and Samani, N.J. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661-678.

29. Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proceedings of the National Academy of Sciences *106*, 7507-7512.

30. Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519-525.

31. Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202-209.

32. Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., and Weir, B.A. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology *30*, 413-421.

33. Chang, C.C., Campoli, M., Restifo, N.P., Wang, X., and Ferrone, S. (2005). Immune selection of hot-spot beta 2-microglobulin gene mutations, HLA-A2 allospecificity loss, and antigen-processing machinery component down-regulation in melanoma cells derived from recurrent metastases following immunotherapy. Journal of immunology *174*, 1462-1471.

34. Chechik, G., Oh, E., Rando, O., Weissman, J., Regev, A., and Koller, D. (2008). Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. Nature biotechnology *26*, 1251-1259.

35. Chen, L., and Flies, D.B. (2013). Molecular mechanisms of T cell co-stimulation and co-inhibition. Nature reviews Immunology *13*, 227-242.

36. Chevrier, N., Mertins, P., Artyomov, M.N., Shalek, A.K., Iannacone, M., Ciaccio, M.F., Gat-Viks, I., Tonti, E., DeGrace, M.M., and Clauser, K.R. (2011). Systematic discovery of TLR signaling components delineates viral-sensing circuits. Cell *147*, 853-867.

37. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology *31*, 213-219.

38. Cleveland, W.S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. The American Statistician *35*, 54.

39. Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nature biotechnology *26*, 1367-1372.

40. Davies, H., Bignell, G.R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M.J., and Bottomley, W. (2002). Mutations of the BRAF gene in human cancer. Nature *417*, 949-954.

41. de Godoy, L.M., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Fröhlich, F., Walther, T.C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature *455*, 1251-1254.

42. Dice, J.F. (1990). Peptide sequences that target cytosolic proteins for lysosomal proteolysis. Trends in biochemical sciences *15*, 305-309.

43. DiGiulio, S. (2013). FDA Approves Xalkori (Crizotinib) for NSCLC. Oncology Times.

44. Donson, A.M., Birks, D.K., Schittone, S.A., Kleinschmidt-DeMasters, B.K., Sun, D.Y., Hemenway, M.F., Handler, M.H., Waziri, A.E., Wang, M., and Foreman, N.K. (2012). Increased immune gene expression and immune cell infiltration in high-grade astrocytoma distinguish long-term from short-term survivors. The Journal of Immunology *189*, 1920-1927.

45. Dunn, G.P., Bruce, A.T., Ikeda, H., Old, L.J., and Schreiber, R.D. (2002). Cancer immunoediting: from immunosurveillance to tumor escape. Nature immunology *3*, 991-998.

46. Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. PLoS computational biology *3*, e39.

47. Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC bioinformatics *10*, 48.

48. Eichelbaum, K., and Krijgsveld, J. (2014). Rapid Temporal Dynamics of Transcription, Protein Synthesis, and Secretion during Macrophage Activation. Molecular & Cellular Proteomics *13*, 792-810.

49. Eichelbaum, K., Winter, M., Diaz, M.B., Herzig, S., and Krijgsveld, J. (2012). Selective enrichment of newly synthesized proteins for quantitative secretome analysis. Nature biotechnology *30*, 984-990.

50. Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., and Lander, E.S. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973.

51. Everts, B., Amiel, E., Huang, S.C.-C., Smith, A.M., Chang, C.-H., Lam, W.Y., Redmann, V., Freitas, T.C., Blagih, J., and van der Windt, G.J. (2014). TLR-driven early glycolytic reprogramming via the kinases TBK1-IKK [epsiv] supports the anabolic demands of dendritic cell activation. Nature immunology *15*, 323-332.

52. Everts, B., Amiel, E., van der Windt, G.J., Freitas, T.C., Chott, R., Yarasheski, K.E., Pearce, E.L., and Pearce, E.J. (2012). Commitment to glycolysis sustains survival of NO-producing inflammatory dendritic cells. Blood *120*, 1422-1431.

53. Fan, S., Ma, Y.X., Gao, M., Yuan, R.Q., Meng, Q., Goldberg, I.D., and Rosen, E.M. (2001). The multisubstrate adapter Gab1 regulates hepatocyte growth factor (scatter factor)-c-Met signaling for cell survival and DNA repair. Molecular and cellular biology *21*, 4968-4984.

54. Fantom Consortium, Pmi, R., Clst, Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Lassmann, T., Itoh, M., *et al.* (2014). A promoter-level mammalian expression atlas. Nature *507*, 462-470.

55. Filbin, M.E., and Kieft, J.S. (2009). Toward a structural understanding of IRES RNA function. Current opinion in structural biology *19*, 267-276.

56. Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., and Omeroglu, A. (2008). Stromal gene expression predicts clinical outcome in breast cancer. Nature medicine *14*, 518-527.

57. Fournier, M.L., Paulson, A., Pavelka, N., Mosley, A.L., Gaudenz, K., Bradford, W.D., Glynn, E., Li, H., Sardiu, M.E., and Fleharty, B. (2010). Delayed correlation of mRNA and protein expression in rapamycin-treated cells and a role for Ggc1 in cellular sensitivity to rapamycin. Molecular & Cellular Proteomics *9*, 271-284.

58. Fridman, W.H., Pagès, F., Sautès-Fridman, C., and Galon, J. (2012). The immune contexture in human tumours: impact on clinical outcome. Nature Reviews Cancer *12*, 298-306.

59. Fritsch, E.F., Hacohen, N., and Wu, C.J. (2014). Personal neoantigen cancer vaccines: The momentum builds. Oncoimmunology *3*, e29311.

60. Futcher, B., Latter, G., Monardo, P., McLaughlin, C., and Garrels, J. (1999). A sampling of the yeast proteome. Molecular and cellular biology *19*, 7357-7368.

61. Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., Tosolini, M., Camus, M., Berger, A., and Wind, P. (2006). Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science *313*, 1960-1964.

62. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., *et al.* (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling *6*, pl1.

63. Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., and Itzhaki, Z. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Molecular cell *47*, 810-822.

64. Garimella, S.V., Gehlhaus, K., Dine, J.L., Pitt, J.J., Grandin, M., Chakka, S., Nau, M.M., Caplen, N.J., and Lipkowitz, S. (2014). Identification of novel molecular regulators of tumor necrosis factor-related apoptosis-inducing ligand (TRAIL)-induced apoptosis in breast cancer cells by RNAi screening. Breast cancer research : BCR *16*, R41.

65. Garofalo, M., Di Leva, G., Romano, G., Nuovo, G., Suh, S.S., Ngankeu, A., Taccioli, C., Pichiorri, F., Alder, H., Secchiero, P., *et al.* (2009). miR-221&222 regulate TRAIL resistance and enhance tumorigenicity through PTEN and TIMP3 downregulation. Cancer cell *16*, 498-509.

66. Geiger, T., Velic, A., Macek, B., Lundberg, E., Kampf, C., Nagaraj, N., Uhlen, M., Cox, J., and Mann, M. (2013). Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. Molecular & Cellular Proteomics *12*, 1709-1722.

67. Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737-741.

68. Graber, T.E., and Holcik, M. (2007). Cap-independent regulation of gene expression in apoptosis. Molecular Biosystems *3*, 825-834.

69. Groh, V., Rhinehart, R., Secrist, H., Bauer, S., Grabstein, K.H., and Spies, T. (1999). Broad tumor-associated expression and recognition by tumor-derived γδ T cells of MICA and MICB. Proceedings of the National Academy of Sciences *96*, 6879-6884.

70. Grønbæk, K., thor Straten, P., Ralfkiaer, E., Ahrenkiel, V., Andersen, M.K., Hansen, N.E., Zeuthen, J., Hou-Jensen, K., and Guldberg, P. (1998). Somatic Fas mutations in non-Hodgkin's lymphoma: association with extranodal disease and autoimmunity. Blood *92*, 3018-3024.

71. Gsponer, J., Futschik, M.E., Teichmann, S.A., and Babu, M.M. (2008). Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science *322*, 1365-1368.

72. GTEx Consortium (2013a). The Genotype-Tissue Expression (GTEx) project. Nature genetics *45*, 580-585.

73. GTEx Consortium (2013b). GTEx Portal (Broad Institute of MIT and Harvard), pp. GTEx_Analysis_RNA-seq_RNA-SeQCv1.1.8_gene_reads__Pilot_2013_2001_2031_patch2011.gct.

74. Guo, W., Jiang, L., Bhasin, S., Khan, S.M., and Swerdlow, R.H. (2009). DNA extraction procedures meaningfully influence qPCR-based mtDNA copy number determination. Mitochondrion *9*, 261-265.

75. Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Molecular and cellular biology *19*, 1720-1730.

76. Hahsler, M., Buchta, C., Hornik, K., Murtagh, F., Brusco, M., Stahl, S., Koehn, H.-F., and Hahsler, M.M. Package 'seriation'.

77. Hamid, O., Robert, C., Daud, A., Hodi, F.S., Hwu, W.-J., Kefford, R., Wolchok, J.D., Hersey, P., Joseph, R.W., and Weber, J.S. (2013). Safety and tumor responses with lambrolizumab (anti–PD-1) in melanoma. New England Journal of Medicine *369*, 134-144.

78. Hamilton, R., Watanabe, C.K., and de Boer, H.A. (1987). Compilation and comparison of the sequence context around the AUG startcodons in Saccharomyces cerevisiae mRNAs. Nucleic acids research *15*, 3581-3593.

79. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674.

80. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC bioinformatics *14*, 7.

81. Herbst, R.S., Soria, J.-C., Kowanetz, M., Fine, G.D., Hamid, O., Gordon, M.S., Sosman, J.A., McDermott, D.F., Powderly, J.D., and Gettinger, S.N. (2014). Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. Nature *515*, 563-567.

82. Hicke, L. (2001). Protein regulation by monoubiquitin. Nature Reviews Molecular Cell Biology *2*, 195-201.

83. Hicklin, D.J., Wang, Z., Arienti, F., Rivoltini, L., Parmiani, G., and Ferrone, S. (1998). beta2-Microglobulin mutations, HLA class I antigen loss, and tumor progression in melanoma. Journal of Clinical Investigation *101*, 2720.

84. Hirano, F., Kaneko, K., Tamura, H., Dong, H., Wang, S., Ichikawa, M., Rietz, C., Flies, D.B., Lau, J.S., and Zhu, G. (2005). Blockade of B7-H1 and PD-1 by monoclonal antibodies potentiates cancer therapeutic immunity. Cancer research *65*, 1089-1096.

85. Hodi, F.S., O'Day, S.J., McDermott, D.F., Weber, R.W., Sosman, J.A., Haanen, J.B., Gonzalez, R., Robert, C., Schadendorf, D., and Hassel, J.C. (2010). Improved survival with ipilimumab in patients with metastatic melanoma. New England Journal of Medicine *363*, 711-723.

86. Hodis, E., Watson, I.R., Kryukov, G.V., Arold, S.T., Imielinski, M., Theurillat, J.P., Nickerson, E., Auclair, D., Li, L., Place, C., *et al.* (2012). A landscape of driver mutations in melanoma. Cell *150*, 251-263.

87. Hsieh, A.C., Liu, Y., Edlind, M.P., Ingolia, N.T., Janes, M.R., Sher, A., Shi, E.Y., Stumpf, C.R., Christensen, C., and Bonham, M.J. (2012). The translational landscape of mTOR signalling steers cancer initiation and metastasis. Nature *485*, 55-61.

88. Huntzinger, E., and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nature Reviews Genetics *12*, 99-110.

89. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nature protocols *7*, 1534-1550.

90. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218-223.

91. Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789-802.

92. Izeradjene, K., Douglas, L., Delaney, A., and Houghton, J.A. (2005). Casein kinase II (CK2) enhances death-inducing signaling complex (DISC) activity in TRAIL-induced apoptosis in human colon carcinoma cell lines. Oncogene *24*, 2050-2058.

93. Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., and Thun, M.J. (2007). Cancer statistics, 2007. CA: a cancer journal for clinicians *57*, 43-66.

94. Ji, R.R., Chasalow, S.D., Wang, L., Hamid, O., Schmidt, H., Cogswell, J., Alaparthy, S., Berman, D., Jure-Kunkel, M., Siemers, N.O., *et al.* (2012). An immune-active tumor microenvironment favors clinical response to ipilimumab. Cancer immunology, immunotherapy : CII *61*, 1019-1031.

95. Joazeiro, C.A., and Weissman, A.M. (2000). RING finger proteins: mediators of ubiquitin ligase activity. Cell *102*, 549-552.

96. Johnson, B.J., Costelloe, E.O., Fitzpatrick, D.R., Haanen, J.B., Schumacher, T.N., Brown, L.E., and Kelso, A. (2003). Single-cell perforin and granzyme expression reveals the anatomical localization of effector CD8+ T cells in influenza virus-infected mice. Proceedings of the National Academy of Sciences of the United States of America *100*, 2657-2662.

97. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., *et al.* (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236-1240.

98. Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., and Raychowdhury, R. (2015). Dynamic profiling of the protein life cycle in response to pathogens. Science *347*, 1259038.

99. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., *et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.

100. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic acids research *32*, D493-496.

101. Khong, H.T., and Restifo, N.P. (2002). Natural selection of tumor variants in the generation of "tumor escape" phenotypes. Nature immunology *3*, 999-1005.

102. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology *14*, R36.

103. Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., and Jain, S. (2014). A draft map of the human proteome. Nature *509*, 575-581.

104. Kloor, M., Michel, S., and von Knebel Doeberitz, M. (2010). Immune evasion of microsatellite unstable colorectal cancers. International journal of cancer *127*, 1001-1010.

105. Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell *44*, 283-292.

106. Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic acids research *15*, 8125-8148.

107. Krawczyk, C.M., Holowka, T., Sun, J., Blagih, J., Amiel, E., DeBerardinis, R.J., Cross, J.R., Jung, E., Thompson, C.B., and Jones, R.G. (2010). Toll-like receptor–induced changes in glycolytic metabolism regulate dendritic cell activation. Blood *115*, 4742-4749.

108. Kristensen, A.R., Gsponer, J., and Foster, L.J. (2013). Protein synthesis rate is the predominant regulator of protein expression during differentiation. Molecular systems biology *9*.

109. Kuan, A.P., Chamberlain, W., Malkiel, S., Lieu, H.D., Factor, S.M., Diamond, B., and Kotzin, B.L. (1999). Genetic control of autoimmune myocarditis mediated by myosin-specific antibodies. Immunogenetics *49*, 79-85.

110. Lackner, D.H., Beilharz, T.H., Marguerat, S., Mata, J., Watt, S., Schubert, F., Preiss, T., and Bähler, J. (2007). A network of multiple regulatory layers shapes gene expression in fission yeast. Molecular cell *26*, 145-155.

111. Landau, D.A., Clement, K., Ziller, M.J., Boyle, P., Fan, J., Gu, H., Stevenson, K., Sougnez, C., Wang, L., and Li, S. (2014). Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. Cancer cell *26*, 813-825.

112. Landowski, T.H., Qu, N., Buyuksal, I., Painter, J.S., and Dalton, W.S. (1997). Mutations in the Fas antigen in patients with multiple myeloma. Blood *90*, 4266-4270.

113. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

114. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495-501.

115. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., *et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

116. Leach, D.R., Krummel, M.F., and Allison, J.P. (1996). Enhancement of antitumor immunity by CTLA-4 blockade. Science *271*, 1734-1736.

117. Lee, M., Topper, S.E., Hubler, S.L., Hose, J., Wenger, C.D., Coon, J.J., and Gasch, A.P. (2011). A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. Molecular systems biology *7*.

118. Lelouard, H., Schmidt, E.K., Camosseto, V., Clavarino, G., Ceppi, M., Hsu, H.-T., and Pierre, P. (2007). Regulation of translation is required for dendritic cell function and survival during activation. The Journal of cell biology *179*, 1427-1439.

119. Lemay, S., Davidson, D., Latour, S., and Veillette, A. (2000). Dok-3, a novel adapter molecule involved in the negative regulation of immunoreceptor signaling. Molecular and cellular biology *20*, 2743-2754.

120. Levine, R.L., Wadleigh, M., Cools, J., Ebert, B.L., Wernig, G., Huntly, B.J., Boggon, T.J., Wlodarska, I., Clark, J.J., and Moore, S. (2005). Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. Cancer cell *7*, 387-397.

121. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

122. Li, J.J., Bickel, P.J., and Biggin, M.D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ *2*, e270.

123. Li, J.J., and Biggin, M.D. (2015). Statistics requantitates the central dogma. Science *347*, 1066-1067.

124. Liggins, A.P., Cooper, C.D., Lawrie, C.H., Brown, P.J., Collins, G.P., Hatton, C.S., Pulford, K., and Banham, A.H. (2007). MORC4, a novel member of the MORC family, is highly expressed in a subset of diffuse large B-cell lymphomas. British journal of haematology *138*, 479-486.

125. Lin, W.-W., and Karin, M. (2007). A cytokine-mediated link between innate immunity, inflammation, and cancer. Journal of Clinical Investigation *117*, 1175.

126. Linnemann, C., van Buuren, M.M., Bies, L., Verdegaal, E.M., Schotte, R., Calis, J.J., Behjati, S., Velds, A., Hilkmann, H., and el Atmioui, D. (2015). High-throughput epitope

discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma. Nature medicine *21*, 81-85.

127. Liu, S.Y., Sanchez, D.J., Aliyari, R., Lu, S., and Cheng, G. (2012). Systematic identification of type I and type II interferon-induced antiviral factors. Proceedings of the National Academy of Sciences of the United States of America *109*, 4239-4244.

128. Liyanage, U.K., Moore, T.T., Joo, H.-G., Tanaka, Y., Herrmann, V., Doherty, G., Drebin, J.A., Strasberg, S.M., Eberlein, T.J., and Goedegebuure, P.S. (2002). Prevalence of regulatory T cells is increased in peripheral blood and tumor microenvironment of patients with pancreas or breast adenocarcinoma. The Journal of Immunology *169*, 2756-2761.

129. Llobet, D., Eritja, N., Encinas, M., Llecha, N., Yeramian, A., Pallares, J., Sorolla, A., Gonzalez-Tallada, F.J., Matias-Guiu, X., and Dolcet, X. (2008). CK2 controls TRAIL and Fas sensitivity by regulating FLIP levels in endometrial carcinoma cells. Oncogene *27*, 2513-2524.

130. Lo, R.S. (2012). Receptor tyrosine kinases in cancer escape from BRAF inhibitors. Cell research *22*, 945-947.

131. Lu, R., Markowetz, F., Unwin, R.D., Leek, J.T., Airoldi, E.M., MacArthur, B.D., Lachmann, A., Rozov, R., Ma'ayan, A., and Boyer, L.A. (2009). Systems-level dynamic analyses of fate change in murine embryonic stem cells. Nature *462*, 358-362.

132. Lund, J.M., Hsing, L., Pham, T.T., and Rudensky, A.Y. (2008). Coordination of early protective immunity to viral infection by regulatory T cells. Science *320*, 1220-1224.

133. Lupfer, C., Thomas, P.G., Anand, P.K., Vogel, P., Milasta, S., Martinez, J., Huang, G., Green, M., Kundu, M., and Chi, H. (2013). Receptor interacting protein kinase 2-mediated mitophagy regulates inflammasome activation during virus infection. Nature immunology *14*, 480-488.

134. Lutz, M.B., Kukutsch, N., Ogilvie, A.L., Rößner, S., Koch, F., Romani, N., and Schuler, G. (1999). An advanced culture method for generating large quantities of highly pure dendritic cells from mouse bone marrow. Journal of immunological methods *223*, 77-92.

135. Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. FEBS letters *583*, 3966-3973.

136. Manghera, M., and Douville, R.N. (2013). Endogenous retrovirus-K promoter: a landing strip for inflammatory transcription factors? Retrovirology *10*, 16.

137. Mantovani, A., Allavena, P., Sica, A., and Balkwill, F. (2008). Cancer-related inflammation. Nature *454*, 436-444.

138. Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764-770.

139. Matsushita, H., Vesely, M.D., Koboldt, D.C., Rickert, C.G., Uppaluri, R., Magrini, V.J., Arthur, C.D., White, J.M., Chen, Y.-S., and Shea, L.K. (2012). Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. Nature *482*, 400-404.

140. Mayer, J., Blomberg, J., and Seal, R.L. (2011). A revised nomenclature for transcribed human endogenous retroviral loci. Mobile DNA *2*, 7.

141. Medema, J.P., de Jong, J., van Hall, T., Melief, C.J., and Offringa, R. (1999). Immune escape of tumors in vivo by expression of cellular FLICE-inhibitory protein. The Journal of experimental medicine *190*, 1033-1038.

142. Mellman, I., and Steinman, R.M. (2001). Dendritic cells: specialized and regulated antigen processing machines. Cell *106*, 255-258.

143. Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome biology *12*, R41.

144. Mertins, P., Qiao, J.W., Patel, J., Udeshi, N.D., Clauser, K.R., Mani, D., Burgess, M.W., Gillette, M.A., Jaffe, J.D., and Carr, S.A. (2013). Integrated proteomic analysis of post-translational modifications by serial enrichment. Nature methods *10*, 634-637.

145. Miranda, M., and Sorkin, A. (2007). Regulation of receptors and transporters by ubiquitination: new insights into surprisingly similar mechanisms. Molecular interventions *7*, 157.

146. Mlecnik, B., Bindea, G., Angell, H.K., Sasso, M.S., Obenauf, A.C., Fredriksen, T., Lafontaine, L., Bilocq, A.M., Kirilovsky, A., and Tosolini, M. (2014). Functional network pipeline reveals genetic determinants associated with in situ lymphocyte proliferation and survival of cancer patients. Science translational medicine *6*, 228ra237-228ra237.

147. Munoz, J., Low, T.Y., Kok, Y.J., Chin, A., Frese, C.K., Ding, V., Choo, A., and Heck, A.J. (2011). The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Molecular systems biology *7*.

148. National Institute of Health. TCGA Data Portal.

149. Nesvizhskii, A.I., Vitek, O., and Aebersold, R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. Nature methods *4*, 787-797.

150. Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G., Peters, B., Sette, A., and Lund, O. (2007a). NetMHCpan, a method for quantitative

predictions of peptide binding to any HLA-A and-B locus protein of known sequence. PloS one *2*, e796.

151. Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Roder, G., Peters, B., Sette, A., Lund, O., *et al.* (2007b). NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PloS one *2*, e796.

152. Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., and Shay, T. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296-309.

153. Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Molecular & cellular proteomics *1*, 376-386.

154. Ong, S.-E., and Mann, M. (2006). A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). Nature protocols *1*, 2650-2660.

155. Oshiumi, H., Sakai, K., Matsumoto, M., and Seya, T. (2010). DEAD/H BOX 3 (DDX3) helicase binds the RIG-I adaptor IPS-1 to up-regulate IFN-beta-inducing potential. European journal of immunology *40*, 940-948.

156. Pages, F., Berger, A., Camus, M., Sanchez-Cabo, F., Costes, A., Molidor, R., Mlecnik, B., Kirilovsky, A., Nilsson, M., Damotte, D., *et al.* (2005). Effector memory T cells, early metastasis, and survival in colorectal cancer. The New England journal of medicine *353*, 2654-2666.

157. Pagliarini, D.J., Calvo, S.E., Chang, B., Sheth, S.A., Vafai, S.B., Ong, S.-E., Walford, G.A., Sugiana, C., Boneh, A., and Chen, W.K. (2008). A mitochondrial protein compendium elucidates complex I disease biology. Cell *134*, 112-123.

158. Parsons, D.W., Jones, S., Zhang, X., Lin, J.C.-H., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.-M., and Gallia, G.L. (2008). An integrated genomic analysis of human glioblastoma multiforme. Science *321*, 1807-1812.

159. Pazdur, R. FDA Expands Approved Use of Opdivo (Nivolumab) for Lung Cancer.

160. Pearce, E.L., and Pearce, E.J. (2013). Metabolic pathways in immune cell activation and quiescence. Immunity *38*, 633-643.

161. Peng, M., Taouatas, N., Cappadona, S., van Breukelen, B., Mohammed, S., Scholten, A., and Heck, A.J. (2012). Protease bias in absolute protein quantitation. Nature methods *9*, 524-525.

162. Powles, T., Eder, J.P., Fine, G.D., Braiteh, F.S., Loriot, Y., Cruz, C., Bellmunt, J., Burris, H.A., Petrylak, D.P., and Teng, S.-l. (2014). MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer. Nature *515*, 558-562.

163. Pozzi, F., Di Matteo, T., and Aste, T. (2012). Exponential smoothing weighted correlations. The European Physical Journal B-Condensed Matter and Complex Systems *85*, 1-21.

164. Preiss, T., and Hentze, M.W. (1998). Dual function of the messenger RNA cap structure in poly (A)-tail-promoted translation in yeast. Nature *392*, 516-520.

165. Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., and Friedman, N. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nature biotechnology *29*, 436-442.

166. Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D.J., Pauli, A., Hacohen, N., Schier, A.F., Blackshear, P.J., and Friedman, N. (2014). High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA Regulatory Strategies. Cell *159*, 1698-1710.

167. Rajasagi, M., Shukla, S.A., Fritsch, E.F., Keskin, D.B., DeLuca, D., Carmona, E., Zhang, W., Sougnez, C., Cibulskis, K., Sidney, J., *et al.* (2014). Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. Blood *124*, 453-462.

168. Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nature protocols *2*, 1896-1906.

169. Ravi, R., and Bedi, A. (2002). Sensitization of tumor cells to Apo2 ligand/TRAIL-induced apoptosis by inhibition of casein kinase II. Cancer research *62*, 4180-4185.

170. Reverte, C.G., Ahearn, M.D., and Hake, L.E. (2001). CPEB degradation during Xenopus oocyte maturation requires a PEST domain and the 26S proteasome. Developmental biology *231*, 447-458.

171. Rizvi, N.A., Hellmann, M.D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J.J., Lee, W., Yuan, J., Wong, P., and Ho, T.S. (2015). Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. Science *348*, 124-128.

172. Robert, C., Karaszewska, B., Schachter, J., Rutkowski, P., Mackiewicz, A., Stroiakovski, D., Lichinitser, M., Dummer, R., Grange, F., and Mortier, L. (2015). Improved overall survival in melanoma with combined dabrafenib and trametinib. New England Journal of Medicine *372*, 30-39.

173. Rogers, S., Wells, R., and Rechsteiner, M. (1986). Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. Science *234*, 364-368.

174. Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G., and Hacohen, N. (2015). Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. Cell *160*, 48-61.

175. Russell, J.H., and Ley, T.J. (2002). Lymphocyte-mediated cytotoxicity. Annual review of immunology *20*, 323-370.

176. Rutledge, W.C., Kong, J., Gao, J., Gutman, D.A., Cooper, L.A., Appin, C., Park, Y., Scarpace, L., Mikkelsen, T., Cohen, M.L., *et al.* (2013). Tumor-infiltrating lymphocytes in glioblastoma are associated with specific genomic alterations and related to transcriptional class. Clinical cancer research : an official journal of the American Association for Cancer Research *19*, 4951-4960.

177. Saiki, Y., Ohtani, H., Naito, Y., Miyazawa, M., and Nagura, H. (1996). Immunophenotypic characterization of Epstein-Barr virus-associated gastric carcinoma: massive infiltration by proliferating CD8+ T-lymphocytes. Laboratory investigation; a journal of technical methods and pathology *75*, 67-76.

178. Samuels, Y., and Ericson, K. (2006). Oncogenic PI3K and its role in cancer. Current opinion in oncology *18*, 77-82.

179. Sanada, T., Takaesu, G., Mashima, R., Yoshida, R., Kobayashi, T., and Yoshimura, A. (2008). FLN29 deficiency reveals its negative regulatory role in the Toll-like receptor (TLR) and retinoic acid-inducible gene I (RIG-I)-like helicase signaling pathway. Journal of Biological Chemistry *283*, 33858-33864.

180. Sato, E., Olson, S.H., Ahn, J., Bundy, B., Nishikawa, H., Qian, F., Jungbluth, A.A., Frosina, D., Gnjatic, S., Ambrosone, C., *et al.* (2005). Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer. Proceedings of the National Academy of Sciences of the United States of America *102*, 18538-18543.

181. Saturno, G., Valenti, M., De Haven Brandon, A., Thomas, G.V., Eccles, S., Clarke, P.A., and Workman, P. (2013). Combining trail with PI3 kinase or HSP90 inhibitors enhances apoptosis in colorectal cancer cells via suppression of survival signaling. Oncotarget *4*, 1185-1198.

182. Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811-1817.

183. Savarrio, L., Gibson, J., Dunlop, D., O'Rourke, N., and Fitzsimons, E. (1999). Spontaneous regression of an anaplastic large cell lymphoma in the oral cavity: first reported case and review of the literature. Oral oncology *35*, 609-613.

184. Scanlan, M.J., Gure, A.O., Jungbluth, A.A., Old, L.J., and Chen, Y.-T. (2002). Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. Immunological reviews *188*, 22-32.

185. Schmidt, E.K., Clavarino, G., Ceppi, M., and Pierre, P. (2009). SUnSET, a nonradioactive method to monitor protein synthesis. Nature methods *6*, 275-277.

186. Schmitt, K., Reichrath, J., Roesch, A., Meese, E., and Mayer, J. (2013). Transcriptional profiling of human endogenous retrovirus group HERV-K(HML-2) loci in melanoma. Genome biology and evolution *5*, 307-328.

187. Schreiber, R.D., Old, L.J., and Smyth, M.J. (2011). Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. Science *331*, 1565-1570.

188. Schrimpf, S.P., Weiss, M., Reiter, L., Ahrens, C.H., Jovanovic, M., Malmström, J., Brunner, E., Mohanty, S., Lercher, M.J., and Hunziker, P.E. (2009). Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. PLoS biology *7*, e1000048.

189. Schumacher, K., Haensch, W., Roefzaad, C., and Schlag, P.M. (2001). Prognostic significance of activated CD8(+) T cell infiltrations within esophageal carcinomas. Cancer research *61*, 3932-3936.

190. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337-342.

191. Schwartz, D.C., and Parker, R. (1999). Mutations in Translation Initiation Factors Lead to Increased Rates of Deadenylation and Decapping of mRNAs inSaccharomyces cerevisiae. Molecular and cellular biology *19*, 5247-5256.

192. Schwitalle, Y., Kloor, M., Eiermann, S., Linnebacher, M., Kienle, P., Knaebel, H.P., Tariverdian, M., Benner, A., and von Knebel Doeberitz, M. (2008). Immune response against frameshift-induced neopeptides in HNPCC patients and healthy HNPCC mutation carriers. Gastroenterology *134*, 988-997.

193. Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58-63.

194. Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., and Lu, D. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236-240.

195. Shankaran, V., Ikeda, H., Bruce, A.T., White, J.M., Swanson, P.E., Old, L.J., and Schreiber, R.D. (2001). IFNγ and lymphocytes prevent primary tumour development and shape tumour immunogenicity. Nature *410*, 1107-1111.

196. Shin, M.S., Kim, H.S., Kang, C.S., Park, W.S., Kim, S.Y., Lee, S.N., Lee, J.H., Park, J.Y., Jang, J.J., and Kim, C.W. (2002a). Inactivating mutations of CASP10 gene in non-Hodgkin lymphomas. Blood *99*, 4094-4099.

197. Shin, M.S., Kim, H.S., Lee, S.H., Lee, J.W., Song, Y.H., Kim, Y.S., Park, W.S., Kim, S.Y., Lee, S.N., and Park, J.Y. (2002b). Alterations of Fas-pathway genes associated with nodal metastasis in non-small cell lung cancer. Oncogene *21*, 4129-4136.

198. Shin, M.S., Park, W.S., Kim, S.Y., Kim, H.S., Kang, S.J., Song, K.Y., Park, J.Y., Dong, S.M., Pi, J.H., and Oh, R.R. (1999). Alterations of Fas (Apo-1/CD95) gene in cutaneous malignant melanoma. The American journal of pathology *154*, 1785-1791.

199. Shumway, S.D., Maki, M., and Miyamoto, S. (1999). The PEST Domain of IκBα Is Necessary and Sufficient forin Vitro Degradation by μ-Calpain. Journal of Biological Chemistry *274*, 30874-30881.

200. Simpson, A.J., Caballero, O.L., Jungbluth, A., Chen, Y.T., and Old, L.J. (2005). Cancer/testis antigens, gametogenesis and cancer. Nature reviews Cancer *5*, 615-625.

201. Soda, M., Choi, Y.L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S.-i., Watanabe, H., Kurashina, K., and Hatanaka, H. (2007). Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. Nature *448*, 561-566.

202. Song, J.J., Kim, J.H., Sun, B.K., Alcala, M.A., Jr., Bartlett, D.L., and Lee, Y.J. (2010). c-Cbl acts as a mediator of Src-induced activation of the PI3K-Akt signal transduction pathway during TRAIL treatment. Cellular signalling *22*, 377-385.

203. Spearman, C. (1904). The proof and measurement of association between two things. The American journal of psychology *15*, 72-101.

204. Spranger, S., Spaapen, R.M., Zha, Y., Williams, J., Meng, Y., Ha, T.T., and Gajewski, T.F. (2013). Up-regulation of PD-L1, IDO, and Tregs in the melanoma tumor microenvironment is driven by CD8+ T cells. Science translational medicine *5*, 200ra116-200ra116.

205. Steen, H., and Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. Nature reviews Molecular cell biology *5*, 699-711.

206. Steinman, R.M., and Banchereau, J. (2007). Taking dendritic cells into medicine. Nature *449*, 419-426.

207. Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T.K., Hein, M.Y., Huang, S.-X., Ma, M., Shen, B., Qian, S.-B., and Hengel, H. (2012). Decoding human cytomegalovirus. Science *338*, 1088-1093.

208. Strauss, D.C., and Thomas, J.M. (2010). Transmission of donor melanoma by organ transplantation. The lancet oncology *11*, 790-796.

209. Subtelny, A.O., Eichhorn, S.W., Chen, G.R., Sive, H., and Bartel, D.P. (2014). Poly (A)-tail profiling reveals an embryonic switch in translational control. Nature.

210. Tang, K.W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M., and Larsson, E. (2013). The landscape of viral expression and host gene fusion and adaptation in human cancer. Nature communications *4*, 2513.

211. Tasaki, T., Sriram, S.M., Park, K.S., and Kwon, Y.T. (2012). The N-end rule pathway. Annual review of biochemistry *81*, 261.

212. Textor, S., Fiegler, N., Arnold, A., Porgador, A., Hofmann, T.G., and Cerwenka, A. (2011). Human NK cells are alerted to induction of p53 in cancer cells by upregulation of the NKG2D ligands ULBP1 and ULBP2. Cancer research *71*, 5998-6009.

213. Topalian, S.L., Hodi, F.S., Brahmer, J.R., Gettinger, S.N., Smith, D.C., McDermott, D.F., Powderly, J.D., Carvajal, R.D., Sosman, J.A., Atkins, M.B., *et al.* (2012). Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. The New England journal of medicine *366*, 2443-2454.

214. Tumeh, P.C., Harview, C.L., Yearley, J.H., Shintaku, I.P., Taylor, E.J., Robert, L., Chmielowski, B., Spasic, M., Henry, G., Ciobanu, V., *et al.* (2014). PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature *515*, 568-571.

215. Uderhardt, S., Herrmann, M., Oskolkova, O.V., Aschermann, S., Bicker, W., Ipseiz, N., Sarter, K., Frey, B., Rothe, T., Voll, R., *et al.* (2012). 12/15-lipoxygenase orchestrates the clearance of apoptotic cells and maintains immunologic tolerance. Immunity *36*, 834-846.

216. UNC Lineberger Comprehensive Cancer Center (2013). TCGA mRNA-seq Pipeline for UNC data (https://cghub.ucsc.edu/docs/tcga/UNC_mRNAseq_summary.pdf) (CGHub).

217. University of California, S.C. (2012). Cancer Genomics Hub (CGHub) (National Cancer Institute).

218. Uyttenhove, C., Pilotte, L., Theate, I., Stroobant, V., Colau, D., Parmentier, N., Boon, T., and Van den Eynde, B.J. (2003). Evidence for a tumoral immune resistance mechanism

based on tryptophan degradation by indoleamine 2,3-dioxygenase. Nature medicine *9*, 1269-1274.

219. Vakkila, J., and Lotze, M.T. (2004). Inflammation and necrosis promote tumour growth. Nature Reviews Immunology *4*, 641-648.

220. Vogel, C., de Sousa Abreu, R., Ko, D., Le, S.Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., and Penalva, L.O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Molecular systems biology *6*.

221. Vogel, C., Silva, G.M., and Marcotte, E.M. (2011). Protein expression regulation under oxidative stress. Molecular & Cellular Proteomics *10*, M111. 009217.

222. Wang, G., Ahmad, K.A., and Ahmed, K. (2006). Role of protein kinase CK2 in the regulation of tumor necrosis factor-related apoptosis inducing ligand-induced apoptosis in prostate cancer cells. Cancer research *66*, 2242-2249.

223. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., and Network, C.G.A.R. (2013). The cancer genome atlas pan-cancer analysis project. Nature genetics *45*, 1113-1120.

224. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., and Marx, H. (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582-587.

225. Wu, G., Nie, L., and Zhang, W. (2008). Integrative analyses of posttranscriptional regulation in the yeast Saccharomyces cerevisiae using transcriptomic and proteomic data. Current microbiology *57*, 18-22.

226. Yan, Z., Cui, K., Murray, D.M., Ling, C., Xue, Y., Gerstein, A., Parsons, R., Zhao, K., and Wang, W. (2005). PBAF chromatin-remodeling complex requires a novel specificity subunit, BAF200, to regulate expression of selective interferon-responsive genes. Genes & development *19*, 1662-1667.

227. Yang, Y.H., and Speed, T. (2002). Design issues for cDNA microarray experiments. Nature Reviews Genetics *3*, 579-588.

228. Yewdell, J.W., and Bennink, J.R. (1999). Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses 1. Annual review of immunology *17*, 51-88.

229. Yosef, N., and Regev, A. (2011). Impulse control: temporal dynamics in gene transcription. Cell *144*, 886-896.

230. Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P.W., and Levine, D.A. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications *4*.

231. Young, G.R., Eksmond, U., Salcedo, R., Alexopoulou, L., Stoye, J.P., and Kassiotis, G. (2012). Resurrection of endogenous retroviruses in antibody-deficient mice. Nature *491*, 774-778.

232. Yu, P., Lubben, W., Slomka, H., Gebler, J., Konert, M., Cai, C., Neubrandt, L., Prazeres da Costa, O., Paul, S., Dehnert, S., *et al.* (2012). Nucleic acid-sensing Toll-like receptors are essential for the control of endogenous retrovirus viremia and ERV-induced tumors. Immunity *37*, 867-879.

233. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhang, C.-Z., Wala, J., and Mermel, C.H. (2013). Pan-cancer patterns of somatic copy number alteration. Nature genetics *45*, 1134-1140.