Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins

Orit Rozenblatt-Rosen1,2,4*, Rahul C. Deo1-4,5,*, Megha Padi1-6,7*, Guillaume Adelmann1-4,8*, Michael A. Calderwood9,10,11, Thomas Rolland1,9,10, Miranda Grace1,12, Amélie Dricot1,9,10, Manor Askenazi2,4,8, Maria Tavares1,12,3,15, Samuel J. Favey9,10,13, Fieda Abderazzaq1-6, Danielle Byrdsong1,9,10, Anne–Ruxandra Carvunis1,9,10, Aylee A. Chen1,12, Jingwei Cheng2,12,13, Mick Correll2,1, Melissa Duarte1,9,11, Chanyu Fan1,9,10, Mariet C. Feltkamp14, Scott B. Faricarlo1,4,8, Rachel Franchi1,9,15, Brijesh K. Gang2,14, Natali Gulbahce1,9,10,16, Tong Hao1,9,10, Amy M. Holthaus1,11, Robert James1,9,10, Anna Korkhin1,2,3, Larissa Litovchick1,2,3, Jessica C. Mar1,6,7, Theodore P. Pak1,18, Sabrina Rabello1,5,9,10,16, Renee Rubio1,6, Yun Shen9,10,16, Saurav Singh1,8, Jennifer M. Spangle1,12,17, Murat Tasan1,4,8, Shelly Wanamaker1,11, James T. Webber4,8, Jennifer Roecklein–Canfield9,15, Eric Johannsen1,11, Albert–László Barabási1,9,16, Rameen Beroukhim1,3,18, Elliott Kiefel1,17, Michael E. Cusick1,17, David E. Hill1,3,10, Karl Münger1,12, Jarrod A. Marto1,4,8, John Quackenbush1,6,7, Frederick P. Roth1,4,9,13, James A. DeCaprio1,2,3 & Marc Vidal9,10,

Genotypic differences greatly influence susceptibility and resistance to disease. Understanding genotype–phenotype relationships requires that phenotypes be viewed as manifestations of network properties, rather than simply as the result of individual genomic variations. Genome sequencing efforts have identified numerous germline mutations, and large numbers of somatic genomic alterations, associated with a predisposition to cancer. However, it remains difficult to distinguish background, or ‘passenger’, cancer mutations from causal, or ‘driver’, mutations in these data sets. Human viruses intrinsically depend on their host cell during the course of infection and can elicit pathological phenotypes similar to those arising from mutations. Here we test the hypothesis that genomic variations and tumour viruses may cause cancer through related mechanisms, by systematically examining host interactome and transcriptome network perturbations caused by DNA tumour virus proteins. The resulting integrated viral perturbation data reflects rewiring of the host cell networks, and highlights pathways, such as Notch signalling and apoptosis, that go awry in cancer. We show that systematic analyses of host targets of viral proteins can identify cancer genes with a success rate on a par with their identification through functional genomics and large-scale cataloguing of tumour mutations. Together, these complementary approaches increase the specificity of cancer gene identification. Combining systems-level studies of pathogen-encoded gene products with genomic approaches will facilitate the prioritization of cancer-causing driver genes to advance the understanding of the genetic basis of human cancer.

Integrative studies of viral proteins have identified host perturbations relevant to the aetiology of viral disease. We examined whether such a strategy, extended systematically across a range of tumour viruses, could shed light on cancers even beyond those directly caused by these pathogens. Our hypothesis is inspired by classical examples in which DNA tumour virus proteins physically target the products of RB1 or TP53, two well-established germline-inherited and somatically inactivated tumour-suppressor genes. We propose that viruses and genomic variations alter local and global properties of cellular networks in similar ways to cause pathological states. Models derived from host perturbations mediated by viral proteins representing the virome should serve as surrogates for network perturbations that result from large numbers of genomic variations, or the variome.

We developed an integrated pipeline to systematically investigate perturbations of host interactome and transcriptome networks induced by gene products of four functionally related, yet biologically distinct, families of DNA tumour viruses: human papillomavirus (HPV), Epstein–Barr virus (EBV), adenovirus (Ad5) and polyomavirus (PyV) (Fig. 1b, Supplementary Table 1 and Supplementary Fig. 1). Applying a stringent implementation of the yeast two-hybrid (Y2H) system, 123 viral open reading frames (viral ORFs) were screened against a collection of about 13,000 human ORFs, resulting in a validated virus–host interaction network of 454 binary interactions involving 53 viral proteins and 307 human proteins (Fig. 1c and Supplementary Table 2). Analysis of our binary interaction map identified 31 host target proteins that showed more binary interactions with viral proteins than would be expected given their ‘degree’ (number of interactors) in our current binary map of the human interactome network (HI-2.0) (Fig. 1c, Supplementary Table 3 and Supplementary Notes), suggesting a set of common mechanisms by which different viral proteins rewire the host interactome network.

To examine perturbations in both the interactome and transcriptome networks directly in human cells, we generated expression constructs fusing each viral ORF to a tandem epitope tag and introduced each construct into IMR-90 normal human diploid fibroblasts. Interactions between viral proteins and the host proteome were identified by tandem affinity purification followed by mass spectrometry (TAP–MS). The
intersection of two independent TAP–MS experiments yielded 3,787 reproducibly mapped virus–host protein complex associations involving 54 viral proteins and the products of 1,079 host genes (Supplementary Table 4). Statistically significant overlaps between the Y2H and the TAP–MS data sets with a positive reference set of curated virus–host interactions were observed, suggesting the quality of the interactome data sets (Supplementary Notes). Host proteins identified as binary interactors or as members of protein complexes showed a statistically significant overlap (P < 0.001) and a statistically significant tendency to interact with each other in HI-2 (P < 0.001), implying that host targets in the virus–host interactome maps tend to fall in the same ‘neighbourhood’ of the host network (Supplementary Fig. 2). Our two complementary interactome data sets highlight specific host biological processes targeted by viral proteins (Supplementary Fig. 3).

To explore the specificity of virus–host relationships, we examined protein complex associations mediated by E6 proteins from six distinct HPV types representing three different disease classes: high-risk mucosal, low-risk mucosal and cutaneous. E6 and E7 proteins encoded by high-risk mucosal HPVs are strongly oncogenic. Multiple host proteins were found to associate with E6 proteins encoded by two or more different HPV types (P < 0.001; Fig. 1d), including the known E6 target UBE3A (E6AP). Among these we observed a statistically significant subgroup of host proteins targeted only by E6 proteins from the same disease class (P < 0.001). The transcriptional regulators CREB-binding protein (CREBBP) and EP300 were found to associate with E6 proteins from both cutaneous HPV types, but not with those from the mucosal classes. In contrast with E6 proteins, no group of host proteins showed class-specific targeting by HPV E7 proteins (Supplementary Fig. 4). These differential associations reflect how rewiring of virus–host interactome networks may relate to the aetiology of viral disease.

In addition to targeting protein–protein interactions, viral proteins also functionally perturb their hosts through downstream effects on gene expression. We profiled the transcriptome of the viral ORF-transduced cell lines to trace pathways through which viral proteins could alter cellular states. Model-based clustering of the 2,944 most frequently perturbed host genes identified 31 clusters, many of which were enriched (P < 0.01) for specific Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Fig. 2a and Supplementary Tables 5 and 6). To uncover transcription factor (TF) binding motifs enriched within the promoters or enhancers of the corresponding genes, a high-confidence map of predicted TF binding sites was generated using information on cell-specific chromatin accessibility and consensus TF-binding motifs (Supplementary Fig. 5). We found a densely interconnected set of 92 TFs (Supplementary Fig. 6 and Supplementary Notes) that either associated with or were differentially expressed in response to viral proteins, and for which target genes were enriched in at least one cluster (Supplementary Fig. 7 and Supplementary Table 7).

The mean expression change of each cluster revealed three distinct groups of viral proteins (Fig. 2a): group I included low-risk and cutaneous HPV E6 proteins, group II contained most of the EBV proteins, and group III included high-risk HPV E6 and E7 proteins and polyomavirus proteins. Consistent with their ability to associate with RB1, group III viral proteins increased the expression of genes that are involved in cell proliferation and whose promoters are enriched in E2F binding sites (clusters C26 and C31). Steady-state levels of these genes are correlated with cellular growth phenotypes (Supplementary Fig. 8 and Supplementary Notes). Similarly, the decreased expression of the p53 signalling pathway probably reflects the ability of group III proteins to bind to and inactivate p53 (cluster C12).

To investigate additional pathways through which viral proteins perturb TFs to reprogram cellular states, we derived a detailed network model containing 58 viral proteins that perturb the activity of 86 TFs, which in turn potentially regulate 30 clusters (Supplementary Fig. 9). This model was predictive of downstream patterns of differential expression (P = 0.003; Fig. 2b) and suggested ways in which viral proteins could regulate many of the biological hallmarks of cancer (Supplementary Notes). For example, we found the regulation of several pathways involved in the response to DNA damage (Fig. 2b), including autophagy potentially through NFE2L2 (cluster C31), the NRF-x-mediated inflammatory response (cluster C23), and the type I interferon response through IRF-1 (cluster C24; Supplementary Fig. 5).

Specific disease outcomes of the three disease classes of HPV might reflect how their respective proteins perturb distinct functional groups of host proteins. Our protein complex interaction map revealed associations between E6 from cutaneous HPVs and Mastermind-like protein 1 (MAML1), EP300 and CREBBP. MAML1 forms a transcriptional activation complex that modulates the expression of Notch target genes in conjunction with the EP300 and CREBBP histone acetyltransferases, the RBPJ transcription factor and the intracellular domain (ICD) of the Notch receptor. Our transcriptome profiling placed the cutaneous
and low-risk HPV E6 proteins in group I apart from the high-risk HPV E6 proteins in group II (Fig. 2a), so we investigated these differential perturbations. Both cutaneous HPV5 and HPV8 E6 proteins co-precipitated MAML1 and EP300, whereas the mucosal HPV E6 proteins did not (Fig. 3a and Supplementary Fig. 10); conversely, HPV6b, HPV11, HPV16 and HPV18 E6 proteins associated with MAML1 and EP300, whereas the mucosal HPV E6 proteins did not.

Perturbations in Notch signalling can confer either oncogenic or tumour-suppressive effects. Because both inhibition of the Notch pathway and the expression of HPV8 E6 promote squamous cell carcinoma, we reasoned that binding of HPV5 and HPV8 E6 to MAML1 might inhibit Notch signalling. To test this, we examined transcript levels of Notch pathway genes and potential Notch target genes with a predicted RBPJ binding site in their promoter across all HPV E6 cell lines as well as in cells depleted for MAML1. Transcript levels of several Notch targets were significantly decreased in IMR-90 cells that were either depleted for MAML1 or expressing either HPV5 or HPV8 E6 (Fig. 3b and Supplementary Fig. 11). These and other results indicate that the association of HPV5 and HPV8 E6 proteins with MAML1 inhibits Notch signalling. Building on these observations and on the established associations between EBV EBNA proteins with MAML1 inhibits Notch signalling. Building on these observations and on the established associations between EBV EBNA proteins and RBPJ, we observed that viral proteins from all four DNA tumour viruses target proteins of the Notch pathway (Fig. 1c).

Viral ORF grouping. Grey blocks show which viral proteins associate with the indicated host proteins. a, Diagram showing how the viral protein–TF–target-gene network was constructed, with paths to three clusters shown. C, cluster; V, viral protein. Null distribution of average fraction of TF target genes differentially expressed in the corresponding cell lines (histogram), along with the observed value (green arrow).
genes (Supplementary Fig. 12; adjusted $P$ for multiple hypothesis testing ($P_{adj} = 0.01$). To optimize the stringency of potential cancer enrichment analyses, we restricted the set of viral protein targets identified by TAP–MS to those identified by three or more unique peptides, a choice corresponding to an experimental reproducibility rate greater than 90% (Supplementary Fig. 13). The resulting stringent candidate set of 947 host target genes (the ‘VirHost’ set; Supplementary Table 9) included 16 proteins encoded by CC genes ($P = 0.007$; Fig. 4a), among which tumour suppressor genes were significantly over-represented ($P = 0.03$).

As a complementary approach to validate our VirHost gene set, we compiled a list of human orthologues of mouse genes recently implicated in tumorigenesis by in vivo transposon mutagenesis screens. Our VirHost data set overlaps significantly with these candidate cancer genes ($P < 0.0001$) (Fig. 4b and Supplementary Table 10). The 156 candidate genes in the overlap were markedly enriched both for CC genes (odds ratio (OR) = 13, $P = 4 \times 10^{-15}$) and for genes implicated in apoptosis, hypoxia response and cell growth pathways ($P_{adj} < 0.05$ for all). Taken together, these observations suggest that our VirHost data set points to previously unknown human cancer-associated genes.

Large-scale tumour-sequencing efforts have the potential to discover new tumour suppressors and oncogenes. To explore how the VirHost data set might be used to interpret these data, we compiled somatic mutations for eight different cancers identified through 12 sequencing projects. Non-synonymous somatic mutations were reported for 10,543 genes, so further prioritization is required for useful identification of candidate causal cancer genes. We therefore scored the likely functional effects of these mutations by using the PolyPhen2 program and generated a cumulative somatic mutation (SM) score for each protein (Supplementary Fig. 14). To compare performance in identifying candidate cancer genes of our VirHost set with that of proteins ranked by SM analysis, we tested a matching number (947) of the top-ranked SM candidates for overlap with CC genes (Fig. 4c and Supplementary Table 11). In comparison with the 16 cancer genes identified in our VirHost set, SM recovered 23 genes linked to ‘regulation of apoptosis’ (OR = 0.017). The intersection also includes plausible contributors to cancer pathogenesis.

Although both strategies showed significant overlap with the reference CC gene set, neither by itself suffices to pinpoint causal genes with high specificity. To overcome this difficulty we exploited the orthogonal nature of the VirHost and SM sets (given $P = 0.58$ for their overlap) by focusing on the 43 proteins at their intersection (the ‘VirHostSM’ subset) (Fig. 4c). In contrast with VirHost (OR = 3.7) or SM (OR = 5.8), the VirHostSM set was markedly enriched in CC genes (five proteins, $OR = 26$, $P = 3 \times 10^{-15}$). Pathway analysis of the 43 proteins revealed 12 proteins implicated in the GO pathway linked to ‘regulation of apoptosis’ (OR = 0.6, $P_{adj} = 0.0017$). The intersection also includes plausible contributors to cancer pathogenesis (Supplementary Fig. 15) such as the oxidative stress response transcription factor NFE2L2.

We compared the ability of VirHost to identify CC genes with two other large-scale genomic approaches: SCNA (somatic copy number alteration)27 analysis of cancer genomes and GWAS (genome-wide association studies) of cancer susceptibility28. The SCNA deletions (SCNA–DEL) and amplifications (SCNA–AMP) and GWAS sets all significantly overlapped with CC genes, but with lower specificity than...
the VirHost overlap with CC (OR = 1.9 for SCNA-DEL, 2.1 for SCNA-AMP, and 3.1 for GWAS, versus 3.7 for VirHost; Fig. 4d–f). The intersections of VirHost with GWAS or SCNA-DEL genes also showed enrichment for cancer genes (Supplementary Table 12). The intersection of VirHost and SCNA-DEL was enriched for genes implicated in apoptosis (GO term ‘programmed cell death’; 15 genes, 

\[ P_{adj} = 0.022, \text{ OR} = 4.3 \] ) Conversely, there was no synergy in the intersection of SCNA-AMP and VirHost, perhaps reflecting the preference of viral proteins in targeting tumour suppressors rather than oncogenes (Fig. 4f).

Our systems-level explorations of viral perturbations facilitate the distinction between driver and passenger mutations in cancer genome sequences. Our data indicate that trans-acting viral products and cis-acting genome variations involved in cancer converge on common pathways.

**METHODS SUMMARY**

Viral ORF entry clones were generated by PCR-based Gateway recombinational clonings. After sequence verification, viral ORFs were transfected by *in vitro* Gateway LR recombinational cloning into expression vectors for Y2H screening and for transduction of IMR-90 cells. Y2H screens were performed against the human ORFeome v5.1 collection of about 13,000 full-length human ORFs. Total RNA was isolated from IMR-90 cells expressing viral ORFs, and gene expression was assayed on Human Gene 1.0 ST arrays. Microarray data were analysed with R/Bioconductor. Viral proteins and associated host proteins were purified by sequential Flag and haemagglutinin immunoprecipitation and analysed by liquid chromatography–tandem mass spectrometry. Virus–host protein complex associations from two independent purifications were analysed. Pathway enrichment was analysed with FuncAssociate. Assessment of statistical significance for overlap between gene sets was performed with Fisher’s exact test or resampling-based approaches.

Received 8 June 2011; accepted 7 June 2012.

**Published online 18 July 2012.**


