Published in final edited form as:
Nat Methods. 2012 December ; 9(12): 1207–1211. doi:10.1038/nmeth.2227.

De novo derivation of proteomes from transcriptomes for transcript and protein identification

Vanessa C. Evans[1], **Gary Barker**[2], **Kate J. Heesom**[3], **Jun Fan**[4], **Conrad Bessant**[4], and **David A. Matthews**[1]

[1]School of Cellular and Molecular Medicine, University of Bristol, University Walk, Bristol. BS8 1TD. UK.

[2]School of Biological Sciences, University of Bristol, University Walk, Bristol. BS8 1TD. UK.

[3]School of Biochemistry, University of Bristol, University Walk, Bristol. BS8 1TD. UK.

[4]Bioinformatics Group, Cranfield Health, Cranfield University, Cranfield, Bedfordshire. MK43 0AL. UK.

Abstract

Identification of proteins by tandem mass spectrometry requires a database of the proteins that could be in the sample. This is available for model species (e.g. humans) but not for non-model species. Ideally, for a non-model species the sequencing of expressed mRNA would generate a protein database for mass spectrometry based identification, allowing detection of genes and proteins using high throughput sequencing and protein identification technologies. Here we use human cells infected with human adenovirus as a complex and dynamic model to demonstrate this approach is robust. Our Proteomics Informed by Transcriptomics technique identifies >99% of over 3700 distinct proteins identified using traditional analysis reliant on comprehensive human and adenovirus protein lists. This facilitates high throughput acquisition of direct evidence for transcripts and proteins in non-model species. Critically, we show this approach can also be used to highlight genes and proteins undergoing dynamic changes in post transcriptional protein stability.

Introduction

Modern deep sequencing techniques capture the transcriptome of any organism in unprecedented detail and there have been substantial breakthroughs in the *de novo* assembly of transcriptomes. Indeed, de novo assembly from raw sequence data has clear benefits for gene identification and the functional annotation of genomes– especially of non-model species[1, 2]. In parallel, improvements in high-throughput Liquid Chromatography coupled tandem mass spectrometry (LC-MS/MS) allows identification of several thousand distinct

Corresponding Author: David A. Matthews, d.a.matthews@bristol.ac.uk.

Author Contributions VE co-wrote the manuscript, prepared infected cells, performed western blots and assisted with immunofluorescence. GB co-wrote the manuscript, wrote software (implement_snp_eff_changes.pl) and assisted with handling the RNAseq data. KH performed the mass spectrometry and assisted with analysis of the MS/MS data. JF helped with the BLAST analysis and wrote some of the BLAST search software. CB co wrote the manuscript and assisted with the MS/MS analysis, BLAST database searches. DM conceived the experiments, conceived the PIT analysis pipeline, led the manuscript writing, wrote software (sam_to_GFF3_and_orfs.pl, pep_to_sam.pl, put_fpkm_values_back.pl, Connect_maxQ_peptides_to_trinity_fasta_files.pl and part of implement_snp_eff_changes.pl and UNIXbatchBlastAndParse_CHO_analysis.pl), assisted with the immunofluorescence and the preparation of infected cells, carried out manual curation, analysis and integration of the data.

Accession codes. The raw sequence reads have been deposited with ArrayExpress at the European Bioinformatics Institute with the accession number E-MTAB-1277. The unprocessed spectra files (in ThermoFisher .raw format) for the human and CHO experiments can be accessed from Canfield University, IP address: 138.250.31.74(port 22), Login: anonymous, Password: anonymous.

Europe PMC Funders Author Manuscripts

proteins from a total cell extract in a single experiment[3, 4]. Combined with quantitative techniques such as Stable Isotope Labelling of Amino acids in Cell culture (SILAC), changes in relative protein levels over time can also be monitored. However, MS/MS data analysis software normally requires an accurate list of proteins that could be present in the sample[5]. This is reasonably straightforward for model organisms (e.g. humans) but not for poorly annotated species, for species where the genome is not yet fully assembled or samples containing proteins from multiple species. Moreover, this approach is not optimised for individual variation caused by SNPs.

A small number of recent publications combine deep sequencing with LC-MS/MS[6, 7]. One study characterised different human cell lines using deep sequencing based transcriptomics (RNAseq) and quantitative proteomics, showing a high correlation between changes in transcript and protein abundance[6]. Another combined sequence analysis of the genome, transcriptome and proteome of human B cells, principally looking for SNP changes[7]. Using a transcriptome to tailor a proteomic analysis would be highly desirable in a range of situations, especially in non-model systems. Research in non-model species is hampered because their transcriptomes and proteomes are, by necessity, annotated predominantly by computationally driven searches for genes and proteins rather than by experimentally derived observations. This clearly has limitations - confidently identifying highly novel proteins in non-model species is particularly challenging with this approach. To alleviate this, proteogenomics is often used to try to improve the identification of proteins in non-model species[8]. This typically relies on the translation of predicted gene models and an all-frames translation of the target genome to generate databases of predicted proteins. These databases are used by ms/ms spectra search engines to positively identify peptides. Whilst these approaches are highly informative they require a good quality copy of the genome in question. Moreover, as the target genome increases in size, the size of the database of possible proteins in all frames becomes increasingly unwieldy. One of the largest proteogenomics analysis attempted to date is on *Medicago truncatula*[9] with a genome of ~0.6 Gbp in size, substantially smaller than the human genome (~3 Gbp) or the wheat genome (~16 Gbp).

We used a highly annotated two genome system (human adenovirus infected human cells) as a robust bench mark to show that RNAseq data can directly inform proteomic analyses allowing the acquisition of transcriptomic plus proteomic data for any given species (and, if present, associated pathogens). Our method is called Proteomics Informed by Transcriptomics (PIT analysis) and it produced extensive, experimentally derived data on transcription and protein content in a complex and dynamic system. PIT analysis provided a high throughput method to derive transcripts, infer proteins from them and show that the proteins are detected by MS/MS - enabling a seamless visualisation of data about the genome, the transcriptome and proteome. We compared the outputs of this technique to what could be generated using standard methods in this well understood and highly annotated system. We recovered the vast majority of information possible from both the virus and human samples in a manner that is independent of pre-existing datasets and, in principle, independent of a copy of the target genome. We also showed that our approach is robust, coping with the transcriptomic plus proteomic data from the virus and the human cell as they evolved over time. Moreover, this integrated approach enabled us to examine the post transcriptional stability of proteins. Adenovirus induces a viral ubiquitin ligase complex that specifically degrades many cellular targets and is key to several aspects of efficient viral replication including underpinning the cancer killing phenotype of the oncolytic adenovirus, ONYX-015[10-12]. Our combined approach identifies both previously described viral ubiquitin ligase targets and high value novel candidates by highlighting proteins that substantially decline in abundance without any corresponding transcriptional downturn.

RESULTS

Sample collection

To collect a matched set of samples we infected human cells with adenovirus and collected samples 8 and 24 hours after infection alongside an uninfected control. The cells were metabolically labelled by SILAC, enabling protein quantitation over time, and at each time point the samples were split in two allowing collection of protein and RNA from the same sample for proteomic and RNAseq analysis. Thus, three flasks of HeLa cells were grown in SILAC culture media saturating all cellular proteins with the appropriate label. One flask was labelled with ^{15}N and ^{13}C labelled arginine and lysine (heavy HeLa), one with ^{13}C labelled arginine and lysine (medium HeLa) and one with normal isotopes (light HeLa). The medium and light HeLa cells were infected with adenovirus and the heavy HeLa cells were mock infected. At 8 hours post infection the light HeLa cells were harvested for protein and RNA. At 24 hours post infection the medium (adenovirus infected) and heavy (mock infected) cells were similarly harvested.

SILAC based quantitative proteomics

The three protein samples were combined on a 1:1:1 ratio before separation by SDS-PAGE and processing for LC-MS/MS analysis. The MS/MS spectra was analysed by MaxQuant software[13] to identify proteins and quantitate abundance changes. Hela cells are cervical carcinoma derived, containing genes from Human papillomavirus type 18 (HPV18) integrated into the cell genome. We searched for HPV18 proteins without success, but we did detect adenovirus proteins. Of 3,818 proteins identified, 3,411 were identified and quantitated by two or more distinct peptides (Supplementary Table 1). Of those, only about 1% showed a twofold or greater increase or decrease in abundance over the first 8 hours of the experiment and just under 8% had altered their abundance by twofold at 24 hours. We were able to detect a threefold increase in *HSPA1A* in the first 8 hours post infection (Hsp70), a gene and protein known to be induced early on by adenovirus infection[14]. We were also able to detect a greater than twofold decrease in *MRE11*, ITGA3 and *RAD50*, all known to be degraded during infection[15, 16]. We previously reported that levels of Upstream Binding Transcription Factor (*UBTF*) remain unchanged during adenovirus infection, something reflected in this dataset[17].

RNAseq analysis of adenovirus infected cells

Cytoplasmic mRNA was harvested from the same three samples of Hela cells because adenovirus inhibits nuclear export of cellular mRNA during infection without inhibiting its production. Each sample was sequenced (56bp paired end read) on an Illumina GAIIx generating a total of ~82 million reads from the three samples.

We imported our data into a locally installed Galaxy NGS software suite[18] and mapped our data using TopHat[19] to the human genome (hg19). Uniquely mapped reads were used for gene expression analysis with Cufflinks[20] using the Ensembl human gene annotation as a guide (v64). Separately we mapped the reads to human adenovirus type 5 (GI:56160529) and HPV18 (GI:30172004) genomes. During the experiment the number of reads mapped to the adenovirus genome raises to about 80% of the total (Table 1) illustrating how the virus transcriptome eventually dominates. We detected HPV18 transcripts at all timepoints noting that adenovirus infection inhibits HPV18 transcription as previously reported[21]. The pattern of reads mapped to the adenovirus genome matches expectations. Thus, at 8 hours post infection most reads map to adenovirus early genes (E1, E2, E3 and E4) whereas by 24 hours most reads map to the late genes derived from the virus major late promoter (Supplementary Figure 1). Turning to human gene expression analysis, comparison of our analysis to a previous experiment using microarrays provides confidence that this approach

is robust. For example, in that experiment[22] one of the few genes upregulated by adenovirus infection at early times is CDC25A and our data reflects this as well as expected rises in HSPA1A expression noted previously. Recently an RNAseq based transcriptomic analysis of human cells infected with adenovirus was published[23]. There are experimental differences, notably a different cell type and using total mRNA extract in that paper instead of cytoplasmic mRNA in this report. However, their conclusions are broadly similar. For example, in the IkappaB family of NFkappaB inhibitors, *NFKBIE* declines whilst *NFKBIB* increases in both data sets.

Proteomics Informed by Transcriptomics (PIT)

We utilised Trinity[2] and a combined set of sequence reads from all three time points for the *de novo* assembly of the transcriptome (Supplementary Dataset 1). We then generated open reading frames (>200 nucleotides) from all six frames of each Trinity generated transcript (Supplementary Dataset 2). This "PIT proteins" list was used as our search database for the MaxQuant package. Comparing the peptides generated by a search of standard human proteomes revealed that a search using the PIT protein list generates almost as many identified peptides (~95%) as that from a canonical list of human proteins from Ensembl or from a non-redundant Swissprot-Uniprot list (Table 2 and Supplementary Table 2). In addition, in the list of peptides identified by searching the PIT proteins dataset we found 360 peptides that belonged to the adenovirus proteome compared to 367 peptides found by searching a standard adenovirus proteome derived from GenBank.

Next we mapped the Trinity transcripts to the human genome using GMAP[24] to generate a Sequence Alignment Map file (Supplementary Dataset 3) and added the identified peptide data from MaxQuant to the Trinity transcripts with in-house software. Finally we used in house software to create a GFF3 format file (Supplementary Dataset 4) combining the data in the SAM file with exon structure information. These SAM and GFF3 files allow us to see which peptides are associated with a transcript, which exon it was derived from, and the transcript's location on the human genome (Figure 1).

Our software also generates a list containing the longest open reading frame associated with each peptide positively identified by MS/MS (Supplementary Dataset 5). Thus, our approach starts with a list of possible proteins derived from the Trinity assembled transcripts and ends up with a list of full length proteins derived from the Trinity transcripts for which there is at least one peptide identified (workflow and software used summarised in Supplementary Figure 2). We derived a list of 7,319 unique proteins in this way, although there will be far fewer truly distinct proteins because any difference will be reported as multiple entries. For example, Trinity may assemble multiple transcripts for the same gene, some longer than others, some may appear to be distinct when in fact it is the same gene and protein. When we searched this list of positively identified proteins using BLAST we determined that all were either human or adenovirus derived proteins (Supplementary Table 3). In fact, our approach confirms the expression of more distinct genes than a traditional approach. We found 3,792 distinct human genes using the PIT proteins dataset vs 3,773 distinct human genes using the Ensembl dataset as the search space with 99.45% overlap in the two lists. There are some small differences in the proteins identified reflecting slight differences in how proteins are included in the canonical Ensembl database compared to Swissprot-Uniprot. Turning to the virus data, although the PIT analysis missed one adenovirus protein (U-exon protein), the PIT approach identified adenovirus proteins not identified by a traditional search of the adenovirus proteome. One example, the "i leader" adenovirus protein is a *bona fide* adenovirus protein[25-27] but is not present in the GenBank list of adenovirus serotype 5 proteins. This illustrates a key advantage of our approach to detecting transcripts and proteins.

PIT analysis of Chinese Hamster Ovary (CHO) cells

To illustrate the potential of the gene and protein identification aspect of PIT we examined CHO cells, which are widely used for protein expression purposes. We obtained a publically available RNAseq data set (European Nucleotide Archive SRP001851) and assembled the transcriptome using Trinity as before (Supplementary Dataset 6). From this we generated a list of proteins (Supplementary Dataset 7) and used this list to search spectra from a total protein extract of CHO cells separated by 1D gel electrophoresis prior to LC-MS/MS. Our PIT analysis shows that a search of the Trinity CHO proteins list compared to the standard UNIPROT list of CHO proteins leads to almost a doubling of the numbers of identified peptides (Supplementary Table 4). Moreover, we were able to infer a list of the largest open reading frames associated with each identified peptide and search these identified ORFS using BLAST to define the nearest homologues in the CHO, mouse and human UNIPROT lists (Supplementary Table 5). This list has 7,333 non-identical transcripts/proteins listed and many are likely to be minor variants of the same protein (indeed, BLAST searching indicates this list maps to approximately 5,672 different homologous mouse proteins). Finally, as with the human data we were able to map Trinity derived transcripts to the CHO genome[28] alongside the locations of identified peptides giving a seamless view of genome, transcript and identified peptides (Supplementary Figure 3 and Supplementary Datasets 8 and 9).

Detecting SNPS in the proteome

We analysed our TopHat alignments using snpEFF (Cingolani, P. “snpEff: Variant effect prediction”, http://snpeff.sourceforge.net, 2012.) to generate a list of non-synonymous SNPs from our RNAseq data. Using this and in house software we derived a list of canonical and variant proteins (Supplementary datasets 10, 11 and 12) to search our MS/MS spectra and to compare to our PIT protein and canonical lists. We were indeed able to correlate 170 SNP changes with detected peptides, 14 peptides where only the canonical sequence (and not the SNP variant) was detected as well as 14 heterologous transcripts where both a canonical and variant peptide were detected by MS/MS analysis (Supplementary Table 6). The majority (149 of 170) SNPs detected by this analysis were also detected by the PIT analysis.

Detecting post transcriptional degradation targets

Adenoviruses boost their replication by inducing the destruction of cellular proteins through modulation of ubiquitin ligase complexes. We wanted to see if known adenovirus induced ubiquitin ligase targets could be identified in our data by looking for proteins that declined in abundance two fold without a corresponding decline in mRNA abundance. Three proteins known to be degraded during adenovirus infection (*Mre11*, *RAD50* and *ITGA3*) were identified as meeting this criteria. From this we developed a short list of proteins (Table 3) whose abundance had fallen proteomically (i.e. ratio of 0.5 or less by 24 hours) without transcriptomic explanation (i.e. mRNA levels at 24 hours at least 0.8 of that in uninfected cells). In addition we checked the half life of these proteins was above 24 hours by consulting the publically available lists of protein half life’s as measured in HeLa cells[29]. Of these, *POLDIP3* (widely known as *SKAR*) was selected for further research since it is proposed to play a role in cellular mRNA export and translation[30] – both known to be affected by adenovirus in a manner dependent on the induction of a novel ubiquitin ligase complex. Indeed, *POLDIP3* is degraded in adenovirus infected cells and the degradation is sensitive to the proteasome inhibitor, MG132 (Figure 2a). In addition, during adenovirus infection, *POLDIP3* is sequestered from a speckled distribution in uninfected cells into track like structures (Figures 2b and 2c) similar to that reported for *MRE11*, a known target of adenovirus induced ubiquitin mediated degradation([31] figure 4 in that paper). Moreover, cells infected with adenovirus mutant dl366[32] which lacks the E4 region required for the

formation of a virally induced ubiquitin ligase complex, do not show any reorganisation of *POLDIP3* (Figure 2d).

Discussion

Here we demonstrated the potential of simultaneous capture of quantitative data on the transcriptome and proteome to study changes in a cell population under dynamic conditions. We show how to use RNAseq data to inform the protein identification process, which, in turn, validates the transcriptomic assembly. PIT analysis has broad utility in the study of a wide range of species where annotation of the genome is suboptimal, but particularly in the field of infections involving zoonosis or arthropod borne infections. This approach may also help to focus research efforts onto post transcriptional events as we have shown by examining proteins that degrade over time without a corresponding decline in mRNA expression.

More importantly, our work shows for the first time that *de novo* transcriptome assembly does generate a practical protein dataset that recovers essentially all of the detectable proteins derived from both the host and virus. By doing this research in a complex but well annotated system we provide confidence that PIT analysis is robust, valid, practical and could be usefully applied to non-model systems where direct experimental evidence of genes and proteins is often lacking. Indeed, in principle a genome is not required for gene and protein identification by PIT analysis.

We also performed a preliminary PIT analysis of proteomic and transcripts data from CHO cells using a Trinity generated list of based on a previously published RNAseq experiment also done in CHO cells[33]. Searching our Trinity derived ORFs list identifies around 70% more peptides compared to searching the current CHO (*Cricetulus griseus)* proteome downloaded from Uniprot (Supplementary Table 4). Moreover, we linked this list of proteins using BLAST to nearest homologues in the Uniprot hamster, mouse and human proteomes (Supplementary Table 5). This preliminary analysis illustrates that our approach does work in non-model systems and that PIT can be used with historical datasets. This analysis provides direct evidence for the transcription and expression of a wide range of genes and proteins in this important cell line (for example see Supplementary Figure 3).

A key primary limitation is the depth of sequencing done at the time and increases in data return from transcriptomic experiments will improve the proportion of proteins captured. We have repeated PIT analysis with declining quantities of raw transcriptomic data (Supplementary Table 7) showing how depth of transcriptome coverage influences data return. Even a reduced dataset of less than 10% of the coverage used here still yields over 70% of the peptides identified by our largest analysis. The importance of this is illustrated by the observation of cases where a peptide was identified by searching a canonical list but not the Trinity list because part of the transcript was missing in the Trinity assembly. PIT analysis may also help refine algorithms for *de novo* transcriptomic assembly – i.e. the best algorithms should yield the largest list of distinct peptides in a subsequent proteomic analysis.

Our PIT analysis also relates each identified peptide to the exon on each transcript and we are currently exploring ways of effectively interrogating the proteomic and transcriptomic data to identify and correlate changes in isoform expression.

As the sensitivity of MS/MS based sequencing increases the proportion of the possible available peptides that can be detected will increase[4]. Our ability to identify SNPs, although currently limited, will improve with improvements in MS/MS based proteomics implying that it will be increasingly attractive to base MS/MS searches on proteins derived from the

Europe PMC Funders Author Manuscripts

transcriptome rather than on canonical lists. We propose that such data is added to the well-established SAM and GFF3 file formats as the most flexible way forward to integrate these data sets since these file formats are widely supported.

Another attractive aspect of PIT is that by interrogating the two data sets in a different way, we identified components of the Double Stranded DNA Break Repair (DSBR) system, *RAD50* and *MRE11*, as being targets for virally mediated degradation. These effects are well established and are functionally important for the virus during replication. We next focussed on *POLDIP3* which is involved in mRNA export and translation[30]. The oncolytic phenotype of the adenovirus derived ONYX-015 virus is linked to the mRNA export pathway - adenovirus interferes with cellular mRNA export and translation in an unknown manner dependent on forming a unique ubiquitin ligase complex[12]. Our data implies that *POLDIP3* is linked to the ONYX-015 phenotype, something we are currently investigating. Critically, our analysis identifies previously known specific viral degradation targets as well as new cellular degradation targets.

This paper shows how to integrate the analysis of transcriptome and proteome offering important new insights, maximising return on the data and providing new tools for the study of both well-established and non-model species and their pathogens. Moreover, being able to rapidly annotate newly sequenced genomes with experimentally derived transcriptomic and proteomic data is highly desirable given the number of genome sequencing projects worldwide. We believe this approach will, alongside current approaches such as proteogenomics, improve gene and protein identification in non-model species as well as refining the application of high throughput technologies to the study of dynamic and/or multi genome systems. Finally, this technique should aid the development of systems approaches to biological research.

Methods

Cell culture, sample harvesting and viruses

HeLa cells were obtained from ECACC and grown in SILAC labelled DMEM with 10% v/v SILAC dialysed Foetal Calf Serum (Dundee Cell Products) for at least 5 population doublings. Approximately 3×10^7 cells were either mock infected or infected with wild type adenovirus serotype 5 at a multiplicity of infection of 30. After 1 hour exposure to the virus, the medium was replaced with fresh SILAC labelled medium and the infection allowed to continue for either 8 or 24 hours.

The cells were washed twice with PBS then treated with trypsin to release the adherent cells, washed a further two times in PBS before splitting the sample in half. One half of the sample was suspended in 0.5 ml of PBS, and 0.1 ml aliquots were stored at −70°C until needed for protein analysis. The other half was immediately processed for extraction of cytoplasmc RNA. Briefly, the cells were re-suspended in 0.5ml 0.1% Triton X-100 to lyse the cytoplasm. The nuclei were spun down and the cytoplasmic fraction was extracted with Trizol to obtain a total cytoplasmic RNA sample.

RNA seq

Prior to further processing for RNAseq, the three samples were used as substrates for PCR based test to confirm the presence of virus transcripts (adenovirus DBP gene) present in both the virus infected samples and not in the uninfected samples (Primer list in Supplementary Table 10). The three samples were labelled UN (uninfected control), T8 (8 hours post infection) and T24 (24 hours post infection). Next, the Trizol extracted RNA was extracted again using RNAeasy (Qiagen) prior to quantitation and processing for poly A+ selection and 56bp paired end sequencing on the University of Bristol Illumina GAIIx using the

manufacturers reagents and protocols. The sequencing data was then uploaded to the Galaxy suite of software for analysis, hosted on a local Galaxy instance at the University of Bristol High Performance Computing resource, BlueCrystal.

The raw sequence reads have been deposited with ArrayExpress at the European Bioinformatics Institute with the accession number E-MTAB-1277.

The paired end sequence data for each time point was initially mapped to a female hg19 (i.e. less the Y chromosome) using TopHat. The following parameters were set: Mean inner distance=80; standard deviation = 15; maximum mismatches in anchor region = 0; minimum intron length = 70; maximum intron length = 500000; allow indel search = yes; maximum insertion length = 3; maximum deletion length = 3; maximum alignments allowed = 40; minimum intron length that may be found during split-segment search = 50; maximum intron length that may be found during split-segment search: = 500000; number of mismatches allowed in the initial read mapping = 2; number of mismatches allowed in each segment alignment for reads mapped independently = 2; minimum length of read segments = 2; own Junctions = no; closure search = yes; exonic hops in splice graph minimum = 50; maximum intron length found by closure search = 5000; minimum intron length found by closure search = 50; coverage search = yes; minimum intron by coverage search = 50; maximum intron by coverage search = 20000.

Mapped reads were then filtered to retain only those reads that map in a proper pair before separating reads that mapped to one location from those that map to more than one location. Gene expression quantitation on uniquely mapping reads was performed using Cufflinks supplied with the Ensembl gtf (v64) as a reference throughout the analysis. The following parameters were set for Cufflinks:

Maximum intron length = 500000; minimum isoform fraction = 0.05; premRNA fraction = 0.05; quartile normalisation = yes; use reference annotation = yes; perform bias correction = yes; set parameters for paired end reads = no.

In addition to mapping to the human genome, Tophat was used to map to the adenovirus type 5 genome (AC_000008.1) and to the human papillomavirus serotype 18 (NC_001357.1) with the same parameters listed above but with the following changes:

Minimum intron length = 30; maximum intron length = 34000 (7000 for papillomavirus); minimum intron length that may be found during split-segment search = 10; maximum intron length that may be found during split-segment search: = 34000 (7000 for papillomavirus).

We also used the Trinity *de novo* assembly software installed on our local copy of the Galaxy suite with default parameters. For this analysis we combined all three time points of data into one large data set comprising ~82 million paired end reads. The output of assembled transcripts (~102,000 entries) was then translated (forward and reverse) into proteins using the EMBOSS tool “getorf” with a minimum nucleotide length of 200 bp between the start and stop codons. Duplicate protein sequences were amalgamated to produce ~80,000 different protein sequences (PIT proteins list) which was then used for the MS/MS analysis. We analysed this list to obtain data on size distribution (Supplementary Table 8) and used BLAST on this file to analyse its relationship to the human proteome (Supplementary Table 9).

Quantitative proteomics

Based on the RNA quantitation, the volume of the three protein samples (T0, T8 and T24) was adjusted to give equal amounts of protein between them. The protein samples were checked by western blotting for the presence of viral proteins (anti DBP) and equal amounts of cellular protein UBTF (see Western Blotting protocol). The three samples were then combined on a 1:1:1 ratio, separated by SDS-PAGE and analysed by LC-MS/MS. The gel lane was cut into 10 slices and each slice subjected to in-gel tryptic digestion using a ProGest automated digestion unit (Digilab UK). A second identical gel lane was run and a series of 4 new slices was taken from a region in the centre of the gel from between 30KDa and 70KDa making a total of 14 slices in all. The resulting peptides were fractionated using a Dionex Ultimate 3000 nanoHPLC system in line with an LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific). In brief, peptides in 1% (vol/vol) formic acid were injected onto an Acclaim PepMap C18 nano-trap column (Dionex). After washing with 0.5% (vol/vol) acetonitrile 0.1% (vol/vol) formic acid peptides were resolved on a 250 mm × 75 μm Acclaim PepMap C18 reverse phase analytical column (Dionex) over a 150 min organic gradient, using 7 gradient segments (1-6% solvent B over 1min., 6-15% B over 58min., 15-32%B over 58min., 32-40%B over 3min., 40-90%B over 1min., held at 90%B for 6min and then reduced to 1%B over 1min.) with a flow rate of 300 nl min^{-1}. Solvent A was 0.1% formic acid and Solvent B was aqueous 80% acetonitrile in 0.1% formic acid. Peptides were ionized by nano-electrospray ionization at 2.3 kV using a stainless steel emitter with an internal diameter of 30 μm (Thermo Scientific) and a capillary temperature of 250°C. Tandem mass spectra were acquired using an LTQ-Orbitrap Velos mass spectrometer controlled by Xcalibur 2.0 software (Thermo Scientific) and operated in data-dependent acquisition mode. The Orbitrap was set to analyze the survey scans at 60,000 resolution (at m/z 400) in the mass range m/z 300 to 2000 and the top six multiply charged ions in each duty cycle selected for MS/MS in the LTQ linear ion trap. Charge state filtering, where unassigned precursor ions were not selected for fragmentation, and dynamic exclusion (repeat count, 1; repeat duration, 30s; exclusion list size, 500) were used. Fragmentation conditions in the LTQ were as follows: normalized collision energy, 40%; activation q, 0.25; activation time 10ms; and minimum ion selection intensity, 500 counts.

The raw data files were processed and quantified using MaxQuant and searched against the databases detailed in the results section. Peptide precursor mass tolerance was set at 10ppm, and MS/MS tolerance was set at 0.8Da. Search criteria included carbamidomethylation of cysteine (+57.0214) as a fixed modification and oxidation of methionine (+15.9949) and appropriate SILAC labels ($^{13}C_6$-Lys, $^{13}C_6$-Arg for duplex and $^{13}C_6$ $^{15}N_2$-Lys and $^{13}C_6$ $^{15}N_4$-Arg for triplex) as variable modifications. Searches were performed with full tryptic digestion and a maximum of two missed cleavages was allowed. The reverse database search option was enabled and all peptide data was filtered to satisfy false discovery rate (FDR) of 1%.

The unprocessed spectra files (in .RAW format) for the human and CHO experiments can be accessed from Canfield University, IP address: 138.250.31.74(port 22), Login: anonymous, Password: anonymous.

Integration of proteomic and RNAseq data

A schematic workflow for our data analysis is given in supplementary figure 2. Briefly there are two aspects, the PIT analysis and the gene expression and protein abundance integration.

In the first step, the list of Trinity derived transcripts is initially mapped to the host cell genome (in this case human) using GMAP to generate a SAM file which reports the Trinity derived identifier of the transcript, the location of the sequence on the genome (or no

location if it is not on the human genome) and how the sequence maps to the target genome (i.e. the exon structure is described). However, GMAP loses the gene expression data at this stage so this information is added back to the SAM file using in house software (put_fpkm_values_back.pl) that also adds a new data field to the SAM file ready for later stages. Next the Trinity transcripts are translated by "getorf" in the EMBOSS package in the Galaxy suite to report all ORFs longer than 200 nt to generate the PIT proteins list. This PIT proteins list is used to search the MS/MS data for positive hits with MaxQuant. Bespoke in-house software (pep_to_sam.pl) is then used to modify the SAM file to add the MaxQuant identified peptides back to the Trinity identified transcripts by adding a series of new data fields (allowed within the SAM format) which contain information on the peptide concerned (e.g. ratio changes and a quality score for the match). In the second step a second in house software tool (sam_to_GFF3_and_orfs.pl) uses the modified SAM format file to generate the GFF3 file. This is done for each transcript which has an identified peptide associated with it irrespective of whether the transcript mapped to the target genome or not. This tool uses the intron exon structure of the transcript reported in the SAM file to determine which exon the first amino acid of the identified peptide is from. In addition, the reading frame and strand that contains the peptide is determined and the longest possible open reading frame (i.e. 5′ most in-frame start codon to the next stop codon after the identified peptide) for that individual peptide is recorded in a FASTA format file together with the name of the Trinity transcript the ORF is derived from. This list of the longest MS/MS identified ORFs is then searched and all identical proteins are amalgamated since, for many transcripts there is more than one identified peptide and each one will generate a separate FASTA entry. This list of longest ORFs is then used later for the BLAST analysis. The GFF3 file this process generates reports the following:

1. The precise location of the start of each identified peptide.
2. A solid region representing the size of the peptide that is colour coded depending on the ratio changes between 0 and 24 hours.
3. The confidence score reported by MaxQuant.
4. All the quantitation ratios derived by MaxQuant.

For the gene expression/protein quantitation integration, the two datasets are integrated using a combination of text file manipulation tools found within Galaxy and manual annotation within Excel. The Cufflinks gene estimation data and the MaxQuant proteomics data was integrated within Galaxy using the common ENSG identifiers present in the gene expression and proteomics data outputs. Thus, we are only able to combine our gene expression data with the protein expression data using the common identifiers provided by Ensembl.

BLAST analysis

The list of longest unique ORFs detected by MS/MS from the Trinity derived dataset was used to BLAST search against two separate databases. First we searched against the Ensembl list of human proteins in order to determine how many distinct human proteins were identified. Secondly the list was searched against the non-redundant protein databases to demonstrate that this approach will identify proteins from multiple species correctly. The BLAST searches were performed using in house software (batchBlastAndParse.pl) and the results manually collated within excel.

PIT analysis of the Chinese Hamster Ovary cells

In essence the analysis pipeline is the same as outlined in supplementary figure 2 using the publically available RNAseq data for CHO cells (European Nucleotide Archive

SRP001851). The main changes are that the Trinity derived transcripts were mapped using GMAP to the CHO genome (RefSeq Assembly ID: GCF_000223135.1). A sample of CHO cells (approximately 200,000 cells) were boiled in SDS-PAGE loading buffer and the proteins separated by SDS-PAGE. The sample was divided into twelve slices which were independently digested with trypsin in-gel and analysed by LC-MS/MS as described above except that the top twenty multiply charged ions in each duty cycle were selected for MS/MS in the LTQ linear ion trap.

The analysis of MS/MS spectra by MaxQuant is as described except that two scripts for connecting the transcriptomic and proteomic data were re-written to take account of the lack of quantitative information in the MaxQuant peptides list. These are pep_to_sam_no_quant.pl and sam_to_GFF3_and_orfs_noquant.pl and directly replace their equivalents used in the human PIT analysis. These form the pipeline that generates the sam and GFF files to allow a seamless view of transcripts and peptides on the CHO genome. Finally we modified our collation and BLAST analysis of the identified peptides using two new scripts. The first is called Connect_maxQ_peptides_to_trinity_fasta_files.pl which takes the Trinity transcripts list and the Trinity derived list of possible ORFS and the peptides.txt list from MaxQuant. This script generates a collated list comprising only those ORFS that have supporting peptides identified by MaxQuant along with the transcript from which it was generated. This collated list (called longest_ORFS_Collated.txt) is then searched sequentially with the script UNIXbatchBlastAndParse_CHO_analysis.pl. This script is designed to take the longest ORF in each line and find the best possible match in the specified protein database using BLAST. We generated specific databases for Chinese Hamster, Mouse and Human by downloading the complete proteome for each species from UNIPROT. We first used the Chinese Hamster database, then the mouse one and finally the human one. After each analysis the results are appended to the beginning of each line of data. The final output is shown in Supplementary Table 5 and allows a researcher to see for every ORF, the transcript it came from, the Trinity name (which can be used to find the location on the CHO genome of the transcript and peptides), the peptides found and the nearest match in the Chinese Hamster, mouse and human proteomes reported by BLAST (or indeed any other proteome).

Searching the data for SNPs and Indels

We used the SNPeff software for our analysis of SNPs in the human data. Initially, we used the identified ENSP amino acid sequence (obtained from BioMart) and derived a list of canonical proteins (supplementary dataset 12). We then corrected the amino acid sequences using our own software (implement_snp_eff_changes.pl) to generate a list of SNP corrected proteins. The two files (supplementary datasets 10 and 11) were combined and used as the search space for MaxQuant along with the PIT proteins list. There were 11,458 unique mutant proteins considered (approximately 10 million amino acids in total) alongside 7,868 uncorrected sequences (approximately 6 million amino acids). The data is then mined manually to find peptides that are only found in the corrected sequences and to find heterogeneous identifications (i.e. where both alleles are apparently expressed). In addition, by searching for SNPs alongside the PIT proteins list we are utilising a database of comparable size to the other searches reported in this manuscript improving the confidence that our identifications are not artefacts (i.e. resulting from a reduced complexity dataset). The outputs are collated manually within excel for ease of viewing.

Transfection, infection and immunofluorescence

HeLa cells were transfected with a plasmid expressing HA tagged *POLDIP3* (also known as SKAR and a generous gift of J. Blenis, Harvard Medical School) using lipofectamine 2000. At the same time the cells were infected at a multiplicity of infection of 1 with either wild

type adenovirus (serotype 5) or dl366 (a generous gift from K. Leppard). After 24 hours the cells were fixed with formaldehyde, permeablised with Triton X-100 and processed for immunofluorescence using either anti HA tag (anti HA serum F-7 from Santa Cruz catalogue number sc-7397) or anti DBP serum together with appropriate Alexa-fluor secondary antibodies (used at 1/200 dilution, alexa fluor 488 and alexa fluor 594).

Western blotting and antibodies

Antibodies used in western blots were anti DBP (used at 1/200 dilution), anti UBTF (anti UBTF serum H-300 from Santa Cruz catalogue number sc-9131 used at 1/100 dilution), anti GAPDH (anti GAPDH serum FL335 from Santa Cruz catalogue number sc-25778 used at 1/100 dilution), anti POLDIP3 (used at 1/50 dilution). In each case new samples of cells infected with adenovirus for 24 hours was obtained and the new samples tested for protein expression alongside the original samples processed for quantitative proteomics. In addition, we treated cells with either DMSO or DMSO containing 10ng/ml MG132 for 8 hours prior to harvesting infected or uninfected HeLa cells to determine the effect of proteasome inhibition on protein abundance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We'd like to thank P. Kellam and A. Palser at the Wellcome Trust Sanger Centre for help and advice throughout. We would also like to thank C. Trapnell (Broad Institute), J. Goeks (Emory University), J. Jackson (Penn State University), P. Cingonlani (McGill University), B. Haas (Broad Institute), T. Wu (Genentech) and J. Robinson (Broad Institute) for very informative and helpful discussions by email. We especially thank I. Goodfellow (Imperial College) for discussions on using proteomic data from BHK and CHO cells. We are grateful to R. T. Hay (University of Dundee) and J. Blenis (Harvard Medical School) for antibodies to DBP and POLDIP3 respectively. We also thank the University of Bristol Transcriptomics facility (especially J. Coghill) and the University of Bristol Wolfson Bioimaging facility for their help. DAM and VCE are funded by the Wellcome trust (Grant number 083604). CB and JF by the BBSRC (BBSRC grant BB/I00095X/1).

References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics. 2009; 10:57–63.
2. Grabherr MG, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011; 29:644–652.
3. Brewis IA, Brennan P. Proteomics technologies for the global identification and quantification of proteins. Adv Protein Chem Struct Biol. 2010; 80:1–44. [PubMed: 21109216]
4. Lamond AI, et al. Advancing cell biology through proteomics in space and time (PROSPECTS). Molecular & cellular proteomics : MCP. 2012; 11 O112 017731.
5. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics. 2010; 73:2092–2123. [PubMed: 20816881]
6. Lundberg E, et al. Defining the transcriptome and proteome in three functionally different human cell lines. Mol Syst Biol. 2010; 6:450. [PubMed: 21179022]
7. Li M, et al. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011; 333:53–58. [PubMed: 21596952]
8. Castellana N, Bafna V. Proteogenomics to discover the full coding content of genomes: a computational perspective. J Proteomics. 2010; 73:2124–2135. [PubMed: 20620248]
9. Volkening JD, et al. A proteogenomic survey of the Medicago truncatula genome. Molecular & cellular proteomics : MCP. 2012

10. O'Shea CC, et al. Late viral RNA export, rather than p53 inactivation, determines ONYX-015 tumor selectivity. Cancer Cell. 2004; 6:611–623. [PubMed: 15607965]
11. Orazio NI, Naeger CM, Karlseder J, Weitzman MD. The adenovirus E1b55K/E4orf6 complex induces degradation of the Bloom helicase during infection. Journal of virology. 2011; 85:1887–1892. [PubMed: 21123383]
12. Woo JL, Berk AJ. Adenovirus ubiquitin-protein ligase stimulates viral late mRNA nuclear export. J Virol. 2007; 81:575–587. [PubMed: 17079297]
13. Cox J, et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. Journal of proteome research. 2011; 10:1794–1805. [PubMed: 21254760]
14. Wu BJ, Hurst HC, Jones NC, Morimoto RI. The E1A 13S product of adenovirus 5 activates transcription of the cellular human HSP70 gene. Molecular and cellular biology. 1986; 6:2994–2999. [PubMed: 3491295]
15. Dallaire F, Blanchette P, Branton PE. A proteomic approach to identify candidate substrates of human adenovirus E4orf6-E1B55K and other viral cullin-based E3 ubiquitin ligases. Journal of virology. 2009; 83:12172–12184. [PubMed: 19759146]
16. Evans JD, Hearing P. Relocalization of the Mre11-Rad50-Nbs1 complex by the adenovirus E4 ORF3 protein is required for viral replication. J Virol. 2005; 79:6207–6215. [PubMed: 15858005]
17. Lam YW, Evans VC, Heesom KJ, Lamond AI, Matthews DA. Proteomics analysis of the nucleolus in adenovirus-infected cells. Mol Cell Proteomics. 2010; 9:117–130. [PubMed: 19812395]
18. Blankenberg D, et al. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. 2010:11–21. Chapter 19, Unit 19 10. [PubMed: 20069539]
19. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
20. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology. 2010; 28:511–515.
21. Swift FV, Bhat K, Younghusband HB, Hamada H. Characterization of a cell type-specific enhancer found in the human papilloma virus type 18 genome. The EMBO journal. 1987; 6:1339–1344. [PubMed: 3038518]
22. Zhao H, Granberg F, Elfineh L, Pettersson U, Svensson C. Strategic attack on host cell gene expression during adenovirus infection. J Virol. 2003; 77:11006–11015. [PubMed: 14512549]
23. Zhao H, Dahlo M, Isaksson A, Syvanen AC, Pettersson U. The transcriptome of the adenovirus infected cell. Virology. 2012; 424:115–128. [PubMed: 22236370]
24. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005; 21:1859–1875. [PubMed: 15728110]
25. Soloway PD, Shenk T. The adenovirus type 5 i-leader open reading frame functions in cis to reduce the half-life of L1 mRNAs. Journal of virology. 1990; 64:551–558. [PubMed: 2296076]
26. Symington JS, et al. Biosynthesis of adenovirus type 2 i-leader protein. Journal of virology. 1986; 57:848–856. [PubMed: 3005631]
27. van den Hengel SK, et al. Truncating the i-leader open reading frame enhances release of human adenovirus type 5 in glioma cells. Virology journal. 2011; 8:162. [PubMed: 21477385]
28. Xu X, et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nature biotechnology. 2011; 29:735–741.
29. Boisvert FM, et al. A quantitative spatial proteomics analysis of proteome turnover in human cells. Molecular & cellular proteomics : MCP. 2011
30. Ma XM, Yoon SO, Richardson CJ, Julich K, Blenis J. SKAR links pre34 mRNA splicing to mTOR/S6K1-mediated enhanced translation efficiency of spliced mRNAs. Cell. 2008; 133:303–313. [PubMed: 18423201]
31. Forrester NA, et al. Serotype-specific inactivation of the cellular DNA damage response during adenovirus infection. Journal of virology. 2011; 85:2201–2211. [PubMed: 21159879]

32. Halbert DN, Cutt JR, Shenk T. Adenovirus early region 4 encodes functions required for efficient DNA replication, late gene expression, and host cell shutoff. Journal of virology. 1985; 56:250–257. [PubMed: 4032537]

33. Birzele F, et al. Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. Nucleic acids research. 2010; 38:3999–4010. [PubMed: 20194116]

Europe PMC Funders Author Manuscripts

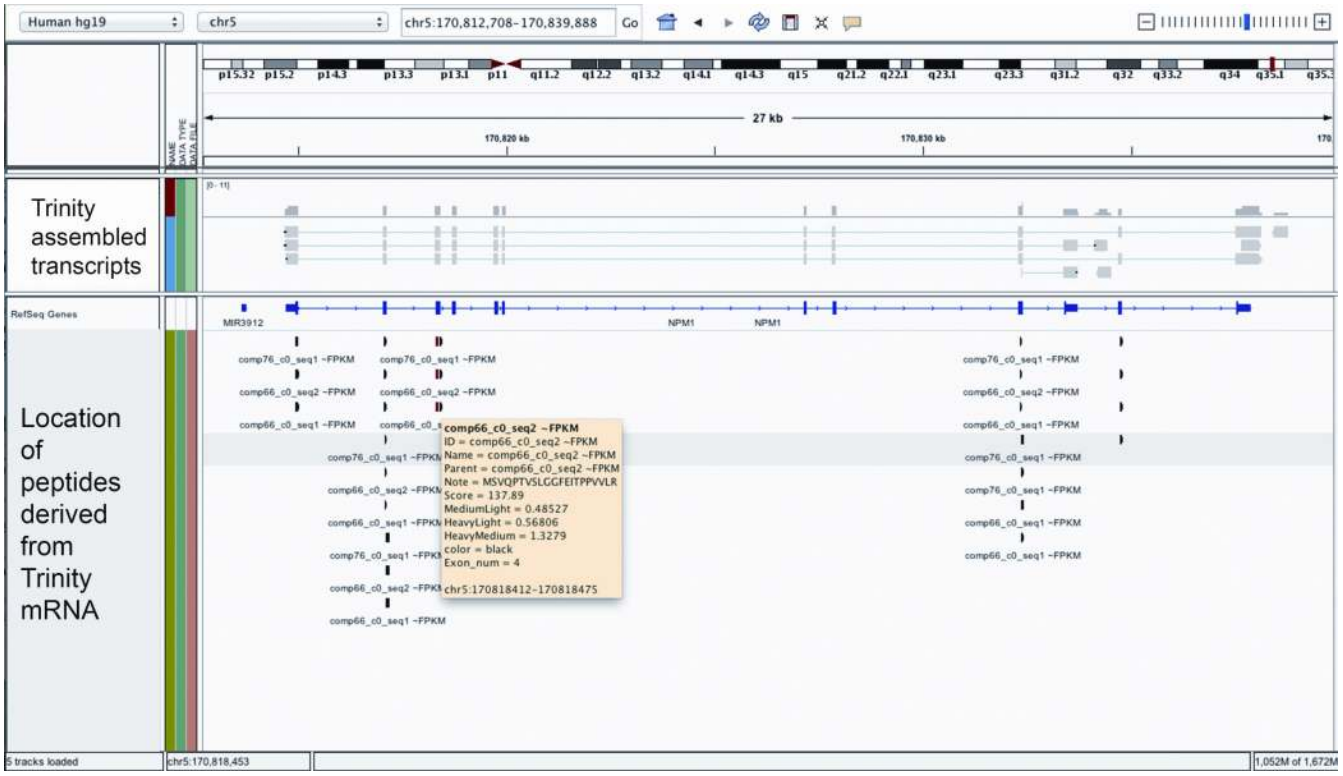

Figure 1. Illustration of data integration between the transcriptome and the proteome

Image taken from the IGV viewer showing a SAM alignment file generated by GMAP using Trinity derived sequences. In addition we show the data from the custom GFF3 file that allows us to see what peptides were identified by MS/MS, their location on the transcript and genome. For each peptide identified the yellow box (arrowed) appears once the mouse pointer is over the peptide and in each case lists the peptide sequence, the confidence score, and the ratios at different time points. In the middle of the screenshot there are the refseq annotated isoforms of NPM1. Note that the same peptide is flagged multiple times as it belongs to one of several Trinity assembled transcripts.

Europe PMC Funders Author Manuscripts

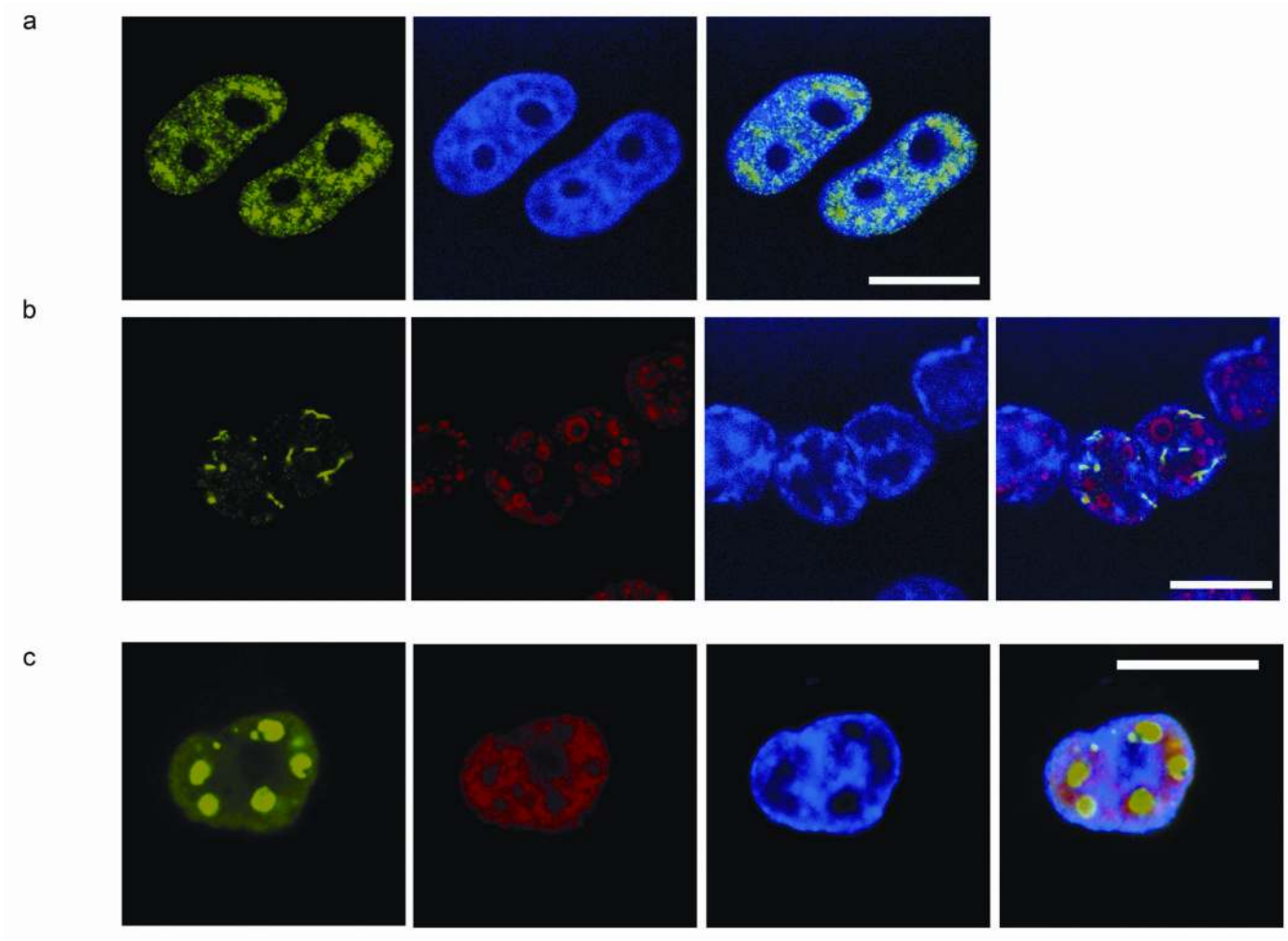

Figure 2. Adenovirus induces degradation of POLDIP3 in a manner sensitive to MG132 and redistribution of POLDIP3 in infected cells

a) Samples of adenovirus infected or uninfected HeLa cells were tested by western blot for expression of POLDIP3. These samples are biological repeats in the presence of either DMSO or MG132 in DMSO. Equivalence of loading is shown by the GAPDH control. The top row of panels (part b) shows the normal distribution of HA tagged POLDIP3 (green) in uninfected HeLa cells. The middle row (c) shows the distribution of HA-POLDIP3 (green) in wild type adenovirus infected cells. The adenovirus DNA binding protein, DBP (in red), is clearly visible in the nuclei of cells. The final row of cells (d) shows the distribution of HA-POLDIP3 (red) in cells infected with adenovirus mutant dl306 which lacks the E4 region of the virus but still expresses DBP (in green). In all cases the infected cells were fixed at 24 hours post infection, the white bar represents 10um and the cell nuclei are stained with DAPI in blue.

Europe PMC Funders Author Manuscripts

Table 1

Reads generated and mapped to the human, adenovirus and papilloma virus genomes.

Total number of paired end reads at each time point is listed along with how many of those reads mapped to a unique site in either a female human genome (hg19 less chromosome Y), the adenovirus type 5 genome or papilloma virus type 18 genome – part of which is integrated into the HeLa cell genome. In all cases we only consider reads where both ends in a pair map to the target genome in the correct orientation and to opposite strands as expected for a correctly mapped pair of sequence reads.

	Uninfected HeLa cells	8 hours post infection	24 hours post infection
Total reads generated	29,552,473	26,220,901	26,251,561
Reads uniquely mapped in a proper pair to female hg19	18,097,929	16,325,343	3,183,200
Reads uniquely mapped in a proper pair to adenovirus type 5	187	521,731	15,134,568
Reads uniquely mapped in a proper pair to papilloma virus type 18	45,088	18,755	634

Table 2

Identification of peptides and proteins using different protein datasets.

Five different lists of proteins were used as the reference list to search the MS/MS spectra using MaxQuant. In all cases the search list included a standard list of known contaminants and a list of reversed proteins to act as a decoy that allowed the false discovery rate to be set at 1%. For the canonical protein lists (Ensembl or Swissprot) we added a list of human adenovirus proteins as well so that we can compare the Trinity list (which will contain adenovirus sequences) on a like for like basis. The adenovirus proteins were derived from the GenBank entry for adenovirus type 5 (AC_000008.1). In each case, the percentage quoted refers to the number of peptides present in both lists as a proportion of the total number of peptides detected in the canonical ENSGs list.

	Canonical ENSGs	ENSGs detected at T0	ENSTs detected at T0	Trinity derived ORFS	SwissProt-Uniprot
Total number of distinct peptides detected	29,371	28,862	28,862	28,827	29,512
As a percentage of detected canonical ENSGs	100%	98.2%	98.2%	95.6%	99.6%
Distinct protein groups reported with at least two peptides detected	3,415	3,373	3,373	3,595	3,443
Peptides detected not in canonical list	0	454	454	754	257
Number of distinct proteins in database	21,173	14,537	29,287	80,648	72,049
Total number of amino acids in dataset	11,633,994	8,828,371	15,690,432	11,305,091	32,897,704
Total number of amino acids found	420,069	418,430	418,430	414,616	421,031