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] & David A Matthews[1]

Identification of proteins by tandem mass spectrometry requires a reference protein database, but these are only available for model species. Here we demonstrate that, for a non-model species, the sequencing of expressed mRNA can generate a protein database for mass spectrometry–based identification. This combination of high-throughput sequencing and protein identification technologies allows detection of genes and proteins. We use human cells infected with human adenovirus as a complex and dynamic model to demonstrate the robustness of this approach. Our proteomics informed by transcriptomics (PIT) technique identifies >99% of over 3,700 distinct proteins identified using traditional analysis that relies on comprehensive human and adenovirus protein lists. We show that this approach can also be used to highlight genes and proteins undergoing dynamic changes in post-transcriptional protein stability.

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 for those of non-model species[1,2]. In parallel, improvements in high-throughput liquid chromatography–coupled tandem mass spectrometry (LC-MS/MS) have allowed for the identification of several thousand distinct proteins from a total-cell extract in a single experiment[3,4]. When combined with quantitative techniques such as stable-isotope labeling of amino acids in cell culture (SILAC), LC-MS/MS can be used to monitor changes in relative protein levels over time. However, MS/MS data analysis software normally requires an accurate list of proteins that could be present in the sample[5]. Obtaining this list is reasonably straightforward for model organisms (for example, humans) but not for poorly annotated species, for species whose genomes are not yet fully assembled or for samples containing proteins from multiple species. Moreover, this approach is not optimized for individual variation caused by single-nucleotide polymorphisms (SNPs).

A small number of recent publications combine deep sequencing with LC-MS/MS[6,7]. One study characterized different human cell lines using deep sequencing–based transcriptomics (RNA-seq) 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]. The use of 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, and confidently identifying novel proteins (that is, new proteins with little or no homology to previously existing proteins) in non-model species is particularly challenging with this approach. To alleviate this problem, 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. Although 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 analyses attempted to date is on *Medicago truncatula*[9], which has a genome ~0.6 gigabase pairs (Gbp) in size, substantially smaller than the human genome (~3 Gbp).

Our PIT analysis constitutes a high-throughput method to derive transcripts, infer proteins from them and use this information to identify peptides detected by MS/MS, enabling a seamless data visualization of genome, transcriptome and proteome. To benchmark PIT, we used a highly annotated two-genome system of human cells infected with adenovirus.

We recovered the vast majority of obtainable information from both the virus and human samples in a manner that is independent of pre-existing data sets and, in principle, independent of a copy of the target genome. We also showed that our approach is robust: it is able to cope 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.

[1]School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK. [2]School of Biological Sciences, University of Bristol, Bristol, UK. [3]School of Biochemistry, University of Bristol, Bristol, UK. [4]Bioinformatics Group, Cranfield Health, Cranfield University, Cranfield, Bedfordshire, UK. Correspondence should be addressed to D.A.M. (d.a.matthews@bristol.ac.uk).

RECEIVED 16 MAY; ACCEPTED 21 SEPTEMBER; PUBLISHED ONLINE 11 NOVEMBER 2012; DOI:10.1038/NMETH.2227

RESULTS

Sample collection

We metabolically labeled HeLa cells by SILAC, either with ^{15}N- and ^{13}C-labeled arginine and lysine (heavy HeLa), with ^{13}C-labeled arginine and lysine (medium HeLa) or 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 h post-infection, we harvested the light HeLa cells for protein and RNA. At 24 h post-infection, the medium and heavy cells were similarly harvested. This enabled protein quantitation over time.

Table 1 | Reads generated and mapped to the human, adenovirus and papilloma virus genomes

	Uninfected HeLa cells	Infected cells, 8 h post-infection	Infected cells, 24 h 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 HPV-18	45,088	18,755	634

The 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 HPV-18 genome—part of which is integrated into the HeLa cell genome. In all cases, we considered only reads in which both ends in a pair mapped to the target genome in the correct orientation and to opposite strands as expected for a correctly mapped pair of sequence reads.

SILAC-based quantitative proteomics

We combined the three protein samples in a 1:1:1 ratio before separating proteins by SDS-PAGE and processing them for LC-MS/MS analysis. We analyzed the MS/MS spectra with MaxQuant software[10] to identify proteins and quantitate abundance changes. HeLa cells are cervical-carcinoma derived, containing genes from human papilloma virus type 18 (HPV-18) integrated into the cell genome. We searched for HPV-18 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 either a twofold-or-greater increase or a 50% decrease in abundance over the first 8 h of the experiment, and just under 8% had altered their abundance by either a twofold increase or a 50% decrease at 24 h post-infection. We were able to detect a threefold increase in expression of *HSPA1A*, which encodes a protein (HSP70) known to be induced early on by adenovirus infection[11], in the first 8 h post-infection. We were also able to detect a greater than 50% decrease in abundance of MRE11, ITGA3 and RAD50, all known to be degraded during infection[12,13]. We previously reported that levels of upstream binding transcription factor (UBTF) remain unchanged during adenovirus infection, which was also reflected in this data set[14].

RNA-seq analysis of adenovirus-infected cells

We harvested cytoplasmic mRNA from the same three samples of HeLa cells because adenovirus inhibits nuclear export of cellular mRNA during infection without inhibiting its production. Viral mRNA export is thus heavily favored. Each sample was sequenced on an Illumina GAIIx for a total of ~82 million paired-end reads 56 bp in length.

We imported our data into a locally installed Galaxy NGS software suite[15] and mapped our data using TopHat[16] to the human genome (hg19). We used uniquely mapped reads for gene expression analysis with Cufflinks[17], using the Ensembl human gene annotation as a guide (v.64). Separately, we mapped the reads to human adenovirus type 5 (GI: 56160529) and HPV-18 (GI: 30172004) genomes. Over the course of the experiment, the number of reads mapped to the adenovirus genome increased to about 80% of the total reads (**Table 1**), illustrating how the virus transcriptome eventually dominates that of the host. We detected HPV-18 transcripts at all time points and observed that over the course of the experiment there was a decline in HPV-18 transcription (data not shown) consistent with previous reports[18]. At 8 h post-infection, most adenovirus-derived reads mapped to adenovirus early genes (E1, E2, E3 and E4), whereas by 24 h, most reads mapped to the late genes derived from the virus major late promoter (**Supplementary Fig. 1**). Regarding human gene expression, we saw results similar to those described using microarray experiments[19]; *CDC25A* was upregulated soon after adenovirus infection, and *HSPA1A* expression increased. Our results were also similar to those of a recent RNA-seq–based transcriptomic analysis of human cells infected with adenovirus[20] even though a different cell type and total mRNA extracts were used rather than cytoplasmic mRNA: for example, in the IκB family of NF-κB inhibitors, expression of *NFKBIE* declined and expression of *NFKBIB* increased in both data sets.

Proteomics informed by transcriptomics

We utilized Trinity[2] and a combined set of sequence reads from all three time points for the *de novo* assembly of the entire (host and virus) transcriptome (**Supplementary Data 1**). We then generated open reading frames (ORFs) (>200 nucleotides) from all six frames of each Trinity-generated transcript (**Supplementary Data 2**). This 'PIT proteins' list was used as our search database for the MaxQuant package. A comparison of the peptides generated by searching standard human proteomes revealed that a search using the PIT protein list generated almost as many identified peptides (~95%) as a search from a canonical list of human proteins from Ensembl or from a nonredundant SwissProt-UniProt list, two standard human-protein lists widely used for this type of analysis (**Table 2** and **Supplementary Table 2**). In addition, in the list of peptides identified by searching the PIT proteins data set, we found 360 peptides that belonged to the adenovirus proteome as compared with 367 peptides found by searching a standard adenovirus proteome derived from GenBank.

Next we mapped the Trinity transcripts to the human genome using GMAP[21] to generate a sequence alignment map (SAM) file (**Supplementary Data 3**) and added the identified peptide data from MaxQuant to the Trinity transcripts with in-house software (**Supplementary Software**). Finally, we used in-house software to create a gene feature format (GFF3) file (**Supplementary Data 4**) combining the data in the SAM file with exon structure information. These SAM and GFF3 files allowed us to see which peptides were associated with a transcript, which exon each was derived from and the transcript's location on the human genome (**Fig. 1**).

Our software also generated a list containing the longest ORF associated with each peptide positively identified by MS/MS (**Supplementary Data 5**). Thus, our approach starts with a list of possible proteins derived from the Trinity-assembled transcripts

Table 2 | Identification of peptides and proteins using different protein data sets

	Canonical ENSGs	ENSGs detected at T_0	ENSTs detected at T_0	Trinity-derived ORFs	SwissProt-UniProt
Total number of distinct peptides detected	29,371	28,862	28,862	28,827	29,512
Detected peptides 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 data set	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

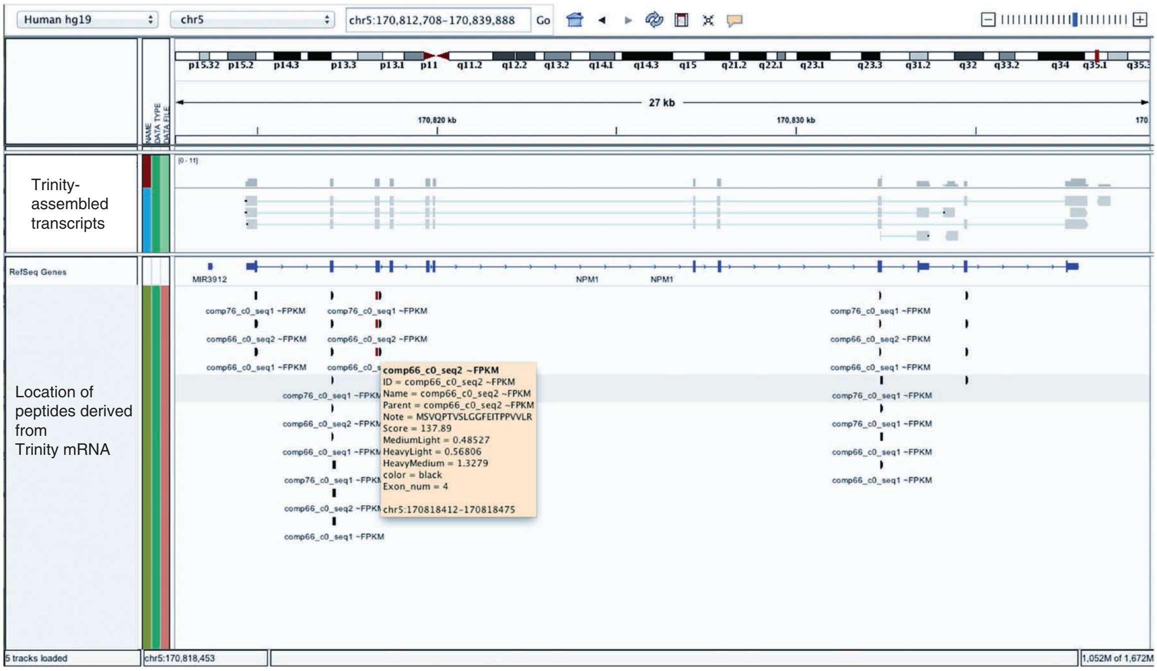
Five different lists of proteins were used as the reference list to search the MS/MS spectra using MaxQuant (ENSG, Ensembl identifier for the gene; ENST, Ensembl identifier for the transcript; T_0, time point 0.). In all cases, the search list included a standard list of known contaminants and a list 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 could compare the Trinity list (which contains 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.

and ends up with a list of full-length proteins derived from the Trinity transcripts for which there is at least one peptide identified (**Supplementary Fig. 2**). We derived a list of 7,319 unique proteins in this way, although it is worth noting that Trinity may report several isoforms for each transcript. For example, Trinity may assemble non-identical transcripts for the same gene, some longer than others; and when translated into proteins, such a list will therefore contain several isoforms of the same protein. When we searched all of this list of positively identified proteins using BLAST, we determined that all were either human or adenovirus-derived proteins (**Supplementary Table 3**). We found 3,792 distinct human genes using the PIT proteins data set, versus 3,773 distinct human found genes using the Ensembl data set, with 99.45% overlap between the two lists. In the viral data, the PIT analysis missed one adenovirus protein (U-exon protein), but it identified proteins not recognized by a traditional search of the adenovirus proteome. As an example, the 'i leader' adenovirus protein is a bona fide adenovirus protein[22–24], but it 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.

We repeated this PIT analysis with declining quantities of the raw transcriptomic data generated here (**Supplementary Table 4**) to examine how depth of transcriptome coverage influenced data return. Even a reduced data set of less than 10% of the coverage used here still yielded over 70% of the peptides identified by our largest analysis. The importance of this is illustrated by cases in which 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. Thus, PIT is limited by the quality and quantity of the RNA-seq data used in the Trinity assembly.

PIT analysis of Chinese hamster ovary cells

To illustrate the potential of the gene and protein identification aspect of PIT for non-model organisms, we examined CHO cells.



comp66_c0_seq2 ~FPKM
ID = comp66_c0_seq2 ~FPKM
Name = comp66_c0_seq2 ~FPKM
Parent = comp66_c0_seq2 ~FPKM
Note = MSVQPTVSLGGFEITPPVVLR
Score = 137.89
MediumLight = 0.48527
HeavyLight = 0.56806
HeavyMedium = 1.3279
color = black
Exon_num = 4

chr5:170818412-170818475

Figure 1 | Data integration between the transcriptome and the proteome. Screenshot of the IGV viewer showing a SAM alignment file generated by GMAP using Trinity-derived sequences. Peptides identified by MS/MS and their location on the transcript and genome are shown in last row. For each peptide identified, the beige box appears when the mouse pointer hovers over the peptide, and in each case, it lists the peptide sequence, confidence score and ratios at different time points.

We obtained a publicly available RNA-seq data set (European Nucleotide Archive no. SRP001851) and assembled the transcriptome using Trinity as before (**Supplementary Data 6**). From this we generated a list of proteins (**Supplementary Data 7**) and used this list to search spectra from a total-protein extract of CHO cells separated by one-dimensional gel electrophoresis before LC-MS/MS. Our PIT analysis showed that a search of the Trinity CHO proteins list, compared to a search of the standard UniProt list of CHO proteins, led to approximately 70% more identified peptides (**Supplementary Table 5**). Moreover, we were able to infer a list of the largest ORFs associated with each identified peptide and to search these identified ORFs, using BLAST to define the nearest homologs in the CHO, mouse and human UniProt lists (**Supplementary Table 6**). This list has 7,333 non-identical transcripts or proteins, and many are likely to be minor variants of the same protein (indeed, BLAST searching indicates that 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[25] alongside the locations of identified peptides, thus providing a seamless view of genome, transcript and identified peptides (**Supplementary Fig. 3** and **Supplementary Data 8** and **9**).

Detecting single-nucleotide polymorphisms in the proteome

We analyzed our TopHat alignments using snpEFF[26] to generate a list of nonsynonymous SNPs from our RNA-seq data. Using this and in-house software, we derived a list of canonical and variant proteins (**Supplementary Data 10–12**) to search our MS/MS spectra and to compare to our PIT-protein and canonical lists. We were able to correlate 170 SNP changes with detected peptides: 14 peptides for which only the canonical sequence (and not the SNP variant) was detected and 14 heterologous transcripts for which both a canonical and variant peptide were detected by MS/MS analysis (**Supplementary Table 7**). The majority of SNPs (149 of 170) 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 by over 50% without a corresponding decline in mRNA abundance. Three proteins known to be degraded during adenovirus infection (MRE11, RAD50 and ITGA3)[27–29] were identified as meeting this criterion. From this we developed a short list of proteins whose abundance had fallen (by over 50% in 24 h) without decreases in transcript levels (mRNA levels at 24 h at least 0.8 of that in uninfected cells). In addition, we confirmed the half-lives of these proteins to be above 24 h (ref. 30). Of these, we selected POLDIP3 (widely known as SKAR) for further research because it is proposed to play a role in cellular mRNA export and translation[31], which are both known to be affected by adenovirus in a manner dependent on the induction of a novel ubiquitin ligase complex. Indeed, POLDIP3 was degraded in adenovirus-infected cells, and the degradation was sensitive to the proteasome inhibitor MG132 (data not shown). In addition, during adenovirus infection, POLDIP3 was sequestered from a speckled distribution in uninfected cells into track-like structures (**Fig. 2a**,**b**) similar to that reported for MRE11, a known target of adenovirus-induced ubiquitin-mediated degradation[32]. Moreover, cells infected with adenovirus mutant dl366 (ref. 33), which lacks the E4 region required for the formation of a virally induced ubiquitin ligase complex, did not show any reorganization of POLDIP3 (**Fig. 2c**).

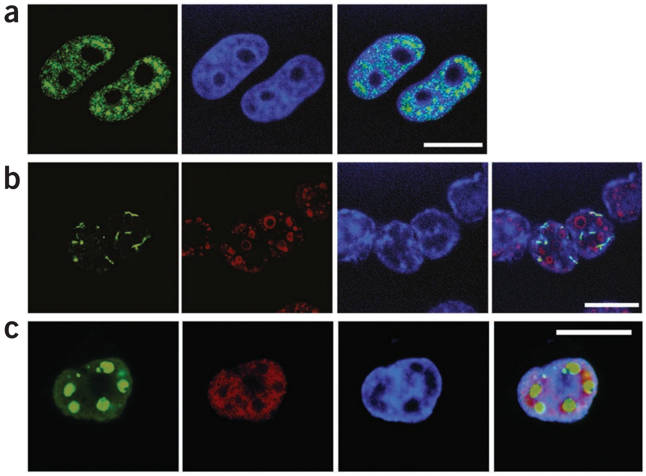


Figure 2 | Adenovirus-induced degradation of POLDIP3 in an MG132-sensitive manner, and redistribution of POLDIP3 in infected cells. (**a**) Normal distribution of hemagglutinin (HA)-tagged POLDIP3 (green) in uninfected HeLa cells. (**b**) Distribution of HA-POLDIP3 (green) in wild-type adenovirus-infected cells. The adenovirus DNA-binding protein, DBP (in red), is clearly visible in the nuclei of cells. (**c**) 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 h post-infection. Cell nuclei are stained with DAPI in blue. Scale bars, 10 µm.

DISCUSSION

PIT analysis has broad utility in the study of a wide range of species for which 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 on post-transcriptional events, as we have shown by examining proteins that degrade over time without a corresponding decline in mRNA expression.

Our PIT analysis of CHO cells illustrates that this approach does work in non-model systems and that PIT can be used with published RNA-seq data sets.

A key limitation is the depth of sequencing, but with higher read depth in transcriptomic experiments, the proportion of proteins identified will improve. PIT analysis may also help refine algorithms for *de novo* transcriptomic assembly: 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 advances in MS/MS-based proteomics, implying that it will be increasingly attractive to base MS/MS searches on proteins derived from the transcriptome rather than on canonical lists. We propose that such data be added to the well-established SAM and

GFF3 file formats as the most flexible way forward to integrate these data sets because 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 previously known as well as new host-cell targets of virally mediated degradation. Our focus on POLDIP3, which is involved in mRNA export and translation[31], is particularly relevant. The oncolytic phenotype of the adenovirus-derived ONYX-015 virus is linked to virus's failure to form a unique ubiquitin ligase complex that modifies the host cell's mRNA export pathway[29]. Our data implies that POLDIP3 is linked to the ONYX-015 phenotype, a hypothesis we are currently investigating.

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, alongside current approaches such as proteogenomics, will improve gene and protein identification in non-model species as well as refine the application of high-throughput technologies to the study of dynamic and/or multigenome systems. Finally, this technique should aid the development of systems approaches to biological research.

METHODS

Methods and any associated references are available in the online version of the paper.

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

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank P. Kellam and A. Palser for help and advice throughout and C. Trapnell, J. Goeks, J. Jackson, P. Cingonlani, B. Haas, T. Wu and J. Robinson for informative and helpful discussions by email. We especially thank I. Goodfellow for discussions on using proteomic data from baby hamster kidney 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. In addition, we thank J. Blenis for the HA-tagged *POLDIP3* expression plasmid and K. Leppard (University of Warwick) for the dl366 adenovirus. We also thank the members of the University of Bristol Transcriptomics facility (especially J. Coghill) and University of Bristol Wolfson Bioimaging facility for their help. D.A.M. and V.C.E. receive funding from the Wellcome trust (grant no. 083604). C.B. and J.F. receive funding from the Biotechnology and Biological Sciences Research Council (grant BB/I00095X/1).

AUTHOR CONTRIBUTIONS

V.C.E. cowrote the manuscript, prepared infected cells, performed western blots and assisted with immunofluorescence. G.B. cowrote the manuscript, wrote software and assisted with handling the RNA-seq data. K.J.H. performed the mass spectrometry and assisted with analysis of the MS/MS data. J.F. helped with the BLAST analysis and wrote some of the BLAST search software. C.B. cowrote the manuscript and assisted with the MS/MS analysis and BLAST database searches. D.A.M. conceived of the experiments and PIT analysis pipeline, led the manuscript writing, wrote software, assisted with the immunofluorescence and the preparation of infected cells, and carried out manual curation, analysis and integration of the data.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2227.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
2. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
3. Brewis, I.A. & Brennan, P. Proteomics technologies for the global identification and quantification of proteins. *Adv. Protein Chem. Struct. Biol.* **80**, 1–44 (2010).
4. Lamond, A.I. *et al.* Advancing cell biology through proteomics in space and time (PROSPECTS). *Mol. Cell. Proteomics* **11**, O112.017731 (2012).
5. Nesvizhskii, A.I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **73**, 2092–2123 (2010).
6. Lundberg, E. *et al.* Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* **6**, 450 (2010).
7. Li, M. *et al.* Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333**, 53–58 (2011).
8. Castellana, N. & Bafna, V. Proteogenomics to discover the full coding content of genomes: a computational perspective. *J. Proteomics* **73**, 2124–2135 (2010).
9. Volkening, J.D. *et al.* A proteogenomic survey of the *Medicago truncatula* genome. *Mol. Cell. Proteomics* **11**, 933–944 (2012).
10. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
11. Wu, B.J., Hurst, H.C., Jones, N.C. & Morimoto, R.I. The E1A 13S product of adenovirus 5 activates transcription of the cellular human HSP70 gene. *Mol. Cell. Biol.* **6**, 2994–2999 (1986).
12. Dallaire, F., Blanchette, P. & Branton, P.E. A proteomic approach to identify candidate substrates of human adenovirus E4orf6-E1B55K and other viral cullin-based E3 ubiquitin ligases. *J. Virol.* **83**, 12172–12184 (2009).
13. Evans, J.D. & Hearing, P. Relocalization of the Mre11-Rad50-Nbs1 complex by the adenovirus E4 ORF3 protein is required for viral replication. *J. Virol.* **79**, 6207–6215 (2005).
14. Lam, Y.W., Evans, V.C., Heesom, K.J., Lamond, A.I. & Matthews, D.A. Proteomics analysis of the nucleolus in adenovirus-infected cells. *Mol. Cell. Proteomics* **9**, 117–130 (2010).
15. Blankenberg, D. *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* **89**, 19.10 (2010).
16. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
17. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
18. Swift, F.V., Bhat, K., Younghusband, H.B. & Hamada, H. Characterization of a cell type-specific enhancer found in the human papilloma virus type 18 genome. *EMBO J.* **6**, 1339–1344 (1987).
19. Zhao, H., Granberg, F., Elfineh, L., Pettersson, U. & Svensson, C. Strategic attack on host cell gene expression during adenovirus infection. *J. Virol.* **77**, 11006–11015 (2003).
20. Zhao, H., Dahlo, M., Isaksson, A., Syvanen, A.C. & Pettersson, U. The transcriptome of the adenovirus infected cell. *Virology* **424**, 115–128 (2012).
21. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
22. Soloway, P.D. & Shenk, T. The adenovirus type 5 i-leader open reading frame functions *in cis* to reduce the half-life of L1 mRNAs. *J. Virol.* **64**, 551–558 (1990).
23. Symington, J.S. *et al.* Biosynthesis of adenovirus type 2 i-leader protein. *J. Virol.* **57**, 848–856 (1986).
24. van den Hengel, S.K. *et al.* Truncating the i-leader open reading frame enhances release of human adenovirus type 5 in glioma cells. *Virol. J.* **8**, 162 (2011).
25. Xu, X. *et al.* The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* **29**, 735–741 (2011).
26. Cingolani, P *et al.* snpEff SNP effect predictor v.3.0 <http://snpeff.sourceforge.net/> (2012).
27. O'Shea, C.C. *et al.* Late viral RNA export, rather than p53 inactivation, determines ONYX-015 tumor selectivity. *Cancer Cell* **6**, 611–623 (2004).
28. Orazio, N.I., Naeger, C.M., Karlseder, J. & Weitzman, M.D. The adenovirus E1b55K/E4orf6 complex induces degradation of the Bloom helicase during infection. *J. Virol.* **85**, 1887–1892 (2011).
29. Woo, J.L. & Berk, A.J. Adenovirus ubiquitin-protein ligase stimulates viral late mRNA nuclear export. *J. Virol.* **81**, 575–587 (2007).
30. Boisvert, F.M. *et al.* A quantitative spatial proteomics analysis of proteome turnover in human cells. *Mol. Cell. Proteomics* **11**, M111.011429 (2012).
31. Ma, X.M., Yoon, S.O., Richardson, C.J., Julich, K. & Blenis, J. SKAR links pre-mRNA splicing to mTOR/S6K1-mediated enhanced translation efficiency of spliced mRNAs. *Cell* **133**, 303–313 (2008).
32. Forrester, N.A. *et al.* Serotype-specific inactivation of the cellular DNA damage response during adenovirus infection. *J. Virol.* **85**, 2201–2211 (2011).
33. Halbert, D.N., Cutt, J.R. & Shenk, T. Adenovirus early region 4 encodes functions required for efficient DNA replication, late gene expression, and host cell shutoff. *J. Virol.* **56**, 250–257 (1985).

ONLINE METHODS

Cell culture, sample harvesting and viruses. HeLa cells were obtained from ECACC and grown in SILAC-labeled DMEM with 10% (v/v) SILAC-dialyzed fetal calf serum (Dundee Cell Products) for at least five 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 h of exposure to the virus, the medium was replaced with fresh SILAC-labeled medium, and the infection was allowed to continue for either 8 h or 24 h.

The cells were washed twice with PBS then treated with trypsin to release the adherent cells. Cells were washed a further two times in PBS before the sample was split 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 cytoplasmic RNA. Briefly, the cells were resuspended in 0.5 ml 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 RNA-seq, the three samples were used as substrates for PCR-based testing to confirm the presence of virus transcripts (adenovirus DBP gene) present in both the virus-infected samples and their absence in the uninfected samples (primer list in **Supplementary Table 8**). The three samples were labeled UN (uninfected control), T8 (8 h post-infection) and T24 (24 h post-infection). Next, the Trizol-extracted RNA was extracted again using RNeasy (Qiagen) before quantitation and processing for poly(A)$^+$ selection and 56-bp paired-end sequencing on the University of Bristol Illumina GAIIx using the manufacturer's reagents and protocols. The sequencing data were 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 paired-end sequence data for each time point were initially mapped to a female hg19 (i.e., less the Y chromosome) using TopHat. The following parameters were set:

Mean inner distance = 80; s.d. = 15; maximum mismatches in anchor region = 0; minimum intron length = 70; maximum intron length = 500,000; 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: = 500,000; 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 = 5,000; minimum intron length found by closure search = 50; coverage search = yes; minimum intron by coverage search = 50; maximum intron by coverage search = 20,000.

Mapped reads were then filtered to retain only those reads that mapped in a proper pair, and then reads that mapped to one location were separated from those that mapped to more than one location. Gene expression quantitation on uniquely mapping reads was performed using Cufflinks supplied with the Ensembl gtf (v.64) as a reference throughout the analysis. The following parameters were set for Cufflinks:

Maximum intron length = 500,000; minimum isoform fraction = 0.05; premRNA fraction = 0.05; quartile normalization = 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 papilloma virus serotype 18 (NC_001357.1) with the same parameters listed above but with the following changes:

Minimum intron length = 30; maximum intron length = 34,000 (7,000 for papilloma virus); minimum intron length that may be found during split-segment search = 10; maximum intron length that may be found during split-segment search: = 34,000 (7,000 for papilloma virus).

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 analyzed this list to obtain data on size distribution (**Supplementary Table 9**) and used BLAST on this file to analyze its relationship to the human proteome (**Supplementary Table 10**).

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 among 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 below). The three samples were then combined in a 1:1:1 ratio, separated by SDS-PAGE and analyzed by LC-MS/MS. The gel lane was cut into ten slices, and each slice was 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 four new slices was taken from a region in the center of the gel from between 30 kDa and 70 kDa, for a total of 14 slices. 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% (v/v) formic acid were injected onto an Acclaim PepMap C18 nano-trap column (Dionex). After a washing step with 0.5% (vol/vol) acetonitrile and 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 seven gradient segments (1%–6% solvent B over 1 min, 6–15% B over 58 min, 15–32% B over 58 min, 32–40% B over 3 min, 40–90% B over 1 min, held at 90% B for 6 min and then reduced to 1% B over 1 min) 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

doi:10.1038/nmeth.2227

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* 300–2,000 and with 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, 30 s; exclusion list size, 500) were used. Fragmentation conditions in the LTQ were as follows: normalized collision energy, 40%; activation q, 0.25; activation time 10 ms; 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 10 p.p.m., and MS/MS tolerance was set at 0.8 Da. Search criteria included carbamidomethylation of cysteine (+57.0214) as a fixed modification and oxidation of methionine (+15.9949) and appropriate SILAC labels ([$^{13}C_6$]lysine, [$^{13}C_6$]arginine for duplex and [$^{13}C_6$, $^{15}N_2$]lysine and [$^{13}C_6$, $^{15}N_4$]arginine 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 were filtered to satisfy false discovery rate (FDR) of 1%.

Integration of proteomic and RNA-seq data. A schematic workflow for our data analysis is given in **Supplementary Figure 2**. Briefly, there are two aspects: the PIT analysis and the integration of gene expression and protein abundance.

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 that 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) that contain information on the peptide concerned (for example, 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 that has an identified peptide associated with it, irrespective of whether the transcript mapped to the target genome. 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 ORF (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 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 because 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: (i) the precise location of the start of each identified peptide, (ii) a solid region representing the size of the peptide that is color-coded depending on the ratio changes between 0 h and 24 h, (iii) the confidence score reported by MaxQuant and (iv) all the quantitation ratios derived by MaxQuant.

For the gene expression–protein quantitation integration, the two data sets are integrated using a combination of text-file manipulation tools found in Galaxy and manual annotation in 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 data set was used for a BLAST search against two separate databases. First, we searched against the Ensembl list of human proteins to determine how many distinct human proteins were identified. Second, the list was searched against the nonredundant 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 were manually collated in 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 publicly available RNA-seq 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 were separated by SDS-PAGE. The sample was divided into 12 slices that were independently digested with trypsin in-gel and analyzed by LC-MS/MS as described above except that the top 20 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 rewritten to take into 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 they 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

doi:10.1038/nmeth.2227

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 6**, and it 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).

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.

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 Data 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 Data 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 was 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 used a database of comparable size to the those of other searches reported in this manuscript, thus improving the confidence that our identifications are not artifacts (i.e., resulting from a reduced-complexity data set). The outputs were 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*) 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 h, the cells were fixed with formaldehyde, permeabilized with Triton X-100 and processed for immunofluorescence using either anti-HA tag (anti HA serum F-7 from Santa Cruz, catalog 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, catalog number sc-9131, used at 1/100 dilution), anti-GAPDH (anti GAPDH serum FL335 from Santa Cruz, catalog number sc-25778, used at 1/100 dilution) and anti-POLDIP3 (used at 1/50 dilution). In each case, new samples of cells infected with adenovirus for 24 h were obtained, and the new samples were tested for protein expression alongside the original samples processed for quantitative proteomics. In addition, we treated cells with either DMSO or DMSO-containing 10 ng ml^{-1} MG132 for 8 h before harvesting infected or uninfected HeLa cells to determine the effect of proteasome inhibition on protein abundance.

doi:10.1038/nmeth.2227