SARS-CoV-2 candidate vaccine ChAdOx1 nCoV-19 infection of human cell lines reveals a normal low range of viral backbone gene expression alongside very high levels of SARS-CoV-2 S glycoprotein expression.

Abdulaziz Almuqrin
University of Bristol
Andrew D. Davidson
University of Bristol
Maia Kavanagh Williamson
University of Bristol
Phil Lewis
University of Bristol
Kate Heesom
University of Bristol
Susan Morris
University of Oxford
Sarah Gilbert
University of Oxford
David A. Matthews (d.a.matthews@bristol.ac.uk)
Universty of Bristol https://orcid.org/0000-0003-4611-8795

Research Article

Keywords: ChAdOx1 nCoV-19, adenovirus, SARS-CoV-2, vaccine, transcriptome, proteome

Posted Date: October 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-94837/v1

Version of Record: A version of this preprint was published on March 15th, 2021. See the published version at https://doi.org/10.1186/s13073-021-00859-1.

Abstract

Background: ChAdOx1 nCoV-19 is a recombinant adenovirus vaccine candidate against SARS-CoV-2. Although replication defective in normal cells, 28kbp of adenovirus genes are delivered to the cell nucleus alongside the SARS-CoV-2 S glycoprotein gene.

Methods: We used direct RNA sequencing to analyse transcript expression from the ChAdOx1 nCoV-19 genome in human MRC-5 and A549 cell lines that are non-permissive for vector replication alongside the replication permissive cell line, HEK293. In addition, we used quantitative proteomics to study over time the proteome and phosphoproteome of A549 and MRC5 cells infected with the ChAdOx1 nCoV-19 vaccine candidate.

Results: The expected SARS-CoV-2 S coding transcript dominated in all cell lines. We also detected rare S transcripts with aberrant splice patterns or polyadenylation site usage. Adenovirus vector transcripts were almost absent in MRC-5 cells but in A549 cells there was a broader repertoire of adenoviral gene expression at very low levels. Proteomically, in addition to S glycoprotein, we detected multiple adenovirus proteins in A549 cells compared to just one in MRC5 cells.

Conclusions: Overall the ChAdOx1 nCoV-19 vaccine's transcriptomic and proteomic repertoire is as expected. The combined transcriptomic and proteomics approaches provide an unparalleled insight into the behaviour of this important class of vaccine candidate and illustrate the potential of this technique to inform future viral vaccine vector design.

Introduction

Since the emergence of SARS-CoV-2 in late 2019 there has been a global effort to develop effective vaccines and at least four different adenovirus vectored vaccines have emerged as promising candidates[1-5]. Adenoviruses are non-enveloped viruses containing a linear double stranded DNA genome of approximately 36 kbp that efficiently deliver their DNA genome to the nuclei of host cells for viral genome replication. Adenovirus based vectors typically have two regions of the virus genome removed, known as E1 and E3[6]. The E1 region contains early genes required to trigger a transcription cascade enabling viral replication; E1 deleted vectors therefore need to be grown in E1 trans-complementing cell lines such as HEK293 cells[7]. HEK293 cells have a 4kbp region of human adenovirus type 5 (HuAd5) integrated into the cellular genome that provides the E1 genes *in trans* enabling efficient virus vector replication and recombinant virus production. The E3 region is comprised of genes encoding proteins that primarily act to subvert the immune response to adenovirus infection and are thus not needed for replication in cell culture and potentially undesirable from a vaccine platform perspective. Usually, the transgene to be expressed is inserted into the virus genome in place of the E1 region under the control of a highly active promoter.

Adenovirus based vaccines are grown in HEK293 cells (or equivalents) to very high titres, purified and then administered to individuals. The virus will attach to a host cell, enter and deliver the recombinant

genome to the nucleus where the desired gene of interest is expressed. Typically, the lack of E1 encoded proteins should ensure that there is no productive viral replication. However, it is noteworthy that removal of the E1 and E3 regions still leaves a number of other viral genes and promoters, including the E2 and E4 regions/promoters as well as the major late promoter, that normally drives expression of transcripts covering the major structural proteins of adenovirus (e.g. Figure 1a). Moreover, background expression of at least some of these adenovirus genes has been previously identified in studies with human adenovirus based vectors[8-11]. Expression of these viral vector genes may lead to the production of adenoviral antigens which in turn would bring the cell to the attention of the adaptive immune response [12-14]. In principle this could lead to vaccine vector infected cells being eliminated before an adaptive response to the vaccine target is raised; especially if repeat re-administration is desired. This has prompted the development of adenovirus vectors based on strains of adenovirus for which there is little or no pre-existing immunity in the general human population[15].

The ChAdOx1 vector virus is derived from chimpanzee adenovirus Y25 and is deleted for E1 and E3 genes. In addition, parts of the E4 region of the Y25 virus have been replaced with the equivalent E4 regions from HuAd5 to facilitate growth in HEK293 cells[16]. We have already shown that this vector is an effective vaccine platform in earlier studies[17-21]. In this study we wanted to determine in detail exactly which adenovirus genes on the ChAdOx1 genome were being expressed in non-permissive cell lines alongside the S glycoprotein from SARS-CoV-2 using the latest transcriptomic and proteomic approaches. We recently developed a pipeline to analyse and characterise transcriptomic data derived from direct RNA sequencing (dRNAseq) on Oxford Nanopore devices and we have used this pipeline to analyse the transcriptome of adenovirus[22] and SARS-CoV-2[23] during productive viral replication. Potentially of significance for adenovirus based vaccine delivery systems, we showed that in a wild type human adenovirus replication cycle in human cells there is a wide range of non-standard transcripts produced as a result of numerous atypical splicing and polyadenylation events[22]. Because of this, we also wanted to understand the transcriptomic repertoire of the S glycoprotein transcript in the ChAdOx1 vector backbone. This is critical because during SARS-CoV-2 replication the S glycoprotein transcript, like all SARS-CoV-2 transcripts, is made in the cytoplasm in the absence of host splicing and polyadenylation machinery[24]. Indeed, in the ChAdOx1 nCoV-19 vector, expression of the S glycoprotein is enhanced by inclusion of an intron between the transcription start site and the S glycoprotein ORF. Inclusion of introns in gene expression cassettes is thought to promote engagement of the nascent mRNA with the host splicing, polyadenylation and mRNA export machinery and has been previously shown to significantly enhance transgene protein expression in adenoviral vectors[25]. We wanted to understand if this addition of an intron before the S glycoprotein ORF could enable unwanted splicing events.

We determined that transcription of the S glycoprotein gene is, as expected, by far the dominant transcript generated in both non-permissive cell lines tested. Interestingly, we noted distinct transcriptomic repertoires from the ChAdOx1 vector were possible in the two non-permissive cell lines and that at a very low-level aberrant splicing and polyadenylation of the SARS-CoV-2 S glycoprotein transcript did occur in all settings. Proteomically, we were able to detect SARS-CoV-2 S glycoprotein and several ChAdOx1

proteins as well as confirming that the SARS-CoV-2 S protein is phosphorylated as we have previously reported[23]. Our findings are likely to be reflected in all adenovirus E1/E3 deleted replication defective type vectors irrespective of the parent adenovirus and represent a comprehensive and informative analysis of the transcriptomic repertoire of this important class of virus vaccine vector.

Materials And Methods

Virus and cells.

Human MRC-5 cells (a genetically normal male human lung fibroblast-like line) A549 cells (a human male lung epithelial-like continuous line derived from carcinomatous tissue) and HEK293 cells (human embryonic kidney epithelial cell line immortalised by human adenovirus E1 region) were obtained from the European Collection of Authenticated Cell Cultures (ECACC). The cells were cultured in DMEM supplemented with 10% foetal bovine serum, 100 U/ml penicillin and 100 ug/ml streptomycin. After reaching confluence, the cells were infected with ChAdOx1 nCoV-19 at a multiplicity of 10 infectious units per cell to ensure infections were synchronous. Duplicate flasks of infected cells were harvested at 24, 48 and 72 hours post infection (hpi) for MRC5-5 and A549 cells. One flask was used to extract total RNA whilst the duplicate was used to extract total protein content. For the HEK293 cells, just one flask was infected for total RNA extraction.

RNA extraction and sequencing

Total RNA was extracted from the infected cells using TRIzol™ reagent (#15596026, Ambion) at 1ml of reagent per 10^7 cells but with two additional washes of extracted RNA with 70% ethanol prior to storage at -80^{o}C under 70% ethanol. Once resuspended in H_2O, extracted RNA was immediately enriched for polyadenylated RNA and then immediately sequenced as we have described previously for dRNAseq of human adenovirus infected cells[22]. As before, the SQK-RNA002 kits and MIN106D R9 version flow cells (Oxford Nanopore Technologies) were used following the manufacturer's protocols exactly. Between 0.8 - 1.5 million QC-passed reads were typically obtained per flow cell over 48 hours.

Data analysis, characterisation of viral transcripts.

As before, our previously described ORF centric data analysis pipeline was used to characterise the RNA derived from the ChAdOx1 nCoV-19 genome[22]. Briefly, the transcripts were mapped to the viral genome with minimap2 and the mapping data was used to try to identify commonly used transcription start and termination locations alongside the splice acceptor/donor sites. Once this is complete, the software then assigns each transcript to a "transcript group" depending on its pattern of transcription start, transcription termination and splice acceptor/donor sites and counts how many transcripts belong to each transcript group. This data is then used by a second in-house script to generate pseudo transcripts based on the genomic sequences alongside a table of features predicted to be present on the vector genome (e.g. ORFS and predicted transcription start sites, Additional file 1). For each transcript group this script then

determines which features (if any) are contained within this transcript group. In addition, it notes which one (if any) of the known ORFs noted in the features table are 5' proximal for each transcript group. In this manner, transcripts with different structures that still code for the same given protein can be counted together. This script also produced GFF files describing the structure of each transcript group and also produces a GFF file describing the dominant transcript type coding for each ORF detailed in the table of features provided. Finally, the pipeline also produces a list of proteins that are 5' proximal for any given transcript group but are not identical to the list of known proteins predicted from the features table, referred to as the "proteins not known" list.

Characterisation of human transcripts

For each dataset minimap2[26] was used to map the transcripts to a list of human transcripts from Ensembl covering all human transcripts but with the fasta header modified so that each transcript included ENSG, ENST and ENSP information (Additional file 2). In this way, dRNAseq transcripts mapped to any given human transcript could be unambiguously assigned to its gene group and to the protein coded by that human transcript. An in-house script counted the number of times dRNAseq transcripts mapped to each curated Ensemble transcript and compared transcript mapping abundance between the mock and each of the time points after adjusting for the total number of mapped transcripts at each time point. This was collated as a simple table of log2 fold changes between mapping abundance of the mock sample transcripts time point and the mapping abundance of the transcripts at each time point. Further statistical evaluation was not applied since there is currently no consensus on how to use dRNAseq data to robustly infer gene expression changes or even the minimum depth of reads required[27].

Total and Phospho proteome analysis

Protein lysates were prepared from MRC5 or A549 cells only following mock infection, or infection with ChAdOx1 nCoV-19 for 24h, 48h or 72h as previously described[23].

Aliquots of 100µg of each sample were digested with trypsin (2.5µg trypsin per 100µg protein; 37°C, overnight), labelled with Tandem Mass Tag (TMT) ten plex reagents according to the manufacturer's protocol (Thermo Fisher Scientific, Loughborough, LE11 5RG, UK) and the labelled samples pooled according to cell line.

For the Total proteome analysis, an aliquot of 50µg of the pooled sample was desalted using a SepPak cartridge according to the manufacturer's instructions (Waters, Milford, Massachusetts, USA). Eluate from the SepPak cartridge was evaporated to dryness and resuspended in buffer A (20 mM ammonium hydroxide, pH 10) prior to fractionation by high pH reversed-phase chromatography using an Ultimate 3000 liquid chromatography system (Thermo Fisher Scientific). In brief, the sample was loaded onto an XBridge BEH C18 Column (130Å, 3.5 µm, 2.1 mm X 150 mm, Waters, UK) in buffer A and peptides eluted with an increasing gradient of buffer B (20 mM ammonium hydroxide in acetonitrile, pH 10) from 0-95% over 60 minutes. The resulting fractions (15 in total) were evaporated to dryness and resuspended in 1%

formic acid prior to analysis by nano-LC MSMS using an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific).

For the Phospho proteome analysis, the remainder of the TMT-labelled pooled sample was also desalted using a SepPak cartridge (Waters, Milford, Massachusetts, USA). Eluate from the SepPak cartridge was evaporated to dryness and subjected to TiO_2-based phosphopeptide enrichment according to the manufacturer's instructions (Pierce). The flow-through and washes from the TiO_2-based enrichment were then subjected to FeNTA-based phosphopeptide enrichment according to the manufacturer's instructions (Pierce). The phospho-enriched samples were again evaporated to dryness and then resuspended in 1% formic acid prior to analysis by nano-LC MSMS using an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific).

High pH RP fractions (Total proteome analysis) or the phospho-enriched fractions (Phospho-proteome analysis) were further fractionated using an Ultimate 3000 nano-LC system in line with an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific). In brief, peptides in 1% (vol/vol) formic acid were injected onto an Acclaim PepMap C18 nano-trap column (Thermo Scientific). 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 (Thermo Scientific) 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 5min., 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 2.0kV using a stainless-steel emitter with an internal diameter of 30 μm (Thermo Scientific) and a capillary temperature of 300°C.

All spectra were acquired using an Orbitrap Fusion Lumos mass spectrometer controlled by Xcalibur 3.0 software (Thermo Scientific) and operated in data-dependent acquisition mode using an SPS-MS3 workflow. FTMS1 spectra were collected at a resolution of 120 000, with an automatic gain control (AGC) target of 400 000 and a max injection time of 100ms. Precursors were filtered with an intensity threshold of 5000, according to charge state (to include charge states 2-7) and with monoisotopic peak determination set to Peptide. Previously interrogated precursors were excluded using a dynamic window (60s +/-10ppm). The MS2 precursors were isolated with a quadrupole isolation window of 0.7m/z. ITMS2 spectra were collected with an AGC target of 10 000, max injection time of 70ms and CID collision energy of 35%.

For FTMS3 analysis, the Orbitrap was operated at 30 000 resolution with an AGC target of 50 000 and a max injection time of 105ms. Precursors were fragmented by high energy collision dissociation (HCD) at a normalised collision energy of 60% to ensure maximal TMT reporter ion yield. Synchronous Precursor Selection (SPS) was enabled to include up to 5 MS2 fragment ions in the FTMS3 scan.

Proteomics Data Analysis

The raw data files were processed using Proteome Discoverer software v2.1 (Thermo Scientific) and searched against a bespoke human database with the fasta headers amended to contain ENSG, ENST and ENSP notation of target human protein concatenated with a custom ChAdOx1 nCoV-19 protein database consisting of a list of “known proteins” predicted to exist by homology with known adenovirus proteins alongside the “proteins not known” list predicted by the transcriptome analysis pipeline described above (Additional file 3). Searches against this database and against an in-house ‘common contaminants’ database were performed using the SEQUEST HT algorithm. Peptide precursor mass tolerance was set at 10 ppm, and MS/MS tolerance was set at 0.6 Da. Search criteria included oxidation of methionine (+15.995 Da), acetylation of the protein N-terminus (+42.011 Da) and methionine loss plus acetylation of the protein N-terminus (-89.03 Da) as variable modifications and carbamidomethylation of cysteine (+57.021Da) and the addition of the TMT mass tag (+229.163) to peptide N-termini and lysine as fixed modifications. For the phospho-proteome analysis, phosphorylation of serine, threonine and tyrosine (+79.966 Da) was also included as a variable modification. Searches were performed with full tryptic digestion and a maximum of 2 missed cleavages were allowed. The reverse database search option was enabled and all data was filtered to satisfy a false discovery rate (FDR) of 1%. Quantitative changes in phosphopeptide abundance were normalised with respect to total protein abundance.

Data integration.

Once the quantitative data for proteome, phosphoproteome and transcriptome were collected an inhouse script was used to integrate the data producing a seamless dataset where for any given gene, data is provided on the number of transcripts mapping to that gene at any time point alongside log2 fold changes in protein or phosphoprotein abundance (if available). For the purposes of noting changes in phospho peptide abundance, if multiple phosphorylation sites were identified for any given protein, in the integrated dataset the largest fold change (positive or negative) is the one reported.

Data Availability

All the raw data files (fastq files, protoemics .RAW files and phophoproteomics .RAW files) are available at Zenodo.org. The fastq data is at http://doi.org/10.5281/zenodo.4044565, the MRC5 proteomics data is at http://doi.org/10.5281/zenodo.4044603 and the A549 proteomics data is at http://doi.org/10.5281/zenodo.4044593 .

Results

Overview of sequencing data outputs

Table 1 illustrates the number of sequence reads obtained for each timepoint and cell line combination alongside the number of reads that mapped to the human transcriptome, the ChAdOx1 nCoV-19 genome or the region of HuAd5 present in HEK293 cells. For MRC-5 cells there was a peak of reads mapping to the ChAdOx1 genome at 48 hpi but overall the proportion of reads does not vary markedly. For A549 cells there was a slight decline over time but notably there were approximately five-fold fewer reads as a

proportion of human mapped transcripts mapping to the ChAdOx1 genome compared to MRC-5 cells. As expected, the proportion of reads that mapped to the ChAdOx1 genome in HEK293 cells, which are permissive for viral genome replication, was significantly higher. We also detected over a thousand reads mapping to the E1 and pIX region in HEK293 cells; this region covers just over 4000 bp from HuAd5 that is integrated into the HEK293 genome[28]. We utilised our previously published ORF-centric analysis pipeline to further characterise the transcripts allowing a qualitative and quantitative understanding of gene expression as we have previously done for wild type isolates of both human adenovirus type 5 and SARS-CoV-2[22,23]. Briefly, this analysis pipeline uses the genome sequence to correct the error prone transcripts generated by the nanopore device. The dRNAseq data is biased towards the 3' polyadenylated end so we further constrained our analysis to transcripts whose mapped start site was within 50 bp of viral promoters as described for our analysis of wild type HuAd5 and SARS-CoV-2. After correction, each transcript was scanned to determine the 5'-most ORF which is compared to the locations and sequences of known ORFs and other features supplied by the user. In this way the pipeline can identify and count the presence of known features on transcripts and ultimately provide quantitative data on the abundance of transcripts that code for each ORF. Moreover, this analysis illustrates the structure of the dominant transcript that codes for each ORF as well as characterising the structure of minor transcripts that could also code for any given ORF (Additional files 4, 5 and 6).

Comparison of ChAdOx1 nCoV-19 adenoviral gene expression levels in non-permissive and permissive cells.

Strikingly, in MRC-5 cells the ChAdOx1 vector backbone genes (i.e. excluding the S glycoprotein gene) were hardly transcribed at any timepoint but in A549 cells we saw a wide range of transcripts that code for almost every protein predicted to be expressed by the ChAdOx1 backbone (Figures 1b, 1c and Table 2). The pattern of dominant transcripts seen in A549 cells was broadly typical of mammalian adenoviruses[29] and was similar to the expression pattern of the virus grown in permissive HEK293 cells (compare Figures 1c and 1d). Looking quantitatively, Table 2 shows that in A549 cells a large number of different ChAdOx1 vector backbone genes were expressed to one degree or another with the adenovirus protein DBP being dominant. Notably, ChAdOx1 vector backbone gene expression declined over the time course analysed.

Comparing the expression of ChAdOx1 gene expression in non-permissive vs permissive HEK293 cells illustrated how deletion of the E1 region affects ChAdOx1 gene expression in A549 cells where some vector backbone gene expression has occurred (Table 2). One key difference was the level of expression of the fibre (L5) transcript which was the dominant transcript in HEK293 cells (9.65% of transcripts mapped to ChAdOx1 nCoV19) after the S glycoprotein transcript but is expressed only at very low levels in A549 cells (0.46% of transcripts mapped to ChAdOx1 nCoV19 at its peak at 48 hpi, Table 2).

Analysis of the transcriptome of ChAdOx1 nCoV-19 in HEK293 cells also allowed the observation of transcription from the region of HuAd5 that is integrated into HEK293 cells (Figure 1e). Notably, there is some expression of HuAd5 pIX from this region (Table 2) which was approximately 100 fold lower than

the ChAdOx1 derived pIX protein. In addition, in the non-permissive cell lines there were transcripts that code for protein IX, which appear to be derived from RNA pol II and have ignored the polyA signal after the S glycoprotein and instead used the pIX poly A signal and spliced out the coding region for S glycoprotein (Figures 1b and 1c). In permissive HEK293 cells, the dominant transcript coding for pIX appeared to revert to the expected transcript for pIX with the transcription start site just upstream of the initiating AUG (Figure 1d). Notably, the type of transcript seen in the non-permissive cells was still present in HEK293 cells, but it was no longer the dominant transcript coding for pIX (Additional file 6).

Expression of SARS-CoV-2 S glycoprotein transcripts

In both permissive and non-permissive cell lines, the most abundant ChAdOx1 nCoV-19 virus vector transcript, as expected, was the transcript for expression of the SARS-CoV-2 S glycoprotein. However, a number of transcripts were identified in all cell lines at low levels which appear to arise from aberrant splicing events. We also found evidence that occasionally some transcripts were spliced to other elements (notably pIX); these would still code for SARS-CoV-2 S glycoprotein but would be polycistronic (Figure 2). We also saw limited evidence of transcripts that extended beyond the pIX polyA signal and deep into the ChAdOx1 vector backbone in HEK293 cells when there was active viral replication. However, it is important to place these rare transcripts in context - the vast majority of transcripts starting at the transcription start site for the S glycoprotein message would generate mRNA that codes for S glycoprotein (Table 3).

Additional polyadenylation site for ChAdOx1 preVI transcripts.

Typically, in human adenoviruses the late transcripts code for most of the structural viral proteins. The late transcripts typically originate at the major late promoter and include three exons (known as the tri-partite leader) and optional i-leader exon and are grouped into five classes called L1 to L5 based on which of the five major polyadenylation sites they use[29,30]. The transcript for preVI is usually regarded as part of the L3 group of transcripts which share a common polyadenylation site but utilise different splice acceptor/donor pairs to place one of three different ORFs (preVI, hexon or 23K) proximal to the 5’ cap for translation. In the case of ChAdOx1 replicating in HEK293 cells, although there were transcripts coding for preVI that would fit this pattern (e.g. utilising the same poly adenylation site as transcripts for hexon and 23K), the dominant transcript for preVI instead utilises an additional polyadenylation site upstream of the start codon for the hexon protein (Figure 3a). This was also the case in A549 cells (Figure 1b) where 12 transcripts utilised this additional polyadenylation site compared to just 4 using the classical L3 polyadenylation site (Additional file 5). Examining the sequence of the viral genome in the region of this additional polyadenylation site reveals a GU rich region preceded by a classical polyadenylation signal[31] (Figure 3b) – this signal is not present in the equivalent region of HuAd5 (Figure 3c) where no such preVI specific polyadenylation site has been observed[22]. However, this additional polyadenylation site is present in human adenovirus type 4 (HuAd4, Figure 3c) which is also a member of the group E mammalian adenoviruses like chimpanzee adenovirus Y25 from which ChaAdOx1 was derived.

Proteomic detection of SARS-CoV-2 S glycoprotein and ChAdOx1 proteins.

As expected, we were able to detect a range of S glycoprotein peptides in MRC5 and A549 cells as well as one phosphorylation site (Ser 1292) derived from the C-terminal portion of the glycoprotein which is internal to the viral particle (Table 4 and Additional files 7 - 14). For the vector backbone expressed proteins we were only able to detect E4ORF3 in MRC5 infected cells whereas in A549 cells in addition to E4ORF3 we also detected DBP and hexon (Table 4). Interestingly in both MRC5 and A549 cells we detected a peptide (SYLTPGDSSSGWTAGAAAY, aa248-265) apparently derived from a single transcript designated mRNA#962. However, this transcript has only one copy present in the transcriptome of the A549 infected cells and a similar transcript was not detected in the MRC-5 cells. Alternatively, we believe it is more likely that the detected peptide resulted from proteolytic activity at aa265 on full length S glycoprotein either from intracellular processes or during sample harvest and preparation. In support of this, an analysis of the proteomics data to include semi tryptic peptides did identify the additional peptide (i.e. on the C-terminal side of an expected cleavage at aa265) which would be consistent with proteolytic activity at this site. The list of possible proteins predicted by translation of the 5'-most ORF of all the observed transcripts is very large and varied and they were all used in the initial searches for peptides. However, this peptide was the only one that could not be predicted by the standard list of expected ChAdOx1 proteins. The proteomic analysis supported the transcriptomic analysis suggesting that the ChAdOx1 nCoV-19 vaccine does not make additional unexpected proteins. Notably, in both A549 and MRC5 cells we observed some of the highest fold increases over time for the S glycoprotein as expected (Additional files 7 and 9).

Analysis of the cellular total transcriptome, proteome and phosphoproteome

Whilst our primary focus was on the transcriptome and proteome of the ChAdOx1 nCov-19 vaccine, we also collected data on the host cell counterparts which are summarised in additional files 15 and 16. The two cell lines appear to have distinct patterns of response to the vaccine virus. We collated lists of proteins that were increased or decreased in abundance at least two-fold at each time point and utilised STRING pathway analysis to determine if cellular pathways were apparently overrepresented (Additional files 15 and 16). Whilst the host cell responses to the vaccine are varied, we did note that in MRC-5 cells there was an apparent enrichment for upregulated proteins involved in the unfolded protein response and ER stress, which was not the case for A549 cells. In A549 cells there was some evidence that ChAdOx1 nCoV-19 infection affected proteins involved in ribosome biogenesis and host mRNA splicing.

Discussion

Recombinant adenovirus vaccines have been developed over many years and have shown great promise as safe and effective vaccine platforms for mass vaccination programmes. Despite their widespread development, this is the first study to directly and comprehensively survey the transcriptomic repertoire of a replication defective adenovirus in a non-permissive host cell. A significant advantage of using simian based adenoviruses like the ChAdOx1 platform to study the transcriptome of E1 deleted adenoviruses is

the absence of replication competent adenoviruses (RCA)[32]. RCAs can arise from homology between the HuAd5 sequences present in the HEK293 cells used to produce the vaccine vector and sequences present in the recombinant vector itself, which are frequently based on HuAd5. Whilst the emergence of an RCA could confound such an analysis in other adenoviral backgrounds, it is not a factor here due to insufficient homology between chimpanzee adenoviruses and the sequences in HEK293 cells[33]. Given the significant advantages in using simian based vectors as a human vaccine platform[34] we wanted to deepen our understanding of these virus vectors on a molecular level. We were especially keen to determine that no additional unanticipated transcripts or proteins were being made. Any one of such proteins could be antigenic with unintended consequences (e.g. generating auto-immune responses for example).

The two host cell lines chosen for this study have distinct properties despite both being derived from male human lungs. The A549 cell line is an immortal cancerous cell line with an average of 66 chromosomes and a deletion in the CDKN2A locus leaving the cell line defective in the p53/ARFp14/MDM2 pathway. By contrast MRC5 cells are genetically normal with a finite cell passage capability. Thus, the distinct transcriptional profiles seen from the ChAdOx1 nCoV-19 vaccine could result from the A549 cells being genetically defective and immortal or it may be connected to unrelated differences in the intracellular environment. Previous work with human adenovirus based vectors has suggested that a cellular E1A-like activity is in some manner connected to pre-existing levels of heat shock proteins, with promyelocytic leukemia protein (PML) also being implicated[35,36]. Notably, in the mock transcriptomic data there are higher levels of sequenced HSPA1A and HSPA1B transcripts in A549 cells compared to MRC-5 cells prior to infection with ChAdOx1 nCoV-19 with the reverse being true for PML transcripts (Additional files 15 and 16). We noted that in MRC5 cells, levels of transcripts for HSPA1A and HSPA1B show increases over time and for HSPA1B there is a corresponding protein increase (Additional files 15 and 16). Whereas in A549 cells there is a decrease in sequenced transcripts for HSPA1A and HSPA1B with no significant change in protein levels for HSPA1A. This supports previous research which implies heat shock (and thus increased expression of heat shock proteins) does not overcome the lack of E1A expression[36] as in MRC5 cells we do not see increasing amounts of ChAdOx1 vector backbone transcription. In addition, NFκB has been implicated in enabling HuAd5 based vector backbone expression in non-permissive cell lines[37]. However, no marked differences were observed between the two cell lines in the transcriptomic abundance data for NFKB1 in our datasets (Additional files 15 and 16). A deeper understanding of the intracellular environments that allow the ChAdOx1 vector to overcome the lack of E1A, even in a limited fashion, could lead to better ways to prevent this which in turn may lead to improved transgene expression. That we observe quite distinct vaccine backbone expression in cell lines from the same tissue site is notable. However, this vaccine vector is currently administered intramuscularly and so longer term it would be useful to examine a wider range of cell types or even biopsy material.

We have previously analysed the proteome of purified adenovirus particles, both wild type HuAd5 and HuAd5 based recombinant vaccine vectors, including a sample of adenovirus manufactured to clinical

grade. We were able to detect non-structural proteins DBP, 100K and E4 14.7K protein in purified virus particles in addition to the expected viral structural proteins[38]. If any cells in a vaccinee did express the full range of adenovirus proteins over time similar to the A549 transcriptomic profile, then T-cell responses to the virus vector could both derive from the incoming viral proteins and from subsequent low-level expression of any of the remaining virus vector backbone genes. That some cells allow low level expression of ChAdOx1 vector backbone transcripts would be consistent with data from human adenovirus vector studies where additional deletions in E2 and E4 or helper-dependent adenovirus vectors (where essentially all the vector backbone genes are removed) were shown to afford longer transgene expression *in vivo*[12,14,39-41]. The continuing low-level expression of vector backbone genes is therefore likely to be the main driver of immune mediated clearance of cells infected with E1/E3 deleted adenoviruses. In addition, low level expression of vector backbone genes in A549 cells may be related to our finding that A549 cells express lower levels of S glycoprotein mRNA as a proportion of total mRNA than MRC5 cells (Table 1).

Our proteomics analysis is consistent with the transcriptomics data in that the S glycoprotein is readily detected by MS/MS and that we detect a slightly wider array of ChAdOx1 proteins in A549 cells compared to MRC-5 cells. We do not detect any evidence of additional unexpected proteins being expressed by the ChAdOx1 nCoV19 vaccine despite a broad search of proteins coded for by all transcripts that map to the ChAdOx1 nCoV-19 genome.

We previously analysed the transcriptomic repertoire of wild type HuAd5 in MRC-5 cells showing that during a productive infection there is significant low-level heterogeneity in the usage of both splice sites and poly adenylation signals[22]. Here, we show that as with the human virus there is similar low-level expression of transcripts with aberrant splice site and polyadenylation signal usage in cell lines infected with the recombinant ChAdOx1 nCoV-19 virus in replication permissive and non-permissive cell lines. Critically, the overwhelming majority of S glycoprotein transcripts have the expected structure with only a small minority of transcripts originating from the S glycoprotein transcription start site being unable to express the S glycoprotein. The most common issue seems to be failure to use the polyadenylation signal immediately after the S glycoprotein ORF. This leads to, for example, usage of the pIX polyadenylation site and the opportunity for splicing events that may disrupt S glycoprotein expression.

Notably in our analysis of wildtype HuAd5 transcripts we proposed that a key aspect of HuAd5 evolution would be the exploration of different splice acceptor/donor sites and initiating codon combinations. Here analysis of expression from the ChAdOx1 backbone identified alternative usage of polyadenylation signals. Typically, the polyadenylation signal for the preVI protein transcript would be shared with the transcripts for hexon and 23K – collectively known as the L3 group since they share this polyadenylation signal. Protein preVI uses two polyadenylation signals, one that is the L3 location used by the hexon and 23K transcripts and a second just after its own stop codon. That this second one is not observed in human adenovirus type 5 is likely due to the lack of a strong polyadenylation signal as shown in Figure 3. However, this arrangement seems to be present both in HuAd4 and chimpanzee adenovirus Y25 (which the ChAdOx1 vector is based on) which are both group E adenoviruses. Indeed, HuAd4 is closely enough

related to chimpanzee adenoviruses to suggest that the human virus may have jumped into chimpanzees at some point in the past[42]. What advantage, if any, the two polyadenylation sites offer the ChAdOx1 virus is not clear but use of the polyadenylation site after the preVI ORF would by necessity reduce the influence of splice site selection in determining whether preVI, hexon or 23K would be expressed from transcripts utilising the canonical L3 polyadenylation site. This could, in principle, bias gene expression in favour of preVI relative to hexon or 23K. Whether this additional polyadenylation site is being selected for or against remains an open question.

Examination of the HEK293 cell data revealed, as expected, expression of HuAd5 transcripts corresponding to E1 region genes dominated by E1b19K, E1a12S and E1a13S as we saw for wild type HuAd5 gene expression at 16hpi infection[22]. As noted in the results, in HEK293 cells we observed a small number of HuAd5 pIX transcripts, approximately 100-fold fewer than ChadOx1 pIX transcripts. However, with 240 copies of protein IX per virus capsid[43-45] it is possible that the ChAdOx1 virus particles grown in HEK293 cells (or derivatives) contain one or two copies of HuAd5-pIX per viral particle assuming the two proteins are equivalently interchangeable on the virus particle.

Conclusions

We believe this kind of analysis provides valuable insight into the transcriptomic and proteomic repertoire of an important vaccine candidate, providing confirmation that the vaccine vector's transcriptome is essentially as intended in these cell lines. Recently, an in-depth multi 'Omic analysis of the herpes virus genome revealed that an oncolytic herpes virus licenced in 2015 under the name Imlygic is in fact deleted for an additional third gene rather than the two intended because of a previously unknown ORF present in the deleted region[46]. Whilst there is no suggestion that this has been a problem, it highlights the importance of utilising state of the art and unbiased approaches to survey genetically modified viruses intended for clinical use. Finally, we argue that this kind of analysis is relatively straightforward and should be routinely incorporated into the early stages of future viral vector evaluation pipelines to allow a robust understanding of the transcriptomic potential of engineered viral vectors.

Declarations

Acknowledgements

D.A.M. and A.D.D. were supported by the BBSRC (grant BB/M02542X/1) and by the United States Food and Drug Administration grant number HHSF223201510104C 'Ebola Virus Disease: correlates of protection, determinants of outcome and clinical management' amended to incorporate urgent COVID-19 studies. SG and SM were funded through VaxHub from the Engineering and Physical Sciences Research Council (EP/R013756/1).

Competing interests.

SG is co-founder of Vaccitech (co-inventors of this vaccine candidate) and named as an inventor on a patent covering use of ChAdOx1-vectored vaccines and a patent application covering this SARS-CoV-2 vaccine. AA, ADD, MKW, PL, KH, SM and DAM declare that they have no competing interests.

References

1 Folegatti, P. M. *et al.* Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. *Lancet* **396**, 467-478, doi:10.1016/S0140-6736(20)31604-4 (2020).

2 Logunov, D. Y. *et al.* Safety and immunogenicity of an rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine in two formulations: two open, non-randomised phase 1/2 studies from Russia. *Lancet*, doi:10.1016/S0140-6736(20)31866-3 (2020).

3 Mercado, N. B. *et al.* Single-shot Ad26 vaccine protects against SARS-CoV-2 in rhesus macaques. *Nature*, doi:10.1038/s41586-020-2607-z (2020).

4 Zhu, F. C. *et al.* Immunogenicity and safety of a recombinant adenovirus type-5-vectored COVID-19 vaccine in healthy adults aged 18 years or older: a randomised, double-blind, placebo-controlled, phase 2 trial. *Lancet* **396**, 479-488, doi:10.1016/S0140-6736(20)31605-6 (2020).

5 Hassan, A. O. *et al.* A Single-Dose Intranasal ChAd Vaccine Protects Upper and Lower Respiratory Tracts against SARS-CoV-2. *Cell*, doi:10.1016/j.cell.2020.08.026 (2020).

6 Wold, W. S. & Toth, K. Adenovirus vectors for gene therapy, vaccination and cancer gene therapy. *Curr Gene Ther* **13**, 421-433, doi:10.2174/15665232133666131125095046 (2013).

7 Graham, F. L., Smiley, J., Russell, W. C. & Nairn, R. Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *J Gen Virol* **36**, 59-74, doi:10.1099/0022-1317-36-1-59 (1977).

8 Saha, B. & Parks, R. J. Human adenovirus type 5 vectors deleted of early region 1 (E1) undergo limited expression of early replicative E2 proteins and DNA replication in non-permissive cells. *PLoS One* **12**, e0181012, doi:10.1371/journal.pone.0181012 (2017).

9 Gorziglia, M. I. *et al.* Elimination of both E1 and E2 from adenovirus vectors further improves prospects for in vivo human gene therapy. *J Virol* **70**, 4173-4178, doi:10.1128/JVI.70.6.4173-4178.1996 (1996).

10 Rittner, K., Schultz, H., Pavirani, A. & Mehtali, M. Conditional repression of the E2 transcription unit in E1-E3-deleted adenovirus vectors is correlated with a strong reduction in viral DNA replication and late gene expression in vitro. *J Virol* **71**, 3307-3311, doi:10.1128/JVI.71.4.3307-3311.1997 (1997).

11 Shimizu, K., Sakurai, F., Machitani, M., Katayama, K. & Mizuguchi, H. Quantitative analysis of the leaky expression of adenovirus genes in cells transduced with a replication-incompetent adenovirus

vector. *Mol Pharm* **8**, 1430-1435, doi:10.1021/mp200121z (2011).

12 Yang, Y. *et al.* Cellular immunity to viral antigens limits E1-deleted adenoviruses for gene therapy. *Proc Natl Acad Sci U S A* **91**, 4407-4411, doi:10.1073/pnas.91.10.4407 (1994).

13 Nakai, M. *et al.* Expression of pIX gene induced by transgene promoter: possible cause of host immune response in first-generation adenoviral vectors. *Hum Gene Ther* **18**, 925-936, doi:10.1089/hum.2007.085 (2007).

14 Morsy, M. A. *et al.* An adenoviral vector deleted for all viral coding sequences results in enhanced safety and extended expression of a leptin transgene. *Proc Natl Acad Sci U S A* **95**, 7866-7871, doi:10.1073/pnas.95.14.7866 (1998).

15 Bangari, D. S. & Mittal, S. K. Development of nonhuman adenoviruses as vaccine vectors. *Vaccine* **24**, 849-862, doi:10.1016/j.vaccine.2005.08.101 (2006).

16 Dicks, M. D. *et al.* A novel chimpanzee adenovirus vector with low human seroprevalence: improved systems for vector derivation and comparative immunogenicity. *PLoS One* **7**, e40385, doi:10.1371/journal.pone.0040385 (2012).

17 van Doremalen, N. *et al.* A single dose of ChAdOx1 MERS provides protective immunity in rhesus macaques. *Sci Adv* **6**, eaba8399, doi:10.1126/sciadv.aba8399 (2020).

18 Stedman, A. *et al.* Safety and efficacy of ChAdOx1 RVF vaccine against Rift Valley fever in pregnant sheep and goats. *NPJ Vaccines* **4**, 44, doi:10.1038/s41541-019-0138-0 (2019).

19 López-Camacho, C. *et al.* Rational Zika vaccine design via the modulation of antigen membrane anchors in chimpanzee adenoviral vectors. *Nat Commun* **9**, 2441, doi:10.1038/s41467-018-04859-5 (2018).

20 Coughlan, L. *et al.* Heterologous Two-Dose Vaccination with Simian Adenovirus and Poxvirus Vectors Elicits Long-Lasting Cellular Immunity to Influenza Virus A in Healthy Adults. *EBioMedicine* **29**, 146-154, doi:10.1016/j.ebiom.2018.02.011 (2018).

21 Antrobus, R. D. *et al.* Clinical assessment of a novel recombinant simian adenovirus ChAdOx1 as a vectored vaccine expressing conserved Influenza A antigens. *Mol Ther* **22**, 668-674, doi:10.1038/mt.2013.284 (2014).

22 Donovan-Banfield, I., Turnell, A. S., Hiscox, J. A., Leppard, K. N. & Matthews, D. A. Deep splicing plasticity of the human adenovirus type 5 transcriptome drives virus evolution. *Commun Biol* **3**, 124, doi:10.1038/s42003-020-0849-9 (2020).

23 Davidson, A. D. *et al.* Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. *Genome*

Med **12**, 68, doi:10.1186/s13073-020-00763-0 (2020).

24 Wang, Y., Grunewald, M. & Perlman, S. Coronaviruses: An Updated Overview of Their Replication and Pathogenesis. *Methods Mol Biol* **2203**, 1-29, doi:10.1007/978-1-0716-0900-2_1 (2020).

25 Matthews, D. A., Cummings, D., Evelegh, C., Graham, F. L. & Prevec, L. Development and use of a 293 cell line expressing lac repressor for the rescue of recombinant adenoviruses expressing high levels of rabies virus glycoprotein. *J Gen Virol* **80 (Pt 2)**, 345-353, doi:10.1099/0022-1317-80-2-345 (1999).

26 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).

27 Soneson, C. *et al.* A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat Commun* **10**, 3359, doi:10.1038/s41467-019-11272-z (2019).

28 Louis, N., Evelegh, C. & Graham, F. L. Cloning and sequencing of the cellular-viral junctions from the human adenovirus type 5 transformed 293 cell line. *Virology* **233**, 423-429, doi:10.1006/viro.1997.8597 (1997).

29 Akusjarvi, G. Temporal regulation of adenovirus major late alternative RNA splicing. *Front Biosci* **13**, 5006-5015, doi:10.2741/3059 (2008).

30 Chow, L. T. & Broker, T. R. The spliced structures of adenovirus 2 fiber message and the other late mRNAs. *Cell* **15**, 497-510, doi:10.1016/0092-8674(78)90019-3 (1978).

31 Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10**, 1001-1010, doi:10.1101/gr.10.7.1001 (2000).

32 Lochmuller, H. *et al.* Emergence of early region 1-containing replication-competent adenovirus in stocks of replication-defective adenovirus recombinants (delta E1 + delta E3) during multiple passages in 293 cells. *Hum Gene Ther* **5**, 1485-1491, doi:10.1089/hum.1994.5.12-1485 (1994).

33 Tatsis, N. *et al.* Chimpanzee-origin adenovirus vectors as vaccine carriers. *Gene Ther* **13**, 421-429, doi:10.1038/sj.gt.3302675 (2006).

34 Morris, S. J., Sebastian, S., Spencer, A. J. & Gilbert, S. C. Simian adenoviruses as vaccine vectors. *Future Virol* **11**, 649-659, doi:10.2217/fvl-2016-0070 (2016).

35 Atwan, Z., Wright, J., Woodman, A. & Leppard, K. N. Promyelocytic leukemia protein isoform II inhibits infection by human adenovirus type 5 through effects on HSP70 and the interferon response. *J Gen Virol* **97**, 1955-1967, doi:10.1099/jgv.0.000510 (2016).

36 Imperiale, M. J., Kao, H. T., Feldman, L. T., Nevins, J. R. & Strickland, S. Common control of the heat shock gene and early adenovirus genes: evidence for a cellular E1A-like activity. *Mol Cell Biol* **4**, 867-874, doi:10.1128/mcb.4.5.867 (1984).

37 Machitani, M. *et al.* NF-κB promotes leaky expression of adenovirus genes in a replication-incompetent adenovirus vector. *Scientific Reports* **6**, 19922, doi:10.1038/srep19922 (2016).

38 Alqahtani, A. *et al.* Analysis of purified wild type and mutant adenovirus particles by SILAC based quantitative proteomics. *J Gen Virol* **95**, 2504-2511, doi:10.1099/vir.0.068221-0 (2014).

39 Rhee, E. G. *et al.* Multiple innate immune pathways contribute to the immunogenicity of recombinant adenovirus vaccine vectors. *J Virol* **85**, 315-323, doi:10.1128/JVI.01597-10 (2011).

40 Coughlan, L. Factors Which Contribute to the Immunogenicity of Non-replicating Adenoviral Vectored Vaccines. *Frontiers in Immunology* **11**, doi:10.3389/fimmu.2020.00909 (2020).

41 Lusky, M. *et al.* In vitro and in vivo biology of recombinant adenovirus vectors with E1, E1/E2A, or E1/E4 deleted. *J Virol* **72**, 2022-2032, doi:10.1128/JVI.72.3.2022-2032.1998 (1998).

42 Purkayastha, A. *et al.* Genomic and bioinformatics analysis of HAdV-4, a human adenovirus causing acute respiratory disease: implications for gene therapy and vaccine vector development. *J Virol* **79**, 2559-2572, doi:10.1128/JVI.79.4.2559-2572.2005 (2005).

43 van Oostrum, J. & Burnett, R. M. Molecular composition of the adenovirus type 2 virion. *J Virol* **56**, 439-448, doi:10.1128/JVI.56.2.439-448.1985 (1985).

44 Reddy, V. S. & Nemerow, G. R. Structures and organization of adenovirus cement proteins provide insights into the role of capsid maturation in virus entry and infection. *Proc Natl Acad Sci U S A* **111**, 11715-11720, doi:10.1073/pnas.1408462111 (2014).

45 Kundhavai Natchiar, S., Venkataraman, S., Mullen, T. M., Nemerow, G. R. & Reddy, V. S. Revised Crystal Structure of Human Adenovirus Reveals the Limits on Protein IX Quasi-Equivalence and on Analyzing Large Macromolecular Complexes. *J Mol Biol* **430**, 4132-4141, doi:10.1016/j.jmb.2018.08.011 (2018).

46 Whisnant, A. W. *et al.* Integrative functional genomics decodes herpes simplex virus 1. *Nat Commun* **11**, 2038, doi:10.1038/s41467-020-15992-5 (2020).

Tables

Due to technical limitations, tables 1-4 are only available as a download in the supplemental files section.

Figures

Figure 1

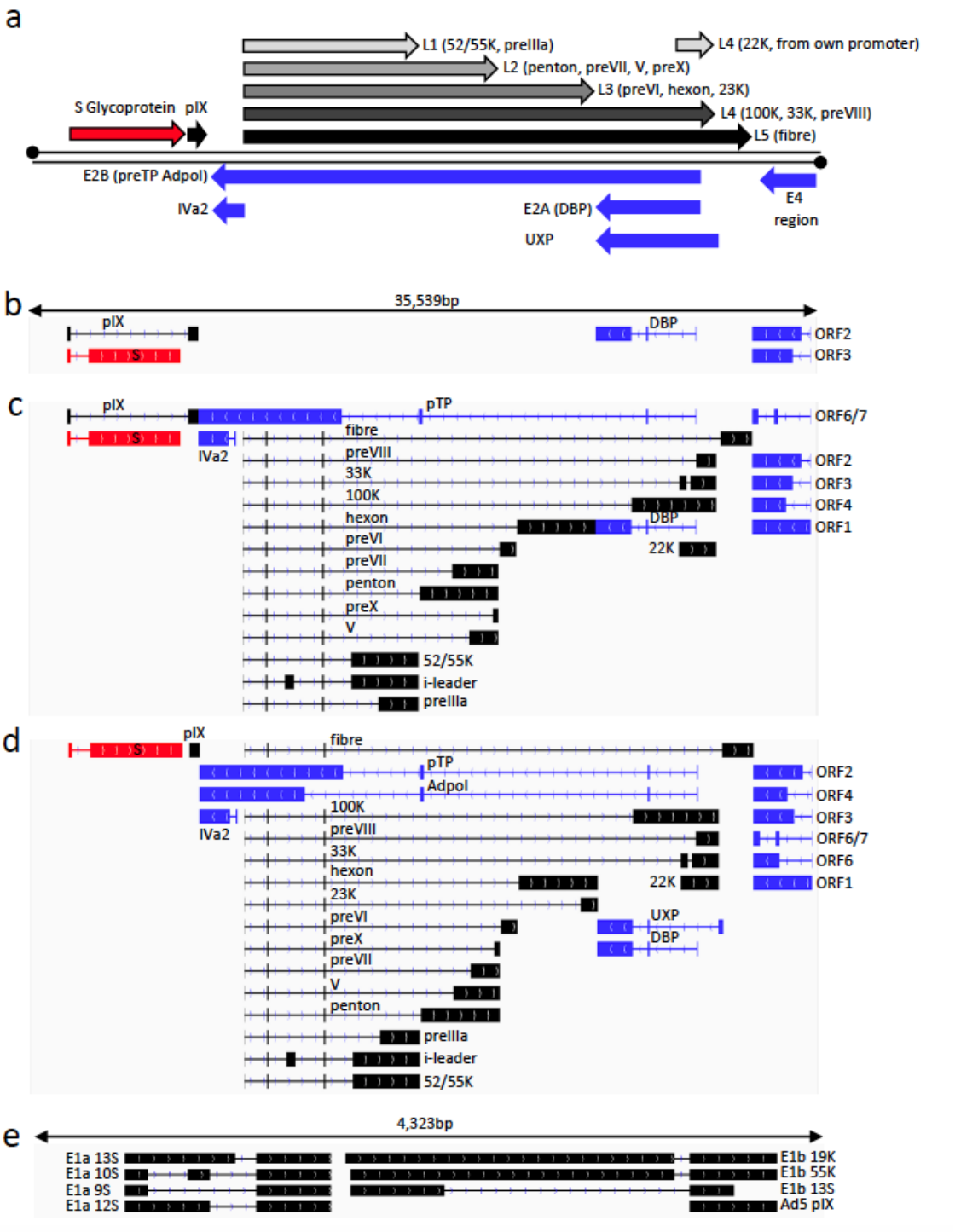

Figure 1

Transcription map of detected transcripts in non-permissive and permissive human cells. Part a shows a simplified schematic of the major transcripts predicted to be made from ChAdOx1 nCoV-19 based on the design of the recombinant vector and from homology with other mammalian adenoviruses. The red arrow indicates the location of the S glycoprotein transcript which is in place of the E1 region of ChAdOx1. Black transcripts are transcribed from the top strand of the dsDNA viral genome and blue

transcripts are transcribed from the complementary strand. Each arrow depicts the expected transcription start and polyadenylation site, importantly there are usually one or more splicing events within this transcript to form the mature mRNA. Note that for the structural proteins of ChAdOx1, most are transcribed starting at the same promoter (known as the major late promoter) and end at one of 5 different polyadenylation sites (known as L1 – L5). However, the 22K gene at least can be expressed from its own promoter as shown. Additional splicing events then place a particular ORF 5' most on the finished transcript for expression. Typically, every member of this late transcript group has three obligate exons at the beginning known as the tripartite leader. In addition, an occasional 4th exon is included containing an ORF known as the i-leader protein. In each subsequent part (b to e), the drawing from IGV viewer illustrates the structure of the dominant transcript that codes for each indicated ORF. The rectangles indicate exons joined by lines indicating introns. In each case, the dominant transcript for the S glycoprotein of SARS-CoV-2 is indicated in red. Transcripts that map to the top strand of the virus genome are in black and those that map to the bottom strand (and are reversed in orientation) are in blue. Parts b, c and d cover the whole 35,539bp of the ChAdOx1 nCoV-19 genome, and part e covers the 4323 bp from the human adenovirus type 5 integrated into the genome of HEK293. Part a illustrates the dominant transcripts for the ORFs identified as being expressed from the ChAdOx1 nCoV-19 genome in MRC5 cells at any of the three time points with parts B and c showing the same data from A549 and 293 cells respectively. Part d illustrates the dominant transcript observed that codes for each of the ORFs coded by the Ad5 region integrated into HEK293 cells.

Figure 2

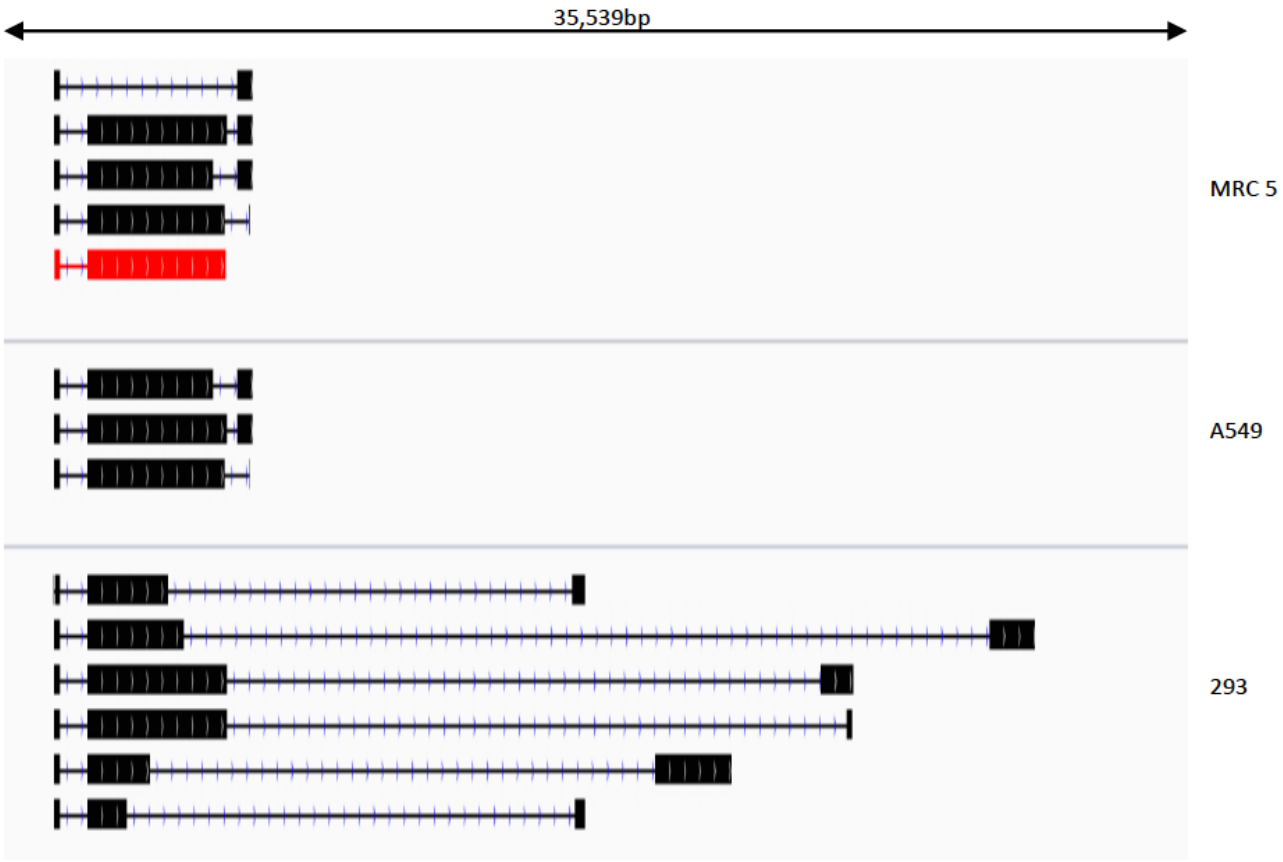

Figure 2

Transcript heterogeneity. This figure illustrates the structure of various transcripts observed at very low levels (less than 10 transcripts per dataset) all of which initiate at the transcription start site that drives expression of the S glycoprotein. Three sets of examples are shown from MRC5, A549 and HEK293 cells, in each case the splicing events depicted arise from use of the canonical GU-AG splice donor/acceptor pair that is used in over 90% of eukaryotic splicing events. For orientation, the structure of the dominant S glycoprotein transcript is shown in red in the examples selected from the MRC5 dataset.

Figure 3

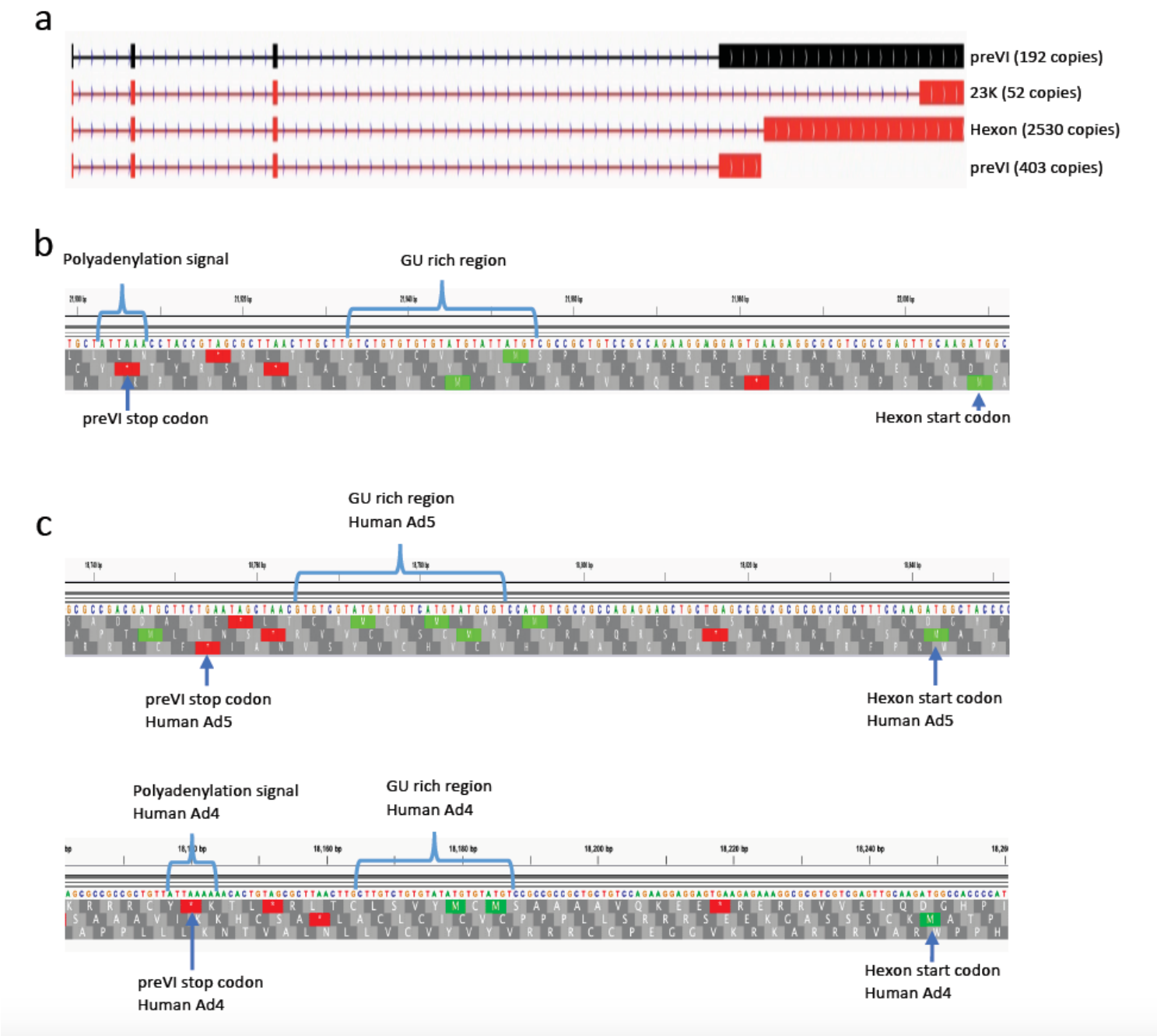

Figure 3

Novel polyadenylation site usage for preVI. Part a illustrates the transcript structure of transcripts coding for three proteins classically considered to be L3 proteins viz: preVI, hexon and 23K proteinase. The three vertical boxes on the left represent the tripartite exons normally included in all major late transcripts generated by mammalian adenoviruses. The L3 transcripts would classically share the same L3 polyadenylation site. In this diagram, the dominant transcripts seen in ChAdOx1 nCoV-19 infected 293 cells for each ORF are coloured red. This figure includes, in black, the transcript structure for observed

transcripts that would both code for preVI and fit the canonical transcript structure for an L3 transcript. In the case of 293 cells however, there are only 192 copies of this transcript compared to over 400 copies of the novel preVI transcript. Part b focusses on the sequences at the proposed polyadenylation site on ChAdOx1 showing the location of the polyadenylation signal and the GU rich region that is often present downstream of a polyadenylation signal. Part c shows the equivalent regions in the genome of human adenoviruses 5 (a group C adenovirus) and human adenovirus type 4 which is, like ChAdOx1, a group E adenovirus.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1chadox293features.txt
- Additionalfile2Hsapiens.GRCh38.cdna.fasta
- Additionalfile3proteinssearchlist.fasta
- Additionalfile4chadoxMRC5characterisedtranscripts.xlsx
- Additionalfile5chadoxA549characterisedtranscripts.xlsx
- Additionalfile6chadox293characterisedtranscripts.xlsx
- Additionalfile7A549Chadox1Totalproteinsidentified.xlsx
- Additionalfile8A549Chadox1PhosphoProteins.xlsx
- Additionalfile9MRC5Chadox1Totalproteinsidentified.xlsx
- Additionalfile10MRC5Chadox1PhosphoProteins.xlsx
- Additionalfile11A549Chadox1TotalPSMsidentified.xlsx
- Additionalfile12A549Chadox1PhosphopeptidesPSMs.xlsx
- Additionalfile13MRC5Chadox1TotalPSMsidentified.xlsx
- Additionalfile14MRC5Chadox1PhosphopeptidesPSMs.xlsx
- Additionalfile15A549transcriptsandproteomics.xlsx
- Additionalfile16MRC5transcriptsandproteomics.xlsx
- Tables.docx