RESEARCH ARTICLE

Regulatory patterns of differentially expressed genes in Ebola and related viruses are critical for viral screening and diagnosis
[version 1; peer review: peer review discontinued]

Daniel Achinko¹ 1-4, Anton Dormer¹ 2,4, Mahesh Narayanan¹ 2, Elton Norman¹ 2, Muneer Abbas 4

¹PepVax, Inc, 10411 Motor City Drive, Bethesda, MD, 20817, USA
²Immage Biotherapeutics Corp, University City Science Center, 3711 Market St., Philadelphia, PA, 19104, USA
³Current affiliation: Computational Biology Branch, NCBI/NLM/NIH, 8600 Rockville Pike, Bethesda, MD, 20894, USA
⁴Current affiliation: National Human Genome Center (NHGC), Cancer Center Building, Howard University, 2041 Georgia Avenue, Washington, D.C., 20060, USA

Abstract

Background
Viral detection techniques and applications are a critical first step to pathogen detection within a given population, especially during outbreaks. Common viral tests currently used are direct specimen examination, indirect examination and serological tests. Serological tests have gained intense interest because they are rapidly performed with patient blood samples for quick diagnosis and treatment. The diagnostic techniques developed around serology are often expensive, require expertise to use and cannot be afforded by developing countries with recurrent viral outbreaks. Therefore exploiting the huge amount of viral data available in various databases is critical to develop affordable and easy-to-use diagnostic tools.

Methods
This study obtained viral sample data from Gene Expression Omnibus database with focus on use of viral glycoprotein for host penetration. Gene relative mean across 34 obtained viral samples were extracted into data tables and used with edgeR statistical software in R version 3.3.1.

Results
Three clusters previously known to be LCK specific (Ebola virus relative viral cluster, EBOVC), CD209 specific (Mean differentiation cluster, MDC) and both LCK and CD209 specific (Kurtosis group cluster, KGC), expressed unique patterns of four proteins of interest (CD209, LCK, IL-2 and MYB). Differential expression analysis showed two cluster patterns on heatmaps, with differentially expressed proteins down-regulated in MDC but up-regulated in KGC and EBOVC for all pairwise cluster comparative analyses performed. Heatmaps showed two distinct immune related patterns, identifying MDC as B-lymphotropic while KGC and EBOVC as T-lymphotropic. Identified pathways were dominantly involved with homeostasis of immune cells and viral cell surface receptors involved in protein kinase activities.

Conclusions
Regulatory proteomic variants identified in clusters suggest transcription repression of HLA class I alleles. This study identified viral expression patterns with screening and therapeutic applications. Given that the viral pathogenetic pathway for Ebola has not been clearly identified yet, assembling its components is vital for vaccine development.

**Keywords**

virus, diagnosis, ebola, immune cells, CD209 receptor, LCK

This article is included in the Disease Outbreaks gateway.

---

**Corresponding author:** Daniel Achinko (achdany@yahoo.co.uk)

**Competing interests:** No competing interests were disclosed.

**Grant information:** The author(s) declared that no grants were involved in supporting this work.

**Copyright:** © 2017 Achinko D et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

**How to cite this article:** Achinko D, Dormer A, Narayanan M et al. Regulatory patterns of differentially expressed genes in Ebola and related viruses are critical for viral screening and diagnosis [version 1; peer review: peer review discontinued] F1000Research 2017, 6:275 (https://doi.org/10.12688/f1000research.10597.1)

**First published:** 15 Mar 2017, 6:275 (https://doi.org/10.12688/f1000research.10597.1)
Introduction
Viruses aim to interrupt and manipulate normal functions of cells, qualifying them as biological agents (Eckert, 1967). As years go by, human viruses are increasingly being studied and recorded, with examples like the Ebolavirus (Sanchez et al., 2016), human immunodeficiency virus (HIV) and the human herpesvirus (HHV) (Zell et al., 2008). Some viruses are considered emerging with epidemic manifestations, and their pathogenic effects have reshaped human response to viral epidemics and their detection methodologies. In the space of two to three years, epidemics resulting from Ebola virus and Zika virus caused the world health organization (WHO) to put in place necessary international public health emergency rules to better deal with such viral threats (Rojek et al., 2016). Adequate preparation for such threats engages both community and health care systems, with better surveillance and response time and better diagnostic development pipelines for vaccines and therapies. WHO recommends that devices used for diagnosis should be sensitive, specific, affordable, equipment-free, robust, rapid and easily delivered, especially in developing countries. Several detection techniques used in the last decade, like PCR and microarrays, have seriously improved upon infectious virus identification (Lau et al., 2010). Given that the safety and quality of blood related tools remains a public health concern, these products require routine screening, good quality control and good manufacturing practices (Albertoni et al., 2014). Development of better serological tests have drastically decreased transmission risk of human immunodeficiency virus (HIV), hepatitis C virus (HCV) and hepatitis B virus (HBV) (Tabor & Epstein, 2002). Antigen detection as a viral quantification method has been demonstrated for gag p24 antigen (Dong et al., 2012) and genome identification (Cotten et al., 2014) but sometimes, the absence of a good dynamic viral detection range results in limitations of this technique. Biological understanding of viral pathogenesis is also critical for developing better detection tools. Several viruses interact with host cell proteins to access and control cell machinery. Viral interference with host regulatory networks affects normal cell processes like gene expression and cell growth (Chatr-aryamontri et al., 2009). Documented information has shown that viruses exploit host phosphorylation mechanisms for their replication and in turn manipulate transcription mechanisms for pathogenic gains (Schang et al., 2002). Due to challenges encountered with current technologies like mass spectrometry, in silico tools are used both in the development of targeted therapies for viral infection through derived antigenic epitopes and for identification of viral phosphorylation sites (Xue et al., 2010). Current genomic viral detection techniques use DNA microarray technology, which is a collection of single-stranded DNA spots attached to a solid surface (Clinical and Laboratory Standards Institute, 2005). This technique has helped detect HIV mutations associated with drug resistance (Kozal et al., 1996), after which several studies have used chips to detect HCV (Xu et al., 2005), CNS infecting viruses (Leveque et al., 2011) and respiratory viruses (Coiras et al., 2005). Previous Ebolavirus and related virus microarray data from the Gene Expression Omnibus (Achinko et al., 2016) identified 3 main viral clusters and 4 main proteins (CD209, LCK, IL-2 and MYB), which play critical roles in host immune activation after viral attacks. Their structural analysis and differential expression analysis between clusters shows viral regulatory patterns of the host genome that are critical to pathogenesis and can be exploited for screening novel viruses and developing novel therapeutic targets.

Methods
Data extraction
Data obtained from the Gene Expression Omnibus profile database as described (Achinko et al., 2016) was used to extract total Ebola virus (EOBV) gene related data (n = 10,296) and virus related data (n = 34) using a unique code (Achinko, 2017). Relative mean (RMEAN) count data was normalized (RCDN) (Achinko et al., 2016) (Dataset 1, Achinko et al., 2017a), and another set was not normalized and considered raw data (RCRD) (Dataset 1, Achinko et al., 2017a). Both sets of data were entered into separate table of values, which were used in R (v.3.3.1) statistical software (R Development Core Team, 2008; RRID SCR_001905) for further analysis.

Structural analysis and domain identification
Sequence variants for four main genes of interest; Lck (protein tyrosine kinase), CD209 (Dendritic Cell-Specific Intercellular adhesion molecule 3 [ICAM-3]-Grabbing Nonintegrin [DC-SIGN]), MYB (MYB proto-oncogene) and IL-2 (Interleukin-2), were obtained from the National Center for Biotechnology Information database (https://www.ncbi.nlm.nih.gov/nuccore). Sequences were imported into Geneious Software platform, version 9.1.6 (Kearse et al., 2012, RRID:SCR_010519). Sequences were aligned using Geneious alignment algorithm with parameters Blosum62 for cost matrix, gap open penalty of 12, gap extension penalty of 3 and refining iterations of 2. Domains predicted for these sequences targeted the PRINTS database (RRID:SCR_003412), PROSITE_PATTERNS database (RRID:SCR_003457), SIGNALP_EUK (Peterson et al., 2010) and PFAM database (RRID:SCR_004726). Figures for differential domain patterns on sequences were exported as png images for viewing.

Statistical analysis
The EBOV related RMEAN gene value (GV) tables, RCDN and RCRD, were processed and all genes with datasum=0 were removed from the table for further analysis. For RCDN, GVgs were summed across samples (GS, gene summed) and a minimal value of 0.001 was added to each gene value per sample (GVps) eliminating all zero value for GVps. The final normalization formula for GVps was Log2 ((GVps+0.001)/GS). Heatmaps consisting of all GVps-related EBOV samples and cluster-based samples (KGC [Kurtosis Group Cluster], MDC [Mean Differentiation Cluster] and EBOVC [Ebolavirus cluster]) for signaling and immune related genes, were developed as described in Achinko et al. (2016). The three main clusters, KGC, MDC and EBOVC served as basis for grouping GVps-related EBOV samples in heatmap, which was further used for differential expression analysis in edgeR (Empirical Analysis of Digital Gene Expression Data in R version 3.3.1) software program (Robinson et al., 2010, RRID:SCR_012802). ANOVA two-way factorial (R Core Team, 2013) was used as basis for pairwise comparative analysis between cluster samples. Given that edgeR software is designed to avoid analysis of data columns with sum = zero, a unique code (Achinko, 2017) was written
to round up GVps in RCRD data table to one decimal place and columns with gene-sum per sample = 0 were removed from the data, bringing total samples to n = 28 (Dataset 2, Achinko et al., 2017b), including sample_1, Sample_2 and sample_3 (KGC); sample_27, sample_28 and sample_29 (MDC); and sample_30, sample_31, sample_32, sample_33 and sample_34 (EBOVC), as described in Achinko et al. (2016). The RCRD table processed in edgeR was split into 3 groups: group 1 related to KGC, group 2 related to EBOVC and group 3 related to MDC. The data was further processed for library size (sequencing depth or gene counts) and typically genes with less than 2 fold expression were filtered out, retaining only genes represented in samples across groups.

Given that groups in our data didn’t reflect true biological replicates but gene profiles observed from the RCDN heatmap showed patterns which could be qualified as such, we analyzed our data using edgeR to include: Principal component analysis (PCA), Biological Coefficient of Variation (BCV), Differential Expression (DE) analyses and gene functional analyses. PCA uses the multidimensional scaling plot of distance between gene expression profiles (MDS). This is a two-dimensional scatter plot showing that distances between samples approximate to log2 fold change. The biological coefficient of variation (BCV) is the square root of the dispersion parameter considered using the negative binomial model. In our case, the algorithm to Estimate Empirical Bayes Tagwise Dispersion values was used to estimate dispersion. Tagwise dispersion uses Bayes empirical method, laying emphasis on a maximum likelihood conditionally weighted approach. Differential expression analysis was performed through genewise negative binomial generalized linear models using quasi-likelihood tests. This test fits a negative binomial generalized log-linear model to count data, then a gene wise statistical test was used to compare genes between groups. The false discovery rate (FDR) calculated for the P-values was based on Benjamini & Hochberg (1995) at $\alpha < 0.05$. Functional analysis was performed using the GO (Gene Ontology, RRID:SCR_006580) database with focus on the biological process, and KEGG (Kyoto Encyclopedia for Genes and Genomes, RRID:SCR_012773).

### Viewing sequence variants

Variants for all genes of interest (CD209, HLA-A, HLA-B, HLA-C, HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DPA2, HLA-DPB1, HLA-DPB2, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRA, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-DRB6, HLA-E, HLA-F-A51, HLA-F, HLA-G, HLA-H, HLA-J, HLA-K, LCK, MYB) were extracted into bed file format using a unique code (Achinko, 2017) and files were mounted into the Integrated Genomics Viewer software (RRID:SCR_011793) and compared with Human genome (hg38) sequences (https://www.ncbi.nlm.nih.gov/genome) and Gencode V24 (Harrow et al., 2012) sequences to view genomic locations and variations for comparative analysis. Genomic matches were exported as image files for viewing.

### Results

#### Cluster determination from extracted and normalized Gene Expression Omnibus data

Through the use of hierarchical clustering, RCDN data showed cluster patterns on the heatmap (Figure 1) following the three main and previously identified clusters, KGC, MDC and EBOVC (Achinko et al., 2016). KGC comprised 5 samples, MDC comprised 16 samples and EBOVC comprised 13 samples. Comparative statistical analysis with ANOVA between Sample 30 (EBOV), Sample1 and Sample 27 then Sample 27 and Sample 1, was significant between Sample 30 and Sample 1 (P-value = 0.0371; F statistic = 4.348, DF=1) Sample 30 and Sample 27 (P-value = 7.51e-05, F-statistic = 15.692, DF=1). No significant difference was seen between Sample 1 and Sample 27 (P-value = 0.3687, F statistic = 0.808). From our previous work (Achinko et al., 2016), heatmaps centric for HLA and signaling proteins showed a similar pattern of CD209 upregulated in KGC and MDC only (Figure 2) but not expressed in EBOVC. Meanwhile, the same couldn’t be said for LCK, MYB and IL-2 proteins, which showed differential expression patterns in the 3 clusters for this analysis, although they were previously observed to be up-regulated in KGC and EBOVC only but down-regulated in Sample 32 (EBOV). The focus on these four proteins (CD209, LCK, MYB and IL-2) was relative to their role in MHC class I and II gene activation. The expression patterns led to identification of variant types expressed by the three cluster groups and it was observed that, apart from CD209 variants (NM_021155 and AF290886) expressed only in KGC and MDC, LCK (NM_005356 and U07236), MYB (NM_005375 and AI357042) and IL-2 (NM_000586) variants were common for KGC and MDC, as opposed to LCK (U23852, M36881 and M26692), MYB (M13666, M15024 and U22376) and IL-2 (S82692 and X00695) variants specific to EBOVC (Figure 3).

For LCK protein variants, PROSITE_PATTERNS and PRINTS domains varied for EBOVC: PROTEIN_KINASE_ATP (PS00107) for AAC50287.1 variant and PROTEIN_KINASE_TYR (PS00109) and TYRKINASE (PR00109) for AAAS9502.1 variant. Both variants presented with SH3DOMAIN (PR00452) and SH2DOMAIN (PR00401), meanwhile KGC and MDC showed both PROSITE_PATTERNS on each protein variant (AAA18225.1 and NP_005347.3). All sequences showed similar PFAM domains and PRINTS domains, which were: SH3_1 (PF00018), SH2 (PF00017) and Pkinase_Tyr (PF07714). For MYB protein variants, PFAM domains varied for EBOVC: PROTEIN KINASE ATP (PS00107) for AAC50287.1 variant and PROTEIN_KINASE_TYR (PS00109) and TYRKINASE (PR00109) for AAAS9502.1 variant. Both variants presented with SH3DOMAIN (PR00452) and SH2DOMAIN (PR00401), meanwhile KGC and MDC showed both PROSITE_PATTERNS on each protein variant (AAA18225.1 and NP_005347.3). All sequences showed similar PFAM domains and PRINTS domains, which were: SH3_1 (PF00018), SH2 (PF00017) and Pkinase_Tyr (PF07714). For MYB protein variants, PFAM domains varied for EBOVC in only six sequences (AAC50232.1, AAB49039.1, AAB49037.1, AAB49035.1 and AAB49034.1), which presented with differential lengths of the Cmyb_C domain (PF09316) and one sequence had PRINTS domain (F139DOMAIN, PR02045). A fraction of PFAM domains (Myb_DNA-binding [PF000249] and LMSTEN [PF07988]) were constant for all EBOVC, KGC and MDC sequences. For IL-2 protein variants, SIGNALP_EUK domains (SignalP-noTM) were present only for EBOVC protein sequence variants (AAB46883.1 and CAA25292.1). All EBOVC sequence variants and NP_005772.2 specific to KGC and MDC, presented similar PROSITE_PATTERNS (INTERLEUKIN_2 [PS00424]), PRINTS domains (INTERLEUKIN2 [PR00265]) and
Figure 1. Heatmap of EBOV related clusters. The viral data for all collected samples was plotted and samples clustered following the three main viral clusters (KGC: Kurtosis Group cluster, MDC: Mean Differentiation cluster, EBOVC: Ebolavirus related cluster) as shown.
Figure 2. HLA centric heatmaps with four main proteins (CD209, LCK, IL-2, MYB). All HLA genes in our dataset were plotted onto a heat map following the three main viral clusters (KGC: Kurtosis Group cluster, MDC: Mean Differentiation cluster, EBOVC: Ebolavirus related cluster). CD209 expressed in KGC and EBOVC but not in MDC. The other proteins (LCK, IL-2 and MYB) showed differential expression patterns within groups.
Figure 3. Structural representation and regulatory patterns of genes of interest LCK, IL-2 and MYB. mRNA sequence variants for the genes of interest were obtained and focus was placed on domains from four databases: PRINTS database (RRID:SCR_003412), PROSITE_PATTERNS database (RRID:SCR_003457), SIGNALP_EUK and PFAM database (RRID:SCR_004726) as diagrammatically represented. (A) Four variants were obtained for LCK Tyrosin-Kinase. The first two sequences belonging to EBOVC showed differential expression patterns of PROSITE_PATTERNS for AAA59502.1 (PROTEIN_KINASE_ATP) and AAA59502.1 (PROTEIN_KINASE_TYR). The other two variants belonged to KGC and MDC and together with EBOVC expressed all the domains for each sequence. All of them never showed the SIGNALP_EUK domain. (B) For IL-2 expressed variants, only EBOVC related protein sequences (AAB46883.1) showed SIGNAL_P-noTM. KGC and MDC variants (NP_000577.2) showed no SignalP-noTM domain. The AAA59140.1 variant was common to all viral clusters and didn’t express the SignalP-noTM domain and the upstream portion of the PRINTS-INTERLEUKIN2 domain. (C) For MYB protein expressed variants, the first 8 variants belonged to EBOVC (AAA52031.1, AAA52032.1, AA49039.1, AA49038.1, AA49037.1, AA49036.1, AA49035.1, AA49034.1). AA49031.1 AAB49038.1, AAB49038.1 sequences showed variable lengths, but lacking the down-stream PFAM-Cmyb_C domain. MYB sequences identified mainly with PFAM domains, and together with the KGC and MDC sequences, they all showed the three identified MYB domains (Myb_DNA-binding, LMSTEN and Cmyb_C) meanwhile only AAB49034.1 variant showed a PRINT-F138DOMAIN further downstream. Yellow=SIGNAL_EUK domain; gold=PRINTS domain; purple=PFAM domain and black=PROSITE_PATTERNS domain.
PFAM domains (IL-2 [PF00715]), except for AAA59140.1 variant common to all viral clusters, which lacked the upstream sequence of the PRINTS domain.

Sample variation analysis
Distance analysis for variation between sample expression profiles using a two dimensional PCA plot (Figure 4A) with log of fold change (logFC) on both axes showed three clusters, with KGC and EBOVC related samples dominantly clustering on the negative X-axis (leading logFC dim 1) and EBOVC located to the bottom left. Clustering patterns observed in the plot suggest common expression patterns in identified viral clusters. MDC dominantly clustered to the positive X and Y axes. EBOV never really grouped in any cluster though aligned with KGC and EBOVC. The estimated coefficient of variation was 8.5958 with a BCV of 2.93, which is much higher than a BCV of 0.4 for a standard controlled experiment. The trend of dispersion fit analysis under the quasi-likelihood pipeline shows an increase in dispersion as the log2CPM increases and the squeezed values show a good fit along the trend (Figure 4B).

Differential expression analyses
The three viral clusters (KGC, MDC and EBOVC) were used in pairs for comparative analyses and variation was considered significant at 5% false discovery rate. The plots were performed for log-fold changes (logFC) against log-counts per million (logCPM) and the blue lines define limits for logFC = ±2, where differentially expressed proteins were highlighted as red point dots. In all three clusters the majority of proteins were downregulated ≤ 5-fold and a few up-regulated ≥ 10-fold (Figure 5A). Pairwise comparative analysis showed a similar pattern for differentially expressed proteins observed in all three clusters when MDC was compared to KGC and EBOVC but when EBOVC was compared to KGC, there were very few differentially expressed genes down-regulated ≤ 10-fold. Differential expression analysis showed that more proteins showed a > ±2 fold change when MDC was individually compared to KGC (Figure 5B) and EBOVC (Figure 5C), while fewer proteins were significantly down regulated when KGC was compared to EBOVC (Figure 5D). This could be confirmed on the heatmaps plotted on estimated tagwise values for the first 100 proteins in each comparative category. Mapping the top 100 differentially expressed proteins for DE on heatmaps showed a very clear split pattern for MDC against KGC (Figure 5E) and EBOVC (Figure 5F), and it could be observed from the pattern that MDC samples clustered together with their proteins down-regulated (green) while KGC and EBOVC samples clustered with up-regulated proteins (red). For EBOVC compared to KGC (Figure 5G), the pattern was a little bit different with MDC samples not completely separated from KGC and EBOVC because some proteins were both up and down-regulated in different samples. This pattern further shows the similarity in EBOVC and KGC and how they differ from MDC.

Pathway analysis for differentially expressed clusters
Pathway analysis was performed on two databases, GO and KEGG. GO identified 20855 terms in total with the first 20 (Table 1) focusing on cell activation and positive and negative regulation of the immune system. Total GO data can be seen in Dataset 3 (Achinko et al., 2017c). KEGG identified 311 pathways with the first 20 (Table 2) involved in phosphate metabolism (glycogenesis and gluconeogenesis), ATP synthesis, fatty acid metabolism, steroid hormone metabolism and DNA metabolism. Total KEGG data can be seen in Dataset 4 (Achinko et al., 2017d). This pattern showed that though the top 20 differentially expressed genes varied as seen in their respective tables, most of them were involved in similar pathways.

Gene variant analysis
Bed files for each of the four genes of interest (CD209, LCK, IL-2, MYB) were extracted. There were two CD209 protein variants with two locations on chromosome 19, but the other genes had protein variants located in related and defined chromosomal positions. For all the HLA class I immune isoforms identified (Figure 6A), they all had sequence variants mapping to HLA-A and these consolidated variants were considered for further genomic analysis (Figure 6B). Focusing on this consolidated region, male sex chromosome (Y) specific protein variants were identified, which belonged to the tetratricopeptide repeat protein (ubiquitously transcribed tetrameric peptide repeat containing, Y-linked [UTY]) and is known to be involved in protein-protein interactions (www.ncbi.nlm.nih.gov/gene/74702). This HLA-A isoform also contained the full length of the UTY gene (Figure 6C) located on chromosome Y and shown to have histone demethylase properties, catalyzing the demethylation of trimethylated Lys-27 (H3K27me3) in histone H3 (Walport et al., 2014). H3K27 is known to be involved with gene regulation, where promoters enriched with K27Me2 and K27Me3 are repressed (Martin et al., 2005). Several variations were observed for each of the HLA class II immune isoforms (Figure 6D). HLA class I and II genes were flanked by several others of interest: RNA-U1 small nuclear 61 (RNU1-61P) pseudogene located between HLA-DRB5 and HLA-DRB6 (chr6:32549940-32550090), microRNA (HRM) 3135B located between HLA-DQA2 and HLA-DQB2 (chr6:32749912-32749979), TAP2 transporter 2, ATP binding cassette subfamily B member in fusion with HLA-DOB (chr6:32821833-32550090), proteasome subunit beta 8 and 9 (PSMB8 and PSMB9) immediately after TAP2 (chr6:32840717-32844935; chr6:32854161-32859851), protein phosphatase 1 regulatory inhibitor subunit 2 pseudogene 1 (PPP1R2P1) after PSMB8 and 9 (chr6:32876475-32880074), NR_037177 (HLA-Z) known as pseudo-HLA class I HLA gene isoform after PPP1R2P1 (chr6:32896402-32896489), Bromodomain containing 2 (BRD2), situated between HLA-DM and HLA-DOA (chr6:32968660-32981505), COL11A2P1 considered pseudo-collagen type 2 gene 1 between HLA-DPA2 and HLA-DPB2 (chr6:33103693-33107144) and COL11A2 protein-coding collagen type XI alpha 2 chain.

Dataset 1. Relative mean (RMEAN) count data normalized (RCDN) table
http://dx.doi.org/10.5256/f1000research.10597.d153382
This data represents relative mean values collected for all 10274 genes identified. EB_Sample: Ebolavirus related sample.
**Figure 4. Principal component analysis of viral sample variation.** (A) There is a distinct pattern of the three viral clusters (KGC: Kurtosis Group Cluster, MDC: Mean Differentiation Cluster, EBOVC: Ebola Virus related Cluster) on the 2D plain. On the positive X and Y-axis is the MDC cluster. To the bottom left is the EBOVC cluster and the top left is the KGC cluster. Ebola virus (EB_Sample_30) was distinct as it grouped with KGC on the left but separated out of the sub groups identified. (B) The trend of dispersion fit analysis shows an increase in dispersion as the log2 count per million (log2CPM) increases and the squeezed values show a good fit along the trend. EB_Sample = Ebola virus related virus sample.

**Dataset 2. edgeR processed data for differential expression analyses**

http://dx.doi.org/10.5256/f1000research.10597.d153383

The data is organized into the following clusters: Kurtosis Group cluster (KGC, rows = 1–5), Mean Differentiation cluster (MDC, rows = 6–16) and Ebolavirus related cluster (EBOVC, rows = 17–28). The table contains 10274 genes identified. EB_Sample: Ebolavirus related sample.

**Dataset 3. Total Gene Ontology pathways identified for differential expression analysis**

http://dx.doi.org/10.5256/f1000research.10597.d153384

The total data consists of 20855 genes. Processes were ranked in order of most to least differentially expressed based on P-value.

**Dataset 4. Total KEGG pathways identified for differential expression analysis**

http://dx.doi.org/10.5256/f1000research.10597.d153385

311 identified pathways were ranked from most important to least important based on P-value.

**Discussion**

Viral genomics is an area of intense research, given that viruses could be regional and specific to the local human host. Viruses like Ebola, Rift Valley fever and Zika have resulted in sudden epidemics in West Africa, East Africa and South America, respectively. This work aimed to look at gene expression patterns of EBOV and related viral species as described in Achinko et al. (2016). Previous work done by Achinko et al. (2016) identified four main genes...
A  DE genes across all viral groups

B  DE genes for MDC vs KGC
Figure 5. Differential expression analyses of related genes in viral clusters. The three viral clusters (KGC, MDC and EBOVC) were used in pairs for comparative analyses. (A) Differential expression of all viral clusters (KGC: Kurtosis Group Cluster, MDC: Mean Differentiation Cluster, EBOVC: Ebola Virus related Cluster) showed that majority of genes were down-regulated ≤ 5-fold and a few up-regulated ≥ 10-fold. Pairwise comparative analysis between (B) MDC and KGC and (C) MDC and EBOVC showed a similar pattern as the one seen in the general comparative analysis between clusters. (D) Pairwise comparative analysis between EBOVC and KGC showed very few DE genes down-regulated ≤ 10-fold. Heatmaps with hierarchical clustering for the top 100 differentially expressed genes identified for MDC and KGC (E) and MDC and EBOVC (F) showed a clear distinct grouping of MDC that was different to KGC and EBOVC, which also grouped together. (G) The heatmap for the top 100 differentially expressed genes identified in EBOVC and KGC showed a pattern which wasn’t distinct for MDC when compared to KGC and EBOVC.
Immune related functions dominate the process. Total data can be viewed in Dataset 2 (Achinko et al., 2017b).

<table>
<thead>
<tr>
<th>Term</th>
<th>biological_process</th>
<th>Ont</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0001775</td>
<td>cell activation</td>
<td>BP</td>
<td>980</td>
</tr>
<tr>
<td>GO:0001867</td>
<td>complement activation, lectin pathway</td>
<td>BP</td>
<td>9</td>
</tr>
<tr>
<td>GO:0001868</td>
<td>regulation of complement activation, lectin pathway</td>
<td>BP</td>
<td>2</td>
</tr>
<tr>
<td>GO:0001869</td>
<td>negative regulation of complement activation, lectin pathway</td>
<td>BP</td>
<td>2</td>
</tr>
<tr>
<td>GO:0002252</td>
<td>immune effector process</td>
<td>BP</td>
<td>797</td>
</tr>
<tr>
<td>GO:0002253</td>
<td>activation of immune response</td>
<td>BP</td>
<td>545</td>
</tr>
<tr>
<td>GO:0002376</td>
<td>immune system process</td>
<td>BP</td>
<td>2651</td>
</tr>
<tr>
<td>GO:0002526</td>
<td>acute inflammatory response</td>
<td>BP</td>
<td>139</td>
</tr>
<tr>
<td>GO:0002576</td>
<td>platelet degranulation</td>
<td>BP</td>
<td>87</td>
</tr>
<tr>
<td>GO:0002673</td>
<td>regulation of acute inflammatory response</td>
<td>BP</td>
<td>70</td>
</tr>
<tr>
<td>GO:0002682</td>
<td>regulation of immune system process</td>
<td>BP</td>
<td>1556</td>
</tr>
<tr>
<td>GO:0002683</td>
<td>negative regulation of immune system process</td>
<td>BP</td>
<td>375</td>
</tr>
<tr>
<td>GO:0002684</td>
<td>positive regulation of immune system process</td>
<td>BP</td>
<td>950</td>
</tr>
<tr>
<td>GO:0002697</td>
<td>regulation of immune effector process</td>
<td>BP</td>
<td>418</td>
</tr>
<tr>
<td>GO:0002698</td>
<td>negative regulation of immune effector process</td>
<td>BP</td>
<td>101</td>
</tr>
<tr>
<td>GO:0002920</td>
<td>regulation of humoral immune response</td>
<td>BP</td>
<td>50</td>
</tr>
<tr>
<td>GO:0002921</td>
<td>negative regulation of humoral immune response</td>
<td>BP</td>
<td>13</td>
</tr>
<tr>
<td>GO:0006508</td>
<td>proteolysis</td>
<td>BP</td>
<td>1599</td>
</tr>
<tr>
<td>GO:0006810</td>
<td>transport</td>
<td>BP</td>
<td>4483</td>
</tr>
</tbody>
</table>

Heatmaps centric to HLA class I and II variants showed similar expression patterns to previous heatmaps (Achinko et al., 2016), where HLA class I isoforms were down-regulated across clusters and CD209 was not expressed for EBOVC but slight variations in expression patterns were observed for LCK, IL-2 and MYB proteins. The expressed variant types identified for each of the four proteins of interest (CD209, LCK, IL-2 and MYB), showed that KGC and MDC expressed unique protein variant types different from those for EBOVC. This observation could confirm the statistically significant difference in expression pattern of EBOVC from KGC and MDC. The observed variants for each protein showed differential domain expression for EBOVC but maintained all domains per variant per protein type for KGC and MDC. This pattern for EBOVC suggests differential regulation due to differential usage of promoter regions on related genes through alternative splicing processes, leading to protein variants with misplaced...
Figure 6. Chromosomal location of immune variants related to viral pathogenesis. (A) Variant distribution of all HLA class I variants identified across all viral samples used. (B) Region on HLA-A common to the majority of HLA class I variants. (C) Ubiquitously transcribed tetratricopeptide repeat containing, Y-linked (UTY) gene identified with HLA class I gene variants with specificity to HLA-A. (D) Distribution of all variants observed for each of the HLA class II immune isoforms identified across all viral samples.
cellular functions (Landry et al., 2003; Pajas et al., 2007). Furthermore, differential usage of EBOV distal promoter regions has been shown to produce non-functional LCK mRNA variants (Takadera et al., 1989), and differential expression of MYB protein variants has been shown to occur at gene promoter regions (O’Rourke & Ness, 2008). Though the expression of these functional protein variants in KGC and MDC, it has been shown that the Epstein-Barr virus (EBV) nuclear antigen 1 (EBNA1) protein (Sample 28) possesses a large Gly-Ala repeat domain affecting proteosomal degradation of EBNA1 polypeptide and preventing HLA I antigenic peptide formation, which is needed to induce CD8+ immune T-cells (Levitskaya et al., 1997). It has also been shown that EBV G protein-coupled receptor BILFI1 engages the HLA-I/β2-microglobulin complex, favoring rapid loss of HLA-I molecules from the cell surface and intense reduction of HLA-I half-life (Zuo et al., 2009). HIV-1 nef protein (Sample 2) either induces degradation of MHC class I complexes in endosomes and lysosomes through RING-finger motifs or targets the degradation of MHC-I complexes through di-leucine motifs (Rowe et al., 2010) common to CD209 protein. EBV has also been shown to interact epigenetically with host genes through histone modifications and promoter hypermethylation, resulting in gene expression variations and tumorigenesis (Tsai et al., 2006). The mechanism of function in KGC and MDC clusters shows a different form of regulation from EBOVC, suggesting that genetic mechanisms underlying the former lead to gene repression or expression while in the latter, it leads to alternative splicing with aim to favor virulence in host cells.

The coefficient of biological variation depicted by PCA was 8.5958, greater than that in controlled biological replication experiments, but the plot grouped EBOVC and KGC to the negative x-axis, centralizing Ebola virus although its related cluster (EBOVC) samples dominantly located to the bottom while MDC samples located to the positive x-axis. This pattern suggests that EBOV could be more heterogenous than the other viral samples and given that it uses sGP to infect its host, there could be different genomic pathways involved in its pathogenesis. The high BCV suggests an uncontrolled biological replication experiment, but the observed expression patterns (Figure 4A) show that individual viral samples were very similar and could be used as replication data.

Differential expression analysis performed on all three clusters showed that the majority of proteins were down-regulated, and those of interest belonged to: human retina cDNA samples, potassium voltage-gated channel interacting protein 4 (KCNIP4), a mucin protein (MUC8) highly expressed in human airways (Shankar et al., 1994), a nucleoporin 36 (nup36) of the nuclear pore complex (NPC) needed for mitotic progression and chromosome segregation (Zuccolo et al., 2007), a suppressor of tumorigenicity protein (ST20), eukaryotic translation initiation factor 5A (EIF5A) whose activation by Kirsten ras oncogene (K-Ras) favors pancreatic ductal adenocarcinoma migration (Fujimura et al., 2015), nitric oxide synthetase (NO), which requires L-arginine and molecular oxygen to synthesize nitric oxide and acts as a neurotransmitter (Bloch et al., 1995), natriuretic peptide A (NPPA) involved in regulation of inflammation by altering cytokine secretion and macrophage function (Mezzasoma et al., 2016), non-coding microRNA 664B known to promote proliferation of human osteosarcoma (Chen et al., 2015) and a basic leucine zipper (MAFF), known to be a proto-oncogene and transcription factor that interacts with the promoter of oxytocin receptor gene. The identification of gene variants associated with human retina cDNA samples could be linked to sample 24 (Munoz-Erazo et al., 2012), which is part of the EBOVC cluster. This sample showed that West Nile Virus (WNV) has various forms of ocular involvement (chorioretinitis, uveitis, and occlusive retinal vasculitis), of which chorioretinitis is associated with retinal pigment epithelium degeneration. In response to WNV infection, interferon-β signaling is induced (Cinatl et al., 2006) through Toll-like receptors.

From all the differentially expressed proteins discussed, we can identify those that directly involve viral interaction, those that are involved with pathways that successfully transmit the virus and those that result in oncogenesis. The general differential expression pattern was mapped onto pairwise comparative analyses of viral clusters for which EBOVC and KGC had very few proteins down regulated. Though the top differentially expressed proteins including the human retina cDNA protein variants and von Willebrand factor (VWF) involved with homeostasis, were down-regulated, some proteins were up-regulated which included but not limited to: wnt family member 1 (WNT1), shown to increase the expression of a scavenger receptor B (CD36) located on endothelial cells, dendrittic cells and macrophages (Matsumoto et al., 2000); and ves-tigial like family member 4 (VLL4), known to act as tumor suppressor in lung cancer by down regulating yes-associated protein 1 (YES)-transcription factor interacting domain (YAP-TEAD) (Zhang et al., 2014). YES protein has also been identified in an interaction pathway with LCK (Achinko et al., 2015). Wolframin (WFS1) is a transmembrane protein mainly located in the endoplasmic reticulum and known to be involved in the regulation of Ca2+ homeostasis (Takei et al., 2006). WAP four-disulfide core domain 8 (WFDC8) functions as a serine protease inhibitor through its Kunitz-inhibitor domain and also plays a role in immune homeostasis (Ferreira et al., 2013). WASP homolog-associated protein, associated with actin, membranes and microtubules, is known to help promote nucleation, which in turn stimulates actin polymerization of actin related protein (Arp2/3 complex) at the golgi and tubular membranes (Campellone et al., 2008). WASP is known to favor proper coordinated transport of viral particles through vesicles into the cell. The up-regulated proteins observed from differential expression analysis of EBOVC and KGC suggest activation of proteins that favor membrane receptors and vesicular transport of viruses into host cells. Comparative analysis between MDC and KGC showed a similar pattern of protein expression as seen in the general expression pattern comparing all three clusters discussed above. Interestingly, when compared to EBOVC, MDC showed a differential expression pattern where the retina cDNA protein variants and VWF were up-regulated and WASP was down-regulated, as opposed to the differential expression pattern seen when comparing EBOVC to KGC where VWF was down-regulated and WASP up-regulated. Down-regulation of VWF suggests a reduced cellular response to viral entry, hence supporting the idea that EBOV uses sGP to penetrate host cells, whilst
up-regulation of VWF suggests an immune response of infected cells after viral pathogenesis. Down-regulation of WASP causes inactivation of vesicles needed to transport MHC class I antigenic epitopes out of the cell (van Endert et al., 2002) and hence causes a downregulation of immune response.

The heatmaps targeting proteins from each differential expression comparative viral cluster analysis presented with a distinct pattern in which proteins were down-regulated in MDC compared to KGC and EBOVC, however this distinct pattern did not show up during comparative viral cluster analysis of differentially expressed proteins from EBOVC and KGC. Previous work on viral clusters MDC, KGC and EBOVC (Achinko et al., 2016) showed that MDC related viral samples were B-lymphotropic as they reflected viral pathogenic expression patterns related to EBV (Schlee et al., 2004). KGC and EBOVC showed patterns related to HIV and EBOV, respectively. HIV and EBOV have been documented to infect dendritic cells, which leads them to direct infection of T lymphocytes, especially CD4+ T cells in HIV-1 (Sanchez et al., 2006). KGC and EBOVC could relate to T cell infection specificity. MDC represents B-lymphotropic viral samples, which use the CD209 receptor route on dendritic cells to penetrate their host whilst KGC and EBOVC represent T-lymphotropic viral samples that are not specific to CD209 receptor as an entry route.

GO identified pathways showed that proteins like VWF, involved in regulation of complement activation and the lectin pathway (GO:0001867, GO:0001869) were of priority to MDC but not EBOVC and KGC. Likewise, KEGG identified pathways showed that genes like WASP protein, involved in glycolysis and ATP synthesis to transport vesicles around the cell, were of priority in KGC and EBOVC but not MDC. These cellular pathways and glycolysis related functions further support a coordinated regulatory pattern in MDC to repress transcription of genes but the same genes are up-regulated for KGC related viral pathogenesis and alternatively spliced for EBOVC related viral pathogenesis. The immune expression patterns observed through bed graphs showed HLA class I and II alleles were identified for each viral infection type and these variants differed from one sample to the other. It was observed that all viruses generated HLA class I alleles with variants cutting across HLA-A, which suggests a common region of interest to be targeted when designing viral antigenic epitopes in silico, with aim to develop a cut across antiviral therapy. The ubiquitously transcribed tetratripeptide repeat containing, Y-linked (UTY) protein is a male-specific demethylase catalizing the demethylation of lysine-27 (Lys27) in histone 3 (H3K27me3) (Walport et al., 2013). A variant of UTY is located on chromosome 6 and has several amino acid repeats in the protein, suggesting a possible translocation of the gene. Its epigenetic role has been shown to non-catalytically regulate transcription by interacting with T-box proteins (Miller et al., 2010). Some KEGG pathways that UTY has been shown to be prominently involved in are: Lysine biosynthesis (path:hsa00300) and Lysine degradation (path:hsa00310). This suggests UTY protein can have an effect on expression of Lysine-containing signaling proteins. With this regulatory pattern in mind, it suggests a possible mechanism of regulation for proteins of interest like: tyrosine-protein kinases (LCK) at the promoter regions with related processes like Tyrosine metabolism (path:hsa00350) and Glycolysis/Gluconeogenesis (path:hsa00010), CD209 with processes like Valine, leucine and isoleucine degradation (path:hsa00280) and Valine, leucine and isoleucine biosynthesis (path:hsa00290), MYB with processes like Purine metabolism (path:hsa00230) and Pyrimidine metabolism (path:hsa00240) and IL-2 with process like Citrate cycle (TCA cycle) (path:hsa00220), TAP2 was found in fusion with the HLA class II gene (HLA-DOB) and was previously shown to be down-regulated in MDC (Achinko et al., 2016). Given that TAP2 is actively involved in HLA class I peptide antigenic processing, transcription of TAP2 in MDC could be repressed. Major HLA class I alleles (HLA-A, HLA-B and HLA-C) were down-regulated in all clusters, but more so in the EBOVC cluster, which suggests important roles of other possible proteins like LCK, which together with IL-2 and MYB were all down-regulated in this cluster (Achinko et al., 2016). Contrasting the role of these proteins in EBOVC and KGC can help to further explain regulatory patterns on host genomes critical for the viral expression patterns observed. The role of proteins with regulatory function on HLA class II chromosomal genes locations needs to be further investigated. Their interplay with HLA class I and II immune genes and how they influence viral pathogenesis will aid in the identification of genetic targets for development of antiviral therapy and vaccine designs.

**Conclusion**

This study was undertaken to further investigate the role of four proteins (CD209, LCK, IL-2 and MYB) previously identified to be critical for viral pathogenesis in related clusters (KGC, MDC and EBOVC) (Achinko et al., 2016). Structural variant analysis showed that variants for these four proteins were unique and specific to EBOVC or KGC and MDC. The domain regulatory pattern suggested differential splicing for EBOVC and gene expression or repression for KGC and MDC, which could be directed by epigenetic programs or differential promoter regulatory patterns on the chromosome. Differential expression analyses suggested that the proteins that vary greatly during viral infection are involved with immune homeostasis and regulation of host cell immune receptors, required by viruses to penetrate the host cells. The heatmaps for these differential expression patterns showed that viral-immune cell interaction patterns were B-lymphotropic for MDC and T-lymphotropic for KGC and EBOVC. Protein expression regulatory patterns observed also suggested an immune regulatory pattern with genes like UTY, which process pattern recognition receptors needed to activate CD4+ and CD8+ T-cells, and specific amino acids on proteins related to proper identification of HLA class I and II peptides by antigen processing cells (APC). Specific differential expression patterns identified in this work set a basis for developing tools necessary for viral screening during epidemic outbreaks and development of antiviral immune therapies to properly stimulate immune expression.
Data and software availability
Latest source code: https://github.com/brack123/Viral_Cluster_Processing-scripts.v1/tree/V1.0.0

Archived source code as at the time of publication: https://doi.org/10.5281/zenodo.226630 (Achinko, 2017)

License: GNU GPL

Dataset 1: Relative mean (RMEAN) count data normalized (RCDN) table. This data represents relative mean values collected for all 10274 genes identified. EB_Sample: Ebolavirus related sample.
DOI, 10.5256/f1000research.10597.d153382 (Achinko et al., 2017a)

Dataset 2: edgeR processed data for differential expression analyses. The data is organized into the following clusters: Kurtosis Group cluster (KGC, rows = 1–5), Mean Differentiation cluster (MDC, rows = 6–16) and Ebolavirus related cluster (EBOVC, rows = 17–28). The table contains 10274 genes identified. EB_Sample: Ebolavirus related sample.
DOI, 10.5256/f1000research.10597.d153383 (Achinko et al., 2017b)

Dataset 3: Total Gene Ontology pathways identified for differential expression analysis. The total data consists of 20855 genes. Processes were ranked in order of most to least differentially expressed based on P-value.
DOI, 10.5256/f1000research.10597.d153384 (Achinko et al., 2017c)

Dataset 4: Total KEGG pathways identified for differential expression analysis. 311 identified pathways were ranked from most important to least important based on P-value.
DOI, 10.5256/f1000research.10597.d153385 (Achinko et al., 2017d)

Author contributions
All authors actively engaged and contributed to the successful production of this manuscript. D.A.A conceived and designed the work, collected the data and did the analysis required to put the manuscript together. D.A.A, D.A.O, M.A, M.N, E.F.N and M.A did the proof reading and edited the manuscript.

Competing interests
No competing interests were disclosed.

Grant information
The author(s) declared that no grants were involved in supporting this work.

References


The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com