Inferring *bona fide* differentially expressed genes and their variants associated with vitamin K deficiency using systems genetics approach

Shalini Rajagopal[1,4,5], Akanksha Sharma[8], Anita Simlot[2], Praveen Mathur[2], Sumita Mehta[7], Sudhir Mehta[2], Jalaja Naravula[5], Krishna Mohan Medicherla[6], Anil Kumar S[5], Uma Kanga[8], Renuka Suravajhala[1], Ramji Bhandari[9], PB Kavi Kishor[5], Bipin G Nair[1] and Prashanth Suravajhala[1,4]

1. Amrita school of Biotechnology, Amrita Vishwa Vidyapeetham, Clappana PO 690525, Kerala, India
2. Department of Medicine, SMS Medical College, JLN Marg, Jaipur 302004, RJ, India
3. Department of Pediatric Surgery, SMS Medical College, JLN Marg, Jaipur, India
4. Bioclues.org, India
5. Department of Biotechnology, Vignan's Foundation for Science, Technology & Research, Vadlamudi, Guntur 522213, AP, India
6. Department of Biotechnology and Bioinformatics, Birla Institute of scientific Research, Statue circle, Jaipur 302001, RJ, India
7. Department of Gynecology and Obstetrics, Babu Jagjivan Ram Memorial Hospital, Delhi, India
8. Department of Transplant Immunology, AIIMS, Delhi, India
9. Department of Biology, University of North Carolina at Greensboro, NC 27412, USA.

Correspondence: prash@am.amrita.edu

Abstract

Systems genetics is key for integrating a large number of variants associated with diseases. Vitamin K (VK) is one of the scarcely studied conditions in lieu of ascertaining either the differentially expressed genes (DEGs) or variants in an individual subpopulation of diseased phenotypes associated with VK, *viz.* myocardial infarction, renal failure, prostate cancer, thrombosis, thrombocytopenia, coagulation related diseases to name a few. In this work, we have screened characteristic DEGs common to three VK-related diseases, *viz. m*yocardial infarction, renal failure and prostate cancer and asked whether or not any DEGs in addition to pathogenic variants are common to these conditions. We attempt to bridge the gap in finding characteristic biomarkers and discuss the role of long noncoding RNAs (lncRNAs) in the biogenesis of VK deficiencies.

Keywords: RNA-Seq, Vitamin K, Comorbidities, Differential Expressed Genes, Variant analysis

1. Introduction

The next generation sequencing (NGS) technologies have paved the way systems genomics is heralded [26]. As NGS has provided scope to understand novel biological mechanisms and molecular underpinning of complex diseases, a thorough genomic and transcriptome analyses are needed [6]. With RNA-seq investigating the dynamic nature of the cell's transcriptome, the component of the genome that is actively translated into RNA molecules, allows researchers to predict when and where genes are turned on or off in a range of cell types/situations. As the number of biological samples investigated using RNA-seq analysis expands, the community has developed a wide range of bioinformatics tools to meet specific demands with highly optimized pipelines for further downstream processing [27]. The advantages of analyzing transcriptome data can be a multitude, for example, finding genomic features such as gene and transcript expression, miRNAs, non-coding RNAs (long non-coding RNAs, and small RNAs) besides predicting variants or mutations in the form of novel isoforms and SNPs (SNVs or indels) with sufficiently high expression levels.

In the recent past, blood disorders have been well studied using RNA-Seq, for example [11] studied the potential gene expression difference in acute respiratory distress syndrome (ARDS) using hematopoietic stem cell transplantation which implies differences in immune response and interferon signaling pathways. Zheng et al recently provided a genome-wide analysis comparing the RNA-seq data from haemorrhoidal diseases [41]. Of late, single-cell transcriptomic strategies have just begun to be understood taking varied diseased phenotypes [22,29] to mention a few. However, vitamin K (VK) deficiency is one of the scarcely studied phenotypes wherein genome-wide transcriptome profiling heralded for understanding the expression profiles are not known. There are, indeed, associated phenotypes such as thrombosis, thrombocytopenia, myocardial infarction, renal failure, CA prostate which are used to ascertain the differential expressed gene (DEG) profiles [39]. What remains intriguing is that calling the somatic or germline variants after the DEG profiles are checked could be a holistic measure for screening and characterizing biomarkers. Our study attempts to fill this gap wherein we have used myocardial, CA prostate (both from sequence read archive) renal (in house) datasets and checked for the DEG profiles and candidate mutations associated with VK deficiency. We discuss the impending effects of the role of DEG profiles in the context of VK and associated deficiencies.

2. Materials and Methods

Datasets

The myocardial data is a strand-specific RNA-Seq dataset for both coding and non-coding RNA profiling from 28 hypertrophic cardiomyopathy (HCM) patients and 9 healthy donors [25]. For our analysis, we have considered three control and three treatment samples from this study through sequence read archive (SRA: ncbi.nlm.nih.gov/sra last accessed on June 9, 2022). The following myocardial samples were taken: SRR8586402; SRR8586407; SRR8586429 (Control) and SRR8586409; SRR8586423; SRR8586431 (Treated). In addition, we used data from the complete transcriptome landscape of prostate cancer (PCa) using RNA-seq from a study [30] :ERR031017; ERR031029; ERR031031 (Control) and ERR031018; ERR031030; ERR031032 (Treated) and also compared with our own PCa datasets from our lab (PRJNA616165). Finally, the renal datasets are divided into three groups: rejection time point, well functioning rejection matched, and post therapy (PRJNA854340). There were four patients in the test group (R2, R4, R5, and R6), all of whom received post-therapy (R2_pt, R4_pt, R5_pt, R6_pt), and three patients in the control group who were in good health (WF2, WF4 and WF6). All samples were conducted on paired-end datasets and were supplemented with three pairs of these datasets (Supplementary Table 1).

2.1 RNA Sequencing analysis, statistics and validation

The reads were checked for quality using FastQC [1] followed by HISAT2 [20,21] to align to the human genome (GRCh38 assembly). Cufflinks-Cuffdiff pipeline was employed to yield significant changes at the level of transcript expression, splicing, and promoters [36], which was later benchmarked in our lab and used to run through the workflow [4]. As Cufflinks treats each pair of fragment reads as a single alignment, there is always an optimal amount of time and energy saved. With an 'overlap graph,' each the largest sets of reads originating from the same isoform result in a minimal set of fragments and this determines transcript abundances using a statistical model [8] (Figure.1).

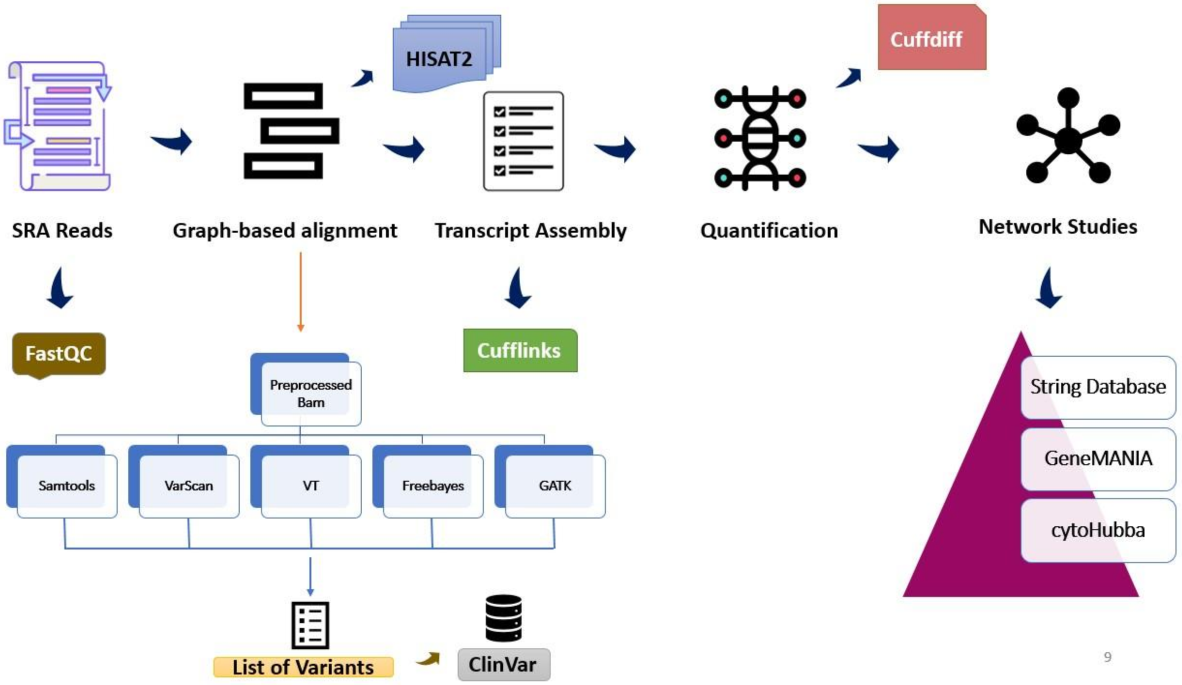

Figure 1: Workflow of the RNA-Seq pipeline. The pipeline for differential gene expression analysis of RNAseq data. In the variant calling step, five different approaches, *viz.* samtools. varscan, vt. freebayes and GATK were utilized to identify significant SNPs.

RNA-Seq reads were trimmed, but there was no significant reduction in size; further aligned reads were processed to generate SAM, BAM and sorted BAM files through a cohort of tools. As the DEG analysis primarily relies on paired samples, we checked paired end reads, i.e. control vs. treated in case of the myocardial, prostate and renal datasets. The resulting tables were filtered (after ensuring cufflinks pipeline was used with -g option to check for novel isoforms) by p and q value <=0.05, and >=2 log2 fold change <=2. The output BAM files were subjected to consensus mapping of SNPs with different tools such as samtools [23,24], varscan [18], freebayes [14], vt [35] and GATK [32] using the default parameters such as a number of criteria, including coverage, read counts, p-value, variant allele frequency, base quality, and the number of strands on which the variant was observed [19]. The filtered variants were then compared with the ClinVar database. Freebayes is a haplotype-based and Bayesian genetic variant detector which calls variants based on the reads aligned to a target, not necessarily with their precise alignment. It can find SNPs, indels and multi-nucleotide polymorphisms(MNPs) besides complex events such as composite insertion and substitution events [14]. To check this, we used a myriad of variant calling tools to screen and reach consensus, for example, vt was used to identify short variations in NGS data [35], varscan on the other hand approaches variant detection by aligning the map to multiple locations even as it screens the unique mapped reads for substitutions, indels thereby detecting multiple reads and converting them into unique SNPs/indels besides determining the total number of reads supporting each allele (reference and variant).

2.2 Clustering coefficient network analysis using cytoHubba

The proteins showing significant changes interacting between all three different datasets were used to build a gene interaction network using STRING-db and later visualized using String [38] and Cytoscape v 3.9.1 [9]. The network was checked for top ranking genes built using the expression correlation plugin with a 0.95 correlation as cut off value. The cluster ID for proteins showing similar abundance values as defined in the hierarchical clustering were used to annotate the network to reveal relationships between the different protein groups. Finally, the network analyzer/cytoHubba was used to define the network measures [3]. Lower contrast (faded yellow/orange) means the rank is lower and bigger the contrast (Red/Maroon) indicates greater is the rank, and we further evaluated the efficacy of network genes for betweenness, closeness, clustering coefficients, and stress centrality for the top 10 DEGs. A commonmost lncRNA and a variant of uncertain significance (VOUS) were checked for the diseased phenotypes.

3. Results and Discussion

Significant DEGs associated with vivid blood disorders were commonly identified across myocardial, renal and prostate

A significant number of DEGs were found common in our analysis. Among them, Apolipoprotein D (APOD ENSG00000189058) is up-regulated in the first set "Sc5-LV and HCM515" and down-regulated in the other "ND2 and HCM273" was found common between the control and treated datasets. It encodes a component of high-density lipoprotein and shares a lot of similarities with plasma retinol-binding protein [10]. This glycoprotein has associations with the lipoprotein acyltransferase enzyme lecithin:cholesterol acyltransferase. Breast cyst and androgen insensitivity syndrome (AIS) are two diseases linked to APOD with cholesterol and sphingolipids transport / recycle to plasma membranes in the lung (normal and CF), and transport of glucose and other sugars, bile salts and organic acids, metal ions, and amine compounds are other associated pathways. CD163 (ENSG00000177575) is downregulated in both myocardial datasets, and is a member of the SRCR superfamily of scavenger receptors known to be associated with monocytes and macrophages. The gene serves as an acute phase-regulated receptor that helps macrophages remove hemoglobin/haptoglobin complexes and endocytose them, potentially protecting tissues from free hemoglobin-mediated oxidative damage. The protein encoding gene was shown to act as a bacterial innate immune sensor and a local inflammatory inducer [31]. Furthermore, it is associated with the multisystem inflammatory syndrome and histiocytic sarcoma in children. Binding and uptake of ligands by scavenger receptors and hematopoietic stem cells (HSCs) and lineage-specific markers are two linked mechanisms with it. Interestingly, we found a processed pseudogene (ENSG00000274295) that is associated with polymerase (DNA Directed), epsilon 2, and accessory subunit (POLE2) in these datasets (Figure 2).

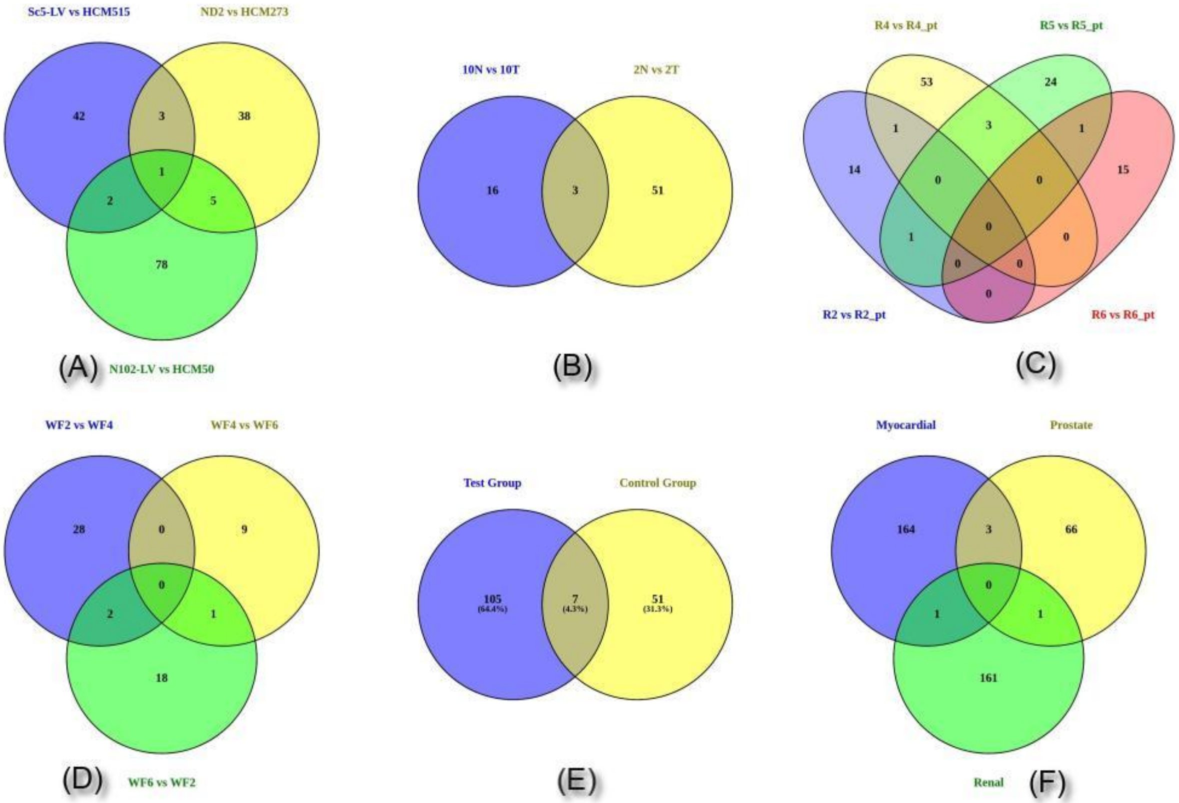

Figure 2 : Differential expressed genes between control and treated in all three datasets, *viz.* (A) myocardial, (B)prostate and (C) test group of renal, (D) control group of renal (E) control and test

group of renal data and (F) common to all the three myocardial, renal and prostate.

Two genes are common between the Sc5-LV/HCM515 and ND2/HCM273; COL1A1, which codes for collagen type I alpha 1 chain (ENSG00000108821) is downregulated in the first dataset and upregulated in the third set; The pro-alpha1 chains of type I collagen, which have two alpha 1 chains and one alpha2 chain, are encoded by this gene. Type I collagen forms fibrils and is found in most connective tissues, including bone, cornea, dermis, and tendon. Osteogenesis imperfecta types I-IV, Ehlers-Danlos syndrome, classical type VIIA,, Caffey disease, and idiopathic osteoporosis are all linked to mutations in this gene. Reciprocal translocations between chromosomes 17 and 22, genes for platelet-derived growth factor beta are located, linked to dermatofibrosarcoma protuberans, a type of skin tumor caused by uncontrolled growth factor expression. Two of its transcripts have been found as a result of the application of alternative polyadenylation signals. Binding and uptake of ligands by scavenger receptors, as well as VEGFR3 signaling in lymphatic endothelium are two linked processes. We also found a significant number of downregulated DEGs in the form of fibroblast growth factor 12 (FGF12; ENSG00000114279) which are associated with activation of apoptotic and synovial fibroblasts pathways regulating a number of biological processes, including embryonic development, cell growth, morphogenesis, tissue repair, tumor growth, and invasion, and have extensive mitogenic and cell survival functions. Although it lacks the N-terminal signal sequence seen in the majority of FGF family members, it does include clusters of basic residues that have been shown to behave as a nuclear localization signal. This protein accumulated in the nucleus but was not secreted when transfected into mammalian cells. What remains interesting is that CPNE5 (ENSG00000124772), which encodes a calcium-dependent protein, is downregulated in all three datasets of myocardial function. It harbors an integrin A domain-like sequence in the C-terminus and may regulate molecular events at the interface of the cell membrane and cytoplasm, and is shown to have several alternatively spliced transcript variants encoding isoforms (see supplementary information) .

The comparison between prostate data, three genes are common. AMACR (ENSG00000242110); PCAT14 ((ENSG00000280623) prostate cancer associated transcript 14) and LTF (Lactotransferrin). AMACR (Alpha-Methylacyl-CoA Racemase) is upregulated in the first two datasets of CA prostate only while the latter dataset didn’t yield any significant DEGs. In addition, various transcript variants with alternative splicing have been identified, for example

C1QTNF3 (C1q and tumor necrosis factor-related protein 3) known to cause bile acid synthesis defect, congenital, 4 and alpha-Methylacyl-Coa racemase deficiency (Supplementary Table 3 and Supplementary information). As LTF (ENSG00000012223) is downregulated in both sets, it is largely associated with cellular growth and differentiation regulation, cancer formation and metastasis. It has been recently discovered to have activity against both DNA and RNA viruses, including SARS-CoV-2 and HIV [28]. The two common genes of OLFM4 (ENSG00000102837) and MMP8 (ENSG00000118113) were downregulated in the renal WF2-WF4 and upregulated in the WF6-WF2. OLFM4 is a novel prognostic predictor as well as therapeutic target for hepatocellular carcinoma [2]. One common lncRNA of OVCH1-AS1 (ENSG00000257599) is upregulated in the WF4-WF6 and downregulated in the WF6-WF2 (Supplementary Table 3 and supplementary information).

3.1 NONHSAT106693 among significantly enriched genes

One of the favorite candidate DEGs often sought are lncRNAs and what remains compelling is the list of lncRNAs that were upregulated and downregulated across the three datasets. Among them, ENSG00000260604 (lncRNA) is upregulated in both datasets whereas ENSG00000276980 (sense intronic complement component 3: C3) and ENSG00000287891 are downregulated. ENSG00000260604 is 1,357 nucleotides long (GeneCards, Ensembl, LNCipedia, and Ensembl/GENCODE) and is a well annotated candidate with the sense intronic C3 sequence forming a product of the genes ENSG00000276980.1, ENSG00000276980, and lnc-GPR108-3 [34]. Whereas C3 (ENSG00000125730) and LINC02208 (ENSG00000250891) are downregulated in the latter two datasets, it helps activate the complement system and the encoded preproprotein is proteolytically processed [17]. Mutations related to this gene are linked to atypical hemolytic uremic syndrome and age-related macular degeneration. The deficiency leads to autosomal recessive and hemolytic uremic syndrome and is widely associated with immune responses, besides Lectin induced complement pathway and peptide ligand-binding receptors. Furthermore, LINC02208 is expressed in tissue samples of the heart [12] even as ENSG00000287891 also identified as a novel lncRNA is downregulated in the latter two myocardial datasets.

Interestingly, a novel lncRNA (ENSG00000285534) is downregulated in both sets of R2-R2_pt and R4-R4_pt while CXCL8, a C-X-C Motif Chemokine Ligand 8 (ENSG00000169429) associated with melanoma and bronchiolitis is downregulated in the R5-R5_pt and up-regulated in the R6-R6_pt; They aid in Immune response CCR3 signaling in eosinophils and cytokine

doi:10.20944/preprints202209.0362.v1

signaling in the and produces a protein that belongs to the CXC chemokine family (encoded by IL-8), a key mediator of the inflammatory response. Mononuclear macrophages, neutrophils, eosinophils, T lymphocytes, epithelial cells, and fibroblasts all produce IL-8 and acts as a chemotactic factor, directing neutrophils to the infection site. In addition to participating in the proinflammatory signaling cascade with other cytokines, it may be likely that the overproduction of such proinflammatory proteins are assumed to be the source of the cystic fibrosis-related lung inflammation which may contribute to coronary artery disease and endothelial dysfunction. Tumor cells release this protein, which promotes tumor motility, invasion, angiogenesis, and metastasis. This chemokine also has angiogenic properties. Higher levels of IL-8 are positively connected with increased severity of numerous illness outcomes, and IL-8 binding to one of its receptors (IL-8RB/CXCR2) enhances blood vessel permeability (eg, sepsis). On the other hand, LEF1 (ENSG00000138795) is upregulated in both datasets (R5 and R6) and three other genes, *viz.* G0S2 (ENSG00000123689), HSD11B1 (ENSG00000227591), PTGS2 (ENSG00000073756) were commonly downregulated in R4 and R5 sets. G0S2 (G0/G1 Switch 2) is a protein coding gene, located in mitochondria involved in extrinsic apoptotic signaling pathway and plays a positive regulation of extrinsic apoptotic signaling pathway regulating Van Der Woude syndrome. HSD11B1 (ENSG00000227591); HSD11B1-AS1 (HSD11B1 Antisense RNA 1) is a lncRNA and is associated with the cortisone reductase deficiency. Besides this, the genes POSTN (ENSG00000133110); EPDR1 (ENSG00000086289); SFRP4 (ENSG00000106483) were upregulated in third set of myocardial and up-regulated in second set of prostate data. Periostin is a secreted protein that induces cell attachment and spreading and plays a role in cell adhesion and its differential expression is known to regulate T2-high asthma, myocardial infarction regulating heparin binding and cell adhesion molecule binding [15]. LTF (ENSG00000012223) is downregulated in the second set of prostate and treated renal data (WF2-WF4) with a novel lncRNA NONHSAT106693 (ENSG00000287903) shown to be upregulated in the third set of myocardial and upregulated in the treated renal data.

3.2 Heatmaps showed distinct gene expression profiles with a variance in principal components

When dealing with multivariate data, it is frequently presented as a matrix in high-throughput experiments for ascertaining gene expression even as specific genetic pathways are of considerable interest. These visualizations help aid in biological prognosis and so we have attempted to identify transcriptome expression by comparing control and treated datasets to check their variance (Supplementary figures 1 and 2). Overall, we obtained 392 DEGs of which

172 genes were upregulated and 220 genes were downregulated (Supplementary Table 3). In myocardial datasets, 10 genes were upregulated in comparing the first datasets (Sc5-LV and HCM515), 6 genes were upregulated and 28 genes were downregulated in the second dataset and 28 genes were upregulated and 24 genes were downregulated in the third datasets (N102_LV and HCM506). Among prostate, 14 genes were upregulated in the first datasets (10N and 10T), 34 genes were upregulated in the second datasets (2N and 2T) and there are no significant DEGs found in the third datasets (3N and 3T). In comparing renal datasets, test group and post therapy were analyzed of which 9 genes were upregulated and 31 genes were downregulated in the first datasets (R2 and R2_pt); 7 genes were upregulated and 30 genes were downregulated in the second datasets (R4 and R4_pt); 13 genes were upregulated and 21 were downregulated in the third datasets (R5 and R5_pt); 19 genes were upregulated and no genes were downregulated in the fourth datasets (R6 and R6_pt). The results of control group includes 9 genes were upregulated in the first datasets (WF2 and WF4); 4 genes were upregulated and 11 genes were downregulated in the second datasets (WF4 and WF6); 19 genes were upregulated and 1 gene downregulated in the third datasets (WF6 and WF2).

3.3 CytoHubba yielded top niche ranks with variant analysis showing no mutations attributed to DEGs

We sought to ask whether the common DEGs form top niche ranks from all three datasets. To check this, we imported the network of DEGs (Figure 3A) into cytoscape and visualized all other networks such as closeness, betweenness, stress, and clustering genes in cytoHubba. The network with the top 10 genes yielded a rank list indicating vivid top niche genes associated through clustering coefficient. While the color indicates the score of the genes interacting in the network, we found that among the top ranking genes, 4 DEGs are known to be associated with VK deficiency (Figure 3B). COL1A2 (Alpha-2 type I collagen) is one of the profibrotic genes that express osteocalcin in the liver [16] and is also used as a marker of cardiac fibrosis [7]. POSTN (Periostin) is one of the VK dependent proteins, which is majorly involved in hematopoiesis [37], myocardial infarction, fibrosis, and bone [40]. SFRP4 (Secreted Frizzled Related Protein 4) is involved in bone mineral density which is related to osteoporosis [5]. EPDR1 (Ependymin Related 1) gene produces a type II transmembrane protein that is related to the protocadherins and ependymoma, two families of cell adhesion molecules. This protein may have a role in calcium-dependent cell adhesion, according to gene expression studies in brain tissue [33]. LTF (Lactotransferrin) has been found to have an effect on host immunological responses and has a potential antagonistic pleiotropy, suggesting that it may be protective against caries besides predisposing to localized aggressive periodontitis [13]. The above comorbidities are related and hence these 4 genes were considered the hub genes associated with VK deficiency. In this proposition, we determined the extent of lncRNAs in the network and as a result we could plot the lncRNA top ranking DEGs as well (Figure 3C; supplementary Table 4). Taken together, we found NONHSAT106693 to be a novel lncRNA with a large expression in testis indicating that it could be associated with all the three, *viz.* VK, renal and PCa (FPKM/TPM: 0.14; Figure 3D). We have reconfirmed the expression in vivid samples of VK in house (data shown).

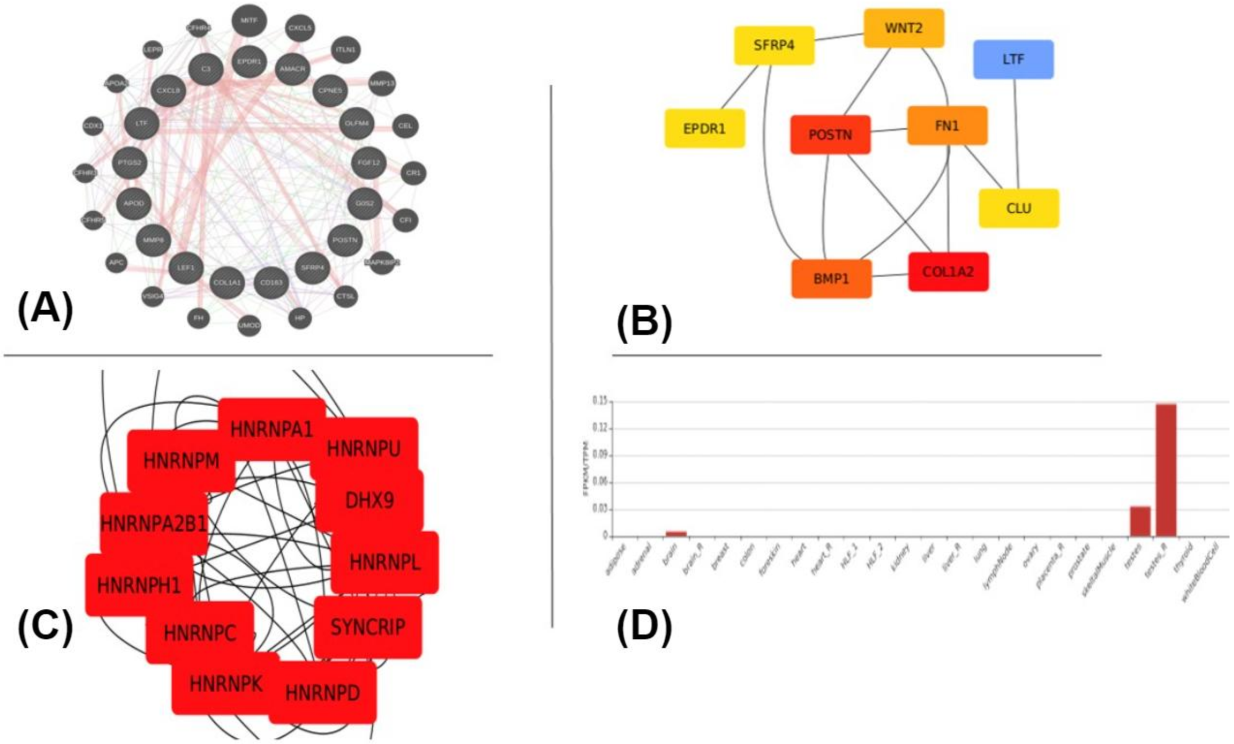

Figure 3: (A) The network illustrates the interaction between the differential expressed genes in all three data sets (myocardial, prostate and renal) and it was constructed with Genemania. (B) The network image represents the clustering coefficient of common DE genes from all three datasets. The image was generated with the cytoHubba. (C) The top ranks of lncRNAs in clustering coefficient networks. (D). The graph shows the tissue expression of common novel lncRNA and it is generated with the noncode database.

4. Discussion

Variants of unknown significance

doi:10.20944/preprints202209.0362.v1

The variants called from these DEGs were further compared with five different tools such as varscan, samtools, freebayes, vt and GATK. To check whether any DEGs harbor the pathogenic variants, we compared ClinVar pathogenic variants of VK with the list of significant number of variants (Supplementary Table 2; Supplementary figure 3). Among them, 37 variants were found to have a match with the ClinVar data, associated with CFTR, ESR1, GGCX, ATP8B1, VWF, GLA, and F8. While CFTR (chr 7) has been afflicted in both myocardial (rs397508397) and prostate control (rs75789129) samples from samtools and vt, ESR1 was shown to be seen in both renal - control and treated samples, and the F8 was seen only in myocardial treated samples. On the other hand, rs563109158 (T>C) found in the gene GGCX, an extremely rare variant of uncertain significance is found to be common to all the three sets implying that this is predisposed in an ostensibly large population (C=0.000318/6 (ALFA)/ C=0./0 (TWINSUK)/C=0.000223/1 (Estonian)/C=0.000404/107 (TOPMED)/C=0.000407/57 (GnomAD)/C=0.000519/2 (ALSPAC)/C=0.001002/1 (GoNL)). This was further considered as a candidate for the classification of disease prevalence and penetrance estimates and was therefore classified as a variant of unknown significance (VOUS). Taken together, none of the DEGs seem to be commonly enriched from our prostate RNA-Seq datasets screened in-house (data unpublished). This VOUS, was common particularly in the control sets (Sc5-LV; ND2) and treated set (HCM506) in myocardial; control set of 2T and treated set of 3T in prostate samples; test group of R4, R5, R6 and post therapy of R4_pt, R5_pt; control group of WF2 in Renal datasets) which indicates that it is poised across diffident phenotypes. The interpretation was reported as the uncertain significance and the variant condition was identified as VK-dependent clotting factors, combined deficiency of type 1 with no citations found in ClinVar, and therefore we have validated using Sanger sequencing from our PCa/VK cohort (data not shown). The other genes ATP8B1, GGCX, GLA, and VWF were identified in all three control and treated samples as the identified results were taken for further analysis of molecular docking and simulation studies (data not shown).

5. Conclusions

Vitamin K (VK) plays an important role in human metabolism. In this work, we investigated whether any common DEGs were significantly enriched among vivid datasets and if so, whether or not the variants in them are noteworthy to VK diseased phenotypes,*viz.* myocardial, renal and prostate cancer. While we found a large number of lncRNAs among the DEGs, NONHSAT106693 was found to be significantly enriched lncRNA across renal and myocardial implying that they play an important role in atypical hemolytic uremic syndrome and age-related macular degeneration. Our work also emphasizes on the role of variants of unknown significance (VOUS) in these phenotypes and especially the common variant, *viz.* rs563109158 seen in GGCX which is associated with VK-dependent clotting factors. There is room for analyzing more datasets associated with VK, coagulation and blood disorders which would set a precedent in screening pathogenic and perhaps unique variants/VOUS as downstream analysis and development of NGS panels for rare blood disorders and VK deficiencies are on the anvil.

Authors' contributions: PS conceived the project. SR and AS wrote the first draft. All the other authors chipped in with lateral sections. PS and PBK proofread the manuscript before all authors agreed to it.

Funding: PS gratefully acknowledge funding from SERB-DST (#IRR/2018/000025)

Data availability: The PCa datasets are deposited at sequence read archive (SRA) with project id: PRJNA616165 and the renal datasets with PRJNA854340.

Acknowledgements: The authors thank Nidhi Shukla for subsidiary works on PCa analysis.

Competing interests: None

References

1. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
2. Ashizawa, Y., Kuboki, S., Nojima, H., Yoshitomi, H., Furukawa, K., Takayashiki, T., Takano, S., Miyazaki, M., & Ohtsuka, M. (2019). OLFM4 Enhances STAT3 Activation and Promotes Tumor Progression by Inhibiting GRIM19 Expression in Human Hepatocellular Carcinoma. Hepatology Communications, 3(7), 954–970.
3. Assenov Y, Ramírez F, Schelhorn SE, Lengauer T, Albrecht M. Computing topological parameters of biological networks. Bioinformatics. 2008 Jan 15;24(2):282-4. doi: 10.1093/bioinformatics/btm554. Epub 2007 Nov 15. PMID: 18006545.
4. Ayam Gupta, Shalin Rajagopal, Sonal Gupta, Ashwani Kumar Mishra, Prashanth Suravajhala. (2021). A bioinformatics pipeline for processing and analysis of whole transcriptome sequence data. dx.doi.org/10.17504/protocols.io.brz8m79w

5. Boudin, E., Fijalkowski, I., Piters, E., & Van Hul, W. (2013). The role of extracellular modulators of canonical Wnt signaling in bone metabolism and diseases. Seminars in Arthritis and Rheumatism, 43(2), 220–240. https://doi.org/10.1016/j.semarthrit.2013.01.004
6. Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1842(10), 1932–1941. https://doi.org/10.1016/j.bbadis.2014.06.015
7. Delbeck, M., Nickel, K. F., Perzborn, E., Ellinghaus, P., Strassburger, J., Kast, R., Laux, V., Schäfer, S., Schermuly, R. T., & von Degenfeld, G. (2011). A role for coagulation factor Xa in experimental pulmonary arterial hypertension. Cardiovascular Research, 92(1), 159–168. https://doi.org/10.1093/cvr/cvr168
8. Dilworth, R. P. (1950). A Decomposition Theorem for Partially Ordered Sets. The Annals of Mathematics, 51(1), 161.
9. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. J Proteome Res. 2019 Feb 1;18(2):623-632. doi: 10.1021/acs.jproteome.8b00702. Epub 2018 Dec 5. PMID: 30450911; PMCID: PMC6800166.
10. Drayna, D. T., McLean, J. W., Wion, K. L., Trent, J. M., Drabkin, H. A., & Lawn, R. M. (1987). Human apolipoprotein D gene: Gene sequence, chromosome localization, and homology to the alpha 2u-globulin superfamily. DNA (Mary Ann Liebert, Inc.), 6(3), 199–204. https://doi.org/10.1089/dna.1987.6.199
11. Englert, J. A., Cho, M. H., Lamb, A. E., Shumyatcher, M., Barragan-Bradford, D., Basil, M. C., Higuera, A., Isabelle, C., Vera, M. P., Dieffenbach, P. B., Fredenburgh, L. E., Kang, J. B., Bhatt, A. S., Antin, J. H., Ho, V. T., Soiffer, R. J., Howrylak, J. A., Himes, B. E., & Baron, R. M. (2019). Whole blood RNA sequencing reveals a unique transcriptomic profile in patients with ARDS following hematopoietic stem cell transplantation. Respiratory Research, 20(1), 15. https://doi.org/10.1186/s12931-019-0981-6
12. Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjöstedt, E., Lundberg, E., Szigyarto, C. A.-K., Skogs, M., Takanen, J. O., Berling, H., Tegel, H., Mulder, J., … Uhlén, M. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Molecular & Cellular Proteomics: MCP, 13(2), 397–406. https://doi.org/10.1074/mcp.M113.035600
13. Fine, D. H. (2015). Lactoferrin: A Roadmap to the Borderland between Caries and Periodontal Disease. Journal of Dental Research, 94(6), 768–776. https://doi.org/10.1177/0022034515577413
14. Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. ArXiv:1207.3907 [q-Bio]. http://arxiv.org/abs/1207.3907
15. Gillan, L., Matei, D., Fishman, D. A., Gerbin, C. S., Karlan, B. Y., & Chang, D. D. (2002). Periostin secreted by epithelial ovarian carcinoma is a ligand for alpha(V)beta(3) and alpha(V)beta(5) integrins and promotes cell motility. Cancer Research, 62(18), 5358–5364.
16. Gupte, A. A., Sabek, O. M., Fraga, D., Minze, L. J., Nishimoto, S. K., Liu, J. Z., Afshar, S., Gaber, L., Lyon, C. J., Gaber, A. O., & Hsueh, W. A. (2014). Osteocalcin Protects

Against Nonalcoholic Steatohepatitis in a Mouse Model of Metabolic Syndrome. Endocrinology, 155(12), 4697–4705. https://doi.org/10.1210/en.2014-1430

17. Herbert, A. (2020). Complement controls the immune synapse and tumors control complement. Journal for Immunotherapy of Cancer, 8(2), e001712. https://doi.org/10.1136/jitc-2020-001712
18. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009 Sep 1;25(17):2283-5. doi: 10.1093/bioinformatics/btp373. Epub 2009 Jun 19. PMID: 19542151; PMCID: PMC2734323.
19. Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research, 22(3), 568–576. https://doi.org/10.1101/gr.129684.111
20. Kim, D., Paggi, J.M., Park, C. *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
21. Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* 2015
22. Li, F., Yan, K., Wu, L., Zheng, Z., Du, Y., Liu, Z., Zhao, L., Li, W., Sheng, Y., Ren, L., Tang, C., & Zhu, L. (2021). Single-cell RNA-seq reveals cellular heterogeneity of mouse carotid artery under disturbed flow. Cell Death Discovery, 7(1), 180.
23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PMID: 19505943; PMCID: PMC2723002.
24. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93. doi: 10.1093/bioinformatics/btr509. Epub 2011 Sep 8. PMID: 21903627; PMCID: PMC3198575.
25. Liu X, Ma Y, Yin K, Li W, Chen W, Zhang Y, Zhu C, Li T, Han B, Liu X, Wang S, Zhou Z. Long non-coding and coding RNA profiling using strand-specific RNA-seq in human hypertrophic cardiomyopathy. Sci Data. 2019 Jun 13;6(1):90. doi: 10.1038/s41597-019-0094-6. PMID: 31197155; PMCID: PMC6565738.
26. Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. PLoS Computational Biology, 13(5), e1005457. https://doi.org/10.1371/journal.pcbi.1005457
27. Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: A tutorial. Molecular Systems Biology, 15(6), e8746. https://doi.org/10.15252/msb.20188746
28. Mansour, R. G., Stamper, L., Jaeger, F., McGuire, E., Fouda, G., Amos, J., Barbas, K., Ohashi, T., Alam, S. M., Erickson, H., & Permar, S. R. (2016). The Presence and Anti-HIV-1 Function of Tenascin C in Breast Milk and Genital Fluids. PloS One, 11(5), e0155261. https://doi.org/10.1371/journal.pone.0155261

29. Ng, C. J., Liu, A., Venkataraman, S., Ashworth, K. J., Baker, C. D., O'Rourke, R., Vibhakar, R., Jones, K. L., & Di Paola, J. (2022). Single-cell transcriptional analysis of human endothelial colony-forming cells from patients with low VWF levels. Blood, 139(14), 2240–2251. https://doi.org/10.1182/blood.2021010683
30. Ren, S., Peng, Z., Mao, J.-H., Yu, Y., Yin, C., Gao, X., Cui, Z., Zhang, J., Yi, K., Xu, W., Chen, C., Wang, F., Guo, X., Lu, J., Yang, J., Wei, M., Tian, Z., Guan, Y., Tang, L., … Sun, Y. (2012). RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Research, 22(5), 806–821. https://doi.org/10.1038/cr.2012.30
31. Ritter, M., Buechler, C., Langmann, T., & Schmitz, G. (1999). Genomic organization and chromosomal localization of the human CD163 (M130) gene: A member of the scavenger receptor cysteine-rich superfamily. Biochemical and Biophysical Research Communications, 260(2), 466–474. https://doi.org/10.1006/bbrc.1999.0866
32. RNAseq-short-variant-discovery-SNPs-Indels. (n.d.). [Https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels-].
33. Shen, E. H., Overly, C. C., & Jones, A. R. (2012). The Allen Human Brain Atlas. Trends in Neurosciences, 35(12), 711–714. https://doi.org/10.1016/j.tins.2012.09.005
34. Siena, Á. D. D., Plaça, J. R., Araújo, L. F., de Barros, I. I., Peronni, K., Molfetta, G., de Biagi, C. A. O., Espreafico, E. M., Sousa, J. F., & Silva, W. A. (2019). Whole transcriptome analysis reveals correlation of long noncoding RNA ZEB1-AS1 with invasive profile in melanoma. Scientific Reports, 9(1), 11350. https://doi.org/10.1038/s41598-019-47363-6
35. Tan, A., Abecasis, G. R., & Kang, H. M. (2015). Unified representation of genetic variants. Bioinformatics, 31(13), 2202–2204. https://doi.org/10.1093/bioinformatics/btv112
36. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology, 31(1), 46–53. https://doi.org/10.1038/nbt.2450
37. Verma, D., Kumar, R., Pereira, R. S., Karantanou, C., Zanetti, C., Minciacchi, V. R., Fulzele, K., Kunz, K., Hoelper, S., Zia-Chahabi, S., Jabagi, M.-J., Emmerich, J., Dray-Spira, R., Kuhlee, F., Hackmann, K., Schroeck, E., Wenzel, P., Müller, S., Filmann, N., … Krause, D. S. (2019). Vitamin K antagonism impairs the bone marrow microenvironment and hematopoiesis. Blood, 134(3), 227–238. https://doi.org/10.1182/blood.2018874214
38. von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., & Snel, B. (2003). STRING: A database of predicted functional associations between proteins. Nucleic Acids Research, 31(1), 258–261. https://doi.org/10.1093/nar/gkg034
39. Wang, Y., Han, S., Ran, R., Li, A., Liu, H., Liu, M., Duan, Y., Zhang, X., Zhao, Z., Song, S., Weng, X., Liu, S.-M., & Zhou, X. (2021). A longitudinal sampling study of transcriptomic and epigenetic profiles in patients with thrombocytopenia syndrome. Nature Communications, 12(1), 5629. https://doi.org/10.1038/s41467-021-25804-z
40. Xiao, H., Chen, J., Duan, L., & Li, S. (2020). Role of emerging vitamin K‑dependent proteins: Growth arrest‑specific protein 6, Gla‑rich protein and periostin (Review).

International Journal of Molecular Medicine, 47(3), 2. https://doi.org/10.3892/ijmm.2020.4835

41. Zheng, T., Ellinghaus, D., Juzenas, S., Cossais, F., Burmeister, G., Mayr, G., Jørgensen, I. F., Teder-Laving, M., Skogholt, A. H., Chen, S., Strege, P. R., Ito, G., Banasik, K., Becker, T., Bokelmann, F., Brunak, S., Buch, S., Clausnitzer, H., Datz, C., … Franke, A. (2021). Genome-wide analysis of 944 133 individuals provides insights into the etiology of haemorrhoidal disease. Gut, gutjnl-2020-323868. https://doi.org/10.1136/gutjnl-2020-323868