Tampere University

Jalmari Kettunen

WHOLE BLOOD VIROME IN RHEUMATOID ARTHRITIS

Faculty of Medicine and Health Technology

Master's thesis

April 2020

ABSTRACT

Jalmari Kettunen: Whole blood virome in rheumatoid arthritis
Master's thesis
Tampere University
Master's Degree Programme in Biomedical Technology
April 2020

Recent research has shown that human body is host to billions of microbes, including viruses. Set of all viruses found in specific environment, such as in single human organ, is called a virome. In addition, certain viruses can infect humans and remain latent for the whole lifetime of their host.

Rheumatoid arthritis (RA) is a common systemic autoimmune disease characterized by chronic inflammation of joints. Its exact aetiology remains unknown which hinders development of effective prevention or cure. Several studies suggest that certain chronic viral infections, such as those of Epstein-Barr virus and human cytomegalovirus, could trigger rheumatoid arthritis in genetically susceptible individuals. However, traditional methods used in these studies can detect only small number of virus species in one experiment and may have left some undetected. Virome sequencing could potentially study whole virome and thus provide overall view of viruses in RA. Until now, virome sequencing approach has never been applied to RA data.

Aim of this thesis was to test whether RA patients have higher viral abundance or more virus species in their blood than healthy individuals. Furthermore, the project tried to detect sequences of possible novel viruses. These objectives were achieved by setting up a bioinformatics pipeline and applying it to a publicly available RNA-sequencing dataset. The dataset was sequenced from whole blood samples of non-treated RA patients (N=5), treated RA patients (N=7) and healthy controls (N=12). Bioinformatics pipeline included quality control, read alignment, de novo assembly and BLAST database searches.

Several bacteriophages were detected. Bacteriophages are viruses infecting prokaryotes. However, neither their abundance nor species richness was significantly different between groups. Further, pipeline suggested 486 sequences to originate from putative novel viruses. It seems that bacteriophages do not associate with rheumatoid arthritis although some viruses may have remained undetected due to limitations of the raw data.

Keywords: rheumatoid arthritis, viruses, metagenomics, transcriptomics, autoimmune diseases

TIIVISTELMÄ

Jalmari Kettunen: Kokoveren viromi nivelreumassa
Pro gradu -tutkielma
Tampereen yliopisto
Bioteknologian maisteriohjelma
Huhtikuu 2020

Tutkimukset ovat osoittaneet, että ihmiskehossa elää miljardeja mikrobeja, mukaan lukien viruksia. Tietyssä rajatussa paikassa, kuten ihmisen elimessä, sijaitsevien virusten joukkoa kutsutaan viromiksi. On huomattu myös, että tietyt virukset voivat tartuttaa ihmisen piilevästi ja näin pysyä isännässään jopa vuosikymmeniä.

Nivelreuma on yleinen systeeminen autoimmuunisairaus, jonka oireisiin kuuluu nivelten krooninen tulehdus. Sen tarkka syy on edelleen tuntematon, mikä estää tehokkaan ehkäisy- tai parannuskeinon löytämisen. Aiemmat tutkimukset viittaavat siihen, että tiettyjen virusten, kuten Epstein-Barrin viruksen tai sytomegaloviruksen, aiheuttama krooninen infektio saattaisi laukaista nivelreuman perinnöllisesti alttiilla henkilöillä. Näissä tutkimuksissa käytetyt menetelmät voivat havaita vain vähän viruslajeja kerrallaan ja saattavat siten jättää havaitsematta taudin kannalta olennaisia viruksia. Sen sijaan viromin sekvensointi voisi antaa kokonaiskuvan nivelreuman ja virusten yhteyksistä. Nivelreumapotilaiden viromia ei ole aiemmin tutkittu sekvensointidatan avulla.

Tämän tutkielman tavoitteena oli selvittää, ovatko veren virusmäärä ja viruslajien lukumäärä suurempia nivelreumapotilailla kuin terveillä henkilöillä. Lisäksi yritettiin tunnistaa ennestään tuntemattomien virusten sekvenssit. Bioinformaattinen dataprosessi pystytettiin ja sitä sovellettiin julkisesti saatavilla olleeseen RNA-sekvensointidataan, joka oli peräisin hoitamattomien potilaiden (N=5), hoidettujen potilaiden (N=7) ja terveiden kontrollien (N=12) kokoverinäytteistä. Bioinformaattiseen dataprosessiin kuului muun muassa laadunvalvonta, sekvenssien linjaus virusgenomeihin, sekvenssien yhdistäminen ja BLAST-tietokantahaut.

Useita bakteriofageja havaittiin. Bakteriofagit ovat esitumallisissa loisivia viruksia. Niiden määrä tai lajikirjo ei kuitenkaan ollut merkitsevästi erilainen ryhmien välillä. Lisäksi havaittiin yhteensä 486 sekvenssiä, jotka saattoivat olla peräisin ennestään tuntemattomista viruksista. Bakteriofageilla ei ilmeisesti ole yhteyttä nivelreumaan, vaikkakin jotkin virukset saattoivat jäädä havaitsematta raakadatan rajoitteiden vuoksi.

Avainsanat: nivelreuma, virukset, metagenomiikka, transkriptomiikka, autoimmuunisairaudet

Acknowledgements

This thesis work was carried out in Microbiology and Immunology research group led by Professor Mikko Hurme. I would like to address my gratitude to Mikko for patient mentoring and arranging funding for this project. Thank you for the opportunity. I would also like to thank my co-supervisor Arttu Autio for helping in bioinformatics analysis. Thank you for great feedback. Also, I want to thank Tapio Nevalainen for many useful conversations and additional help. I am grateful for funding provided by Tampere Tuberculosis Foundation and Competitive State Research Financing of the Expert Responsibility area of Tampere University Hospital. I also thank CSC for providing computational resources for this project. Finally, I thank my family for undying support during all my study years.

Iisalmi, April 2020

Jalmari Kettunen

Table of contents

1 Introduction 1
2 Literature review 3
2.1 Rheumatoid arthritis 3
2.2 Possible viral mechanisms to alter autoimmune response 6
2.2.1 Viral mechanisms protecting from autoimmunity 6
2.2.2 Viral mechanisms inducing autoimmunity 7
2.3 Rheumatoid arthritis and viruses 11
2.3.1 Epstein-Barr virus and rheumatoid arthritis 11
2.3.2 Human cytomegalovirus and rheumatoid arthritis 12
2.3.3 Other highly prevalent viruses and rheumatoid arthritis 14
2.4 High-throughput sequencing of human virome 15
2.4.1 Challenges of virome research 15
2.4.2 Healthy human blood virome 17
3 Objectives 19
4 Materials and Methods 20
4.1 Origin of data 20
4.2 Read alignment 20
4.3 Analysis of unaligned reads 21
4.4 Qualitative analysis of contigs 21
4.5 Statistical analysis of quantitative viral matches 22
5 Results 24
5.1 Quantitative results 24
5.2 Qualitative results 28
6 Discussion 30
7 Conclusion 33
8 References 34
9 Appendices 42

Abbreviations

ACPA	anti-citrullinated protein antibody
ACR	American College of Rheumatology
AD	autoimmune disease
BLAST	Basic Local Alignment Search Tool
CMV	human cytomegalovirus
CD	cluster of differentiation
CSC	Finnish IT Centre for Science
DMARD	disease modifying anti-rheumatic drug
EBNA-1	Epstein-Barr virus nuclear antigen 1
EBV	Epstein-Barr virus
ELS	ectopic lymphoid structure
EULAR	European League Against Rheumatism
HC	healthy control
HERV	human endogenous retrovirus
HHV	human herpesvirus
HLA	human leukocyte antigen
HML	human endogenous murine mammary tumour virus -like
IgG	class G immunoglobulin
IgG1Fc	Fragment crystallizable part of type 1 immunoglobin G
IGV	Integrative Genomics Viewer
IL-10	interleukin-10
HMM	hidden Markov model
MAPQ	mapping quality value
MHC	major histocompatibility complex
MS	multiple sclerosis
NCBI	National Centre of Biological Information
NSAID	non-steroid anti-inflammatory drug
ntRA	non-treated rheumatoid arthritis patient
PAD	peptidylarginine deiminase
PBMC	peripheral blood mononuclear cell
PCR	polymerase chain reaction
PD1	programmed cell death receptor
PD-L1	programmed cell death-ligand 1

RA	rheumatoid arthritis
SLE	systemic lupus erythematosus
T1D	type 1 diabetes
TGF-β1	tissue growth factor beta 1
TLR	Toll-like receptor
tRA	treated rheumatoid arthritis patient
VLP	virus-like particle

1 Introduction

Traditionally, it was believed that human body is a sterile environment. Recent studies have shown that this is not the case: billions of viruses and other microbes reside in multiple organs. This human microbiome can affect our health with various mechanisms, but much is still unknown [1]. In addition, it has been observed that many viruses can infect humans and stay latent for the whole lifetime of their host [2]. Viral composition of microbiome, also known as virome, is the less-studied part of the microbiome [3]. Many viruses remain unidentified which is why this part of virome is called viral dark matter [4].

Rheumatoid arthritis (RA) is one of the most common autoimmune diseases in the world. It causes chronic inflammation of joints. Exact cause of arthritis is unclear. This inhibits development of cure or prevention although RA causes substantial morbidity globally [5]. More knowledge on RA pathogenesis is needed in order to find better treatments.

Certain viruses, such as Epstein-Barr virus (EBV) and human cytomegalovirus (CMV), have been associated with development of RA. These viruses can establish chronic latent infection in humans. Many studies suggest that a virus could induce rheumatoid arthritis in genetically susceptible individuals [6]. If viral aetiology were proven correct, this would have remarkable consequences. In theory, many RA cases could be prevented with a large-scale vaccination campaign. However, much more research is needed. Previous publications studying the connection of RA and viruses have often used polymerase chain reaction (PCR) or immunoassays to measure viruses in blood [7,8]. In these methods, only those viruses for which primers or antibodies have been designed are detected. Thus, these methods cannot study whole virome in one experiment.

High-throughput transcriptome sequencing can potentially detect both RNA and DNA viruses. It may also detect viruses that have not been associated with RA previously. So far, rheumatoid arthritis has never been studied with high-throughput virome sequencing. The aim of this thesis is to define whole blood viral transcriptome in RA patients and in control groups. This is achieved by applying bioinformatics pipeline to a publicly available dataset [9]. This project

seeks to measure if number of virus species and viral abundances are significantly different between RA patients and control groups. In addition, amount of viral dark matter is identified.

2 Literature review

2.1 Rheumatoid arthritis

Autoimmune diseases (AD) are a heterogenous group of diseases in which the body elicits an immune response against its own tissues or cells. This requires breaking of physiological tolerance against natural tissue molecules, also known as self-antigens. It has been observed that even healthy individuals have small number of autoreactive T cells in circulation. This is because some of them can escape negative selection in thymus if they bind self-molecules in low affinity. It is believed that a secondary, peripheral breaking of tolerance is required to induce an autoimmune disease. ADs differ both in manifestation of disease and in prevalence [10].

Hayter et al. summarized that there exists 81 autoimmune diseases [11]. However, exact definition of an autoimmune disease is under debate. For instance, some authors do not define psoriasis, Crohn's disease or celiac disease as autoimmune diseases because immune response in these are induced by bacteria or food substance rather than by human cells [12]. Hayter et al. proposed criteria where disease is an autoimmune disease if it has evidence for at least 3 major criteria or at least 3 minor criteria [11].

The most common autoimmune diseases are rheumatoid arthritis, Hashimoto's thyroiditis and Graves' disease, each with over 0.5 % global prevalence [11]. Other autoimmune diseases include type 1 diabetes (T1D), multiple sclerosis (MS) and systemic lupus erythematosus (SLE). Incidence of many autoimmune diseases, such as type 1 diabetes and celiac disease, have increased in several countries recently [13,14].

Rheumatoid arthritis (RA) is an autoimmune disease that causes pain, stiffness and swelling of joints, most commonly in joints of hand. Also, systemic inflammation is often present and therefore rheumatoid arthritis is classified as systemic autoimmune disease. Synovial membrane (synovium) is infiltrated by B cells and other leukocytes [15]. At microscopic level, B cells often form clusters called ectopic lymphoid structures (ELS) [16]. Risk factors for RA include smoking, lower educational level, high birth weight, obesity [5], certain genotypes and

epigenetic modifications [17]. Two thirds of RA patients are women [5,17] and although patients can be of any age, elderly have higher prevalence. There is an ongoing debate about whether the incidence of RA is increasing or decreasing. Many studies reported declining incidence during the second half of the 20th century but the incidence may have risen after 1995 [5].

Untreated RA leads to synovial joint destruction. Treatment includes surgery, physiotherapy and immunosuppressive drugs. Examples of used pharmacological agents include steroid drugs such as methotrexate and many biological drugs such as infliximab. Together, these drugs are called disease-modifying anti-rheumatic drugs (DMARD) [18]. Remission after cease of drug treatment is rare but possible [5].

Diagnosis of RA is based on multiple factors, such as assessment by both clinician and patient but also biomarkers. Traditionally, rheumatoid factor has been used as biomarker for diagnosis. Rheumatoid factor is a general name for any antibody that recognizes Fc part of human IgG antibodies [5]. However, titer of anti-citrulline protein antibodies (ACPA) has been reported to be a more specific biomarker than rheumatoid factor [8]. It has been observed that 50-67% of RA patients are positive for ACPAs. In addition, half of RA patients have ACPAs several years before the onset of symptoms [19]. Another study claimed that even up to 90 % of RA patients have antibodies against neutrophil-derived citrullinated histones [20]. On average, ACPA positive patients have worse prognosis than ACPA negative patients. ACPAs are strong predictors of joint erosion [20]. Thus, RA patients are often classified into these two groups [5].

Physiologically, peptidylarginine deiminase (PAD) enzymes transform arginine residues of various proteins into citrulline residues. The proteins can be fibrin, for example. There are five PAD isoforms from which PAD2 and PAD4 are active in RA [20]. It is known that one transcript variant of PAD4 is more stable than others. The allele which encodes this variant is also a risk factor for RA [5]. Unnaturally stable enzyme could produce large amounts of citrullinated proteins and in this way promote RA pathogenesis. It must be mentioned here that a pan-PAD inhibitor decreased severity of collagen-induced arthritis but also joint and serum protein citrullination in mice [21]. However, citrullination of proteins is very common process and thus it is still an unsolved mystery how immune system breaks its tolerance for citrullinated

proteins. Citrullinated proteins can increase TNF-alpha production of macrophages by themselves but more studies are needed. In any case, effects of ACPAs are clear. They activate inflammatory macrophages and differentiation of osteoclasts [20].

Genetic risk factors determine 50-60 % of risk developing RA [22]. Indeed, smoking is a risk factor for developing ACPA-positive RA only if patient is genetically susceptible to RA [23]. The most significant genetic factors are variants in HLA-DRB1 gene. Class II human leukocyte antigen receptors (HLA II receptors), also known as human major histocompatibility complex type II receptors (MHC II receptors), are highly divergent receptors encoded by HLA-D gene region. They are present on cellular membranes of antigen-presenting cells, such as dendrocytes, macrophages and B cells. They bind peptides, i.e. possible antigens, processed by the cell. When other leukocytes recognize the HLA-peptide-complex on antigen-presenting cell, leukocytes can be activated and promote immune response against the antigen. Class II HLA receptor consists of four different peptide chains (Figure 1). For example, gene HLA-DRB1 encodes beta chain of type 1 belonging to gene group R of HLA class II [24].

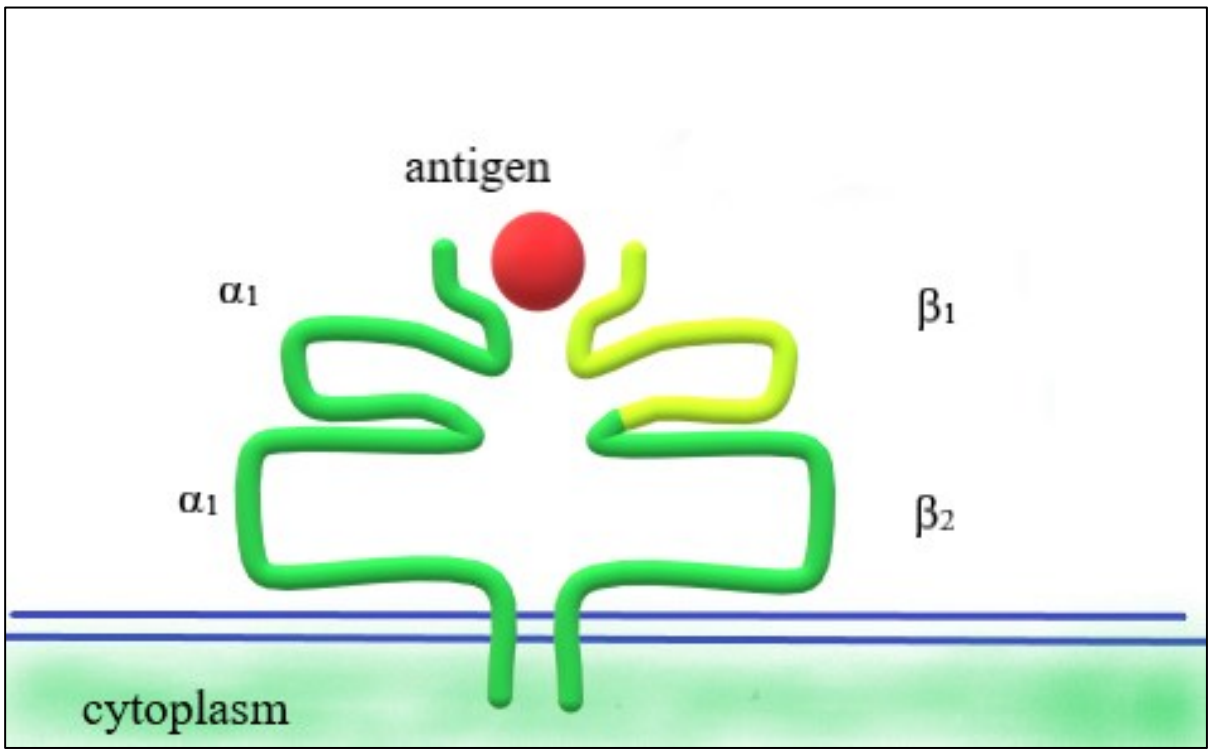

Figure 1. HLA class II receptor on antigen-presenting cell. Variants of so called 'shared epitope' take place in beta 1 peptide chain which is highlighted in yellow in this schematic figure.

All RA associated HLA-DR alleles encode the same motif located in the third hypervariable region (HVR3) of the beta 1 chain. Amino acid content of this motif is QKRAA, QRRAA or RRRAA. The different variants are commonly called as shared epitope [8,20]. Seventy percent of RA patients express HLA–DR molecules containing the shared epitope. This suggests that antigen-presenting is important in RA pathogenesis [8]. It is interesting that citrullination and HLA-polymorphism have a connection. Shared epitope appears to be a risk factor for ACPA production in RA rather than an independent risk factor for RA development [20]. In another study, citrullinated fibrinogen induced arthritis in transgenic mice with shared epitope [25]. When splenic T cells or plasma were transferred from these mice to naïve recipients, recipient mice developed arthritis [26].

Based on discussed associations, many novel pharmacological agents for RA have been tested on rats and mice [20]. Despite these efforts, RA continues to cause substantial morbidity globally [27]. No certain RA treatment is in use because precise mechanisms of pathogenesis are unknown. New research methods are needed to uncover them.

2.2 Possible viral mechanisms to alter autoimmune response

2.2.1 Viral mechanisms protecting from autoimmunity

Traditionally, healthy human organs were believed to be sterile. New research has proven that this is not the case. Human body compartments are swarming with different viruses and other microbes. In addition, many of these species are remarkably prevalent in humans. For example, some studies estimated that over 90 % of human population is infected by viruses from family *Anelloviridae*. Other highly prevalent viruses include herpesviruses such as human herpesvirus 6, human herpesvirus 7, Varicella zoster virus (human herpesvirus 3), human cytomegalovirus (human herpesvirus 5 or CMV) and Epstein-Barr virus (human herpesvirus 4 or EBV) [2,28].

How can a chronic viral infection persist a lifetime without any obvious symptoms? Answer is far from complete, but this can happen with many mechanisms. Viruses can remain latent in cells for years. Herpes simplex viruses infecting neurons are a common example of this. Also, immune system can limit its response to a virus or virus evades the immune system successfully.

These adaptations arguably result from coevolution between host and virus that has lasted millions of years [2].

During prolonged viral infection, T and B cells can lose their function. In addition, inhibitory receptors of T cells, immunoregulatory cytokines and regulatory T cells can restrict immune response. Restriction of immune response is reasonable because this limits tissue damage caused by cytotoxic leukocytes and antibodies. Additionally, unrestricted exponential growth of T cells would be harmful [2].

Once a virus can express their genes, it can evade the immune system in many ways. Viral products can decrease expression of recognition molecules of T cells and natural killer cells or inhibit antigen presentation. They can also act as antagonists of proinflammatory cytokines and chemokines or block their intracellular effects [2].

When viral infection modifies immune system, it might protect us from autoimmunity as a side effect. It has been proposed that EBV and hepatitis B viruses could protect from T1D and SLE, respectively. These claims are based on observations where healthy controls had more virus-specific antibodies in circulation than AD patients [29]. Similarly, Coxsackie group B viruses, belonging to enteroviruses, were observed to protect mice from development of type 1 diabetes. Protective effect resulted from upregulation of PD-L1 in lymphocytes, inhibition of PD1-producing CD8+ T cells and production of CD4+CD25+ regulatory T cells [29].

This is interesting also considering the hygiene hypothesis [30]. Viral infections might modify immune system in such way that it does not respond to irrelevant stimuli which would cause autoimmunity disorder. Viruses might prevent ADs in this way.

2.2.2 Viral mechanisms inducing autoimmunity

Although a few protective effects have been observed, most studies propose exactly the opposite: viral infections are risk factors for autoimmune diseases. It is well-known fact that autoimmune diseases and aging can make individuals more susceptible to viral infections but

what if viruses caused autoimmunity? This is newer idea and many theories have been proposed how this could happen: viral persistence [17], molecular mimicry, epitope spreading, bystander activation and polyclonal activation [29].

Viral persistence is perhaps the simplest mechanism. Viral infection activates intracellular Toll-like receptors (TLR) which leads to release of inflammatory cytokines and interferons. Adaptive immune system can also respond and deploy cytotoxic T cells which destroy infected cells. Because immune system never succeeds to clear the infection despite the tissue damage, symptoms of autoimmune disease are seen [17,31]. TLR-activation which affects T1D has been observed with rotavirus [32] and a Coxsackie B virus [33]. This very direct mechanism seems unlikely in many cases, but other mechanisms discussed here may require chronic viral infection as initial step.

In molecular mimicry, virus antigen happens to structurally resemble a self-antigen. Antigen-presenting cells, such as dendritic cell, take up viral antigens and present them to T cells and B cells. Antibodies produced by B cells may bind to self-antigens because of structural similarity and in this way cause destruction of tissue. In addition, matching autoreactive T cells may encounter antigen-presenting cells and proliferate which would cause wider immune response [6,10]. A well-known bacterial example of this is mimicry between M protein of Streptococcus pyogenes and cardiac myosin which leads to rheumatic fever [10,29].

In bystander activation, non-specific antiviral response destroys tissue and subsequently releases self-antigens which would not be available to leucocytes otherwise. In inflammatory environment full of pro-inflammatory cytokines, these self-antigens may induce activation of even ‘ignorant’ autoreactive T cells. Subsequently, this also results in activation of B cells [6].

Epitope spreading resembles bystander activation. Here, antiviral response leads to increased release of molecules. Some of these self-molecules and viral antigens are taken up by B cells and processed. T cells activated from viral antigens activate B cells that present viral antigens on their surface. However, B cells may mistake to produce antibodies against self-antigens

instead of viral antigens because they are present in the same vesicles of the B cell. Secreted antibodies cause the immune response [6,10].

Polyclonal activation is physiological equivalent of epitope spreading but it may also contribute to autoimmunity. When a large viral antigen activates humoral immune system, many different B cells produce many different antibodies against it. This is called polyclonal activation. However, it increases the probability of producing antibodies that recognize self-molecules, especially if humoral response lasts long periods [29]. This has been demonstrated in connection of EBV and Graves' disease [34].

Many additional mechanisms have been proposed that could explain how viruses cause autoimmunity disorders. EBV may cause immortalization of autoreactive B cells [35]. Several studies indicate that hepatitis B virus could cause polyarteritis nodosa via formation of immune complexes in blood vessel walls. Other evidence suggests Coxsackie B viruses could induce type 1 diabetes by dysregulating microRNAs in Langerhans islets [29]. However, more research is needed to proof these claims.

Even more puzzling is that Coxsackie B virus can be either induce or inhibit T1D, depending on context. One explanation was that this results from different ages of laboratory mice [29]. Table 1 shows viruses that are hypothesized to cause ADs. Although many associations have been found, exact molecular mechanisms are still unclear and should be studied separately for every combination of virus and AD.

Table 1. Examples of viral infections that have been associated with autoimmune diseases. Table is modified from Smatti et al [6].

virus	**positive correlation with AD**	**negative correlation with AD**	**proposed mechanism(s) for positive correlation**	**references**
Chikungunya virus	symmetric polyarthritis		epitope spreading	[36]
Coxsackie B viruses	myocarditis, T1D	T1D	bystander activation, molecular	[6,17,29]

			mimicry, viral persistence (T1D)	
Dengue virus	SLE		epitope spreading	[6]
Epstein-Barr virus	systemic sclerosis, myasthenia gravis, SLE, Sjögren's syndrome, thyroiditis, autoimmune hepatitis, Graves' disease, Hashimoto's disease, MS, RA	T1D	molecular mimicry (MS, SLE, RA), epitope spreading (RA, SLE), polyclonal activation (Graves' disease), immortalization of autoreactive B cells	[6,8,17,29,34,35]
hepatitis B virus	polyarteritis nodosa, antiphospholipid syndrome	SLE		[29]
hepatitis C virus	type 2 autoimmune hepatitis, cryoglobulinemia, polyarthritis, Sjögren's syndrome, thrombocytopenia, thyroiditis, vasculitis		bystander activation (Sjögren's syndrome, thyroiditis)	[6,17]
Herpes simplex viruses	autoimmune encephalitis, stromal keratitis		molecular mimicry (encephalitis)	[6]
HHV-6	MS, autoimmune thyroiditis			[6,17]
human cytomegalovirus	systemic sclerosis, Sjögren's syndrome, MS, RA, SLE, T1D		molecular mimicry (MS), epitope spreading (RA, SLE)	[6,7]
human endogenous retroviruses	MS, RA, SLE, Sjögren's syndrome		molecular mimicry	[37,38]
human T-lymphotropic virus type 1	myelopathy			[6]
Influenza A virus	Acute disseminated encephalomyelitis, T1D, type 1 narcolepsy		bystander activation (T1D, encephalomyelitis), molecular mimicry (narcolepsy, encephalomyelitis)	[6]
measles virus	MS			[6]
mumps virus	T1D, MS			[6,39]
rotavirus	T1D			[6,17]
rubella virus	T1D, MS			[6,39]
Varicella zoster virus	MS			[40]

West Nile virus	Myasthenia gravis		molecular mimicry, bystander activation	[6]
zikavirus	Guillain-Barre syndrome		molecular mimicry	[6]

2.3 Rheumatoid arthritis and viruses

2.3.1 Epstein-Barr virus and rheumatoid arthritis

Epstein-Barr virus is a human herpesvirus that has estimated prevalence of 80-90% in adult human population [2,8]. It has double-stranded DNA genome of 172 kilobases and belongs to genus *Lymphocryptovirus*. In general, EBV follows life cycle of mature B lymphocyte but it can infect also epithelial cells. It can reside as latent infection in resting B cells for the whole lifetime of its host. It has been associated with many cancers, including Burkitt's lymphoma and Hodgkin's disease [8]. This is not surprising considering that it can immortalize B cells [35]. In addition, it has been associated with RA, thyroiditis, Sjögren's syndrome [8], MS, Graves' disease, systemic sclerosis, myasthenia gravis and SLE [29] but it is suspected to correlate negatively with T1D [29].

Prior EBV infection is not more frequent in RA patients than in healthy individuals [41]. This is reasonable since the virus is extremely common in population. However, in vitro study observed that CD8+ cells are less responsive to produce interferons against EBV peptides in RA patients than in healthy EBV carriers [42]. Reason for this is uncertain. Perhaps CD8+ cells restrict their response against chronic viral infection? This deficiency might lead to EBV overload.

Studies utilizing PCR (polymerase chain reaction) have not proved that RA synovium would have higher DNA load than healthy synovium. Data has been very variable, to say the least [8]. In contrast to synovium, EBV DNA is detected more frequently in peripheral blood mononuclear cells (PBMC), synovial fluid and saliva of RA patients than in healthy controls [8]. Further, other PCR studies found that EBV load of PBMCs was 7-10-fold higher in RA patients if compared to controls [43,44].

Apparently, these high amounts of EBV can provoke immune response. Alspaugh et al. observed that many RA patients have antibodies against Epstein-Barr virus nuclear antigen 1 (EBNA-1). In addition, these antibodies were present in high titers [45], although a more recent study gave different results [46]. Serum reactivity to EBNA-1 was observed in 67% of patients with RA whereas only 8% of control group reacted to it [45]. Another study observed that high proportion of CD8+ T cells from RA joints recognize key proteins of EBV lytic infection [47].

How could EBV trigger rheumatoid arthritis? There are several connections between EBV proteins and self-antigens relevant to RA. For example, citrullinated EBNA-1 can cross-react with citrullinated human fibrin which is highly prevalent in RA patients [48]. Also, ELS in RA synovium contain plasma cells that produce ACPAs. These same cells are infected with EBV [16]. Johansson et al. found that ACPAs appear years before onset of RA, more often in future patients than in controls. These antibodies were able to bind citrullinated EBV proteins [49]. When ELS-containing RA synovia were transplanted onto immunodeficient mice, the mice produced antibodies against citrullinated EBV proteins [16]. Notably, Roudier et al. observed that some healthy individuals with prior EBV infection have T cells that recognize peptides containing the shared epitope [50]. T cell recognition means that the antigen can potentially cause immune response.

Evidence suggests that EBV proteins could become citrullinated by PAD enzymes. They would be presented on HLA receptors. If individual had shared epitope, T cells would mistakenly recognize citrullinated EBV protein as citrullinated fibrin, for example. This could initiate immune response against citrullinated fibrin, an ubiquitous self-molecule. Although evidence looks strong, final details are missing.

2.3.2 Human cytomegalovirus and rheumatoid arthritis

Human cytomegalovirus is a herpesvirus and one of several viruses in genus *Cytomegalovirus*. As all herpesviruses, CMV have a large double-stranded DNA genome of 230 kilobases and can encode hundreds of proteins, many of which are antigens. It infects 40–99% of the adult population depending on ethnic and socioeconomic conditions [2,7,51]. Infection in immunocompetent individuals normally proceeds unnoticed and lasts a lifetime. After primary infection in one of many possible cell types (epithelial cells of liver, lungs, kidney, salivary

glands or large intestine, placenta endothelial cells, smooth muscle cells, fibroblasts, neuronal cells or various myeloid cells), CMV remains latent in CD34+ myeloid progenitors, from which virus can reactivate and replicate in lytic way [7].

Even though a healthy immune system controls CMV replication, the virus cannot be eliminated by immune functions or by antiviral drugs [7], such as valganciclovir or ganciclovir [51]. Upon failure or reduced efficiency of immune system, opportunistic CMV infections may lead to severe or even fatal illness. Immunodeficiencies occur in infants, AIDS patients or graft recipients, for instance. Symptoms of active CMV replication include fatigue, hepatitis, enterocolitis, encephalitis, pneumonitis, retinitis and sensorineural hearing loss [7].

What is the role of CMV in rheumatoid arthritis? Pierer et al. noticed that RA patients with IgG antibodies against CMV had more severe joint destruction than seronegative patients [52]. However, it has not been proven that CMV-specific antibodies would be more frequent [7] or higher [46] among RA patients than among healthy controls. When we discuss links between CMV and RA, CD4+CD28- cells must be mentioned too.

During aging, proportion of CD4+CD28- cells among T cells expands. These cells are memory T cells that have lost their CD28 receptor after repetitive cell division. Despite being senescent, they can potentially produce large amounts of pro-inflammatory cytokines before apoptosis. CD4+CD28- cells are associated with mortality [53] and age-related decrease in immune functions, i.e. immunosenescence. Notably, RA patients have increased CD4+CD28- compartment [7] and a high amount of these cells predicts faster destruction of joints [54]. In addition, Thewissen et al. associated HLA-DR4 alleles with an increase in CD28- CD4+ compartment [55].

Interestingly, CMV infection has been associated with RA and CD4+CD28- cells. Although RA patients have expanded CD4+CD28- compartment, number of cells can be 3-10-fold higher if CMV infection is also present. Numerous studies have demonstrated this [7,15]. Effect of chronological age seems small compared to these results [56]. CD4+CD28- cell frequencies rarely exceed 1 %, even in very old individuals, unless CMV infection is present [15]. In RA,

CMV infection might activate autoreactive CD4+CD28- cells which would cause inflammation of joints.

Despite these many associations, no CMV peptide has been proven to trigger RA [7,15]. HLA receptor would present this peptide to leukocytes which would initiate immune response. It must be noted that association does not equal causation. CMV and CD4+CD28- cells might only worsen RA that has already been initiated. Furthermore, CMV seems to have immunosuppressive abilities which do not make investigations any easier: it can cause apoptosis of leukocytes and production of immunosuppressive cytokines TGF-β1 and IL-10 [57,58].

2.3.3 Other highly prevalent viruses and rheumatoid arthritis

Human herpesviruses 6 and 7 has been reported to be highly prevalent in population [2]. Broccolo et al. found that AD patients had more frequently cell-free HHV-6 viremia and higher HHV-6 specific IgG-levels than controls [59]. However, target group included many other diagnoses than just RA. Furthermore, Kholodnyuk et al. did not detect HHV-6 or HHV-7 in blood of RA patients [60]. All in all, there is not enough evidence to link these herpesviruses to RA pathogenesis.

Human endogenous retroviruses (HERV) have integrated into human genome over the course of millions of years and subsequently have lost their ability to infect. However, some of these proviruses have active expression and can therefore affect human health. Members of HERV-K family are one of the most active HERVs. Could they associate with rheumatoid arthritis? Many HERVs are located near genes encoding MHC I and II receptors [37]. According to Krzysztalowska-Wawrzyniak et al, HERV-K113 is slightly more prevalent in RA patients than in healthy controls but prevalence does not associate with clinical features [38]. A more studied HERV is HERV-K10 whose Gag gene can encode many proteins. HERV-K10 have high mRNA transcription in PBMCs and synovium fibroblasts of RA patients [61]. It also seems that RA patients have increased antibody levels against Gag matric protein of HERV-K10 [37]. Bioinformatics suggested reason for this: Gag matrix protein is antigenic and structurally resemble IgG1Fc, a target of rheumatoid factor. Indeed, RA patients showed reactivity to certain peptide sequence of this protein [62]. These interesting connections need more research.

Anelloviridae is a recently found, genetically diverse [63] family of circular single-stranded DNA viruses. It is thought that anelloviruses infect most humans latently by the end of childhood. Indeed, their estimated prevalence in human population is 70-100 % depending on geographical area [2,63,64]. Researchers have tried to link Torque teno virus, a well-studied anellovirus, to many diseases. Maggi et al. showed that RA patients have higher Torque teno virus loads than healthy controls [65] although virus load might be simply a consequence of disease.

More research is needed to illuminate relationship of viruses and RA pathogenesis. Most of the mentioned studies used PCR or immunoassays to detect viruses. A more high-throughput method could reveal much more viruses in single experiment.

2.4 High-throughput sequencing of human virome

2.4.1 Challenges of virome research

High-throughput sequencing has revolutionized study of viruses, including those inhabiting human body. In principle, sequencing can detect all nucleic acids in sample whereas quantitative PCR, the current gold standard method [66,67], can detect only a handful of viruses for which PCR primers are designed [68]. This makes high-throughput sequencing methods very good at screening and narrowing down number of pathogens. Sequencing methods can also discover viruses previously unknown to science. Indeed, this metagenomics approach has discovered that many human organs have a distinct collection of viruses, virome [3]. However, sequencing approach has many challenges that are discussed here.

Proportion of viral sequences in metagenomics sample is often very small. This is a challenge because sequencing has been reported to be less sensitive than PCR [66]. Many wet lab methods have been used to selectively increase number of viral sequences. Culturing virus in host cells increases concentration of viral nucleic acids and improves computational results. Second possible method is virus-like particle (VLP) purification which can include filtering, nuclease treatment and gradient ultracentrifugation, for example. Viral tagging is an enhancement to VLP purification where viral DNA is labeled with fluorescent dye. Double-stranded RNA

viruses can be concentrated using nucleases, antibodies or chromatography [4]. Finally, virome capture sequencing was introduced as a sequencing system customized for virome sequencing [69]. However, all these enrichment methods are applicable to only certain viruses or they concentrate some virus groups more than others. This can lead to biased results [4].

Current short read alignment programs, such as Bowtie2 [70], assign sequencing reads very efficiently to correct genome. Unfortunately, viruses mutate very rapidly and therefore genomes of even the same species can differ. If a short read has any mismatches in alignment process, it likely remains unaligned. Because of this, de novo assembly programs are used to combine reads into contiguous sequences, also known as contigs, based on genomic context [4]. De novo assemblers are not perfect, though. All sequences from the same species are not always combined into the same contig which wastes computational resources downstream in bioinformatics pipeline. Similarity compression algorithms can alleviate this issue by identifying highly similar contigs and preserving only the longest one of them [71]. Another standard procedure is to use more lenient alignment algorithms, such as BLAST, instead of strict short read aligners [4].

Many viruses share similarity only in amino acid level. For this reason, nucleic acid sequences are often translated into peptide sequences in metagenomics studies. When alignment algorithms calculate similarity between sequences, they penalize for mismatch equally in every position. However, multiple sequence alignment studies show that some protein domains have been conserved better than others during evolution. Hidden Markov model (HMM) profiles can take these domains into consideration and are interesting if the goal is to discover novel virus species [4]. HMMER is an example of software which utilizes HMM profiles [72].

All mentioned computational methods rely on sequence similarity. However, some viruses lack direct sequence similarity to all viruses in databases. One reason for this is that current databases are biased toward pathogenic human viruses [4]. Because of this, many alignment-independent methods have been developed. It is known that viral sequences can be separated from other sequences based on frequencies of oligomers, i.e. k-mers. Methods based on k-mers can distinguish viruses of different Baltimore classification [73], viruses of different hosts [74] and even human-infecting viruses from other viruses [75]. VirFinder was developed to detect viral

nucleic acid sequences in metagenomics data. It uses logistic regression and lasso regularization in its classification [76]. Similarly, Seguritan et al. applied k-mer approach to detect viral structural proteins [77]. Some pipelines, such as CLARK [78] and KrakenUniq [79], combine k-mer information with databases to classify viruses taxonomically but they are extremely memory-intensive.

2.4.2 Healthy human blood virome

Metagenomics studies have illuminated what could be the physiological composition of human blood virome. Anelloviruses are constantly detected in these studies and are regarded as the most prevalent viruses in humans. Moustafa et al. sequenced DNA blood virome in over 8,000 apparently healthy individuals. They found anelloviruses in 8 % of individuals [28]. In fact, one DNA virome study found solely anelloviruses in a control group consisting of 10 healthy adults [80]. Anelloviruses were found also in studies where total nucleic acid [81,82] or solely transcriptome [67] was extracted. This suggests that they are actively transcribed in blood cells. Prevalence of *Anelloviridae* varied from 6 % to 93 % in these studies. In general, more recent studies have higher *Anelloviridae* prevalences which suggests advancements in methodology.

Herpesviridae is another commonly detected virus family in blood. These viruses were the most frequently detected by Moustafa et al, with combined prevalence of over 40 % [28]. Total nucleic acid experiments have also detected them [81,82] but not as frequently as traditional estimates (up to 90 %) suggested [2].

Pegivirus C is RNA virus belonging to family *Flaviviridae*. It can infect lymphocytes persistently, but the infection can be eliminated. It has not proven to be detrimental to health. In fact, it may inhibit disease progression in HIV patients [83]. Whole nucleic acid and whole transcriptome sequencing has detected pegivirus C in blood of healthy Nigerian, Japanese and Chinese populations [67,82,84]. A few samples were positive in each study which supports suggested prevalence of 1-4 % [2].

Bacteriophages (phages for short) are viruses infecting prokaryotes. Notably, many blood virome studies detected phages, such as those from families *Myoviridae*, *Inoviridae* and

Siphoviridae [28,67,85]. It is well-established that these DNA virus families are abundant in human gut [3]. Phages were detected with both DNA and transcriptome extraction protocols. Even so, authors have regarded bacteriophages as contaminants [28,67]. Other viruses detected by whole virome sequencing are generally supported only by individual studies.

3 Objectives

As medication used in RA causes differential gene expression in blood cells [18], it is also possible that this treatment affects transcription machinery of viruses. Therefore, the aim of this thesis is to measure viral transcriptome of whole blood in three groups: non-treated RA patients, treated RA patients and healthy controls. Project should answer following issues: 1) Is total viral abundance per individual different between groups? 2) Are abundances of individual viruses different between groups? 3) Is number of virus species per individual different between groups? Additionally, putative viral sequences are identified among unidentified sequences. Hypothesis is that non-treated RA patients have more virus species and higher viral abundances than treated patients or healthy controls.

4 Materials and Methods

4.1 Origin of data

The RNA-sequencing dataset analysed in this work originates from 12 RA patients and 12 healthy individuals. Each sample originates from one person. Origin of dataset was described previously [9]. Twelve female patients from Karolinska University Hospital (Solna, Sweden) had RA that corresponded to the American College of Rheumatology (ACR) 1987 criteria and the European League Against Rheumatism (EULAR)/ACR 2010 criteria for RA, according to assessment by trained rheumatologists. Of the 12 patients with RA, 5 were non-treated and had not previously used antirheumatic drugs (patients with early RA with symptom duration less than 1 year), and 7 were receiving anti-rheumatic treatment (either methotrexate or biological agents). Non-treated RA patients could still receive non-steroid anti-inflammatory drugs (NSAID). Twelve matched healthy female individuals were included as a control group. Whole blood from all patients and controls was collected in PAXgene Blood RNA Tubes according to the manufacturers protocol and saved at -20 °C. RNA was extracted using PAXgene Blood miRNA kit (PreAnalytiX, Hombrechtikon, Switzerland). TruSeq library construction was performed before RNA-sequencing with Illumina HiSeq 2000 platform (Illumina, San Diego, USA). Sequencing produced 20 million 101 base pairs long paired-end reads per sample. Data was deposited in NCBI Sequence Read Archive (GEO:GSE90081).

4.2 Read alignment

Bioinformatics pipeline (Figure 2) modified from previous study [67] was run in Puhti supercomputer cluster of CSC (Espoo, Finland). RNA-sequencing reads of 24 samples were downloaded from Sequence Read Archive in FASTQ format. Illumina Universal Adapters were trimmed with TrimGalore (v0.6.4; https://github.com/FelixKrueger/TrimGalore; 26.2.2020). Other quality filtering was done with following parameters of PRINSEQ (lite v0.20.4) [86]: read length ≥ 70 nucleotides, mean quality score of read ≥ 25, proportion of ambiguous bases ≤ 1 %, filter all kinds of duplicates, DUST score measuring low complexity ≤ 7. Quality filtering was confirmed with FastQC (v0.11.8; https://www.bioinformatics.babraham.ac.uk/projects/fastqc; 26.2.2020). Quality filtered reads were subtracted sequentially by aligning them with Bowtie2 (v2.3.5.1) [70] against human reference genome (GCF_000001405.26_GRCh38_genomic.fna from NCBI), human reference

transcriptome (GCF_000001405.26_GRCh38_rna.fna from NCBI) and latest Human Microbiome Project genomes [87] (2,236 archae, bacterial and fungi genomes downloaded 15.11.2019 from NCBI). Subtracted reads were aligned against latest complete viral Refseq genomes [88] (8,992 genomes downloaded 12.12.2019 from NCBI). Number of reads aligning to each virus was quantified with idxstats tool of SAMtools (v1.9) [89]. Finally, viral alignments were viewed in IGV genome browser [90].

4.3 Analysis of unaligned reads

Unaligned reads were de novo assembled into contigs with SPAdes run in RNA-seq mode (v3.13.0) [91]. Contigs were compressed at 90 % similarity and with word size 7 by CD-HIT-EST (v4.8.1) [71]. Compressed contigs were searched against NCBIs non-redundant nucleotide database (nt) by BLASTN algorithm (v2.9.0) [92] with e-value threshold 1 x 10^{-5}. Those contigs that did not have BLASTN matches were translated in six frames of standard genetic code with Transeq (EMBOSS toolkit v6.5.7.0) [93]. Resulting peptide contigs were searched against NCBIs non-redundant protein database (nr) by BLASTP algorithm (v2.9.0) with e-value threshold 1 x 10^{-5}. The best BLAST search match for each contig was assumed to be the true organism of that contig. Reads corresponding to each contig was quantified by aligning unaligned reads against compressed contigs and running idxstats tool of SAMtools.

4.4 Qualitative analysis of contigs

Latter part of pipeline did not identify viruses at species level or measure their abundances but rather tried to find novel virus sequences. Therefore, its results are called qualitative results. Those peptide contigs that did not have BLASTP matches were searched for remote protein homologs with e-value threshold 1 x 10^{-5} with HMMER (v3.2.1) [72] which used HMM profiles built from vFam [94] (4,156 high-performance profiles, updated February 2014), pVOG [95] (9,518 profiles, updated May 2016) and Pfam [96] (17,929 profiles in version 32, updated 30.8.2018) databases. If a peptide contig matched several profiles, the profile with the highest full-sequence bit score was chosen. Those contigs whose corresponding peptide contigs did not have any HMMER matches were searched for possible viral sequences with machine learning method VirFinder (v1.1) [76] run under model ‘VF.modEPV_k8.rda’. This model had been trained with a positive set of over 7,300 prokaryotic and eukaryotic viruses and with equally sized negative set of prokaryotic hosts (https://github.com/jessieren/VirFinder; 26.2.2020).

Putative viral contig (p-value < 0.05) was qualified only if it had high complexity (DUST score ≤ 7 according to PRINSEQ) and no tandem repeats detected by Tandem Repeat Finder (v4.09 with recommended parameters) [97]. Custom Python scripts and Seqtk (v1.3-r106; https://github.com/lh3/seqtk; 26.2.2020) were utilized in text processing throughout the pipeline.

4.5 Statistical analysis of quantitative viral matches

Viral reads resulting from read alignment and BLAST searches could be assigned to certain species and are therefore called quantitative results. For each virus in each sample, reads aligned to Refseq viruses were summed with reads corresponding viral contig. Following previous study [67], virus was considered detected in a sample if its total read count in the sample was ≥ 5. Then, read count for each virus in each sample was normalized to reads per million quality filtered reads:

$$virus\ abundance = \frac{virus\ reads}{total\ quality\ reads\ in\ sample} \times 10^6$$

First, samples were divided in 3 groups: non-treated RA patients, treated RA patients and healthy controls. Difference between groups was tested with Kruskal-Wallis non-parametric test in three matters: total viral abundance per sample, viral species per sample and virus abundances of individual viruses per sample (Benjamini-Hochberg adjusted p-values). Later, samples were divided in 2 groups: RA patients and healthy controls. Three above mentioned attributes were studied again for difference but this time with asymptotic Wilcoxon-Mann-Whitney test (package coin). All statistical tests were done in R software (v3.6.1; https://www.r-project.org; 26.2.2020).

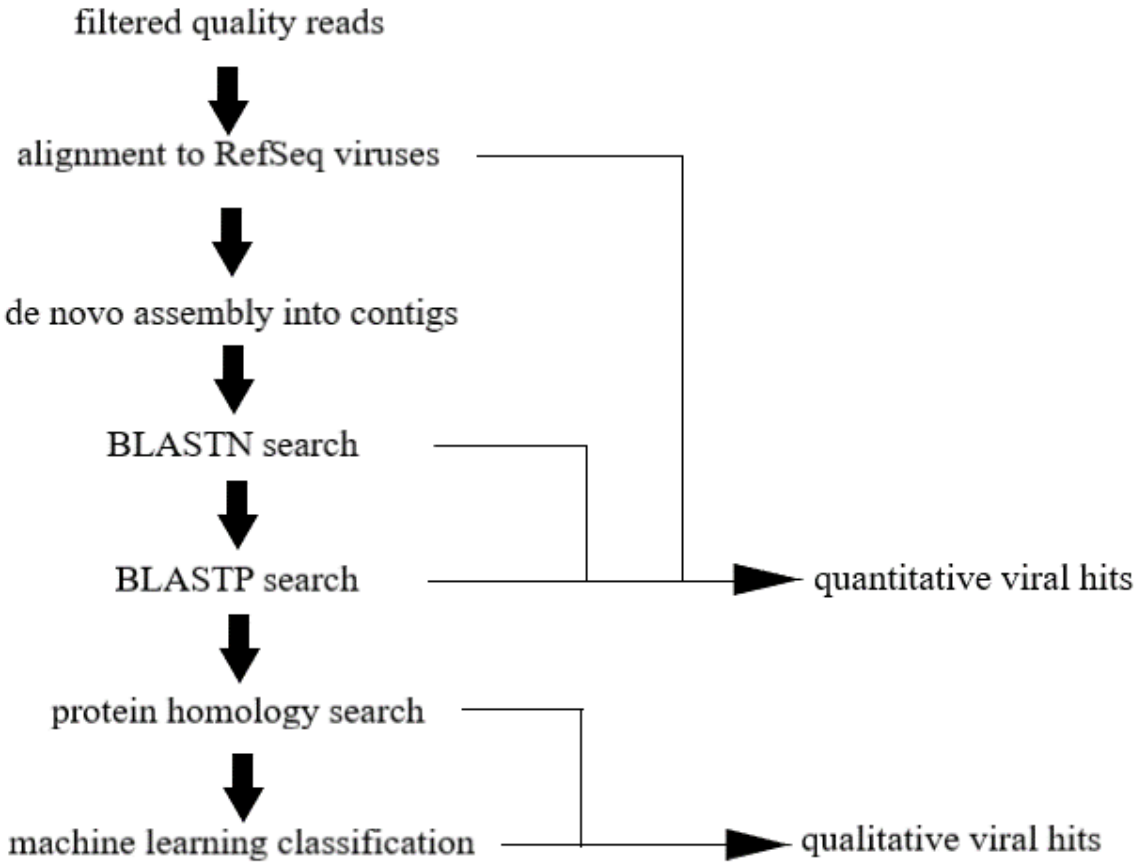

Figure 2. Simplified schematic of the used bioinformatics pipeline. The pipeline was modified from previous study [67]. Please note division into quantitative and qualitative results.

5 Results

5.1 Quantitative results

Viral reads resulting from read alignment and BLAST searches could be assigned to certain species and are therefore called quantitative results. On average, each sample had 11,389,225 reads after quality control. There were 273,341,387 quality reads in total. On average, 83.5 % of reads aligned to human genome, 14.7 % aligned to human transcriptome and 0.38 % aligned to human microbiome genomes. A total of 16,434 reads aligned to viral Refseq genomes. Therefore, there were a total of 10,592,900 unaligned reads available for de novo assembly (3.9 % of original quality reads). On average, each sample had 441,371 unaligned reads available for de novo assembly.

Of all unaligned reads, 50.7 % were able to assemble into contigs. This means that 1.9 % of all quality reads remained unidentified. Out of 452,351 contigs in total, 449,872 remained after similarity compression, of which 8,024 were longer than 1 kilobase. Length of compressed contigs varied from 49 to 7,964 nucleotides and median length was 193 nucleotides (Figure 3). 36 contigs matched viral sequences in BLASTN search (Appendix A). Majority of BLAST searches matched human sequences which was expected from previous study [67]. Some bacterial matches were also detected. 1,630 contigs did not have matches in BLASTN search and these were translated into peptide contigs in six frames (9,780 in total). 23 peptide contigs matched viral proteins in BLASTP search (Appendix B). Viral read counts from read alignment and BLAST searches were combined for each virus and each sample.

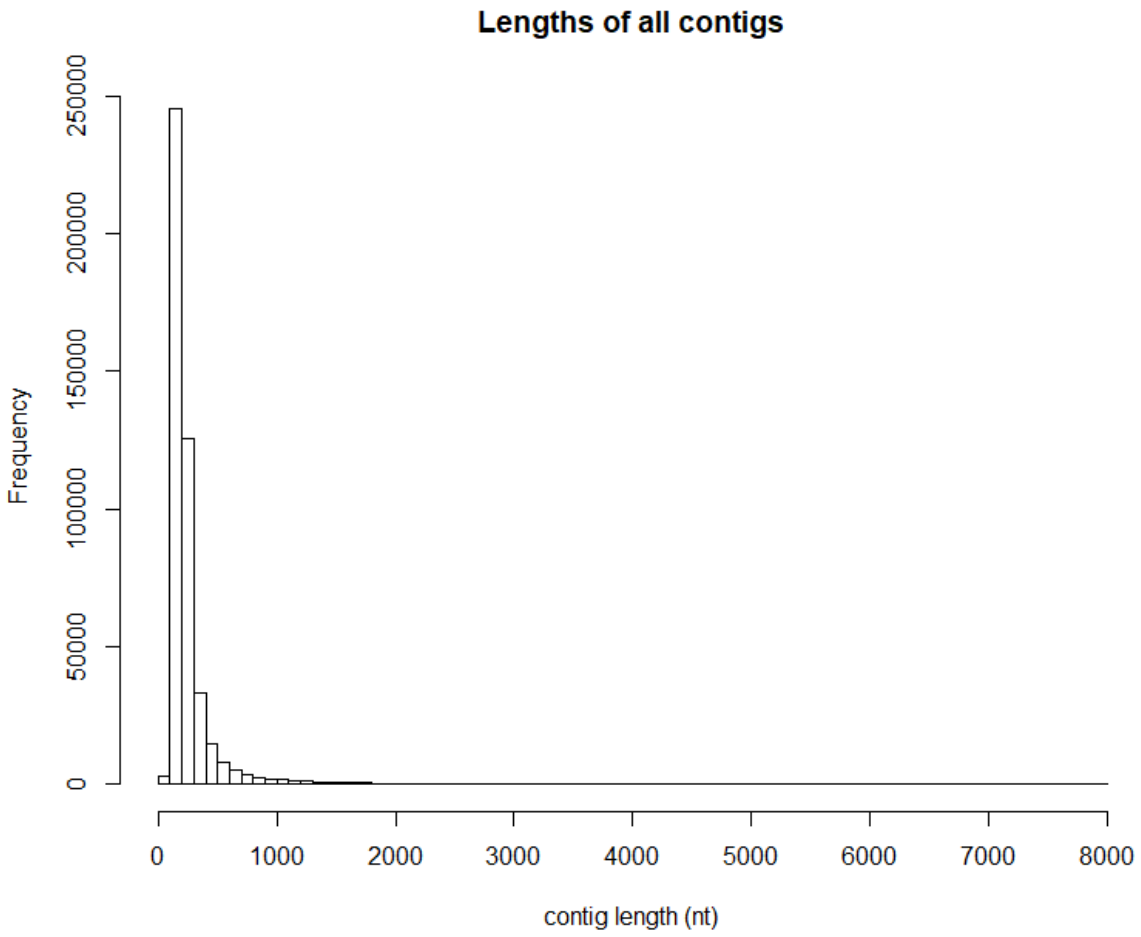

Figure 3. Length distribution of all assembled contigs. SPAdes assembled surprisingly short contigs. Bin size of histogram is 100 nucleotides. Maximum length was 7,964 nucleotides.

With a detection limit of 5 reads in a sample, 13 viruses were detected across all samples. Among these, most abundant ones on average were Proteus phage VB_PmiS-Isfahan and coliphage phi-X174. Six viruses were detected in all samples. In general, bacteriophages dominated virus species (Figure 4).

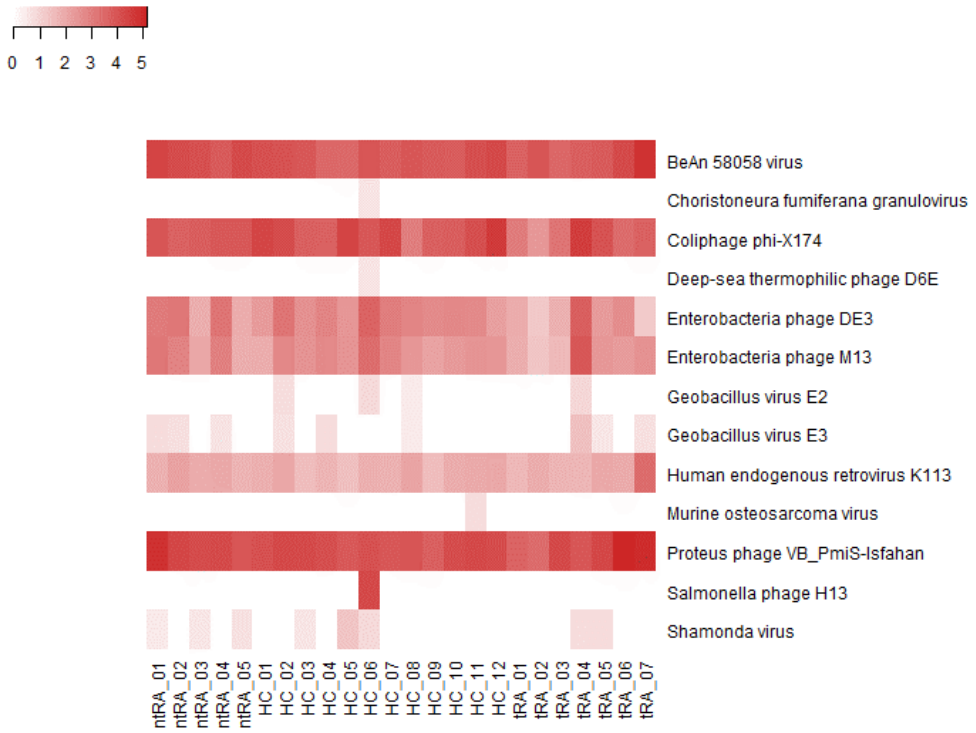

Figure 4. Heatmap of viral abundances. Normalized read counts of detected viruses (see Materials and Methods) were transformed ($\log_2$(abundance+1)) before plotting them as heatmap (R package heatmap3) [98]. HC: healthy control. ntRA: non-treated rheumatoid arthritis patient. tRA: treated rheumatoid arthritis patient.

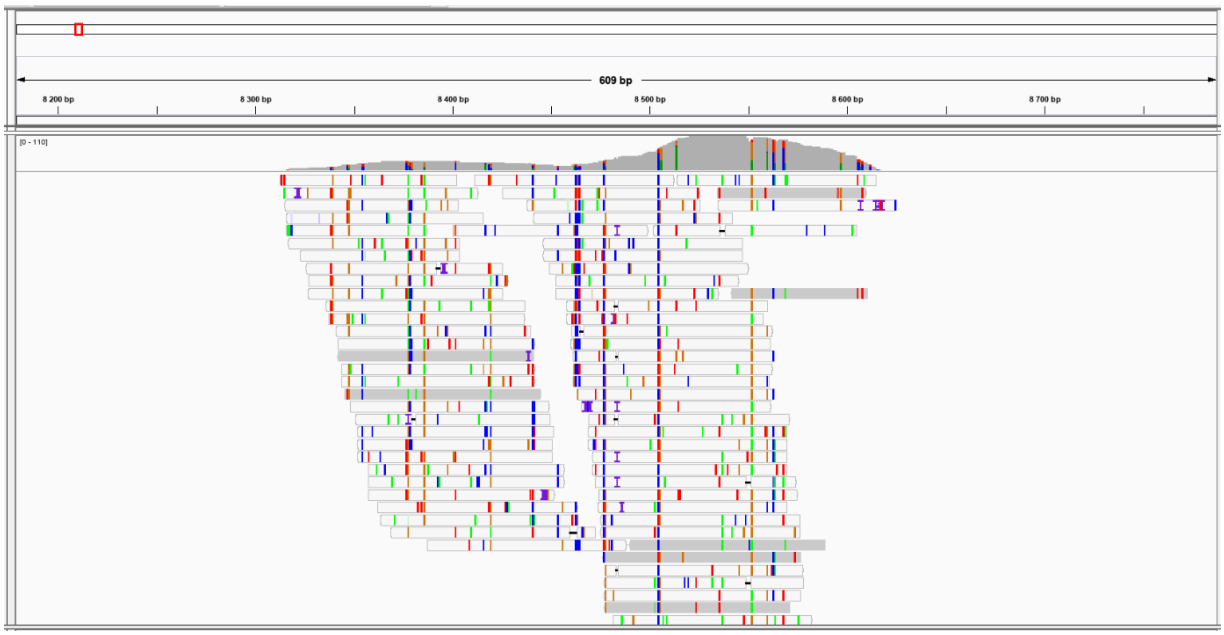

Figure 5. Alignment to BeAn 58058 virus genome in healthy control sample #06. This is an example of poor alignment quality. Transparent reads have MAPQ value of zero.

When comparing 3 groups, there was no significant difference in total viral abundance per individual (Kruskal-Wallis p-value = 0.82) (Figure 6). Also, number of detected viral species for each individual was calculated but these were not significantly different between groups (Kruskal-Wallis p-value = 0.82). In addition, none of the individual viruses had significant difference between groups (Kruskal-Wallis p-values in range 0.13 - 0.72). Outcome did not change if experimental design was changed to comparison of two groups: RA patients versus healthy controls. Neither total viral abundance nor species richness was significantly different between groups (Wilcoxon-Mann-Whitney test p-values 0.86 and 0.98, respectively). Finally, none of the individual viruses had differential abundance between two groups (Wilcoxon-Mann-Whitney p-values in range 0.09 - 0.61).

When viral alignments were inspected in genome browser, many reads had very poor mapping quality (Figure 5). This indicate misalignment. When poor read alignment results were discarded, only 6 out of 13 viruses remained as true positive: coliphage phi-X174, Enterobacteria phage DE3, Enterobacteria phage M13, HERV-K113, Proteus phage VB_PmiS-Isfahan and Salmonella phage H13. This conclusion did not change outcome of statistical tests (p-values in range 0.09-0.90).

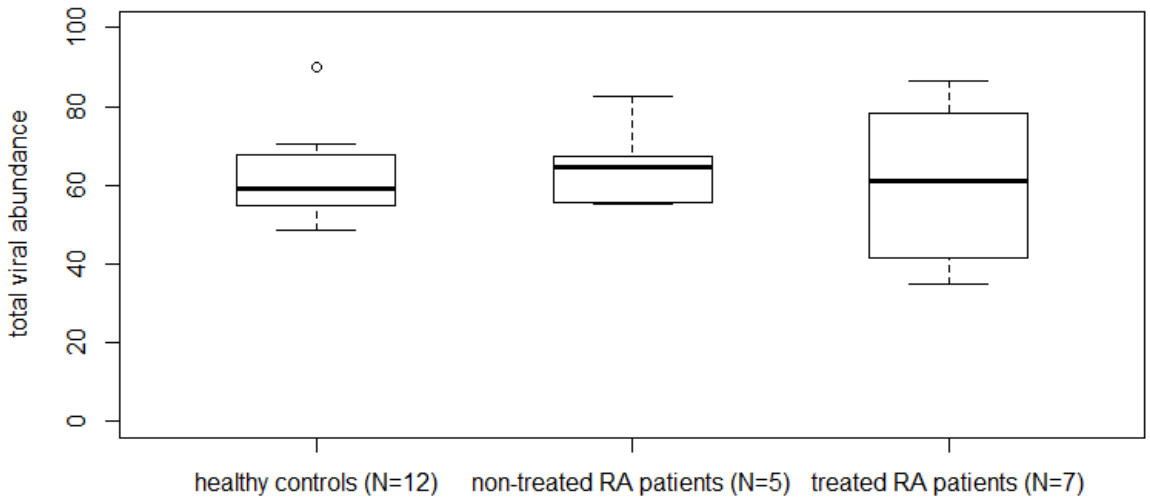

Figure 6. Boxplots of total viral abundance per individual. Three groups were not significantly different.

5.2 Qualitative results

Latter part of pipeline did not identify viruses at species level or measure their abundances but rather tried to find novel virus sequences. Therefore, its results are called qualitative results. Peptide contigs without BLASTP matches (8,585 contigs) were submitted in remote homology search. In total, 33 peptide contigs matched some hidden Markov Model profile but only 15 of these could be regarded as viral. Viral matches were dominated by bacteriophage families such as *Myoviridae* and *Siphoviridae* (Table 2). Nucleotide contigs whose corresponding peptide contigs did not have any matches (1,597 nucleotide contigs) were input into VirFinder. VirFinder suggested 758 contigs to be viral. Only 471 of these qualified after low complexity and tandem repeat filtering. Number of putative viral contigs per sample varied in range 6-164. In total, HMMER and VirFinder suggested 486 viral contigs. Out of all nucleotide contigs, the pipeline left 1,126 with unknown origin.

Table 2. Viral matches from HMMER. HC: healthy control. ntRA: non-treated rheumatoid arthritis patient. tRA: treated rheumatoid arthritis patient.

sample	contig	HMM profile matching the contig	virus families of the profile
ntRA_1	NODE_16622_length_172_cov_0.845528_g16461_i0_3	VOG8053	Myoviridae
ntRA_4	NODE_18986_length_163_cov_0.912281_g18797_i0_2	VOG8284	Myoviridae, Siphoviridae
HC_2	NODE_14479_length_167_cov_0.872881_g14359_i0_2	VOG6036	Myoviridae, Siphoviridae
HC_2	NODE_14537_length_167_cov_0.864407_g14417_i0_6	VOG6472	Siphoviridae
HC_4	NODE_12442_length_171_cov_0.852459_g12276_i0_4	VOG8609	Siphoviridae
HC_5	NODE_7516_length_180_cov_0.793893_g7445_i0_6	VOG5579	Siphoviridae
HC_5	NODE_10428_length_166_cov_0.888889_g10357_i0_4	VOG6472	Siphoviridae
HC_6	NODE_15320_length_164_cov_0.895652_g15187_i0_5	VOG4662	Siphoviridae
HC_6	NODE_9792_length_184_cov_1.488889_g9659_i0_1	vFam_4416	Nyamiviridae
HC_8	NODE_19767_length_162_cov_0.920354_g19609_i0_5	VOG5461	Myoviridae, Siphoviridae
HC_8	NODE_11572_length_188_cov_2.438849_g11414_i0_3	vFam_6570	Phycodnaviridae
tRA_5	NODE_9114_length_193_cov_0.715278_g8942_i0_4	VOG0382	Myoviridae, Siphoviridae
tRA_5	NODE_6670_length_217_cov_0.916667_g6498_i0_4	VOG1026	Bicaudaviridae, Inoviridae, Myoviridae, Siphoviridae
tRA_5	NODE_11048_length_184_cov_0.948148_g10876_i0_5	VOG10310	Myoviridae
tRA_6	NODE_16440_length_165_cov_0.887931_g16300_i0_5	VOG0985	Myoviridae, Podoviridae, Siphoviridae

6 Discussion

This study compared whole blood viral transcriptomes between non-treated RA patients, treated RA patients and healthy controls. No significant difference was found between groups in viral abundance or in number of virus species. Notably, no human exogenous viruses were detected. Absence of anelloviruses and herpesviruses was especially surprising.

It was unexpected to find HERV-K113 because its sequences should have been filtered out during subtraction of host reads as they are part of human genome. When it comes to sequence content, HERVs have long terminal repeats in both ends of their sequences. Large proportion of human genome consists of repetitive elements. It is very challenging for alignment software to map reads correctly on this type of genomic region. Apparently, alignment software left HERV-K113 reads unaligned to human genome and therefore they passed read subtraction phase of the pipeline. HERV-K113 is a member of HML-2 subgroup of HERV-K family [37,99]. It has been estimated that HERV-K113 integrated into human genome within last 5 million years which makes it relatively recent invader. It has active transcription as it retains open reading frame in genes gag, pol and env [37]. One study discovered that HERV-K113 insertion was slightly more prevalent in RA patients than in healthy controls [38]. However, this project did not find such difference.

Other detected viruses than HERV-K113 were bacteriophages. Proteus phage VB_PmiS-Isfahan is a lytic bacteriophage which was recently discovered from sewage water. It is classified under family *Siphoviridae* [100]. Its bacterial host, *Proteus mirabilis*, is part of human gut normal flora [101] but it is also common cause of urinary tract infection [100].

Salmonella phage H13 was found from one individual of control group. This suggests that the individual was asymptomatic carrier of Salmonella bacteria at time of sampling. Phages might have spread from their host bacteria to circulation through gut epithelia (see discussion below).

Viral families of bacteriophages M13, DE3 and phi-X174 are *Inoviridae*, *Siphoviridae* and *Microviridae*, respectively. *Escherichia coli* is a common host of these 3 phages

(https://www.genome.jp/virushostdb; 26.2.2020). *E. coli* is one of the most abundant bacteria in normal flora of gut, but several strains are pathogenic [102].

It could be stated that the discussed bacteriophages are simply contaminants. Bacteria and consequently their bacteriophage parasites can contaminate laboratory equipment. However, many independent sequencing studies have detected the same bacteriophage families in blood. Also, prevalence of many bacteriophages in this study was not sporadic but rather consistent across all samples. It was observed already in 1980s that bacteriophages may penetrate human gut epithelia and travel to bloodstream [103]. Later animal models and in vitro studies support this [104–106]. In addition, Asplund et al. tested a plethora of laboratory equipment in case of contaminating effect in metagenomics studies. They could not associate viruses detected in this project to any laboratory equipment that was used here [107]. It seems therefore that these 5 bacteriophages were, in fact, present in the blood. Indeed, Barr suggested that blood cells could transiently express engulfed phage DNA because phages are known to be efficient transporters of genetic material [106]. This could explain detection of bacteriophage transcripts.

Vaahtovuo et al. observed that RA patients had few bifidobacteria in their faecal microbiota [108]. Thus, it has been suggested that factors like diet, bacterial normal flora and aging could increase permeability of the gut [109,110]. Also, it has been proposed that increased transfer of gut bacteriophages into blood could be the cause of autoimmune diseases [111]. However, it must be noted that RA patients in this project did not differ from healthy controls in their phage abundance or species richness.

Two methods, remote protein homology search and k-mer based classification, were used to find sequences of possible novel viruses. Results of these methods are taken as qualitative results because they obviously do not identify viruses at species level. HMMER results showed that peptide contigs had domains common with bacteriophage families *Bicaudaviridae*, *Inoviridae*, *Myoviridae*, *Podoviridae* and *Siphoviridae*. *Myoviridae*, *Podoviridae* and *Siphoviridae* are classified under the same order, *Caudovirales*. Viruses of this order have distinctive tails and double-stranded DNA genomes [112]. Two contigs from healthy controls matched eukaryotic virus families. *Nyamiviridae* and *Phycodnaviridae* seem random mismatches since known hosts of these viruses are seabirds [113] and algae [114], respectively.

In addition, *Phycodnaviridae* has been associated with contamination from laboratory equipment [107]. Despite this, it seems likely that human blood may still contain unidentified viruses.

The current project has several limitations. First, RNA-sequencing detects only RNA viruses and DNA viruses with active transcription. If a DNA virus has low expression or it is integrated in human genome, RNA-sequencing may not detect it. Secondly, no viral enrichment was performed in laboratory before sequencing. This gives unbiased expression values but viruses with low expression may remain undetected. Thirdly, each sample was sequenced to have 20 million raw reads which is relatively low sequencing depth (https://www.encodeproject.org/about/experiment-guidelines; 26.2.2020). Again, this lowers sensitivity and could explain why anelloviruses or herpesviruses were not detected. Low sequencing depth may also explain why SPAdes software failed to assemble very long contigs. Short contigs make database matching hard and therefore hinder much of the used pipeline [115]. Indeed, BLAST searches with contigs did not contribute much to the final read counts (Appendix A and B). In addition, used pipeline was not able to filter all human reads out. This was demonstrated by BLAST searches where majority of contigs matched human sequences. This was a waste of computational resources. Finally, presence of all viruses should be confirmed with quantitative PCR, which is considered to be the gold standard method in virus detection [66,67]. Use of another method could lower chance of false positives. However, acquiring original samples and PCR testing them were out of scope of this thesis. All in all, this project shows us that virome sequencing is a very delicate process.

Although detected viruses did not correlate with RA disease state, it does not mean they would be biologically insignificant. They could affect immune response which could be seen in overall gene expression. Gene ontology analysis could illuminate this issue. Also, it could be interesting to reveal virome of other organs if the pipeline were enhanced. In the case of RA, synovium and synovial fluid could be interesting research topics.

7 Conclusion

Bioinformatics pipeline modified from earlier study was used to measure viral transcriptome in whole blood of non-treated RA patients, treated RA patients and healthy controls. Publicly available RNA-sequencing dataset was utilized. The pipeline detected many bacteriophages. However, no significant difference was found between groups in total viral abundance per individual, in abundances of individual viruses or in number of species. This project supports views that bacteriophages of gastrointestinal tract could transfer to bloodstream, but they do not seem to associate with rheumatoid arthritis. In addition, 486 contiguous sequences suggested presence of novel viruses in human blood.

8 References

1. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch S V, Knight R (2018) Current understanding of the human microbiome. *Nat Med* **24**: 392–400.

2. Virgin HW, Wherry EJ, Ahmed R (2009) Redefining Chronic Viral Infection. *Cell* **138**: 30–50.

3. Zarate S, Taboada B, Yocupicio-Monroy M, Arias CF (2017) Human Virome. *Arch Med Res* **48**: 701–716.

4. Krishnamurthy SR, Wang D (2017) Origins and challenges of viral dark matter. *Virus Res* **239**: 136–142.

5. van der Woude D, van der Helm-van Mil AHM (2018) Update on the epidemiology, risk factors, and disease outcomes of rheumatoid arthritis. *Best Pract Res Clin Rheumatol* **32**: 174–187.

6. Smatti MK, Cyprian FS, Nasrallah GK, Al Thani AA, Almishal RO, Yassine HM (2019) Viruses and autoimmunity: A review on the potential interaction and molecular mechanisms. *Viruses* **11**: 762.

7. Halenius A, Hengel H (2014) Human cytomegalovirus and autoimmune disease. *Biomed Res Int* **2014**: 472978.

8. Balandraud N, Roudier J (2018) Epstein-Barr virus and rheumatoid arthritis. *Jt Bone Spine* **85**: 165–170.

9. Shchetynsky K, Diaz-Gallo L-M, Folkersen L, Hensvold AH, Catrina AI, Berg L, Klareskog L, Padyukov L (2017) Discovery of new candidate genes for rheumatoid arthritis through integration of genetic association data with expression pathway analysis. *Arthritis Res Ther* **19**: 19.

10. Murphy K, Mowat A, Weaver C (2012) *Janeway's immunobiology*. Garland Science, New York.

11. Hayter SM, Cook MC (2012) Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmun Rev* **11**: 754–765.

12. Fry L, Baker BS, Powles A V, Engstrand L (2015) Psoriasis is not an autoimmune disease? *Exp Dermatol* **24**: 241–244.

13. DiMeglio LA, Evans-Molina C, Oram RA (2018) Type 1 diabetes. *Lancet (London, England)* **391**: 2449–2462.

14. Popp A, Maki M (2019) Changing Pattern of Childhood Celiac Disease Epidemiology: Contributing Factors. *Front Pediatr* **7**: 357.

15. Bano A, Pera A, Almoukayed A, Clarke THS, Kirmani S, Davies KA, Kern F (2019) CD28null CD4 T-cell expansions in autoimmune disease suggest a link with cytomegalovirus infection. *F1000Research* **8**: 327.

16. Croia C, Serafini B, Bombardieri M, Kelly S, Humby F, Severa M, Rizzo F, Coccia EM, Migliorini P, Aloisi F, et al. (2013) Epstein-Barr virus persistence and infection of autoreactive plasma cells in synovial lymphoid structures in rheumatoid arthritis. *Ann Rheum Dis* **72**: 1559–1568.

17. Wang L, Wang F-S, Gershwin ME (2015) Human autoimmune diseases: a comprehensive update. *J Intern Med* **278**: 369–395.

18. Sumitomo S, Nagafuchi Y, Tsuchida Y, Tsuchiya H, Ota M, Ishigaki K, Suzuki A, Kochi Y, Fujio K, Yamamoto K (2018) Transcriptome analysis of peripheral blood from patients with rheumatoid arthritis: a systematic review. *Inflamm Regen* **38**: 21.

19. Nielen MMJ, van Schaardenburg D, Reesink HW, van de Stadt RJ, van der Horst-Bruinsma IE, de Koning MHMT, Habibuw MR, Vandenbroucke JP, Dijkmans BAC (2004) Specific autoantibodies precede the symptoms of rheumatoid arthritis: a study of serial measurements in blood donors. *Arthritis Rheum* **50**: 380–386.

20. Sakkas LI, Daoussis D, Liossis S-N, Bogdanos DP (2017) The Infectious Basis of ACPA-Positive Rheumatoid Arthritis. *Front Microbiol* **8**: 1853.

21. Willis VC, Gizinski AM, Banda NK, Causey CP, Knuckley B, Cordova KN, Luo Y, Levitt B, Glogowska M, Chandra P, et al. (2011) N-alpha-benzoyl-N5-(2-chloro-1-iminoethyl)-L-ornithine amide, a protein arginine deiminase inhibitor, reduces the severity of murine collagen-induced arthritis. *J Immunol* **186**: 4396–4404.

22. Frisell T, Saevarsdottir S, Askling J (2016) Family history of rheumatoid arthritis: an old concept with new developments. *Nat Rev Rheumatol* **12**: 335–343.

23. Klareskog L, Stolt P, Lundberg K, Källberg H, Bengtsson C, Grunewald J, Rönnelid J, Erlandsson Harris H, Ulfgren A-K, Rantapää-Dahlqvist S, et al. (2006) A new model for an etiology of rheumatoid arthritis: Smoking may trigger HLA–DR (shared epitope)–restricted immune reactions to autoantigens modified by citrullination. *Arthritis Rheum* **54**: 38–46.

24. Klein J, Sato A (2000) The HLA system. First of two parts. *N Engl J Med* **343**: 702–709.

25. Yang H, Biermann MH, Brauner JM, Liu Y, Zhao Y, Herrmann M (2016) New Insights into Neutrophil Extracellular Traps: Mechanisms of Formation and Role in Inflammation. *Front Immunol* **7**: 302.

26. Yue D, Brintnell W, Mannik LA, Christie DA, Haeryfar SMM, Madrenas J, Chakrabarti S, Bell DA, Cairns E (2010) CTLA-4Ig blocks the development and progression of citrullinated fibrinogen-induced arthritis in DR4-transgenic mice. *Arthritis Rheum* **62**: 2941–2952.

27. Cross M, Smith E, Hoy D, Carmona L, Wolfe F, Vos T, Williams B, Gabriel S, Lassere M, Johns N, et al. (2014) The global burden of rheumatoid arthritis: estimates from the global burden of disease 2010 study. *Ann Rheum Dis* **73**: 1316–1322.

28. Moustafa A, Xie C, Kirkness E, Biggs W, Wong E, Turpaz Y, Bloom K, Delwart E, Nelson KE, Venter JC, et al. (2017) The blood DNA virome in 8,000 humans. *PLoS Pathog* **13**: e1006292–e1006292.

29. Shamriz O, Shoenfeld Y (2018) Infections: a double-edge sword in autoimmunity. *Curr Opin Rheumatol* **30**: 365–372.

30. Strachan DP (1989) Hay fever, hygiene, and household size. *BMJ* **299**: 1259–1260.

31. Fujinami RS, von Herrath MG, Christen U, Whitton JL (2006) Molecular mimicry, bystander activation, or viral persistence: infections and autoimmune disease. *Clin*

Microbiol Rev **19**: 80–94.

32. Pane JA, Webster NL, Coulson BS (2014) Rotavirus Activates Lymphocytes from Non-Obese Diabetic Mice by Triggering Toll-Like Receptor 7 Signaling and Interferon Production in Plasmacytoid Dendritic Cells. *PLOS Pathog* **10**: e1003998.

33. McCall KD, Thuma JR, Courreges MC, Benencia F, James CBL, Malgor R, Kantake N, Mudd W, Denlinger N, Nolan B, et al. (2015) Toll-like receptor 3 is critical for coxsackievirus B4-induced type 1 diabetes in female NOD mice. *Endocrinology* **156**: 453–461.

34. Nagata K, Kumata K, Nakayama Y, Satoh Y, Sugihara H, Hara S, Matsushita M, Kuwamoto S, Kato M, Murakami I, et al. (2017) Epstein-Barr Virus Lytic Reactivation Activates B Cells Polyclonally and Induces Activation-Induced Cytidine Deaminase Expression: A Mechanism Underlying Autoimmunity and Its Contribution to Graves' Disease. *Viral Immunol* **30**: 240–249.

35. Nanbo A, Inoue K, Adachi-Takasawa K, Takada K (2002) Epstein-Barr virus RNA confers resistance to interferon-alpha-induced apoptosis in Burkitt's lymphoma. *EMBO J* **21**: 954–965.

36. Goupil BA, Mores CN (2016) A Review of Chikungunya Virus-induced Arthralgia: Clinical Manifestations, Therapeutics, and Pathogenesis. *Open Rheumatol J* **10**: 129–140.

37. Trela M, Nelson PN, Rylance PB (2016) The role of molecular mimicry and other factors in the association of Human Endogenous Retroviruses and autoimmunity. *APMIS* **124**: 88–104.

38. Krzysztalowska-Wawrzyniak M, Ostanek M, Clark J, Binczak-Kuleta A, Ostanek L, Kaczmarczyk M, Loniewska B, Wyrwicz LS, Brzosko M, Ciechanowicz A (2011) The distribution of human endogenous retrovirus K-113 in health and autoimmune diseases in Poland. *Rheumatology (Oxford)* **50**: 1310–1314.

39. Virtanen JO, Jacobson S (2012) Viruses and multiple sclerosis. *CNS Neurol Disord Drug Targets* **11**: 528–544.

40. Sotelo J, Corona T (2011) Varicella Zoster Virus and Relapsing Remitting Multiple Sclerosis. *Mult Scler Int* **2011**: 214763.

41. Ball RJ, Avenell A, Aucott L, Hanlon P, Vickers MA (2015) Systematic review and meta-analysis of the sero-epidemiological association between Epstein-Barr virus and rheumatoid arthritis. *Arthritis Res Ther* **17**: 274.

42. Klatt T, Ouyang Q, Flad T, Koetter I, Bühring H-J, Kalbacher H, Pawelec G, Müller CA (2005) Expansion of peripheral CD8+ CD28- T cells in response to Epstein-Barr virus in patients with rheumatoid arthritis. *J Rheumatol* **32**: 239–251.

43. Balandraud N, Meynard JB, Auger I, Sovran H, Mugnier B, Reviron D, Roudier J, Roudier C (2003) Epstein-Barr virus load in the peripheral blood of patients with rheumatoid arthritis: accurate quantification using real-time polymerase chain reaction. *Arthritis Rheum* **48**: 1223–1228.

44. Lunemann JD, Frey O, Eidner T, Baier M, Roberts S, Sashihara J, Volkmer R, Cohen JI, Hein G, Kamradt T, et al. (2008) Increased frequency of EBV-specific effector memory CD8+ T cells correlates with higher viral load in rheumatoid arthritis. *J*

Immunol **181**: 991–1000.

45. Alspaugh MA, Henle G, Lennette ET, Henle W (1981) Elevated Levels of Antibodies to Epstein-Barr Virus Antigens in Sera and Synovial Fluids of Patients with Rheumatoid Arthritis. *J Clin Invest* **67**: 1134–1140.

46. Sherina N, Hreggvidsdottir HS, Bengtsson C, Hansson M, Israelsson L, Alfredsson L, Lundberg K (2017) Low levels of antibodies against common viruses associate with anti-citrullinated protein antibody-positive rheumatoid arthritis; implications for disease aetiology. *Arthritis Res Ther* **19**: 219.

47. Scotet E, David-Ameline J, Peyrat MA, Moreau-Aubry A, Pinczon D, Lim A, Even J, Semana G, Berthelot JM, Breathnach R, et al. (1996) T cell response to Epstein-Barr virus transactivators in chronic rheumatoid arthritis. *J Exp Med* **184**: 1791–1800.

48. Cornillet M, Verrouil E, Cantagrel A, Serre G, Nogueira L (2015) In ACPA-positive RA patients, antibodies to EBNA35-58Cit, a citrullinated peptide from the Epstein–Barr nuclear antigen-1, strongly cross-react with the peptide β60-74Cit which bears the immunodominant epitope of citrullinated fibrin. *Immunol Res* **61**: 117–125.

49. Johansson L, Pratesi F, Brink M, Arlestig L, D'Amato C, Bartaloni D, Migliorini P, Rantapaa-Dahlqvist S (2016) Antibodies directed against endogenous and exogenous citrullinated antigens pre-date the onset of rheumatoid arthritis. *Arthritis Res Ther* **18**: 127.

50. Roudier J, Petersen J, Rhodes GH, Luka J, Carson DA (1989) Susceptibility to rheumatoid arthritis maps to a T-cell epitope shared by the HLA-Dw4 DR beta-1 chain and the Epstein-Barr virus glycoprotein gp110. *Proc Natl Acad Sci* **86**: 5104–5108.

51. Dag MS, Turkbeyler IH, Ozturk ZA, Kısacık B, Tutar E, Kadayıfçı A (2015) Cytomegalovirus ileocolitis in a rheumatoid arthritis patient: case report and literature review. *Reumatismo* **67**: 13–16.

52. Pierer M, Rothe K, Quandt D, Schulz A, Rossol M, Scholz R, Baerwald C, Wagner U (2012) Association of anticytomegalovirus seropositivity with more severe joint destruction and more frequent joint surgery in rheumatoid arthritis. *Arthritis Rheum* **64**: 1740–1749.

53. Pawelec G, McElhaney JE, Aiello AE, Derhovanessian E (2012) The impact of CMV infection on survival in older humans. *Curr Opin Immunol* **24**: 507–511.

54. Goronzy JJ, Matteson EL, Fulbright JW, Warrington KJ, Chang-Miller A, Hunder GG, Mason TG, Nelson AM, Valente RM, Crowson CS, et al. (2004) Prognostic markers of radiographic progression in early rheumatoid arthritis. *Arthritis Rheum* **50**: 43–54.

55. Thewissen M, Somers V, Venken K, Linsen L, van Paassen P, Geusens P, Damoiseaux J, Stinissen P (2007) Analyses of immunosenescent markers in patients with autoimmune disease. *Clin Immunol* **123**: 209–218.

56. Pera A, Caserta S, Albanese F, Blowers P, Morrow G, Terrazzini N, Smith HE, Rajkumar C, Reus B, Msonda JR, et al. (2018) CD28(null) pro-atherogenic CD4 T-cells explain the link between CMV infection and an increased risk of cardiovascular death. *Theranostics* **8**: 4509–4519.

57. Spencer J V, Cadaoas J, Castillo PR, Saini V, Slobedman B (2008) Stimulation of B lymphocytes by cmvIL-10 but not LAcmvIL-10. *Virology* **374**: 164–169.

58. Michelson S, Alcami J, Kim SJ, Danielpour D, Bachelerie F, Picard L, Bessia C, Paya C, Virelizier JL (1994) Human cytomegalovirus infection induces transcription and secretion of transforming growth factor beta 1. *J Virol* **68**: 5730–5737.

59. Broccolo F, Drago F, Cassina G, Fava A, Fusetti L, Matteoli B, Ceccherini-Nelli L, Sabbadini MG, Lusso P, Parodi A, et al. (2013) Selective reactivation of human herpesvirus 6 in patients with autoimmune connective tissue diseases. *J Med Virol* **85**: 1925–1934.

60. Kholodnyuk, Kadisa, Svirskis, Gravelsina, Studers, Spaka, Sultanova, Lejniece, Lejnieks, Murovska (2019) Proportion of the CD19-Positive and CD19-Negative Lymphocytes and Monocytes within the Peripheral Blood Mononuclear Cell Set is Characteristic for Rheumatoid Arthritis. *Medicina (B Aires)* **55**: 630.

61. Freimanis G, Hooley P, Ejtehadi HD, Ali HA, Veitch A, Rylance PB, Alawi A, Axford J, Nevill A, Murray PG, et al. (2010) A role for human endogenous retrovirus-K (HML-2) in rheumatoid arthritis: investigating mechanisms of pathogenesis. *Clin Exp Immunol* **160**: 340–347.

62. Nelson PN, Roden D, Nevill A, Freimanis GL, Trela M, Ejtehadi HD, Bowman S, Axford J, Veitch AM, Tugnet N, et al. (2014) Rheumatoid arthritis is associated with IgG antibodies to human endogenous retrovirus gag matrix: a potential pathogenic mechanism of disease? *J Rheumatol* **41**: 1952–1960.

63. Spandole S, Cimponeriu D, Berca LM, Mihăescu G (2015) Human anelloviruses: an update of molecular, epidemiological and clinical aspects. *Arch Virol* **160**: 893–908.

64. Giacconi R, Maggi F, Macera L, Pistello M, Provinciali M, Giannecchini S, Martelli F, Spezia PG, Mariani E, Galeazzi R, et al. (2018) Torquetenovirus (TTV) load is associated with mortality in Italian elderly subjects. *Exp Gerontol* **112**: 103–111.

65. Maggi F, Andreoli E, Riente L, Meschi S, Rocchi J, Delle Sedie A, Vatteroni ML, Ceccherini-Nelli L, Specter S, Bendinelli M (2007) Torquetenovirus in patients with arthritis. *Rheumatology* **46**: 885–886.

66. Kramná L, Kolářová K, Oikarinen S, Pursiheimo J-P, Ilonen J, Simell O, Knip M, Veijola R, Hyöty H, Cinek O (2015) Gut virome sequencing in children with early islet autoimmunity. *Diabetes Care* **38**: 930–933.

67. Li G, Zhou Z, Yao L, Xu Y, Wang L, Fan X (2019) Full annotation of serum virome in Chinese blood donors with elevated alanine aminotransferase levels. *Transfusion* **59**: 3177–3185.

68. Zhang H, Morrison S, Tang Y-W (2015) Multiplex polymerase chain reaction tests for detection of pathogens associated with gastroenteritis. *Clin Lab Med* **35**: 461–486.

69. Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ, Ian Lipkina W (2015) Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *MBio* **6**: e01491-15.

70. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

71. Li W, Fu L, Niu B, Wu S, Wooley J (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief Bioinform* **13**: 656–668.

72. Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**: 431.

73. Simmonds P, Xia W, Baillie JK, McKinnon K (2013) Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla--selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. *BMC Genomics* **14**: 610.

74. Kapoor A, Simmonds P, Lipkin WI, Zaidi S, Delwart E (2010) Use of Nucleotide Composition Analysis To Infer Hosts for Three Novel Picorna-Like Viruses. *J Virol* **84**: 10322–10328.

75. Zhang Z, Cai Z, Tan Z, Lu C, Jiang T, Zhang G, Peng Y (2019) Rapid identification of human-infecting viruses. *Transbound Emerg Dis* **66**: 2517–2522.

76. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**: 69.

77. Seguritan V, Alves Jr. N, Arnoult M, Raymond A, Lorimer D, Burgin Jr. AB, Salamon P, Segall AM (2012) Artificial Neural Networks Trained to Detect Viral and Phage Structural Proteins. *PLOS Comput Biol* **8**: e1002657.

78. Ounit R, Wanamaker S, Close TJ, Lonardi S (2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16**: 236.

79. Breitwieser FP, Baker DN, Salzberg SL (2018) KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol* **19**: 198.

80. Li S-K, Leung RK-K, Guo H-X, Wei J-F, Wang J-H, Kwong K-T, Lee S-S, Zhang C, Tsui SK-W (2012) Detection and identification of plasma bacterial and viral elements in HIV/AIDS patients in comparison to healthy adults. *Clin Microbiol Infect* **18**: 1126–1133.

81. Kim KW, Horton JL, Pang CNI, Jain K, Leung P, Isaacs SR, Bull RA, Luciani F, Wilkins MR, Catteau J, et al. (2019) Higher abundance of enterovirus A species in the gut of children with islet autoimmunity. *Sci Rep* **9**: 1749.

82. Furuta RA, Sakamoto H, Kuroishi A, Yasiui K, Matsukura H, Hirayama F (2015) Metagenomic profiling of the viromes of plasma collected from blood donors with elevated serum alanine aminotransferase levels. *Transfusion* **55**: 1889–1899.

83. Chivero ET, Stapleton JT (2015) Tropism of human pegivirus (formerly known as GB virus C/hepatitis G virus) and host immunomodulation: insights into a highly successful viral infection. *J Gen Virol* **96**: 1521–1532.

84. Stremlau MH, Andersen KG, Folarin OA, Grove JN, Odia I, Ehiane PE, Omoniwa O, Omoregie O, Jiang P-P, Yozwiak NL, et al. (2015) Discovery of Novel Rhabdoviruses in the Blood of Healthy Individuals from West Africa. *PLoS Negl Trop Dis* **9**: e0003631.

85. Dinakaran V, Rathinavel A, Pushpanathan M, Sivakumar R, Gunasekaran P, Rajendhran J (2014) Elevated Levels of Circulating DNA in Cardiovascular Disease Patients: Metagenomic Profiling of Microbiome in the Circulation. *PLoS One* **9**: e105221.

86. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863–864.

87. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, et al. (2017) Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**: 61–66.

88. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**: D61-5.

89. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

90. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**: 178–192.

91. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD (2019) rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**: giz100.

92. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

93. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

94. Skewes-Cox P, Sharpton TJ, Pollard KS, DeRisi JL (2014) Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS One* **9**: e105067.

95. Grazziotin AL, Koonin E V, Kristensen DM (2017) Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res* **45**: D491–D498.

96. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**: D279-85.

97. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

98. Zhao S, Guo Y, Sheng Q, Shyr Y (2014) Advanced heat map and clustering analysis using heatmap3. *Biomed Res Int* **2014**: 986048.

99. Garcia-Montojo M, Doucet-O'Hare T, Henderson L, Nath A (2018) Human endogenous retrovirus-K (HML-2): a comprehensive review. *Crit Rev Microbiol* **44**: 715–738.

100. Yazdi M, Bouzari M, Ghaemi EA (2019) Genomic analyses of a novel bacteriophage (VB_PmiS-Isfahan) within Siphoviridae family infecting Proteus mirabilis. *Genomics* **111**: 1283–1291.

101. Oduyebo OO, Odugbemi TO, Idewu A, Adefule-Ositelu A, Aibinu IE, Ogunro A (2010) Incidence of postoperative eye infections in a private eye hospital in Lagos, Nigeria. *Nig Q J Hosp Med* **20**: 138–143.

102. Vogt RL, Dippold L (2005) Escherichia Coli O157:H7 Outbreak Associated with Consumption of Ground Beef, June–July 2002. *Public Health Rep* **120**: 174–178.

103. Weber-Dabrowska B, Dabrowski M, Slopek S (1987) Studies on bacteriophage penetration in patients subjected to phage therapy. *Arch Immunol Ther Exp (Warsz)* **35**: 563–568.

104. Nguyen S, Baker K, Padman BS, Patwa R, Dunstan RA, Weston TA, Schlosser K, Bailey B, Lithgow T, Lazarou M, et al. (2017) Bacteriophage transcytosis provides a mechanism to cross epithelial cell layers. *MBio* **8**: e01874-17.

105. Navarro F, Muniesa M (2017) Phages in the Human Body. *Front Microbiol* **8**: 566.

106. Barr JJ (2017) A bacteriophages journey through the human body. *Immunol Rev* **279**: 106–122.

107. Asplund M, Kjartansdottir KR, Mollerup S, Vinner L, Fridholm H, Herrera JAR, Friis-Nielsen J, Hansen TA, Jensen RH, Nielsen IB, et al. (2019) Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clin Microbiol Infect* **25**: 1277–1285.

108. Vaahtovuo J, Munukka E, Korkeamaki M, Luukkainen R, Toivanen P (2008) Fecal microbiota in early rheumatoid arthritis. *J Rheumatol* **35**: 1500–1505.

109. Hurme M (2019) Viruses and immunosenescence – more players in the game. *Immun Ageing* **16**: 13.

110. Fasano A (2012) Leaky gut and autoimmune diseases. *Clin Rev Allergy Immunol* **42**: 71–78.

111. Tetz G, Tetz V (2018) Bacteriophages as New Human Viral Pathogens. *Microorganisms* **6**: 54.

112. Sutton TDS, Hill C (2019) Gut Bacteriophage: Current Understanding and Challenges. *Front Endocrinol (Lausanne)* **10**: 784.

113. Mihindukulasuriya KA, Nguyen NL, Wu G, Huang H V, da Rosa APAT, Popov VL, Tesh RB, Wang D (2009) Nyamanini and midway viruses define a novel taxon of RNA viruses in the order Mononegavirales. *J Virol* **83**: 5109–5116.

114. Yolken RH, Jones-Brando L, Dunigan DD, Kannan G, Dickerson F, Severance E, Sabunciyan S, Talbot Jr CC, Prandovszky E, Gurnon JR, et al. (2014) Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. *Proc Natl Acad Sci U S A* **111**: 16106–16111.

115. Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C (2019) Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **7**: 12.

9 Appendices

Appendix A. Viral BLASTN matches. HC: healthy control. ntRA: non-treated rheumatoid arthritis patient. tRA: treated rheumatoid arthritis patient.

sample	contig	match	read count
ntRA_01	NODE_474_length_1019_cov_2.192784_g405_i0	Escherichia virus phiX174, complete genome	40
ntRA_01	NODE_16902_length_171_cov_0.852459_g16741_i0	Geobacillus virus E3, complete genome	2
ntRA_02	NODE_4750_length_313_cov_2.306818_g4512_i0	Escherichia virus phiX174, complete genome	11
ntRA_03	NODE_255_length_1206_cov_1.897148_g204_i0	Escherichia virus phiX174, complete genome	42
ntRA_03	NODE_24030_length_163_cov_1.184211_g23880_i0	Escherichia virus phiX174, complete genome	3
ntRA_04	NODE_12491_length_183_cov_1.664179_g12302_i0	Escherichia virus phiX174, complete genome	4
ntRA_04	NODE_14869_length_175_cov_0.817460_g14680_i0	Deep-sea thermophilic phage D6E, complete genome	2
ntRA_05	NODE_22137_length_166_cov_0.837607_g21929_i0	Geobacillus virus E3, complete genome	2
HC_02	NODE_9398_length_187_cov_0.739130_g9278_i0	Deep-sea thermophilic phage D6E, complete genome	2
HC_04	NODE_3020_length_283_cov_1.871795_g2854_i0	Escherichia virus phiX174, complete genome	8
HC_04	NODE_11807_length_174_cov_0.824000_g11641_i0	Geobacillus virus E3, complete genome	2
HC_05	NODE_4529_length_204_cov_2.548387_g4458_i0	Enterobacteria phage phiX174 isolate 10A90, complete genome	8
HC_06	NODE_2538_length_303_cov_5.944882_g2405_i0	Salmonella phage H13, complete genome	27

HC_06	NODE_5926_length_215_cov_1.481928_g5793_i0	Escherichia virus phiX174, complete genome	5
HC_06	NODE_11503_length_177_cov_0.804688_g11370_i0	Deep-sea thermophilic phage D6E, complete genome	2
HC_06	NODE_12700_length_172_cov_1.105691_g12567_i0	Geobacillus virus E2, complete genome	2
HC_06	NODE_16296_length_155_cov_33.094340_g16163_i0	Salmonella phage H13, complete genome	151
HC_07	NODE_3821_length_244_cov_2.523077_g3698_i0	Escherichia virus phiX174 strain evolved J1, complete genome	11
HC_07	NODE_5018_length_220_cov_2.479532_g4895_i0	Escherichia virus phiX174, complete genome	8
HC_08	NODE_15839_length_173_cov_0.838710_g15681_i0	Deep-sea thermophilic phage D6E, complete genome	2
HC_08	NODE_15840_length_173_cov_0.838710_g15682_i0	Geobacillus virus E3, complete genome	2
HC_08	NODE_19977_length_162_cov_0.884956_g19819_i0	Enterobacteria phage phiX174 isolate 10B90, complete genome	2
HC_09	NODE_745_length_567_cov_1.664093_g636_i0	Escherichia virus phiX174, complete genome	18
HC_10	NODE_12895_length_170_cov_0.859504_g12802_i0	Geobacillus virus E3, complete genome	2
HC_11	NODE_2019_length_365_cov_3.481013_g1865_i0	Proteus phage VB_PmiS-Isfahan, complete genome	9
HC_11	NODE_15258_length_171_cov_0.836066_g15103_i0	Geobacillus virus E3, complete genome	2
HC_12	NODE_330_length_840_cov_2.250316_g289_i0	Escherichia virus phiX174, complete genome	32
tRA_01	NODE_7548_length_206_cov_1.936306_g7397_i0	Escherichia virus phiX174, complete genome	7

tRA_01	NODE_13632_length_172_cov_1.747967_g13481_i0	Escherichia virus phiX174, complete genome	4
tRA_02	NODE_23235_length_168_cov_1.773109_g22988_i0	Escherichia virus phiX174, complete genome	5
tRA_03	NODE_13147_length_171_cov_1.491803_g13014_i0	Escherichia virus phiX174, complete genome	5
tRA_04	NODE_7205_length_182_cov_0.766917_g7139_i0	Geobacillus virus E3, complete genome	2
tRA_05	NODE_755_length_631_cov_1.951890_g614_i0	Enterobacteria phage phiX174 isolate 10A90, complete genome	24
tRA_06	NODE_3617_length_287_cov_1.025210_g3477_i0	Escherichia virus phiX174, complete genome	6
tRA_07	NODE_19136_length_166_cov_0.888889_g18957_i0	Geobacillus virus E3, complete genome	2
tRA_07	NODE_19976_length_164_cov_0.904348_g19797_i0	Geobacillus virus E3, complete genome	2

Appendix B. Viral BLASTP matches. HC: healthy control. ntRA: non-treated rheumatoid arthritis patient. tRA: treated rheumatoid arthritis patient.

sample	**contig**	**match**	**read count**
ntRA_01	NODE_11913_length_189_cov_0.714286_g11752_i0_1	putative DNA primase [Geobacillus virus E3]	2
ntRA_01	NODE_13973_length_181_cov_0.780303_g13812_i0_6	single-stranded-DNA-specific exonuclase RecJ [Bacillus phage PBC2]	2
ntRA_01	NODE_18458_length_167_cov_0.872881_g18297_i0_4	hypothetical protein BCD7_0090 [Bacillus phage BCD7]	2
ntRA_02	NODE_16796_length_186_cov_0.751825_g16558_i0_2	DNA ligase [Bacillus phage vB_BcoS-136]	2

ntRA_04	NODE_9479_length_197_cov_0.959459_g9290_i0_4	hypothetical protein E3_0103 [Geobacillus virus E3]	3
ntRA_04	NODE_10505_length_191_cov_0.732394_g10316_i0_3	site-specific tyrosine recombinase [Bacillus phage pW2]	2
ntRA_04	NODE_18155_length_165_cov_0.896552_g17966_i0_1	ATP-dependent Clp protease proteolytic subunit 2 [Bacillus phage vB_BcoS-136]	2
ntRA_04	NODE_18711_length_164_cov_0.878261_g18522_i0_1	putative structural protein [Bacillus phage pW2]	2
HC_01	NODE_15069_length_169_cov_0.858333_g14958_i0_6	putative DNA gyrase [Geobacillus virus E3]	2
HC_02	NODE_9144_length_188_cov_0.748201_g9024_i0_2	hypothetical protein EalM132_00086 [Exiguobacterium phage vB_EalM-132]	2
HC_03	NODE_14629_length_176_cov_0.811024_g14473_i0_2	putative single-stranded DNA binding protein [Bacillus phage PBC2]	2
HC_03	NODE_16767_length_169_cov_0.825000_g16611_i0_4	putative nucleotide-binding protein [Brevibacillus phage Sundance]	2
HC_06	NODE_9958_length_184_cov_0.748148_g9825_i0_4	hypothetical protein PBC2_213 [Bacillus phage PBC2]	2
HC_06	NODE_11500_length_177_cov_0.804688_g11367_i0_6	deoxynucleoside kinase [Bacillus phage PBC2]	2
HC_06	NODE_11942_length_175_cov_0.825397_g11809_i0_5	putative holiday junction	2

		resolvase [Bacillus phage vB_BspM_MarvelLand]	
HC_07	NODE_14639_length_162_cov_0.902655_g14516_i0_3	hypothetical protein PBC2_242 [Bacillus phage PBC2]	2
HC_09	NODE_10565_length_181_cov_0.780303_g10449_i0_2	hypothetical protein PBC2_077 [Bacillus phage PBC2]	2
tRA_02	NODE_16102_length_189_cov_0.707143_g15855_i0_1	hypothetical protein, partial [Bacillus phage phiKir1]	2
tRA_02	NODE_22217_length_171_cov_0.844262_g21970_i0_5	tyrosine recombinase [Bacillus phage vB_BcoS-136]	2
tRA_02	NODE_24055_length_167_cov_0.864407_g23808_i0_1	hypothetical protein vBBcoS136_00253 [Bacillus phage vB_BcoS-136]	2
tRA_03	NODE_11063_length_180_cov_0.786260_g10930_i0_5	DNA ligase [Bacillus phage vB_BcoS-136]	2
tRA_03	NODE_15148_length_165_cov_0.853448_g15015_i0_5	DNA polymerase [Salmonella phage Astrid]	2
tRA_04	NODE_6632_length_185_cov_0.764706_g6566_i0_4	major head protein [Bacillus phage 0305phi8-36]	2