Characterizing viral diversity in Respiratory tract infection:

Emphasizing on Microarray technology as genomic sensor in clinical diagnosis

By

Saadia Noreen

A thesis presented to Medical faculty of Norwegian University of Science and Technology in partial fulfillment of requirement for the degree of MSc.

Molecular Medicine

Acknowledgments:

Working on this master Project was an overwhelming experience as through this I learned with how to write papers and proposals, give talks, work in a group, stay up until the birds start singing, stay focus and DNA microarray technology. My first gratitude is to my Allah who always helped me. **Allah bless me always.** Than I would like to say thanks to many people for huge reasons. It seems impossible to finish in one page. Because I cant resist my digressions.

It takes me back to one year , when **Dr. Jorn klein** announced this project as last minute call and one of my friend named jo bhai showed me this call. When I saw this call I was so happy. The day was more precious to me when I was selected for this project among many students. So I am deeply grate ful to Jorn Klein my supervisor for having a trust on me. You have supported me, in difficulties, and always listen to my ideas in discussions, and expending those with your key insights. You taught me how to work independently, and to find ways. You are an appropriate guider. I cant forget your words never give up, and keep your head up, one step after the other.

I have been very privileged to get to know and to collaborate with many other people like **Torfinn.** I learned a lot of techniques from you. Your technical excellence and tremendous grasp of experimental issues had a great impact on me. You teach me about technical ground work which I cant forget. Your help means a lot to me. Thank you Torfinn. I would like to Thank **Hilde** for helping me in initial experiments. I have greatly enjoyed the opportunity to work with **Konika,** whose mathematical and system biological insights are impressive. A big thank to you Konika for bearing me in your home with silly questions, always irritates you. But you always replied to that. Thanks for being friendly to me. You gave me understandings of data analysis which significantly improved my work.

My dearest thanks to my family, **my parents,** my **big brother Asif**, my sweet bhabi **Amna**, my nieces, **Zoya and Zemal** for their support.

I would like to thank to **Towfiq** for proof reading my thesis, taking out grammar mistakes. I would like to thanks **Jo bhai, Naila, Lindi** for their constant support. You guys always encouraged me. Thanks for that.

Finally I will not say thank to my sister, **my best friend my baji Riffat** . Because according to me this thanks word seems small in front of her. I will just say. **I LOVE YOU**. You are solution of all my problems. You are the ONE AND ONLY person whom I would like to dedicate my thesis.

Contents

Lists of Figures

List of tables:

Abstract

Background

Acute respiratory tract infection is common illness of human with significant morbidity and mortality. In pediatrics, viruses are the major cause if this illness. There is an imperative need to develop a diagnostic tool to measure viral diversity for preventing contraproductive treatments. This present study focuses on evaluating viruses from clinical samples of respiratory tract infection by using advanced diagnostic method such as microarray technology.

Methods

Target was amplified using random amplification. Indirect method of hybridization was used to fluorescently label target with Cy3. A previously developed LLMDA subarray 2 (GPL13407) was demonstrated as detectichip. This chip comprise of 58,000 probes. The detectichip was designed by Agilent technologies. Samples were hybridized on this chip. The resulting fluorescent produce after hybridization was explored and digitized using gene pix pro software. Data was normalized with two methods named as 1) within array control method 2) with whole negative control array. Log 2 fold change was calculated. Significant testing was also performed. Detecti V software was used to perform these tasks.

Results:

Detected viral species were arranged according to their log2 fold change. The higher log 2 fold change indicated the abundance of viral species in sample. More over graphs and figures were also drawn to indicate the detection of viral species. Significant testing indicates the presence of high level viral families according to their p-value and t testing.

Conclusion:

Detectichip can successfully characterize viruses frequently found in clinical samples. By applying both normalization method it can be stated that this detectichip able to identify broad spectrum of viral family, viral species, bacteriophages, plant virus. The likely viral RTI were adenovirus , influenza virus, rhino virus. The rare virus associated with RTI was human papilloma virus and mammalian orthoreovirus. This chip confirms that in sample 1 adenoviridae family was significantly present where as in sample 2 the likely specie is Influenza .

1-Introduction

1.1. Respiratory tract Infections

Acute respiratory tract infections (RTI) are leading causes of childhood morbidity and mortality worldwide, resulting in nearly 2 million deaths annually[3-5]. The worldwide disease burden is estimated at almost 98,000,000 disability adjusted life years (DALYs) and more than 4,000,000 deaths per year[6]. Even in the Western World, RTI contributes more to the disease burden and health care costs higher than any other infectious disease. Despite its abundance and potential severity, we still have limited information about causes, pathophysiology and the best ways to treat various RTI. In children, more than three-fourth of all RTI is caused by viruses[7,9,10] and today over 200 various respiratory viruses are known. Nevertheless, in recent studies on childhood RTI, despite the use of comprehensive diagnostic approaches more than one fourth of respiratory samples are virus negative. Thus, it is likely that more respiratory viruses do exist, but they may escape detection from present diagnostic tests.

Respiratory infections can be termed according to the area of involvement, namely upper respiratory tract infections and lower respiratory tract infections. In children with RTI, a likely pathogen (virus or bacteria) may be detected in respiratory secretions from the upper airways[15]. Initial symptoms of upper respiratory tract infection are coryza, cough and hoarseness[16]. Rhinitis, laryngitis, pharyngitis (sore throat) are commonly observed. These are infections of nasopharynx, larynx, tensile, sinuses, ears[17]. Lower respiratory tract infections refers to infections of the trachea, bronchus, and alveolus, results tracheitis (causing hoarseness, painful breathing in and out[17]), bronchitis, bronchiolitis, bronchopneumonia asthma exacerbations, acute otitis media as shown in figure below[18]. Lower respiratory infections are generally more severe than upper respiratory infection[17]. e.g. when infection results in necrosis of epithelial lining the bronchioles, it can spread to lungs causes bronchiolitis and pneumonia[17].

Virus infection starts when sufficient viruses are available, site of infection is permissive to the virus and local host antiviral defense system is ineffective. Viral pathogenesis is the result of molecular interactions between the cell and virus. N-acetyl neuraminic acid, glycosaminoglycans and glycolipids, ICAM integrin receptors and molecules of the Major Histocompatibility complex are involved in interactions with epithelial cells of human respiratory tract.

Some viruses require co-receptor molecules to penetrate the cell e.g Adenovirus. Respiratory viruses are mainly transmitted by direct contact with the contaminated secretions[19,20], aerosols or fomites (objects or surface. Most viruses that infect humans enter into the body through the respiratory tract as aerosols produced by coughing. Sneezing is a source of discharging droplets containing virus particles and leading to initiate the infection. Large viral particles are usually trapped in the turbinates and sinuses and could cause upper respiratory infections. Smaller viral particles can reach the alveolar spaces and cause infections in the lower respiratory[11].

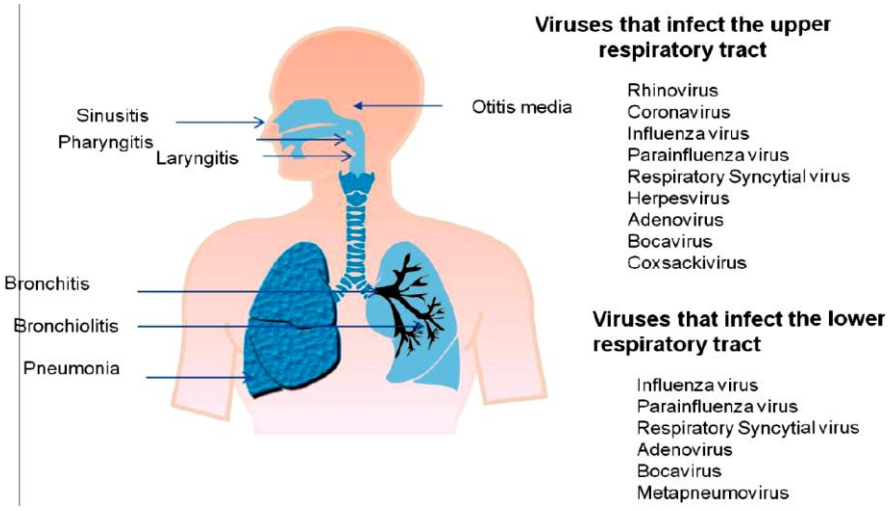


Figure 1: The Respiratory tract infection caused by viruses: Viruses can infect the upper respiratory tract occasionally some of them can cause infections in the lower respiratory tracts. Others enter through the respiratory tracts but they move to other organs. This figure is adapted from [11].

1.1.1. Respiratory Tract Infection pathogens and prevalence of infection:

Several bacteria, viruses, fungus, and other pathogens are the source of respiratory tract infections (RIs)[21]. Influenza virus type A, B, C, Para influenza virus type 1,2,3, Respiratory syncytial virus (RSV), Adenovirus and Rhino virus are the most common cause of RTI[22]. Laryngitis is mostly caused by Para influenza virus[17].

In recent studies about pediatrics respiratory infections, more than one-fourth of respiratory samples are virus negative. Earlier studies also hypothesized that no etiological agent can be identified in ~30% of patients suffering from respiratory disease due to unknown viruses[6]. With

regards to prevalence of viruses in several western countries the data revealed that adeno virus 6 strains (A-F) prevalence is approximately 2-14%[23]. Human metapneumovirus have been discovered in the Netherlands in 2001[24], affecting mainly children less than 5 years[25] and causing both upper and lower respiratory tract infection[26]. Corona virus accounts for 5-30% virus infection. Human Bocavirus was discovered in Sweden in 2005 with prevalence approximately 2-11%[27].

Table 1: Respiratory tract Infections (RTIs) pathogens: Studies reflects the mechanical disorder due to major viruses, families, types. Majority of the respiratory tract infection viruses are RNA viruses.

Virus family	Viral agents	Virus types	Nucleic acid	Disease type
Orthomyxoviridae	Influenza A and B	Several	RNA	Acute Bronchitis
	Parainfluenza	1,2,3,4	RNA	Laryngitis
	Respiratory syncytial virus	1	RNA	Bronchiolitis, pneumonia
Picornaviridae	Rhinovirus	Several	RNA	Bronchitis
	Enterovirus		RNA	
	Coxakievirus (24 types)	A21	RNA	Upper respiratory infection
	Echovirus(34 types)	11,20	RNA	
Coronaviridae	Human corona virus	0C43	RNA	Bronchitis
	Human corona virus	229E	RNA	Bronchitis
	Human corona virus	HKU1	RNA	Bronchitis
Adenoviridae	Adeneovirus(41type)	5-10 types	DNA	Pharyngitis, Bronchitis
Parvoviridae	Bocavirus		DNA	
	Polyomavirus WU/K1		DNA	
	Parvovirus		DNA	Lower respiratory infection
Paramyxoviridae	Human metapneumovirus		RNA	

This table is modified from[17,22,28] .

1.2. Conventional diagnostics methods

Respiratory tract infection viruses diagnosis began in 1933[29]. *In vitro culturing methods* have first been used to diagnose respiratory virus infection[22]. Propagation of viruses in culture is

visualized by morphological changes termed as cytopathic effect[9]. These effects can be observed by microscopy.

Primary monkey kidney cell lines are usually used for isolating respiratory viruses[9] (vero cells). Some viruses are not culturable in vitro, i.e human bocavirus and human corona virus[6]. It might take 10 days or more to develop a cytopathic effect. Delay in achieving results make tests laborious to work with[16].

Serological methods are based on the rise in antibody concentration (usually increase IgG responses) in infected host[6,16,28,30]. Hemagglutination test (HIT), complement fixation and immunoassays (EIA) belong to those serological methods. Two weeks are usually required to develop antibody response in infected host[16]. Antigen detection methods are particularly useful for those viruses which require long time to propagate in cell culture such as respiratory syncytial virus, parainfluenza viruses, adenoviruses. However these methods are not useful for detecting those viral species that have antigenic heterogeneity e.g. rhinovirus[9].

Serological methods give results rapidly, e.g. direct florescent antibody staining tests (DFA) give results in 3 hours. In this DFA technique fluorescein labeled antibody detect pathogenic antigen and immunofluorescence pattern can be observed with microscopy[19]. In general, those conventional methods have low sensitivity and specificity and hollowistic view on viral flora is not give. These methods do not show global picture[31].

Microscopic diagnosis is also useful but to some extent. It is used to diagnose inclusion bodies. But the presence of these inclusion bodies is not true for all respiratory tract infection viruses. Electron microscopy is useful as it reveals presence of virus particles in patient and investigating corona viruses. This is also helpful in identifying viruses from formalin fixed tissues if fresh tissue is not available. Tissue preparation and increased detection time are disadvantages[28] .

1.3. Molecular diagnostic Methods:

Molecular diagnostic methods are able to identify nucleic acids of specific pathogens in specimen[30]. Molecular methods are important as these methods deals with the genes. Whereas conventional methods works on basis of phenotypic characteristics.

The development of polymerase chain reaction provides a rapid and sensitive method, but requires for testing specific pathogens that we know before. This technique allows the amplification of single or multiple copies of DNA to produce thousands of copies of DNA by an enzyme-reaction (shown in figure 2). The major disadvantage of PCR is that it is unable to detect novel viruses.

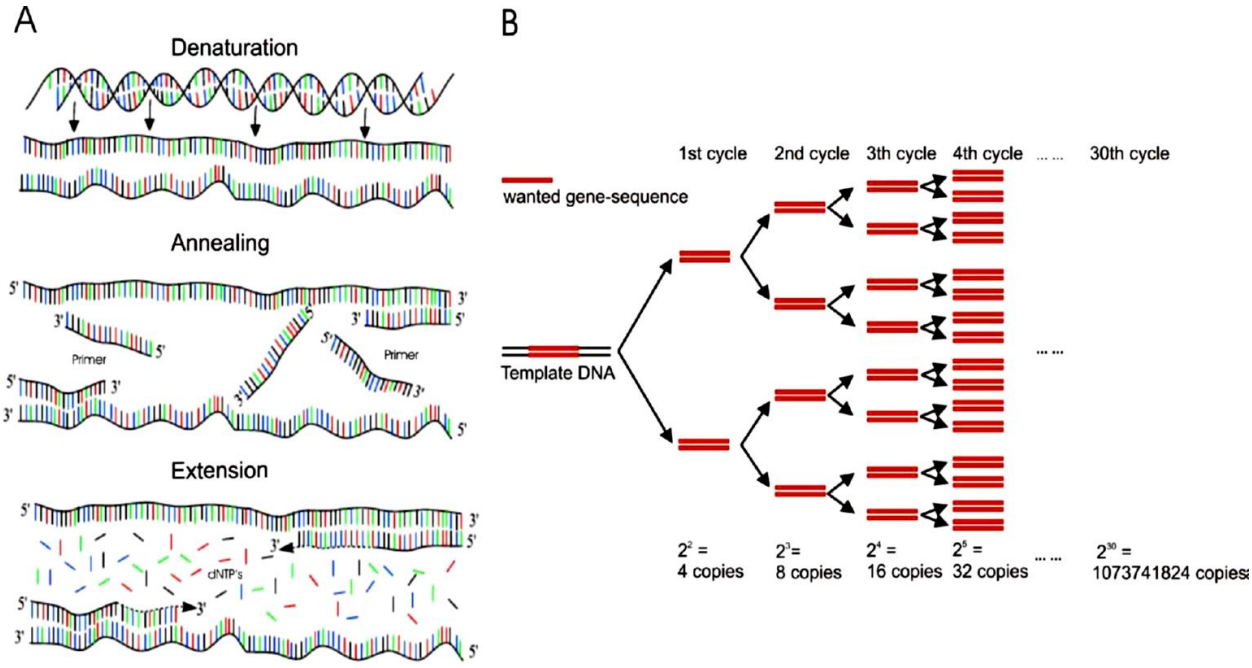


Figure 2: Principle of polymerase chain reaction (A) Major steps of PCR (B) Exponential growth of amplicon in the PCR amplification[32].

PCR is based on the thermal denaturation of double stranded DNA and the specific annealing of oligonucleotides (primers) at the 5'- and 3' end of the target DNA area. These oligonucleotides become elongated in the extension phase by a DNA-dependent DNA polymerase, using free deoxynucleoside-triphosphate (dNTP) .This process take place in a thermal cycler and the temperature profile is repeated until enough copies of the target sequences are produced, e.g. 30 cycles produce theoretically Each PCR must be adapted to the actual application. Some important parameters are:

• **Primers**

To find the approbate primers for the target is a crucial point for using PCR as detection method. The first step is to find a conserved target sequence of a defined length unique for the organism. Primers are usually 12 – 50 bases in length. Depending on the base composition of the primers the annealing temperature (TA) is calculated. The optimal TA is defined approximately 5 °C lower than melting temperature (TM) of the oligonucleotides[33]

TM = 4 • (G+C) + 2 • (A+T) only valid for oligonucleotides up to 20 bases !

• **Polymerase**

A broad variety of available DNA polymerases exists and can be chosen depending on the used PCR technique. For PCR as detection method, a thermostable polymerase with 3' → 5' nuclease activity, allowing the replacement of incorrectly integrated nucleotides (proofreading function), is useful. To avoid unspecific annealing and polymerase activity at room temperature, the use of a so called Hot-start polymerase is an advantage. Hot start polymerases become active only after an activation step of 10 minutes at 95 °C.

•**Number of cycles**

The number of cycles necessary to obtain a sufficient amount of an amplicon depends on the concentration of the DNA template. Usually 30 – 40 cycles are enough to get a good detectable result The extreme sensitivity of PCR may also be a disadvantage, because of the ability to generate millions of copies from one single template. Thus, a contaminating DNA template can produce a false positive result. The risk of contamination can be reduced by the physical separation of the reaction preparation from the area of reaction product analysis, use of negative controls, use of pipette tips with barriers preventing contamination.

1.3.1. Other PCR methods:

In *Reverse transcriptase PCR* RNA target first reverse transcribed in to cDNA.

Nested PCR uses two pairs of primers for two rounds of PCR.

If multiple primer are used in PCR amplifying different targets then it is known as *multiplex PCR*[34].

Real time PCR can be modified as quantitative PCR as the sample which has high a viral load will be amplified sooner as compared to sample with low viral load[34]. Real time PCR allows more sensitive quantitative detection by florescence detection technology.

FRET (fluorescent resonance energy transfer) probes, dual hybridization probes that are sequence specific probes are used in Real time PCR which are mixed with PCR reaction mixture[35]. PCR amplification can be monitored as the probes react with the PCR product to give an increase in the fluorescence signal[34].

Random PCR uses a primer that has unique universal sequence at 5' and at the 3' end contain hexa or heptamer sequence. A second primer is complimentary to first primer universal sequence [34].

1.3.2. Other nucleic acid based methods

Loop mediated isothermal amplification is a technology that does not require thermal cycling. Amplification of sample can be carried out at a single specific temperature.

Sanger sequencing is the conventional method of sequencing DNA. It is based on chain termination method. The Basic principle is that in four separate reaction mixtures, DNA template , dNTPs, DNA polymerase, and dideoxy nucleotide triphosphate is added. The later would terminate the synthesis of corresponding fluorescently labeled dNTP results termination of extension [36].

1.3.3. Requirement of new diagnostic method:

Some viruses escape from the detection with diagnostic method. There is a need of new diagnostic method.

Next generation sequencing techniques reveals unbiased and in-depth information but is expensive[37,38]. High throughput method such as 454 (roche), Solexa (Illumina), or SOLid (Life technologies are useful for metagenomic analysis. Millions of unbiased sequences can be obtained[38]

Another hightroughput technology is microarray technology, which will be described in deep below.

1.4. Microarray technology as advanced diagnostic method

Nucleic acid based methods complementing conventional methods, such as immunological assays, biochemical and culture based assays. However those conventional assays are

monoparametric (determining single parameter). In contrast to that microarray technology, performs a multiparametric assays[39].

The first report on detecting viruses by using microarray technology was published in 2002[40]. Southern blot and dot blots are known as radio labeled macroarrays. Those *macroarrays* comprise larger size spots of 300 microns, where as in microarray sample a spot is less than 200 microns[31] A microarray is a collection of specific DNA probes on miniaturized device to produce either quantitative data or qualitative data[31,37,41]. Thousands of molecules are analyzed in a high throughput fashion to see global picture of system under study[31]. The position of particular probes is designated as features or spot[42].

Genotyping to analyze the genomic DNA for characterizing viruses can be done with the help of microarray technology and can therefore be used as genomic sensors[43].

Microarrays are divided into several categories depending upon density, number of probes, length of probes, number of dyes, and type of target as well as on the material used for depositing probes.

High density arrays contain thousands to millions of spots. Those arrays are not yet used for daily laboratory practices.

Low density microarrays contain few hundred to thousands probes and are used for in vitro diagnostics of viral pathogens[39,44].

Double stranded DNA microarray comprise of PCR product, or cDNA which have high sensitivity and low specificity and lack palindromic sequences[41]. *Oligonucleotide microarrays* consist of 25 to 80 nucleotides rather than the whole gene[43].

Protein microarray, tissue microarray, cell microarray are other types of array[31] but DNA microarray array is the most widely used type[42].

Microarray technology focuses on high throughput analysis for pathogen detection.[45], identifying pathogen viability[45], enumerating pathogenic agents[45], antimicrobial resistance monitoring and strain typing. Therefore microarrays are good diagnostic platforms[41] for public health medicine and veterinary diagnostics[45]. Microarrays occupies a place between low cost multiplex PCR and

expensive high throughput sequencing technologies[37]. It was reported that microarray identified pathogen with 94% accuracy (76% sensitivity and 100%specificity) in 36 patient specimen[46]. Cost of a microarray is approximately 100-500 dollars[38].

The target molecule is labeled with fluorescent dyes and gives a signal upon hybridization[47]. Indirect fluorochrome labeling increases the sensitivity of detection[40](details are given in labeling sample section) . Microarray were scanned at 5 um resolution through agilent scanner[46].

Arrays can be divided into subarrays allows testing of multiple samples on one slide. Duplicate spots are also act as control[43]. Spotting same probe several times on array is a kind of replicating the array[42]. The replication is important to have back up in form of duplicate spots and for improving data quality[42]. In this present study 2 glass slides contain two subarrays each with 105000 probes.

Sequence recovery is required for obtaining novel virus sequences for emerging infectious diseases[48]. It can be done by putting 100 µl of water at 90°C to array. Recovered eluate amplified with specific primer, cloned with plasmid vector. After transformation colonies are screened by sequencing[49].

1.4.1. Design of microarray experiments:

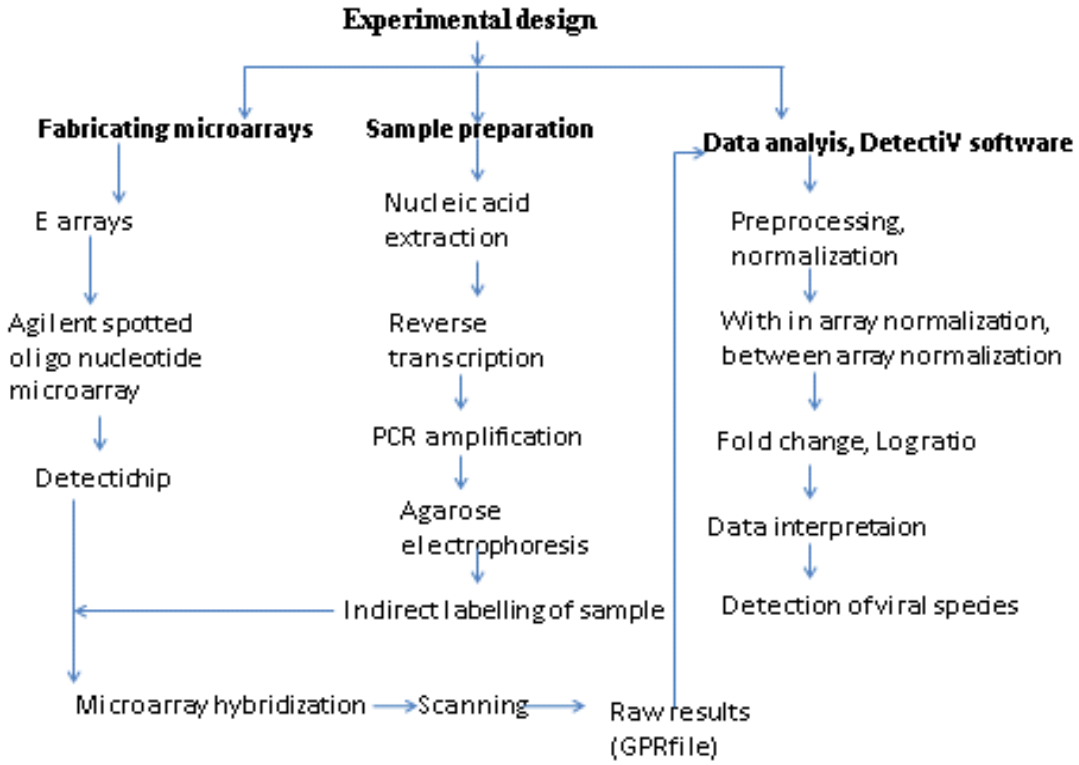


Figure 3: Flow of microarray experiment: Begins with a microarray design. In first step microarray chips are printed. Sample is extracted labeled with fluorescent dye , hybridized to microarray. It is than scanned to acquire fluorescent image. Image analysis is performed to get raw signal data. Raw data is further normalized. Statistical tests are performed and meaningful data is interpreted.

1.4.2. Probes characteristics

Probes are short nucleic acids with a known sequence and identity[50]**.** These sequences are frequently selected from nucleotide sequences found in the 3'end of the transcript[47] . There are two types of probes , gene probes and oligonucleotide probes. Gene probes are longer than 500 nucleotides and require a whole gene sequence which is obtained from Genbank, EMBL or DDBJ sources. Longer probes provides more sensitivity and increase hybridization strength but are less specific [41] because they might bind randomly to potential target sequences. Appropriate probe length is critical to determine [41].

Oligonucleotide probes targeting specific sequences with in a gene. They detect genes with slight differences in sequences. This will prevent hybridization to closely related sequences[51] .Usually 0.1-1nano liter probe is spotted on the slide[47].

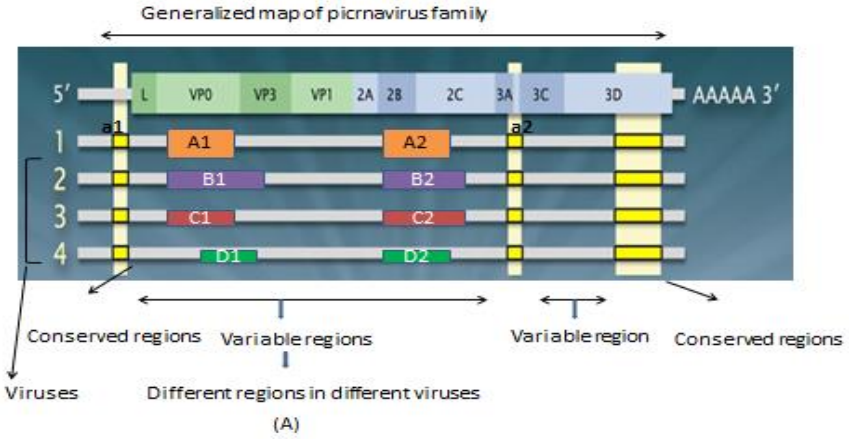
Conserved regions (highly similar sequences) are generally used for probe design. The reason is that probes specific for a given genus but not with other genera can discriminate between the intended target and all other target[52]

Oligonucleotide probes should have 40–60% G-C, without complementary regions and repeats of more than four single base should be avoided. For probe designing amino acid sequence conservation is also utilized[53].

1.4.3. Logic of finding known and novel virus on viro chip:

Viral genome consists of variable and conserved regions. The latter have not changed much during evolution, whereas variable region are constantly changed. Variable regions are used for distinguishing viral strain of same species[12].

Virochip detect known and novel viruses by conserve sequence homology. It was reported that probes designed with conserved sequences within a pathogen genus or family and help in determining those pathogen in disease of ambiguous etiology[40] and reveal novel viruses within the same family[37].



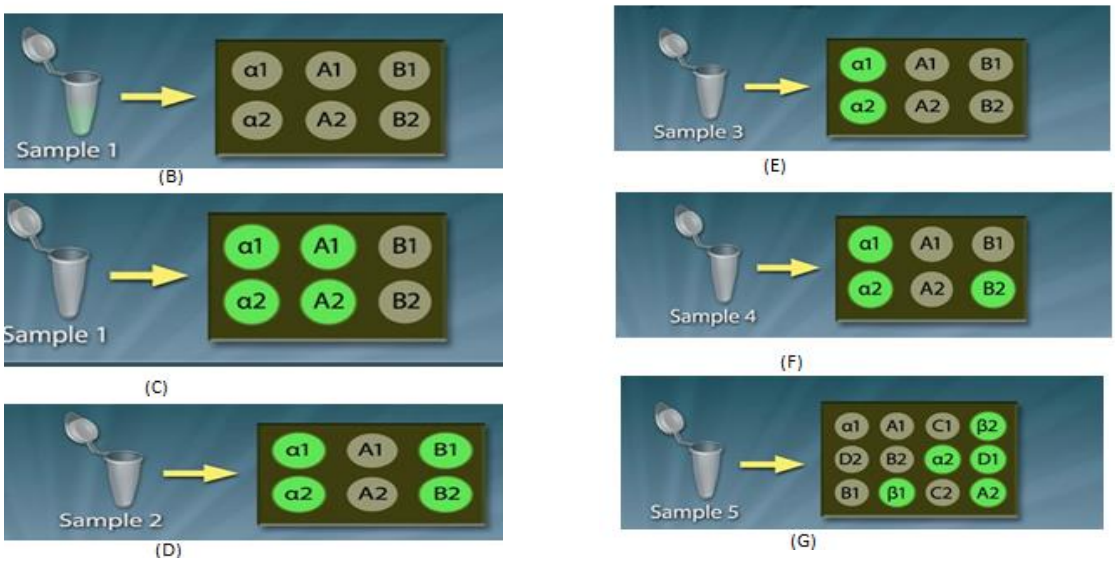


Figure 4: Concept of finding known and unknown viruses using probes sequences: Figure A: indicates generalized map of a virus family represented by 4 virus species. Each virus species have conserved and variable regions. 6 spots in figure B-F represents probe of unique sequences. **Figure C:** After putting unknown sample 1 on these spots, green fluorescence will be given after activation by laser. One can conclude that unknown sample is Virus A because it contain family a1 conserved sequences and virus A variable. **Figure D**: But if unknown sample shows a type of hybridization in which a1, a2, B1, B2 sequences show fluorescence it means this time sample contain virus type B. **Figure E**: If the hybridization pattern gives fluorescence at a1, a2 which are the conserved sequences present in known viruses but variable sequence did not match any known viruses and indicate that virus is new member of the same family. **Figure F**: If the fluorescence is present in a1, a2, B2 position than it can be concluded that sample contain virus of same family but it will be a new virus that is more closely related to virus B but not a virus B **Figure G:** If expanding the spots on chip to 12 and if the fluorescence is observed on a2, D1, b2, b1. Results indicate a new virus formed between recombination between virus A and D . These figures are modified from [12].

1.5. Microarray platforms for printing slides:

Different platforms are using different materials for printing the microarrays (e,g, glass, silicon). Glass materials provides non porous, transparent, stable, rigid material that allows fluorescent detection, less background fluorescent and efficient kinetics during the hybridization[41,47]. Physical delivery techniques such as inkjet or microjet printing are used for depositing probes on high density microarray[54]. Printing and insitu synthesis are two major methods for depositing probes on materials[31]. Examples are given below

1.5.1. NimbleGen Systems

This system employs maskless photo mediated system for printing microarrays[41]. In this technology virtual masks are produced with the help of a digital micrometer device for creating a

U.V pattern. On the microarray surface there are nucleotides containing photo labile protecting groups. The protecting groups are removed after exposing to UV light. This allows synthesis of 60 to 100 nucleotides as shown in figure below.

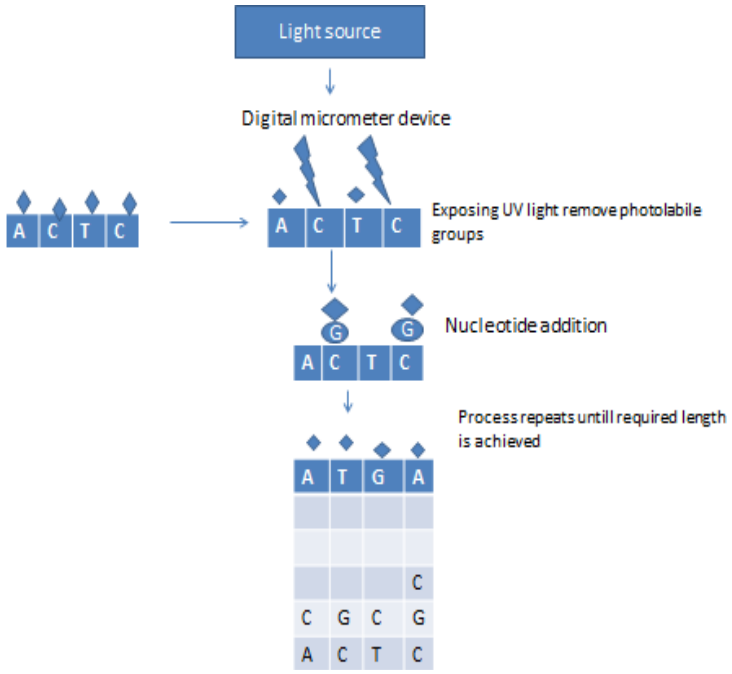


Figure 5: General principles of NimbleGen oligonucleotide microarray: light deprotection and nucleotide addition as shown.

1.5.2. Affymatrix

This platform synthesizes oligonucleotide in situ chemically on a quartz surface by photolithographic technology[55]. Oligonucleotides of 25 bases fabricated directly on to the surface by masking, light exposure and coupling of oligonucleotide[56]. After hydroxyl group attached to glass slides, linker groups were also attached to photoliable groups. These groups are deprotected after exposure to U.V light allowing addition of nucleoside phosphoramidite monomer[54].

1.5.3. Suspension bead array

This is a type of three dimensional array consist of beads as solid support. Different sets of these beads are created with unique ID, contain different red to infra red ratio. Targets are amplified with biotinylated primer and than they hybridized to microspheres. The interaction can be monitored with flow cytometer. [41].

1.5.4. Electronic microarray

This type of microarray controls nucleic acid transport by active hybridization via electric current. Complementary semiconductor technology is use for electronic addressing of nucleic acid. The negative charged nucleic acid hybridized with positively charged test sites Advantages of this technology is multiplexing, minimum waste production, low cost. This is sufficient for majority diagnostic approaches [41].

1.5.5. Agilent technologies

The Agilent platform utilizes inkjet technology for synthesizing and spotting the probes on slides.[41] There is no need for lithographic or digital masks. Synthesizing 60mer oligonucleotide accomplished by using five ink (4 bases plus catalyst) and involves coupling and deprotection step[41]. Phosphoramidite chemistry maintains coupling efficiency for the synthesis of full length oligonucleotides directly on the surface on the microarray [50,56].

Digital sequence files allow pico liter spotting of oligos base by base. Formats of agilent microarray slides are 1x244000probes, 2x105000probes, 4x44000probes, 8x15000 probes,8x60000 probes. In the present study 2x105000 probes format was selected which means that slide contain 2 subarrays and each have probes of 105000 probes

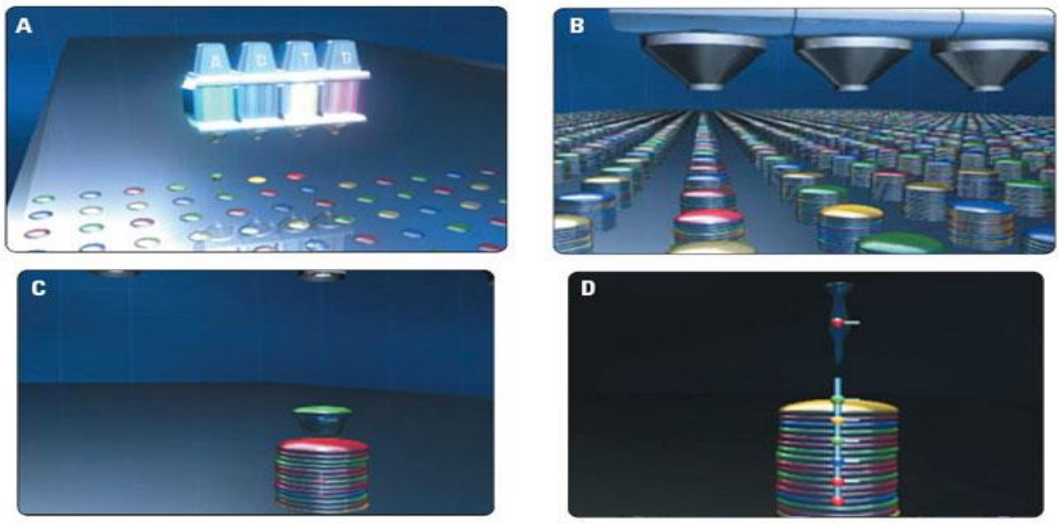


Figure 6: Oligo synthesis mechanism via inkjet printing: The first layer of nucleotides is deposited on the activated microarray surface. B: growth of the oligos is shown after multiple **l**ayers of nucleotides have been precisely printed. C: close-up of one oligo as a new base is being added to the chain, which is shown in figure D. adapted from[8].

1.6. Target or sample preparations:

A "target" is the free nucleic acid sample whose identity and/or abundance is being detected[50]. Different platforms follow different methods for amplifying the target. Target amplification is required for clinical samples, because the concentration of viruses are often below the detection limit in clinical samples [57].

One of those method is **Random phi 29 amplification** generate enough viral material for hybridization[57]. However this method can not amplify short DNA or RNA fragments. A preliminary step for detecting RNA viruses requires whole transcriptome amplification to produce cDNA fragments. This cDNA can be efficiently amplified by Phi29[57] . This phi 29 amplification was used to amplify targets in Lawrence Livermore microbial detection array(LMDA) platform which is described later [57].

Random PCR amplification is a broader approach for amplifying high density arrays [34]. This has an additional advantage to identify unknown pathogen[46]. Majority of RNA viruses lack polyA tail therefore it is advantageous to use random primer.

1.7. Sample labeling:

Samples can be labeled with different florescent dyes, mixed and hybridized to probes on the array. Measured fluorescent intensity is proportional to abundance of target. Therefore equal labeling per molecule is important.

The most commonly use dye for direct and indirect labeling are Cyanine-3 (Cy-3) and Cyanine-5 (Cy-5) dyes. They act as reporter molecules because they indicate the presence of cDNA.

Because Cy-3 and Cy-5 are light sensitive it is important to perform all working steps in the dark, to prevent degradation of dyes. Cyanine dyes gives intense strong fluorescence as these dye will not bind to the chip[42]. Three major labeling methods are

1.7.1. Direct labeling method:

This method involves the enzymatic incorporation Cy-3and Cy-5 labeled dCTP and dUTP in the target. In addition dATP and dGTP are also added. Major disadvantage is the inherent bias due to presence of different sizes of Cy3 and Cy5. The rate of labeled nucleotide incorporation is not same[56].

1.7.2. Indirect labeling method:

Indirect labeling involves first incorporates modified nucleotide sequence, that is aminoallyl linker followed by labeling with fluorescent dye. Rate of incorporation will be the same for modified and unmodified dNTPs and thereby avoiding the bias[56]. The modified dNTP amino group reacts with the NHS of mono reactive Cy-3 dye [42,47,58].

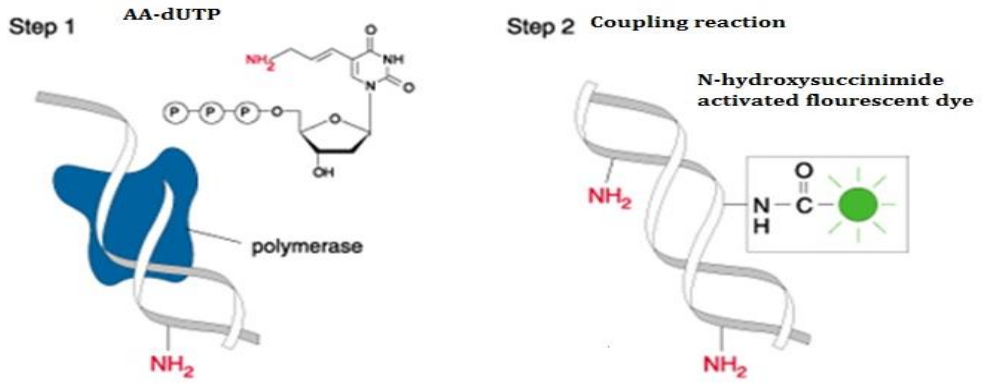


Figure 7: Indirect method of Cy 3 dye incorporation: In step 1 aminoallyl dUTP is added. Fluorescent dye attached to amine group modified in the second step. This figure is modified from[1,2].

1.7.3. Indirect dendimer labeling method:

The dendimer is complex nucleic acids branched structure comprise of hundreds of fluorescent molecules. Greenchip established indirect dendimer labeling method[59]. The dendimer method requires very less amount of RNA[60].

1.8. Hybridization:

The process of binding probe and target is known as hybridization[47]. Classical or automatic procedures are adopted for array hybridization. [42,56]. The details are described in material and method section.

1.9. Scanning

To measure the fluorescent signals scanning is an important step. It is done with scanner at the particular resolution[42,56]. The details are given in material and method section.

1.10. Previous developed virus detection arrays:

1.10.1 Virochip:

Chen et al. developed a panviral microarray assay named virochip to detect known and novel viruses[37]. The 70 mer sequences unique to a taxonomic family fabricated using agilent ink jet array platform[37]. This platform detect novel viruses SARS coronavirus, rhinovirus clade, retrovirus associated with prostate virus called XMRV, avian borna virus causing wasting disease in parrots, cardiovirus in children associated with respiratory and diarrheal illness. Present version of virochip is composed of 36.000 viral probes. These probes are formed from 1500 viruses present in genbank in 2009[61].

In virochip probes are designed by sequence derived from multiple genomic regions Due to the legth of probes 70 mers. It is tolerant for sequence mismatch[62].

1.10.2. Resequencing pathogen microarray:

This pathogen detection arrays is based on affymatrix arrays. For each target gene 4 sequence probe of legth 25 to 29 nucleotide is selected. This platform is for tropical and emerging

agents [37].

1.10.3. GreeneChip

GreeneChip Vr has been developed for detection of targeting respiratory viruses[38]. This is a high density array based on Agilent ink jet system.

Greene chip pm version 1 composed of 9477 probes for viruses. GreeneChip vr is extensively reviewed in[49] composed of 29495 60 mer oligo. Probe sequence was selected in a way that they cover viral species present in gen bank allowing five or fever mismatches [37].

1.10.4. Lawrence Livermore microbial detection array:

It is the most comprehensive array to date. Probe lengths between 50 and 65 nucleotide. This array is based on maskless photo mediated system. This platform is able to identify viruses and bacteria. The conserved sequences with in the family were used for probe designing. 10 to 50 number of different probes were selected for each microbial genome sequence[37]. 50 sequences per virus and 15 probes per bacteria were made as compared to other chips in which less probes

er targets were designed[38]. Version 1 of chip contain only viral sequence, second version designed with viral and bacterial probes[38].

2-Objectives

For applying the best possible treatment it is however important to answer the clinical question: which pathogens are present at the disease site, how important is the presence, and which influence has this presence on the microbial ecology and subsequent pathogenesis? Lacking this information only imperial therapy is possible, which may not be effective or even contraproductive (the unnecessary use of antibiotics).Therefore to answer the question of who is present and who is why responsible for given infectious respiratory disease, a good diversity of viral population that contribute to pathology is require, to know relationship between organism and disease and to prevent secondary spread of infection. The advanced diagnosis is requiring for characterizing viral diversity for this purpose high through put technology DNA microarray was used as a diagnostic tool.

- The major goal is to identify broad panel of respiratory viral pathogen in clinical samples of pediatric respiratory tract infection by applying microarray technology as a diagnostic tool.
- To be able to understand DNA microarray technology
- To visualize the global picture of detected pathogen by using software.
- Minor aims include gain experience in the planning and performing research experiment.

3- Material and Methods:

3.1. Place of work:

To work with respiratory tract infection biosafety 2 level facilities are required. This level is appropriate for working with agents that causes hazards to individual and the environment. Initial steps were carried out in BSL-2 safety levels faculty of medicine, NTNU. Final steps of experiment were performed at NTNU Cell & Molecular Biology laboratory in Trondheim.

3.2. Experimental stages:

Experimental work done in this study can be described in three stages.

1) Design of specific microarray-detectichip

2) Sample preparation

3) Data analysis

3.3. Design of specific microarray-detectichip:

3.3.1.Name of my microarray chip:

Microarray designed for this particular work is based on Agilent platform. The name given to this microarray was detectichip because this chip was able to detect viruses through detecti V software so I gave the name to the chip as detecti chip.

3.3.2. Technology used:

Detectichip that is microarray chip contains desired probes sequence was printed from Agilent. The reason for selecting Agilent as platform is its reliability, good service by its product engineers and moreover the scanner in NTNU is of Agilent which can be used for scanning printed slides from Agilent.

3.3.3. Probe selection:

Probes were selected from Geodata base. The details of Geodata base is given below.

3.3.4. Gene expression omnibus

The Gene expression omnibus (GEO) launched by NCBI is a public repository for storing and maintaining microarray data.

3.3.5. Data retrieving

Data in GEO consist of platform, sample, series, and raw data with their proper accession number[63,64].

GEO Platform (GPLxxx) : Provides a probe list.

GEO Sample (GSMxxx) : signifies a set of the molecule being probed.

GEO Series (GSExxx): Arranges samples into meaningful data that make up an experiment.

GEO Data set (GDSxxx): Reassembled data from geo staff.

Data can be downloaded as Simple Omnibus Format in Text (SOFT) format or viewing as Hyper Text Marke up Language (HTML)[65]. SOFT format contains data table and descriptive information in form of text and matrix format. SOFT matrix is arranged in excels sheets.

3.4. Searching for probe sequences:

Different platforms were observed for selecting probe sequence.

1-MegaViro platform (GPL1834), contained ~11,000 oligonucleotides were first observed. But it was rejected because it contains less probe sequence relative to other platform.

2- Platform (GPL3429, Viro3) was than selected for probes sequences. As these probe sequences were specifically designed for pediatric respiratory tract infections containing~22,000 oligonucleotides derived from ~1200 viral species. But again this probe sequence selection was rejected because these sequences were of 70 nucleotide long. Agilent technologies stop producing 70 mer long probes. There was option to select another platform if agilent will not print the slide or if these probes from another platform should be selected so that agilent can synthesize detcetichip. So this viro 3 probe platform was not selected due to technical reasons.

3.5. Obtaining required probe sequences from Geodatabase:

The other platform (GPL13407) was selected. This is a subcollection of larger array describe in an article from Gardner et al[38]. The reason for choosing this platform was its ability to detect viruses and bacteria. Majority of probes were of 60 nucleotide long. Some of the probe sequences were also of more than 60 mer sequences. These probes were modified by deleting the nucleotide from the edges. This was just to make the compromise to keep those probes on the

array. This platform provide 58131 probe sequences containing 33263 unique probe sequences out of which 24868 are redundant. These unique probes consist of specific sequences representing virus families, nonconforming sequences (not associated with family ranked taxonomic node), bacteria, plasmid, human and random control sequences[38].

After selecting the required probes the next step was to deposit that probes on chip. For this purpose Agilent technology was used.

3.5.1. E-array for ordering the detectichip

E array is a web based tool was used to create a microarray design for uploading probe. This tool consists of 5 major steps for uploading probes.

1) In this step probe parameter and file details were set out. The details about handling the uploaded probes and to set the probe parameters were specified. Appropriate format was selected. Either MiNIML abbreviation format or complete format is chosen. MiNIML format consist of probe ID column and Probe sequence column.Complete format contains seven column. The files are uploaded were uploaded in xls format.
2) Probes were uploaded and previewed
3) After uploading probe design were defined. Microarray name type such as spotted oligonucleotide arrays and the format of 2 ×105000 probes was selected. This format means that one glass slide will contain 2 subarrays, each subarray will contain 105000 spots. Duplicate probes were selected due to enhance quality control measures Linker sequences were also added. These sequences are added for those sequences which are shorter than 60 nucleotides. Moreover the purpose of these linker sequences are to decrease steric hindrance and make probe sequence available for target sequences.
4) Layout of probes is selected in which positive or negative control probes are added.
5) Finally microarray design was saved and ordering details were received[36].

3.5.2. Agilent technologies printing the detectichip:

Different platform follows different methods to fabricate oligonucleotide on chip. For example, Affymetrix uses 25-mer, MWG uses 50-mer, Agilent uses 60-mer, Operon uses 70-mer, and Clontech uses 80-mer for their oligonucleotide microarray fabrication. Photolithographic approaches followed by Affymatrix platform synthesize oligonucleotide on silicon wafer.

NimbleGen systems utilize Maskless Array Synthesizer (MAS) technology to fabricate oligonucleotide on glass slide (Principles of each technology is given in introduction part of thesis).

Ink-jetting printing technology has also been used to synthesize oligonucleotide probes on the microarray by Agilent. Printing can be done with robotic devices contain pins [66].This technology fabricates 60 mers oligonucleotide probes base by base that are present in digital sequence files enabling politer of spotting. Phosphoramidite chemistry maintains coupling efficiency for the synthesis of full length oligonucleotide[8].

After probes were successfully uploaded the next step was to fabricate detectichip done by Agilent. Detectichip is an array encompassing probe sequences of 60 mer synthesize by Agilent technologies. Spotted oligonucleotide microarray was chosen for array fabrication, because this technology provides multiplexing, rapid, cost effective, unbiased detection of infectious agents that leads to higher sensitivity and specificity. The basic difference between spotted oligonucleotide and cDNA microarray is the probe length. Oligonucleotide probes are shorter in length in contrast to cDNA nucleotide.

3.6. Sample preparation:

Two Clinical samples were obtained from nasal swabs of hospitalized children. One was sample 1 (known for adenovirus presence). Second sample was sample 2 (unknown). Target sequences present in clinical samples were randomly amplified with random primers. The details are given below.

3.6.1. Primers:

Primers were synthesized by Eurogentec (Belgium). Oligo dt primer, random primer primer serve as starting point for the polymerase to add the bases to make up a strand complementary to the template.

Oligo dt primer initiate reverse transcription at the 3' end of poly A tail. Total RNA can be used as template material. Random primers are require to transcribe non polyadenylated RNA. These primers bind throughout the target. They are less likely to give 3' bias in resulting cDNA[67].

In this study random primers are used because poly A tail are not present in the majority of the viruses. The random primers binds to multiple sites along the entire transcript to generate short, partial-length cDNAs[67]. The principle of random priming is described in figure below.

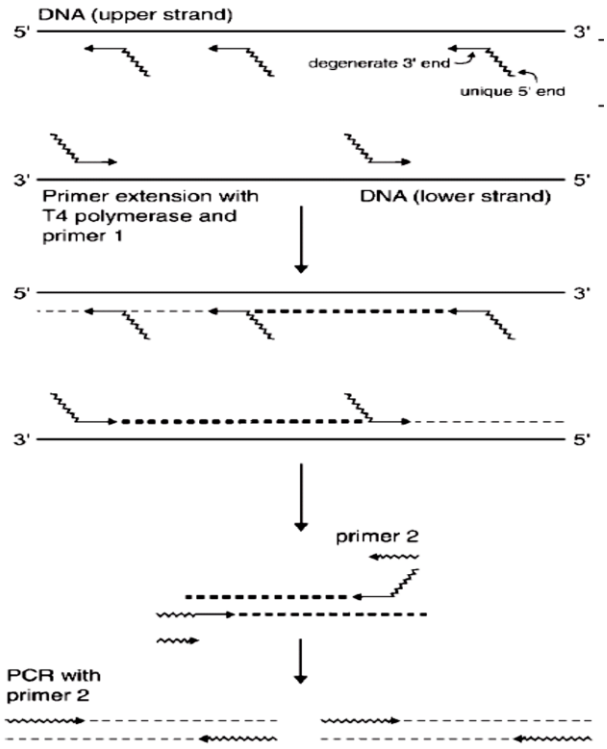


Figure 8: Principle of Random PCR: A primer 5' end with unique sequence indicated by wavy line and 3'degenerate end sequence used in PCR reaction to amplify viral DNA. The degenerate part of the primer anneals to complementary sequence. The primer is extended using Ligase. Generated double strands separated by denaturation. A primer representing only unique sequences of the first primer is used for subsequent amplification of target sequences. figure is adapted from[14].

Primer A Sequence

5'-GTTTCCCAGTCACGATA-(N9)-3'

Primer B sequence:

5'-GTTTCCCAGTCACGATA-3'

3.6.2. Reconstitution of Primers:

Primers were obtained in lyophilized form and reconstituted in water. To make 100µM stock solution of Primer A 1.37ml water was added in vial. 40pmol/µl working reagent was prepared

by taking 40 µl stock and 60 µl water. Primer B was reconstituted first in 337µl water to make 100pmol/µl concentration.

3.7. Pretreatment and extraction of Nucleic acid:

Qiagen Ultrasens Virus Kit (Heldin, Germany) was used to extract nucleic acid from nasal aspirate sample which were preserved in medium. The vials in kit based on Silica gel membrane technology to concentrate viral nucleic acid. The kit allows downstream detection of low viral titres, in contrast to other isolation kits.

Briefly 0.8ml buffer AC (to inactivate RNAs) and 5.6 µl carrier RNA was added to sample, incubated for 10 minutes followed by centrifugation for 3 minutes to pallet nucleic acid. Pallet was resuspended in buffer AR, Proteinase K and incubated for 10 minutes at 40°C to digest the protein. Buffer AB 300 µl and 700 µl cell lysate was applied to QIAamp spin column. The extracted viral nucleic acid binds to membrane after centrifugation. Washing was done with buffer AW1 and AW2. Pure nucleic acids was eluted in 60 µl buffer AVE and stored at -80°C.

Extracted nucleic acid was treated with three rounds(A,B,C) of enzymatic reaction to amplify the target.

Round A: cDNA was synthesized with the primer A and reverse transcriptase.

Round B: specific primer B is used to amplify template of round A sample.

Round C: Round B sample was further amplified target by PCR cycles followed by integrating aminoallyl dUTP [68,69].

The details of these steps are given below:

3.8. Reverse Transcription and First strand synthesis (Round A):

Reverse transcription is a process of forming cDNA. This step of reverse transcription is required because extracted RNA is likely to be destroy so reverse transcribed into cDNA is performed to get stable form of DNA[70]. Moreover most respiratory viruses are RNA. To detect RNA viruses this step is required.[67].

First-strand reverse transcription was initiated with a random nonamer linked to a primer sequence. Briefly 1 µl (40pmol/µlit) Primer A (eurogenetec) and 4 µl extracted nucleic were heated to 65°C for 5 minute in a thermo T100™ thermal Cycler Bio-Rad followed by cooling at room temperature for 5 minutes. This step is require to eliminate secondary structure, primer binding and relaxing RNA.

In above 5 µl of reaction mixture, 1 (10X) RT buffer , 1 µl (12.5mM) dNTP mix (Promega, Madison USA), DEPC treated water 2 µl, SIII reverse transcriptase 0.5 µl (Invitrogen), 0.5 µl dithiothretrol (Invitrogen) was added followed by incubation in T100™ thermal Cycler Bio-Rad for 60 minutes at 42°C. At this temperature superscript III RT enzyme provide increase specificity and higher yield of cDNA . Reaction mixture was heated at 94°C for 2 minutes (to abort reverse transcriptase) in thermocycler followed by cooling at 10°C for 2 minutes. Then the reaction mixture was pulse centrifuged for 5 seconds.

SuperScript reverse transcriptase (S III RT) is the version of M-MLV RT that has been engineered to reduce RNase H activity and provide increased thermal stability. The enzyme is used to synthesize cDNA at a temperature range of 42–55°C, providing increased specificity, higher yields of cDNA, and more full-length product than other reverse transcriptase. SuperScript reverse transcriptase (SIII RT) is not significantly inhibited by ribosomal and transfer RNA, it is used to synthesize first strand cDNA from a total RNA preparation Dithiothretrol (DTT) is added to stabilize enzyme.

3.9. Second Strand Synthesis:

For second strand synthesis sequenase™ Version 2.0 (USB products, Affymatrix, Inc, USA) was used. It is a genetically engineered form of T7 DNA polymerase that has no 3'→5' exonuclease activity and is highly processive (incorporate nucleotides). In above reaction mixture tube sequenase mix contain (1 µl (5X) Sequenase reaction buffer, 3.8 µl water, 0.15 µl (13units/µl) was added and ramp T100™ thermal Cycler Bio-Rad for 10 to 37°C. It was heated at 94°c for 2 minutes followed by cooling to 10°C. The resulting product cDNA was stored at -20°C.

3.10. PCR amplification of randomly primed cDNA: (Round B)

Polymerase chain reaction amplify DNA by denaturation, primer annealing, and primer elongation[71]. It is based on thermal denaturation of double stranded DNA and annealing

oligonucleotide at 5' and 3'. Nucleotide extension is done with polymerase. Thermo cycler maintains temperature profile to generate target copies.

To randomly primed DNA, In PCR tube round A sample 5µl with 45 µl of master mix containing (5 µl (10X) Klentaq PCR buffer, 1 µl (12.5 mM dNTP mix (Promega, Medison USA), 1 µl (100pmol/µl) primer B (Eurogentec), 1 µl (50X) Klentaq LA enzyme (Clontech United States Canada), and 37µlwater was added.

Klentaq LA is combination of klentaq-1 DNA with small amount of DNA polymerase which tends to increase fidelity, yield and length of product. PCR protocol was run as follows: 94°C, 2 min →(**25** cycles of 94°C, 30 s / 50°C, 45 s / 72°C, 1 min) →72°C, 5 min →10°C (hold).After this round B sample was checked on 1.5% agarose gel.

3.11. Agarose gel electrophoresis:

This is a technique to separate 100bp to 25kb fragments. Agarose is extracted from sea weed[51]. Molecular sieving depends on pore size that depends on concentration of agarose. Higher concentration of gel smaller will be the pore size and smaller DNA fragments can be separated. Lower the concentration corresponds to larger DNA fragments separation. In this study 1.5% agarose gel was used. Gel was made in 40 ml 0.5X TBE. After adding 0.6 g agarose in 35 ml of 0.5 ml TBE buffer. It was heated for 1 minute and incubated overnight or incubated for half an hour. Gel was poured in gel apparatus containing combs. After 30 minutes it was solidified and remove comb straight upwards and slowly. Then TBE buffer was added. Electric current was set at 60 minutes and 70 volts. 10X 2 µl loading buffer was added in 20 µl sample .Loading dye serves to concentrate and color the sample. After running the gel bands are appeared on the gel. Put that gel in gel red staining solution for 45 minutes followed with water for 5 minutes. DNA fragments of 765 bp, 880 bp 1022 bp were separated on a 1.5% agarose gel . The gel was exposed to UV light and the picture taken with a gel documentation system. After separation, the resulting DNA fragments are visible as clearly defined bands. The DNA standard or ladder should be separated to a degree that allows for the useful determination of the sizes of sample bands.

3.12. Fluorescent dye incorporation (Round C)

Fluorescent dye are mostly used for labeling purposes by following either direct or indirect

methods. Dyes are considered as reporter molecule because they indicate presence of cDNA bound to microarray. Amarsham Cye 3 dye monoreactive (GE health care) was used in labeling experiment. It was resuspended in 45µl of DMSO and stored at 20°C. Labeling efficiency also depends on handling Dye. It should always wrapped in aluminum foil, providing dark by keeping sample in reaction box and incubating for appropriate time to prevent degradation. In this present study indirect method of labeling was used because rate of incorporation will be the same for modified and unmodified dNTPs.

First aminoallyl dNTP (Invitrogen) mix was made containing aminoallyl dUTP. Aminoallyl dUTP (AA-dUTP) is responsible for the production of amine labeled DNA. AA-dNTP mix was made (12.5mM 12.5 µl dATP,dCTP, dGTP),5 mM 5 µl dTTP and 15mM,15 µl AA.dUTP and water was added to make final volume 100 µl in a tube. 1 µl aminoallyl dNTP mix was added in PCR tube containing 5µl of round B sample and 45 µl of master mix containing PCR buffer, 1 µl primer B, 1 µl Klentaq LA (Clontech United States Canada). PCR program was run as follows 94°C, 2 min →(**15**cycles of 94°C, 30 s / 50°C, 45 s / 72°C, 1 min) →72°C, 5 min →10°C (hold).

3.12.1. Cleaning of Round C sample

DNA clean and concentrator kit from (zymogen) was used for cleaning round C sample. This kit is used for rapid purification and concentration of high quality of DNA from PCR and post RT cDNA clean up. In a 1.5 ml microcentrifuge 50 µl round C sample and 250 µl DNA binding buffer was added in to zymo spin column. Followed by vertxing. Mixture was transferred to spin column in a collection tube. Centrifuged and flow through was discarded. 200 µl of DNA wash buffer was added to column. This was centrifuged for 30 second. Washing step was repeated. Than 10 µl of DNA elution buffer was added directly to column matrix and incubate at room temperature for one minute. Column was transferred to a 1.5 ml microcentrifuge and centrifuged for 30 second to elute DNA. Ultra pure DNA is ready for further use. 1 µl of 1M bicarbonate and 1 µl of Cy3 dye was added to the sample. Incubated for 1 hour in the dark. Cy3 labeled sample was again clean with DNA clean and concentrator kit by using the same kit.

3.12.2. Measuring Dye incorporation:

After indirect labeling the amount of dye incorporated in the sample was measured by using nanodrop before hybridization because it is helpful in setting up correct hybridization reaction. It was done using Nano Drop 1000 Spectrophotometer which allows measurements using 1 µl of

DNA or RNA. Appropriate value 6 pmol/µg DNA is recommended. It was determined using the formula below

Dye incorporation= pmol/µl CY3

ng/µl (concentration of DNA in ug)

3.13. Hybridization of fluorescently labeled target to detectichip:

Hybridization is done in chamber contain slide. A spot indicates hybridization if sample contain a complementary sequence to that spot which is detectable in fluorescence. So every spot provides a way of independent assay.

Master mix was prepared by blocking reagent 105µl and 21 µl fragmentation buffer. Than hybridization mix was prepared by 10µl round C sample, 85.5 µl water, 30 µl master mix, 125µl hybridization buffer. 250µl of sample of hybridization mix was loaded on to the gasket slide. Agilent printed detectichip array was placed on the above side of gasket slide (Agilent label facing down and its numeric barcode side should be up). Chamber cover was placed on this and thumbscrew was tighten and rotated in anticlockwise direction in order to wet the gasket slide. It was to be sure that detectichip is coated with hybridization mix. It was hybridized in oven at 60C° at 10 rpm speed for 17 hours. Hybridization times in high ionic strength buffer and high temperature tends to increase hybridization mixing increases hybridization kinetics[47].

3.13.1. Washing array:

Gasket/array slide was taken apart in gene expression buffer1 at room temperature. Then it was washed in preheated buffer 2 for one minute. Slide was removed from buffer 2 in 5 to 10 seconds and it was placed in slide holder and insert into array scanner facing Agilent barcode facing up.

Stringent washing is required in final washing steps. It is done by decreasing ionic strength of buffer or increasing temperature[42]. Stringent washing is done to remove cross hybridization[47].

3.14. Scanning

Detectichip is ready to scan after washing. Reporter molecule emits detectable light when stimulated with laser and its intensity was recorded. More bound spots give more intensity. Scanner might pick up light from other resources these comprise of some sample hybridized

nonspecifically, unwashed labeled sample might present on slide. Background is due to extra light. Agilent high resolution scanner was used for data processing.

To measure the fluorescent signals scanning is an important step. It was done with scanner at 5micro meter resolution scanner[42] . Fluorescent signals might weaken with time so it is important to scan the slides immediately after hybridization.

Laser is used to scan array. 150µm sized feature require 15um resolution[56]. At 532 or 635nm scanner forms grey scale images in a lossless tagged image file format called TIFF file[56]. A major characteristic of most scanners are to select the color depth of 2 byte. This means that each pixel can assumes 65,535 different intensity levels[42]. Signal intensity corresponds to target or sample bound to spot[31]. As the diameter of spot is usually less than 200 microns in diameter. Most scanners resolution is 5 to 10 microns. This resolution gives the signal intensity upto hundred of pixels in 2D image file[31].

So maximum value for pixel will be like 2^8-1=255, 2^{12}-1=4095, 2^{16}-1=65535 presented as dynamic range of scanner[31].

3.14.1. Image Analysis:

Image analysis software places feature indicator in semi automatic or automatic manner. Manually controlling of the placement is preferably better[42]. Feature indicator consist of foreground and background[42].

Quantification of image data is done by calculating the mean or median of foreground and background intensity[42]. GAL file (gene pix array list) file is used to define features. Image analysis software should differentiae a spot from its background, enabling quantification of overall intensity for each spot[31]. Many statistical programs machine learning and visualization techniques are used in to make a biological sense of massive data [31].

3.14.2. Gene pix pro software for scanning and analyzing arrays

Gene pix pro software was used for scanning and analyzing arrays. Image of each Cy3 dye channel was saved as 16 bit TIF file. This data was extracted using Gene pix pro software. Some limitations are that data cant normalized and multiple experiment cant viewed. An image of array was loaded by opening TIF file. After loading the TIF file GAL (Gene array list) file was loaded

as shown in figure. Then gene pix result file (GPR) file was opened. Spots which shows defects should be flagged (excluded from analysis).

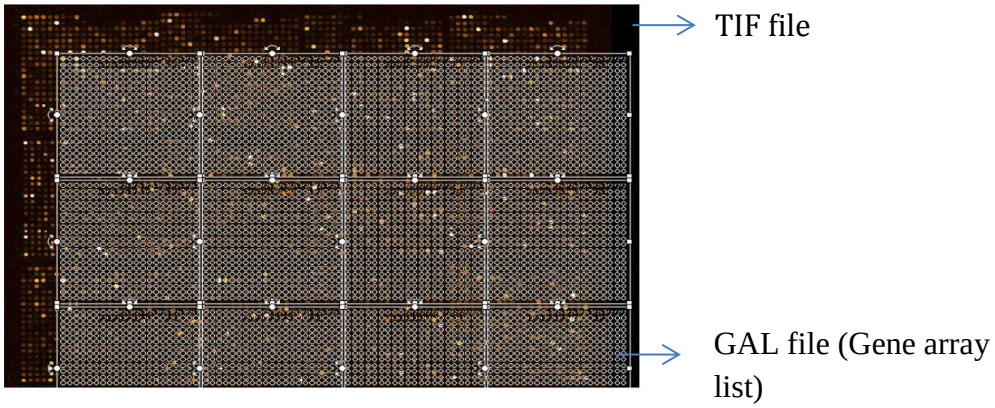


Figure 9: Loading GAL file on TIF image array by load array list [7]

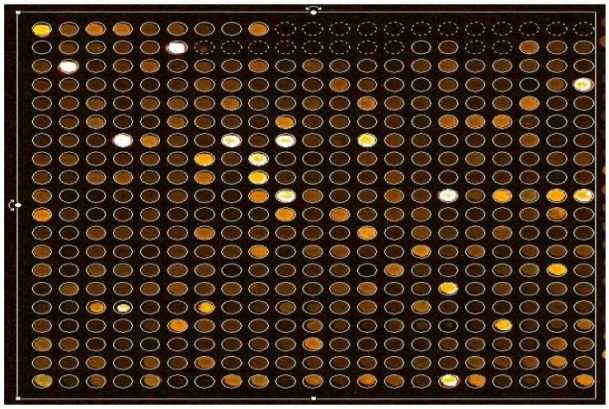


Figure 10: By using block mode : GAL file blocks combined with image indicated spot identification inside block[7]

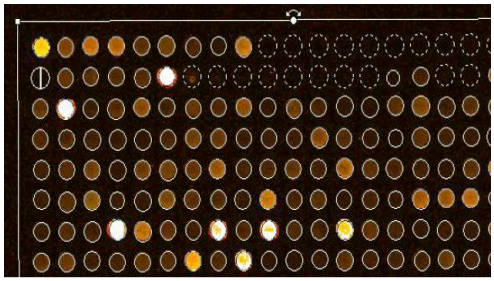


Fig 11: Spot identification inside each block: This figure indicated different spots such as saturated, unsaturated, absent, blank and undetected spot[7].

Pixels were converted in spot intensities by using launch analysis option. These result files were saved as GPR (Gene pix result file) which was used further for analysis.

4-Data analysis method

4.1. Different data analysis software

Data analysis can be done using different software. Below are the different software used by different platforms to analyze the data.

4.1.1. Composite likelihood maximization:

This algorithm is developed to analyze LLMDA data. But it can also be applied for virochip data analysis[37]. LLMDA chip data was analyzed using maximum likelihood method[57]. A forward selection algorithm was applied to find the most likely targets which are present in the sample. For pathogen analysis on array both positive and negative probe signals are also considered. positive signals above certain threshold allows pathogen detection[38].

4.1.2. E- predict

This software is developed for virochip data analysis. This compares probe intensity vectors for each array. Comparison is done against theoretical hybridization free energy vector for each sequence in target database. Subsequent Pearson correlation coefficient done for normalization. E-predict edict identifies multiple targets using iterative statistics[38]. The major advantage to use E-predict is to produce quantitative reliable data. However this software package lacks visualization tools and work in unix/Linux platform[13].

4.1.3. Greene LAMP algorithm

Green lamp stands for Log Transformed Analysis of Microarray using P value to analyze greene chip data[37]. This assumes probe intensity should be normally distributed.

This algorithm use BLAST alignments for assigning probes to target taxa. Probe intensity of each probe is determined with p value calculations and Z test[38].

4.1.4. VIPR :

This software stands for Viral Identification using Probabilistic Algorithm[37]. VIPR estimates on and off intensity distribution for each probe[37]. It is not designed to handle complex samples but can test for range of species[37]. VIPR estimated quantitative data for each probes. In context of my thesis this software is not used because the software is not able to handle complex high density arrays data of thesis.

4.1.5. Phylodetect:

In phylodetect software probe intensities are reduced to binary indicators. Scores are displayed in tree structure format, and therefore can be easily interpreted. However it is not designed to handle complex samples therefore not useful for high density arrays[37].

4.2. General Data analysis steps:

Due to highthroughput technologies large amount of data is generated needing computational analysis. In general data analysis refers to obtaining raw data by image reading software (e.g. genepix), normalization and further identification of the targets.

Normalized data of samples are compared with control samples to produce log ratio with P value calculations to determine its significance[72]. T-tests, ANOVAs, Gene Ontology (GO) overrepresentations, Bonferroni corrections, Fishers exact test are some examples of statistical methods used for complex data analysis[73]. In this study fold changes was calculated followed by T-test significant analysis to identify viral species in the sample. DetectiV software was used in this studies, steps are described below.

4.2.1. Preprocessing (From image to numbers):

Preprocessing is required to convert raw data to useful biological data. The steps of preprocessing include image analysis, background correction, normalization and data transformation[74]

4.2.2. Background correction:

The purpose of background correction is to eliminate background noise. The background noise is usually introduced by incomplete hybridization or washing steps. Background correction is adjusted by subtracting background intensity from foreground intensity[42] for each probe.

4.2.3. Normalization

In this process microarray spot intensities are adjusted in order to examine the variability across different experiments[74]. The goal of normalization is to remove systematic variation from the data. Systematic bias arises due to different labeling efficiencies scanning parameters[70]. Normalization results in consistent data[74] so that each feature is comparable to all chips[75]. Data is normalized in such a way that data is independent of particular experiment and technology used.

Normalization methods available in DetectiV are:

1)Within array normalization:

Housekeeping genes, which have constant expression or external controls also known as negative control are require for normalization process[42] Negative controls do not have any complementarity to the genes or genomes, and therefore can be used for normalization purposes. This type of normalization is generally used for adjusting signals of one single microarray. In this type of normalization each signal of array was divided by average intensities of all control probes per sample followed by log transformation.

2) Between array normalization:

This type of normalization is utilized for comparing intensities between different samples of arrays. It is assumed that intensities or log ratios have similar distribution across a set of arrays, making them comparable. In this process the probe intensities of the amplified labeled samples are divided by the intensities of same probes intensities of the negative control array probes. This gives the log fold changes in expression level of probes with respect to a negative control. An obvious example for a negative control array may be RNA from a known uninfected animal or it contains water rather than sample. The negative control array therefore has a value for each specific probe representing that value we would expect to see if that specific probe has not hybridized to anything.

3) Normalization with median:

DetectiV offers an addition method that is the median method which calculates the global median value for each array. It is based on the assumption that most probes will not hybridize to anything. If this assumption is false then this method should not be used. However, if the assumption holds, then the median is a good representation of that value which we would expect

to see from probes that have not hybridized to anything. Since there was enough raw probe intensities detected both on control and amplified samples, this method could not give specific results and was thus discarded in this study. All type of normalized data gives probes of high fold changes and are also tested for significance. Traditional statistical tests are used to test if any groups of probes or virus families in this case, are significantly different from zero. In all instances, after taking the log2, groups of probes that have not hybridized to anything should be normally distributed and have mean zero.

4.2.4. Fold change, Log transformation and P values:

Fold change is defined as the ratio of florescent intensity or it can be a difference of intensities in control and experiment[76,77]. This approach is applied to find the differentially expressed genes between control and experiment. An arbitrary threshold is selected and the fold change is considered if it is larger than threshold[42,74]. A cut off value of two fold up or down regulation is selected generally to distinguish differential expression and to produce reproducible results[42].

By applying specific mathematical functions data are transformed into different forms. The most common transformation in microarray studies is log2 of fold change or log ratio[76,77] .

Importance of logarithmic transformation is to provide values that are interpretable, more meaningful from the biological point of view and also eliminate the misleading disproportion between two relative changes. It is often more convenient to use logarithms of the expression ratios than the ratios themselves because effects on intensity of microarray signals tend be multiplicative. The logarithm transformation converts these multiplicative effects (ratios) into additive effects (differences), which are easier to model [76,78].

P-value is a measure of the evidence against the null hypothesis in a statistical test. It is the probability of the occurrence of a test statistic equal to, or more extreme than, the observed value under the assumption that the null hypothesis is true[76].

4.3. DetectiV software for analyzing Detectichip:

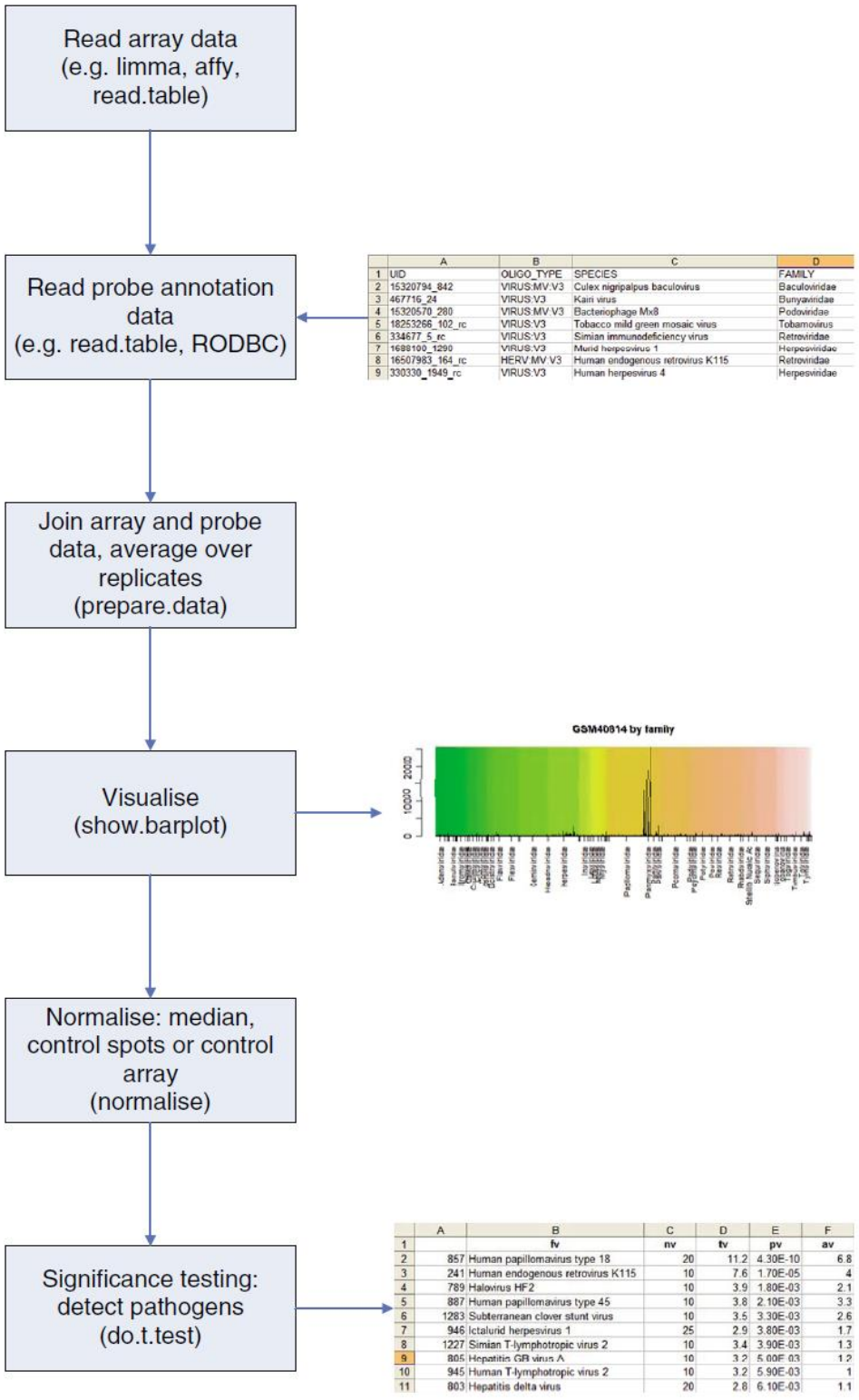
DeteciV is a software package for exploratory data analysis, written in R language. Exploratory data analysis aid the process of analysis by hiding certain aspects of data and making visible important characters rather than analyzing whole spread sheet. It offers both graphical and non-

graphical analysis and includes visualization tools for univariate (looking at one column at one time) and multivariate (looking at two or more than two column) data.

The function of this software is to provide visualization, normalization and significant testing[13] The reason for using DetectiV software to analyze the detectichip in this study are that the software is basically collection of R functions and it thus open to all computer platforms like windows and Unix/Linux. This is independent of microarray technology used (e.g. Illumina or Affy chips) for instance. Its major advantage over E-predict is powerful visualization tools in form of bar plot[51].

4.3.1. Principle of DetectiV software:

Like any other programming tools the first stage is to read the array data. This can be accomplished by the functions from limma, marray or other bioconductor packages specific for reading genepix files. Further advantages of DetectiV offers various inbuilt functions for further processing. Normalization may be carried out using the **normalise()** function, and data can be visualized using **show.barplot()** function. Finally, the **do.t.test()** function may be used to find significantly expressed pathogens group by families or names. It perform a t- test for each virus family and looks at the top ten ordered by p-value, filtered to have an average log2 scaled value >=1. Figure (12)shows the flowchart of the steps involved in data analysis.



	A	B	C	D
1	UID	OLIGO_TYPE	SPECIES	FAMILY
2	15320794_842	VIRUS:MV:V3	Culex nigripalpus baculovirus	Baculoviridae
3	467716_24	VIRUS:V3	Kairi virus	Bunyaviridae
4	15320570_280	VIRUS:MV:V3	Bacteriophage Mx8	Podoviridae
5	18253266_102_rc	VIRUS:V3	Tobacco mild green mosaic virus	Tobamovirus
6	334677_5_rc	VIRUS:V3	Simian immunodeficiency virus	Retroviridae
7	1698100_1290	VIRUS:V3	Murid herpesvirus 1	Herpesviridae
8	16507983_164_rc	HERV:MV:V3	Human endogenous retrovirus K115	Retroviridae
9	330330_1949_rc	VIRUS:V3	Human herpesvirus 4	Herpesviridae

	A	B	C	D	E	F
1		fv	nv	tv	pv	av
2	857	Human papillomavirus type 18	20	11.2	4.30E-10	6.8
3	241	Human endogenous retrovirus K115	10	7.6	1.70E-05	4
4	789	Halovirus HF2	10	3.9	1.80E-03	2.1
5	887	Human papillomavirus type 45	10	3.8	2.10E-03	3.3
6	1283	Subterranean clover stunt virus	10	3.5	3.30E-03	2.6
7	946	Ictalurid herpesvirus 1	25	2.9	3.80E-03	1.7
8	1227	Simian T-lymphotropic virus 2	10	3.4	3.90E-03	1.3
9	805	Hepatitis GB virus A	10	3.2	5.00E-03	1.2
10	945	Human T-lymphotropic virus 2	10	3.2	5.90E-03	1
11	803	Hepatitis delta virus	20	2.8	6.10E-03	1.1

Figure 12: Principle of DetectiV software: This figure is adapted from[13]

4.4. Data analysis workflow in R

R is a program for statistical computing which works in command driven language, typing commands). Bioconductor project is open source for developing open software[79] dedicated to biological data analysis. Bioconductor core packages were installed in R. The most commonly known Bioconductor[79] packages are *limma, and marray.*

4.4.1. Packages

Limma is a package for differential expression analysis of microarray data[80]. It can read many microarray scanner output formats and offer functions for their analysis. For one channel array, data can be represented as data matrix in which rows represents probes and column represents arrays or treatments [80].

marray package is for analysis of two color arrays and provide functions for reading, normalization and graphical display [81]

Bioconductor packages need to be installed on R environment once and then can be loaded as libraries each time the scripts are run. The packages are installed by using **biocLite()** command on R which utilizes an R script present on Bioconductor website initially loaded by the command **source().**The libraries for installed packages are loaded by the **library()** command.

```
source("http://www.bioconductor.org/biocLite.R")
biocLite("limma")
biocLite("marray")
library(limma)
library(marray)
```

The functions provided by DetectiV were assembled in one file called “detectiv_all_packages.R” and read it in our program by the function source() such as:

```
source("detectiv_all_packages.R")
```

4.4.2. Reading Targets

After installing libraries of required packages, the image analysis files (gpr files) produced from genepix pro software[7]are read in R. A target file is created with the names and properties of the raw data files and it is read in the variable ‘targets’ using the command **readTargets().**

```
targets <- readTargets("targets.txt")
```

4.4.3. Preprocessing

The gene pix result files(GPR) files have two columns for representing foreground and background probe intensities named F532 Mean and B532 Mean respectively. These are read into a variable of class RG List using the **read.maimages()** function. RG list is a simple list for storing red and green channel foreground and background intensities for a batch of spotted array. The intensities represented by the red color are ignored since only one color was used in this assay. R, Rb, G, Gb represents column names for Red, Red background, Green and Green background intensities respectively.

```
Cy3 <- "F532 Mean"
b<-"B532 Mean"
RG                                                                    <-
read.maimages(targets,source="genepix",columns=list(R=Cy3,G=Cy3,Rb=b,G
b=b))
RG$R <- NULL
RG$Rb <- NULL
```

Prepare.data() function is used to create a data files with the extracted expression values and combining them with the annotation values for the probes present on the chip. The probe annotation files consist of viral species names, probe sequences etc.

```
grouping <-read.table(file= "Probe_annotation_full_control.txt",header
= TRUE,sep="\t")
gdata <- prepare.data(RG$G, RG$genes$ID, grouping, "ID")
```

Further in this study known sample for adenovirus presence (sample 1) was assayed twice, thus an average of its probe intensities is calculated before further analysis and other samples are renamed accordingly.

```
gdata<-cbind(gdata,0)
gdata[,17]<-rowMeans(gdata[,4:5])
colnames(gdata)[17]<-'Sample 1 (Known)'
```

```
colnames(gdata)[2]<-'Sample 2 (Unknown)'
colnames(gdata)[3]<-'Negative sample (control)'
```

4.4.4. Visualization

Graphs for visualization of various intensities are made with the help of show.barplot() function. Each black bar represents a unique probe ID, which is grouped together according to user-defined annotation such as virus family names. The plots are saved in pdf format with the function pdf().

```
#making raw data plots
pdf("plots/raw_data_plots/raw_control.pdf")
show.barplot(gdata,"Name",3)
dev.off()
```

4.5. Visualizing raw probe intensities for each virus family

Below the figure for three sample demonstrate the raw probe intensities

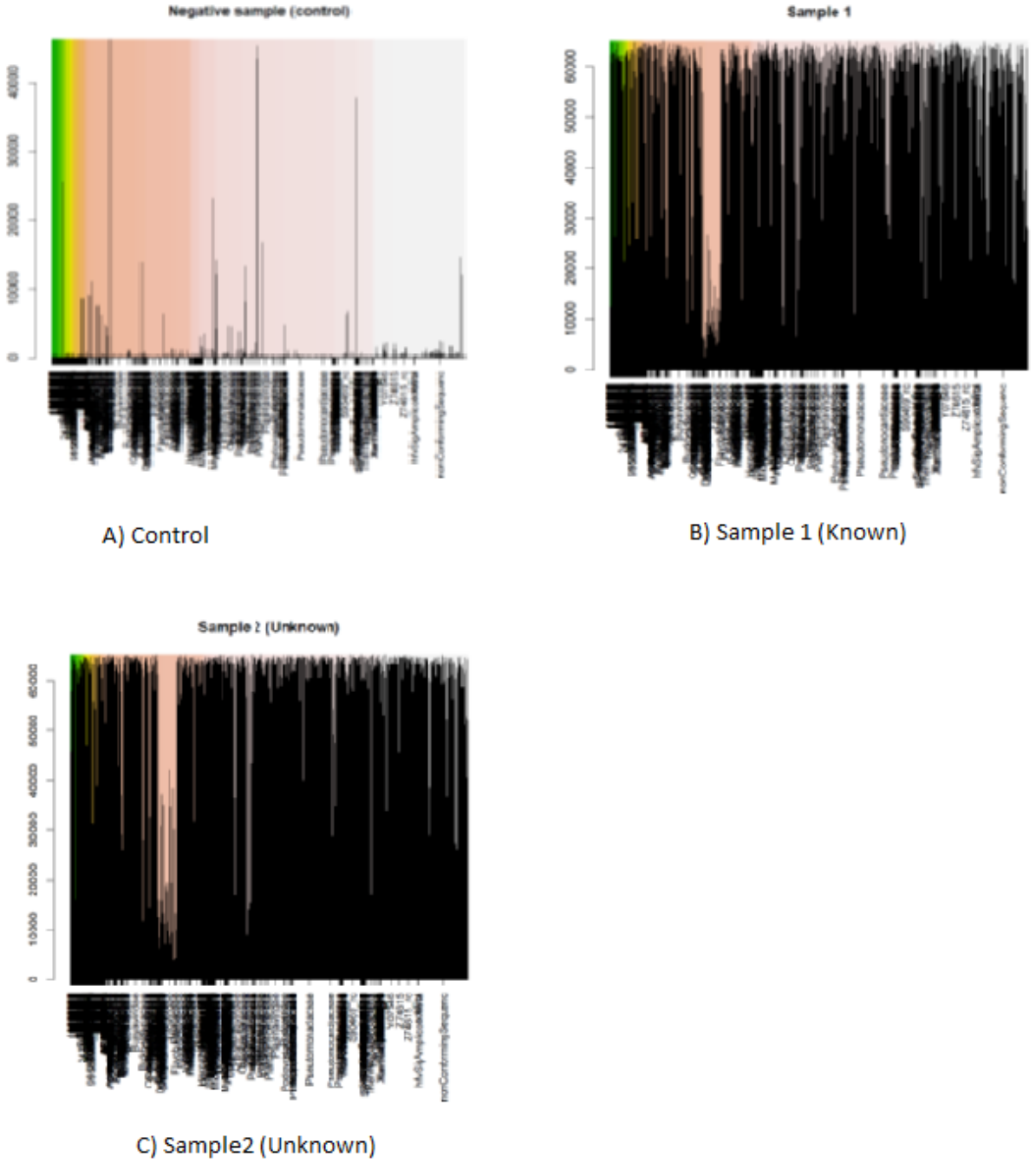


Figure 13: Raw probe intensities for each virus family in the three samples: Virus family names are shown on x axis and are also represented by background colors. The y axis represents raw probe intensities. Length of the peaks represents abundance of the corresponding virus family in the sample. Graph A represents Control sample, graph B represents Known or sample 1 and graph C represents Unknown or sample 2. Due to the high amount of positive signals. The data is not readable but shown here to represent because of certain reasons. These included that graphical representation appeared showed that software is working well and is favorable for further processes. The Peaks are not much visible because the data was not preprocessed yet. More black color and less peaks are due to that lot

viral families are present in both sample. Sharp peaks were observed in control which indicates that viral families are present in negative control but in very less amount as it was expected.

4.6. Background Correction of detectichip

For background correction RG$G-RG$Gb is added as a parameter in the prepare.data() function and the mean for sample 1 is recalculated. Then we draw plots for background corrected samples. The plots are now drawn for only the probes with positive intensities and the probes with negative intensities were discarded.

```
gdata <- prepare.data(RG$G-RG$Gb, RG$genes$ID, grouping, "ID")

gdata<-cbind(gdata,0)
gdata[,17]<-rowMeans(gdata[,4:5])
colnames(gdata)[17]<-'Sample 1 (Known)'
colnames(gdata)[2]<-'Sample 2 (Unknown)'
colnames(gdata)[3]<-'Negative sample (control)'

pdf("plots/background_corrected_plots/bg_sample2.pdf")
show.barplot(gdata[which(gdata[,2]>0),],"Name",2)
dev.off()
```

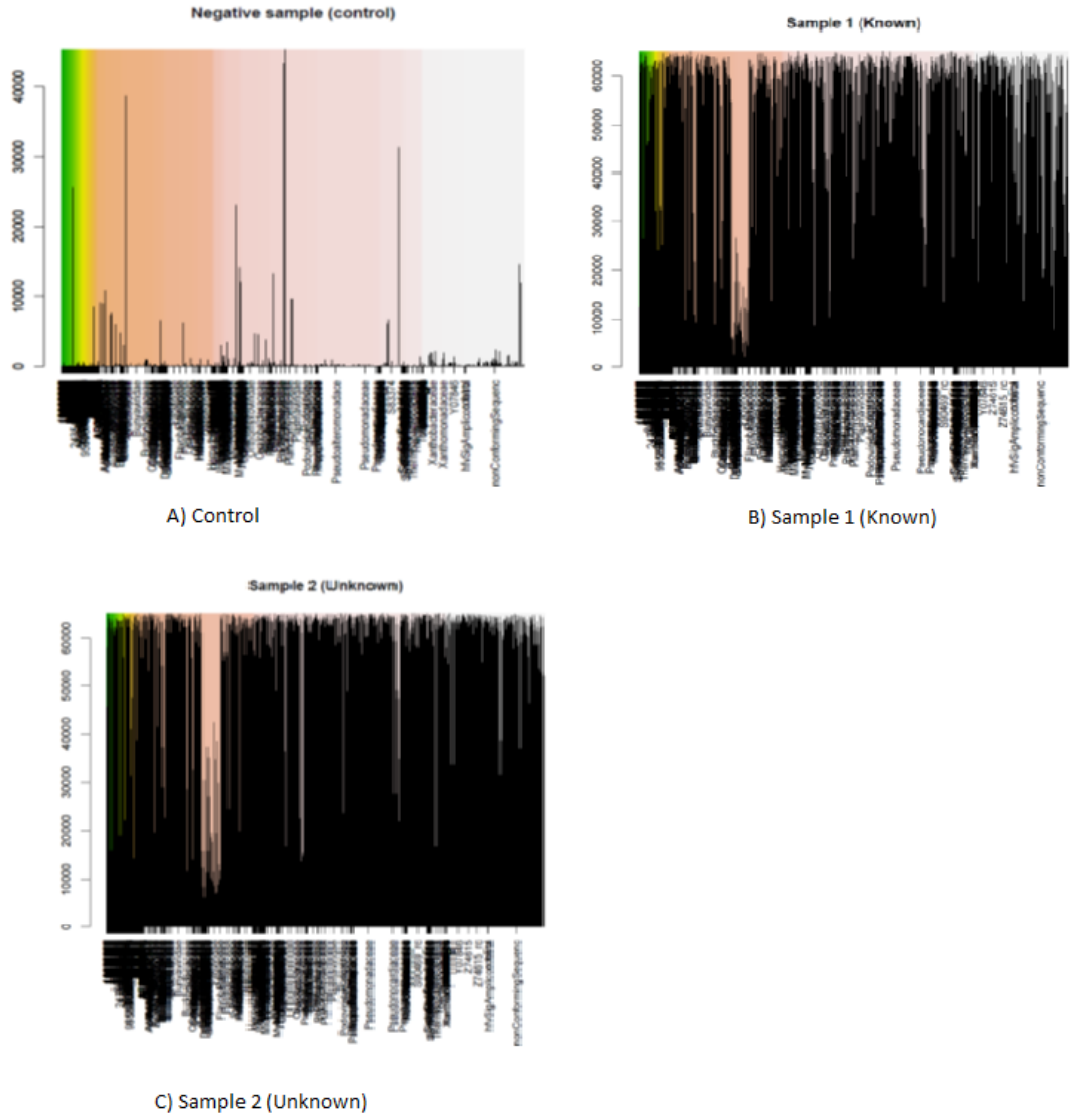


Figure 14: Background corrected intensities averaged for each virus family in three samples: Virus family names are shown on x axis and are also represented by background colors. The y axis represents raw probe intensities. Length of the peaks represents abundance of the corresponding virus family in the sample. Graph A represents Control sample, graph B represents Known(Adenovirus) or sample 1 and graph C represents Unknown or sample 2.

There is not much difference in raw data plots and background corrected plots. It is because here in this case data is not normalized yet. More black color and less peaks are due to that lot viral families are present in both sample. Sharp peaks were observed in control which indicate that viral families are present in negative control but in very less amount as it was expected.

4.7. Normalization and significance testing methods:

4.7.1. With in array control method

Normalization with negative controls presents on the arrays which ideally should have no intensities and thus their mean intensities are calculated to normalize other probes. The resulting data is again plotted for visualization. The results data is also written to an excel sheet to simplify results interpretation by using the command **write.table().**

```
controls <- grep("control", gdata$Name)
ndata <- normalise(gdata, c(2:3,17), controls)

pdf("plots/norm_with_controls/n_norm_sample2.pdf")
show.barplot(ndata[which(ndata[,2]>0),],"Name",2)
dev.off()

write.table(ndata,"ndata_control_on_array_3.xls",sep="\t",row.names=FA
LSE,quote=FALSE
```

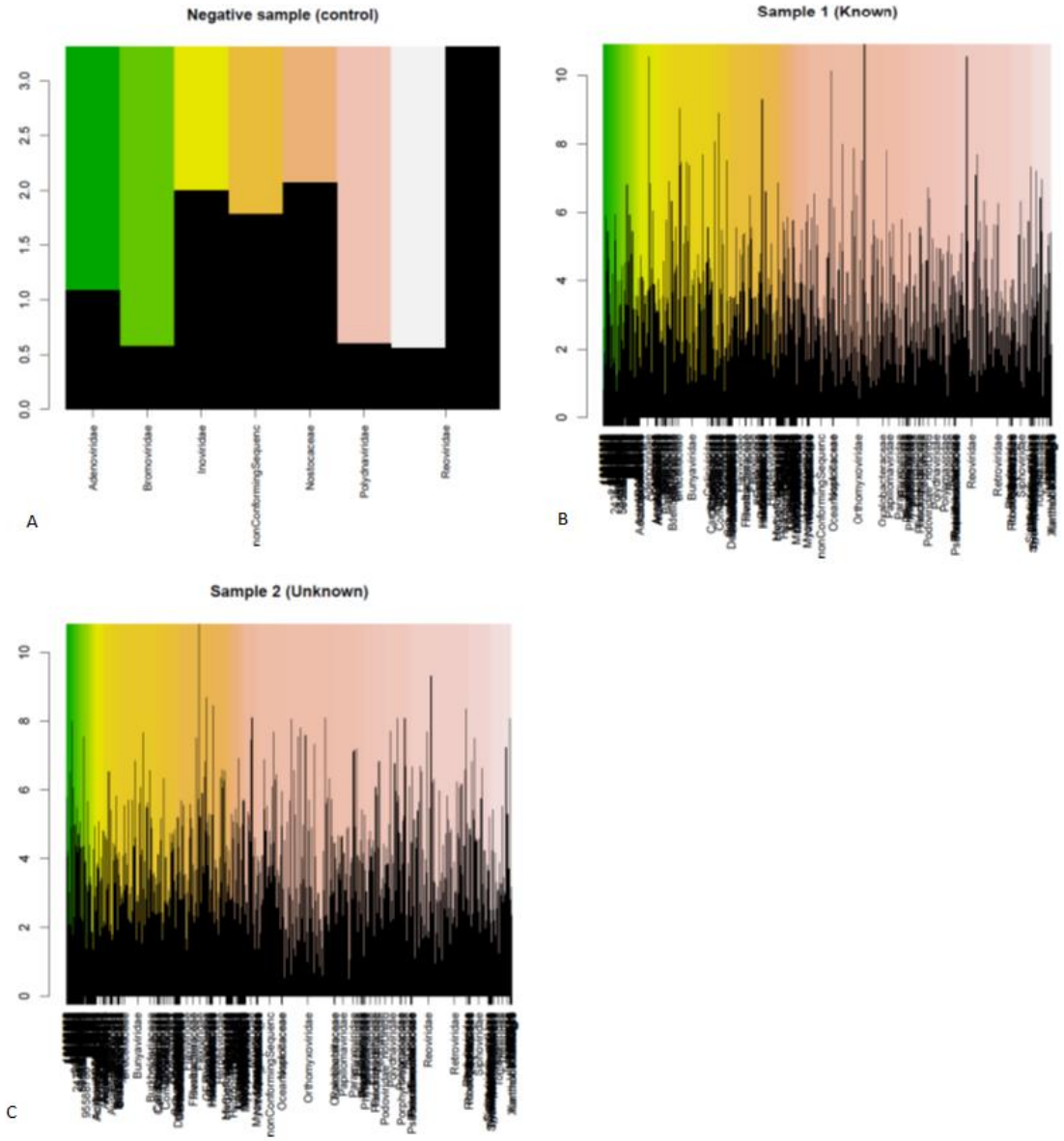


Figure 15: Graphs for all the three samples normalized with average of the negative controls in samples. The data is grouped by virus families which are shown on x axis and are also represented by background colors. The y axis represents log2 FC of normalized probe intensities. Length of the peaks represents abundance of the corresponding virus family in the samples. Graph A represents Control sample, graph B represents Known (adenovirus) or sample 1 and graph C represents Unknown or sample 2. The three graphs indicated that in negative sample 7 families are showing increase log 2 fold change. In sample 2 and sample 1 peaks indicating many virus families are present there.

4.7.2. Normalization with Whole negative control array method

Similarly the expression data is also normalized with the whole array as control sample to take into account the virus probes present in the known and unknown samples with comparison to the control array done with water. It is also visualized with plots and written to file for biological interpretation.

```
#code for normalization with control array, plotting and writing data.

adata <- normalise(gdata, c(2:5,17), carray=gdata[,3])
adata <- normalise(gdata, 2:5, carray=gdata[,3])
pdf("plots/norm_with_array/a_norm_sample2.pdf")
show.barplot(adata[which(adata[,2]>0),],"Name",2,ylim=c(0,50))
dev.off()
pdf("plots/norm_with_array/a_norm_control.pdf")
show.barplot(adata[which(adata[,3]>=0),],"Name",3,ylim=c(0,50))
dev.off()
pdf("plots/norm_with_array/a_norm_sample-avg.pdf")
show.barplot(adata[which(adata[,17]>0),],"Name",17,ylim=c(0,50))
dev.off()

write.table(adata,"adata_normalization_with_negative_sample_2.txt",sep
="\t",row.names=FALSE,quote=FALSE)
```

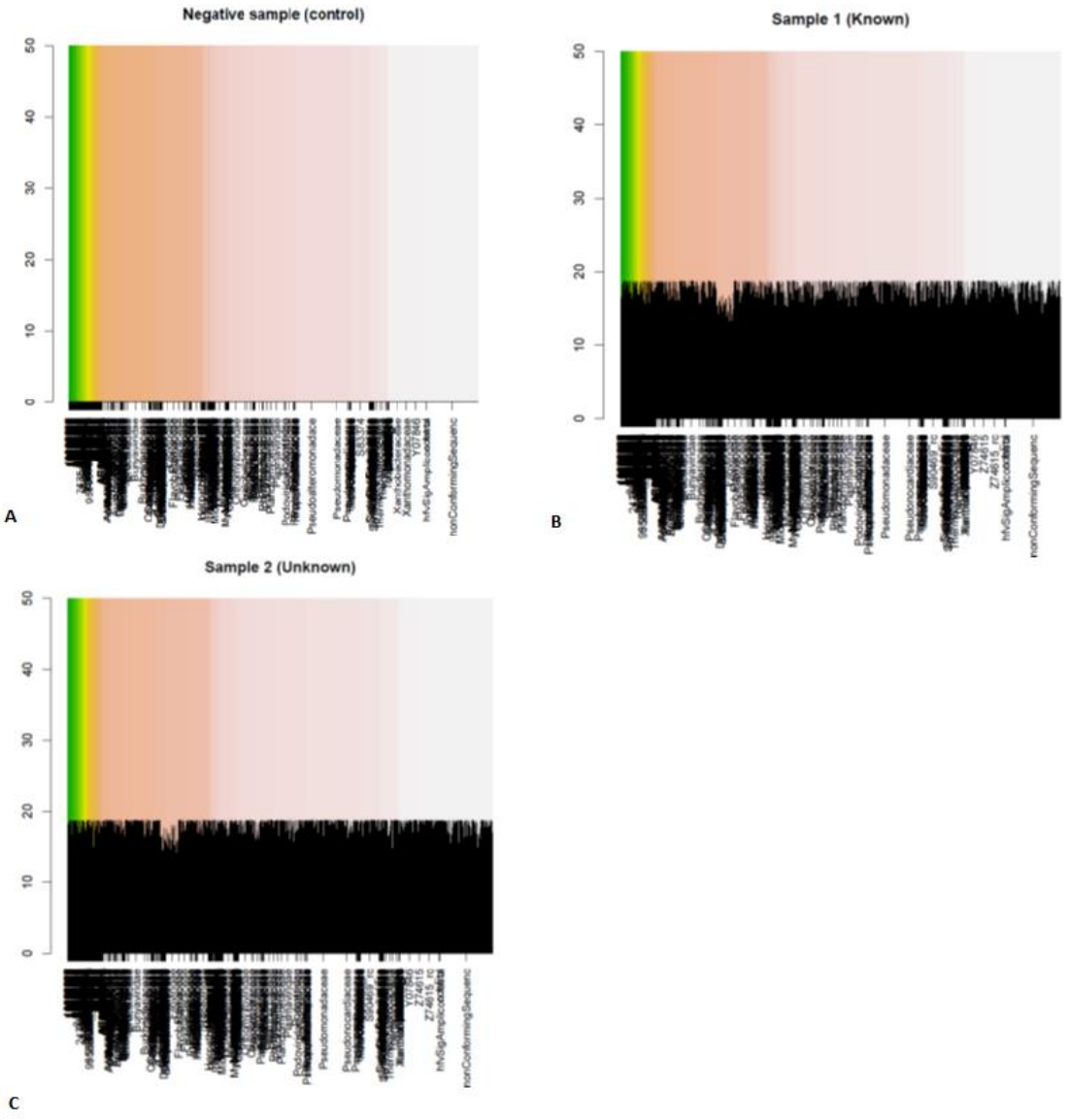


Figure 16: Graphs for all the three samples normalized with probe intensities in the controls sample. The data is grouped by virus families which are shown on x axis and are also represented by background colors. The y axis represents log2 FC of normalized probe intensities. Length of the peaks represents abundance of the corresponding virus family in the samples. Graph A represents Control sample, graph B represents Known or sample 1 and graph C represents Unknown or sample 2.

4.8. Significance testing

Significance testing for each sample and for each of the virus families present in them is done by the function do.t.test(). This function provides a measure (P value) of how significant are the intensities of virus families considering they are represented by multiple probes in one array and have variable intensities. Due to lack of probe annotation like virus species for all probes and the technical limitation for retrieving them, significant testing could not be done for any individual virus species. Thus it only indicates which virus families are significantly present in the control, known and unknown samples after they are normalized. The resulting files with Pvalues and average intensities are written to excel files.

```
#sample code for t test calculation

tt <- do.t.test(ndata, ndata$Name, 2)
ignore <- ("control")
tt <- tt[!tt$fv %in% ignore,]
tt[tt$av>=1,][1:10,]
write.table(tt,"results/sample2_significant_norm_control.xls",sep="\t"
,row.names=FALSE,quote=FALSE)
```

5-Results:

5.1. Gel electrophoresis of PCR Product:

The figure shows the PCR results of round B sample which are visualized on 1.5% agarose gel electrophoresis, sufficient to separate the fragments from 300-3000bp fragments. Round B sample was obtained taking cDNA (product of round A) as a template and than further amplifying with primer B. A smear was visualized around 200-1000bp indicated that primer is randomly bind to the target.

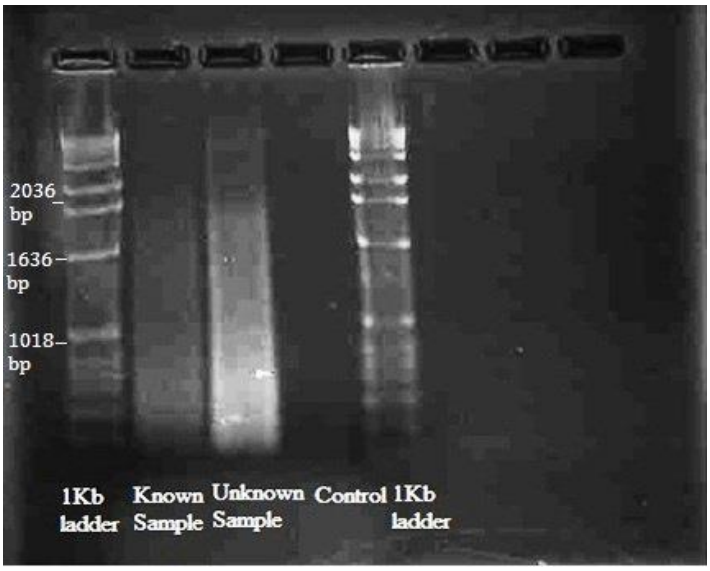


Figure 17: Gel electrophoresis after random PCR amplification of round B samples: Indicated the appearance of bands in (known for adenovirus presence) sample 1 and unknown samples2. In known sample appearance of weaker band might be due to less concentrated sample, still showed that primer attached randomly to the nucleic acids. In known sample a smear of 200-1000bp indicated that primer attached randomly to the nucleic acid. In fourth well there is no appearance of band as there is no any sample. In first and last lane 1 Kb ladder is present.

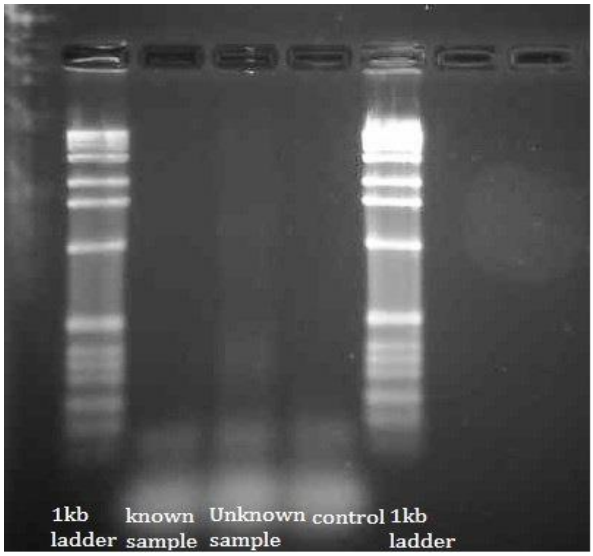


Figure 18: PCR results by using T7 primers: No fragments are observed because T7 promer did not bind to nucleic acid

5.2. Detection of viral species within array control normalized data

For deeper insights in the results, the probes with highest fold changes in the known and unknown samples were selected. Virus species of these probes were identified by aligning the sequences with nucleotide database with the BLAST tool. This resulted in some duplication of virus species as they are represented by multiple probes too. This redundancy was removed by manually selecting the representative probes based on higher fold changes. Any negative fold change values of the selected probes in other samples were set to zero for better visualization and interpretation.

The species detected as highly abundant in sample 1 and sample 2 after normalization through the negative controls present sample. Graph is drawn for comparing the species .

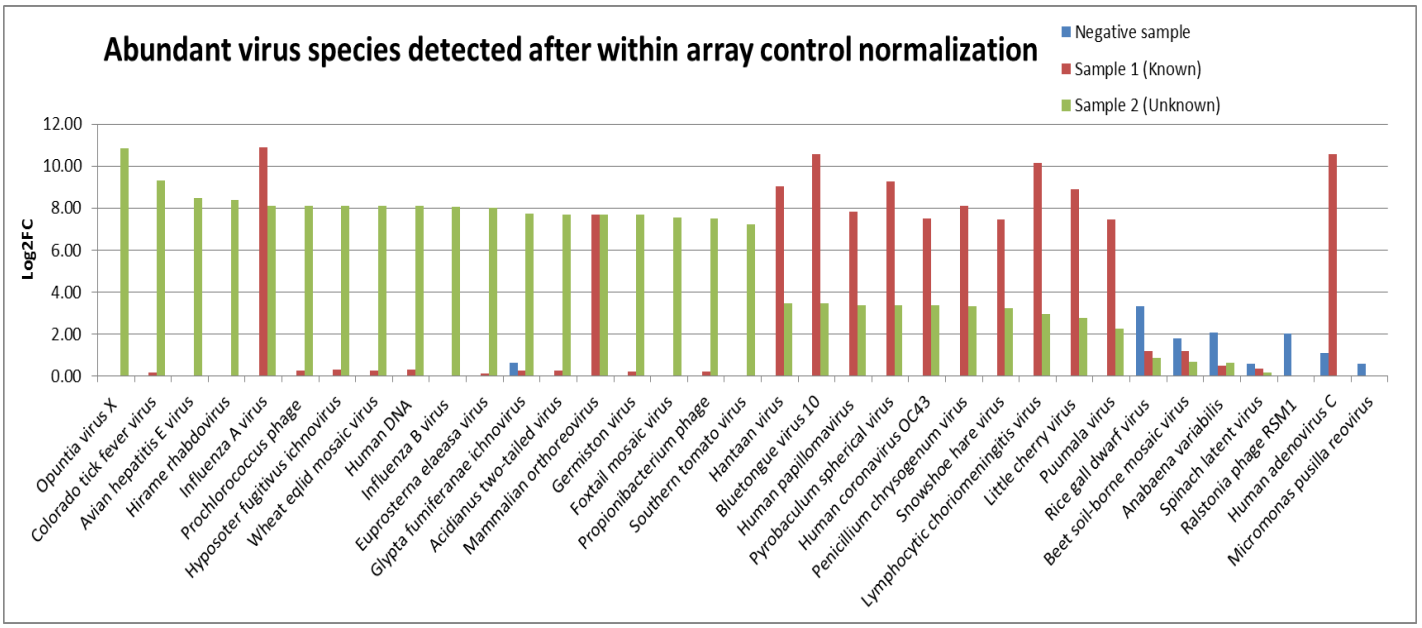


Figure 19: Virus species detection with the highest log2 FC values by with in array normalization method: in unknown, known and control samples in the data normalized with controls within array.

This table was made to show the higher Log2 fold change values and corresponding species in three samples. Yellow color indicates that those species were present in particular sample. Table 2 shown below the graph. A comparison of their values is shown in figure 18

Species	Negative sample	Sample 1 (Known)	Sample 2 (Unknown)
Opuntia virus X	0.00	0.00	10.84
Colorado tick fever virus	0.00	0.13	9.31
Avian hepatitis E virus	0.00	0.00	8.46
Hirame rhabdovirus	0.00	-3.18	8.36
Influenza A virus	0.00	10.90	8.11
Prochlorococcus phage	0.00	0.25	8.09
Hyposoter fugitivus ichnovirus	0.00	0.30	8.08
Wheat eqlid mosaic virus	0.00	0.22	8.08
Human DNA	0.00	0.29	8.08
Influenza B virus	0.00	-0.04	8.06
Euprosterna elaeasa virus	0.00	0.11	8.01
Glypta fumiferanae ichnovirus	0.60	0.23	7.70
Acidianus two-tailed virus	0.00	0.23	7.69
Mammalian orthoreovirus	0.00	7.69	7.69
Germiston virus	0.00	0.17	7.66
Foxtail mosaic virus	0.00	-0.38	7.52
Propionibacterium phage	0.00	0.19	7.51
Southern tomato virus	0.00	-0.75	7.19
Hantaan virus	0.00	9.01	3.46
Bluetongue virus 10	0.00	10.54	3.45
Human papillomavirus	0.00	7.82	3.37
Pyrobaculum spherical virus	0.00	9.27	3.37
Human coronavirus OC43	0.00	7.50	3.36
Penicillium chrysogenum virus	0.00	8.07	3.29
Snowshoe hare virus	0.00	7.45	3.22
Lymphocytic choriomeningitis virus	0.00	10.12	2.92
Little cherry virus	0.00	8.88	2.75
Puumala virus	0.00	7.43	2.25
Rice gall dwarf virus	3.31	1.16	0.85
Beet soil-borne mosaic virus	1.78	1.17	0.67
Anabaena variabilis	2.07	0.48	0.59
Spinach latent virus	0.57	0.33	0.14
Ralstonia phage RSM1	2.00	-1.99	-1.22
Human adenovirus C	1.09	10.54	-5.87
Micromonas pusilla reovirus	0.56	-6.41	-7.42

Table 2: Highly abundant virus species found after normalization through normalization with controls within arrays. The species detected as highly abundant in sample 1 and sample 2 after normalization through the negative controls present. They were made unique and the sample for which they were selected is highlighted in yellow. The table is sorted on the fold change values in sample 2 (unknown sample).

Thus the highly abundant species in sample 2 are

- Opuntia virus X
- Colorado tick fever virus
- Avian hepatitis E virus
- Hirame rhabdovirus
- Influenza A virus
- Prochlorococcus phage
- Hyposoter fugitivus ichnovirus
- Wheat eqlid mosaic virus
- Human DNA
- Influenza B virus

Highly abundant species in known sample are

- Influenza A virus
- Human adenovirus C
- Bluetongue virus 10
- Lymphocytic choriomeningitis virus

5.2.1. Likely respiratory virus species in three samples by with in array control method:

Likely respiratory virus species were detected from the with in array control method according to the high Log 2 fold change

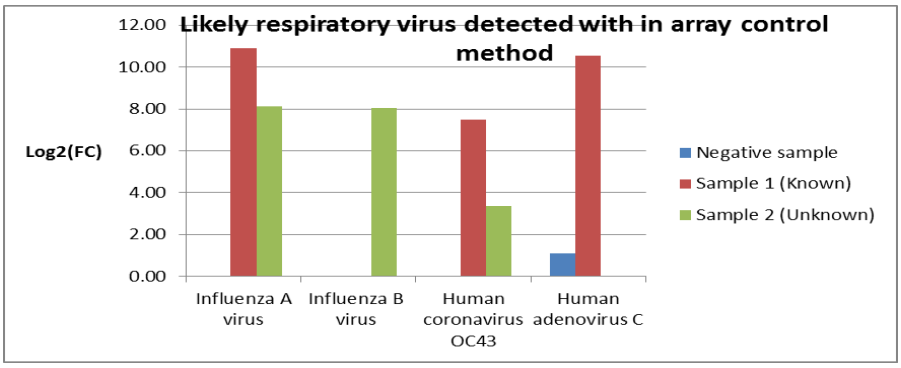


Figure 20: Likely respiratory viruses in three samples after normalization with in array control method: The figure demonstrates Influenza A, B, Human corona virus were detected species in sample 2(unknown). Whereas in addition to theses viruses Human adenovirus C was detected in sample 1(known).

Table 3: Indicted the detected viral species in three samples normalized with in array control method . In column of sample 1 and sample 2 (unknown) the values of log2 fold change are shown

Species	Negative sample	Sample 1 (Known)	Sample 2 (Unknown)
Influenza A virus	0.00	10.90	8.11
Influenza B virus	0.00	0.00	8.06
Human coronavirus OC43	0.00	7.50	3.36
Human adenovirus C	1.09	10.54	0.00

5.2.2. Rare respiratory virus species detection by with in array control method

The Detected viruses were further characterize as rare respiratory virus as there relevance with respiratory tract infection

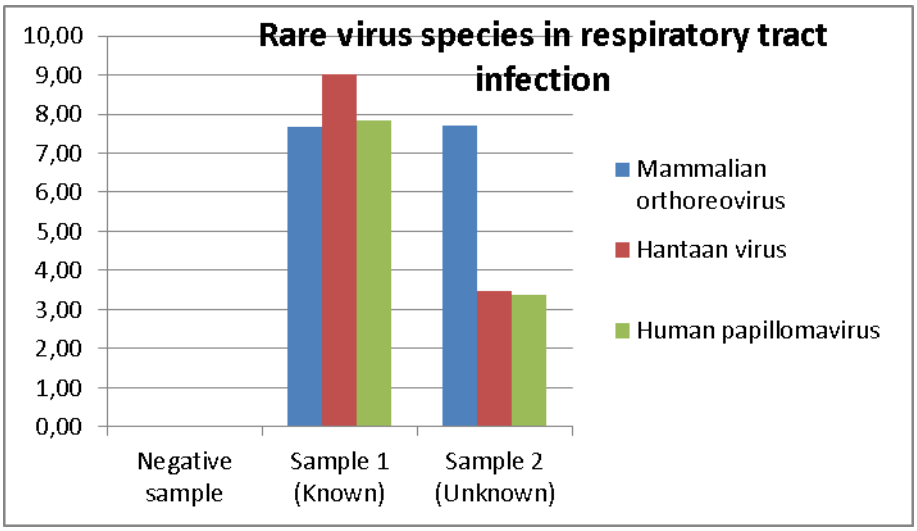


Figure 21: Detection of rare viral species in respiratory tract infection after normalization with in array control method: Mammalian orthoreovirus, Hantaan viruses and human papilloma viruses were detected in both samples.

Below the table show the Log 2 fold change values correspond to species.

Table 4: Log 2 fold change values for rare respiratory tract viruses (with in array control method)

Species	Negative sample	Sample 1 (Known)	Sample 2 (Unknown)
Mammalian orthoreovirus	0.00	7.69	7.69
Hantaan virus	0.00	9.01	3.46
Human papillomavirus	0.00	7.82	3.37

5.3. Viral species detection after normalized with whole negative control array:

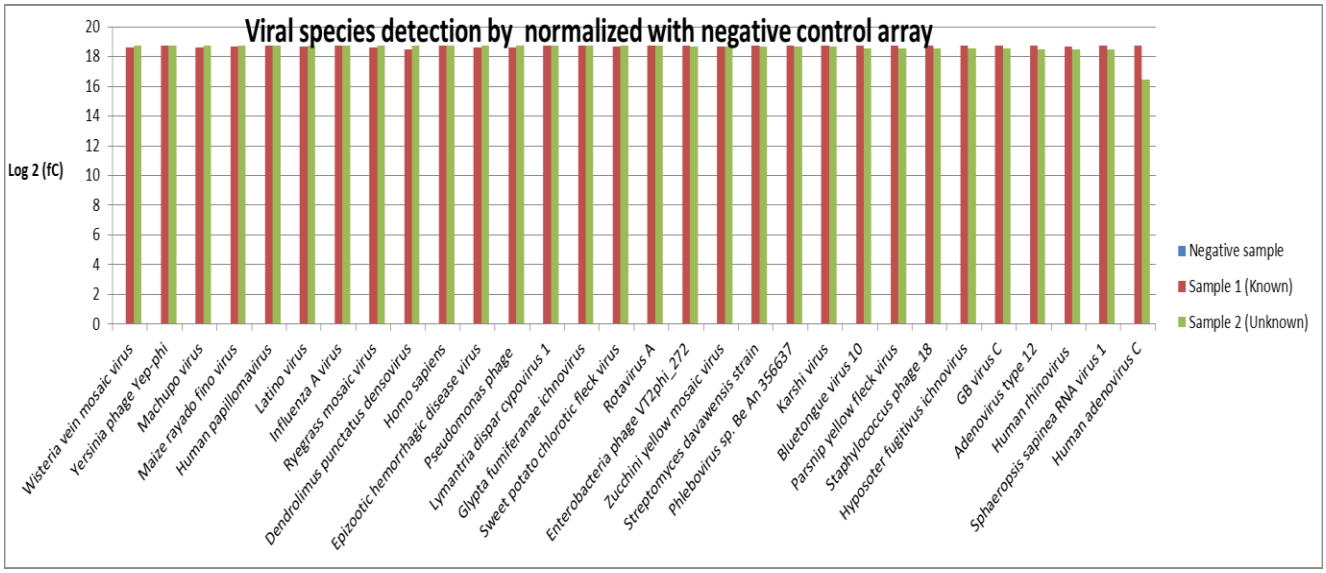


Figure 22: Graph displaying the virus species with highest log2 fold changes from the data normalized with whole array as control. We see both known and unknown sample to have large number of different virus species.

Thus the highly abundant species found in unknown sample are

- Wisteria vein mosaic virus
- Yersinia phage Yep-phi
- Machupo virus
- Maize rayado fino virus
- Latino virus
- Human papillomavirus
- Influenza A virus

The highly abundant species found in the known sample are

- Rotavirus A
- Adenovirus type 12

- Enterobacteria phage
- GB virus C
- Hyposoter fugitivus ichnovirus
- Human adenovirus

Below the table indicate the Log 2 fold change value of top species.

Species	Negative sample	Sample 1 (Known)	Sample 2 (Unknown)
Wisteria vein mosaic virus	0	18.60	18.73
Yersinia phage Yep-phi	0	18.71	18.73
Machupo virus	0	18.61	18.73
Maize rayado fino virus	0	18.70	18.73
Latino virus	0	18.70	18.73
Human papillomavirus	0	18.71	18.73
Influenza A virus	0	18.71	18.73
Ryegrass mosaic virus	0	18.59	18.72
Dendrolimus punctatus densovirus	0	18.46	18.72
Homo sapiens	0	18.71	18.72
Epizootic hemorrhagic disease virus	0	18.60	18.72
Pseudomonas phage	0	18.61	18.72
Lymantria dispar cypovirus 1	0	18.71	18.72
Glypta fumiferanae ichnovirus	0	18.71	18.71
Sweet potato chlorotic fleck virus	0	18.70	18.71
Rotavirus A	0	18.73	18.71
Enterobacteria phage VT2phi_272	0	18.72	18.70
Zucchini yellow mosaic virus	0	18.70	18.69
Streptomyces davawensis strain	0	18.72	18.68
Phlebovirus sp. Be An 356637	0	18.71	18.68
Karshi virus	0	18.72	18.64
Bluetongue virus 10	0	18.71	18.58
Parsnip yellow fleck virus	0	18.71	18.57
Staphylococcus phage 18	0	18.71	18.56
Hyposoter fugitivus ichnovirus	0	18.72	18.55
GB virus C	0	18.72	18.52
Adenovirus type 12	0	18.72	18.50
Human rhinovirus	0	18.70	18.49
Sphaeropsis sapinea RNA virus 1	0	18.71	18.46
Human adenovirus C	0	18.72	16.48

Table 5 : Highly abundant virus species found after normalization through normalization with whole negative array control. The species detected as highly abundant in sample 1 and sample 2 after normalization through the whole array (negative sample). They were made unique and the sample for which they were selected is highlighted in yellow. The table is sorted on the fold change values in sample 2 (unknown sample).

5.3.1. Likely virus species detection after whole negative control array normalization:

The detected species were extracted from the above data due to their likely relevance with disease

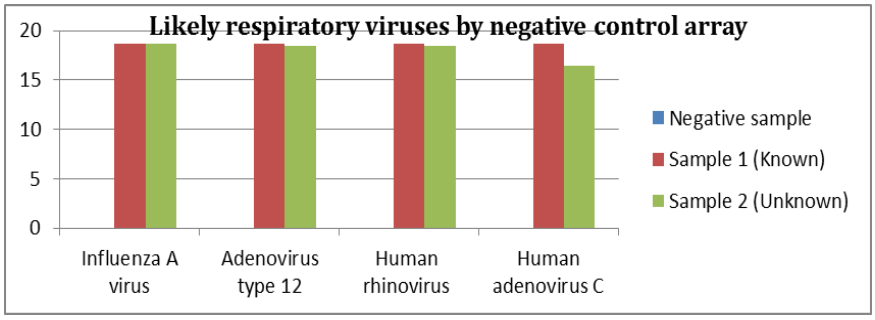


Figure 23: Detection of likely respiratory tract viruses after whole negative control array normalization: This demonstrates the green and red bars indicated that Influenza A virus, Adeno virus 12, Human rhino virus is abundant in sample 1 and 2. In addition to these Human adeno virus C is present more in sample 1(known for adeno virus) as it was expected.

The table is drawn to show the Log 2 fold change value for likely respiratory viruses

Table 6: The values of Log 2 Fold change of likely viruses for each sample: Indicating that influenza A virus is present more in unknown sample

Influenza A virus	0	18.71	18.73
Adenovirus type 12	0	18.72	18.50
Human rhinovirus	0	18.70	18.49
Human adenovirus C	0	18.72	16.48

5.3.2. Rare virus species detection after whole negative control array normalization

The detected species were extracted from the above data due to their rare relevance with disease.

The rare respiratory virus was detected in sample after whole arrays as negative control method.

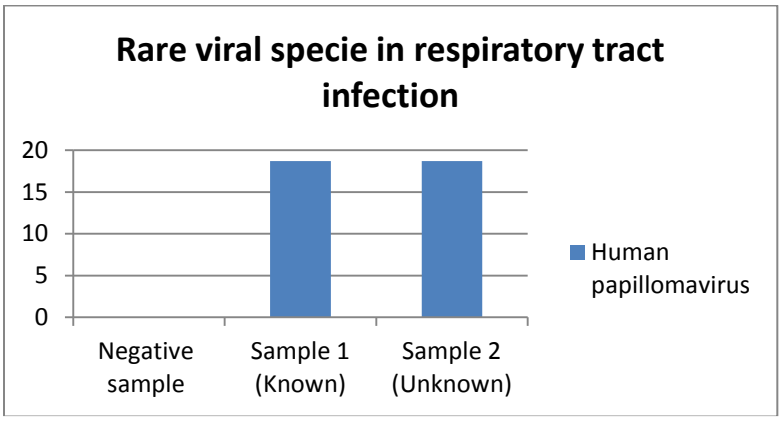


Figure 24: Rare viral specie in respiratory tract infection: Normalization with whole negative control array methods indicated that Human papilloma virus is rare respiratory tract virus

Table 7: Log2 fold change values for each sample for rare respiratory virus: Indicating abundant human papilloma virus

Species	Negative sample	Sample 1 (Known)	Sample 2 (Unknown)
Human papillomavirus	0	18.71	18.73

5.4. Significant Testing

5.4.1. Significant testing for sample 1 n data

The n data was generated after normalization with in array control for each sample. The n data was further used for significant testing for sample 1. After significant testing different t value, p value statistical test generated results that are shown table 8. This significant testing was used to identify the increase expression for presence of significant viral family . Those viral families were selected which gave P value calculation less than 0.1, and also increase expression value.

Table 8 : Significant virus families for samle 1 ndata: P value <0.1, sorted on the average expression of probes representing adenoviridae family significantly present in known sample

fv	nv	tv	pv	av
Human_32	3	47150	0.07	12785
32127331	7	11720	0.03	0.82

3217331=(adenoviridae)

Where

fv=The name of the group, from grouping

tv=The value of the t statistic

pv=The p-value from the t-test

nv=The number of observations in the group

av=The mean value of the observations in column for each group

5.4.2. Significant testing for sample 2 n data

The n data was further used for significant testing for sample 2. After significant testing different t value, p value statistical test generated results that are shown table 9. This significant testing was used to identify the increase expression for presence of significant viral family . Those viral families were selected which gave P value calculation less than 0.1, and also increase expression value. This test indicated that retroviridae family present significantly in sample 2 as shown in table 9.

Table 9: Significant virus families with P value <0.1, sorted on the average expression of probes representing this family in unknown sample.

fv	nv	tv	pv	av
8439395	26	2.65	0.01	1.32
Human_32	3	2.26	0.08	1.05
59976	4	2.72	0.04	0.82

8439395= Human endogenous retrovirus (Retroviridae)

59976=Human endogenous retrovirus(Retroviridae)

T-test on control indicated absence of any significant family as expected

5.4.3. Significant testing for sample 1 a data :

The a data was generated after whole negative array normalization method. The a data was further used for significant testing for sample 2. After significant testing different t value, p value statistical test generated results. Top families are shown here in table 10. This significant testing was used to identify the increase expression for presence of significant viral family . Those viral families were selected which gave P value calculation less than 0.1, and also increase expression value. This test indicate that adenoviridae present significantly in sample 1.

Table 10 : Following a data table of top 10 significantly present virus families in the known sample. The families were sorted on the average intensities of all the probes representing them in descending order. The P values were very low and are rounded off to 0. There were 271 families which were significant in sample 1, the most abundant being Adenoviridae.

fv	nv	tv	pv	av
Adenoviridae	664	15.13	0	10493.77
Bunyaviridae	2084	19.22	0	5578.24
Retroviridae	1988	20.66	0	4658.59
Papillomaviridae	1778	15.75	0	4618.71
Herpesviridae	1402	14.15	0	4558.59
Flaviviridae	1642	13.84	0	4321.58
Siphoviridae	2446	16.64	0	3397.58
Reoviridae	4546	20.2	0	3363.49
Flexiviridae	1598	14.77	0	3336.77
Orthomyxoviridae	7224	26.16	0	3312.67

5.4.4. Significant testing for Sample 2 (unknown) a data:

A data was generated after whole array as a normalization method. When significant testing were performed on sample 2 a data. The top 19 significantly virus families are shown in table 11.

Table 11 : Following A data, table of top 19 significantly present virus families in the unknown sample for a data. The families are sorted on the average intensities of all the probes representing them in descending order. Some of the P values were very low and are rounded off to 0. There were 793 families which were significant in sample 2, the most abundant being Tymoviridae.

fv	nv	tv	pv	av
Tymoviridae	608	113.869579	0	12.9506741
Herpesviridae	1402	155.769671	0	11.851081
Adenoviridae	664	87.4240537	0	11.2707644
Podoviridae	986	103.319339	0	11.2477293
Flexiviridae	1598	123.434843	0	11.0393644
Retroviridae	1988	136.72417	0	11.0249599
Myoviridae	1312	121.513615	0	10.9755774
Caliciviridae	764	83.7965791	0	10.9603144
Picornaviridae	1200	110.253123	0	10.9045657
Potyviridae	946	88.1833074	0	10.7392151
Bunyaviridae	2084	118.93673	0	10.7119302
nonConformingSequence	2579	145.845537	0	10.6817859
Siphoviridae	2446	142.875122	0	10.6571059
Papillomaviridae	1778	107.715272	0	10.5602706
Baculoviridae	694	77.262997	0	10.4649585
Flaviviridae	1642	107.802757	0	10.1799022
Polydnaviridae	1736	103.03283	0	10.0007581
Orthomyxoviridae	7224	213.306011	0	9.96593966
Reoviridae	4546	139.463508	0	9.12695138

6-Discussion

In this report the demonstration of detecting viruses were reported. The goal was to detect respiratory viruses with DNA microarray technology. For this purpose GPL13407 platform was utilized to fabricate the chip by Agilent technologies, as these probes were used in the previous studies[38,57]. The probes were selected due to comprehensive nature of LLMDA platform.

6.1. Gel electrophoresis of PCR product

The second step was to amplify samples. Two publication were found appropriate regarding amplification of samples which have used virochip assays[48,82]. Random amplification process was done in this study rather than phi29 amplification because random primer amplification protocol was simple. Further more other primers T7 primer were also checked for amplification but these primers were not able to bind the nucleic acid because many viruses lack poly A tail so, these primers require poly A tail to bind with. The results are shown in (figure 18 on page 49).

After amplifying with random primers the results were further evaluated with the gel electrophoresis. Interestingly smear of bands were detected indicated that primer randomly bind the target and amplified it. It is shown in (figure 17, page 48) proving usefulness of random primer for amplification. But in sample 1 (known for adenovirus) the smear was light as compared to appearance of smear in the sample 2. There might be two possible reasons for this. Less titers of viral nucleic acid is present in sample 1(Known for Adenovirus) as compared to unknown sample 2. Moreover the dimer and hair pin formation is the consequences of intrapair secondary structure formation which can be formed between 17mer tag and a nonamer results, partial binding of the primer to its target in case of known sample 1(known for adenovirus). The appearance of light smear might be due to this reason.

6.2. Preprocessing and visualization:

The data was normalized with two method. 1) Normalization within array control method in which signal of array was divided by average intensities of all control probes per sample followed by log transformation. The data first visualized by making plots for three sample in detectiV software. Normalization with this method (figure 15-page 44) only shows that the known and unknown sample contain abundant viral families as indicated by there black color peaks).

The second normalization method was the 2) Normalization with whole negative control array was done by taking the average of one full negative control array intensities (control sample was used) on this array and the values were divided by each probe intensity for sample 1 and sample 2. As the whole negative control array the average intensities were too much low. When they were divided by each probe intensities of other sample. The result of log 2 fold change was again higher for sample 1 and sample 2 as shown in plots of (Figure 16 on page 46). The possible reason might be that whole one negative control array the water was used as negative sample. But if the sample from an healthy person will be used as negative sample than there might be a possibility that healthy person sample contain viruses more than presence of viruses in water. When the value for average intensity will be divided by the intensity values for the probes of infected person than less viral family will be appear that can visualized with plot. But as the negative sample on whole array control was water and it contain less viral families. So this results abundance viral families in other samples.

6.3. Detection of virus species according Log2 fold change:

After normalization a those probe sequences were selected which gave the high log 2 fold change. As there was no information at the species level in annotation file. So the species with higher fold change were detected by aligning the sequences in BLAST. The sequences of those probes were selected which gave high log 2 fold change. The detected viral species were shown in (figure 19, page 50, by using with in array control normalization data) and (figure 22 page 55 whole negative control array) representing by green red peaks. These figures indicate the abundance species in samples. Many animal viruses, plant viruses , phages were also detected in the samples by those methods. Detection of phages on the array suggest a strategy of indirect detection of bacteria.

Tables (table 2, page 51, and table 5, page 56) are also drawn below these figures so that one can easily compare the values for log 2 fold change for corresponding species in three samples.

6.4. Detection of Likely respiratory viruses

As the samples were from nasal aspirates therefore those species were further characterized for likely respiratory viruses as the goal of this study was to characterize viral diversity. The likey

viruses figures are shown (Figure 20, Page 53 and figure 23 on page 57). By comparing both methods of normalization it is indicated that sample 1(known for adenovirus) is abundant in **adenovirus C species** as it was expected. In sample 2 (unknown) there was **Influenza A virus** because of its higher Log 2 fold change value. Here Log2 fold change is higher 8.11 (shown in table 3, page 57). These two viral species are more likely respiratory viruses as demonstrated in previous studies indicating that Influenza virus infects 5-15 % global population[83].

Finding suggests that influenza viruses are one the major causes of acute respiratory tract infection in pediatric sample. Influenza A virus was detected with both methods indicated that children (unknown sample 2) were more disposed to influenza A in this study.

Second method of normalization with whole negative control array suggests that two types of adenoviruses were present in sample 1 (Figure 23 page 57) confirming that the patient of sample 1 is more disposed to adenovirus attack. This finding suggests that the correct method was performed throughout, as this sample 1 was previously tested for **Adenovirus presence**. The results of DNA microarray indicates that adenovirus is present in sample 1.

Human corona virus was detected as a likely virus in both samples by following, with in array control normalization method but not with whole negative array normalization method. The presence of human corona virus in both samples with in array control method suggests that this virus could be a causative agent[84] .This causes cold like illnesses and majorly cause upper respiratory tract infection in children and adult [85,86] .

Human rhino virus was detected by using whole negative control array normalization method as shown in (figure 23, page 57). This is a RNA virus, belong to family picornaviridae, and cause upper respiratory tract infection[87].

6.5. Detection of rare respiratory viruses:

Data was further characterized for the presence of rare viruses associated with respiratory tract infection. Moreover the data for rare viral RTI suggests that the samples normalized with in array control method and with whole array as negative control method indicated the detection of Human papilloma virus as shown in (figure 21, page 54 and figure 24 page 58). The results from both methods indicated that **Human papilloma virus** exists as rare virus in patients of

respiratory tract infection. Previous studies indicated that recurrent respiratory papillomatosis is a rare condition caused by human papilloma virus. These viruses cause respiratory distress and hoarsness. The papillomas of aerodigestive tract are observed in the patients[88]. Human papillomaviruses have been been detected previously in the respiratory tract and are commonly found in tumors in the lungs and the oropharynx[89,90].

Mammalian orthoreovirus is also present as rare viral species with high log 2 fold change with in array control method (figure 21, page 54, and figure 24 page 58). Previous studies suggests that Mammalian orthoreovirus is considered as a human respiratory disease agent able to infect man due to zoonotic bat borne transmission[91]. Reovirus stands for (Respiratory enteric orphan virus). They were first isolated in 1950 and were named as orphan virus because these viruses were not associated with any disease at that time. These viruses have RNA genome and belong to reoviridae family[92] and assume to be a rare virus in respiratory tract.

Hanta virus with high fold change is present in both samples as a rare viral RTI. Hantavirus causes hemorrhagic fever with renal syndrome. In 1933 a severe respiratory disease occurred in South Western USA. In those cases non cardiogenic pulmonary oedema was present. These infection were assumed due to hantavirus[93]. So it can be a possible explanation that these samples in this study might have hantavirus as rare virus for respiratory tract.

6.6. Significant testing

Significant T tests were performed on virus family level, due to lack of species level annotation. It gives an idea as to probes of which viral families are significantly expressed in the samples. Significance level (pv) of <0.1 was considered as a threshold for identification of virus families.

Significant testing was performed after normalization within array control on sample 1 (known for presence of adenovirus), two hits were shown to be significant (table 8, page 58), one of which had very high average expression (av). This was represented by Human_32 indicating the presence of human genome in the sample as expected. The other significant hit showed that adenoviridae family is significantly present in sample 1(known for adenovirus).

Unknown sample (table 9, page 59) significant testing performed on n data showed that 3 hits were significant as there P-values were <0.1. This indicated that that Retroviridae (human

endogenous retro virus) was significantly present which is likely as these retro virus are integrated in human and become a part of human genome .

By following significant tests on the negative control sample. The results suggested no viral family are significantly present in it. This was expected because control do not contain any sample but water.

When significant testing was performed on a data (whole negative array control normalization) indicated many viral families which were significantly present.. For sample 1 (known for adenovirus presence) the data indicated probes of adenoviridae familiy to be significantly expressed and with highest expression values (table 10, page 60). The next viral family which was significantly present was the Bunyaviridae family. Hantaan virus also belongs to this family. Other family significantly present was retroviridae family. The Papillomaviridae family was also present significantly as the P value is zero. Reoviridae family and orthomyxoviridae (influenza virus belongs to this virus family) were among significant families.

For unknow sample 2 the viral family significantly present were Tymoviridiidae and Herpesviridae (shown in table 11 on page 60). These are plant virus families unlikely present in respiratory tract infection sample. The othe viral family that was significantly expresed was podoviridae family. Bacteriophage is belong to this family. Flexiviridae and retroviridae are also expressed significantly.

By comparing both methods it can be suggested that whole negative control array normalized data produced reliable results. As this method able to charatarize number of viruses in both species level and family level.

6.7. Other pathogen detection:

DNA microarray can also be utilized for nonviral pathogen detection. Interestingly in this study detectichip detected bacteriophages, like yersenia phage, pseudomonas phage, Enterobacteria phage, staphylococcus phage by whole negative array normalization method (were detected as shown in table 5 page 56). The phage is detected and possible explanation is that humans are constantly exposed to bacterial viruses.

6.8. Other aspects of detecti-chip:

The majority of viruses detected with detecti chip are RNA viruses such as rhinoviruses and coronaviruses. Many previous studies have been focused on the characterization of RNA viruses in the respiratory tract[94] . But DNA virus was also detected on this detectichip.

Major drawback of selecting fold change method is the arbitrary nature of cutoff value, lack of statistical measures and potential of biasing detract from its appeal [95]. Other aspects like biases in random PCR amplification, cross hybridization effects, insufficiency of microarray data analysis poses difficulties in detection process .Although the cost of assay is high. Considering the cost on per test basis steps are taken to reduce chip cost so that point-of-care system would result in reduction of cost in near future. Future developments may leads to overcome the limitations of microarray technology such as designing more comprehensive coverage of respiratory pathogen, improving primer selection, merging chip device into portable device gives benefit to point of care field. More over there is need to develop more reliable statistical tools automated software so that one can easily interpret microarray data. Microarray is highly sensitive as it can detect thousands of pathogens in less sample. This method is also highly specific as it can detect intended target. But sensitivity and specificity profile can be increase by carefully designing the experiment as each step tends to increase these profiles. Designing probes for specific target reduce cross hybridization and increase specificity.

7-Conclusion:

In the clinical diagnostic and research DNA microarray technology serves as a genomic sensor to detect viruses in samples of respiratory tract infection. Due to broad panel of detection, high throughput technology seems most promising because microarry provides a sensitive and specific way to detect multiple pathogen. Despites its sensitive detection any challenges need to be addressed regarding start up cost, miniaturization, analysis tools.

Investigations proposed that the selected statistical or fold change cut-off suggest that microarray analysis can offer essentially more than one answer, inferring data interpretation complex. My findings detected that Influenza A virus is present more in unknown sample 2 with two normalized method indicating that children of unknown sample are more disposed to influenza attack. Significant testing after whole array normalization confirms that adenoviridae family is significantly present in unknown sample. More over rare viruses were human papilloma virus which was detected with both methods in both sample. By comparing both methods the whole array normalization method was more accurate because this method provide accurate significant test for known sample for adenovirus.

To my knowledge, this is the first demonstration of feasibility of using detectiV software implemented on LLMDA probes and amplified sample by following virochip protocol for identification of respiratory tract viruses. The overall approach will facilitate rapid identification for broad spectrum of pathogens.

8-References

1 http://www.invitrogen.com/etc/medialib/en/images/ics_organized/References/the-handbook/Nucleic-Acid-Detection/Labeling-Oligos-Nucleic-Acids.

2 Nimmakayalu, M., Henegariu, O., Ward, D. C. & Bray-Ward, P. Simple method for preparation of fluor/hapten-labeled dUTP. *BioTechniques* **28**, 518-522 (2000).

3 Welte, T. & Kohnlein, T. Global and local epidemiology of community-acquired pneumonia: the experience of the CAPNETZ Network. *Seminars in respiratory and critical care medicine* **30**, 127-135, doi:10.1055/s-0029-1202941 (2009).

4 Murtagh, P., Giubergia, V., Viale, D., Bauer, G. & Pena, H. G. Lower respiratory infections by adenovirus in children. Clinical features and risk factors for bronchiolitis obliterans and mortality. *Pediatr Pulmonol* **44**, 450-456, doi:10.1002/ppul.20984 (2009).

5 McCarthy, J. E. & Evans-Gilbert, T. Descriptive epidemiology of mortality and morbidity of health-indicator diseases in hospitalized children from western Jamaica. *Am J Trop Med Hyg* **80**, 596-600 (2009).

6 Pyrc, K. *et al.* Use of sensitive, broad-spectrum molecular assays and human airway epithelium cultures for detection of respiratory pathogens. *PLoS One* **7**, e32582, doi:10.1371/journal.pone.0032582 (2012).

7 *Image analysis after microarray scan*, <http://transcriptome.ens.fr> (

8 <http://www.agilent.com> (

9 Lodes, M. J. *et al.* Identification of upper respiratory tract pathogens using electrochemical detection on an oligonucleotide microarray. *PLoS One* **2**, e924, doi:10.1371/journal.pone.0000924 (2007).

10 Chambers, J., (New York: Springer-Verlag, 1998).

11 Ma. Eugenia Manjarrez-Zavala, D. P. R.-O., Luis Horacio Gutiérrez-González, R. O.-D. a. & Cabello-Gutiérrez, C. Pathogenesis of Viral Respiratory Infection. *INTECH (open science)*.

12 *Virochip DNA Microarray Howard Hughes Medical Institute*, <http://www.hhmi.org/biointeractive/disease/Virochip/16.html> (

13 Watson, M., Dukes, J., Abu-Median, A. B., King, D. P. & Britton, P. DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data. *Genome Biol* **8**, R190, doi:10.1186/gb-2007-8-9-r190 (2007).

14 Ambrose, H. E. & Clewley, J. P. Virus discovery by sequence-independent genome amplification. *Reviews in medical virology* **16**, 365-383, doi:10.1002/rmv.515 (2006).

15 Lodes, M. J. *et al.* Identification of Upper Respiratory Tract Pathogens Using Electrochemical Detection on an Oligonucleotide Microarray. *PLoS ONE* **2**, e924, doi:10.1371/journal.pone.0000924 (2007).

16 Tregoning, J. S. & Schwarze, J. Respiratory Viral Infections in Infants: Causes, Clinical Symptoms, Virology, and Immunology. *Clinical microbiology reviews* **23**, 74-98, doi:10.1128/cmr.00032-09 (2010).

17 Mims, C., Wekelin, D., Playfair, J. & Roitt, I. in *Medical Microbiology* 182-218 (1998).

18 Bicer, S. *et al.* Virological and clinical characterizations of respiratory infections in hospitalized children. *Italian journal of pediatrics* **39**, 22, doi:10.1186/1824-7288-39-22 (2013).

19 Mahony, J. B. Nucleic acid amplification-based diagnosis of respiratory virus infections. *Expert review of anti-infective therapy* **8**, 1273-1292, doi:10.1586/eri.10.121 (2010).

20 Beka, H. *et al.* Frequency of common viruses in etiology of acute respiratory tract infections. *Indian journal of pediatrics* **80**, 91-96, doi:10.1007/s12098-012-0880-z (2013).

21 Reddington, K., Tuite, N., Barry, T., O'Grady, J. & Zumla, A. Advances in multiparametric molecular diagnostics technologies for respiratory tract infections. *Current opinion in pulmonary medicine* **19**, 298-304, doi:10.1097/MCP.0b013e32835f1b32 (2013).

22 Mahony, J. B., Petrich, A. & Smieja, M. Molecular diagnosis of respiratory virus infections. *Critical reviews in clinical laboratory sciences* **48**, 217-249, doi:10.3109/10408363.2011.640976 (2011).

23 Zheng, X. *et al.* Identification of Adenoviruses in Specimens from High-Risk Pediatric Stem Cell Transplant Recipients and Controls. *Journal of Clinical Microbiology* **46**, 317-320, doi:10.1128/jcm.01585-07 (2008).

24 Wilczynski, J. & Litwinska, B. [Human metapneumovirus--new identified virus infecting human respiratory tract]. *Przeglad epidemiologiczny* **58**, 325-333 (2004).

25 van den Hoogen, B. G. *et al.* A newly discovered human pneumovirus isolated from young children with respiratory tract disease. *Nature medicine* **7**, 719-724, doi:10.1038/89098 (2001).

26 Kahn, J. S. Human metapneumovirus: a newly emerging respiratory pathogen. *Current opinion in infectious diseases* **16**, 255-258, doi:10.1097/01.qco.0000073776.11390.bb (2003).

27 Arnold, J. C., Singh, K. K., Spector, S. A. & Sawyer, M. H. Human bocavirus: prevalence and clinical spectrum at a children's hospital. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **43**, 283-288, doi:10.1086/505399 (2006).

28 Thurlbeck, W. M. & Churg, A. in *Thurlbeck's Pathology Of The Lung* (eds A. M Churg *et al.*) 197-222 (New York, Thieme,, 2005).

29 Jartti, T., Soderlund-Venermo, M., Hedman, K., Ruuskanen, O. & Makela, M. J. New molecular virus detection methods and their clinical value in lower respiratory tract infections in children. *Paediatric respiratory reviews* **14**, 38-45, doi:10.1016/j.prrv.2012.04.002 (2013).

30 Loeffelholz, M. & Chonmaitree, T. Advances in diagnosis of respiratory virus infections. *International journal of microbiology* **2010**, 126049, doi:10.1155/2010/126049 (2010).

31 Shi1, L., , W. H., , Z. S., , X. L. & , a. W. T. Microarrays: Technologies and Applications *Applied Mycology & Biotechnology* **Volume 3** (2003).

32 *Principle of PCR*, <http://users.ugent.be/~avierstr/principles/pcr.html> (

33 Griffin H., G. A.

34 Jerome, K. R. Lennette'sLaboratoryDiagnosis ofViral Infections. *Infectious Diseaseand Therapy Series* **50** (2010).

35 Cobo, F. Application of molecular diagnostic techniques for viral testing. *The open virology journal* **6**, 104-114, doi:10.2174/1874357901206010104 (2012).

36 *Create a microarray design by uploading probes (Wizard)*, <https://earray.chem.agilent.com/earray/helppages/Index.htm#How_do_I_Log_In.htm> (

37 McLoughlin, K. S. Microarrays for Pathogen Detection and Analysis. *Briefings in Functional Genomics*, doi:10.1093/bfgp/elr027 (2011).

38 Gardner, S., Jaing, C., McLoughlin, K. & Slezak, T. A microbial detection array (MDA) for viral and bacterial detection. *BMC Genomics* **11**, 668 (2010).

39 Weile, J. & Knabbe, C. Current applications and future trends of molecular diagnostics in clinical bacteriology. *Analytical and bioanalytical chemistry* **394**, 731-742, doi:10.1007/s00216-009-2779-8 (2009).

40 Tiberini, A., Tomassoli, L., Barba, M. & Hadidi, A. Oligonucleotide microarray-based detection and identification of 10 major tomato viruses. *Journal of virological methods* **168**, 133-140, doi:10.1016/j.jviromet.2010.05.003 (2010).

41 Miller, M. B. & Tang, Y. W. Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical microbiology reviews* **22**, 611-633, doi:10.1128/cmr.00019-09 (2009).

42 Ehrenreich, A. DNA microarray technology for the microbiologist: an overview. *Appl Microbiol Biotechnol* **73**, 255-273, doi:10.1007/s00253-006-0584-2 (2006).

43 Bryant, P. A., Venter, D., Robins-Browne, R. & Curtis, N. Chips with everything: DNA microarrays in infectious diseases. *Lancet Infect Dis* **4**, 100-111, doi:10.1016/s1473-3099(04)00930-2 (2004).

44 Leveque, N., Renois, F. & Andreoletti, L. The microarray technology: facts and controversies. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* **19**, 10-14, doi:10.1111/1469-0691.12024 (2013).

45 Call, D. R. Challenges and opportunities for pathogen detection using DNA microarrays. *Critical reviews in microbiology* **31**, 91-99, doi:10.1080/10408410590921736 (2005).

46 Wong, C. *et al.* Optimization and clinical validation of a pathogen detection microarray. *Genome Biology* **8**, R93 (2007).

47 Dufva, M. in *DNA Microarrays for Biomedical Research* Vol. 529 *Methods in Molecular Biology* (ed Martin Dufva) Ch. 1, 1-22 (Humana Press, 2009).

48 Wang, D. *et al.* Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol* **1**, E2, doi:10.1371/journal.pbio.0000002 (2003).

49 Palacios, G. *et al.* Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerging infectious diseases* **13**, 73-81, doi:10.3201/eid1301.060837 (2007).

50 Shi Leming, H. w., Su Zhenqiang, Lu Xianping and Weida Tong. Microarrays: Technologies and Applications. *Applied Mycology & BiotechnologyAn International Series. Fungal Genomics* **3** (2003).

51 Lee, P. Y., Costumbrado, J., Hsu, C. Y. & Kim, Y. H. Agarose gel electrophoresis for the separation of DNA fragments. *J Vis Exp*, doi:10.3791/3923 (2012).

52 Mur, M. A. d. in *Medical Biomethods Handbook* (ed J. M. Walker and R. Rapley) (Humana Press, Inc., Totowa, NJ).

53 Jabado, O. J. *et al.* Comprehensive viral oligonucleotide probe design using conserved protein regions. *Nucleic acids research* **36**, e3, doi:10.1093/nar/gkm1106 (2008).

54 Heller, M. J. DNA microarray technology: devices, systems, and applications. *Annual review of biomedical engineering* **4**, 129-153, doi:10.1146/annurev.bioeng.4.020702.153438 (2002).

55 Wang, H. Y. *et al.* Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. *Genome Biol* **4**, R5 (2003).

56 Ehrenreich, A. DNA microarray technology for the microbiologist: an overview. *Applied Microbiology and Biotechnology* **73**, 255-273, doi:10.1007/s00253-006-0584-2 (2006).

57 Erlandsson, L., Rosenstierne, M. W., McLoughlin, K., Jaing, C. & Fomsgaard, A. The Microbial Detection Array Combined with Random Phi29-Amplification Used as a Diagnostic Tool for Virus Detection in Clinical Samples. *PLoS ONE* **6**, e22631, doi:10.1371/journal.pone.0022631 (2011).

58 DeRisi, J. Amino-allyl Dye Coupling Protocol. (2001).

59 Quan, P. L. *et al.* Detection of respiratory viruses and subtype identification of influenza A viruses by GreeneChipResp oligonucleotide microarray. *J Clin Microbiol* **45**, 2359-2364, doi:10.1128/jcm.00737-07 (2007).

60 Manduchi, E. *et al.* Comparison of different labeling methods for two-channel high-density microarray experiments. *Physiological Genomics* **10**, 169-179, doi:10.1152/physiolgenomics.00120.2001 (2002).

61 Chen, E. C., Miller, S. A., DeRisi, J. L. & Chiu, C. Y. Using a pan-viral microarray assay (Virochip) to screen clinical samples for viral pathogens. *J Vis Exp*, doi:10.3791/2536 (2011).

62 Chiu, C. Y. *et al.* Diagnosis of a Critical Respiratory Illness Caused by Human Metapneumovirus by Use of a Pan-Virus Microarray. *Journal of Clinical Microbiology* **45**, 2340-2343, doi:10.1128/jcm.00364-07 (2007).

63 Barrett, T. & Edgar, R. Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. *Methods in molecular biology (Clifton, N.J.)* **338**, 175-190, doi:10.1385/1-59745-097-9:175 (2006).

64 Barrett, T. *et al.* NCBI GEO: archive for high-throughput functional genomic data. *Nucleic acids research* **37**, D885-890, doi:10.1093/nar/gkn764 (2009).

65 Barrett, T. & Edgar, R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods in enzymology* **411**, 352-369, doi:10.1016/s0076-6879(06)11019-8 (2006).

66 Leming, S., Weiming, H., Zhenqiang, S., Xianping, L. & Weida, T. in *Applied Mycology and Biotechnology* Vol. Volume 3 (eds K. Arora Dilip & G. Khachatourians George) 271-293 (Elsevier, 2003).

67 Stangegaard, M., Dufva, I. H. & Dufva, M. Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. *BioTechniques* **40**, 649-657 (2006).

68 Bohlander, S. K., Espinosa, R., 3rd, Le Beau, M. M., Rowley, J. D. & Diaz, M. O. A method for the rapid sequence-independent amplification of microdissected chromosomal material. *Genomics* **13**, 1322-1324 (1992).

69 Wang, D. *et al.* Viral Discovery and Sequence Recovery Using DNA Microarrays. *PLoS Biol* **1**, e2, doi:10.1371/journal.pbio.0000002 (2003).

70 Khondoker, M. R. *Statistical Methods for Preprocessing Microarray Gene Expression Data*, (2006).

71 Saiki, R. K. *et al.* Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science (New York, N.Y.)* **230**, 1350-1354 (1985).

72 http://books.google.no/books?id=2z-Wqnvo0CwC&pg=PA114&lpg=PA114&dq=always+log+spot+intensities+and+ratios&source=. *Methods of enzymology, DNA Microarrays, Part B: Databases and Statistics: Databases and Statistics*. Vol. 411.

73 Dalman, M. R., Deeter, A., Nimishakavi, G. & Duan, Z. H. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics* **13 Suppl 2**, S11, doi:10.1186/1471-2105-13-s2-s11 (2012).

74 Allison, D., Cui, X., Page, G. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* **7**, 55 - 65 (2006).

75 Owzar, K., Barry, W. T., Jung, S. H., Sohn, I. & George, S. L. Statistical challenges in preprocessing in microarray experiments in cancer. *Clin Cancer Res* **14**, 5959-5966, doi:10.1158/1078-0432.ccr-07-4532 (2008).

76 Cui, X. & Churchill, G. A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* **4**, 210 (2003).

77 Tibshirani, R. A comparison of fold-change and the t-statistic for microarray data analysis. (2007).

78 Draghici, S. *Data analysis tools for DNA microarrays*. (2001).

79 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).

80 Smyth, G. K. Limma: Linear Models for Microarray Data.

81 Yang, Y. H. marray: Exploratory analysis for two-color spotted microarray data. R package version 1.38.0. (2009).

82 Chen, E. C., Miller, S. A., DeRisi, J. L. & Chiu, C. Y. Using a Pan-Viral Microarray Assay (Virochip) to Screen Clinical Samples for Viral Pathogens. *J Vis Exp*, e2536, doi:doi:10.3791/2536 (2011).

83 Mazumdar, J. *et al.* Burden of respiratory tract infections among paediatric in and out-patient units during 2010-11. *European review for medical and pharmacological sciences* **17**, 802-808 (2013).

84 Birch, C. J. *et al.* Human coronavirus OC43 causes influenza-like illness in residents and staff of aged-care facilities in Melbourne, Australia. *Epidemiology and infection* **133**, 273-277 (2005).

85 Vabret, A., Mourez, T., Gouarin, S., Petitjean, J. & Freymuth, F. An outbreak of coronavirus OC43 respiratory infection in Normandy, France. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **36**, 985-989, doi:10.1086/374222 (2003).

86 Vabret, A., Mourez, T., Gouarin, S., Petitjean, J. & Freymuth, F. An Outbreak of Coronavirus OC43 Respiratory Infection in Normandy, France. *Clinical Infectious Diseases* **36**, 985-989, doi:10.1086/374222 (2003).

87 Rollinger, J. M. & Schmidtke, M. The human rhinovirus: human-pathological impact, mechanisms of antirhinoviral agents, and strategies for their discovery. *Medicinal research reviews* **31**, 42-92, doi:10.1002/med.20176 (2011).

88 Doyle, D. J., Gianoli, G. J., Espinola, T. & Miller, R. H. Recurrent respiratory papillomatosis: juvenile versus adult forms. *The Laryngoscope* **104**, 523-527 (1994).

89 Klein, F., Amin Kotb, W. F. & Petersen, I. Incidence of human papilloma virus in lung cancer. *Lung cancer (Amsterdam, Netherlands)* **65**, 13-18, doi:10.1016/j.lungcan.2008.10.003 (2009).

90 Zawadzka-Glos, L., Jakubowska, A., Chmielik, M., Bielicka, A. & Brzewski, M. Lower airway papillomatosis in children. *International journal of pediatric otorhinolaryngology* **67**, 1117-1121 (2003).

91 Kohl, C. *et al.* Isolation and characterization of three mammalian orthoreoviruses from European bats. *PLoS One* **7**, e43106, doi:10.1371/journal.pone.0043106 (2012).

92 Chua, K. B. *et al.* A previously unknown reovirus of bat origin is associated with an acute respiratory disease in humans. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 11424-11429, doi:10.1073/pnas.0701372104 (2007).

93 Snell, N. J. New treatments for viral respiratory tract infections--opportunities and problems. *The Journal of antimicrobial chemotherapy* **47**, 251-259 (2001).

94 See, H. & Wark, P. Innate immune response to viral infection of the lungs. *Paediatric respiratory reviews* **9**, 243-250, doi:10.1016/j.prrv.2008.04.001 (2008).

95 Cui, X. & Churchill, G. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4**, 210 (2003).