UPTEC X 10 002
Examensarbete 30 hp
Februari 2010

Bioinformatic analysis of RNA data from large-scale sequencing of adenovirus infected human cells

Martin Dahlö

UPPSALA UNIVERSITET

Bioinformatics Engineering Program

Uppsala University School of Engineering

UPTEC X 10 002	Date of issue 2010-02

Author

Martin Dahlö

Title (English)

Bioinformatic analysis of RNA data from large-scale sequencing of adenovirus infected human cells

Title (Swedish)

Abstract

The gene usage in adenovirus infected human cells was compared to the gene usage in uninfected cells. This was done by studying the mRNA levels in the cells. The mRNA was extracted, sequenced and aligned to a human reference genome, and changes in coverage between the samples were found.

Keywords

Differential expression, mRNA, R, Bowtie, adenovirus, Solexa, time series, data analysis, human cells

Supervisors

Anders Isaksson
Uppsala universitet

Scientific reviewer

Hans Ellegren
Uppsala universitet

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **17**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 471 4687

Bioinformatic analysis of RNA data from large-scale sequencing of adenovirus infected human cells

Martin Dahlö

Sammanfattning

Människor kommer dagligen i kontakt med virus av olika slag, men tack vare vårt immunförsvar får sjukdomar sällan fäste. Vi håller oss oftast friska, men ibland är oturen framme och infektionen är ett faktum.

Syftet med detta examensarbete har varit att studera vad som händer inne i cellen när virus tagit sig in och ställt om cellens funktion till en massproducerande ”virusfabrik”. Genom att studera hur användningen av DNA:t i cellen förändras när virus kommer in kan vi lära oss mer om hur virus arbetar och hur man kan försvåra / förhindra dess arbete.

För att hitta de mest förändrade områdena jämfördes DNA:ts användningsnivåer (mRNA) vid tre tillfällen: oinfekterat tillstånd, 12 timmar och 24 timmar efter infektion. Av dessa drygt 1400 områden var 80% i redan kända gener (exoner och introner). 20% var i områden vars funktioner ännu ej är kända (intergeniska).

Data har tagits fram med hjälp av en sekvenseringsmaskin (Solexa). Dessa data har analyserats och program har skrivits för att hitta förändrade områden. I examensarbetet ingick också att presentera resultaten på ett begripligt sätt, samt hitta och lära sig använda de verktyg (Linux, Bowtie, R) med vars hjälp analysen genomfördes.

Examensarbete 30hp
Civilingenjörsprogrammet bioinformatik
Hösten 2009

Contents

1 Introduction

The aim of this project was to create a software pipeline that would analyze a time series of large scale sequencing data of RNA from adenovirus infected human cells (See section 7.1 for more information about adenovirus). The RNA in a human cell culture was collected and sequenced at different time points after infection, and genes differentially expressed between a time point and mock infected cells were identified. The sequencing technology used was Solexa sequencing (See section 7.2 for more information). Since Solexa produces much more data than older technologies it was important to analyze it fast and efficiently.

One of the most exciting things about using Solexa sequencing is that you do not have to decide which RNA you want to sequence, since it sequences all poly(A) selected RNA in the cell. This means that even non-coding RNA and RNA from unknown genes will be sequenced, so that you can see things you were not even looking for. Hence your findings will not be limited by previous knowledge.

There have been previous studies[1] done with microarrays, but a major drawback with that technology is that you have to choose which mRNAs you want to study. Since these studies have included expression profiles, gene onthology analysis and more time points than this study, the focus of this study will be to find the unannotated regions that have not been discovered yet.

Understanding which genes a virus regulates can teach us how the virus acts once inside a cell. This can in turn teach us how to make it harder or impossible for it to complete the infection cycle. There are not many different efficient antiviral drugs on the market today, so the world is quite vulnerable to drug/vaccine resistant viruses. Learning from antibiotics and antibiotic resistant bacteria,[2] a greater number of different types of drugs and vaccines will be needed in the future.

2 Summary

This section will give a brief overview of how the pipeline works and what results were obtained. For more details on how the pipeline works, see section 3.

As seen in Figure 1, the project can be broken down to four important steps.

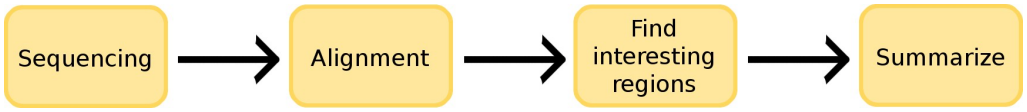

Figure 1: Flowchart of the pipeline.

The sequencing step was the practical part of the project where human cultures were lysed and the poly(A) selected RNA in the cells collected. There were three cultures involved in this step, two of them infected with adenovirus at different time points and one uninfected. The RNA was then converted to a cDNA library that was sequenced and the resulting data sets was the raw data for this project.

The alignment step was done with a publicly available alignment program and was the step where all the small reads from the sequencing step was aligned to human and virus reference genomes.

The third step was to find the interesting regions in the genomes. This was defined as regions where the expression level differed between the data sets.

The last step was the step where the results were sorted and presented in a way that humans could understand.

The final run of the pipeline resulted in ∼1 400 interesting regions, where the viral infection had changed the expression level of the genome. The majority of these regions were in annotated genes, but about one hundred of them were in unannotated intergenic regions. These regions are extra interesting since not that much is known about them.

3 Materials and Methods

This section will cover how all the steps in the pipeline were carried out in more detail.

3.1 Sequencing

The wet lab part of the project was performed by Hongxing Zhao at the Department of Genetics and Pathology, Uppsala University. Three human cell cultures were prepared and two of them were infected by adenovirus (See section 7.1 for more information about adenovirus). The three cultures were handled identically, and one of the infected cultures was taken out after 12 hours. The cells in the culture were lysed and the RNA was collected. The second infected culture was allowed to incubate for a total of 24 hours after

infection and the RNA from this culture was collected the same way as the 12 hour culture. The uninfected culture was also collected at this point.

The three samples with RNA from each culture were sent to a sequencing service[3] at Akademiska Sjukhuset in Uppsala for Solexa sequencing. The resulting data sets from the sequencing, one for each time point, was the raw data used in the rest of the project. Each data set was roughly around 9 gigabyte, or ∼53 000 000 reads, which equals about 3 975 000 000 base pairs.

Unfortunately the sequencing of the 24 hour data set was delayed so much that it was unable to be a part of the analysis in this project. It was however included in the analysis after the completion of this project.

3.2 Alignment

The data from the sequencing needed some preprocessing before it could be aligned. The Solexa data was not in the standard FASTA format so it was converted using a third-party software.[4]

The next thing that needed to be done was to align all the ∼106 000 000 reads to a reference genome. This was done with the program Bowtie.[5] The reference genome used was a combination of a human reference genome[6] and an adenovirus reference genome.[7] As seen in figure 2, the two genomes were combined so that the adenovirus genome was treated as an extra chromosome in the human genome. Since the standard FASTA file format is quite flexible, this was not a problem.

Figure 2: The adenovirus was inserted as an extra chromosome (red) in the human reference genome.

There were some reads that Bowtie did not find a unique position for, but ∼60% of the reads were aligned. A possible explanation to this is that the reads are from cDNA and the mapping is done to a ordinary DNA reference genome. Reads that are overlapping splice junctions will only map uniquely to the exonic parts of the reference genome. The intronic regions that have been spliced out will interfere and the alignment program will assume that the read does not fit in that position.

Another explanation is that a portion of the reads were of such low technical quality that the alignment program filtered them out. This is a problem that all sequencing technologies struggle with, and Solexa is no exception. If these reads were to be filtered out before running the alignment program, the

statistics would look better for the alignment, but the resulting alignment would be the same.

3.3 Finding Interesting Regions

The resulting alignments, one for each time point, were loaded into the statistics program R[8] for further analysis. The only data that was extracted from the alignments was the chromosome number and start point of each aligned read, since the sequence itself was irrelevant for the kind of analysis that would be done.

Each read's position was compared with a gene list[9] to determine if the read was in an exonic, intronic or intergenic region, and was then sorted into three subsets accordingly. This separation made later steps in the analysis easier and faster.

To be able to compare regions of the genome at different time points, the genome was divided into regions 200 base pairs long. These regions were handled one by one, throughout the whole genome. For each time point, the number of reads starting in each region was counted, and only interesting regions were kept for the next step in the analysis. See figure 3 for a graphical representation. This was done because the goal was to find differentially expressed regions, not genes. If genes were the focus of the project, one could simply extract the coverage for each of the ∼26 000[9] known genes and compare the time points with each other. However, to be able to compare every nucleotide in the genome, some kind of discretization had to be done.

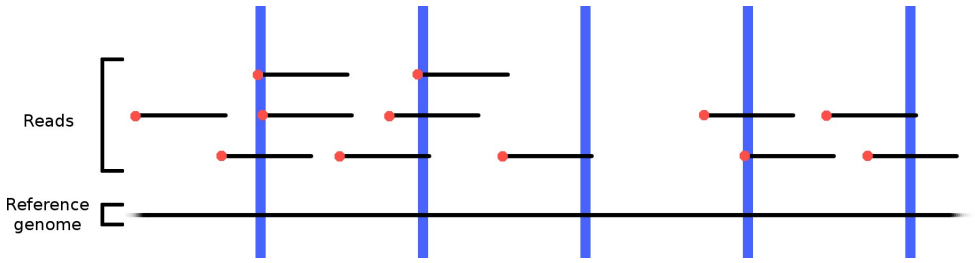

Figure 3: The reference genome was divided into regions 200 base pairs long (blue bars), and the number of reads starting (red dots) in each region was counted. Only regions with a certain amount of reads in them were kept.

What defines an interesting region is a high enough coverage and differences in expression between the time points. This is not to say that the region *is* differentially expressed, but to say that it is *a good candidate* for being differentially expressed. The number of reads in the region at the time

point was normalized in regard to the total number of aligned reads in the data set from that time point, and the length of the area investigated. See formula 1. This was done to be able to compare different data sets with different number of aligned reads with each other.

$$normalizedReads = \frac{\#readsInBin}{binSizeInKb \times millionMappedReads} \quad (1)$$

The coverage criteria was determined by introducing a normalized cutoff value that at least one of the time points must exceed. This was to ensure that there was a strong expression in at least one time point at the interesting regions. The lower coverage you have, the more susceptible to random fluctuation the number is. Hence, it is very hard to say anything about the differential expression if all time points have low coverage. The limit chosen for this was the normalized average expression level, in the areas that had expression, of the mock infected sample. With the used data sets, this equaled about 290 reads per bin.

The difference between time points criteria was based on how many times bigger the highest expressed time point was compared to the least expressed time point. Since only two time points was available, it was not possible to calculate any statistical cutoff value. The somewhat arbitrary chosen two fold change seemed significant enough, especially in combination with the coverage criteria. See figure 4 for a graphical representation.

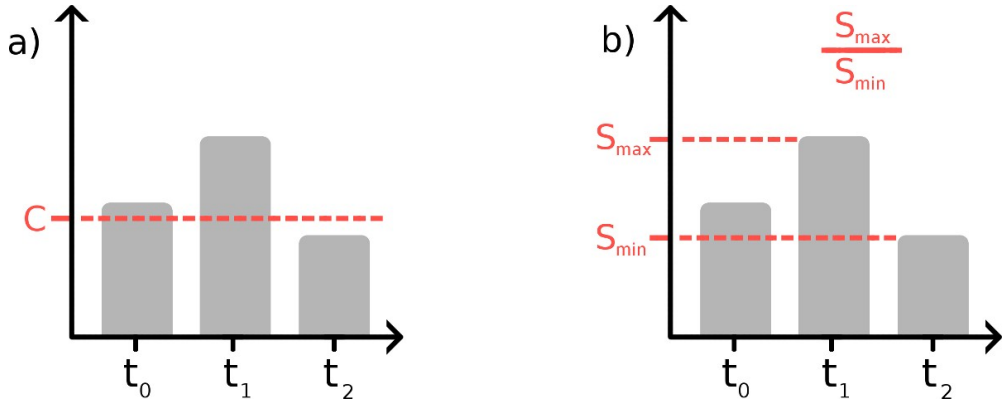

Figure 4: The three bars (t0, t1 and t2) in each plot represents the expression level in one region at three different time points. a) At least one of the time points must be above the coverage cutoff value. b) The ratio between the highest expressed time point and the lowest expressed time point, which must be higher than a certain cutoff value.

3.4 Summarize

Once all the interesting regions had been found, adjacent regions were combined into larger regions to avoid multiple hits for the same gene or region. The sort of information that was sought was among other things position of the region, how big the differential expression was and the coverage for each time point. The information about each interesting region was saved to a tab separated text file sorted by chromosome and position and this was the final result.

4 Results

The final run of the pipeline using the following cutoff values resulted in ∼1 400 interesting regions.

- Normalized coverage cutoff: 44.94219 reads per 1000 base pairs and million mapped reads
- Minimum spread cutoff: 2 fold change

As seen in figure 5, 1042 (73%) of the interesting regions were located in annotated exonic areas of the genome, 114 (8%) in annotated intronic areas and 271 (19%) in unannotated intergenic areas.

Figure 5: Graphical representation of the distribution of exonic, intronic and intergenic reads, in the found interesting regions.

See section 7.3 for a sample and a complete explanation of the output file.

5 Discussion and Conclusion

So what were these interesting regions that were found? Most of the reads in the interesting regions were inside annotated exons, which was comforting since the goal of the project was to find differentially expressed genes. Seeing which genes were affected by the infection will help understanding what

happened once the virus entered the cell. That kind of knowledge is vital when trying to develop new ways of successfully stopping the infection.

Even more exciting was the fact that 271 of the interesting regions were located in non protein coding regions. These regions are the primary source of the 'things I did not know I did not know' kind of knowledge, which in some cases turns out to be the most interesting kind of knowledge. Since the exonic regions have previously been investigated with microarrays,[1] these intergenic regions are what we hoped to find with the program.

The support for the result produced by the pipeline could be considered strong. The coverage in the interesting regions was mostly much higher than the set cutoff limit, often a thousand fold sequencing coverage. The cutoff for differential expression, that was set to a two fold change in expression, was really noticeable when viewed at those expression levels.

Since there were only two time points available it would be hard to statistically motivate a cutoff value for the differential expression. The two fold change in expression was deemed reliable through discussions with colleagues. A two fold change as the limit would probably miss a portion of modestly differentially expressed regions, but since the goal of the project was to find the most differentially expressed this was of the least concern.

Now that the interesting regions had been found it remains to investigate them in more detail to see what kind of genes and other things that were affected by the viral infection. Comparing the results of the exonic regions with the microarray studies, or making a gene onthology analysis, could also be interesting as future plans. These plans, however, are out of scope for this project.

6 Acknowledgements

During this project I had the pleasure to work in Anders Isakssons group at Uppsala University. Anders was my supervisor and helped me with everything from settling in at the office to discuss the more biological aspects of the project. Markus Rasmussen was in the same group and I am sure my project would not have been the same without his technical expertise and programming experience.

Ulf Pettersson and Hongxing Zhao at Rudbeck Laboratory provided new view points and helped steering the project in the right direction, and my scientific reviewer Hans Ellegren at the Evolutionary Biology Centre made sure this report was scientifically correct.

References

[1] F. Granberg, "Global profiling of host cell gene expression during adenovirus infection." http://uu.diva-portal.org/smash/record.jsf?pid=diva2:169144, 2006. (As of November 26, 2009).

[2] J. Turnidge and K. Christiansen, "Antibiotic use and resistance — proving the obvious," *The Lancet*, vol. 365, pp. 548–549, February 2005.

[3] "Uppsala university snp technology platform." http://www.medsci.uu.se/molmed/snpgenotyping/, November 2009. (As of December 3, 2009).

[4] "Maq: Mapping and assembly with qualities." http://maq.sourceforge.net/, April 2008. (As of January 12, 2010).

[5] "Bowtie: and ultrafast memory-efficient short read aligner." http://bowtie-bio.sourceforge.net, October 2009. (As of November 14, 2009).

[6] "Genome reference consortium: Human genome assembly build 37." http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data/index.shtml, June 2009. (As of December 12, 2009).

[7] GeneBank, "Human adenovirus 2, complete genome." http://www.ncbi.nlm.nih.gov/nuccore/56160492?report=genbank&log$=seqview, December 2008. (As of August 11, 2009).

[8] "The r project for statistical computing." http://www.r-project.org, October 2009. (As of November 14, 2009).

[9] "Ucsc genome browser." http://genome.ucsc.edu/cgi-bin/hgTables?command=start, December 2009. (As of December, 2009).

[10] MicrobiologyBytes, "Adenoviruses." http://www.microbiologybytes.com/virology/Adenoviruses.html, April 2009. (As of August 11, 2009).

[11] D. Curiel and J. T. Douglas, *Adenoviral vectors for gene therapy*. Academic Press, 2002.

[12] L. Stannard, "Adenovirus." http://web.uct.ac.za/depts/mmi/stannard/adeno.html. (As of August 28, 2009).

[13] Illumina, "technology: illumina sequencing technology." http://www.illumina.com/pages.ilmn?ID=203, 2009. (As of August 11, 2009).

[14] Illumina, "Illumina sequencing technology." http://www.illumina.com/downloads/SS_DNAsequencing.pdf, April 2009. (As of August 24, 2009).

7 Appendix

7.1 Human Adenovirus Type 2

There are many different types of adenovirus and at least 51 of them cause infections in humans. It is mostly the respiratory organs that get infected, but some of the types will also spread to other organs like eyes and digestive system.[10]

The virus particles are non-enveloped icosahedral particles, 60-90 nm in diameter (See Figure 6). Inside the particle is the double stranded viral genome which is 30-38 kilo base pair long and consists of 30-40 genes, depending on which type it is.[11] Human adenovirus type 2 more specifically has a 36 kilo base pair long genome and 37 genes.[7]

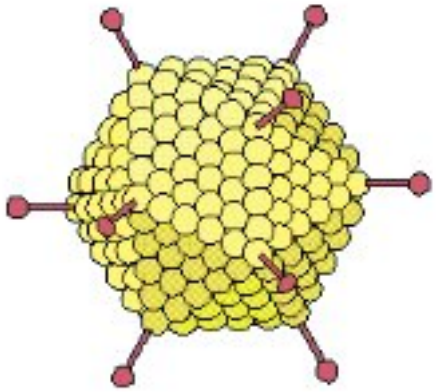

Figure 6: [12] A virus particle, 60-90 nanometer in diameter.

Adenovirus is a popular model organism to use when studying virus-host interactions, since it is easy to work with, and if there were to happen an accident the resulting infection would not be worse than an ordinary cold.

7.2 Solexa Sequencing

The sequencing technology developed by Illumina is, like other high throughput sequencing methods, based on first fragmenting the DNA, then sequencing the fragments and then reconstructing the whole sequence in a computer.[13]

The way Solexa works is that it divides the DNA into 70-75 base pair random fragments and attaches adaptor sequences at each end of the frag-

ments. The fragments are then distributed randomly onto a surface which is already covered with small pieces of DNA that is complementary to the adaptor sequences. (See Figure 7)

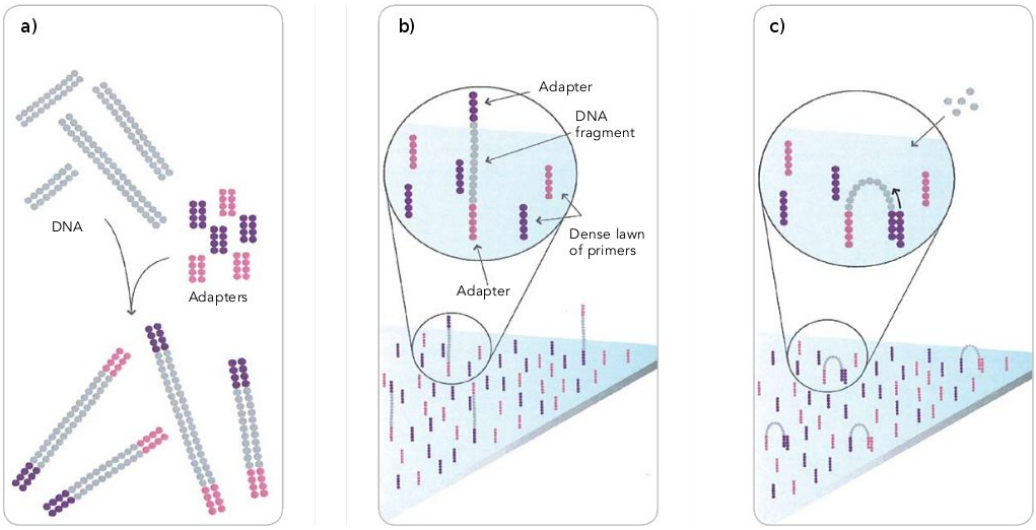

Figure 7: [14] a) Adaptor sequences are attached to the ends of the fragments. b) The fragments are randomly distributed and attached across the surface. c) Bridge amplification is initiated and the fragments free adaptor end binds to a complementary sequence attached to the plate.

The fragments get attached to the surface and using a technique called bridge amplification the fragments are multiplied to form small clusters of single stranded fragments, since each new copy of the fragment also gets attached right next to the original fragment. After the bridge amplification there is no single fragments sparsely distributed over the surface, but densely packed cluster of fragments, where each cluster consists of many single stranded copies of the same fragment.[13] (See Figure 8)

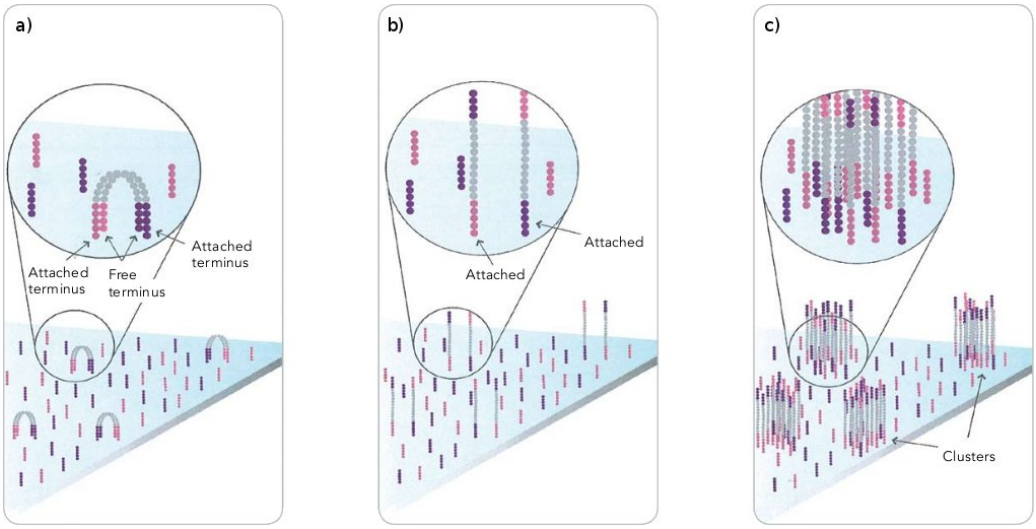

Figure 8: [14] a) Bridge amplification is complete b) The strands separate and the cycle is repeated. c) A cluster of fragments is formed after numerous iterations.

When the clusters of fragments have been created, the process of sequencing-by-synthesis begins. A mix of all nucleotides and DNA polymerase is added to the surface, so the nucleotides will get incorporated on the fragments. The nucleotides are not ordinary nucleotides, they are also reversible terminators marked with removable fluorescent dyes, so when one nucleotide is attached, replication stops. After each incorporation of a nucleotide a laser scans the surface, and the fluorescent dyes emits light of different colors, one color for each type of nucleotide. A camera records the color of the light being emitted by each cluster of fragments, and this is where the actual sequencing takes place.[13] (See Figure 9)

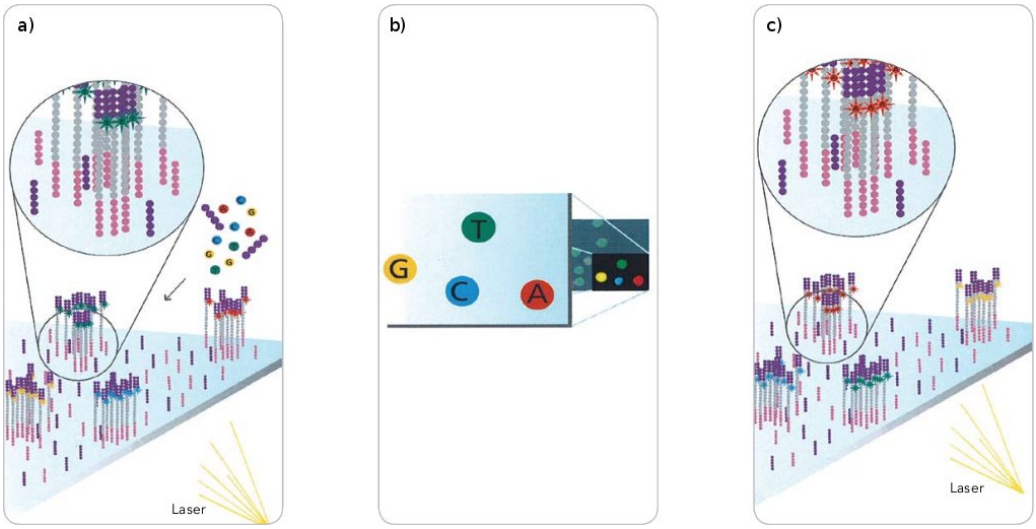

Figure 9: [14] a) Sequencing-by-synthesis begins. Once a nucleotide is attached, replication stops. b) A laser excites the fluorescent dyes, and a camera sees which colors the clusters emit. c) The reversible terminators are removed and the cycle is repeated.

Since the nucleotides are reversible terminators it is easy to remove the terminator part and also the dyes, and do the process again. This process will keep on repeating until the whole length of the fragments have been replicated/sequenced. When all the fragments have been sequenced, the reconstruction begins. The reconstruction is usually that all the fragments get aligned to a reference sequence to form a consensus sequence. This consensus sequence can then be used for whichever further analysis.[13] (See Figure 10)

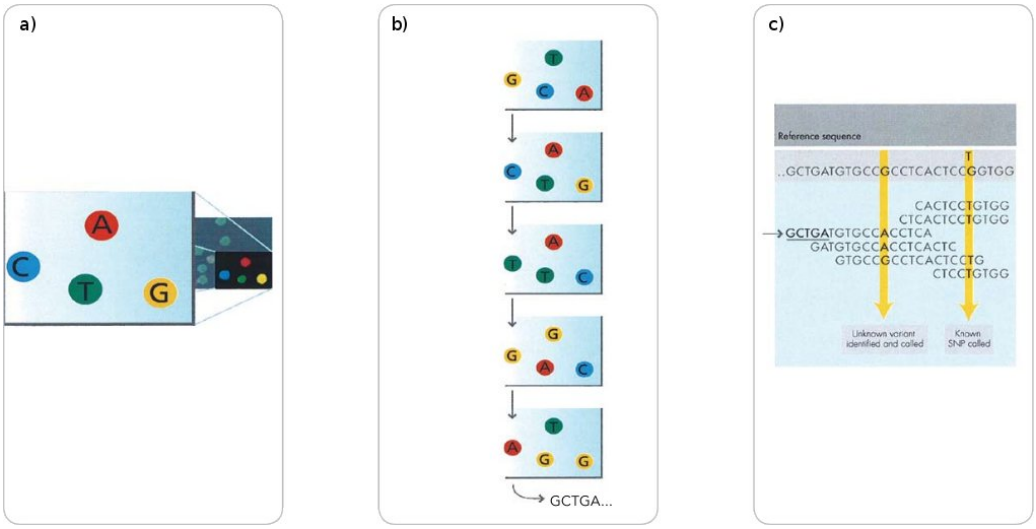

Figure 10: [14] a) The process is repeated and the second step is also recorded. b) The remaining steps are also recorded and a sequence can be read. c) The individual fragments are aligned to a reference genome, and differences can be detected.

7.3 Output

See figure 11 for a sample of the output file. Each row represents a unique interesting region that was found. Had it not been for sequencing problems there would have been 4 more columns with information about the 24 hours after infection sample, with the prefix Ad24. Here is an explanation of the columns in the output file:

- Chr: the number of the chromosome the region is located on.
- Start: the start point on that chromosome where the region begins.
- End: the end point on that chromosome where the region ends.
- Genes: the name of the gene/genes that overlaps the region.
- Type: specifies if the reads covering the region is in an annotated exonic or intronic area, or in an unannotated intergenic area.
- MaxRat: the absolute value of the most differentially expressed sample's normalized expression level's log_2 ratio with the uninfected sample's normalized expression level. $(abs(log_2(\frac{MostDiffExpLvl}{MockNorm})))$

- Ad12Log2Rat: the 12 hours after infection sample's normalized expression level's log_2 ratio with the uninfected sample's normalized expression level. ($log_2(\frac{Ad12Norm}{MockNorm})$)
- Ad12NormDiff: the difference between the 12 hours after infection sample's normalized expression level and the uninfected sample's normalized expression level. ($Ad12Norm - MockNorm$)
- MockNorm: the uninfected sample's normalized expression level.
- Ad12Norm: the 12 hours after infection sample's normalized expression level.
- MockRaw: the uninfected sample's unnormalized expression level. (The actual number of reads covering that region)
- Ad12Raw: the 12 hours after infection sample's unnormalized expression level. (The actual number of reads covering that region)

ID	Chr	Start	End	Genes	Type	MaxRat	Ad12.BTLog2Rat	Ad12.BTNormDiff	Mock.BTNorm	Ad12.BTNorm	Mock.BTRaw	Ad12.BTRaw
314	4	119381400	119381600	-	intergenic	6.32	6.32	172.55	2.19	174.74	14	1152
1053	17	29361600	29361800	-	intergenic	5.68	5.68	54.88	1.1	55.97	7	369
636	9	97109600	97109800	-	intergenic	5.21	5.21	78.81	2.19	81	14	534
1065	17	39215800	39216000	-	intergenic	4.93	-4.93	-80.52	83.25	2.73	532	18
506	7	22771000	22771200	IL6	exonic	4.76	-4.76	-59.38	61.66	2.28	394	15
814	12	19610200	19610400	AEBP2	intronic	4.62	4.62	268.88	11.42	280.31	73	1848
82	1	152006200	152006400	S100A11	exonic	4.37	4.37	52.25	2.66	54.91	17	362
1246	20	49457200	49457400	BCAS4	intronic	4.21	4.21	90.7	5.16	95.86	33	632
1259	21	30370200	30370400	-	intergenic	4.11	4.11	50.87	3.13	54	20	356
677	10	32520800	32521000	-	intergenic	4	4	74.93	5.01	79.94	32	527

Figure 11: The top 10 most differentially expressed regions that were found. Sorted by the MaxRat column.