Report Number: WUCSE-2013-1

2013

Scanner: An Efficient and Accurate Trimming Tool for Illumina Next Generation Sequencing Reads

Xiang Zhou

Recent advances in High-Throughput Sequencing (HTS) technology have greatly facilitated the researches in bioinformatics field. With the ultra-high sequencing speed and improved base-calling accuracy, Illumina Genome Analyzer is currently the most widely used platform in the field. To use the raw reads generated from the sequencing machine, the 3' adapter sequence attached to the real read in the process of ligation needs to be correctly trimmed. This is often done by some inhouse scripts or different packages with various parameters. They either use the Smith-Waterman algorithm or search for an exact match of the 3' adapter sequence. In this... Read complete abstract on page 2.

Scanner: An Efficient and Accurate Trimming Tool for Illumina Next Generation Sequencing Reads

Xiang Zhou

Complete Abstract:

Recent advances in High-Throughput Sequencing (HTS) technology have greatly facilitated the researches in bioinformatics field. With the ultra-high sequencing speed and improved base-calling accuracy, Illumina Genome Analyzer is currently the most widely used platform in the field. To use the raw reads generated from the sequencing machine, the 3' adapter sequence attached to the real read in the process of ligation needs to be correctly trimmed. This is often done by some inhouse scripts or different packages with various parameters. They either use the Smith-Waterman algorithm or search for an exact match of the 3' adapter sequence. In this report, I investigated methodologies as well as the strengths and weaknesses of five representative mainstream adapter trimming tools in order to suggest a direction for other researchers. Furthermore, four sets of detailed analysis were performed to evaluate the performances of these tools. I demonstrated that my adapter trimming method is flexible, accurate and efficient for Next Generation Sequencing (NGS) analysis.

2013-1

LCScanner: An Efficient and Accurate Trimming Tool for Illumina Next Generation Sequencing Reads

Authors: Xiang Zhou

Abstract: Recent advances in High-Throughput Sequencing (HTS) technology have greatly facilitated the researches in bioinformatics field. With the ultra-high sequencing speed and improved base-calling accuracy, Illumina Genome Analyzer is currently the most widely used platform in the field. To use the raw reads generated from the sequencing machine, the 3' adapter sequence attached to the real read in the process of ligation needs to be correctly trimmed. This is often done by some inhouse scripts or different packages with various parameters. They either use the Smith-Waterman algorithm or search for an exact match of the 3' adapter sequence.

In this report, I investigated methodologies as well as the strengths and weaknesses of five representative mainstream adapter trimming tools in order to suggest a direction for other researchers. Furthermore, four sets of detailed analysis were performed to evaluate the performances of these tools. I demonstrated that my adapter trimming method is flexible, accurate and efficient for Next Generation Sequencing (NGS) analysis.

Type of Report: MS Project Report

Master's Project Report

LCScanner: An Efficient and Accurate Trimming Tool for Illumina Next Generation Sequencing Reads

by

Xiang Zhou

Master of Engineering in Computer Science

Washington University in St. Louis, 2012

Research Advisor: Professor Weixiong Zhang

Abstract

Recent advances in High-Throughput Sequencing (HTS) technology have greatly facilitated the researches in bioinformatics field. With the ultra-high sequencing speed and improved base-calling accuracy, Illumina Genome Analyzer is currently the most widely used platform in the field.

To use the raw reads generated from the sequencing machine, the 3' adapter sequence attached to the real read in the process of ligation needs to be correctly trimmed. This is often done by some in-house scripts or different packages with various parameters. They either use the Smith-Waterman algorithm or search for an exact match of the 3' adapter sequence.

In this report, I investigated methodologies as well as the strengths and weaknesses of five representative mainstream adapter trimming tools in order to suggest a direction for other researchers. Furthermore, four sets of detailed analysis were performed to evaluate the performances of these tools. I demonstrated that my adapter trimming method is flexible, accurate and efficient for Next Generation Sequencing (NGS) analysis.

1. Introduction

Recent advances in High-Throughput Sequencing (HTS) technology have greatly facilitated the researches in bioinformatics world. With the ultra-high sequencing speed and improved base-calling accuracy, Illumina Genome Analyzer is currently the most widely used platform in the field [1].

To use the raw reads generated from the Illumina sequencing machine, the 3' adapter sequence attached to the actual read in the process of ligation needs to be correctly trimmed. However, the current available methods for dealing with the raw reads are somewhat arbitrary. This is often done by some in-house scripts or different packages with various parameters. They either use a Smith-Waterman algorithm [2] or search for an exact match to the 3' adapter sequence.

There are two categories of adapter trimming tools, stand-alone programs and integrated ones with other functions such as alignment, quality score tuning, etc. The most popular stand-alone programs include Cutadapt (http://code.google.com/p/cutadapt/), Fastx_clipper from FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and Vectorstrip from EMBOSS package (http://embossgui.sourceforge.net/demo/manual/vectorstrip.html). The most widely used integrated programs include Novoalign (http://www.novocraft.com/main/page.php?s=novoalign), MAQ (http://maq.sourceforge.net/) and SOAP (http://soap.genomics.org.cn/).

The purpose of this tool, **LCScanner**, is to facilitate researchers to better preprocess Illumina sequencing reads by providing an integrated and easy-to-use pipeline, thus making small RNA sequences ready for downstream analysis and processing. LCScanner offers shorter computational time, more flexibility and more accurate trimming results than other tools.

In this study, four sets of detailed analysis were performed (two on plant and two on mammalian using one-adapter and three-adapter sequences) respectively to evaluate the performances of these tools in the following sections.

In summary, four existing tools were chosen in the following analysis to compare with my program: Cutadapt, Clipper, Vectorstrip and Novoalign. In the next few sections, I demonstrated results as well as some interesting observations, analyzed performances and discussed the biological significance of the results processed by these five tools respectively.

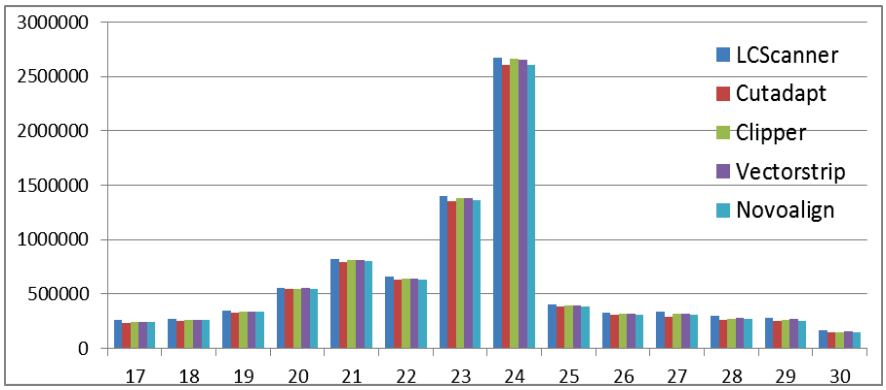
2. Experimental Results

In this section, I performed a thorough analysis on five adapter trimming tools using wide-type Arabidopsis and Human datasets. Due to the sequencing time and cost, it has been increasingly common to generate multiple libraries using a single lane of the Illumina sequencing machine. In order to test the multi-adapter trimming function, two wide-type Arabidopsis datasets were used and analyzed respectively.

2.1 Single Adapter – Arabidopsis

A total of 10,155,795 raw reads sequenced from wide-type Arabidopsis were processed by the five programs using similar parameters: one mismatch for LCScanner, 10% error rate for Cutadapt and Vectorstrip, and default parameters for Clipper and Novoalign. Among them, an average of 84.4% reads can be trimmed by various adapter trimming programs, indicating a relatively high quality of the library. Considering the typical small RNA length, I chose 17-30nt as the length range for investigation. Figure 1 showed the length distributions of the qualified reads (adapter trimmed reads of length 17nt to 30nt) processed by the five programs. Without any surprise, the majority of the trimmed reads rested within 21-24nt, which accords with canonical miRNA length very well [3].

(A)



(B)

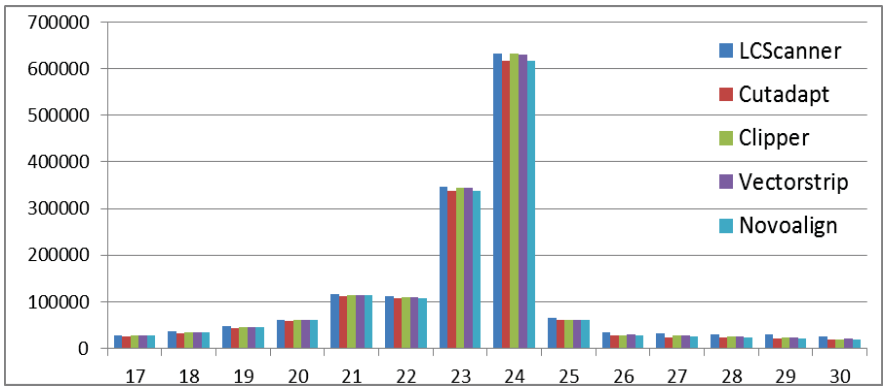


Figure 1: Length distributions of the (A) qualified (adapter trimmed reads of length 17nt to 30nt) and (B) unique qualified reads processed by the five programs with similar parameters.

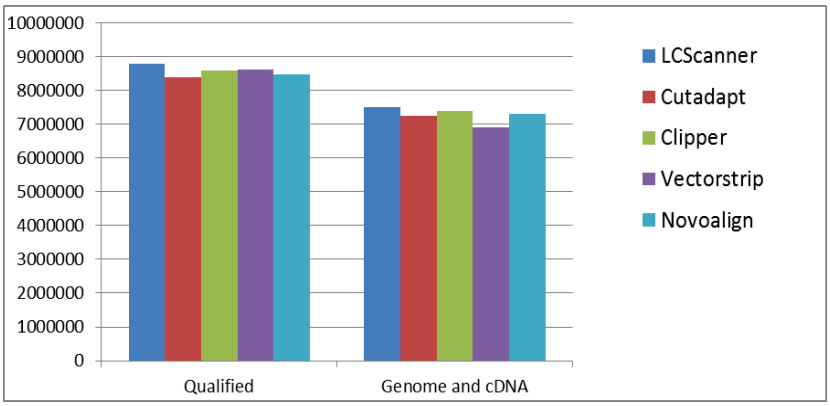
In summary, LCScanner found the most qualified reads, while Vectorstrip came the second (with 2.0% less). The remaining three programs came behind as Clipper, Novoalign and Cutadapt. Although Vectorstrip performed best among all the five programs in terms of total qualified reads, it yielded lowest genome and cDNA mappable reads in the subsequent analysis. This was partly due to the rigid alignment method used by Vectorstrip and the difficulty of adjusting trimming parameters. One notable observation here was the time-cost of these programs: Novoalign was much slower than all other four programs, possibly due to the quadratic time complexity of its mapping procedure (See Performance evaluation).

For the qualified reads, I then mapped them to Arabidopsis genome (TAIR 10) and cDNA sequences as well as miRNA precursors (miRBase v18). The numbers of genome and cDNA mappable reads were similar to the qualified reads, in that LCScanner generated 2% more mappable reads than Clipper and 3-9% more than other three programs.

Further analysis on the length distributions of the miRNA mappable reads (Figure 2) indicated that LCScanner saved not only more qualified reads, but also more miRNA mappable reads which may be involved in various biological functions. Detailed analysis on their lengths showed that my program found slightly more reads of above 22nt. Long miRNAs (lmiRNAs) with these lengths are of extremely importance due to their biological significance. Many of them are DCL3, RDR2 and Pol IV dependent small RNAs, representing the typical heterochromatic small interfering RNA (hc-

siRNA) pathway. They were miRNA derived siRNAs associated with AGO4 and guide DNA methylation at some target loci [10].

(A)



(B)

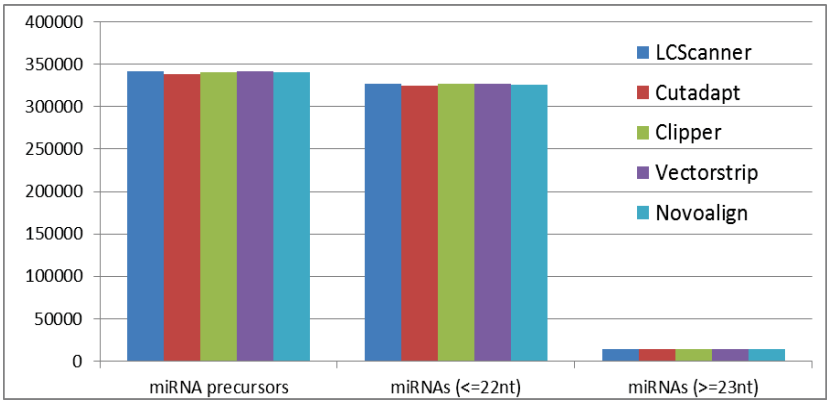


Figure 2: Numbers of the (A) qualified, genome and cDNA mappable reads, as well as (B) reads (with read length separated into <=22nt and >=23nt) mappable to Arabidopsis miRNA precursors and mature miRNAs (with 2nt extension on both ends). Read numbers for miRNAs with length greater or equal to 23nt cannot be shown due to the small quantity of the reads. Although there were no significant differences, the numbers of mappable reads generated from LCScanner were slightly more than those from other programs.

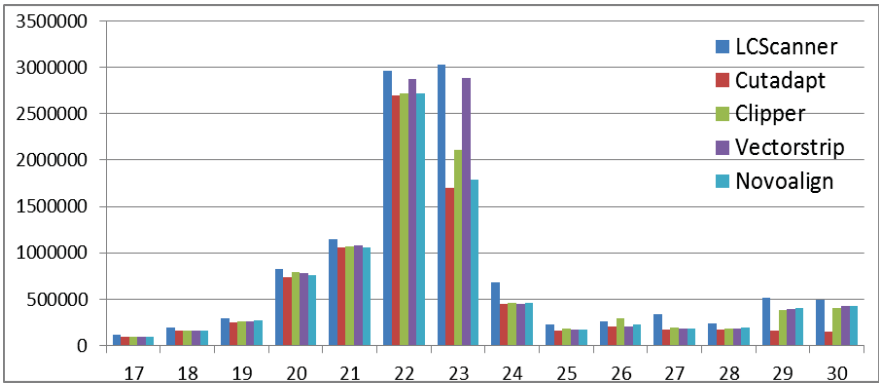
In summary, the above analysis on wide-type Arabidopsis suggested that my adapter trimming program, LCScanner, can adequately and accurately trim Illumina raw sequences, thus saving more biologically usable reads than all other four programs by 2%-9% on a single plant dataset. If an inappropriate adapter trimming tool was chosen, up to one tenth of the qualified reads might be lost in further analysis.

2.2 Single Adapter – Human

A good adapter trimming program should have a versatile ability of dealing with various datasets. A good result on plant domain does not necessarily guarantee a good performance on mammalian domain, since the small RNAs or other biologically functional sequences generated from sequencing machines are different [11]. Therefore, I conducted another test on Human normal skin dataset.

Figure 3. showed the length distributions of the qualified reads processed by the five abovementioned tools. As we can see, the trimming results based on this human dataset has much more variance between programs than those based on the plant dataset. Still, LCScanner took the lead with 11.6-38.8% more qualified reads than the other four programs. On average, it excelled a little bit (7-8%) on 21-22nt length reads, but a lot (43-51%) on 23-24nt length reads compared to the other four programs. Furthermore, LCScanner found more longer reads than all other programs did, which is important for the discovery of the long non-protein coding small RNAs [12].

(A)



(B)

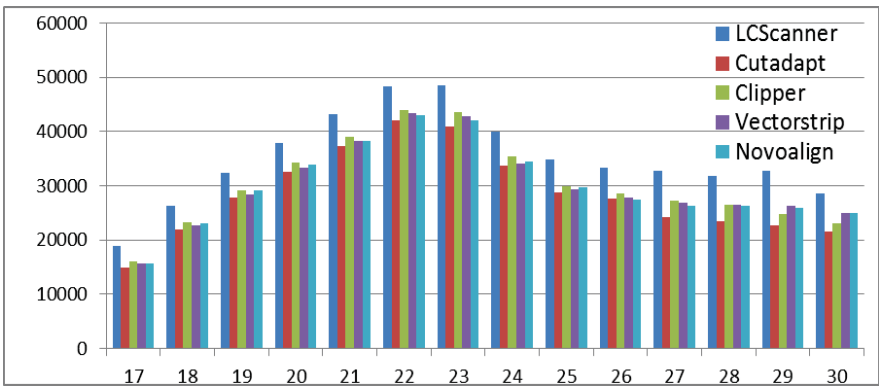
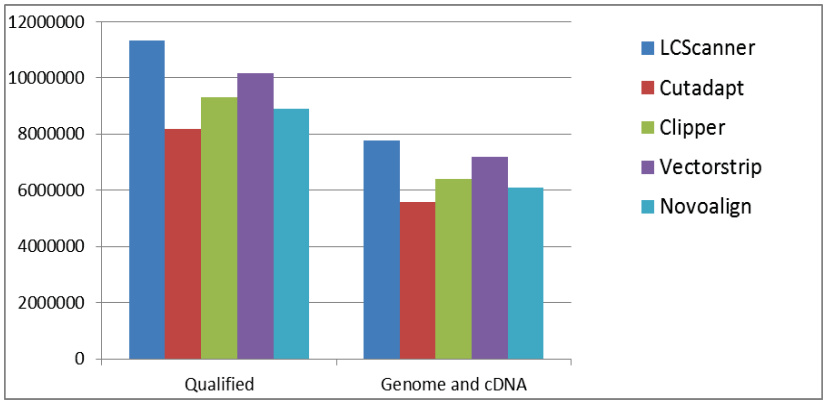


Figure 3: Length distributions of the (A) qualified (adapter trimmed reads of length 17nt to 30nt) and (B) unique qualified reads processed by the five programs with similar parameters.

Reads mappable to human genome and cDNA libraries were plotted in Figure 4. It did not surprise us much that although genome mappable reads differed a lot, miRNA mappable reads had very small variance. Many of the reads saved by LCScanner may map to other functional particles such as tRNAs, rRNAs, snoRNAs, snRNAs, piwi-RNAs and transposons. Unlike the results in Arabidopsis, Vectorstrip remained the second place in terms of both qualified and genome and cDNA mappable reads in human dataset. However, the reads mappable to miRNAs still did not vary much.

(A)



(B)

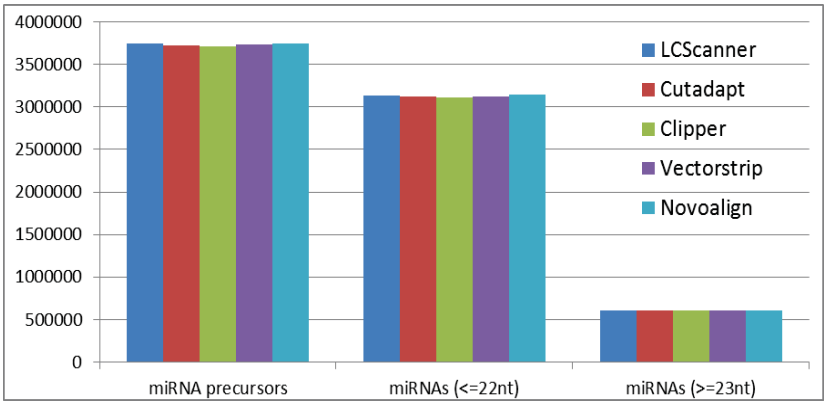


Figure 4: Numbers of the (A) qualified, genome and cDNA mappable reads, as well as (B) reads (with read length separated into <=22nt and >=23nt) mappable to human miRNA precursors and mature miRNAs (with 2nt extension on both ends). Read numbers for miRNAs with length greater or equal to 23nt cannot be shown due to the small quantity of the reads. The numbers of mappable reads generated from LCScanner were more than those from other programs.

In conclusion, by combining the results on these two datasets, I demonstrated that LCScanner excelled all other four mainstream adapter trimming tools. Solely based on the mapping results in the two species above, I believe that Clipper and Novoalign may currently be the best choices if time constraint is not critical.

Novoalign is distinctive among these five tools in that it is the only tool that offers an integrated trimming and mapping functions. However, LCScanner can be easily combined with Bowtie or other popular alignment tools for further downstream analysis.

2.3 Multiple Adapters – Arabidopsis

Of all these five programs, only LCScanner and Cutadapt can deal with those datasets containing multiple adapters, i.e. one or more different adapters may be ligated to a single read to differentiate various libraries in a single lane in order to save sequencing cost. All other three programs have to run multiple times to trim the adapters by feeding them one by one. Moreover, complex extra procedures were also needed to separate the result files. In particular, trimmed reads generated by Clipper, Vectorstrip or Novoalign may exist in multiple files if a read can be differently trimmed by two or more adapters. Even Cutadapt cannot always correctly assign a read to one library once the read can theoretically be trimmed by multiple adapters. Therefore, I wrote two pieces of codes to parse the files generated from Cutadapt and Clipper in order to correctly assign each qualified read to one and only one library.

A total of 3,400,277 raw sequencing reads containing three adapters from Arabidopsis were processed by the three programs with their default parameters (one mismatch for LCScanner, 10% error rate for Cutadapt and default error rate for Clipper). LCScanner found a total of 3,200,518 qualified reads, with 1,648,841, 1,071,771 and 479,906 distinct reads in three libraries respectively. In contrast, Cutadapt and Clipper found a total of 3,343,628 and 2,855,573 reads with duplicates! After the separation process for duplicated reads, the two programs generated 2,780,814 and 2,796,567 qualified reads, which were 13.1% and 12.6% worse than my results. In terms of genome/cDNA

mappable reads and miRNA mappable reads, LCScanner also outperformed the other two programs (Figure 5), although not by much (4%-7%).

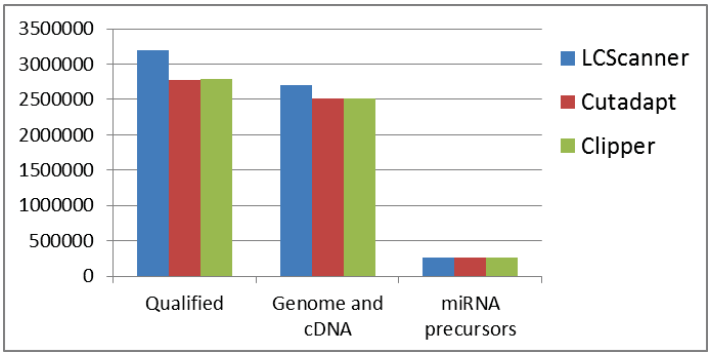


Figure 5: Numbers of the qualified, genome/cDNA mappable and miRNA mappable reads in three-adapter Arabidopsis library, using three trimming programs.

In summary, the results above indicated that the numbers of mappable reads generated by Cutadapt and Clipper were comparable. However, these numbers were still smaller than those generated by LCScanner, since many usable reads were wrongly trimmed or discarded. Seeing that Cutadapt is more user-friendly and has the ability to handle multiple adapters simultaneously as LCScanner does, I conducted a final analysis using these three programs.

2.4 Human Psoriasis Datasets and Its Biological Significance

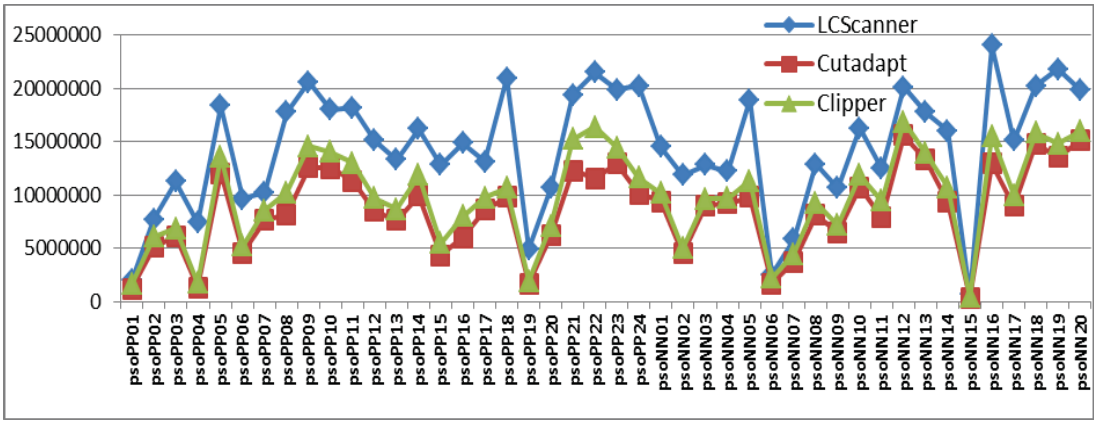
To get a detailed view of how, and by how much, different methods can affect biological results, I conducted a final thorough test on the largest Human psoriasis datasets to date (as of June, 2012) available on NCBI GEO database. In this analysis, I gathered a total of 44 libraries, 24 from psoriatic lesion (PP) and 20 from normal persons (NN).

Since previous results strongly suggested that my tool (LCScanner) and Clipper performed the best, while Cutadapt was ideal for multiple adapters, I chose these three tools for the final analysis.

2.4.1 Datasets Analysis

Figure 6A and Figure 7A showed the numbers of qualified reads processed by these three programs. It was obvious that the blue line was higher than the red and green in almost every library. LCScanner's genome/cDNA mappable reads were also much more than those from Cutadapt (74.7% of my result) or Clipper (85.9% of my result). However, in Figure 7B, not much difference could be observed for miRNA mappable reads (99.2% and 99.7% of the reads can be generated by Cutadapt and Clipper respectively compared to LCScanner).

(A)



(B)

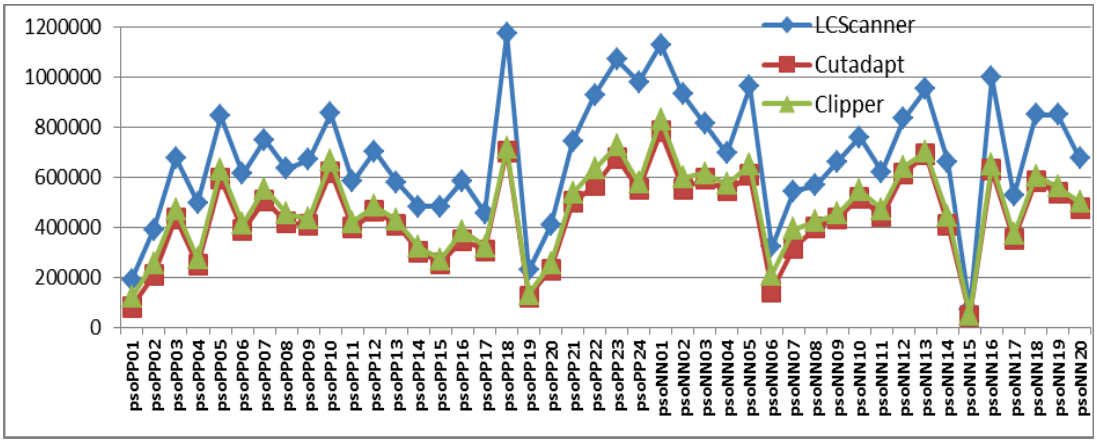
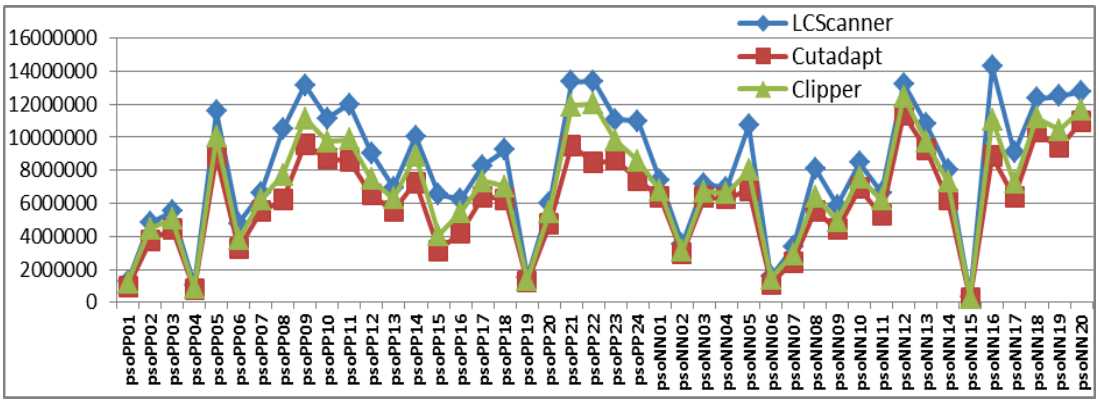
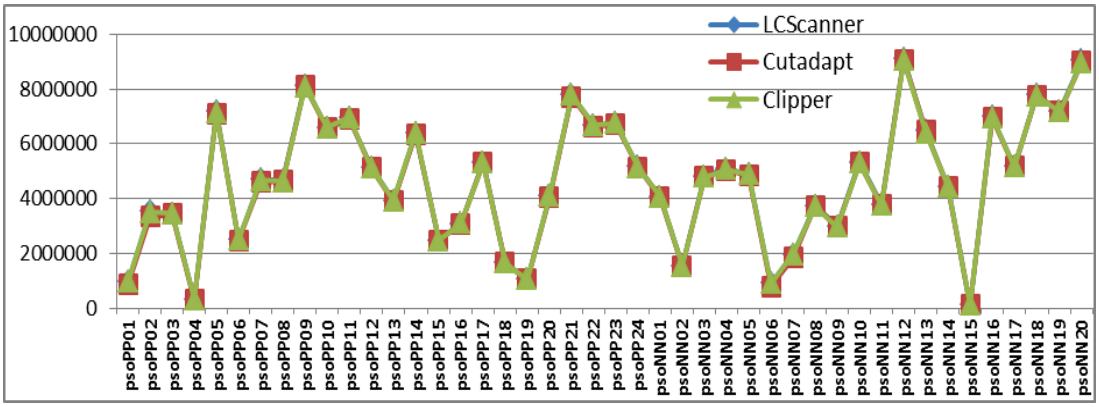


Figure 6: Numbers of (A) all and (B) unique qualified sequences trimmed by three tools (LCScanner, Cutadapt and Clipper). The x-axis indicates the library names, while the y-axis indicates the number of qualified reads.

(A)



(B)



(C)

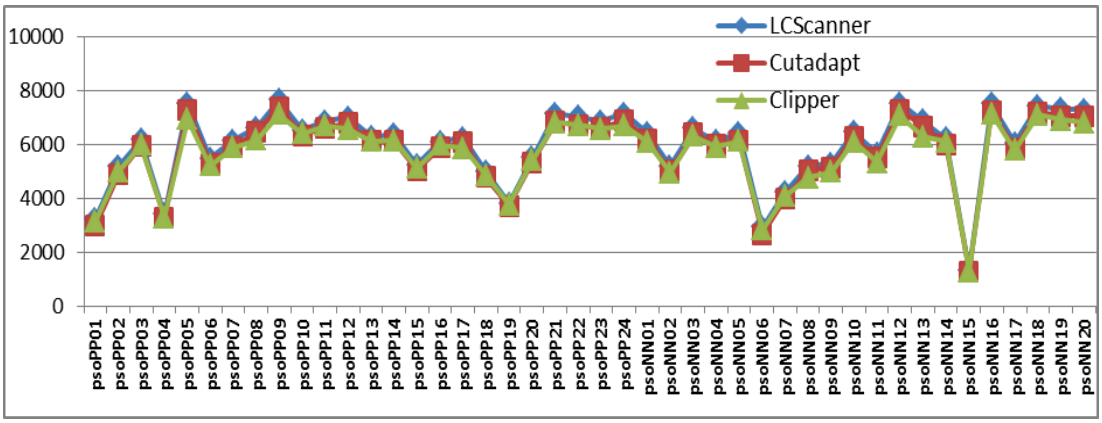


Figure 7: Numbers of (A) all genome and cDNA mappable, (B) all miRNA precursor mappable and (C) unique miRNA precursor mappable reads generated by three tools (LCScanner, Cutadapt and Clipper). The x-axis indicates the library names, while the y-axis indicates the number of mappable reads.

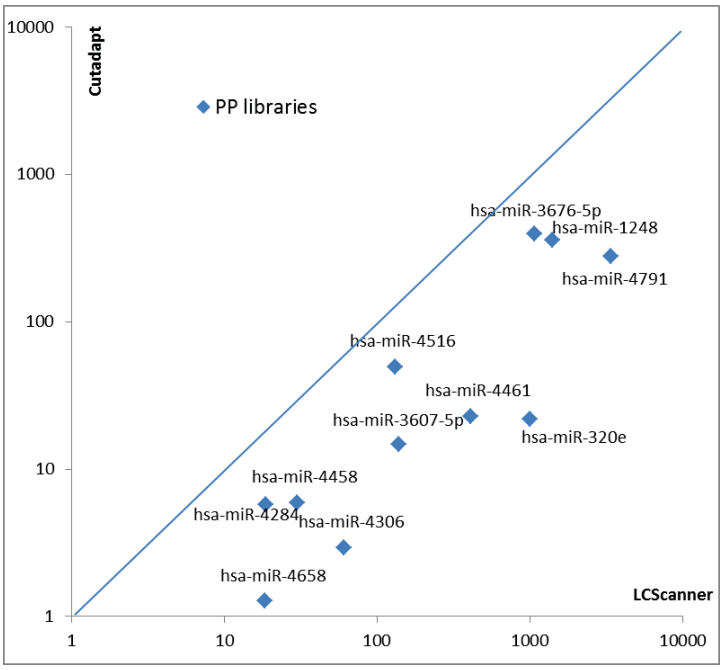
Further analysis of the unique reads (Figure 7C) revealed that the differences indeed existed. LCScanner still generated 3.7% and 4.9% more miRNA mappable reads than Cutadapt and Clipper

respectively. This observation suggested that although other two tools found similar amount of miRNA mappable reads compared to mine, it was actually achieved by producing many duplicate sequences while neglecting reads with very few copies. Only LCScanner can save more unique miRNA mappable reads.

I then examined the expression levels of miRNAs generated by each program. In miRBase v18, there were a total of 1921 miRNAs in the database. Here I define the term "hit" as follows: if there exist at least 10 copies of a sequence (after normalization) in a particular library that can be mapped to miRNA-X, we say that miRNA-X "hits" this library or the library is "hit" by miRNA-X. This term is vital for many experiments, such as differential expression (DE) analysis, novel miRNA discovery, etc. It implies the existence of a particular miRNA mappable reads in a group of datasets. If a program wrongly generated a small "hit" number for a particular read in a dataset, it meant that some libraries contained 0 copies of such read. Therefore, these libraries would contribute to a large DE ratio compared to the actual ratio or hamper the discovery of novel miRNAs due to the low abundances, which is typically the case when finding miRNA star sequences [16, 17, 18]. Since miRNA* sequence supports the release of miRNA duplex from the predicted fold-back structure, the presence of miRNA and its corresponding miRNA* sequences in a dataset would provide compelling evidence for the annotation of novel miRNAs [19, 20].

Figure 8 represented the most differentially expressed (both up- and down-regulated) miRNA abundances generated by LCScanner and Cutadapt. All the points were well below the diagonal line, indicating that the miRNAs generated by LCScanner had higher abundances than those generated by Cutadapt. Further analysis into these individual miRNAs revealed important biological functionalities for human body.

(A)



(B)

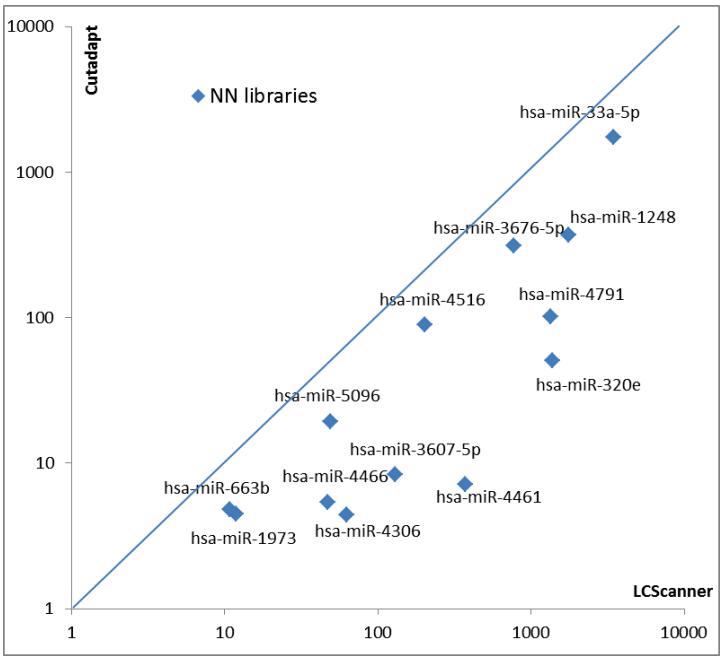


Figure 8: Plots of the most differentially expressed miRNAs from LCScanner and Cutadapt in PP (A) and NN (B) libraries. The x-axis indicates the expression level of miRNAs from LCScanner, while the y-axis indicates the expression level of miRNAs from Cutadapt. The diagonal line represents the points where the expression levels of miRNAs are the same in two programs.

Of all human miRNAs, I was particularly interested in those that have large variances in differential expression (DE) ratios between psoriatic (PP) and normal (NN) samples processed by LCScanner and Cutadapt. A total of 6 miRNAs were picked out according to the above criteria (differing by at least 2-fold in DE ratios and with at least 10 copies in PP or NN libraries) for further analysis. Specifically, hsa-miR-3182 and hsa-miR-3123 have much higher DE ratios, while hsa-miR-4466, hsa-miR-3178, hsa-miR-4461 and hsa-miR-663b have much lower DE ratios in my results compared to Cutadapt's.

2.4.2 Biological Significance

I then performed miRNA target predictions (miRDB, http://mirdb.org/miRDB/) [21, 22] on these 6 miRNAs. Target genes with top 5 target scores of hsa-miR-3182 were: ZDHHC20, RAB2A, NEU3, USP46 and GDA. Gene association studies (NextBio, http://www.nextbio.com/b/nextbio.nb) [23, 24] showed that three of these top five genes (ZDHHC20, NEU3 and GDA) presented up-regulation in psoriatic lesions versus normal controls in previous studies. Cutadapt presented no fold change on hsa-miR-3182, while LCScanner's results showed a top increase by 6.97 folds in PP library among all miRNAs. Similarly, hsa-miR-3123 demonstrated a 4.14-fold increase in PP library processed by LCScanner, but there was no change in Cutadapt's results. Indeed, three out of top five target genes (FNDC3B, ABL2 and TJP1) of hsa-miR-3123 were over-expressed in PP versus NN libraries in previous studies.

Overexpression of FNDC3B can increase the activation of the signal transducer and activator of transcription 3 (STAT3) signaling pathway to promote cell proliferation and tumor formation [25]. v-abl Abelson murine leukemia viral oncogene homolog 2 (ABL2) is also up-regulated in psoriatic lesions. Abl2 non-receptor tyrosine kinase acts downstream of the EGF receptor and Src tyrosine kinases, leading to uncontrolled cell division and proliferation [26]. TJP1 gene (tight junction protein 1, zona occludens 1), which encodes a protein located on a cytoplasmic membrane surface of intercellular tight junctions and plays an important role in the regulation of cell migration, is overexpressed in various types of carcinomas [27]. Up-regulation is observed in human melanoma cells, which seems to be related to invasiveness of the cells [28, 29].

For hsa-miR-4466, LCScanner's results showed no significant change of the expression levels in PP versus NN libraries. However, Cutadapt mistakenly treated it as a remarkable overexpression of 5.85-fold change from NN to PP libraries, while hsa-miR-4466 is putatively considered to be down-regulated in psoriatic skins [30, 31].

Similarly, reduced expression of hsa-miR-3178 (0.35 fold in PP versus NN library) [32] and invariance in the expression level of hsa-miR-4461 (1.09 folds in PP versus NN library) in previous studies both perfectly coincided with LCScanner's results. In contrast, Cutadapt treated hsa-miR-3178 as no change in its expression level between two libraries and hsa-miR-4461 as overexpression (1.89-fold increase). Notably, hsa-miR-663b exhibited a significant increase in both LCScanner's (1.04-fold increase) and Cutadapt's results (3.09-fold increase).

All these DE miRNAs discussed above consolidated the credibility of my trimming program. Nonetheless, due to the numerous uncertainties and exceptions in this biological field, we cannot take the results of any computation tools for granted. Experimental results must be gathered before any conclusions can be drawn.

In this section, I have demonstrated and explained results and observations from five mainstream adapter trimming tools using various metrics. Evidences showed that my program, LCScanner, was the best of these five in terms of versatility, adaptability and usability of the trimmed reads. In this paper, I further buttressed these results by describing the trimming procedure in the Material and Methods section.

3. Performance Evaluation

In this section, I performed two sets of analyses to evaluate the efficiency of the five mainstream adapter trimming programs. The first test was on Arabidopsis dataset and the other one was on human psoriasis datasets.

Detailed analysis results of Arabidopsis and human samples were shown in Figure 9. We can easily observe that LCScanner performed the best in both datasets. It only used 73.4% of the time, on average, to finish trimming process compared to the second-best program, Vectorstrip. However, the problems with Vectorstrip mentioned in previous sections make it an inferior adapter trimming tool for Illumina Next Generation Sequencing samples. It did not surprise us much that Novoalign ranked the lowest in this experiment, since it used CPU intensive algorithm for alignment and completed mapping together with trimming.

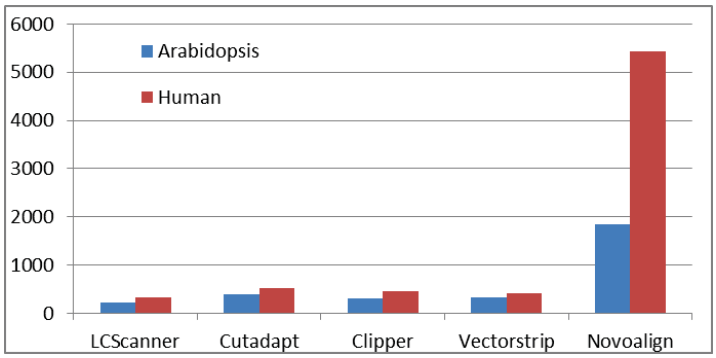


Figure 9: Time cost (in seconds) of the five trimming programs on the Arabidopsis and a single human normal skin datasets using similar trimming parameters.

Detailed analysis of time complexity of the three programs on the 44 human psoriasis datasets was shown in Figure 10. The result showed that my adapter trimming program, LCScanner, ranked second among the three, with Clipper performing the best and Cutadapt performing the worst. I also noticed that in some NN samples, LCScanner performed similarly or even better than the best publicly available tool, Cutadapt.

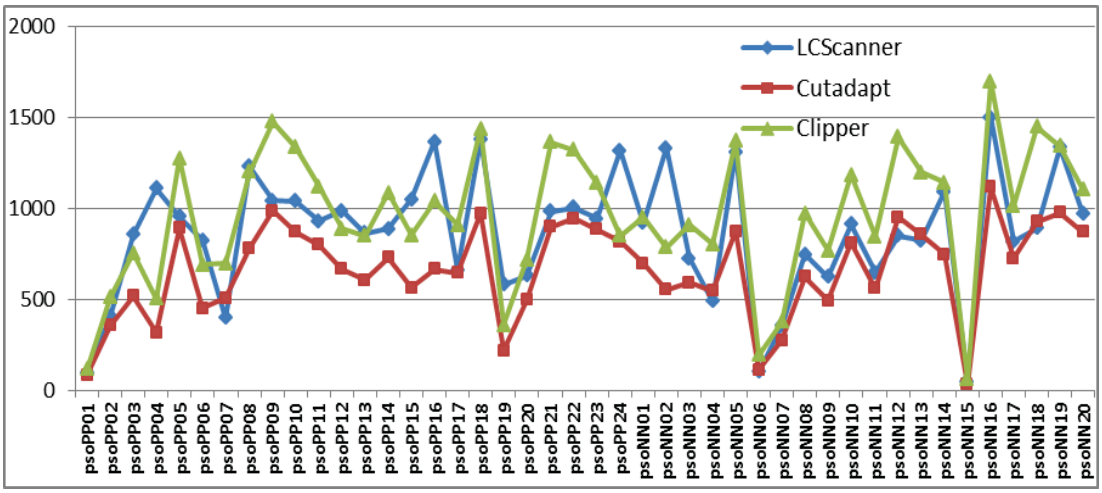


Figure 10: Time cost (in seconds) of the three trimming programs on the 44 human psoriasis datasets using similar trimming parameters.

The memory consumption of these five programs was comparable, with Novoalign taking the greatest amount of memory (data not shown).

4. Material and Methods

Individual deep sequencing libraries were downloaded from NCBI GEO database (http://www.ncbi.nlm.nih.gov/geo/). Samples were processed to remove sequencing adapters using default parameters. Adapter-trimmed reads were then mapped to Arabidopsis genome/cDNA TAIR databases (TAIR9, http://www.tair.org) and human genome sequences (UCSC genome browser, hg19, GRCh37, http://genome.ucsc.edu/) respectively. The sequences of mature miRNAs were obtained from miRBase v18 (http://www.mirbase.org/).

Read alignment were done by Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) [34] allowing zero mismatches and finding the best alignment.

4.1 Trimming of the Adapters

I applied a two-step cascading trimming procedure in my algorithm – it first finds the Longest Common Substring (LCS), starting at the first nucleotide of the given adapter sequence, between the raw read and the adapter. If a read satisfies the abovementioned criteria, I call it "pre-qualified read". Then for those remaining raw reads (this is typically much smaller than the original dataset), the program keeps finding the LCS between the raw read and the adapter starting from the next nucleotide of the given adapter sequence. Finally I trim the satisfying read with initial "N"s and check if there are still any other "N"s in the middle. If yes, I throw the read away, otherwise, put it in the pre-qualified read pool unless it is less than the user-specified (or default) length.

Note that the finding of the LCS can be replaces by finding the best n-mismatch common substrings (n can be set to 1, 2 or 3). The program filters out trimmed reads based on a user-defined sequence length. If no value is specified, a default value of 17nt will be used. LCScanner also has a user-specified quality control parameter to provide the function of removing low quality reads based on Phred quality score [38, 39] and trimming "N"s from the reads.

Seeing that the proposed method, LCScanner, considers quality score when doing the trimming, is not alignment based, and implemented in pure C language, it excels several mainstream tools in terms of accuracy (correctness) and time complexity.

4.2 Differential Expression Analysis

I downloaded human miRNAs from miRBase v18 (http://www.mirbase.org/). I used an in-house script to extend the miRNA sequences with 2nt on both ends. I then mapped the adapter trimmed sequences in each library to the extended miRNA sequences with perfect match. If a read can be mapped to multiple miRNAs, I counted each of them as a valid mapping. Read counts were then normalized to the average number of genome and cDNA mappable reads to adjust for variations between samples. Specifically,

Normalized # of read = (Raw # of read) * (average number of genome and cDNA mappable reads) / (number of genome and cDNA mappable reads in this sample).

Simple fold change was then calculated for each miRNA sequence. I selected differentially expressed miRNAs with at least two fold changes and ten copies in combined PP or NN libraries.

5. Conclusion

This report presents three major contributions to the current small RNA research community. First, I developed a flexible, accurate and efficient adapter trimming tool for Next Generation Sequencing analysis. Second, I performed a thorough analysis and comparison on the accuracy and efficiency of several mainstream adapter trimming programs. The results will hopefully serve as a preliminary note for other researchers. Third, the biological significance observed in the experiments will benefit and shed some light on future studies.

In conclusion, LCScanner, as a part of an in-house small RNA analyzing pipeline in our laboratory (Zhang's Computational Biology Lab, Washington University in St. Louis), has already been used in various large projects focused on Moss, Arabidopsis, Rice, Mouse and Human. It has been thoroughly and extensively tested and proved to be easy to use on any Unix-based machines.

Acknowledgments

I would like to thank Professor Weixiong Zhang for his constant help and guidance over the months I spent working on this report. I would also like to thank Professor Jeremy Buhler and Dr. Sharlee Climer for serving on my master's project committee.

Additionally, I would like to take this opportunity to thank my friends and family, especially my parents for their continuous support and encouragement.

A special thanks goes to the many graduate students and distinguished faculty within my department who have reviewed this report and helped support the related research.

Xiang Zhou

Washington University in St. Louis
December 2012

References

[1] Elaine R. Mardis. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9: 387-402, 2008.

[2] T. F. Smith and M. S. Waterman. Comparison of Biosequences. *Adv. Appl. Math*, vol. 2, pp. 482-498, 1981.

[3] Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281-297, 2004.

[4] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*, 17(1):10–12, 2011.

[5] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*, 38:1767-71, 2010.

[6] FASTQ format, http://en.wikipedia.org/wiki/FASTQ_format

[7] Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, 16(6), 276–277, 2000.

[8] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11), 1851–1858, 2008.

[9] Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5), 713–714, 2008.

[10] Chellappan P., Xia J., Zhou X., Gao S., Zhang X., Coutino G., Vazquez F., Zhang W., Jin H.. siRNAs from miRNA sites mediate DNA methylation of target genes. *Nucleic Acids Res*, 38: 6883–6894, 2010.

[11] Carrington JC, Ambros V. Role of microRNAs in plant and animal development. *Science*, 301: 336-338, 2003.

[12] Katiyar-Agarwal S., Gao S., Vivian-Smith A., Jin H. A novel class of bacteria-induced small RNAs in Arabidopsis. *Genes & Dev*, 21: 3123–3134, 2007.

[13] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic Local Alignment Search Tool. *J. Mol. Biol.*, vol. 215, pp. 403-410, 1990.

[14] Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, 442:199-202, 2006.

[15] Kawaji H, Nakamura M, Takahashi Y, Sandelin A, Katayama S, et al. Hidden layers of human small RNAs. *BMC Genomics*, 9: 157, 2008.

[16] Lu C, Kulkarni K, Souret FF, Muthuvalliappan R, Tej SS, et al. MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Res*, 16:1276-1288, 2006.

[17] Rajagopalan R, Vaucheret H, Trejo J, Bartel DP. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev*, 20:3407-3425, 2006.

[18] Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, Carrington JC. High-throughput sequencing of Arabidopsis microRNAs: Evidence for frequent birth and death of MIRNA Genes. *PLoS ONE*, 2:e219, 2007.

[19] R. Sunkar, X. Zhou, Y. Zheng, W. Zhang and J-K. Zhu, Identification of novel and candidate miRNAs in rice by high throughput sequencing. *BMC Plant Biology*, 8:25, 2008.

[20] G. Jagadeeswaran, Y. Zheng, N. Sumathipala, H. Jiang, E. Arese, J.L. Soulages, W. Zhang and R. Sunkar. Deep sequencing of small RNA libraries reveals dynamic regulation of conserved and novel microRNAs and microRNA-stars during silkworm development. *BMC Genomics*, 11:52, 2010.

[21] Xiaowei Wang and Issam M. El Naqa. Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, 24(3):325-332. 2008.

[22] Xiaowei Wang. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA*, 14(6):1012-1017, 2008.

[23] Disease Atlas, http://www.nextbio.com/b/search/da.nb

[24] Ilya Kupershmidt, Qiaojuan Jane Su, Anoop Grewal, Suman Sundaresh, Inbal Halperin, James Flynn, Mamatha Shekar, Helen Wang, Jenny Park, Wenwu Cui, Gregory D Wall, Robert Wisotzkey, Satnam Alag, Saeid Akhtari, Mostafa Ronaghi. Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS One*, 5(9). pii: e13066, Sep 2010.

[25] Chen CF, Hsu EC, Lin KT, Tu PH, Chang HW, Lin CH, et al. Overlapping high-resolution copy number alterations in cancer genomes identified putative cancer genes in hepatocellular carcinoma. *Hepatology*, 52: 1690-1701, 2010.

[26] Gil-Henn H, Patsialou A, Wang Y, Warren MS, Condeelis JS, Koleske AJ. Arg/Abl2 promotes invasion and attenuates proliferation of breast cancer in vivo. *Oncogene*, Epub, July 9, 2012.

[27] Matter K, Aijaz S, Tsapara A, Balda MS. Mammalian tight junctions in the regulation of epithelial differentiation and proliferation. *Curr Opin Cell Biol*, 17:453-458, 2005.

[28] K.S. Smalley, P. Brafford, N.K. Haass, J.M. Brandner, E. Brown, M. Herlyn. Up-regulated expression of zonula occludens protein-1 in human melanoma associates with N-cadherin and contributes to invasion and adhesion. *Am. J. Pathol.*, 166, pp. 1541-1554, 2005.

[29] H. Schluter, I. Moll, H. Wolburg, W.W. Franke. The different structures containing tight junction proteins in epidermal and other stratified epithelial cells, including squamous cell metaplasia. *Eur J Cell Biol*, 86, pp. 645-655, 2007.

[30] Abraira, V.E., T. Satoh, D.M. Fekete, and L.V. Goodrich. Vertebrate Lrig3-ErbB interactions occur in vitro but are unlikely to play a role in Lrig3-dependent inner ear morphogenesis. *PLoS One*, 5:e8981, 2010.

[31] Hong KK, Cho HR, Ju WC, Cho Y, Kim NI. A study on altered expression of serine palmitoyltransferase and ceramidase in psoriatic skin lesion. *J Korean Med Sci*, 22:862-867, 2007.

[32] Cattaruzza, M., K. Schafer, and M. Hecker. Cytokine-induced down-regulation of zfm1/splicing factor-1 promotes smooth muscle cell proliferation. *J. Biol. Chem*, 277:6582-6589, 2002.

[33] Minoche, A. E., J. C. Dohm and H. Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome biology*, 12:R112. DOI: 10.1186/gb-2011-12-11-r112, 2011.

[34] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10:R25, 2009.

[35] Kircher, M, Stenzel U, Kelso, J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome. Biol.*, 10, R83, 2009.

[36] Altshuler D., Pollara V., Cowles C., Van Etten W., Baldwin J., Linton L., Lander E., Pollara V., Cowles C., Van Etten W., Baldwin J., Linton L., Lander E.,

Cowles C., Van Etten W., Baldwin J., Linton L., Lander E., Van Etten W., Baldwin J., Linton L., Lander E., Baldwin J., Linton L., Lander E., Linton L., Lander E., Lander E. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407:513-516, 2000.

[37] Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., Russ, C., Lander, E. S., Nusbaum, C., and Jaffe, D. B. Quality scores and snp detection in sequencing-by-synthesis systems. *Genome Res*, 18(5):763-770, 2008.

[38] Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, 8 (3): 175-185, 1998.

[39] Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 8 (3): 186-194. doi:10.1101/gr.8.3.186, 1998.