Selection of differentiating features in RNA-Sequencing of thyroid cancer samples.

Tomasz Stokowy, *Silesian University of Technology, Institute of Automatic Control; Cancer Center and Institute of Oncology, Nuclear Medicine and Endocrine Oncology Department;Gliwice, Poland*
(18.08.2011, *dr hab. inż. Krzysztof Fujarewicz, Silesian University of Technology*)

Abstract

Gene expression profiles are frequently differentiating between types of cancer. In thyroid cancer such differentiation has been already shown with DNA microarrays. Unfortunately markers that could describe some thyroid tumor types are still missing. To reveal new prognostic markers we have analysed thyroid tumor samples with RNA sequencing.

Quality control, filtering, alignment with splice junction mapping, gene expression calling and comparative analysis were performed. Results were presented as a list of differentiating features sorted by Fragments Per Kilobase of exon per Million reads (FPKM).

Significantly changed transcriptome regions were found and annotated by chromosome location. Biological interpretation of the results has proved detection of already known markers as well as novel ones. New potential revealed markers can be divided into following groups: genes, pseudogenes, expressed intronic regions, expressed promoter fragments, novel exons detected.

Conclusion: RNA sequencing data analysis results will significantly improve current knowledge of molecular markers of thyroid cancer.

1. Introduction

Thyroid tumor types has been already investigated with many methods of molecular biology. It has been shown that most frequent in the human population papillary thyroid (PTC) cancer differs significantly in the manner of gene expression [1] as well as short RNA expression [2]. Other thyroid tumor types – follicular thyroid cancer(FTC) and benign follicular thyroid adenoma (FA) has not been described well yet, however some studies on this problem can be found in the literature [3]. Pathologists still find it difficult to differentiate between them with classical medical approaches. We are taking into consideration also benign cold thyroid nodules (CTN) and autonomous and active hot thyroid nodules (HTN) together with their surrounding tissues (ST).

There were analysed 8 samples with Illumina RNA sequencing to evaluate usefulness of this method in molecular markers of thyroid cancer detection. The samples investigated were already analysed with other molecular biology methods so some tumor related markers are already known.

In case there could be satisfactory results of this pilot study obtained, further experiments with statistically significant number of samples will be performed to find markers differentiating between FTC and FA.

2. Data

Data acquired for 8 samples was described as 76bp long single end reads (2PTC and 2FTC) and 101bp paired end reads (1HTN, 1CTN, 2ST). All the sequences were stored in fastq format to allow flexible analysis. To investigate the quality of particular base read in all the sequences Sanger Phred quality score was used:

$$Q_{Sanger} = -10\log_{10} p$$

where p is a probability of single base to be read incorrectly. There was selected threshold of mean Q = 20 for a sequence to be accepted in quality control which stands for p=0.01. On the figure 1 a per base quality plot for HTN sample is presented. Observed boxplots confirm known Illumina pipeline effects like cluster reannotation after a few cycles which results in quality increase and dropdown of average quality in the last reads. Finally, after quality control about 63 million paired reads were accepted, which stands for ~6.5 billion of bases read per sample (in a single end). Additionally other indicators such as GC content, equal in average per sample to 45% and not significant amount of over expressed sequences per read indicated good quality of data. The quality score investigations were performed with FastQC software [4] and filtering with appropriate perl scripts.

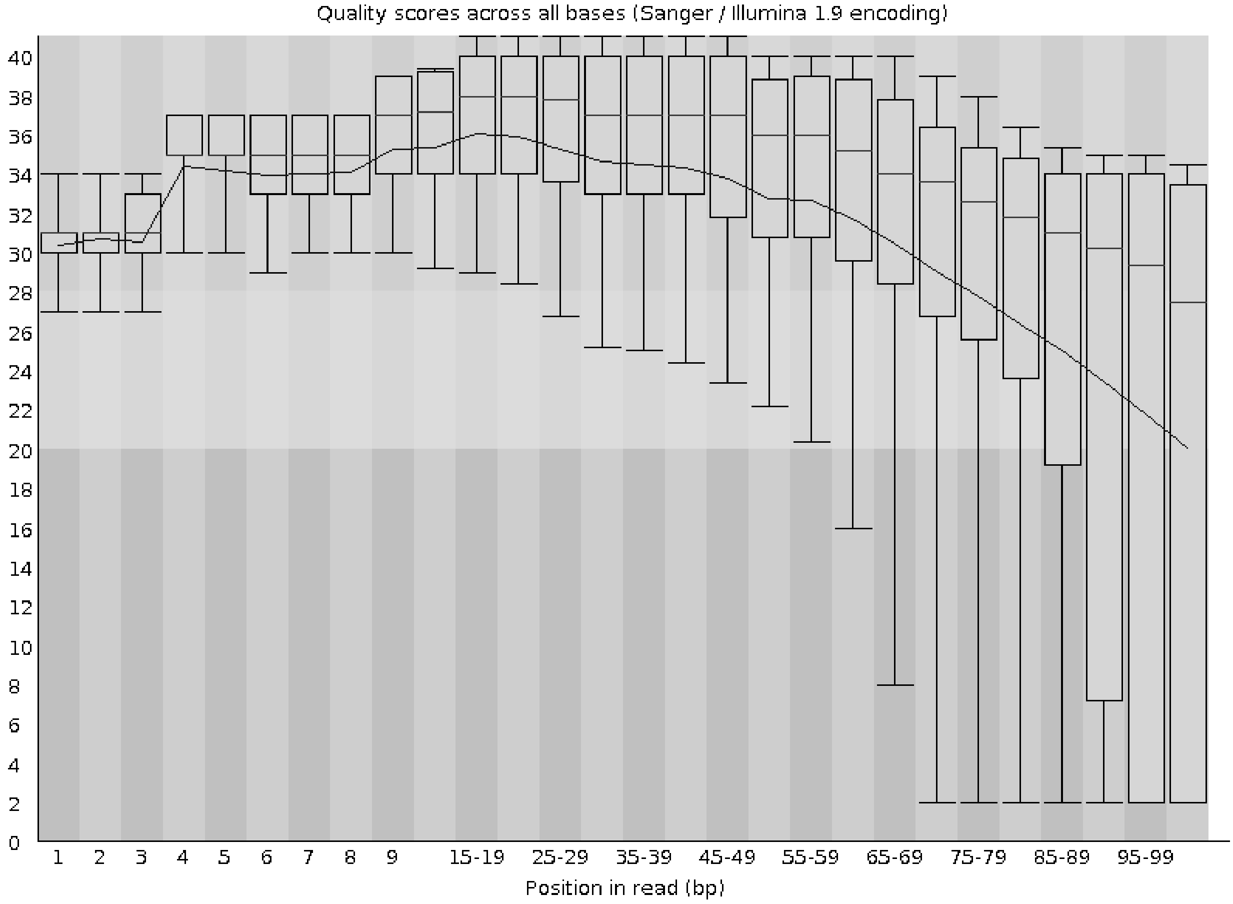

Fig.1. Per base sequence quality in HTN sample.

3. Analysis

The fastq files for each sample were mapped against human genome hg19 with the tophat program [5]. For paired end reads additionally the –r parameter was set to 400 indicating the distance between forward and reverse read as suggested from the medium fragment size detected during library preparation . There was 1 mismatch allowed per sequence during mapping to allow alignment of SNPs containing fragments and take into account sequencing errors.

Application of tophat led to mapping of both, unspliced and spliced reads. Exemplary reads spanning exons in ITIH5 gene are presented in figure 2. Aligned reads are marked in grey and spanning reads in black. In the reference track on the bottom of the graph blocks depict exons, lines introns and arrows indicate the direction of the transcription.

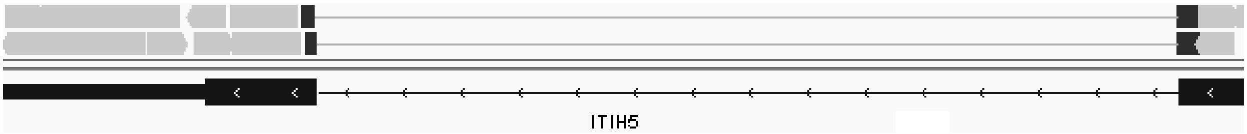

Fig.2. Reads spanning exons of ITIH5 gene.

In more than 50% of all reads mapping found at least one match to the reference . Aligned sequences distribute along the whole genome reflecting gene expression in the respective sample. To browse for the known effects observed in the samples Integrated Genome Viewer (IGV) [6] were used. Results of mapping are presented in figure 3. The different tracks indicate coverage of reads (1), splice junctions (2), compilation of mapped reads (3) also shown in the upper part of figure 2 and finally the RefSeq reference (4).

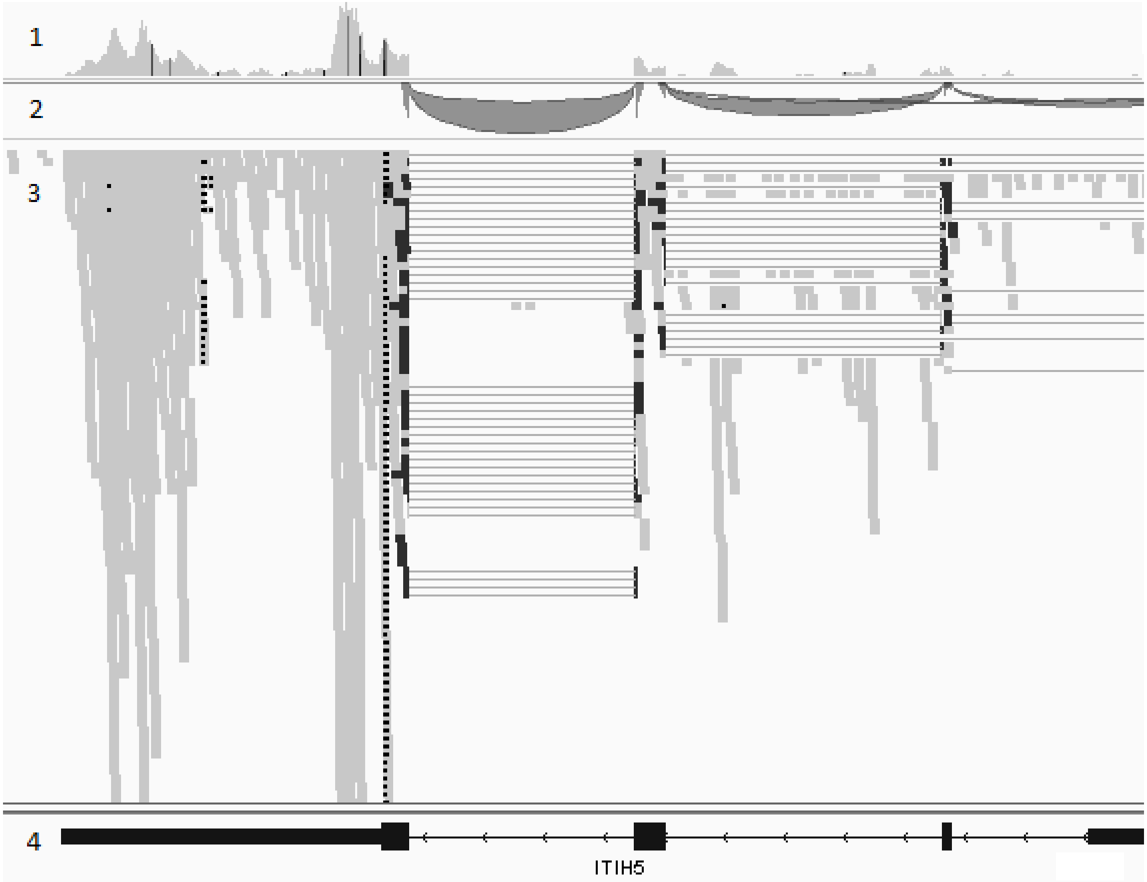

Fig.3. Reads mapped to the reference in part of ITIH5 gene.

For each of the samples the tophat algorithm was applied with the same mapping parameters. Unfortunately the coverage of the reads can not be directly used as an expression measure because the over all number of accepted reads for each sample is usually different. Additionally, reads sequenced on different flowcells need an universal transcript abundance measure that could account for a possible batch effects in the study. To account for the total number of reads and to address the potential bias a measure of transcript abundance has been proposed as an universal mRNA expression marker.

The measure calculates relative transcript abundance as Fragments per kilobase of exon per million reads mapped (FPKM). This particular indicator gives the local coverage independently of the total number of reads. Assembly of transcripts and calculation of their abundance s has been performed with Cufflinks v. 1.0.3 [7].

Unfortunately, further comparative analysis of the samples was significantly limited at this point, due to the small number of samples in each class. FPKM fold change values were calculated to observe the differences between particular samples. However, our sequencing data allowed the comparison to known effects from other biological experiments that included material from our samples.

4. Results

For all 8 samples reads acquired from RNA sequencing has been mapped. Known down regulated genes in FTC versus FA and normal thyroid tissue [3] like CA4, ITIH5, ELMO1 show almost no coverage in RNA-Seq samples from FTC. In figure 4 two FTC samples (1,2) were compared to a RNA-Seq sample from healthy thyroid (3) studied in the Illumina Body Map project. Both coverage and splice junction tracks are shown for all 3 samples.

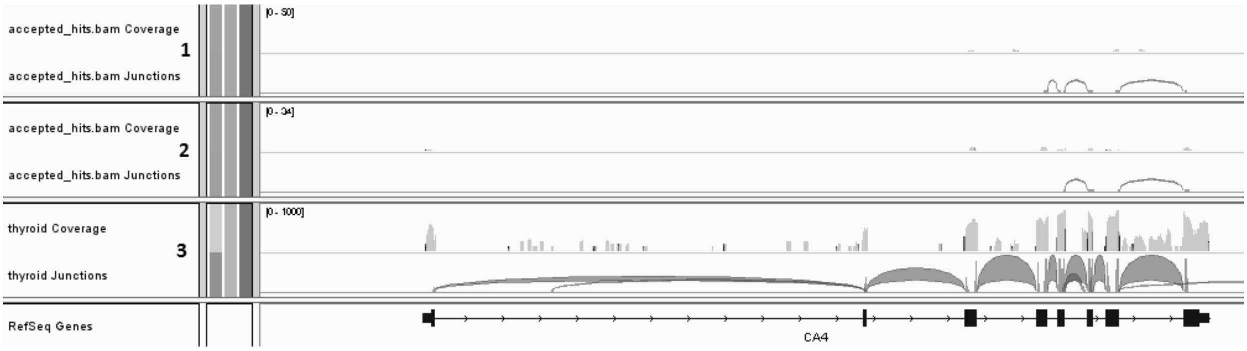
accepted_hits.bam Coverage
1
accepted_hits.bam Junctions
[0 - 50]
accepted_hits.bam Coverage
2
accepted_hits.bam Junctions
[0 - 34]
thyroid Coverage
3
thyroid Junctions
[0 - 1000]
RefSeq Genes
CA4

Fig.4. Comparison of FTC samples (1,2) and healthy thyroid sample (3) in the region of CA4 gene.

Similar differential expression was determined for FTC versus PTC samples on the base of calculated FPKM values. 1253 regions with a |FPKM log ratio| > 2 were found. All these regions were annotated with the precise chromosome location. One such region with a log ratio ≈ 5.5 was found in the area of the LRP5L gene. In this example not only the expression level is different between FTC and PTC but also the amount and location of splice junctions within that gene (figure 5). The ratio of changes between FTC and PTC entities is a confirmation of known biological differences between follicular and papillary thyroid cancer.

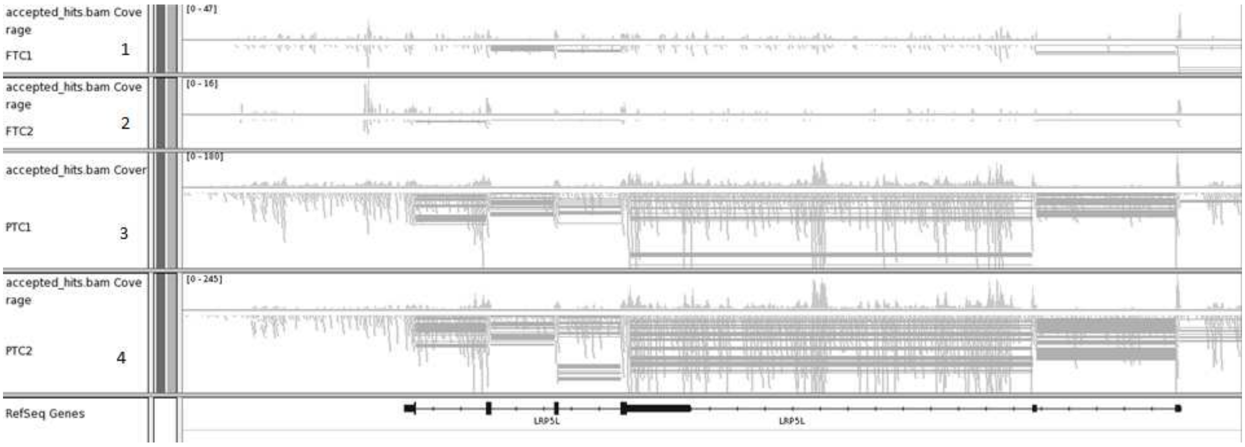

Fig.5. Differentiation of FTC samples (1,2) and papillary thyroid samples (3,4) in the region of LRP5L gene.

Regions described as differentially expressed between samples were not only annotated as genes. There have been also strong effects between samples in the coverage of RefSeq exons, as well as outside of genes. Less than 30% of all events could be described with Ensembl data base as pseudogenes, retrogenes or exons of genes not described by RefSeq.

High resolution RNA sequencing allows to measure expression and investigate SNP/mutation effects at once. Single nucleotide polymorphisms are easily recognizable from sequencing errors, especially for the regions with a coverage > 20. An example of a sequencing error (1) compared to a SNP (2) can be observed in figure 6.

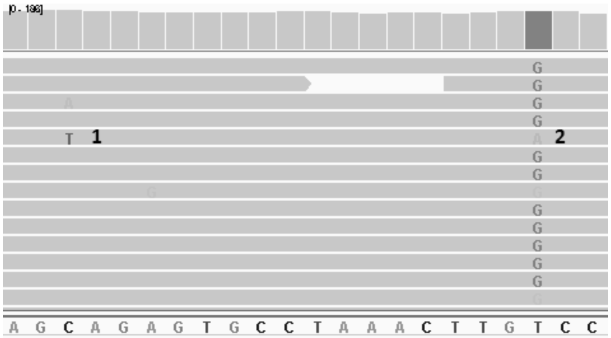

Fig.6. Sequencing error(1) versus SNP(2) in high zoom browsing of RNA-Seq data

5. Discussion

Sequencing errors are usually contained in sequences with low quality (low Phred score) and could therefore be detected even before mapping. It is worth mentioning that sequencing results for the same tissue are very reproducible. Expression pattern of thyroid nodules and their surrounding tissues are very similar. Figure 7 shows differences between one patient (line 1,2) an another (line 3,4) but almost no difference between samples from the same tissue : (1) versus (2) and (3) versus (4).

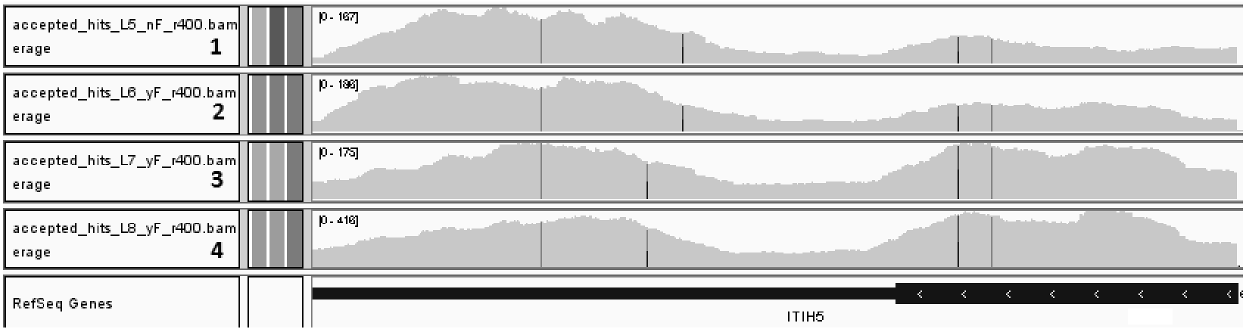
accepted_hits_L5_nF_r400.bam
erage
1
[0 - 167]
accepted_hits_L6_yF_r400.bam
erage
2
[0 - 198]
accepted_hits_L7_yF_r400.bam
erage
3
[0 - 175]
accepted_hits_L8_yF_r400.bam
erage
4
[0 - 416]
RefSeq Genes
ITIH5

Fig.7. Similarities in expression pattern between samples coming from the same tissue – patient A (1 and 2), patient B (3 and 4).

6. Conclusions

RNA sequencing data has been investigated with a set of tools and software to proof its usefulness in the detection of molecular markers for thyroid cancer discovery. Genes differentiating between FTC in FA in microarrays has been confirmed to be differentially changed in RNA sequencing data. It has been shown that there are many approaches of data analysis, however a clear workflow has not been established so far. All the available software is still under development by the leading bioinformatics departments around the world. However, it is expected that standardized analysis algorithms will be clarified soon.

From the biological point of view sequencing data shows huge amount of information that has not been fully understood yet. On the other hand resolution and accuracy of RNA sequencing data is significantly better than data from microarray experiments. Additionally, there are many factors indicating that sequencing data may become a source of new prognostic markers in the discovery of cancer.

7. Bibliography and author

[1] Barbara Jarząb et al.: *Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications*, Cancer research 2005; 65(4):1587-97.

[2] He H, Jazdzewski K et al.: *The role of microRNA genes in papillary thyroid carcinoma.* Proc Natl Acad Sci U S A. 2005 Dec 27;102(52):19075-80. Epub 2005 Dec 19.

[3] Borup R et al.: *Molecular signatures of thyroid follicular neoplasia.* Endocr Relat Cancer. 2010 Jul 28;17(3):691-708. Print 2010 Sep.

[4] Andrews S, *FastQC: A quality control tool for high throughput sequence data.*

[5] http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

[6] Trapnell C, Pachter L, Salzberg SL. *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics doi:10.1093/bioinformatics/btp120

[7] Integrative Genome Browser (IGV v.2.0.4) http://www.broadinstitute.org/igv/home

[8] Trapnell C et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation,* Nature Biotechnology doi:10.1038/nbt.1621

Author:

MSc. Tomasz Stokowy

Silesian University of Technology
Automatic Control Department
ul. Akademicka 10
44-100 Gliwice
tel. (032) 237 28 99
email: *tomasz.stokowy@polsl.pl*