Variant Calling and Microarray Expression Analysis in Pancreatic Islet Samples: Master’s Thesis in Bioinformatics

Olof Asplund

June 2013 - January 2014

Contents

1 Introduction to Master's Thesis Project

Diabetes mellitus: definitions and pathology

There are between 314 and 382 million adults with *diabetes mellitus* world-wide.[1] This group of related conditions are characterized by defects in the production and/or response to insulin, leading to an increase in blood glucose levels(hyperglycemia).

Type 1 *diabetes mellitus*(T1DM) is characterized by the loss of beta cells in the pancreatic islets, which ceases the production of the hormone insulin. Type 2 *diabetes mellitus*(T2DM), in contrast, is characterized by a decreased response to insulin in skeletal muscle, adipose tissue and liver cells of the body.[2] Both lead to the decrease of glucose transporter protein 4(GLUT4) mediated transport of glucose into these cell types,[3] resulting in hyperglycemia.

Genetics of Diabetes

Both T1DM and T2DM have polygenetic hereditary components, with at least 20 genes contributing to T1DM and at least 36 contributing to T2DM susceptibility.[4,5] Especially for T2DM, the genetic mechanisms are poorly understood, with only 10 percent of the hereditary susceptibility being explainable by discovered gene variants.[6] Identifying and characterizing risk variants is thus important for drug development, diagnostics and treatment, and understanding of the pathogenesis of diabetes. Both T1DM and T2DM is characterized by endocrine dysregulation which affects the expression of many genes. As such, studying gene expression in diabetic and non-diabetic individuals and animal models also provides valuable insight into the mechanisms behind these highly prevalent diseases.

Lund University Diabetes Center

The Lund University Diabetes Center(LUDC), located within the Clinical Research Center(CRC) in Malmö University Hospital, performs extensive research to investigate genetic and environmental factors contributing to *diabetes mellitus* and its underlying factors. Several large-scale population studies of diabetic individuals from Finland, Scania in Sweden and other locations are performed, in conjunction with *in vitro* and animal studies.

A large computational facility, LUDC-calc, is available for researchers to enable high-throughput processing of sample data. Here, I have performed two different studies which involve two important fields in bioinformatics: genomic variant discovery from high-throughput sequencing data, and comparative gene

expression analysis. The studies provide insight into technological and statistical considerations for these types of analyses.

References

[1] G. Danaei, M. M. Finucane, Y. Lu, G. M. Singh, M. J. Cowan, C. J. Paciorek, J. K. Lin, F. Farzadfar, Y.-H. Khang, G. A. Stevens, M. Rao, M. K. Ali, L. M. Riley, C. A. Robinson, and M. Ezzati, "National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants," *The Lancet*, vol. 378, pp. 31–40, July 2011.

[2] K. Alberti and P. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. part 1: diagnosis and classification of diabetes mellitus. provisional report of a WHO consultation," *Diabetic Medicine*, vol. 15, no. 7, p. 539–553, 1998.

[3] R. Govers, A. C. F. Coster, and D. E. James, "Insulin increases cell surface GLUT4 levels by dose dependently discharging GLUT4 into a cell surface recycling pathway," *Molecular and Cellular Biology*, vol. 24, pp. 6456–6466, July 2004. PMID: 15226445 PMCID: PMC434240.

[4] M. A. Kelly, C. H. Mijovic, and A. H. Barnett, "Genetics of type 1 diabetes," *Best Practice & Research Clinical Endocrinology & Metabolism*, vol. 15, pp. 279–291, Sept. 2001.

[5] C. Herder and M. Roden, "Genetics of type 2 diabetes: pathophysiologic and clinical relevance," *European journal of clinical investigation*, vol. 41, pp. 679–692, June 2011. PMID: 21198561.

[6] S. H. Kwak and K. S. Park, "Genetics of type 2 diabetes and potential clinical implications," *Archives of pharmacal research*, vol. 36, pp. 167–177, Feb. 2013. PMID: 23377708.

2 Comparison of RNA-seq and Exome Sequencing Data for Variant Calling

Abstract

This article describes the creation of a pipeline for variant calling from high-throughput next-generation exome and RNA sequencing data using commonly used bioinformatics tools. High-throughput sequencing data from six pancreatic islet cell samples were analyzed using the pipeline, and the resulting variant calls were validated against chip genotyping data from the same individuals. The results indicate that variant calling can be applied to RNA-seq and exome sequencing data to identify genetic variants in exons and coding regions with high precision, while the recall was relatively low. In other words, identified genotypes seem to have a high probability of being correct, but only part of the present variants are picked up. This is especially true for RNA-seq.

Abbreviations

BAM: Binary Alignment Map; BASH: Bourne Again Shell, a command processor in Linux; BLAST: Basic Local Alignment Search Tool; DNA: deoxyribonucleic acid; FASTA: a commonly used sequence file format; FASTQ: an extension of FASTA with added sequence qualities; GATK: Genome Analysis Toolkit; HTS: High-Throughput Sequencing; RNA: ribonucleic acid; SAM: Sequence Alignment Map; SNP: Single Nucleotide Polymorphism; UCSC: University of California Santa Cruz; UTR: untranslated region; VCF: Variant Call Format.

3 Introduction

Next generation sequencing

The development of massively parallel, high-throughput sequencing(HTS) has revolutionized the field of molecular biology, enabling high-coverage sequencing of samples at the genome, exome or transcriptome level for a relatively low cost. In genomics, this is a powerful tool which can be used for characterizing cancer genomes, identifying neutral and disease-causing variants, and studying population-wide genetic variation.[7] In transcriptomics, HTS allows for high-precision differential gene expression analysis and the characterization of known or novel splicing forms.[8]

Massively parallel sequencing results in a large number of sequences, referred to as *reads*. Attached to each read are base qualities for each nucleotide in the read, showing the predicted error rate for each base. For paired-end sequencing, which is used in this project, the sequences come in pairs of two reads sequenced from opposite ends of a longer sequence of DNA, with a specific amount of unknown sequence between them. This gives positional information which can be used to detect small- and large-scale genomic insertions and deletions.[9]

The reads are typically organized in FASTQ files, a flat file format similar to the FASTA format, but also including base qualities. The sequences may originate from several different sources. Exome sequencing is performed on a library of source DNA from which known exon sequences are captured, using, for instance, a solid surface with attached exon-specific probes, and sequenced.[10] Additionally, some of the current library preparation kits also capture many 5'- and 3'-untranslated regions(UTRs). RNA sequencing, or RNA-seq, is typically performed by extracting the cellular RNA, removing rRNAs, potentially isolating the poly-A-tagged mRNA transcripts, and using reverse transcriptase to construct a DNA library, which is then sequenced.

Alignment

To analyze the output of next-generation sequencing, the reads have to be either assembled into longer contiguous sequences(*de novo* assembly) or aligned to a reference genome or transcriptome. The former is mainly used for characterization of species without a good reference genome, or for small genomes, such as those of bacteria. In humans, the latter is typically used, since the human genome is well-characterized.

The speed of each alignment needs to be very high to match the large number of reads generated by high-throughput sequencing. Frequently used general-purpose alignment algorithms, such as the Needleman-Wunsch algorithm and the BLAST algorithm, are too slow for this purpose. For the alignment of millions of reads per sample, this would take too long to be of practical use. Instead, specialized algorithms are used, which radically increase alignment speed. To achieve the increase in speed, a key method is indexing the reference sequence in a way which makes it possible to quickly match the reads against the reference. Other optimizations include usage of efficient low-level instructions, parallel computation across multiple processors and speed-efficient memory management.

Paired-end sequencing provides advantages in alignment. The known genomic distance between the pairs can be used to identify insertions and deletions relative to the reference sequence. If two ends map far away from each other on the reference genome, it can be inferred that a deletion has occurred between them. Conversely, an insertion has occurred if two reads in a pair map closely to each other.

For RNA sequencing, splicing hinders mapping reads directly to the genome. Due to each transcript potentially being derived from several different discontinuous fragments of a gene, mapping the reads straight to the genome can potentially introduce error by mapping reads to the wrong regions or discarding reads due to poor alignment. Thus, programs such as Tophat and STAR have been developed which take splicing into account. To do this, a *splice junction database* is used along with the reference genome; the junction database provides the programs with the positions of splice junctions across which splicing takes place. This additional information enables the aligners to detect intra- and intergenic splicing events which have taken place inside reads, and thus correctly mapping the reads to the reference.

Variant calling

Variant calling is the process of identifying genetic variation in sequencing data, such as single nucleotide variants, copy number variations, structural variants,

such as indels and inversions, and fusion genes. In this project, we only look at single nucleotide variants.

After the reads have been mapped to the reference genome, and the results have been processed to remove alignment errors, special programs, such as the UnifiedGenotyper in the Genome Analysis Toolkit from Broad Institute,[11] are used which iterate through the aligned reads, identifying genomic loci where reads are aligned with high confidence, but differ from the reference genome. This way, alternative alleles can be identified. By further inspecting the aligned reads at these positions, genotypes can be inferred. For instance, if 50 out of 100 reads have a different base at a certain locus, this might indicate that the individual in question is heterozygous at this locus. If all reads have a different base, the individual might be homozygous for the allele in question.

The set of SNPs that can be found for a given experiment depends on the source of DNA. For instance, exome sequencing only targets exons and the untranslated regions surrounding genes. RNA-seq targets untranslated regions, exons, and sporadically introns. It has been shown that, on average, 39 percent of the genome is expressed as primary transcripts,[12] which means that a substantial part of the genome is likely to be inaccessible by RNA-seq without pooling data from multiple tissues. In addition, sequencing *depth* or *coverage*, which is the number of reads overlapping a specific region of the reference, will vary for different genes based on gene expression levels. Inside genes, depth for each exon will vary depending on splicing patterns. A previous study of samples from the 1000 Genomes dataset by Quinn *et.al.* has shown that RNA-seq data can be used for variant calling with around 90 percent specificity and sensitivity as compared to whole genome sequencing data.[13] That study contained few samples, however, and only used RNA-seq data.

4 Materials and Methods

Materials

Six human pancreatic islet cell samples which had been sequenced using RNA sequencing and exome sequencing, and for which chip genotyping had been performed, were selected for analysis. The samples were taken post-mortem and stored frozen. The read libraries were 101 bp paired-end reads with a fragment length of 300 bp. For exome sequencing, the average number of reads per sample was 74,442,454 reads(± 5,988,360). The average per-sample sequencing depth across the targeted intervals was 48X(± 6X) pre-variant calling, for BWA alignments. For RNA-seq, the mean number of reads per sample was 65,000,909 reads(± 5,985,383).

Methods

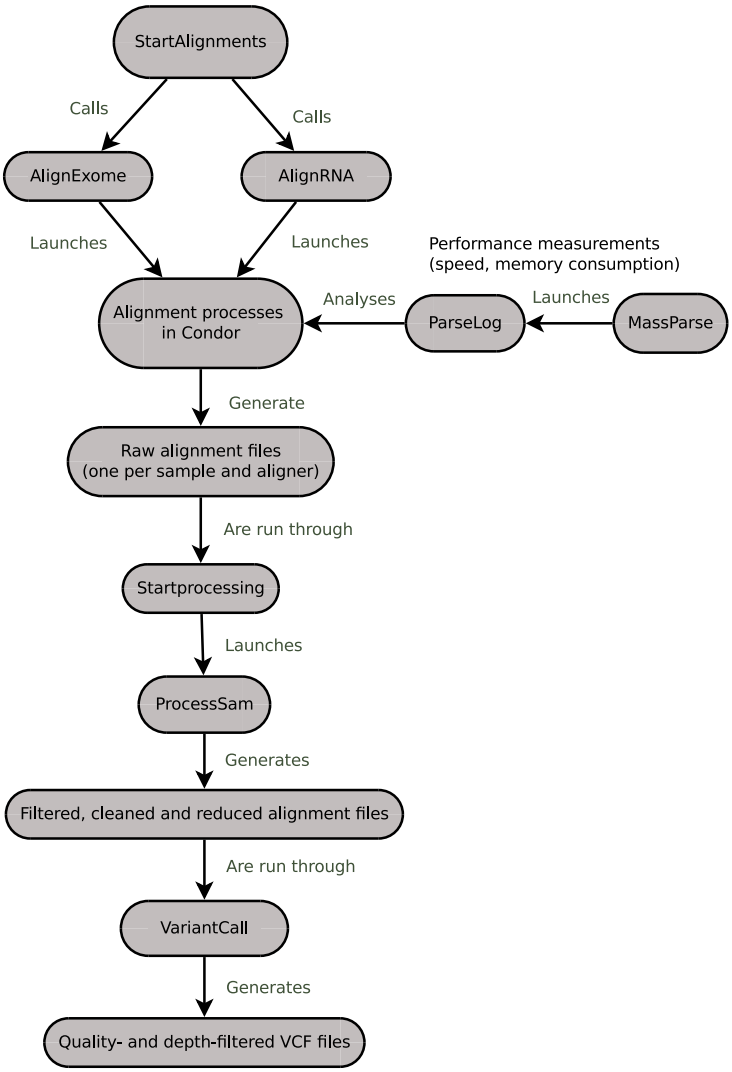

Figure 1: Outline of analysis workflow.

A pipeline for aligning and processing the raw FASTQ reads was constructed, using BASH shell scripts. The pipeline consists of several modular scripts which arrange the parallel deployment of calculation jobs, and ensure that the output files from each step are arranged in a highly organized fashion(figure 1). In each step, several samples are processed in parallel through Condor, the calculation job management system used on the calculation servers at Lund University Diabetes Centre(LUDC). When variants are called, all samples are processed together.

Alignment

For RNA-seq data, the raw reads were aligned with three different aligners: Tophat,[14]STAR[15] and GSNAP,[16] but due to technical difficulties, variant calls were only acquired from Tophat(described in the Results section). For exome data, the raw reads were aligned with four different aligners: GSNAP,Bowtie2,[17]BWA[18](using the MEM algorithm) and Novoalign.[19] All exome aligners were configured to only report one alignment per read, and where applicable, fragment and insert sizes were defined. All RNA aligners were configured to only report one alignment per read, to only search for known splice junctions, and to report only uniquely mapping reads, and were otherwise left on the default settings. Each alignment was run on 10 cores. The Condor log files were used to extract performance information, such as peak memory consumption and execution time. All aligners except BWA used hg19 as the reference sequence. For BWA, b37d5 was used as the reference sequence, since this index is commonly used in-house. For RNA-seq alignment, splice junctions were taken from RefGene from the UCSC Genome Table Browser.[20]

Processing

The alignment files were processed to improve variant calling and to correctly format the data. Reads with low($\leq$40) mapping quality were filtered out using the Picard Toolkit for all aligners except Tophat, which does not supply mapping qualities. Reads mapping to non-autosomal chromosomes were removed. This was done to exclude sex-specific SNPs and to remove the effect of varying numbers of X-chromosomes between samples. The alignment files were then coordinate-sorted, duplicate reads(reads mapping to the same genomic coordinates) were removed and the files were reordered in the correct order in respect to chromosomes. Local realignment around indels was performed to remove mismatches at the edges of indels, and base quality recalibration was performed. Read reduction(a process which removes extraneous information from alignment files before variant calling) was also performed.

Variant Calling

A list of SNPs which are unambiguously annotated as located in exons and 5'- and 3'-UTRs in HapMap was acquired through the R package *biomaRt*.[21] In order to give a fair comparison between exome sequencing and RNA-seq, only SNPs in this list were used for validation. The reduced aligner output was run through the UnifiedGenotyper. A variant call format(VCF) file with HapMap SNPs was used to search for SNPs. The resulting raw calls were filtered to remove SNPs with a read depth(DEPTH field in VCF file) below 10 and quality

score(QUAL field in VCF file) below 10.

Validation

To make sure that only the most reliable chip genotyping data was used for validation, the raw chip genotyping file was filtered in several steps using the PLINK software package[2223] and R.[24] First, the list of SNPs available on the chip was extracted. The intersecting set of SNPs between this list and the list of SNPs in gene regions mentioned above was calculated. Data from these SNPs were extracted from the HapMap VCF annotation file used for variant calling. The list of SNPs was then filtered to only retain SNPs with a single alternative allele, thus removing SNPs with multiple alleles. After that, the resulting list of SNPs was filtered to only retain SNPs which had the same alleles in the VCF file and the raw chip genotyping file. Chip calls from that list(table 1) was used to validate the variant calling results.

Coding		Untranslated regions		
Non-synonymous	Synonymous	3'-UTR	5'-UTR	Total
5,747(31.1%)	4,612(24.9%)	7,117(38.5%)	1,009(5.5%)	18,485

Table 1: Distribution of SNPs used for analysis.

To compare the variant calls produced by each aligner to the reference chip genotype calls, each VCF(variant call) file was loaded into R. Calls without a SNP ID were removed, as were all SNPs where any of the calls showed a third allele. SNP calls with a sample-wise depth below 10 were filtered out. For the exome sequencing data, calls with a genotype quality below 30 were also filtered out. SNPs which were in the list in table 1 were then validated against the chip genotyping data.

To measure the performance of each aligner, precision and recall was calculated for each genotype.

$$Precision = \frac{N_{true\ positives}}{N_{true\ positives} + N_{false\ positives}} \tag{1}$$

$$Recall = \frac{N_{true\ positives}}{N_{true\ positives} + N_{false\ negatives}} \tag{2}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

In this context, precision measures the probability of a call for a certain genotype being correct when compared to the chip. Recall, also known as sensitivity, shows the probability of a correct call being made for a genotype which is

present on the chip. The harmonic mean of precision and recall gives a combined measurement, known as the F_1-score, which can be used to quickly compare different genotypes and aligners.

False and true positives and negatives are defined in table 2. We here use two different versions of false negatives and call them "global" or "local" recall. Recall under the local definition is limited to the test space of calls which were made, while the global definition extends into all calls which are on the chip(in the list of considered SNPs and samples). As such, global recall is preferable. However, in order to study how recall changes depending on different factors related to the calls made, such as sequencing depth and quality, local recall has to be used.

Call type	Definition
True positive	Called,correct genotype
False positive	Called,incorrect genotype
False negative(global)	Not called or incorrect call for another genotype
False negative(local)	Incorrect call for another genotype

Table 2: How calls from each genotype were classified.

5 Results and Discussion

Aligner Performance

Exome

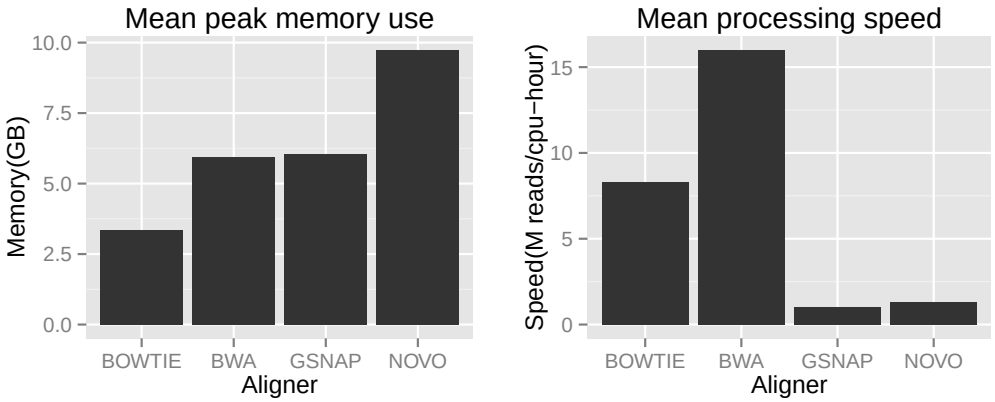

Figure 2: Peak memory usage and processing speed for exome alignment.

Memory usage and processing speed varies between programs, with Bowtie and BWA having both a relatively low memory footprint and high processing speed(Figure 2). BWA is around twice as fast as the second fastest aligner, Bowtie.

RNA-seq

For RNA sequencing, of the three aligners used, only Tophat produced output which worked for variant calling with GATK. Files produced by GSNAP with spliced alignment contained a large number of problems or discrepancies to what GATK accepts as input, resulting in GATK not accepting them as valid files. STAR alignment files also caused errors when run through the pipeline. For Tophat, 2 out of 6 samples failed to be processed by GATK, since a GATK tool which was part of the processing pipeline encountered a fatal error in these files and terminated.

Variant Calls

Exome

Table 3 shows some general statistics about exome sequencing calls. BWA appears to be the best aligner, producing more calls with higher F_1-scores than the other aligners.

	Aligner	Multiple alleles	Outside list	Filtered-out calls	Calls analysed	SNPs analysed
1	bowtie	33	28,251	2,460	70,276	12,063
2	bwa	34	28,965	1,892	71,493	12,217
3	gsnap	34	28,994	1,959	71,322	12,191
4	novo	34	28,569	1,882	70,351	12,025

Table 3: Results of filtering of exome SNP calls for each aligner. "Calls analyzed" represents the number of calls where data from both platforms were used after filtering. "SNPs analyzed" represents the number of SNPs which were used for analysis after filtering.

Aligner	Genotype	Precision	Recall	F1-Score
bowtie	0/0	0.982	0.454	0.621
	1/0	0.999	0.889	0.941
	1/1	0.979	0.837	0.903
bwa	0/0	0.982	0.461	0.628
	1/0	0.999	0.904	0.949
	1/1	0.979	0.851	0.911
gsnap	0/0	0.982	0.460	0.627
	1/0	0.999	0.902	0.948
	1/1	0.979	0.849	0.909
novo	0/0	0.982	0.453	0.620
	1/0	0.999	0.891	0.942
	1/1	0.980	0.839	0.904

Table 4: Genotype-wise precision and recall(global) for each exome aligner.

RNA-seq

Due to technical issues with Tophat processing, 2 out of 6 samples are excluded from the analysis of RNA-seq. 4 out of 6 samples aligned using Tophat are used to represent RNA-seq alignment and variant calling.

Aligner	Multiple alleles	Outside list	Filtered-out calls	Calls analyzed	SNPs analyzed
tophat	21	16,940	4,943	22,277	6,545

Table 5: Results of filtering of RNA-seq SNP calls for each aligner. "Calls analyzed" represents the number of calls where data from both platforms were used after filtering. "SNPs analyzed" represents the number of SNPs which were used for analysis after filtering.

Aligner	Genotype	Precision	Recall	F1-Score
tophat	0/0	0.983	0.183	0.308
	1/0	0.999	0.475	0.644
	1/1	0.977	0.431	0.598

Table 6: Genotype-wise precision and recall(global) for each RNA-seq aligner.

Comparison of RNA-seq and exome alignments

RNA sequencing generated fewer calls and found fewer SNPs than exome sequencing(table 5). In addition, more low-quality calls were made and filtered

out. Precision was similar for exome sequencing and RNA-seq(table 6), while recall was higher for exome sequencing. For both platforms, depth does not have a clear correlation with precision or (local) recall, but remains quite stable(see figure 5 and supplementary materials). Higher quality is correlated with higher accuracy, although a large proportion of the calls have the maximum quality value, 99, corresponding to an error p-value of $1.259 \cdot 10^{-10}$(data not shown).

Of all SNPs detected by BWA, representing exome sequencing, and Tophat, representing RNA-seq, 6,302 SNPs are common between the two, while 5,915 are uniquely found in exome sequencing and 243 are found uniquely in RNA-seq(figure 3). There are slight differences in the distributions of SNPs across different locations and variant types(figure 4), which may partly explain this.

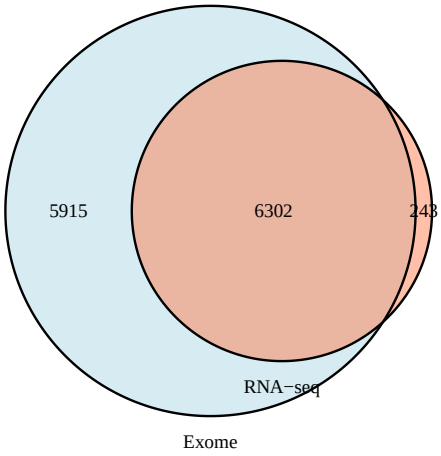

Figure 3: Overlap between SNPs detected by each platform. BWA represents exome sequencing, and Tophat represents RNA-seq.

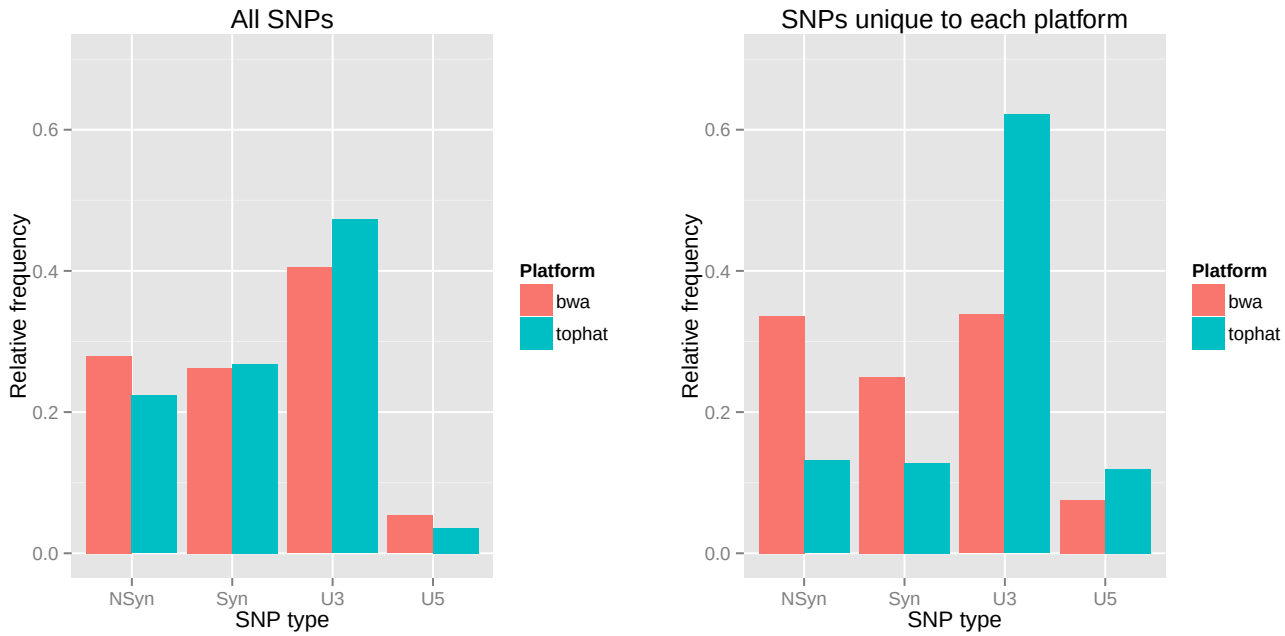

Figure 4: Frequencies of different types of SNPs called from exome(BWA) and RNA-seq(Tophat) data."NSyn" represents non-synonymous coding SNPs, "Syn" represents synonymous coding SNPs, and "U3" and "U5" represent 3'-UTR and 5'-UTRs, respectively. The height of the bars represents the relative frequency in each aligner's SNP set. The second graph shows the type distribution of SNPs which are only found by RNA-seq.

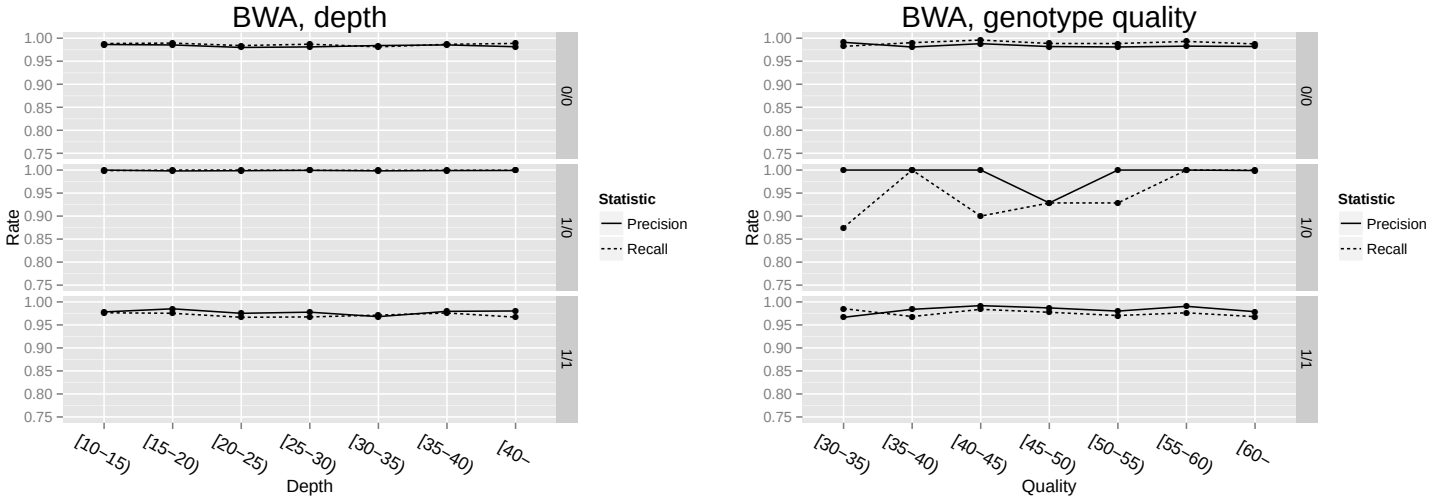

Figure 5: Operating characteristics for BWA.

The calls made from BWA, representing exome sequencing, and Tophat, representing RNA-seq, were compared to study differences in calls made. Out of the 71,493 exome sequencing calls and the 22,277 RNA-seq calls, 21,048 calls were common for both platforms. Of these, 99.87%(21,020 calls) predicted the same genotype. 98.94% of these were correct. The relation between RNA-seq and exome-seq calls is seen in table 7.

		RNA-seq		
		0/0	0/1	1/1
exome	0/0	7,413	0	0
	0/1	12	8,315	15
	1/1	0	1	5,292

Table 7: Counts of calls from RNA-seq and exome sequencing in respect to genotype.

Discussion

This project shows that variant calling can be performed with high precision from next-generation sequencing data. The precision is the highest for heterozygote calling. The Hardy-Weinberg equilibrium shows that, for biallelic loci with an allele frequency of 0.5, the proportion of genotypes are 25 percent for each homozygous genotype, and 50 percent for the heterozygote genotype.[25] Thus, a partial explanation is that calls for heterozygosity may be more likely to be correct due to chance alone for common alleles.

BWA seems to be the best aligner for exome sequencing data, with more calls made, higher accuracy and much higher speed. It should be noted that BWA was used with a slightly different reference genome, which could account for part of the difference. A likely explanation for the extreme differences in speed may be that different aligners were more or less able to utilize the multiple cores assigned to them. That would especially explain the big difference between BWA and Bowtie, and GSNAP and Novoalign. An important point to mention is that minimal adjustments to the configuration of each aligner were made. As such, it is not known to what degree the differences in accuracy and speed are the result of the underlying algorithms, or due to the default set of parameters used for alignments being more appropriate for this project. All aligners have many different options pertaining to, for instance, alignment scoring, read-reference mismatch tolerance and performance. It is likely that the differences between aligners shown here would increase or decrease if the parameters of each aligner were optimized.

Interestingly, calls for homozygosity for the alternative allele have both slightly lower precision and higher recall than for other genotypes. The lower precision might be due to mis-mapped reads which are misinterpreted as variants, insufficient depth at the variant site, or monoallelic expression. The low recall for the homozygous reference allele is likely the result of the variant calling program not calling homozygous reference alleles unless another sample has been called with any alternative allele for that particular SNP.

Apart from some technical difficulties, RNA-seq fared well in this comparison, generally performing as well as exome sequencing in regard to precision. As expected, the recall is lower, although the difference in sample size makes it difficult to know how big the real difference would be. Assuming that each sample contributes with approximately the same amount of calls, the number of expected calls for Tophat would be around $22,277 \cdot 1.5 = 33,415$ calls for all six samples, which is less than half of the number of calls from exome sequencing data.

The high agreement between platforms, as discussed in the previous section shows that the main difference between platforms is not the quality of generated SNP calls, but their localization. Specifically, around half of the SNPs only detected in only RNA-seq were from the 3'-UTR. The majority of the regions targeted by exome sequencing are exons, whereas for RNA-seq, since the cDNA library was constructed using poly-A-targeted enrichment, every transcript in the library will carry 5'- and 3'-UTRs. The human 3'-UTR is on average four times as long($\sim$ 800 bp) as the 5'-UTR($\sim$ 200 bp),[26] which might explain why a higher proportion of 3'-UTR SNPs are uniquely found by RNA-seq. For exome sequencing, the list of uniquely found SNPs are shown to be more uniformly distributed.

As a technical note, the analysis pipeline is shown to work well for exome sequencing data, but less well for RNA-seq data, due to alignment data from spliced alignment not being completely compatible with the Genome Analysis Toolkit. This has to be fixed for the pipeline to be viable for RNA-seq data. While feature-rich, well-documented, and built with parallelization in mind, GATK is very sensitive to the structure of its input, and usually cannot recover from problems.

In addition, processing the alignments is computation-intensive and takes a substantial amount of time(data not shown). Most of this time is taken up by indel realignment and base quality recalibration. Optimization of these steps would decrease the total processing time.

This project has laid the foundations of a pipeline for variant calling from exome and RNA-seq data. There is a number of possible extensions of the study:

- Correct the problems encountered with RNA sequencing. Most urgently,

the pipeline has to be adjusted so that all samples are properly processed. Additionally, being able to use STAR for alignment would substantially decrease alignment time.

- Optimize the sample processing step for speed. In its current state, the pipeline is very slow.
- Increase the sample size to increase statistical power.
- Attempt to increase the relatively low recall.
- Use the resulting data for -omics studies, for instance of allelic imbalance of expression and RNA editing.

6 Acknowledgements

I would like to thank my supervisors, Joao Fadista, Petter Storm and Leif Groop, for guidance and support with this project.

7 Supplementary material: detailed operating characteristics

This section shows the precision and recall for different aligners depending on depth and genotype quality. The corresponding graph for BWA is shown in Figure 5 .
Please note that in these graphs, recall is calculated differently. Only calls which were made are considered, which is the reason for the much higher recall values seen here compared to the rest of the report. As such, recall is defined for each genotype as

$$\frac{N_{Correct}}{N_{Correct} + N_{Calls\ for\ other\ genotypes}}. \tag{4}$$

Exome sequencing

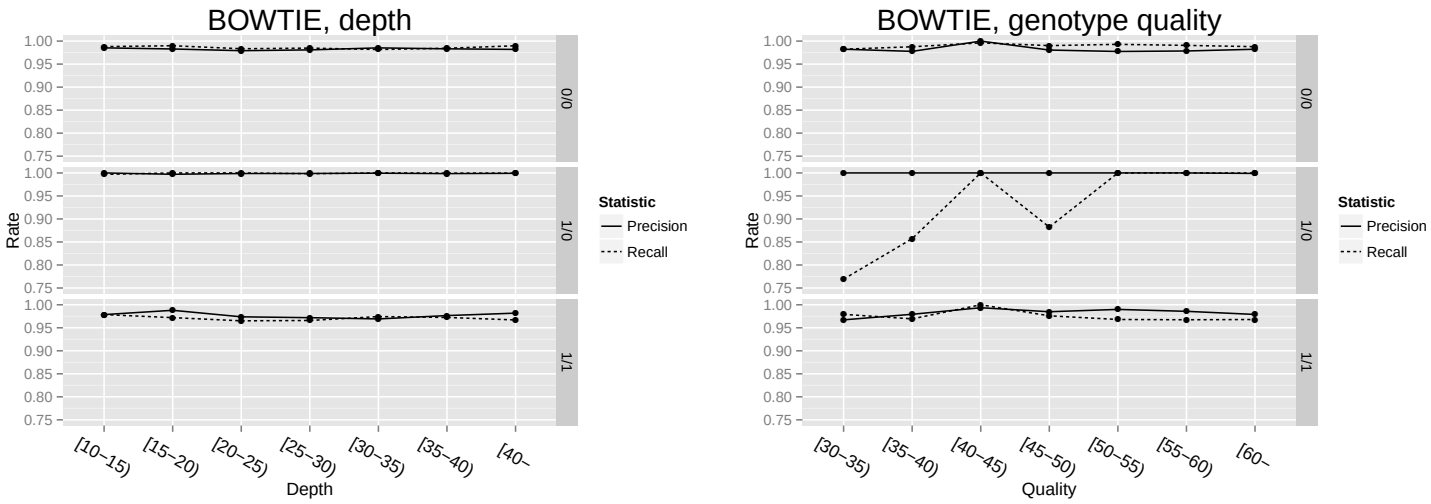

Figure 6: Operating characteristics for Bowtie.

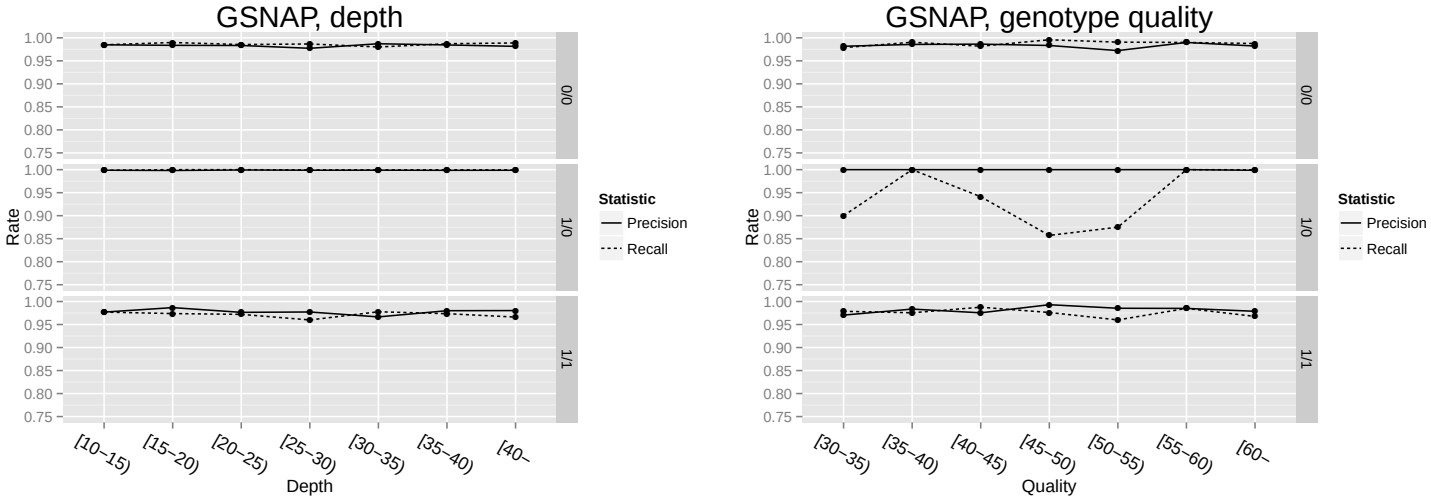

Figure 7: Operating characteristics for GSNAP.

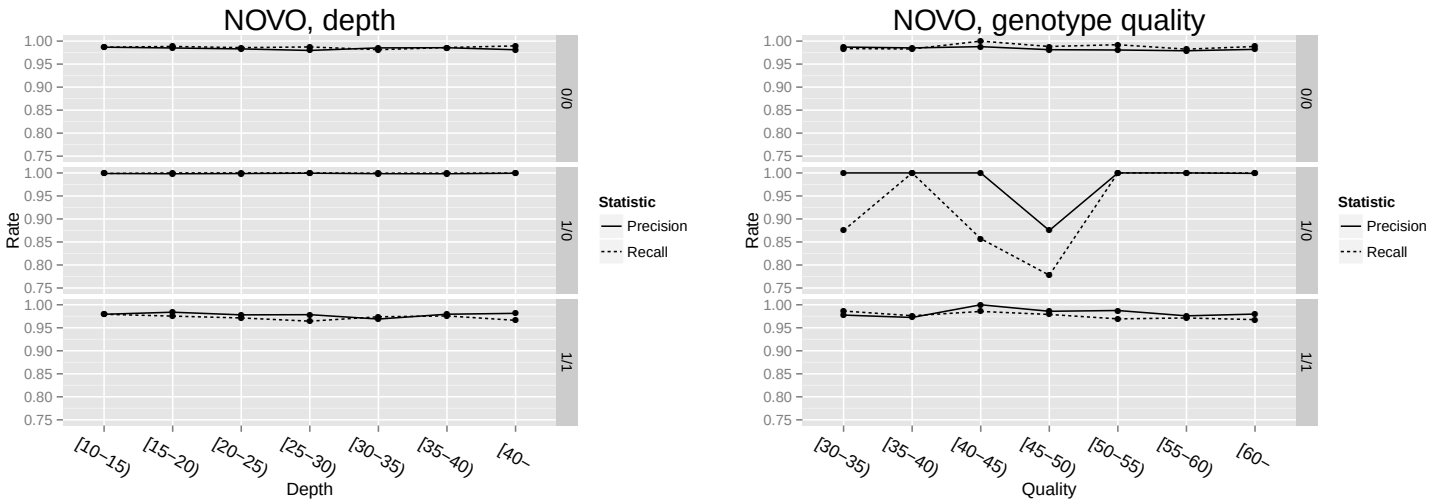

Figure 8: Operating characteristics for GSNAP.

RNA sequencing

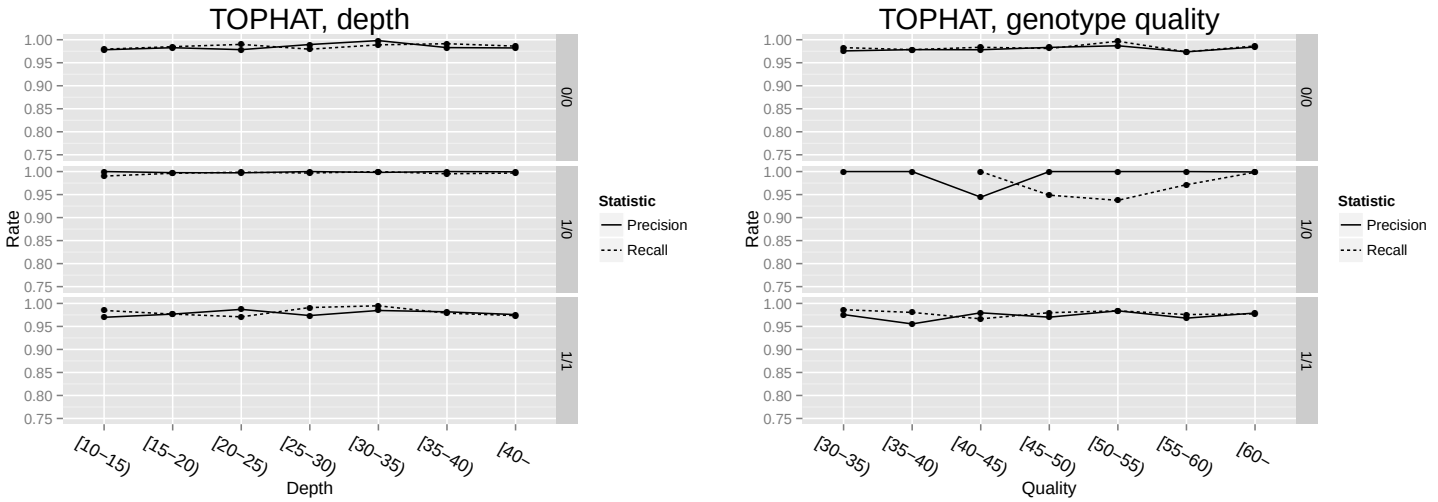

Figure 9: Operating characteristics for Tophat.

8 Supplementary material: Software commands

This section shows the command parameters used for key programs in the pipeline.

Alignment

SAMPLE1 and SAMPLE2 refer to the two files for each paired-end sample. samplename, lane, and adapter refer to different designations used for read-group assignments. GSNAP:

```
gsnap -A sam --pairmax-dna 300 --pairexpect 202 --pairdev 30 \
--read-group-id=${samplename}_${lane}_${adapter} --read-group-name=$samplename \
--read-group-platform=$PLATFORM_NAME -n 1 -B 5 -t $N_THREADS -D $GMAP_INDEX \
-d hg19 $SAMPLE1 $SAMPLE2
```

BWA:

```
bwa mem -M -t $N_THREADS \
-R @RG\tID:${samplename}_${lane}_${adapter}\tSM:$samplename\tPL:$PLATFORM_NAME \
$BWA_INDEX $SAMPLE1 $SAMPLE2
```

Bowtie:

```
bowtie2 -p $N_THREADS --end-to-end -I 270 -X 330 \
--rg-id ${samplename}_${lane}_${adapter} --rg SM:$samplename \
--rg PL:$PLATFORM_NAME -S ${samplename}_exome_bowtie.sam \
$BOWTIE_INDEX -1 $SAMPLE1 -2 $SAMPLE2
```

Novoalign:

```
novoalign -k -c $N_THREADS -o SAM \
@RG\tID:${samplename}_${lane}_${adapter}\tSM:$samplename\tPL:$PLATFORM_NAME \
-i PE 202,30 -r Random -d $NOVO_INDEX -f $SAMPLE1 $SAMPLE2
```

Tophat:

```
tophat --max-multihits 1 --num-threads $N_THREADS \
--rg-id ${samplename}_${lane}_${adapter} --rg-sample $samplename \
--rg-platform $PLATFORM_NAME --mate-inner-dist 98 --mate-std-dev 30 \
--transcriptome-index=$BOWTIE2_TRANSCRIPTOME --no-novel-juncs $BOWTIE_INDEX \
$SAMPLE1 $SAMPLE2
```

Post-processing

This portion performs quality-based filtering(MAPQ $\geq$ 40) and removal of non-autosomal reads. The awk script portion checks and only prints each record if the RNAME fields in each entry in the SAM/BAM file is of the form "(one or more digits)"(for BWA alignment files, which use a reference with chromosomes named 1, 2, etc.), or "chr(one or more digits)"(for all other aligners). The header lines at the start are always printed. The last portion reads the input stream of SAM data and encodes it as a BAM file. BWA alignment files:

```
samtools $viewcommand -q 40 $1 | \
awk '{if($0 !~ /^@/){if($3 ~ /^[[:digit:]]+$/){print $0}}else print $0}' | \
samtools view -Sb - > $PWD/$base.filt.bam
where $viewcommand is "view -h" for BAM files and "view -Sh" for SAM files,
and $1 is the alignment file in question.
```

Files from other aligners:

```
samtools $viewcommand -q $MIN_MAPQ $1 | \
awk '{if($0 !~ /^@/){if($3 ~ /^chr[[:digit:]]+$/){print $0}}else print $0}' | \
samtools  view -Sb - > $PWD/$base.filt.bam
where $viewcommand is "view -h" for BAM files and "view -Sh" for SAM files,
and $1 is the alignment file in question.
```

Subsequent filtering:

```
#Sort, index and convert SAM file to BAM
java -Xmx$JAVA_MEM -jar ~/Picard-Tools/SortSam.jar CREATE_INDEX=true \
INPUT=$PWD/$base.filt.bam OUTPUT=$PWD/$base.sorted.bam \
SORT_ORDER=coordinate

#Use MarkDuplicates to remove duplicate reads.
#Also writes removal info to $base.dedup.info and creates index for output.
java -Xmx$JAVA_MEM -jar ~/Picard-Tools/MarkDuplicates.jar \
CREATE_INDEX=true INPUT=$PWD/$base.sorted.bam \
OUTPUT=$PWD/$base.dedup.bam METRICS_FILE=$base.dedup.info \
REMOVE_DUPLICATES=true

#Reorder contigs to make sure they're in the correct order for GATK(chrM,chr1...)
java -Xmx$JAVA_MEM -jar ~/Picard-Tools/ReorderSam.jar \
CREATE_INDEX=true INPUT=$PWD/$base.dedup.bam \
OUTPUT=$PWD/$base.dedup.ordered.bam REFERENCE=$REFERENCE_FASTA

#Create indel realigner targets
java -Xmx$JAVA_MEM -jar ~/GATK/GenomeAnalysisTK.jar \
-nt $N_THREADS -T RealignerTargetCreator -R $REFERENCE_FASTA \
-I $PWD/$base.dedup.ordered.bam  -o $PWD/$base.dedup.ordered.intervals

#Realign indels
java -Xmx$JAVA_MEM -jar ~/GATK/GenomeAnalysisTK.jar \
-compress 0 -T IndelRealigner -R $REFERENCE_FASTA \
-I $PWD/$base.dedup.ordered.bam -targetIntervals \
$PWD/$base.dedup.ordered.intervals \
-o $PWD/$base.realigned.bam

#Base quality recalibration(calculation)
java -Xmx$JAVA_MEM -jar ~/GATK/GenomeAnalysisTK.jar \
-nct $N_THREADS -T BaseRecalibrator -I $PWD/$base.realigned.bam \
-R $REFERENCE_FASTA -knownSites $DBSNP_VCF -o $PWD/$base.realigned.table

#Base quality recalibration(applying changes)
java -Xmx$JAVA_MEM -jar ~/GATK/GenomeAnalysisTK.jar -nct $N_THREADS \
-T PrintReads -R $REFERENCE_FASTA -I $PWD/$base.realigned.bam \
-BQSR $PWD/$base.realigned.table -o $PWD/$base.recalibrated.bam

#reduce reads
```

```
java -Xmx$JAVA_MEM -jar ~/GATK/GenomeAnalysisTK.jar -T ReduceReads \
-R $REFERENCE_FASTA -I $PWD/$base.recalibrated.bam -o $PWD/$base.reduced.bam
```

Variant calling

```
samples=''
#Create a long string with the path of each reduced BAM file in the directory
for samfile in $(ls *.reduced.bam);do
echo "Found sample $samfile."
samples="$samples-I $PWD/$samfile "
done
#Use UnifiedGenotyper to call variants
#The data is downsampled to 1000 reads/position.
#Only SNPs in hapmap(3.3) are considered.
java -Xmx$JAVA_MEM -Djava.io.tmpdir=$OUT_DIR/tmp -jar $GATK_BIN \
-nt 2 -nct 5 --genotyping_mode GENOTYPE_GIVEN_ALLELES --alleles $VCF_HAP \
-T UnifiedGenotyper -R $REFERENCE_FASTA $samples --dbsnp $VCF_HAP \
-o $OUT_DIR/${aln}_${tech}_raw.vcf -dcov 1000
#Filter vcf file
#vcf-annotate does the filtering, while the awk script removes
#all variants which did not pass filtering.
cat $OUT_DIR/${aln}_${tech}_raw.vcf|vcf-annotate --filter MinDP=10/Qual=10 | \
awk '{if($0 ~ "^#"){print $0}else{if($7 ~ "PASS")print $0}}' > \
$OUT_DIR/${aln}_${tech}_filt.vcf
```

9 Software tools and versions

The following tools were used in the analysis:

- Picard-Tools[27] 1.58
- samtools[28] 0.1.19
- vcftools[29] 0.1.9
- Genome Analysis Toolkit(GATK)[11] 2.2-16
- PLINK[22, 23] 1.07
- Sequence aligners:
  - BWA[18] 0.7.5a
  - Bowtie[17] 2.0.6
  - Tophat[14] 2.0.7
  - GSNAP[16] 2013-06-27
  - STAR[15] 2.3.0e
  - Novoalign[19] 3.00.05
- R language environment[24] 2.15.1
- R packages:

- reshape2[30] 1.2.2
- ggplot2[41] 0.9.3.1
- VennDiagram[43] 1.6.5

References

[7] D. Koboldt, K. Steinberg, D. Larson, R. Wilson, and E. R. Mardis, "The next-generation sequencing revolution and its impact on genomics," *Cell*, vol. 155, pp. 27–38, Sept. 2013.

[8] K.-O. Mutz, A. Heilkenbrinker, M. Lönne, J.-G. Walter, and F. Stahl, "Transcriptome analysis using next-generation sequencing," *Current Opinion in Biotechnology*, vol. 24, pp. 22–30, Feb. 2013.

[9] J. O. Korbel, A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, B. E. Taillon, Z. Chen, A. Tanzer, A. C. E. Saunders, J. Chi, F. Yang, N. P. Carter, M. E. Hurles, S. M. Weissman, T. T. Harkins, M. B. Gerstein, M. Egholm, and M. Snyder, "Paired-end mapping reveals extensive structural variation in the human genome," *Science (New York, N.Y.)*, vol. 318, pp. 420–426, Oct. 2007. PMID: 17901297 PMCID: PMC2674581.

[10] J. K. Teer and J. C. Mullikin, "Exome sequencing: the sweet spot before whole genomes," *Human Molecular Genetics*, vol. 19, pp. R145–R151, Oct. 2010. PMID: 20705737 PMCID: PMC2953745.

[11] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature genetics*, vol. 43, pp. 491–498, May 2011. PMID: 21478889.

[12] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L.-H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigó, and T. R. Gingeras, "Landscape of transcription in human cells," *Nature*, vol. 489, pp. 101–108, Sept. 2012.

[13] E. M. Quinn, P. Cormican, E. M. Kenny, M. Hill, R. Anney, M. Gill, A. P. Corvin, and D. W. Morris, "Development of strategies for SNP detection in RNA-Seq data: Application to lymphoblastoid cell lines and evaluation using 1000 genomes data," *PLoS ONE*, vol. 8, Mar. 2013. PMID: 23555596 PMCID: PMC3608647.

[14] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, pp. 1105–1111, Mar. 2009.

[15] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics (Oxford, England)*, vol. 29, pp. 15–21, Jan. 2013. PMID: 23104886.

[16] T. D. Wu and S. Nacu, "Fast and SNP-tolerant detection of complex variants and splicing in short reads," *Bioinformatics (Oxford, England)*, vol. 26, pp. 873–881, Apr. 2010. PMID: 20147302.

[17] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature Methods*, vol. 9, pp. 357–359, Apr. 2012.

[18] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows-wheeler transform," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1754–1760, July 2009. PMID: 19451168.

[19] *Novoalign, by Novocraft(http://www.novocraft.com).*

[20] *UCSC Genome Table Browser(http://genome.ucsc.edu/cgi-bin/hgTables).*

[21] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt," *Nature protocols*, vol. 4, no. 8, pp. 1184–1191, 2009. PMID: 19617889.

[22] S. Purcell, *PLINK version 1.07(http://pngu.mgh.harvard.edu/purcell/plink/).*

[23] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *American journal of human genetics*, vol. 81, pp. 559–575, Sept. 2007. PMID: 17701901 PMCID: PMC1950838.

[24] R. Team, "R: A language and environment for statistical computing(http://www.R-project.org)," 2004.

[25] G. H. Hardy, "MENDELIAN PROPORTIONS IN a MIXED POPULATION," *Science (New York, N.Y.)*, vol. 28, pp. 49–50, July 1908. PMID: 17779291.

[26] F. Mignone and G. Pesole, "mRNA untranslated regions (UTRs)," in *eLS*, John Wiley & Sons, Ltd, 2001.

[27] *Picard-Tools software project(http://picard.sourceforge.net).*

[28] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, pp. 2078–2079, Aug. 2009. PMID: 19505943 PMCID: PMC2723002.

[29] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, pp. 2156–2158, Aug. 2011. PMID: 21653522.

[30] H. Wickham, "Reshaping data with the reshape package," *Journal of Statistical Software*, vol. 21, no. 12, pp. 1–20, 2007.

[31] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[32] H. Chen, *VennDiagram: Generate high-resolution Venn and Euler plots*, 2013. R package version 1.6.5.

10 Cross-platform comparison of array gene expression data

Abstract

The Affymetrix GeneChip microarray and the Illumina Human BeadChip array have big differences on both the technological and analytical level. The aim of this project was to investigate the level of correspondence between these two platforms. Expression data from three different sources, Affymetrix, Illumina and RNA sequencing, were processed and compared using expression values, linear models and gene set enrichment analysis for 12 islet cell samples. While the two array platforms show big differences at the raw intensity level, it is shown that the differences between the platforms decrease with higher levels of analysis. This gives support for the possibility of combining microarray results across platforms.

Abbreviations

BMI: body mass index; DNA: deoxyribonucleic acid; FFPE: Formalin-fixed paraffin-embedded; GSEA: gene set enrichment analysis; MDS: multidimensional scaling; PAGE: parametric analysis of gene set enrichment; RMA: robust multiarray average; RNA: ribonucleic acid; SNP: single nucleotide polymorphism; VST: variance-stabilizing transformation.

11 Introduction

Microarrays

DNA microarrays are an established technology for genetic analysis of biological samples with many uses, including single nucleotide polymorphism(SNP) analysis, differential gene expression analysis, and transcript splicing studies, among many other(see Plomin et al.[33] for a brief review of the technology). Because of the low cost, the availability of well-established analysis tools and the extensive body of knowledge from many years of use as a primary tool in labs, the microarray is a useful technology in research.

The typical microarray, exemplified by the commonly used GeneChip arrays manufactured by Affymetrix, consists of a flat surface on which oligonucleotide probes specific to a genomic or transcriptomic target sequence are immobilized in a specific pattern. Complementary DNA(cDNA) fragments tagged with fluorescent molecules are applied to the microarray and bind to their antisense target probes. The array is then scanned, and the fluorescence measured at each spot reflects the amount of target sequence in the sample. Probes can be used to target different features, most prominently SNPs as used in chip genotyping, and transcripts as used in gene expression profiling.

The raw scans have to be processed in order to make between-array comparisons of expression levels possible and to inspect the quality of each array. A commonly used method for between-array normalization of Affymetrix data is the robust multi-array average(RMA)[34] method. This includes background correction, $\log_2$-transformation and quantile normalization. The goal is to normalize the mean expression values and variances to put all intensities on the same scale.

The BeadArray technology, manufactured by Illumina, in contrast, puts the probes on microscopic beads, which are then put in wells arranged in a hexagonal grid. The arrangement of beads and the number of replicate beads per probe is random and varies between arrays. As with standard high-density DNA microarrays, the resulting raw data consists of measures of signal strengths mapped

to specific probe sequences, but the different design affects how the data is analyzed.

In order for data to be used across different platforms, it is important to know the level of correspondence between results from different platforms. A previous study by Zhang et al. performed on a mitochondrial disease model in mice showed a relatively good agreement between Affymetrix and Illumina expression arrays,[37] although the correlation between platforms on the raw probe intensity level was non-linear.

RNA-seq

Massively parallel high-throughput sequencing, commonly known as next-generation sequencing, is a very powerful tool for genetic analysis. In transcriptome sequencing[38](also known as RNA-seq), all RNA or a subset of RNA types from a sample are sequenced. This allows for the study of transcript isoforms resulting from alternative splicing, fusion transcripts resulting from trans-splicing(between genes) events, and gene expression profiling with a higher dynamic range and sensitivity than expression microarrays. This method is more expensive than microarrays, however, and is more complicated and computationally intensive to analyze.

Goals of this project

Twelve samples for which Affymetrix, Illumina and RNA-seq data was available were analyzed. Differences between Affymetrix and Illumina were identified by examining the raw signal intensities, through differential expression analysis by comparing different patient groups, and through gene set enrichment analysis.

12 Materials and Methods

Materials

Islet cell samples from twelve individuals were profiled using Affymetrix HuGene ST microarray GeneChips, one 12-sample HumanHT-12 bead-array BeadChip, and transcriptome sequencing. The sample group included both sexes, all patients were non-diabetic, and had a mean BMI of 26.8(S.D. 3.5). The mean age(measured as the date difference between the birth date and the date of isolation) was 51.8(S.D. 10.9). For Affymetrix arrays, raw CEL-files were used as the starting point of analysis. For Illumina, non-normalized, non-background-corrected probe-level summary data, produced using GenomeStudio 2011.1, was

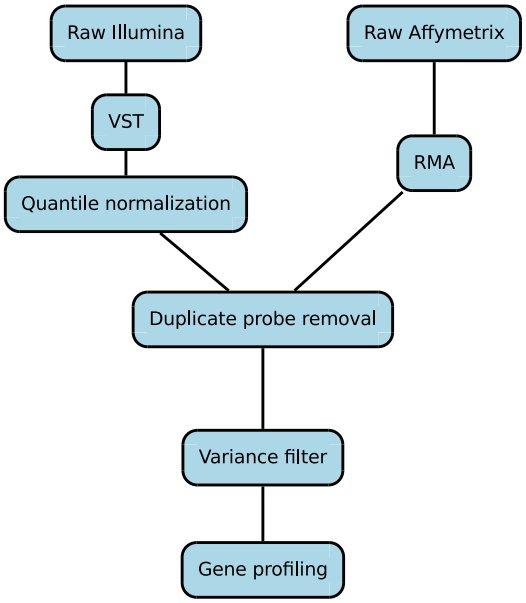

Figure 10: The preprocessing steps performed on the raw microarray data.

used. For RNA-seq data, gene-level, length-normalized counts produced using Tophat and Cufflinks transcriptome assembly were used.

Methods

All calculations were performed using the statistics software *R*[39] and the bioinformatics framework *Bioconductor*.[40] Plots were generated using the packages *ggplot2*,[41] *reshape2*,[42] and *VennDiagram*.[43]

Microarray Preprocessing

The raw Affymetrix data was RMA-normalized using the *affy*[44] package, while the raw Illumina data was transformed using the variance-stabilizing transformation method(VST)[45] from the *lumi*[46] package, followed by quantile normalization, as recommended by Ritchie et al.[47] Affymetrix and Illumina data was then filtered in two steps. First, probes without Entrez ID annotation were removed, and all probes mapping to the same Entrez ID's were reduced to the probes with the highest variance. The resulting set of probe intensities were used for expression level comparisons between platforms. Affymetrix data was annotated using the annotation databases *hugene10stcdf*[48] and *hugene10sttranscriptcluster.db*.[49] Illumina data was annotated using the annotation database *lumiHumanAll.db*.[50] In a second filtering step, probes were variance-filtered, only retaining probes

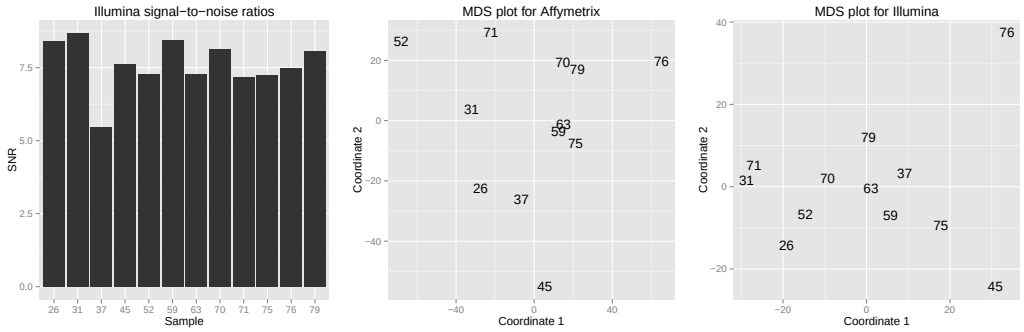

Figure 11: The per-section signal to noise ratio was calculated from the metrics file produced from raw data. In addition, multi-dimensional scaling(MDS) plots of $\log_2$-transformed data from each platform were used to identify potential clustering factors between samples.

with variance above the median. The motivation for this is that low variability across samples indicates non-expressed probes. The resulting data was used for gene expression profiling(table 8). Quality assessment plots were made to compare the signal intensities before and after filtering(figure 12).

Platform	Before	Duplicates	No annotation	Low variance	Left
Affymetrix	32321	1364	11072	9943	9942
Illumina	47230	10406	15998	10413	10413

Table 8: Results of nonspecific probe filtering.

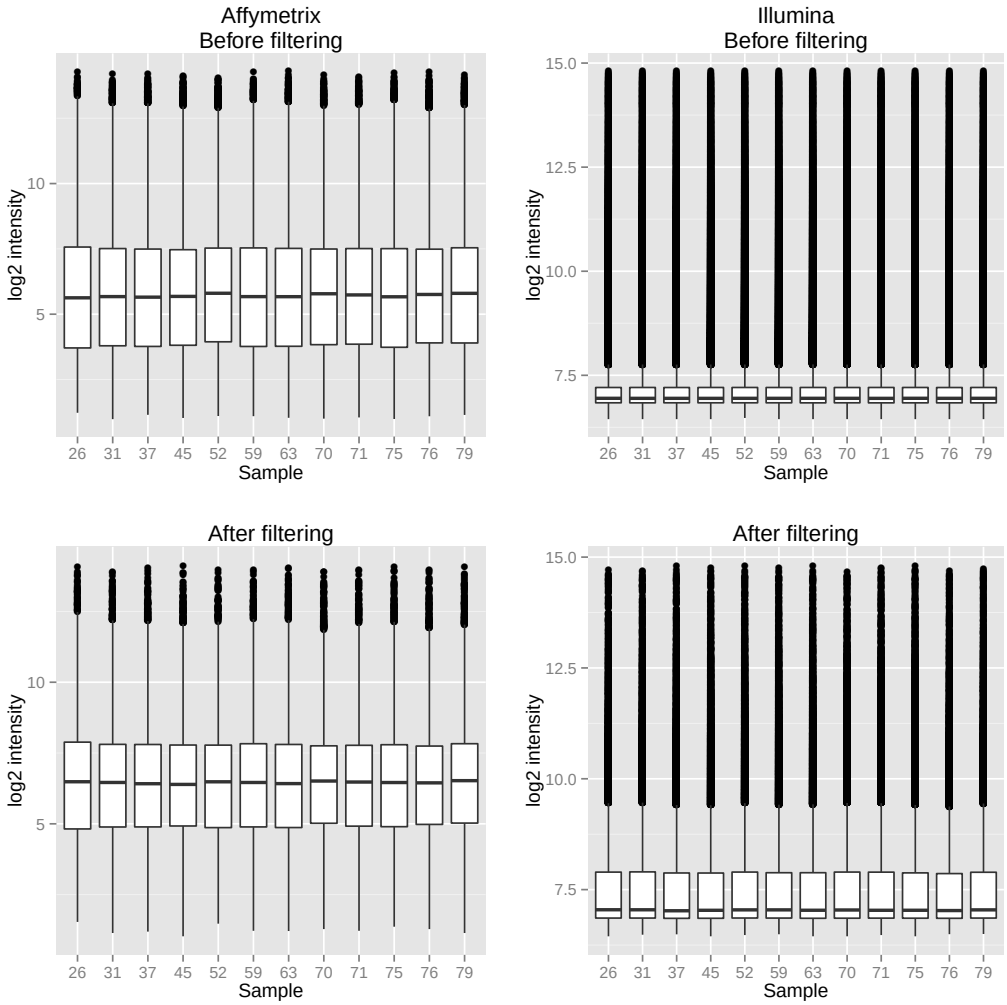

Figure 12: Boxplots of $\log_2$-transformed data before and after filtering.

Cross-platform expression level correlation

As a measurement of within-platform correlation of $\log_2$-transformed data, the Spearman correlations between each sample and all other samples were calculated and visualized using boxplots(figure 13).

To measure between-platform correlations, the Spearman correlations between Affymetrix versus Illumina, Affymetrix versus RNA-seq and Illumina versus RNA-seq per sample for all common genes were calculated and visualized with scatterplots(see figure 14 and appendix).

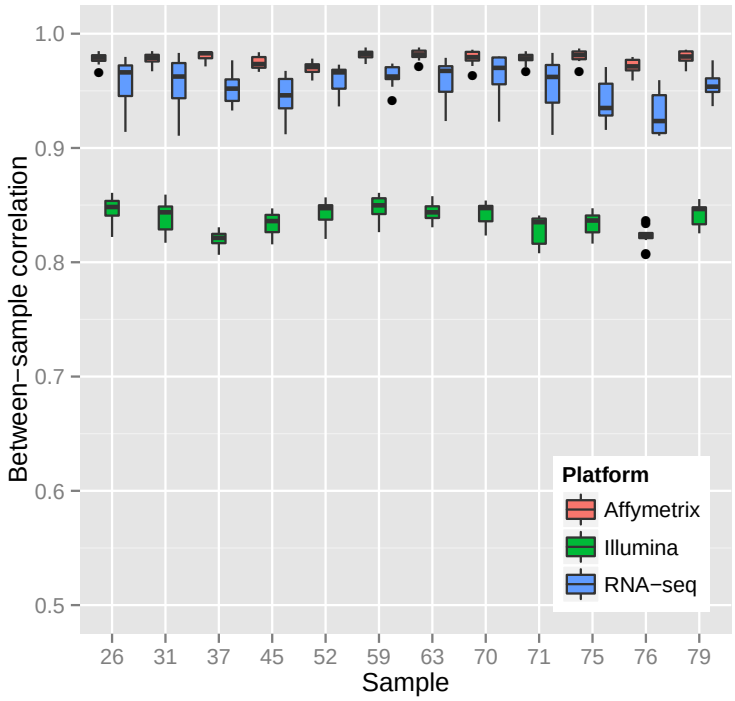

Figure 13: Within-platform intensity correlations for the three platforms. Each sample was compared to every other sample, resulting in 11 data points per sample.

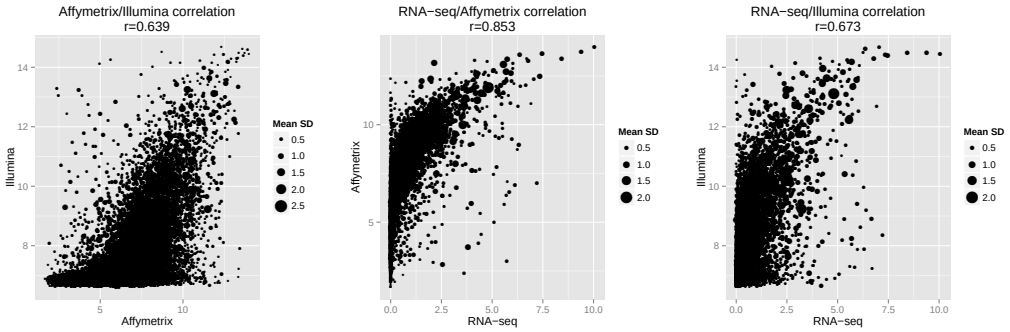

Figure 14: The mean $\log_2$-scale intensities per Entrez gene ID were calculated for each platform. In the above scatterplot, each dot represents one Entrez ID. The size of the dot corresponds to the mean of the standard deviations for both platforms for the Entrez ID in question. Scatterplots for each individual sample can be found in the appendix of this paper.

Differential expression analysis

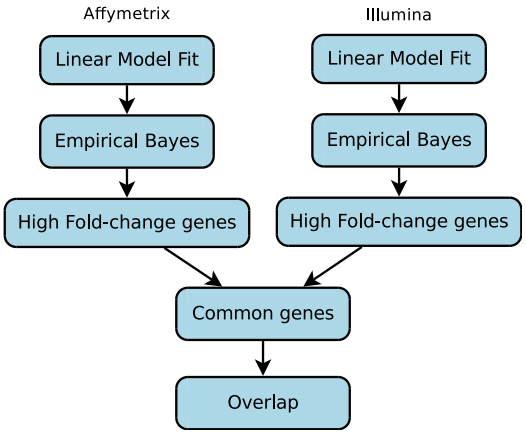

Figure 15: Differential expression workflow. Each array platform was analyzed using the same statistical model, and the correlation and overlap between array platforms was compared.

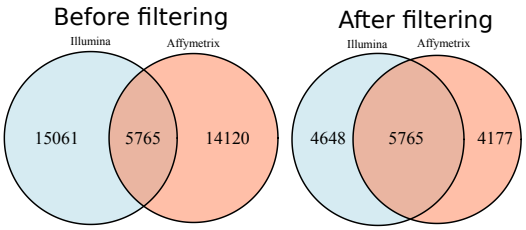

Figure 16: Overlap between Entrez IDs targeted by each platform before and after variance filtering.

As preparation for analysis of differential expression, the overlap between the Entrez IDs targeted by each raw and processed dataset was calculated,resulting in a list of 5,574 Entrez IDs targeted by both platforms after processing(figure 16).

For the purpose of evaluating the correlation between and inside platforms, 3 unpaired, two-class comparisons were made using the R package *limma*:[51]

- males and females(8 versus 4 samples),
- samples above and below the median BMI(26.25) in the dataset(6 versus 6 samples), and
- age ≥ 60 and < 60(8 versus 4 samples).

The different groups were chosen with the aim of giving measurable differences in gene expression: for the gender comparison, mainly Y-linked genes; for BMI, genes affected by body mass and food intake,such as genes associated with metabolism; and for age, potential age-related differences in gene expression. In addition, between-platform fold-change correlation plots were constructed from all overlapping genes.

RNA-seq analysis

The duplicate-filtered array data was remapped to gene symbols in order to enable comparison to the RNA-seq counts, which were mapped to gene symbols rather than Entrez gene identifiers. MDS plots were made from the normalized RNA-seq counts using the plotMDS function in edgeR.[52] Between-platform correlations were calculated between $\log_2$-scale counts from each sequencing platform and the $\log_2$-scale RNA-seq counts, with a pseudocount of 1 added to all counts to set values on the same scale as the array data.

For each group comparison, RNA-seq counts were variance-normalized using the *voom*[53] function in limma. Linear model fits were then performed in the same way as for microarray data. Between-platform $\log_2$ fold-change correlation plots of the gene symbols common between each platform and RNA-seq were created for each group design.

Gene set enrichment analysis

Gene set enrichment was analysed for the BMI and age comparison using parametric analysis of gene set enrichment(PAGE) from the *PGSEA*[54] package. The linear fits for both array platform created earlier were used to perform gene set enrichment analysis on the age and BMI comparisons. In the PGSEA analysis, high-significance($p \leq 0.01$) enriched gene sets in the high-age and high-BMI

groups were calculated for each platform. The Spearman correlation coefficients for pairs of Z-scores common between platforms were calculated, and the overlap of the sets of enriched gene sets between platforms were calculated.

13 Results and Discussion

Affymetrix versus Illumina comparison

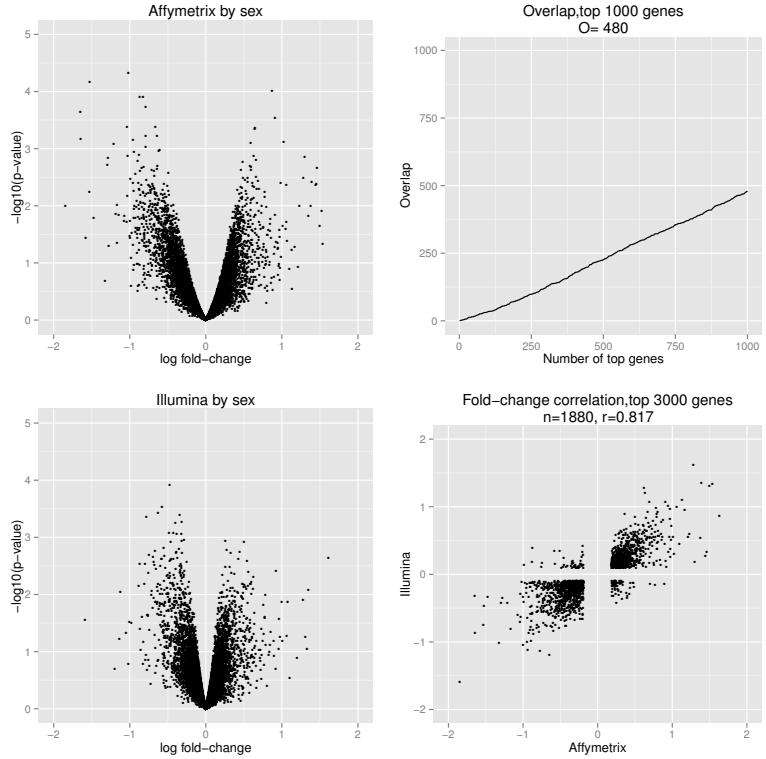

Figure 17: Results of sex comparison. Left side: Volcano plots for each platform. Right side, top: Number of overlapping genes in the highest fold-change list of genes from each platform. Right side, bottom: Fold-change correlation of the 1,880 genes common between the lists of the 3,000 highest fold-change genes of each platform.

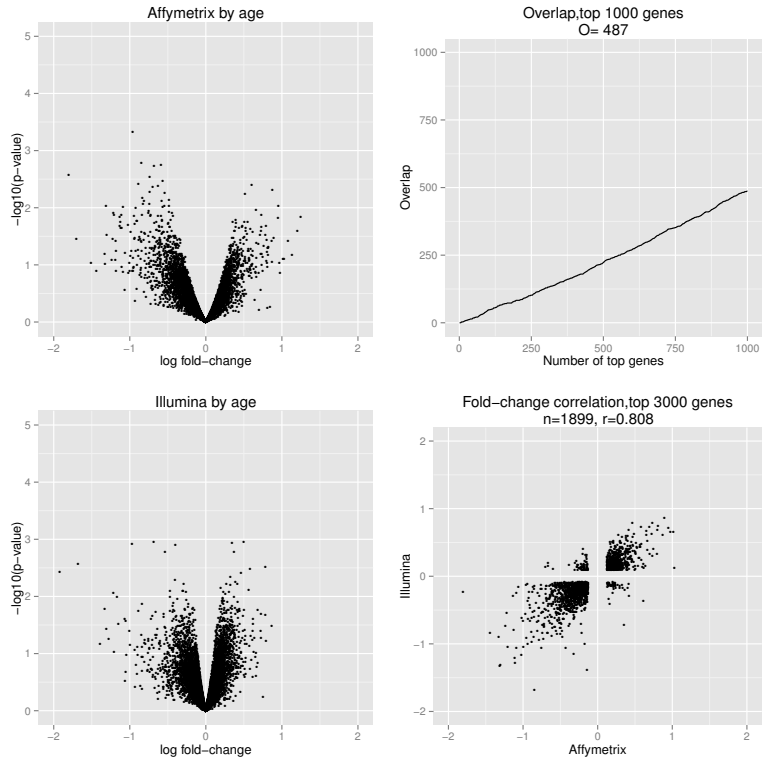

Figure 18: Results of age comparison. Left side: Volcano plots for each platform. Right side, top: Number of overlapping genes in the highest fold-change list of genes from each platform. Right side, bottom: Fold-change correlation of the 1,899 genes common between the lists of the 3,000 highest fold-change genes of each platform.

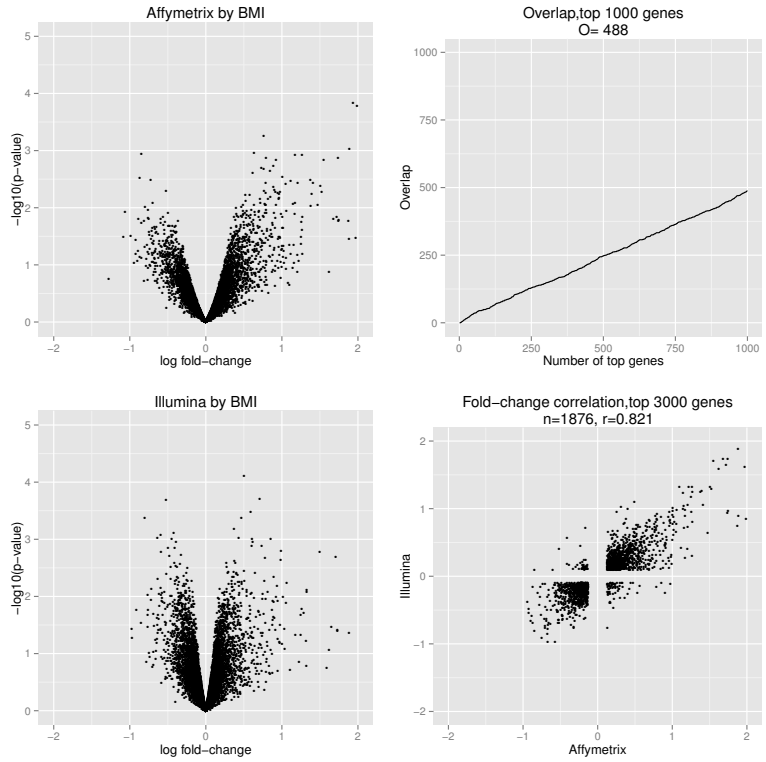

Figure 19: Results of BMI comparison. Left side: Volcano plots for each platform. Right side, top: Number of overlapping genes in the highest fold-change list of genes from each platform. Right side, bottom: Fold-change correlation of the 1,876 genes common between the lists of the 3,000 highest fold-change genes of each platform.

Between-platform correlations show a strong skewness in the distribution of Illumina $\log_2$-scale values towards the lower end, with a mean Affymetrix-Illumina per-probe correlation of 0.639. The skewness seen is similar to results seen in a study by Zhang et al.[37] Experimental error may have an additional negative influence on correlation, which is supported by the mean signal-to-noise ratio being below 10 for the array, whereas a ratio of 10 is recommended as a threshold for what should be considered to be a good-quality array.[55]

All three comparisons show an overlap for the top 1000 highest fold-change genes of slightly below 50%, and with fold-change correlations between the platforms for high-ranking genes ranging from 0.808 to 0.821. In other words, a gene with a high fold-change on one platform is likely to show a high fold-change on another, while the internal ranking of even the highest fold-change genes differs between platforms, and some genes will be shown as highly differentially expressed on only one of the two platforms. This is in agreement with the findings of a study by Cheadle et al., which found a low overlap between high-scoring genes between platforms, while results from gene set enrichment analyses such as GSEA and PAGE are largely consistent between platforms.[56]

There are important limitations in regard to what conclusions can be drawn from this comparison:

- The sample size is small, resulting in a lack of statistical power to detect differential expression, and overall high adjusted and unadjusted p-values.
- The group designs used may be limited in regard to truly differentially expressed genes, which would introduce noise in the top-ranking tables. This seems especially to be the case for the sex comparison.
- The annotation available and the imposed limitation to genes detectable on both platforms causes a loss of information, since genes unique to each platform and poorly annotated genes cannot be fully compared. It is also likely that some probes are mismapped.
- Quality assessment suggests that the array had a relatively low signal-to-noise ratio.

Several studies directly map the probe sequences to current versions of the genome, which might improve the stability of results across platforms by removing the effect of incorrect annotations.

RNA-seq analysis

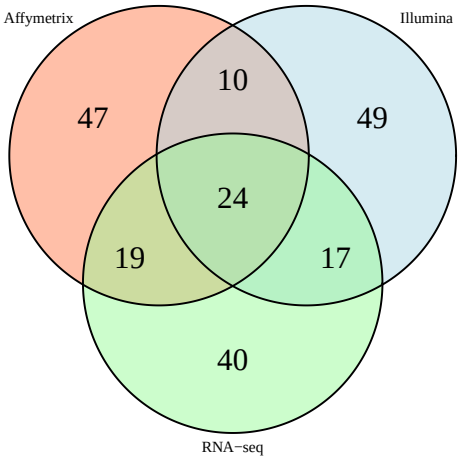

Figure 20: Overlap between gene symbols in the top 100 highest average-expression genes for all platforms.

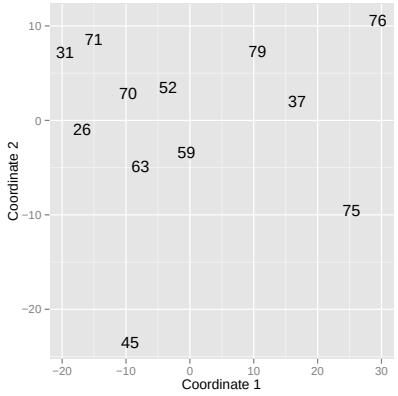

Figure 21: MDS plot for RNA-seq.

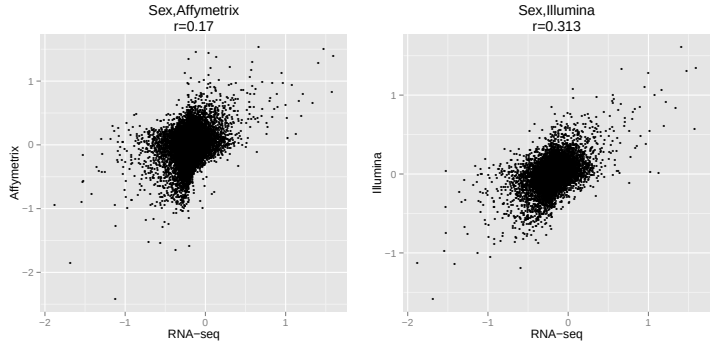

Figure 22: RNA-seq/array fold-change correlations for sex comparison.

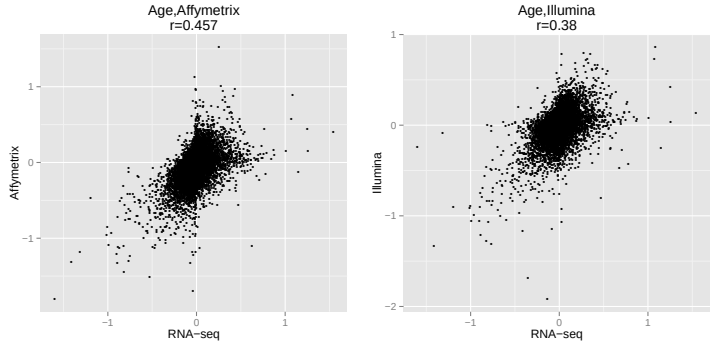

Figure 23: RNA-seq/array fold-change correlations for age comparison.

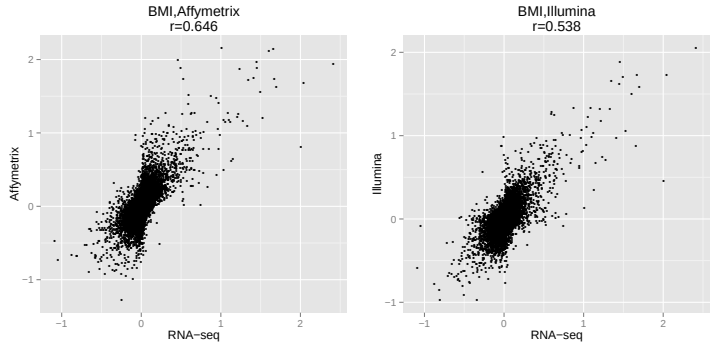

Figure 24: RNA-seq/array fold-change correlations for BMI comparison.

The comparisons between RNA-seq and microarrays show bigger differences than those which are found when comparing microarrays(figure 22 to 24). Between-platform correlations versus RNA-seq for Affymetrix vary from 0.273

to 0.642, while correlations for Illumina vary from 0.329 to 0.529. Both platforms show the highest correlation to RNA-seq for the BMI comparison.

Gene set enrichment analysis

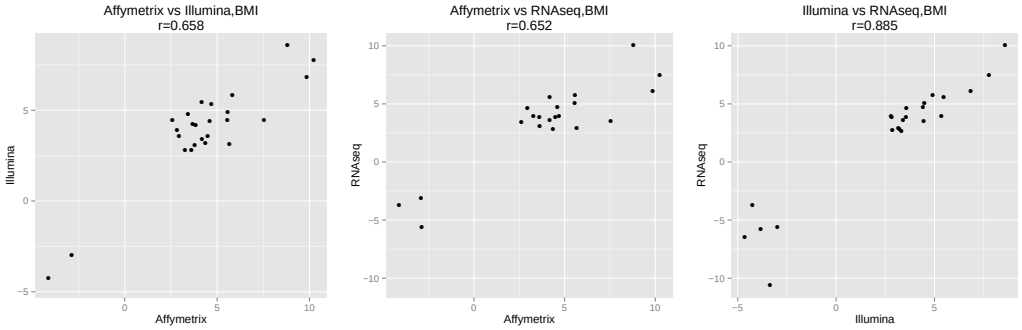

Figure 25: Enrichment score correlation for BMI.

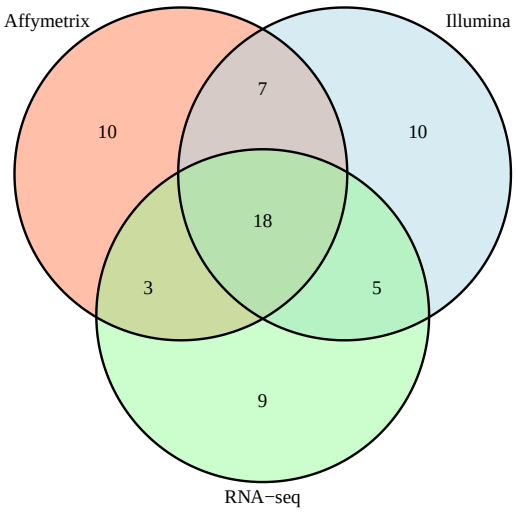

Figure 26: Overlap between enriched gene sets, BMI.

Gene set	Affymetrix	Illumina	RNA-seq
ribosome		-3.363	-10.623
oxidative phosphorylation		-4.651	-6.447
parkinsons disease		-3.830	-5.768
spliceosome	-2.877	-2.988	-5.577
huntingtons disease			-4.293
proteasome			-4.127
purine metabolism	-4.127	-4.252	-3.723
aminoacyl trna biosynthesis	-2.927		-3.130
alzheimers disease			-2.996
rna degradation			-2.725

Table 9: Top ten downregulated gene sets according to RNA-seq data, in BMI comparison.

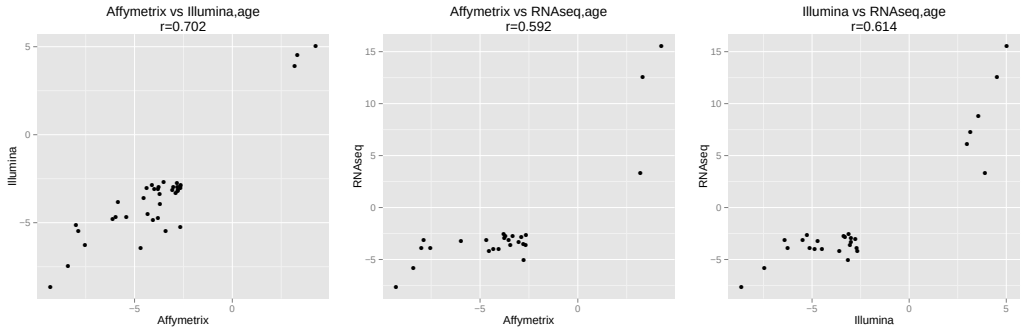

Figure 27: Enrichment score correlation for age.

Gene set	Affymetrix	Illumina	RNA-seq
complement and coagulation cascades	8.798	8.589	10.065
metabolism of xenobiotics by cytochrome p450	10.239	7.778	7.506
drug metabolism cytochrome p450	9.854	6.835	6.073
cytokine cytokine receptor interaction	5.577	4.909	5.784
chemokine signaling pathway	4.172	5.474	5.576
nod like receptor signaling pathway	5.562	4.486	5.051
intestinal immune network for iga production	4.595	4.415	4.747
glutathione metabolism	2.941	3.567	4.649
systemic lupus erythematosus			4.033
apoptosis	3.272	2.792	3.942

Table 10: Top ten downregulated gene sets according to RNA-seq data, in BMI comparison.

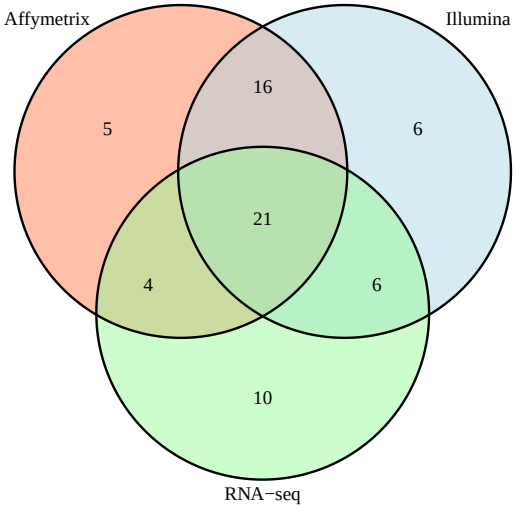

Figure 28: Overlap between enriched gene sets, age.

Gene set	Affymetrix	Illumina	RNA-seq
metabolism of xenobiotics by cytochrome p450	-9.296	-8.630	-7.605
steroid hormone biosynthesis	-8.401	-7.460	-5.795
propanoate metabolism	-2.769	-3.152	-5.039
pathogenic escherichia coli infection		-2.673	-4.186
arrhythm. right ventr. cardiomyop.	-4.536	-3.596	-4.144
tight junction	-4.319	-4.481	-4.030
leukocyte transendothelial migration	-4.048	-4.866	-3.984
valine leucine and isoleucine degradation		-2.705	-3.909
drug metabolism cytochrome p450	-7.535	-6.247	-3.889
starch and sucrose metabolism	-7.993	-5.128	-3.869

Table 11: Top ten downregulated gene sets according to RNA-seq data, in age comparison.

Gene set	Affymetrix	Illumina	RNA-seq
ribosome	4.273	5.030	15.526
oxidative phosphorylation	3.328	4.522	12.603
parkinsons disease		3.564	8.810
huntingtons disease			7.805
alzheimers disease		3.155	7.312
lysosome			6.117
type i diabetes mellitus		2.976	6.089
graft versus host disease			5.409
allograft rejection			5.347
autoimmune thyroid disease			4.760

Table 12: Top ten downregulated gene sets according to RNA-seq data, in age comparison.

PAGE analysis shows between-platform correlations between 0.592 and 0.885. Illumina detects two additional "true positive" enriched gene sets(i.e. gene sets which are shown to be enriched for RNA-seq) for the BMI comparison(figure 26). In the age comparison(figure 28), Illumina detects two additional true positives as well, while also getting one additional false positive. False positive rates are consistent between Illumina and Affymetrix. The combined number of true positives is roughly equal to the combined number of false positives.

14 Conclusion

The results show that fold-change correlation is moderately high($\sim$0.8) although ranks of top genes are relatively discordant(less than 50% overlap). Gene set enrichment analysis seems to show a moderate agreement between platforms, but there is a substantial number of gene sets which show enrichment uniquely for each array platform. As such, it seems like data from the two platforms is comparable, although it is probably a good idea to conduct additional studies and develop methods to adjust for the differences.

There are some interesting extensions of this work that could be made:

- Refinements can be made to the analytical procedure. An example would be to improve the mapping between probes and genes, for instance by mapping probe sequences to current versions of the reference transcriptome.
- The main goal of this project was to investigate how well data from the Illumina and Affymetrix array platforms correspond. Attempting to com-

bine data across platforms is a natural step forward.

- Earlier studies, and this project, show a non-linear relationship between signal intensities for each gene. It is possible that the data from the two microarray platforms and the RNA-seq data could be used as the basis for approximating a function which converts intensities from one array platform to equivalent intensities on another array platform.

15 Acknowledgements

I would like to thank my supervisors, Joao Fadista, Petter Storm and Leif Groop, for guidance and support with this project.

16 Supplementary material:Between-Platform correlations

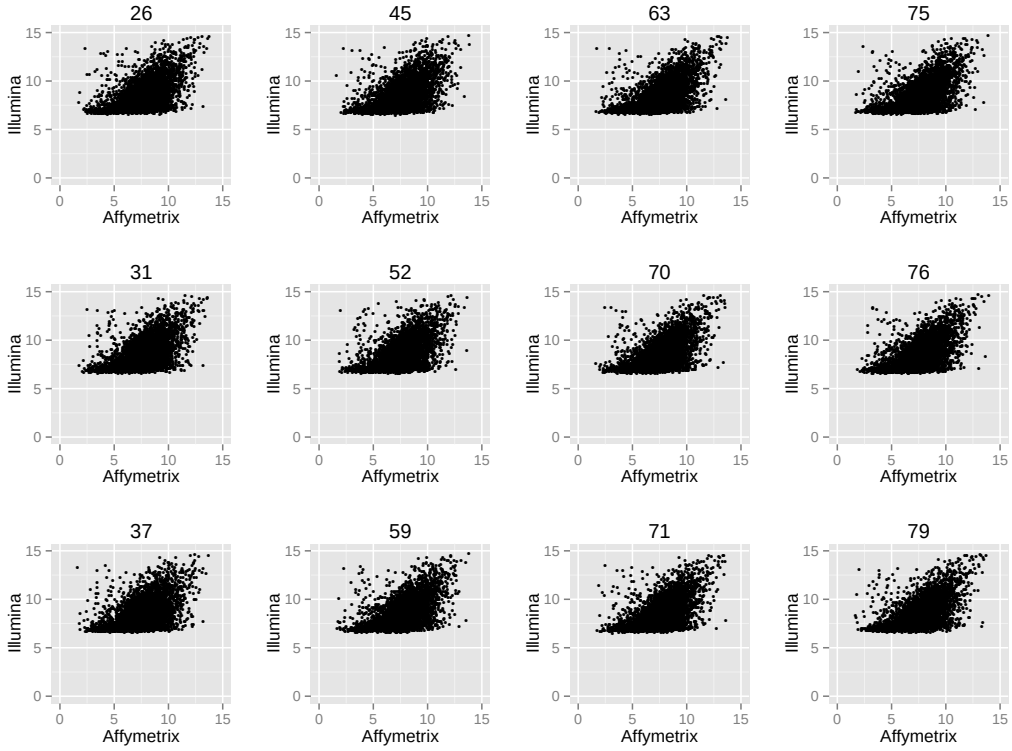

Figure 29: Affymetrix/Illumina correlation.

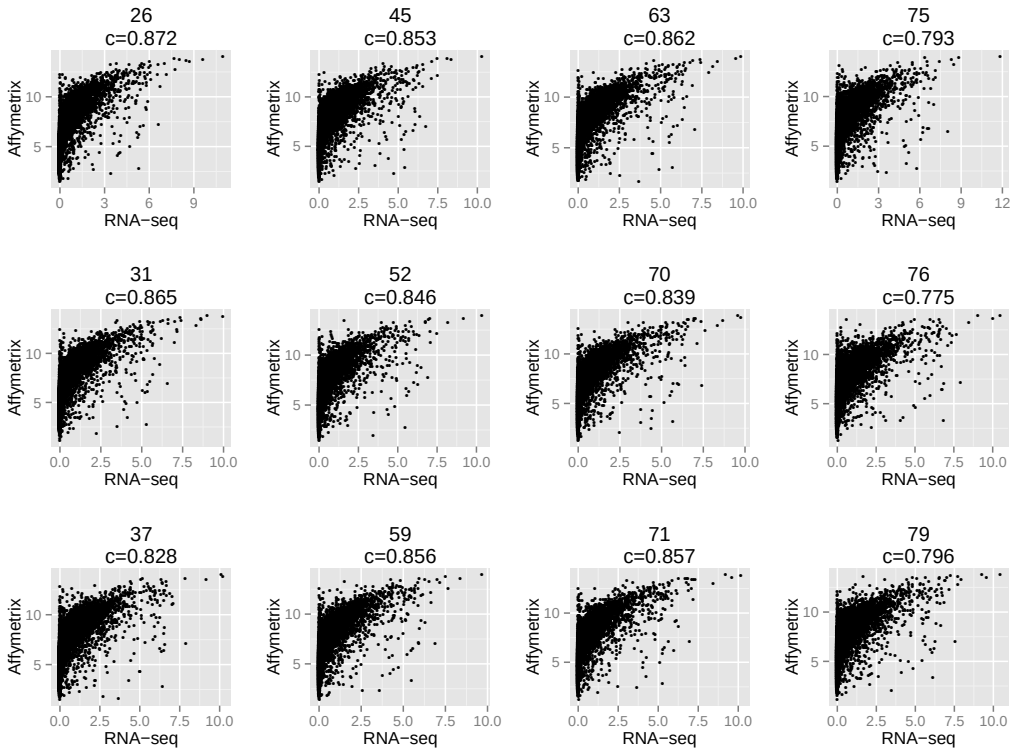

Figure 30: RNA-seq/Affymetrix correlation.

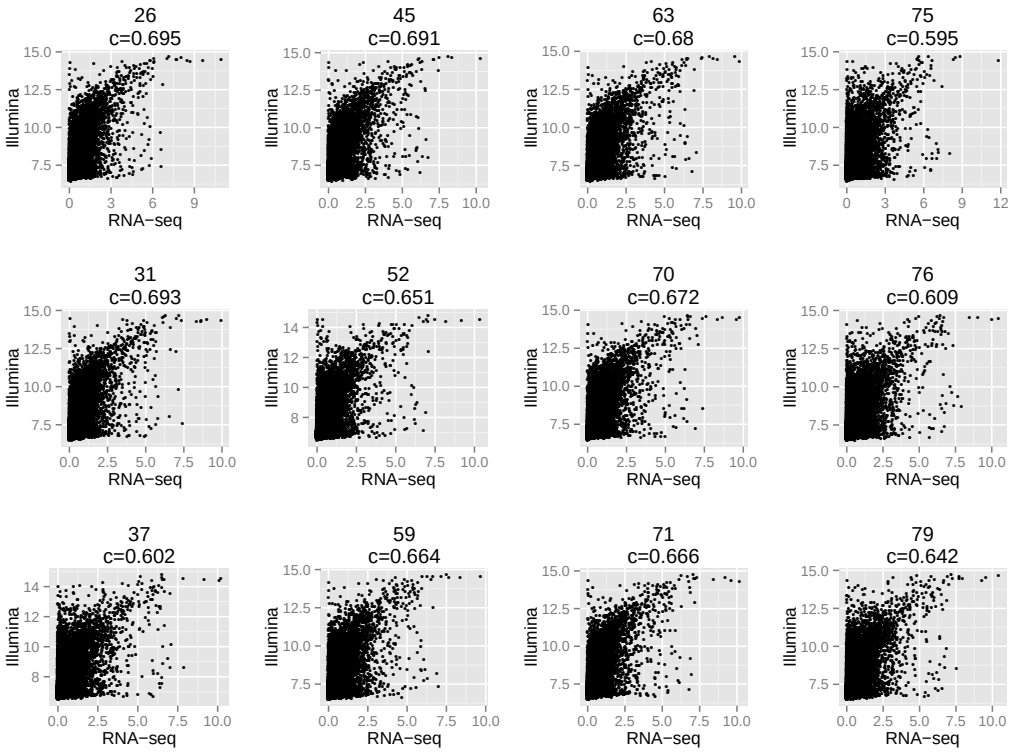

Figure 31: RNA-seq/Illumina correlation.

17 Software tools and versions

The following tools were used in the analysis:

- Biobase[40] 2.18.0
- Array loading and processing:
  - lumi[46] 2.10.0
  - affy[44] 1.36.1
- limma[51] 3.14.4
- Annotation:
  - lumiHumanAll.db[50] 1.18.0
  - hugene10sttranscriptcluster.db[49] 8.0.1
  - hugene10stv1cdf[48] 2.11.0
- PGSEA[54] 1.32.0
- Plotting:
  - VennDiagram[43] 1.6.5
  - reshape2[42] 1.2.2
  - ggplot2[41] 0.9.3.1

References

[33] R. Plomin and L. C. Schalkwyk, "Microarrays," *Developmental Science*, vol. 10, p. 19–23, Jan. 2007. PMID: 17181694 PMCID: PMC2776927.

[34] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics (Oxford, England)*, vol. 4, pp. 249–264, Apr. 2003. PMID: 12925520.

[35] "exodiab.se: Human tissue lab." http://www.exodiab.se/human-tissue-lab/. Accessed: 2014-01-07.

[36] N. Masuda, T. Ohnishi, S. Kawamoto, M. Monden, and K. Okubo, "Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples.," *Nucleic Acids Research*, vol. 27, p. 4436–4443, Nov. 1999. PMID: 10536153 PMCID: PMC148727.

[37] Z. Zhang, D. L. Gasser, E. F. Rappaport, and M. J. Falk, "Cross-platform expression microarray performance in a mouse model of mitochondrial disease therapy," *Molecular genetics and metabolism*, vol. 99, pp. 309–318, Mar. 2010. PMID: 19944634.

[38] K.-O. Mutz, A. Heilkenbrinker, M. Lönne, J.-G. Walter, and F. Stahl, "Transcriptome analysis using next-generation sequencing," *Current opinion in biotechnology*, vol. 24, p. 22–30, Feb. 2013. PMID: 23020966.

[39] R. Team, "R: A language and environment for statistical computing(http://www.R-project.org)," 2004.

[40] R. C. Gentleman, V. J. Carey, D. M. Bates, and others, "Bioconductor: Open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, p. R80, 2004.

[41] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[42] H. Wickham, "Reshaping data with the reshape package," *Journal of Statistical Software*, vol. 21, no. 12, pp. 1–20, 2007.

[43] H. Chen, *VennDiagram: Generate high-resolution Venn and Euler plots*, 2013. R package version 1.6.5.

[44] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "affy—analysis of affymetrix genechip data at the probe level," *Bioinformatics*, vol. 20, no. 3, pp. 307–315, 2004.

[45] M. J. Dunning, M. E. Ritchie, N. L. Barbosa-Morais, S. Tavare, and A. G. Lynch, "Spike-in validation of an illumina-specific variance-stabilizing transformation," *BMC Research Notes*, vol. 1, p. 18, June 2008. PMID: 18710543 PMCID: PMC2518281.

[46] P. Du, W. A. Kibbe, and S. M. Lin, "lumi: a pipeline for processing illumina microarray," *Bioinformatics (Oxford, England)*, vol. 24, pp. 1547–1548, July 2008. PMID: 18467348.

[47] M. E. Ritchie, M. J. Dunning, M. L. Smith, W. Shi, and A. G. Lynch, "BeadArray expression analysis using bioconductor," *PLoS Computational Biology*, vol. 7, Dec. 2011. PMID: 22144879 PMCID: PMC3228778.

[48] T. B. Project, *hugene10stv1cdf: hugene10stv1cdf*. R package version 2.11.0.

[49] A. Li, *hugene10sttranscriptcluster.db: Affymetrix Human Gene 1.0-ST Array Transcriptcluster Revision 8 annotation data (chip hugene10sttranscriptcluster)*. R package version 8.0.1.

[50] M. Carlson, S. Falcon, H. Pages, and N. Li, *lumiHumanAll.db: Human Illumina annotation data (chip lumiHumanAll)*. R package version 1.18.0.

[51] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, eds.), pp. 397–420, New York: Springer, 2005.

[52] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, pp. 139–140, Jan. 2010. PMID: 19910308.

[53] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "Voom! precision weights unlock linear model analysis tools for rna-seq read counts," tech. rep., 2013.

[54] K. Furge and K. Dykema, *PGSEA: Parametric Gene Set Enrichment Analysis*, 2012. R package version 1.32.0.

[55] "Beadarray r package manual – beadarray use cases." http://www.bioconductor.org/packages/devel/data/experiment/vignettes/BeadArrayUseCases/inst/doc/BeadArrayUseCases.pdf. Accessed: 2014-01-17.

[56] C. Cheadle, K. G. Becker, Y. S. Cho-Chung, M. Nesterova, T. Watkins, r. Wood, William, V. Prabhu, and K. C. Barnes, "A rapid method for microarray cross platform comparisons using gene expression signatures," *Molecular and cellular probes*, vol. 21, pp. 35–46, Feb. 2007. PMID: 16982174.

Corrections

- Table 1 has been recalculated.
-