Quality Control and Analysis of RNA-seq Data from Breast Cancer Tumor Samples

Christian Brueffer*[1]

Supervisor: Lao Saal[1]

[1]Translational Oncogenomics Unit, Canceromics Branch, Division of Oncology, BMC C13, SE-221 84 Lund, Sweden,

Email: Christian Brueffer*- christian.brueffer@med.lu.se;

*Corresponding author

Abstract

Background: Breast cancer is the most common kind of cancer among women in Sweden. While the short- to mid-term survival chances are good, the long-term survival chances are poor, and a large number of women are also likely being overtreated and thus suffer from unnecessary side effects. The South Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative aims at improving breast cancer outcome by developing new diagnostics and predictive tests based on RNA sequencing (RNA-seq) technology. With RNA-seq being a complicated technology with many error sources, quality control is needed to gain confidence in the obtained data.

Results: During this project an RNA-seq quality control pipeline was built and integrated into the existing SCAN-B RNA-seq analysis pipeline. The quality control pipeline was used to evaluate the quality of 2547 RNA-seq libraries. The evaluation showed good overall quality of the data. While the quality of the first sequenced libraries is not optimal, quality has increased steadily and settled in on a high level.

Conclusions: Quality control is essential for the RNA-seq analysis process. The metrics used in this project provide good insight into the quality of the evaluated datasets. However, cancer cells feature a distinct genomic landscape which can make the interpretation of metrics difficult. Thus, care has to be taken when drawing conclusions about the quality of RNA-seq data from cancer-derived samples.

Background

Since the publication of the first draft sequence of a human genome by the Human Genome Project [1] and Celera Corporation [2] in 2001, the idea of personalized medicine has received increased interest from many medical practitioners and scientists. Personalized medicine refers to medical treatment which is customized to each patient. The human genome draft sequence as well as sequencing technologies which are rapidly getting faster yet cheaper provide an unprecedented opportunity to implement this idea.

One of the diseases whose treatment could benefit from personalized medicine is cancer. Cancer is characterized by uncontrolled cell growth and proliferation. These characteristics are due to an

accumulation of aberrations at the genome level. In a normal cell, genes involved in stimulating and inhibiting cell growth and proliferation are in balance. When these genes get modified due to mutations, structural rearrangements or modifications in their regulation, this balance is destroyed. Several other processes (i.e., angiogenesis and apoptosis evasion) are involved in stimulating tumor growth and cancer progression, summarized by Hanahan and Weinberg in their "hallmarks of cancer" [3].

Breast Cancer

Breast cancer is a type of cancer which arises in cells of the mammary gland. The vast majority of cases affect women, with only 1% of cases affecting men. While most breast cancer incidence is usually sporadic, a hereditary risk exists in 5%-10% of cases, of which about half are caused by *BRCA1/BRCA2* germline mutations [4, 5].

Classically, breast cancers have been classified in two ways: (1) by histopathology such as morphological subtype (growth pattern of the tumor when viewed under the light microscope), grade (appearance of cancer cells compared to normal cells) and receptor status (expression of receptors on the surface of the tumor cells), and (2) stage (tumor size and extend of spread within the body). Cancer cells may have different receptors. The ones most commonly used for classification are estrogen receptor (ER), progesterone receptor (PR) and the human epidermal growth factor 2 (HER2) receptor, because of their associations to outcome and response to therapy [6]. Patients whose tumors are positive for both ER and PR have a better prognosis than any of the other ER/PR combinations [7].

In recent years a new way of classifying breast cancer tumors based on molecular profiles has been introduced [8, 9]. These subtypes have been determined by gene expression profiling, where the genes whose expression varies most between tumors have been identified. The major subtypes found this way are luminal, HER2 and basal. Luminal tumors show high gene expression of hormone receptors and associated genes, while HER2 tumors have high expression of HER2 and other genes in the *ERBB2* amplicon (17q11.2-12). Basal tumors are characterized by high expression of basal cytokeratins such as CK5/14, as well as low expression of ER, PR and HER2. These tumors are also called triple-negative tumors [10].

Primary treatment for breast cancer is surgical removal of the tumor, which may be followed by systemic therapy with anti-hormonal agents and chemotherapy, as well as radiation treatment. Common therapeutic agents include tamoxifen (ER antagonist), trastuzumab (monoclonal antibody interfering with HER2) and different aromatase inhibitors (block estrogen synthesis), i.e. anastrozole, letrozole and exemestane. Chemotherapeutic agents include cisplatin (induces apoptosis by DNA-crosslinking), doxorubicin (DNA intercalating agent) and paclitaxel (interferes with cell division) [11].

Of 27,688 new cancer cases in women reported in Sweden in 2011, 8382 were identified as breast cancer. This equals 30.1% of new cases, making breast cancer the most common kind of cancer among women in Sweden [12]. With rates of 98% 1-year survival and 88.5% 5-year survival [13], the short- to mid-term survival chances are good. However, the long-term survival chances are much worse. Brenner and Hakulinen show these to be about 60% at 15-year followup and about 50% at 20-year followup [14]. Furthermore, a significant fraction of women are likely cured by surgery alone, or may only need milder systemic therapy, but are being overtreated today and thus suffer from unnecessary side effects. In addition to its health and psychosocial effects, overtreatment also poses a significant economic burden on healthcare systems [15].

The South Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative

The *South Sweden Cancerome Analysis Network - Breast* (SCAN-B) project was started in 2010 by Professor Åke Borg; my thesis mentor Lao Saal is a founding member of the steering group and oversees the research aspects. The project is a joined effort of "surgeons, pathologists, oncologists, radiologists, nurses, and biologists who strive to improve survival and quality of life for breast cancer patients" [16]. SCAN-B currently has seven participating hospitals in Malmö, Lund, Helsingborg, Växjo, Halmstad, Kristianstad and Karlskrona, all located in the South Sweden healthcare region. The project's goals are threefold [16]:

- Introduce gene expression and genomic tumor

profiling into the clinical routine for breast cancer

- Improve tumor classification, diagnosis, prognostication and prediction of treatment effects
- Eventual health care implementation, clinical trials, cooperation with drug and biotech industry, and an accelerated pipeline towards personalized care

Within SCAN-B, a tumor sample is taken from each consenting patient. In addition, blood samples are taken before the surgery and at defined follow-up intervals. The samples are sent to Oncology Department of Lund University for further analysis. Currently the analysis process consists of performing RNA sequencing (RNA-seq) of the tumor samples. The resulting RNA-seq data is the basis for the work presented in this thesis.

As of May 2013, 3513 patients have consented to be part of SCAN-B. From these patients, 3399 blood samples and 2780 tumor samples have been collected.

Additional retrospective cases have been incorporated in SCAN-B: approximately 450 tumor samples from the All Breast Cancer in Malmö (ABIM) project collected between 2007 and 2010, as well as approximately 85 samples from Lund and 60 hereditary *BRCA1*/*BRCA2* cases.

RNA Sequencing (RNA-seq)

RNA sequencing (RNA-seq) [17] is a tool for transcriptome analysis. It employs high-throughput analysis to determine the RNA sequences and their abundance in a sample. In most cases the RNA is reverse-transcribed into complementary DNA (cDNA) first, which is then sequenced. Direct RNA sequencing is possible, however the techniques are currently immature and therefore less commonly used [18, 19].

For a sample to be sequenced, it has to be transformed into a sequencing library. This involves series of steps termed a library preparation protocol. Different protocols with varying properties exist. One of the most important properties is strandedness, which determines whether the protocol retains the information of which DNA strand transcripts were transcribed from. This is particularly important when a transcript has arisen from a section of DNA which has overlapping genes on the opposing strands.

Levin et al. [20] compared different stranded protocols regarding several metrics. The dUTP protocol [21] emerged as the clear winner from this comparison, performing best in most metrics. Another comparison was performed within the Department of Oncology, comparing dUTP with the Illumina TruSeq and the Epicentre ScriptSeq protocols. Based on these two comparisons, the dUTP protocol was chosen to be used for RNA-seq library preparations within the SCAN-B initiative. The currently employed protocol is a customized and optimized version of the protocol by Parkhomchuk et al. [21]. The sequencing itself is performed with an Illumina [22] HiSeq 2000 sequencer.

SCAN-B dUTP-based Library Preparation Protocol

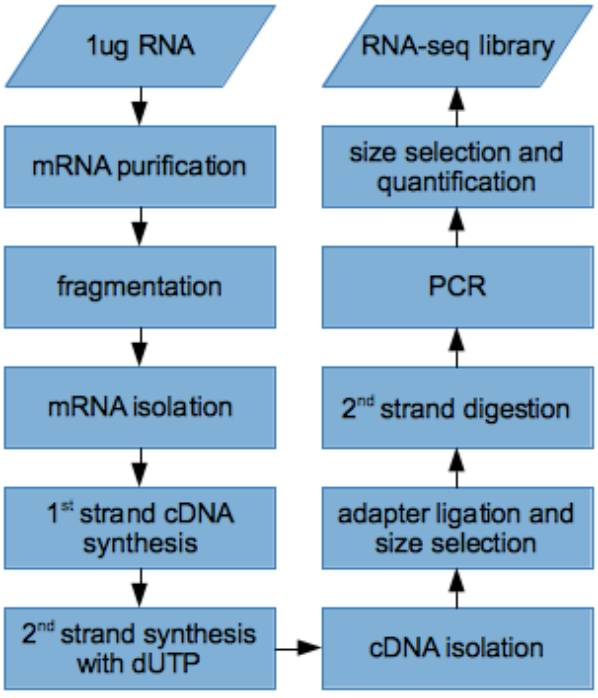

Figure 1: Simplified dUTP process flow.

The steps involved in the customized dUTP library preparation protocol are illustrated in Figure 1. As source material, $1\mu g$ of total RNA is used. First, the mRNA is purified using poly-dT Dynabeads. Then, it is fragmented into approximately 240 base-pair

fragments catalyzed by zinc cations and the fragments isolated using a Zymo spin column. Synthesis of the first cDNA strand is induced by adding random hexamers which act as primers, as well as reverse transcriptase and nucleotides. After removal of excess reagents, the second strand is synthesized with an admixture containing dUTP in place of dTTP, as well as polymerase and nucleotides. The resulting double-stranded cDNA is again isolated using a Zymo spin column. In the next step, different Illumina TruSeq adapters are ligated to the 5' and 3' ends of the cDNA. The adapters are needed for the actual sequencing process to work later on, but also include barcode sequences to enable matching of the sequenced reads to source samples. The mixture is size-selected to remove excess free adapters. So that the resultant adapter-ligated cDNA maintains strandedness, the second strand which includes dUTP is then specifically digested using uracil-DNA glycosylase. The resulting cDNA fragments are amplified by polymerase chain reaction (PCR) and isolated using size selection. Subsequent quantification is performed to estimate the amount of cDNA for basic quality control.

The result of the library preparation is double-stranded cDNA, where one half of the duplex has suitable adapter sequences on its 5' and 3' ends which is ready for sequencing.

Illumina Sequencing by Synthesis Method

Illumina HiSeq 2000 sequencing machines utilize the sequencing by synthesis method [23]. The machine performs the actual sequencing process in a "flow cell". A flow cell is a compartment containing lanes, in which reagents flow in from one side and leave on the other side. The Illumina HiSeq 2000 can operate two flow cells at the same time, with flow cells currently having eight lanes. In the SCAN-B RNA-seq used herein, 16 sequencing libraries resulting from the library preparation are mixed into one pool. Each pool is applied to two lanes on two flow cells, respectively. Consequently, 4 pools containing 64 libraries are sequenced in parallel. Through the pooling, every library is sequenced on 2 lanes per flow cell to provide redundancy against lane failure.

Using the sequencing adapters ligated during the library preparation, the DNA fragments are randomly immobilized onto baits, which are attached to the surface of the lanes. Baits are short oligionucleotide sequences which are complementary to the free end of the adapters. Hundreds of millions of "clusters" of identical DNA molecules, initiated by a single clonal DNA sequence, are formed *in situ* in the flow cell using bridge amplification. Bridge amplification works by attached DNA fragments bending and hybridizing to a nearby bait with its free adapter, forming an arch. The bait acts as primer for PCR with the DNA fragment being the template. The resulting double-stranded DNA is then denatured, leading to two single-stranded surface-attached templates. This process, also known as solid-phase amplification, creates clusters containing up to 1000 identical copies out of each DNA fragment in the prepared library [23, 24].

During the actual sequencing process, modified nucleotides which are labeled with different dyes and include a terminating group are used. First, primers, modified nucleotides and DNA polymerase are added to the flow cell. Once a primer attaches to a DNA template, the DNA polymerase starts second strand synthesis by incorporating the first nucleotide. To read which nucleotide has been incorporated, the sequencer excites the clusters with a laser, inducing fluorescence. The fluorescent spots are detected with a charge-coupled device (CCD) sensor. Since a flow cell lane usually contains many millions of clusters, the resolution of the sensor has to be high and the sequencer's image recognition software has to keep track of millions of spots at the same time. From the detected spots, the actual bases of the DNA sequences are then called. Due to each cluster having the same nucleotide incorporated at the same time, the signal is strong enough for the sequencer to detect reliably. After the detection, the terminating group in the nucleotides is cleaved, so a new incorporation and detection cycle can start.

Using the sequencing by synthesis method, only relatively short sequence stretches can be sequenced with acceptable accuracy [25]. The Illumina HiSeq 2000 is currently restricted to a maximum read length of 100 bases [26], which can lead to problems in post-processing steps such as mapping. This disadvantage is alleviated by the high throughput of sequencing as well as paired-end sequencing. After the clusters have been sequenced, they are regenerated in such a way that the DNA fragments attach to the flow cell surface with the other sequencing adapter. Following this, sequencing starts again leading to DNA which has been sequenced from both ends. Reads generated this way are called paired-end reads. The two individual reads in a paired-end

read are called "mates" or "ends". Paired-end libraries are designated by their read length, i.e. 2x50 or 2x100.

Quality Control

An RNA-seq experiment consists of many steps, spanning from sample extraction over library preparation to the actual sequencing. Each step has the possibility of influencing the experimental data in undesirable ways and introducing errors and biases. Quality control of the output data is a crucial step needed to quantify these problems. While basic quality control measures are part of the library preparation process, comprehensive quality control can only be performed after the sequencing process.

To evaluate different strand-specific library preparation protocols, Levin et al. [20] identified five key metrics:

- *Library complexity*
  Library complexity refers to the number of unique start and end positions of read pairs in the library. This value should be high as it reflects the random shearing of DNA.

- *Strand specificity*
  Strand specificity describes the percentage of reads which in sense direction. This value should ideally be 100%.

- *Evenness of coverage*
  Evenness of coverage is the the coefficient of variation (CV) for the distribution of reads along a transcript. A low CV is preferred, as it indicates an even distribution.

- *Continuity of coverage*
  Continuity of coverage refers to whether or not gaps exists in the coverage of an annotated transcript. Ideally, coverage should be completely continuous. This metric is defined as the percentage of gaps per transcript.

- *Coverage at 3' and 5' ends*
  Annotated transcript coverage at the 3' and 5' ends is an indicator for how well transcription boundaries are covered. High coverage is preferred.

However, these metrics only describe the properties of the prepared library. Additional metrics are needed to assess the quality of the sequencing process, as well as the alignment of the library reads to a reference genome. The most common metric for sequencing quality is the base quality as estimated by the sequencing machine. This metric is commonly expressed as a Phred score [27,28]. The Phred score Q is defined as $Q = -10 * log_{10}P$, where P is the probability of calling a wrong base. A Phred score of 30 therefore equals base calling accuracy of 99.9%.

For the alignment to a reference genome, it is possible to calculate the mapping rate (number of aligned reads divided by all reads), the number of reads that align uniquely and the number of reads that align in multiple places in the genome.

The goal of this project was twofold: To integrate quality control into the RNA-seq pipeline used for the SCAN-B initiative, and to perform systematic evaluation of the quality of RNA-seq datasets which have already been sequenced as part of SCAN-B.

Methods

SCAN-B RNA-seq Pipeline

RNA-seq experiments produce massive amounts of data. The SCAN-B data consists of paired-end reads which need to be further processed before they are useful for analysis. The current SCAN-B processing pipeline starts with the sequencing using an Illumina HiSeq 2000 sequencer. Since each sequencing library is typically spread out over multiple lanes on both flow cells, the data needs to be demultiplexed and merged into a single data file in the FASTQ format. The sequencing is performed using a read length of 2x100 bases. In case a read length of 2x50 is desired, the 2x100 data is truncated. Downsampling is performed, if requested. Next, the read file is filtered for ribosomal RNA (rRNA), human DNA repeat regions, and phiX virus DNA. The filtered reads are then aligned to a reference genome using the TopHat2 splice junction mapper [29,30], with an option turned on to enable finding of *de-novo* transcripts.

The utilized reference genome is a custom genome. It mainly consists of chromosomes 1-22, X and mitochondrial DNA from the hg19 human genome assembly [1], as well as the Y chromosome from the b37 assembly [31]. Added to this are extra human genome sequences which have not been

added to the official assembly yet, as well as decoy sequences [32]. Like the extra sequences, the decoy sequences have not been added to the official human genome assembly yet. They mostly contain satellite, simple and interspersed repeat regions. The purpose of these decoy sequences is to improve discovery of single nucleotide polymorphisms (SNP).

The last step of the pipeline is the use of the Cufflinks software [33] for transcript expression estimation. This uses a transcript model which includes information about known transcripts. The model used is the UCSC hg19 refGene model from 2013-01-29 in GTF format [34], containing 44,202 transcripts.

Quality Control Package Selection

Several different RNA-seq quality control packages exist. Some packages provide rudimentary metrics, while others are more comprehensive. Wang et al. [35] produced a table which compares the functionality of common software which provides some form of RNA-seq quality control metrics.

The three most comprehensive packages are RNA-SeQC [36], RSeQC [35] and Qualimap [37]. García-Alcalde et al. [37] provide a comparison table which shows the three packages are very similar in terms of functionality.

After an evaluation process, RNA-SeQC was selected as the quality control package to use for SCAN-B. Although the packages are functionally similar, there are differences. RNA-SeQC's biggest advantage over RSeQC are it's pipeline-ready summary metrics file. While Qualimap is pipeline-ready, it does not provide library complexity metrics, a true multi-sample mode and expression correlation. Even though the latter two are currently unused in the SCAN-B pipeline, they are likely to prove useful in the future.

A downside of RNA-SeQC is its lack of flexibility regarding which metrics to run. While it is possible to disable certain metrics, it is not possible to run only a specific subset of metrics to decrease execution time.

In summary, any of the three comprehensive quality control packages would have been a suitable choice for the SCAN-B pipeline. RNA-SeQC was chosen due to minor advantages over the other two packages.

Quality Control Pipeline

The quality control pipeline consists of two steps: data preprocessing and execution of the RNA-SeQC quality control package.

Data Preprocessing

TopHat, Picard [38] and the Genome Analysis Toolkit (GATK) [39] all operate with files following the Sequence Alignment/Map (SAM) file format [40] as well as its binary version, the Binary Alignment/Map (BAM) format. However, while the file format is standardized, there are differences in the expectations and interpretations of certain fields in a SAM record.

TopHat writes separate files for reads which map to the reference genome (`accepted_hits.bam`) and reads which do not map (`unmapped.bam`). For most analyses, it is sufficient to only work with the mapped reads. However, for quality control it is desirable to use all reads to obtain a view of the entire sequencing library. Experimentation revealed that various steps were necessary to merge the mapped and unmapped files, and to make the merged file compatible with RNA-SeQC. Most incompatibilities are rooted in the unmapped file. The following changes to reads in this file are needed for downstream preprocessing steps to succeed:

- Remove /1 and /2 suffixes from read names (only needed for TopHat up to version 2.0.6)
- Set the "next segment in the template unmapped" bit (0x8) in the FLAGS field if both mates are unmapped
- Set mapping quality (MAPQ field) to zero
- For unmapped reads with a mapped mate:
  - Set the RNAME and RNEXT fields to the value of the respective mapped mate
  - Set the POS field to the value of the mate's POS field
  - Set the PNEXT field to zero

The requirement to set the "next segment in the template unmapped" bit in the FLAGS field manually stems from a bug in the TopHat software which has been identified and reported during the course of this work, but is still present as of TopHat version 2.0.8. TopHat sets this bit correctly when one

mate is mapped and the other one is unmapped. However, it fails to set the bit when both mates are unmapped, leading to wrong assumptions in downstream processing software like Picard.

A script has been developed using the Python programming language [41] and the PySAM library [42], which applies the changes outlined above and which has been utilized as part of this QC preprocessing pipeline. The script has been published online [43] and has been successfully used by other people facing similar challenges [44].

After correcting the unmapped file it has to be sorted (Picard `ReorderSam.jar`) in the order of the reference genome, before it can be merged (`samtools merge`) with the mapped file. The merged file needs a read group (RG) SAM header, which is added using Picard `AddOrReplaceReadGroups.jar`. All reads are assigned to a single read group. As the last step, an index has to be created for the BAM file (`samtools index`).

The result of this pipeline is a BAM file which contains both mapped and unmapped reads and therefore represents all reads which served as input for the TopHat alignment stage in the SCAN-B RNA-seq pipeline. The file is compatible with the RNA-SeQC quality control software package.

RNA-SeQC Metrics

The result of the quality control pipeline is a directory structure containing the quality metrics for one RNA-seq dataset. The metrics are present in two forms; as an HTML-report (`report.html`) which can be used for manual inspection via a web browser, and as a tab-delimited file (`metrics.tsv`) for pipeline use. The HTML report contains not only global metrics about the dataset, but also links to coverage plots and detailed per-transcript statistics.

RNA-SeQC includes the following metrics in its `metrics.tsv` file:

- *Total Purity Filtered Reads Sequenced*
  The total number of reads, excluding *Failed Vendor QC Check* and *Alternative Alignments*. All metrics referring to "total reads" refer to this value.

- *Alternative Alignments*
  Duplicate read entries providing alternative coordinates (bit 0x100 set in SAM format FLAGS field).

- *Failed Vendor QC Check (Failed Reads)*
  The number of reads with bad quality (bit 0x200 set in SAM format FLAGS field).

- *Read Length*
  The maximum length found for all reads.

- *Estimated Library Size*
  The number of expected fragments in the sequenced library based upon *Total Purity Filtered Reads Sequenced* and duplication rate assuming a Poisson distribution. This is an approximation of the library complexity.

- *Mapped*
  The total number of mapped reads (unique and duplicate).

- *Mapping Rate*
  All mapping reads (unique and duplicate) divided by the total number of reads.

- *Mapped Unique*
  Number of reads which are aligned as well as non-duplicate.

- *Mapped Unique Rate of Total*
  Mapped Unique divided by *Total Purity Filtered Reads Sequenced*.

- *Unique Rate of Mapped*
  Mapped unique reads divided by all mapped reads.

- *Duplication Rate of Mapped*
  Mapped duplicate reads divided by all mapped reads.

- *Base Mismatch Rate*
  The number of aligned bases not matching the reference divided by the total number of aligned bases.

- *rRNA*
  Non-duplicate and duplicate reads aligning to rRNA regions as defined in the transcript model.

- *rRNA rate*
  The number of rRNA reads divided by the total number of reads.

- *Mapped Pairs*
  The total number of paired-end reads for which both ends map.
- *End 1 Mapping Rate*
  The number of mapped end 1 mates of paired-end reads divided by the total number of reads.
- *End 2 Mapping Rate*
  The number of mapped end 2 mates of paired-end reads divided by the total number of reads.
- *End 1 Mismatch Rate*
  The number of bases from end 1 mates of paired-end reads not matching the reference divided by the total number of mapped bases.
- *End 2 Mismatch Rate*
  The number of bases from end 2 mates of paired-end reads not matching the reference divided by the total number of mapped bases.
- *Fragment Length Mean*
  The mean length of all library fragments, as determined by the mapping positions of paired-end reads.
- *Fragment Length StdDev*
  The standard deviation of all library fragments, as determined by the mapping positions of paired-end reads.
- *Chimeric Pairs*
  The number of paired-end reads where both mates map to different genes.
- *Intragenic Rate*
  The fractions of reads mapping in within genes (within introns or exons).
- *Exonic Rate*
  The fraction of reads mapping within exons.
- *Intronic Rate*
  The fraction of reads mapping within introns.
- *Intergenic Rate*
  The fraction of reads mapping in the genomic space between genes.
- *Expression Profiling Efficiency*
  The ratio of exonic reads to total reads.
- *Transcripts Detected*
  The number of transcripts with at least 5 reads.
- *Genes Detected*
  The number of genes with at least 5 reads.
- *End 1 Sense*
  The number of end 1 mates of paired-end reads that were sequenced in the sense direction.
- *End 1 Antisense*
  The number of end 1 mates of paired-end reads that were sequenced in the antisense direction.
- *End 2 Sense*
  The number of end 2 mates of paired-end reads that were sequenced in the sense direction.
- *End 2 Antisense*
  The number of end 2 mates of paired-end reads that were sequenced in the antisense direction.
- *End 1 % Sense*
  The number of end 1 mates of paired-end reads that were sequenced in the sense direction divided by the number of all end 1 reads.
- *End 2 % Sense*
  The number of end 2 mates of paired-end reads that were sequenced in the sense direction divided by the number of all end 2 reads.
- *Mean Per Base Cov.*
  Mean per base coverage for the 1000 middle expressed transcripts.
- *Mean CV*
  Mean coefficient of variation across the 1000 (default setting) middle expressed transcripts.
- *No. Covered 5'*
  Number of 5' ends whit at least one mapped read.
- *5' Norm*
  Mean per base coverage of 5' read ends of middle expressed transcripts.
- *3' Norm*
  Mean per base coverage of 3' read ends of middle expressed transcripts.
- *Num. Gaps*
  Number of stretches with at least 5 bases having zero coverage.
- *Cumulative Gap Length*
  Sum of gap length of all middle expressed transcripts.

- *Gap %*
  The total cumulative gap length divided by the total cumulative transcript lengths of middle expressed transcripts.

In addition to the coverage statistics for middle expressed reads, the HTML report also includes these statistics for the top (default 1000) and bottom (default 1000) expressed transcripts. The report also includes graphs for coverage and GC bias.

From the metrics included in the `metrics.tsv` file, one more metrics can be calculated:

- *Strand Specificity*
  The mean of the percentage of End 1 reads mapping in antisense direction (100 − *End 1 % Sense*) and *End 2 % Sense*.

- *Total Reads*
  The sum of *Total Purity Filtered Reads Sequenced*, *Alternative Alignments* and *Failed Vendor QC Check*.

SCAN-B Data

The RNA-seq data available within the SCAN-B project is divided into three subsets, based on the RNA source and the purpose of the data:

- SCAN-B dataset (`scanb`)
- Validation dataset (`validation`)
- Quality Control SCAN-B dataset (`qcscanb`)

The libraries in all three subsets have been sequenced using a read length of 2x100 base pairs. In addition to this, every library is *in silico* truncated to a read length of 2x50 base pairs. Sequencing runs were organized by multititer "plates" containing up to 64 libraries from the `scanb` and `validation` datasets. Plates have been sequenced in chronological order, starting with plate 1. If a sequencing run failed for a plate, the plate was excluded. Thus, some plate numbers are skipped in the final dataset.

The number of reads per library in the datasets range from 55,684 to 504 million, with a median of 85.8 million and a mean of 86.67 million.

The `scanb` dataset consists of sequencing libraries from samples which are part of the SCAN-B patient set and preserved using the RNAlater preservative. The set consists of 99 sequenced libraries from 85 patients. In most cases, each library has been sequenced once. From 9 of these patient's samples, two libraries were prepared and sequenced. For one patient, two samples were available and subsequently sequenced. In addition, three technical replicates are part of this dataset. In addition, for each of these libraries, a downsampled library with 30 million reads was produced.

The samples sequenced for the `validation` dataset originate from the retrospective ABIM and Lund collections of snap-frozen tumor biopsies. The dataset consists of 570 sequenced libraries from 511 patients. Most of these libraries have been sequenced once, however 37 libraries were sequenced twice and 22 technical replicates were produced. The dataset is targeted at potential inclusion in the "validation project", which aims to develop RNA-seq-based gene expression classifiers for the five conventional biomarkers used in clinical practice today, ER, PR, HER2, grade, and Ki67.

Both `scanb` and the `validation` tumor samples were processed for RNA identically using the Qiagen AllPrep method.

The `qcscanb` dataset consists of 44 samples for which downsampled libraries with different amounts of reads have been created. The full libraries which serve as base for the downsamples have been selected from the `scanb` and `validation` datasets. The downsamples consists of 16 steps ranging from 1 million reads to 80 million reads in 0.5 million and 1 million increments. Overall, the dataset consists of 1126 2x100 base pair libraries and 1249 2x50 base pair libraries. The purpose of this dataset is to ascertain the differences in gene expression quantification and quality control metrics between the full 2x100 read data and the truncated 2x50 read data, as well as the differences between using all available reads and datasets downsampled to smaller numbers of reads *in silico*.

Quality Control Processing

Table 1: Number of processed libraries per dataset and read length.

Dataset	Number of Sequencing Libraries 2x50	2x100
`scanb`	99	99
`validation`	570	569
`qcscanb`	526	684

The quality control pipeline was applied to all subsets of the SCAN-B datasets. Two multi-core compute nodes with 100 gigabyte of main memory were used for the processing of the libraries. Due to the massive amounts of data per library, as well as the need for costly preprocessing steps, the runtimes for the analysis of one a typical SCAN-B library was approximately 80 minutes for a 2x50 library and 110 minutes for a 2x100 library. Table 1 shows the number libraries quality control runs were performed for per dataset and library type. The size of one RNA-SeQC output directory lies between 200 and 400 megabytes.

Aggregation and Plotting of Metrics Data

RNA-SeQC produces metrics in two forms, one of which is a tab-delimited file (`metrics.tsv`). This file is intended to be used in pipelines. A Python script was developed to traverse through all given RNA-SeQC output directories and gather the content of the `metrics.tsv` in one tab-delimited file. In addition, the data in the file is used to calculate the metrics *strand specificity* and *total reads*.

Using the aggregated quality control metrics in tab-delimited form as input, all plotting has been performed by Python scripts using the matplotlib [45] 2D plotting library.

Results and Discussion

The work performed during this project can be divided into four parts: (1) development of a quality control pipeline, (2) quality evaluation for the combined `scanb` and `validation` dataset, (3) quality evaluation for the `qcscanb` dataset as well as (4) comparison of SCAN-B RNA-seq quality evaluation to the Levin et al. *Saccharomyces cerevisiae* data using the same bioinformatics methods. The results of the four parts will be presented and discussed separate from each other in this section.

Development of an RNA-seq Quality Control Pipeline

The first and fundamental step for the subsequent parts of this project was the development of a pipeline for the quality control of RNA-seq datasets. This required evaluating and selecting a quality control software package, and making it work with the existing SCAN-B RNA-seq pipeline and sequencing data. The RNA-SeQC package was subsequently selected as quality control package. However, making the existing pipeline and sequencing data compatible with RNA-SeQC required a lot of experimentation with the BAM file format. Details about this process can be found in the Methods section.

The result of this part of the project are several scripts written in the Python and Bourne Shell programming languages, which make the existing RNA-seq pipeline as well as sequencing files compatible with RNA-SeQC. As part of this work, a bug in the popular splice junction mapper TopHat was identified, reported and worked around [43, 44].

SCAN-B and Validation Dataset Quality

In this section, plots for the metrics identified as important in the *Quality Control* section are going to be shown and discussed in detail. Since the data and plots for the 2x100 and 2x50 read length data are comparable for these metrics, only 2x50 plots are included here.

All plots shown in this section consist of boxplots of one metric, respectively. Each plot includes several boxplots which for libraries per-plate, as well as one boxplot for all `validation` libraries as well as one boxplot for all `scanb` libraries. The lower x-axis reflects the plate number as well as the respective dataset. The upper x-axis shows the number of values included in the respective boxplot. Table 2 shows the minimum, maximum, median and mean values of the analyzed metrics.

Number of Reads

Figure 2a shows the number of reads in the `validation` and `scanb` libraries, filtered for failed reads and alternative alignments. Table 2 shows the characteristical values for this data. The median is 80.53 million reads, with two outlier libraries above 450 million reads on plate 2.

Failed Reads

Figure 2b shows the percentage of reads which were marked as bad by the sequencing machine. The minimum, maximum and median values are 0.013%,

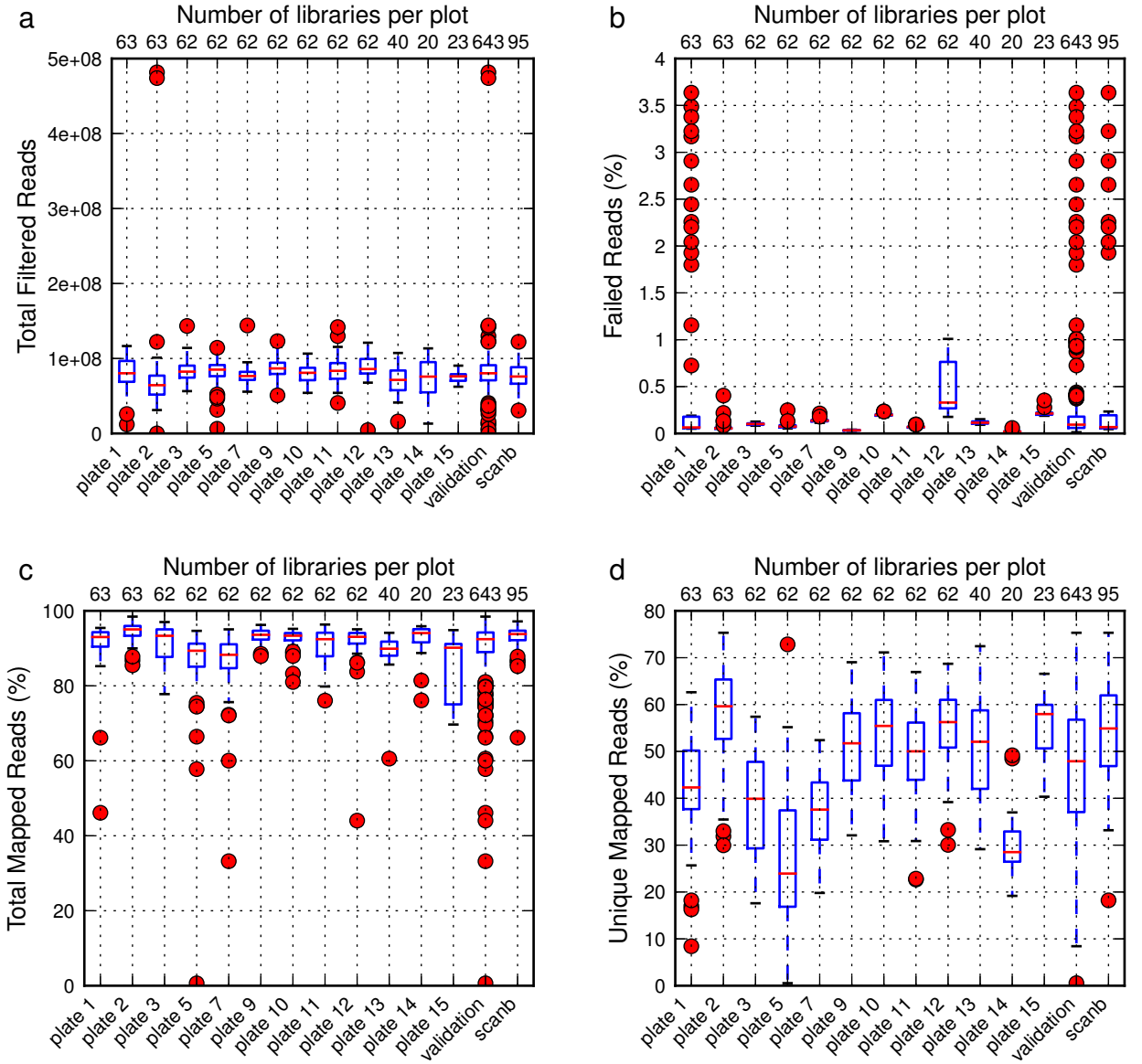

Figure 2: **a**) Total filtered reads, **b**) failed reads in percent, **c**) total mapped reads in percent, **d**) unique mapped reads in percent.

3.6% and 0.094%, respectively. The plot shows drastic numbers of failed reads in libraries on plate 1. Generally, the percentage of failed reads is stable and well below 0.25% in most cases. Plate 12 also shows an elevated level of bad reads, although on a much lower scale than on plate 1. The means and quartiles of the whole `validation` and `scanb` datasets well below 0.25%, with the vast majority of outliers coming from plates 1 and 12.

Mapping Rate

Figure 2c shows the total mapped reads in percent. The minimum, maximum and median values are 0.63%, 98.5% and 92.5% respectively. The median mapping rate is high across all samples, being above 90%. There are outliers, most notably a library with 0.6% on plate 5. Since the input reads to the quality analysis have already been filtered for rRNA as part of the analysis pipeline, the mapping rate of the

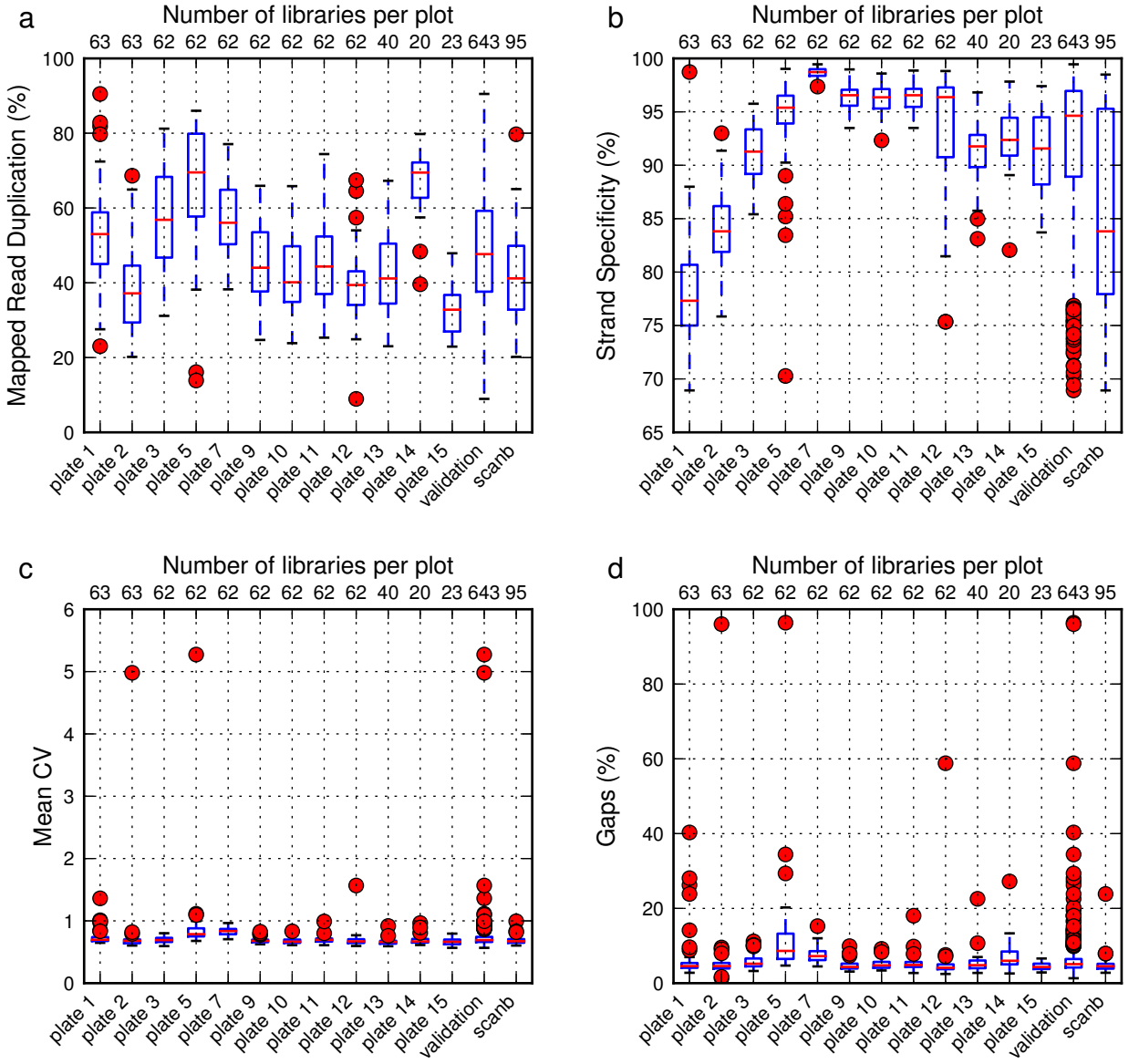

Figure 3: **a**) Mapped read duplication, **b**) strand specificity in percent, **c**) mean coefficient of variation, **d**) transcript gaps in percent.

whole dataset would be higher.

Library Complexity

Figure 2d shows the library complexity of the `validation` and `scanb` libraries. It is measured as the unique mapped reads percent of total filtered reads. The unique mapped reads vary between 0.55% and 75.3%. The median across all libraries is 48.5%. Plates 1, 3, 5 and 7 have median values well below the overall median, and libraries ranging into low percentage regions of unique mapped reads. From plate 9 on, the percentage is continuously on a high level, with the exception of plate 14.

Factors which can influence this measure is the quality of the RNA which enters the library preparation process,

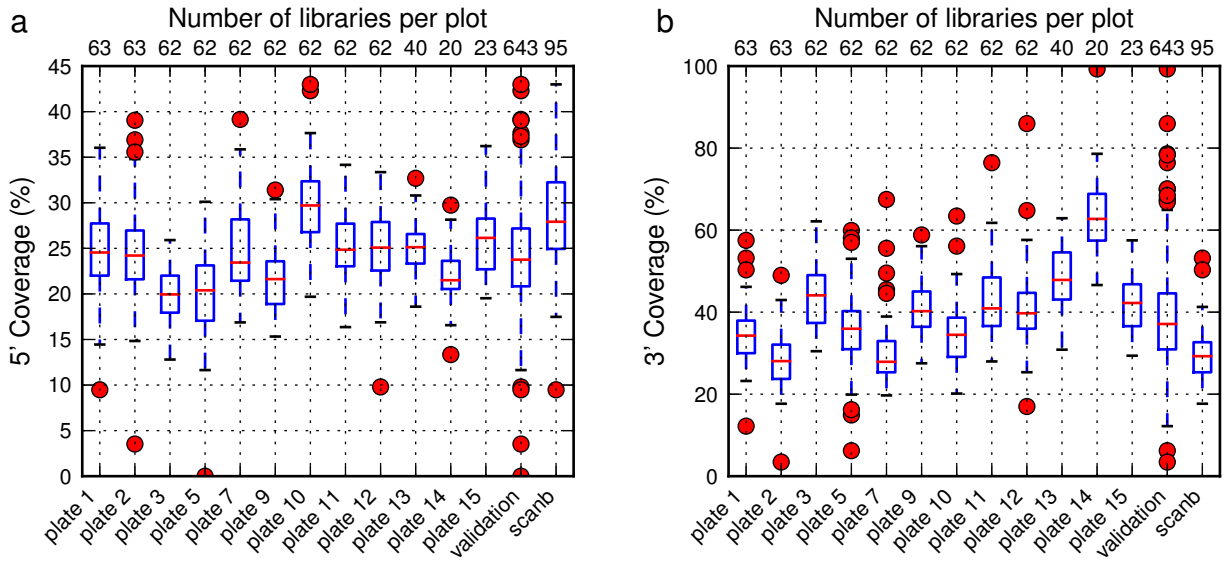

Figure 4: **a**) Coverage at 5' and **b**) coverage at 3' transcript ends.

Table 2: Combined `scanb` and `validation` dataset metrics properties.

Metric	Minimum	Maximum	Median	Mean
Total filtered reads	51,815	481.7 million	80.5 million	81.4 million
Failed reads (%)	0.013	3.6	0.094	0.19
Total mapped reads (%)	0.6	98.5	92.5	90.6
Unique mapped reads (%)	0.6	75.3	48.5	46.3
Mapped read duplication (%)	8.9	90.5	47.4	48.8
Strand specificity (%)	68.9	99.5	94.6	91.9
Mean CV	0.57	5.27	0.69	0.73
Transcript Gaps (%)	1.3	96.4	5	6.3
5' Coverage (%)	0	43	23.7	24
3' Coverage (%)	3.4	99.3	37.3	38.7

Mapped Read Duplication

Figure 3a shows the duplicated mapped reads in percent. The minimum, maximum and median values are 8.9%, 90.5% and 47.4% respectively. Plates 1, 3, 5 and 7 have a higher percentage of duplicated mapped read than the global median, indicating worse library quality. Compared to these plates, the plates from 9 on show a lower duplication percentage.

Strand Specificity

Figure 3b shows strand specificity. The minimum, maximum and median values across all libraries are 68.9%, 99.5%, and 94.6% respectively. On plate 1, strand specificity is generally low for a stranded library preparation protocol. However, with increasing plate number the strand specificity increases, peaking with plate 7 and continuing on a high level for plates 8-11 with values consistently over 95%. Plate 12 includes some less stranded libraries with values dropping below 85%. From plate 12 on, the values are steady on a high level with medians around 92%. The boxplot for the `validation` dataset shows a much higher median value than `scanb`. The reason for this is that the included `scanb` libraries were predominantly sequenced on plates 1 and 2 which show the worst strand specificity.

Evenness of Coverage

Figure 3c shows the mean coefficient of variation, which measures the evenness of coverage across transcripts. The minimum, maximum and median values across all samples are 0.57, 5.27 and 0.69, respectively. Across all plates, the median values are stable well below a CV of 1 with the quartiles being close to the respective medians. There are some outliers, however most of them also lie below a CV of 1. Two notable exceptions are libraries on plates 2 and 5, which show a CV of 5 and higher, thus having a very unevenly distributed reads.

Continuity of Coverage

Figure 3d shows the percentage of the transcripts which are uncovered by reads. The minimum, maximum and median values are 1.3%, 96.4% and 5%, respectively. The median values are stable around 5% across all transcripts, with the quartiles being close to the respective medians. There are many outliers, particularly on plate 1. Two outliers on plates 2 and 5 show a percentage of uncovered regions of above 95%, indicating massive problems in these libraries.

Coverage at 5' and 3' ends

Figures 4a and 4b show the coverage of 5' and 3' ends, respectively. Across all plates, the minimum, maximum and median values are 0%, 43% and 23.7% for the 5' case, as well as 3.4%, 99.3% and 37.3% for the 3' case. Generally, the coverage at both 5' and 3' ends is homogenous across plates. There are few outliers in the negative direction, the two worst ones again being on plates 2 and 5.

Discussion

Overall, the evaluated metrics show good results across the two datasets. Metrics like the library complexity, strand specificity and mapping rate show quality problems of the first five sequenced plates. However the sequencing quality increases steadily with plates 7 to 11 showing the most favorable quality across all metrics. The quality of the libraries consistently improving also shows the learning process of the laboratory staff in library preparation and the sequencing process.

The quality metrics can be correlated to observations made about problems during the sequencing process by the laboratory staff. During sequencing of plate 1, one flow cell gradually malfunctioned; eventually leading to many second reads of paired-end reads being sequenced with 38 bases instead of 100. This may also have impacted sequencing error rate, leading to the large number of failed reads in the impacted libraries. During the sequencing of plates 3 and 5, one lane and the chiller module malfunctioned, respectively. On the other hand, no events have been recorded during sequencing of plates 7 to 11, being reflected in these libraries performing well regarding the quality metrics.

The two worst outliers on plates 2 and 5 in the plots for mean CV and 5'/3' coverage are the same libraries in all cases (`2008264.1.l.r.lib` and `2009092.1.l.r.lib`). This shows these metrics are indicators for an underlying factor, namely general problems with these libraries. Generally, libraries performing badly in these metrics are easy to identify, leading to the possibility to re-sequence them.

Dependent on the metric, bad quality can have different sources. Low numbers of reads or high numbers of failed reads most likely stem from errors during the sequencing process. This is illustrated by the high numbers of failed reads on plate 1, combined with the observations the the laboratory staff. The mapping rate depends on the quality of the base calling and the reference genome. Mapping software like TopHat provide a vast amount of options which can also have an influence. For the other discussed metrics, problems during the library preparation are the most likely explanation.

Library complexity could be influenced by the quality of the RNA entering preparation process. If the RNA is degraded, i.e. due to conservation procedures, it has already lost part of its potential for generating unique reads.

Duplication is caused by the PCR step in library preparation. While PCR is needed to provide enough cDNA for the sequencing process, letting the amplification process run for too long can lead to excessive duplicate reads. However, the amount of source RNA which enters into the PCR is likely another important factor, which is also connected to the library complexity. If the rate of unique fragments in the input material is low, the number of duplications per unique fragment is high. Thus, all processed which lead to loss of input material (i.e., mRNA purification and size selection) may play a role.

In the dUTP protocol, strand specificity depends on the incorporation of dUTP into the second cDNA strand. Any process that leads to less dUTP being incorporated into the second strand (i.e., due to no or not enough dUTP, as well as bad quality dUTP) could be responsible for low strand specificity. The second step in reaching strand specificity is the digestion of the second cDNA strand. If there is no, little or bad dUTP incorporated into the cDNA strand, digestion may be ineffective, leading to unstranded fragments. Another reason for this to happen would be bad uracil-DNA glycosylase.

One obvious reason for low coverage low sequencing depth. The reason for this could be low PCR amplification levels, malfunctions of flow cell lanes, or misconfiguration of the sequencer.

The causes for bias in 5' or 3' coverage likely lies in the library preparation process. Random hexamer priming has been shown to introduce such bias [46]. If the input RNA is degraded, bias could also be introduced due to the mRNA purification process. The process selects for mRNA based on their poly-A tails. If the mRNA is degraded, this leads to more 3' ends being pulled down and subsequently being sequenced.

Quality of qcscanb Dataset

As for the `validation` and `scanb` datasets, plots were generated for all metrics described in the *RNA-SeQC Metrics* section. Plots for the `qcscanb` metrics identified as important in the *Quality Control* section are going to be shown and discussed in detail. Each plots shows one metric for all downsamples of the included samples, with the downsamples on the x-axis from least reads to most reads. Each sample has a unique combination of marker and color. Since the data and plots for the 2x100 and 2x50 read length data are comparable for these metrics, only 2x50 plots are included here.

The plots show 16 different downsampling steps for 44 libraries. The lowest number of reads is 1 million, the highest number is 80 million. Not all downsampling steps have been performed on all libraries.

Metrics of Downsample Data

Figures 5a, 5d, 6a, 6c and 6d show the total filtered reads, unique mapped reads, mapped read duplications, mean coefficient of variation and percentage of transcript coverage gaps, respectively. All of these plots show smooth, monotonic curves. The total filtered reads and mapped read duplication increase with increasing number of reads per downsample, while the mean CV and the percentage of transcript gaps decrease. Figures 5b, 5c and 6b show percentage of failed reads, percentage of total mapped reads and strand specificity, respectively. All three plots show constant values with increasing number of reads per downsample. All of these metrics show consistent and expected behavior when following them over several random sampling steps.

Figures 7a and 7b show the coverage of transcript 5' and 3' ends, respectively. While the plots generally show the expected behavior of increasing coverage with increased number of reads per downsample, they are very uneven. Especially the 3' ends show noisy curves for the downsamples of several libraries.

All in all, the metrics show the downsampling process works and the metrics follow the expected behavior. Considering this, the 3' mapping reads and to a lesser extend the 5' mapping reads do not seem to be distributed completely randomly in the sequence library files. The plots suggest a small bias during the sampling process.

Quality Control Metrics in the Cancer Context

The RNA-SeQC quality control package contains many metrics useful for assessing a dataset's characteristics. However, cancer cells have some unique characteristics which decrease the usefulness of some metrics. Due to cancer being a genomic disease, cancer cells can have a vastly different genomic landscape compared to normal cells. The main processes which shape this landscape are structural rearrangements, copy number changes and point mutations.

Structural rearrangements are caused by chromosome breakage which leads to the rearrangement of chromosome parts in a modified order. This causes problems for metrics which make use of a transcript model, since reads may no longer map in the genomic coordinates that are expected from the transcript model. Examples for this are read coverage, coverage gaps and consequently the mean covariance across transcripts.

Copy number changes are caused by the deletion or amplification of chromosomal parts or even whole chromosomes. A genomic sequence where reads usually map to uniquely may be amplified, leading to an increased number of alternative alignments for these

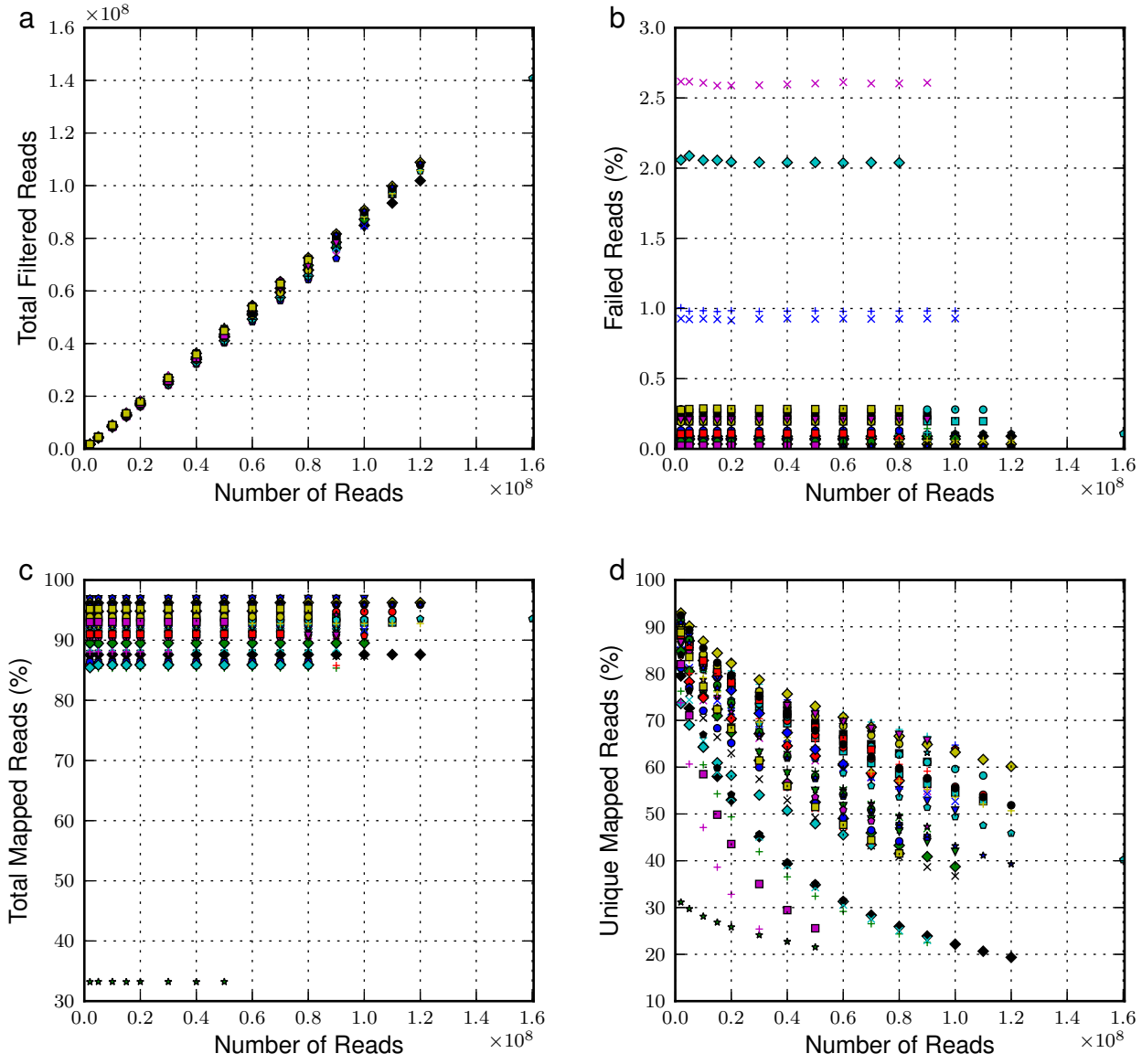

Figure 5: **a**) Total filtered reads, **b**) failed reads in percent, **c**) total mapped reads in percent, **d**) unique mapped reads in percent.

reads. Conversely, a region may have been deleted, leading to reads not being mapped at all.

Point mutations are changes of single bases within the genome. Their effect on quality metrics is likely less severe than that of structural rearrangements and copy number changes. However, they could still cause reads to fail to align correctly to the reference genome. The effects of SNPs are similar to those of point mutations. However, SNPs are present to approximately equal amounts in any cell, making their effect on metrics unbiased.

Other mutations, i.e. insertions and deletions ("indels"), can have drastic effects on the alignment of reads to the reference genome, particularly as they get larger.

The described effects on metrics are neither caused by problems during library preparation, nor by sequencing errors or mapping errors. They are merely an artifact of the unique genomic landscape in cancer cells. Therefore, metrics from cancer cell

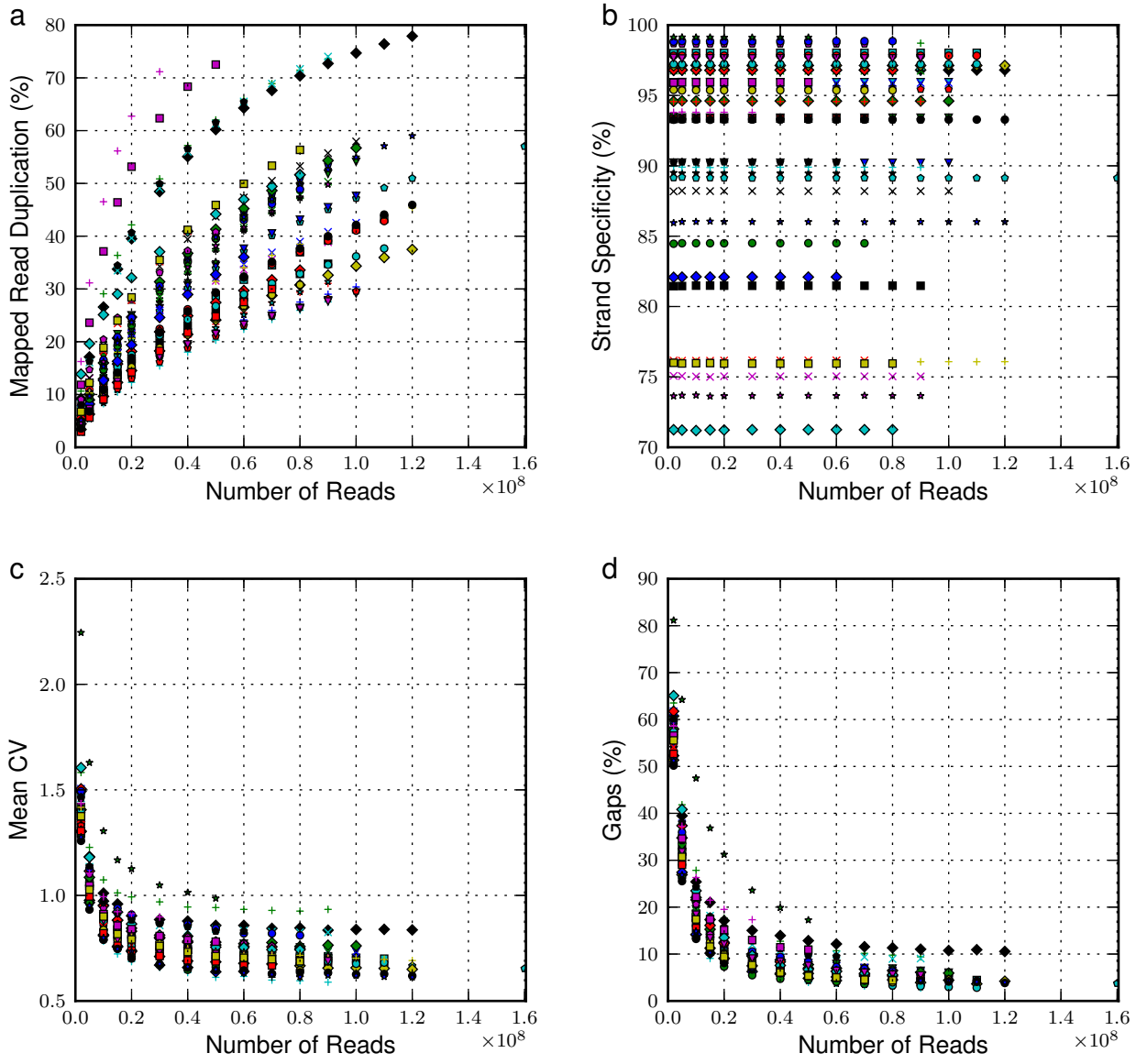

Figure 6: **a**) Mapped read duplication, **b**) strand specificity in percent, **c**) mean coefficient of variation, **d**) transcript gaps in percent.

data always has to be interpreted with its context in mind.

Alignment and Quality Control of Levin et al. Saccharomyces cerevisiae Data

The raw data Levin et al. [20] used for their sequencing library comparison consists of *Saccharomyces cerevisiae* RNA-seq reads which available in short read archive (SRA) format under accession number SRR059176. To enable comparisons between the quality metrics gathered during this work, it was desirable to run the RNA-SeQC software on this data. The SRA archive does not include reference genome mapping information, nor could this information be obtained from the original authors.

Therefore, the reads were aligned to the UCSC sacCer2 genome [34] using TopHat 2.0.8. During this process, a 86.5% of reads could be aligned. The number reported by Levin et al. was 86.61%, making

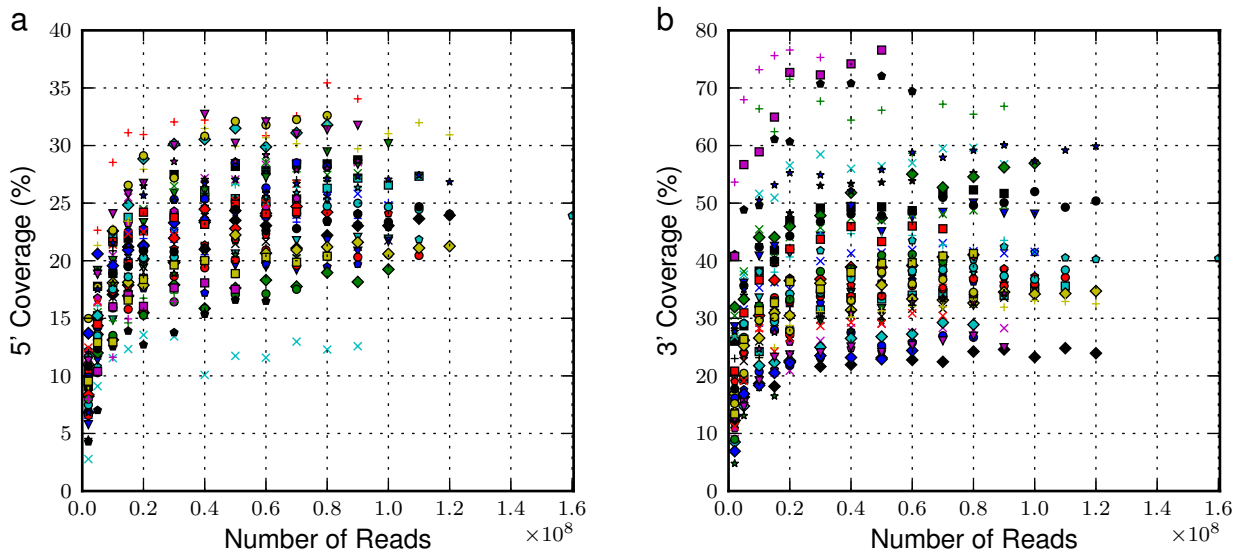

Figure 7: **a**) Coverage at 5' and **b**) coverage at 3' transcript ends.

the alignment comparable. The same preprocessing as for the SCAN-B was used to make the data file compatible with RNA-SeQC.

Unfortunately, RNA-SeQC currently crashes when running it with the resulting BAM file as input, thus no comparisons could be performed as of this writing.

Conclusions

Quality control is an essential step in the RNA-seq analysis process. Considering the complexity of the library preparation, sequencing and reference genome mapping, there are numerous possibilities of introducing errors or biases.

The quality evaluation showed the overall quality of the already sequenced SCAN-B data to be good. While the quality is not optimal in the first few sequenced plates, it has continuously increased and settled in on a high level. Some quality deficiencies correlate well with the presence of recorded problems which occurred during sequencing, i.e. malfunctioning flow cells or lanes.

The quality metrics for the downsampled data showed that the random sampling during the downsampling process works as expected. Metrics correlate well with the number of reads. The plots of 5' and especially 3' coverage showed noisier curves than expected, suggesting bias in the distribution of 5' and 3' mapping reads in the data. This should be investigated further.

Quality metrics have to be interpreted with the datasets in mind from which they were generated from. Cancer cells feature a genomic landscape which is vastly different from that of normal cells. Structural rearrangements, copy number changes and point mutations can lead to seemingly bad quality metrics, even though the RNA-seq analysis process is not at fault.

With this in mind, the metrics used in this project appear to provide good insight into the quality of RNA-seq datasets from tumor samples. The metrics described by Levin et al. [20] cover the quality of the prepared library, while the percentage of failed reads and the mapping rate provide insight into the sequencing process and the mapping of reads onto the reference genome. Overall, the results correlate well with observations by the laboratory staff during the sequencing process and the increasing level of training and experience of the staff.

Acknowledgements

I would like to thank my thesis supervisor Lao Saal, as well as Christof Winter, Jari Häkkinnen and Johan Vallon-Christersson at the Department of Oncology of Lund University. During the course of my thesis project they taught me a lot and patiently answered any questions I had. Thank you also to the SCAN-B meeting group for interesting discussions, Martin Lauss for letting me use his compute node to increase the quality control processing throughput as well as the Translational Oncogenomics Unit, Sofia Gruvberger-Saal and Jill Howlin for their support.

References

1. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860–921.
2. Venter C, et al.: **The Sequence of the Human Genome**. *Science* 2011, **291**(5507):1304–1351.
3. Hanahan D, Weinberg RA: **Hallmarks of Cancer: The Next Generation**. *Cell* 2011, **144**(5):646–674.
4. Malone KE, Daling JR, Thompson JD, O'Brien CA, Francisco LV, Ostrander EA: **BRCA1 Mutations and Breast Cancer in the General PopulationAnalyses in Women Before Age 35 Years and in Women Before Age 45 Years With First-Degree Family History**. *JAMA* 1998, **279**(12):922–929.
5. Apostolou P, Fostira F: **Hereditary Breast Cancer: The Era of New Susceptibility Genes**. *Biomed Res Int* 2013.
6. Visvanathan K, et al.: **American Society of Clinical Oncology Clinical Practice Guideline Update on the Use of Pharmacologic Interventions Including Tamoxifen, Raloxifene, and Aromatase Inhibition for Breast Cancer Risk Reduction**. *Journal of Clinical Oncology* 2009, **27**(19):3235–3258.
7. Dunnwald LK, Rossing MA, Li CI: **Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients**. *Breast Cancer Research* 2007, **9**.
8. Perou CM, et al.: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**(6797):747–752.
9. Sørlie T, et al.: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *Proc Natl Acad Sci U S A* 2001, **98**(19):10869–10874.
10. Schnitt SJ: **Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy** 2010.
11. American Cancer Society Guide to Cancer Drugs: *http://www.cancer.org/treatment/treatmentsandsideeffects/guidetocancerdrugs*.
12. Socialstyrelsen: **Cancer Incidence in Sweden 2011** 2012.
13. Coleman MP, et al.: **Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data**. *The Lancet* 2010, **377**(9760):127–138.
14. Brenner H, Hakulinen T: **Very-Long-Term Survival Rates of Patients With Cancer**. *Journal of Clinical Oncology* 2002, **20**(21):4405–4409.
15. Masood S: **Focusing on Breast Cancer Overdiagnosis and Overtreatment: The Promise of Molecular Medicine**. *Breast J.* 2013, **19**(2):127–129.
16. The SCAN-B Initiative: *http://scan.bmc.lu.se/index.php/Main_Page*.
17. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature Reviews Genetics* 2009, **10**:57–63.
18. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM: **Direct RNA sequencing**. *Nature* 2009, **461**(7265):814–818.
19. Ozsolak F, Milos PM: **Single-molecule direct RNA sequencing without cDNA synthesis**. *Wiley Interdisciplinary Reviews: RNA* 2011, **2**(4):565–570.
20. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A: **Comprehensive comparative analysis of strand-specific RNA sequencing methods**. *Nat Methods* 2010, **7**(9):709–15.
21. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A: **Transcriptome analysis by strand-specific sequencing of complementary DNA**. *Nucleic Acids Res* 2009, **37**(18):e123.
22. Illumina, Inc: *http://www.illumina.com*.
23. Illumina: **Illumina Sequencing Technology**. *http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf* 2010.
24. Metzker ML: **Sequencing technologies - the next generation**. *Nature Reviews Genetics* 2010, **11**:31–46.
25. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV: **The challenges of sequencing by synthesis**. *Nature Biotechnology* 2009, **27**(11):1012–1023.
26. Illumina HiSeq Comparison: *http://www.illumina.com/systems/hiseq_comparison.ilmn*.
27. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment**. *Genome Research* 1998, **8**(3):175–185.
28. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Research* 1998, **8**(3):186–194.
29. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105–1111.

30. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biology* 2013, **14**(4).

31. 1000 Genomes Project: *http://www.1000genomes.org*.

32. Li H: **The missing human sequences (version 5)**. *ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.slides.pdf* 2011.

33. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, and LP: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nature Biotechnology* 2010, **28**(5):511–515.

34. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool**. *Nucleic Acids Research* 2004, **32**(suppl 1):493–496.

35. Wang L, Wang S, Li W: **RSeQC: quality control of RNA-seq experiments**. *Bioinformatics* 2012, **28**(16):2184–2185.

36. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G: **RNA-SeQC: RNA-seq metrics for quality control and process optimization**. *Bioinformatics* 2012, **28**(11):1530–1532.

37. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A: **Qualimap: evaluating next-generation sequencing alignment data**. *Bioinformatics* 2012, **28**(20):2678–2679.

38. Picard: *http://picard.sourceforge.net/*.

39. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Research* 2010, **20**(9):1297–1303.

40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078–2079.

41. Python Programming Language: *http://www.python.org*.

42. PySAM: *http://code.google.com/p/pysam/*.

43. Fixup script for TopHat unmapped read files: *https://github.com/cbrueffer/misc_bioinf/blob/master/fix_tophat_unmapped_reads.py*.

44. SEQanswers forum thread regarding TopHat unmapped files: *http://seqanswers.com/forums/showthread.php?t=28155*.

45. Hunter JD: **Matplotlib: A 2D graphics environment**. *Computing In Science & Engineering* 2007, **9**(3):90–95.

46. Hansen KD, Brenner SE, Dudoit S: **Biases in Illumina transcriptome sequencing cause by random hexamer priming**. *Nucleic Acids Research* 2010, **38**(12).