From the Department for Cell and Molecular Biology
and the Ludwig Institute for Cancer Research
Karolinska Institutet, Stockholm, Sweden

PROGRESSION OF RNA-SEQUENCING TO SINGLE-CELL APPLICATIONS

Daniel Ramsköld

Stockholm 2014

Cover art: Genetic “prayer flags” with cDNA sequences for marker genes for diseases, e.g. *HTT* (in Huntington’s disease). By Joe Davis at MIT, with permission.

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Åtta.45 Tryckeri

ISBN 978-91-7549-554-5

Progression of RNA-sequencing to single-cell applications

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Daniel Ramsköld

Principal Supervisor:
Docent Rickard Sandberg
Karolinska Institutet
Department of Cell and Molecular Biology
and Ludwig Institute for Cancer Research

Co-supervisor:
Dr. Qiaolin Deng
Karolinska Institutet
Ludwig Institute for Cancer Research

Opponent:
Dr. Aviv Regev
Massachusetts Institute of Technology
Department of Biology and Broad Institute

Examination Board:
Docent Lars Feuk
Uppsala University
Department of Immunology, Genetics and Pathology

Professor Jussi Taipale
Karolinska Institutet
Department of Biosciences and Nutrition

Professor Camilla Sjögren
Karolinska Institutet
Department of Cell and Molecular Biology

The defense of the thesis takes place in the CMB lecture hall (Berzelius väg 21, Solna) on Wednesday 28 May 2014, at 13:00.

ABSTRACT

New methods enable new discoveries. My time as a PhD student has run in parallel with the maturation of the RNA-seq method, and I have used it to discover basic properties of gene expression and transcriptomes. My part has been bioinformatics – the computer analysis of biological data.

RNA-seq quantifies gene expression for all genes in one experiment, allowing discoveries without prior knowledge, as opposed to single-gene hypothesis testing. When I started my PhD, this was done by microarray followed by qRT-PCR validation, which can be arduous. In contrast to microarrays, RNA-seq quantifies expression with little ambiguity of which gene each expression value corresponds to, and in absolute terms. But at the time, data analysis of RNA-seq was full of unknowns and there were little software available. Nowadays, partly the result of my work, the data analysis is much less complicated, and RNA-seq can be performed on diminutive samples, down to single cells, which was not viable using microarrays.

My first study (Paper I) used one of the very first RNA-seq datasets to study general features of transcriptomes, such as mean mRNA length (~1,500 nt) and the number of genes expressed per tissue (~13,000). I also found special features of some tissues: the liver transcriptome is dominated by a few highly expressed gene, brain expresses especially long mRNAs and testis expresses many more genes than other tissues.

Following this tissue RNA-seq study, I evaluated a new library preparation method for single-cell RNA-seq (Paper III), developed before the prevalence of single-cell RNA-seq. I used technical replicates to show that the method was accurate and reliable for the more highly expressed genes at single-cell RNA levels, and with input RNA amounts corresponding to >50 cells it produced as good quality data as bulk RNA-seq. Then the method was applied on melanoma cells isolated from human blood, and I listed surface antigen genes that distinguished these circulating tumour cells from other cells in the blood.

This single-cell RNA-seq method was then applied on pre-implantation embryo cells (Paper IV). Using first-generation crosses between two mouse strains, I could separate the expression from the maternal and the paternal copies of the genes. I found that 12-24% of the genes express only one of their two copies in any given cell, in a random manner that affects almost all the expressed genes. I also found that the two copies are expressed independently from each other.

Finally, I studied Sox transcription factors during neural development (Paper II), combining RNA-seq and microarray data for different cell types with ChIP-seq data for transcription factor binding and histone modifications. I found that Sox proteins bind to the enhancers active in the stem cells where the Sox proteins are active, but also to enhancers specific to subsequent cells in development. I also found that different Sox factors bind to much the same enhancers, and that they can induce histone modifications.

In conclusion, my work has advanced the RNA-seq method and increased the understanding of transcriptional regulation and output.

SAMMANFATTNING PÅ SVENSKA

Nya metoder möjliggör nya upptäckter. Min tid som doktorand har gått parallellt med mognaden av metoden RNA-sekvensering, och jag har använt den för att upptäcka grundläggande egenskaper hos valet av vilka gener som används i en cell och populationen av de RNA-molekyler som gener gör upphov till när de är aktiva.

RNA-sekvensering mäter aktiviteten för alla gener i ett experiment, vilket möjliggör upptäckter utan förkunskaper, till skillnad från hypotesprövning med enskilda gener. När jag startade som doktorand gjordes detta genom tekniken microarray. Man fick sedan följa upp med tekniken qRT-PCR, eftersom microarrayer inte var tillförlitliga nog. RNA-sekvensering behöver inte detta. Däremot var data-analysen av RNA-sekvenseringsdata full av oklarheter och det var ont om metoder och datorprogram. Delvis på grund av mitt arbete är data-analysen numera betydligt mindre komplicerat, och dessutom kan RNA-sekvensering utföras på mycket små prover, ner till enstaka celler.

Denna avhandling sammanfattar de fyra huvudprojekt jag arbetade på under min doktorandtid. I det första studerade jag antalet gener olika vävnader använder och om de likheter som finns mellan vävnader i gen-användning. Nästa arbete beskriver en metod för att använda RNA-sekvensering på enskilda celler. Ett tredje arbete använder denna encellsmetod för att studera slumpmässighet i gen-aktivitet, där jag såg att gener ganska ofta bara använder bara den maternella eller bara den paternella kopian av genen i en cell. Det sista arbetet handlar om steget före en gen är uttryckt, då proteiner binder till DNA en lång bit ifrån de gener de ska påverka; här fann jag att en grupp sådana proteiner (Sox) binder ett bra tag i förväg innan genen ska bli aktiv under embryonalutvecklingen.

Sammanfattningsvis har mitt arbete avancerat RNA-sekvenseringsmetoden och ökat förståelsen för valet av genaktivering inom celler.

LIST OF SCIENTIFIC PAPERS

I. **Daniel Ramsköld**, Eric T Wang, Christopher B Burge, Rickard Sandberg.
An abundance of ubiquitously expressed genes revealed by tissue transciptome sequence data.
PLoS Computational Biology. 5: e1000598. 2009.

II. Maria Bergsland*, **Daniel Ramsköld***, Cécile Zaouter, Susanne Klum, Rickard Sandberg, Jonas Muhr.
Sequentially acting Sox transcription factors in neural lineage development.
Genes and Development. 25: 2453-2464. 2011.

III. **Daniel Ramsköld***, Shujun Luo*, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faradini, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, Rickard Sandberg.
Full-length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells.
Nature Biotechnology. 30: 777-782. 2012.

IV. Qiaolin Deng*, **Daniel Ramsköld***, Björn Reinius, Rickard Sandberg.
Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells.
Science. 343: 193-196. 2014.

* these authors contributed equally

ADDITIONAL PUBLICATIONS

Other publications during doctoral studies (not included in the thesis)

- **Daniel Ramsköld***, Erşen Kavak*, Rickard Sandberg.
  How to analyze gene expression using RNA-sequencing data.
  Methods in Molecular Biology. 802: 259-271. 2012. (protocol)
- Shaobo Jin*, Anders P Mutvei*, Indira V Chivukula, Emma R Andersson, **Daniel Ramsköld**, Rickard Sandberg, Kian Leong Lee, Pauliina Kronqvist, Veronika Mamaeva, Päivi Östling, John-Patrick Mpindi, Olli Kallioniemi, Isabella Screpanti, Lorenz Poellinger, Cecilia Sahlgren, Urban Lendahl.
  Non-canonical Notch signaling activates IL-6/JAK/STAT signaling in breast tumor cells and is controlled by p53 and IKKα/IKKβ.
  Oncogene. 32: 2892-4902. 2012.
- Helena Storvall, **Daniel Ramsköld**, Rickard Sandberg.
  Efficient and comprehensive representation of uniqueness for next-generation sequencing by minimum unique length analyses.
  PLoS ONE 8: e53822. 2013.
- Banafsheh Kadkhodaei, Alexandra Alvarsson, Nicoletta Schintu, **Daniel Ramsköld**, Nikolaos Volakakis, Eliza Joodmardi, Takashi Yoshitake, Jan Kehr, Mickael Decressac, Anders Björklund, Rickard Sandberg, Per Svenningsson, Thomas Perlmann.
  Transcription factor Nurr1 maintains fiber integrity and nuclear-encoded mitochondrial gene expression in dopamine neurons.
  Proceedings of the National Academy of Sciences. 110: 2360-2365. 2013.
- Zhi Xiong Chen, Karin Wallis, Stuart M Fell, Veronica R Sobrado, M Charlotte Hemmer, **Daniel Ramsköld**, Ulf Hellman, Rickard Sandberg, Rajappa S Kenchappa, Tommy Martinsson, John I Johnsen, Per Kogner, Susanne Schlisio.
  RNA helicase A is a downstream mediator of KIF1Bβ tumor suppressor function in neuroblastoma.
  Cancer Discovery. CD-13-0362. 2014.

* these authors contributed equally

CONTENTS

LIST OF ABBREVIATIONS

Blast	basic local alignment search tool
bp	base pair
cDNA	complementary deoxyribonucleic acid
CEBPA	CCAAT/enhancer binding protein alpha
ChIP	chromatin immunoprecipitation
DNA	deoxyribonucleic acid
FDR	false discovery rate
FISH	fluorescence in situ hybridisation
Fox	forkhead box
FPKM	fragments per kilobase and million mapped reads
HNF4A	hepatocyte nuclear factor 4 alpha
IGV	integrated genome viewer
mRNA	messenger ribonucleic acid
NCBI	national center for biotechnology information
ng	nanogram
PCR	polymerase chain reaction
pg	picogram
PGM	personal genome machine
PLoS	public library of science
poly(A)	polyadenylate
pre-mRNA	precursor messenger ribonucleic acid
qPCR	quantitative polymerase chain reaction

qRT-PCR	quantitative reverse transcription polymerase chain reaction
RNA	ribonucleic acid
RPKM	reads per kilobase and million mapped reads
SAGE	serial analysis of gene expression
seq	sequencing
SMART	switching mechanism at 5' end of RNA template
SNP	single nucleotide polymorphism
SOLiD	sequencing by oligonucleotide ligation and detection
Sox	sex determining region Y box
Sry	sex determining region Y

INTRODUCTION

Transcription

The different cell types (muscle cells, liver cells etc) of the body share the same DNA and therefore the same set of genes. But the set of genes each cell type actually uses differs, both by its cell type and by its responses to the local environment. A gene needs to copy (transcribe) itself into RNA to have an effect. Apart from being transcribed into RNA or not, the number of RNA molecules can be important since e.g. more RNA for a particular protein can produce more of that protein. One measures RNA amounts to learn if and how much active different genes are, in a particular cell or in a larger group of cells. This can tell which genes are involved in reacting to a stimulus or differ in activity during a disease, as well as providing a signature for each cell type.

Whether a gene should be transcribed or not is controlled by DNA regions called enhancers[1]. Their DNA binds proteins called transcription factors, for which over a thousand exist in human[2]. Not the same transcription factors are present in all cell types and under all external conditions, therefore not all enhancers are active in every cell but only a limited subset. Each gene has many enhancers that control it, located nearby on the DNA. When an enhancer is bound by a sufficient number of proteins, it can come in proximity to the start of the gene by DNA looping. There, it coaxes the protein RNA polymerase into moving down along the gene (RNA polymerase has a molecular motor function), transcribing the gene as it moves[3].

Methods and history of RNA quantification

RNA amounts could be measured already in the 1970s. Northern blot, from 1977[4], is a method that can measure RNA amount by using a radioactively labeled RNA probe[5]. Soon there was also RNA-FISH[6], a method which uses a fluorescently labeled oligonucleotide to make microscopy pictures where RNAs with a particular sequence light up as dots, or as a smear depending on resolution[7]. This method dates from 1982[7], and can both determine if the gene is expressed, where the RNA is located and sometimes even quantify how much RNA there is. Ten years later[8] came qRT-PCR, which converts RNA to cDNA and measures total DNA amount during DNA amplification as the intensity of a DNA-binding dye. Compared to Northern it is faster and more sensitive, the latter a result of including DNA amplification (i.e. PCR)[9]. The first quantitative sequencing-based method, SAGE, came in 1995[10]. Because it used short (9 bp at the beginning in 1995) sequences to identify genes, there were difficulties with ambiguity. The tags were produced by restriction enzymes that cut a set distance from their binding sites. The sequencing itself was done by concatenating the small DNA fragments and running Sanger sequencing[10]. Gene expression microarrays also date from 1995[11] and became the dominant method for quantifying RNA from thousands of genes in an experiment, rather than SAGE. Microarrays typically rely on oligonucleotide probes (DNA that is tens of bases long) to bind fluorescently labeled cDNA, with the oligonucleotides arranged in a pre-determined pattern so that spots will light up at known places when a cDNA for a particular gene is present, with light intensity proportional to the amount of cDNA. The ability of microarrays to measure thousands of genes at the same time brought discoveries that methods that only measure a few genes at a time could not have achieved[12].

Sequencing of fragments of RNA, called expressed sequence tags, has been around for decades, and was used to discover genes[13] rather than for RNA quantification. The history of RNA-sequencing starts when the DNA sequencing machines by Solexa/Illumina allowed sequencing of millions of RNA pieces per sample. Before these machines, sequencing of random RNA fragments was limited to gene discovery, because of cost and low sequence depth[14]. With the new technology, the number of sequenced fragments per gene was dramatically increased such that the counts would correlate with the expression level of the gene to reflect how much RNA from each gene there was in the sample. The first publications came out in May 2008[15,16]. At this time the cost associated with RNA-seq was high, which meant that studies used few or no replicates (that is, preparing the same type of material twice or more and sequencing separately) to control for biological and technical variation. Yet the improvement in data quality, allowing genes to be quantified in absolute terms (e.g. one gene could be determined to be expressed twice as much as another gene), meant it was a technology that could stand competition already from the start, and produce new insights.

DNA sequencing machines

The last few years the rapid decrease in sequencing costs have markedly slowed down[17], and Illumina's technology has taken most of the market[18,19]. I have seen the suggestion that Illumina is like Intel in the microprocessor business, only releasing improvements often enough to stay better than their competitors[20]. In 2008/2009, the SOLiD sequencing system was a close competitor, with a slightly lower price per sequence read but with lower quality and more difficult data analysis (Illumina was slightly better when we did the comparison in spring 2009, due to the large fraction of unmapped reads in SOLiD data). Since then, SOLiD has fallen behind[21]. Nowadays, Ion Torrent's technology is the main competitor to Illumina, and seems pretty close at least for machines (Ion Torrent PGM, Illumina MiSeq) that have lower throughput and high cost per read but higher speed and lower purchase cost[22,23].

The latest development is the Illumina HiSeq X 10, for which Illumina calculates a cost per whole human genome sequence at just below US$1000, at the standard 30x coverage[24]. This particular price per human genome has for long been a goal[25,26]. There was even a prize, later canceled, for the first to reach this cost per human genome, called the Archon X prize[27]. Because of the high initial purchase cost and throughput, the HiSeq X 10 is suited only for large genomics centres[28]. Excluding the HiSeq X 10, the current sequencing cost for a human genome is around US$5000[29].

RNA-sequencing

RNA-seq analysis starts with RNA extraction from the biological sample. For single-cell protocols, this is merely a matter of placing the cell in a lysis solution, whereas for many bulk samples, the cells are dissociated from each other and there is an extra RNA purification step. The next step is library preparation, where the RNA is reverse transcribed into cDNA, which is then fragmented and universal adapter sequences, and DNA barcodes, are added to the ends of each fragment. The fragmentation step uses a separate kit (Nextera transposase) or machine (Covaris sonication) from the rest of the library preparation steps, and causes the fragment ends that will be sequenced to be distributed across the length of the gene and is needed to make the molecules short enough to work with common DNA sequencing machines (e.g. by Illumina or Ion Torrent[30]).

Finally the cDNA library is sequenced, producing millions of so-called reads. A read is a partial (often 50bp) sequence of a cDNA fragment. Because DNA sequencing machines are expensive and thus should be constantly in use, the sequencing step is often done by a core facility or other service provider. One lane gives more than enough reads for an RNA-seq experiment, so several samples (often over ten in a lane) are generally sequenced together. They then need to be separated after sequencing (demultiplexing), using the DNA barcodes that were added during library preparation[31].

After the demultiplexing step, the sequences are aligned to the genome. The genome sequence contains introns whereas the reads mostly originate from spliced RNA. Therefore the alignment step needs to either include gene sequences with introns removed, or the alignment program has to be able to identify when reads cross from one exon to the next. Many such programs are freely and publicly available, e.g. TopHat[32,33], RNA-STAR[34] or GSNAP[35].

Quality control on sequenced samples can be done by a number of methods. The alignment success percent and absolute number of mapped reads I find are the best ways to spot trouble. Running the program FastQC[36] can help discovering overrepresented sequence. Running NCBI Blast[37] on a few reads (if a low fraction of the reads aligned to the genome of the species that should have been sequenced) to align against a broad variety of species can tell you if there is contamination by DNA from a different organism. Because FASTQ files from Illumina's machines are sorted by location on the flow cell, it is a good idea to avoid the start and end of the file, where I have noticed that the error rate is higher[38]. If the reads align at a high frequency, it can next be a good idea to look at a few samples in a genome browser such as IGV[39], to look for e.g. signs of DNA contamination (little mounds of alignment that do not follow genes and differ in location between samples) and alignment/assembly problems (such as regions with many mismatches next to each other). Finally, the fraction of reads aligning to exons, introns and intergenic regions tells you something about pre-mRNA fraction (gives more reads in introns), DNA contamination and alignment error (both these two increase the fraction of reads in intergenic regions).

By counting the number of reads per gene, and knowing the gene length and the sequencing depth (the number of total reads for the library), the alignments can be summarised into one value per gene, using a metric called RPKM or FPKM (reads/fragments per kilobase and million mapped reads). I wrote my own program for RPKM calculation for Paper I, but several others are available. Cuffdiff[40-42] appears to be the most popular one, and performs well for all but the shortest transcripts[43].

Some standard analysis methods to run are hierarchical clustering (I prefer Spearman correlation as metric and have seen the advice to use complete linkage), principal component analysis, and statistical testing for differential expression (DESeq[44] is what I generally use, but several exist[45]). Hierarchical clustering provides additional quality control, as good samples all correlate well with one another since many genes are intrinsically highly or lowly expressed (e.g. ribosomal proteins that are highly expressed in all cell types). Bad samples show up as outliers in the hierarchical clustering.

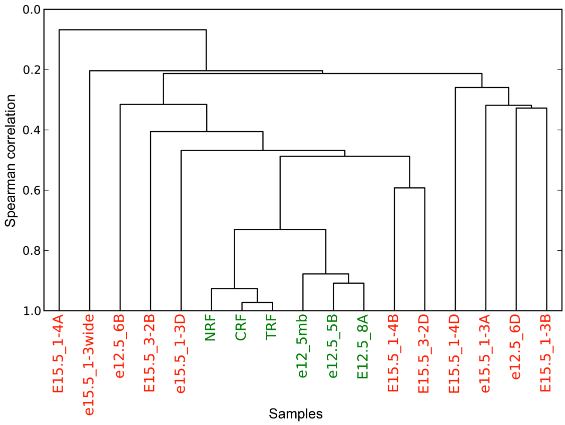


Figure 1. Illustration of hierarchical clustering, which can tell apart low quality samples (red) from better quality samples (green). The RF samples are from fibroblast cell culture, the rest are laser capture microdissected brain samples.

Principal component analysis allows you not only to find clusters of samples, but it also gives you a list of genes, weighted by importance (the U matrix of singular value decomposition[46], if that is the method used), that drive the separation in a particular principal component. It also informs which separations are the strongest ones and which are the less informative ones, as each principal component contains a particular percent of the total variance. And it can show if some of the differences between groups are linearly dependent (they then line up one the same principal component). The main downside is the lack of P-values, making it hard to select clusters or know where to cut off the gene list coming from a principal component. If the first principal component separates the two groups you are interested in comparing, then a differential expression test solves that. It either gives a P-value for each gene, typically adjusted for multiple testing (the Benjamini-Hochberg FDR method[47] is the most common) or it will tell you there are no significant differences, in the case when the clusters were not actually separate clusters.

To identify functional groups among the differentially expressed gene, they typically go into a gene ontology analysis tool, such as DAVID[48] or ToppFun[49]. It is also possible to use the genes with the highest or lowest weights in a principal component, or from fold change rankings, but it is not the standard way. Gene ontology analysis compares the input list of genes against a large number of pre-made gene lists (e.g. a biological function such as transcription, a location such as the mitochondrion or a pathway such as Wnt signaling) and tests for significant overlap between gene lists.

Finally, most projects benefit from data analysis designed for just that one study. Here, programming is important, resulting in a bioinformatician writing a large number of small, single-purpose programs, in languages such as R, python or perl[50]. And this is where most time is spent.

ChIP-sequencing

ChIP-seq is a DNA sequencing-based method to list genomic binding sites for a DNA-binding protein, by chromatin immunoprecipitation (ChIP). The first step is the application of formaldehyde to covalently bind proteins to adjacent DNA (fixation). The cells are then lysed to release their DNA, which is then fragmented by sonication to ~200-400 bp[51]. Some of this DNA is set aside as a negative control, with the fixation reversed. The rest is run through antibody capture for the protein of interest, with non-captured DNA washed away. Thereby DNA bound by that DNA-binding protein will be overrepresented in the sample. The formaldehyde crosslink is reversed and the sample becomes a sequencing library by adding universal adapters to the ends. Then one end of each DNA fragment (but not the adapters themselves) is sequenced.

Alignments for ChIP-seq use simpler alignment programs than RNA-seq, since there is no intron splicing to account for. Bowtie[52,53] has long been the most common program choice[54], bwa[55,56] is also popular. The steps after alignment are peak call and peak annotation. Peak calling algorithm search for "peaks", hill-like regions of read density about the same width as the DNA fragment length (a few hundred base pairs). During peak annotations, those peaks are associated with genes. Generally the closest gene within a set maximum distance is chosen.

A new development for ChIP-seq analysis is the statistical concept *irreproducible discovery rate*, which uses replicates[57]. This is in contrast to the false discovery rate, which gets a distribution from a negative sample[58] or in some cases a simulated random read distribution[59].

RNA-seq and ChIP-seq compared to qPCR alternatives

Microarrays are still used[60,61]. But the main competitor to sequencing-based measurements is quantitative PCR, in the forms of qRT-PCR which quantifies RNA for a single gene, and ChIP-qPCR which quantifies binding of a protein to a specific DNA stretch. In the early days, ChIP-seq and RNA-seq were often validated using ChIP-qPCR and qRT-PCR, respectively. Thankfully this habit seems to be disappearing. The problem is that qRT-PCR only provides technical validation (and only for the sequencing step), typically giving a strong correlation with fold changes from RNA-seq yet without validating biological findings[62] unless it is performed across a larger panel of biological samples.

For single-cell measurements, RNA-seq and qRT-PCR been reported to be equally good[63]. For measuring something lowly expressed in a bulk of much more highly expressed genes, I would trust qRT-PCR to have better sensitivity than RNA-seq, as highly expressed genes eat up the number of fragments that are sequenced in a sequencing run, leaving little sequencing depth for genes that are more lowly expressed by orders of magnitude. This is not a problem for single-cell RNA-seq, or even most tissue RNA-seq, as the dynamic range of expression levels is only so large (3 to 4 orders of magnitude[16]) but would make it hard to sequence from e.g. a few bacteria hidden in a tissue[64].

In a comparison between ChIP-seq and ChIP-qPCR, ChIP-seq has advantages even for single sites. Its genome-wide nature provides it with negative controls in the same experiment, making it much more reliable for finding a lack of binding. Also, ChIP-qPCR has to throw out negative results because they are inconclusive, which I suspect causes a bias (at weakly binding sites) that ChIP-seq would not have. But ChIP-seq suffers from being slow (sequencing queue time) and more expensive. For the purpose they are used – genome-wide discovery for ChIP-seq, single-site testing for ChIP-qPCR – ChIP-seq is said to require better antibodies to get sufficient sensitivity in comparison with ChIP-qPCR.

A sufficient sequencing depth

There has been a tradition for RNA-sequencing to focus on the amount of sequence reads per sample, placing less emphasis on the number of replicates. Early studies did not always use replicates[65], and currently, Encode recommends 30 million reads per RNA-seq sample as a minimum (corresponding to 20-25M mapped reads), but only 2 replicates. Lately the importance of replicates has been emphasized over the sequencing depth for individual samples[66]. My own opinion is based on analyses in papers I and III, where I found that detection (in bulk RNA-seq) and accuracy (in single-cell RNA-seq) of gene expression values from RNA-seq saturates after a few million reads. Thus I believe most studies will not require tens of millions of reads per sample, but should instead focus on increasing the number of replicates.

Single-cell RNA-seq protocols

Currently there are two kits available by different manufacturers (Clontech Smart-seq and Sigma-Aldrich TransPlex). One of them, Smart-seq, has a single cell isolation machine built for it, the Fluidigm $C_{1.}$ There are also several published protocols: Quartz-seq[67], Tang et al. 2010[68], STRT[69], Smart-seq2[70], CEL-seq[71] and MARS-seq[72]. The C_1 + Smart-seq system appears to be the most commonly used one.

Single-cell RNA-seq protocols can use even small amounts of total input RNA, such as 10 pg. RNA-seq library preparation kits for bulk RNA (i.e. for a population of cells) have a much higher input requirement: 500 pg for Nugen Ovation v2, 50 ng for Agilent SureSelect, 100 ng for Illumina TruSeq. As a result, the single-cell RNA-seq protocols are being used not only for single-cell samples, but also for samples with, for example, ~100 cells, which is sufficiently many cells to average out cell-to-cell variation and give a low variation between replicate samples (Paper III, and seminar by Rahul Satija). The single-cell protocols lack the strand-specificity of some bulk RNA-seq protocols, but that is no great loss, while the gain is the ability to sort cells more carefully for the cell type you want to study, or to be able to use small biopsies.

Improvement needs

Early on (2008), issues for RNA-seq included a requirement for large amounts of RNA per sample (as much as for microarrays), lack of strand-specificity and to some extent a 3' bias. The last two were soon solved for bulk RNA-seq, and single-cell RNA-seq has now solved the first one. Early on there were also little published in the way of algorithms and programs for RNA-seq, e.g. there was no purpose-built alignment program for RNA-seq before TopHat[32] in March 2009.

Still today, the RNA sequencing protocols have room for improvement. Reducing the RNA loss – by reducing technical losses and by making them more foolproof against RNA degradation – would make the data easier to interpret and reduce sample variability that obscures biological differences. Adding multiplexing (preparing many samples together), which is done by tagging cDNA fragments with DNA barcodes, could help somewhat to increase the number of samples, but would soon run into the DNA sequencing machine's limit on throughput, which at the moment is increasing rather slowly[17].

On the bioinformatics side, alignment is today the most mature, suffering only from high RAM usage (but with the continuous computer hardware improvements, even laptops will have enough). Aligning two samples for the same tissue to the genome but with different alignment programs (TopHat[32], RNA-STAR[34] and GSNAP[35] are the ones I have tested) gives you strikingly similar results. Quality control and clustering can both be done without much difficulty with current methods. Statistical methods for determining the most significant gene expression differences between two sample groups is the least developed part, and better algorithms are needed. This is especially the case for single-cell data, which is rich in zeros (as expression measurements) both due to biological stochasticity and technical losses. One problem I have encountered is that assigning samples to groups randomly can sometime give significant genes, perhaps due to an outlier value in one sample for a gene that gets called significant. Other issues are biases based on gene expression level as well as big differences between algorithms' outputs. More options for paired experimental setups would also be helpful. But it is hard to say how much of a problem these issues represent, before there RNA-seq studies are attempted to be replicated (in the sense of same question and choice of sample types, but new samples and people analysing them). Microarrays have a history of producing gene sets that differ too much between studies[73], so RNA-seq studies that produce differential expression gene might suffer from a similar problem until the statistical understanding improves.

AIMS

- To determine general features about gene activity in cells, such as "How many genes are needed for housekeeping functions, how many do specialised functions in a cell type?"

- To improve computational analysis of RNA-seq data, e.g. defining detection limits and required sequencing depth

- To understand how gene regulation over developmental time is achieved by transcription factors with similar DNA binding properties.

- To evaluate single-cell RNA-seq, including its accuracy, sensitivity and gene body coverage.

- To clarify aspects of pre-implantation development, in particular the replacement of maternal RNA by embryonic transcription

- To characterise allelic expression patterns in individual cells

RESULTS AND DISCUSSION

An abundance of ubiquitously expressed genes

Findings in Paper I:

- Tissues usually express 11,000-13,000 protein-coding genes
- Most of these expressed genes can be found expressed in any tissue
- RNA-seq has background, at below 0.3 RPKM
- A few million reads are enough to saturate gene detection
- The liver transcriptome is dominated by a few genes, unlike most tissues

Before RNA-seq, the main methods to measure the activity of genes were qPCR and microarrays. Neither qPCR nor microarrays measure absolute expression values, unless a standard curve for each gene is prepared. RNA-seq differs in this aspect: it can, fairly well, tell that e.g. RNA from gene A is twice as abundant as RNA from gene B. Thus the arrival of RNA-seq provided an opportunity to study the ”shape” of the transcriptome. RNA-seq also has the advantage of being able to tell which nucleotides of a gene are expressed, whereas previous methods generally interrogated a pre-selected part of the gene being measured. This allows RNA-seq to tell which parts of the gene structure are being expressed. Nowadays it can also discover genes without help of prior annotation[74], but in 2008 and 2009, the read lengths were shorter and there were fewer assembly programs (and none designed for RNA)[75], which are the programs needed to put together reads outside known genes into new gene structures and tell these reads apart from various types of background.

Paper I deals with questions about the structure of the transcriptome (all the expressed RNA): How many genes are active in a tissue? Which genes are always needed? Why are some messenger RNAs short and others long? In addition it deals with technical questions, such as background level and required sequencing depth.

I counted the number of expressed genes per tissue in Paper I. Microarrays, the only previous genome-wide method, do not have as good sensitivity as RNA-seq. The 200 intensity threshold for detection on Affymetrix arrays corresponds to 3-5 RNA copies per cell[76,77] whereas RNA-seq, as it turned out, could detect expression at 0.1 copies per cell (I calculated in the paper that 0.3 RPKM was an appropriate threshold, Mortazavi et al 2008[16] had found that 3 RPKM in a liver cell, with nearly the same protocol, corresponded to 1 copy per cell). I found that 11,000-

13,000 different (protein-coding) genes were expressed for most of the tissue samples (8,000 of those genes were detected in all the tissues I had RNA-seq data for). Though ~12,000 is not that far from what the most thorough microarray study[78] had found, which was 8,200 on average. I am pleased to see that the numbers I calculated have been used for validation of transcriptome assemblies[79-81]. The list of ubiquitously expressed genes has also proven useful for focusing on tissue-specific genes when testing for differential expression[82].

Paper I also analysed the length of the untranslated regions of mRNAs. Among the tissues, brain had by far the longest untranslated regions. The paper only shows the result for mouse, but it was the same in human. I never included the human samples because these had a problem with RNA degradation, which shortens the mRNAs and which I indirectly measured using the read density ratio between the 3' and the 5' ends of the coding regions of all genes. Tellingly, breast and fat tissue were the samples unaffected by RNA degradation.

There are studies from the 1970s that found a multi-modal distribution of gene expression values[83-86], with two or three quite small groups of highly expressed genes giving distinct peaks in the gene expression distribution. I mentioned in Paper I that these groups were not present in the RNA-seq data, but it was all a mono-modal distribution. Those groups of highly expressed genes have never been found again (they were probably created by bad maths). However, a study published in 2011[87] argues that the bump in the low end of the distribution, which I had dismissed as non-expressed background from alignment or DNA contamination, were actually transcribed genes, though non-functional. That would be a reason for a biological, rather than technical, gene expression cut-off, located at somewhere 1-5 RPKM. However, I have seen the bimodal distribution with a low-expression group under 1 transcript per cell (first commented on in a email by Quin Wills) even for single-cell RNA-sequencing data (of papers III and IV), and heard from others (Sten Linnarsson) about a similar bump in their expression level distributions, so it does seems to be a technical issue.

The most common question I get on this paper has been how to calculate a threshold RPKM level by the method in Paper I. It seems that providing the algorithm as text and formulas only got people partway to being able to implementing it (though some papers seem to have reimplemented the algorithm without further help than the paper). Sending them code did work. But it means fewer will have tried the algorithm then would have otherwise tried. Hopefully the less short-handed descriptions in Papers II-IV (their methods sections can be as long as the rest of the paper combined) reduces this issue of difficulty of reuse that Paper I had.

Sequentially acting Sox transcription factors

Findings in Paper II:

- Sox2, Sox3 and Sox11 bind to mostly to the same enhancers
- Sox proteins can bind enhancers long before the enhancer becomes active
- Sox3 can induce histone modifications (histone 3 lysine 4 and 27 trimethylation)

The Sox gene family of transcription factors has expanded during vertebrate evolution, to 20 genes in human and mouse, which can be categorized into 9 groups, e.g. the SoxB1 group is Sox1, Sox2 and Sox3[88]. The first reported Sox gene, Sry, was identified by its role in sex-determination. The rest of the family was than classified as Sox proteins by containing the same, well-conserved DNA-binding domain as Sry. Other than the DNA-binding domain, the groups are not similar and appear to have domains that function differently outside the DNA-binding domain.

Plenty of the Sox proteins are involved in brain development. The weakly activating SoxB1 are active early in development, and function to maintain neural stem and progenitor cells, which are the self-renewing cell types of the brain and spinal cord. SoxB2 proteins (Sox14, Sox21) are also present in these cells, but are repressive transcription factors, counteracting the role of the SoxB1 genes. As cells differentiate, SoxC (Sox4, Sox11, Sox12) proteins become active in differentiating neurons, and SoxE (Sox8, Sox9, Sox10) proteins are important in the astrocytes and oligodendrocyte lineages, which have supporting roles in the brain.

Paper II investigated how Sox proteins perform their role during brain and spinal cord development, by looking at the sites in the mouse genome where Sox2, Sox3 and Sox11 bind in the relevant cell types, which were neural progenitor cells for Sox2 and Sox3, and early neurons for Sox11. There was already a published dataset for Sox2 in mouse embryonic stem cells, which I included in the analysis.

Excluding the embryonic stem cell dataset, the binding of Sox2, Sox3 and Sox11 overlapped very well, with essentially no independent binding sites for Sox2 and Sox11, and overlap at 70% of the Sox3 sites (pretty much the same Sox3 sites had Sox2 and Sox11 binding, so there is a 30% subset of Sox3-specific sites, though that could be the higher quality of the Sox3 dataset). The high overlap with Sox11 was not an expected result. The cell type was not the same, and Sox11 has a rather different function, driving differentiation instead of inhibiting differentiation as the SoxB1 genes do when over-expressed. But it did fit with ideas about Sox proteins as transcription factors that bind early to keep a place on the DNA available for later binding by another transcription factor, such as another Sox protein, in a later cell type. For example, Sox2 binding in embryonic stem cells, at one enhancer, had been shown to allow later Foxd3 binding[89], and some Fox transcription factors have been shown to keep chromatin open for other Fox proteins[90].

Paper II contains both microarray, RNA-seq and ChIP-seq data, but not the array equivalent of ChIP-seq, which is chip-on-chip. The project started with its own microarray data, and later added microarray data from gene expression omnibus, a database of published microarrays. However, due to quality concerns, we later added RNA-seq expression datasets of our own. The reason was quality. Microarrays get signals from genes mixed together and have a low dynamic range, producing problems for highly and lowly expressed genes. And RNA-seq was not much more expensive at the time (the difference has since been reduced further). For the choice between ChIP-seq and chip-on-chip, the advantages of the sequencing technique are much greater[91]. In 2008 chip-on-chip was already a technique on the way out, as it either gives up the discovery potential, by limiting probes to a set of sites, or is very expensive, if the entire genome is covered, and still has worse sensitivity and resolution.

Midway through the course of the project, I heard the results of a study[92] that compared the DNA binding of two proteins (CEBPA and HNF4A) in five different species, which concluded that very few of the binding sites are conserved during evolution. Only 7-14% of binding sites were the same between human and mouse, and nothing says Sox proteins would have more conserved binding. Their finding was backed up by a previous paper of theirs[93] which performed ChIP-seq (for 3 evolutionary unrelated transcription factors: HNF1A, HNF4A, HNF6) in a mouse with an added human chromosome 21. This study[93] found that the human chromosome bound the transcription factors as they do in people, whereas the homologous mouse regions (which consist mainly of chromosome 16) bound them differently.

This means the sites we find may not be that useful for human studies, as they will not be at the homologous locations. Yet the principles for how Sox genes function, should nonetheless hold true in other species such as mouse, and can probably be generalized to other transcription factor families.

Some ChIP-seq samples were lower-quality, and gave a thousand or so peaks by de novo peak calling. They were still helpful, because I could test binding sites known from better-quality ChIP-seqs, tens of thousands of them, on the data from the worse-quality ChIP-seq. To do so, I adopted the signal/background comparison method of paper I, using either peak-sized regions near the peaks, or read counts from the input as the negative set. Though with a low sequencing depth (only millions of reads) it could be hard to be sure that regions that seem negative really are that.

mRNA-seq from single cell levels of RNA

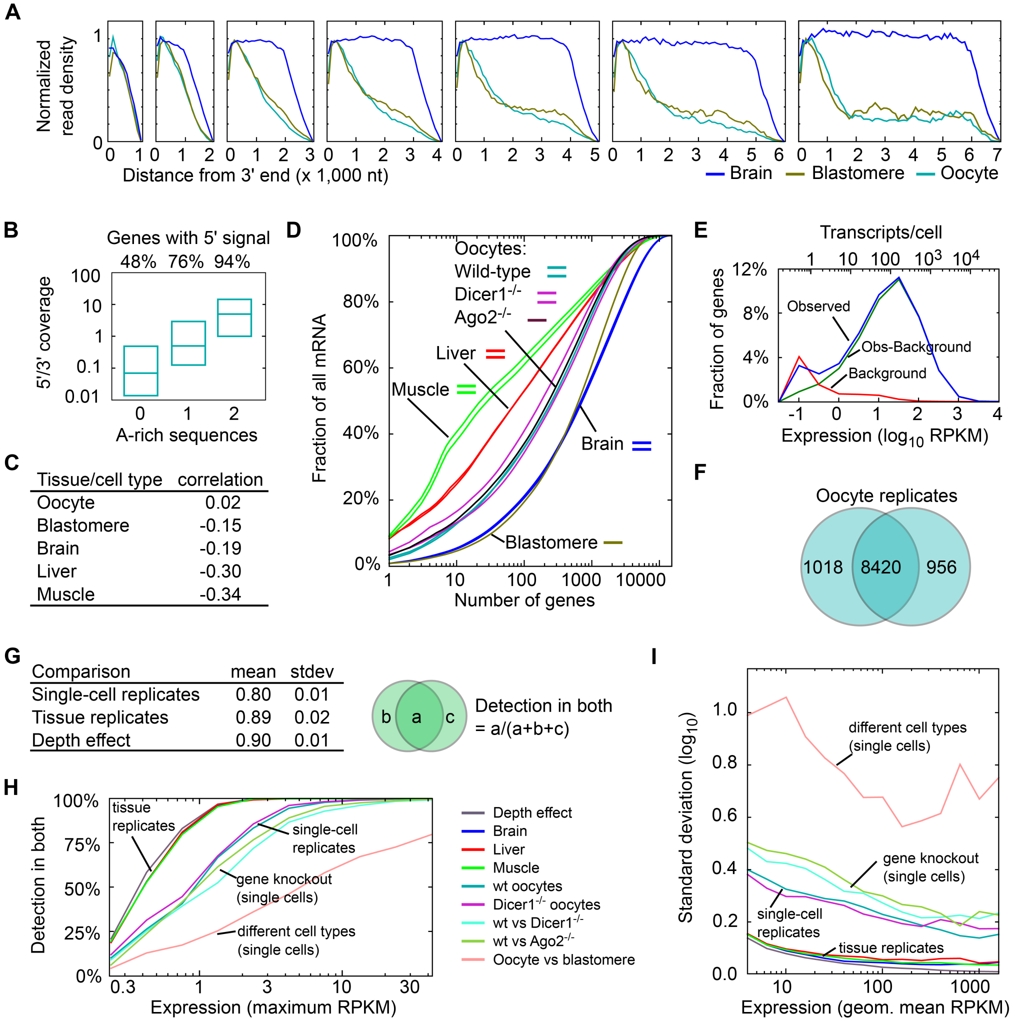
Findings in Paper III:

- This is a new RNA-seq library preparation method that works even for RNA from only one cell
- With medium amounts of input RNA (1 ng) this method works as good as bulk RNA-seq
- Reads from this method are distributed across the length of each gene
- Isolation by the marker NG2 can find cells in the blood migrating from a melanoma

While my two first papers are based on samples with at least hundreds of thousands of cells, the aim was to do a study on pre-implantation development (which is the first week of pregnancy in human), where the number of cells is rather more limited, between one and about a hundred cells per embryo. So I would need single-cell RNA-sequencing, and therefore my PhD project plan included that I take part in single-cell RNA-sequencing methods development.

The first single-cell RNA-seq dataset came in April 2009[94], using a method that was a slightly modified version of a single-cell microarray protocol[95]. The cell was an oocyte, which is much larger than normal cells (~50 times) but still its ~1 ng of total RNA is much less than the recommended minimum of 100 ng in Illumina's standard RNA-seq protocol called TruSeq (though people seem to get those to work with just tens of nanograms). Rickard Sandberg (my PhD supervisor) had a single-cell RNA-sequencing project on pre-implantation mouse embryos in mind since he started in 2008, so this was exciting. I evaluated the oocyte data to see if the protocol had worked well, and concluded that it did. One thing deviated from the description in the paper[94]: it rarely produced full-length cDNA, but mostly <500 bp, and sometimes when it seemed to have done so, it was actually priming on poly(A) stretches within the RNA. It was nonetheless exiting news that an apparently working single-cell RNA-seq protocol was around. Both I and a master thesis student in our group, Sigrid Karstorp, tried the protocol[95] but it would not work; neither of us got amplified cDNA out of it. Meanwhile, I tried to get my re-analysis of the Tang et al. 2009[94] data published, but failed (rejected by two journals, then we gave up). It later formed the nucleus of Paper III.

A year later, near the end of 2010, Rickard Sandberg showed my single-cell RNA-seq data re-analysis to Gary Schroth at the company Illumina. As a result, our lab was invited to beta test a new protocol for single-cell RNA-seq that Illumina was developing, and we were sent data from test runs of their protocol, which I could quickly analyse because I already had the tools developed. In the end we set up collaboration, where Illumina supplied their test data, and we were allowed to analyse it essentially without interference from them and with the option to publish what we found.



Tissue/cell type	correlation
Oocyte	0.02
Blastomere	-0.15
Brain	-0.19
Liver	-0.30
Muscle	-0.34

Comparison	mean	stdev
Single-cell replicates	0.80	0.01
Tissue replicates	0.89	0.02
Depth effect	0.90	0.01

Figure 2. This single figure was prepared but never published for my re-analysis of the Tang et al. 2009 single-cell RNA-seq data. (**A**) Read density across the length of genes. The first plot represent genes 0-1,000 bp long, the second 1,000-2,000 bp etc. (**B**) 3' bias as a function of poly(A) sequences in the gene. (**C**) Spearman correlation between the gene length and expression for all genes expressed >1 RPKM. (**D**) The contribution of the most expressed genes to the total number of mRNA molecules, where the x axis show the cumulative number of genes sorted by expression. (**E**) Gene expression in genes and background regions in the blastomere sample. (**F**) Venn diagram showing the number of genes detected (>0.2 RPKM) in oocytes, and the concurrence of replicates. (**G**) The agreement ("detection in both") between single-cell or tissue replicates, at fixed sequence depth. (**H**) The fraction of genes detected in two samples as a function of gene expression level. (**I**) Variation in expression level estimates using between two samples as a function of gene expression.

Paper III presents a new method for RNA-sequencing from small amounts of RNA, down to RNA from single cells, and data analysis of how well the method works. For example we compared how high the technical noise level was at different gene-expression levels in comparison to biological differences, and we looked at how far along the length of the RNA there was cDNA formed that was then sequenced. To show that the method works on a biologically/medically relevant sample, we included an analysis of melanoma cells which had migrated into the bloodstream of a person with cancer, and I generated a list of genes for surface antigens that were specific to some degree to theses melanoma cells.

Manufacturer's descriptions of their protocols can be understated, presumably not wanting to disappoint customers who would try the protocol at the limit of its ability, at least RNA input amounts are conservative. Scientists on the other hand can be a bit too optimistic about method performance, or maybe that is a result of pressure from journal editors. We are aware that the description “full-length” was a bit of an overstatement and it would have been more accurate to call it “nearly full-length”, since the reverse transcriptase does not perfectly extend to full-length cDNA, missing the end for 60% of the time. The title and the associated news&views article[96] proclaims “full-length” as the main thing, pushed by our expectations of journal opinion and by one reviewer, even though our own conclusion was rather that, based on that the data quality and the kits that worked for Qiaolin Deng in our group right off, it was the first single-cell RNA-seq protocol which was worth using.

The main factor for sample quality of single-cell cDNA libraries is how much of the starting RNA contributes to the sequenced cDNA, i.e. yield. Paper III did not include the spike-in RNA controls (bacterial RNA added in known amounts to the sample before reverse transcription and amplification) needed to calculate the yield. One recent paper[97] calculated 10% as its Smart-seq yield per starting RNA, although with quite a bit of variation between individual RNA spikes. Based on the two least abundant spikes in Paper IV, I calculated a ~15% yield, which is in good agreement. It is a fairly low number, and as a result, both lowly and medium (<100 RPKM) expressed genes suffered from lots of technical variation. A few people in our group, Simone Picelli in particular, tried to further develop the Smart-seq protocol[70,98] and got the yield up, to 40% (see calculation in Paper IV figure S26 and the data in its figure 3A).

One type of graphical tool that I developed turned out to be especially useful, for Paper III and for other RNA-seq protocol development projects (including Picelli et al. 2013[98] in the group. It bins genes by RPKM expression number, and shows a “detection” percent for each expression bin of genes. It is calculated on a pair of samples, as the fraction of genes detected in either sample that are detected in both. If there are more than two, they are matched into pairs and the average is shown Because the technical dropout of genes is a function of cDNA molecule number, and RPKM implicitly includes a normalisation for the total amount of cDNA, the line in the plot shifts to the right (more cDNA) or left (less cDNA) depending on how many molecules there are reads from, which is a function of yield and input RNA amount.

The Smart-seq protocol/kit (SMARTer Ultra Low Input RNA Kit for Illumina Sequencing) is, as far as my February 2014 literature search can tell, the most popular single-cell RNA-seq protocol. I can spot 9 papers, excluding technical comparisons and evaluations, that use it. It help that this kit was well tested before release, as evidenced by sending it out for beta testing. Another factor is the Fluidigm C_1 machine. The C_1 is a microfluidics system that takes a single cell suspension, gets cells into separate wells (96 wells per run) and performs Smart-seq library preparation within the machine, i.e. it automates the Smart-seq procedure.

Dynamic, random monoallelic gene expression

Findings in Paper IV:

- Randomness in transcription causes 12-24% of the genes to only be represented in the transcriptome by one of two gene copies
- The two copies of a gene are independently transcribed from one another
- The paternal X chromosome is initially activated in 2-cell stage embryos, but soon gets inactivated
- By the four-cell embryonic stage, there is little remaining maternal RNA

To some extent, Paper IV is about embryonic development during the first few days, where our plan was to learn more about the degradation of RNA that is left from the oocyte and by what time cells begin to become different cell types. But in particular, this paper deals with an unexpected phenomenon we observed, that in the individual cell there will not always be RNA from both copies of each gene, but sometimes only from one, in a way that is random, rapidly changing and affecting virtually all expressed genes. There are several mechanisms that can underlie this observation. One is transcriptional bursting, i.e. that genes are transcribed into several RNA at a time and then is silent for a long time[99]. Another mechanism is inherited random monoallelic expression, where a randomly chosen copy of a gene is the only one expressed over many cell divisions, as many as 15 divisions[100]. This type of monoallelic expression was recently shown to be rarer in embryonic stem cells than elsewhere[101], perhaps because whatever the changes to chromatin and transcription factor binding that causes it, those are changing at a faster pace in embryonic stem cells and its related cell types during pre-implantation development. I did see a few percent of the monoallelicly expressed genes (over expected) with a coordinated, same allelic choice as their neighbour genes (Paper IV figure S4E), which suggests that some of the monoallelic expression was driven by the same thing as whatever causes inherited monoallelic expression. But mostly what I observed must be a consequence of transcriptional bursting, and this implies that the burst frequencies are low compared to RNA degradation rates, causing both monoallelic expression and, for something like 1% of the genes (based on 20% monoallelic expression: (1-0.1)·(1-0.1)-0.1·0.1=0.2, 0.1·0.1=0.01) that "should" be expressed, loss of all gene expression. But there is also a third suggested mechanism[102]: that replication-induced supercoiling at the promoter can randomly hinder or aid transcription of each allele.

Half or more of the genes appeared to have monoallelic expression in each of our Smart-seq single-cell samples. But from a simulation of molecule loss (Paper IV figure S4D) I could see that while some of the monoallelic expression must be real, a great part was a technical artifact. A key experiment (demanded by a reviewer and soon thereafter published in a slightly different, ten cell version[97]) was to split the contents of a lysed cell into two tubes and prepare and sequence two separate samples. While the ten cell version of it was inconclusive[97], with the single split cell version I could make an algorithm (Paper IV figure S26) that looks at how often the two samples say the same thing and from that calculate loss frequency and what the monoallelic fraction of gene expression was before those losses.

The implications of random monoallelic expression lie in the understanding of the effects of heterozygous mutations and SNPs. The same heterozygous genotype at a locus can have effects that vary much between individuals (variable expressivity) or randomly gives one of two phenotypes (penetrance). Perhaps it is because development takes a slightly different path depending on which allele of a gene is expressed in a critical cell and developmental time point.

One of the greatest challenges turned out to be to get a good list of genetic differences between the two mouse strains we were using. The genomes for both strains had been sequenced (one of them, for C57BL/6, is the mouse reference genome), and there was a database where I could get a list of genetic differences between two chosen strains. But the SNPs (single base pair differences, pronounced snips) need to be trustworthy. If I list a position as A in the strain C57BL/6 and G in the strain CAST/Ei, but that position is A in both, then I will get false detection of the C57BL/6 allele when the CAST/Ei allele at that position is expressed. This type of error turned out to be very common in the database, presumably due to sequencing errors in the CAST/Ei genome that were interpreted as SNPs. As a result, the initial read assignment to alleles got it wrong about half the time (some of my samples were from a single strain background, these acted as controls and gave error rates). I first derived my own set from the pure CAST/Ei samples and the reference genome, but eventually we used the database SNPs but filtered them for SNPs that were "heterozygous" in the sum of our samples. A sign that it worked is the similar number of genes with only CAST/Ei or only C57BL/6 allele expression (errors would give a C57BL/6 bias), visible in the figures in Paper IV that show maternal and paternal monoallelic as separate bars.

The paper ended up with much less focus on embryonic development than originally planned. I did determine e.g. where the cells started to become different from the other cells in the same embryo (between 8-cell and 16-cell stage), but that time point was not a great surprise[103]. The data is publicly available though[104] for anyone who wants to know which genes are expressed when during mouse pre-implantation development.

Future perspectives

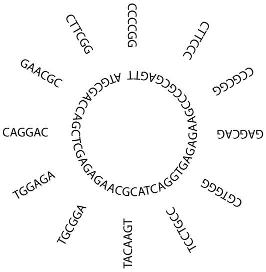
The transcriptional dynamics is something you need to have an idea about to make use of single-cell transcriptome data, for example to differentiate between on-off changes and changes from one expression level to the other, or to understand what level of variation is technical, biological within a cell type, or indicative of several cell types. But it is as a tool for a wide range of exploratory biological studies I see the use of this kind of measurements in the near future. Methods that use measure on the RNA from a large number of cells miss some features of the sample: the on-off dynamics of genes[105], and cell type composition. They are also cumbersome to use on specific cell types, requiring cell culture and limiting to what extent cells can be selected for a particular sub-population.

Single-cell RNA-seq will be useful particularly for cell type discovery. One strategy would be to dissociate a tissue into cells and perform single-cell RNA-seq on many of them, and then find clusters among their gene expression profiles[105]. Another strategy would be to narrow down cell types using other gene expression profiling methods that can only measure a small number of genes at a times, and then use single-cell RNA-seq to verify that the cell-type is homogeneous and different from other cell types that are sampled at the same time, or equally important to find that this is not the case and that the presumed cell type may have subtypes or unexpected similarity to another cell type.

For medical diagnostics, the use of RNA-sequencing is still far off – microarrays are so far only used for one test (MammaPrint)[106], so there is a considerable lag compared to biological research. However, hospitals are likely to procure DNA sequencing machines (the kind with moderate throughput but fast speed, like the Illumina MiSeq) in the near future, for testing for mutations and aneuploidy. Thus RNA-seq does not suffer from the same equipment problem that makes microarrays slow, this should help.

In vitro fertilisation is a field that is unusually open to innovation and trying out new techniques. Because RNA is an amplification step away from DNA, and much of it is conveniently tagged with a common sequence at one end (the poly(A) tail) that makes primer binding easy, RNA sequencing has some advantages over the more direct DNA sequencing for mutations. And it measures the outcome of gene regulation, in a way that looking and the regulatory DNA elements will not help with (the knowledge about them is way too small).

Cheers for reading this far!



ACKNOWLEDGEMENTS

Plenty of people to thank, since doing research all on your own wouldn't work so well.

To **Rickard** who is a great teacher, for example making me learn python and RNA-seq analysis in a week or two. And who might have had a hand in a few projects...

To **Erşen** for helping me learn statistics and for feedback on projects, and to **Jun** for teaching me Linux skills.

To **Qiaolin** for being great to work with. And **Ilgar** and **Helena** too.

To **Maria** and **Jonas** because I loved that Sox project, and for plenty of good ideas and leadership from both of you. Though occasionally **Cécile** was the one to ask for theory and for her sense of detail.

To **Ingemar** for teaching me about interdisciplinary work and a basic understanding of life and science.

To the researchers at **Illumina** for generously handing over data.

To **Qiaolin** again for teaching me the practical aspects of working with mice and to **Björn** for explaining its magical workings. To **Johannes** too. And to **Maria**, **Lizzy** and **Alex** for donating adorable mice.

To **Rickard** and **Lizzy** for giving me the opportunity to teach at courses, which was awesome.

And to the rest of the **Sandberg group**, all the people in the **Stockholm Ludwig Institute** and to the **Lendahl group**, **Nino** included.

More than a few people helped style-checking/fact-checking the thesis summary: Rickard, Björn, Omid, Helena, Qiaolin, Ilgar, Gösta, Daniel Edsgärd and Hagey, Tanya, Julianna and Louise.

I was located first in the Karolinska Institutet Department of Cell and Molecular Biology and then at the Ludwig Institute for Cancer Research (Stockholm). I was funded first by a Karolinska Institutet KID-grant and then by the European Research Council.

REFERENCES

1. Walters, M. C. *et al.* Enhancers increase the probability but not the level of gene expression. *Proceedings of the National Academy of Sciences* **92,** 7125–7129 (1995).
2. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics* **10,** 252–263 (2009).
3. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152,** 1237–1251 (2013).
4. Alwine, J. C., Kemp, D. J. & Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences* **74,** 5350–5354 (1977).
5. Gene Expression. Wikipedia. https://en.wikipedia.org/w/index.php?title=Gene_expression&oldid=596880257.
6. Levsky, J. M. & Singer, R. H. Fluorescence in situ hybridization: past, present and future. *Journal of Cell Science* **116,** 2833–2838 (2003).
7. Singer, R. H. & Ward, D. C. Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog. *Proceedings of the National Academy of Sciences* **79,** 7331–7335 (1982).
8. Weis, J. H., Tan, S. S., Martin, B. K. & Wittwer, C. T. Detection of rare mRNAs via quantitative RT-PCR. *Trends in Genetics* **8,** 263–264 (1992).
9. Freeman, W. M., Walker, S. J. & Vrana, K. E. Quantitative RT-PCR: pitfalls and potential. *BioTechniques* **26,** 112–122, 124–125 (1999).
10. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270,** 484–487 (1995).
11. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270,** 467–470 (1995).
12. Flintoft, L. Milestone 21 (1995) The microarray revolution. Nature milestones 1 December 2005. http://www.nature.com/milestones/geneexpression/milestones/articles/milegene21.html
13. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651–1656 (1991).
14. Bainbridge, M. N. *et al.* Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7,** 246 (2006).
15. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* **5,** 613–619 (2008).
16. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
17. Wetterstrand KA. DNA sequencing costs: data from the NHGRI genome sequencing program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed 1 April 2014.
18. Toner, B. In sequence survey: Illumina holds two-thirds of sequencing market, splits desktop share with Ion PGM. 2 October 2012. http://www.genomeweb.com/sequencing/sequence-survey-illumina-holds-two-thirds-sequencing-market-splits-desktop-share.

19. Nisen, M. *How Illumina's gene sequencing technology could transform health care*. Business Insider. 17 October 2013.

20. GenoMax, *New thing?* Seqanswers 15 January 2014. http://seqanswers.com/forums/showthread.php?t=39890&page=2.

21. Liu, L. *et al.* Comparison of next-generation sequencing systems. *BioMed Research International* **2012,** (2012).

22. Herper, M. *Analyst: the better desktop DNA sequencer may be losing the marketing war*. Forbes 15 August 2012.

23. Seqanswers. *Ion Torrent PGM vs Illumina MiSeq* http://seqanswers.com/forums/showthread.php?t=19432.

24. *Illumina announces the thousand dollar genome - Bio-IT World*.

25. Jacob, H. *The $1,000 genome is here*. MIT technology review. 18 February 2014.

26. Meeting webcast, Beyond the beginning: the future of genomics. 12–14 December 2001. http://www.genome.gov/10001294.

27. Archon genomics X prize, http://genomics.xprize.org/.

28. Check Hayden, E. Is the $1,000 genome for real? *Nature* (2014).

29. Check Hayden, E. Technology: The $1,000 genome. *Nature* **507,** 294–295 (2014).

30. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13,** 341 (2012).

31. lllumina website. http://www.illumina.com/technology/multiplexing_sequencing_assay.ilmn accessed 17 Mar 2014.

32. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

33. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14,** R36 (2013).

34. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

35. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26,** 873–881 (2010).

36. FastQC, Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

37. NCBI Blast, http://blast.ncbi.nlm.nih.gov/Blast.cgi.

38. Yang, X. *et al.* HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics* **14,** 33 (2013).

39. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29,** 24–26 (2011).

40. Trapnell, C. *et al.* Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology* **28,** 511–515 (2010).

41. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7,** 562–578 (2012).

42. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology* **31,** 46–53 (2013).

43. Storvall, H., Ramsköld, D. & Sandberg, R. Efficient and comprehensive representation of uniqueness for next-generation sequencing by minimum unique length analyses. *PLoS ONE* **8,** e53822 (2013).

44. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11,** R106 (2010).

45. Seyednasrollah, F., Laiho, A. & Elo, L. L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics* bbt086 (2013).

46. Hendler, R. W. & Shrager, R. I. Deconvolutions based on singular value decomposition and the pseudoinverse: a guide for beginners. *Journal of Biochemical and Biophysical Methods* **28,** 1–33 (1994).

47. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57,** 289–300 (1995).

48. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2008).

49. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research* **37,** W305–W311 (2009).

50. Napolitano, F., Mariani-Costantini, R. & Tagliaferri, R. Bioinformatic pipelines in Python with Leaf. *BMC Bioinformatics* **14,** 201 (2013).

51. Schmidt, D. *et al.* ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions. *Methods* **48,** 240–248 (2009).

52. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10,** R25 (2009).

53. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9,** 357–359 (2012).

54. Lindner, R. & Friedel, C. C. A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq. *PLoS ONE* **7,** e52403 (2012).

55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

56. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

57. Bailey, T. *et al.* Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol* **9,** e1003326 (2013).

58. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9,** 1–9 (2008).

59. Jothi, R. *SISSRs manual*. 25 November 2008, http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs/SISSRs-Manual.pdf.

60. Masterson, L. *et al.* Gene expression differences predict treatment outcome of merkel cell carcinoma patients. *Journal of Skin Cancer* **2014,** (2014).

61. Sikora, M. J. *et al.* Invasive lobular carcinoma cell lines are characterized by unique estrogen-mediated gene expression patterns and altered tamoxifen response. *Cancer Res.* **74,** 1463–1474 (2014).

62. Fang, Z. & Cui, X. Design and validation issues in RNA-seq experiments. *Briefings in Bioinformatics* **12,** 280–287 (2011).

63. Chi, K. R. Singled out for sequencing. *Nature Methods* **11,** 13–17 (2014).

64. Matsuda, K., Tsuji, H., Asahara, T., Kado, Y. & Nomoto, K. Sensitive quantitative detection of commensal bacteria by rRNA-targeted reverse transcription-PCR. *Applied and Environmental Microbiology* **73,** 32–39 (2007).

65. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).
66. Hansen, K. D., Wu, Z., Irizarry, R. A. & Leek, J. T. Sequencing technology does not eliminate biological variability. *Nature Biotechnology* **29,** 572–573 (2011).
67. Sasagawa, Y. *et al.* Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome biology* **14,** R31 (2013).
68. Tang, F. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nature Protocols* **5,** 516–535 (2010).
69. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research* (2011).
70. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9,** 171–181 (2014).
71. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports* **2,** 666–673 (2012).
72. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343,** 776–779 (2014).
73. Zhang, M. *et al.* Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics* **24,** 2057–2063 (2008).
74. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29,** 644–652 (2011).
75. Zhang, W. *et al.* A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS ONE* **6,** e17915 (2011).
76. Su, A. I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences* **99,** 4465–4470 (2002).
77. Xinmin, L., Kim, J., Zhou, J., Gu, W. & Quigg, R. Use of signal thresholds to determine significant changes in microarray data analyses. *Genetics and Molecular Biology* **28,** 191–200 (2004).
78. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences* **101,** 6062–6067 (2004).
79. Spradling, K. D., Glenn, J. P., Garcia, R., Shade, R. E. & Cox, L. A. The baboon kidney transcriptome: analysis of transcript sequence, splice variants, and abundance. *PLoS ONE* **8,** e57563 (2013).
80. Esteve-Codina, A. *et al.* Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. *BMC Genomics* **12,** 552 (2011).
81. Guo, S. *et al.* Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. *BMC Genomics* **11,** 384 (2010).
82. Hackett, N. R. *et al.* RNA-Seq quantification of the human small airway epithelium transcriptome. *BMC Genomics* **13,** 82 (2012).
83. Hastie, N. D. & Bishop, J. O. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **9,** 761–774 (1976).
84. Bishop, J. O., Morton, J. G., Rosbash, M. & Richardson, M. Three abundance classes in HeLa cell messenger RNA. *Nature* **250,** 199–204 (1974).
85. Quinlan, T. J. *et al.* The concept of mRNA abundance classes: a critical reevaluation. *Nucleic Acids Research* **5,** 1611–1625 (1978).
86. Goldberg, R. B., Hoschek, G., Tam, S. H., Ditta, G. S. & Breidenbach, R. W. Abundance, diversity, and regulation of mRNA sequence sets in soybean embryogenesis. *Developmental Biology* **83,** 201–217 (1981).

87. Hebenstreit, D. *et al.* RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular Systems Biology* **7,** (2011).

88. Guth, S. I. E. & Wegner, M. Having it both ways: Sox protein function between conservation and innovation. *Cellular and Molecular Life Sciences* **65,** 3000–3018 (2008).

89. Liber, D. *et al.* Epigenetic Priming of a pre-B cell-specific enhancer through binding of Sox2 and Foxd3 at the ESC stage. *Cell Stem Cell* **7,** 114–126 (2010).

90. Zaret, K. S., Wandzioch, E., Watts, J. & Xu, J. Pioneer factors, genetic competence, and inductive signaling: programming liver and pancreas progenitors from the endoderm. *Cold Spring Harb Symp Quant Biol (*2008).

91. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10,** 669–680 (2009).

92. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328,** 1036–1040 (2010).

93. Wilson, M. D. *et al.* Species-specific transcription in mice carrying human chromosome 21. *Science* **322,** 434–438 (2008).

94. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6,** 377–382 (2009).

95. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Research* **34,** e42 (2006).

96. Goetz, J. J. & Trimarchi, J. M. Transcriptome sequencing of single cells with Smart-Seq. *Nature Biotechnology* **30,** 763–765 (2012).

97. Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Research* gr.161034.113 (2013).

98. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10,** 1096–1098 (2013).

99. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135,** 216–226 (2008).

100. Gendrel, A.-V. *et al.* Developmental dynamics and disease potential of random monoallelic gene expression. *Developmental Cell* **28,** 366–380 (2014).

101. Eckersley-Maslin, M. A. *et al.* Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Developmental Cell* **28,** 351–365 (2014).

102. Yu, H. & Dröge, P. Replication-induced supercoiling: a neglected DNA transaction regulator? *Trends in Biochemical Sciences* (2014)

103. Johnson, M. H. & Ziomek, C. A. The foundation of two distinct cell lineages within the mouse morula. *Cell* **24,** 71–80 (1981).

104. Gene Expression Omnibus, accession number GSE45719. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45719.

105. Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods* **11,** 22–24 (2014).

106. Gampenrieder, S. P. *et al.* Multi-gene signatures in breast cancer: actual clinical outcome. *memo - Magazine of European Medical Oncology* **7,** 16–21 (2014). http://link.springer.com/article/10.1007/s12254-014-0130-3.