Applications of DNA sequencing in transcriptome studies

Dave Ting Pong Tang

June 22, 2014

Abstract

The mammalian genome encodes both coding and non-coding transcripts that work in concert to orchestrate a wide range of cellular functions. By profiling the complete set of transcripts expressed, i.e., the transcriptome, the dynamics between transcripts can be inferred.

List of publications

1. Sofia Francia, Flavia Michelini, Alka Saxena, Dave Tang, Michiel de Hoon, Viviana Anelli, Marina Mione, Piero Carninci, and Fabrizio dAdda di Fagagna. Site-specific dicer and drosha rna products control the dna-damage response. *Nature*, 488(7410):231–235, 2012

2. Alka Saxena, Dave Tang, and Piero Carninci. pirnas warrant investigation in rett syndrome: An omics perspective. *Disease markers*, 33(5):261–275, 2012

3. Dave T. P. Tang, Charles Plessy, Md Salimullah, Ana Maria Suzuki, Raffaella Calligaris, Stefano Gustincich, and Piero Carninci. Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. *Nucleic Acids Research*, 41(3):e44, 2013

Contents

1 Introduction **7**
- 1.1 History of DNA and genes . . . 8
- 1.2 DNA sequencing . . . 10
  - 1.2.1 Sanger and Maxam-Gilbert sequencing . . . 10
  - 1.2.2 Automated DNA sequencing . . . 11
- 1.3 The human genome . . . 11
- 1.4 Applications of sequencing . . . 13
  - 1.4.1 Transcriptomics . . . 14
  - 1.4.2 Transcription . . . 14
  - 1.4.3 Transcript regulation . . . 16
  - 1.4.4 Characterising the transcriptome . . . 16
  - 1.4.5 Cap Analysis Gene Expression . . . 17
  - 1.4.6 Non-coding RNAs . . . 17
- 1.5 The repetitive genome . . . 18
- 1.6 Stem cells . . . 20
- 1.7 The role of bioinformatics in genomics . . . 22

2 Template switching artifacts **24**

3 Role of small RNAs in DNA damage repair **37**

4 Rett syndrome and piRNAs **46**

5 CCL2 activates hypoxia related genes **62**

6 Regulated expression of repetitive elements **63**

7 The blood transcriptome **64**

8 General discussion **65**
  - 8.0.1 Experimental differences . . . 65
  - 8.0.2 Signal versus noise . . . 65
  - 8.0.3 Coding versus non-coding transcripts . . . 66
  - 8.0.4 Transcriptome profiling using CAGE . . . 66

List of Figures

1.1 The structure of DNA 9
1.2 Developments in next generation sequencing 11
1.3 Condensation of DNA 12
1.4 DNA transcription 15
1.5 Coding probability of FANTOM3 mouse cDNAs 17
1.6 Coverage of repetitive elements in vertebrate genomes 19
1.7 Vertebrate genomes sizes 20
1.8 Tally of repetitive element families in the hg19 genome 21

List of Abbreviations

3D	three-dimensional.
A	adenine.
ASC	Adult stem cell.
BRE	TFIIB recognition element.
C	cytosine.
CAGE	Cap analysis gene expression.
cDNA	Complementary DNA.
ChIP	Chromatin immunoprecipitation.
DNA	Deoxyribonucleic acid.
ENCODE	Encyclopedia Of DNA Elements.
ERV	Endogenous retrovirus.
ESC	Embryonic stem cell.
FANTOM	Functional annotation of the mammalian genome.
G	guanine.
GO	Gene ontology.
HGP	Human Genome Project.
iPSC	Induced pluripotent stem cell.
LINE	Long interspersed elements.
lncRNA	Long non-coding RNA.
LTR	Long terminal repeat.
miRNA	Micro RNA.
mRNA	Messenger RNA.
ncRNA	Non-coding RNA.
piRNA	Piwi-interacting RNA.

Pol I	RNA polymerase I.
Pol II	RNA polymerase II.
Pol III	RNA polymerase III.
RNA	Ribonucleic acid.
RT	Reverse transcriptase.
SAM	Sequence Alignment/Map.
SINE	Short interspersed elements.
T	thymine.
TF	Transcription factor.
TFBS	Transcription factor binding site.
tRNA	Transfer RNA.
TS	Template switching.
TSS	Transcription start site.
U	uracil.

Chapter 1

Introduction

In the winter of 1868/9, Swiss physician and biologist, Johannes Friedrich Miescher isolated an unknown substance from the nuclei of cells[4]. This substance was unlike anything he had observed before; it was resistant to protease, lacked sulphur, and contained large amounts of phosphorous. He recognised that he had isolated a novel substance and as it was isolated from the nucleus, he named it nuclein. Today, we know of this substance as nucleic acid (deoxyribonucleic and ribonucleic acid), the molecule of heredity.

Although Miescher speculated that nuclein may have had a role in heredity, he later rejected the idea. It wasn't until 1944, when Oswald Avery, Colin MacLeod, and Maclyn McCarty characterised the transforming factor in Griffith's Experiment[5] as deoxyribonucleic acid (DNA), that it was hypothesised that DNA was the genetic material[6]. In order to identify DNA as the transforming factor, they used an enzyme that only degraded DNA (deoxyribonucleodepolymerase) and showed that it was able to eliminate the transforming power. It was later confirmed in 1952 by Alfred Hershey and Martha Chase that DNA was indeed the genetic material by performing a series of experiments in bacteriophages[7]. They demonstrated that when bacteriophages infected bacteria, only their DNA would enter into the cytoplasm of the bacteria, while their protein would remain outside. This demonstrated that the genetic material that infected the bacteria was DNA, in a time when the consensus was that protein was the genetic material due to the variety of amino acids.

Fast forward to the 21st century and not only have we sequenced the entire collection of DNA, the genome, of our very own species [8, 9], we have the capacity to sequence an entire human genome in a matter of days by using high-throughput sequencers. We have also just recently arrived in the $1,000 genome era[10], whereby we can sequence the entire genome of an individual for around $1,000 US dollars (USD). In contrast, the Human Genome Project (HGP), which gave us the first glimpse of the human genome costed approximately 2.7 billion fiscal year 1991 US dollars[11]. In roughly 60 years, we have gone from establishing DNA as the molecule of heredity to being able to sequence a human genome in a few days for around 1,000 USD. Developments in molecular genetics are definitely moving at a very fast pace.

High-throughput sequencing is, however, not limited to genome sequencing. The discovery of reverse transcriptase[12, 13] allowed the generation of complementary DNA (cDNA) from an ribonucleic acid (RNA) template; thus we

are able to sequence the entire collection of RNA, known as the transcriptome, from biological samples. Other applications of sequencing has allowed us to capture in a genome-wide manner DNA-protein interactions, DNA methylation patterns, and histone modifications[14]. Many of these applications are named with a "Seq" suffix; for example RNA-Seq refers to RNA sequencing and ChIP-Seq refers to chromatin immunoprecipitation followed by sequencing. There are a large number of *Seq protocols[15], which is set to increase with the decrease in sequencing costs and the development of innovations in molecular biology.

This thesis focuses primarily on the study of transcriptomes using high-throughput sequencing and bioinformatics. While the genome is assumed to be more or less identical in the 37 trillion cells that make up the human body[16], their transcriptomes are markedly different. It is the differential use of the genome that makes it possible to give rise to over 400 different cell types[17] that make up different tissues, organs, and systems. Differential usage may also be indicative of disease; for example, transcriptomes isolated from cancerous cells are quite different from normal cells. In this work, we have studied and compared transcriptomes from various biological samples and systems: in whole blood samples (Chapter 2 and 7), in the FANTOM5 panel of samples (Chapter 6), in a DNA damage response (DDR) system (Chapter 3), in a mouse model for Rett syndrome (Chapter 4), and in human induced pluripotent stem cells (iPSCs) (Chapter 5). Collectively, these studies demonstrate the applicability of transcriptome sequencing and bioinformatics in gaining insight in various biological systems.

1.1 History of DNA and genes

The classical definition of a gene dates back to Gregor Mendel when he demonstrated in his plant breeding experiments that discrete traits could be inherited from parents to offspring. While the mechanism of inheritance wasn't clear at that time, these heritable traits were termed a gene. The definition of a gene has evolved with our increasing knowledge of genetics and biochemistry[18]. The idea that genes were blueprints for proteins was perceived when mutations in Neurospora genes could cause defects in steps in metabolic pathways[19]. This became the "one gene, one polypeptide" hypothesis; the idea was that each gene was responsible for producing a single enzyme/protein in a biochemical pathway. The association of DNA with genes began with Griffith's Experiment[5], where it was demonstrated that heritable traits could be transferred between dead and live bacteria. It was later demonstrated that this heritable material was DNA[6].

DNA was discovered by Johannes Friedrich Miescher who subsequently named it nuclein. In 1881, Albrecht Kossel determined that nuclein was composed of five bases: adenine (A), cytosine (C), guanine (G), thymine (T), and uracil (U) and in 1889, Richard Altmann discovered that nuclein was acidic and renamed it as nucleic acid. The basic component of DNA was deduced by Phoebus Levene in 1909, where he discovered that DNA consisted of an acid, an organic base, and a sugar; he also showed that these components were linked together as phosphate-sugar-base to form units, which he termed nucleotides. This sugar-phosphate backbone forms the structural framework of nucleic acids and make DNA highly stable. While Levene proposed that DNA was made up

equal amounts of A, C, G, and T, it was discovered by Erwin Chargaff that DNA should have a 1:1 ratio of pyrimidine (C, T, and U) and purine (A and G) bases, which is known as Chargaff's rules. Chargaff's rules was one of the observations that led to the eventual deduction of the three-dimensional (3D) structure of DNA.

The 3D structure of DNA showed how adenines paired with thymines and cytosines paired with guanine[20]; this is now known as Watson-Crick base pairing (Figure 1.1). The base pairing explained how genetic information could be copied as they famously wrote in their paper:

> "It has not escaped our attention that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."
>
> — Watson and Crick

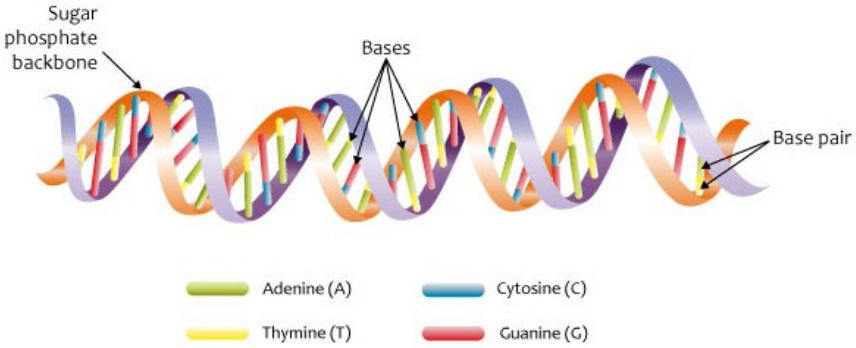

Figure 1.1: The structure of DNA is based on the repeated pattern of a sugar and phosphate group, known as the sugar phosphate backbone, and the base pairing of the four bases, adenine (A), cytosine (C), guanine (G), and thymine (T). Image by NHS National Genetics and Genomics Education Centre licensed under the Creative Commons license.

The relationship between DNA and proteins was demonstrated by the use of artificial ribonucleic acid (RNA) and bacterial cells[21]. RNA is synthesised by RNA polymerase, using DNA as a template, which is a process known as transcription. By using RNA artificially created to be composed of entirely uracils, Matthaei and colleagues produced a protein composed entirely of the amino acid phenylalanine. This experiment demonstrated that nucleic acids contained a code for amino acids. This code, known as the genetic code, was later cracked three years later[22] and defined how information is encoded in the genetic material. Nirenberg and colleauges worked out that three nucleotides defined a codon and is translated into one of the 20 standard amino acids. This flow of information from nucleic acid to protein was described as the "Central Dogma"[23].

At that period of time, a gene was known as a section of DNA that was used in the production a particular protein. In each gene, the nucleotides were

arranged in a specific order to facilitate the production of the corresponding amino acids and eventual protein. However with the advent of DNA sequencing and high-throughput sequencing, the definition of a gene further evolved.

1.2 DNA sequencing

The process of DNA sequencing is determining the exact order of nucleotides within a DNA molecule. The first generation of DNA sequencing methods consisted of the electrophoretic methods (Sanger and Maxam-Gilbert sequencing) and were very labour intensive, requiring 4 separate reactions for the determination of each base. One of the first innovations was the use of different fluorophores with Sanger sequencing, which allowed the reactions to be pooled together during gel electrophoresis and eliminated the use of radioactive material. Automation of the DNA sequencing process was possible with the development an apparatus that could detect the fluorescence emitted by the chain terminated fragments. This development was key towards the success of the HGP. For over 25 years since its inception, Sanger sequencing was the method of choice for DNA sequencing.

The next wave of DNA sequencing methods, the so-called next generation or second generation sequencing, started with the development of pryosequencing and bridge amplification.

1.2.1 Sanger and Maxam-Gilbert sequencing

Two of the earliest protocols for DNA sequencing were Sanger sequencing[24], and Maxam-Gilbert sequencing[25]. These two methods, which were developed in the 1970s, were considered the first generation of DNA sequencing methods. Both methods require the use polyacrylamide gel electrophoresis, which can resolve DNA fragments at a 1 bp resolution, for the determination of the DNA bases. The generation of the DNA fragments is different between the two methods.

The key feature of Sanger sequencing is the use of dideoxynucleotide triphosphates (ddNTPs) and a purified DNA polymerase enzyme to synthesise DNA. The structure of a normal nucleotide (dNTP), consists of a 3' hydroxyl (OH) group in the pentose sugar. The chain-terminating ddNTPs lack the OH group that is necessary for the formation of the phosphodiester bond between one nucleotide and the next during DNA strand elongation. If a ddNTP is incorporated into a growing DNA strand, the strand elongation is terminated. The idea is to set up a reaction with a mixture of dNTPs [deoxyadenosine triphosphate (dATP), deoxyguanosine triphosphate (dGTP), deoxycytidine triphosphate (dCTP), deoxythymidine triphosphate (dTTP)] and one particular ddNTP in a ratio of 300:1. Most of the times, the DNA will be elongated but once the ddNTP is incorporated the strand stops. This results in a number of DNA fragments of varying lengths and depending on the ddNTP used, the last base of the fragment corresponds to that base. This reaction is carried out for the other ddNTPs and all the fragments are separated using polyacrylamide gel electrophoresis. By reading the ladder, the DNA bases can be deduced.

The Maxam-Gilbert sequencing method relies on the use of chemicals that can cleave specific bases. Dimethyl sulfate is used to cleave purine bases (A and

G) and hydrazine is used to cleave pyrimidine bases (C and T). To distinguish the purines, an adenine-enhanced cleavage step is carried out, which cleaves adenines preferentially. To distinguish the pyrimidines, NaCl is used with hydrazine to suppress the reaction of thymines. As with Sanger sequencing, the DNA fragments are separated using polyacrylamide gel electrophoresis, and the DNA bases are deduced by reading the gel.

1.2.2 Automated DNA sequencing

Sanger sequencing became the *de facto* method for DNA sequencing due to its comparative ease and the use of fewer toxic materials than Maxam-Gilbert sequencing. A further improvement to Sanger sequencing replaced the need to radioactively label the DNA fragments by using chemically synthesised fluorescent oligonucleotide primers[26]. Four different fluorophores were used for each ddNTP reaction allowing all four reactions to be co-electrophoresed and the DNA sequence was deduced by reading the fluorescence colours. The development of a fluorescence detection apparatus linked to a computer that processed the data created the world's first partially automated DNA sequencer[26].

(Figure 1.2)

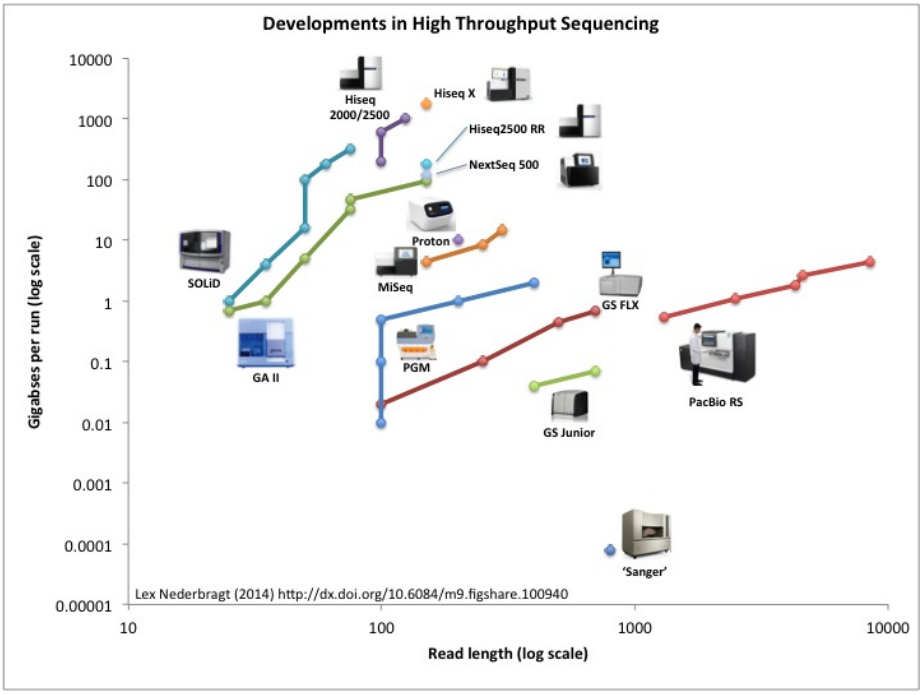

Figure 1.2: Log read length versus log gigabases per run for various next generation sequences[27].

1.3 The human genome

Inside every somatic nucleated cell in the human body are 23 pairs of chromosomes; these chromosomes make up the human genome. A chromosome is a long DNA molecule that is condensed so that it may physically fit inside the nucleus (Figure 1.3). We can estimate the physical length the human genome by multiplying the number of base pairs contained on each chromosome by the length of a base pair. The haploid human genome contains around 3×10^9 base pairs of DNA; therefore there is a total of 6 billion base pairs of DNA per cell. Given that each base pair of DNA is about 0.34 nanometers long or 3.4×10^{-10} meters[28], each diploid cell contains 3.4×10^{-10} * 3×10^9 * 2 or 2.04 meters of DNA. Given that the estimated number of cells in the human body is around 3.7×10^{13} or 37 trillion[16], a typical person contain around 74 trillion meters of DNA. Certain proteins, called histones, help compact the vast amounts of DNA inside of us (specifically, inside the eukaryotic nucleus). The histones are a family of small, positively charged proteins that provide the energy, in the form of electrostatic interactions, to fold negatively charged DNA (due to the phosphate groups in its phosphate-sugar backbone). The resulting DNA-protein complex is called chromatin.

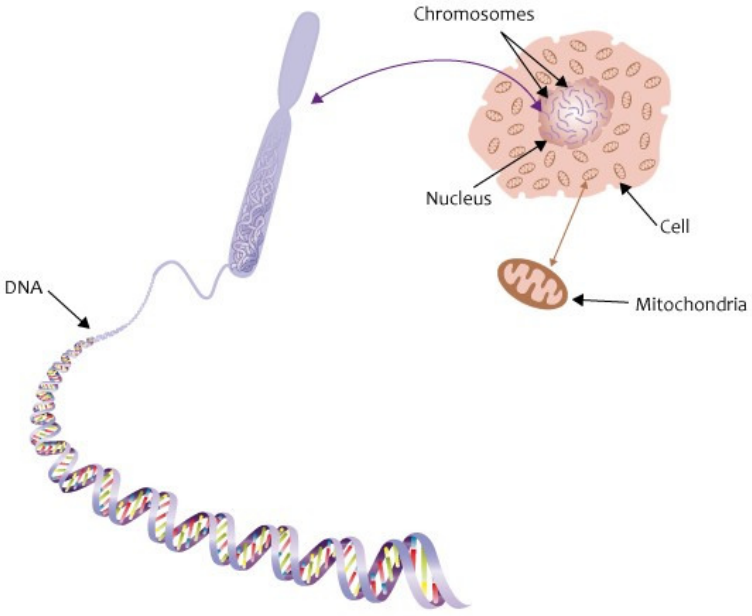

Figure 1.3: DNA is condensed into chromosomes by forming DNA-protein complexes known chromatin, which is further coiled into thicker fibers called 30nm fibers. The chromosomes reside inside the nucleus of a cell. Mitochondria also contain its own DNA. Image by NHS National Genetics and Genomics Education Centre licensed under the Creative Commons license.

Chromatin possesses a fundamental repeating structure[29], known as nucleosomes and the packaging of DNA into nucleosomes shortens DNA about sevenfold. However despite this, chromatin is still too long to fit inside the nucleus, which is approximately 10 to 20 microns in diameter. Chromatin is further coiled into a thicker fiber, called the 30nm fiber, because it is roughly 30 nanometers in diameter. Processes such as transcription and replication require the two strands of DNA to come apart temporarily, thus allowing polymerases access to the DNA template. However, the presence of nucleosomes and the folding of chromatin into 30-nanometer fibers pose barriers to the enzymes that unwind and copy DNA. It is therefore important for cells to have means of opening up chromatin fibers and/or removing histones transiently to permit transcription and replication to proceed. Generally speaking, there are two major mechanisms by which chromatin is made more accessible:

1. Histones can be enzymatically modified by the addition of acetyl, methyl, or phosphate groups.

2. Histones can be displaced by chromatin remodelling complexes, thereby exposing underlying DNA sequences to polymerases and other enzymes.

It is important to remember that these processes are reversible, so modified or remodelled chromatin can be returned to its compact state after transcription and/or replication are complete.

1.4 Applications of sequencing

The age of high throughput sequencing and the application of this technology to studying the transcriptome (also epigenome and interactome).

One of the first technologies that allowed the simultaneous profiling of thousands of transcripts was microarrays and with the development of whole-genome arrays, the whole repertoire of known transcripts could be assayed. However, with the advent of high-throughput sequencers, RNA sequencing (RNA-Seq) became an attractive alternative to microarrays. RNA-Seq is an application of high-throughput sequencing that is used to study RNA expression levels in a discovery-based manner, since it does not rely on any *a priori* knowledge of transcripts. RNA-Seq gives digital gene expression (DGE) profiles, which provides a large dynamic range of expression values and is much more quantitive than microarrays. Furthermore, RNA-Seq does not suffer from cross-hybridisation issues when profiling transcripts from repetitive regions of the genome.

RNA-Seq is relatively new compared to microarray technology, which has been around since 1995, but is gaining more widespread use due to the rapid drop in sequencing costs. As such RNA-Seq protocols are continually being developed and improved upon to avoid any biases. Methods used for storing, analysing, and visualising RNA-Seq data are also being heavily researched upon. In the near future, RNA-Seq may be used as routinely as microarrays for performing transcriptome profiling.

Non-coding RNAs, which include small and long non-coding RNAs, contribute to a significant portion of the transcriptome in mammalian genomes. However, the functional significance of this wide-spread occurrence of ncRNAs

for organismal development and differentiation is unclear. While coding transcripts serve mainly as templates for protein synthesis, the functions of some well known non-coding transcripts are diverse. Furthermore, the major portion of many mammalian genomes are occupied by DNA sequences that are repetitive.

Technologies for the interrogation of nucleic acids have made it possible to investigate different biological questions.

1.4.1 Transcriptomics

The transcriptome is the collection of RNA molecules derived from the genome at a particular time and location. The diversity of cells is determined by the expression patterns of genes and transcription factors (TFs).

1.4.2 Transcription

The genome is a store of biological information that requires the coordinated activity of enzymes and proteins to bring it to life. Transcription is a highly regulated mechanism that processes DNA into a single-stranded RNA molecule and begins with RNA polymerase attaching to the template DNA strand (Figure 1.4). In some cases, the RNA molecule is the functional product, however, the RNA can be further translated into a protein molecule. There are three different types of RNA polymerases in eukaryotic cells: Pol I transcribes the genes that encode most of the ribosomal RNAs (rRNAs); Pol II transcribes the messenger RNAs (mRNAs); and Pol III transcribes the genes for small regulatory RNA molecules, such as transfer RNA (tRNA).

Messenger RNAs are transcripts of protein-coding genes and are translated into protein

RNA polymerase usually binds upstream to the DNA that will be transcribed; this region is known as the promoter. Promoters are classified by their distance to the transcription start site (TSS): the core promoter is the closest and is required for transcription; the proximal promoter is the next closest ($\sim$ 250 bp) and usually contains primary regulatory elements; distal promoters contain additional regulatory elements and is further upstream. Commonly, mRNAs have a TATA box 25 to 35 bases upstream of the TSS to which the RNA polymerase binds to. In addition to RNA polymerase, a number of essential cofactors (known as general transcription factors) are required for transcription, such as TFIIB and TFIID; TFIID recognises the TATA box and TFIIB recognises the TFIIB recognition element (BRE). Transcription with Pol I and Pol III are similar, but the promoter sequences and activator proteins differ.

As transcription takes place, the DNA double helix unwinds and RNA polymerase reads the template strand; the RNA sequence will have the same sequence as the coding strand. The termination of transcription relies on terminator sequences that are found close to the ends of transcripts. For Pol I transcripts, transcription is stopped by a termination factor that unwinds the DNA-RNA hybrid formed during transcription. For Pol III transcripts, inverted repeats are transcribed and these sequences can form hairpin loops that pauses the RNA polymerase. For Pol II transcripts, a polyadenylation signal (AAUAAA) at the end of the sequence is necessary for termination[30]. The nascent RNA is cleaved at the polyadenylation site and a poly-adenine tail

(poly(A)) is added; the poly(A) tail is important for the stability of the RNA, translation, and for nuclear export.

Processing of precursor RNA include the addition of a cap and poly-A tail, splicing, and RNA editing. Removal of the cap is considered to the first step towards mRNA degradation. The capping procedure is thought to occur only in the nucleus, however a cytoplasmic form of a capping enzyme has been identified; the role of cytoplasmic re-capping.

A specialised nucleotide cap is added to the 5' end and this is the site recognised by the ribosomes in protein synthesis. The ribosome binds to the cap and reads along the 5' untranslated region (UTR) until it reaches an initiation codon.

Some introns are removed by specific proteins known as splicing factors, where other introns are removed independently through autocatalysis.

The default state of transcription in eukaryotes is off, meaning that genes are not constantly transcribed. Structural properties of DNA make it inaccessible to the PIC; chromatin structure and nucleosome positioning are altered in order for the transcriptional machinery to access parts of the genome for transcription. Transcription factors (TFs) are regulatory proteins that can activate the transcription of DNA by binding to specific DNA sequences (they are also known to negatively regulate transcription but this is less common). For example, the core promoter elements include the TATA box, the initiator element (Inr), the downstream promoter element (DPE), the TFIIB recognition element (BRE) and the CpG island (CGI). Each of these elements are necessary for transcription; for example the TATA binding protein (TBP) binds to the TATA box and is involved in DNA strand separation during transcription.

1.4.3 Transcript regulation

Control of the frequency and rate of transcription and by the stability of the transcripts and translational efficiency for protein-coding transcripts.

Methylation of specific bases in the DNA.

1.4.4 Characterising the transcriptome

Microarrays [31]

Studies on the yeast transcriptome using DNA microarrays showed that although mRNAs are being degraded and re-synthesised all the time, the composition of the yeast transcriptome undergoes very little change if the environment remains constant[32] However on switching from aerobic to anaerobic respiration, the levels of over 700 mRNAs increase by a factor of two or higher, and another 1,000 mRNAs decline to less than half their original amount Studying the transcriptome of acute lymphoblastic leukemia cells with respect to acute myeloid leukemia cells revealed differences that made it possible to distinguish the two types of leukemia

- Pervasive transcription describes the observation that a large percentage of DNA in mammalian genomes is transcribed.
- Tiling arrays and full length cDNA sequencing suggested that most of the genome is transcribed

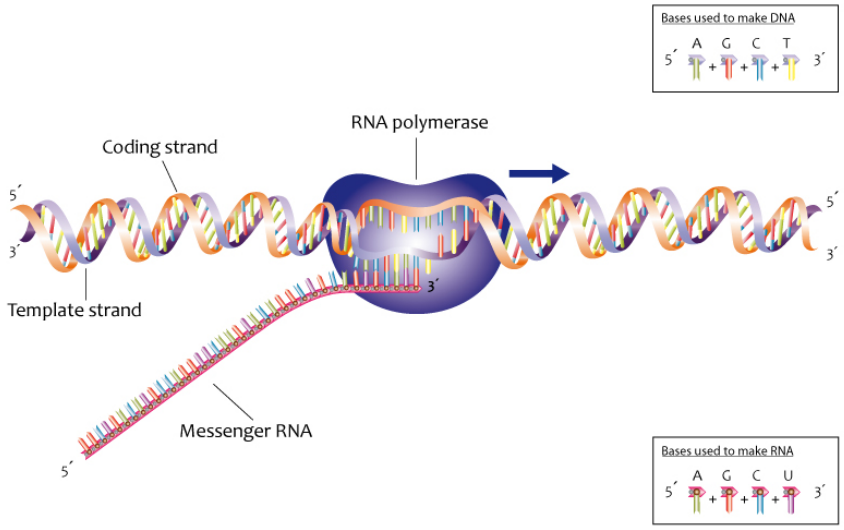

Figure 1.4: The process of transcription begins with RNA polymerase, which reads the DNA sequence (the template strand) in the 5' to 3' direction and produces a complementary RNA strand called the messenger RNA (mRNA). The mRNA has the same sequence as the coding strand, except that thymines are replaced by uracils. Image by NHS National Genetics and Genomics Education Centre licensed under the Creative Commons license.

- The ENCODE pilot project provided multiple lines of evidence that most of the mammalian genome is associated with at least one primary transcript, i.e., pervasive transcription of the genome
- Are the majority of detected low-level transcription due to technical artifacts and/or background biological noise?
- Protocols that capture only poly-A transcripts as opposed to poly-A minus and cytoplasmic versus nuclear enrichment libraries
- Sequencing depth and sampling of RNA molecules; absolute transcript quantification will help (such as using unique molecule identifiers and non-PCR based methods)
- Functional transcriptomics in the post-ENCODE era, specifically what is the criteria for functionality
- Collection of mouse full length cDNAs by the FANTOM consortium revealed many transcripts of unknown function (TUFs)
- 38.6% of the FANTOM3 mouse full-length cDNAs have very low coding potential (CPAT coding probability of less than 0.2) (Figure 1.5)
- The technologies that allowed full length cDNA sequencing

- PacBio sequencing has read lengths greater than 10kb and can be used for full length cDNA sequencing

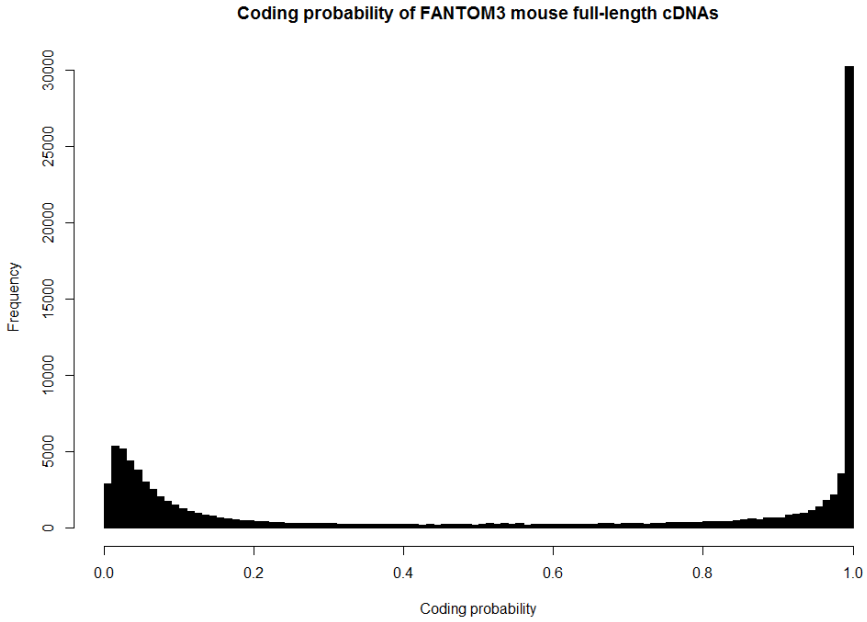

Figure 1.5: Distribution of coding probabilities of FANTOM3 mouse full-length cDNAs as predicted by CPAT[33].

1.4.5 Cap Analysis Gene Expression

Transcription start sites (TSSs) are the nucleotides that are the first to be transcribed by RNA polymerase into a particular RNA. One such technology for capturing TSSs is called Cap Analysis Gene Expression (CAGE), which essentially captures RNA that have a cap structure (specifically biotinylation of diol groups of RNA[34]) and sequences them. The cap is right at the start of a precursor transcript, thus this position represents the TSS. The output of CAGE is a set of short tags that correspond to the capped RNAs, with the number of tags reflecting the abundance of the RNAs.

- The transcriptional landscape of the mammalian genome[35]
- Cap Analysis Gene Expression (CAGE) for genome-wide identification of transcription start sites
- CAGE captures promoter architecture[36]

1.4.6 Non-coding RNAs

Diverse class of non-coding RNAs, broadly broken down into long and short non-coding RNA

Classical non-coding RNA found in both prokaryotes and eukaryotes are ribosomal RNAs and transfer RNAs, which are both involved in protein synthesis

Micro-RNAs (miRNAs) were first first observed in 1993 in Caenorhabditis elegans[37] as a small double-stranded RNA that were bound to the 3' untranslated region (UTR) of an mRNA. The binding of the miRNA inhibited the translation of the mRNA and is one of the mechanisms by which miRNAs post-transcriptionally regulate gene expression. The biogenesis of miRNAs begins with the cleavage by an enzyme called Dicer[38], which is part of the RNase III family. miRNAs inhibit translation by recruiting a ribonucleoprotein complex called RNA-induced silencing complex (RISC). Short-interfering RNAs are also cleaved by Dicer and can also bind to mRNA but require complete complementarity. MiRNAs are products of dsRNAs encoded in genes in our genome and do not require full complementarity to bind to a target mRNA thus one miRNA may regulate several genes.

A single-stranded nucleic acid molecule will tend to fold up on itself to form localised double-stranded regions, producing structures called hairpins or stem-loop structures. This is due to the hydrophobic nature of the bases, which means that the bases are unstable when exposed to an aqueous environment. Pairing of the bases enables them to be removed from interaction with the surrounding water and stabilising the DNA helix.

Assembly of non-coding RNAs with proteins as ribonucleoprotein (RNP) structures Interaction of non-coding RNAs with chromatin

1.5 The repetitive genome

Since the release of the draft human genome sequence[8, 9], it was established that only a small fraction of the genome is made up of protein-coding sequences and the majority of the genome was made up with repetitive elements. As the human genome contains the entire instruction set that is necessary for the development of a human, the decoding of the genome was thought to provide answers to many outstanding biological questions. However, there are several paradoxes that are still currently unresolved:

1. K-value paradox: complexity does not correlate with the number of chromosomes
2. C-value paradox: complexity does not correlate with genome size
3. N-value paradox: complexity does not correlate with the number of protein coding genes

If we measure organismal complexity in terms of mental cognition, we observe that complexity is not correlated to the number of chromosomes, the size of genomes, and the number of protein coding genes. Humans have 46 chromosomes and some species of butterflies have over hundreds of chromosomes, such as *Polyommatus atlantica*. In terms of genome size, the species *Polychaos dubium*, a freshwater amoeboid, has one of the largest genomes known with around 670 gb of DNA sequence; humans on the other hand have a genome size of roughly 3 gb. And lastly humans have roughly the same number of protein-coding genes as *Caenorhabditis elegans*, roughly 20,000 versus 21,000,

respectively. However, the *C. elegans* genome is about 100 mb[39], which is approximately 30 times smaller than the human genome, despite having a similar number of genes. One of the discrepancies between genome sizes despite having a similar number of genes is due to repetitive elements, which makes up roughly 50% of the human genome. Among 66 vertebrate genomes, the percentage coverage of repetitive elements in the human genome is quite high (Figure 1.6).

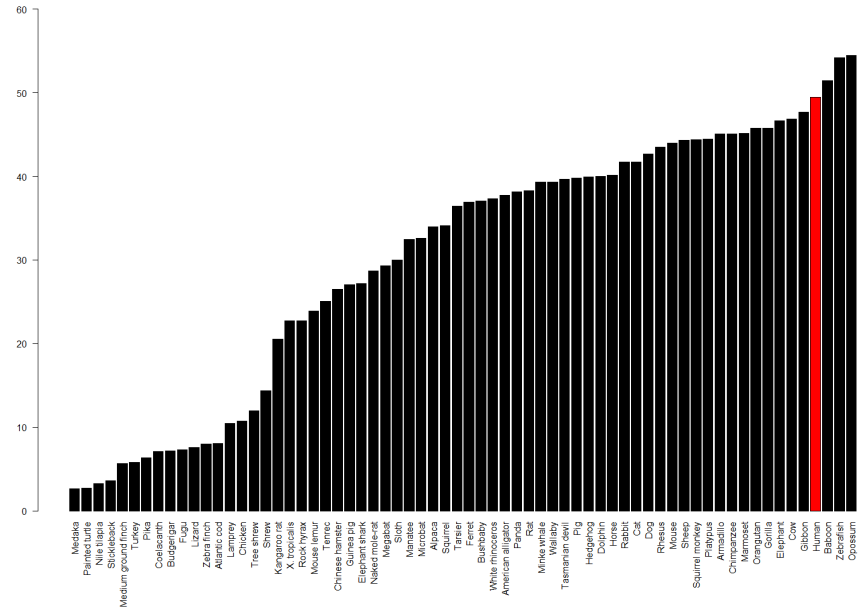

Figure 1.6: The total coverage of repetitive element in 66 vertebrate genomes as annotated by RepeatMasker in the respective genomes[40].

However, the larger vertebrate genomes do not always contain the highest percentage of repetitive elements (Figure 1.7). At least in humans, transposons make up the majority of the repetitive elements that make up the human genome. In particular class I transposons (retrotransposons), which are able to transcribed and inserted into the genome, make up a large portion of the human genome. W.Ford Doolittle and Carmen Sapienza wrote in 1980[41]: "When a given DNA, or class of DNAs, of unproven phenotypic function can be shown to have evolved a strategy (such as transposition) which ensures its genomic survival, then no other explanation for its existence is necessary" and Leslie Orgel and Francis Crick, wrote that junk DNA has little specificity and conveys little or no selective advantage to the organism[42].

The origin of the term "junk DNA" is usually attributed to Susumu Ohno, who used it to describe pseudogenes, which are gene copies that have no known biological function. In its modern day usage, "junk DNA" is used to describe DNA sequence that goes not play a functional role in an organism. Dr. Ohno estimated that there would be an upper limit to the number of functional loci in mammalian genomes based on mutational load and a fixed mutation rate.

He predicted that mammalian genomes could not have more than 30,000 loci under selection as this would guarantee a progressive decline in fitness, leading to extinction.

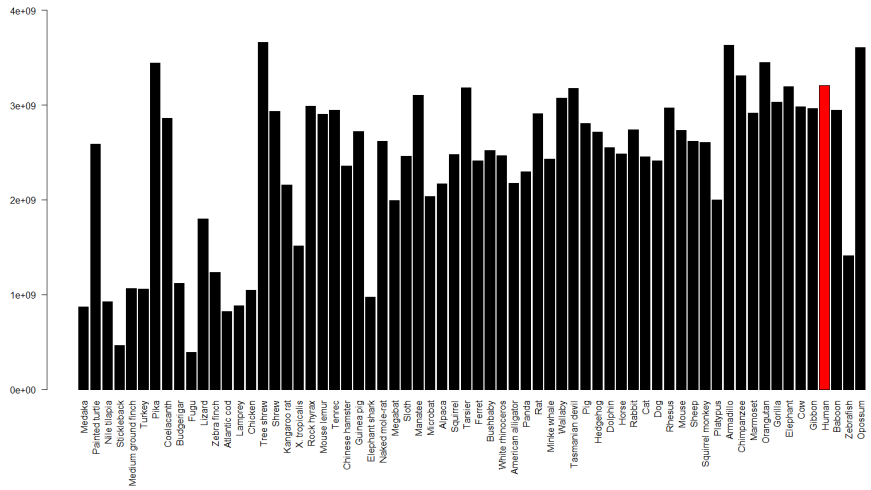

Figure 1.7: The genome sizes of 66 vertebrate genomes, sorted from the lowest to highest percent of repetitive element coverage[43].

- Repeat-associated binding sites (RABS) are over-represented in proximity of regulated genes and that the binding motifs within these repeats have undergone evolutionary selection
- Indeed, studies conducted both at the gene and genome levels have uncovered TE insertions that seem to have been co-opted - or exapted - by providing transcription factor binding sites (TFBSs) that serve as promoters and enhancers, leading to the hypothesis that TE exaptation is a major factor in the evolution of gene regulation.
- The transcription of repetitive elements, especially transposable elements, in specific tissues
- Repetitive elements are usually highly methylated, however differentiation methylation patterns of repetitive elements was observed in specific tissues
- Functions of some GWAS candidates in intergenic regions (such as http://www.nature.com/nature/journ
- The regulated retrotransposon transcriptome of mammalian cells[44]

FANTOM5[45]

1.6 Stem cells

The existence of stem cells was discovered in 1961 and the stem cell theory was published in 1963[46]. In this work, Canadians researchers James Till and

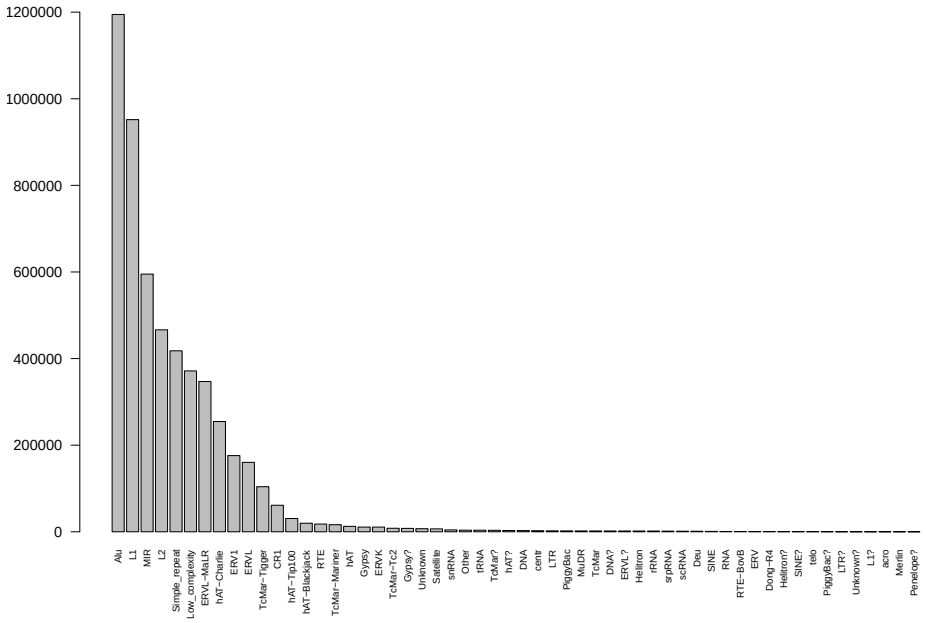

Figure 1.8: The number of repetitive element families determined by RepeatMasker in the hg19 genome.

Ernest McCulloch proved that stem cells were present in bone marrow. They exposed mice to high doses of radiation that killed off the mouse's blood and immune forming system and then injected bone marrow cells into some of the mice. The mice that didn't receive the transplants died and the mice that received the transplants survived because the bone marrow cells rebuilt their blood and immune forming systems. The bone marrow cells were able to reproduce themselves as well as generating different cell types. Human bone marrow transplants are still routinely used nowadays to treat leukaemia and other kinds of blood disorders.

Stem cells have two key characteristics that differ from other cells: they can reproduce themselves for long periods of time, known as self-renewal, and they can differentiate or specialise into specific cell types under certain conditions. What Till and McCulloch discovered were actually adult or tissue stem cells (ASCs), which are multipotent; this means that they have limited differentiation ability and are only able to generate specific types of differentiated cells. For example, haematopoietic stem cells, which are a type of adult stem cells that reside in the medulla of the bone (bone marrow), are only able to give rise to different mature blood cell types and tissues. In 1981, researchers were able to derived embryonic stem cells (ESCs) from mouse embryos (specifically the inner cell mass of blastocysts) and grow them on Petri dishes[47, 48]. In normal development, the inner cell mass begins growing into all the different cell types of the fully developed body and as such ESCs are able to differentiate into all the cell types that make up the body and they are pluripotent. Cells with the highest potency or with the most differentiation potential, are known as

totipotent cells and are able to form all the cell types in a body, including the placental cells. Embryonic cells within the first couple of cell divisions after fertilisation are the only cells that are totipotent.

In 2006, a new methodology was invented by Takahashi and Yamanaka to generate another type of stem cell, called induced pluripotent stem cells (iPSCs)[49]. In their work, they discovered a way to generate pluripotent stem cells from somatic cells, which have no differentiation potential. Since their discovery, many researchers wondered about the development potential of iPSCs and in 2009 a Chinese group were successful in creating a live mice from iPSCs that were able to produce offspring[50]. In one of the first studies to show the clinical application of human iPSCs to human disease, a team derived iPSCs from fibroblasts of patients suffering from alpha 1 antitrypsin deficiency, corrected the disease causing mutation, and used the corrected stem cells to produce liver cells that were transplanted into the liver via intrasplenic injection[51]. Since the cells are genetically identical to the patient, there is no worry about rejection or immune problems, and is one reason why iPSCs are an attractive option for regenerative medicine.

1.7 The role of bioinformatics in genomics

Modern day high-throughput sequencers generate large amounts of data; for example in the blood transcriptome project (Chapter 7), one lane of sequencing on the HiSeq2000 produced 74.5 million reads (using 15 gigabytes of storage space when uncompressed). To deal with data at this scale, informatics tools for storing, managing, and analysis such data sets are absolutely necessary. Bioinformatics can be thought of as informatics for biological data, though historically it was defined as "the study of informatic processes in biotic systems"[52]. The HGP was one of the first large scale international research efforts and bioinformatics was crucial towards the successful completion of the HGP[53]. An important principle established during the HGP, called the "Bermuda Principles", was set by an international assortment of genome-research leaders towards the rapid and public sharing of human genome information. These set of commitments left a lasting legacy in large genomic science projects such as The International HapMap Project, ENCODE and modENCODE, and The Cancer Genome Atlas where data was made freely available prior to publication[54]. The public availability of data from these projects have resulted in major advances in genomics and bioinformatics.

The FASTQ format was formally defined in 2010[55] and has become the *de facto* format for storing raw sequencing output. FASTQ is similar to the FASTA format with the addition of storing quality scores for each nucleotide. The format is simplistic, which may be the reason for its popularity before being formalised. The establishment of a standard format is important as developers can produce programs, such as sequence aligners, that can process output from different sequencing machines. Another standard file format for storing sequence alignments is the Sequence Alignment/Map (SAM) format[56].

The reads generated from the second generation of sequencers were much shorter and much more numerous than traditional Sanger sequencing reads. The Illumina GAII can produce up to 100 million reads of 50 bp in a single run. Traditional sequence alignment tools using the NeedlemanWunsch algorithm or

SmithWaterman algorithm were simply too slow and new developments were necessary. The popular short-read alignment tool, BWA[57], implements the BurrowsWheeler transform to deal with the vast number of short reads.

Comparing genomic features using BEDTools[58].

Expression data sets are commonly stored as matices; for example if we let A be an $m \times n$ matrix, where a_{ij} are elements of A, the i^{th} row would represent the transcriptional response of the i^{th} gene and the j^{th} column would represent the expression profile of the j^{th} assay:

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ . & & . & & . \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{in} \\ . & & . & & . \\ a_{m1} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}$$

A differential expression analysis can be easily performed on an expression matrix, such as by using the edgeR package[59] from Bioconductor[60]. Multidimensional scaling methods, such as Singular Vector Decomposition (SVD), can also be applied on the matrix to find underlying patterns in the data. A very popular visualisation method of expression data is the use of heatmaps that reflect the relative expression strength of each element in the matrix (a_{mn}), which is commonly used in tandem with hierarchical clustering. Co-expression matrices can be further derived from the expression matrix to find transcripts with a similar transcriptional response or assays with similar expression profiles. Networks or graphs can be built from such co-expression matrices for visual purposes or for analysing the structure of co-expression patterns.

Chapter 2

Template switching artifacts

Published online 24 November 2012

Nucleic Acids Research, 2013, Vol. 41, No. 3 ***e44***
doi:10.1093/nar/gks1128

Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching

Dave T. P. Tang[1], Charles Plessy[1], Md Salimullah[1], Ana Maria Suzuki[1], Raffaella Calligaris[2], Stefano Gustincich[2] and Piero Carninci[1,*]

[1]Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan and [2]Sector of Neurobiology, International School for Advanced Studies (SISSA), via Bonomea 265, 34134 Trieste, Italy

Received March 13, 2012; Revised September 27, 2012; Accepted October 23, 2012

ABSTRACT

Template switching (TS) has been an inherent mechanism of reverse transcriptase, which has been exploited in several transcriptome analysis methods, such as CAGE, RNA-Seq and short RNA sequencing. TS is an attractive option, given the simplicity of the protocol, which does not require an adaptor mediated step and thus minimizes sample loss. As such, it has been used in several studies that deal with limited amounts of RNA, such as in single cell studies. Additionally, TS has also been used to introduce DNA barcodes or indexes into different samples, cells or molecules. This labeling allows one to pool several samples into one sequencing flow cell, increasing the data throughput of sequencing and takes advantage of the increasing throughput of current sequences. Here, we report TS artifacts that form owing to a process called strand invasion. Due to the way in which barcodes/indexes are introduced by TS, strand invasion becomes more problematic by introducing unsystematic biases. We describe a strategy that eliminates these artifacts *in silico* and propose an experimental solution that suppresses biases from TS.

INTRODUCTION

Reverse transcriptase (RT) has been widely used for the construction of cDNA libraries since its discovery (1,2) and has been subsequently used for gene expression studies. One intrinsic property of RT is that once it has reached the 5′ end of a RNA molecule, the 7-methylguanosine at the cap site is reverse transcribed to cytosine residues (3). This activity at the cap site has also been previously demonstrated on RNAs with an artificial adenosine cap, which was reverse-transcribed to thymidine (4). In addition to this mechanism, RT also exhibits terminal transferase activity that allows the addition of non-templated nucleotides (predominantly cytidines) once it reaches the 5′ end of a RNA molecule, especially in the presence of manganese (5). Combined, these two mechanisms form a cytosine overhang at the 3′ end of the cDNA after reverse transcription and serves as a useful marker for the 5′ site of the RNA. These properties have been taken advantage of in the construction of full-length cDNA libraries (6). More specifically, the library construction method uses oligonucleotides incorporating a stretch of consecutive ribo-guanosine nucleotides, $r(G)_3$, at the 3′ end of the first strand cDNA that allows for the hybridization of the oligonucleotide with the cytosine overhang. Once hybridized, the RT then switches templates and starts polymerizing the oligonucleotide, thereby incorporating the oligonucleotide sequence with the cDNA sequence. This process is known as the template-switching (TS) mechanism.

Following original cDNA cloning protocols (6,7), several high-throughput transcriptome analyses protocols have incorporated the TS mechanism (8–12). The TS oligonucleotide used for the hybridization to the cytosine overhang is further used for incorporating priming sites for downstream steps in the respective protocols. Furthermore, in the experiments conducted by Plessy *et al.* (9) and Islam *et al.* (10), the TS oligonucleotide was used to incorporate DNA barcode sequences (also known as DNA indexes) into its cDNA libraries, allowing for pooled or multiplexed reactions. By including a set of known sequences (i.e. barcodes) directly upstream of the $r(G)_3$ in the TS oligonucleotide, these sequences become identifiers for different samples. The pooling of several samples into a single sequencing reaction is a common strategy towards minimizing costs and labor (13) and increases the data throughput.

*To whom correspondence should be addressed. Tel: +81 45 503 9222; Fax: +81 45 503 9216; Email: carninci@riken.jp

Given the constant increase of number of reads per sequencer run, techniques for multiplexing libraries are flourishing. For example, the current protocol of the HiSeq 2000 sequencer can produce up to 3 billion single reads that pass filtering on a single flow cell run (http://www.illumina.com/systems/hiseq_systems/hiseq_2000_1000/performance_specifications.ilmn). Methods that measure transcript expression levels by their 5′-end such as STRT (14), CAGE (15) or nanoCAGE (16) have a reduced complexity compared with RNA-Seq, and therefore take a particular advantage of multiplexing. In addition to TS, there are ligation- and polymerase chain reaction (PCR)-based methods that have been used for introducing barcodes into samples for multiplexed experiments. In single-read libraries using restriction enzymes to cleave sequence tags, the barcode is often added by ligation at the 5′ or 3′ end of the construct, like for CAGE (15), the cleaved version of nanoCAGE (9), SAGE protocols such as HT-SuperSAGE (17) or small RNA libraries (18). However, studies have demonstrated that ligation-based methods are heavily biased due to RNA ligases having sequence-specific biases (19,20). One strategy used for dealing with ligation-based biases has been to standardize the sequence at the end of the RNA adaptor that will be ligated (18). Another proposed strategy was to use a pool of RNA adaptors (20); however, Alon *et al.* (19) have further suggested that barcodes should be introduced via PCR-based methods, such as Illumina's industry standard known as TruSeq. TruSeq uses 6-nt barcodes, which are detected as a separate step after sequencing the forward read or its mate pair. Read indexes are primed with a separate oligonucleotide, which gives a lot of flexibility in their placement in the 5′ and 3′ linkers. The designers of TruSeq protocols took this opportunity to place the index far from the reaction sites, usually in the tail of the primers. However, the indexes are introduced at a late step in the reaction, as there are no universal primers that would amplify the libraries and keep the indexes at the same time. As a consequence, it does not allow the pooling of the samples at early preparation steps, and for this reason, strategies where barcodes can be introduced as early as possible, such as via TS or ligation-based methods, are still preferred in situations that strongly benefit in terms of cost or logistics from early pooling. The question of which multiplexing approach to take is highly dependent on the nature of the research. For example, in a study by Kivioja *et al.* (21), they describe a method for introducing unique molecular identifiers via TS for quantifying transcript numbers. These identifiers are random bases in the TS oligonucleotides and function like random barcodes that index RNAs molecules instead of indexing samples. Double-stranded ligation and PCR are ruled out as alternatives for introducing indexes. In the case of ligation, it would be too difficult to produce the double-stranded adaptors because random sequences will not be reverse complementary. Indexing via PCR would be too late, as the purpose of these identifiers is to detect PCR duplicates. Lastly, Kivioja *et al* (21) have envisioned that unique molecular identifiers can be combined with sample barcodes.

One of the main advantages of using TS is the lack of purification and adaptor ligation steps, which eliminates ligation-introduced biases and also minimizes the loss of material. This has made TS highly suitable in studies working with a limited amount of RNA (9,10,12,22,23). Although TS is an inherent property of RTs, and is therefore only implemented in transcriptome studies, we may see an increase in the use of TS due to the growing interest in single cell transcriptomics (24). There are, however, intrinsic problems associated with the TS mechanism, such as the concatenation of TS oligonucleotides due to cycles of terminal transferase activity and TS oligonucleotide hybridization (25). Another issue that we address here is the interruption of first strand synthesis via strand invasion. Although TS is most efficient when RT has reached the end of the RNA template, the TS oligonucleotide may hybridize to the first strand cDNA due to sequence complementarity before the RT has finished polymerizing. This creates first strand cDNAs that are artificially shorter than the RNA due to the incomplete reverse transcription process. Furthermore, although this is usually a systematic bias, this becomes more problematic in protocols using varied TS oligonucleotides for barcoding purposes, as the strand invasion process is dependent on the oligonucleotide sequence. We study in detail the artifacts and biases created by strand invasion in a protocol using the TS mechanism and demonstrate how it is possible to remove such artifacts *in silico*. Lastly, we propose possible experimental strategies that may help reduce such artifacts and biases in protocols that use TS, and demonstrate it with the nanoCAGE protocol.

MATERIALS AND METHODS

NanoCAGE libraries were prepared from total RNA isolated from human whole blood samples (200 ng per sample) and rat whole body RNA (500 ng per sample) according to a previously published protocol (16), and sequenced using the Illumina GAIIx instrument on five (four for blood samples and one for rat samples) sequencing lanes. These quantities of starting material are well above the recommended quantity of 50 ng, and we therefore expected that the difference would not cause one set of samples to underperform compared with the other set. Blood samples were collected in PAXgene blood RNA tubes (PreAnalytix) following manufacturer's instructions from seven donors (four male and three females) of the same ethnicity with an average age of 67 years and a standard deviation of 6.6 years and were labeled as 14–20P. Blood samples were collected following a fasting period and at the same hour of the day to help reduce variability. The rat whole body RNA were a generous donation from Dr. Alistair Forrest and are commercially available from BioChain (http://www.biocat.com/products/R4434567-1-BC).

We processed all five lanes of sequencing from the nanoCAGE libraries as follows. Using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), we extracted raw tags and distributed them into their respective samples based on their barcode sequence. Raw reads that

did not match a barcode sequence were discarded mainly owing to poor or ambiguous base was calling. Barcode sequences (and the common spacer sequence in the rat libraries) and the leading guanosines were trimmed off from the sequenced read. Next, we filtered out artifactual reads using TagDust (26), a program that filters out reads resembling the primer, linker and adaptor sequences used during library construction, using a false discovery rate of 0.01. Lastly, reads mapping to the ribosomal sequences U13369.1, NR_003285.2, NR_003286.2 and NR_003287.2 with ≤2 mismatches were considered to be ribosomal sequences and removed (Supplementary Table S1). After all pre-processing stages, reads were mapped to the hg19 or rn4 genome depending on the sample using BWA (27) with a mismatch threshold of 2. Using SAMTools (28), we selected reads with a mapping quality (MAPQ) of 10 (90% accuracy) or better.

Identifying barcode-biased tags in the whole blood libraries

For the 21 human whole blood libraries, we used 14 barcodes, consisting of 12 unique barcode sequences (ACATAC, AGTACG, ATCACG, CACGAT, CGATAC, GAGACG, GCTATA, GCTCAG, GTAGTG, GTATAC, TATGTG and TCGACG) where the mean Hamming distance between all pairwise barcodes is 4.32. For each sample, three libraries were made using two barcodes, i.e. one technical replicate was made with one barcode sequence, and the other two technical replicates made with the other barcode sequence (Table 1).

Next, using a present/absent criterion, we identified reads among technical replicates that were only present when using one barcode and absent when the other barcode is used. We used a threshold of ≥21 raw reads for our present criteria (Supplementary Table S4). Marioni *et al.* (29) reported that if technical replicates are sequenced, the read counts for a particular feature should vary according to the Poisson distribution. Thus, it is unlikely that our selected reads are a consequence of natural variation but rather are attained by the use of a different barcode sequence. Sequence logos (30) were created by extracting the nine nucleotides upstream of these mapped reads, and the sequence enrichment was calculated using unique upstream sequences.

Filtering strand invasion artifacts

From our selection of barcode-biased reads, we observed that the upstream region of these reads showed sequence complementary to the tail of the TS oligonucleotide (Figure 2), which is a consequence of strand invasion. This served as a marker for strand invasion artifacts, which was subsequently used as our strategy for their removal. Thus, once reads were mapped to a reference, the nine nucleotides immediately upstream were extracted, and using a global alignment approach (31), they were aligned to the last nine nucleotides of the TS oligonucleotide used for construction of that particular library. The edit distance was used as a metric for the alignment, and a single mismatch or gap constituted an edit distance of one. A perfect alignment would thus have zero edits.

Table 1. A summary of the biological and technical replicates used in this study, along with the barcodes and the number of reads that were mapped at a MAPQ of ≥10

Samples	Technical replicate	Barcodes	Number of reads mapping at q10
Human			
14P	1	GCTATA	597909
	2	CACGAT	711936
	3	CACGAT	960204
15P	1	GTAGTG	445901
	2	CGATAC	592336
	3	CGATAC	674823
16P	1	TATGTG	1040935
	2	GAGACG	1163416
	3	GAGACG	756476
17P	1	ACATAC	722023
	2	GCTCAG	538660
	3	GCTCAG	695706
18P	1	ATCACG	685146
	2	GTATAC	889014
	3	GTATAC	884897
19P	1	CACGAT	371069
	2	TCGACG	663186
	3	TCGACG	420775
20P	1	CGATAC	741908
	2	AGTACG	1195816
	3	AGTACG	1431334
Rat	1	ACAGAT	927429
	2	ATCGTG	849609
	3	CACGAT	793598
	4	CACTGA	810155
	5	CTGACG	863029
	6	GAGTGA	895320
	7	GTATAC	1005221
	8	TCGAGC	823343

We observed the enrichment of at least two guanosine nucleotides directly upstream of where a strand invasion artifact mapped. Thus, we imposed this criterion to our filtering strategy; reads were only considered to be artifacts if two of the three nucleotides directly upstream were guanosines. Lastly, as an indication of the edit distance threshold to use for data filtering, we filtered libraries using edit distances of zero to five and measured the Spearman's rank correlation coefficient between technical replicated libraries at each threshold. The filtering strategy was implemented using Perl, and an executable version of the script is available as supplementary data.

Specificity and sensitivity

The specificity of a method relates to the ability of identifying negative results, assessed by the number of false positives. We created a negative set, i.e. putatively non-biased reads, by selecting for the least variable reads among technical triplicates. We first normalized reads by tags per million, and selected the top 20% of least variable reads among replicates. In contrast, the sensitivity refers to the ability of identifying positive results, assessed by the number of true positives. We created a positive set, i.e. strand invasion artifacts, in the same manner that

we identified barcode-biased tags described above. With our negative and positive sets, we then applied our barcode filtering scheme described above with an edit distance of four. The specificity was calculated as the 'number of true negatives/(number of true negatives + number of false positives)', and the sensitivity was calculated as the 'number of true positives/(number of true positives + number of false negatives)'.

Differential expression analysis

For the comparison of different libraries, we used a previously developed read/tag clustering method (32), as opposed to comparing individual reads. The clustering method aggregates reads that are mapped within a window of 20 nucleotides into single entity clusters; the expression of the cluster is the summation of all tags within the cluster. We conducted our differential expression analyses on tag clusters present among technical replicates using the edgeR_2.4.1 package (33) on R version 2.14.1. Within a technical triplicate set, technical replicates made with one barcode were tested against technical replicates made with the other barcode. For the comparison of the rat libraries, we arbitrarily tested the libraries made with the ACAGAT, ATCGTG, CACGAT and CACTGA barcodes against the libraries made with the CTGACG, G AGTGA, GTATAC and TCGAGC barcodes. We used an independent filtering criterion (34), selecting for tag clusters with $\geq$10 raw reads. The standard edgeR pipeline was carried out using a common dispersion approach (and tag-wise dispersion for the rat libraries) and the Benjamini and Hochberg's (35) approach for controlling the false discovery rate. Tag clusters with an adjusted *P*-value of $\leq$0.01 were defined as differentially expressed.

RNA-Seq data sets

We processed two independently produced RNA-Seq data sets, made using two different protocols (10,36). Briefly, Islam *et al.* analysed the single cell transcriptomes of mouse embryonic fibroblasts and embryonic stem cells. The Islam *et al.* RNA-Seq libraries, which was made using TS and in a manner very similar to nanoCAGE, was downloaded directly from the author's website and was processed in the same manner as our nanoCAGE libraries owing to the similarity between the protocols. Briefly, Guttman *et al.* produced RNA-Seq libraries from mouse embryonic stem cells, neuronal precursor cells and lung fibroblasts by mRNA fragmentation and random-primed reverse transcription. The Guttman *et al.* data set was downloaded from the DNA Data Bank of Japan under the accession number SRP002325, and the sequenced reads were mapped using TopHat (37) on the default settings. After all pre-processing steps, we compared the derived transcript structures between the fibroblasts libraries made by Islam *et al.* and by Guttman *et al.* In addition, we also compared different fibroblast libraries made with different barcodes in the Islam *et al.* data set.

RESULTS

Barcode specific reads in nanoCAGE libraries

Total RNA, isolated from whole blood samples derived from seven donors, was used to prepare 21 separate nanoCAGE libraries where each sample was made in triplicate (Table 1). Furthermore, libraries were prepared together to help limit batch effects. To study the effect of using different TS oligonucleotides and thus the barcode sequence, we prepared the same sample identically except for the TS oligonucleotides used; two barcodes were used per technical triplicate. As there are an odd number of replicates, two of the three replicates were prepared with one barcode and the remaining replicate prepared with the other barcode. NanoCAGE libraries were prepared following a previously published protocol (16). The 21 nanoCAGE libraries were then sequenced in multiplex using Illumina's GAIIx instrument on four sequencing lanes.

Sequenced reads in the nanoCAGE protocol represent the site at which TS occurred (Figure 1), which represents the 5′ end of a RNA molecule and thus the putative transcriptional starting site (TSS) (9). Hence, to identify artifacts, we could compare nanoCAGE reads that do not map to known promoters of transcripts, although these could represent previously uncharacterized transcripts. A more definitive approach not requiring transcript annotations is to search for intra sample differences, i.e. reads present only in one set of barcoded technical replicates. To correctly identify the corresponding transcript for a sequenced read, we selectively analysed 16 281 067 reads that could be mapped to the genome with 90% confidence (MAPQ of $\geq$10) (27). Finally, from this set, we identified 132 980 barcode specific reads, i.e. reads present only in one set of technical replicates using a particular barcode and not the other, where the variance is unlikely due to Poisson noise (see 'Materials and Methods' section).

From our barcode specific nanoCAGE reads, we analysed the region directly surrounding the reads. Interestingly, the upstream sequence of these barcode biased reads revealed an enrichment of nucleotides that resembled the 3′ end of the TS oligonucleotide used for that library (Figure 2). The sequence logos illustrate an enrichment of guanosines at positions −1 to −3, which corresponds to the $r(G)_3$ tail of the TS oligonucleotide, whereby positions −4 to −9 show a varied enrichment of nucleotides that resemble the barcode used to produce the library, especially positions −4 to −6 (Figure 2). These results suggest the hybridization of the TS oligonucleotide to a complementary region on the first strand cDNA, i.e. strand invasion, and thus produces TS artifacts in a barcode dependent manner (Figure 1B). Although the $r(G)_3$ tail of the TS oligonucleotide preferentially binds to the cytosine overhang created by the RT (9), the increase in sequence complementarity in the 3′ tail of the TS oligonucleotide may increase the hybridization of the TS oligonucleotide to the first strand cDNA (Figure 1B).

Filtering out strand invasion artifacts

Artifactual reads need to be removed before they are used for further downstream analyses (26). The TS mechanism

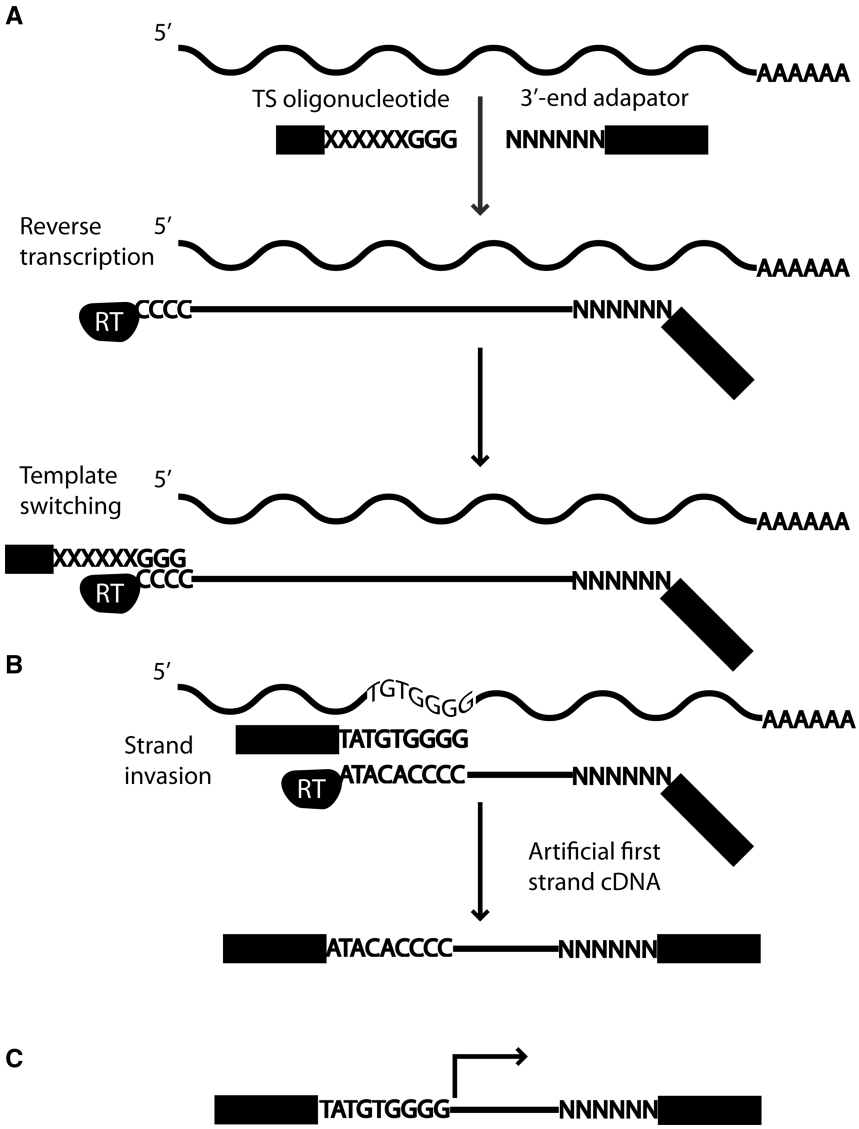

Figure 1. (**A**) The TS mechanism is used for first strand cDNA synthesis. First, an oligonucleotide hybridizes to the RNA molecule, and RT starts polymerizing. Once the RT reaches the 5′ end of the RNA, a cytosine overhang is formed. The TS oligonucleotide containing three riboguanosines hybridizes to the cytosine overhang and the RT switches template and polymerizes the TS oligonucleotide. (**B**) However, during RT synthesis, if the polymerized region has sequence complementary with the 3′ tail of the TS oligonucleotide, it may invade and hybridize with the first strand cDNA. RT then switches template and polymerizes the TS oligonucleotide. However, this strand invasion process has resulted in a cDNA that is shorter than the RNA. (**C**) With the nanoCAGE protocol, sequencing begins just upstream of the site of TS, which includes the barcode and the riboguanosine linker sequence. The barcode and linker sequences are trimmed off during processing steps, and the final read sequence is indicated by the black arrow.

is expected to occur at the cytosine overhang created by the RT (Figure 1A); thus, the sequence immediately upstream of a nanoCAGE tag should exhibit sequence complementarity only on a random basis, although this is largely dependent on the makeup of the genome. Under these assumptions, we devised a strategy for removing strand invasion artifacts by aligning the sequence immediately upstream of reads to the 3′ tail of the TS oligonucleotide. We chose to align the nine nucleotides directly upstream of a read (Figure 1C) to the last nine nucleotides of the TS oligonucleotide owing to the enrichment profiles previously observed (Figure 2).

Next, we analysed a range of sequence complementarity scores to determine the optimal threshold for classifying reads as artifacts. First, we carried out a global alignment (31) between the sequence upstream of a read and the TS oligonucleotide tail for all libraries. We directly used the edit distance of an alignment as a measurement of the sequence complementarity, where gaps and mismatches were individually constituted as one edit; a perfect alignment would thus have zero edits. In addition, we only classified reads as artifacts if two or more of the three nucleotides directly upstream were composed mainly of guanosines (see 'Materials and Methods' section). Lastly, we filtered out reads on a range of edit distances, from zero to five, and found that by removing such noise, we had technical replicates that correlated better with each other (Supplementary Table S2).

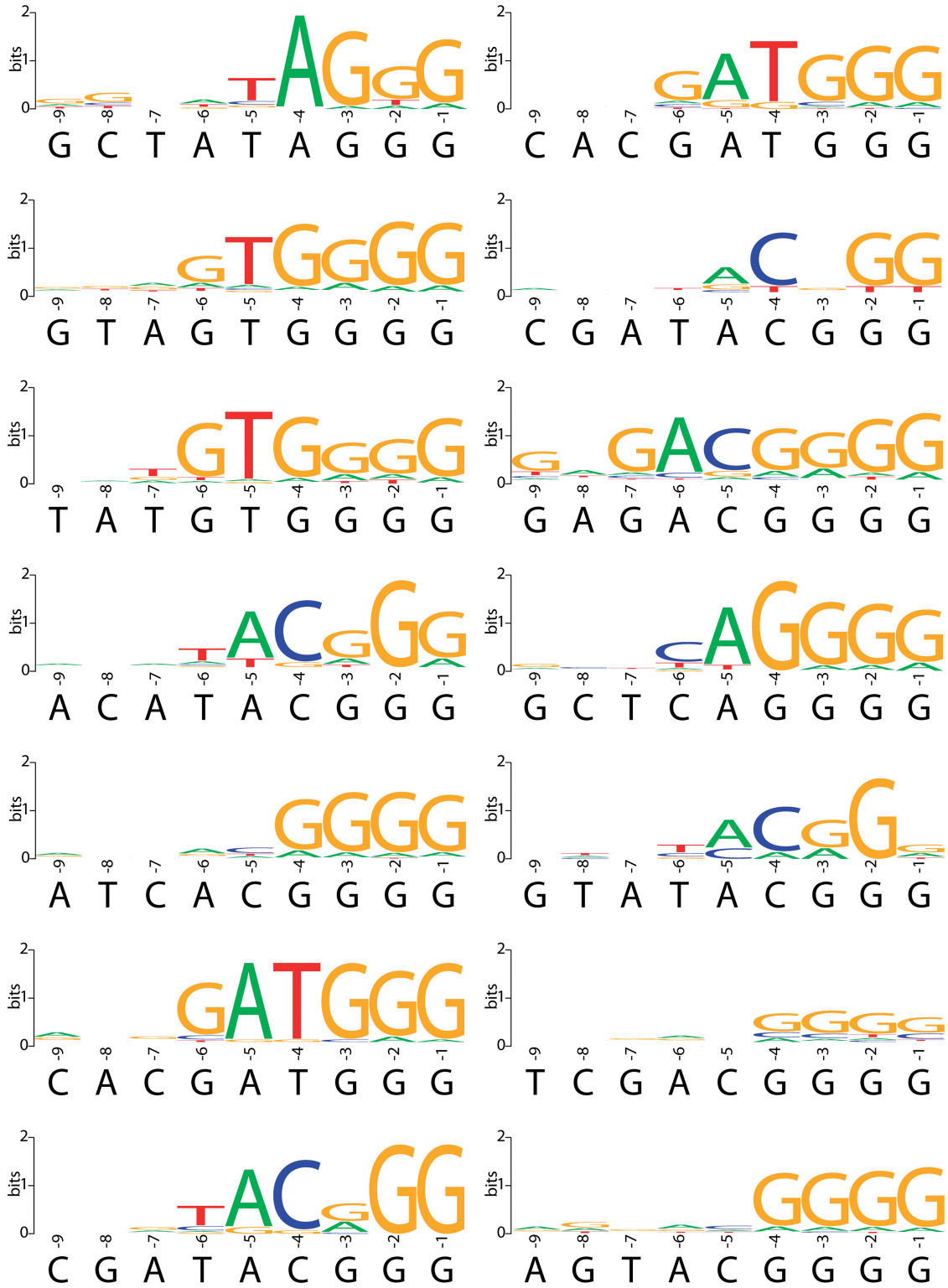

Figure 2. By preparing technical triplicates with different barcodes, we could identify reads present in a barcode specific manner, i.e. barcode-biased reads. The sequence logos were created using the sequence directly upstream of individually mapped barcode biased reads. The barcode sequence used to prepare each library and the three riboguanosines of the TS oligonucleotide are shown directly below the corresponding sequence logos. The enrichment profiles closely resemble the tailing sequence of the TS oligonucleotide used to construct that particular library.

Effects of artifact filtering on library correlation

A common metric used to determine library similarity is by correlation. We assessed the correlation of technical replicates after our filtering strategy at various thresholds. Given the nature of CAGE reads and promoters, we first clustered reads into what are known as 'tag clusters' (32), enabling us to measure the correlation of libraries. Tag clusters are representative of putative promoter regions whereby the number of reads mapping to these regions represents the level of expression (see 'Materials and Methods' section). We calculated the Spearman's rank correlation coefficient between all technical replicates, given the skewed expression rate of blood transcripts, i.e. the distribution of transcript expression is not linear; so we used a non-parametric measure of correlation. The distribution of transcripts in blood is largely skewed by the presence of globin transcripts, which resulted in the under sequencing of other transcripts. This under sampling resulted in an increase of noise, especially for transcripts that are lowly expressed, and subsequently lower correlations between replicates. The removal of globin transcripts would not significantly affect the Spearman's rank correlation coefficient owing to the way the correlation is calculated. For technical replicates made with different barcodes, we observed a general increase in correlation between the libraries as we relaxed the similarity threshold, i.e. increasing the edit distance (Supplementary Table S2). The increase in correlation was the direct consequence of removing library specific reads, i.e. TS artifacts. The opposite effect, a decrease in correlation after read filtering, was observed when comparing technical replicates with the same barcode (Supplementary Table S2). The correlation of technical replicates was inflated owing to TS artifacts, and the removal of these reads decreased the correlation. For the comparison of libraries with different barcodes, the majority of correlations increased until an edit distance of four. This was due to the decrease in stringency, which resulted in real signal being removed by random chance that a loose alignment could be formed between the upstream sequence and the tail sequence of the TS oligonucleotide. Although the correlations between technical replicates are considered moderate, we demonstrated that we are able to identify TS artifacts, and the removal of these artifacts resulted in higher correlations between technical replicates.

Effects of artifact filtering on differential expression detection

One of the core analyses conducted on transcriptome data is a statistical test that detects differential expression of transcripts. An observed difference is statistically significant only when the observed difference is greater than expected from random variation. Transcripts may be spuriously detected as differentially expressed owing to the introduction of experimental variations such as from using different barcodes. We tested this notion by conducting differential expression analyses using edgeR (33) on technical replicated libraries before artifact filtering, after filtering and after randomly removing reads (Figure 3). Given that our analyses were carried out on technical replicates, we would expect to find very few tag clusters that are detected as differentially expressed. However, a fraction of tag clusters were detected as differentially expressed between technical replicates before filtering (Supplementary Table S3). In all cases, the removal of strand invasion artifacts decreased the number of differentially expressed candidates (Supplementary Table S3). Using an edit distance of four for barcode filtering, on average, we observed a roughly 10-fold decrease in the number of differentially expressed candidates. In contrast, removing random reads resulted in a slight decrease of 1.2-fold in differentially expressed candidates.

Sensitivity and specificity of artifact filtering

We have experimentally prepared our libraries in such a way that we can identify TS artifacts. Using a set of nanoCAGE reads that were identified as TS artifacts (see 'Materials and Methods' section), i.e. true positives, we applied our filtering strategy to measure the sensitivity of the method. Reads not detected as artifacts in this set were considered as false negative. Using an edit distance metric of four, the average sensitivity was ~94.3% across the entire data set; we could detect 125 357 of the 132 980 true positives (Supplementary Table S4).

The specificity of a method gives an estimate of the number of false positives. This measure is important for quantifying the potential amount of signal that is removed due to the random chance that the upstream region of a read resembles the 3′ tail of the TS oligonucleotide. To determine a true negative set, i.e. not barcode biased, we selected reads with the lowest amount of variance between the technical replicates (see 'Materials and Methods' section), and if any of these reads were filtered out, they were considered false positives. Of the subset of reads we considered to be a true negative set ($n = 135\,613$), on average 6.7% +/− 2.1 of these reads were considered false positives (Supplementary Table S5). However, one should consider that even our true negative set may contain strand invasion artifacts, i.e. a false positive in the sense of being a true negative, and we only examined a small proportion of the total number of reads in a library; in reality, the false positive rate is likely to be much lower.

Degree of bias from different barcode sequences

Strand invasion occurs during first strand cDNA synthesis, and successful hybridization depends on the degree of sequence complementarity between the cDNA and the 3′ tail of the TS oligonucleotide. Therefore, the number of TS artifacts becomes a function of the number of RNA molecules that contains sequence complementarity to the TS oligonucleotide. Barcode sequences that occur more prominently among RNA molecules would result in a higher number of TS artifacts. To test this hypothesis, we scanned the genome in a sliding window manner. Given that the last six nucleotides of the TS oligonucleotide are the most important for strand invasion (Figure 2), we tallied the number of all possible 6-mers that end in GGG (total of 64 6-mer combinations) across the human genome (hg19) on both strands. For the sake of simplicity

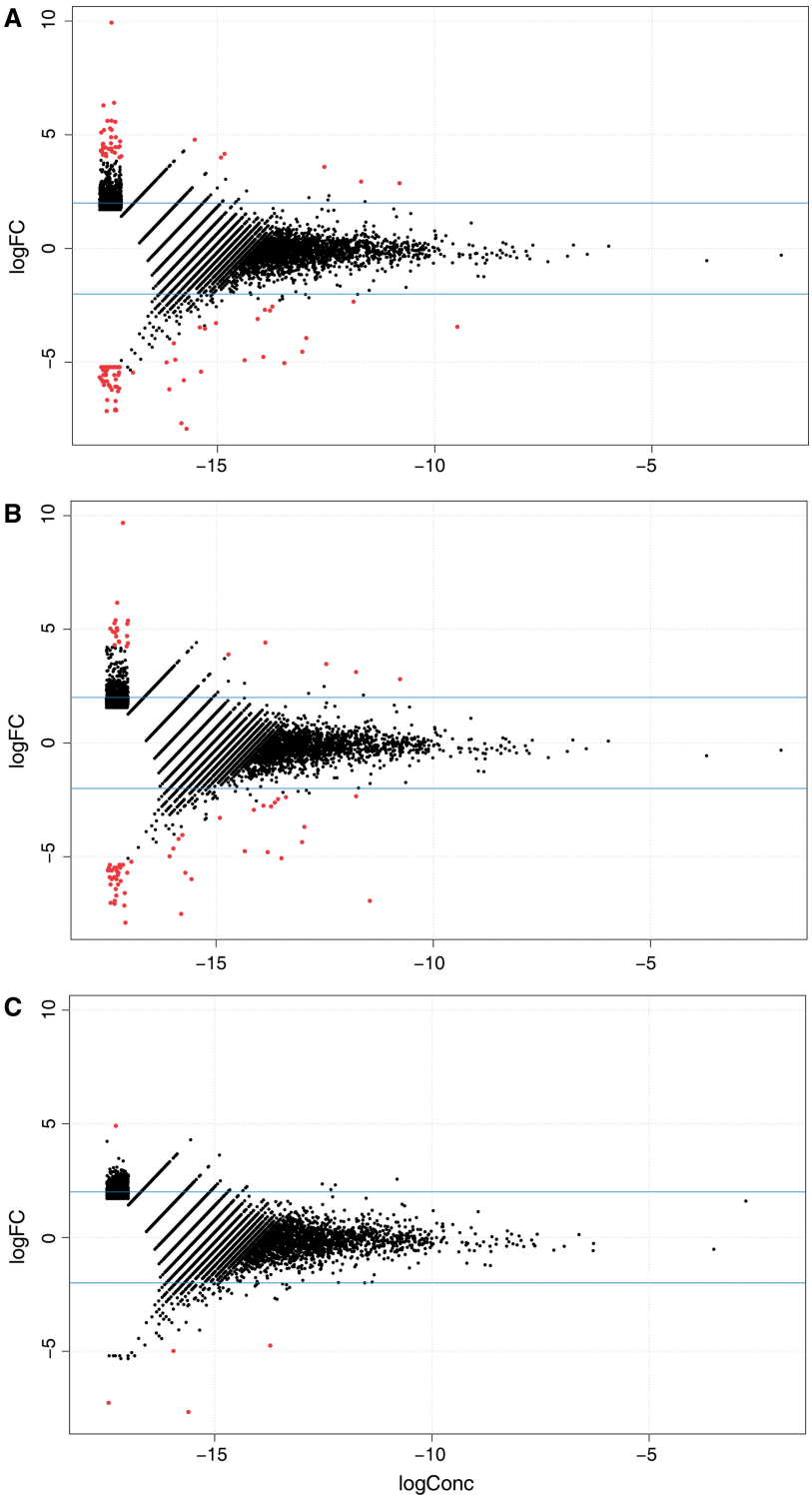

Figure 3. Differential expression analyses were carried out between technical replicates made from different barcode sequences using edgeR. The scatter plots show the log-fold change (*y*-axis) against the log concentration (*x*-axis) for tag clusters present among the 14P technical replicates. In red are tag clusters that were detected as differentially expressed at an adjusted *P*-value ≤ 0.01. Three separate analyses were carried out: test for differential expression (**A**) before strand invasion artifacts were filtered out, (**B**) when random reads were removed and (**C**) after strand invasion artifacts were filtered out. By removing artifacts, fewer tag clusters were detected as differentially expressed compared with no filtering and removing random reads.

and owing to a lack of a complete transcriptome, we chose to tally this number across the genome as opposed to a defined transcriptome. We previously defined a set of TS artifacts by examining technical replicates made with different barcodes and used this number as an estimate of the number of TS artifacts. We then calculated the Spearman's rank correlation coefficient between the number of TS artifacts and the tallied number for the 6-mer that corresponded to the sequence at the end of the TS oligonucleotide. As expected, a positive correlation (Spearman's rho ~ 0.67) was observed (Supplementary Table S6), which supported the hypothesis that choosing a barcode sequence, which is present more often in the genome, leads to a larger number of strand invasion artifacts.

An experimental strategy for suppressing artifacts and barcode bias

Our results have suggested that first strand cDNAs with regions of sequence complementarity to the last six nucleotides of the TS oligonucleotide are potential sites for strand invasion. Given this information, intuitively it is expected that if the last six nucleotides of the TS oligonucleotide occur more frequently amongst RNA molecules, there would be a higher number of TS artifacts. We established this hypothesis that barcode sequences that are present more frequently in the genome have more strand invasion artifacts (Supplementary Table S6). So to suppress the number of artifacts, one should select barcodes less frequent in the genome. We have also observed that barcodes that end with a guanosine, thus creating a sequence tail of four guanosines in the TS oligonucleotide, have much higher number of TS artifacts (Supplementary Table S6). Furthermore, at transcriptional starting sites, libraries made with a barcode ending with a guanosine have much higher counts of certain transcripts compared with barcodes that do not end with a guanosine (Figure 4); this type of bias cannot be mitigated by our artifact filterer. To suppress this barcoding bias, it is necessary that the sequence directly upstream of the riboguanosines is standardized, a strategy similar to standardizing the adaptor sequence in ligation-based barcoding (18). Additionally, our strategy for the choice of a standard spacer is one that occurs less frequently in the genome. Thus, we can potentially suppress the extent of strand invasion and systematically remove the barcode bias effect.

To test this strategy, we redesigned the TS oligonucleotide to include a 6-nucleotide spacer (GCTATA) directly upstream of the riboguanosines. We produced eight nanoCAGE libraries using eight barcodes from rat whole body RNA, i.e. technical replicates, and sequenced them on one lane on the Illumina GAIIx platform. We processed these libraries in the same manner as our blood nanoCAGE libraries and obtained around 8 million reads in total after processing (Supplementary Table S7). Next, using the tag-clustering method previously described, we aggregated our reads and measured the pairwise correlations of each library; the average Spearman's rank correlation coefficients was $\sim$0.75 (Supplementary Table S7), a vast improvement to the blood nanoCAGE libraries. To investigate how much of the variance in the data is explained by sequencing noise, for each tags per million - normalized tag cluster, we calculated the mean and the exact 95% confidence intervals (CIs) for the mean assuming a Poisson distribution. For each tag cluster and the respective library expression, we tallied the number of times an expression value was inside the CIs; approximately 92% of the total expression values fall inside the 95% CIs. The nanoCAGE protocol is designed to work with few nanograms of total RNAs and require a relatively large number of semi-suppressive PCR cycle, which in addition to the Poisson noise, may account for points that fall outside of the 95% CIs. Semi-suppressive PCR allows the use of random primers, which can capture non-coding RNAs, but, however, shows suboptimal yields at each PCR cycles. Additionally, a second PCR reaction is needed to add sequencing adapters after the semi-suppressive PCR. In summary, although nanoCAGE can also identify non-polyA RNAs from low starting material (9), it requires two PCR cycles, which may be a source of noise.

When we applied the filtering strategy on the libraries made with the common spacer, we found that on average 4.5% +/− standard deviation of 0.12 (Supplementary table S7) of the total reads were detected as putative TS artifacts compared with an average of 11.1% +/− standard deviation of 6.65 (Supplementary Table S1) for the libraries made without the common spacer. By using a common spacer, all libraries had roughly the same number of putative artifacts, i.e. very low standard deviation, which is also lower than the number of putative artifacts detected in most libraries made without the common spacer. Although the older data set detected an average of $\sim$11%, the number of artifacts is highly dependent on the barcode (Supplementary Table S1), which is the reason for a much higher strand deviation. For example, by using the GCTCAG barcode without a spacer, up to $\sim$25% of the reads were detected as artifacts. By using a common spacer, the biases will affect the same transcripts in the same way in different samples. This is particularly important when conducting differential expression analyses and because of this, it is not necessary to filter out putative artifacts. To test this, we performed a differential expression analysis on the common spacer libraries, and indeed none of the tag clusters were detected as significantly differentially expressed.

DISCUSSION

The TS mechanism has been exploited in full-length cDNA library construction owing to its technical simplicity (6), in transcriptome analyses due to its ability to mark the 5′ end of a RNA molecule (9) and its flexibility in incorporating DNA barcodes for multiplexing (10) and for incorporating DNA fingerprints for quantifying the absolute number of molecules (21). Owing to the elimination of adaptor mediated steps, RNA material can be conserved, making TS an attractive choice when

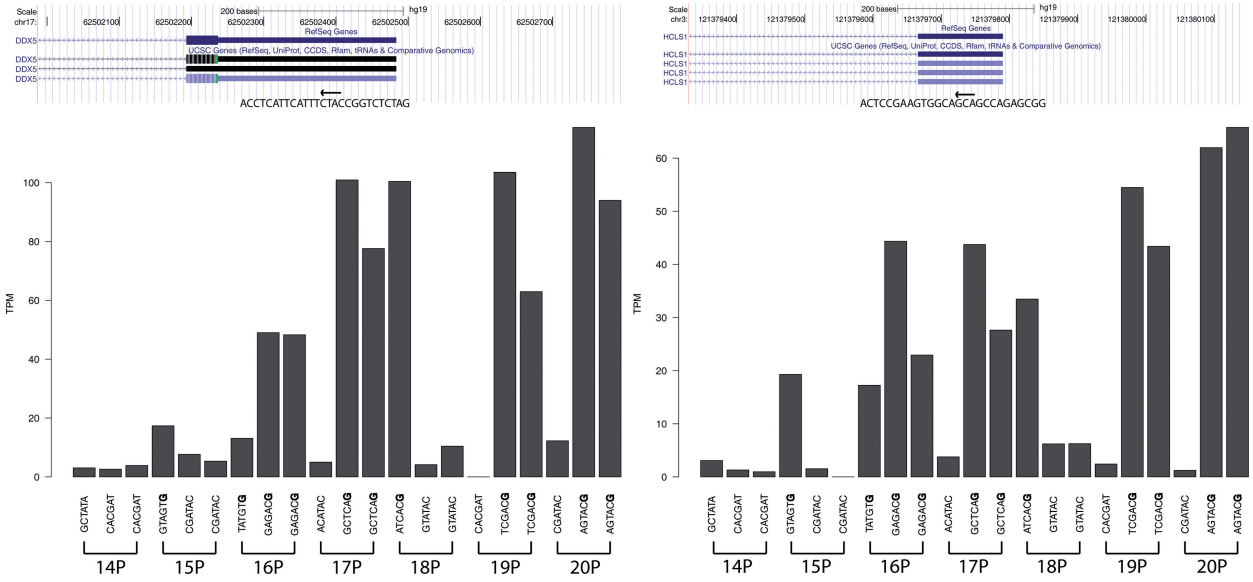

Figure 4. Barcode bias in nanoCAGE technical replicated libraries. The choice of barcode sequence can affect the read count pertaining to the transcriptional starting site. Here, we show two examples for the genes DDX5 and HCLS1, where the read count fluctuates according to the barcode sequence and not by the sequencing depth or by the library. Libraries made with barcodes ending with a guanosine (shown in bold in the bar plot) have a much higher tags per million (TPM) count than barcodes that end with other nucleotides.

working with limited amounts of sample, such as with single cell type analyses (10,12). Furthermore, the decreasing costs of DNA sequencing are driven by an increase of throughput in a constant number of sequencing lanes, which makes multiplex sequencing determinant for cost and time efficiency. Although TS has been used to incorporate barcodes during reverse transcription for multiplexing, one particular drawback of this approach is that the barcode sequences may skew the following reactions, in particular PCR, in favor of one sample. For this reason, strategies where the barcode is added at a last step are sometimes preferred, for instance in the protocols of Illumina's TruSeq product line. Nevertheless, this has the drawback that samples cannot be pooled as early as with TS, which increase the work load and cost of the experiment. In addition, because these two methods of barcoding are directed at different parts of the library constructs, they can be used together to implement combinatorial multiplexing. By combining two barcodes together, the index diversity is greatly increased, and this allows the unique labeling of all transcripts in a sample (38). This approach takes advantage of the very high throughput of current sequencers, which can also be applied to labeling several thousands of low complexity single cell libraries.

Here, we have characterized and investigated a source of bias that is inherent to TS: the production of spurious, sequence-specific reads owing to strand invasion and different hybridization rates as a consequence of choosing different barcodes. We have shown that the extent of strand invasion depends highly on the sequence of the TS oligonucleotide, especially the last six nucleotides. All oligonucleotides in the reverse transcription reactions can interrupt the first-strand cDNA synthesis by strand invasion, and we have previously observed oligo-dT primers being template switched at the 5′ of T-rich regions of the mRNAs (data not shown). This strand invasion becomes more problematic when different sets of TS oligonucleotides are used to barcode specific samples, as this subjects the samples to different degrees of bias. In these multiplexed libraries, the strand invasion artifacts will produce sample-specific signals, which will can wrongly suggest correlations or in contrary mask the similarity between related samples. For example, samples using two barcodes that end in the same six nucleotides may artificially cluster together irrespective of the sample condition. Even in non-multiplexed libraries, strand invasion produce shortened cDNAs that can systematically bias expression levels and create artifacts that do not reflect the transcriptome. We compared two different RNA-Seq protocols, one using TS (and with the same multiplexing strategy as nanoCAGE) (10) and another by conventional RNA fragmentation on mouse fibroblasts and observed different coverage patterns (Supplementary Figure S1). The transcript profile observed in the TS RNA-Seq protocol is likely a consequence of strand invasion. Moreover, we have also observed different transcript profiles in biologically replicated samples that were made from different barcodes (Supplementary Figure S2). As we have demonstrated in our work, it is crucial to control strand invasion products especially with respect to introducing barcodes by TS.

It is possible to identify strand invasion products *in silico* and consequently have them removed. We have shown that by analysing the sequence upstream of where a sequenced read maps, artifactual reads could be identified with high specificity and sensitivity. Importantly, by removing such noise, replicated libraries made

using different barcodes correlated better to each other. By performing a differential expression analysis on the filtered data sets, on average, a 10-fold decrease in the number of tag clusters called as differentially expressed was observed. The removal of strand invasion artifacts, which contribute to an increased variance among samples, is crucial in differential expression analyses using digital gene expression data such as CAGE and RNA-Seq. However, it is ideal to design an experimental protocol that limits as much as possible biases that are a consequence of the barcoding strategy (19). We proposed a strategy, which we tested in the nanoCAGE method, by updating the sequence of the TS oligonucleotide by inserting a 6-nucleotide long standard spacer between the barcode and the ribo-guanosines. In addition, we chose a spacer sequence that had less potential for strand invasion. The main purpose of the common spacer is to ensure that any TS bias systematically affects all libraries in the same manner. We confirmed this by conducting a differential expression analysis on the libraries made with the common spacer, and indeed no tag clusters were detected as significantly differentially expressed (Supplementary Table S7). A potential downside to the common spacer approach is the addition of six more nucleotides to a sequenced read. However, when sequenced on a HiSeq instrument with the standard read length of 50 nucleotide, the resulting libraries can be aligned accurately with standard tools such as BWA (27), as 35 informative bases are remaining after removing the barcode, the spacer and the linker. Alternative strategies could be conceived, and the spacer could be extended or replaced by a random sequence (21,39).

We have described in this article an inherent problem that exists with the TS mechanism, which we could suppress by combining experimental and computational strategies; however, TS artifacts cannot be entirely abolished. What distinguishes the artifacts from *bona fide* full-length cDNAs is the presence of the remaining 5′ part of the mRNA as a possibly long tail in the mRNA/cDNA/oligonucleotide triplex. By using an experimental protocol called CAP Trapper (40), which is used in CAGE protocols, it is possible to identify this triplex due to the presence of a 7-methylguanosine cap, therefore accurately identifying transcriptional starting sites as opposed to strand invasion products. This concept of combining TS and CAP Trapper has been shown to produce multiplexed libraries that capture promoters with high fidelity (41). However, as this methodology requires additional preparation steps and is not favored in most TS protocols, where the starting material is limited such as in single cell analyses. Despite the remaining artifacts, our proposed strategy allows one to directly compare different samples, such as between normal and diseased samples. Given that TS is garnering interest again, as seen by the number of recent publications that have used TS, it is important that investigators become aware of TS artifacts. It is clear that more investigations are needed to fully understand the TS mechanism, especially with respect to the types of biases that could potentially be introduced.

DATA DEPOSITION

Sequence data have been deposited in the DNA Data Bank of Japan under accession code DRA000552.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–7 and Supplementary Figures 1 and 2.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Michiel de Hoon for assistance with the statistical analyses and Dr. Alistair Forrest for the rat samples.

FUNDING

The European Union Seventh Framework Programme under grant agreement [FP7-People-ITN-2008-238055] (‘BrainTrain’ project) (to P.C.); the Research Grant for RIKEN Omics Science Center from Ministry of Education, Culture, Sports, Science and Technology; the Grant-in-Aids for Scientific Research (A) No. 20241047 for nanoCAGE (to P.C.); the 7th Framework Programme Dopaminet Project from the EU (to P.C.); the Telethon grant [GGP10224] (to S.G.). Funding for open access charge: the European Union Seventh Framework Programme under grant agreement [FP7-People-ITN-2008-238055] (‘BrainTrain’ project) (to P.C.).

Conflict of interest statement. None declared.

REFERENCES

1. Baltimore,D. (1970) RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, **226**, 1209–1211.
2. Temin,H.M. and Mizutani,S. (1970) RNA-dependent DNA polymerase in virions of *Rous sarcoma* virus. *Nature*, **226**, 1211–1213.
3. Hirzmann,J., Luo,D., Hahnen,J. and Hobom,G. (1993) Determination of messenger RNA 5′-ends by reverse transcription of the cap structure. *Nucleic Acids Res.*, **21**, 3597–3598.
4. Ohtake,H., Ohtoko,K., Ishimaru,Y. and Kato,S. (2004) Determination of the capped site sequence of mRNA based on the detection of cap-dependent nucleotide addition using an anchor ligation method. *DNA Res.*, **11**, 305–309.
5. Schmidt,W.M. and Mueller,M.W. (1999) CapSelect: a highly sensitive method for 5′ CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Res.*, **27**, e31.
6. Zhu,Y.Y., Machleder,E.M., Chenchik,A., Li,R. and Siebert,P.D. (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques*, **30**, 892–897.
7. Matz,M., Shagin,D., Bogdanova,E., Britanova,O., Lukyanov,S., Diatchenko,L. and Chenchik,A. (1999) Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res.*, **27**, 1558–1560.
8. Cloonan,N., Forrest,A.R., Kolle,G., Gardiner,B.B., Faulkner,G.J., Brown,M.K., Taylor,D.F., Steptoe,A.L., Wani,S., Bethel,G. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.
9. Plessy,C., Bertin,N., Takahashi,H., Simone,R., Salimullah,M., Lassmann,T., Vitezic,M., Severin,J., Olivarius,S., Lazarevic,D. *et al.* (2010) Linking promoters to functional transcripts in small

samples with nanoCAGE and CAGEscan. *Nat. Methods*, **7**, 528–534.
10. Islam,S., Kjallquist,U., Moliner,A., Zajac,P., Fan,J.B., Lonnerberg,P. and Linnarsson,S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, **21**, 1160–1167.
11. Ko,J.H. and Lee,Y. (2006) RNA-conjugated template-switching RT-PCR method for generating an *Escherichia coli* cDNA library for small RNAs. *J. Microbiol. Methods*, **64**, 297–304.
12. Ramskold,D., Luo,S., Wang,Y.C., Li,R., Deng,Q., Faridani,O.R., Daniels,G.A., Khrebtukova,I., Loring,J.F., Laurent,L.C. *et al.* (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**, 777–782.
13. Maeda,N., Nishiyori,H., Nakamura,M., Kawazu,C., Murata,M., Sano,H., Hayashida,K., Fukuda,S., Tagami,M., Hasegawa,A. *et al.* (2008) Development of a DNA barcode tagging method for monitoring dynamic changes in gene expression by using an ultra high-throughput sequencer. *Biotechniques*, **45**, 95–97.
14. Islam,S., Kjallquist,U., Moliner,A., Zajac,P., Fan,J.B., Lonnerberg,P. and Linnarsson,S. (2012) Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nat. Protoc.*, **7**, 813–828.
15. Takahashi,H., Lassmann,T., Murata,M. and Carninci,P. (2012) 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.*, **7**, 542–561.
16. Salimullah,M., Sakai,M., Plessy,C. and Carninci,P. (2011) NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb. Protoc.*, **2011**, pdb prot5559.
17. Matsumura,H., Yoshida,K., Luo,S., Kimura,E., Fujibe,T., Albertyn,Z., Barrero,R.A., Kruger,D.H., Kahl,G., Schroth,G.P. *et al.* (2010) High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. *PLoS One*, **5**, e12010.
18. Kawano,M., Kawazu,C., Lizio,M., Kawaji,H., Carninci,P., Suzuki,H. and Hayashizaki,Y. (2010) Reduction of non-insert sequence reads by dimer eliminator LNA oligonucleotide for small RNA deep sequencing. *Biotechniques*, **49**, 751–755.
19. Alon,S., Vigneault,F., Eminaga,S., Christodoulou,D.C., Seidman,J.G., Church,G.M. and Eisenberg,E. (2011) Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Res.*, **21**, 1506–1511.
20. Jayaprakash,A.D., Jabado,O., Brown,B.D. and Sachidanandam,R. (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res.*, **39**, e141.
21. Kivioja,T., Vaharautio,A., Karlsson,K., Bonke,M., Enge,M., Linnarsson,S. and Taipale,J. (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72–74.
22. Goetz,J.J. and Trimarchi,J.M. (2012) Transcriptome sequencing of single cells with Smart-Seq. *Nat. Biotechnol.*, **30**, 763–765.
23. Fan,J.B., Chen,J., April,C.S., Fisher,J.S., Klotzle,B., Bibikova,M., Kaper,F., Ronaghi,M., Linnarsson,S., Ota,T. *et al.* (2012) Highly parallel genome-wide expression analysis of single mammalian cells. *PLoS One*, **7**, e30794.
24. Wang,D. and Bodovitz,S. (2010) Single cell analysis: the new frontier in 'omics'. *Trends Biotechnol.*, **28**, 281–290.
25. Kapteyn,J., He,R., McDowell,E.T. and Gang,D.R. (2010) Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples. *BMC Genomics*, **11**, 413.
26. Lassmann,T., Hayashizaki,Y. and Daub,C.O. (2009) TagDust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics*, **25**, 2839–2840.
27. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
28. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
29. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
30. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
31. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
32. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G., Frith,M.C. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
33. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
34. Bourgon,R., Gentleman,R. and Huber,W. (2010) Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci. USA*, **107**, 9546–9551.
35. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B*, **57**, 289–300.
36. Guttman,M., Garber,M., Levin,J.Z., Donaghey,J., Robinson,J., Adiconis,X., Fan,L., Koziol,M.J., Gnirke,A., Nusbaum,C. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.
37. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
38. Shiroguchi,K., Jia,T.Z., Sims,P.A. and Xie,X.S. (2012) Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl Acad. Sci. USA*, **109**, 1347–1352.
39. Konig,J., Zarnack,K., Rot,G., Curk,T., Kayikci,M., Zupan,B., Turner,D.J., Luscombe,N.M. and Ule,J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, **17**, 909–915.
40. Carninci,P., Kvam,C., Kitamura,A., Ohsumi,T., Okazaki,Y., Itoh,M., Kamiya,M., Shibata,K., Sasaki,N., Izawa,M. *et al.* (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, **37**, 327–336.
41. Batut,P.J., Dobin,A., Plessy,C., Carninci,P. and Gingeras,T.R. (2012) High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.*, **23**, 169–180.

Chapter 3

Role of small RNAs in DNA damage repair

LETTER

doi:10.1038/nature11179

Site-specific DICER and DROSHA RNA products control the DNA-damage response

Sofia Francia[1,2], Flavia Michelini[1], Alka Saxena[3], Dave Tang[3], Michiel de Hoon[3], Viviana Anelli[1]†, Marina Mione[1]†, Piero Carninci[3] & Fabrizio d'Adda di Fagagna[1,4]

Non-coding RNAs (ncRNAs) are involved in an increasingly recognized number of cellular events[1]. Some ncRNAs are processed by DICER and DROSHA RNases to give rise to small double-stranded RNAs involved in RNA interference (RNAi)[2]. The DNA-damage response (DDR) is a signalling pathway that originates from a DNA lesion and arrests cell proliferation[3]. So far, DICER and DROSHA RNA products have not been reported to control DDR activation. Here we show, in human, mouse and zebrafish, that DICER and DROSHA, but not downstream elements of the RNAi pathway, are necessary to activate the DDR upon exogenous DNA damage and oncogene-induced genotoxic stress, as studied by DDR foci formation and by checkpoint assays. DDR foci are sensitive to RNase A treatment, and DICER- and DROSHA-dependent RNA products are required to restore DDR foci in RNase-A-treated cells. Through RNA deep sequencing and the study of DDR activation at a single inducible DNA double-strand break, we demonstrate that DDR foci formation requires site-specific DICER- and DROSHA-dependent small RNAs, named DDRNAs, which act in a MRE11–RAD50–NBS1-complex-dependent manner (MRE11 also known as MRE11A; NBS1 also known as NBN). DDRNAs, either chemically synthesized or *in vitro* generated by DICER cleavage, are sufficient to restore the DDR in RNase-A-treated cells, also in the absence of other cellular RNAs. Our results describe an unanticipated direct role of a novel class of ncRNAs in the control of DDR activation at sites of DNA damage.

Mammalian genomes are pervasively transcribed, with most transcripts apparently not associated with coding functions[4,5]. An increasing number of ncRNAs have been shown to have a variety of relevant cellular functions, often with very low estimated expression levels[6–8]. DICER and DROSHA are two RNase type III enzymes that process ncRNA hairpin structures to generate small double-stranded RNAs[9] (see Supplementary Information).

Detection of a DNA double-strand break (DSB) triggers the kinase activity of ATM, which initiates a signalling cascade by phosphorylating the histone variant H2AX (γH2AX) at the DNA-damage site and recruiting additional DDR factors. This establishes a local self-feeding loop that leads to accumulation of upstream DDR factors in the form of cytologically detectable foci at damaged DNA sites[3,10]. The DDR has been considered to be a signalling cascade made up exclusively of proteins, with no direct contributions from RNA species to its activation.

Oncogene-induced senescence (OIS) is a non-proliferative state characterized by a sustained DDR[11] and senescence-associated heterochromatic foci (SAHF)[12]. Because ncRNAs participate in heterochromatin formation[13], we investigated whether they could control SAHF and OIS. We used small interfering RNAs (siRNAs) to knockdown DICER or DROSHA in OIS cells and monitored SAHF and cell-cycle progression. Knockdown of either DICER or DROSHA, as well as ATM as control[14], restored DNA replication and entry into mitosis (Supplementary Figs 1 and 2); we did not detect overt SAHF changes, however (Supplementary Fig. 3a, b). Instead, we observed that DICER or DROSHA inactivation significantly reduced the number of cells positive for DDR foci containing 53BP1, the autophosphorylated form of ATM (pATM) and the phosphorylated substrates of ATM and ATR (pS/TQ), but not γH2AX, without decreasing the expression of proteins involved in the DDR (Supplementary Fig. 3a–c). Importantly, the simultaneous inactivation of all three GW182-like proteins, TNRC6A, B and C, essential for the translational inhibition mediated by microRNAs (miRNAs; canonical DICER and DROSHA products involved in RNAi)[15], does not affect DDR foci formation (Supplementary Fig. 4).

We next asked whether DICER or DROSHA inactivation also affects ionizing-radiation-induced DDR activation. We transiently inactivated DICER or DROSHA by siRNA in human normal fibroblasts (HNFs), exposed cells to ionizing radiation, and monitored DDR foci. We observed that a few hours after exposure to ionizing radiation, DICER or DROSHA inactivation impairs the formation of pATM, pS/TQ and MDC1, but not γH2AX, foci without decreasing their protein levels (Fig. 1a, b and Supplementary Fig. 5). Furthermore, at an earlier time point (10 min) after ionizing radiation, 53BP1 foci were significantly reduced (Supplementary Fig. 6a). Using an RNAi-resistant form of DICER in DICER knockdown cells, we observed that re-expression of wild-type DICER, but not of a DICER endonuclease mutant (DICER44ab)[16], rescues DDR foci formation (Supplementary Fig. 6b–d). The simultaneous knockdown of TNRC6A, B and C, or DICER has a comparable impact on a reporter system specific for miRNA-dependent translational repression[17], but only DICER inactivation reduces DDR foci formation (Supplementary Fig. 7). To confirm further the involvement of DICER in DDR activation, we used a cell line carrying a hypomorphic allele of *DICER* (*DICER*^{exon5}) defective in miRNA maturation[18]. In *DICER*^{exon5}-irradiated cells, pATM, pS/TQ and MDC1, but not γH2AX, foci formation is impaired without a decrease in their protein levels, and 53BP1 foci formation is delayed compared to the DICER wild-type parental cell line (Supplementary Fig. 8). These defects could be reversed by the re-expression of wild-type DICER but not of the mutant form DICER44ab (Supplementary Fig. 9). By immunoblotting, we confirmed that ATM autophosphorylation is reduced in DICER or DROSHA knockdown HNFs, and in *DICER*^{exon5} cell lines (Supplementary Fig. 10). These results indicate that DICER and DROSHA RNA products control DDR activation and act independently from canonical miRNA-mediated translational repression mechanisms.

DDR signalling enforces cell-cycle arrest at the G1/S and G2/M checkpoints[3]. We observed that DNA-damage-induced checkpoints were impaired in DICER- or DROSHA-inactivated cells and that

[1]IFOM Foundation - FIRC Institute of Molecular Oncology Foundation, Via Adamello 16, 20139 Milan, Italy. [2]Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia, at the IFOM-IEO Campus, Via Adamello 16, 20139 Milan, Italy. [3]Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [4]Istituto di Genetica Molecolare, Consiglio Nazionale delle Ricerche, Pavia 27100, Italy. †Present addresses: Departments of Surgery and Medicine, Weill Cornell Medical College and New York Presbyterian Hospital, 1300 York Avenue, New York, New York 10065, USA (V.A.); Institute of Toxicology and Genetics, Karlsruhe Institute of Technology, 76344 Karlsruhe, Germany (M.M.).

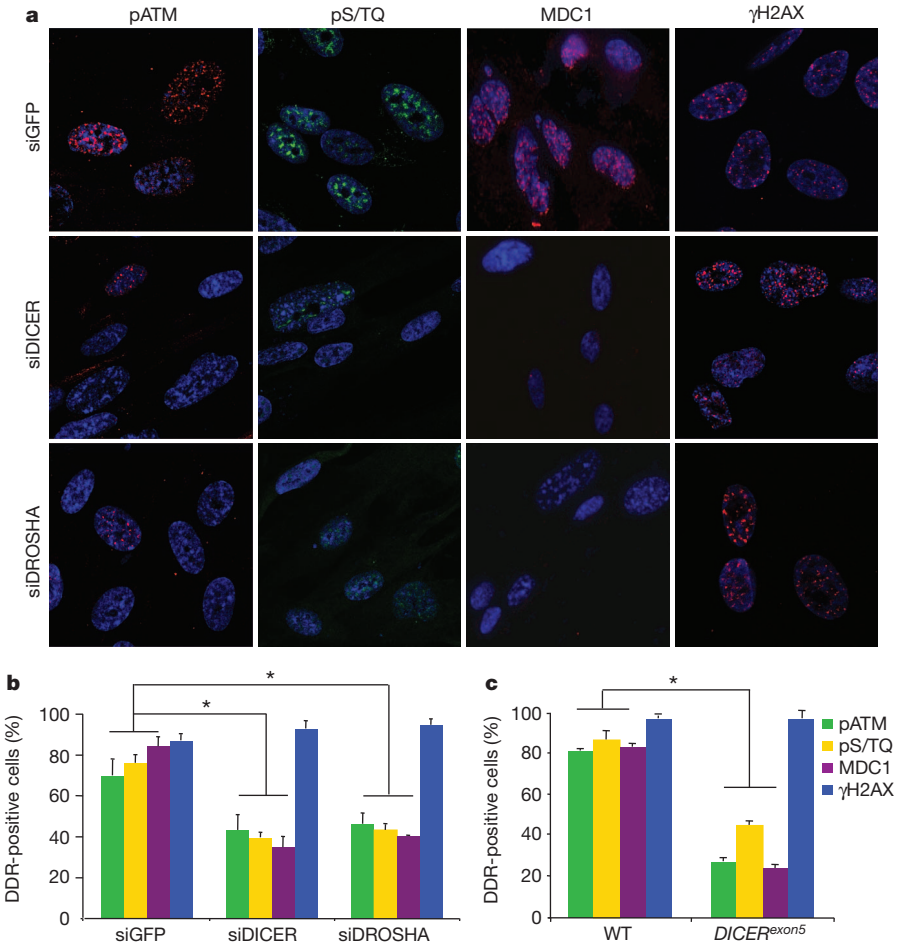

Figure 1 | DICER or DROSHA inactivation impairs DDR foci formation in irradiated cells. **a**, DICER or DROSHA knockdown WI-38 cells were irradiated (10 Gy) and fixed 7 h later. Original magnification, ×250. **b**, Histogram shows the percentage of cells positive for pATM, pS/TQ, MDC1 and γH2AX foci. **c**, Wild-type (WT) and *DICER*^{exon5} cells were irradiated (2 Gy) and fixed 2 h later. Histogram shows the percentage of cells positive for pATM, pS/TQ, MDC1 and γH2AX foci. Error bars indicate s.e.m. ($n \geq 3$). Differences are statistically significant (*P value < 0.01).

wild-type DICER re-expression in *DICER*^{exon5} cells restores checkpoint functions whereas two independent mutant forms of DICER fail to do so (Supplementary Figs 11–13). Thus, DICER and DROSHA are required for DNA-damage-induced checkpoint enforcement.

To test the role of DICER in DDR activation in a living organism, we inactivated it by morpholino antisense oligonucleotide injection in *Danio rerio* (zebrafish) larvae[19]. Such Dicer inactivation results in a marked impairment of pAtm and zebrafish γH2AX accumulation in irradiated larvae as detected both by immunostaining and immunoblotting of untreated or Dicer morpholino-injected larvae and of chimaeric animals (Supplementary Figs 14 and 15).

Previous reports have shown that mammalian cells can withstand transient membrane permeabilization and RNase A treatment, enabling investigation of the contribution of RNA to heterochromatin organization and 53BP1 association to chromatin[20,21]. We used this approach to address the direct contribution of DICER and DROSHA RNA products in DDR activation. Irradiated HeLa cells were permeabilized and treated with RNase A, leading to degradation of all RNAs, without affecting protein levels (Supplementary Fig. 16a). We observed that 53BP1, pATM, pS/TQ and MDC1 foci become markedly reduced in number and intensity upon RNA degradation whereas, similarly to DICER- or DROSHA-inactivated cells, γH2AX is unaffected (Fig. 2a and Supplementary Fig. 16b). Notably, 53BP1, MDC1 and γH2AX triple staining shows that RNA degradation reduces 53BP1 and MDC1 accumulation at unperturbed γH2AX foci (Supplementary Fig. 16c). When RNase A is inhibited, DDR foci progressively reappear within minutes and α-amanitin prevents this (Supplementary Fig. 17a, b), suggesting that DDR foci stability is RNA polymerase II dependent.

We tested whether DDR foci can reform upon addition of exogenous RNA to RNase-A-treated cells. We observed that DDR foci robustly reform in RNase-A-treated cells following their incubation with total RNA purified from the same cells, but not with transfer RNA (tRNA) control (Fig. 2b–d). Similar conclusions were reached using an inducible form of PpoI and AsiSI site-specific endonucleases[22,23] (data not shown).

Next, we attempted to characterize the length of the RNA species involved in DDR foci reformation, which we refer to as DDRNAs. We observed that an RNA fraction enriched by chromatography for species <200 nucleotides was sufficient to restore DDR foci (Supplementary Fig. 17c–e). To attain better size separation, we resolved total RNA on a polyacrylamide gel and recovered RNA fractions of different lengths (Supplementary Fig. 17f, g). Using equal amounts of each fraction, we observed that only the 20–35-nucleotide fraction could restore DDR foci (Fig. 2b), consistent with the size range of DICER and DROSHA RNA products.

To test the hypothesis that DDRNAs are DICER and DROSHA products, we tested DDR foci restoration with total RNA extracted from wild-type or *DICER*^{exon5} cells. Although RNA extracted from wild-type cells restores pATM, pS/TQ and 53BP1 foci, RNA from *DICER*^{exon5} cells does not (Fig. 2c, d). Importantly, RNA from

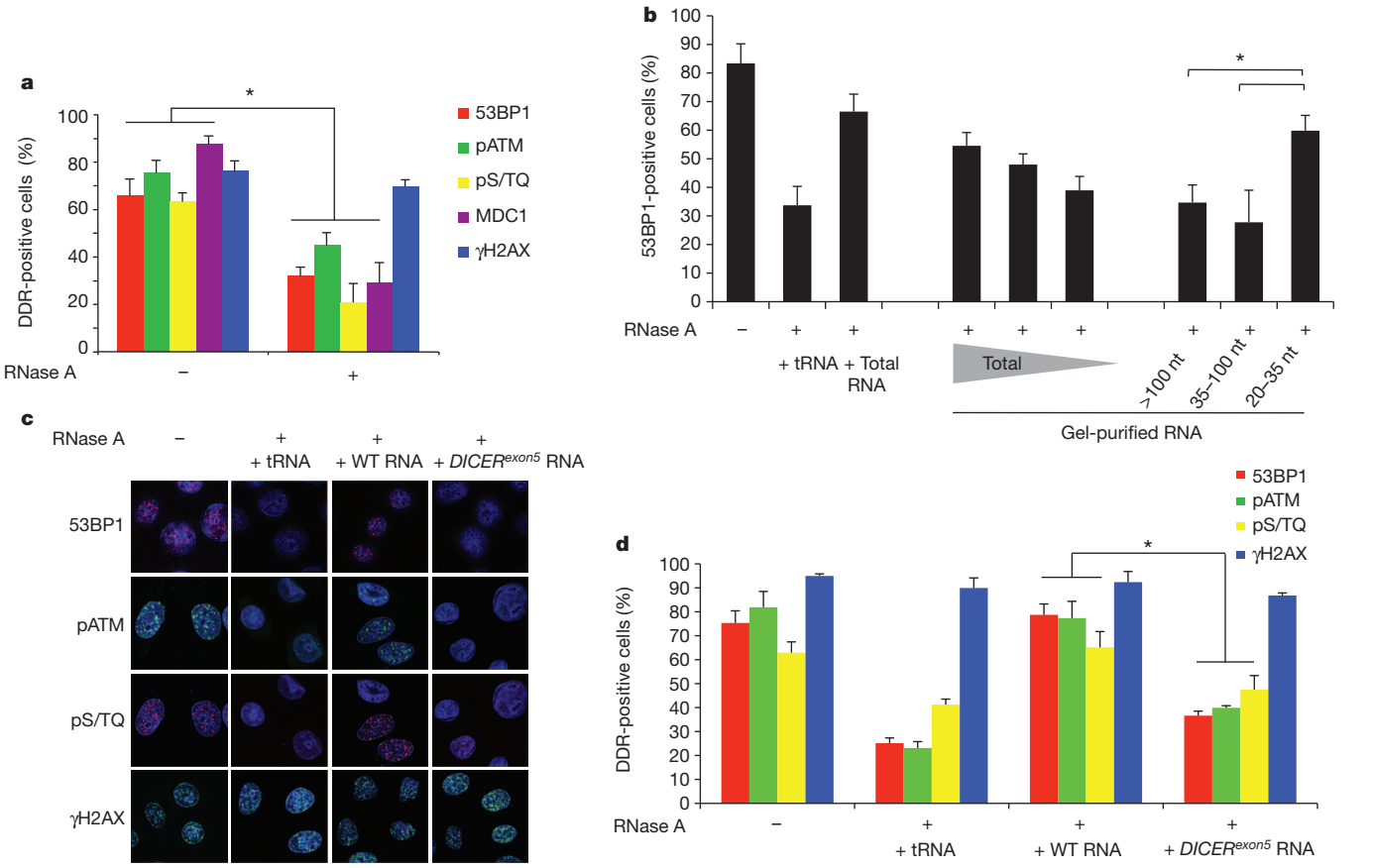

Figure 2 | Irradiation-induced DDR foci are sensitive to RNase A treatment and are restored by small and DICER-dependent RNAs. **a**, Irradiated HeLa cells (2 Gy) were treated with PBS (−) or RNase A (+) and probed for 53BP1, pATM, pS/TQ, MDC1 and γH2AX foci. Histogram shows the percentage of cells positive for DDR foci. **b**, 100, 50 or 20 ng of gel-extracted total RNA and 50 ng of RNA extracted from each gel fraction (>100, 35–100 and 20–35 nucleotides (nt)) were used for DDR foci reconstitution after RNase treatment. **c**, 53BP1, pS/TQ and pATM foci are restored in RNase-treated cells when incubated with RNA of wild-type cells but not with RNA of *DICER*^{exon5} cells or tRNA. Original magnification, ×350. **d**, Histogram shows the percentage of cells positive for DDR foci. Error bars indicate s.e.m. ($n \geq 3$). Differences are statistically significant (*P value < 0.01).

DICER^{exon5} cells re-expressing wild-type, but not endonuclease-mutant, DICER allows DDR foci reformation (Supplementary Fig. 18a, b). These results were reproduced using RNA extracted from cells transiently knocked down for DICER or DROSHA (Supplementary Fig. 18c–f).

Ionizing radiation induces DNA lesions that are heterogeneous in nature and random in their genomic location. To reduce this complexity, we studied a single DSB at a defined and traceable genomic locus. We therefore took advantage of NIH2/4 mouse cells carrying an integrated copy of the I-SceI restriction site flanked by arrays of Lac- or Tet-operator repeats at either sites[24]. In this cell line, the expression of the I-SceI restriction enzyme together with the fluorescent protein Cherry-Lac-repressor allows the visualization of a site-specific DDR focus that overlaps with a focal Cherry-Lac signal (cut NIH2/4 cells). No DDR focus formation is observed overlapping with the Cherry-Lac signal in the absence of I-SceI expression (uncut NIH2/4 cells). Also in this system, RNase A treatment causes the disappearance of the 53BP1, but not the γH2AX, focus at the I-SceI-induced DSB; total RNA addition from cut cells restores 53BP1 focus formation in a dose-dependent manner (Fig. 3a, b). Therefore, a DDR focus generated on a defined DSB can disassemble and reassemble in an RNA-dependent manner.

To determine whether DDRNAs are generated at the damaged locus or elsewhere in the genome, we took advantage of the fact that the I-SceI-induced DSB is generated within an integrated exogenous sequence, which is not present in the parental cell line. As RNAs extracted from NIH2/4 or parental cells are expected to differ only in the potential presence of RNA transcripts generated at the locus, we used these two RNA preparations to attempt to restore 53BP1 focus formation at the I-SceI-induced DSB in RNase-A-treated cells. The formation of the 53BP1 focus was efficiently recovered only by RNA purified from NIH2/4 cells and not from parental cells (Fig. 3c), indicating that DDRNAs originate from the damaged genomic locus.

The MRE11–RAD50–NBS1 (MRN) complex is necessary for ATM activation[25], and pATM and MRE11 foci formation is sensitive to RNase A treatment in the NIH2/4 cell system (Supplementary Fig. 19a, b). To probe the molecular mechanisms by which RNA modulates DDR focus formation, we used a specific MRN inhibitor[26], mirin, which prevents ATM activation also in the NIH2/4 system (Supplementary Fig. 19d). In the presence of mirin, NIH2/4 RNA is unable to restore 53BP1 or pATM focus formation (Fig. 3d, e), indicating that DDRNAs act in a MRN-dependent manner.

To detect potential short RNAs originating from the integrated locus, we deep-sequenced libraries generated from short (<200 nucleotides) nuclear RNAs of cut or uncut NIH2/4 cells, as well as from parental cells expressing I-SceI as negative control. Sequencing revealed short transcripts arising from the exogenous locus (Supplementary Fig. 20a–e), 47 reads in cut cells, 20 reads in uncut cells and none in parental cells, indicating that even an exogenous integrated locus lacking mammalian transcriptional regulatory elements is transcribed and can generate small RNAs.

To test whether the identified locus-specific small RNAs are biologically active and have a causal role in DDR activation, we chemically synthesized four potential pairs among the sequences obtained and used them to attempt to restore the DDR focus in RNase-A-treated

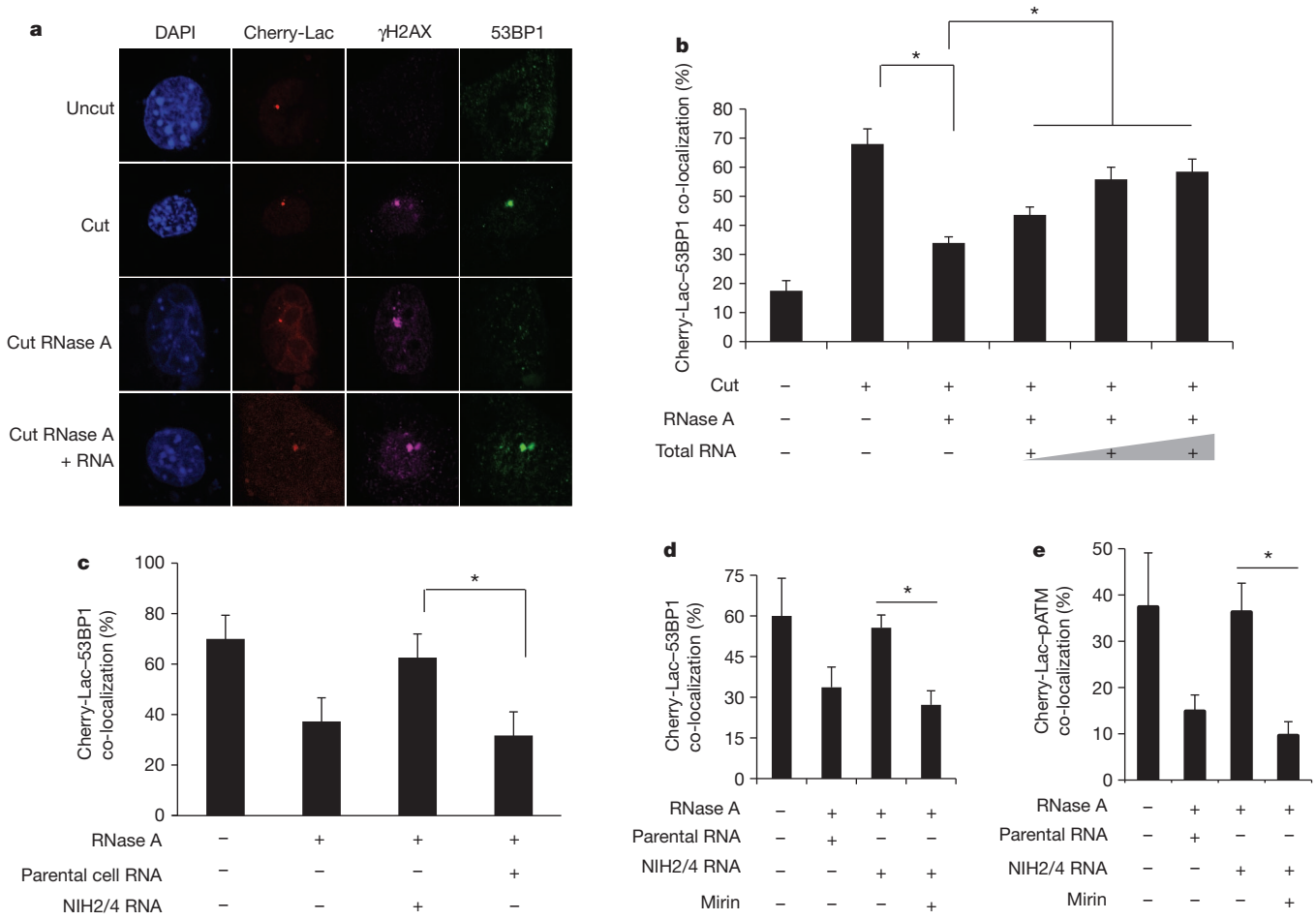

Figure 3 | Site-specific DDR focus formation is RNase A sensitive and can be restored by site-specific RNA in a MRN-dependent manner. **a**, Cut NIH2/4 cells display a 53BP1 and γH2AX focus co-localizing with a Cherry-Lac focus. 53BP1, but not γH2AX, focus is sensitive to RNase A and is restored by incubation with total RNA. DAPI, 4′,6-diamidino-2-phenylindole. Original magnification, ×450. **b**, Histogram shows the percentage of cells in which 53BP1 and Cherry-Lac foci co-localize. Addition of 50, 200 or 800 ng of RNA purified from cut NIH2/4 rescues 53BP1 foci formation in a dose-dependent manner. **c**, RNA purified from cut NIH2/4 restores 53BP1 focus whereas RNA from parental cells expressing I-SceI does not. **d**, **e**, RNase-A-treated cut NIH2/4 cells were incubated with RNA from cut NIH2/4 cells, or parental ones, to test 53BP1 or pATM focus reformation in the presence of the MRN inhibitor mirin (100 μM). Histogram shows the percentage of cells positive for a DDR focus. Error bars indicate s.e.m. ($n \geq 3$). Differences are statistically significant (*P value < 0.05).

cells. Notably, we observed that addition of locus-specific synthetic RNAs, but not equal amounts of control RNAs, triggers site-specific 53BP1 focus reformation over a large range of concentrations in the presence, but also in the absence, of total RNA from parental cells (Fig. 4a and Supplementary Fig. 20f). To show further the biological activity of RNAs processed by DICER, we *in vitro* transcribed both strands of the sequence spanning the locus, or a control one, and processed the resulting RNAs with recombinant DICER. *In vitro*-generated locus-specific DICER RNA products, but not control RNAs, allowed DDR focus reformation in RNase-A-treated cells even in the absence of parental RNA (Fig. 4b and Supplementary Fig. 20g, h). Overall, these results indicate that DDRNAs are small RNAs with the sequence of the damaged locus, which have a direct role in DDR activation.

To investigate the biogenesis of such RNAs *in vivo*, we performed deeper sequencing of small nuclear RNAs from cut and uncut wild-type as well as DICER or DROSHA knockdown NIH2/4 cells (Supplementary Fig. 21). As expected, DICER or DROSHA knockdown significantly reduced reads mapping to the known miRNAs (Supplementary Fig. 22). Our statistical analyses revealed that the percentage of 22–23-nucleotide RNAs arising from the locus significantly increases in the wild-type cut sample compared to the uncut one and that DICER inactivation significantly reduces it (Supplementary Fig. 23a, b); the detectable decrease in DROSHA-inactivated cells did not reach statistical significance. Because the fraction of 22–23-nucleotide RNAs from the locus is significantly higher with respect to that of non-miRNA genomic loci, the RNAs detected are very unlikely to be random degradation products (Supplementary Fig. 23c). Finally, 22–23-nucleotide RNAs at the locus tend to have an A/U at their 5′ and a G at their 3′ end (Supplementary Fig. 23d), a nucleotide bias significantly different from the originating locus and from the rest of the genome.

In summary, we demonstrate that different sources of DNA damage, including oncogenic stress, ionizing radiation and site-specific endonucleases, activate the DDR in a manner dependent on DDRNAs, which are DICER- and DROSHA-dependent RNA products with the sequence of the damaged site. DDRNAs control DDR foci formation and maintenance, checkpoint enforcement and cellular senescence in cultured human and mouse cells and in different cell types in living zebrafish larvae. They act differently from canonical miRNAs, as inferred by their demonstrated biological activity independent of other RNAs and of GW182-like proteins.

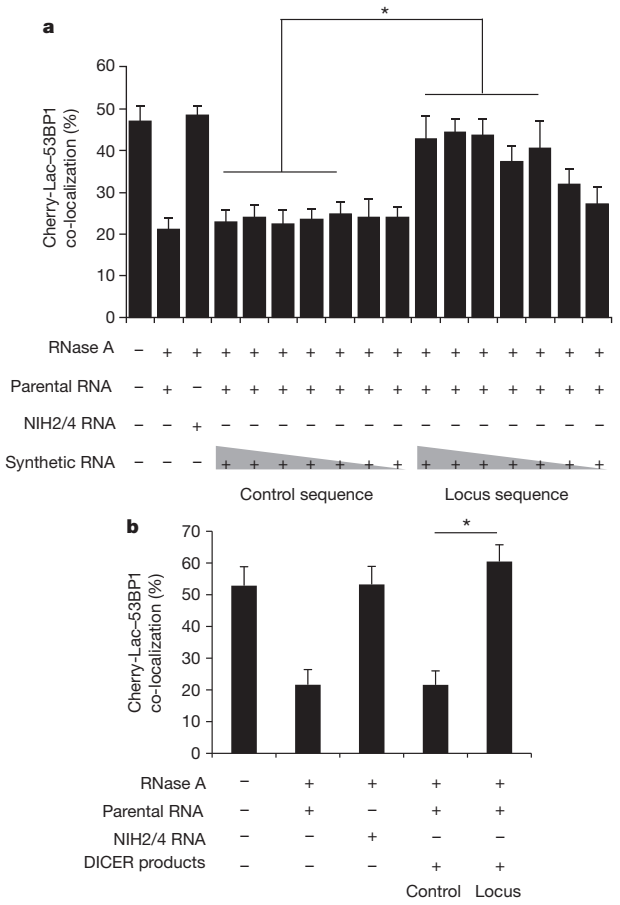

Figure 4 | Chemically synthesized small RNAs and *in vitro*-generated DICER RNA products are sufficient to restore DDR focus formation in RNase-A-treated cells in a sequence-specific manner. **a**, Chemically synthesized RNA oligonucleotides were annealed and were tested to restore DDR focus formation in RNase-A-treated cut NIH2/4 cells. Mixed with a constant amount (800 ng) of parental cell RNA, a concentration range (1 ng μl^{-1} to 1 fg μl^{-1}, tenfold dilution steps) of locus-specific or GFP RNAs was used. Locus-specific synthetic RNAs (down to 100 fg μl^{-1}) allow site-specific DDR activation. **b**, Small double-stranded RNAs generated by recombinant DICER were tested to restore DDR focus formation in RNase-A-treated cut NIH2/4 cells. 1 ng μl^{-1} RNA was tested mixed with 800 ng of parental cell RNA. Locus-specific DICER RNAs, but not control RNAs, allow site-specific DDR activation. Histograms show the percentage of cells positive for DDR focus. Error bars indicate s.e.m. ($n \geq 3$). Differences are statistically significant (*P value < 0.05).

METHODS SUMMARY

Details of cell cultures, plasmids, siRNAs and antibodies used, as well as descriptions of methods for immunofluorescence, immunoblotting, checkpoint assays, real-time quantitative polymerase chain reaction (PCR), zebrafish injection and transplantation, RNase A treatment, small RNA extraction and purification from gel, RNA sequencing and statistical analyses are provided in Methods.

Full Methods and any associated references are available in the online version of the paper.

Received 8 February 2010; accepted 4 May 2012.
Published online 23 May 2012.

1. Esteller, M. Non-coding RNAs in human disease. *Nature Rev. Genet.* **12,** 861–874 (2011).
2. Krol, J., Loedige, I. & Filipowicz, W. The widespread regulation of microRNA biogenesis, function and decay. *Nature Rev. Genet.* **11,** 597–610 (2010).
3. Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* **461,** 1071–1078 (2009).
4. Clark, M. B. *et al.* The reality of pervasive transcription. *PLoS Biol.* **9,** e1000625 (2011).
5. Wilusz, J. E., Sunwoo, H. & Spector, D. L. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* **23,** 1494–1504 (2009).
6. Wang, X. *et al.* Induced ncRNAs allosterically modify RNA-binding proteins *in cis* to inhibit transcription. *Nature* **454,** 126–130 (2008).
7. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322,** 750–756 (2008).
8. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nature Rev. Genet.* **10,** 155–159 (2009).
9. Kim, V. N., Han, J. & Siomi, M. C. Biogenesis of small RNAs in animals. *Nature Rev. Mol. Cell Biol.* **10,** 126–139 (2009).
10. Lukas, J., Lukas, C. & Bartek, J. More than just a focus: the chromatin response to DNA damage and its role in genome integrity maintenance. *Nature Cell Biol.* **13,** 1161–1169 (2011).
11. d'Adda di Fagagna, F. Living on a break: cellular senescence as a DNA-damage response. *Nature Rev. Cancer* **8,** 512–522 (2008).
12. Narita, M. *et al.* Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell* **113,** 703–716 (2003).
13. White, S. A. & Allshire, R. C. RNAi-mediated chromatin silencing in fission yeast. *Curr. Top. Microbiol. Immunol.* **320,** 157–183 (2008).
14. Di Micco, R. *et al.* Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. *Nature* **444,** 638–642 (2006).
15. Tritschler, F., Huntzinger, E. & Izaurralde, E. Role of GW182 proteins and PABPC1 in the miRNA pathway: a sense of deja vu. *Nature Rev. Mol. Cell Biol.* **11,** 379–384 (2010).
16. Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E. & Filipowicz, W. Single processing center models for human Dicer and bacterial RNase III. *Cell* **118,** 57–68 (2004).
17. Nicoli, S. *et al.* MicroRNA-mediated integration of haemodynamics and Vegf signalling during angiogenesis. *Nature* **464,** 1196–1200 (2010).
18. Cummins, J. M. *et al.* The colorectal microRNAome. *Proc. Natl Acad. Sci. USA* **103,** 3687–3692 (2006).
19. Wienholds, E. *et al.* MicroRNA expression in zebrafish embryonic development. *Science* **309,** 310–311 (2005).
20. Maison, C. *et al.* Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nature Genet.* **30,** 329–334 (2002).
21. Pryde, F. *et al.* 53BP1 exchanges slowly at the sites of DNA damage and appears to require RNA for its association with chromatin. *J. Cell Sci.* **118,** 2043–2055 (2005).
22. Berkovich, E., Monnat, R. J. Jr & Kastan, M. B. Roles of ATM and NBS1 in chromatin structure modulation and DNA double-strand break repair. *Nature Cell Biol.* **9,** 683–690 (2007).
23. Iacovoni, J. S. *et al.* High-resolution profiling of γH2AX around DNA double strand breaks in the mammalian genome. *EMBO J.* **29,** 1446–1457 (2010).
24. Soutoglou, E. *et al.* Positional stability of single double-strand breaks in mammalian cells. *Nature Cell Biol.* **9,** 675–682 (2007).
25. Stracker, T. H. & Petrini, J. H. The MRE11 complex: starting from the ends. *Nature Rev. Mol. Cell Biol.* **12,** 90–103 (2011).
26. Dupré, A. *et al.* A forward chemical genetic screen reveals an inhibitor of the Mre11–Rad50–Nbs1 complex. *Nature Chem. Biol.* **4,** 119–125 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. Soutoglou, W. C. Hahn, M. Kastan, V. Orlando, R. Shiekhattar, J. Amatruda, T. Halazonetis, E. Dejana, P. Ng and F. Nicassio for sharing reagents, M. Fumagalli and F. Rossiello for reading the manuscript, M. Dobreva, V. Matti and F. Pezzimenti for technical support, G. D'Ario for help with statistical analyses, B. Amati, M. Foiani, V. Costanzo and F.d.A.d.F. group members for help and discussions. The F.d.A.d.F. laboratory was supported by Fondazione Italiana Ricerca Sul Cancro (FIRC), Associazione Italiana Ricerca sul Cancro (AIRC) European Community's 7th Framework Programme (FP7/2007-2013) under grant agreement no. 202230, acronym "GENINCA", HFSP, AICR, the EMBO Young Investigator Program. The initial part of this project was supported by Telethon grant no. GGP08183. P.C. was supported by 7th Framework of the European Union commission to the Dopaminet consortium, a Grant-in-Aids for Scientific Research (A) no. 20241047, Funding Program for the Next Generation World-Leading Researchers (NEXT Program) to P.C. and a Research Grant to RIKEN Omics Science Center from MEXT. S.F. is supported by Center for Genomic Science of IIT@SEMM (Scuola Europea di Medicina Molecolare) and AIRC. M.M. was supported by Cariplo (grant no. 2007-5500) and AIRC. A.S. is supported by a JSPS fellowship P09745 and grant in aid by JSPS, and D.T. is supported by the European Union 7th Framework Programme under grant agreement FP7-People-ITN-2008-238055 ("BrainTrain" project) to P.C.

Author Contributions A.S., D.T. and P.C. planned, generated and analysed the genomics data presented in Supplementary Figs 20a–e, 21, 22b and 23. M.d.H. performed statistical analysis of the genomics data. A.S. and P.C. also edited the manuscript. M.M. and V.A. generated the data presented in Supplementary Figs 14 and 15. F.M. generated the data shown in Figs 2b, 3d, e, 4b and Supplementary Figs 2b, e, 3e, 4b, 5f, g, 6b–d, 7d, 9, 13d–f, 14d, f, 17f, g, 18a, b, 19, 20g, h and 22a and generated RNA for deep sequencing; contributed to: Supplementary Figs 16a, 5d, e, 11c, d and edited the manuscript. S.F. generated the data shown in remaining figures, contributed to experimental design and edited the manuscript. F.d.A.d.F. conceived the study, designed the experiments and wrote the manuscript.

Author Information Sequence data have been deposited in the DNA Data Bank of Japan under accession code DRA000540. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to F.d.A.d.F. (fabrizio.dadda@ifom-ieo-campus.it).

METHODS

Cultured cells. Early-passage WI-38 cells (ATCC) were grown under standard tissue culture conditions (37 °C, 5% CO_2) in MEM supplemented with 10% fetal bovine serum, 1% L-glutamine, 1% non-essential aminoacids, 1% Na pyruvate. HeLa, Phoenix ecotrophic and HEK293T cell lines were grown under standard tissue culture conditions (37 °C, 5% CO_2) in DMEM, supplemented with 10% fetal bovine serum, 1% glutamine, 1% penicillin/streptomycin. *DICER*^{exon5} colon cancer cell lines[18] were cultured in McCoy's 5A medium plus 10% fetal calf serum, 1% penicillin/streptomycin. NIH2/4 cells[24] were grown in DMEM, supplemented with 10% fetal bovine serum, 1% glutamine, gentamicine (40 μg ml^{-1}) and hygromycin (400 μg ml^{-1}).

H-RasV12-overexpressing senescent BJ cells were generated as described previously[14]. BrdU incorporation assays were carried out at least 1 week after cultures had fully entered the senescent state, as determined by ceased proliferation, DDR activation and SAHF formation. Ionizing radiation was induced by a high-voltage X-ray generator tube (Faxitron X-Ray Corporation). In general, WI-38 cells were exposed to 5 Gy and transformed cells (RKO, HCT116 and HeLa) to 2 Gy for the DDR foci formation studies. 5 Gy were used for the G2/M checkpoint assays and 10 Gy for the G1/S checkpoint assays.

Cherry-Lac and I-SceI-restriction endonuclease expressing vectors were transfected by lipofectamine 2000 (Invitrogen) in a ratio of 3:1. Sixteen hours after transfection around 70% of the cells were scored positive for DDR markers in the Lac array. For generation of DICER and DROSHA knockdown, NIH2/4 cells were infected with lentiviral particles carrying pLKO.1, shDICER or shDROSHA vectors. After 48 h cells were superinfected with Adeno Empty Vector (gift from E. Dejana) or Adeno I-SceI (gift from P. Ng). Nuclei were isolated the day after the adenoviral infection.

Antibodies. Mouse anti-γH2AX, anti-H3K9me3, rabbit polyclonal anti-pH3 (Upstate Biotechnology); anti-pS/TQ (Cell Signaling Technology); anti-H2AX, anti-H3 and anti-DICER (13D6) (Abcam); rabbit polyclonal anti-53BP1 (Novus Biological); mouse monoclonal anti-53BP1 (gift from T. Halazonetis); anti-MRE11 (gift from S. P. Jackson); anti-BrdU (Becton Dickinson); rabbit polyclonal anti-MCM2 (gift from M. Melixetian); anti-MRE11 rabbit polyclonal raised against recombinant MRE11; anti-pATM (Rockland); mouse monoclonal anti-ATM and anti-MDC1 (SIGMA); anti-Lamin A/C (Santa Cruz); anti-vinculin (clone hVIN-1), anti-β-tubulin (clone AA2) and anti-Flag M2 monoclonal antibodies (Sigma).

Indirect immunofluorescence. Cells were grown on poly-D-lysinated coverslips (poly-D-lysine was used at 50 μg ml^{-1} final concentration) and plated (15–20 × 10^3 cells per cover) 1 day before staining. DDR and BrdU staining was performed as described previously[14]. Cells were fixed in 4% paraformaldehyde or methanol:acetone 1:1. NIH2/4 mouse cells were fixed by 4% paraformaldehyde as described previously[24]. Images were acquired using a wide field Olympus Biosystems Microscope BX71 and the analySIS or the MetaMorph software (Soft Imaging System GmbH). Comparative immunofluorescence analyses were performed in parallel with identical acquisition parameters; at least 100 cells were screened for each antigen. Cells with more than two DDR foci were scored positive. Confocal sections were obtained with a Leica TCS SP2 or AOBS confocal laser microscope by sequential scanning.

Plasmids. DICER–Flag, DICER44ab–Flag and DICER110ab–Flag were a gift from R. Shiekhattar. DICER110ab–Flag and DICER44ab–Flag double mutants carry two amino acid substitutions in the RNase III domains of DICER (Asp 1320 Ala and Asp 1709 Ala for 44ab, and Glu 1652 Ala and Glu 1813 Ala for 110ab mutant; both mutants were reported to be deficient in endonuclease activity[16]). pLKO.1 shDICER-expressing vector was a gift from W. C. Hahn. Short hairpin sequence for DICER is: CCGGCCACACATCTTCAAGACTTAACT CGAGTTAAGTCTTGAAGATGTGTGGTTTTTG. pRETROSUPER shp53 was as described previously[14]. Short hairpin sequence for p53 was: AGTAGATTAC CACTGGAGTCTT. Cherry-Lac-repressor and I-SceI-restriction endonuclease expressing vectors were gifts from E. Soutoglou[24]. shRNA against mouse DICER- and DROSHA-expressing vectors were a gift from W. C. Hahn. shRNA for mouse DICER: CCGGGCCTCACTTGACCTGAAGTATCTCGAGA TACTTCAGGTCAAGTGAGGCTTTTT. shRNA for mouse DROSHA: CCGG CCTGGAATATGTCCACACTTTCTCGAGAAAGTGTGGACATATTCCAGG TTTTTG.

siRNA. The DHARMACON siGENOME SMARTpool siRNA oligonucleotide sequences for human 53BP1, ATM, DICER, DROSHA were as follows. 53BP1: GAGAGCAGAUGAUCCUUUA; GGACAAGUCUCUCAGCUAU; GAUAUC AGC UUAGACAAUU; GGACAGAACCCGCAGAUUU. ATM: GAAUGUU GCUUUCUGAAUU; AGACAGAAUUCCCAAAUAA; UAUAUCACC UGUU UGUUAG; AGGAGGAGCUUGGGCCUUU. DICER: UAAAGUAGCUGGAA UGAUG; GGAAGAGGCUGACUAUGAA; GAAUAUCGAUCCUAUGUUC; GAUCCUAUGUUCAAUCUAA. DROSHA: CAACAUAGACUACACGAUU; CCAACUCCCUCGAGGAUUA; GGCCAACUGUUAUAGAAUA; GAGUAG GCUUCGUGACUUA.

The DHARMACON siGENOME si RNA sequences for human TNRC6A, B and C were as follows. GW182/TNRC6A: GAAAUGCUCUGGUCCGCUA; GCCUAAAUAUUGGUGAUUA. TNRC6B: GCACUGCCCUGAUCCGAUA; GGAAUUAAGUCGUCGUCAU. TNRC6C: CUAUUAACCUCGCCAAUUA; GGUAAGUCCUCCAUUGAUG.

siRNA against human DICER 3′ UTR: CCGUGAAAGUUUAACGUUU. siRNA against GFP: AACACUUGUCACUACUUUCUC. siRNA against luciferase: CAUUCUAUCCUCUAGAGGAUGdTdT; dTdTGUAAGAUAGGAGAUC UCCUAC.

siRNAs were transfected by Oligofectamine or Lipofectamine RNAi Max (Invitrogen) at a final concentration of 200 nM in OIS cells and 100 nM in HNFs. In the siRNA titration experiment OIS cells were transfected in parallel with 20 nM and 200 nM siRNA oligonucleotides. For siRNA transfection with deconvolved siRNA oligonucleotides we used 50 nM for smart pools and 12.5 nM for deconvolved siRNAs.

Real-time quantitative PCR. Total RNA was isolated from cells using TRIzol (Invitrogen) or RNAeasy kit (Qiagen) according to the manufacturer's instructions, and treated with DNase before reverse transcription. For small RNA isolation we used the *mir*Vana miRNA Isolation Kit (Ambion). cDNA was generated using the Superscript II Reverse Transcriptase (Invitrogen) and used as a template in real-time quantitative PCR analysis. TaqMan MicroRNA Assays (Applied Biosystems) were used for the evaluation of mature miR-21 and rnu44 and rnu19 expression levels (assay numbers 000397, 001094 and 001003). 18S or β-actin was used as a control gene for normalization. Real-time quantitative PCR reactions were performed on an Applied Biosystems ABI Prism 7900HT Sequence Detection System or on a Roche LightCycler 480 Sequence Detection System. The reactions were prepared using SyBR Green reaction mix from Roche. Ribosomal protein P0 (RPP0) was used as a human and mouse control gene for normalization.

Primer sequences for real-time quantitative PCR. RPP0: TTCATTGTGGGAG CAGAC (forward), CAGCAGTTTCTCCAGAGC (reverse); human endogenous DICER: AGCAACACAGAGATCTCAAACATT (forward), GCAAAGCAGG GCTTTTCAT (reverse); human endogenous and overexpressed DICER: TGTTCCAGGAAGACCAGGTT (forward), ACTATCCCTCAAACACTCT GGAA (reverse); human DROSHA: GGCCCGAGAGCCTTTTATAG (forward), TGCACACGTCTAACTCTTCCAC (reverse); human GW182: CAGCCAGTCA GAAAGCAGTG (forward), TGTGAGTCCAGGATCTGCTACTT (reverse); mouse DICER: GCAAGGAATGGACTCTGAGC (forward), GGGGACTTCG ATATCCTCTTC (reverse); mouse DROSHA: CGTCTCTAGAAAGGTCCTAC AAGAA (forward), GGCTCAGGAGCAACTGGTAA (reverse).

RNase A treatment and RNA complementation experiments. Cells were plated on poly-D-lysinated coverslips and irradiated with 2 Gy of ionizing radiation. One hour later HeLa cells were permeabilized with 2% Tween 20 in PBS for 10 min at room temperature while I-SceI-transfected NIH2/4 cells were permeabilized in 0.5% Tween 20 in PBS for 10 min at room temperature. RNase A treatment was carried out in 1 ml of 1 mg ml^{-1} ribonuclease A from bovine pancreas (Sigma-Aldrich catalogue no. R5503) in PBS for 25 min at room temperature. After RNase A digestion, samples were washed with PBS, treated with 80 units of RNase inhibitor (RNaseOUT Invitrogen 40 units μl^{-1}) and 20 μg ml^{-1} of α-amanitin (Sigma) for 15 min in a total volume of 70 μl. For experiments with mirin, NIH2/4 cells were incubated at this step also with 100 μM mirin (Sigma) or DMSO for 15 min. Then, RNase-A-treated cells were incubated with total, small or gel-extracted RNA, or the same amount of tRNA, for an additional 15 min at room temperature. If using mirin, NIH2/4 cells were incubated with total RNA in the presence of 100 μM mirin or DMSO for 25 min at room temperature. Cell were then fixed with 4% paraformaldehyde or methanol:acetone 1:1.

In complementation experiments with synthetic RNA oligonucleotides, eight RNA oligonucleotides with the potential to form four pairs were chosen among the sequences that map at the integrated locus in NIH2/4 cells, obtained by deep sequencing. Synthetic RNA oligonucleotides were generated by Sigma with a monophosphate modification at the 5′ end. Sequences map to different regions of the integrated locus: two pairs map to a unique sequence flanking the I-SceI restriction site, one to the Lac-operator and one to the Tet-operator repetitive sequences. Two paired RNA oligonucleotides with the sequences of GFP were used as negative control. Sequences are reported below.

Oligonucleotide 1: 5′-AUAACAAUUUGUGGAAUUCGGCGC-3′, oligonucleotide 2: 5′-CGAAUUCCACAAAUUGUUAUCC-3′, oligonucleotide 3: 5′-AU UUGUGGAAUUCGGCGCCUCUAGAGUCGAGG-3′, oligonucleotide 4: 5′-CC UCGACUCUAGAGGCG-3′, oligonucleotide 5: 5′-AGCGGAUAACAAUUU GUGGCCACAUGUGGA-3′, oligonucleotide 6: 5′-UGUGGCCACAAAUUG UU-3′, oligonucleotide 7: 5′-ACUCCCUAUCAGUGAUAGAGAAAAGUGA

AAGU-3′, oligonucleotide 8: 5′-CUUUCACUUUUCUCUAUCACUGAUAGG GAGUG-3′. GFP 1: 5′-GUUCAGCGUGUCCGGCGAGUU-3′, GFP 2: 5′-CU CGCCGGACACGCUGAACUU-3′.

RNAs were resuspended in 60 mM KCl, 6 mM HEPES, pH 7.5, 0.2 mM $MgCl_2$, at the stock concentration of 12 μM, denatured at 95 °C for 5 min and annealed for 10 min at room temperature.

DICER RNA products were generated as follows. A 550-bp DNA fragment carrying the central portion of the genomic locus studied (three Lac repeats, the I-SceI site and two Tet repeats) was flanked by T7 promoters at both ends and was used as a template for *in vitro* transcription with the TurboScript T7 transcription kit (AMSBIO). The 500-nucleotide-long RNAs obtained were purified and incubated with human recombinant DICER enzyme (AMSBIO) to generate 22–23-nucleotide RNAs. RNA products were purified, quantified and checked on gel. As a control, the same procedure was followed with a 700-bp construct containing the RFP DNA sequence. Equal amounts of DICER RNA products generated in this way were used in a complementation experiment in NIH2/4 cells following RNase A treatment.

Small RNA preparation. Total RNA was isolated from cells using TRIzol (Invitrogen) according to the manufacturer's instructions. To generate small RNA-enriched fraction and small RNA-devoid fraction we used the *mir*Vana microRNA Isolation Kit (Ambion) according to the manufacturer's instructions. The *mir*Vana microRNA isolation kit uses an organic extraction followed by immobilization of RNA on glass-fibre (silica-fibres) filters to purify either total RNA, or RNA enriched for small species. For total RNA extraction ethanol is added to samples, and they are passed through a filter cartridge containing a glass-fibre filter, which immobilizes the RNA. The filter is then washed a few times and the RNA is eluted with a low ionic-strength solution. To isolate RNA that is highly enriched for small RNA species, ethanol is added to bring the samples to 25% ethanol. When this lysate/ethanol mixture is passed through a glass-fibre filter, large RNAs are immobilized, and the small RNA species are collected in the filtrate. The ethanol concentration of the filtrate is then increased to 55%, and it is passed through a second glass-fibre filter where the small RNAs become immobilized. This RNA is washed a few times, and eluted in a low ionic strength solution. Using this approach consisting of two sequential filtrations with different ethanol concentrations, an RNA fraction highly enriched in RNA species ≤200 nucleotides can be obtained[18,27].

RNA extraction from gel. Total RNA samples (15 ng) were heat denatured, loaded and resolved on a 15% denaturing acrylamide gel (1× TBE, 7 M urea, 15% acrylamide (29:1 acryl:bis-acryl)). Gel was run for 1 h at 180 V and stained in GelRed solution. Gel slices were excised according to the RNA molecular weight marker, moved to a 2 ml clean tube, smashed and RNA was eluted in 2 ml of ammonium acetate 0.5 M, EDTA 0.1 M in RNase-free water, rocking overnight at 4 °C. Tubes were then centrifuged 5 min at top speed, the aqueous phase was recovered and RNA was precipitated and resuspended in RNase free water.

G1/S checkpoint assay. WI-38 cells were irradiated with 10 Gy and 1 h afterwards incubated with BrdU (10 μg ml^{-1}) for 7 h; HCT116 cells were irradiated at 2 Gy and incubated with BrdU for 2 h. Cells were fixed with 4% paraformaldehyde and probed for BrdU immunostaining. At least 100 cells per condition were analysed.

G2/M checkpoint assay. HEK 293 calcium phosphate transfected cells were irradiated with 5 Gy and allowed to respond to ionizing-radiation-induced DNA damage in a cell culture incubator for 12, 24 or 36 h. Then, at these three time points after irradiation, together with not irradiated cells, 1×10^6 cells were collected for fluorescence activated cell sorting (FACS) analysis, fixed in 75% ethanol in PBS, 30 min on ice. Afterwards, cells were treated 12 h with 250 μg ml^{-1} of RNase A and incubated for at least 1 h with propidium iodide (PI). FACS profiles were obtained by the analysis of at least 5×10^5 cells. In the complementation experiments HEK 293 cells were transfected using Lipofectamine RNAi Max (Invitrogen) and 48 h later irradiated with 5 Gy. Cells were then treated as explained above.

Immunoblotting. Cells were lysed in sample buffer and 50–100 μg of whole cell lysate were resolved by SDS–PAGE, transferred to nitrocellulose and probed as previously described[14].

For zebrafish immunoblotting protein analysis, 72 h post-fertilization (hpf) larvae were deyolked in Krebs Ringer's solution containing 1 mM EDTA, 3 mM PMSF and protease inhibitor (Roche complete protease inhibitor cocktail). Embryos were then homogenized in SDS sample buffer containing 1 mM EDTA with a pestle, boiled for 5 min and centrifuged at 13,000 r.p.m. for 1 min. Protein concentration was measured with the BCA method (Pierce) and proteins (50—900 μg) were loaded in an SDS-12% (for γH2AX and H3) and SDS-6% polyacrylamide gel (for pATM and ATM), transferred to a nitrocellulose membrane, and incubated with anti-γH2AX (1:2,000, a gift from J. Amatruda[28]), H3 (1:10,000, Abcam), pATM (1:1,000, Rockland), ATM (1:1,000, Sigma). Immunoreactive bands were detected with horseradish-peroxidase-conjugated anti-rabbit or anti-mouse IgG and an ECL detection kit (Pierce). Protein loading was normalized to equal amounts of total ATM and H3.

Zebrafish embryo injection, cell transplantation and staining. Zebrafish embryos at the stage of 1–2 cells were injected with a morpholino against Dicer1[29] diluted in Danieau buffer. The morpholino oligonucleotide was injected at a concentration of 5 ng nl^{-1}, and a volume of 2 nl per embryo. To assess the efficiency of the morpholino to block miRNA maturation, we co-injected the morpholino with *in vitro* synthesized mRNA, encoding for red fluorescent protein (RFP) and carrying three binding sites for miR126 in the 3′ UTR[17]. The oligonucleotides carrying the binding sites for miR126 used for construction of the pCS2:RFPmiR126 sensor are: 5′-GCATTATTACTCACGGTACGAATAAGG CATTATTACTCACGGTACGAATAAGGCATTATTACTCACGGTACGA-3′ and 5′-CGTAATAATGAGTGCCATGCTTATTCCGTAATAATGAGTGCCA TGCTTATTCCGTAATAATGAGTGCCATGCT-3′. The construct was verified by sequencing and used to synthesize mRNA *in vitro* using the mMessage Kit (Ambion). mRNA encoding for RFPmiR126 sensor was injected alone or in combination with Dicer1 morpholino at a concentration of 10 pg nl^{-1}. For cell transplantation experiments, we injected donor embryos with a mixture of *dicer1* morpholino and mRNA encoding for GFP (5 pg nl^{-1}). Approximately 20 cells were transplanted from donor embryos at dome stage (5 hpf) to uninjected host at the same stage. Successfully transplanted larvae (displaying GFP+ cells) were irradiated as described below. Mature miRNAs were reverse transcribed to produce six different cDNAs for TaqMan MicroRNA assay (30 ng of total mRNA for each reaction; Applied Biosystems). Real-time PCR reactions based on TaqMan reagent chemistry were performed in duplicate on ABI PRISM 7900HT Fast Real-Time PCR System (Applied Biosystems). The level of miRNA expression was measured using C_T (threshold cycle). Fold change was calculated as 2^{-CT}.

For immunofluorescence in zebrafish larvae, 72 hpf larvae were irradiated with 12 Gy, fixed in 2% paraformaldehyde for 2 h at room temperature. After equilibration in 10 and 15% sucrose in PBS, larvae were frozen in OCT compound on coverslips on dry ice. Sections were cut with a cryostat at a nominal thickness of 14 μm and collected on Superfrost slides (BDH). Antisera used were zebrafish γH2AX (gift from J. Amatruda[28]) and pATM (Rockland). GFP fluorescence in transplanted embryos was still easily visible in fixed embryos. Images were acquired with a confocal (Leica SP2) microscope and ×63 oil immersion lens.

RNA sequencing. Nuclear RNA shorter than 200 nucleotides was purified using *mir*Vana microRNA Isolation Kit. RNA quality was checked on a small RNA chip (Agilent) before library preparation. For Illumina hi Seq Version3 sequencing, spike RNA was added to each RNA sample in the RNA: spike ratio of 10,000:1 before library preparation and libraries for Illumina GA IIX were prepared without spike. An improved small RNA library preparation protocol was used to prepare libraries[30]. In brief, adenylated 3′ adapters were ligated to 3′ ends of 3′-OH small RNAs using a truncated RNA ligase enzyme followed by 5′ adaptor ligation to 5′-monophosphate ends using RNA ligase enzyme, ensuring specific ligation of non-degraded small RNAs. cDNA was prepared using a primer specific to the 3′ adaptor in the presence of dimer eliminator and amplified for 12–15 PCR cycles using a special forward primer targeting the 5′ adaptor containing additional sequence for sequencing and a reverse primer targeting the 3′ adaptor. The amplified cDNA library was run on a 6% polyacrylamide gel and the 100 bp band containing cDNAs up to 33 nucleotides long was extracted using standard extraction protocols. Libraries were sequenced after quality check on a DNA high sensitivity chip (Agilent). Multiplexed barcode sequencing was performed on Illumina GA-IIX (35 bp single end reads) and Illumina Hi seq version3 (51 bp single end reads).

Statistical analyses. Results are shown as means ± s.e.m. *P* value was calculated by Chi-squared test. Quantitative PCR with reverse transcription results are shown as means of a triplicate ± standard deviation (s.d.) and *P* value was calculated by Student's *t*-test as indicated. *n* stands for number of independent biological experiments.

Statistical analysis of small RNA sequencing data. Statistical significance of downregulation of normalized miRNAs in DICER and DROSHA knockdown samples was calculated using the Wilcoxon signed-rank test.

The differences in the fraction of 22–23 nucleotides versus total small RNAs at the locus between the wild-type, DICER knockdown and DROSHA knockdown before and after cut were calculated by fitting a negative binomial model to the small RNAs count data and performing a likelihood ratio test, keeping the fraction of 22–23-nucleotide versus total small RNAs at the locus fixed across conditions under the null hypothesis and allowing it to vary between conditions under the alternative hypothesis.

27. Duchaine, T. F. *et al.* Functional proteomics reveals the biochemical niche of *C. elegans* DCR-1 in multiple small-RNA-mediated pathways. *Cell* **124,** 343–354 (2006).
28. Sidi, S. *et al.* Chk1 suppresses a caspase-2 apoptotic response to DNA damage that bypasses p53, Bcl-2, and caspase-3. *Cell* **133,** 864–877 (2008).

29. Wienholds, E., Koudijs, M. J., van Eeden, F. J., Cuppen, E. & Plasterk, R. H. The microRNA-producing enzyme Dicer1 is essential for zebrafish development. *Nature Genet.* **35,** 217–218 (2003).

30. Kawano, M. *et al.* Reduction of non-insert sequence reads by dimer eliminator LNA oligonucleotide for small RNA deep sequencing. *Biotechniques* **49,** 751–755 (2010).

Chapter 4

Rett syndrome and piRNAs

Disease Markers 33 (2012) 261–275
DOI 10.3233/DMA-2012-0932
IOS Press

piRNAs warrant investigation in Rett Syndrome: An omics perspective

Alka Saxena*, Dave Tang and Piero Carninci
RIKEN Omics Science Center, Yokohama, Japan

Abstract. Mutations in the *MECP2* gene are found in a large proportion of girls with Rett Syndrome. Despite extensive research, the principal role of MeCP2 protein remains elusive. Is MeCP2 a regulator of genes, acting in concert with co-activators and co-repressors, predominantly as an activator of target genes or is it a methyl CpG binding protein acting globally to change the chromatin state and to supress transcription from repeat elements? If MeCP2 has no specific targets in the genome, what causes the differential expression of specific genes in the *Mecp2* knockout mouse brain? We discuss the discrepancies in current data and propose a hypothesis to reconcile some differences in the two viewpoints. Since transcripts from repeat elements contribute to piRNA biogenesis, we propose that piRNA levels may be higher in the absence of MeCP2 and that increased piRNA levels may contribute to the mis-regulation of some genes seen in the *Mecp2* knockout mouse brain. We provide preliminary data showing an increase in piRNAs in the *Mecp2* knockout mouse cerebellum. Our investigation suggests that global piRNA levels may be elevated in the *Mecp2* knockout mouse cerebellum and strongly supports further investigation of piRNAs in Rett syndrome.

Keywords: Rett Syndrome, MeCP2, piRNAs, LINE 1, short RNAs

Rett Syndrome (RTT), a severe neurodevelopmental disorder, leads to intellectual disability in girls. After a normal prenatal and postnatal period, patients usually present with developmental delay between 6 and 18 months of age, followed by the development of stereotypic hand movements and loss of acquired skills including voluntary hand use, language and communication. This regression is characteristic of Rett Syndrome which, in 97% of clinically diagnosed classic cases and 70% of atypical cases, is caused by mutations in the methyl CpG binding protein 2 gene, (*MECP2*) [1]. In some patients with atypical Rett Syndrome, where some but not all clinical features are seen, mutations in *CDKL5* [2] or *FOXG1* [3] are found, albeit infrequently. Diagnosis of Rett Syndrome is based on clinical criteria [4] and confirmed upon detection of a mutation in *MECP2*, *CDKL5* or *FOXG1*. However, in approximately 20% of girls clinically diagnosed with classic or atypical Rett syndrome, mutations cannot be detected in either of these genes.

MECP2 gene undergoes X chromosome inactivation (XCI) [5], which means that in any cell with two X chromosomes, RNA transcripts arise only from the *MECP2* gene on the active X chromosome. This is because the *MECP2* gene on the inactive X chromosome has been silenced. Chaumeil et al. demonstrated, through in-situ hybridization in mouse ES cells, that the *Mecp2* gene moves inside the silencing compartment of *Xist* on the 4th day after differentiation [6]. Although some genes are known to escape X inactivation in humans and mouse [7–10], Carrel et al. showed using rodent/human somatic hybrid cell lines that in humans MECP2 transcripts are not expressed from the inactive X chromosome [7]. Due to the random nature of X inactivation, part of the clinical variability in RTT is attributed to the differences in the X inactivation status of patients [11–13], however more recent data suggest X inactivation status may not adequately explain the phenotypic variations [14].

MECP2 gene is composed of 4 exons and generates two transcripts which encode two nearly identical protein isoforms [15,16]. MeCP2_e1, which commences translation from exon 1, is encoded from a transcript encompassing exons 1, 3 and 4; and MeCP2_e2, which starts translation from the end of exon 2, is generated from a transcript arising from exons 1, 2, 3 and 4, where exon 1 and most of exon 2 form the 5'UTR [15,

*Corresponding author: Alka Saxena, RIKEN Omics Science Center, Yokohama, Japan. E-mail: alka@gsc.riken.jp.

ISSN 0278-0240/12/$27.50

16]. Although *MECP2* is expressed in all tissues, semi-quantitative PCR analysis has shown that *Mecp2_e1* may have a higher expression level over *Mecp2_e2* in the brain [16]. Mutations in exons 3 and 4 affect both protein isoforms and are frequently found in Rett patients. Mutations in exon 1 can cause Rett syndrome despite the fact that mutations in exon 1 do not affect the coding region of the MeCP2_e2 protein. Interestingly, mutations in exon 2 of the gene, which have the potential to affect the MeCP2_e2 isoform alone, have so far not been found in patients. While earlier work on an Australian patient with a recurrent deletion in exon 1 of *MECP2* gene demonstrated the absence of MeCP2_e2 protein correlated with X inactivation status, suggesting translational interference from the mutation [17], a recent publication found no evidence of loss of MeCP2_e2 protein in a Canadian patient with a similar mutation indicating that some patients may present with clinical features of Rett syndrome even in the presence of a fully functional MeCP2_e2 isoform [18]. Interestingly, this data also suggests that despite high sequence similarities there is no functional redundancy between the two protein isoforms.

Due to its property to bind methylated DNA with high affinity and its association with repressor complexes consisting of HDAC1/2 and Sin3A, MeCP2 was believed to function as a transcriptional repressor [19, 20]. The MeCP2-DNA interaction was shown to result in chromatin compaction, which is also correlated with silencing of chromatin [21]. Subsequent studies revealed that MeCP2 had binding affinity to methylated DNA as well as non-methylated DNA [22]. Absence of MeCP2 in mouse brains also results in an increase in H3Ac levels, suggesting a role for MeCP2 in chromatin modification [23,24]. Recent data suggest that MeCP2 protein may be a regulator of transcription, acting in concert with activators as well as repressors to regulate gene expression. Yasui et al. first reported that promoter occupancy by MeCP2 may not result in gene silencing [25]. Using a custom tiling array of selected chromosomal regions totalling 26.3 Mb, they performed ChIP-chip analysis on SH-SY5Y cells with antibodies against MeCP2 and RNA polymerase II. The data revealed co-occupancy of MeCP2 and RNA Polymerase II at selected promoters suggesting that MeCP2 binding may not be correlated to gene repression [25]. Using ChIP-chip assays for 24,275 promoters, they demonstrated that only 2600–4300 promoters were occupied by MeCP2, of which 1534 promoters showed strongest enrichment. Comparison with gene expression arrays in the same cell lines revealed that almost 63% of the "strongest" promoters were expressed in SH-SY5Y cells. Subsequent MeDIP-ChIP analysis revealed that just 2.2% of methylated promoters were occupied by MeCP2 [25]. These data were supported in part by Chahrour et al. who used microarrays to determine differentially expressed genes in the hypothalamus of 6 week old *Mecp2* knockout (KO) mouse and in the hypothalamus of a mouse model that overexpressed *Mecp2* under its endogenous promoter (Tg) [26]. Combining their data from the KO and Tg models, they identified 2561 genes as direct targets of MeCP2, of which ∼85% were activated by MeCP2 and ∼ 15% were repressed by MeCP2 [26]. Using mass spectrometry on proteins co-immunoprecipitated with an anti-MeCP2 antibody, they identified CREB1 as a co-activator associated with MeCP2 and demonstrated co-occupancy of the two proteins at an activated target *Sst [26]*. Together, these data established MeCP2 as an activator of transcription [25,26]. Thus transcriptional mis-regulation is believed to underlie the phenotype seen in patients with mutations in the *MECP2* gene. In view of the fact that FOXG1 is a member of the forkhead family of transcription regulators, it is likely that in patients carrying mutations in *FOXG1,* mis-regulation of genes may contribute to the phenotypic features. The molecular pathology leading to the clinical phenotype of Rett Syndrome in mutation negative patients remains unknown. While much has been reported on the mis-regulation of specific genes after MeCP2 knockdown (KD) or KO, such studies have not yet been reported for FOXG1. Other studies reveal subtle changes in the expression levels of specific genes after MeCP2 KD or in the *Mecp2* KO mouse brain rather than genome wide transcriptional mis-regulation [27–29].

However, a recent report suggests that the absence of a functional MeCP2 may result widespread mis-regulation of repeat elements. Skene et al. investigated MeCP2 binding on selected loci in the mature mouse brain using ChIP-qPCR and demonstrated that MeCP2 was enriched all across the loci, but the enrichment was reduced over CpG islands, which are generally methylation free [24]. With bisulfite modification and sequencing of selected loci they demonstrated the recovery of predominantly methylated chromatin from the MeCP2 ChIP, re-emphasizing the role of MeCP2 as a methyl CpG binding protein [24]. Based on their investigation of the histone acetylation status by Western blotting and H3Ac ChIP-qPCR at 100 loci, they concluded that the association of MeCP2 with chromatin causes a genome-wide decrease in histone acetylation [24]. To investigate the binding sites of MeCP2

genome wide, Skene et al. performed MeCP2-ChIP-sequencing on the whole brain. Despite deep sequencing, they did not find peaks of MeCP2 occupancy, but found reads which coincided with methylated regions of the genome. Since they did not uncover specific binding targets of MeCP2 in the genome, they hypothesized that MeCP2 may act at a global level, most likely to supress transcription from the repeat regions of the genome [24]. Using qPCR, they demonstrated a 1.6 fold increase in transcripts arising from repeat sequences such as LINE-1, intra-cisternal A particles (IAPs) and major satellite DNA in the nuclear fraction of the *Mecp2* KO mouse brain. Based on their data they proposed that MeCP2 functions to repress spurious transcription of repeat elements [24] rather than to activate specific gene targets. An earlier investigation into the association between MeCP2 and LINE-1 and Alus had revealed that MeCP2 repressed LINE-1 expression and transposition, but activated Alu expression [30]. The role of MeCP2 in repressing transcription and transposition of LINE-1 elements was also corroborated by independent studies from the Gage Lab that showed that LINE-1 is over expressed in neuron progenitor cells after KD of MeCP2 and in neurons derived from Rett patients [31]. The data from Yasui et al. suggests that MeCP2 displays limited binding to methylated sites [25] and from Chahrour et al. proposes that MeCP2 acts as a transcriptional activator of specific targets [26]. In contrast, the data from Skene et al. suggests that MeCP2 binds methylated DNA, is a modulator of global chromatin state and may not have specific gene targets [24]. We note that some of these studies were conducted using microarrays or custom tiling arrays, which are generally limited to gene specific probes and exclude repeat sequences. Despite contradictory inferences on MeCP2 function, the fact remains that specific genes are mis-expressed and repeat elements are over-expressed in *Mecp2* KO mice. To reconcile the two opposing views, an alternative model would suggest that the key role of MeCP2 is to silence LINEs and similar repeat elements globally and that the observed mis-expression of genes is a downstream consequence of mis-expressed repeat elements.

Several recent reports suggest that non-coding RNAs play a role in the regulation of transcription through epigenetic modifications, [for a review see [32]]. Repeat elements, particularly LINEs, are known to participate in the silencing of genes on the X chromosome [33]. Expression of LINEs in the vicinity of genes is instrumental for their inclusion into the Xist silencing compartment [33]. It is not yet known whether transcripts from LINE elements are associated with chromatin remodelling complexes to mediate epigenetic changes and fine-tune gene transcription, but we note that the elevated repeat elements were found in the nuclear compartment of the *Mecp2* KO mouse brain cells [24] and there is emerging evidence of enrichment of LINEs in nuclear and chromatin fraction of cells [34]. A recent report using a retrotransposon capture sequencing technique (RC-seq) reveals that somatic transposition of LINE-1 (L1) in the hippocampus results in insertions, predominantly in exons and introns of protein coding genes [35]. Comparing microarray data with their RC-seq data, Baillie et al. reported that intronic L1 insertions are likely to cause overexpression of such genes in the brain, suggesting a regulatory role for L1 [35]. It is not clear if the increase in retrotransposon expression in the *Mecp2* KO brain leads to their active transposition even in post mitotic neurons. Random integration of transposons is suppressed in differentiated somatic cells by transcriptional [36] and post-transcriptional mechanisms [37]. It would be interesting to investigate if somatic retrotransposition is increased in the *Mecp2* KO brain and whether overexpressed genes identified in *Mecp2* KO mouse show novel intronic L1 insertion events.

Retrotransposons such as LINEs can be further processed into short 21–24 nucleotide double stranded siRNAs [38] or into single stranded 24–31 nucleotide long piRNAs [39,40]. Watanabe et al. described in mouse oocytes dicer dependent double stranded endogenous siRNAs mapping exclusively to retrotransposons or expressed mRNA transcripts [38]. While the presence of endogenous siRNAs has not been demonstrated in the mouse brain, given that such short RNAs are shown to regulate the expression levels of specific genes and specific retrotransposons [38,41], and that dysfunctional MeCP2 may result in the overexpression of LINE-1 [24,31], it would be interesting to investigate the presence of endogenous siRNAs in the MeCP2 KO mouse brain.

piRNAs are germ line specific short RNAs of size 24 to 31 nucleotides generated through dicer independent processing of long single strand RNA transcripts. piRNAs interact with the PIWI proteins (MILI, MIWI and MIWI2 in mouse) [39,40] and their function, though not fully understood, appears to relate to silencing of transposons especially LINE-1 [39,40,42] intracisternal A particles [39,40] and specific genes through DNA hypermethylation [43]. In mouse testis, 17% of piRNAs bound to the MIWI protein map to repeats including LINEs, SINEs and LTRs [40]. In addition, through

a unique ping-pong cycle, piRNAs are amplified from existing retrotransposon transcripts, mostly LINE-1 elements [44,45]. This amplification cycle also results in the depletion of LINE-1 in germ line cells and is believed to deplete the levels of retrotransposon transcripts after differentiation [42]. Thus piRNAs regulate expression of LINEs and Intracisternal A particles both transcriptionally and post transcriptionally. Interestingly, piRNAs have recently been shown to regulate expression of a single imprinted gene in an imprinted locus in mouse spermatogonia via DNA methylation by piRNA targeting of a non coding RNA (pitRNA) arising from the locus [43]. It is as yet not known if such specific targeting by piRNAs is a widespread phenomenon, nevertheless it highlights a mechanism through which piRNAs may regulate expression of specific genes.

Until recently, piRNAs and their associated proteins were presumed to be germ line specific in mouse, but a report published last year confirmed the presence of MIWI and its associated piRNAs in the mouse hippocampus through sequencing, RIP-qPCR, northern blots, western blots and in situ hybridization studies [46]. Through bioinformatics analysis, Lee at al. identified specific piRNAs expressed in the brain and showed through piRNA inhibition studies, that one piRNA in the brain, DQ541777, may play a role in regulating the size of dendritic spines [46]. It appears plausible that in the absence of MeCP2, over expression of repeat elements, particularly LINE-1 may result in an increase in piRNA amplification from transposons. It would be interesting to investigate whether akin to germ line cells, in brain also, the increase in piRNAs result in depletion of retrotransposon transcript levels through transcriptional and post transcriptional silencing. The mechanism of transcriptional silencing by piRNAs through DNA methylation may require recruitment of repressor complexes by proteins that bind methylated DNA, including MeCP2, thus highlighting a feedback loop for MeCP2 requirement.

To investigate our hypothesis that piRNAs may be overexpressed in the *Mecp2* KO mouse brain, we analysed a short RNA library made from mouse cerebellum [47]. To identify miRNAs differentially expressed in the cerebellum, Wu et al. performed short RNA sequencing of pooled 6 week old pre-symptomatic wildtype and *Mecp2* KO cerebellum ($n = 4$ in each pool) using the SOLiD version 2 sequencer [47]. We downloaded the pooled libraries from the DDBJ database (DDBJ accession number SRP005132). ncRNAs were downloaded from NONCODE version 3 [48] and a total of 75,814 mouse piRNAs were extracted from this database. As the SOLiD reads correspond to the 5' ends of small RNAs, we directly mapped the respective reads from the pooled WT and KO libraries, using SHRiMP version 2.2.2 [49] with the default parameters, to the mouse piRNAs. After mapping we corrected tag numbers for reads multi-mapping to more than one piRNA, so that if a read mapped equally well to 2 or more individual piRNA sequences, the tag numbers were divided by the number of times it multi mapped, followed by equal assignment to all the piRNAs. For expression analysis we did not take into consideration multi mapped reads that mapped to 5 or more piRNAs and also filtered out reads that had less than 5 tags in the KO samples. The tag numbers were normalized by tags per million before comparison between wildtype and KO samples. Our very preliminary analysis of piRNAs in the cerebellum reveals 357 piRNAs in the cerebellum libraries (Supplementary Table 1). While 81% (287) of the individual piRNAs found in the cerebellum have a higher expression in KO, 59% (208) piRNAs show an expression change of over 1.5 fold in the KO cerebellum compared with the wildtype (Fig. 1B and supplementary Table 1). Overall, we found a striking 1.9 fold increase in the total piRNAs in the KO cerebellum in comparison with the wildtype cerebellum (Table 1 and Fig. 1A) suggesting a global increase in piRNAs in the *Mecp2* KO sample. We next investigated whether the 20 most abundant piRNAs identified by Lee et al. in the mouse hippocampus were represented in the mouse cerebellum [46]. We found 19 out of the 20 piRNAs reported by Lee et al. in the cerebellum libraries (Table 2) including DQ 541777 (the piRNA implicated in regulating the size of dendritic spines) and of these, 12 piRNAs (60%) revealed a fold change of over 1.5 in the KO cerebellum (Table 2). Interestingly, DQ541777 is the 5^{th} most abundant piRNA in the cerebellum libraries, the two most highly abundant piRNAs in the cerebellum libraries map to rRNA loci, which were incidentally excluded in the hippocampal analysis [46].

Based on our preliminary findings, we suggest a model for Rett Syndrome where, in the absence of a functional MeCP2, the over-expressed repeat elements lead to an increase in the total piRNAs. The over-represented piRNAs may function, not only to deplete the load of repeat transcripts in cells, but also to fine-tune the expression level of specific genes. Thus piRNA mis-regulation may contribute to some of the differences in gene expression seen in the *Mecp2* KO mouse brain.

While our analyses provide preliminary evidence of genome wide piRNA over-expression in the *Mecp2*

Table 1
piRNA analysis in the short RNA libraries made from 6 week old pooled cerebellum from the wildtype mouse and the *Mecp2* knockout mouse [38] (DDBJ accession number SRP005132)

	WT cerebellum	KO cerebellum
Read ID	SRR089647	SRR089648
Total sRNA reads	3660124	2789136
Total reads mapped to piRNAs (no filter)	362089	522238
Reads mapped to piRNAs (filter out < 5 reads in KO)	356283	518589
piRNA mapped tags normalized by TPM	97341.74762	185931.8441
Fold change (tpm normalized piRNAs over WT)	1	1.910093548

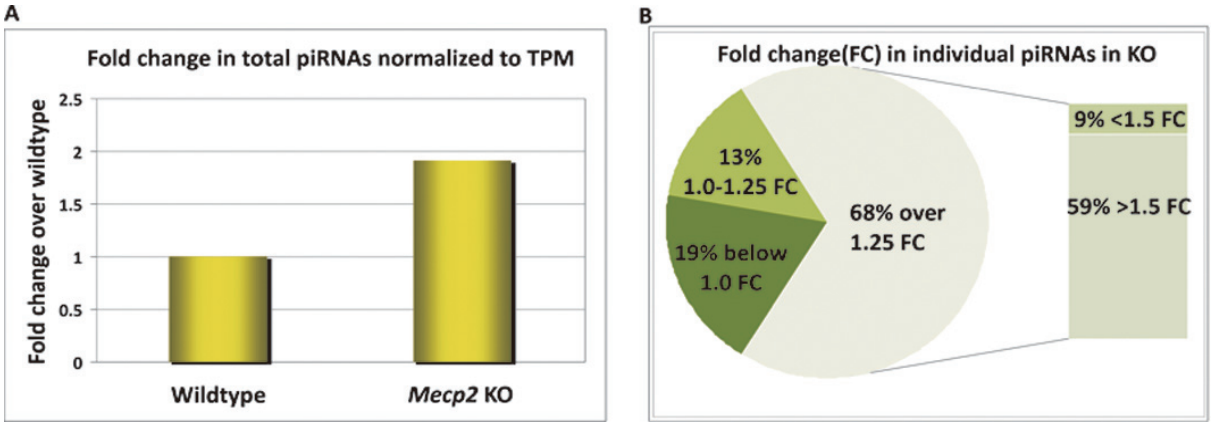

Fig. 1. piRNA levels are elevated in *Mecp2* KO cerebellum. Read numbers for individual piRNAs found in the wildtype (WT) and *Mecp2* knockout (KO) samples were normalized to tags per million as described in the text. After filtering out the piRNAs with less than 5 reads in the KO sample, the total piRNA reads were summed up. The histogram in panel A shows that the total numbers of piRNAs are almost doubled (1.9 fold) in the KO sample suggesting a global rise in piRNAs. The fold change relative to WT was calculated for each individual piRNA in KO. The pie chart in panel B reveals that 81% of piRNAs show a higher expression level in the KO sample. Of these, 59% have a fold change of over 1.5 in the KO sample (see supplementary Table 1). (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/DMA-2012-0932)

KO cerebellum, this data was generated from libraries without replicates. Thus additional detailed investigations are warranted to affirm the over-representation of piRNAs and gain insights into the extent of their contribution to the gene mis-regulation seen in the Rett mouse model. We did not venture into the identification of gene targets of mis-regulated piRNAs and their intersection with the known mis-regulated genes or repeats. Such data may provide insights into the mis-regulation of some genes and the biogenesis of the over-represented piRNAs. Notably, in humans and mouse, an absence of MeCP2 results in fewer dendritic spines when compared with wildtype neurons [50,51]. While inhibition of DQ541777 was reported to cause a decrease in spine density, whether the overexpression of piRNAs, including DQ541777, can cause such morphological changes in the brain is not yet known. Further, recent reports have demonstrated that unlike mRNAs, miRNAs are stable in extracellular environments including blood serum [52,53]. Although such analyses have not been conducted for piRNAs, if piRNAs are found to be stable in extracellular fluids, differentially expressed piRNAs may potentially represent clinical biomarkers for the diagnosis and prognosis of Rett Syndrome.

There is overwhelming complexity in unravelling the molecular pathogenesis of the phenotype seen in Rett syndrome. Although the miRNA repertoire of *Mecp2* KO cerebellum has been investigated using next generation sequencing approaches, the field would benefit in re-identifying mis-regulated transcripts through deep, long and short (to identify sRNAs other than miRNAs), RNA sequencing of specific brain regions. Additionally, RC-seq conducted to identify somatic integration events would reveal whether the overexpressed genes are correlated with intronic LINE-1 insertion events. And while it has been established for the *Mecp2* KO mouse model, that investigating a specific brain region is more fruitful than investigating the whole brain [26], analysis of neuronal subtypes would be even more insightful. Until now, the isolation of neuronal subtypes from adult mouse brain using high throughput techniques such as FACS sorting was challenging, yielding few nuclei and poor quality RNA. A recently pub-

Table 2
Comparison of the top twenty piRNAs reported in the hippocampus [38] with the piRNAs found in the cerebellum

Top 20 piRNAs in hippocampus	WT hippocampus tag numbers	WT cerebellum tag numbers	KOcerebellum tag numbers	WT cerebellum TPM	KO cerebellum TPM	FC KO/WT
DQ541777	16130	1995.50	2411.00	545.20	864.43	1.59
DQ705026	6257	154.00	377.00	42.08	135.17	3.21
DQ555094	3439	202.00	140.00	55.19	50.19	0.91
DQ719597	2459	168.00	306.00	45.90	109.71	2.39
DQ689086	1514	65.00	78.00	17.76	27.97	1.57
DQ540285	1433	457.40	548.23	124.97	196.56	1.57
DQ540981	1360	126.50	124.00	34.56	44.46	1.29
DQ720186	849	336.00	251.00	91.80	89.99	0.98
DQ555093	775	189.50	129.50	51.77	46.43	0.90
DQ540862	639	21.50	33.50	5.87	12.01	2.04
DQ540284	635	456.90	548.23	124.83	196.56	1.57
DQ541506	580	523.90	627.40	143.14	224.94	1.57
DQ539915	304	35.50	28.00	9.70	10.04	1.04
DQ540861	252	20.50	31.50	5.60	11.29	2.02
DQ715526	207	20.00	20.00	5.46	7.17	1.31
DQ543676	182	438.90	518.70	119.91	185.97	1.55
DQ722288	175	2.00	13.00	0.55	4.66	8.53
DQ551351	168	Not found	Not found	Not found	Not found	Not found
DQ550765	118	10.75	9.50	2.94	3.41	1.16
DQ708131	115	3.00	6.00	0.82	2.15	2.62

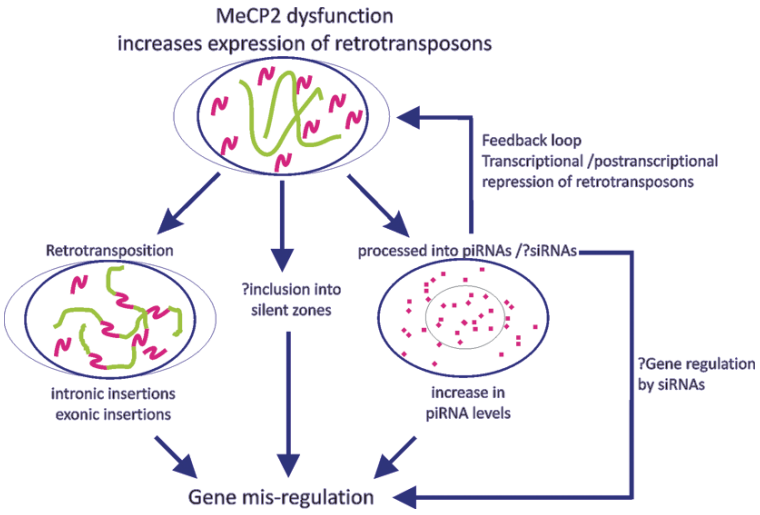

Fig. 2. Schematic of the proposed model showing that changes in the expression level of some genes may be a consequence of the increase in expression of retrotransposons. DNA is depicted in green, retrotransposon transcripts in pink and piRNAs as pink dots. See text for details. (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/DMA-2012-0932)

lished trehalose enhanced technique for FACS sorting individual neuronal subtypes could help isolate high quality RNA from *Mecp2* null neuronal subtypes for transcriptome sequencing [54].

In conclusion, we propose that overexpression of LINE-1 may contribute to the mis-regulation of some genes in Rett syndrome, mediated through insertional events or by an increase in piRNAs (Fig. 2). Our preliminary data suggests that piRNA expression levels may be altered globally in the absence of MeCP2. Application of next generation sequencing technologies may resolve some key questions regarding MeCP2 function and the downstream consequences of a dysfunctional MeCP2.

Acknowledgements

A.S. is supported by the Funding Program for Next Generation World-Leading Researchers by MEXT to

P.C and a MEXT Grant-in-Aid 321-EA-O-77; D.T. is supported by the European Union 7th Framework Programme under grant agreement FP7-People-ITN-2008-238055 ("BrainTrain" project) to P.C. and a Research Grant for RIKEN Omics Science Center from MEXT.

References

[1] Amir, R.E., et al., Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet, 1999. **23**(2): 185-8.
[2] Weaving, L.S., et al., Mutations of CDKL5 cause a severe neurodevelopmental disorder with infantile spasms and mental retardation. Am J Hum Genet, 2004. **75**(6): 1079-93.
[3] Ariani, F., et al., FOXG1 is responsible for the congenital variant of Rett syndrome. Am J Hum Genet, 2008. **83**(1): 89-93.
[4] Neul, J.L., et al., Rett syndrome: revised diagnostic criteria and nomenclature. Ann Neurol, 2010. **68**(6): 944-50.
[5] Adler, D.A., et al., The X-linked methylated DNA binding protein, Mecp2, is subject to X inactivation in the mouse. Mamm Genome, 1995. **6**(8): p. 491-2.
[6] Chaumeil, J., et al., A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. Genes Dev, 2006. **20**(16): 2223-37.
[7] Carrel, L. and H.F. Willard, X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature, 2005. **434**(7031): 400-4.
[8] Lingenfelter, P.A., et al., Escape from X inactivation of Smcx is preceded by silencing during mouse development. Nat Genet, 1998. **18**(3): 212-3.
[9] Xu, J., X. Deng and C.M. Disteche, Sex-specific expression of the X-linked histone demethylase gene Jarid1c in brain. PLoS One, 2008. **3**(7): e2553.
[10] Yang, F., et al., Global survey of escape from X inactivation by RNA-sequencing in mouse. Genome Res, 2010. **20**(5): 614-22.
[11] Hoffbuhr, K.C., et al., Associations between MeCP2 mutations, X-chromosome inactivation, and phenotype. Ment Retard Dev Disabil Res Rev, 2002. **8**(2): 99-105.
[12] Weaving, L.S., et al., Effects of MECP2 mutation type, location and X-inactivation in modulating Rett syndrome phenotype. Am J Med Genet A, 2003. **118A**(2): 103-14.
[13] Huppke, P., et al., Very mild cases of Rett syndrome with skewed X inactivation. J Med Genet, 2006. **43**(10): 814-6.
[14] Xinhua, B., et al., X chromosome inactivation in Rett Syndrome and its correlations with MECP2 mutations and phenotype. J Child Neurol, 2008. **23**(1): 22-5.
[15] Kriaucionis, S. and A. Bird, The major form of MeCP2 has a novel N-terminus generated by alternative splicing. Nucleic Acids Res, 2004. **32**(5): 1818-23.
[16] Mnatzakanian, G.N., et al., A previously unidentified MECP2 open reading frame defines a new protein isoform relevant to Rett syndrome. Nat Genet, 2004. **36**(4): 339-41.
[17] Saxena, A., et al., Lost in translation: translational interference from a recurrent mutation in exon 1 of MECP2. J Med Genet, 2006. **43**(6): 470-7.
[18] Gianakopoulos, P.J., et al., Mutations in MECP2 exon 1 in classical Rett patients disrupt MECP2_e1 transcription, but not transcription of MECP2_e2. Am J Med Genet B Neuropsychiatr Genet, 2012. **159B**(2): 210-6.
[19] Nan, X., et al., Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature, 1998. **393**(6683): 386-9.
[20] Jones, P.L., et al., Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet, 1998. **19**(2): 187-91.
[21] Georgel, P.T., et al., Chromatin compaction by human MeCP2. Assembly of novel secondary chromatin structures in the absence of DNA methylation. J Biol Chem, 2003. **278**(34): 32181-8.
[22] Nikitina, T., et al., Multiple modes of interaction between the methylated DNA binding protein MeCP2 and chromatin. Mol Cell Biol, 2007. **27**(3): 864-77.
[23] Shahbazian, M., et al., Mice with truncated MeCP2 recapitulate many Rett syndrome features and display hyperacetylation of histone H3. Neuron, 2002. **35**(2): 243-54.
[24] Skene, P.J., et al., Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. Mol Cell, 2010. **37**(4): 457-68.
[25] Yasui, D.H., et al., Integrated epigenomic analyses of neuronal MeCP2 reveal a role for long-range interaction with active genes. Proc Natl Acad Sci U S A, 2007. **104**(49): 19416-21.
[26] Chahrour, M., et al., MeCP2, a key contributor to neurological disease, activates and represses transcription. Science, 2008. **320**(5880): 1224-9.
[27] Urdinguio, R.G., et al., Mecp2-null mice provide new neuronal targets for Rett syndrome. PLoS One, 2008. **3**(11): e3669.
[28] Yakabe, S., et al., MeCP2 knockdown reveals DNA methylation-independent gene repression of target genes in living cells and a bias in the cellular location of target gene products. Genes Genet Syst, 2008. **83**(2): 199-208.
[29] Smrt, R.D., et al., Mecp2 deficiency leads to delayed maturation and altered gene expression in hippocampal neurons. Neurobiol Dis, 2007. **27**(1): 77-89.
[30] Yu, F., et al., Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. Nucleic Acids Res, 2001. **29**(21): 4493-501.
[31] Muotri, A.R., et al., L1 retrotransposition in neurons is modulated by MeCP2. Nature, 2010. **468**(7322): 443-6.
[32] Saxena, A. and P. Carninci, Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. Bioessays, 2011. **33**(11): 830-9.
[33] Chow, J.C., et al., LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell, 2010. **141**(6): 956-69.
[34] Djebali S, et al., Landscape of transcription in human cells. Nature, 2012. doi:10.1038/nature11233.
[35] Baillie, J.K., et al., Somatic retrotransposition alters the genetic landscape of the human brain. Nature, 2011. **479**(7374): 534-7.
[36] Slotkin, R.K. and R. Martienssen, Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet, 2007. **8**(4): 272-85.
[37] Yang, N. and H.H. Kazazian, Jr., L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nat Struct Mol Biol, 2006. **13**(9): 763-71.
[38] Watanabe, T., et al., Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature, 2008. **453**(7194): 539-43.
[39] Aravin, A., et al., A novel class of small RNAs bind to MILI protein in mouse testes. Nature, 2006. **442**(7099): 203-7.
[40] Girard, A., et al., A germline-specific class of small RNAs

binds mammalian Piwi proteins. Nature, 2006. **442** (7099): 199-202.

[41] Chen, L., et al., Naturally occurring endo-siRNA silences LINE-1 retrotransposons in human cells through DNA methylation. Epigenetics, 2012. **7**(7): 758-71.

[42] Reuter, M., et al., Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature, 2011. **480**(7376): 264-7.

[43] Watanabe, T., et al., Role for piRNAs and noncoding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. Science, 2011. **332**(6031): 848-52.

[44] Brennecke, J., et al., Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell, 2007. **128**(6): 1089-103.

[45] Gunawardane, L.S., et al., A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science, 2007. **315**(5818): 1587-90.

[46] Lee, E.J., et al., Identification of piRNAs in the central nervous system. RNA, 2011. **17**(6): 1090-9.

[47] Wu, H., et al., Genome-wide analysis reveals methyl-CpG-binding protein 2-dependent regulation of microRNAs in a mouse model of Rett syndrome. Proc Natl Acad Sci U S A, 2010. **107**(42): 18161-6.

[48] Bu, D., et al., NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Res, 2012. **40** (Database issue): D210-5.

[49] David, M., et al., SHRiMP2: sensitive yet practical SHort Read Mapping. Bioinformatics, 2011. **27**(7): 1011-2.

[50] Belichenko, P.V., et al., Widespread changes in dendritic and axonal morphology in Mecp2-mutant mouse models of Rett syndrome: evidence for disruption of neuronal networks. J Comp Neurol, 2009. **514**(3): 240-58.

[51] Chapleau, C.A., et al., Dendritic spine pathologies in hippocampal pyramidal neurons from Rett syndrome brain and after expression of Rett-associated MECP2 mutations. Neurobiol Dis, 2009. **35**(2): 219-33.

[52] Wang, K., et al., Export of microRNAs and microRNA-protective protein by mammalian cells. Nucleic Acids Res, 2010. **38**(20): 7248-59.

[53] Turchinovich, A., et al., Characterization of extracellular circulating microRNA. Nucleic Acids Res, 2011. **39**(16): 7223-33.

[54] Saxena, A., et al., Trehalose-enhanced isolation of neuronal sub-types from adult mouse brain. Biotechniques, 2012. **52**(6): 381-5.

Supplemental material

Supplementary Table 1

List of piRNAs, found in the cerebellum samples in wildtype and Mecp2 KO samples with their tag numbers and normalized tags per million (Tpm) values. This list was generated after filtering out piRNAs with less than 5 tags in the Mecp2 KO sample. Top 20 piRNAs reported to be present in the hippocampus [38] are highlighted in **bold**. Tpm was calculated by dividing the tag numbers by the total number of reads in the library (see Table 1, 3660124 for wildtype (WT) and 2789136 for Mecp2 KO). Fold change was calculated by dividing the Tpm normalized tag numbers for Mecp2 KO with WT

piRNA accession	Length	Number of tags in wildtype cerebellum	Number of tags in Mecp2 KO cerebellum	WT_Tpm	Mecp2 KO_Tpm	Fold_change_ KO/WT	Reported as top 20 piRNAs in brain
DQ546606	29	147865.67	229912.17	40399.09	82431.32	2.04	No
DQ540966	30	146001.83	228493.17	39889.86	81922.56	2.05	No
DQ540188	25	29252.92	24557.00	7992.33	8804.52	1.10	No
DQ558990	30	6166.50	11048.00	1684.78	3961.08	2.35	No
DQ541777	**30**	**1995.50**	**2411.00**	**545.20**	**864.43**	**1.59**	**Yes**
DQ703900	32	4252.00	1621.50	1161.71	581.36	0.50	No
DQ540229	31	1495.50	1033.33	408.59	370.49	0.91	No
DQ719271	21	1355.00	1009.83	370.21	362.06	0.98	No
DQ540053	29	810.00	822.00	221.30	294.71	1.33	No
DQ708952	22	694.00	720.00	189.61	258.14	1.36	No
DQ541506	**28**	**523.90**	**627.40**	**143.14**	**224.94**	**1.57**	**Yes**
DQ542358	31	446.00	549.50	121.85	197.01	1.62	No
DQ540285	**32**	**457.40**	**548.23**	**124.97**	**196.56**	**1.57**	**Yes**
DQ540284	**31**	**456.90**	**548.23**	**124.83**	**196.56**	**1.57**	**Yes**
DQ540283	31	456.90	548.23	124.83	196.56	1.57	No
DQ543676	**31**	**438.90**	**518.70**	**119.91**	**185.97**	**1.55**	**Yes**
DQ701563	26	384.00	489.50	104.91	175.50	1.67	No
DQ541630	25	680.50	464.00	185.92	166.36	0.89	No
DQ705026	**29**	**154.00**	**377.00**	**42.08**	**135.17**	**3.21**	**Yes**
DQ551624	28	196.00	349.33	53.55	125.25	2.34	No
DQ551625	29	196.00	349.33	53.55	125.25	2.34	No
DQ701020	19	281.00	339.00	76.77	121.54	1.58	No
DQ715971	25	461.17	327.00	126.00	117.24	0.93	No
DQ710909	23	428.17	312.00	116.98	111.86	0.96	No
DQ719597	**28**	**168.00**	**306.00**	**45.90**	**109.71**	**2.39**	**Yes**
n200793	31	463.83	302.50	126.73	108.46	0.86	No
DQ711996	22	313.83	295.00	85.74	105.77	1.23	No
DQ724236	18	264.83	277.00	72.36	99.31	1.37	No
DQ541689	30	401.30	273.15	109.64	97.93	0.89	No
DQ720186	**23**	**336.00**	**251.00**	**91.80**	**89.99**	**0.98**	**Yes**
DQ714752	31	115.30	242.15	31.50	86.82	2.76	No
DQ559312	26	380.00	224.00	103.82	80.31	0.77	No
n202644	19	173.00	222.50	47.27	79.77	1.69	No
DQ725273	26	335.80	221.70	91.75	79.49	0.87	No
n204765	21	145.00	210.17	39.62	75.35	1.90	No
DQ558144	25	376.00	209.00	102.73	74.93	0.73	No
DQ696996	22	206.50	192.00	56.42	68.84	1.22	No
DQ707524	20	191.50	191.00	52.32	68.48	1.31	No
DQ553318	26	181.00	163.00	49.45	58.44	1.18	No
DQ714439	28	178.47	147.18	48.76	52.77	1.08	No
DQ555094	**32**	**202.00**	**140.00**	**55.19**	**50.19**	**0.91**	**Yes**
DQ716505	21	80.25	138.50	21.93	49.66	2.26	No
DQ719430	23	79.25	137.00	21.65	49.12	2.27	No
DQ706273	22	79.25	137.00	21.65	49.12	2.27	No
DQ712837	23	53.00	135.00	14.48	48.40	3.34	No
DQ549760	29	191.00	133.00	52.18	47.69	0.91	No
DQ555093	**29**	**189.50**	**129.50**	**51.77**	**46.43**	**0.90**	**Yes**
DQ550329	28	126.00	127.00	34.43	45.53	1.32	No
DQ540952	28	91.70	124.27	25.05	44.55	1.78	No
DQ540981	**30**	**126.50**	**124.00**	**34.56**	**44.46**	**1.29**	**Yes**
DQ709462	30	90.67	108.33	24.77	38.84	1.57	No

Supplementary Table 1, continued

piRNA accession	Length	Number of tags in wildtype cerebellum	Number of tags in Mecp2 KO cerebellum	WT_Tpm	Mecp2 KO_Tpm	Fold_change_ KO/WT	Reported as top 20 piRNAs in brain
DQ540984	30	62.93	103.50	17.19	37.11	2.16	No
DQ713872	22	81.00	97.00	22.13	34.78	1.57	No
DQ552696	30	120.00	91.00	32.79	32.63	1.00	No
DQ719488	24	248.00	90.00	67.76	32.27	0.48	No
DQ541470	27	127.72	89.05	34.89	31.93	0.91	No
DQ551913	27	57.33	80.25	15.66	28.77	1.84	No
DQ689086	**27**	**65.00**	**78.00**	**17.76**	**27.97**	**1.57**	**Yes**
DQ717385	22	89.50	64.33	24.45	23.07	0.94	No
DQ548183	28	65.50	60.00	17.90	21.51	1.20	No
DQ540859	30	77.50	58.50	21.17	20.97	0.99	No
DQ556354	30	20.00	58.50	5.46	20.97	3.84	No
DQ696831	31	33.00	54.00	9.02	19.36	2.15	No
DQ540872	32	107.50	54.00	29.37	19.36	0.66	No
DQ540944	26	47.68	53.08	13.03	19.03	1.46	No
DQ542796	28	49.00	53.00	13.39	19.00	1.42	No
DQ546708	30	43.00	52.00	11.75	18.64	1.59	No
DQ541352	25	44.93	51.08	12.28	18.32	1.49	No
DQ540974	29	66.50	49.50	18.17	17.75	0.98	No
DQ540526	28	32.00	49.00	8.74	17.57	2.01	No
DQ551739	28	22.67	47.22	6.19	16.93	2.73	No
DQ540403	30	24.00	46.00	6.56	16.49	2.52	No
DQ540915	30	54.00	46.00	14.75	16.49	1.12	No
DQ539904	28	45.00	45.50	12.29	16.31	1.33	No
DQ699095	24	20.00	44.50	5.46	15.95	2.92	No
DQ723924	20	20.00	44.00	5.46	15.78	2.89	No
DQ541614	28	18.38	43.58	5.02	15.63	3.11	No
DQ564913	30	29.33	41.92	8.01	15.03	1.88	No
DQ553409	27	45.50	41.50	12.43	14.88	1.20	No
DQ547181	28	18.00	41.00	4.92	14.70	2.99	No
DQ545450	28	32.00	41.00	8.74	14.70	1.68	No
DQ687520	26	41.93	39.08	11.46	14.01	1.22	No
DQ689768	30	26.00	39.00	7.10	13.98	1.97	No
DQ540964	25	42.50	39.00	11.61	13.98	1.20	No
DQ701846	29	46.00	38.00	12.57	13.62	1.08	No
DQ540058	27	17.00	37.00	4.64	13.27	2.86	No
DQ706530	31	34.00	36.17	9.29	12.97	1.40	No
DQ710928	30	54.00	36.00	14.75	12.91	0.87	No
DQ706818	28	52.00	36.00	14.21	12.91	0.91	No
DQ546549	25	33.83	36.00	9.24	12.91	1.40	No
DQ540988	28	23.17	35.50	6.33	12.73	2.01	No
DQ540862	**30**	**21.50**	**33.50**	**5.87**	**12.01**	**2.04**	**Yes**
DQ696491	29	41.72	32.92	11.40	11.80	1.04	No
DQ540861	**27**	**20.50**	**31.50**	**5.60**	**11.29**	**2.02**	**Yes**
DQ563946	30	5.00	31.00	1.37	11.11	8.14	No
DQ562907	29	9.73	30.70	2.66	11.01	4.14	No
DQ702901	29	13.00	30.67	3.55	11.00	3.10	No
DQ707624	26	14.93	30.07	4.08	10.78	2.64	No
DQ540780	26	16.33	30.00	4.46	10.76	2.41	No
DQ540412	27	37.00	29.00	10.11	10.40	1.03	No
DQ539915	**32**	**35.50**	**28.00**	**9.70**	**10.04**	**1.04**	**Yes**
DQ565590	31	1.00	28.00	0.27	10.04	36.74	No
DQ698557	30	10.93	27.65	2.99	9.91	3.32	No
DQ689686	22	28.75	27.50	7.85	9.86	1.26	No
DQ541113	31	15.50	27.50	4.23	9.86	2.33	No
DQ555883	30	59.75	27.30	16.32	9.79	0.60	No
DQ711586	30	27.00	27.00	7.38	9.68	1.31	No
DQ564866	27	14.83	27.00	4.05	9.68	2.39	No
DQ562906	27	8.73	26.70	2.39	9.57	4.01	No

Supplementary Table 1, continued

piRNA accession	Length	Number of tags in wildtype cerebellum	Number of tags in Mecp2 KO cerebellum	WT_Tpm	Mecp2 KO_Tpm	Fold_change_ KO/WT	Reported as top 20 piRNAs in brain
DQ564777	27	11.00	26.70	3.01	9.57	3.19	No
DQ562139	27	29.00	26.00	7.92	9.32	1.18	No
DQ725665	32	27.67	26.00	7.56	9.32	1.23	No
n205237	30	28.17	25.50	7.70	9.14	1.19	No
DQ702517	30	76.58	25.33	20.92	9.08	0.43	No
DQ725422	27	22.50	25.00	6.15	8.96	1.46	No
DQ540976	27	41.50	24.30	11.34	8.71	0.77	No
DQ564776	26	8.00	24.20	2.19	8.68	3.97	No
DQ709768	28	7.00	24.00	1.91	8.60	4.50	No
DQ540175	32	42.00	24.00	11.48	8.60	0.75	No
DQ540963	29	21.32	23.83	5.82	8.55	1.47	No
DQ691499	30	41.00	23.50	11.20	8.43	0.75	No
DQ701440	28	14.00	23.00	3.83	8.25	2.16	No
DQ539926	27	17.00	22.50	4.64	8.07	1.74	No
DQ724091	30	17.00	22.50	4.64	8.07	1.74	No
DQ715990	30	18.33	22.17	5.01	7.95	1.59	No
DQ719178	29	21.50	22.00	5.87	7.89	1.34	No
DQ687463	26	17.00	22.00	4.64	7.89	1.70	No
DQ552936	30	57.00	22.00	15.57	7.89	0.51	No
DQ703911	30	14.00	22.00	3.83	7.89	2.06	No
DQ555802	28	19.00	22.00	5.19	7.89	1.52	No
DQ540860	30	35.00	22.00	9.56	7.89	0.82	No
DQ707092	26	7.93	21.65	2.17	7.76	3.58	No
DQ555882	29	53.75	21.30	14.69	7.64	0.52	No
DQ721541	28	16.33	21.17	4.46	7.59	1.70	No
DQ715697	31	16.33	21.17	4.46	7.59	1.70	No
n199527	26	17.33	21.00	4.74	7.53	1.59	No
DQ709916	25	11.00	21.00	3.01	7.53	2.51	No
DQ565303	30	11.30	20.82	3.09	7.46	2.42	No
DQ718173	30	11.30	20.82	3.09	7.46	2.42	No
DQ692434	29	11.30	20.82	3.09	7.46	2.42	No
DQ691288	24	20.92	20.50	5.71	7.35	1.29	No
n199120	16	38.00	20.50	10.38	7.35	0.71	No
DQ715526	**28**	**20.00**	**20.00**	**5.46**	**7.17**	**1.31**	**Yes**
DQ567738	29	8.00	20.00	2.19	7.17	3.28	No
DQ720914	31	15.00	20.00	4.10	7.17	1.75	No
DQ540280	30	14.50	20.00	3.96	7.17	1.81	No
DQ540869	29	16.52	19.93	4.51	7.15	1.58	No
DQ540975	26	27.50	19.50	7.51	6.99	0.93	No
DQ541100	28	15.50	19.00	4.23	6.81	1.61	No
DQ712821	24	21.00	19.00	5.74	6.81	1.19	No
DQ545972	30	19.00	19.00	5.19	6.81	1.31	No
DQ702126	29	18.15	18.25	4.96	6.54	1.32	No
DQ703459	30	13.00	18.17	3.55	6.51	1.83	No
DQ699219	30	21.00	18.00	5.74	6.45	1.12	No
DQ540965	32	15.17	18.00	4.14	6.45	1.56	No
DQ707037	28	8.55	17.60	2.34	6.31	2.70	No
DQ541147	28	10.50	17.50	2.87	6.27	2.19	No
DQ555881	28	43.25	17.30	11.82	6.20	0.52	No
DQ568996	31	7.00	17.00	1.91	6.10	3.19	No
n202030	32	11.67	17.00	3.19	6.10	1.91	No
DQ703779	25	7.00	17.00	1.91	6.10	3.19	No
DQ545225	30	5.00	17.00	1.37	6.10	4.46	No
DQ541000	28	13.18	16.93	3.60	6.07	1.69	No
n197343	38	13.00	16.50	3.55	5.92	1.67	No
DQ540868	26	13.18	16.10	3.60	5.77	1.60	No
DQ541218	26	79.00	16.00	21.58	5.74	0.27	No
DQ718197	27	10.50	16.00	2.87	5.74	2.00	No

Supplementary Table 1, continued

piRNA accession	Length	Number of tags in wildtype cerebellum	Number of tags in Mecp2 KO cerebellum	WT_Tpm	Mecp2 KO_Tpm	Fold_change_ KO/WT	Reported as top 20 piRNAs in brain
DQ714526	30	18.07	15.67	4.94	5.62	1.14	No
DQ540689	29	29.50	15.50	8.06	5.56	0.69	No
DQ540059	26	11.00	15.50	3.01	5.56	1.85	No
DQ724251	22	11.75	15.40	3.21	5.52	1.72	No
DQ714788	30	26.50	15.33	7.24	5.50	0.76	No
DQ540867	25	11.93	15.10	3.26	5.41	1.66	No
DQ693545	30	4.67	15.00	1.28	5.38	4.22	No
DQ717257	29	17.00	14.50	4.64	5.20	1.12	No
DQ544489	29	8.92	14.37	2.44	5.15	2.11	No
DQ551953	28	12.23	14.25	3.34	5.11	1.53	No
DQ541776	29	3.65	14.23	1.00	5.10	5.12	No
DQ541806	27	19.00	14.00	5.19	5.02	0.97	No
DQ540253	31	27.00	14.00	7.38	5.02	0.68	No
DQ686298	21	7.00	14.00	1.91	5.02	2.62	No
DQ706110	30	23.50	14.00	6.42	5.02	0.78	No
DQ545604	27	17.00	14.00	4.64	5.02	1.08	No
n202750	30	12.00	14.00	3.28	5.02	1.53	No
DQ696259	29	6.00	14.00	1.64	5.02	3.06	No
DQ723396	30	6.00	13.50	1.64	4.84	2.95	No
DQ555880	27	24.25	13.30	6.63	4.77	0.72	No
DQ716469	31	5.40	13.08	1.48	4.69	3.18	No
DQ722288	**28**	**2.00**	**13.00**	**0.55**	**4.66**	**8.53**	**Yes**
DQ558886	26	4.00	13.00	1.09	4.66	4.26	No
DQ548138	27	6.00	13.00	1.64	4.66	2.84	No
DQ566603	30	9.00	13.00	2.46	4.66	1.90	No
DQ694433	25	19.00	13.00	5.19	4.66	0.90	No
DQ559729	29	16.00	13.00	4.37	4.66	1.07	No
DQ541629	26	10.18	12.93	2.78	4.64	1.67	No
DQ540202	31	16.50	12.50	4.51	4.48	0.99	No
DQ710188	27	2.00	12.00	0.55	4.30	7.87	No
DQ541627	28	11.00	12.00	3.01	4.30	1.43	No
DQ725115	29	46.00	12.00	12.57	4.30	0.34	No
DQ715208	24	0.50	12.00	0.14	4.30	31.49	No
DQ718174	22	5.50	12.00	1.50	4.30	2.86	No
DQ690565	31	9.00	12.00	2.46	4.30	1.75	No
DQ721627	30	36.00	12.00	9.84	4.30	0.44	No
DQ563182	31	4.00	12.00	1.09	4.30	3.94	No
DQ725966	19	12.33	11.67	3.37	4.18	1.24	No
DQ717747	32	14.00	11.33	3.83	4.06	1.06	No
DQ698641	27	1.00	11.00	0.27	3.94	14.44	No
DQ705481	22	7.00	11.00	1.91	3.94	2.06	No
DQ702236	29	7.00	11.00	1.91	3.94	2.06	No
DQ697536	31	9.10	10.92	2.49	3.91	1.57	No
DQ693633	30	17.67	10.50	4.83	3.76	0.78	No
DQ692951	30	9.07	10.50	2.48	3.76	1.52	No
DQ558403	26	8.33	10.42	2.28	3.73	1.64	No
DQ716691	28	11.87	10.33	3.24	3.70	1.14	No
DQ548430	30	11.17	10.08	3.05	3.62	1.18	No
DQ709071	29	8.17	10.00	2.23	3.59	1.61	No
DQ568824	30	4.00	10.00	1.09	3.59	3.28	No
DQ709273	31	11.00	10.00	3.01	3.59	1.19	No
DQ719680	26	5.00	10.00	1.37	3.59	2.62	No
DQ719784	21	12.00	10.00	3.28	3.59	1.09	No
DQ691624	22	12.60	10.00	3.44	3.59	1.04	No
DQ719096	29	6.10	9.98	1.67	3.58	2.15	No
DQ693813	30	13.17	9.98	3.60	3.58	1.00	No
DQ709946	26	4.83	9.83	1.32	3.53	2.67	No
DQ540134	28	3.00	9.83	0.82	3.53	4.30	No

Supplementary Table 1, continued

piRNA accession	Length	Number of tags in wildtype cerebellum	Number of tags in Mecp2 KO cerebellum	WT_Tpm	Mecp2 KO_Tpm	Fold_change_ KO/WT	Reported as top 20 piRNAs in brain
DQ551740	30	4.80	9.75	1.31	3.50	2.67	No
DQ551741	32	5.00	9.75	1.37	3.50	2.56	No
DQ708438	31	0.50	9.50	0.14	3.41	24.93	No
n204129	26	10.50	9.50	2.87	3.41	1.19	No
DQ698521	27	11.50	9.50	3.14	3.41	1.08	No
DQ550765	**31**	**10.75**	**9.50**	**2.94**	**3.41**	**1.16**	**Yes**
DQ707509	31	22.50	9.25	6.15	3.32	0.54	No
DQ712474	28	22.50	9.25	6.15	3.32	0.54	No
DQ712486	30	22.50	9.25	6.15	3.32	0.54	No
DQ724038	21	7.23	9.17	1.98	3.29	1.66	No
DQ717846	29	7.83	9.08	2.14	3.26	1.52	No
DQ727278	29	3.00	9.00	0.82	3.23	3.94	No
DQ703937	31	3.17	9.00	0.87	3.23	3.73	No
DQ553641	29	5.00	9.00	1.37	3.23	2.36	No
DQ698382	31	7.00	9.00	1.91	3.23	1.69	No
DQ547060	30	5.83	8.92	1.59	3.20	2.01	No
DQ554102	30	4.83	8.57	1.32	3.07	2.33	No
DQ541631	30	5.75	8.50	1.57	3.05	1.94	No
DQ694480	28	4.42	8.33	1.21	2.99	2.48	No
DQ562379	25	8.83	8.17	2.41	2.93	1.21	No
DQ703493	26	7.05	8.13	1.93	2.92	1.51	No
DQ699015	30	8.00	8.00	2.19	2.87	1.31	No
DQ565695	31	4.00	8.00	1.09	2.87	2.62	No
DQ687262	30	16.00	8.00	4.37	2.87	0.66	No
DQ540853	26	1.00	8.00	0.27	2.87	10.50	No
DQ541719	29	4.00	8.00	1.09	2.87	2.62	No
DQ721809	29	5.00	8.00	1.37	2.87	2.10	No
DQ686705	26	3.00	8.00	0.82	2.87	3.50	No
DQ719574	27	9.50	8.00	2.60	2.87	1.11	No
DQ698794	29	11.00	8.00	3.01	2.87	0.95	No
DQ569658	30	13.67	7.67	3.73	2.75	0.74	No
DQ684777	30	16.33	7.67	4.46	2.75	0.62	No
DQ548429	28	8.67	7.58	2.37	2.72	1.15	No
DQ569911	31	3.83	7.50	1.05	2.69	2.57	No
DQ714299	28	2.25	7.50	0.61	2.69	4.37	No
DQ705429	28	2.02	7.48	0.55	2.68	4.87	No
DQ700644	30	10.00	7.33	2.73	2.63	0.96	No
DQ542520	32	8.40	7.27	2.30	2.61	1.14	No
DQ725358	22	7.88	7.22	2.15	2.59	1.20	No
DQ715253	28	4.57	7.08	1.25	2.54	2.04	No
DQ540081	28	5.00	7.00	1.37	2.51	1.84	No
DQ559964	27	1.00	7.00	0.27	2.51	9.19	No
DQ555047	31	13.00	7.00	3.55	2.51	0.71	No
DQ546953	32	7.00	7.00	1.91	2.51	1.31	No
DQ697785	18	18.00	7.00	4.92	2.51	0.51	No
DQ712498	27	14.00	7.00	3.83	2.51	0.66	No
DQ703255	29	2.00	7.00	0.55	2.51	4.59	No
DQ703079	27	5.00	7.00	1.37	2.51	1.84	No
DQ559781	28	5.00	7.00	1.37	2.51	1.84	No
DQ726397	22	1.00	7.00	0.27	2.51	9.19	No
DQ540951	28	106.00	7.00	28.96	2.51	0.09	No
DQ707442	32	1.17	7.00	0.32	2.51	7.87	No
DQ695733	31	1.17	7.00	0.32	2.51	7.87	No
DQ550956	30	2.00	6.67	0.55	2.39	4.37	No
DQ704045	30	2.00	6.67	0.55	2.39	4.37	No
DQ698371	26	4.50	6.50	1.23	2.33	1.90	No
DQ540939	26	4.50	6.50	1.23	2.33	1.90	No
DQ686264	26	5.33	6.33	1.46	2.27	1.56	No

Supplementary Table 1, continued

piRNA accession	Length	Number of tags in wildtype cerebellum	Number of tags in Mecp2 KO cerebellum	WT_Tpm	Mecp2 KO_Tpm	Fold_change_ KO/WT	Reported as top 20 piRNAs in brain
DQ550027	30	1.20	6.33	0.33	2.27	6.93	No
DQ543701	30	10.17	6.33	2.78	2.27	0.82	No
DQ554152	31	3.85	6.18	1.05	2.22	2.11	No
DQ550614	31	7.42	6.13	2.03	2.20	1.09	No
DQ557476	29	5.08	6.08	1.39	2.18	1.57	No
DQ723756	31	0.83	6.08	0.23	2.18	9.58	No
DQ694583	30	2.00	6.00	0.55	2.15	3.94	No
DQ541882	32	2.25	6.00	0.61	2.15	3.50	No
DQ563242	30	1.00	6.00	0.27	2.15	7.87	No
DQ714655	21	2.50	6.00	0.68	2.15	3.15	No
DQ708131	**27**	**3.00**	**6.00**	**0.82**	**2.15**	**2.62**	**Yes**
DQ565679	30	2.50	6.00	0.68	2.15	3.15	No
DQ695473	27	3.50	6.00	0.96	2.15	2.25	No
DQ551797	29	5.00	6.00	1.37	2.15	1.57	No
n202533	30	3.00	6.00	0.82	2.15	2.62	No
DQ719363	29	3.50	6.00	0.96	2.15	2.25	No
DQ564810	31	3.00	6.00	0.82	2.15	2.62	No
DQ542432	29	3.50	6.00	0.96	2.15	2.25	No
DQ687025	30	2.00	6.00	0.55	2.15	3.94	No
DQ705141	29	3.00	6.00	0.82	2.15	2.62	No
DQ550009	32	2.50	6.00	0.68	2.15	3.15	No
DQ712132	27	3.92	5.92	1.07	2.12	1.98	No
DQ710634	26	3.92	5.92	1.07	2.12	1.98	No
DQ688047	31	0.90	5.87	0.25	2.10	8.55	No
DQ540133	27	3.00	5.83	0.82	2.09	2.55	No
DQ724045	31	5.17	5.83	1.41	2.09	1.48	No
DQ717785	29	3.67	5.67	1.00	2.03	2.03	No
DQ697835	30	0.90	5.62	0.25	2.01	8.19	No
DQ550424	30	6.95	5.62	1.90	2.01	1.06	No
DQ699107	32	5.08	5.55	1.39	1.99	1.43	No
DQ692222	31	4.08	5.55	1.12	1.99	1.78	No
DQ559632	29	5.03	5.52	1.38	1.98	1.44	No
DQ539909	31	3.50	5.50	0.96	1.97	2.06	No
DQ699690	29	3.00	5.50	0.82	1.97	2.41	No
DQ694268	30	6.17	5.50	1.68	1.97	1.17	No
DQ704413	30	3.50	5.50	0.96	1.97	2.06	No
DQ540217	30	2.50	5.50	0.68	1.97	2.89	No
DQ705397	27	3.00	5.50	0.82	1.97	2.41	No
DQ541518	32	2.83	5.42	0.77	1.94	2.51	No
DQ561657	31	2.87	5.30	0.78	1.90	2.43	No
DQ713688	20	6.47	5.27	1.77	1.89	1.07	No
DQ695662	27	0.20	5.25	0.05	1.88	34.45	No
DQ554873	32	1.92	5.03	0.52	1.80	3.45	No
DQ718455	31	1.92	5.03	0.52	1.80	3.45	No
DQ569362	29	2.00	5.00	0.55	1.79	3.28	No
DQ711635	30	2.00	5.00	0.55	1.79	3.28	No
DQ541249	31	5.00	5.00	1.37	1.79	1.31	No
DQ718385	28	6.00	5.00	1.64	1.79	1.09	No
DQ701776	29	38.00	5.00	10.38	1.79	0.17	No
DQ561111	29	2.00	5.00	0.55	1.79	3.28	No
DQ697860	31	8.00	5.00	2.19	1.79	0.82	No
DQ541101	27	5.50	5.00	1.50	1.79	1.19	No
DQ691503	28	0.00	5.00	0.00	1.79	5.00	No
DQ701204	27	6.00	5.00	1.64	1.79	1.09	No
DQ695609	26	8.50	5.00	2.32	1.79	0.77	No
DQ545504	27	10.00	5.00	2.73	1.79	0.66	No
DQ696494	26	0.00	5.00	0.00	1.79	5.00	No
DQ726803	30	1.00	5.00	0.27	1.79	6.56	No

Supplementary Table 1, continued

piRNA accession	Length	Number of tags in wildtype cerebellum	Number of tags in Mecp2 KO cerebellum	WT_Tpm	Mecp2 KO_Tpm	Fold_change_ KO/WT	Reported as top 20 piRNAs in brain
DQ695702	32	0.00	5.00	0.00	1.79	5.00	No
DQ563823	30	1.00	5.00	0.27	1.79	6.56	No
DQ551895	27	5.00	5.00	1.37	1.79	1.31	No
DQ567700	26	5.50	5.00	1.50	1.79	1.19	No
DQ688232	31	6.50	5.00	1.78	1.79	1.01	No
DQ700848	28	3.00	5.00	0.82	1.79	2.19	No
DQ710935	30	1.00	5.00	0.27	1.79	6.56	No
DQ563772	27	23.00	5.00	6.28	1.79	0.29	No
DQ691599	28	6.00	5.00	1.64	1.79	1.09	No
DQ721750	21	2.00	5.00	0.55	1.79	3.28	No
DQ722234	31	0.50	5.00	0.14	1.79	13.12	No
DQ709374	31	5.25	5.00	1.43	1.79	1.25	No
DQ709264	32	8.00	5.00	2.19	1.79	0.82	No
DQ554082	30	3.50	5.00	0.96	1.79	1.87	No
DQ541574	29	4.50	5.00	1.23	1.79	1.46	No
DQ552422	31	2.50	5.00	0.68	1.79	2.62	No
		356282.87	518589.20	97341.75	185931.84	1.91	19

Chapter 5

CCL2 activates hypoxia related genes

Chapter 6

Regulated expression of repetitive elements

Chapter 7

The blood transcriptome

Chapter 8

General discussion

Small non-coding RNAs have been implicated in many biological processes, such as messenger RNA regulation or transposon silencing. We identified a role of small non-coding RNAs in the DNA damage repair mechanism. The inactivation of dicer and drosha, which are required for the biogenesis of small RNAs, leads to a loss of DNA damage repair. We sequenced the small RNAs from cells that had induced DNA damage and found small RNAs arising from the vicinity of the DNA double-strand break. Importantly, synthetic small RNAs mimicking these small RNAs could drive the DNA damage response [1].

Parkinsons disease (PD) is a slowly progressive disease in which dopamine neurons in the substantia nigra degenerate undetected for years before clinical symptoms develop. The lack of clinical symptoms highlights the necessity of a laboratory test, such as an assay for biomarkers, which can correlate subjects with PD risk. We have profiled the RNAs in the whole blood sample of PD patients and age-matched controls using high-throughput deepCAGE sequencing. By comparing the RNA profiles between PD patients and controls, we aim to discover novel biomarkers that are present in whole blood, which may be further developed into a non-invasive clinical test for PD.

8.0.1 Experimental differences

poly-A versus random primers cytoplasmic versus nuclear enrichment CAGE versus RNA-seq

8.0.2 Signal versus noise

biological variance and technical variance expression profiles from sample of cells versus single cell highly expressed versus lowly expressed transcripts

8.0.3 Coding versus non-coding transcripts

Transcripts of Unknown Function (TUFs) The role of non-coding RNAs The role of repetitive elements in the genome

8.0.4 Transcriptome profiling using CAGE

Unbiased profiling of total transcripts Comparison with different environmental conditions Gene ontology enrichment

Acknowledgements

> "Twenty years from now you will be more disappointed by the things that you didn't do than by the ones you did do. So throw off the bowlines. Sail away from the safe harbor. Catch the trade winds in your sails. Explore. Dream. Discover."
>
> — Mark Twain

The culmination of work presented in this thesis would not have been possible without the guidance and support of many. But my involvement in this PhD project would not have even begun if not for my supervisor who accepted me into the program and a former colleague who forwarded the opening to me. I still remember the predicament I faced over 4 years ago when I had to decide whether or not I would commit to the idea of potentially working in Japan. I had already brushed off the idea once but in the end I was convinced that it was a tremendous opportunity and I would have regretted it if I let it slip by. In the end, I decided to set my sails to explore an entirely new world.

To this day, I have absolutely no regrets for embarking on this journey; I had to leave behind close friends of many years but I still remember their responses when I asked for their advice regarding the opportunity. Their answers were unanimous; clearly, they had had enough of me. I arrived in Japan knowing just one single person, who I got to know through a Perl mailing list. Though we were complete strangers, he helped me settle into Japan and I'm forever thankful. At work, I met a lot of awesome people. The first order of business was to find out where I could play basketball and I was introduced to my basketball buddy for the last 3 years. Thank you for preventing me from suffering from basketball withdrawal! There are two colleagues who despite my hermit tendencies still occasionally ask me to hang out. Thank you for not giving up on me, yet. I also had the pleasure of meeting some highly dedicated researchers whom I greatly admire, for their scientific knowledge and personality. I have learned a lot from you all and I have tried to shape myself in your footsteps. I would also like to acknowledge the Good Samaritans who answer questions on discussion forums and mailing lists; where would I be without them.

Finally, thank *you* from the bottom of my heart.

Bibliography

[1] Sofia Francia, Flavia Michelini, Alka Saxena, Dave Tang, Michiel de Hoon, Viviana Anelli, Marina Mione, Piero Carninci, and Fabrizio dAdda di Fagagna. Site-specific dicer and drosha rna products control the dna-damage response. *Nature*, 488(7410):231–235, 2012.

[2] Alka Saxena, Dave Tang, and Piero Carninci. pirnas warrant investigation in rett syndrome: An omics perspective. *Disease markers*, 33(5):261–275, 2012.

[3] Dave T. P. Tang, Charles Plessy, Md Salimullah, Ana Maria Suzuki, Raffaella Calligaris, Stefano Gustincich, and Piero Carninci. Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. *Nucleic Acids Research*, 41(3):e44, 2013.

[4] Ralf Dahm. Discovering dna: Friedrich miescher and the early years of nucleic acid research. *Human genetics*, 122(6):565–581, 2008.

[5] Fred Griffith. The significance of pneumococcal types. *Journal of Hygiene*, 27(02):113–159, 1928.

[6] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *The Journal of experimental medicine*, 79(2):137–158, 1944.

[7] Alfred D Hershey and Martha Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of general physiology*, 36(1):39–56, 1952.

[8] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[9] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[10] Erika Check Hayden. Is the $1,000 genome for real? *Nature*, Jan 2014.

[11] http://www.genome.gov/11006943/. Accessed: 2014-05-21.

[12] D. Baltimore. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, 226(5252):1209–1211, Jun 1970.

[13] H. M. Temin and S. Mizutani. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, 226(5252):1211–1213, Jun 1970.

[14] http://www.nature.com/nrg/series/nextgeneration/index.html. Accessed: 2014-05-21.

[15] Lior Pachter. http://liorpachter.wordpress.com/seq/. Accessed: 2014-06-02.

[16] E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, and S. Canaider. An estimation of the number of cells in the human body. *Ann. Hum. Biol.*, 40(6):463–471, 2013.

[17] M. K. Vickaryous and B. K. Hall. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol Rev Camb Philos Soc*, 81(3):425–455, Aug 2006.

[18] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. What is a gene, post-ENCODE? History and updated definition. *Genome Res.*, 17(6):669–681, Jun 2007.

[19] E. L. Tatum and G. W. Beadle. Genetic Control of Biochemical Reactions in Neurospora: An "Aminobenzoicless" Mutant. *Proc. Natl. Acad. Sci. U.S.A.*, 28(6):234–243, Jun 1942.

[20] J. D. WATSON and F. H. C. CRICK. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737738, Apr 1953.

[21] J. H. MATTHAEI, O. W. JONES, R. G. MARTIN, and M. W. NIRENBERG. Characteristics and composition of RNA coding units. *Proc. Natl. Acad. Sci. U.S.A.*, 48:666–677, Apr 1962.

[22] M. Nirenberg, P. Leder, M. Bernfield, R. Brimacombe, J. Trupin, F. Rottman, and C. O'Neal. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc. Natl. Acad. Sci. U.S.A.*, 53(5):1161–1168, May 1965.

[23] Francis H Crick. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138, 1958.

[24] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–5467, Dec 1977.

[25] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.*, 74(2):560–564, Feb 1977.

[26] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679, 1986.

[27] Lex Nederbragt. developments in NGS. `http://dx.doi.org/10.6084/m9.figshare.100940`, 12 2012.

[28] S. Arnott, R. Chandrasekaran, D. L. Birdsall, A. G. Leslie, and R. L. Ratliff. Left-handed DNA helices. *Nature*, 283(5749):743–745, Feb 1980.

[29] K.E. Van Holde, C.G. Sahasrabuddhe, and Barbara Ramsay Shaw. A model for particulate structure in chromatin. *Nucleic Acids Research*, 1(11):1579–1586, 1974.

[30] J. Logan, E. Falck-Pedersen, J. E. Darnell, and T. Shenk. A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. *Proc. Natl. Acad. Sci. U.S.A.*, 84(23):8306–8310, Dec 1987.

[31] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, Oct 1995.

[32] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, Oct 1997.

[33] Dave Tang. Coding probability of mouse full-length cDNAs from FANTOM3. `http://dx.doi.org/10.6084/m9.figshare.1046601`, 06 2014.

[34] P. Carninci, C. Kvam, A. Kitamura, T. Ohsumi, Y. Okazaki, M. Itoh, M. Kamiya, K. Shibata, N. Sasaki, M. Izawa, M. Muramatsu, Y. Hayashizaki, and C. Schneider. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, 37(3):327–336, Nov 1996.

[35] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, et al. The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563, Sep 2005.

[36] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, 38(6):626–635, Jun 2006.

[37] R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, Dec 1993.

[38] E. Bernstein, A. A. Caudy, S. M. Hammond, and G. J. Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–366, Jan 2001.

[39] The C. elegans Sequencing Consortium. Genome sequence of the nematode c. elegans: A platform for investigating biology. *Science*, 282(5396):2012–2018, 1998.

[40] Dave Tang. Percentage coverage of repetitive elements in vertebrate genomes. http://dx.doi.org/10.6084/m9.figshare.1033768, 05 2014.

[41] W Ford Doolittle and Carmen Sapienza. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–3, 1980.

[42] Leslie E Orgel and Francis H Crick. Selfish dna: the ultimate parasite. *Nature*, 284(5757):604–607, 1980.

[43] Dave Tang. Vertebrate genome sizes sorted by repetitive elements percentage. http://dx.doi.org/10.6084/m9.figshare.1033808, 05 2014.

[44] G. J. Faulkner, Y. Kimura, C. O. Daub, S. Wani, C. Plessy, K. M. Irvine, K. Schroder, N. Cloonan, A. L. Steptoe, T. Lassmann, K. Waki, N. Hornig, T. Arakawa, H. Takahashi, J. Kawai, A. R. Forrest, H. Suzuki, Y. Hayashizaki, D. A. Hume, V. Orlando, S. M. Grimmond, and P. Carninci. The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, 41(5):563–571, May 2009.

[45] A. R. Forrest, H. Kawaji, M. Rehli, J. K. Baillie, M. J. de Hoon, et al. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470, Mar 2014.

[46] A. J. BECKER, E. A. McCULLOCH, and J. E. TILL. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature*, 197:452–454, Feb 1963.

[47] M. J. Evans and M. H. Kaufman. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292(5819):154–156, Jul 1981.

[48] G. R. Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 78(12):7634–7638, Dec 1981.

[49] K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, Aug 2006.

[50] X. Y. Zhao, W. Li, Z. Lv, L. Liu, M. Tong, T. Hai, J. Hao, C. L. Guo, Q. W. Ma, L. Wang, F. Zeng, and Q. Zhou. iPS cells produce viable mice through tetraploid complementation. *Nature*, 461(7260):86–90, Sep 2009.

[51] K. Yusa, S. T. Rashid, H. Strick-Marchand, I. Varela, et al. Targeted gene correction of 1-antitrypsin deficiency in induced pluripotent stem cells. *Nature*, 478(7369):391–394, Oct 2011.

[52] P. Hogeweg. The roots of bioinformatics in theoretical biology. *PLoS Comput. Biol.*, 7(3):e1002021, Mar 2011.

[53] Lincoln Stein. How Perl Saved the Human Genome Project. http://www.bioperl.org/wiki/How_Perl_saved_human_genome. Accessed: 2014-06-06.

[54] Jorge L Contreras. Bermudas legacy: Policy, patents, and the design of the genome commons. *Minn. JL Sci. & Tech.*, 12:61–97, 2011.

[55] P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, 38(6):1767–1771, Apr 2010.

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

[57] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[58] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010.

[59] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan 2010.

[60] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5(10):R80, 2004.