High-throughput sequencing and transcriptomics

Dave Ting Pong Tang

November 10, 2014

Abstract

Recent developments in DNA sequencing, has led to the development of high-throughput sequencing, which has massively scaled up the process of DNA sequencing in a time- and cost-effective manner. With the completion of the various mammalian genome projects, the focus has shifted towards identifying functional elements in genomes. The field of transcriptomics aims to identify, annotate, and analyse the transcribed products of the genome. The application of high-throughput sequencing to transcriptome profiling has allowed accurate unbiased profiling of transcripts and revealed the complexity of mammalian transcriptomes. However, these applications are relatively immature, as they have been developed within the last 5-6 years, and therefore are constantly being investigated and improved upon. This thesis focuses on the analysis of transcriptomes through the use of high-throughput sequencing and bioinformatic methods.

In the first study of this thesis, biases and technical artefacts in a transcriptome technology known as nano cap analysis gene expression (nanoCAGE) were investigated. We found that biases were introduced through the use of molecular barcodes and developed several strategies for coping with such biases. In the second study, we captured and analysed the transcriptional output of cells with induced DNA damage using small RNA sequencing. We observed a previously uncharacterised class of RNAs that formed near the DNA break site and were necessary in establishing the DNA damage response. In the third and last study, we focused on transcripts that initiated from repetitive elements (REs) and characterised their expression patterns across a wide panel of cell lines, tissues, and primary cells. We demonstrated that REs may drive the expression of long non-coding RNAs and enhancer RNAs, and their expression patterns are tissue specific.

High-throughput sequencing has transformed the field of transcriptomics by revealing a much more complex picture of transcription than previously anticipated. The layers of complexity include the expression of multiple classes of RNA species, each performing various regulatory roles, random transcriptional events or transcriptional noise, and complexity perpetrated from technical artefacts. In order to separate noise from signal requires an understanding of the different technologies, careful scrutiny of the data, and the application of appropriate bioinformatic methods.

List of publications

1. 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

2. 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

3. Dave Tang and Piero Carninci. The regulated expression of repetitive elements across human cell types and tissues. *In preparation*

Other publications

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

5. Yuki Hasegawa, Dave Tang, Naoko Takahashi, Yoshihide Hayashizaki, Alistair R. R. Forrest, the FANTOM consortium, and Harukazu Suzuki. Ccl2 enhances pluripotency of human induced pluripotent stem cells by activating hypoxia related genes. *Sci. Rep.*, 4, Jun 2014

6. Dave Tang, Ana Maria Suzuki, Raffaella Calligaris, Stefano Gustincich, and Piero Carninci. Deep transcriptome sequencing of whole blood samples from Parkinson's disease patients. *In preparation*

Contents

1 Introduction **9**
- 1.1 A brief history of DNA . . . 9
- 1.2 The Central Dogma of Molecular Biology . . . 10
- 1.3 Transcription . . . 11
  - 1.3.1 Transcriptional regulation . . . 12
  - 1.3.2 DNA accessibility . . . 14
- 1.4 DNA sequencing . . . 15
  - 1.4.1 Next-generation sequencing . . . 17
  - 1.4.2 Third generation sequencing and beyond . . . 18
- 1.5 Expression analysis . . . 20
  - 1.5.1 Transcriptome profiling . . . 20
  - 1.5.2 Transcriptional complexity . . . 21
  - 1.5.3 Defining a gene . . . 23
  - 1.5.4 Non-coding RNAs . . . 23
- 1.6 Repetitive mammalian genomes . . . 24
  - 1.6.1 Junk DNA . . . 25
  - 1.6.2 Impact of transposable elements on genomes . . . 26
- 1.7 Bioinformatics and genomics . . . 27
  - 1.7.1 High-throughput sequencing data . . . 28
  - 1.7.2 Analysing expression datasets . . . 29

2 Template switching artifacts **31**

3 Role of small RNAs in DNA damage repair **44**

4 Regulated expression of repetitive elements **53**

5 General discussion **70**

List of Figures

1.1 DNA base pairing 10
1.2 The central dogma 11
1.3 Core promoter elements 12
1.4 DNA transcription 13
1.5 DNA packaging 14
1.6 Radioactively labelled sequencing gel 16
1.7 Sanger sequencing 17
1.8 Developments in next generation sequencing 19
1.9 Cap Analysis Gene Expression protocol 22
1.10 Coverage of repetitive elements in vertebrate genomes 26

List of Tables

1.1 Table of histone modifications 16
1.2 Table of non-coding RNAs 24

List of Abbreviations

3D	Three-dimensional.
A	Adenine.
ANOVA	Analysis of variance.
ASC	Adult stem cell.
BP	Base pair.
BRE	B recognition element.
C	Cytosine.
CAGE	Cap analysis gene expression.
CCL2	chemokine (C-C motif) ligand 2.
cDNA	Complementary DNA.
CGI	CpG islands.
ChIP	Chromatin immunoprecipitation.
CPAT	Coding-potential assessment tool.
CRT	Cyclic reversible termination.
dATP	Deoxyadenosine triphosphate.
DBD	DNA-binding domain.
dCTP	Deoxyguanosine triphosphate.
ddNTP	Dideoxynucleotide triphosphate.
DDR	DNA damage response.
dGTP	Deoxycytidine triphosphate.
DNA	Deoxyribonucleic acid.
dTTP	Deoxythymidine triphosphate.
emPCR	Emulsion PCR.
ENCODE	Encyclopedia of DNA elements.
ERV	Endogenous retrovirus.
ESC	Embryonic stem cell.
FANTOM	Functional annotation of the mammalian genome.
FDR	False discovery rate.
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.
MecP2	Methyl CpG binding protein 2.
miRNA	Micro RNA.
mRNA	Messenger RNA.
NaCl	Sodium chloride.
nanoCAGE	Nano Cap analysis gene expression.
ncRNA	Non-coding RNA.
PAGE	Polyacrylamide gel electrophoresis.
PCR	Polymerase chain reaction.
PET	Paired-end ditag.
piRISC	piRNA-induced silencing complex.
piRNA	Piwi-interacting RNA.
Pol I	RNA polymerase I.
Pol II	RNA polymerase II.
Pol III	RNA polymerase III.
pre-miRNA	Precursor miRNA.
pri-miRNA	Primary miRNA.
QC	Quality control.
qRT-PCR	Quantitative real-time polymerase chain reaction.
RABS	Repeat-associated binding sites.
RAP	Repeat Analysis Pipeline.
RE	Repetitive elements.
RIDL	Repeat Insertion Domains of LncRNAs.
RNA	Ribonucleic acid.
RNA pol	RNA polymerase.
RT	Reverse transcriptase.
SAGE	Serial analysis gene expression.
SAM	Sequence alignment/map.
SBL	Sequencing by ligation.
SBS	Sequencing by synthesis.
SINE	Short interspersed elements.
SNA	Single-nucleotide addition.
SOLiD	Sequencing by Oligonucleotide Ligation and Detection.
T	Thymine.

TBP	TATA binding protein.
TE	Transposable element.
TF	Transcription factor.
TFBS	Transcription factor binding site.
TFIIB	Transcription factor IIB.
TMM	Trimmed mean of M values.
tRNA	Transfer RNA.
TS	Template switching.
TSS	Transcription start site.
TU	Transcriptional unit.
TUF	Transcripts of unknown function.
U	Uracil.

Chapter 1

Introduction

1.1 A brief history of DNA

In the winter of 1868/9, Swiss physician and biologist, Johannes Friedrich Miescher isolated an unknown substance from the nuclei of cells[7]. This substance was unlike anything he had observed before; it was resistant to protease, lacked sulphur, and contained a large amount of phosphorous. He recognised that he had isolated a novel substance and as it was from the nucleus, he 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). Later in 1889, Richard Altmann discovered that nuclein was acidic (due to the presence of phosphorous) and renamed nuclein to nucleic acid. The basic component of deoxyribonucleic acid (DNA) was deduced by Phoebus Levene in 1909, who discovered that DNA consisted of an acid, an organic base, and a sugar. Levene 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 makes DNA highly stable. In 1928, Frederick Griffith demonstrated that heritable traits could be transferred between dead and live bacteria and that provided the first clue that a "transforming factor" existed[8]. It wasn't until 1944, when Oswald Avery, Colin MacLeod, and Maclyn McCarty demonstrated that deoxyribonucleodepolymerase (an enzyme that degrades DNA) destroyed the "transforming factor", that it was hypothesised DNA was the genetic material[9]. This was later confirmed in 1952 by Alfred Hershey and Martha Chase, by demonstrating that when bacteriophages infected bacteria, only their DNA would enter into the cytoplasm of the bacteria, while their protein remained outside[10].

While Levene proposed that DNA was made up of equal amounts of A, C, G, and T, it was later discovered by Erwin Chargaff that DNA had a one-to-one ratio of pyrimidine (C, T, and U) and purine (A and G) bases[11, 12]; this became known as Chargaff's rules. This observation by Chargaff and insights gained from Rosalind Franklin were necessary for the deduction of the three-dimensional (3D) structure of DNA by Francis Crick and James Watson in 1953[13]. The 3D structure of DNA demonstrated how adenines paired with

thymines and cytosines paired with guanine (Figure 1.1); this became known as Watson-Crick base pairing and explained how genetic information could be copied due to the complementary nature of DNA.

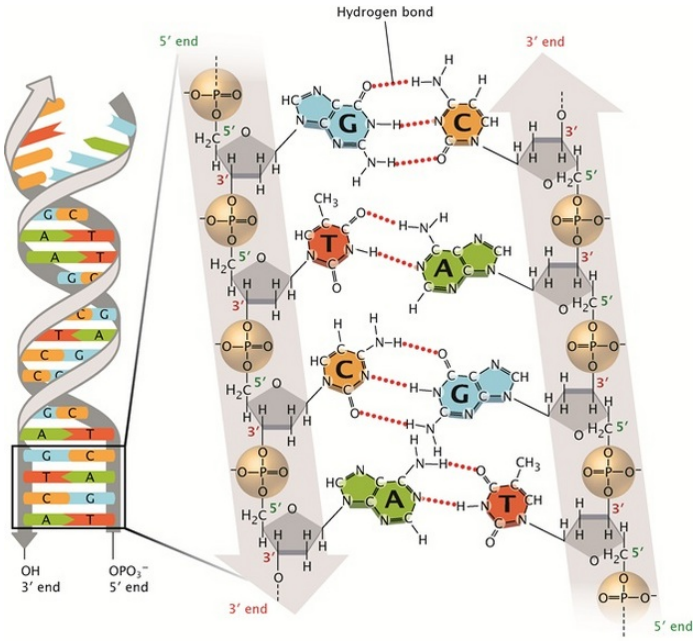

Figure 1.1: The structure of DNA is based on the repeated pattern of deoxyriboses and phosphate groups, forming the sugar-phosphate backbone, and the base pairing of the four bases, adenine (A), cytosine (C), guanine (G), and thymine (T). Two hydrogen bonds connect A to T and three hydrogen bonds connect C to G. Image used with permission from Nature Education 2013.

1.2 The Central Dogma of Molecular Biology

In 1958, Francis Crick wrote a seminal paper on protein synthesis, where he described the importance of proteins in living organisms and first proposed the central dogma of molecular biology[14]. Crick described how DNA or ribonucleic acid (RNA) could be used as templates for proteins and further described the possible directions of information flow between DNA, RNA, and protein. However, he noted that once information had been transferred from either DNA or RNA to protein, it was not possible for information to flow back to nucleic acids (Figure 1.2). In 1970, an enzyme known as reverse transcriptase (RT) was discovered[15, 16], which allowed RNA to be used as a template for producing DNA. In light of this and due to the misunderstanding of the central dogma, Crick restated the central dogma[17]: "The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid."

Prior to proposing the central dogma, Crick had predicted the existence of

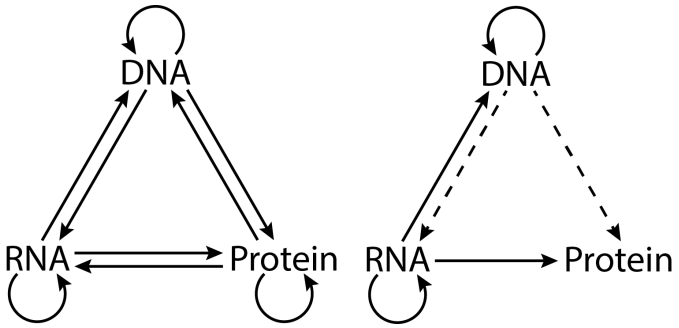

Figure 1.2: All possible information transfer pathways between DNA, RNA, and protein are shown on the left. Probable (solid arrows) and possible (dotted arrows) information transfer pathways, as originally proposed in 1958 by Francis Crick[14], are shown on the right. The Central Dogma of Molecular biology states that once information has been transferred to a protein, it is not possible for this information to be transferred back to the nucleic acids.

"adaptors" that would transfer information from RNA to protein in 1955[18]. Crick proposed that there were twenty adaptors and special enzymes, one for each amino acid; the enzymes would join a particular amino acid to its own special adaptor. This theory was later confirmed by the discovery of transfer RNA (tRNA) in 1958[19]. The discovery of messenger RNA (mRNA) in 1961 [20] demonstrated the flow of information from DNA to RNA and from RNA to protein. In 1962, the DNA code used for encode the amino acids of proteins was deduced[21]. Matthaei and colleagues demonstrated that an artificially created RNA, composed entirely of uracils, would produce a protein composed entirely of phenylalanine. The full code, known as the genetic code, was cracked three years later in 1965[22] and defined how information was encoded in DNA to produce amino acids. Nirenberg and colleagues deduced that three nucleotides defined a codon, which are translated into one of the 20 standard amino acids.

1.3 Transcription

Transcription is the process by which a particular segment of DNA is processed into RNA by the enzyme RNA polymerase (RNA pol). There are three different types of RNA polymerases in eukaryotic cells: Pol I transcribes DNA that encode most of the ribosomal RNAs (rRNAs); Pol II transcribes DNA that encode mRNAs and other non-coding RNAs; and Pol III transcribes the genes for small regulatory RNA molecules, such as tRNAs. The first step in transcription is initiation, whereby RNA pol binds upstream of the DNA to be transcribed, at a region known as the promoter (Figure 1.4). A promoter can be classified by their distance from the transcription start site (TSS), which are the first nucleotides transcribed by RNA pol. The core promoter for a region to be transcribed, i.e. the transcript, by Pol II is usually found immediately upstream of the TSS and contains specific DNA sequences or elements that are necessary for transcription. The core promoter elements include the TATA box (usually located 25 to 35 bases upstream of the TSS), the TFIIB recognition element [also known as the B recognition element, (BRE)], the initiator element (Inr), the

downstream promoter element (DPE), and CpG islands (CGIs) (Figure 1.3). The proximal promoter lies $\sim$ 250 bp of the TSS and contains primary regulatory elements. Distal promoters do not have a fixed distance from the TSS but are usually further upstream and contain additional regulatory elements.

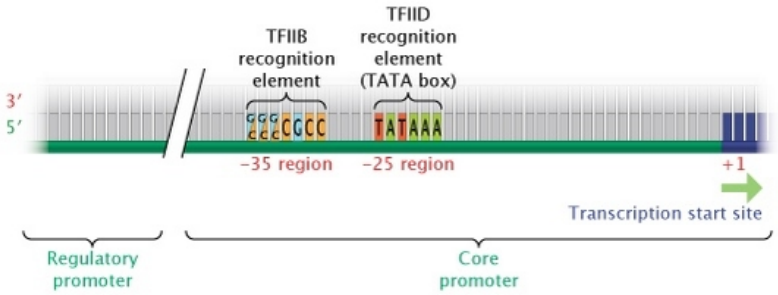

Figure 1.3: Core promoter elements recognised by RNA Pol II include the TFIIB recognition element and the TATA box, which are located around 35 and 25 bp upstream of the transcription start site (TSS), respectively. Various regulatory elements that regulate transcription may lie further upstream. Image used with permission from Nature Education 2014.

Once transcription has initiated, RNA pol and its associated proteins unwind the DNA double helix; once unwound, RNA pol reads the template DNA strand and adds nucleotides to the 3' end of a nascent RNA transcript. Transcription is terminated when the RNA polymerase reaches the termination site and the mRNA transcript and RNA pol are released (Figure 1.4). Transcription results in two main classes of RNA transcripts: (1) Protein-coding transcripts, where the RNA known as mRNA can be further translated into a protein molecule and (2) Non-coding transcripts, where the RNA molecule is the functional product.

1.3.1 Transcriptional regulation

The regulation of transcription ensures that transcripts are expressed in the correct spatial and temporal manner; this is necessary for maintaining cellular identity and for responding appropriately to environmental cues. Transcriptional regulation is achieved through interactions directly related to the DNA sequence and modifications not directly related to the DNA sequence. TFs are regulatory proteins that can activate or enhance the transcription of DNA by binding to specific DNA sequences and recruiting RNA polymerase[23]. TFs contain DNA-binding domains (DBDs) that allow them to bind specifically to DNA regions; these sites are known as transcription factor binding site (TFBS). One particular group of regulatory DNA that TFs bind to are known as enhancer sequences, which as the name suggests, enhances in the rate of transcription. Enhancer sequences can be located thousands of nucleotides away from the promoter they interact with, as they are brought into proximity to the promoter by the physical looping of DNA. In addition, enhancers may be positioned in both forward and reverse orientations, and located either upstream or downstream from its associated promoter and still affect transcription.

DNA is a physical entity inside cells and physical factors that act on DNA

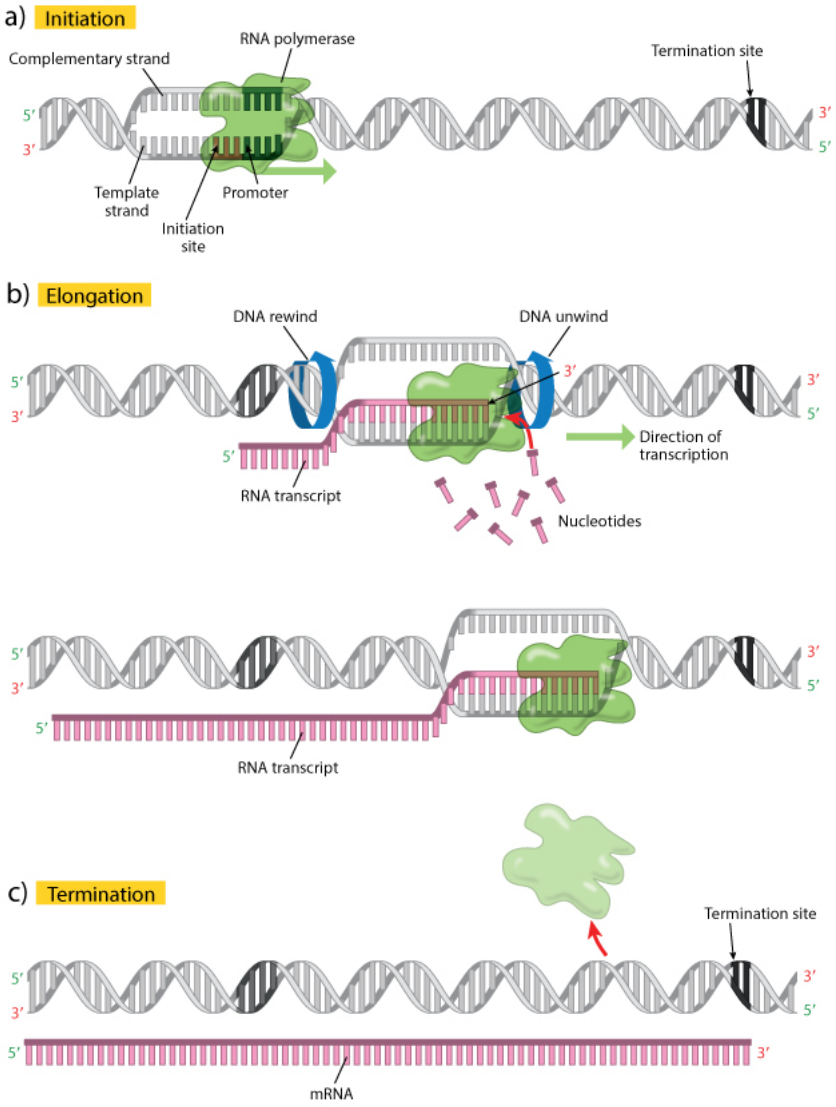

Figure 1.4: The process of transcription can be broadly grouped into three stages: a) initiation, b) elongation, and c) termination. Initiation involves the binding of RNA polymerase (shown as a large green blob) to the promoter and the DNA double helix starts to separate. RNA polymerase starts reading the sequence on the template strand in the 5' to 3' direction (green arrow). The elongation step involves the movement of RNA polymerase along the DNA strand producing a growing RNA transcript chain, which continually closes and opens the DNA strand. The nucleotides are shown as pink T-shaped molecules and the red arrow indicates that they are added at the 3' end of the nascent transcript. The termination step occurs once the RNA polymerase reaches the termination site, and the RNA transcript and RNA polymerase are separated from the DNA. Image used with permission from Nature Education 2013.

also regulate transcription; this is known as epigenetics. Biochemical modifications on the DNA, such as DNA methylation, which is the covalent addition of a methyl group to the 5 position of cytosine. DNA methylation is an important regulator of gene transcription and high levels of methylation in the promoter region of genes results in gene silencing. The structural compaction of DNA into chromosomes also regulates transcription, by limiting the accessibility of DNA to TFs and RNA pol. This compaction is achieved mainly via histones, which are a family of small and positively charged proteins that fold negatively charged DNA in the form of electrostatic interactions; this folding helps condense DNA and the resulting DNA-histone complex is called chromatin. Chromatin possesses a fundamental repeating structure[24], known as the nucleosome, which is the structural and functional unit of chromatin. Nucleosomes are structured with two of each of the following histones: H2A, H2B, H3, and H4, and forms a histone octamer that binds and wraps about 146 base pairs of DNA. The H1 histone protein binds to DNA that links nucleosomes, called linker DNA, wrapping another 20 bps of DNA and stabilising the linker DNA. Chromatin is found in two varieties: heterochromatin, which features DNA tightly wrapped into a 30 nm fibre, and euchromatin where DNA is lightly packed as nucleosomes (Figure 1.5).

1.3.2 DNA accessibility

Chromatin structure and nucleosome positioning are altered in order for the transcriptional and replication machinery to be able to access parts of the genome for transcription. Chromatin structure can be relaxed by biochemically modifying histones, to strengthen or weaken its association with DNA. Generally speaking, there are two major mechanisms by which chromatin is made more accessible via histone modifications:

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.

Importantly, these two processes are reversible, so modified or remodelled chromatin can be returned to its compact state after transcription and/or replication are complete. The nomenclature for histone modifications is defined by the name of the histone, followed by the single-letter amino acid abbreviation and its position, and then an abbreviation of the enzymatic modification; for example, H3K27ac indicates the acetylation of lysine 27 on H3. Specific histone modifications are associated with different biological states; for example, acetylation removes the positive charge on histones, thereby decreasing the interaction between histones and DNA, loosening chromatin, and allowing transcriptional activation to take place. On the other hand, the tri-methylation of lysine 27 on histone H3, i.e. H3K27me3, is associated with the inhibition of transcription[25]. Given that distinct histone modifications can either be implicated in the activation or repression of transcription, a "histone code" has been proposed[26] and the profiling of the histone states provides insights into the transcriptional state of a DNA region. Table 1.1 summarises a list of histone modifications and variants profiled by the ENCODE project[27].

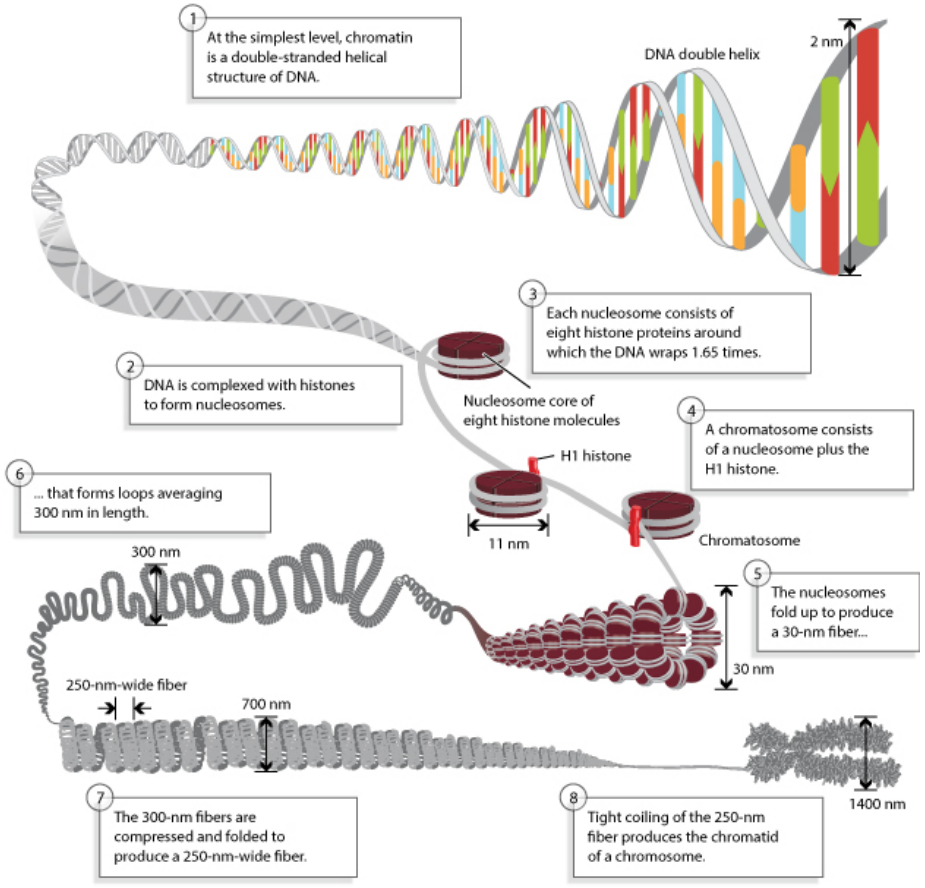

Figure 1.5: DNA is compacted at various levels: 1) as a double-stranded helical structure, 2 and 3) as nucleosomes, which consists of eight histone molecules with DNA wrapped around 1.65 times, 4) as a chromatosome, which consists of the nucleosome and a H1 histone, 5) as a 30 nm fiber of folded nucleosomes, 6) as looped nucleosomes that are on average 300 nm in length, 7) as compressed 300 nm fibers, producing a 250 nm fiber, and finally 8) as chromatids of a chromosome. Image used with permission from Nature Education 2013.

1.4 DNA sequencing

DNA sequencing is the process of determining the exact order of nucleotides within a DNA molecule. The first generation of DNA sequencing methods (Sanger and Maxam-Gilbert sequencing) were developed in the 1970s and were very labour intensive, requiring four separate polyacrylamide gel electrophoresis (PAGE) runs, for the determining the sequence of each base. The key feature of Sanger sequencing [28] was the use of chain-terminating dideoxynucleotide triphosphates (ddNTPs). The structure of a normal nucleotide (dNTP), consists of a 3' hydroxyl (OH) group in the pentose sugar; 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. The

Histone modification or variant	Signal characteristics	Putative functions
H2A.Z	Peak	Histone protein variant (H2A.Z) associated with regulatory elements with dynamic chromatin
H3K4me1	Peak/region	Mark of regulatory elements associated with enhancers and other distal elements, but also enriched downstream of transcription starts
H3K4me2	Peak	Mark of regulatory elements associated with promoters and enhancers
H3K4me3	Peak	Mark of regulatory elements primarily associated with promoters/transcription starts
H3K9ac	Peak	Mark of active regulatory elements with preference for promoters
H3K9me1	Region	Preference for the 5 end of genes
H3K9me3	Peak/region	Repressive mark associated with constitutive heterochromatin and repetitive elements
H3K27ac	Peak	Mark of active regulatory elements; may distinguish active enhancers and promoters from their inactive counterparts
H3K27me3	Region	Repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes
H3K36me3	Region	Elongation mark associated with transcribed portions of genes, with preference for 3 regions after intron 1
H3K79me2	Region	Transcription-associated mark, with preference for 5 end of genes
H4K20me1	Region	Preference for 5 end of genes

Table 1.1: Summary of histone modifications and variants profiled by the ENCODE project[27].

idea was to set up a reaction with a mixture of dNTPs [deoxyadenosine triphosphate (dATP), deoxyguanosine triphosphate (dGTP), deoxycytidine triphosphate (dCTP), deoxythymidine triphosphate (dTTP)] and a particular ddNTP in a ratio of 300:1. Most of the times, the DNA will be elongated but if a ddNTP is incorporated into the growing DNA strand, strand elongation is terminated. This results in DNA fragments of varying lengths, where the last base of these fragments corresponding to the ddNTP used. By performing the same reaction for the other three ddNTPs and loading the fragments of each reaction onto separate PAGE lanes, the DNA bases can be deduced by reading the four lanes (Figure 1.6).

The Maxam-Gilbert sequencing method[29] relies on the use of chemicals that can cleave specific bases in contrast to chain-terminating ddNTPs. Dimethyl sulfate was used to cleave purine bases (A and G) and hydrazine was 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 PAGE, and the DNA bases are deduced by reading the gel.

Sanger sequencing became the *de facto* method for DNA sequencing due to

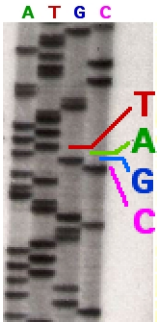

Figure 1.6: PAGE, which has a 1 bp resolution, is used to separate the radioactively labelled DNA fragments from each reaction using specific chain-terminating ddNTPs. By reading the DNA fragments, the sequence of the DNA can be deduced. Image used under the terms and agreement of the Wikipedia GFDL.

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[30]. 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 (Figure 1.7). The development of a fluorescence detection apparatus linked to a computer that processed the data created the world's first partially automated DNA sequencer[30]; this development was key towards the success of the Human Genome Project (HGP). For over 25 years since its inception, Sanger sequencing was the method of choice for DNA sequencing.

1.4.1 Next-generation sequencing

The next wave of DNA sequencing techniques, the so-called next-generation (next-gen) or second generation sequencing, started with various strategies that relied on a combination of template preparation, sequencing, and imaging that allowed thousands to billions of sequencing reactions to be performed simultaneously[31]. Next-gen sequencing relies on the clonal amplification of templates and uses *in vitro* cloning rather than bacterial cloning; the two most common methods of clonal amplification are emulsion polymerase chain reaction (emPCR)[32] and solid-phase amplification[33]. With emPCR individual DNA molecules are isolated with primer-coated beads in water-in-oil microreactors and clonal amplification leads to thousands of copies of the DNA molecule in an emulsion. 454 pyrosequencing and Sequencing by Oligonucleotide Ligation and Detection (SOLiD) sequencing employ emPCR and the amplification products are deposited into individual wells for sequencing. Solid-phase amplification relies on a lawn of high-density primers that are covalently attached on a slide surface (also known as a flow cell) and bind to DNA molecules that have been ligated with sequencing adaptors. The two methods allow each DNA template to be spatially separated and allow massively parallel sequencing to take place.

Sequencing can take place via the use of DNA polymerase, which is com-

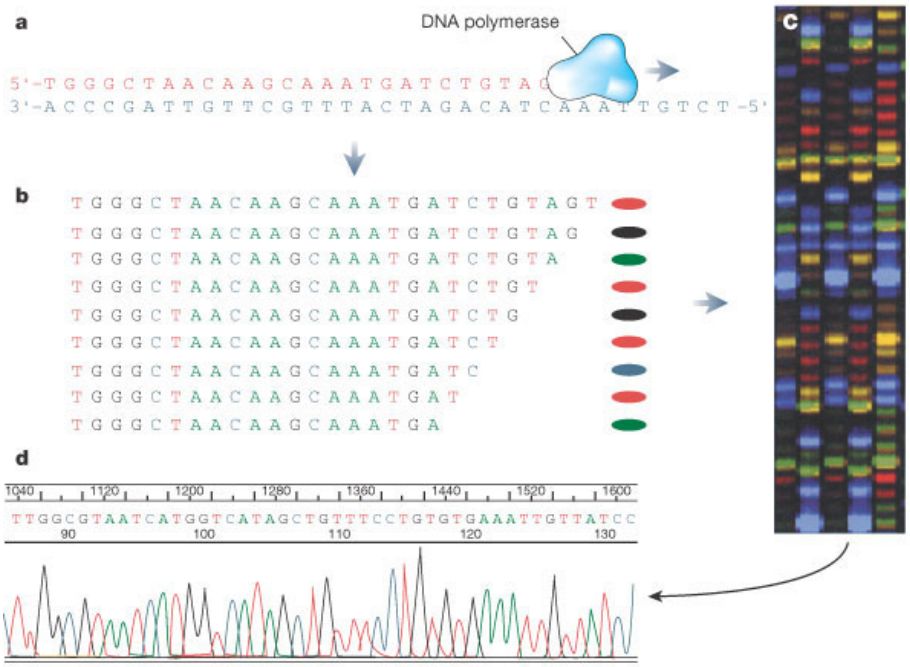

Figure 1.7: a) DNA polymerase synthesises a complementary strand of DNA, however when a b) fluorescently labelled chain-terminating ddNTP base is incorporated, synthesis terminates producing DNA fragments of various sizes. c) As each terminator are fluorescently labelled with different dyes, each fragment will fluoresce a particular colour and the d) sequence trace is read by a computer that determines the sequence based on the coloured peaks.

monly known as sequencing-by-synthesis (SBS), or via the use of DNA ligase, which is known as sequencing-by-ligation (SBL). SBS can be further classified into cyclic reversible termination (CRT), single-nucleotide addition (SNA), and real-time sequencing[31]. CRT uses reversible terminators and initial developments used the dideoxynucleotides chain terminators used in Sanger sequencing. The concept of CRT is that a DNA polymerase incorporates one fluorescently modified nucleotide, which has a reversible terminator that terminates DNA synthesis. Unincorporated nucleotides are washed away and fluorescence imaging takes place to determine the identity of the incorporated nucleotide. The last step removes or cleaves the reversible terminator and the fluorescent dye, and the cycle is repeated. The CRT method is used in Solexa/Illumina and Helicos single-molecule fluorescent sequencing. SBL relies on DNA ligase and uses either one-base-encoded or two-base-encoded probes that are fluorescently labelled. The probes hybridise to its complementary sequence on the primed template and DNA ligase is added to join the probe to the primer. Non-ligated probes are washed away followed by fluorescence imaging and cleavage of the fluorescent dye and the cycle is repeated. The SBL method is used in SOLiD sequencing.

1.4.2 Third generation sequencing and beyond

The third generation of sequencing refers to single-molecule sequencing technologies, which has the capacity for generating longer read lengths at potentially cheaper costs[34]. One of the major advantages of single-molecule sequencing is that polymerase chain reaction (PCR) is not required, and therefore amplification biases and PCR mutations are eliminated. Furthermore, by employing third generation sequencing, quantitative applications of sequencing, such as RNA sequencing, can give a much more representative picture of the true abundance of RNA molecules. The HeliScope sequencer was the first commercially available single-molecule sequencer, which was based on the work of Stephen Quake and colleagues[35]. HeliScope sequencing utilises billions of primed single-molecule templates that are covalently attached to a solid support and uses CRT but with slight differences from Solexa/Illumina sequencing. HeliScope sequencing uses Helicos Virtual Terminators, which differ from the reversible terminators used in Solexa/Illumina sequencing and dye labelled nucleotides are added individually in the predetermined order of C, T, A, and G, followed by fluorescence imaging.

With the advent of high-throughput sequencing we now have the capacity to sequence an entire human genome in a matter of days. In addition, we have just recently arrived in the $1,000 genome era, whereby we can sequence the entire genome of an individual at a 30x depth (the minimum depth required for clinical applications) for around 1,000 US dollars (USD). In contrast, the Human Genome Project (HGP), which gave us the first glimpse of the human genome[36] costed approximately 2.7 billion fiscal year 1991 US dollars[37]. Further developments in sequencing by various companies are aiming towards longer read lengths at a higher output (Figure 1.8). Currently, different sequencers either have very long reads but at a low-throughput or have a high-throughput of shorter reads; as such, each sequencer fills a particular niche. *De novo* assembly of genomes requires longer reads for less ambiguity and the quantification of RNA requires higher throughput in order to accurately sample the vast RNA population.

1.5 Expression analysis

Transcription of a region of DNA results in the expression of an RNA transcript. A transcript may be constitutively expressed, i.e constantly expressed, or expressed according to the current cellular requirements. By comparing transcript levels between different conditions, insight can be gained on the possible function of a particular transcript. Of note is that the amount of a specific transcript in a cell at a given time is not only influenced by the rate of transcription but also by the stability of the transcript; a rapidly degraded transcript may appear to be lowly transcribed. Northern blotting[39] was one of the first methods for quantifying the expression level of specific RNA transcripts. This technique involves the electrophoretic separation of purified RNA, followed by immobilising the RNA onto a blotting membrane; detection of the transcript is achieved by hybridising a specific probe that is complementary to part of the transcript. The relative amount of a specific transcript can be estimated by comparing the strength of signals from different samples; this estimate assumes that an equiv-

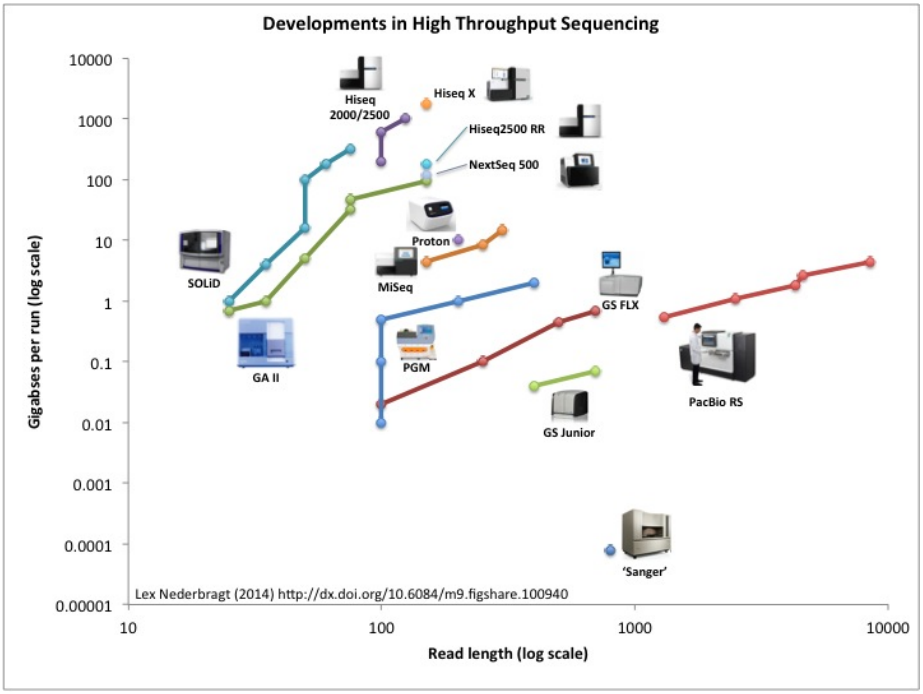

Figure 1.8: Read length (log_{10}) versus gigabases per run (log_{10}) for various high-throughput sequencer[38]. Currently, the HiSeq X, the sequencer that bought us into the $1,000 genome era, provides the highest throughput. Image used under the terms and agreement of the CC-BY license.

alent amount of total RNA was used per sample and this is verified by using a probe that detects a constitutively expressed transcript. Northern blotting is also useful for determining the size of a specific transcript and is commonly used for detecting alternatively spliced transcripts; however, Northern blotting is not very sensitive for estimating transcript abundance.

The most sensitive method of detecting specific transcript levels is quantitative real-time polymerase chain reaction (qRT-PCR), which is able to detect sparse levels of transcripts, down to the level present in single cell. The method employs the use of a fluorescent dye and PCR, where the primer pairs are designed to be specific for a given transcript. By measuring the number of cycles that is required for the formation of a detectable amount of product, which is known as the cycle threshold (C_t) value, the transcript levels can be estimated by comparing the results to qRT-PCR experiments performed using standard samples[40]; lower C_t values indicate higher amounts of initial template. It should be ensured that the PCR step is equally efficient for different transcripts by proper primer design, as a more efficient PCR will result in a lower C_t value compared to a less efficient PCR.

1.5.1 Transcriptome profiling

Transcriptome profiling refers to the expression profiling of the complete collection of transcripts in a cell at a specific time point. Transcripts associated with a particular biological process or disease state can be identified by profiling all transcripts and in addition, the interplay between transcripts can be inferred by studying transcripts with similar expression patterns. One of the first technologies that allowed the simultaneous profiling of thousands of transcripts at once were microarrays[41]. DNA probes that are complementary to specific DNA sequences, such as complementary DNAs (cDNA) or genomic regions, are attached to a solid surface and fluorescently labelled target sequences are hybridised onto the surface. Target sequences that complement the probe sequences hybridise to the probe and the signal intensity provides a measure of the expression strength of a particular transcript. In one of the first application of microarrays, researchers were able to observe the change in expression levels of 700 mRNAs during a switch from aerobic to anaerobic respiration in yeast cells[42]. However, microarrays have several limitations, which includes requiring *a priori* knowledge of the genome or transcript sequences, high background levels from cross-hybridisation[43], and a limited dynamic range in quantifying expression.

In contrast to the hybridisation approach of microarrays, sequencing-based approaches have been developed for transcriptome profiling. Prior to the advent of next-gen sequencing, sequencing approaches were based on the sequencing of short tags; these tagging approaches were cost-effective, as only short fragments of cDNAs were sequenced. Typically type IIS restriction enzymes were used to create tags, which were then concatenated, cloned, and sequenced. A technology called Serial Analysis of Gene Expression (SAGE)[44] was the first tag-based approach, which created 9 to 10 bp long tags that generally corresponded to the 3' end of the transcripts. A similar technology known as Massive Parallel Signature Sequencing (MPSS), was later developed, which is similar to SAGE but employs different biochemical steps and a different sequencing approach[45]. MPSS was an improvement to SAGE in that it produced longer tags (16-20 bp) and libraries that were 20 times larger than typical SAGE libraries[45]. Another tagging method known as the paired-end ditag (PET) approach, prepared ditags that corresponded to the 5 and 3 end of the same full-length cDNA[46]. The PET approach allows the mapping of cDNA boundaries, helps resolve ambiguous tag mappings by using the paired-end information, and has the potential to detect unconventional fusion transcripts and rearrangement events. The SAGE approach also gave rise to Cap Analysis Gene Expression (CAGE)[47], which combines the tagging strategies of SAGE with a molecular technique known as Cap-Trapper[48, 49]. The CAGE protocol captures all capped transcripts and sequences a short tag (20 or 27 nt depending on which restriction enzyme is used) that corresponds to the 5' end of an RNA transcript (Figure 1.9).

With the arrival of next-gen sequencing, the SAGE and CAGE methods were adapted to high-throughput sequencers[50, 51]. The short-read and high-throughput nature of next-gen sequencing suited tag-based approaches aptly, as the tag lengths were in the size range of reads produced by the sequencer. Whole transcriptome shotgun sequencing or simply RNA sequencing (RNA-Seq) methods were later developed to sequence entire populations of RNA. RNA-Seq refers to the fragmentation of RNA followed by deep-sequencing on next-gen

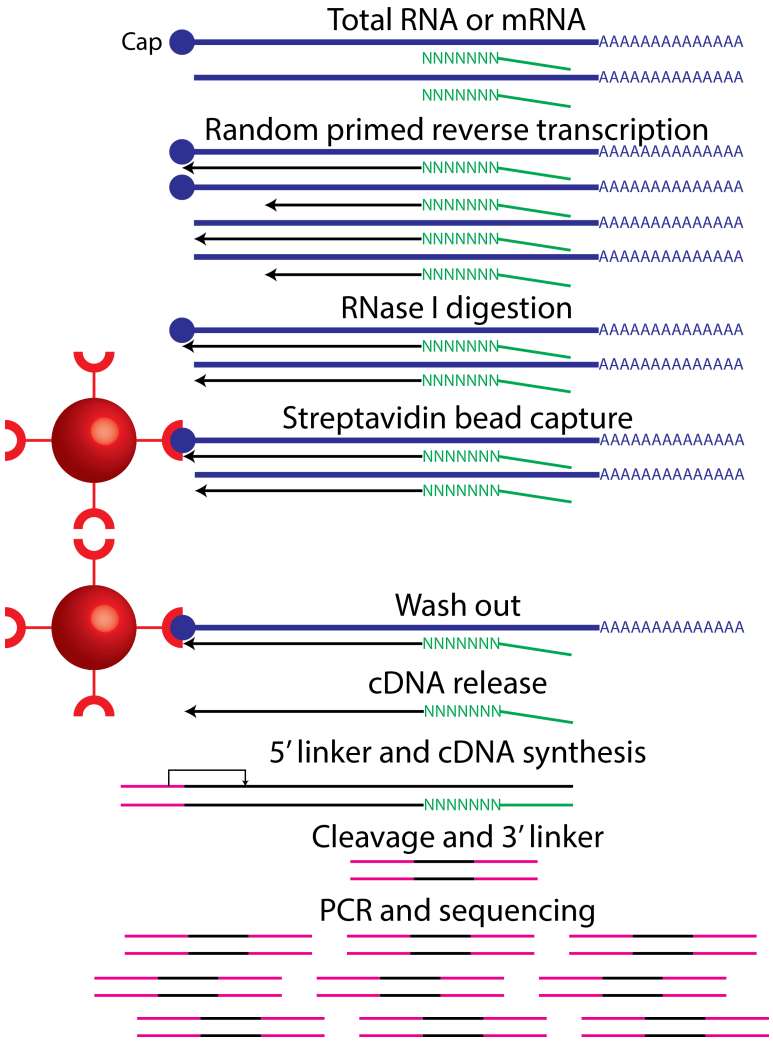

Figure 1.9: The Cap Analysis Gene Expression (CAGE) protocol starts with synthesising cDNA from either total RNA or mRNA by using random or oligo dT primers (only random primers are shown here). Reverse transcription takes place in RNAs with or without a cap and to full or partial completion; the RNase I digestion removes partially reverse transcribed RNA as they are not protected by a full double strand. The 5' end of cDNAs are selected by streptavidin beads and unbound cDNA are washed out. After release from the bead, a linker is attached to the 5' end of the single-stranded cDNA; this linker contains recognition sites that allow the endonuclease cleavage. Lastly a linker is attached to the 3' end of the tag sequence, which is amplified and directly sequenced.

sequencing platforms[52]; the fragmentation step is required due to the short-read nature of next-gen sequencers. Typically, RNA-Seq methods enrich for transcripts with a poly-A tail; this enrichment step is carried out to avoid rRNA sequences, which make up a large fraction of the total RNA population. One

major difference between RNA-Seq and the tag-based approaches is that the entire length of the RNA is sequenced in RNA-Seq; the advantage in this is that alternative splicing patterns can be inferred. However tag-based approaches, such as CAGE, can be used in a complementary nature to RNA-Seq[53], since transcript boundaries are defined clearly in CAGE.

1.5.2 Transcriptional complexity

Several landmark studies have revealed that the transcriptional landscape is much more complex than previously anticipated. The functional annotation of the mammalian genome (FANTOM) project, which began as an initiative to sequence and functionally annotate mouse full-length complementary DNA (cDNA)[54], revealed that the transcriptome was dominated by transcripts that had no apparent coding potential[55]. The FANTOM consortium also revealed massive antisense transcription[56], which is transcription arising from the strand opposite the sense strand, and extensive alternative promoter usage[57]. Tiling arrays, which are microarrays designed to interrogate a genome at evenly spaced intervals, revealed that a large fraction of genomic bases were transcribed[58, 59, 60]. This observation that a large percentage of mammalian genomes are transcribed became known as pervasive transcription[61], however these claims made on the basis of tiling arrays were questioned[62]. However, the Encyclopaedia of DNA elements (ENCODE) project, which endeavours to identify all functional elements in the human genome, observed that mammalian genomes were pervasively transcribed[63, 27]. These claims were based on the use of various genome-wide biochemical assays and multiple lines of evidence. However, it is not known whether these products of transcription are functional or not, as they do not overlap known genes.

1.5.3 Defining a gene

The idea of a gene dates back to Gregor Mendel and his plant breeding experiments that demonstrated that discrete traits could be inherited from parents to offspring. The term "gene" was coined in 1909 by Wilhelm Johannsen to describe the Mendelian units of heredity. Genes were later described as the precursors to proteins, in the "one gene, one polypeptide" hypothesis, when it was observed that mutations in Neurospora genes would cause defects in different steps of metabolic pathways [64]. After the determination of the genetic code (see section 1.2), a gene was recognised as a stretch of DNA that coded for a protein in an open reading frame (ORF). The discovery of introns[65, 66], altered the ORF concept, in that genes were now composed of both protein-coding regions (exons) and the non-coding regions (introns). The trend had been that each time a major discovery had been made, the definition of a gene was revised. Thus in light of pervasive transcription members of the ENCODE consortium suggested that the definition of a gene needs to be updated yet again[67]. The definition they proposed was: "A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products"[67]. This definition is similar to the idea of a transcriptional unit (TU), proposed by the FANTOM consortium, which is a segment of the genome that shares common core of genetic information and is able to generate transcripts[55].

1.5.4 Non-coding RNAs

The protein-centric view of genes segregated RNA as either being coding RNA, the mRNAs, or RNA with no coding potential, known as non-coding RNAs (ncRNAs). Historically, it was believed that there were only a few ncRNA families, such as the tRNAs, rRNAs, and the spliceosomal RNAs, all of which existed to aid protein translation. NcRNAs exist as single-strand nucleic acid molecules, which tend to fold on themselves due to the hydrophobic nature of bases, to form localised double-stranded regions that form structures called hairpins or stem-loop structures. NcRNAs can be broadly classified as short or small RNAs, which are defined as being <200 nucleotides long, or as long non-coding RNAs (lncRNAs), which are >200 nucleotides in length; these size cut-offs correspond to the commonly used size selections in biochemical fractionations and does not have any biological implications.

One of the most well studied classes are the micro-RNAs (miRNAs), which were first observed in *Caenorhabditis elegans*[68]. MiRNAs are typically 20-24 nucleotides in length and function as regulators of expression by base-pairing with complementary sequences within mRNAs (commonly to the 3' untranslated region (UTR)). Another class of well known ncRNAs are the Piwi-interacting RNA (piRNA), which were first observed in drosophila[69]. PiRNA are typically between 24 and 32 nucleotides long and are thought to be involved in gene silencing, especially in silencing transposable elements (TEs), by forming the piRNA-induced silencing complex (piRISC). The characterisation of full-length cDNAs revealed another class of RNAs known as long non-coding RNAs (lncRNAs)[55]. Unlike miRNAs and piRNAs, they are not as well characterised and are usually described with respect to protein-coding genes as sense, antisense, bidirectional, intronic, and intergenic[70]. LncRNAs have been dismissed due to lack of sequence conservation[71] and have been suggested to be products of transcriptional noise[72]. While there have been several well documented cases of lncRNAs, it has been suggested that more experimental work needs to be performed on lncRNAs to find out how many are indeed functional[73].

The list of ncRNAs (Table 1.2) has increased over the last 20 years due to technological advances, such as RNA-Seq. These include, but are not limited to, promoter upstream transcripts[74], long intergenic non-coding RNAs[75, 76, 77], transcription start site-associated RNA[78], enhancer RNAs[79], promoter-associated short RNAs and termini-associated short RNAs[61], double strand break-induced small RNAs[80] or DNA damage RNAs[2], competing endogenous RNAs[81], and transcription initiation RNAs[82]. Despite an abundance of ncRNA classes, many of these ncRNAs are supported only by transcriptional data, since validation studies have only been performed in a low-throughput manner. As the genome is pervasively transcribed, transcription by itself is not enough evidence to support function. Newer technologies, such as parallel analysis of RNA structure[83] and fragmentation sequencing[84], that focus on the structural properties of RNA may provide more clues to the extent of functionality of ncRNAs.

Non-coding RNA	Typical size (bp)	Putative functions
Transcription initiation RNAs	17-18	Transcriptional regulation
Micro-RNAs	20-24	Silencing of targeted messenger RNA
DNA damage RNA	22-23	Establishing the DNA damage response
Promoter-associated short RNAs	22-200	Unknown
Piwi-interacting RNA	24-32	Silencing of transposable element
Enhancer RNA	50-2000	Transcriptional regulation
Transfer RNA	73-94	Translation of messenger RNA
Long intergenic non-coding RNA	>200	Chromatin modification
Ribosomal RNA	1,900 and 5,000	Protein synthesis
Competing endogenous RNA	Undefined	Micro-RNA sponge

Table 1.2: A list of various non-coding RNAs, their size profiles, and putative functions.

1.6 Repetitive mammalian genomes

The discovery that cells contained a large fraction of repetitive DNA was made by measuring the re-association rates of DNA strands after denaturation[85]. The characterisation of repetitive elements (REs) was possible with the release of the mouse[86] and human[87, 36] genome sequences, which showed that these genomes are indeed largely made up of REs. The two major groups of REs are tandem repeats, which include different classes of satellite repeats, and interspersed repeats, which are mostly made up of transposable elements (TEs)[88]. Within TEs are two main classes: Class I TEs or retrotransposons, which are DNA elements that are transcribed into RNA, reverse transcribed back to DNA, and transposed to a new location in the genome (a copy-and-paste mechanism), and Class II TEs or DNA transposons, which simply excise their DNA sequence from one location to another via transposase enzymes (a cut-and-paste mechanism). As retrotransposons are able to produce a copy of themselves before propagation, they are more numerous than DNA transposons. The retrotransposon known as the Alu element, has an estimated copy number of more than one million, making them the most abundant RE in the human genome[36].

There are various methods for identifying REs in genomes, which can be broadly categorised into *de novo*, homology, structure, and comparative genomic based methods[89]. For the initial mouse and human genome sequencing projects, REs were catalogued using a popular software called RepeatMasker, which identifies REs by using homology-based methods to search against a database of consensus repeats[90], such as the Repbase Update database[91]. One of the main drawbacks of identifying REs in this manner is the reliance on sequence homology and a single consensus sequence, resulting in missing REs with an extensive number of mutations. Recently, RepeatMasker has incorporated the use of profile Hidden Markov Models to annotate REs[92]; profile methods use an alignment of multiple representative sequences rather than a single consensus and are more sensitive than single sequence searches. However, REs not contained within databases may still be missed and *de novo* methods may be the key to identifying these repeats. It has been proposed that up to two-thirds of the human genome may be made up of repetitive elements based on a *de novo* identification method[93].

1.6.1 Junk DNA

Historically, TEs have been labelled as purely selfish elements that have no function or provide no selective advantage to an organism[94, 95] and were considered as junk DNA. The term "junk DNA", was popularised by Susumu Ohno[96], 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 does not play a functional role in an organism. The question of how much of the human genome is functional was also addressed by Susumu Ohno, who used a fixed mutation rate (each locus has a 10^{-5} probability of sustaining a deleterious mutation) to estimate the number of functional loci[96]. Given this mutation rate, he predicted that the human genome could not have more than 30,000 loci under selection, as this would guarantee a progressive decline in genetic fitness, leading to mutational meltdown[97]. In stark contrast to the estimation made by Susumu Ohno, the ENCODE project reported that 80% of the human genome has a biochemical function[27], to which junk DNA was immediately dismissed[98]. This lead to several critiques[99, 100, 101], which all raised the issue of whether biochemical activity is enough to assert function.

One peculiar observation among eukaryotic genome sizes, known as the C-value enigma, is the lack of correlation between genome sizes and organismal complexity[102]. The genome size variation among eukaryotes can be partially explained by the presence or absence of TEs[103], which raises the question of whether or not these sequences are necessary or simply junk. To answer this question, the *Fugu rubripes* genome was sequenced to provide a useful reference for annotating functional elements in the human genome[104]. The fugu has one of the smallest vertebrate genomes; at 390 Mb less than 10% of the genome is made up of REs, compared to 50% in the human genome (Figure 1.10). An even more extreme example, is the genome of the carnivorous bladderwort plant, *Utricularia gibba*, which is 82 Mb in size and is almost devoid of REs[105]. At least in these two organisms, their genomes suggest that "junk DNA" is not essential.

1.6.2 Impact of transposable elements on genomes

Genome evolution has been largely affected by TEs, which have the ability to move around within genomes. The activity of TEs have led to insertion mutations and genomic instability, but have also contributed to genetic innovation[107]. Despite the dismissal that TEs are purely selfish and are non-essential elements, the possibility that some TEs may become useful was speculated:

> "It would be surprising if the host genome did not occasionally find some use for particular selfish DNA sequences, especially if there were many different sequences widely distributed over the chromosomes. One obvious use ... would be for control purposes at one level or another. This seems more than plausible."
>
> — Orgel and Crick 1980

While the estimated number of active TEs in the human genome is less than 0.05%[108], it is clear that TEs have impacted the evolution of genomes. There

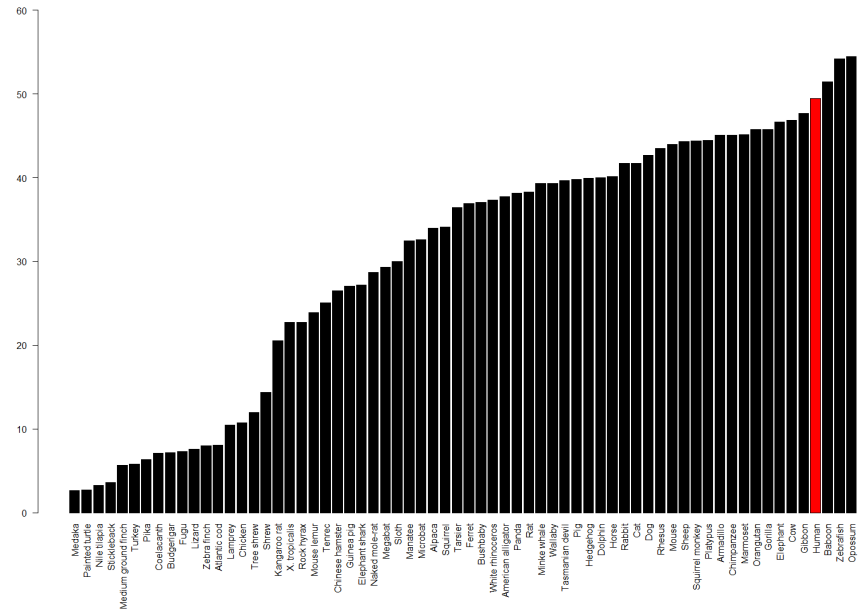

Figure 1.10: The total coverage, in percentage, of repetitive elements in 66 vertebrate genomes as annotated by RepeatMasker; the human fraction is shown as a red bar[106]. Image used under the terms and agreement of the CC-BY license.

is increasing evidence that TEs have served as a source for evolutionary innovation by contributing to the evolution of regulatory networks[109], influenced alternative splicing[110], drove the evolution of lncRNAs[111], and provided alternative promoters for genes[112]. These examples illustrate the process of exaptation, where a character (the TE) becomes co-opted for a new function that it was not originally functional for[113]. As speculated by Orgel and Crick, there may be many other examples of TE exaptation, however, large-scale analysis of TE expression have been limited. This was due to cross-hybridisation problems with repetitive sequences in microarrays and the short read nature of current high-throughput sequencers.

In one of the first genome-wide profiling studies of TE expression, it was revealed that a large number of transcriptional events initiate within TEs in a tissue specific manner[114]. Many of these TSSs were validated experimentally to show that TEs served as alternative promoters. In order to deal with the multi-mapping nature of reads corresponding to TEs, a probabilistic method called multi-map rescue was implemented[115], which allowed reads to be assigned probabilistically based on the nearby region. Another genome-wide study examining the methylation patterns of TEs revealed tissue-specific methylation patterns[116]. In this study, the authors developed the Repeat Analysis Pipeline (RAP), to deal with ambiguously mapped reads, in a manner that aggregates reads to families of repeats[117]. Both of these studies illustrate the importance of developing specific methodologies for analysing TEs and that TEs have been

important in regulating the expression of genes.

1.7 Bioinformatics and genomics

Modern day high-throughput sequencers generate a large amount of data and dedicated informatics tools for storing, managing, and the analysis of such data sets are absolutely necessary. Bioinformatics can be thought of as a subset of informatics that deals with biological data, though historically it was defined as "the study of informatic processes in biotic systems"[118]. The HGP was one of the first large scale international research efforts, which demonstrated how bioinformatics was crucial towards the successful completion of the project[119]. Furthermore, the HGP also set the stage for data sharing by establishing important principles, known collectively as the "Bermuda Principles", that promoted 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 are made freely available prior to publication[120]. By opening such resources, researchers are able to integrate and leverage these datasets for generating testable hypothesises.

While many of the foundations in bioinformatics were related to molecular evolution and population genetics, the availability of complete genome sequences has opened up an entire research discipline called genomics. The field of genomics is also closely tied with bioinformatics, as genomics studies typically deals with large amounts of data, corresponding to features of genomes. There are several sub-branches within genomics; the field of functional genomics focuses on the dynamic aspects of genomes, such as transcription, interactions between proteins and genomic regions, and DNA methylation patterns. Functional genomics, as the name suggests, attempts to discover and establish function to elements in the genome and usually employing the use of high-throughput methods for genome-wide screening.

1.7.1 High-throughput sequencing data

The FASTQ format was formally defined in 2010[121] and has become the *de facto* format for storing raw high-throughput sequencing data. FASTQ is similar to the FASTA format but with the addition of quality scores, known as the Phred quality score, for each sequenced nucleotide and is usually the starting point of bioinformatic pipelines. Typically, quality control (QC) steps are carried out next to remove potential artefacts and low quality reads; one such tool known as TagDust[122], removes reads that match sequences used during the library preparation, such as primer and adaptor sequences. Other QC steps include removing reads containing undetermined bases, as they indicate poor overall read quality; for sequencers that output reads of different lengths, reads outside the main length distribution are removed; quantifying highly over-represented 10-mers can also be implemented as a QC step[123], to identify potential artefactual sequences.

In order to put sequencing reads into context, reads are mapped onto their corresponding reference genome; various tools are available for aligning high-throughput sequencing reads. Traditional tools such as BLAST[124] and BLAT[125]

are unable to cope with the large quantity and short length of reads from high-throughput sequencers. One popular short-read alignment tool, BWA[126], implements the BurrowsWheeler transform to deal with millions to billions of short reads. The Burrows-Wheeler transform allows a large mammalian genome, for example human, to be indexed and stored efficiently into memory[127]. The Sequence Alignment/Map (SAM) format[128] is the standard file format for storing sequence alignments and contains all the information for reconstructing an alignment. In addition the SAM format contains information on where a read maps on the genome, the quality of the mapping, and depending on the alignment program, other mapping statistics. The open source program SAMTools[129], provides various utilities for the processing and analysis of alignments stored in the SAM format. In order to save disk space, SAM files are typically stored as BAM files, which are simply their binary equivalent. Recent developments have introduced a newer format known as CRAM, based on a newer compression method[130], which further compresses BAM files but are still processable using SAMTools.

The BED format[131] is another standard file format used for storing the location of a set of features (or even sequencing reads) with respect to a reference genome and has been popularised by the UCSC Genome Browser[132]. Due to the popularity of the BED format, a suite of tools released as BEDTools[133], provides various routines for comparing genomic features stored in BED format. A common task in processing RNA-Seq data involves intersecting mapped reads to known genomic features, such as genes, to associate a read to a gene and thereby quantifying its expression. In addition to this, the proximity of elements on a chromosome may indicate potential functional interactions, thus reads may be annotated with respect to the physical distance, i.e. spatially, to genomic features. For example, promoters are usually upstream and nearby the genes that it initiates transcription for, thus CAGE reads are associated with nearby gene models in this manner.

1.7.2 Analysing expression datasets

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

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

Typically, two or more groups of assays are produced, such as a set of control experiments versus a set of treated experiments, and the aim is to find differentially expressed transcripts between the two conditions. Data normalisation is required prior to comparing assays to ensure that experimental factors (not including the experimental treatment) that cause differences are accounted for. For example, a library that has been sequenced twice as much as another library will have most of its transcript as seemingly expressed twice as much. One way

to account for this is to normalise reads by counts or tags per million (TPM); to normalise by TPM the i^{th} gene in the j^{th} assay:

$$TPM_{a_{ij}} = \frac{a_{ij} \times 1000000}{\sum_{i=1}^{m} a_{mj}}$$

Other methods include normalising the expression by the length of a transcript, known as reads per kilobase of transcript per million mapped reads or fragments per kilobase of transcript per million mapped reads, for RNA-Seq experiments where reads are generated throughout the length of the transcript, quantile normalization, which is a technique for making two distributions identical in statistical properties[134], and trimmed mean of M values (TMM), which is a normalisation method that takes into account differences in the RNA population from different samples by scaling the expression[135].

After the appropriate normalisation methods have been carried out, the testing of differential expression is performed by comparing the amount of variation between groups against the amount of variation within groups for a particular transcript, which is similar to carrying out a t-test or analysis of variance (ANOVA). However, the simple assumptions of these tests are not met due to the heteroscedastic and non-normal distribution properties of high-throughput sequencing data. RNA-Seq expression is measured by digital counts of reads, meaning that expression levels are discrete, and the variance is modelled using discrete probability distributions. In a pioneering study examining the reproducibility of RNA-Seq, it was noted that the variation between technical replicates was close to the shot noise limit[136]. Thus it was suggested that the Poisson model was sufficient in modelling the variance and used for testing differential expression. However, it was demonstrated that the Poisson model underestimated the effects of biological variability, i.e. the variation between two different biological samples is greater than Poisson variation[137]. This can be accounted for by modelling variance under a negative binomial model, which has been implemented for differential expression analysis in the edgeR package[138] from Bioconductor[139].

The correlation of transcriptional responses or expression profiles can be calculated to identify transcripts expressed in a similar manner or assays with a similar profile, respectively. Correlation measures such as Pearson's product-moment correlation coefficient or Spearman's rank correlation coefficient are typically used and can be used as a measure of co-expression, i.e. transcripts with a similar transcriptional response are assumed to be co-expressed. Other measures of similarity or rather dissimilarity include metrics such as the Euclidean, maximum, Manhattan, Canberra, Jaccard, and Minkowski distances. Hierarchical clustering is usually performed in an agglomerative manner to reveal the topology of distance matrices and visualised using dendrograms to revealing the most similar transcripts or assays. The visualisation of expression datasets include heatmaps, which transforms the expression matrix into colours representing the relative expression strength and are commonly arranged according to the hierarchical clustering structure[140]. Graphs can also be used to represent associations of transcripts or assays, which can leverage methodologies developed in graph theory.

Expression datasets are usually large and therefore subjected to the multiple testing problem, whereby significant p-values are observed due to the large

number of statistical inferences that are carried out. Several techniques are used to account for this, such as the Bonferroni correction or false discovery rate (FDR) control[141] and these techniques generally require a higher p-value significance threshold to compensate for the number of inferences made. A FDR of 10% means that it is expected that 10% of our statistical inferences are false positives and the FDR is the p-value at which to draw a threshold. It is important to account for multiple testing when performing numerous statistical tests, which is common when analysing high-throughput expression datasets.

Chapter 2

Template switching artifacts

In this work, we studied the transcriptional landscape in whole blood using a technology called nano Cap Analysis Gene Expression (nanoCAGE). While it was expected that the variability between samples would be high due to the heterogeneous nature of whole blood, unexpectedly, samples prepared with the same or similar barcodes had very similar expression profiles regardless of biological origin. This affected differential expression analyses whereby the estimation of variability was increased due to biological replicates having very different expression profiles. In order to adjust for the barcode bias, we identified template-switching artefacts that caused the bias, and developed a bioinformatic solution for removing these artefacts and improved the nanoCAGE protocol to mitigate the barcode bias. This work laid the groundwork for analysing transcriptome profiling using high-throughput sequencing, by understanding how technical artefacts and biases are able to form and affect downstream analyses.

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

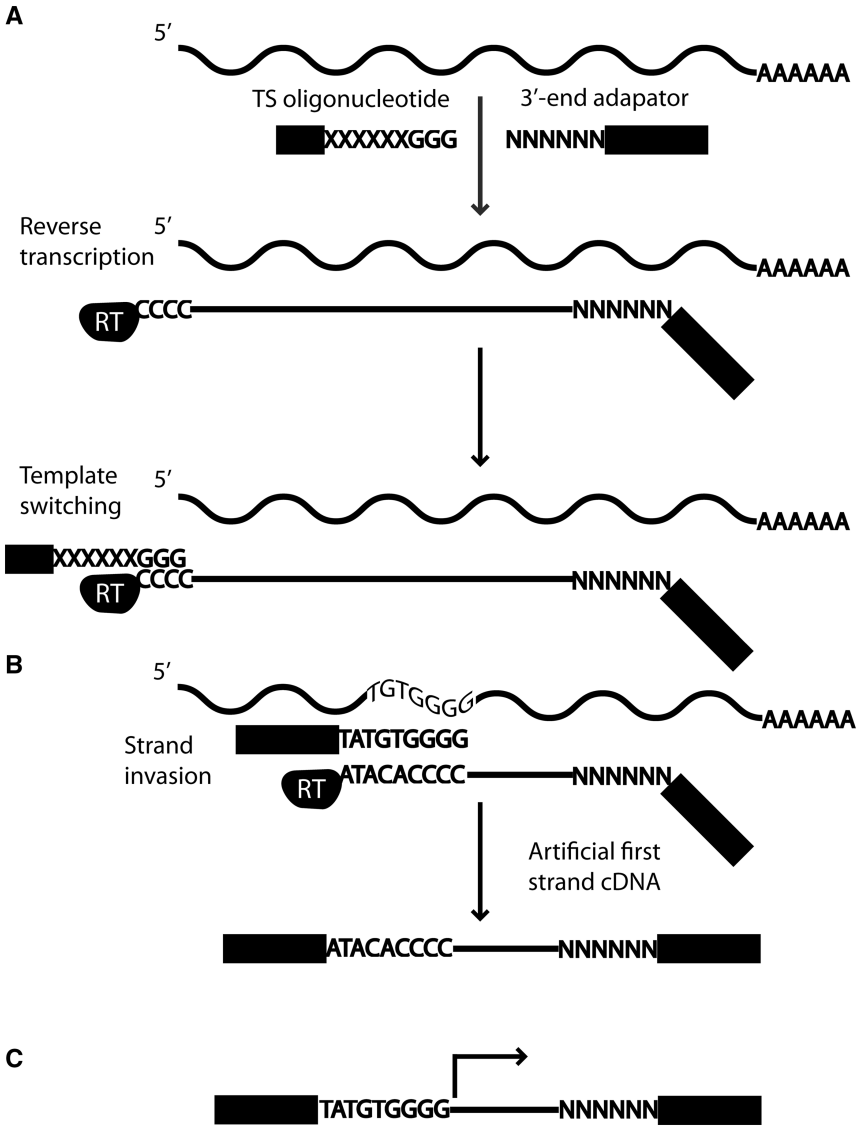

Figure 1. (**A**) The TS mechanism is used for first strand cDNA synthesis. First, an oligonucleotide hybridizes to the RNA molecule, and RT starts polymerizing. Once the RT reaches the 5′ end of the RNA, a cytosine overhang is formed. The TS oligonucleotide containing three riboguanosines hybridizes 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).

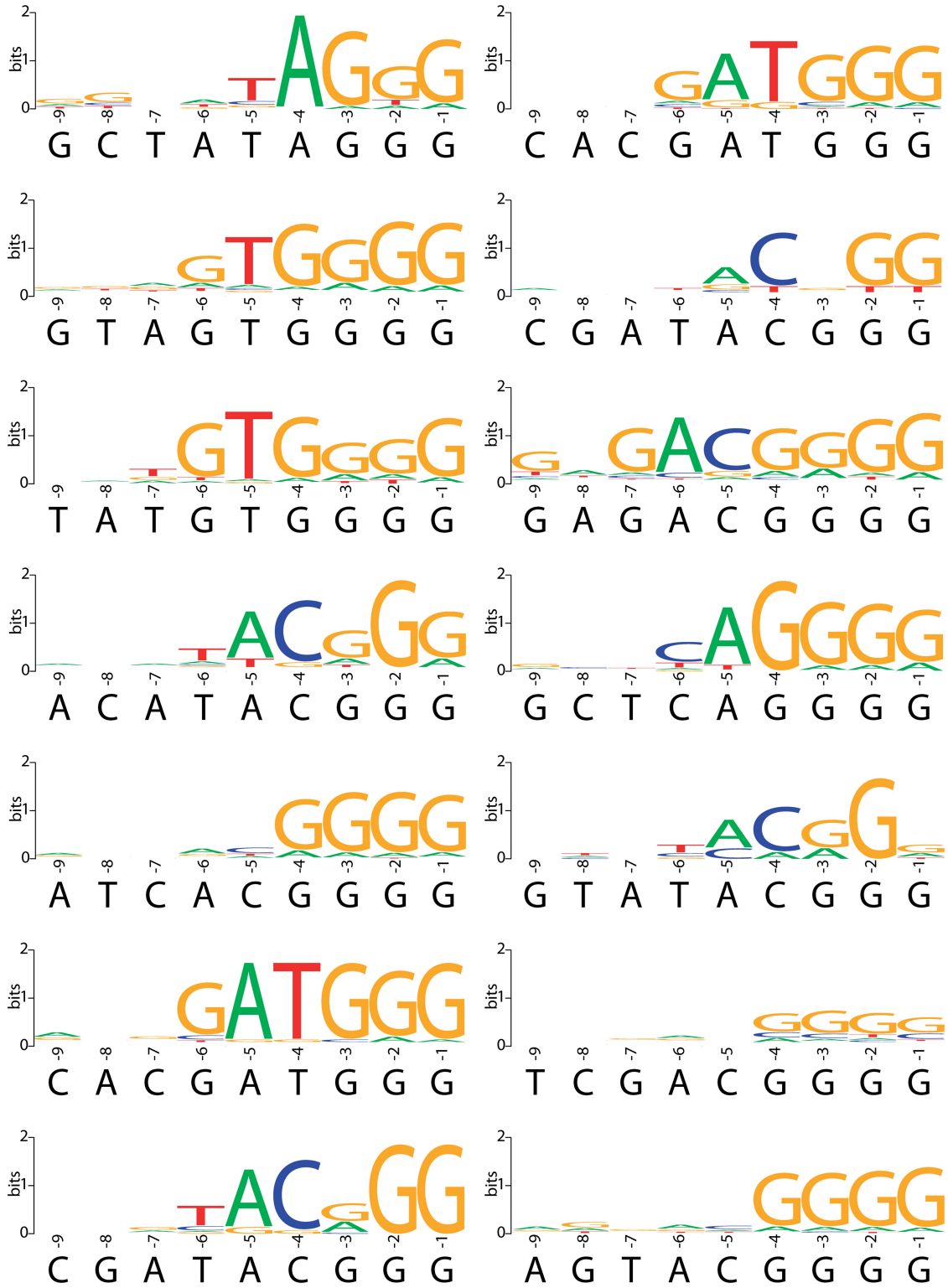

Figure 2. By preparing technical triplicates with different barcodes, we could identify reads present in a barcode specific manner, i.e. barcode-biased reads. The sequence logos were created using the sequence directly upstream of individually mapped barcode biased reads. The barcode sequence used to 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

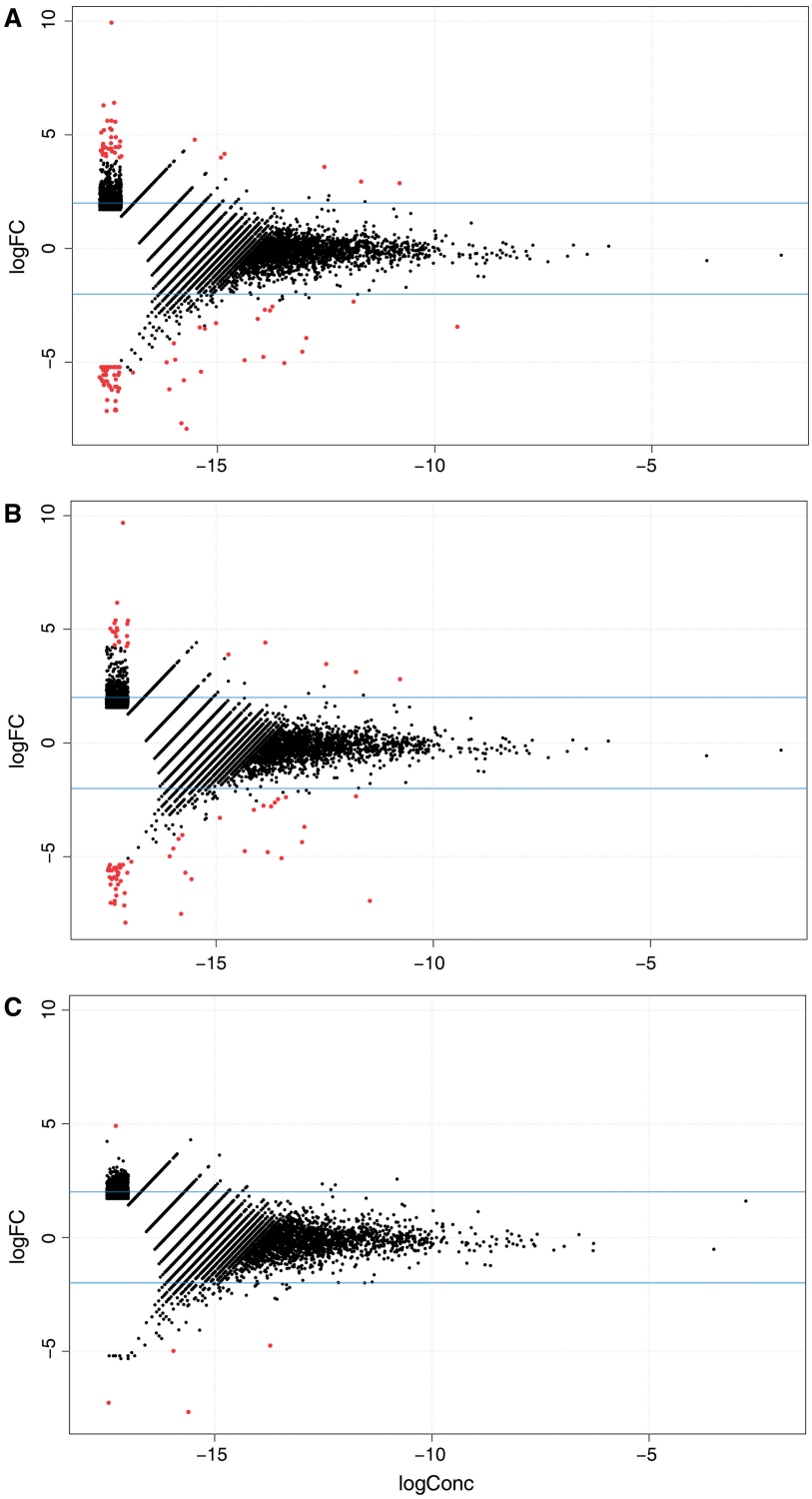

Figure 3. Differential expression analyses were carried out between technical replicates made from different barcode sequences using edgeR. The scatter plots show the log-fold change (*y*-axis) against the log concentration (*x*-axis) for tag clusters present among the 14P technical replicates. In red 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

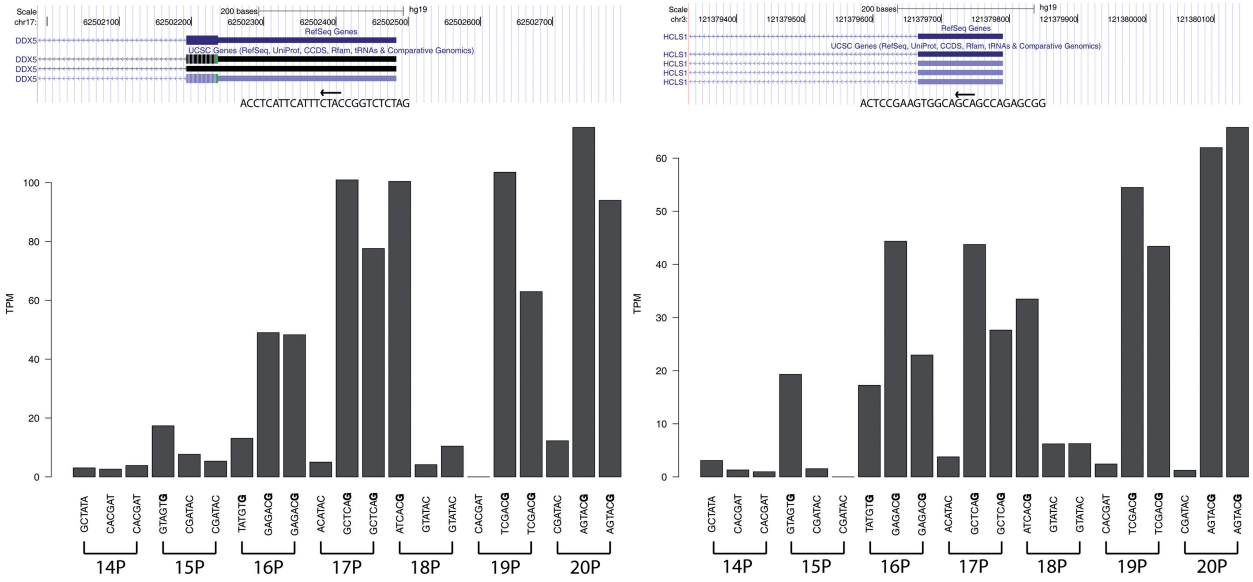

Figure 4. Barcode bias in nanoCAGE technical replicated libraries. The choice of barcode sequence can affect the read count pertaining to the transcriptional starting site. Here, we show two examples for the genes DDX5 and HCLS1, where the read count fluctuates according to the barcode sequence and not 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

It has been previously reported that small RNAs play a role in establishing the DNA damage response[142]. We leveraged the discovery potential of small RNA sequencing to gain some insight into the potential interplay between small RNAs and DNA damage. Intriguingly, we discovered that small RNAs formed near the site of DNA damage. We considered alternative hypotheses that could explain the presence of these small RNAs such as RNA degradation or technical artefacts. We characterised certain features of the DNA damage small RNAs (DDRNAs) to the population of sequenced small RNAs and observed that the bulk of the DDRNAs had a characteristic size and nucleotide profile, suggesting that they are not random degradation products. Furthermore, we performed experiments to suggest possible biogenesis pathways and the knock-down of Dicer and Drosha, which are enzymes involved with small RNA processing, affected the profile of DDRNAs and impaired the DNA damage response. This work has important implications to the observation that the genome is pervasively transcribed, as it offers a hypothesis that pervasive transcription is necessary for maintaining DNA integrity.

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.).

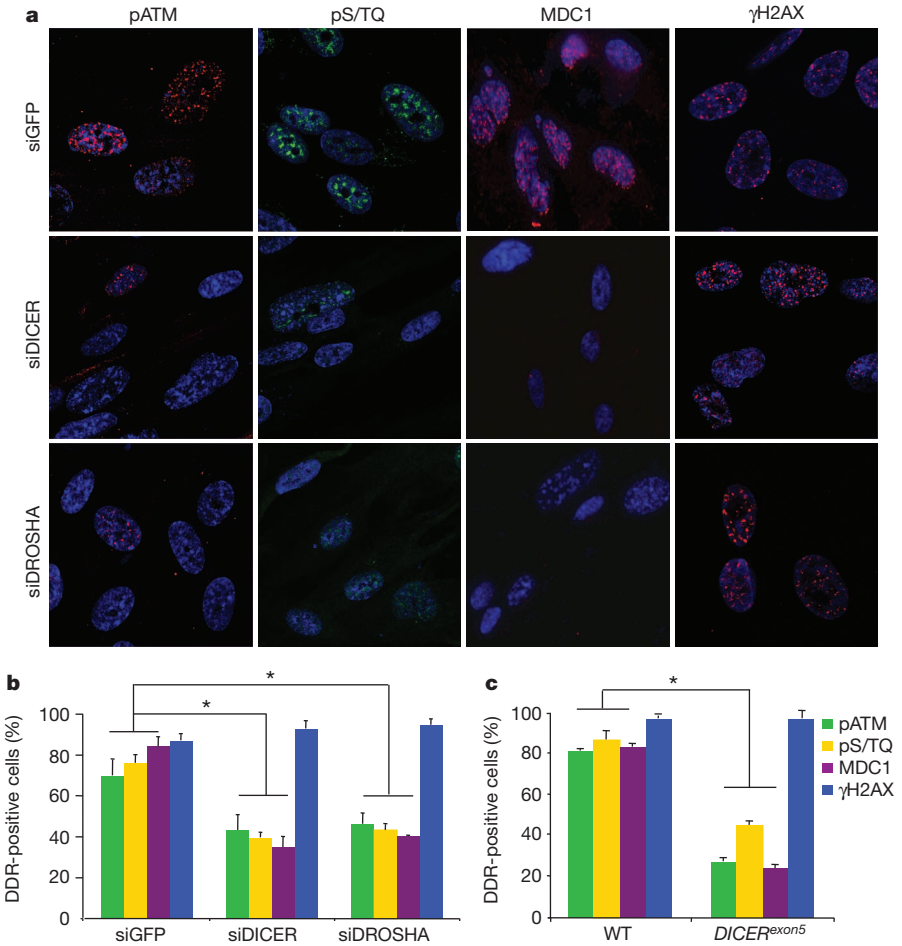

Figure 1 | DICER or DROSHA inactivation impairs DDR foci formation in irradiated cells. **a**, DICER or DROSHA knockdown WI-38 cells were irradiated (10 Gy) and fixed 7 h later. Original magnification, ×250. **b**, Histogram shows the percentage of cells positive for pATM, pS/TQ, MDC1 and γH2AX foci. **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

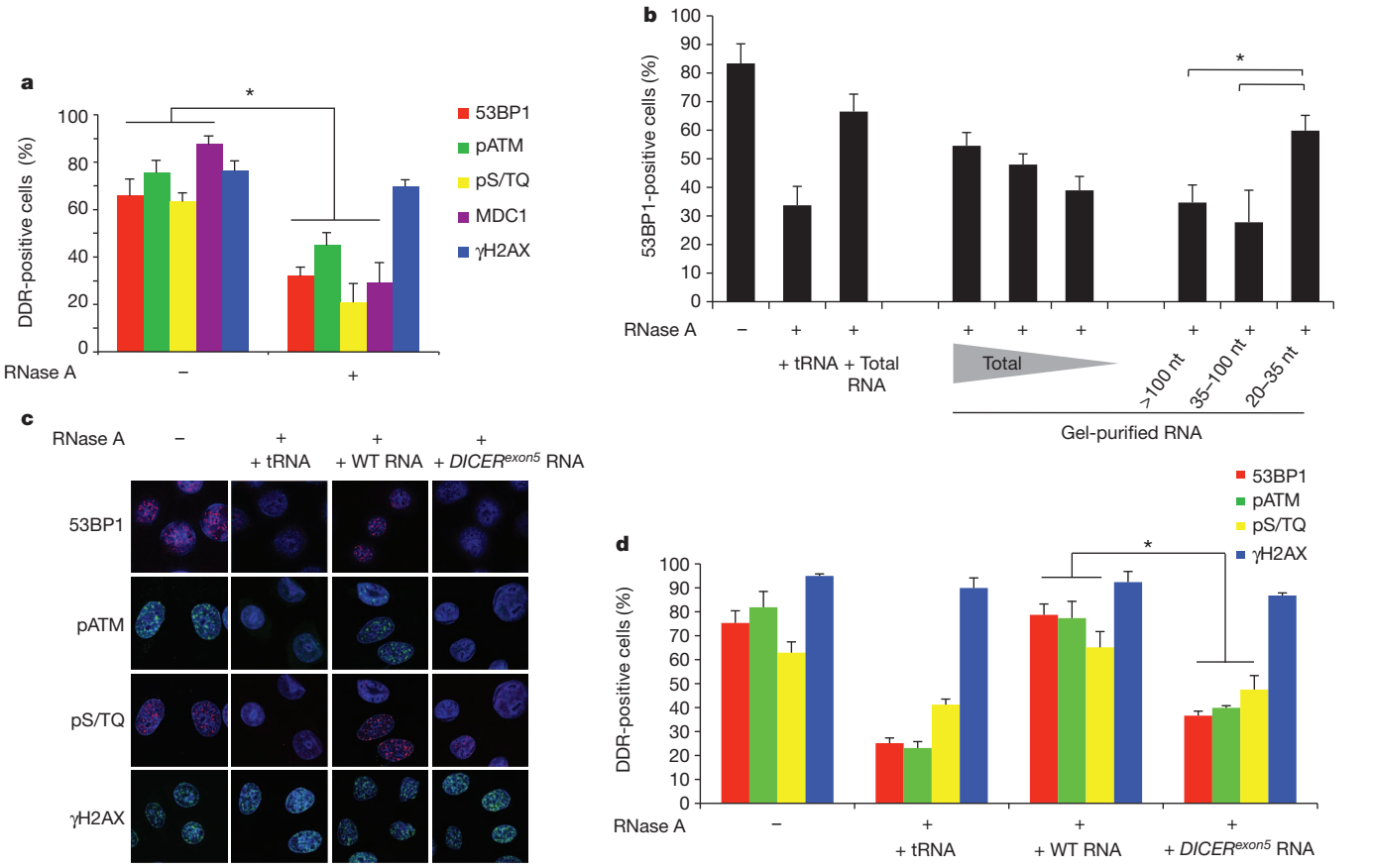

Figure 2 | Irradiation-induced DDR foci are sensitive to RNase A treatment and are restored by small and DICER-dependent RNAs. **a**, Irradiated HeLa cells (2 Gy) were treated with PBS (−) or RNase A (+) and probed for 53BP1, pATM, pS/TQ, MDC1 and γH2AX foci. Histogram shows the percentage of cells positive 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

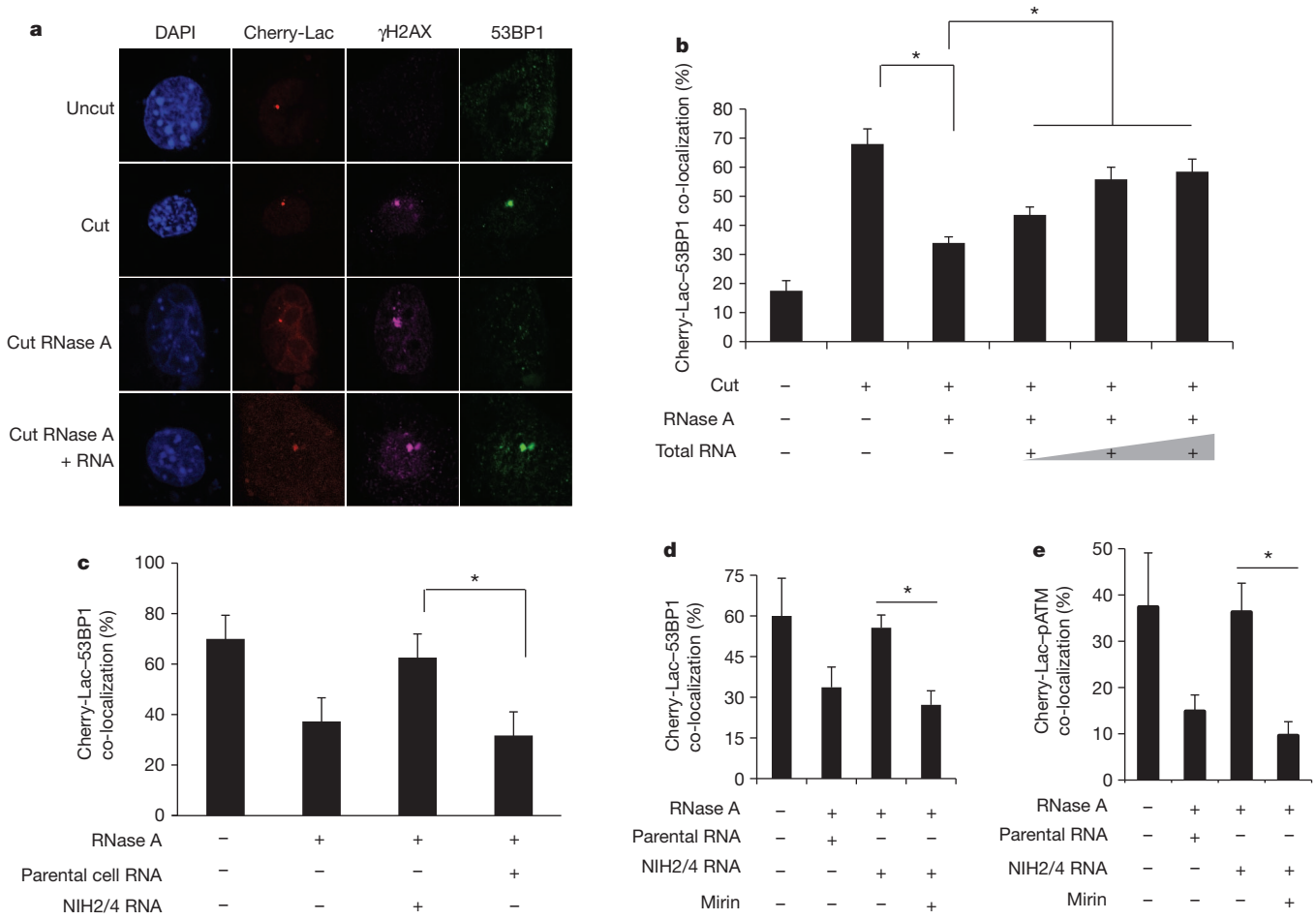

Figure 3 | Site-specific DDR focus formation is RNase A sensitive and can be restored by site-specific RNA in a MRN-dependent manner. **a**, Cut NIH2/4 cells display a 53BP1 and γH2AX focus co-localizing with a Cherry-Lac focus. 53BP1, but not γH2AX, focus is sensitive to RNase A and is restored by incubation 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.

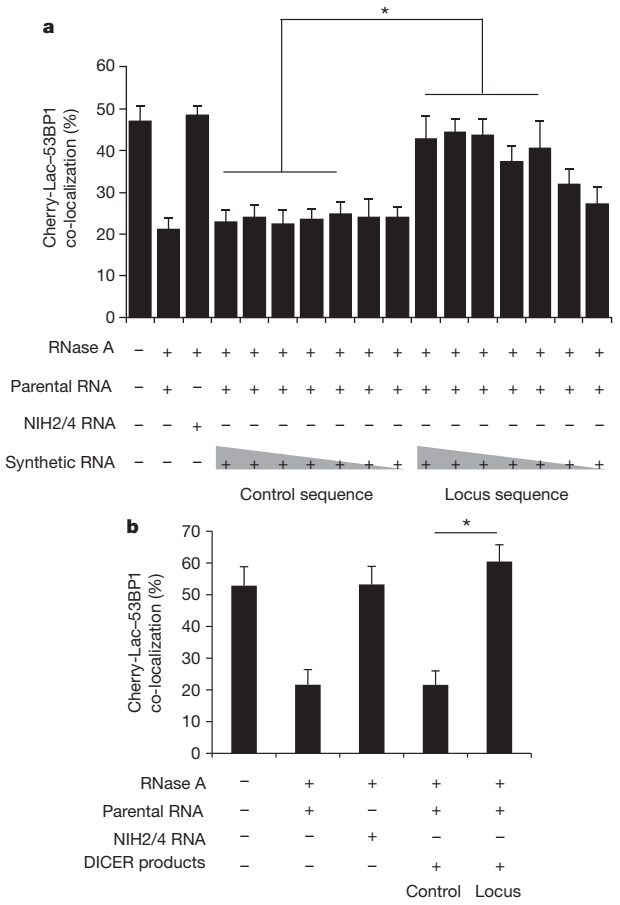

Figure 4 | Chemically synthesized small RNAs and *in vitro*-generated DICER RNA products are sufficient to restore DDR focus formation in RNase-A-treated cells in a sequence-specific manner. **a**, Chemically synthesized RNA oligonucleotides were annealed and were tested to restore DDR focus formation 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

Regulated expression of repetitive elements

It is well established that the human genome is pervasively transcribed and this has been attributed to a "leaky" transcriptional system. Our observation that small RNAs are formed near the vicinity of DNA damaged sites and are required for establishing the DNA damage response, implies that all RNAs are potentially useful, since any site in the genome is vulnerable to DNA damage. A class of RNA that is often neglected from study are those that form from the repetitive portion of the genome. It is estimated that at least half of the human genome is made up of repetitive elements (REs) and transcription initiation has been observed within these elements[114]. However, the general impression is that these elements are non-functional and are simply transcriptional noise. Motivated by the hypothesis that all RNAs could be potentially useful, we quantified and catalogued expression signal from REs using the FANTOM5 atlas.

The regulated expression of repetitive elements across human cell types and tissues

Dave Tang*[1], the FANTOM5 consortium, and Piero Carninci†[1]

[1] *RIKEN Center for Life Science Technologies (Division of Genomic Technologies)*

November 10, 2014

Abstract

A large fraction of the human genome is composed of repetitive elements (REs), many of which are transposable elements (TEs). Most of these TEs are molecular fossils that have long lost the ability to transpose. However, transcription can initiate from these elements as they contain requisite signals necessary for transcription. However, it is unclear whether these transcriptional products have any functional purpose or are just transcriptional by-products. In order to address this question, we performed the most comprehensive survey of transcriptional events arising from TEs by using the FANTOM5 human atlas of Cap Analysis Gene Expression (CAGE) libraries, which consists of a large panel of human cell and tissues types (totalling 988 samples). We observed that TEs were pervasively transcribed, lowly expressed on average compared to protein-coding genes, have restrictive expression patterns, overlap enhancer sequences, serve as precursors to small RNAs, and drive the transcription of long noncoding RNA. Furthermore, based on the expression patterns of TEs, we could associated to families of TEs to specific ontological categories ascribed to the FANTOM5 samples. This hints at the fact that characteristics of certain families of TEs, such as their sequence composition, may be associated to factors that are specific to particular cell types. These lines of evidence support the hypothesis that various TEs have become exapted and have a functional role in the human genome.

*Email: `dave.tang@riken.jp`
†Email: `carninci@riken.jp`; Corresponding author

1 Introduction

Roughly half of the human genome is made up of repetitive elements (REs), consisting mainly of completely inactive transposable elements (TEs)[1]. The estimated number of active TEs in the human genome is less than 0.05%[2], which is as expected given that TEs are a major factor in causing mutations that lead to human disease[3]. However, despite the detrimental effects of TEs, their ability to move around has shaped the evolution of the genomes in which they reside[4]. For example, TEs were found to contribute transcription factor binding sites (TFBSs) to various mammalian transcription factors[5] and it has even been suggested that many TFBSs had originated from TEs[6]. Furthermore, in a study that examined the methylation status of TEs, it was demonstrated that methylation patterns in TEs were tissue-specific and displayed enhancer hallmarks[7]. A separate study examining transcription initiation events arising within TEs, observed that TEs were used as alternative promoters and are expressed in a tissue-specific manner[8], corroborating with tissue-specific methylation patterns. These studies imply that TEs are developmentally regulated and may contribute to driving the large diversity of cell types.

Numerous RNA sequencing studies have revealed a large set of long non-coding RNAs (lncRNAs) in the human genome[9, 10, 11]. In a study examining TEs and lncRNAs, it was shown that lncRNAs are enriched with TE sequences, leading to the hypothesis that TEs have contributed to the origin of many lncRNAs[12]. As TEs naturally contain transcriptional regulatory signals, this is not surprising; for example, the long terminal repeats (LTRs) of retrotransposons have been found to drive the expression of nearby genes[13] and the transcription start site (TSS) of the lncRNAs *BANCR*, *lnc-RoR*, and *lncRNA-ES3* all overlap the LTR sequence of separate endogenous retroviruses[12]. In addition, these lncRNAs only exist in the genomes of organisms that descended from the last common ancestor which contains the retrovirus, clearly supporting the idea that these lncRNAs arose from TEs[**?**]. Furthermore, a specific TE family was found to drive the expression of stem cell specific lncRNAs[14], suggesting that TEs are able confer tissue-specificity possibly by associating with specific regulatory signals.

However, the consideration that TEs may be functional products of the genome is generally met with scepticism due to historical views that TEs are simply selfish elements that serve no purpose apart from propagating itself[15] and provides little to no selective advantage to organisms[16]. Furthermore, the elaborate silencing systems that repress the transposition of mobile genetic elements in the human genome suggest that they molecular pests that need to be silenced[17]. However, in their original paper, Orgel and Crick suggested at the hypothesis that these selfish DNA sequences may become exapted for control purposes[16]. Furthermore, the technical difficulties associated with REs have limited the number of genome-wide studies investigating their potential roles. In the era of microarray technologies, cross-hybridisation issues made it difficult to study the expression of REs. High-throughput sequencing technology has made it possible to quantify the expression of REs, however the short read nature of

these technologies has made it difficult to map reads to REs.

It has been previously reported that 75% of reads that are 25 nucleotides long can be mapped uniquely to the human genome and at 60 nucleotides long 95% uniqueness can be achieved[18]. However, this estimate assumes an equal distribution of reads in a given library, which is typically not the case. When dealing with a read that maps to multiple places, there are usually three choices on what to do: a) Discard the read, b) Take the best alignment and if there are multiple best hits, take one randomly, and c) Report all alignments or report up to a certain number[19]. The third choice can also be incorporated with a strategy that uses uniquely mapped reads to probabilistically weight multi-mapping reads[20]. The basic premise is that a region that is transcriptional active is more likely to have given rise to a read than a transcriptional inert region. Other strategies include mapping to a consensus database of REs, such as RepBase Update[21], or mapping to regions of a genome that has been annotated as being repetitive[7], or combining reference genome mapping with consensus sequence mapping[22]. These strategies aim to utilise as many reads as possible to obtain a more representative expression profile.

In this study, we processed over two billion reads from 988 FANTOM5 CAGE libraries and overlaid them to REs annotated using profile hidden Markov models. We found that REs are expressed in a tissue-specific manner and expression signal from REs can be used to produce biologically meaningful clusters. Furthermore, expression profiles of specific REs can be associated to specific ontologies, suggesting that specific families of REs may be involved in specific functions. Lastly, by examining the genomic locality of expressed REs, we observed that they overlapped genomic regions known to produce small RNAs and lncRNAs, as well as enhancer regions more often than expected by chance.

2 Methods

2.1 Annotating repeats in the human genome

RepeatMasker[23] is a program that screens DNA sequences for repetitive elements (REs); the tool relies on a search algorithm and a database of RE profiles. Traditionally, homology-based tools such as cross_match and variants of BLAST have been used to screen DNA sequence against RE consensus sequences, the most commonly used database being Repbase Update[21]. Recently a database of REs based on profile hidden Markov models was developed, called Dfam[24], which allowed screening of REs using a hidden Markov model search tool, called nhmmer[25]. It has been reported that screening for REs using nhmmer and Dfam is more sensitive and specific than consensus sequence based approaches[24]. For this reason, we annotated REs in the human genome (hg19) using RepeatMasker (4.0.3), nhmmer (hmmer-3.1b1), and Dfam (1.2). Specifically, we ran the command `RepeatMasker -e hmmer -species human -s -xsmall -pa 8 chr.fa`, for each assembled chromosome. REs were classified by class, family, and individual element names.

2.2 Aggregating CAGE reads to repetitive elements

The details describing the preparation of the Cap Analysis Gene Expression (CAGE) libraries for the FANTOM5 project are described elsewhere[26]. Briefly, a CAGE protocol optimised for the HeliScope Genetic Analysis System was developed and used to prepare 988 FANTOM5 libraries. A high-throughput short read sequence alignment program called Delve was used to map the CAGE reads to the human genome (hg19). Delve is able to recognise sequencing biases or increased error rates in homopolymer stretches, which makes it suitable for the HeliScope sequencer. Mapping qualities following the Phred scale were provided for each mapped read, where the qualities are probabilities that a mapped read is incorrect[27].

To aggregate CAGE reads to REs, the coordinates of the mapped reads were intersected with the RE coordinates using intersectBed from the BEDTools suite[28]; parallelisation of the computations was achieved using GNU parallel[29]. For each repeat class (1087 in total), we tallied the number of reads that intersected that class; thus for each FANTOM5 library, a tally was produced for each repeat class resulting in a 1087×988 matrix. We performed this aggregation step using reads thresholded at various mapping qualities $(0, \ldots, 10)$. Finally, tallies for each library were normalised by tags per million (TPM); library size was the total number of reads that intersected the repeat classes.

2.3 Markov clustering

We calculated the Spearman's rank correlation coefficient between all repeat classes (590,241 pairwise calculations) and all libraries (487,578 pairwise calcu-

lations) using the matrix of aggregated CAGE reads. Correlations between REs and libraries were represented as a graph, where the nodes or vertices represented a single RE class and the edges or connections represented a correlation between the two nodes. We used the Markov clustering (MCL) algorithm[30] to reveal natural groups within the graphs using only nodes that had a correlation of 0.96 or better to another node. The algorithm simulates flow within a graph and promotes flow in a highly connected region and demotes less connected regions. The MCL algorithm takes one parameter, the inflation parameter, which adjusts the granularity of the clusters. We tested various inflation parameters between two to ten, and used four as this was a good compromise between the number of clusters and cluster sizes. The graphs were visualised using the Cytoscape software [31].

2.4 FANTOM5 sample ontology enrichment analysis

Structured ontologies were developed and used to annotate the FANTOM5 libraries, allowing the identification of enriched biological properties based on CAGE expression profiles[32]. Specifically, samples were annotated using the The Open Biological and Biomedical Ontologies (`http://www.obofoundry.org/`) and the structured ontologies, were grouped into hierarchical cellular, anatomical, disease and experimental ontologies. Each ontological term can be used to separate libraries in a binary manner, where a library either has membership (x) or no membership (y) to a particular ontology. To test for ontology enrichment, the TPM expression of libraries in x were compared to the expression of libraries in y using a Mann-Whitney-Wilcoxon test; this was performed on all 846 ontologies and for all all repeat classes (1,087). The p-values were adjusted following the Benjamini & Hochberg method[33] and resulted in a 1087×846 matrix of adjusted p-values. Each element of this matrix corresponds to a p-value indicating whether the expression profile of a particular repeat class enriches a particular ontology.

2.5 Parametric tag clustering of CAGE reads

The FANTOM5 CAGE libraries were previously clustered using a decomposition-based peak identification (DPI) method that utilised a stringent mapping quality threshold of 20[32]. This criteria removes signal arising from REs and is inappropriate for studying the expression of REs. We used a tag clustering method known as parametric clustering[34], which uses maximal scoring segments to clusters reads. Specifically for every maximal scoring segments, the minimum and maximum values of the density parameter, d, are reported and used to assess whether a cluster is robust and not formed due to random fluctuations in the data set. We kept tag clusters with at least ten raw tags, a maximum density / minimum density ratio of at least two, and limited tag clusters to a length of 200 bps. We performed tag clustering using reads thresholded at various mapping qualities $(0, \ldots, 10)$ and to simplify the large number of tag clusters, we

took the largest tag cluster that could encompass all other tag clusters, which we called non-overlapping tag clusters.

2.6 Tag cluster annotation

We used the intersectBed tool from the BEDTools suite to annotate tag clusters to GENCODE (v19) transcripts[35], REs, FANTOM5 permissive enhancers[36], FANTOM5 small RNAs, and long intergenic non-coding RNAs[9]. We separated the genome into 5 separate classes based on the GENCODE annotations and annotated each tag cluster hierarchically in the order: promoter, exon, intron, repetitive elements, and intergenic region. Genomic regions +/- 200 bp around the starting site of a GENCODE transcript was considered the promoter region of that transcript. Exonic regions were defined as regions overlapping the exons of GENCODE transcripts. Intronic regions were defined as the region remaining from the subtraction (using substractBed) between a GENCODE gene model and the exonic regions. Intergenic regions were defined as the remaining region from the subtraction between gene models and the genome sequence. FANTOM5 small RNA libraries were clustered in the same manner as the CAGE data (see section 2.5).

We used the GenometriCorr package[37] to calculate potential correlations between two sets of genomic features: a query and a reference set. Relative and absolute distances between query and reference features are tested against a uniform distribution of distances. The significance of overlap between two sets of features is tested using a projection test, using a binomial test, where the probability of overlap is based on the coverage of reference features and by using the Jaccard index defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

which measures the amount of overlap between two features. To test where observed intersections are statistically significant, a null distribution was created by measuring the Jaccard index of 1,000 permutations of the query features.

2.7 Measuring expression specificity

The Shannon entropy has been previously used as a metric to characterise the overall expression specificity of a gene amongst a panel of samples[38]. We used the Shannon entropy as a measure for expression specificity and calculated the metric as follows:

$$-\sum_{i=1}^{n} x_i \cdot log_2 x_i$$

where i is the library index and x_i the expression of a feature in a particular library. In addition, we normalised the expression of a feature in a single library by the total expression of the features in all libraries. The Shannon entropy,

measured in bits, ranges from zero for transcripts expressed only in a single sample to $log_2 n$ for features that are expressed uniformly across all n samples.

In addition, we used the h-index[39] as a measure of expression ubiquity, which is defined as:

$$H(x) = max\{i = 1, \ldots, n : x_i \geq i\}$$

where the h-index, $H(x)$, is the maximum value in the set of sorted expression values, i, such that the expression of the x_i library is greater than or equal to i. For example, a tag cluster with a h-index of one is supported by one CAGE read in at least one library. A tag cluster with a h-index of two is supported by at least two CAGE reads in at least two libraries, and so on.

All code used for this work is available at `https://github.com/davetang/fantom5_repeat`.

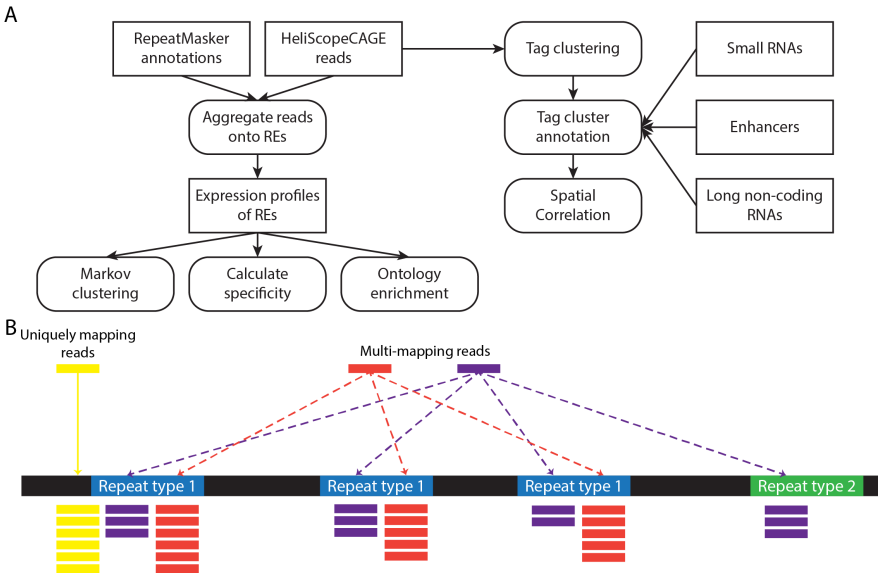

Figure 1: Summary of the methods. A) The pipeline of this study; rectangles show data sources and rounded rectangles show bioinformatic procedures. B) The aggregation strategy used to assign reads to repeat types. In the example above, 3 scenario are shown: the first scenario shows the unambiguous assignment of reads to a repeat type, the second scenario shows how multi-mapping reads are all randomly assigned to the same repeat type, and the third scenario shows how multi-mapping reads can be randomly assigned to different repeat types, which can occur in repeats belonging to a similar class.

3 Results

3.1 Shannon entropy

We quantified the expression of REs, in two independent ways: by aggregating reads onto REs and by tag clustering and intersection (Figure 1). In the first approach, CAGE reads were tallied across 1,087 RE classes (Figure 1B) to measure the overall expression strength of REs. This approach of aggregation is robust against multi-mapping reads, which are likely to multi-map to the same RE class; we compared the expression matrices tallied with reads at different mapping qualities and they showed very high correlations. The lowest correlation (Spearman's rho = 0.83) was between the expression matrices prepared using all reads against using reads thresholded with a mapping quality of 10 or better. Next, we used the Shannon entropy to measure the specificity of RE expression and this revealed several RE classes that had a much more restrictive expression pattern across the FANTOM5 libraries (Figure 2). Expression of LTR7 was restricted to pluripotent and embryonic stem cells, which has been previously reported[14]. Expression of MER74C was found to be restricted to blood samples and MER41E expression was restricted to placental samples. In order to associate the expression of REs to biological functions, we performed a sample ontology enrichment analysis.

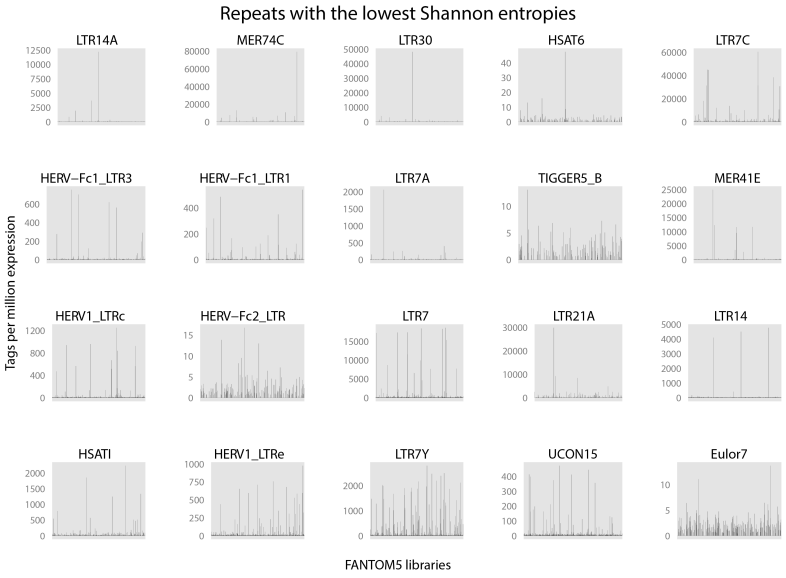

Figure 2: The 20 repetitive element classes with the lowest Shannon entropies; the x-axis shows each individual FANTOM5 library and the y-axis shows the tags per million expression.

3.2 Sample ontology enrichment analysis

FANTOM5 libraries have been associated to particular ontologies[32], which allowed libraries to be separated into two groups for each ontology: those that are associated with the ontology and those that are not. Based on this separation, we tested whether the RE expression between the two groups was statistically different, i.e. sample ontology enrichment. This resulted in 919,602 tests (Figure 3), of which 69,038 associations were statistically significant (adjusted p-value <0.05).

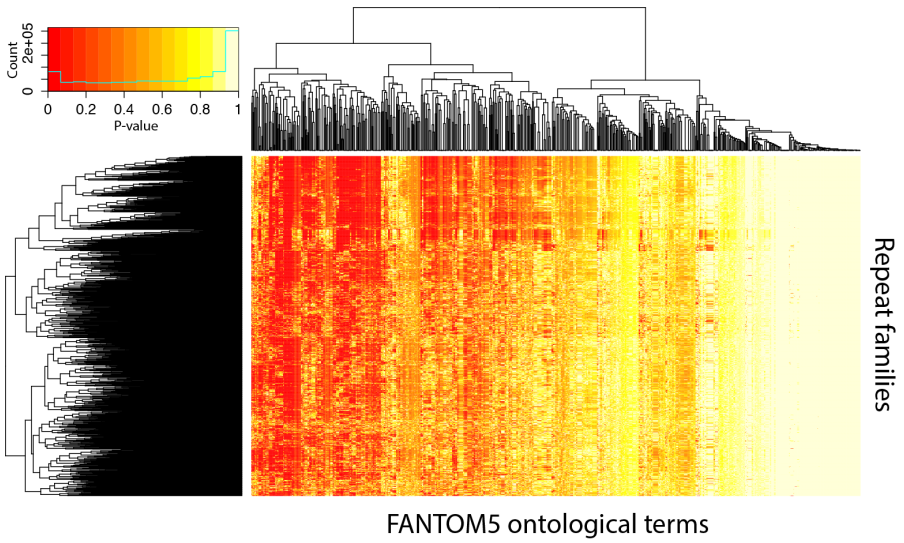

Figure 3: Heatmap of p-values from the sample ontology enrichment analysis; rows represents particular repeat families and columns represent FANTOM5 ontologies.

LTR7 had 80 sample ontologies that were significantly enriched including "embryonic stem cell" and "H9 embryonic stem cell line" (adjusted p-value for both ˜0.00693). MER74C had 59 ontologies that were significantly enriched including "blood" (adjusted p-value ˜0.001952) and "whole blood" (adjusted p-value ˜0.00012). The sample ontology "cancer", contained the most number of enriched REs with 656 out of a total of 1087, which suggests that REs are over-expressed in cancerous samples.

3.3 Markov clustering

We constructed a network of FANTOM5 libraries based on the aggregated expression of REs to determine whether the expression patterns from RE is enough to form biological meaningful clusters. FANTOM5 libraries were connected based on the expression profile correlations of each library against every other

library (Figure 4A). The highly connected network, showed that most repeats were expressed in a similar manner across all libraries. To reveal any natural groups within this graph, we performed Markov clustering (MCL), which is an unsupervised cluster algorithm based on simulation of stochastic flow in graphs. The natural groups revealed by the MCL algorithm (Figure 4B), were FANTOM libraries that were biological related. For example, induced pluripotent stem cells clustered with human embryonic stem cells and embryoid bodies (Supplementary figure 2). This suggests that different cell or tissue types have a pronounced RE expression pattern.

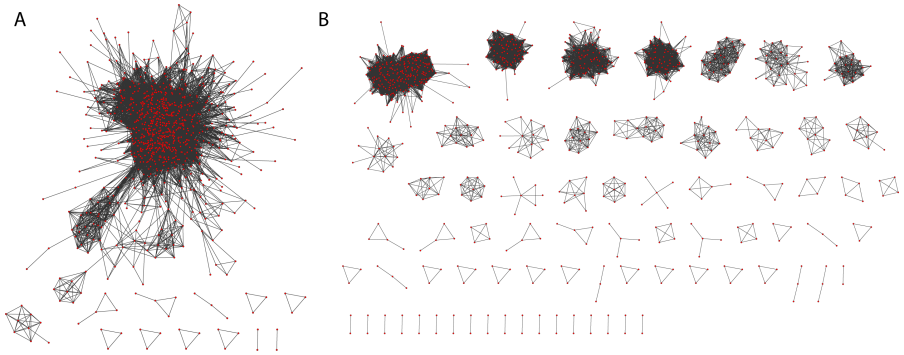

Figure 4: Representing the expression patterns of REs as graphs. A) Each red node represents a FANTOM5 library and each edge represents a Spearman's correlation of 0.96 between two nodes B) Markov clustering revealed the natural groups present in the correlation graph.

3.4 Tag clustering

In the second approach of quantifying the expression of REs, we clustered all mapped FANTOM5 CAGE reads, at various mapping qualities, using a parametric clustering methods[34]. This tag clustering approach, as opposed to the aggregation method, allows REs to be put into context of other genomic features, such as known transcript models. We annotated tag clusters to GENCODE transcripts in a hierarchical manner and to avoid confounding signal; thus tag clusters annotated as RE, do not overlap known transcripts. We performed this annotation step using reads mapped at various mapping qualities to assess the impact of mapping qualities (Figure 5).

We decided to use tag clusters using reads at a mapping quality of ten or better, despite losing a large number of tag clusters.

3.5 Genomic correlations

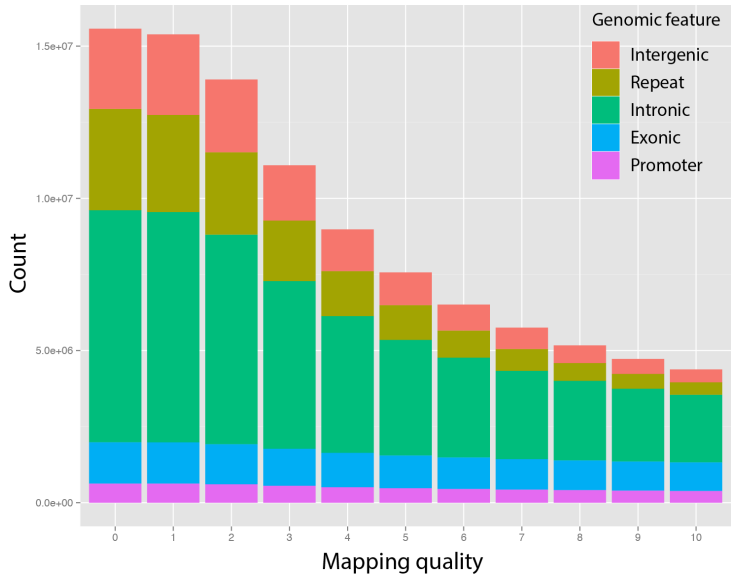

Figure 5: GENCODE annotation of tag clusters using reads at various mapping qualities.

Genomic feature	Tag clusters	Mean expression	Median expression
Promoter	390162	9914	52
Exonic	937236	391	33
Intronic	2219891	87	20
Repetitive	410313	210	20
Intergenic	421534	126	21

Table 1: Tag cluster statistics using reads with a mapping quality of 10 or better.

4 Discussion

	sRNAs	Enhancers	lincRNAs
Number of genomic features	610246	43011	14281
Relative distance p-value	0	0	1.63e-14
Absolute distance p-value	<0.001	<0.001	<0.001
Projection test p-value	0	0	0.00264
Jaccard measure p-value	<0.001	<0.001	<0.001

Table 2: P-values of statistical tests carried out by the GenometriCorr package.

5 Data deposition

All CAGE data has been deposited at DDBJ DRA under accession number DRA000991

6 Funding

FANTOM5 was made possible by a Research Grant for RIKEN Omics Science Center from MEXT to Y.Hayashizaki and a Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to Y. Hayashizaki. D. Tang is supported by the European Union Seventh Framework Programme under grant agreement FP7-People-ITN-2008-238055 ("BrainTrain" project) to P. Carninci.

7 Authors' contributions

All authors read and approved the final manuscript.

References

[1] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[2] R. E. Mills, E. A. Bennett, R. C. Iskow, and S. E. Devine. Which transposable elements are active in the human genome? *Trends Genet.*, 23(4):183–191, Apr 2007.

[3] P. A. Callinan and M. A. Batzer. Retrotransposable elements and human disease. *Genome Dyn*, 1:104–115, 2006.

[4] R. Cordaux and M. A. Batzer. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.*, 10(10):691–703, Oct 2009.

[5] G. Bourque, B. Leong, V. B. Vega, X. Chen, Y. L. Lee, K. G. Srinivasan, J. L. Chew, Y. Ruan, C. L. Wei, H. H. Ng, and E. T. Liu. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.*, 18(11):1752–1762, Nov 2008.

[6] C. Feschotte. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.*, 9(5):397–405, May 2008.

[7] M. Xie, C. Hong, B. Zhang, R. F. Lowdon, X. Xing, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat. Genet.*, 45(7):836–841, Jul 2013.

[8] G. J. Faulkner, Y. Kimura, C. O. Daub, S. Wani, C. Plessy, et al. The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, 41(5):563–571, May 2009.

[9] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, 25(18):1915–1927, Sep 2011.

[10] A. M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 106(28):11667–11672, Jul 2009.

[11] M. Guttman, J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, 477(7364):295–300, Sep 2011.

[12] A. Kapusta, Z. Kronenberg, V. J. Lynch, X. Zhuo, L. Ramsay, G. Bourque, M. Yandell, and C. Feschotte. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.*, 9(4):e1003470, Apr 2013.

[13] C. J. Cohen, W. M. Lock, and D. L. Mager. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene*, 448(2):105–114, Dec 2009.

[14] D. Kelley and J. Rinn. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.*, 13(11):R107, 2012.

[15] W. F. Doolittle and C. Sapienza. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–603, Apr 1980.

[16] L. E. Orgel and F. H. Crick. Selfish DNA: the ultimate parasite. *Nature*, 284(5757):604–607, Apr 1980.

[17] N. Yang and H. H. Kazazian. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat. Struct. Mol. Biol.*, 13(9):763–771, Sep 2006.

[18] N. Whiteford, N. Haslam, G. Weber, A. Prugel-Bennett, J. W. Essex, P. L. Roach, M. Bradley, and C. Neylon. An analysis of the feasibility of short read sequencing. *Nucleic Acids Res.*, 33(19):e171, 2005.

[19] T. J. Treangen and S. L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, 13(1):36–46, Jan 2012.

[20] G. J. Faulkner, A. R. Forrest, A. M. Chalk, K. Schroder, Y. Hayashizaki, P. Carninci, D. A. Hume, and S. M. Grimmond. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, 91(3):281–288, Mar 2008.

[21] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110(1-4):462–467, 2005.

[22] D. S. Day, L. J. Luquette, P. J. Park, and P. V. Kharchenko. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol.*, 11(6):R69, 2010.

[23] A. F. A. Smit, R. Hubley, and P. Green. RepeatMasker Open-3.0, 1996-2004.

[24] T. J. Wheeler, J. Clements, S. R. Eddy, R. Hubley, T. A. Jones, J. Jurka, A. F. Smit, and R. D. Finn. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.*, 41(Database issue):70–82, Jan 2013.

[25] T. J. Wheeler and S. R. Eddy. nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 29(19):2487–2489, Oct 2013.

[26] M. Kanamori-Katayama, M. Itoh, H. Kawaji, T. Lassmann, S. Katayama, et al. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.*, 21(7):1150–1159, Jul 2011.

[27] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–1858, Nov 2008.

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

[29] O. Tange. Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47, Feb 2011.

[30] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30(7):1575–1584, Apr 2002.

[31] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, Nov 2003.

[32] 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.

[33] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[34] M. C. Frith, E. Valen, A. Krogh, Y. Hayashizaki, P. Carninci, and A. Sandelin. A code for transcription initiation in mammalian genomes. *Genome Res.*, 18(1):1–12, Jan 2008.

[35] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, 22(9):1760–1774, Sep 2012.

[36] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, Mar 2014.

[37] A. Favorov, L. Mularoni, L. M. Cope, Y. Medvedeva, A. A. Mironov, V. J. Makeev, and S. J. Wheelan. Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput. Biol.*, 8(5):e1002529, May 2012.

[38] J. Schug, W. P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, 6(4):R33, 2005.

[39] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.*, 102(46):16569–16572, Nov 2005.

Chapter 5

General discussion

The work carried out as part of this thesis had the underlying goal of interpreting the transcriptional output from different biological systems. In the various studies, different transcriptome profiling methods employing the use of high-throughput sequencers were used to catalogue and quantify the expression of transcripts. While the genome is largely static, the transcriptome is highly dynamic and the cohort of expressed transcripts at a specific development stage or physiological condition will allow us to understand development and disease on a molecular level. However, two problems complicate the study of transcriptomes: 1) the presence of artefacts and biases and 2) pervasive transcription. In order to properly interpret transcriptional output, it is necessary to filter out the noise from the biological signal.

Transcriptome profiling falls under the category of functional genomics, in which the aim is to identify functional elements in the genome by evidence of transcription. The "read out" from such experiments represents the expression of transcripts, which by itself is not entirely informative. The raw signal needs to be annotated or in order words, put into perspective, usually with respect to the genome. In the case of transcriptome profiling via high-throughput sequencing, the raw signal, i.e. sequenced reads, are first processed and mapped back to the genome. The processing step involves quality control steps that remove reads with missing features, ambiguously called bases, and reads that are composed of the primer or adaptor sequences used in the library preparation. Furthermore, different thresholds are usually implemented after read mapping to remove potentially spurious events; these threshold can be based on the mapping confidence or mapping quality, and expression strength. The mapped reads are then annotated to genomic features, such as protein-coding genes and repetitive elements, or to genomic properties, such as sequence conservation to other genomes and the nucleotide composition of a region. Finally, the annotated data is interpreted with respect to the experiment design; "sanity checks", which are basic tests that can quickly assess the validity of the results, are commonly performed.

The identification of artefacts and biases is important prior to the analysis of data; while tools[122, 143] are available for performing quality control steps, they are unable to deal with idiosyncratic biases from specific protocols. In chapter 2, we demonstrated a particular type of bias that was introduced with the use of molecular barcodes. DNA barcodes are used to allow different RNA

sources to be pooled at an early stage of sample preparation and are easily introduced by modifying the sequences of primers or linkers. This has two main benefits in that the starting material is increased and different samples can be sequenced on a single flowcell, increasing the cost-effectiveness of the experiment. In our study, we observed that the efficiency of primer annealing was dependent on the barcode sequence[1], thus creating a barcode bias. Furthermore, we noticed a particular type of artefact that was artificially produced during the template-switching step of the experimental protocol[1]. Specifically, these template-switching artefact, which are a result of strand invasion, interrupted the process of reverse transcription when the cDNA has increased sequence complementary to the template-switching primer. The bias was identified when libraries prepared with the same or similar barcodes clustered together, regardless of the biological origin of the library. We mitigated this bias by introducing a spacer sequencer into the template-switching primer, which standardised the template-switching efficiency.

Another complication with interpreting transcriptional output is pervasive transcription. There have been two surprises in the genomics field since the completion of the draft human genome sequence[36, 87]. The first was the number of genes in the human genome, which many genome biologists had over-estimated prior to the sequencing of the human genome[144]. This came as a surprise, as it was thought that organismal complexity was a consequence of the number of genes; however, humans have roughly the same number of genes as *Caenorhabditis elegans*. The next surprise was the observation that most of the human genome was transcribed[59, 63], i.e. pervasive transcription. This phenomenon was first observed using tiling arrays[61] and later by various transcriptome profiling methods[145]. However, whether pervasive transcription consists mainly of background transcriptional noise or is of functional importance has been a matter of debate. The low fidelity of RNA polymerase II[146] and promiscuous binding of transcription factors[147] may give rise to "leaky" transcription. Furthermore, the existence of various systems that suppress pervasive transcription[148] also raises the question of whether these products of pervasive transcription are functional.

Despite questions over their functionality, pervasive transcription is known to produce various classes of non-coding RNAs (ncRNAs)with characteristic features. These include various classes of small RNAs that are associated to promoters[149], long non-coding RNAs (lncRNAs)[150], and long intergenic non-coding RNAs (lincRNA)[151]. Evidence of context-specific transcription, e.g. transcription at a specific developmental stage, tissue type, or against a particular stimulus, has been observed in these classes of ncRNA. In chapter 3, we profiled the small RNA population of cells that were induced for DNA damage, as a role for small RNAs in the DNA damage response (DDR) had been previously observed in *Neurospora crassa*[142]. Our work and the work of an independent group, demonstrated that small RNAs would form in the vicinity of the double strand break[2, 80]. Furthermore, these RNAs, which we named as DDRNAs, had specific size and nucleotide profiles that differed from other classes of small RNAs. While it is unclear what role these DDRNAs serve in the DDR, knock-down of Dicer and Drosha[2] and mutations in *Ago2*[80] impaired the DNA repair efficiency. As a signalling cascade is initiated from the site of DNA damage[152], it was hypothesised that AGO2, which are known to bind to different classes of small RNAs, recruits the DNA damage repair complex[80].

The guiding and recruitment of various complexes to specific genomic loci is becoming an emerging theme when discussing the potential role of various ncRNAs. Apart from the classical example of miRNAs being associated with the RNA-induced silencing complex, many lincRNAs were shown to be bound to chromatin-modifying complexes and may function to guide these complexes to regions that need to undergo epigenetic regulation[153, 154]. Intriguingly, many enzymatic members of chromatin remodelling complexes do not contain DNA binding domains but possess RNA binding domains. For example, *Xist* has been shown to recruit chromatin silencing proteins within the Polycomb complex to induce gene silencing via histone methylation[155]. Other ncRNAs including *RepA*, *Air*, *Kcnq1ot1*, and *Hotair* have also been shown to be associated with Polycomb and direct the complex to specific loci[153]. Another mode by which ncRNAs can regulate chromatin is through the act of transcription, whereby RNA Pol II activity causes nucleosome rearrangements and local chromatin remodelling[156]. Thus while the transcriptional product may be inert, it serves its purpose by maintaining a open chromatin conformation. An extreme case of this mode of function is a study that suggests that non-coding transcription can send "ripples of transcription", which causes remodelling of the expression landscape and thereby influences the expression of neighbouring loci[157].

One of the major points that argues against ascribing function to products of pervasive transcription is because the human genome is largely occupied by repetitive DNA sequences, made up of mainly transposable elements (TEs). These elements have been historically considered as simply selfish products that have propagated themselves very successfully[94, 95]. However, it has been demonstrated that transcripts derived from TEs are able to regulate chromatin structure in centromeres and neocentromeres[158]. In a recent study, it was demonstrated that interspersed repeat sequences, such as LINE-1, are associated with euchromatin and loss of these repeat sequences caused aberrant chromatin distribution and condensation[159]. Furthermore, a large number of transcriptional events was shown to be initiated within TEs and served as alternative promoters[114]. In addition, TEs have been observed to be part of many lncRNAs[160], which led to the Repeat Insertion Domains of LncRNAs (RIDLs) hypothesis[161], whereby TEs act in a manner similar to structural domains in proteins. Furthermore, it has been suggested that TEs have contributed to the origin of many lncRNAs[111, 162].

In chapter 4, we profiled the expression of TEs in a large panel of samples (988 libraries), given the many potential functional roles of TEs. As previously reported, we observed that on average, expression from TEs are more tissue-specific and lowly expressed compared with protein-coding transcripts. The relatively low expression strength of pervasively transcribed products has been used to support the claim that these products are the consequence of degradation or background noise. However, it should be pointed out that unlike mRNAs, which require a longer half-life in order to exported into the cytoplasm for translation, ncRNA exert their function immediately in the nucleus. Many transcriptome profiling methods do not enrich for nuclear fraction, where many transcripts arising from TEs exist[163]. Furthermore, in the case that ncRNA are expressed in small quantities, like transcription factors, they are still able to trigger an amplified cascade of downstream events. However, despite their lower expression patterns, we were able to cluster biologically similar libraries together using the expression profiles from TEs. We also showed that transcriptional

events initiating within TEs are overlap small RNAs, enhancers, and lincRNAs more often than chance. These lines of evidence support the notion that some TEs have become exapted in the human genome.

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 friends and colleagues. 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 and it has been the best experience of my life (thus far). 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 in Japan. Their answers were unanimous; clearly, they had had enough of me. I've made many new friends in Japan and I'd like to think that they know who they are. Thanks guys and gals for enduring me! I would also like to acknowledge the Good Samaritans who answer questions on discussion forums and mailing lists; where would I be without them.

Bibliography

[1] 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.

[2] 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.

[3] Dave Tang and Piero Carninci. The regulated expression of repetitive elements across human cell types and tissues. *In preparation.*

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

[5] Yuki Hasegawa, Dave Tang, Naoko Takahashi, Yoshihide Hayashizaki, Alistair R. R. Forrest, the FANTOM consortium, and Harukazu Suzuki. Ccl2 enhances pluripotency of human induced pluripotent stem cells by activating hypoxia related genes. *Sci. Rep.*, 4, Jun 2014.

[6] Dave Tang, Ana Maria Suzuki, Raffaella Calligaris, Stefano Gustincich, and Piero Carninci. Deep transcriptome sequencing of whole blood samples from Parkinson's disease patients. *In preparation.*

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

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

[9] 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.

[10] 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.

[11] E. Chargaff, R. Lipshitz, and C. Green. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *J. Biol. Chem.*, 195(1):155–160, Mar 1952.

[12] D. ELSON and E. CHARGAFF. On the desoxyribonucleic acid content of sea urchin gametes. *Experientia*, 8(4):143–145, Apr 1952.

[13] 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.

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

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

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

[17] Francis Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, Aug 1970.

[18] The Crick Papers: 1955 From DNA to protein. `http://genome.wellcome.ac.uk/doc_WTD022319.html`.

[19] M. B. Hoagland, M. L. Stephenson, J. F. Scott, L. I. Hecht, and P. C. Zamecnik. A soluble ribonucleic acid intermediate in protein synthesis. *J. Biol. Chem.*, 231(1):241–257, Mar 1958.

[20] S. BRENNER, F. JACOB, and M. MESELSON. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190(4776):576–581, May 1961.

[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] T. I. Lee and R. A. Young. Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, 34:77–137, 2000.

[24] 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.

[25] M. D. Young, T. A. Willson, M. J. Wakefield, E. Trounson, D. J. Hilton, M. E. Blewitt, A. Oshlack, and I. J. Majewski. ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res.*, 39(17):7415–7427, Sep 2011.

[26] T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293(5532):1074–1080, Aug 2001.

[27] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[28] 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.

[29] 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.

[30] 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.

[31] M. L. Metzker. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11(1):31–46, Jan 2010.

[32] D. Dressman, H. Yan, G. Traverso, K. W. Kinzler, and B. Vogelstein. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U.S.A.*, 100(15):8817–8822, Jul 2003.

[33] M. Fedurco, A. Romieu, S. Williams, I. Lawrence, and G. Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.*, 34(3):e22, 2006.

[34] E. E. Schadt, S. Turner, and A. Kasarskis. A window into third-generation sequencing. *Hum. Mol. Genet.*, 19(R2):R227–240, Oct 2010.

[35] I. Braslavsky, B. Hebert, E. Kartalov, and S. R. Quake. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U.S.A.*, 100(7):3960–3964, Apr 2003.

[36] 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.

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

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

[39] J. C. Alwine, D. J. Kemp, and G. R. Stark. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5350–5354, Dec 1977.

[40] K. J. Livak and T. D. Schmittgen. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*, 25(4):402–408, Dec 2001.

[41] 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.

[42] 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.

[43] M. J. Okoniewski and C. J. Miller. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 7:276, 2006.

[44] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, Oct 1995.

[45] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, 18(6):630–634, Jun 2000.

[46] P. Ng, C. L. Wei, W. K. Sung, K. P. Chiu, L. Lipovich, C. C. Ang, S. Gupta, A. Shahab, A. Ridwan, C. H. Wong, E. T. Liu, and Y. Ruan. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods*, 2(2):105–111, Feb 2005.

[47] T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.*, 100(26):15776–15781, Dec 2003.

[48] 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.

[49] P. Carninci, A. Westover, Y. Nishiyama, T. Ohsumi, M. Itoh, S. Nagaoka, N. Sasaki, Y. Okazaki, M. Muramatsu, C. Schneider, and Y. Hayashizaki. High efficiency selection of full-length cDNA by improved biotinylated cap trapper. *DNA Res.*, 4(1):61–66, Feb 1997.

[50] H. Matsumura, S. Reich, A. Ito, H. Saitoh, S. Kamoun, P. Winter, G. Kahl, M. Reuter, D. H. Kruger, and R. Terauchi. Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc. Natl. Acad. Sci. U.S.A.*, 100(26):15718–15723, Dec 2003.

[51] Hazuki Takahashi, Timo Lassmann, Mitsuyoshi Murata, and Piero Carninci. 5 end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nature Protocols*, 7(3):542–561, Feb 2012.

[52] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, Jan 2009.

[53] H. Kawaji, M. Lizio, M. Itoh, M. Kanamori-Katayama, A. Kaiho, H. Nishiyori-Sueki, J. W. Shin, M. Kojima-Ishiyama, M. Kawano, M. Murata, N. Ninomiya-Fukuda, S. Ishikawa-Kato, S. Nagao-Sato, S. Noma, Y. Hayashizaki, A. R. Forrest, and P. Carninci. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res.*, 24(4):708–717, Apr 2014.

[54] J. Kawai, A. Shinagawa, K. Shibata, M. Yoshino, M. Itoh, et al. Functional annotation of a full-length mouse cDNA collection. *Nature*, 409(6821):685–690, Feb 2001.

[55] Y. Okazaki, M. Furuno, T. Kasukawa, J. Adachi, H. Bono, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 420(6915):563–573, Dec 2002.

[56] S. Katayama, Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, et al. Antisense transcription in the mammalian transcriptome. *Science*, 309(5740):1564–1566, Sep 2005.

[57] 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.

[58] P. Bertone, V. Stolc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306(5705):2242–2246, Dec 2004.

[59] P. Kapranov, S. E. Cawley, J. Drenkow, S. Bekiranov, R. L. Strausberg, S. P. Fodor, and T. R. Gingeras. Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, 296(5569):916–919, May 2002.

[60] J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308(5725):1149–1154, May 2005.

[61] P. Kapranov, J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316(5830):1484–1488, Jun 2007.

[62] H. van Bakel, C. Nislow, B. J. Blencowe, and T. R. Hughes. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.*, 8(5):e1000371, May 2010.

[63] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, et al. Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, Jun 2007.

[64] G. W. Beadle and E. L. Tatum. Genetic control of biochemical reactions in neurospora. *Proceedings of the National Academy of Sciences*, 27(11):499–506, 1941.

[65] L. T. Chow, J. M. Roberts, J. B. Lewis, and T. R. Broker. A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. *Cell*, 11(4):819–836, Aug 1977.

[66] A. J. Berk and P. A. Sharp. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell*, 12(3):721–732, Nov 1977.

[67] 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.

[68] 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.

[69] A. A. Aravin, N. M. Naumova, A. V. Tulin, V. V. Vagin, Y. M. Rozovsky, and V. A. Gvozdev. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Curr. Biol.*, 11(13):1017–1027, Jul 2001.

[70] C. P. Ponting, P. L. Oliver, and W. Reik. Evolution and functions of long noncoding RNAs. *Cell*, 136(4):629–641, Feb 2009.

[71] J. Wang, J. Zhang, H. Zheng, J. Li, D. Liu, H. Li, R. Samudrala, J. Yu, and G. K. Wong. Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature*, 431(7010):1 p following 757; discussion following 757, Oct 2004.

[72] A. Huttenhofer, P. Schattner, and N. Polacek. Non-coding RNAs: hope or hype? *Trends Genet.*, 21(5):289–297, May 2005.

[73] J. T. Kung, D. Colognori, and J. T. Lee. Long noncoding RNAs: past, present, and future. *Genetics*, 193(3):651–669, Mar 2013.

[74] P. Preker, J. Nielsen, S. Kammler, S. Lykke-Andersen, M. S. Christensen, C. K. Mapendano, M. H. Schierup, and T. H. Jensen. RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, 322(5909):1851–1854, Dec 2008.

[75] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, 25(18):1915–1927, Sep 2011.

[76] I. Ulitsky, A. Shkumatava, C. H. Jan, H. Sive, and D. P. Bartel. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, 147(7):1537–1550, Dec 2011.

[77] N. L. Kedersha and L. H. Rome. Isolation and characterization of a novel ribonucleoprotein particle: large structures contain a single species of small RNA. *J. Cell Biol.*, 103(3):699–709, Sep 1986.

[78] E. Valen, P. Preker, P. R. Andersen, X. Zhao, Y. Chen, C. Ender, A. Dueck, G. Meister, A. Sandelin, and T. H. Jensen. Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nat. Struct. Mol. Biol.*, 18(9):1075–1082, Sep 2011.

[79] T. K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187, May 2010.

[80] W. Wei, Z. Ba, M. Gao, Y. Wu, Y. Ma, S. Amiard, C. I. White, J. M. Rendtlew Danielsen, Y. G. Yang, and Y. Qi. A role for small RNAs in DNA double-strand break repair. *Cell*, 149(1):101–112, Mar 2012.

[81] Y. Tay, J. Rinn, and P. P. Pandolfi. The multilayered complexity of ceRNA crosstalk and competition. *Nature*, 505(7483):344–352, Jan 2014.

[82] R. J. Taft, E. A. Glazov, N. Cloonan, C. Simons, S. Stephen, et al. Tiny RNAs associated with transcription start sites in animals. *Nat. Genet.*, 41(5):572–578, May 2009.

[83] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107, Sep 2010.

[84] J. G. Underwood, A. V. Uzilov, S. Katzman, C. S. Onodera, J. E. Mainzer, D. H. Mathews, T. M. Lowe, S. R. Salama, and D. Haussler. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, 7(12):995–1001, Dec 2010.

[85] R. J. Britten and D. E. Kohne. Repeated sequences in dna. *Science*, 161(3841):529–540, Aug 1968.

[86] R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, Dec 2002.

[87] 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.

[88] J. Jurka. Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.*, 8(3):333–337, Jun 1998.

[89] Casey M. Bergman and Hadi Quesneville. Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, 8(6):382–392, 2007.

[90] M. Tarailo-Graovac and N. Chen. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*, Chapter 4:Unit 4.10, Mar 2009.

[91] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110(1-4):462–467, 2005.

[92] T. J. Wheeler, J. Clements, S. R. Eddy, R. Hubley, T. A. Jones, J. Jurka, A. F. Smit, and R. D. Finn. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.*, 41(Database issue):70–82, Jan 2013.

[93] A. P. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.*, 7(12):e1002384, Dec 2011.

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

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

[96] S. Ohno. So much "junk" DNA in our genome. *Brookhaven Symp. Biol.*, 23:366–370, 1972.

[97] A. F. Palazzo and T. R. Gregory. The case for junk DNA. *PLoS Genet.*, 10(5):e1004351, May 2014.

[98] Elizabeth Pennisi. Encode project writes eulogy for junk dna. *Science*, 337(6099):1159–1161, 2012.

[99] D. Graur, Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol*, 5(3):578–590, 2013.

[100] W. F. Doolittle. Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U.S.A.*, 110(14):5294–5300, Apr 2013.

[101] Sean R. Eddy. The c-value paradox, junk dna and encode. *Current Biology*, 22(21):R898–R899, Nov 2012.

[102] T. R. Gregory. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol Rev Camb Philos Soc*, 76(1):65–101, Feb 2001.

[103] M. G. Kidwell. Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115(1):49–63, May 2002.

[104] S. Aparicio, J. Chapman, E. Stupka, N. Putnam, J. M. Chia, et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, 297(5585):1301–1310, Aug 2002.

[105] E. Ibarra-Laclette, E. Lyons, G. Hernandez-Guzman, C. A. Perez-Torres, L. Carretero-Paulet, et al. Architecture and evolution of a minute plant genome. *Nature*, 498(7452):94–98, Jun 2013.

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

[107] R. Cordaux and M. A. Batzer. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.*, 10(10):691–703, Oct 2009.

[108] R. E. Mills, E. A. Bennett, R. C. Iskow, and S. E. Devine. Which transposable elements are active in the human genome? *Trends Genet.*, 23(4):183–191, Apr 2007.

[109] C. Feschotte. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.*, 9(5):397–405, May 2008.

[110] G. Lev-Maor, O. Ram, E. Kim, N. Sela, A. Goren, E. Y. Levanon, and G. Ast. Intronic Alus influence alternative splicing. *PLoS Genet.*, 4(9):e1000204, 2008.

[111] A. Kapusta, Z. Kronenberg, V. J. Lynch, X. Zhuo, L. Ramsay, G. Bourque, M. Yandell, and C. Feschotte. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.*, 9(4):e1003470, Apr 2013.

[112] C. J. Cohen, W. M. Lock, and D. L. Mager. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene*, 448(2):105–114, Dec 2009.

[113] Stephen Jay Gould and Elisabeth S. Vrba. Exaptation; a missing term in the science of form. *Paleobiology*, 8(1):4–15, 1982.

[114] 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.

[115] G. J. Faulkner, A. R. Forrest, A. M. Chalk, K. Schroder, Y. Hayashizaki, P. Carninci, D. A. Hume, and S. M. Grimmond. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, 91(3):281–288, Mar 2008.

[116] M. Xie, C. Hong, B. Zhang, R. F. Lowdon, X. Xing, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat. Genet.*, 45(7):836–841, Jul 2013.

[117] D. S. Day, L. J. Luquette, P. J. Park, and P. V. Kharchenko. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol.*, 11(6):R69, 2010.

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

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

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

[121] 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.

[122] T. Lassmann, Y. Hayashizaki, and C. O. Daub. TagDust–a program to eliminate artifacts from next generation sequencing data. *Bioinformatics*, 25(21):2839–2840, Nov 2009.

[123] T. Lassmann, Y. Hayashizaki, and C. O. Daub. SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics*, 27(1):130–131, Jan 2011.

[124] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct 1990.

[125] W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome Res.*, 12(4):656–664, Apr 2002.

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

[127] C. Trapnell and S. L. Salzberg. How to map billions of short reads onto genomes. *Nat. Biotechnol.*, 27(5):455–457, May 2009.

[128] 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.

[129] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[130] M. Hsi-Yang Fritz, R. Leinonen, G. Cochrane, and E. Birney. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.*, 21(5):734–740, May 2011.

[131] The BED format. `http://genome.ucsc.edu/FAQ/FAQformat.html#format1`.

[132] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, Haussler, and David. The human genome browser at ucsc. *Genome Research*, 12(6):996–1006, 2002.

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

[134] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003.

[135] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, 11(3):R25, 2010.

[136] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18(9):1509–1517, Sep 2008.

[137] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[138] 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.

[139] 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.

[140] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95(25):14863–14868, Dec 1998.

[141] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1995.

[142] H. C. Lee, S. S. Chang, S. Choudhary, A. P. Aalto, M. Maiti, D. H. Bamford, and Y. Liu. qiRNA is a new type of small interfering RNA induced by DNA damage. *Nature*, 459(7244):274–277, May 2009.

[143] FastQC. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

[144] Making a Genesweep: It's Official! `http://www.bio-itworld.com/archive/071503/genesweep`.

[145] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, Sep 2012.

[146] K. Struhl. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.*, 14(2):103–105, Feb 2007.

[147] F. Spitz and E. E. Furlong. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, 13(9):613–626, Sep 2012.

[148] T. H. Jensen, A. Jacquier, and D. Libri. Dealing with pervasive transcription. *Mol. Cell*, 52(4):473–484, Nov 2013.

[149] A. Jacquier. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.*, 10(12):833–844, Dec 2009.

[150] L. Yang, J. E. Froberg, and J. T. Lee. Long noncoding RNAs: fresh perspectives into the RNA world. *Trends Biochem. Sci.*, 39(1):35–43, Jan 2014.

[151] M. J. Hangauer, I. W. Vaughn, and M. T. McManus. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.*, 9(6):e1003569, Jun 2013.

[152] S. P. Jackson and J. Bartek. The DNA-damage response in human biology and disease. *Nature*, 461(7267):1071–1078, Oct 2009.

[153] A. Saxena and P. Carninci. Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. *Bioessays*, 33(11):830–839, Nov 2011.

[154] A. M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 106(28):11667–11672, Jul 2009.

[155] A. Wutz and J. Gribnau. X inactivation Xplained. *Curr. Opin. Genet. Dev.*, 17(5):387–393, Oct 2007.

[156] L. Chakalova, E. Debrand, J. A. Mitchell, C. S. Osborne, and P. Fraser. Replication and transcription: shaping the landscape of the genome. *Nat. Rev. Genet.*, 6(9):669–677, Sep 2005.

[157] M. Ebisuya, T. Yamamoto, M. Nakajima, and E. Nishida. Ripples from neighbouring transcription. *Nat. Cell Biol.*, 10(9):1106–1113, Sep 2008.

[158] A. C. Chueh, E. L. Northrop, K. H. Brettingham-Moore, K. H. Choo, and L. H. Wong. LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. *PLoS Genet.*, 5(1):e1000354, Jan 2009.

[159] L. L. Hall, D. M. Carone, A. V. Gomez, H. J. Kolpa, M. Byron, N. Mehta, F. O. Fackelmayer, and J. B. Lawrence. Stable C0T-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes. *Cell*, 156(5):907–919, Feb 2014.

[160] D. Kelley and J. Rinn. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.*, 13(11):R107, 2012.

[161] R. Johnson and R. Guigo. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA*, 20(7):959–976, Jul 2014.

[162] A. Kapusta and C. Feschotte. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet.*, 30(10):439–452, Oct 2014.

[163] A. Fort, K. Hashimoto, D. Yamada, M. Salimullah, C. A. Keya, et al. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.*, 46(6):558–566, Jun 2014.