Article

Single cell transcriptomes: how low can you go?

Chris Harris

Centre for Translational Cancer Research, Department of Biochemistry, University of Otago, PO Box 56, Dunedin 9054 (email: *harch909@student.otago.ac.nz*)

Keywords: *single cell*, *transcriptome*, *PCR*, *linear amplification*

Chris graduated with an LLB and BSc(Hons) in 2013 from the University of Otago. He is currently completing an MSc at Otago focussed on studying single cancer cells with the view to improving diagnostic tests. Chris hopes to continue his studies to improve outcomes for cancer patients.

Introduction

Biologists have been studying the cell since the discovery of this indivisible building block of life in 1665 by Robert Hooke. They are the basic units of structure and reproduction. Their importance is self-evident; however, the technology to study life at this indivisible level in great detail has only emerged recently.

The morphology of single cells has been well characterised, but their inner workings less so. The important instructions for the cell to make its products (e.g. proteins) are made of DNA and are stored in the nucleus. The DNA instructions (genome) cannot be moved outside the nucleus to the cellular factories, so copies of the instructions need to be made for delivery. These copies are made from a molecule called messenger RNA (mRNA). All the mRNA in a cell at any one time is called the 'transcriptome'. Studying the transcriptome, rather than the genome, will enable us to determine what the cell is actively producing and potentially in what quantities. For example, cancer happens because of mutations in DNA, but not all DNA mutations cause cancer. Studying the transcriptome will allow us to determine the functional impact of DNA mutations.

In the last four years, the advent of next-generation sequencing (e.g. RNA-seq) combined with whole transcriptome amplification methods have enabled us to study massive quantitative data sets. These methods have been largely applied to whole tissues rather than individual cells. When looking at the characteristics of any kind of population – people or cells – studying them as a whole, as opposed to individuals, can leave out important information. For example, a rugby team might score 20 points, but this gives no indication that one player scores 80% of those points.

Similarly, studying cells at an individual level can reveal clinically important extremes. Tumours often begin from a single initiator cell and grow to become a heterogenous mass of cells with different sub-populations that often co-operate to evade therapies and our natural defences.[1-3] There may be a small number of cells within a tumour that are integral to driving uncontrolled growth, (e.g. cancer stem cells). We need to study these cell types at a single cell transcriptome level to understand what they are doing and how to treat them.

In the last two years, efficient methods for studying individual cells at the transcriptome level have been developed. The importance of studying single cell transcriptomes to further our understanding of cancer and biology in general warrants greater understanding and improvement of these methods.

The challenge of single cell transcriptome amplification

It has been demonstrated that the average gene expression across a population of cells is significantly biased by a small sub-population of cells with high expression.[4] Specific transcripts can vary as much as 1000 fold between presumably equivalent single cells.

A single cell contains about 1-50 pg of RNA.[5] Diluting this amount of RNA in 10 µl of water (the volume used in some single cell reactions) is roughly the equivalent of sprinkling less than a pinch of salt in an Olympic-size swimming pool. Furthermore, only 1-5% of the total RNA is mRNA (0.01-2.5 pg per cell). It is not yet possible to sequence mRNA directly from a single cell. The mRNA needs to be amplified approximately 10^7 fold to be analysed with RNA-seq or microarray technology. Two broad amplification strategies have been described:

- Exponential amplification through a **reverse transcription (RT) - polymerase chain reaction (PCR)** based method
- Linear amplification through **RT - *in vitro* transcription (IVT)** or **phi29** based methods.

The first PCR method was described by Norman Iscove and the first IVT method was described by James Eberwine, both around 20 years ago.[6,7] Since that time, several variations of these strategies have been developed. Most of them are able to detect the expression of thousands of genes in single cells. They all attempt to optimise four elements for effective single cell transcriptome analysis:

1. High sensitivity (i.e. amplification of very limited) potentially very degraded, starting RNA amounts. This is particularly relevant for clinical samples.
2. Reproducibility between cells of the same type.
3. Coverage of all mRNA fragments, including rare species.

4. Fidelity of amplification (i.e. the ability to preserve the relative abundance and content of the starting mRNA population).

Many single-cell studies focus on reproducibility, whereas fewer consider fidelity. This is surprising, given the entire purpose of the amplification exercise is to determine the quantity of each mRNA species the cell is in fact producing. A couple of studies have compared amplified mRNA to unamplified mRNA to evaluate the fidelity of single cell methods. One evaluated both a PCR and IVT method[8] and the other focused on IVT.[9] Although the amplified and unamplified mRNA may be from the same source, determining fidelity from this comparison is significantly limited by sampling error. The challenge is to develop a method to determine what mRNA is in fact amplified.

Islam *et al.* determined the absolute quantity of original mRNA transcripts by tagging each mRNA molecule with a unique molecular identifier (random 5bp sequence) before amplification.[10] The number of original molecules for each mRNA species should correspond to the number of unique molecular identifiers present in the amplified population for each species. For example, if only two kinds of unique molecular identifier are present among 10,000 amplified copies of a specific mRNA molecule, that means there were only two copies of the mRNA molecule to begin with.

Amplifying the single cell transcriptome by PCR

The first step of all current amplification methods involves converting the mature mRNA into complementary DNA (cDNA) by reverse transcriptase (RTase). The RTase is primed from a synthesised poly-thymine oligonucleotide (oligo dT) primer annealed to the poly-adenine (poly A) tail of the mRNA (Fig. 1). This synthesised oligonucleotide contains the forward PCR primer. The reverse primer can be attached to the 5' end of the cDNA in a variety of ways.

A procedure was developed by Chenchik *et al.* (1998) to add a reverse PCR primer to the 5' end of cDNA with the help of the 'template-switch' (TS) effect.[11] This method takes advantage of the ability of Moloney Murine Leukemia Virus (MMLV) reverse transcriptase to add a few non-templated cytosines to the 3' end of the newly synthesised cDNA strand. This usually occurs when the reverse transcriptase reaches the capped 5' end of the mRNA. An oligonucleotide containing the reverse PCR primer sequence with three guanines on its 3' end is added to the reaction (the template switch oligonucleotide).

Several laboratories have taken advantage of the MMLV RT enzyme 'template switch' phenomenon, while adding their own modifications. One of the simpler methods is SMART-seq 2, developed by Picelli *et al.* (2014).[12] A similar method called single-cell tagged reverse transcription (STRT) was developed previously by Islam *et al.* (2011), but notably is able to analyse dozens of single cells in a single PCR.[10] This is possible because Islam *et al.* tag a unique barcode to the template switch oligonucleotide for each sample, enabling the samples to be distinguished in downstream analysis. Multiplexing will be a necessary feature in methods designed to elucidate the nature of heterogeneity of populations of cells, requiring the analysis of thousands of cells.

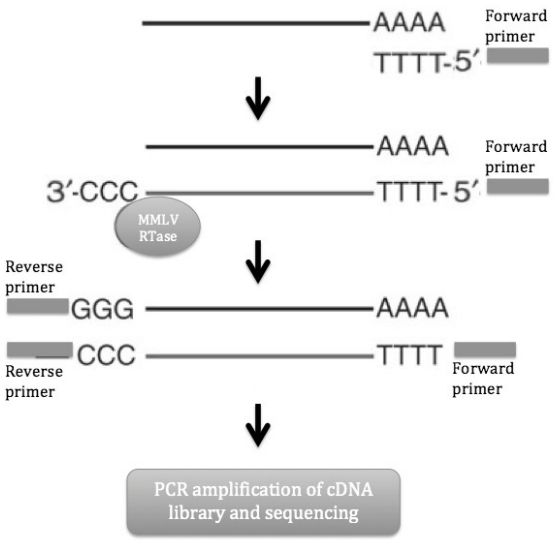

Fig. 1. The addition of three cytosines to the 5' end of the cDNA by the RTase enables an oligonucleotide with the reverse primer attached to three guanosines to base pair with the cDNA.

PCR by poly A tailing methods

An alternative to the template switching approach is the ligation of a poly A tail to the 3' end of the newly synthesised cDNA template by a terminal deoxynucleotidyltransferase (TdT).[13] An oligo dT with the reverse primer sequence is then able to base pair to this poly A tail attached to the 3' cDNA end. This approach is more flexible than the MMLV capped switch approach given that it does not require the cDNA template to extend to the capped 5' end of the RNA. This is potentially useful for degraded RNA samples that have lost a large proportion of capped ends. However, it requires more reaction steps compared to template switching methods, potentially increasing the level of technical noise. Poly A tailing techniques such as Quartz-seq by Sasagawa *et al.* demonstrate impressive reproducibility.[14] Quartz-seq produced a Pearson correlation constant (PCC) of 0.93 when performed on triplicate diluted RNA samples (10 pg) from mouse embryonic stem cells. It remains to be seen how this method directly compares to SMART-seq 2.

PCR problems and proposed solutions

According to a prevalent hypothesis in the literature, as the number of PCR cycles increase, the fidelity decreases as a result of the exponential increase in stochastic amplification and differing efficiencies for various template lengths and abundances.[14] Iscove *et al.* (2002) tested this hypothesis by comparing the amplified cDNA from an IVT linear amplification method and poly A tailing exponential amplification with unamplified RNA sample (10 µg).[15] Contrary to expectations, they found that the exponential amplification better preserved the relative abundance of mRNA transcripts than a single round of linear amplification. Further investigation is required to confirm this finding.

The TS approach has significant intrinsic problems relating to the production of background templates. This background 'noise' can consist of amplification of small artefacts produced by interruption of first strand synthesis and concatenation of TS oligonucleotides.[16] First strand interruption can be caused by the TS oligo hybridising to the first strand cDNA by sequence complementarity, before it has reached the 5' cap. Plessy *et al.* (2013) discourage first strand interruption by adding a 6- nucleotide long spacer between the barcode and the ribo-guanosines in the TS oligonucleotide.[17] Concatenation occurs when the MMLV RTase enzyme continues to add cytosines to the end of the complementary strand to the TS oligo. This results in a continuous string of linked TS oligos. Kapteyn *et al.* (2010) resolve concatenation simply by adding blocking isomeric bases to the 5' end of the TS oligo, preventing any extension from that end.[18]

A TS approach combining elements of both MMLV RTase enzyme cytosine addition and poly A tailing was devised by Schmidt and Mueller (1999).[19] The poly A tail is ligated to the non-templated cytosines (Fig. 2). A TS oligonucleotide with a 3' string of ribo-thymines as well as ribo-guanines is required to base pair to the cDNA 3' overhang. This is designed to make it more difficult for the TS oligo to prematurely base pair and cause first strand interruption. Another approach incorporates a 'locked nucleic acid' (LNA) on the ribo-guanine end of the TS oligo.[20] Picelli reported that this doubled the cDNA yield compared to the conventional ribo-guanine ends when the SMART-seq 2 protocol was used. It was speculated that LNA increases the efficiency of TS oligo binding to the cytosine cap, probably because LNA:DNA base pairs have greater thermal stability.

Template switch by poly A tailing also has intrinsic byproduct issues. The TdT modifies the RT primer to produce <200 bp fragments.[13] Sasagawa *et al.* completely eliminated these byproducts by a combination of minimal oligo dT concentration, exonuclease I treatment, restricted poly A tailing and an optimised suppression PCR, all in a single tube reaction.[14] Interestingly, Sasagawa *et al.* found that topoisomerase V treatment alone could suppress byproduct synthesis. Topoisomerase V relaxes supercoiled DNA by cleaving and rotating a single strand in a double helix, and then ligating it back together. It is not known how it reduces byproduct synthesis in this case.

PCR without template switching or poly A tailing

While template switching remains an attractive option for PCR, whether by 'template switching' or poly A tailing, there are notable alternatives where a TS oligo is unnecessary.

Designed-primer seq (DP-seq) involves a PCR with a suite of specific primers that are designed to amplify the majority of the transcriptome.[21] Bhargava *et al.* (2013) developed an algorithm to identify 44 primer sequences that preferentially amplified unique transcripts of a reference mouse transcriptome, while minimising mis-priming and primer dimerisation. This enabled an RT reaction using the RT primer only, but required a more complicated PCR process. The PCR consisted of a 37°C Klenow polymerase incubation step, followed by a PCR step with a 72°C extension, in order to amplify partially hybridised primers.

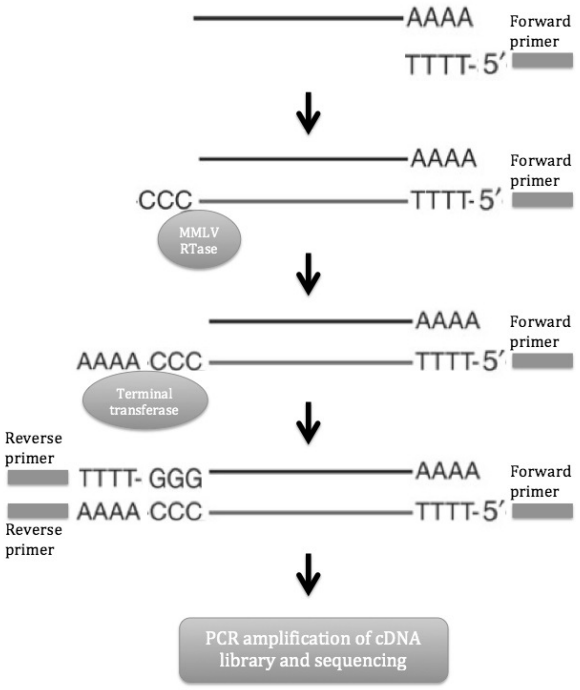

Fig. 2. A terminal transferase, or TdT, adds a poly A tail to the cytosines already added by the MMLV RTase. This combination increases the binding specificity by the complementary bases tagged to a reverse primer for PCR.

Bhargava *et al.* were able to amplify more than 80% of the transcriptome from 50 pg of mouse embryonic stem cell (mESC) RNA. Amplifying samples smaller than 50 pg resulted in a substantial increase of technical noise from PCR artefacts and significant deviations in the observed expression of low copy genes. The dominant bias was due to primers being blocked by cDNA secondary structure. This was despite efforts to predict secondary structure using UNAfold software. The designed primer method is better able to quantify rare transcripts than random primer methods and it has a high technical quantitative reproducibility (R^2=0.86 from 50 pg RNA samples).

Amplifying the single cell transcriptome by linear methods

The second broad category of amplifying cDNA transcribed from mRNA is by linear methods. Linear amplification methods tend to only amplify a few hundred base pairs of template. This has several theoretical advantages over exponential amplification by PCR. PCR amplification is most efficient over small regions of a few hundred base pairs, so that longer cDNA strands are not amplified to the same degree as shorter strands. The differential PCR amplification of a 1 kb cDNA fragment compared to a 6kb fragment can be as much as 1300 fold after only 25 cycles.[22] Also, very small sub-populations of RNA can miss being amplified in PCR, and those that are amplified can become over-represented exponentially.[23,24] Therefore, in theory, linear amplification methods are better at preserving the relative abundances of mRNA species in the original sample.

Linear amplification methods for single cells have been

largely based on the T7 RNA polymerase (T7) and the phi29 DNA polymerase (phi29), both derived from bacteriophages. The T7 catalyses the formation of first strand anti-sense RNA (aRNA) from template cDNA (known as *in vitro* transcription or IVT). T7 can amplify RNA up to 1000 fold in one round of amplification.[25] It has the advantage of being extremely specific to a T7 promoter sequence. Limitations include the low processivity of T7 (only synthesises up to 1500bp of RNA per binding event,[26] though this isn't necessarily relevant for quantification) and a requirement for double stranded promoter sequence. This method is also unsuitable for the sense strand probes of spotted oligo microarrays because the aRNA needs to be converted back to sense strand cDNA.[27]

T7 RNA polymerase methods

CEL-seq (Cell Expression by Linear amplification and Sequencing) is a T7 polymerase method developed by Hashimshony *et al.* (2012).[28] IVT on single cells has an inherent limitation in that at least 400 pg of starting RNA is required. Hashimshony *et al.* remove this problem by pooling samples that have been barcoded with a unique sequence in the oligo dT after the RT, while also providing a multiplex capability similar to the PCR STRT method. Hashimshony's group compared the data gathered from their method with the data from Islam's PCR method, and found their linear method was more reproducible, amplified more genes than STRT, and was better able to distinguish gene expression levels between different cell types.

The Chum-RNA method is another T7 IVT method developed by Tougan *et al.* (2008) that uses a different approach to overcome the 400 pg hurdle.[29] Key enzymes involved in the IVT reaction, such as RTase, T7 RNA polymerase, DNA polymerase I and DNA ligase, require at least 1 µM of substrate to work efficiently. This is about one million times more than what is available from a single cell. Tougan *et al.* resolve this by adding dummy or 'Chum' RNA (41bp strands including a poly A tail) to the single cell amount of RNA. This allows the enzymes to be in contact with sufficient substrate, enabling efficient enzyme reaction. 'Chum' refers to the 'friendly' action of the RNA.[29] The whole RNA population is converted to cDNA and amplified efficiently, including the single cell RNA. The dummy cDNA population is then selectively removed by column chromatography, leaving the population of interest.

The Chum-RNA method allows the amplification of down to 0.49 femtograms of mRNA (or 730 molecules of RNA), corresponding to a sub-population of mRNA in a single cell. This is the lowest amount of RNA that has been amplified by any single cell method, demonstrating the Chum-RNA's impressive sensitivity. Fluorescent labeling and microarray analysis showed that this method did not introduce any significant bias. The Pearson correlation of gene expression between a single cell and 1 million cells was 0.98, potentially indicating strong fidelity.

phi29 DNA polymerase methods

In contrast to T7 polymerase, phi29 polymerase synthesises DNA from cDNA and is characterised by the highest processivity among known DNA polymerases (70kb insertions per binding event).[30] It is also characterised by its ability to amplify DNA by multiple strand displacement. Multiple strand displacement begins with random primers annealing to the template to initiate polymerisation by phi29. When phi29 reaches the next primer on the strand, it causes the newly formed double strand to become single stranded at that point, i.e. it becomes displaced (Fig. 3). This allows further primer annealing to occur on the displaced, newly synthesised, strand. The process repeats on the newly synthesised strand, resulting in a hyper-branched network of polymerisation and a high yield of product. The Klenow polymerase used in DP-seq also has strand displacement activity.

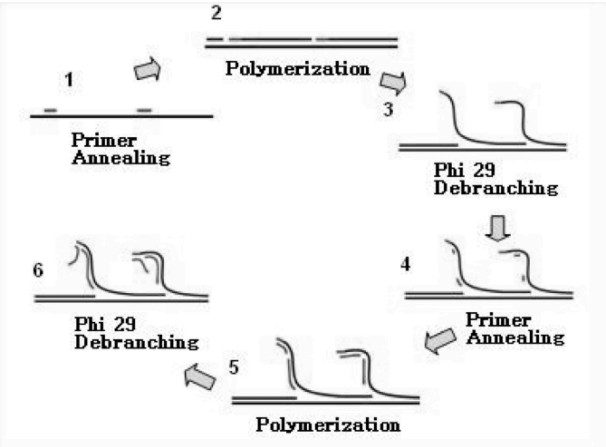

Fig. 3. The phi29 DNA polymerase elongates DNA from the primer annealed to the cDNA template strand (steps 1 and 2). Once the polymerase reaches an adjacent primer, the new strand is partially displaced (debranched) from the template (step 3). This enables more primers to anneal to the newly synthesised strand (step 4) and further DNA polymerisation occurs using the new strand as the template (step 5). This process is repeated.

Pan *et al.* (2012) developed the phi29-mRNA amplification (PMA) method that was adapted from their whole genome amplification procedure.[31] The cDNA is circularised by intra-molecular ligation before amplification by phi29. This enables multiple strand displacement and amplification to occur in a rolling circle, so that both ends of the cDNA are captured (Fig. 4).

Pan *et al.* used PMA to detect ~5000 transcripts from a single cell from an erythroleukemic cell line, with coverage of most or all codons. However, PMA was unable to detect many low abundance transcripts. The low abundance sequences that were picked up were sometimes missing their 3' ends. This suggests these transcripts were being lost before or after cDNA circularisation because of incomplete ligation. Although PMA is less sensitive than an exponential amplification method, it gives full-length trancripts. This may be important for delineating novel splice variants.

Discussion

There are a multitude of methods currently available to amplify the transcriptome of single cells. Only a few of the more mainstream methods have been discussed here and there are many others.[32] The next steps will be to determine systematically which methods have the best sensitivity, coverage, reproducibility, and fidel-

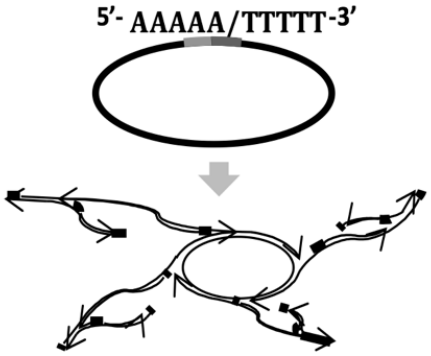

Fig. 4. The cDNA template is circularised, and strand displacement with further primer annealing for phi29 DNA polymerase initiation occurs in the same fashion as in Fig. 3.

ity, and how to improve upon these elements. Very few studies attempt to quantify biological stochasticity and stochasticity caused by technical error.[33]

All of these methods are limited to amplifying poly-adenylated RNA transcripts, i.e. predominantly mRNA. Future methods should encompass non poly-adenylated mature RNA species that have a role in gene expression regulation and carcinogenesis, such as micro-RNA.[34] Poly A tailing to these non poly-adenylated molecules is a possible solution. Improving sensitivity will also be important for studying long non-coding RNA with important regulatory roles, despite being present in low numbers.[35]

Expense and labour are also factors often overlooked in the literature. It may be necessary to sequence the transcriptome of hundreds or thousands of cells from a single tissue and the costs and labour quickly skyrocket. The future will likely be in automated microfluidic lab-on-a-chip methods, such as those developed by Fluidigm® that use a PCR based approach.[36] The *Fluidigm® C1™* Single-Cell Auto Prep System is capable of processing up to 96 individual cells at a time. Linear amplification methods tend to be more time consuming than PCR-based methods, and therefore perhaps less suited to clinical use.[23]

As single cell transcriptome amplification methods improve, so will our understanding of biological processes involved in carcinogenesis and organ or embryo development. We will only reach a fuller understanding of these processes when we go beyond poly-adenylated transcripts to all cellular transcripts. It is hoped that one day these methods will be applied in the clinic to improve the sensitivity and specificity of diagnosis, determine prognosis and improve the treatment of cancer and other diseases.

Acknowledgements

Thanks to Dr Robert Day for his thoughtful mentorship in the single cell arena; Professor Parry Guilford for his leadership and vision; and to everyone in the Cancer Genetics Laboratory for helping it all come together.

References

1. Nowell, P. C. *Science* **1976,** *194*, 23–28.
2. Merlo, L. M.; Pepper, J. W.; Reid, B. J.; Maley, C. C. *Nat. Rev. Cancer* **2006,** *6*, 924–935.
3. Axelrod R.; Axelrod D. E.; Pienta K. J. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13474–13479.
4. Bengtsson M.; Ståhlberg A.; Rorsman P.; Kubista M. *Genome Res.* **2005**, *15*, 1388-1392.
5. Boon W.C.; Petkovic-Duran K.; Zhu Y.; Manasseh R.; Horne M. K.; Aumann T. D. *J. Vis. Exp*. **2011**, *11*, e3144.
6. Brady, G.; Barbara, M.; Iscove, N. *Methods Mol. Cell. Biol*. **1990,** *2*, 17-25.
7. Eberwine J.; Yeh H.; Miyashiro K.; Cao Y.; Nair S.; Finnell R.; Zettel M.; Coleman P. *Proc. Natl. Acad. Sci.* **1992,** *89*, 3010–3014.
8. Iscove, N. N.; Barbara, M.; Gu, M.; Gibson, M.; Modi, C; Winegarden, N. *Nat. Biotechnol.* **2002**, *20*, 940-943.
9. Subkhankulova, T.; Livesey, F. J. *Genome Biol.* **2006**,*7*, R18.

Islam, S.; Zeisel, A.; Joost, S.; Manno, G.; Zajac, P.; Kasper, M.; Lönnerberg, P.; Linnarsson, S. *Nat. Methods* **2014**, *11*, 163-166.

11. Chenchik, A.; Zhu, Y.; Diatchenko, L.; Li., R.; Hill, J.; Siebert, P. In *RT-PCR Methods for Gene Cloning and Analysis* (Ed.: Siebert, P.; Larrick, J.), BioTechniques Books, MA, 1998, 305–319.
12. Picelli, S.; Faridani, O. R.; Björklund, A. K.; Winberg, G.; Sagasser, S.; Sandberg, R. *Nat. Protoc.* **2014**, *9*, 171-181.
13. Tang, F.; Barbacioru, C.; Nordman, E.; Li, B.; Xu, N.; Bashkirov, V. I.; Lao, K.; Surani, M. A. *Nat Protoc*. **2010**, *5*, 516-535.
14. Sasagawa, Y.; Nikaido, I.; Hayashi, T.; Danno, H.; Uno, K. D.; Imai, T.; Ueda, H. R. *Genome Biol.* **2013**, *14*, R31.
15. Iscove, N. N.; Barbara, M.; Gu, M.; Gibson, M.; Modi , C.; Winegarden, N. *Nat. Biotechnol.* **2002**, *20*, 940-943.
16. Tang, D. T.; Plessy, C.; Salimullah, M.; Suzuki, A. M.; Calligaris, R.; Gustincich, S.; Carninci, P. *Nucleic Acids Res.* **2013**, *41*, e44.
17. Plessy, C.; Desbois, L.; Fujii, T.; Carninci, P. *Bioessays* **2013**, *35*, 131-140.
18. Kapteyn, J.; He, R.; McDowell, E. T.; Gang, D. R. *BMC Genomics* **2010**, *11*, 413.
19. Schmidt, W.; Mueller, M. *Nucleic Acids Res.* **1999**, *27,* e31.
20. Picelli, S.; Björklund, A. K.; Faridani, O. R.; Sagasser, S.; Winberg, G.; Sandberg, R. *Nat. Methods* **2013**, *10*, 1096-1098.
21. Bhargava, V.; Ko, P.; Willems, E.; Mercola, M.; Subramaniam, S. *Sci Rep*. **2013**, *3*, e1740.
22. Jeffreys, A. J.; Wilson, V.; Neumann, R.; Keytel, J. *Nucleic Acids Res.* **1988,** *16*, 10953–10971.
23. Day, R.; McNoe, L.; Macknight, R. *Int. J. Plant Genomics* **2007**, *2007*, e61028.
24. Lauss, M.; Vierlinger, K.; Weinhaeusel, A.; Szameit, S.; Kaserer, K.; Noehammer, C. *Virchows Arch.* **2007**, *451*, 1019-1029.
25. Wilhelm, J.; Muyal, J.; Best, J.; Kwapiszewska, G.; Stein, M.; Seeger, W.; Bohle, R.; Fink, L. *Clin. Chem.* **2006**, *52*, 1161-1167.
26. Skinner, G.; Baumann, C.; Quinn, D.; Molloy, J.; Hoggett, J. *J. Biol. Chem.* **2004**, *279*, 3239-323244.
27. Goff, L.; Bowers, J.; Schwalm, J.; Howerton, K.; Getts, R.; Hart, R. *BMC Genomics* **2004**, *5*, e76.
28. Hashimshony, T.; Wagner, F.; Sher, N.; Yanai, I. *Cell Rep*. **2012**, *2*, 666-673.
29. Tougan, T.; Okuzaki, D.; Nojima, H. *Nucleic Acids Res.* **2008**, *36*, e92.
30. Blanco, L.; Bernad, A.; Lázaro, J.; Martín, G.; Garmendia, C.; Salas, M. *J. Biol. Chem.* **1989**, *264*, 8935-8940.
31. Pan, X.; Durrett, R. E.; Zhu, H.; Tanaka, Y.; Li, Y.; Zi, X.; Marjani, S. L.; Euskirchen, G.; Ma, C.; Lamotte, R. H.; Park, I.-H. H.; Snyder, M. P.; Mason, C. E.; Weissman, S. M. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 594-599.
32. Saliba, A.-E.; Westermann, A.; Gorski, S.; Vogel, J. *Nucleic Acids Res.* **2014**, gku555.
33. Marinov, G.; Williams, B.; McCue, K.; Schroth, G.; Gertz, J.; Myers, R.; Wold, B. *Genome Res.* **2014**, *24*, 496-510.
34. Saini, H. K.; Griffiths-Jones, S.; Enright, A. J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17719–17724.
35. Derrien, T.; Johnson, R.; Bussotti, G.; Tanzer A. *Genome Res.* **2012,** 22, 1775–1789.
36. Streets, A.; Zhang, X.; Cao, C.; Pang, Y.; Wu, X.; Xiong, L.; Yang, L.; Fu, Y.; Zhao, L.; Tang, F.; Huang, Y. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 7048-7053.