[Frontiers in Bioscience 19, 1015-1027, June 1, 2014]

Insights into the next generation of cancer stem cell research

Daniel V. Brown[1], Theo Mantamadiotis[1]

[1]Cancer Signaling Laboratory 1, Department of Pathology, The University of Melbourne, Parkville, VIC 3010, AUSTRALIA

TABLE OF CONTENTS

1. ABSTRACT

The understanding of how cancer stem cells (CSCs) or tumor-initiating cells (TICs) behave is important in understanding how tumors are initiated and how they recur following initial treatment. More specifically to understand how CSCs behave, the different signaling mechanisms orchestrating their growth, cell cycle dynamics, differentiation, trans-differentiation and survival following cytotoxic challenges need to be deciphered. Ultimately this will advance the ability to predict how these cells will behave in individual patients and under different therapeutic conditions. Second or next-generation sequencing (NGS) capabilities have provided researchers a window into the molecular and genetic clockwork of CSCs at an unprecedented resolution and depth, with throughput capabilities allowing sequencing of hundreds of samples in relatively short timeframes and at relatively modest costs. More specifically NGS gives us the ability to accurately determine the genomic & transcriptomic nature of CSCs. These technologies and the publicly available cancer genome databases, together with the ever increasing computing power available to researchers locally or via cloud-based servers, are changing the way biomedical cancer research is approached.

2. INTRODUCTION

This review will present the current knowledge and discuss the use of NGS technologies in understanding cancer stem cell (CSC) biology and explore the advantages and challenges of applying next generation sequencing (NGS) and bioinformatics techniques. Particular emphasis will be placed upon the use of these technologies to decipher the dynamic genomic changes of the transcriptome (see Table 1. Glossary of terms) and epigenome of CSCs. We discuss the important biological considerations relevant to CSC genomic analysis, as well as illustrate bioinformatic approaches and considerations when analyzing NGS data. The value of these approaches is finally considered and the ultimate therapeutic benefits arising are discussed.

3. CANCER STEM CELLS

3.1. Cancer stem cell identity and development

It is important to note that the nature of CSCs or tumor -initiating cells (TICs) is like a set of moving goal posts; the moment a particular characteristic is discovered and attributed to these cells, it is not long before another perhaps competing view is presented. The earliest model of

Table 1. Glossary of terms used in next generation sequencing and bioinformatics

Term	Definition
Algorithm	A precise set of instructions for performing a certain task. Often represented as a computer program to transform a given input into a desired output
Alignment	The process of mapping a DNA sequence back to a reference genome to identify the locus from which it originated from
Cloud computing	A loosely coupled network of computers usually housed off-site and administered by a separate entity. Often provided as a service for a fee
Cluster computing	A tightly coupled network of computers usually administered in house for the purposes of data storage and analysis
De novo alignment	The process of mapping a DNA sequence without a reference genome using only the DNA sequences themselves
DNA-seq	The process of using massively parallel sequencing to identify the precise order of nucleotides that comprise a pool of DNA fragments
Epigenome	The genome-wide catalogue of heritable, reversible chemical modifications to DNA and histones in a population of cells
Gene expression signature	A specific pattern of a given subset genes expressed by a population of cells that is characteristic of a given state such as cancer
Machine learning	The use of computer algorithms that learn from data such as gene expression to facilitate pattern recognition and classification to predict parameters such clinical response
Multiplexing	The use of molecular barcodes comprised of a short nucleotide sequence not present in the target genome, to uniquely label a sample. Multiple samples with distinct barcodes are then pooled and sequenced in a single reaction. Samples are then deconvoluted in the computational analysis
Pair-end sequencing	The process of sequencing both ends of a DNA fragment. Improves the performance of alignment algorithms. Post alignment, the presence of discordantly mapping coordinates for each end indicates mRNA splicing or an insertion/ deletion has occurred
Personalized medicine	A paradigm in clinical practice that advocates the tailoring of medical decisions and treatments based on an individual's specific genetic features such as being mutant or wildtype for an oncogene
Pipeline	The chaining together of distinct computer programs where the output of the previous program is the input of the next program. Facilitates the execution of non-dependent software in parallel to increase the speed of data analysis
RNA-seq	The process of sequencing a pool of cDNA fragments that have been generated by reverse transcription of RNA, typically mRNA
Sequence deconvolution	The computational procedure of sorting a pool of DNA fragments into sample-specific or species-specific reads based on the sequence
Sequencing bias	In the context of NGS, the presence of technical noise that obscures the biological signal. Can arise during the construction of the DNA/ cDNA library or the from sequencing technology utilized
Sequencing read	The linear sequence of nucleotides originating from a single DNA fragment
Subtype	In the context of genomics, a subset of all the instances of a disease in a population such as cancer. Defined by the a priori pattern of molecular features such as gene expression, epigenetic marks or somatic mutations
Transcriptome	The full complement of RNAs transcribed from the collective genome of a population of cells. Includes rRNA, tRNA, mRNA and miRNA
Algorithm	A precise set of instructions for performing a certain task. Often represented as a computer program to transform a given input into a desired output
Alignment	The process of mapping a DNA sequence back to a reference genome to identify the locus from which it originated from
Cloud computing	A loosely coupled network of computers usually housed off-site and administered by a separate entity. Often provided as a service for a fee
Cluster computing	A tightly coupled network of computers usually administered in house for the purposes of data storage and analysis
De novo alignment	The process of mapping a DNA sequence without a reference genome using only the DNA sequences themselves
DNA-seq	The process of using massively parallel sequencing to identify the precise order of nucleotides that comprise a pool of DNA fragments
Epigenome	The genome-wide catalogue of heritable, reversible chemical modifications to DNA and histones in a population of cells
Gene expression signature	A specific pattern of a given subset genes expressed by a population of cells that is characteristic of a given state such as cancer
Machine learning	The use of computer algorithms that learn from data such as gene expression to facilitate pattern recognition and classification to predict parameters such clinical response
Multiplexing	The use of molecular barcodes comprised of a short nucleotide sequence not present in the target genome, to uniquely label a sample. Multiple samples with distinct barcodes are then pooled and sequenced in a single reaction. Samples are then deconvoluted in the computational analysis
Pair-end sequencing	The process of sequencing both ends of a DNA fragment. Improves the performance of alignment algorithms. Post alignment, the presence of discordantly mapping coordinates for each end indicates mRNA splicing or an insertion/ deletion has occurred
Personalized medicine	A paradigm in clinical practice that advocates the tailoring of medical decisions and treatments based on an individual's specific genetic features such as being mutant or wildtype for an oncogene
Pipeline	The chaining together of distinct computer programs where the output of the previous program is the input of the next program. Facilitates the execution of non-dependent software in parallel to increase the speed of data analysis
RNA-seq	The process of sequencing a pool of cDNA fragments that have been generated by reverse transcription of RNA, typically mRNA
Sequence deconvolution	The computational procedure of sorting a pool of DNA fragments into sample-specific or species-specific reads based on the sequence
Sequencing bias	In the context of NGS, the presence of technical noise that obscures the biological signal. Can arise during the construction of the DNA/ cDNA library or the from sequencing technology utilized
Sequencing read	The linear sequence of nucleotides originating from a single DNA fragment
Subtype	In the context of genomics, a subset of all the instances of a disease in a population such as cancer. Defined by the a priori pattern of molecular features such as gene expression, epigenetic marks or somatic mutations
Transcriptome	The full complement of RNAs transcribed from the collective genome of a population of cells. Includes rRNA, tRNA, mRNA and miRNA

CSCs presented a hierarchical model where a multipotent progenitor cell is able to both self renew and differentiate into more restricted progeny (1, 2). The implication of this model is that eliminating the stem cell fraction will lead to exhaustion of the regenerative capacity of the tumor and regression. However, genomic analysis of clinical tumor samples have uncovered substantial heterogeneity that cannot have been generated via the hierarchical CSC model (3-6). The ierarchical model also imposes the limitation that only the stem cell fraction is able to generate tumors. There are some tumors that can be initiated by injection of a few or even a single cell indicating tumorigenic potential is not rare for these cancers (5, 7, 8). These studies were performed with more rigorous transplantation procedures, such as severely immunocompromised mice or syngeneic models, compared with the earlier studies that conceived the stem cell hypothesis (9-11)

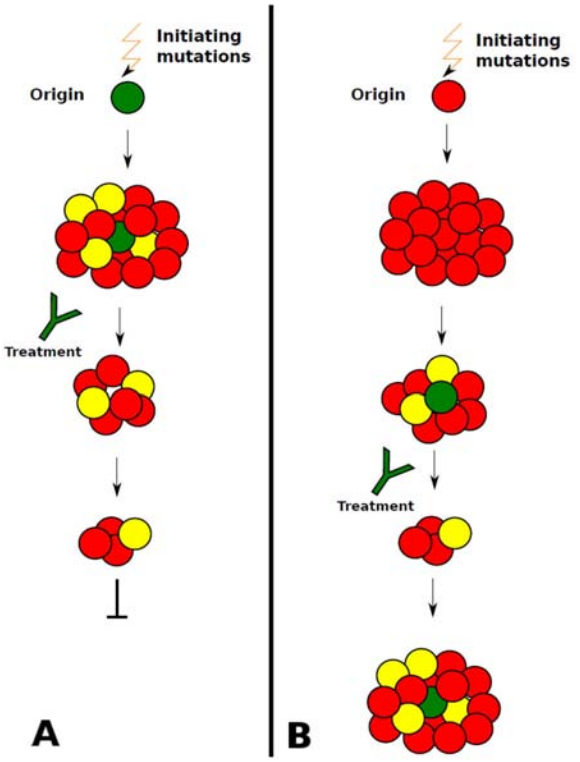

Figure 1. Models of cancer stem cell development. A) The hierarchical model claims the proliferative potential of a malignant tumor resides in a rare subpopulation (green) with a resemblance to normal stem cells. The CSC has the ability to self-renew and differentiate into the cell types that dominate the tumor mass. Upon treatment with an agent that targets the stem cell fraction, the bulk of the tumor ultimately exhausts its replicative ability. B) The plasticity model claims the stem-like subpopulation (green) arises through infrequent, stochastic dedifferention of cells in the tumor bulk. Upon treatment with an agent that targets the CSCs, the stem-like fraction can be regenerated by the remainder of the tumor bulk and recurrence occurs.

3.2. Dedifferentiation model

There is evidence of interconversion of non-stem cells to stem cells. Human mammary epithelial stem cells (HMECs) grow adherently in culture but a small sub population grow in suspension (12). Flow cytometry analysis of the cells in suspension revealed both a CSC (CD44+) and non-CSC fraction (CD44-). Single cell cloning and *in vivo* transplantation revealed that the non-stem cell and stem cells can readily inter-convert, albeit at a low rate (12). The shuttling between non-CSC and CSC had parallels with epithelial to mesenchymal transition (EMT) and was regulated by TGF beta and the epigenetic configuration at the ZEB1 promoter. Both proteins have a well described role in EMT (13).

The cancerous state has been proposed to increase the probability of dedifferentiation from non-stem cells to stem cells which naturally occurs *in vivo* (14). This de-differentiation mechanism has been shown to occur for the aggressive brain tumor glioblastoma multiforme (GBM) *in vivo* (15, 16). Using GFAP-Cre mice, the tumor supressors PTEN and p53 were specifically knocked down in mature glial cells using lentiviruses stereotaxically injected directly into the hippocampus, generating gliomas of varying grades (15). Synapsin I-Cre mice were also injected with lentiviruses targeting H- Ras V12 and shp53 or shNF1 and shp53, into the cortex demonstrating that fully mature neural cells can give rise to tumors (16). Tumor cells were isolated and had phenotypic characteristics of stem cells when cultured *in vitro*. Immunofluorescence analysis of tumors taken from mice at different time points showed an increase of stem cell markers and a decrease of differentiation markers over time (16). Therefore it is still not clear if the cell of origin for CSC-driven tumors is a normal stem cell that subsequently differentiates when initiated to generate the multiple cell types present in a tumor, or a mature cell type that dedifferentiates to a stem-like state (Figure 1).

4. NEXT GENERATION SEQUENCING TECHNOLOGY

Second generation or next-generation sequencing (NGS) is the massively paralleled determination of short reads of DNA sequence from a pool of DNA fragments (17). It has been used to detect rearrangements in cancer genomes (18), somatic mutations (19) and germ-line mutations (20). By using an individual's own germline sequence as a baseline, deviations from this baseline, ie. somatic mutations, can be identified at an accuracy comparable to Sanger sequencing at vastly reduced cost by economies of scale (2, 21, 22). RNA sequencing (RNA-seq) enables unbiased digital measurement of transcription from cells and tissues at a higher resolution and dynamic range than for array-based techniques. (23-27). RNA-seq demonstrates high reproducibility between technical replicates and consistency with microarray results (25, 28). However, RNA-seq outcomes may be inaccurate due to biases arising from gene length (25, 29) and first-strand synthesis efficiency/errors (26, 30). Statistical analysis of RNA-seq data is currently an ongoing area of research (26, 31). The analysis pipeline for RNA-seq data is more complex than that for microarray. Short reads are first aligned to a reference genome and/or transcriptome, although de novo methods for transcript assembly exist (27, 32). Aligned reads are then summarized to discrete genomic locations, such as genes for differential expression testing (4, 6, 33).

There are various sequencing-based technologies to analyze the epigenome. For classic epigenetic analysis of modified DNA by methylation there are two widely used approaches: bisulfite-based methods (BS-Seq) and methylated DNA immunoprecipitation sequencing (MeDIP-Seq) Bisulfite-based methods provide single base pair resolution but because of the reduced complexity of the sequence introduced by the bisulfite conversion, alignment is more difficult than MeDIP-seq (27, 34). Conversely MeDIP-Seq enriches for

methylated DNA and therefore suffers from lower resolution as the recovered fragment is typically 150 – 200bp (27, 35).

4.1. Advantages of NGS RNA-seq over microarray and tiling arrays

In the late 1990s the thorough analysis of cell transcriptomes was made possible by the development of techniques such as microarray and serial analysis of gene expression (SAGE) (36, 37). These techniques enabled the analysis of transcriptomes of tumor tissue or defined stem cell populations (38, 39) and led to the identification of genes involved in the specific cell behaviors in normal and pathological states. In less than a decade, the development of NGS has gradually become the genomic analysis tool of choice. NGS analysis is equally applicable to DNA sequencing as well as RNA sequencing. Unlike microarrays, and similar to SAGE, RNA-seq does not require knowledge of predetermined expressed sequences, thus provides an unbiased approach to transcript discovery, and because of the sensitivity of the technique it provides the ability to not only discover novel transcripts but to also reveal novel isoforms or fusions that would be missed on an array. Compared with SAGE, NGS offers much higher throughputs and therefore provides increased coverage of the genome.

4.2. Deciphering CSC transcriptomes by RNA-seq

The sensitivity of NGS is especially useful in the analysis of stem cell transcriptomes. Stem cells often express very low levels of many genes indicating a transcriptional state that keeps the cell ready to initiate a variety of differentiation programs. Such low level gene expression is difficult to detect by microarray-based methods, but with the sensitivity and ability to perform deep sequencing with RNA-seq, low level transcript identification is feasible, even at the single cell level (1, 40-42). Because of the relatively short time that RNA-seq has been implemented and the limitations of identifying cancer-specific stem/-initiating cells, there are only limited examples of transcriptome analyses. Most applications have focused on embryonic stem cells and non-stem tumor cells. Transcriptome analysis using RNA-seq has revealed regulatory networks in embryonic stem cells.RNA-seq technology has also been used to study the transcriptomes of liver CSCs and breast CSCs (3, 5, 40). A combination of methodologies using NGS and microarray-based gene expression profiling of CD44+ breast CSCs isolated from primary ER-alpha+ breast cancer showed that CSCs exhibited the expected overexpression of genes involved in stem cell maintenance, but also showed higher level expression of numerous genes driving the PI3K pathway suggesting that ERα+ breast cancer CSCs require an active PI3K pathway and revealing a possible PI3K-dependent mechanism for the endocrine resistance of this tumor subtype (5, 41). Different stages of tumor cell differentiation correlating with disease progression have been identified using RNA-seq. Jiang and colleagues showed that during blast crisis, the final phase of chronic myeloid leukemia (CML), progenitors have an increased IFN-gamma pathway gene expression as well as BCR-ABL amplification. Furthermore, during CML progression there was an upregulation in the expression of the IFN-responsive adenosine deaminase acting on RNA enzyme (ADAR1) isoform which correlated with the expression of a mis-spliced form of GSK3-beta which in turn is implicated in leukemia stem cell self-renewal (9, 43).

4.3. NGS analysis of stem cell transcriptomes

In addition to studying the genome and mRNA transcriptome of stem cells, NGS has also been used to study the broader RNA transcriptome. Post-transcriptional regulation especially through microRNA (miRNA) and long noncoding RNA is an important regulator of gene expression and cellular function. In humans the majority of transcribed RNA, aside from ribosomal RNA, is involved in regulation of the coding genes; these are the ncRNAs (non-coding RNAs), such as microRNAs (miRNAs), long noncoding RNAs (lncRNAs) and circular RNAs. MicroRNAs have diverse roles in regulating gene expression and controlling development and disease (7, 8, 44). Many miRNAs are located in genomic regions that are deleted or amplified in various cancer types, suggesting they might play an important role in cancer progression. RNA-seq is well suited to ncRNA analysis since techniques to enrich for small transcripts can be applied which exclude mRNA transcripts and rRNA. Although not specifically a study on CSCs, work investigating the role of Dicer in endothelial precursor cells (EPCs) in an orthotopic breast tumor mouse model revealed, through NGS analysis, two EPC intrinsic VEGF-responsive miRNAs, miR-10b and miR-196b, which led to a decrease in circulating EPCs and a significant defect in angiogenesis-mediated tumor growth in mice (10, 11, 45).

A study comparing the differences between normal neural stem cells and glioblastoma stem cells uncovered 10 differentially expressed microRNAs that predominately act on the p53 pathway facilitating the tumorigenic phenotype of the glioblastoma (12, 46). Because of their small size, microRNAs are unable to be sequenced in a traditional RNA-seq protocol and must be size-selected and prepared separately. However, the relatively low number of microRNAs, approximately 2,500 at present, lends itself to multiplexed sequencing of many samples in a single sequencing reaction (12, 47). Unfortunately early studies were confounded by sequencing bias introduced by nucleotide barcodes in the multiplexing step (13, 48). This bias has since been resolved but it underlies the importance of inspecting the raw data and appropriate quality control measures in NGC experiments.

A further layer of regulation in the mRNA – miRNA axis was added by the discovery of circular RNAs (14, 49-51). Having previously been described as rare in mammalian cells, thousands of circular RNAs were identified as alignments with a non-linear sequence of exons generated by a “backsplicing” mechanism (15, 16, 50, 51). Computational analysis of the circular RNAs revealed tens of miRNA-binding sites within the circular molecule and in-vitro-binding experiments confirmed an association of miRNA with circular RNA (15, 49, 52).

4.4. Challenges in applying NGS to CSC biology

Although NGS can help us better understand the biology of stem cells, there are significant challenges that need to be overcome to adopt the widespread use of this technology for studying stem cells. The first hurdle is the cost of the technology. Although the cost per base for sequencing is rapidly declining, the technology is still expensive compared with medium throughput technologies like microarray. Moreover, the high throughput sequencing instruments are expensive and require specialized training to operate. The large amount of data generated by these machines also requires a sizable investment in infrastructure, such as storage servers and computing clusters. However, the advent of smaller bench top sequencers (MiSeq and Ion proton) with lower throughput, but reduced running costs, should make the technology affordable to an individual laboratory. Also the rise of core facilities and private companies that can sequence samples with a rapid turnaround time and a small fee will eventually make sequencing affordable to small labs.

Another challenge when using these technologies for stem cell analysis is the amount of genetic material required. Most stem cells are extremely rare populations and it is difficult or often impossible to get large numbers of stem cells without significantly altering their properties. This poses a major technological hurdle to using NGS technology for studying rare stem cell populations. However it is now possible to sequence single cells and this opens up exciting possibilities for stem cell research (16, 53-55). Many stem cells are clonal in origin and analysis of genome, transcriptome and epigenome of such single stem cells can yield valuable information about the heterogeneity of these cells (16, 56).

The third hurdle for the widespread use of NGS technologies for studying stem cells is the relatively immature data analysis pipeline for NGS platforms (Figure 2). Unlike microarrays the technology and therefore the approach to data analysis itself is constantly changing, e.g. the adoption of read lengths in excess of 100 bases and pair-end sequencing. Such technological developments have influenced the way aligners work, as modern aligners incorporate more traditional BLAST like algorithms after seeding to a locus (17, 57). Such differences also make it difficult to compare studies between different data analysis pipelines.

In spite of these challenges, the use of NGS technology promises to further our understanding of stem cell function. The scope, sensitivity and specificity of this technology offer unparalleled exciting tools to study stem cell function, differentiation and behavior under different conditions. We can get a glimpse of the power of this technology in studying stem cells by looking at the extent to which microarray has been routinely used in understanding stem cell biology. As the technology platform matures, it becomes easier to use and the hurdles in the widespread adoption of the platform are overcome. Eventually NGS will be as common as microarray to interrogate stem cell function. The advent of low cost third generation sequencing technology like the Oxford Nanopore™ will certainly help promote this trend (18, 58).

5. CURRENT KNOWLEDGE OF THE CSC GENOMIC LANDSCAPE

5.1. Transcriptome of CSCs by microarray

A number of studies have examined gene expression profiles of GBM stem cells to gain insight in the differences between CSCs and normal stem cells and to sub classify GBM stem cells into clinically meaningful subtypes (23, 25-27, 41). The gene expression profiles of a panel of GBM stem cell lines grown from human tumors were compared to profiles obtained from normal neural cells of various origins including astrocytic stem cells, adult neural stem cells and fetal neural stem cells (25, 42). The normal neural stem cells broadly clustered into 3 categories, whilst the GBM specific stem cells were exclusively grouped into 2 of these clusters. GBM stem cells that had low expression of CD133 clustered with adult neural stem cells, and, in contrast, GBM stem cells with a high expression of CD133 clustered with fetal neural stem cells. A 24 gene signature was able to differentiate between adult-like and fetal-like GSCs *ex vivo* (25, 42).

Genes differentially expressed between CD133+ and CD133− glioblastoma-initiating cells were also used to generate a clinically useful gene expression signature that could more accurately predict patient outcome compared to histological or molecular classification (26, 59, 60). The CD133+ overexpressed genes corresponded to genes involved in proliferation, whereas genes with decreased expression corresponded to immune regulation. The candidate signature corresponded to higher grade proneural subtype gliomas and was also associated with poorer survival in The Cancer Genome Atlas dataset (TCGA). Interestingly, glioma samples enriched for the CD133 positive signature contained 3 times as many mutations compared with other samples. However, the genes mutated in the CD133-enriched samples were randomly distributed and no pattern could be discerned (26, 61).

Tag-seq, a precursor technique to RNA-seq, was used in a similar study to profile a panel of normal and glioma CSCs (27, 62). Similar to SAGE, a 17-base pair sequence downstream of the NlaIII restriction site was sequenced in a massively parallel fashion. A panel of glioma-initiating cells were compared with fetal-derived neural stem cells (27, 62-64). The differentially expressed list of genes had high predictive power for survival and tumor grade and was correlated with age. The glioblastoma stem signature was more predictive than IDH status, a biomarker that is entering clinical use (27, 62). Therefore there is potential for a signature of cancer and ‘stemness’ being informative in the clinic.

5.2. Clinical significance of a CSC signature

The clinical significance of a single stem cell marker such as CD133 is controversial. Immunohistochemical screening of GBM CSCs or breast tumors does not indicate a statistically significant link between CD133 or CD15 expression and survival (36, 65). This is in contrast to studies at the mRNA level demonstrating that CD133 expression is a significant negative prognostic factor for both progression-free and

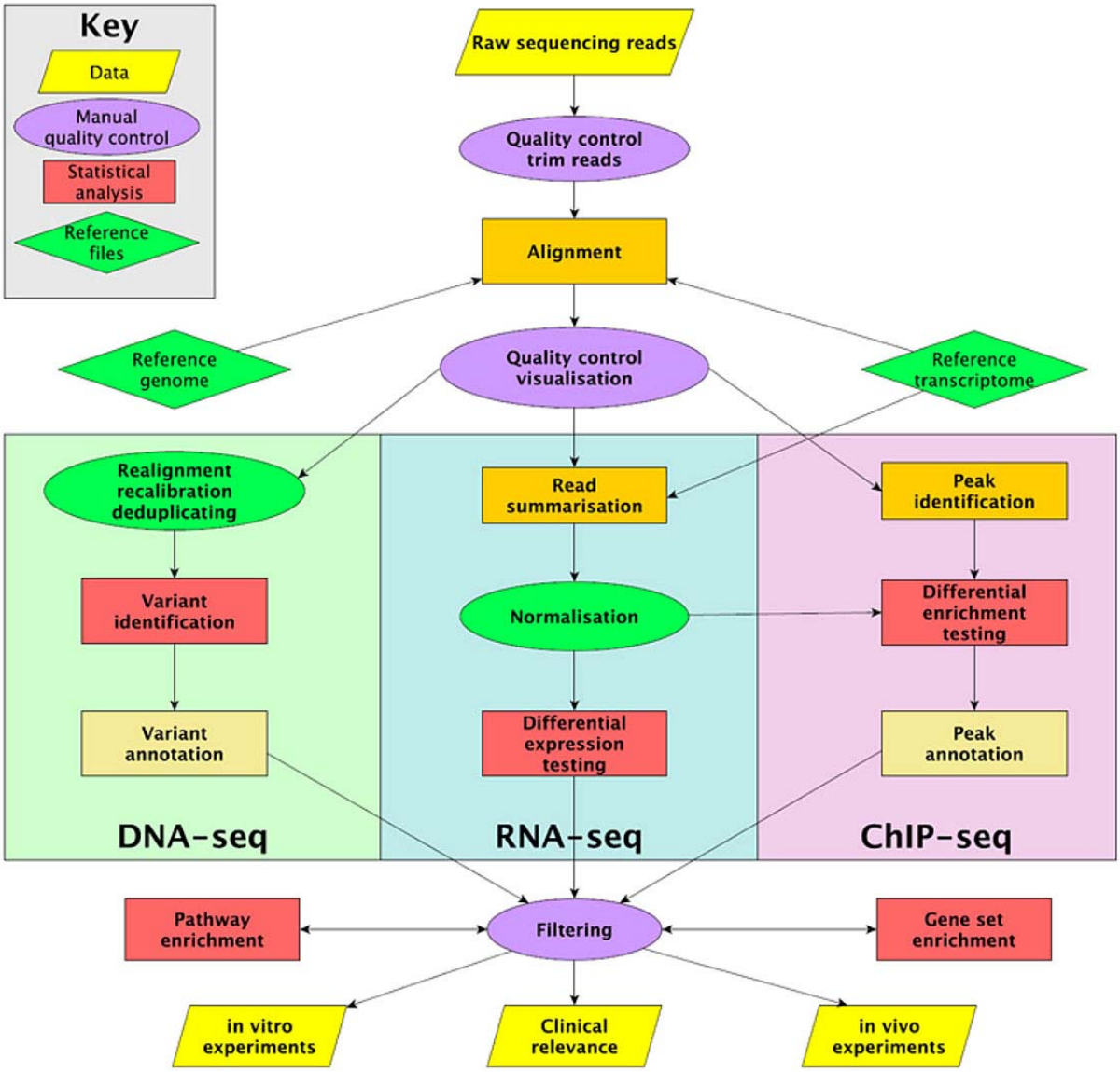

Figure 2. Typical NGS analysis pipeline. Bioinformatic analysis of NGS data can be broken down into discrete steps which involves capturing an input and emitting an output (parallelogram). Both free and commercial software are available to perform analysis with free software, usually requiring more computational and statistical expertise (rectangle). The most frequent types of NGS experiments: DNA-seq, RNA-seq and ChIP-seq share upstream and downstream steps. Frequent manual quality control (ellipse) is employed to ensure systematic technical, and biological errors are accounted for. A common goal for NGS experiments when studying CSCs is the generation of hypotheses for further functional research.

overall survival in GBM (27, 38, 66, 67). More success has been derived from using a subset of genes termed a 'gene signature', to define the CSC phenotype (40-42, 68, 69). In one study, the EphB2 surface marker was used to enrich for murine intestinal stem cells. A variety of gene signatures were derived including a gene signature characteristic of intestinal stem cells. This gene signature correlated with tumor grade and was able to predict recurrence after therapy in humans with colorectal cancer (40, 70). A similar strategy was used in the generation of a stem cell signature predictive for $Her2^+$ breast tumors (41, 71). In this study, mammary tumor-initiating cells were collected from MMTV-Her2/Neu transgenic mice and compared by microarray to the initiating-cell-depleted (non-stem cell) population. Using a machine learning approach on a human dataset to retain stem cell genes predictive in a human $Her2^+$ breast cancer cohort, a 17 gene signature was able to classify $Her2^+$ patient tumors into a 'good' and 'poor' survival group. The signature was also predictive of a response to standard chemotherapy and the Her2-specific neutralizing antibody Trastuzumab™. The signature performed better than the currently available generic

Mammaprint™ signature which tests for 70 genes and predicts recurrence (43, 72), although it was not predictive for other subtypes of breast cancer (41, 72). These types of approaches using stem cell gene signatures are showing promise in many tumor types. A stem cell signature derived from adult cancer and embryonic stem cell transcriptomes was used in a classification analysis on an ovarian cancer dataset (42). Expression of stem cell genes were characteristic of type II ovarian cancer, which is typically the serous subtype and is more aggressive compared to low grade, heterogeneous type I ovarian cancer (42).

5.3. Molecular subtyping by high throughput experiments

Although gene expression profiling in cancer research is well developed, the usefulness in general cancer diagnosis and prediction is still coming to the fore. Already molecular pathology gene expression profiling complements breast cancer diagnosis and prediction for subtype classification (59, 60). Gene expression (transcriptome) profiling was able to predict relapse, treatment response and metastatic potential. A similar approach was applied to GBM, with the aim of assisting in the classification of different grades of gliomas. Quantification of gene expression by microarray combined with unsupervised machine learning techniques was able to differentiate between aggressive GBM, low grade astrocytoma and oligodendrogliomas (61). Clustering of the GBM subset of gliomas based on gene expression revealed 4 previously unrecognized subtypes (62). These 4 groups were found to resemble specific stages in neuronal development. The proneural subtype resembled a differentiated neural cell type and displayed the best prognosis. Conversely, the mesenchymal subtype resembled a primitive cell type and was associated with shortest survival (62-64). Adding to the expression profile, traditional DNA sequencing of 601 genes identified mutations characteristic to each subtype (62).The number of tumor subtypes one can define seems to be limited only by the size of the training set (number of specimens), with one of the last major microarray studies finding 10 breast cancer subgroups by integrating copy number and gene expression from 2,000 tumors (65). NGS with its greater resolution and accuracy has the potential to define further patient subgroups, ultimately paving the way for personalized medicine. This is the premise behind the large cancer genomics projects of TCGA and International Cancer Genome Consortium (ICGC).

5.4. Cancer genome consortia and publicly available cancer genomic data

The field of cancer genomics has been rapidly accelerated by the advent of NGS. Enormous consortia consisting of clinicians, biologists and bioinformaticians have combined to profile hundreds of tumors to identify driving events in tumor initiation and maintenance. TCGA and the ICGC have selected specific cancer types for broad sequencing projects. DNA, methylated DNA, RNA and microRNA data have been generated and are available to researchers. Although these large projects have focused on characterization of heterogeneous tumor tissue, subtypes of tumors exist that have a CSC signature (27, 66, 67). Internet-based tools, such as cBio, are making the results of these analyses accessible to biologists and clinicians (68, 69). Similar tools exist for genome-wide experiments relating to stem cell research (70).

After completion of the human genome project, the next logical step was characterization of the functional elements of the genome. The encyclopedia of DNA elements (ENCODE) project set out to catalog the sequences in the genome that were transcribed and regions involved in transcriptional and epigenetic regulation (71). The project initially started with microarray technology and later utilized NGS. Samples were split into 3 tiers of different coverage. 'Tier 1' is the most thoroughly interrogated set of samples that includes H1 hESC, which is one of the most common human stem cell lines used in biomedical research (72). Several different induced pluripotent stem cells (iPSCs) are represented in 'Tier 2'. The key findings from this large study were that 62% of the genome was represented in RNA sequencing reads, 8.1% of the genome is capable of binding proteins and 3.9% of the genome contains chromatin accessible to proteins, implying that these regions are regulated. Overall 84.4% of the genome is covered by an experiment undertaken by the ENCODE consortium (72). This data are publically available and represent a rich resource for further analysis by researchers with specific cancer biology questions.

6. FUTURE PERSPECTIVES

6.1. Epigenetic contribution to CSC phenotype

Recently the tumorigenic properties imparted by the epigenetic status of CSCs was investigated (19, 73). The epigenome of glioblastoma CSCs was reset to an embroyonic state by transduction of classical induced pluriopotency genes Oct4 and Klf4 (2 of the 4 so-called Yamanaka iPSC genes). A small proportion (2 out of 14 lines) were successfully reprogrammed to a pluripotent state and were able to be maintained in culture, similar to reprogramming frequency in normal cells. These reprogrammed CSCs were more similar in gene expression profile to iPSCs than the parental cells, yet retained the genomic alterations (mutations) of the parent cell. Likewise, their epigenetic profile was similar to iPSCs and not to the parental tumor. These reprogrammed CSCs formed teratomas in mice, while the parental tumors formed gliomas. The majority of the resulting teratomas were neural-like, with a mixture of other cell lineages present at low frequency. This indicates that there is some genomic memory retained in reprogrammed CSCs, independent of its epigenetic status (20, 73). This work was conducted with array technology and it is possible that the more unbiased nature of NGS can uncover epigenetic loci important in cancer developmental memory.

6.2. Unraveling tumor heterogeneity

Heterogeneity is a major complicating factor in the interpretation of cancer and stem cell genomics (74-76). It is well documented that analysis of genetic alterations of tumors is complicated by contamination of the tumor with normal stroma and immune cells (24, 77). A similar analogy can be applied for experiments involving stem

cells, as only a fraction of cells *in vivo* or in culture will occupy a stem cell state, being surrounded by its much more numerous differentiated progeny. NGS has been used to study the tumor — stromal interface in studies of human tumor cells xenografted into mice (28, 78). NGS is then used to deconvolute species specific gene expression by distinguishing human and mouse sequences during alignment, although this approach has its technical challenges (29, 79). With the ability of NGS to perform pair-end sequencing with longer reads than was previously available, it is likely that it will be possible to study complex interactions between various tumor microenvironments, including emerging research investigating interactions between stromal cells, immune cells and tumor cells *in vivo*. Models of stem cell driven tumors will benefit from such approaches, as it will be possible to investigate signaling in the tumor stem cell niche *in vivo* using xenograft models.

6.3. Single-cell sequencing technologies

A relatively new technological development to study the prevalence and differences of a potential stem cell subpopulations in a tumor is single-cell sequencing (30, 54). Microfluidic instruments combined with 'single-tube' protocols and multiple rounds of PCR are able to generate sufficient amounts of DNA and RNA that can be sequenced by NGS. For RNA-seq this technique can only be performed with a poly-A tail capture, resulting in a 3'-end bias dependent on the length of the transcript (31, 54). The extensive PCR steps also result in bias in the sequences amplified and propagation of PCR artifacts. The statistical analysis of single-cell RNA-seq is currently in the early stages, with the stochastic nature of gene expression at the single-cell level requiring to be modeled for the analysis (32, 80). Despite these issues, RNA-seq and single-cell resolution analysis of individual cell lineages derived from the embryonic inner cell mass has been applied and has revealed mechanisms regulating the epigenome of specific embryonic stem cells during early development (81, 82).

Single nucleus sequencing of breast cancer tumors indicate the diversity in copy number profiles observed is consistent with the anatomical origin of the cells within the tumor bulk *(34, 53)*. Phylogenetic reconstruction of the evolution of individual cells indicate divergence of the major subclones when the tumor was small. This suggests an initial aneuploid generating event followed by clonal expansion and divergence. No common ancestors cells were identified that link the emergent subclones together. Given that 100 individual nuclei were sequenced, these rare cells may have been missed. Complicating matters are the presence of multiple, distinct pseudodiploid cancer cells throughout the tumor with no obvious relationship to each other or the main tumor lineage. These cells may have arisen from the initial insult that generated the genomic instability in the first place and are continually being generated and selected against in the background *(35, 53)*.

6.4. Dynamic plasticity of CSC phenotype

Most genomic experiments at present use a design that takes a snapshot of the population at a single point in time. The kinetics of the CSC state is of particular interest in the context of cancer development. It may be that the potential of tumor cells to dedifferentiate to a CSC state over time is more relevant to the clinical course of a cancer, rather than the number of CSCs at steady state (13, 37). Multipoint genome-wide 'kinetic' measurements will be expensive to perform and more complicated to analyze compared with a static experimental design, but have the potential to reveal much about the dynamic nature of CSCs. Coupled with single-cell sequencing one could track crucial intermediate states between non-CSCs and CSCs, such as those recently shown to be epigenetically modulated in iPSC reprogramming (39, 83, 84).

6.5. How will NGS help improve the way patients are treated?

For all the advances in the technology and the understanding of the biology of CSCs, the ultimate reason to be conducting this research, is to improve cancer patient treatment by providing a real improvement in both quality of life and survival. Having said this, the inevitable cost-benefit calculations come into play, otherwise new therapies would simply be unaffordable to most patients. NGS seems to provide real hope that will enable cancer researchers to further delve into the subtleties of CSC biology, eg. the identity of the molecular features distinguishing CSCs from other cells.. NGS will also enable unprecedented disease-specific treatments to be delivered to patients quickly and affordably, by informing clinicians on how to intervene with new and existing combinations of drugs.

7. ACKNOWLEDGEMENTS

The work is supported by funds from the Department of Pathology, The University of Melbourne. We thank A/Prof Andrew Lonie, A/Prof Frederic Hollande and Paul Daniel for helpful discussions. We also thank Gulay Filiz and Tina Isaakidis for critical reading of the manuscript

8. REFERENCES

1. Łabaj, P. P., G. G. Leparc, B. E. Linggi, L. M. Markillie, H. S. Wiley and D. P. Kreil: Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics* 27, 383–91 (2011)

2. Bonnet, D. and J. E. Dick: Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med* 3, 730–737 (1997)

3. Ho, D. W. Y., Z. F. Yang, K. Yi, C. T. Lam, M. N. P. Ng, W. C. Yu, J. Lau, T. Wan, X. Wang, Z. Yan, H. Liu, Y. Zhang and S. T. Fan: Gene Expression Profiling of Liver Cancer Stem Cells by RNA-Sequencing. *PLoS ONE* 7, 37159 (2012)

4. Navin, N., A. Krasnitz, L. Rodgers, K. Cook, J. Meth, J. Kendall, M. Riggs, Y. Eberling, J. Troge, V. Grubor, D. Levy, P. Lundin, S. Maner, A. Zetterberg, J. Hicks and M.

Wigler: Inferring tumor progression from genomic heterogeneity. *Genome Res* 20, 68–80 (2010)

5. Hardt, O., S. Wild, I. Oerlecke, K. Hofmann, S. Luo, Y. Wiencek, E. Kantelhardt, C. Vess, G. P. Smith, G. P. Schroth, A. Bosio and J. Dittmer: Highly sensitive profiling of CD44+/CD24- breast cancer stem cells by combining global mRNA amplification and next generation sequencing: evidence for a hyperactive PI3K pathway. *Cancer Lett* 325, 165–174 (2012)

6. Shackleton, M., E. Quintana, E. R. Fearon and S. J. Morrison: Heterogeneity in Cancer: Cancer Stem Cells versus Clonal Evolution. *Cell* 138, 822–829 (2009)

7. Kelly, P. N., A. Dakic, J. M. Adams, S. L. Nutt and A. Strasser: Tumor growth need not be driven by rare cancer stem cells. *Science* 317, 337 (2007)

8. Quintana, E., M. Shackleton, M. S. Sabel, D. R. Fullen, T. M. Johnson and S. J. Morrison: Efficient tumour formation by single human melanoma cells. *Nature* 456, 593–598 (2008)

9. Jiang, Q., L. A. Crews, C. L. Barrett, H.-J. Chun, A. C. Court, J. M. Isquith, M. A. Zipeto, D. J. Goff, M. Minden, A. Sadarangani, J. M. Rusert, K.-H. T. Dao, S. R. Morris, L. S. B. Goldstein, M. A. Marra, K. A. Frazer and C. H. M. Jamieson: ADAR1 promotes malignant progenitor reprogramming in chronic myeloid leukemia. *Proc Natl Acad Sci USA* 110, 1041–1046 (2013)

10. Singh, S. K., I. D. Clarke, M. Terasaki, V. E. Bonn, C. Hawkins, J. Squire and P. B. Dirks: Identification of a cancer stem cell in human brain tumors. *Cancer Research* 63, 5821–5828 (2003)

11. Al-Hajj, M., M. S. Wicha, A. Benito-Hernandez, S. J. Morrison and M. F. Clarke: Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci USA* 100, 3983–3988 (2003)

12. Chaffer, C. L., I. Brueckmann, C. Scheel, A. J. Kaestli, P. A. Wiggins, L. O. Rodrigues, M. Brooks, F. Reinhardt, Y. Su, K. Polyak, L. M. Arendt, C. Kuperwasser, B. Bierie and R. A. Weinberg: Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state. *Proc Natl Acad Sci USA* 108, 7950–7955 (2011)

13. Chaffer, C. L., N. D. Marjanovic, T. Lee, G. Bell, C. G. Kleer, F. Reinhardt, A. C. D'Alessio, R. A. Young and R. A. Weinberg: Poised Chromatin at the ZEB1 Promoter Enables Breast Cancer Cell Plasticity and Enhances Tumorigenicity. *Cell* 154, 61–74 (2013)

14. Marusyk, A., V. Almendro and K. Polyak: Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* 12, 323–334 (2012)

15. Chow, L. M. L., R. Endersby, X. Zhu, S. Rankin, C. Qu, J. Zhang, A. Broniscer, D. W. Ellison and S. J. Baker: Cooperativity within and among Pten, p53, and Rb pathways induces high-grade astrocytoma in adult brain. *Cancer Cell* 19, 305–316 (2011)

16. Friedmann-Morvinski, D., E. A. Bushong, E. Ke, Y. Soda, T. Marumoto, O. Singer, M. H. Ellisman and I. M. Verma: Dedifferentiation of Neurons and Astrocytes by Oncogenes Can Induce Gliomas in Mice. *Science* (2012)

17. Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M. Rothberg: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380 (2005)

18. Campbell, P. J., P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J. W. Teague, A. Menzies, I. Goodhead, D. J. Turner, C. M. Clee, M. A. Quail, A. Cox, C. Brown, R. Durbin, M. E. Hurles, P. A. W. Edwards, G. R. Bignell, M. R. Stratton and P. A. Futreal: Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Publishing Group* 40, 722–729 (2008)

19. Cancer Genome Atlas Research Network: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068 (2008)

20. Reid, S., D. Schindler, H. Hanenberg, K. Barker, S. Hanks, R. Kalb, K. Neveling, P. Kelly, S. Seal, M. Freund, M. Wurm, S. D. Batish, F. P. Lach, S. Yetgin, H. Neitzel, H. Ariffin, M. Tischkowitz, C. G. Mathew, A. D. Auerbach and N. Rahman: Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. *Nat Genet* 39, 162–164 (2007)

21. Chin, E. L. H., C. da Silva and M. Hegde: Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. *BMC Genet* 14, 6 (2013)

22. McCourt, C. M., D. G. McArt, K. Mills, M. A. Catherwood, P. Maxwell, D. J. Waugh, P. Hamilton, J. M. O'Sullivan and M. Salto-Tellez: Validation of Next Generation Sequencing Technologies in Comparison to Current Diagnostic Gold Standards for BRAF, EGFR and KRAS Mutational Analysis. *PLoS ONE* 8, 69604 (2013)

23. Beier, D., P. Hau, M. Proescholdt, A. Lohmeier, J. Wischhusen, P. J. Oefner, L. Aigner, A. Brawanski, U. Bogdahn and C. P. Beier: CD133(+) and CD133(-)

glioblastoma-derived cancer stem cells show differential growth characteristics and molecular profiles. *Cancer Research* 67, 4010–4015 (2007)

24. Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein and M. Snyder: The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1349 (2008)

25. Lottaz, C., D. Beier, K. Meyer, P. Kumar, A. Hermann, J. Schwarz, M. Junker, P. J. Oefner, U. Bogdahn, J. Wischhusen, R. Spang, A. Storch and C. P. Beier: Transcriptional profiles of CD133+ and CD133- glioblastoma-derived cancer stem cell lines suggest different cells of origin. *Cancer Research* 70, 2030–2040 (2010)

26. Yan, X., L. Maa, D. Yia, J.-G. Yoonb, A. Diercksa, B. Gregory Foltza, N. D. Pricec, Leroy E Hooda and Q. Tian: A CD133-related gene expression signature identifies an aggressive glioblastoma subtype with excessive mutations. *Proc Natl Acad Sci USA* 1–13 (2011)

27. Engstrom, P. G., D. Tommei, S. H. Stricker, C. Ender, S. M. Pollard and P. Bertone: Digital transcriptome profiling of normal and glioblastoma-derived neural stem cells identifies genes associated with patient survival. *Genome Med* 4, 76 (2012)

28. Marioni, J. C., 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, 1509–1517 (2008)

29. Oshlack, A. and M. J. Wakefield: Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 4, 14 (2009)

30. Hansen, K. D., S. E. Brenner and S. Dudoit: Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* 38, 131 (2010)

31. Dillies, M.-A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézicon behalf of The French StatOmique Consortium: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics* (2012)

32. Martin, J. A. and Z. Wang: Next-generation transcriptome assembly. *Nature Reviews Genetics* 12, 671–682 (2011)

33. Anders, S., D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber and M. D. Robinson: Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* 8, 1765–1786 (2013)

34. Chatterjee, A., P. A. Stockwell, E. J. Rodger and I. M. Morison: Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Research* 40, 79 (2012)

35. Clark, C., P. Palta, C. J. Joyce, C. Scott, E. Grundberg, P. Deloukas, A. Palotie and A. J. Coffey: A comparison of the whole genome approach of MeDIP-seq to the targeted approach of the Infinium HumanMethylation450 BeadChip for methylome profiling. *PLoS ONE* 7, 50233 (2012)

36. Kim, K.-J., K.-H. Lee, H.-S. Kim, K.-S. Moon, T.-Y. Jung, S. Jung and M.-C. Lee: The presence of stem cell marker-expressing cells is not prognostically significant in glioblastomas. *Neuropathology* 31, 494–502 (2011)

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

38. Metellus, P., I. Nanni-Metellus, C. Delfino, C. Colin, A. Tchogandjian, B. Coulibaly, F. Fina, A. Loundou, M. Barrie, O. Chinot, L. Ouafik and D. Figarella-Branger: Prognostic impact of CD133 mRNA expression in 48 glioblastoma patients treated with concomitant radiochemotherapy: a prospective patient cohort at a single institution. *Ann Surg Oncol* 18, 2937–2945 (2011)

39. Georgantas, R. W., V. Tanadve, M. Malehorn, S. Heimfeld, C. Chen, L. Carr, F. Martínez-Murillo, G. Riggins, J. Kowalski and C. I. Civin: Microarray and serial analysis of gene expression analyses identify known and novel transcripts overexpressed in hematopoietic stem cells. *Cancer Research* 64, 4434–4441 (2004)

40. Merlos-Suárez, A., F. M. Barriga, P. Jung, M. Iglesias, M. V. Céspedes, D. Rossell, M. Sevillano, X. Hernando-Momblona, V. da Silva-Diz, P. Muñoz, H. Clevers, E. Sancho, R. Mangues and E. Batlle: The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* 8, 511–524 (2011)

41. Liu, J. C., V. Voisin, G. D. Bader, T. Deng, L. Pusztai, W. F. Symmans, F. J. Esteva, S. E. Egan and E. Zacksenhaus: Seventeen-gene signature from enriched Her2/Neu mammary tumor-initiating cells predicts clinical outcome for human HER2+:ERα- breast cancer. *Proc Natl Acad Sci USA* 109, 5832–5837 (2012)

42. Schwede, M., D. Spentzos, S. Bentink, O. Hofmann, B. Haibe-Kains, D. Harrington, J. Quackenbush and A. C. Culhane: Stem cell-like gene expression in ovarian cancer predicts type II subtype and prognosis. *PLoS ONE* 8, 57799 (2013)

43. Paik, S., S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant and N. Wolmark: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 351, 2817–2826 (2004)

44. Reinhart, B. J., F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz and G. Ruvkun: The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* 403, 901–906 (2000)

45. Plummer, P. N., R. Freeman, R. J. Taft, J. Vider, M. Sax, B. A. Umer, D. Gao, C. Johns, J. S. Mattick, S. D. Wilton, V. Ferro, N. A. J. McMillan, A. Swarbrick, V. Mittal and A. S. Mellick: MicroRNAs regulate tumor angiogenesis modulated by endothelial progenitor cells. *Cancer Research* 73, 341–352 (2013)

46. Lang, M.-F., S. Yang, C. Zhao, G. Sun, K. Murai, X. Wu, J. Wang, H. Gao, C. E. Brown, X. Liu, J. Zhou, L. Peng, J. J. Rossi and Y. Shi: Genome-Wide Profiling Identified a Set of miRNAs that Are Differentially Expressed in Glioblastoma Stem Cells and Normal Neural Stem Cells. *PLoS ONE* 7, 36248 (2012)

47. Creighton, C. J., J. G. Reid and P. H. Gunaratne: Expression profiling of microRNAs by deep sequencing. *Briefings in Bioinformatics* 10, 490–497 (2009)

48. Alon, S., F. Vigneault, S. Eminaga, D. C. Christodoulou, J. G. Seidman, G. M. Church and E. Eisenberg: Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Res* 21, 1506–1511 (2011)

49. Memczak, S., M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. le Noble and N. Rajewsky: Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338 (2013)

50. Jeck, W. R., J. A. Sorrentino, K. Wang, M. K. Slevin, C. E. Burd, J. Liu, W. F. Marzluff and N. E. Sharpless: Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19, 141–157 (2013)

51. Salzman, J., C. Gawad, P. L. Wang, N. Lacayo and P. O. Brown: Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE* 7, 30733 (2012)

52. Hansen, T. B., T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard and J. Kjems: Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388 (2013)

53. Navin, N., J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks and M. Wigler: Tumour evolution inferred by single-cell sequencing. *Nature* 472, 90–94 (2011)

54. Ramsköld, D., S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth and R. Sandberg: Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* 30, 777–782 (2012)

55. Zong, C., S. Lu, A. R. Chapman and X. S. Xie: Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338, 1622–1626 (2012)

56. Dalerba, P., T. Kalisky, D. Sahoo, P. S. Rajendran, M. E. Rothenberg, A. A. Leyrat, S. Sim, J. Okamoto, D. M. Johnston, D. Qian, M. Zabala, J. Bueno, N. F. Neff, J. Wang, A. A. Shelton, B. Visser, S. Hisamori, Y. Shimono, M. van de Wetering, H. Clevers, M. F. Clarke and S. R. Quake: Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 29, 1120–1127 (2011)

57. Liao, Y., G. K. Smyth and W. Shi: The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research* 41, 108 (2013)

58. Pareek, C. S., R. Smoczynski and A. Tretyn: Sequencing technologies and genome sequencing. *J Appl Genet* 52, 413–435 (2011)

59. van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536 (2002)

60. Sørlie, T., C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning and A. L. Børresen-Dale: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 98, 10869–10874 (2001)

61. Shai, R., T. Shi, T. J. Kremen, S. Horvath, L. M. Liau, T. F. Cloughesy, P. S. Mischel and S. F. Nelson: Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene* 22, 4918–4923 (2003)

62. Verhaak, R. G. W., K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, D. N. HayesCancer Genome Atlas Research Network: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110 (2010)

63. Zorn, K. K., T. Bonome, L. Gangi, G. V. R. Chandramouli, C. S. Awtrey, G. J. Gardner, J. C. Barrett, J. Boyd and M. J. Birrer: Gene expression profiles of serous, endometrioid, and clear cell subtypes of ovarian and endometrial cancer. *Clin Cancer Res* 11, 6422–6430 (2005)

64. Phillips, H. S., S. Kharbanda, R. Chen, W. F. Forrest, R. H. Soriano, T. D. Wu, A. Misra, J. M. Nigro, H. Colman, L. Soroceanu, P. M. Williams, Z. Modrusan, B. G. Feuerstein and K. Aldape: Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 9, 157–173 (2006)

65. Curtis, C., S. P. Shah, S. Chin, G. Turashvilli, O. M. Rueda, M.J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Graf, G. Ha, G. Haffari, A. Bashashati, R.Russel, S. McKinney, METABRIC group, A. Langerod, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A. Borresen-Dale, J. D. Brenton, S. Tarvare, G. Caldas and S. Aparicio: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature 486, 346-352* (2012)
.
66. Cheng, W.-Y., J. J. Kandel, D. J. Yamashiro, P. Canoll and D. Anastassiou: A multi-cancer mesenchymal transition gene expression signature is associated with prolonged time to recurrence in glioblastoma. *PLoS ONE* 7, 34705 (2012)

67. Andreopoulos, B. and D. Anastassiou: Integrated Analysis Reveals hsa-miR-142 as a Representative of a Lymphocyte-Specific Gene Expression and Methylation Signature. *Cancer Inform* 11, 61–75 (2012)

68. Gao, J., B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander and N. Schultz: Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 6, l1 (2013)

69. Cerami, E., J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander and N. Schultz: The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov* 2, 401–404 (2012)

70. Kong, L., K.-L. Aho, K. Granberg, R. Lund, L. Järvenpää, J. Seppälä, P. Gokhale, K. Leinonen, L. Hahne, J. Mäkelä, K. Laurila, H. Pukkila, E. Närvä, O. Yli-Harja, P. W. Andrews, M. Nykter, R. Lahesmaa, C. Roos and R. Autio: ESTOOLS Data atHand: human stem cell gene expression resource. *Nature Publishing Group* 10, 814–815 (2013)

71. ENCODE Project Consortium: The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636–640 (2004)

72. ENCODE Project Consortium, B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter and M. Snyder: An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012)

73. Stricker, S. H., A. Feber, P. G. Engstrom, H. Carén, K. M. Kurian, Y. Takashima, C. Watts, M. Way, P. Dirks, P. Bertone, A. Smith, S. Beck and S. M. Pollard: Widespread resetting of DNA methylation in glioblastoma-initiating cells suppresses malignant cellular behavior in a lineage-dependent manner. *Genes & Development* 27, 654–669 (2013)

74. Gerlinger, M., A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal and C. Swanton: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 366, 883–892 (2012)

75. Xu, X., Y. Hou, X. Yin, L. Bao, A. Tang, L. Song, F. Li, S. Tsang, K. Wu, H. Wu, W. He, L. Zeng, M. Xing, R. Wu, H. Jiang, X. Liu, D. Cao, G. Guo, X. Hu, Y. Gui, Z. Li, W. Xie, X. Sun, M. Shi, Z. Cai, B. Wang, M. Zhong, J. Li, Z. Lu, N. Gu, X. Zhang, L. Goodman, L. Bolund, J. Wang, H. Yang, K. Kristiansen, M. Dean, Y. Li and J. Wang: Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 148, 886–895 (2012)

76. Buganim, Y., D. A. Faddah, A. W. Cheng, E. Itskovich, S. Markoulaki, K. Ganz, S. L. Klemm, A. van Oudenaarden and R. Jaenisch: Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* 150, 1209–1222 (2012)

77. Cibulskis, K., M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander and G. Getz: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 31, 213–219 (2013)

78. Bradford, J. R., M. Farren, S. J. Powell, S. Runswick, S. L. Weston, H. Brown, O. Delpuech, M. Wappett, N. R. Smith, T. H. Carr, J. R. Dry, N. J. Gibson and S. T. Barry: RNA-Seq Differentiates Tumour and Host mRNA Expression Changes Induced by Treatment of Human Tumour Xenografts with the VEGFR Tyrosine Kinase Inhibitor Cediranib. *PLoS ONE* 8, 66003 (2013)

79. Valdes, C., P. Seo, N. Tsinoremas and J. Clarke: Characteristics of cross-hybridization and cross-alignment of expression in pseudo-xenograft samples by RNA-Seq and microarrays. *J Clin Bioinforma* 3, 8 (2013)

80. Kim, J. K. and J. C. Marioni: Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol* 14, R7 (2013)

81. Tang, F., C. Barbacioru, S. Bao, C. Lee, E. Nordman, X. Wang, K. Lao and M. A. Surani: Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* 6, 468–478 (2010)

82. Xue, Z., K. Huang, C. Cai, L. Cai, C.-Y. Jiang, Y.

Feng, Z. Liu, Q. Zeng, L. Cheng, Y. E. Sun, J.-Y. Liu, S. Horvath and G. Fan: Genetic programs in human and mouse early embryos revealed by single-cell RNA□sequencing. *Nature* 500, 593-597 (2013)

83. Chen, J., H. Liu, J. Liu, J. Qi, B. Wei, J. Yang, H. Liang, Y. Chen, J. Chen, Y. Wu, L. Guo, J. Zhu, X. Zhao, T. Peng, Y. Zhang, S. Chen, X. Li, D. Li, T. Wang and D. Pei: H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. *Nature Publishing Group* 45, 34–42 (2012)

84. Rais, Y., A. Zviran, S. Geula, O. Gafni, E. Chomsky, S. Viukov, A. A. Mansour, I. Caspi, V. Krupalnik, M. Zerbib, I. Maza, N. Mor, D. Baran, L. Weinberger, D. A. Jaitin, D. Lara-Astiaso, R. Blecher-Gonen, Z. Shipony, Z. Mukamel, T. Hagai, S. Gilad, D. Amann-Zalcenstein, A. Tanay, I. Amit, N. Novershtern and J. H. Hanna: Deterministic direct reprogramming of somatic cells to pluripotency. *Nature* (2013)

Key Words: Stem Cell, Cancer, Cancer Stem Cells, Gene Expression, Next Generation Sequencing, Systems Biology, Bioinformatics, Review

Send correspondence to: Theo Mantamadiotis, Department of Pathology, The University of Melbourne, Parkville, VIC 3010, Australia, Tel: 61383445861, Fax: 6138344004, E-mail: theom@unimelb.edu.au