Published in final edited form as:

Expert Opin Drug Discov. 2012 March ; 7(3): 231–243. doi:10.1517/17460441.2012.660145.

Third Generation Sequencing Techniques and Applications to Drug Discovery

Fatih Ozsolak, Ph.D.[Senior Scientist]

Helicos BioSciences Corporation, One Kendall Square, Cambridge, MA, 02139, Phone: 617-264-1840, Fax: 617-264-1700, fatihozsolak@gmail.com

Abstract

Introduction—There is an immediate need for functional and molecular studies to decipher differences between disease and “normal” settings to identify large quantities of validated targets with the highest therapeutic utilities. Furthermore, drug mechanism of action and biomarkers to predict drug efficacy and safety need to be identified for effective design of clinical trials, decreasing attrition rates, regulatory agency approval process and drug repositioning. By expanding the power of genetics and pharmacogenetics studies, next generation nucleic acid sequencing technologies have started to play an important role in all stages of drug discovery.

Areas covered—This article reviews the first and second generation sequencing technologies (SGSTs) and challenges they pose to biomedicine. The article then focuses on the emerging third generation sequencing technologies (TGSTs), their technological foundations and potential contributions to drug discovery.

Expert Opinion—Despite the scientific and commercial success of SGSTs, the goal of rapid, comprehensive and unbiased sequencing of nucleic acids has not been achieved. TGSTs promise to increase sequencing throughput and read lengths, decrease costs, run times and error rates, eliminate biases inherent in SGSTs, and offer capabilities beyond nucleic acid sequencing. Such changes will have positive impact in all sequencing applications to drug discovery.

Keywords

Drug discovery; Epigenetics; Genetics; Next generation sequencing; Pharmacogenetics; Third generation sequencing

1. Introduction

The completion of the first draft of human genome sequence ten years ago, subsequent developments in “omics” technologies and their application to drug discovery led to an initial wave of “genomics drugs” entering the clinical world [1]. Despite these recent successes, the sequencing of the human genome did not bring the much anticipated revolution and productivity rise in drug discovery. Instead, it underlined our limited knowledge of the biology of diseases and the need for extensive functional studies to decipher connections and alterations between disease and “normal” settings using the tools of genomics, epigenomics, transcriptomics, metabolomics and proteomics in an integrated manner to identify large quantities of validated targets with therapeutic utility. Challenges in drug discovery extend well beyond discovery and validation of targets. Drug mechanism of action and biomarkers to predict drug efficacy and safety need to be identified for the purposes of effective design of clinical trials, decreasing attrition rates, regulatory agency approval process and drug repositioning to other disease conditions [2, 3].

Today, drug companies face diminishing research and development productivity, increasingly strict regulatory requirements, key patent expirations and payers requiring decreased costs [3]. The pressures due to these challenges are rapidly increasing and already at a point which prompted several prominent thought-leaders to question the survival of the drug industry [4]. To maintain profitability, drug companies need to decrease attrition rates, reduce time and costs of clinical trials, form a healthy drug pipeline and develop processes to maximize success and efficiency in each phase of the drug discovery process, from the early target-to-hit phase to product launch and postmarketing surveillance. Due to the enabling nature of technology in drug discovery, benefiting from the latest technological advances is necessary to maximize drug development efficiency.

During the last decade, microarray technologies have become integrated into the drug development process primarily due to their ability to monitor global gene expression levels and to enable investigations of the causative genetic states through genome-wide association studies (GWAS) [5]. Origins of systems approaches in drug discovery owe much of their progress to the comprehensive profiles generated with microarray-based technologies [6]. With the introduction of second-generation sequencing technologies (SGSTs) in 2005 [7], the domination of microarrays came to an end due to SGSTs' superior technical performance and cost advantages. This review will briefly summarize the current state of SGSTs, and focus on third generation sequencing technologies (TGSTs) and benefits they may bring to drug discovery process over existing technologies.

2. Second Generation Sequencing Technologies (SGSTs) and Applications

The era of first generation sequencing dominated by Sanger sequencing [8] quickly came to an end with the introduction of the Roche/454 platform [7] in 2005, followed by Solexa/Illumina system [9] in 2006, Applied Biosystems SOLiD system [10] in 2007, Complete Genomics [11] human genome sequencing service in 2008 and Ion Torrent system [12] in 2010 (Table 1). These SGSTs generate hundreds of thousands to billions of 25–800 nucleotide-long reads within days in a low-cost manner compared to Sanger sequencing. Following a shotgun approach, several of these technologies today can re-sequence a human genome in 10–30× coverage for under \$5,000 in reagent costs in 8–10 days (Figure 1). Recent reviews describe the technological foundations and working principles of these technologies [13, 14].

SGSTs have enjoyed a warm welcome from the academic and industrial scientists. With such an enthusiastic customer base and investments from the users and SGST companies in technology and applications development, many novel capabilities have been achieved within a very short time frame after their launch. SGST applications in drug discovery have primarily been whole genome sequencing, targeted (e.g. exome) sequencing, digital gene expression, DNA methylation sequencing and copy-number variation, and have been expanding to chromatin immunoprecipitation sequencing (ChIP-Seq), RNA immunoprecipitation sequencing (RIP-Seq), whole transcriptome analyses, single cell analyses, nucleic acid structure determination, chromatin conformation (Hi-C) analysis and many others. SGSTs technologies have also been adapted beyond nucleic acid sequencing and, for instance, to protein-DNA affinity measurements [15], which may find applications in drug discovery in the future. The details of each of these applications are beyond the scope of this review and have been reviewed elsewhere [16, 17].

SGSTs employ various chemistries and detection mechanisms (such as luminescence- and fluorescent-based detection). The common principle is their need for clonally amplified molecules for sequencing signal detection because of their lack of single molecule sensitivity. Clonal amplification imposes restrictions on the sample properties (fragment

length, adaptor sequences and location, etc.) for optimal sequencing performance, necessitating multiple nucleic acid manipulations (ligation, amplification etc.) before and after sample loading to the sequencer. As sequencing cycles progress, the clonal strands get out of phase with one another due to chemistry inefficiencies. The result is limited read lengths, higher error rates (compared to Sanger sequencing), and below optimal sequencing efficiency [18]. These challenges prevent unbiased data to be obtained and complicate downstream analyses (Figure 2). One challenge is the lack of *de novo* genome assembly capability and even complete coverage with SGSTs. SGSTs can cover 94–98% of the human genome. The coverage is limited by repeat element fraction, type and length, and coverage biases (originating from GC content, nucleic acid secondary structure and other factors). The sample preparation steps like amplification and reverse transcription introduce various artifacts [14, 19], such as generation of artificial “fused” transcripts due to template switching that complicate the identification of genomic and transcriptomic rearrangements [20]. The relatively high error rate of SGST platforms impose restrictions on rare cell mutation profiling studies [21], which is particularly relevant for biomarker discovery and target validation studies. The error rates and read length challenges also place limitations on species differentiation. Many experimental systems and samples used for drug discovery contain sequences from multiple species, such as human cancer xenografts in mouse and bacterial or other cell types in human specimens [22]. SGST capabilities impose limitations on the separation of true human signal from other species. The high cost of obtaining and operating SGSTs and the required infrastructure prevents their wide-spread adaptation to drug discovery, other biomedical areas and diagnostics. Furthermore, the required sample manipulations often necessitate the purchase of additional equipment and automation infrastructure. The recent introduction of lower-cost versions of these sequencers (MiSeq from Illumina and Ion Torrent from Life Technologies) are steps in the right direction, but these come at the expense of throughput and other capabilities of higher end platforms. Among the SGSTs, Ion Torrent perhaps attracts the most attention today [12]. It uses clonally-amplified DNA templates and multiplexed ion sensors to detect the release of hydrogen atoms as nucleotides are incorporated in the sequencing-by-synthesis reaction. Ion Torrent technology may be considered as a TGST because of its electronic signal detection capability. Because of its simplicity, semiconductor-based detection system, low reagent/hardware cost and potential rapid throughput improvement, it is expected to achieve the \$1000 (and perhaps \$100) genome in the near future.

3. Third Generation Sequencing Technologies (TGSTs)

Despite the advances in SGSTs, the goal of rapid, comprehensive and unbiased sequencing of DNA has not been achieved. The available technologies rely on multiple manipulation steps to covert natural DNA in a form that can be detected using electrical or chemical signals by various sensing mechanisms. It is now clear that each manipulation introduced to DNA causes artifacts and inaccuracies in DNA measurements. Therefore, novel technologies in development today, such as nanopore-based systems (discussed below), are being designed in ways that require minimal or no manipulation of natural DNA molecules.

Just like DNA, ability to quickly and accurately quantify and decode the sequence composition of all cellular RNA molecules directly without reverse transcription or other manipulations has long been desired. The initial attempts on determining RNA sequence in the 1970s by Helen Donnis-Keller, Alan M. Maxam and Walter Gilbert did not involve intermediate complementary DNA (cDNA) synthesis or other steps, and relied on the tendency of various RNases to cut RNA molecules at certain nucleotides [23]. Subsequent evolution of technologies such as quantitative real-time PCR, Sanger or SGSTs required an ensemble of molecules for signal detection purposes because of their lack of single molecule sensitivity. This limitation necessitated the generation of thousands of copies of each RNA

molecule prior to measurements. Because molecular mechanisms that can enable RNA amplification was not known or at least not experimentally feasible, conversion of RNA to cDNA and cDNA amplification became the standard steps for almost all RNA analysis approaches available today.

The ideal nucleic acid sequencing technology should: i) sequence natural DNA/RNA fragments of various sizes fully (as short as a few bases to as long as several hundred megabases) with minimal (or no) errors, sample manipulation and run time for detection of haplotype phasing on a chromosome scale; ii) generate no sequence-dependent or other biases for optimal quantification and genotyping performance; iii) require minimum quantities of input material, ideally single cells; iv) have high/complete efficiency in all of the sequencing steps (e.g. target DNA/RNA capture to the sequencing system and sequencing process); v) detect modified nucleotides (such as methylated cytosines) directly; vi) operate in laboratories and low-resource settings without special external environment requirements such as temperature and humidity control; and vii) be sufficiently affordable and economically viable to allow worldwide adoption and operation in research and clinical settings. With the introduction of the SGSTs and subsequent rapid improvement in capabilities, we are now closer, but still far away, from the ultimate goal of rapid, comprehensive and unbiased sequencing of DNA/RNA molecules. Emerging single molecule TGSTs may provide at least partial solutions to the challenges posed by SGSTs.

3.1 Single Molecule DNA Sequencing (SMDS) and Direct RNA Sequencing (DRS) Technologies

With the recent advances in automated single molecule imaging and fluidics technologies, the first commercially available TGST was launched by Helicos BioSciences, which paved the way towards the ability to analyze unmanipulated, natural DNA and RNA molecules in a massively parallel manner with the Single Molecule DNA Sequencing (SMDS) [24] and Direct RNA Sequencing (DRS) [25, 26] technologies. SMDS and DRS are based on the sequencing-by-synthesis system developed by the Stephen R. Quake laboratories at Caltech and Stanford a decade ago [27]. The system is the first and currently the only system that can sequence RNA directly (e.g. without cDNA synthesis) in addition to DNA and cDNA molecules. The sequencing system's flow cells are coated with poly(dT) oligonucleotides (Figure 3A). The requirement for DNA and RNA templates is the presence of 3' poly(A) tail that is "blocked" at its 3' terminus with a 3' deoxy nucleotide. 3' polyadenylation and blocking is performed with A-tailing enzymes. For the characterization of RNA species that naturally contain a poly(A) tail, such as messenger RNAs, additional polyA-tailing may not be required. After template capture to flow cells, the sequencing-by-synthesis is initiated with incubating the templates with one of the four Virtual Terminator™ (VT) nucleotide analogs [28] and polymerase in a cyclical manner, with alternating VTs in each cycle. VTs contain a fluorescent dye and chemically-cleavable groups that prevent the addition of another nucleotide in their intact form until the single molecule images are taken and a cleavage solution is introduced at the end of each cycle. Repeating addition-imaging cycles many times provides a large set of images from which the sequence information is generated for each individual molecule. The system allows up to 50 channels to be sequenced in parallel, each channel having the capacity to generate up to 30 million aligned reads. Actual yield depends on the run time (2–9 days), number of channels and throughput requested by the user.

The system's single molecule nature and low input DNA/RNA requirements (400–1000 attomoles) are particularly advantageous for settings with limited cell/nucleic acid quantities [29, 30]. Its avoidance of extensive manipulation steps result in better coverage, minimal or no GC bias and higher quantification and representation range for DNA and RNA compared to SGSTs, even in ancient, formalin-fixed paraffin-embedded tissue (FFPET) and other

highly degraded samples [31]. DRS today is the only technology that can avoid well-known reverse transcription and sample manipulation biases for RNAs. One advantage of DRS is the universality of sample preparation steps for different applications. In other words, unlike cDNA methods which require different cDNA synthesis and sample manipulation steps for short and long RNAs, DRS requires only 3' polyadenylated templates. Thus, both short and long RNAs can be sequenced together in a single experiment. The system's short read lengths, up to 55 nucleotides (nts), lack of paired read capability and error rate (3–4%) poses challenges in certain applications such as alternative splicing detection, microbiome sequencing and *de novo* genome sequencing. These challenges along with its high cost are perhaps the main reasons that have limited Helicos' commercial success. The system may be altered to enable elevated read lengths and paired reads in theory through minor modifications, but the errors may not be improved significantly, because the majority of the errors are caused by dye blinking during the imaging step and there is no apparent remedy for this issue. The system is capable of single step target selection and sequencing [32], which provides advantages in targeted sequencing and diagnostics applications focusing on gene sequencing. A modified version of the system has been extended to proteomics [33].

3.2 Single Molecule Real Time (SMRT) Sequencing

The origins of the SMRT technology go back to the mid 1990s, when Steve Turner, a physicist, and Jonas Korlach, a biochemist, at Cornell University hypothesized observing the real-time incorporation of fluorescent nucleotides into growing strands of DNA via a polymerase as an alternative way to rapidly sequence DNA. The challenge of observing fast incorporation kinetics of polymerases at a single molecule scale and in a multi-color fashion was resolved with the integration semiconductor processing, photonics and biochemistry fields and the development of the zero-mode waveguide (ZMW) technology [34]. ZMWs are microwells that are nanometers in diameter and placed in a metal film on a silica surface (Figure 3B). Due to the light travel behavior in small wells, illumination of the bottom of the ZMW allows detection of single molecules which are near the bottom of the well without significant interference from upper parts of the well. In the SMRT technology [35], a single polymerase molecule is immobilized to the bottom glass surface of the well in the imaged area. During sequencing cycles, the DNA templates to be sequenced are loaded to polymerases, followed by polymerase-driven incorporation of fluorescently-labeled nucleotides and observation of this event in real-time by the ZMW imaging system. The SMRT nucleotide analogs contain distinct fluorescent dyes that are attached to their phosphate groups. When the nucleotides are incorporated by the polymerase, the fluorescent dye is cleaved with the phosphate group, leaving a natural growing chain of nucleotides hybridized to the template. Through diffusion, the nucleotides find the polymerase-template complex and cleaved fluorescent dye leave the imaged area. Diffusion takes places on the order of microseconds, while the nucleotide incorporation by polymerase is in the order of milliseconds. This difference enables detection of nucleotide incorporation signal over the noise created by diffusing fluorescent groups. Currently, 75,000 ZMWs can be monitored simultaneously to obtain sequence information.

The system benefits from the natural high polymerase processivity, currently generating reads with a median length of 2000 nucleotides, with the potential to increase in the future. One of the limiting factors is the polymerase damage induced by the relatively high-strength lasers used for imaging. Perhaps polymerase evolution studies and chemistry improvements to stabilize and protect the polymerase may alleviate such issues. The technology requires several sample preparation steps and minimal sequencing reagent quantities, and run times are on the order of 30–60 minutes. One challenge is the limitations imposed on SMRT by the current state of the optical imaging technologies, which cannot observe polymerase-driven nucleotide incorporation at the natural speed (up to 1000 nucleotides per second).

This limitation requires polymerization to be slowed down to up to 5 nucleotides per second. Imaging technologies also limit the number of ZMWs that can be imaged per run in an economically feasible manner. The company estimates that improvements in imaging technologies by 2015 may allow company to create a second version of the system that can observe more than a million ZMWs per run. Because of this latter limitation, the system currently has limited throughput (~40–90 megabases of useful sequence per run), presenting challenges particularly for the analysis of species with medium- and large-sized genomes, such as human and mouse, and in quantification applications. Another challenge is the SMRT's error rates (15–20%) [35]. This level is the highest among the commercially available sequencing technologies today. It is thought to be due to the challenges of observing such fast single molecule dynamics in real time and is not expected to improve significantly in the near future. Template circularization and sequencing each template 10–15 times can reduce the error rates, but this comes at the expense of throughput, read lengths and restrictions on template sizes. Recent studies underlined the system advantages in infectious diseases and microbial pathogen identification due to its short run time and long reads [36, 37]. However, because these studies used cultured bacteria isolates, it is unclear how useful the technology with its current low throughput can be for actual clinical samples that tend to be a mixture of nucleic acids from human and other bacterial species at varying ratios. The hardware cost is high and require a controlled environment for operation. These properties will likely limit its application in resource-limited settings. The system has recently been used to identify methylated bases directly [38], but it is not clear if the technology can differentiate these and other modified bases comprehensively and directly in natural DNA samples. The technology has also been used to study ribosome-transfer RNA kinetics [39], and, at least in theory, can be extended to other applications, such as direct RNA sequencing.

3.3 Nanopore Sequencing Technologies

Nanopore sequencing generally refers to sequence determination as single molecules of nucleic acids electrophoretically pass through nanoscale pores (Figure 3C). The origins of nanopore-based sequencing approaches dates back to a landmark publication by the Deamer group at the University of California, Santa Cruz [40]. This study showed that, when a voltage bias was introduced across a nanopore, the electric field could drive single stranded nucleic acids through the nanopore and the passage could be detected as ionic current and duration changes across the pore. Many different implementations of this discovery are currently being pursued by academic and industrial scientists across the world to achieve nanopore sequencing. Among SGSTs and TGSTs available and in development today, nanopore technologies are particularly noteworthy because very low quantities of nucleic acid molecules can theoretically be analyzed from cells without any manipulation to obtain very long reads (perhaps megabases if not more) and sequence information can be obtained directly from RNA and even modified nucleotides at high efficiency and throughput in minimal time and cost. Much of nanopore's advantage comes from the fact that original DNA/RNA templates can be detected directly, unlike other technologies that rely on sequencing-by-synthesis and thus the copy of the template that is synthesized during sequencing is detected. The possibility that each nanopore can be used multiple times is another advantage for nanopores. Perhaps the nanopore's potential to be applicable beyond sequencing, in areas such as small molecule detection and biodefense [41, 42], is another reason fueling the much interest in this area. It is hard to determine how close the field is to providing a functional nanopore sequencing system and what capabilities nanopore sequencing offers today. There is only limited information disclosed by the companies operating in this area. Based on the published information, it is clear that challenges still lie ahead [43].

The current nanopore research activities focus on finding the optimal pore characteristics that allow controlled translocation and detection of nucleic acids. Nanopores can be constructed from biological or solid state materials, or a combination of both [44]. Much focus has been devoted to *Staphylococcus aureus* α -hemolysin, a transmembrane heptamer, pores formed in lipid bilayers [45], as pioneered by Dr. Hagan Bayley at the University of Oxford and pursued by Oxford Nanopore Technologies for commercialization. The stability and ease of production of α -hemolysin, along with the available structure information and ability to mutate the protein to achieve desired pore characteristics and interactions with nucleic acids are among the reasons for this choice. Recently however, the focus has been gradually shifting towards solid-state nanopores [46, 47], such as those made from graphene [48] and silicone nitride, because of production capabilities at better economies of scale and higher potential throughput compared to biological pores, along with the potentially better ability to control pore structure, such as thickness and width, and signal-to-noise ratio. Translocation speed of nucleic acids through pores can be controlled in a variety of ways, such as altered viscosity and salt concentrations [49], chemical modifications of the pore, and using a polymerase [50] or exonuclease [51] to direct translocation through the nanopore. Electrical nanopore signal detection is the desired mode of detection due to its lower-cost nature, simplicity and scale, but optical detection systems [52] have also been proposed. The latter approach, current being pursued by NobleGen, require prior conversion of nucleic acids to a fluorescently-labeled and “coded” form that can be detected optically as the fluorescent “code” passes through the pores. Hybridizing oligonucleotides of known sequences to single-stranded nucleic acids and detecting the location of such double-stranded regions with nanopores is an alternative way of sequence determination pursued by Nabsys. With support from Roche and NIH, IBM scientists are working on a “DNA transistor” solid-state nanopore system that controls DNA translocation and does sequence determination through voltage modulations [53, 54]. The chemistry-free nature of the system is appealing, although, according to available public information, proof of concept has not yet been achieved.

It is unclear which combination of these various nanopore approaches has the highest potential. It is likely that each combination will have advantages and disadvantages for each biomedical application. Among the nanopore-focused companies, Oxford Nanopore Technologies may be the closest to offer a commercialized system. With a highly modular and scalable hardware architecture that can be easily adjusted to meet the needs of customers from large genome centers to small laboratories and clinical centers, the system design seems to be able to fit with the needs of a large array of potential customers both in the developed as well as the developing world. Yet still, much of their success will depend on the differentiating attributes, sequencing performance and cost-benefit structure they can offer relative to existing SGSTs.

3.4 Other Potential TGSTs

Other TGSTs are currently being pursued by academic and industrial scientists, but still years away from commercial applications. One approach in development by the Block group at Stanford involves measuring RNA polymerase movement across DNA templates stretched between two optical traps in a reaction containing three nucleotides at high concentrations and one nucleotide at limiting concentrations [55]. This system has the resolution to detect polymerase movement in angstrom resolution. Movement of the polymerase will exhibit pauses as the limiting nucleotide is being added during *in vitro* transcription. Such pauses indicate the locations of the limiting nucleotide. By performing this reaction four times, with each of the four nucleotides being limiting in each reaction, sequencing information can be obtained. In their 2006 publication, the system could achieve

read lengths of 30 bases. The read lengths are probably limited by the polymerase processivity, which can be several thousand bases.

Companies including ZS Genetics, Halcyon Molecular and Reveo aim to sequence DNA by direct imaging using scanning tunneling or transmission electron microscopy. Proof-of-concept studies have not been published with these technologies, but these systems may be capable of achieving very long reads (up to millions of bases) with short sequencing run times.

4. TGST Applications to Drug Discovery

TGSTs promise to increase throughput and read lengths, decrease costs, run times and error rates, eliminate biases inherent in SSGT approaches, require minimal input material and offer capabilities beyond nucleic acid sequencing. Such changes will have positive impact in all applications of sequencing technology to drug discovery. For instance, epigenomics studies have been gaining ground in drug discovery. Because epigenetics profiles are cell-type dependent, not permanent (e.g. can be reversible) and are created by both genetic and environmental factors, exploring epigenetic mechanisms may help address a larger proportion of disease aetiology than genetic factors alone and lead to novel biomarkers and drugs. Epigenetics-based biomarkers (e.g. septin 9 methylation status for colorectal cancer from Epigenomics [56]) and drugs (e.g. Vorinostat from Merck, a histone deacetylase inhibitor, for the treatment of cutaneous T-cell lymphoma) are already available and more are in clinical trials particularly for cancer treatment [57, 58]. TGSTs may expedite drug discovery efforts in this area by not only allowing methylation status and other modified nucleotides to be detected directly, but also by enabling other epigenomic components, such as histone modifications and chromatin organization, to be studied in an integrated manner with genomic and transcriptomic mechanisms by extended use of assays such as ChIP and Hi-C [59].

Another area where TGSTs may contribute to drug discovery is by enabling observations at the single cell level. The need for understanding pharmacodynamic processes at the basic unit of life, namely cells, rather than on a population level is clear in drug discovery. One recent example showing the benefit of single cell studies involves determination of the gene expression mechanisms induced by three vaccines [60]. The differences in vaccine-induced responses could only be detected in single cell studies, not with traditional approaches measuring the average signal from large numbers of cells. SSGTs have been applied for single cell measurements [61, 62], but the field is very young and required cell manipulation steps introduce significant data quality concerns. With their single molecule sequencing capabilities and low input quantity requirements, TGSTs may provide a direct, unbiased and economical solution for single cell measurements of DNA and RNA molecules and enable the studies of rare cell populations of therapeutic value, such as circulating tumor cells [63].

With the realization that the majority of genes are modulated by noncoding RNAs (ncRNAs) and ncRNA role in disease [64, 65], drug companies, such as Miragen, Regulus and Santaris, are now focusing on ncRNA-based therapies. The importance of sequencing technologies in this area will only increase since sequencing is the primary mode of assay for detection and quantification in ncRNA-focused discovery efforts. With the long reads, high throughput and direct RNA sequencing capabilities of TGSTs, exploring ncRNA mechanisms, target identification and validation will be significantly expedited.

TGSTs may also be useful in aptamer-based drug discovery efforts. Aptamer field has already produced a treatment for age-related macular degeneration (Pegaptanib marketed by Pfizer, an anti-VEGF aptamer) and further efforts are in place from multiple drug companies. Aptamers are generally discovered using a procedure called SELEX (systemic

evolution of ligands by exponential enrichment). The procedure involves multiple cycles of selection of randomized DNA/RNA constructs based on their binding to targets of interests, followed by amplification of enriched nucleic acids to be used in the next enrichment cycle [66]. Because amplification/ligation events are known to be sequence-dependent and not preserve absolute quantities of nucleic acids, the use of TGSTs with their low input quantity requirement and direct sequencing capability may improve the success of SELEX screens and reduce the number of enrichment cycles needed [66].

Another opportunity for TGSTs lies in whole genome sequencing applications. Certain repeat regions in the genomes cannot be analyzed with SGSTs due to read length and sequencing bias limitations [67]. With TGSTs' potential read lengths up to megabases, access to such regions and complete *de novo* genome assemblies may finally be enabled. As the number of studies underlining repeat element role in disease and their value as biomarkers increases [68], TGSTs can be the much needed tool to studies focusing on the role of repeat elements in normal physiology and disease.

Human microbiota substantially affects many aspects of human physiology [69] and may regulate drug metabolism, absorption, tolerance and efficacy. Altered microbiota have been linked to obesity [70, 71] and other disorders [72]. SGSTs accelerated research in this area by allowing rapid analysis and identification of microbial communities in human body and other research settings through genome and, particularly, 16S ribosomal DNA/RNA sequence analyses. It has recently become clear that, due to amplification/ligation biases, artificial chimeric fragment formation, amplicon choice differences, read length restrictions and computational challenges, microbial sequencing studies are prone to errors in species determination [73–75] and have limited quantitative power and errors in measuring species abundance. TGST potential for long reads and direct sequencing may alleviate some of the difficulties in this field. TGSTs may also be important for the study of antibiotic mechanism of action, antibiotic effects to patients, and antibiotic resistance through microbial whole genome sequencing studies [76, 77].

5. Conclusion

In order for TGSTs' limitless potential in drug discovery to be realized, TGST companies and service providers need to achieve economical viability. From a business standpoint, TGST companies are at a disadvantaged position compared to the position of SGST companies in the early 2000s before the sequencing boom. With significant investments from the pharmaceutical and academic research facilities particularly during the past three years, the sequencing market may already be saturated with high-throughput SGST sequencers. Even major SGST providers expect much of their sales in the near future to be driven by the lower-cost benchtop SGST systems that are more suited to budget-conscious research environments and clinical diagnostic facilities. This transition is part of the trend towards realizing the fruits of genomics and pharmacogenomics studies and moving towards lower cost and greater access to sequencing and personalized medicine. However, this also means that the bar is set very high for TGSTs. The achievement and commercial viability of TGST companies and affect on drug discovery will depend on their ability to differentiate and offer highly demanded capabilities and economical advantages that cannot be matched by other sequencing technologies.

6. Expert Opinion

Second generation sequencing technologies (SGSTs) truly revolutionized many fields of biomedical sciences and drug discovery, and are about to bring much needed capabilities to clinical diagnostics. Major SGST platforms are expected to make sequencing of human

genomes in a day with only a few thousands of dollars in reagent costs possible within the year 2012. Considering the thirteen years and hundreds of millions of dollars it took to sequence the first human genome, the progress made in the last decade and the achievements of scientists and engineers in the sequencing technologies field are worth much admiration. Time will show that the dramatic advances we have seen particularly in the last five years mark only the beginning of the upcoming genomics revolution in medicine. Sequencing field will keep rapidly evolving with the frequent improvements in SGSTs and additional capabilities third generation sequencing technologies (TGST) promise to bring. With projected TGST advantages in costs and time-to-result, and applications beyond DNA/cDNA/RNA sequencing, sequencing will stop being a “luxury” item and enter all phases of drug discovery. Perhaps the biggest revolution in drug discovery will come from systems-based approaches aiming to integrate data from genomics and other disciplines to build comprehensive and predictive molecular models of disease and treatment pathways. The lack of sophisticated statistical algorithms and software solutions for SGST data analysis and integration with other dimensions of biological complexity such as proteomics or cell-based high content screening studies prevents us from gaining the full benefits of SGSTs. Even for GWAS (genome-wide association studies)-type approaches that has been in use for more than a decade and aim to find the connections between phenotype to genotype, computational approaches available does not address our needs, especially for the study of complex diseases caused by multiple genes with low penetrance and perhaps factors beyond genetics. The situation will only get worse as the data storage and analysis requirements increase with the vast quantities of data TGSTs are expected to bring, and as the applications of TGSTs extend to agriculture, biosafety, alternative biofuel and other research areas. The need and the resulting business opportunities have been recognized by multiple companies, including large established corporations such as Microsoft, IBM and Amazon, and non-profit organizations such as Sage Bionetworks. However, achieving such integrated in-depth multi-dimensional views of biological states will not be easy due to the variations in platforms used for each molecular assessment and the high variability in experimental and analysis needs in different areas and stages of drug discovery. Investing significant resources to establish and maintain in-house computational capabilities and developing custom solutions for the particular needs of each drug discovery unit, while not desirable, may be the only way, at least in the near term, for drug companies wishing to benefit from the fruits of sequencing and other rapidly evolving “omics” technologies.

Article Highlights

- Sequencing of the first human genome a decade ago underlined our limited knowledge of the biology of diseases and the need for extensive functional studies to decipher connections and alterations between disease and “normal” settings using the tools of genomics, epigenomics, transcriptomics, metabolomics and proteomics in an integrated manner to identify large quantities of validated targets with therapeutic utility.
- Second generation sequencing technologies (SGSTs) truly revolutionized many fields of biomedical sciences and drug discovery, and are about to bring much needed capabilities to clinical diagnostics. However, SGSTs still present multiple technological challenges.
- With the advantages of third generation sequencing technologies (TGSTs) in costs and time-to-result, and applications beyond DNA/cDNA/RNA sequencing, sequencing will stop being a “luxury” item and enter all phases of drug discovery.

- The achievement and commercial viability of TGST companies and affect on drug discovery will depend on their ability to differentiate and offer highly demanded capabilities and economical advantages that cannot be matched with other sequencing technologies.

Abbreviations

FGST	First Generation Sequencing Technology
SGST	Second Generation Sequencing Technology
TGST	Third Generation Sequencing Technology
cDNA	Complementary DNA
qPCR	Quantitative Polymerase Chain Reaction
SMRT	Single Molecule Real Time
SMDS	Single Molecule DNA Sequencing
DRS	Direct RNA Sequencing
SELEX	Systemic Evolution of Ligands by Exponential Enrichment
ncRNA	Non-coding RNA
GWAS	Genome-wide Association Studies
RIP-Seq	RNA Immunoprecipitation Sequencing
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
Hi-C	Chromosome Conformation Assay
VT	Virtual Terminator
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
ZMW	Zero-mode Waveguide
FFPET	Formalin-fixed Paraffin-embedded Tissue
NIH	National Institutes of Health
nt	Nucleotide
Mb	Megabase
Gb	Gigabase
miRNA	microRNA

Acknowledgments

The author apologizes to authors whose work could not be cited owing to space constraints.

Declaration of Interest

F Ozsolak is an employee of the Helicos Biosciences Corporation and is grateful to the US National Human Genome Research Institute for their grant support (R01 HG005230 and R44 HG00527).

References

- Hall J, Dennler P, Haller S, et al. Genomics drugs in clinical trials. *Nature reviews*. 2010; 9(12):988.
- Beckman RA, Clark J, Chen C. Integrating predictive biomarkers and classifiers into oncology clinical development programmes. *Nature reviews*. 2011; 10(10):735–748.
- Paul SM, Mytelka DS, Dunwiddie CT, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature reviews*. 2010; 9(3):203–214.
- Booth B, Zimmel R. Prospects for productivity. *Nature reviews*. 2004; 3(5):451–456.
- Debouck C, Goodfellow PN. DNA microarrays in drug discovery and development. *Nature genetics*. 1999; 21 1 Suppl:48–50. [PubMed: 9915501]
- Iskar M, Zeller G, Zhao XM, et al. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Current opinion in biotechnology*. 2011
- Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005; 437(7057):376–380. [PubMed: 16056220]
- Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*. 1975; 94(3):441–448. [PubMed: 1100841]
- Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008; 456(7218):53–59. [PubMed: 18987734]
- Valouev A, Ichikawa J, Tonthat T, et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome research*. 2008; 18(7):1051–1063. [PubMed: 18477713]
- Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science (New York, NY)*. 2010; 327(5961):78–81.
- Rothberg JM, Hinz W, Rearick TM, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011; 475(7356):348–352. [PubMed: 21776081]
- Metzker ML. Sequencing technologies - the next generation. *Nature reviews*. 2010; 11(1):31–46..
** A comprehensive technical review describing second generation sequencing technologies.
- Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nature reviews*. 2011; 12(2):87–98.
- Nutiu R, Friedman RC, Luo S, et al. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature biotechnology*. 2011; 29(7):659–664.
- Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet*. 2008; 24(3):133–141. [PubMed: 18262675]
- Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. *Genomics*. 2008; 92(5):255–264. [PubMed: 18703132]
- Linsen SE, de Wit E, Janssens G, et al. Limitations and possibilities of small RNA digital gene expression profiling. *Nature methods*. 2009; 6(7):474–476. [PubMed: 19564845]
- Mamanova L, Andrews RM, James KD, et al. FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nature methods*. 2010; 7(2):130–132. [PubMed: 20081834] . ** The manuscript exemplifies several of the shortcomings of commonly used molecular techniques in sequencing applications and proposes solutions.
- McManus CJ, Duff MO, Eipper-Mains J, Graveley BR. Global analysis of trans-splicing in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107(29):12975–12979. [PubMed: 20615941]
- Flaherty P, Natsoulis G, Muralidharan O, et al. Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic acids research*. 2011
- Lin MT, Tseng LH, Kamiyama H, et al. Quantifying the relative amount of mouse and human DNA in cancer xenografts using species-specific variation in gene length. *BioTechniques*. 2010; 48(3):211–218. [PubMed: 20359302]
- Donis-Keller H, Maxam AM, Gilbert W. Mapping adenines, guanines, and pyrimidines in RNA. *Nucleic acids research*. 1977; 4(8):2527–2538. [PubMed: 409999]
- Harris TD, Buzby PR, Babcock H, et al. Single-molecule DNA sequencing of a viral genome. *Science (New York, NY)*. 2008; 320(5872):106–109.

25. Ozsolak F, Kapranov P, Foissac S, et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*. 2010; 143(6):1018–1029. [PubMed: 21145465]
26. Ozsolak F, Platt AR, Jones DR, et al. Direct RNA sequencing. *Nature*. 2009; 461(7265):814–818. [PubMed: 19776739]
27. Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*. 2003; 100(7):3960–3964. [PubMed: 12651960]
28. Bowers J, Mitchell J, Beer E, et al. Virtual terminator nucleotides for next-generation DNA sequencing. *Nature methods*. 2009; 6(8):593–595. [PubMed: 19620973]
29. Goren A, Ozsolak F, Shores N, et al. Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nature methods*. 2010; 7(1):47–49. [PubMed: 19946276]
30. Ozsolak F, Ting DT, Wittner BS, et al. Amplification-free digital gene expression profiling from minute cell quantities. *Nature methods*. 2010; 7(8):619–621. [PubMed: 20639869]
31. Orlando L, Ginolhac A, Raghavan M, et al. True single-molecule DNA sequencing of a pleistocene horse bone. *Genome research*. 2011; 21(10):1705–1719. [PubMed: 21803858]
32. Thompson JF, Reifenger JG, Giladi E, et al. Single-step capture and sequencing of natural DNA for detection of BRCA1 mutations. *Genome research*. 2011
33. Tessler LA, Reifenger JG, Mitra RD. Protein quantification in complex mixtures by solid phase single-molecule counting. *Analytical chemistry*. 2009; 81(17):7141–7148. [PubMed: 19601620]
34. Levene MJ, Korlach J, Turner SW, et al. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science (New York, NY)*. 2003; 299(5607):682–686.
35. Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. *Science (New York, NY)*. 2009; 323(5910):133–138.
36. Chin CS, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. *The New England journal of medicine*. 2011; 364(1):33–42. [PubMed: 21142692]
37. Rasko DA, Webster DR, Sahl JW, et al. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. *The New England journal of medicine*. 2011; 365(8):709–717. [PubMed: 21793740]
38. Flusberg BA, Webster DR, Lee JH, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods*. 2010; 7(6):461–465. [PubMed: 20453866]
39. Uemura S, Aitken CE, Korlach J, et al. Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature*. 2010; 464(7291):1012–1017. [PubMed: 20393556]
40. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America*. 1996; 93(24):13770–13773. [PubMed: 8943010] . ** Groundbreaking work that set the stage for nanopore nucleic acid sequencing.
41. Guan X, Gu LQ, Cheley S, et al. Stochastic sensing of TNT with a genetically engineered pore. *Chembiochem*. 2005; 6(10):1875–1881. [PubMed: 16118820]
42. Wu HC, Bayley H. Single-molecule detection of nitrogen mustards by covalent reaction within a protein nanopore. *Journal of the American Chemical Society*. 2008; 130(21):6813–6819. [PubMed: 18444650]
43. Branton D, Deamer DW, Marziali A, et al. The potential and challenges of nanopore sequencing. *Nature biotechnology*. 2008; 26(10):1146–1153.
44. Hall AR, Scott A, Rotem D, et al. Hybrid pore formation by directed insertion of alpha-haemolysin into solid-state nanopores. *Nature nanotechnology*. 2010; 5(12):874–877.
45. Howorka S, Cheley S, Bayley H. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nature biotechnology*. 2001; 19(7):636–639.
46. Dekker C. Solid-state nanopores. *Nature nanotechnology*. 2007; 2(4):209–215.
47. Gershow M, Golovchenko JA. Recapturing and trapping single molecules with a solid-state nanopore. *Nature nanotechnology*. 2007; 2(12):775–779.
48. Garaj S, Hubbard W, Reina A, et al. Graphene as a subnanometre trans-electrode membrane. *Nature*. 2010; 467(7312):190–193. [PubMed: 20720538]

49. Kawano R, Schibel AE, Cauley C, White HS. Controlling the translocation of single-stranded DNA through alpha-hemolysin ion channels using viscosity. *Langmuir*. 2009; 25(2):1233–1237. [PubMed: 19138164]
50. Lieberman KR, Cherf GM, Doody MJ, et al. Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *Journal of the American Chemical Society*. 2010; 132(50):17961–17972. [PubMed: 21121604]
51. Clarke J, Wu HC, Jayasinghe L, et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotechnology*. 2009; 4(4):265–270.
52. McNally B, Singer A, Yu Z, et al. Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. *Nano letters*. 2000; 10(6):2237–2244. [PubMed: 20459065]
53. Luan B, Martyna G, Stolovitzky G. Characterizing and controlling the motion of ssDNA in a solid-state nanopore. *Biophysical journal*. 2011; 101(9):2214–2222. [PubMed: 22067161]
54. Luan B, Peng H, Polonsky S, et al. Base-by-base ratcheting of single stranded DNA through a solid-state nanopore. *Physical review letters*. 2010; 104(23):238103. [PubMed: 20867275]
55. Greenleaf WJ, Block SM. Single-molecule, motion-based DNA sequencing using RNA polymerase. *Science (New York, NY)*. 2006; 313(5788):801.
56. Grutzmann R, Molnar B, Pilarsky C, et al. Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. *PloS one*. 2008; 3(11):e3759. [PubMed: 19018278]
57. Delmore JE, Issa GC, Lemieux ME, et al. BET bromodomain inhibition as a therapeutic strategy to target c-Myc. *Cell*. 2011; 146(6):904–917. [PubMed: 21889194]
58. Kaiser J. Epigenetic drugs take on cancer. *Science (New York, NY)*. 2010; 330(6004):576–578.
59. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, NY)*. 2009; 326(5950):289–293.
60. Flatz L, Roychoudhuri R, Honda M, et al. Single-cell gene-expression profiling reveals qualitatively distinct CD8 T cells elicited by different gene-based vaccines. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108(14):5724–5729. [PubMed: 21422297]
61. Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. *Nature*. 2011; 472(7341):90–94. [PubMed: 21399628]
62. Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*. 2009; 6(5):377–382. [PubMed: 19349980]
63. Yu M, Stott S, Toner M, et al. Circulating tumor cells: approaches to isolation and characterization. *The Journal of cell biology*. 2011; 192(3):373–382. [PubMed: 21300848]
64. Esteller M. Non-coding RNAs in human disease. *Nature reviews*. 2011; 12(12):861–874.
65. Wahlestedt C. Natural antisense and noncoding RNA transcripts as potential drug targets. *Drug discovery today*. 2006; 11(11–12):503–508. [PubMed: 16713901]
66. Zimmermann B, Gesell T, Chen D, et al. Monitoring genomic sequences during SELEX using high-throughput sequencing: neutral SELEX. *PloS one*. 2010; 5(2):e9169. [PubMed: 20161784]
67. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews*. 2011
68. Ting DT, Lipson D, Paul S, et al. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science (New York, NY)*. 2011; 331(6017):593–596.
69. Li M, Wang B, Zhang M, et al. Symbiotic gut microbes modulate human metabolic phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105(6):2117–2122. [PubMed: 18252821]
70. Backhed F, Ding H, Wang T, et al. The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences of the United States of America*. 2004; 101(44):15718–15723. [PubMed: 15505215]
71. Vijay-Kumar M, Aitken JD, Carvalho FA, et al. Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science (New York, NY)*. 2010; 328(5975):228–231.

72. Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. *Genome medicine*. 2011; 3(3):14. [PubMed: 21392406]
73. Haas BJ, Gevers D, Earl AM, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research*. 2011; 21(3):494–504. [PubMed: 21212162]
74. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. Removing noise from pyrosequenced amplicons. *BMC bioinformatics*. 2011; 12:38. [PubMed: 21276213]
75. Zhang Q, Doak TG, Ye Y. Artificial functional difference between microbial communities caused by length difference of sequencing reads. *Pacific Symposium on Biocomputing*. 2012:259–270. [PubMed: 22174281]
76. Comas I, Borrell S, Roetzer A, et al. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nature genetics*. 2011
77. Toprak E, Veres A, Michel J, et al. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature genetics*. 2011
78. Levy S, Sutton G, Ng PC, et al. The diploid genome sequence of an individual human. *PLoS biology*. 2007; 5(10):e254. [PubMed: 17803354]
79. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*. 2008; 452(7189):872–876. [PubMed: 18421352]
80. Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. *Nature biotechnology*. 2009; 27(9):847–850.
81. Kim JI, Ju YS, Park H, et al. A highly annotated whole-genome sequence of a Korean individual. *Nature*. 2009; 460(7258):1011–1015. [PubMed: 19587683]
82. Ley TJ, Mardis ER, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*. 2008; 456(7218):66–72. [PubMed: 18987736]
83. Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. *Nature*. 2008; 456(7218):60–65. [PubMed: 18987735]
84. Li Y, Wang J. Faster human genome sequencing. *Nature biotechnology*. 2009; 27(9):820–821.

Figure 1. Reduction in the cost of human genome sequencing

Each circle area is proportional to the sequencing cost (not including labor, sample preparation, data storage and analysis). The figure does not include the recent advances brought by the Illumina HiSeq 2000 system, which can sequence 5–6 human genomes per two weeks at ~\$3000 per genome reagent costs. The first human genome and Craig Venter Genome was sequenced with the Sanger technology [78]. Jim Watson genome was sequenced with Roche/454 [79]. Pushkarev *et al.* genome was sequenced with the Helicos SMDS [80]. The remaining genomes were sequenced with the Illumina/Solexa technology [9, 81–83]. This figure was reprinted from [84] with permission from the Nature Publishing Group.

Figure 2. SGST quantification biases

473 human miRNA sequences were synthetically produced and combined at equimolar ratio. The miRNA abundance was measured with Life Technologies SOLiD, Illumina/Solexa sequencing and quantitative real-time Polymerase Chain Reaction (qPCR) analyses using standard or commercial sample preparation kits. The miRNA quantities measured were normalized relative to the median level and ranked from highest to lowest levels. The dashed line indicates the ideal case, where each measured miRNA quantity is the same. The lines mark the hsa-let-7a to hsa-let-7g and hsa-let-7i miRNA isoforms (represented by the labels a–g and i). The observed bias and wide distribution of miRNA abundances away from the ideal equimolar case can be due to a combination of biases introduced at various steps, such as cDNA synthesis, ligation, amplification and sequencing. This figure was reprinted from [18] with permission from the Nature Publishing Group.

Figure 3. TGST platforms

A) Helicos SMDS/DRS single-molecule sequencing. DNA/RNA (purple) isolated from cells is fragmented and tailed to produce polyA-tailed templates. The templates are hybridized to the flow cell surfaces coated with poly(dT) oligonucleotides (green), which also serve as primers. The sequencing process involves cyclically incorporating each of the four fluorescently labeled Virtual Terminator nucleotides with a polymerase, imaging the labeled nucleotides and, finally, cleaving off the fluorescent labels. This figure was reprinted from [84] with permission from the Nature Publishing Group. **B) Pacific Biosciences SMRT sequencing system.** The polymerase loaded with a template molecule to be sequenced (the white chain) is immobilized at the bottom of a ZMW microwell (the gray block). Real time laser excitation and emission is recorded from below the ZMW surface as the fluorescently-labeled nucleotides are being incorporated by the polymerase in the sequencing-by-synthesis reaction. This figure was reprinted from [35] with permission from the American Association for the Advancement of Sciences. **C) Nanopore Sequencing.** The graph shows the ionic current amplitude change between the “open pore” and the “blocked pore” states as the single stranded DNA molecule (yellow chain) passes through the α -hemolysin pore (pink, blue and white structure) formed on lipid bilayers. This figure was reprinted from [43] with permission from the Nature Publishing Group.

SGST and TGST Companies

Table 1

Throughput indicates the number of usable (e.g. aligned) bases per run. Run times are indicated in hours or days, and do not include strand formation, data download, sample preparation and run setup times. The information presented in this table was collected from company websites and announced public information. There is limited information available for the companies that are still in the development phase and have not released a commercial version of their technology. For such companies, the sequencing technology type listed in this table may be different than the actual technology used if/when their technology are commercialized. The information that is not available or could not be located is indicated as “NA” (not available). In cases where a technology offers a range of read length, throughput and/or run time options, the highest values were reported with the “up to” statement. For technologies offering paired read capability, “2X(Read Length)” notation was used to indicate lengths of each paired reads. “nt”: nucleotides; “Mb”: megabase; “Gb”: gigabase.

	Generation	Released	Sequencing Chemistry	Read Length	Throughput	Run Time	Raw Error Rate
Roche/454 GS FLX Titanium XL+	SGST	Yes	Pyrosequencing, Emulsion PCR	500–1000nt	700Mb	23 Hours	0.4–1.5%
Roche/454 GS Junior	SGST	Yes	Pyrosequencing, Emulsion PCR	400nt	35Mb	10 Hours	0.4–1.5%
Ion Torrent Chip 318	SGST	Yes	Sequencing by Synthesis, Emulsion PCR	200nt	1Gb	2 Hours	1–3%
Illumina HiSeq 2000	SGST	Yes	Sequencing by Synthesis, Reversible Terminator, Bridge PCR	up to 2X100nt	up to 600Gb	up to 11 Days	0.5–2%
Illumina MiSeq	SGST	Yes	Sequencing by Synthesis, Reversible Terminator, Bridge PCR	up to 2X150nt	up to 2Gb	up to 27 Hours	0.5–2%
Life Technologies 5500xl	SGST	Yes	Sequencing by Ligation, Emulsion PCR	2X60nt	up to 200Gb	up to 7 Days	0.20%
Gnubio	SGST	No	Sequencing by Hybridization, Amplification in Solution	NA	NA	NA	NA
Complete Genomics	SGST	Yes	Sequencing by Ligation, Rolling Circle Amplification in Solution	up to 2X35nt	NA	NA	0.5–1.5%
HellScope	TGST	Yes	Single Molecule Sequencing by Synthesis	up to 55nt	up to 35Gb	up to 8 Days	3–5%
Pacific BioSciences	TGST	Yes	Single Molecule Read Time Sequencing by Synthesis	up to 6000nt	up to 90Mb	1 Hour	13–15%
Oxford Nanopore Technologies	TGST	No	Nanopore Sequencing, Biological and Solid-State Pore, Electrical Readout	NA	NA	NA	NA
NobleGen	TGST	No	Nanopore Sequencing, Solid-State Pore, Optical Readout	NA	NA	NA	NA
Nabsys	TGST	No	Nanopore Sequencing, Solid-State Pore, Electrical Readout	NA	NA	NA	NA
Genia	TGST	No	Nanopore Sequencing, Biological Pore, Electrical Readout	NA	NA	NA	NA
Electronic BioSciences	TGST	No	Nanopore Sequencing, Biological and Solid-State Pore, Electrical Readout	NA	NA	NA	NA
ZS Genetics	TGST	No	Transmission Electron Microscopy	NA	NA	NA	NA
Halcyon Molecular	TGST	No	Transmission Electron Microscopy	NA	NA	NA	NA
Reveo	TGST	No	Scanning Tunneling Microscopy	NA	NA	NA	NA
IBM	TGST	No	Nanopore Sequencing, Solid-State Pore, Electrical Readout	NA	NA	NA	NA