Published in final edited form as:
Expert Rev Mol Diagn. 2009 October ; 9(7): 659–666. doi:10.1586/erm.09.50.

Emergence of single-molecule sequencing and potential for molecular diagnostic applications

Patrice M Milos, PhD
Helicos BioSciences, 1 Kendall Square, Building 700, Cambridge, MA 02139, USA, Tel.: +1 617 264 1800, Fax: +1 617 264 1700, pmilos@helicosbio.com

Abstract

The effective demonstration of single-molecule sequencing at scale over the last several years offers the exciting opportunity for a new era in the field of molecular diagnostics. As we aim to personalize and deliver cost-effective healthcare, we must consider the need to fully integrate genomics into decision-making. We must be able to accurately and cost effectively obtain a complete genome sequence for disease diagnosis, interrogate a molecular signature from blood for therapeutic monitoring, obtain a tumor mutation profile for optimizing therapeutic choice – each molecular diagnostic measurement utilized to better inform patient care. Would a physician or molecular pathology laboratory want to utilize a PCR process in which millions of DNA copies of a patient's nucleic acid are created when an alternative approach allowing direct measurement of the nucleic acids is possible? I would suggest not! In this article we will focus on the emergence of single-molecule sequencing, the single-molecule sequencing methodologies in the marketplace or under development today, as well as the importance of these methods for molecular characterization and diagnosis of disease with the ultimate application for molecular diagnostics.

Keywords

DNA; microbial and viral screening; nanopore; noninvasive fetal screening; quantitation; real-time DNA sequencing; RNA; sequencing by synthesis; single-molecule sequencing; tumor genome

In 2001, culminating 13 years of effort at a cost of US$2.7 billion, the initial sequencing of a human genome provided the impetus for a technological revolution aimed at bringing the cost of individual genome sequencing to a price that would allow the complete molecular characterization of an individual's genome [1,2]. The pace of science and technology has quickened – resulting in the 2008 launch of the '1000 Genomes Project', a global project of grand scale [3–5]. While one may debate the various price points that accompany the vision for aligning genomic information aimed at the personalization of healthcare, the ever declining price of complete genome sequencing will reach a tipping point at which benefit outweighs cost and routine application will occur.

Genome knowledge, gained to inform an individual's propensity for disease risk and better diagnosis at the time of disease symptoms, as well as the integration with effective therapeutic

Financial & competing interests disclosure

Patrice M Milos is an employee of Helicos BioSciences Corporation, Cambridge, MA, USA. The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed

No writing assistance was utilized in the production of this manuscript.

treatment for one's disease or perhaps even prevention of disease, if proven to marry vision with results, will drive the cost–benefit ratio more and more favorably. Yet, the personalization of healthcare requires significant research investment and new, innovative technologies to achieve these desired end points. Vision provides the way forward for the technological revolution necessary, for without the vision the path ahead would never be pursued. This intimately relates to the field of molecular diagnostics and perhaps the underlying hope for this technological revolution lies in the field of single molecule sequencing and, thus, a timely perspective topic.

Consider in the next 5–10 years, microarrays, a recent addition to the molecular diagnostics community for both DNA and RNA measurements, will likely be replaced by single-molecule sequencing, quantitative PCR for molecular signature discovery will be replaced by quantitative measurements of nucleic acids using cDNA or direct RNA sequencing, and microbial and pathogen detection will become as simple as obtaining a serum or urine sample followed by direct single-molecule sequencing to detect the infectious agent, strain and genotype. And finally, within the next 5–10 years, we will quickly cross the threshold whereby the cost of sequencing an individual's genome using single-molecule sequencing will be minimal in comparison to the potential benefit to the patient. In all, this is a truly exciting time for the field of molecular diagnostics.

This year, 2009, celebrates the fifth anniversary of the awarding of the National Human Genome Research Institute's (NHGRI) Advanced DNA Sequencing Technology grant program [101]. Following quickly on the heels of the remarkable success of the Human Genome Project and the intimate role NHGRI played in this worldwide initiative, this program was established by NHGRI as a cornerstone for public/private investment necessary to fuel this technology revolution and represented a bold attempt to speed this revolution to patients [6]. While not solely focused on single-molecule sequencing, a good majority of the more than US$100 million invested in this program shepherded this maturing field of single-molecule sequencing. A uniting principle for these grants was the need to minimize the complex nucleic acid sample preparation found with existing sequencing methodologies, the need to obtain ultra-high throughput at low cost, as well as the ability to truly reflect the cellular nucleic acid. These investments have included various methodologies, such as sequencing by synthesis to monitor the growing strand of DNA synthesized during cyclic addition of fluorescently labeled nucleotides, and visual detection of the incorporation events, real-time monitoring of DNA polymerase-directed nucleotide incorporation through imaging of fluorescently labeled nucleotides and the measurement of ionic current passing through nanopores to detect the sequence of nucleic acid molecules passing through the various pore complexes. Yet while the Helicos™ Genetic Analysis System, the first commercial instrument for single-molecule sequencing, has just recently become available, the field is in its infancy and offers unbounded opportunity for application to molecular diagnostics. This article, a vision, will provide insight into what we might expect to emerge from single-molecule sequencing, which is likely to develop quickly during the next 5–10 years.

Emergence of short-read technologies

Initial versions of the massively parallel next-generation sequencing technologies, including the 454 Genome Sequencer 20, the Illumina Genome Analyzer and the more recent Life Technologies/Applied Biosystems SOLiD™ system have reinvigorated the research and diagnostic research communities interest in the power of massive-scale sequencing for whole-genome sequences and begun to be utilized to address important biological questions [7–9]. We are now entering an era in which the vision described in the open paragraphs of this perspective seems within our grasp and single-molecule sequencing will be playing an intimate

role by providing the next technological leap required to pursue whole-genome sequencing and genomic analyses in a cost-effective manner.

The key advantages for single-molecule sequencing as applied to molecular diagnostics include subnanogram sample quantity requirements, the simplicity by which sample preparation is achieved, the lack of PCR amplification providing unbiased sequence information, the shear number of molecules that can be interrogated to allow an accurate and quantitative view of the genomic measurements of interest, and the potential for read lengths that extend well above 5–10 kb and far exceed current methods. Several approaches for single-molecule sequencing are currently in differing states of maturity and offer great potential for application to molecular diagnostics. A brief survey of three different and distinct single-molecule methods that demonstrate unique attributes of single-molecule sequencing follows, with relevant distinguishing features of each technology and it's most appropriate applications.

Approaches to single-molecule sequencing

Sequencing by synthesis, commercialized by Helicos BioSciences Corporation in the form of the HeliScope™ utilizes visual imaging of fluorescently labeled nucleotides and a glass flow cell system. The flow cell contains the individual DNA molecules, captured by oligonucleotides that are complementary to the 3′ end of DNA molecules of interest and are deposited on the surface of a flow cell. Helicos currently uses an oligo dT-50 surface to capture complementary poly-A tail sequences, which have been synthetically added to the 3′ end of genomic DNA using terminal transferase [10–12]. An alternative approach could also involve a flow-cell surface prepared with a sequence-specific oligonucleotide that would be complementary to naturally occurring DNA sequences adjacent to specific genomic regions of interest to allow sequencing through such regions within the genome. Once captured on the flow cell surface, the sequencing by synthesis reaction is initiated through the cyclic addition of fluorescently labeled Virtual Terminator™ nucleotides using alternating cyclic addition of dA, dC, dG and dTs into the channels of the flow cell [12]. DNA polymerase in solution catalyzes the complementary nucleotide incorporation followed by laser excitation of the fluor present on the VT nucleotide and subsequent total internal reflection imaging to capture the presence of the nucleotides added to the billions of growing strands of DNA. Subsequent cleavage of the terminating moiety on the nucleotide then enables the next cyclic nucleotide addition to proceed [12]. With the current ability to capture nearly 3 billion molecules of DNA on the two HeliScope flow cell surfaces, each with 25 channels, sequence yields more than 1B DNA molecules and extends well above 25–28 gigabases of useable sequence per run. A recent demonstration of the sequencing of an individual human genome using this method at a cost below $50,000 [13] provides more impetus for further developments in needed areas such as improvements in read lengths, the implementation of paired reads to accurately map complex genomic regions and continuing accuracy in sequence reads. With this unprecedented depth of genomic information, the unique quantitative information as well as the ability to interrogate archival and forensic nucleic acid samples, sequencing by synthesis provides an unbiased and unparalleled view of copy number variation, transcriptome quantitation and stored tissue samples to name a few applications. Perhaps uniquely suited for the massive scale of sequencing by synthesis, the massive numbers of molecules offer important opportunities for digital and quantitative measurements critical in diagnostics. The recent description of quantitative transcriptome measurements using yeast poly-A+ RNA and Helicos single-molecule digital gene expression application demonstrates an important, cost-effective solution for transcript profiling [14]. This application has now been extended to mammalian tissue RNAs, cell lines and tumor tissue [Raz T & Lipson D, Unpublished Data]. Considering the simplicity of the sample preparation method, combined with the per channel reagents costs, a single HeliScope channel can provide a highly reproducible transcriptome measurement for a cost approaching $350 per sample, making this a very viable approach for clinical application.

Furthermore, with the potential to double and triple yields in the coming year with chemistry and surface-density improvements, the technology provides a continued path to $1000 genome performance.

Real-time single-molecule sequencing utilized by Pacific Biosciences utilizes virtual zero-mode wave guides in which a modified DNA polymerase and template DNA are affixed to the well surface to allow DNA to interact with the polymerase in real time. The system allows the real-time monitoring of the incorporation of the four different phospholinked, fluorescently labeled dNTPs, which are circulating free in solution, into the growing strand of DNA followed by optical detection of the molecular events [15–18]. The sequencing system under development contains several thousand reaction wells in which a thin metal film is deposited onto an optical zero-mode wave guide and optical constraint allows the direct and real-time imaging of fluorescently labeled molecules as they are incorporated into the growing strands of DNA. Sample preparation required for sequencing with this system requires the circularization of the template DNA through the use of bar-bell-shaped nucleic acid adapters that are ligated onto the ends of genomic DNA or cDNA to allow the multipass sequencing required to overcome the current error profiles inherent in their system. These ligations have the potential to create biases that are unique to such a method but will not be fully revealed until more sequencing of complex genomic targets has been completed. While inherent limitations currently exist for this method, the potential for long read lengths of several thousand base pairs far exceeds any of the currently available sequencing platforms available today. Furthermore, the real-time speed with which the sequencing occurs offers a potentially fascinating opportunity for point-of-care nucleic acid sequencing. While the actual instrument and sequencing costs remain to be fully described, the method offers the possibility of sequencing long molecules in real time for a cost-effective approach to obtaining genomic sequence information in complex genomic regions.

A final area of significant longer term investment involves the use of nanopore sequencing, reviewed extensively by Branton *et al*. [19] and being commercialized by Oxford Nanopore Technologies. Early studies demonstrated the basic concept of the passage of single-stranded nucleic acid molecules through nanopores, nano-meter-sized pores, or biological pores created through use of pore-forming membrane proteins as in the case of α-hemolysin or the porin MspA of *Mycobacterium smegmatis* and detection of the DNA movement through the pores [20–24]. Owing to the pore size constraint, individual nucleotide elements of the nucleic acid strands pass through the pores and can be detected using two different approaches: movement of ionic current allowing direct electronic measurements or optical resolution of the molecules as they traverse through the pore. While further from commercial application, the technology offers the potential for very long read lengths with few chemistry requirements and offers the potential to go well beyond any currently available read lengths as well as continuing to dramatically drive down sequencing costs of these long reads.

The current next-generation sequencing technologies that we described earlier, while not single-molecule sequencing, are currently paving the way for the introduction of these new single-molecule sequencing platforms into the diagnostics realm. In particular, the 454 Genome Sequencer 20 and FLX systems have proved to be useful in the analysis of bacterial and viral strains in clinical studies, detecting HIV viral mutations early in the appearance of these emerging isolates [25,26]. While many research areas will benefit from single-molecule sequencing, three immediate areas where single-molecule sequencing offers advantages over non-single-molecule-based methodologies as well as importance for molecular diagnostics are described.

Analysis of nucleic acids from body fluids

Over the last 10 years, there has been a growing realization that the presence of circulating, cell-free nucleic acids in human serum or urine provides new insight into the utility of these nucleic acids for detecting and diagnosing a variety of human conditions [27–30]. Interesting applications emerging have included the examination of low levels of circulating tumor DNA, DNA circulating following transplantation indicative of early graft rejection, RNA molecules found in serum and the growing recognition that fetal DNA present in maternal blood offers the potential for noninvasive fetal screening.

While various studies investigating circulating nucleic acids are emerging [27,28], recent studies involving noninvasive fetal screening have demonstrated the feasibility of using massively parallel sequencing for prenatal screening of chromosomal abnormalities [29–32]. In such cases, maternal and fetal plasma DNA was isolated from women during early pregnancy. The DNA was then sequenced using current next-generation sequencing technology to provide short sequence reads. The sequence reads were mapped to the genome and quantitated at the chromosomal level to allow an assessment of abnormal chromosomal read count ratios indicative of chromosomal aneuploidy, including trisomy 21. Two of these studies correctly identified the presence of trisomy 21 in pregnant women identified to be carrying trisomy 21 fetuses via standard invasive diagnostic methods. A challenge seen with the sequence reads, however, is the clear bias in the nonuniform distribution of sequence reads due to the differing G + C content of the sequence reads when using the PCR-based Illumina technology. This is particularly relevant for chromosomal abnormality detection given the widely varying percentages of G + C content found across human chromosomes, and reads must map quantitatively and accurately across the array of chromosomal genomic content in order to detect a diverse array of chromosomal abnormalities. Currently, this technology has required data obtained from multiple channels of an instrument to provide the coverage and statistical corrections needed to adjust for the genomic content bias seen, which thus results in increased costs for sequencing. In addition, a control genomic sample must also be run to allow appropriate normalization of the chromosomal data.

Single-molecule sequencing offers tremendous promise for this application. Requiring only limited amounts of circulating DNA – mid-picogram amounts – the isolated nucleic acid, which is already optimized for sequencing due to the fragmented nature of the circulating nucleic acid, makes the sample preparation simple and highly amenable to routine and simple sample processing for a molecular diagnostic laboratory. Single-molecule sequencing also provides the potential for more precise measurements due to the lack of G + C bias inherent in this technology, with sequencing by synthesis methods being the preferred choice of technology since the deeper the view of the fetal/maternal DNA possible, the more accurate and precise the measurement; one would want to be able to distinguish differences as low as a 10% deviation from normal chromosomal read counts to allow as early a window for detection as possible. In fact, with some 10–20 million sequence reads that accurately map to the genome, and demonstrate an even distribution across the diverse genomic content of the human genome, one should reliably be able to detect deviations of the normal chromosomal content in the range required during the first trimester of pregnancy. With cost becoming an important attribute of molecule diagnostics, this offers the potential to utilize one channel on an existing commercial single-molecule sequencing platform with minimal upfront sample preparation costs suggesting a per sample cost well below the $1000 pricing, an attractive and noninvasive alternative to existing methods. Furthermore, while the data provides insight into large chromosomal aneuploidy, sequence information can further provide additional information for the mother.

In a similar way to circulating fetal DNA can be monitored against the background of endogenous maternal DNA, circulating tumor DNA may provide insight into early detection of an abnormal tissue state long before overt symptoms result in a visit to your physician. The challenge for patient surveillance will be the requirement for accuracy in ones ability to detect the presence of mutated sequences of relevant tumor gene sequences at a very high fidelity. This application will likely require the ability to detect a mutation found in one out of 100,000 DNA molecules to perhaps even one out of 1 million DNA sequence molecules [33,34]. Here pristine accuracy will be required in order to overcome the underlying error rate of single-molecule-sequencing technologies or alternatively creative strategies that allow one to resequence individual molecules to build a consensus sequence on the same individual DNA molecule will become increasingly important.

Microbial & pathogen detection

Additional extensions in the use of body fluids for molecular diagnostics include the ability to detect the presence of a pathogenic organism and identify the particular pathogenic strain, as well as identifying sequence information relevant to drug sensitivity or resistance in order to better inform therapeutic treatment. Thus, the physician or clinical pathologist has a variety of diverse needs with respect to their ability to make accurate molecular diagnoses. At present, the field of pathogen and microbial diagnostics utilizes both immunoassay-based measurements as well as nucleic acid-based methods, the latter of which are rapidly growing in utility. Immunoassays can examine the presence of an immune response in an individual or they can detect antigens produced directly by the pathogen; however, they often lack the sensitivity required for definitive diagnosis, particularly in early stages of an infection, often making therapy choices difficult [35]. Early utilization of nucleic acid measurements has focused on PCR-based amplification to allow enrichment of the pathogen genome signal in regions of pre-defined interest and initially focused on viral sequences [36,37], with newer emphasis turning to microbial or viral arrays, which allow the interrogation of selected species that can be difficult to culture and are of biomedical importance [38]. Yet, in every case, the sequence of the organism needs to be well defined and assays in place for detection versus a global view of the pathogenic state of the patient in which the infection is fully characterized at the molecular level.

Single-molecule sequencing has the potential to dramatically increase the sensitivity as well as specificity of pathogen detection as well as uncover new emerging resistant strains, as well as new strains or pathogens. In addition, due to the highly quantitative nature of the single-molecule-sequencing nucleic acid measurements, a more global picture of the complement of microbial and/or viral nucleic acid is entirely possible. The ability to obtain human serum samples or urine samples and directly isolate the nucleic acid from these biological fluids either through isolation of viral particles or bacteria, or in many cases direct measurements of the circulating nucleic acid, provides the simple substrate for subsequent single-molecule sequencing. Following nucleic acid isolation, single-molecule-sequencing methods can examine DNA molecules, or nucleic acid converted to first strand cDNA in the case of pathogenic RNA viruses, to detect and characterize bacterial or viral infections. Sequencing by synthesis platforms that examine billions of nucleic acid molecules will allow a deep view into the isolated or circulating nucleic acid, allowing detection of very low quantities of circulating viruses or bacterial organisms. After nucleic acid isolation, low picogram quantities of nucleic acid are poly-A tailed and hybridized to an oligo dT flow cell surface, capturing millions to billions of DNA or cDNA strands depending on flow cell usage. At present, read lengths of 35–55 nt allow one to sequence the entire genome of most bacterial genomes and newer methods for paired reads allow you to maximize the placement of sequence tags in repetitive genomic regions with special emphasis on ribosomal genes, which are often highly duplicated in tandem repeats. Costs again become important and will require continued

optimization to bring the cost down to a desired price range for diagnostic applications as one might imagine requiring a very low cost requirement for routine use approaching well below the $100 price range. Real-time single-molecule sequencing also offers an interesting opportunity for the rapid detection of bacterial and viral sequences, with the major hurdle to optimize sample preparation and achieve error rates that allow an accurate complete genome sequence. In addition, with extended read lengths obtained via real-time single-molecule sequencing or with emerging nanopore technologies, novel or highly rearranged pathogenic strains can easily be assembled.

An integral part of microbial and viral nucleic acid diagnostics using single-molecule sequencing will be the ability to accurately and rapidly perform *de novo* genome assembly or assembly of critically important regions from the sequence information obtained to allow accurate diagnosis. New and improving assembly tools being developed for *de novo* genome assembly of short reads, including ALLPATHS [39] the Sanger Center's Velvet [40] 454's Newbler assembler [41] and the Celera Assembler reviewed by Chaisson and Pevzner [42], are making rapid progress for enhancing small genome assembly and thus will greatly facilitate improved use in molecular diagnostics.

Examining the tumor genome

The field of oncology is positioned to be truly transformed by the promise of single-molecule sequencing assuming continually improving accuracy rates given some of the unique challenges poised by somatic mutations. The hallmark of tumorigenesis involves the underlying somatic changes that occur in the cell, which leads to perturbation of the cellular responses and often unchecked cell proliferation. The occurrence of these somatic mutations and their importance in tumorigenesis has been extensively studies over the last several years [43–46]. These hallmark studies have provided the foundation of significant investment in public initiatives like the Cancer Genome Atlas [47,102]. The potential to examine the complete tumor genome at the time of diagnosis in a cost-effective manner utilizing single-molecule sequencing offers the promise to marry mutational status directly with clinical care. Key here will be the continued improvements in error rates to allow low level detection of somatic events in the background of normal DNA, which may be reflected in only one out of every 100 DNA, or perhaps even one in 1000 DNA molecules sequenced. But one must also consider the cost–effectiveness of sequencing whole tumor genomes. While one might consider early studies aiming to demonstrate the value of tumor genome knowledge, the long-term ability to integrate this information into the true diagnostic setting for patient care requires a significant reduction in costs that may only be achievable by the promises of single-molecule sequencing. Evidence of success using these current and emerging methods will be watched closely as the race to the $1000 genome continues.

Novel methods for sequencing individual strands of DNA multiple times also dramatically reduce the single-read error rates and, therefore, also offer benefit to such molecular screening [11]. In addition, while PCR amplification or targeted DNA capture are currently utilized to select those regions of the genome of highest interest for mutational investigation, the methods all require cumbersome sample preparation including ligation and additional rounds of amplification [48–52]. The ability to simplify the selection methods to avoid all such sample manipulation, with the exception of shearing the genomic DNA of interest and selection of the regions of interest followed by direct sequencing of the picogram quantities of selected nucleic acid, makes this another feature that greatly benefits from single-molecule sequencing.

The ability to identify somatic events early in the pathogenesis of the disease, perhaps even prior to aberrant cellular growth, through monitoring of nucleic acid circulating in blood once again may provide unique insight into the molecular events occurring at early stages in cancer.

The ability to utilize the shear simplicity and scale once again comes into play as a regular monitoring of circulating DNA or RNA for mutational screening to detect the presence of mutational events occurring throughout the human body. The ability to detect these rare events, however, will require the ability to detect mutation events that may be well below the 1:1,000,000 events in your total DNA or RNA population [33,34]. At present this high level of accuracy will require continued improvements in the mechanisms that lead to errors, which, in single-molecule sequencing, are predominated by dark nucleotides resulting in apparent deletion events that are overcome routinely by sequence coverage in the current methods for sequencing using single-molecule sequencing.

While much of this perspective has focused on the examination of DNA, RNA transcriptome studies are similarly poised to benefit from the application of single-molecule sequencing. This includes full transcriptome sequencing in which first-strand cDNA is synthesized using random hexamer priming or digital gene expression in which a poly-U primer is used for hybridization to the poly-A tail of messenger RNA and first strand cDNA extension follows – both of which are followed by a simple dA tailing reaction followed by direct sequencing [53]. While these methods can commonly be used for examination of RNA isolated from tumor tissue or normal adjacent tissue, new and emerging interest again points to the importance of circulating small RNA for measurements in the field of oncology [54–56]. With an unbiased and highly quantitative measurement, RNA studies, whether cellular or circulating, will be dramatically improved by single-molecule sequencing.

With vision comes a note of caution

In any perspective article, one must also strive to provide insight into the pitfalls that await this emerging field of single-molecule sequencing. As indeed the potential is huge, one does need to be cautious of the continued introduction of new technologies and the pace of introduction for clinical applications. Consider when comparative genomic hybridization (CGH) arrays were first utilized for the assessment of cytogenetic changes occurring in tumors. The first publication on the use of CGH arrays to identify genome-wide copy number changes appeared in 1999 [57] as purely a research result with important research findings on cytogenetic changes. Yet, it took 5 years until the use of array CGH for clinical studies led to a shift in the marketplace with array vendors moving from the academic researcher to diagnostics laboratories. The potential now for sequencing to replace array CGH measurements highlights, once again, the fact that technological paradigm shifts continue unabated. Key are cost, accuracy and depth of information, as well as ease of use and interpretation. While we have now seen the publication of a single-molecule sequencing of a human genome at a cost that is below $50,000, routine usage will require costs well below this range for practical application in the molecular diagnostic setting.

Until now, single-molecule sequencing has presented a significant, new challenge for the molecular diagnostic community. With the simplicity of the methods, many of the upfront challenges that laboratories have faced are definitively streamlined. However, the downstream hurdles might become quite enormous. The wealth of data generated by single-molecule sequencing methods and the computational and analytical tools required to analyze and make sense of the sequences becomes a significant hurdle if not addressed immediately. User interfaces that allow the marriage of sequence information with a molecular and clinical interpretation require significant development on the part of both the technology developers and the user community. The sooner attention is paid to this facet of the field the quicker the uptake will be by the pathologist, medical center researcher and the molecular diagnostic community. I have absolute belief that our greatest hope for the individual healthcare benefits lies in our ability to effectively integrate the knowledge gained from genomic information with

an individual's medical care and thus deliver a true state of personalized healthcare throughout one's lifetime.

Expert commentary

Single-molecule sequencing of nucleic acids holds the potential to address many of the fundamental issues that have challenged the field of genomic-based biomarkers, including the ability to effectively standardize measurements, and to allow quantitative and qualitative comparisons within and across diverse patient cohorts. By limiting the inherent bias present in PCR-based measurements and with other nonsingle-molecule sequencing platforms by eliminating the complicated sample preparation required by many technologies, single-molecule sequencing allows the direct interrogation of the nucleic acid contained within tissues and cells. While originally focused on DNA and cDNA sequencing measurements, new research that allows the direct interrogation of RNA paves the way to a future of transcriptomic measurements of unparalleled depth and free from the cumbersome challenges many cDNA-based measurements. Single-molecule sequencing thus offers the potential promise to usher in a new era in molecular diagnostics.

Five-year view

Over the next 5 years, single-molecule sequencing will emerge as a major component to molecular diagnostics. The simplicity and bias-free measurement provides new impetus for transitioning from indirect to direct measurements that truly reflect the biology of the patient. Such technological advancement will usher in an era where the knowledge gained from molecular information outweighs the cost of sequencing an individual's genome and with that comes the opportunity to truly revolutionize the personalization of healthcare.

Key issues

- Single-molecule sequencing will usher in the next technological revolution of the era of the genome.
- Single-molecule sequencing will provide a more definitive and accurate measurement necessary for molecular diagnostics.
- Prenatal, pathogenic and oncology diagnostics will be enabled in new ways through the use of single-molecule sequencing.
- Technical improvements to drive down error rates and increase read lengths remain; however, with continued investment from both public and private sources of investment, these technological solutions will emerge.

Acknowledgments

Special thanks to Patrick Terry for providing insight into the many principles and challenges of genomics and molecular diagnostics and all my colleagues at Helicos for their expertise and passion to achieve the vision described herein.

References

1. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921. [PubMed: 11237011]
2. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291:1304–1351. [PubMed: 11181995]
3. Kaiser J. A plan to capture human diversity in 1000 genomes. Science 2008;319:395. [PubMed: 18218868]

4. Kuehn BM. 1000 genomes project promises closer look at variation in human genome. JAMA 2008;300 (23):2715. [PubMed: 19088343]
5. Wise J. Consortium hopes to sequence genome of 1000 volunteers. BMJ 2008;336(7638):237. [PubMed: 18244979]
6. Collins FS, Green ED, Guttmacher AE, et al. A vision for the future of genomics research. Nature 2003;422:835–847. [PubMed: 12695777]
7. 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]
8. 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]
9. Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. Nature 2008;456 (7218):60–65. [PubMed: 18987735]
10. Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. Proc. Natl Acad. Sci. USA 2003;100(7):3960–3964. [PubMed: 12651960]
11. Harris TD, Buzby PR, Babcock H, et al. Single-molecule DNA sequencing of a viral genome. Science 2008;320(5872):106–109. [PubMed: 18388294]
12. Bowers J, Mitchell M, Beer E, et al. Virtual terminator™ nucleotides for next generation DNA sequencing. Nat. Methods 2009;6(8):593–595. [PubMed: 19620973]
13. Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. Nat. Biotechnol 2009;27(9):847–852. [PubMed: 19668243]
14. Lipson DL, Raz T, Kieu A, et al. Quantification of the yeast transcriptome by single-molecule sequencing. Nat. Biotechnol 2009;27:652–658. [PubMed: 19581875]
15. Levene MJ, Korlach J, Turner SW, et al. Zero-mode waveguides for single-molecule analysis at high concentrations. Science 2003;299(5607):682–686. [PubMed: 12560545]
16. Korlach J, Marks PJ, Cicero RL, et al. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc. Natl Acad. Sci. USA 2008;105(4):1176–1181. [PubMed: 18216253]
17. Korlach J, Bibillo A, Wegener J, et al. Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. Nucleosides Nucleotides Nucleic Acids 2008;27(9): 1072–1083. [PubMed: 18711669]
18. Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. Science 2009;323(5910):133–138. [PubMed: 19023044]
19. Branton D, Deamer DW, Marziali A, et al. The potential and challenges of nanopore sequencing. Nat. Biotechnol 2008;26(10):1146–1153. [PubMed: 18846088]
20. Kasianowicz JJ, Brandin E, Branton D, et al. Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl Acad. Sci. USA 1996;93:13770–13773. [PubMed: 8943010]
21. Braha O, Walker B, Cheley S, et al. Designed protein pores as components for biosensors. Chem. Biol 1997;4:497–505. [PubMed: 9263637]
22. Clarke J, Wu HC, Jayasinghe L, et al. Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol 2009;(4):265–270. [PubMed: 19350039]
23. Stoddart D, Heron AJ, Mikhailova E, et al. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proc. Natl Acad. Sci. USA 2009;106:7702–7707. [PubMed: 19380741]
24. Butler TZ, Pavlenok M, Derrington I, et al. Single-molecule DNA detection with an engineered MspA protein nanopore. Proc. Natl Acad. Sci. USA 2008;105(52):20647–20652. [PubMed: 19098105]
25. Simen BB, Simons JF, Hullsiek KH, et al. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. J. Infect. Dis 2009;99(5):693–701. [PubMed: 19210162]
26. Archer J, Braverman MS, Taillon BE, et al. Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4 (CXCR4)-using HIV-1 with ultra-deep pyrosequencing. AIDS 2009;23(10):1209–1218. [PubMed: 19424056]
27. Scalzo PL, Ikuta N, Cardoso F, et al. Quantitative plasma DNA analysis in Parkinson's disease. Neurosci. Lett 2009;452(1):5–7. [PubMed: 19444939]

28. Nakamura T, Sunami E, Nguyen T, et al. Analysis of loss of heterozygosity in circulating DNA. Methods Mol. Biol 2009;520:221–229. [PubMed: 19381958]
29. Chiu RW, Cantor CR, Dennis Lo YM. Non-invasive prenatal diagnosis by single molecule counting technologies. Trends Genet 2009;25(7):324–331. [PubMed: 19540612]
30. Heung MM, Jin S, Tsui NB, et al. Placenta-derived fetal specific mRNA is more readily detectable in maternal plasma than in whole blood. PLoS ONE 10 2009;4(6):E5858.
31. Chiu RW, Chan KCA, Gao Y, et al. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc. Natl Acad. Sci. USA 2008;105:20458–20463. [PubMed: 19073917]
32. Fan HC, Blumenfeld YJ, Chitkara U, et al. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc. Natl Acad. Sci. USA 2008;105:16266–16271. [PubMed: 18838674]
33. Schwarzenbach H, Alix-Panabières C, Müller I, et al. Cell-free tumor DNA in blood plasma as a marker for circulating tumor cells in prostate cancer. Clin. Cancer Res 2009;15(3):1032–1038. [PubMed: 19188176]
34. Ahlquist DA, Sargent DJ, Loprinzi CL, et al. Stool DNA and occult blood testing for screen detection of colorectal neoplasia. Ann. Intern. Med 2008;149(7):441–450. [PubMed: 18838724]
35. Jortani SA, Pugia MJ, Elin RJ, et al. Sensitive noninvasive marker for the diagnosis of probable bacterial or viral infection. J. Clin. Lab. Anal 2004;18(6):289–295. [PubMed: 15543565]
36. Varani S, Stanzani M, Paolucci M, et al. Diagnosis of bloodstream infections in immunocompromised patients by real-time PCR. J. Infect 2009;58(5):346–351. [PubMed: 19362374]
37. Palka-Santini M, Cleven BE, Eichinger L, et al. Large scale multiplex PCR improves pathogen detection by DNA microarrays. BMC Microbiol 2009;9:1. [PubMed: 19121223]
38. Lin B, Malanoski AP. Resequencing arrays for diagnostics of respiratory pathogens. Methods Mol. Biol 2009;529:231–257. [PubMed: 19381976]
39. Butler J, MacCallum I, Kleber M, et al. ALLPATHS, *de novo* assembly if whole-genome shotgun reads. Genome Res 2008;18(5):810–820. [PubMed: 18340039]
40. Zerbino DR, Birney E. Velvet, algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res 2008;18(5):821–829. [PubMed: 18349386]
41. Miller JR, Delcher AL, Koren S, et al. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 2008;24(24):2818–2824. [PubMed: 18952627]
42. Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. Genome Res 2008;18 (2):324–330. [PubMed: 18083777]
43. Leary RJ, Lin JC, Cummins J, et al. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. Proc. Natl Acad. Sci. USA 2008;105(42): 16224–16229. [PubMed: 18852474]
44. Jones S, Zhang X, Parsons DW, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 2008;321(5897):1801–1806. [PubMed: 18772397]
45. Parsons DW, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. Science 2008;321(5897):1807–1812. [PubMed: 18772396]
46. Greenman C, Stephens P, Smith R, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446(7132):153–158. [PubMed: 17344846]
47. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455(7216):1061–1068. [PubMed: 18772890]
48. Hodges E, Xuan Z, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. Nat. Genet 2007;39(12):1522–1527. [PubMed: 17982454]
49. Krishnakumar S, Zheng J, Wilhelmy J, et al. A comprehensive assay for targeted multiplex amplification of human DNA sequences. Proc. Natl Acad. Sci. USA 2008;105(27):9296–9301. [PubMed: 18599465]
50. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol 2009;27(2):182–189. [PubMed: 19182786]

51. Hodges E, Rooks M, Xuan Z, et al. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat. Protoc 2009;4(6):960–974. [PubMed: 19478811]

52. Turner EH, Lee C, Ng SB, et al. Massively parallel exon capture and library-free resequencing across 16 genomes. Nat. Methods 2009;6(5):315–316. [PubMed: 19349981]

53. Lipson D, Raz T, Kieu A, et al. Quantitation of the yeast transcriptome by single molecule sequencing. Nat. Biotechnol 2009;27(7):652–658. [PubMed: 19581875]

54. Mitchell PS, Parkin RK, Kroh EM, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proc. Natl Acad. Sci. USA 2008;105(30):10513–10518. [PubMed: 18663219]

55. Tewari M, Krishnamurthy A, Shukla HS. Predictive markers of response to neoadjuvant chemotherapy in breast cancer. Surg. Oncol 2008;17(4):301–311. [PubMed: 18467090]

56. Wyman SK, Parkin RK, Mitchell PS, et al. Repertoire of microRNAs in epithelial ovarian cancer as determined by next generation sequencing of small RNA cDNA libraries. PLoS ONE 2009;4 (4):E5311. [PubMed: 19390579]

57. Pollack JR, Perou CM, Alizadeh AA, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat. Genet 1999;23(1):41–46. [PubMed: 10471496]

Websites

101. National Human Genome Research Institute's Advanced DNA Sequencing Technology grant program. www.genome.gov/27527585

102. The Cancer Genome Atlas TCGA. http://cancergenome.nih.gov