Leveraging Whole Human Genome Sequencing in Cancer

Illumina sequencing is enabling whole-genome sequencing studies to become a mainstay of disease research.

Introduction

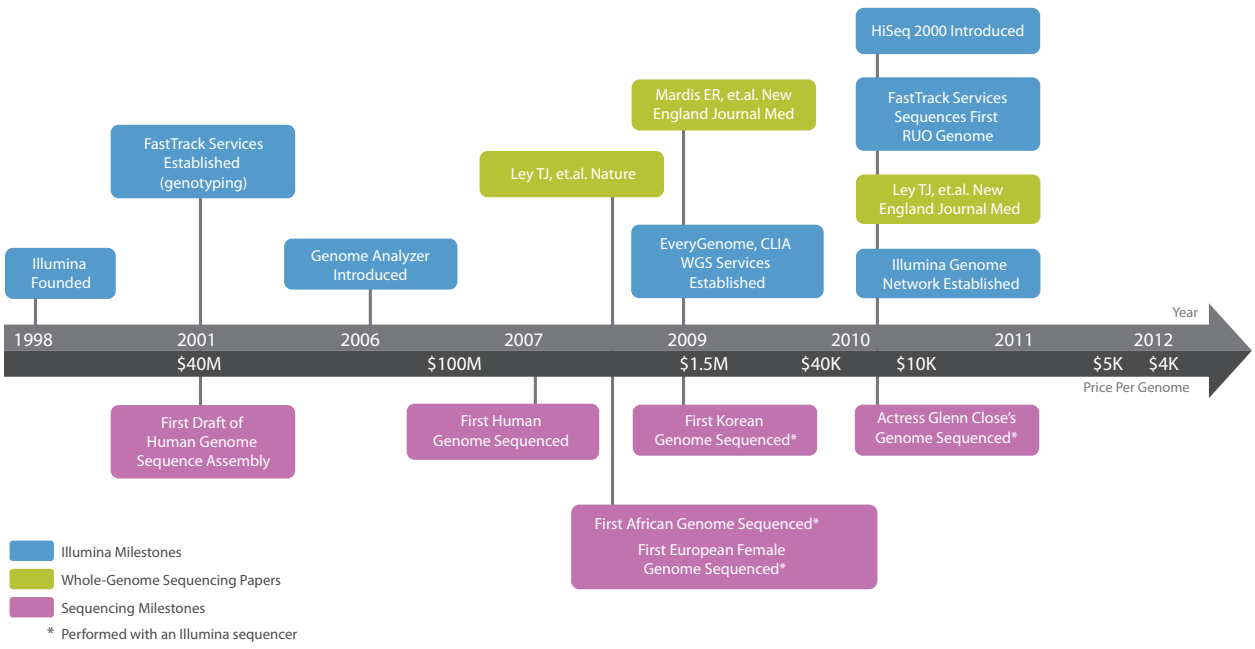
The first human genome was fully sequenced in 2003, enabling researchers to analyze the differences and similarities between individual genomes during normal and disease states. Yet, the initial cost of performing sequencing studies was high, putting a strain on research budgets and limiting its routine use. Advances made by Illumina in the development of massively parallel sequencing systems, such as the Genome Analyzer™ and the HiSeq® systems, have reduced the cost of whole human genome sequencing studies, while providing the coverage, sequencing depth, and high-quality results necessary to identify gene mutations associated with disease. Based on industry-leading TruSeq™ technology, the Genome Analyzer and HiSeq systems are enabling whole human genome sequencing to become an integral part of disease research, whether it is performed by researchers in their own laboratories, or by Illumina and its partners through the Illumina Genome Network.

Whole Human Genome Sequencing Studies Evolve

In the years since it was introduced, the Genome Analyzer and its reagents have been continually enhanced to deliver higher throughput and sequencing depth, improving the system's ability to identify gene mutations in whole-genome sequencing (WGS) studies. Over the last four years, a Washington University research team has published three scientific papers that demonstrate the evolution of WGS on the Genome Analyzer and the system's delivery of increasingly higher clarity TruSeq results in identifying somatic mutations in acute myeloid leukemia (AML) (Figure 1).

M1 AML is the most common of eight subtypes of the disease, representing approximately 20% of all AML cases. Previous technologies identified very few recurrent somatic copy number variants, suggesting that there may be yet undiscovered mutations that contribute to the disease. The cost-effectiveness, accuracy, and low-sample input requirement of the Genome Analyzer led the

Figure 1: Advances in Whole Human Genome Sequencing



In just 15 years, whole human genome sequencing has moved from a project that took 20 groups collaborating more than a year to generate one billion bases covering 94% of the genome and costing $40 million, to a study that can be performed on a HiSeq 2000 system for less than $10,000. After identifying an initial group of somatic mutations in AML using the original Genome Analyzer (2008), researchers at Washington University have leveraged the increasingly higher clarity of the Genome Analyzer$_{IIx}$ (2009, 2010) to dig deeper into the genome to obtain a broader picture of the role somatic mutations have in the disease.

research team to choose it to perform WGS on M1 AML patient samples, with the goal of identifying any relevant mutations that contribute to the molecular pathogenesis of this common subtype.

2008—Pioneering the Use of Whole Human Genome Sequencing to Identify Somatic Mutations in Cancer

Using the first-generation Genome Analyzer, the research team performed WGS on tumor and normal (skin) samples from an AML-afflicted individual with normal cytogenetics[1]. At the time, this type of analysis was relatively novel and this study established the validity of using next-generation sequencing technology to investigate somatic mutations in cancer.

In the study, four tumor and three normal libraries were loaded on to 10 Genome Analyzer flow cells, with 98 full runs needed to achieve 90% coverage on the tumor and 34 full runs necessary to achieve 80% coverage on the normal samples. Of the > 2.6 million single-nucleotide variants (SNVs) found in the tumor and > 2.5 million detected in the normal samples, eight SNVs were found to be highly significant heterozygous, non-synonymous somatic variations found only in the tumor samples, and therefore most probable for involvement in disease.

The authors stated, "…massively parallel short-read sequencing provides an effective method for examining single nucleotide and short indel variants by comparison of the aligned reads to a reference genome sequence. By sequencing our patient's tumour genome to a depth of 30-fold coverage, and gauging our ability to detect known heterozygous positions across the genome, we have produced a sufficient depth and breadth of sequence coverage to comprehensively discover somatic genome variants. A slightly lower coverage of the normal genome from this individual helped to identify nearly 98% of potential variants as being inherited, a critical filter that allowed us to more readily identify the true somatic mutations in this tumour."

According to the authors, the results strongly supported "the notion that hypothesis driven (for example, candidate gene-based) examination of tumour genomes by PCR-directed or capture-based methods is inherently limited, and will miss key mutations. A further and important consideration is the demand for large amounts of genomic DNA by these techniques; this is a serious limitation when precious clinical samples are being studied. The Illumina technology requires only 1 µg of DNA per library, enabling the study of primary tumour DNA rather than requiring the use of tumour cell lines, which may contain genetic changes and adaptations required for immortalization and maintenance in tissue culture conditions."

The authors believe that the eight SNVs identified would not have been discovered without the use of whole-genome sequencing, supporting the use of the technology for uncovering mutations that are relevant in the pathogenesis of other types of cancer.

2009—Higher Coverage Identifies More Somatic Mutations

A year later, the research team initiated a whole-genome sequencing study of another primary, cytogenetically normal, *de novo* M1 AML genome and a matched normal skin genome from an AML-afflicted patient[2]. The enhanced Genome Analyzer$_{IIx}$ enabled the research team to obtain a higher level of coverage (approximately 98% vs. 91% coverage) with fewer sequencing runs (16.5 vs. 98), resulting in a dramatically reduced cost of data generation.

"...the eight SNVs identified would not have been discovered without the use of whole-genome sequencing, supporting the use of the technology for uncovering mutations that are relevant in the pathogenesis of other types of cancer."

Figure 2: Washington University WGS Cancer Research Parameters and Current Illumina Genome Network Service Capabilities

	Washington University WGS Research Papers			Illumina Genome Network Service Capabilities
Year	**2008[1]**	**2009[2]**	**2010[3]**	**2010–2012**
System	**Genome Analyzer (Solexa)**	**Genome Analyzer$_{IIx}$**	**Genome Analyzer$_{IIx}$**	**Genome Analyzer$_{IIx}$ / HiSeq 2000**
Project	WGS	WGS and targeted resequencing on 16 samples	WGS and targeted resequencing of 1 sample	All project sizes (minimum of 5µg DNA input)
Diploid Coverage	Tumor: 46×–65× Normal: 27×–42×			30×–200×
PE or Single Reads	Tumor: 32 bp single read Normal: 35 bp single end	36 or 51 bp paired end	Tumor: 32 bp single end Normal: 35 bp single end	2×100 bp paired end
Insert Size	150–200 bp	200–250 bp	150–200 bp	300–350 bp
Cost	n/a	n/a	n/a	$4,000/sample (50+ samples)

The more comprehensive coverage identified twelve somatic mutations within gene coding sequences, along with 52 somatic point mutations in conserved or regulatory portions of the genome. Of the 64 mutations, four were found to occur in other AML patient samples. While mutations in MRAS and NPM1 had been previously identified in AML patients, this WGS study uncovered never before identified mutations in the IDH1 gene and a nongenic mutation in a highly conserved regulatory genomic region.

2010—Identification of Recurring Somatic Mutations Yields a Broader Picture

Two years after their original WGS studies with AML samples, the Washington University research team decided to reevaluate the first AML case. System and reagent advancements enabled them to achieve 99.6% coverage of tumor samples and identify several previously undetected non-synonymous tier-1 mutations[3]. Of the mutations identified, one base pair deletion in DNMT3A (not identified in the two previous studies) encoding a DNA methyltransferase was verified using targeted resequencing in a larger sample set.

"The discovery of highly recurrent mutations in DNMT3A may provide a tool for the classification of intermediate-risk AML," stated the authors. "If these data are reproduced in other series, clinical trials designed to assess the effects of early intensification of treatment in patients with DNMT3A mutations may be warranted. The careful analysis of many additional AML samples with unbiased genomic methods may reveal changes in methylation or expression that help to define the mechanisms of action of DNMT3A."

Summary

In the span of just five years, whole human genome sequencing with Illumina sequencers has become a mainstay of disease research, cost-effectively delivering highly accurate results that are uncovering gene:disease associations never before recognized. While the Washington University team performed its own WGS studies with a Genome Analyzer, other researchers can take advantage of the WGS services provided by the Illumina Genome Network to obtain the same high-quality TruSeq genome, with data provided in industry-standard format to streamline collaborations and enable integrated follow-on studies.

Figure 2 illustrates the advances the Washington University research team was able to make using the Genome Analyzer, and also shows how similar studies could have been performed through the Illumina Genome Network.

References

1. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 6: 66–72.
2. Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, et al. (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med 10: 1058–1066.
3. Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, et. al. (2010) DNMT3A mutations in acute myeloid leukemia. N Engl J Med 16: 2424–2433.

Illumina, Inc. • 5200 Research Place, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2011 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, BaseSpace, BeadArray, BeadXpress, cBot, CSPro, DASL, DesignStudio, Eco, Genetic Energy, GAIIx, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, Sentrix, SeqMonitor, Solexa, TruSeq, VeraCode, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.
Pub. No. 770-2011-031 Current as of 07 December 2011

illumina®