Roche

Diagnostics

Genome Sequencer 20 System

First to the Finish

Technology 4-5

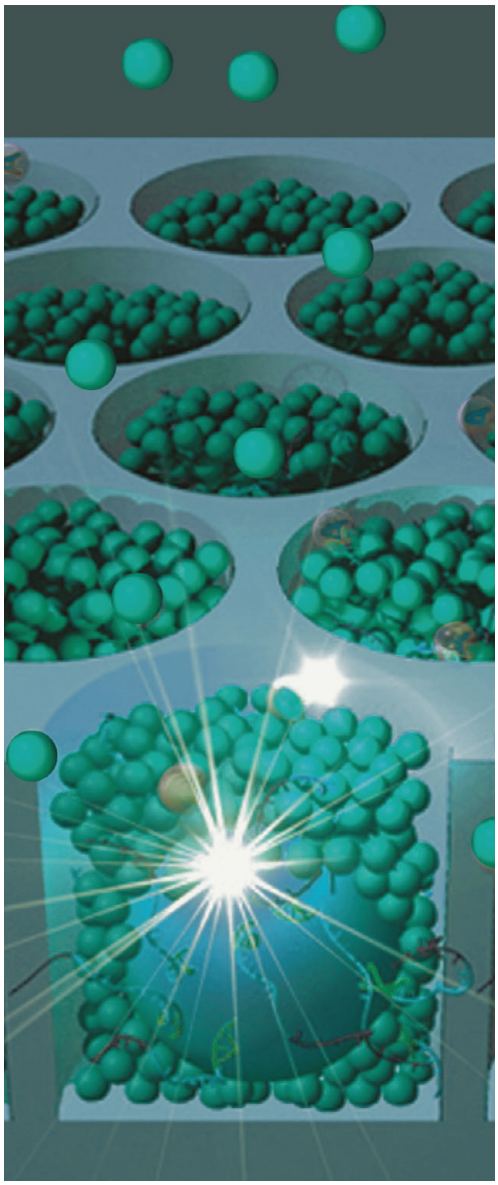
Process Steps 8-11
- DNA Library Preparation 8
- emPCR Amplification 9
- Sequencing-by-Synthesis 10-11

The Genome Sequencer 20 System 6-7

Software 12-13

Applications 14-36
- Whole Genome Sequencing 14-19
  - – Resequencing 14-15
  - – *De Novo* Sequencing 16-17
  - – Paired End Assembly 18-19
- Transcriptome and Gene Regulation Studies 20-27
- Amplicon Analysis 28-35

Ordering Information 37-39



Illumina Exhibit 1028, Page 2 of 40

The Genome Sequencer 20 System

The newest revolution in sequencing today

The Genome Sequencer 20 System uses a revolutionary technology.

- Accurately decipher more than 20 million bases per 5.5-hour instrument run.
- Eliminate cloning and colony picking.
- Generate complete libraries with no cloning bias.

Perform innovative applications that are not possible with other techniques.

Whole Genome Sequencing (shotgun)

- *De novo* sequence or resequence microbial genomes and BACs – in days, not weeks or months.
- Prepare a Paired End library to order and orient the contigs from your *de novo* sequencing project.

Transcriptome/Gene Regulation Studies

- Perform gene identification and quantification studies based on high-throughput sequencing of cDNA fragments (short tags, ESTs, miRNA).
- Identify transcription factor binding sites (ChIP libraries).
- Study DNA methylation patterns.

Amplicon Analysis

- Discover somatic mutations in complex samples for cancer research.
- Accelerate SNP discovery.

Illumina Exhibit 1028, Page 3 of 40

Technology – Workflow

From DNA to bioinformatics

Technology

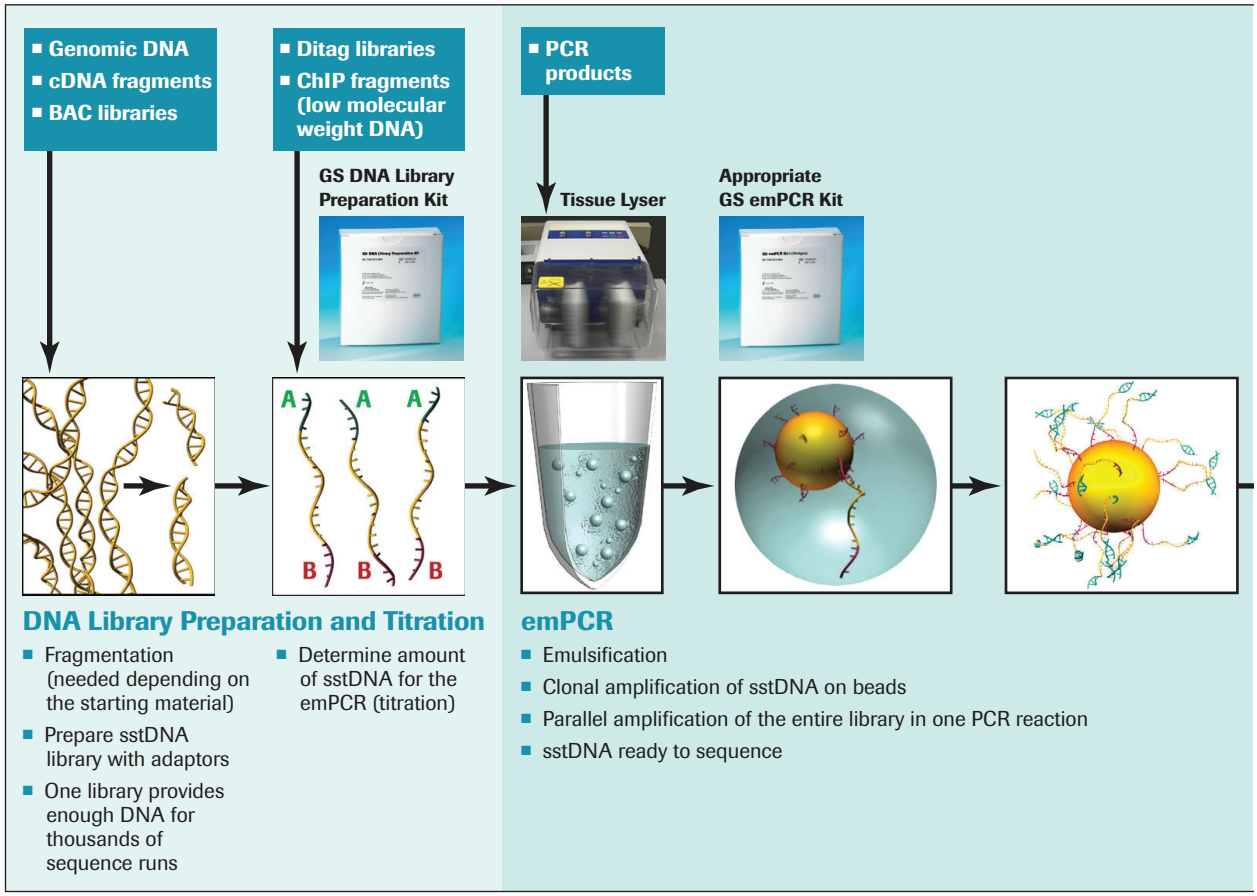
From library preparation to bioinformatics – experience this high-speed, complete solution for efficient high-throughput sequencing.

The Genome Sequencer 20 System, developed using the novel 454 technology,[1] eliminates the need for large-scale robotics for traditional sample preparation. Not only is clonal bias removed, but the need for colony picking and microplate handling is reduced to a simple preparation step.

Through parallelization, state-of-the art image processing, and unique data analysis, tens of megabases of sequence data are now produced in hours from a single instrument run.

Finish your research project in record time with fast, accurate, and cost-effective high-throughput sequencing – a dramatic difference compared to the traditional Sanger technology (Figure 1).

Benefit from the versatility of this technology to sequence diverse sample types, such as genomic DNA, cDNA, BAC libraries, or PCR products (Figure 2), supporting multiple applications from whole genome sequencing to amplicon and transcriptome analysis.



Illumina Exhibit 1028, Page 4 of 40

Comparison of high-throughput Sanger technology to the 454 technology used by the Genome Sequencer 20 System, in whole genome sequencing

Sanger technology employs	454 technology employs
7 days*	**2.5 days**
■ shotgun fragmentation of the genome	■ shotgun fragmentation of the genome ■ adaptor ligation on DNA fragments ■ titration/quantification
■ cloning of the fragments into bacteria ■ colony picking, microplate handling ■ DNA purification from the clones	■ clonal amplification of DNA fragments on beads (emPCR) ■ DNA-bead enrichment
Weeks**	**1 day†**
■ sequencing by dideoxy chain termination ■ electrophoresis	■ sequencing-by-synthesis on a PicoTiterPlate device ■ image and signal processing
■ whole genome mapping or assembly	■ whole genome mapping or assembly

* Assumes high-throughput robotics and several technicians are in place.
** For example, approximately 150 runs (1 run/2 hours) for a 2-million-base genome at 6x coverage.
† For example, 1 run (1 run/5.5 hours) for a 2-million-base genome at 10x coverage.

Figure 1: Comparison of Sanger technology with 454 technology for whole genome sequencing a two-million-base bacterium.

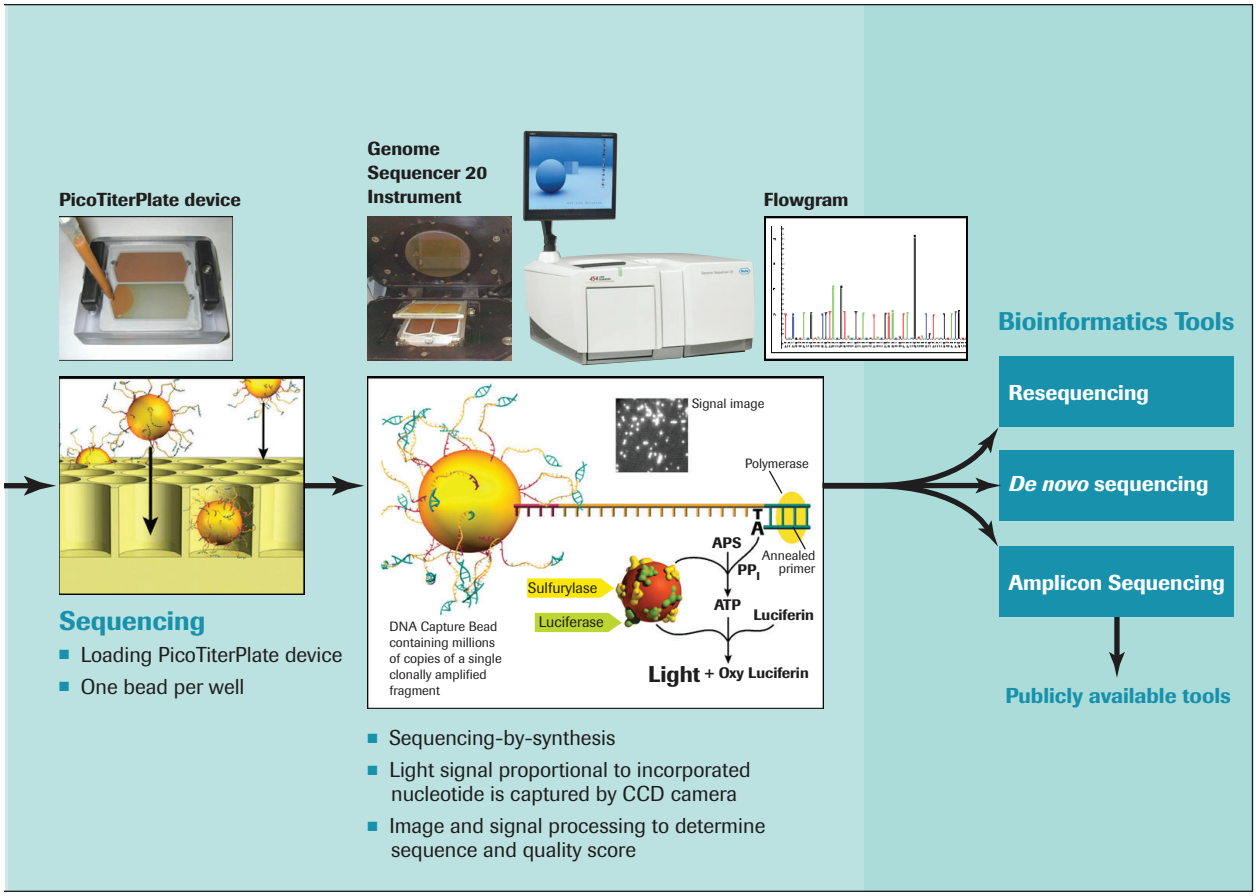


Figure 2: Genome Sequencer 20 System Workflow Overview

Technology

Illumina Exhibit 1028, Page 5 of 40

The Sequencing Technology

Enhance your sequencing process — from genome to sequence in record time

Process Steps

Generate tens of millions of bases per run with the straightforward workflow of the Genome Sequencer 20 System (Figures 3-6).

DNA Library Preparation

Sample preparation is dependent on the type of starting material used. The preparation process comprises a series of enzymatic steps to produce single-stranded template DNA (sstDNA) incorporating primer and binding adaptors. For example, genomic DNA (gDNA) is fractionated into smaller fragments (300-800 base pairs) that are subsequently polished (blunted). Short Adaptors (A and B) are then ligated onto the ends of the fragments. These adaptors provide priming sequences for both amplification (emPCR) and sequencing of the sample-library fragments, and contain a streptavidin binding site for sample purification. Low molecular weight DNA is used without fragmentation and sample preparation begins with adaptor ligation. The A and B adaptors can also be added during PCR by using the appropriate primers (provided in GS emPCR Kit II (Amplicon A, Paired End) and GS emPCR Kit III (Amplicon B). The sstDNA library produced at the end of this preparation step is assessed for its quality, and the optimal amount (DNA copies per bead) needed for emPCR is determined by a titration run.

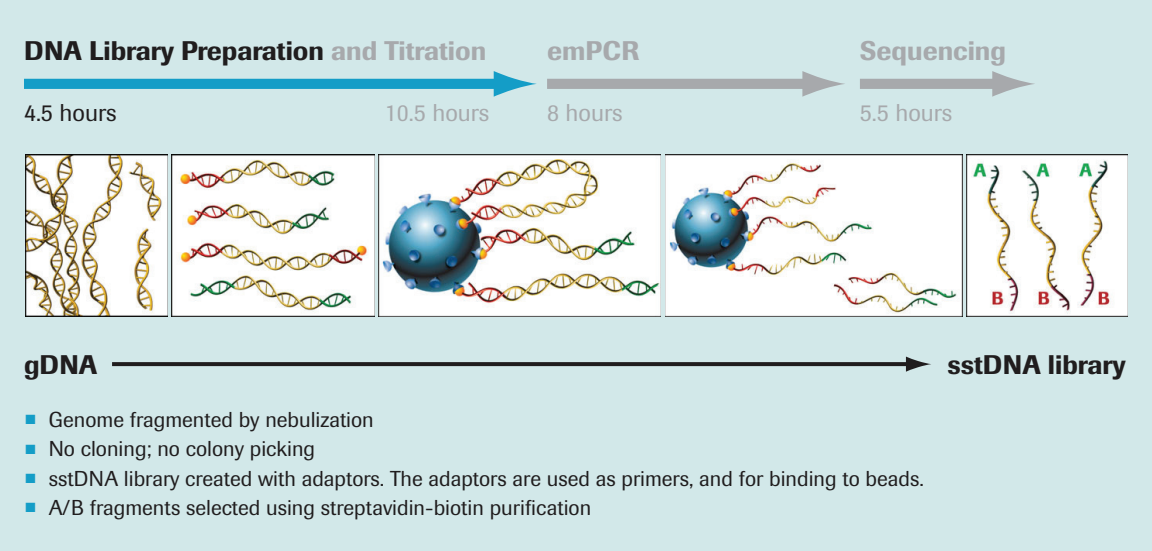


Figure 3: DNA library preparation with the Genome Sequencer 20 System.

Illumina Exhibit 1028, Page 6 of 40

emPCR Amplification

The sstDNA library is immobilized onto specially designed DNA Capture Beads. Each bead carries a single sstDNA library fragment. The bead-bound library is emulsified with amplification reagents in a water-in-oil mixture. Each bead is separately captured within its own microreactor for PCR amplification. Amplification is performed in bulk, resulting in bead-immobilized, clonally amplified DNA fragments that are specific to each bead.

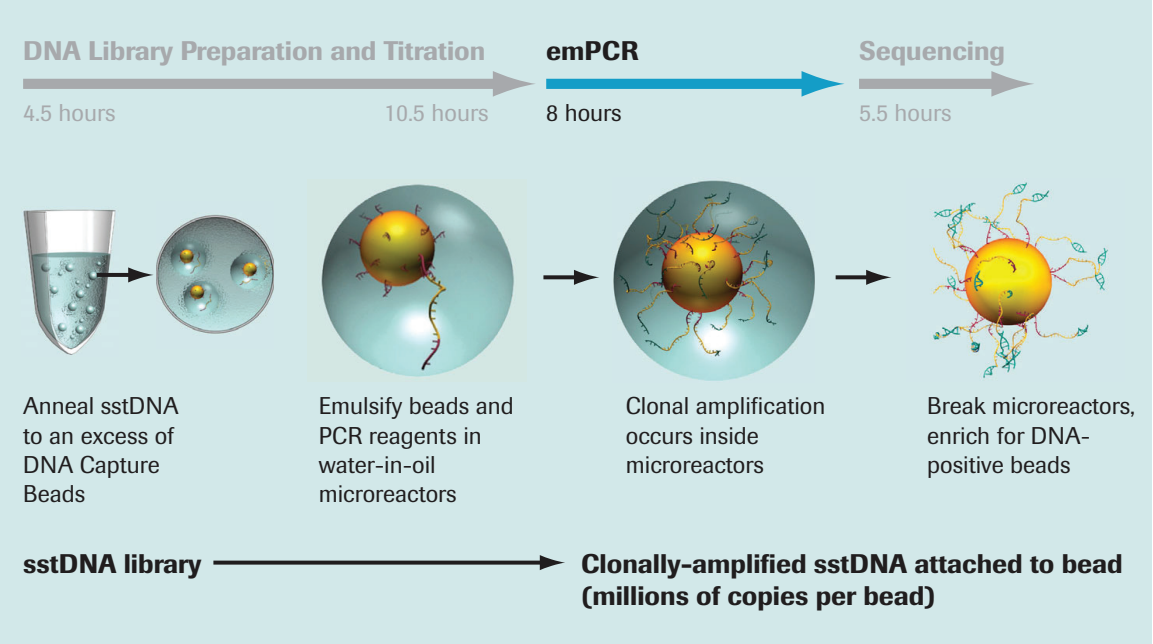


Figure 4: Overview of emulsion-based clonal amplification (emPCR) with the Genome Sequencer 20 System.

The Sequencing Technology

Enhance your sequencing process — from genome to sequence in record time

Process Steps

Sequencing-by-Synthesis

Sequencing starts with the preparation of a PicoTiterPlate device; during this step, a combination of beads, sequencing enzymes, and an sstDNA library is deposited into the wells of the device. The bead-deposition process maximizes the number of wells that contain an individual sstDNA library bead.

The loaded PicoTiterPlate device is placed into the Genome Sequencer 20 Instrument. The fluidics subsystem flows sequencing reagents (containing buffers and nucleotides) across the wells of the plate. Each sequencing cycle consists of flowing individual nucleotides in a fixed order (TACG) across the PicoTiterPlate device. During the nucleotide flow, each of the hundreds of thousands of beads with millions of copies of DNA is sequenced in parallel.

If a nucleotide complementary to the template strand is flowed into a well, the polymerase extends the existing DNA strand by adding nucleotide(s). Addition of one (or more) nucleotide(s) results in a reaction that generates a chemiluminescent signal that is recorded by the CCD camera in the Genome Sequencer 20 Instrument. The signal strength is proportional to the number of nucleotides incorporated in a single nucleotide flow.

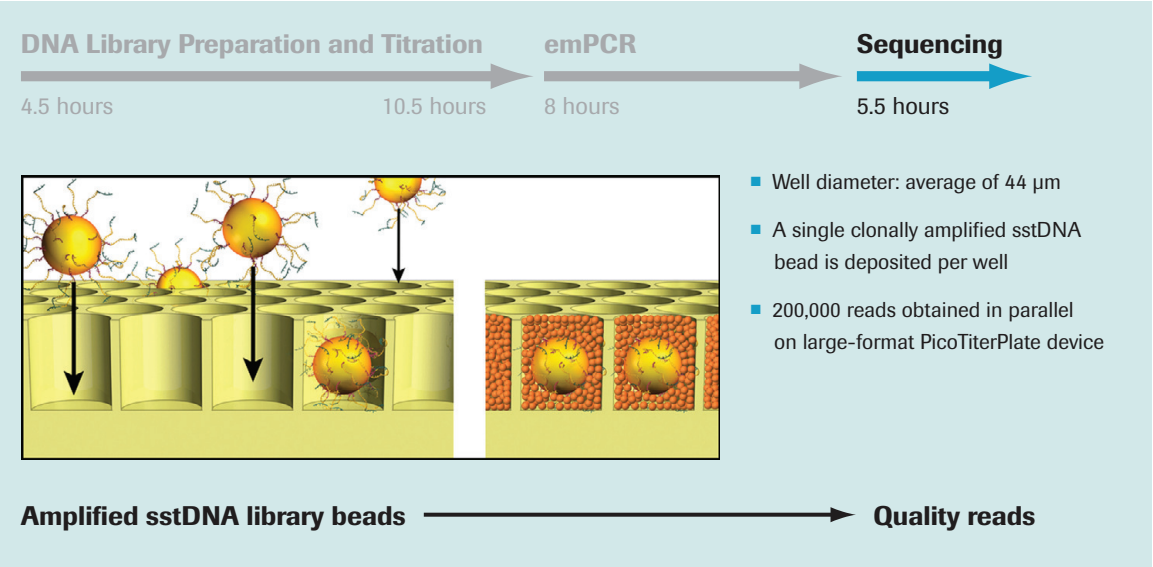


Figure 5: Deposition of DNA beads into the PicoTiterPlate device.

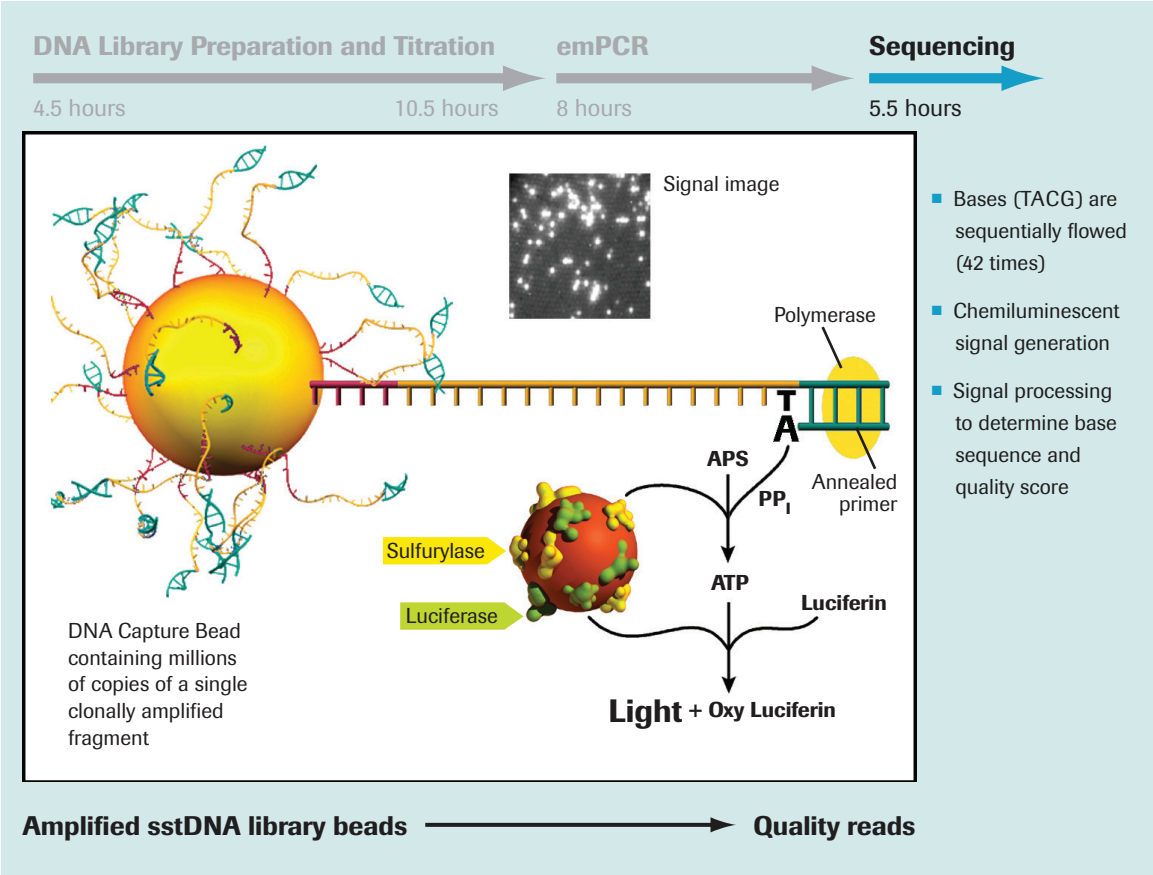


Figure 6: Sequencing reaction of the Genome Sequencer 20 System.

The Genome Sequencer 20 Instrument

Perform ultra-high-throughput DNA sequencing

Genome Sequencer 20 System

The Genome Sequencer 20 System revolutionizes DNA sequencing, delivering sequence data in a massively parallel fashion.

The Genome Sequencer 20 System includes:

- Instrument and accessories
- Reagents and consumables for library construction, amplification, and sequencing
- Analysis software for resequencing, *de novo* assembly, and amplicon sequencing.

The instrument (Figure 7) is the centerpiece of the Genome Sequencer 20 System. It comprises both optics and fluidics subsystems, which are controlled by a computer subsystem.

The fluidics subsystem consists of a reagents cassette, a sipper manifold, pumps, valves, and debubblers. It ensures accurate reagent dispensing and flows the sequencing reagents across the wells of the PicoTiterPlate device.

The optics subsystem includes a CCD camera, which captures the light signal resulting from the sequencing reaction (Figure 8).

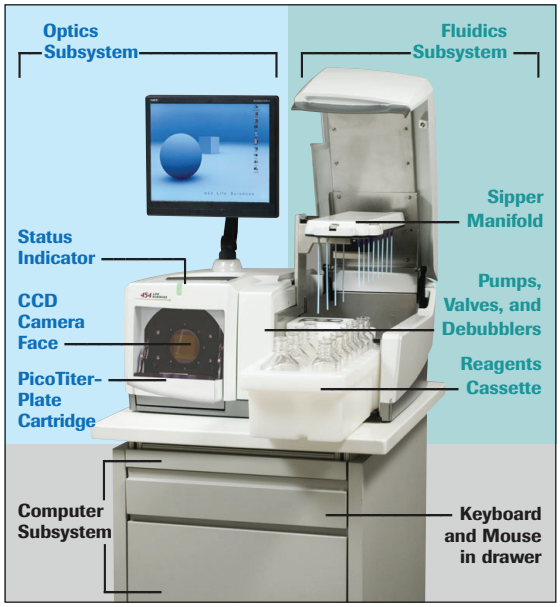


Figure 7: Open view of the Genome Sequencer 20 Instrument.

Instrument space requirement: 30 in (77 cm) Wide x 36 in (92 cm) Deep x 69 in (176 cm) High (cart and instrument).

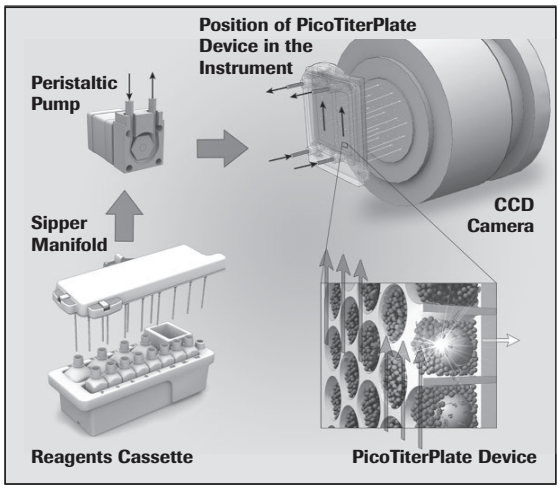


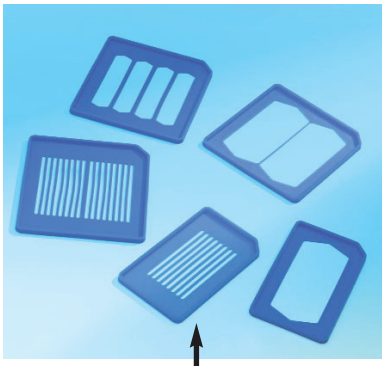
Figure 8: Expanded view of the Genome Sequencer 20 Instrument components. Arrows represent reagent flow.

Illumina Exhibit 1028, Page 10 of 40

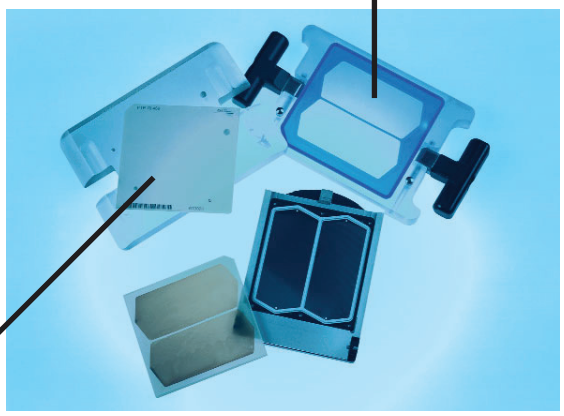
PicoTiterPlate Devices and Gaskets

Sequence more than 20 million bases in a single run with the specially designed PicoTiterPlate device (Figure 9). Choose from two different plate sizes and five different gaskets to meet your specific sequencing throughput needs (Table 1).

- The PicoTiterPlate device is created from fiberoptic bundles that are etched to produce individual wells in picoliter format.
- Each well is only able to accept a single DNA bead.
- Signals generated by reactions in the wells are captured by the CCD camera.
- Gaskets are used to divide the PicoTiterPlate device into separate regions and create loading areas for different throughput needs.



Available formats of PicoTiterPlate device gaskets.



Close-up view of a PicoTiterPlate device.

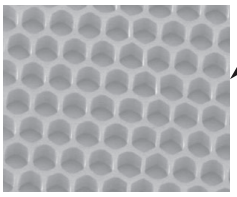


Figure 9: PicoTiterPlate device and accompanying accessories.

Loading region size	PicoTiterPlate device size	Number of regions per PicoTiterPlate device	Run throughput* per region	Run throughput per PicoTiterPlate device
Large: 30 x 60 mm	70 x 75 mm	2	10 Mbp	20 Mbp
	40 x 75 mm	1		10 Mbp
Medium: 14 x 43 mm	70 x 75 mm	4	3.3 Mbp	13 Mbp
Small: 2 x 53 mm	70 x 75 mm	16	0.63 Mbp	10 Mbp
	40 x 75 mm	8		5 Mbp

Table 1: Sequencing throughput with various combinations of PicoTiterPlate devices and gaskets.
* minimum achievable throughput Mbp = million base pairs

Illumina Exhibit 1028, Page 11 of 40

Software

Benefit from a fully integrated software package

Software

The combination of signal intensity and positional information generated across the PicoTiterPlate device allows the Linux-based software to determine the sequence of hundreds of thousands of individual reactions simultaneously, producing millions of bases of sequence per hour from a single run (Figure 10).

Flowgrams and Base Calling

A Flowgram is the graphic representation of the sequence of flowgram signals from a single well of the PicoTiterPlate device, which will be translated into the nucleotide sequence for an individual read.

The signal intensity is proportional to the number of nucleotides incorporated. The bases of the consensus sequence are called by averaging all flowgram signals of the individual reads.

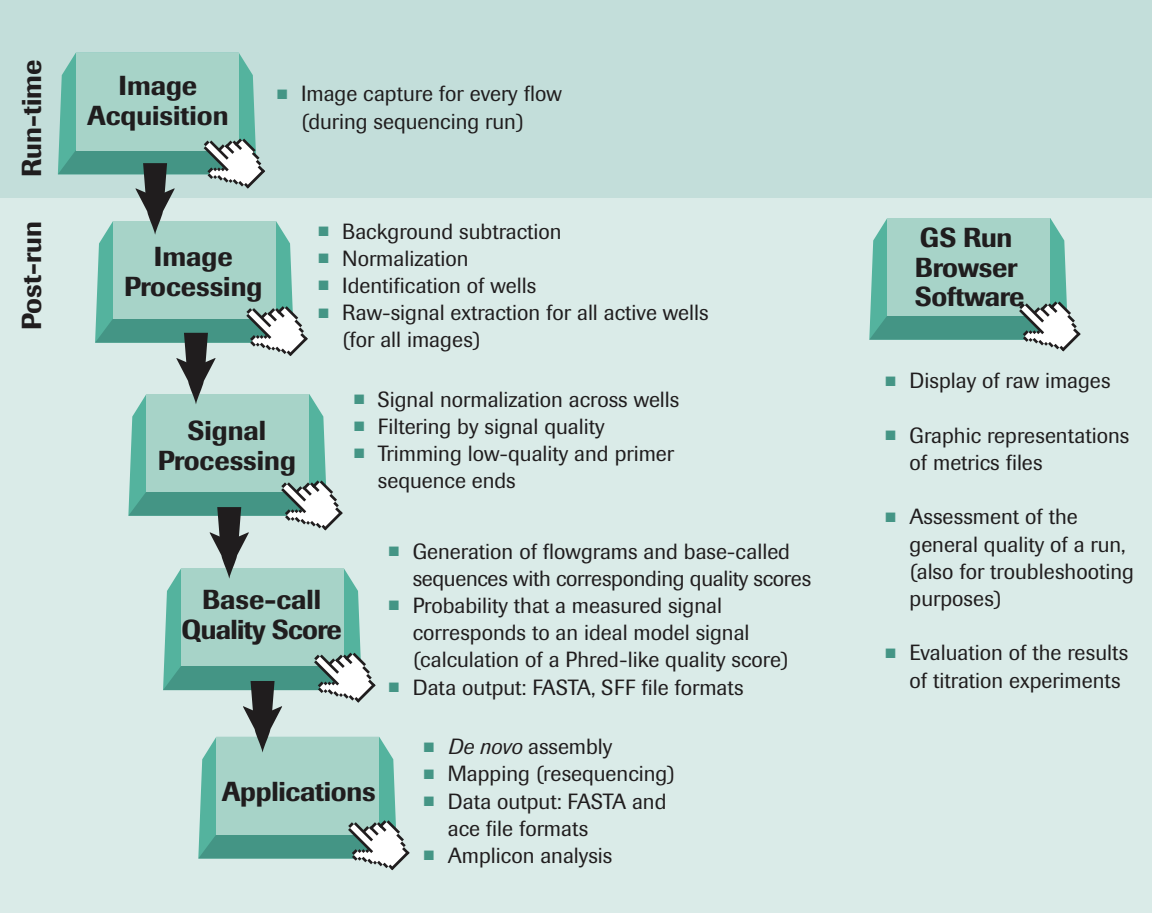


Figure 10: Data processing and analysis output of the Genome Sequencer 20 System.

Illumina Exhibit 1028, Page 12 of 40

Data Processing and Analysis Output

After the signal is captured by the CCD camera, it undergoes image processing. The data-processing output for mapping and assembly includes normalized signals across the wells, flowgrams, and base-called sequences. In addition, a Phred-like quality score is calculated. The data-analysis output results in a consensus sequence based on the flowgram information from all wells of the run or a pool of runs, followed by base calling of the consensus sequence. All raw data is accessible, and the file formats are compatible with publicly available sequencing analysis tools (Figure 11).

Results Assessment

GS Run Browser software is part of the standard software package provided with the Genome Sequencer 20 Instrument. It is an inter-active application that allows the user to view the results of a GS 20 sequencing run, and it displays raw images and graphic representations of various metrics files.

GS Run Browser software can be used to assess the general quality of a run, and therefore is a useful tool for troubleshooting if problems are observed. The application also facilitates evaluation of the results of titration experiments. Most of the data generated by GS Run Browser software can also be exported to an Excel spreadsheet.

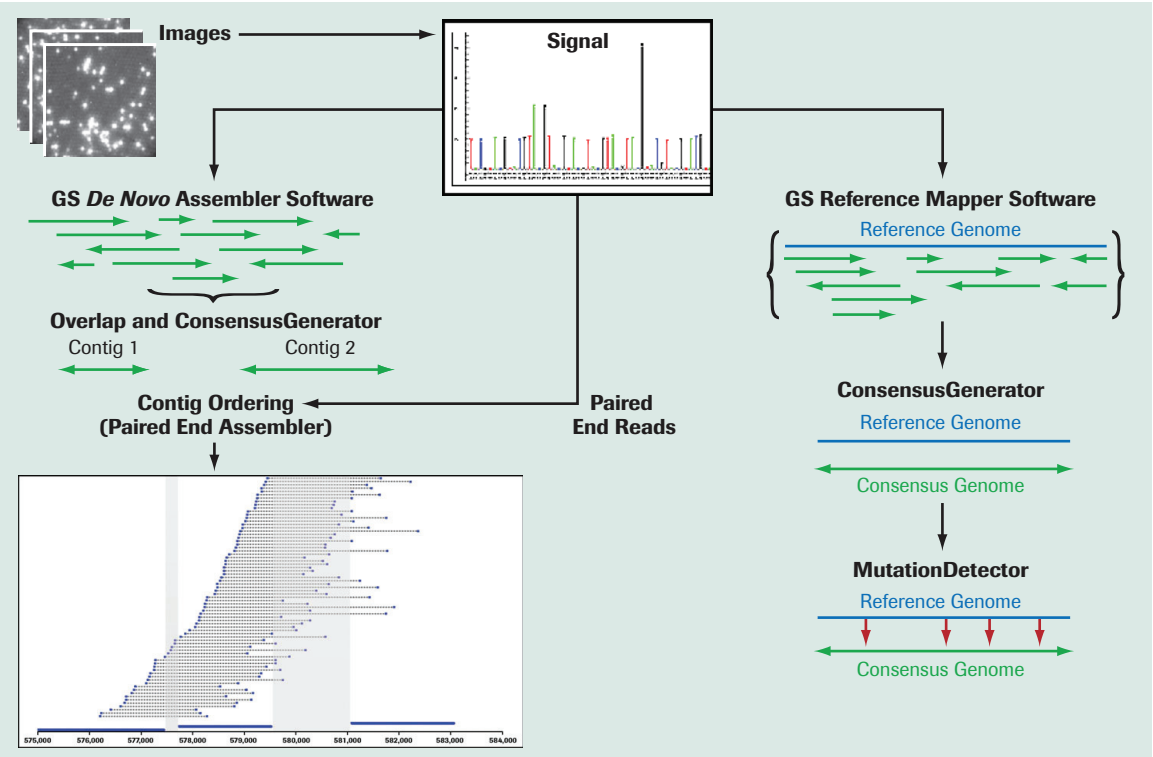


Figure 11: Bioinformatics flow process.

Software

Illumina Exhibit 1028, Page 13 of 40

Applications

Expand your versatility

Applications

The Genome Sequencer 20 System uses a revolutionary technology, deciphering more than 20 megabases in 5.5 hours on a single instrument. This powerful system enables the following applications:

Whole genome sequencing

- *De novo* whole genome shotgun sequencing and resequencing of microbial genomes and BAC clones
- Organization of contigs into scaffolds by a paired-end assembly approach

Transcriptome and gene regulation studies

- High-throughput transcriptome analysis based on short tags, ESTs, ChIP, or GIS-PET sequencing, or the genome-wide identification of miRNA sequences
- Investigation of gene regulation by studying DNA methylation patterns

Amplicon analysis

- Ultra-deep sequencing of PCR products (resequencing for medical research) for
  - — Identification of somatic mutations in complex cancer samples
  - — High-confidence SNP discovery on a population level

Whole Genome Mapping (Resequencing)

Introduction

An integrated software pipeline performs whole genome mapping on the Genome Sequencer 20 Instrument. This pipeline consists of three modules — GS Reference Mapper software, ConsensusGenerator, and a high-confidence MutationDetector (Figure 11).

The GS Reference Mapper software allows users to map individual reads to a reference genome up to one megabase in size. Runs can be combined: for example, up to 200 million bases can be mapped, yielding up to 20x coverage of a 10-million-base genome. The GS Reference Mapper software creates an ideal flowgram signal space genome from the reference genome for comparison with the reads. By operating in flowgram signal space, the mapping process is able to utilize the volume of information contained in the flowgram signals (such as the information inherent in the negative flows as well as positive flows) that is partially lost after base calling (conversion to nucleotide space). After mapping the individual reads, the ConsensusGenerator program combines the individual read information to create a higher confidence base call using overlapping reads. The MutationDetector sorts through the consensus base calls to list high-confidence differences from the reference genome.

Resequencing – Applications

— Comparative genomics
- Identify single base mutations
- Identify mutation hotspots and conserved regions
- Identify inserted or deleted genes
- Assess gene correlations or sequence deviations with observable traits (*e.g.*, understand the genetic basis for drug resistance)[2]

Illumina Exhibit 1028, Page 14 of 40

- Study virulence prediction based on gene sequence variation
- Perform epidemiological analysis
- Understand the genetic difference between industrial producer strains and their corresponding parental strains as the basis for producer strain development
- Perform metagenomic analysis based on shotgun sequencing of environmental DNA and subsequent mapping against known microbial genome sequences[3, 4, 5]
- Sequence ancient DNA (*e.g.*, shotgun sequencing of the woolly mammoth genome)[6]

Example: Resequencing of a variety of bacterial genomes

Resequencing results obtained with the Genome Sequencer 20 System across a variety of bacterial genomes are presented in Table 2. After whole genome sequencing of bacterial strains up to an x-fold coverage, consensus contigs were generated based on mapping of raw reads against the corresponding reference sequence. As the results demonstrate, the Genome Sequencer 20 System achieves a high degree of coverage across the genomes with a high degree of concordance with the published genomes. This approach was used, for example, by Johnson & Johnson Pharmaceutical Research and Development to find point mutations in multiple bacteria.[2] Breaks in the scaffold are the result of incomplete coverage of the genome due to random chance and repeat regions that are longer than the raw sequence reads, and thus cannot be uniquely anchored. The contigs generated by the GS Reference Mapper software are provided in standard file formats and are readily incorporated into standard sequence viewers or assembly tools.

	M. genitalium	*B. licheniformis*	*E. coli*	*S. pneumoniae*	*S. coelicolor*	*S. cerevisiae*
Genome Size (bases)	580,069	4,222,645	4,639,675	2,160,837	8,667,507	12,070,820
Coverage Depth (x-fold)	20.9	22	23.1	21.6	19.8	21.1
Number of Contigs	8	86	119	116	349	694
Average Contig Size (kb)	72.404	48.54	38.293	17.71	24.468	16.547
Size of Largest Contig (kb)	168.532	256.696	327.137	109.203	160.315	266.004
Total Genomic Coverage (%)	99.86	98.88	98.22	95.95	98.52	95.14
Total Genomic Coverage of Non-Repeat Regions (%)	100.00	100.00	100.00	100.00	100.00	100.00
Total Accuracy of Consensus Sequence (%)	99.9921	99.9954	99.9964	99.9960	99.9864	99.9439
Number of Runs	0.3	2.5	3.0	1.5	6.0	11.0

Table 2: The uniformity of coverage — achieved with the Genome Sequencer 20 System — on a number of bacterial genomes. The results shown are mapping of reads to a known genome. As Coverage of Non-Repeat Regions demonstrates, the Genome Sequencer 20 System process achieves a high degree of coverage across the genomes. The repeats are excluded from the mapping results, as they are not uniquely mapped with 100 bp reads.

Illumina Exhibit 1028, Page 15 of 40

Whole Genome Sequencing
Expand your versatility

Applications

De Novo Assembly of Whole Genomes

GS *De Novo* Assembler software is the new *de novo* assembly software for use with the Genome Sequencer 20 Instrument. Exploiting the inherent advantages of the GS 20 Instrument's performance, the GS *De Novo* Assembler software operates in flowgram signal space, as opposed to the standard nucleotide space. By operating in flowgram signal space, GS *De Novo* Assembler software is able to utilize the abundant information stored in the flowgram signals that is lost after base calling (conversion to nucleotide space).

GS *De Novo* Assembler software has three main functions: overlap generation, contig layout, and consensus generation. The overlap generator aligns raw reads in flowgram signal space using a proprietary algorithm. Consensus generation is based on signal averaging where all aligned flowgram signals at each position are averaged and the final base call is performed on the averaged signal. The signal averaging allows higher quality consensus base calls (Figure 11).

De novo Sequencing – Applications

— Unknown microorganisms up to 50 Mb
- Generate an overview of the genome structure
- Study DNA sequence organization, distribution, and information content
- Conduct gene surveys: novelty, locations, and functions
- Compare to other organisms and correlate with observable traits

— BACs, YACs
- Sequence BAC clones, for example, as the basis for whole genome sequencing of plants and animals

— Unknown viruses

— Paired End Assembly

Example: _De novo_ sequencing of a variety of bacterial genomes

The results for *de novo* sequencing of a variety of genomes are shown in Table 3, page 17. The genome assemblies are nearly as comprehensive as the mapping results shown in Table 2, page 15, with the vast majority of bases in the assemblies correct (as measured by their concordance with the published reference genome). As with the mapping results, the breaks in the contigs occur as a result of random chance and at the boundaries of repeats. The contigs generated by the GS *De Novo* Assembler software are provided in standard file formats and are readily incorporated into standard sequence viewers or assembly tools.

Illumina Exhibit 1028, Page 16 of 40

	M. genitalium	*B. licheniformis*	*E. coli*	*S. pneumoniae*	*S. coelicolor*	*S. cerevisiae*
Genome Size (bases)	580,069	4,222,645	4,639,675	2,076,278	8,641,205	12,070,820
Coverage Depth (x-fold)	20.68	21.98	23.5	22.37	20.5	25
Number of Contigs	20	136	139	229	1013	717
Average Contig Size (kb)	28.008	30.657	32.603	8.802	8.383	15.817
Size of Largest Contig (kb)	154.741	200.162	163.595	59.578	70.534	98.656
Total Genomic Coverage (%)	96.57%	98.63%	97.45%	92.99%	96.96%	92.86%
Total Genomic Coverage of Non-Repeat Regions (%)	99.42	100.0000	100.0000	100.00	97.9097	100.00
Total Accuracy of Consensus Sequence (%)	99.9940	99.9970	99.9990	99.9950	99.9920	99.9780
Number of Runs	0.3	2.5	3	1.5	6	11
Misassemblies	0	2	3	4	36	16

Table 3: The performance of GS *De Novo* Assembler software, the Genome Sequencer 20 System *de novo* assembler, for sequencing of several bacterial genomes. The genome sequences from all six bacteria are publicly available in GenBank.

Using the Genome Sequencer 20 System, *de novo* sequencing of more than 100 bacterial artificial chromosomes (BACs) per month is feasible. This makes the system a perfect platform for performing large-scale BAC sequencing projects within the framework of

- metagenomics projects;
- resequencing portions of eukaryotic genomes; or
- whole genome sequencing based on a BAC-to-BAC approach (*e.g.*, plant genomes) (Table 4).

	BAC 1	*BAC 2*	*BAC 3*	*BAC 4*
Number of Contigs >500 bp (all contigs)	6 (9)	8 (18)	4 (9)	8 (12)
Total Contig Size (kb)	139.794	119.438	123.051	150.849
Average Contig Size (kb)	23.299	14.929	30.762	18.856
Largest Contig Size (kb)	102.356	35.375	78.305	51.957
Bases >PHRED 40 (%)	99.9	99.9	99.9	99.8
Coverage (%)	29.7	40.2	34.8	20.6

Table 4: Performance of the Genome Sequencer 20 System — *de novo* sequencing of *Brassica napa* BACs.

Illumina Exhibit 1028, Page 17 of 40

Applications

Paired-End Assembly

Facilitate finishing of the high-quality draft sequence

Standard whole genome *de novo* assembly uses the GS *De Novo* Assembler software to assemble reads into contigs. Thereafter, paired-end reads are used to order and determine the relative positions of contigs produced by *de novo* shotgun sequencing and assembly.

The GS Paired End Adaptor Kit provides reagents for the creation of a paired-end library of fragments from a DNA sample. The generated paired-end reads are DNA fragments that have a 44-mer adaptor sequence in the middle flanked by a 20-mer sequence on each side. The two flanking 20-mers are segments of DNA that were originally located approximately 2.5 kb apart in the genome of interest (Figure 13). The ordering and orienting of contigs generates scaffolds which provide a high-quality draft sequence of the genome and simultaneously facilitate finishing of the genome (Figures 12 and 14).

Benefits

- Prepare a paired-end library to order and orient the contigs from your *de novo* sequencing project.
- Generate a paired-end DNA library for many sequencing runs.
- Facilitate finishing of the high-quality draft sequence.

The genomes of three different organisms were shotgun sequenced. The number of assembled contigs was reduced by adding paired-end data from additional sequencing using the Genome Sequencer 20 System. This resulted in a higher coverage of the entire genome (Table 5).

	E. coli	*B. licheniformis*	*S. cerevisiae*
Genome Size	4.6 Mb	4.2 Mb	12.2 Mb
Oversampling	22	27	23
Number of Contigs (unoriented)	140	98	821
Number of Paired-End Reads	112,000	255,000	395,000
Number of Scaffolds	24	9	153
Coverage of Genome (%)	98.6	99.2	93.2

Table 5: Paired-end data of different organisms.

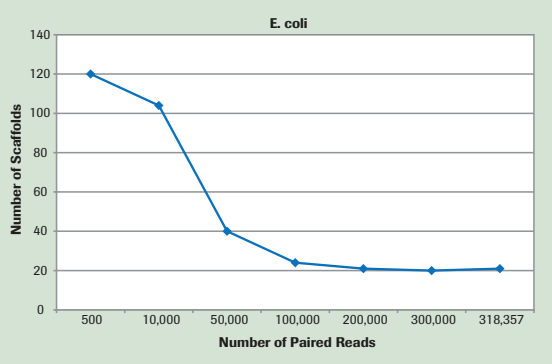


Figure 12: Incremental assembly from paired-end reads and whole genome shotgun reads. The graph shows the number of reads achieved versus the number of scaffolds that were obtained. Using more than 200,000 paired-end reads does not result in further reduction of the number of scaffolds.

Illumina Exhibit 1028, Page 18 of 40

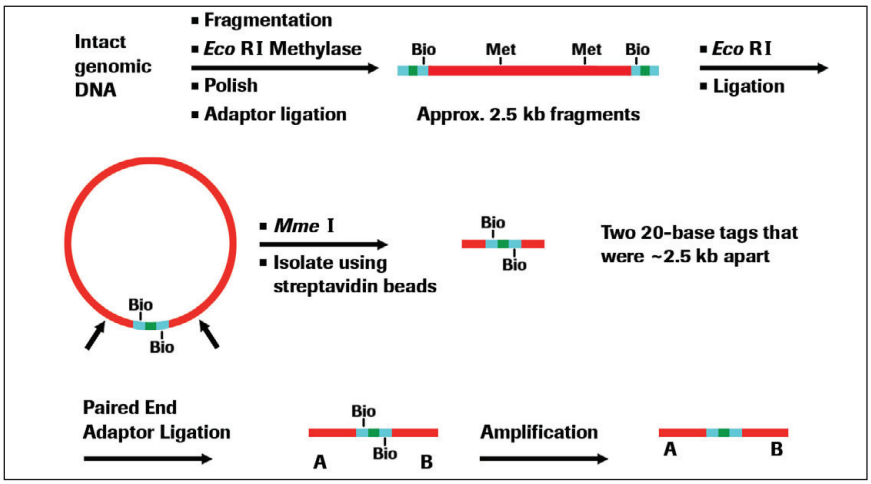


Figure 13: Generation of a paired-end library. Intact genomic DNA is fragmented to yield an average length of 2.5 kb using hydroshearing). The fragmented genomic DNA is methylated with *Eco* RI methylase to protect the *Eco* RI restriction sites. The ends of the fragments are blunt-ended, polished, and an adaptor DNA oligo is blunt-end ligated onto both ends of the digested DNA fragments. Subsequent digestion with *Eco* RI cleaves a portion of the adaptor DNA, leaving sticky ends. The fragments are circularized and ligated, resulting in 2.5 kb circular fragments. The adaptor DNA contains biotin tags and two *Mme* I restriction sites; after treatment with *Mme* I, the circularized DNA is cleaved 20 nucleotides away from the restriction sites in the adaptor DNA. This digestion generates small DNA fragments that have the adaptor DNA in the middle and 20 nucleotides of genomic DNA that were once approximately 2.5 kb apart on each end. These small, biotinylated DNA fragments are purified from the rest of the genomic DNA using streptavidin beads. The purified paired-end fragments are processed using the standard library-preparation protocol for the Genome Sequencer 20 System (see page 6, Figure 3).

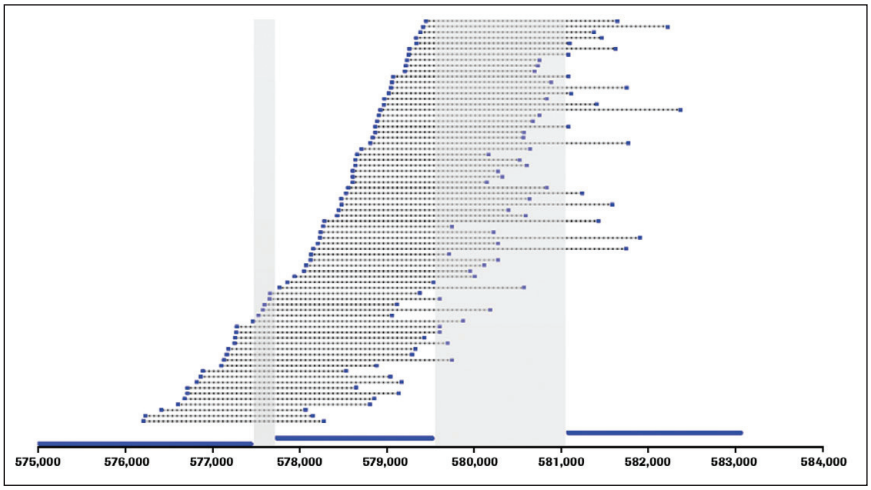


Figure 14: Schematic view of paired-end reads that are used to orient and order contigs and build scaffolds.

Transcriptome and Gene Regulation Studies

Expand your versatility

Applications

Introduction

The Genome Sequencer 20 System enables the study of transcriptomes with outstanding depth of coverage and sensitivity. This is due to the system's massively parallel sequencing technology which generates a high number of sequence reads (minimum of 200,000 ESTs per five-hour run). As a result, sequencing of transcriptomes is now possible up to a previously unattainable sequence coverage in a very short period of time, facilitating the identification of previously unknown transcripts.[7] Another application using Genome Sequencer 20 technology is the genome-wide identification of small non-coding RNAs. Several publications prove that this new technology offers a straightforward method for genome-wide identification of completely unknown groups of small non-coding RNAs.[3, 8, 9]

In addition, the Genome Sequencer 20 technology facilitates gene expression studies. Ditags or longer tags (*e.g.*, EST tags) can be sequenced very rapidly at a very high throughput, providing information on the types of genes expressed and alternative start or termination sites, as well as revealing information about expression levels.

Binding sites of DNA-binding proteins, such as transcription factors, can now be identified without the use of microarrays[7]; DNA fragments that include binding-site sequences can be isolated after immunoprecipitation with their associated transcription factors and characterized using high-throughput sequencing.

Overview of Transcription and Gene Sequencing Applications using the Genome Sequencer 20 System

Identification and quantification of transcripts	Gene regulation
- Sequence small cDNA tags or GIS-PET[10] tags - Sequence larger cDNA tags such as 5´ or 3´ ESTs (100 base pairs)	- Identify miRNAs, siRNAs, or other small non-coding RNAs (sncRNAs) on a genome-wide level - Study changes in methylation patterns - Identify transcription factor binding sites - Identify chromatin binding sites

Sample preparation

At least 200,000 sequence reads can be generated in a five-hour run with the Genome Sequencer 20 System; therefore, 200,000 ESTs can be sequenced during a single run. Generally, the substrate used for sequencing consists of low molecular weight cDNA molecules (*e.g.*, ditags or short PCR products carrying the sequence information of the 5´ or the 3´ end of mRNA molecules). For example, during genome-wide identification and quantification of small non-coding RNA (sncRNA) molecules (*e.g.*, miRNA or siRNA), short cDNA molecules generated from small RNA (isolated using gel fractionation methods) are used as input.

Illumina Exhibit 1028, Page 20 of 40

Two methods can be employed to add the A and B adaptor sequences (needed for emulsion PCR and sequencing) to the cDNA molecules (Figure 15).

1. Ligation of the 44-mer standard A and B adaptors (included in the GS DNA Library Preparation Kit) to cDNA molecules
2. Incorporation of A and B adaptor sequences during cDNA generation

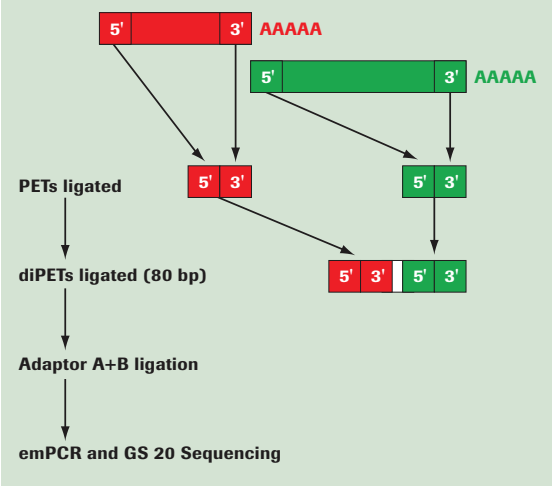


Figure 15: Substrates for the high-throughput sequencing pipeline are typically low molecular weight molecules representing transcripts (in this case, paired-end tags [PETs] carrying information on the 5´ and the 3´ end of mRNAs). Once the Genome Sequencer 20-specific adaptors are ligated to these tags, straightforward sequencing of hundreds of thousands of these molecules is possible.

Develop a Superior Understanding of Transcriptomes — Benefits at a Glance

Make New Discoveries

- Discover expressed genes in known and unknown genomes at an unprecedented degree of depth with low cost per clonal read.
- Uncover altered start and termination sites, translocations, and trans-splicing products with exceptional sensitivity and accuracy.
- Perform gene identification and quantification studies simultaneously, accelerating the analysis of unknown transcriptomes.

Drive the understanding of gene regulation

- Identify genome-wide transcription factor binding sites with unprecedented efficiency and sensitivity.
- Analyze gene transcription control by identifying and characterizing DNA methylation patterns more accurately.
- Generate a comprehensive genome-wide map of small non-coding RNAs (*e.g.*, miRNA) and their different classes in days.

Illumina Exhibit 1028, Page 21 of 40

Transcriptome and Gene Regulation Studies

Expand your versatility

Identification and Quantification of Transcripts

I. Sequencing of ESTs (Expressed Sequence Tags)

The Genome Sequencer 20 System is perfectly suited for analysis of 5' as well as 3' ends of cDNA molecules. The 5' sequencing of full-length cDNA molecules was developed by Yutaka Suzuki (University of Tokyo, Department of Medical Genome Science). This method extends the DNA tag sequencing and mapping strategy of short ditags by using the full-read capability of GS 20 to generate 100 bp ESTs at the 5' end of cDNA molecules (Figure 16). The advantages of having longer tags include

- Improved mapping (sequence identification) of cDNAs to known reference sequences (quantification and genome annotation – mapping transcription start sites/splice variants)
- Information on unknown expressed genes (protein functions) from unknown genomes or newly identified transcripts

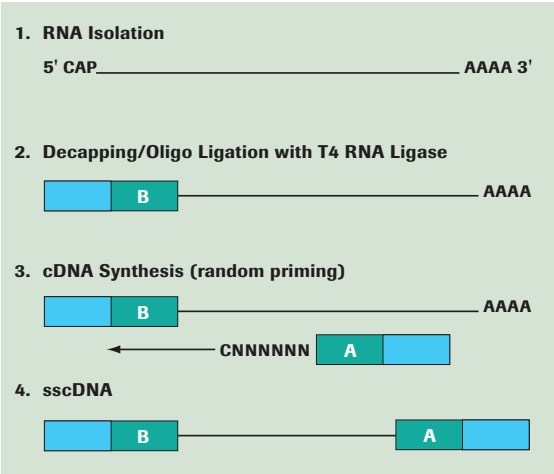


Figure 16: Overview of EST sequencing procedure. A specific oligo sequence (including the B adaptor sequence at the 3' end of the oligo) is added to the 5' end of full-length mRNAs by T4 RNA ligase (Oligo Capping). Subsequently, cDNA synthesis is randomly primed by a hybrid oligonucleotide where the 3' portion of the oligo consists of a random hexamer sequence and the 5' end consists of the specific 19 bp sequence identical to the A adaptor sequence. The resulting single-stranded cDNA (sscDNA) is used directly in the emPCR reaction to generate DNA beads for GS 20 sequencing.

Illumina Exhibit 1028, Page 22 of 40

II: Simultaneous sequencing of 5´ and 3´ ends of full-length cDNA molecules – Genome Identification Signature Sequencing (GIS-PET)

The Genome Identification Signature (GIS) technique was developed by the Singapore Genome Center[10] (Figure 17). GIS is a DNA-tag sequencing and mapping strategy in which 5' and 3' signatures of full-length cDNAs are accurately extracted into paired-end ditags (PETs) that are concatenated for efficient sequencing. The PETs are subsequently mapped to genome sequences to demarcate the transcription boundaries of every gene. GIS analysis is potentially 30-fold more efficient than standard (Sanger) cDNA sequencing approaches or transcriptome characterization. Using this approach, it is also possible to uncover intergenically spliced and unusual fusion transcripts. Paired-end ditagging for transcriptome analysis can also be applied to whole genome analysis of cis-regulatory and other DNA elements and represents an important technological advance for genome annotation.

Summary of Results

- Mapping of GIS-PETs to the human genome
  - 462,626 PETs generated (single run on GS 20)
  - 22,992 PET clusters – 92% map to unique loci
- Identification of altered start and termination sites
  - 125,986 selected PETs (100%)
  - 13% showed altered 3' or 5' start or termination sites (alternative or truncated)
- Identification of novel transcripts
  - 21,153 (100%) unique PET clusters studied
  - 7% novel genes

Data provided by Chia-Lin Wei and Yijun Ruan – Genome Institute of Singapore.

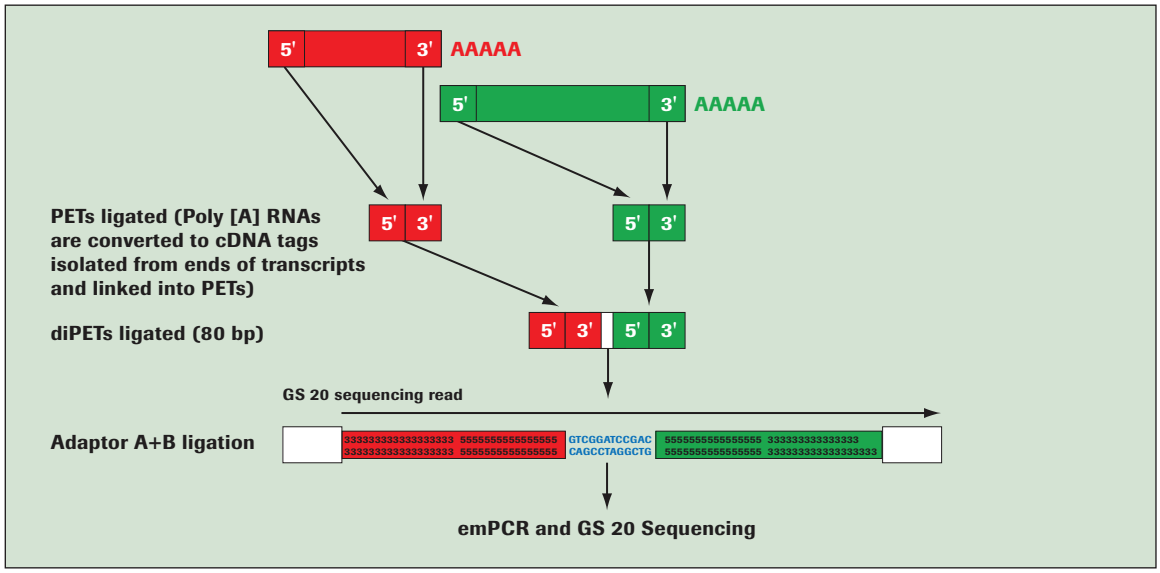


Figure 17: Schematic representation of Genome Identification Signature Sequencing with paired-end ditags (GIS-PET).
Data provided by Chia-Lin Wei and Yijun Ruan – Genome Institute of Singapore.

Illumina Exhibit 1028, Page 23 of 40

Transcriptome and Gene Regulation Studies

Expand your versatility

Applications

Gene Regulation Studies

I. Genome-wide identification of small non-coding (snc) RNAs

Two methods are used to prepare small RNAs for sequencing on the Genome Sequencer 20 Instrument. In one method (Figure 18, Method 1), described in the publication by Henderson *et al.*,[3] small RNA (after conversion into cDNA) is processed using the standard GS library preparation protocol. The GS process-specific A and B adaptors (44-mer adaptors) are blunt-end ligated onto the double-stranded cDNA copy of the small RNA molecules. Afterwards, a single-stranded library is produced (page 6) and transferred to the emPCR step for clonal amplification and subsequent sequencing.

The second method adds GS process-specific A and B adaptors onto the small RNA during a PCR step (Figure 18, Method 2).

During sequencing of sncRNAs using the Sanger approach, double-stranded cDNA copies of the RNA are concatemerized in order to make sequencing more economical. In comparison, the GS 20 approach is much more straightforward because the often complicated concatemerization step can be omitted. Moreover, cost per clonal read is much lower for the Genome Sequencer 20 System, thus providing a real basis for screening for sncRNA on a genome-wide level.

For example, Girad *et al.*[8] used the system in order to characterize a new class of small RNAs, called piRNAs, in mouse testis. More than 87,000 reads were generated, approximately 53,000 of which would be classified as candidate piRNAs. Other examples where the GS 20 System has been successfully used for the characterization of sncRNAs include the genome-wide analysis of an *Arabidopsis thaliana* dicer mutant[3] or the characterization of the piRNA complex from rat testes.[9]

II. Identification of binding sites of transcription factors or other DNA-binding proteins

To gain a better understanding of the regulation of transcription in the transcriptome, the ability to derive a whole-genome map of transcription factor binding sites (TFBS) is crucial.

Using the Genome Sequencer 20 System, it is now possible to combine chromatin immunoprecipitation (ChIP) with high-throughput sequencing. For this application, cells are treated with formaldehyde in order to crosslink genomic DNA and DNA binding proteins, such as transcription factors. The crosslinked cells are then lysed and sonicated. Subsequently, the transcription factors of interest (or other DNA binding proteins) are immunoprecipitated together with the corresponding DNA (including the binding site) using transcription factor-specific monoclonal antibodies. After removal of the proteins, paired-end tags of the remaining DNA fragments can be generated and sequenced in a high-throughput manner. Using this approach, 57 p53 binding sites were identified in HCT116 cells on the genome level.[7]

Illumina Exhibit 1028, Page 24 of 40

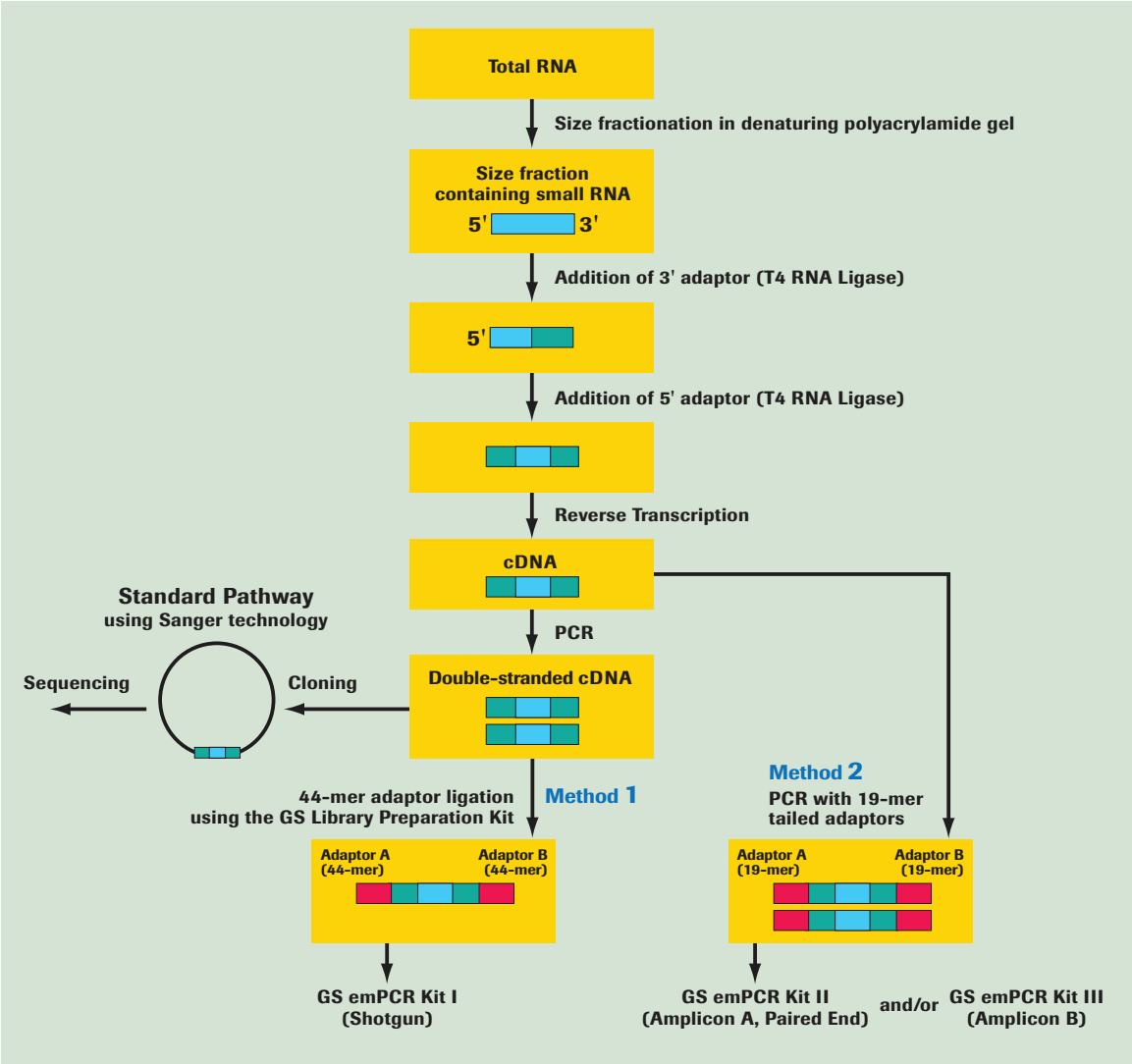


Figure 18: Sequencing of small non-coding (snc) RNA. The figure illustrates two methods for preparing sncRNA for sequencing. In each method, adaptors are added to the sncRNA and transcribed into cDNA.

Method 1. The second strand is synthesized and Adaptors A and B (included in the GS DNA Library Preparation Kit) are ligated onto the double-stranded cDNA. Subsequently, emPCR is performed using the GS emPCR Kit I (Shotgun).

Method 2. PCR is performed using cDNA, and primers that include the sequences for the A and B Adaptors. Subsequently, emPCR is performed with the GS emPCR Kit II (Amplicon A, Paired End) or GS emPCR Kit III (Amplicon B), or both kits, depending on application needs (see Table 7, page 37). Each of the methods results in a clonally amplified sstDNA fragment attached to a bead, ready for analysis in a sequencing run.

Illumina Exhibit 1028, Page 25 of 40

Applications

III. Analysis of DNA methylation patterns using the Genome Sequencer 20 System

Loss of methylation, as well as hypermethylation, of CpG islands within promoter regions is known to be a very important regulation mechanism for many genes. Genome methylation occurs at cytosine residues located 5' to a guanosine in a CpG dinucleotide. Dense areas of CpG dinucleotides within promoter regions are organized into CpG islands.

Applying a known bisulfite-treatment procedure, 454 Life Sciences Corporation has established a sequencing-based technology to quantitatively characterize the methylation state of each CpG dinucleotide in a given target genomic sequence (Figure 20).

To better understand how the bisulfite chemistry and GS 20 sequencing perform with clinical research samples, eight colorectal cancer (CRC) tumor samples and their matched normal adjacent tissue (NAT) were analyzed (Figure 21). The frequency of methylation is determined by the formula Fmeth=1-(FreqCtoT), where FreqCtoT is the frequency of conversion of the cytosine to a thymine in a CpG island.

The results obtained in this experiment are supported by the published literature, and demonstrate that a significant percentage of CRCs shows methylation of the p16 CpG island.[11, 12] This proves that the described GS 20 application, which is based on an open system without bias caused by CpG dinucleotides, is a straightforward technique to obtain accurate data regarding the quantification of methylation at each CpG dinucleotide addressed. This data can then be analyzed using powerful statistical tools to determine what residues are important in the development and maintenance of cancer.

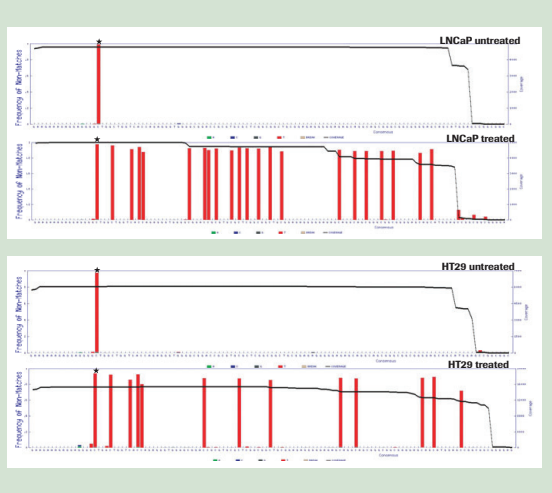


Figure 19: Consensus assembly data from p16 amplicon sequencing from bisulfite-treated genomic DNA from cancer-derived cell lines. The graphs represent the sequencing results from untreated (A and C) and sodium bisulfite-treated (B and D) genomic DNAs derived from an LNCaP prostate cell line (A and B) and an HT29 CRC cell line (C and D). Red bars represent C to T conversion as a result of the bisulfite treatment, and their frequency in the sequence is indicated on the left vertical axis. The horizontal axis is the reference sequence for the p16 gene amplicons used in these studies. The right vertical axis and the black horizontal line indicate the number of sequences that were used in the assembly, and thus cover a given position along the amplicon. Absence of a converted C to T suggests protection from bisulfite treatment due to methylation of the cytosine. Comparison of D to B reveals several sites in D where there is no conversion of cytosine to thymine, representing methylation of those cytosines. Note that the untreated samples (A and C) show no conversion of C to T, demonstrating the specificity of the sequencing. The red bars topped with black stars (★) indicate the nucleotides within the primer used in the PCR; a C to T change from the reference sequence was engineered into those primers.

Illumina Exhibit 1028, Page 26 of 40

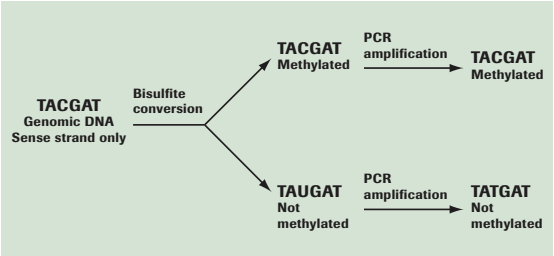


Figure 20: Following extraction from tissue or cells, genomic DNA is treated with sodium bisulfite, which prevents changes to the methylation status of the sample. Treatment of DNA with sodium bisulfite results in the deamination of unmethylated cytosines to uracils, while methylated cytosines remain unchanged. PCR amplification of the converted cytosine (to uracil) results in the substitution of thymine for the uracil. Comparison of the sequence obtained from the bisulfite-treated amplicon to the published sequence using the GS Amplicons Variant Analysis software allows for identification of any differential methylation.

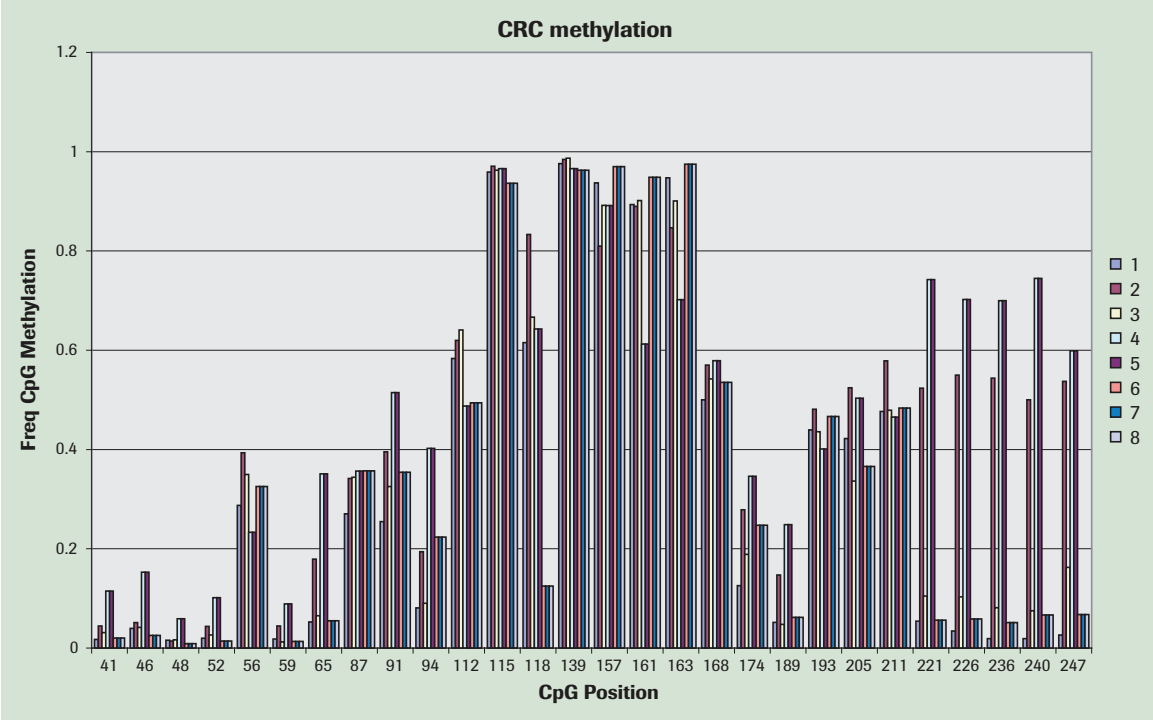


Figure 21: CpG Methylation Frequency in the p16 CpG island in colorectal cancer (CRC) samples. This figure represents the methylation frequencies for each of the CpG dinucleotides assayed in the p16 promoter region in eight separate CRC tumor samples. Frequency of methylation is frequency of C to T conversion subtracted from 1.0, which would represent 100% conversion of C to T when there is no protection of a C from the bisulfite treatment. Some positions (*e.g.*, 139) show a nearly complete methylation of CpG in all the samples, while others show almost no methylation (*e.g.*, position 59). The variability in methylation between the eight samples and from one CpG dinucleotide to another is striking (see positions 221 through 247).

Illumina Exhibit 1028, Page 27 of 40

Amplicon Analysis

Expand your versatility

Applications

Introduction

The power of the Genome Sequencer 20 System derives from utilizing its novel technology to perform many different applications. The new amplicon sequencing tools address scientific questions in medical research, such as the identification of somatic mutations in complex cancer samples or the analysis of genetic variability in microbial, human, animal, or plant populations. These and most other applications using the new emPCR kits for amplicon analysis and corresponding software modules could not be addressed before, due to technical or economic reasons.

Overview of Amplicon Sequencing Applications using the Genome Sequencer 20 System

Cancer Research	Medical Research	Population Genetics
Somatic mutations – Discover rare somatic mutations in complex samples (*e.g.*, cancer biopsies) based on ultra-deep* sequencing of amplicons	**SNP disovery** – Uncover SNPs in pools of genomic DNA using ultra-deep* sequencing of amplicons derived from these pools	**Genomic diversity** – Study genetic diversity in populations based on ultra-deep* sequencing of amplicons derived from pools of genomic DNA

* High-coverage sequencing of PCR products: For example, a 100 bp PCR product is sequenced up to a 1,000-fold coverage if 1,000 single 100 bp reads are generated from this PCR product.

Amplicon Sequencing Technology

Generation of PCR products (sequencing templates)

During PCR with specially designed primers, A and B adaptors are incorporated into the amplicon, and are used as primers for the sequencing reactions. This construct enables sequencing in either A or B direction or in both directions (Figure 22).

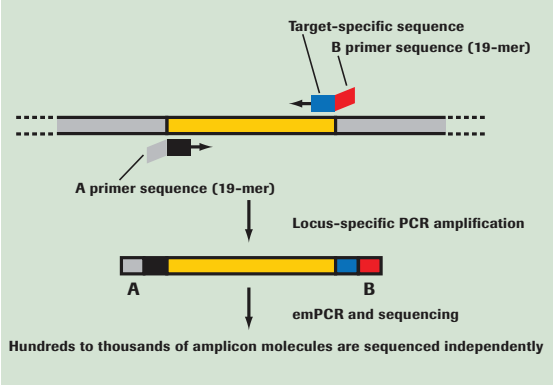


Figure 22: Schematic presentation of amplicon construction. First, the region of interest is amplified using PCR primers containing target-specific sequences and A and B sequences necessary for the sequencing process (locus-specific PCR amplification). Subsequently, emPCR and high-throughput sequencing of single molecules can be performed (hundreds to thousands of amplicon molecules are sequenced independently).

Illumina Exhibit 1028, Page 28 of 40

emPCR ensures sequencing from single molecules (clonal sequencing)

Typically, in order to detect low-frequency mutations in complex samples such as cancer biopsies, the target (*e.g.*, an exon) is amplified first. If the mutation frequency is expected to be less than 20%, subcloning of molecules of the PCR product prior to sequencing is mandatory — this is very tedious and expensive.

With the Genome Sequencer 20 System, molecules of PCR products can be sequenced individually, without the need to subclone in *E. coli* (Figure 23).

This is due to one of the main advantages of the emPCR process, which is that, while the entire procedure is performed batch-wise on the whole library, clonality is achieved by the physical segregation of the DNA-carrying beads in an emulsion during *in vitro* amplification. As a result, the procedure does not require prior biological cloning of the template fragments. Moreover, the achieved clonality avoids the tedious and very expensive manual editing of mixed sequence chromatograms during resequencing projects for medical research, and provides the basis for haplotype identification without subcloning.

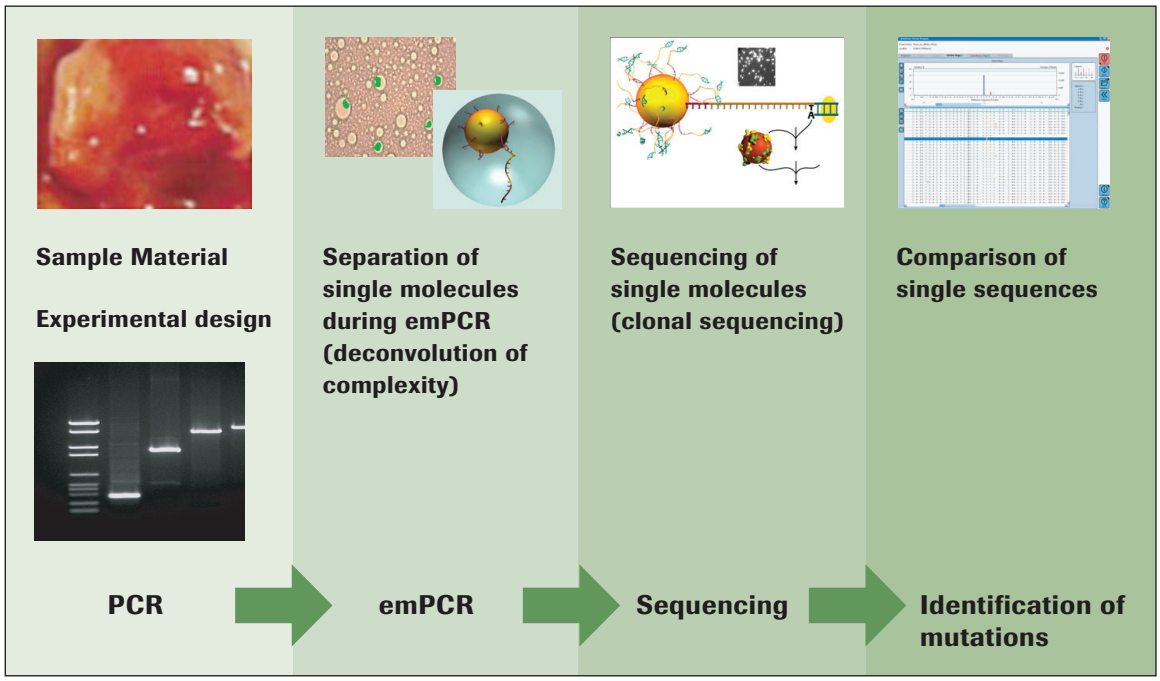


Figure 23: Overview of the amplicon-sequencing process using the Genome Sequencer 20 System. Molecules of PCR products are sequenced individually. This is ensured through separation of the molecules during emPCR, which replaces tedious and expensive cloning in *E. coli*. Hence, the emPCR step provides the basis for massively parallel sequencing from single molecules (clonal sequencing) and very rapid, cost-effective, highly sensitive, and accurate identification of sequence deviations at a single-molecule level.

Illumina Exhibit 1028, Page 29 of 40

Amplicon Analysis

Expand your versatility

Amplicon Sequencing

The new GS emPCR kits for amplicon analysis enable the sequencing of a PCR product from both directions (Figure 24). Consequently, the total read length per amplicon amounts to 200 bases. Sequencing from the A-sequence site toward the bead is possible by using emPCR beads that carry oligonucleotides complementary to the B sequence of the PCR product. Sequencing from the B-sequence site is ensured by using emPCR beads that carry an oligonucleotide which is complementary to the A sequence on the strand to be sequenced.

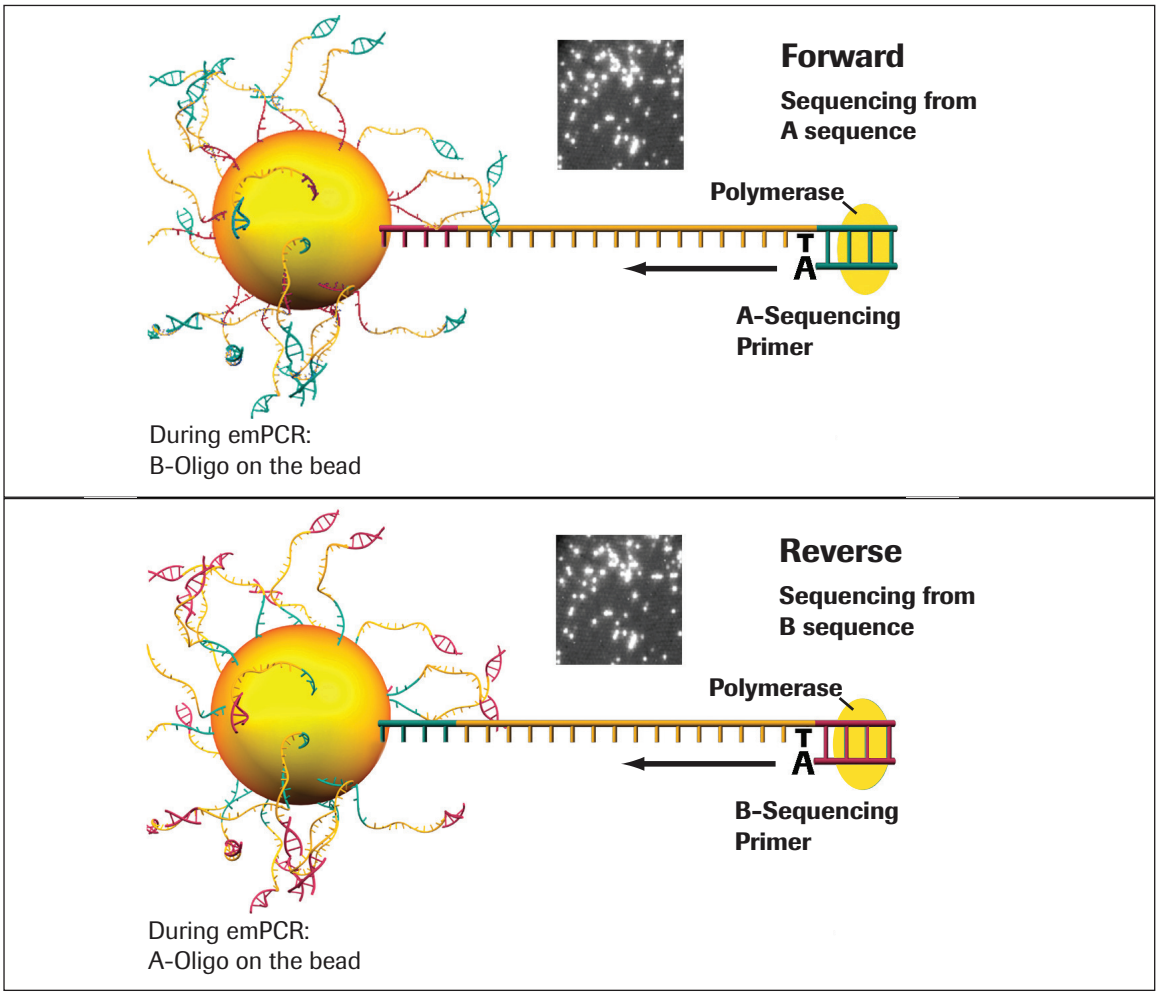


Figure 24: Schematic presentation of GS 20 amplicon sequencing. With the new GS emPCR (amplicon) kits, it is possible to sequence a PCR product in either A or B direction or in both directions (200 base pairs per PCR product).

Illumina Exhibit 1028, Page 30 of 40

Experimental setup

The processing and sequencing of amplicons with the Genome Sequencer 20 System is quite flexible and allows for a wide range of experimental design. The researcher has many options regarding the length of amplicons, the number of amplicons pooled together, the number of reads desired for a given amplicon pool, and whether to read from the A end, the B end, or both. Although the setup for a given experiment will depend on the specific project goals, there are a few general guidelines that will ensure the best possible result:

- Use a high-fidelity polymerase in the amplicon-generation step. Roche's FastStart High Fidelity PCR System features high fidelity coupled with robust amplification of a wide array of input templates.
- Obtain the highest confidence in low frequency variation by using bi-directional reads.
- Perform amplification, pooling, and sequencing as described in the workflow overview below.

Workflow overview

- Design overlapping PCR amplicons (if the region of interest is greater than 100 base pairs in length).
- Generate each amplicon separately.
- Purify and quantify the amplicons.
- Dilute to 200,000 copies per microliter and pool in equimolar ratios.
- Perform emPCR using the GS emPCR Kits II and III in separate reactions.
- Pool the GS emPCR Kit II and III emulsions after completion of the emPCR amplification, breaking, and enrichment procedures.
- Sequence with a format that will generate the desired number of reads/depth of coverage.
- Use the amplicon software to identify and quantify known and *de novo* variations.

Sensitivity

In principle, there is no limit to the number of amplicons that can be pooled. The level of multiplexing and the sequencing format are determined by the desired sensitivity. The following general guidelines regarding desired sensitivity/depth of coverage will determine experimental setup:

- 5% on single base changes and multi-base deletions
  — 1000x
- 1% variation of single base changes and multi-base deletions
  — 5000x

As a result, somatic mutations in complex cancer samples can be identified with unprecedented sensitivity and speed. A recent publication reported on the identification of somatic mutations that were completely missed using the classical Sanger approach.[13]

Illumina Exhibit 1028, Page 31 of 40

Amplicon Analysis

Expand your versatility

Applications

Amplicon Analysis Software

The Genome Sequencer 20 System includes a software module that is specifically designed for amplicon sequencing analysis. This GS Amplicons Variant Analysis software compares the amplicon reads to a reference sequence. Based on this comparison, relative mutation frequencies can be calculated (*e.g.,* 5% of all the sequence reads derived from a cancer biopsy research sample differs from the reference sequence, indicating certain mutations, insertions, or deletions).

Overview of the features of the GS Amplicons Variant Analysis software

- Variation plot for quick identification of sequence variation
- Summary table of variants and their frequencies
- Ability to select each mutation in the variation plot and display the sequences supporting the reported frequency
  - — Ability to access the underlying data set and view the individual reads
  - — Ability to view the individual flowgram(s)
- Enables importation of data into existing analysis tools
  - — FASTA file format of trimmed or untrimmed sequence
  - — Base calls with Phred-equivalent quality scoring for individual reads

In the global alignment screen (Figure 25), the variation histogram plot shows the variations of the consensus sequence in comparison to a given reference sequence. The percentage of the variation and the coverage can be easily identified.

In addition to viewing the consensus reads, users can view the alignment of all individual sequencing reads to consensus reads.

The Tri-Flowgram view of an individual read can be displayed by selecting the read sequence and choosing "open flowgrams". The differences compared to a reference sequence can also be analyzed using the Tri-Flowgram screen (Figure 26).

Accelerate Amplicon Analysis — Benefits at a Glance

Accelerate cancer research

- Identify and quantify somatic cancer mutations which were previously undetectable.
- Identify and quantify somatic mutations with unprecedented speed and accuracy.
- Investigate cancer development on the genomic DNA level like never before.

Drive SNP discovery

- Accelerate SNP discovery by screening pooled genomic DNA samples.
- Detect heterozygotes and identify haplotypes without subcloning PCR products.
- Focus on the science as weeks of tedious manual post-run analysis are eliminated.

Revolutionize population genetics

- Uncover genetic variations in populations of microbes, humans, animals, or plants with exceptional speed, accuracy, and sensitivity. Roche Applied Science has broken down the technological and economic hurdles for you.

Illumina Exhibit 1028, Page 32 of 40

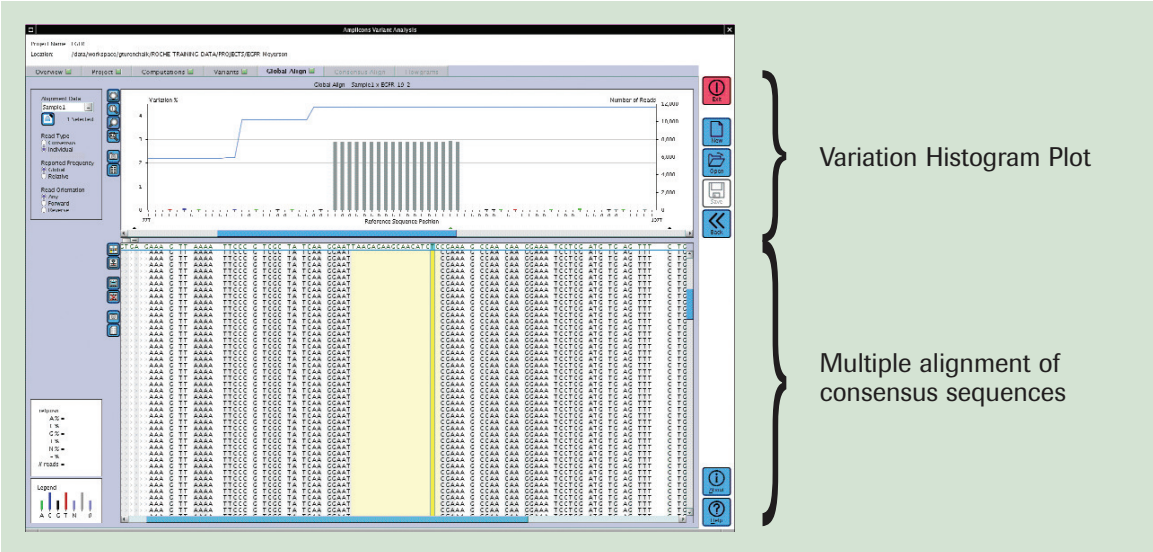


Figure 25: Global Alignment Screen of GS Amplicons Variant Analysis software. The data in this histogram show the Del-4 mutation in exon 19 at a 3% abundance from the experiment described on page 34.

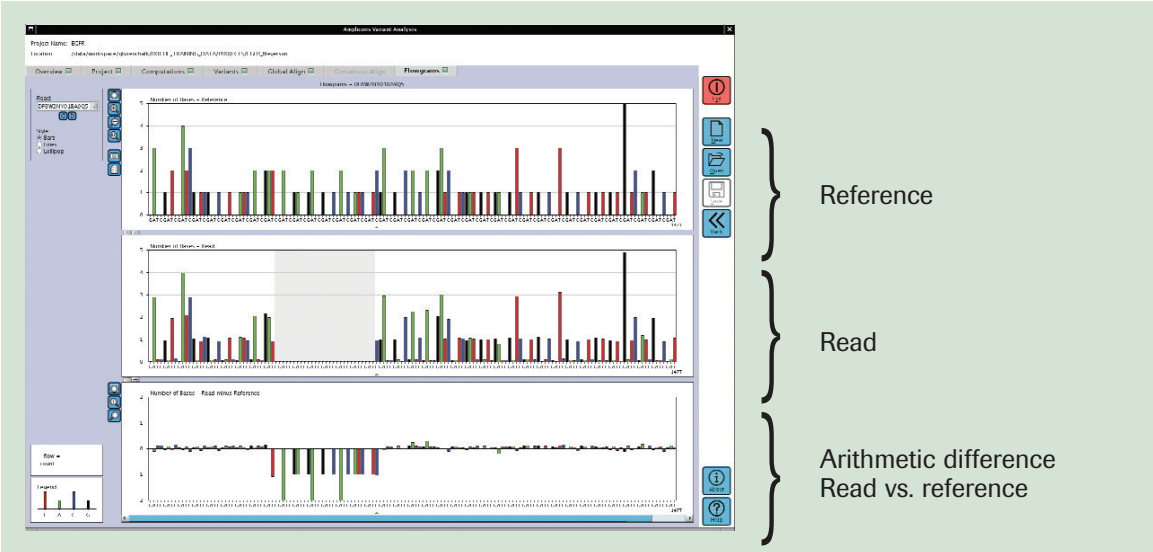


Figure 26: Tri-Flowgram Screen. In this screen, the flowgrams of the read and the reference sequences and the subtraction of the reference from the read sequence are displayed. The data show the Del-4 mutation in exon 19 at a 3% abundance from the experiment described on page 34.

Illumina Exhibit 1028, Page 33 of 40

Amplicon Analysis

Expand your versatility

Examples of Amplicon Sequencing

I. Ultra-deep sequencing to analyze drug resistance in cancer research

Paraffin-embedded pleural effusion material (from a lung adenocarcinoma subject who initially responded strongly to erlotinib treatment) was used as starting material. Sequencing of EGFR exons 18-22, performed as shown in Figure 27, revealed an 18 bp deletion in exon 19 at a frequency of 0.28%, which might account for the subject's early strong TKI response (data not shown).

Following 12.5 months of erlotinib treatment, the subject relapsed with a massive pleural effusion. Pathological examination showed very low tumor content in the isolated sample.

A sample from the same subject was sequenced again and revealed the Del-4 mutation at 3% abundance in exon 19 (Figures 25 and 26).

In addition, a T790M mutation (shown in previous studies to confer resistance to TKI inhibitors) was found at 2% abundance in exon 20 (Figure 28).

The data demonstrates fast and accurate identification of cancer-associated mutations from complex samples at a sensitivity and speed that is unprecedented.[13]

Data provided by M. Meyerson, Department of Medical Oncology, Dana-Faber Cancer Institute, Harvard Medical School, Boston, in cooperation with 454 Life Sciences Corporation.

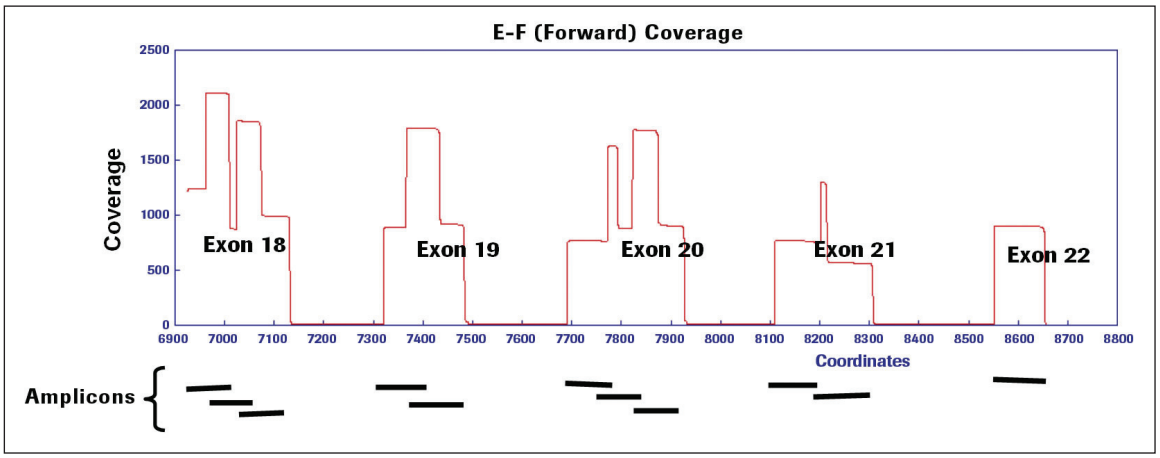


Figure 27: Eleven PCR amplicons, ranging in size between 85 and 156 base pairs, were generated to cover exons 18-22 of the EGF receptor. Each target region was individually amplified and quantified. Prior to emPCR amplification and sequencing, all amplicons were pooled in equimolar ratios.

Illumina Exhibit 1028, Page 34 of 40

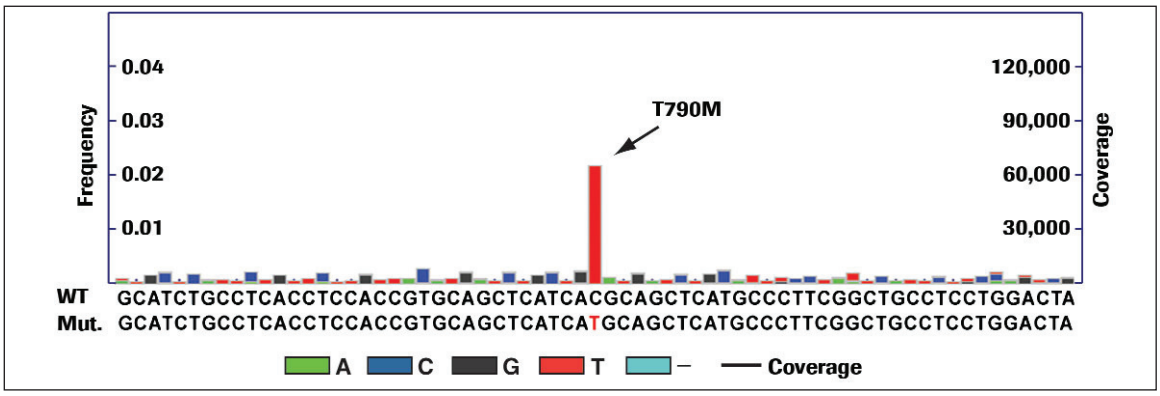


Figure 28: Variation Histogram Plot. The histogram shows a substitution mutation in exon 20, T790M. It is present at a relative allele frequency of approximately 2%. Figure published in *Nature Medicine* - **12**, 852 - 855 (2006) [13]

II: Example of acceleration of SNP discovery by pooling of samples

Typically, SNP discovery is performed based on resequencing of individually amplified genomic regions and direct sequencing of the resulting PCR products — a very time-consuming and expensive approach. Using the Genome Sequencer 20 System and its single-molecule sequencing capabilities, it is now possible to discover SNPs on the population level first (Figure 29). For this application, genomic DNA samples from multiple (*e.g.*, 100) test subjects are pooled. Subsequently, regions of interest, such as exons, are amplified and sequenced at a very high coverage (ultra deep) and sequence variations compared to reference sequence are detected using the GS Amplicons Variant Analysis software (Figure 30). Depending on the type of project, the confirmation of SNPs on the individual level can be performed using the GS 20 System or a SNP genotyping platform such as the LightCycler® Real-Time PCR System.

Data provided by McPherson *et al.* (Human Genome Sequencing Center, Baylor College of Medicine)

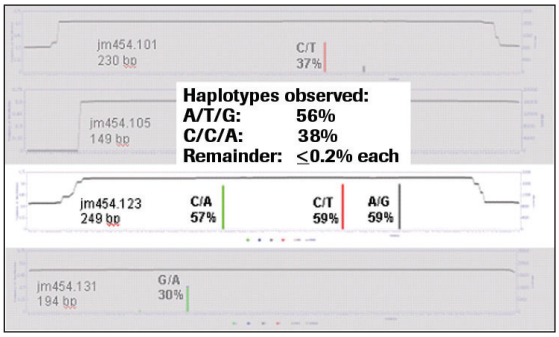


Figure 29: SNP discovery in a population of 92 individuals. The in-depth analysis revealed the corresponding haplotypes.

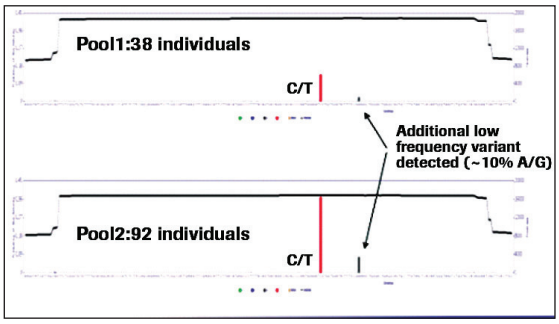


Figure 30: Detection of low frequency variants by ultra-deep sequencing.

Applications

Expand your versatility

Application	GS DNA Library Preparation Kit**	GS Paired End Adaptor Kit	GS emPCR Kit I (Shotgun)**	GS emPCR Kit II (Amplicon A, Paired End)	GS emPCR Kit III (Amplicon B)
Genome Sequencing (including BACs, YACs, plasmids, cosmids, *etc.*)					
Shotgun sequencing	X		X		
Paired-end sequencing (no titration required)		X		X	
Transcriptome/Gene Regulation Analysis					
Small non-coding RNA					
Protocol: direct ligation of cDNA	X		X		
Protocol: PCR with tailed primers				O	O
ESTs (expressed sequence tags)					
Protocol: direct ligation of cDNA	X		X		
Protocol: PCR with tailed primers				O	O
DiTags					
Protocol: direct ligation of cDNA	X		X		
Protocol: PCR with tailed primers				O	O
ChIP fragments* **	X		X		
Protocol: direct ligation of cDNA	X		X		
Amplicon Analysis					
PCR fragments (no titration required)*					
Up to 100 base pairs				O	O
200-500 base pairs				O	O
Long-Range PCR					
PCR products over 1500 base pairs, nebulized	X		X		

Table 6. Overview of applications using the Genome Sequencer 20 System and corresponding kits.

O = Option. Read length averages 100 base pairs. For targets shorter than 100 bp, select either GS emPCR Kit II (Amplicon A, Paired End) or GS emPCR Kit III (Amplicon B) and sequence from only one direction, or use both kits for forward and reverse sequencing in order to increase accuracy. For targets longer than 100 bp, use both GS emPCR Kit II (Amplicon A, Paired End) and GS emPCR Kit III (Amplicon B) (forward and reverse) to read sequence from both ends. As target length extends beyond 200 bases, the ability to sequence the entire length of the amplicon will be diminished.

* Use PCR products that are either shorter than 500 bp for direct use in emPCR, or longer than 1500 bp so they can be sheared using the nebulization protocol prior to emPCR.

** The GS DNA Library Preparation Kit must be used in combination with the GS emPCR Kit I (Shotgun).

Illumina Exhibit 1028, Page 36 of 40

The Genome Sequencer 20 System
Ordering Information

Product	Cat. No.	Pack Size
Genome Sequencer 20 Instrument★	04 760 085 001	1 instrument including accessories
Reagent Kits		
GS DNA Library Preparation Kit	04 852 265 001	1 kit for 10 library preparations
GS Paired End Adaptor Kit	04 891 457 001	1 kit for 10 library preparations
GS emPCR Kit I (Shotgun)	04 852 290 001	1 kit for 16 amplification reactions
GS emPCR Kit II (Amplicon A, Paired End)	04 891 384 001	1 kit for 16 amplification reactions
GS emPCR Kit III (Amplicon B)	04 891 392 001	1 kit for 16 amplification reactions
GS 20 Sequencing Kit (70x75)★	04 853 342 001	1 kit for 1 sequencing run
GS 20 Sequencing Kit (40x75)★	04 852 419 001	1 kit for 1 sequencing run
GS PicoTiterPlate Kit (70x75)★	04 852 427 001	1 plate with accessories
GS PicoTiterPlate Kit (40x75)★	04 746 813 001	1 plate with accessories
GS 20 Maintenance Wash Kit	04 852 397 001	1 kit for 1 maintenance wash run
Additional Accessories		
70x75 Bead Deposition Device (2 large regions)	04 777 069 001	1 bottom plate and 1 top plate
70x75 Bead Deposition Device (4 medium regions)	04 840 712 001	1 bottom plate and 1 top plate
70x75 Bead Deposition Device (16 small regions)	04 840 682 001	1 bottom plate and 1 top plate
40x75 Bead Deposition Device (1 large region)	04 777 034 001	1 bottom plate and 1 top plate
40x75 Bead Deposition Device (8 small regions)	04 777 042 001	1 bottom plate and 1 top plate
Counterweight for Bead Deposition Devices	04 777 115 001	1 counterweight
GS 20 Reagents Cassette	04 825 527 001	1 reagents cassette
GS 20 Reagent Insulator Cover	04 825 551 001	1 reagent insulator cover
Camera Faceplate Guard	04 777 018 001	1 faceplate guard
Nebulizer Holder	04 777 026 001	1 nebulizer holder
Pre-wash Tube Insert	04 865 227 001	1 pre-wash tube insert

Illumina Exhibit 1028, Page 37 of 40

References

1. Margulies, M. *et al.*, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature* **437** (2005): 376-380.

2. Andries, K. *et al.*, "A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*," *Science* **307** (2005):223-227.

3. Henderson, I.R. *et al.*, "Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning," *Nature Genetics* **38**, no. 6 (June 2006): 721-725; advance online publication (doi:10.1038/ng1804).

4. Sogin, M.L. *et al.*, "Microbial diversity in the deep sea and the underexplored 'rare biosphere'" *PNAS* **103**, no. 32 (August 2006):12115-12120.

5. Krause, L. *et al.*, "Finding novel genes in bacterial communities isolated from the environment," *Bioinformatics* **22**, no. 14 (July 2006):e281-e289 (doi:10.1093/bioinformatics/btl247).

6. Poinar, H.N. *et al.*, "Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA," *Science* **311**, no. 5759 (January 2006):392-394.

7. Ng, P. *et al.*, "Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes," *Nucleic Acids Research* **34**, no. 12 (July 2006): e84; advance online publication (doi:10.1093/nar/gkl444).

8. Girard, A. *et al.*, "A germline-specific class of small RNAs binds mammalian Piwi proteins," *Nature* **442** (July 2006):199-202; advance online publication (doi:10.1038/nature04917).

9. Lau, N.C. *et al.*, "Characterization of the piRNA complex from rat testes," *Science* **313**, no. 21 (July 2006):363-367; advance online publication June 15, 2006 (doi:10.1126/science.1130164).

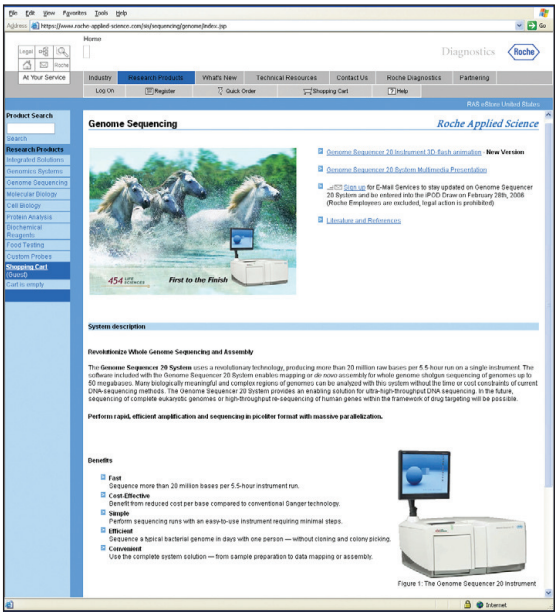
10. Ng, P. *et al.*, "Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation," *Nature Methods* **2** (2005):105-111; advance online publication (doi:10.1038/nmeth733).

11. Kim, B.N. *et al.*, "Methylation and expression of p16INK4 tumor suppressor gene in primary colorectal cancer tissues," *Int J Oncol.* **26** (2001):1217-1226.

12. Herman, J.G. *et al.*, "Methylation-specific PCR: A novel PCR assay for methylation status of CpG islands," *Proc. Natl. Acad Sci.* **93** (1996):9821-9826.

13. Thomas, R.K. *et al.*, "Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing," *Nature Medicine* **12**, no. 7 (July 2006):852-855; advance online publication (doi:10.1038/nm1437).

Illumina Exhibit 1028, Page 38 of 40



For detailed information, references, multimedia presentation, *etc.*, visit ***www.genome-sequencing.com***

For additional information, call **800 262 4911** (for U.S. customers only), or contact your local sales representative.

Limited Label Licenses, Disclaimers, and Trademarks

PICOTITERPLATE is a trademark of 454 Life Sciences Corporation, Branford, CT, USA.

FASTSTART and LIGHTCYCLER are trademarks of Roche.

Other brands or product names are trademarks of their respective holders.

The technology used for the LightCycler® System is licensed from Idaho Technology, Inc., Salt Lake City, UT, USA.

★ **RESTRICTION ON USE:** Purchaser is only authorized to use the Genome Sequencer Instrument with PicoTiterPlate devices supplied by 454 Life Sciences Corporation and in conformity with the procedures contained in the Operator's Manual.

For life-science research only. Not for use in diagnostic procedures.

454 LIFE SCIENCES

Roche Diagnostics

Roche Diagnostics GmbH
Roche Applied Science
68298 Mannheim
Germany

09/06 - 04866452990

www.roche-applied-science.com

© 2006 Roche Diagnostics GmbH. All rights reserved.