

Introduction to Complete Genomics' Sequencing Technology

Complete Genomics Service Advantages

- Accurate human genome data and analysis
- Specific workflows and output for individual samples and tumor-normal pairs
- Flexibility in coverage options; $\geq 40X$ or $\geq 80X$ coverage available
- Affordable for all study sizes
- Simple. Fast. Complete.

Introduction

With advances in nanotechnology, optics, hardware, software, and large-scale computing, it has finally become affordable to sequence a whole human genome at high-quality for genomic-based disease research. Today, most researchers are burdened with purchasing sequencing instruments and developing a data management and analysis infrastructure to handle the massive data from these instruments, instead of focusing on the biological analysis of the data. To get to the next level of genomic discovery and to achieve transformational insights about the genetic basis of human disease, scientists need to be able to conduct genomic research on a large scale. There needs to be a comprehensive, end-to-end service that manages large-scale human genome sequencing in a simple, cost-effective way. An ideal solution should streamline sample preparation, sequencing, data management, and data analysis.

The Complete Genomics Approach

Complete Genomics has developed a sequencing platform capable of generating whole human genome sequence data at an unprecedented level of throughput and low cost. The company deploys this platform through its commercial-scale, fully automated human genome sequencing center and offers comprehensive human genome sequencing services. Complete Genomics provides two primary services, the Standard Sequencing Service and the Cancer Sequencing Service. With each service, customers receive high-quality data including reports on summary statistics; variants including SNPs, indels, copy number variants, structural variants and mobile element insertion events; and reads, scores and mappings. These services allow customers to more efficiently characterize the full spectrum of genetic variants that exist in large numbers of human subjects. This ability to conduct large-scale human genome studies enables customers to elucidate further the genetic underpinnings of complex diseases and drug responses.

Complete Genomics' Sequencing Technology

Complete Genomics has brought together diverse technologies to create a comprehensive solution for large-scale studies of whole human genomes. This solution integrates a sequencing platform that is the combination of technology advancements in libraries, arrays, sequencing assay, instruments and software (Figure 1).

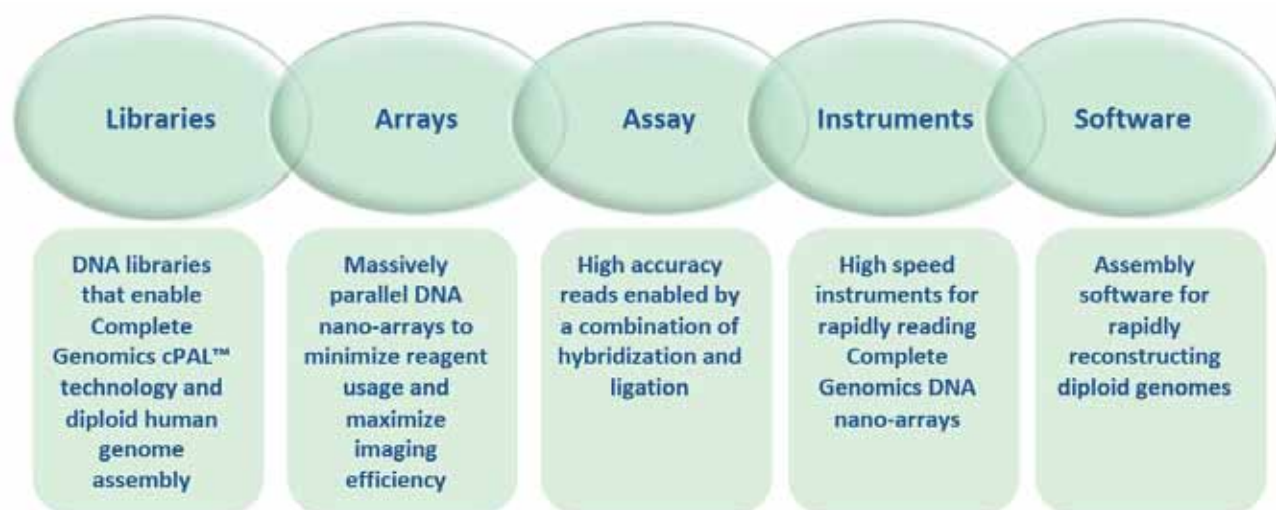


Figure 1. Complete Genomics' Sequencing Technology.

The low reagent usage and high imaging efficiency of Complete Genomics' sequencing platform enable sequencing of whole human genomes at a fraction of the cost of alternative approaches. The accuracy of Complete Genomics' novel sequencing chemistry, combined with custom mapping and assembly techniques, enable the company to resolve many of the complexities of the human genome, and thereby provide the high-quality human genome data sets required to understand complex diseases and drug responses.

Furthermore, Complete Genomics has combined its in-house developed high-throughput sequencing instruments with an enterprise-class data center to create a high-throughput commercial human genome sequencing center. This facility enables Complete Genomics' customers to sequence hundreds or thousands of human genomes efficiently and cost-effectively. With this approach, customers are not burdened with the operational, computational, and capital purchase costs of owning and operating the instruments. Additionally, customers do not need to expend the computing resources necessary for large-scale sequencing of whole human genomes.

Technology Components

Each technology component of Complete Genomics' solution will be discussed in detail in this paper. Complete Genomics' DNA libraries, sequencing platform, high-speed instruments, and suite of base-calling, mapping, assembly, and analysis software enable low reagent use and delivery of high-quality human genome data sets.

Libraries

Complete Genomics' DNA libraries, in conjunction with its proprietary Combinatorial Probe-Anchor Ligation (cPAL™) technology, are used for human genome sequencing. Currently, 35-base mate pair reads are generated from approximately 500-base pair genomic fragments. This fragment size is sufficient to span very common repetitive elements, in particular Alu repeats, which comprise 10% of the genome.

Library Construction Process

Complete Genomics' DNA libraries consist of genomic DNA fragments with known synthetic DNA sequences (called adaptors) interspersed within the genomic DNA at regular intervals. The adaptors act as starting points for reading up to 10 bases from each adaptor-genomic DNA junction.

Complete Genomics uses a proprietary library construction process to insert four adaptors into each DNA fragment (Figure 2). A four-adaptor approach supports 70-base reads (35 bases per mate pair). The read length may be increased by inserting more adaptors.

Arrays

Complete Genomics has developed ultra-high density DNA arrays that can be read with standard fluorescent chemistry and imagers constructed from commercial components, thus minimizing the cost of both reagents and imaging. Unlike alternative approaches, clonal DNA amplification is not performed in emulsions or on

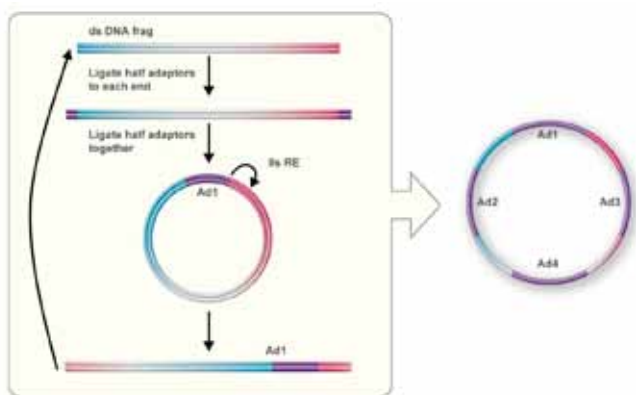


Figure 2. Multiple Adaptor Library Construction Process.

surfaces. The amplification process occurs in solution and in a single reaction chamber, resulting in higher density and lower reagent usage. Additionally, because the DNA nanoball (DNB™) production process inherently produces clonal amplicons, it is not subject to the stochastic variation from limiting dilution inherent in alternative approaches.

Clonal DNA Amplification in Solution – DNA Nanoballs (DNBs)

Sequencing is performed on amplified DNA clusters termed DNA nanoballs (DNBs). The amplification avoids the cost and challenges of relying on single fluorophore measurements used by single-molecule sequencing systems.

Starting with a small circular DNA template (Figure 3) consisting of approximately 80 bases of genomic DNA and four synthetic adaptors, Complete Genomics generates a head-to-tail concatemer consisting of more than 200 copies of the circular template. Complete Genomics has developed a variety of proprietary techniques for forming this concatemer into a ball as well as controlling its size, density and binding affinity to

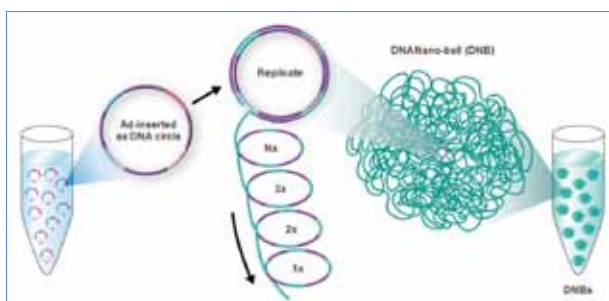


Figure 3. DNA Nanoball Formation.

surfaces and to other DNBs. One milliliter (mL) of reaction volume generates over 10 billion DNBs, sufficient for sequencing an entire human genome.

Patterned Substrates

Complete Genomics produces patterned substrates (Figure 4) with two-dimensional arrays of spots that are activated to capture and hold DNBs. The patterned surfaces are produced using standard silicon processing techniques.

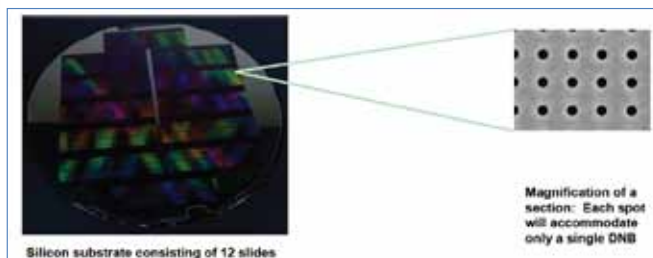


Figure 4. Patterned Substrate.

Self-assembling DNB Arrays

Complete Genomics makes a DNB array by introducing the DNBs to the patterned surface (Figure 5). The DNBs stick to the activated, or “sticky,” spots, and do not stick to the fields between the spots. When a single DNB sticks to a spot, it repels other DNBs, resulting in at most one DNB per spot. DNBs are three-dimensional, resulting in more DNA copies per square nanometer of binding surface than traditional DNA arrays. This unique three-dimensional quality further reduces the quantity of sequencing reagents required, resulting in brighter spots and more efficient imaging. In practice, DNB array occupancies exceed 90%. A high-density DNB array thus “self-assembles” from DNBs in solution, eliminating one of the most costly aspects of producing traditional patterned oligo or DNA arrays.

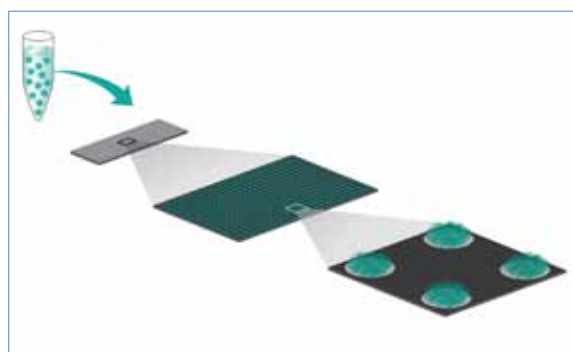


Figure 5. Slide Preparation.

Assays

The historical drawback of sequencing by ligation has been short read length, which is typically limited to approximately six bases from the ligation site. Complete Genomics increased the read length to 10 bases and, by inserting multiple adaptors into each genomic fragment, each of which has two ligation sites, multiple adjacent 10-base segments of genomic DNA can be read.

Combinatorial Probe-Anchor Ligation (cPAL™): Ligation-based DNA Sequencing

Complete Genomics' approach combines hybridization and ligation to produce high-accuracy reads with minimal reagent usage. Complete Genomics' sequencing assay, called combinatorial Probe-Anchor Ligation (cPAL™), has many of the advantages of sequencing by hybridization (SBH) including DNA array parallelism, independent and non-iterative base reading, and the capacity to read multiple bases per reaction. In addition, cPAL resolves two SBH limitations—the inability to read simple repeats and the need for intensive computation.

cPAL uses pools of probes labeled with four distinct dyes (one per base) to read the positions adjacent to each adaptor (Figure 6). Each read position has a separate pool of probes. Complete Genomics' proprietary approach allows 10 contiguous bases to be read from

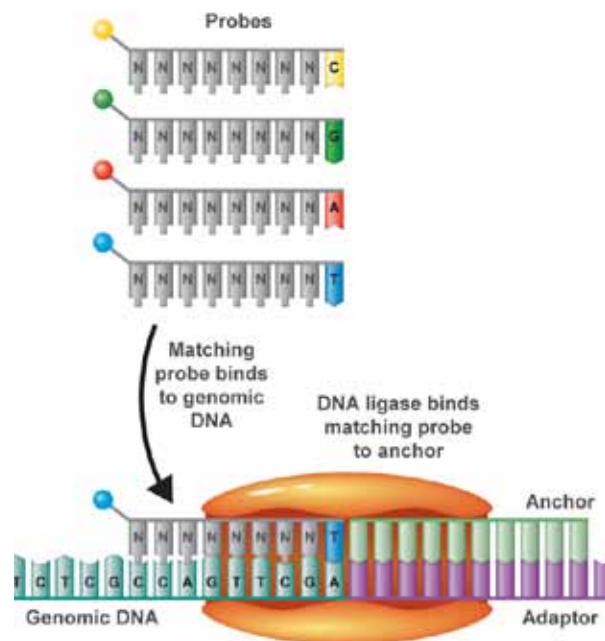


Figure 6. Combinatorial Probe-Anchor Ligation (cPAL™).

each end of an adaptor. Ligating the matching probes with the adjacent anchors dramatically improves the full-match specificity of the probe binding as compared to hybridization without ligation. Under optimal conditions, the raw error rate of this assay can be below 0.1%.

After each base is read, the entire anchor-probe complex is washed away. The next anchor is then hybridized, and the next probe is ligated to the anchor. There is no chaining of consecutive probes and, thus, no accumulation of errors.

One of the unique advantages of cPAL is random access (independent and non-iterative base reading). Each base-read cycle does not depend on the completeness of any of the previous cycles. This process provides excellent fault tolerance qualities—if a base read fails, it does not prevent interpretation of the remaining reads for that DNB; and, if desired, the failed base can simply be re-assayed.

Another key advantage of independent base reading is the tolerance to low ligation yield per cycle. This tolerance dramatically reduces the required probe and enzyme concentrations, thereby substantially reducing reagent costs. cPAL further allows the ability to read multiple positions per cycle, which is not possible with sequencing by synthesis. Reading multiple positions per cycle decreases the number of cycles, thus reducing reagent consumption and imaging time.

Instruments

Complete Genomics' high-speed instrument design is highly modular and allows rapid reading of sub-micron DNA nanoarrays. Each of its components may be independently upgraded as suppliers release new, improved versions. By relying on standardized components, the company is able to track its suppliers' technology roadmaps to deliver state-of-the-art performance, while leveraging continuous cost. High-volume purchases likewise enable the company to work with suppliers to improve performance of the Complete Genomics' sequencing instrument.

Complete Genomics' sequencing instrument (Figure 7) consists of three loosely coupled standardized sub-systems:

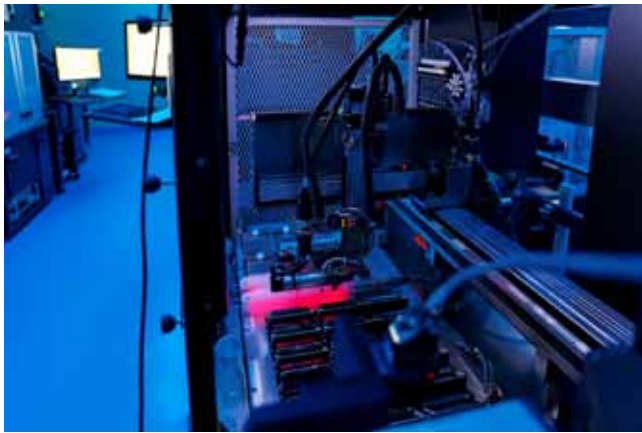


Figure 7. Complete Genomics Sequencing Instrument.

- DNA nanoarrays, packaged into flow slides
- Standard liquid-handling robot
- High-speed imager

This modular design enables Complete Genomics to adjust components easily as specifications or performance criteria change and to rapidly reconfigure hardware as technologies advance.

Flow Slides

Complete Genomics has developed a powerful flow-slide platform to minimize reagent use and simplify fluorescence imaging (Figure 8). Micro-channels formed on top of the patterned substrates enable efficient reagent delivery and eliminate dead volume while simultaneously satisfying the optical requirements for high-resolution imaging. Process capacity (DNA spots measured per cycle) may be increased by adding more flow slides to the liquid handling deck. This addition ensures that increases in imager speed are matched by



Figure 8. Flow Slides.

increases in process capacity.

Fluidics Robot

Complete Genomics uses standard, off-the-shelf, liquid handling robots to pipette reagents to the flow slides. When reactions are complete and a flow slide is ready to be imaged, a robotic arm transfers the slide from the liquid handling deck to the imager stage. Each instrument can run 2 to 16 slides in parallel—while one slide is imaging, the remaining slides are in various stages of preparation for imaging.

Imager

The imager is constructed from off-the-shelf components to form a four-color fluorescence microscope. The main components are the illuminator, filter changer, microscope objective, tube lens, motorized stage, and detector.

Data Analysis and Software

Complete Genomics has developed its own suite of base-calling, mapping, assembly, and analysis software to rapidly reconstruct genomes from billions of mate pair reads. Because Complete Genomics sequences large numbers of human genomes, its innovative software is optimized for assembling and analyzing human genomes. This software is fast and accurate, aligning reads and assembling human genomes in less than a day.

The base-calling software receives data from the imager after each reaction cycle. Images are processed to determine the bases at each position on a DNB array. The called bases for each DNB are collated to form raw read data. Mapping, assembly, and analysis software operate on read data and produce a variety of outputs, including reads aligned to a reference genome and consensus sequence assembly of overlapping DNB reads. Customers are provided with high-quality data including reports on summary statistics; variants including SNPs, indels, copy number variants, structural variants and mobile element insertion events; and reads, scores and mappings.

Base-Calling Software

Four images, one for each color dye, are generated for each queried genomic position. The position of each

spot in an image and the resulting intensities for each of the four colors is determined by adjusting for crosstalk between dyes and background intensity. A quantitative model is fit to the resulting four-dimensional data set. A base is called for a given spot, with a quality score that reflects how well the four intensities fit the model.

Read Data Format

Read data include both a called base and a quality score. The quality score is correlated with base accuracy. Analysis software, including sequence assembly software, uses the score to determine the contribution of evidence from individual bases within a read.

Reads are “gapped” due to the DNB structure (Figure 9). Gap sizes vary (usually +/-1 base) due to the variability inherent in enzyme digestion. Due to the random-access nature of cPAL, reads can occasionally be an unread base (‘no-call’) in an otherwise high-quality DNB. Read pairs are mated as described in the DNA libraries section.

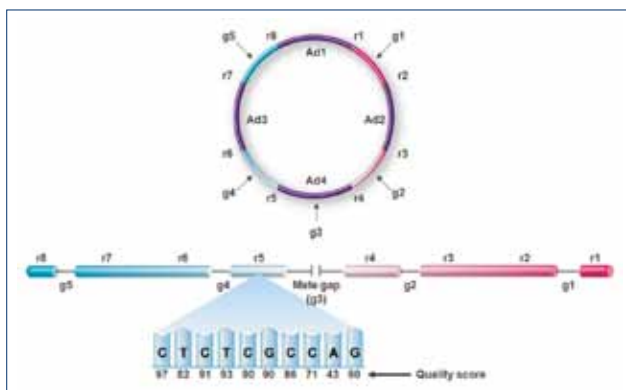


Figure 9. Read Data Format.

Mapping

Complete Genomics has developed high-speed mapping software capable of aligning read data to a reference sequence. The software runs on a standard Linux cluster and scales horizontally with more CPUs. The mapping is tolerant of small variations from a reference sequence, such as those caused by individual genomic variation, read errors, or unread bases. To support assembly of larger variations, including large-scale structural changes or regions of dense variation, each 35-base arm of a DNB is mapped separately, with mate pairing constraints applied after alignment.

Assembly and Analysis

Complete Genomics has developed assembly and

analysis software (Analysis Pipeline) that scales horizontally with more CPUs using a standard Linux cluster.

The Analysis Pipeline supports the DNB read structure (mated, gapped reads with non-called bases). The Complete Genomics assembler currently calls SNPs and short insertions, deletions, and block substitutions up to approximately 50 base pairs resulting in highly accurate variant detection on both alleles (SNPs and small Indels) on over 90% of the genome. The algorithm uses a combination of evidential (Bayesian) reasoning and de Bruijn graph-based algorithms. Using a statistical model, that is empirically calibrated to each data set, allows all read data to be used without pre-filtering or data trimming. In addition, copy number variation and structural variation data is provided as well as mobile element insertions. Complete Genomics employs both read-depth and paired-end mapping to detect a broader range of structural variation events and sizes. Read-depth analysis is better able to identify CNVs in complex regions of the genome rich in segmental duplications

Complete Genomics annotates the variants called in each genome with entries from a variety of public databases. There are two classes of annotations that are provided with our variant calls: 1) those that put the results in a biological context (gene information, known disease associate, etc.) and 2) those that will help filter the list of variants, facilitating drilling down to the variants that explain disease (e.g., known variants in dbSNP, overlap with repetitive elements, etc.). Both are important to extracting biological knowledge from sequencing data.

Our biological annotations are derived from a number of public annotation databases including:

1. RefSeq alignments in NCBI's annotation builds for gene and functional annotation of variants (e.g., current build that we use is NCBI 37.2). This determines if a variant overlaps a particular gene and what functional impact the variant may have (e.g., missense mutation, gene fusion event, etc.).
2. miRBase, to identify whether variants overlap microRNAs.
3. COSMIC, to identify variants previously detected in cancer samples.
4. dbSNP is used to annotate variants that overlap

dbSNP entries, and these entries are now annotated if disease association is known.

5. DGV (Database of Genomic Variants) is used to annotate CNVs.

Data analysis on customer genomes is performed using Complete Genomics' secure computing infrastructure. Once a genome sequence has completed analysis and passes quality control (QC), the completed customer data set is transmitted on a secure network connection to Amazon Web Services for delivery to customers.

Data Quality

As reported in the Science article (Science 1 January 2010: Vol. 327. no. 5961, pp. 78 – 81), Complete Genomics' scientists generated high-quality diploid base calls in as much as 95 percent of the genomes sequenced, identifying 3.2 million to 4.5 million sequence variants per genome processed. Detailed validation of one genome data set demonstrated a sequence accuracy of just one false variant per 100 kilobases. Customer studies have also demonstrated an accuracy of one Mendelian Inheritance Error in 300 kilobases (Science 30 April 2010: Vol. 328. no. 5978, pp. 636 - 639).

In order to enable the research community to validate Complete Genomics' sequencing performance and to further improve data analysis and interpretation methods, whole human genome sequence data sets are available on an FTP server (<ftp2.completegenomics.com>) for free download and general use. These data result from the sequencing of 69 standard, non-diseased samples as well as two matched tumor and normal sample pairs.

Summary

Complete Genomics provides whole human genome

sequencing and analysis as a service to academic and biopharmaceutical researchers at an unprecedented quality, cost, and scale without requiring investment in in-house sequencing instruments, high-performance computing resources, or specialized personnel.

Complete Genomics offers two primary services: the Standard Sequencing Service and the Cancer Sequencing Service. The Complete Genomics Standard Sequencing Service provides whole human genome sequencing, with sophisticated assembly, variant detection, and annotation across all variation types including SNPs, indels, copy number variations (CNVs), structural variations (SVs), and mobile element insertions (MEIs). The Complete Genomics Cancer Sequencing Service provides whole human genome sequencing of cancer pairs or cancer trios with assembly, variant detection, and annotation as noted above, and each tumor is compared to both the normal match submitted in the pair or trio and to the human reference genome. Customers utilizing either service receive a data package of fully mapped reads, summary files of all variations found (including somatic variants for cancer samples), and evidence files that support the variations.

As the world's first company dedicated to large-scale human genome sequencing and analysis, Complete Genomics is uniquely positioned to enable scientists to conduct human disease research on thousands of genomic samples, thus accelerating genomic-based discovery.

www.completegenomics.com info@completegenomics.com
2071 Stierlin Court, Mountain View, CA 94043 USA Tel 650.943.2800



Copyright© 2011 Complete Genomics, Inc. All rights reserved. Complete Genomics and the Complete Genomics logo are trademarks of Complete Genomics, Inc. All other brands and product names are trademarks or registered trademarks of their respective holders.

Complete Genomics data is for Research Use Only and not for use in the treatment or diagnosis of any human subject.
support@completegenomics.com Toll-free: 1-855-CMPLETE (1-855-267-5383) or 1-650-943-2600
Information, descriptions and specifications in this publication are subject to change without notice.

Published in U.S.A., December 2011, WPTO-01