The goal of Next Generation Sequencing (NGS) is to create large, biologically meaningful contiguous regions of the DNA sequence—the building blocks of the genome—from billions of short fragment data pieces. Whole genome “shotgun sequencing” is the best approach based on costs per run, compute resources, and clinical significance. Shotgun sequencing is the random sampling of read sequences from NGS instruments with optimal coverage. NGS coverage is defined as: Number of reads x (Read Length/Length of Genome). The number of reads is usually in the millions with the read length and length of genome quoted in base pairs. The length of the human genome is about 3 billion base pairs.

The steps in shotgun sequencing are:
1. Extract and fragment DNA.
2. Clone DNA and sequence both ends of clone.
4. Assemble sequence by creating a De Bruijn graph (which can have 100 million-plus nodes with segments called k-mers) and its sub-graphs; detect overlaps and join; reduce graph and create scaffolds.
5. Finish sequence by closing gaps—higher coverage means fewer and smaller gaps.
6. Find phenotypic and clinical significance via Single-Nucleotide Polymorphisms (SNPs), inserts/deletions (indels), variants, Copy Number Variations (CNVs), and mutations.

The architectural components of NGS are the instrument(s), High Performance Computing (HPC), storage, and the network. The goal of HPC in NGS is to reduce latency and maximize DNA sequence data processed in unit time. The slowest components in an HPC environment are the network and disk. Several reasons for not achieving total parallelization include:

- **Algorithmic limitations**: These occur due to mutual dependencies or portions of the process that can only be performed sequentially.

- **Bottlenecks**: Data access is the largest bottleneck in HPC workflows. Because loop-based algorithms move large amounts of data in and out of the CPU, on-chip resources tend to be underutilized and performance is limited by the slowest data paths to memory and its storage. Access to a shared resource—execution paths in core, shared memory paths in multi-core chips, and I/O devices—serializes the execution. This affects concurrency as well.

- **Startup overhead**: Base calling and other large numbers of small file writes also incur this overhead before caching.

- **Communication**: Full concurrency between parts of a parallel system is more theoretical than practical. Since communication is the cornerstone of parallel algorithms, there is always some serialization involved. Other bottlenecks, however small, are based on driver and version “jitters” within the OS and network.
The faster-than-Moore’s-Law advances in genomics science in the past five years have led to the acceleration of Molecular Diagnostics (MDx), which Dr. Leroy Hood describes as “P4 Medicine”: personalized, predictive, preventive, and participatory. The use-cases to enable clinical genomics using NGS are shown in Table 1 immediately below.

<table>
<thead>
<tr>
<th>Name</th>
<th>Nucleic acid population</th>
<th>Analysis strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq</td>
<td>RNA (may be poly-A, mRNA, or total RNA)</td>
<td>Alignment of reads to “genes”; variations for detecting splice junctions and quantifying abundance</td>
</tr>
<tr>
<td>Small RNA sequencing</td>
<td>Small RNA (often miRNA)</td>
<td>Alignment of reads to small RNA references (e.g. miRNA), then to the genome; quantify abundance</td>
</tr>
<tr>
<td>ChIP-Seq</td>
<td>DNA bound to protein, captured via antibody (ChIP=Chromatin ImmunoPrecipitation)</td>
<td>Align reads to reference genome, identify peaks and motifs</td>
</tr>
<tr>
<td>Structural Variation Analysis</td>
<td>Genomic DNA, with two reads (mate-pair reads) per DNA template</td>
<td>Align mate-pairs to reference sequence and interpret structural variants</td>
</tr>
<tr>
<td>de novo Sequencing</td>
<td>Genomic DNA, possibly with external data (e.g. cDNA, genomes of closely related species, etc.)</td>
<td>Piece together reads to assemble contigs, scaffolds, and (ideally) whole-genome sequence</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>Entire RNA or DNA from a (usually microbial) community</td>
<td>Phylogenetic analysis of sequences</td>
</tr>
</tbody>
</table>

**Table 1: NGS use-cases and analysis strategy**

The analysis strategy is driven mostly by open source software as shown in Table 2 below. However, as the clinical genomics field matures, supported platforms like Avadis and CLCbio are gaining traction due to documentation, audit, and regulatory requirements. Techniques like smallRNA, ChIPseq, and RNAseq—along with de novo sequencing—are showing clinical promise.

<table>
<thead>
<tr>
<th>Steps in the bioinformatics process</th>
<th>Integrated Toolkits</th>
<th>Align/Assemble to Reference Genome</th>
<th>de novo Alignment and Assembly</th>
<th>SNP/indel Discovery</th>
<th>Annotation, Browser</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• CASAVA</td>
<td>• BoWTie</td>
<td>• AbysS</td>
<td>• SOAPsnp</td>
<td>• Avadis</td>
</tr>
<tr>
<td></td>
<td>• CLCbio Genomics Workbench</td>
<td>• BWA</td>
<td>• SOAPdenovo</td>
<td></td>
<td>• EagleView</td>
</tr>
<tr>
<td></td>
<td>• Galaxy</td>
<td>• ELAND</td>
<td></td>
<td></td>
<td>• Harvard Genotator</td>
</tr>
<tr>
<td></td>
<td>• GA4</td>
<td>• Maq</td>
<td></td>
<td></td>
<td>• NCBI MapView</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• SAMtools</td>
<td></td>
<td></td>
<td>• SAMtools</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• BFAST</td>
<td></td>
<td></td>
<td>• UCSC Genome Browser</td>
</tr>
</tbody>
</table>

**Table 2: Commonly used analysis tools in NGS**

**ILLUMINA HiSeq**

The NGS architecture shown in Figure 2 alongside is for a single production sequencer example—the Illumina HiSeq. With “sequencing as a service” gaining popularity, the researcher focuses on the Binary Alignment Map (BAM) file to begin the sequence analysis from a functional perspective: SNPs, inserts/deletions (indels), variants, CNVs, and mutations.

**ILLUMINA HISEQ 2000 DATA THROUGHPUT**

Throughput approx. 30 to 50 Gbp per day

- Read length of about 100bp, approx. 50x to 75x coverage
- ~ 8 TB raw data per run (excluding images, log files, including base call files)
- ~ 100 GB results data per run
- ~ 2 runs per week per instrument
- ~ 8 TB per week per instrument
- 4 whole genome sequences (WGS) or 16 exomes per week per instrument
- Realistic throughput of 160 whole genomes or 480 exomes per year per instrument. This is about 350 TB per year.

**Figure 2: Reference architecture**
EMC ISILON ONEFS

EMC® Isilon® OneFS combines the three layers of traditional storage architectures—the file system, volume manager, and RAID—into one unified software layer, creating a single intelligent distributed file system that runs on one storage cluster. The advantages of OneFS for NGS are many:

- **Scalable**: Scale out as needs grow. Linear scale with increasing capacity—from 18 TB to 16 PB in a single filesystem and a single global namespace.
- **Predictable**: Dynamic content balancing is performed as nodes are added, upgraded, or capacity changes. No added management time is required because this process is simple.
- **Available**: OneFS is “self-healing.” It protects your data from power loss, node or disk failures, and loss of quorum and storage rebuild by distributing data, metadata, and parity across all nodes.
- **Efficient**: Compared to the average 50 percent efficiency of traditional RAID systems, OneFS provides over 80 percent efficiency, independent of CPU compute or cache. This efficiency is achieved by tiering the process into three types as shown in the figure alongside and by the pools within these node types.
- **Enterprise Ready**: Administration of the storage clusters is via an intuitive Web-based UI. Connectivity to your process is through standard file protocols: CIFS, SMB, NFS, FTP/HTTP, iSCSI, and HDFS. Standardized authentication and access control is available at scale: AD, LDAP, and NIS.

Note that either the X- or S-Series and the NL-series should be procured together to build a balanced system.

**Assumptions:**

- a. Raw files are handled separately
- b. Process starts at base call files
- c. Common archival layer

**Output files:**

- BAM files: 30 bytes per read + 2 bytes per base pair; approximately 100 GB to 250 GB for human WGS
- SRA files: 10 Bytes per base pair (~ 30 GB for human genome WGS)

Analysis and Interpretation times not included—these vary from a week to several months.

### Table 3: Two storage reference architectures

<table>
<thead>
<tr>
<th>One-to-One Architecture</th>
<th>Many-to-One architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Sequencer to One Storage Cluster ≤ 3 sequencers</td>
<td>Many Sequencers to One Storage Cluster ≥ 3 sequencers</td>
</tr>
</tbody>
</table>

**Example with Single Sequencer**

- EMC Isilon X-Series 18 TB raw (3x 6 TB) with minimum 400 GB SSD, 48 GB RAM
- Move to archival layer after every run using SmartPools or SnapShotIQ
- Scratch area (tmp folders) on Storage Cluster
- IB backend and minimum 1 Gbps front-end Ethernet (2x 1 Gbps LACP or 10 Gbps for improved performance)

**Scale this as sequencers are added**

Archival Layer:

- NL-Series 108 TB raw (3x 36 TB): Design with 324 TB raw data (3x 108 TB) if planning for one year of storage.

**Example with 3 Sequencers**

- EMC Isilon S-Series 36 TB raw (3x 12 TB) with minimum 800 GB SSD, 144 GB RAM
- Scratch area (tmp folders) on HPC Cluster if NGS Analysis software allows this configuration; otherwise Scratch area on the S-Series Cluster
- IB backend and 10 Gbps front-end Ethernet

**Scale this as sequencers are added**

Archival Layer:

- Batched movement using SmartPool rules or SnapShotIQ NL-Series 324 TB raw (3x 108 TB): Design with 2x NL clusters if planning for one year of storage.
DATA FLOW AND PERFORMANCE

NGS DATA FLOW AND SIZING

Despite the similarity of NGS components—i.e., sequencer(s), HPC, and storage—the workflows can be quite different as shown in Table 1. It is imperative that the research and the IT team understand and plan the HPC and storage architecture accordingly. Figure 4 alongside illustrates one such workflow with its data paths, file, and data sizes. This is representative of most WGS workflows. It is also critical to understand the instrument throughput.

PERFORMANCE METRICS

As with understanding the workflow, the degree of parallelization of the algorithms in the process is critical for performance. The other main factors are RAM (3 GB/core), NFS tuning, and jumbo frames (TCP MTU 9000). Thread tuning for SGE and NFS could also help.

Test setup:

- 72-core (2.6 GHz) HPC platform with 144 GB RAM on CentOS 6.2
- Son-of Grid Engine and PDSH, parallel make with GCC
- EMC Isilon S-Series node with 12 TB raw, 540 GB SSD and 144 GB RAM with IB backend on OneFS 6.5.4
- 1 Gbps front-end Ethernet, NFSv4
- Illumina CASAVA 1.8 analysis platform
- Human whole genome sequence from Illumina

CONCLUSION

Next Generation Sequencing is a complex and multivariate endeavor—with the instrument end changing and evolving faster than the storage or HPC layers—as the process moves into clinical genomics. EMC Isilon makes the storage design, implementation, and upgrade process for NGS painless and simple.

CONTACT US

To learn more about how EMC Isilon products, services, and solutions help solve your business and IT challenges, contact your local representative or authorized reseller—or visit us at www.EMC.com/Isilon.

References: