WHITE PAPER

Analyzing Whole Human Genomes for as Little as $22

QIAGEN Bioinformatics and Intel Reference Architecture for High-Volume Genome Analysis

Introduction

Advances in next-generation sequencing (NGS) technologies are challenging the bioinformatics industry to develop more powerful tools for accurate detection and interpretation of variants, while making it faster and more cost-effective. By sequencing whole human genomes at rates of up to 18,000 per year—an average of one genome every 32 minutes—new sequencers have broken the $1,000 genome barrier[1] and opened the door to population studies and clinical usage models that have not been possible before.

This high-volume genome sequencing generates an enormous amount of data that must be analyzed as fast as it is produced. Assuming 700 GB of data for a whole human genome at 30x coverage, keeping pace with a high-end NGS system could require analyzing more than 30 terabytes of data per day. According to Illumina, which manufactures the industry's highest throughput NGS system to date, it would take an 85-node high performance computing (HPC) cluster to meet these heavy processing requirements.[2]

QIAGEN Bioinformatics and Intel offer a more cost-effective alternative. Working together, the two companies developed a cluster reference architecture that could meet these same throughput requirements with only 32 compute nodes, while dramatically reducing space, energy, infrastructure, and management costs.

Based on the performance tests and total cost of ownership (TCO) study described in this paper, the reference architecture could potentially reduce annual costs by as much $1.3 million versus the 85-node cluster defined by Illumina, while delivering the same level of throughput.[3] Depending on sequencing volumes and data center efficiency, this solution could enable full analysis of whole human genomes for as little as $22 each.

Table of Contents

> “NGS technologies are transforming the field of bioinformatics by generating raw sequencing data at the speeds and volumes needed to provide cost-effective support for population studies and clinical usage models.”

An Optimized Hardware and Software Stack

The QIAGEN Bioinformatics and Intel reference architecture is specifically optimized to deliver high throughput for genomic workloads, while containing infrastructure and operating costs. It is also designed to scale on-demand at each tier—compute, networking, and storage—so organizations can grow their analytic capability in a straightforward and cost-effective manner. The reference architecture is described below.

Intel® Scalable System Framework

The compute cluster for the QIAGEN Bioinformatics and Intel reference architecture includes 32 two-socket servers based on the Intel® Xeon® processor E5 v3 family. These processors are ideal for today’s highly parallel and data-intensive genomic workloads. They provide up to 50 percent more cores, cache, and system bandwidth than the prior generation of Intel Xeon processors. They also support Intel® Advanced Vector Extensions 2.0 (Intel® AVX2). By enabling a single instruction to operate simultaneously on up to 256 data points, Intel AVX2 can significantly improve performance for some of the most data- and compute-intensive stages of genome analysis.

QIAGEN Biomedical Genomics Server Solution

The performance results documented by Illumina for its 85-node cluster were based on a variant calling workflow of raw data using open source BWA-GATK software.[2] The 32-node QIAGEN Bioinformatics and Intel reference architecture uses Biomedical Genomics Server solution, which consists of the CLC Genomics Server software with the Biomedical Genomics Server Extension. Biomedical Genomics Server solution provides all the advanced tools and capabilities of CLC Genomics Workbench and CLC Biomedical Genomics Workbench, but is designed specifically for HPC clusters. Geneticists benefit from a powerful workbench that provides sophisticated analytic and pipeline capabilities with intuitive interfaces for increased productivity. A number of additional advantages make Biomedical Genomics Server solution particularly appropriate for high-volume genome analysis.

- **Optimized Performance.** QIAGEN Bioinformatics developers work closely with Intel engineers to optimize CLC Genomics Workbench, Biomedical Genomics Workbench, and CLC Genomics Server for every new Intel Xeon processor generation. To maximize the value of these efforts, QIAGEN Bioinformatics uses Intel® Parallel Studio as a key component in its software development environment. Intel updates its software development tools regularly to take advantage of the latest hardware innovations, which helps QIAGEN Bioinformatics achieve higher performance gains with less effort.
- **High Quality Results.** QIAGEN Bioinformatics invests considerable resources to optimize the quality of the results provided by Biomedical Genomics Server solution. By providing higher sensitivity and a lower percentage of false positives than many alternative solutions,[4] Biomedical Genomics Server solution can help clinical researchers to focus in more quickly on promising new candidates for follow-up (see the sidebar on page 3, Proven Accuracy–Results You Can Trust).
- **Simplified Cluster Management.** QIAGEN Biomedical Genomics Server is designed to insulate users from the complexities of cluster computing. It provides geneticists with the speed and throughput of a powerful HPC cluster, while reducing their dependence on HPC specialists when planning, scheduling, and monitoring their analyses.

A Fast, Scalable Cluster Fabric

The QIAGEN Bioinformatics and Intel reference architecture includes a high-speed interconnect fabric—based on Intel® True Scale Fabric—that connects the compute nodes to each other and to a centralized storage system. Intel True Scale Fabric is designed specifically for HPC. It provides up to 40 Gbps of bandwidth per port, and uses open source communications protocols that are optimized for fast MPI message rates and efficient storage communications.

Intel True Scale Fabric also provides a simple pathway to next-generation Intel® Omni-Path Architecture (Intel® OPA), which will deliver additional improvements in fabric performance, scalability, and cost models. Intel OPA will provide 100 Gbps port bandwidth and advanced features for traffic optimization and resilience. It will also include a new family of 24- and 48-port edge switches. The high bandwidth and port densities of these switches will make it particularly easy and cost-effective to scale fabric performance as organizations grow their clusters.

Massively Scalable, Centralized Storage

Keeping all the nodes, cores, and threads of a large compute cluster operating at high efficiency requires fast access to massive amounts of raw data. One way to deliver fast data access is by using local storage drives in each server of an HPC cluster. However, local storage complicates data management, an issue that can be particularly problematic when many terabytes of data must be ingested, stored, analyzed, and archived daily.

To address this issue, the QIAGEN Bioinformatics and Intel reference architecture uses Intel® Enterprise Edition for Lustre* to support a centralized, 165 terabyte storage system. Lustre is the leading parallel storage system in the world,[5] largely because it scales readily to deliver extreme capacity and performance (up to hundreds of petabytes of storage and more than two terabytes per second of aggregate I/O throughput[6]). Intel packages this open source software with powerful management tools that simplify implementation and management.

To deliver high performance while containing costs, the storage reference architecture combines commodity hard disk drives (HDDs) with a small number of high-speed Intel® Solid State Drives (Intel® SSDs). The HDDs provide high capacity at low cost, while the Intel SSDs accelerate the operations that are most important for fast genome analysis.

Further improvements in TCO and overall performance can be achieved through the use of PCIe* NVMe based Intel SSDs. The market leading performance of the Intel® P3600 and P3700 Series DC SSDs provide reduced cost through a 6:1 consolidation, while increasing overall performance. Available in up to 4 TB capacity, one P Series drive will typically deliver similar or greater performance of at least 6 SATA SSDs.[7]

Designing and managing a distributed storage solution, such as Lustre, is more complex than deploying a traditional, scale-up storage system. Intel® Manager for Luster* simplifies installation, configuration, and management through integrated monitoring and an easy-to-use web-based dashboard. Integrated support for hierarchical storage management (HSM) provides transparent support for heterogeneous drives, so mixing Intel SSDs and HDDs is simple and straightforward.

HSM helps to simplify both initial implementations and future performance scaling. The performance benefits of Intel Data Center SSDs for NVMe are realized out-of-the-box, with little or no need for manual configuration. HSM also enables transparent integration with backup and archiving solutions, such as high-volume tape storage. Users and administrators can access both current and archived data using a common set of tools and interfaces. Intel also offers 24/7 support for Lustre, to help organizations maintain high reliability and uptime in production environments.

Proven Accuracy – Results You Can Trust

While efficiency and cost-effectiveness is an important factor for NGS data analysis, accuracy in both variant calling and interpretation is most crucial. This is made possible by QIAGEN Bioinformatics solutions, comprised of the Biomedical Genomics Workbench and Server and Ingenuity Variant Analysis. In an indirect comparison with the gold standard variants from Genome in a Bottle using the GCAT portal, sensitivity and specificity were comparable with 97.34 percent and 99.9995 percent respectively. In comparison to an open source pipeline with BWA and Samtools, CLC's variant calling pipeline had a more than 4x lower false positive rate with 0.0005 percent compared to 0.0023 percent.

The sensitivity of 97.34 percent for CLC's variant calling pipeline yields a false negative rate of 2.66 percent. Furthermore, internal benchmarking for causal variant identification, using WGS data from genetic odyssey cases, showed 30x enrichment in biologically relevant variants using Ingenuity Variant Analysis that leverages millions of biological findings manually curated from the peer-reviewed literature or integrated from third-party databases, from the Ingenuity Knowledge Base, to enable fast and accurate discovery of rare disease-causing variants within human genome data.

To learn more about the benchmarking study on accuracy using QIAGEN Biomedical Genomics Workbench and Ingenuity Variant Analysis visit us at **qiagenbioinformatics.com/accuracy**.

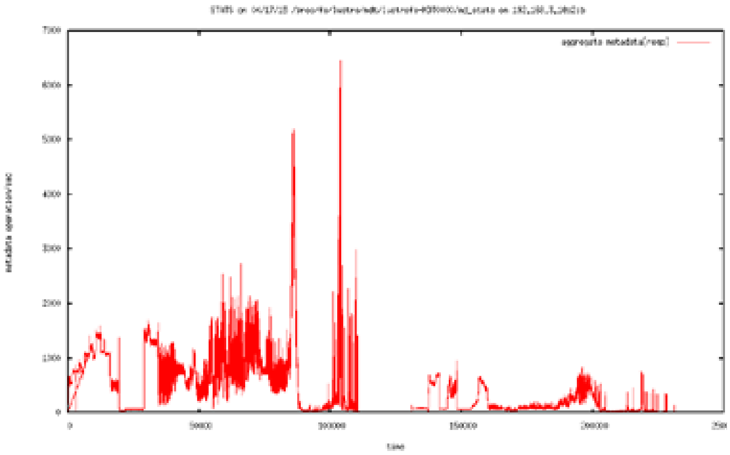


Figure 1. I/O read demands on the metadata server are heavy, particularly during earlier stages of analysis (recorded using Intel® Manager for Lustre*).

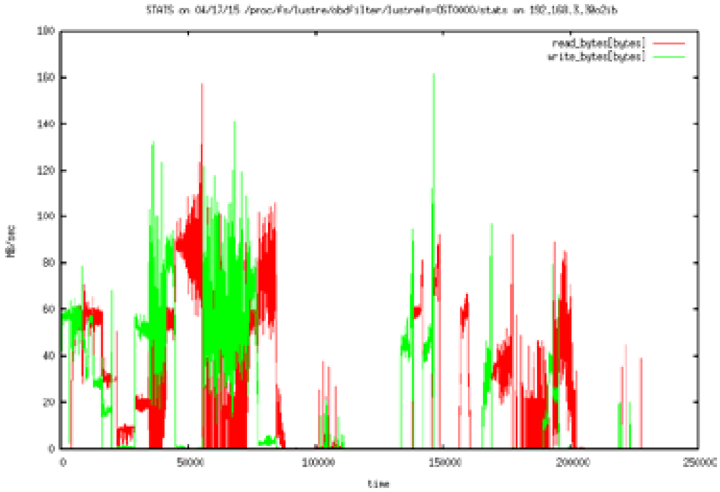


Figure 2. I/O read and write demands are heavy on the object storage servers (recorded using Intel® Manager for Lustre*).

Intel Resources for Lustre* Users

Intel offers comprehensive resources for organizations that are evaluating or deploying Lustre for high-performance, massively scalable storage.

- **Intel® Solutions for Lustre* Software**
  lustre.intel.com
  This Intel web site provides access to a range of Intel products, resources, and training opportunities that can help Lustre users achieve better results with less effort and reduced risk.
- **Architecting a High Performance Storage System**
  This Intel reference architecture describes a step-by-step approach to designing a Lustre storage system.

Architecting the Lustre Storage System

A Lustre storage system includes one or more metadata servers (MDSs), which manage file names, directories, access permissions, and file layout, and one or more object storage servers (OSSs), which store and serve up the data files. Each metadata server can have multiple metadata targets (MDTs) and each object storage server can have multiple object storage targets (OSTs). Since metadata servers are not involved directly in I/O operations, the metadata and object layers of a Lustre storage solution can be scaled independently.

Most HPC applications depend primarily on fast reads of large, sequential data files. Genome analysis workloads are somewhat different. Many stages of analysis require the execution of relatively small operations across large numbers of data files, which places heavy loads on metadata servers (Figure 1). These operations also generate temporary data that must be stored, so writes and reads on the object storage servers can be simultaneous and heavy (Figure 2).

To address these extreme workload requirements, the storage solution is architected as follows (Table 1 on page 5).

- **One metadata server with four Intel SSDs** in a RAID 10 configuration. The Intel SSDs accelerate performance for the metadata-intensive stages of genome analysis. The high performance of the drives enables a single metadata server to handle the heavy demands of the workload, which simplifies implementation, management, and scaling. (A second, hot-standby metadata server will typically be required in production environments.)
- **Four object storage servers, each with four object storage targets.** Each object storage target controls four 4 TB SATA drives in a RAID 5 configuration. The data capacity of the system is equal to the sum of the capacities of the object storage drives, which is 256 TB for this design. The use of RAID 5 reduces the usable data capacity to 165 TB, but provides parity data protection, which is needed to reliably recover high volume data in the event of a disk or server failure. For organizations deploying larger Lustre configurations, the reference architecture design team suggests implementing a RAID 6 layout using 10 near-line (NL) SAS HDDs.

Scaling Storage Performance

One way of scaling I/O performance in a Lustre storage system is by "striping" individual data files across multiple drives. When the striped file is accessed, the data is read simultaneously across the multiple drives. By including four drives per object storage target, the reference design supports up to 4x striping per file, which can increase I/O throughput by a factor of four versus a single drive. Striping can be configured for each file when it is created.

QIAGEN Bioinformatics and Intel Reference Architecture for High-Volume Genome Analysis

Intel® Scalable System Framework – 32 compute nodes + 3 management nodes	
Compute Node (each)	2 x Intel® Xeon® E5-2697 v3 processor (14 core, 2.60 GHz), 128 GB DDR4 memory (2133 MHz), 500 GB HDD 10K RPMH
Software	CLC Genomics Server, version 7.1 (configured to submit jobs using DRMAA to the SLURM workload manager)
Interconnect Fabric	
Host Connect Adaptors (HCAs)	2 x Intel® True Scale Single Port HCAs (2x QLE7340, QDR-80) per node
Switch	Intel® True Scale 12300 QDR InfiniBand Switch (36 port, 40 Gbps)
Storage System	165 TB useable disk capacity (256 TB total)
Software	Intel® Enterprise Edition for Lustre* 2.3
1 Metadata Server (MDS)	2 x Intel® Xeon® E5-2697 v3 processor (14 core, 2.60 GHz), 128 GB DDR4 memory (2133 MHz), RAID Controller (RAID 10), 4 x 800 GB Intel® Solid State Drive Data Center P3700 Series
4 Object Storage Servers (OSSs)	2 x Intel® Xeon® E5-2697 v3 processor (14 core, 2.60 GHz), 128 GB DDR4 memory (2133 MHz)
4 Object Storage Targets (OSTs) per OSS	4 x 4 TB SATA HDDs, RAID Controller (RAID 5)

Table 1.

There are many additional ways to scale a Lustre storage system. In this implementation, QIAGEN Bioinformatics and Intel engineers found that 4x striping using high-capacity HDDs was a cost-effective approach. As demands on the system grow, capacity can be scaled by adding more object storage servers, without compromising performance for data access times.

If analytic workloads change over time and it becomes necessary to increase I/O performance, additional HDDs could be added to each object storage target to support increased striping. However, it's important to note that increased striping puts heavier pressure on the metadata server, which must manage and track file layout across the larger number of disks.

To increase I/O performance without increasing metadata workloads, one or more Intel SSDs could be added to each object storage target. Depending on requirements, either cost-optimized SATA SSDs or high-performance Intel DC SSDs for NVMe could be used.

The Intel DC SSDs for NVMe are PCIe based. By implementing the NVMe protocol, these Intel SSDs only consume half the number of CPU cycles per I/O operation compared with SATA or SAS SSDs.

Configuring Lustre* and Intel® True Scale Fabric

Lustre provides an enormous number of metrics that can be monitored during runtime and used to tune the system for specific workloads. Appropriate tuning is important, particularly for applications such as genome analysis that are different from typical HPC workloads. Intel Manager for Lustre helps to simplify performance monitoring and tuning, by providing performance metrics through a user friendly interface. Intel also offers training that can help customers get up to speed more quickly. (For details on Intel training opportunities for Lustre, visit lustre.intel.com.

Some configuration is also required for the interconnect fabric, such as configuring it to support remote direct memory access (RDMA) for fast streaming of large files. The following settings provide a good starting point for configuring Lustre and Intel True Scale Fabric within the QIAGEN Bioinformatics and Intel reference architecture. For high-volume production environments, it is recommended that administrators perform additional fine-tuning to achieve fully optimized results for their specific workloads.

- **Lustre Configuration (starting points):**
  ```
  lctl set_param osc.*.max_rpcs_in_flight=256
  lctl set_param osc.*.max_dirty_mb=1024
  ```
- **Intel True Scale Configuration (starting points):**
  ```
  options ko2iblnd peer_credits=128 map_on_demand=32
  ```

Verifying Performance for High-Volume NGS Environments

To verify the performance of the reference architecture, QIAGEN Bioinformatics and Intel performed a series of tests with CLC Genomics Server running on a 16-node cluster and the 165 TB Lustre storage system described above. The tests were performed using publicly available whole human genome sequences at 30x coverage.[9]

Since the 16-node test cluster was half the size of the proposed cluster reference architecture, success was determined by the cluster's ability to analyze 24 genomes within 24 hours. This level of throughput would confirm that 48 genomes per day (an average of one every 30 minutes) could be analyzed using two 16-node clusters or a single 32-node compute cluster. The analysis was performed using a standard WGS variant calling workflow starting from the raw reads, and including full QC.

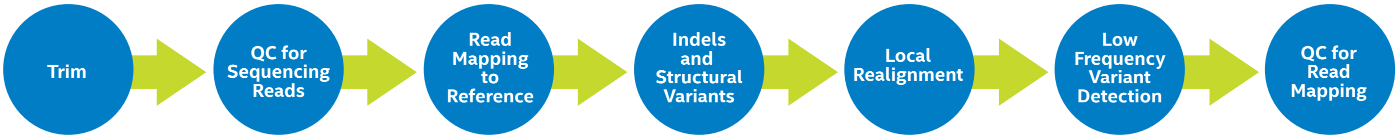

Trim
QC for Sequencing Reads
Read Mapping to Reference
Indels and Structural Variants
Local Realignment
Low Frequency Variant Detection
QC for Read Mapping

Figure 3: Standard Qiagen WGS Variant Calling Workflow.

Measuring a Performance Baseline

To begin testing, a single workflow was run on the cluster to provide a baseline for performance and to determine which stages of the analysis consumed most time and resources on the cluster (see Table 2). Analysis for the single workflow was completed in 59,600 seconds, or approximately 16.5 hours.

Single Node Performance Baseline – 1 WGS

Pipeline Stage	Runtime (seconds)
Import Raw Reads	2,800
Trim Sequences	1,977
Read Mapping	14,729
Insertion and Deletion Identification	4,123
Local Realignment	13,183
QC for Read Mapping	2,284
Variant Detection	22,586
Removal of Potential False Positives	184
Creation of Genome Browser View	10
Total Workflow Elapsed Time	59,600

Table 2.

Scaling to 24 Simultaneous Workflows

A series of tests was then performed to determine scalability as additional workloads were run simultaneously on the cluster. Initial tests showed acceptable scalability for up to 10 simultaneous workloads, although completion times rose as more workloads were added. When 16 workloads were run simultaneously, completion times exceeded the 24 hour limit.

Based on monitoring data obtained using Intel Lustre Manager, the increasing runtimes appeared to be the result of storage I/O contention during the early stages of analysis, when high volume data access is most pronounced. A number of strategies were combined to address that bottleneck.

- **4x Striping.**
  Striping was set at 4x to take advantage of all four drives within each object storage target.
- **Staggered workflows.**
  Adding a 30-minute delay before starting each successive workflow helps to ensure that the initial I/O-intensive stages of analysis do not occur concurrently for successive genomes. This strategy not only improves performance, but also aligns well with the data output of a high-volume sequencer.
- **Tweaks to Lustre.**
  The Lustre network configuration was optimized to take better advantage of the performance capability of the Intel True Scale Fabric host adapters. The configuration of the Lustre filesystem was also optimized for the many small files[10] of a whole human genome dataset.

In addition to these adjustments, the team tried using local HDDs for the temporary storage of data and calculations during analysis, a strategy that is often used in HPC environments. In this instance, however, the use of local HDDs did not improve performance. Using Lustre for temporary storage was just as effective. Using Lustre also provides a simpler and more cost-effective solution for delivering the right amount of temporary storage per compute node, without overprovisioning physical disk capacity.

Achieving Fastest Time-to-Results per Genome

Using the Lustre optimizations above, the fastest time to results were achieved by running seven workflows simultaneously on the cluster with no degradation in runtimes versus the baseline test.[3] That is, each workflow completed within roughly 16.5 hours of the time it was started. This approach can help organizations optimize total throughput for genome analysis, while maintaining the shortest possible runtime for each analysis.

Achieving Highest Total Throughput

Using the Lustre optimizations and staggered workflows, the cluster was able to run 24 workflows concurrently, with all of the analyses completed within 86,400 seconds (24 hours) of the time they were started.[3] This result validates the hypothesis that two 16-node clusters, or a single 32-node cluster, could keep pace with the output of today's highest volume next-generation sequencer operating at full capacity, completing analysis of 1 WGS every 30 minutes.

Additional tests were run to measure performance for analyzing whole human exomes. (For details, see the sidebar, Analyzing 720 Whole Human Exomes per Day, on next page.)

Total Cost of Ownership

A TCO study was performed by QIAGEN Bioinformatics and Intel to estimate the cost of analyzing genomes using the 32-node QIAGEN Bioinformatics and Intel reference architecture and the 85-node cluster defined by Illumina. Hardware, software, maintenance, and data center costs were estimated over a four year lifecycle. Hardware and software component costs were based on currently available pricing at the time of the study.

To help ensure a conservative TCO estimate for the QIAGEN Bioinformatics and Intel reference architecture, cluster costs were estimated for a total of 35 nodes to reflect additional management nodes, and the storage solution was sized at 500 TB. Although this higher storage capacity was not necessary for the performance tests, the analysts felt it was more in line with the requirements of a high volume genomic environment. The TCO estimate also included costs for five additional Biomedical Genomics Workbench clients.

The results showed a 4-year TCO of $2,862,085 for the 85-node cluster architecture and $1,518,747 for the 35-node QIAGEN Bioinformatics and Intel reference architecture.[3] Based on this study, the 35-node cluster provides savings of approximately $1.3 million over four years, while delivering roughly the same level of performance. It should be noted that data center staffing costs were not included in the TCO analysis due to the high degree of variability between institutions and across different geographical regions.

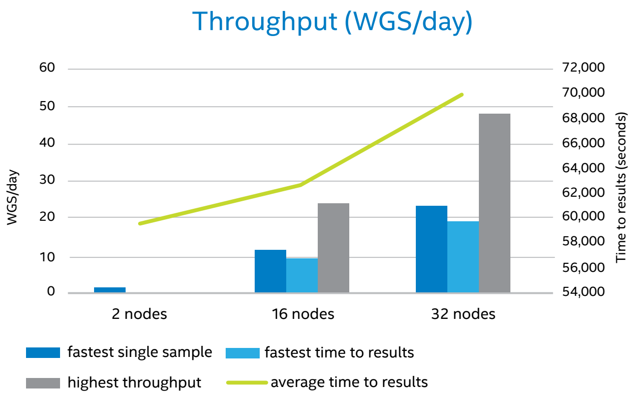


Figure 4.

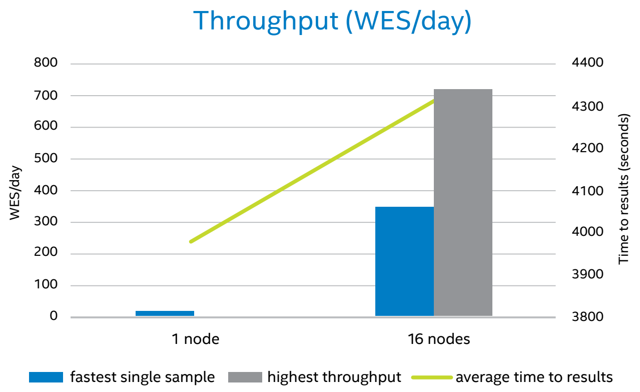
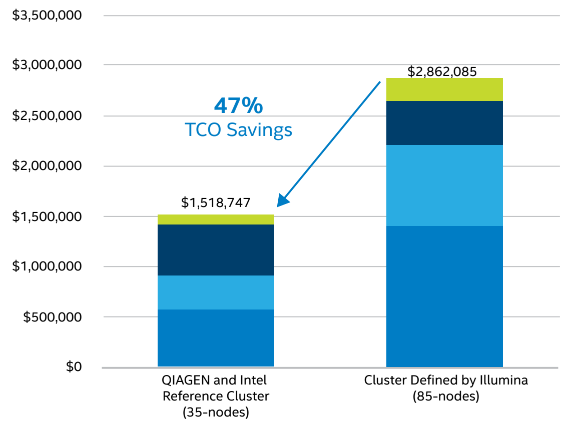


Figure 5.

Total Cost of Ownership over 4 years, Processing 48 Whole Genomes Per Day[2]



	QIAGEN and Intel Reference Cluster (35-nodes)	Cluster Defined by Illumina (85-nodes)
Infrastructure and Utilities	$99,357	$211,615
OS and Software	$509,880	$441,660
Server Maintenance	$335,860	$815,660
Server Hardware Acquisition	$573,650	$1,393,150

Figure 6.

Analyzing 720 Whole Human Exomes per Day[3]

Using the same 16-node test cluster and publicly available data[a], the QIAGEN Bioinformatics and Intel team was able to analyze approximately 720 whole human exomes within 24 hours, which equates to submitting an exome for analysis every two minutes. Most individual exome analyses were completed within 80 minutes, though a few took as long as 2-3 hours.

Based on these results, a single node of the cluster would be able to analyze the output of an Illumina HiSeq 2500* system operating at full capacity[b], delivering an average of up to 30 analyzed exomes per day.

[a] Public Illumina HiSeq 2000 exome sequencing data from A.C. Nichols et al. This dataset can be accessed from the European Nucleotide Archive (ENA) under the accession number SRP018669. The data represents a tumor/normal matched sample pair from a massive acinic cell carcinoma of the parotid gland.

[b] Based on the published maximum output of the Illumina HiSeq 2500 (150 samples per 5 days). http://www.illumina.com/systems/hiseq_2500_1500/system.html

Conclusion

NGS technologies are transforming the field of bioinformatics by generating raw sequencing data at the speeds and volumes needed to provide cost-effective support for population studies and clinical usage models. The QIAGEN Bioinformatics and Intel reference architecture define an HPC cluster, interconnect fabric, and storage solution that can match the output of today's fastest NGS devices, while delivering fully analyzed results for as little as $22 per whole human genome.

Based on QIAGEN Biomedical Genomics Server solution and Intel® Architecture-based compute, fabric, and storage components, the reference architecture delivers a powerful, state-of-the-art genomics workbench for geneticists that delivers high quality results, while masking the complexity of cluster computing. The reference architecture is also designed to scale incrementally at each layer, so organizations can continue to grow their analytics capability in a straightforward and cost-effective manner.

For More Information:

Learn more about QIAGEN Bioinformatics solutions:
qiagenbioinformatics.com
sampletoinsight.com

Learn more about Intel in Health & Life Sciences:
intel.com/healthcare/bigdata
intel.com/healthcare/optimizecode

[1] Based on the published output capacity of the Illumina HiSeq X Ten next-generation sequencer. http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf

[2] Source: Gold standard BWA-GATK pipeline specified in Illumina HiSeq X System Lab Setup and Site Prep Guide, Document 15050093 v03, January 2016, available at https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/hiseqx/hiseq-x-lab-setup-and-site-prep-guide-15050093-03.pdf

[3] Based on internal performance tests and a total cost of ownership analysis performed by QIAGEN Bioinformatics and Intel. Performance tests were conducted on a 16-node high performance computing (HPC) cluster. Each node was configured with 2 x Intel® Xeon® processor E5-2697 v3 (2.6 GHz, 14 core), 128 GB memory, and a 500 GB storage drive. All nodes shared a 165 TB storage system based on Intel® Enterprise Edition for Lustre, 256 TB of 1000 RPM disk storage and 4 x 800 GB Intel® Solid State Drive Data Center S3700 Series. The interconnect fabric featured 2x Intel® True Scale Single Port HCAs (2x QLE7340, QDR-80) per node and a 36-port Intel True Scale switch 12300 (40 Gbps). The TCO analysis was performed using an internal Intel tool and publicly available product pricing and availability as of October 9, 2015. The TCO for the test cluster was estimated over a 4-year period and compared with the estimated TCO of an 85-node cluster, as described in the Illumina HiSeq X System Lab Setup and Site Prep Guide, Document # 15050093 v01, September 2015. https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/hiseqx/hiseq-x-lab-setup-and-site-prep-guide-15050093-01.pdf. To quantify the TCO comparison, specific products were chosen that would fulfill the general specifications defined within the Illumina guide. Support costs for both systems were estimated as 60 percent of TCO. The performance and TCO results should only be used as a general guide for evaluating the cost/benefit or feasibility of a future purchase of systems. Actual performance results and economic benefits will vary, and there may be additional unaccounted costs related to the use and deployment of the solution that are not or cannot be accounted for. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance.

[4] For more information about the data quality advantages of QIAGEN Bioinformatics software, visit https://www.QIAGENbioinformatics.com/solutions/bx-iva-cancer-solution/benefits/.

[5] Lustre is used by nearly 70 percent of the world's fastest high performance computing (HPC) clusters. Source: www.top500.org

[6] Source: "Big Data Meets High Performance Computing," an Intel white paper. http://www.intel.com/content/www/us/en/software/intel-lustre-big-data-wp.html

[7] Learn more about Intel® NVM Express SSDs at: https://www.youtube.com/watch?v=I7Cic0Rb7D0

[8] Reference architecture represents the minimum recommended configuration for these WGS & WES workloads. Suggestions for further performance gains are noted in the text, including the latest generation processors, network/fabric, and solid state drives. The TCO analysis utilized the reference architecture including 3 additional manageability nodes for a total cluster size of 35 nodes and 500 TB total storage. The specific performance benchmark configuration is captured in footnote 3.

[9] The WGS dataset used in the performance tests is NA12878D, which was generated with Illumina's TruSeq Nano kit using 350bp inserts and sequenced on a single lane of an Illumina HiSeq X with > 87% bases and with quality > Q30. Average coverage: 35.57%, Read lengths: 2 x 151, 120 Gb. This dataset can be accessed at https://dnanexus-rnd.s3.amazonaws.com/NA12878-xten.html.

[10] For further Lustre tuning and DSS features, please see: http://blogs.intel.com/intellabs/2014/11/19/dss/.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order. Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's Web site at www.intel.com.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at http://www.intel.com/content/www/us/en/benchmarks/intel-product-performance.html.

Printed in USA 0316/DW/HBD/PDF ♲ Please Recycle 334003-001US