



# Genomics Research Overview

## From Genome Sequencing to Watson

**Kathy Tzeng, PhD, World-wide Technical Lead, Genomic Solutions, STG**  
**Janis Landry-Lane, World-wide Director, Genomic Solutions, STG**





- <http://www.mskcc.org/videos/mskcc-and-ibm-collaborate-applying-watson-technology-help-oncologists>

# Evolution of DNA Sequencers



\$149,000,

Life Technology Ion Proton Sequencer (late 2012)  
<http://www.youtube.com/watch?v=OKhxoGcr4Rk>

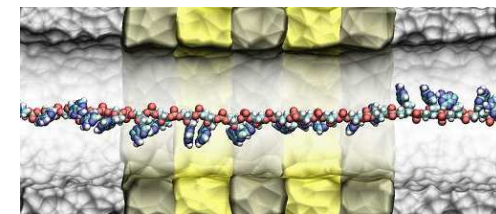


\$740,000,

Illumina HiSeq 2500



Oxford Nanopore GridION (2014?)



IBM DNA Transistor Nanopore based (?)

Source: <http://blueseq.com/knowledgebank/sequencing-platforms/>

# Path for Genomic Medicine

NHGRI, a branch of NIH, has defined 5 steps for genomic medicine.  
(source: E. Green et al., Nature 470, 204–213)

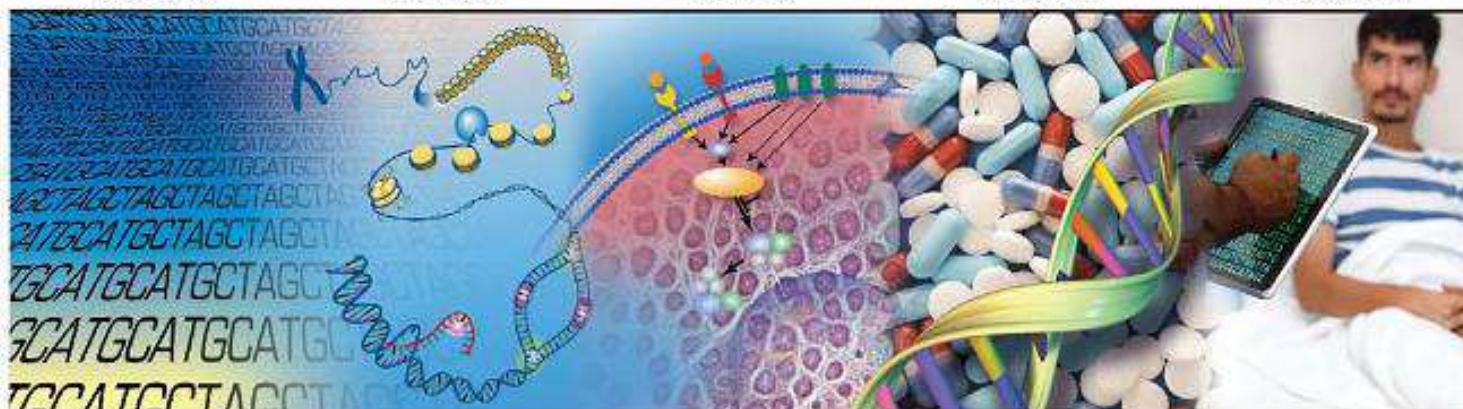
Understanding  
the Structure of  
Genomes

Understanding  
the Biology of  
Genomes

Understanding  
the Biology of  
Disease

Advancing  
the Science of  
Medicine

Improving the  
Effectiveness of  
Healthcare



What needs to be performed	▪Sequencing of genome (not just humans but other organisms)	▪Transcriptome analysis	▪Genotype-Phenotype relationship (GWAS, Epistatic analysis)	▪Chemical genomics	▪Diagnostic
	▪Genome assembly	▪Epigenome analysis	▪QTL Analysis	▪Genome based drug discovery	▪Genetic counseling
	▪Variant call (SNPs identifications)	▪Metagenome analysis	▪Biomolecule interactions (pathway)	▪RNAi development	▪Personalized treatment
	▪Human genetic variation analysis	▪functional genomics	▪Modeling (systems biology)	▪Stem Cell research	▪Prognosis
	▪Structural variants	▪ comparative genomics		▪Protein simulation	▪Preventive
		▪genomics annotation		▪Multi-scale organ simulation	▪Long term life care

Our solution steps:

Sequencing

Translational Medicine

Personalized Healthcare

## GENOMICS– saving lives

---

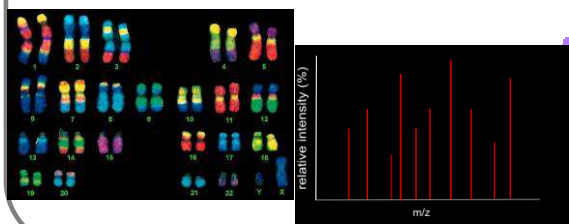


- 2002: Lukas Wartman about to graduate from WashU Medical School
- 2003: Dr. Lukas Wartman, diagnosed with Leukemia
- 2004: chemotherapy, cancer in remission
- 2011: cancer relapsed, marrow transplant didn't work
- 2011: WashU doctors started genomic medicine effort
- **Aug31: WashU Genome Center started sequencing Dr. Wartman's cancer DNA and RNA**
- **Sept: RNA sequencing revealed overactive FLT3 gene in cancer cell**
- **October: Sutent was administered and Leukemia in full remission**

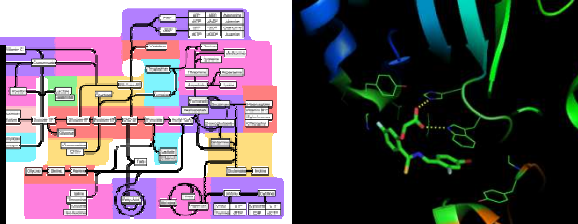


# The Diverse Biospace

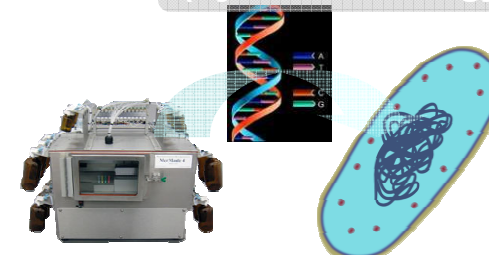
## 'Omics Big Data analytics



## Modeling (Systems Biology)



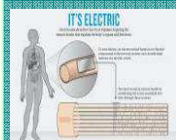
## Bioengineering (Synthetic Biology)



### Healthcare: Genomic Medicine



Genomic based  
personalized  
and preventive  
healthcare



Advanced  
biosensors and  
bioelectronics



Accelerate drug  
discovery,  
development,  
and  
manufacturing

### Agriculture and Food



Salt, drought,  
diseases tolerant  
crop to expand  
arable land

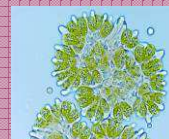


Efficient, Safe,  
and healthy  
meat  
production



Probiotics and  
neutraceuticals  
for disease  
prevention and  
aging care

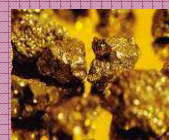
### Energy, Environment, and Natural resources



Next  
generation  
inexpensive  
biofuel



Carbon capture  
and  
Bioremediation  
for clean air,  
water and soil



Rare earth,  
precious metal  
collection

### Chemical, Pharmaceutical and Consumer Products



Green  
Chemistry:  
Bioplastics and  
enzymes



Functional  
polymers such  
as spider silk  
and tires



Cosmetics and  
personal care  
products

# A Large and Growing Market for HPC

## Next Generation

**Genomics:** *Advances that will Transform Life, Business, and the Global Economy*

"Next-generation genomics marries advances in the science of sequencing and modifying genetic material with the latest big data analytics capabilities"

**Potential economic impact, 2025**

**Low High**



**700B-1.6T (US\$)**

### Next Generation Genomics



Fast, low-cost gene sequencing, advanced big data analytics, and synthetic biology ("writing" DNA)

Source: Disruptive Technologies: Advances that will Transform Life, Business, and the Global Economy. McKinsey Global Institute, May 2013

## Insights from an Industry Expert

"Through all the 'omics revolutions, biology is turned into an Information Science"

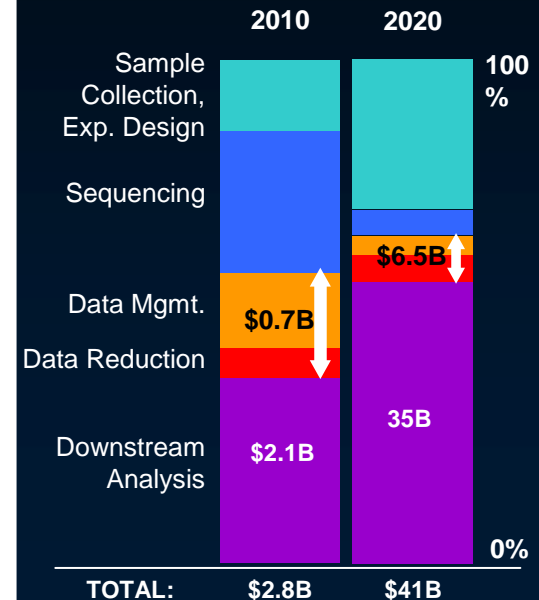
"We have massive amounts of information at the top in the cloud, with users at the bottom in different industries. The biggest barrier to any real use is the lack of standardization and connectivity"

Burrill and Company

July 8, 2013

## IBM Academy of Technology Study on Genomic Medicine

Entire segments of genome analysis are expected to grow. In particular, downstream analysis is growing fast



Source: Sboner et al. *Genome Biology* 2011, 12:125

## **Eric Schadt, Director, Icahn Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine**

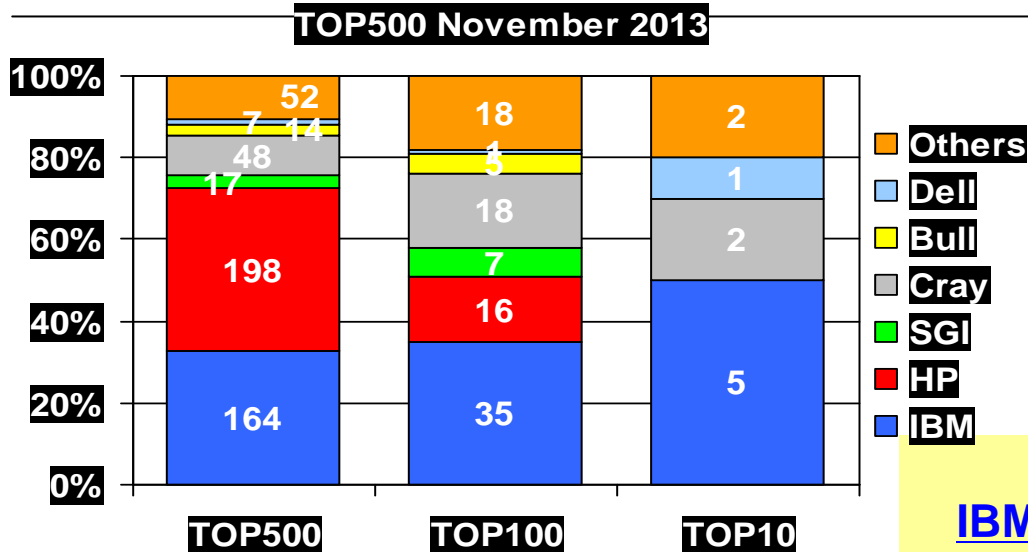
---



- **How does Supercomputing prove useful in Medical research?**
- **There are two main paths. One is in managing the amount of data that can be generated today in the medical arena, things like DNA sequencing. For example, a whole genome sequence of a cancer patient would generate a terabyte of data. If you imagine doing many hundreds of thousands of individuals, you're now into the peta-and even exabyte scales of data. Managing and processing that information down to something that can be medically actionable requires supercomputing infrastructure and expertise.**
- **And then the other path would be coming up with predictive models of disease based on the subtype of disease you have and what treatments may best target that subtype of disease, employing very sophisticated mathematical algorithms that require supercomputing to execute in a timely fashion.**

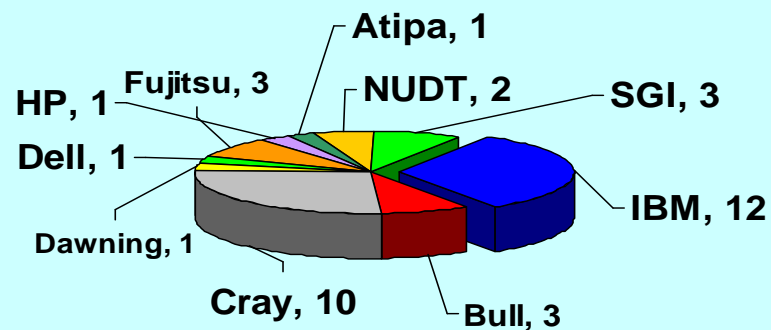


# IBM Supercomputing Leadership



*Semiannual independent ranking of the top 500 supercomputers in the world*

**June 2014 Number of PetaFlop Systems**



## IBM supercomputing leadership ...

- ✓ Most installed aggregate throughput with over 79 Petaflops out of 250 Petaflops (31.6%)
- *IBM leads for 29 Lists in a row*
- ✓ Most in TOP 10 with 5)
- ✓ Most in TOP 20 with 7, most in TOP 100 with 35
- ✓ Most 1 Petaflop or greater systems with 11 out of 31
- ✓ Fastest Intel based system [x86-only]
- ✓ **20 of 25 most energy-efficient systems**

## Washington University Genome Institute

---

**As one of only three NIH funded large-scale sequencing centers in the United States, the Genome Institute is helping to lead the way in high-speed, comprehensive genomics. Since its inception in 1993, the institute has played a vital role in the field of genome sequencing, receiving over \$800 million in funding. The Genome Institute began as a key player in the Human Genome Project – an international effort to decode all 6 billion letters of our genetic blueprint – ultimately contributing 25 percent of the finished sequence.**

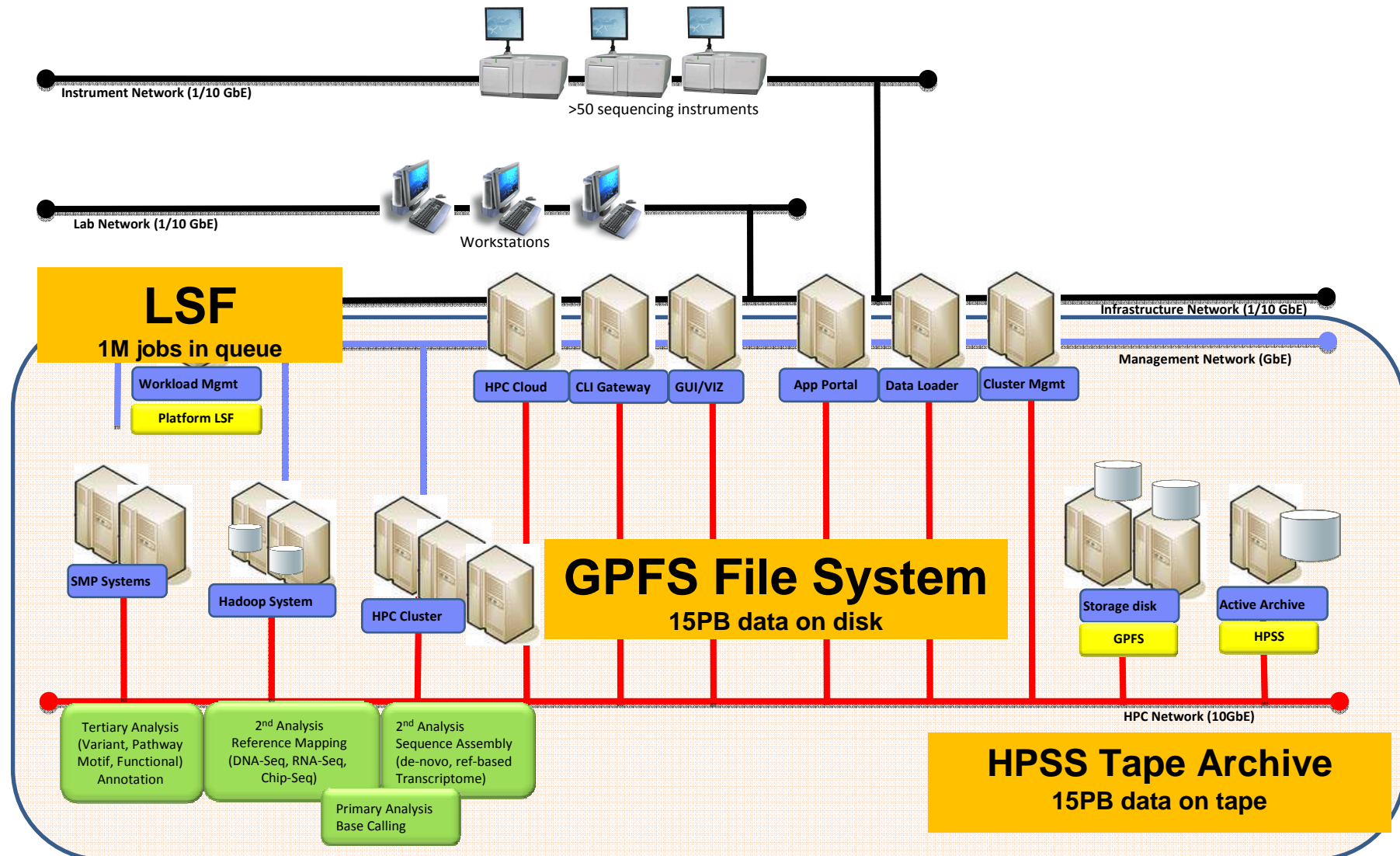


### **Leader in cancer genomics**

**A major goal of the Genome Institute is to advance the emerging field of cancer genomics. In 2008, the Genome Institute became the first to sequence the complete genome of a cancer patient — a woman with leukemia — and to trace her disease to its genetic roots. GI has since sequenced the genomes of many cancer patients including those with breast, lung, ovarian and brain tumors. The Genome Institute has also initiated a major landmark project with St. Jude Children's Research Hospital to sequence the genomes of several hundred pediatric cancer patients.**



# HPC Infrastructure for High-throughput Genomics Research— very large genomics center customer



## MD Anderson “Moonshot”



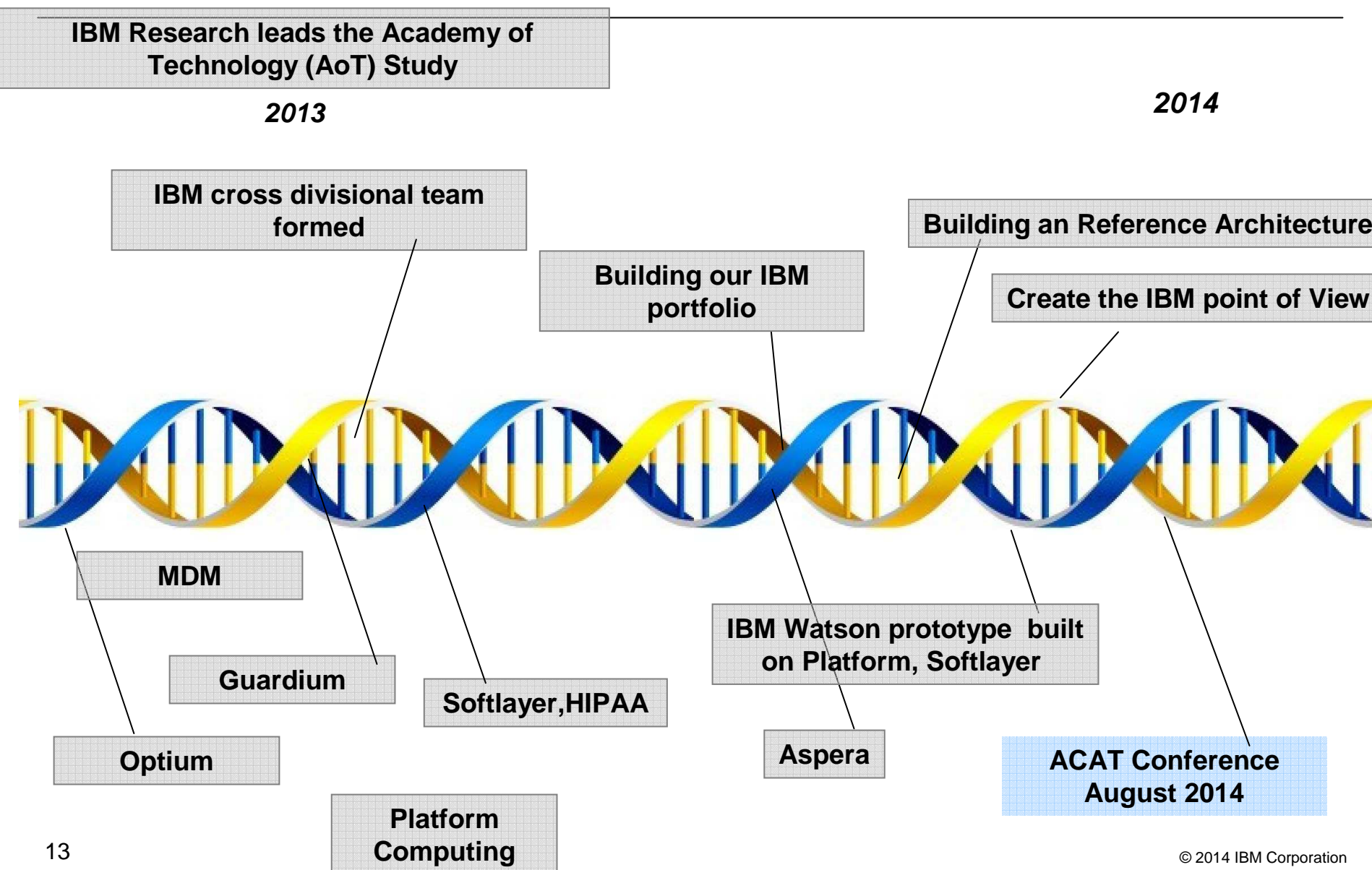
### ***MD Anderson Cancer Center Launches Massive 'Moonshot' Effort Against 8 Cancers***

The teams will focus on personalizing treatment according to an individual's **tumor genes**, assessment of the **effectiveness of therapies** being tried, better diagnoses and early detection and **reducing side effects** of treatment.

IBM's Watson technology is expected to play a key role for the “adaptive learning environment” that MD Anderson is developing, enabling iterative and continued learning between clinical care and research. There is standardization of longitudinal collection, ingestion and integration of patient's medical and clinical history, laboratory data as well as research data into MD Anderson's centralized patient data warehouse. Once aggregated, this complex data is linked and made available for deep analyses by advanced analytics to extract novel insights that can lead to improved effectiveness of care and better patient outcomes. “One unique aspect of the MD Anderson Oncology Expert Advisor is that it will not solely rely on established cancer care pathways to recommend appropriate treatment options,” explained Lynda Chin, M.D., professor and chair of Genomic Medicine, at MD Anderson. “Our cancer patients will be automatically matched to appropriate clinical trials by the Oncology Expert Advisor.

With **genetic information** and more precise drugs, “we have many of the tools we need to pick the fight of the 21st century” and find ways to defeat these cancers, DePinho said.

# Brief History of Genomic Medicine Activities at IBM





## Common Challenges– middleware is important

---

- **Genomic pipeline is a series of HPC steps that need to run at the right time and in the right order to maintain throughput.**
- **Critical processes need to complete with no tolerance for failure in the clinical environment.**
- **There are inter-dependencies between steps**
- **Tools remove tedious manual tasks**
- **In the analytics, there are new dependencies on external, distributed events**
- **Hard to diagnose why the workflow fails**

**Increasingly large, diverse and unstructured datasets**

## Our IBM Genomic Ongoing work:

---

### ■ Address the Community's Preferences:

- Address the need for a HPC Cloud offering for genomics, translational science, and personalized healthcare (IBM Platform Cloud Services in Softlayer– built on our HPC strength)
- Build a coherent public/private cloud offering (which burst capability)
- Start modestly and scale -- add to existing IT resources

### ■ Work Closely with our Partners in Genomic Medicine:

- Partner with Best of Breed Application Providers
- Build the best of breed translational platform with partners
- Middleware, Hardware and the CLOUD are the roles of IBM in genomics
- Utilize IBM Software, Services, and Watson when appropriate
- Create an ecosystem of applications that support the science of genomic medicine

### Contribute to the community:

- IBM Research and IBM STG working alongside practitioners

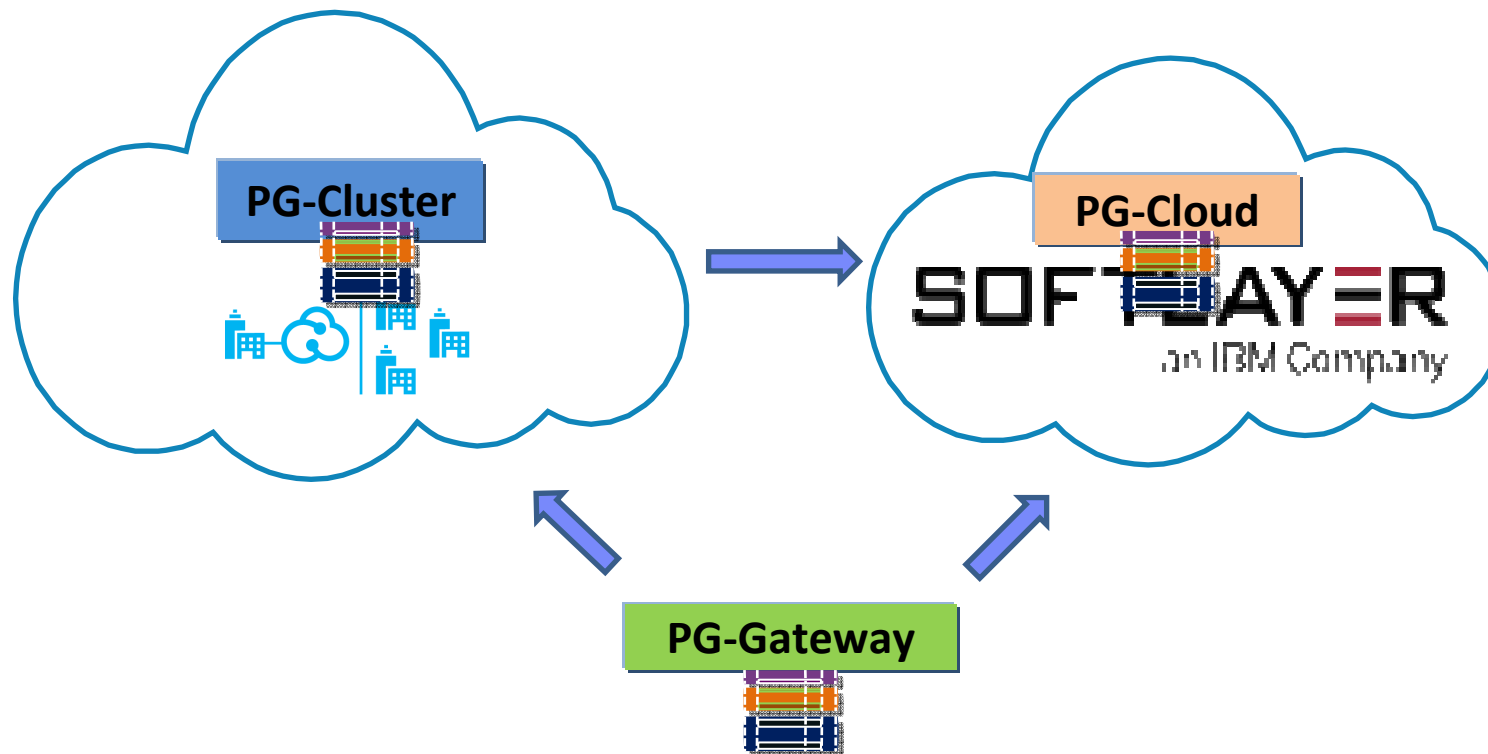
*Speed matters, workload throughput matters, the amount of data stored matters, long term retention matters*

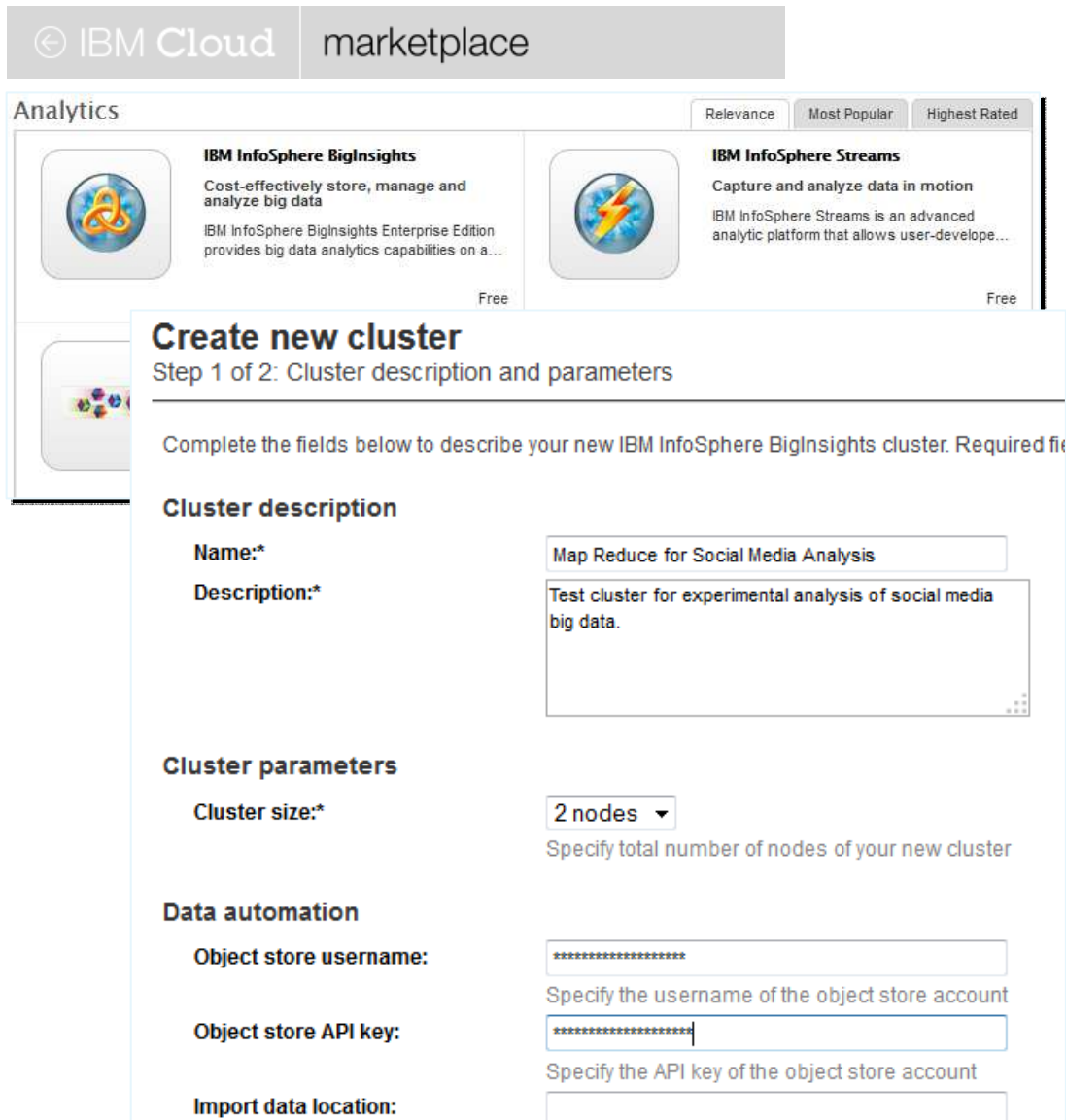
## Utilize our HPC Best Practices to Address Customer Requirements:

---

- Utilize HPC best practices for this HPC workload in-house
  - Total cost of ownership now includes power and cooling
  - There are economies of scale that are important– scaling is **not Moore's law**
- Utilize our consulting arm (GBS) to integrate scientific requirements to a solution
- Utilize middleware to maximize the utilization, and improves workflow
- Design an infrastructure that avoids multiple copies of the data, and have an archiving solution that meets scientific and funding requirements. Can we effectively compress the data?

# The PowerGENE Reference Architecture (PG)





The screenshot displays the IBM Cloud marketplace interface. At the top, there's a navigation bar with 'IBM Cloud' and 'marketplace'. Below this, the 'Analytics' section is visible, featuring two offerings: 'IBM InfoSphere BigInsights' and 'IBM InfoSphere Streams'. Both are marked as 'Free'. A modal window titled 'Create new cluster' is open, showing 'Step 1 of 2: Cluster description and parameters'. The form includes sections for 'Cluster description' (Name, Description), 'Cluster parameters' (Cluster size), and 'Data automation' (Object store username, Object store API key, Import data location).

Analytics

Relevance Most Popular Highest Rated

**IBM InfoSphere BigInsights**  
Cost-effectively store, manage and analyze big data  
IBM InfoSphere BigInsights Enterprise Edition provides big data analytics capabilities on a...  
Free

**IBM InfoSphere Streams**  
Capture and analyze data in motion  
IBM InfoSphere Streams is an advanced analytic platform that allows user-developers...  
Free

**Create new cluster**  
Step 1 of 2: Cluster description and parameters

Complete the fields below to describe your new IBM InfoSphere BigInsights cluster. Required fields are marked with an asterisk.

**Cluster description**

Name:\* Map Reduce for Social Media Analysis

Description:\* Test cluster for experimental analysis of social media big data.

**Cluster parameters**

Cluster size:\* 2 nodes  
Specify total number of nodes of your new cluster

**Data automation**

Object store username: \*\*\*\*\*  
Specify the username of the object store account

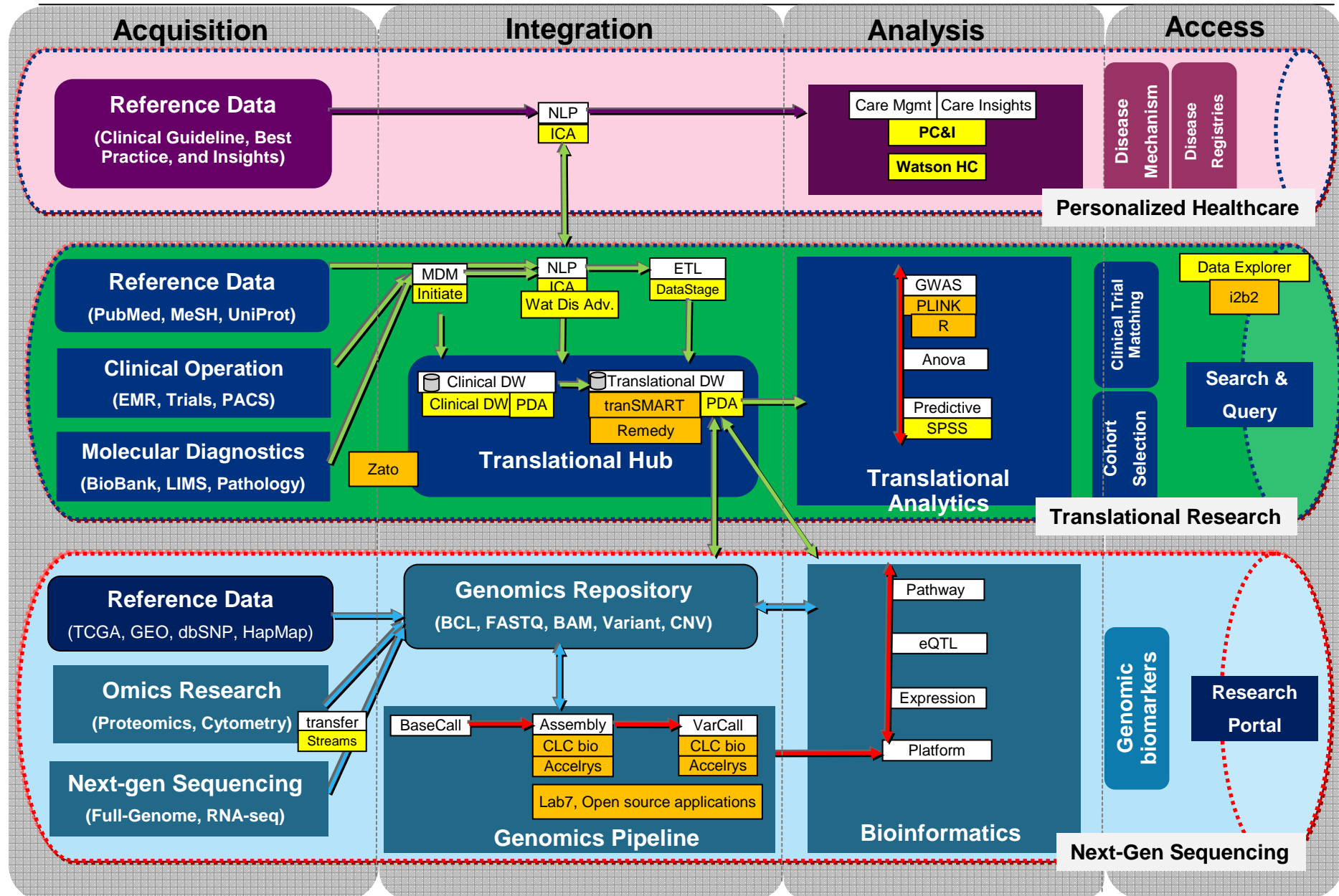
Object store API key: \*\*\*\*\*  
Specify the API key of the object store account

Import data location:

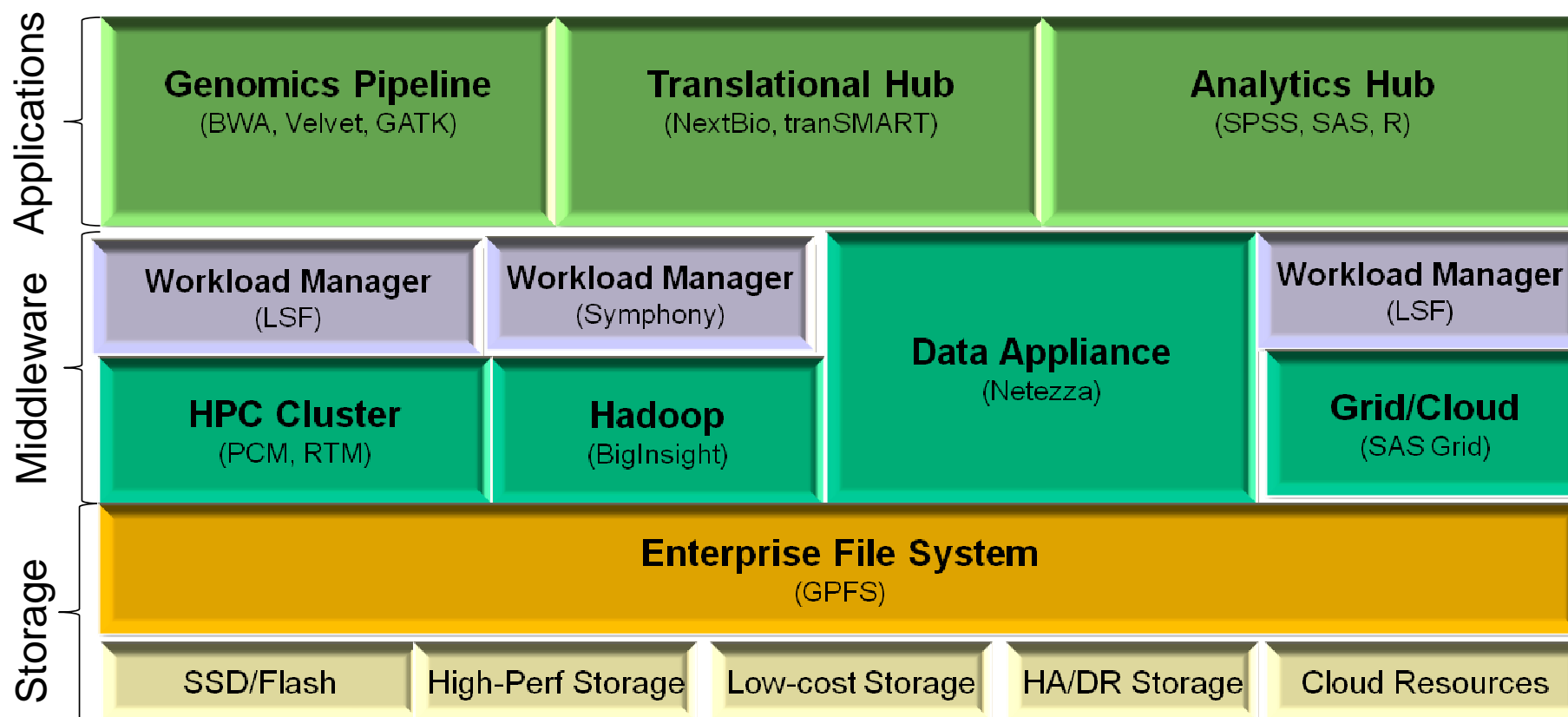
- The offering includes analytics clusters with software and a management portal
- High performance analytics clusters with monthly pricing for large or ongoing BigData analytics; Virtual Analytics Clusters with hourly pricing for short term use
- For Big Data and Analytics users: access to enterprise-grade solutions, running on the IBM Cloud
- For Big Data and Analytics vendors: your solution provisioned on the IBM Cloud for enterprise customers
- Available 4Q in the IBM Cloud marketplace



# IBM Genomic Medicine Reference Architecture

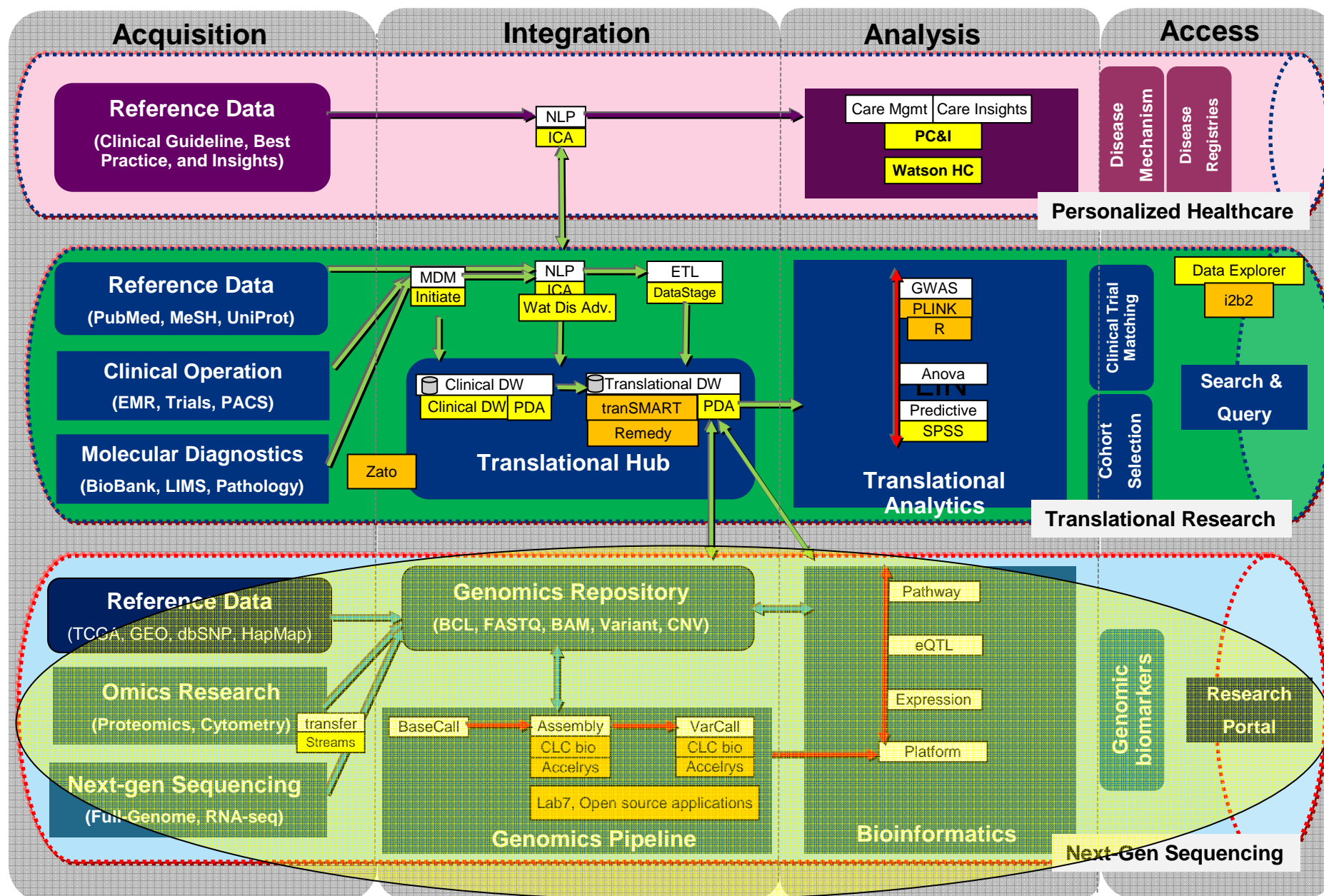


# IBM Genomics Medicine Reference Architecture



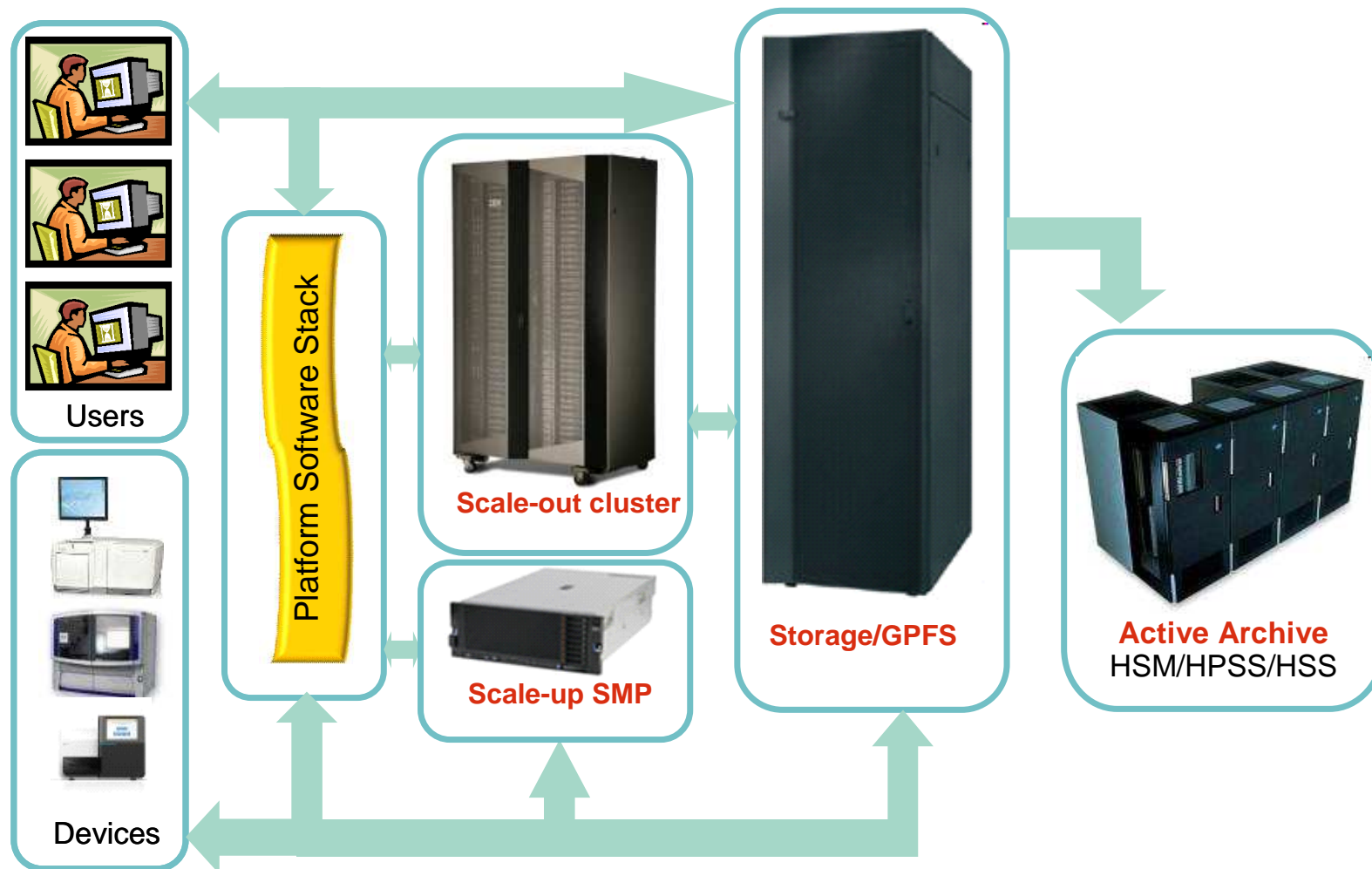
# IBM Genomic Medicine Reference Architecture

Technical Computing





## NGS System Architecture



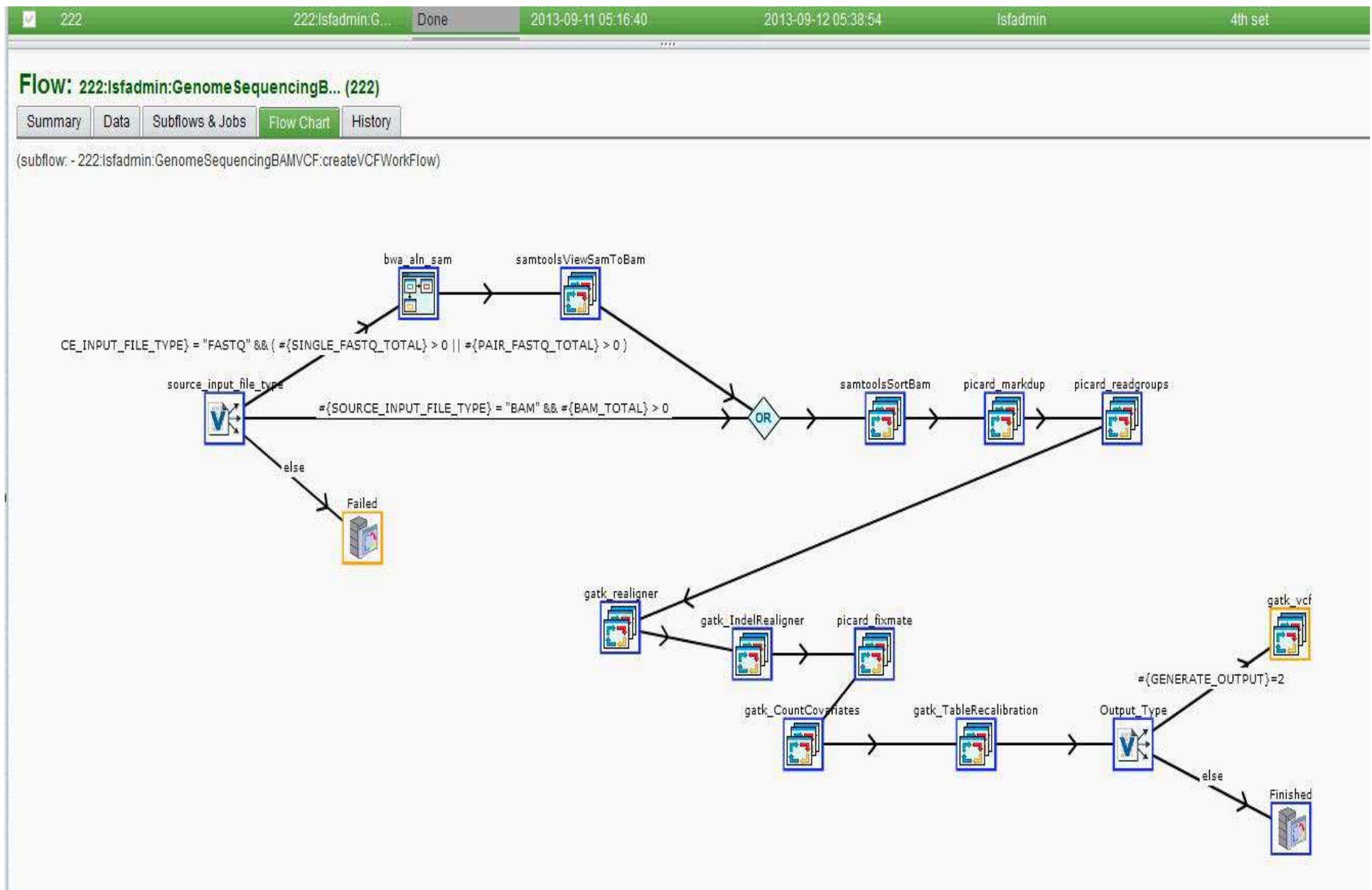
# Sequencing Workflow – Command Line



- `bwa aln -t 12 -l 40 -n 3 -k 2`
- `bwa sampe -a 700 -P -o 1000`
- `samtools view -bt`
- `samtools sort`
- Picard: `java -Xmx8g -Djava.io.tmpdir MarkDuplicates.jar METRICS_FILE=metrics CREATE_INDEX=true VALIDATION_STRINGENCY=LENIENT REMOVE_DUPLICATES=true ASSUME_SORTED=true TMP_DIR`
- Picard: `java -Xmx8g -Djava.io.tmpdir AddOrReplaceReadGroups.jar SORT_ORDER=coordinate RGID=sample_lane RGLB=sample RGPL=illumina RGPU=lane RGSM=sample RGCN=center_name CREATE_INDEX=True VALIDATION_STRINGENCY=LENIENT TMP_DIR`
- Gatk lite: `java -Xmx8g -Djava.io.tmpdir -T RealignerTargetCreator -nt 1`
- Gatk lite: `java -Xmx8g -Djava.io.tmpdir -T IndelRealigner -targetIntervals -known 1000G_biallelic.indels.hg19.vcf`
- Picard: `java -Xmx8g -Djava.io.tmpdir FixMateInformation.jar SO=coordinate VALIDATION_STRINGENCY=LENIENT CREATE_INDEX=true TMP_DIR`
- Gatk lite: `java -Xmx#{JAVA_REQMEM}g -Djava.io.tmpdir -T CountCovariates -recalFile -knownSites:dbsnp,VCF /gpfs/gpfs1/GENOME/SNP_INDEL_VCF/dbsnp_137.hg19.vcf -cov ReadGroupCovariate -cov QualityScoreCovariate -cov CycleCovariate -cov DinucCovariate`
- Gatk lit: `java -Xmx8g -Djava.io.tmpdir -T TableRecalibration -recalFile -sMode SET_Q_ZERO -solid_nocall_strategy THROW_EXCEPTION -nback 7 --baq RECALCULATE`
- Gatk lite: `java -Xmx4g -jar $GATK_BIN/GenomeAnalysisTK.jar -glm BOTH -R $REFERENCE -T UnifiedGenotyper -l recalibrated.bam`



# Platform Process Manager WORKFLOW



## Platform LSF Scheduler Maximizes the Utilization of IT resource

Data Set: 37x coverage of whole human genomes

Workflow Input: 74 fastq.gz files, Workflow Output: Recalibrated Bam file

Dependency steps = Using LSF bsub-w option

Runs	1 <sup>st</sup> Set	2 <sup>nd</sup> Set	3 <sup>rd</sup> Set	4 <sup>th</sup> Set	Total Sets
1 set on 8 nodes	10.06 hrs	-----	-----	-----	<b>10.06 hrs</b>
4 sets on 8 nodes	19.02 hrs	20.9 hrs	21.26 hrs	25.07 hrs	<b>25.10 hrs</b>



# Lab7 and IBM deliver reliable NGS management system for clinicians



## Wholly-Integrated Data and Workflow Management for NGS Using Lab7 ESP™ on IBM Power Systems

As the costs associated with Next Generation Sequencing (NGS) continue to drop, the technique is becoming increasingly attractive to researchers and clinicians who previously would not have considered it, transforming the life science and healthcare industries, and making personalized medicine a reality. Traditionally, NGS laboratories have been in large research institutions that have enjoyed the luxury of having well-established IT infrastructures and bioinformatics teams that could manage the flood of data generated by NGS instruments. The tools and methods developed in these institutions, however, are not easily replicated to work well for the newest adopters of NGS, so an integrated hardware and software solution is needed to make data management and analysis more approachable to these newcomers.

The landscape for the management, analysis, and reporting of NGS data is strewn with an increasingly large set of software tools that independently address specific portions of the overall workflow. While this collection of tools offers flexibility to researchers, there are

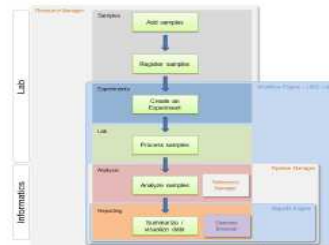


Figure 1: Typical NGS data workflow and Lab7 ESP functional components

significant hurdles to implementing collections of these tools in production environments. For instance, each tool is typically developed under conditions that are completely independent to other tools, using different programming

languages, file types, and references. As a consequence, data flow through these disparate tools is far from streamlined, and provenance of data, critical to regulated environments, is difficult to maintain.

With the Lab7 Enterprise Sequencing Platform™ (ESP), the basic principles of enterprise software are brought to the NGS community by bringing together the disparate functions required to operate an NGS lab under one functional umbrella. The Lab7 ESP combines sample tracking and protocol management with powerful data analysis, reporting, and visualization tools that previously had to be run

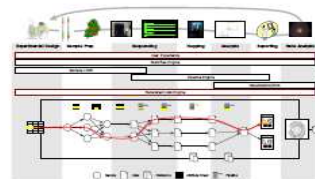


Figure 2: Data provenance through the NGS workflow

independently. All of these functional elements are built upon a robust workflow engine that provides the link between each functional component to ensure rigorous data provenance throughout the workflow process.

### The IBM / Lab7 Advantage for NGS

In addition to the complexities of software for data management, genomic sequencing and analysis requires tremendous compute power and storage capacity.



A compute architecture using IBM® Power Systems and IBM's flagship General Parallel File System (GPFS™) delivers outstanding performance for such demanding requirements.

The Lab7 ESP, a comprehensive enterprise NGS software, is now fully enabled on IBM's Power Systems and integrated with Platform Load Sharing Facility (LSF). This combination allows large amounts of NGS data and end-to-end workflows to be tracked, managed, analyzed, and visualized with efficiency and reliability through the inclusion of the following functional components:

- LIMS Lite - NGS sample tracking and protocol management
- Pipeline Manager - Data analysis management
- Reporting and Visualization - Custom report generation
- Genomic Data Manager - Genomic reference management (e.g. genomes, genome versions, annotations, ontologies)

### The IBM Power Systems Solution

IBM Power Systems deliver trusted, state of the art technology at the small, mid, and enterprise computing levels, and are broadly deployed in production environments worldwide. These systems compete with x86 servers on cost, while delivering greater performance, higher utilization, and superior availability.

Higher performance per core achieved on Power Systems through:

- Massive parallelism (threads) compared to x86
- Higher clock frequencies
- 4-way SMT per core
- Larger POWER L3 on-chip cache
- PowerVM placement optimization

	Intel Ivy Bridge	POWER7+
Clock rates per processor	2.7 GHz	3.61 GHz
Symmetric multi-threading per core	2	4
On-Chip L3 Cache	30 MB	80 MB
Max threads per server	48	128

Also Power Systems are larger servers that have up to 128 threads per server and superior memory performance. With so much compute power consolidated together, Power Systems can provide outstanding performance for NGS workloads, especially for workloads that are highly parallelized and/or have large memory footprints.

Besides superior performance, POWER systems also provide:

- Industry leading RAS features that reduce downtime, and
- Industry leading virtualization features that reduce energy and IT costs and better manage growth without adding complexity.

### IBM GPFS

The growth of data has placed a strain on life science research as organizations add more storage hardware. Traditional network-attached storage solutions are restricted in performance, security and scalability. The IBM General Parallel File System (GPFS) overcomes these issues by not only enabling high-performance, file-based storage access, it also can help in optimizing data management. For IO intensive workloads in NGS file systems and storage play a big role in performance. GPFS's scalable I/O performance significantly benefit various NGS workloads. It also provides benefits such as seamless capacity expansion, improved enterprise wide efficiency, commercial-grade reliability, business continuity and the flexibility of supporting a wide variety of platforms.

### IBM Platform LSF



# The Biovia (Accelrys) and IBM Partnership Delivers Optimized NGS Solution

IBM Systems and Technology Group  
Solution Brief



## Highlights

**Faster time-to-market for new drugs** - Faster NGS allows companies to bring drugs to the market earlier, saving them millions of dollars.

**Preventive and personalized medicine** - Faster NGS fundamentally transforms the discovery and development of drugs and the delivery of therapies.

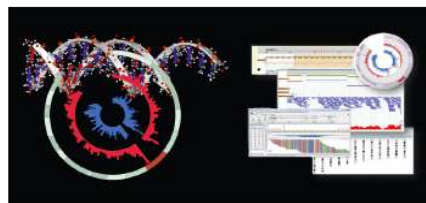
**Agile and Versatile** - Next Generation Sequencing (NGS) Collection provides researchers with a versatile and agile platform to analyze and interpret massive datasets generated by DNA sequencing instruments.

**Extraordinary Performance** - Compute architecture using IBM System x iDataPlex systems and IBM SONAS delivers outstanding performance for demanding NGS applications.

**Reduced Costs** - Half-depth design provides power and cooling efficiencies and ultimate data center space savings and is easy to deploy, integrate, service and manage.

## Accelrys Enterprise Platform NGS Collection Delivers Superior Performance on IBM iDataPlex and IBM SONAS - Complete Human Genome Mapping in a Couple of Hours, Not Days.

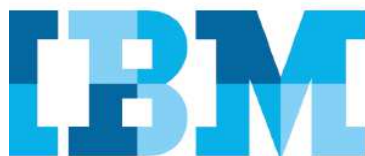
Research teams using next generation sequencing (NGS) technologies are facing the daunting challenge of supporting compute-intensive analysis methods against petabytes (PB) of data while simultaneously keeping pace with rapidly evolving algorithmic best practices. NGS users can now solve these challenges by deploying the Accelrys® Enterprise Platform™ (AEP) and the NGS Collection on new systems offerings from IBM®.



Analysis of whole genome data requires massive compute power to drive diverse algorithms and analysis tools. Accelrys Pipeline Pilot running on IBM hardware provides an optimal environment for the rapid extraction of biological meaning from NGS data.

Faster and affordable NGS is fundamentally transforming the healthcare and life sciences industries. By improving time-to-market for preventive and personalized medicine, companies can save millions of dollars in drug discovery and development while delivering innovative therapies.

The Next Generation Sequencing (NGS) Collection provides researchers with a versatile and agile platform to analyze and interpret the massive datasets generated by current DNA sequencing instruments. Using pre-built protocols, researchers can perform common computational workflows such as De novo sequencing, mapping to reference sequences, and variation detection. With the component collection, they can easily create other NGS protocols tailored to their needs. The AEP NGS Collection supports native data formats from all the major sequencing vendors - allowing researchers to exploit the strengths of each sequencing platform and even combine results to augment analyses and interpretation.

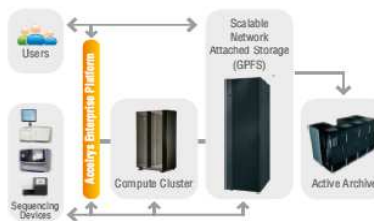


IBM Systems and Technology Group  
Solution Brief

Through the use of a flexible data management repository, the AEP NGS Collection components access reference sequences, mapped reads, and genomic features. The data reader and writers support common formats, such as SAM, BAM, GFF3, or FASTQ, as appropriate, to optimize integration of industry-leading algorithms in this rapidly evolving domain. By leveraging the integration capabilities of the Accelrys Enterprise Platform, other NGS applications and algorithms can be easily integrated into existing data pipelines. This helps researchers take advantage of the latest computational methods while minimizing the effort required to modify existing pipelines.

## The IBM - Accelrys Advantage for NGS

With the cost of sequencing a human genome close to \$1000, faster computing not only saves money (each day's delay in bringing a drug to market costs millions of dollars), but also makes NGS analyses affordable and therefore transforms the healthcare and life sciences industries.



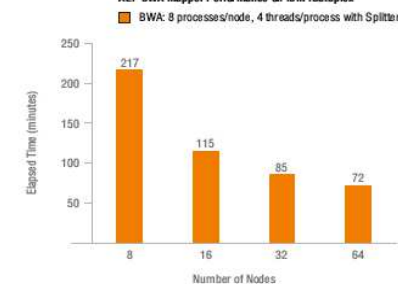
IBM reference architecture for Accelrys Enterprise Platform NGS collection

	GPFS	NFS
Bowtie2	119	437

For example, it takes about two to three days to complete human genome mapping with typical 30x coverage with the widely used BWA mapping algorithm on a single node.

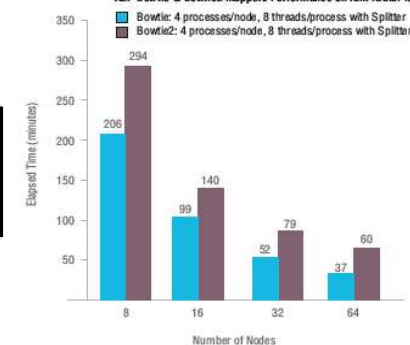
By running BWA with the AEP NGS Collection on the IBM compute architecture, researchers have achieved significant improvement in runtime.

AEP BWA Mapper Performance on IBM iDataPlex



Similarly, it takes a few days to complete human genome mapping with typical 30x coverage using the open source algorithm Bowtie or Bowtie2 on a single node. Running Bowtie or Bowtie2 with the AEP NGS Collection on the IBM compute architecture, researchers have achieved significant improvement in runtime.

AEP Bowtie & Bowtie2 Mappers Performance on IBM iDataPlex







# IBM and CLC bio Deliver Turnkey Solution II 12/2013

IBM Systems and Technology  
IBM Technical Computing

Life Sciences



## Highlights

- IBM and CLC bio genomics sequencing analytics solution for faster time to results
- IBM Application Ready Solution for CLC bio provides optimized IBM hardware/software platform and reference architecture
- IBM Business Partner Re-Store offers turnkey solution

## IBM and CLC bio deliver genomics sequencing analytics solution

IBM and CLC bio provide an accelerated genomics research platform to convert sequencer data to usable genomic insight.

Imagine a world where medical diagnoses and treatment regimens are based on a person's specific genetic makeup—reducing side effects and improving patient outcomes. That's the promise of personalized medicine, which is rapidly becoming a reality through advances in genomic sequencing and analysis.

### APPLYING GENOMIC SEQUENCING TO THERAPEUTICS

Dr. Lukas Warman has firsthand experience with the power of genomic sequencing. A genetics researcher at Washington University in St. Louis, Missouri, Dr. Warman ended up contracting the very disease he was studying: adult acute lymphoblastic leukemia. His condition deteriorated rapidly, and there was no known treatment for the cancer.

His colleagues decided to fully sequence the genes of both his cancerous cells and healthy cells using the High Performance Computing cluster housed in the Genome Institute at Washington University. They discovered something completely unexpected: one of Dr. Warman's normal genes, FLT3, was malfunctioning, producing massive quantities of a protein that was feeding the cancer.

The team found a drug typically used to control the overactive FLT3 gene in patients with kidney cancer. Dr. Warman became the first person to take this drug for leukemia, and his cancer is now in remission. Dr. Warman's case demonstrates how genomic sequencing enables researchers to understand the role of genes in fueling a specific cancer. Consequently, cancer treatment could be customized with drugs that target a gene rather than the tumor or tissue where the cancer first appears.

IBM Systems and Technology  
IBM Technical Computing

Life Sciences

### ESTABLISHING HIGH-THROUGHPUT PERFORMANCE

Because each human genome comprises over three billion base pairs, whole genomic sequencing requires tremendous processing power and storage capacity in order to correlate the variants in the genome with the relevant patient symptoms. Facing increased demand for sequencing, the industry is challenged to drive down cost while speeding up the assembly, mapping and analysis involved in the sequencing process.

To address these issues, IBM and CLC bio have undertaken a joint effort to develop the IBM Application Ready Solution for CLC bio, a next-generation sequencing (NGS) platform. The system was built for practitioners, requiring little IT administration, yet it is scalable, flexible and extendable. This end-to-end solution integrates a computing cluster built on advanced IBM hardware and software, CLC Genomics Server software for high-throughput sequencing, and CLC Genomics Workbench client/desktop software for analyzing and visualizing NGS data.

The cluster compute nodes consist of IBM® Flex System™ x240 powered by Intel® Xeon® E5-2680v2 processors. These nodes are connected to an IBM Storwize® V7000 Unified network attached storage system that consolidates block and file workloads. The Storwize V7000 Unified system also has a single, easy-to-use management interface that supports both block and file storage, helping to simplify administration.

Storwize V7000 Unified system supports file data storage using the IBM General Parallel File System (GPFS™). With its leading file system performance and its ability to scale based on customer needs, GPFS is used in the world's largest high-performance computing (HPC) installations in addition to mainstream technical computing environments. Plus, CLC bio software uses a shared-disk file management solution that provides fast, reliable access to NGS data for optimizing performance.

To simplify the deployment and management of the cluster, IBM Platform™ HPC provides a complete set of technical and high performance computing (HPC) management capabilities in a single product. The rich set of out-of-the-box features reduces the complexity and cost of managing and running an optimized genomics sequencing cluster. Integrated workload management features have been designed to help improve time-to-results and asset utilization.

### PROVIDING A SCALABLE, TURNKEY SOLUTION

IBM Application Ready Solution for CLC bio has been developed in partnership with CLC bio to deliver a scalable, high performance genomics sequencing platform based on an IBM reference architecture. A turnkey solution is available from IBM business partner Re-Store, LLC. It comes pre-integrated with CLC Genomics Server and CLC Genomics Workbench and includes global support and service. The solution is easy to deploy and use, simplifying IT administration and boosting productivity. It has also been designed to scale as workloads expand over time. The solution provides up to 90 TB of effective storage capacity, and administrators can easily add storage extensions and more compute nodes as necessary.

These three analytics solutions have been benchmarked for their mapping, variant calling and filtering performance. CLC Genomics Workbench 6.5 and Platform HPC enabled Genomics Server 5.5 were installed on an IBM server under Storwize V7000 Unified and GPFS. The benchmark was executed using the 37x coverage human genome data set (1,415,483,596 reads, 100 bp/read) and 150x coverage Exome reads (NA12878) from Illumina Genome Analyzer II. Benchmarking showed that the change to Analytics Solutions will perform as follows (see Figures 1 on page 3).



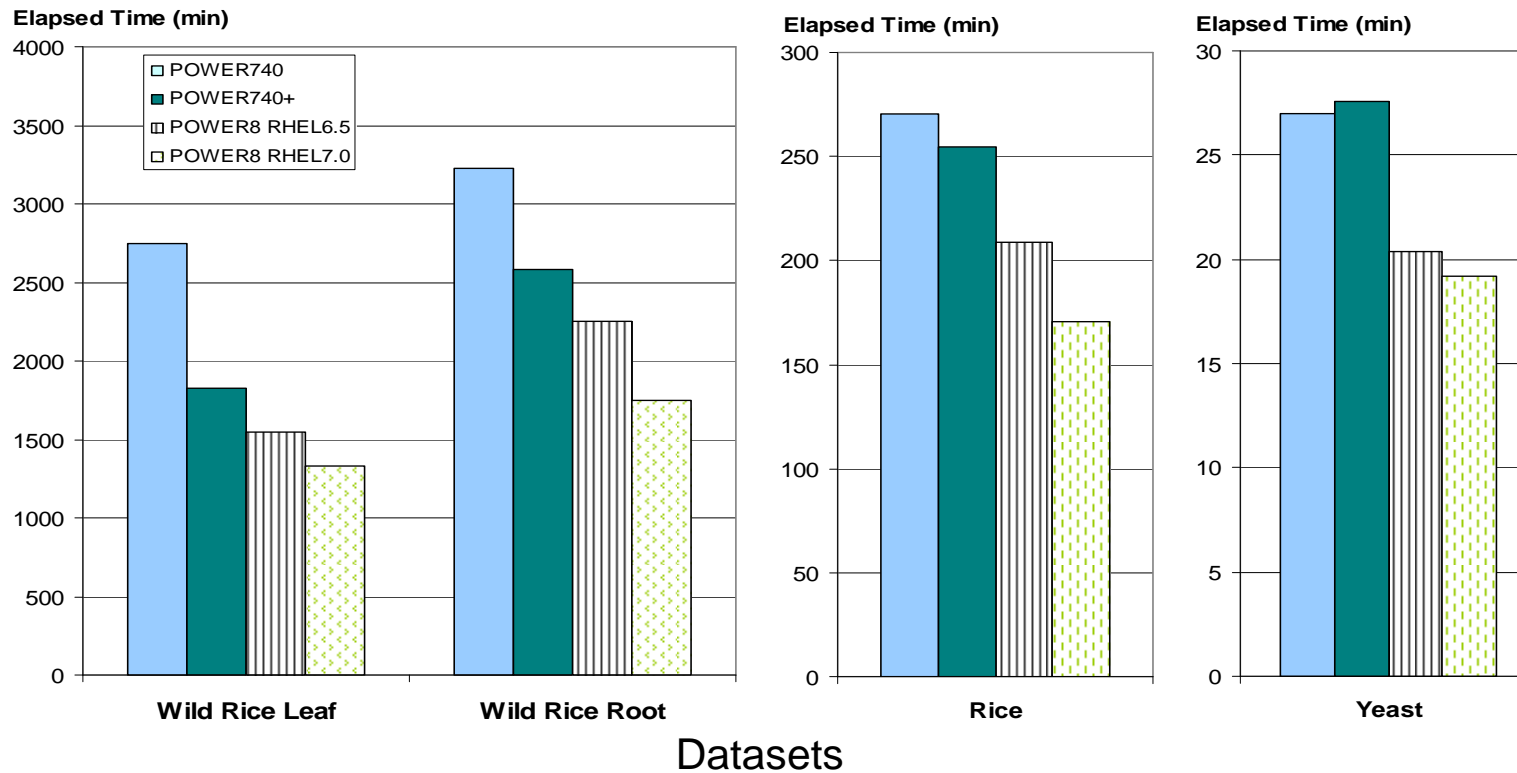


## Comparison of Turnkey Solution Version I & II

	Turnkey Solution Version I	Turnkey Solution Version II
<b>small</b>	7 full human genomes per week	15 full human genomes (37x) per week or 120 human exome (150x) per week
<b>medium</b>	14 full human genomes per week	30 full human genomes (37x) per week or 240 human exome (150x) per week
<b>large</b>	28 full human genomes per week	60 full human genomes (37x) per week or 480 human exome (150x) per week

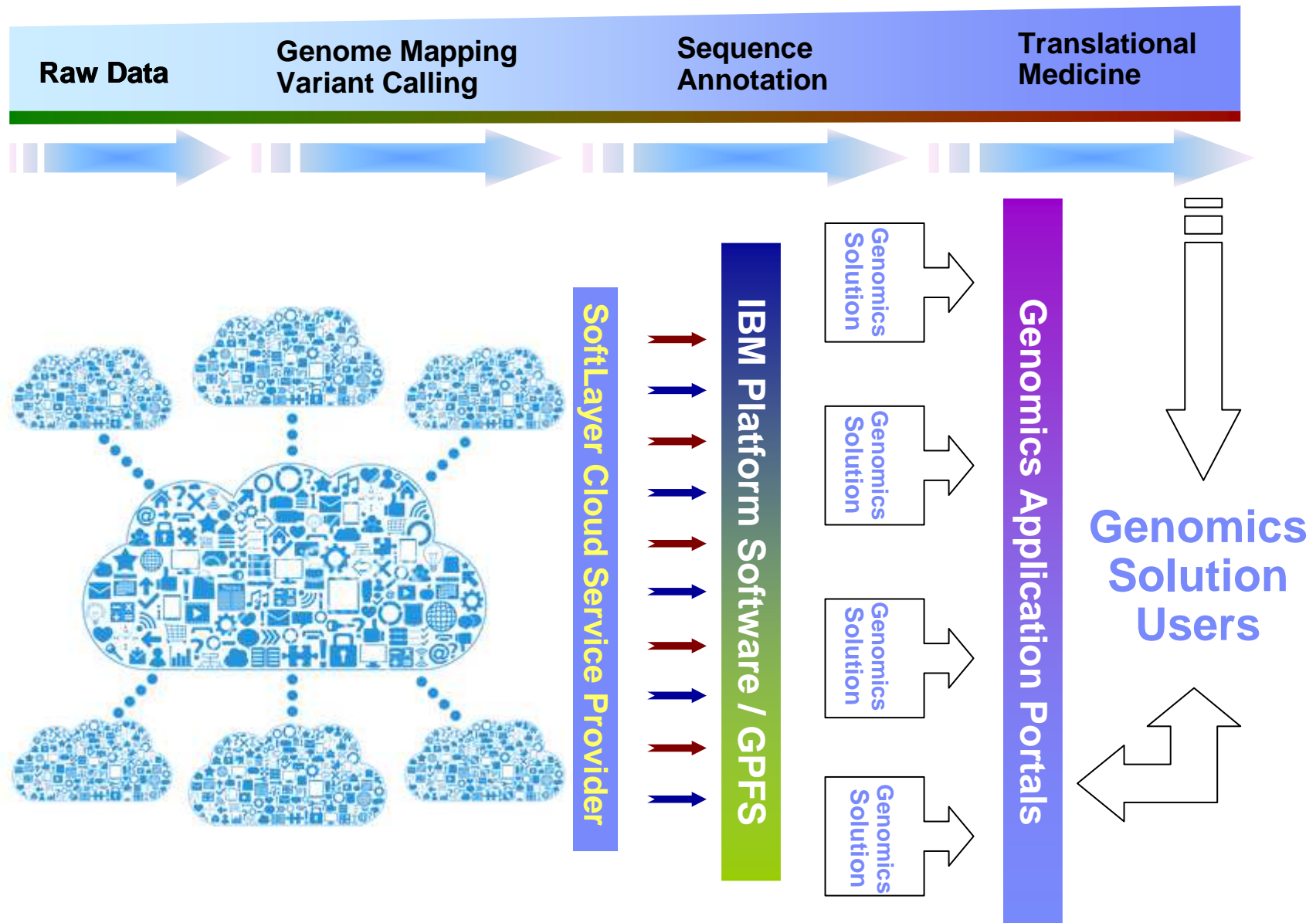
# Performance Benchmarks of Trinity on Power Systems

## -RNA *de novo* Assembly Benchmark



- *De novo* RNA-seq assembly benchmark run on 16-core Power7 740, Power7+ 740 and Power8 S824 running either RHEL6.5 or RHEL7.0
- All four benchmark cases show significantly faster execution time on Power8, with 2x, and 1.5x better performance than P7 and P7+ with large cases (wild rice), respectively.
- Benchmarks on Power8 running RHEL7.0 performs 20-30% faster than on Power8 running RHEL6.5, which does not fully support P8 architecture features.

# Genomics Solution-as-a-Service on SoftLayer



## Benchmark Result of BIOVIA NGS Workflow on SoftLayer

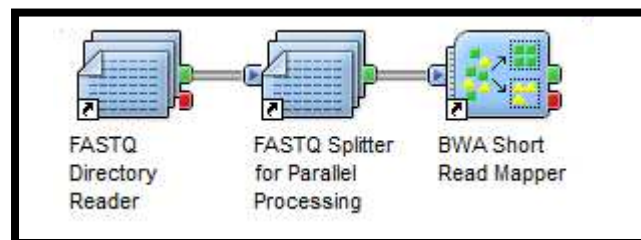
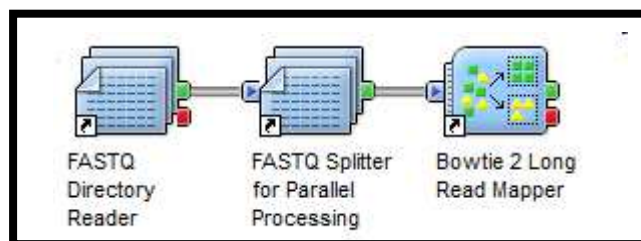
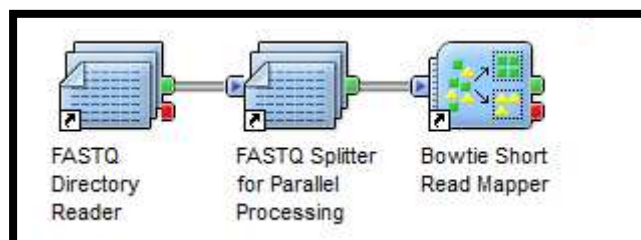
### Human Genome Reference Mapping Benchmark

**SRX000600 experiment:** a paired-end Illumina sequencing of HapMap: NA18507.

**Data Size:**

FASTQ files: 213 runs totaling ~158 GB compressed (~600 GB uncompressed);

Reference: human reference genome build 37.3: ~3GB data.



Benchmark Time (hh:mm)	4 Nodes	8 Nodes
Bowtie	05:01	03:19
Bowtie2	06:45	03:53
BWA	04:01	03:22

The workloads that usually needs days to finish **on single node** now can be completed in hours using **BIOVIA NGS collection** on SoftLayer HPC cloud.

## Benchmark on a single SoftLayer node using CLC bio

Benchmark Time (hh:mm:ss)		37x Coverage Human Genome Benchmark	150x Coverage Human Exome NA12878
Mapping	Preprocessing	00:16:40	00:02:38
	Mapping	03:28:34	00:25:55
	Post processing	02:52:13	00:19:07
	Total	06:37:27	00:47:30
Variant Detections		02:47:50	00:08:57
Filter Known Variants		00:06:56	00:02:13
Total Time		09:25:17	00:58:40

**Note: The 37 coverage workflow benchmark was completed in 10 hours.**

## Benchmark results of CLC bio 4 workflows on 4 SoftLayer nodes simultaneously

Benchmark Time (hh:mm:ss)	37x Coverage Human Genome			
	1	2	3	4
Read Mapping	07:43:13	07:22:32	07:41:32	07:43:03
Variant Detections	02:45:28	03:04:21	02:54:32	03:00:08
Filter vs. Known Variants	00:06:56	00:07:32	00:09:10	00:07:15
<b>Total Time</b>	<b>10:35:37</b>	<b>10:34:25</b>	<b>10:45:14</b>	<b>10:50:37</b>

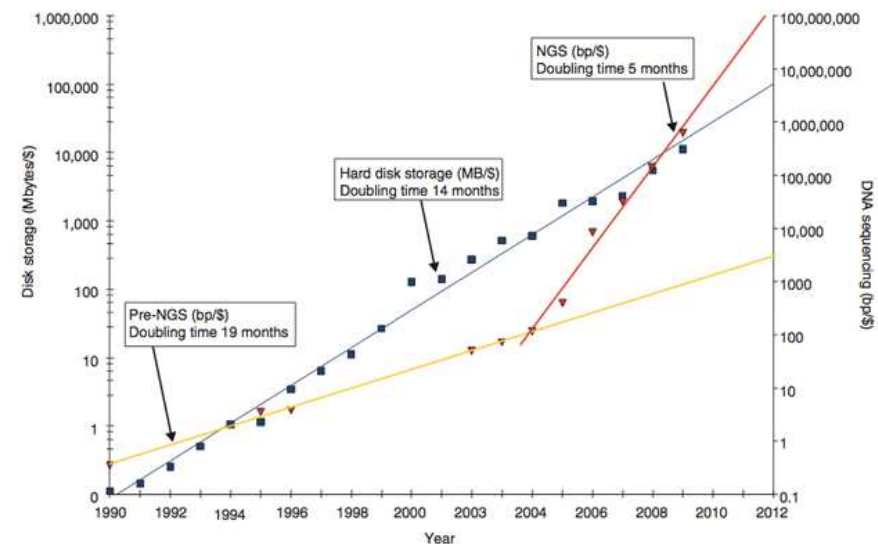
**Note: This is to test the I/O performance on GPFS**



## Use less on-line storage with data compression

IBM Research project	Compression ratio (lossless)	Speed/throughput
gzip on Power7 with FPGA board	On average 1:3 for fastq files	2.5GB/s on average (200 GB fastq can be compressed in 80 second)
Parallelized CRAM	1:2 to 1:4 with respect to BAM files depending on the sequencing depth and other factors. (from FASTQ to compressed BAM ratio is 16X)	Achieved beyond 10 times speed up using 12 cores (approximately 0.5GB/min) FPGA acceleration is ongoing.

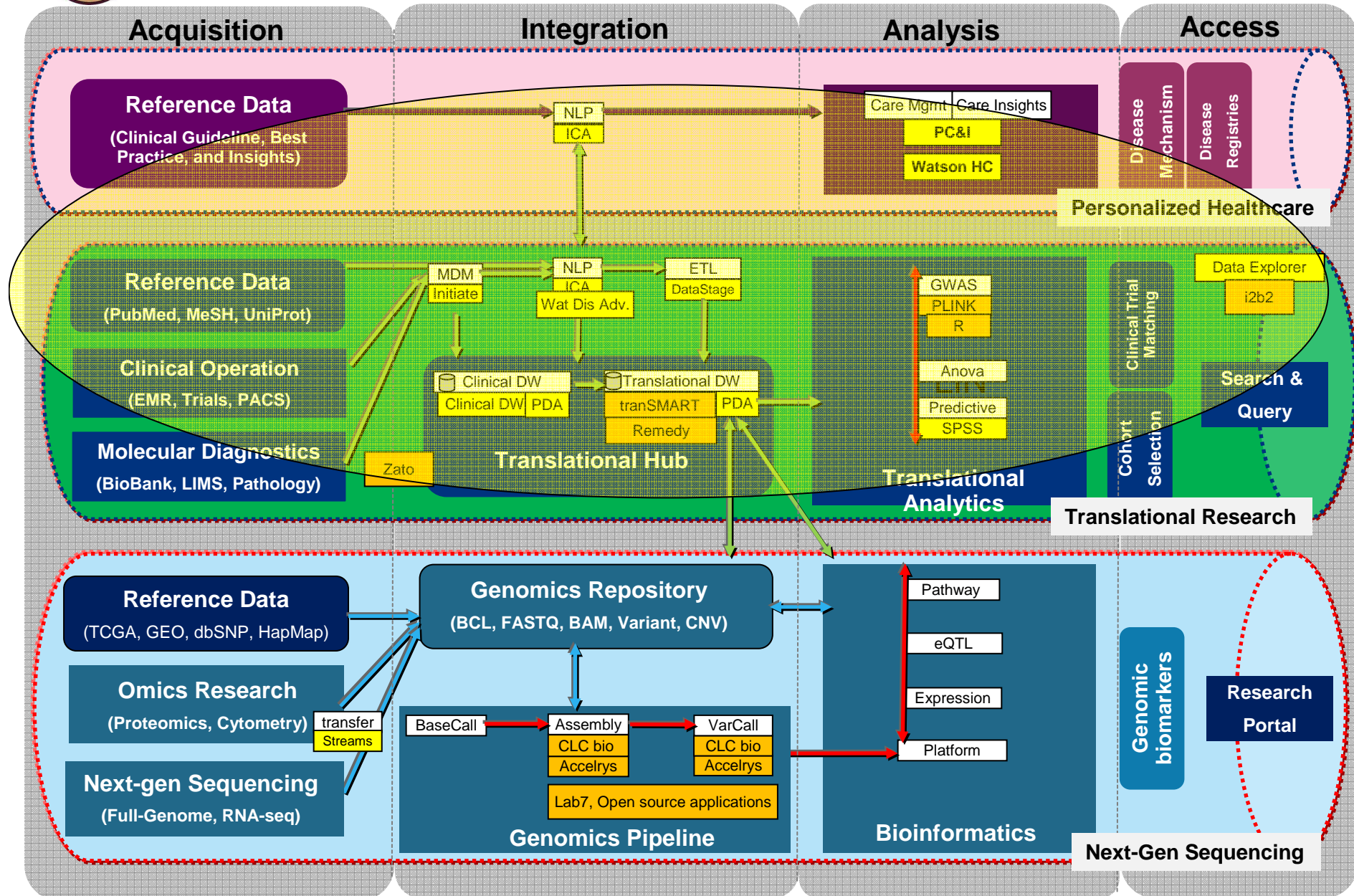
- Data increase is a serious issue in genomics. Since 2004, the data is doubling in every 5 months.
- Compression has been very slow and scientists are reluctant using compression.
- Pistoia compression contest was held in 2012. James Bonfield of Sanger Institute won with 1:9 compression ratio and 0.1GB/min
- CRAM is released late 2012 to compress BAM file by EBI.
- IBM Research is working on improving compression for genomics data.

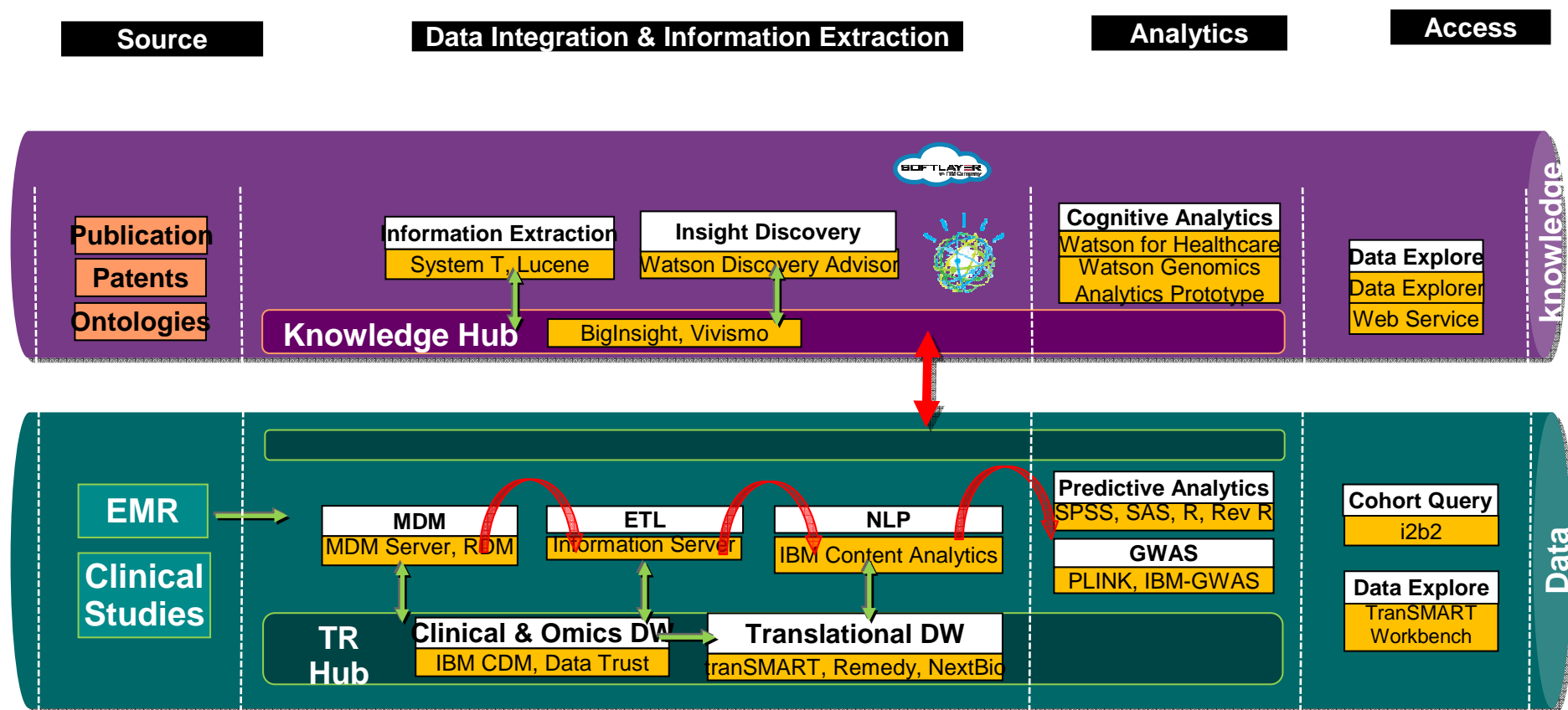


Source: Baker M., *Nature Methods* 7, 495 - 499 (2010)



# IBM Genomic Medicine Reference Architecture





# Case study: Gene-environment interaction analyses

School of Pharmacy and Pharmaceutical Sciences, SUNY Buffalo

## Challenge

- Scanning genomes of Multiple Sclerosis patients to identify gene variations that contribute to the risk of developing MS by searching genetic and environment factors
- Developed a technique they call AMBIENCE but was taken up 27 hours to run a simple study
- Need supercomputer-level processing to handle combinatorial explosion involved in latest data mining methods for gene interaction studies

## Solution


- Deployed an IBM PureData data warehousing appliance as their research analytics infrastructure

## Benefits

- Process exponentially complex gene-environment interactions more than **6,000X as fast** as large high-performance computing systems
- 27 hours to less than 12 minutes
- Using the Appliance embedded R functional
- Carry out their research with little to no database administration.
- Proceed to more complex studies, build more robust models
- Forbes Article : [SUNY Searches Big Data For Multiple Sclerosis Causes](#)







# Vanderbilt Medical Research Hospital discovering connections between drugs, disease, and genetics to provide better care

## Challenge

Vanderbilt University School of Medicine (VUSM) is doing groundbreaking work to identify the genetic basis of diseases and drug response, to make new medical discoveries.

Called the Synthetic Derivative, VUSM created a database from a combination of 20 years of EMR data and their DNA bank (BioVU). This provides researchers with a rich resource to research disease patterns and treatment protocols.

Traditional IT systems were impeding Vanderbilt's work. Each iteration of an analysis often took weeks

## Benefits

Reduce research timeline from up to 1 year to weeks – a 10 fold improvement

Helps researchers and clinicians connect genetic and phenotypic markers to health outcomes.

Helps physicians correlate gene variance to the effectiveness of specific drugs or the likelihood of an adverse reaction.

Researchers can test more theories – even long shots. Often this type of work can lead to unexpected insight



# IBM's Watson analyzes the human genome to battle brain cancer



Glioblastoma, an aggressive and malignant brain cancer, kills more than

13,000

people in the U.S. each year.

New York Genome Center and IBM are partnering in a first-of-a-kind program to accelerate the race to personalized, life-saving treatment for cancer patients.

.....

This Watson application is a cloud-based system that compares genetic data against comprehensive biomedical literature and drug databases.





- NEED Analytics
- NEED High Performance Computing
- NEED High Performance Storage



© 2014 IBM Corporation