Addressing the Data Deluge: How to effectively leverage NGS Data for Pharmaceutical Development and Clinical Research with the accurate, scalable and easy-to-use Bina Genomic Management Solutions.

Sharon Barr, Hugo Y.K. Lam and Narges Bani Asadi, Bina Technologies, Inc., Redwood City, CA, 94065.

ABSTRACT & INTRODUCTION

Next Generation Sequencing Datasets are providing the research and medical community with unprecedented visibility to the origins of health and disease and have the potential of transforming the drug development and the practice of medicine. However these datasets have introduced several new and fundamental IT challenges for the organizations. Overcoming these IT and Analytics challenges is necessary before we can leverage the data effectively for discovery and clinical applications.

Bina Genomics Management System (GMS) is developed to address these challenges, streamlining the processing of genomics information for different functions within an organization to facilitate effective collaboration and communication in the IT, Bioinformatics, Science, and Clinical teams. Meanwhile, GMS is proven to considerably improve the speed, throughput, and the quality of analysis and enable the organization to leverage the massive volumes of data becoming available through clinical and research studies.

Specifically here are some unique values our technology brings to Pharmaceutical, Research, and Medical Institutes facing ever-growing volumes of genomics information:

1. Providing secure effective collaboration for different consumers of NGS data

The genomics information need to be managed, analyzed, and interpreted by many different groups within an organization (as an example researcher, geneticist, IT, lab technician). Bina GMS portal provides very secure environment for managing roles, and permissions, sharing and collaboration, and automatic notifications.

2. Optimized Execution Framework

Bina infrastructure software incorporates our proprietary data modeling and execution engine, leveraging modern technologies such as Hadoop, HBase, and NoSQL that have been optimized for scalable and efficient processing, indexing, and real-time querying of NGS datasets. Our infrastructure has improved the throughput and speed of processing 10-100x times compared to alternative available solutions.

3. Secure and Scalable Data Management

NGS datasets are unstructured and include many different formats : BCL, FASTQ, BAMs, VCFs, Rerports.. At the moment these datasets are sitting in silos without an effective way to be searched, queried, compared, and shared. Through Bina secure data management layer

the organizations will meet their requirements on security, compliance, audits, and provenance while accessing the unique indexing and analytics capabilities of Bina platform.

4. Secure and Flexible Deployment

With the dynamic landscape of IT and security regulations organizations are deploying their software and datasets in hybrid and heterogenous environments including local/private compute environment as well as public cloud solutions. Bina GMS Solution supports all the different deployment options and provides integration across different platforms.

5. Supported and maintained best in class analysis tools

(Open and Extensible, Optimized Workflows and Analytics)

Bina solution provides the most comprehensive and the best-in-class tools for the analysis of NGS data in each step. Our platform is open and extensible to include user tools and datasets as needed.

- Processing and QC of raw data (FASTQ)
- Processing of data for alignment, Variant Calling, CNV, SV, and Somatic analysis supporting WGS, WES, RNA-Seq samples (BAM, VCF)
- Processing the data for annotation and analytics supporting 200 annotations and growing including drug interactions, protein functions, pathways, disease associations, and population studies. (Annotation)
- Processing the cohorts or family datasets for clinical findings or discovery (Analytics)

Bina Genomic Management Solutions (GMS) & Deployment Flexibility

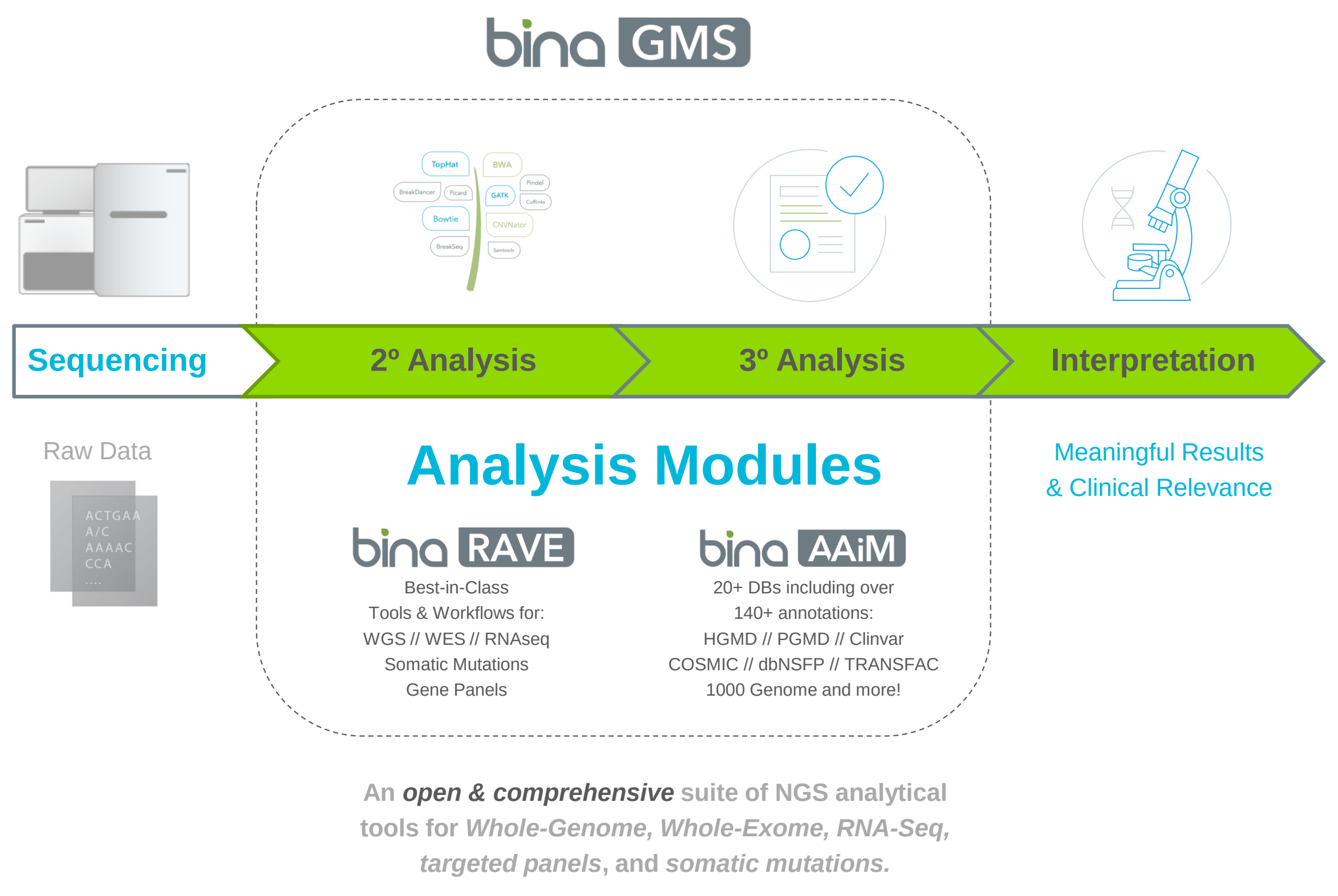
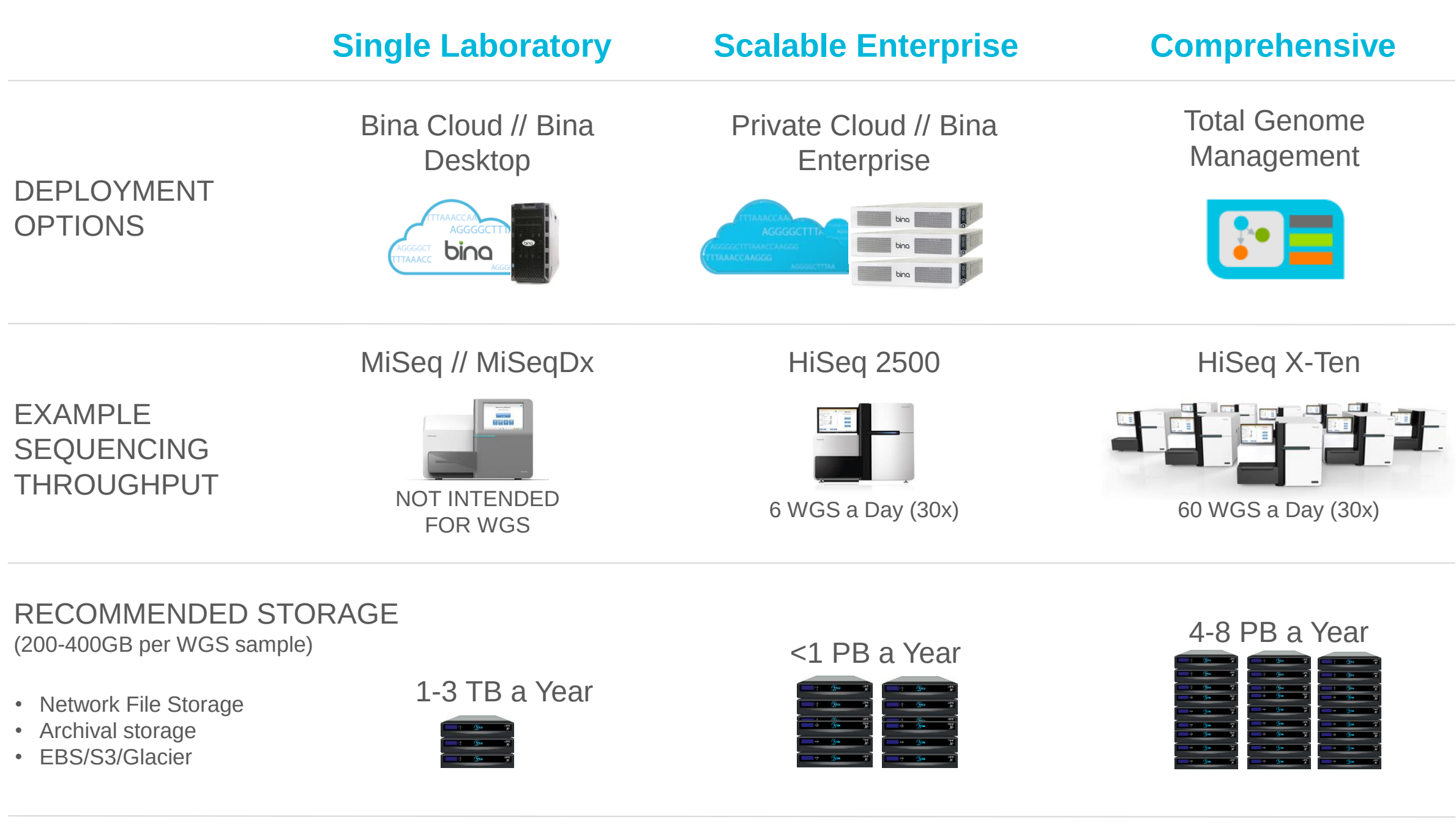


Figure 1. The Comprehensive Bina Genome Management System (Bina-GMS) includes analysis modules for secondary analysis (Bina-RAVE) and annotation (Bina-AAiM) for NGS data analysis, annotation and interpretation.

Scalable Deployment Options



	Single Laboratory	Scalable Enterprise	Comprehensive
DEPLOYMENT OPTIONS	Bina Cloud // Bina Desktop	Private Cloud // Bina Enterprise	Total Genome Management
EXAMPLE SEQUENCING THROUGHPUT	MiSeq // MiSeqDx NOT INTENDED FOR WGS	HiSeq 2500 6 WGS a Day (30x)	HiSeq X-Ten 60 WGS a Day (30x)
RECOMMENDED STORAGE (200-400GB per WGS sample) • Network File Storage • Archival storage • EBS/S3/Glacier	1-3 TB a Year	<1 PB a Year	4-8 PB a Year

Figure 2. The Comprehensive Bina-GMS deployment options support a broad range of linear scalability needs and implementation strategies to accommodate growth in sequencing throughput, including on-premises and cloud-based computing.

Bina RAVE: Secondary Analysis

bina RAVE Example GMS Workflow (data view)

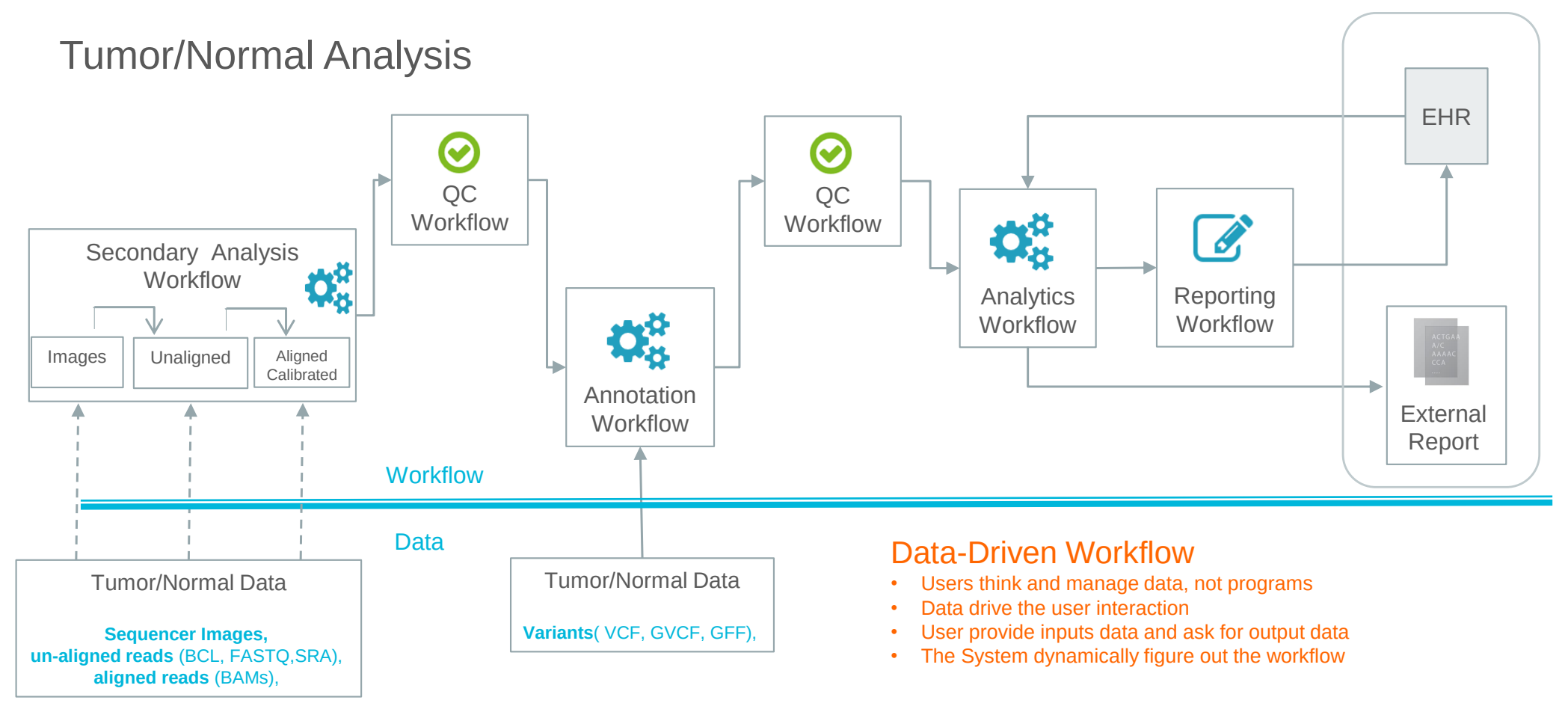

Tumor/Normal Analysis
Secondary Analysis Workflow
Images
Unaligned
Aligned Calibrated
QC Workflow
Annotation Workflow
QC Workflow
Analytics Workflow
Reporting Workflow
EHR
External Report
Workflow
Data
Tumor/Normal Data
Sequencer Images, un-aligned reads (BCL, FASTQ,SRA), aligned reads (BAMs),
Tumor/Normal Data
Variants(VCF, GVCF, GFF),
Data-Driven Workflow
Users think and manage data, not programs
Data drive the user interaction
User provide inputs data and ask for output data
The System dynamically figure out the workflow

Figure 3. NEED A LEGEND Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Table 1. Speed and throughput of Bina RAVE when analysis was run using a single rack-mountable Bina Rack for on-premises hardware deployment.

		WGS	WES	RNA-Seq
Deployed on 1 Bina Rack	Turnaround Time	~6.7 hrs	0.5 hr	4 hrs
	Max Throughput	~4.5/day	117/day	24/day
	Annual Throughput	~1600 WGS/year	~42K WES/year	~8.7K RNA/year

Data specification

Samples:

- WGS: Illumina Platinum Genomes for the NA12878 (30X)
- WES: Illumina Nextera Capture for the NA12878 from GiaB (100X)
- RNA: Illumina Paired-end Sequencing for the hESC from ENCODE (50M reads)

Pipeline:

- WGS: BWA-MEM 0.7.5a, GATK 3.1 (Best Practices with Haplotype Caller & VQSR), 4 SV callers
- WES: BWA-MEM 0.7.5a + GATK 3.1 (Best Practices with Haplotype caller & VQSR)
- RNA: Tophat 2.0, Bowtie 2 2.1.0, Cufflinks 2.1.1

Hardware / VM specification:

Local HW: 4 nodes

Each node has:
16 Cores, 128GB RAM, 1TB HDD

Virtual - AWS:
5 CR1.8xlarge VMs

Each VM has:
32 vCPU, 244GB RAM, 240GB SSD

Bina AAiM: Tertiary Analysis

bina AAiM Annotation, Analytics, Intelligence Module

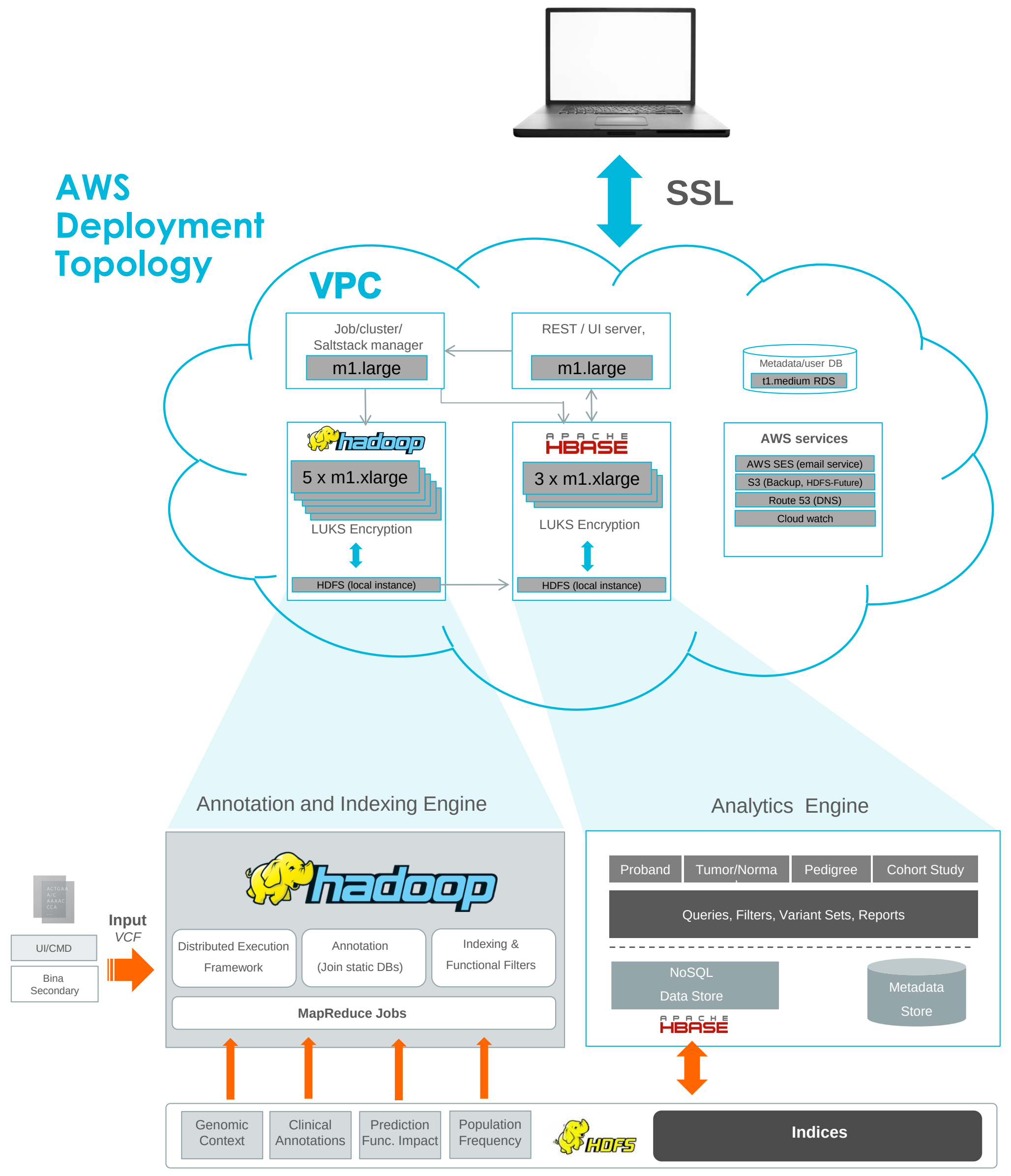
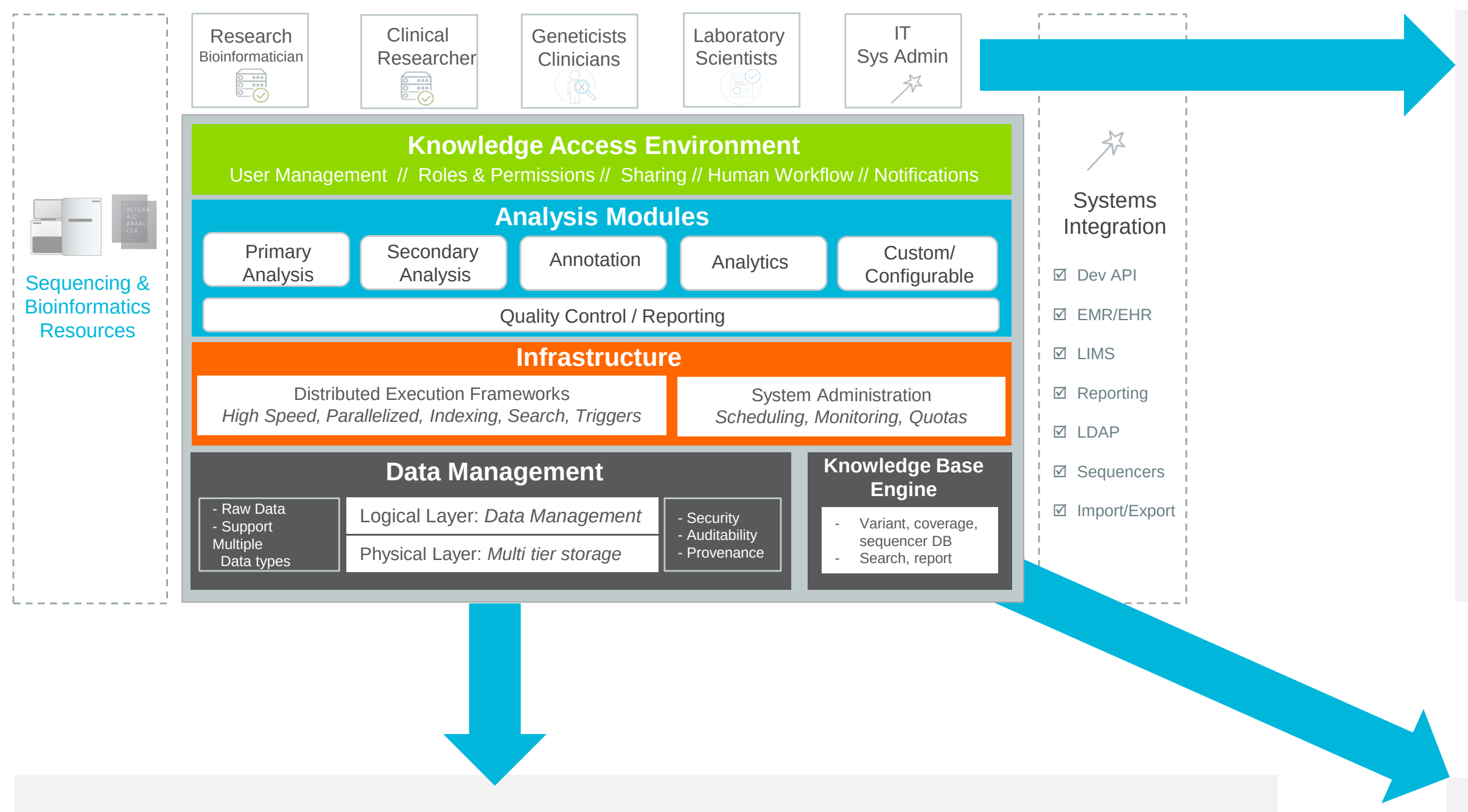


Figure 4. Bina AAiM deployment model on the cloud with an AWS deployment option as an example

Bina GMS: A scalable data management and collaboration platform

bina GMS Bina Genomic Management System Architecture


Research Bioinformatician
Clinical Researcher
Geneticists Clinicians
Laboratory Scientists
IT Sys Admin
Sequencing & Bioinformatics Resources
Knowledge Access Environment
User Management // Roles & Permissions // Sharing // Human Workflow // Notifications
Analysis Modules
Primary Analysis
Secondary Analysis
Annotation
Analytics
Custom/ Configurable
Quality Control / Reporting
Infrastructure
Distributed Execution Frameworks
High Speed, Parallelized, Indexing, Search, Triggers
System Administration
Scheduling, Monitoring, Quotas
Data Management
- Raw Data
- Support Multiple Data types
Logical Layer: Data Management
Physical Layer: Multi tier storage
- Security
- Auditability
- Provenance
Knowledge Base Engine
- Variant, coverage, sequencer DB
- Search, report
Systems Integration
☑ Dev API
☑ EMR/EHR
☑ LIMS
☑ Reporting
☑ LDAP
☑ Sequencers
☑ Import/Export

Personas

Bioinformatician

- New tools
- Best practices
- Define workflows
- Visualization
- Knowledge Base Engine Access

Clinical Researcher

- Annotation
- Definition queries
- Aggregate Analysis
- Case Control
- Visualization

Geneticists Clinicians

Laboratory Scientists

- Execute queries
- Validate workflow
- Reporting

IT Sys Admin

- Integration
- Security
- Compliance
- Different deployments
- Cloud elasticity
- User-defined personas

Data Management Layer

- **Data types abstraction**
  - Sequencer Images, Un-aligned reads (BCL, FASTQ,SRA), Aligned reads (BAMs), Variants(VCF, GVCF, GFF), SVs, Annotations (VCF, MAF)
  - Expressions (FPKM), isoforms (GFF)
  - Reports' raw data (CSV), reports (PDFs)
- **Logical data management**
  - Separation of the data access and storage
  - Provenance
  - Auditability
  - Sharing
  - API to different data types
- **Physical layer**
  - Multi tier storage:
    - Local: Memory, SSD, spinning, archive
    - AWS: VM resources, EBS, S3, Glacier, Zone awareness
  - Geographically distributed
  - Data transition policies (IT persona)
- **Security**
  - In transit encryption
  - On disk encryption
  - Secret key management
- **Compliance**
  - US: HIPAA, FISMA, FedRAMP
  - Country specifics regulations

Knowledge Base Engine

- **Knowledge data**
  - Variants database
  - Coverage database
  - Sequencers database
- **Central data repository**
  - Projects (raw and processed data, reports, results)
  - General search capacities
- **Feedback loop**
  - RAVE
  - AAiM
  - Improve research and discovery over time

Figure 5. Bina GMS consist of few layers including Knowledge access (UI and usability), Analysis modules (RAVE and AAiM), Distributed execution infrastructure, Data Management layer which consists of logical and Physical sub layers and provides security, auditing, provenance, etc, and Knowledge Base Engine which aggregate knowledge of samples and feed it back to the analysis modules

REFERENCES:

1. REF1
2. REF2
3. REF3
4. REF4

www.bina.com

Contact Us
rd@bina.com