

# **Bigger and Better Data**

## **Lessons from Frontlines of Precision Medicine**

Frank Lee PhD  
IBM Global Industry Leader for Systems Group

**BIG DATA**  
**& BUSINESS ANALYTICS GROUP**  
[bigdata.wayne.edu](http://bigdata.wayne.edu)



WAYNE STATE  
UNIVERSITY

*Getting Your Transformation Right*

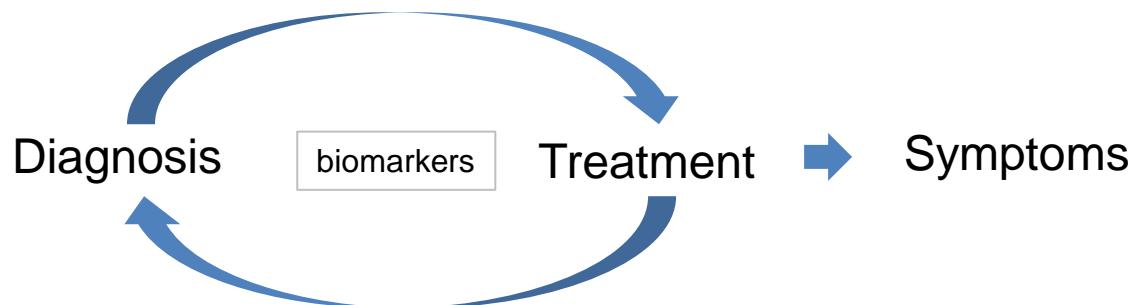
**SYMPOSIUM**

March 22-23, 2018

# Precision Medicine: A Case Study for Speed, Smart & Scale

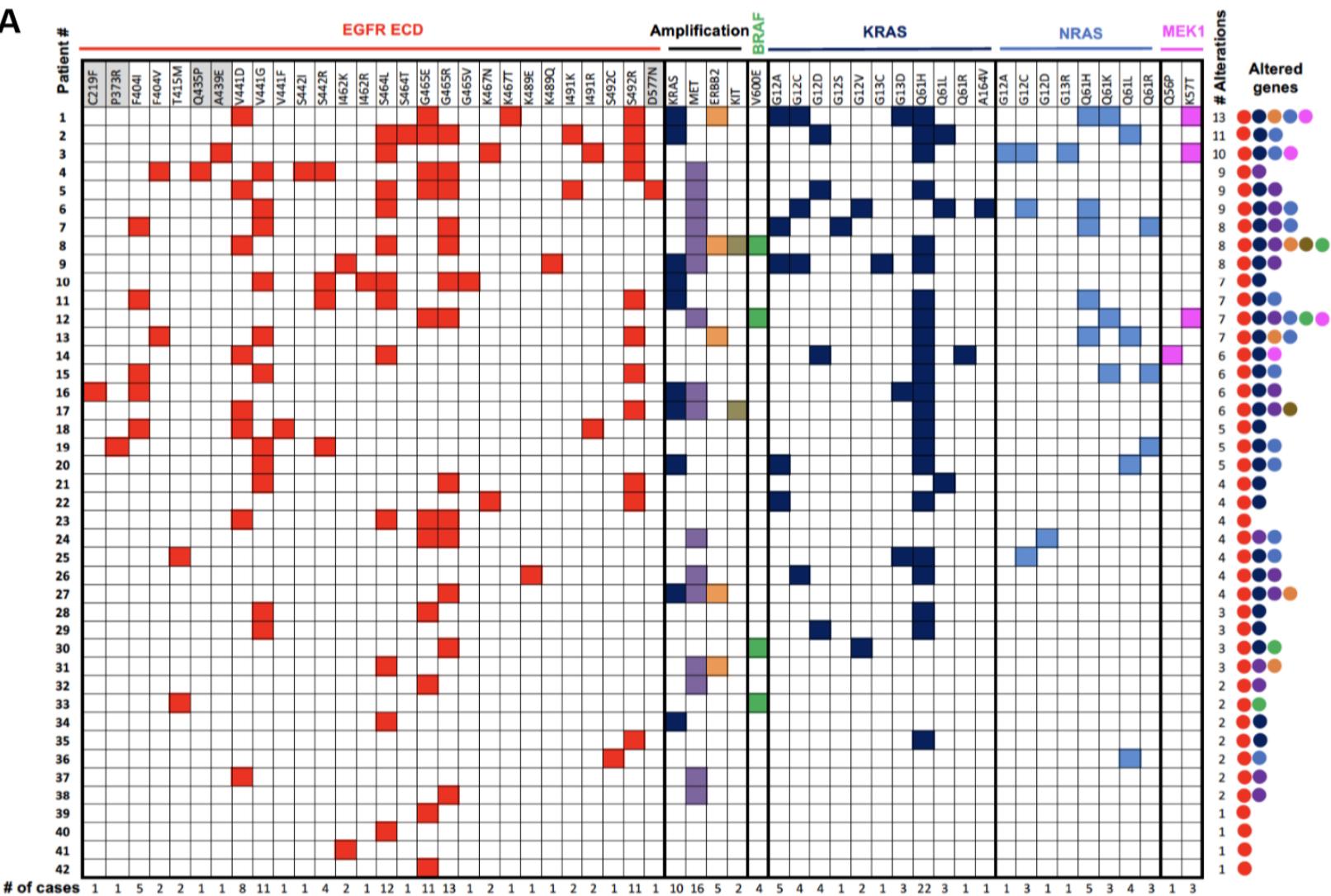
# Paradigm Shift

1<sup>st</sup> Symptom → 1<sup>st</sup> Diagnosis → 1<sup>st</sup> Treatment  
2<sup>nd</sup> Symptoms → 2<sup>nd</sup> Diagnosis → 2<sup>nd</sup> Treatment  
3<sup>rd</sup> Symptoms → 3<sup>rd</sup> Diagnosis → 3<sup>rd</sup> Treatment



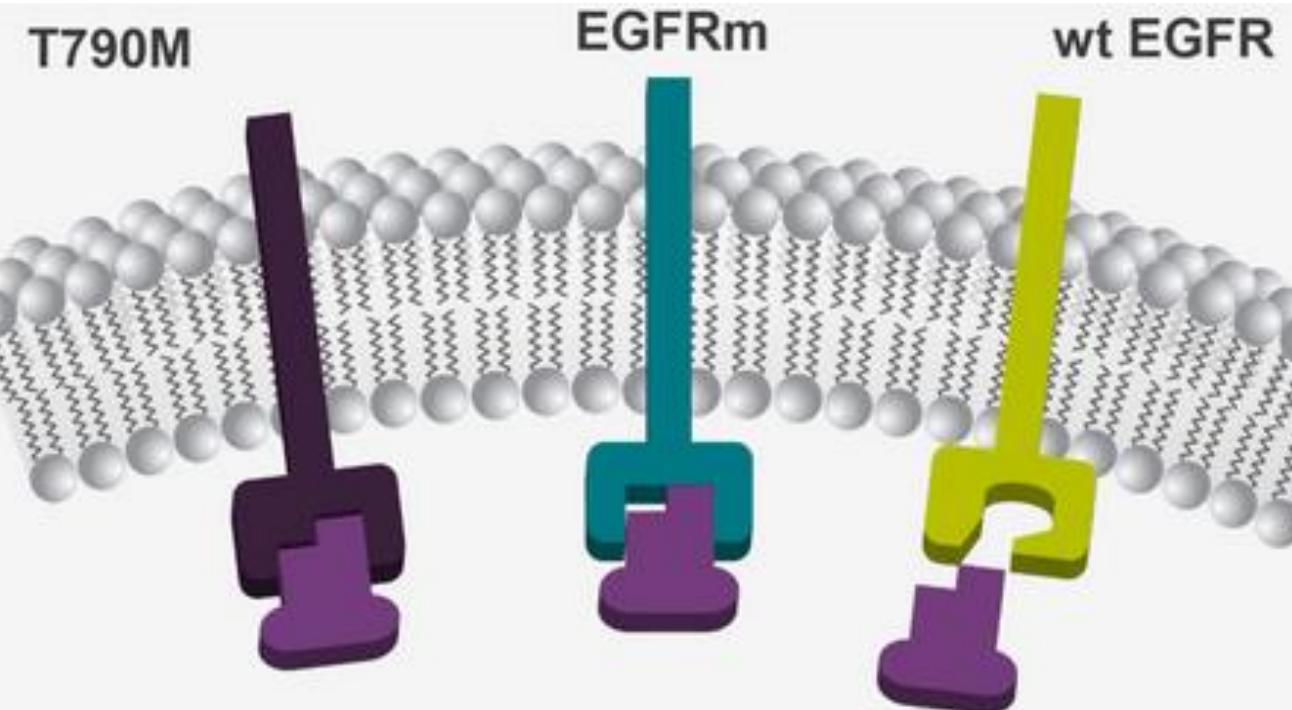
# Precision: Genomics Landscape of Colorectal Cancer

A



Strickler et al: Genomic landscape of cell-free DNA in patients with colorectal cancer

- Automation and instrumentation for decoding DNA (Next-Gen Sequencing)
- Early and frequent sequencing to monitor cancer (Liquid Biopsy)
- Single cell sequencing from tissue biopsy
- Tracking all related biomarkers (methylation, expression, protein etc)
- Connected to clinical phenotypes (symptoms) and outcome



**8 Hours**

from tissue isolation to  
sequencing test results

Inhibition of T790M

Inhibition of EGFRm

Low activity of wt EGFR



AstraZeneca

**TAGRISSO**



**2.5 Years**

from start of clinical trial to  
FDA approval (Nov 2015)



3730x DNA analyzer

### Automated DNA sequencer

- Capillary electrophoresis
- Costs reduced by 90%
- Human operation 15 min/day/machine
- 1 million bp/day



*Proc. Natl. Acad. Sci. USA*  
Vol. 96, pp. 9745–9750, August 1999  
Genetics

**Radiation hybrid mapping of the zebrafish genome**

NEIL A. HUKRIEDE<sup>1</sup>\*, LUCILLE JOLY<sup>†</sup>, MICHAEL TSANG<sup>‡</sup>, JENNIFER MILES<sup>†</sup>, PATRICIA TELLIS<sup>§</sup>,  
JONATHAN A. EPSTEIN<sup>¶</sup>, WILLIAM B. BARBAZU<sup>§</sup>, FRANK N. LI<sup>§</sup>, BARREI PAV<sup>¶</sup>, JOHN H. POSTLETHWAITE<sup>¶</sup>,  
THOMAS J. HUDSON<sup>¶</sup>, LEONARD I. ZON<sup>§</sup>, JOHN D. MCPHERSON<sup>§</sup>, MARIO CHEVRETT<sup>‡</sup>, IGOR B. DAVID<sup>\*</sup>,  
STEPHEN L. JOHNSON<sup>§</sup>, AND MARIE ECKER<sup>†,¶</sup>

<sup>1</sup>Laboratory of Molecular Genetics, Unit of Biological Conservation, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892; <sup>2</sup>Loeb-Harrington Research Center, Ottawa Hospital, Department of Medicine, University of Ottawa, Ottawa, Canada, K1Y 4E9; <sup>3</sup>Montreal General Hospital Research Institute and Department of Surgery, McGill University, Montreal, H3G 1AA; <sup>4</sup>Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110; <sup>5</sup>Howard Hughes Medical Institute and Department of Hematology, Children's Hospital, Boston, MA 02115; and <sup>6</sup>Institute of Neuroscience, University of Oregon, Eugene, OR 97403

Contributed by Igor B. David, June 14, 1999

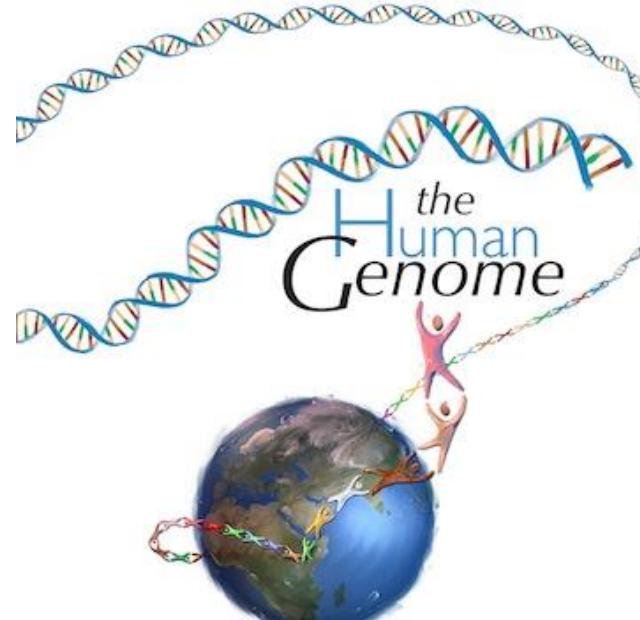
**ABSTRACT** The zebrafish is an excellent genetic system for the study of vertebrate development and disease. In an effort to provide a rapid and robust tool for zebrafish gene mapping, a panel of radiation hybrids (RH) was produced by fusion of irradiated zebrafish AB9 cells with mouse B78 cells. The overall retention of zebrafish sequences in the 93 RH cell lines was 88%. Comparison of the 93 RH panel to a subset of the LNS4 panel with 849 simple sequence length polymorphism markers, 84 cloned genes and 122 expressed sequence tags allowed the production of an RH map whose total size was 11,591 centiRays. From this value, we estimated the average translocation frequency of the LNS4 RH panel to correspond to 1 centilocus per 488 kilobases. Placement of a group of 235 unbiased markers on the RH map suggests that the map generated for the LNS4 panel at present, covers 88% of the zebrafish genome. Comparison of marker positions in RH and metacentric maps indicated a 96% concordance. Mapping expresses sequence tags and cloned genes by using the LNS4 panel should prove to be a valuable method for the identification of candidate genes for specific mutations in zebrafish.

Somatic-cell hybrids and radiation hybrids (RHs) have played a key role in the mapping of human and mouse genes (1–7). Cell hybrids provide one of the best expression systems for assigning genes to chromosomes or chromosome segments because mapping with cell hybrids does not require gene polymorphism. RHs are generated by irradiating cells from a donor species, causing random chromosomal breaks, and

We have previously shown that stable transfer of zebrafish chromosomes or chromosome segments to a rodent cell line was possible (17). Markers from the simple sequence-length polymorphism (SSLP) meiotic map could be anchored on a panel of zebrafish/mouse somatic-cell hybrids (14). Furthermore, Kwok *et al.* (18) demonstrated that RH technology could be used to generate a map of the zebrafish genome. Finally, we report characterization of LNS4, a zebrafish RH panel composed of 93 cell lines. We characterized the panel for 1,053 markers, including 84 genes and 122 ESTs, generating a map that we compared with a meiotic map by using a set of common markers.

**MATERIALS AND METHODS**

**Production of RHs.** We fused irradiated zebrafish fin AB9 cells to mouse B78 somatic cells. The B78 recipient cell line is not deficient in an enzyme that could be used to select for zebrafish chromosomal elements in hybrids. Therefore, zebrafish chromosomes were tagged with the aminoglycoside phosphotransferase gene that confers resistance to G418, as described (17). More than 400 independent G418-resistant AB9 clones were pooled for fusion experiments. Briefly,  $3 \times 10^7$  G418-resistant cells were irradiated with x-ray doses between 2,000 and 9,000 rad, mixed with an equal number of B78 cells, and fused in the presence of polyethylene glycol as described (17). One milliliter of 10% FBS was added to each dish. No colonies were observed in the controls (irradiated AB9 cells; unfused B78 cells; and irradiated AB9 and B78 cells



10 year & \$3 billion

# Speed: From Days to Hours

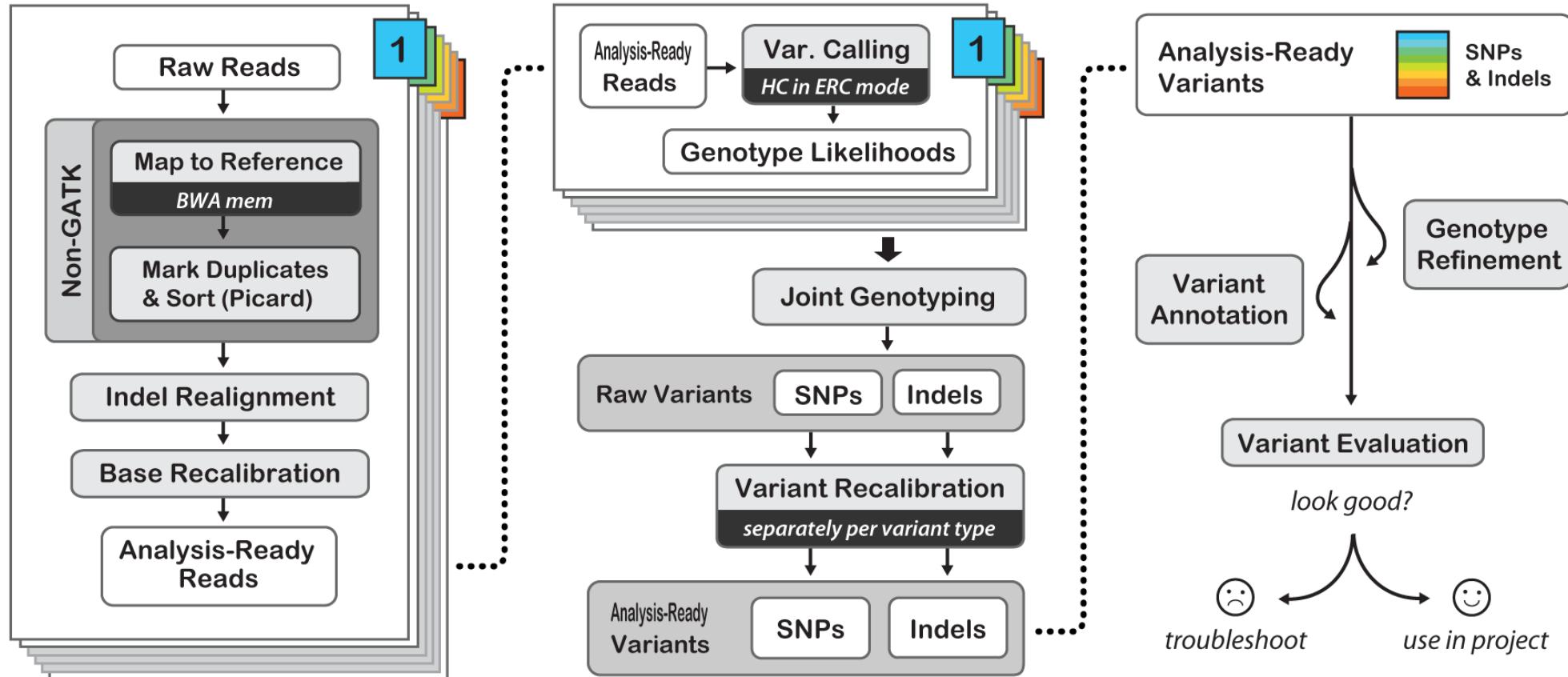
## Data Pre-processing

>>

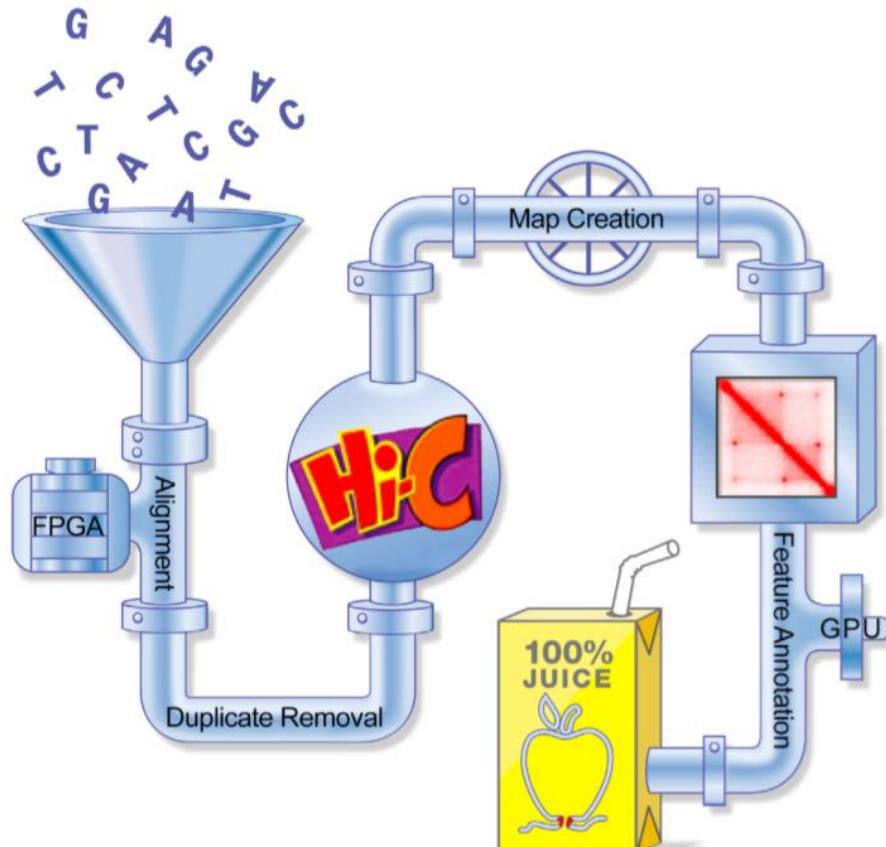
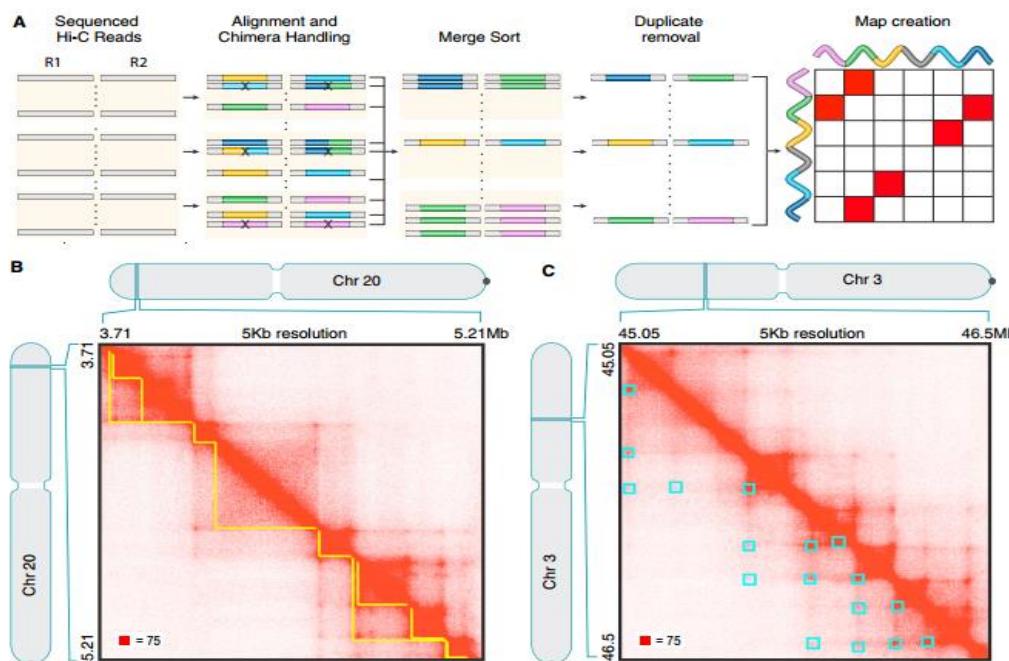
## Variant Discovery

>>

## Preliminary Analyses

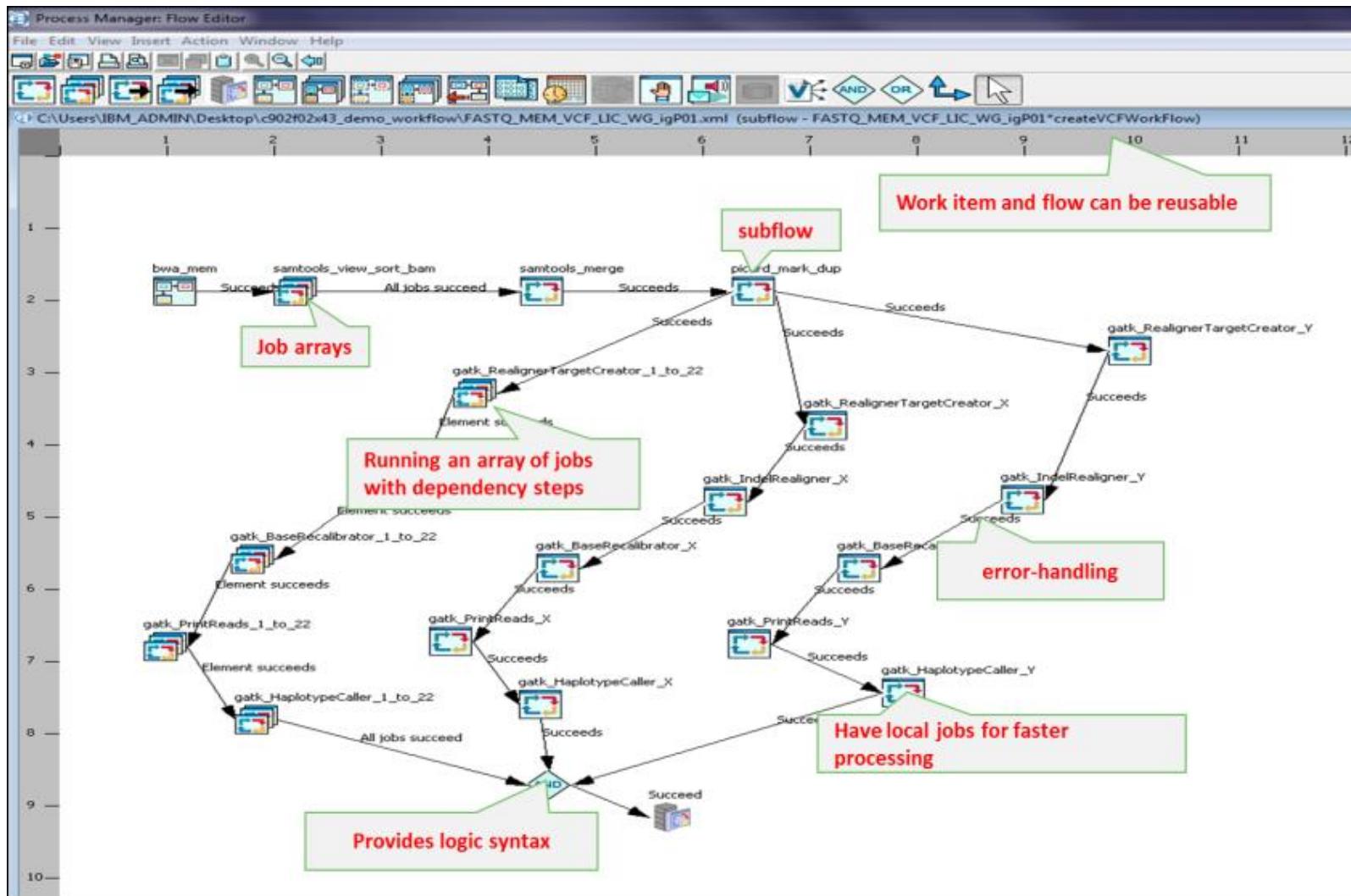


# Speed+: 3D Genomics for De Nova Assembly

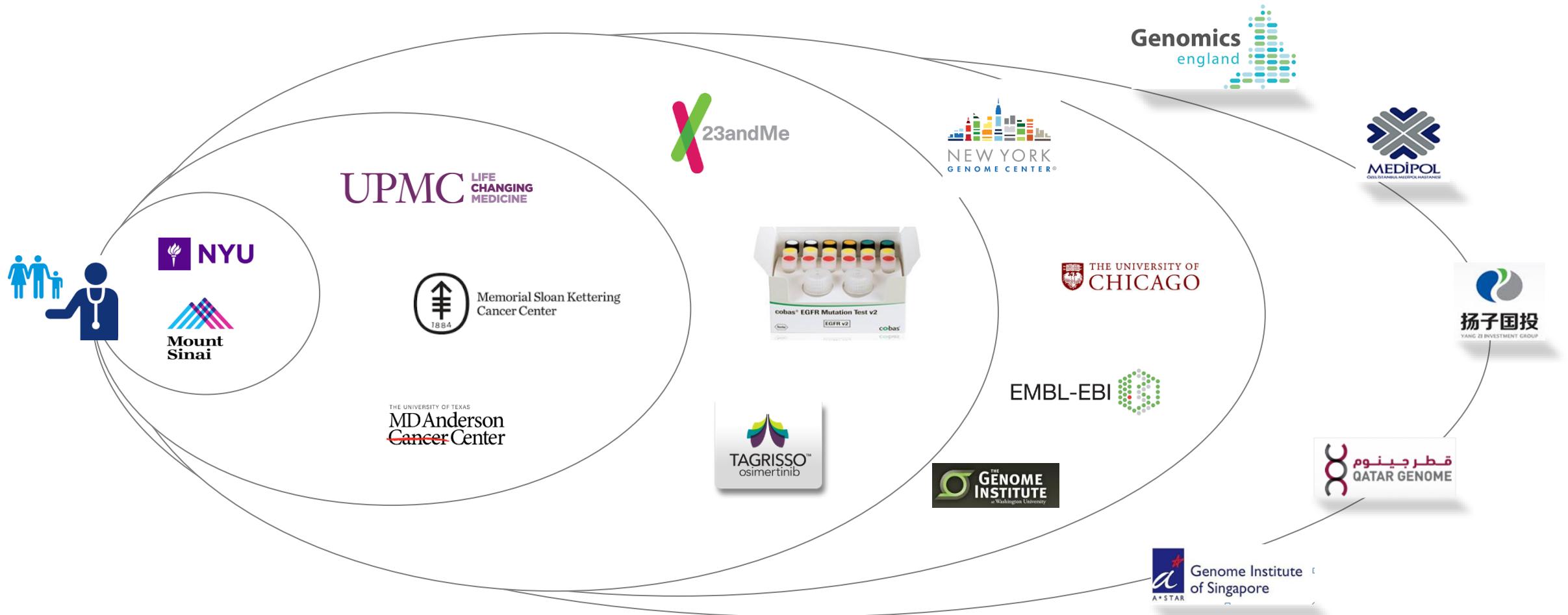


**REPORT**  
De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds  
Olu Pfeifer,<sup>1,2</sup> Sushil Rana,<sup>1,2</sup> Jason D. Ober,<sup>1,2</sup> Scott A. Prokop,<sup>1,2</sup> Michael A. Hickey,<sup>1,2</sup> Daniel L. Green,<sup>1,2</sup> and Brian L. Lovett,<sup>1,2</sup>  
<sup>1</sup> Ohio Department of Health, Division of Vector-Borne Diseases, Columbus, OH, USA;<sup>2</sup> Ohio State University, Columbus, OH, USA  
This is the authors' version of the work. Changes resulting from the peer review process may differ slightly from the published version. The final published version of the article may be available at <https://doi.org/10.1101/10.1101/2017.09.14.159402>. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a [aCC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

# Smart: Auto-drive Analytical Pipeline



# Scale: Institutional, Regional, Global



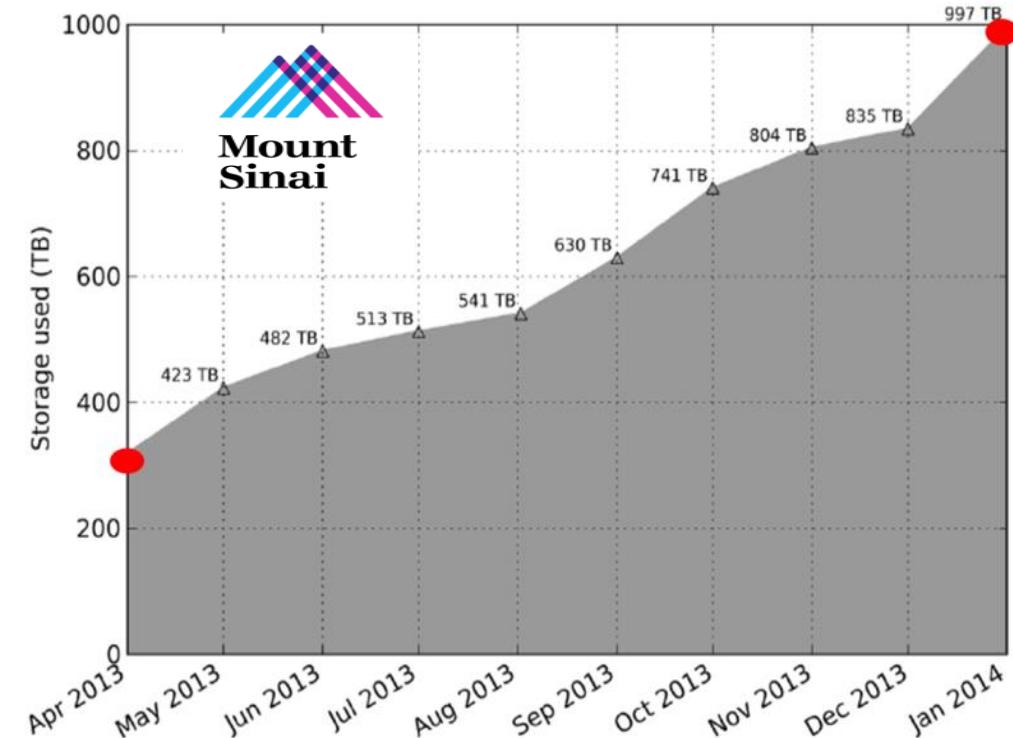
# Instrumenting Data: Challenges & Opportunities

# Processing Data Faster Than Instruments

*High Performance data landing and analysis*



- **Byte:** 1 Grain of Rice
- **Terabyte:** 2 Container Ships
- **Petabyte:** Blankets Manhattan
- **Exabyte:** Blankets US West Coast States
- **Zettabyte:** Fills Pacific Ocean
- **Yottabyte:** AN EARTH SIZE BALL OF RICE

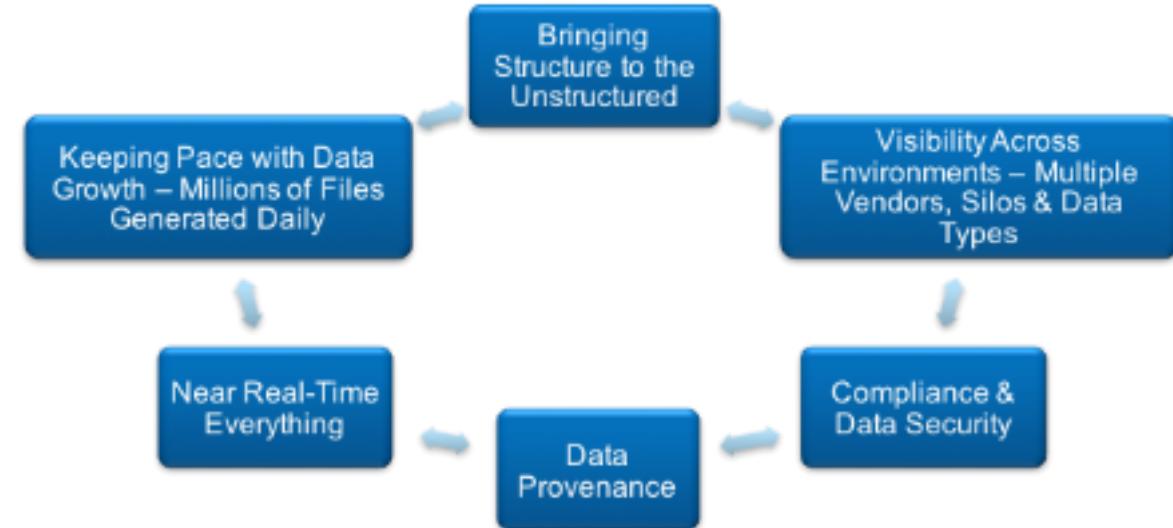


# Finding & Tracking Data

*Finding needles in haystacks, lots of them and growing faster ever*



(human API)

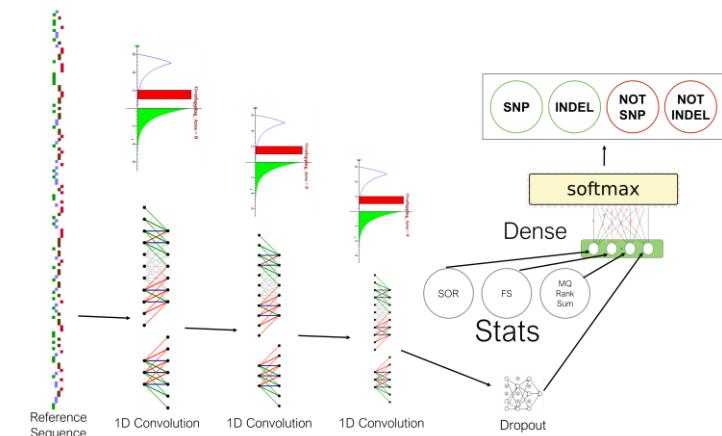
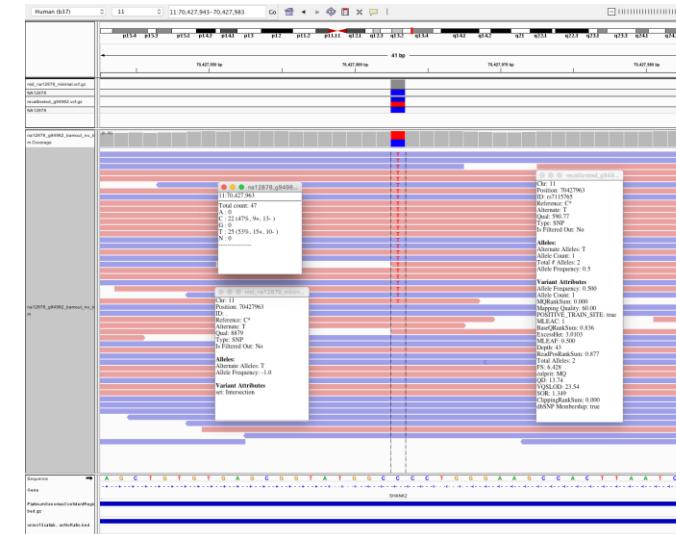


# Using My Data To Solve Your Problem

*Using Deep Learning to Train Model based on Bigger & Better Dataset*



*"I think you'll find that mine is bigger..."*



Training a CNN to filter genomic variants based on high-quality results, then apply it to low-coverage raw data

# Aggregating Our Data to Solve One Problem

*Combining datasets into large cohorts for analysis*



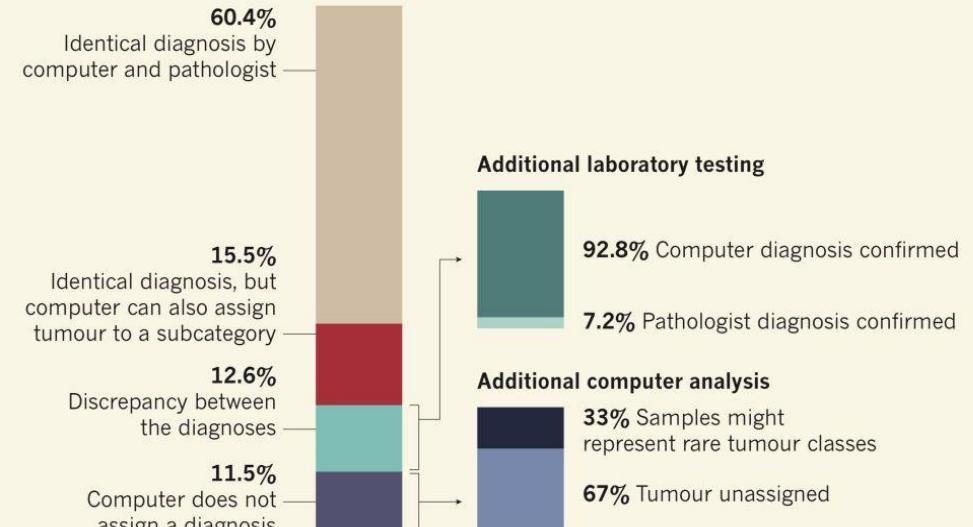
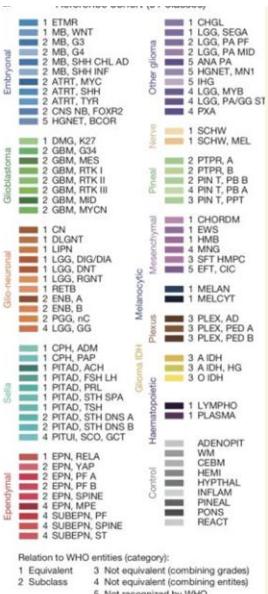
## ARTICLE

doi:10.1038/nature26000

### DNA methylation-based classification of central nervous system tumours

A list of authors and their affiliations appears in the online version of the paper.

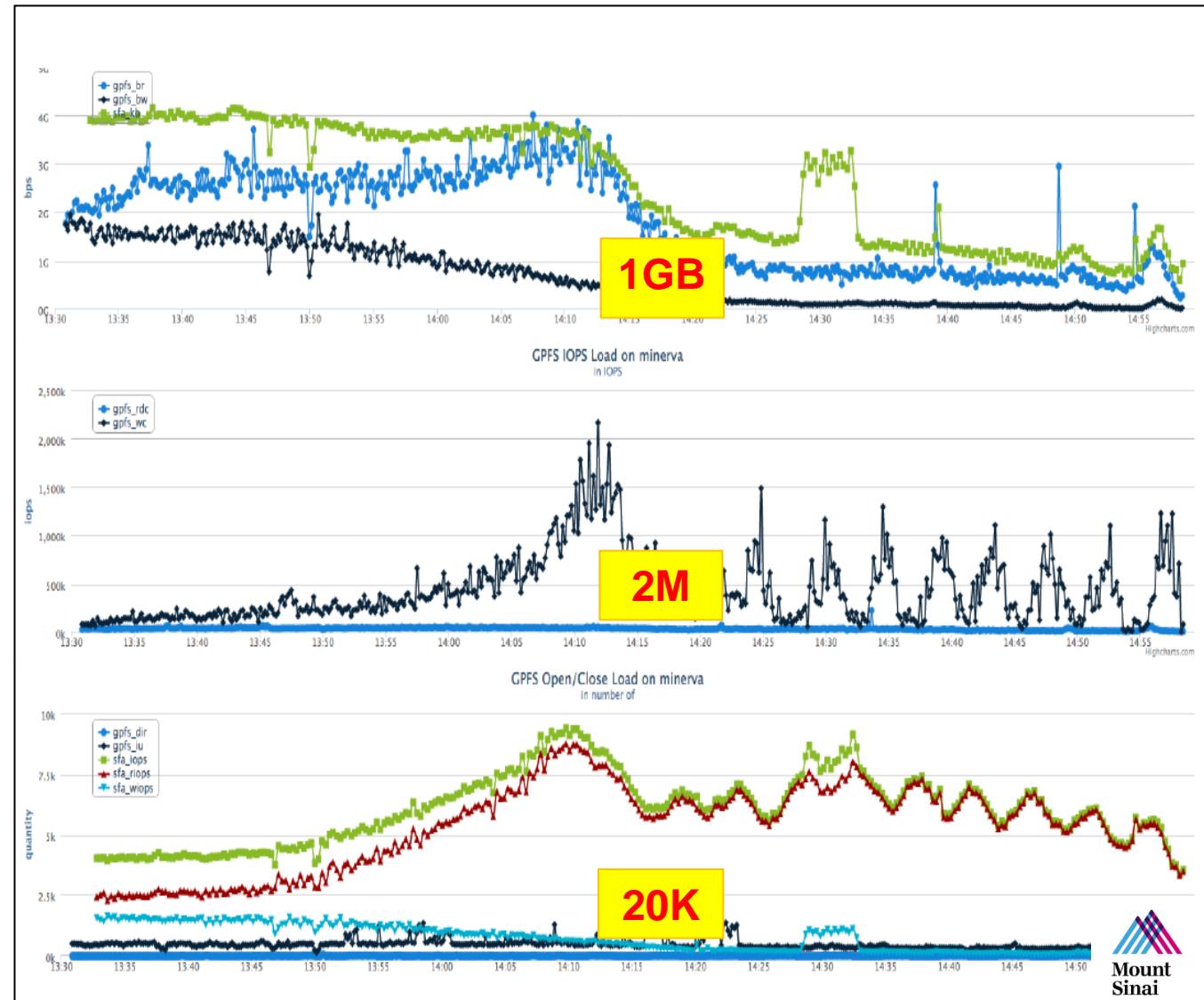
Accurate pathological diagnosis is crucial for optimal management of patients with cancer. For the approximately 100 known tumour types of the central nervous system, standardization of the diagnostic process has been shown to be particularly challenging—with substantial inter-observer variability in the histopathological diagnosis of many tumour types. Here we present a comprehensive approach for the DNA methylation-based classification of central nervous system tumours across all entities and age groups, and demonstrate its application in a routine diagnostic setting. We show that the availability of this method may have a substantial impact on diagnostic precision compared to standard methods, resulting in a change of diagnosis in up to 12% of prospective cases. For broader accessibility, we have designed a free online classifier tool, the use of which does not require any additional onsite data processing. Our results provide a blueprint for the generation of machine-learning-based tumour classifiers across other cancer entities, with the potential to fundamentally transform tumour pathology.



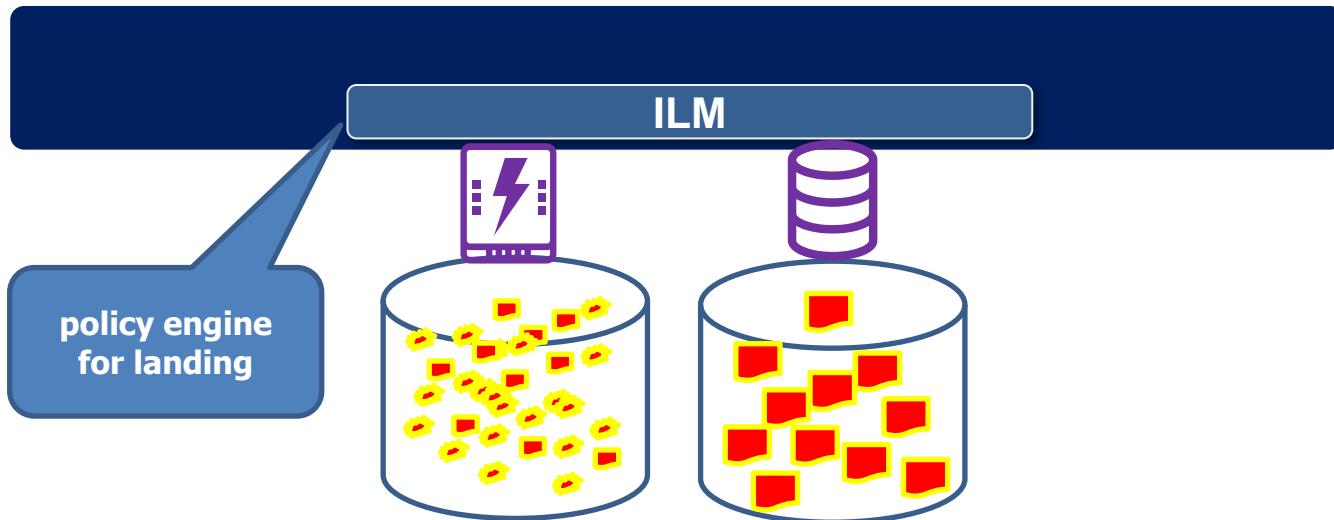
## How to Instrument Data?

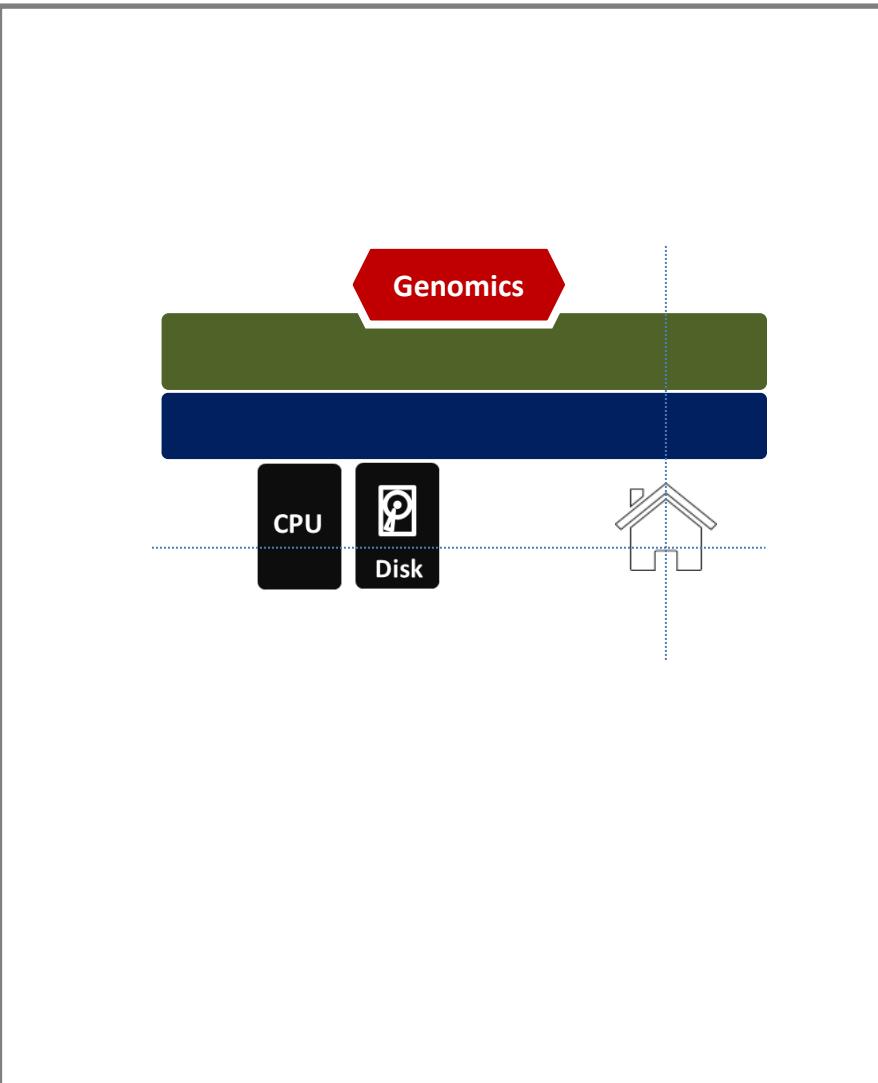
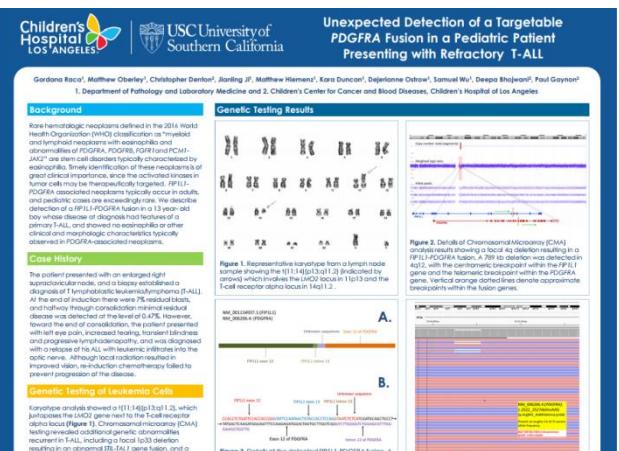
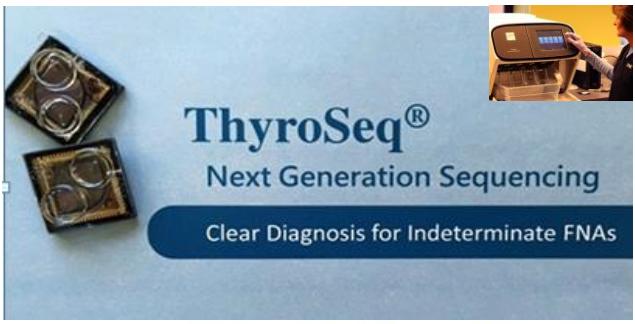
- **Fast Data Landing**
- Smart Data Tiering
- Flexible Data Accessing
- Global Data Peering
- Data Cataloguing

# High Speed & Throughput



# High Performance Data Lander





	Processing Stage	IBM Runtime
BWA	4.26	
Samtools	2.08	
MarkDuplicates	6.86	
RealignTargets	1.06	
IndelRealigner	1.06	
GATK BaseRecalibrator	1.49	
GATK PrintReads+Index	2.55	
GATK HaplotypeCaller	2.64	
Total	20.94	
	Total Runtime	9.85

**GATK 3.5 Best Practice Pipeline 50% Speedup**  
**GATK 3.8 Removing 1TB memory requirement**

Before	After
50 hours using 1 Node ~24cores, 1 QDR link, 256GB RAM	5 hours using 1 Node ~12cores, 1 FDR Link, 64GB RAM

## What

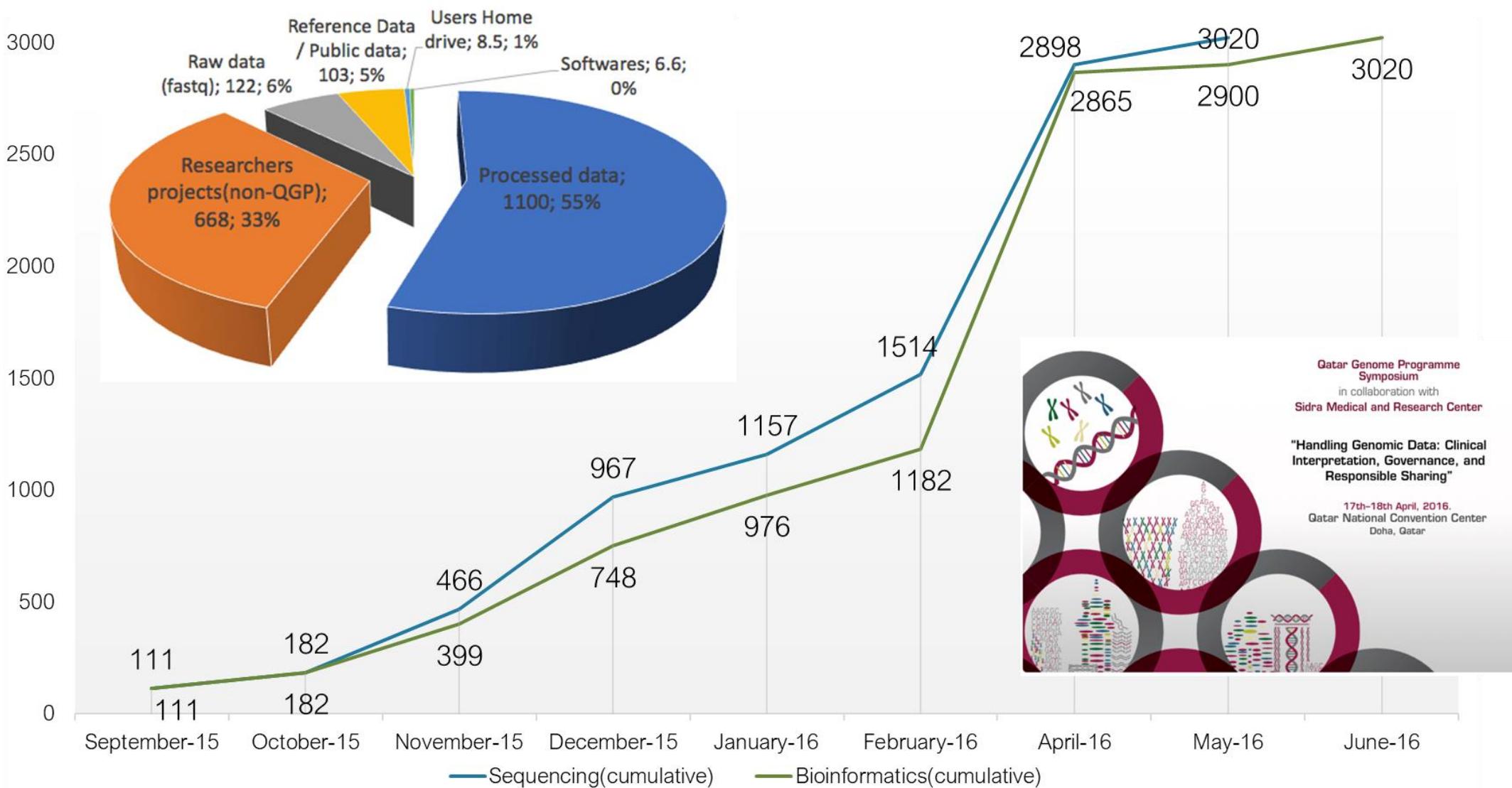
- File name
- File size
- Filer owner
- File path
- Filesystem name
- File set name
- File inode ID
- File permission
- File ctime
- File atime
- File mtime

## When

- Cluster name
- Global file ID
- Job submission user
- Job ID
- Job name
- Flow ID Job status
- Job start time
- Job finish time
- Job working directory
- Input files
- User variables

- **Fast Data Landing**
- **Smart Data Tiering**
- **Flexible Data Accessing**
- **Global Data Peering**
- **Data Cataloguing**

# Scale of Bigger Data

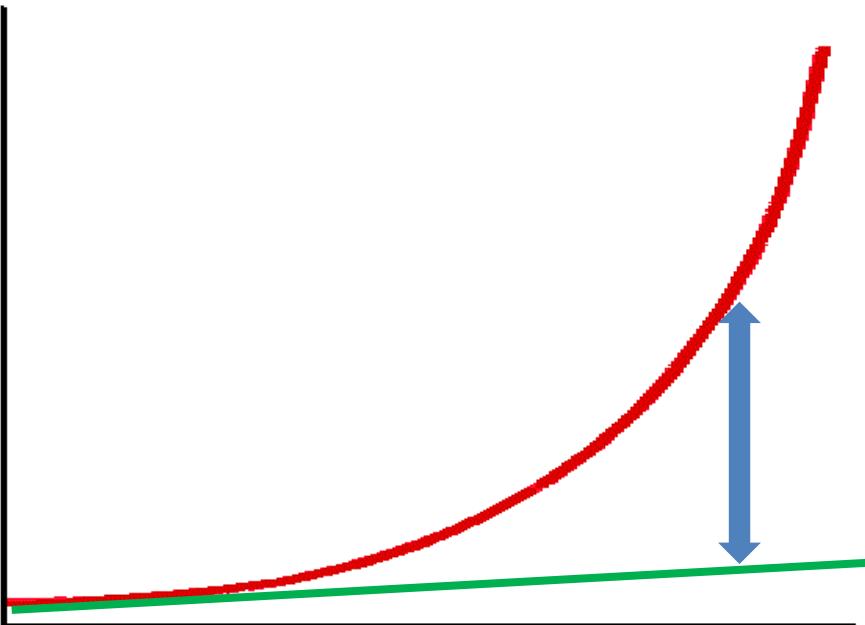


Qatar Genome Programme  
Symposium  
in collaboration with  
Sidra Medical and Research Center

"Handling Genomic Data: Clinical  
Interpretation, Governance, and  
Responsible Sharing"

17th-18th April, 2016.  
Qatar National Convention Center  
Doha, Qatar

# "Bending the Curve"

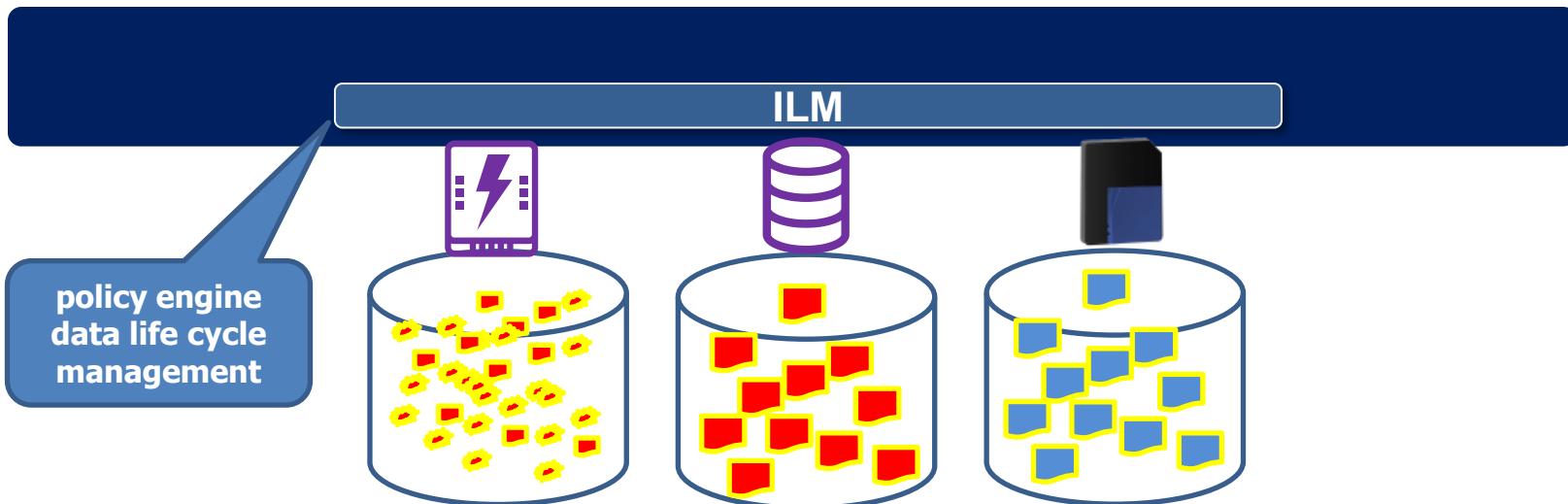


THE UNIVERSITY OF TEXAS  
MD Anderson  
Cancer Center

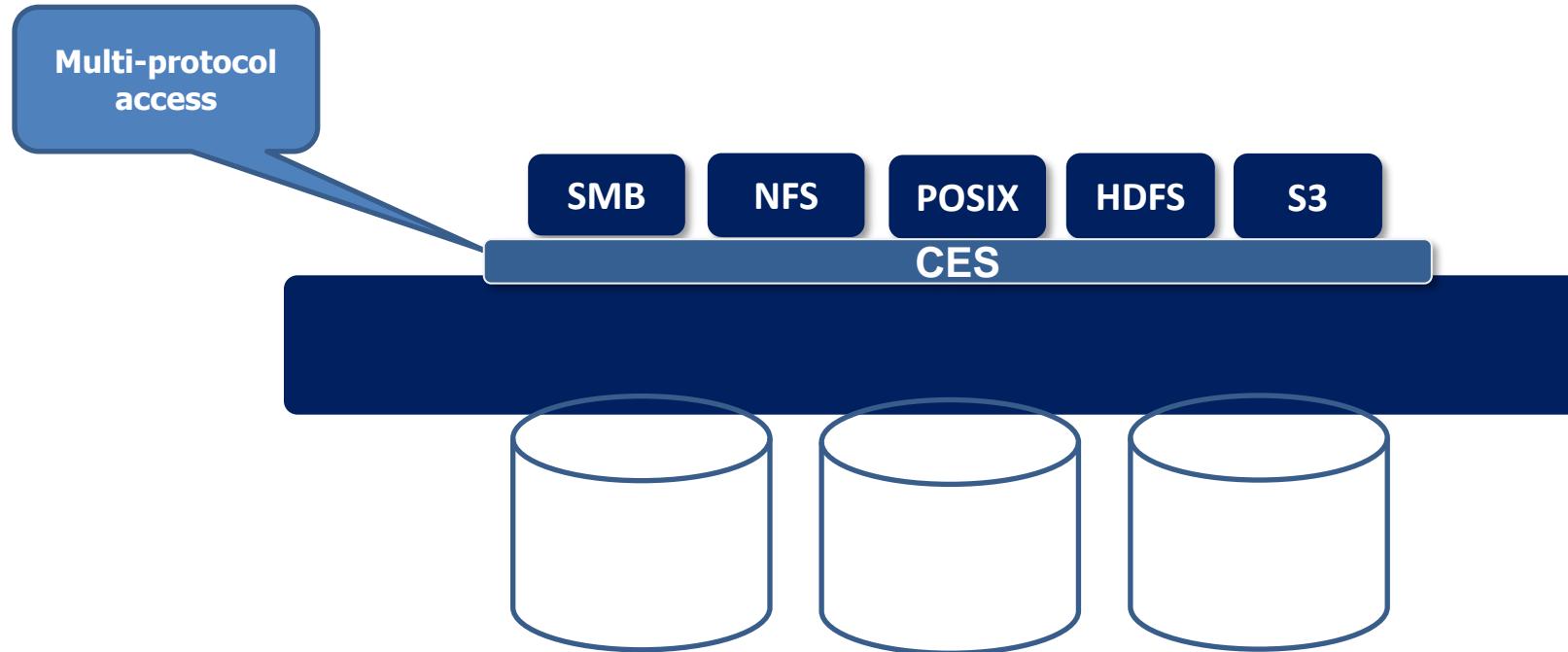
HIMSS18



# Peta-scale Data Management to Lower Cost



- Fast Data Landing
- Smart Data Tiering
- **Flexible Data Accessing**
- Global Data Peering
- Data Cataloguing





NFS

POSIX

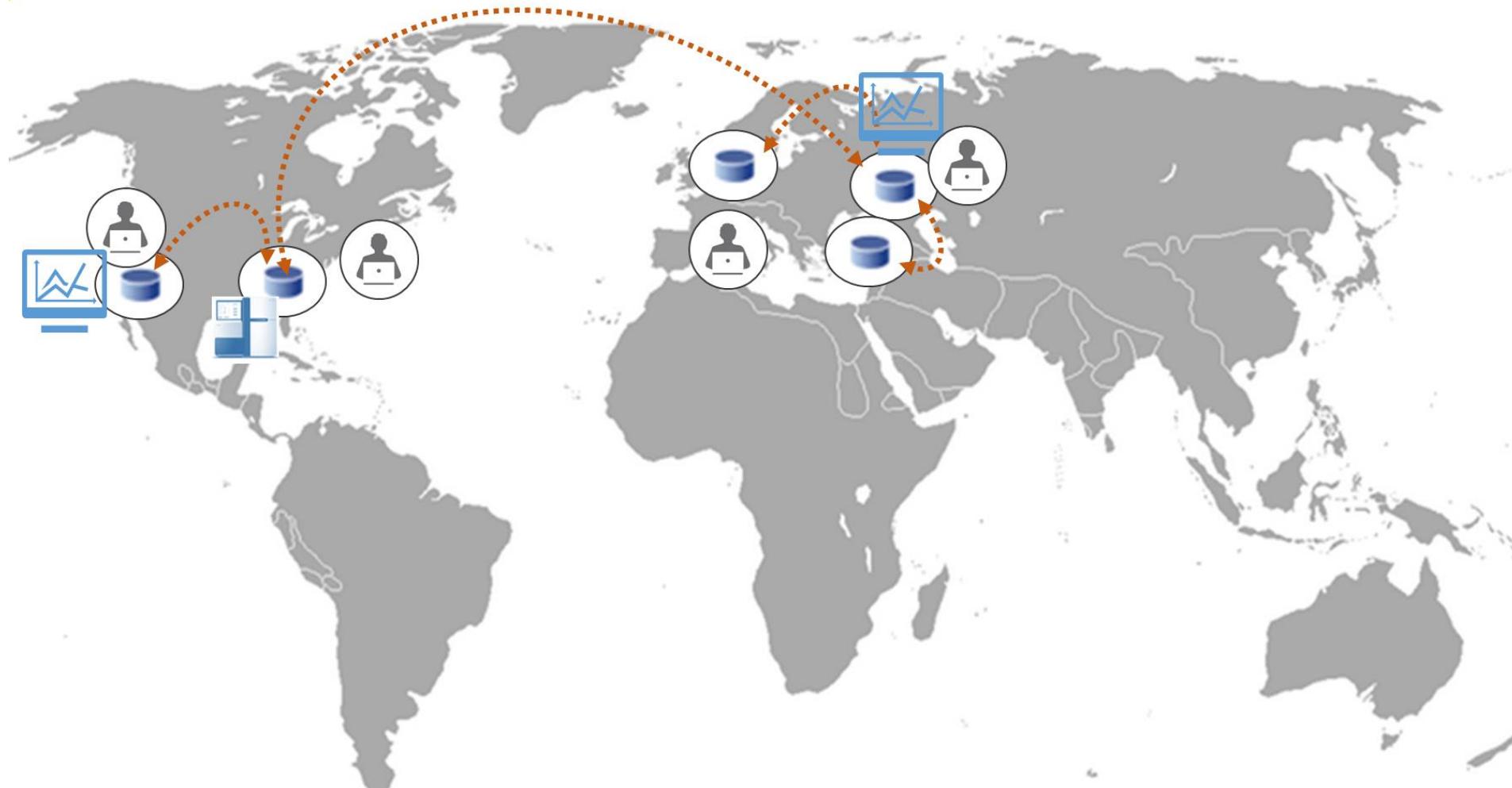


HDFS

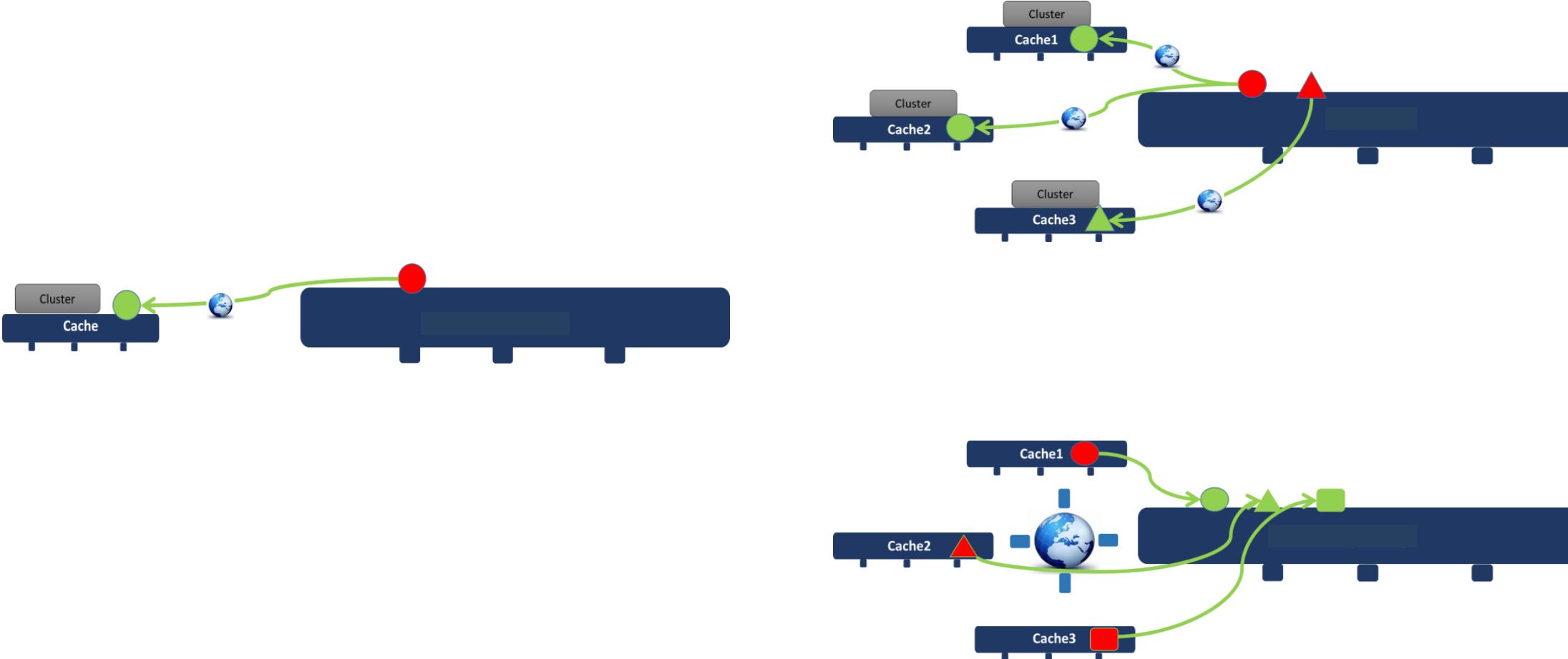


SMB

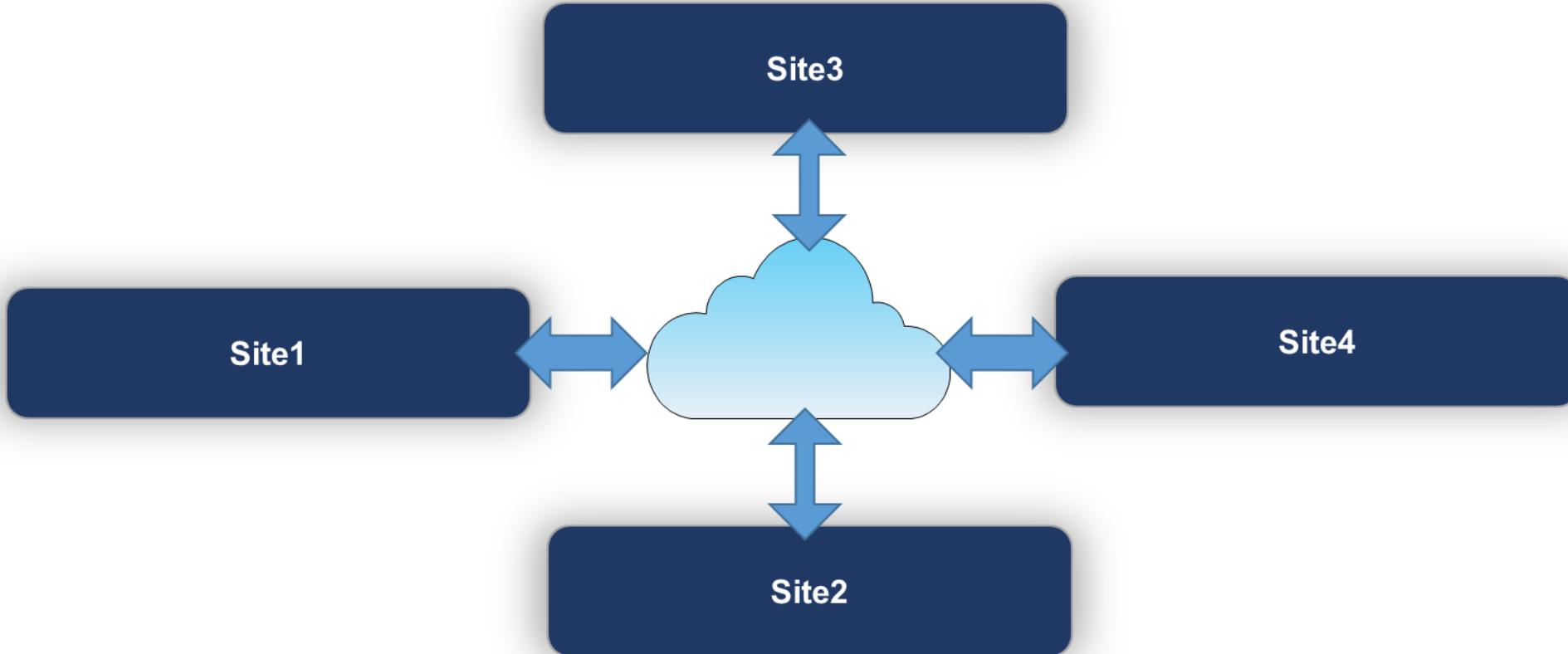
- Fast Data Landing
- Smart Data Tiering
- Flexible Data Accessing
- **Global Data Peering**
- Data Cataloguing



# Global Data Peering with Cache

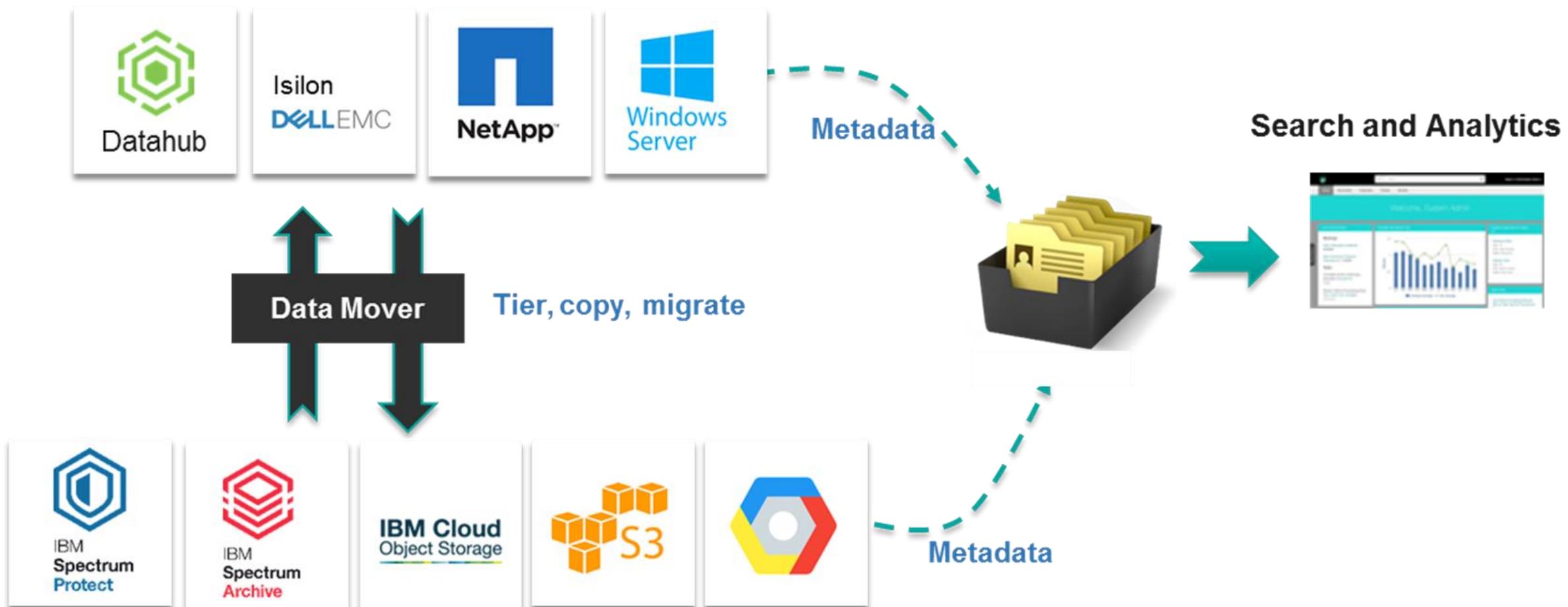


# Global Data Peering with Cloud

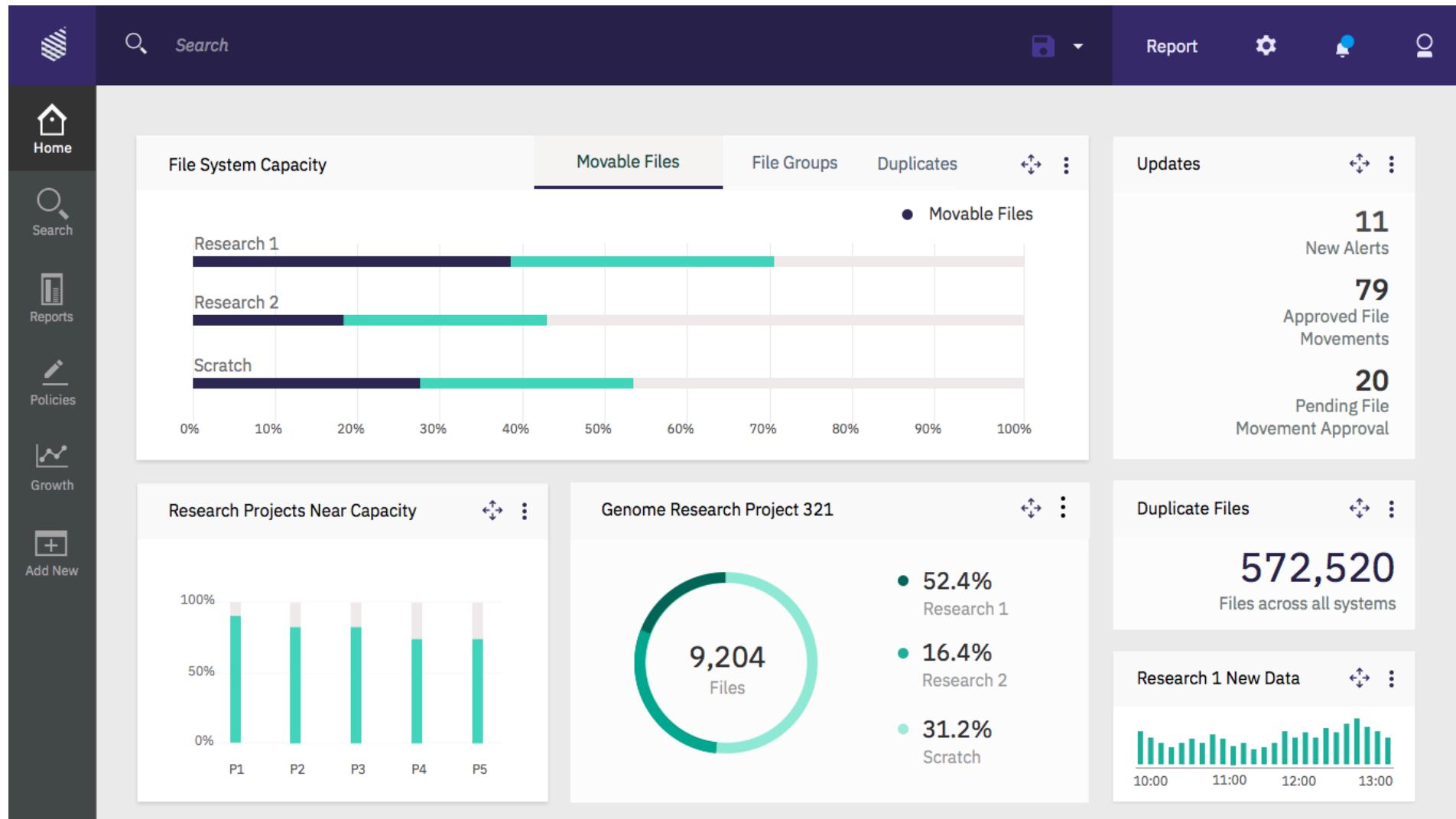


- **Fast Data Landing**
- **Smart Data Tiering**
- **Flexible Data Accessing**
- **Global Data Peering**
- **Data Cataloguing**

# Metadata Catalogue



# Data Dashboard



# Metadata for Dataset

Logical view of genomic variation dataset,  
data come from **different VCF files**.

Hundreds of millions of mutations, some meta data needed: **Variant annotation**

- Clinical info
- Consequence types
- Conservation scores
- Population frequencies
- ...

**Genomics England** project:

- **200M variants x 100K samples**, about **20 trillion** points
- With different layers of data, about **80-100 trillion** points
- a lot of meta data for variants and samples
- about **400TB** to be indexed

		Samples						
		var_1	28	32	29	28	35	32
var_1	16,12	16,17	18,14	12,14	16,14	16,12		
var_2	C/C	C/G	C/C	C/G	C/C	G/G		
"	"	"	"	"	"	"		
"	"	"	"	"	"	"		
"	"	"	"	"	"	"		
"	"	"	"	"	"	"		
"	"	"	"	"	"	"		
"	"	"	"	"	"	"		
"	"	"	"	"	"	"		
"	"	"	"	"	"	"		
"	"	"	"	"	"	"		
var_n	"	"	"	"	"	"		

**Different layers** of information:

- Genotype for samples
- Allele counts
- Quality scores
- Phase information
- ...

Meta data: **Sample annotation**

- Phenotype
- Family and population pedigree
- Clinical variables
- ...

**Heterogeneous data analysis and algorithms**, different technologies and solutions required:

- Search and filter using data and meta data
- Data mining, correlation
- Statistic tests
- Machine learning
- Interactive analysis
- Network-based analysis
- Visualization
- Encryption
- ...

**Applications:**

- Personalized medicine
- Trait association
- ...

# Genomic Database

Search

Clear No filters selected

Study i

Studies Filter

In (AND) ▾

- 1kG\_phase3
- ESP6500
- EXAC
- 1kG\_phase3\_chrY
- 1kG\_phase3\_chrMT
- GONL
- UK10K\_ALSPAC
- UK10K\_TWINSUK
- MGP

Genomic i

Chromosomal Location

3:444-55555, 1:1-100000

Feature IDs (gene, transcript, SNP, ...)

Search for Gene Symt +

BRCA2, ENSG000000100010

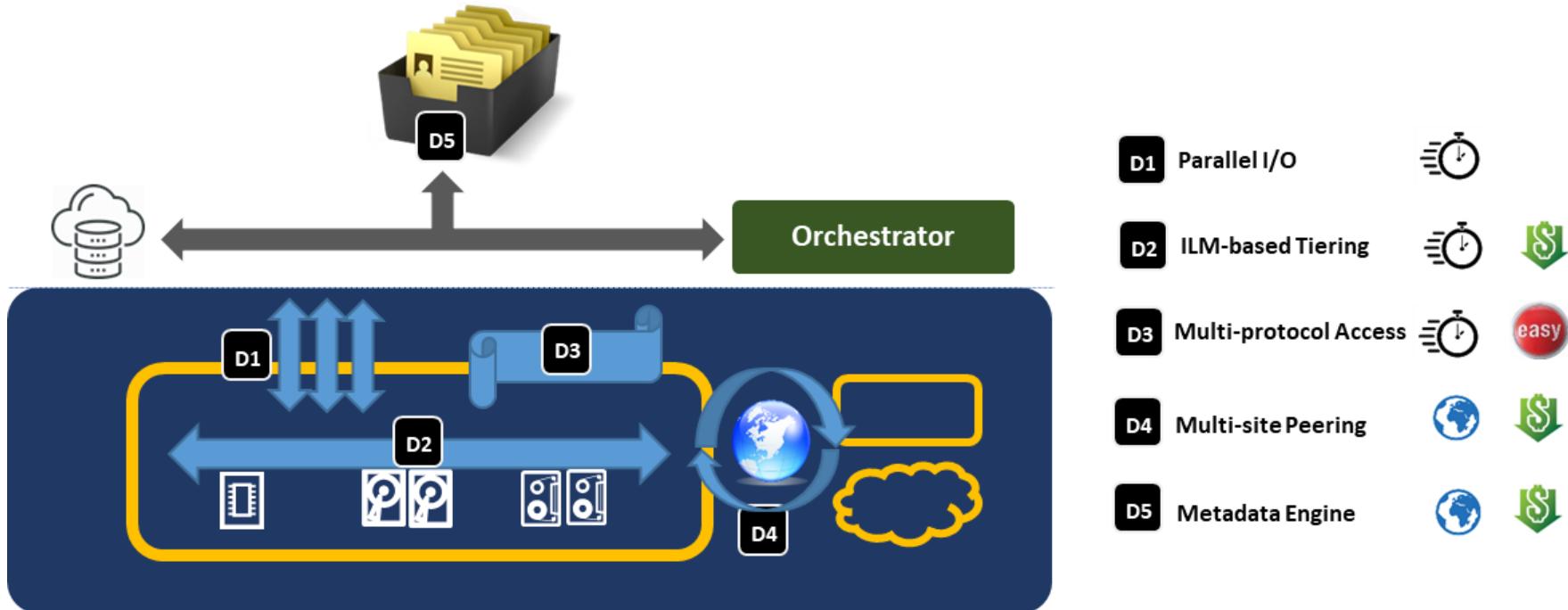
Showing 1-10 of -1 variants

Table Result Summary (Beta) Genome Browser (Beta)

Variant	SNP Id	Genes	Type	Consequence Type	Deleteriousness i			Conservation i			Population Frequencies i		
					SIFT	Polyphen	CADD	PhyloP	PhastCons	GERP	1000 Genomes	ExAC	ESP6500
13:19323853 G/A	rs112913900		SNV	intergenic_variant	-	-	1.52	0.154	0.140	0.000			
13:19323907 C/T	rs149770383		SNV	intergenic_variant	-	-	1.18	0.154	0.016	0.000			
13:19323918 G/A	rs375201401		SNV	intergenic_variant	-	-	0.97	-1.342	0.005	0.000			
13:19323936 C/T	rs372351635		SNV	intergenic_variant	-	-	1.75	0.154	0.096	0.000			
13:19323983 G/A	rs75074214		SNV	intergenic_variant	-	-	1.57	0.154	0.021	0.000			
13:19747991 C/T	rs3764135	SMPD4P2,TUBA3C	SNV	3_prime_UTR_variant	-	-	7.40	0.194	0.166	0.728			
13:19748001 C/T	rs201712054	SMPD4P2,TUBA3C	SNV	3_prime_UTR_variant	-	-	9.21	0.194	0.312	-2.440			
13:19748024 A/T	rs36216910	SMPD4P2,TUBA3C	SNV	synonymous_variant	-	-	0.05	0.163	0.581	-1.960			
13:19748031 G/C	rs139316426	SMPD4P2,TUBA3C	SNV	missense_variant	tolerated	-	9.67	0.194	0.576	1.220			
13:19748038 C/T	rs1803092	SMPD4P2,TUBA3C	SNV	missense_variant	tolerated	-	19.85	0.194	0.652	1.220			

http://hgva.opencb.org

# Architecture of Instrumented Data



## **Instrumented Data: Early Applications**

# Metadata-driven Information Sharing

Search all beacons for allele

GRCh37 ▾

13 : 32936732 G > C

Search

Response	All	None
<input checked="" type="checkbox"/> Found	10	
<input type="checkbox"/> Not Found	33	
<input type="checkbox"/> Not Applicable	25	

Organization	All	None
<input checked="" type="checkbox"/> AMPLab, UC Berkeley		
<input checked="" type="checkbox"/> Australian Genomics He...		
<input checked="" type="checkbox"/> Belgian Medical Genomi...		
<input checked="" type="checkbox"/> BGI		
<input checked="" type="checkbox"/> Bioinformatics Area, Fun...		
<input checked="" type="checkbox"/> BioReference Laboratori...		
<input checked="" type="checkbox"/> Brazilian Initiative on Pre...		
<input checked="" type="checkbox"/> BRCA Exchange		
<input checked="" type="checkbox"/> Broad Institute		
<input checked="" type="checkbox"/> Centre for Genomic Reg...		
<input checked="" type="checkbox"/> Centro Nacional de Anal...		
<input checked="" type="checkbox"/> Children's Mercy Hospital		
<input checked="" type="checkbox"/> Curoverse		
<input checked="" type="checkbox"/> DNAstack		
<input checked="" type="checkbox"/> ELIXIR		
<input checked="" type="checkbox"/> EMBL European BioInfor...		

- |   |  |               |       |
|---|--|---------------|-------|
|    | <b>BRCA Exchange</b>                           | Show Metadata | Found |
|   | Hosted by BRCA Exchange                        |               |       |
|    | <b>Cafe Variome</b>                            |               | Found |
|   | Hosted by University of Leicester              |               |       |
|    | <b>Cafe Variome Central</b>                    |               | Found |
|   | Hosted by University of Leicester              |               |       |
|    | <b>HGMD Public</b>                             |               | Found |
|   | Hosted by University of California, Santa Cruz |               |       |
|  | <b>Kaviar</b>                                  |               | Found |
|   | Hosted by Institute for Systems Biology        |               |       |
|  | <b>Leiden Open Variation</b>                   |               | Found |
|   | Hosted by University of California, Santa Cruz |               |       |

<https://beacon-network.org>

United States Patent

9,354,922

Lee

May 31, 2016

Metadata-driven workflows and integration with genomic data processing systems and techniques

### Abstract

Systems, methods and computer program products configured to provide and perform metadata-based workflow management are disclosed. The inventive subject matter includes a computer readable storage medium having computer readable program instructions embodied therewith. The computer readable program instructions are configured to: initiate a workflow configured to process data; associate the data with metadata; and drive at least a portion of the workflow based on at least some of the metadata. The metadata include anchoring metadata; common metadata; and custom metadata. Inventive subject matter also encompasses a method for managing genomic data processing workflows using metadata includes: initiating a workflow; receiving a request to manage the workflow using metadata comprising: anchoring metadata, common metadata, and custom metadata, associating the metadata with the data; and driving at least a portion of the workflow based on the metadata. The workflow involves genomic analyzes.

Inventors: Lee; Frank N. (Sunset Hills, MO)

Applicant: Name City State Country Type

International Business Machines Corporation Armonk NY US

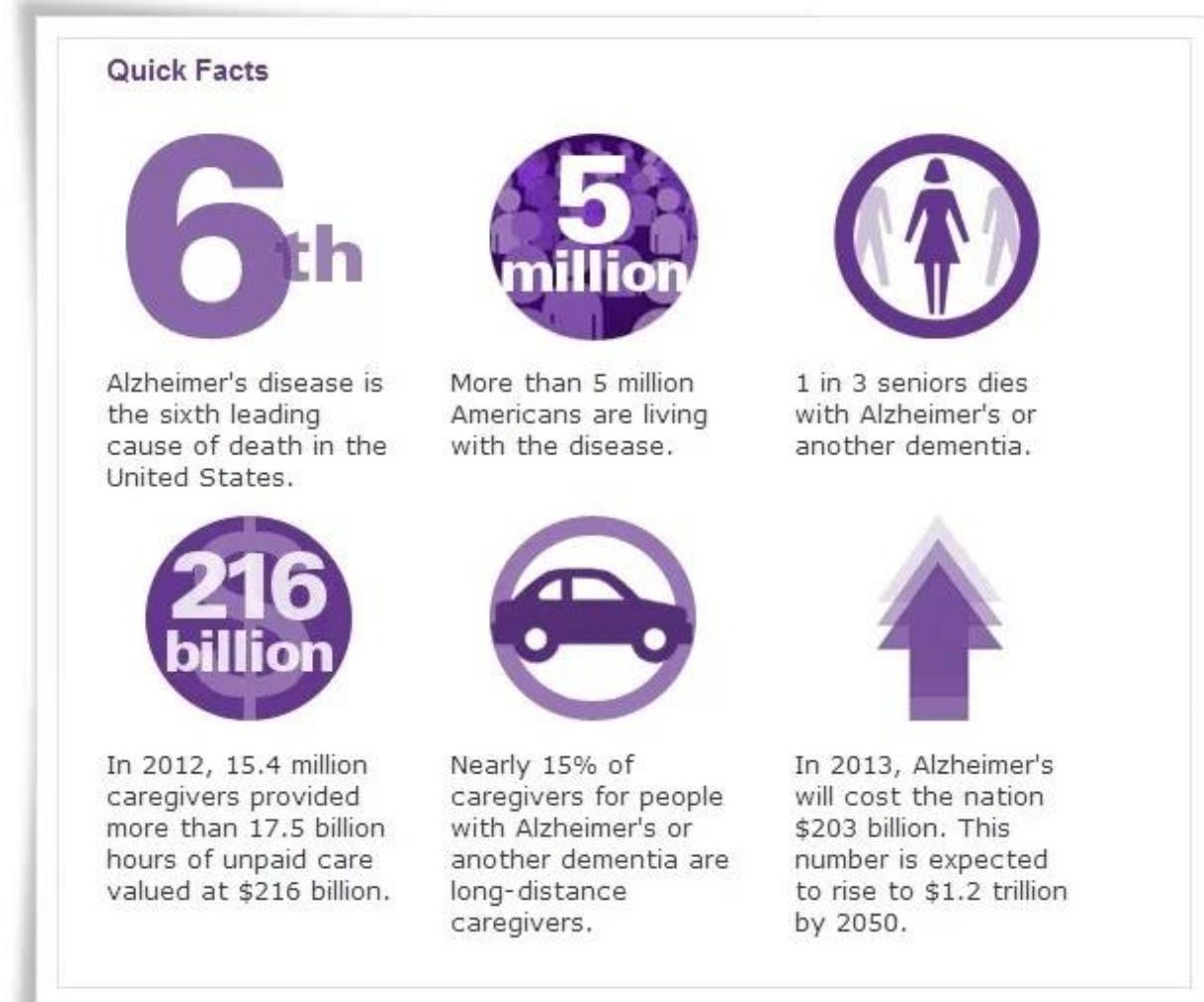
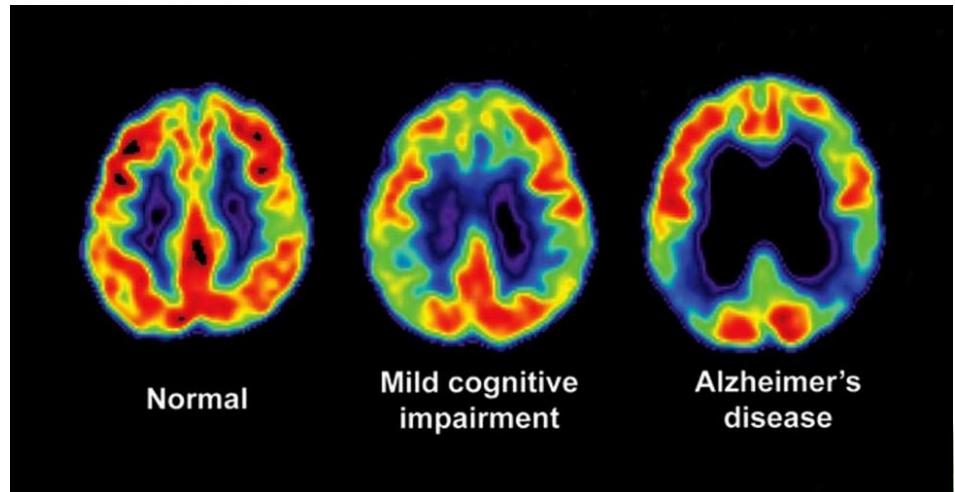
Assignee: International Business Machines Corporation (Armonk, NY)

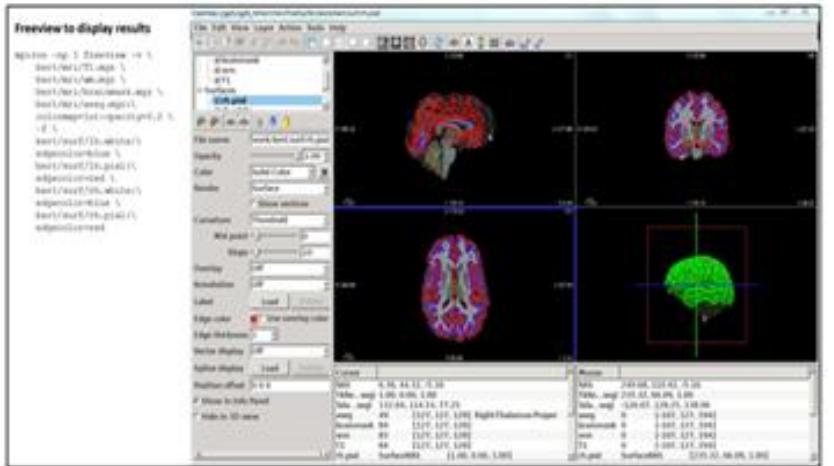
Family ID: 1000001877202

Appl. No.: 14/243,301

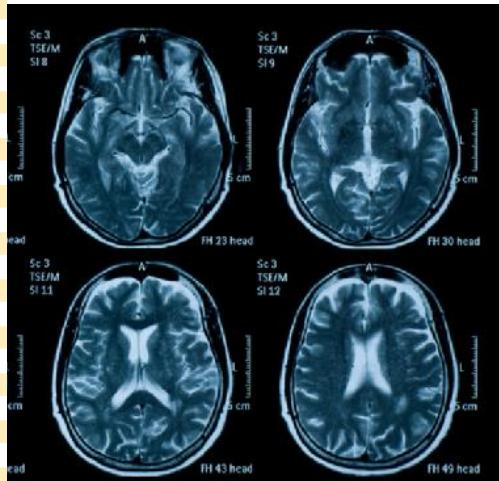
Filed: April 2, 2014

# **Bouncing Back**

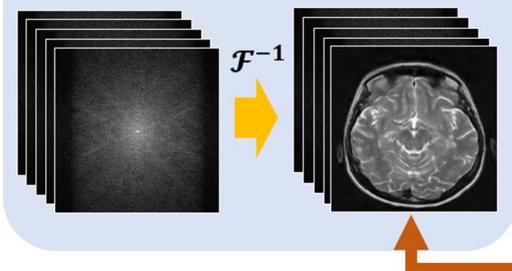




0020,9518	AcquisitionIndex
0020,9529	ContributingSOPInstancesRefSeq
0020,9536	ReconstructionIndex
0021,1003	SeriesFromWhichPrescribed
0021,1005	GenesisVersionNow
0021,1007	SeriesRecordChecksum
0021,1018	GenesisVersionNow
0021,1019	AgreconRecordChecksum
0021,1020	TableStartLocation
0021,1035	SeriesFromWhichPrescribed
0021,1036	ImageFromWhichPrescribed
0021,1037	ScreenFormat
0021,104A	AnatomicalReferenceForScout
0021,104F	LocationsInAcquisition
0021,1050	GraphicallyPrescribed
0021,1051	RotationFromSourceXRot
0021,1052	RotationFromSourceYRot
0021,1053	RotationFromSourceZRot
0021,1054	ImagePosition
0021,1055	ImageOrientation
0021,1056	IntegerSlop

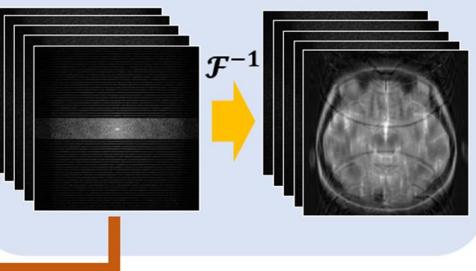


**Full-sampling**  
**scan time: 30 min**



$\mathcal{F}^{-1}$

**Undersampling**  
**scan time: 10 min**



The goal of undersampled MRI is to develop this map.

