

Bigger and Better Data

Lessons from Frontlines of Precision Medicine

Frank Lee PhD
IBM Global Industry Leader for Systems Group

5th Annual Big Data & Business Analytics Symposium – March 22-23, 2018

BIG DATA
& BUSINESS ANALYTICS GROUP
bigdata.wayne.edu



WAYNE STATE
UNIVERSITY

Getting Your Transformation Right

SYMPOSIUM

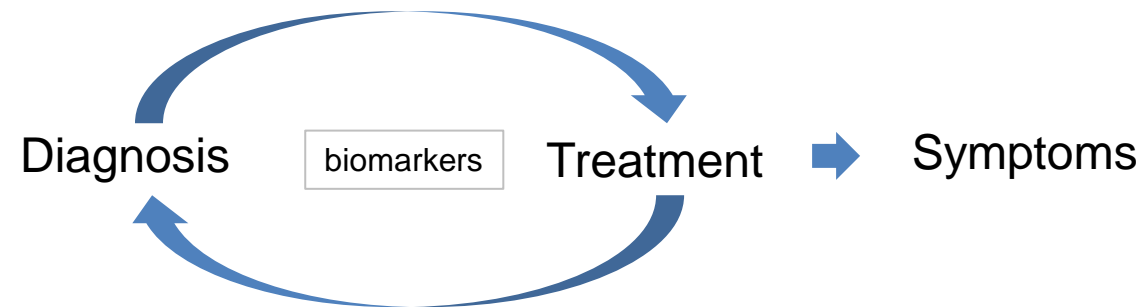
March 22-23, 2018

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

1st Symptom → 1st Diagnosis → 1st Treatment

2nd Symptoms → 2nd Diagnosis → 2nd Treatment

3rd Symptoms → 3rd Diagnosis → 3rd Treatment

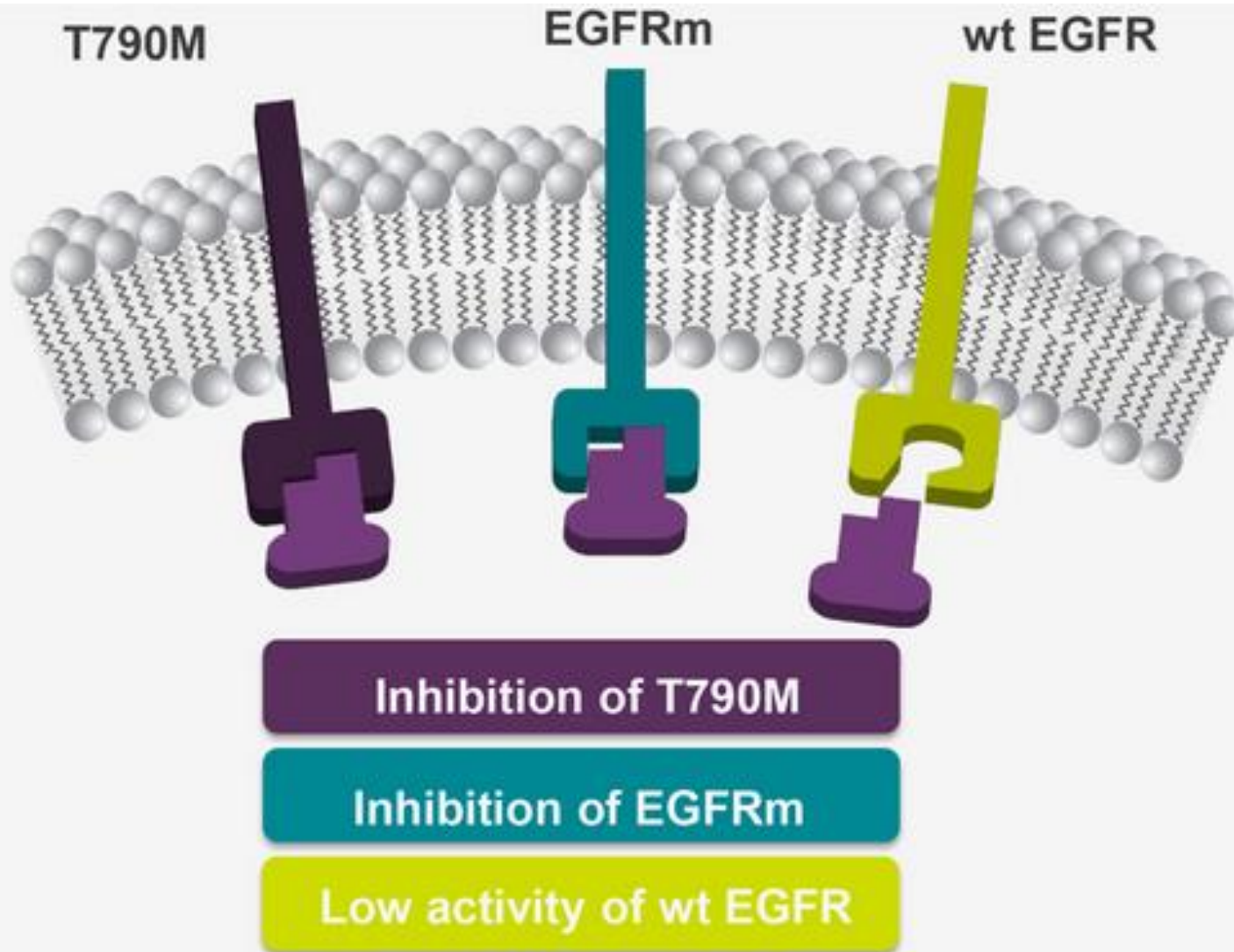


- Automation and instrumentation for decoding DNA (Next-Gen Sequencing)
- Early and frequent sequencing to monitor cancer (Liquid Biopsy)
- Single cell sequencing from tissue biopsy
- Tracking all related biomarkers (methylation, expression, protein etc)
- Connected to clinical phenotypes (symptoms)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



8 Hours

from tissue isolation to
sequencing test results



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*, LUCILLE JOLY¹, MICHAEL TSANG*, JENNIFER MILES¹, PATRICIA TELLS¹,
JONATHAN A. EPSTEIN*, WILLIAM B. BARBAZUK², FRANK N. LI³, BARRY PAW⁴, JOHN H. POSTLETHWAIT⁵,
THOMAS J. HUDSON⁶, LEONARD I. ZON⁷, JOHN D. MCPHERSON⁸, MARIO CHEVRETTE⁹, IGOR B. DAWID*,
STEPHEN L. JOHNSON¹⁰, AND MARC ECKER^{11*}

*Laboratory of Molecular Genetics and Unit on Biological Computation, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892; ¹Loeb Health Research at the Ottawa Hospital, Department of Medicine, University of Ottawa, Ottawa, Canada, K1Y 4K9; ²Montreal General Hospital Research Institute and Department of Surgery, McGill University, Montreal, Canada, H3G 1A6; ³Department of Genetics, Washington University Medical School, St. Louis, MO 63110; ⁴Vassar Hughes Medical Institute and Department of Hematology, Children's Hospital, Boston, MA 02115; and ⁵Institute of Neuroscience, University of Oregon, Eugene, OR 97403

Contributed by Igor B. Dawid, June 14, 1999

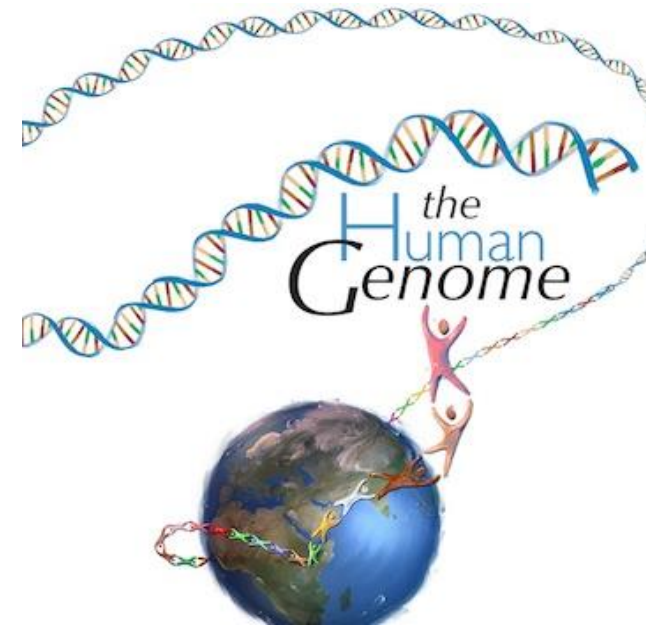
ABSTRACT The zebrafish is an excellent genetic system for the study of vertebrate development and disease. In an effort to provide a rapid and robust tool for zebrafish gene mapping, a panel of radiation hybrids (RH) was produced by fusion of irradiated zebrafish AB9 cells with mouse B78 cells. The overall retention of zebrafish sequences in the 93 RH cell lines that constitute the LNS4 panel is 22%. Characterization 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 11,501 centiRays. From this value, we estimated the average breakpoint frequency of the LNS4 RH panel to correspond to 1 centiRay = 148 kilobase. 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 meiotic maps indicated a 96% concordance. Mapping expressed 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 constitute one of the most expedient methods 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, Kwak *et al.* (18) demonstrated that RH technology could be used for nonmammalian vertebrates. In the present study, 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.

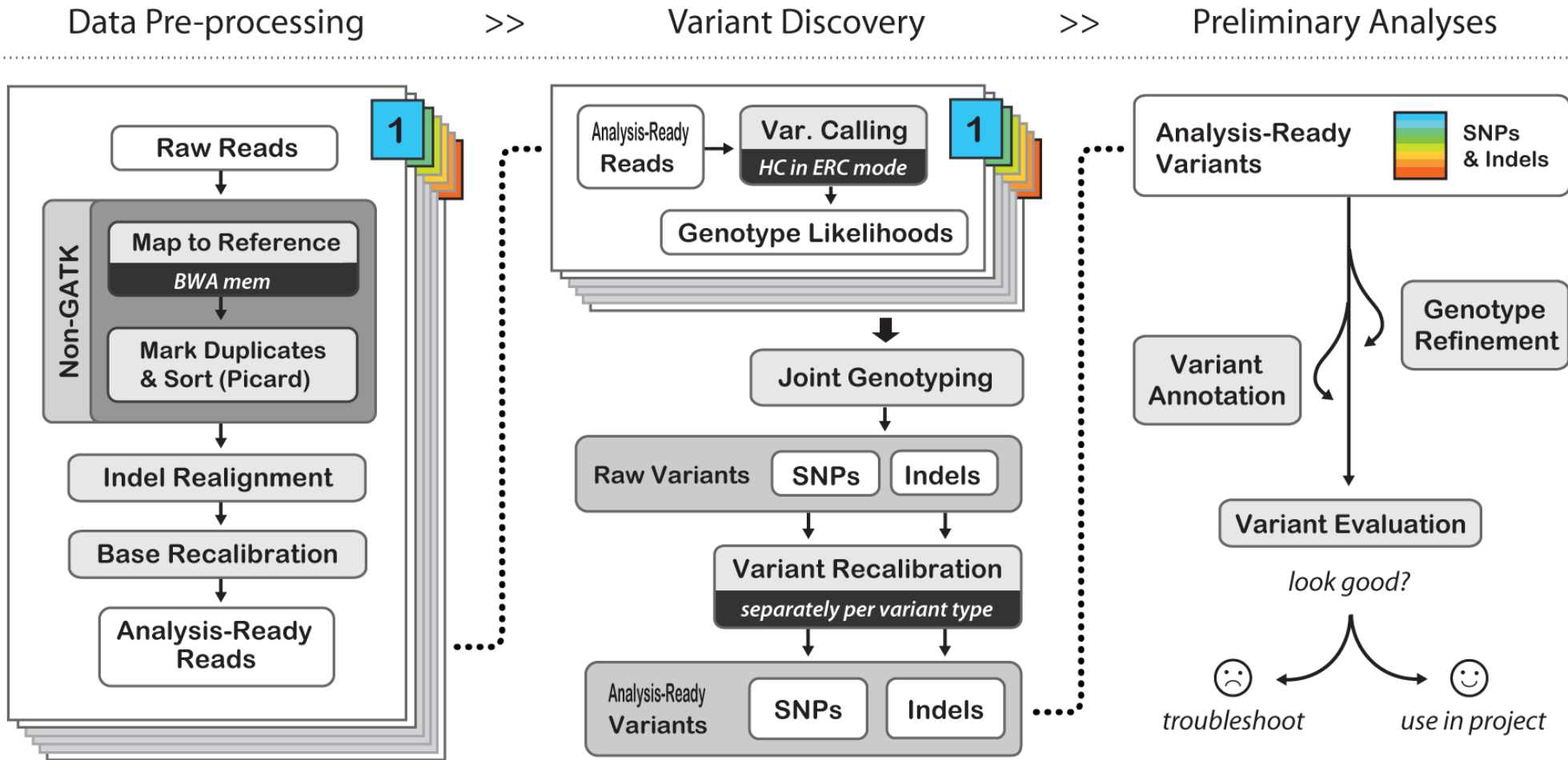
MATERIALS AND METHODS

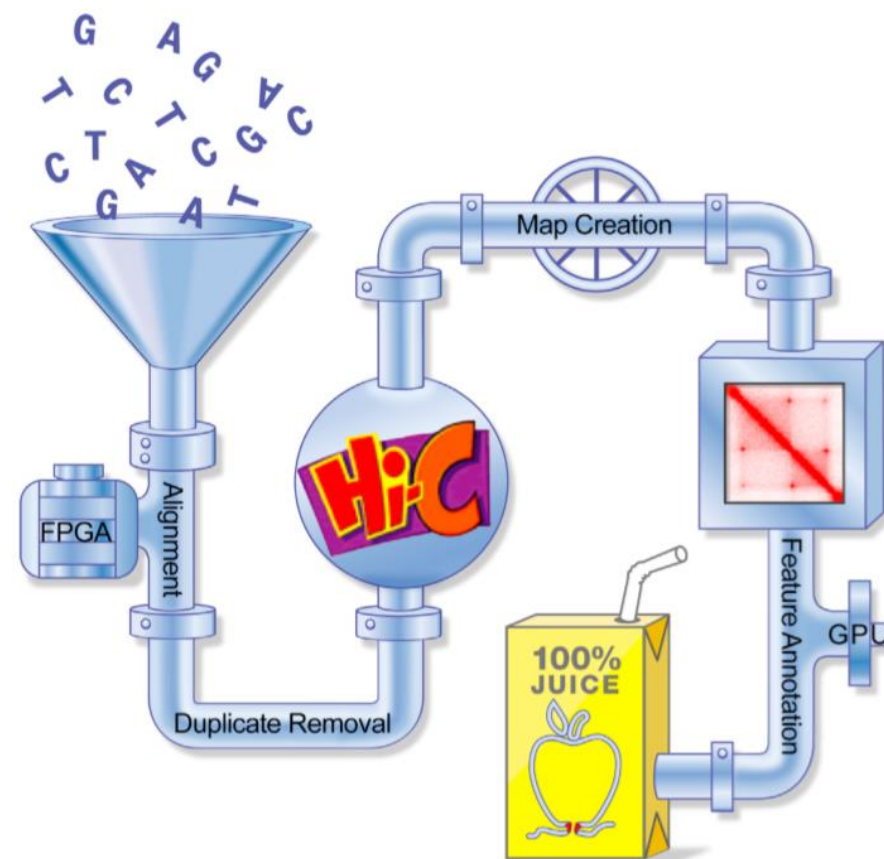
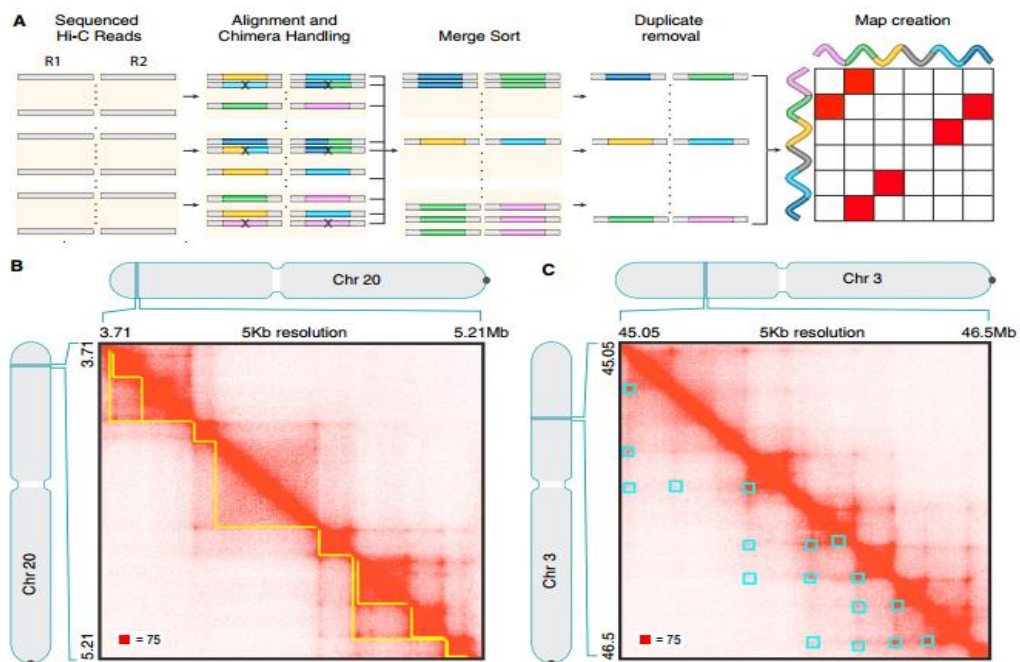
Production of RHs. We fused irradiated zebrafish fin AB9 cells to mouse B78 melanoma cells. The B78 recipient cell line is not deficient in an enzyme that could be used to select for zebrafish chromosomal elements in hybrids. Therefore, zebrafish chromosomes were tagged with the aminoglycoside phosphotransferase gene that confers resistance to G418, as described (17). More than 400 independent G418-resistant AB9 clones were pooled for fusion experiments. Briefly, 3×10^7 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). G418 (800 μ g/ml) was added 24 h after fusion. No colonies were observed in the controls (irradiated AB9 cells, unfused B78 cells, and irradiated AB9 and B78 cells



10 year & \$3 billion

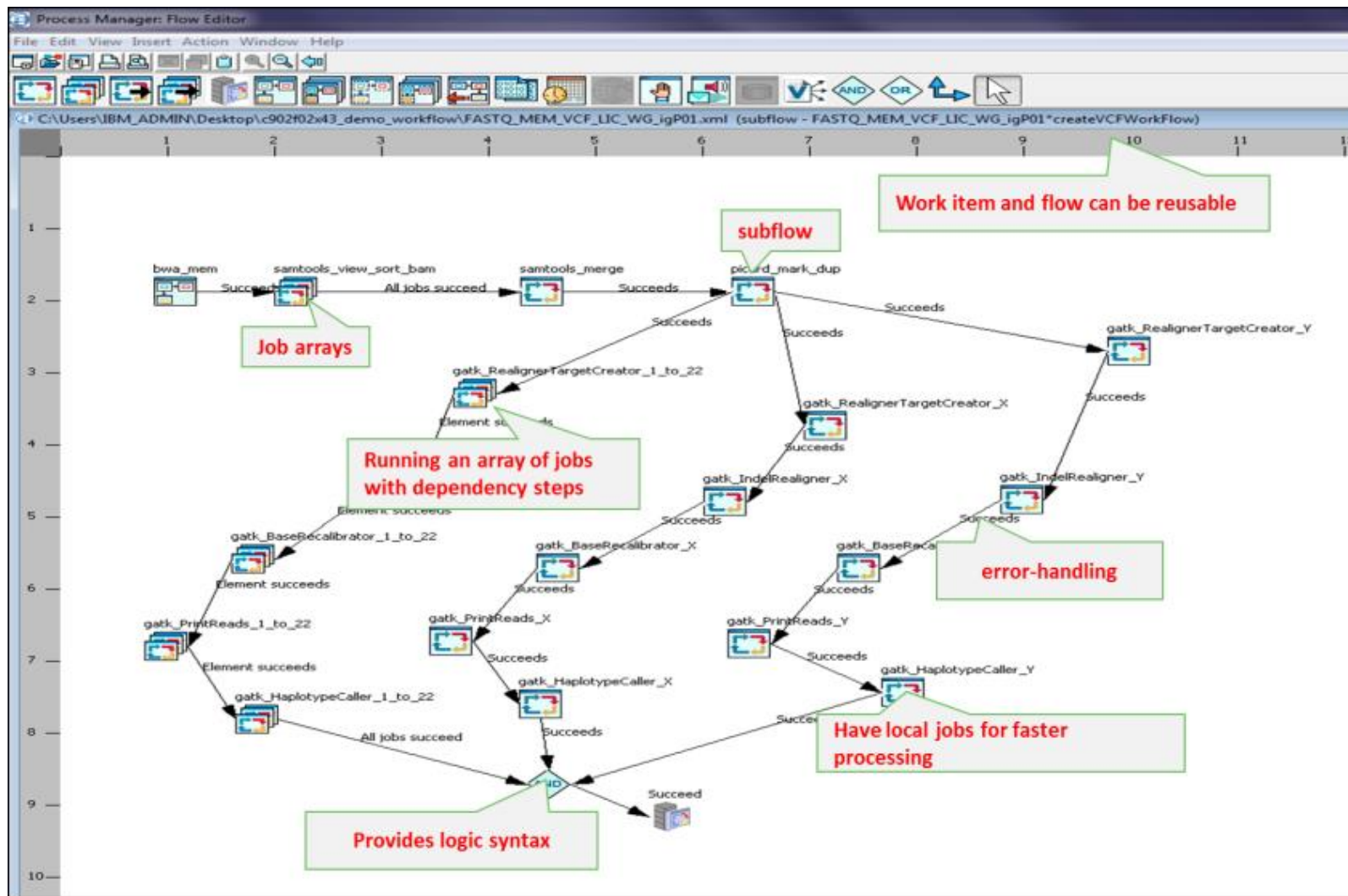
Speed: From Days to Hours

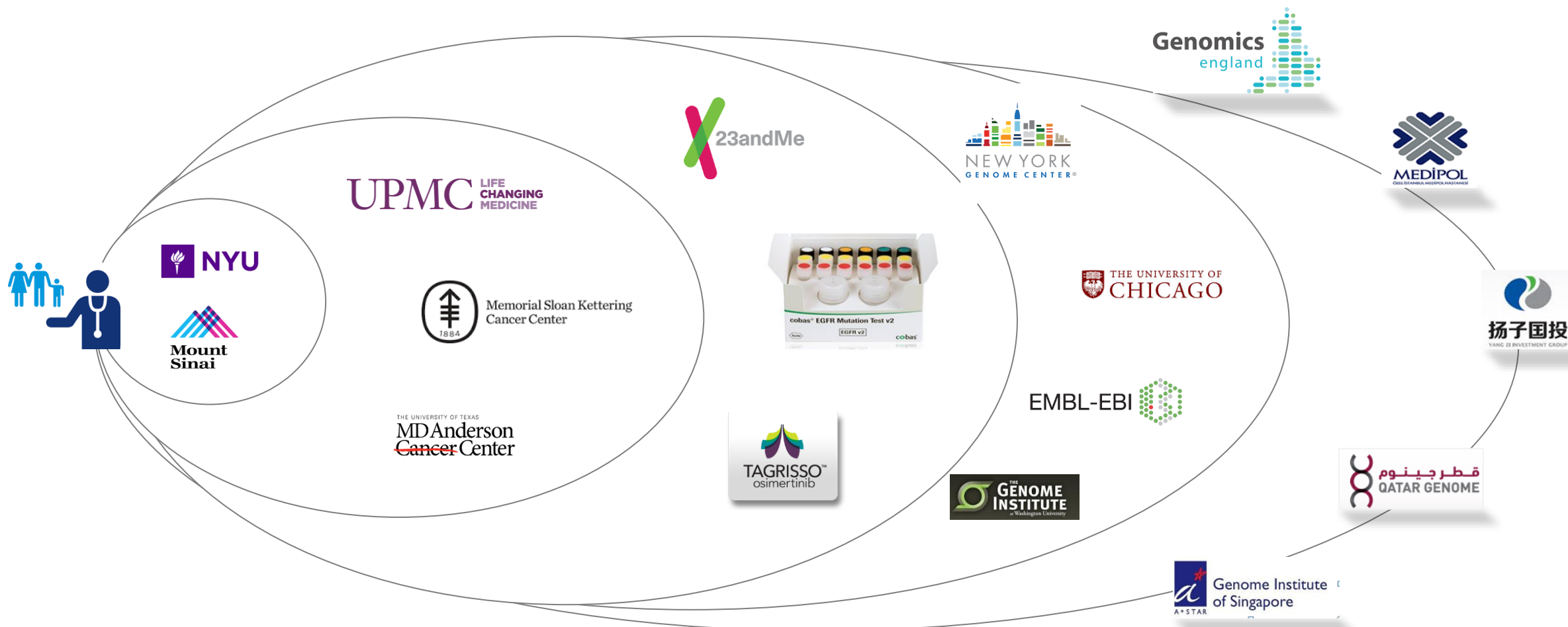




REPORT
GENOME ASSEMBLY
De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds
 Udo Hoffmann,^{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100}
 The *Aedes aegypti* genome is the most important vector for the spread of dengue, Zika, and chikungunya viruses. We present a high-quality, chromosome-length assembly of the *Aedes aegypti* genome using Hi-C data. This assembly is the most complete and accurate to date, and it provides a valuable resource for the study of the *Aedes aegypti* genome and the diseases it transmits.

Smart: Auto-drive Analytical Pipeline





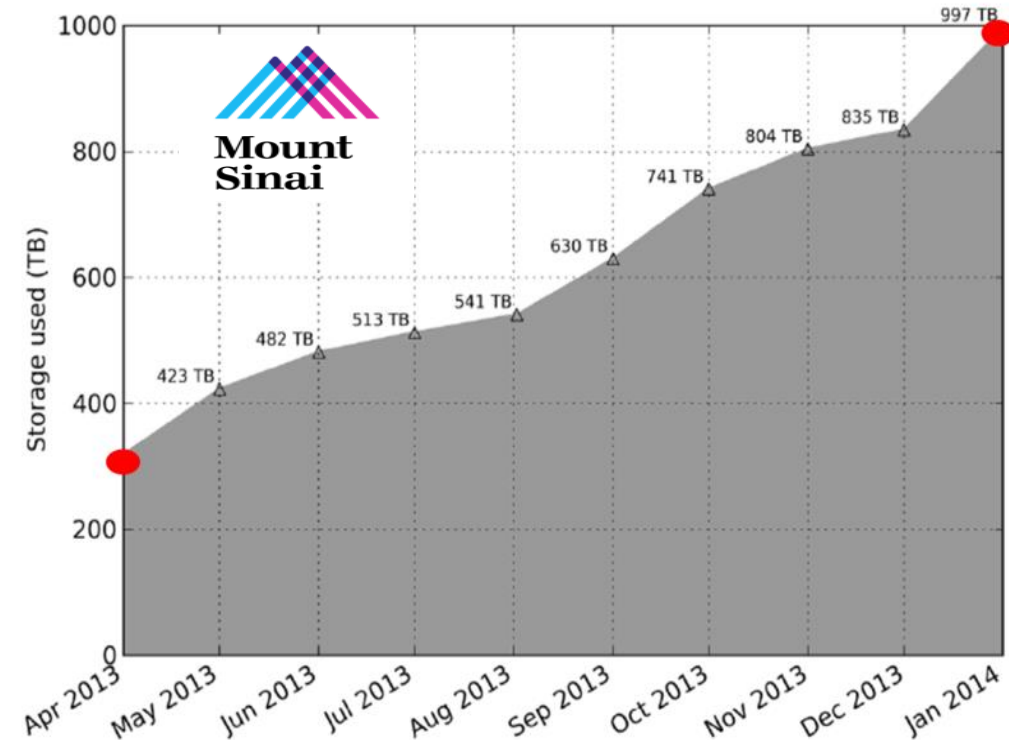
Instrumenting Data: Challenges & Opportunities

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

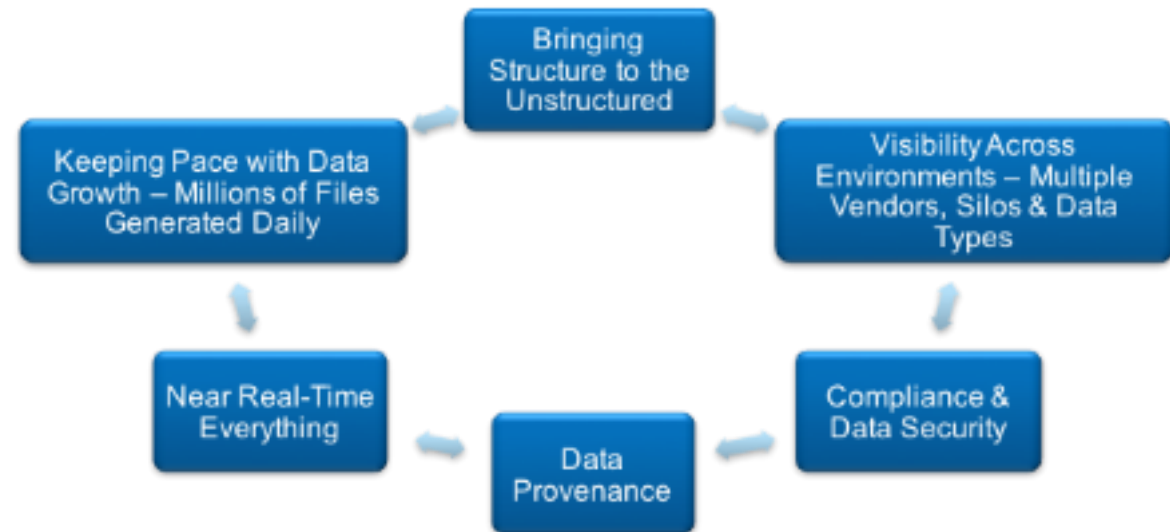


Finding & Tracking Data

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

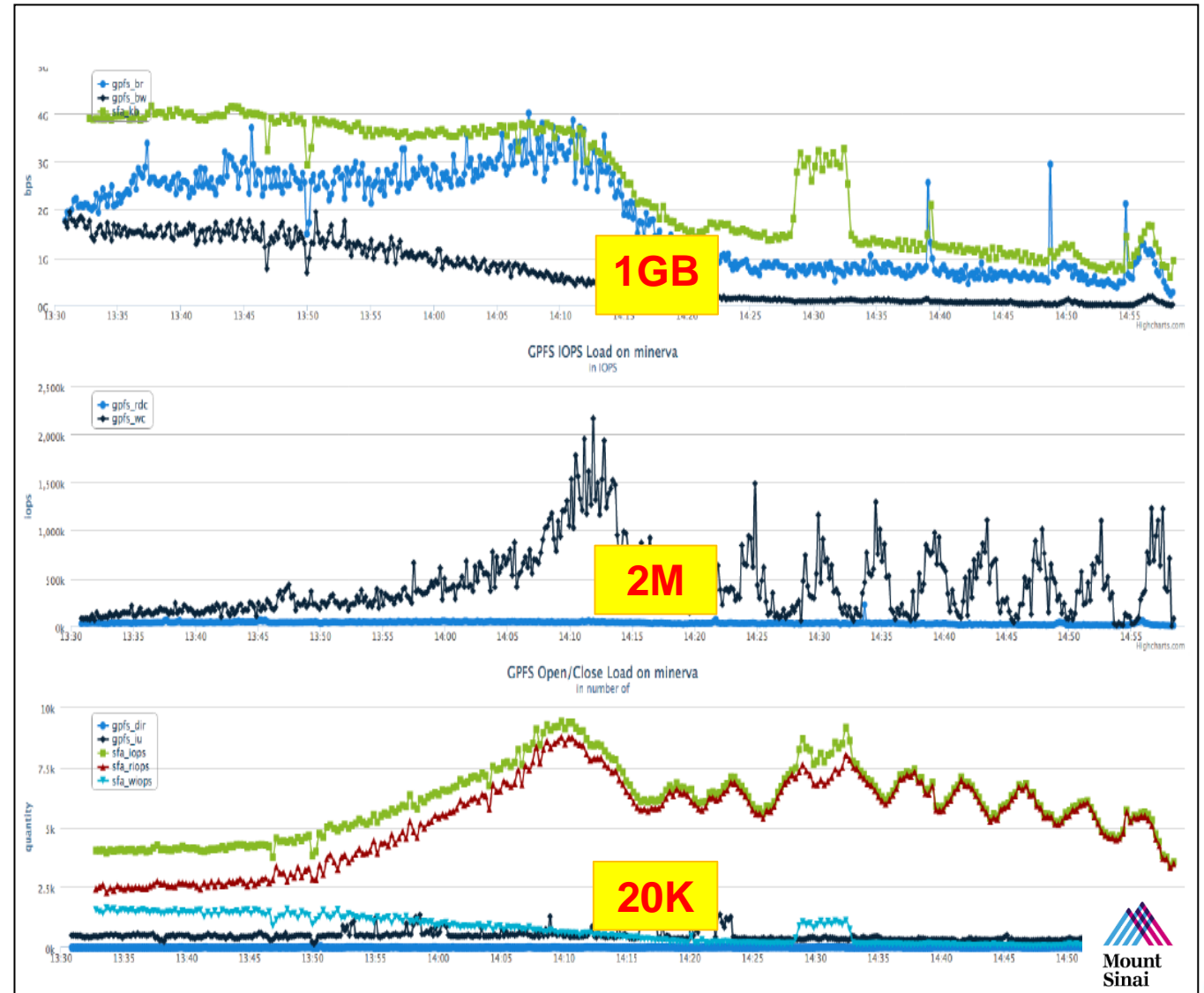


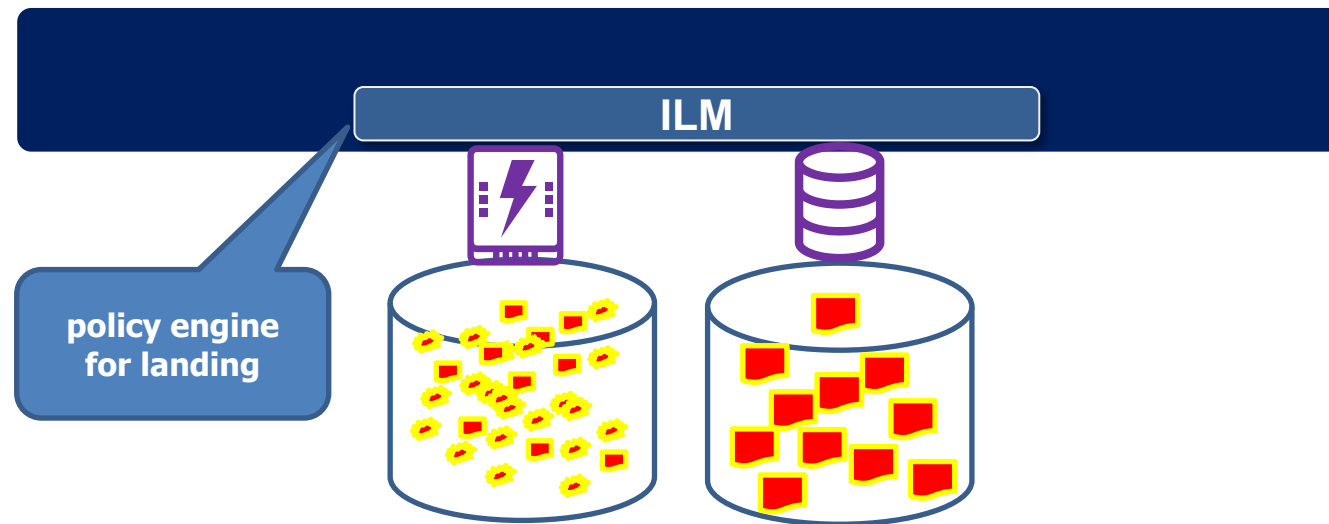
(human API)

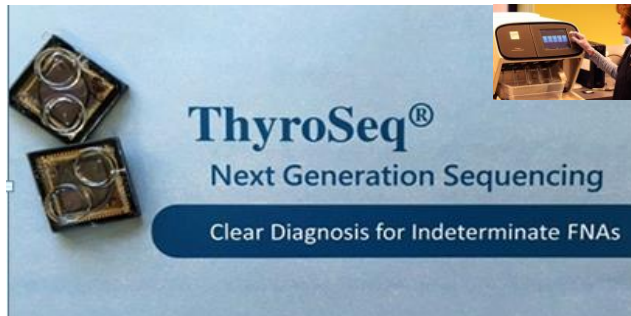


How to Instrument Data?

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

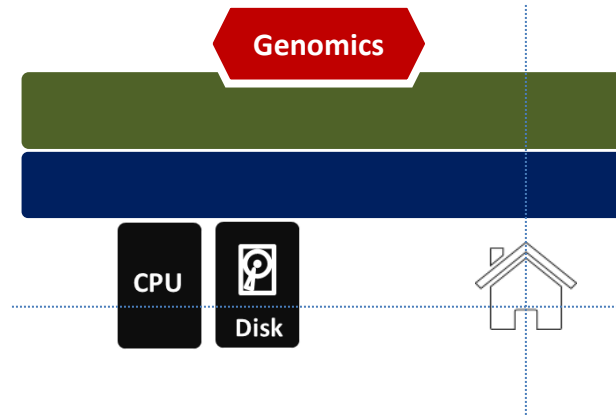






ThyroSeq®
Next Generation Sequencing
Clear Diagnosis for Indeterminate FNAs

MOLECULAR & GENOMIC PATHOLOGY LABORATORY



Children's Hospital of Los Angeles | USC University of Southern California

Unexpected Detection of a Targetable PDGFRA Fusion in a Pediatric Patient Presenting with Refractory T-ALL

Gordana Raca¹, Matthew Oberley¹, Christopher Denton¹, Jiarling Ji¹, Matthew Hemen¹, Kara Duncan¹, Debrahne Oboasi¹, Samuel Wu¹, Deepa Bhargava¹, Paul Gaynon¹
1. Department of Pathology and Laboratory Medicine and 2. Children's Center for Cancer and Blood Diseases, Children's Hospital of Los Angeles

Background
Rare hematologic neoplasms defined in the 2016 World Health Organization (WHO) classification as "myeloid" and "lymphoid neoplasms with neoplasmic and abnormalities of PDGFRA, PDGFRB, FGFR3 and PCCL1-JAK2" are stem cell disorders typically characterized by eosinophilia. Early identification of these neoplasms is of great clinical importance, since the activated tyrosine kinase receptors may be therapeutically targeted. PPLU-PDGFRB-associated neoplasms typically occur in adults, and pediatric cases are exceedingly rare. We describe detection of a PPLU-PDGFRB fusion in a 13-year-old boy whose disease of diagnosis had features of a primary T-ALL, and showed no neoplasmic or other clinical or morphologic characteristics typically observed in PDGFRA-associated neoplasms.

Case History
The patient presented with an enlarging right supraorbital nodule, and a biopsy established a diagnosis of T lymphoblastic leukemia/lymphoma (T-ALL). At the end of induction there were 78 residual blasts, and halfway through consolidation relapsed disease was detected at the level of 0.4%. However, toward the end of consolidation, the patient presented with left eye pain, increased tearing, transient eddches and progressive lymphadenopathy, and was diagnosed with a relapse of his ALL with leukemic infiltrates into the optic nerve. Although local radiation resulted in improved vision, re-induction chemotherapy failed to prevent progression of the disease.

Genetic Testing of Leukemia Cells
Karyotype analysis showed a t(11;14)(p13;q11.2), which juxtaposes the JAK2 gene next to the T-cell-receptor alpha locus (Figure 1). Chromosomal microarray (CMA) testing revealed additional genetic abnormalities recurrent in T-ALL, including a focal 1035 deletion resulting in a deleted PPLU-PDGFRB fusion, and a deletion involving the CXCR2/ACD102B locus. SangerSeq, CMA testing also showed a PPLU deletion.

Genetic Testing Results

Figure 1. Representative karyotype from a lymph node sample showing the t(11;14)(p13;q11.2) [indicated by arrow] which involves the JAK2 locus in 11p13 and the T-cell receptor alpha locus in 14q11.2.

Figure 2. Details of Chromosomal Microarray (CMA) results showing a focal 45 deletion resulting in a PPLU-PDGFRB fusion. A PPLU deletion was detected on 11p13 with the centromeric breakpoint within the PPLU gene and the telomeric breakpoint within the PDGFRA gene. Vertical orange dashed lines denote approximate breakpoints within the fusion genes.

Figure 3. Details of the detected PPLU-PDGFRB fusion. A schematic representation and the exact sequence of the fusion reads are shown in the panels A and B.

CHLA OncoKids Cancer Panel

		Processing Stage	IBM Runtime
BWA	4.26	BWA mem	4.54
Samtools	2.08	SortSam	0.18
MarkDuplicates	6.86	Mark Duplicate	
RealignTargets		GATK BaseRecalibrator	1.15
IndelRealigner	1.06	GATK PrintReads	1.18
GATK BaseRecalibrator	1.49	GATK HaplotypeCaller	1.43
GATK PrintReads+Index	2.55	Total Runtime	9.85
GATK HaplotypeCaller	2.64		
Total	20.94		

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

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

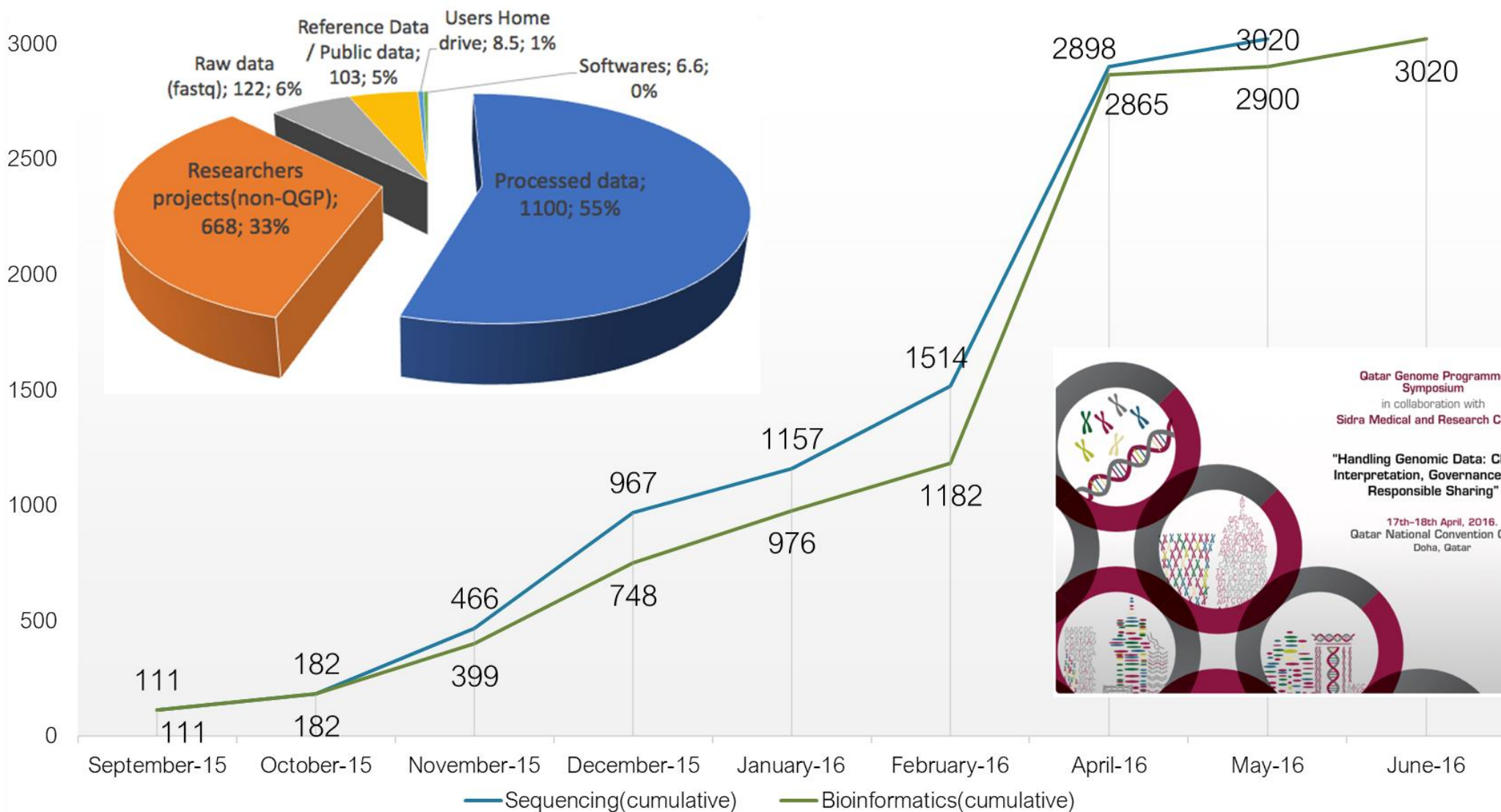
What

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

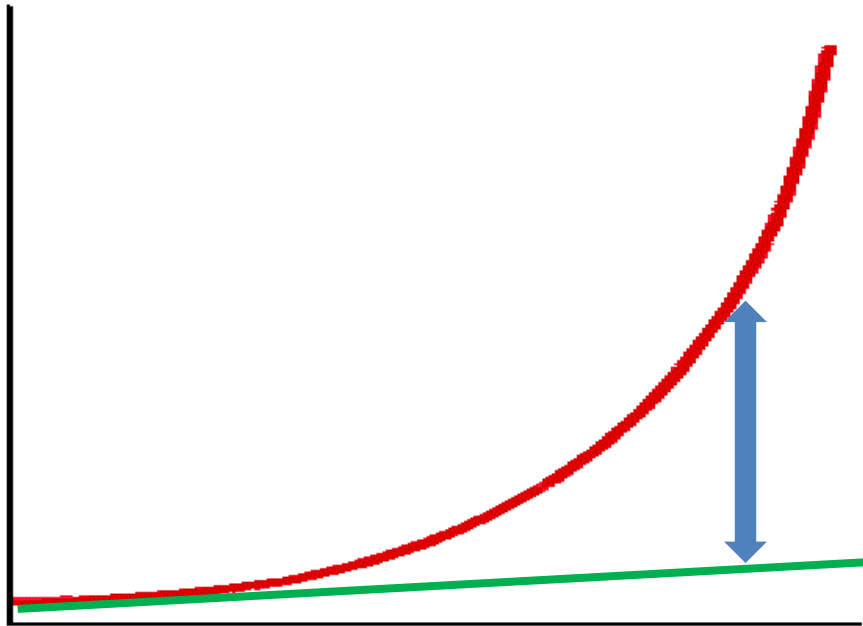
When

- 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

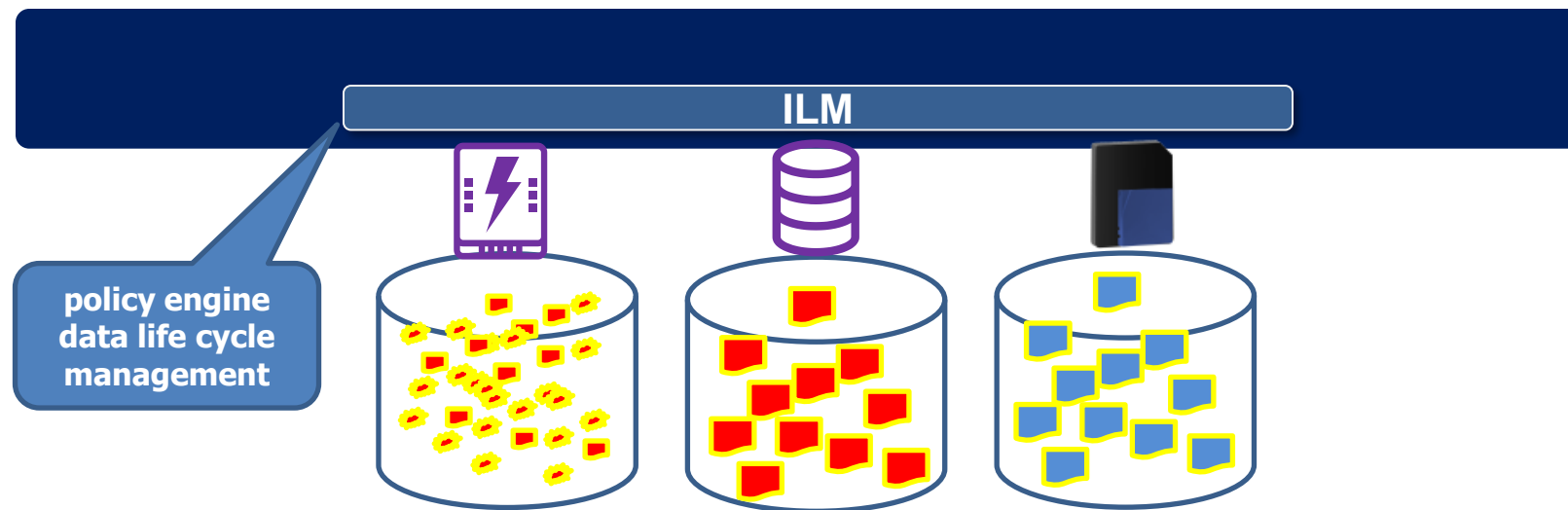


"Bending the Curve"

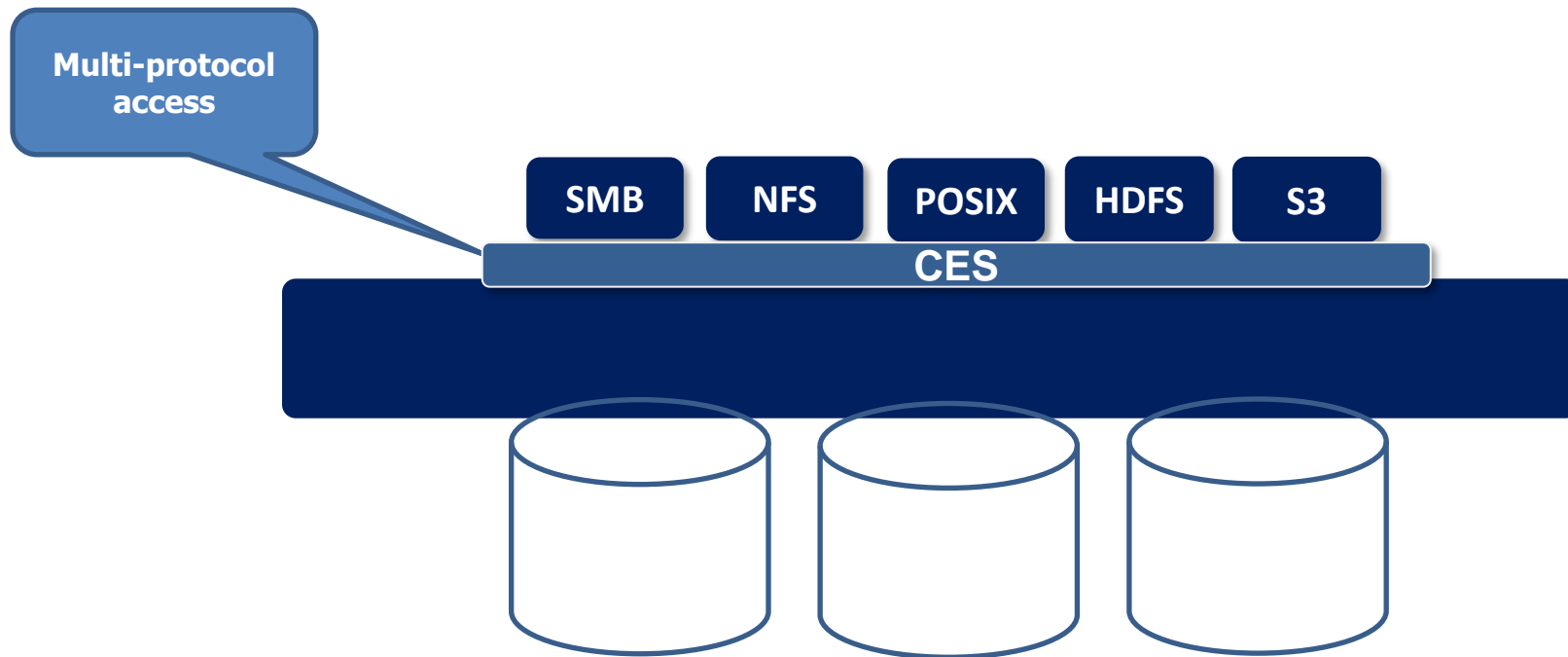


THE UNIVERSITY OF TEXAS
MDAnderson
Cancer Center

HIMSS18



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





NFS

POSIX

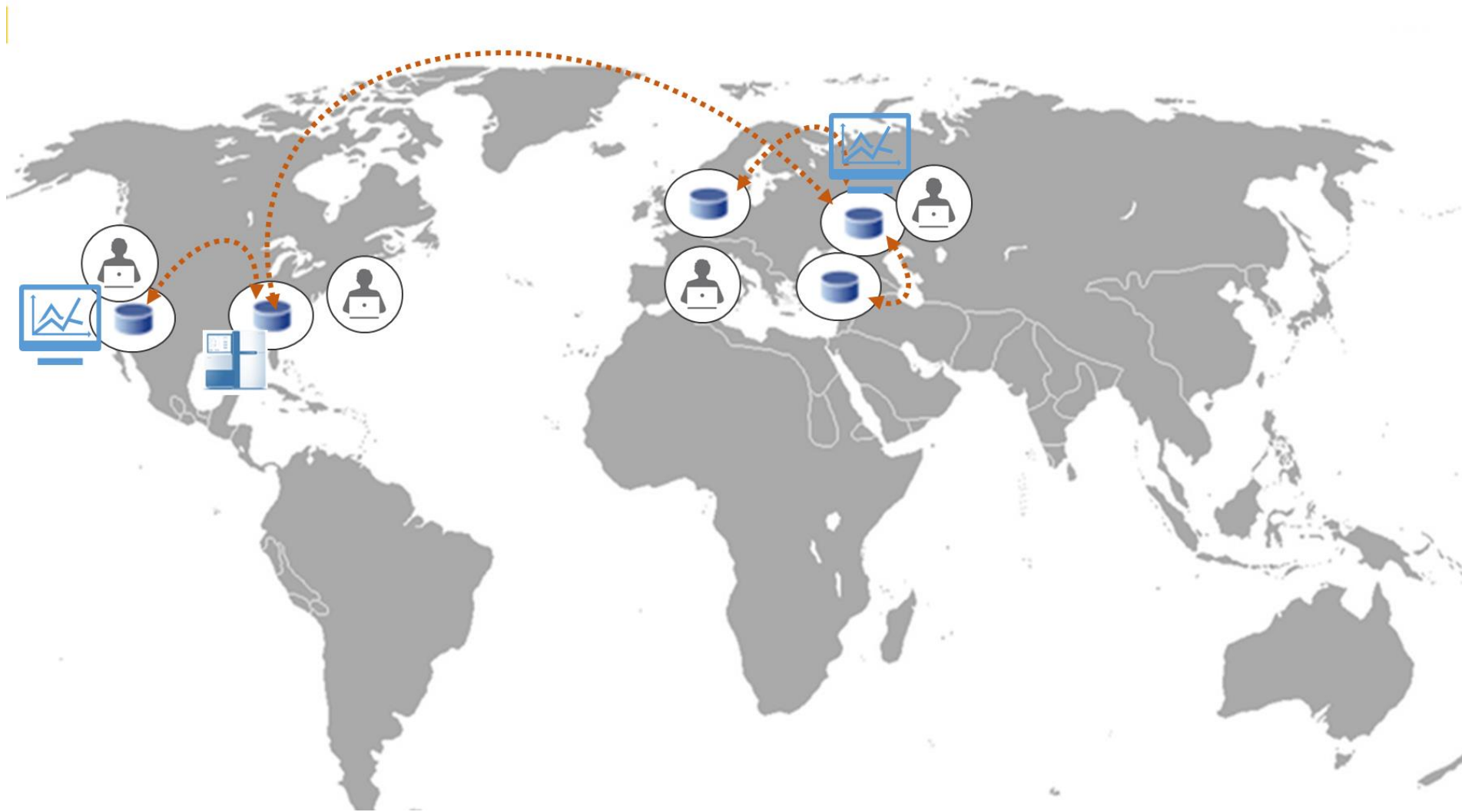


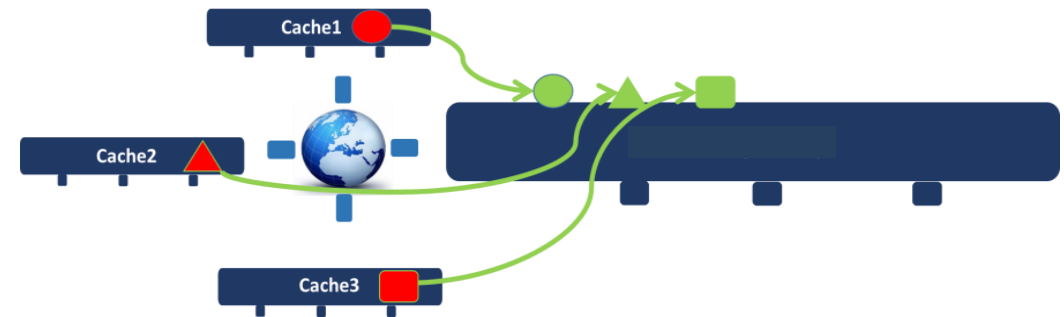
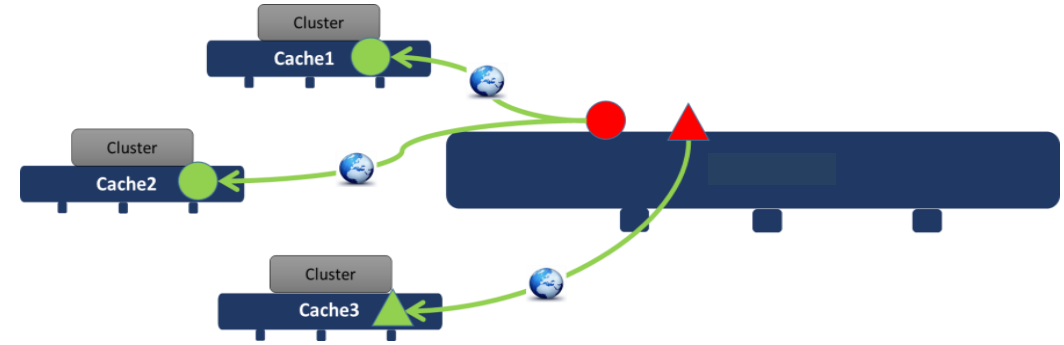
HDFS

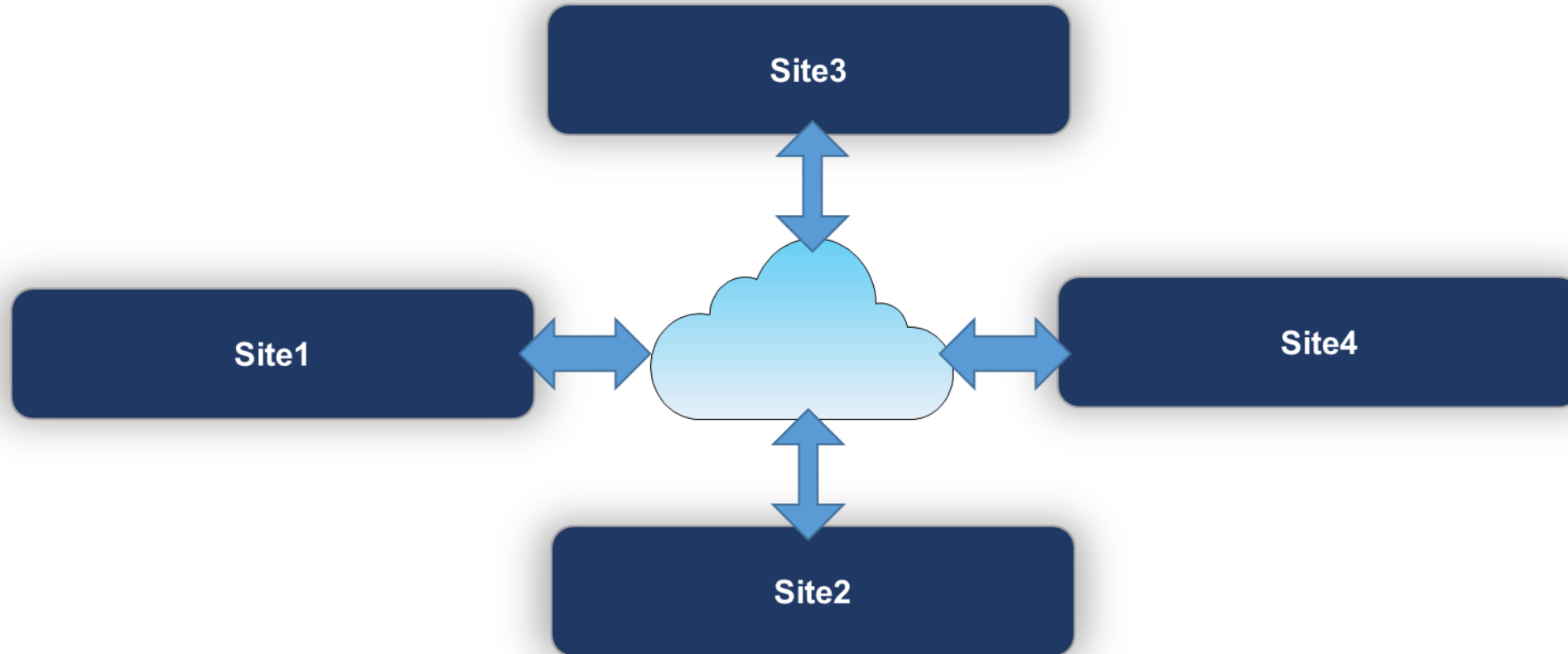


SMB

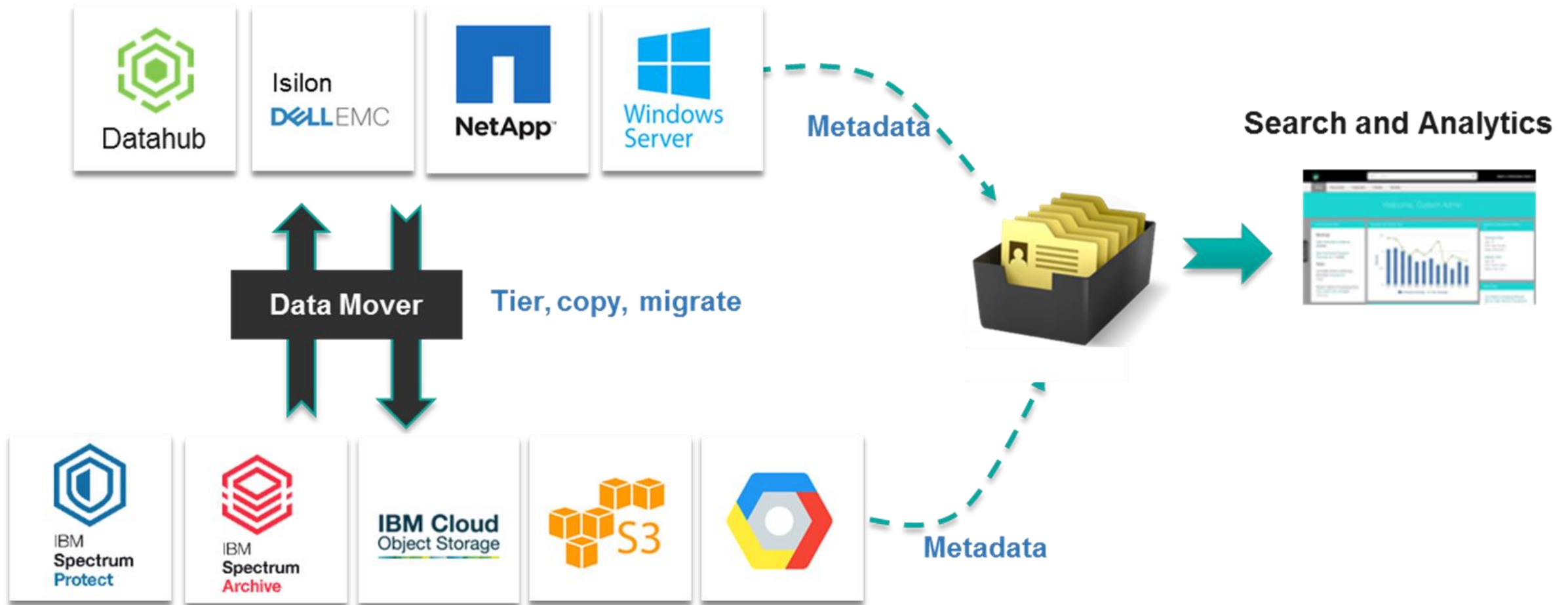
- Fast Data Landing
- Smart Data Tiering
- Flexible Data Accessing
- **Global Data Peering**
- Data Cataloguing

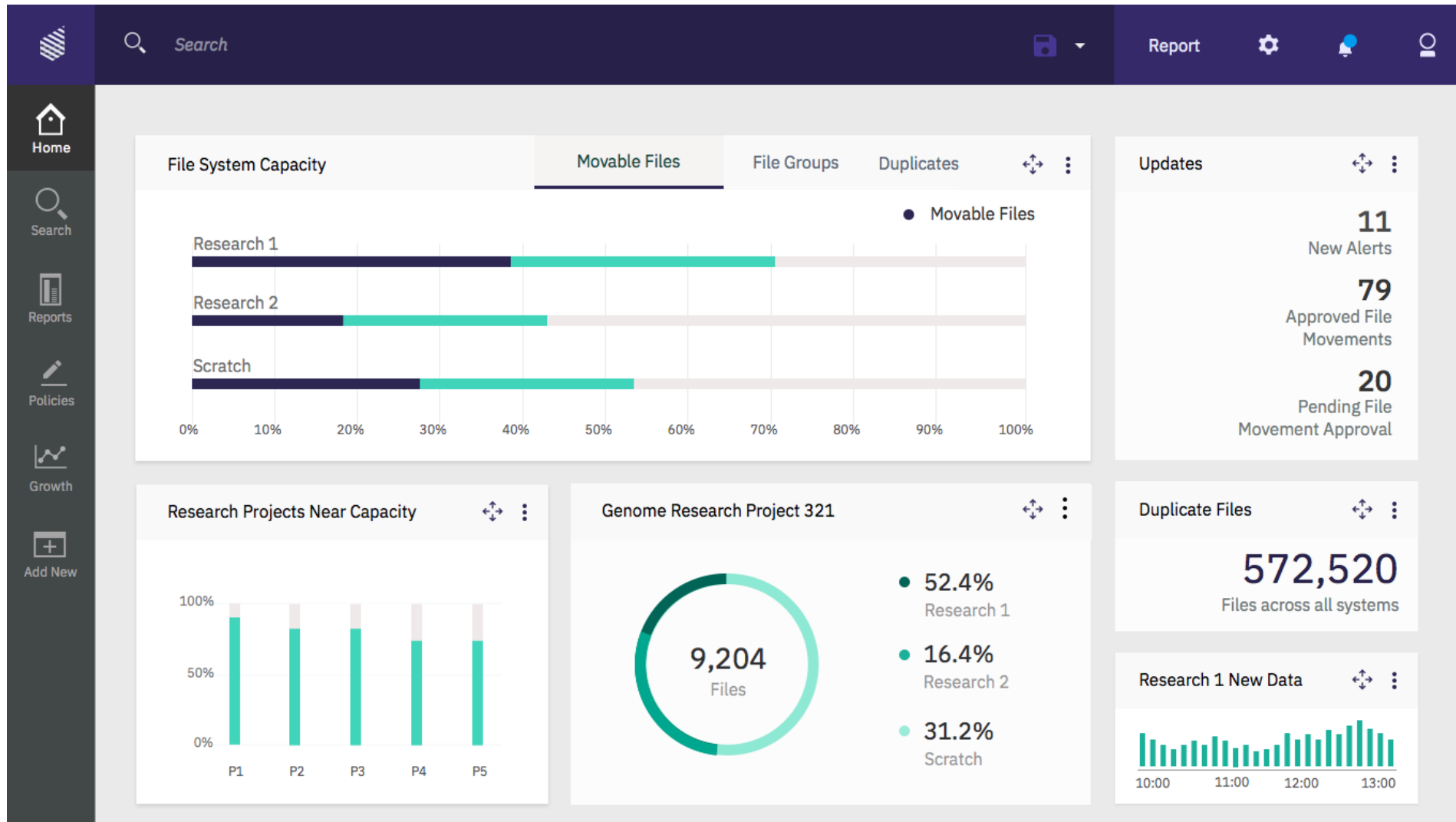






- Fast Data Landing
- Smart Data Tiering
- Flexible Data Accessing
- Global Data Peering
- **Data Cataloguing**





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

Genomic Variants

Samples

	var_1	28	32	29	28	35	32
var_1	A/T	A/A	A/T	T/T	A/A	A/T	...
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
- ...

Clear
No filters selected

Table Result
Summary (Beta)
Genome Browser (Beta)

Showing 1-10 of -1 variants Download

Variant	SNP Id	Genes	Type	Consequence Type	Deleteriousness			Conservation			Population Frequencies		
					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			

Study

Studies Filter

In (AND):

- 1kG_phase3
- ESP6500
- EXAC
- 1kG_phase3_chrY
- 1kG_phase3_chrMT
- GONL
- UK10K_ALSPAC
- UK10K_TWINSUK
- MGP

Genomic

Chromosomal Location

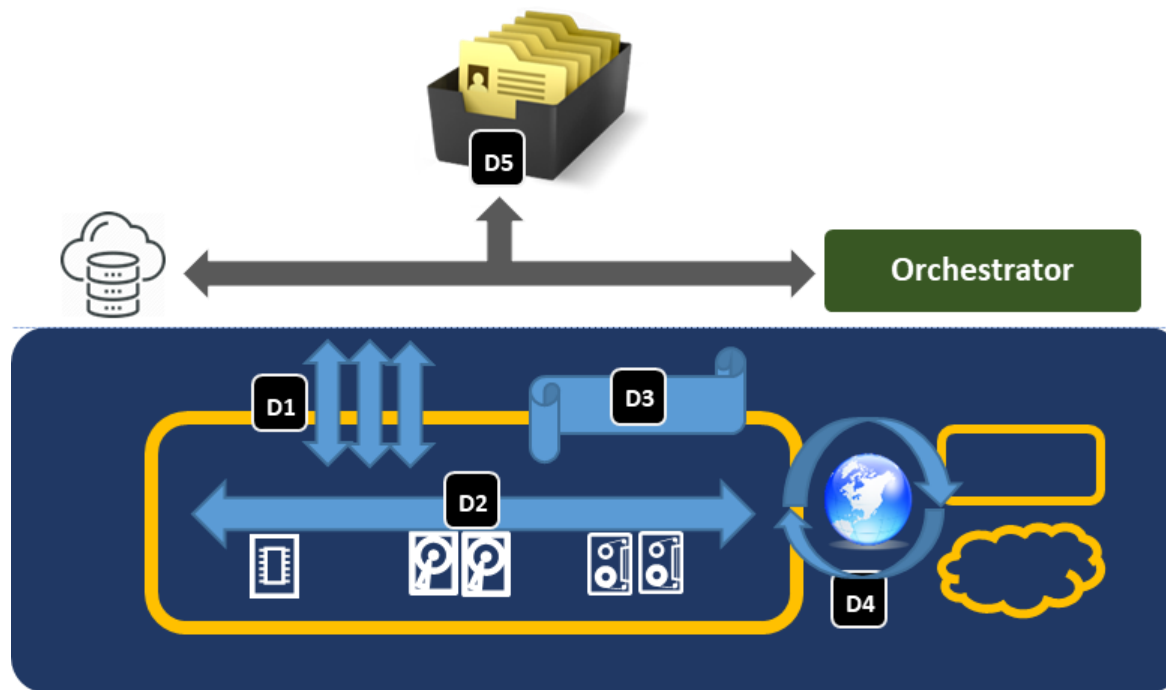
3:444-55555, 1:1-100000

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

Search for Gene Synt

BRCA2, ENSG00000139618

<http://hgva.opencb.org>



- D1** Parallel I/O
- D2** ILM-based Tiering
- D3** Multi-protocol Access
- D4** Multi-site Peering
- D5** Metadata Engine

Instrumented Data: Early Applications

Search all beacons for allele







GRCh37 ▾ 13 : 32936732 G > C Search

Response All None

- Found 10
- Not Found 33
- Not Applicable 25

Organization All None

- AMPLab, UC Berkeley
- Australian Genomics He...
- Belgian Medical Genomi...
- BGI
- Bioinformatics Area, Fun...
- BioReference Laboratori...
- Brazilian Initiative on Pre...
- BRCA Exchange
- Broad Institute
- Centre for Genomic Reg...
- Centro Nacional de Anal...
- Children's Mercy Hospital
- Curoverse
- DNASTack
- ELIXIR
- EMBL European Bioinfor...

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

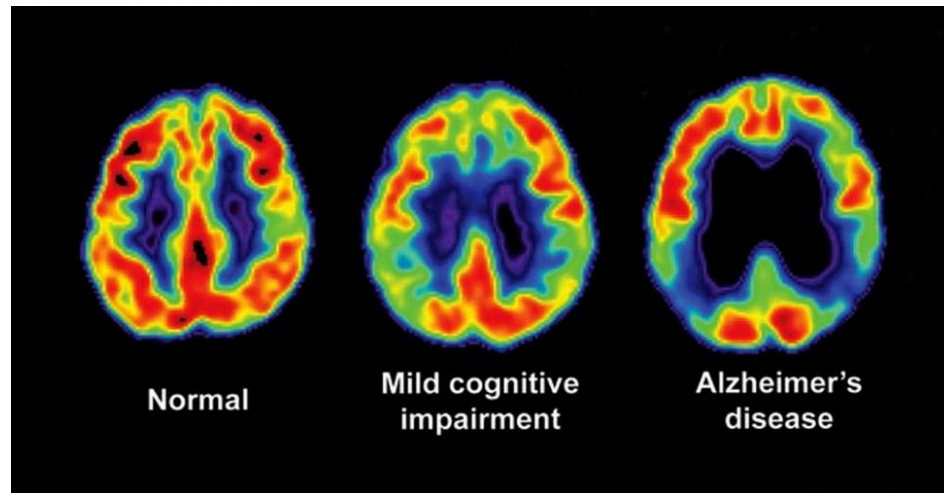
Assignee: International Business Machines Corporation (Armonk, NY)

Family ID: 1000001877202

Appl. No.: 14/243,301

Filed: April 2, 2014

Bouncing Back



Quick Facts

6th

Alzheimer's disease is the sixth leading cause of death in the United States.

5 million

More than 5 million Americans are living with the disease.



1 in 3 seniors dies with Alzheimer's or another dementia.

216 billion

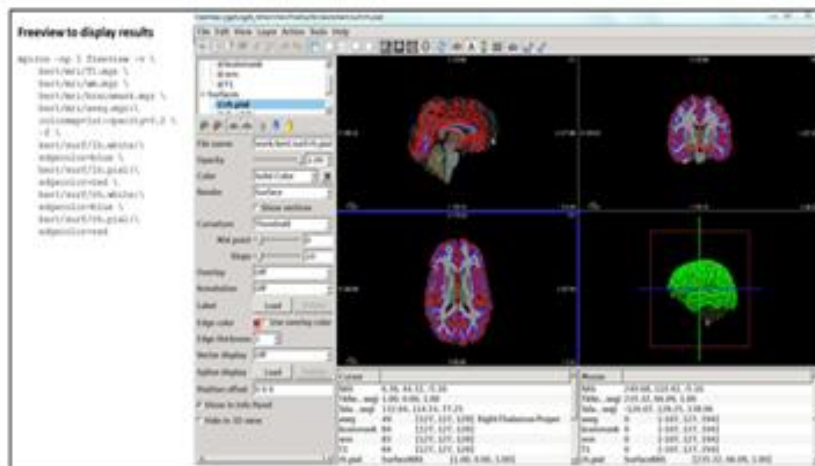
In 2012, 15.4 million caregivers provided more than 17.5 billion hours of unpaid care valued at \$216 billion.



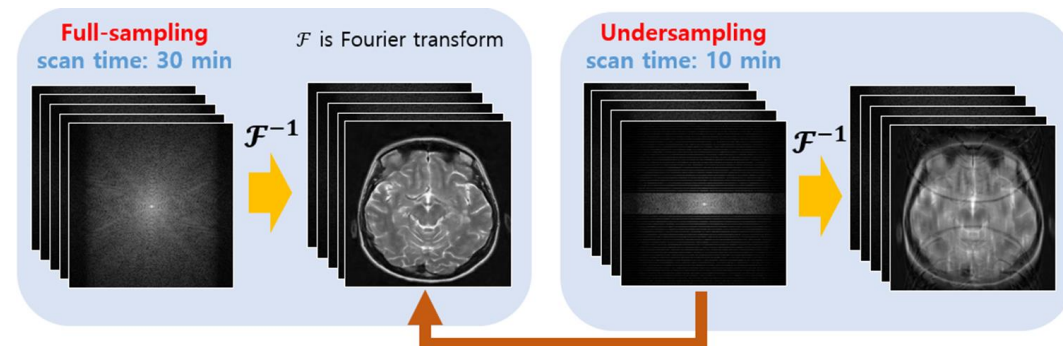
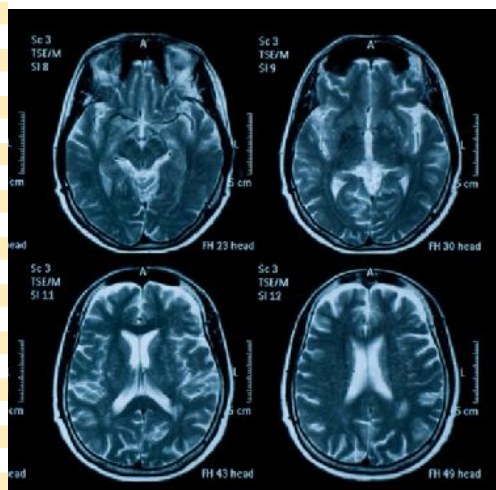
Nearly 15% of caregivers for people with Alzheimer's or another dementia are long-distance caregivers.



In 2013, Alzheimer's will cost the nation \$203 billion. This number is expected to rise to \$1.2 trillion by 2050.



- 0020,9518 AcquisitionIndex
- 0020,9529 ContributingSOPInstancesRefSeq
- 0020,9536 ReconstructionIndex
- 0021,1003 SeriesFromWhichPrescribed
- 0021,1005 GenesisVersionNow
- 0021,1007 SeriesRecordChecksum
- 0021,1018 GenesisVersionNow
- 0021,1019 AcqreconRecordChecksum
- 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



The goal of undersampled MRI is to develop this map.

