Standardizing the Next Generation of Bioinformatics Software Development With BioHDF (HDF5)

BioHDF BoF SC09

geospiza™

Overview

- Driver: bioinformatics challenges in Next Generation DNA Sequencing (NGS)
- BioHDF project and examples
- HDF5 (Hierarchical Data Format)

Contributors

- Cornell
  - Christopher Mason
  - Paul Zumbo
- Yale
  - Stephen Sanders
- The HDF Group
  - Mike Folk
  - Dana Robinson
  - Ruth Aydt
- Tahoe Informatic
  - Martin Gollery
- Geospiza
  - Mark Welsh
  - N. Eric Olson
  - Todd Smith
- Funding
  - NIH STTR HG003792

Next Generation DNA Sequencing

"Transforms today's biology"

"Democratizing genomics"

NGS is Powerful

"Changing the landscape"

"Genome center in a mail room"

"The beginning of the end for microarrays"

geospiza

Example: Measuring Gene Expression

dbEST - Jan 20, 2009

Total Organisms	1,683
Total ESTs	59,498,205
Human	8,163,902
Mouse	4,850,605
Maize	2,018,337
Arabidopsis	1,526,124
Cattle	1,517,143
Pig	1,476,771
Soybean	1,380,071
Zebra Fish	1,380,017
5 others	>1,000,000

Other Technologies

Experiment	Measurements
SAGE	1,000-100,000
Microarrays	1.4 M (probes) 0.8 M (probes) 48 K (transcripts)

Next Generation Sequencing

Instrument	Measurements
454 Titanium	**1,000,000**
Illumina GA	**80,000,000**
SOLiD V3	**180,000,000**

Greater sensitivity, higher dynamic ranges
+ Qualitative data: isoforms, alleles, …

geospiza

NGS is Daunting

"Prepare for the deluge"

"Byte-ing off more than you can chew"

"These sequencers are going to totally screw you"

"If the data problem is not addressed, ABI's SOLiD, 454's GS FLX, Illumina's GAII or any of the other deep sequencing platforms will be destined to sit in their air-conditioned rooms like a Stradivarius without a bow."

geospiza

NGS Data are Analyzed in Three Phases

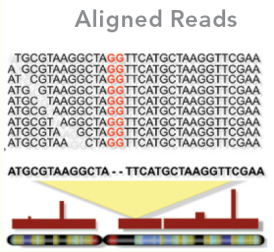
Primary Data Analysis - Images to bases

Sequences +
Quality values
Run quality

Ref Seq +
Aligner

Assembler

Secondary Data Analysis



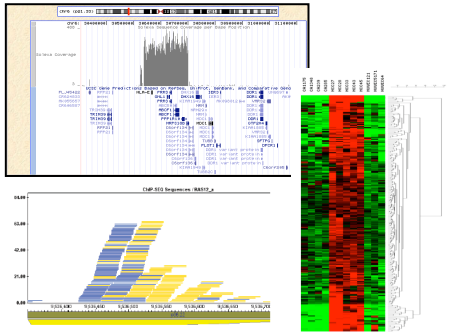
Gene lists
Read Density
Variant list
Sample, run quality

Secondary Data Production
***De novo* assembly =>**

One or more
Data sets

Contigs + Annotation

Tertiary Data Analysis



Differential expression
Methylation sites
Gene association
Genomic structure
Experiment, science

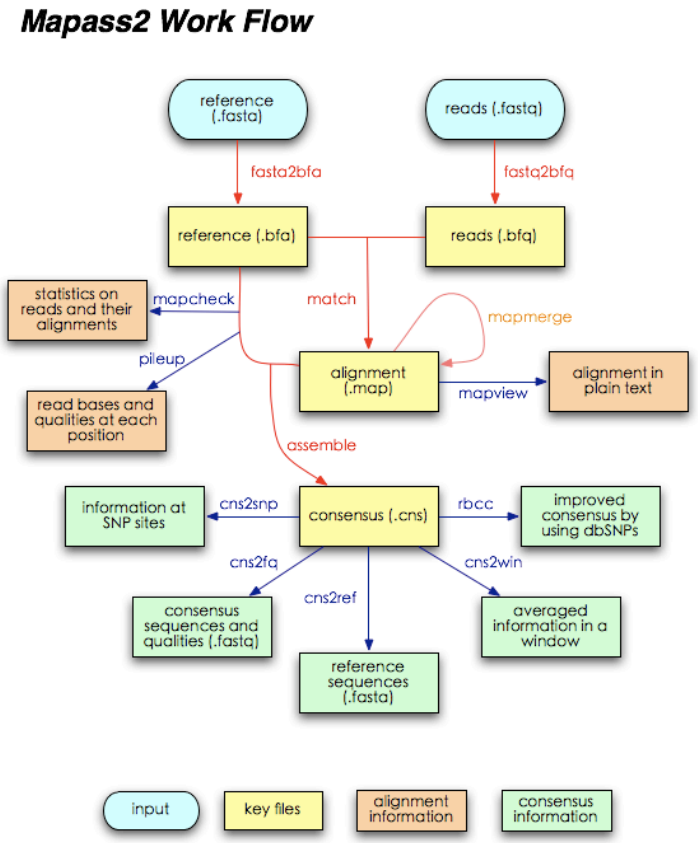
geospiza

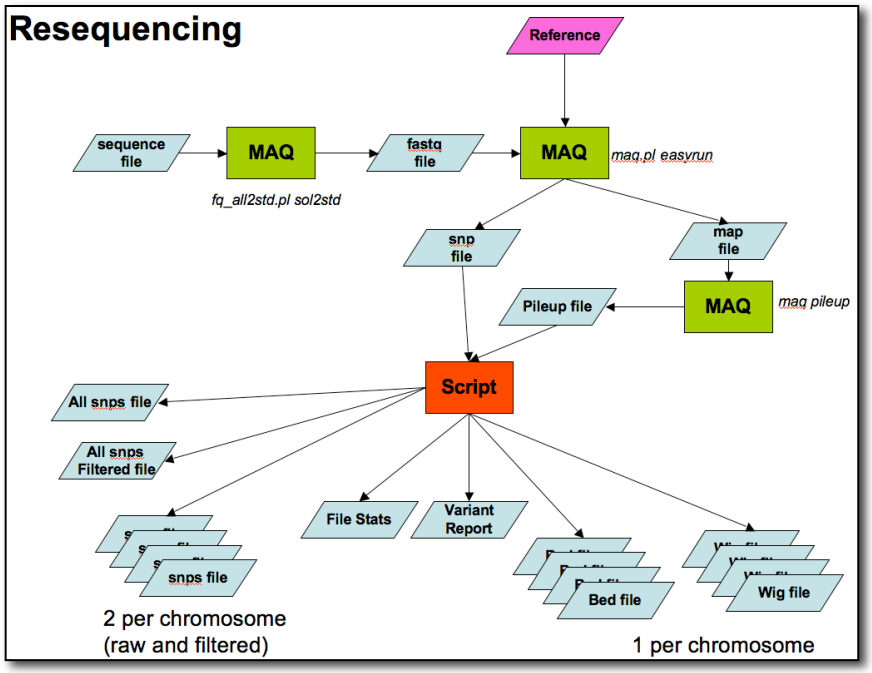
Secondary Analysis is Complex

Examples: MAQ - http://maq.sourceforge.net

Secondary Analysis for:

ChIP-Seq Tag profiling Resequencing





Story repeats for BWA, Bowtie, TopHat, Mapreads, SOAP …

geospiza

Complexity Limits Scale and Productivity

- Data are fundamentally unstructured
- Solve problems with redundant data processing
  - Incremental processing with data filtering
  - New question - rerun alignments
- Each question needs a new output format
  - One file for tables of alignments
  - Another file with bases aligned to see mismatches
  - Another file to ask statistical questions
  - More files and images for visualization
  - Files are linked by virtue of being in the same directory
  - Perl hashes fill up and keep running out of disk space

Makes Getting Answers Difficult

Process

10 - 100 million reads

↓

Align to reference data

↓

Parse files, reformat data, create reports

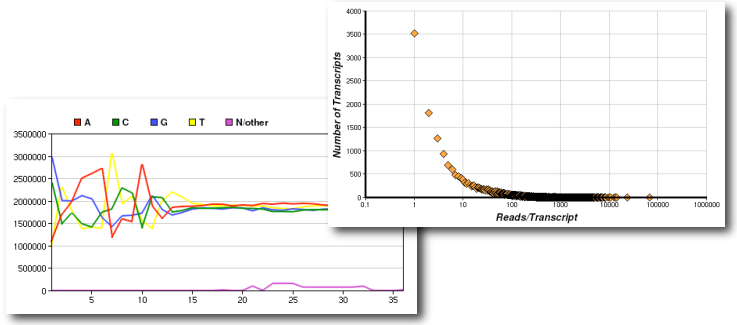
↓

Review results, make decisions

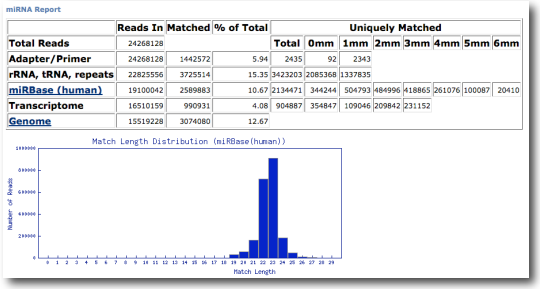
Applications

Gene Expression

Count	RPM	Read Map	RefSeq ID	Title	Gene ID	Chrom.	Type
67898	16880.64		NR_003287	28S ribosomal RNA	LOC100008589	12	rRNA
23231	5775.64		NM_002055	glial fibrillary acidic protein	GFAP	17	mRNA
14380	3575.12		NR_003286	18S ribosomal RNA	LOC100008588	21|Un	rRNA
13416	3335.45		NM_006888	calmodulin 1 (phosphorylase kinase, delta)	CALM1	14	mRNA
12661	3147.75		NR_002819	metastasis associated lung adenocarcinoma transcript 1 (non-protein coding)	MALAT1	11	miscRNA
10649	2647.53		NM_005909	microtubule-associated protein 1B	MAP1B	5	mRNA
10584	2631.37		NR_002715	RNA, 7SL, cytoplasmic 1	RN7SL1	14	miscRNA
10053	2499.35		NM_000533	proteolipid protein 1	PLP1	X	mRNA
9594	2385.24		NM_199478	proteolipid protein 1	PLP1	X	mRNA
9346	2323.58		NM_001101	actin, beta	ACTB	7	mRNA
8460	2103.31		NM_152793	chromosome 7 open reading frame 41	C7orf41	7	mRNA
8137	2023.00		NM_005184	calmodulin 3 (phosphorylase kinase, delta)	CALM3	19	mRNA
8026	1995.41		NM_006087	tubulin, beta 4	TUBB4	19	mRNA
7829	1946.43		NM_002373	microtubule-associated protein 1A	MAP1A	15	mRNA
7364	1830.82		NM_004321	kinesin family member 1A	KIF1A	2	mRNA
7230	1797.51		NM_130811	synaptosomal-associated protein, 25kDa	SNAP25	20	mRNA
7185	1786.32		NM_001961	eukaryotic translation elongation factor 2	EEF2	19	mRNA
7075	1758.97		NM_001831	clusterin	CLU	8	mRNA
6921	1720.68		NR_001568	brain cytoplasmic RNA 1 (non-protein coding)	BCYRN1	2	miscRNA



Small RNA



Epi-Genomics

Reference	Reads	Max. Density	Data	View	Overview
chr1	115109	36	WIG	Genome Browser	
chr2	100061	3729	WIG	Genome Browser	
chr3	63867	12	WIG	Genome Browser	
chr4	69857	64	WIG	Genome Browser	

Variation Analysis

Reference	Reads	Variants (raw)	Variants (filtered)	Data	View	Overview
chr1	106843	2464	152	BED WIG	Genome Browser	
chr2	76870	2436	138	BED WIG	Genome Browser	
chr3	90614	2535	89	BED WIG	Genome Browser	
chr4	72492	2180	141	BED WIG	Genome Browser	

geospiza

And Comparing Between Samples Hard

Process

10 - 100 million reads

↓

Align to reference data

↓

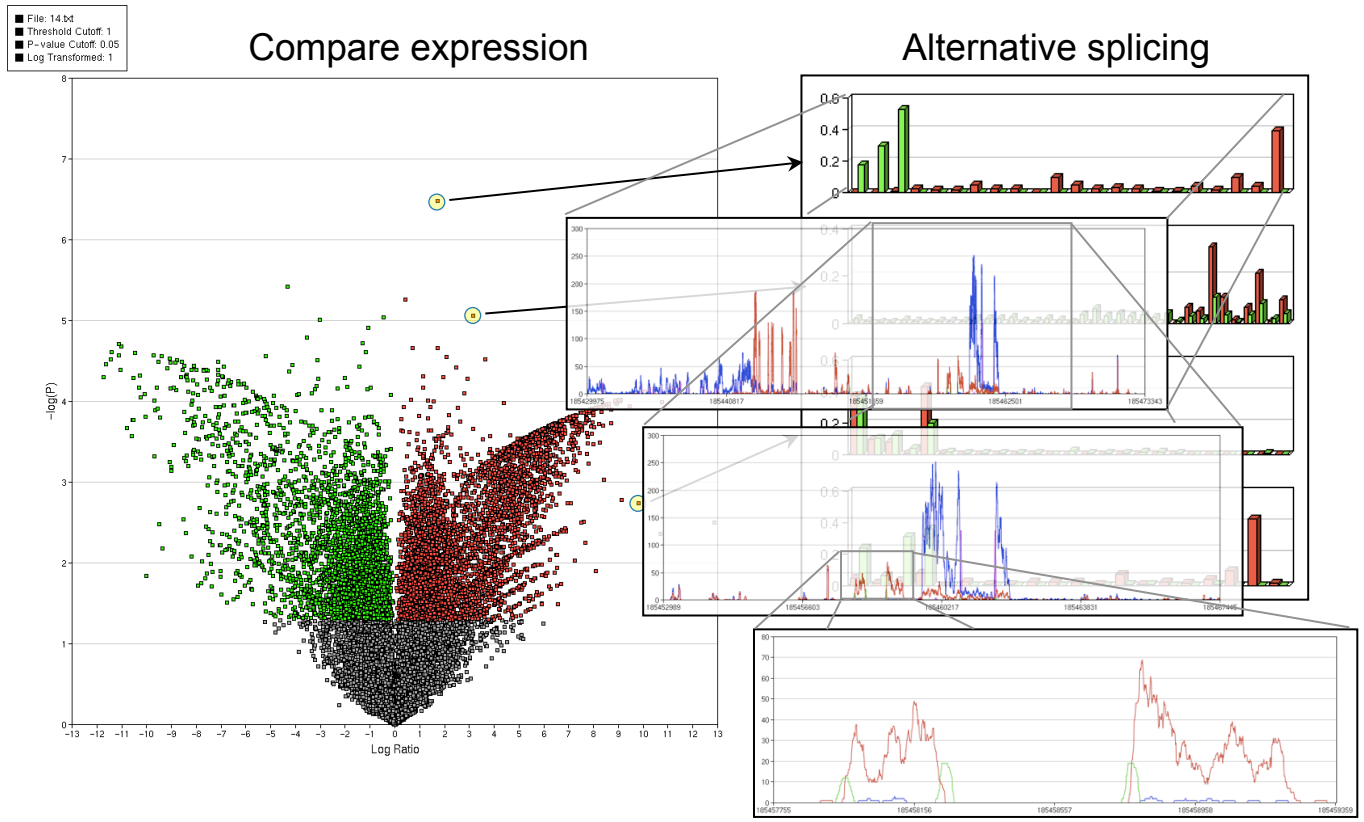
Review results, make decisions

↓

Repeat n times, With n samples

Explore Data Between Samples

And drill into details

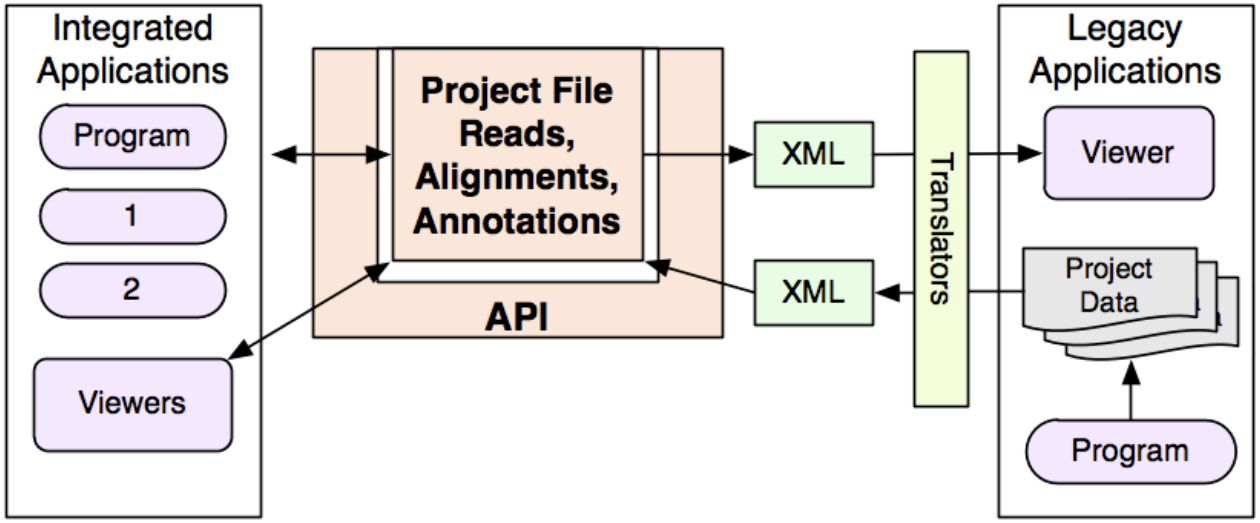


geospiza

What is Desired

1. Scalable systems with smoothly operating user interfaces
2. Summarize results and drill into details for single samples
3. Compare results between samples and within groups
4. Integrate multiple high-dimension datatypes

Data must be structured, indexed, and annotated



Need a better way to work with NGS data and information

geospiza

Current Efforts

- SAM (Sequence, Alignment, Mapping) format
  - 1000 genomes
  - Text-base format to hold alignment data
- BAM and SAMTools
  - Binary SAM
  - Designed to improve data handling performance
  - Indexes data, compression
- Useful model and first binary implementation standard (bigBed, bigWig)
- Developed as specific point solutions
- Benchmarking standard

Li H., et. al. 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Performance is Only One Requirement

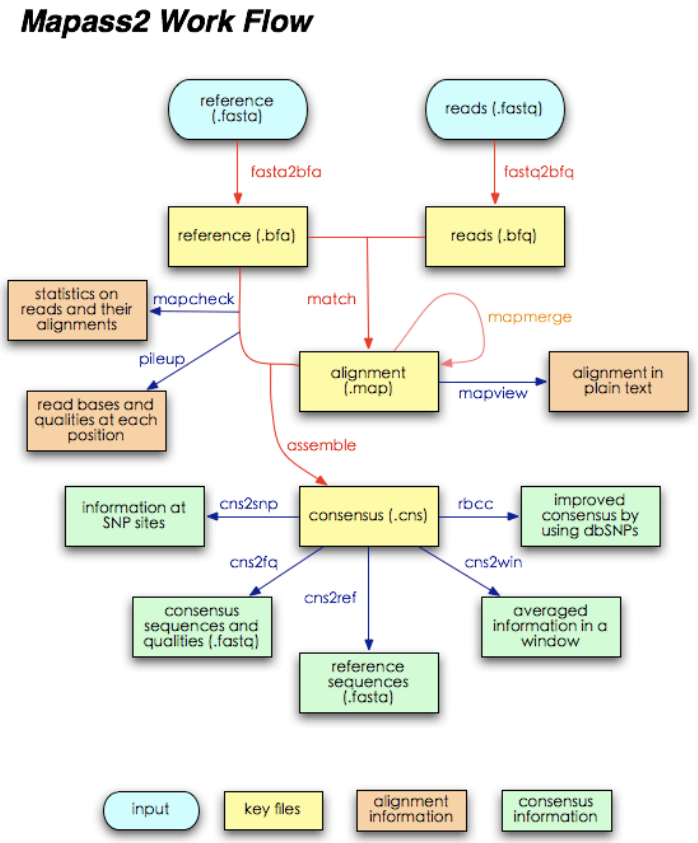
- Reduce systems complexity
  - Separate data model and implementation
  - Development and debugging tools
- Integrate diverse data and information
- Platform maturity
- Broad, stable support
- Extensible systems that can meet the unknown requirements

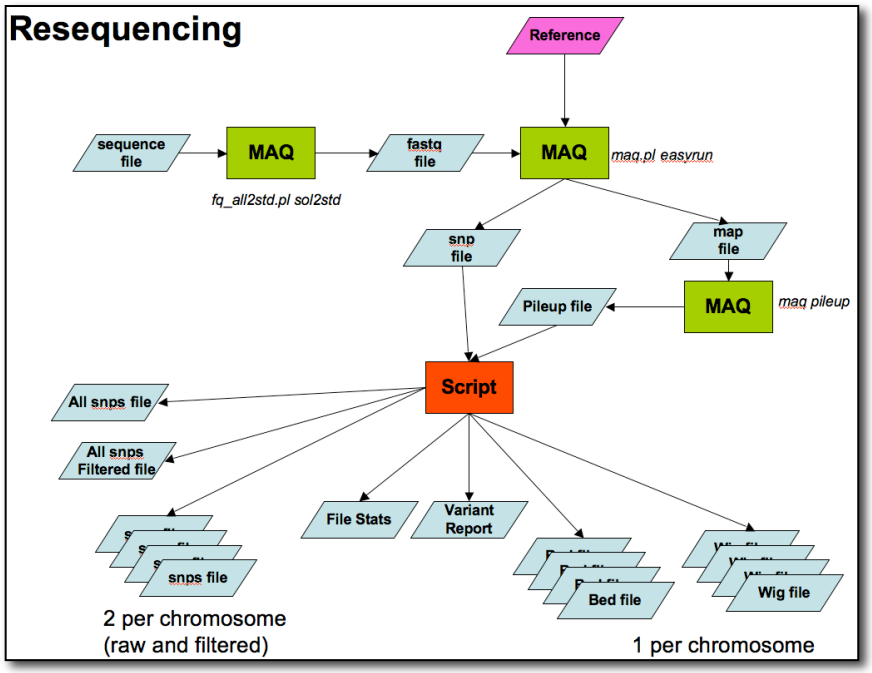
Current Efforts Address Performance

Examples: MAQ - http://maq.sourceforge.net

Secondary Analysis for:

ChIP-Seq Tag profiling Resequencing





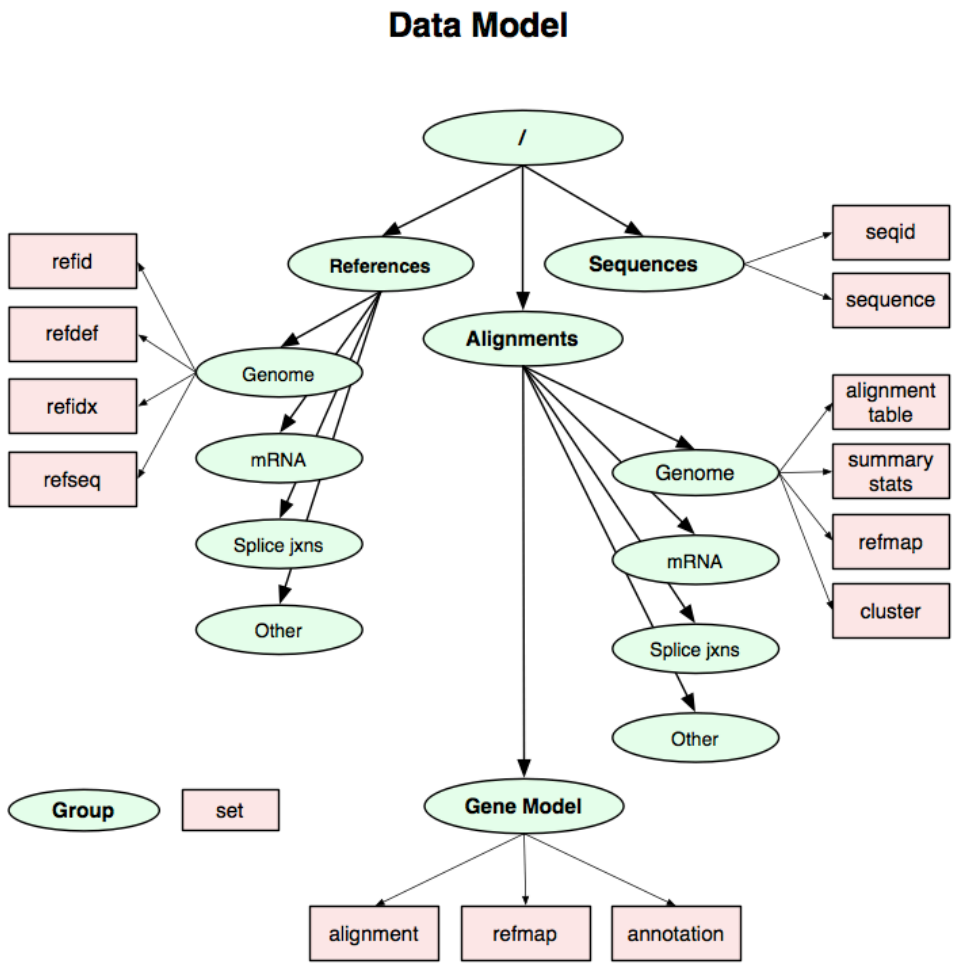
Story repeats for BWA, Bowtie, TopHat, Mapreads, SOAP …

geospiza

BioHDF Project

- **NIH STTR**
  - Geospiza, Seattle WA
  - The HDF Group, Urbana/Champaign IL
- ***Goal: Use an existing proven technology to move bioinformatics problems from organizing and structuring data to asking questions and visualizing data***
  - Develop data models and tools to work with NGS data in HDF (Hierarchical Data Format)
  - Create HDF5 domain-specific extensions and library modules to support the unique aspects of NGS data => BioHDF
  - Integrate BioHDF technologies into Geospiza products
- **Deliver core BioHDF technologies to the community as open-source software**

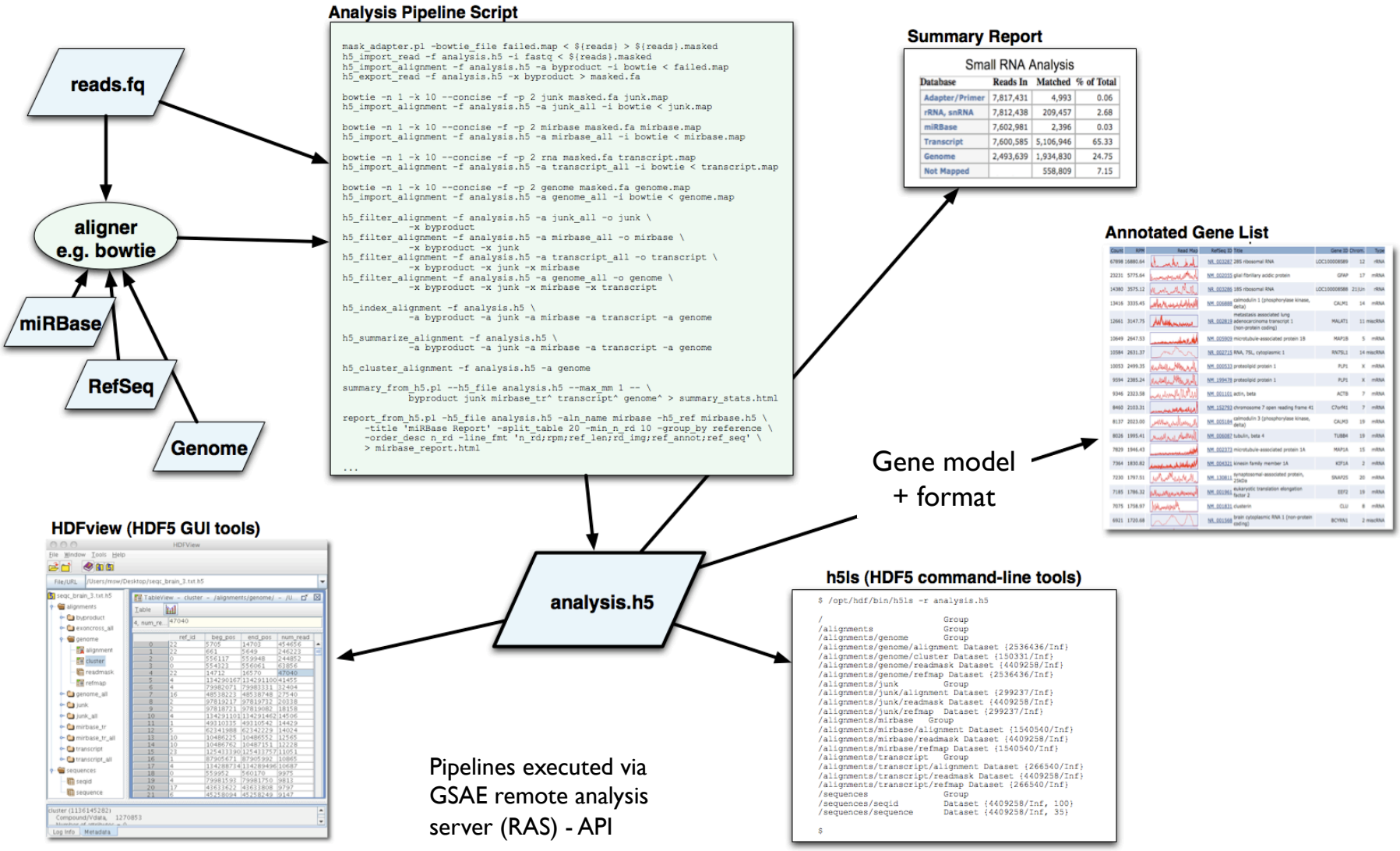
Data Model



- HDF
  - Groups
  - Datasets
  - Attributes
- BioHDF
  - Sequences
  - Reference DBs
  - Alignments
  - Gene Models (annotations)
  - Single or Multiple files

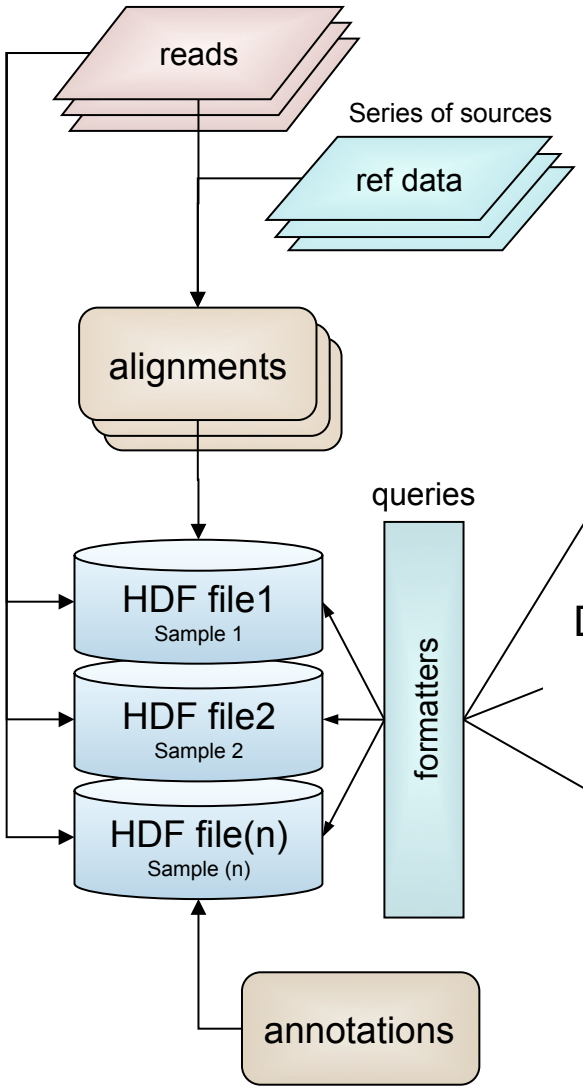
geospiza

Example Implementation



geospiza

Enables Deeper Integration: Data and Results

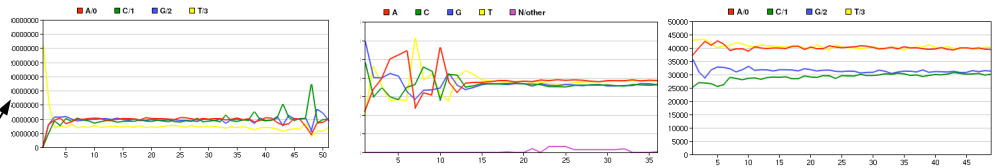


Base comp.

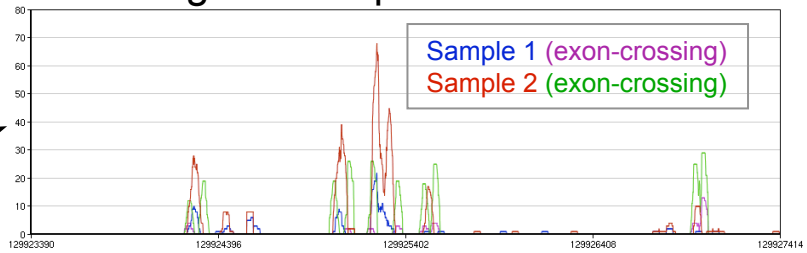
Data by range

.wig
.bed
file

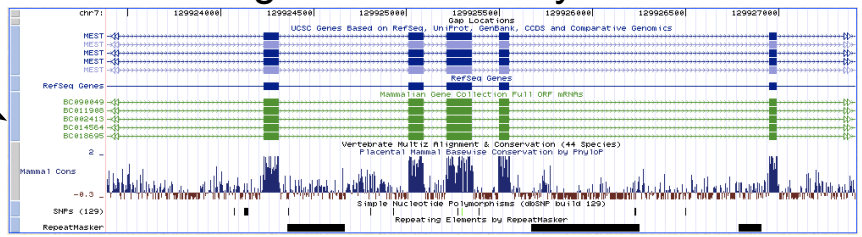
Integrate data across platforms



Integrate samples / annotations

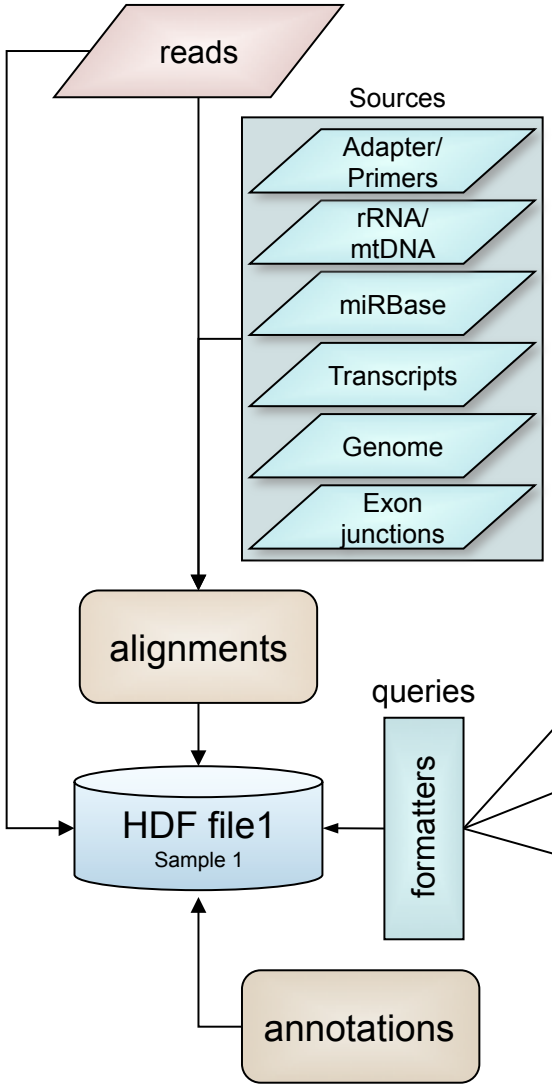


Integrate between systems



geospiza

Simplifies Data Mining



"Subtractive" question

"Biology" question

Examine match quality

One alignment step, different questions

Small RNA Analysis

Database	Reads In	Matched	% of Total
Adapter/Primer	7,817,431	4,993	0.06
rRNA, snRNA	7,812,438	209,457	2.68
miRBase	7,602,981	2,396	0.03
Transcript	7,600,585	5,106,946	65.33
Genome	2,493,639	1,934,830	24.75
Not Mapped		558,809	7.15

Splicing / Exon Analysis

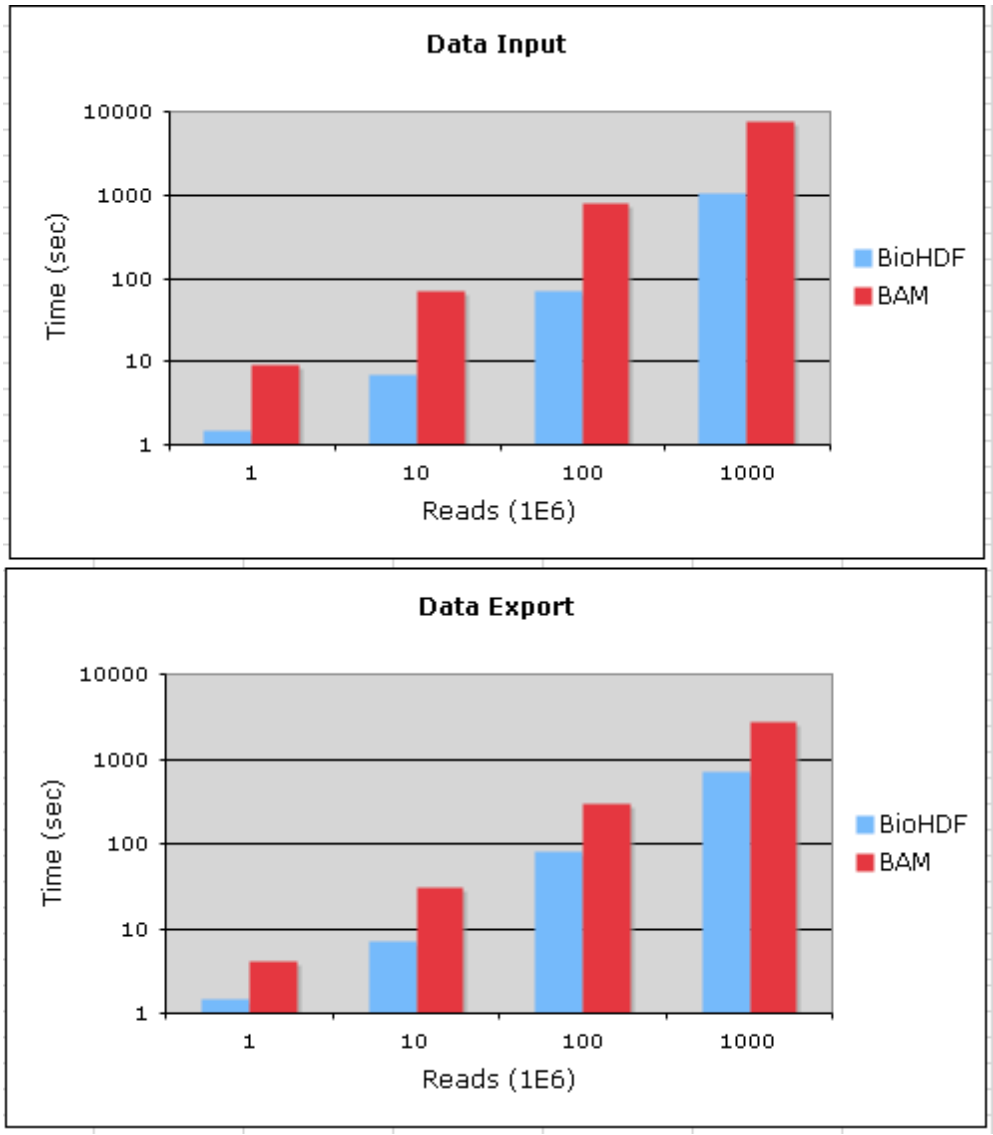
Database	Reads In	Matched	% of Total
Adapter/Primer	7,817,431	4,993	0.06
rRNA, snRNA	7,812,438	209,457	2.68
Genome	7,602,981	6,616,055	84.63
Exon Boundary	986,926	420,285	5.38
Not Mapped		566,641	7.25

Splicing / Exon Analysis

Database	Reads In	Matched	% of Total			
				0mm	1mm	2mm
Adapter/Primer	7,817,431	4,993	0.06	4,993	0	0
rRNA, snRNA	7,812,438	209,457	2.68	165,523	34,863	9,071
Genome	7,602,981	6,616,055	84.63	5,396,006	999,514	220,535
Exon Boundary	986,926	420,285	5.38	343,536	62,943	13,806
Not Mapped	7,817,431	566,641	7.25			

geospiza

HDF vs BAM Performance



- Avg. 8x import improvement
- Avg. 4x export improvement
- Improved compression
- Improved organization
- Consistent scaling

geospiza

Additional Performance

Test Case: 9.3 million GA reads aligned to HG build 36.1

(4-core 3GHz Intel Xeon)

	Flat File World	HDF5 World
fasta file	609 MB - no random access	143 MB - compressed, random access
Bowtie Alignments = fasta + alignment	1033 MB - no random access	284 MB - index, 374 MB + index
Export Alignments (ch5): 100 Mbase region 10 Mbase region 1 Mbase region 0.1 Mbase region	~1 M alignments 1470 ms 450000 alignments 735 ms 44000 alignments 735 ms 4000 alignments 735 ms 600 alignments 735 ms	~1 M alignments 1100 ms 450000 alignments 540 ms 44000 alignments 62 ms 4000 alignments 19 ms 600 alignments 15 ms
Development	> Months - develop file formats, indices, access libraries, and debug to make efficient	Days, Weeks - write I/0 code - parsers, loaders, and access methods

HDF improves storage, access, and development efficiency without adding computational overhead

geospiza

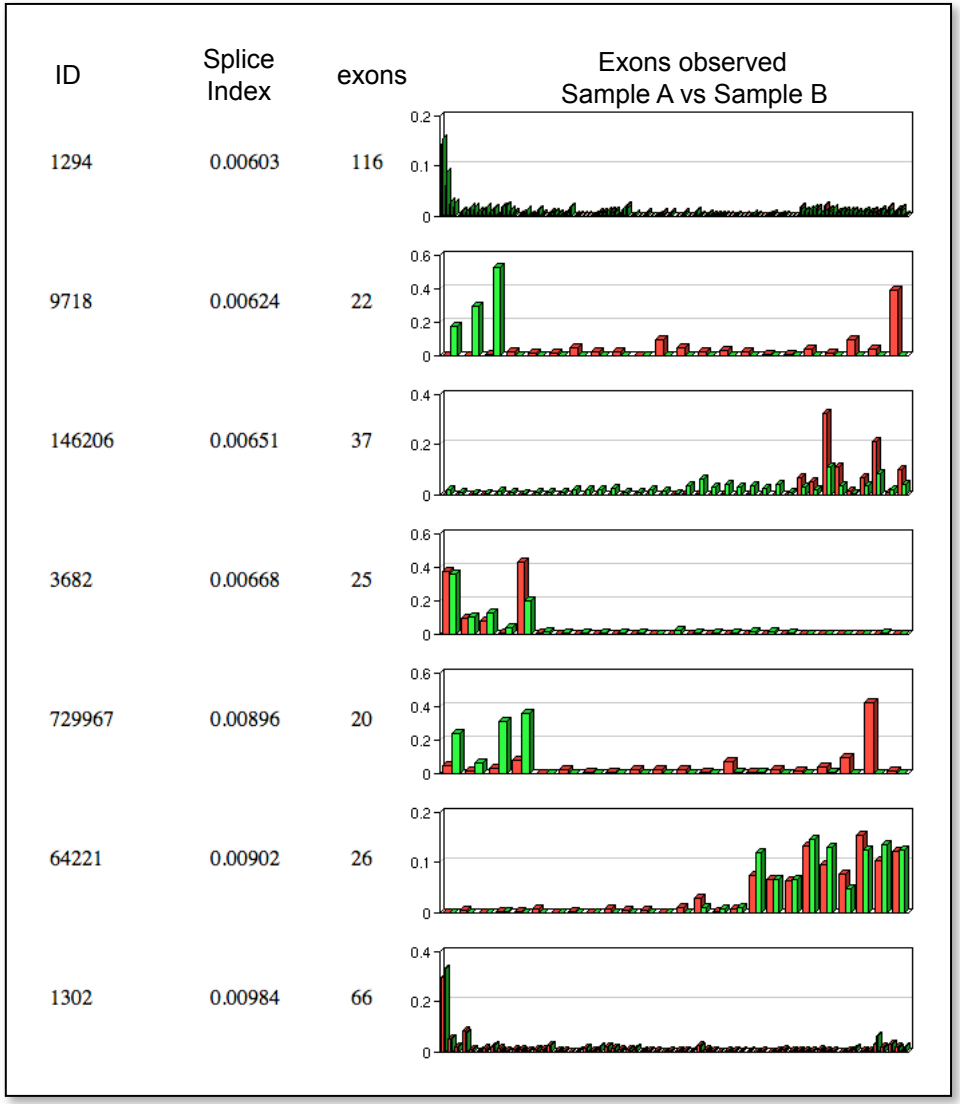
Value of Development Time

Instead of:

- Developing and debugging low level infrastructures to support “novel” binary data formats
- Optimizing high-end hardware system
- Tuning and redesigning RDBMS and other implementations

Focus on:

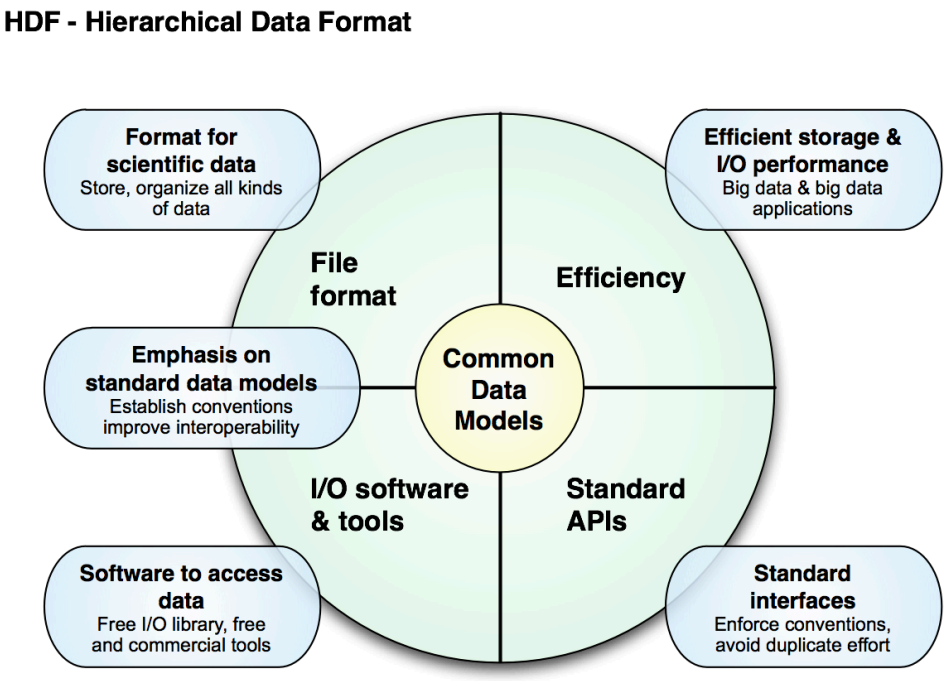
- Working with 100s of million reads for 100s of samples
- Measuring gene expression
- Identifying isoforms
- Observing sequence and structural variation
- Drilling into details from summaries



geospiza

Why HDF?

HDF: 20 Years in Physical Sciences



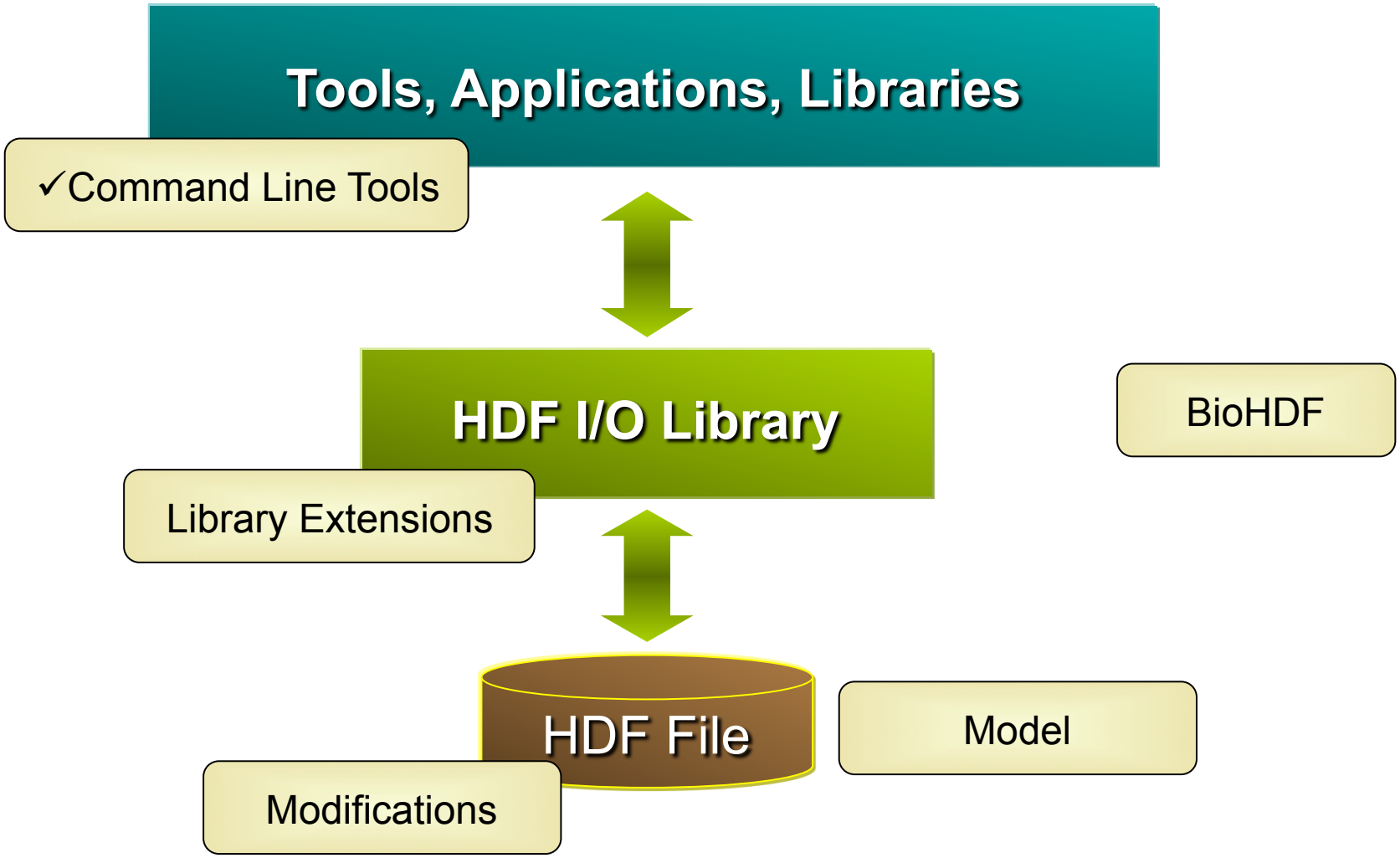
A platform for creating software to work with many kinds of *scientific* data

- ✓ Arrays, rich data types, groups accommodate every kind of data
- ✓ Store any combination of data objects in one container.
- ✓ Performance: fast random access **and** efficient, scalable storage
- ✓ Portability, data sharing: platform independent, self describing, common data models
- ✓ Tools for viewing, analysis: HDFview, MATLAB, others
- ✓ Widespread: used in academia, govt, industry - MATLAB, IDL, NASA-Earth Observing System

HDF5 Platform

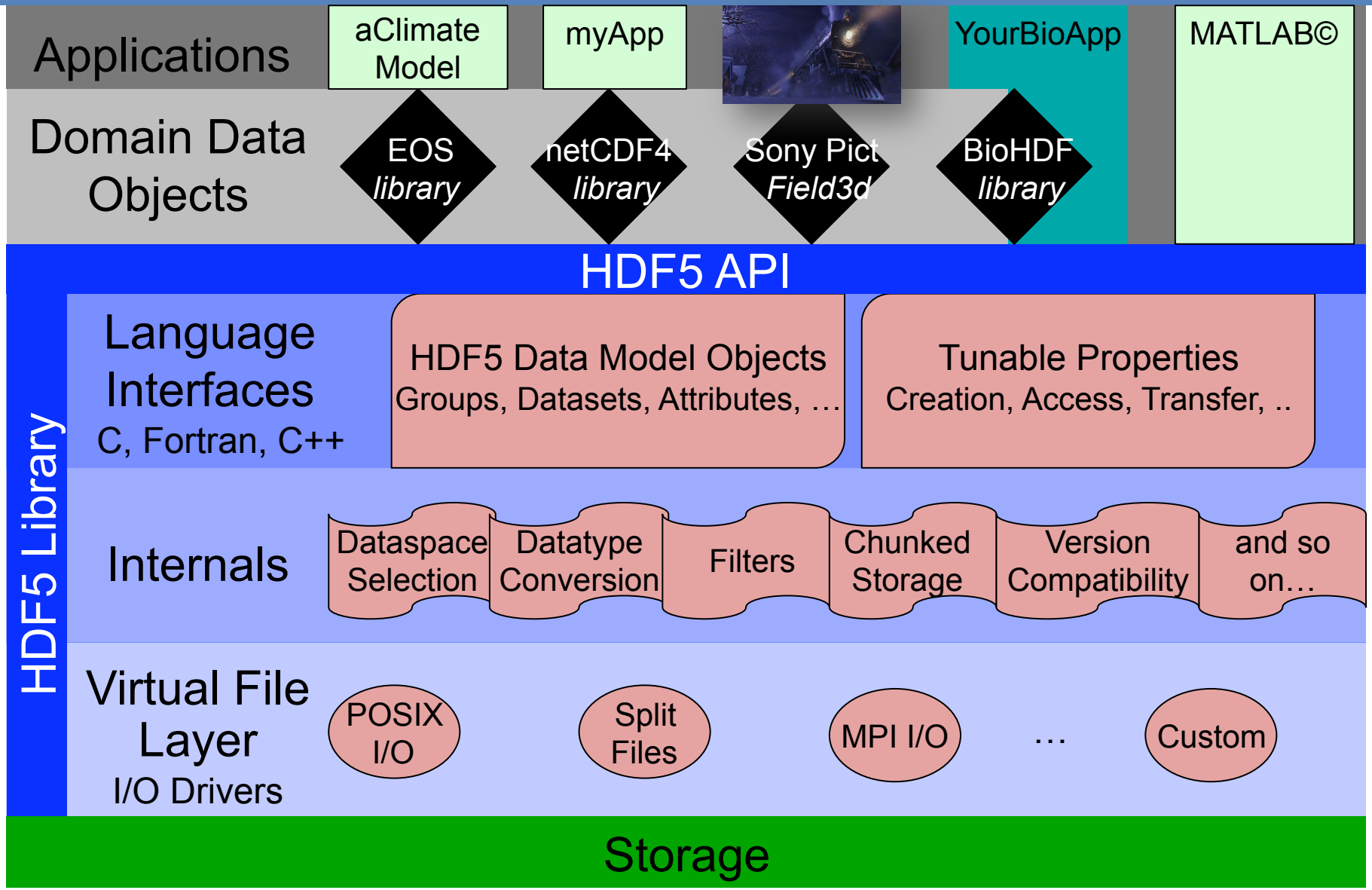
- HDF5 Abstract Data Model
  - Defines the “building blocks” for data organization and specification
  - Files, Groups, Links, Datasets, Attributes, Datatypes, Dataspaces
- HDF5 Software
  - Tools
  - Language Interfaces
  - HDF5 Library
- HDF5 Binary File Format
  - Bit-level organization of HDF5 file
  - Defined by HDF5 File Format Specification

HDF Software



geospiza

HDF5 API and Applications



Prepared to Manage Systems Biology Data

practice

DOI:10.1145/1562764.1562781

Article development led by queue.acm.org

The biosciences need an image format capable of high performance and long-term maintenance. Is HDF5 the answer?

BY MATTHEW T. DOUGHERTY, MICHAEL J. FOLK, EREZ ZADOK, HERBERT J. BERNSTEIN, FRANCES C. BERNSTEIN, KEVIN W. ELICEIRI, WERNER BENGER, CHRISTOPH BEST

Unifying Biological Image Formats with HDF5

THE BIOLOGICAL SCIENCES need a generic image format suitable for long-term storage and capable of handling very large images. Images convey profound ideas in biology, bridging across disciplines. Digital imagery began 50 years ago as an obscure technical phenomenon. Now it is an indispensable computational tool. It has produced a variety of incompatible image file formats, most of which are already obsolete.

Several factors are forcing the obsolescence: rapid increases in the number of pixels per image; acceleration in the rate at which images are produced; changes in image designs to cope with new scientific instrumentation and concepts; collaborative requirements for interoperability of images collected in different labs on different instruments; and research metadata dictionaries that must support frequent and rapid extensions. These problems are not unique to the biosciences. Lack of image standardization is a source of delay, confusion, and errors for many scientific disciplines.

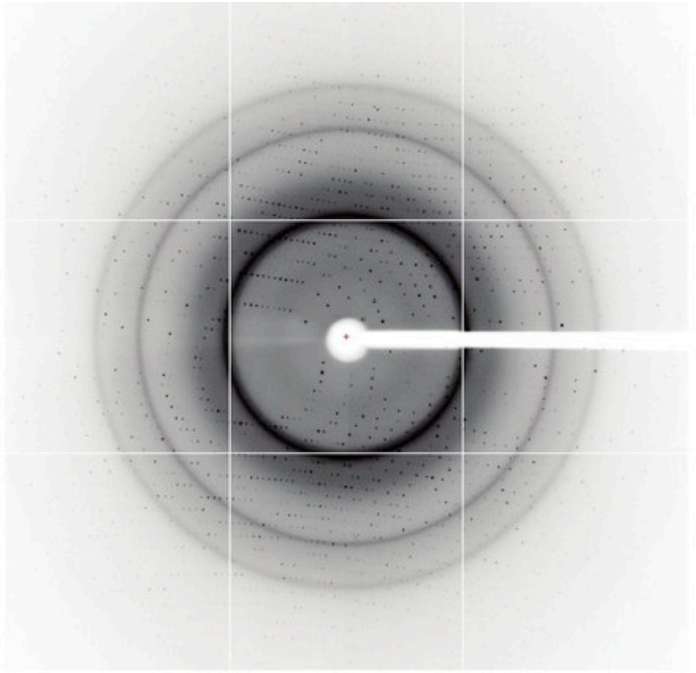
There is a need to bridge biological and scientific disciplines with an image framework capable of high computational performance and interoperability. Suitable for archiving, such a framework must be able to maintain images far into the future. Some frameworks represent partial solutions: a few, such as XML, are primarily suited for interchanging metadata; others, such as CIF (Crystallographic Information Framework),[2] are primarily suited for the database structures needed for crystallographic data mining; still others, such as DICOM (Digital Imaging and Communications in Medicine),[1] are primarily suited for the domain of clinical medical imaging.

What is needed is a common image framework able to interoperate with all of these disciplines, while providing high computational performance. HDF (Hierarchical Data Format)[6] is such a framework, presenting a historic opportunity to establish a coin of the realm by coordinating the imagery of many biological communities. Overcoming the digital confusion of incoherent bio-imaging formats will result in better science and wider accessibility to knowledge.

Semantics: Formats, Frameworks, and Images

Digital imagery and computer technology serve a number of diverse biological communities with terminology differences that can result in very different perspectives. Consider the word *format*. To the data-storage community the hard-drive format will play a major role in the computer performance of a community's image format, and to some extent, they are inseparable. A format can describe a standard, a framework, or a software tool; and formats can exist within other formats.

Image is also a term with several uses. It may refer to transient electrical signals in a CCD (charge-coupled device), a passive dataset on a storage device, a location in RAM, or a data structure written in source code. Another example is *framework*. An image framework might implement an image standard, resulting in image files created by a software-imaging tool. The framework, the standard, the files, and the tool, as in the case of HDF,[6] may be so interrelated that they represent different facets of the same specification. Because these terms are so ubiquitous and varied due to perspective, we shall use them interchangeably, with the emphasis on the storage and management of pixels throughout their lifetime, from acquisition through archiving.



An x-ray diffraction image taken by Michael Soltis of LSAC on SSRL BL9-2 using an ADSC Q315 detector (SN901).

Hierarchical Data Format Version 5

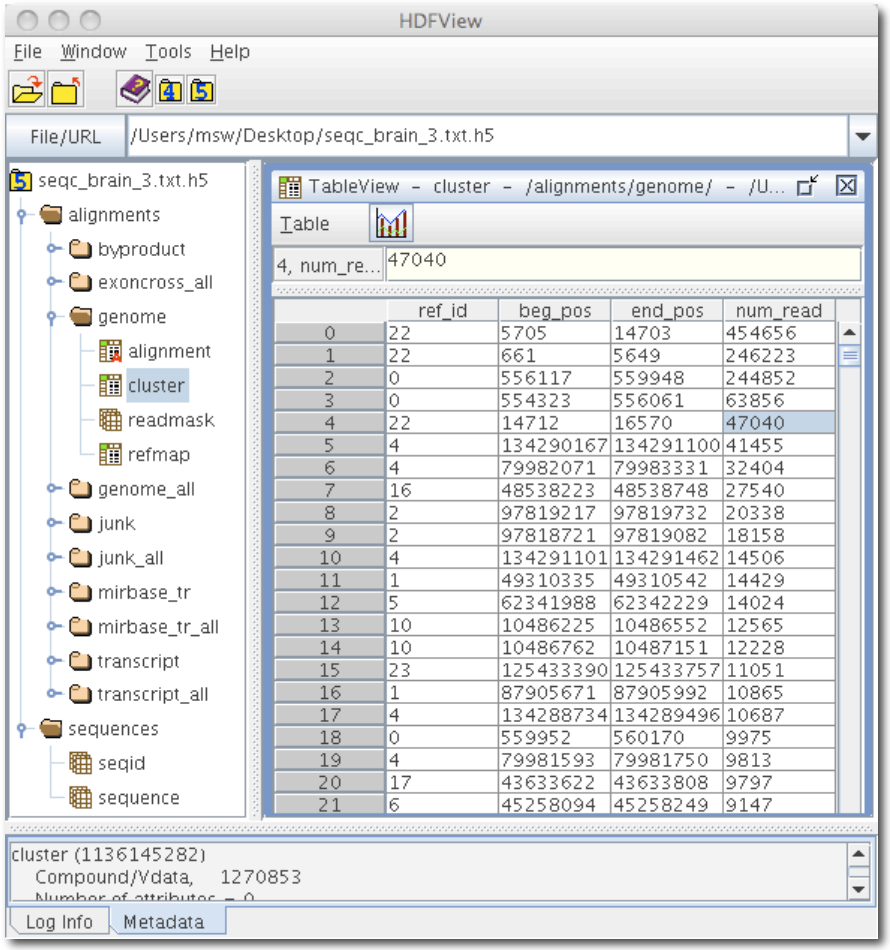
HDF5 is a generic scientific data format with supporting software. Introduced in 1998, it is the successor to the 1988 version, HDF4. NCSA (National Center for Supercomputing Applications) developed both formats for high-performance management of large heterogeneous scientific data. Designed to move data efficiently between secondary storage and memory, HDF5 translates across a variety of computing architectures. Through support from NASA (National Aeronautics and Space Administration), NSF (National Science Foundation), DOE (Department of Energy), and others, HDF5 continues to support international research. The HDF Group, a nonprofit spin-off from the University of Illinois, manages HDF5, reinforcing the long-term business commitment to maintain the format for purposes of archiving and performance.

Because an HDF5 file can contain almost any collection of data entities in a single file, it has become the format of choice for organizing heterogeneous collections consisting of very large and complex datasets. HDF5 is

COMMUNICATIONS OF THE ACM – October 2009

Benefits

- Separates the model, implementation, and view of the data
- Combines data from multiple samples
- Compression and other performance advantages
- Rapid prototyping environment
- *Significant reduction in development time*
- *Approaching problems differently*

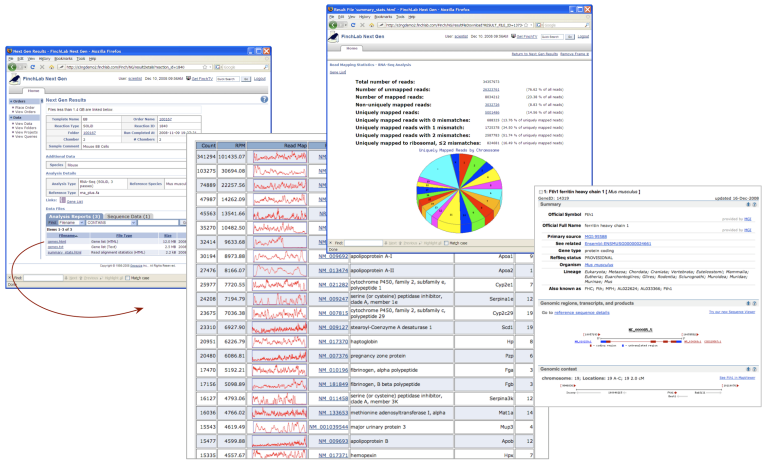


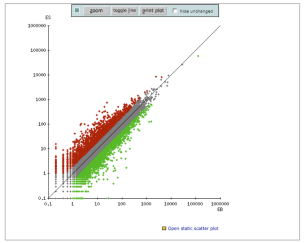
Only had to define the model, data importers, and export tools

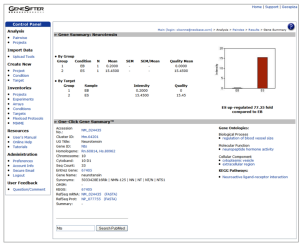
geospiza

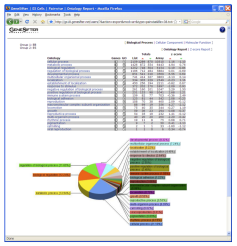
Geospiza - Bioinformatics since 1997

GeneSifter™ Laboratory and Analysis Software Systems









- **For:** Core, Service, Data Production Laboratories and Research Scientists
- **Working with:** Sanger Sequencing, Microarray, Next Generation Sequencing, and (or) other platforms
- **GeneSifter supports:** Laboratory operations, Data Management, Multiple Levels of Data Analysis
- **Delivered cost effectively:** through hosted and on-site delivery models.

From Samples to Results™

geospiza

Next Steps

More information, getting involved

Contact: todd@geospiza.com

Dana Robinson derobins@hdfgroup.org

http://www.biohdf.org

geospiza