

BUSINESS TECHNOLOGY



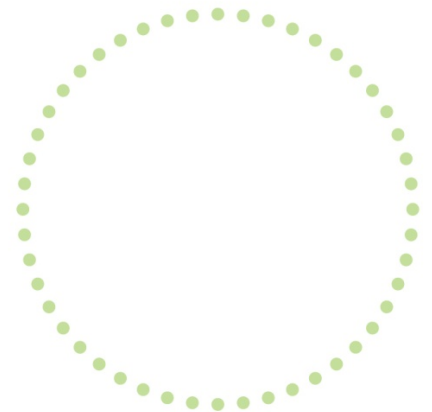
Pfizer Informatics for NGS

Giles Day & Adam Kraut
Pfizer BioTherapeutics R&D – BioTeam, Inc.
XGEN Mar 15th 2010



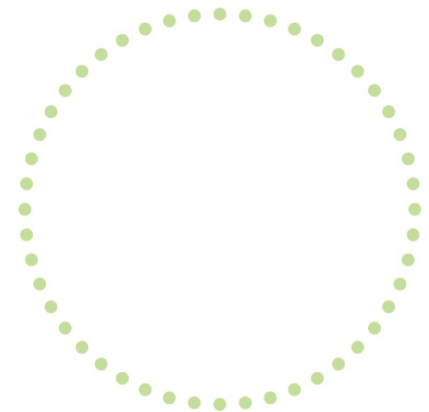


- Small biotech purchased by Pfizer 2006
- Located in South San Francisco
- Antibody Therapeutics
- 454 Sequencing Team co-located
- Approx 100 people





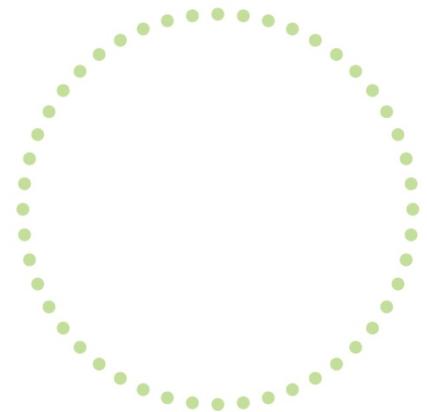
- Capture of knowledge & decisions
 - Why did we do that, who did that, when did they do it
- Integration of structured & un-structured information
- Discovery of relevant information
- Foster collaboration



Overview of NGS Strategy



- New target discovery
 - GWAS
 - In-house & external
- Support of existing therapeutic programs
 - Phage display library analysis
 - Patient stratification



Solutions Come and Go



- In-house relational platforms
- SharePoint
- Semantic Solutions
- Google/Search Devices
- Pathway Platforms



What They Are



- Inflexible
 - Unable to adapt to changing research processes & organisations
 - Can never capture everything
 - Difficult to administer
- Culture Defiant
 - I'm not ready to share
 - What benefit do I get
 - It's not my solution
- Traditional
 - Software Development cycles are too long – even agile ones
 - Too expensive

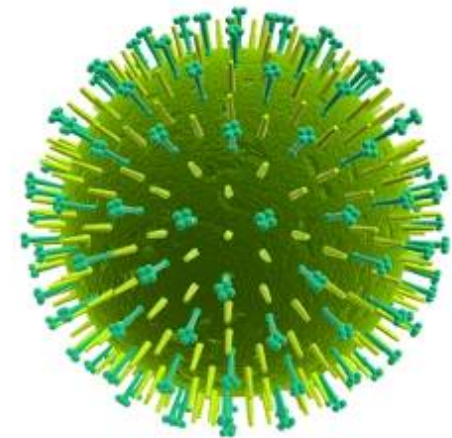


What They Need To Be



- Useful return more benefit than cost of construction
- Intuitive low adoption barrier
- Fast rapid prototyping
- Flexible modular, able to retool to new uses
- Talkative communicates easily with other technologies
- Cheap design cycle not limited by cost of development
- Open open source, transparent & easily modified
- Adopted Community of developers expanding functionality

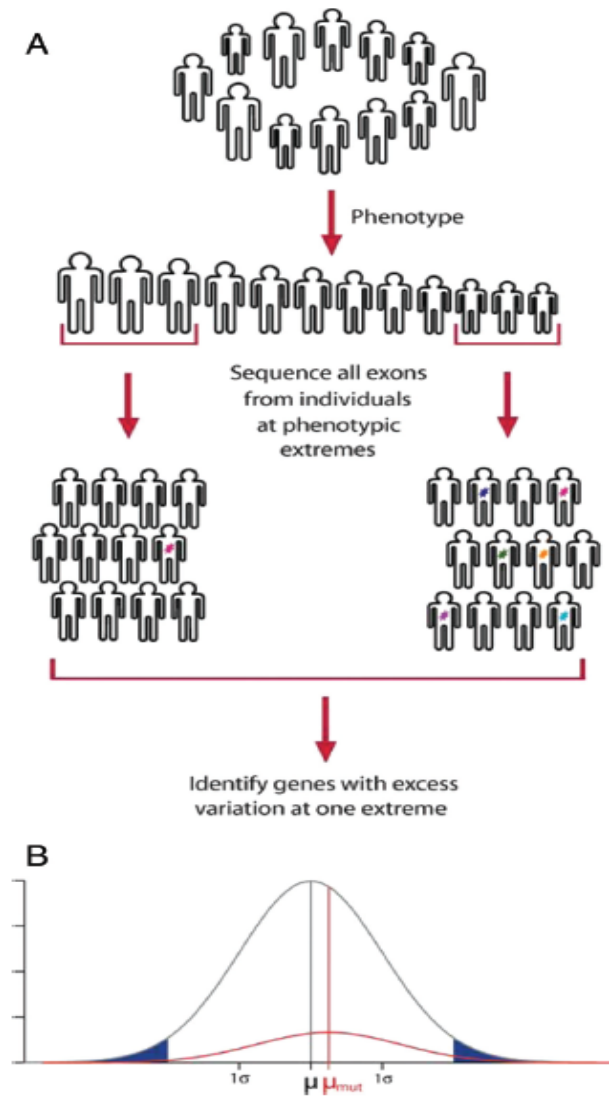
Informatics must co-evolve with scientific innovation



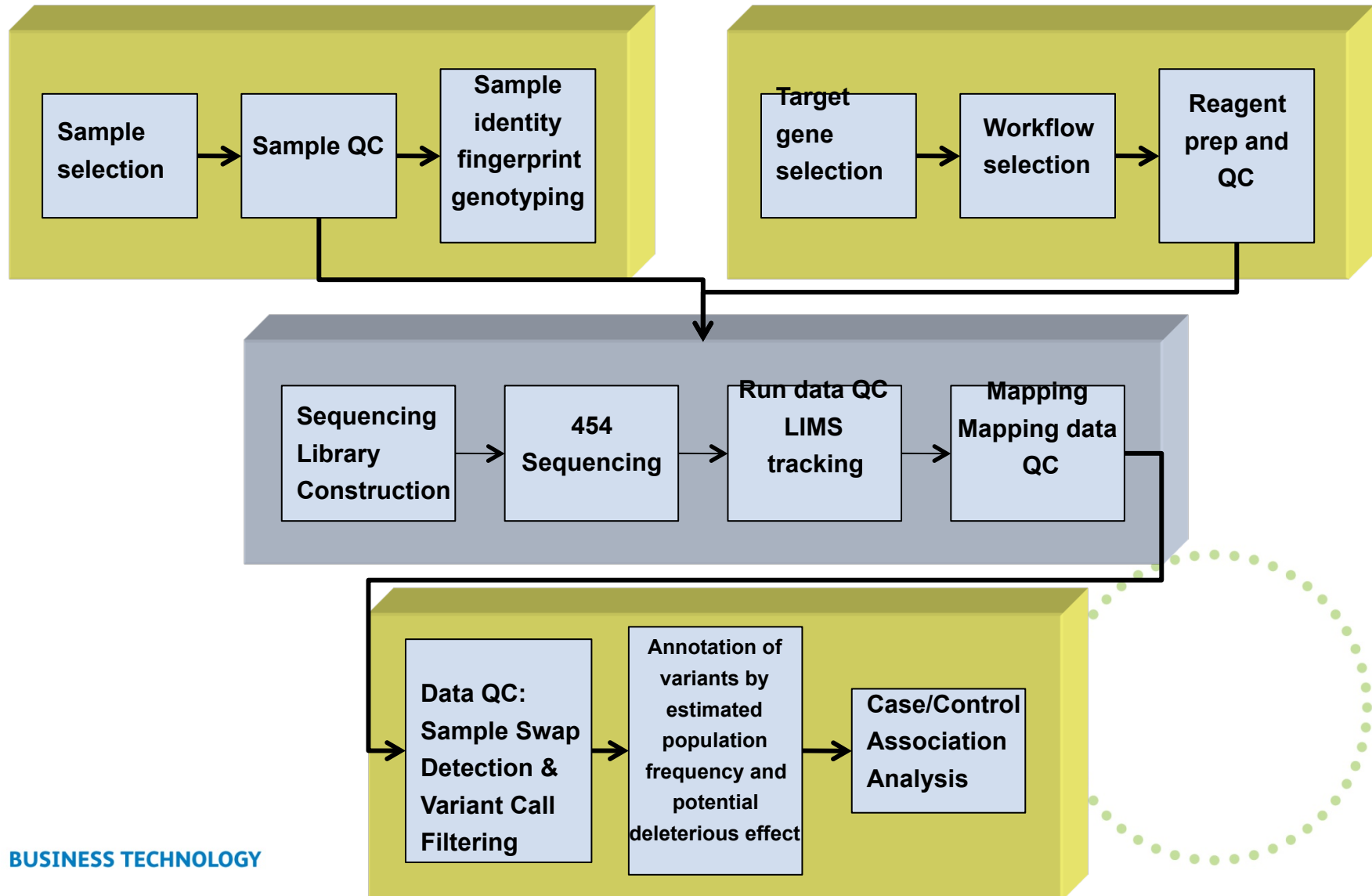
How to Translate Bases to Clinical Outcomes?



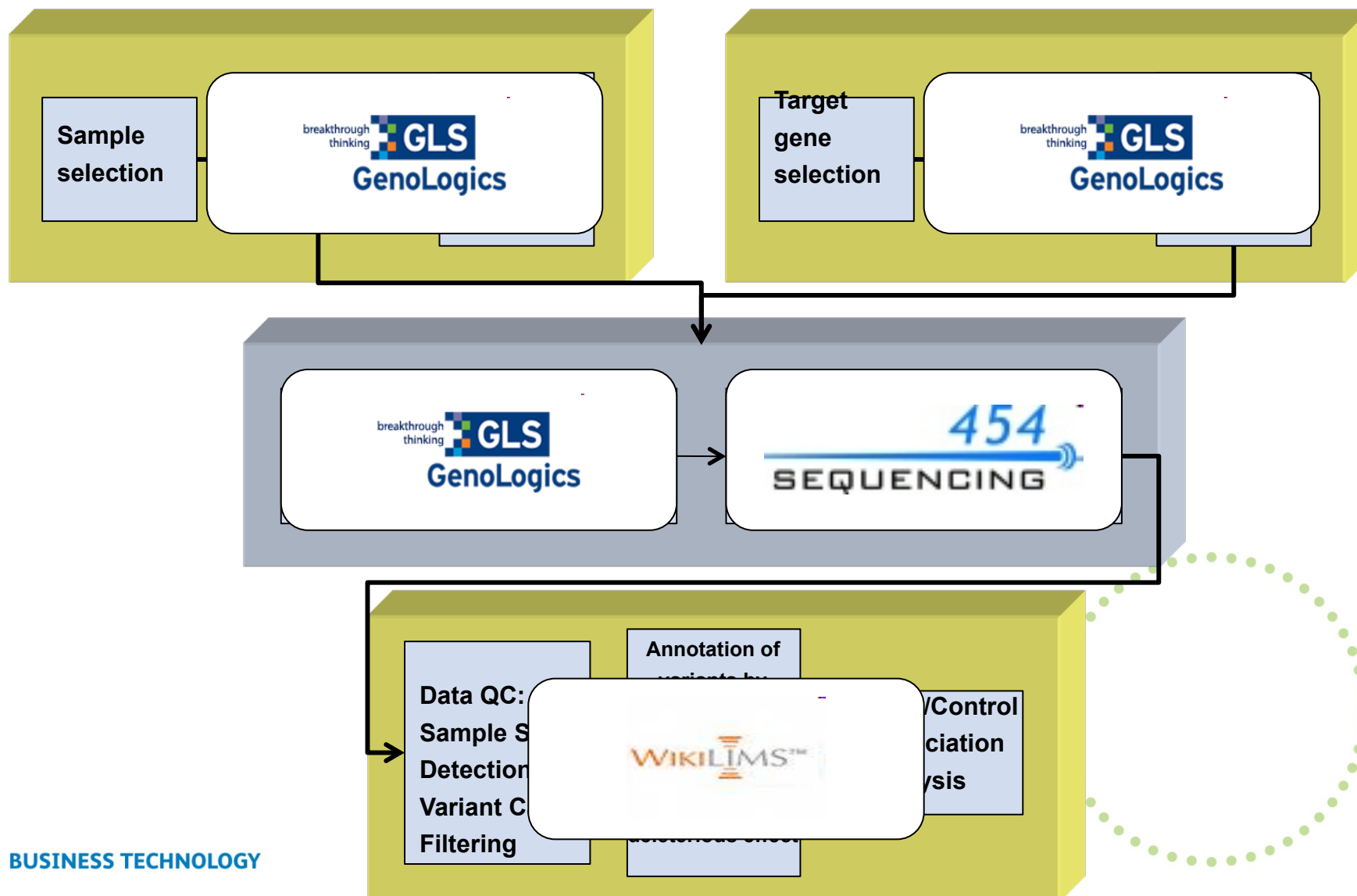
- GWAS Approach
 - 100s to 1000s samples
 - Well characterized phenotypes
 - Sequence genes/genomes in extreme phenotypes
 - Sequence deep to discover variants
 - Rare loss / gain of function variants
 - Small (additive) effects of many variants expected
 - Annotation
 - Validation of putative SNPs on larger cohorts
 - Identify drug targets
 - Patient stratification, Dx



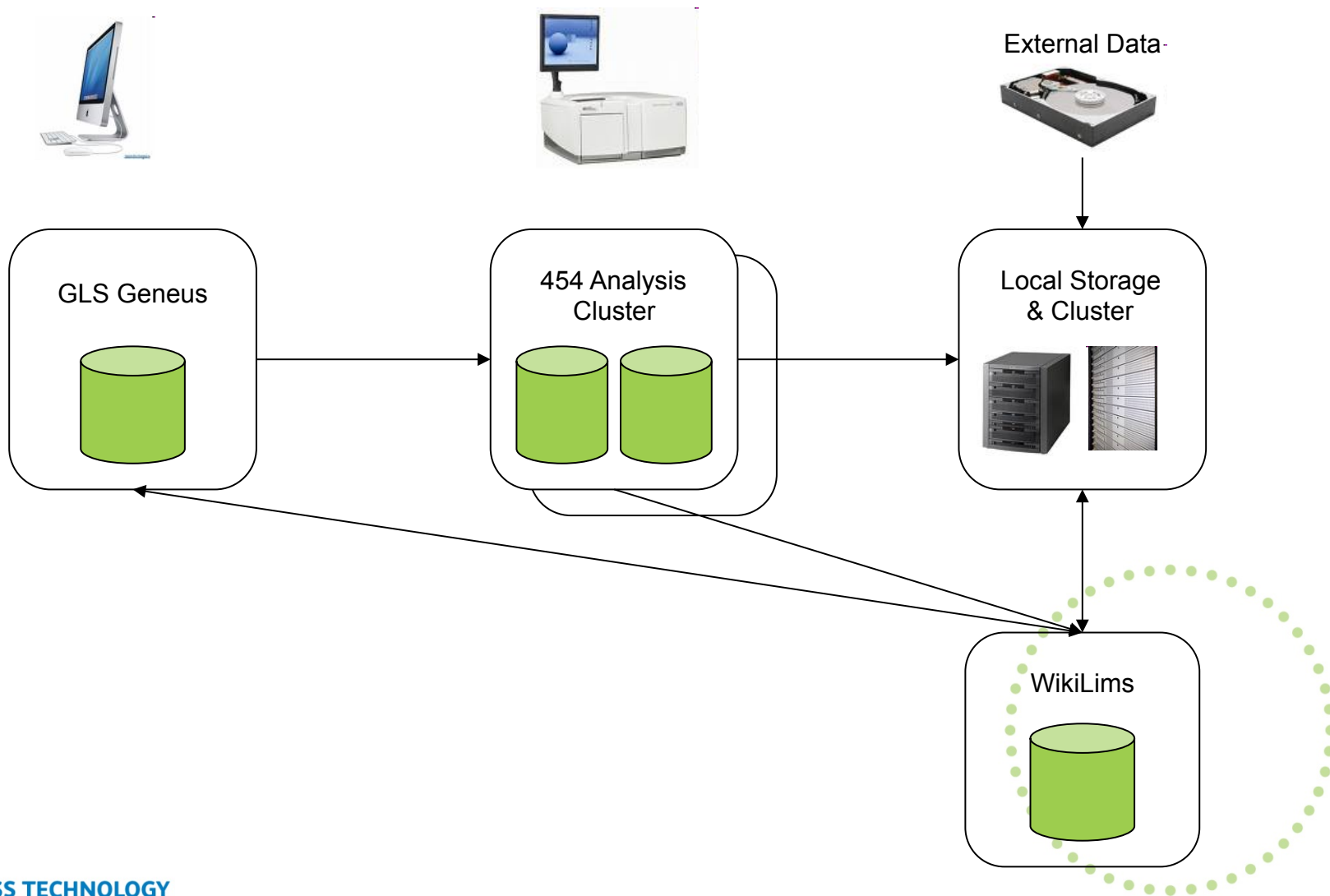
Overview of NGS Workflow



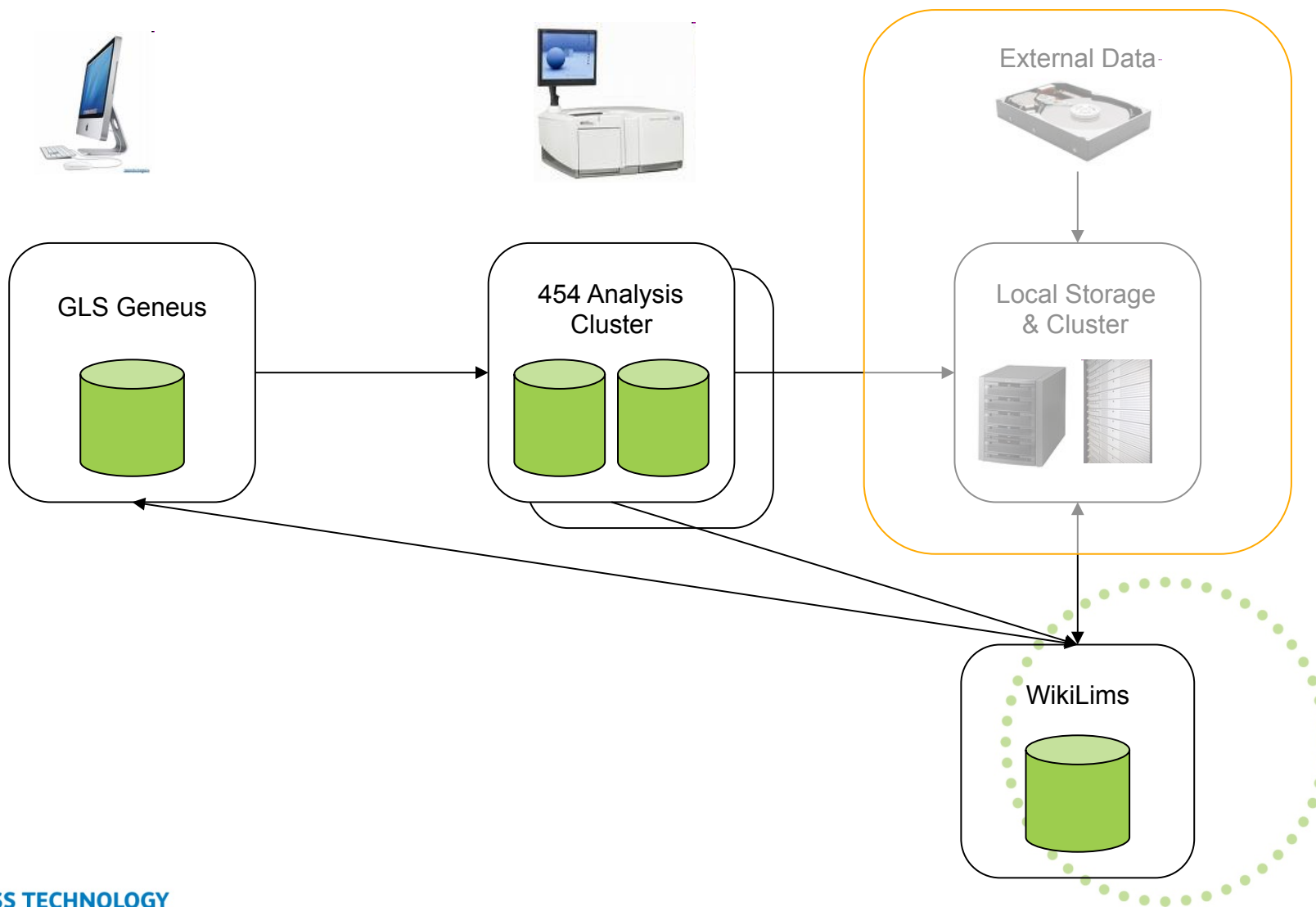
Informatics Solutions



Informatics Infrastructure

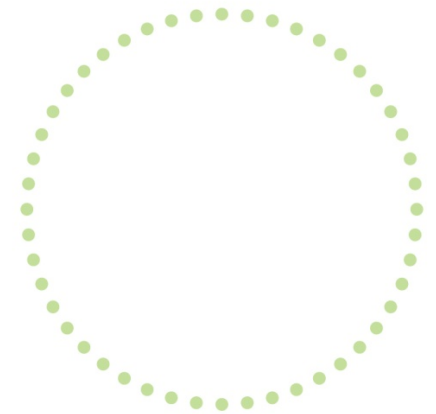


Future Directions

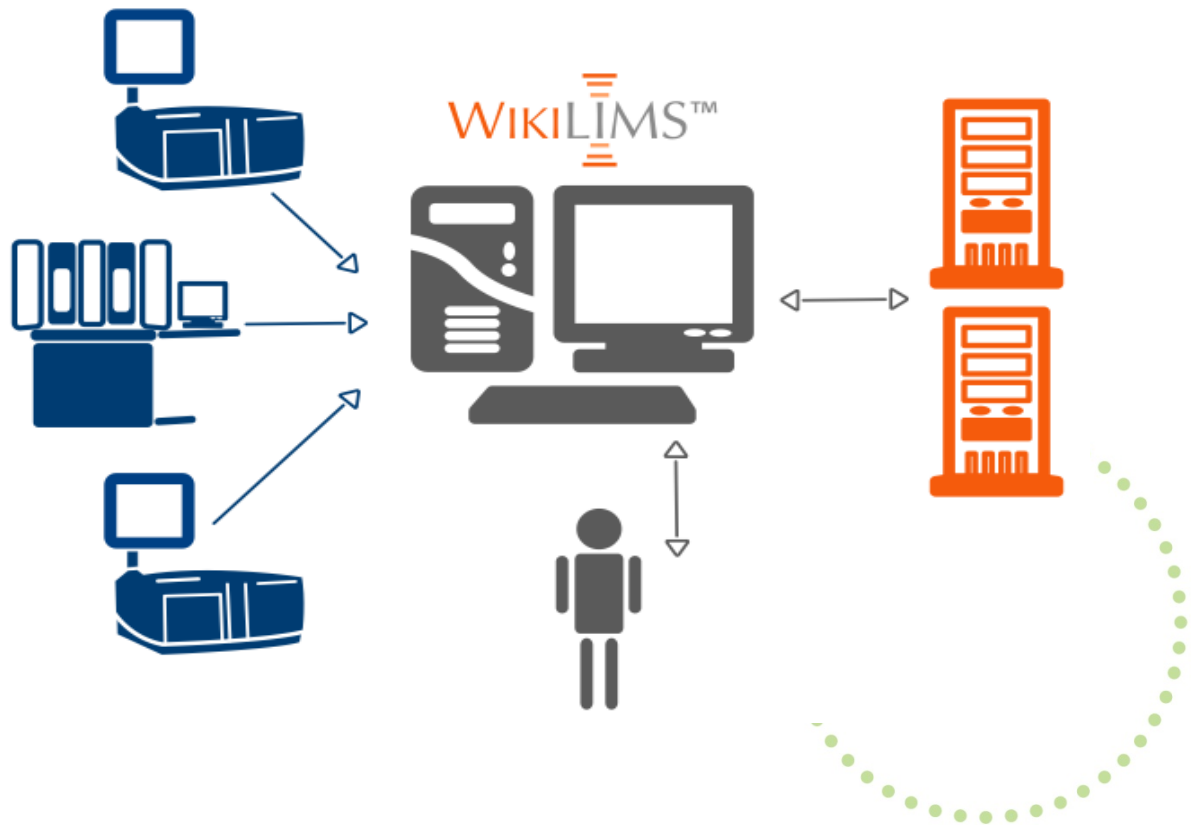




The Details



- A system that is quickly constructed and easily revised
- A system that can handle many data types
- A modular and transparent system
- A Web-based system



Simplified tracking information



← Older edit

Line 5:

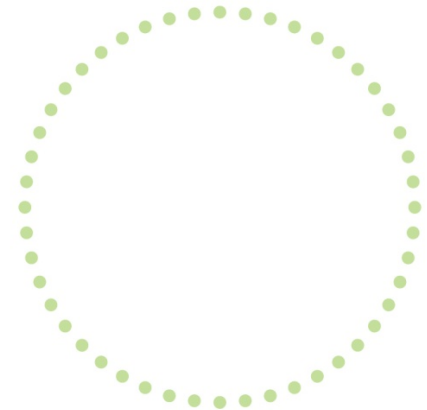
```
ILibrary type=pool
ILab member assigned to prepare library=Bioteam

}}
```

Line 5:

```
ILibrary type=pool
ILab member assigned to prepare library=Bioteam
+ ISample receipt confirmed=Yes
+ ISample QA confirmed=No
}}
```

- Email and RSS notifications for every step in workflow
- Wiki Revision Control explains **who** did **what** and **when**
- Lab Managers can revert and undo tasks



Monitoring Sequencing Runs– Calendar view



2009-10-23	UCSF01_01_Run5
2009-10-20	UCSF01_01_Run4

Showing 1 to 10 of 30 entries

First Previous 1 2 3 Next Last

January 2010

◀ Today ▶

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14 RDT01 03 Run1	15	16
17	18 Run01 03 Run2	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6



Intelligent Annotation: Leveraging Web Services



External Data

Annotations

AL3A2 HUMAN

WikiSysop my talk admin links my preferences my watchlist my contributions log out

page discussion edit history delete move protect watch refresh

AL3A2 HUMAN

Name(s) Fatty aldehyde dehydrogenase

Domain Transmembrane

Function Oxidoreductase

Organism Homo sapiens (Human)

Catalytic activity An aldehyde + NAD(+) + H(2)O = an acid + NADH.

Length 485 AA

Uniprot <http://www.uniprot.org/uniprot/P51648>

Sequence

MELEVRRVRQAFLSGRSRPLRFLRQQLALRRMVQEREKDLTAIAADLCKSEFNVSQE
VITVLGEIDFMLENLPEWVTAKPVKKNVLTMLDEAYIQPQLGVLIIGAWNYPFVLTIQ
PLIGAIAAGNAVIIKPSSESENTAKILAKLLPQYLDQDLIVINGGVEETTELLKQRFDH
IFYTGNTAVGKIVMEAAKHLTPVTLELGGKSPCYIDKDCDLIVCRITWGKYMNCGQT
CIAPDYILCEASLQNIQVWKIKETVKEFYGENIKESPDYERIIINLRHFKRILSLLEGQKI
AFGGGETDEATRYIAPTFLTVDVDPKTKVMQEEIFGPILPIVPVKNVDEAINFINEREKPLA
LYVFSHNHKLKRMIDETSSGGVTGNDVIMHFTLNSFPFGVGSSGMGAYHGKHSFDTFS
HQRPCLLKSLKREGANKLRYPPNSQSKVDWGKFFLLKRFNKEKLGLLLLTFLGIVAVALV
KAEYY

Category: Proteins

Combining raw data and reporting



UCSF01 01 Run6

Date: 2009-10-26
Chemistry: FLX

▼ D_2009_10_27_02_28_10_SSF-FLX-cluster_fullProcessingAmplicons

View Files

Lane Number	Project Phase Name	Library Name	Copies per Bead	Bead Count	Total Raw Wells	Total Keypass Wells	Total Keypass %	Con Ave Len
1	UCSF01.01	UCSF01.01_SFLP011	0.5	750,000	444,624	437,127	98.3	2
2	UCSF01.01	UCSF01.01_SFLP012	0.5	750,000	482,013	475,751	98.7	2
Total					926,637	912,878	98.5	2

Category: 454Runs

Name	Last modified	Size	Description
Parent Directory		-	
1.ATG.454Reads.fna	27-Oct-2009 01:24	3.0M	
1.ATG.454Reads.qual	27-Oct-2009 01:24	7.7M	
1.TCA.454Reads.fna	27-Oct-2009 01:24	74M	
1.TCA.454Reads.qual	27-Oct-2009 01:24	190M	
2.ATG.454Reads.fna	27-Oct-2009 01:24	2.3M	
2.ATG.454Reads.qual	27-Oct-2009 01:24	6.0M	
2.TCA.454Reads.fna	27-Oct-2009 01:24	72M	
2.TCA.454Reads.qual	27-Oct-2009 01:24	185M	
454BaseCallerMetrics.csv	27-Oct-2009 01:24	6.4K	
454BaseCallerMetrics.txt	27-Oct-2009 01:24	20K	
454DataProcessingDir.xml	27-Oct-2009 01:24	351	
454QualityFilterMetrics.csv	27-Oct-2009 01:24	954	
454QualityFilterMetrics.txt	27-Oct-2009 01:24	2.1K	
454RuntimeMetricsAll.csv	27-Oct-2009 01:24	8.3K	
454RuntimeMetricsAll.txt	27-Oct-2009 01:24	17K	
DM_2009_10_27_14_37_03-F4PLOWH01/	27-Oct-2009 12:57	-	
DM_2009_10_27_14_39_14-F4PLOWH02/	27-Oct-2009 12:57	-	
P_2009_10_27_12_39_37_mapping_target_set_1-F4PLOWH01/	27-Oct-2009 09:59	-	
P_2009_10_27_12_49_07_mapping_target_set_1-F4PLOWH02/	27-Oct-2009 11:05	-	
SigProcRunBackupComplete	27-Oct-2009 01:25	0	
dataRunParams.xml	27-Oct-2009 01:24	6.0K	
gsRunProcessor.log	27-Oct-2009 01:24	9.2K	

Simple Integrations



[page](#) [discussion](#) [view source](#) [history](#)

RIN06 02 001

Date: 2010-01-29
Chemistry: Titanium

▼ D_2010_01_30_01_35_18_SSF-FLX-cluster_fullProcessing

View Files

Lane Number	Project Phase Name	Library Name	Copies per Bead	Bead Count	Total Raw Wells	Total Keypass Wells	Total Keypass %	Control Average Length	Control Keypass Wells	Controls with 98% accuracy over 200 bp %	Control Pass Filter %	Control Mixed %	C
1					995,379	983,282	98.8	443.6	8,427	93.0	94.8	0.8	
2					1,033,895	1,023,075	99.0	440.7	8,417	90.5	91.8	0.8	
Total					2,029,274								

[page](#) [discussion](#) [view source](#) [history](#)

View source

for RIN06 02 001

You do not have permission to edit this page, for the following reason:
The action you have requested is limited to users in the group: **Users**.
You can view and copy the source of this page:

```
{{454Run|id=33|run_dt=0000-00-00}}  
  
== Primer Design ==  
  
[[Image:primer.design.titanium.png]]  
  
== 6Mer MIDs ==  
  
<PRE>  
+-----+-----+  
| name   | sequence |  
+-----+-----+  
| PFMID1 | ACGTGT   |  
| PFMID2 | ACTCTC   | **  
| PFMID3 | AGATCG   |  
| PFMID4 | AGTGAC   | **
```

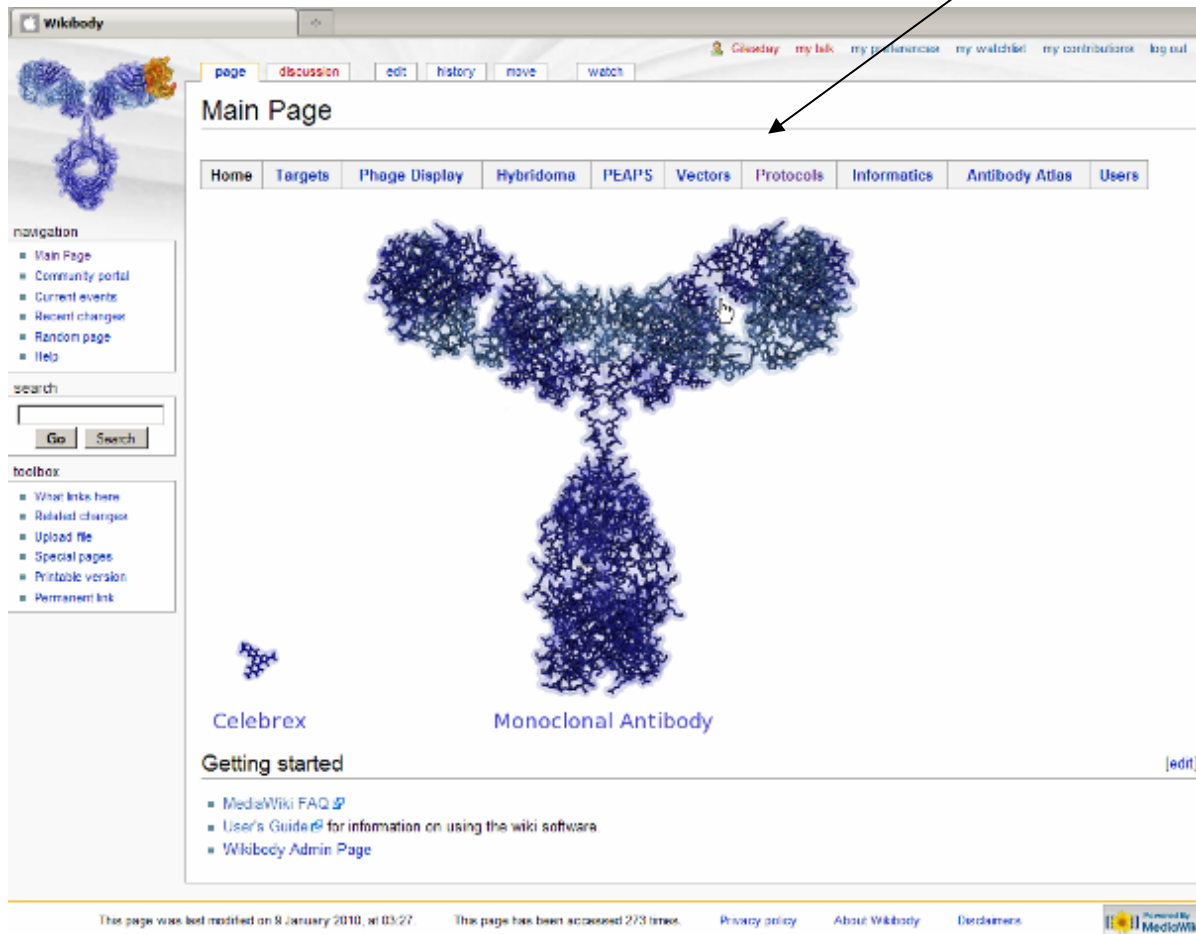
Data retrieved from database via 1 line of code

Getting Started

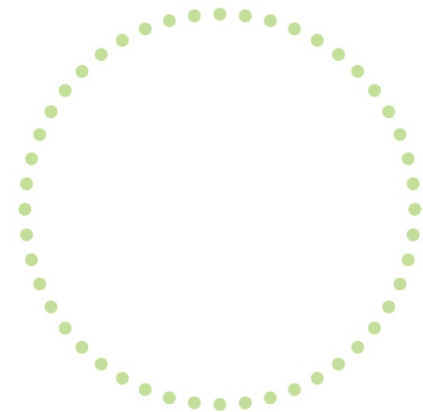


Familiar Wiki layout

Tabs separate concept areas



Supports all areas for
Rinat research



Project Overview



Targets - Wikibody

Gilesday my talk my preferences my watchlist my contributions

page discussion edit history move watch

Targets

Home Targets Phage Display Hybridoma PEAPS Vectors Protocols Informatics Antibody Atlas Users

Target Name	Uniprot	Structure	Antibodies	Functional	Humanized	Epitope Known	In-vivo efficacy
	_HUMAN	yes	yes	yes	yes	yes	yes
	_HUMAN	yes	yes	yes	yes	no	yes
	_HUMAN	yes	yes	yes	no	some	yes
	_HUMAN	yes	no	no	no	no	no
	_HUMAN	yes	yes	yes	yes	yes	yes
	_HUMAN	yes	yes	?	?	?	?
	_HUMAN	-	yes	?	no	no	no
	_HUMAN	yes	yes	?	?	?	?
	_HUMAN	yes	yes	?	?	?	?
	_HUMAN	yes	yes	?	?	?	?

Active projects

Decisions and current status

Super Simple Editing



Rich Editor

Editing Targets

[Disable rich editor](#) [\[Open Rich editor in new window\]](#)

Wikikitext

Format

Target Name	Uniprot	Structure	Antibodies	Functional	Humanized	Epitope Known	In-vivo efficacy
		yes	yes	yes	yes	yes	yes
		yes	yes	yes	yes	no	yes
		yes	yes	yes	no	some	yes
		yes	no	no	no	no	no
		yes	yes	yes	yes	yes	yes
		yes	yes	?	?	?	?
		-	yes	?	no	no	no
		yes	yes	?	?	?	?
		yes	yes	?	?	?	?
		yes	yes	?	?	?	?
		no	yes	?	no	no	?

Regular HTML table

Individual Target Page



Public data

Mutagenesis experiments

The screenshot shows the Wikibody interface for an individual target page. The page is titled "Wikibody" and has a navigation sidebar on the left. The main content area is divided into sections: "1 Target", "2 Antibodies", "3 Mutagenesis", "4 Efficacy", and "5 Vectors". The "3 Mutagenesis" section is expanded, showing subsections: "3.1 Mutagenesis by Amisha & Jake", "3.2 Loop Replacement by Javier", "3.3 Mutagenesis by Javier and Dan", "3.3.1 Biacore results 1 - Test mutant binding to", "3.3.2 Biacore results 2 - Test binding of 5 mouse IgGs to wildtype and mutant", "3.3.3 Biacore results 3 - Test binding of 7 human scFvs to wildtype and mutant", "3.4 Resurfaced Mutants", and "3.5 Epitope". The "5 Vectors" section is also visible. The "Target" section at the bottom provides a description of the target's role in regulating the level of a protein and its activity. The "Sequence" section is also present. A 3D molecular structure of the target is shown on the right side of the page.

Reagents

Binding data

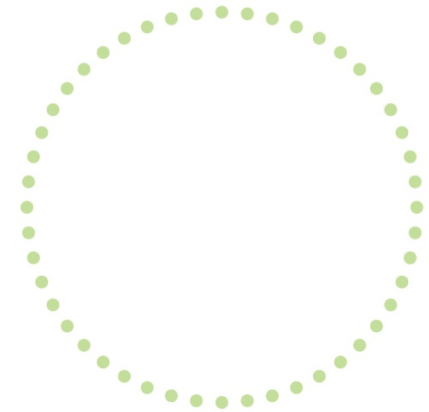
“Small things, loosely joined, written fast”



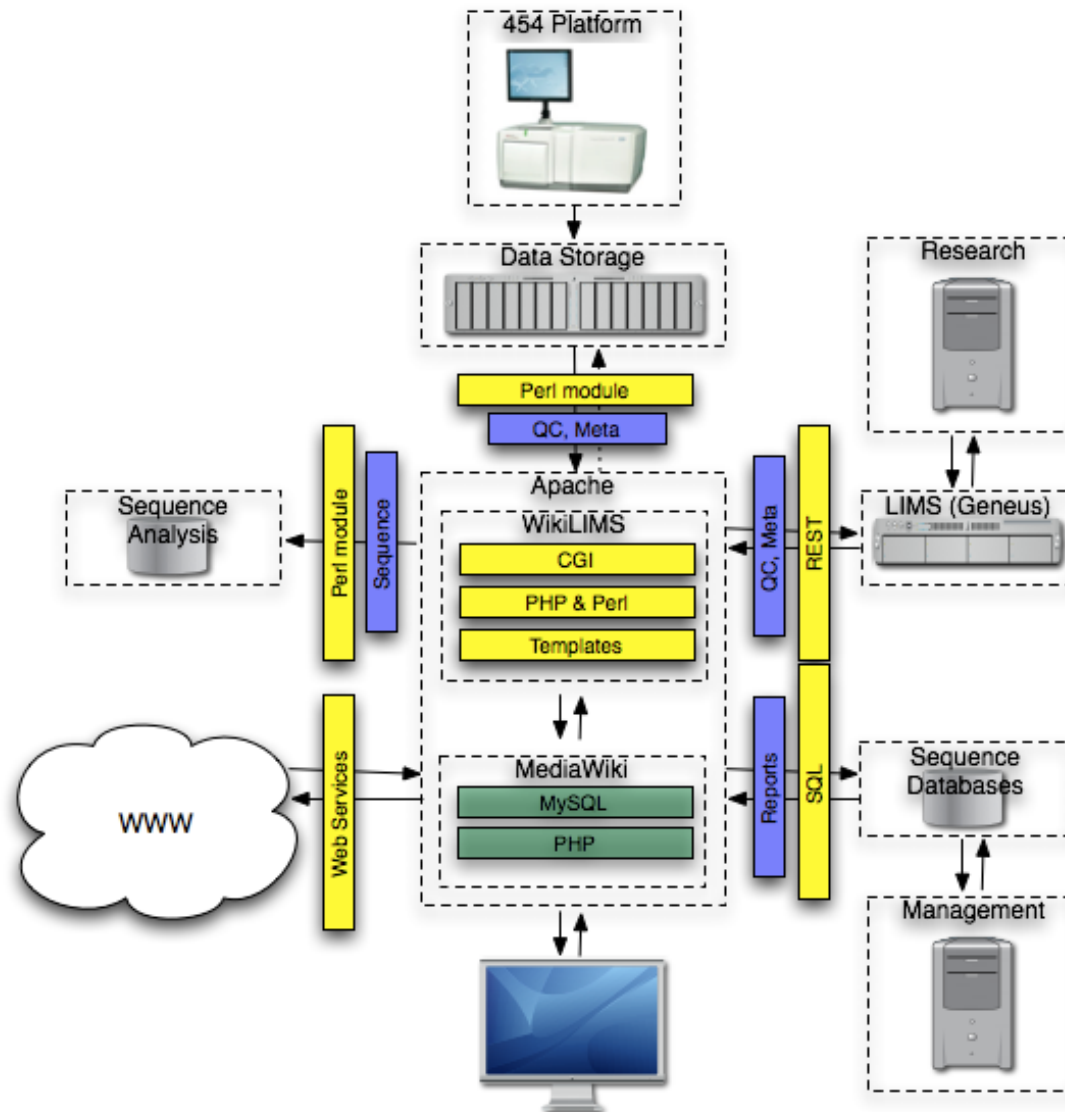
The screenshot shows a web application interface for Pfizer. On the left, there's a sidebar with links: Projects, Runs, New Project, Add Samples, View Samples. The main content area has a header with the Pfizer logo and text: "UCSF01 Rare variant screening of candidate genes - UCSF01.01" and "New sample_shipment". Below this is a calendar for January 2010. To the right of the calendar is a table of sequencing data. A green arrow labeled "APIs" and "HTTP" points from the calendar to the table.

Container Type	Container Barcode	Well Location	Geneus Sample Name	Manifest Donor ID	Manifest DNA Accession	Sequencing DB ID
Tube	NullTmplCtCtrl		NullTmplCtCtrl_11	0000	0000	SID-201
96 well plate	Plate 1	A01	Plate 1_A1	8251	2883	SID-1
96 well plate	Plate 1	A02	Plate 1_A2	2880	3763	SID-9
96 well plate	Plate 1	A03	Plate 1_A3	9010	4115	SID-17
96 well plate	Plate 1	A04	Plate 1_A4	8952	4212	SID-25
96 well plate	Plate 1	A05	Plate 1_A5	4903	4994	SID-33
96 well plate	Plate 1	A06	Plate 1_A6	6878	5214	SID-41
96 well plate	Plate 1	A07	Plate 1_A7	7271	5278	SID-49
96 well plate	Plate 1	A08	Plate 1_A8	4075	5449	SID-57
96 well plate	Plate 1	A09	Plate 1_A9	5079	5596	SID-65

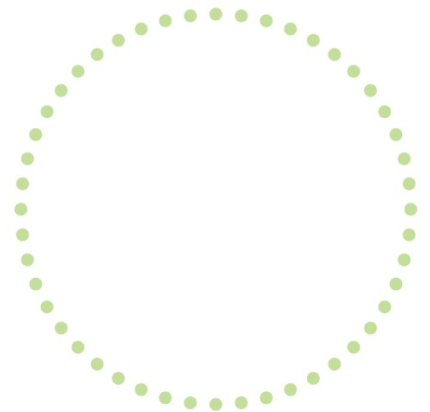
- Bring together diverse systems with loose interfaces
- Design and leverage **APIs** wherever possible
- Avoid all-in-one solutions
- Scale components **independently**



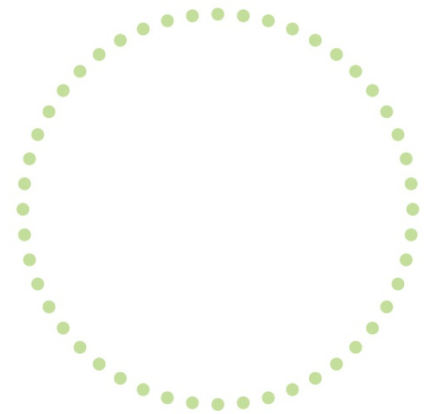
WikiLIMS@Pfizer



Data Analysis



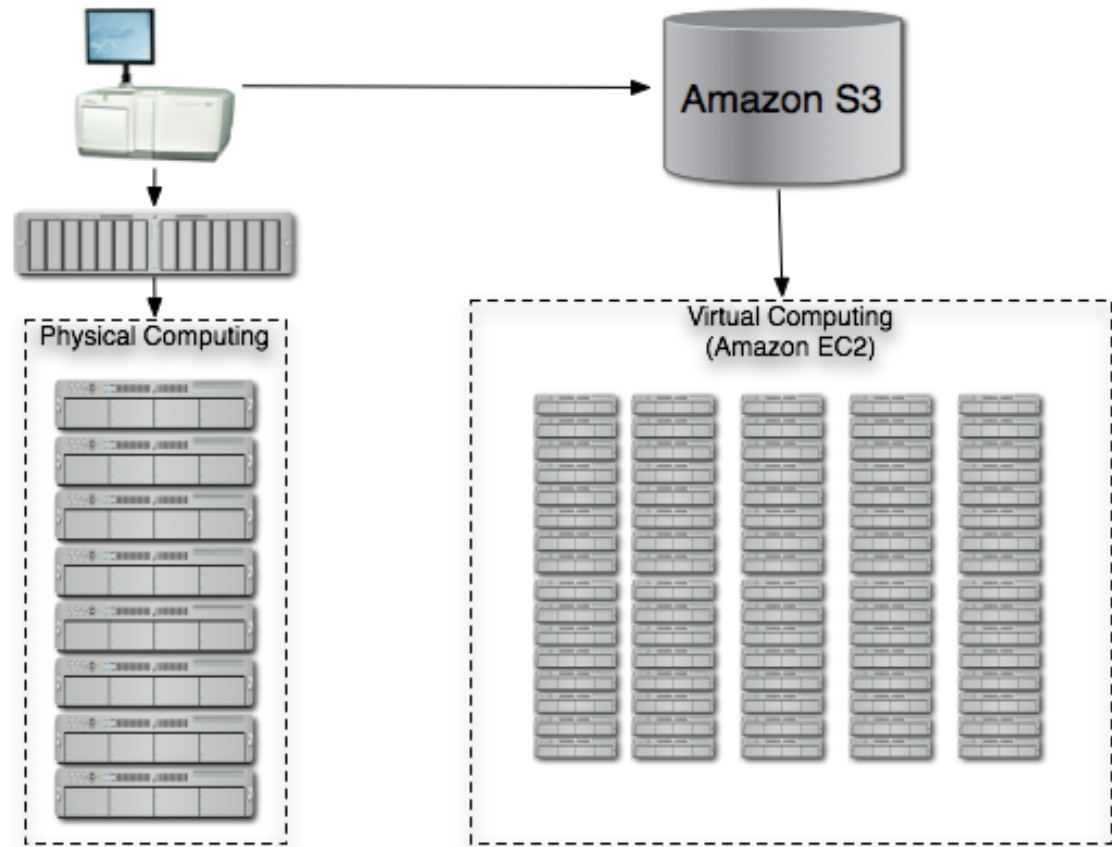
Infrastructure as a Service (IaaS) for NGS Data



Balancing Trade-offs...



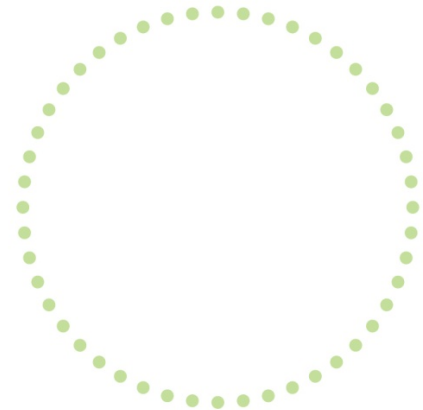
- Physical computing
- Virtual computing
- Availability
- Performance
- Investment
- Cost savings



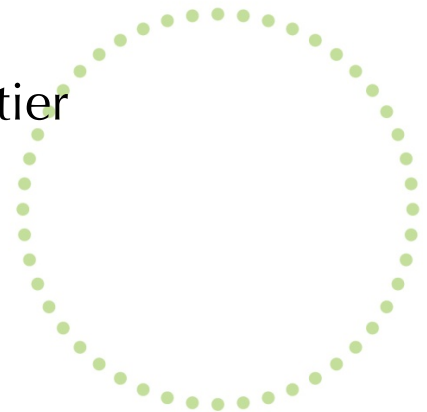
Amazon S3 Bulk Data Ingest/Export



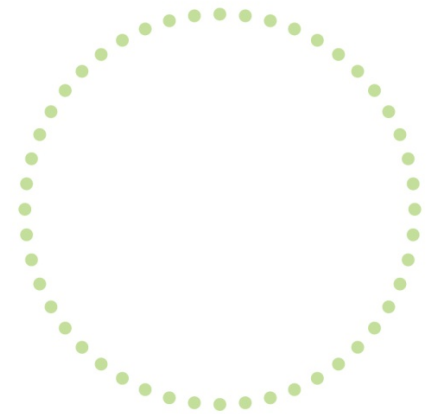
- **How do I forklift 200TB of data into the cloud?**
 - Fast network connectivity is expensive
 - Ship physical disks to Amazon to load into S3
- **What this means:**
 - S3 storage is cheap, deep, and geographically replicated
 - Primary analysis data moved into remote utility storage
 - Data would rarely come back
 - Need to reprocess or analyze?
 - Spin up “cloud” servers and re-analyze in-situ



- Bootstrapping new resources can happen in **minutes** instead of **months**
- Shift capital expenditure to operating expenditure (CapEx/OpEx)
- Primary analysis done on-site or commercial sequencing facility; data moved to remote utility storage
- Easier to distribute and collaborate across multiple sites
- Data would rarely need to move back
- Re-analyze data 'in the cloud' on virtual clusters
- Drop common analysis pipelines into a shared service tier



- Security concerns should not be overlooked
 - Regulatory compliance
- Amazon may be the only provider of bulk import/export services
 - What if Amazon goes out of business?
- Life Science tends to be I/O bound; I/O performance in the cloud is still lacking
- IT staff needs to understand 'the cloud' and learn to use it
- Critical to quantify your own internal operating costs



- Organic – mirroring Pfizerpedia
- Providing information carrots
- Restricted to Rinat groups
- Open source?



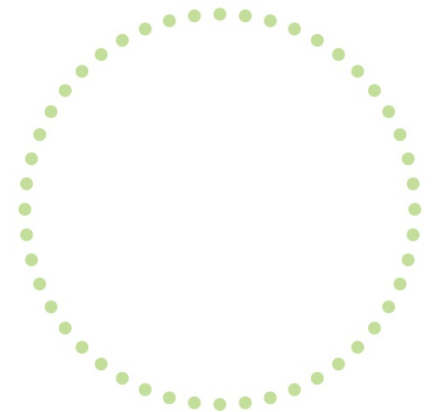
Advantages



- Flexibility
- Cost
- Complexity
- Adoption



- Jacob Glanville
- Adam Kraut (BioTeam)
- Steve Pitts
- Andrea Rossi



Questions?



- That's it
- Giles Day giles.mr.day@pfizer.com
- Adam Kraut kraut@biotam.net

