




# 2009 GMOD Meeting

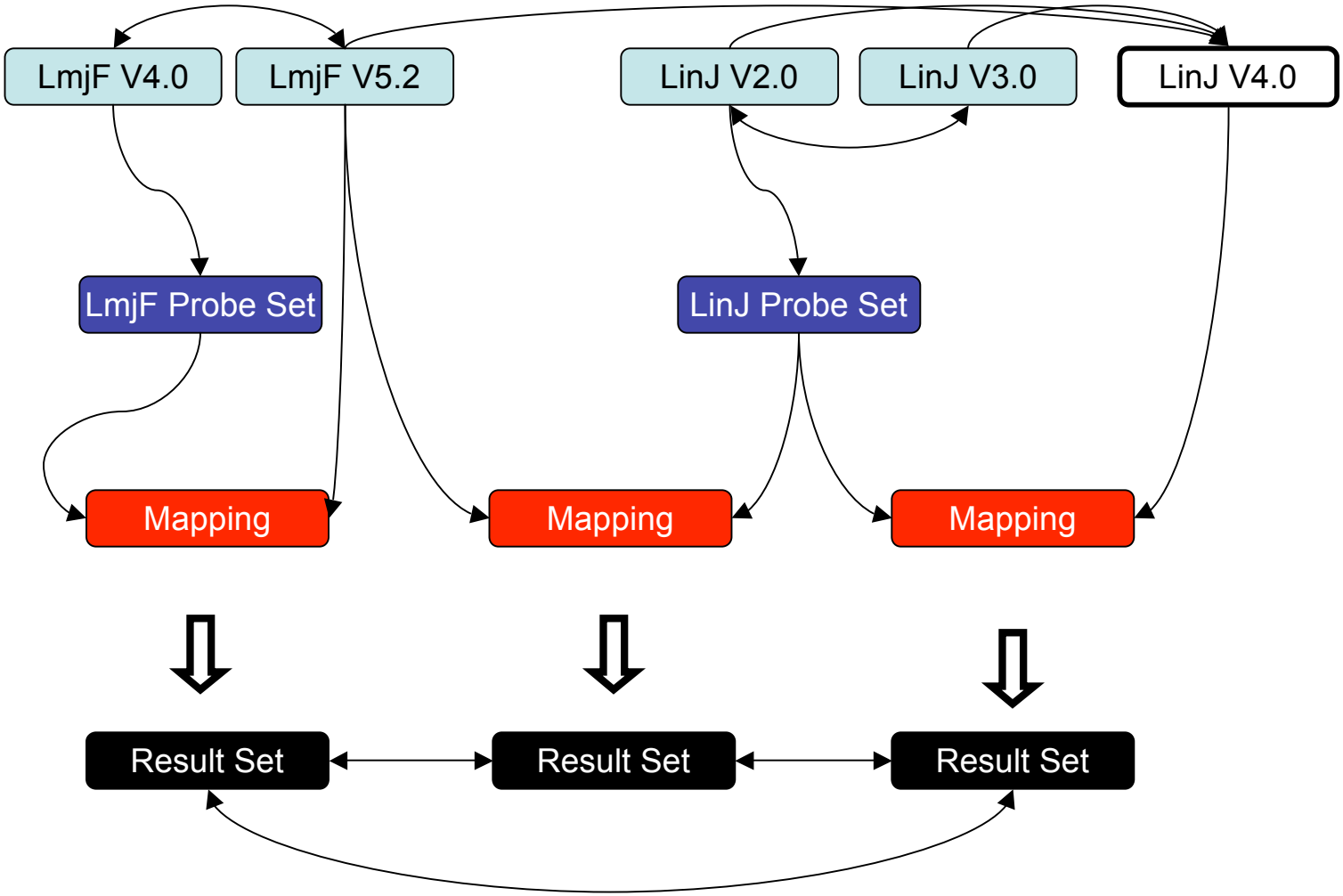
Dhileep Sivam & Isabelle Phan

Seattle Biomedical Research Institute

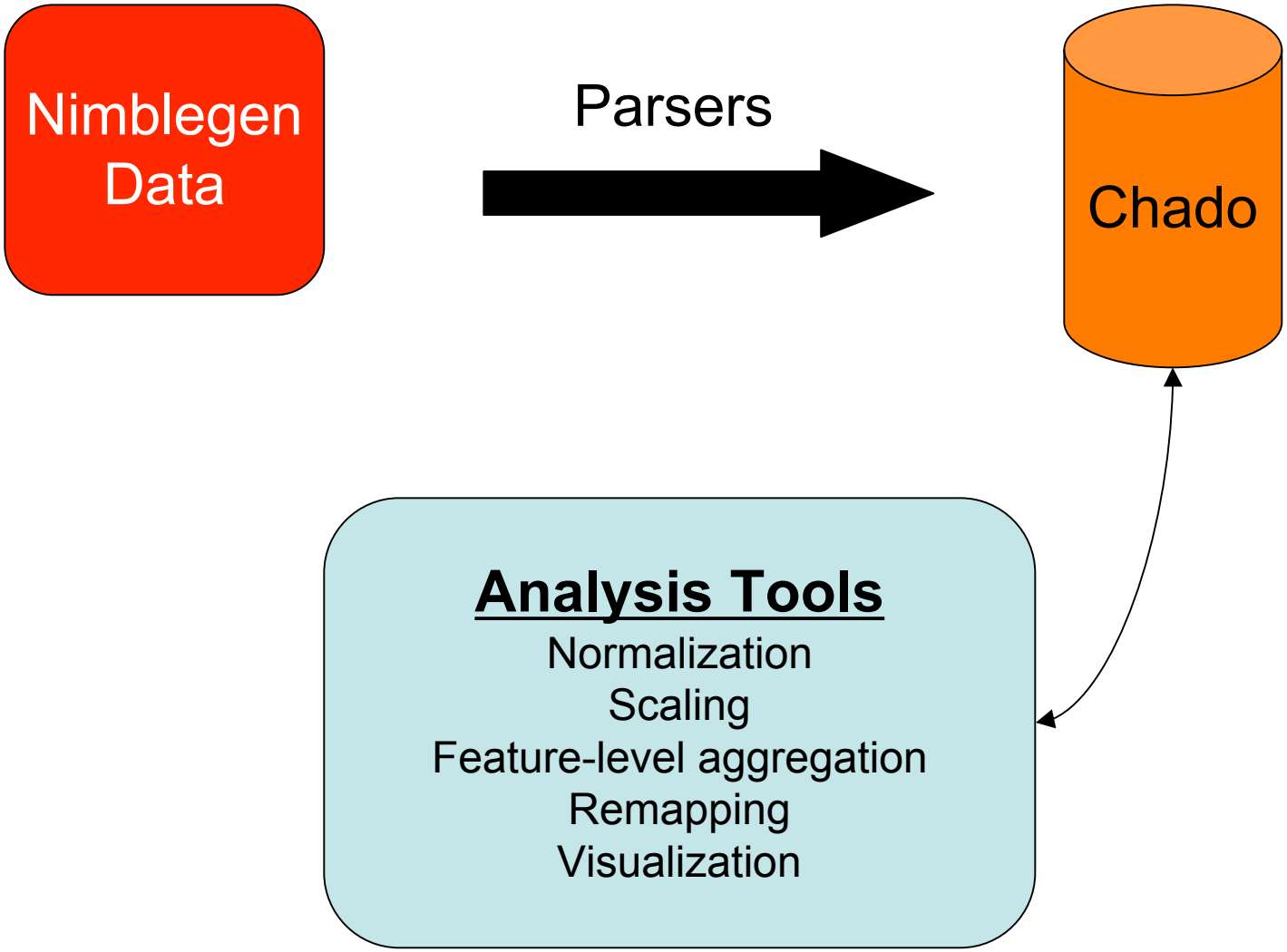
# Seattle Biomedical Research Institute (SBRI)

- Founded in 1976
- About 250 full-time staff
- Focus on infectious disease
- 13 Labs
- Strong ties to the University of Washington
- Bioinformatics Core

# How we first came to use Chado



# Microarray Project

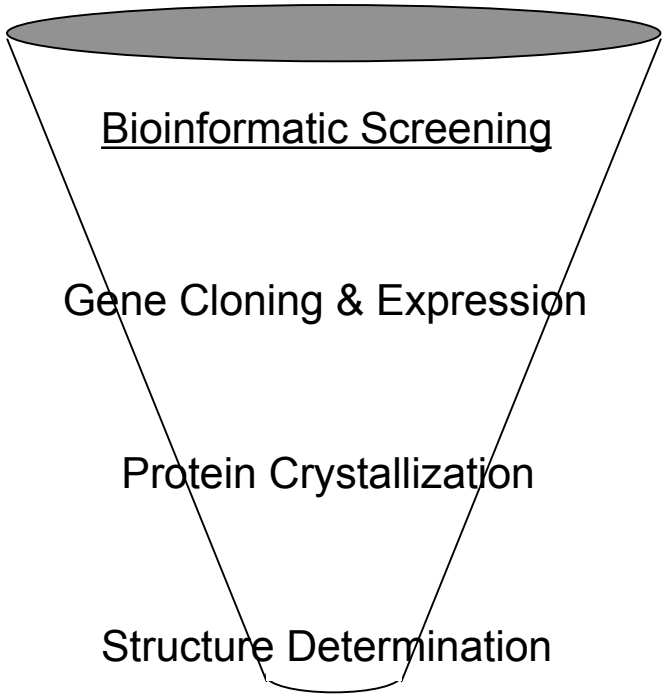


# Use Case: SSGCID

## Seattle Structural Genomics Center for Infectious Disease

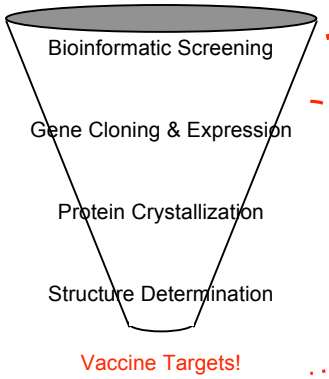
### Project Aim

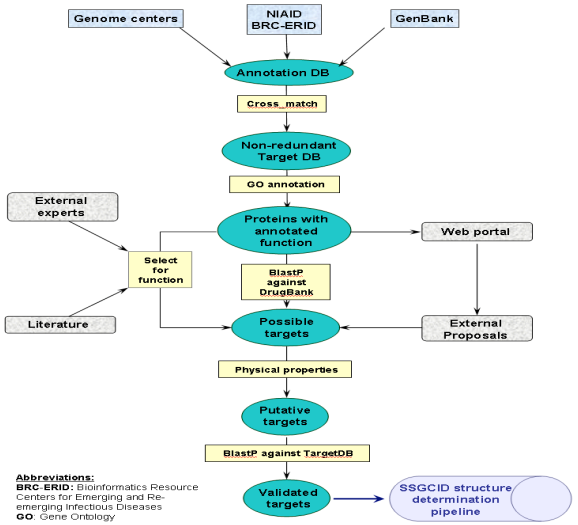
- 3D Protein Structure
- NIAID Emerging and re-emerging priority pathogens
- Structures will serve as a starting point for drug development
- Multi-center

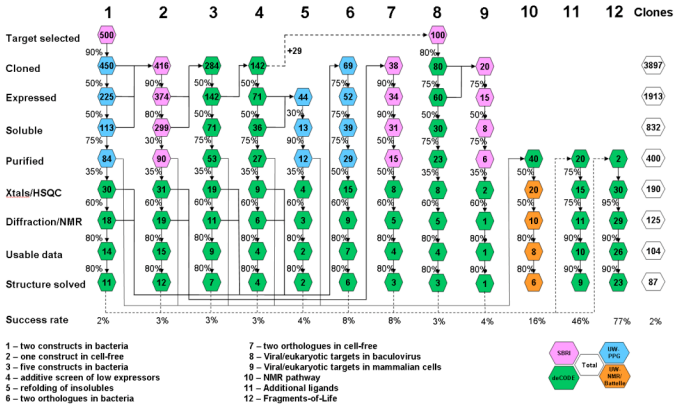



Vaccine Targets!

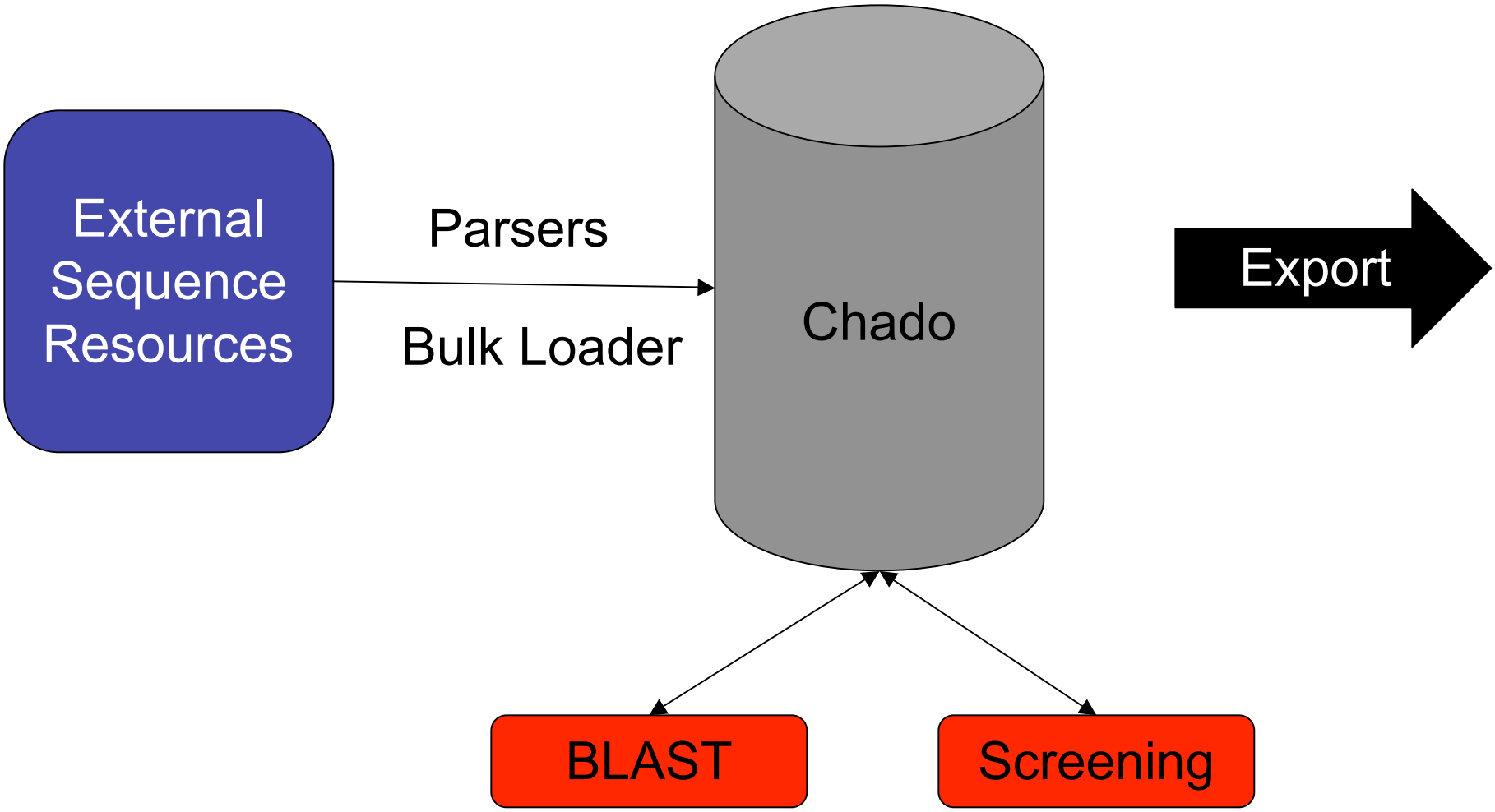
# SSGCID


Bioinformatic Screening
Gene Cloning & Expression
Protein Crystallization
Structure Determination
Vaccine Targets!

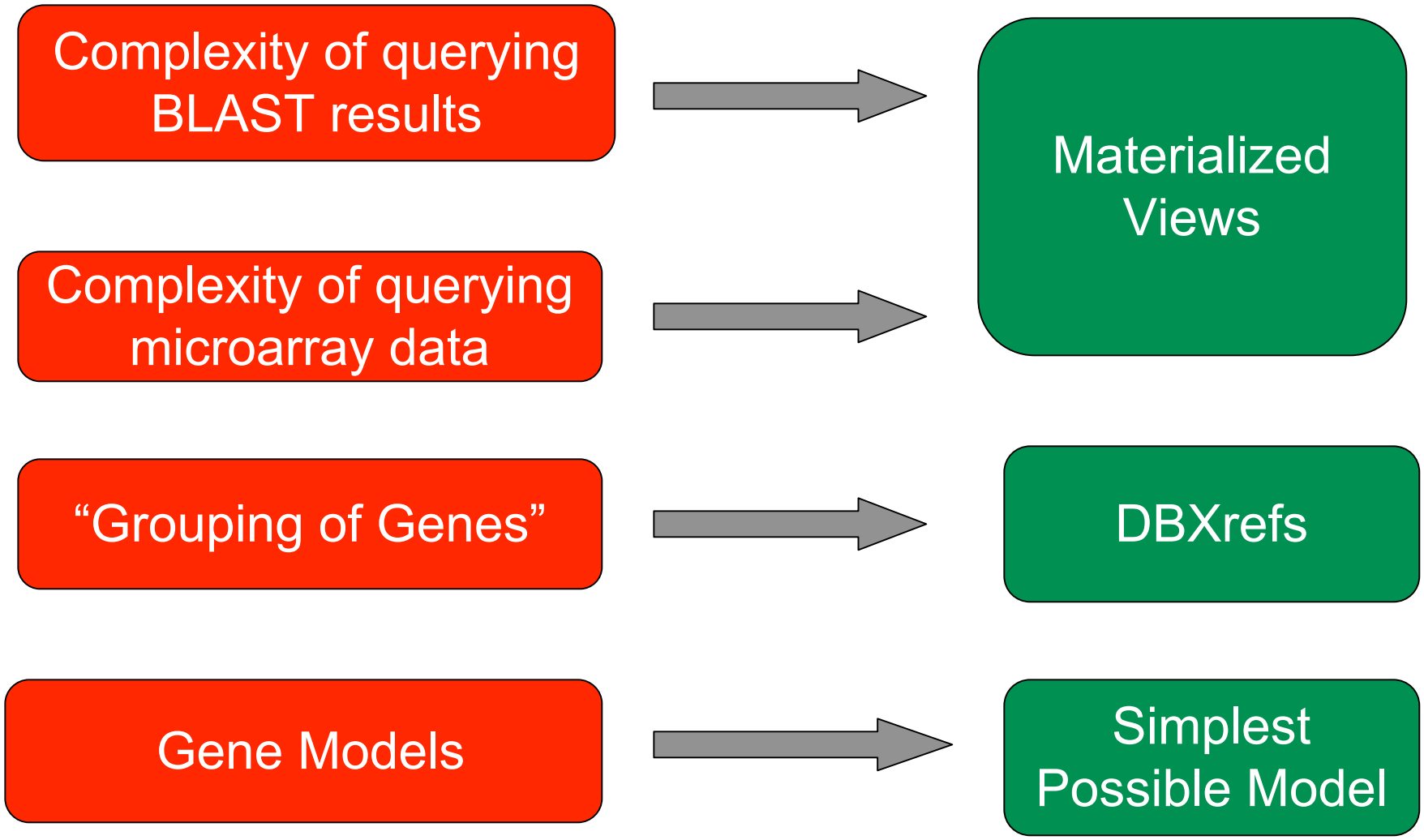

Genome centers
NIAID BRC-ERID
GenBank
Annotation DB
Cross_match
Non-redundant Target DB
GO annotation
External experts
Proteins with annotated function
Web portal
Select for function
BlastP against DrugBank
Literature
Possible targets
External Proposals
Physical properties
Putative targets
BlastP against TargetDB
Validated targets
SSGCID structure determination pipeline
Abbreviations:
BRC-ERID: Bioinformatics Resource Centers for Emerging and Re-emerging Infectious Diseases
GO: Gene Ontology


1 2 3 4 5 6 7 8 9 10 11 12 Clones
Target selected
Cloned
Expressed
Soluble
Purified
Xtals/HSQC
Diffraction/NMR
Usable data
Structure solved
Success rate
1 – two constructs in bacteria
2 – one construct in cell-free
3 – five constructs in bacteria
4 – additive screen of low expressors
5 – refolding of insolubles
6 – two orthologues in bacteria
7 – two orthologues in cell-free
8 – Viral/eukaryotic targets in baculovirus
9 – Viral/eukaryotic targets in mammalian cells
10 – NMR pathway
11 – Additional ligands
12 – Fragments-of-Life

# SSGCID

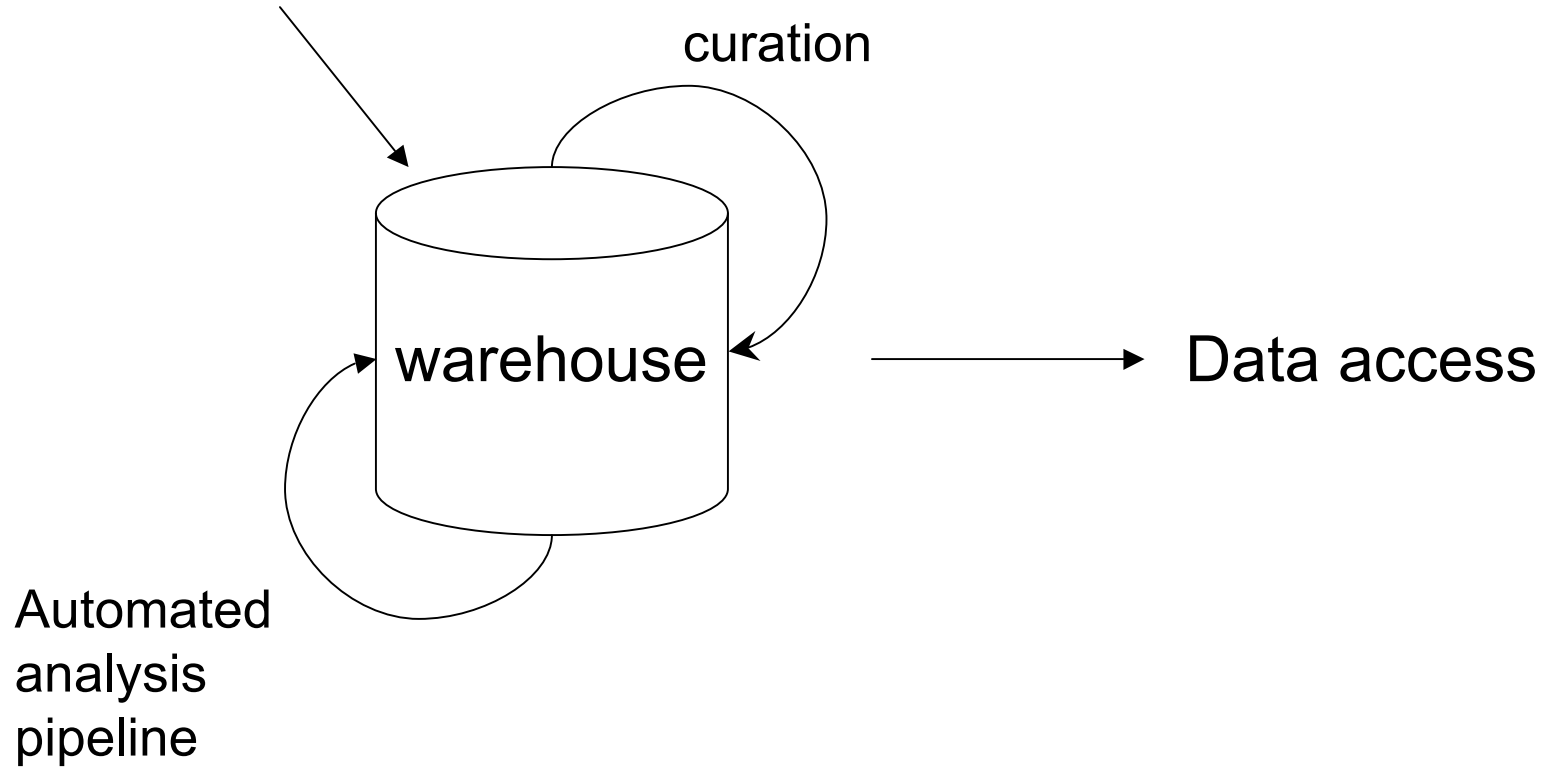

External Sequence Resources
Parsers
Bulk Loader
Chado
Export
BLAST
Screening

# Things that have come up…



Sequence data management at SBRI

Proteomics
Microarray
*Structural genomics*

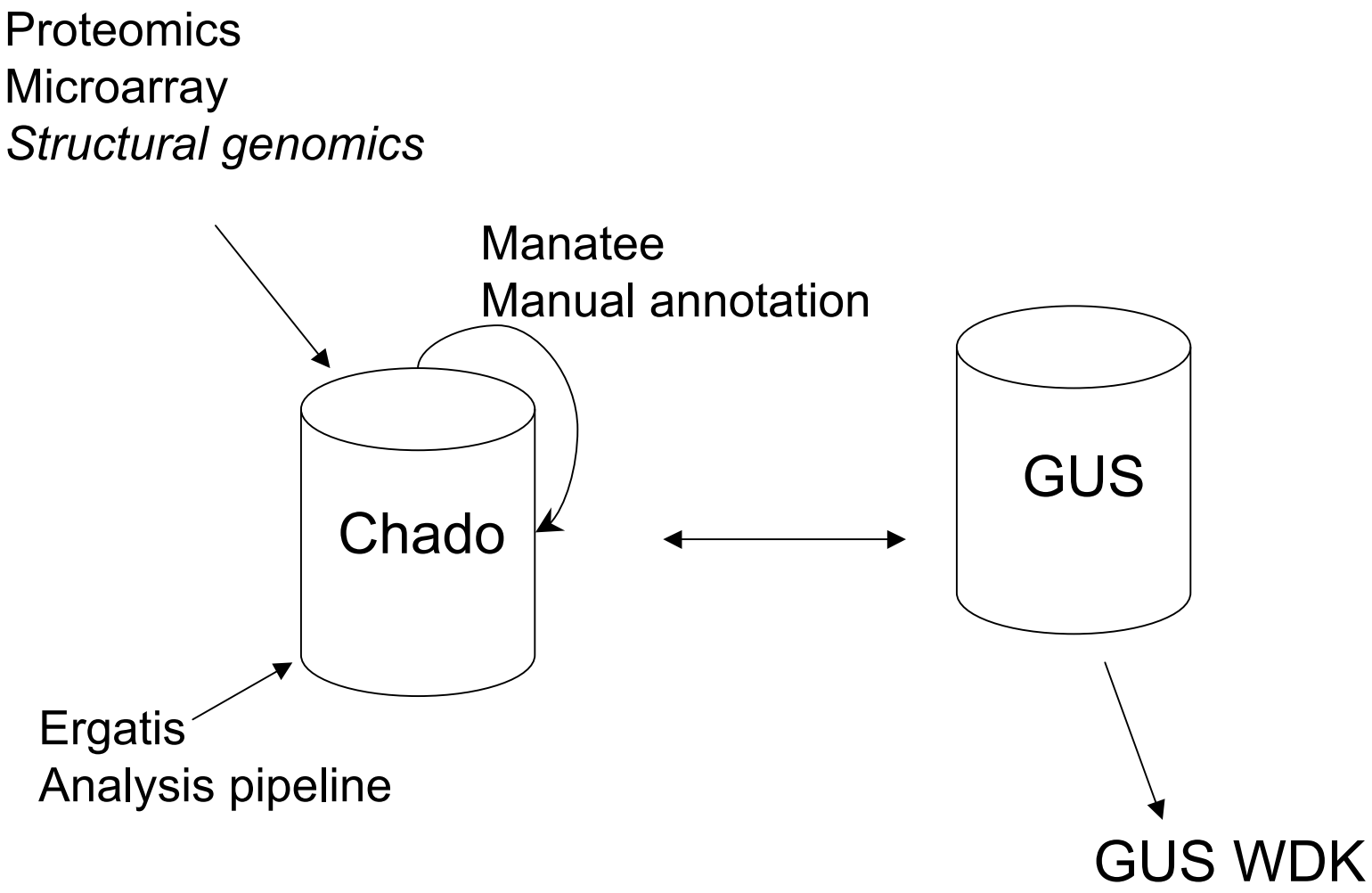


# Chado + GUS: why do we need both?

- Chado
  - Collaboration with IGS
  - Annotation tools: Manatee (apollo), Ergatis
    - Internal data production

- Gus
  - Collaboration with UPenn
  - Web front end
    - External data access

Sequence data management at SBRI



# Chado2GUS: Lost in translation

- Chado
  - Denormalized schema
    - Polymorphism
  - Mysql (IGS Chado)
- GUS
  - Normalized schema
    - Subclassing
  - Postgres port from Oracle

performance issue

performance issue

# Picking the best of two worlds

- Chado
  - Biological data model
  - Flexibility
- GUS
  - Software engineering
  - Flexibility

# The future?

- SQL-free data production
  - Instead of custom wrappers over raw SQL:
    - ORMs: Chado Hibernate, ActiveRecords
    - Unified object model
- RDBMS-free data mining
  - Instead of GUS predefined query + set combination
    - Biomart + Galaxy
    - RDF + triple store + sparql (object store + Lucene)