

Establishing a Central Resource of Data from Genome Sequencing Projects

Scope of the NIH Data Sets

June 5/6 2012

Inventory of NIH Projects *Projected through end of ~2012)*

<http://www.genome.gov/27545796>

Institute or Center	Project	Whole Exome (yes/no)	Whole Genome (yes/no)	Depth of Coverage (3X, 20X, etc)	Sequencing Platform	Date Sequencing Initiated	Date Sequencing (to be) Completed	Date Data to be Shared	Venue for Data Sharing	Number of Individuals (to be) Sequenced	Criteria for Selection of Individuals	Primary Phenotype	Age Range (yr)	Race/Ethnic Comp
NCI-CCR	DRCT Whole Exome	Yes	No	>100	SOLiD	Jan-12	Mar-12	Upon Publication	dbGaP	25		Desmoplastic Small Round Cell Tumor	Child	Any
NCI-CCR	EWS Whole Genome	No	Yes	>40	Complete Genomics	May-11	Sep-11	Upon Publication	dbGaP	6		Ewing's Sarcoma	Child	Any
NCI-CCR	Genome structure of DLBCL		Yes	Unknown	Complete genomics	Mar-11	Jul-11	2012	dbGAP	2	One patient with ABC DLBCL and One with GCB DLBCL	Development of DLBCL	38-67	Caucasian
NCI-CCR	GIST Whole Exome	Yes	No	>100	SOLiD	Mar-11	Apr-11	Upon Publication	dbGaP	3	GIST with Family Member	Gastrointestinal Stromal Tumor	Child	Any
NCI-CCR	HBL Whole Exome	Yes	No	>100	SOLiD	Nov-11	Jan-12	Upon Publication	dbGaP	25		Hepatoblastoma	Child	Any

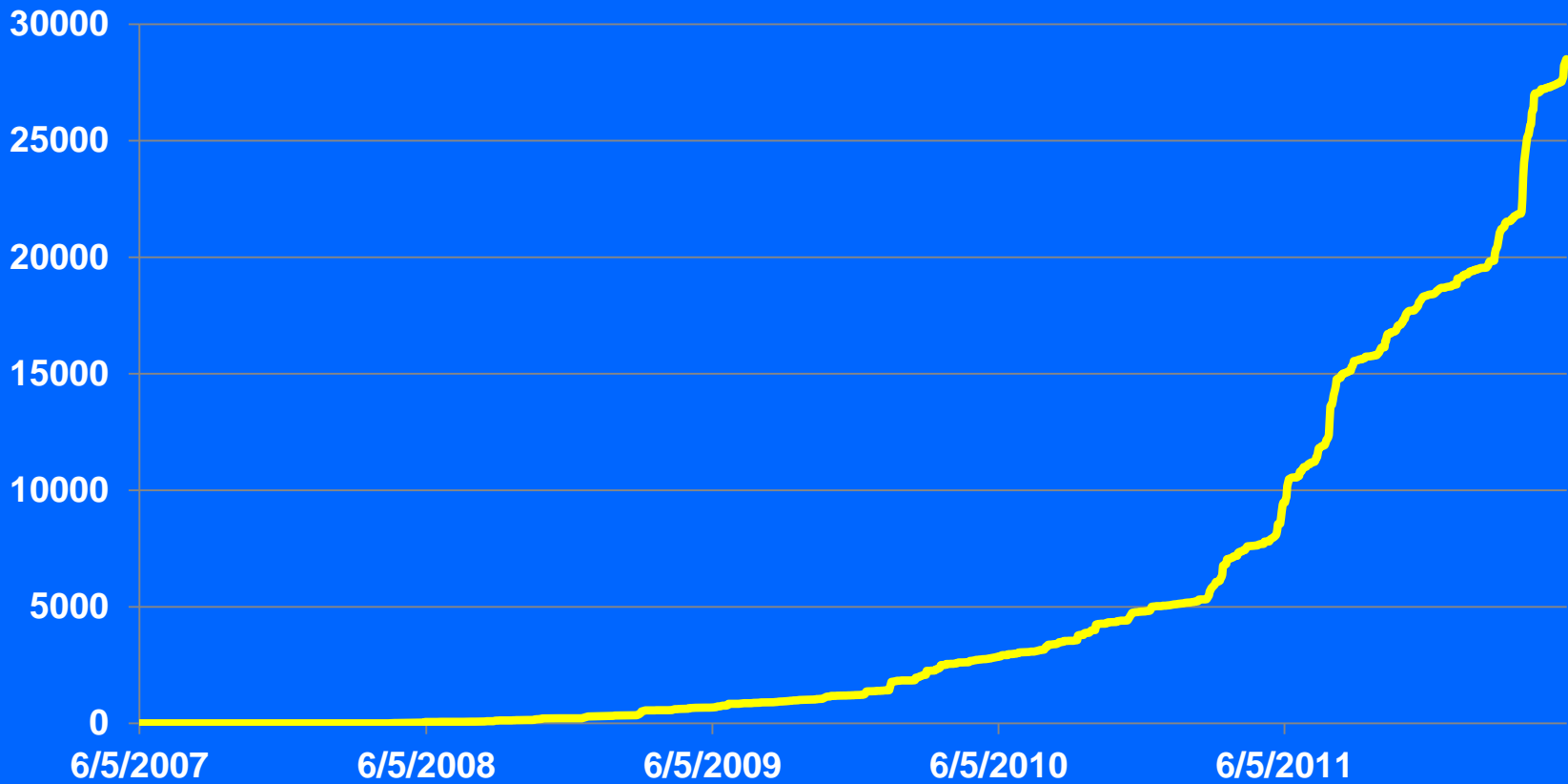
Inventory High Level Numbers:

By end of year:

- Projects: ~190
- Samples: ~68,800 (~18K as WGS?)

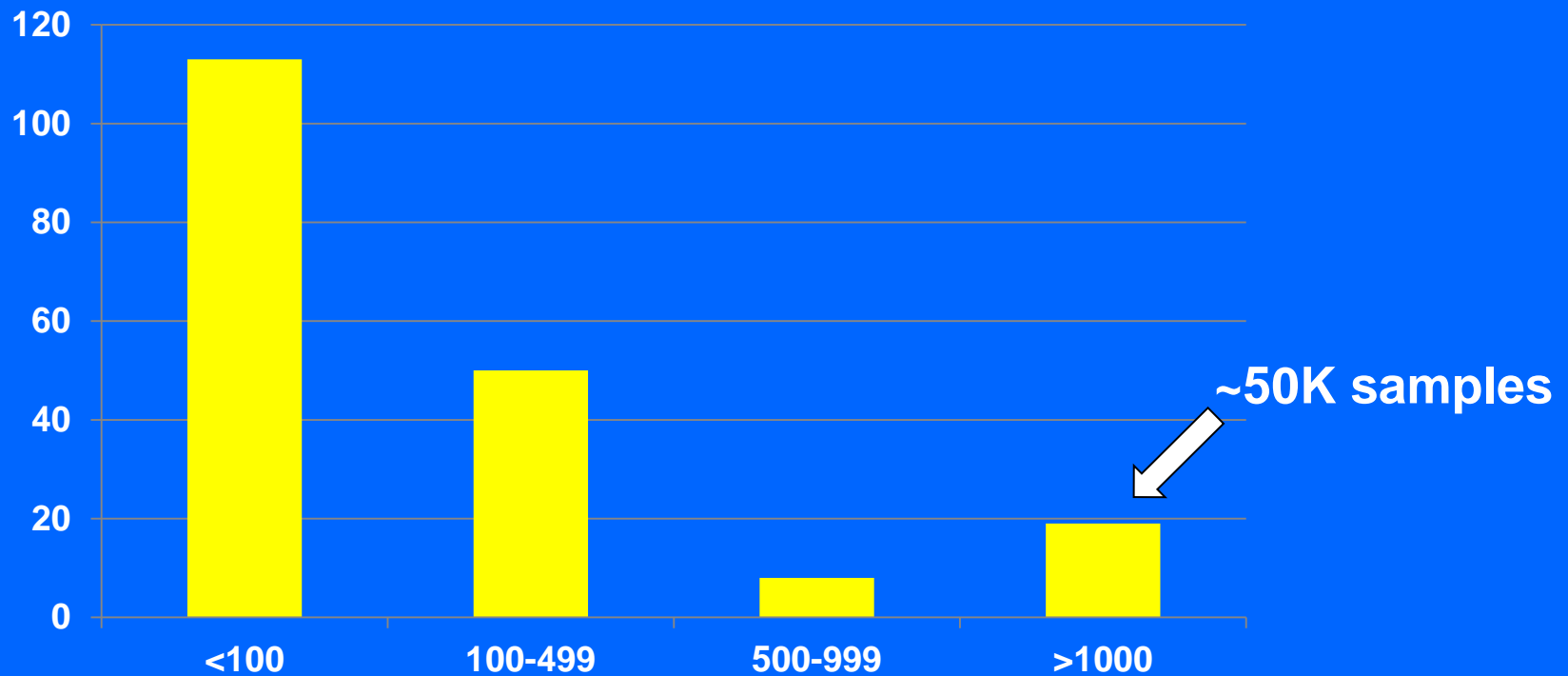
Growing.....

Current samples in dbGaP

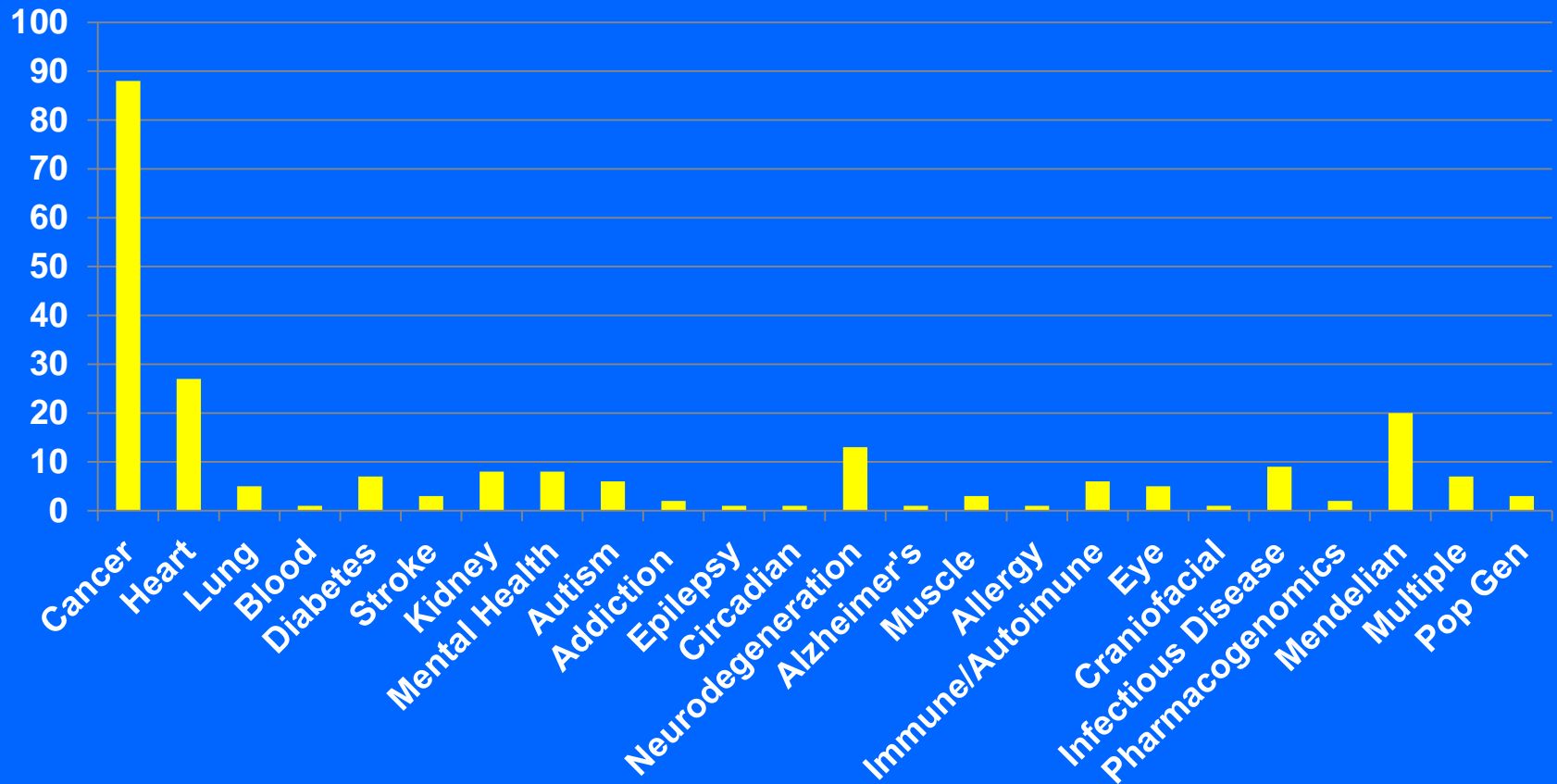


Data Are Organized into Projects

Projects by sample size



Projects by Disease Area



“High-Value” Samples?

Out of ~68,800 total

~26,000 samples have no data use limitations. (~2600 are completely public...)

~15,900 samples - participant recontact is permitted (could *in principal* be re-phenotyped)

~Physical samples actually available – 25K?

Other Things We ~~may~~ Want To Know

- What phenotype data?
- What exposure data?
- What population?
- What age?

These data are all at
NCBI/dbGaP, right? So what's
the problem?



~200 projects; ~400 Consent Groups; Go
through multiple DAC's to access;
Inconsistent metadata, etc.

And, Not All the Data are in dbGaP

- Some of the NIH-funded data are in CGHub
- Disease-specific databases will proliferate
- More data outside US

Summary

- A lot of data: easy to imagine 100K+ samples and 1000 Tbases just from NIH by end of 2013.
What number of samples should we plan for? (1M?)
- If we do nothing, also easy to imagine these being divided into many projects
- Large numbers of samples are concentrated in a few studies. If “high value” (eg., no restrictions, recontactable, etc.) then could be useful for aggregate analyses in short term. But far from optimal, and not scalable

Many Thanks To

Nicholas Clegg – Inventory digest and slides

Steve Sherry, Eugene Yaschenko, Martin Shumway – dbGaP
summary data

Teri Manolio, Ian Marpuri – Inventory

Lisa Brooks – general

end

These data are (mostly) Available via dbGaP

Home - dbGaP - NCBI - Mozilla Firefox

File Edit View History Bookmarks Tools Help

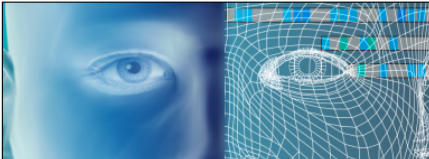
http://www.ncbi.nlm.nih.gov/gap

Most Visited Getting Started Latest Headlines

NCBI Resources How To My NCBI Sign In

dbGaP dbGaP Search

Limits Advanced Help



dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype.

Getting Started

- [dbGaP Tutorial](#)
- [Overview](#) [FAQ](#)
- [How to Submit](#)
- [Browse Top Level Studies](#)

Access dbGaP Data

- [Apply for Controlled Access Data](#)
- [Public Data via ftp Download](#)
- [Association Results Browser](#)
- [Phenotype-Genotype Integrator](#)

Important Links

- [dbGaP RSS Feed](#)
- [Code of Conduct](#)
- [Security Procedures](#)
- [Contact Us](#)

Latest Studies

Study	Embargo Release	Details	Participants	Type Of Study	Links	Platform
phs000414.v1.p1 Whole Genome Sequencing of CBF-Leukemia	Version 1: 2012-05-30	V D A S	17	Case Set, Tumor vs. Matched-Normal, Whole Genome Sequencing	Links	Genome Analyzer II HiSeq 2000
phs000413.v1.p1 Whole Genome Sequencing of Pediatric Acute Megakaryoblastic Leukemia	Version 1: 2012-05-30	V D A S	4	Case Set, Cohort, Tumor vs. Matched-Normal	Links	HiSeq 2000
phs000409.v1.p1 Sequencing of Medulloblastoma	Version 1: 2012-05-30	V D A S	93	Case Set, Whole Genome Sequencing, Tumor vs. Matched-Normal	Links	AFFY_6.0
phs000341.v1.p1 Genome-Wide Analysis of Hypodiploid ALL	Version 1: 2012-05-30	V D A S	20	Case Set, Tumor vs. Matched-Normal, Xenograft	Links	HiSeq 2000
phs000291.v2.p1 NHLBI GO-ESP: Lung Cohorts Exome Sequencing Project (Lung Health Study of Chronic Obstructive Pulmonary Disease)	Version 1: passed embargo Version 2: 2012-08-21	V D A S	337	Longitudinal, Exome Sequencing	Links	Genome Analyzer IIX

[List Top Level Studies](#)

Done