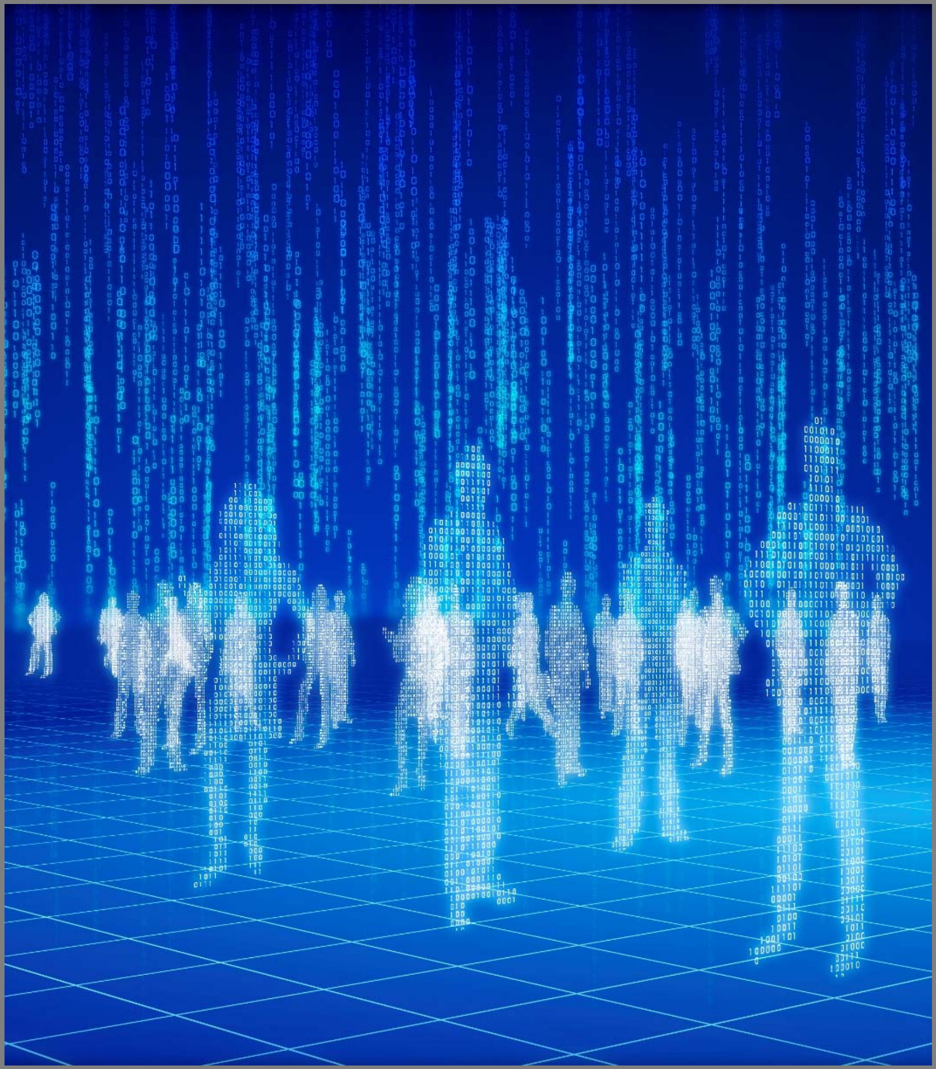


Maximizing Public Data Sources for Sequencing and GWAS

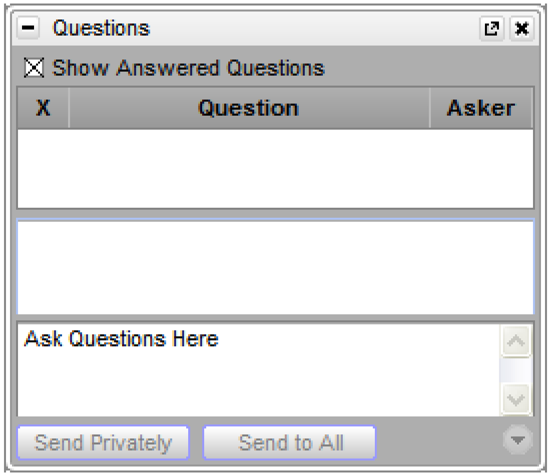
February 4, 2014

G Bryce Christensen
Director of Services

GOLDEN HELIX
Accelerating the Quest for Significance™

Questions during the presentation

Use the Questions pane in your GoToWebinar window



Agenda

1 Why Use Public Data?

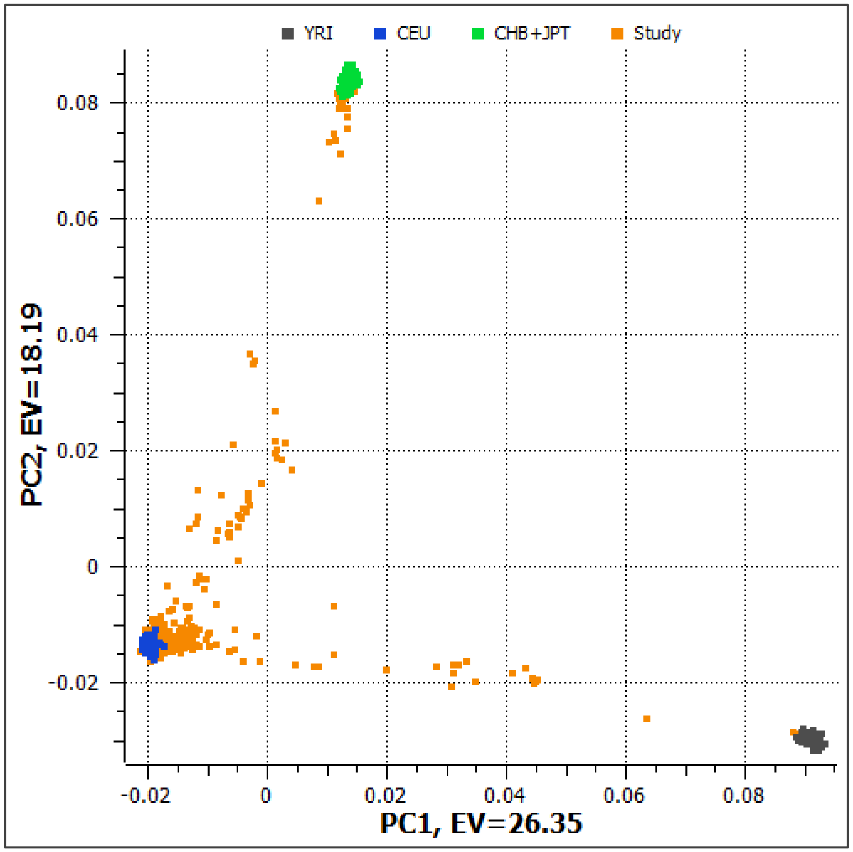
2 Where to Find Public Data

3 Tips for Using Public Data

4 Manipulating Public Data in SVS

Why Use Public Data?

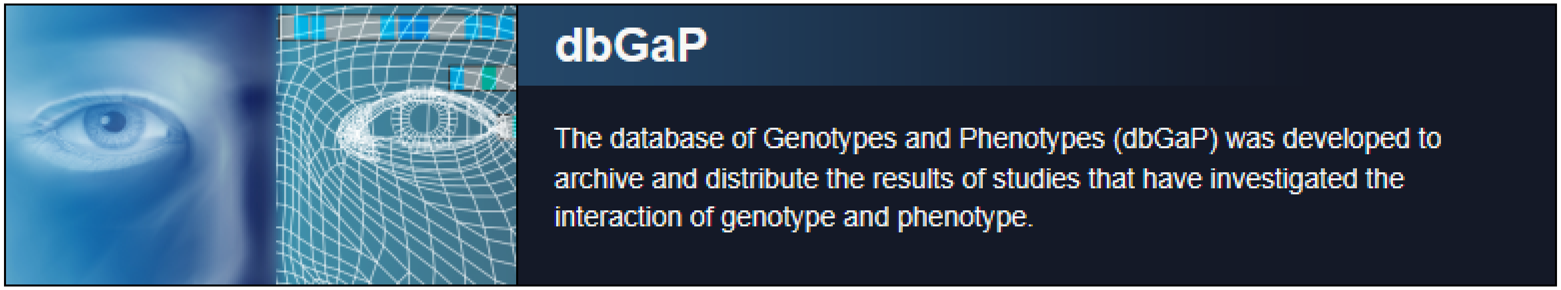
- Reference samples for assessing population structure in GWAS
- Replicating results of your own GWAS or other research
- Meta-analysis or Mega-analysis
- Testing new analytical methods
- Reference data for SNP imputation
- Increase study size with public controls



Sources of Public Data

- **NCBI**
  - dbGaP
  - GEO
  - SRA
- **EGA**
- **HapMap Project**
- **1000 Genomes Project**
- **Hardware vendors**
- **Software vendors**
- **All over the internet…**

dbGaP



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.

- “Database of Genotypes and Phenotypes”
- 435 studies in database (as of January 28th)
- Known primarily as a GWAS database, but NGS content is growing
- Freely view and download results for many studies
- Access to raw phenotype and genotype data requires application process

435 Studies in dbGaP (January 28th)

GWAS Platforms

- **Affymetrix**

SNP-6.0:	51
500k:	15

- **Illumina**

HumanHap550:	37
HumanHap300:	13
HumanCNV370:	11
Human610:	35
Human660:	26
Omni1:	22
Omni2.5:	14
Human_1M:	12

- **Perlegen**

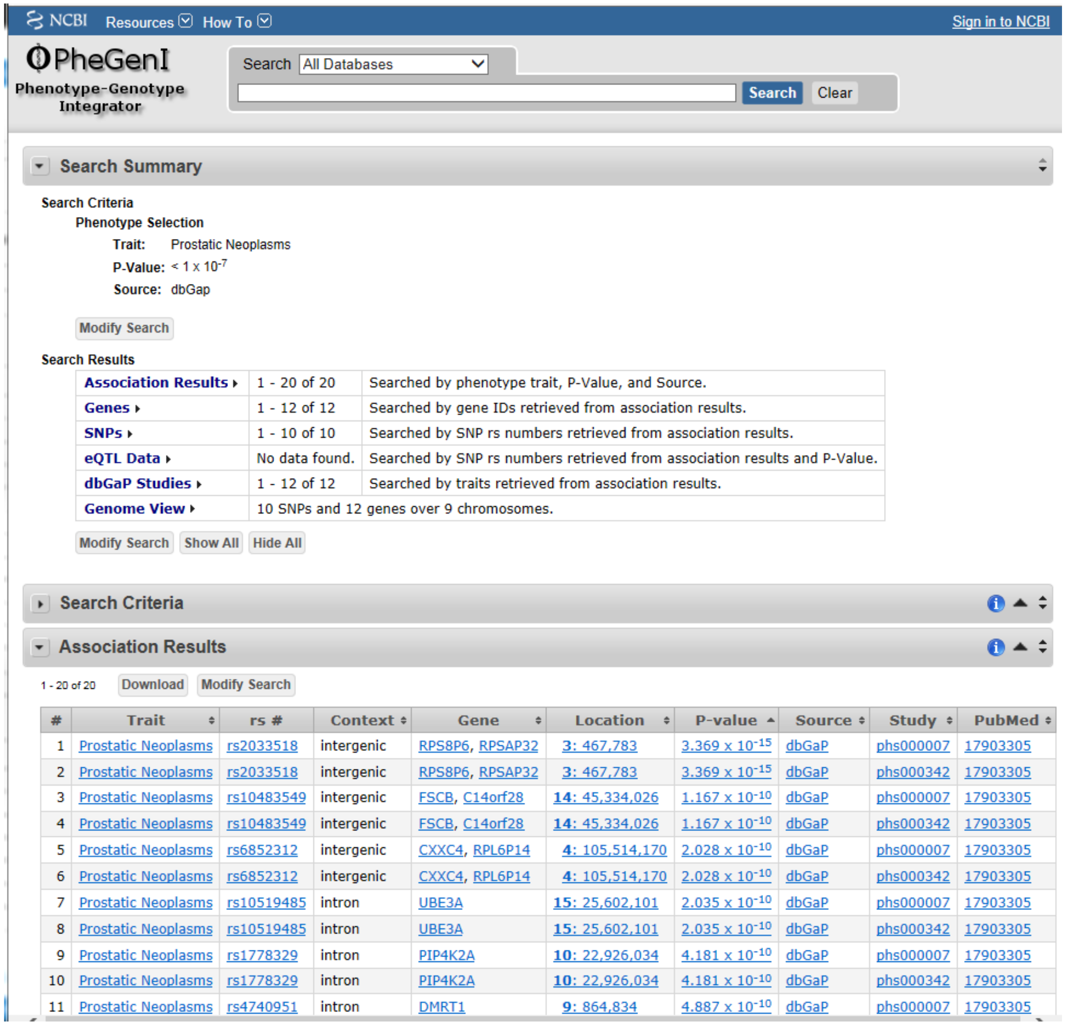
600k:	4

NGS Platforms

- **454: 22**
- **GA-II: 49**
- **HiSeq 2000: 72**
- **HiSeq 2500: 3**

dbGaP Tools

- **GaP Browser**
  - View GWAS study results in context of other genomic annotations
- **GaP Genome Browser**
  - Karyotype views of GWAS study results
- **PheGenI**
  - “Phenotype-Genotype Integrator”
  - Search NHGRI and dbGaP study results by phenotype or by gene
  - Annotated results with links to abstracts and/or dbGaP study pages.



Search Summary

Search Criteria

Phenotype Selection

Trait: Prostatic Neoplasms

P-Value: $< 1 \times 10^{-7}$

Source: dbGap

Search Results

Association Results	1 - 20 of 20	Searched by phenotype trait, P-Value, and Source.
Genes	1 - 12 of 12	Searched by gene IDs retrieved from association results.
SNPs	1 - 10 of 10	Searched by SNP rs numbers retrieved from association results.
eQTL Data	No data found.	Searched by SNP rs numbers retrieved from association results and P-Value.
dbGaP Studies	1 - 12 of 12	Searched by traits retrieved from association results.
Genome View	10 SNPs and 12 genes over 9 chromosomes.	

Search Criteria

Association Results

1 - 20 of 20

#	Trait	rs #	Context	Gene	Location	P-value	Source	Study	PubMed
1	Prostatic Neoplasms	rs2033518	intergenic	RPS8P6, RPSAP32	3: 467,783	3.369×10^{-15}	dbGaP	phs000007	17903305
2	Prostatic Neoplasms	rs2033518	intergenic	RPS8P6, RPSAP32	3: 467,783	3.369×10^{-15}	dbGaP	phs000342	17903305
3	Prostatic Neoplasms	rs10483549	intergenic	FSCB, C14orf28	14: 45,334,026	1.167×10^{-10}	dbGaP	phs000007	17903305
4	Prostatic Neoplasms	rs10483549	intergenic	FSCB, C14orf28	14: 45,334,026	1.167×10^{-10}	dbGaP	phs000342	17903305
5	Prostatic Neoplasms	rs6852312	intergenic	CXXC4, RPL6P14	4: 105,514,170	2.028×10^{-10}	dbGaP	phs000007	17903305
6	Prostatic Neoplasms	rs6852312	intergenic	CXXC4, RPL6P14	4: 105,514,170	2.028×10^{-10}	dbGaP	phs000342	17903305
7	Prostatic Neoplasms	rs10519485	intron	UBE3A	15: 25,602,101	2.035×10^{-10}	dbGaP	phs000007	17903305
8	Prostatic Neoplasms	rs10519485	intron	UBE3A	15: 25,602,101	2.035×10^{-10}	dbGaP	phs000342	17903305
9	Prostatic Neoplasms	rs1778329	intron	PIP4K2A	10: 22,926,034	4.181×10^{-10}	dbGaP	phs000007	17903305
10	Prostatic Neoplasms	rs1778329	intron	PIP4K2A	10: 22,926,034	4.181×10^{-10}	dbGaP	phs000342	17903305
11	Prostatic Neoplasms	rs4740951	intron	DMRT1	9: 864,834	4.887×10^{-10}	dbGaP	phs000007	17903305

Applying for dbGaP data

- **Each application is reviewed by a “DAC,” or data access committee**
  - I’ve seen approval time range from 1 to 8+ weeks.
- **Keep proposals relatively simple**
  - Read the instructions and be sure that your application is complete before submitting
  - Contact DAC before submitting if you have special needs or concerns
- **Some datasets require IRB approval to access**
  - Waiver letter is often sufficient
- **Pay attention to data embargoes**
- **External collaborators and contractors must apply separately for access**
- **Pay attention to consent groups**
  - General research use
  - Non-commercial use
  - Disease-specific use

Using dbGaP Data

- **Know what you are getting—read the documentation!**
  - Original study design
  - Data processing and formats
- **Be patient and thorough as you explore the data--treat it like fresh new data and don't assume that it is "clean."**
- **Phenotype data is usually stored in text files, often with a separate data dictionary.**
  - Read the documentation!
- **Carefully review phenotype data for completeness and consistency.**
  - Data from multi-center projects can be particularly problematic

Using dbGaP Data, continued

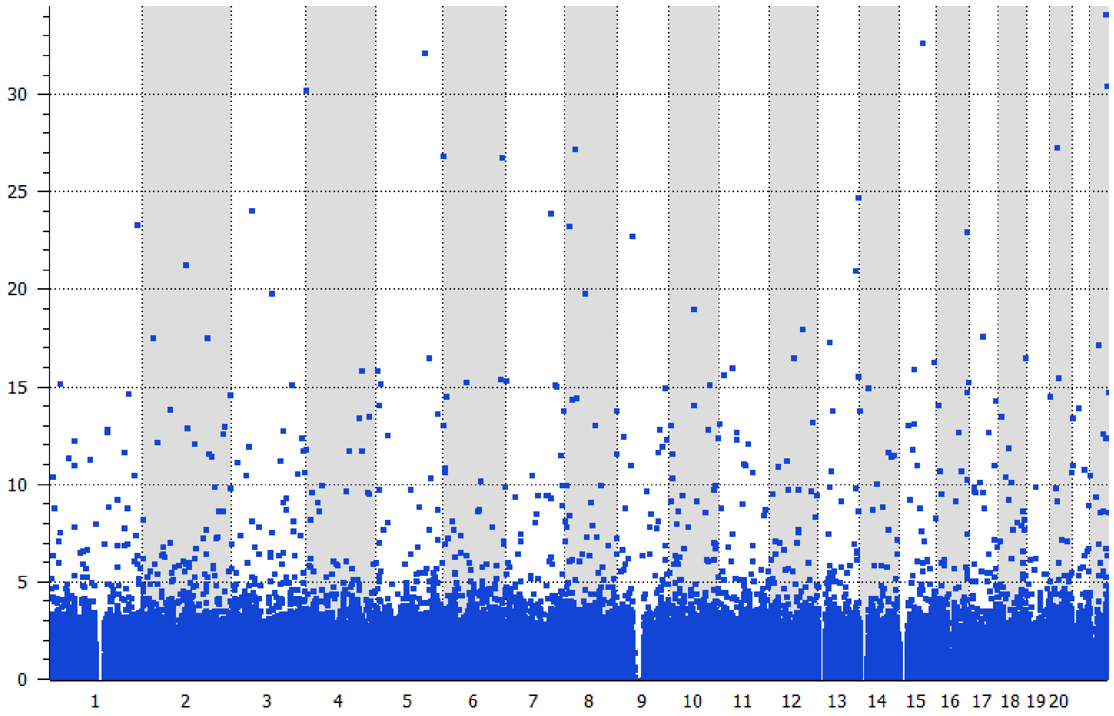
- **Many studies include three levels of genotype data:**
  - Raw data
    - CEL or iDat files
    - Hardest to use
  - Processed data
    - Genotype calls or Log Ratio values
    - Individual and/or matrix formats
  - QC’ed data
    - As used for the public analysis results
    - Easiest to use (usually in a format supported directly by SVS)
- **Start from the raw or minimally processed data and do your own QC whenever possible.**

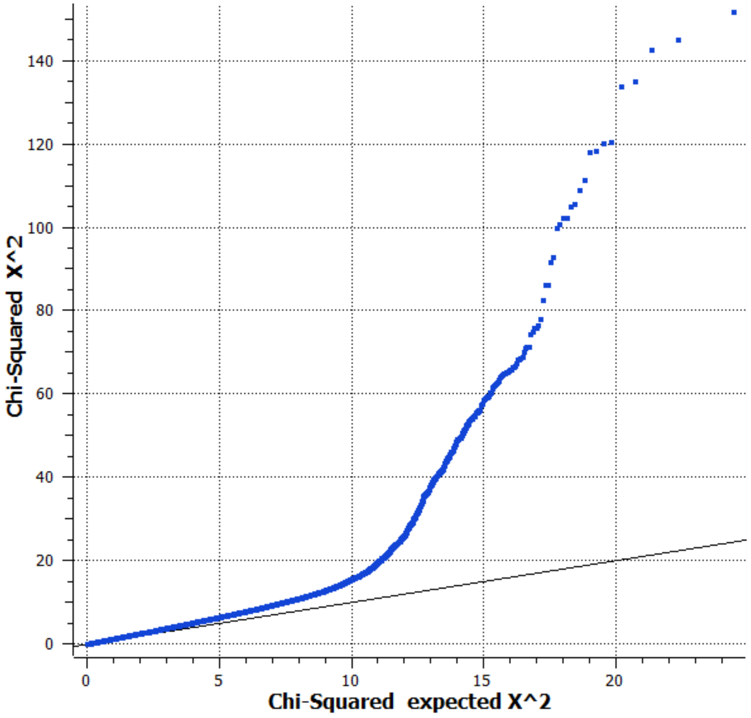
The “Gotchas”

A sampling of issues GHI has observed in dbGaP and elsewhere:

- **Gender discrepancies**
- **Cryptic relatedness**
- **Phenotype data formatted differently between sample groups in a study**
- **Incomplete matching of subjects between raw and processed genotype data.**
  - Example: 500 with raw data, 510 with processed data, 495 with both.
- **Nsp/Sty mismatches in Affy 500k data**
- **Batch effects processed genotypes**

Example of Batch Effects in a Multi-Center Study





- **Caucasian controls from one center have very different allele frequencies than the Caucasian controls from another center…**

GEO – Gene Expression Omnibus

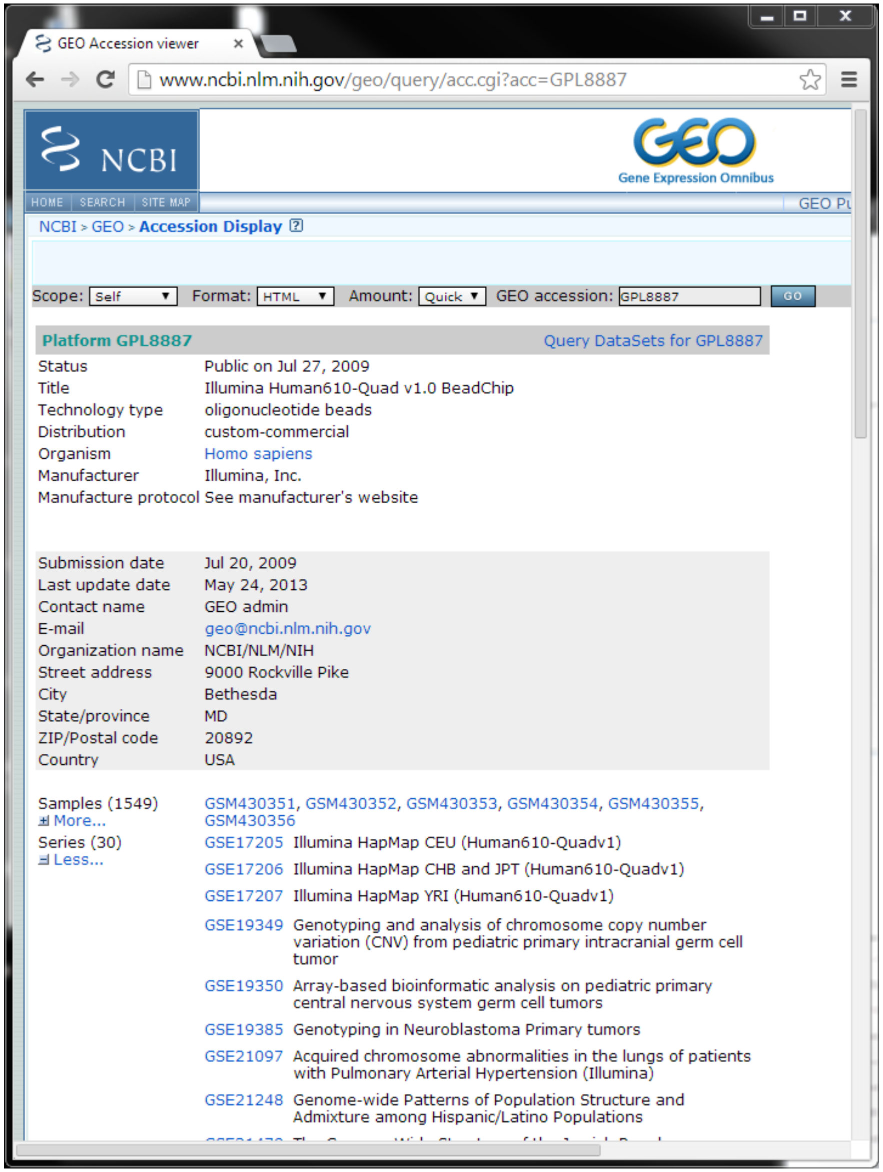
- **“GEO is a public functional genomics data repository… Tools are provided to help users query and download experiments and curated gene expression profiles.”**
- **Primarily a gene expression database, but also includes extensive genotype data**
- **Data access:**
  - “Anybody can access and download public GEO data. There are no login requirements.”
  - “NCBI places no restrictions on the use or distribution of the GEO data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted.”

GEO Data Profile

- **3413 studies, 1335 with human data (1239 mouse, 311 rat, etc.)**
- **Genotype data among the human datasets:**
  - 730 datasets flagged as containing some SNP array data
    - 11,715 samples among 200 data series for Affy 6.0
    - 9689 samples in 152 series for Affy 250k-Nsp
    - 1757 samples in 26 series for Illumina Omni-1
    - 1245 samples in 11 series for Illumina 550k
- **Sample sizes are generally much smaller than with dbGaP**
- **Many studies are based on somatic tissues**
- **GEO database structure is sample oriented, very detailed, and very different from dbGaP**

GEO: Browsing the Database

- **Browse data by platform to get data for every sample or study to use a particular chip.**
  - 762 samples in 24 studies using Illumina Human1M-Duo.
- **Browse by study design to get data for similar types of studies.**
  - 403 results for “SNP genotyping by SNP array.”
  - 654 results for “Genome variation profiling by SNP array.”



NCBI > GEO > **Accession Display**

Scope: Self Format: HTML Amount: Quick GEO accession: GPL8887

Platform GPL8887 Query DataSets for GPL8887

Status	Public on Jul 27, 2009
Title	Illumina Human610-Quad v1.0 BeadChip
Technology type	oligonucleotide beads
Distribution	custom-commercial
Organism	Homo sapiens
Manufacturer	Illumina, Inc.
Manufacture protocol	See manufacturer's website

Submission date	Jul 20, 2009
Last update date	May 24, 2013
Contact name	GEO admin
E-mail	geo@ncbi.nlm.nih.gov
Organization name	NCBI/NLM/NIH
Street address	9000 Rockville Pike
City	Bethesda
State/province	MD
ZIP/Postal code	20892
Country	USA

Samples (1549) GSM430351, GSM430352, GSM430353, GSM430354, GSM430355, GSM430356
More...

Series (30)
Less...

GSE17205	Illumina HapMap CEU (Human610-Quadv1)
GSE17206	Illumina HapMap CHB and JPT (Human610-Quadv1)
GSE17207	Illumina HapMap YRI (Human610-Quadv1)
GSE19349	Genotyping and analysis of chromosome copy number variation (CNV) from pediatric primary intracranial germ cell tumor
GSE19350	Array-based bioinformatic analysis on pediatric primary central nervous system germ cell tumors
GSE19385	Genotyping in Neuroblastoma Primary tumors
GSE21097	Acquired chromosome abnormalities in the lungs of patients with Pulmonary Arterial Hypertension (Illumina)
GSE21248	Genome-wide Patterns of Population Structure and Admixture among Hispanic/Latino Populations

Using GEO Data

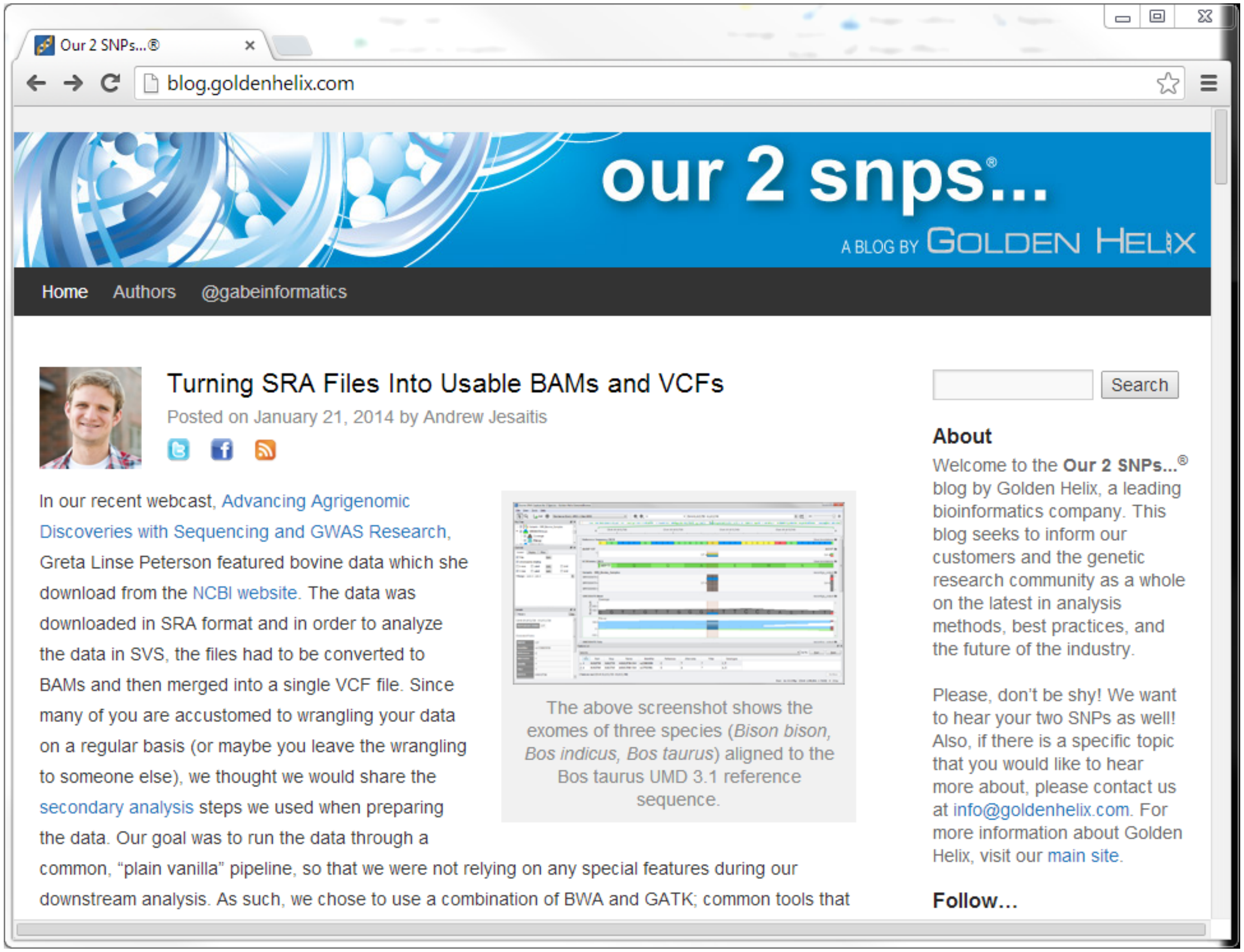
- **GEO is a good resource for test data and reference data.**
- **There are a few large GWAS studies, but not many.**
- **GEO has several human diversity reference panels available for various genotyping arrays.**
  - Illumina posts HapMap data there for many of their arrays.
  - Other diversity panels from NIA, Mayo, others.
- **Raw and processed data formats are usually available.**
- **“Series Matrix File” is a plain text format that is fairly easy to work with.**

SRA: Sequence Read Archive

Sequence Read Archive

- **SRA…**
  - “Archives raw oversampling NGS data for various genomes from several platforms”
  - “Shares NGS data with EMBL and DDBJ”
  - “Serves as a starting point for ‘secondary analysis’”
  - Provides access to data from human clinical samples to authorized users who agree to the dataset’s privacy and usage mandates.”
- **SRA primarily stores reads reads (SRA/fastq) and alignments (BAM)**
- **SRA hosts sequence data for some dbGaP and EGA studies**
  - Data not part of public SRA, but searchable summaries do appear on SRA.
- **PubMed abstracts can be linked to research data on SRA**

Our Team’s Experience with SRA



Turning SRA Files Into Usable BAMs and VCFs

Posted on January 21, 2014 by Andrew Jesaitis

In our recent webcast, Advancing Agrigenomic Discoveries with Sequencing and GWAS Research, Greta Linse Peterson featured bovine data which she download from the NCBI website. The data was downloaded in SRA format and in order to analyze the data in SVS, the files had to be converted to BAMs and then merged into a single VCF file. Since many of you are accustomed to wrangling your data on a regular basis (or maybe you leave the wrangling to someone else), we thought we would share the secondary analysis steps we used when preparing the data. Our goal was to run the data through a common, “plain vanilla” pipeline, so that we were not relying on any special features during our downstream analysis. As such, we chose to use a combination of BWA and GATK; common tools that

The above screenshot shows the exomes of three species (*Bison bison, Bos indicus, Bos taurus*) aligned to the Bos taurus UMD 3.1 reference sequence.

About

Welcome to the **Our 2 SNPs...**® blog by Golden Helix, a leading bioinformatics company. This blog seeks to inform our customers and the genetic research community as a whole on the latest in analysis methods, best practices, and the future of the industry.

Please, don’t be shy! We want to hear your two SNPs as well! Also, if there is a specific topic that you would like to hear more about, please contact us at info@goldenhelix.com. For more information about Golden Helix, visit our main site.

Follow...

- **A recent Golden Helix webcast featured bison and cattle sequence data from SRA. Read about it on our blog!**

EGA: European Genome-Phenome Archive

european genome-phenome archive

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

- **European equivalent of dbGaP**
- **Many EGA datasets are searchable on dbGaP**
- **May be most familiar as the repository for the WTCCC GWAS data**
- **From 2013 IGES talk by Justin Paschall:**
  - Over 450 studies in EGA
  - Extensive sequence data, including 110k BAM files and 35k fastq
  - Current submission rate of about 30TB/month
- **From personal experience: don't forget to request the decryption key…**

A Few More Sources

- **Illumina provides example data for most of their genotyping chips**
  - Complete HapMap Phase 2 populations for some, subset for others
- **Major imputation software developers have 1000 Genomes reference panels available in their preferred input formats**
  - Beagle
  - Impute2
  - MACH
- **Golden Helix offers several public datasets for download from within SVS**
  - HapMap data for various genotyping chips
  - 1000 Genomes
  - Complete Genomics

Agenda

1 Why Use Public Data?

2 Where to Find Public Data

3 Tips for Using Public Data

4 Manipulating Public Data in SVS

Final Tips for Using Public Data

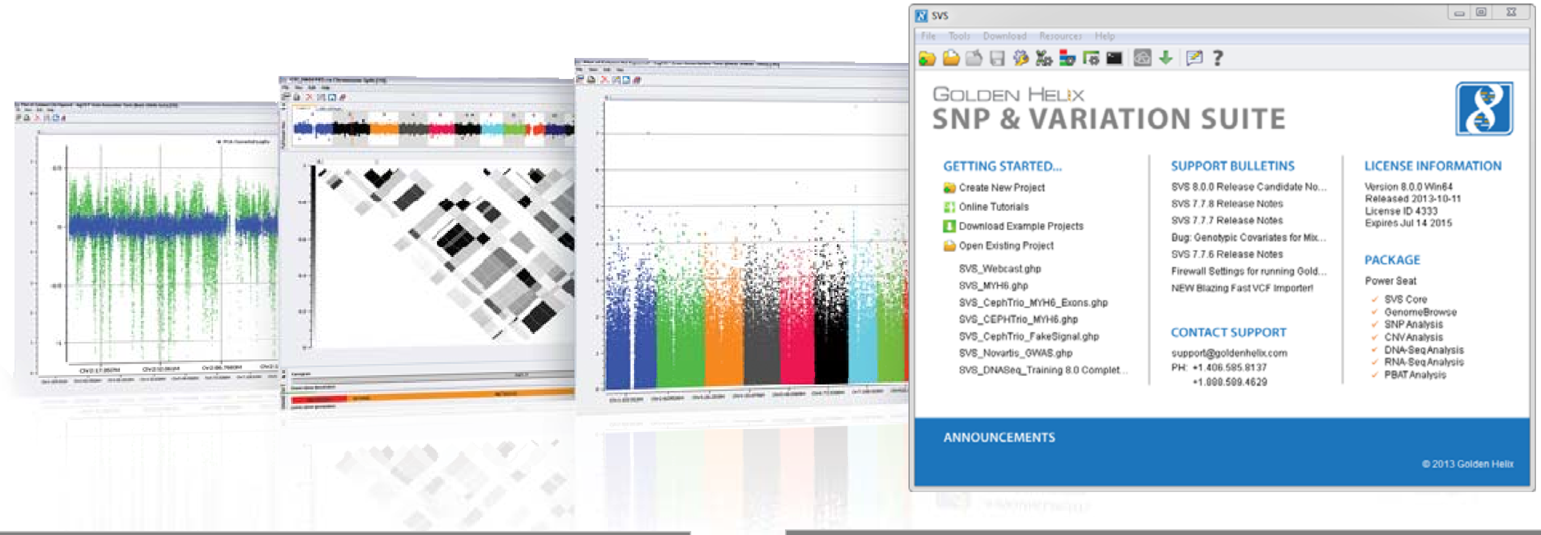
- **Read the documentation BEFORE you download the full archive**
- **Be vigilant with QC**
- **You can't be too careful, especially when combining data from multiple sources**
  - Start from raw data and process each source with a standard protocol. Re-calling genotypes is never a bad idea.
  - Pay special attention to strand orientation
  - Best if all sources were genotyped with the same array, but consider using imputation to combine data from mismatched arrays
  - Always adjust statistical tests for the data source
- **Examine results carefully before reporting or publishing**
  - Give special attention to results involving rare alleles.
  - If something seems fishy, it probably is.

Challenges of Public Data

Some of the challenges we hear about at Golden Helix:

- **“These files are really big!”**
  - Welcome to the world of bioinformatics. Small hard drives need not apply.
- **“Do I need a Linux computer to work with dbGaP data?”**
  - No, but if you’re in Windows, you will find that a Linux emulator like CygWin is very useful for manipulating the data. Compression utilities like WinRar and 7-Zip may also be helpful.
- **“There are a bunch of different data formats here…”**
  - Many of the standard formats you find on dbGaP and elsewhere can be read by SVS. Contact us if you’re not sure about a particular file—we might already have an import script that will work with it.
- **“I can read the data in text files, but it needs some serious manipulation before I can use it.”**
  - Data manipulation? That’s one of the most powerful features in SVS…

SNP & Variation Suite (SVS)



GOLDEN HELIX
SNP & VARIATION SUITE

GETTING STARTED...

- Create New Project
- Online Tutorials
- Download Example Projects
- Open Existing Project

SVS_Webcast.ghp
SVS_MYH6.ghp
SVS_CephTrio_MYH6_Exons.ghp
SVS_CEPHTrio_MYH6.ghp
SVS_CephTrio_FakeSignal.ghp
SVS_Novartis_GWAS.ghp
SVS_DNASeq_Training 8.0 Complet...

SUPPORT BULLETINS

SVS 8.0.0 Release Candidate No...
SVS 7.7.8 Release Notes
SVS 7.7.7 Release Notes
Bug: Genotypic Covariates for Mix...
SVS 7.7.6 Release Notes
Firewall Settings for running Gold...
NEW Blazing Fast VCF Importer!

CONTACT SUPPORT

support@goldenhelix.com
PH: +1.406.585.8137
+1.888.589.4629

LICENSE INFORMATION

Version 8.0.0 Win64
Released 2013-10-11
License ID 4333
Expires Jul 14 2015

PACKAGE

Power Seat

- ✓ SVS Core
- ✓ GenomeBrowse
- ✓ SNP Analysis
- ✓ CNV Analysis
- ✓ DNA-Seq Analysis
- ✓ RNA-Seq Analysis
- ✓ PBAT Analysis

ANNOUNCEMENTS

© 2013 Golden Helix

Core Features

- Powerful Data Management
- Rich Visualizations
- Robust Statistics
- Flexible
- Easy-to-use

Applications

- Genotype Analysis
- DNA sequence analysis
- CNV Analysis
- RNA-seq differential expression
- Family Based Association

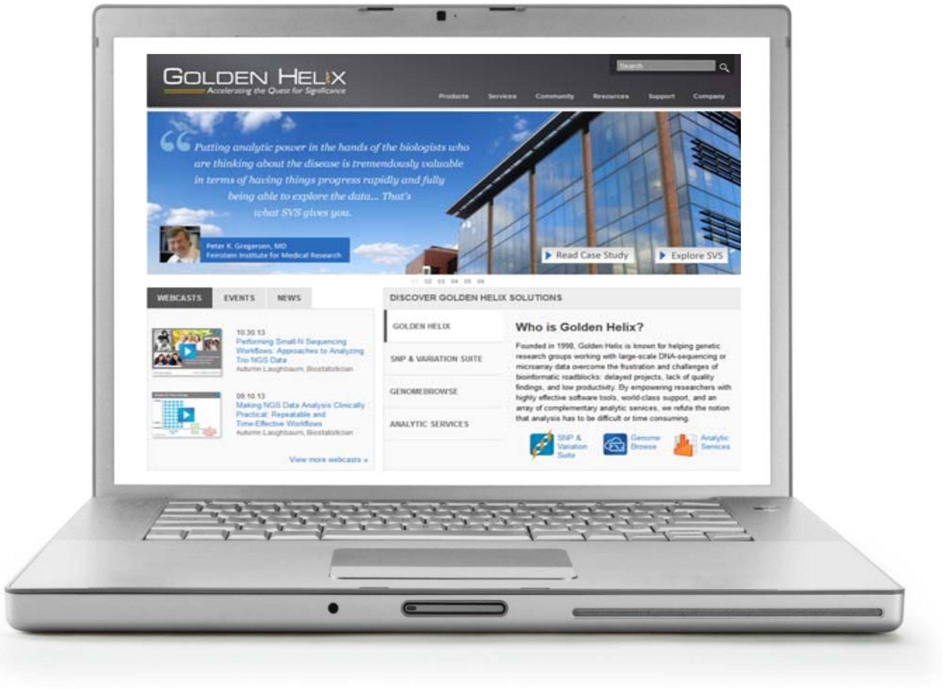
GOLDEN HELIX
Accelerating the Quest for Significance™

GOLDEN HELIX
SNP & VARIATION SUITE 8

[Demonstration]

Questions or more info:

- Email info@goldenhelix.com
- Request an evaluation of the software at www.goldenhelix.com



"Putting analytic power in the hands of the biologists who are thinking about the disease is tremendously valuable in terms of having things progress rapidly and fully being able to explore the data... That's what SVS gives you.

Peter K. Gregersen, MD
Feinstein Institute for Medical Research

10.30.13
Performing Small-N Sequencing Workflows: Approaches to Analyzing Trio NGS Data
Autumn Laughbaum, Biostatistician

09.10.13
Making NGS Data Analysis Clinically Practical: Repeatable and Time-Effective Workflows
Autumn Laughbaum, Biostatistician

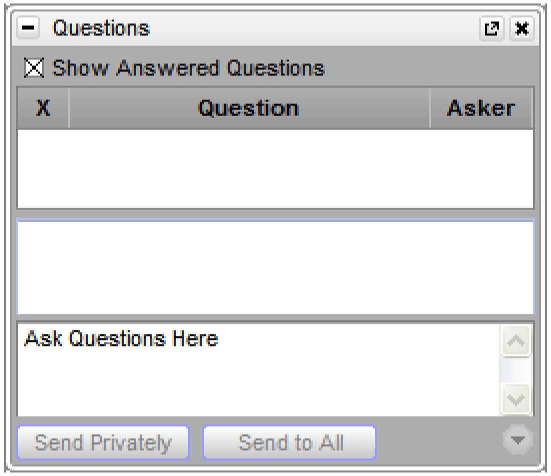
View more webcasts »

Who is Golden Helix?

Founded in 1998, Golden Helix is known for helping genetic research groups working with large-scale DNA-sequencing or microarray data overcome the frustration and challenges of bioinformatic roadblocks: delayed projects, lack of quality findings, and low productivity. By empowering researchers with highly effective software tools, world-class support, and an array of complementary analytic services, we refute the notion that analysis has to be difficult or time consuming.

Questions?

Use the Questions pane in your GoToWebinar window


Questions
☒ Show Answered Questions
X
Question
Asker
Ask Questions Here
Send Privately
Send to All