



WAREHOUSE

A scalable genetic data warehouse for VarSeq.

February 3, 2016

Andreas Scherer, PhD
President & CEO

Gabe Rudy
VP Product & Engineering



1 Overview Golden Helix

2 Introduction Genetic Data Warehousing

3 VSWarehousing: Concepts and Use Cases

4 Outline Early Adopter Program

Golden Helix – Who We Are



Golden Helix is a global bioinformatics company founded in 1998.

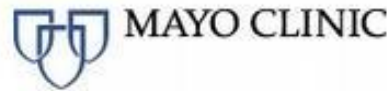


**Filtering and Annotation
Clinical Reports
Pipeline
Data Warehousing**

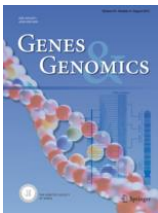
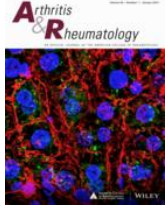
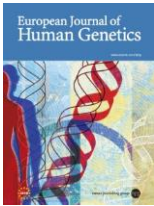
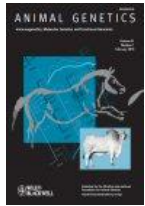


**GWAS
Genomic Prediction
Large-N-Population Studies
RNA-Seq
CNV-Analysis**

Over 300 customers globally



Cited in over 900 peer-reviewed publications

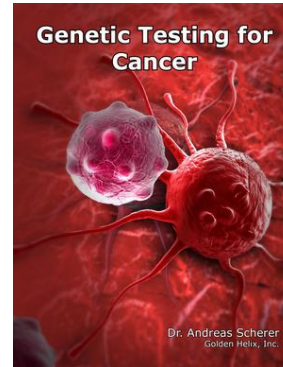


Golden Helix – Who We Are



When you choose a Golden Helix solution, you get more than just software

- REPUTATION
- TRUST
- EXPERIENCE



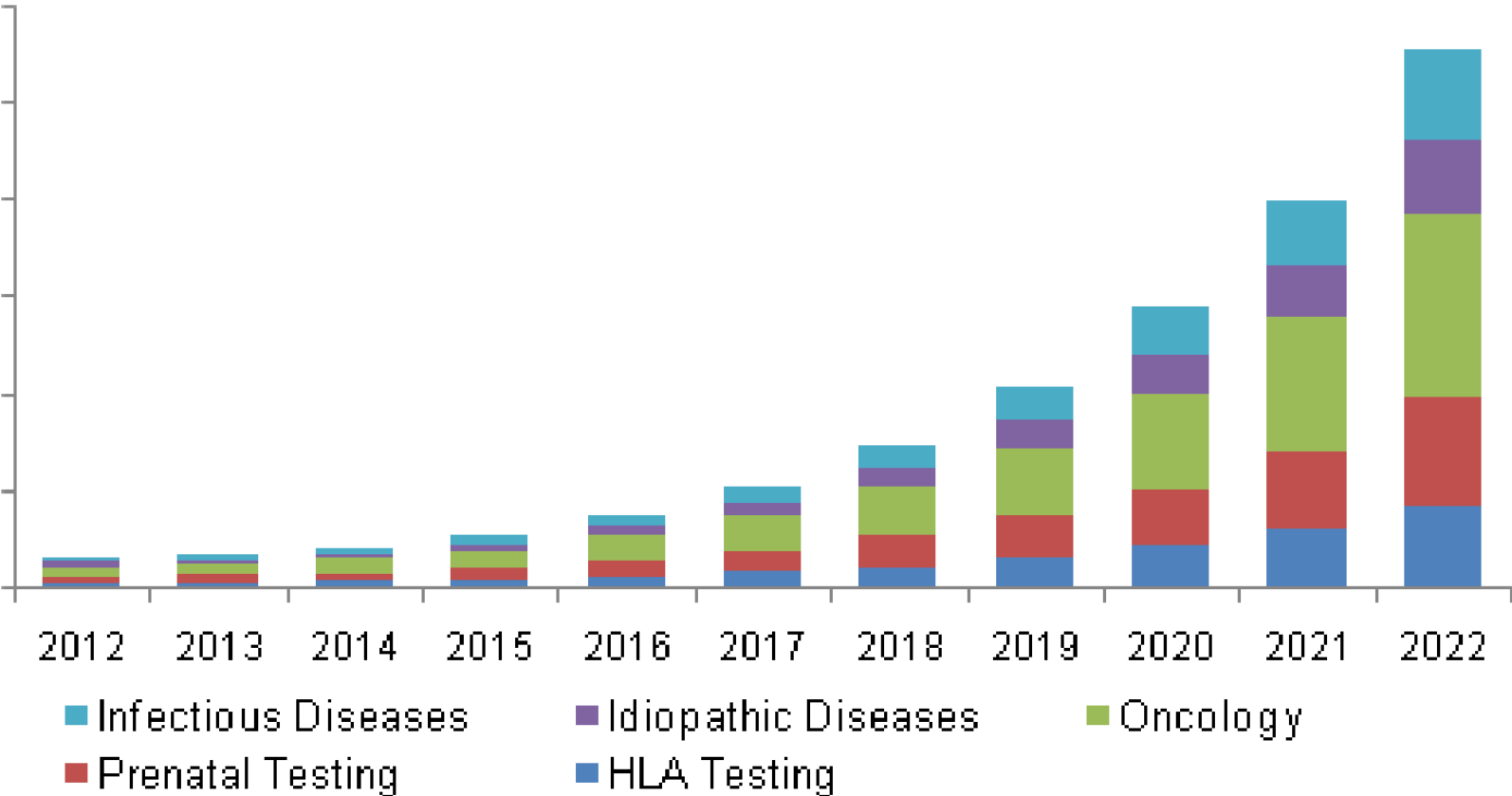
- INDUSTRY FOCUS
- THOUGHT LEADERSHIP
- COMMUNITY

- TRAINING
- SUPPORT
- RESPONSIVENESS



- TRANSPARENCY
- INNOVATION and SPEED
- CUSTOMIZATIONS

Precision Medicine unfolding

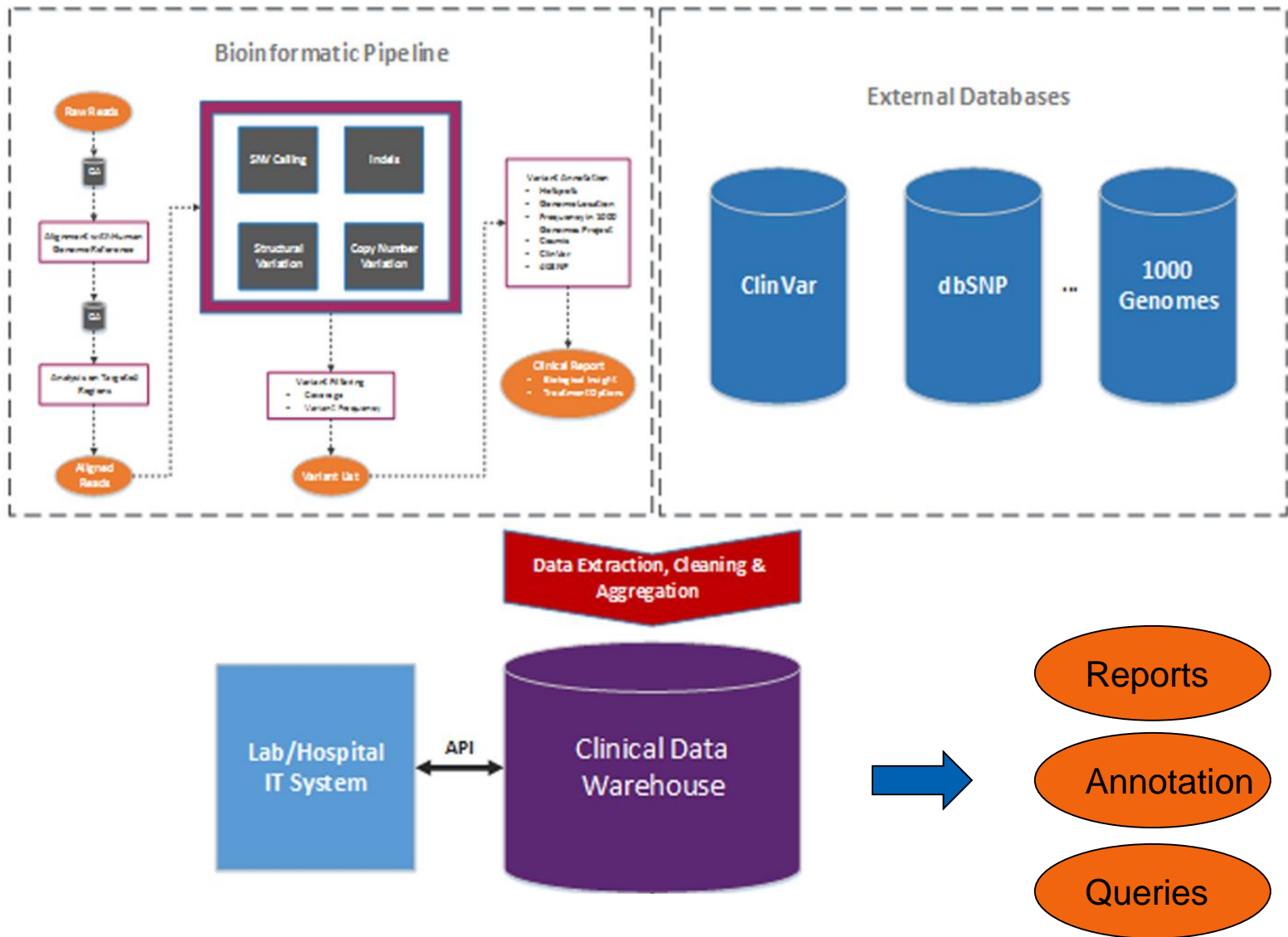


Labs have a lot of data – and more to come!



Billions of Variants
Terra Bytes of Data
+40% Annual Growth
Rate

Genetic Data Warehouse



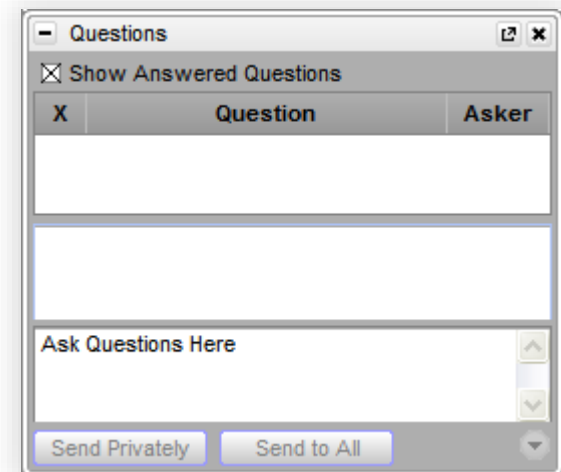


- **Annotation Source:** Have I seen this variant before? If so, at what frequency?
- **Conducting Research:** Capturing samples, reports: allowing to extract affected and unaffected study participants to conduct further research on a genomic level
- **Connecting with other legacy systems:** Integration point between lab and other hospital systems. Sharing data with other labs and research organizations.



Questions during the presentation

Use the Questions pane in your GoToWebinar window





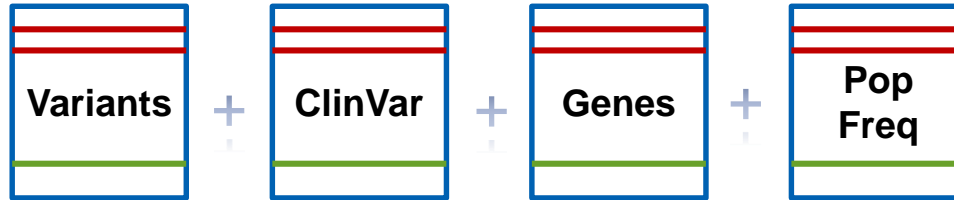
Test Launched

Ongoing Data Science, Re-Annotation, Medical Archiving

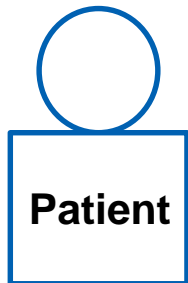
Annotate, Filter, Interpret Workflow

VCF

Variant Call File



Annotated Variants: Marked for Reporting



- Phenotype
- Lab Info
- Referring Info



Name

D.O.B.

Pathogenic ▼

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut ...

Incidental ▼

Ut enim ad minim veniam, quis

Structured



Golden Labs 203 Enterprise Blvd Phoenix, AZ 85016 Phone: 602.967.8137 Fax: 602.955.5555		Provider Information Physician: Dr. Nabilo Aminola Institution: ACME General Hospital Case Id: 01-1024	
Patient Information Name: Harrison Solo Gender: Male Date of Birth: 7/13/1942 Id: 1234		Sample Sample Site: Blood Sample Type: Blood Collection Date: 2/22/16 Collection Met.: Peripheral Draw Panel Coverage: 86.25% Report Date: 2/22/16	
Results Positive Mutations with an established somatic link detected.			
Interpretation Summary Although BRAF is most commonly associated with malignant melanoma, Lee et al. (2004) showed that BRAF is occasionally mutated in leukemias. As the patient presented acute leukemia and a mutation associated with leukemias was found in the BRAF gene, we recommend treatment take advantage of known drugs targeting mutations in this gene.			
Recommendations The recommended drugs targeting the BRAF mutation are included in the table below as well as 10 of the clinical trials currently underway. Investigation of the incidental findings may also lead to a mutation that can be targeted with a drug to treat the cancer.			
Incidental Findings: Mutations with a possible somatic link detected. This is a missense variant located in the TP53 gene. The transcription factor p53 responds to diverse cellular stresses to require target genes that induce cell cycle arrest, apoptosis, senescence, repress, or changes in metabolism. In addition, p53 appears to induce apoptosis through nontranscriptional cytoplasmic processes. In untreated p53 is kept inactive essentially through the action of the ubiquitin ligase MDM2 (123233), which inhibits p53 transcriptional activity and about p53 to promote its degradation. Numerous posttranslational modifications modulate p53 activity, most notably phosphorylation and acetylation. Several less abundant p53 isoforms also modulate p53 activity. Activity of p53 is ubiquitously low in human cancer either by mutation of the gene itself or by loss of cell signaling upstream or downstream of p53 (Cisplatin and MDM2: Sources: 1001; Spontaneous and iatrogenic, 2003). This gene has been observed to exhibit Autosomal recessive, Autosomal dominant, Somatic mutation, and Multifactorial inheritance pattern. It has been associated with Adrenal cortical carcinoma, Breast cancer; Choroid plexus papilloma, Colorectal cancer, Hepatocellular carcinoma, Li-Fraumeni syndrome, Nasopharyngeal carcinoma, Osteosarcoma, Pancreatic cancer, Basal cell carcinoma 7, and Glioma susceptibility 1. Li-Fraumeni syndrome (LFS) is a clinically and genetically heterogeneous inherited cancer syndrome. LFS is characterized by autosomal dominant inheritance and early onset of tumors, multiple tumors within an individual, and multiple affected family members. In contrast to other inherited cancer syndromes, which are predominantly characterized by site-specific cancers, LFS presents with a variety of tumor types. The most cor			

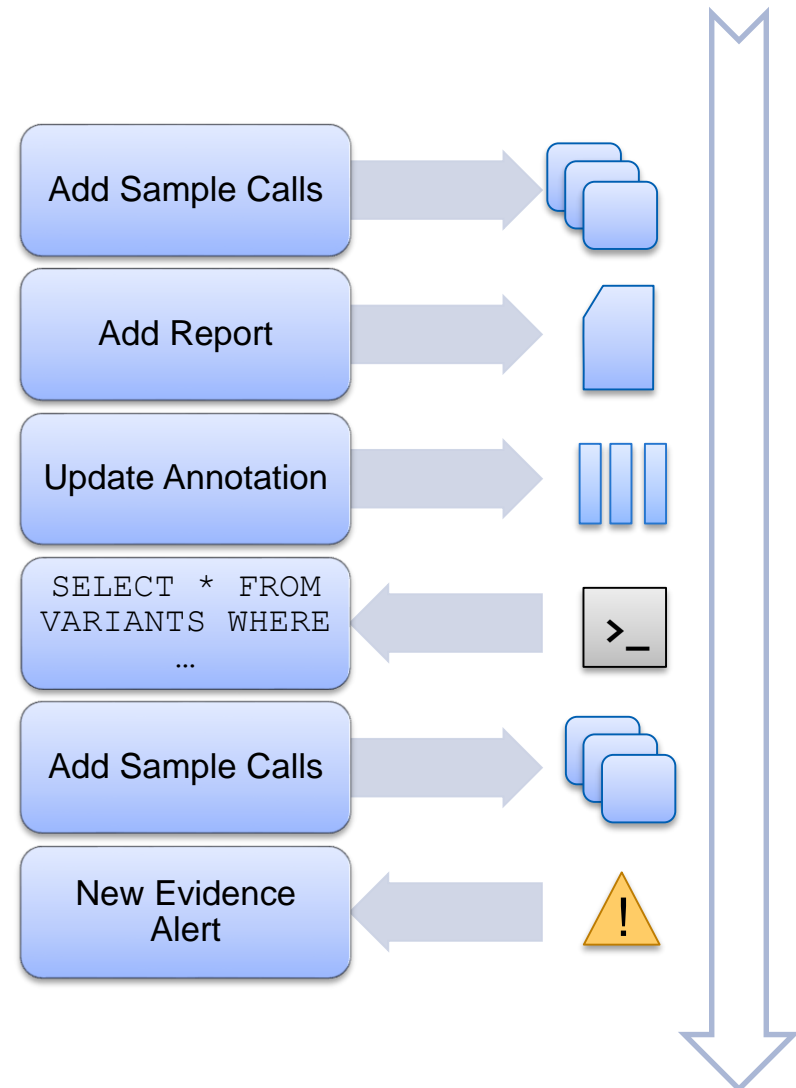
Rendered



Requirements for Variant Warehousing



- A place to archive full VCFs of every sequenced sample
- Query and retrieve subsets of data at any time
- Ask the Variant Warehouse:
 - Have I ever seen this variant in my previous test samples?
 - What proportion of samples? Frequency and genotype counts.
 - Does this gene contain other rare variants in my cohort?
 - Have I classified this variant and put it into a report for any previous samples?
 - ClinVar's monthly release has new and updated variant classifications. Are any of those variants I reported differently?



Variant Sites

CHR	POS	ID	...
-----	-----	----	-----

Variant Call Fields

GT	GQ	RD	AD	...
S1				

Annotations and Algs

ClinVar	Gene	Pop Freq	Allele Counts
---------	------	----------	---------------



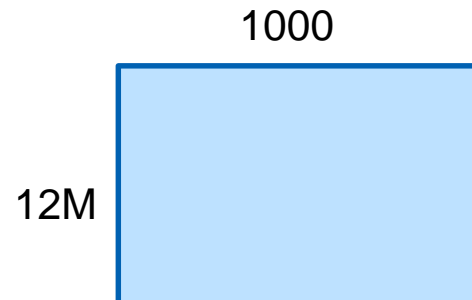
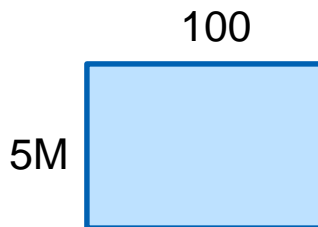
Samples

S1	Affected
S2	Affected
S3	Control
S4	Control
...	

Samples

Variants

Var Calls
Rows



Traditional RDBMS

- Insert/Update/Delete Rows
- Scales poorly
- Powerful query language (SQL)
- Unintuitive star topology (lots of joins)

NoSQL

- Add/Update/Remove value blobs for keys
- Scales well
- Poor query interface, only on manually indexed dimensions

Data Warehouse

- Batch updates to read-optimized matrixes
- Scales well (disk and query speeds)
- Intuitive unified topology
- Powerful query interface possible (SQL front-end)



ORACLE
DATABASE

The traditional row-based data storage approach is **dead**, as row-based storage will never match column-based storage's performance increase by factor 100x



Michael Stonebraker

Building on VarSeq's Genomic Toolset



■ Mature, Scalable Technology

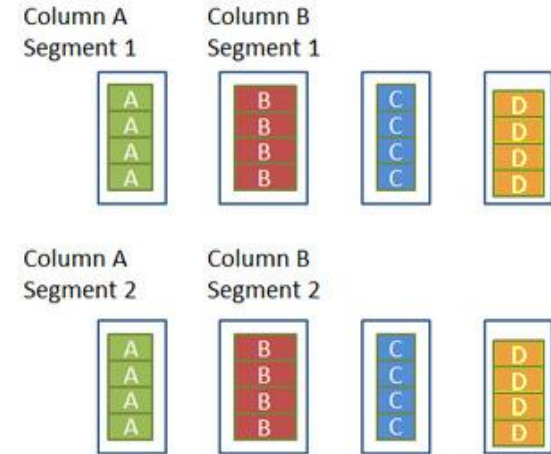
- Chunked, compressed column-store TSF
- Efficient querying, transparent joining
- Golden Helix core technology for 4 years

■ VarSeq Solves NGS Challenges

- Merge, shift and normalize variant calls
- Richly annotate genes, describe variants
- Current and relevant public annotations
- Licensed and private annotations
- QC and Cohort genomic algorithms

■ Integrated Experience

- Interpretation and visualization
- Report authoring and previewing
- Sample QC and cataloging
- Subsetting and exporting



Each field is chunked and compressed

Technology	Filter on Gene Effect + Sample Read Depth	Storage Size in Tables
PostgreSQL 9.4	6300 ms	2.5 GB
VSWarehouse	560 ms	150MB
Improvement	11x	16x



WAREHOUSE



Genomic Coordinate	RefAlt	Gene Names	HGVS p. (Clinically Relevant)
1.2494329	G/A	TNFRSF14	NP_003811.2 p.Val241Ile
1.11190803	C/T	MTOR	NP_004949.1 p.Glu1796Lys
1.27105927	-G	ARID1A	NP_006066.3 p.Gly1848fs
1.27105930	G/-	ARID1A	NP_006066.3 p.Asp1850fs
1.43814978	G/A	MPL	NP_005364.1 p.Ser505Asn

Web: Query, View, Export

- Uses “Harvest” biomedical toolkit from CHOP
- Open source, extendable
- Access to every variant and sample table

Research Exome Warehouse

- V2 (100 samples x 12M variants)
- V1 (50 samples x 6.5M variants)

Tumor Gene Panel Warehouse

- V1 (22 samples x 6k variants)

Exomes Dx Report (3 samples)

- Report – Genetic Variants (5 var)
- Report – Incident Findings (2 var)

Cancer Panel Report (6 samples)

- Report – Genetic Variants (22 var)
- Report – Incident Findings (46 var)

Research Exome Freq Tumor Gene Panel Freq

Common False Positives Abby Labs Collab Findings

Projects on VSWarehouse Server:

	# Samples	# Vars	Ver	Size
	25	131870	2	
	17	461613	1	

Update Create

Upload Samples from Projects

Genetic Variants

Primary Findings

Variant: 1:43814967 T/C (MPL)

Classification: Pathogenic

*Interpretation: **B I U**

This is a Missense Variant located in the MPL gene.

The MPL gene encodes the receptor for thrombopoietin (THPO; 600044), a

Input Report Interpretations, Sync



Viz and Annotate Warehouse Variants

Chr 1: 115256530 - G/T (Existing Record)

False Positive: Variant is False Positive

Custom Assessment Catalogs

```
warehouse=# SELECT count(*) FROM
p_exomes_1 WHERE
EffectCombined= 'LoF'
and HomoVar > 1;

count
-----
    305
(1 row)
```

API – SQL, REST, Python

Access

Store

Update



varSEQ™

Illumina TruSight Myeloid

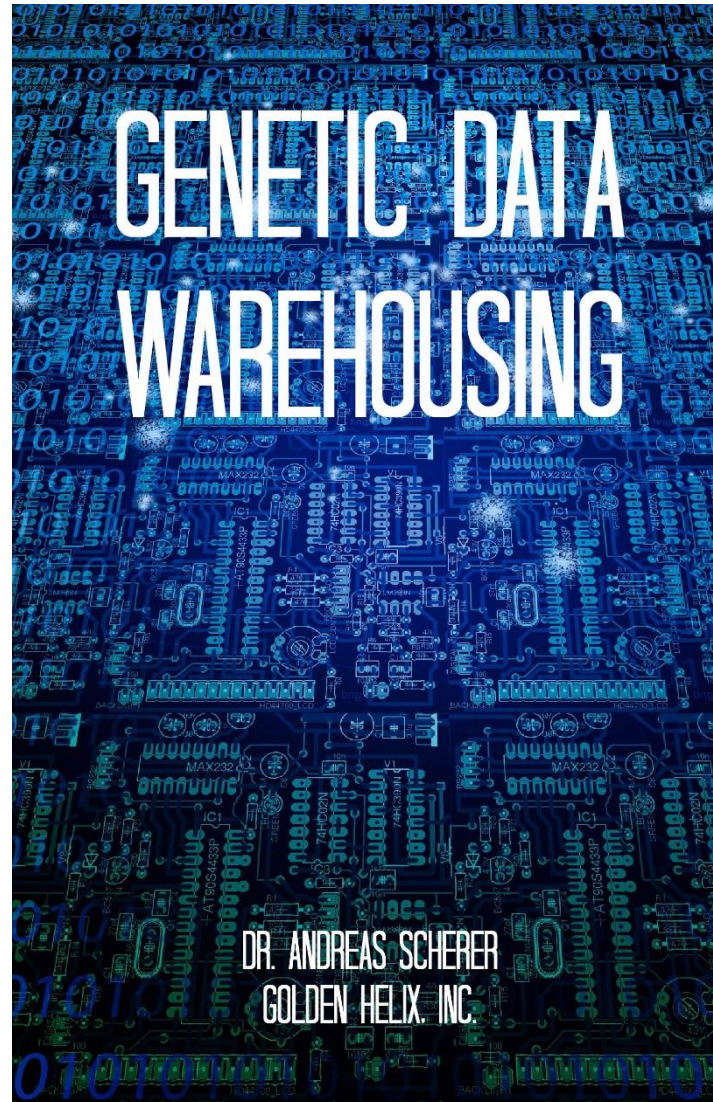
WAREHOUSE

Query Reports and Variants

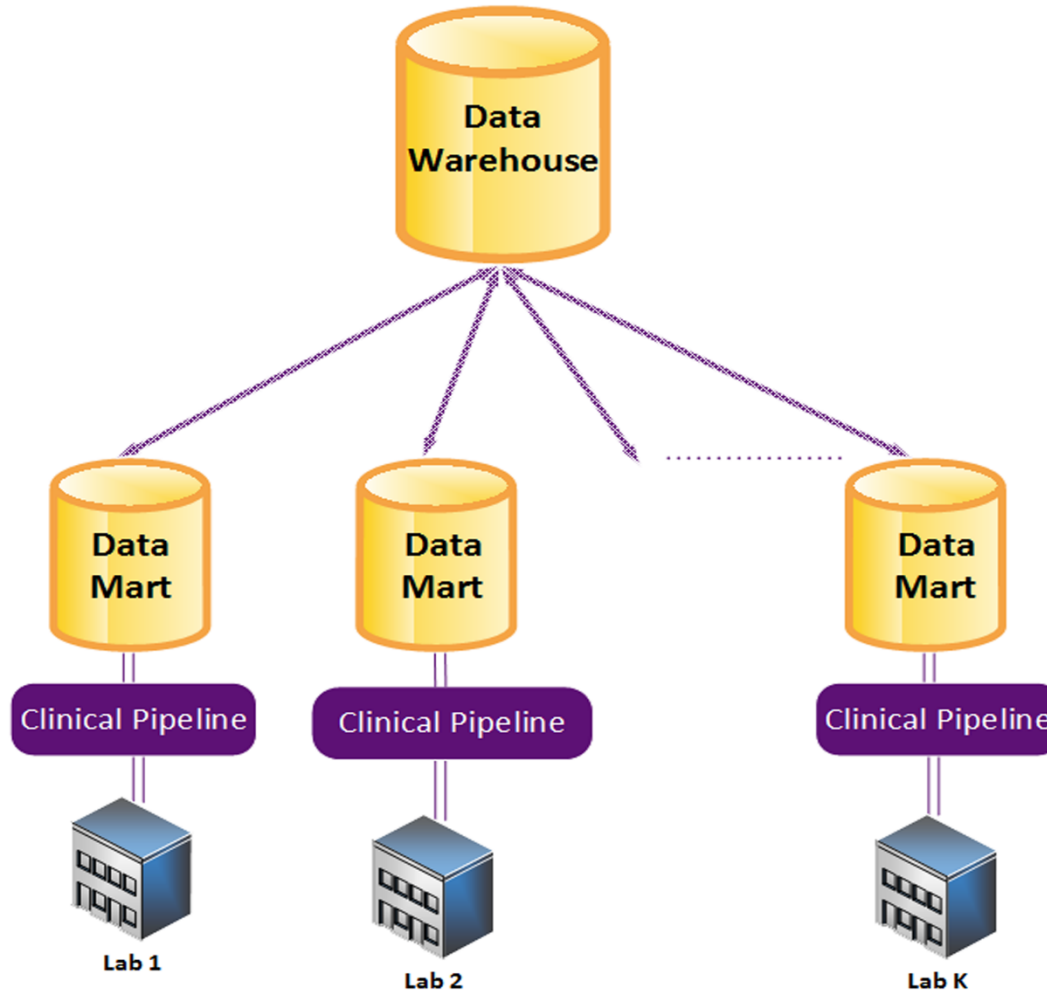
varSEQ™

Exome Trio Diagnostic

New e-book!



Architectures: Hub-Spoke Model



Launch timing



The Early Adopter Program



■ Our goal

- We want to build highly referenceable customers
- We are interested in a variety of use cases
- Case studies

■ What is in it for you?

- The ability to influence the product roadmap. Key features can be prioritized.
- Access to our complete tech stack
 - VarSeq, VSPipeline, VSReports at no charge
 - VSWarehouse at 50% discount
 - 15 month license
 - We are willing to commit to these terms for 3 years

- **Important fine-print: Commitment until March 15 2016.**
We limit this offer to three customers.



Questions or more info:

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





Questions?

Use the Questions pane in your GoToWebinar window

