WAREHOUSE

A scalable genetic data warehouse for VarSeq.

February 3, 2016

Andreas Scherer, PhD
President & CEO

Gabe Rudy
VP Product & Engineering

GOLDEN HELIX
Accelerating the Quest for Significance™

Agenda

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.

gsk GlaxoSmithKline

varSEQ™

Filtering and Annotation
Clinical Reports
Pipeline
Data Warehousing

GOLDEN HELIX SNP & VARIATION SUITE

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

Over 300 customers globally

SickKids
THE HOSPITAL FOR SICK CHILDREN

Cincinnati Children's
Hospital Medical Center

The Children's Hospital of Philadelphia

Stanford University

UBC

Ucla

Yale

The Feinstein Institute for Medical Research
North Shore-Long Island Jewish Health System

North Shore LIJ

PREVENTION GENETICS
DISEASE PREVENTION THROUGH GENETIC TESTING

MAYO CLINIC

BabyGenes™

LINEAGEN

MOFFITT CANCER CENTER

Dartmouth-Hitchcock
NORRIS COTTON CANCER CENTER

BAYER

Lilly

abbvie

TEVA

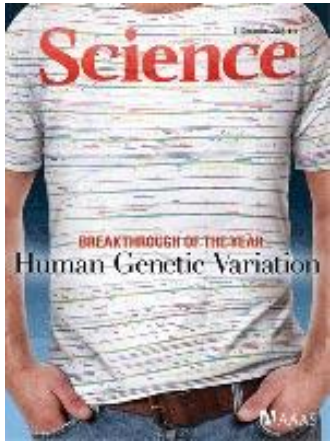
NATIONAL CANCER INSTITUTE

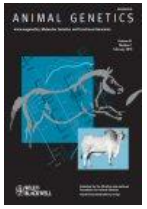
USDA
USDA Agricultural Research Service

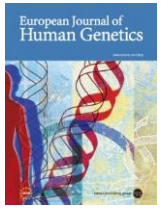
FDA

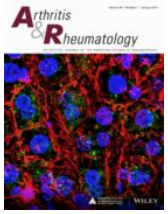
GOLDEN HELIX
Accelerating the Quest for Significance™

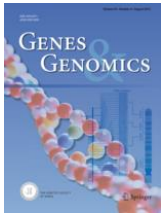
Cited in over 900 peer-reviewed publications

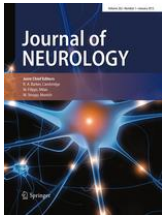








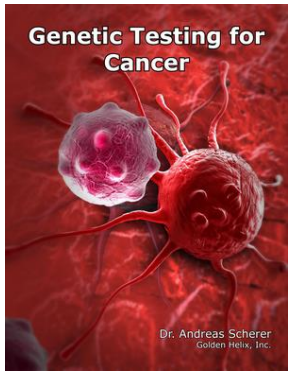




Golden Helix – Who We Are

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

- REPUTATION
- TRUST
- EXPERIENCE

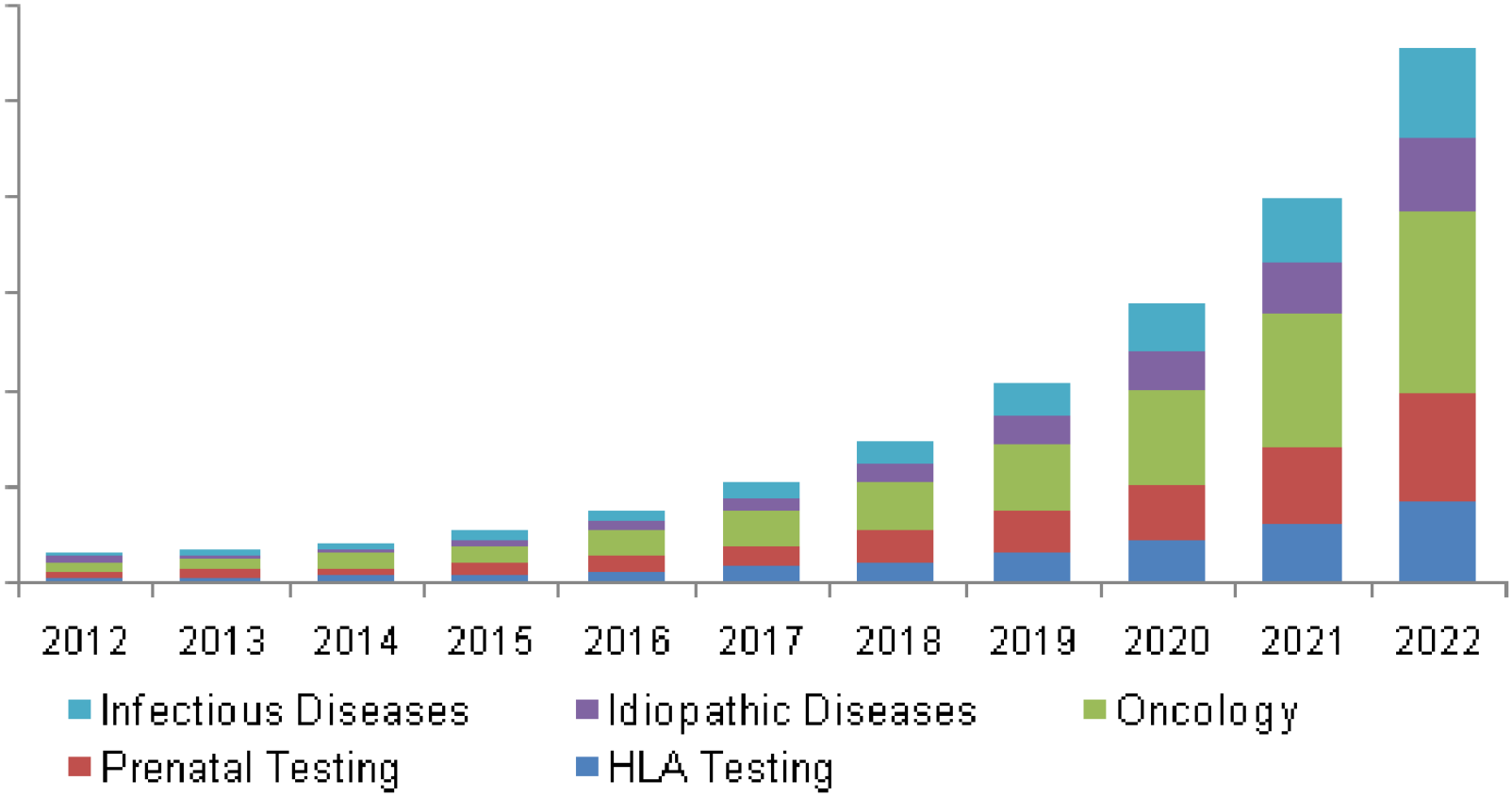


- INDUSTRY FOCUS
- THOUGHT LEADERSHIP
- COMMUNITY

- TRAINING
- SUPPORT
- RESPONSIVENESS

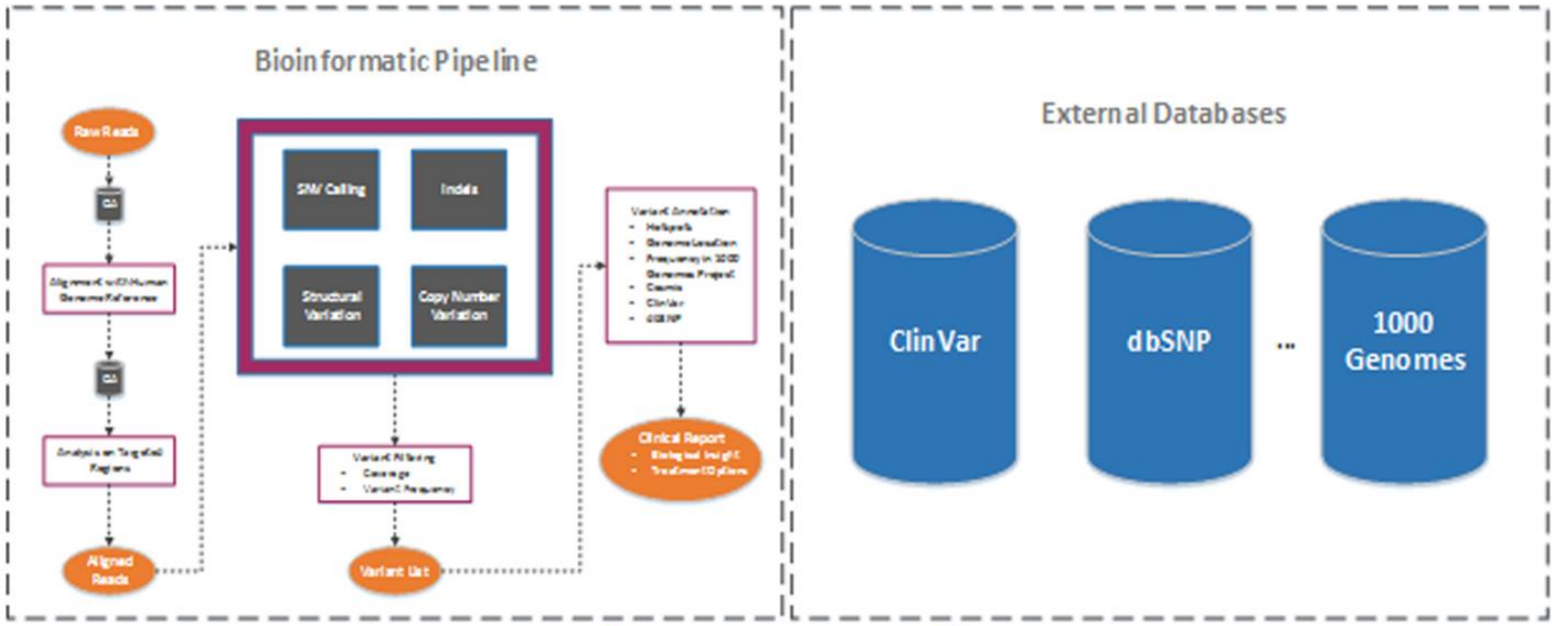
- TRANSPARENCY
- INNOVATION and SPEED
- CUSTOMIZATIONS

Precision Medicine unfolding



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

Genetic Data Warehouse



Data Extraction, Cleaning & Aggregation

Lab/Hospital IT System

API

Clinical Data Warehouse

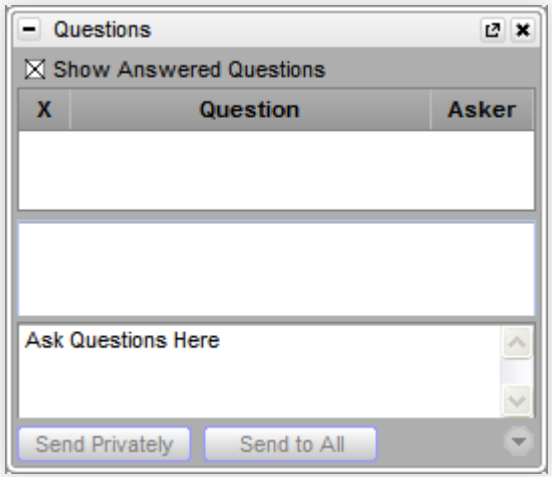
- Reports
- Annotation
- Queries

Use Cases

- **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



R&D Genetic Test Workflow

Cataloging of Tested Samples

Test Launched

Ongoing Data Science, Re-Annotation, Medical Archiving

Annotate, Filter, Interpret Workflow

VCF

Variant Call File

Variants + **ClinVar** + **Genes** + **Pop Freq**

Annotated Variants: Marked for Reporting

Patient

- 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

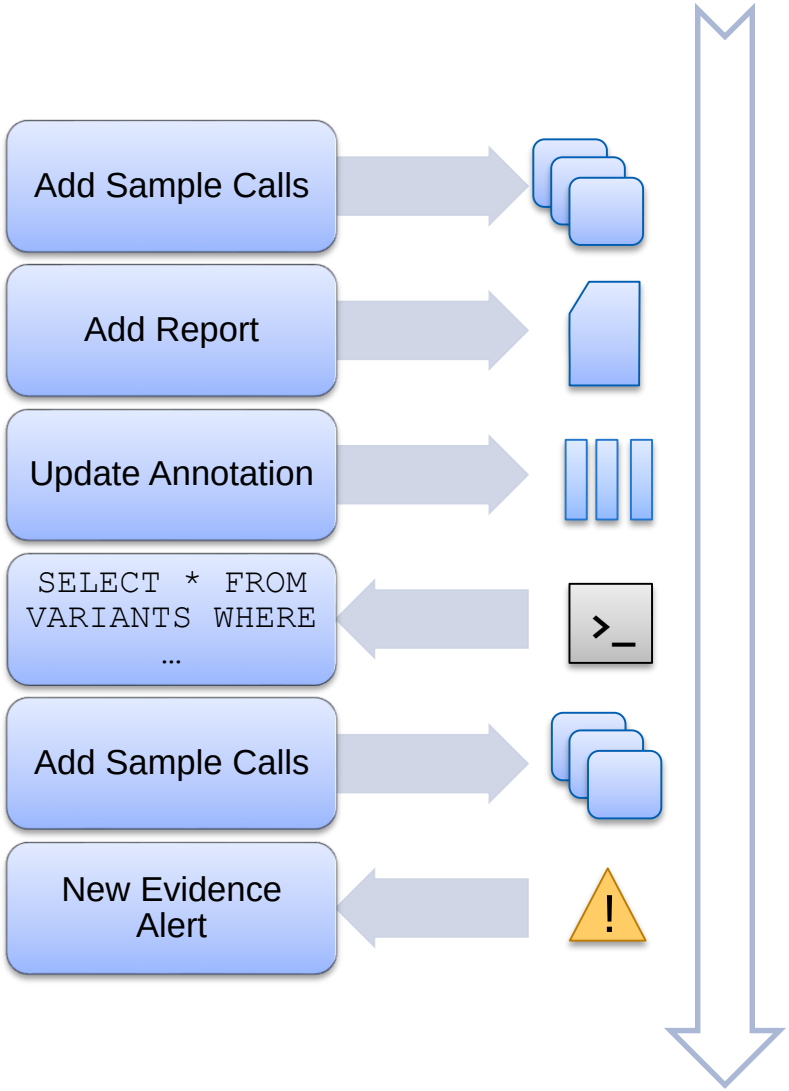
Ut enim ad minim veniam, quis

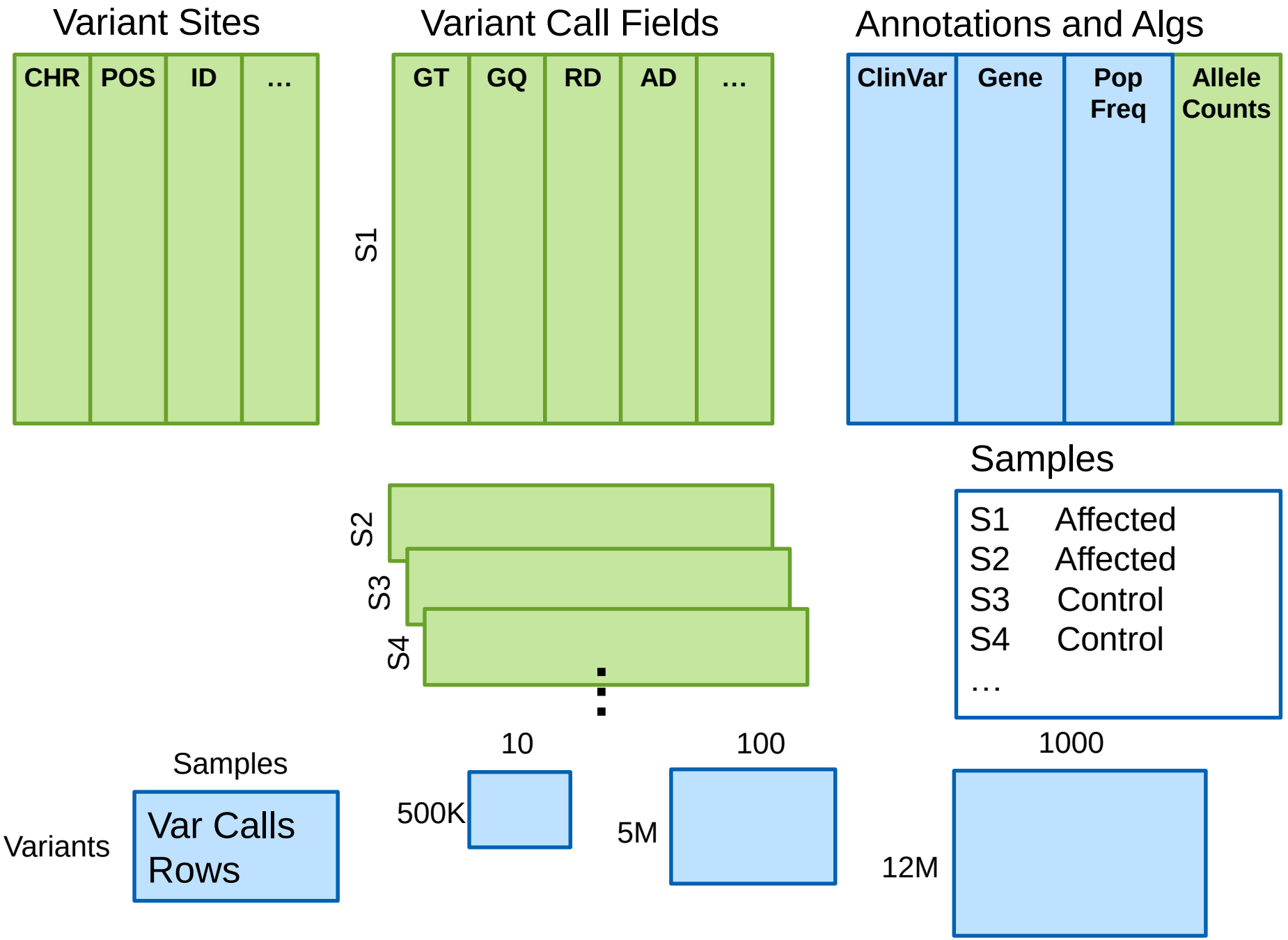
Structured

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
…
S2
S3
S4
10
100
1000
Samples
Variants
Var Calls Rows
500K
5M
12M

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)

MySQL PostgreSQL mongoDB CouchDB

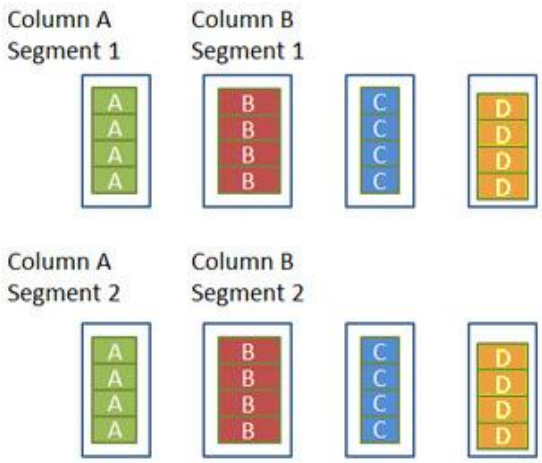
ORACLE DATABASE

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

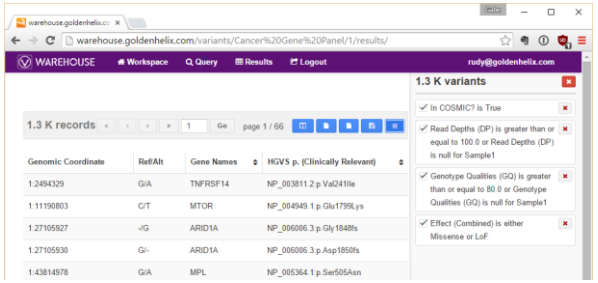


Each field is chunked and compressed

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

varSEQ™



Web: Query, View, Export

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

```
[warehouse=# SELECT count(*) FROM
             p_exomes_1 WHERE
             EffectCombined= 'LoF'
             and HomoVar > 1;
 count
-------
   305
(1 row)
```

API – SQL, REST, Python

Access

Store

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

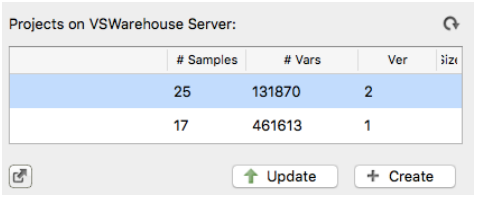
Tumor Gene Panel Warehouse
- V1 (22 samples x 6k variants)

Exomes Dx Report (3 samples)
- Report – Genetic Variants (5 var)
- Report – Incident Findings (2 var)

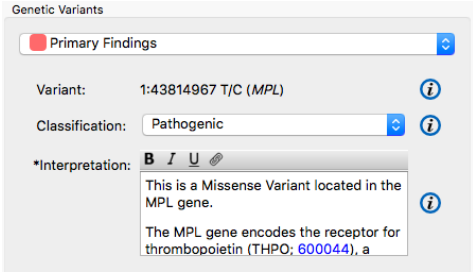
Cancer Panel Report (6 samples)
- Report – Genetic Variants (22 var)
- Report – Incident Findings (46 var)

Research Exome Freq
Tumor Gene Panel Freq

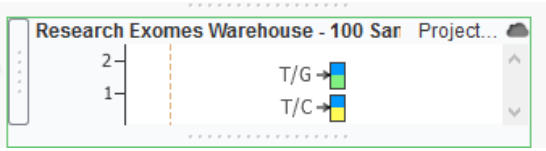
Common False Positives
Abby Labs Collab Findings



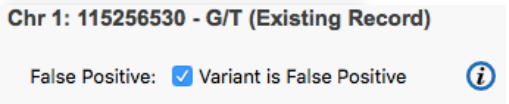
Upload Samples from Projects



Input Report Interpretations, Sync



Viz and Annotate Warehouse Variants



Custom Assessment Catalogs

Update

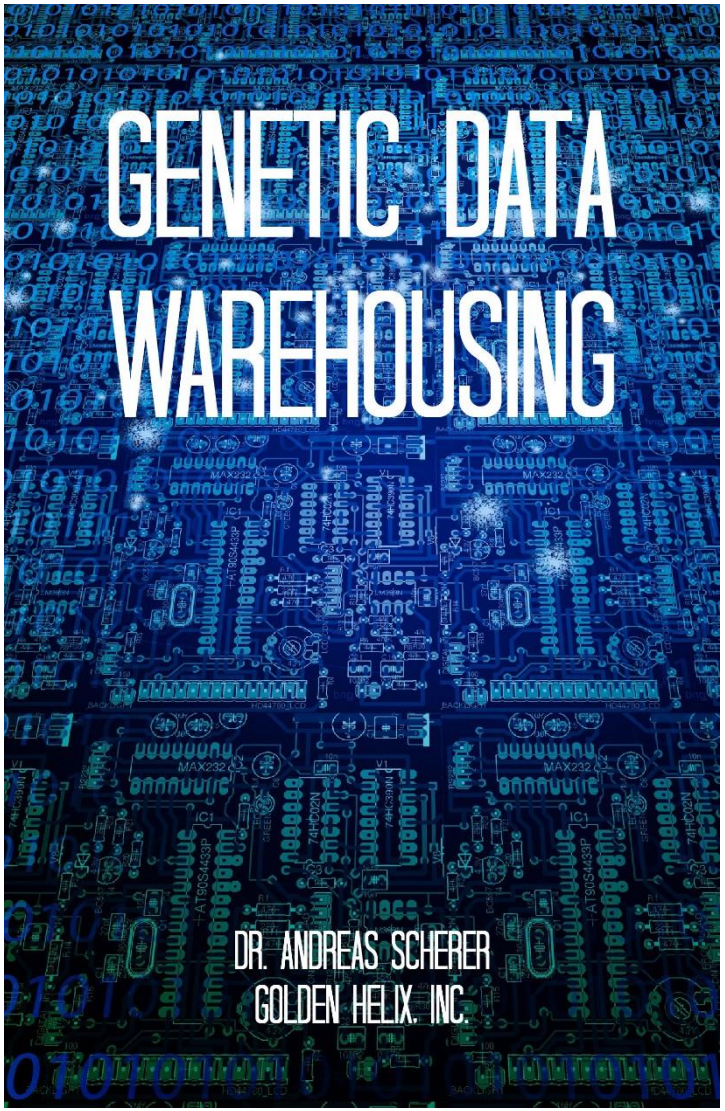
[Demonstration]

Illumina TruSight Myeloid

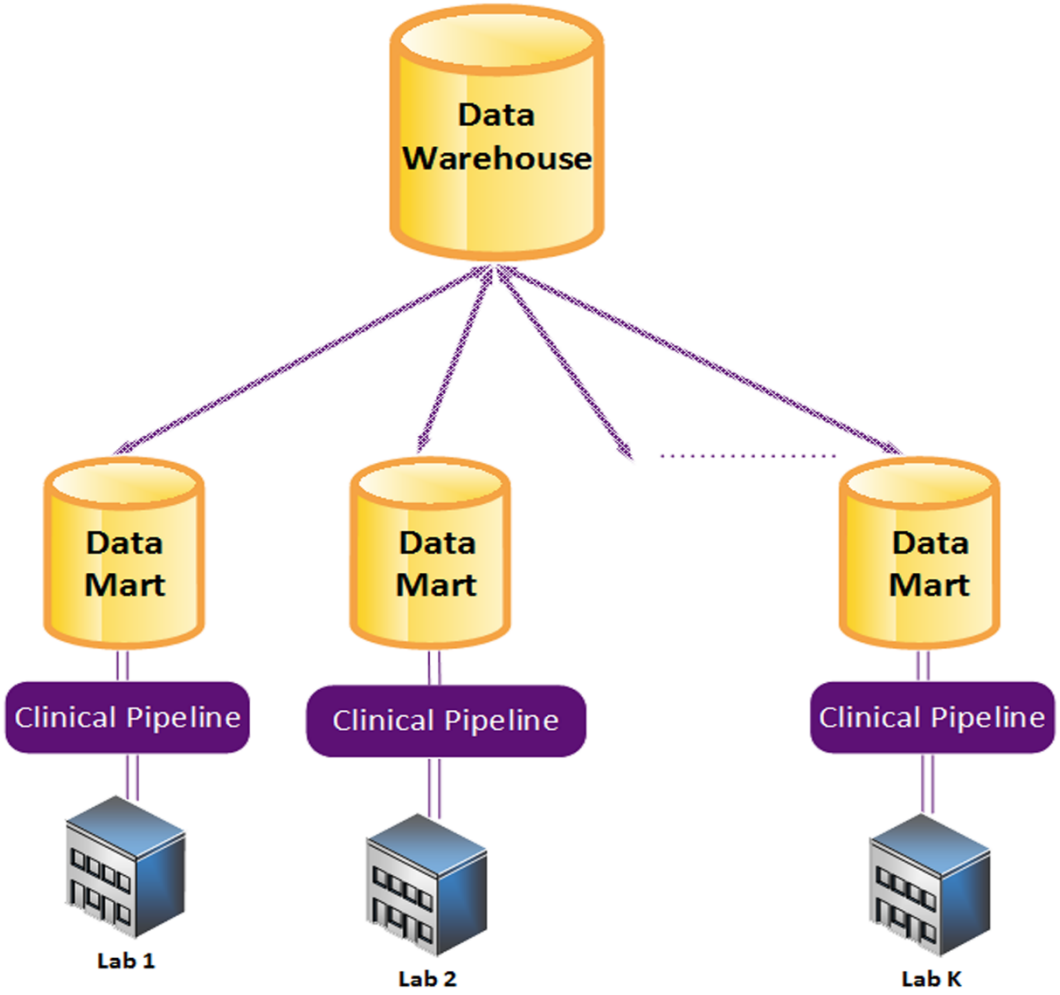
Query Reports and Variants

Exome Trio Diagnostic

New e-book!


GENETIC DATA WAREHOUSING
DR. ANDREAS SCHERER
GOLDEN HELIX, INC.

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