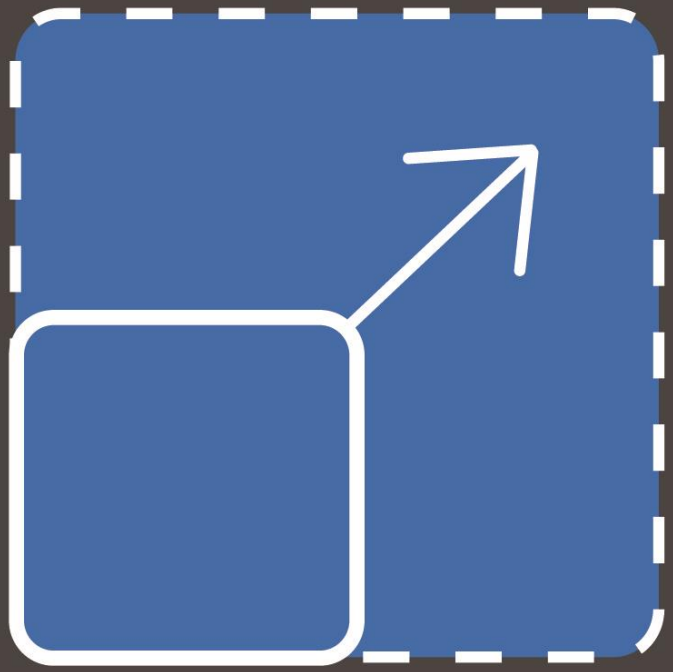


Big Data at Golden Helix: Scaling to Meet the Demand of Clinical and Research Genomics

September 21, 2016

Gabe Rudy
VP Product & Engineering

GOLDEN HELIX
Accelerating the Quest for Significance™

Agenda

1. Overview Golden Helix
2. Big Data in Genomics
3. Big Data at Golden Helix
4. Use Cases and Questions

Golden Helix – Who We Are

Golden Helix is a global bioinformatics company founded in 1998.

GlaxoSmithKline

varSEQ®

- Filtering and Annotation
- Clinical Reports
- Pipeline: Run Workflows

VARIANT WAREHOUSE

- Variant Warehouse
- Centralized Annotations
- Hosted Reports
- Sharing and Integration

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

LINEAGEN

MOFFITT CANCER CENTER

Dartmouth-Hitchcock
NORRIS COTTON CANCER CENTER

BAYER

Lilly

abbvie

TEVA

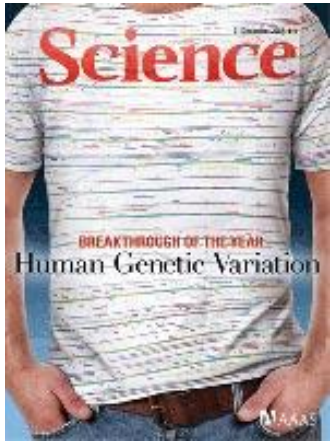
NATIONAL CANCER INSTITUTE

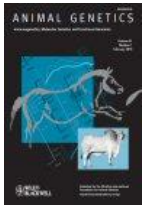
USDA
USDA Agricultural Research Service

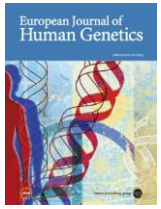
FDA

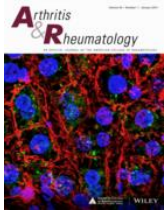
GOLDEN HELIX
Accelerating the Quest for Significance™

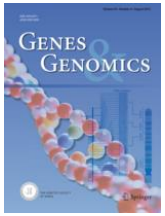
Cited in over 1000 peer-reviewed publications







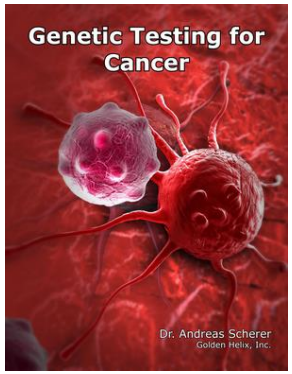




Golden Helix – Who We Are

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

- REPUTATION
- TRUST
- EXPERIENCE

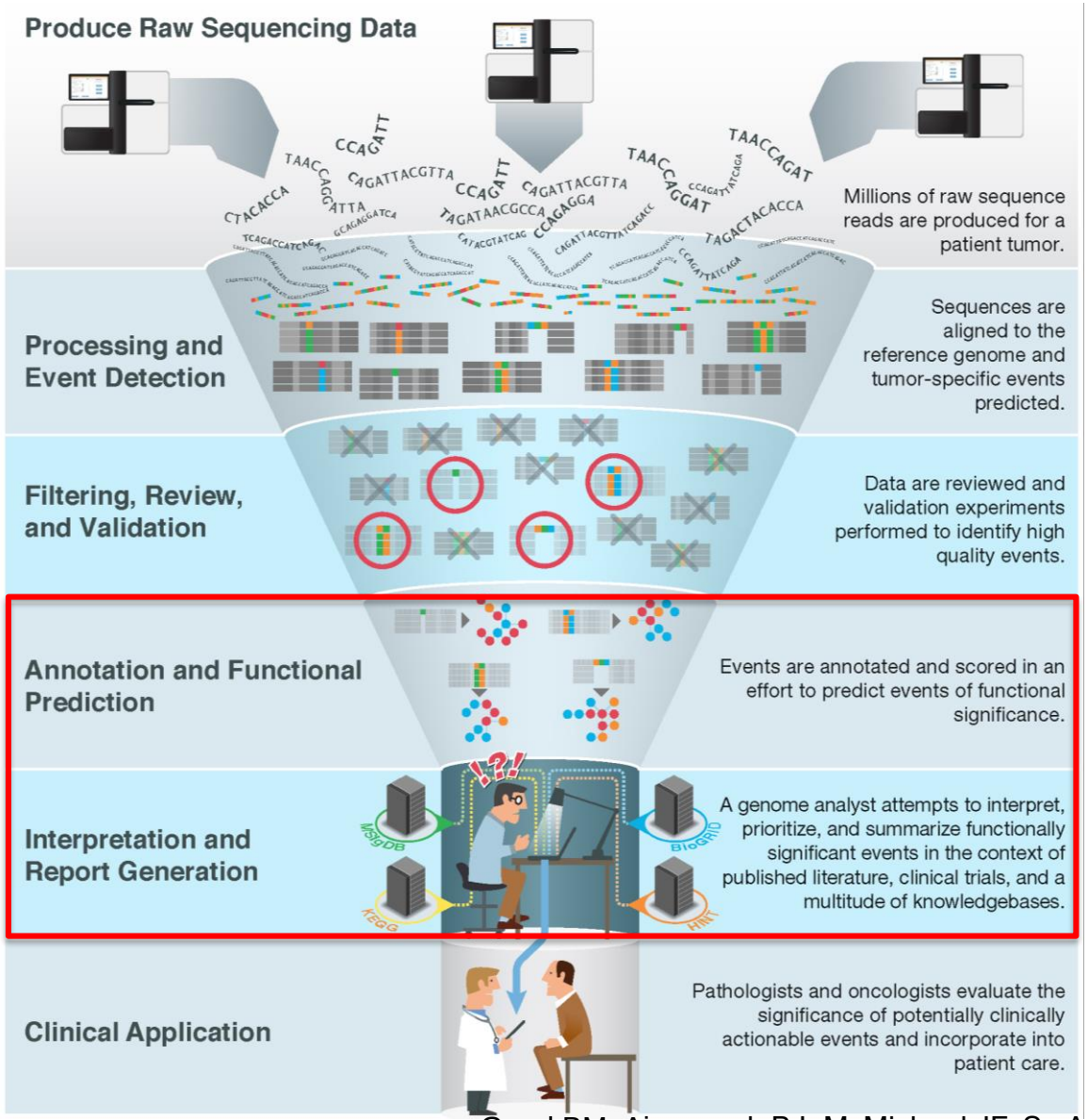


- INDUSTRY FOCUS
- THOUGHT LEADERSHIP
- COMMUNITY

- TRAINING
- SUPPORT
- RESPONSIVENESS

- TRANSPARENCY
- INNOVATION and SPEED
- CUSTOMIZATIONS

Path of Data to the Clinic



FASTQ Files:
Per Sample ~100GB

BAM Files:
Per Sample ~100GB

gVCF Files:
Per Sample ~2-5GB

Annotated Variants:
Per Sample ~100MB

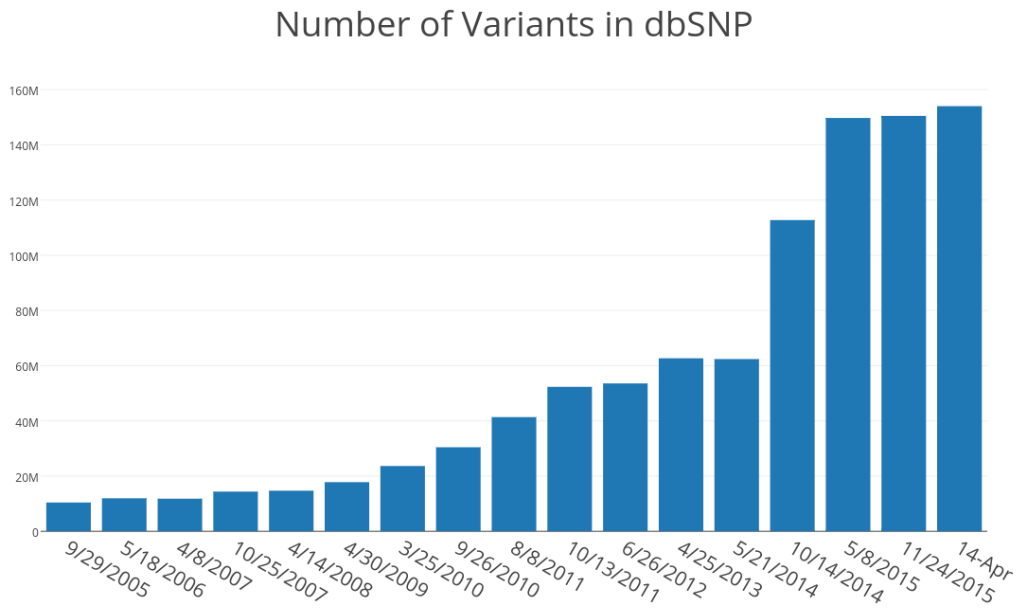
Clinically Assessments:
Per Sample ~10MB

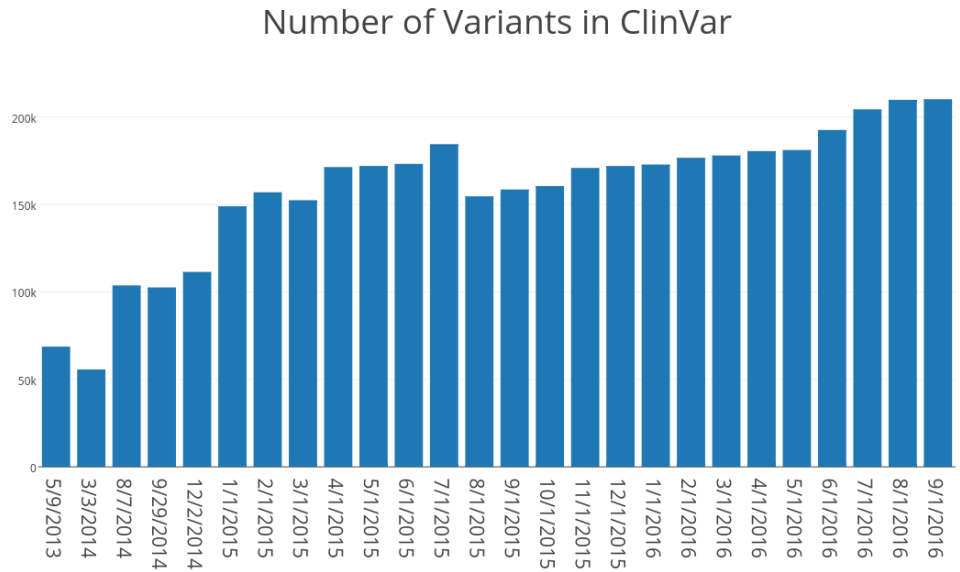
Clinical Reports:
Per Sample ~1MB

Good BM, Ainscough BJ, McMichael JF, Su AI†, Griffith OL†. 2014. Genome Biology. 15(8):438.

NGS Driving Our Catalog of Small Variations

Project	Samples	Vars
1KG Phase 1	1,094	39M
1KG Phase 3	2,504	84M
NHLBI ESP	6,500	2M
BROAD ExAC	61,486	10M





GOLDEN HELIX
Accelerating the Quest for Significance™

Big Data at Golden Helix

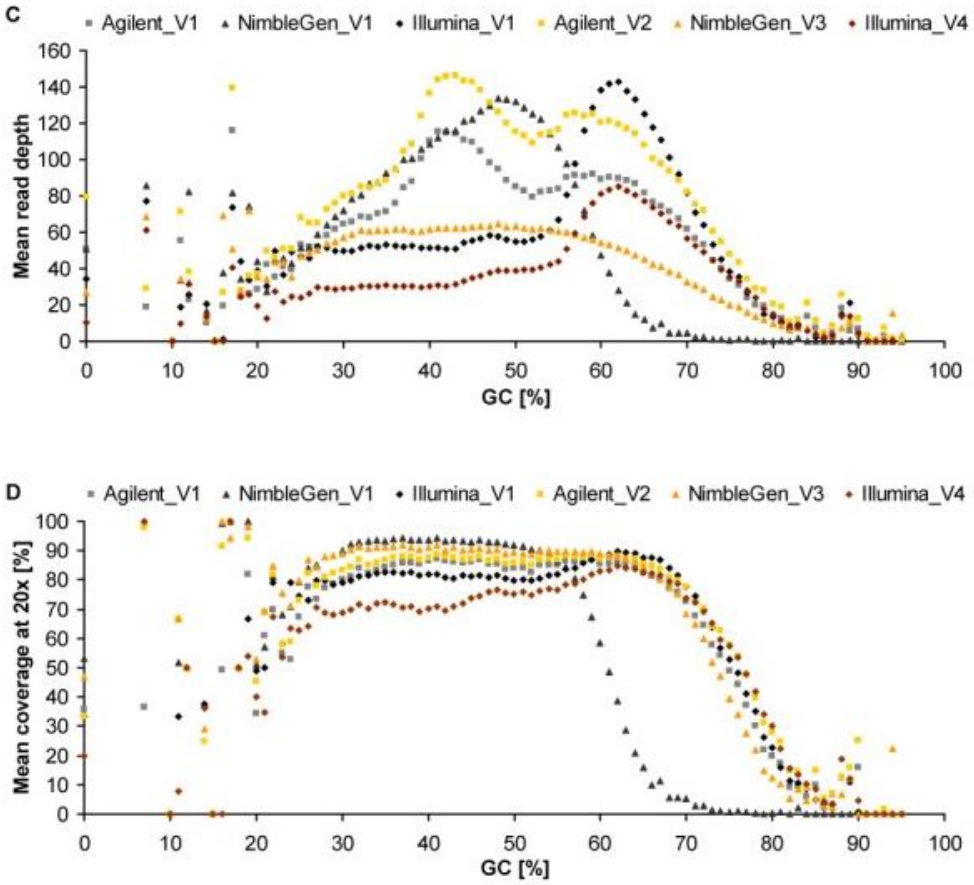
varSEQ™ **Whole Genome Sequencing**

WAREHOUSE **Usage Examples on Big Data**

SNP & VARIATION SUITE **Agrigenomics Prediction**

Whole Genomes Provide a Better Exome

- **Whole Genome PCR Free**
  - Less sensitive to GC content
- **Not limited to target design**
- **Uniform coverage allows for CNV detection**
- **Downsides (other than cost):**
  - Often lower coverage (~50X vs 150X)
  - Doesn't make sense for tumors where you need very high read depth
  - Only loosing ~0.5% of variants[1]
  - Can your tools handle WGS?

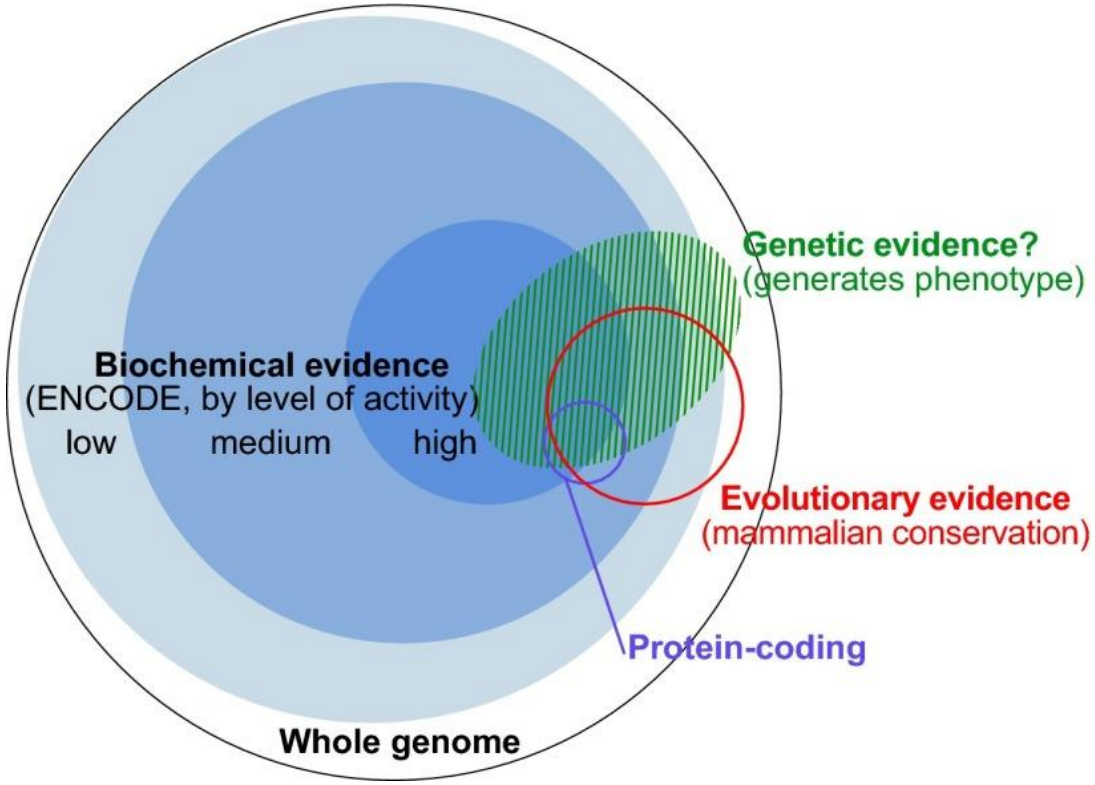


Means of 6 Samples Run on 6 Exon Kits

[1] Meienberg J, Zerjavic K, Keller I, et al. New insights into the performance of human whole-exome capture platforms. Nucleic Acids Res. 2015;43:e76. doi: 10.1093/nar/gkv216

Variants Outside Exome Target Capture Too!

- **Intronic variants can be of clinical significance**
  - ClinVar has ~10K intronic variants (387 P or LP)
- **Many PGX and other clinical trait association variants are intergenic**
- **Need annotation sources outside genes:**
  - CADD
  - Conservation scores
  - Splice site predictions



Kellis M, Wold B, Snyder MP, et al. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*. 2014;111(17):6131-6138. doi:10.1073/pnas.1318948111.

The Amazing 17!

- 17 Supercentenarian
- Sequenced using CGI WGS
- 1 Male, 16 Females
- Found DSC2 Pathogenic Mutation
- Found weak TSHZ3 rare variant burden over controls
- Lets do some similar analysis!
  - Filter to ClinVar Pathogenic variants for LoF variants
  - Count per gene presence of rare, functional Homozygous variants

OPEN ACCESS Freely available online

PLOS ONE

Whole-Genome Sequencing of the World's Oldest People

Hinco J. Gierman[1], Kristen Fortney[1], Jared C. Roach[2], Natalie S. Coles[3,4], Hong Li[2], Gustavo Glusman[2], Glenn J. Markov[1], Justin D. Smith[1], Leroy Hood[2], L. Stephen Coles[3,4], Stuart K. Kim[1]*

1 Depts. of Developmental Biology and Genetics, Stanford University, Stanford, CA, United States of America, 2 Institute for Systems Biology, Seattle, WA, United States of America, 3 Gerontology Research Group, Los Angeles, CA, United States of America, 4 David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States of America

Abstract

Supercentenarians (110 years or older) are the world's oldest people. Seventy four are alive worldwide, with twenty two in the United States. We performed whole-genome sequencing on 17 supercentenarians to explore the genetic basis underlying extreme human longevity. We found no significant evidence of enrichment for a single rare protein-altering variant or for a gene harboring different rare protein altering variants in supercentenarian compared to control genomes. We followed up on the gene most enriched for rare protein-altering variants in our cohort of supercentenarians, TSHZ3, by sequencing it in a second cohort of 99 long-lived individuals but did not find a significant enrichment. The genome of one supercentenarian had a pathogenic mutation in DSC2, known to predispose to arrhythmogenic right ventricular cardiomyopathy, which is recommended to be reported to this individual as an incidental finding according to a recent position statement by the American College of Medical Genetics and Genomics. Even with this pathogenic mutation, the proband lived to over 110 years. The entire list of rare protein-altering variants and DNA sequence of all 17 supercentenarian genomes is available as a resource to assist the discovery of the genetic basis of extreme longevity in future studies.

Citation: Gierman HJ, Fortney K, Roach JC, Coles NS, Li H, et al. (2014) Whole-Genome Sequencing of the World's Oldest People. PLoS ONE 9(11): e112430. doi:10.1371/journal.pone.0112430

Editor: Patrick Lewis, UCL Institute of Neurology, United Kingdom

Received July 22, 2014; **Accepted** September 29, 2014; **Published** November 12, 2014

Copyright: © 2014 Gierman et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. The data are available from supercentenarians.stanford.edu and from Google Genomics dataset 1825457193295669773. Apply for data access at http://goo.gl/MGcyJS.

Funding: This work was supported by the Ellison Medical Foundation/American Federation for Aging Research Fellowship, Stanford Dean's Fellowship, The Paul Glenn Foundation Biology of Aging Seed Grant, National Institute of General Medical Sciences Center for Systems Biology (P50 GM076547) and the University of Luxembourg – Institute for Systems Biology Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: stuartkm@stanford.edu

Introduction

Supercentenarians are the world's oldest people, living beyond 110 years of age [1]. As would be expected for people that reach this age, supercentenarians have escaped many age-related diseases [2–5]. For example, there is a 19% lifetime incidence of cancer in centenarians compared to 49% in the normal population [6]. Similarly, supercentenarians have a lower incidence of cardiovascular disease and stroke than controls [5].

The genetic component of human lifespan based on twin studies has been estimated to be around 20–30 percent in the normal population [7], but higher in long-lived families [8–10]. Furthermore, siblings, parents, and offspring of centenarians also live well beyond average [11,12]. Lifestyle choices in terms of smoking, alcohol consumption, exercise, or diet does not appear to differ between centenarians and controls [13]. Taken together, these findings provide ample evidence that extreme longevity has a genetic component .

Several gene association studies have compared cohorts of long-lived subjects to controls. Analysis of candidate genes has shown that polymorphisms in the Insulin-like Growth Factor 1 Receptor gene (IGF1R) and the FOXO3 transcription factor gene are associated with extreme longevity [14,15]. Genome-wide association studies have shown that the ApoE4 haplotype is depleted in centenarians [16–18]. Sebastiani et al. compiled a list of 281 independent single-nucleotide polymorphisms (SNPs) that showed strong associations with extreme longevity (though none were genome-wide significant except for an ApoE SNP) [17]. They then showed that a genetic signature that combines information from these 281 SNPs is predictive for extreme longevity, indicating that at least some of these SNPs are truly associated with longevity. However, specific variants associated with longevity have not yet been identified [18,19].

More recently, studies have begun to use whole-exome sequencing and whole-genome sequencing (WGS) of centenarians to find variants associated with extreme longevity [19–21]. Ye et al. compared the genome sequence of a pair of 100-year-old twins to a pair of 40-year-old twins and found no evidence of accumulation of somatic mutations during aging [20]. By sequencing blood cells of a supercentenarian, Holstege et al. first identified somatic mutations and then used this information to infer clonal lineages in hematopoietic stem cells. They found that white blood cells in this individual were derived from only two clones of hematopoietic stem cells [21].

Here, we have sequenced the genomes of 17 supercentenarians. We limited the majority of our analyses to the thirteen genomes from Caucasian females. From this small sample size, we were unable to find rare protein-altering variants significantly associated with extreme longevity. However, we did find that one supercentenarian carries a pathogenic variant associated with arrhyth-

http://dx.doi.org/10.1371/journal.pone.0112430

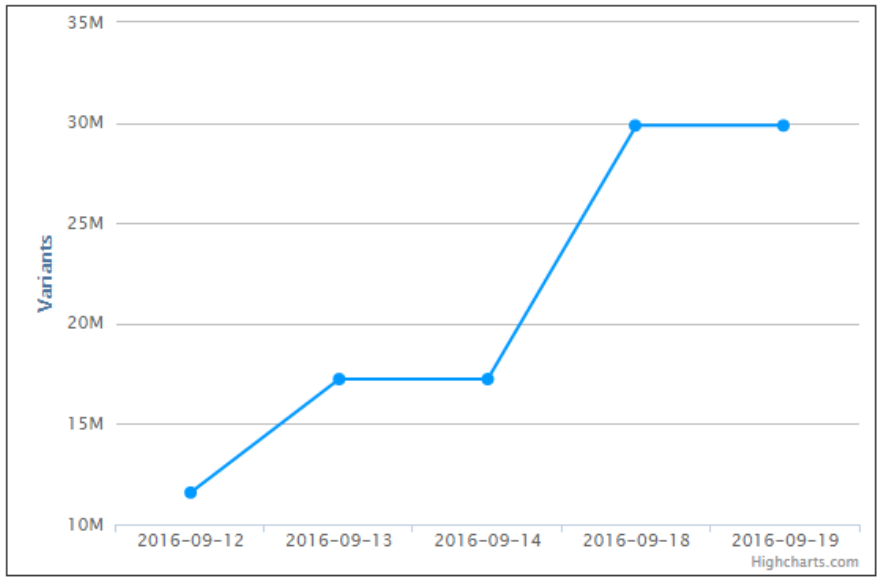
VarSeq Demonstration

VSWarehouse: Treasure your Samples

- **Scalable Infrastructure of VarSeq as a Multi-User Growing Repository of your NGS Samples**
  - Query variants and annotations
  - Exports to Text, Excel, VCF
- **Aggregate Samples:**
  - Targeted Gene Panels
  - Exomes
  - Genomes
- **Cancer and Germline Workflows**
- **Deep integration with VarSeq for annotation, reporting and hands on assessment/classification**

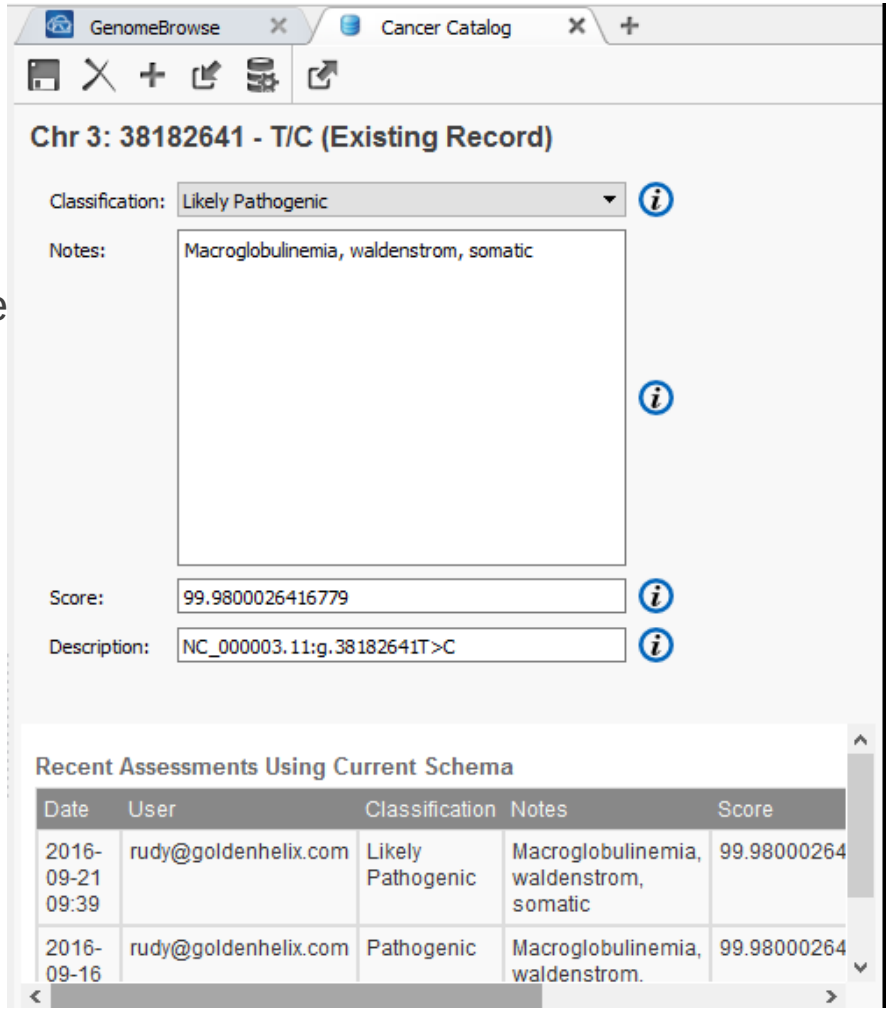
WAREHOUSE

Scalable Variant Warehouse for VarSeq ®



New in VSWarehouse 1.2

- **Assessment catalogs**
  - Store your variant curations, assessments, classifications
  - Can also be used for cataloging common false positives, other custom annotations
  - Updated by users with version history
  - Can “bulk import” existing knowledgebase
- **Optimized imports, especially adding samples to existing warehouse projects**
- **Improved query speeds**



VarSeq + VSWarehouse Demostration

- **Cancer Gene Panel Demo Data**
- **Lets Connect it to a VSWarehouse Installation**
  - Report
  - Catalog
  - Annotations
- **Lets Extract our Samples from VSWarehouse**
  - Query rare, functional variants in our cohort
  - Export to Excel

SNP & Variation Suite

- **Mature Research Analysis Platform**
- **Designed for Large Datasets, both in Samples and Markers/Variants**
- **Agrigenomics is growing market, pushing the sample limit “N”**
- **Certain matrix operations are computed on NxN matrixes**

Kinship Matrices & GBLUP

- The GBLUP method computes a genomic relationship matrix and from that computes the “Genomic Best Linear Unbiased Predictor” (GBLUP) of additive genetic merits by sample and of allele substitution effects (ASE) by marker.
- GBLUP can be used to predict Estimated Breeding Values (EBV) for all samples in a dataset which allows for the identification of samples with the highest EBV to carry forward in breeding programs.
- It can also be used to identify influential loci for the phenotype of interest that can then be used for a targeted assay for diagnostic purposes.

Beating the System

- We use:
  - Out-of-memory scratch buffers
  - Piece wise large data matrix operations
  - Adaptation of method[1] for approximating matrix decompositions of large matrices using random numbers.

Mode	Approach	Max Iterations per Batch	Early Exit Precision
Slow	Large N	30	10^{-7}
Medium	Large N	12	10^{-5}
Quick	Large N	5	10^{-5}
Quickest	Large N	2	10^{-3}
Exact	Small N	N/A	N/A

# Samples	Slow	Medium	Quick	Quickest	Exact (Small N Algorithms)
2k	~1 min	1 min	~1 min	~1 min	~1 min
4k	~10 min	7 min	5 min	3 min	~2 min
8k	76 min	80 min	32 min	19 min	~5 min
20k	1133 min / ~19 hrs	823 min / ~14 hrs	469 min / ~8 hrs	234 min / ~4 hrs	~57 min
40k	Not computed	Not computed	3556 min / ~59 hrs / ~2.5 days	2042 min / ~34 hrs / ~1.5 days	~5796 min / ~4 days

Halko et al. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. arXiv:0909.4061 [math.NA]

SVS Demonstration

GOLDEN HELIX
SNP & VARIATION SUITE

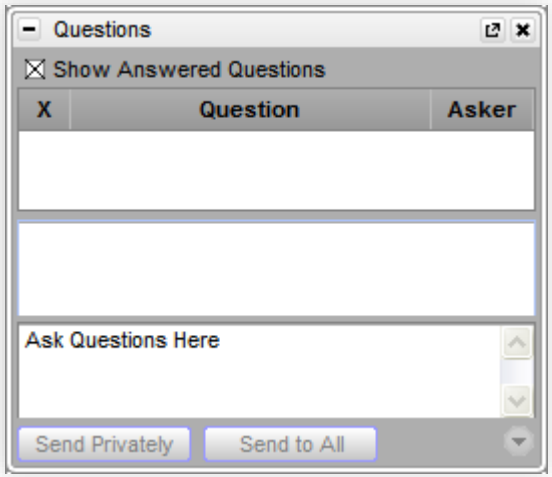
[Demonstration]

Use Cases for VSWarehouse and VarSeq

- **Researchers:** Store and aggregate research samples. Re-run queries against new annotations and research findings over time. Share and collaborate.
- **Small Labs:** Build up in-house knowledge base of variant assessments. Have warehouse of clinical samples to draw on for quality and frequency filtering. Have versioned snapshots of warehouse to reference from archived reports.
- **Core Labs:** Provide institution wide population frequencies. Access controlled projects for individual groups of users. Customized workflows for different use cases.
- **Consortiums:** Secure cloud based repository for users to submit new samples. Versioned history, customized annotations and per-cohort statistics. Allow users to query and extract the subsets necessary for analysis.

Questions during the presentation

Use the Questions pane in your GoToWebinar window



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