GOLDEN HELIX

Enabling Precision Medicine

Advantages of VarSeq's Annotation Capabilities

Darby Kammeraad - Field Application Scientist

CIOReview
20 most promising Biotech Technology Providers

pharma TECH OUTLOOK
Top 10 Analytics Solution Providers

Gartner
Hype Cycle for Life sciences

Golden Helix – Who We Are

Golden Helix is a global bioinformatics company founded in 1998.

GlaxoSmithKline

varSEQ®

Variant Calling
Filtering and Annotation
Clinical Reports
CNV Analysis
Pipeline: Run Workflows

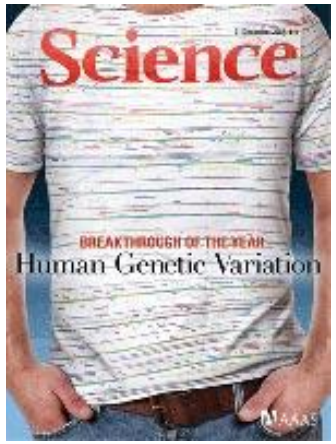
VARIANT WAREHOUSE

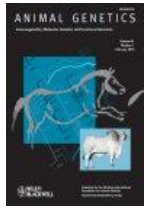
Variant Warehouse
Centralized Annotations
Hosted Reports
Sharing and Integration

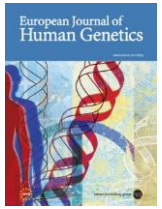
SNP & VARIATION SUITE

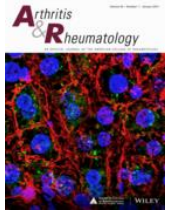
GWAS
Genomic Prediction
Large-N-Population Studies
RNA-Seq
Large-N CNV-Analysis

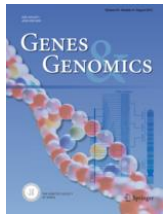
Cited in over 1100 peer-reviewed publications

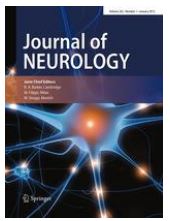












GOLDEN HELIX
Enabling Precision Medicine

Over 350 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
Agricultural Research Service

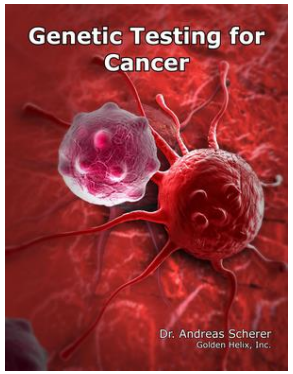
FDA

GOLDEN HELIX
Enabling Precision Medicine

Golden Helix – Who We Are

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

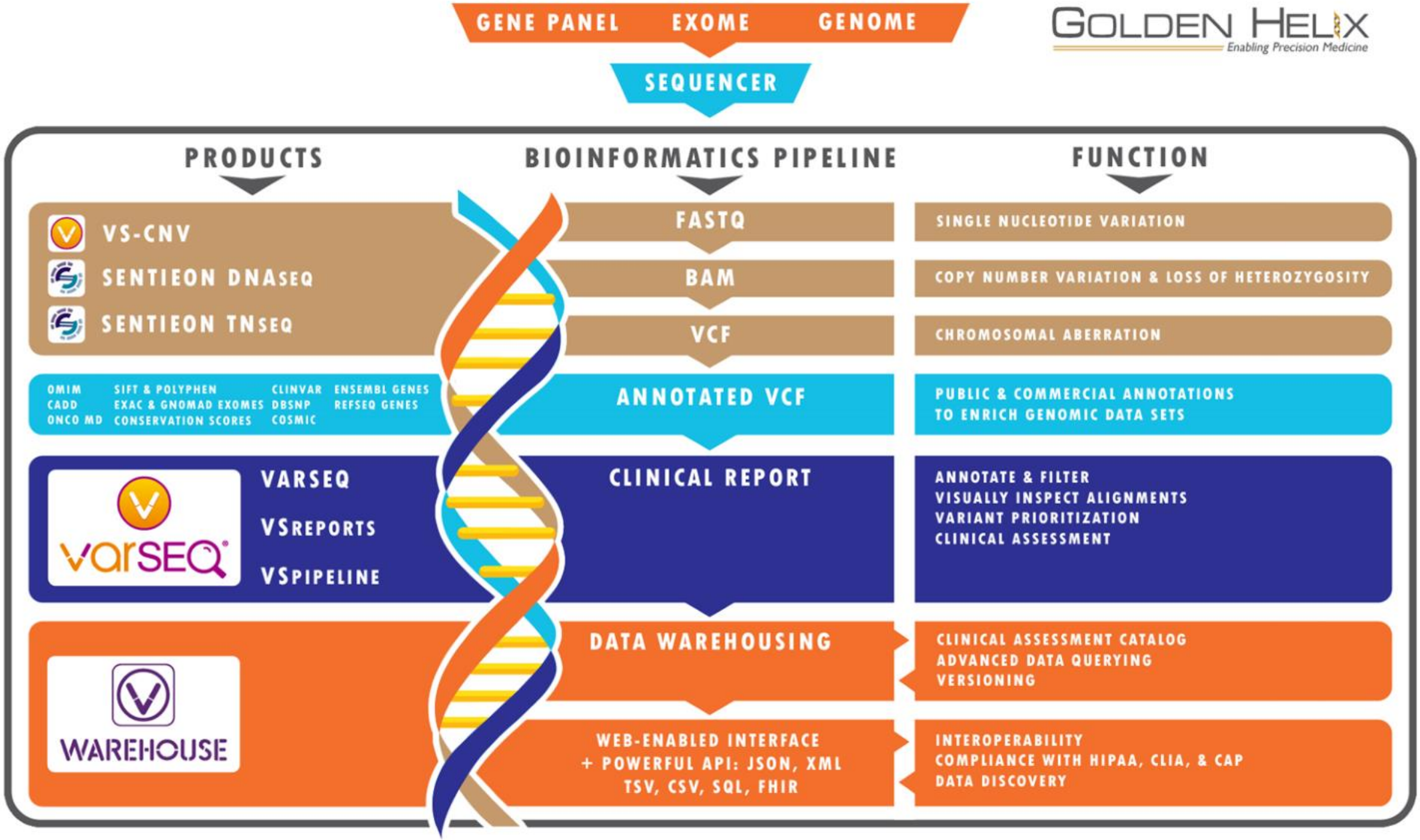
- REPUTATION
- TRUST
- EXPERIENCE



- INDUSTRY FOCUS
- THOUGHT LEADERSHIP
- COMMUNITY

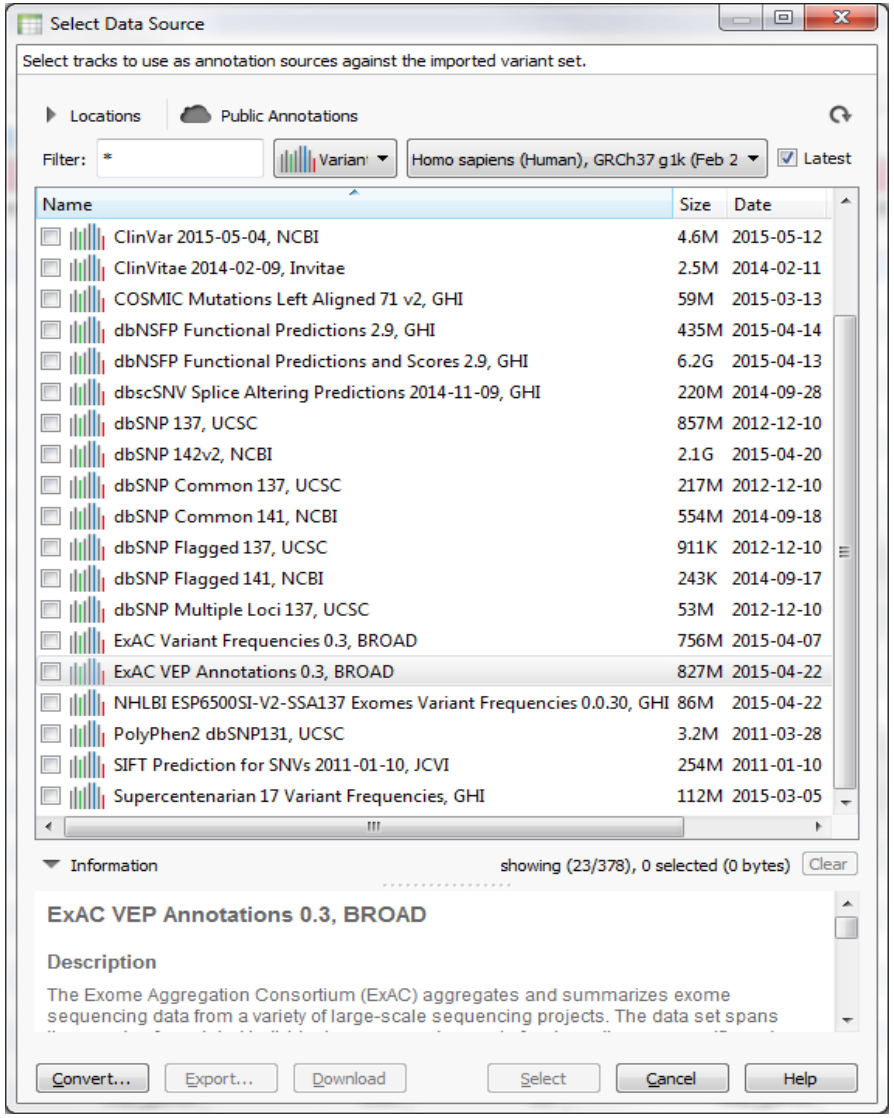
- TRAINING
- SUPPORT
- RESPONSIVENESS

- INNOVATION and SPEED
- CUSTOMIZATIONS


GENE PANEL
EXOME
GENOME
SEQUENCER
GOLDEN HELIX
Enabling Precision Medicine
PRODUCTS
BIOINFORMATICS PIPELINE
FUNCTION
VS-CNV
SENTIEON DNASEQ
SENTIEON TNSEQ
FASTQ
BAM
VCF
SINGLE NUCLEOTIDE VARIATION
COPY NUMBER VARIATION & LOSS OF HETEROZYGOSITY
CHROMOSOMAL ABERRATION
OMIM
CADD
ONCO MD
SIFT & POLYPHEN
EXAC & GNOMAD EXOMES
CONSERVATION SCORES
CLINVAR
DBSNP
COSMIC
ENSEMBL GENES
REFSEQ GENES
ANNOTATED VCF
PUBLIC & COMMERCIAL ANNOTATIONS
TO ENRICH GENOMIC DATA SETS
varSEQ
VARSEQ
VSREPORTS
VSPIPELINE
CLINICAL REPORT
ANNOTATE & FILTER
VISUALLY INSPECT ALIGNMENTS
VARIANT PRIORITIZATION
CLINICAL ASSESSMENT
WAREHOUSE
DATA WAREHOUSING
CLINICAL ASSESSMENT CATALOG
ADVANCED DATA QUERYING
VERSIONING
WEB-ENABLED INTERFACE
+ POWERFUL API: JSON, XML
TSV, CSV, SQL, FHIR
INTEROPERABILITY
COMPLIANCE WITH HIPAA, CLIA, & CAP
DATA DISCOVERY

Annotation Options in VarSeq

Types	Description	Popular Examples in VarSeq
Gene Tracks	Gene and effect on transcript(s)	RefSeq, Ensemble, dbNSFP
Assemblies	Refence sequence and alignment review	GRCh 37 hg19, GRCh 38/37 g1k Low complexity regions
Microarray Probe Maps	Matching variant with microarray probe location	Affymetrix Cytogenetic/500K/SNP
Variant/Function	Allele frequencies and functional predictions	gnomAD, ExAC, ICGC, CADD, OMIM, dbSNP, dbNSFP, OncoMD, ClinVar, COSMIC (cancer)
Targeted Panels	Disease specific regions	TruSight (Cancer/Cardio/Autism), Ion AmpliSeq Disease Panel



Data Curation – A Peek Under The Hood

- Frequently update annotations – monthly for most (ClinVar, OncoMD, & others)
- From many disparate sources, researching the best representation of the raw data sources
- Variant normalization and transformation ensures the precision and sensitivity in matching genomic data source
- We work with creators of annotation sources providing feedback
- Substantial savings for clients – multiple Full Time Equivalents

Clinical assessment - ClinVar

- **ClinVar –** features 414,708 variants

  - This public archive from NCBI
  - Collaboration of many clinical labs (both commercial and academic)
  - Reports the relationship among human variations and phenotypes (supporting evidence from dbSNP)
  - Variants found in patient samples, their clinical significance, submitter information, and other supporting data
  - Alleles mapped to reference sequences and use HGVS standards
  - Submissions can be review by an expert panel.

ClinVar 2017-09-05, NCBI							
Ref/Alt	Accession	Gene Names	HGVS g. Name	Clinical Significance	MedGen	Disease Name	ClinVar Review Status
T/-	RCV000007567.3	CFTR	NC_000007.13:g.117171108delT	Pathogenic	C0010674	Cystic fibrosis	(0 Stars) Not classified by submitter
G/A	RCV000114996.3	ZPR1,APOA5	NC_000011.9:g.116660686G>A	Risk Factor	C2676231	Hypertriglyceridemia, susc...	(0 Stars) Not classified by submitter
A/-	RCV000029281.1	ABCC9	NC_000012.11:g.21958999delA	Uncertain Significance	C0878544	Cardiomyopathy	(1 Star) Classified by single submit...
G/A	RCV000337303.1	TBX3	NC_000012.11:g.115117337G>A	Likely Benign	C1866994	Ulnar-mammary syndrome	(1 Star) Classified by single submit...

Clinical assessment - OMIM

- **OMIM (updated Monthly)**
  - Contains information from all known Mendelian disorders
  - Variants (features 20,527 variants) These are specific variant assertions with clinical annotations and references
  - Genes (features 14,825 variants) Includes linked phenotypes and their inheritance pattern, with full HTML descriptions
  - Phenotypes (features 4,370 variants) Linked genes, alternative phenotype names, descriptions, and references

OMIM Variants 2017-04-01, GHI

Ref/Alt	Phenotype	Gene Name	GeneOMIMID	Entrez Gene ID	PubMed ID	HasPubMedID	Name	dbSNP	Description	References
A/C	INSULIN ...	HNF1A	142410	6927	12788852,...	True	HNF1A, IL...	rs1169288	<p><a hr...	1. Babaya ...
G/C	CODON 7...	TP53	191170	7157	11403041,...	True	TP53, PR...	rs1042522	<p><a hr...	1. Aaltone...

OMIM Genes 2017-04-01, GHI

Gene Name	OMIM ID	PubMed ID	Title	Description	Gene Status	Disorders
HNF1A	142410	12788852,1707...	HNF1 HOMEO...	?	Confirmed	Diabetes mellit...
TP53	191170	11403041,8673...	TUMOR PROTE...	<p>The transc...	Confirmed	Adrenal cortica...

OMIM Phenotypes 2017-04-01, GHI

Gene Names	Cytogenetic Locations	Entrez Gene IDs	GeneOMIMIDs	Inheritance	OMIM ID	PubMed ID	HasPubMedID	Title	Alternative Title(s)	Description	References
GPD2,NE...	2q24.1,2q32,2q36,3p...	2820,4760,366...	138430,601...	Autosomal dominant,Multifactorial,A...	125853,14...	17726085,...	True,True,Tr...	DIABETES...	DIABETES MELLITU...	<p>Mole...	16. Elbein ...
RAD54L,C...	1p32,2q33,2q34-q35,...	8438,841,580,...	603615,601...	Autosomal dominant,Somatic mutati...	114480,11...	19330027,...	True,True,Tr...	BREAST C...	BREAST CANCER F...	<p>Breas...	9. Anzick ...

Annotations for Cancer

- **CIViC (updated monthly)** – features 634 variants
  - Variant Clinical Evidence Summaries & Region Clinical Evidence Summaries (exon and gene deletions/gains).
  - CIViC accepts public knowledge contributions but requires that experts review these submissions.
  - Evidence statements & records (response to therapy, prognostic, diagnostic, or predisposing for cancer.

CIViC - Region Clinical Evidence Summaries 2017-08-01, WUSTL

Gene Name	Representative Transcript	Variant Type	Disease	Disease Ontology ID	Drugs	Clinical Significance	Evidence Direction	Evidence Level	Trust Rating
APC,APC	ENST00000457016.1,ENST0000...	MUTATION,...	Colon Carcin...	1520,9256	JW55,G007-LK	Sensitivity,Sensitivity	Supports,Supports	D - Preclinical,D - P...	3 out of 5 Stars,4 out ...
PTCH1,PTCH1	ENST00000331920.6,ENST0000...	MUTATION,L...	Brain Medull...	0060105,0060105	Vismodegib,...	Sensitivity,Sensitivity	Supports,Supports	B - Clinical,B - Clini...	4 out of 5 Stars,2 out ...
TP53,TP53,T...	ENST00000269305.4,ENST0000...	DELETERIOU...	Head And N...	5520,5520,3748,7061,0...	Chemothera...	Poor Outcome,Poor ...	Supports,Does Not S...	B - Clinical,B - Clini...	3 out of 5 Stars,3 out ...

- **COSMIC Mutations Left Aligned 71** – features 2,151,007 variants
  - Catalogs somatic variants discovered in cancer samples.
  - Provides details about the frequency, tumor types and histology
  - Provides gene level annotations with relevant summary and curated oncology details
  - COSMIC breaks out each sample-variant pair into a record
    - VarSeq provides the fields in COSMIC with relevant hyperlinks.

COSMIC Mutations Left Aligned 71 v2, GHI

Ref/Alt	Mutation ID	Mutation CDS	Mutation AA	Gene Name	Transcript ID	Gene CDS Length	HGNC ID	Primary Site	Mutation Description	Mutation Zygosity
A/C	430522	c.79A>C	p.I27L	HNF1A	ENST0000025...	1896	11621	Prostate (2),...	Substitution - Missense	Heterozygous (1)
G/C,G/C,...	250061,376...	c.215C>G,c....	p.P72R,p.P...	TP53,TP5...	ENST0000026...	1182,1182,1041,1...	11998,?,?,?,?	Upper aerod...	Substitution - Missense,Substitution - Misse...	Homozygous Varia...

Annotations for Cancer

- **ICGC Simple Somatic Mutations 22** – features 47,879,813 variants
  - Collection of data from across 89 committed projects currently
  - Goals related to quality
    - Ensure that most cancer genes with frequency of >3% are discovered
    - High sequence level resolution
    - High quality standards
    - Control based data (tumor/normal pairs)
  - Somatic mutations in 21 primary cancer sites in 21k donors
  - Primary Site and affected donor frequency.

ICGC Simple Somatic Mutations 22, GHI								
Ref/Alt	Identifier	AffectedDonorsForAllProjects	Project Count	Project ID	Primary Site	Affected Donors	Total Samples	Affected Donor Frequency
G/C	MU151094	1	1	COAD-US	Colorectal	1	216	0.00463
A/G	MU156543	1	1	COAD-US	Colorectal	1	216	0.00463
T/C	MU3888690	1	1	THCA-SA	HeadAndNeck	1	129	0.00775
A/G	MU112255	1	1	COAD-US	Colorectal	1	216	0.00463

Annotations for Cancer

- **OncoMD (updated Monthly)**
  - Variant and Gene Summaries
    - Cancer related genes (onco and tumor suppressor genes)
    - Effect on protein
    - Publications/studies associated with the variant
    - Drug Targeting Mutations
    - List of open clinical trials

OncoMD Clinical Trials							OncoMD Studies with Variant					
Gene Symbol	Cancer Type	Country	Drugs	Inclusion Criterion	Status	Trial Number	Ref/Alt	Gene Symbol	PubMed ID	Study Type	Title	SampleCount
ALK,ALK	Brain and ...	Canada,U...	crizotinib,crizoti...	ALK MUTATION,A...	Recruitin...	NCT00939...	G/C	ALK	?	No Study ...	?	1
ALK,ALK	Brain and ...	Canada,U...	crizotinib,crizoti...	ALK MUTATION,A...	Recruitin...	NCT00939...	?	?	?	?	?	?

Frequency Tracks – From ExAC to gnomAD

- **ExAC** – features 10,324,246 variants
- **gnomAD** – features 17,439,605 variants

  - Major changes from ExAC –
    - Genome (15,496) and exome (123,136)
    - Gnomad is a new product (data processing perspective)
    - Cohort wider selection of ethnicities (Ashkenazi Jewish)
    - New/novel ways of flagging low quality variants

ExAC Variant Frequencies 0.3, BROAD					gnomAD Exomes Variant Frequencies 2.0.1 v2, BROAD				
Ref/Alt	Identifier	Filter	Alt Allele Freq (AF)	Alt Allele Counts (AC)	Ref/Alt	Filter	Alt Allele Prob (RF)	Alt Allele Freq (AF)	Ashkenazi Jewish Allele Count (AC_ASJ)
G/A	rs72975710	PASS	0.0001978	24	G/A	PASS	0.95295	0.000199914	0
C/T	rs72996036	PASS	3.295e-05	4	C/T	PASS	0.953006	2.4375e-05	0
A/G	rs421016	VQSRTrancheSNP99.60to99.80	0.003155	383	A/G	PASS	0.11371	0.00130657	26
G/A	rs73035708	PASS	0.0001977	24	G/A	PASS	0.945595	0.000138133	0
G/T	rs72914988	PASS	0.001466	178	G/T	PASS	0.954732	0.00167319	21
C/T	rs73477443	PASS	8.242e-06	1	C/T	PASS	0.892885	2.47519e-05	2
?	?	?	?	?	C/A	RF	0.0074095	4.10826e-06	0
?	?	?	?	?	A/G	PASS	0.95369	4.06121e-06	0

Frequency Tracks cont… – NHLBI and 1kgenome

- **NHLBI** - Features 2,029,948 variants
  - Current release is taken from 6503 samples
  - Focus on heart, lung, and blood disorders

NHLBI ESP6500SI-V2-SSA137 Exomes Variant Frequencies 0.0.30, GHI							
Ref/Alt	Identifier	All AAF	European American AAF	African American AAF	All MAF	All HomoVar GTC	All Het GTC
G/A	rs72975710	0.000461326	0	0.00136178	0.000461326	0	6
C/T	rs72996036	7.68876e-05	0	0.000226963	7.68876e-05	0	1
A/G	rs421016	0.00030755	0.000465116	0	0.00030755	0	4
G/A	rs73035708	0.000615101	0	0.00181571	0.000615101	0	8
G/T	rs72914988	0.00284484	0.000930233	0.00658193	0.00284484	0	37

- **1kGenome** - Features 85,823,495 variants
  - Project ran from 2008 to 2015. One of the largest catalogs
  - Goal: ID variants with at least 1% frequencies

1kG Phase3 - Variant Frequencies 5b, GHI						
Ref/Alt	Identifier	All Indiv Freq	European Allele Freq (EUR_AF)	African/African American Allele Freq (AFR_AF)	American Allele Freq (AMR_AF)	South Asian Allele Freq (SAS_AF)
G/A	rs72975710	0.000798722	0	0.003	0	0
C/T	rs72996036	0.000399361	0	0.0008	0.0014	0
A/G	rs421016	0.00339457	0.0119	0.0015	0	0.002
G/A	rs73035708	0.000998403	0	0.0038	0	0
G/T	rs72914988	0.00319489	0.002	0.0061	0.0086	0
C/T	rs73477443	0.000199681	0	0.0008	0	0
C/A	rs73297817	0.000199681	0	0.0008	0	0

Functional Prediction Annotations

- **dbNSFP Functional Predictions and Scores 3.0** – features 82,832,027 variants
  - 14 classifier/prediction algorithms: SIFT, Polyphen2, LRT, MutationTaster, MutationAssessor, FATHMM, MetaSVM, MetaLR, VEST, PROVEAN, FATHMM-MKL coding and fitCons
  - 8 conservation scores (phyloP46way_primate, phyloP46way_placental, phyloP100way_vertebrate, phastCons46way_primate, phastCons46way_placental, phastCons100way_veterbrate, GERP++ and SiPhy)

dbNSFP Functional Prediction Voting							
N of 6 Predicted Tolerated	N of 6 Predicted Damaging	SIFT Pred (C)	Polyphen2 HVAR Pred (C)	MutationTaster Pred (C)	MutationAssessor Pred (C)	FATHMM Pred (C)	FATHMM MKL Coding Pred (C)
0 of 6 Predicted as Tolerated	6 of 6 Predicted as Damaging	Damaging	Possibly damaging	Damaging	Predicted functional (medium)	Damaging	Damaging
1 of 6 Predicted as Tolerated	5 of 6 Predicted as Damaging	Damaging	Probably damaging	Damaging	Predicted functional (medium)	Tolerated	Damaging
0 of 6 Predicted as Tolerated	6 of 6 Predicted as Damaging	Damaging	Possibly damaging	Damaging	Predicted functional (medium)	Damaging	Damaging
0 of 6 Predicted as Tolerated	6 of 6 Predicted as Damaging	Damaging	Probably damaging	Damaging	Predicted functional (medium)	Damaging	Damaging
2 of 6 Predicted as Tolerated	4 of 6 Predicted as Damaging	Damaging	Possibly damaging	Damaging	Predicted non-functional (neutral)	Tolerated	Damaging
2 of 6 Predicted as Tolerated	4 of 6 Predicted as Damaging	Damaging	Probably damaging	Damaging	Predicted non-functional (low)	Tolerated	Damaging
2 of 6 Predicted as Tolerated	4 of 6 Predicted as Damaging	Damaging	Possibly damaging	Damaging	Predicted non-functional (low)	Tolerated	Damaging
1 of 6 Predicted as Tolerated	5 of 6 Predicted as Damaging	Damaging	Probably damaging	Damaging	Predicted non-functional (low)	Damaging	Damaging

- **dbscSNV Splice Altering Predictions 1.1** – features 15,030,435 variants
  - Predicts all snps −3 to +8 at the 5’ splice site and −12 to +2 at the 3’ splice site
  - Two ensemble predictions scores, I can provide cut-offs for 95% specificity in calling splice altering mutations

dbscSNV Splice Altering Predictions 1.1, GHI								
Ref/Alt	RefSeq?	Ensembl?	RefSeq Region	RefSeqG...	Ensembl Region	Ensembl Gene	Ada Score	RF Score
A/T	True	True	splicing	CEP104(...	splicing	ENSG00000116198(...	0.999946	0.962
C/T	True	True	splicing	CEP104(...	splicing	ENSG00000116198(...	0.999934	0.958
C/A	True	True	splicing	AK2(N...	splicing	ENSG00000004455(...	0.995877	0.864
C/T	True	True	splicing	CLSPN(...	splicing	ENSG00000092853(...	0.99999	0.938

Functional Prediction Annotations cont…

- **GWAS Catalog 2015-12-29** – features 22,373 variants
  - Identifies location of SNPs
  - Lists associated publication where the SNP (assay <100,000 SNPS)

- **CADD – Interpreting Variants of Clinical Significance**
  - Provides C-scores of “deleteriousness” for SNVs and indels in the human genome.
  - Also scores coding/non-coding regions
  - Score based on multiple annotation types:
    - Conservation, population frequency, regulatory, functional/structural

CADD Scores 1.3			
Ref/Alt	Raw Score	PHRED Score	Estimated?
G/C	-1.39267	0.003	False
A/G	-0.044528	2.178	False
T/C	-0.486219	0.238	False
A/G	-1.0317	0.014	False
G/T	1.24973	12.01	False
T/A	1.48553	13.24	False
T/C	-0.586188	0.135	False

Transcript Annotations

- **RefSeq** – features 84,950 variants
  - Includes genomic DNA, transcripts, and proteins
  - Effect of transcripts
  - HGVS notation
  - Sequence ontology of variant in all transcripts in database

RefSeq Genes 105v2, NCBI					
Gene Names	SequenceOntologyCombined	Effect (Combined)	TranscriptNameClini...	HGVS c. (Clinically Relevant)	HGVS p. (Clinically Relevant)
ALK	missense_variant	Missense	NM_004304.4	NM_004304.4:c.4587C>G	NP_004295.2:p.Asp1529Glu
ALK	missense_variant	Missense	NM_004304.4	NM_004304.4:c.1427T>C	NP_004295.2:p.Val476Ala
EPCAM	missense_variant	Missense	NM_002354.2	NM_002354.2:c.344T>C	NP_002345.2:p.Met115Thr
MLPH	missense_variant	Missense	NM_024101.6	NM_024101.6:c.1040A>G	NP_077006.1:p.His347Arg

- **Ensembl** – features 215,170 variants
  - Joint effort from EBI and WTSI
  - Annotate, analyze, and display

Ensembl Genes 75v2, Ensembl					
Gene Names	Sequence Ontology (Combined)	Effect (Combined)	Transcript Name (Clinically Relevant)	HGVS c. (Clinically Relevant)	HGVS p. (Clinically Relevant)
CFTR	disruptive_inframe_deletion	Missense	ENST00000003084	ENST00000003084:c.1520_1...	p.Phe508del
HFE	missense_variant	Missense	ENST00000357618	ENST00000357618:c.187C>G	p.His63Asp
TRIM63	missense_variant	Missense	ENST00000374272	ENST00000374272:c.709A>G	p.Lys237Glu

VarSeq Demonstration