

Variant Analysis and Interpretation

Deanna M. Church
Senior Director of Genomics and Content
Personalis, Inc



 @deannachurch

Short Course in Medical Genetics 2013

Analyzing and interpreting variants

Analytical Validity

Our ability to reliably identify variants

Clinical Validity

Our ability to reliably associate variants
with a disease.

Analyzing and interpreting variants



Analyzing and interpreting variants

Analytical Validity

Standard Exomes and Genomes are not “Finished”

Performance comparison of exome DNA sequencing technologies

Michael J Clark, Rui Chen, Hugo Y K Lam, Konrad J Karczewski, Rong Chen, Ghia Euskirchen, Atul J Butte & Michael Snyder

Performance comparison of whole-genome sequencing platforms

Hugo Y K Lam, Michael J Clark, Rui Chen, Rong Chen, Georges Natsoulis, Maeve O'Huallachain, Frederick E Dewey, Lukas Habegger, Euan A Ashley, Mark B Gerstein, Atul J Butte, Hanlee P Ji & Michael Snyder

Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing

Jason O'Rawe^{1,2}, Tao Jiang³, Guangqing Sun³, Yiyang Wu^{1,2}, Wei Wang⁴, Jingchu Hu³, Paul Bodily⁵, Lifeng Tian⁶, Hakon Hakonarson⁶, W Evan Johnson⁷, Zhi Wei⁴, Kai Wang^{8,9*} and Gholson J Lyon^{1,2,9*}

Variant Concordance - "illumina-100bp-pe-exome-30x"

Novoalign+Gatk_UG Bowtie+Gatk_UG Bwa+Gatk_UG



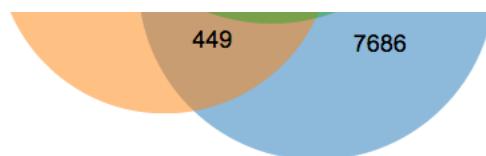
Daniel MacArthur

@dgmcarthur

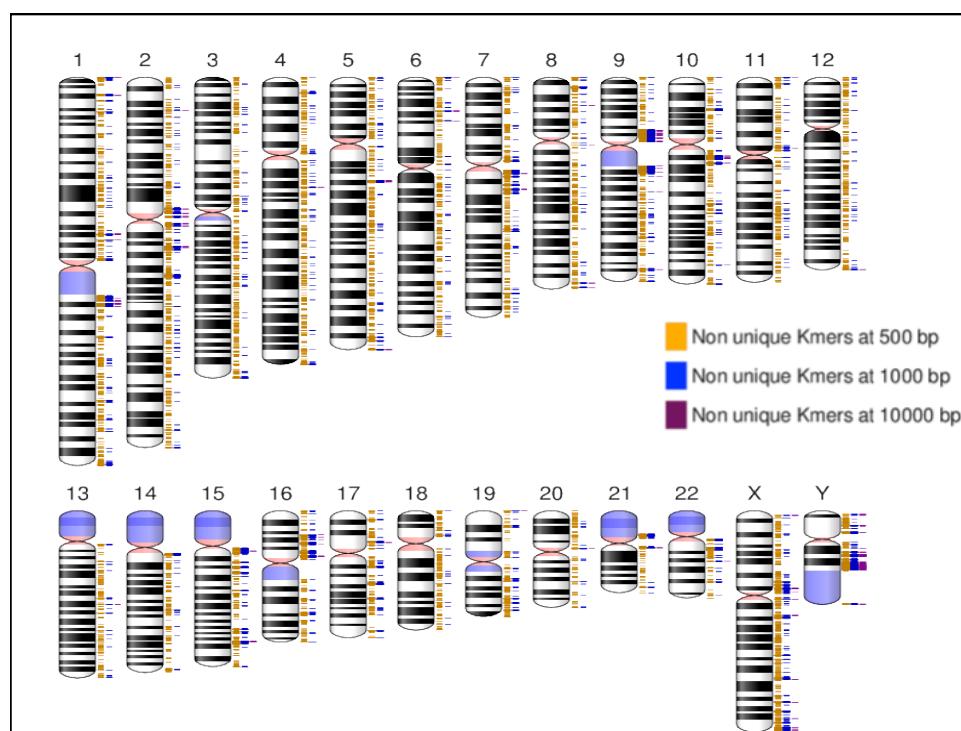
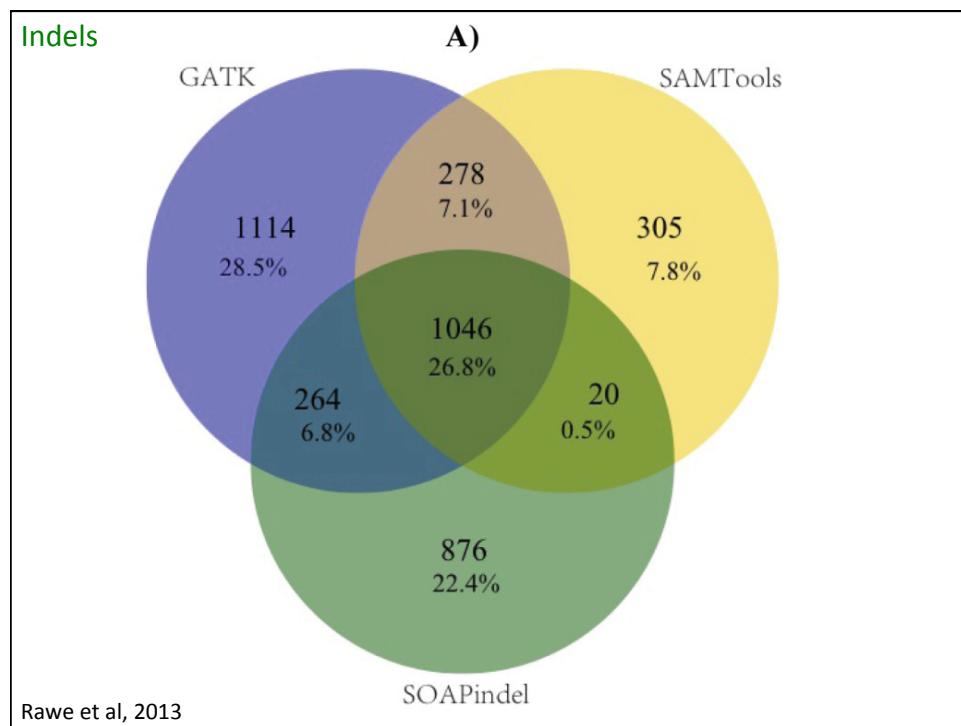


Following

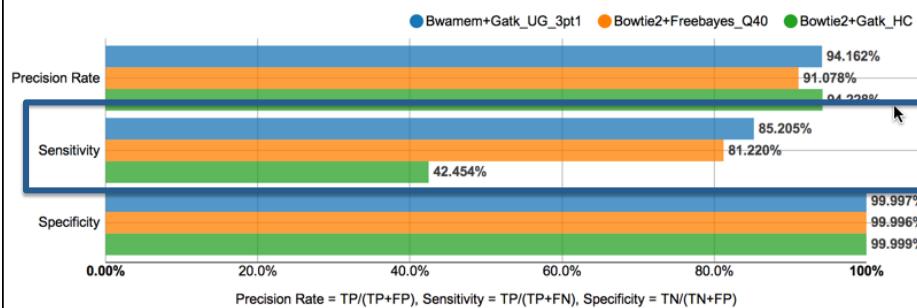
Rong Chen (Personalis): even when the same genome is sequenced twice with 50x Illumina, 6% of variants differ (!!). #apga13



<http://www.bioplanet.com/gcat>

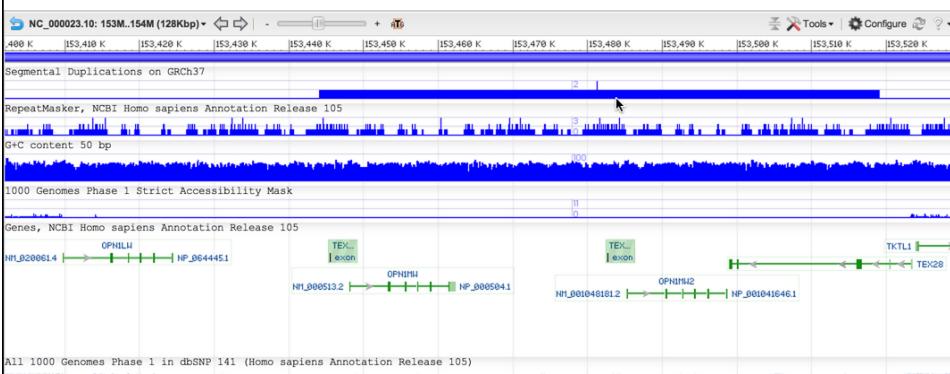


Most research pipelines tuned to limit false discovery

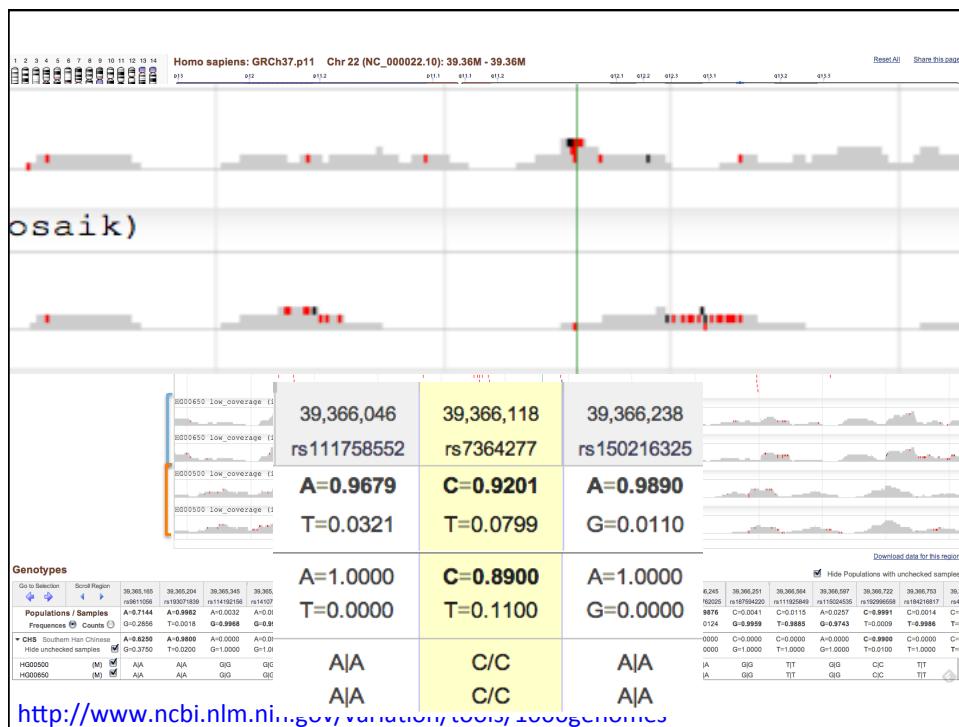
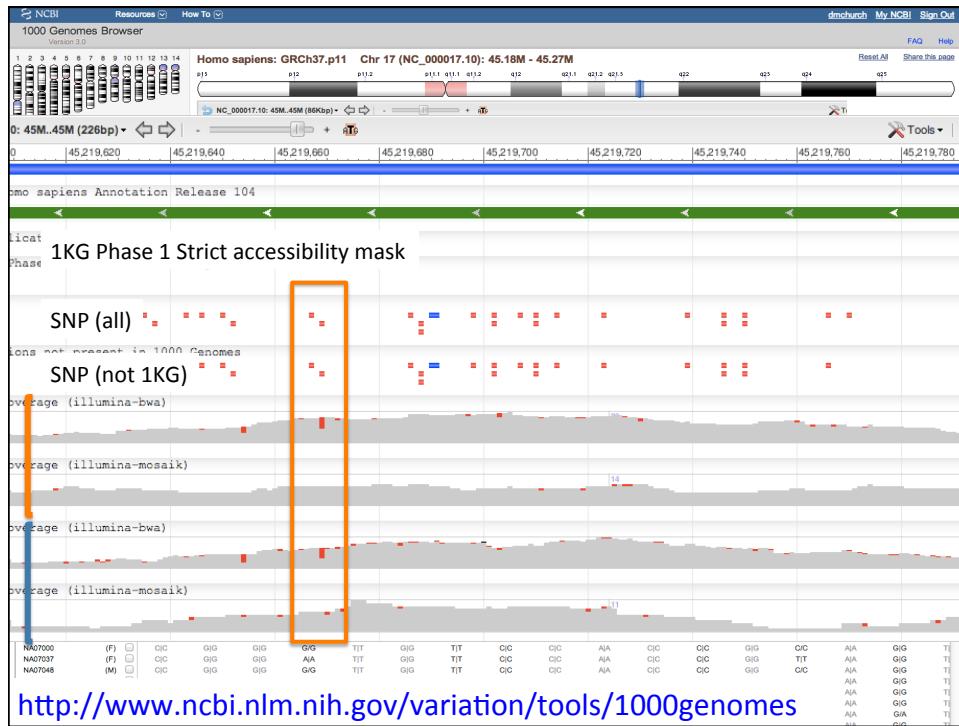


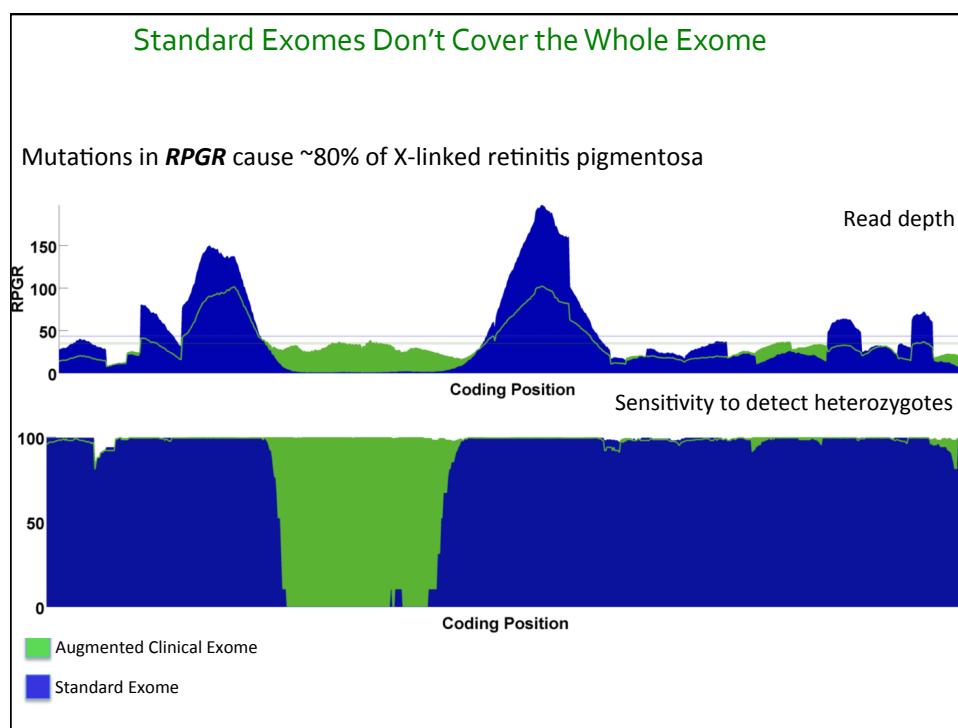
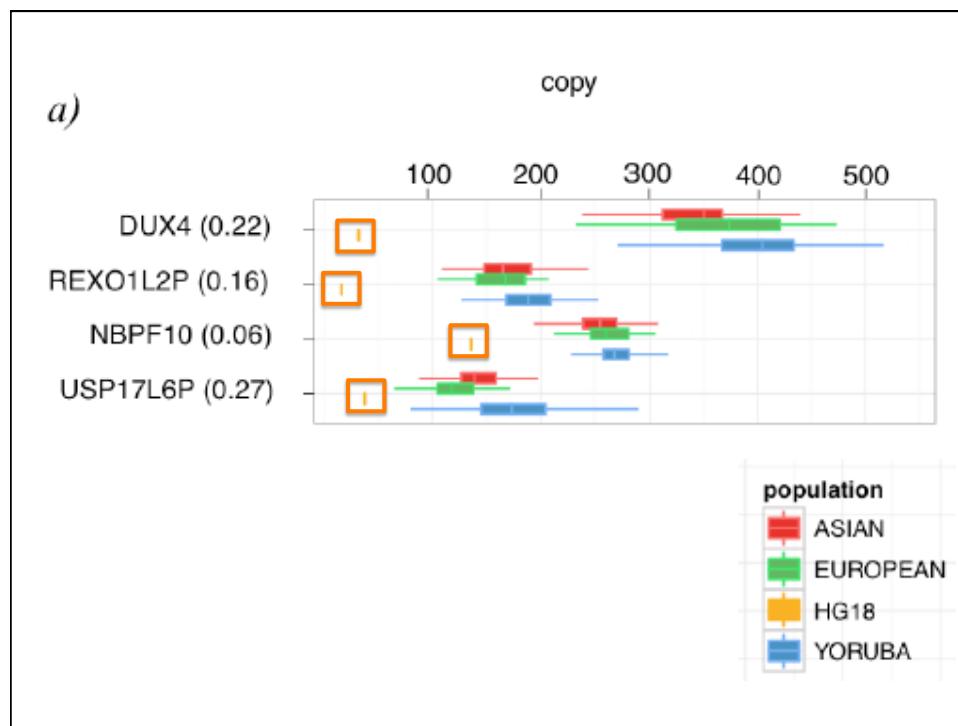
<http://www.bioplanet.com/gcat>

GRCh37.p13 (GCF_000001405.25)Chr X (NC_000023.10):153.4M - 153.5M



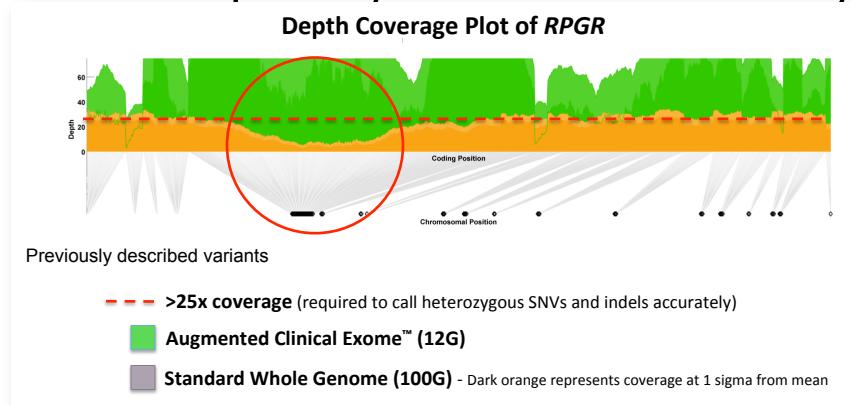
[http://www.ncbi.nlm.nih.gov/variation/view/?
cfg=NCID_1_5296314_130.14.18.128_9146_1406167515_2716676379](http://www.ncbi.nlm.nih.gov/variation/view/?cfg=NCID_1_5296314_130.14.18.128_9146_1406167515_2716676379)



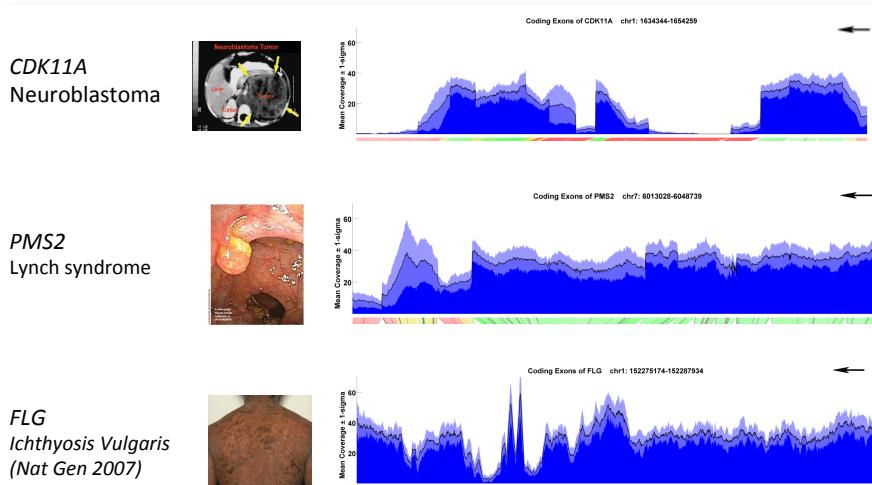




Whole Genome Sequencing Does Not Completely Solve the Accuracy



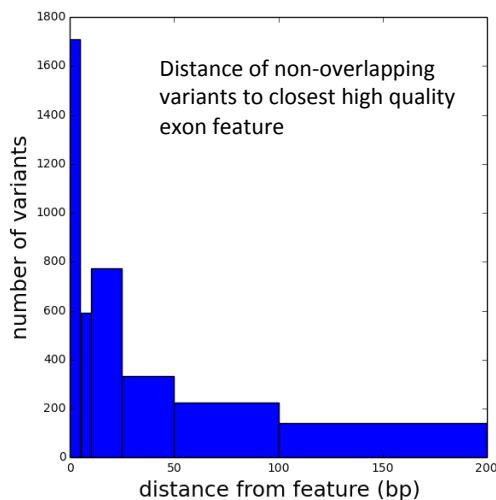
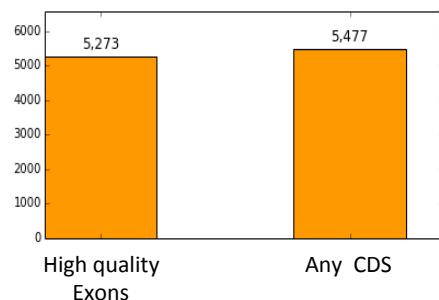
Standard Genomes Don't Cover the Whole Genome



Most of what we know is in the context of genes

ClinVar: 55,467 variants*

Variants with no overlap



* With a genomic location as of 03/07/2014

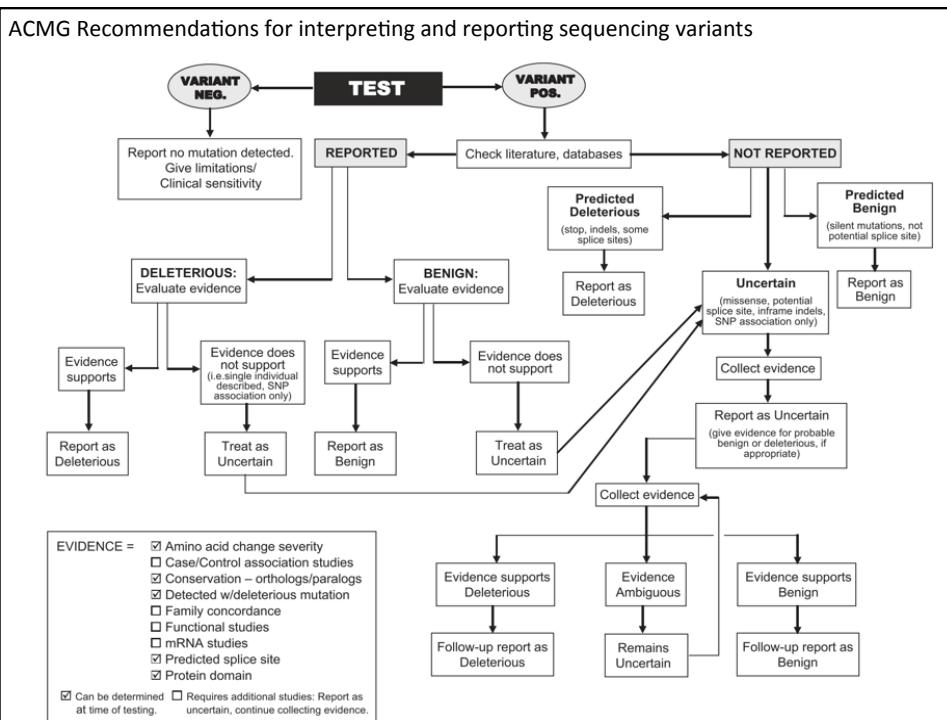
Standard Exomes and Genomes are not “Finished”



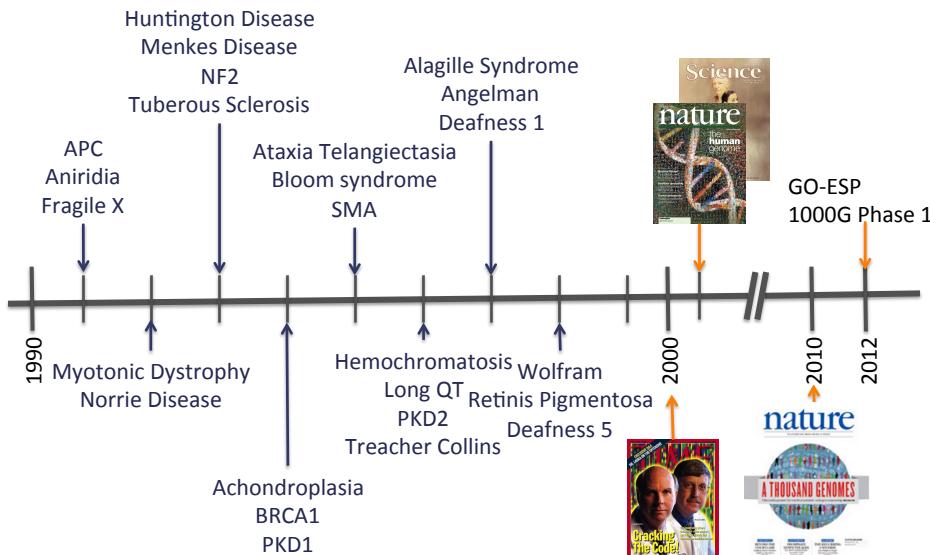
- WGS cost > WES cost.
- Still need to supplement WGS to get full gene coverage.
- Most interpretation is in the context of genes.

Analyzing and interpreting variants

Clinical Validity



Integrating the literature



Adapted from <http://www.ndsu.edu/pubweb/~mcclean/plsc431/homework/positional-cloning/>

Gene Panels

GTR: GENETIC TESTING REGISTRY

All GTR Tests Conditions/Phenotypes Genes Labs GeneReviews Advanced search for tests

Find tests by searching test names, disease names, phenotypes, gene symbols and names, protein names, laboratory names, directors and locations.

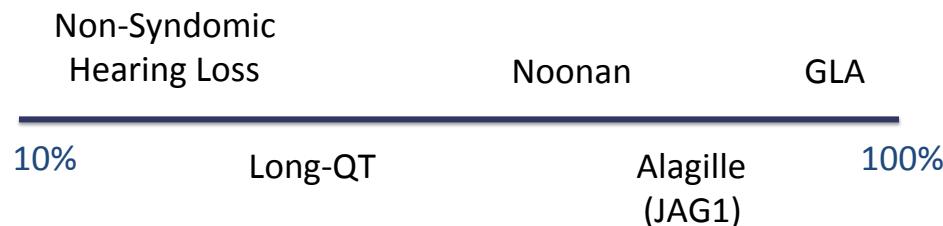
YouTube GTR Tutorials

IMPORTANT NOTE: NIH does not independently verify information submitted to the GTR; it relies on submitters to provide information that is accurate and not misleading. NIH makes no endorsements of tests or laboratories listed in the GTR. GTR is not a substitute for medical advice. **Patients and consumers** with specific questions about a genetic test should contact a health care provider or a genetics professional.

Sanger sequencing of entire coding sequence: 9,589
NGS sequencing of entire coding sequence: 1,451

<http://www.ncbi.nlm.nih.gov/gtr/>

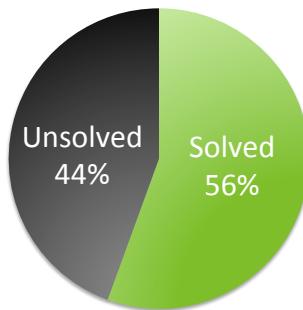
Diagnostic Yield: Gene Panels



Laboratory for Molecular Medicine
Lieve et al., 2013

Diagnostic Yield: Exome

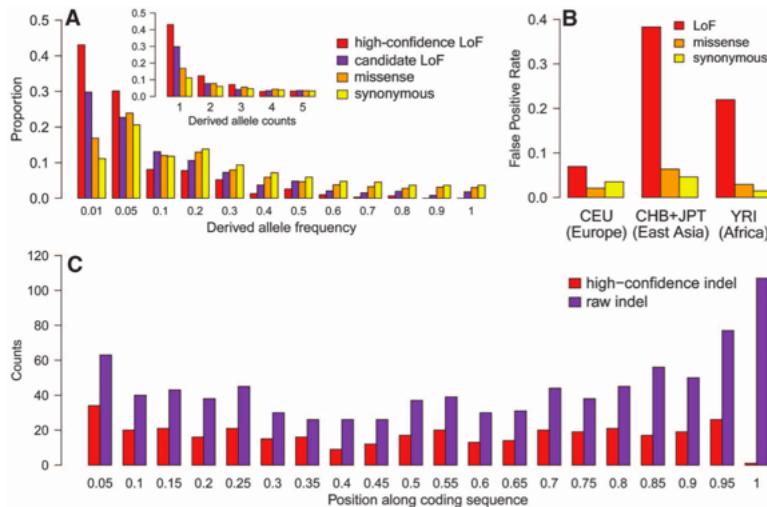
Personalis Data (56 cases) Yang et al., 2013
Diagnostic Yield Numerous ACMG abstracts



Numbers reported across labs

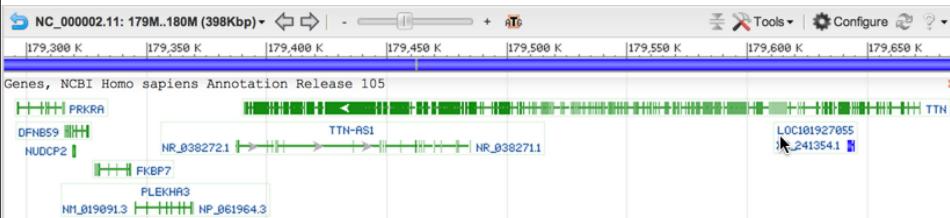
25-56%

A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes



MacArthur et al., 2012

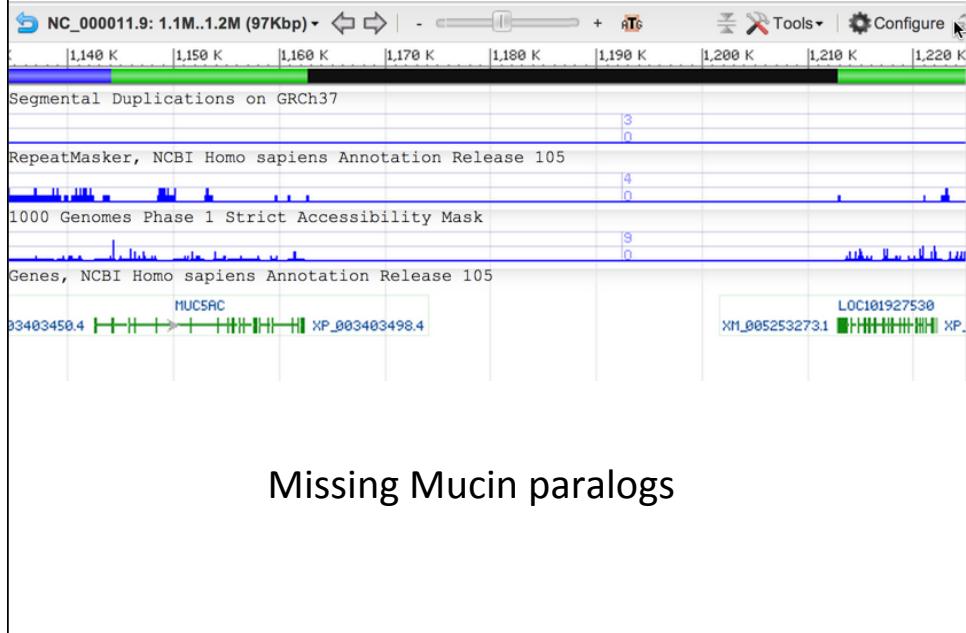
Statistical models to evaluate significance



Guidelines for investigating causality of sequence variants in human disease

D. G. MacArthur^{1,2}, T. A. Manolio³, D. P. Dimmock⁴, H. L. Rehm^{5,6}, J. Shendure⁷, G. R. Abecasis⁸, D. R. Adams^{9,10}, R. B. Altman¹¹, S. E. Antonarakis^{12,13}, E. A. Ashley¹⁴, J. C. Barrett¹⁵, L. G. Biesecker¹⁶, D. F. Conrad¹⁷, G. M. Cooper¹⁸, N. J. Cox¹⁹, M. J. Daly^{1,2}, M. B. Gerstein^{20,21}, D. B. Goldstein²², J. N. Hirschhorn^{2,23}, S. M. Leaf²⁴, L. A. Pennacchio^{25,26}, J. A. Stamatoyannopoulos²⁷, S. R. Sunyaev^{28,29}, D. Valle³⁰, B. F. Voight³¹, W. Winckler^{2†} & C. Gunter^{18†}

Statistical models to evaluate significance



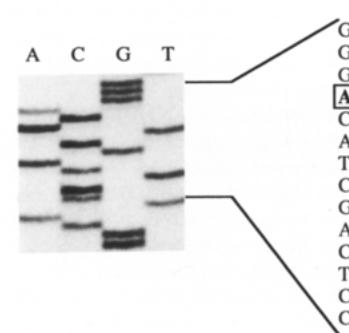
Integrating the literature

1994

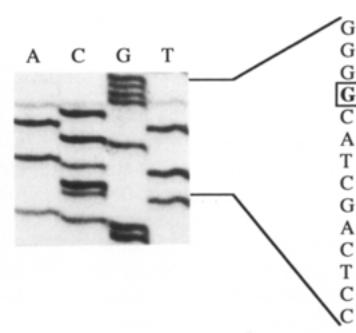
Mutations in the Transmembrane Domain of FGFR3 Cause the Most Common Genetic Form of Dwarfism, Achondroplasia

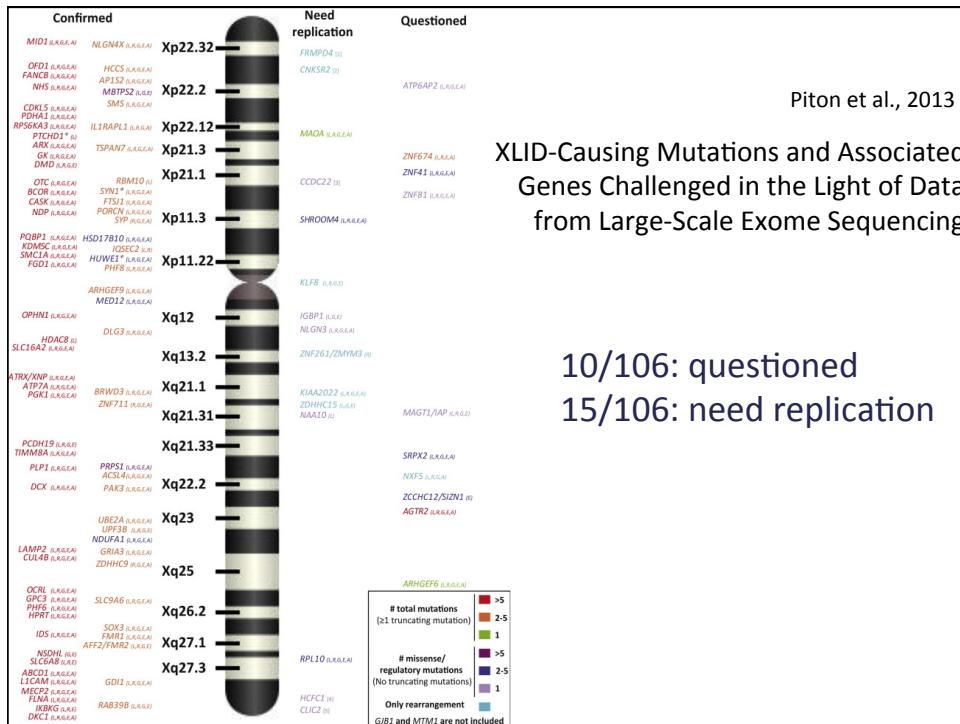
B

Mutant allele



Normal Allele





Integrating the literature

New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants

Charlotte Andreasen^{1,2,5}, Jonas B Nielsen^{1,2,5}, Lena Refsgaard^{1,2}, Anders G Holst^{1,2}, Alex H Christensen^{1,2}, Laura Andreasen^{1,2}, Ahmad Sajadieh³, Stig Haunso^{1,2,4}, Jesper H Svendsen^{1,2,4} and Morten S Olesen^{*,1,2}

Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale Resequencing

Yali Xue,¹ Yuan Chen,¹ Qasim Ayub,¹ Ni Huang,¹ Edward V. Ball,² Matthew Mort,² Andrew D. Phillips,² Katy Shaw,² Peter D. Stenson,² David N. Cooper,² Chris Tyler-Smith,^{1,*} and the 1000 Genomes Project Consortium

Integrating the literature

Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants

Alon Keinan^{1*} and Andrew G. Clark^{1,2}

Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants

Wenqing Fu¹, Timothy D. O'Connor¹, Goo Jun², Hyun Min Kang², Goncalo Abecasis², Suzanne M. Leal³, Stacey Gabriel⁴, David Altshuler⁴, Jay Shendure¹, Deborah A. Nickerson¹, Michael J. Bamshad^{1,5}, NHLBI Exome Sequencing Project* & Joshua M. Akey¹

Rare variants

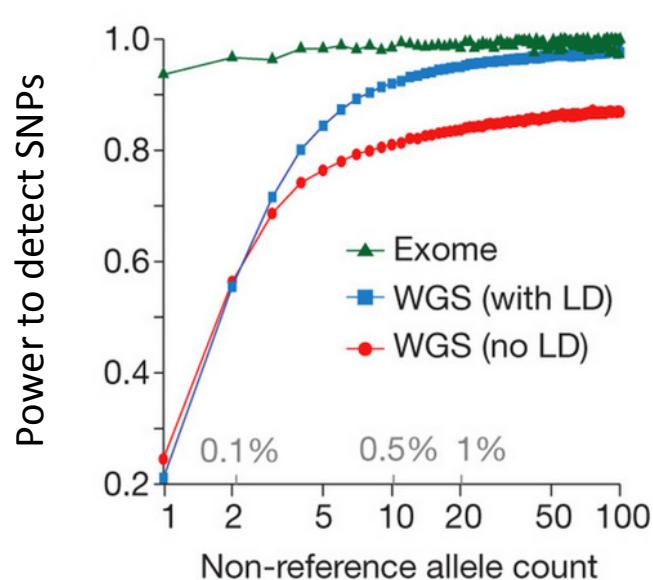


Figure 1b. 1000 Genomes Consortium, Nature 2012

ClinVar

NCBI Resources How To

ClinVar Advanced

Home About Data use and maintenance Using the website How to submit Statistics FTP site

ACTGATGGTATGGGCCAAGAGATATCT
CAGGTACGGCTGTCACTTAGACCTCAC
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC
CCATGGTGCATCTGACTCTCTGA~~G~~GAGAAGT
GCAGGTTGGTATCAAGGTACAAGACAGGT
GGCACTGACTCTCTGCTATTGGTCTAT

ClinVar

ClinVar aggregates information about sequence variation and its relationship to human health.

Using ClinVar

- About ClinVar
- Data Dictionary
- Downloads/FTP site
- FAQ
- Contact Us
- RSS feed/What's new?
- Factsheet

Tools

- ACMG Recommendations for Reporting of Incidental Findings
- Clinical Remapping service
- RefSeqGene/LRG
- Submissions
- Variation Reporter
- Variation Viewer

Related Sites

- ClinGen
- GeneReviews®
- GTR®
- ICCG
- MedGen
- OMIM®
- Variation

<http://www.ncbi.nlm.nih.gov/clinvar/>

ClinVar

BRCA1:c.4035delA (p.Glu1345Glu=fs)

BRCA1:c.4035delA (p.Glu1345Glu=fs)

Go to:

Variant type: Deletion

Cytogenetic location: 17q21.3

Genomic location: Chr17:43091496 (on Assembly GRCh38)
Chr17:41243513 (on Assembly GRCh37)

Other names: Glu1346LysfsX20
4154delA

HGVS: NG_005905.2:g.126488delA
NM_007294.3:c.4035delA
NM_007298.3:c.788-464delA
[...more](#)

Links: Breast Cancer Information Core (BIC) (BRCA1):
4154&base_change=del A
dbSNP: [80357711](#)

NCBI 1000 Genomes Browser: [rs80357711](#)

Molecular consequence: NM_007294.3:c.4035delA: frameshift variant
SO:0001589
NM_007298.3:c.788-464delA: intron SO:0001627
NR_027676.1:n.4171delA: non-coding transcript variant
SO:0001619

Clinical significance
BRCA1:c.4035delA (p.Glu1345Glu=fs)

Clinical significance: Pathogenic/Likely pathogenic

Review status: ★★★★☆

Number of submission(s): 4

Condition(s)

Familial cancer of breast [GeneReviews - MedGen - OMIM]
Breast-ovarian cancer, familial 1 [MedGen - Orphanet - OMIM]
BRCA1 and BRCA2 Hereditary Breast and Ovarian Cancer [MedGen]

See supporting ClinVar records

1 Affected Gene

breast cancer 1, early onset (BRCA1) [Gene - OMIM - Variation view]

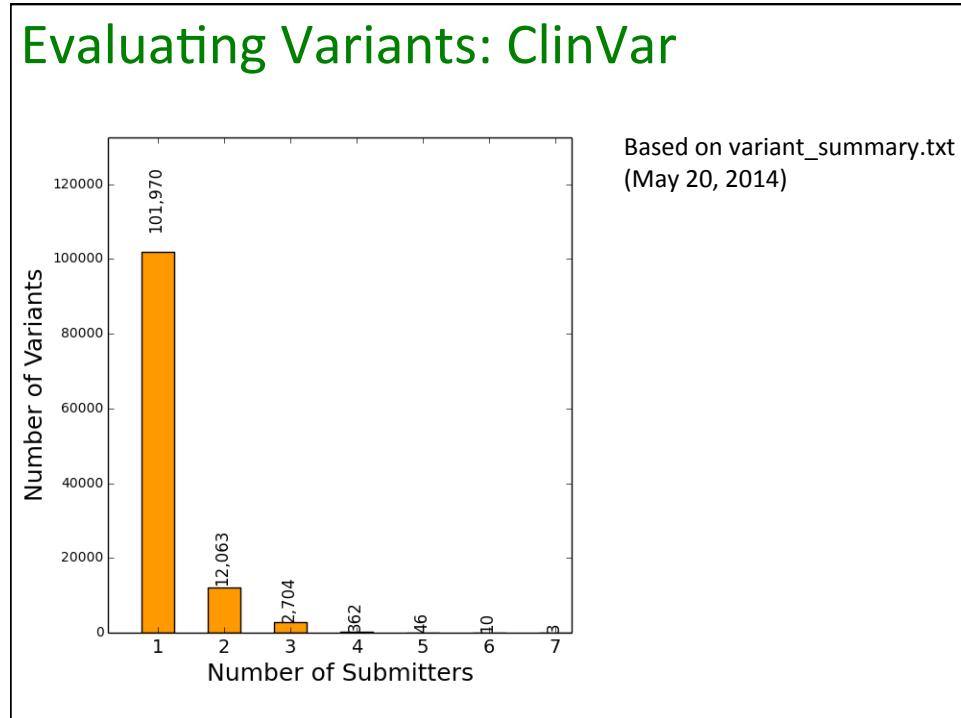
Search ClinVar for variants within BRCA1
Search ClinVar for variants including BRCA1

Browser views

RefSeqGene
GRCh38 Assembly
GRCh37 Assembly

Key is the variant-phenotype relationship
Variants defined in archives (dbSNP/dbVar)

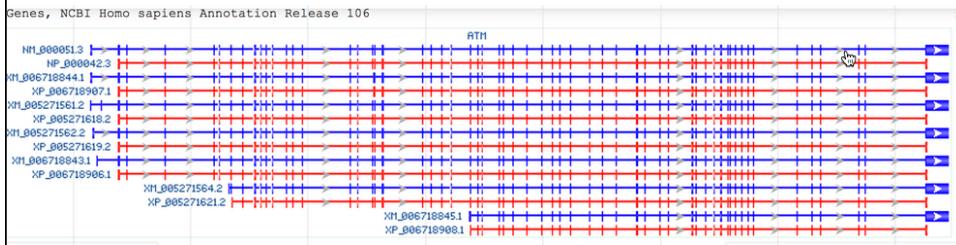
ClinVar								
Germline								
Clinical significance (Last evaluated)	Review status (Assertion method)	Collection method	Condition(s) (Mode of inheritance)	Origin	Citations	Submitter (Last submitted)	Submission accession	
Pathogenic (Jun 18, 2012)	classified by single submitter (clinical testing)	clinical testing	Breast-ovarian cancer, familial 1 [MedGen Orphanet OMIM]	germline		Sharing Clinical Reports Project (SCRIP) (Dec 30, 2013)	SCV000053740	
Pathogenic (Jun 11, 2014)	classified by single submitter (clinical testing)	clinical testing	BRCA1 and BRCA2 Hereditary Breast and Ovarian Cancer [MedGen]	germline	PubMed (1)	Invitae (Jun 11, 2014)	SCV000076426	
Pathogenic	classified by single submitter (clinical testing)	clinical testing	Familial cancer of breast [GeneReviews MedGen OMIM]	germline		GeneDx (Nov 8, 2013)	SCV000108672	
Pathogenic (May 29, 2002)	classified by single submitter (clinical testing)	clinical testing	Breast-ovarian cancer, familial 1 [MedGen Orphanet OMIM]	germline, not provided		Breast Cancer Information Core (BIC) (BRCA1) (Mar 28, 2014)	SCV000144938	



Communicating knowledge

ATM: 5762ins137

“137 nt added between exons 40 and 41”



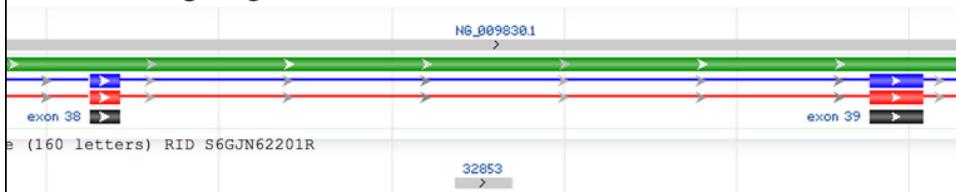
McConville et al., 1996

Communicating knowledge

```

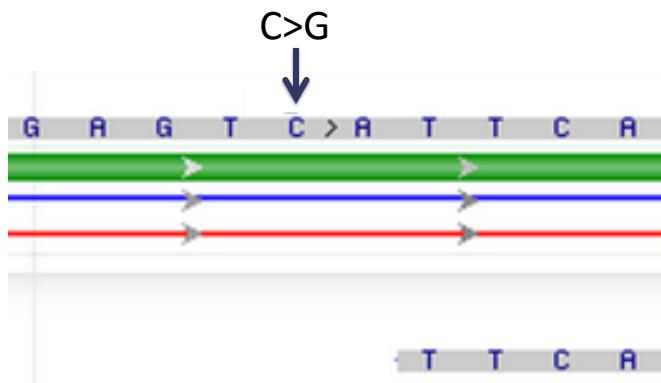
1 ttcatttcag ACTTACCATT GGGGTAGAAT GGTGTTGACA TTAGCAGGAG
51 TTGGAGAAGA TAAAAATTCT TCATATTCCA TTTTAATGCT GAATAAGATC
101 CTGAAGAATA TTCCTGGAGA AAGTATGAAT GGGATATAGA AAAACGGta
151 aagacatgca

```



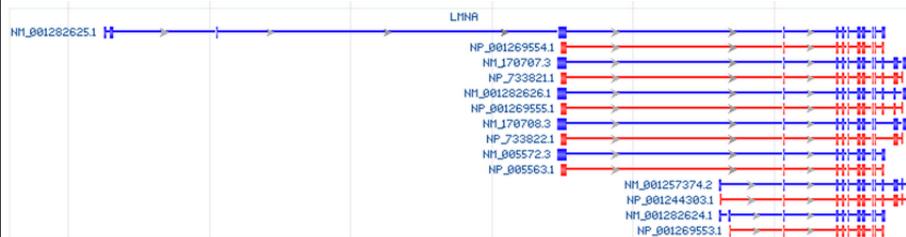
Communicating knowledge

NC_000011.9 (chr11) 108179684



Communicating knowledge

LMNA:c.82C>T; p.R28W



Turk et al., 2013

Communicating knowledge

Ref AGATTCAC
 Alt AGATTTCAC

Valid VCF (left shift):

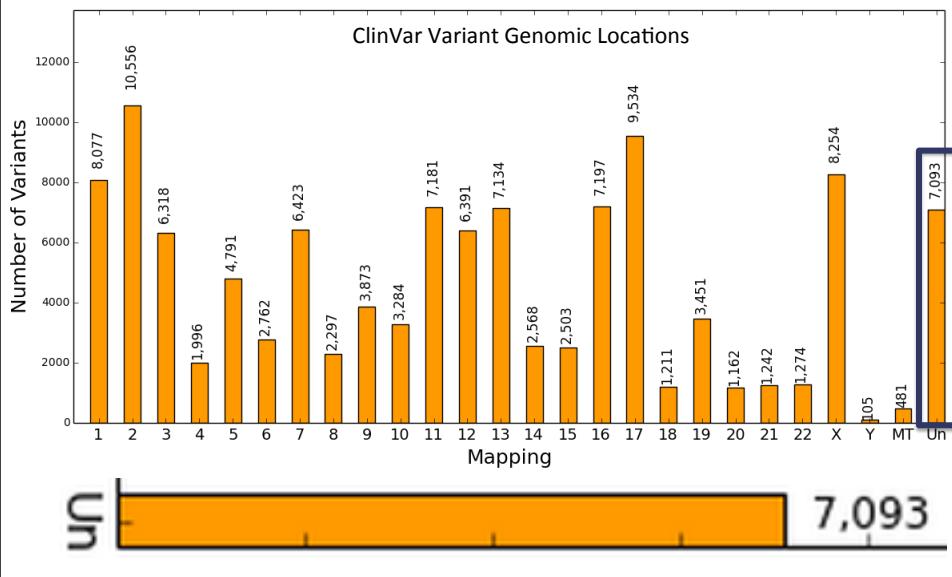
```
#CHROM POS ID REF ALT QUAL FILTER INFO
22 3 123 A AT . . .
```

Valid HGVS (right shift):

chr22:g.5dupT

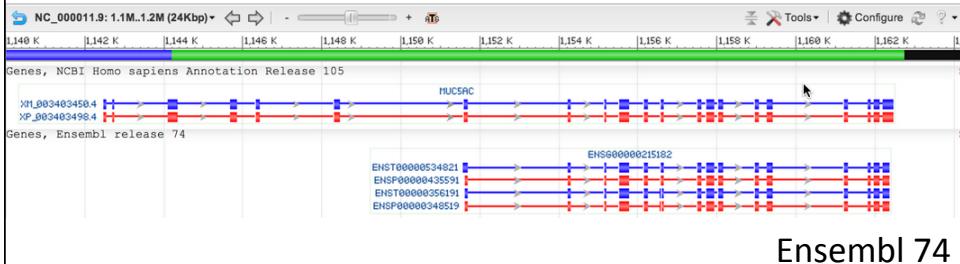
NOTE: Better to use NC_000022.10 instead of chr22/22

Communicating knowledge



Communicating knowledge

NCBI Homo sapiens annotation run 105



McCarthy DJ, et al., 2014

Take home messages

- More Sequence
- Better Phenotyping
- More precise data handling