Title

How and why do scientists reuse others' data to produce new knowledge?

Authors

Borgman, Christine
Pasquetto, Irene V.

Publication Date

2018-09-15

How and why do scientists reuse others' data to produce new knowledge? Background, Foreground, and Beyond

Christine L. Borgman, PhD

Distinguished Research Professor
Center for Knowledge Infrastructures
University of California, Los Angeles
http://christineborgman.info
https://knowledgeinfrastructures.gseis.ucla.edu
@scitechprof

Irene V. Pasquetto, PhD

Research Affiliate
UCLA Center for Knowledge Infrastructures

Fringe Event
Cochrane Colloquium
Edinburgh, 15 September 2018

Cochrane

Trusted evidence.
Informed decisions.
Better health.

Christine Borgman

Bernie Boscoe

Peter Darch

Milena Golshan

Irene Pasquetto

Michael Scroggins

Cheryl Thompson

Morgan Wofford

ALFRED P. SLOAN FOUNDATION 1934

UCLA Center for Knowledge Infrastructures

THE WHITE HOUSE
WASHINGTON

Data sharing policies

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

- Research Councils of the UK
- European Union
- U.S. Federal research policy
- Australian Research Council
- Individual countries, funding agencies, journals, universities

JST JAPAN SCIENCE AND TECHNOLOGY AGENCY

中央研究院
中央研究院 ACADEMIA SINICA 1928

中国科学院 CHINESE ACADEMY OF SCIENCES

Supported by
wellcometrust

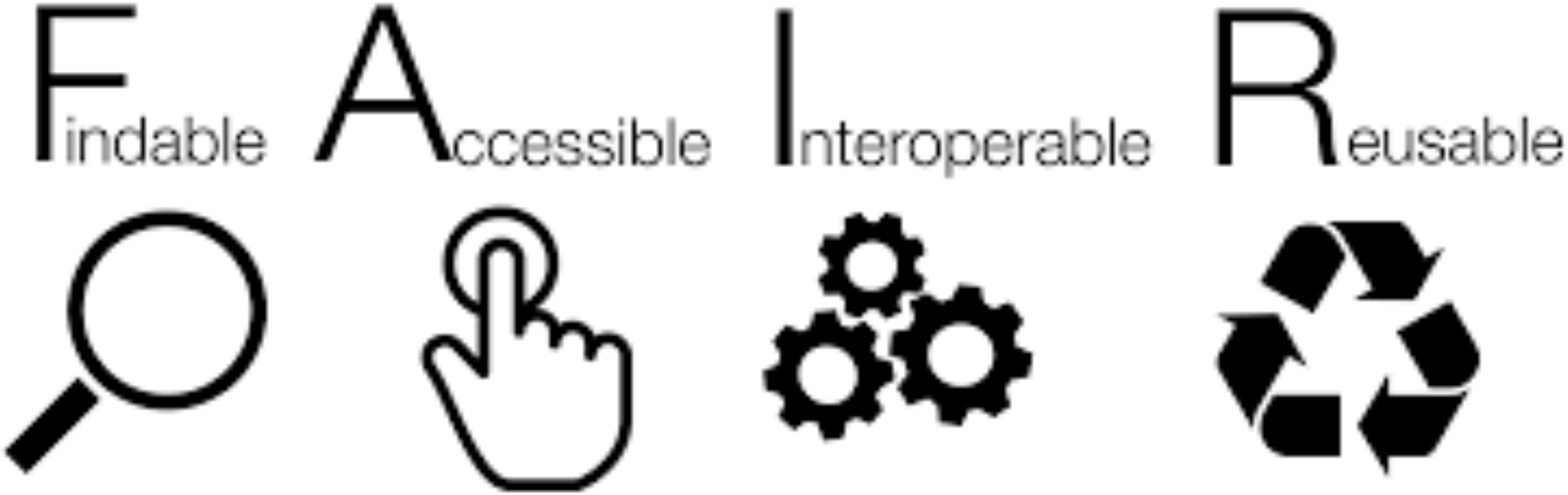
Australian Government
National Health and Medical Research Council

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

NIH National Institutes of Health
Turning Discovery Into Health

Policy RECommendations for Open Access to Research Data in Europe
RECODE
DETAILS

Data Stewardship: The Ideal



Wilkinson, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, http://dx.doi.org/10.1038/sdata.2016.18

3

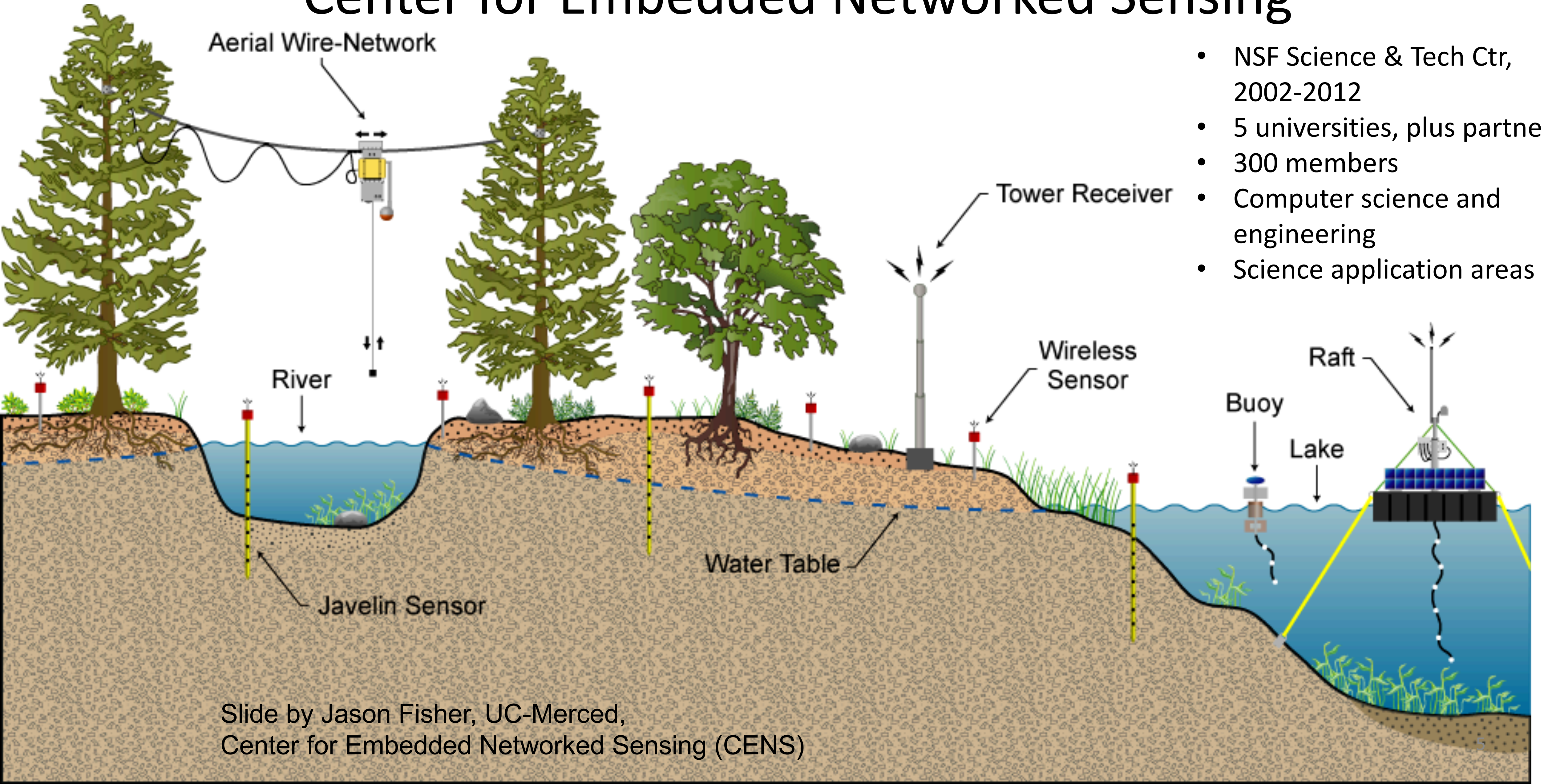
What is “data reuse”?

Image source: https://datasupport.researchdata.nl

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). **On the Reuse of Scientific Data**. *Data Science Journal*, *16*. https://doi.org/10.5334/dsj-2017-008

Center for Embedded Networked Sensing

- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas

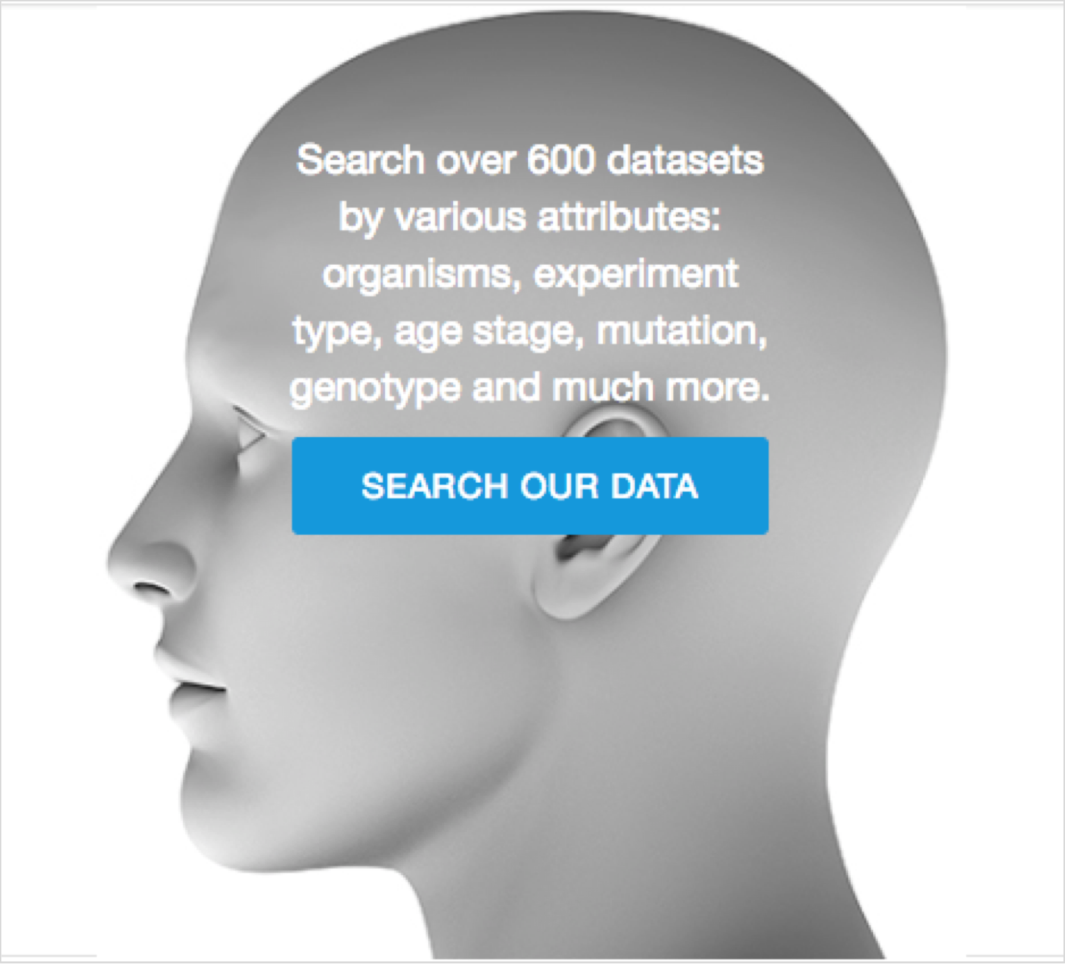


Slide by Jason Fisher, UC-Merced,
Center for Embedded Networked Sensing (CENS)

Background and Foreground Reuses of data at CENS

Images: CKI and NSF archives

The DataFace Consortium for Data Sharing



Search over 600 datasets by various attributes: organisms, experiment type, age stage, mutation, genotype and much more.

GOAL:

Collect and release high-throughput “hypothesis free” biomedical data related to the genetics of facial formation and development in humans and animals.

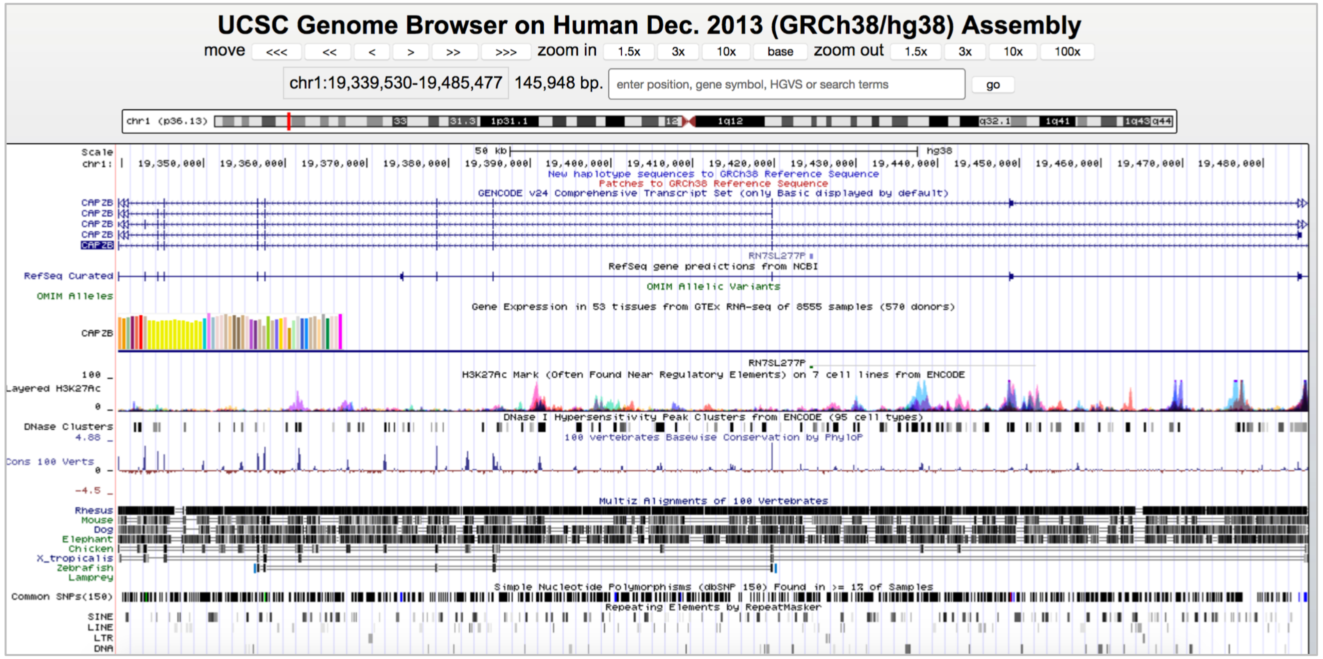
DATA TYPES:

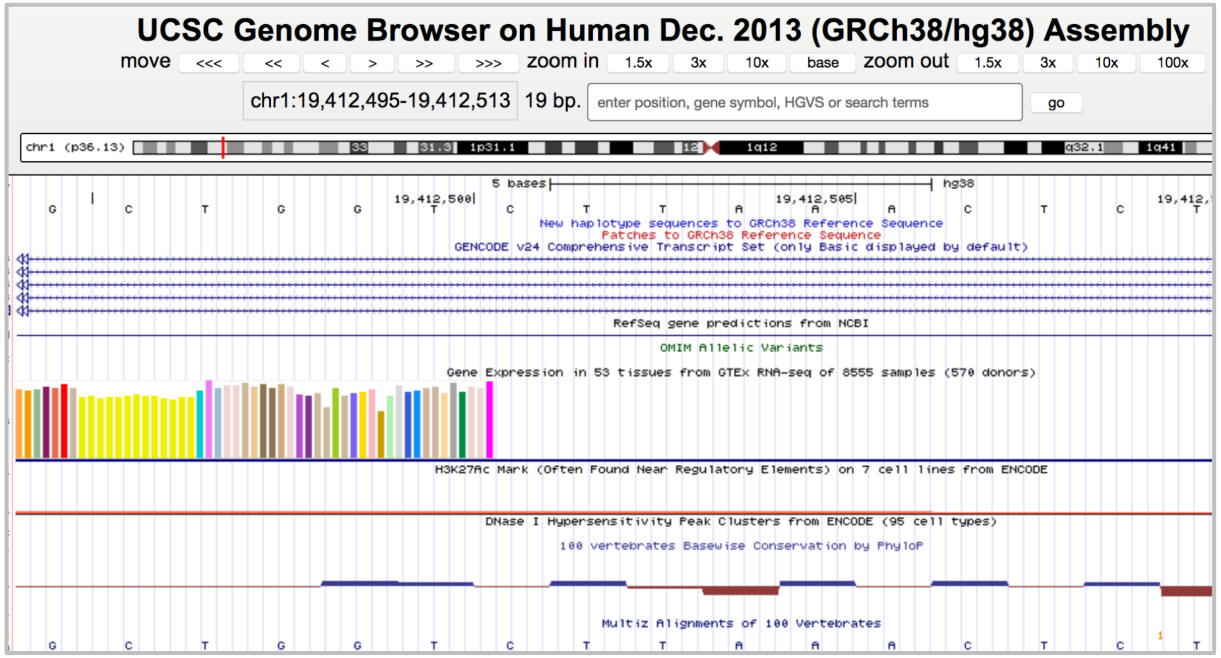
Whole genome sequences, gene expression data from ChiP-seq, RNA-seq, and microarrays, genotypes and phenotypes from GWAS studies, etc.

DOMAINS:

Developmental biologists, evolutionary experts, human geneticists, computational biologists, surgeons, etc.

Background Reuse at DataFace: Comparison, control, verification. (I)


UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr1:19,339,530-19,485,477 145,948 bp.
enter position, gene symbol, HGVS or search terms
go
chr1 (p36.13)
Scale
chr1:
50 kb
hg38
19,350,000 19,360,000 19,370,000 19,380,000 19,390,000 19,400,000 19,410,000 19,420,000 19,430,000 19,440,000 19,450,000 19,460,000 19,470,000 19,480,000
New haplotype sequences to GRCh38 Reference Sequence
Patches to GRCh38 Reference Sequence
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
CAPZB
RN7SL277P
RefSeq Curated
RefSeq gene predictions from NCBI
OMIM Alleles
OMIM Allelic Variants
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
100
0
DNase Clusters
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
4.88
-4.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
Common SNPs(150)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA

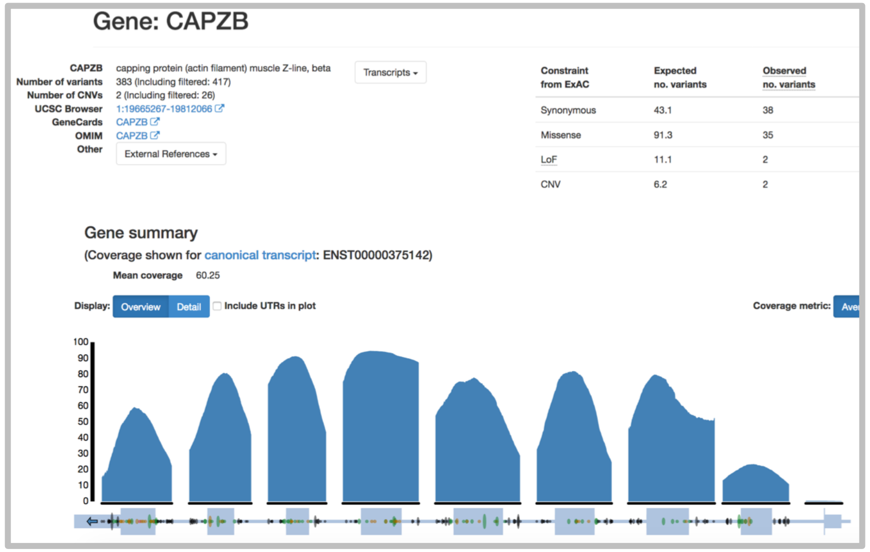

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr1:19,412,495-19,412,513 19 bp.
enter position, gene symbol, HGVS or search terms
go
chr1 (p36.13)
5 bases
hg38
19,412,500
19,412,505
G C T G G T C T T A A A C T C T
New haplotype sequences to GRCh38 Reference Sequence
Patches to GRCh38 Reference Sequence
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
RefSeq gene predictions from NCBI
OMIM Allelic Variants
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates

UCSC Genome Browser – Search example (CAPZB gene)

UCSC Genome Browser – Zoom IN

Irene Pasquetto, DataFace Study, 2018

Background Reuse at DataFace: Comparison, control, verification. (II)



Gene: CAPZB

CAPZB	capping protein (actin filament) muscle Z-line, beta
Number of variants	383 (Including filtered: 417)
Number of CNVs	2 (Including filtered: 26)
UCSC Browser	1:19665267-19812066
GeneCards	CAPZB
OMIM	CAPZB
Other	External References ▾

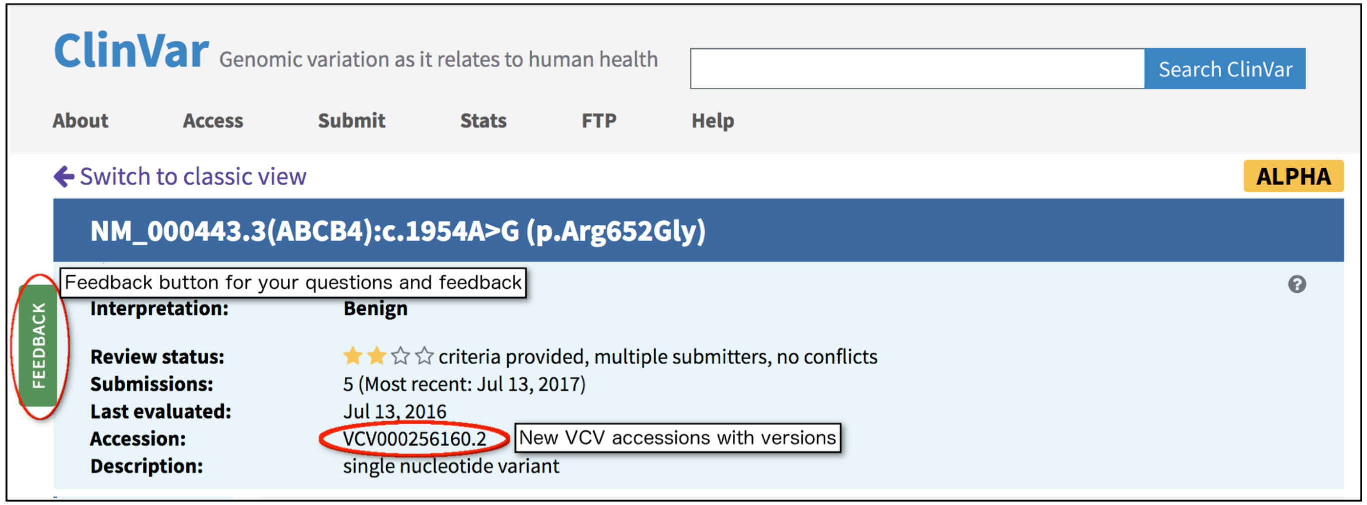
Constraint from ExAC	Expected no. variants	Observed no. variants
Synonymous	43.1	38
Missense	91.3	35
LoF	11.1	2
CNV	6.2	2

Gene summary

(Coverage shown for canonical transcript: ENST00000375142)

Mean coverage 60.25

Display: Overview Detail □ **Include UTRs in plot** **Coverage metric:** Ave

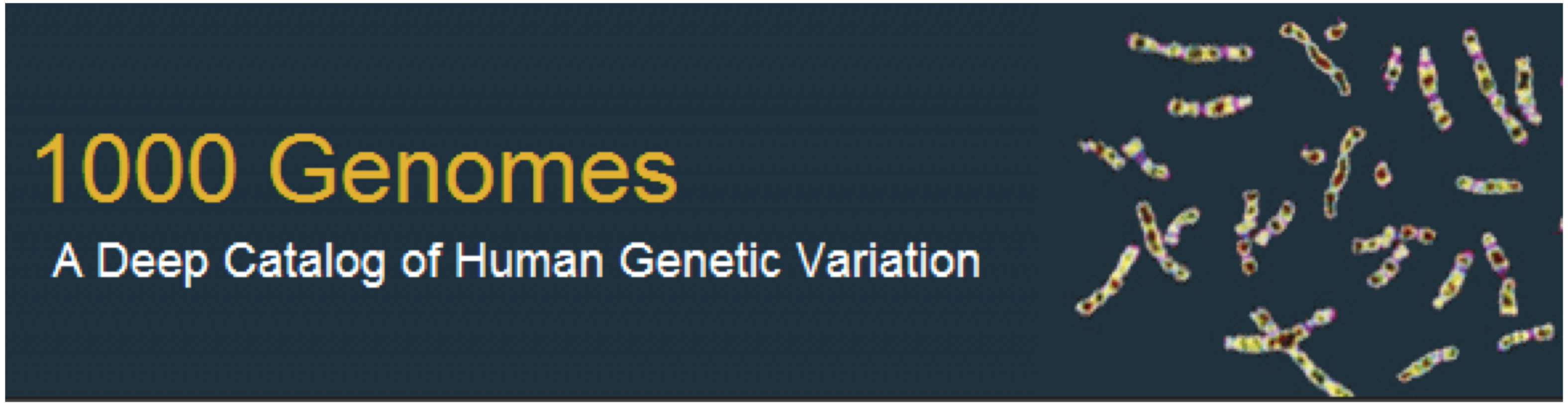


NM_000443.3(ABCB4):c.1954A>G (p.Arg652Gly)

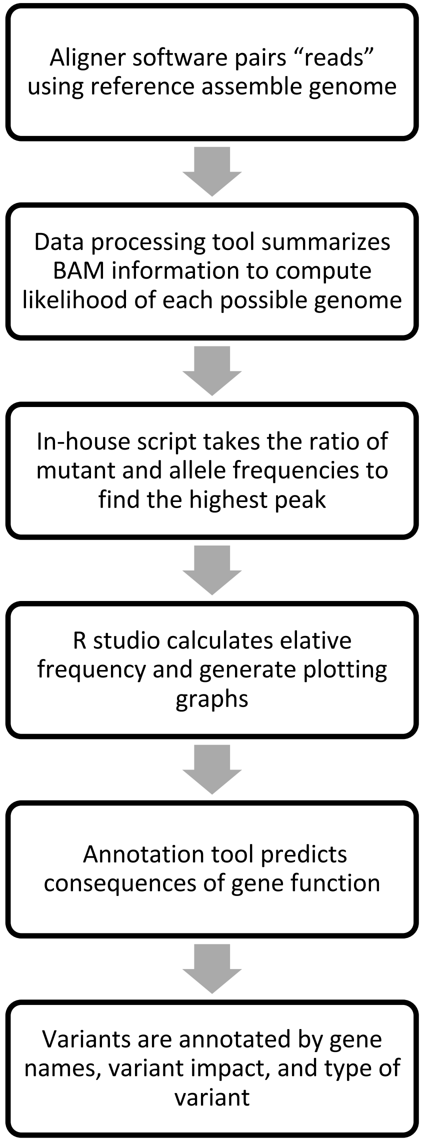
Feedback button for your questions and feedback

Interpretation:	**Benign**
Review status:	★★☆☆ criteria provided, multiple submitters, no conflicts
Submissions:	5 (Most recent: Jul 13, 2017)
Last evaluated:	Jul 13, 2016
Accession:	VCV000256160.2
Description:	single nucleotide variant

New VCV accessions with versions



Foreground Reuse at DataFace: Data Analysis



Aligner software pairs "reads" using reference assemble genome

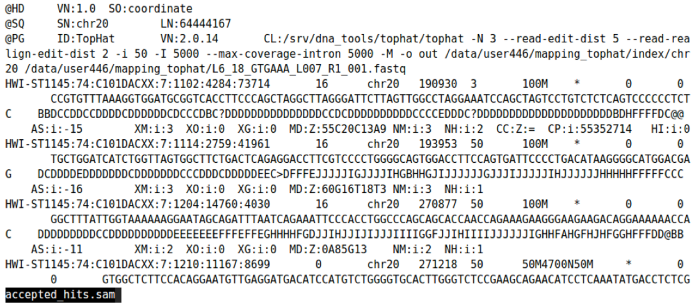
Data processing tool summarizes BAM information to compute likelihood of each possible genome

In-house script takes the ratio of mutant and allele frequencies to find the highest peak

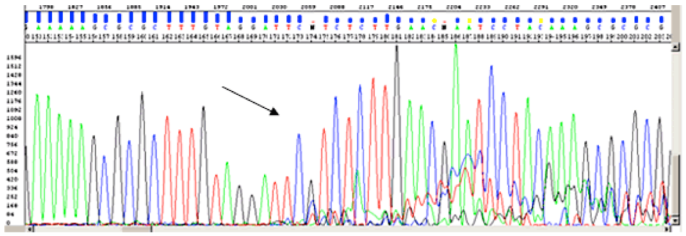
R studio calculates elative frequency and generate plotting graphs

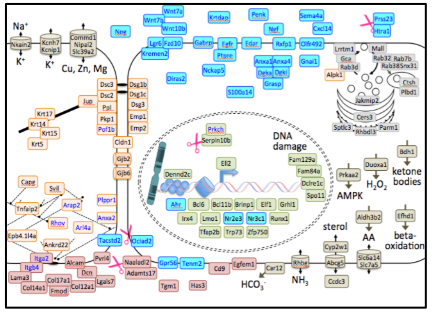
Annotation tool predicts consequences of gene function

Variants are annotated by gene names, variant impact, and type of variant



```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```





"RAW" DATA

Pipeline

RESULTS

Having access to the contact information of those who collected the data increases rates of foreground reuse.

Image source: https://www.123rf.com

The “Data Creator Advantage”

- Creator has most current annotations about the dataset
- Creator has most specialized knowledge of relevant literature
- Creator may have software pipelines locally customized for the dataset

Image source: https://www.siteminder.com

	BACKGROUND Reuse of Data	**FOREGROUND Reuse of data**
Goal of reuse	“Ground truthing:” calibrate, compare, confirm	Analysis: identify patterns, correlations, causal relationships
Example of reuse	Instrument calibration, sequence annotation, review summary-level data	Meta-analyses, novel statistical analyses
Frequency of reuse	Frequent, routine practice	Rare, emergent practice

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (in review). The Who, What, When, and Why of Reusing Data in Scientific Practice. *Harvard Data Science Review*.

	BACKGROUND Reuse of Data	**FOREGROUND Reuse of data**
Goal of reuse	“Ground truthing:” calibrate, compar confirm	Analyses: identi patterns, co ns, causal r
Example of reuse	Instrume , sequ on, r mary-level da	Met s.
Frequency of reuse	Freq nt - routine practice	Rare - emergent practice

INDEPENDENT REUSE OF DATA

COLLABORATIVE REUSE WITH DATA CREATORS

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (in review). The Who, What, When, and Why of Reusing Data in Scientific Practice. *Harvard Data Science Review*.

Questions: Trusted Evidence?

- When to reuse open data independently?
- When to collaborate with data creators?
- What information is needed, when, to trust evidence?

Questions: Informed decisions?

- What do you need to know about the data to inform decisions?
- When are data sufficient for decision making?
- When is further information about about data needed?
- How should data sharing and reuse be governed?

Questions: Better health?

- Where should community invest in data sharing and reuse?
- How should data resources be governed?
- Who should be responsible for sustaining access to health data?
- What are reasonable licensing agreements?
- What are appropriate funding models for data resources?

Acknowledgements

UCLA Center for Knowledge Infrastructures

Christine Borgman

Bernie Boscoe

Peter Darch

Milena Golshan

Irene Pasquetto

Michael Scroggins

Cheryl Thompson

Morgan Wofford