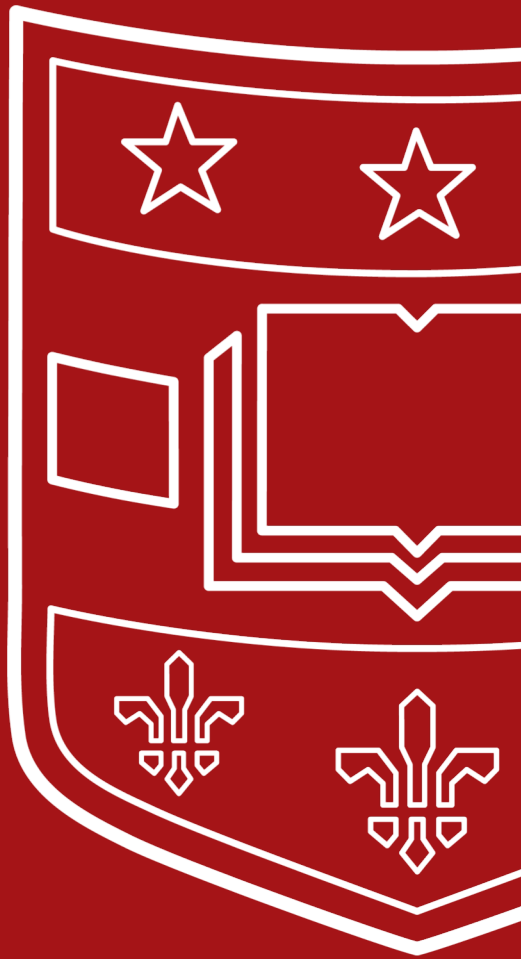
Clinical Interpretation Knowledgebases

and Learning Bioinformatics

Alex H Wagner, PhD
December, 2019 – UTH BIO390



Part I – Clinical Interpretation Knowledgebases

@handlerwagner

SIB
Swiss Institute of Bioinformatics

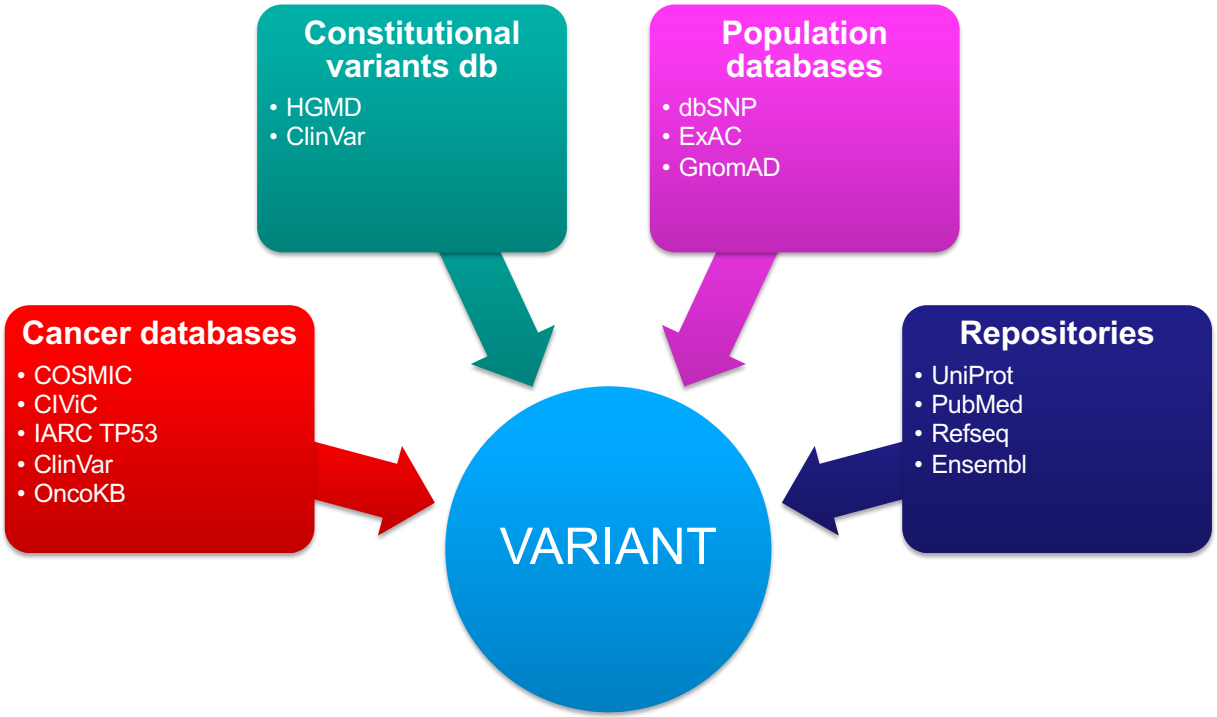
Introduction to bioinformatics: Clinical Bioinformatics

Valérie Barbié, head of SIB Clinical Bioinformatics

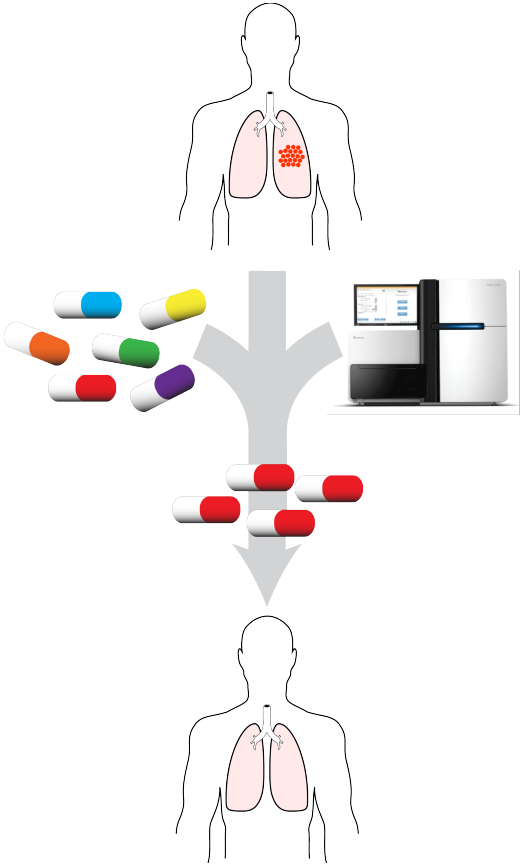
Zürich, 26 November 2019

Annotating a variant: knowledgebases

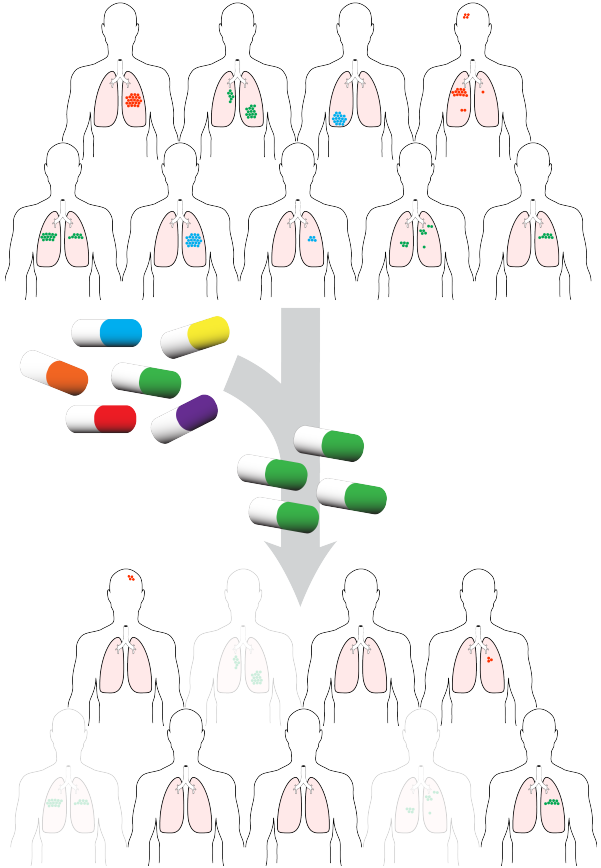
not exhaustive



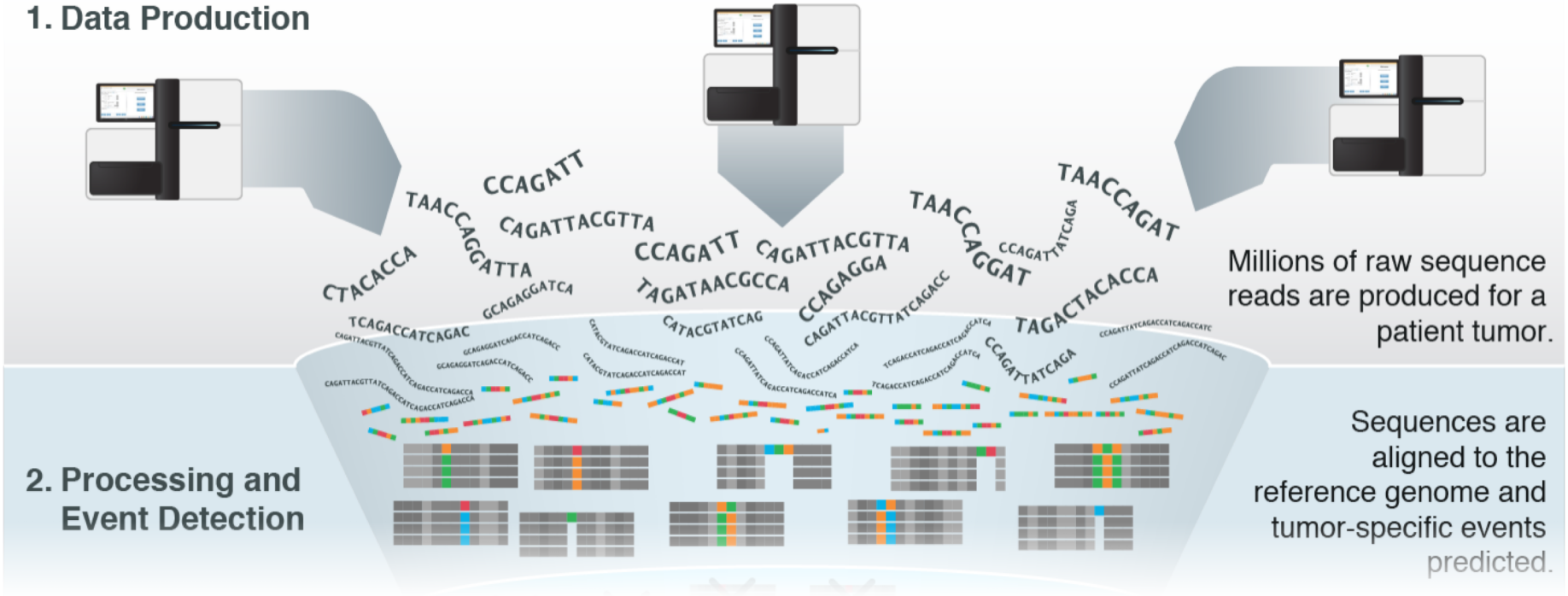
Precision medicine



Conventional therapies

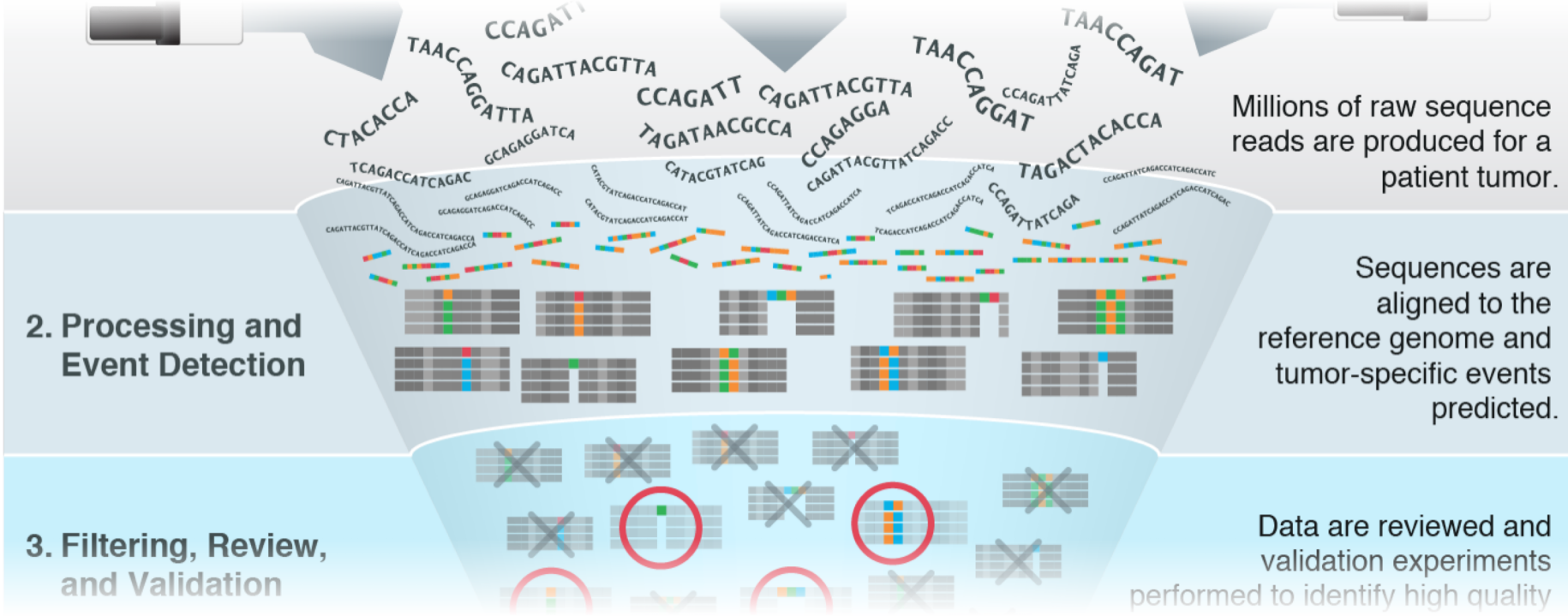


Precision Oncology Bottleneck



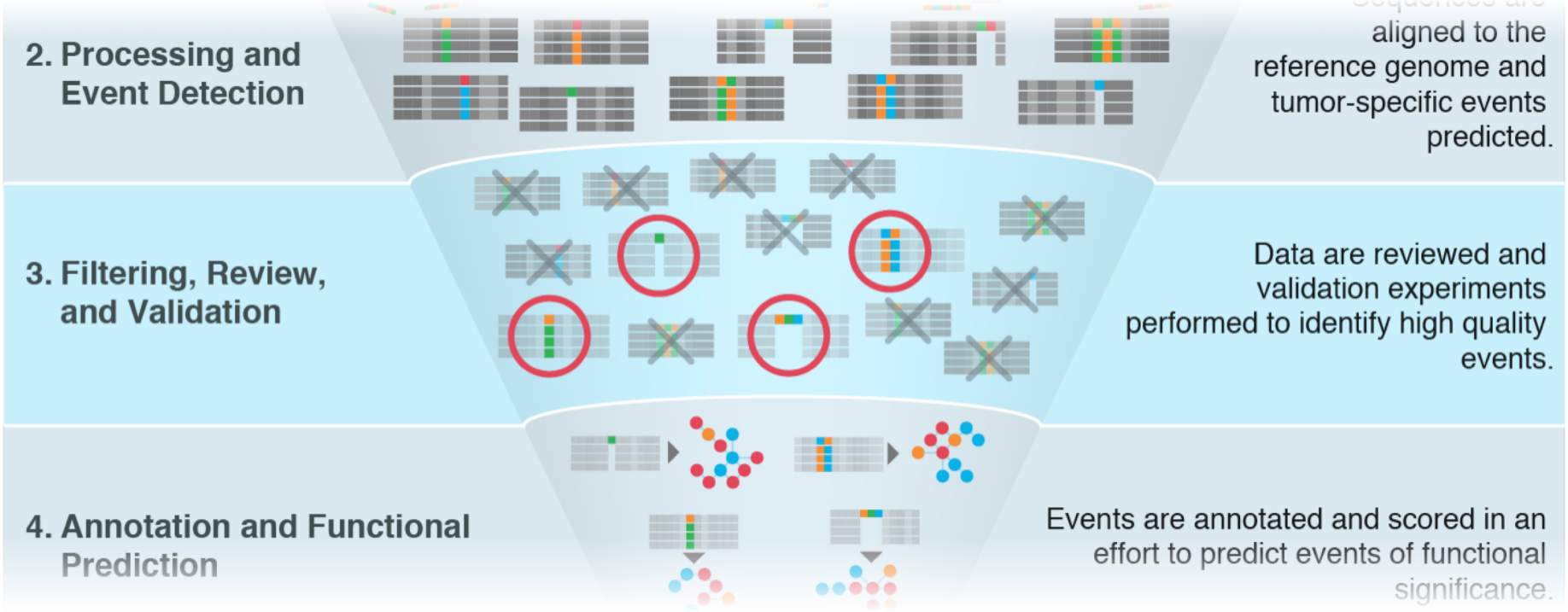
Good BM, Ainscough BJ, McMichael JF, Su AI†, Griffith OL†. 2014. Genome Biology. 15(8):438.

Precision Oncology Bottleneck



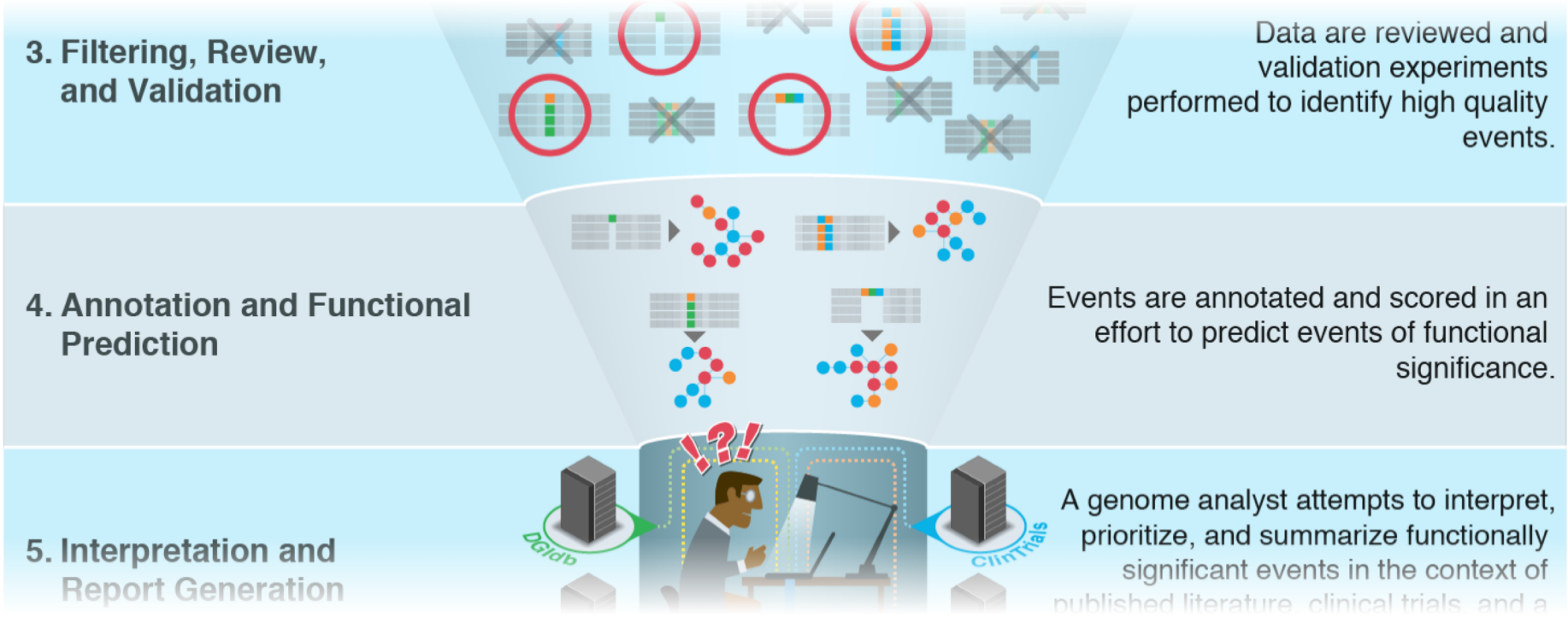
Good BM, Ainscough BJ, McMichael JF, Su AI†, Griffith OL†. 2014. Genome Biology. 15(8):438.

Precision Oncology Bottleneck



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

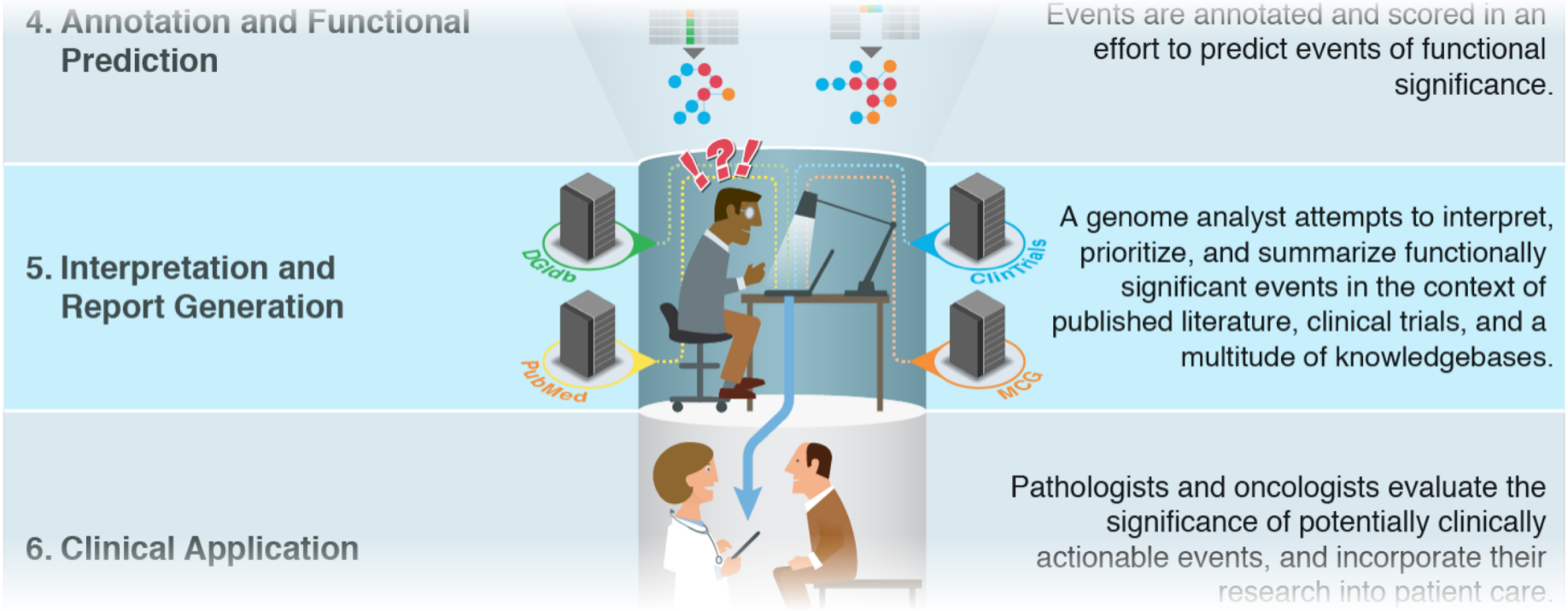
Precision Oncology Bottleneck



@handlerwagner

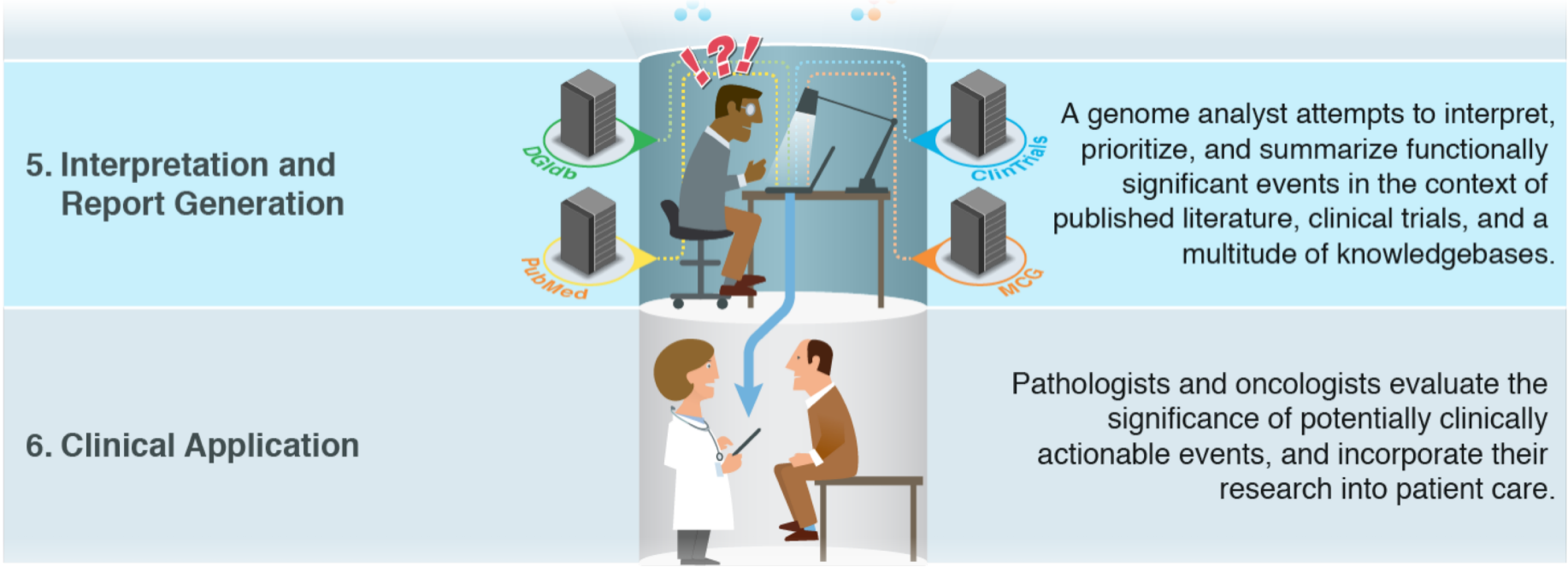
Good BM, Ainscough BJ, McMichael JF, Su AI†, Griffith OL†. 2014. Genome Biology. 15(8):438.

Precision Oncology Bottleneck



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

Precision Oncology Bottleneck



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

Precision Oncology Bottleneck

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

Precision Oncology Bottleneck

How do we alleviate this bottleneck?

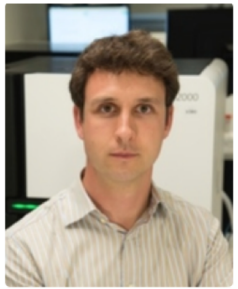
@handlerwagner

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

CIViC

CLINICAL INTERPRETATIONS OF VARIANTS IN CANCER

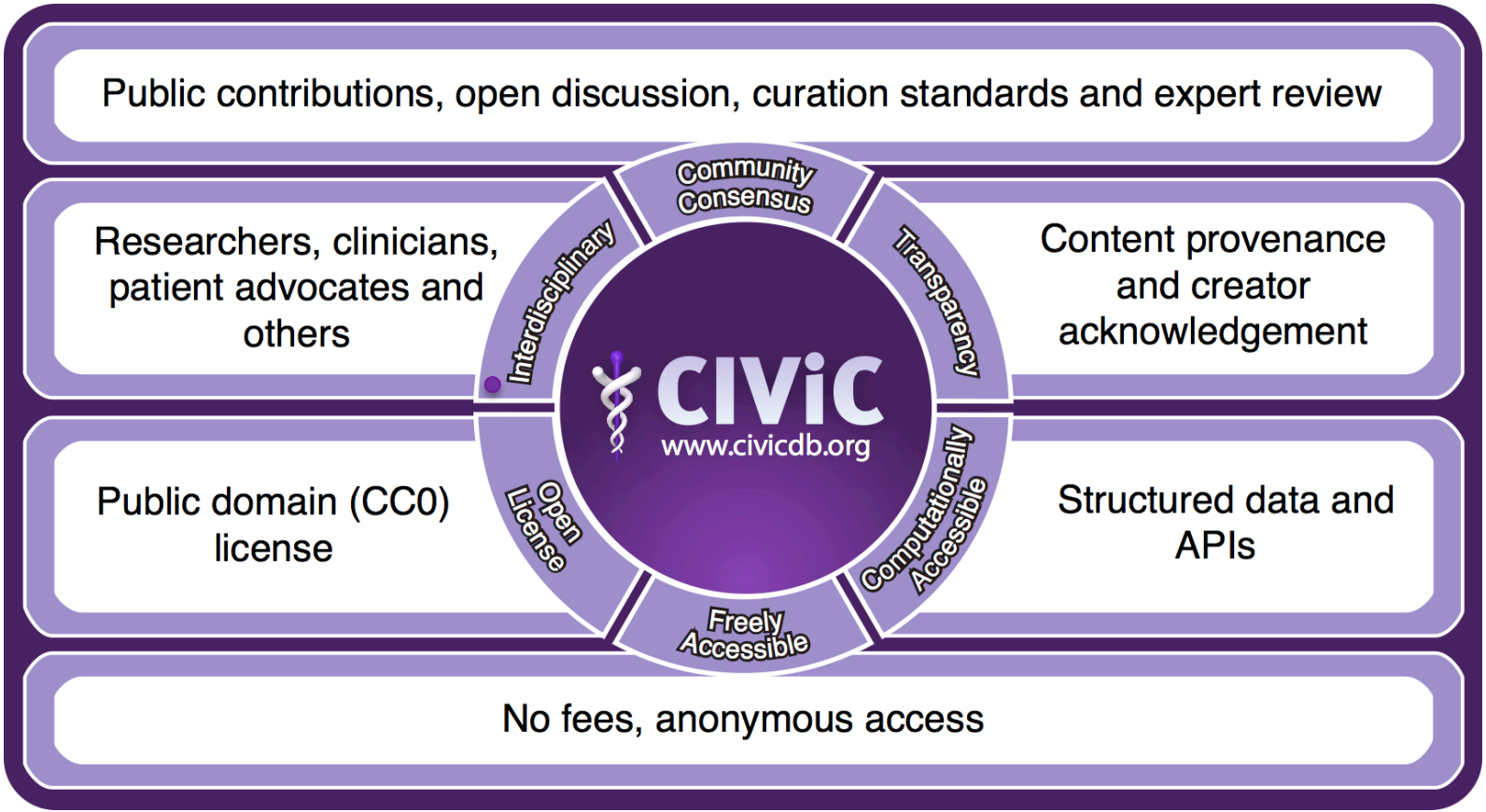
civicdb.org



Obi Griffith

Malachi Griffith

CIViC principles emphasize collaborative, open sharing of interpretations



What is a variant of *significance* in cancer?

Prognostic of survival change

- *Biallelic CEBPA mutations are associated with improved overall survival in patients with acute myeloid leukemia*

Predictive of therapeutic response

- *BRAF V600E predicts sensitivity to vemurafenib*

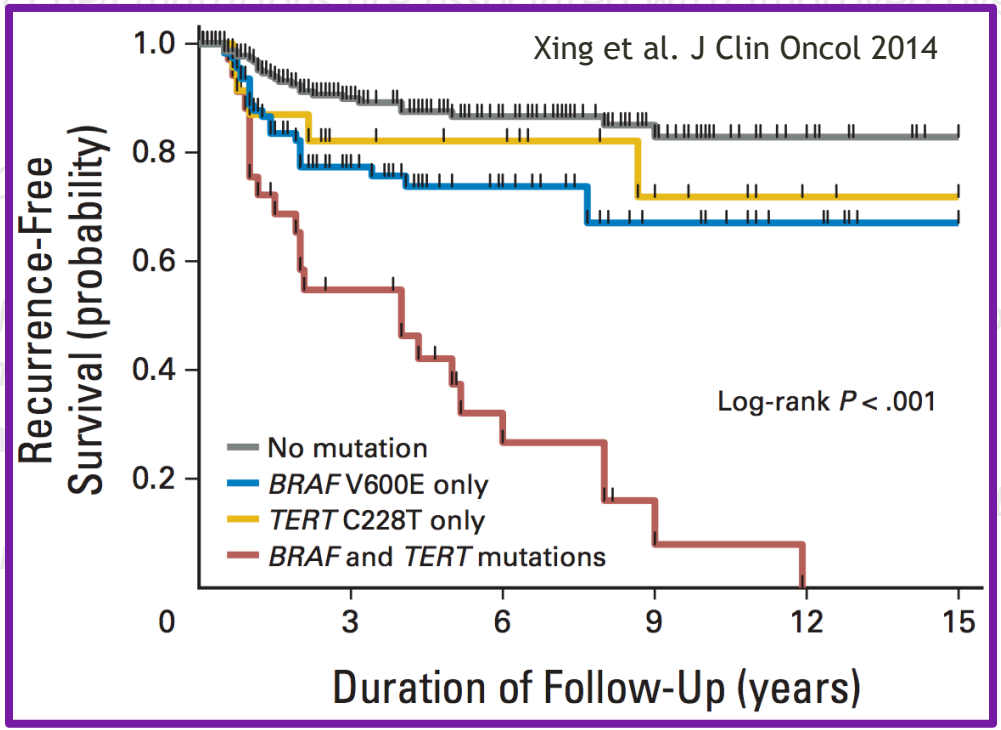
Diagnostic of tumor subtype

- *DNAJB1-PRKACA fusion differentiates fibrolamellar hepatocellular carcinoma from conventional HCC*

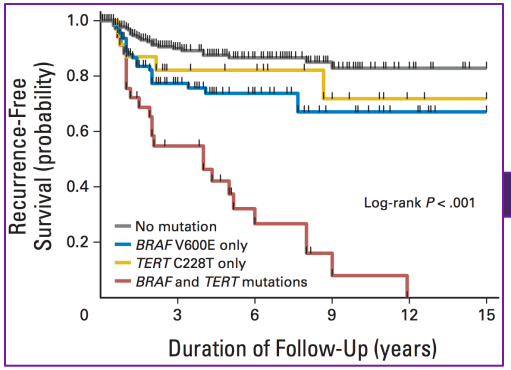
Predisposing for cancer development

- *Patients with the RUNX1 Y260* mutation are associated with increased risk of developing acute myeloid leukemia*

What is a variant of *significance* in cancer?



CIViC provides a literature curation interface for the interpretation of clinically-relevant variants



Structured Evidence Item

EVIDENCE EID656

Evidence Summary | Evidence Talk

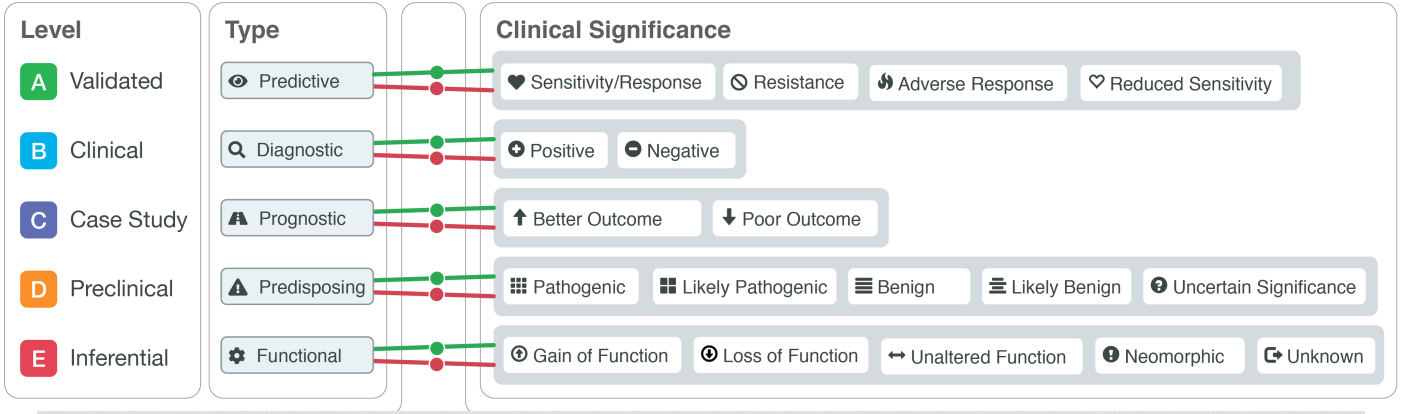
Submitted by gatoravi | Last Modified by kkrysiak | Last Reviewed by ahwagner | Accepted by NickSpies

In patients with papillary thyroid cancer harboring both BRAF V600E and the TERT promotor mutation C228T (N=35), recurrence-free survival is worse than in patients harboring one of these mutations (N=159 BRAF, N=26 TERT promoter mutated) or no mutations in either gene (N=287)(P<0.001).

Evidence Level: B - Clinical
Evidence Type: Prognostic
Evidence Direction: Supports
Clinical Significance: Poor Outcome
Variant Origin: Somatic Mutation

Disease: Papillary Thyroid Carcinoma
Associated Phenotype: --
Source: Xing et al., 2014, J. Clin. Oncol.
PubMed ID: 25024077
Clinical Trial: --
Trust Rating: ★★★★★

Curated evidence items are the foundational unit of CIViC



EVIDENCE EID656

Evidence Summary | Evidence Talk

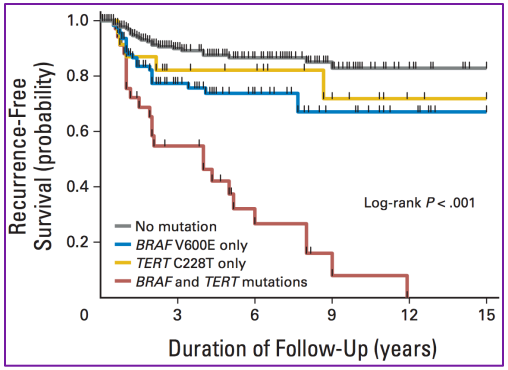
Submitted by gatoravi | Last Modified by kkrysiak | Last Reviewed by ahwagner | Accepted by NickSpies

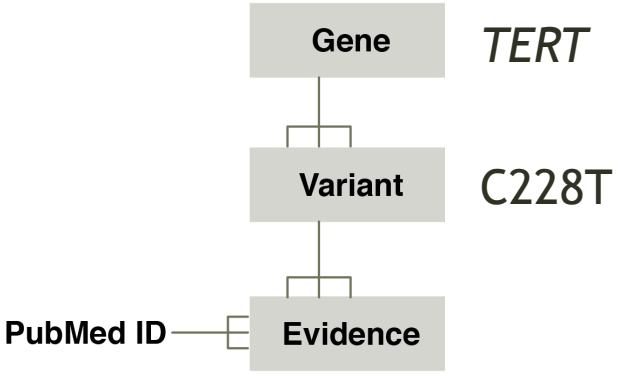
In patients with papillary thyroid cancer harboring both BRAF V600E and the TERT promotor mutation C228T (N=35), recurrence-free survival is worse than in patients harboring one of these mutations (N=159 BRAF, N=26 TERT promoter mutated) or no mutations in either gene (N=287)(P<0.001).

Evidence Level: B - Clinical
Evidence Type: Prognostic
Evidence Direction: Supports
Clinical Significance: Poor Outcome
Variant Origin: Somatic Mutation

Disease: Papillary Thyroid Carcinoma
Associated Phenotype: --
Source: Xing et al., 2014, J. Clin. Oncol.
PubMed ID: 25024077
Clinical Trial: --
Trust Rating: ★★★★★

Curated evidence items are the foundational unit of CIViC





EVIDENCE EID656

Evidence Summary | Evidence Talk

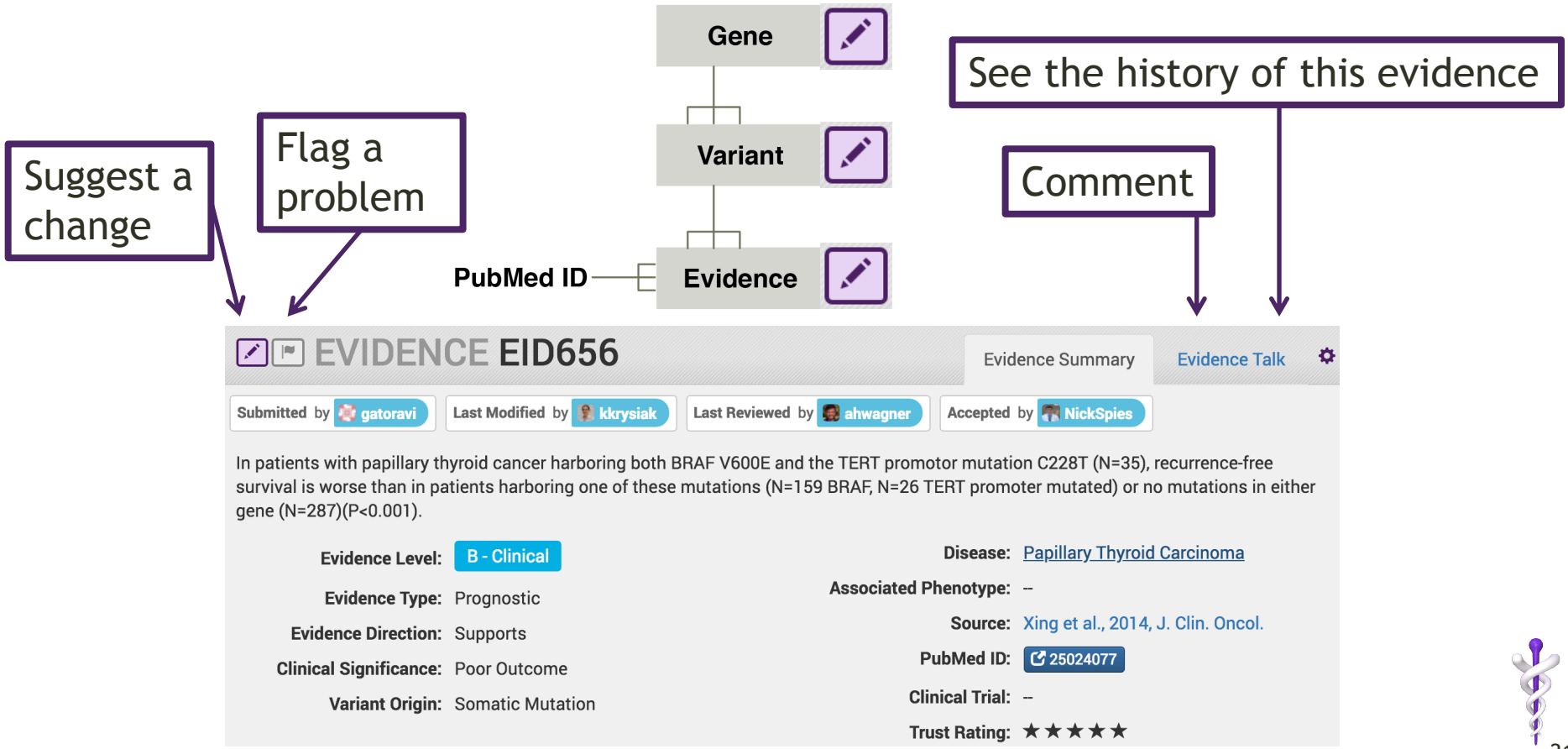
Submitted by gatoravi | Last Modified by kkrysiak | Last Reviewed by ahwagner | Accepted by NickSpies

In patients with papillary thyroid cancer harboring both BRAF V600E and the TERT promotor mutation C228T (N=35), recurrence-free survival is worse than in patients harboring one of these mutations (N=159 BRAF, N=26 TERT promoter mutated) or no mutations in either gene (N=287)(P<0.001).

Evidence Level: B - Clinical
Evidence Type: Prognostic
Evidence Direction: Supports
Clinical Significance: Poor Outcome
Variant Origin: Somatic Mutation

Disease: Papillary Thyroid Carcinoma
Associated Phenotype: --
Source: Xing et al., 2014, J. Clin. Oncol.
PubMed ID: 25024077
Clinical Trial: --
Trust Rating: ★★★★★

Curated evidence items are the foundational unit of CIViC

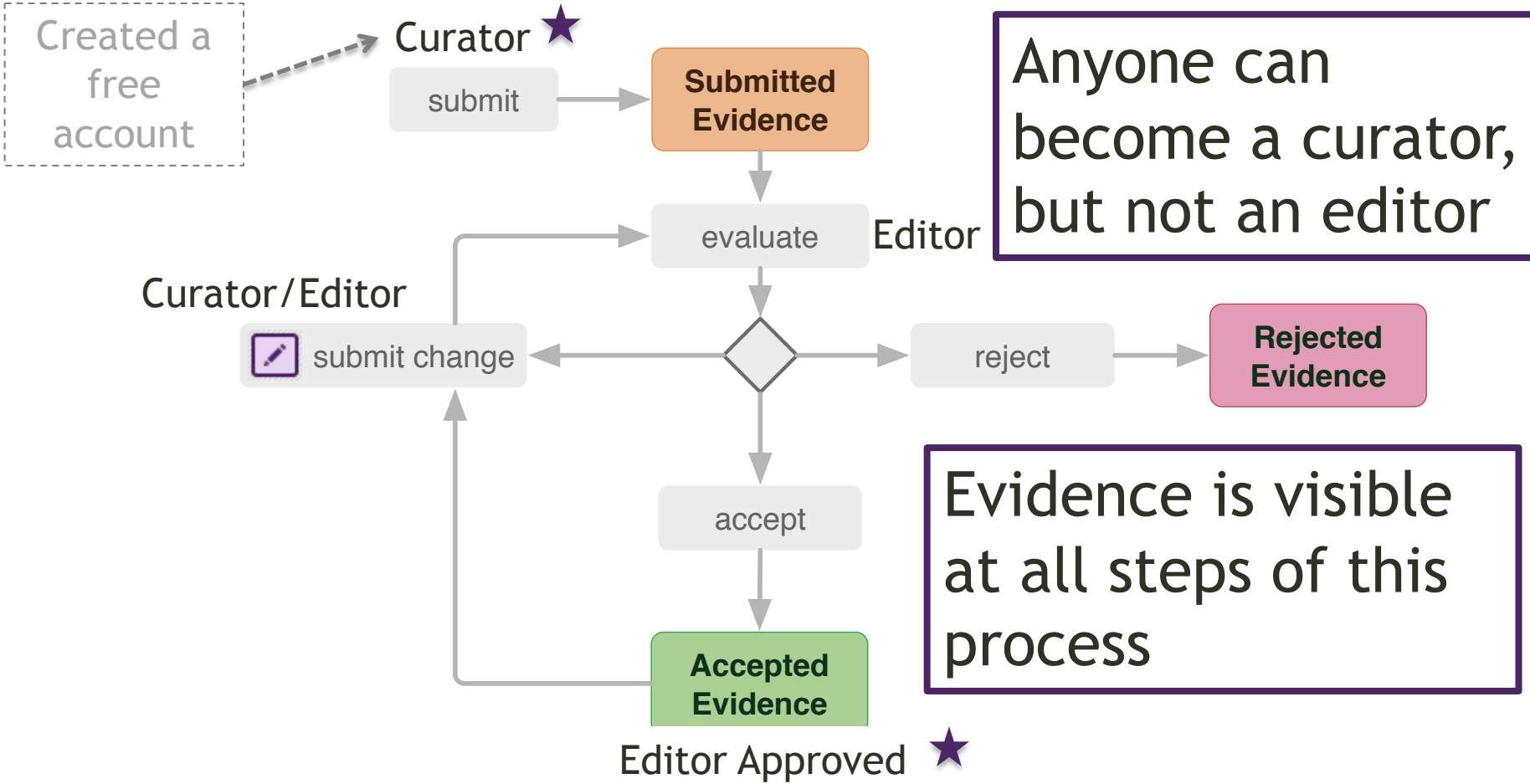


Fundamental requirements of CIViC evidence

- Clinical relevance to cancer
- More than simple observation of a variant in a tumor

- Must not plagiarize the source material
- Published, traceable knowledge (PubMed ID or ASCO abstract)
  - Data available to support assertions
- Must not include personal health information
  - Published peer reviewed case reports are acceptable

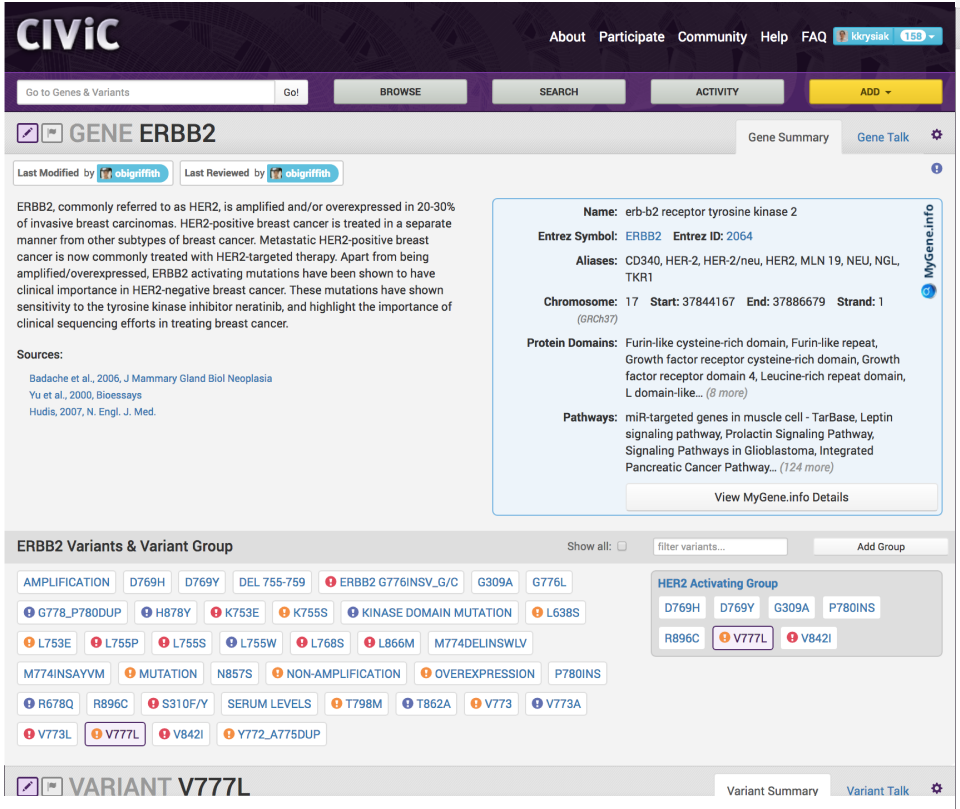
The CIViC evidence item life cycle is designed to promote quality, provenance, transparency and adaptation as knowledge evolves

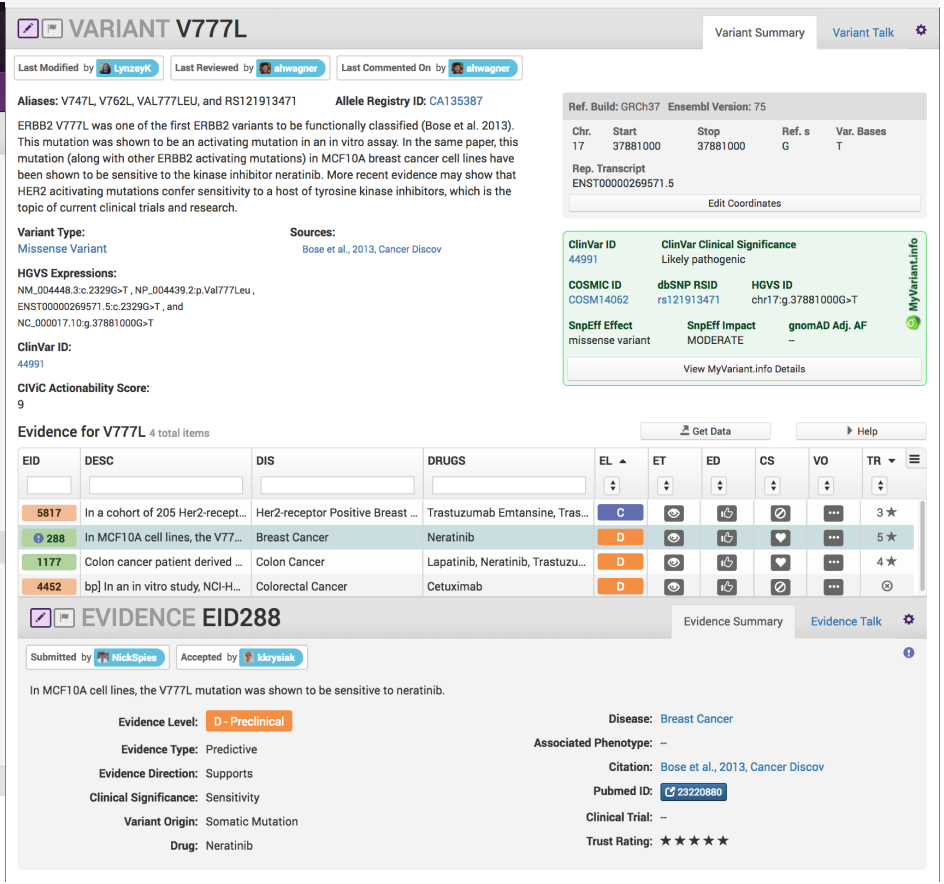


Anyone can become a curator, but not an editor

Evidence is visible at all steps of this process

The goal of curating evidence is to document the clinical relevance of genes and variants, and ultimately to support final assertions





Assertions are built from a collection of evidence items

- Each assertion aggregates multiple CIViC evidence items into a single clinical consensus
  - Specific variant, disease, evidence type, [drug]

- Underlying data is carefully organized and transparent, allowing for rapid updates and reassessment as new information emerges

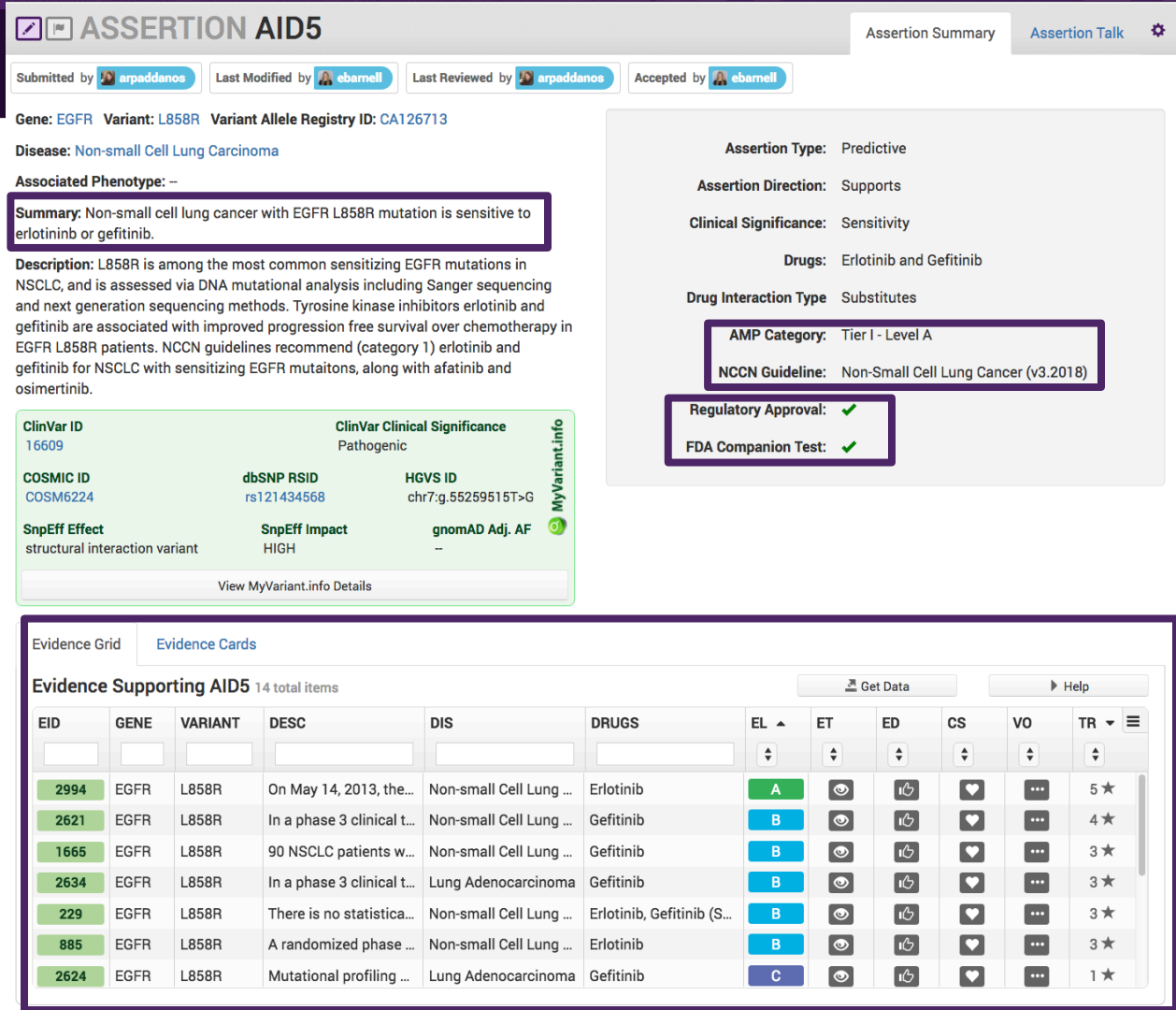
- Once finalized, assertions are submitted to ClinVar

Assertions - example

NSCLC with EGFR L858R mutation is sensitive to erlotinib and gefinitib

- NCCN guidelines
- FDA approval

...



ASSERTION AID5

Assertion Summary | Assertion Talk

Submitted by arpaddanos | Last Modified by ebarnell | Last Reviewed by arpaddanos | Accepted by ebarnell

Gene: EGFR **Variant:** L858R **Variant Allele Registry ID:** CA126713

Disease: Non-small Cell Lung Carcinoma

Associated Phenotype: --

Summary: Non-small cell lung cancer with EGFR L858R mutation is sensitive to erlotininb or gefitinib.

Description: L858R is among the most common sensitizing EGFR mutations in NSCLC, and is assessed via DNA mutational analysis including Sanger sequencing and next generation sequencing methods. Tyrosine kinase inhibitors erlotinib and gefitinib are associated with improved progression free survival over chemotherapy in EGFR L858R patients. NCCN guidelines recommend (category 1) erlotinib and gefitinib for NSCLC with sensitizing EGFR mutaitons, along with afatinib and osimertinib.

ClinVar ID 16609 — **ClinVar Clinical Significance** Pathogenic

COSMIC ID COSM6224 — **dbSNP RSID** rs121434568 — **HGVS ID** chr7:g.55259515T>G

SnpEff Effect structural interaction variant — **SnpEff Impact** HIGH — **gnomAD Adj. AF** --

MyVariant.info

View MyVariant.info Details

Assertion Type: Predictive

Assertion Direction: Supports

Clinical Significance: Sensitivity

Drugs: Erlotinib and Gefitinib

Drug Interaction Type Substitutes

AMP Category: Tier I - Level A

NCCN Guideline: Non-Small Cell Lung Cancer (v3.2018)

Regulatory Approval: ✔

FDA Companion Test: ✔

Evidence Grid | Evidence Cards

Evidence Supporting AID5 14 total items

Get Data | Help

EID	GENE	VARIANT	DESC	DIS	DRUGS	EL	ET	ED	CS	VO	TR
2994	EGFR	L858R	On May 14, 2013, the...	Non-small Cell Lung ...	Erlotinib	A					5★
2621	EGFR	L858R	In a phase 3 clinical t...	Non-small Cell Lung ...	Gefitinib	B					4★
1665	EGFR	L858R	90 NSCLC patients w...	Non-small Cell Lung ...	Gefitinib	B					3★
2634	EGFR	L858R	In a phase 3 clinical t...	Lung Adenocarcinoma	Gefitinib	B					3★
229	EGFR	L858R	There is no statistica...	Non-small Cell Lung ...	Erlotinib, Gefitinib (S...	B					3★
885	EGFR	L858R	A randomized phase ...	Non-small Cell Lung ...	Erlotinib	B					3★
2624	EGFR	L858R	Mutational profiling ...	Lung Adenocarcinoma	Gefitinib	C					1★

Gene and Variant must exist in CIViC

ASSERTION AID5

Assertion Summary | Assertion Talk

Submitted by arpaddanos | Last Modified by ebarnell | Last Reviewed by arpaddanos | Accepted by ebarnell

Gene: EGFR **Variant:** L858R **Variant Allele Registry ID:** CA126713

Disease: Non-small Cell Lung Carcinoma

Associated Phenotype: --

Summary: Non-small cell lung cancer with EGFR L858R mutation is sensitive to erlotininb or gefitinib.

Description: L858R is among the most common sensitizing EGFR mutations in NSCLC, and is assessed via DNA mutational analysis including Sanger sequencing and next generation sequencing methods. Tyrosine kinase inhibitors erlotinib and gefitinib are associated with improved progression free survival over chemotherapy in EGFR L858R patients. NCCN guidelines recommend (category 1) erlotinib and gefitinib for NSCLC with sensitizing EGFR mutaitons, along with afatinib and osimertinib.

ClinVar ID 16609	**ClinVar Clinical Significance** Pathogenic	MyVariant.info
COSMIC ID COSM6224	**dbSNP RSID** rs121434568	**HGVS ID** chr7:g.55259515T>G
SnpEff Effect structural interaction variant	**SnpEff Impact** HIGH	**gnomAD Adj. AF** --

View MyVariant.info Details

Assertion Type:	Predictive
Assertion Direction:	Supports
Clinical Significance:	Sensitivity
Drugs:	Erlotinib and Gefitinib
Drug Interaction Type	Substitutes
AMP Category:	Tier I - Level A
NCCN Guideline:	Non-Small Cell Lung Cancer (v3.2018)
Regulatory Approval:	✔
FDA Companion Test:	✔

Evidence Grid | Evidence Cards

Evidence Supporting AID5 14 total items

Get Data | Help

EID	GENE	VARIANT	DESC	DIS	DRUGS	EL	ET	ED	CS	VO	TR
2994	EGFR	L858R	On May 14, 2013, the...	Non-small Cell Lung ...	Erlotinib	A					5★
2621	EGFR	L858R	In a phase 3 clinical t...	Non-small Cell Lung ...	Gefitinib	B					4★
1665	EGFR	L858R	90 NSCLC patients w...	Non-small Cell Lung ...	Gefitinib	B					3★
2634	EGFR	L858R	In a phase 3 clinical t...	Lung Adenocarcinoma	Gefitinib	B					3★
229	EGFR	L858R	There is no statistica...	Non-small Cell Lung ...	Erlotinib, Gefitinib (S...	B					3★
885	EGFR	L858R	A randomized phase ...	Non-small Cell Lung ...	Erlotinib	B					3★
2624	EGFR	L858R	Mutational profiling ...	Lung Adenocarcinoma	Gefitinib	C					1★

Gene and Variant must exist in CIViC

Evidence must be fully curated

ASSERTION AID5

Assertion Summary | Assertion Talk

Submitted by arpaddanos | Last Modified by ebarnell | Last Reviewed by arpaddanos | Accepted by ebarnell

Gene: EGFR **Variant:** L858R **Variant Allele Registry ID:** CA126713

Disease: Non-small Cell Lung Carcinoma

Associated Phenotype: --

Summary: Non-small cell lung cancer with EGFR L858R mutation is sensitive to erlotininb or gefitinib.

Description: L858R is among the most common sensitizing EGFR mutations in NSCLC, and is assessed via DNA mutational analysis including Sanger sequencing and next generation sequencing methods. Tyrosine kinase inhibitors erlotinib and gefitinib are associated with improved progression free survival over chemotherapy in EGFR L858R patients. NCCN guidelines recommend (category 1) erlotinib and gefitinib for NSCLC with sensitizing EGFR mutaitons, along with afatinib and osimertinib.

Assertion Type:	Predictive
Assertion Direction:	Supports
Clinical Significance:	Sensitivity
Drugs:	Erlotinib and Gefitinib
Drug Interaction Type	Substitutes
AMP Category:	Tier I - Level A
NCCN Guideline:	Non-Small Cell Lung Cancer (v3.2018)
Regulatory Approval:	✔
FDA Companion Test:	✔

MyVariant.info

ClinVar ID	ClinVar Clinical Significance	
16609	Pathogenic	
COSMIC ID	**dbSNP RSID**	**HGVS ID**
COSM6224	rs121434568	chr7:g.55259515T>G
SnpEff Effect	**SnpEff Impact**	**gnomAD Adj. AF**
structural interaction variant	HIGH	--

View MyVariant.info Details

Evidence Grid | Evidence Cards

Evidence Supporting AID5 14 total items

Get Data | Help

EID	GENE	VARIANT	DESC	DIS	DRUGS	EL	ET	ED	CS	VO	TR
2994	EGFR	L858R	On May 14, 2013, the...	Non-small Cell Lung ...	Erlotinib	A					5★
2621	EGFR	L858R	In a phase 3 clinical t...	Non-small Cell Lung ...	Gefitinib	B					4★
1665	EGFR	L858R	90 NSCLC patients w...	Non-small Cell Lung ...	Gefitinib	B					3★
2634	EGFR	L858R	In a phase 3 clinical t...	Lung Adenocarcinoma	Gefitinib	B					3★
229	EGFR	L858R	There is no statistica...	Non-small Cell Lung ...	Erlotinib, Gefitinib (S...	B					3★
885	EGFR	L858R	A randomized phase ...	Non-small Cell Lung ...	Erlotinib	B					3★
2624	EGFR	L858R	Mutational profiling ...	Lung Adenocarcinoma	Gefitinib	C					1★

RULES

Gene and Variant must exist in CIViC

Critical rationale / decisions are described

Evidence must be fully curated

ASSERTION AID5

Assertion Summary | Assertion Talk

Submitted by arpaddanos | Last Modified by ebarnell | Last Reviewed by arpaddanos | Accepted by ebarnell

Gene: EGFR **Variant:** L858R **Variant Allele Registry ID:** CA126713

Disease: Non-small Cell Lung Carcinoma

Associated Phenotype: --

Summary: Non-small cell lung cancer with EGFR L858R mutation is sensitive to erlotininb or gefitinib.

Description: L858R is among the most common sensitizing EGFR mutations in NSCLC, and is assessed via DNA mutational analysis including Sanger sequencing and next generation sequencing methods. Tyrosine kinase inhibitors erlotinib and gefitinib are associated with improved progression free survival over chemotherapy in EGFR L858R patients. NCCN guidelines recommend (category 1) erlotinib and gefitinib for NSCLC with sensitizing EGFR mutaitons, along with afatinib and osimertinib.

ClinVar ID 16609	**ClinVar Clinical Significance** Pathogenic	MyVariant.info
COSMIC ID COSM6224	**dbSNP RSID** rs121434568	**HGVS ID** chr7:g.55259515T>G
SnpEff Effect structural interaction variant	**SnpEff Impact** HIGH	**gnomAD Adj. AF** --

View MyVariant.info Details

Assertion Type: Predictive

Assertion Direction: Supports

Clinical Significance: Sensitivity

Drugs: Erlotinib and Gefitinib

Drug Interaction Type Substitutes

AMP Category: Tier I - Level A

NCCN Guideline: Non-Small Cell Lung Cancer (v3.2018)

Regulatory Approval: ✔

FDA Companion Test: ✔

Evidence Grid | Evidence Cards

Evidence Supporting AID5 14 total items

Get Data | Help

EID	GENE	VARIANT	DESC	DIS	DRUGS	EL	ET	ED	CS	VO	TR
2994	EGFR	L858R	On May 14, 2013, the...	Non-small Cell Lung ...	Erlotinib	A					5★
2621	EGFR	L858R	In a phase 3 clinical t...	Non-small Cell Lung ...	Gefitinib	B					4★
1665	EGFR	L858R	90 NSCLC patients w...	Non-small Cell Lung ...	Gefitinib	B					3★
2634	EGFR	L858R	In a phase 3 clinical t...	Lung Adenocarcinoma	Gefitinib	B					3★
229	EGFR	L858R	There is no statistica...	Non-small Cell Lung ...	Erlotinib, Gefitinib (S...	B					3★
885	EGFR	L858R	A randomized phase ...	Non-small Cell Lung ...	Erlotinib	B					3★
2624	EGFR	L858R	Mutational profiling ...	Lung Adenocarcinoma	Gefitinib	C					1★

Gene and Variant must exist in CIViC

Critical rationale / decisions are described

Evidence must be fully curated

ASSERTION AID5

Assertion Summary | Assertion Talk

Submitted by arpaddanos | Last Modified by ebarnell | Last Reviewed by arpaddanos | Accepted by ebarnell

Gene: EGFR **Variant:** L858R **Variant Allele Registry ID:** CA126713

Disease: Non-small Cell Lung Carcinoma

Associated Phenotype: --

Summary: Non-small cell lung cancer with EGFR L858R mutation is sensitive to erlotininb or gefitinib.

Description: L858R is among the most common sensitizing EGFR mutations in NSCLC, and is assessed via DNA mutational analysis including Sanger sequencing and next generation sequencing methods. Tyrosine kinase inhibitors erlotinib and gefitinib are associated with improved progression free survival over chemotherapy in EGFR L858R patients. NCCN guidelines recommend (category 1) erlotinib and gefitinib for NSCLC with sensitizing EGFR mutaitons, along with afatinib and osimertinib.

ClinVar ID 16609	**ClinVar Clinical Significance** Pathogenic	MyVariant.info
COSMIC ID COSM6224	**dbSNP RSID** rs121434568	**HGVS ID** chr7:g.55259515T>G
SnpEff Effect structural interaction variant	**SnpEff Impact** HIGH	**gnomAD Adj. AF** --

View MyVariant.info Details

Assertion Type:	Predictive
Assertion Direction:	Supports
Clinical Significance:	Sensitivity
Drugs:	Erlotinib and Gefitinib
Drug Interaction Type	Substitutes
AMP Category:	Tier I - Level A
NCCN Guideline:	Non-Small Cell Lung Cancer (v3.2018)
Regulatory Approval:	✔
FDA Companion Test:	✔

At least 1 Evidence item with 3+ stars

Evidence Grid | Evidence Cards

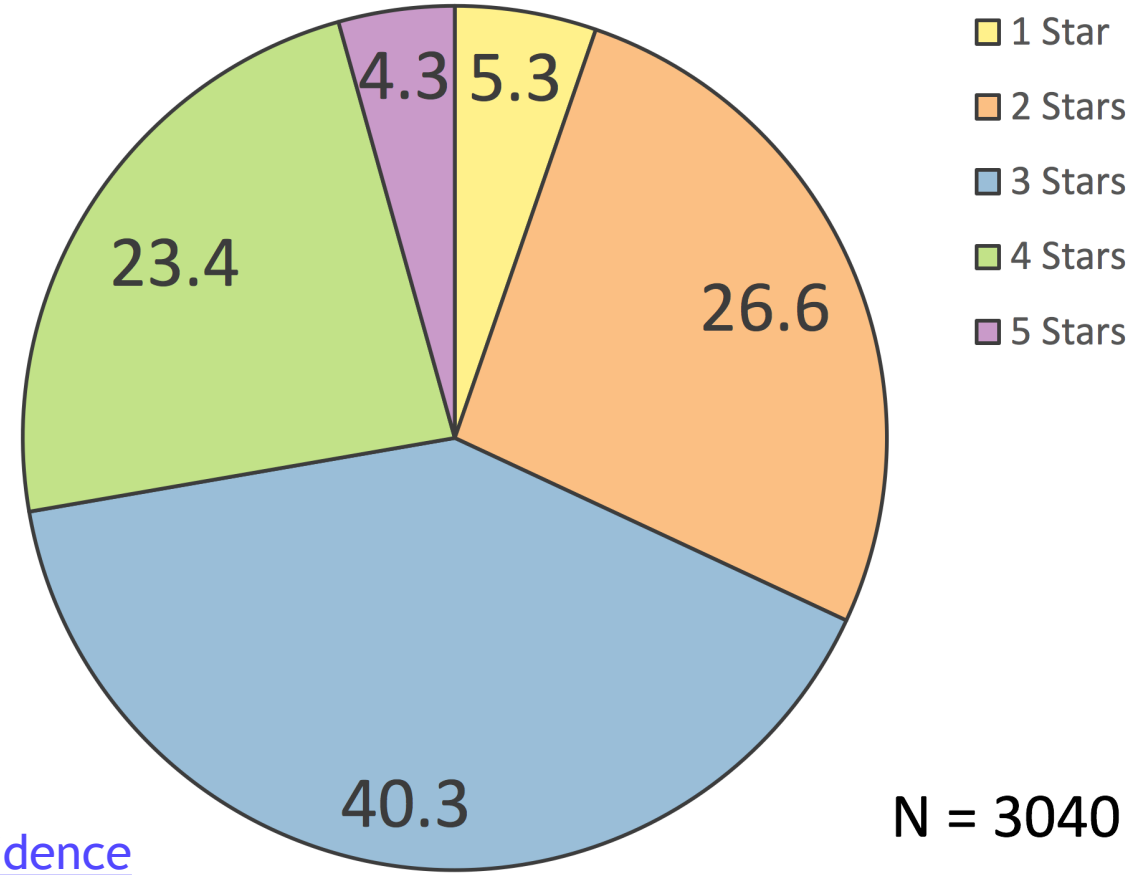
Evidence Supporting AID5 14 total items

Get Data | Help

EID	GENE	VARIANT	DESC	DIS	DRUGS	EL	ET	ED	CS	VO	TR
2994	EGFR	L858R	On May 14, 2013, the...	Non-small Cell Lung ...	Erlotinib	A					5★
2621	EGFR	L858R	In a phase 3 clinical t...	Non-small Cell Lung ...	Gefitinib	B					4★
1665	EGFR	L858R	90 NSCLC patients w...	Non-small Cell Lung ...	Gefitinib	B					3★
2634	EGFR	L858R	In a phase 3 clinical t...	Lung Adenocarcinoma	Gefitinib	B					3★
229	EGFR	L858R	There is no statistica...	Non-small Cell Lung ...	Erlotinib, Gefitinib (S...	B					3★
885	EGFR	L858R	A randomized phase ...	Non-small Cell Lung ...	Erlotinib	B					3★
2624	EGFR	L858R	Mutational profiling ...	Lung Adenocarcinoma	Gefitinib	C					1★

Star ratings help guide Assertion creation

- Moderated evidence is appropriately distributed

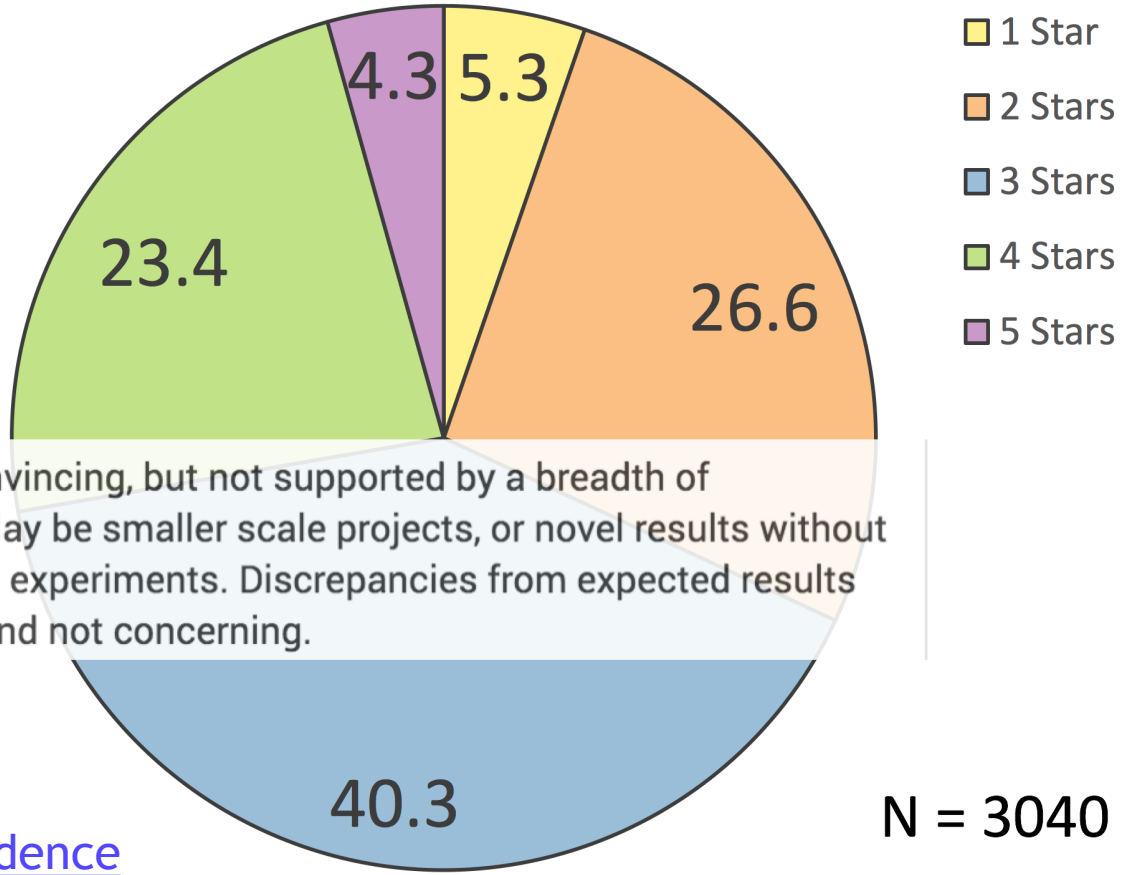


N = 3040

https://civicdb.org/statistics/evidence

Star ratings help guide Assertion creation

- Moderated evidence is appropriately distributed



★★★☆☆ Evidence is convincing, but not supported by a breadth of experiments. May be smaller scale projects, or novel results without many follow-up experiments. Discrepancies from expected results are explained and not concerning.

N = 3040

https://civicdb.org/statistics/evidence

ASSERTION AID5

Assertion Summary | Assertion Talk

Submitted by arpaddanos | Last Modified by ebarnell | Last Reviewed by arpaddanos | Accepted by ebarnell

Gene: EGFR **Variant:** L858R **Variant Allele Registry ID:** CA126713

Disease: Non-small Cell Lung Carcinoma

Associated Phenotype: --

Summary: Non-small cell lung cancer with EGFR L858R mutation is sensitive to

Assertion Type: Predictive

Assertion Direction: Supports

Tier I: Variants of Strong Clinical Significance

Therapeutic, prognostic & diagnostic

Level A Evidence

FDA-approved therapy

Included in professional guidelines

Level B Evidence

Well-powered studies with consensus from experts in the field

Tier II: Variants of Potential Clinical Significance

Therapeutic, prognostic & diagnostic

Level C Evidence

FDA-approved therapies for different tumor types or investigational therapies

Multiple small published studies with some consensus

Level D Evidence

Preclinical trials or a few case reports without consensus

Tier III: Variants of Unknown Clinical Significance

Not observed at a significant allele frequency in the general or specific subpopulation databases, or pan-cancer or tumor-specific variant databases

No convincing published evidence of cancer association

Tier IV: Benign or Likely Benign Variants

Observed at significant allele frequency in the general or specific subpopulation databases

No existing published evidence of cancer association

8)

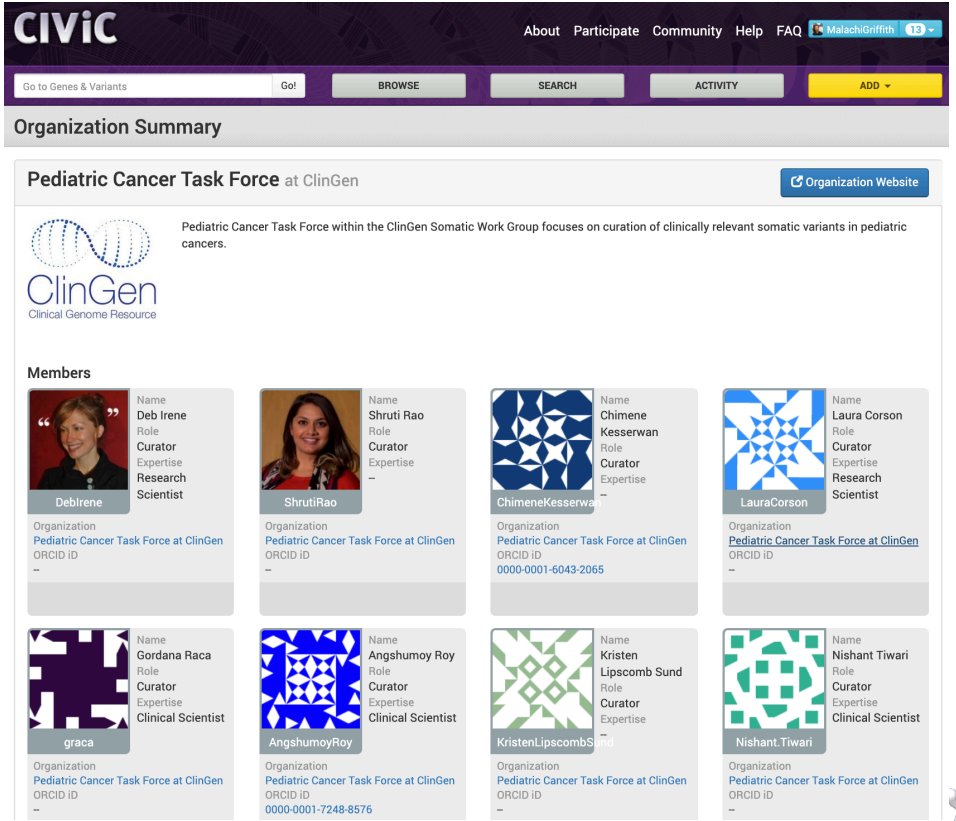
elp

TR

5★
4★
3★
3★
3★
3★
1★

How do we scale up without compromising quality (ClinGen Somatic)?

- Collaboration with the ClinGen Somatic Working Group is critical to achieving broad engagement by domain experts
  - Subha Madhavan and Shruti Rao
- Focused working groups forming (disease, gene, variant type, etc.)
- >40 members
- >250 evidence items submitted



Creation of standards, SOPs, and training materials is needed

Correspondence | Open Access | Published: 29 November 2019

Standard operating procedure for curation and clinical interpretation of variants in cancer

Arpad M. Danos, Kilannin Krysiak, Erica K. Barnell, Adam C. Coffman, Joshua F. McMichael, Susanna Kiwala, Nicholas C. Spies, Lana M. Sheta, Shahil P. Pema, Lynzey Kujan, Kaitlin A. Clark, Amber Z. Wollam, Shruti Rao, Deborah I. Ritter, Dmitriy Sonkin, Gordana Raca, Wan-Hsin Lin, Cameron J. Grisdale, Raymond H. Kim, Alex H. Wagner, Subha Madhavan, Malachi Griffith & Obi L. Griffith

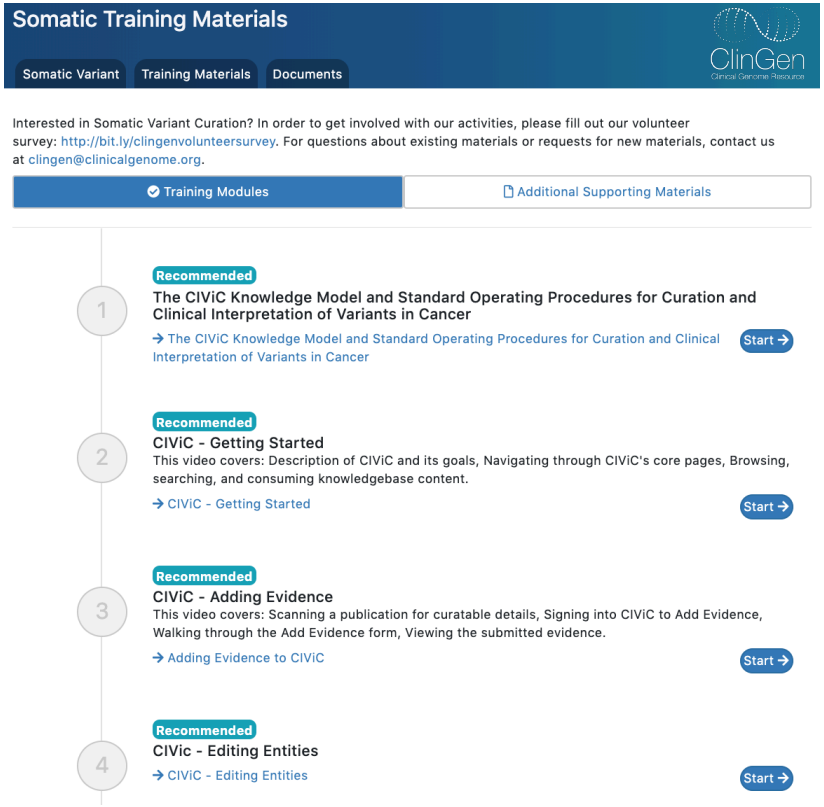
Genome Medicine **11**, Article number: 76 (2019) | Cite this article

380 Accesses | **4** Altmetric | Metrics

Abstract

Manually curated variant knowledgebases and their associated knowledge models are serving an increasingly important role in distributing and interpreting variants in cancer. These knowledgebases vary in their level of public accessibility,

https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-019-0687-x

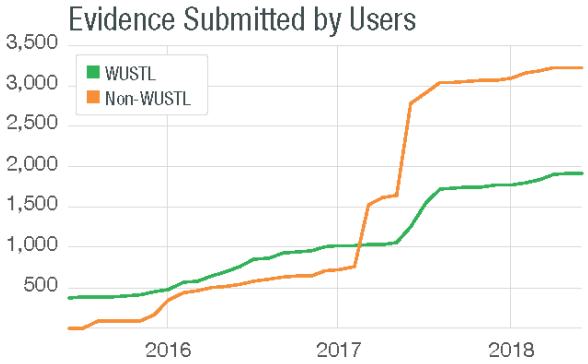


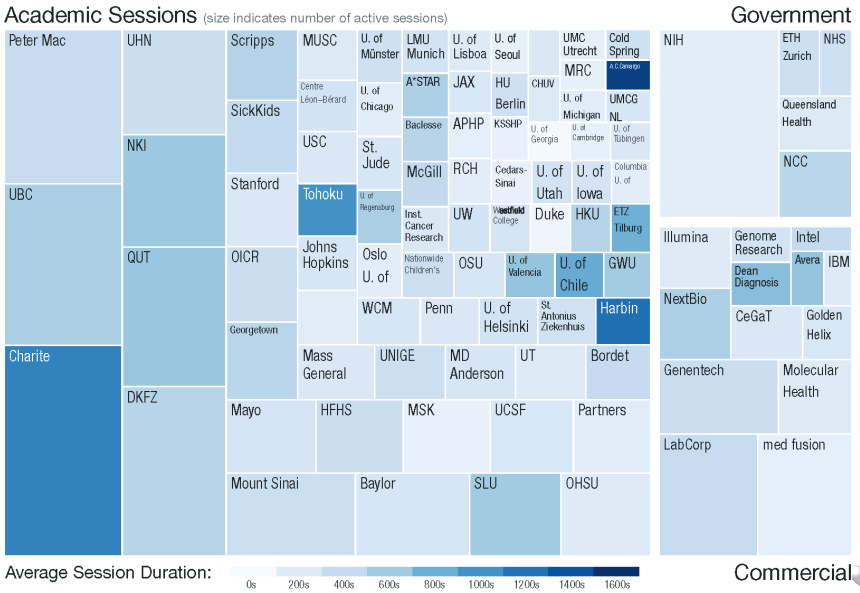
https://clinicalgenome.org/curation-activities/somatic/training-materials/

How do we measure success?

- 6,468 evidence lines curated for 2,357 variants, 402 genes, 274 cancer types, 2,374 papers. 24 assertions to date.
- >3,000 users per month
- 191 contributors to date

- >750,000 API requests per month





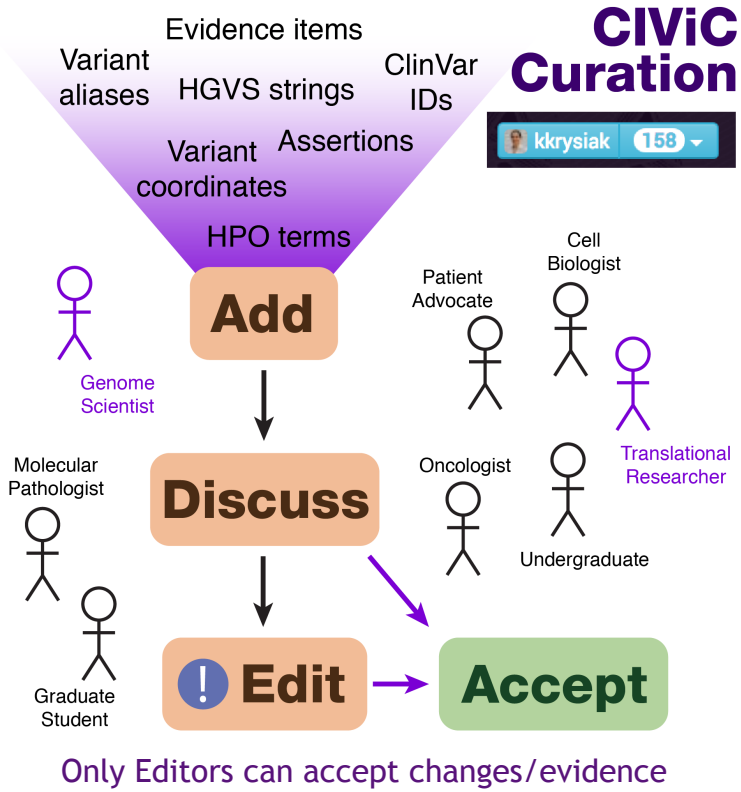
Open model promotes adoption by commercial and academic data clients

Genome Browser

Alissa Clinical Informatics Platform
Alissa Interpret

CIViC has >35 known data clients (https://civicdb.org/about#data-clients)

CIViC's open access and crowd-sourcing model has advantages and disadvantages



- **Advantages**
- Scalable
- Promotes transparent consensus building (with provenance)
- Diverse community (remote, asynchronous, global contributions)
- **Disadvantages**
- Quality (perceived and real)
- Difficult to focus curation effort
- Funding

Problem: CIViC is a Silo in a Diverse Knowledge Ecosystem

@handlerwagner

Variant interpretation knowledge remains siloed

Established Interpretation Knowledgebases (somatic)

- **CIViC (WashU)**
- **OncoKB (MSKCC)**
- **JAX-Clinical Knowledgebase (Jackson lab)**
- **Cancer Genome Interpreter (Barcelona)**
- **MyCancerGenome (Vanderbilt)**
- PMKB (Cornell)
- KnowledgeBase for Precision Oncology (MD Anderson)
- CanDL (Ohio State)
- COSMIC (Sanger)

Germline

- PharmGKB
- ClinVar
- ClinGen Evidence Repository

Additionally...

- Many ad hoc “databases” (acadamic centers / hospitals)
- Many industry examples

Collaborative standards and interoperability work is needed: Variant Interpretation for Cancer Consortium (VICC) - GA4GH Driver Project

Goals/Principles:

- Global integration of clinical cancer variant interpretation
- Standards and guidelines
- Open content for sharing

- Facilitate cross-knowledgebase queries

search.cancervariants.org

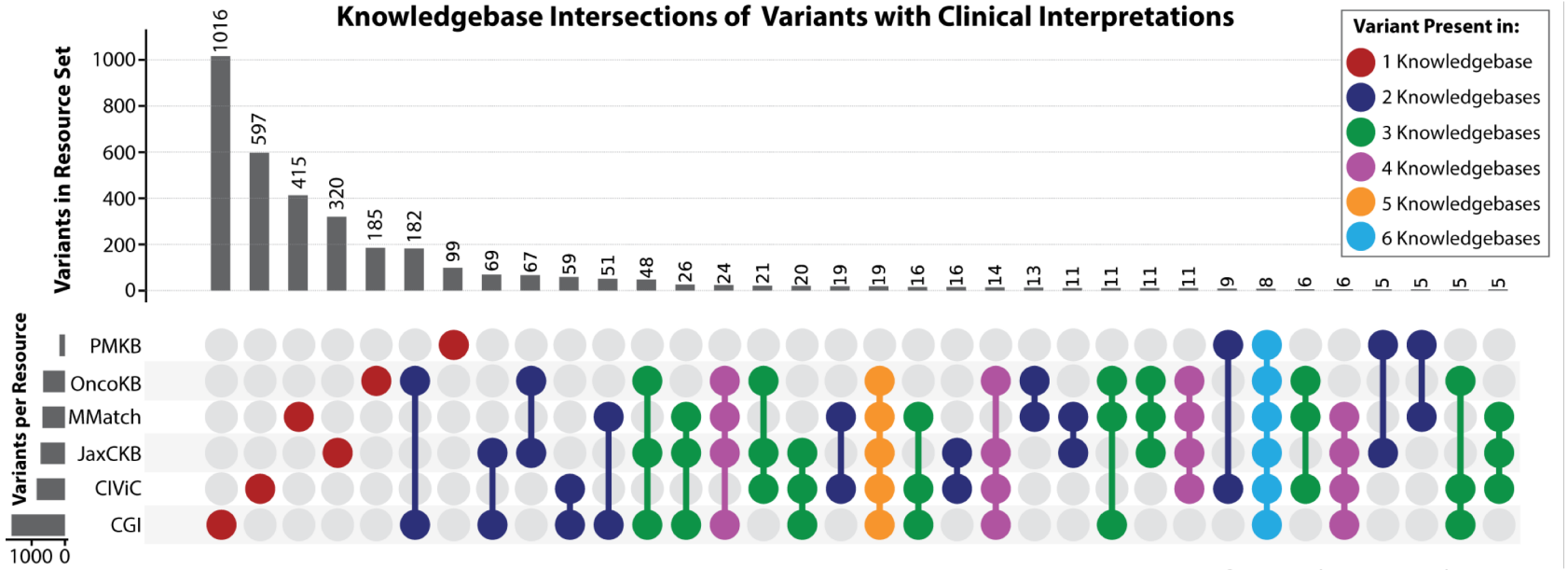
cancervariants.org

[ga4gh.org](ga4gh.org)

Variant Interpretations are Heterogeneous

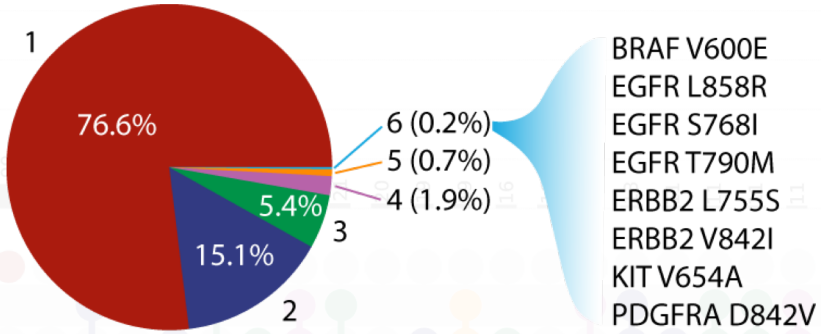
@handlerwagner

Variant Representation Across Knowledgebases



Wagner AH, et al. *Nat. Genet.* (accepted) 2019

Variant Representation Across Knowledgebases



Wagner AH, et al. *Nat. Genet.* (accepted) 2019

Poor overlap of variants creates
inconsistent clinical interpretations

This can be addressed by
aggregating curated knowledge

Variant Interpretations are Structurally Disparate

@handlerwagner

Diversity in Structure of Variant Interpretations

	OncoKB	CIViC	PMKB
Gene	BRAF	BRAF (Entrez ID: 673)	BRAF
Isoform	ENST00000288602 / NM_004333.4	ENST00000288602.6	ENST00000288602
Variant	V600E (????)	V600E (chr7:g.140453136A>T)	V600E (7:140453136-140453136)
Disease	Melanoma	Skin Melanoma (DOID:8923)	Tumor: Melanoma / Tissue: Skin
Drug	Dabrafenib	Dabrafenib + Trametinib	?
Clinical Significance	Known Effect: Sensitive	Supports Sensitivity	?
Evidence Level	2B	A - Validated	Tier 1
Statement	Approved Indications: Dabrafenib is FDA-approved for BRAF V600E mutant unresectable or metastatic melanoma	Open-label, randomized phase 3 trial with 704 patients with metastatic melanoma with a BRAF V600 mutation. Patients were randomized ...	… Various B-Raf inhibitors (Vemurafenib, Dabrafenib) have been FDA approved for melanoma therapy in certain settings.

@handlerwagner

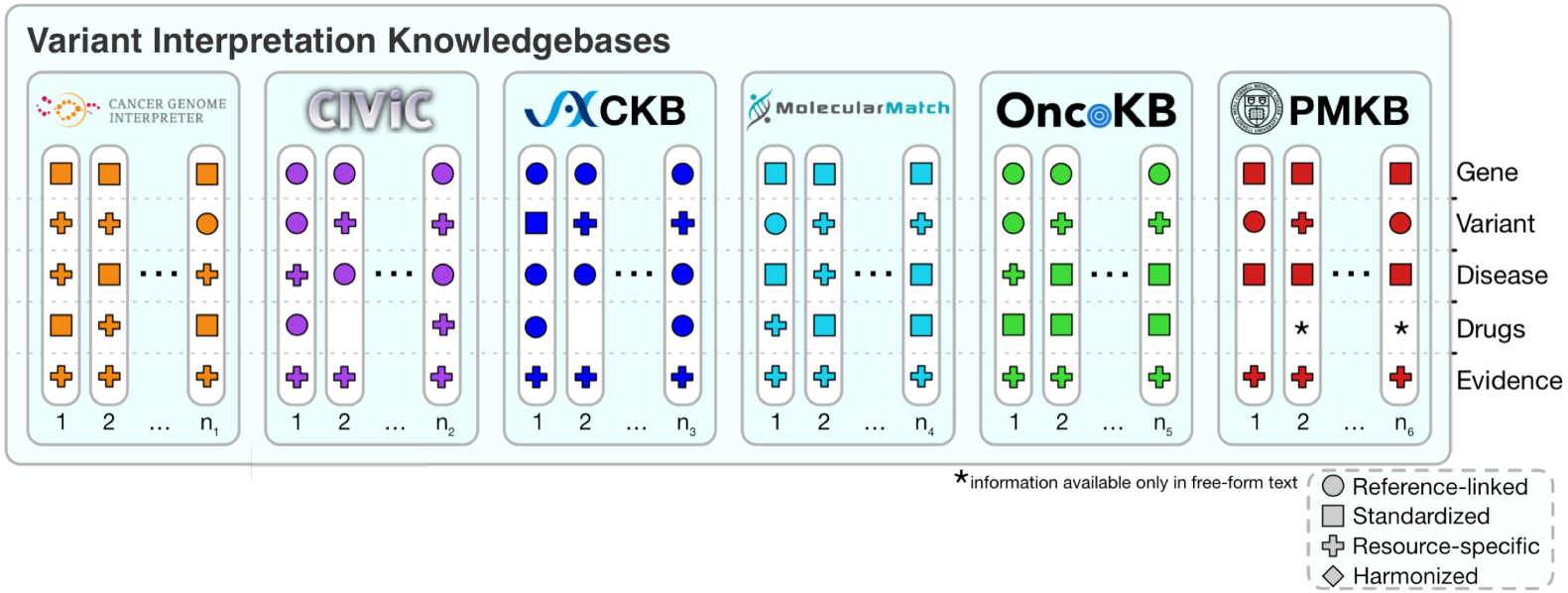
Poor overlap of variants creates
inconsistent clinical interpretations

We can aggregate, but...

Clinical variant interpretation
knowledgebases are not interoperable

This can be addressed by harmonizing interpretations

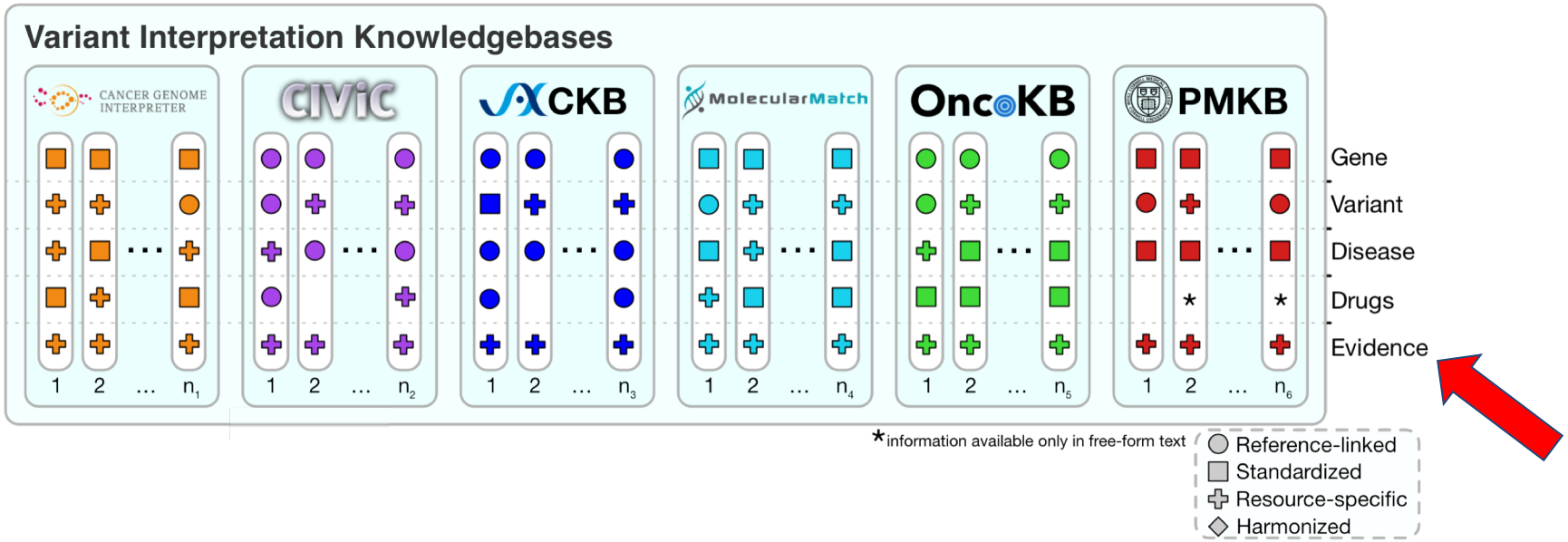
Structure of a Variant Interpretation



Wagner AH, et al. *Nat. Genet.* (accepted) 2019

@handlerwagner

Structure of a Variant Interpretation



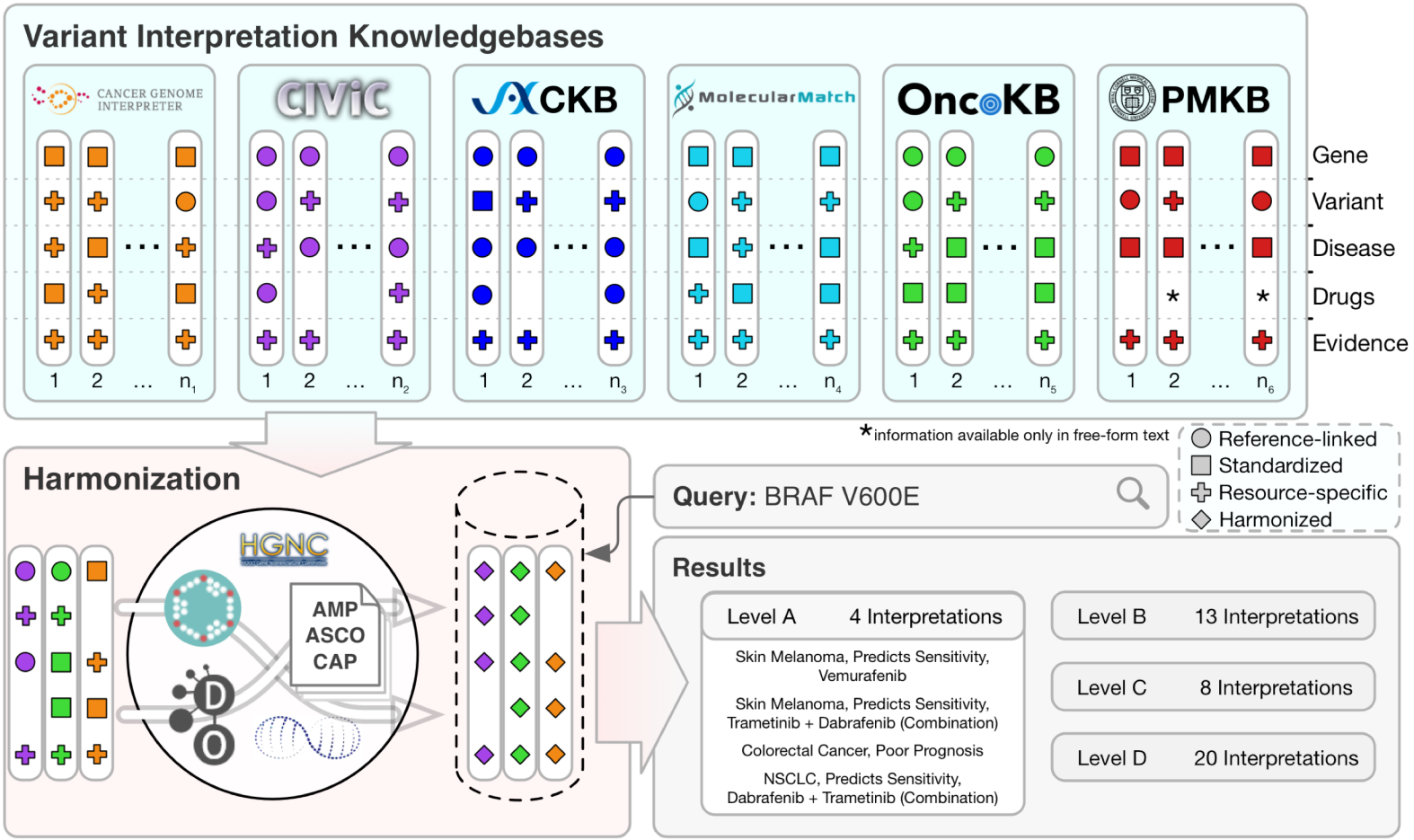
Wagner AH, et al. *Nat. Genet.* (accepted) 2019

@handlerwagner

AMP/ASCO/CAP Variant Evidence Guidelines							
Evidence Level	**Defining Characteristics**	**CIViC**	**OncoKB**	**JAX-CKB**	**CGI**	**MMatch**	**PMKB**
Level A (Tier I)	*Evidence from professional guidelines or FDA-approved therapies relating to a biomarker and disease.*	Level A	Level 1 / 2A /R1	Guideline / FDA Approved	Clinical Practice	Level 1A	Tier 1
Level B (Tier I)	*Evidence from clinical trials or other well-powered studies in clinical populations, with expert consensus.*	Level B	Level 3A	Phase III	Clinical Trials III-IV	Level 1B	
Level C (Tier II)	*Evidence for therapeutic predictive markers from case studies, or other biomarkers from several small studies. Also evidence for biomarker therapeutic predictions for established drugs for different indications.*	Predictive Level C	Level 2B, Level 3B	Clinical Study/ Phase I / Phase II	Clinical Trials I-II, Case Reports	Level 2C	Tier 2
Level D (Tier II)	*Preclinical findings or case studies of prognostic or diagnostic biomarkers. Also includes indirect findings.*	Non-predictive Level C / Level D / Level E	Level 4	Phase 0, Pre-clinical	Pre-clinical Data	Level 2D	

Wagner AH, et al. *Nat. Genet.* (accepted) 2019

@handlerwagner



Wagner AH, et al. *Nat. Genet.* (accepted) 2019

@handlerwagner

Harmonizing Variants: An **Expressive** and **Computable** specification for Variation Representation (VR-spec / VR)

Reece Hart

@handlerwagner

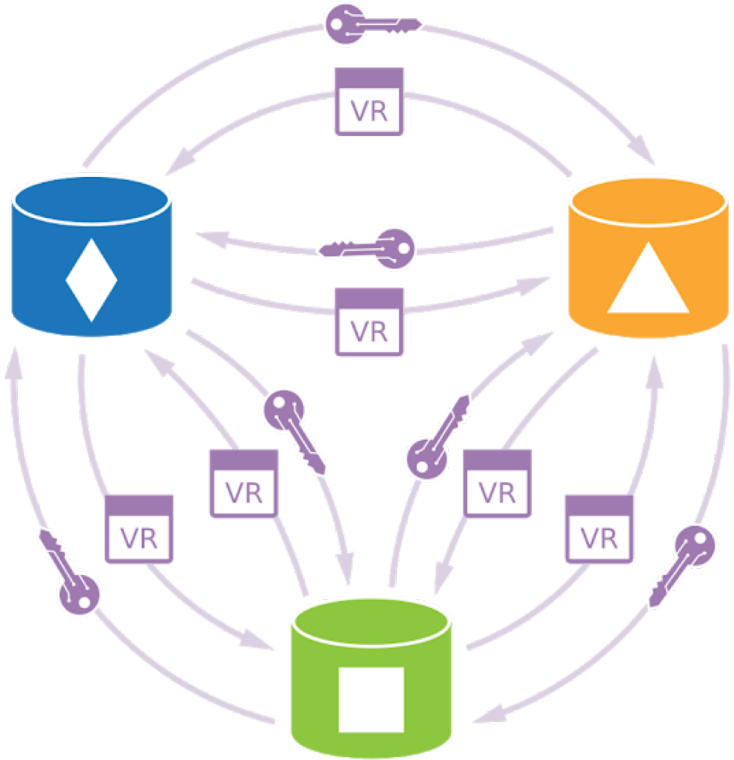
Overview

Global Alliance for Genomics & Health

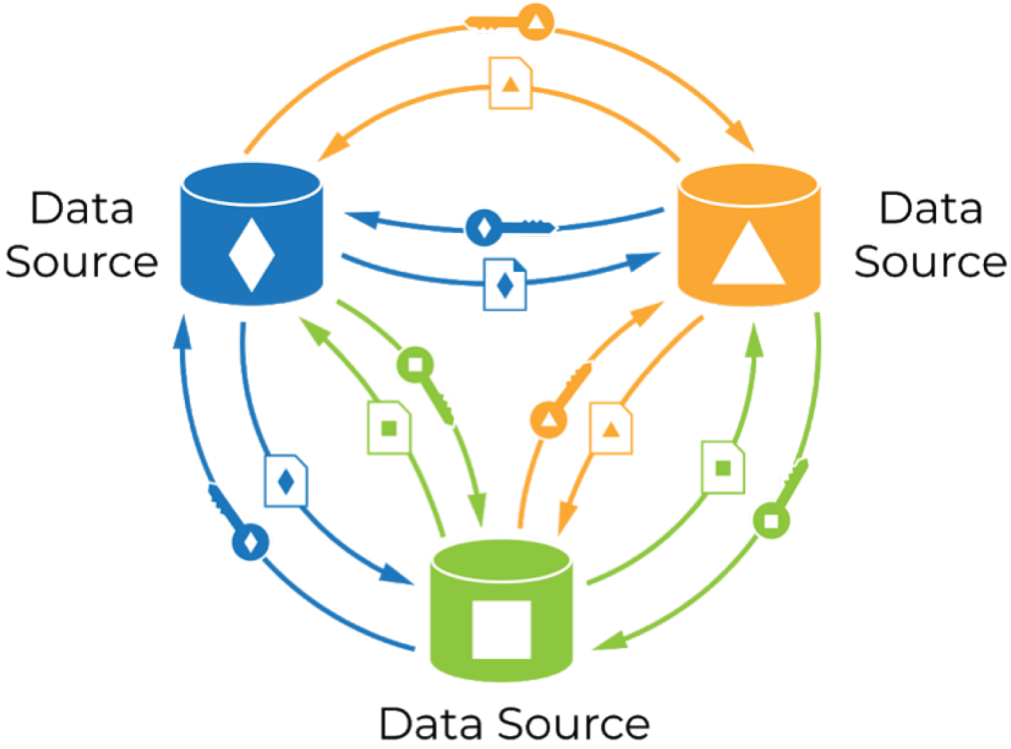
VR aims to standardize the definition, sharing, and identification of biological variation within and between systems.

VR 1.0 Aims:

- An extensible schema for variation
- Minimization of representational ambiguity
- Shared identifiers for variation
- Example Implementation



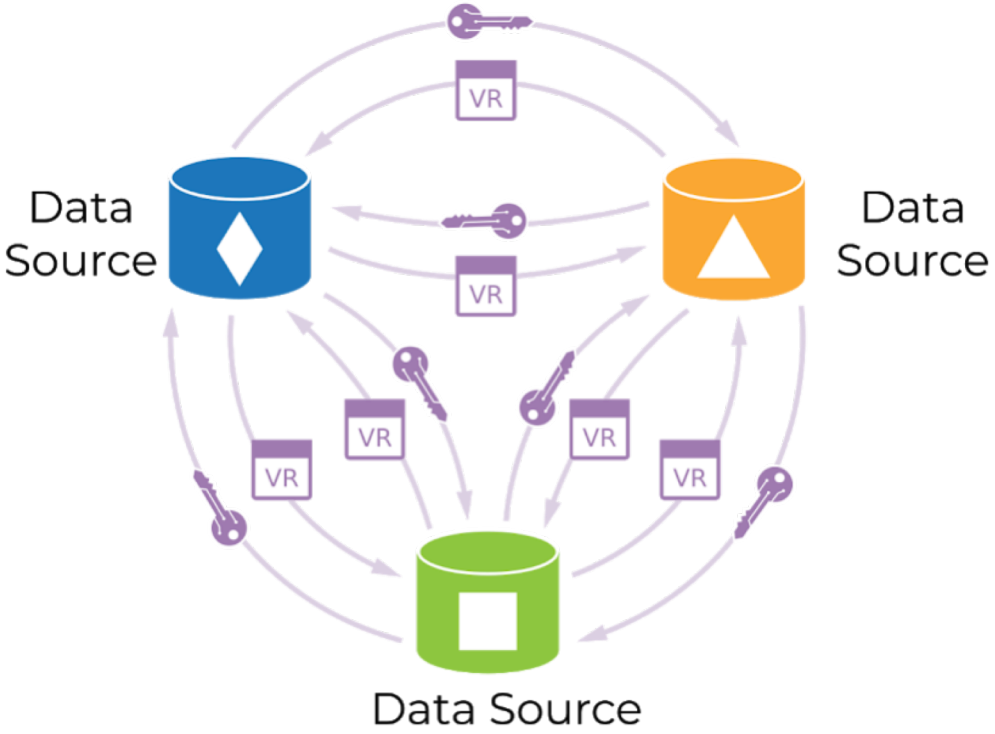
CURRENTLY...



PAIRS OF SYSTEMS COORDINATE KEYS AND FORMATS IN ORDER TO SHARE VARIATION DATA. ADDING A NEW SYSTEM IS DIFFICULT.

Rendering by Stephanie Li

WITH THE VR SPECIFICATION...



SYSTEMS USE A COMMON IDENTIFIER, COMPUTED FROM THE DATA ITSELF, AND A COMMON DATA FORMAT. ADDING A NEW SYSTEM IS MUCH EASIER.

Rendering by Stephanie Li

Presentation ≠ Representation

NM_080588.2:c.139_140insC
ENST00000367279:c.139_140insC
NM_080588.2:c.139dup
ENST00000367279:c.139dup


```
{
  "sequence_id": "ga4gh:SQ.0a1b2c3d",
  "location": (139,139),
  "state": "C"
}
```

Multiple human names

due to choice of accession, normalization, ins/dup, etc.

Structured representations
Clear conventions

= VR

One conceptual variant

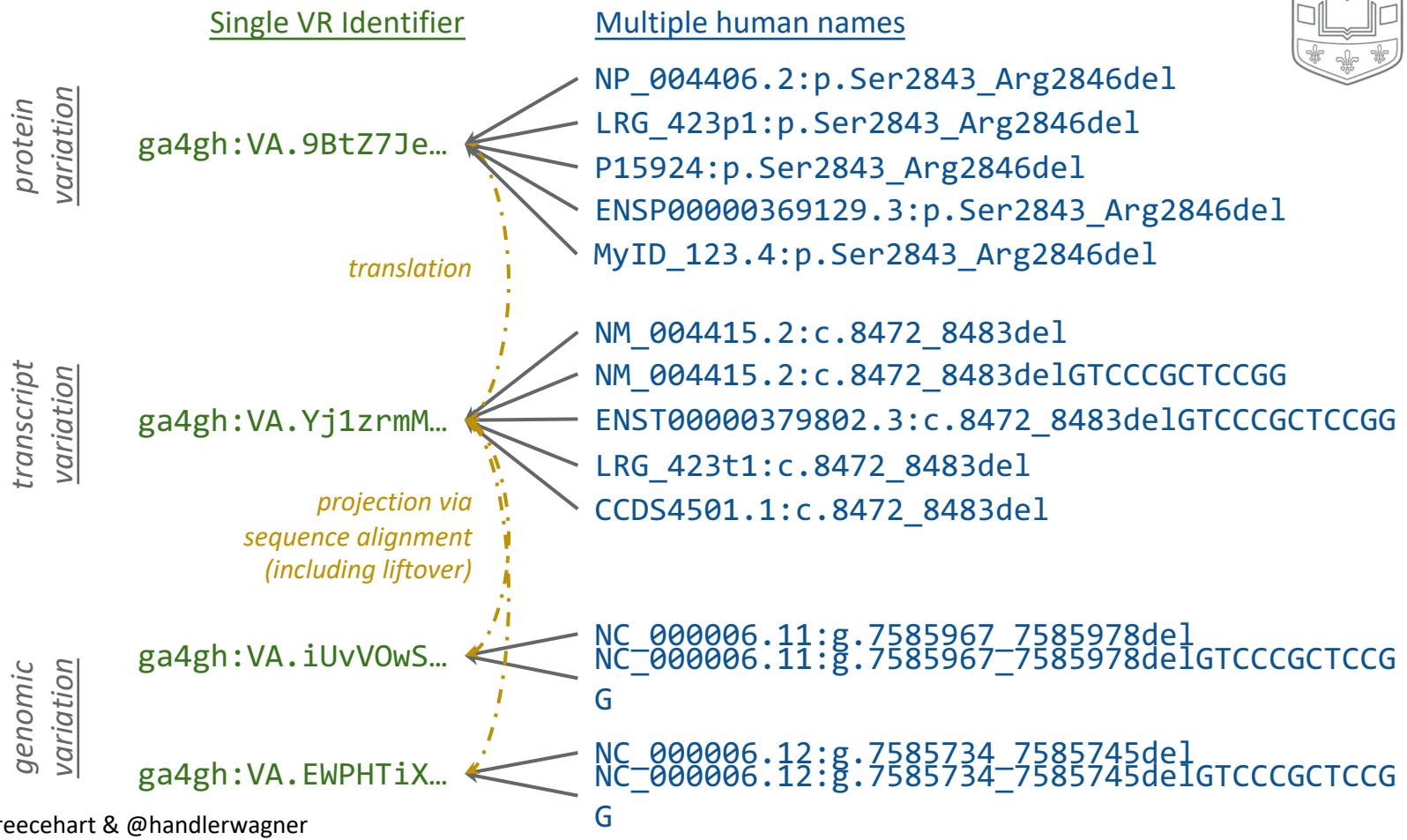
The mission of the VR is to

standardize the representation of all classes of biological variation

to enable

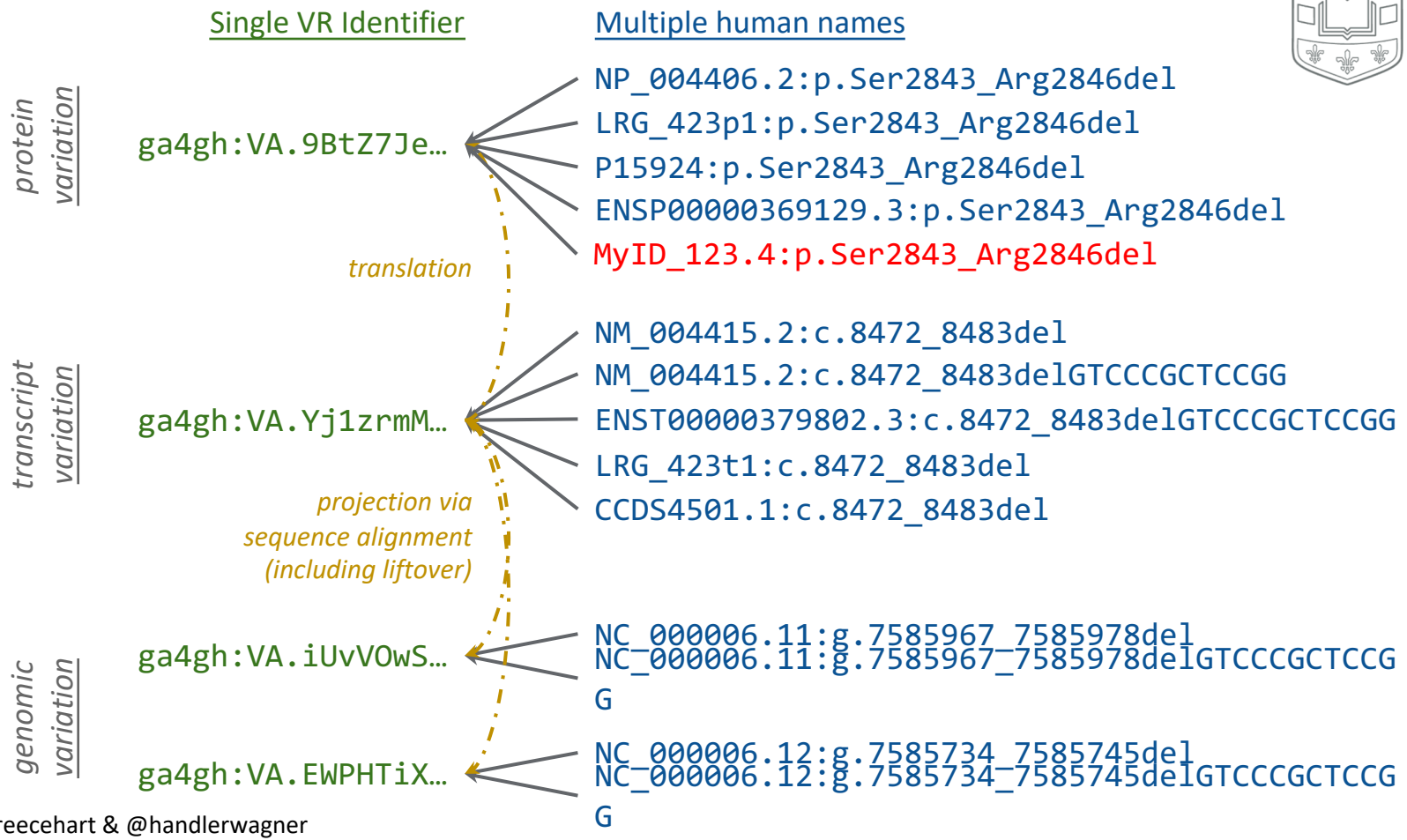
accurate curation and reliable computed interpretation.

VR identifiers: precise, unique names for variation

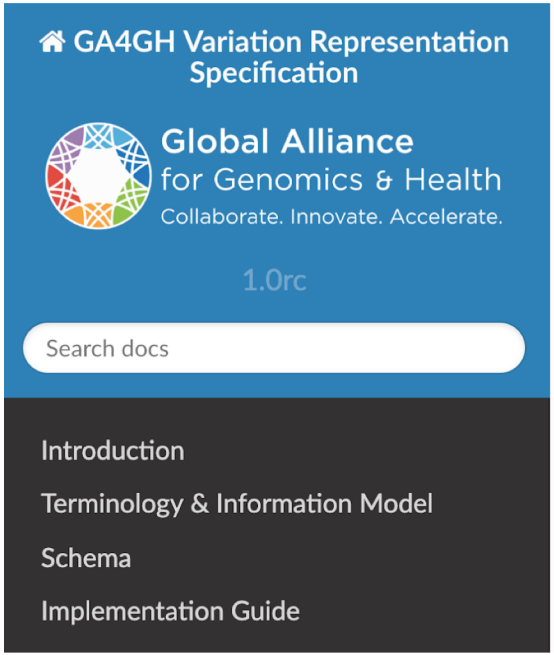


@reecehart & @handlerwagner

VR identifiers: precise, unique names for variation



@reecehart & @handlerwagner



Docs » GA4GH Variation Representation Specification

Edit on GitHub

GA4GH Variation Representation Specification

The Variation Representation Specification (VR-Spec) is a standard developed by the Global Alliance for Genomic Health to facilitate and improve sharing of genetic information. The Specification consists of a JSON Schema for representing many classes of genetic variation, conventions to maximize the utility of the schema, and a Python implementation that promotes adoption of the standard.

- Introduction
- Terminology & Information Model

https://vr-spec.readthedocs.io

@handlerwagner

Part III – How to Learn Bioinformatics

Learning bioinformatics **usually requires solving computational problems** of varying difficulty that are extracted from real challenges of molecular biology.

-Rosalind Bioinformatics Education Platform

Learning bioinformatics **usually requires solving computational problems** of varying difficulty that are extracted from real challenges of molecular biology.

-Rosalind Bioinformatics Education Platform

The theme of **learning by doing** is probably the one that I suggest most to people.

-Nicholas J. Loman

Learning bioinformatics **usually requires solving computational problems** of varying difficulty that are extracted from real challenges of molecular biology.

-Rosalind Bioinformatics Education Platform

The theme of **learning by doing** is probably the one that I suggest most to people.

-Nicholas J. Loman

Learning by doing is not ideal but that's the reality.

-Mario Caccamo

Learning bioinformatics **usually requires solving computational problems** of varying difficulty that are extracted from real challenges of molecular biology.

-Rosalind Bioinformatics Education Platform

The theme of **learning by doing** is probably the one that I suggest most to people.

-Nicholas J. Loman

Learning by doing is not ideal but that's the reality.

-Mario Caccamo

The answer depends on **what you want to do**.

-Manoj Samanta, Homolog.us blog

Learning bioinformatics **usually requires solving computational problems** of varying difficulty that are extracted from real challenges of molecular biology.

-Rosalind Bioinformatics Education Platform

The theme of **learning by doing** is probably the one that I suggest most to people.

-Nicholas J. Loman

Learning by doing is not ideal but that's the reality.

-Mario Caccamo

The answer depends on **what you want to do**.

-Manoj Samanta, Homolog.us blog

...get the raw data (SRA) from an a lab evo paper (few muts) and **reproduce results**.

-Tami Lieberman

Learning bioinformatics **usually requires solving computational problems** of varying difficulty that are extracted from real challenges of molecular biology.

-Rosalind Bioinformatics Education Platform

The theme of **learning by doing** is probably the one that I suggest most to people.

-Nicholas J. Loman

Learning by doing is not ideal but that's the reality.

-Mario Caccamo

The answer depends on **what you want to do**.

-Manoj Samanta, Homolog.us blog

First thing I tell them is **close MS Excel**.

-Aylwyn Scally

...get the raw data (SRA) from an a lab evo paper (few muts) and **reproduce results**.

-Tami Lieberman

The theme of **learning by doing** is probably the one that I suggest most to people.

-Nicholas J. Loman

Learning bioinformatics **usually requires solving computational problems** of varying difficulty that are extracted from real challenges of molecular biology.

-Rosalind Bioinformatics Education Platform

Learning bioinformatics is the process of trying something new, making mistakes, and remembering how not to repeat them.

-Mario Caccamo

The answer depends on **what you want to do**.

-Manoj Samanta, Homolog.us blog

...get the raw data (SRA) from an a lab evo paper (few muts) and **reproduce results**.

-Tami Lieberman

First thing I tell them is **close MS Excel**.

-Aylwyn Scally

Trying Something New

Programming

Programming - UNIX

1. Familiarity with OS operations
    a. creating, moving, deleting files and directories
    b. I/O streams
    c. managing permissions

2. Learning common tools
    a. streaming text manipulation
        i. sed / awk / tr
        ii. head / tail
        iii. more / less
    b. compression
        i. tar / gzip
        ii. tabix
    c. text editing
        i. nano
        ii. emacs / vim
    d. remote operations
        i. ssh
        ii. scp
        iii. wget / curl

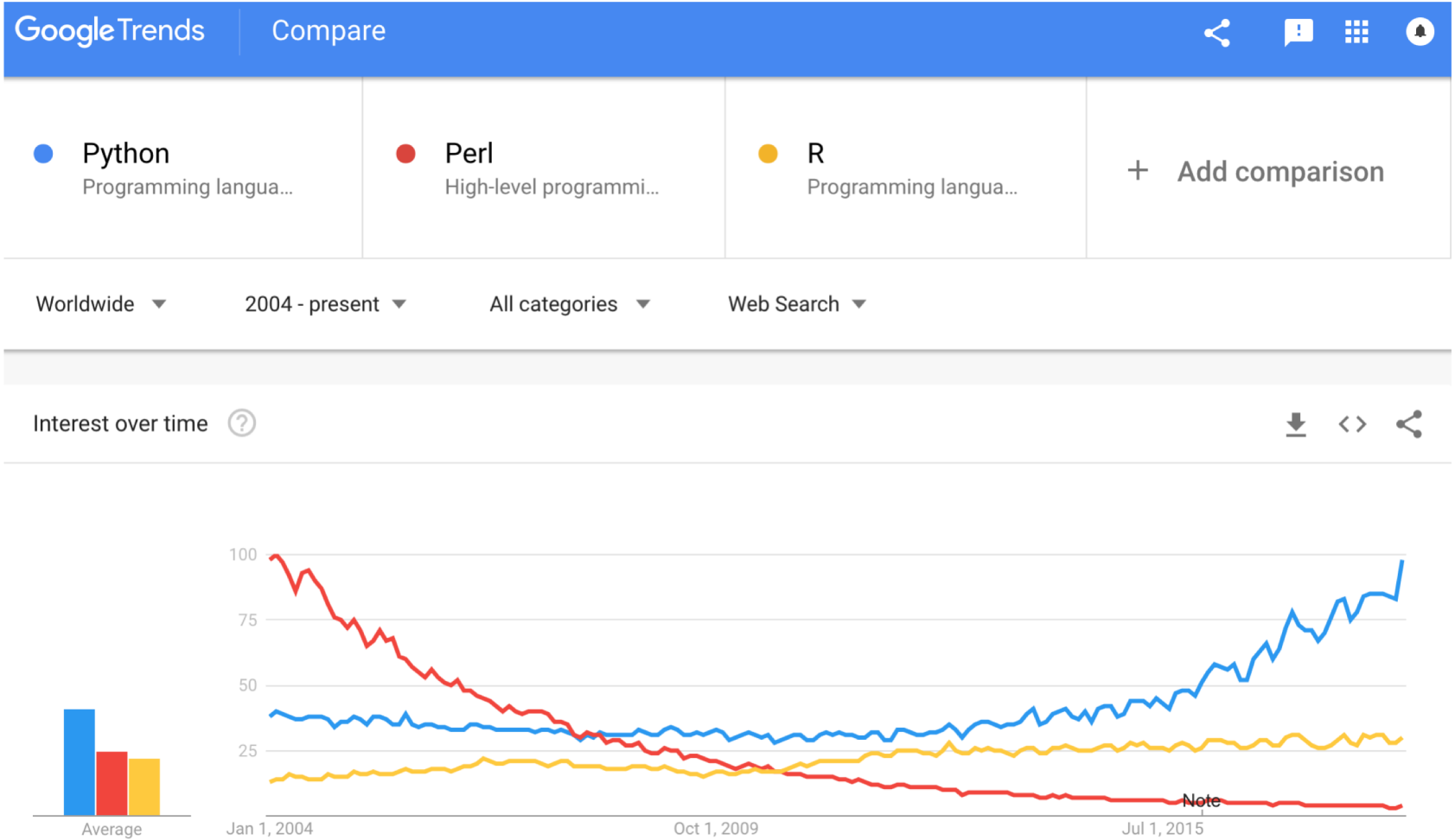
Programming - UNIX

rnaseq.wiki: a broad suite of tutorials covering an end-to-end bioinformatics workflow for RNA-seq analysis.
Includes a UNIX “bootcamp” tutorial:
https://github.com/griffithlab/rnaseq_tutorial/wiki/Unix-Bootcamp

tutorialspoint: generic programming tutorials repository.
Contains a very thorough UNIX introduction module:
www.tutorialspoint.com/unix/index.htm

Learnshell.org: Interactive UNIX shell programming tutorial.

Programming


Google Trends
Compare
Python
Programming langua...
Perl
High-level programmi...
R
Programming langua...
+ Add comparison
Worldwide
2004 - present
All categories
Web Search
Interest over time
100
75
50
25
Note
Average
Jan 1, 2004
Oct 1, 2009
Jul 1, 2015

Programming - Python

- General purpose language
  - Readability focused
  - Forgiving interpreter
- Broad applications
  - Machine Learning
  - Scripting
  - Data visualization
  - Web applications
- Great for Reproducibility
  - Environment management
  - Jupyter notebooks
  - PyCharm community IDE

Programming - Python

Hitchhiker's Guide to Python: an opinionated guide to getting started with Python. Includes commonly used extensions.

ROSALIND: bioinformatics learning platform with numerous "learn by doing" exercises, starts with a Python primer.

The Python Tutorial: Provided by the Python foundation, extremely comprehensive survey of the core library.

Programming - Python

Problem

A **string** is simply an ordered collection of symbols selected from some **alphabet** and formed into a word; the **length** of a string is the number of symbols that it contains.

An example of a length 21 **DNA string** (whose alphabet contains the symbols 'A', 'C', 'G', and 'T') is "ATGCTTCAGAAAGGTCTTACG."

Given: A DNA string s of length at most 1000 nt.

Return: Four integers (separated by spaces) counting the respective number of times that the symbols 'A', 'C', 'G', and 'T' occur in s.

Sample Dataset

```
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
```

Sample Output

```
20 12 17 21
```

http://rosalind.info/problems/dna/

Programming - Python

```
>>> from collections import Counter
>>> dna = 'AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC'
>>> nt_count = Counter(dna)
>>> nt_count
Counter({'T': 21, 'A': 20, 'G': 17, 'C': 12})
>>> out_list = [str(nt_count[nucleotide]) for nucleotide in ('A','C','G','T')]
>>> out_list
['20', '12', '17', '21']
>>> print(' '.join(out_list))
20 12 17 21
```

http://rosalind.info/problems/dna/

Programming - R

- Statistical Computing Language
- Strengths
  - Machine Learning
  - Data visualization
    - ggplot2
- Broad bioinformatics community
  - historical depth with Bioconductor
- Actively maintained IDE
  - RStudio

Programming - R

SWIRL: interactive instruction in R programming and data science in the R console.

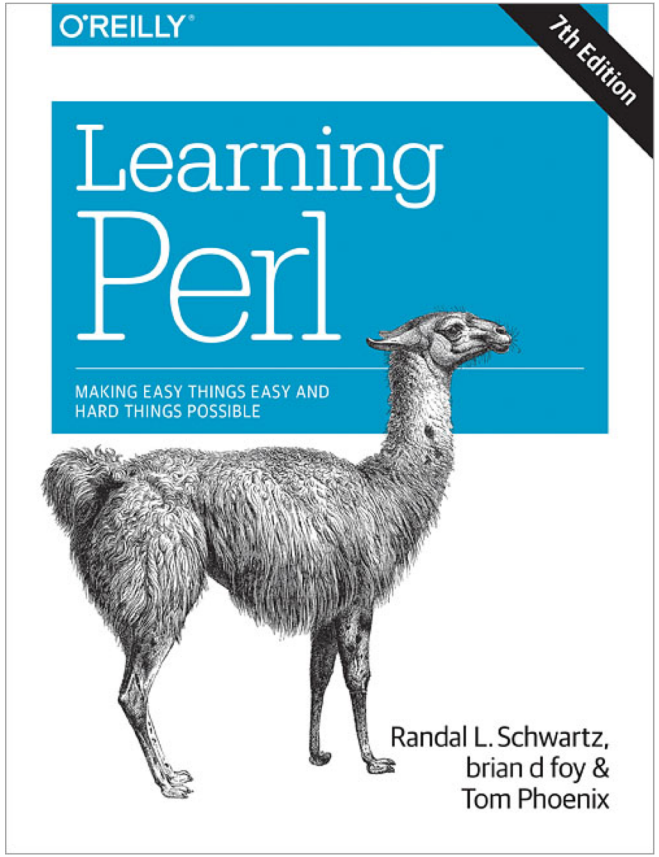
R-bloggers: R news and tutorials contributed by R blogger community.

GenViz.org: Genomic Visualizations with R tutorial, contains an R primer.

Programming

Programming - Perl

- Historically popular
- Very useful for text manipulation
- Utility as scripting language in decline



Making Mistakes

Making Mistakes - Resources

Question: Snp, Dip, Snv Notation

4

A nomenclature question. Considering:

- a **SNP** is a **single** nucleotide **polymorphism**, so a nucleotide change **within-species** of a single position, and
- a **DIP** is a **deletion/insertion polymorphism**, so an **indel within-species** of a single genomic position, how would one call:
- a nucleotide change **between-species** of a single homologous genomic coordinate?
- how about an **indel between-species** of a single homologous genomic coordinate? Is there an official or recognized nomenclature for them?

7.1 years ago by 2184687-1231-83- • **4.9k**

snp • 21k views

ADD COMMENT • link • Not following ▾ modified 3.6 years ago by scott.a.fay • **0** • written 7.1 years ago by 2184687-1231-83- • **4.9k**

2 do you have a reference to that definition of SNV?

ADD REPLY • link written 7.1 years ago by Jeremy Leipzig ♦ **17k**

1 I thought the difference between SNP and SNV was one of population frequency (SNP being a frequently observed SNV) not of intra versus inter species differences

ADD REPLY • link written 7.1 years ago by Russh • **1.2k**

11

It might clarify the SNP and SNV by reading this blog - "SNP vs. SNP"?

Quote:

> *SNP (single nucleotide polymorphism) vs. SNV (single nucleotide variant) As their name suggests, both are concerned with aberrations at a single nucleotide. However, a SNP is when an aberration is expected at the position for any member in the species – for example, a well characterized allele. A SNV on the other hand is when there is a variation at a position that hasn't been well characterized – for example, when it is only seen in one individual. It is really all a question of frequency of occurrence.*

7.1 years ago by Boboppie • **520**
Cambridge, UK

Remembering

Remembering - git

- Speed
- Simple Design
- Strong Support for non-linear development (thousands of branches)
- Fully distributed
- Scalable to very large projects (like the Linux Kernel)

https://git-scm.com/book/en/v1/Getting-Started

Remembering - GitHub

- Hosts Git Repositories
- Documentation Markdown
- Wikis
- Issue Tracking
- Repo Management
- Collaborative Workflow

ahwagner / civic_pmids.ipynb

Created 10 months ago

Edit | Delete | Star 0

Code | Revisions 1

Embed | <script src="https://gist. | Download ZIP

Jupyter Notebook for extracting all PMIDs in CIViC

civic_pmids.ipynb | Raw

In [1]:
```
import requests
```

In [2]:
```
url = "https://civic.genome.wustl.edu/api/sources"
```

In [3]:
```
pmids = set()
next_page = url
while next_page:
    resp = requests.get(next_page)
    resp.raise_for_status()
    json = resp.json()
    for record in json['records']:
        pmid = record.get('pubmed_id', None)
        if pmid is None:
            continue
        pmids.add(pmid)
    next_page = json['_meta']['links']['next']
```

In [4]:
```
len(pmids)
```

Out[4]: 1752

In [5]:
```
with open('pmids.txt', 'w') as out:
    for pmid in pmids:
        print(pmid, file=out)
```

Remembering - GitHub

- Hosts Git Repositories
- Documentation Markdown
- Wikis
- Issue Tracking
- Repo Management
- Collaborative Workflow

https://guides.github.com/activities/hello-world/

The theme of **learning by doing** is probably the one that I suggest most to people.

-Nicholas J. Loman

Learning bioinformatics **usually requires solving computational problems** of varying difficulty that are extracted from real challenges of molecular biology.

-Rosalind Bioinformatics Education Platform

Learning bioinformatics is the process of trying something new, making mistakes, and remembering how not to repeat them.

-Mario Caccamo

The answer depends on **what you want to do**.

-Manoj Samanta, Homolog.us blog

...get the raw data (SRA) from an a lab evo paper (few muts) and **reproduce results**.

-Tami Lieberman

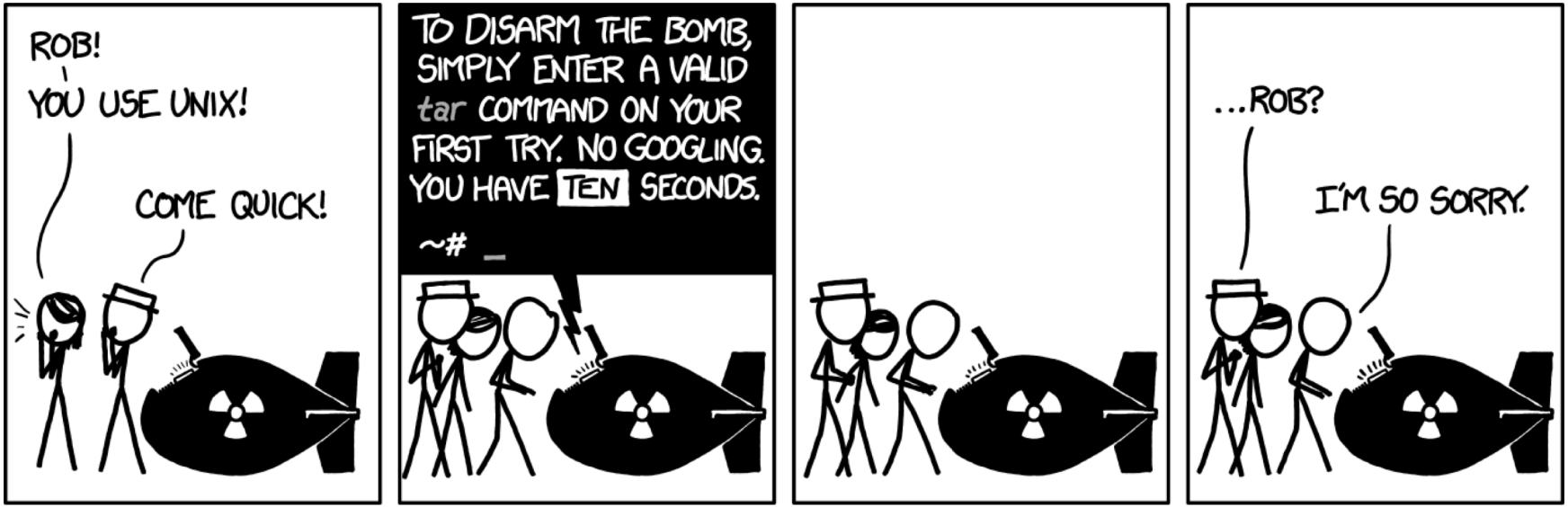
First thing I tell them is **close MS Excel**.

-Aylwyn Scally

First thing I tell them is **close MS Excel**.

-Aylwyn Scally

Questions?



https://xkcd.com/1168/