Clinical Genome Knowledge Base and Linked Data technologies

ClinGen

Clinical Genome Resource

Aleksandar Milosavljevic

BCM

Baylor College of Medicine

Topics

1. ClinGen Resource project

2. Building the Clinical Genome Knowledge Base

3. Linked Data technologies

4. Using Linked Data technologies to enable pattern discovery in the Clinical Genome Knowledge Base

Topics

1. ClinGen Resource project

1. Building the Clinical Genome Knowledge Base

2. Linked Data technologies

3. Using Linked Data technologies to enable pattern discovery in the Clinical Genome Knowledge Base

ClinGen
Clinical Genome Resource

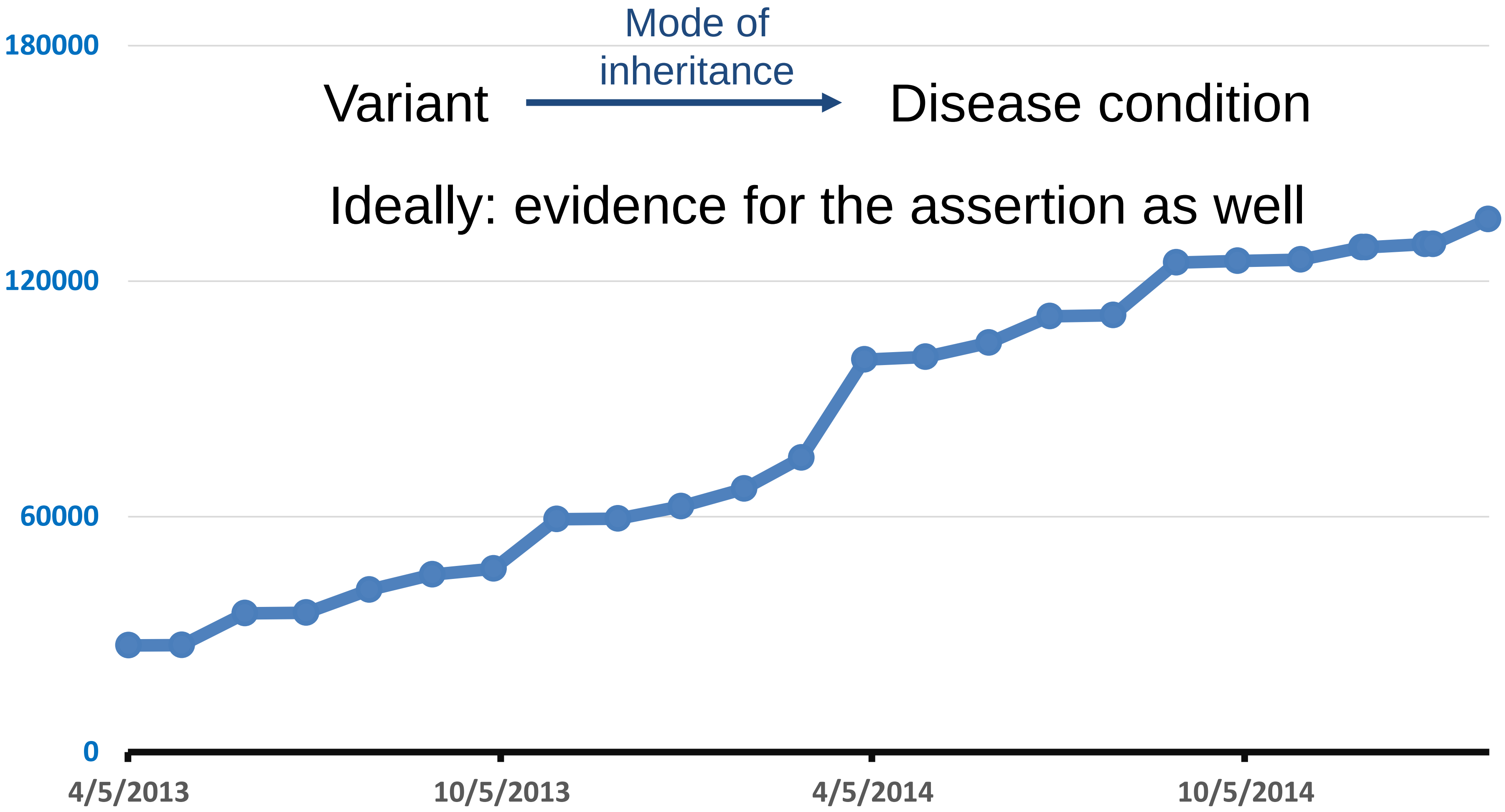
www.clinicalgenome.org

- Engage the clinical genomics community in data sharing efforts (sharing of data from genetic testing labs)

- Develop the infrastructure and standards to support curation of the knowledge about the role of genes and genetic variants in human diseases

- Incorporate machine-learning approaches to speed the identification of clinically relevant variants

Variants in ClinVar
(mostly from Diagnostic Genetic Testing)

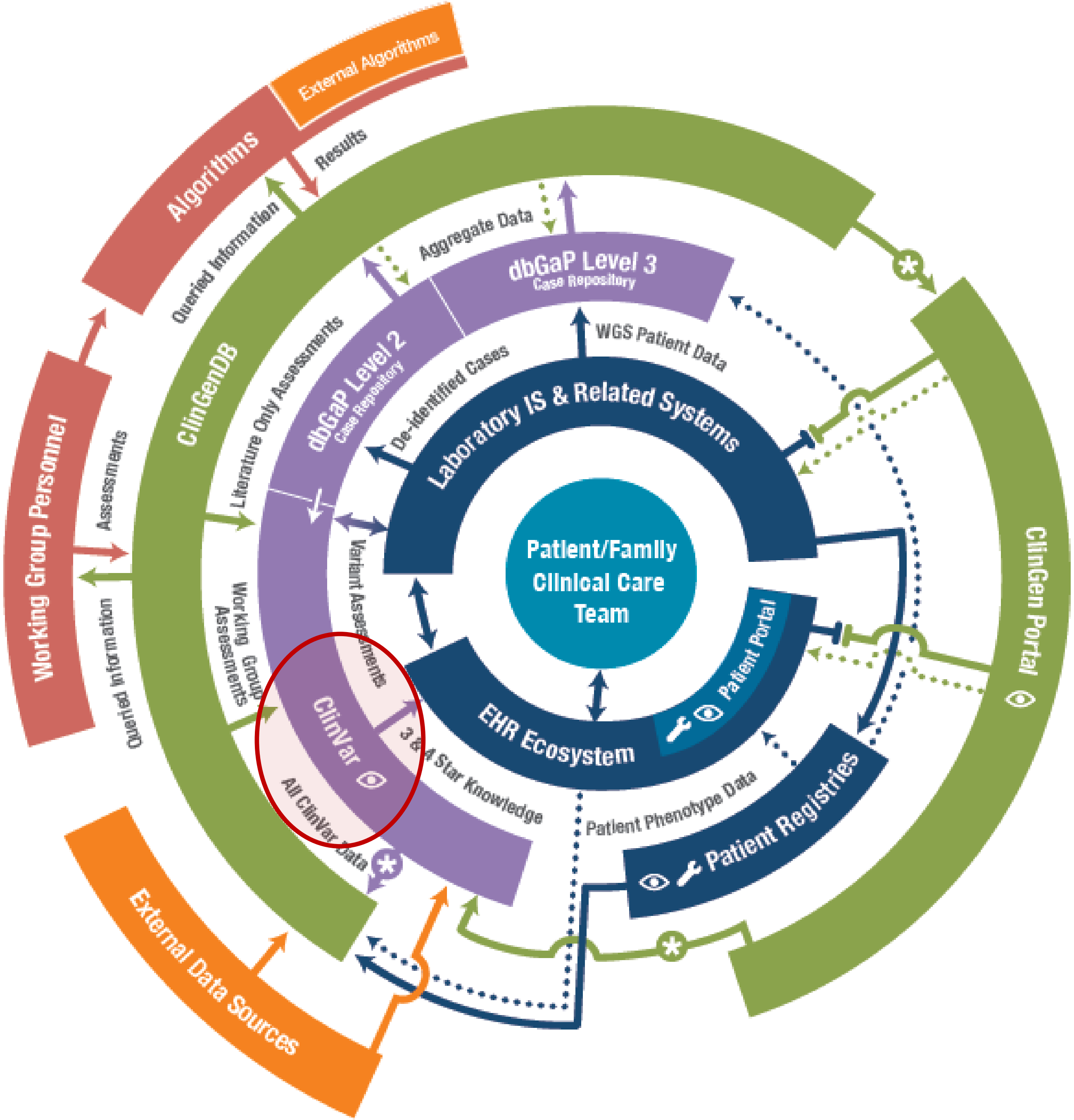


Variant $\xrightarrow{\text{Mode of inheritance}}$ Disease condition

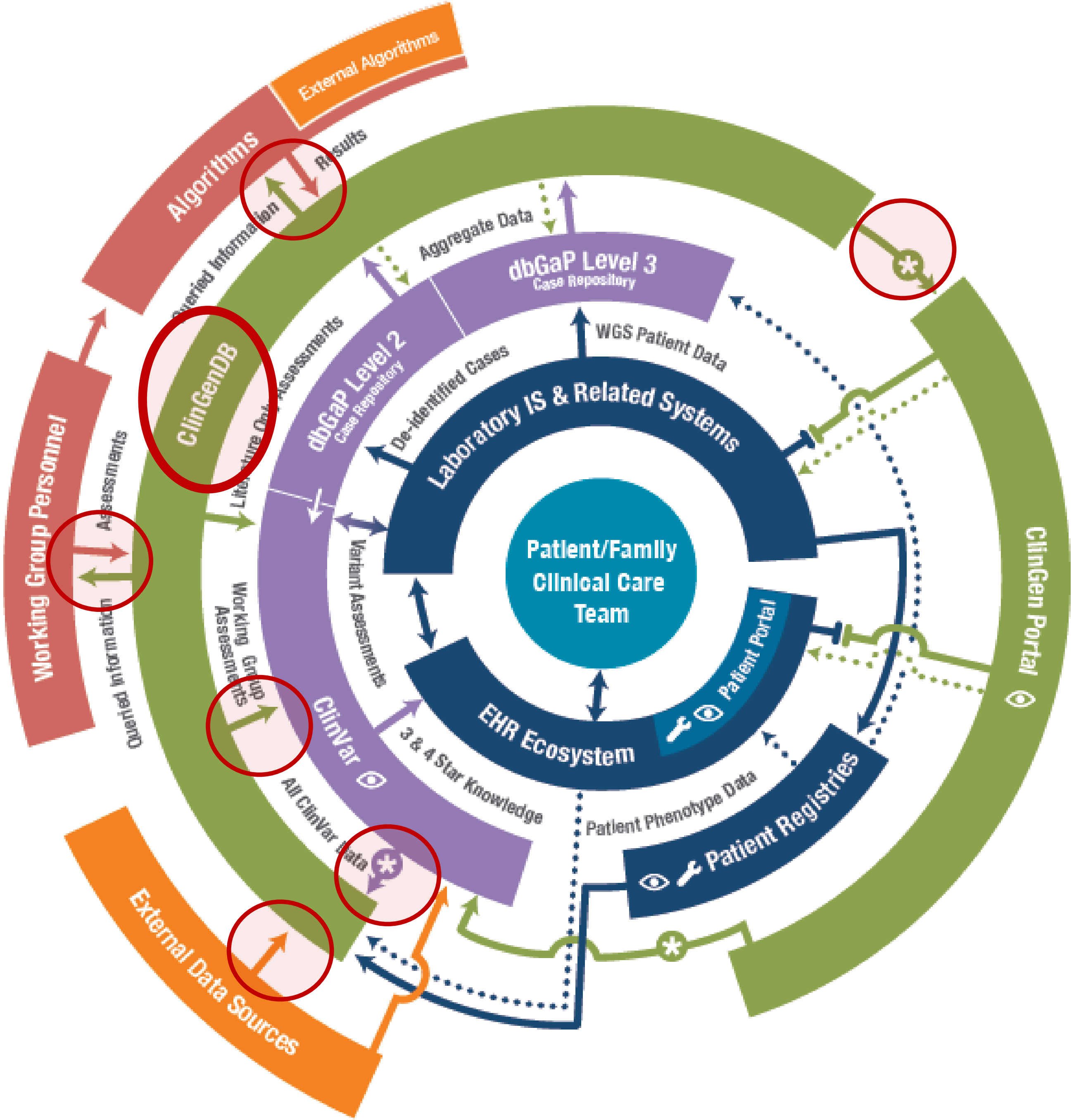
Ideally: evidence for the assertion as well

(Source: ClinVar)

ClinGen
Clinical Genome Resource



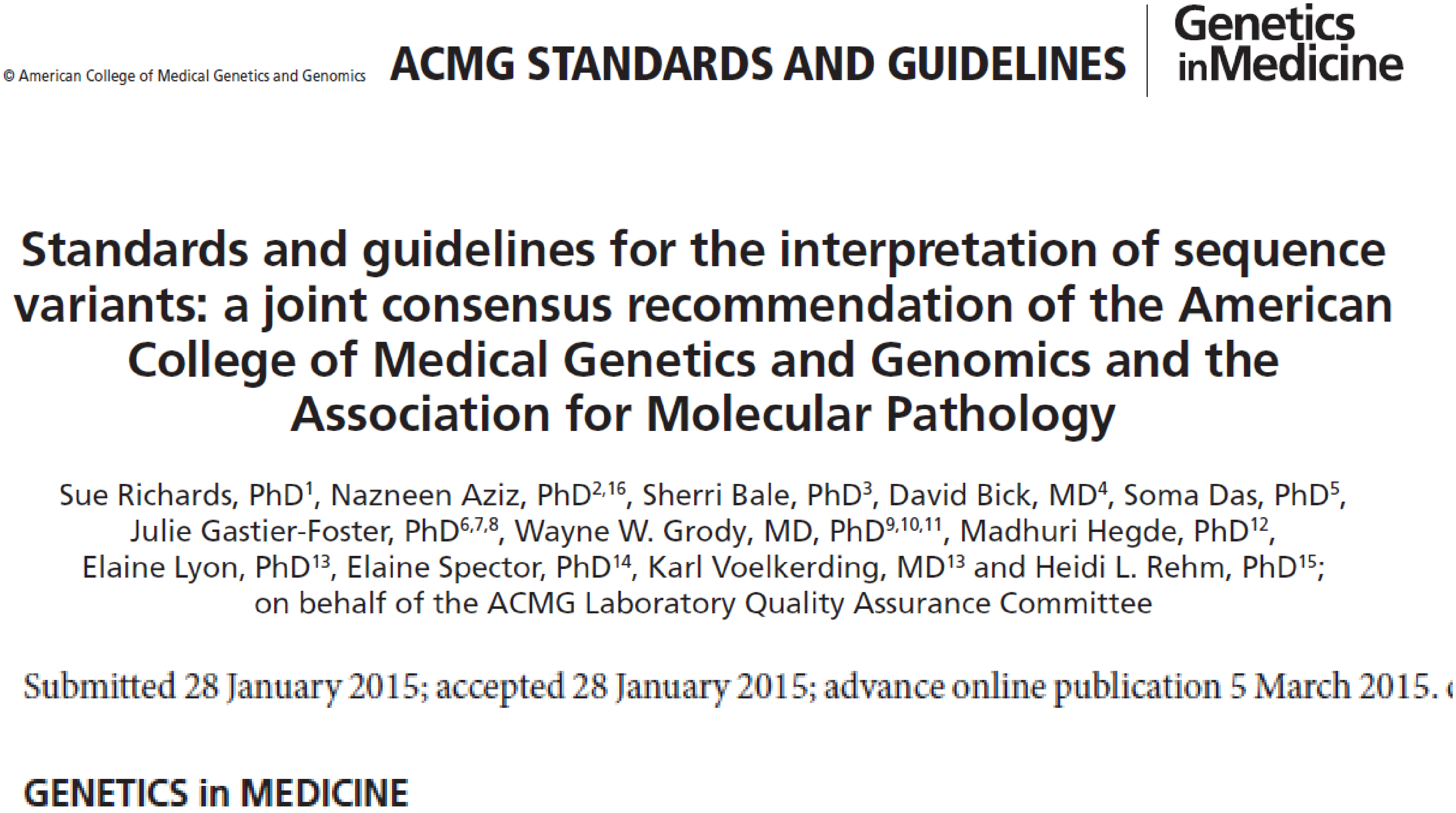
ClinGen
Clinical Genome Resource



Topics

1. ClinGen Resource project
2. Building the Clinical Genome Knowledge Base
3. Linked Data technologies
4. Using Linked Data technologies to enable pattern discovery in the Clinical Genome Knowledge Base

Evaluating pathogenicity of genetic variants for specific diseases



ACMG STANDARDS AND GUIDELINES | Genetics inMedicine

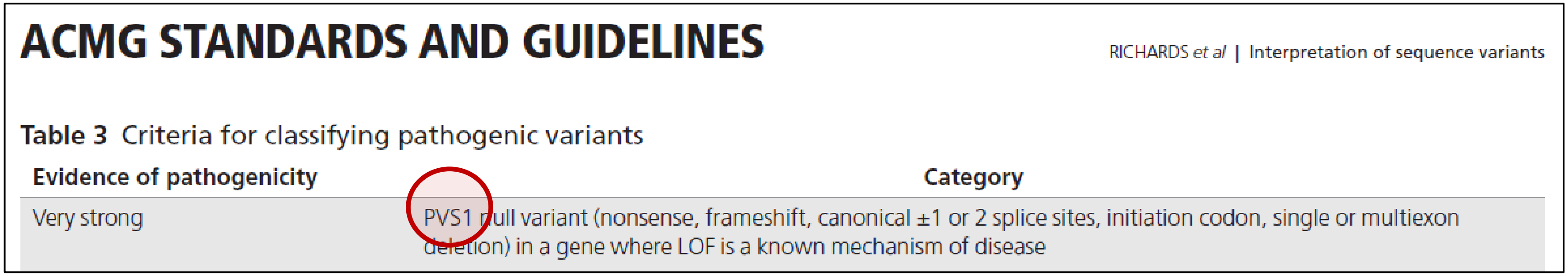
Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5], Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12], Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15]; on behalf of the ACMG Laboratory Quality Assurance Committee

Submitted 28 January 2015; accepted 28 January 2015; advance online publication 5 March 2015.

GENETICS in MEDICINE

Evaluating pathogenicity of genetic variants for specific diseases

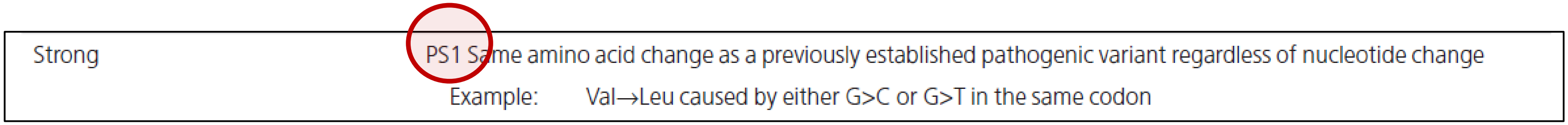


ACMG STANDARDS AND GUIDELINES

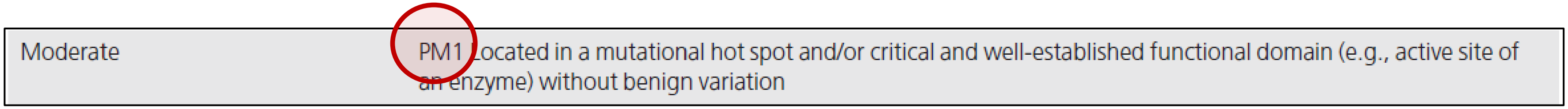
RICHARDS *et al* | Interpretation of sequence variants

Table 3 Criteria for classifying pathogenic variants

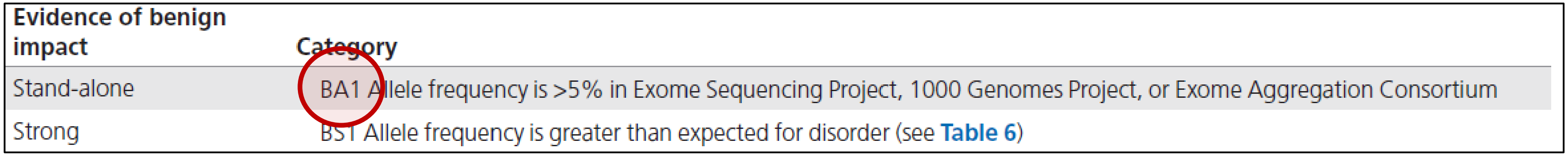
Evidence of pathogenicity	Category
Very strong	PVS1 null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease



Strong	PS1 Same amino acid change as a previously established pathogenic variant regardless of nucleotide change Example: Val→Leu caused by either G>C or G>T in the same codon

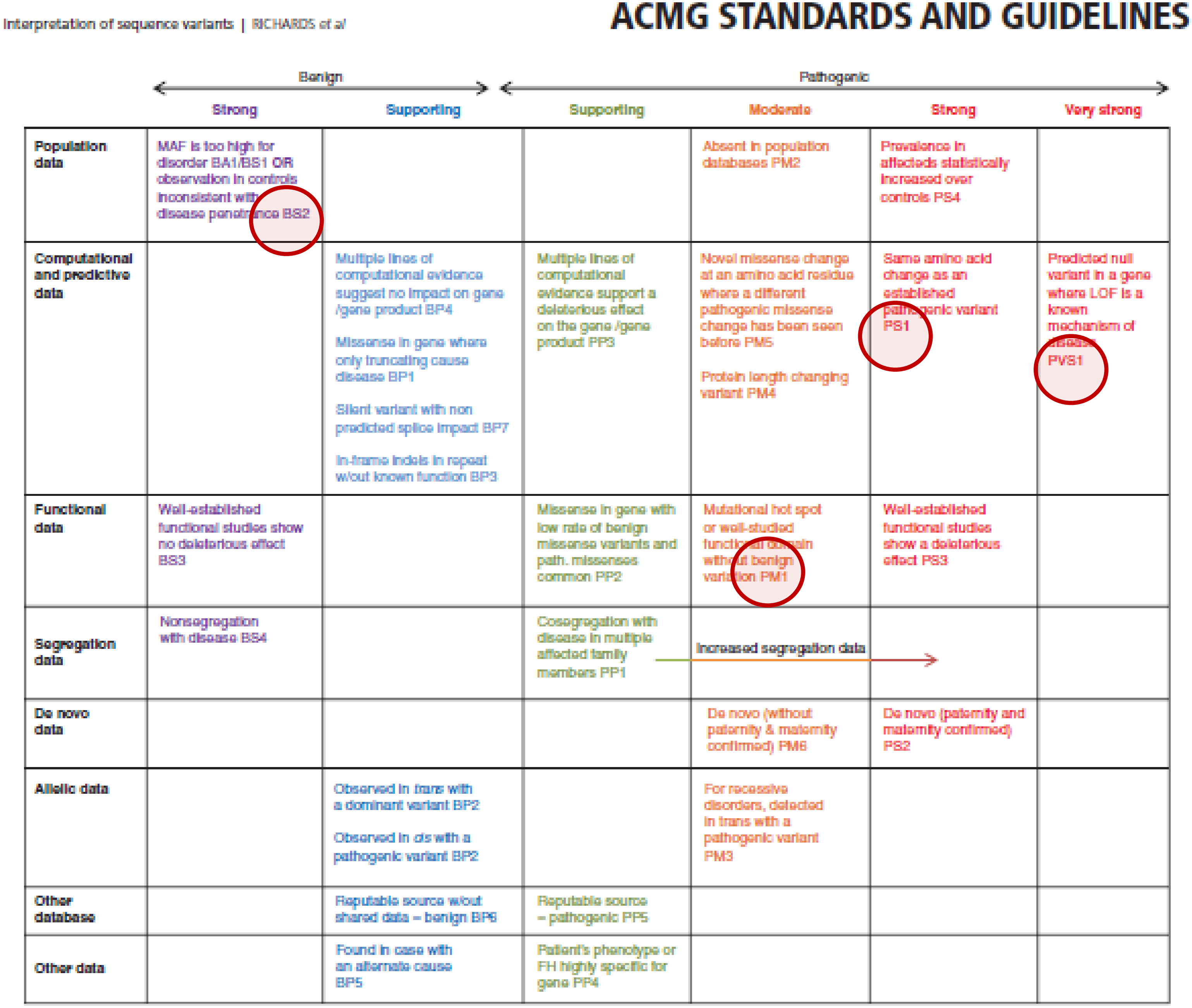


Moderate	PM1 Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation



Evidence of benign impact	Category
Stand-alone	BA1 Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium
Strong	BS1 Allele frequency is greater than expected for disorder (see **Table 6**)

ACMG STANDARDS AND GUIDELINES



Interpretation of sequence variants | RICHARDS *et al*

	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Evaluating pathogenicity of genetic variants for specific diseases

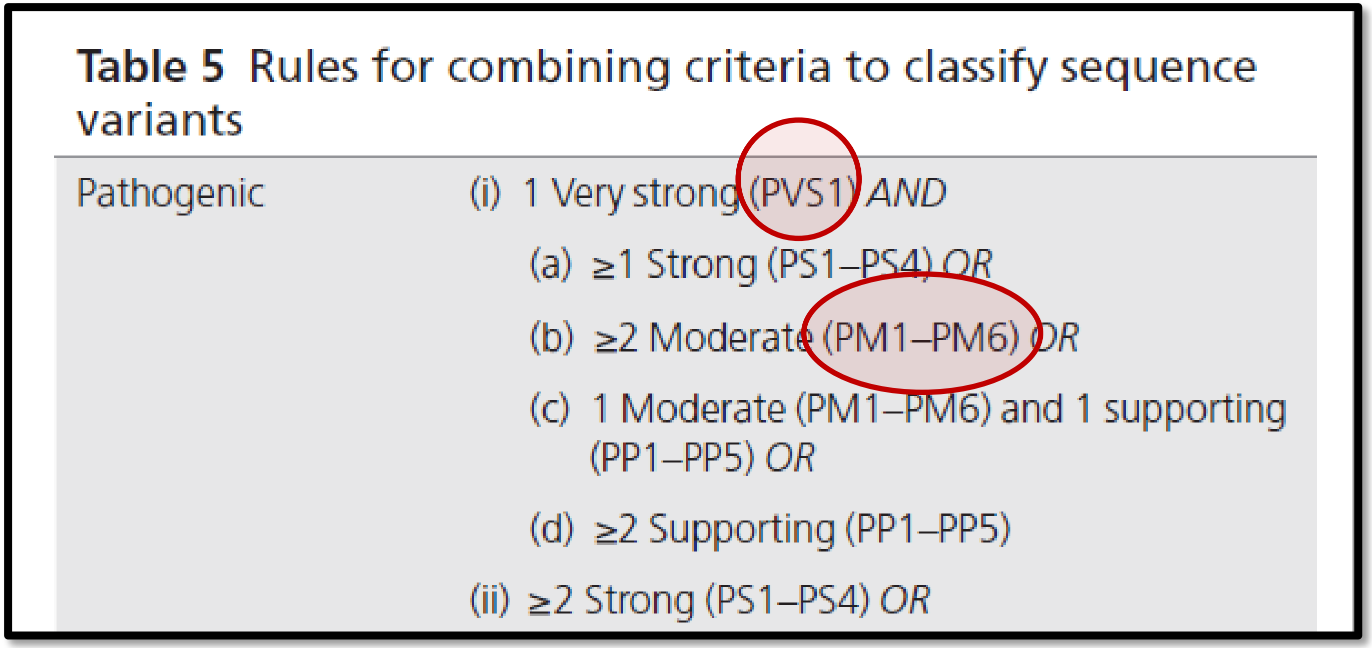
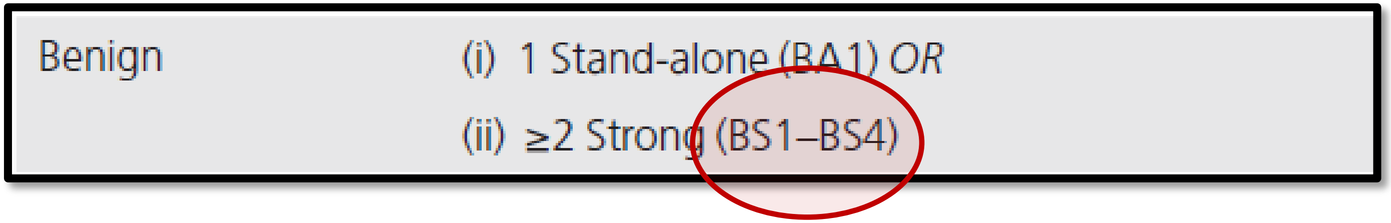


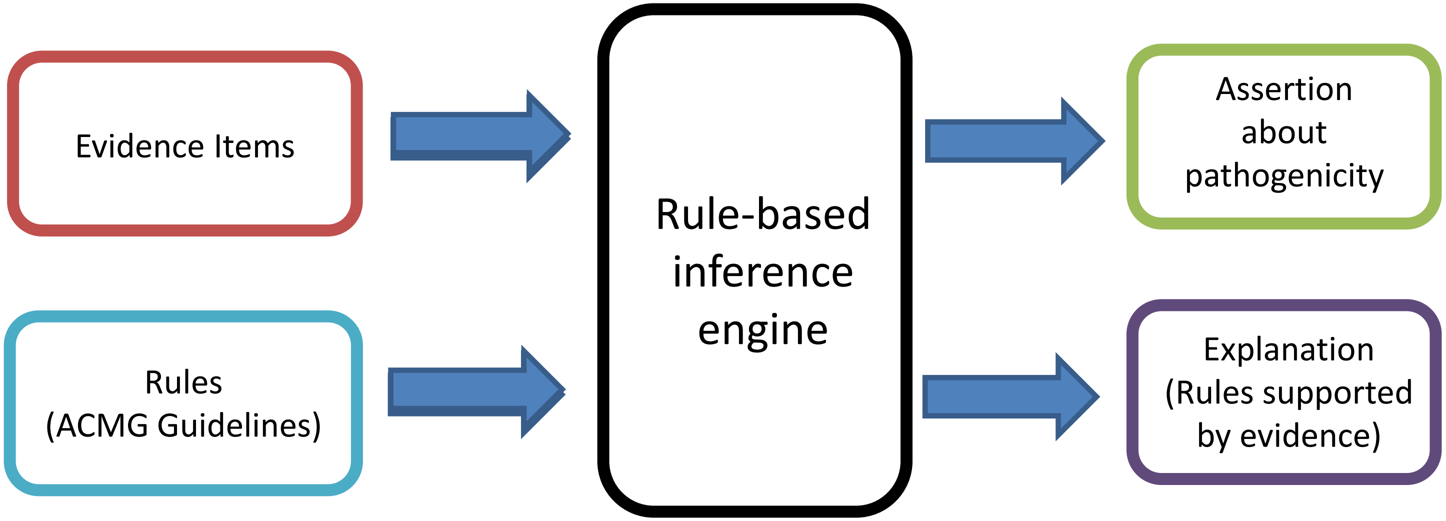
Table 5 Rules for combining criteria to classify sequence variants

Pathogenic	(i) 1 Very strong (PVS1) *AND* (a) ≥1 Strong (PS1–PS4) *OR* (b) ≥2 Moderate (PM1–PM6) *OR* (c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) *OR* (d) ≥2 Supporting (PP1–PP5) (ii) ≥2 Strong (PS1–PS4) *OR*



Benign	(i) 1 Stand-alone (BA1) *OR* (ii) ≥2 Strong (BS1–BS4)

Evaluating pathogenicity of genetic variants for specific diseases



Evidence Items

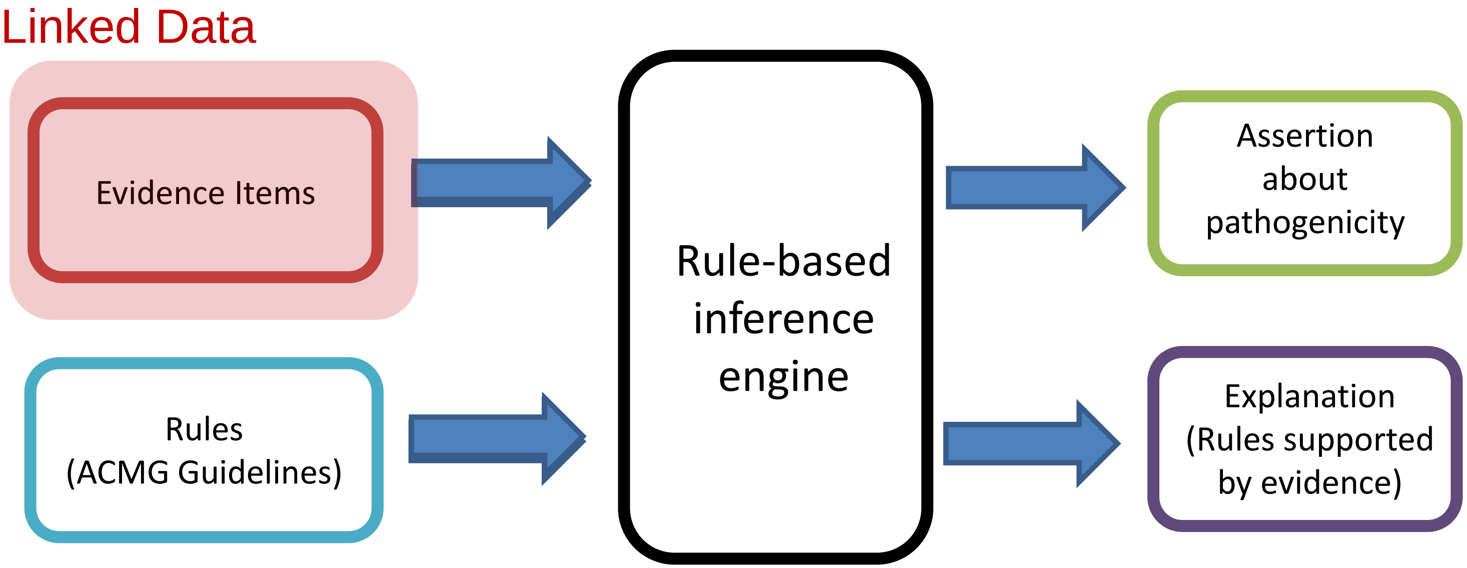
Rules (ACMG Guidelines)

Rule-based inference engine

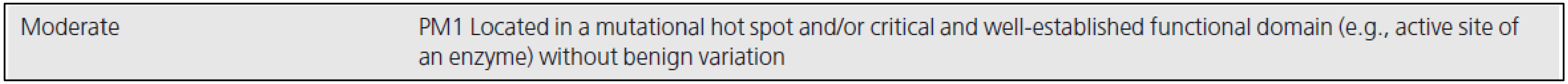
Assertion about pathogenicity

Explanation (Rules supported by evidence)

Evaluating pathogenicity of genetic variants for specific diseases

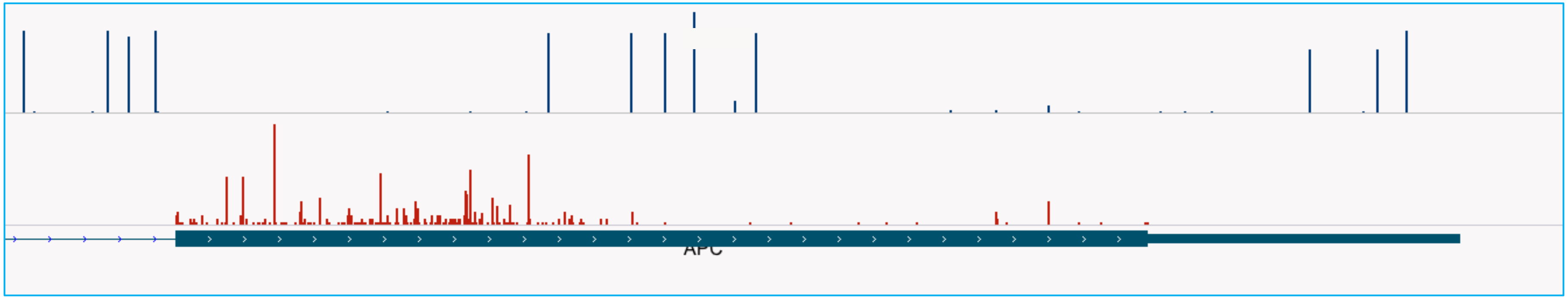

Linked Data
Evidence Items
Rules
(ACMG Guidelines)
Rule-based inference engine
Assertion about pathogenicity
Explanation
(Rules supported by evidence)

Evidence code PM1: mutational hotspot



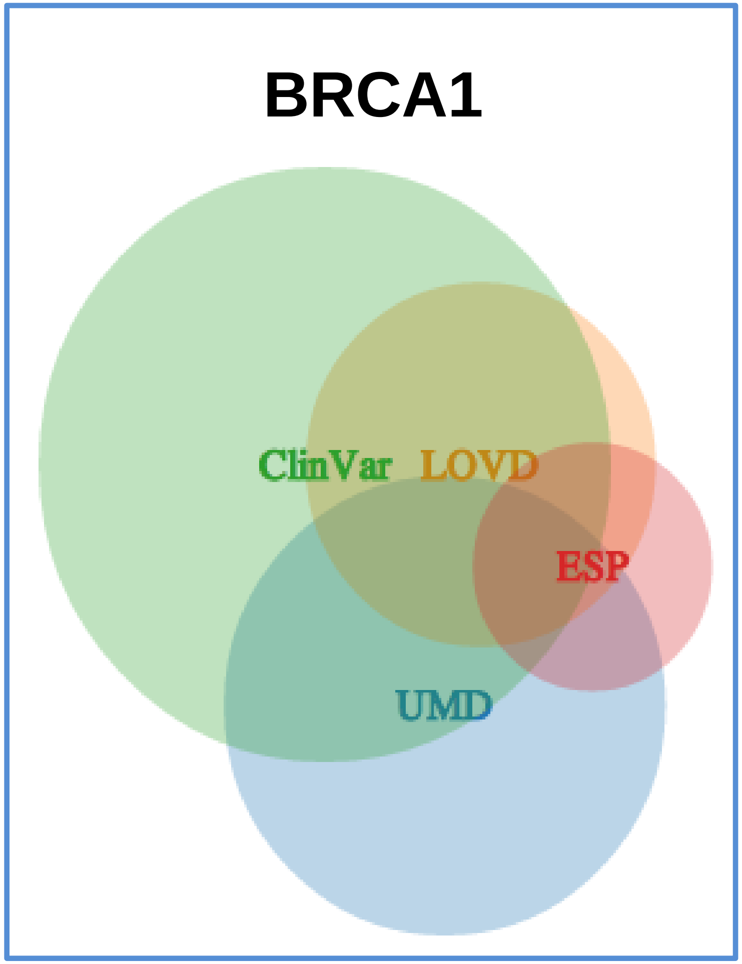
Moderate	PM1 Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation

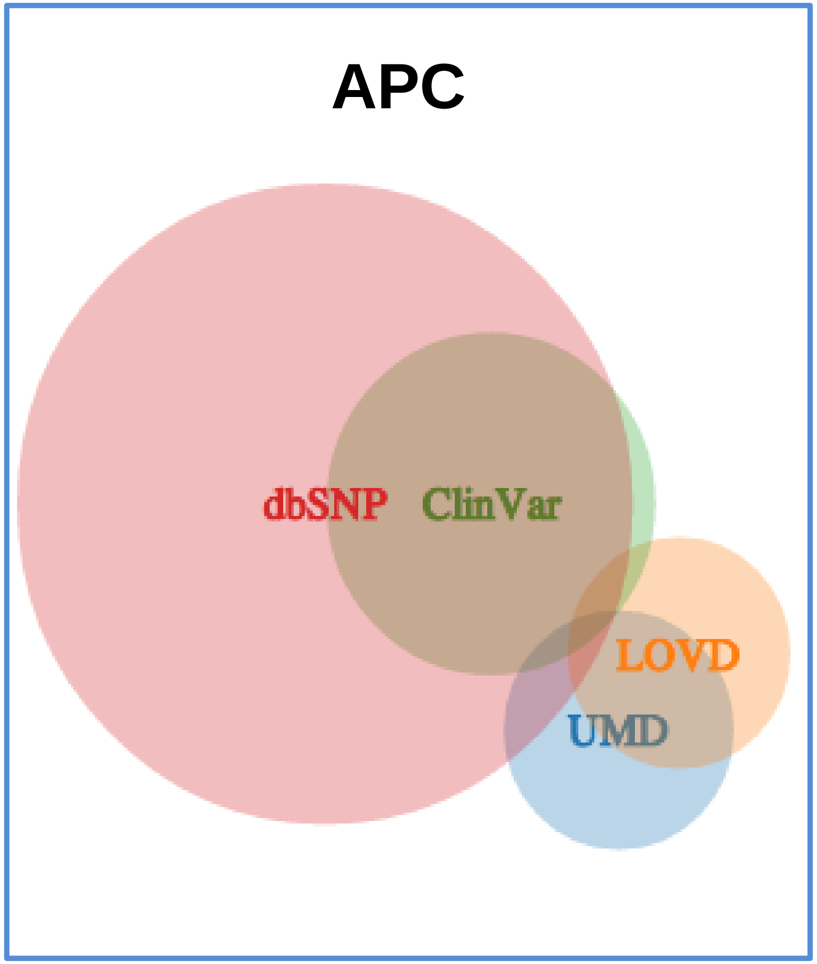
Variant frequency profile in normal population



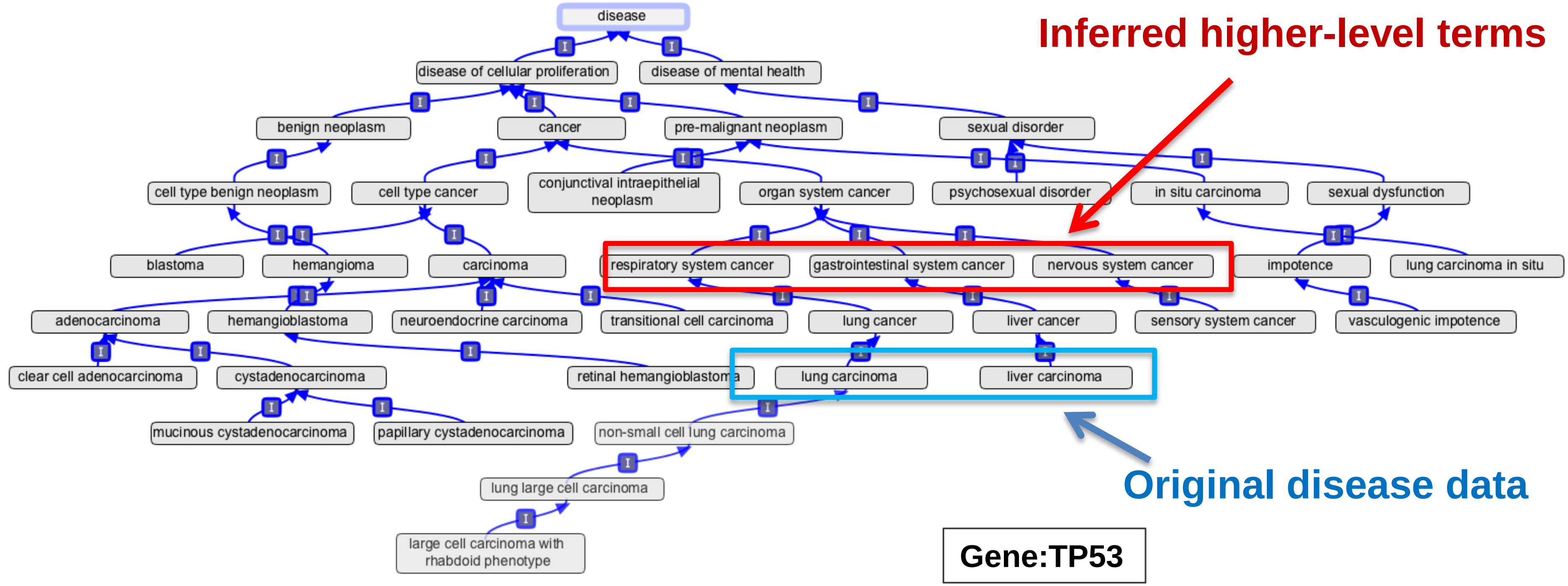
Cancer-predisposing variants

Collating information about variants from public sources





Ontology traversal to collate variants linked to related disease conditions



Inferred higher-level terms

Original disease data

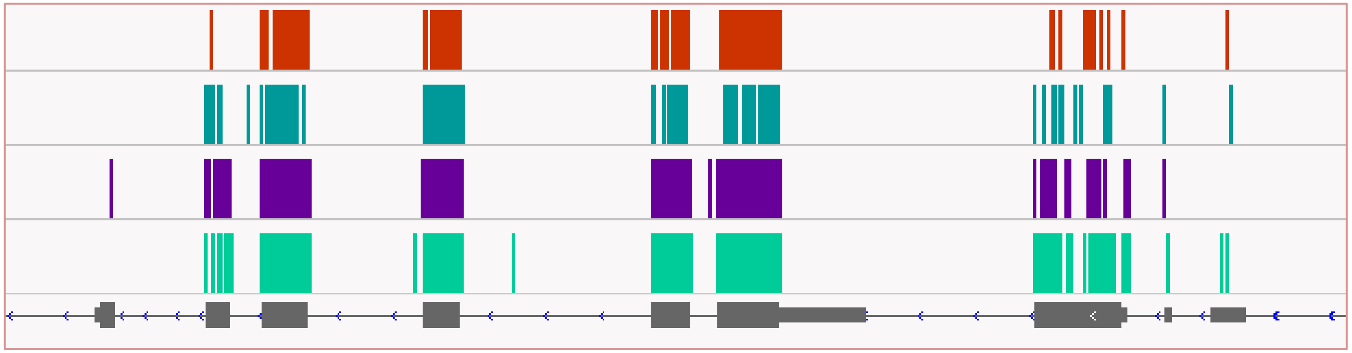
Gene:TP53

Reproductive organ cancer

Integumentary system cancer

Gastrointestinal system cancer

Respiratory system cancer



Topics

1. ClinGen Resource project
2. Building the Clinical Genome Knowledge Base
3. Linked Data technologies
4. Using Linked Data technologies to enable pattern discovery in the Clinical Genome Knowledge Base

RDF: Resource Description Framework

- Originally: metadata data model describing web resources

- Web resource = anything identifiable by a URL

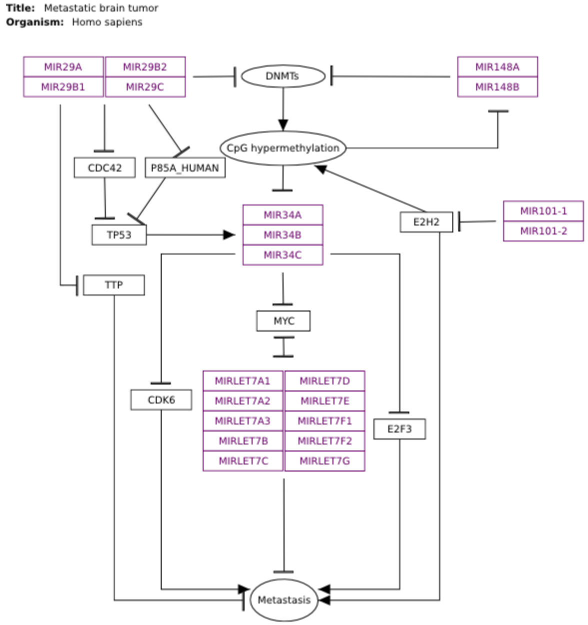
- Subject-predicate-object expressions
 (related to classical entity-property-value)

- Subject = anything identifiable by a URL / URI

- Expression = edge in a “Knowledge Graph”

- Wkipathways “serializes” biological pathway data as RDF

Example:
Metastatic brain tumor pathway



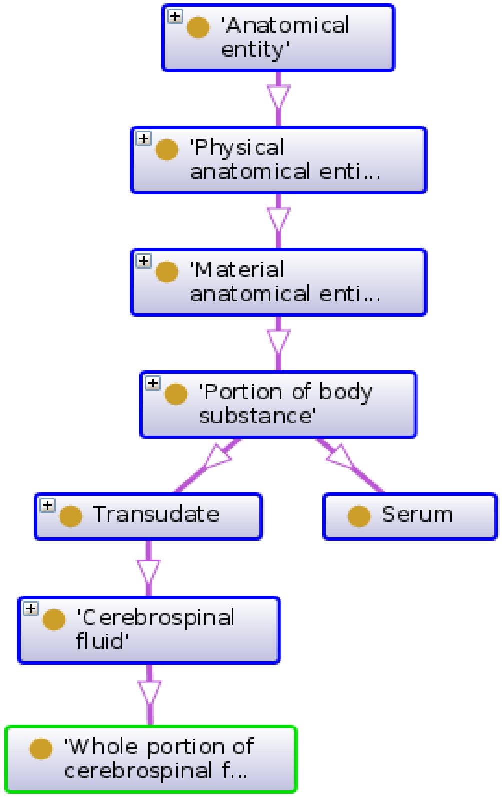
RDFs: RDF Schema

Uses RDF to express data model (data schema)

Some RDFs predicates:

- subClassOf – the subject is a subclass of a class
- subPropertyOf – the subject is a subproperty of a property
- domain – domain of the subject property
- range - range of the subject property

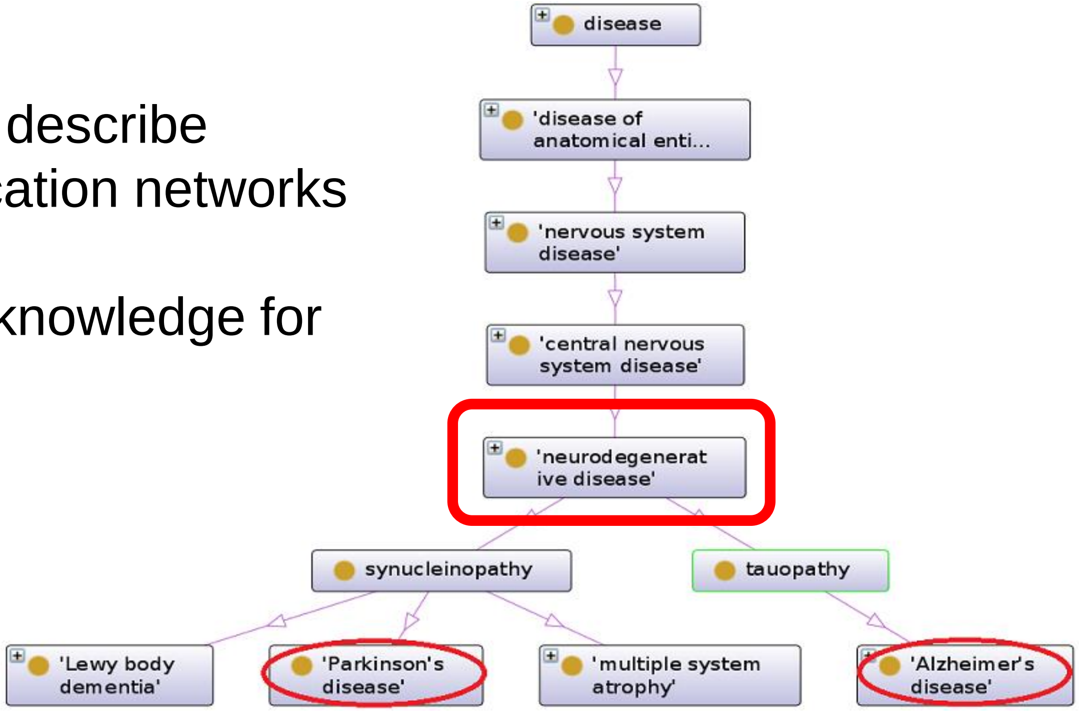
Example:
classification of body fluids


'Anatomical entity'
'Physical anatomical enti...
'Material anatomical enti...
'Portion of body substance'
Transudate
Serum
'Cerebrospinal fluid'
'Whole portion of cerebrospinal f...

OWL: Web Ontology Language

Example: “ontology traversal”
to annotate variants
to more general disease categories

- Uses RDF to express to describe taxonomies and classification networks
- Defines the structure of knowledge for various domains



Some Linked Data projects

“RDF-like graphs”:

- Facebook Social Graph -- Open Graph API

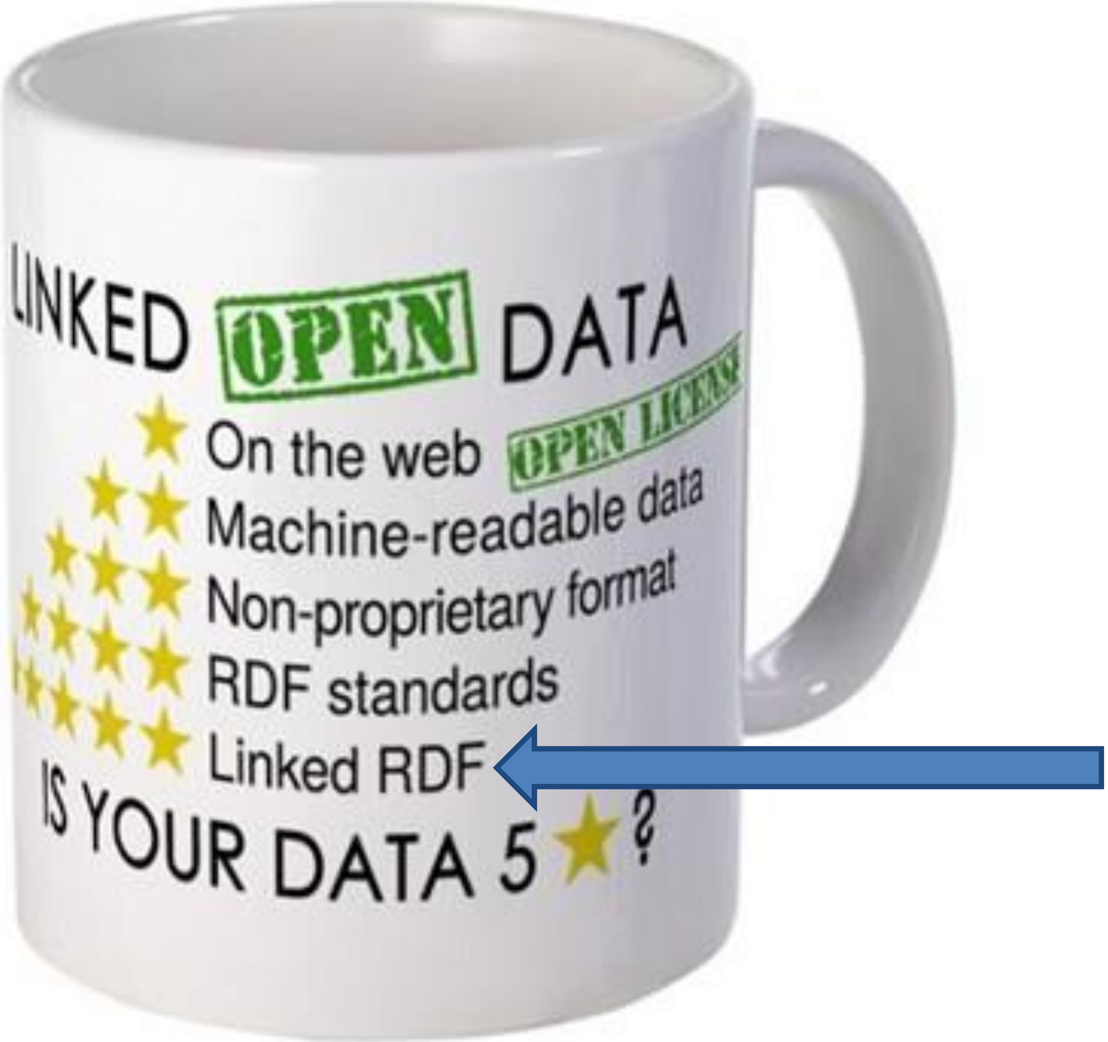
- Google Knowledge Graph / Vault

- WikiData

RDF graphs:

- data.gov
- Bio2RDF
- Bioontologies / Bioportal -- RDF, RDFs, OWL

What is Linked *Open* Data?



Linked RDF Standards

1999-2014 RDF, RDFs, OWL *-- graph syntax and semantics*

2014 JSON-LD *-- JSON -- data format for RDF graphs*

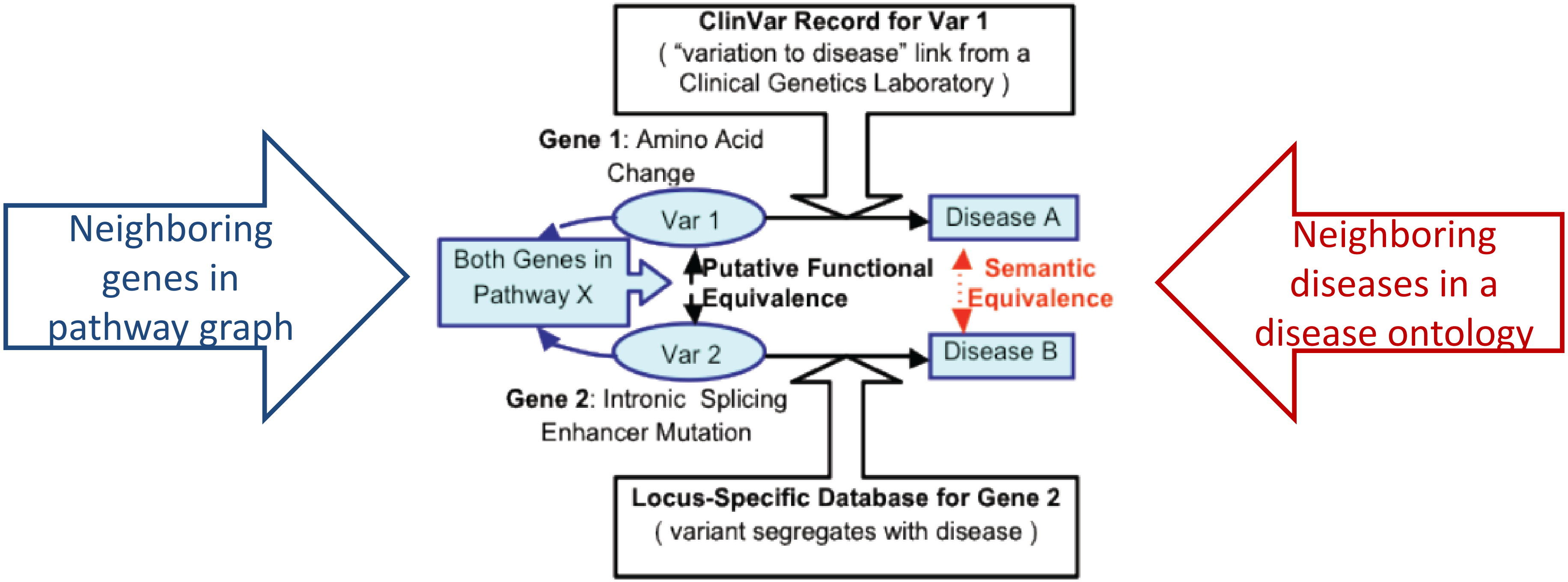
Feb 26 2015 Linked Data Platform 1.0 – *HTTP operations*

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
- Include links to other URIs, so that they can discover more things

Topics

1. ClinGen Resource project
2. Building the Clinical Genome Knowledge Base
3. Linked Data technologies
4. Using Linked Data technologies to enable pattern discovery in the Clinical Genome Knowledge Base

Use Case #1: Discovering “match” patterns



Neighboring genes in pathway graph

ClinVar Record for Var 1
(“variation to disease” link from a Clinical Genetics Laboratory)

Gene 1: Amino Acid Change

Var 1

Disease A

Both Genes in Pathway X

Putative Functional Equivalence

Semantic Equivalence

Var 2

Disease B

Gene 2: Intronic Splicing Enhancer Mutation

Locus-Specific Database for Gene 2
(variant segregates with disease)

Neighboring diseases in a disease ontology

Use Case #2: Discovering Recurrent Mutually Exclusive (RME) mutational patterns in cancer

Hallmark phenotypes of cancer are acquired by positive selection

(Hanahan and Weinberg, Cell 2001 and 2011)

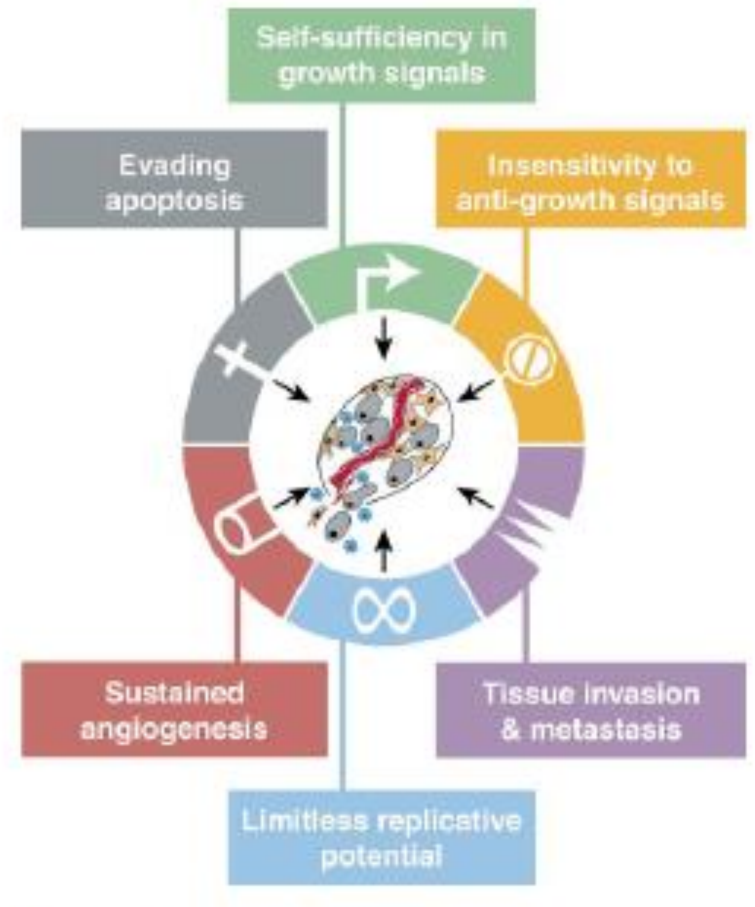
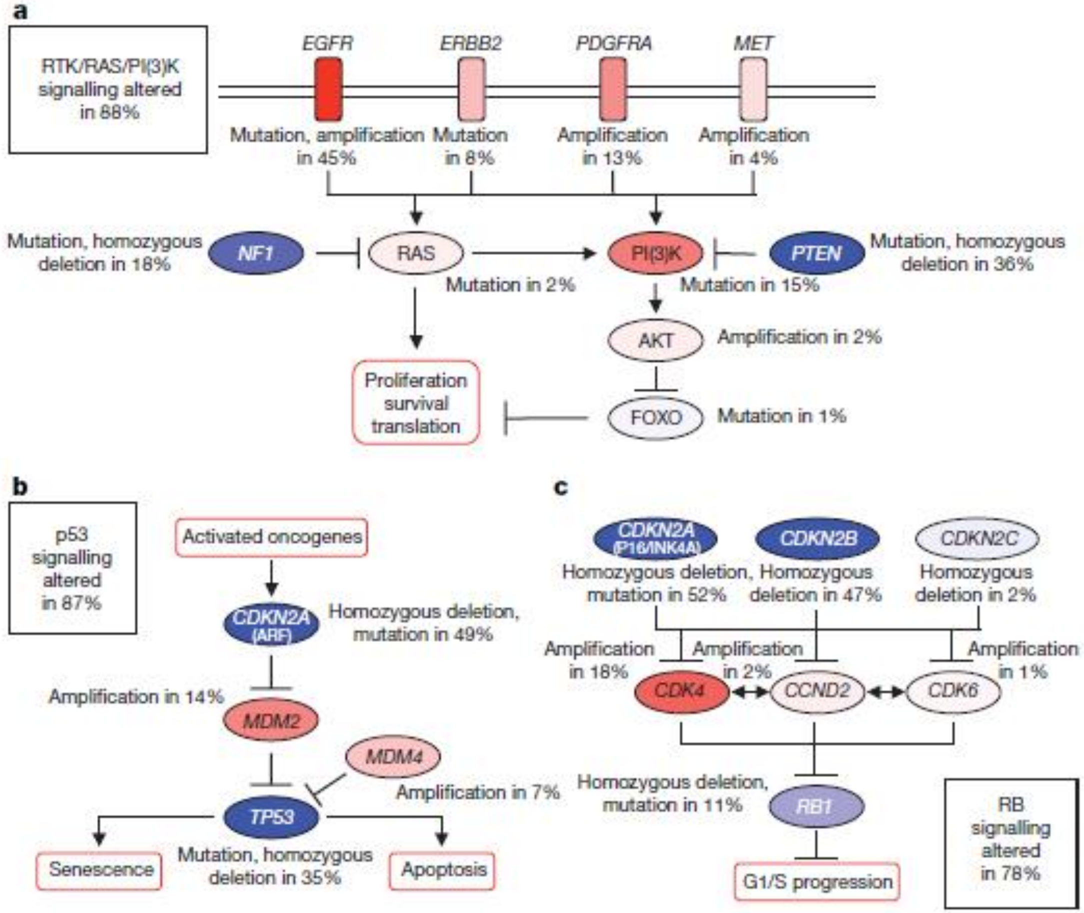

Self-sufficiency in growth signals
Evading apoptosis
Insensitivity to anti-growth signals
Sustained angiogenesis
Tissue invasion & metastasis
Limitless replicative potential

Figure 1. Acquired Capabilities of Cancer

Recurrent mutually exclusive mutational patterns identified in key glioblastoma pathways

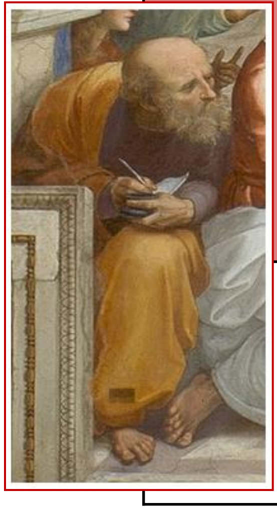
Comprehensive genomic characterization defines human glioblastoma genes and core pathways TCGA Research Network, *Nature* **455**, 1061-1068 (2008)


a
RTK/RAS/PI(3)K signalling altered in 88%
EGFR
ERBB2
PDGFRA
MET
Mutation, amplification in 45%
Mutation in 8%
Amplification in 13%
Amplification in 4%
Mutation, homozygous deletion in 18%
NF1
RAS
Mutation in 2%
PI(3)K
Mutation in 15%
PTEN
Mutation, homozygous deletion in 36%
AKT
Amplification in 2%
Proliferation survival translation
FOXO
Mutation in 1%
b
p53 signalling altered in 87%
Activated oncogenes
CDKN2A (ARF)
Homozygous deletion, mutation in 49%
Amplification in 14%
MDM2
MDM4
Amplification in 7%
TP53
Mutation, homozygous deletion in 35%
Senescence
Apoptosis
c
CDKN2A (P16/INK4A)
Homozygous deletion, mutation in 52%
CDKN2B
Homozygous deletion in 47%
CDKN2C
Homozygous deletion in 2%
Amplification in 18%
Amplification in 2%
Amplification in 1%
CDK4
CCND2
CDK6
Homozygous deletion, mutation in 11%
RB1
G1/S progression
RB signalling altered in 78%

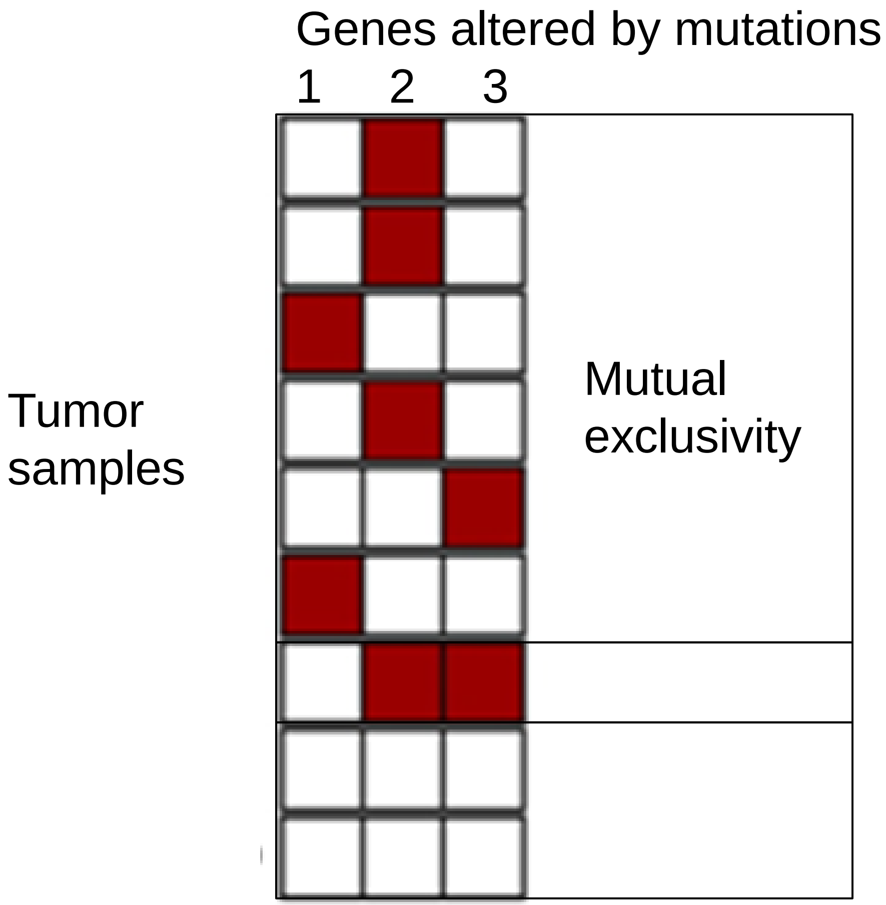
Aristotle’s Zoology: either horns or tusks

…[Aristotle] believed that **nature was economical** and gave no animal too many gifts, observing that **no animal possessed both horns and tusks**

Acquired capability of animals: defense	
Horns	Tusks



Discovering recurrent mutually exclusive patterns



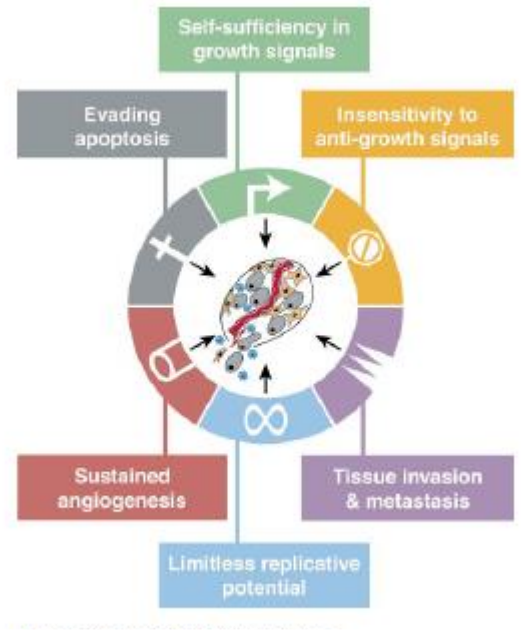
Genes altered by mutations

Tumor samples

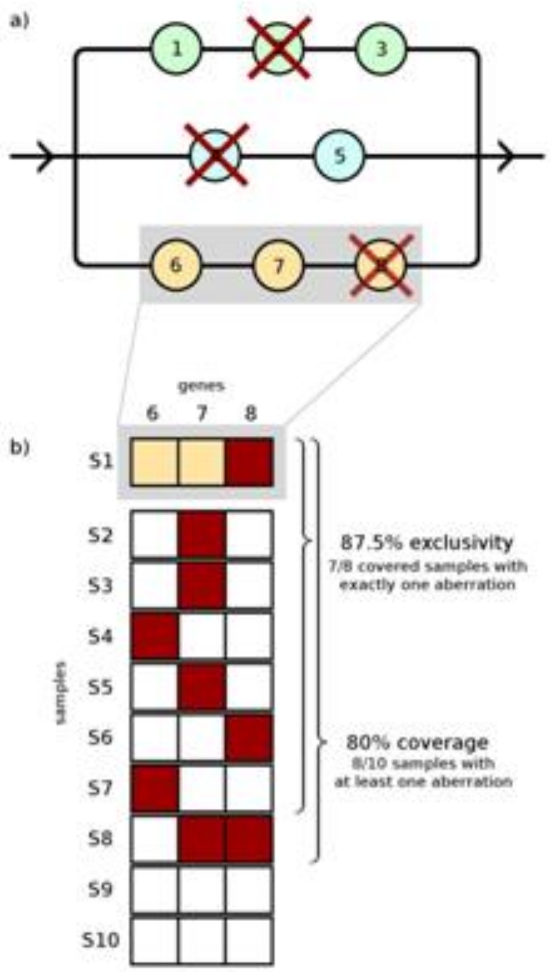
1	2	3
	🟥	
	🟥	
🟥		
	🟥	
		🟥
🟥		
	🟥	🟥

Mutual exclusivity

Discovering recurrent mutually exclusive patterns


Self-sufficiency in growth signals
Evading apoptosis
Insensitivity to anti-growth signals
Sustained angiogenesis
Tissue invasion & metastasis
Limitless replicative potential

(Hanahan and Weinberg, Cell 2001 and 2011)

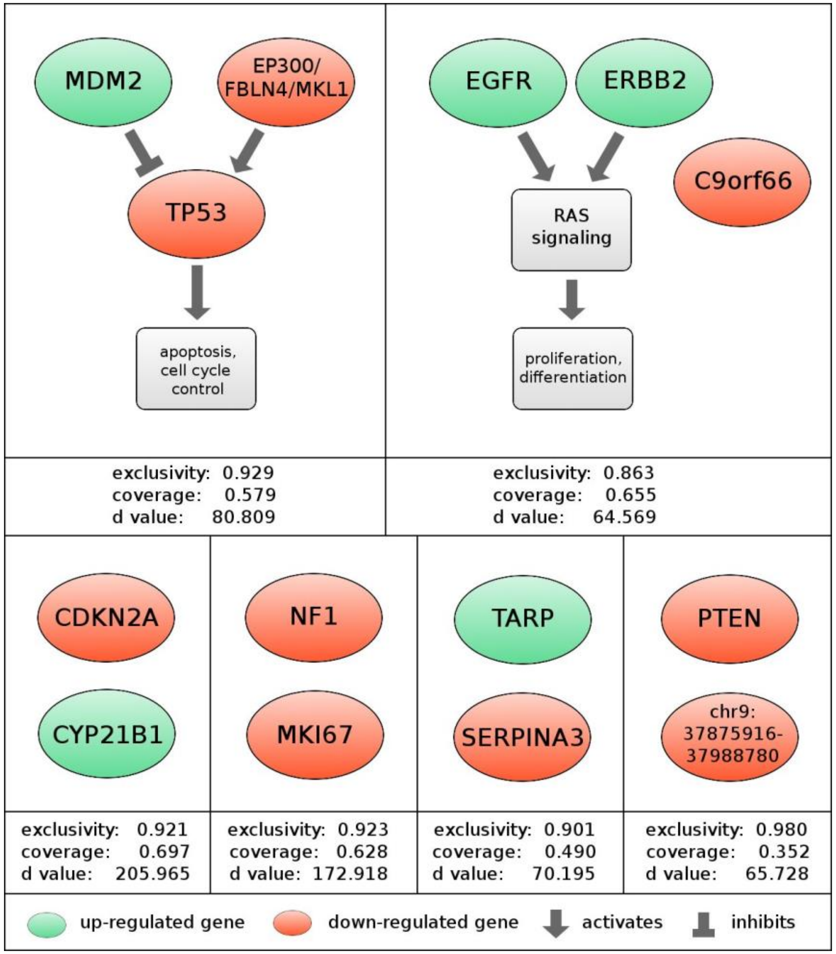

a)
1
3
5
6
7
b)
genes
6
7
8
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
samples
87.5% exclusivity
7/8 covered samples with exactly one aberration
80% coverage
8/10 samples with at least one aberration

Can key modules / pathways be discovered based on RME patterns alone ?

(Miller CA et al. BMC Med Genomics. 4(1):34 2011)

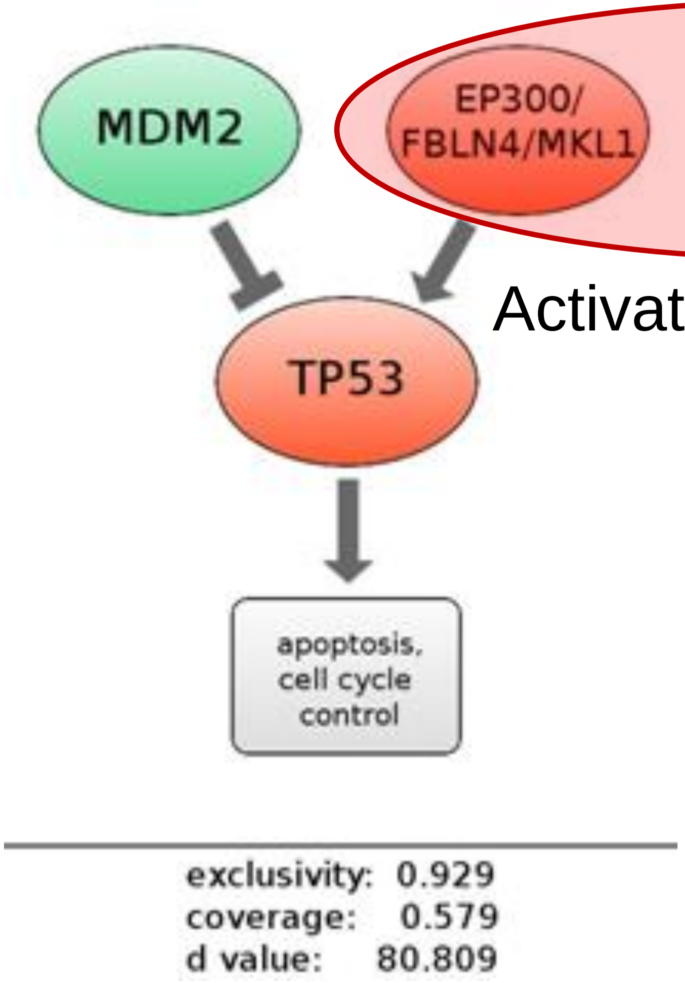
RME algorithm applied to glioblastoma TCGA collection (145 samples)

Rediscovered modules within all 3 pathways in glioblastoma reported by the TCGA Research Network, *Nature* **455**, 1061-1068 (2008).



(Miller CA et al. BMC Med Genomics. 4(1):34 2011)

RME algorithm applied to glioblastoma TCGA collection (145 samples)



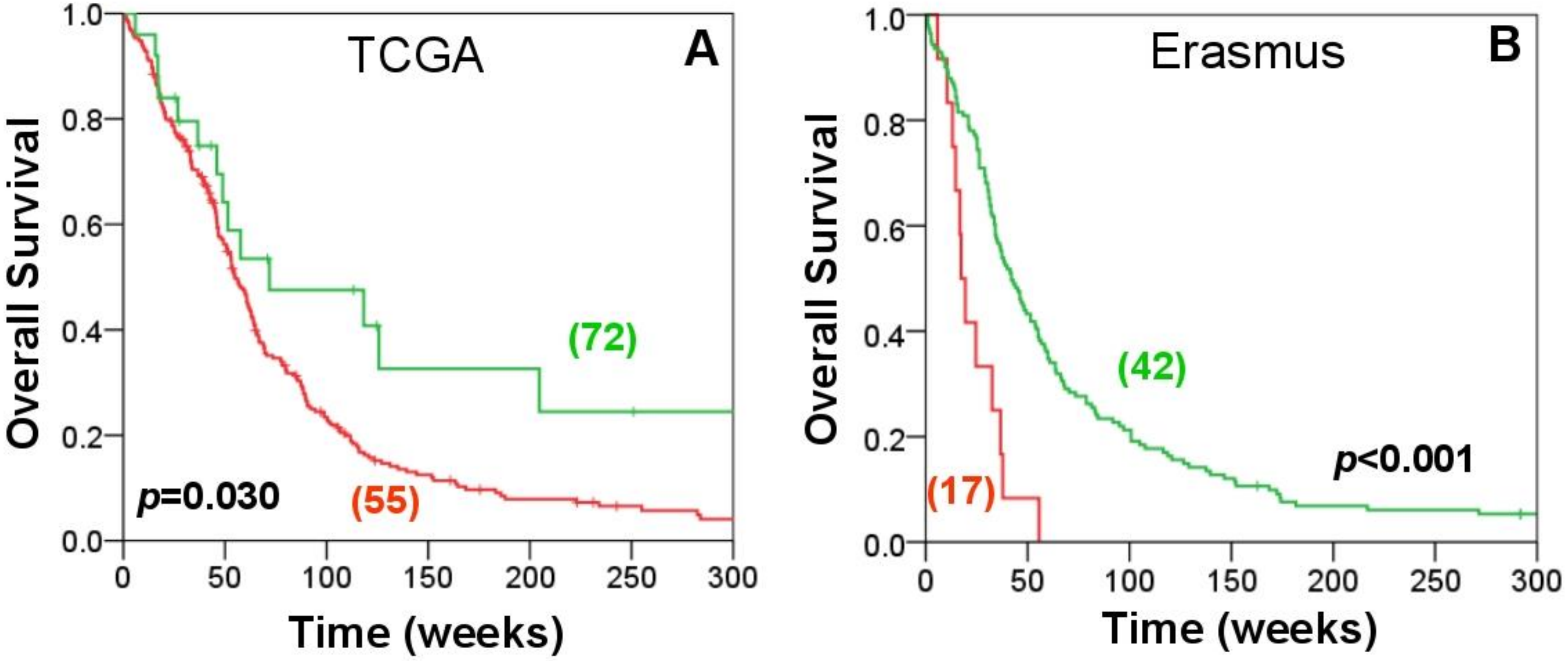
exclusivity: 0.929
coverage: 0.579
d value: 80.809

Role in Glioblastoma?

Activation by acetylation

(Miller CA et al. BMC Med Genomics. 4(1):34 2011)

Effect of *EP300* expression on grade-independent and age-independent survival in glioblastoma



(Miller CA et al. BMC Med Genomics. 4(1):34 2011)

Use Case #2: Discovering Recurrent Mutually Exclusive (RME) mutational patterns in cancer

Step 1: Identify variants associated with a specific phenotype (cancer)
> Disease ontology traversal

Step 2: Find groups of genes that form RME pattern

Step 3: Examine connectedness of the genes within pathways/networks
> Evaluate pattern of connectedness within networks or pathways

Topics

1. ClinGen Resource project

2. Building the Clinical Genome Knowledge Base

3. Linked Data technologies

4. Using Linked Data technologies to enable pattern discovery in the Clinical Genome Knowledge Base

Conclusion

The **Web of hyperlinks** (Web 1.0, 2.0) has greatly amplified the impact of the Human Genome Project

The **Web of Linked Data** (Web 3.0) will enable creation of a computable Clinical Genome Knowledge Base to inform

- Clinical interpretation of genomic variation
- Pattern discovery leading to new hypotheses

Acknowledgements

BCM Baylor College of Medicine

Sharon Plon
Andrew R. Jackson
Xin Feng
Ronak Patel
Rajarshi Ghosh

ClinGen
Clinical Genome Resource

National Human Genome Research Institute