

Gene Normalizer: A Tool to Resolve Genetic Ambiguity Through Data Harmonization

Anastasia Bratulin¹, James Stevenson², Kori Kuzma², Matthew Cannon², Wesley Goar², Alex Wagner²
¹The Ohio State University, Columbus, OH, ²The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH

Problem 1: Alias-Gene Intersections

Alias-Gene Intersections describe when the same gene symbol is used in a primary gene symbol category as well as an alias category for two distinct genes.

| HGNC Gene Symbol | Aliases |
|------------------|--|
| NRAS | NS6, CMNS, KRAS , N-ras, NCMS, NRAS1, ALPS4 |
| KRAS | NS, NS3, OES, CFC2, RALD, K-Ras, KRAS1, KRAS2... |

Figure 1. Alias-Gene Intersection illustration using two genes.

| Source | Total # of gene records | # of records with alias-gene intersections |
|----------------|-------------------------|--|
| Ensembl | 40354 | 266 (0.6%) |
| HGNC | 43164 | 483 (1.1%) |
| NCBI Gene Info | 75346 | 2394 (3.2%) |

Figure 2. Evidence of Alias-Gene Intersections across three different sources.

Problem 2: Alias-Alias Intersections

Alias-Alias Intersections describe when the same gene symbol is used in the alias category for multiple distinct genes.

| HGNC Gene Symbol | Aliases |
|------------------|-------------------------|
| IGHM | AGM1, MU, VH |
| IGHV6-1 | IGHV61, VH |
| SLC7A4 | HCAT3, CAT-4, VH |

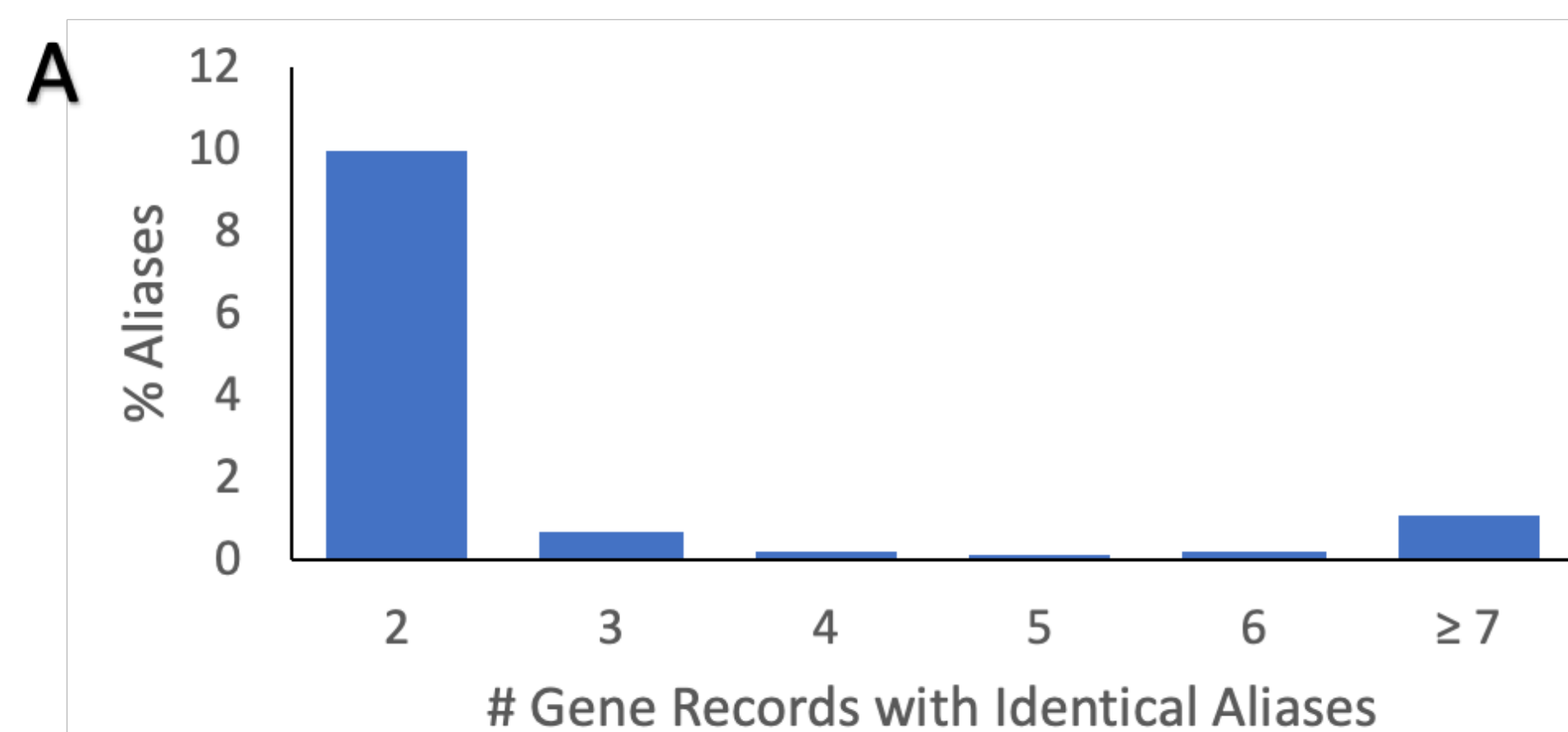
Figure 3. Alias- Alias Intersection illustration using three genes.

| Source | Total # of gene records | # of records with a shared alias |
|----------------|-------------------------|----------------------------------|
| Ensembl | 40354 | 3075 (7.6%) |
| HGNC | 43164 | 2084 (4.83%) |
| NCBI Gene Info | 75346 | 2957 (3.92%) |

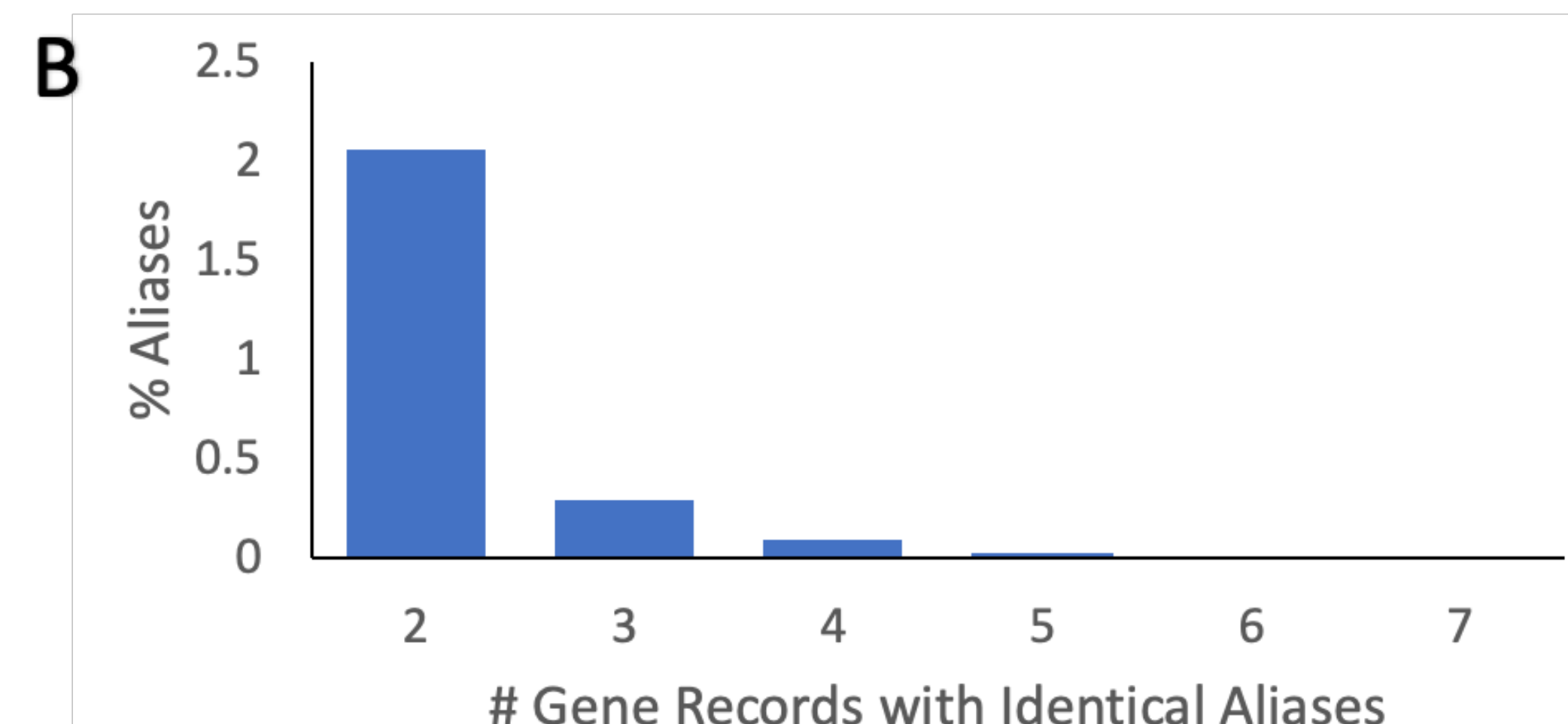
Figure 4. Evidence of Alias-Alias Intersections across three different sources.

Alias-Alias Intersection Distribution per Data Source

Ensembl



HGNC



NCBI Gene Info

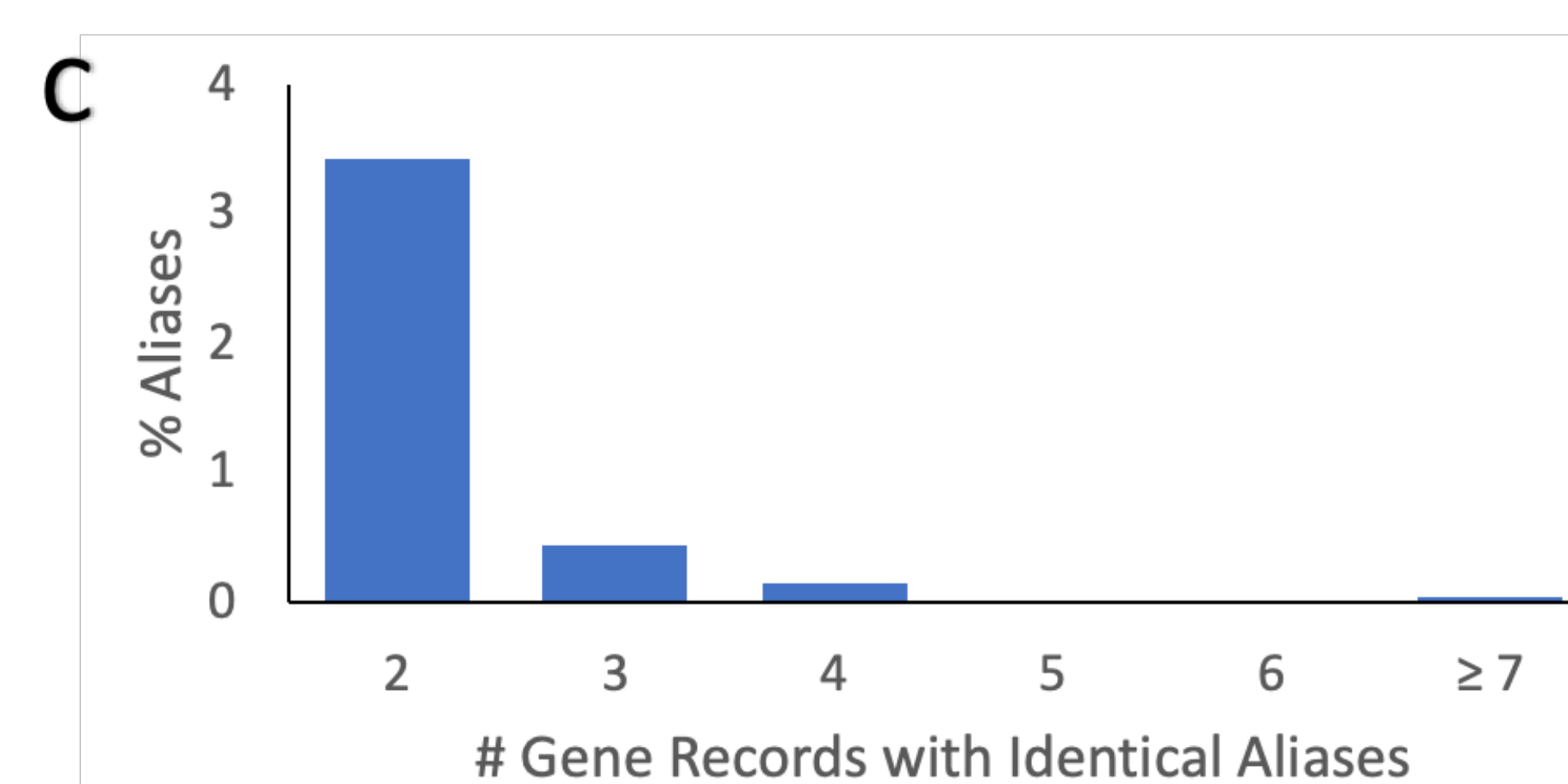


Figure 5. Percentage of aliases in each source that are shared amongst two gene records or more.

Future Directions: Gene Normalizer

Using policies that prioritize the gene symbol category (Problem 1) and identify aliases that are family groups (Problem 2), allows the Gene Normalizer to systematically blend data from different sources into one cohesive construct.

Gene Record from HGNC

Approved symbol ? BRAF

Approved name ? B-Raf proto-oncogene, serine/threonine kinase

Locus type ? gene with protein-coding transcript

HGNC ID ? HGNC:1097

Symbol status ? Approved

Previous names ? " v-raf murin

Alias symbols ? **BRAF1; BRAF-1**

Chromosomal location ? 7q34

Gene groups ? **RAF family**
Mitogen-act

Gene Record from NCBI

BRAF B-Raf proto-oncogene, serine/threonine kinase [*Homo sapiens* (human)]

Gene ID: 673, updated on 2-Jul-2023

Summary

Official Symbol BRAF provided by HGNC

Official Full Name B-Raf proto-oncogene, serine/threonine kinase provided by HGNC

Primary source HGNC:HGNC:1097

See related Ensembl:ENSG00000157764 MM:164757; AllianceGenome:HGNC:1097

Gene type protein coding

RefSeq status REVIEWED

Organism *Homo sapiens*

Also known as NS7; B-raf; BRAF1; RAFB1; B-RAF1; BRAF-1

Summary This gene encodes a protein belonging to the RAF family of serine/threonine protein kinases. It is involved in cell division, differentiation, and secretion. Mutations in this gene, most commonly the V600E mutation, are associated with various other cancers as well, including non-Hodgkin lymphoma, colorectal cancer, lung. Mutations in this gene are also associated with cardiocirculatory, Noonan, and Costello syndromes. [provided by RefSeq, Aug 2017]

Expression Ubiquitous expression in testis (RPKM 6.6), brain (RPKM 5.3) and 25 other tissues See more

Orthologs mouse all

Figure 6. Gene records from different sources.

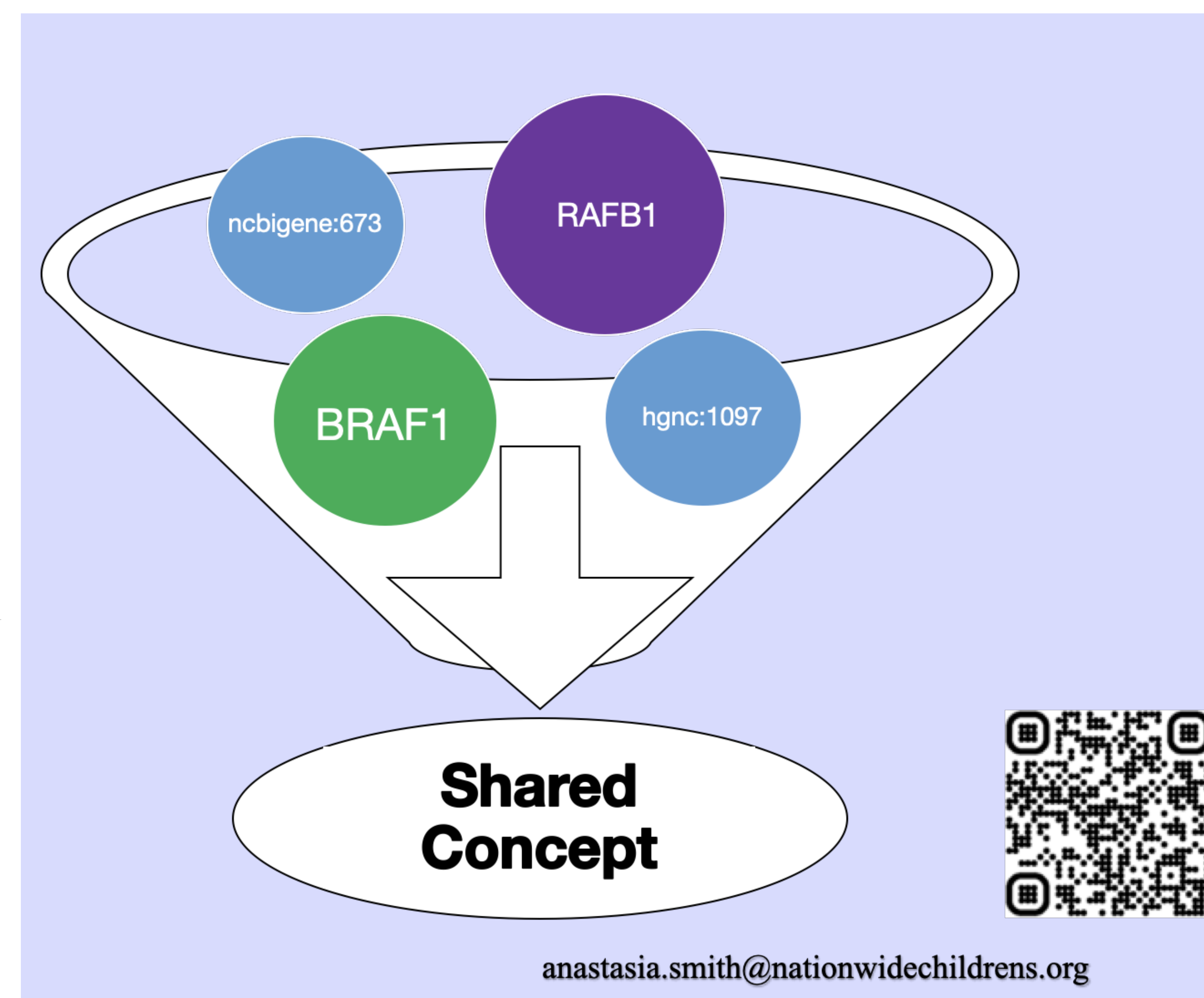


Figure 7. The Gene Normalizer - identifying a shared concept supports a publicly accessible normalization service at normalize.cancervariants.org/gene