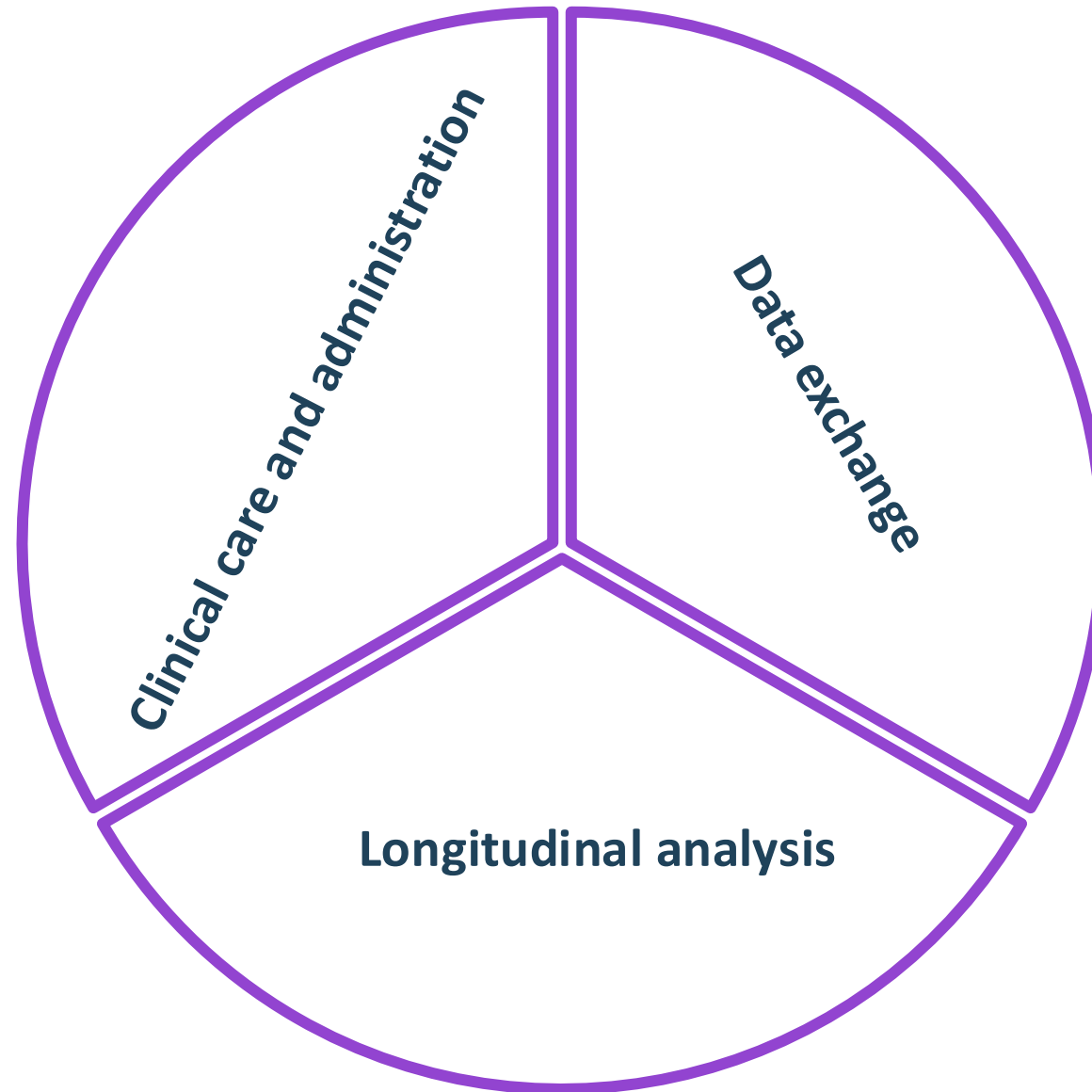
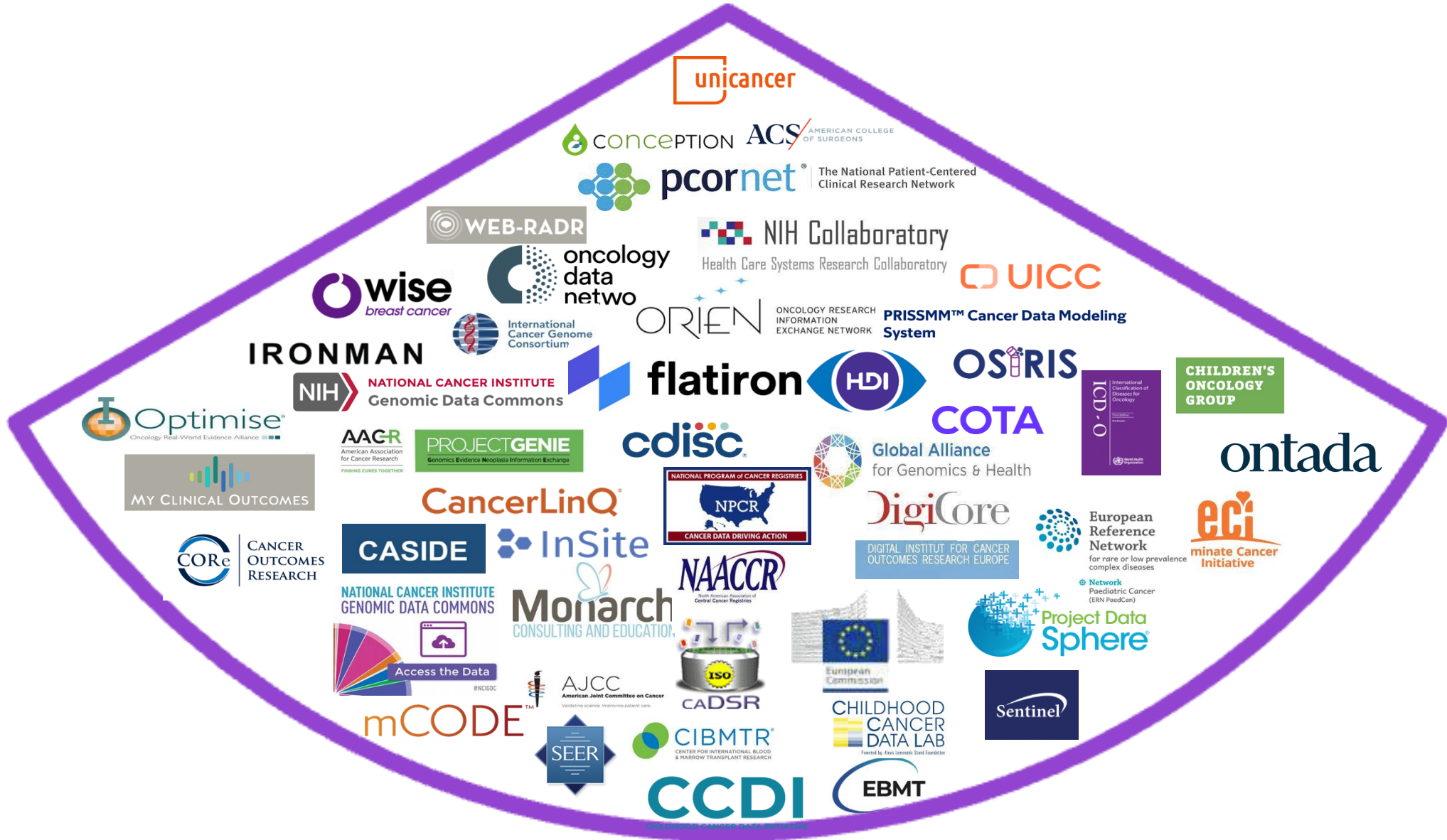




# Genomic Data in a Closed World Environment

Asieh Golozar, MD PhD MPH MHS



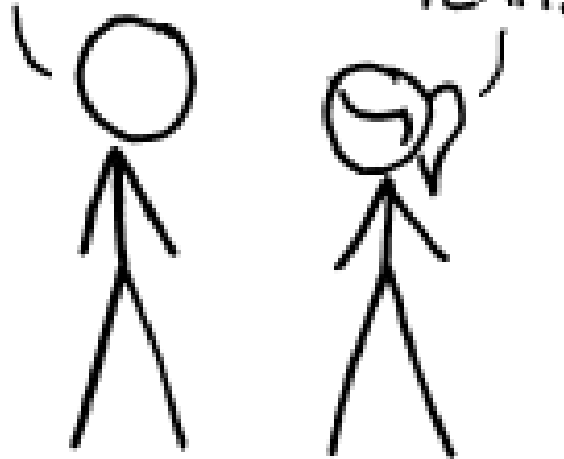




## HOW STANDARDS PROLIFERATE: (SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)

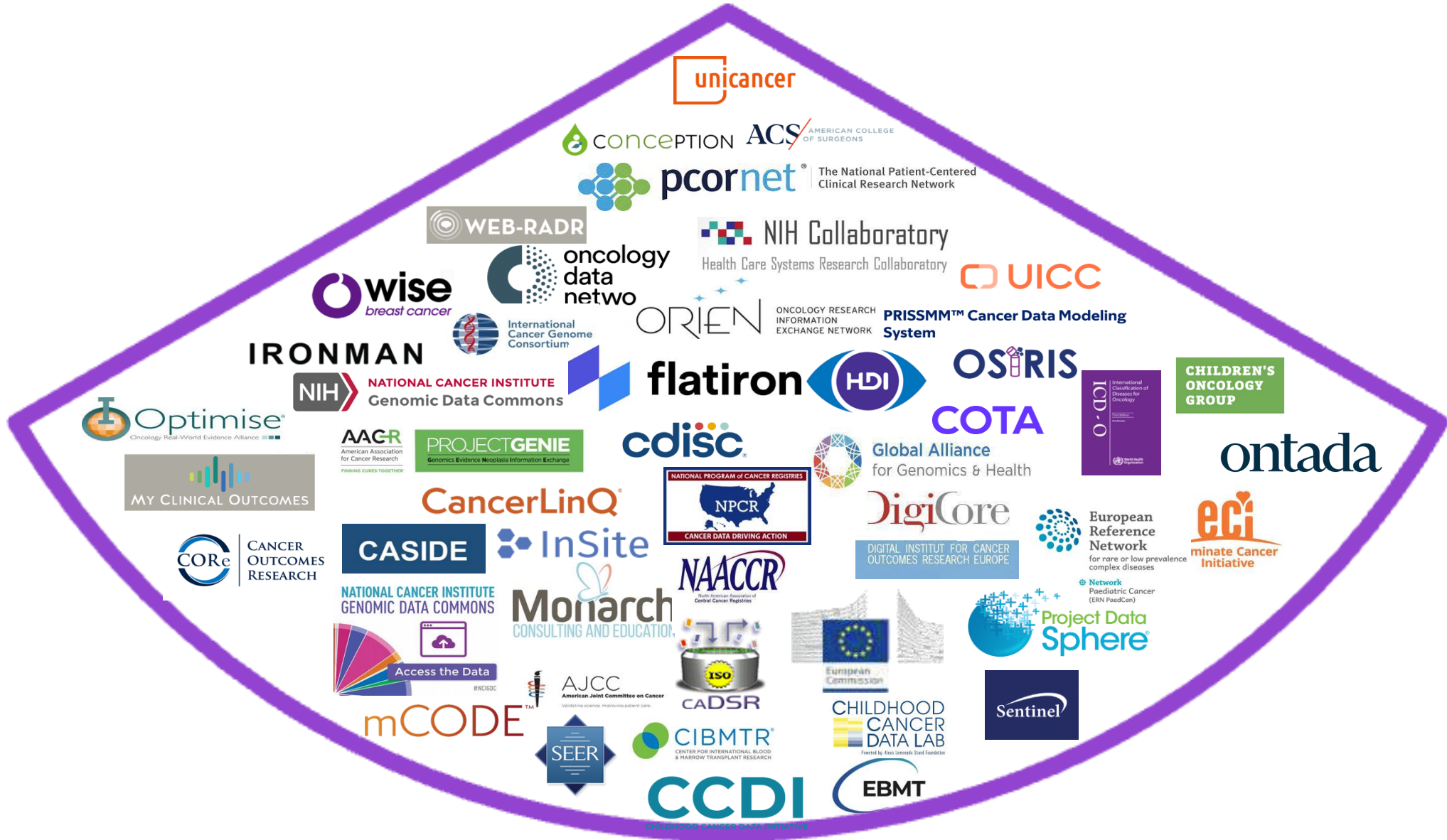
SITUATION:  
THERE ARE  
14 COMPETING  
STANDARDS.

14?! RIDICULOUS!  
WE NEED TO DEVELOP  
ONE UNIVERSAL STANDARD  
THAT COVERS EVERYONE'S  
USE CASES.



SOON:

SITUATION:  
THERE ARE  
15 COMPETING  
STANDARDS.





# Formal logic in knowledge representation

## Open-world assumption

1. What we know – is true
2. What we don't know – we don't know

## Closed-world assumption

1. What we know – is true
2. What we don't know – is false

→ RWD is Closed World **implicitly**

Example:

- ICD-10-CM has all diseases
- Our analyses, like rates, incidence, causal effect expect, we expect:
  - The code if the patient suffers from it
  - The code to be absent if not
- There is no such thing as “No MI today”.



# Closed and Open World Domains

## Closed

- Conditions, including
  - Histology
  - Topology
  - Metastases
  - Lymph nodes
  - Stages
  - Grades
- Drugs
- Procedures
- Devices

## Open

- Images
- Wave forms
- Omic variants
- Locations
- Survey answers
- Free text



So, if it is not a new standard,  
what do we need?

Genomic Data in a Closed World Environment





# Precision Oncology

If we ever want to apply epidemiological methods used in clinical research to the study genomic variation, it

... requires genomic variants that are:

- unique,
- comprehensive,
- searchable,
- which means, standardized



# However, genomic variant data...

- ... are not standardized for interoperability
- ... are too narrow when measured through pre-defined panels and overwhelmed with meaningless information when obtained through Next Generation Sequencing
- ... are obtained through a myriad of different detection methods
- ... occur in different modalities, which are somewhat interconnected



# They have their own world of representation

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT TUMOR NORMAL
1 156761535 1:156761535 T/T C 0 . NS=2;DP=885;DPB=885;6;AC=0;AN=6;AF=0;RO=860;AO=14;PRO=0;PAO=0;QR=33971;QA=561;PQR=0;PQA=0;SRF=416;SRR=444;SAF=7;SAR=7;SRP=4.98987;SAP=3.0103;AB=0;ABP=0;RUN=1;RPP=5.49198;RPPR=4.23238;RPL=9;R
R=5;EPP=18.5208;EPPR=10.3731;DPRA=2.26568;ODDS=248.622;GTI=0;TYPE=ins;CIGAR=1M19M;NUMALT=2;MEANALT=3;LEN=1;MQM=70;MQMR=70;PAIRED=0.928571;PAIREDR=0.965116;technology.ILLUMINA=1;OLD_VARIANT=1:156761535:TCCCCCCCCA/
TCCCCCCCCA;ANN=TC|frameshift_variant|HIGH|PRCC|PRCC|transcript|NM_005973.4|protein_coding|4/7|c.1138dupC|p.Gln380fs|1429/2123|1139/1476|380/491||INFO_REALIGN_3_PRIME;LOF=(PRCC|PRCC|1|1.00);AA=p.Q380fs*12
GT:GQ:DP:AD:RO:QR:AO:QA:SRF:SRR:SAF:SAR 0/0/0:160.002:614:592,6:592:23419:14:561:284:308:7:7 0/0/0:160.002:271:268,2:268:10552:0:0:132:136:0:0
10 123307688 10:123307688 C/T C T 0 . NS=2;DP=1057;DPB=1057;AC=0;AN=6;AF=0;RO=1040;AO=15;PRO=0;PAO=0;QR=42032;QA=615;PQR=0;PQA=0;SRF=724;SRR=316;SAF=15;SAR=0;SRP=35.5824;SAP=35.5824;AB=0;ABP=0;RUN=1;RPP=35.5824;RPPR=9.09877;RPL=0;
RPR=15;EPP=35.5824;EPPR=24.7334;DPRA=1.9691;ODDS=337.572;GTI=0;T
YPE=snp;CIGAR=1X;NUMALT=1;MEANALT=1;LEN=1;MQM=70;MQMR=70;PAIRED=0.066667;PAIREDR=0.503846;technology.ILLUMINA=1;ANN=T|intron_variant|MODIFIER|FGFR2|FGFR2|transcript|NM_022970.3|protein_coding|5/17|c.624+3116G>
A||||||,T|intron_variant|MODIFIER|FGFR2|FGFR2|transcript|NM_022970.3|protein_coding|3/15|c.357+3116G>
A||||||,T|intron_variant|MODIFIER|FGFR2|FGFR2|transcript|NM_022970.3|protein_coding|4/16|c.357+3116G>
A||||||,T|intron_variant|MODIFIER|FGFR2|FGFR2|transcript|NM_022970.3|protein_coding|5/15|c.624+3116G>
A||||||,T|intron_variant|MODIFIER|FGFR2|FGFR2|transcript|NM_022970.3|protein_coding|6/16|c.926+3116G>
GT:GQ:DP:AD:RO:QR:AO:QA:SRF:SRR:SAF:SAR 0/0/0:160.002:701:686,
11 92564925 11:92564925 G/T G T 0 . NS=2;DP=73
0;RPR=15;EPP=35.5824;EPPR=3.06621;DPRA=1.9521;ODDS=315.102;GTI=
E=snp;CIGAR=1X;NUMALT=1;MEANALT=1;LEN=1;MQM=70;MQMR=70;PAIRED=
4557|| GT:GQ:DP:AD:RO:QR:AO:QA:SRF:SRR:SAF:SAR 0/0/0:138.986:0
11 108196957 11:108196957 G/GT G GT 0 . NS=2;DP=73
R=3;EPP=4.78696;EPPR=18.0389;DPRA=1.78626;ODDS=197.677;GTI=0;T
GT|TTTTTTTA;ANN=GT|intron_variant|MODIFIER|ATM|ATM|transcript|
0:160.002:262:257,4:257:10295:0:0:97:160:0:0
12 111844023 12:111844023 G/T G 6.87107e-15 . NS
;RPR=0;EPP=5.18177;EPPR=22.3561;DPRA=1.08477;ODDS=47.333;GTI=0
CIGAR=1X;NUMALT=1;MEANALT=1;LEN=1;MQM=70;MQMR=70;PAIRED=0.75;P
GT:GQ:DP:AD:RO:QR:AO:QA:SRF:SRR:SAF:SAR 0/0/0:148.008:105:101,
13 114292584 13:114292584 G/T G T 0 . NS=2;DP=10
;RPR=0;EPP=31.2394;EPPR=3.18851;DPRA=2.125;ODDS=314.123;GTI=0;
PE=snp;CIGAR=1X;NUMALT=1;MEANALT=1;LEN=1;MQM=70;MQMR=69.9757;P
T|TTTT|,T|intron_variant|MODIFIER|TFDP1|TFDP1|transcript|NR_0
0:160.002:320:320,0:320:12696:0:0:157:163:0:0
16 30991329 16:30991329 C/A C A 9.07071e-14 . NS=2;D
4;EPP=3.0103;EPPR=3.38469;DPRA=1.0411;ODDS=36.7629;GTI=0;TYPE=
GAR=1X;NUMALT=1;MEANALT=1;LEN=1;MQM=70;MQMR=70;PAIRED=0.75;PAI
GT:GQ:DP:AD:RO:QR:AO:QA:SRF:SRR:SAF:SAR 0/0/0:136.801:76:72,
16 50745398 16:50745398 A/C A C A 0 . NS=2;DP=66
3.02379;RPL=6;RPR=6;EPP=3.73412;EPPR=3.2261;DPRA=1.69512;ODDS=
ACCCCCA;ANN=A|frameshift_variant|HIGH|NOD2|NOD2|transcript|NM_
/11|c.1502delC|p.Pro501fs|1648/4446|1502/3042|501/1013||INFO_
0:160.002:246:242,0:242:9726:0:0:117:125:0:0
16 67127271 16:67127271 GTTTT/G GTTTT G 384.394 . NS=2;DP=92;DPB=284.607;AC=6;AN=6;AF=1;RO=43;AO=5;PRO=75.4167;PAO=41.2;QR=1675;QA=157;PQR=288.82;PQA=1549.57;SRF=1;SRR=42;SAF=2;SAR=3;SRP=87.8997;SAP=3.44459;AB=0;ABP=0;RUN=1;RPP=
6.91895;RPPR=17.6046;RPL=4;RPR=1;EPP=3.44459;EPPR=21.2406;DPRA=2.83333;ODDS=
61.5392;GTI=0;TYPE=del;CIGAR=1M50D22M;NUMALT=1;MEANALT=11;LEN=5;MQM=70;MQMR=70;PAIRED=0.4;PAIREDR=0.55814;technology.ILLUMINA=1;OLD_VARIANT=16:67127271:GTTTTTTTTTTTTTTTTTTTTTTT/G
GT|TTTTTTTTTTTTTTTTTTT;ANN=G|intron_variant|MODIFIER|CBFB|CBFB|transcript|NM_022845.2|protein_coding|5/5|c.496-5320,496-5316delTTT|TTTT|INFO_REALIGN_3_PRIME GT:GQ:DP:AD:RO:QR:AO:QA:SRF:SRR:SAF:SAR 1/1/1:129.116:68:31,5:31:1210:5:157:0:31:2:3 1/1/1:129.116:24:12,0:12:465:0:0:1:11:0:0
17 40273273 17:40273273 C/G C G 0 . NS=2;DP=400;DPB=400;AC=0;AN=6;AF=0;RO=391;AO=9;PRO=0;PAO=0;QR=14902;QA=218;PQR=0;PQA=0;SRF=250;SRR=141;SAF=0;SAR=9;SRP=68.9931;SAP=22.5536;AB=0;ABP=0;RUN=1;RPP=22.5536;RPPR=222.94;RPL=0;RPR=9;EPP
22.5536;EPPR=18.6105;DPRA=2.27869;ODDS=125.177;GTI=0;TYPE=snp;CIGAR=1X;NUMALT=1;MEANALT=1;LEN=1;MQM=70;MQMR=70;PAIRED=0.713555;technology.ILLUMINA=1;ANN=G|mismatch_variant|MODERATE|BCL3|BCL3|transcript|NM_005178.4|protein_coding|7/9|c.976G>T|p.Arg326Cys|1046/1864|976/1365|326/
/18|c.506G>C|p.Arg17Pro|110/3112|50/2514|17/837||G|upstream_gene_variant|MODIFIER|HSPB9|HSPB9|transcript|NM_033194.2|protein_coding||c.-1596C>
G|11483|G|downstream_gene_variant|MODIFIER|RAB5C|RAB5C|transcript|NM_001252039.1|protein_coding||c.*4528G>C|113721|G|downstream_gene_variant|MODIFIER|RAB5C|RAB5C|transcript|NM_004583.3|protein_coding||c.*4528G>
C|113721|G|downstream_gene_variant|MODIFIER|RAB5C|RAB5C|transcript|NM_007052.1|pseudogene||n.-327C>T|113721| GT:GQ:DP:AD:RO:QR:AO:QA:SRF:SRR:SAF:SAR 0/0/0:141.459:125:120,5:120:4768:5:205:50:70:2:3 0/0/
0:160.002:122:122,0:122:4634:0:0:70:52:0:0
19 18888194 19:18888194 C/T C T 5.27969e-14 . NS=2;DP=407;DPB=407;AC=0;AN=6;AF=0;RO=396;AO=11;PRO=0;PAO=0;QR=15955;QA=424;PQR=0;PQA=0;SRF=141;SRR=255;SAF=0;SAR=11;SRP=74.2741;SAP=26.8965;AB=0;ABP=0;RUN=1;RPP=26.8965;RPPR=161.484;RPL=
11;RPR=0;EPP=26.8965;EPPR=17.8377;DPRA=2.13077;ODDS=101.941;GTI=0;TY
PE=snp;CIGAR=1X;NUMALT=1;MEANALT=1;LEN=1;MQM=70;MQMR=70;PAIRED=0.0909091;PAIREDR=0.924242;technology.ILLUMINA=1;ANN=T|3_prime_UTR_variant|MODIFIER|CRTCI|CRTCI|transcript|NM_001098482.1|protein_coding|15/15|c.*2C>
T|11|2|,T|3_prime_UTR_variant|MODIFIER|CRTCI|CRTCI|transcript|NM_015321.2|protein_coding|14/14|c.*2C>T|11|2| GT:GQ:DP:AD:RO:QR:AO:QA:SRF:SRR:SAF:SAR 0/0/0:139.152:277:266,11:266:10705:11:424:94:172:0:11 0/0/
0:139.152:130:130,0:130:5250:0:0:47:83:0:0
19 45261587 19:45261587 C/T C T 3.10393e-14 . NS=2;DP=187;DPB=187;AC=0;AN=6;AF=0;RO=182;AO=5;PRO=0;PAO=0;QR=7252;QA=205;PQR=0;PQA=0;SRF=79;SRR=103;SAF=2;SAR=3;SRP=9.88265;SAP=3.44459;AB=0;ABP=0;RUN=1;RPP=3.44459;RPPR=3.2012;RPL=2;RPR
=3;EPP=3.44459;EPPR=3.05802;DPRA=2.01613;ODDS=39.4674;GTI=0;TYPE=snp
;CIGAR=1X;NUMALT=1;MEANALT=1;LEN=1;MQM=70;MQMR=70;PAIRED=1;PAIREDR=0.994505;technology.ILLUMINA=1;ANN=T|mismatch_variant|MODERATE|BCL3|BCL3|transcript|NM_005178.4|protein_coding|7/9|c.976G>T|p.Arg326Cys|1046/1864|976/1365|326/
454||T|upstream_gene_variant|MODIFIER|MIR8085|MIR8085|transcript|NR_107052.1|pseudogene||n.-327C>T|113721| GT:GQ:DP:AD:RO:QR:AO:QA:SRF:SRR:SAF:SAR 0/0/0:141.459:125:120,5:120:4768:5:205:50:70:2:3 0/0/
0:160.002:62:62,0:62:2484:0:0:29:33:0:0
2 121746517 2:121746517 G/A G A 1.29283e-14 . NS=2;DP=261;DPB=261;AC=0;AN=6;AF=0;RO=255;AO=6;PRO=0;PAO=0;QR=10173;QA=238;PQR=0;PQA=0;SRF=129;SRR=126;SAF=3;SAR=3;SRP=3.08694;SAP=3.0103;AB=0;ABP=0;RUN=1;RPP=4.45795;RPPR=7.51504;RPL=4;R
PR=2;EPP=3.0103;EPPR=3.22319;DPRA=1.45226;ODDS=55.4693;GTI=0;TYPE=sn
PR=2;CIGAR=1X;NUMALT=1;MEANALT=1;LEN=1;MQM=70;MQMR=70;PAIRED=1;PAIREDR=0.976471;technology.ILLUMINA=1;ANN=A|synonymous_variant|LOW|GLT2|GLT2|transcript|NM_005270.4|protein_coding|13/13|c.3027G>A|p.Leu1009Leu|3057/6769|3027/4761|1009/1586||
GT:GQ:DP:AD:RO:QR:AO:QA:SRF:SRR:SAF:SAR 0/0/0:145.263:155:149,6:149:5937:6:238:75:74:3:3 0/0/0:145.263:106:106,0:106:4236:0:0:54:52:0:0
20 32077959 20:32077959 C/T C T 0.1495 . NS=2;DP=100;DPB=100;AC=1;AN=6;AF=0;RO=96;AO=4;PRO=0;PAO=0;QR=3801;QA=164;PQR=0;PQA=0;SRF=51;SRR=45;SAF=2;SAR=2;SRP=3.8246;SAP=3.0103;AB=0.0909091;ABP=66.97;RUN=1;RPP=5.18177;RPPR=7.443
72;RPL=3;RPR=1;EPP=5.18177;EPPR=4.45795;DPRA=0.785714;ODDS=3.351
```







# → ETLing is easier said than done

## We get:

- Different modalities, which are often mixed
- Different sources
- Different detection methods
- Flat files with complex syntax requirements but lax compliance



## We need:

- Precise variants
- Standardized variants
- Meaningful hierarchy
- Closed World assumption



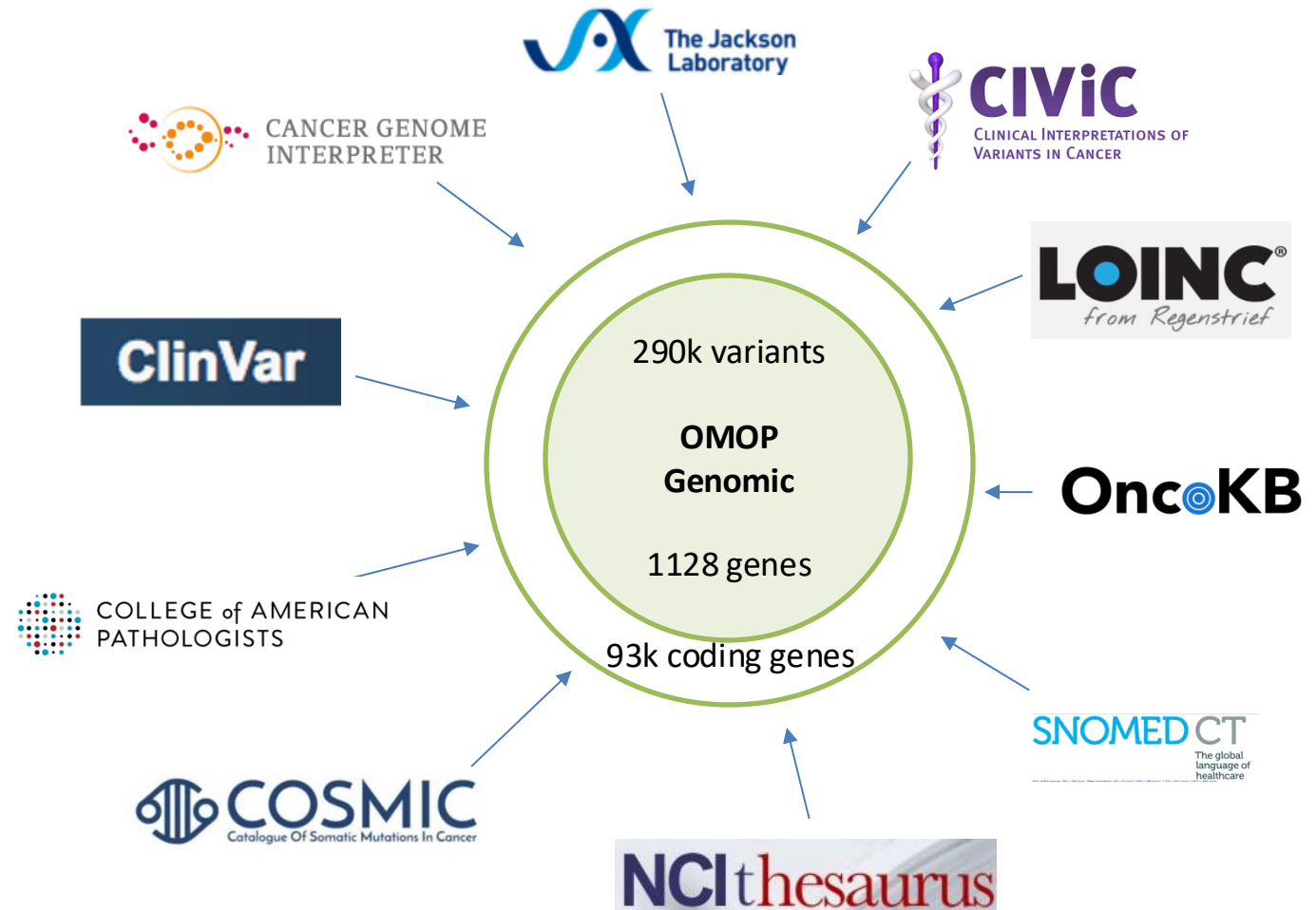
## Recording Variants in the OMOP CDM

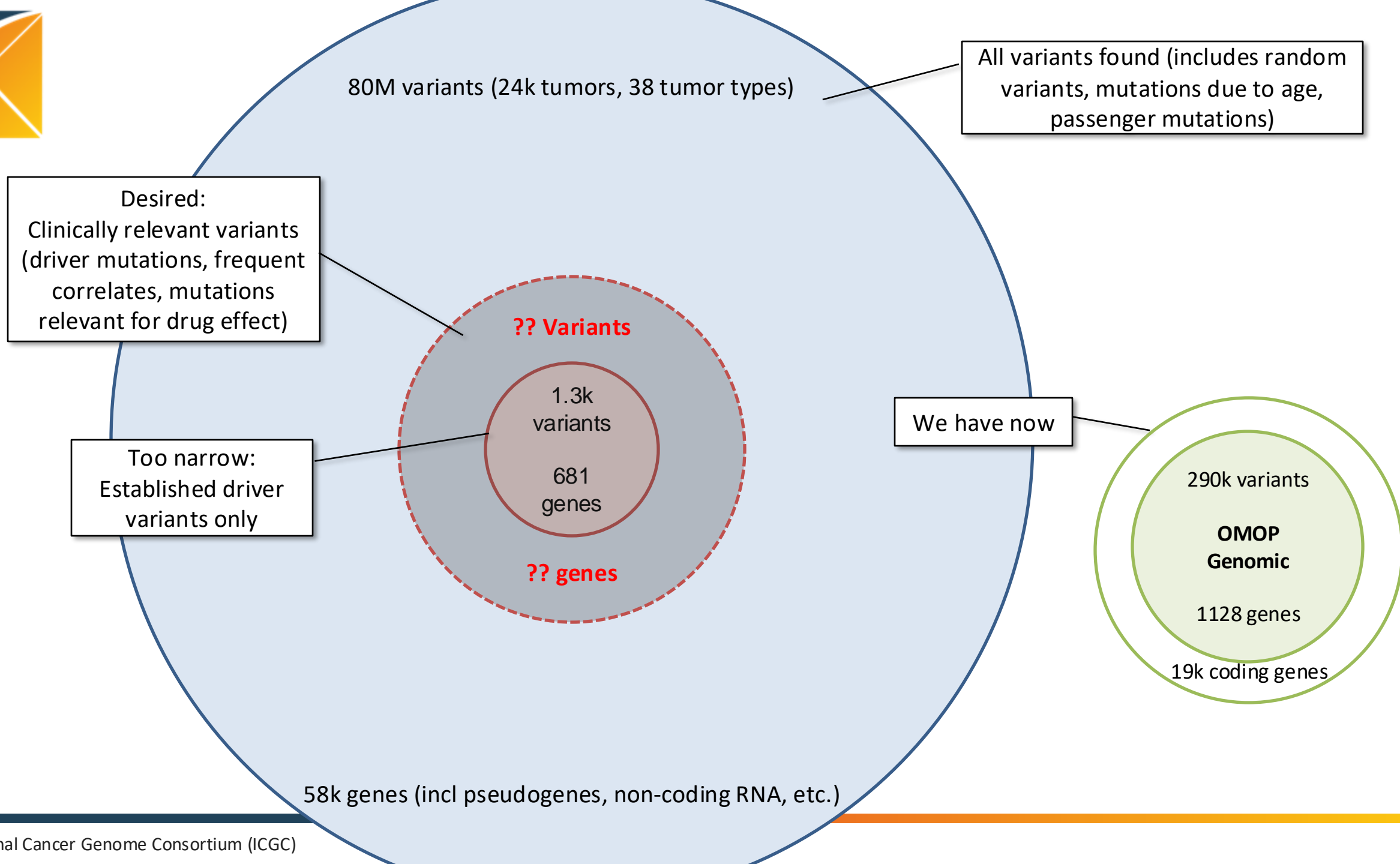


# OMOP Genomic is built from relevant sources

... by

- Combining public repositories
- Deduping them

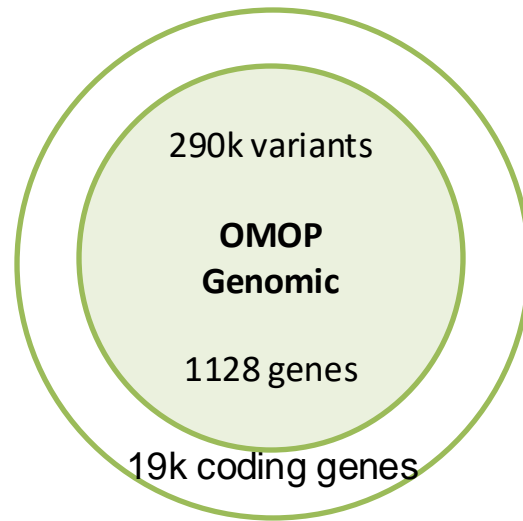








# OMOP Genomic contains



Genetic Variation: 19,297

Structural Variant: 3,226

Gene DNA Variant: 83,460

Gene RNA Variant: 92,988

Gene Protein Variant: 89,795



# Hierarchical relationships inside OMOP Genomic

Logical gene

DNA

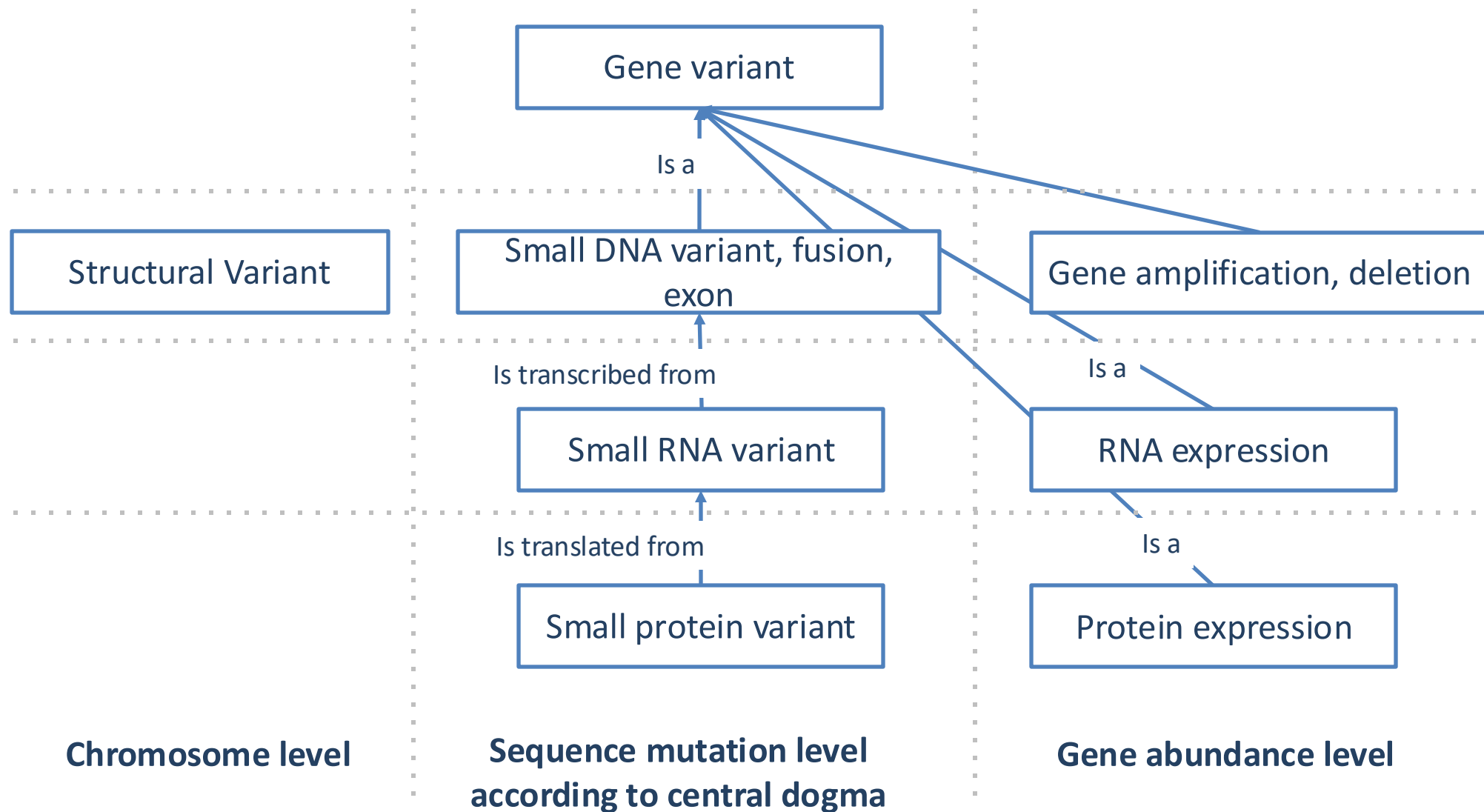
RNA

Protein

Chromosome level

Sequence mutation level  
according to central dogma

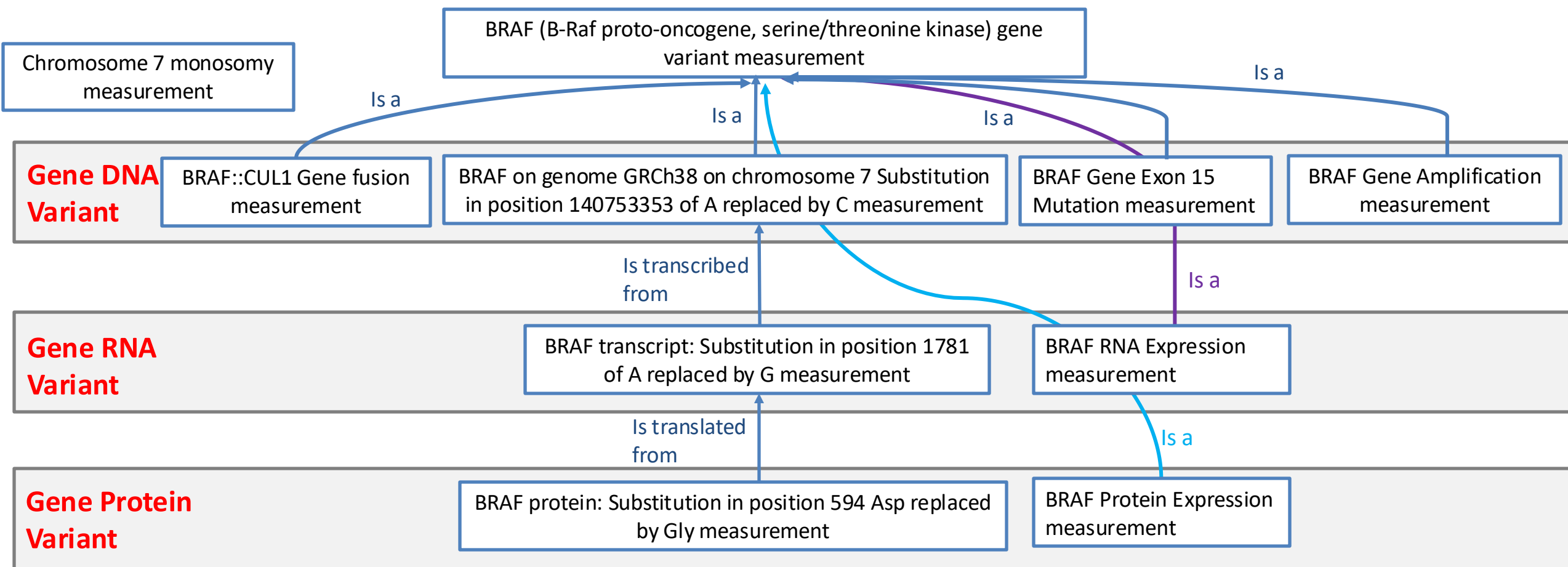
Gene abundance level





# Hierarchical relationships inside OMOP Genomic

## EXAMPLE





# Summary

1. Genomic variants must enter epidemiological research.
2. For that, they need to become uniquely identifiable features.
3. For that, we need them to be
  - Well, unique
  - Well, identifiable
  - Comprehensive (to cancer)
  - Finite
4. OMOP gives you all that.
5. [Join the Journey](#)

= Closed World