Using Machine Learning and Expert Human Guidance to Automate Clinical Data Integration to CDISC Standards

Timothy Danford, Tamr Inc.

Files, Names, & Standards

Please Don’t Make a Programmer Cry

File Formats are Complicated

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
1       13372   .       G       C       608.91  PASS    AC=3;AC_AFR=0;AC_AMR=0;AC_Adj=2;AC_EAS=0;AC_FIN=0;AC_Het=0;AC_Ho
m=1;AC_NFE=0;AC_OTH=0;AC_SAS=2;AF=6.998e-05;AN=42870;AN_AFR=770;AN_AMR=134;AN_Adj=8432;AN_EAS=254;AN_FIN=16;AN_NFE=2116;
AN_OTH=90;AN_SAS=5052;BaseQRankSum=0.727;ClippingRankSum=1.15;DP=139843;FS=0.000;GQ_MEAN=12.48;GQ_STDDEV=15.18;Het_AFR=0
;Het_AMR=0;Het_EAS=0;Het_FIN=0;Het_NFE=0;Het_OTH=0;Het_SAS=0;Hom_AFR=0;Hom_AMR=0;Hom_EAS=0;Hom_FIN=0;Hom_NFE=0;Hom_OTH=0
;Hom_SAS=1;InbreedingCoeff=-0.0844;MQ=35.72;MQ0=0;MQRankSum=0.727;NCC=60853;QD=23.42;ReadPosRankSum=0.727;VQSLOD=-1.687e
+00;culprit=MQ;DP_HIST=14728|2455|2120|518|121|499|534|314|111|21|10|2|2|0|0|0|0|0|0|0,1|0|0|0|1|0|0|0|0|0|0|0|0|0|0|0
|0|0|0;GQ_HIST=1012|14971|172|100|3161|259|127|30|8|9|5|16|1162|274|59|45|17|2|3|3,0|0|0|0|0|0|1|0|0|0|0|0|0|0|1|0|0|0|0|0
|0;CSQ=C|ENSG00000223972|ENST00000456328|Transcript|non_coding_transcript_exon_variant&non_coding_transcript_variant|620
||||||1||1|DDX11L1|HGNC|37102|processed_transcript|YES|||||||||3/3|||ENST00000456328.2:n.620G>C|||||||||||||||||||,C|ENSG
00000223972|ENST00000450305|Transcript|splice_region_variant&non_coding_transcript_exon_variant&non_coding_transcript_va
riant|412||||||1||1|DDX11L1|HGNC|37102|transcribed_unprocessed_pseudogene|||||||||5/6|||ENST00000450305.2:n.412G>C||||||
|||||||||||||,C|ENSG00000223972|ENST00000515242|Transcript|non_coding_transcript_exon_variant&non_coding_transcript_vari
ant|613||||||1||1|DDX11L1|HGNC|37102|transcribed_unprocessed_pseudogene|||||||||3/3|||ENST00000515242.2:n.613G>C|||||||
|||||||||||,C|ENSG00000223972|ENST00000518655|Transcript|intron_variant&non_coding_transcript_variant|||||||1||1|DDX11L1
|HGNC|37102|transcribed_unprocessed_pseudogene|||||||||||2/3||ENST00000518655.2:n.482-31G>C|||||||||||||||||||,C||ENSR000
00528767|RegulatoryFeature|regulatory_region_variant|||||||1||||||regulatory_region|||||||||||||||||||||||||||||||
1       13380   .       C       G       7829.15 VQSRTrancheSNP99.60to99.80      AC=41;AC_AFR=22;AC_AMR=1;AC_Adj=23;AC_EA
S=0;AC_FIN=0;AC_Het=23;AC_Hom=0;AC_NFE=0;AC_OTH=0;AC_SAS=0;AF=1.209e-03;AN=33902;AN_AFR=578;AN_AMR=104;AN_Adj=6922;AN_EA
S=192;AN_FIN=8;AN_NFE=1684;AN_OTH=74;AN_SAS=4282;BaseQRankSum=0.018;ClippingRankSum=0.264;DP=106203;FS=4.314;GQ_MEAN=12.
26;GQ_STDDEV=20.80;Het_AFR=22;Het_AMR=1;Het_EAS=0;Het_FIN=0;Het_NFE=0;Het_OTH=0;Het_SAS=0;Hom_AFR=0;Hom_AMR=0;Hom_EAS=0;
Hom_FIN=0;Hom_NFE=0;Hom_OTH=0;Hom_SAS=0;InbreedingCoeff=-0.0794;MQ=31.93;MQ0=0;MQRankSum=0.00;NCC=67643;QD=8.10;ReadPosR
ankSum=-1.580e-01;VQSLOD=-3.041e+00;culprit=MQ;DP_HIST=11485|1978|1863|671|104|288|337|152|58|11|2|2|0|0|0|0|0|0|0,3|1
1|7|6|5|2|1|1|0|0|0|1|0|0|0|0|0|0|0|0;GQ_HIST=1048|11415|159|57|2884|298|140|17|3|5|1|4|570|210|55|45|16|3|4|17,0|2|1|0|
2|0|0|0|0|4|0|4|2|0|1|1|2|1|1|16;CSQ=G|ENSG00000223972|ENST00000456328|Transcript|non_coding_transcript_exon_variant&non
_coding_transcript_variant|628||||||1||1|DDX11L1|HGNC|37102|processed_transcript|YES|||||||||3/3|||ENST00000456328.2:n.62
8C>G|||||||||||||||||||,G|ENSG00000223972|ENST00000450305|Transcript|splice_region_variant&intron_variant&non_coding_tra
nscript_variant|||||||1||1|DDX11L1|HGNC|37102|transcribed_unprocessed_pseudogene|||||||||5/5||ENST00000450305.2:n.414+6
C>G|||||||||||||||||||,G|ENSG00000223972|ENST00000515242|Transcript|non_coding_transcript_exon_variant&non_coding_transc
ript_variant|621||||||1||1|DDX11L1|HGNC|37102|transcribed_unprocessed_pseudogene|||||||||3/3|||ENST00000515242.2:n.621C>
G|||||||||||||||||||,G|ENSG00000223972|ENST00000518655|Transcript|intron_variant&non_coding_transcript_variant|||||||1||
1|DDX11L1|HGNC|37102|transcribed_unprocessed_pseudogene|||||||||||2/3||ENST00000518655.2:n.482-23C>G|||||||||||||||||||,G
||ENSR00000528767|RegulatoryFeature|regulatory_region_variant|||||||1||||||regulatory_region|||||||||||||||||||||||||||||||
|||
```

Format Specification is Hard

```
     /*-------------------------------------------------------------*/
     /* Name: ieee2xpt                                              */
     /* Purpose: converts IEEE to transport                         */
     /* Usage: rc = ieee2xpt(to_ieee,p_data);                       */
     /* Notes: this routine is an adaptation of the wzctdbl routine */
     /* from the Apollo.                                            */
     /*-------------------------------------------------------------*/

void ieee2xpt(ieee,xport)

unsigned char *ieee; /* ptr to IEEE field (2-8 bytes) */
   unsigned char *xport; /* ptr to xport format (8 bytes) */
   {

 register int shift;
   unsigned char misschar;
   int ieee_exp;
   unsigned long xport1,xport2;
   unsigned long ieee1 = 0;
   unsigned long ieee2 = 0;
```

Are Your Formats Self-Describing?

- Are you formats even defined at all?
  - Programmers have a lot of tools to describe formats, they're called **grammars & semantics**
- Where is the format "flexible?"
  - Your data is not a **special snowflake**.
- Can the format describe itself?
  - How much communication is **out-of-band** (i.e. outside the format itself)?

Naming Is Complicated

Reference and alternative alleles of a multi nucleotide polymorphism (MNP)

```
REF    GGGCATGGG
ALT    GGGTGCGGG
```

Genome Reference		Variant Call Format			
	GGGGCATGGGG	POS	REF	ALT	
REF	GCAT	4	GCAT	GTGC	Not left trimmed
ALT	GTGC				
REF	CATG	5	CATG	TGCG	Not right trimmed
ALT	TGCG				
REF	GCATG	4	GCATG	GTGCG	Not left and right trimmed
ALT	GTGCG				
REF	CAT	5	CAT	TGC	Normalized
ALT	TGC				

Alleles represented against the human genome reference. Allele pairs are colored the same, all are representations of the same variant.

Alleles represented in Variant Call Format, all are representations of the same variant.

http://genome.sph.umich.edu/wiki/Variant_Normalization

Study Data Tabulation Model Implementation Guide:
Pharmacogenomics/Genetics
Version 1.0 (Provisional)

Prepared by the
CDISC PGx Team

[redacted]

Notes to Readers

- This the provisional version 1.0 of the implementation guide for Pharmacogenomics and Pharmacogenetics. It is intended to correspond to version 1.5 of the CDISC Study Data Tabulation Model.
- Because SDTM v1.5 has not yet been published, this document remains a provisional release only.

"Multiple Names in Your Standard" Means You Have Multiple Standards

6. Genetic Variation:
    a. PFTESTCD and PFTEST generally identify the type of test performed by specifying the type of material assessed, such as nucleotides or amino acid, or the level of granularity, such as codon or allele. See the table below for suggested PFTESTCD values:

PFTESTCD	PFTEST	Notes
NUC	Nucleotide	Observes nucleotide sequences or values. Performed on DNA or RNA.
CDN	Codon	Observes nucleotide values, reported in groups of three, in which the position of the third nucleotide is divisible by three. Performed on DNA or RNA.
AA	Amino Acid	Observes amino acid sequences or values. Performed on proteins, or inferred from nucleotide (or codon) results.
ALE	Allele	Identifies the allele (version) of the gene/genetic region of interest. The result is the name for that allele according to the gene's nomenclature committee.

    b. --STRESC holds a "standardized" result. The usual cases where STRESC is different from ORRES are unit conversion and scoring of results (as in questionnaires). PF uses a different convention, where STRESC is the result in a format from an established nomenclature. This standard result often includes information that is parsed out into several SDTM variables, such as those for genetic location (PFGENLOC), observed result (PFORRES), and expected result (PFORREF).
        i. Nucleotide and amino acid location (position) values in PFGENLOC should follow the same rules as in PFSTRESC.
        ii. Unless a more appropriate nomenclature exists, the standard format for nucleotide and amino acid results in PFSTRESC follows the rules of HGVS nomenclature.
        iii. When PFTESTCD is ALE, results in PFSTRESC should follow the nomenclature system specified by the relevant gene's committee.

HGVS is a controversial subject in research genomics!

If You Can't Name It, You Don't Know It

- Naming is Important!
  - Inevitably, someone is going to set up a database and make a name into a **key**

- Not All Names Are Created Equal

  normalized?

  unique?

  structured?

- External standards may feel like a constraint
  - **But they are a blessing in disguise.**

Standards vs. Research

Daniel MacArthur
@dgmacarthur

Follow

Standards are defined by those who have the greatest tolerance for interminably long conference calls about standards.

RETWEETS 88 LIKES 95

6:05 AM - 17 Nov 2015

What Should You Do?

1. File Formats are not “a detail.”
2. It’s easy to get Naming wrong.
3. Standards Bodies need to include basic researchers as much as possible.

How can we organize our data for both **today** and **tomorrow**?

Please Don’t Make Your Programmers Cry