H. Pan
September 30, 2009

Data Set Integrity Check
CRISP Genetics Study

Reference Publication: Rossetti, S. et al. for the CRISP Study Group. Comprehensive molecular diagnostics in autosomal dominant polycystic kidney disease, **J Am Soc Nephrol.** 2007 Jul;18(7):2143 -60.

The Consortium for Radiological Imaging Studies of Polycystic Kidney Disease (CRISP) studied the progression of autosomal dominant polycystic kidney disease (ADPKD), and compared radiological techniques for measuring increases in renal volume during the progression of ADPKD. The CRISP study tested whether magnetic resonance (MR) can detect changes in renal volume, cyst volume, or changes in % cystic involvement in ADPKD individuals over a short period of time (1 to 2 years). As a partial check of the integrity of the CRISP genetics datasets archived in the NIDDK data repository, a dataset integrity check (DSIC) was performed to verify that selected published results from the CRISP papers can be reproduced using the archived datasets. The DSIC consists of a small number of analyses performed to duplicate published results on molecular analysis of the mutation classification for the cohort of 202 probands with ADPKD, as reported by Rossetti, S. et al in 2007 in the **J Am Soc Nephrol.** Results of the DSIC are described below.

The intent of this DSIC is to provide confidence that the data distributed by the NIDDK repository is a true copy of the study data. Our intent is ***not*** to assess the integrity of the data analyses reported by study investigators. As with all data analyses of complex datasets, complete replication of a set of analysis results should not be expected on a first (or second) exercise in secondary analysis. This occurs for a number of reasons including differences in the handling of missing data, restrictions on cases included in samples for a particular analysis, software coding used to define complex variables, etc. Experience suggests that most discrepancies can ordinarily be resolved by consultation with the study DCC; however this process is labor-intensive for both DCC and Repository staff. It is thus not our policy to resolve every discrepancy that is observed in an integrity check. Thus, we do not attempt to resolve minor or inconsequential discrepancies with published results or discrepancies that involve complex analyses unless staff of the NIDDK Repository suspect that the observed discrepancy suggests that the dataset may have been corrupted in storage, transmission, or processing by repository staff. We do, however, document in footnotes to the integrity check those instances in which our secondary analyses produced results that were not fully consistent with those reported in the target publication.

I. Background

The Genetic study is an ancillary study for CRISP. The Genetics data was provided to the repository by the ancillary study Principal Investigator and includes:

1) “Athena” genetics mutation report,
2) Raw chromatogram files sequenced in “Athena”. These records have a 7-digit “Athena ID”, which can be mapped to the 6-digit “CRISP ID”. This mapping file is included as part of the data archive.

3) A summary mutation data file, “CRISP mutation data 92007.xls”, for 230 participants

Analysis datasets used for published results were not provided at the time of this DSIC. Dr. Harris was available for general questions regarding study and data protocols including the genomic sequences for PKD1 and PKD2; however, reconstruction of the analysis dataset was left primarily to Repository analysts. Results of analyses on the reconstructed analysis dataset were compared to published results. Although there is only one table in the publication, this table summarizes the results listed from 12 tables in the Appendix. We performed a 2-step process for this replication to assure the quality of the archived datasets:

- Examine that data for all the subjects (by CRISP ID) is present
- Select randomly 1 reported mutation for each of the PKD1 and PKD2 genes from each of the 12 tables, with a total of 24 mutations to be examined

II. Results

1. Assess if raw data for each subject is present

To assess if raw data for each subject is present, we checked the IDs of raw data against the genetic study summary report - Athena Report. While majorities of the raw data are chromatograms, the raw data for eight subjects are image files of Multiplex Ligation-Dependent Probe Amplification (MLPA) results in PDF format. The CRSIP IDs for these raw MLPA result files are listed below:

- 116413
- 120395
- 186714
- 259940
- 301157
- 393936
- 407132
- 493328

Replication of Table 1 in the Rossetti, S. et al publication

Table 1 in the publication listed “Summary of mutations in the CRISP families” in 17 rows. However, we could only replicate the first three rows, because the only subjects(CRISP ID) we had knowledge of are those listed in Tables IA, IB, and IC in the Appendix. As listed below, the numbers are consistent between the publication and the archived data. The remaining rows were not able to be replicated since the subject (CRISP ID) is not directly annotated for each mutation listed.

	Published Results			**Computed from Archive**			Corresponding Table
Mutation Type	PKD1	PKD2	Total	PKD1	PKD2	Total	**Appendix**
Definite (MG = A)	106	21	127	106	21	127	Table IA
Highly likely (MG = B)	34	3	37	34	3	37	Table IB
Likely (MG = C)	13	3	16	13	3	16	Table IC

Replication of Table I-A/B/C in the Appendix of the Rossetti, S. et al publication

Of the 180 CRISP IDs listed, 152 have raw chromatogram files, and 28 have raw chromatogram files with different CRISP IDs from members of the same family. A mapping table for these 28 CRISP IDs from the publication and the CRISP IDs for their family members are provided below.

Published	Archived Chromatograms
CRISP ID	CRISP ID
Table IA - PKD1	
400002	492327
200002	256171
100001	186714
400003	406726
100005	183417
100008	106311
300003	374068
400004	408493
100007	100797
200005	264225
300001	359308
100003	175611
300005	313195
300004	343097
300008	318562
100004	101618
Table IA - PKD2	
200006	243560
140005	182337
200008	290336
300002	327325
300006	313307
Table IB - PKD1	
200001	216086
300007	316110
400001	414128
200007	273225
Table IB - PKD2	
120010	187456
Table IC - PKD1	
200004	206816
200003	215052

Replication of Table II, III, IV in the Appendix of the Rossetti, S. et al publication

These 3 tables are not assessed in this DSIC since the subject (CRISP ID) is not directly annotated for each mutation listed.

Replication of Table V in the Appendix of the Rossetti, S. et al publication

The data archive has the raw chromatogram files for all but the first 3 of the 22 CRISP IDs listed. The chromatogram files of the members from the same family were used

instead for these 3 CRISP IDs. The mapping for the CRISP IDs for family members is provided below.

Published	Archived
CRISP ID	CRISP ID
100002	168556
100006	142175
100009	116413, 120777, 147564, 156643, 187840

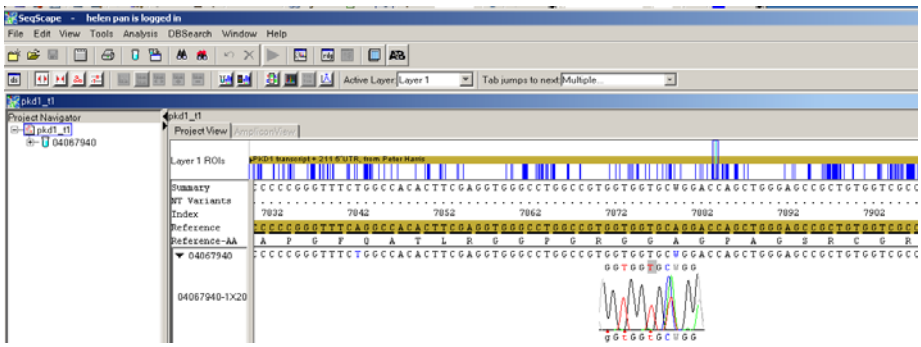
2. Assess if selected reported mutation is present in the raw data

To replicate the variations reported in the report, we used the chromatograms delivered to us, which were assembled using SeqScape Genetic Variation Analysis Software by Applied Biosystems (ABI). CRISP ID was randomly selected from the tables and chromatograms were assembled using "Basecaller-3100POP6SR.bcp" for "Basecaller", and "DP3100POP6{BD-21M13}v1.mob" for "DyeSet/Primer" setting, as suggested from the chromatogram files.

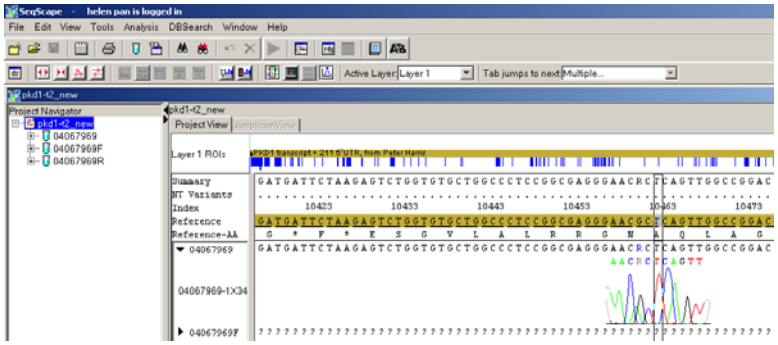
The reference sequences used in the Athena report are not specified in the paper or deposited in NCBI GenBank. These reference sequences were provided to the repository by the Genetics Ancillary study PI. In this DISC report, "PKD1 transcript" and "PKD1 genomic" sequences are used to replicate the SNP basecall for the PKD1 gene, and "PKD2 transcript" sequence is used for the PKD2 gene. The sequence data was numbered from the start of the transcript in the reference sequence, but the analysis report in the publication was numbered from the start of translation. Since the 5'UTR is 210bp, we added 210 to the transcript positions reported in the publication to compare with the reference sequence used here.

Replication of Table I-A/B/C in the Appendix of the Rossetti, S. et al publication

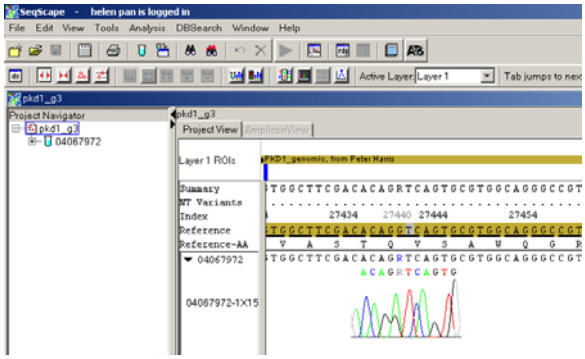
CRISP ID 100505 (Athena ID 4067940), 7666C→T is compared at (7666+210=) 7,876 and the mutation is observed in the archived data.



CRISP ID 157925 (Athena ID 4067969), 10251G→T is compared at (10251+210=) 10,461 and the mutation is observed in the archived data.



CRISP ID 166508 (Athena ID 4067972), IVS15+1G→T (27,440 in genomic sequence) is observed in the archived data.

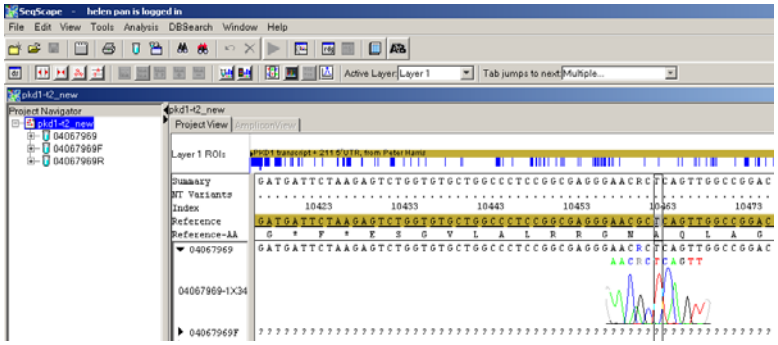


Replication of Table II, III, IV in the Appendix of the Rossetti, S. et al

These three tables are not assessed in this DSIC since the subject (CRISP ID) is not directly annotated for each mutation listed.

Replication of Table V in the Appendix of the Rossetti, S. et al

CRISP ID 157925 (Athena ID 4067969), 10251G→T is compared at (10251+210=) 10,461 and the mutation is observed in the archived data.



III. Summary of the findings for the replication

Assess if raw data for each subject is present

Raw data for all subjects are present, though a few have different data format (PDF instead of chromatogram) and three subjects have data from family members' data instead of their own.

Assess if selected mutation reported is present in the raw data

A few randomly selected mutations were analyzed and the reported mutations could be replicated, which suggests the archived dataset is a true copy of the study dataset.

ATTACHMENT 1

Full Text of Article

Rossetti, S. et al. for the CRISP Study Group. Comprehensive molecular diagnostics in autosomal dominant polycystic kidney disease, **J Am Soc Nephrol.** 2007 Jul; 18(7):2143-60.