Understanding the Exome Kit Used in Genomic Studies is Critical to Minimizing Information Gaps in the Diagnostic Analysis

Amir K. Yaghoobi[1], Sarathbabu Krishnamurthy MS[1], Dr Hila Milo Rasouly PhD[1], Shiraz Bhed[1] , Sandy Yang[1], Dr Ali Gharavi MD[1]
Columbia Division of Nephrology[1]

INTRODUCTION AND OBJECTIVE: One percent of the human genome is comprised of exomes, the protein-coding regions of genes. Because variants within exonic regions are more likely to affect phenotype, exome kits were developed to efficiently sequence these regions. The goal of this exploration was to determine if popular exome kits—specifically the earlier version of Roche Exome Kit (SeqCap_EZ_Exome_v3_capture) used since 2016—have the potential to miss disease-causing genomic regions.

METHODS: Using Python and the pybedtools library, the regions present within a Roche bed file were removed from a Consensus Coding Sequence (CCDS) file. The result was run through ANNOTSV using the GRCh37 genome assembly. All regions corresponding to genes associated with an entry in OMIM (Online Catalog of Human Genes and Genetic Disorders) were selected and the result was then filtered for regions related to a list of genes for Mendelian kidney and genitourinary disorders. Finally, summary information was constructed to highlight key findings.

RESULTS: This analysis found that 275/625 kidney genes contained at least one pathogenic or likely pathogenic variant excluded from the Roche Exome Kit. Additionally, 245 of those genes contained at least one pathogenic variant. The Roche Exome Kit excluded 1,871 pathogenic or likely pathogenic variants from these genes, with 1,331 of the excluded variants being classified as pathogenic. For a more detailed analysis, see Tables 1 and 2. Please note: featured in Table 1 are the five genes with the most pathogenic or likely pathogenic variants missed by the Roche Exome Kit.

Table 1: Top 5 Genes With Regions Missing From the Roche Exome Kit When Sorted by Pathogenic Variant Count

Count	PKD1	KMT2D	MSH6	PALB2	NSD1
Pathogenic	84	62	49	38	34
Likely Pathogenic	19	18	3	0	2
Uncertain Significance	137	158	288	181	60
Likely Benign	42	172	121	67	36
Benign	9	104	3	1	15

Table 2: Summary Statistics

Type	Mean	SD	Min	25%	50%	75%	Max
Pathogenic	3.422	7.532	0	0	1	4	84
Likely Pathogenic	1.388	2.826	0	0	0	2	24
Uncertain Significance	20.82	37.023	0	4	9	21	366
Likely Benign	11.699	19.295	0	2	5	14	172
Benign	1.781	5.834	0	0	1	2	104

CONCLUSIONS: Before performing genomic analyses, researchers should understand the exome kit utilized when sequencing the genomic data they intend to use. Failing to

do so may lead a genomic study to produce incomplete results, as potential regions of interest could be excluded.