# Miles forward, but one step back: The impact of methodological differences in whole exome sequencing centers on the depth of coverage of coding variants with known clinical impact

Murilo G Borges[1,2], Vera N Solferini[3], Benilton S Carvalho[2,4] and Iscia Lopes-Cendes[1,2]
[1]University of Campinas, School of Medical Sciences, Department of Medical Genetics. [2]Brazilian Institute of Neuroscience and Neurotechnology (BRAINN). [3]University of Campinas, Institute of Biology, Department of Genetics, Evolution and Bioagents. [4]University of Campinas, Institute of Mathematics, Statistics and Computing Science, Department of Statistics.

**Introduction:** The coding region of the human genome corresponds to less than 2% of its entirety and it is known as exome. This portion of the human genome concentrates most of the pathologic variations, which are known to cause disease in humans. For a better interpretation of these variants, evidence-based databases, such as ClinVar, compiles data on the presumed relationships between DNA variants and phenotypes. In the present work, we aim to investigate the pattern of base-specific depth in variants present in ClinVar, within the exome definition, in subjects who had the exome captured by different approaches in different sequencing centers by the 1000 Genomes Project.
**Materials and Methods:** We used public data from the 1000 Genomes Project Consortium[1] to investigate the depth of coverage variations in 1112 whole exome sequenced (WES) samples from sequencing phase 3. We extracted 282,453 variants from ClinVar (built 20170801, GRCh37.p13) and performed variant annotation using the Ensembl Variant Effect Predictor (VEP version 84). 4,543 of the total number of variants were exonic and had any impact on transcription as well (121 were classified as high, 2,166 as moderate, 1,641 as low and 615 as modifier). We used "samtools depth" (version 1.0) to estimate the base-by-base depth of the 4,543 considered variants. We conducted all further analyses using the R statistical environment (version 3.3.2). We tested the assumption of no difference among the density of depths for each sequencing center with a pairwise Wilcoxon Test with a subsequent Bonferroni correction. We also applied multidimensional scaling (MDS) to compare the groups, addressing the data high-dimensionality issue and obtained a low-dimensional representation of the data.
**Results:** Depth distribution varies significantly ($p < 0.001$ among each sequencing center), with an average of 82.8±67.6 for BCM, 123.0±85.6 for BGI, 86.6±79.2 for BI and 49.4±33.8 for WUGSC. Multidimensional scaling analysis confirmed that samples depth patterns clusters according to the center they were sequenced in, with 69.0% of the explained variance for the first two principal components. This signals that protocol advancement and intrinsic methodological differences in each one of the sequencing centers directly affect the patterns of coverage in the set of variants analyzed. Through the depth distribution of the 450 variants with higher variance, we could correctly assign 96.9% of the samples to their sequencing centers when considering 5 clusters to the dendrogram branches.
**Discussion:** The originality of the present study lies in the fact that this study compared samples only based on their depth of coverage. The present work shows for the first time that it is possible to distinguish samples based only on their depth patterns, showing that the capture reaction differences in whole exome sequencing directly reflect on final analysis results. Technical integration is challenging while trying to link genetic variations to disease, mainly for federated sequencing initiatives[2,3].
**Conclusion:** We conclude that WES depth in samples from different sequencing centers is liable to technical differences. Our results are not unexpected given that the initial step for a WES experiment is the capture of the target regions to be subsequently enriched and sequenced.
**References:** 1. Consortium, 1000 Genomes Project *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015). 2. Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A. & Tarczy-Hornoch, P. Data integration and genomic medicine. *Journal of Biomedical Informatics* **40,** 5–16 (2007). 3. Ginsburg, G. S. & Willard, H. F. Genomic and personalized medicine: foundations and applications. *Translational Research* **154,** 277–287 (2009).