20th World Congress in Fetal Medicine

Automated bulk-sorting of molecular karyotypes and classification of copy number variants

Gadsbøll K, Rasmussen S, Hui L, Pynaker C, Petersen OB, Vogel I
Center for Fetal Medicine, Pregnancy and Ultrasound, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark

Objective

Manual sorting and classification of copy number variants (CNVs) from large cytogenetic datasets is time-consuming and prone to variations in laboratory practices. The aim of this study was to develop and validate a universal open-source script to bulk-sort karyotype results and to map them to known syndromes for research use.

Methods

From conventional and molecular karyotypes from the Danish Cytogenetic Central Register (n=25,397) a previously published bulk-sorting algorithm in SAS and R was further developed (PMID: 29080253). From this algorithm, karyotypes were categorized as normal or assigned to one of eight abnormal categories. The algorithm was validated using an international sample of 2159 deidentified karyotypes from the Victorian Prenatal Diagnosis Database in Australia. All karyotypes used the ISCN nomenclature. In a newly developed algorithm, further subclassification was applied to CNVs. First, all CNVs were converted into Browser Extensible Data (BED) format, isolating the chromosome number, proximal and distal chromosomal breakpoints. Hg18 karyotypes were converted into hg19. CNVs were then mapped to three reference datasets: 1) DECIPHER (deciphergenomics.org/disorders/syndromes/list); 2) CNV syndromes published in Weise et al. (PMID: 22396478); and 3) Central Denmark Region clinical database of pathogenic (class 5) CNVs.

Results

We developed universal and modifiable scripts in statistical software SAS and R that generated 3 output datasets: 1) Karyotypes categorized as normal or assigned to one of eight abnormal categories (Figure 1); 2) An overview of each CNV, including molecular karyotype, aberration size (Mb), largest overlap with syndromes in the reference datasets, and categorization as pathogenic or not from overlap with reference datasets; 3) For CNVs mapping to a known syndrome, a detailed description of CNV type (del/dup), inheritance, syndrome name, and percentage overlap with the reference dataset was produced. The algorithm can be modified to accommodate additional reference datasets to allow for local customization.

Conclusion

We have developed an open-source algorithm to automate the sorting and description of multiple CNVs, and to generate a detailed summary of their genomic characteristics mapped to known syndromes. This algorithm will increase the efficiency of data analysis and facilitate future research collaborations.