CytoGPS: A Web-Enabled Karyotype Analysis Tool for Cytogeneticists and Biomedical Data Scientists

Zachary B. Abrams, PhD\textsuperscript{1*}; Lin Zhang, MA\textsuperscript{2*}; Ricky Rodriguez\textsuperscript{2}; Lynne Abruzzo, MD, PhD\textsuperscript{3}; Kevin R. Coombes, PhD\textsuperscript{1}; Phillip R. O. Payne, PhD\textsuperscript{2}

\textsuperscript{*}These authors contributed equally to this work

1 Department of Biomedical Informatics, The Ohio State University, Columbus, OH
2 Institute for Informatics, Washington University School of Medicine, St. Louis, MO
3 Dept. of Pathology, Wexner Medical Center, The Ohio State University, Columbus, OH

Background
Karyotyping is the practice of visually examining and recording chromosomal abnormalities and is commonly used to diagnose diseases of genetic origin, including cancers. Karyotypes are written in a language known as the International System for Human Cytogenomic Nomenclature (ISCN). Analyzing karyotypes is currently done in a manual, non-computational manner due to the structure of the ISCN. The ISCN however is computationally intractable and there is no mature software that automatically parses karyotypes, thus precluding the potential of these genomic data from being fully realized. ISCN standards have been in place for almost half a century and, due to its widespread use in the field of medicine (both academia and industry), a tool that can unlock the wealth of genomic data from ISCN karyotypes is badly needed.

Software
We developed a software program, CytoGenetic Pattern Sleuth (CytoGPS), to fill this gap. Built using the tools of ANTLR (Another Tool for Language Recognition)\textsuperscript{1} and Java, CytoGPS first examines the validity of an inputted karyotype based on the latest ISCN manual. Our software is designed to automatically suggest a revised karyotype when an incorrect but correctible karyotype is entered; for example, incorrect use of brackets and/or commas. This effective spell check system is a useful user-friendly feature since human errors and typos are ubiquitous in the process of inputting medical files. This form of data quality assurance can aid in both data entry and in processing large retrospective data sets.

CytoGPS then extracts important and relevant biological information embedded in the karyotypes. Such information includes 1) the number of occurrences of loss, gain, and fusion of each band (or sub-band) on each chromosome, 2) detailed systems for characterizing derivative chromosomes, and 3) the relationship of clones in the karyotypes composed of multiple related clones. The loss gain fusion (LGF) model is a means of representing a karyotype in vector space, enabling downstream computational analysis. Thus, CytoGPS allows cytogeneticists and biomedical data scientists access to new analyses using karyotype data. By converting text-based karyotypes into a computational format, CytoGPS makes it possible to combine them with complementary clinical and genomic data for knowledge discovery purposes.

Deployment
For the ease of use of the CytoGPS tool, we developed a website (http://www.cytogps.org), which is free to use for any nonprofit purpose. We adopted the Ideogram API for data visualization. Users can examine single karyotypes for details or upload a text file for batch analysis. Details of our web portal can be found in our recent publication\textsuperscript{2}. We believe CytoGPS will gain in popularity among cytogeneticists and biomedical data scientists.

References