CytoConverter: a web-based tool to convert karyotypes to genomic coordinates

Janet Wang and Thomas LaFramboise*

Abstract

Background: Cytogenetic nomenclature is used to describe chromosomal aberrations (or lack thereof) in a collection of cells, referred to as the cells’ karyotype. The nomenclature identifies locations on chromosomes using a system of cytogenetic bands, each with a unique name and region on a chromosome. Each band is microscopically visible after staining, and encompasses a large portion of the chromosome. More modern analyses employ genomic coordinates, which precisely specify a chromosomal location according to its distance from the end of the chromosome. Currently, there is no tool to convert cytogenetic nomenclature into genomic coordinates. Since locations of genes and other genomic features are usually specified by genomic coordinates, a conversion tool will facilitate the identification of the features that are harbored in the regions of chromosomal gain and loss that are implied by a karyotype.

Results: Our tool, termed CytoConverter, takes as input either a single karyotype or a file consisting of multiple karyotypes from several individuals. All net chromosomal gains and losses implied by the karyotype are returned in standard genomic coordinates, along with the numbers of cells harboring each aberration if included in the input. CytoConverter also returns graphical output detailing areas of gains and losses of chromosomes and chromosomal segments.

Conclusions: CytoConverter is available as a web-based application at https://jxw773.shinyapps.io/Cytogenetic__software/ and as an R script at https://sourceforge.net/projects/cytoconverter/. Supplemental Material detailing the underlying algorithms is available.

Keywords: Cytogenetics, Chromosomal abnormalities, Karyotypes, Text parsing

Background

Many human diseases, particularly cancer, are caused by or driven by gains and losses of chromosomes or chromosomal segments [2]. In cancer, oncogenes are often found within regions of gain, while tumor suppressor genes are frequently deleted [13]. Chromosomal abnormalities also cause many known syndromes, and may be suspected as causes of undiagnosed diseases [14]. As such, testing for gains and losses is an important component of both research and clinical practice.

Before the advent of higher-resolution techniques, karyotype summarizes the state of the genetic material in a collection of cells. The naming system for describing chromosomal changes – cytogenetic nomenclature – is dictated by the International System for Human Cytogenomic Nomenclature (ISCN) [11]. It describes the total number of chromosomes in the cell, the unmodified sex chromosomes, and the chromosome(s) with abnormalities present. Karyotyping relies on the use of banding techniques, which have allowed researchers to describe the locations of microscopic-level changes such as translocations and band-sized or larger deletions and duplications. The completion of the human genome sequence in the early 2000s, however, enabled description of chromosomal aberrations in terms of genomic coordinates, which can specify locations at nucleotide-level resolution [12]. Technological advances such as genotyping microarrays (reviewed in LaFramboise, 2009 [8]) and, more recently, next-generation sequencing (NGS) have

* Correspondence: Thomas.LaFramboise@case.edu
Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH 44106, USA

© The Author(s). 2019 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
spurred the community to more commonly frame dele-
tions and gains in terms of genomic coordinates. How-
ever, karyotyping is still used today [9]. The use of
cytogenetics remains popular for analyzing blood samples
[5], for example. Additionally, there is copious archival
data (e.g. the Mitelman Database of Chromosome Aberra-
tions and Gene Fusions in Cancer) recorded in cyto-
genetic nomenclature. Large numbers of these karyotyped
samples have not been subjected to microarray or NGS-
based analyses, and in many cases the DNA is no longer
available. In order to allow researchers to more easily
identify the genomic entities such as genes and regulatory
loci that are harbored in the gained and lost regions of
karyotyped samples, we have developed a converter that
parses the karyotypes and returns the corresponding gains
and losses in terms of genomic coordinates.

A handful of tools are currently available to computa-
tionally analyze cytogenetic nomenclature. CyDAS [7] can
characterize the loss and gain of chromosome material and
and can create karyograms from cytogenetic nomenclature.
However, the program does not appear to have updated
since 2004, and gives information using band locations, not
genomic coordinates. The National Center for Biotechnol-
gy Information (NCBI) provides a tool to convert
cytogenetic bands into genomic coordinates. However, this
tool does not characterize losses and gains, and only con-
verts individual bands and not karyotypes (https://www.
cbi.nlm.nih.gov/genome/tools/cyto_convert/).

Implementation
CytoConverter is written in R. It accepts single karyo-
type strings and/or text, tab-delimited tables as input.
An input table must have two columns, one indicating
sample names and the other containing the karyotypes.
The rules of cytogenetic nomenclature should be
followed according to the ISCN 2016. CytoConverter
outputs gains and losses, labelled by sample name, in
hg18, hg19, or GRCh38 coordinates. If cell count infor-
mation is provided in the input karyotype, CytoConver-
ter also reports the numbers of cells harboring each
aberration, and the total number of cells in the sample.
More specifically, CytoConverter divides a karyotype into
individual clones, where applicable, reporting results ac-
cording to each clone’s karyotype. It will output a table of
losses and gains, with each row consisting of the sample
name (adding “_n” to the sample name for the n
th clone in
the karyotype, if multiple clones are present), chromosome
number, beginning base position, ending base position,
whether the region was lost or gained, and number of cells in the clone out of total number of cells in the sample if cell number information is provided in the input karyotype. CytoConverter also provides graphical output displaying a heatmap in which rows represent distinct samples or clones and columns represent chromosomal locations. Gains are displayed in red, hemizygous losses in blue, and homozygous losses in yellow. An overview of the algorithm used to parse cytogenetic nomenclature is provided in Fig. 1.

CytoConverter uses the coordinates of cytobands as specified by the cytoband.txt file (resolution 850 bands, the maximum available) from the UCSC Genome Browser R/Builds folder. A web interface was created using Rshiny (Chang et al., 2018 [4]), an R package used to build interactive web applications that run R in the background.

Results
Here we provide a set of examples demonstrating CytoConverter’s capabilities.

Single karyotype example
Figure 2 (top panel) shows an example of the interface and output of CytoConverter, applied to the karyotype of AML-193, a cell line derived from a female acute myeloid leukemia (AML) patient. The karyotype for AML-193 is given as 49 < 2n>,XX,+ 3,+ 6,+ 8,+ 13,i (17q) by the Leibniz Institute DSMZ- German Collection of Microorganisms and Cell Cultures (https://www.dsmz.de/catalogues/details/culture/ACC-549.html). As part of the Cancer Cell Line Encyclopedia [1], AML-193 was subjected to microarray-based copy number analysis, the results of which are shown in Fig. 2 (bottom panel). As shown, the copy number changes inferred by CytoConverter from the karyotype match well with the copy number lesions revealed from arrays, except for the more focal changes that are below the limited resolution available on the karyotype level. In this example, the number of cells present are not indicated in the karyotype, but CytoConverter is set up to accommodate complex karyotypes wherein different cells harbor different chromosomal aberrations (see Example involving multiple clones subsection below).

Multiple karyotype example
We acquired karyotypes for 943 pediatric patients from the TARGET (“Therapeutically Applicable Research To Generate Effective Treatments”; [10]) project. We uploaded a text, tab-delimited file (Additional file 2: Table
S1) containing all 943 karyotypes to CytoConverter to automatically extract the implied gains and losses. CytoConverter’s tabular output is given in Additional file 3: Table S2. Recurrent duplication of chromosome 8 is apparent in CytoConverter’s graphical representation (Fig. 3). In general, it is expected the graphical output will be useful for visual detection of frequent loss or gain of specific chromosomal loci in associated samples.

Example involving multiple clones
Standard cytogenetic nomenclature can accommodate samples with multiple distinct clones, each having a different karyotype. The nomenclature can also be used to indicate the number of visible cells corresponding to each clone karyotype. For example, 47,X,+X [30]/46,XX, + 7,+ 9[50] indicates a sample with two clones, one with 30 visible cells harboring an extra copy of the X chromosome, and the other with 50 visible cells harboring extra copies of chromosomes 7 and 9. As noted above, CytoConverter handles multiple clones by appending “_n” to the sample name for its n-th clone’s karyotype. Table 1 gives example input for three patient samples, the last of which has the two-clone karyotype 47,X,+X [30]/46,XX, + 7,+ 9[50]. CytoConverter’s tabular and graphical output are shown in Table 2 and Fig. 4, respectively.

Table 1 Example input to CytoConverter
| Sample Name and Karyotype | Karyotype |
|--------------------------|-----------|
| ABC                      | 46,XY,der(1;19)(q10;p10) |
| DEF                      | 47,XX,+der(10)(10;21)(p13;q21) |
| P_10                     | 47,X[30]/48,XX,+ 7,+ 9 [50] |

Table 2 CytoConverter’s tabular output from input given in Table 1
| Sample ID | Chr | Start | End    | Type   | Cells Present |
|-----------|-----|-------|--------|--------|---------------|
| ABC_1     | chr1 | 0     | 125,000,000 | Loss   | unknown       |
| ABC_1     | chr19| 26,500,000 | 59,128,983 | Gain   | unknown       |
| DEF_1     | chr10| 12,200,000 | 135,534,747 | Gain   | unknown       |
| DEF_1     | chr21| 16,400,000 | 48,129,895  | Gain   | unknown       |
| P_10_1    | chrX | 0     | 155,270,560 | Gain   | 30 of 80      |
| P_10_2    | chr7 | 0     | 159,138,663 | Gain   | 50 of 80      |
| P_10_2    | chr9 | 0     | 141,213,431 | Gain   | 50 of 80      |

Fig. 3 Graphical output by CytoConverter derived from karyotypes of 943 patients from the TARGET consortium [10]
Additional cytogenetic lesions and examples
There is a substantial diversity of cytogenetic terms that CytoConverter is able to handle beyond those shown above. We detail CytoConverter’s approach to parsing each of these in the Additional file 1, where we also provide results from its parsing of several more example karyotypes.

Conclusions
In summary, we have developed a user-friendly web-based tool that allows users to input any number of human karyotypes, and obtain the genomic coordinates of all gains and losses implied by each of the karyotypes. We anticipate that this tool will be of considerable value to the community for analyzing archival patient samples, as well as samples for which higher-resolution copy number data is unavailable. It should be noted that CytoConverter only reports net gain and loss of chromosomal material relative to the normal, diploid 2n. In keeping with the standards for array-based copy number inference, we do not report, for example, balanced translocations or inversions. This is an area for future development.

Availability and requirements
Project name: CytoConverter.
Project home page: https://jxw773.shinyapps.io/Cytogenetic__software/

Operating system(s): Platform independent.
Programming language: R.
Other requirements: R 3.5.3 or higher.
License: GNU v.3.0.
Any restrictions to use by non-academics: None.

Additional files

Additional file 1: Additional material provided describing uses and caveats of CytoConverter including example input. (DOCX 136 kb)
Additional file 2: Sample table containing karyotypes for 943 pediatric patients from the TARGET ("Therapeutically Applicable Research To Generate Effective Treatments"; [10]) project. (TXT 54 kb)
Additional file 3: Table output generated by CytoConverter from sample input from Additional file 2. (TXT 56 kb)

Abbreviations
AML: Acute myeloid leukemia; ISCN: International System for Human Cytogenomic Nomenclature; NCBI: National Center for Biotechnology Information; NGS: Next-generation sequencing

Acknowledgements
This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University.

Authors’ contributions
TL conceptualized the work and helped draft the manuscript. JW wrote the software and helped draft the manuscript. Both authors have read and approved the final manuscript.
Funding
Thank you to the Case Western Reserve Support of Undergraduate Research and Creative Endeavors (SOURCE) for funding part of the project. SOURCE did not have any influence on the direction of the research.

Availability of data and materials
Some data analyzed was derived from cBioPortal from the CCLE cell line data. https://www.cbioportal.org/patient?studyId=cellline_ccle_broad&caseId=AML-193. Karyotype was derived from https://www.dsmz.de/catalogues/details/culture/ACC-549.html

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 15 April 2019 Accepted: 29 August 2019
Published online: 11 September 2019

References
1. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, ... Garraway LA, et al. The cancer cell line encyclopedia enables predictive modeling of anticaner drug sensitivity. Nature. 2012;483(7391):603-607. doi:https://doi.org/10.1038/nature11003.
2. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463(7283):899-905. https://doi.org/10.1038/nature08832.
3. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy B, ... Schultz N, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discovery. 2012;2(5):401-404. doi:https://doi.org/10.1158/2159-8290.CD-12-0095.
4. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. Shiny: web application framework for R. R package version 1.1.0; 2018.
5. Ferguson-Smith M. History and evolution of cytogenetics. Mol Cytogenet. 2015;8:169. https://doi.org/10.1186/s13039-015-0125-8.
6. Gao J, Aksoy B, Dogrusoz U, Dresdner G, Gross B, Sumer SO, ... Schultz N, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6(269), pl1. doi: https://doi.org/10.1126/scisignal.2004088.
7. Hiller B, Bradikje J, Balz H, Rieder H. CyDAS: a cytogenetic data analysis system. Bioinformatics. 2005;21(7):182-3.
8. LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res. 2009;37(13):4181-93. doi:https://doi.org/10.1093/nar/gkp552.
9. Li S, Garrett-BaleSinme FE, Chung SS, Sanders MA, Hirck T, Rapaport F, et al. Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. Nat Med. 2016;22(7):792-9. https://doi.org/10.1038/nmm4125.
10. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukemias and solid tumours. Nature. 2018;555(7696):371-6. https://doi.org/10.1038/nature25795.
11. McGowan-Jordan J, and Schmidt M. ISCN 2016 An International System for Human Cytogenomic Nomenclature. Reprint of: Cytogenetic and Genome Research 2016;148:1, Karger, S.
12. Naidoo N, Pawitan Y, Soong R, Cooper DN, Ku C. Human genetics and genomics a decade after the release of the draft sequence of the human genome. Human Genomics. 2011;5(6):577-622. https://doi.org/10.1186/1479-7364-5-6-577.
13. Smith JC, Sheltzer JM. Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. eLife. 2018;7:e39217. https://doi.org/10.7554/eLife.39217.
14. Theisen A, Shaffer LG. Disorders caused by chromosome abnormalities. Appl Clin Genet. 2010;3:159-74. https://doi.org/10.2147/ACG.S8884.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.