InferCNV, a python web app for copy number inference from discrete gene-level amplification signals noted in clinical tumor profiling reports [version 1; peer review: awaiting peer review]

Paraic A. Kenny

1 Kabara Cancer Research Institute, Gundersen Medical Foundation, La Crosse, WI, 54601, USA
2 Department of Medicine, University of Wisconsin-Madison, Madison, WI, 53705, USA

**Abstract**
As somatic next-generation sequencing gene panel analysis in advanced cancer patients is becoming more routine, oncologists are frequently presented with reports containing lists of genes with increased copy number. Distinguishing which of these amplified genes, if any, might be driving tumor growth and might thus be worth considering targeting can be challenging. One particular issue is the frequent absence of genomic contextual information in clinical reports, making it very challenging to determine which reported genes might be co-amplified and how large any such amplicons might be. We describe a straightforward Python web app, InferCNV, into which healthcare professionals may enter lists of amplified genes from clinical reports. The tool reports (1) the likely size of amplified genomic regions, (2) which reported genes are co-amplified and (3) which other cancer-relevant genes that were not evaluated in the assay may also be co-amplified in the specimen. The tool is accessible for web queries at http://infercnv.org.

**Keywords**
cancer, genetic testing, copy number variation, gene amplification, oncology, targeted therapy

This article is included in the Python collection.
Corresponding author: Paraic A. Kenny (pakenny@gundersenhealth.org)

Author roles: Kenny PA: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project was supported by the Gundersen Medical Foundation. P.K. holds the Dr. Jon & Betty Kabara Endowed Chair in Precision Oncology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Kenny PA. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Kenny PA. InferCNV, a python web app for copy number inference from discrete gene-level amplification signals noted in clinical tumor profiling reports [version 1; peer review: awaiting peer review] F1000Research 2019, 8:807 https://doi.org/10.12688/f1000research.19541.1

First published: 06 Jun 2019, 8:807 https://doi.org/10.12688/f1000research.19541.1
Introduction
Focal somatic gene copy number changes are a widespread event in tumor evolution. Although these regions of amplification may be large, encompassing many hundreds of genes, typically only one or a small number of genes within the amplified regions are involved in driving tumor growth. Identification of the key driver genes within recurrent amplicons has led to the approval of some therapies that have changed clinical practice (e.g. anti-ERBB2 agents\(^3\)); however, targeting other amplified genes such as FGFR family members\(^4\), \(EGFR\)\(^5\) or \(KIT\)\(^6\) has frequently proved disappointing. Nevertheless, even some of the more negative trials include occasional strong responses, indicating that sub-populations of patients with amplification of these oncogenes may experience clinical benefit if they can be identified.

With the goal of individualizing treatment for cancer patients, next-generation sequencing from tumor specimens is becoming widely adopted\(^7\). In addition to somatic point mutations, several of these assays report copy number changes in assayed genes. Reports for physicians typically present a list of amplified genes without providing a genomic context, leaving physicians and molecular tumor boards to hypothesize which of the listed genes might be driver genes suitable for therapeutic targeting. Given the poor response rates that have often been observed in clinical studies with amplified genes (compared to targeting genes activated by point mutation or fusion), physicians are often appropriately cautious about deciding whether a reported amplified gene may be actionable. Thus, many patients are spared receiving ineffective therapies, but a subgroup of patients who may experience clinical benefit do not get that opportunity.

Here we provide an easy-to-use web tool for analyzing clinical genomics reports of amplified genes. It determines (1) the likely size of amplified genomic regions, (2) which reported genes are co-amplified and (3) which other cancer-relevant genes that were not evaluated in the assay may also be co-amplified in the specimen.

The primary goals are to allow healthcare professionals to determine whether the amplification region surrounding a particular oncogene is relatively small and lacking in other likely candidate cancer drivers (which may indicate increased likelihood that the analyzed gene is a driver) from larger amplicons with additional candidate driver genes (which would suggest a reduced probability that the reported gene is a driver). The approach was developed to analyze the widely used Foundation One test (329 genes) provided by Foundation Medicine but, by simply editing the target gene list, it can be generalized to tests from other vendors which report copy number variation throughout the genome.

Methods
Implementation
InferCNV is written in Python 2.7 with Flask and implemented as a web service running on the Google App Engine (http://infercnv.org). Additional supplied requirements are (1) the coordinates of genes in the human genome, ‘coordinates.txt’ (hg38, UCSC genome browser), (2) The gene list from the assay of interest, ‘foundationone.txt’ and (3) a file listing genes recurrently altered in cancer from COSMIC\(^8\) (retrieved 5/3/2018), ‘cosmic.txt’.

An html page with a single query window allows the user to enter a comma-delimited list of genes reported as being amplified. The entry is passed to the script and parsed into individual gene names. An error check is performed to confirm that all entered gene names correspond to gene names in the genome used. The entered genes are considered to be amplified, while the other genes in the assay are considered to be not amplified.

A simplified schema of how the algorithm works is presented in Figure 1, which depicts a chromosomal region containing 30 genes. Seven cancer-relevant genes are present, five of which are evaluated by the genomic assay (Figure 1A). In this test example, three genes were reported to be amplified (Figure 1B). Running the algorithm identifies these amplified genes (8, 13, 20) as well as the nearby assayed genes that are not reported amplified (4, 27). The algorithm considers all genes located between genes reported as amplified to also be amplified (Figure 1C, red shaded region). Because not every gene is assayed, precisely delineating the boundaries of an amplicon is not possible. To address this, the algorithm determines the nearest non-amplified gene at each end of the amplicon and infers that the genes located up to, but not including that gene may be possibly amplified (Figure 1D). The script then returns an html report page listing the entered genes, the amplicons into which they fall (in many cases, several discrete genes will be consolidated into a single amplicon), and also the other cancer-relevant genes within these regions that may be co-amplified with the reported genes. All genes reported include hyperlinks to that gene’s page on COSMIC.

Operation
InferCNV is accessed via a web browser and has been tested on commonly used browsers such as Chrome, Firefox, Internet Explorer and Microsoft Edge.

Use cases
Three use cases taken from genomic reports of patients at our clinic are presented:

Use case 1
A case of esophageal adenocarcinoma with eight reported amplified genes (Table 1), which were resolved by InferCNV to four amplicons. The co-amplification of \(FGF3\), \(FGF4\) and \(FGF10\) with \(CCND1\) (which is likely the driver gene in this amplicon\(^9\)) might indicate that consideration of FGFR inhibitors may not be helpful if these FGF genes are simply co-amplified passenger genes.

Use case 2
A case of soft tissue sarcoma with five reported amplified genes (Table 2) which were resolved into three amplicons. In the absence of genomic context information, both \(PDGFR\) and
**Figure 1. Schematic representation of amplicon boundary inference approach.** (A) Schematic diagram of a model genomic region with 30 numbered genes, which include a total of 7 cancer-relevant genes. (B) Input scenario for algorithm: a clinical genomics report noting amplification of three genes in this region. (C) Copy number inference for genes in regions bounded by reported amplified genes. (D) Copy number inference for genes surrounding regions bounded by reported amplified genes.

**Use case 3**

The third example is a breast cancer case from our clinic with three reported amplified genes (Table 3). The report highlights one region on chromosome 5, and two regions on chromosome 7. The latter predicted amplicons share a nearby boundary at 7q22.3 suggesting the possibility that there is a regional amplification on 7q encompassing both sets of genes. In this case, MET was judged to be a possible driver amplicon, and the patient had a very strong response to a MET inhibitor\(^9\).

**Discussion**

We have described a straightforward tool to provide additional genomic context to aid interpretation of amplifications in somatic cancer sequencing reports. Use of this tool may aid decision-making by healthcare professionals about therapeutic options.
Table 1. Use case – Esophageal adenocarcinoma with reported amplification of CCND1, MAP2K1, RICTOR, FGF10, FGF19, FGF3, FGF4 and MCL1.

<table>
<thead>
<tr>
<th>Gene reported amplified (chromosomal location)</th>
<th>Number of potentially co-amplified genes (chromosomal region)</th>
<th>Genes annotated as recurrently altered in cancer by COSMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCL1 (Chr1q21.2)</td>
<td>8 (Chr1p12–1q23.1)</td>
<td>BCL9, PDE4DIP, ARNT, MLLT11, TPM3, MUC1, LMNA, PRCC</td>
</tr>
<tr>
<td>RICTOR (Chr5p13.1), FGF10 (Chr5p12)</td>
<td>75 (Chr5p13.2–5q11.2)</td>
<td>LIFR, IL6ST</td>
</tr>
<tr>
<td>CCND1 (Chr11q13.3), FGF19 (Chr11q13.3), FGF4 (Chr11q13.3), FGF3 (Chr11q13.3)</td>
<td>214 (Chr11q13.1–11q13.5)</td>
<td>CCND1, NUMA1</td>
</tr>
<tr>
<td>MAP2K1 (Chr15q22.31)</td>
<td>224 (Chr15q15.1–15q22.31)</td>
<td>B2M, USP8, MYO5A, C15orf65, TCF12, MAP2K1</td>
</tr>
</tbody>
</table>

Table 2. Use case – Soft tissue sarcoma with reported amplification of KIT, PDGFRA, MDM2, RICTOR and FGF10.

<table>
<thead>
<tr>
<th>Gene reported amplified (chromosomal location)</th>
<th>Number of potentially co-amplified genes (chromosomal region)</th>
<th>Genes annotated as recurrently altered in cancer by COSMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDGFRA (Chr4q12), KIT (Chr4q12)</td>
<td>97 (Chr4p15.31–4q12)</td>
<td>SLC34A2, RHOH, PHOX2B, FIP1L1, CHIC2, PDGFRA, KIT</td>
</tr>
<tr>
<td>RICTOR (Chr5p13.1), FGF10 (Chr5p12)</td>
<td>75 (Chr5p13.2–5q11.2)</td>
<td>LIFR, IL6ST</td>
</tr>
<tr>
<td>MDM2 (Chr12q15)</td>
<td>48 (Chr12q14.1–12q15)</td>
<td>LRIG3, WIF1, HMGa2, MDM2</td>
</tr>
</tbody>
</table>

Table 3. Use Case – Triple Negative breast cancer with reported amplification of RICTOR, CDK6 and MET.

<table>
<thead>
<tr>
<th>Gene reported amplified (chromosomal location)</th>
<th>Number of potentially co-amplified genes (chromosomal region)</th>
<th>Genes annotated as recurrently altered in cancer by COSMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RICTOR (Chr5p13.1)</td>
<td>44 (Chr 5p13.2–5p12)</td>
<td>LIFR</td>
</tr>
<tr>
<td>CDK6 (Chr7q21.2)</td>
<td>190 (Chr 7q21.12–7q22.3)</td>
<td>AKAP9, CDK6, TRRAP, CUX1</td>
</tr>
<tr>
<td>MET (Chr7q31.2)</td>
<td>91 (Chr 7q22.3–7q32.1)</td>
<td>MET, POT1, SND1</td>
</tr>
</tbody>
</table>

The method relies on the accuracy with which test vendors report gene amplification calls. In testing, we identified a small number of cases in which two amplicons were inferred in very close proximity (e.g. Use case 3), which raises the possibility that the assayed gene between the two regions is erroneously not called as amplified. In cases with two or more closely co-located amplicons, users should consider that there is a strong possibility of a regional amplification encompassing both predicted amplicons. Future assays with larger number of genes or more sensitive amplification calling algorithms will likely permit more accurate refining of the boundaries of individual amplicons.

Because the coverage across the genome is somewhat sparse, refining the amplicon boundaries is more challenging than with a more high-density approach like SNP arrays. The primary purpose is to list genes that are potentially co-amplified with a gene identified by a test vendor as possibly actionable in order to allow healthcare professionals to gain further insight into the likelihood that the listed gene is truly the driver gene in that amplicon. Accordingly, we do not distinguish in the report between genes that are likely co-amplified (red genes, Figure 1) from the boundary region genes which are possibly co-amplified (green genes, Figure 1). In any case in which a healthcare professional might consider targeting a non-assayed gene predicted by this algorithm to be amplified (e.g. LIFR in Use case 2 and Use case 3), further clinical testing to directly confirm gene amplification would be warranted.

**Data availability**

All data underlying the results are available as part of the article and no additional source data are required.

**Software availability**

Source code available from: https://github.com/paraickenny/inferCNV.


License: MIT License.

References


Grant information
This project was supported by the Gundersen Medical Foundation. P.K. holds the Dr. Jon & Betty Kabara Endowed Chair in Precision Oncology.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Soon after publishing this article, we realized that the "inferCNV" name was already in prior use for an unrelated bioinformatics tool. To avoid confusion, the tool described in this article has been renamed "InferAMP" and can be accessed at http://inferamp.org. The name of the tool will be updated in the next version of the article.

**Competing Interests:** No competing interests were disclosed.