SOFTWARE TOOL ARTICLE

DangerTrack: A scoring system to detect difficult-to-assess regions [version 1; referees: 2 approved, 1 approved with reservations]

Igor Dolgalev [1*], Fritz Sedlazeck[2*], Ben Busby[3]

[1]New York University School of Medicine, New York, NY, 10016, USA
[2]Department of Computer Science, Johns Hopkins University, Baltimore, MD, 21202, USA
[3]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA

[*] Equal contributors

v1 **First published:** 07 Apr 2017, **6**:443 (doi: 10.12688/f1000research.11254.1)
Latest published: 07 Apr 2017, **6**:443 (doi: 10.12688/f1000research.11254.1)

Abstract

Over recent years, multiple groups have shown that a large number of structural variants, repeats, or problems with the underlying genome assembly have dramatic effects on the mapping, calling, and overall reliability of single nucleotide polymorphism calls. This project endeavored to develop an easy-to-use track for looking at structural variant and repeat regions. This track, DangerTrack, can be displayed alongside the existing Genome Reference Consortium assembly tracks to warn clinicians and biologists when variants of interest may be incorrectly called, of dubious quality, or on an insertion or copy number expansion. While mapping and variant calling can be automated, it is our opinion that when these regions are of interest to a particular clinical or research group, they warrant a careful examination, potentially involving localized reassembly. DangerTrack is available at https://github.com/DCGenomics/DangerTrack.

This article is included in the Hackathons channel.

Open Peer Review

Referee Status: ? ✓ ✓

	Invited Referees 1	2	3
version 1 published 07 Apr 2017	? report	✓ report	✓ report

1 **Melissa A. Gymrek**, University of California, San Diego USA, University of California USA, **Nima Mousavi**, University of California USA

2 **Justin M. Zook**, National Institute of Standards and Technology (NIST) USA

3 **Andrew Carroll**, DNAnexus USA

Discuss this article

Comments (0)

Corresponding author: Ben Busby (ben.busby@gmail.com)

How to cite this article: Dolgalev I, Sedlazeck F and Busby B. **DangerTrack: A scoring system to detect difficult-to-assess regions [version 1; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2017, **6**:443 (doi: 10.12688/f1000research.11254.1)

Grant information: F.S. was supported through a National Science Foundation award (DBI-1350041) and National Institutes of Health award (R01-HG006677). B.B. was supported by the Intramural Research Program of the National Library of Medicine.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 07 Apr 2017, **6**:443 (doi: 10.12688/f1000research.11254.1)

Introduction

The advent of next generation sequencing has enabled the comparison of cells, organisms, and even populations at the genomic level. Whole genome sequencing experiments are run worldwide on a daily basis with various aims, from exploring novel genomes to diagnosing complex variations in high-ploidy cancer samples. A common step in all of these studies is the mapping of the sequence to a reference genome or assembly to identify variations (whole genome sequencing) or expression (RNA sequencing) of the sample.

Multiple studies so far have suffered from mapping artifacts typically occurring in highly variable regions, including single nucleotide polyporphisms (SNPs) and structural variants (SVs), which may be repetitive regions or regions that are not correctly represented by the reference genome (Degner *et al.*, 2009). Multiple methods have been suggested to overcome this bias, including constructing a personalized reference genome (Satya *et al.*, 2012), sequencing the parental genomes (Graze *et al.*, 2012), building graph genomes over all known variants (Dilthey *et al.*, 2015), or carefully reconciling particular subregions. The latter includes discarding reads using a mapping quality filter, realigning reads locally, or computing a localized *de novo* assembly using the Genome Analysis Toolkit to improve the quality of SNP calls. However, all these methods often depend on the sample quality (e.g. coverage, error rate), may result in additional expenses, and are often optimized only for human genome data.

Here, we present DangerTrack, the first approach to automatically classify difficult-to-assess regions by combining annotated features, such as mappability and SV calls. DangerTrack can be applied to any genome and organism of interest. It runs within minutes and provides a Browser Extensible Data (BED) file with a score for every 5 kb region. The height of the score indicates the trustworthiness of the region in terms of SNP calling, and thus how difficult an accurate mapping can be. DangerTrack represents a flexible and easy to use method to detect hard-to-analyze regions with a pure mapping approach. We compared the results of DangerTrack to the blacklisted regions of ENCODE (https://personal.broadinstitute.org/anshul/projects/encode/rawdata/blacklists/hg19-blacklist-README.pdf), as well as to the list of problematic regions from the National Center for Biotechnology Information (NCBI).

Methods

Incorporation of structural variation data sets

We downloaded the SVs dataset from the 1000 Genomes Project (Sudmant *et al.*, 2015) (1KG) from dbVar (estd219; https://www.ncbi.nlm.nih.gov/dbvar/studies/estd219/), as well as a 16-candidate SV callset from the Genome in a Bottle (GIAB; Zook *et al.*, 2016) (Ashkenazi son dataset available at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/latest; NIST Reference Material 8391: HG002 and NA24385). These Variant Call Format (VCF) datasets were converted into BED files using SURVIVOR (Jeffares *et al.*, 2017) (available from https://github.com/fritzsedlazeck/SURVIVOR). Each SV was represented by two entries in the BED file, listing the breakpoints of each reported SV.

Next, we binned the breakpoints in 5 kb windows and counted the number of SVs in these windows. The number of SV breakpoints per window was normalized by the 99% quantile number of breakpoints within a window across the whole genome. Thus, the higher the ratio, the more SV breakpoints are in a given window, and therefore the less trustworthy the reference seems to be.

Incorporation of mappability tracks

We downloaded the 50 bp and 100 bp mappability tracks from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/).

These mappability tracks contain a measurement for each base of the reference genome. These tracks were generated using different window sizes, with high signals indicating areas where the sequence is unique. The GEM (GEnome Multitool) mapper (http://big.crg.cat/services/gem_aligner) was used to generate CRG k-mer alignability. The method is equivalent to mapping sliding windows of k-mers back to the genome. For each window, a mappability score is computed as 1 divided by the number of matches found in the genome. Thus, a score of 1 indicates one match in the genome, 0.5 indicates two matches in the genome, and so on.

Next, we computed the score for uniqueness of regions. This was done by subtracting the average mappability value from 1. Thus, a value of 0 represents a unique region. Similarly to the SVs computation method, we summarized the average uniqueness score per 5 kb window, obtained by simple average across the window for both 50 bp and 100 bp tracks.

DangerTrack score calculation

We computed the DangerTrack score by combining all four features with a uniform weighting schema. Note that our score operates between 0 and 1, where 0 means a unique, easy-to-assess region, and 1 means a region that is repetitive and enriched for structural variations.

The resulting genome-wide DangerTrack score in BED and bedGraph formats are available at: https://github.com/DCGenomics/DangerTrack. The repository also contains the bash and R scripts for downloading, cleaning, and summarizing the data, so the score can be computed independently or for different window sizes. The code can be adapted for use with other genomes assuming comparable mappability and structural variation data sets are available.

Comparison to blacklist regions from NCBI and ENCODE

We downloaded the Blacklisted Regions that are defined as problematic by ENCODE Data Analysis Center (https://www.encodeproject.org/annotations/ENCSR636HFF/), as well as the list from the Genome Reference Consortium (ftp.ncbi.nlm.nih.gov/pub/grc/human/GRC/Issue_Mapping/) of regions that either underwent manual curation from GRch37 to GRch38 or are listed

as problematic for future versions of the human genome. For the comparison, we binned the list of regions similar to our approach in 5 kb regions. Next, we compared the values between our track and the generated ENCODE and GRC tracks.

Results

Data exploration

To assess the ability of DangerTrack to highlight suspicious regions, we computed the DangerTrack score over the human reference genome (hg19) using data from the 1000 Genomes Project and GIAB, as well as mappability tracks from UCSC. We downloaded 72,432 SVs from the 1000 Genomes Project data and 135,653 SVs from the GIAB database, for a total of 363,234 breakpoints. Figure 1 shows the histogram of the number of SV breakpoints within each 5 kb bin. As expected, the SV events predicted by the 1000 Genomes Project (labeled 1KG in Figure 1) and GIAB data highlight regions in the genome with high structural variability. However, very few regions exist that incorporate more than 20 events within 5 kb. Interestingly, these two tracks are not very similar, with a correlation of only 0.06 over a subsample of 10,000 bins.

For the mappability data, we naturally expect a high correlation, since the regions that are not unique within a 100 bp region will also not be unique given a 50 bp sequence. The correlation over the 10,000 subsampled 5 kb regions is therefore high (0.95). We chose these two mappability tracks as they reassemble the often-used read length and also take into account local alignment-based clipping of reads.

Evaluation of DangerTrack

Next, we compared the DangerTrack score to manually curated regions from ENCODE and NCBI. These regions represent areas along the genome that are either discarded due to their problematic mapping from previous experiences during the ENCODE project, updated in GRCh38, or still under manual curation for future genome releases. Figure 2 and Figure 3 represent the comparison between the DangerTrack score and the listed regions for ENCODE and NCBI, respectively. We observe a very high correlation for both tracks, highlighting that the DangerTrack score captures these regions. Figure 4 represents the overlap of the DangerTrack score and the annotated regions from Encode and NCBI for chromosomes 1–8.

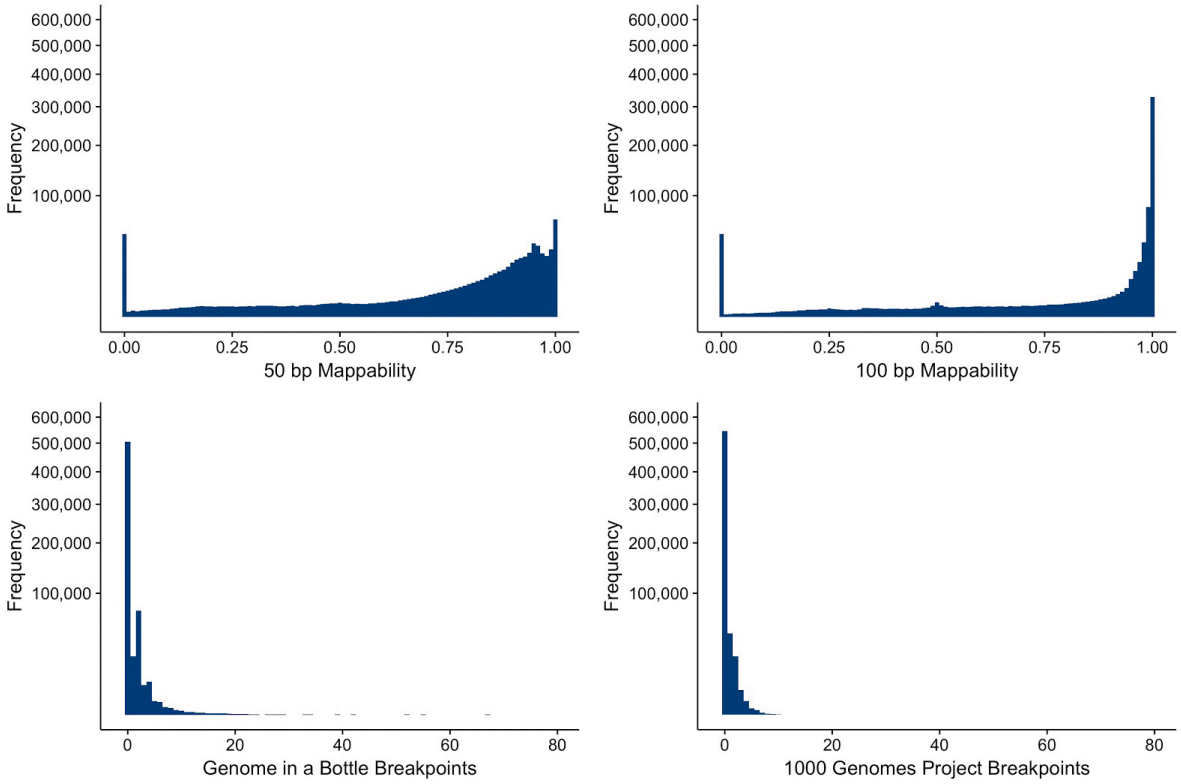

Figure 1. Distributions of the observed frequencies over the binned datasets. (A) Histogram over the hg19 genome of mappability with respect to 50 bp. (B) Histogram over the hg19 genome of mappability with respect to 100 bp. (A and B) are obviously closely related, with the exception that (A) (50 bp regions) includes more regions that have on average a lower score. (C) Distribution of SVs across the hg19 genome based on 16 SV data sets. (D) Distribution of SVs across the hg19 genome based on the 1000 Genomes Project call set.

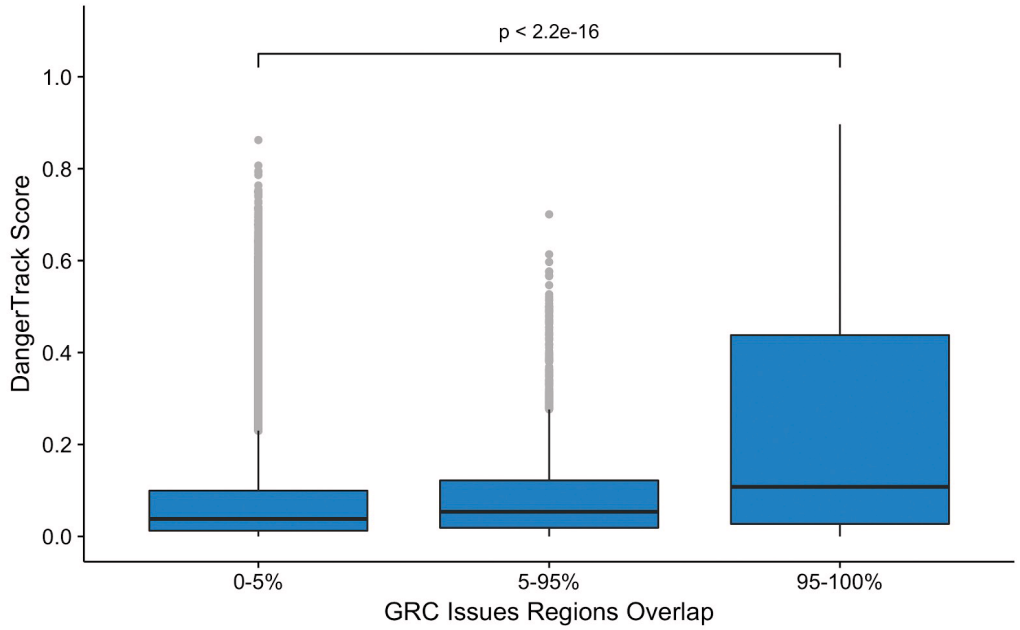

Figure 2. Comparison between DangerTrack score and GRC updated regions or regions that are still manually vetted.

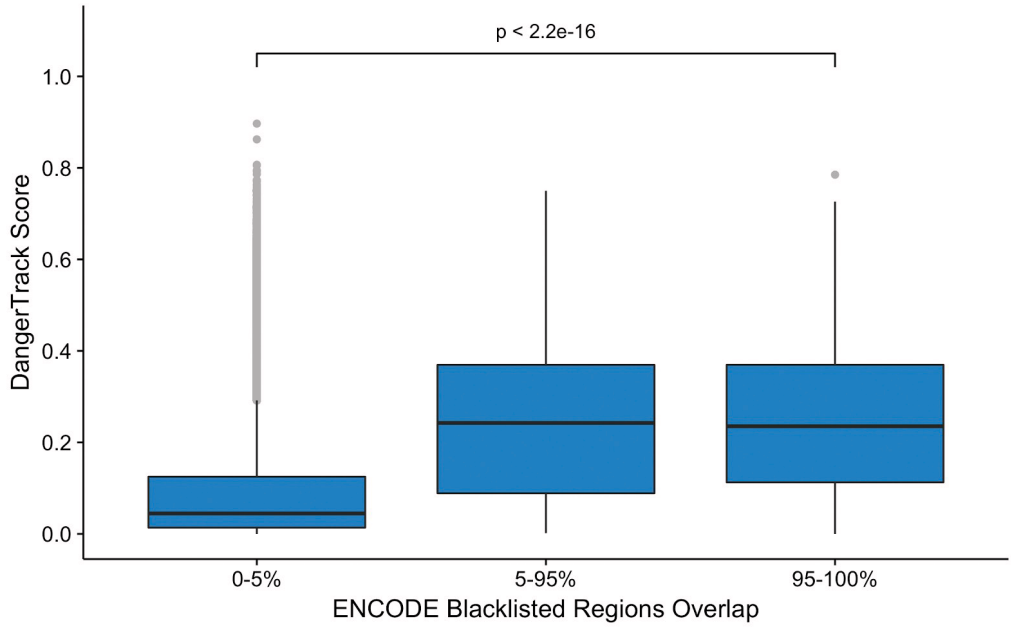

Figure 3. Comparison between DangerTrack score and ENCODE blacklisted regions.

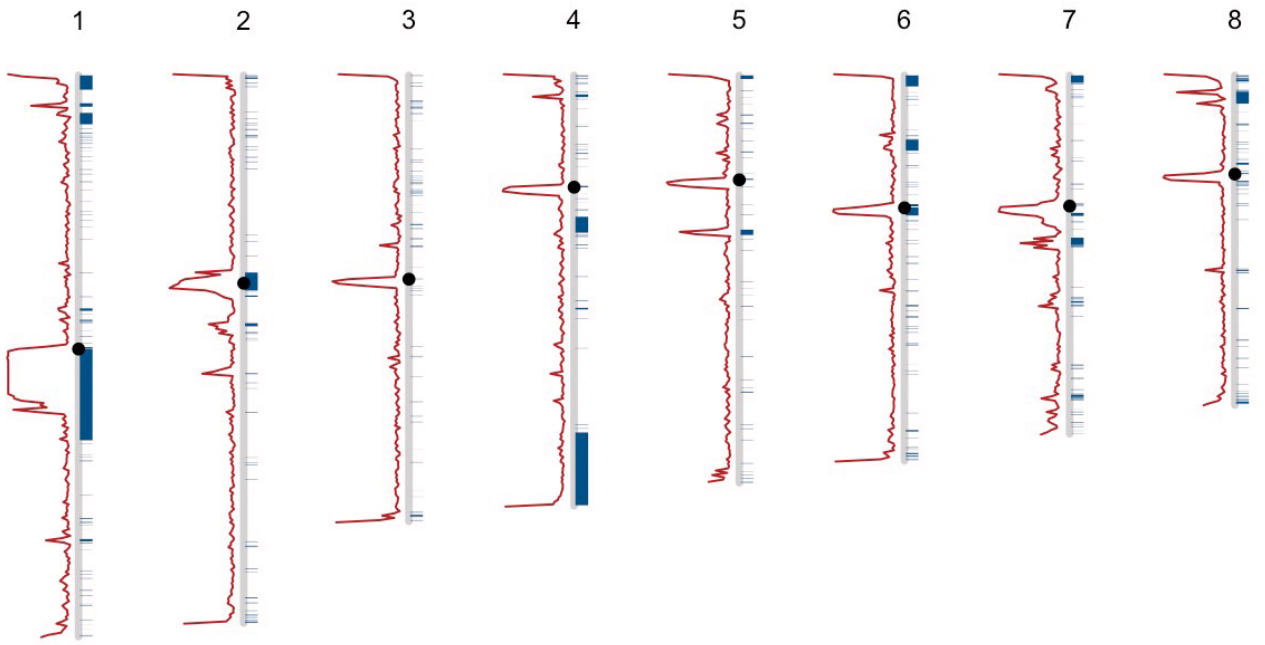

Figure 4. Comparison of DangerTrack score (red) and known blacklisted regions of ENCODE and GRC (blue) along chromosomes 1–8.

Conclusions and next steps

The results of DangerTrack overlap with previously-established troubling regions from the ENCODE blacklist and with regions of assembly error identified by the Genome Reference Consortium. Furthermore, we identified 48,891 5 kb regions (7.9% of all regions) that are not trustworthy. Thus, the mappability score and the concentration of SV breakpoints in a region indicate that the region is less reliable for SNP calling alone. This difficulty may be due to a high degree of difference in the reference sequence or the number of unresolved regions. While we showed that DangerTrack is capable of capturing these challenges for hg19, this method isuniversally applicable regardless of organism. The mappability tracks can be established easily and SV calls from other organisms can be incorporated. Nevertheless, DangerTrack is only a first step in understanding the underlying complexity of certain regions. Future work will include a revised weighting of the individual tracks.

Software availability

The code for the pipeline and the resulting genome-wide DangerTrack score are publically available at: https://github.com/DCGenomics/DangerTrack

Archived source code as at time of publication: doi, 10.5281/zenodo.438344 (igor & DCGenomics, 2017).

License: MIT

Author contributions

I.D., F.J.S., and B.B participated in designing the study, carrying out the research, and preparing the manuscript. I.D., F.J.S., and B.B were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

F.S. was supported through a National Science Foundation award (DBI-1350041) and National Institutes of Health award (R01-HG006677). B.B. was supported by the Intramural Research Program of the National Library of Medicine.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors wish to thank NCBI Hackathon organizers, Cold Spring Harbor Labs, Mike Schatz, Vamsi Kodapalli. The authors also thank Lisa Federer, NIH Library Writing Center, for manuscript editing assistance.

References

Degner JF, Marioni JC, Pai AA, *et al.*: **Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data.** *Bioinformatics.* 2009; **25**(24): 3207–3212.
PubMed Abstract | Publisher Full Text | Free Full Text

Dilthey A, Cox C, Iqbal Z, *et al.*: **Improved genome inference in the MHC using a population reference graph.** *Nat Genet.* 2015; **47**(6): 682–688.
PubMed Abstract | Publisher Full Text | Free Full Text

Graze RM, Novelo LL, Amin V, *et al.*: **Allelic imbalance in *Drosophila* hybrid heads: exons, isoforms, and evolution.** *Mol Biol Evol.* 2012; **29**(6): 1521–1532.
PubMed Abstract | Publisher Full Text | Free Full Text

Igor, DCGenomics: **NCBI-Hackathons/DangerTrack: DangerTrack Release 1.1 [Data set].** *Zenodo.* 2017.
Data Source

Jeffares DC, Jolly C, Hoti M, *et al.*: **Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast.** *Nat Commun.* 2017; **8**: 14061.
PubMed Abstract | Publisher Full Text | Free Full Text

Satya RV, Zavaljevski N, Reifman J: **A new strategy to reduce allelic bias in RNA-Seq readmapping.** *Nucleic Acides Res.* 2012; **40**(16): e127.
PubMed Abstract | Publisher Full Text | Free Full Text

Sudmant PH, Rausch T, Gardner EJ, *et al.*: **An integrated map of structural variation in 2,504 human genomes.** *Nature.* 2015; **526**(7571): 75–81.
PubMed Abstract | Publisher Full Text | Free Full Text

Zook JM, Catoe D, McDaniel J, *et al.*: **Extensive sequencing of seven human genomes to characterize benchmark reference materials.** *Sci Data.* 2016; **3**: 160025.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ? ✓ ✓

Version 1

Referee Report 24 April 2017

doi:10.5256/f1000research.12141.r21627

✓ **Andrew Carroll**
DNAnexus, Mountain View, CA, USA

This work uses the frequency of detected structural variant events as a proxy for reference quality and thus as a measure of trustworthiness for variant calls in a region. The foundation for the work is a consensus set of multiple SV methods applied in both Genome in a Bottle and 1000 genomes.

The concept is interesting and well applied. The analysis is clearly communicated, and the code is both openly available and interpret-able.

I do have a minor issue with the communication of the issue. The article doesn't get specific about why performing analysis in these genomic regions might be "dangerous", which I would categorize as:

1. Regions which represent a problem in the reference itself (either regions of mis-assembly, or locations where a rare SV event was present in the sequence used which the rest of the population doesn't have)

2. Regions which may contain "SV hotspots"

3. Low complexity regions, centromeric regions, telomeric regions.

4. Segmental duplications with insufficient divergence, mobile elements.

5. Regions of high heterozygosity in the population - which may or may not be SV in origin but due to their diversity may manifest as SV events.

Since the paper doesn't specifically break down these possibilities, I feel that the tone of the discussion would imply to a naive reader that these regions are "dangerous" because they are SV hotspots, when I think the authors would agree with me that reference assembly issues, low complexity regions, and segmental duplications are probably responsible for most of the "dangerous" regions.

It might be worthwhile to spend a few sentences to explain why a region might be dangerous, though the paper is written so tightly around the problem that it might distract the paper from its core point.

Separately, another distinction it might be worthwhile to make is that the SVs from the 1000 genomes project are quite different than the HG002 SVs. In one case, we have SVs generated over a population using more limited sequencing technology. HG002 is very different, where we are instead generating a wealth of calls from a single sample. Both approaches are worthwhile and it is appropriate to combine

them. It would be interesting to note whether there are differences between the types of regions the two approaches identify.

From this point, one could ask a number of additional questions beyond the scope of the manuscript. For example - which of these regions are no longer SV hotspots when using longer reads.

Hopefully authors of SV tools will begin to take this DangerTrack into account when developing their methods - either to give quality values to their calls, or as a means to separate calling on hard and easy regions of the genome.

Overall, the work is sound and the paper very well written.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 24 April 2017

doi:10.5256/f1000research.12141.r21624

Justin M. Zook
Material Measurement Laboratory, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA

The authors present a tool that finds potentially problematic regions for small variant calling, based on mappability and previously called SVs. It's an interesting and useful tool, and I have a few suggestions for improvement:

1. It is worth noting that the 1000 Genomes SVs were discovered using only short reads in thousands of individuals, and the GIAB SVs were discovered using short and long reads in only one

mother-father-son trio. There are nuances to this that would be useful to discuss. In particular, GIAB SV locations may or may not be SVs in other individuals, and because they are not highly curated, they may contain FPs around repetitive regions, but these are still good regions to identify as problematic for small variant calling

2. I don't see the script for running SURVIVOR commands and generating the dangertrack bed files in the GitHub site, which would be useful to include to reproduce the results.

3. When the authors say "Furthermore, we identified 48,891 5 kb regions (7.9% of all regions) that are not trustworthy," what dangertrack score threshold do they use (e.g., ==1, >0.9, >0.5, ...)?

4. It may be interesting to compare to the GIAB high-confidence bed file for NA12878 (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.2/GRCh37/HG001_GR) and the Platinum Genomes bed file for NA12878 (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/technical/platinum_genomes/2016-1.0/hg19/small_variants/), since these are both regions where they purport to make high-confidence small variant calls.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 24 April 2017

doi:10.5256/f1000research.12141.r21625

? **Melissa A. Gymrek**[1,2], **Nima Mousavi**[3]
[1] Department of Medicine, University of California, San Diego, La Jolla, CA, USA
[2] Department of Computer Science and Engineering, University of California, La Jolla, CA, USA
[3] Department of Electrical and Computer Engineering, University of California, La Jolla, CA, USA

The authors present DangerTrack, a method to score regions of the genome that are likely to be problematic for variant calling. The method combines information from structural variant catalogs and mappability scores across the genome into a single track whose score is meant to correlate with the “trustworthiness” of a region.

Assessing which regions of the genome are likely to be error-prone for variant calling is indeed an important issue, especially for the use case mentioned of clinicians and biologists interested in particular variants. The manuscript is for the most part well-written and the method is clearly described. However, the rationale for developing a new annotation track on top of existing annotations such as the ENCODE “black list” is not well described, and the authors do not provide sufficient evidence of the claim that DangerTrack successfully classifies “difficult” regions of the genome for variant calling. These and other concerns are outlined in more detail below.

Major comments:

- The rationale and goal are not clearly defined: What was the primary rationale for developing DangerTrack? It was not immediately clear why another annotation is needed, given that tracks such as the ENCODE blacklist and the NCBI problematic region list already exist. One potential reason is that those lists are inadequate, and we need a list that is better at picking out truly problematic regions for variant calling. If that was indeed the goal, the authors do not present sufficient evidence that their annotation is any better than the existing lists. On the other hand, another reasonable motivation to create this tool is that the ENCODE/NCBI lists were created manually, and could not be easily constructed for a new genome or a new individual. If that is the case, the authors should explicitly state early on that this was their primary motivation. Finally, there are other automated tools/tracks such as RepeatMasker and dustmasker that might be used to filter likely low quality variants. How does DangerTrack compare to those?

- Insufficient evaluation: The authors claim that their score, based on the # SV breakpoints/5kb, tracks with SNP call quality. However, this is never backed up with any evidence. Thus, it is impossible to tell whether this track actually adds any value in filtering low quality SNP calls. One potential validation would be to look at SNP quality scores or SNP call accuracy stratified by DangerTrack value, and show a relationship. If DangerTrack does a better job of classifying incorrect vs. true SNP calls than other tracks, then that would be clear evidence that it gives value added over existing tools. Similarly the discussion states that the authors identify ~48K “untrustworthy” regions, but there is no data to back up the statement that those regions are indeed enriched for incorrect calls.

Minor comments:

- Last paragraph of introduction, suggest to change “height of the score” to “magnitude of the score”

- Last sentence before “Evaluation of DangerTrack”, change “reassemble” to “resemble”

- Same sentence, how are the mappability tracks related to base-clipped reads?

- How did the authors decide on the weighting scheme to combine different features?

- The low overlap between 1000 Genomes and GIAB breakpoints raises concerns over how reproducible DangerTrack will be and how sensitive it is to the quality of the SV catalog used as input.

- Figures 2 and 3 are not well described, it was not clear what is being depicted.
- The authors indicate that there is “very high correlation” with the ENCODE blacklist track. This should be stated more quantitatively, in terms of e.g. correlation or % overlap.

Is the rationale for developing the new software tool clearly explained?
No

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

Referee Expertise: Bioinformatics

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.