Check for updates

SOFTWARE TOOL ARTICLE

The third international hackathon for applying insights into large-scale genomic composition to use cases in a wide range of organisms [version 1; peer review: 1 approved, 3 approved with reservations]

Kimberly Walker[1], Divya Kalra[1], Rebecca Lowdon[2], Guangyi Chen[3,4], David Molik[5], Daniela C. Soto[6], Fawaz Dabbaghie[3,7], Ahmad Al Khleifat[8], Medhat Mahmoud[1], Luis F Paulin[1], Muhammad Sohail Raza[9], Susanne P. Pfeifer[10], Daniel Paiva Agustinho[11], Elbay Aliyev[12], Pavel Avdeyev[13], Enrico R. Barrozo[14], Sairam Behera[1], Kimberley Billingsley[15], Li Chuin Chong[16], Deepak Choubey[17], Wouter De Coster[18,19], Yilei Fu[20], Alejandro R. Gener[21], Timothy Hefferon[22], David Morgan Henke[23], Wolfram Höps[24], Anastasia Illarionova[25], Michael D. Jochum[14], Maria Jose[26], Rupesh K. Kesharwani[1], Sree Rohit Raj Kolora[27], Jędrzej Kubica[28], Priya Lakra[29], Damaris Lattimer[30], Chia-Sin Liew[31], Bai-Wei Lo[32], Chunhsuan Lo[33], Anneri Lötter[34], Sina Majidian[35], Suresh Kumar Mendem[36], Rajarshi Mondal[37], Hiroko Ohmiya[38], Nasrin Parvin[37], Carolina Peralta[39], Chi-Lam Poon[40], Ramanandan Prabhakaran[41], Marie Saitou[42], Aditi Sammi[43], Philippe Sanio[44], Nicolae Sapoval[20], Najeeb Syed[12], Todd Treangen[20], Gaojianyong Wang[45], Tiancheng Xu[20], Jianzhi Yang[46], Shangzhe Zhang[47], Weiyu Zhou[48], Fritz J Sedlazeck[1], Ben Busby[49]

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, 77030, USA
[2]Bayer Crop Science, Chesterfield, MO, 63017, USA
[3]Drug Bioinformatics, Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Saarbrücken, Germany
[4]Center for Bioinformatics, Saarland University, Saarbrücken, Germany
[5]Tropical Crop and Commodity Protection Research Unit, Pacific Basin Agricultural Research Center, Hilo, HI, 96720, USA
[6]Biochemistry & Molecular Medicine, Genome Center, MIND Institute, University of California, Davis, Davis, CA, 95616, USA
[7]Institute for Medical Biometry and Bioinformatics, University hospital Düsseldorf, Düsseldorf, Germany
[8]Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK
[9]CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Beijing, China
[10]Center for Evolution and Medicine, Arizona State University, Tempe, AZ, USA
[11]Department of Molecular Microbiology, Washington University in St. Louis School of Medicine, St. Louis, MO, 63110, USA
[12]Research Department, Sidra Medicine, Doha, Qatar
[13]Computational Biology Institute, The George Washington University, Washington, DC, 20052, USA
[14]Department of Obstetrics & Gynecology, Baylor College of Medicine, Houston, TX, 77030, USA
[15]Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD,

USA
[16]Beykoz Institute of Life Sciences and Biotechnology, Bezmialem Vakif University, Beykoz, Istanbul, Turkey
[17]Department of Technology, Savitribai Phule Pune University, Pune, Maharashtra, India
[18]Applied and Translational Neurogenomics Group, VIB Center for Molecular Neurology, Antwerp, Belgium
[19]Applied and Translational Neurogenomics Group, Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium
[20]Department of Computer Science, Rice University, Houston, TX, USA
[21]Association of Public Health Labs, Centers for Disease Control and Prevention, Downey, CA, USA
[22]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20892, USA
[23]Department Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, 77030, USA
[24]EMBL Heidelberg, Genome Biology Unit, Heidelberg, Germany
[25]German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany
[26]Centre for Bioinformatics, Pondicherry University, Pondicherry, India
[27]University of California Berkeley, Berkeley, CA, USA
[28]University of Warsaw, Warsaw, Poland
[29]Department of Zoology, University of Delhi, Delhi, India
[30]University of Applied Sciences Upper Austria - FH Hagenberg, Mühlkreis, Austria
[31]Center for Biotechnology, University of Nebraska-Lincoln, Lincoln, Nebraska, 68588, USA
[32]Department of Biology, University of Konstanz, Konstanz, Germany
[33]Human Genetics Laboratory, National Institute of Genetics, Japan, Mishima City, Japan
[34]Department of Biochemistry, University of Pretoria, Pretoria, South Africa
[35]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland
[36]ICAR-NIVEDI, Bangalore, Karnataka, India
[37]Department of Biotechnology, The University of Burdwan, West Bengal, India
[38]Genetic Reagent Development Unit, Medical & Biological Laboratories Co., Ltd., Tokoyo, Japan
[39]Max Planck Institute for Evolutionary Biology, Plon, Germany
[40]Weill Cornell Medicine, New York, NY, USA
[41]Hoffmann-La Roche Limited, Regions, Diagnostics & Research (RDR), Mississauga, Canada
[42]Center of Integrative Genetics (CIGENE),Faculty of Biosciences, Norwegian University of Life Sciences, As, Norway
[43]School of Biochemical Engineering, Indian Institute of Technology (BHU), Varanasi, Uttar Pradesh, India
[44]University of Applied Sciences Upper Austria - FH Hagenberg, Hagenberg im Mühlkreis, Austria
[45]Max Planck Institute for Molecular Genetics, Berlin, Germany
[46]Department of Quantitative and Computational Biology,, University of Southern California, Los Angeles, CA, USA
[47]School of Biology, University of St Andrews, St Andrews, UK
[48]Department of Statistical Science, George Mason University, Fairfax, Virginia, USA
[49]DNAnexus, Mountain View, CA, USA

v1 **First published:** 16 May 2022, **11**:530
https://doi.org/10.12688/f1000research.110194.1
Latest published: 16 May 2022, **11**:530
https://doi.org/10.12688/f1000research.110194.1

Abstract

In October 2021, 59 scientists from 14 countries and 13 U.S. states collaborated virtually in the Third Annual Baylor College of Medicine & DNANexus Structural Variation hackathon. The goal of the hackathon was to advance research on structural variants (SVs) by prototyping and iterating on open-source software. This led to nine hackathon projects focused on diverse genomics research interests, including various SV discovery and genotyping methods, SV sequence reconstruction, and clinically relevant structural variation, including SARS-CoV-2 variants. Repositories for the projects that participated in the hackathon are available at

Open Peer Review

Approval Status ? ✓ ? ?

	1	2	3	4
version 1 16 May 2022	? view	✓ view	? view	? view

1. **Rodolfo Aramayo**, Texas A&M University, College Station, USA
2. **Nguyen Quoc Khanh Le**, Taipei Medical University, Taipei, Taiwan

https://github.com/collaborativebioinformatics.

Keywords
Structural variants, k-mer, Covid-19, Long-reads, Tomatoes, Cancer, Viral integration, Hackathon, NGS

This article is included in the Agriculture, Food and Nutrition gateway.

This article is included in the Emerging Diseases and Outbreaks gateway.

This article is included in the Sidra Medicine gateway.

This article is included in the Python collection.

This article is included in the Max Planck Society collection.

This article is included in the Hackathons collection.

3. **Pedro G. Ferreira**, University of Porto, Porto, Portugal

4. **Quan Long**, University of Calgary, Calgary, Canada

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Kimberly Walker (kw5@bcm.edu), Divya Kalra (divyak@bcm.edu), Rebecca Lowdon (rebecca.lowdon@bayer.com), Guangyi Chen (Guangyi.Chen@helmholtz-hips.de), Fritz J Sedlazeck (fritz.sedlazeck@bcm.edu), Ben Busby (bbusby@dnanexus.com)

Author roles: Walker K: Formal Analysis, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Kalra D**: Formal Analysis, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Lowdon R**: Formal Analysis, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Chen G**: Formal Analysis, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Molik D**: Formal Analysis, Methodology, Software, Visualization, Writing – Review & Editing; **Soto DC**: Formal Analysis, Methodology, Software, Visualization, Writing – Review & Editing; **Dabbaghie F**: Formal Analysis, Methodology, Software, Writing – Review & Editing; **Khleifat AA**: Formal Analysis, Methodology, Software, Writing – Review & Editing; **Mahmoud M**: Formal Analysis, Methodology, Software, Writing – Review & Editing; **Paulin LF**: Formal Analysis, Methodology, Software, Writing – Review & Editing; **Raza MS**: Formal Analysis, Methodology, Software, Writing – Review & Editing; **Pfeifer SP**: Formal Analysis, Methodology, Software, Writing – Review & Editing; **Agustinho DP**: Formal Analysis, Methodology, Software; **Aliyev E**: Formal Analysis, Methodology, Software; **Avdeyev P**: Formal Analysis, Methodology, Software; **Barrozo ER**: Formal Analysis, Methodology, Software; **Behera S**: Formal Analysis, Methodology, Software; **Billingsley K**: Formal Analysis, Methodology, Software; **Chong LC**: Formal Analysis, Methodology, Software; **Choubey D**: Formal Analysis, Methodology, Software; **De Coster W**: Formal Analysis, Methodology, Software; **Fu Y**: Formal Analysis, Methodology, Software; **Gener AR**: Formal Analysis, Methodology, Software; **Hefferon T**: Formal Analysis, Methodology, Software; **Henke DM**: Formal Analysis, Methodology, Software; **Höps W**: Formal Analysis, Methodology, Software; **Illarionova A**: Formal Analysis, Methodology, Software; **Jochum MD**: Formal Analysis, Methodology, Software; **Jose M**: Formal Analysis, Methodology, Software; **Kesharwani RK**: Formal Analysis, Methodology, Software; **Kolora SRR**: Formal Analysis, Methodology, Software; **Kubica J**: Formal Analysis, Methodology, Software; **Lakra P**: Formal Analysis, Methodology, Software; **Lattimer D**: Formal Analysis, Methodology, Software; **Liew CS**: Formal Analysis, Methodology, Software; **Lo BW**: Formal Analysis, Methodology, Software; **Lo C**: Formal Analysis, Methodology, Software; **Lötter A**: Formal Analysis, Methodology, Software; **Majidian S**: Formal Analysis, Methodology, Software; **Mendem SK**: Formal Analysis, Methodology, Software; **Mondal R**: Formal Analysis, Methodology, Software; **Ohmiya H**: Formal Analysis, Methodology, Software; **Parvin N**: Formal Analysis, Methodology, Software; **Peralta C**: Formal Analysis, Methodology, Software; **Poon CL**: Formal Analysis, Methodology, Software; **Prabhakaran R**: Formal Analysis, Methodology, Software; **Saitou M**: Formal Analysis, Methodology, Software; **Sammi A**: Formal Analysis, Methodology, Software; **Sanio P**: Formal Analysis, Methodology, Software; **Sapoval N**: Formal Analysis, Methodology, Software; **Syed N**: Formal Analysis, Methodology, Software; **Treangen T**: Formal Analysis, Methodology, Software; **Wang G**: Formal Analysis, Methodology, Software; **Xu T**: Formal Analysis, Methodology, Software; **Yang J**: Formal Analysis, Methodology, Software; **Zhang S**: Formal Analysis, Methodology, Software; **Zhou W**: Formal Analysis, Methodology, Software; **Sedlazeck FJ**: Conceptualization, Formal Analysis, Methodology, Resources, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Busby B**: Conceptualization, Formal Analysis, Methodology, Resources, Software, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: Ben Busby is a full-time employee of DNAnexus. Rebecca Lowdon is a full-time employee of Bayer Crop Sciences. Ramanandan Prabhakaran is a full-time employee of Hoffmann-La Roche Limited. Luis F Paulin is sponsored by Genentech, Inc. Wouter De Coster has received travel reimbursement and free consumables from ONT. FJS received research support from ONT and PacBio. Alejandro Rafael Gener is an editorial board member of AIDS, and has received poster bursaries from ONT in 2019.

Grant information: Tim Heffernon is supported by the intramural research program of the National Library of Medicine. AAK is funded by ALS Association Milton Safenowitz Research Fellowship, The Motor Neurone Disease Association (MNDA) Fellowship (Al Khleifat/Oct21/975-799) and The NIHR Maudsley Biomedical Research Centre. SPP is supported by a National Science Foundation CAREER grant (DEB-2045343). Marie Saitou is supported by The Research Council of Norway (SalmoSV, grant number 325874) Wouter De Coster is supported by a postdoctoral fellowship from the FWO (1233221N). SM (Sina Majidian) is supported by the Swiss National Science Foundation, Grant number 186397. David Molik (DCM) is supported by the USDA Agricultural Research Service HQ Research Associate program in Big Data Shangzhe Zhang is funded by China Scholarship Council PhD scholarship 202106180022 ARG: "This publication was supported by Cooperative Agreement Number NU60OE000104-02, funded by the Centers for Disease Control and Prevention through the Association of Public Health Laboratories. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Association of Public Health Laboratories."
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Walker K *et al*. 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. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

How to cite this article: Walker K, Kalra D, Lowdon R *et al.* **The third international hackathon for applying insights into large-scale genomic composition to use cases in a wide range of organisms [version 1; peer review: 1 approved, 3 approved with reservations]** F1000Research 2022, **11**:530 https://doi.org/10.12688/f1000research.110194.1

First published: 16 May 2022, **11**:530 https://doi.org/10.12688/f1000research.110194.1

Introduction

One of the processes by which genomes incur deleterious changes are commonly linked to the genetic signatures known as structural variants (SVs). SVs are large genomic alterations, where large is typically (and somewhat arbitrarily) defined as encompassing at least 50 base pairs (bp). These genomic variants are typically classified as deletions, duplications, insertions, inversions, and translocations describing different combinations of DNA gains, losses, or rearrangements. Copy number variations (CNVs) are a particular subtype of SVs mainly represented by deletions and duplications. SVs are typically described as single events, although more complex scenarios involving combinations of SV types exist.[1,2] Understanding how and why SVs occur can help gain a deeper understanding of evolutionary processes driving species divergence and phenotypic adaptation, genomic processes leading to genetic variation and etiologies of plant and animal diseases.[3] With a recent deluge of available genomic data, SVs are an optimal target for computational biology research.[4]

In October 2021, 59 researchers from 14 countries participated virtually in the third Baylor College of Medicine & DNAnexus hackathon, focusing on interrelated topics such as SVs, short tandem repeats (STRs), *k*-mer profiling, viruses, reference refinement and annotation. The hackathon groups addressed questions around: the use of SVs in the localization and understanding of quantitative trait loci (QTL), reference-free analysis of SVs, parallelization of SV workflows, the assessment and refining the quality of detected SVs, use of SVs in the understanding of adaptation in viruses, and understanding genetic signatures of diseases through SVs. The international hackathon focused on nine softwares to answer these questions; eight of which we present in this paper: STRdust, kTom, INSeption, GeneVar2, cov2db, K-var, Imavirus, and a Reference Panel Generator (RPG) for diverse sequencing data analysis. Several emergent themes became apparent over the course of the hackathon.

QTLs link a phenotypic trait to a local genomic region, and in its broadest definition, a molecular change affecting a phenotype.[5] A direct connection can be drawn between some SVs and QTLs. Linking traits and their genetic underpinnings is a common practice in the fields of agricultural genomics, molecular evolution, and genetic disease research.[6] Structural variation is one possible genomic change that could result in a QTL. This year's hackathon featured work on tomatoes and other plants which provided an alternative viewpoint to the generally human-focused research of previous hackathons. Such cross-disciplinary research allows disparate groups working on similar problems to push the envelope of what is possible with current technologies.

Nucleotide sequence substrings of length *k* (*k*-mers) continue to prove useful in SV work and in genomics, however, the time needed to assess the frequency of SVs presents a resource problem.[7] The reduction of the computational resources required to complete an SV assessment in a genome would allow greater amounts of SV data to be processed in genomic workflows. Many bioinformatic tools currently used to locate genomic SVs use a sliding window alignment technique, which can be time-consuming.[8,9] However, implementing a *k*-mer based approach to create a pool of reference *k*-mers of known SVs, the annotation speed of variation in new genomes might be increased.[10,11] *k*-mers have also been used in alignment-free methods, bypassing the need for reference genomes.[12]

A portion of the hackathon focused on virus work. At the time of the hackathon, the COVID-19 pandemic was ongoing and the question of what SVs are present, and how they might change the behavior of SARS-CoV-2 was unresolved.

Together the projects of this hackathon represent a range of fields, a range of academic, industry, and government researchers, and a range of desired impacts in the field of SV analysis. Topical introductions to the specific work of each group can be found below, except from "nibSV" which was reported previously[11] and did not achieve significant progress.

STRdust: Detect and genotype short tandem repeats

Short tandem repeats (STRs) (*i.e.,* repeated instances of short 2-6 bp DNA motifs) are widespread in the genomes of most organisms. Due to their highly polymorphic nature, STRs are frequently employed in population and evolutionary genomic studies ranging from genealogy to forensics and disease diagnostics.[13] For example, in humans, expansions in functional STRs have been linked to many neurological and developmental disorders[14,15] whereas in plants, STRs have been found to impact several traits important to agriculture including growth rate and yield.[16] Yet, despite their importance, STRs remain relatively poorly characterized in most species. On the one hand, second-generation sequencing platforms (*e.g.,* Illumina[17] (RRID:SCR_010233)) are limiting our view of STR variation within the read length due to both the short length of sequencing reads produced as well as frequent amplification biases (such as GC-biases and over−/under-representation of certain reads on a genome-wide scale). On the other hand, third-generation sequencing platforms (namely, PacBio (PacBio Sequel II System,[18] (RRID:SCR_017990)) and Oxford Nanopore Technologies (ONT)[19] (RRID:SCR_003756)) allow for the generation of single-molecule reads spanning tens to hundreds of kilobases in length but error rates (~1% in PacBio HiFi reads and ~ 10–15% in ONT[20]) continue to exacerbate reliable STR

detection. To mitigate this issue, several long-read STR calling methods have been developed in recent years, including PacmonSTR[21](RRID:SCR_002796), NanoSatellite,[22] TRiCoLOR[23](RRID:SCR_018801), and Straglr[24] – however, their usability remains limited due to platform and/or computational demands. In order to address these shortcomings, we introduce STRdust, a tool to accurately detect and genotype STRs from long reads.

kTom: k-mers for profiling tomato introgressions

The success of commercially cultivated vegetables requires a balance of selection for domestication traits while maintaining genomic diversity and quality characteristics, and this is particularly true for tomato breeding programs.[25] Many desirable traits for crops are obtained by crossing elite breeding germplasm to wild relatives that carry a trait of interest (*e.g.*, disease resistance or fruit flavor). This process of moving a genomic region from one species or distantly-related species into another is called introgression.[25]

Tomato is an important crop and indispensable in the diet of many cultures and regions. The demand for fresh and processed tomatoes makes them one of the most important vegetables grown globally, with >180 million tons of tomatoes produced in 2019 worldwide (FAOSTAT).[26]

Genetic traits have been moved into cultivated tomatoes over the past several decades of tomato breeding through trait introgression. Identifying and tracking introgressed traits is a crucial function of modern tomato breeding.[25] The introgression of traits often occurs as large presence/absence structural variants with novel genes or sequences. Some introgressions can be completely defined by *de novo* sequencing and assembly, but this can be expensive for many samples and is not always successful for more complex genomic introgressions.[2] These complex structural variation patterns, coupled with the lack of reference genomes for many wild tomato relatives, complicate the efforts to locate or characterize the introgressed traits in the elite germplasm's genome. Consequently, most marker sets today rely exclusively on SNPs, which do not always track diverse tomato genetics.[27]

Here we present kTom, a tool to characterize the *k*-mer content of re-sequenced genomes and to identify *k*-mers that are unique to traited samples. kTom is a collection of off-the-shelf tools arranged to allow for a tractable characterization of *k*-mer frequencies in a population. We used re-sequenced tomato accessions for this demonstration, but the same approach can work for any species. Having a reference-free method to characterize and track introgression sequences will give researchers more agility to understand the nature of important traits.[28]

INSeption: Polishing structural variants

Some types of SVs, such as insertions, play a crucial role in shaping the genome and thus the function of each gene. For example, more than 50 percent of mammalian genomes include a repeating DNA sequence known as transposable elements.[29] Additionally, insertions can indicate an early tumorigenic event,[30] demonstrating a role in disease, making it crucial to accurately identify them.

Read-based SV calling methods broadly fall into the categories of alignment- and assembly-based approaches.[2] In alignment-based approaches, SVs are inferred from patterns of abnormal read mapping on an existing reference sequence.[2] Alignment-based approaches pose a popular method for calling SVs both from short-reads and long-reads, with a multitude of tools developed for both read mapping (*e.g.,* BWA[31](RRID:SCR_010910), Minimap2[32](RRID:SCR_018550), and NGMLR[33](RRID:SCR_017620)) and SV detection (*e.g.,* DELLY[34](RRID:SCR_004603) and SNIFFLES[33](RRID:SCR_004603)). A downside of alignment-based SV detection lies in the incomplete resolution of complex or large genomic rearrangements or insertions exceeding common read lengths.[35] By contrast, assembly-based approaches utilize *de novo* sequence assemblies computed directly on the sampled reads, circumventing any biases introduced by the use of reference sequences.[2] SVs are thereby called by aligning such assemblies against a reference and identifying local incongruencies. Commonly used tools include Canu[36] (RRID:SCR_015880) and Flye[37] (RRID:SCR_017016) for sequence assembly, Minimap2 and BlasR[38] for alignment against a reference and SGVar[35] and Paftools[32] for SV calling. Assembly-based approaches can resolve even complex rearrangements and long insertions, but the construction of high-quality, haplotype-resolved assemblies requires thorough quality control and typically a high quality and diversity of data.[39]

GeneVar2: Gene-centric data browser for structural variants

Next-generation sequencing (NGS) technologies can be a powerful source in uncovering underlying genetic causes of diseases, but significant challenges still remain for SV interpretation and clinical analysis for clinicians.[40] Although various tools are available to predict the pathogenicity of a protein-changing variant—a list of these is available at OpenCRAVAT—they do not always agree, further compounding the problem.[41]

Here we present GeneVar2: an open access, gene-centric data browser to support structural variant analysis. There are two ways to interact with GeneVar2. First, GeneVar2 takes an input of a gene name or an ID and produces a report that informs the user of all SVs overlapping the gene and any non-coding regulatory elements affecting expression of the gene. Second, users can upload variant call format (VCF) files from their analysis pipelines as input to GeneVar2. GeneVar2 will output clinically relevant information as well as provide useful visualizations of disease ontology and enrichment pathway analysis based on SV types.

cov2db: A low frequency variant database for SARS-CoV-2

Global SARS-CoV-2 sequencing efforts have resulted in a massive genomic dataset availability to the public for a variety of analyses. However, the two most common resources are genome assemblies (deposited in GISAID[41] (RRID:SCR_018251) and GenBank[42] (RRID:SCR_002760), for example) and raw sequencing reads. Both of these limit the quantity of information, especially with respect to variants found within the SARS-CoV-2 populations. Genome assemblies only contain common variants, which is not reflective of the full genomic diversity within a given sample (even a single patient derived sample represents a viral population within the host[43–46]). Raw sequencing reads on the other hand require further analyses in order to extract variant information, and can often be prohibitively large in size.

Thus, we propose cov2db; a database resource for collecting low frequency variant information for available SARS-CoV-2 data. As of October 2021 there were more than 1.2 million SARS-CoV-2 sequencing datasets in the Sequence Read Archive (SRA)[47](RRID:SCR_004891) and European Nucleotide Archive (ENA)[48] (RRID:SCR_006515). Our goal is to provide an easy to use query system, and contribute to a database of VCF files that contain variant calls for SARS-CoV-2 samples. We hope that such interactive databases will speed up downstream analyses and encourage collaboration.

K-var: A "fishing" expedition for phenotype associated k-mers

k-mers are commonly used in bioinformatics for genome and transcriptome assembly, error correction of sequencing reads, and taxonomic classification of metagenomes.[49,50] More recently, *k*-mers have been used for genotyping of structural variations in large datasets in a mapping-free manner.[51] Sample comparison based on *k*-mers profiles provides a computationally efficient mapping-free way to address key differences between two biological conditions, avoiding the limitations of reference bias, mappability and sequencing errors.[52–54] Of particular interest are case-control studies, that allow to pinpoint genetic loci putatively implicated with a phenotype or a disease.

Here we develop a pipeline that takes a sample's sequencing data from two distinct conditions (ideally control vs. treatment or two different conditions) as input and compares their *k*-mer profiles in order to highlight *k*-mers associated with the phenotype. This approach was tested in a panel of cancer cell lines from the NCI-60 dataset (RRID:SCR_003057) contrasting primary and metastatic tissues to highlight mutational signatures underlying cancer progression.

Imavirus: Virus integration in disease

Viral infections impact human health as they can lead to short- and long-term diseases,[55] including cancers. Different forms of cancer are caused by viruses such as human papillomaviruses[56] and hepatitis B virus capable of integrating into the host genome.[57] Other viruses such as human immunodeficiency viruses (HIV) integrate into the host genome as a normal part of viral replication, contributing to cancer indirectly, and less commonly directly through insertional mutagenesis.[58] Knowing exactly where the integration events occur can help researchers and ultimately clinicians to better understand the effect of virus integration in disease.

Common assumptions about integrations are that they are single copy and show an absence of additional structural variability.[58] Different mechanisms might lead to different insertion site topology. For example, one would expect a difference between natural HIV-1 p31 integrase-mediated integration (insertion + tandem duplication of five bases of host target site) vs. insertion of viral genomic content (after reverse transcription in case of retroviruses like HIV) with host cell's DNA repair machinery. Such differences might include conservation of viral terminal repeat elements with virus-specific insertion signatures[59] vs. divergence[60] from this pattern.

When considering model insertion sites for assay evaluation, insertion site location heterogeneity exists to varying degrees in natural infection (with different mechanisms such as virus-dependent integration vs. host-dependent insertion contributing differently) vs. transgenic model organism (in the case of the Tg26 HIV-1 transgenic mouse, pronuclear injection and insertion of restriction enzyme-digested pNL4–3.[61] NL4–3 is the most common lab strain of HIV-1.[62]

With advances in sequencing technologies,[63,64] high-throughput sequencing data is available to explore viral genome integration space. Integration sites can be detected through identification of breakpoints between host and virus genome(s).[65] Some integrating viruses can produce run-on transcripts or may participate in trans-splicing between virus exon and downstream host exons.[66] Integration events have been previously detected by identifying these and other signatures such as chimeric reads in short-read sequencing (single-end and paired-end) and long-read sequencing.[65,67–75]

Here, we suggest tools and a general workflow that can be used for virus integration detection and discuss current caveats in using publicly available datasets for this type of analysis.

RPG: Reference Panel Generator

Despite great advances in our knowledge of NGS data analysis, a diverse complete reference genome sequence is lacking for humans. This leads to lack of sensitivity for detecting small insertions and deletions (INDELs) and structural variation, incomplete architecture of large polymorphic CNVs and correctly calling single nucleotide variants (SNVs) at complex genomic regions. High-quality Telomere-to-Telomere (T2T) CHM13 long-read genome assembly from T2T consortium[76] could be utilized as a reference panel to universally improve read mapping and variant calling.

Currently, we aim to provide a revised version of CHM13 reference panels along with an RPG pipeline based on 1000 Genomes Project[77] (RRID:SCR_006828) common allele calls and those abnormally avoided stop codons. Overall, such reference panels will greatly improve future population-scale diverse sequencing data analysis and correctly identify hundreds of thousands of novel per-sample variants in clinical settings.

Methods

DNAnexus (RRID:SCR_011884), a cloud platform, was used to run the code developed at the hackathon. It provides flexibility to run a wide array of software applications either on a cloud workstation (default number of cores = 8) or on an interactive environment such as a Jupyter notebook (default number of cores = 16). One of these two resources were used to run the software during the hackathon, unless otherwise specified.

STRdust

STRdust[142] parses the CIGAR (a compressed representation of an alignment that is used in the SAM file format) of each read, either genome-wide or in user-specified loci, in order to identify sufficiently large (>15 bp) insertions or soft-clipped bases which could indicate the presence of an enlarged STR. The sequence of those candidate-expansions is extracted, along with 50 bp of flanking sequence. Leveraging the phased input data, such insertions are combined per haplotype when multiple of these are found close by (within 50 bp) across multiple reads. The combination is done using spoa 4.0.7,[78] which generates a multiple sequence alignment and from that a consensus sequence. The obtained consensus sequence, in which inaccuracies inherent to the long read sequencing technologies should be reduced, is then used in mreps 6.2.01,[79] which will assess the repetitive character of the sequence and identify the repeat unit (Figure 1).

STRdust was tested against simulated STR datasets produced by SimiSTR. SimiSTR modified the GRCh38 (human) and SL4.0 (tomato) reference genome assemblies. Additional variation (SNVs) was introduced with SURVIVOR 1.0.7[80] at a rate of 0.001.

Long reads were simulated using SURVIVOR[80] for the GRCh38 (human) and SL4.0 (tomato) STR-modified genomes. Mapping was performed with Minimap2[32] 2.24 two-fold (with and without the -Y parameter), and phasing was done with longshot 0.4.1.[81] Default parameters were used for all tools, if not otherwise mentioned. STRdust results were compared to TRiCoLOR 1.1,[23] and Straglr 1.1.1[24] using default parameters. Figure 1 shows the workflow of STRdust described in this section.

STRdust is very easy to implement. One can, simply input the bam file after cloning the python script as follows: `python3 STRdust/STRDust.py mapped_long_reads.bam -o results_dir`. For further details on installation and implementation, review our github page.

kTom

kTom (*k*-mers for profiling Tomato introgressions)[143] aims to use *k*-mers to tag introgressions in elite tomato germplasm.

Current implementation

The kTom workflow (Figure 2a) processes re-sequenced genomes (only tested with Illumina short reads to date) to generate *k*-mer profiles per sample and calculates the population frequencies of these *k*-mers. Our use case is focused on

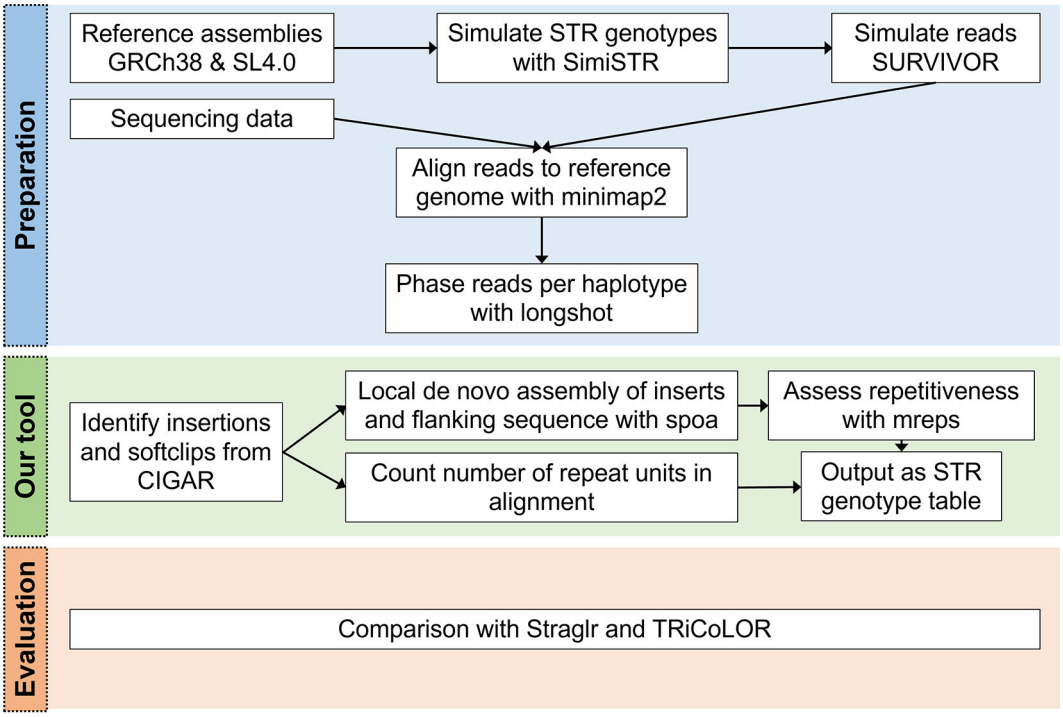

Figure 1. STRdust workflow. During the preparation phase, reads (either simulated or sequenced) are aligned to the corresponding reference genome with Minimap2[32] and the mapped reads are then phased using longshot. Next, STRdust identifies insertions and soft-clips from the Concise Idiosyncratic Gapped Alignment Report (CIGAR) string which identify regions of possible short tandem repeats (STR) expansion. These regions are further analyzed by performing *de novo* assembly using spoa and assessing the repetitiveness of the region with mreps. STRdust outputs the STR genotype as a tab separated table for further analysis. We evaluated STRdust by comparing the results of simulated STR expansions produced by SimiSTR based on the human (Genome Reference Consortium Human Build 38, GRCh38) and tomato (*Solanum lycopersicum* 4.0, SL4.0) reference genomes, to two novel tools: Straglr[24] and TRiCoLOR.[23]

k-mers with low-mid range frequencies, which we believe should capture *k*-mers unique to introgressed traits in our test population. Therefore, we use these *k*-mers to generate a distance matrix and understand the relatedness of samples.

To prototype the kTom workflow, we used 40 Whole Genome Shotgun (WGS) datasets from the 84 tomato or wild species accessions generated by The 100 Tomato Genome Sequencing Consortium[82] (BioProject PRJEB5235).

Data processing

Raw FASTQ files were quality-checked with FastQC version 0.11.9[83](RRID:SCR_014583) and trimmed with Flexbar version 1.4.0[84](RRID:SCR_013001), clipping five bases on 5′ and 3′ ends and keeping reads with quality score > 20 and a minimum length of 50. *k*-mers were counted using functions in Jellyfish version 2.3.0[85](RRID:SCR_005491) (`jellyfish count` followed by `jellyfish histo`) with `kmersize = 21`. The *k*-mers histogram was generated with Genomescope version 1.0.0[86](RRID:SCR_017014). *k*-mer counts for individual samples were then aggregated into a *k*-mer frequency matrix of *k*-mers as rows and samples as columns. This frequency matrix can be visualized as an interactive heatmap (example Figure 2b) by running `kmer_heatmap.R` which uses ComplexHeatmap version 2.8.0[87] (RRID:SCR_017270), InteractiveComplexHeatmap version 1.1.3[88] and tidyverse v1.3.1[89] (RRID:SCR_019186) R packages.

INSeption

INSeption[144] was tested using HiFi reads for sample HG002 (RRID:CVCL_1C78) retrieved from the genome in a bottle (GIAB) project.[90] The reads were aligned against GRCh37 using Minimap2[32] and Sniffles 1.012[33] was used to call SVs. We filtered out SVs that were supported by less than 10 reads using bcftools 1.12[91] (RRID:SCR_005227). We extracted insertions that are larger than 999 nucleotides. No reads span the entire insertion. Additionally, we filtered reads that were not aligned to reference using samtools 1.14[91] (RRID:SCR_002105), with the -f 4 option. Finally, we extracted reads that

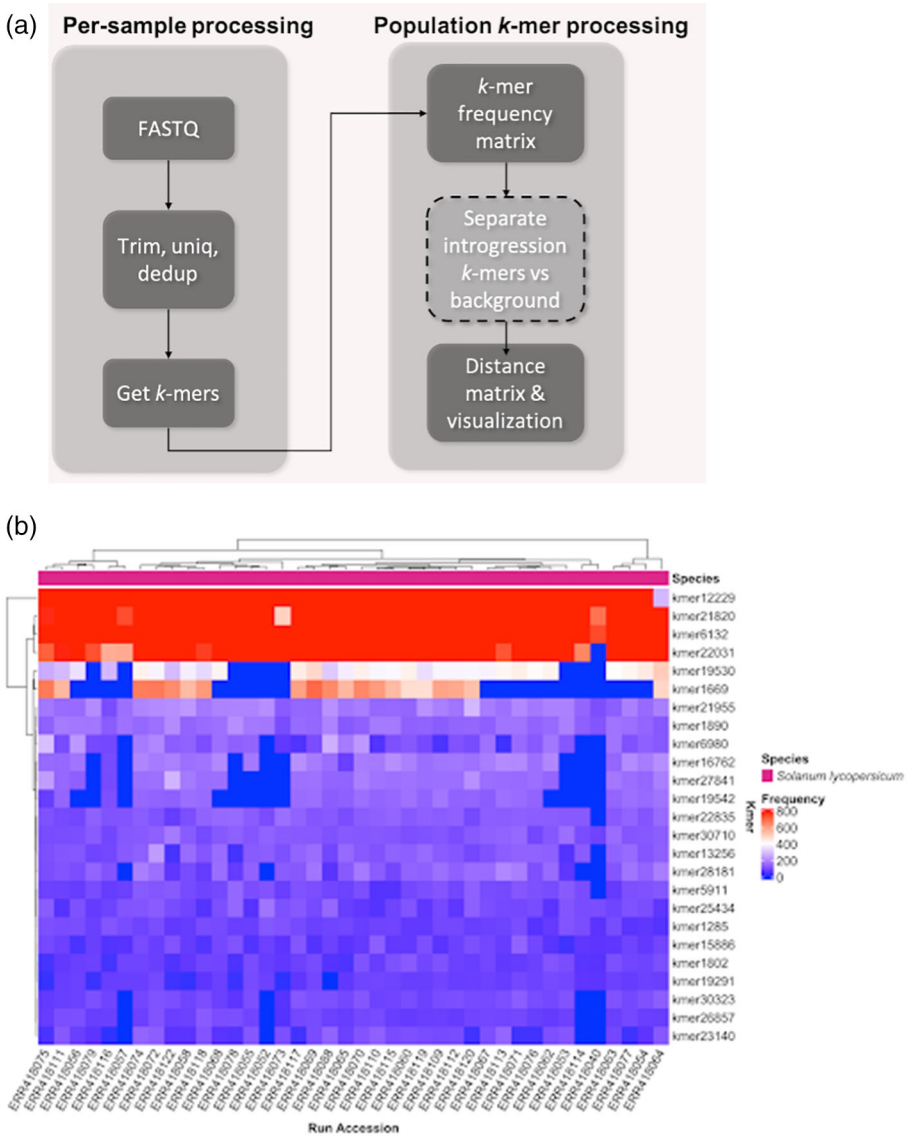

Figure 2. (a). kTom workflow, with major steps for individual sample and population data processing. (uniq = get unique reads; dedup = deduplicate reads). (b). *k*-mer frequency heatmap from kTom. Frequency of selected *k*-mers in each accession analyzed. Differential *k*-mer frequencies are apparent in this view. Depending on the nature of the accessions, this view may provide a first glimpse into genetic sequences underlying structural variations that differentiate the accessions.

support each insertion studied: first, we extracted read names from the SV file using bcftools and grouped them using SV ID, followed by extracting the FASTA sequence from the binary alignment map (BAM) file using samtools and awk (Figure 3a, left-hand side).

Allele frequency

For an analysis of the allele frequency (AF) for each mutation type, we created a Python[92,93] (RRID:SCR_008394) script (SVStat.py) that takes a VCF as input. For each SV type, it stores the AF and how often this AF was encountered. This data is then being visualized in *n* different plots (with *n* representing the number of SV types), where the x-axis represents the AF and the y-axis represents the number of times each SV type occurs.

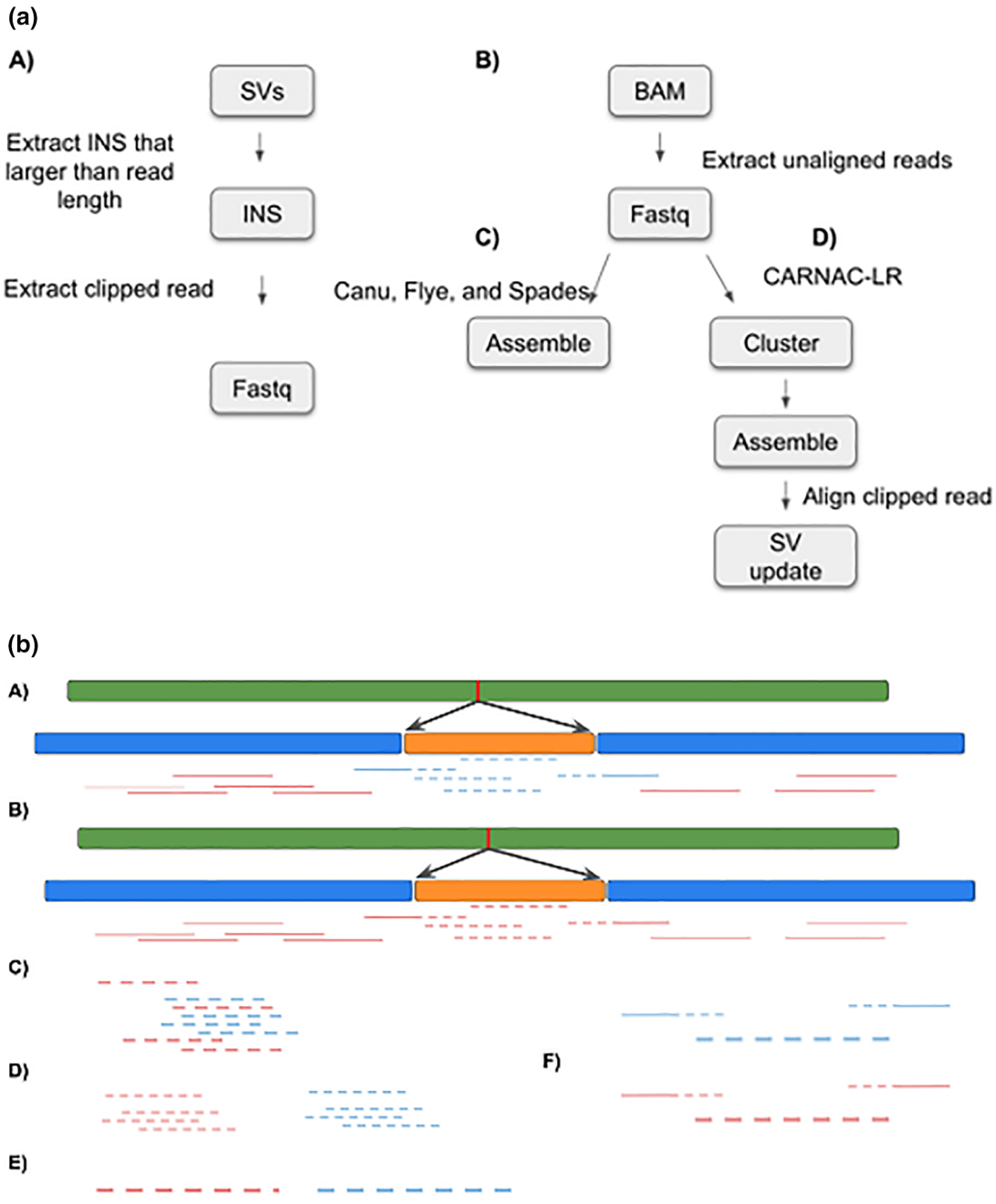

Figure 3. (a). INSeption workflow. Showing the tools used in the pipeline to detect insertion by extracting clipped reads (A), extracting unaligned reads (B), and then assembly (C) or clustering, assembling and aligning (D). SV: structural variant, INS:insertion, BAM: binary alignment map. (b). INSeption workflow, a graphical representation of the pipeline in (3a) showing two insertions, red and orange, in (A) and (B) we extract the unaligned reads (C), cluster them into groups (D), assemble each cluster (E) and finally align clipped reads to the assembled cluster (F).

Clustering unmapped reads

To be able to assemble a sequence from all unmapped reads, we tried several approaches. We attempted to identify clusters of reads using the LROD version 1.0[94] package, which we found unsuitable for our purposes due to long runtimes. More successfully, we used the program CARNAC-LR version 1.0.0[95] to build clusters of reads using Minimap2 version 2.22 aligner[32] and a subsequent *k*-mer based clustering approach. As output, for each cluster, all sequences and their IDs were exported into a FASTA file. On our testing dataset, we identified 64 such clusters. These clustered read files are then the basis for the next step for subsequent sequence assembly (Figure 3a right-hand side).

Delegate read clusters to the sequence assembler

All cluster.fasta files were loaded into the assembler programs (Flye version 2.9[37] and Spades version 3.15.3,[96] see software availability for input parameters) with another python script (clusterAssemble.py). This script has the ability to run a single cluster.fasta file or a whole batch within a directory. The inputs are the program location, program name, an optional flag: multi (for running the batch of clusters), an input directory or an input file, and an output directory (Figure 3a right-hand side continued).

Identifying integration sites for assembled clusters

Having successfully assembled contigs for N = 15 read clusters using Canu v2.2[36](RRID:SCR_015880), we searched for overlap of these contigs with the breakpoint regions of 30 previously identified long insertion sites. We reasoned that for each assembled contig which represents an insertion sequence, reads supporting the insertion breakpoint should also overlap with that specific contig. To find such contigs of interest, we first extracted the sequence reads (n = 604) which support a long inversion and therefore overlap at least one insertion breakpoint. This set of reads was then aligned against all 15 assembled contigs using Minimap2 (parameters: -x map-hifi -P), and using the contigs as a 'pseudo' reference. Finally, we manually inspected the resulting alignments to identify long (>3 kbp) contigs overlapping reads (Figure 3b).

GeneVar2

GeneVar2[145] is an update of GeneVar,[11] to help inform clinical interpretation of structural variants (Figure 4). It has expanded options allowing users to upload a VCF file, while maintaining its search functionality—based on gene name- on its web interface. GeneVar2 annotates the uploaded VCF file with a number of items which can then be downloaded by the user. Annotations include: SV allele frequency from gnomAD-SV[85](RRID:SCR_014964) and probability of being loss-of-function intolerant (pLI) from gnomAD; transcripts and coding regions of the impacting gene from GENCODE (v35)[97]; the gene associations with corresponding phenotype annotation from OMIM[100]; and known clinical SVs and their pathogenicity from dbVar.[86]

Additionally, when a user uploads a VCF file, an option to download graphs for visualizing SVs in the dataset, is available. There is an alternate format, comma-separated values (CSV), available to download with an annotated VCF. GeneVar2, written in R, is available on GitHub (Software availability section) with detailed instructions on installation and usage. GeneVar2 is a web-based application that can also be installed by an individual on their platform to run on the command line and launch locally. Instructions on how to build and run GeneVar2 on DNAnexus can be found here.

When users launch GeneVar2 as a web-application, they can enter individual gene names (HGNC[98](RRID: SCR_002827)), Ensembl[99] (RRID:SCR_002344) gene accession (ENSG) or Ensembl transcript accession (ENST) for extracting various SVs overlapping their gene of choice. GeneVar2 will output the gene-level summary with detailed information about the SVs within the gene. It links the gene information to databases such as OMIM[100] (RRID: SCR_006437), GTEx[101](RRID:SCR_013042), gnomAD and allele frequency is reported based on gnomAD genomes and exomes.

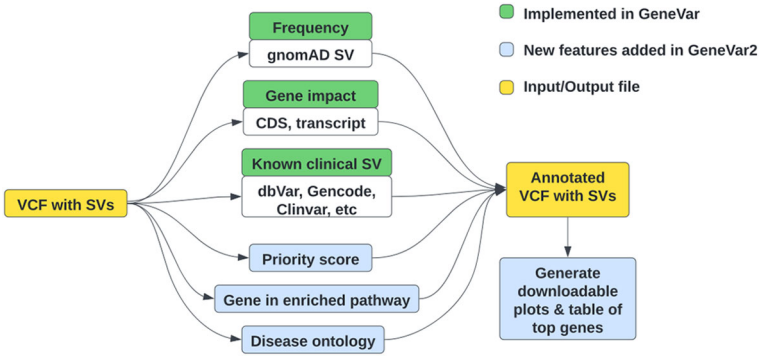

Figure 4. High-level outline of GeneVar2 workflow. Green boxes represent the initial features of GeneVar, implemented last year, while blue boxes represent new features implemented in GeneVar2 during this hackathon. (VCF: variant call format, SV: structural variation, CDS: coding sequence).

If users first need to call SVs on their samples, the developers recommend Parliament2[102](RRID:SCR_019187). Parliament2 runs a combination of tools to generate structural variant calls on whole-genome sequencing data. It can run the following callers: Breakdancer[103](RRID:SCR_001799), Breakseq2,[104] CNVnator[105](RRID:SCR_010821), Delly2,[34] Manta,[106] and Lumpy[107](RRID:SCR_003253). Because of synergies in how the programs use computational resources, these are all run in parallel. Parliament2 will produce the outputs of each of the tools for subsequent investigation. See the Parliament2 GitHub page for further details.

After users upload a VCF file containing SVs, GeneVar2 annotated each entry with the genes overlapping the SV, allele frequency from gnomAD-SV, and assigns a clinical rank to all the SVs in the VCF relative to each other. This is accomplished using the main annotation script *annotate_vcf.R.* The final annotated file is available for download as a VCF and CSV format. For Gene and Disease ontology and pathway analysis, *GeneAnnotationFromCSV. R* supports the enrichment analysis using KEGG[108–110](RRID:SCR_012773), Disease Ontology (DO),[111] Network of Cancer Gene[112] and Disease Gene Network (DisGeNET)[113] (RRID:SCR_006178). In addition, several visualization methods were provided by Bioconductor package *clusterprofiler*[114] (RRID:SCR_016884) and *enrichplot*[115] to help interpreting enrichment and disease ontology results.

Alternatively, if users prefer they can run GeneVar2 on the command line, by installing it on their platform. Users should have R version 4.1 or higher installed. In addition, you will need to have *sveval,* a custom R library, installed which can be accessed via *BiocManager* using *'jmonlong/sveval'*. Scripts and instructions can be found on GeneVar2's Github repository in the software availability section.

cov2db

cov2db[146] is implemented as a set of modular scripts which enable the user to annotate and reformat their original VCF files into mongoDB (RRID:SCR_021224) ready JavaScript object notation (JSON) documents. Namely, there are three key components provided within the code repository[1]: the VCF annotation and processing framework, together with the relevant software and scripts[2]; a sample set of annotated VCFs that can be used as a starting point for a SARS-CoV-2 iSNV database[3]; an R Shiny[116] (RRID:SCR_001626) app to facilitate a graphical user interface (GUI) for the interactions and quick summaries of the data within the database (Figure 5). The fields to query the cov2db database, such as annotation and variant information, are listed in the readme on our Github page. All of the above can be used to spin up an independent instance of cov2db and provide a user interface to interact with it. Minimal system requirements for a local cov2db instance are dictated by the mongoDB requirements with the key limiting factor being RAM used. Large variant databases will consume substantial amounts of RAM, and we suggest hosting those on dedicated high memory compute servers. Cov2db can run on x86 *nix-style platforms as is. We have not tested the software on ARM architectures or Windows based hosts. End users can interact with a hosted database from any web browser.

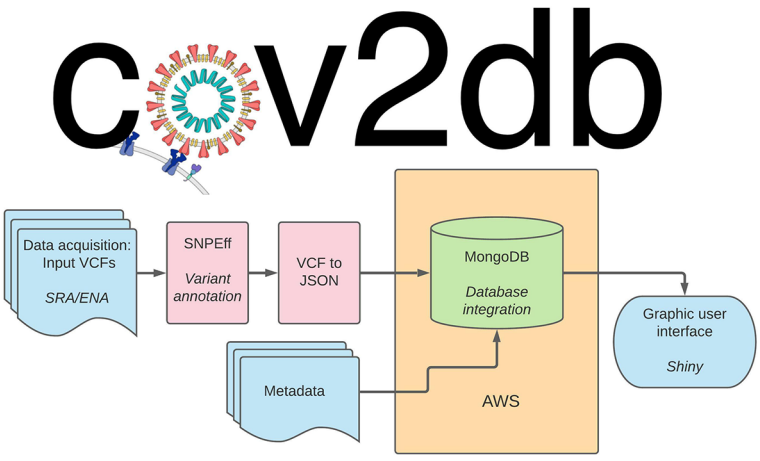

Figure 5. Cov2db workflow architecture. User provided variant call format (VCF) (or iVar output) files are annotated and ultimately converted into JavaScript object notation (JSON). The resulting JSON files serve as the primary input into the database. Secondary input can be provided by supplying any relevant metadata with the sample accession numbers serving as key. The resulting database can be queried directly via mongoDB command-line interface (CLI) or summarized and presented visually via the corresponding R Shiny app. AWS: Amazon web services.

Our current design supports input VCFs generated by LoFreq[117] (RRID:SCR_013054) or converted into VCFs from the iVar[118] output via provided script. These files are subsequently annotated with snpEff[119] (RRID:SCR_005191) using the SARS-CoV-2 reference, and resulting information is recorded as an annotated VCF. Finally, we provide an additional script to convert the annotated VCFs into JSON files that can be directly integrated into the mongoDB database. Metadata intake for the database is separate, and linking between the metadata for the samples and the variant call data is done within the database via the accession number keys.

K-var

As a proof of concept for K-var,[147] we used whole exome sequencing of the NCI-60 dataset, a panel of 60 different human tumor cell lines widely used for the screening of compounds to detect potential anticancer activity (Figure 6). *k*-mer frequencies were obtained for each sample, using the tool Jellyfish version 2.3.0. First, counts of *k*-mers of size 31 were obtained with jellyfish count. Using a custom script, *k*-mers sequence and counts were tabulated to facilitate downstream analyses. The frequency distribution was plotted using R v3.6.3[120] (RRID:SCR_000432), and low frequency *k*-mers likely arising from sequencing errors were removed. We measured the relevance of *k*-mers to the condition using TF-IDF (term frequency-inverse document frequency) with pre-defined control and test datasets. *k*-mers significantly correlated to the disease are extracted using logistic regression followed by ranking and/or classification of the significant *k*-mers. The genomic positions of the disease associated *k*-mers were identified and these positions were run through the ensembl-VEP pipeline to detect probable biological consequences.

Imavirus

There's an abundance of public high-throughput sequencing data (e.g. via the National Center for Biotechnology Information Sequence Read Archive). Some integrating viruses can produce run-on transcripts or may participate in trans-splicing between virus exon and downstream host exons.[121] Others have shown that it is possible to identify integration events by identifying chimeric reads in single-end short-read and paired-end short-read sequencing, as well as long read sequencing.[65,67–75] Others have not yet interrogated available large public datasets with current iterations of mapping.[148]

We sought to do so by scoping out the available data and exploring at least one control dataset. We then generated a non-exhaustive list of relevant human pathogenic viruses and evaluated tools for unbiased interrogation of paired-end short-read data. Minimap2 version 2.22,[32] HISAT2 version 2.2.1[122] (RRID:SCR_015530), and STAR version 2.7.9a[123] (RRID:SCR_004463) were evaluated on paired-end short-read RNA-seq from the Tg26 mouse model with HIV believed

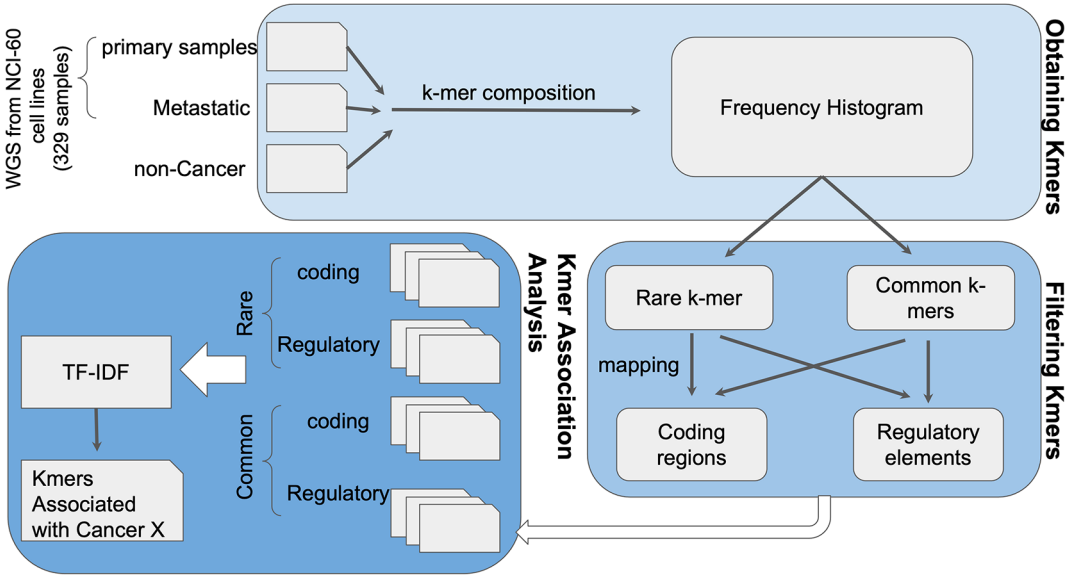

Figure 6. K-var workflow. The *k*-mer composition of whole-genome sequencing (WGS) sequencing data from cases and controls is obtained using Jellyfish. Rare and common *k*-mers are identified based on their frequency across samples, and mapped to a reference genome to assess their putative functional impact. Selected *k*-mers are then compared between cases and controls using term frequency-inverse document frequency (TF-IDF) statistical modeling to evaluate association with the phenotype of interest. As a proof of concept, K-var was implemented using cancer samples from the NCI-60 dataset.

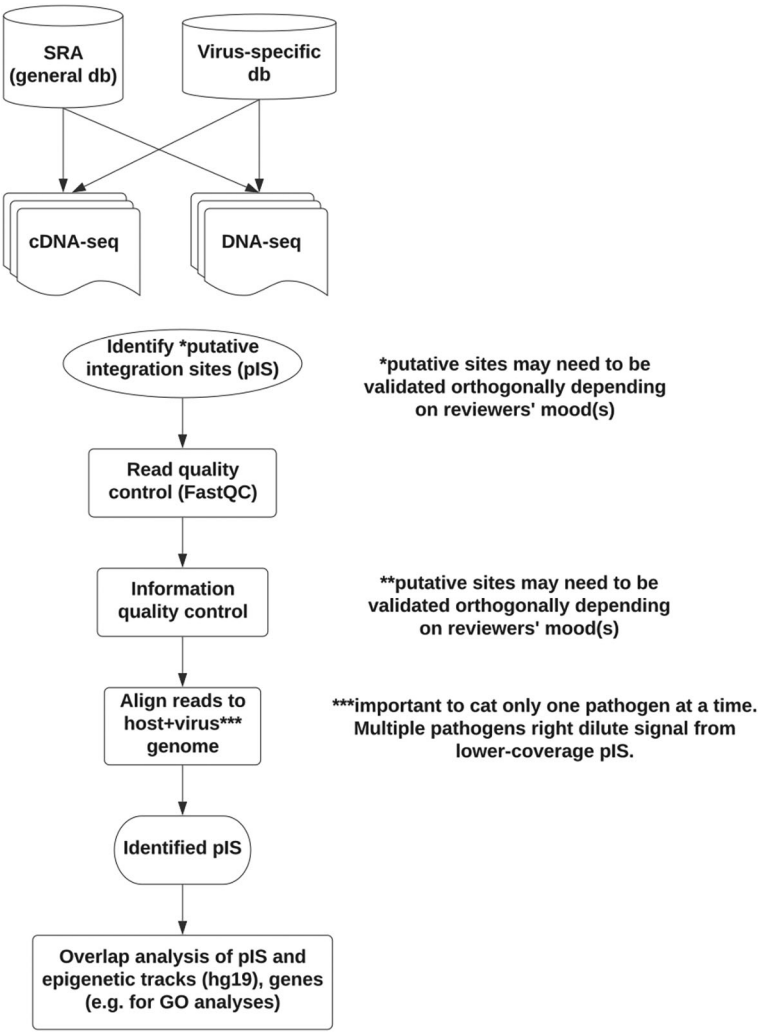

Figure 7. Imavirus workflow. To scope out the samples relevant for viral integration studies, human viruses known to integrate were chosen, along with viruses believed not to integrate (negative control set). Not shown, a dataset to contain human immunodeficiency virus (HIV) sequence (Tg26) and to express HIV protein was used as a positive control for pipeline development. Sequence Read Archive (SRA) was evaluated for the presence of RNA-seq (expression) and DNA-seq (host genomic DNA) from relevant viruses. A generic pipeline was evaluated on the positive control dataset with the goal of processing viral samples in SRA. Future work would also evaluate identified insertion/integration sites for possible clinical relevance. (GO: Gene Ontology).

to be inserted as a transgene. Minimap2 did not work for visual exploration by default, possibly because it treats paired-end reads as single-end. Mapped reads were viewed in IGV colored by orientation and with "view as pairs" selected. HISAT2 and STAR, both split-read mappers, worked to identify at least one previously identified insertion site on mouse chr8.[124] Finally, we refined this approach using human plus individual virus genomes (Figure 7).

The mouse model used includes two "insertion sites" on chr8, one on chr18, two on chrX, and a camouflaged one on chr4 embedded in a LINE element (the last site validated by long-read sequencing and deep paired-end 150 genomic DNA sequencing). These sites segregated together when multiple animals were genotyped and sequenced.[125,126] This behavior is suggestive of a yet to be defined complex structural variation encompassing multiple HIV transgene "copies" together with parts of different mouse chromosomes. The Tg26 HIV-1 transgenic mouse model[61] illustrates the current limitations of using short-read sequencing, which may only capture virus:host junctions (insertion/integration half-sites) in the absence of recapitulating the entire insertion site unambiguously. When deriving putative viral integration sites from RNA-seq, sites may be more likely to be detected if coming from highly expressed loci.

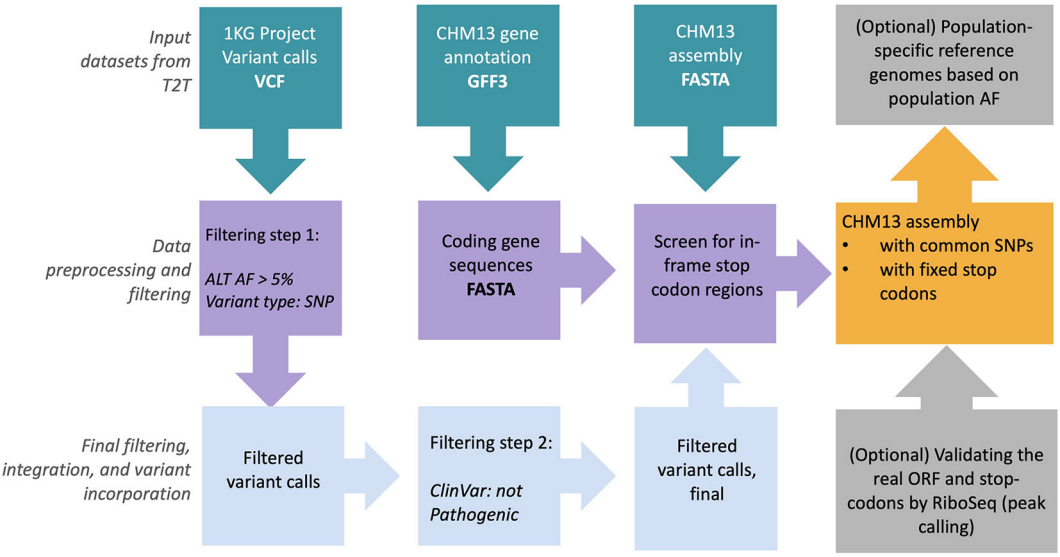

Figure 8. Overview of the reference panel generator pipeline for revising CHM13 reference panel. CHM13 genome sequence (FASTA), gene annotations (GFF3), and combined 1000 Genomes Project single nucleotide variants (SNVs) and insertion/deletion (INDEL) call sets in variant call format (VCF) are retrieved from Amazon-AWS[127] (RRID:SCR_012854) cloud. Only common alleles (>5% allele frequency (AF)) in the variant call set are retained. ClinVar[128] database was used to annotate variant calls with any clinical significance. Subsequently, common allele calls are replaced with CHM13 rare alleles in CHM13 FASTA genome sequence. Finally, screen-out in-frame stop-codon sites from genome sequence in order to generate the final reference panel files in FASTA format.

RPG

RPG[149] is a scalable and easy to apply pipeline that utilizes input genome assembly (FASTA format) and gene annotations (GFF3 format), and outputs reference panels based on the 1000 Genomes Project (1KGP) common allele calls and those abnormally avoided stop codons. Currently, the RPG pipeline is tested on the T2T-CHM13 genomic data set provided by T2T consortium in an effort to provide high-quality reference panels for diverse sequencing data analysis (Figure 8). The generation of this panel is described in Figure 8 and the accompanying figure legend.

The resultant output T2T genome features completeness (i.e. filled gaps in its genomic sequence) compared to previously available GRCh38 releases. It further harbors 1KGP common alleles and avoids stop codons. Such T2T genomic sequence can be utilized in the 'read mapping' and 'variant calling' steps while processing whole genome sequencing (WGS) data and has important applications in improving structural variant identification. The output files generated by RPG pipeline are available in GitHub repository (Software availability section) along with supplementary pre-processing scripts.

Use cases

Please refer to the Methods section for implementation details of the software including its input/output options and dependencies.

STRdust

Identification and characterization of STR using short-read sequencing data have been met with shortcomings including biases introduced by polymerase chain reaction (PCR) amplification. Long-read sequencing can identify STRs more accurately than short-read sequencing as reads can span across the entire repeat region, however, they still exhibit a high error rate. Although tools have been developed to address this problem, they still have limitations such as not being able to consider multiple STRs in a single read. To address this, our tool STRdust is capable of detecting and genotyping STRs in long-read sequencing data in both mammals and plants without prior genome annotation. As a proof of concept we simulated STRs expansions using the human and tomato reference genomes and current annotations and applied STRdust, which only requires a long read sequence alignment (see Methods). This tool can be used by plant breeders to accurately genotype STRs and develop linkage maps, which are essential for mapping quantitative trait loci and intelligently selecting for combinations of traits of interest in the offspring.

kTom

In plant breeding, important traits are often moved into elite breeding material through traditional plant breeding methods of crossing and back-crossing with phenotypic selection to retain the trait of interest. In the era of genomics, genotype markers can be used to track the introgression of traits into different lines. However, for traits with complex underlying genome biology, including structural variations, SNP-based markers are often insufficient to discover or track traits reliably in a breeding pipeline. This is particularly relevant for identifying and tracking disease resistance loci, which have been introgressed from wild tomato relatives into elite tomatoes over decades of breeding[25] and higher-resolution tracking of those loci could accelerate tomato breeding. To circumvent the SNP-based limitations for finding and following trait introgressions, the kTom tool uses a *k*-mer approach to characterize re-sequenced genomes and identify potential *k*-mer tags for trait introgressions. The kTom tool enables the user to understand the *k*-mer profile of the resequenced genome (from Illumina WGS reads) and compare that to a background *k*-mer profile (e.g. the reference genome for that line or a known genome without the trait of interest) to identify novel *k*-mers. kTom can enable population-level analysis of structural variation, including establishing an alternate (non-SNP) genotyping method to profile introgressions within a population and investigating and visualizing the history of introgressions. A derivation of kTom data can facilitate understanding tomato population structure with a data type more able to account for SVs. In addition, the output of kTom should be able to form the basis for a *k*-mer GWAS approach.[28] The kTom tool was designed with plant breeding problems in mind, but it can be applied to any resequencing dataset without the need for a reference genome.

INSeption

Insertions play an important role in human genetic variability and diseases, and therefore their accurate identification is key for genetic analyses and clinical studies. However, comprehensively identifying sequence-resolved insertions can be challenging, especially when the read length is not sufficient to span the whole inserted sequence. In those cases, SV callers will identify the insertion's location but not its sequence. INSeption is a bioinformatics workflow that addresses this issue by reconstructing the inserted sequence utilizing the unaligned portions of reads (i.e. hanging reads). After retrieving a sample's unaligned reads, INSeption builds a consensus sequence to provide sequence-resolved insertions. This information allows scientists to better assess the impact of an insertion on gene function and genome organization.

GeneVar2

SVs account for more genetic differences between humans than other types of variation and are the underlying genetic cause of several traits and diseases.[33,129] Although SV discovery has become more readily available, its interpretation is particularly challenging for those outside the immediate field of genetics.[40,41] GeneVar2 is an extremely fast and computationally efficient platform for the analysis, visualization, and interpretation of structural variation data. It is designed to provide a powerful and easy-to-use tool for applications in biomedical research and diagnostic medicine at minimal computational cost. Its comprehensive approach brings the analyses of structural variation within the reach of non-specialist laboratories and to centers with limited computational resources available.

cov2db

Cataloging viral mutations within a sample (intra-host variation) and across samples (inter-host variation) provides critical insights to understanding the dynamics of viral evolution during the COVID-19 pandemic.[130] The SARS-CoV-2 virus has been shown to have high genomic diversity[45,131]), and mutations can change the fitness of the virus[132] by increasing its transmission or pathogenicity potential.[133,134] SNVs can also result in dramatically different protein function and recognition,[135,136] and studies have shown persistent intra-host evolution of SARS-CoV-2 in immuno-compromised hosts.[137] cov2db represents an integrative platform and complementary database for active monitoring of SARS-CoV-2 strain variants specific to circulating SARS-CoV-2 lineages and will facilitate efficient and sensitive tracking of both inter-host and intra-host SARS-CoV-2 variation.

Input to cov2db consists of a single or multiple VCF file(s) in the format output by the LoFreq variant caller. Cov2db does not provide an output, but allows its users to interact with a mongoDB database instance containing the variant calling information provided by the users.

K-var

The identification of phenotype-associated biomarkers is crucial for precision medicine, crop breeding, and answering evolutionary questions. Based on the area of interest, K-var can be applied to identify mutational signatures that help distinguish between conditions using phenotype associations with low bias. Short-read sequencing data is used as input to estimate *k*-mer frequencies per sample, followed by statistical correlation to a known phenotype across two distinct conditions. The output is a ranked table of significant phenotype-associated *k*-mers that can be used to fish for genomic regions experiencing mutations. Precisely identifying these genomic locations will help in downstream analysis to infer

biological consequences. During the hackathon, we ran K-var using as input metastatic and non-metastatic breast cancer whole-exome sequencing (non-metastatic $n = 7$; metastatic $n = 5$) from the NCI-60 dataset. K-var delivered a ranked list of 44,884 *k*-mers, where the score indicates the relevance of each sequence (calculated by TF-IDF analysis) in differentiating metastatic and non-metastatic (primary) sequences. The top-ranked *k*-mer identified by K-var impacted methyl-methanesulfonate sensitivity 19 (*MMS19*), a component of the Fe-S assembly machinery involved in the production of proteins associated with genomic stability, such as DNA polymerase and DNA repair proteins. This gene has been reported as a breast cancer candidate gene in familial studies in Tunisian individuals.[138] The next highest-ranked *k*-mer impacted 1-Acylglycerol-3-Phosphate O-Acyltransferase 4 (*AGPAT4*), which has been proposed as required for triple-negative breast cancer progression.[139] This ranked *k*-mer list and genes impacted provide a resource to "fish" for genes relevant to breast cancer metastasis.

Imavirus

When deriving putative viral integration sites from RNA-seq, sites may be more likely to be detected if coming from highly expressed loci. The Tg26 mouse reanalyzed in the present study was made from pronuclear injection of pNL4–3 restriction products.[61] Such insertion depends entirely on host DNA repair machinery on nuclear DNA with nuclear topology distinct from human cells which HIV more easily infects. As such, many of the sites missed during the RNA-seq interrogation may have been missed due to low levels of expression at those loci, or those parts of the complex insertion site. Many viruses have the capability of integrating into host genomes, leading to DNA damage and gene disruption. Accurately identifying virus integration sites and potentially disrupted genes is important to fully understand their impact on disease severity. However, identifying virus integration sites from genomic DNA is challenging and there are not many bioinformatics tools available to reliably detect viral presence or integration events. Here, we developed Imavirus, a bioinformatics approach that identifies putative virus integration sites (pIS) in public data. Using unbiased RNA-seq datasets, Imavirus aimed to identify pIS and to pinpoint clinically relevant viral integration sites, especially those that may affect cell function and possibly contributing to disease and/or antiviral responses and possibly contributing to virus fitness. Imavirus is a community resource that aids researchers by enhancing our knowledge of viral infection and improving disease severity prediction of viral infections. During the hackathon we were able to verify a previously reported integration site on mouse chr8 which can be seen in our GitHub repository cited in the Software Availability section.

Future work should explore the datasets we scoped out in SRA for more physiological systems such as animal models or stable cell lines to identify more putative insertion sites. Another important limitation of the positive control set explored here is that the Tg26 mouse has approximately 15 copies of HIV integrated at the same loci.[125,126] While HIV signal may be coming from multiple loci, when considering junctions, most of the signal seemed to be coming from an HIV copy with run-on transcripts in chr8. Natural human infections would have distinct insertion sites, making them harder to spot with these approaches.

RPG

The human reference genome has served as a foundation for human genetics and genomics studies. Despite its countless applications, the current human reference genome assembly (GRCh38) harbors several gaps and missing nucleotides ('N' characters) that hinder comprehensive analysis. Therefore, complete T2T genome references are essential to make sure all the genomic variants are discovered and analyzed. Here, we implemented a pipeline that incorporates 1000 Genome Project common alleles and avoids stop codons into the T2T-CHM13 genome sequence, in order to provide a complete human reference sequence for diverse sequencing data analyses.

Conclusions

The results of the 2021 Baylor College of Medicine/DNAnexus hackathon described here represent novel work that pushes the field forward for human, plant, and viral genome SV detection. All are needed to further current findings about diversity and the complexity of organisms and their genotypes. To further facilitate this progress in a FAIR-compliant manner, 59 people, across the world with different professional backgrounds, came together in October 2021 to complete or further eight groundbreaking prototypes.

Next steps

Some directions that we think will be impactful in the future are:

k-mer analyses avoid reference and mapping biases through a reference-free approach. We endeavor to create tools for *k*-mer analysis which work with both short- and long- read sequencing technologies. An example use case is the identification of rare variants can help in diagnostic purposes to identify disease biomarkers and therapeutic targets for personalized medicine. Along similar lines, crop breeding research can benefit from identifying markers associated with disease resistance.

Table 1. Lists the data source utilized by each tool developed during the hackathon.

Tool name	Data source utilized
STRdust	GRCh38 human reference genome: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.fna.gz SL4.0 tomato genome: https://solgenomics.net/ftp//tomato_genome/assembly/build_4.00/S_lycopersicum_chromosomes.4.00.fa.gz
kTom	100 Tomato Consortium: whole genome data of 84 tomatoes, BioProject PRJEB5235 - https://www.ncbi.nlm.nih.gov/bioproject/236988
INseption	GIAB HiFi data set (fastq files): https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb/
GeneVar2	Gene association with phenotype disorders in OMIM: https://maayanlab.cloud/static/hdfs/harmonizome/data/omim/gene_list_terms.txt.gz R clusterProfiler annotation for Disease Ontology, DisGeNET, Network of Cancer Gene, Gene Ontology, KEGG pathway: https://guangchuangyu.github.io/software/clusterProfiler/ dbVar, known clinical SV annotation, GRCh38: http://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_study/tsv/nstd102.GRCh38.variant_call.tsv.gz GENCODE v35 gene annotation: http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_35/gencode.v35.annotation.gff3.gz gnomAD pLI information: https://azureopendatastorage.blob.core.windows.net/gnomad/release/2.1.1/constraint/gnomad.v2.1.1.lof_metrics.by_gene.txt.bgz gnomAD-SV BED file with allele frequencies: https://datasetgnomad.blob.core.windows.net/dataset/papers/2019-sv/gnomad_v2.1_sv.sites.bed.gz
K-var	Whole exome sequencing data of NCI-60 dataset, BioProject PRJNA523380. Breast cancer accession numbers: SRR8619035, SRR8619036, SRR8619038, SRR8619044, SRR8619110, SRR8619113, SRR8619154, SRR8619076, SRR8619133, SRR8619134, SRR8618981, SRR8619186 GRCh38 human reference genome: http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/GRCh38.primary_assembly.genome.fa.gz GRCh38 Gencode gene annotations: http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.primary_assembly.annotation.gff3.gz
Imavirus	SRA RNA-seq test data: SRR10302267. The accession for the pNL4–3 used to make the Tg26 mouse is GenBank:AF324493.2, and this was used to explore data in IGV. The mm10 mouse genome was used to visualize cognate integration site(s) on mouse chr8. Accession lists in GitHub repository listed in the software availability section.
RPG	Complete genome of CHM13 T2T v2.0, BioProject PRJNA559484: https://www.ncbi.nlm.nih.gov/assembly/GCA_009914755.4

Specifically, we would improve the kTom tool to enable quantification of *k*-mers to detect potential copy-number changes, and this would be particularly relevant for disease resistance loci, which often contain gene copy-number variations.[140]

To achieve these outcomes, additional modules can be added to the kTom code base. For example, for disease-resistance introgression that is known to be in some but not all of samples in a collection, the *k*-mer frequency matrix can be filtered to keep low-middle frequency *k*-mers. With or without this filtering step, a distance matrix can be computed from the *k*-mer frequency matrix and used for hierarchical clustering to suggest sets *k*-mers introgressed together. The resulting output can then be used for validation complemented by curated and known loci, and later for phenotypic association.

For clearer virus integration site mapping, long read DNA sequencing is preferable to short reads, but these types of data are sparse in major public repositories. The present hackathon scoped out sequences from the Sequence Read Archive and subset viruses and controls relevant for integration studies. Therefore, future work is needed to compare short-read datasets to long-read generated stable vs transient insertion sites in order to improve our understanding of the effects on viral replication, host gene regulation, and disease.

To inform clinical significance of SVs for clinicians and researchers, GeneVar2 is a comprehensive tool for understanding the impact of SVs on disease. To expand users' ability to identify and communicate key SV findings, Samplot,[141] a multi-sample structural variant visualizer, will be integrated with GeneVar2. Subsequent development will focus on further cloud integration and new output options, such as research reports and the ability to use the application off-line.

The annual nature of this hackathon has seeded teams and projects that are often ongoing for multiple years, resulting in mature software products. Other annual hackathons, particularly the NBDC/DBCLS (Japan) and ELIXIR (Europe) bio-hackathons, have seen the same. In this vein, we expect to see many of the projects that have been seeded here continue next year, and possibly in other hackathons.

Data availability

Underlying data

The data used for these projects were obtained from publicly accessible repositories and are available in Table 1.

Software availability

STRdust

Source code available from: https://github.com/collaborativebioinformatics/STRdust

Release version: 0.2.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.6467829.[142]

License: MIT.

kTom

Source code available from: https://github.com/collaborativebioinformatics/kTom

Release version: 0.2.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.6467823.[143]

License: MIT.

INSeption

Source code available from: https://github.com/collaborativebioinformatics/InSeption

Release version: 0.2.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.6467818.[144]

License: MIT.

GeneVar2

Source code available from: https://github.com/collaborativebioinformatics/GeneVar2

Release version: 0.2.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.6467837.[145]

License: MIT.

Cov2db

Source code available from: https://github.com/collaborativebioinformatics/cov2db

Release version: 0.2.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.6467825.[146]

License: MIT.

K-var

Source code available from: https://github.com/collaborativebioinformatics/kvar

Release version: 0.2.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.6467850.[147]

License: MIT.

Imavirus

Source code available from: https://github.com/collaborativebioinformatics/imavirus

Release version: 0.2.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.6467774.[148]

License: MIT.

RPG

Source code available from: https://github.com/collaborativebioinformatics/RPG_Pikachu

Release version: 0.2.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.6467816.[149]

License: MIT.

Acknowledgements

The hackathon was sponsored by Pacific Biosciences of California, Inc. and Oxford Nanopore Technologies Limited. We'd like to thank Richard Gibbs and the Baylor College of Medicine Human Genome Sequencing Center for hosting the hackathon. Computation was performed on the DNAnexus platform, which donated compute and storage to the hackathon.

The U.S. Department of Agriculture is an equal opportunity lender, provider, and employer.

Mention of trade names or commercial products in this report is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

References

1. Ho SS, Urban AE, Mills RE: **Structural variation in the sequencing era.** *Nat. Rev. Genet.* 2020 Mar; **21**(3): 171–189.
PubMed Abstract | Publisher Full Text

2. Mahmoud M, Gobet N, Cruz-Dávalos DI, *et al.*: **Structural variant calling: the long and the short of it.** *Genome Biol.* 2019 Nov 20; **20**(1): 246.
PubMed Abstract | Publisher Full Text

3. Sanchis-Juan A, Stephens J, French CE, *et al.*: **Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing.** *Genome Med.* 2018 Dec 7; **10**(1): 95.
PubMed Abstract | Publisher Full Text

4. Hurles ME, Dermitzakis ET, Tyler-Smith C: **The functional impact of structural variation in humans.** *Trends Genet.* 2008 May; **24**(5): 238–245.
PubMed Abstract | Publisher Full Text

5. Serba DD, Daverdin G, Bouton JH, *et al.*: **Quantitative trait loci (QTL) underlying biomass yield and plant height in switchgrass.** *Bioenerg. Res.* 2015 Mar; **8**(1): 307–324.
Publisher Full Text

6. Hartl DL: *A primer of population genetics and genomics.* Oxford University Press; 2020.
Publisher Full Text

7. Ge J, Guo N, Meng J, *et al.*: **K-mer Counting for Genomic Big Data**. Chin FYL, Chen CLP, Khan L, *et al.*, editors. *Big data – bigdata 2018.* Cham: Springer International Publishing; 2018. p. 345–51.

8. Tajima F: **Determination of window size for analyzing DNA sequences.** *J. Mol. Evol.* 1991 Nov; **33**(5): 470–473.
PubMed Abstract | Publisher Full Text

9. Wellenreuther M, Mérot C, Berdan E, *et al.*: **Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification.** *Mol. Ecol.* 2019 Mar; **28**(6): 1203–1209.
PubMed Abstract | Publisher Full Text

10. Ebler J, Clarke WE, Rausch T, *et al.*: **Pangenome-based genome inference.** *BioRxiv.* 2020 Nov 12.

11. Mc Cartney AM, Mahmoud M, Jochum M, *et al.*: **An international virtual hackathon to build tools for the analysis of structural variants within species ranging from coronaviruses to vertebrates.** *F1000Res.* 2021 Mar 26; **10**: 246.
Publisher Full Text

12. Zielezinski A, Vinga S, Almeida J, *et al.*: **Alignment-free sequence comparison: benefits, applications, and tools.** *Genome Biol.* 2017 Oct 3; **18**(1): 186.
Publisher Full Text

13. Fan H, Chu J-Y: **A brief review of short tandem repeat mutation.** *Genomics Proteomics Bioinformatics.* 2007 Feb; **5**(1): 7–14.
Publisher Full Text

14. Pearson CE, Nichol Edamura K, Cleary JD: **Repeat instability: mechanisms of dynamic mutations.** *Nat. Rev. Genet.* 2005 Oct; **6**(10): 729–742.
Publisher Full Text

15. Mirkin SM: **Expandable DNA repeats and human disease.** *Nature.* 2007 Jun 21; **447**(7147): 932–940.
Publisher Full Text

16. Zhu L, Wu H, Li H, *et al.*: **Short Tandem Repeats in plants: Genomic distribution and function prediction.** *Electron. J. Biotechnol.* 2021 Mar; **50**: 37–44.
Publisher Full Text

17. Bentley DR, Balasubramanian S, Swerdlow HP, *et al.*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature.* 2008 Nov 6; **456**(7218): 53–59.
PubMed Abstract | Publisher Full Text

18. Wenger AM, Peluso P, Rowell WJ, *et al.*: **Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome.** *Nat. Biotechnol.* 2019 Oct; **37**(10): 1155–62.
PubMed Abstract | Publisher Full Text

19. Shafin K, Pesout T, Lorig-Roach R, *et al.*: **Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes.** *Nat. Biotechnol.* 2020 Sep; **38**(9): 1044–1053.
PubMed Abstract | Publisher Full Text

20. Dohm JC, Peters P, Stralis-Pavese N, *et al.*: **Benchmarking of long-read correction methods.** *NAR Genom Bioinform.* 2020 Jun; **2**(2): lqaa037.
PubMed Abstract | Publisher Full Text

21. Ummat A, Bashir A: **Resolving complex tandem repeats with long reads.** *Bioinformatics.* 2014 Dec 15; **30**(24): 3491–3498.
PubMed Abstract | Publisher Full Text

22. De Roeck A, De Coster W, Bossaerts L, *et al.*: **NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION.** *Genome Biol.* 2019 Nov 14; **20**(1): 239.
PubMed Abstract | Publisher Full Text

23. Bolognini D, Magi A, Benes V, *et al.*: **TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data.** *Gigascience.* 2020 Oct 7; **9**(10).
PubMed Abstract | Publisher Full Text

24. Chiu R, Rajan-Babu I-S, Friedman JM, *et al.*: **Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences.** *Genome Biol.* 2021 Aug 13; **22**(1): 224.
PubMed Abstract | Publisher Full Text

25. Lin T, Zhu G, Zhang J, *et al.*: **Genomic analyses provide insights into the history of tomato breeding.** *Nat. Genet.* 2014 Nov; **46**(11): 1220–6.
PubMed Abstract | Publisher Full Text

26. FAOSTAT: **Choice Reviews Online.** 2011 Jan 1; **48**(05): 48–2430–48–2430.

27. Schouten HJ, Tikunov Y, Verkerke W, *et al.*: **Breeding has increased the diversity of cultivated tomato in the netherlands.** *Front. Plant Sci.* 2019 Dec 10; **10**: 1606.
PubMed Abstract | Publisher Full Text

28. Voichek Y, Weigel D: **Identifying genetic variants underlying phenotypic variation in plants without complete genomes.** *Nat. Genet.* 2020 May; **52**(5): 534–540.
PubMed Abstract | Publisher Full Text

29. Hancks DC, Kazazian HH: **Roles for retrotransposon insertions in human disease.** *Mob. DNA.* 2016 May 6; **7**: 9.
PubMed Abstract | Publisher Full Text

30. Cajuso T, Sulo P, Tanskanen T, *et al.*: **Retrotransposon insertions can initiate colorectal cancer and are associated with poor survival.** *Nat. Commun.* 2019 Sep 6; **10**(1): 4022.
PubMed Abstract | Publisher Full Text

31. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009 Jul 15; **25**(14): 1754–1760.
PubMed Abstract | Publisher Full Text

32. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018 Sep 15; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text

33. Sedlazeck FJ, Rescheneder P, Smolka M, *et al.*: **Accurate detection of complex structural variations using single-molecule sequencing.** *Nat. Methods.* 2018 Jun; **15**(6): 461–468.
PubMed Abstract | Publisher Full Text

34. Rausch T, Zichner T, Schlattl A, *et al.*: **DELLY: structural variant discovery by integrated paired-end and split-read analysis.** *Bioinformatics.* 2012 Sep 15; **28**(18): i333–i339.
PubMed Abstract | Publisher Full Text

35. Tian S, Yan H, Klee EW, *et al.*: **Comparative analysis of de novo assemblers for variation discovery in personal genomes.** *Brief Bioinformatics.* 2018 Sep 28; **19**(5): 893–904.
PubMed Abstract | Publisher Full Text

36. Koren S, Walenz BP, Berlin K, *et al.*: **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.** *Genome Res.* 2017 May; **27**(5): 722–736.
PubMed Abstract | Publisher Full Text

37. Kolmogorov M, Yuan J, Lin Y, *et al.*: **Assembly of long, error-prone reads using repeat graphs.** *Nat. Biotechnol.* 2019 May; **37**(5): 540–546.
PubMed Abstract | Publisher Full Text

38. Chaisson MJ, Tesler G: **Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory.** *BMC Bioinformatics.* 2012 Sep 19; **13**: 238.
PubMed Abstract | Publisher Full Text

39. Ebert P, Audano PA, Zhu Q, *et al.*: **Haplotype-resolved diverse human genomes and integrated analysis of structural variation.** *Science.* 2021 Apr 2; **372**(6537).
PubMed Abstract | Publisher Full Text

40. Iacoangeli A, Al Khleifat A, Sproviero W, *et al.*: **ALSgeneScanner: a pipeline for the analysis and interpretation of DNA sequencing data of ALS patients.** *Amyotroph Lateral Scler Frontotemporal Degener.* 2019 May; **20**(3–4): 207–215.
PubMed Abstract | Publisher Full Text

41. Iacoangeli A, Al Khleifat A, Sproviero W, *et al.*: **DNAscan: personal computer compatible NGS analysis, annotation and visualisation.** *BMC Bioinformatics.* 2019 Apr 27; **20**(1): 213.
PubMed Abstract | Publisher Full Text

42. Sayers EW, Cavanaugh M, Clark K, *et al.*: **GenBank.** *Nucleic Acids Res.* 2020 Jan 8; **48**(D1): D84–D86.
PubMed Abstract | Publisher Full Text

43. Lythgoe KA, Hall M, Ferretti L, *et al.*: **SARS-CoV-2 within-host diversity and transmission.** *Science.* 2021 Apr 16; **372**(6539).
PubMed Abstract | Publisher Full Text

44. Wang Y, Wang D, Zhang L, *et al.*: **Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients.** *Genome Med.* 2021 Feb 22; **13**(1): 30.
PubMed Abstract | Publisher Full Text

45. Sapoval N, Mahmoud M, Jochum MD, *et al.*: **SARS-CoV-2 genomic diversity and the implications for qRT-PCR diagnostics and transmission.** *Genome Res.* 2021 Apr; **31**(4): 635–644.
PubMed Abstract | Publisher Full Text

46. Armero A, Berthet N, Avarre J-C: **Intra-Host Diversity of SARS-Cov-2 Should Not Be Neglected: Case of the State of Victoria, Australia.** *Viruses.* 2021 Jan 19; **13**(1).
PubMed Abstract | Publisher Full Text

47. Leinonen R, Sugawara H, Shumway M: **International Nucleotide Sequence Database Collaboration. The sequence read archive.** *Nucleic Acids Res.* 2011 Jan; **39**(Database issue): D19–D21.
PubMed Abstract | Publisher Full Text

48. Leinonen R, Akhtar R, Birney E, *et al.*: **The european nucleotide archive.** *Nucleic Acids Res.* 2011 Jan; **39**(Database issue): D28–D31.
PubMed Abstract | Publisher Full Text

49. Compeau PEC, Pevzner PA, Tesler G: **How to apply de Bruijn graphs to genome assembly.** *Nat. Biotechnol.* 2011 Nov 8; **29**(11): 987–991.
PubMed Abstract | Publisher Full Text

50. Zhao L, Xie J, Bai L, *et al.*: **Mining statistically-solid k-mers for accurate NGS error correction.** *BMC Genomics.* 2018 Dec 31; **19**(Suppl 10): 912.
PubMed Abstract | Publisher Full Text

51. Breitwieser FP, Baker DN, Salzberg SL: **KrakenUniq: confident and fast metagenomics classification using unique k-mer counts.** *Genome Biol.* 2018 Nov 16; **19**(1): 198.
PubMed Abstract | Publisher Full Text

52. Rahman A, Hallgrímsdóttir I, Eisen M, *et al.*: **Association mapping from sequencing reads using k-mers.** *elife.* 2018 Jun 13; **7**.
PubMed Abstract | Publisher Full Text

53. Mehrab Z, Mobin J, Tahmid IA, *et al.*: **Reference-free Association Mapping from Sequencing Reads Using k-mers.** *Bio Protoc.* 2020 Nov 5; **10**(21): e3815.
PubMed Abstract | Publisher Full Text

54. Wang Y, Chen Q, Deng C, *et al.*: **KmerGO: A Tool to Identify Group-Specific Sequences With k-mers.** *Front. Microbiol.* 2020 Aug 25; **11**: 2067.
PubMed Abstract | Publisher Full Text

55. Woolhouse M, Scott F, Hudson Z, *et al.*: **Human viruses: discovery and emergence.** *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 2012 Oct

19; **367**(1604): 2864–2871.
PubMed Abstract | Publisher Full Text

56. Nkili-Meyong AA, Moussavou-Boundzanga P, Labouba I, *et al.*: **Genome-wide profiling of human papillomavirus DNA integration in liquid-based cytology specimens from a Gabonese female population using HPV capture technology.** *Sci. Rep.* 2019 Feb 6; **9**(1): 1504.
PubMed Abstract | Publisher Full Text

57. Tu T, Budzinska MA, Vondran FWR, *et al.*: **Hepatitis B Virus DNA Integration Occurs Early in the Viral Life Cycle in an in vitro Infection Model via Sodium Taurocholate Cotransporting Polypeptide-Dependent Uptake of Enveloped Virus Particles.** *J. Virol.* 2018 Jun 1; **92**(11).
PubMed Abstract | Publisher Full Text

58. Bushman FD: **Retroviral insertional mutagenesis in humans: evidence for four genetic mechanisms promoting expansion of cell clones.** *Mol. Ther.* 2020 Feb 5; **28**(2): 352–356.
PubMed Abstract | Publisher Full Text

59. Marchand C, Johnson AA, Semenova E, *et al.*: **Mechanisms and inhibition of HIV integration.** *Drug Discov. Today Dis. Mech.* 2006 Jul 1; **3**(2): 253–260.
PubMed Abstract | Publisher Full Text

60. Huang R, Zhou P-K: **DNA damage repair: historical perspectives, mechanistic pathways and clinical translation for targeted cancer therapy.** *Signal Transduct. Target. Ther.* 2021 Jul 9; **6**(1): 254.
PubMed Abstract | Publisher Full Text

61. Dickie P, Felser J, Eckhaus M, *et al.*: **HIV-associated nephropathy in transgenic mice expressing HIV-1 genes.** *Virology.* 1991 Nov; **185**(1): 109–119.
PubMed Abstract | Publisher Full Text

62. Gener AR, Zou W, Foley BT, *et al.*: **Reference Plasmid pHXB2_D is an HIV-1 Molecular Clone that Exhibits Identical LTRs and a Single Integration Site Indicative of an HIV Provirus.** *Res Sq.* 2021 Apr 6.

63. Shendure J, Balasubramanian S, Church GM, *et al.*: **DNA sequencing at 40: past, present and future.** *Nature.* 2017 Oct 19; **550**(7676): 345–353.
PubMed Abstract | Publisher Full Text

64. Stark R, Grzelak M, Hadfield J: **RNA sequencing: the teenage years.** *Nat. Rev. Genet.* 2019 Nov; **20**(11): 631–656.
Publisher Full Text

65. Cameron DL, Jacobs N, Roepman P, *et al.*: **Virusbreakend: viral integration recognition using single breakends.** *Bioinformatics.* 2021 May 11; **37**: 3115–3119.
PubMed Abstract | Publisher Full Text

66. Desfarges S, Ciuffi A: **Viral integration and consequences on host gene expression**. Witzany G, editor. *Viruses: essential agents of life.* Dordrecht: Springer Netherlands; 2012. p. 147–75.

67. Artesi M, Hahaut V, Cole B, *et al.*: **PCIP-seq: simultaneous sequencing of integrated viral genomes and their insertion sites with long reads.** *Genome Biol.* 2021 Apr 6; **22**(1): 97.
PubMed Abstract | Publisher Full Text

68. Zhuo Z, Rong W, Li H, *et al.*: **Long-read sequencing reveals the structural complexity of genomic integration of HBV DNA in hepatocellular carcinoma.** *NPJ Genom. Med.* 2021 Oct 12; **6**(1): 84.
PubMed Abstract | Publisher Full Text

69. Stephens Z, O'Brien D, Dehankar M, *et al.*: **Exogene: A performant workflow for detecting viral integrations from paired-end next-generation sequencing data.** *PLoS One.* 2021 Sep 22; **16**(9): e0250915.
PubMed Abstract | Publisher Full Text

70. Ramirez R, van Buuren N, Gamelin L, *et al.*: **Targeted Long-Read Sequencing Reveals Comprehensive Architecture, Burden, and Transcriptional Signatures from Hepatitis B Virus-Associated Integrations and Translocations in Hepatocellular Carcinoma Cell Lines.** *J. Virol.* 2021 Sep 9; **95**(19): e0029921.
PubMed Abstract | Publisher Full Text

71. Yang W, Liu Y, Dong R, *et al.*: **Accurate detection of HPV integration sites in cervical cancer samples using the nanopore minion sequencer without error correction.** *Front. Genet.* 2020 Jun 26; **11**: 660.
PubMed Abstract | Publisher Full Text

72. Patro SC, Brandt LD, Bale MJ, *et al.*: **Combined HIV-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors.** *Proc. Natl. Acad. Sci. USA.* 2019 Dec 17; **116**(51): 25891–25899.
PubMed Abstract | Publisher Full Text

73. Iwase SC, Miyazato P, Katsuya H, *et al.*: **HIV-1 DNA-capture-seq is a useful tool for the comprehensive characterization of HIV-1 provirus.** *Sci. Rep.* 2019 Aug 23; **9**(1): 12326.
PubMed Abstract | Publisher Full Text

74. Nguyen N-PD, Deshpande V, Luebeck J, *et al.*: **ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer.** *Nucleic Acids Res.* 2018 Apr 20; **46**(7): 3309–3325.
PubMed Abstract | Publisher Full Text

75. Xia Y, Liu Y, Deng M, *et al.*: **Detecting virus integration sites based on multiple related sequencing data by VirTect.** *BMC Med. Genet.* 2019 Jan 31; **12**(Suppl 1): 19.
PubMed Abstract | Publisher Full Text

76. Aganezov S, Yan SM, Soto DC, *et al.*: **A complete reference genome improves analysis of human genetic variation.** *BioRxiv.* 2021 Jul 13.

77. 1000 Genomes Project ConsortiumAuton A, Brooks LD, *et al.*: **A global reference for human genetic variation.** *Nature.* 2015 Oct 1; **526**(7571): 68–74.
PubMed Abstract | Publisher Full Text

78. Vaser R, Sović I, Nagarajan N, *et al.*: **Fast and accurate de novo genome assembly from long uncorrected reads.** *Genome Res.* 2017 May; **27**(5): 737–746.
PubMed Abstract | Publisher Full Text

79. Kolpakov R, Bana G, Kucherov G: **mreps: Efficient and flexible detection of tandem repeats in DNA.** *Nucleic Acids Res.* 2003 Jul 1; **31**(13): 3672–3678.
PubMed Abstract | Publisher Full Text | Free Full Text

80. 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 Jan 24; **8**: 14061.
PubMed Abstract | Publisher Full Text

81. Edge P, Bansal V: **Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing.** *Nat. Commun.* 2019 Oct 11; **10**(1): 4660.
PubMed Abstract | Publisher Full Text

82. 100 Tomato Genome Sequencing ConsortiumAflitos S, Schijlen E, *et al.*: **Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing.** *Plant J.* 2014 Oct; **80**(1): 136–148.
Publisher Full Text

83. Andrews S: **FastQC: a quality control tool for high throughput sequence data.** 2010.
Reference Source

84. Dodt M, Roehr JT, Ahmed R, *et al.*: **FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms.** *Biology (Basel).* 2012 Dec 14; **1**(3): 895–905.
PubMed Abstract | Publisher Full Text

85. Marçais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics.* 2011 Mar 15; **27**(6): 764–770.
PubMed Abstract | Publisher Full Text

86. Vurture GW, Sedlazeck FJ, Nattestad M, *et al.*: **GenomeScope: fast reference-free genome profiling from short reads.** *Bioinformatics.* 2017 Jul 15; **33**(14): 2202–2204.
PubMed Abstract | Publisher Full Text

87. Gu Z, Eils R, Schlesner M: **Complex heatmaps reveal patterns and correlations in multidimensional genomic data.** *Bioinformatics.* 2016 Sep 15; **32**(18): 2847–2849.
PubMed Abstract | Publisher Full Text

88. Gu Z, Hübschmann D: **Make interactive complex heatmaps in R.** *Bioinformatics.* 2021 Dec 2.

89. Wickham H, Averick M, Bryan J, *et al.*: **Welcome to the tidyverse.** *JOSS.* 2019 Nov 21; **4**(43): 1686.
Publisher Full Text

90. Zook JM, McDaniel J, Olson ND, *et al.*: **An open resource for accurately benchmarking small variant and reference calls.** *Nat. Biotechnol.* 2019 May; **37**(5): 561–566.
PubMed Abstract | Publisher Full Text

91. Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *Gigascience.* 2021 Feb 16; **10**(2).
PubMed Abstract | Publisher Full Text

92. Van Rossum G, Drake FL Jr: *Python reference manual.* Centrum voor Wiskunde en Informatica Amsterdam; 1995.

93. Van Rossum G, Drake FL: *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace; 2009.

94. Luo J, Chen R, Zhang X, *et al.*: **LROD: An Overlap Detection Algorithm for Long Reads Based on k-mer Distribution.** *Front. Genet.* 2020 Jul 11; **11**: 632.
PubMed Abstract | Publisher Full Text

95. **CARNAC-LR: Clustering coefficient-based Acquisition of RNA Communities in Long Reads - Archive ouverte HAL.** [cited 2022 Feb 15].
Reference Source

96. Prjibelski A, Antipov D, Meleshko D, *et al.*: **Using SPAdes de novo assembler.** *Curr. Protoc. Bioinformatics.* 2020; **70**(1): e102.
PubMed Abstract | Publisher Full Text

97. Frankish A, Diekhans M, Jungreis I, *et al.*: **GENCODE 2021.** *Nucleic Acids Res.* 2021 Jan 8; **49**(D1): D916–D923.
PubMed Abstract | Publisher Full Text

98. Tweedie S, Braschi B, Gray K, *et al.*: **Genenames.org: the HGNC and VGNC resources in 2021.** *Nucleic Acids Res.* 2021 Jan 8; **49**(D1): D939–D946.
PubMed Abstract | Publisher Full Text

99. Howe KL, Achuthan P, Allen J, *et al.*: **Ensembl 2021.** *Nucleic Acids Res.* 2021 Jan 8; **49**(D1): D884–D891.
PubMed Abstract | Publisher Full Text

100. Sayers EW, Beck J, Bolton EE, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res.* 2021 Jan 8; **49**(D1): D10–D17.
PubMed Abstract | Publisher Full Text

101. GTEx Consortium: **The Genotype-Tissue Expression (GTEx) project.** *Nat. Genet.* 2013 Jun 1; **45**(6): 580–585.

102. English AC, Salerno WJ, Hampton OA, *et al.*: **Assessing structural variation in a personal genome-towards a human reference diploid genome.** *BMC Genomics.* 2015 Apr 11; **16**: 286.
PubMed Abstract | Publisher Full Text

103. Fan X, Abbott TE, Larson D, *et al.*: **BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping.** *Curr. Protoc. Bioinformatics.* 2014; **45**: 15.6.1–15.6.11.
Publisher Full Text

104. Joshi I, DeRycke J, Palmowski M, *et al.*: **Genome-wide mapping of DNA double-strand breaks from eukaryotic cell cultures using Break-seq.** *STAR Protocols.* 2021 Jun 18; **2**(2): 100554.
PubMed Abstract | Publisher Full Text

105. Abyzov A, Urban AE, Snyder M, *et al.*: **CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.** *Genome Res.* 2011 Jun; **21**(6): 974–984.
PubMed Abstract | Publisher Full Text

106. Chen X, Schulz-Trieglaff O, Shaw R, *et al.*: **Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications.** *Bioinformatics.* 2016 Apr 15; **32**(8): 1220–1222.
PubMed Abstract | Publisher Full Text

107. Layer RM, Chiang C, Quinlan AR, *et al.*: **LUMPY: a probabilistic framework for structural variant discovery.** *Genome Biol.* 2014 Jun 26; **15**(6): R84.
PubMed Abstract | Publisher Full Text

108. Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res.* 2000 Jan 1; **28**(1): 27–30.
PubMed Abstract | Publisher Full Text | Free Full Text

109. Kanehisa M: **Toward understanding the origin and evolution of cellular organisms.** *Protein Sci.* 2019 Nov; **28**(11): 1947–1951.
PubMed Abstract | Publisher Full Text

110. Kanehisa M, Furumichi M, Sato Y, *et al.*: **KEGG: integrating viruses and cellular organisms.** *Nucleic Acids Res.* 2021 Jan 8; **49**(D1): D545–D551.
PubMed Abstract | Publisher Full Text

111. Schriml LM, Arze C, Nadendla S, *et al.*: **Disease Ontology: a backbone for disease semantic integration.** *Nucleic Acids Res.* 2012 Jan; **40**(Database issue): D940–D946.
PubMed Abstract | Publisher Full Text

112. Repana D, Nulsen J, Dressler L, *et al.*: **The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens.** *Genome Biol.* 2019 Jan 3; **20**(1): 1.
PubMed Abstract | Publisher Full Text

113. Piñero J, Queralt-Rosinach N, Bravo À, *et al.*: **DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes.** *Database (Oxford).* 2015 Apr 15; **2015**: bav028.
PubMed Abstract | Publisher Full Text

114. Wu T, Hu E, Xu S, *et al.*: **clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.** *Innovation (N Y).* 2021 Aug 28; **2**(3): 100141.
PubMed Abstract | Publisher Full Text

115. Yu G: **enrichplot: Visualization of Functional Enrichment Result. R package version 1.14.1.** 2021.
Reference Source

116. Chang W, *et al.*: **"Shiny: web application framework for R." R package version 1.5 (2017).** 2017.

117. Wilm A, Aw PPK, Bertrand D, *et al.*: **LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets.** *Nucleic Acids Res.* 2012 Dec; **40**(22): 11189–11201.
PubMed Abstract | Publisher Full Text

118. Grubaugh ND, Gangavarapu K, Quick J, *et al.*: **An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar.** *Genome Biol.* Jan 8, 2019; **20**(1): 8.
PubMed Abstract | Publisher Full Text

119. Cingolani P, Platts A, Wang LL, *et al.*: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3.** *Fly (Austin).* 2012 Jun; **6**(2): 80–92.
PubMed Abstract | Publisher Full Text

120. R Core Team: *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2017.
Reference Source

121. Telwatte S, Morón-López S, Aran D, *et al.*: **Heterogeneity in HIV and cellular transcription profiles in cell line models of latent and productive infection: implications for HIV latency.** *Retrovirology.* 2019 Nov 11; **16**(1): 32.
PubMed Abstract | Publisher Full Text

122. Kim D, Paggi JM, Park C, *et al.*: **Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.** *Nat. Biotechnol.* 2019 Aug 2; **37**(8): 907–915.
PubMed Abstract | Publisher Full Text

123. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013 Jan 1; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text

124. Gharavi AG, Ahmad T, Wong RD, *et al.*: **Mapping a locus for susceptibility to HIV-1-associated nephropathy to mouse chromosome 3.** *Proc. Natl. Acad. Sci. USA.* 2004 Feb 24; **101**(8): 2488–2493.
PubMed Abstract | Publisher Full Text | Free Full Text

125. AlejandroR G, Washington T, Hyink D, *et al.*: **3264 - The Multiple HIV-1 Transgenes in the Murine Model of HIV-Associated Nephropathy Fail to Segregate as Expected.** *American Society of Human Genetics Annual Meeting.* 2020 Jan 1.

126. Gener A, Fan Y, Das GC, *et al.*: **PEA0011 - Insights from HIV-1 Transgene Insertions in the Murine Model of HIV-Associated Nephropathy. 23rd International AIDS Conference (AIDS2020).** 2020 Jan 1.

127. Amazon Web Services, Inc: **Amazon Web Services.**
Reference Source

128. Landrum MJ, Lee JM, Benson M, *et al.*: **ClinVar: improving access to variant interpretations and supporting evidence.** *Nucleic Acids Res.* 2018 Jan 4; **46**(D1): D1062–D1067.
PubMed Abstract | Publisher Full Text

129. Al Khleifat A, Iacoangeli A, van Vugt JJFA, *et al.*: **Structural variation analysis of 6,500 whole genome sequences in amyotrophic lateral sclerosis.** *NPJ Genom. Med.* 2022 Jan 28; **7**(1): 8.
PubMed Abstract | Publisher Full Text

130. van Dorp L, Acman M, Richard D, *et al.*: **Emergence of genomic diversity and recurrent mutations in SARS-CoV-2.** *Infect. Genet. Evol.* 2020 Sep; **83**: 104351.
PubMed Abstract | Publisher Full Text

131. Karamitros T, Papadopoulou G, Bousali M, *et al.*: **SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies.** *J. Clin. Virol.* 2020 Oct; **131**: 104585.
PubMed Abstract | Publisher Full Text

132. Plante JA, Liu Y, Liu J, *et al.*: **Spike mutation D614G alters SARS-CoV-2 fitness.** *Nature.* 2021 Apr; **592**(7852): 116–121.
PubMed Abstract | Publisher Full Text

133. Hou YJ, Chiba S, Halfmann P, *et al.*: **SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo.** *Science.* 2020 Dec 18; **370**(6523): 1464–1468.
PubMed Abstract | Publisher Full Text

134. Davies NG, Jarvis CICMMID COVID-19 Working Group, *et al.*: **Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7.** *Nature.* 2021 May; **593**(7858): 270–274.
PubMed Abstract | Publisher Full Text

135. Syed AM, Taha TY, Tabata T, *et al.*: **Rapid assessment of SARS-CoV-2-evolved variants using virus-like particles.** *Science.* 2021 Dec 24; **374**(6575): 1626–1632.
PubMed Abstract | Publisher Full Text

136. Teng S, Sobitan A, Rhoades R, *et al.*: **Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity.** *Brief. Bioinformatics.* 2021 Mar 22; **22**(2): 1239–1253.
PubMed Abstract | Publisher Full Text

137. Choi B, Choudhary MC, Regan J, *et al.*: **Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host.** *N. Engl. J. Med.* 2020 Dec 3; **383**(23): 2291–2293.
PubMed Abstract | Publisher Full Text

138. Hamdi Y, Boujemaa M, Ben Rekaya M, *et al.*: **Family specific genetic predisposition to breast cancer: results from Tunisian whole exome sequenced breast cancer cases.** *J. Transl. Med.* 2018 Jun 7; **16**(1): 158.
PubMed Abstract | Publisher Full Text

139. Lee ST, Feng M, Wei Y, *et al.*: **Protein tyrosine phosphatase UBASH3B is overexpressed in triple-negative breast cancer and promotes invasion and metastasis.** *Proc. Natl. Acad. Sci. USA.* 2013 Jul 2; **110**(27): 11121–11126.
PubMed Abstract | Publisher Full Text

140. van Wersch S, Li X: **Stronger When Together: Clustering of Plant NLR Disease resistance Genes.** *Trends Plant Sci.* 2019 Aug; **24**(8): 688–699.

141. Belyeu JR, Chowdhury M, Brown J, *et al.*: **Samplot: a platform for structural variant visual validation and automated filtering.** *Genome Biol.* 2021 May 25; **22**(1): 161.
PubMed Abstract | Publisher Full Text

142. Wouter De Coster A, Paulin L, Avdeyev P, *et al.*: **collaborativebioinformatics/STRdust: Release 0.2 (0.2).** *Zenodo.* 2022.
Publisher Full Text

143. Liew CS, Busby B, Rlowd: **collaborativebioinformatics/kTom: Release 0.2 (0.2).** *Zenodo.* 2022.
Publisher Full Text

144. Medhat DM, Philippe S, Busby B: **collaborativebioinformatics/INSeption: Release 0.2 (0.2).** *Zenodo.* 2022.
Publisher Full Text

145. Ilovericenoodle K, Kesharwani R, Al Khleifat A, *et al.*: **collaborativebioinformatics/GeneVar2: Release 0.2 (0.2).** *Zenodo.* 2022.
Publisher Full Text

146. Agustinho DP, Sapoval N, Ramanandan SKM, *et al.*: **collaborativebioinformatics/cov2db: Release 0.2 (0.2).** *Zenodo.* 2022.
Publisher Full Text

147. Albin D, Rohit K, Soto DC, *et al.*: **collaborativebioinformatics/kvar: Release 0.2 (0.2).** *Zenodo.* 2022.
Publisher Full Text

148. Gener G, Busby B, Peralta C: **collaborativebioinformatics/imavirus: Release 0.2 (0.2).** *Zenodo.* 2022.
Publisher Full Text

149. Lo CH, Zhang S, Xu T, *et al.*: **collaborativebioinformatics/RPG_Pikachu: Release 0.2 (0.2).** *Zenodo.* 2022.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓ ? ?

Version 1

Reviewer Report 19 October 2022

https://doi.org/10.5256/f1000research.121770.r149877

© **2022 Long Q.** This is an open access peer review report 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.

?

Quan Long
University of Calgary, Calgary, Canada

In the manuscript, the authors have reported "The third hackathon for applying insights into large-scale genomic composition to use cases in a wide range of organisms" that happened in Baylor College of Medicine in 2021.

Detailed descriptions of eight software packages focusing on structural variation detection have been provided. The data for benchmarking are also mentioned. However, I do not see any quantitative results regarding to the evaluation of the software. For instances: what are the true/false positives for the tools? What are the advantages/disadvantages of different tools when comparing to each other? The manuscript focuses more on describing the algorithms and the data, which are important. However, without an evaluation, this seems to be an unfinished work. If the evaluation could be added, this looks an excellent review and comparison of SV tools.

Also it is clear that different authors have written their pieces and then submitted. The work looks quite segmented without smoothing.

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

Is the description of the software tool technically sound?
Partly

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?

No

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.

Reviewer Expertise: Genomics, Genetics, Machine Learning

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

Reviewer Report 12 October 2022

https://doi.org/10.5256/f1000research.121770.r149876

© 2022 Ferreira P. This is an open access peer review report 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.

?

Pedro G. Ferreira
University of Porto, Porto, Portugal

Kimberly Walker and colleagues presents a manuscript that reports the results of a workshop meeting where different groups have applied their developed tools for the analysis of genomic variation.

The manuscript is written in an unconventional way, which is hard to follow.

My main concern with the manuscript is that I did not find a conducting line or a specific goal, rather than reporting the methods that were applied. The title mentions a wide range of organisms. In the introduction, it is mentioned a focus on tomatoes, plants, Sars-cov2, and some methods like INSeption analyse human data from Genome in a Bottle project. Overall, that is too confusing.

The text focuses on describing the tools and their workflows. It would be interesting to discuss the results from a qualitative and quantitative point-of-view. This way, one would obtain a better intuition of the relevance and utility of the tool.

What are the results of applying these tools in tomato genome? It would be very interesting to summarize the novel findings on the tomato genome or the human genome from the different views of applying all these tools. For instance, in the INSeption section it is mentioned allele frequency analysis. I'm very curious about these results, but these were not discussed at all.

From the quantitative point of view, what are the computational demands of these tools? What are the performance gains of using a k-mer approach when compared with other approaches?

Overall, without a clear focus on the goals and the reporting of the insights gained, the manuscript is of little use. Therefore, I recommend a substantial re-organization and improvement of the manuscript.

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

Is the description of the software tool technically sound?
Partly

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?
No

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.

Reviewer Expertise: Transcriptomics

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

Reviewer Report 23 September 2022

https://doi.org/10.5256/f1000research.121770.r149878

© 2022 Le N. This is an open access peer review report 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.

Nguyen Quoc Khanh Le
Taipei Medical University, Taipei, Taiwan

In this study, the authors briefly went through all algorithms in the third international hackathon for applying insights into large-scale genomics composition. Overall, it is well-written and holds

potential for indexing. I just have some minor comments as follows:

1. I'm wondering whether the authors included all projects from the hackathon or not. Any reasons to exclude the other projects out of this analysis?

2. Adding more use cases to each algorithm would be more interesting.

3. The authors should add more discussions to show the biological insights of each tool/algorithm.

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.

Reviewer Expertise: Genomics analysis

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

Reviewer Report 22 June 2022

https://doi.org/10.5256/f1000research.121770.r139299

© 2022 Aramayo R. This is an open access peer review report 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.

?

Rodolfo Aramayo
Department of Biology, Texas A&M University, College Station, TX, USA

Introduction:

In this manuscript, Kimberly Walker and colleagues describe and summarize the results of *"The third hackathon for applying insights into large-scale genomic composition to use cases in a wide range of organisms."*

This virtual hackathon took place in October 2021 and encompassed a total of 59 scientists from 14 countries and 13 U.S. states.

In this manuscript, the authors describe the evaluation of eight software packages aimed at identifying Structural Variants (SVs) by using mainly k-mers-based tools, among others.

This manuscript addresses key problems in Computational Genomics. Importantly, unlike before, the 2021 hackathon features work on plants, which provides an important alternative point of view to the human-focused research of previous hackathons.

Below you will find a series of general and specific recommendations that will, in my opinion, significantly improve the readability and overall quality of this work, while correcting a series of issues I have found associated with both the online version and the downloaded PDF version of the manuscript.

General Comments:
I hope you will agree with me that this is, by no means, a "standard" manuscript. Yet, the manuscript was written trying to adjust the logic of the text to that of a "standard" manuscript. It is very hard to follow the logic of what was described in the text, as the text is presented.

Let me explain: any potential reader that is trying to learn, understand and follow what was done with any one of the software packages here described would have an extremely hard time trying follow the logic of a given tool under the current manuscript format.

The current format is fragmented. Information related to a given software package is scattered around the text. It is hard to follow.

This is why I strongly suggest to reorganize the text of the manuscript as follows:

1. Abstract
2. Introduction to the general topic **only** (not to the different packages used)
3. Methods. Describe the methods general to all the work (Cloud Computing)
4. STRdust: Detect and genotype short tandem repeats
4.1. Introduction to STRdust
4.2. Experimental Rationale, Methods and Discussion for STRdust
4.3. Next Steps or Recommended Future Directions for STRdust
5. kTom: k-mers for profiling tomato introgressions
5.1. Introduction to kTom
5.2. Experimental Rationale, Methods and Discussion for kTom
5.3. Next Steps or Recommended Future Directions for kTom
6. cov2db
6.1. Introduction to cov2db
6.2. Experimental Rationale, Methods and Discussion for cov2db
6.3. Next Steps or Recommended Future Directions for cov2db

...
...
...
X. Conclusions
Y. Tables
Y.1. Table 1. Lists the data source utilized by each tool developed during the hackathon.
Y.2 Table 2. A table compiling the Data, and Software Availability and others.
V. Acknowledgements
Z. References

Regarding what I call in the example above section Y.2, I would like to see the data associated Data, Software availability, and others organized as a single table with the following format or something similar:

==========
Table 2. Software availability

	Row 1	Row 2
Column 1: Tool Name	STRdust	
Column 2: Source code	https://github.com/collaborativebioinformatics/STRdust	
Column 3: Version No.	0.2	
Column 4: Link to Archive	https://doi.org/10.5281/zenodo.6467829	
Column 5: License	MIT	

==========

In summary, I think that the format described will be easier to read, as it will contain all the information and figures relating to a piece of software together. It will follow a logical flow. As far as I am concerned, this is a better way to present this work. It is not a "conventional" way, but then again, this is not a "conventional" manuscript.

Specific Comments:

Comment 01:
The paragraph:
"These genomic variants are typically classified as deletions, duplications, insertions, inversions, and translocations describing different combinations of DNA gains, losses, or rearrangements. Copy number variations (CNVs) are a particular subtype of SVs mainly represented by deletions and duplications. SVs are typically described as single events, although more complex scenarios involving combinations of SV types exist.1,2"

Should read:
These genomic variants are typically classified as deletions, duplications, insertions, inversions, and translocations describing different combinations of DNA losses, gains and/or rearrangements. SVs are typically described as single events, although more complex scenarios involving combinations of SV types exist. Copy number variations (CNVs) are a particular subtype of SVs mainly represented by deletions and duplications. 1,2

Rationale: As originally written, the paragraph starts describing Structural Variants (SVs), then switches to Copy Number Variations (CNVs), and the returns to SVs. I would like to see the ideas of SVs presented together.

Comment 02:
The paragraph:
"In October 2021, 59 researchers from 14 countries participated virtually in the third Baylor College of Medicine & DNAnexus hackathon, focusing on interrelated topics such as SVs, short tandem repeats (STRs), k-mer profiling, viruses, reference refinement and annotation."

Should read:
In October 2021, 59 researchers from 14 countries participated virtually in the third Baylor College of Medicine & DNAnexus hackathon, focusing on interrelated topics such as k-mer profiling, short tandem repeats (STRs), SVs, reference refinement, annotation and viruses.

Rationale: I would like to see the presented topics sorted according to those that are related to large genomes and then small genomes (i.e., viruses). This order should also be applied to the rest of the manuscript.

Comment 03:
The paragraph:
"The international hackathon focused on nine softwares to answer these questions; eight of which we present in this paper: STRdust, kTom,INSeption, GeneVar2, cov2db, K-var, Imavirus, and a Reference Panel Generator (RPG) for diverse sequencing data analysis. Several emergent themes became apparent over the course of the hackathon."

Should read:
The international hackathon focused on nine software packages (eight of which we present in this paper), to answer these questions: K-var, kTom, STRdust, INSeption, GeneVar2, cov2db, RPG, and Imavirus. Several emergent themes became apparent over the course of the hackathon.

Rationale: I think it reads better and lists the virus-related package last (see Comment 02).

Comment 04:
The paragraph:

"Nucleotide sequence substrings of length k (k-mers) continue to prove useful in SV work and in genomics, however, the time needed to assess the frequency of SVs presents a resource problem.7 The reduction of the computational resources required to complete an SV assessment in a genome would allow greater amounts of SV data to be processed in genomic workflows. Many bioinformatic tools currently used to locate genomic SVs use a sliding window alignment technique, which can be time-consuming.8,9 However, implementing a k-mer based approach to create a pool of reference k-mers of known SVs, the annotation speed of variation in new genomes might be increased.10,11 k-mers have also been used in alignment-free methods, bypassing the need for reference genomes.12"

Needs serious rearrangement. I read this paragraph as follows:

Here you are telling me that k-mers are useful for SV work:

"Nucleotide sequence substrings of length k (k-mers) continue to prove useful in SV work and in genomics,"

Here you are telling me that generating SV is computationally expensive:

"however, the time needed to assess the frequency of SVs presents a resource problem.7"

Here you are telling me that if we were to reduce the time it takes to compute them, they would then be more likely to be incorporated in genomic workflows:

"The reduction of the computational resources required to complete an SV assessment in a genome would allow greater amounts of SV data to be processed in genomic workflows."

Here you are describing that time consuming sliding windows alignment techniques are currently used to locate genomic SVs:

"Many bioinformatic tools currently used to locate genomic SVs use a sliding window alignment technique, which can be time-consuming.8,9"

Here you are telling me that if we were to implement a k-mer-based approach, the annotation speed would increase:

"However, implementing a k-mer based approach to create a pool of reference k-mers of known SVs, the annotation speed of variation in new genomes might be increased.10,11"

Here you are stating that k-mer have been used in alignment-free methods:

"k-mers have also been used in alignment-free methods, bypassing the need for reference genomes.12"

I would like to see these ideas presented as follows (re-do the text):

First, introduce k-mers and how k-mers are being used in genomics:

"k-mers have also been used in alignment-free methods, bypassing the need for reference genomes.12"

Second, state that k-mers are useful:

"Nucleotide sequence substrings of length k (k-mers) continue to prove useful in SV work and in genomics,"

State that the way SV are calculated is time-consuming:

"Many bioinformatic tools currently used to locate genomic SVs use a sliding window alignment technique, which can be time-consuming.8,9"

Third, state that calculating SV is expensive:

"however, the time needed to assess the frequency of SVs presents a resource problem.7"

Fourth, tell us why using k-mers might be better:

"However, implementing a k-mer based approach to create a pool of reference k-mers of known SVs, the annotation speed of variation in new genomes might be increased.10,11"

Finally, conclude how faster calculation is likely to result on SV data to be readily incorporated in existing genomic workflows, thus improving the overall genome annotation:

"The reduction of the computational resources required to complete an SV assessment in a genome would allow greater amounts of SV data to be processed in genomic workflows."

STRdust Specific Comments:

Comment 05:

In following Resource IDs:

(RRID:SCR_010233)
(RRID:SCR_017990)
(RRID:SCR_003756)
(RRID:SCR_002796)
(RRID:SCR_018801)

their links do not work because the RESEARCH RESOURCE IDENTIFICATION PORTAL has changed. The only way I have found to access the resource is by accessing the legacy RRID website (https://scicrunch.org/resources-legacy/).

Regarding resource SCR_002796, it points to a github repository (https://github.com/alibashir/pacmonstr), which, in my opinion is the link it should be used (added as a reference?). The same can be said about resource SCR_018801, that points to the https://github.com/davidebolo1993/TRiCoLOR, repository.

In general, I think that GitHub repositories that have not been archived via databases like Zenodo, should be linked directly to GitHub.

Comment 06:

In the following paragraph:

"To mitigate this issue, several long-read STR calling methods have been developed in recent years, including PacmonSTR21(RRID:SCR_002796), NanoSatellite,22 TRiCoLOR23(RRID:SCR_018801), and Straglr24 -- however, their usability remains limited due to platform and/or computational demands"

You state that either the usability or computational demands of the packages: PacmonSTR, NanoSatellite, TRiCoLOR, and Straglr, are problematic.

In fairness to the authors of those packages, you should provide data supporting that statement.

In the legend of Figure 1 you clearly state:

"We evaluated STRdust by comparing the results of simulated STR expansions produced by SimiSTR based on the human (Genome Reference Consortium Human Build 38, GRCh38) and tomato (Solanum lycopersicum 4.0, SL4.0) reference genomes, to two novel tools: Straglr24 and TRiCoLOR.23"

And in the text, you state:

"STRdust results were compared to TRiCoLOR 1.1,23 and Straglr 1.1.124 using default parameters."

But you never really showed any data that would allow an independent observer to reach the same conclusion you claim to have reached.

In my opinion, this is not fair to the developers of the other packages and should be corrected.

kTom Specific Comments:

Comment 07:

Please define: *"k-mers with low-mid range frequencies"*

Comment 08:

Please state the length or range of lengths of the Illumina reads used in these experiments.

It is not clear to me how the read length affects the outcome of this analysis (has anyone tested this parameter?).

Comment 09:

Also, it is important to state the k-mer length or k-mer length range that was used for this analysis.

What is the minimum k-mer length that is calculated/used by this package?

Comment 10:

As I suspect that the read coverage will have a profound effect on these experiments, therefore, I would like to see a table (or Supplementary table), where the coverage for the different datasets analyzed is presented.

Comment 11:

Where the FastQ read lengths normalized in these experiments?

If they were not, then it is hard for me to evaluate the meaning of the results displayed in Figure 2. Again, data requested in Comment 09 is important.

Comment 12:

The link to resource RRID:SCR_014583, is not working. Also, this resource should link directly to the FastQC web site (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Comment 13:

Similarly, link to resources SCR_013001, SCR_005491, SCR_017014, and SCR_017270 are not working.

Comment 14:

When searched manually, resource SCR_013001 links to the Sourceforge link: https://sourceforge.net/projects/flexbar/, but in there the Flexbar version 1.4.0, used in the work is not present. The oldest version present there is for version 2.2. Also, the official Flexbar website has been moved to GitHub (https://github.com/seqan/flexbar). Version 1.4.0, is not available there either, thus a person wanting to reproduce these results using the exact same versions presented in this manuscript, would be unable to do so.

This needs to be corrected.

Comment 15:

Resources SCR_005491 and SCR_017014, should link directly to GitHub (https://github.com/gmarcais/Jellyfish and https://github.com/schatzlab/genomescope, respectively), and resources SCR_017270 and SCR_019186, should link directly to Bioconductor (https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html) and CRAN (https://cran.r-project.org/web/packages/tidyverse/index.html), respectively.

Comment 15:

The resolution of Figure 2 must be increased. When the PDF file is displayed on a large screen the figure is extremely pixelated and hard to interpret.

INseption Specific Comments:

Comment 16:
The resources for
SCR_010910 - BWA,
SCR_018550 - Minimap2,
SCR_017620 - NGMLR,
SCR_004603 - DELLY,
SCR_004603 - SNIFFLES \<==Note wrong ID!,
SCR_015880 - Canu,
SCR_017016 - Flye,

SCR_005227 - BCFTools,
SCR_002105 - SamTools,
SCR_008394 - Python, and
SCR_015880 - Canu

Have links not working and, again, in my opinion should point to their GitHub or SourceForge repositories.

I really appreciate the work of trying to make uniform the web addresses citations, but in the long term, as the repositories get updated, it will create a problem.

Please consider adding these links as part of the references.

Modern Reference Managers like Paperpille (https://paperpile.com/app), do an excellent job doing this.

Comment 17:
You state that: *"We filtered out SVs that were supported by less than 10 reads using bcftools 1.12"*

Bcftools is a complex set of scripts. Please provide the exact command used.

Comment 18:
Figure 3 is very hard to understand. The problem is that you are using lower-case letters and upper-case letters to refer to the different panels. Do not do that. Use Panel A, SubPanel I, etc.

This figure needs to be edited or replaced and the figure legend needs to be re-written so as to make it digestible. It is not clear what are you trying to show: are you showing haplotypes? Are you showing reference and experimental regions? Why is the F panel on the right? And the C, D, and E panels on the left? This is confusing. I understand you are trying to save space, but the logic must be presented sequentially.

Comment 19:
In the section: *"Clustering unmapped reads"*

You state:

"To be able to assemble a sequence from all unmapped reads, we tried several approaches. We attempted to identify clusters of reads using the LROD version 1.094 package, which we found unsuitable for our purposes due to long runtimes. More successfully, we used the program CARNAC-LR version 1.0.095 to build clusters of reads using Minimap2 version 2.22 aligner32 and a subsequent k-mer based clustering approach."

Here, you need to define what you mean by *"due to long runtime"*.

Also, you need to define and describe what you mean by:

"and a subsequent k-mer based clustering approach."

You need a reference to a publication or to a script name or to a repository.

Comment 20:

In the section: *"Identifying integration sites for assembled clusters"*

You state:

"Having successfully assembled contigs for N = 15 read clusters using Canu v2.236(RRID:SCR_015880), we searched for overlap of these contigs with the breakpoint regions of 30 previously identified long insertion sites."

How were those 30 regions previously identified?

Comment 21:

Also related to Comment 20, are you stating that you re-identified regions that were previously identified as containing insertions?

If this is the case, can your experimental logic be applied to identify unknown insertion points?

If the answer is yes, then why have you not done that?

Can you expand Figure 3 to clarify this point?

Comment 22:

Please comment on the Genome Coverage needed for INseption to work as expected.

Has anyone tested the effects of Genome Coverage on INseption's performance?

Comment 23:
The most important part that must be addressed in this section is the fact that while you have explained the logic of using sample reads to detect insertions in the reference genome, by definition, in comparing genomes from the same population, an insertion in Genome A is a deletion in Genome B and an insertion in Genome B, is a deletion in Genome A. So, it is not clear to me how your experimental logic would be applied to a deletion present in the reference genome and an insertion in the sample reads. As things are explained here, it looks to me that you seem to have only solved 50% of the problem.

GeneVar2 Specific Comments:

Comment 16 applies to this section.

Comment 24:
How is GeneVar2 different or better or what are its advantages when compared to seeing this same information in a good old genome browser like Ensembl? Please elaborate.

Comment 25:
Also, I think it would really help if you could add a figure showing the how a gene like BRCA2 (ENSG00000139618), displays in GeneVar2. That would help people not familiar with the **GeneVar2** browser to be able to see it in action.

cov2db Specific Comments:

Comment 16 applies to this section.

Comment 26:
In the statement:
"Minimal system requirements for a local cov2db instance are dictated by the mongoDB requirements with the key limiting factor being RAM used. Large variant databases will consume substantial amounts of RAM, and we suggest hosting those on dedicated high memory compute servers."

You need to be more specific. Define *large database,* and be more specific about RAM usage. Also, when recommending to use dedicated high memory compute servers, please be more specific on the minimal requirements needed.

Comment 27:
In the statement:
"Our current design supports input VCFs generated by LoFreq117 (RRID:SCR_013054) or converted into VCFs from the iVar118 output via provided script."

Please give the name of the *"provided script".*

Comment 28:
To truly appreciate the package cov2db in action, I would recommend the authors present a screenshot of an actual data display of a region of the virus in the R Shiny app.

K-var Specific Comments:

Comment 16 applies to this section.

Comment 29:
The URL (http://www.dtp.nci.nih.gov/) present in the resource RRID:SCR_003057 (when searched manually, because the link does not work...) point to a dead web site.

Comment 30:
The statement:
"k-mer frequencies were obtained for each sample, using the tool Jellyfish version 2.3.0."

Needs clarification. What kind of data was used to obtain the k-mers? FastQ files? BAM files? Tables?

Comment 31:
You mention that you used *"whole exome sequencing of the NCI-60 dataset"*, but based on the data you gave me I was unable to find that dataset. Even when I searched the NCBI-SRA website using

the terms *"NCI-60 cancer"* I was not able to easily identify the exome datasets in question.

Please give specific links to the data you used in these experiments.

Comment 32:
It is totally unclear to me how the data coming from these different samples compare to each other in terms of coverage/number of reads.

It would help if you were to provide a table with this data. K-mer frequency is dependent on coverage.

Comment 33:
Please provide the name of the *"custom script"* used to tabulate the data.

Is this script part of your submission?

Comment 34:
Please define *"low frequency k-mers"*

Comment 35:
How were the control and test datasets pre-defined? And how you implemented the TF-IDF test? (R Script?, Python Script?).

Imavirus Specific Comments:

Comment 16 applies to this section.

Comment 36:
The statement:
"During the hackathon we were able to verify a previously reported integration site on mouse chr8 which can be seen in our GitHub repository cited in the Software Availability section."

How did you do this? What commands did you use? Where is your computational pipeline?

Comment 37:
The statement:
"Using unbiased RNA-seq datasets, Imavirus aimed to identify pIS..."

Please define *"unbiased"*

Comment 38:
The statement:
"Future work should explore the datasets we scoped out in SRA for more physiological systems such as animal models or stable cell lines to identify more putative insertion sites."

What do you mean by: *"more physiological systems such as animal models or stable cell lines to identify more putative insertion sites"*

This statement makes no sense to me. A bacterial cell does not have less Physiology than a Human Liver cell. It has a *different* physiology. Please rephrase this idea.

Comment 39:
Based on what was presented in this manuscript Imavirus is an idea, not a tool. The authors are not presenting a logical code-based pipeline to complete this analysis.

The Zenodo repository contains tables and figures, not code, and although this is a nice idea, I cannot understand why this was included on this manuscript. The tool Imavirus as presented is an idea and should not be part of this report. Alternatively, the authors should provide a viable computational pipeline that can be used to perform a similar analysis. I was unable to find not a single draft text page containing a list of commands.

As this *"pipeline"* is presented is completely unreproducible and useless.

This section is by far below the standards of the other tools presented. It should either be removed or re-done. As it is, it hurts the manuscript as it does not follow the F1000 Condition for publication:

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

In addition, the authors make no mention of competing tools like ViFi, VirTect, and VIRUSBreakend.

In fact, when they mention:
"However, identifying virus integration sites from genomic DNA is challenging and there are not many bioinformatics tools available to reliably detect viral presence or integration events."

Without mentioning competing tools, is not fair to the authors of those other competing tools and, in my opinion, it is misleading.

I have to conclude that, after reading this section I am left with the taste that Imavirus is an idea that is being presented so as to be "sold" to be incorporated into a commercial product.

RPG Specific Comments:

Comment 16 applies to this section.

Comment 40:
In the legend of Figure 8 you state:
"Only common alleles (\>5% allele frequency (AF)) in the variant call set are retained."
and
"Subsequently, common allele calls are replaced with CHM13 rare alleles in CHM13 FASTA genome sequence."

Please elaborate how you did that.

Also elaborate if you do that for both haplotypes or not.

Comment 41:
It would help to have a table with the actual numbers of variants calls, non-pathogenic variants, pathogenic variants, number of in-frame stop codons, etc.

Comment 42:
It is not clear to me the origin of the list of 1KGP common alleles. Where those downloaded or calculated? And, if they were calculated, can you provide their list?

Conclusion:
This is a good piece of work. Here, I am offering a series of suggestions that, I am sure, will improve the overall quality of this work.

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?
No

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?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genetics, Genomics, Computational Genomics, Epigenetics, Fungal Genetics and Biology, Metazoan Genome Organization and Digital Biology

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

Reviewer Response 19 Oct 2022

Rodolfo Aramayo, Texas A&M University, College Station, USA

As requested from the authors, I am expanding some of my previous comments so as to make them more clear. Below, I will post the original comment and its associated expanded

comment(s).

kTom Specific Comments:

Comment 10:

As I suspect that the read coverage will have a profound effect on these experiments, therefore, I would like to see a table (or Supplementary table), where the coverage for the different datasets analyzed is presented.

Expanded Comment.

1. After Quality Control, where the resulting read lengths distributions observed for the different genome samples equivalent?

2. After Quality Control, was the Sample Coverage (No. Reads x Length per Read) equivalent between the different samples?

My main concern is with the possibility of some genomes having a significantly lower Sample Coverage, which could result in under-counting Kmers for that particular genome.

The requested table in question should present the result of adding the lengths of all the reads corresponding to a given genomic dataset, divided by the total number of reads for such dataset. This “Sample Coverage” calculation should be performed for each genome and the results summarized on a table. These numbers could also be ‘normalized’ in relation to the genome with the lowest sample coverage.

INseption Specific Comments:

Comment 22:

Please comment on the Genome Coverage needed for INseption to work as expected.

Has anyone tested the effects of Genome Coverage on INseption's performance?

Expanded Comment.

As presented, the authors first aligned HiFi reads to GRCh37 using Minimap2 and then used Sniffles to call SVs. The authors also report that they filtered out SVs that were supported by less than 10 reads using bcftools.

My question about the Genome Coverage needed for INseption to work as expected, is:

How did you test or controlled for chromosomal regions that were either not-covered or had low coverage after Minimap2 mapping?

Given that the existence of such regions could potentially be the source of false negatives SV in your analysis.

Related to the question, if anyone has tested the effects of Genome Coverage on INseption's performance, I would like to know how many HiFi reads spanning a SV are needed for Sniffles to call an SV.

Also, given that the shorter the read, the higher the probability that the read in question could be assigned to two different genomic positions by Minimap2, thus generating different SAM flags, has anyone tested how these parameters affect Inseption's performance? What was the minimal insert size accepted for analysis?

Comment 23:

The most important part that must be addressed in this section is the fact that while you have explained the logic of using sample reads to detect insertions in the reference genome, by definition, in comparing genomes from the same population, an insertion in Genome A is a deletion in Genome B and an insertion in Genome B, is a deletion in Genome A. So, it is not clear to me how your experimental logic would be applied to a deletion present in the reference genome and an insertion in the sample reads. As things are explained here, it looks to me that you seem to have only solved 50% of the problem.

Expanded Comment.

This is related to the potential existence of regions in the reference genome that have no mappers, not because of low coverage, but because the donor genome of the reads in question have a deletion in a region that is present in the reference genome. Like I said before: For the same synthenic region, if those regions have deletions/insertions, an insertion in Genome A is a deletion in Genome B and an insertion in Genome B, is a deletion in Genome A. Both genomes could have a deletion (in relation to a third genome), or an insertion (again in relation to a third genome), and, despite that, be considered to be identical. But if Genome A has an insertion in relation to Genome B, it is also possible that Genome B has an insertion in relation to Genome A. Detecting insertions in Genome A in relation to Genome B, does not detect insertions in Genome B in relation to Genome A.

The way the data was presented, it was not clear to me your approach took both possibilities into consideration.

K-var Specific Comments:

Comment 32:

It is totally unclear to me how the data coming from these different samples compare to each other in terms of coverage/number of reads.

It would help if you were to provide a table with this data. K-mer frequency is dependent on coverage.

Expanded Comment.

The calculation of Kmers starts with Short-read sequencing data. The authors state they used: 7 non-metastatic and 5 metastatic samples. My question is: How those different samples compare in terms of number of reads and the total number of bases present in that particular sample.

To answer that question, the authors need to calculate the result of adding the lengths of all the reads corresponding to a given sample, divided by the total number of reads for such dataset. Such calculation will reveal the theoretical "sample coverage" (not genome coverage), of each sample and allow us to compare if such coverage between different samples is equivalent.

For example, I can see how a given sample whose sample coverage is half of that another sample, could potentially generate a different number of Kmers.

It follows that if one were to compare two samples one metastatic with one non-metastatic both having exactly the same sample coverage, would you be able to re-identify the same genes? And, how reducing sample coverage would affect the resulting kmer table?

RPG Specific Comments:

Comment 41:

It would help to have a table with the actual numbers of variants calls, non-pathogenic variants, pathogenic variants, number of in-frame stop codons, etc.

Expanded Comment. What I am requesting is a table, similar to the one you presented in your github repository (3. Biologically annotated variants (CHM13 based).), where you would summarize the percentage of common variants that resulted from your annotation.

Imavirus Specific Comment:

In re-reading the manuscript I became concerned about how potentially misleading is to present a GITHUB repositiory that supposedly present a computational pipeline, when in fact such said repository does not contain a single line of code.

Competing Interests: None

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research