Human Genetics (2023) 142:201–216
https://doi.org/10.1007/s00439-022-02494-1

ORIGINAL ARTICLE

SCIP: software for efficient clinical interpretation of copy number variants detected by whole-genome sequencing

Qiliang Ding[1,2,3] · Cherith Somerville[1,2,3] · Roozbeh Manshaei[1,3] · Brett Trost[3,4] · Miriam S. Reuter[3,4,5] · Kelsey Kalbfleisch[1,2] · Kaitlin Stanley[1] · John B. A. Okello[1,2,3,6] · S. Mohsen Hosseini[1,7] · Eriskay Liston[1,2] · Meredith Curtis[1] · Mehdi Zarrei[3,4] · Edward J. Higginbotham[3,4,8] · Ada J. S. Chan[3,4] · Worrawat Engchuan[3,4] · Bhooma Thiruvahindrapuram[3] · Stephen W. Scherer[3,4,9] · Raymond H. Kim[1,2,10,11] · Rebekah K. Jobling[1,2,8]

Received: 26 May 2022 / Accepted: 9 October 2022 / Published online: 14 November 2022

Abstract

Copy number variants (CNVs) represent major etiologic factors in rare genetic diseases. Current clinical CNV interpretation workflows require extensive back-and-forth with multiple tools and databases. This increases complexity and time burden, potentially resulting in missed genetic diagnoses. We present the Suite for CNV Interpretation and Prioritization (SCIP), a software package for the clinical interpretation of CNVs detected by whole-genome sequencing (WGS). The SCIP Visualization Module near-instantaneously displays all information necessary for CNV interpretation (variant quality, population frequency, inheritance pattern, and clinical relevance) on a single page—supported by modules providing variant filtration and prioritization. SCIP was comprehensively evaluated using WGS data from 1027 families with congenital cardiac disease and/or autism spectrum disorder, containing 187 pathogenic or likely pathogenic (P/LP) CNVs identified in previous curations. SCIP was efficient in filtration and prioritization: a median of just two CNVs per case were selected for review, yet it captured all P/LP findings (92.5% of which ranked 1st). SCIP was also able to identify one pathogenic CNV previously missed. SCIP was benchmarked against AnnotSV and a spreadsheet-based manual workflow and performed superiorly than both. In conclusion, SCIP is a novel software package for efficient clinical CNV interpretation, substantially faster and more accurate than previous tools (available at https://github.com/qd29/SCIP, a video tutorial series is available at https://bit.ly/SCIPVideos).

Abbreviations

ASD	Autism spectrum disorder
B/LB	Benign or likely benign
CGC	Cardiac Genome Clinic
CNV	Copy number variation
GUI	Graphical user interface
HI	Haploinsufficient/haploinsufficiency
IGV	Integrative Genomics Viewer
OMIM	Online Mendelian Inheritance in Man database
pext	Proportion expressed across transcripts

✉ Raymond H. Kim
raymond.kim@sickkids.ca

✉ Rebekah K. Jobling
rebekah.jobling@sickkids.ca

1 Ted Rogers Centre for Heart Research, Cardiac Genome Clinic, The Hospital for Sick Children, Toronto, ON, Canada

2 Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, Toronto, ON, Canada

3 The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada

4 Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON, Canada

5 CGEn, The Hospital for Sick Children, Toronto, ON, Canada

6 MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA

7 Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

8 Genome Diagnostics, Department of Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, ON, Canada

9 Department of Molecular Genetics and the McLaughlin Centre, University of Toronto, Toronto, ON, Canada

10 Fred A. Litwin Family Centre in Genetic Medicine, University Health Network, Toronto, ON, Canada

11 Department of Medicine, University of Toronto, Toronto, ON, Canada

pLI	Probability of being loss-of-function intolerant
P/LP	Pathogenic or likely pathogenic
popmax	The continental population with the highest allele frequency (gnomAD)
RAM	Random access memory
SCIP	Suite for CNV Interpretation and Prioritization
SNV	Single-nucleotide variants
SV	Structural variation
TCAG	The Centre for Applied Genomics
TS	Triplosensitive/triplosensitivity
WGS	Whole-genome sequencing

Introduction

Deletions or duplications of human genomic regions are collectively termed copy number variants (CNVs). CNVs range in size from 50 base-pairs (bp) to mega base-pairs (Mb). CNVs can exist as benign variations, e.g., > 20,000 CNVs with an allele frequency > 1% have been catalogued (Collins et al. 2020; Feuk et al. 2006; Iafrate et al. 2004; MacDonald et al. 2014; Sebat et al. 2004); however, they are also a major contributor to genetic disease as numerous contiguous gene syndromes have been documented (Amberger et al. 2019; Yuen et al 2017; Cerruti Mainardi 2006; Costa et al. 2022; McDonald-McGinn et al. 2015; Oskoui et al. 2015; Pereira and Marion 2018; Zarrei et al. 2019). Whole-gene deletions and intragenic deletions or duplications can cause loss-of-function, while whole-gene duplications result in local triploidy. To date, > 300 genes have been curated by the ClinGen Consortium (Rehm et al. 2015) as haploinsufficient.

Whole-genome sequencing (WGS) is increasingly being recommended as a first-line test for suspected rare genetic disorders (Lionel et al. 2018; Manickam et al. 2021; Marshall et al. 2020; NICUSeq Study Group et al. 2021). One major advantage of WGS is the ability to detect single-nucleotide variants (SNVs), indels, CNVs, and copy-neutral structural variants (SVs) genome-wide in a single test. Furthermore, unlike karyotyping and chromosomal microarray analysis (CMA), which have lower bounds on the CNV sizes they can detect, WGS identifies CNVs of all sizes. In recent studies, WGS was found to have superior sensitivity in detecting CNVs compared with CMA (Gross et al. 2019; Jiang et al. 2013; Trost et al. 2018). CNV detection in paired-end WGS data is supported by three major types of evidence: read depth (Abyzov et al. 2011; Handsaker et al. 2011; Zhu et al. 2012), paired-end reads with abnormal insert size and/or orientation, and split reads (Figure S1), with the latter two commonly referred to as anomalous reads (Chen et al. 2016).

Several factors make the clinical interpretation of CNVs challenging. Caused by the relatively higher false detection rate of CNVs compared with SNVs and indels, greater emphasis must be placed on variant quality assessment. Tools such as the Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al. 2013) are used to visually inspect read depth and/or anomalous reads at the putative CNV. However, this can be time-consuming, as it can take one minute or more per CNV for IGV to display alignments (depends on read depth and size). In addition, visualization relative to other annotations is essential to interpret CNVs. During this step, an analyst must query multiple databases, e.g., gnomAD-SV (Collins et al. 2020) and DGV (MacDonald et al. 2014) for benign variation, ClinGen dosage sensitivity curations for haploinsufficient (HI) and triplosensitive (TS) regions, genome browsers for genes, and ClinVar (Landrum et al. 2018) and DECIPHER (Firth et al. 2009) for pathogenic variation. This process is complex and error-prone, as it involves back-and-forth maneuvers and synthesizing evidence across multiple webpages.

Several publicly available tools have been developed to address these inefficiencies. ClinSV is a CNV/SV analysis pipeline that uses custom IGV tracks to display binned read depth and mapping quality, anomalous reads, and variants from both an internal database and DGV (Minoche et al. 2021). Another tool, samplot, pre-generates read depth and anomalous read plots for manual review of CNV quality (Belyeu et al. 2021). CNspector is an interface for interactive viewing of copy number changes at scales from single-exon to genome-wide, particularly suitable for cancer genomes (Markham et al. 2019). CNVxplorer allows users to query biological annotations (e.g., pathway enrichment, KO models, regulatory regions) of a CNV, but could not be used for variant quality assessment (Requena et al. 2021). AnnotSV (with visualization provided by knotAnnotSV) is a web-based tool that performs annotation, prioritization, and visualization for human CNVs and SVs (Geoffroy et al. 2021). However, it is not capable of visualizing variant quality or incorporating it for prioritization. In summary, some of these tools are suitable for CNV quality assessment, while others are designed to explore clinical relevance; however, no publicly available tool exists that allows for investigation of both aspects simultaneously in a unified and integrated environment.

Here we present the Suite for CNV Interpretation and Prioritization (SCIP), which provides a Visualization Module that near-instantaneously displays all information necessary for clinical CNV interpretation. The Visualization Module is supported by a backend, providing variant filtration and prioritization. SCIP was rigorously evaluated using 1027 families ascertained for congenital cardiovascular disease and/or autism spectrum disorder (ASD). SCIP performed substantially better than a spreadsheet-based manual workflow and AnnotSV (Geoffroy et al. 2021).

Materials and methods

Computational requirements and implementation of SCIP

The SCIP backend was implemented at the high-performance computing facility of SickKids Research Institute. Each instance of the SCIP backend was run on a single core on a server with a 2.3-GHz Skylake Intel Xeon processer and 8 GB of memory (unless for CNVs > 10 Mb in size [up to 64 GB of memory was used]) running CentOS 7. Required software include `Perl` (v5.16), `R` (v3.5.1), `samtools` (v1.10), `bedtools` (v2.26), and `tabix` (v0.2.5). The Visualization Module was implemented on the analysts' personal computers, running macOS Big Sur or Windows 10 using Google Chrome. Required software includes `R`, RStudio, and three R packages (`shiny`, `DT`, and `plotrix`) and their dependencies. See Supplementary Materials for details on generating the annotation files used by SCIP.

The spreadsheet-based manual workflow for clinical CNV interpretation

After variant calling, all CNVs in a given sample were processed by an in-house annotation pipeline, which outputs a spreadsheet where each row is a CNV and each column is an annotation. The following types of annotation were included: variant quality, overlap with common (variant frequency > 1%) and rare CNVs in gnomAD-SV, DGV, and internal control databases, overlap with genes and curated dosage sensitive regions, gene constraint information (e.g., gnomAD and ExAC pLI Karczewski et al. 2020; Lek et al. 2016)), and gene–phenotype/disease association. This spreadsheet-based workflow was applied to both the CGC and MSSNG samples. For the MSSNG samples, an additional column indicates whether the CNV was considered as "high quality rare".

The spreadsheet was reviewed by an analyst with an advanced degree in genetics, tasked with identifying potentially reportable CNVs. Filters, typically a combination of variant quality and gene constraints, were used for CNV prioritization. Regions that harbour candidates (putative CNVs that appears to be reportable based on the information in the spreadsheet, if their variant qualities are satisfactory, i.e., if they are true positive CNV calls) were visualized in IGV to assess read depth and anomalous reads. This process is time-consuming, as loading read alignments may take a long time (especially for large CNVs) and additional efforts may be required to inspect anomalous reads (particularly split reads). The analyst would consult various online tools to confirm the information in the spreadsheet and/or gather additional information. Classification was based on the synthesis of all available information, considering the patient's clinical manifestation(s).

For the CGC samples, the CNVs identified by the analysts were further reviewed by a panel of clinical and molecular geneticists and genetic counsellors. For the MSSNG samples, the CNVs identified by the analysts were further reviewed by at least three clinical genetics experts with advanced degrees and/or postgraduate experience in human genetics.

Cohorts used for the evaluation of SCIP

The CGC samples ($n = 316$ families) were sequenced at the Cardiac Genome Clinic of The Hospital for Sick Children for congenital cardiovascular diseases, primarily congenital heart defects. Participant recruitment and genome analysis procedures were described previously (Reuter et al. 2020). Briefly, the samples were sequenced on the Illumina HiSeq X or NovaSeq 6000 platforms at The Centre for Applied Genomics (TCAG) in Toronto, Canada to generate 150-bp paired-end reads at ≥ 30 × coverage. Reads were mapped to the GRCh37/hg19 reference genome using BWA (Li and Durbin 2009). CNVs were identified using ERDS (Zhu et al. 2012) and CNVnator (Abyzov et al. 2011) calls, with a window size of 500 bp. High-quality CNVs were identified as those detected by both methods with > 50% overlap (Trost et al. 2018). Manta was also used to identify CNVs based on anomalous reads (Chen et al. 2016).

We also analyzed WGS data from the MSSNG Project, which contains nearly 3000 families sequenced for autism spectrum disorder (Trost et al. 2022). These samples were analyzed by The Centre for Applied Genomics at The Hospital for Sick Children, aligned to GRCh38/hg38. CNVs were detected using the algorithms ERDS (Zhu et al. 2012) and CNVnator (Abyzov et al. 2011) based on a previously described workflow (Trost et al. 2018). A CNV was deemed rare if its frequency was < 1% in MSSNG parents and in 1000 Genomes Project population controls according to both algorithms. A CNV was deemed to be "high quality rare" if it was rare, was detected by both ERDS and CNVnator with at least 50% reciprocal overlap, and less than 70% of the CNV overlapped assembly gaps, centromeres, and segmental duplications. CNVs underwent at least three rounds of manual curation by experienced scientists to identify P/LP CNVs and reportable VUS that may be responsible for autism spectrum disorder. As families with no reportable CNV add little value in evaluating SCIP, we randomly excluded about 2/3 of such families from this study, resulting in a collection of 711 families.

CNVs in both cohorts were extensively curated. In both the CGC and MSSNG cohorts, P/LP CNVs were identified using the spreadsheet-based manual approach described above. The CGC and MSSNG cohorts collectively form the cohort used to evaluate SCIP. For this study, all previously identified P/LP CNVs were re-interpreted by an analyst using the ACMG/ClinGen guidelines (Yuen et al 2017; Riggs et al. 2020), ensuring that they met the threshold of LP (0.90 points). CNVs not meeting the threshold were downgraded to VUS.

This highly diverse and expertly curated cohort included 1,027 families, 316 aligned to hg19 and 711 aligned to hg38 (Figure S2a), among which 174 had one or more P/LP CNVs (Figure S2b). The P/LP CNVs were diverse: 121 deletions and 66 duplications (Table S1), size ranging from 2.51 kb to 77.01 Mb, a mixture of de novo and inherited variants, covered several recurrent regions (e.g., distal 1q21.1, 16p11.2, 22q11.2), and reached P/LP by different ACMG/ClinGen rule combinations (Figures S2c, S2d, and S2e).

Comparison with AnnotSV

We also compared the performance of SCIP with a recently published CNV/SV prioritization tool, AnnotSV (Geoffroy et al. 2021). We selected 15 ASD cases with a good diversity of P/LP CNV type and size, including two cases with no P/LP CNV. For each case, we uploaded the full list of CNVs (the same list provided to the SCIP backend) to the AnnotSV web server (https://lbgi.fr/AnnotSV/runjob). The svtBEDcol option was set to 4. For phenotype-driven analysis, we used the HPO term HP:0000717 (autism). All other options were kept at default. We did not upload the optional SNV VCF file, as we found that it had little, if any, effect on prioritization. CNVs with a score of 4 or 5 were considered prioritized. Performance was measured by the number of prioritized CNVs requiring manual review, as well as the rank of the P/LP CNV (if any) among the reviewable CNVs.

Results

Implementation and software architecture of SCIP

SCIP is composed of three modules: Variant Filtration, Prioritization, and Visualization. The Variant Filtration and Prioritization Modules together form the SCIP backend, while the Visualization Module is the frontend (Fig. 1). The backend was implemented using Perl and R, while the

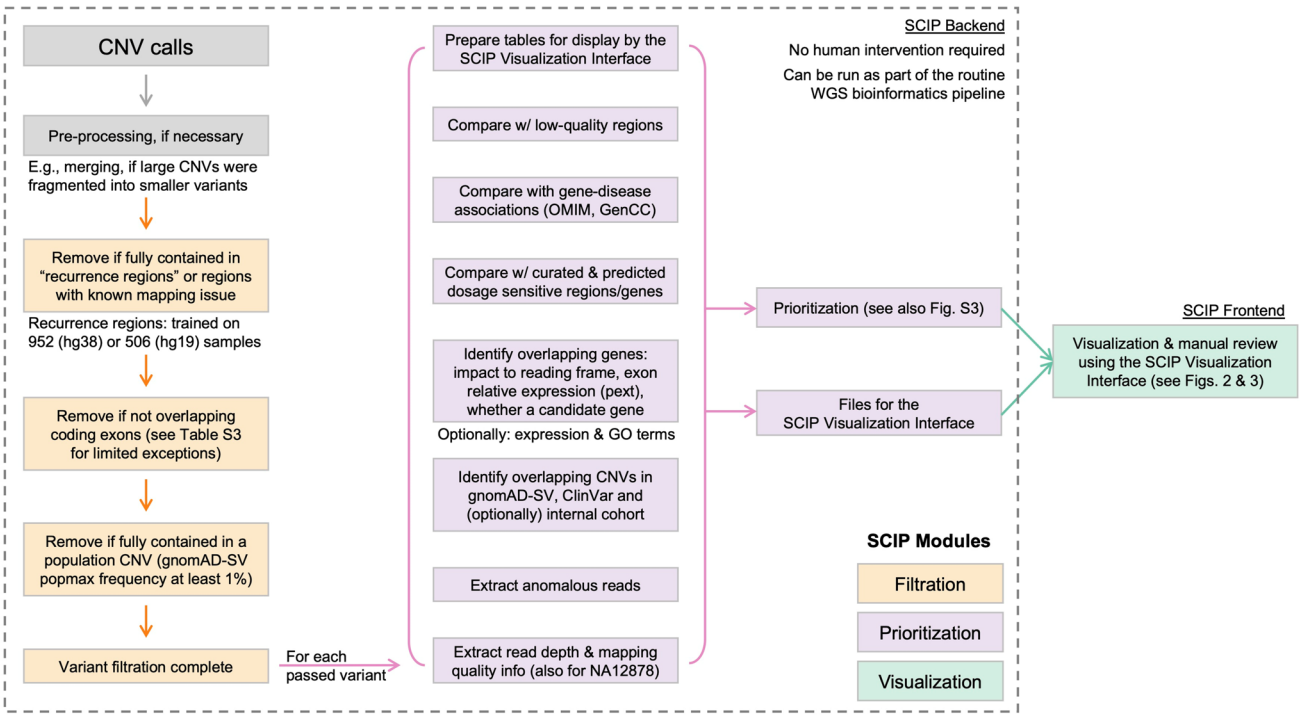

Fig. 1 Overall Software Architecture of SCIP. SCIP is composed of three modules. The Variant Filtration and Prioritization Modules collectively form the SCIP backend (within the dotted rectangle). CNV calls, after necessary pre-processing (e.g., merging), pass through the three-step Variant Filtration Module (orange). The remaining variants are then analyzed by the Prioritization Module, which calculates a priority score (Figure S3) and generates files for the Visualization Module. User may also opt to perform their own filtering and skip the SCIP Variant Filtration Module

frontend was implemented as a R Shiny application. The SCIP backend can be implemented on a high-performance computing server or a personal computer, while the Visualization Module is implemented on the analyst's personal computer. Files generated by the Prioritization Module must be available to the Visualization Module.

For a single sample, the minimum input file requirements are a list of CNVs and a BAM/CRAM alignment file. Optionally, if genome sequencing has been performed on multiple related individuals, CNVs detected in those individuals may be used as input for the purpose of inheritance analysis. SCIP requires annotations from external databases and the alignment file of a reference sample (e.g., NA12878, a widely used control sample (Zook et al. 2016)) (Table S3). Variants passing Filtration are processed by the Prioritization Module, then visualized in the Visualization Module, sorted by priority and size. SCIP is compatible with both hg19 and hg38 genome builds. SCIP is available on GitHub (https://github.com/qd29/SCIP). The SCIP GitHub webpage also includes step-by-step instructions (also see Supplementary Texts 1 and 2) and a demo of the SCIP Visualization Module. To further improve usability, a video tutorial series covering the setup and usage of SCIP is on YouTube at https://bit.ly/SCIPVideos.

SCIP variant filtration and prioritization modules

CNVs are first processed by the SCIP Variant Filtration and Prioritization Modules (collectively, the SCIP backend; Fig. 1). While CNVs from all callers (based on read depth or anomalous reads) are acceptable, interval merging is necessary if large CNVs were broken into fragments (e.g., by gaps), as SCIP may not be able to handle substantial undercalling (see Discussion and Supplementary Materials).

Variant filtration has three steps for optimal efficiency. In the first step, CNVs fully contained in regions with known issues (gaps, centromeres, repeats) and recurrence regions (for details, see Supplementary Materials) are removed. In the second step, CNVs that overlap coding sequences are retained, in addition to CNVs that overlap genes with clinically relevant non-coding variation (Table S2). Finally, CNVs are removed if they are fully contained in population variations (i.e., same type of CNV seen in > 1% frequency in a gnomAD-SV population). The remaining CNVs are passed to the Prioritization Module. If users opt to perform their own filtration, they may start directly with the Prioritization Module.

The Prioritization Module (Table S3) annotates and calculates a priority score for each CNV. For a given CNV, the following types of annotations are generated by the Prioritization Module: (1) overlapping CNVs in gnomAD-SV (including their allele frequencies), (2) overlapping CNVs in ClinVar (including their pathogenicity interpretations, associated conditions, allele origins, and gene contents), (3) overlapping CNVs in the internal cohort (if provided), (4) overlapping ClinGen dosage sensitive regions and genes, and (5) overlapping genes (including whether the overlap is full or partial, strand information, associated conditions provided by OMIM and GenCC, gnomAD gene constraints, exons and transcripts affected by the CNV, and exon-level relative expression data [i.e., pext scores]).

It is important to note that some annotation files used by the Prioritization Module, specifically (1) ClinGen dosage sensitivity curations, (2) ClinVar CNV information, (3) OMIM and GenCC gene-disease associations, and (4) the non-coding pathogenic regions, require periodic updates. The current guidelines recommend updating items 1–3 at least quarterly (Austin-Tse et al. 2022), and update item 4 if new pathogenic non-coding regions are discovered. See Table S3 for additional instructions.

The priority score is the sum of two components: clinical relevance and adverse information. Lower scores denote higher priority. The clinical relevance score is based on whether the CNV overlaps any ClinGen-curated dosage sensitive region, genes with substantial loss-of-function constraints, and/or genes associated with genetic conditions. While the theoretical range of the clinical relevance score is from 1 to 99, currently we only use 1–5 and 99 for deletions, and 1–7 and 99 for duplications. The default clinical relevance score (for CNVs without known clinical relevance) is 99. In other words, a CNV with any clinical relevance will have a score between 1 and 5 for deletions, and 1 and 7 for duplications. For example, a deletion that fully contains a ClinGen-curated HI region has a clinical relevance score of 1. The adverse information score is based on whether there is evidence against the CNV being true and/or pathogenic. This score is binary: if adverse information exists, the score will be 100; otherwise, the score will be zero. A priority score below 99 (with clinical relevance and without adverse information, currently 1–5 for deletions and 1–7 for duplications) denotes high priority. A priority score of 99 (without clinical relevance or adverse information) denotes moderate priority. A priority score above 99 (any clinical relevance and with adverse information) denotes low priority. Only CNVs with high priority need further review. The scoring scheme was described in detail in Figure S3. The Prioritization Module also generates files to be used by the Visualization Module for rendering tables and plots. The SCIP backend is fully programmatic and can be seamlessly integrated into the clinical WGS bioinformatics pipeline.

Overview of the SCIP visualization module/interface

The SCIP Visualization Module is a web-based graphical user interface (GUI) developed using the R Shiny

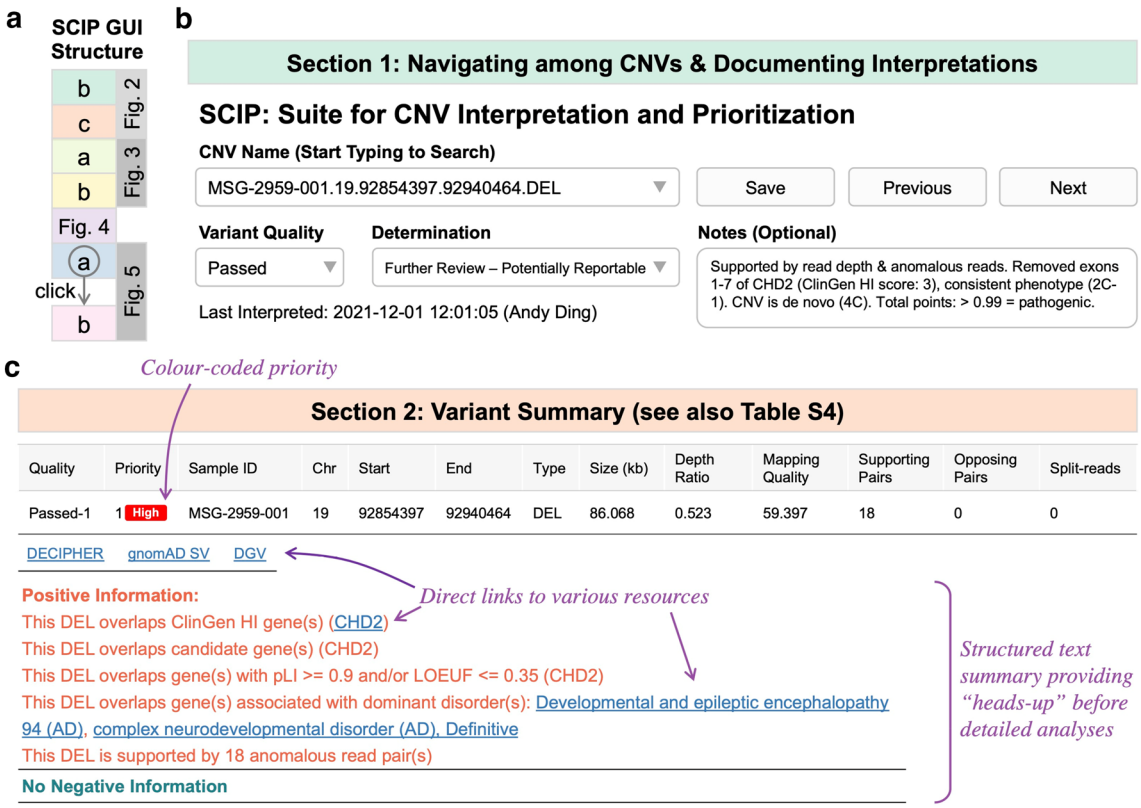

Fig. 2 SCIP Visualization Module, Part 1. **a** Schematic of the SCIP Visualization Module. This panel provides an overview of the SCIP Graphical User Interface (GUI, i.e., the Visualization Module). This panel illustrates that the SCIP Visualization Module displays multiple sections sequentially. The details of these sections are shown in additional figure panels (the names and colour codes of which are indicated). One of the sections (Sect. 6.1) of the SCIP Visualization Module is toggled by a mouse click (for more details, see Fig. 5), as indicated by the "click" wording in this panel. **b** Sect. 1 allows navigation through CNVs, either using the searchable CNV Name drop-down menu or the previous/next buttons. A user may view, modify, or enter interpretations. **c** The Variant Summary section offers an overview of a CNV, facilitating precision analysis. For this deletion, this section displays that it overlaps *CHD2*, a curated HI gene, and is well-supported by anomalous reads

package. Parameters are specified in a text configuration file (Table S5).

The GUI contains six sections (Fig. 2a), logically organized based on the manual workflow and the ACMG/ClinGen CNV interpretation guidelines (Riggs et al. 2020). The first section (Fig. 2b) allows the user to navigate through the CNVs in a sample, view previous interpretations (if any), and enter or modify interpretations. Interpretations are centrally stored in a text file.

The second section (Fig. 2c) provides an overview of a given CNV, intended to improve the efficiency of interpretation by highlighting key evidence. Priority calculated by the Prioritization Module is colour-coded (Figure S3). Structured sentences list positive (e.g., fully contains HI region) and negative (e.g., substantial overlap with population variants) findings (Table S4). In addition, the ratio of median read depth within vs. flanking the variant, median mapping quality within the variant, and a quality score (guidance only, Supplementary Materials) are displayed in the table.

The third and fourth sections facilitate variant quality assessment, as a more efficient alternative to IGV. The third section (Figs. 3a and S4) plots binned read depth and mapping quality of the CNV and its flanking regions (50% of variant size, minimum 100 kb, on both sides). All plots in the Visualization Module are interactive, allowing parameter adjustments and zooming in and out. By default, the plots have data from NA12878 overlaid (as a reference) but may be adjusted to display the clinical sample only. Common spurious read depth changes (e.g., regions with mapping issues) visible in the reference sample can be identified and excluded.

The fourth section (Figs. 3b and S5) displays colour-coded anomalous reads: read pairs with inward (normal) orientation but very large or small insert size (red), read pairs facing outward (blue) or in the same direction (cyan), and split reads (purple). Due to technical differences, the number of supporting reads plotted may be slightly lower than displayed in the second section. When zoomed-in, partial

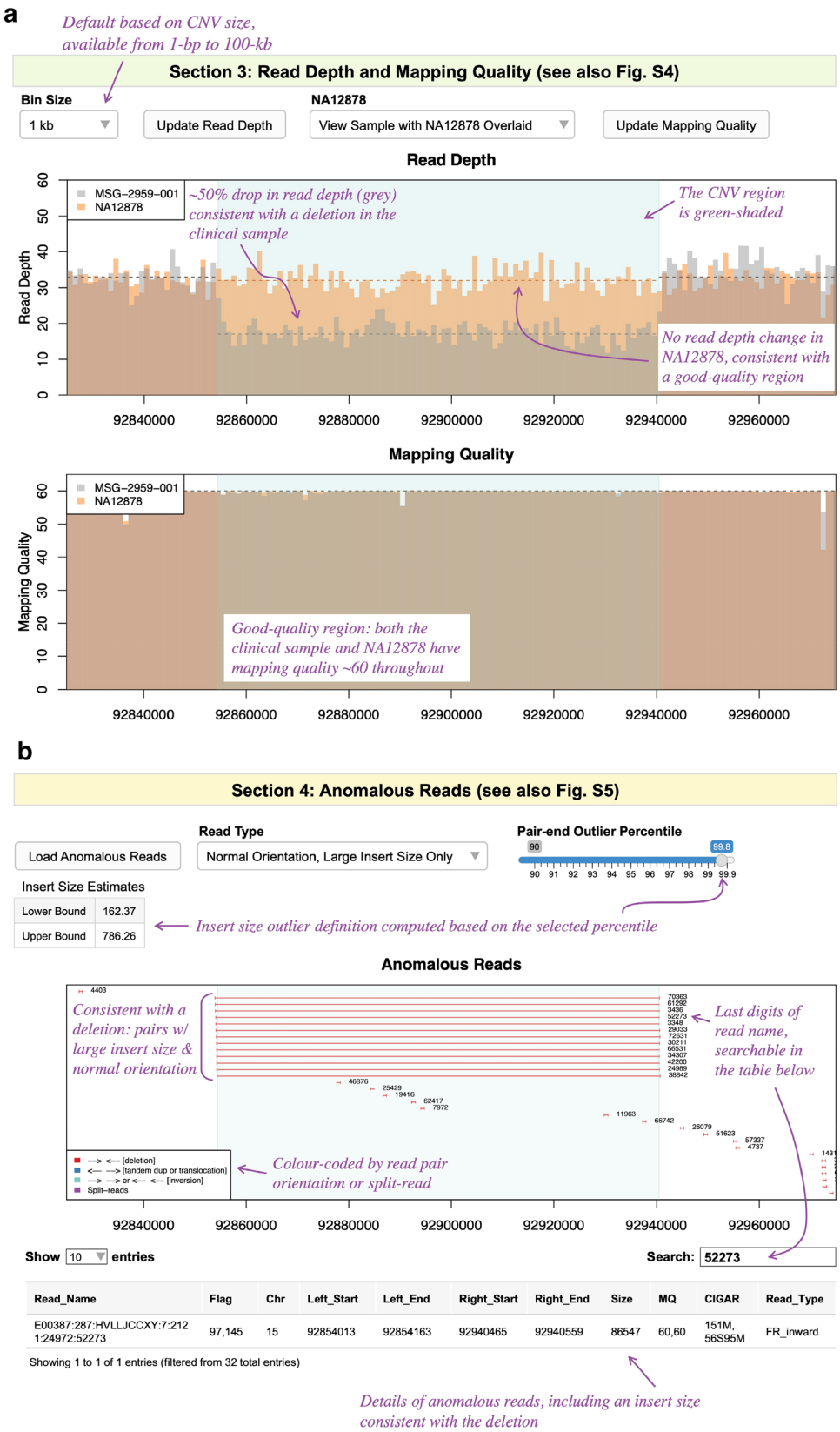

Fig. 3 SCIP Visualization Module, Part 2. **a** Sect. 3 plots binned read depth and mapping quality information. **b** Sect. 4 plots anomalous reads in the region flanking the variant. The upper/lower bound of insert size (for outlier detection) is set as 99.5/0.5-percentile of that of all read pairs by default and is adjustable. A user may view specific kinds of anomalous reads depending on the CNV type. For example, this plot only shows read pairs with normal orientation but unusually large insert size (expected for a deletion). Part of their names is displayed alongside the reads, which are searchable in the table below. Together with Sect. 3 (panel a), they offer a more efficient alternative to IGV in assessing CNV quality. Purple text in the figure highlights additional features of the interface

read names are displayed next to the reads, which are searchable in the accompanying table. This table shows details of anomalous reads, which is helpful when determining the exact CNV breakpoints. Using the Read Type drop-down menu, users may view a subset of anomalous reads specific to the variant type (Figures S1 and S5), e.g., inward pairs with large insert size when interpreting a deletion.

After a putative CNV is determined to be of satisfactory quality using Sects. 3 and 4, a user will then proceed to Sect. 5 (Figs. 4 and S6). This section visualizes the CNV-of-interest relative to known CNVs, allowing the identification of benign population variants (gnomAD-SV (Collins et al. 2020)) and recurrent pathogenic variants that have been interpreted previously (ClinVar (Landrum et al. 2018)). If the user provided CNV data from other family members, their CNVs will be coloured in red in the Internal Cohort panel, facilitating the study of inheritance patterns. Figure 4 shows a de novo variant (trio sequenced where the variant was absent in parents), and Figure S6 demonstrates a maternally inherited variant. When zoomed-in, IDs are displayed alongside the variants, which are searchable in the tables below that contain additional information (e.g., links to databases, gene content for ClinVar variants).

The sixth section (Figs. 5 and S7) provides the biological and clinical context of the CNV relative to genomic annotations. Tailored to the ACMG/ClinGen guidelines (Riggs et al. 2020), the following are plotted: dosage sensitivity curations (ClinGen) and constraints (gnomAD pLI (Karczewski et al. 2020)), coding exons (see Table S2 for limited exceptions), and relative exon expression (gnomAD pext (Cummings et al. 2020)). The pext score is helpful to evaluate the biological relevance of affected exons (Abou Tayoun et al. 2018). The accompanying tables contain a wealth of information, including direct links to databases and Google

Fig. 4 SCIP Visualization Module, Part 3. The External and Internal Variant Databases section compares the variant-of-interest with known CNVs, as well as the internal cohort (if provided). In the gnomAD-SV panel, names and popmax allele frequencies are displayed next to the variants. This is supplemented by a table with links to gnomAD-SV. For this CNV, no gnomAD-SV variants overlapping *CHD2* (purple box) were observed. Filtering of variants by popmax frequency is available. ClinVar variants are colour-coded by consequence (see legend) and may be filtered by consequence or size. The accompanying table displays gene content of the ClinVar variants (including whether full or partial overlap), allowing comparison with the CNV-of-interest. In the Internal Cohort panel, variants from the same family as the proband are coloured in red. There are no red-coloured similar-sized variants in this panel (despite trio sequenced), indicating that the variant is de novo

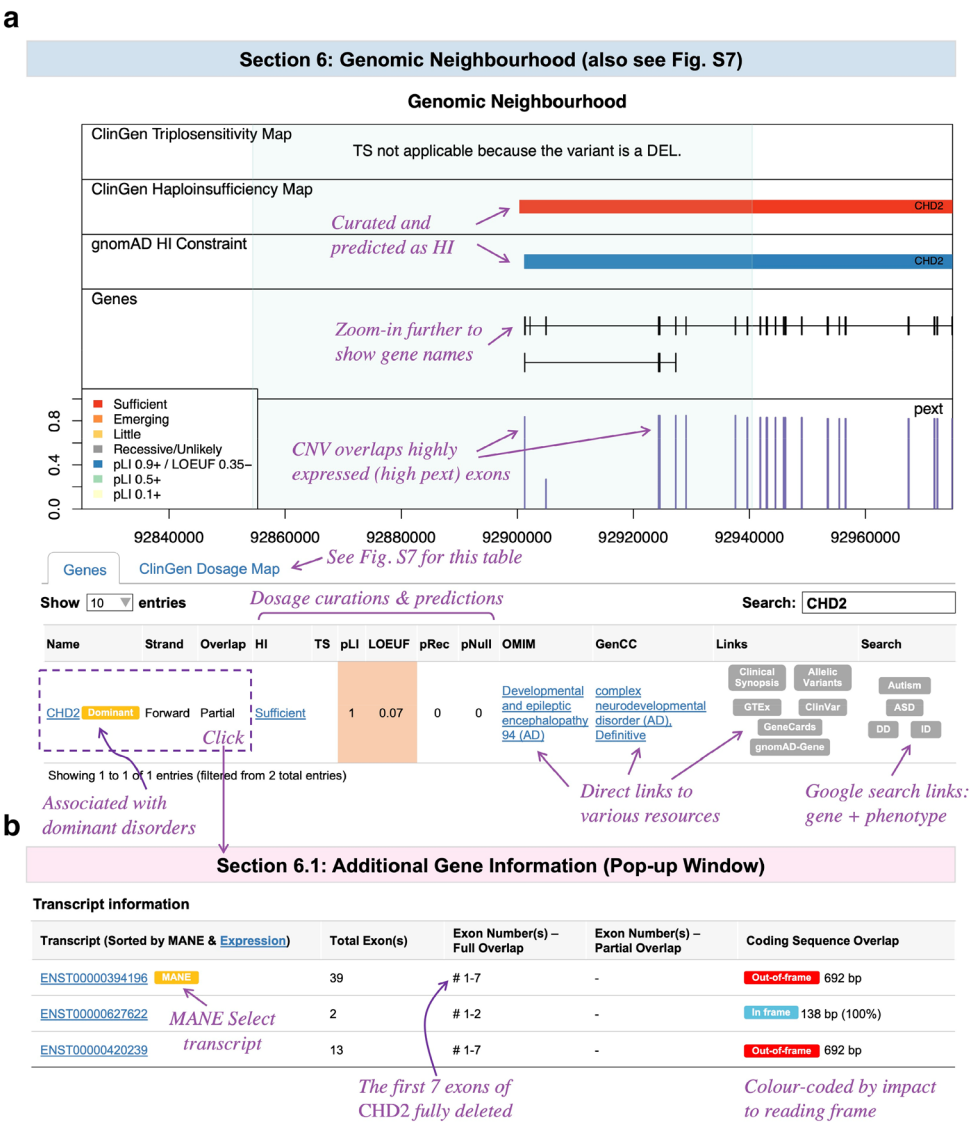

Fig. 5 SCIP Visualization Module, Part 4. **a** The Genomic Neighbourhood section plots dosage sensitivity curations and constraints, genes, and pext (relative exon expression) scores. Dosage information is colour-coded (see legend). The Genes table comprises a wealth of information, including links to external resources (e.g., OMIM clinical synopsis and allelic variants pages, GTEx, GeneCards, Google search terms). **b** The Transcript Information table shows exons in biologically relevant transcripts affected by the CNV. This pop-up table can be toggled by clicking one of the first three columns of the Genes table. The queried CNV removed exons 1–7 of the MANE Select transcript of *CHD2*, supporting its pathogenicity

search terms (gene + phenotype), to facilitate additional exploration of genes overlapping the CNV. Further, the first three columns of the Genes table function as clickable buttons that lead to a pop-up window containing affected exon information (Fig. 5b). This is helpful in investigating the impact of intragenic CNVs on reading frames.

Taken together, the SCIP Visualization Module offers a logically organized workflow, guiding users through a thorough clinical interpretation of CNVs—quality assessment, relationship to known benign and pathogenic variants, inheritance, and genomic context. Taking advantage of being web-based, the interface also provides many direct links, e.g., to DGV, OMIM, GeneCards, and Google. These links allow users to explore databases rapidly and seamlessly, saving time and reducing errors. A video walkthrough of the SCIP Visualization Module can be found on YouTube at https://bit.ly/SCIPVideos (video #3).

To further demonstrate how SCIP facilitates efficient identification of clinically reportable CNVs and exclusion of non-relevant putative CNVs, we presented four typical use cases of SCIP in Supplementary Text 3 (a pathogenic deletion, a pathogenic duplication, a population variant, and a CNV not affecting biologically relevant transcript).

Computational performance of the SCIP backend

We evaluated the computational burden of the SCIP backend using a single core on a server with a 2.3-GHz Skylake Intel Xeon processer. Variant filtration took 0.18 min per sample (i.e., 0.58 min per 10,000 CNVs), while the Prioritization Module used a median of 10.47 min (IQR: 10.06, range 0.52–136.77) per sample. Files generated for the Visualization Module occupied a median of 64.90 MB (IQR: 64.67, range 4.04–1039.49) of storage per sample. While RAM

usage was not monitored, 8 GB was adequate for the Variant Filtration Module and most CNVs processed by the Prioritization Module (although CNVs > 10 Mb may require more than 8 GB of RAM). These results indicate that the SCIP backend had a minimal computational burden. Furthermore, the Variant Filtration (by chromosome) and Prioritization Modules (by CNV) support parallelization, allowing speed improvements when resources permit.

SCIP was efficient at variant filtration and prioritization

We next evaluated SCIP using a large collection of clinical WGS samples. We assembled this collection from two sources at The Hospital for Sick Children: a cohort of patients with cardiovascular anomalies (primarily congenital heart defects) from the Cardiac Genome Clinic (CGC, $n = 316$ families) (Reuter et al. 2020), and a cohort of patients with autism spectrum disorder from the MSSNG Project and analyzed at The Centre for Applied Genomics (TCAG) ($n = 711$ families) (Trost et al. 2022; Yuen et al. 2016), for a total of 1027 families. Because some families had multiple sequenced siblings, the 1027 families harboured 1188 non-parental WGS samples. Before filtering, they contained a median of 3,222 CNVs (inter-quartile range [IQR]: 930.75, range: 816–17,586). After the application of the Variant Filtration Module, a median of 12 CNVs per sample remained (IQR: 8, range: 1–84), reflecting > 99.5% in reduction (Fig. 6a).

The remaining CNVs were processed through the SCIP Prioritization Module, by which only a median of two CNVs per sample (IQR: 3, range: 0–40) were classified as high priority (score less than 99, Figure S3) requiring manual review. This represents an 81% further reduction (Fig. 6a). More than 13% (159/1,188) of the samples had no high priority CNVs (Fig. 6b). Taken together, the SCIP Variant Filtration and Prioritization Modules were highly efficient at CNV filtration, with < 0.1% of all CNVs requiring manual review.

SCIP-prioritized CNVs captured all previously identified P/LP findings

We then focused on the 174 families (183 non-parental samples) with previously identified P/LP CNV findings (Figure S2). Reassuringly, all 187 previously identified P/LP CNVs

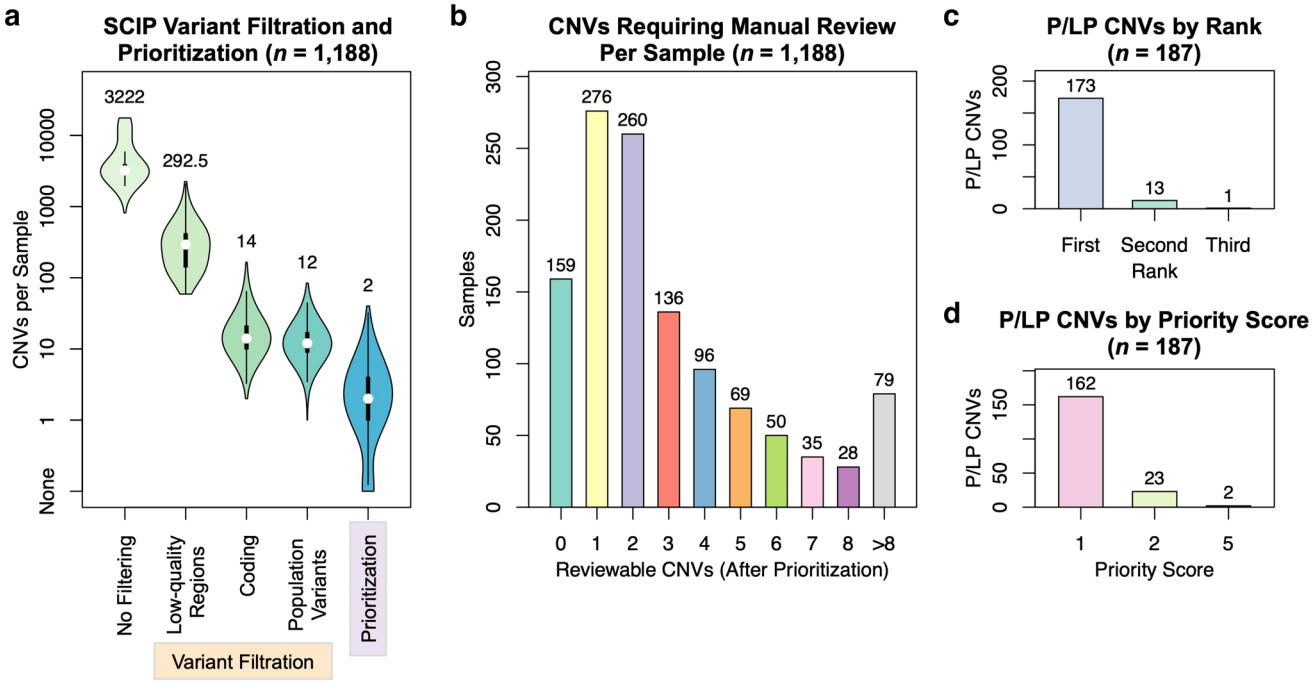

Fig. 6 SCIP was Highly Efficient at Filtration and Prioritization of P/LP CNVs. **a** The SCIP Variant Filtration and Prioritization Modules were effective, reducing the median number of variants per sample from 3222 (pre-filtering) to two (after prioritization). CNVs remaining per sample after each step of variant filtration are also plotted. **b** Distribution of the number of variants requiring manual review per sample. The majority (695/1188) of the samples had two or fewer reviewable CNVs, while nearly 95% (1109) had no more than eight reviewable CNVs. **c** SCIP further prioritized P/LP variants among reviewable CNVs, with 92.5% of them ranked first in the respective sample. CNVs were ranked by priority score and size. **d**All but two previously identified P/LP CNVs had priority scores of 1 or 2 (as determined by the SCIP Prioritization Module). While we currently select variants with priority scores < 99 for manual review, this finding indicates that further efficiencies in selecting reviewable CNVs may be possible

were classified by SCIP as high priority for manual review, indicating that SCIP was non-inferior in sensitivity than the spreadsheet-based workflow at identifying P/LP CNVs (Materials and Methods). SCIP further prioritized the P/LP findings among reviewable CNVs: 92.5% (173/187) of them ranked first in the respective sample (Fig. 6c; $p = 6.37 \times 10^{-26}$, one-tailed Wilcoxon rank-sum test). In summary, SCIP was effective at prioritizing P/LP CNVs.

SCIP had higher sensitivity than the manual workflow

Given the complexity of the manual workflow, we hypothesized that some P/LP CNVs may have remained undetected. Therefore, we re-interpreted CNVs in all 1027 families (Figure S2b) using SCIP. We identified an additional 835.20-kb pathogenic deletion at 15q11.2 (BP1–BP2) that fully contained HI region ISCA-37448, in a patient with autism spectrum disorder. This ISCA region was recurrent and previously reported in patients with autism. This result suggests that SCIP, benefitting from efficient variant filtration and user-friendly visualization, is more sensitive than the spreadsheet-based manual workflow in identifying P/LP CNVs.

SCIP substantially reduced time burden of clinical CNV interpretation

We performed a blinded, timed, head-to-head comparison (SCIP vs. the spreadsheet-based manual workflow) of CNV interpretation with 15 ASD cases (Table S6). They were selected for good representation of CNV types and sizes (including no CNV finding) and randomized. Two analysts experienced in SCIP (1 year of experience) and the manual workflow (1.75 years of experience), respectively, blinded to case selection, was tasked with identifying reportable findings and timing the analyses using the corresponding approach. Each case was assigned different IDs for SCIP and the manual workflow.

The analysts were able to identify all reportable findings, or lack thereof, using either approach. However, SCIP was substantially faster (Fig. 7a): it took SCIP a median of 2 min 21 s (IQR: 90.5, range: 23–268 s) per case, corresponding to an 80.7%-reduction (median, IQR: 12.2, range: 64.3–90.0) in time burden compared with the manual workflow. This was statistically significant ($p = 3.05 \times 10^{-5}$, paired one-tailed Wilcoxon rank-sum test). In addition, SCIP was consistently faster across diverse scenarios (Fig. 7b). Thus, SCIP achieved the designed goal of substantially reducing the time burden of CNV interpretation without compromising efficacy.

SCIP substantially outperformed AnnotSV in CNV prioritization

We then benchmarked SCIP against AnnotSV (Geoffroy et al. 2021), a recently published tool for annotation, prioritization, and tabular visualization (with the knotAnnotSV tool) of CNVs detected in clinical cases. Because AnnotSV has minimal support for variant quality assessment, we compared the performance of CNV prioritization between SCIP and AnnotSV. The 15 ASD cases selected for the above head-to-head comparison (with the spreadsheet-based workflow) were used in this analysis. Unfiltered lists of CNVs detected in these samples were provided to both SCIP and AnnotSV. We then compared the number of prioritized CNVs and rank of the P/LP variant (if any) among prioritized CNVs between the two tools.

SCIP significantly outperformed AnnotSV in both metrics (Fig. 7c). Among the 15 cases, a median of two CNVs per case were prioritized by SCIP, which was significantly less than the median of eight CNVs by AnnotSV ($p = 8.88 \times 10^{-4}$, paired one-tailed Wilcoxon rank-sum test). In addition, we found that in one case, AnnotSV failed to include the pathogenic variant among the prioritized CNVs. In the remaining cases, the median rank of the P/LP variant was 1 and 4.5 for SCIP and AnnotSV, respectively ($p = 6.66 \times 10^{-3}$, paired one-tailed Wilcoxon rank-sum test). Taken together, these findings indicate that SCIP was substantially superior to AnnotSV in prioritizing P/LP CNVs in clinical cases.

Discussion

SCIP is substantially better than the manual workflow and published tools

SCIP streamlines the workflow for clinical interpretation of CNVs. We rigorously evaluated SCIP using more than 1,000 families containing nearly 200 P/LP CNVs. It is noteworthy that this cohort size and the number of P/LP CNVs are almost unprecedented. Most tools were evaluated using a small number of P/LP CNVs (a few dozen or less), or only with benign or simulated variants (Belyeu et al. 2021; Minoche et al. 2021).

We revealed that SCIP was more sensitive than the spreadsheet-based manual workflow, not only selecting all previously identified P/LP CNVs for manual review, but also discovering one pathogenic CNV overlooked by previous curations. Meanwhile, CNV interpretation using SCIP was more than 60% faster per case than the manual workflow. We also showed that SCIP was significantly more effective than AnnotSV (Geoffroy et al. 2021) in CNV prioritization. These results convincingly indicate that SCIP is substantially better than previous workflows and tools.

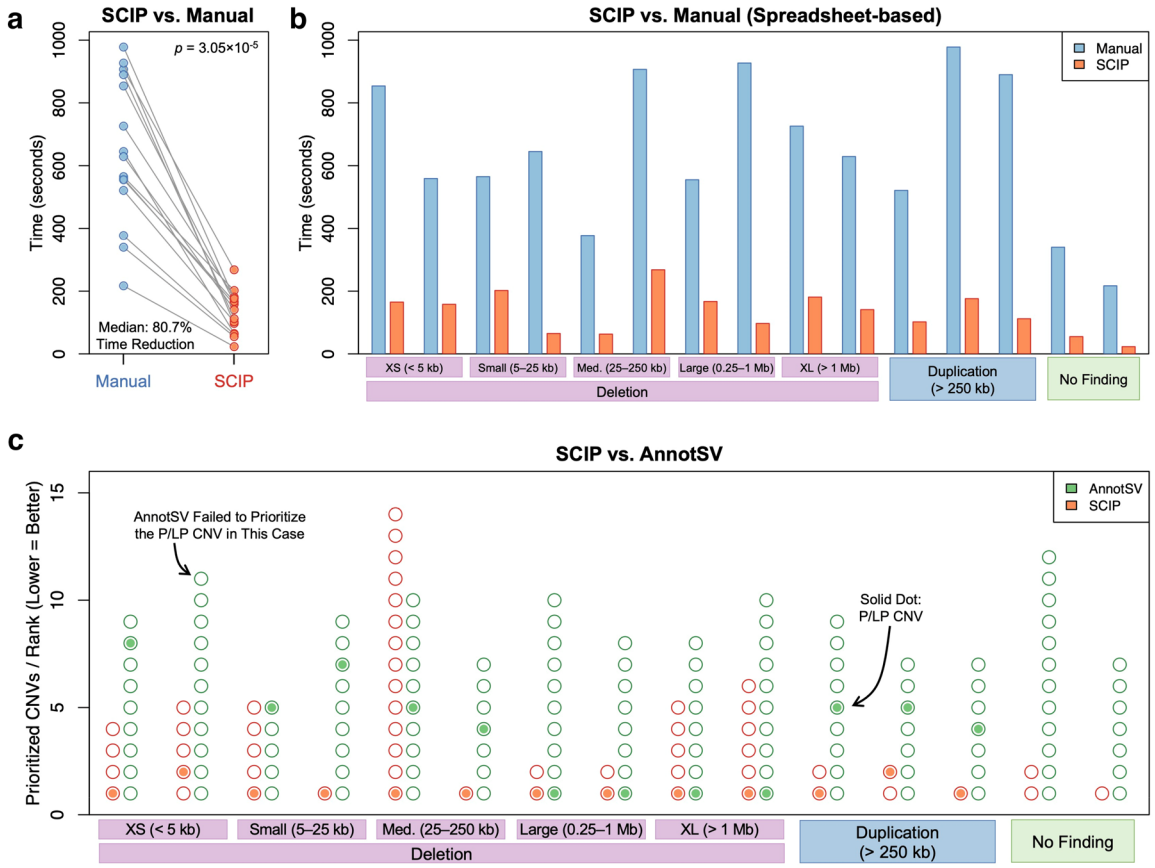

Fig. 7 SCIP was Substantially Superior to Previous Approaches for CNV Interpretation. **a, b** In a head-to-head comparison with the spreadsheet-based manual workflow using 15 samples sequenced for autism spectrum disorder, SCIP was 80.7% (median) faster than the manual approach currently used by the CGC and TCAG at The Hospital for Sick Children. **b** The observed time savings of SCIP was consistent across a diverse range of scenarios, including deletions of varying sizes ($n=2$ for each category), duplications ($n=3$), and two cases with no reportable CNV findings. **c** SCIP was statistically significantly more effective at CNV prioritization than AnnotSV. Each case is represented by two columns of circles (one orange and one green). Circles indicate prioritized CNVs, while solid dots indicate P/LP CNVs. Rank of the P/LP variant among prioritized CNVs can be inferred using the Y-axis. Compared with AnnotSV, SCIP had significantly lower number of prioritized CNVs per case and better ranking for P/LP CNVs

All components of SCIP contributed to its performance. The Variant Filtration and Prioritization Modules efficiently selected < 0.1% of all variants for manual review. While the Visualization Module provided an integrated interface displaying most, if not all, information needed for CNV interpretation, keeping the back-and-forth among multiple tools to a minimum. When working in unison, fewer variants will be presented to an analyst in a more user-friendly manner, resulting in time reduction without sacrificing sensitivity.

SCIP has superior visualization capabilities compared to previously published tools, such as ClinSV (Minoche et al. 2021), samplot (Belyeu et al. 2021), CNVxplorer (Requena et al. 2021), and AnnotSV/knotAnnotSV (Geoffroy et al. 2021). Most importantly, none of these tools provide a feature analogous to Sect. 2 of the SCIP Visualization Module, i.e., summarizing evidence that may support or refute CNV quality and/or pathogenicity. The summary feature within the SCIP Visualization Module is useful in quickly showing analysts where to focus their attention during the detailed review. For example, an analyst might need to spend more time on quality assessment when alerted that the CNV was not supported by anomalous reads. Further, being a web-based tool, SCIP provides many direct links to external resources. Most other tools are unable to do so as they are not web-based. We found this feature greatly reduced time burden and error. For example, using SCIP, only one click is required to view the CNV region in the gnomAD browser, while three separate steps are needed otherwise (open the main page, enter the coordinates [error-prone], then click search). In addition, while some tools are good at variant quality assessment, and others are good at visualizing biological context, none are ideal for both. SCIP, in contrast, displays all information, including variant quality and biological context, in one unified interface.

Limitations and future developments

Improvements may be possible in selecting CNVs for manual review. CNVs with priority scores below 99 require manual review, while nearly 99% (185/187) of the P/LP CNVs had priority score 1 or 2 (Fig. 6d). Further refinements may reduce the number of reviewable CNVs per case while maintaining sensitivity. A small subset (6.6%) of samples analyzed in this study had relatively larger numbers (> 8) of reviewable CNVs. We found that the SCIP Filtration Module had lower efficiency for these samples, the root cause of which remains to be determined. In addition, SCIP currently does not use patient phenotypes in prioritization. Incorporating patient phenotypes may further prioritize P/LP variants among the reviewable CNVs.

While most known pathogenic CNVs directly impact protein-coding genes, emerging evidence reveals that non-coding CNVs, particularly those overlapping regulatory elements, may also be causative for Mendelian disease (Flöttmann et al. 2018). SCIP currently has limited capability in analyzing non-coding CNVs. This limitation is inherent to the incomplete understanding of clinical relevance of non-coding CNVs. For well-established non-coding pathogenic regions, SCIP uses an exception list (Table S2), i.e., essentially treating them as coding. We suggest updating the exception list if new pathogenic non-coding regions are discovered in the future. Updated lists will be posted on the SCIP GitHub site. We plan to accommodate non-coding CNVs when guidelines become available in the future.

SCIP is designed for CNVs detected by WGS for constitutional genetic disorders, therefore its applicability in other scenarios may be limited (e.g., somatic CNVs in cancer or CNVs detected on exome sequencing or gene panels). Furthermore, SCIP is not intended for the visualization of full-chromosome aneuploidies. SCIP had been primarily tested in patients with congenital cardiovascular disease and/or ASD, and we plan to further evaluate SCIP in patients with other rare genetic disorders. Additionally, SCIP does not yet support the identification of compound heterozygous variants involving both CNV and SNV. This functionality will be incorporated in a forthcoming sister tool of SCIP for the clinical interpretation of SNVs.

Finally, SCIP may be unable to handle substantial CNV under-calling (identifying variants smaller than their actual sizes). This typically results from the fragmentation of a large (0.5–1 + Mb) variant, with the CNV caller identifying it as multiple discrete but nearby smaller variants. To avoid this issue, merging nearby CNVs before using SCIP is recommended (see the "Pre-processing: Merging Under-called CNVs" section in Supplementary Materials for details). We did not encounter any issue with under-calling using this approach, despite more than 50 P/LP CNVs being initially fragmented.

Conclusions

We designed and implemented SCIP, a tool that effectively reduces the complexity and improves the efficiency of clinical CNV interpretation. SCIP was evaluated on an unparalleled cohort of more than 1000 WGS samples containing nearly 200 P/LP CNVs. SCIP had superior performance than previous workflows and tools. SCIP is fully available for implementation in clinical diagnostic laboratories, and we are confident that it will substantially improve clinical CNV interpretation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-022-02494-1.

Acknowledgements The authors wish to acknowledge the resources of MSSNG (www.mss.ng), Autism Speaks and The Centre for Applied Genomics at The Hospital for Sick Children, Toronto, Canada. We thank the generosity of the donors who supported the MSSNG program. We also thank all participating families for their time and contributions. We thank Dr. James Stavropoulos for feedback on the manuscript.

Author contributions Conceptualization: QD and CS; methodology: QD, CS and RM; software: QD; formal analysis: QD and CS; investigation: QD, CS, RM, BT, MSR, KK, KS, JBO, SMH, EL, MC, MZ, EJH, AJC and WE; writing—original draft preparation: QD; writing—review and editing: QD, CS, RM, BT, KK, KS, JBO, SMH, EL, MC, MZ, AJC, WE, SWS and RKJ; visualization: QD; supervision: SWS, RHK and RKJ; project administration: KK, KS, EL and MC; funding acquisition: RHK and RKJ.

Funding This work was funded by generous support from the Ted Rogers Centre for Heart Research at The Hospital for Sick Children.

Declarations

Conflict of interest None declared.

Ethical approval Data used in this study were obtained as part of research protocols approved by the Research Ethics Boards at The Hospital for Sick Children (1000053844, 0019980189), the University Health Network (16–6282), and the Mount Sinai Hospital (19–0320-E).

References

Abou Tayoun AN, Pesaran T, DiStefano MT, Oza A, Rehm HL, Biesecker LG, Harrison SM, Working CSVI, G, (2018) Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Hum Mutat 39:1517–1524. https://doi.org/10.1002/humu.23626

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21:974–984. https://doi.org/10.1101/gr.114876.110

Amberger JS, Bocchini CA, Scott AF, Hamosh A (2019) OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res 47:D1038–D1043. https://doi.org/10.1093/nar/gky1151

Austin-Tse CA, Jobanputra V, Perry DL, Bick D, Taft RJ, Venner E, Gibbs RA, Young T, Barnett S, Belmont JW, Boczek N, Chowdhury S, Ellsworth KA, Guha S, Kulkarni S, Marcou C, Meng L, Murdock DR, Rehman AU, Spiteri E, Thomas-Wilson A, Kearney HM, Rehm HL (2022) Best practices for the interpretation and reporting of clinical whole genome sequencing. NPJ Genom Med 7:27. https://doi.org/10.1038/s41525-022-00295-z

Belyeu JR, Chowdhury M, Brown J, Pedersen BS, Cormier MJ, Quinlan AR, Layer RM (2021) Samplot: a platform for structural variant visual validation and automated filtering. Genome Biol 22:161–161. https://doi.org/10.1186/s13059-021-02380-5

Cerruti Mainardi P (2006) Cri du Chat syndrome. Orphanet J Rare Dis 1:33. https://doi.org/10.1186/1750-1172-1-33

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32:1220–1222. https://doi.org/10.1093/bioinformatics/btv710

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, Watts NA, Solomonson M, O'Donnell-Luria A, Baumann A, Munshi R, Walker M, Whelan CW, Huang Y, Brookings T, Sharpe T, Stone MR, Valkanas E, Fu J, Tiao G, Laricchia KM, Ruano-Rubio V, Stevens C, Gupta N, Cusick C, Margolin L, Genome Aggregation Database Production T, Genome Aggregation Database C, Taylor KD, Lin HJ, Rich SS, Post WS, Chen Y-DI, Rotter JI, Nusbaum C, Philippakis A, Lander E, Gabriel S, Neale BM, Kathiresan S, Daly MJ, Banks E, MacArthur DG, Talkowski ME (2020) A structural variation reference for medical and population genetics. Nature 581:444–451. https://doi.org/10.1038/s41586-020-2287-8

Costa CIS, da Silva Montenegro EM, Zarrei M, de Sá ME, Silva IMW, de Oliveira SM, Wang JYT, Zachi EC, Branco EV, da Costa SS, Lourenço NCV, Vianna-Morgante AM, Rosenberg C, Krepischi ACV, Scherer SW, Passos-Bueno MR (2022) Copy number variations in a Brazilian cohort with autism spectrum disorders highlight the contribution of cell adhesion genes. Clin Genet 101:134–141. https://doi.org/10.1111/cge.14072

Cummings BB, Karczewski KJ, Kosmicki JA, Seaby EG, Watts NA, Singer-Berk M, Mudge JM, Karjalainen J, Satterstrom FK, O'Donnell-Luria AH, Poterba T, Seed C, Solomonson M, Alföldi J, Genome Aggregation Database Production T, Genome Aggregation Database C, Daly MJ, MacArthur DG (2020) Transcript expression-aware annotation improves rare variant interpretation. Nature 581:452–458. https://doi.org/10.1038/s41586-020-2329-2

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97. https://doi.org/10.1038/nrg1767

Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP (2009) DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. Am J Hum Genet 84:524–533. https://doi.org/10.1016/j.ajhg.2009.03.010

Flöttmann R, Kragesteen BK, Geuer S, Socha M, Allou L, Sowińska-Seidler A, Bosquillon de Jarcy L, Wagner J, Jamsheer A, Oehl-Jaschkowitz B, Wittler L, de Silva D, Kurth I, Maya I, Santos-Simarro F, Hülsemann W, Klopocki E, Mountford R, Fryer A, Borck G, Horn D, Lapunzina P, Wilson M, Mascrez B, Duboule D, Mundlos S, Spielmann M (2018) Noncoding copy-number variations are associated with congenital limb malformation. Genet Med 20:599–607. https://doi.org/10.1038/gim.2017.154

Geoffroy V, Guignard T, Kress A, Gaillard JB, Solli-Nowlan T, Schalk A, Gatinois V, Dollfus H, Scheidecker S, Muller J (2021) AnnotSV and knotAnnotSV: a web server for human structural variations annotations, ranking and analysis. Nucleic Acids Res 49:W21-w28. https://doi.org/10.1093/nar/gkab402

Gross AM, Ajay SS, Rajan V, Brown C, Bluske K, Burns NJ, Chawla A, Coffey AJ, Malhotra A, Scocchia A, Thorpe E, Dzidic N, Hovanes K, Sahoo T, Dolzhenko E, Lajoie B, Khouzam A, Chowdhury S, Belmont J, Roller E, Ivakhno S, Tanner S, McEachern J, Hambuch T, Eberle M, Hagelstrom RT, Bentley DR, Perry DL, Taft RJ (2019) Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. Genet Med: off J Am Coll Med Genet 21:1121–1130. https://doi.org/10.1038/s41436-018-0295-y

Handsaker RE, Korn JM, Nemesh J, McCarroll SA (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43:269–276. https://doi.org/10.1038/ng.768

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951. https://doi.org/10.1038/ng1416

Jiang Y-h, Yuen RKC, Jin X, Wang M, Chen N, Wu X, Ju J, Mei J, Shi Y, He M, Wang G, Liang J, Wang Z, Cao D, Carter MT, Chrysler C, Drmic IE, Howe JL, Lau L, Marshall CR, Merico D, Nalpathamkalam T, Thiruvahindrapuram B, Thompson A, Uddin M, Walker S, Luo J, Anagnostou E, Zwaigenbaum L, Ring RH, Wang J, Lajonchere C, Wang J, Shih A, Szatmari P, Yang H, Dawson G, Li Y, Scherer SW (2013) Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. Am J Hum Genet 93:249–263. https://doi.org/10.1016/j.ajhg.2013.06.012

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Genome Aggregation Database C, Neale BM, Daly MJ, MacArthur DG (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581:434–443. https://doi.org/10.1038/s41586-020-2308-7

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46:D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation C (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291. https://doi.org/10.1038/nature19057

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (oxford, England) 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM, Thiruvahindrapuram B, Merico D, Jobling R, Nalpathamkalam T, Pellecchia G, Sung WWL, Wang Z, Bikangaga P, Boelman C, Carter MT, Cordeiro D, Cytrynbaum C, Dell SD, Dhir P, Dowling JJ, Heon E, Hewson S, Hiraki L, Inbar-Feigenberg M, Klatt R, Kronick J, Laxer RM, Licht C, MacDonald H, Mercimek-Andrews S, Mendoza-Londono R, Piscione T, Schneider R, Schulze A, Silverman E, Siriwardena K, Snead OC, Sondheimer N, Sutherland J, Vincent A, Wasserman JD, Weksberg R, Shuman C, Carew C, Szego MJ, Hayeems RZ, Basran R, Stavropoulos DJ, Ray PN, Bowdin S, Meyn MS, Cohn RD, Scherer SW, Marshall CR (2018) Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med: off J Am Coll Med Genet 20:435–443. https://doi.org/10.1038/gim.2017.119

MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW (2014) The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 42:D986–D992. https://doi.org/10.1093/nar/gkt958

Manickam K, McClain MR, Demmer LA, Biswas S, Kearney HM, Malinowski J, Massingham LJ, Miller D, Yu TW, Hisama FM, Directors ABo (2021) Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American college of medical genetics and genomics (ACMG). Genet Med 23:2029–2037. https://doi.org/10.1038/s41436-021-01242-6

Markham JF, Yerneni S, Ryland GL, Leong HS, Fellowes A, Thompson ER, De Silva W, Kumar A, Lupat R, Li J, Ellul J, Fox S, Dickinson M, Papenfuss AT, Blombery P (2019) CNspector: a web-based tool for visualisation and clinical diagnosis of copy number variation from next generation sequencing. Sci Rep 9:6426–6426. https://doi.org/10.1038/s41598-019-42858-8

Marshall CR, Chowdhury S, Taft RJ, Lebo MS, Buchan JG, Harrison SM, Rowsey R, Klee EW, Liu P, Worthey EA, Jobanputra V, Dimmock D, Kearney HM, Bick D, Kulkarni S, Taylor SL, Belmont JW, Stavropoulos DJ, Lennon NJ, Medical Genome I (2020) Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. NPJ Genom Med 5:47–47. https://doi.org/10.1038/s41525-020-00154-9

McDonald-McGinn DM, Sullivan KE, Marino B, Philip N, Swillen A, Vorstman JAS, Zackai EH, Emanuel BS, Vermeesch JR, Morrow BE, Scambler PJ, Bassett AS (2015) 22q11.2 deletion syndrome. Nat Rev Dis Primers 1:15071. https://doi.org/10.1038/nrdp.2015.71

Minoche AE, Lundie B, Peters GB, Ohnesorg T, Pinese M, Thomas DM, Zankl A, Roscioli T, Schonrock N, Kummerfeld S, Burnett L, Dinger ME, Cowley MJ (2021) ClinSV: clinical grade structural and copy number variant detection from whole genome sequencing data. Genome Med 13:32–32. https://doi.org/10.1186/s13073-021-00841-x

NICUSeq Study Group, Krantz ID, Medne L, Weatherly JM, Wild KT, Biswas S, Devkota B, Hartman T, Brunelli L, Fishler KP, Abdul-Rahman O, Euteneuer JC, Hoover D, Dimmock D, Cleary J, Farnaes L, Knight J, Schwarz AJ, Vargas-Shiraishi OM, Wigby K, Zadeh N, Shinawi M, Wambach JA, Baldridge D, Cole FS, Wegner DJ, Urraca N, Holtrop S, Mostafavi R, Mroczkowski HJ, Pivnick EK, Ward JC, Talati A, Brown CW, Belmont JW, Ortega JL, Robinson KD, Brocklehurst WT, Perry DL, Ajay SS, Hagelstrom RT, Bennett M, Rajan V, Taft RJ (2021) Effect of whole-genome sequencing on the clinical management of acutely Ill infants with suspected genetic disease: a randomized clinical trial. JAMA Pediatr 175:1218–1226. https://doi.org/10.1001/jamapediatrics.2021.3496

Oskoui M, Gazzellone MJ, Thiruvahindrapuram B, Zarrei M, Andersen J, Wei J, Wang Z, Wintle RF, Marshall CR, Cohn RD, Weksberg R, Stavropoulos DJ, Fehlings D, Shevell MI, Scherer SW (2015) Clinically relevant copy number variations detected in cerebral palsy. Nat Commun 6:7949–7949. https://doi.org/10.1038/ncomms8949

Pereira E, Marion R (2018) Contiguous gene syndromes. Pediatr Rev 39:46–49. https://doi.org/10.1542/pir.2016-0073

Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS, ClinGen (2015) ClinGen–the clinical genome resource. N Engl J Med 372:2235–2242. https://doi.org/10.1056/NEJMsr1406261

Requena F, Abdallah HH, García A, Nitschké P, Romana S, Malan V, Rausell A (2021) CNVxplorer: a web tool to assist clinical interpretation of CNVs in rare disease patients. Nucleic Acids Res 49:W93–W103. https://doi.org/10.1093/nar/gkab347

Reuter MS, Chaturvedi RR, Liston E, Manshaei R, Aul RB, Bowdin S, Cohn I, Curtis M, Dhir P, Hayeems RZ, Hosseini SM, Khan R, Ly LG, Marshall CR, Mertens L, Okello JBA, Pereira SL, Raajkumar A, Seed M, Thiruvahindrapuram B, Scherer SW, Kim RH, Jobling RK (2020) The cardiac genome clinic: implementing genome sequencing in pediatric heart disease. Genet Med: off J Am Coll Med Genet 22:1015–1024. https://doi.org/10.1038/s41436-020-0757-x

Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, Raca G, Ritter DI, South ST, Thorland EC, Pineda-Alvarez D, Aradhya S, Martin CL (2020) Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American college of medical genetics and genomics (ACMG) and the clinical genome resource (ClinGen). Genet Med: off J Am Coll Med Genet 22:245–257. https://doi.org/10.1038/s41436-019-0686-8

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528. https://doi.org/10.1126/science.1098918

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192. https://doi.org/10.1093/bib/bbs017

Trost B, Walker S, Wang Z, Thiruvahindrapuram B, MacDonald JR, Sung WWL, Pereira SL, Whitney J, Chan AJS, Pellecchia G, Reuter MS, Lok S, Yuen RKC, Marshall CR, Merico D, Scherer SW (2018) A Comprehensive workflow for read depth-based

identification of copy-number variation from whole-genome sequence data. Am J Hum Genet 102:142–155. https://doi.org/10.1016/j.ajhg.2017.12.007

Trost B, Thiruvahindrapuram B, Chan AJS, Engchuan W, Higginbotham EJ, Howe JL, Loureiro LO, Reuter MS, Roshandel D, Whitney J, Zarrei M, Bookman M, Somerville C, Shaath R, Abdi M, Aliyev E, Patel RV, Nalpathamkalam T, Pellecchia G, Hamdan O, Kaur G, Wang Z, MacDonald JR, Wei J, Sung WWL, Lamoureux S, Hoang N, Selvanayagam T, Deflaux N, Geng M, Ghaffari S, Bates J, Young EJ, Ding Q, Shum C, D'abate L, Bradley CA, Rutherford A, Aguda V, Apresto B, Chen N, Desai S, Du X, Fong MLY, Pullenayegum S, Samler K, Wang T, Ho K, Paton T, Pereira SL, Herbrick J-A, Wintle RF, Fuerth J, Noppornpitak J, Ward H, Magee P, Baz AA, Kajendirarajah U, Kapadia S, Vlasblom J, Valluri M, Green J, Seifer V, Quirbach M, Rennie O, Kelley E, Masjedi N, Lord C, Szego MJ, MnH Z, Lang M, Strug LJ, Marshall CR, Costain G, Calli K, Iaboni A, Yusuf A, Ambrozewicz P, Gallagher L, Amaral DG, Brian J, Elsabbagh M, Georgiades S, Messinger DS, Ozonoff S, Sebat J, Sjaarda C, Smith IM, Szatmari P, Zwaigenbaum L, Kushki A, Frazier TW, Vorstman JAS, Fakhro KA, Fernandez BA, Lewis MES, Weksberg R, Fiume M, Yuen RKC, Anagnostou E et al (2022) Genomic architecture of autism spectrum disorder from comprehensive whole-genome sequence annotation. MedRxiv. https://doi.org/10.1101/2022.05.05.22274031

Yuen RKC, Merico D, Cao H, Pellecchia G, Alipanahi B, Thiruvahindrapuram B, Tong X, Sun Y, Cao D, Zhang T, Wu X, Jin X, Zhou Z, Liu X, Nalpathamkalam T, Walker S, Howe JL, Wang Z, MacDonald JR, Chan AJS, D'Abate L, Deneault E, Siu MT, Tammimies K, Uddin M, Zarrei M, Wang M, Li Y, Wang J, Wang J, Yang H, Bookman M, Bingham J, Gross SS, Loy D, Pletcher M, Marshall CR, Anagnostou E, Zwaigenbaum L, Weksberg R, Fernandez BA, Roberts W, Szatmari P, Glazer D, Frey BJ, Ring RH, Xu X, Scherer SW (2016) Genome-wide characteristics of de novo mutations in autism. NPJ Genom Med 1:16027. https://doi.org/10.1038/npjgenmed.2016.27

Yuen RKC, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel RV, Whitney J, Deflaux N, Bingham J, Wang Z, Pellecchia G, Buchanan JA, Walker S, Marshall CR, Uddin M, Zarrei M, Deneault E, D'Abete L, Chan AJS, Koyanagi S, Paton T, Pereira SL, Hoang N, Engchuan W, Higginbotham EJ, Ho K, Lamoureux S, Li W, MacDonald JR, Nalpathamkalam T, Sung WWL, Tsoi FJ, Wei J, Xu L, Tasse A-M, Kirby E, Van Etten W, Twigger S, Roberts W, Drmic I, Jilderda S, Modi BM, Kellam B, Szego M, Cytrynbaum C, Weksberg R, Zwaigenbaum L, Woodbury-Smith M, Brian J, Senman L, Iaboni A, Doyle-Thomas K, Thompson A, Chrysler C, Leef J, Savion-Lemieux T, Smith IM, Liu X, Nicolson R, Seifer V, Fedele A, Cook EH, Dager S, Estes A, Gallagher L, Malow BA, Parr JR, Spence SJ, Vorstman J, Frey BJ, Robinson JT, Strug LJ, Fernandez BA, Elsabbagh M, Carter MT, Hallmayer J, Knoppers BM, Anagnostou E, Szatmari P, Ring RH, Glazer D, Pletcher MT, Scherer SW (2017) Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci 20:602–611. https://doi.org/10.1038/nn.4524

Zarrei M, Burton CL, Engchuan W, Young EJ, Higginbotham EJ, MacDonald JR, Trost B, Chan AJS, Walker S, Lamoureux S, Heung T, Mojarad BA, Kellam B, Paton T, Faheem M, Miron K, Lu C, Wang T, Samler K, Wang X, Costain G, Hoang N, Pellecchia G, Wei J, Patel RV, Thiruvahindrapuram B, Roifman M, Merico D, Goodale T, Drmic I, Speevak M, Howe JL, Yuen RKC, Buchanan JA, Vorstman JAS, Marshall CR, Wintle RF, Rosenberg DR, Hanna GL, Woodbury-Smith M, Cytrynbaum C, Zwaigenbaum L, Elsabbagh M, Flanagan J, Fernandez BA, Carter MT, Szatmari P, Roberts W, Lerch J, Liu X, Nicolson R, Georgiades S, Weksberg R, Arnold PD, Bassett AS, Crosbie J, Schachar R, Stavropoulos DJ, Anagnostou E, Scherer SW (2019) A large data resource of genomic copy number variation across neurodevelopmental disorders. NPJ Genom Med 4:26–26. https://doi.org/10.1038/s41525-019-0098-3

Zhu M, Need AC, Han Y, Ge D, Maia JM, Zhu Q, Heinzen EL, Cirulli ET, Pelak K, He M, Ruzzo EK, Gumbs C, Singh A, Feng S, Shianna KV, Goldstein DB (2012) Using ERDS to infer copy-number variants in high-coverage genomes. Am J Hum Genet 91:408–421. https://doi.org/10.1016/j.ajhg.2012.07.004

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, Henaff E, McIntyre ABR, Chandramohan D, Chen F, Jaeger E, Moshrefi A, Pham K, Stedman W, Liang T, Saghbini M, Dzakula Z, Hastie A, Cao H, Deikus G, Schadt E, Sebra R, Bashir A, Truty RM, Chang CC, Gulbahce N, Zhao K, Ghosh S, Hyland F, Fu Y, Chaisson M, Xiao C, Trow J, Sherry ST, Zaranek AW, Ball M, Bobe J, Estep P, Church GM, Marks P, Kyriazopoulou-Panagiotopoulou S, Zheng GXY, Schnall-Levin M, Ordonez HS, Mudivarti PA, Giorda K, Sheng Y, Rypdal KB, Salit M (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 3:160025–160025. https://doi.org/10.1038/sdata.2016.25