bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

VariCarta: a comprehensive database of harmonized genomic variants found in ASD sequencing studies

Manuel Belmadani[1,2], Matthew Jacobson[1,2], Nathan Holmes[1,2], Minh Phan[1,2], Paul Pavlidis[1,2], Sanja Rogic[1,2]*

1) Michael Smith Laboratories, UBC, Vancouver, BC
2) Department of Psychiatry, UBC, Vancouver, BC

*Corresponding author

177 Michael Smith Laboratories
2185 East Mall
University of British Columbia
Vancouver BC V6T1Z4, Canada
604 827 4211
rogic@msl.ubc.ca

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Background

Recent years has seen a boom in the application of the next-generation sequencing technology to the study of human diseases, including Autism Spectrum Disorder (ASD), where the focus has been on identifying rare, possibly causative genomic variants in ASD individuals. Because of the high genetic heterogeneity of ASD, a large number of subjects is needed to establish evidence for a variant or gene ASD-association, thus aggregating data across cohorts and studies is necessary. However, methodological inconsistencies and subject overlap across studies complicate data aggregation.

Description

Here we present VariCarta, a web-based database developed to address these challenges by collecting, reconciling and consistently cataloguing literature-derived genomic variants found in ASD subjects using ongoing semi-manual curation. The careful manual curation combined with a robust data import pipeline rectifies errors, converts variants into a standardized format, identifies and harmonizes cohort overlaps and documents data provenance. The harmonization aspect is especially important since it prevents the potential double-counting of variants which can lead to inflation of gene-based evidence for ASD-association.

Conclusion

VariCarta is the largest collection of systematically curated, harmonized and comprehensively annotated literature-derived ASD-associated. The database currently contains 35,615 variant events from 8,044 subjects, collected across 50 publications, and

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reconciles 6,057 variants that have been reported in literature multiple times. VariCarta is freely accessible at http://varicarta.msl.ubc.ca.

Background

The genetics risk of ASD is based on the interplay between common inherited variants and rare, often *de novo*, variants. While the aggregated contribution of common variants is estimated to be high, ranging up to 50% (Klei et al., 2012; Gaugler et al., 2014), their individual effect sizes are small and only a few common risk loci have yet been identified so far, partially due to the limited cohort sizes used in genome-wide association studies up to now (He et al., 2013a; Grove et al., 2017; Kosmicki et al., 2017; Vorstman et al., 2017). On the other hand, rare variants with large effect sizes, such as copy-number variants or likely gene damaging single nucleotide variants (SNVs), were found in many ASD individuals; it is estimated that rare genetic variants, both *de novo* and inherited, are causal in at least 10–30% of cases (Iossifov et al., 2012, 2014; Sanders et al., 2015; Turner et al., 2017a; Vorstman et al., 2017; Yuen et al., 2017). Consequently, most of the effort towards understanding the etiology of ASD in the last half a decade has gone into employing next-generation sequencing technology to identify rare, possibly causative mutations in ASD individuals.

There have been a few dozen whole exome sequencing (WES) or whole genome sequencing (WGS) studies of various ASD cohorts, reporting on identified variants and genes containing them. The overlap within and between the studies' findings at the gene and, especially, variant level has been limited, highlighting the genetic heterogeneity of the disorder. Thus a key strategy for identifying ASD candidate genes is to sequence

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

large numbers of individual genomes to identify likely impactful events, especially *de novo* mutations, recurring in the same gene in unrelated individuals (Buxbaum et al., 2012; Sanders et al., 2012a). This approach resulted in the discovery of many ASD candidate genes. Current estimates for the total number of ASD-associated genes range from several hundreds to over a thousand (Iossifov et al., 2012, 2014; Sanders et al., 2012b, 2015; He et al., 2013b; Ronemus et al., 2014).

While the sequencing efforts continue, there is a need to aggregate and consolidate the findings. The analysis and reporting of variants across studies can vary in many respects, such as reference assembly, transcript set, variant calling and annotation software used, as well as the choice of variant nomenclature, and without careful curation and standardization they are not easily comparable and integrable (Deans et al.; McCarthy et al., 2014; Yen et al., 2017). At the same time, some cohorts (or individuals within cohorts) have been repeatedly sequenced and analyzed, and the same variants have been reported multiple times across publications, but not necessarily in a consistent format. If pooled across studies, this can lead to double-counting and inflation of gene-based evidence for ASD association. With these challenges in mind, we developed VariCarta with the goal of collecting, reconciling and accurately cataloguing literature-derived ASD-associated variants. We employ precise, systematic curation of the data, standardized processing and reporting, identification of overlaps, and comprehensive annotation. The harmonized data is available to researchers for querying and download at http://varicarta.msl.ubc.ca.

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Construction and content

We searched the literature for publications reporting rare genomic variants, SNVs and InDels, found in subjects with ASD diagnosis. We prioritized whole-genome and whole-exome studies (37) over candidate gene studies (13). Papers currently included in VariCarta are listed in Supplemental File 1.

For each publication, we applied the following curation procedures, with an intermediate goal of organizing all relevant variant information in a tabular format that is ready for import. The first step in this process is to copy the relevant text from the source file (typically a supplementary file) as-is in a template import document. The completed document is composed of a set of predefined worksheets, which contain the publication's metadata, variant data and a description of the steps needed to automatically extract, transform and load the data into a uniform variant data model. This document is parsed by a computational pipeline, which validates and stores the data into a relational database. The link to the pipeline's source code as well as an example import document are found at varicarta.msl.ubc.ca/downloads.

There are different variant reporting conventions used in in the literature, so careful inspection is needed to determine the format that was used in a publication, and steps needed to convert it into a standardized format VariCarta uses for displaying variants. These steps are described in the import document and may include conversion of non-genomic to genomic variant coordinates using TransVar (Zhou et al., 2015), biocommons UTA and HGVS packages (Hart et al., 2015), UCSC liftOver of genomic coordinates to GRCh37/hg19 human genome assembly, conversion from zero-based to one-based

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

indexing system, correction of erroneous reference bases and so on. Once the genomic coordinates have been resolved and the uniform formatting has been applied, we use ANNOVAR (Wang et al., 2010) to annotate variants. This annotation includes gene-level information, such as genomic context, transcript accession number, and functional effect of the variant within the gene. We also include variant-level annotations using CADD (Kircher et al., 2014) for variant deleteriousness prediction and ExAC (Lek et al., 2016) for allele frequencies in the general population. At this stage, we exclude variants that are found in control subjects or subjects not reported to have an ASD diagnosis (for studies that report on cohorts with different neurodevelopmental phenotypes). Variants were excluded if they were reported as having failed to validate using an orthogonal sequencing technology, whose coordinates and/or reference alleles cannot be confidently disambiguated or were problematic for some other reason detailed in the curator notes. Although the trend in ASD sequencing studies is to report mainly rare coding genomic variants, the final list of imported variants may include common and non-coding variants if these were present in the original report. These are easily distinguished based on ExAC frequencies or genomic context.

The last phase of variant processing is harmonization, which ensures that variant events are not double-counted. We define a “variant event” as a unique combination of a reference allele, its genomic location and alternative allele belonging to a single individual. We also define a “complex event” as a grouping of two or more variant events from the same individual that differ but have overlapping or adjacent genomic coordinates. This indicates that the grouped variant events might be describing the same underlying genotype but are incongruent due to the heterogeneity of formats used across

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

papers. For example, the same several-base long substitution can get reported either as a single substitution variant or a series of SNVs.

Because subject IDs are used to define variant events, we take special care when handling cohorts that are already present in VariCarta to ensure that subject IDs between the studies are consistent. Occasionally, the same subject could have slightly different IDs between studies: for example, Yuen et al. 2015 (Yuen et al., 2015) uses "-" (dash), while Yuen et al. 2017 (Yuen et al., 2017) uses "_" (underscore) in otherwise identical subject IDs from the same cohort. Another example is the omission of "p1" suffix, meant to indicate a proband, from a Simons Simplex Collection (SSC) family ID in Iossifov et al. (2012) (Iossifov et al., 2012), Iossifov et al. (2014) (Iossifov et al., 2014) and Ji et al. (2016) (Ji et al., 2016). We attempt to rectify such discrepancies during the curation stage. This is facilitated by the introduction of internal VariCarta-specific subject IDs that can be linked with more than one original study ID. However, we were not always able to resolve overlaps, such as in cases were subject IDs are not provided or are non-mappable to the original cohort identifiers. While these variants appear in the database as uniquely reported, the curation notes on the publication details page, would indicate the possibility of overlap.

Utility and discussion

VariCarta web site features

The central feature of the website is its capability of querying variants by gene name/symbol or by genomic region. The resulting variant table shows the summarized information about all the variants found in the query region and is the launching point for

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

accessing gene information from the Ensembl and NCBI databases, genomic annotation of the region using the UCSC Genome browser, information about papers that reported the variants and original, published variant data. A variant event is displayed only once in the variant table, with the Sources column listing IDs of all publications reporting it. The publication IDs are tagged to indicate the scope of sequencing study ([W] whole genome sequencing, [E] whole exome sequencing, [T] targeted sequencing). A complex event is initially displayed as one row that can be expanded to show the information about each grouped variant event (an example of a complex event is shown in Table 1). All query results are available for download as plain text/csv format from the spreadsheet icon in the table navigation header bar.

Table 1: An example of a complex event in VariCarta

Location	**Ref**	**Alt**	**Effect**	**Source**
chr5:134059281-134059281	C	A	nonsynonymous SNV	Iossifov2014, Kosmicki2017, Krupp2017
chr5:134059282-134059282	C	G	nonsynonymous SNV	Iossifov2014, Ji2016, Kosmicki2017
chr5:134059281-134059282	CC	AG	non-frameshift substitution	Krumm2015

A two-base substitution found in SSC subject 14606.p1 has been reported in five different papers included in VariCarta. Iossifov et al. (Iossifov et al., 2014) and Kosmicki et al. (Kosmicki et al., 2017) report it as two contiguous SNVs (C->A and C->G) at positions 134059281 and 134059282 on chr5, while Krumm et al. (Krumm et al., 2015) reports it as a non-frameshift substitution (CC->AG). Both representations describe the

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

same resulting genotype. The other two papers report just one of the contiguous SNVs each (Krupp et al. (Krupp et al., 2017): C->A and Ji et al. (Ji et al., 2016): C->G). This complex variant can also be viewed directly in VariCarta at varicarta.msl.ubc.ca/variant?chr=5&start=134059281&stop=134059281.

The original data as presented in the source publication for each variant event can be accessed by clicking on a magnifying glass icon, which launches a pop-up window displaying the variant information parsed from the publication of origin or inferred during the curation stage. This allows users to confirm the relationship between the source and our representation of it in the VariCarta system.

The list of publications that have been curated and included in VariCarta is shown on the Publications page (varicarta.msl.ubc.ca/publications; also see Supplemental File 1). The basic publication information is listed in a table with links to the original publications. Clicking on a book icon in the Details column opens a page with more detailed information about the study, including methodology used, size and type of the cohort, and types of variants reported. We also provide curation notes, which detail issues that had to be resolved during the curation stage, or other noteworthy information regarding the study. Clicking on the variant event count will display all the variants from that publication that are available in VariCarta.

The Statistics page (varicarta.msl.ubc.ca/stats) offers several gene rankings based on different criteria. These rankings are not intended to be used for a prioritization of ASD candidate genes because they can include common variants (non-coding variant are

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

excluded from ranking calculations), as well as variants from targeted studies, which are typically excluded for the prioritization purposes. The Statistics page also shows the distribution of variants across publications, functional effects and genomic features. Finally, it includes a heatmap of the variant overlap between publications (Figure 1), which illustrates the extent of variant double-reporting across the literature. The variants reported in two different papers can be accessed by clicking on the overlap number in the heatmap.

Any VariCarta updates will be immediately available on the development version of the database (dev.varicarta.msl.ubc.ca), while the main site will be updated on regular bases. Previous releases of the database can be downloaded from varicarta.msl.ubc.ca/downloads. The source code for the web application and the variant processing pipeline is open source (Apache 2 license) and available on GitHub (link from varicarta.msl.ubc.ca /downloads). Questions, comments and requests should be sent to pavlab-support@msl.ubc.ca.

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

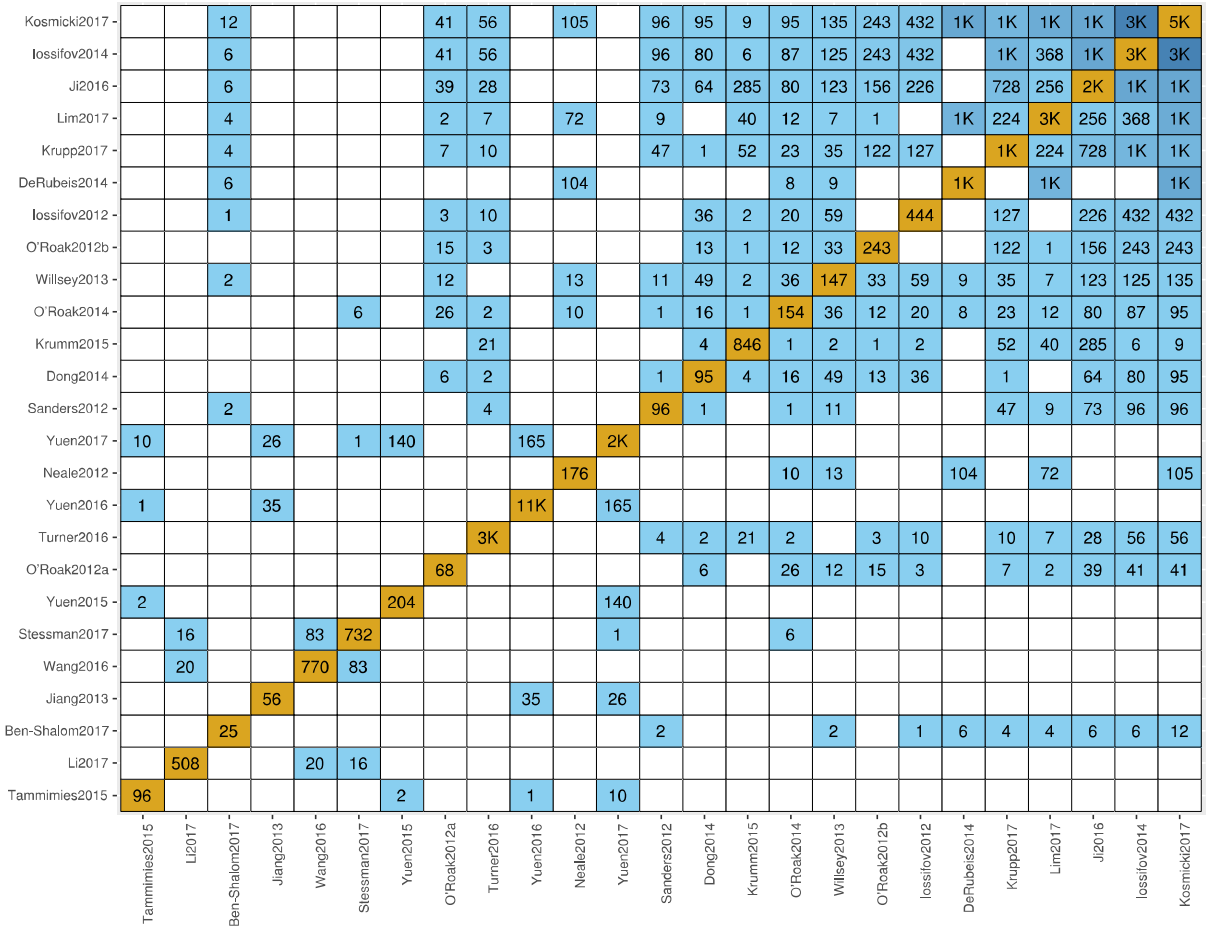

Figure 1: Heatmap showing the overlap between variant events reported by different papers. Cells along the diagonal represent the total number of variant events originating from each paper. Only a subset of the data is shown, for the full version see varicarta.msl.ubc.ca/stats.

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

VariCarta statistics

VariCarta currently contains 35,615 variant events from 8,044 subjects, collected from 50 publications. All the subjects are affected individuals (VariCarta does not collect variants from control individuals), majority of them with a reported clear ASD diagnosis and only a few with an unclear diagnosis. The latter are from publications reporting on ASD subjects among others, but without clear diagnosis label provided. Thirty-seven publications are from whole genome or whole exome sequencing studies and the remainder are from more targeted (i.e. candidate gene) studies. The variants are affecting 12,416 genes, with SCN2A being the gene with most reported variants (103 variants).

Comparison to similar resources

We are aware of three other public databases that collect ASD variant data from scholarly literature. One of them is considered to be ASD-specific (SFARI), while the other two are more general, containing *de novo* mutations associated with multiple phenotypes (denovo-db and NPdenovo).

The SFARI gene database, a licensed version of AutDB (Basu et al., 2009) by MindSpec, is a web portal for cataloguing ASD candidate genes. The central SFARI Human Gene module contains information about human genes that have been associated with ASD, relevant references from articles, and different kinds of genetic datasets that provide evidence for linking the genes to ASD and help categorize and score genes based on the strength of that evidence. These data include genetic variants, both rare and common, reported to be associated with the disease. However, the SFARI database is not entirely ASD-specific since many of the reported variants were found in subjects not diagnosed with ASD, but other disorders, such as intellectual disability, epilepsy and general

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

developmental delay. Since the SFARI variant table does not show the disease association, the access to ASD-specific variants is non-trivial. The variant annotation in SFARI also lacks transcript information and genomic coordinates, which makes it difficult to identify the exact genomic location of the variant and its predicted functional effect. Another concern is that, due to the multiple WES/WGS publications on the same cohorts of ASD individuals, some variants are reported more than once potentially leading to the inflation of the ASD-association evidence. Finally, the variant information in SFARI is not available for bulk download and thus cannot be used for any kind of high-throughput analysis.

denovo-db (Turner et al., 2017b) is another database of germline *de novo* variants, not limited to neuropsychiatric disorders. The latest release, from August 2018, contains data from 16 ASD studies (compared to 50 studies in VariCarta). The database provides variants' comprehensive position annotation, damage scores, population frequency and validation status. Turner et al. (Turner et al., 2017b) discuss the variant over-reporting problem in the literature, especially in the case of heavily studied SSC cohort for which the sequencing information has been reported in multiple publications. Their solution to the problem is to avoid duplication of samples in their database, thus when a new study is imported, any samples that are pre-existing in the database would be ignored.
NPdenovo (Neuropsychiatric Disorder De novo Mutations Database) (Li et al., 2016) collects *de novo* mutations found in subjects with neuropsychiatric disorders and their sibling controls. The database provides functional annotation of the variants, information about associated genes, their brain expression patterns and co-expression. According to the website, NPdenovo was most recently updated in 2015, although their publication

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

table (http://www.wzgenomics.cn/NPdenovo/home_detail.php) includes a few papers published in more recent years. The table lists 13 ASD-related studies, however for three of these some or all variant data were not included in the database due to the cohort overlap (as explained in the footnote of the table).

Discussion

Here we present VariCarta, a database of literature-derived genomic variants found in ASD subjects. The motivation for developing VariCarta arose from our own research interests and inability to find a comprehensive collection of published, reconciled and well-annotated ASD-associated variants. Among the main features of VariCarta are the careful curation of variants combined with the added robustness of a data import pipeline into a relational database, with the goals of rectifying errors, standardizing reporting format, harmonizing cohort overlap, appending comprehensive annotation and tracking provenance to the original report.

Even though the prevalence of ASD is fairly high (Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators and Centers for Disease Control and Prevention (CDC), 2014) the number and size of cohorts with genetic samples available for sequencing is still limited. As a consequence, some cohorts have been extensively studied. The most prominent example is the SSC cohort of 2600 families (Fischbach and Lord, 2010), which has subjects that have been reported on in at least 19 studies currently included in VariCarta. Since it is not always clear what cohorts were used in a study and there are inconsistencies in subject ID and variant reporting formats across studies, the overlap between studies' results is not always obvious. This can lead to double-counting of variants and inflation of gene-based evidence for ASD

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

association. Having the precise counts of reported variants per gene is important, especially considering that most ASD candidate genes have been established based on the observation of clusters of damaging mutations in multiple unrelated individuals.

One way to address this problem, as adopted by denovo-db and NPdenovo, is to avoid studies reporting on a cohort that has been used in another already curated study. While this approach may work for the obvious cases of cohort overlaps, there are many partial and more complex overlaps between studies, as illustrated in Figure 1, and consolidating these cases requires more systematic approach. In addition, studies that report on the same subjects, do not necessarily report the same variants, either due to differences in variant processing (e.g. applying different filters) or heterogeneity of variant reporting formats. Simply excluding variants from one of the studies can result in the omission of valuable information. Finally, linking variants to all the papers that report them is important for documenting data provenance and accessing additional information provided in the publication.

VariCarta puts special emphasis on harmonizing variants derived from the same cohorts of subjects by carefully linking the variants together, while still retaining the original publication and annotation data. Currently VariCarta contains 6,057 variants that have been reported more than once; some of them has been published in as many as nine papers (varicarta.msl.ubc.ca/variant?chr=20&start=49510027&stop=49510027).

Although we made a substantial effort to convert all variants into the same format sometimes the differences cannot be completely resolved; the example in Table 1 is one of the simpler cases. In these cases, we still link them together as complex events,

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

indicating that they have been reported as overlapping variants in the same individual, and are likely to represent the same event. There are currently 327 complex events in VariCarta.

Conclusions

To the best of our knowledge, VariCarta is the largest collection of systematically curated, harmonized and annotated literature-derived variants that is specific to ASD. Although variants can go through multiple transformation steps as they are being converted to the VariCarta uniform format, these steps are documented and the original variant information is retained and readily accessible. Our curators are always on the lookout for new relevant publications, which are being continuously added to the database. While whole genome and whole exome studies are of primary interest, we will continue to add gene-targeted studies that are aligned with our research interests or requested by the research community. We hope that the ASD research community as well as clinicians working with ASD subjects will find VariCarta to be a useful resource.

Funding

This project was supported by Simons Foundation (SFARI 368406 to PP).

Authors' contributions

MB implemented the data processing pipeline and contributed to software development. MJ designed and engineered the web application and relational database model. NH participated in the initial data collection. JP carried out curation of all publications and performed data processing tasks. SR and PP conceived the study. PP provided overall project leadership. SR provided project management and led database design and testing.

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SR drafted the manuscript with input from PP and MB. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank the authors of publications included in VariCarta who communicated with us to provide clarification or additional information.

References

Basu SN, Kollu R, Banerjee-Basu S. 2009. AutDB: a gene reference resource for autism research. Nucleic Acids Res 37:D832–D836.

Buxbaum JD, Daly MJ, Devlin B, Lehner T, Roeder K, State MW, Autism Sequencing Consortium. 2012. The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. Neuron 76:1052–1056.

Deans ZC, Fairley JA, Dunnen JT den, Clark C. HGVS Nomenclature in Practice: An Example from the United Kingdom National External Quality Assessment Scheme. Hum Mutat 37:576–578.

Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators, Centers for Disease Control and Prevention (CDC). 2014. Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010. Morb Mortal Wkly Rep Surveill Summ Wash DC 2002 63:1–21.

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fischbach GD, Lord C. 2010. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron 68:192–195.

Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, Mahajan M, Manaa D, Pawitan Y, Reichert J, Ripke S, Sandin S, et al. 2014. Most genetic risk for autism resides with common variation. Nat Genet 46:881–885.

Grove J, Ripke S, Als TD, Mattheisen M, Walters R, Won H, Pallesen J, Agerbo E, Andreassen OA, Anney R, Belliveau R, Bettella F, et al. 2017. Common risk variants identified in autism spectrum disorder. bioRxiv 224774.

Hart RK, Rico R, Hare E, Garcia J, Westbrook J, Fusaro VA. 2015. A Python package for parsing, validating, mapping and formatting sequence variants using HGVS nomenclature. Bioinformatics 31:268–270.

He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD, State MW, Devlin B, et al. 2013a. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet 9:e1003671.

He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD, State MW, Devlin B, et al. 2013b. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet 9:e1003671.

Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, Smith JD, Paeper B, et al. 2014. The

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

contribution of de novo coding mutations to autism spectrum disorder. Nature 515:216–221.

Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee Y, Narzisi G, Leotta A, Kendall J, Grabowska E, et al. 2012. De Novo Gene Disruptions in Children on the Autistic Spectrum. Neuron 74:285–299.

Ji X, Kember RL, Brown CD, Bućan M. 2016. Increased burden of deleterious variants in essential genes in autism spectrum disorder. Proc Natl Acad Sci 113:15054–15059.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310.

Klei L, Sanders SJ, Murtha MT, Hus V, Lowe JK, Willsey AJ, Moreno-De-Luca D, Yu TW, Fombonne E, Geschwind D, Grice DE, Ledbetter DH, et al. 2012. Common genetic variants, acting additively, are a major source of risk for autism. Mol Autism 3:9.

Kosmicki JA, Samocha KE, Howrigan DP, Sanders SJ, Slowikowski K, Lek M, Karczewski KJ, Cutler DJ, Devlin B, Roeder K, Buxbaum JD, Neale BM, et al. 2017. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. Nat Genet 49:504–510.

Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, Raja A, Coe BP, Stessman HA, He Z-X, Leal SM, Bernier R, et al. 2015. Excess of rare, inherited truncating mutations in autism. Nat Genet 47:582–588.

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Krupp DR, Barnard RA, Duffourd Y, Evans SA, Mulqueen RM, Bernier R, Rivière J-B, Fombonne E, O'Roak BJ. 2017. Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. Am J Hum Genet 101:369–390.

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, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285.

Li J, Cai T, Jiang Y, Chen H, He X, Chen C, Li X, Shao Q, Ran X, Li Z, Xia K, Liu C, et al. 2016. Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. Mol Psychiatry 21:290–297.

McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier J-B, Donnelly P. 2014. Choice of transcripts and software has a large effect on variant annotation. Genome Med 6:26.

Ronemus M, Iossifov I, Levy D, Wigler M. 2014. The role of de novo mutations in the genetics of autism spectrum disorders. Nat Rev Genet 15:133–141.

Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, Murtha MT, Bal VH, Bishop SL, Dong S, Goldberg AP, Jinlu C, et al. 2015. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron 87:1215–1233.

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, et al. 2012a.

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485:237–241.

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, et al. 2012b. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485:237–241.

Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, Kronenberg ZN, Hormozdiari F, Raja A, Pennacchio LA, Darnell RB, Eichler EE. 2017a. Genomic Patterns of De Novo Mutation in Simplex Autism. Cell 171:710-722.e12.

Turner TN, Yi Q, Krumm N, Huddleston J, Hoekzema K, F. Stessman HA, Doebley A-L, Bernier RA, Nickerson DA, Eichler EE. 2017b. denovo-db: a compendium of human de novo variants. Nucleic Acids Res 45:D804–D811.

Vorstman JAS, Parr JR, Moreno-De-Luca D, Anney RJL, Nurnberger Jr JI, Hallmayer JF. 2017. Autism genetics: opportunities and challenges for clinical translation. Nat Rev Genet 18:362–376.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164–e164.

Yen JL, Garcia S, Montana A, Harris J, Chervitz S, Morra M, West J, Chen R, Church DM. 2017. A variant by any name: quantifying annotation discordance across tools and clinical databases. Genome Med 9:7.

bioRxiv preprint doi: https://doi.org/10.1101/608356; this version posted April 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Yuen RKC, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel RV, Whitney J, Deflaux N, Bingham J, Wang Z, Pellecchia G, Buchanan JA, et al. 2017. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci 20:602–611.

Yuen RKC, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N, Chrysler C, Nalpathamkalam T, Pellecchia G, Liu Y, Gazzellone MJ, D'Abate L, et al. 2015. Whole-genome sequencing of quartet families with autism spectrum disorder. Nat Med 21:185–191.

Zhou W, Chen T, Chong Z, Rohrdanz MA, Melott JM, Wakefield C, Zeng J, Weinstein JN, Meric-Bernstam F, Mills GB, Chen K. 2015. TransVar: a multilevel variant annotator for precision genomics. Nat Methods 12:1002–1003.