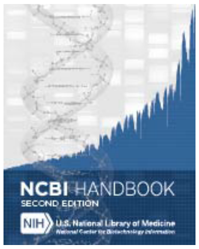
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Kitts A, Church D, Hefferon T, et al. dbVar. 2014 Oct 26. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/


NCBI HANDBOOK
SECOND EDITION
NIH U.S. National Library of Medicine
National Center for Biotechnology Information

dbVar

Adrienne Kitts, M.S., Deanna Church, Ph.D., Tim Hefferon, Ph.D., and Lon Phan, Ph.D.

Created: October 26, 2014.

Scope

The Database of Genomic Structural Variation (dbVar) is a National Center for Biotechnology Information (NCBI) archival database that manages sequence variation. dbVar complements dbSNP by archiving copy number variants (CNV), insertions, deletions, inversions, and translocations (1) that are longer than 50 base pairs (bp). The database is organized around the studies that have identified these variants, and includes variations from research-based and clinical submissions. Structural variants that have asserted germline or somatic clinical significance should be submitted to Clinvar which will forward appropriate portions of the data to dbVar for accessioning.

dbVar and the European Bioinformatics Institute's (EBI's) Database of Genomic Variants archive (DGVa) share the same data model and exchange data regularly. Together they represent the largest and most comprehensive archive of structural variation in the world.

Structural variation can be detected with a variety of experimental methods. Most of the data dbVar receives have been generated by Next-Generation Sequencing (NGS) or microarray (either oligo or SNP array) technologies. These methods alone can detect a wide variety of variant types, but they vary in the degree of precision and certainty they can provide with respect to breakpoint location and copy number change. A complete list of structural variation detection methods and analysis types can be found in dbVar's online documentation.

dbVar's primary tasks are to support the submission and organization of structural variations to aid researchers in the study of a wide range of biological problems . dbVar assigns stable, traceable identifiers to the structural variants found in each study, calculates locations on newer assemblies as appropriate, provides some validation of content, and integrates the structural variant data submitted to us with a wide array of NCBI tools and data.

Medical Genetics

Advances in next-generation sequencing technologies have allowed researchers to generate massive amounts of sequence data. When clinical samples are sequenced using these technologies, novel structural variants that have causative roles in disease may be identified.

Structural variants have been implicated in complex human disorders that include cancer, neurological diseases, and developmental disabilities such as Down, Turner, and Prader-Willi Syndromes, as well as increased susceptibility to disorders including HIV, Crohn disease, and lupus(4). dbVar accepts submissions from disease resources to maximize our representation of complex genetic disorders. One example are the submissions from the International Collaboration for Clinical Genomics (ICCG), which is part of ClinGen, a larger NIH-funded effort.

dbVar manages and organizes structural variation data such as location and variation type from all submitters and provides clinicians and researchers a point of access to both clinical cases and control subjects. dbVar also provides value-added data such as confirmed validation status and current clinical phenotype interpretation through our integrated relationship with ClinVar.

Association Studies

Structural variant submissions that contain sensitive clinical information or do not have informed consent from the originating sample donor are submitted to NCBI's Database of Genotypes and Phenotypes (dbGaP). Variants are assigned dbGaP accessions, stripped of identifying information, and then submitted to dbVar. dbVar in turn, provides users access to variant location, variant type, and summary variant data stripped of identifying information. All sensitive information contained in the dbGaP submission remains in dbGaP and can only be retrieved through dbGaP's controlled access system.

dbVar's annotated records and catalog of common structural variations can be used to inform the design of Genome-wide Association Studies (GWAS), create variation arrays used in these studies, and interpret GWAS study results.

Evolutionary Biology

Although it is known that structural variation plays an important role in species and disease evolution(5), until recently, the technology required to produce structural variant maps with the degree of resolution needed to trace the evolutionary history of a species or gene has not been available. The advent of high throughput sequencing technologies has made the comparison of structural variation between related species possible (6), so the time when evolutionary analysis of structural variation will be more commonplace is approaching. dbVar currently houses studies that have evolutionary implications, including studies that compare structural variation observed in different breeds of dog (nstd10, nstd13), among the Great Apes (nstd82, estd193), and structural variant studies that have implications in gene and gene function evolution (estd199, nstd78). dbVar's variant catalog can be used to inform evolutionary studies through its use in the selection of candidate variants for both species and gene evolutionary studies.

History

Creation and Growth

Following the discovery that healthy human subjects contained copy number variants (CNV) wide spread throughout the genome, the Database of Genomic Variants (DGV) was founded in 2004 at the Center for Applied Genomics in Toronto, Canada, to organize, manage, and provide access to the initial data produced by early structural variation studies (7). It was soon discovered that a more comprehensive and permanent archive would be needed to work in conjunction with DGV. The National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) collaboratively launched dbVar and DGVa (Database of Genomic Variants archive) in 2009 to meet this need.

The collaboration between dbVar and DGVa was designed to be close; the two resources communicate and exchange data on a regular basis, and share a uniform data model as well as similar database schemas. dbVar and DGVa together assumed DGV's role in the organization and management of structural variation data and augmented this role by providing increased capacity for structural variant storage as well as integration with a host of other genetic databases and tools.

Database Development Milestones

After its initial launch in 2009, dbVar grew slowly in size and began integrating its data with that of other NCBI resources. By 2012, dbVar underwent its first major improvement with a database schema overhaul conducted in collaboration with DGVa, which vastly improved mutual data exchange capabilities. 2012 was also the year that dbVar released the dbVar Genome Browser, which dramatically improved our users' ability to interpret and analyze dbVar data by allowing a side-by-side comparison of variants from any study in dbVar superimposed on an annotated set of chromosome ideograms from multiple assemblies. By 2013, dbVar saw a great advance in its ability to provide its users with deeper clinical insight when dbVar became fully integrated with ClinVar. The integrated relationship between dbVar and ClinVar provides dbVar users with access to in-depth information about clinical assertions associated with structural variants. With the release of the Variation Viewer upgrade in 2013, dbVar users could also explore structural variants side by side with small-scale variants (dbSNP) and clinical variations (ClinVar) in a genomic context. dbVar also integrated with NCBI's Variation Reporter at about the same time, allowing users to submit variation calls to find metadata for known variants, or see the predicted molecular consequence of the call if the submitted variant is novel.

The discovery of structural variants that have multiple breakpoints derived from complex genetic rearrangements spurred dbVar's development of a database model to capture and display complex structural variants associated with cancer and other illnesses. dbVar's ability to capture both simple and complex structural variants was greatly improved in the fall of 2014 when it began accepting dbVar submissions in Variant Call Format (VCF). dbVar's VCF submission specification is very similar to the 1000 Genomes VCF specification v.4.2, which allows for annotation of both simple and complex structural variants. dbVar's VCF specification contains modifications that allow our users to submit additional information for their structural variants with ease.

Evolution in Submitted Content

Upon its initial release, dbVar was populated with historical structural variation data that was mined from available research literature. Because the first paper describing the prevalence of genomic structural variation data in normal human subjects was released in 2004 (8), there wasn't very much historic data available, which necessarily meant that the initial population of structural variants in dbVar was limited to human. Initial submissions to dbVar in 2009 included data from human as well as from model species such as *Mus musculus*, and agriculturally important species such as *Bos taurus* and *Sus scrofa.*

Currently, dbVar archives and provides access to 4.3 million variant regions from 127 studies and 16 taxa that include plants, invertebrates, and vertebrates.

Data Model

dbVar stores data in a three-level nested hierarchy:

The topmost level in the dbVar nested hierarchy is called the Study (**std**), which can be either a particular publication that produced a set of data, or a community resource that submits new data on a regular basis. Regardless of which it is, a study is a record that serves to indicate a group of data (sv and their supporting ssv) generated in a particular series of experiments, and provide general information about those experiments.

The next level down in the hierarchy is called a Variant Region (**sv**: structural variant), which is formed when multiple Variant Calls (ssv) in a particular region from one study are grouped together under the same identifier.

Note: dbVar does not currently integrate data across studies, so what appears to be the same region will not have the same sv identifier if those regions were identified in different studies (std).

The base level of dbVar's nested data storage hierarchy is called a "Variant Call" or **ssv** (supporting structural variant). Variant Calls are the individual variant placements that contain the actual data used to place submitted structural variants.

Accessions

If the Study, the Variant Region, or Variant Call was originally submitted to dbVar at NCBI, the accession numbers are given an "n" prefix (i.e., nstd, nsv, and nssv, respectively). If the Study, Variant Region or Variant Call was originally submitted to DGVa at EBI, the accession numbers are given an "e" prefix (i.e., estd, esv, and essv).

The quality of the data in dbVar depends on the quality of the data submitted to us, so we provide the following consistency checks:

- dbVar performs data validation during submission processing that will catch certain types of errors, such as inconsistent data, invalid placements, or invalid entries. Serious errors will cause the validation processes to stop submission loading, and dbVar will determine at that point whether it is necessary to contact the submitter for corrections.
- If a variant submission has a noticeably incorrect placement (e.g., coordinates located at the end of a chromosome or within an assembly gap), we will return the submission and ask the submitter to check and/or correct the location.

Since dbVar is minimally curated and reviews submission for obvious errors only, it is important that all submissions to dbVar contain high quality data, and the responsibility for maintaining data quality rests ultimately with our submitters.

Note: 1000Genomes records in dbVar now contain a data quality indicator.

Study (std, nstd, or estd)

A dbVar Study is a record that serves to group together Variant Region (sv) and Variant Call (ssv) data with descriptive metadata including organism, study type, submitter, project, and any associated publications.

Although the fields that characterize a study rarely change, the Variant Regions and Variant Calls in studies submitted by an ongoing project such as 1000Genomes or ICCG (International Collaboration for Clinical Genomics) may change if the submitter updates a Variant Region or submits new Variant Calls.

Study Submitted Content

Content for a study includes general information including author contact information, study identifiers, a brief description of the study, the study type (e.g., control set, case-control, matched-normal, etc.), associated IDs from PubMed, Taxonomy, the Entrez Genome Project, and dbGaP, as well as a description and identifiers for all samples and sample sets used in the study. Complete information regarding methods and analyses used for variant discovery must also be included in a study submission as well as any validation data generated for the experiment.

The Study is a dbVar user's entry point to structural variation data, which is why the primary search mode of dbVar is the "Study Browser". Because data contained within a Study record are found at the same time by the same authors in the same laboratory (or laboratories) using a specific set of methods and analyses, you can use any of these characteristics as search terms in the dbVar Study Browser to find studies, variant regions, and variant calls whose metadata contain the entered search term.

Variant Region (sv, nsv, or esv)

A Variant Region (sv) record is composed of Variant Calls (ssv) located at or near the same location that have been grouped together, or "merged"—either by the submitter at his or her discretion, or by dbVar in the case of calls with identical coordinates and call types. You can consider the Variant Region record as a parent to the grouped ssv records that are its children in the dbVar nested record hierarchy. It should be noted that if two Variant Regions from a study overlap, the submitter can merge them into a single Variant Region. Each submitted Variant Region is assigned a unique sv (structural variant) ID.

Variant Region records provide variant location and placement information, the evidence used to place the variant, available validation and clinical assertion data, as well as links to associated publications.

Reference Variants

As stated in the previous section, a Variant Region, like a refSNP, is a marker on the genome that denotes a group of variants found in a specific region, but this is where the similarity ends. Variant Regions group variations of different types that may occur in the same position, but because of breakpoint uncertainty in the identification of variation boundaries, the variations may also occur scattered throughout the same identified genomic span. Because an exact location for each structural variant cannot be ascertained, true reference variants cannot be established at this time.

We will move closer to the identification of reference variants as sequencing continues, the comparison of genomes continues to expand, and current breakpoint ranges are narrowed down until they are much more defined.

Submitted Content

Merging Variant Calls into Variant Regions

Consolidation of similar or identical Variant Calls into a Variant Region, or merging overlapping Variant Regions into each other can be performed during submission of structural variants to dbVar, but submitter-performed merging is completely optional.

dbVar will merge those Variant Calls submitted without merge data with other Variant Calls at the same location and of the same type and will assign a Variant Region ID during submission processing. Likewise, dbVar will merge overlapping Variant Regions into each other if a submitter chooses not do so during submission.

Linked Variant Call / Linked Variant Region IDs

Submitters who choose to merge their Variant Calls or Variant Regions will need to provide a list of IDs for previously submitted Variant Calls or a list of Variant Region IDs that are to be merged. As an alternative, in the case of a Variant Call merge, submitters can provide the ID of the Variant Region parent in the Variant Call portion of the submission, while in the case of a Variant Region merge, submitters can provide the ID of the parent Variant Region instead of a list of the merging Variant Region IDs.

Assertion Method

Submitters who choose to merge Variant Calls or Variant Regions will need to provide details of the method they used to assert that a group of Variant Calls or Variant Regions require merging. Submitters can provide any algorithms they used to establish that the Variant Calls or Variant Regions require merging, or they can provide a simple statement such as "reciprocal 50% overlap", "Calls with identical coordinates merged", or "Region is identical to call, no merge performed".

Note: Since dbVar is an archive, we do not validate submitted assertion methods, and as such, we rely on our submitters to use reliable assertion methods for merging groups of Variant calls into Variant Regions, or merging Variant Regions together.

Genotype data

dbVar has begun accepting genotype data in VCF format, and is already in receipt of Genotype data from the 1000Genomes Project submitted via VCF. dbVar will soon begin accepting genotypes in other formats via an updated submission template. If you have questions about submitting structural variation genotype data to dbVar, please contact us at: dbvar@ncbi.nlm.nih.gov

Variant Call (ssv)

A Variant Call record represents an independent instance of a variation produced by an experiment as well as its subsequent analysis. It includes data indicating the location, type, and size of a detected structural variant. Each submitted Variant Call is assigned an ssv (supporting structural variant) ID. If the variant call was submitted to dbVar at NCBI, the ssv ID is given the prefix "nssv". If the variant call was submitted to DGVa at EBI, the ssv ID is given the prefix "essv".

Variant Calls are comparable in nature to alleles, but depending on the particular experiment, a variant may or may not actually be an allele. For example, an experiment may yield a variant call that is an allele, but a second, different analysis may yield a completely different call for that variant. In such a case, the variant isn't actually an allele—it is an artifact. Therefore, given this possible difference in analytical outcome, if a call was generated from a pooled sample, it may or may not be an allele.

Sequence Requirements

dbVar requires that all Variation Calls must be made on an assembly sequence that has already been submitted to an International Nucleotide Sequence Database Collaboration (INSDC) database, which includes Genbank, the European Nucleotide Archive, or the DNA Database of Japan (DDBJ).

Variant Call Boundaries

Structural variation can be difficult to represent because current structural variation detection technologies seldom provide the base pair resolution necessary to determine variant breakpoints. This introduces an element of uncertainty into the identification of breakpoint boundaries. The extent of this uncertainty depends on the experimental methods that were used to detect the variant, which in turn influences what data submitters will provide to us:

- Detection methods such as arrayCGH and SNP array produce only a range of coordinates within which the breakpoints likely occur, so the submitter can define a minimal region that is definitely involved in the variation, but cannot define precise breakpoints.
- Detection methods such as Paired-End Mapping and Optical Mapping will produce just the precise location for the outer boundaries between which the variant breakpoints must fall, so the submitter can define the region of the genome known to contain the variant, but not the exact location of the variant or its breakpoints.
- Detection methods such as long read sequencing technology or 2nd generation sequencing reads may or may not provide break point resolution. In those cases where breakpoint resolution is achieved, the submitter provides the breakpoint coordinates. In those cases where sequence detection does not give precise breakpoints, the submitter can provide a range of breakpoint coordinates.

When structural variants are submitted to dbVar, we ask the submitter to provide a specific set of data that will capture all the available information we need—including the degree of breakpoint uncertainty present—regardless of the detection method used to find the variant.

Validation

Because dbVar is an archive, we report variants as they are submitted to us and accept (but do not require) validation data used to confirm variant calls. To be considered "validated", a variant must be confirmed as valid by one or more separate methods. The number of calls validated for a variant region, validation methods and analysis will be part of the Study (std) page, while the Variant Region (sv) and Variant Call (ssv) records will contain summary validation data.

Dataflow

New Submissions and the Start of a New Release

Submissions to dbVar are currently accepted via email to dbvar@ncbi.nlm.nih.gov, and will eventually be accepted through a direct upload using the NCBI Variation Submission Portal, which will allow submitters to track the progress of their submissions and will allow for direct communication between dbVar and the submitter should an error be found during submission processing.

Most data submitted to dbVar are data associated with a recently published study, or a study associated with a publication currently in review. Submission updates to existing studies are generally limited to large ongoing studies like 1000Genomes.

Data Conversion

Data submitted to dbVar is received in the form of an Excel spreadsheet and is converted from Excel to dbVar's XML format. dbVar's converter software first converts the Submitted Excel spreadsheet into a series of tab separated, text-based files, and then during a subsequent step, the text files are converted into dbVar XML submission files.

The dbVar data converter contains a series of validation steps that scan the data for errors during each step in the conversion process. If the validation processes finds minor errors during either conversion step (e.g., data is not in the right form, etc.), the dbVar submissions team will correct the error in the original submission file and put it through the conversion process again.

If an error found during the conversion process is more complex (e.g., coordinates that extend beyond the length of the chromosome), then a member of the dbVar team will contact the submitter, explain the issue and ask the submitter to fix it and resubmit. When the corrected data are received from the submitter, dbVar loads the corrected data through the converter process again.

This process is repeated until the conversion process no longer generates error messages for the submission.

Data Testing

Once the submission has been successfully converted to dbVar XML submission files, the files are loaded into a test version of dbVar. The loading process itself has its own set of validations, which scan the data for errors. If the detected errors are simple, the dbVar submission team will correct them, and if the errors are complex, a member of the dbVar submission team will contact the submitter, explain the issue and ask the submitter to fix it and resubmit.

Once the submission loads to the dbVar test site successfully, the data are reviewed to verify that the data is appropriate to the submission and that the data are being displayed correctly. If the data as shown on the test site is incorrect or does not display properly, dbVar will troubleshoot any difficulties with the data.

It should be noted that we are in the process of shifting the validation processes so that all validations will take place during the conversion process.

Merging Calls into Regions

Each submission to dbVar will contain calls (nssv) which the submitter may or may not have grouped into regions (nsv) since the region portion of the submission is optional. If the submitter decides to create and submit a region or regions, dbVar will check the method used to create the region to insure that the grouping is accurate.

If the submission does not contain any regions, the converter will automatically merge all calls that are the same type and have the same coordinates.

If a call or calls within the submitted region extend beyond the coordinates defined by the submitted range, an error warning will be generated, and the dbVar staff will review the error. If the merging error is simple, the dbVar team will resolve the issue, but if the merging error is complex, a member of the dbVar team will contact the submitter, explain the problem and send the submission back to be fixed and resubmitted.

Clinical Assertions

Structural variants with clinical significance are submitted to ClinVar, which will then process and accession the data, and sends the data in the ClinVar XML format to dbVar. dbVar maps the ClinVar XML formatted data to dbVar XML format, validates the data, which then proceeds through the dbVar test site load process.

It should be noted that the integration between dbVar and ClinVar is still relatively new, so dataflow between these two resources is still in process and may change.

Association Studies

Structural variants contained within an association study are submitted to dbGaP. Before sending structural variant data from an association study to dbVar, dbGaP removes any sensitive data, and sends the data as tab files to dbVar.

Data Exchange with DGVa

dbVar exchanges data with DGVa every month using an exchange XML format agreed upon by both databases. Because of this shared database schema, the recipient archive usually experiences very few problems loading to their own database, and the exchanged data are available for viewing and download on the recipient's site within a week of their exchange.

Quality Control

If the data displayed on the dbVar test site is approved, the submission is loaded to our quality assurance (QA) database and cross-checked for errors. If the results of the QA testing show no errors, then the data is released to the dbVar public site.

Remapping

dbVar annotates Variant Regions (the non-redundant set of variations) on reference genome genomic sequences, chromosomes, mRNAs, and proteins as part of the NCBI RefSeq project. We then use Assembly-Assembly remapping to project features from the reference assembly coordinate system to selected assembly coordinate systems using genomic alignments. dbVar performs a base-by-base analysis of each feature on the source sequence in order to project the feature through the alignment to the new sequence.

Build Integration

dbVar updates the links between dbVar and BioProject, ClinVar, dbGaP, dbSNP, Gene, HomoloGene, MedGen, Nucleotide, OMIM, Protein, PubMed, PubMed Central, Taxonomy Variation Viewer, and Variation Reporter.

dbVar also maintains links from dbVar records to resources outside of NCBI's Entrez system such as ClinGen, GeneReviews, HPO, and OMIM. Links to all resources related to a particular dbVar record can be found in on the upper right-hand corner of the dbVar record under "Links".

Public Release

A dbVar public release involves an update to the public database and the production of a new set of files on the dbVar FTP site. dbVar currently makes an announcement on the dbVar News and Announcements RSS feed and NCBI News (http://www.ncbi.nlm.nih.gov/news) when a dbVar release is made publicly available, but intends to eventually move its public release announcements to the ncbi-announce list and will make these announcements on a monthly basis in the future. dbVar announcements are also posted on NCBI's Twitter account and Facebook page to take advantage of the visibility social networking can provide.

Access

dbVar can be queried directly from the search bar at the top of the dbVar homepage, by using the links to dbVar resources and search options located on the homepage (including FTP), or by accessing related NCBI resources that link to dbVar data.

dbVar Home Page

dbVar is a part of the Entrez integrated information retrieval system and may be searched either by using an ID number query, or by using combinations of different search fields and qualifiers.

Single Record Query

Use the search bar at the top of the dbVar homepage to find variations using dbVar record identifiers. The record identifiers currently supported for single record queries are the study ID (nstd or estd), the Variant Region ID (nsv or esv) and the variant call ID (nssv or essv).

Complex Entrez Query

Use the dbVar Advanced Search Builder page to construct a complex search using combinations of different search fields and qualifiers. The Advanced Search Builder allows you to construct a query by selecting multiple search terms from a large number of fields and qualifiers. See the Advanced Search Builder video tutorial for information about how to find existing values in fields and combine them to achieve a desired result.

Study Browser

The dbVar Study Browser displays all available dbVar studies, and allows the displayed studies to be filtered by organism, study type, method, and variant size. Once the number of available studies has been narrowed, links to publications and individual study pages allow a more in-depth search of available data.

Genome Browser

The dbVar Genome Browser searches dbVar data within the framework of a selected genomic assembly using a location, gene name, or phenotype as search terms and presents the result in a graphic display showing the variant in relation to genes located in the region. All dbVar variant region (nsv or esv) report pages are linked to the dbVar Genome Browser via the "Genome View" tab, which presents a graphic display of variant regions from other studies that overlap the variant displayed in the variant region report.

Variation Reporter

Variation Reporter matches submitted variation calls to references in ClinVar and to variants housed in dbVar or dbSNP, thereby allowing access via a Web search or through an application programming interface (API) to all data and metadata that dbVar has for the matching variants. If you submit novel variants and there are no matches between your data and the variants housed in dbVar or dbSNP, the Variation Reporter will provide the predicted consequence of each submitted variant.

Variation Viewer

Variation Viewer allows users to access variation data from dbVar, dbSNP, and ClinVar in relation to a specific gene or chromosomal location, and will allow the user to display data from any of these sources in an integrated navigable map. Users can search by dbVar accessions, gene, phenotype or disease, and chromosome positions (http://www.ncbi.nlm.nih.gov/variation/view/help/#search).

Search via ClinVar, Gene, or PubMed

There are multiple databases in NCBI that maintain links to dbVar. Related dbVar records are located by following links in the "Related Information" section of a record.

dbVar FTP Site

NCBI supports the public distribution of dbVar data by providing compressed data dumps in a number of different formats. Access to the NCBI FTP site is available via the World Wide Web (ftp://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/) in data formats that include CSV, GVF, TAB, VCF, and XML.

ADA Section 508-Compliance

All links provided on the dbVar homepage are also provided in text format at the bottom of the page to support browsing by text-based Web browsers. Suggestions for improving database access by disabled persons should be sent to the dbVar development group at dbvar@ncbi.nlm.nih.gov

Related Tools and Studies

Remapping

NCBI's Genome Remapping Service (Remap) supports the conversion of genomic locations from one sequence to another based on alignments. Use Remap if you have identified the location of variation on an assembly, or on a RefSeqGene/LRG, and want to determine the location on a different assembly (or on the genome in the case of the RefSeqGene). dbVar remaps data in all its submissions to and from recent assemblies (e.g., from GRCh37 to GRCh38).

Association Studies

dbGaP archives and distributes data from studies that examine the relationship between phenotype and genotype. Such studies include Genome-wide Association Studies (GWAS), medical sequencing, and molecular diagnostic assays. Links are available from dbGaP controlled access records to related variation data in dbVar, but there are no reciprocal links from dbVar records to dbGaP unless the aggregate data are public.

Variation as Related to Citations, Genes, Phenotypes, and other NCBI Databases

Multiple databases in NCBI can be used to identify variation that meets certain criteria. They may either reference dbVar ID numbers explicitly, or provide links from their records to records in dbVar.

Variation Reporter

Variation Reporter matches submitted variation call data to variants housed in dbVar or dbSNP, allowing access to all data and metadata that dbVar has for any known matching variants. If you submit novel variants to the Variation Reporter, and there are no matches between your data variants housed in dbVar or dbSNP, the Variation Reporter will provide the predicted consequence of each submitted variant.

Variation Viewer

Variation Viewer allows users to access variation data from dbVar, dbSNP, and ClinVar in relation to a specific gene or chromosomal location, and will allow the user to display data from any of these sources in an integrated navigable map.

1000 Genomes Browser

The 1000 Genomes Browser provides access to 1000 Genomes data including variations, genotypes, and sequence read alignments within the context of GRCh37, the reference assembly used by the 1000 Genomes Project for analysis. The browser allows you to configure the display to include multiple data tracks of interest and provides links to related data housed in various NCBI resources. The 1000 Genomes Browser allows users to quickly view alignments supporting a particular variant call and can be used to download and read variant data for small genomic regions of interest.

References

1. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Gen. 2006 Feb;7(2):85–97. PubMed PMID: 16418744.
2. She X, Cheng Z, Zöllner S, Church DM, Eichler EE. Mouse segmental duplication and copy number variation. Nat.Genet. 2008 Jul;40:909–914. PubMed PMID: 18500340.
3. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science. 2008 Jun 20;320(5883):1629–3. PubMed PMID: 18535209.
4. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Gene. 2009;10:451–81. PubMed PMID: 19715442.
5. Quinlan AR, Hall IM. Characterizing complex structural variation in germline and somatic genomes. Trends Genet. 2012 Jan;28(1):43–53. PubMed PMID: 22094265.
6. Sindi SS, Raphael BJ. Identification of Structural Variation, In Maria Poptsova, Genome Analysis: Current Procedures and Applications. Caister Academic Press, pp. 1-19, 2014

7. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014 Jan;42(Database issue):D986–92. PubMed PMID: 24174537.
8. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the Human Genome. Nat Genet. 2004 Sep;36(9):949–51. PubMed PMID: 15286789.
9. Church DM, Lappalainen I, Sneddon TP, Hinton J, Maguire M, Lopez J, Garner J, Paschall J, DiCuccio M, Yaschenko E, Scherer SW, Feuk L, Flicek P. Public data archives for genomic structural variation. Nat Genet. 2010 Oct;42(10):813–4. PubMed PMID: 20877315.
10. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, Chen C, Maguire M, Corbett M, Zhou G, Paschall J, Ananiev V, Flicek P, Church DM. DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. 2013 Jan;41(Database issue):D936–41. PubMed PMID: 23193291.