The Database of Genomic Variants
User Tutorial

http://dgv.tcag.ca/

Updated, June 2013

About the Database

- The objective of the Database of Genomic Variants is to provide a comprehensive summary of structural variation in the human genome. We define structural variation as genomic alterations that involve segments of DNA that are larger than 50bp. The content of the database only represents structural variation identified in healthy control samples.
- The Database of Genomic Variants provides a useful catalog of control data for studies aiming to correlate genomic variation with phenotypic data. The database is continuously updated with new data from peer reviewed research studies.

The specific aims of DGV include:

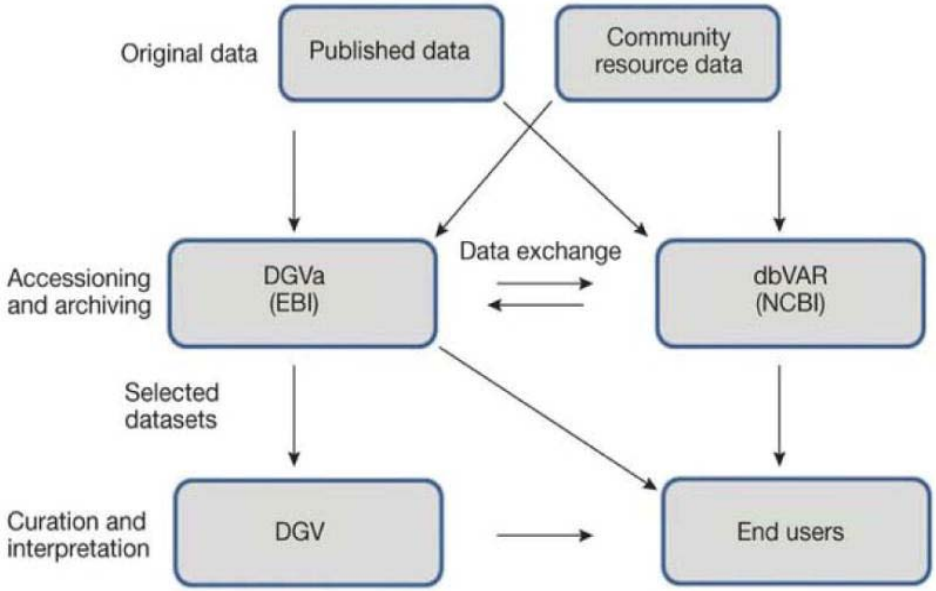
- To provide information about **genomic variants identified in control samples**
- To serve as a **resource** to both **clinical and research labs**
- To show **variation in genomic context** in a simple genome browser
- To transparently provide **access to underlying source data** (downloadable)
- To be **publicly available at no cost**

The “New” DGV

http://dgv.tcag.ca/

- The Database of Genomic Variants (DGV) has been working in partnership with the new database archives at EBI (DGVa; http://www.ebi.ac.uk/dgva/) and NCBI (dbVar; http://www.ncbi.nlm.nih.gov/dbvar) to collect, organize and curate genome-wide information on copy number variation. The goal was to provide a fully integrated, standardized and accessioned set of variants.

- A schematic of how this data is generated and shared between the groups is highlighted here.



(Church DM et. al. Nat Genet. 2010 Oct;42(10):813-4)

Database Content

The database currently contains information from 53 different peer reviewed studies. Corresponding to ~ 265,000 CNV regions and over 1,100 inversions.

The majority of data sets in DGV are from four different types of studies:

- Array based comparative genomic hybridization (CGH) and comparative intensity analysis (SNP/CNV arrays)
- Identification of deletions based on statistical analysis of SNP data
- Clone end sequence mapping
- Sequence trace mapping

What DGV is not…

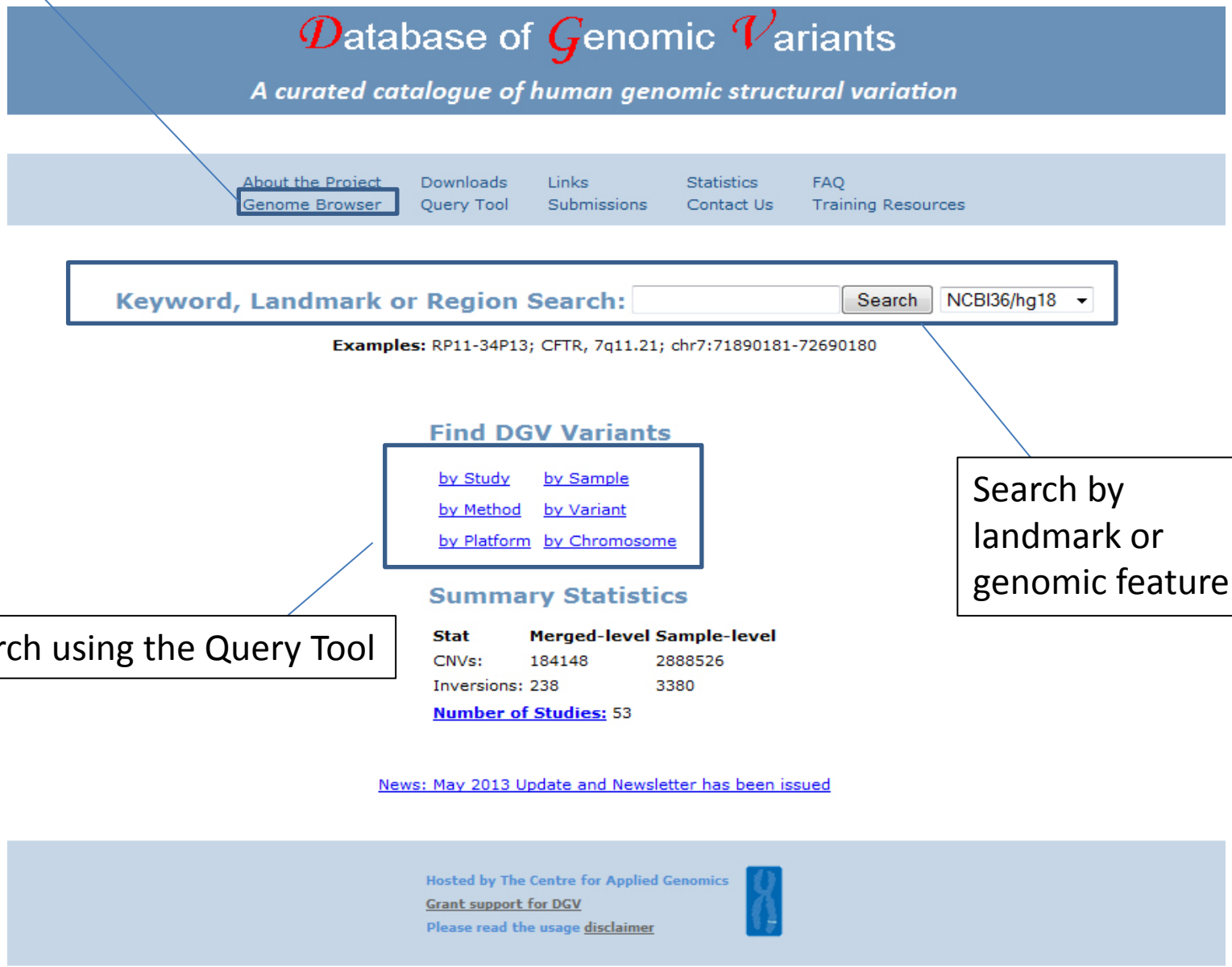
- An uncurated repository of primary data
- A substitute for a well-designed control experiment
  – The database is still limited in content
  – About 6,500 individuals represented; not ethnically matched
- Due to biases in studies to date, the content is almost certainly not an accurate representation of structural variation on a population scale.

DGV Home Page

http://dgv.tcag.ca/dgv/app/home

Enter the Genome Browser

Points of Entry into the Database



Search by landmark or genomic feature

Search using the Query Tool

Genome Browser

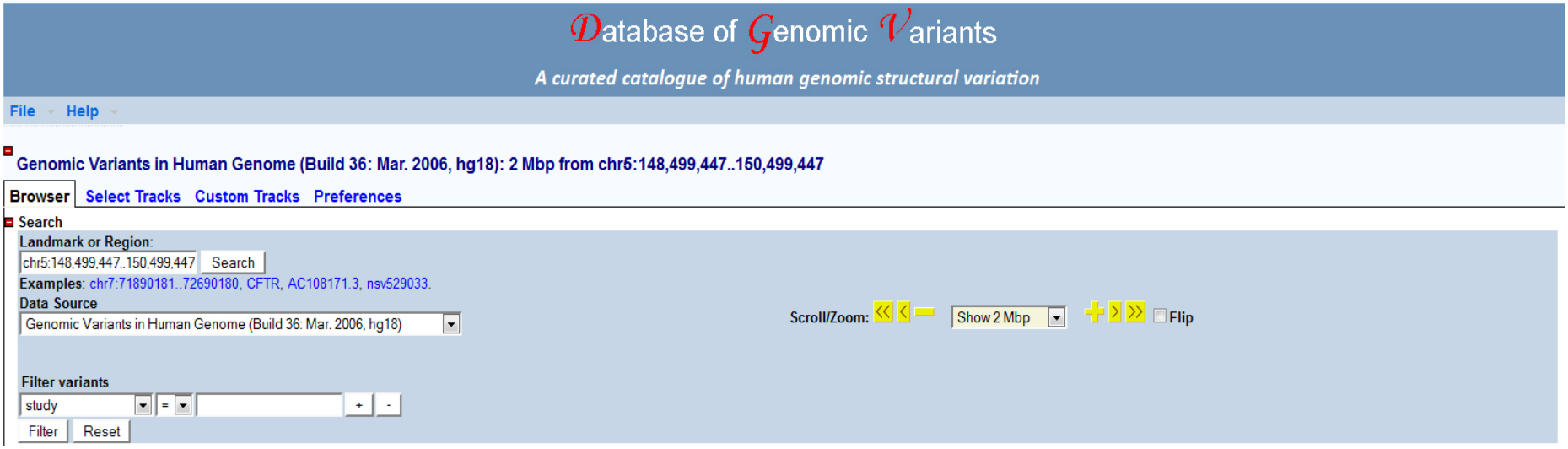
- DGV uses the Generic Genome Browser (http://gmod.org/wiki/GBrowseGMOD) to provide a visual display of structural variation data.

- In addition to the structural variation data, we have provided numerous additional tracks of relevant annotations.

- The following slides will help to provide an introduction to the various features and content in our genome browser.

Genome Browser

Many features of the genome browser are the same, with some functional differences and some new tools/options.

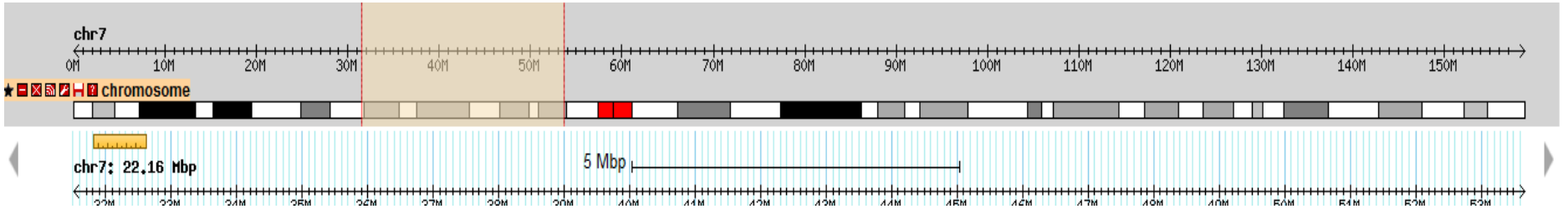


At the top of the browser window, there are options modifying the display in the current browser

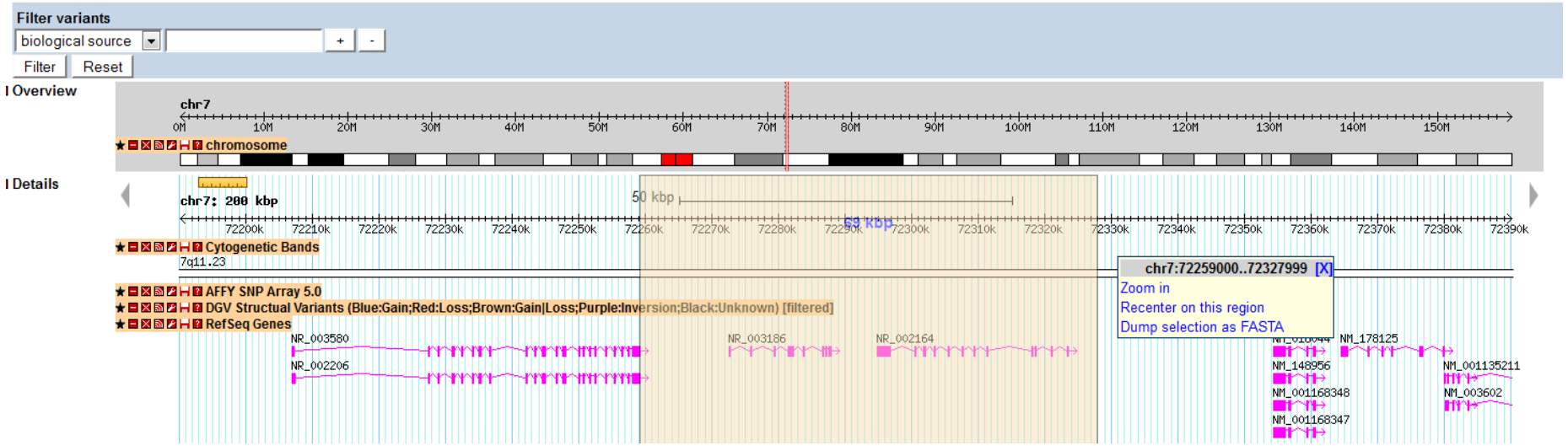
1. To turn on/off various tracks and to modify the display, chose the “Select Tracks” tab.
2. To upload your own custom annotations, or to modify the display, chose the appropriate tabs.
3. To filter the data displayed in the browser, to only show a subset of the variants, use the “Filter Variants” option, select the data type you want to filter by, and add a keyword to the box. To filter by greater than one data type, use the “+” key to add another row.

Genome Browser Options

1. Click and drag on the chromosome to select a particular region

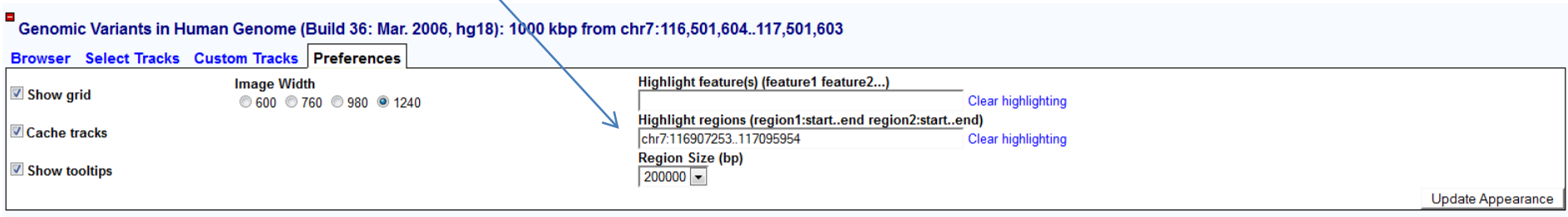


2. Click and drag on the genomic position bar to zoom in or centre on a region of interest.

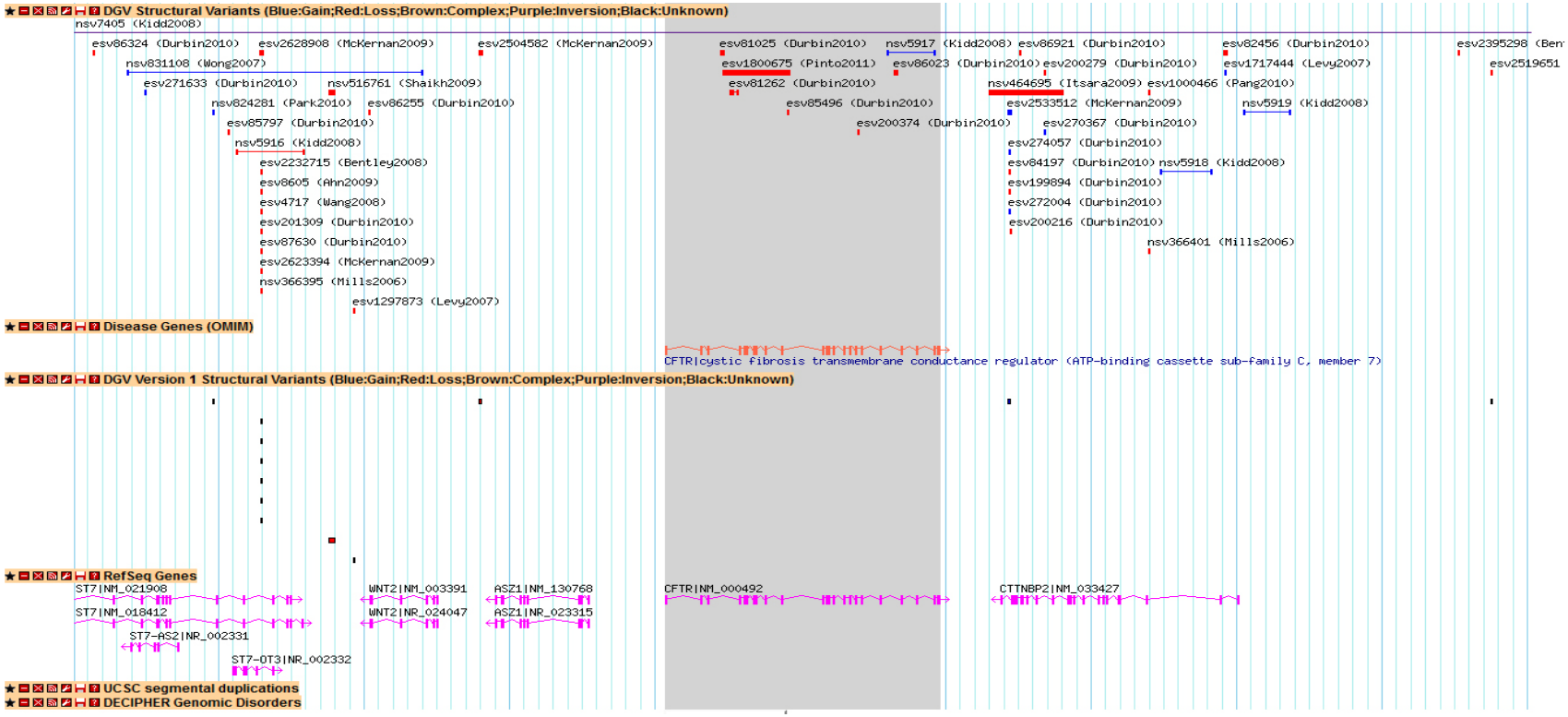


Genome Browser Options

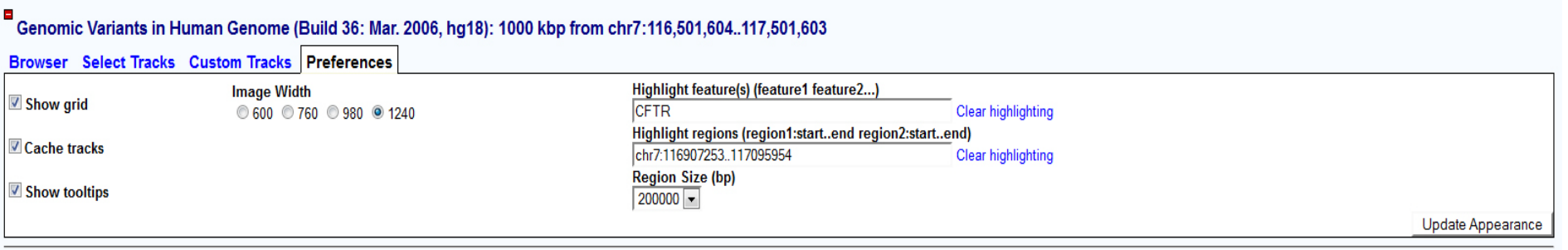
To highlight a region of interest in the browser, select the Preferences tab, and add the target region into this box.



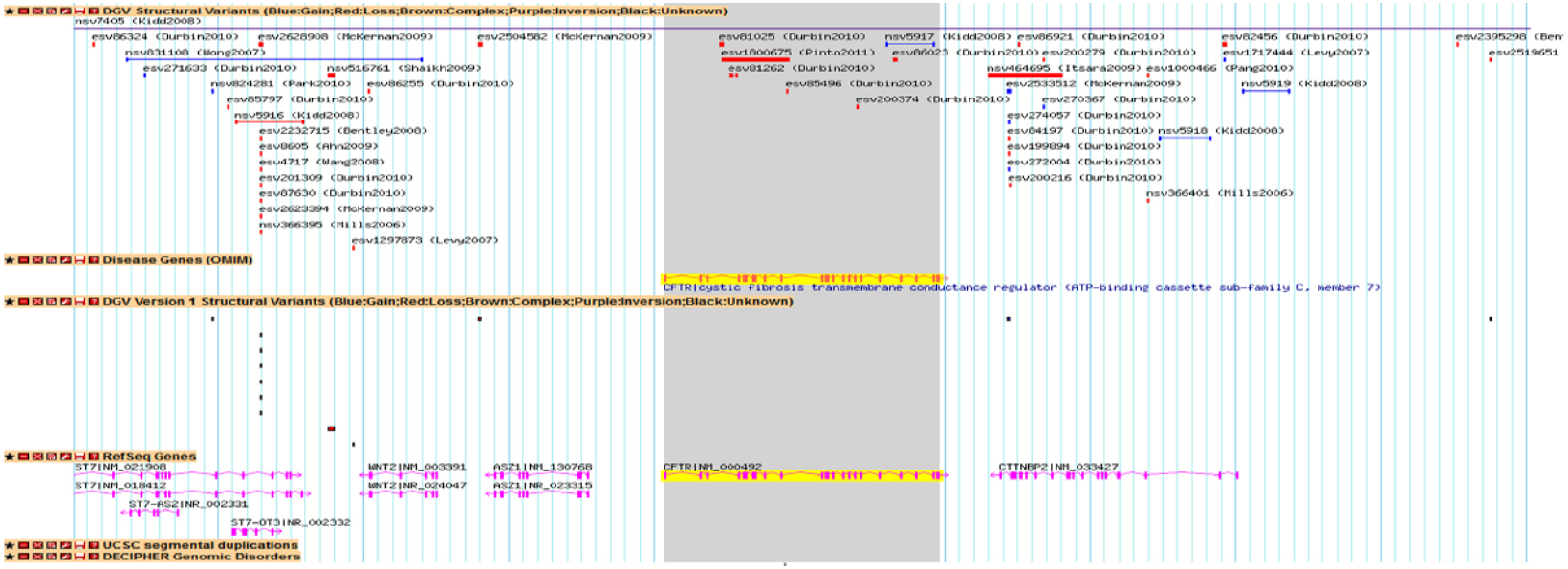
The region will appear as a shaded grey box as seen below.



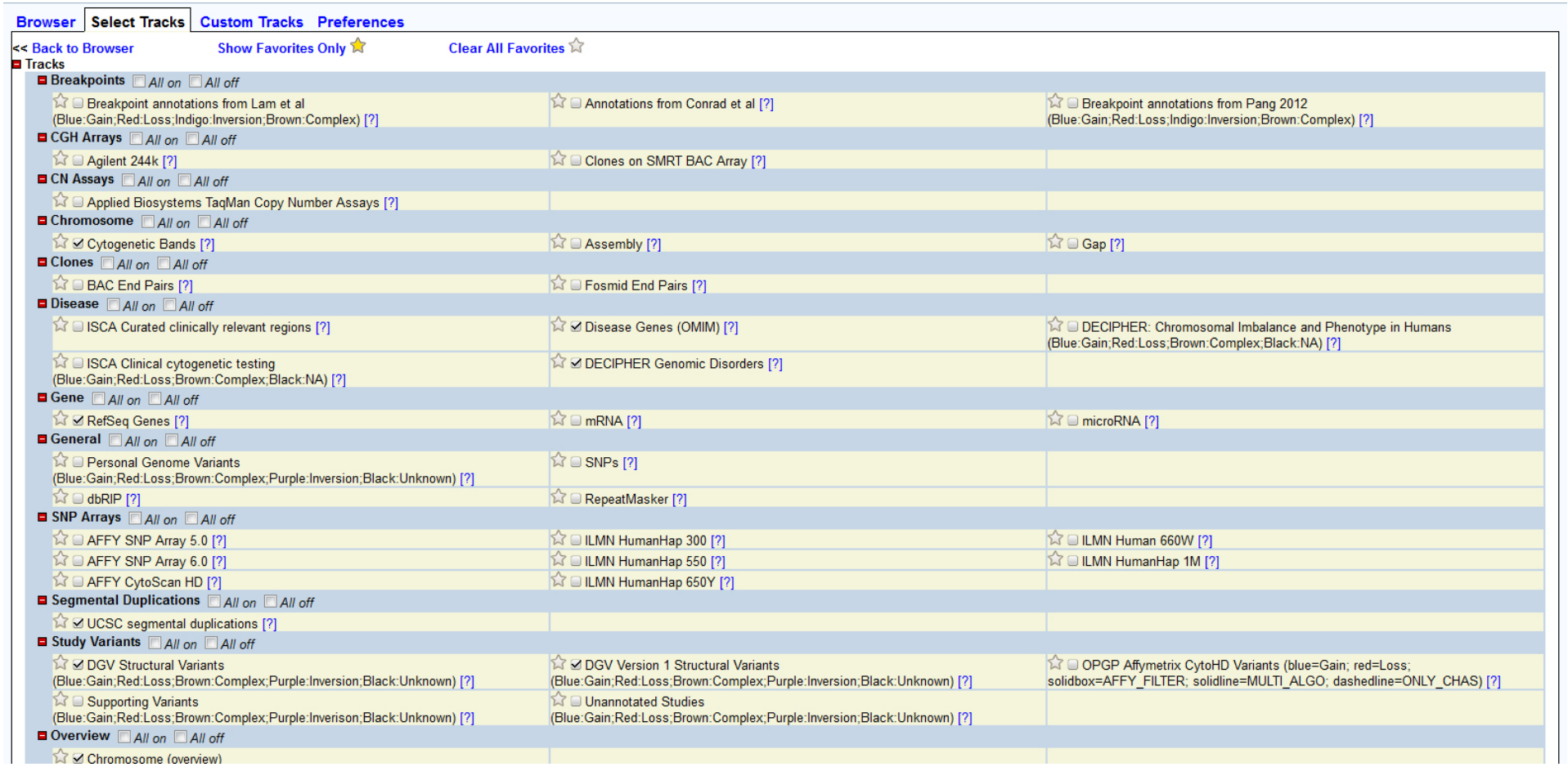
Genome Browser Options



In addition to adding a shaded area to the genome browser, the Preferences tab will allow you to modify the display on your computer by selecting the desired Image Width. You can Also turn the grid pattern on or off, and you may also highlight specific entries in the genome browser that you may be interested in. For example, if you input the term CFTR in the Highlight feature(s) box, you'll get the following display.

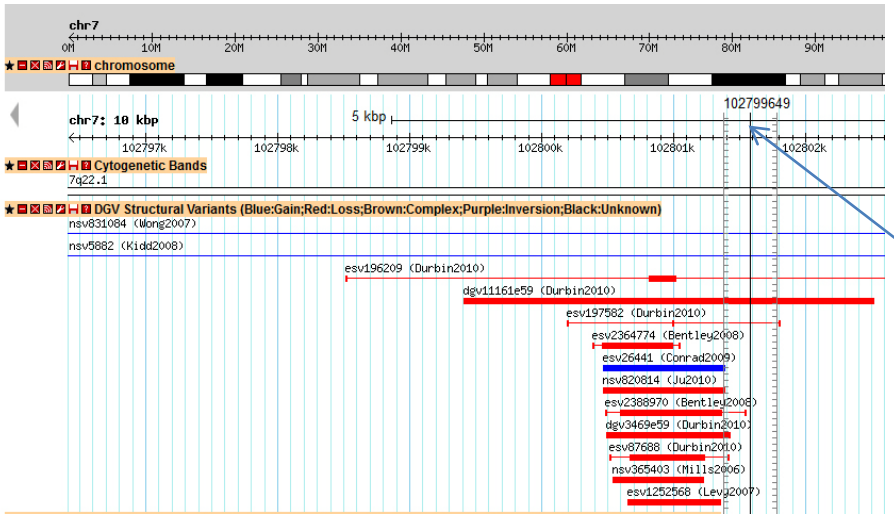


Genome Browser Options

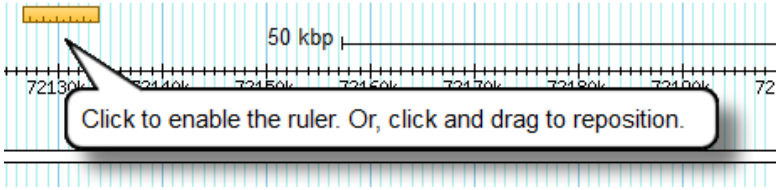


To select the tracks you want shown in the browser, navigate to the “Select Tracks” tab. Check the box beside the track name to turn the track on or off. To turn all the tracks on/off within a group (i.e. Disease), use the “All on” or “All off” options. The “?” beside the tracks will provide information regarding the content of the data displayed in this track.

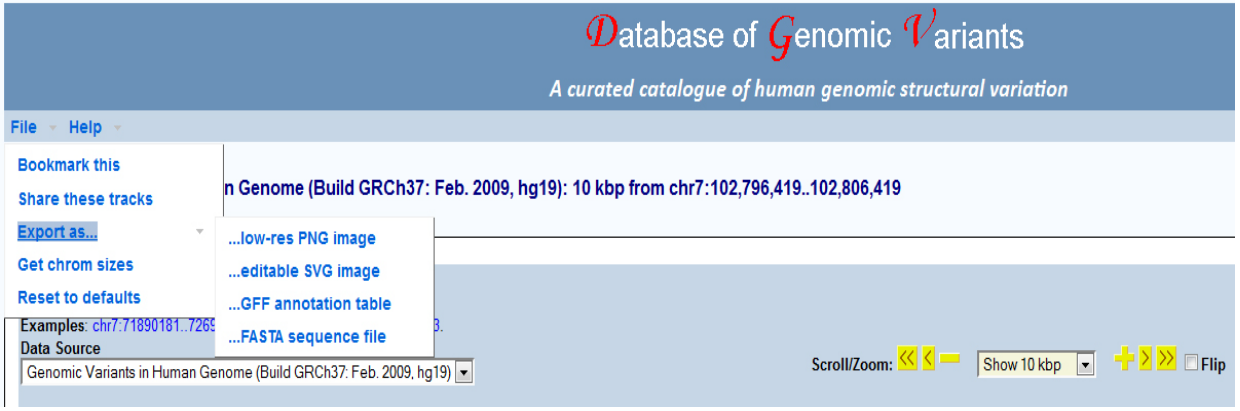
Genome Browser Options



Click on the ruler to add a vertical marker to align features.



File option allows users to export graphics for Figures, and can access GFF tables.



Genome Browser Options
Track Options

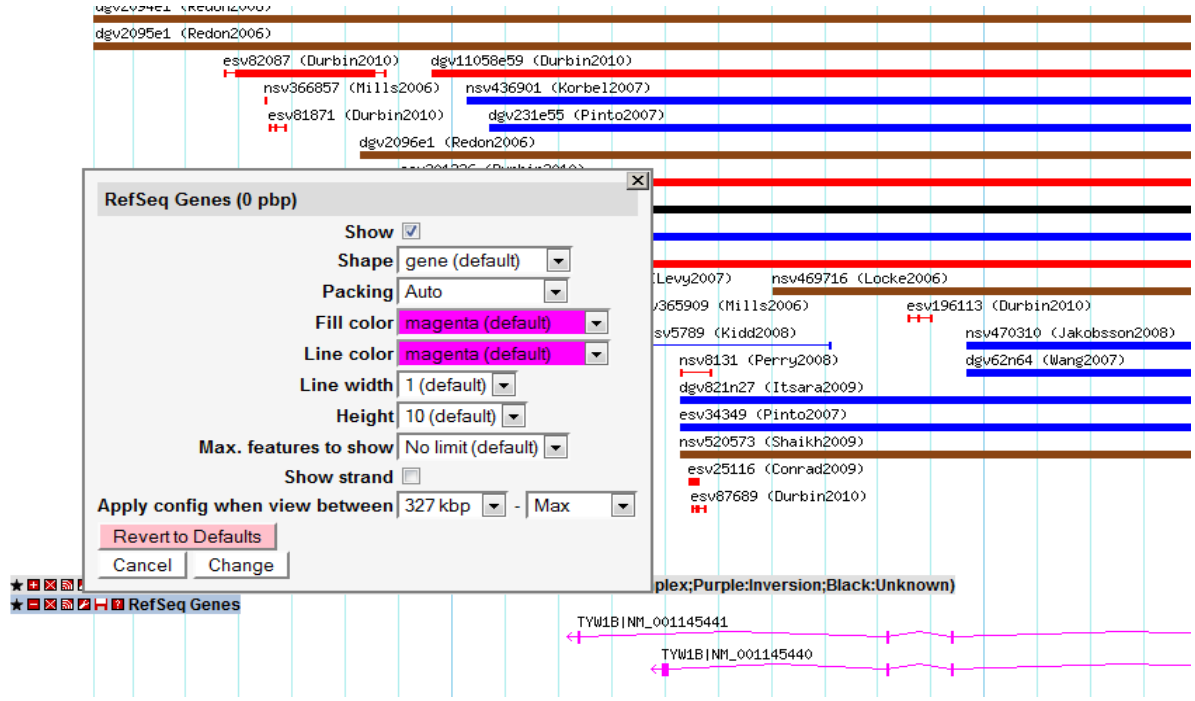
Minimize the track

Displays information about the data
In the track.

Close the track.

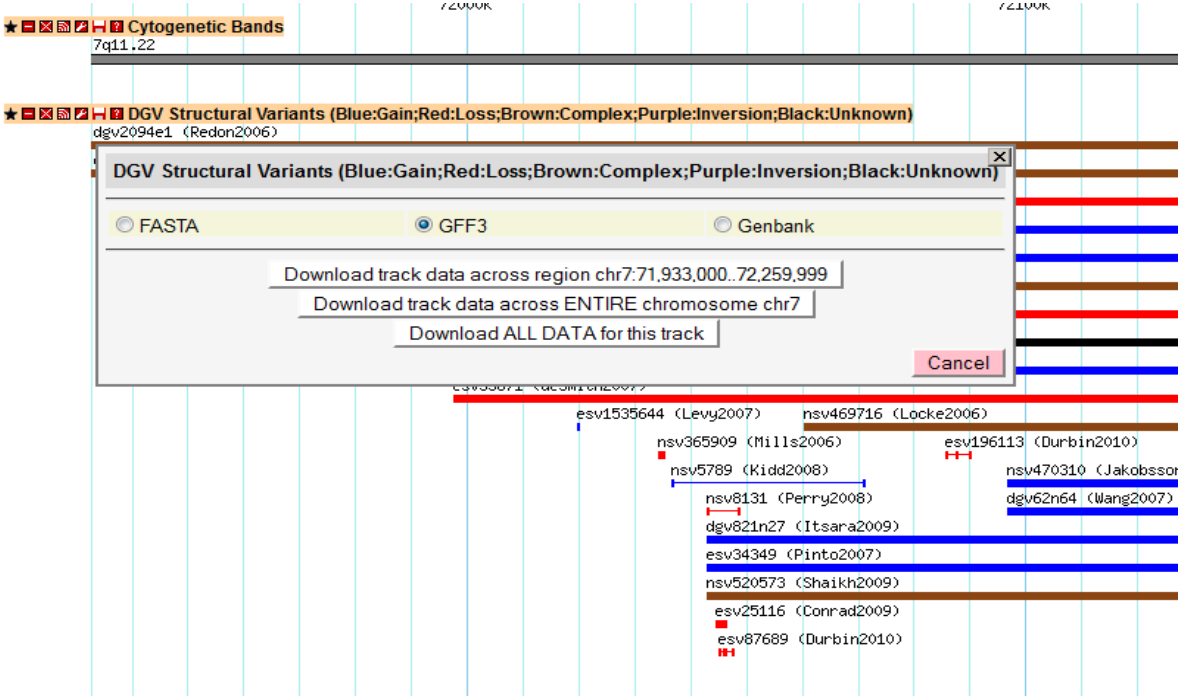
Save the data from this track
This allows the user to download or save the underlying data as a GFF file.
User may select the region displayed in the window, the whole chromosome or genome wide (see next slide).

Allows users to change how the data are displayed (see next slide).



Change Display
(click on wrench icon)

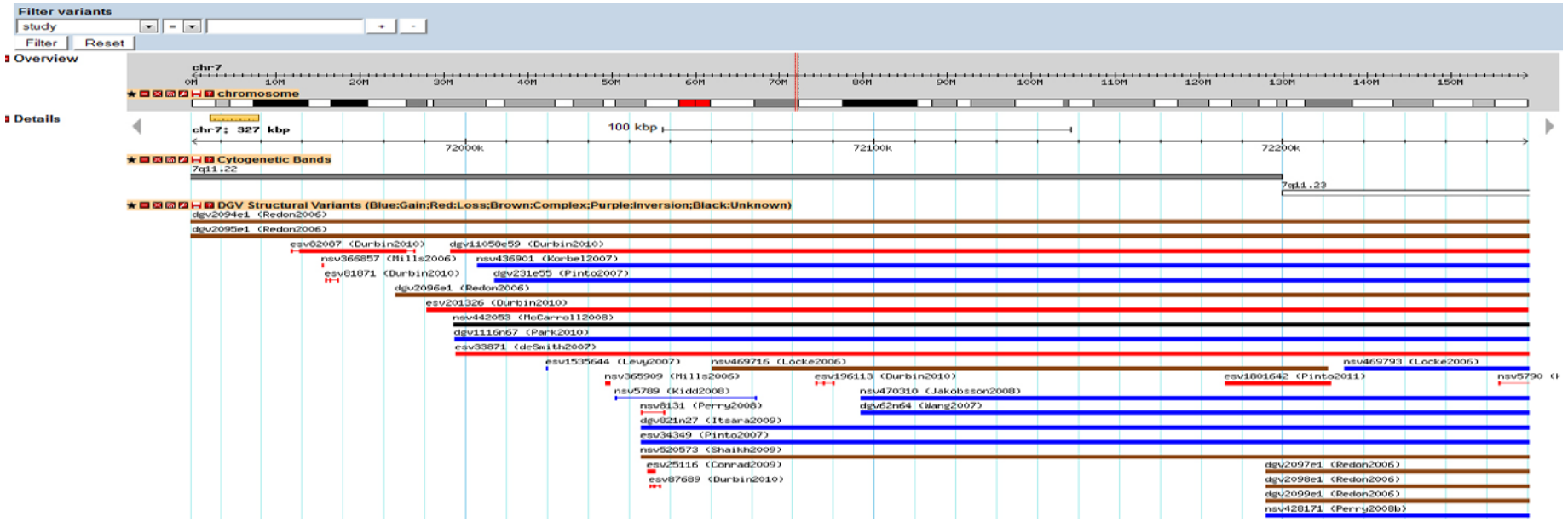
Example:
-can force option to show strand/direction for genes regardless of the size of the window being used.



Save the Data
(click on “save icon”)

Example:
-quick and easy method to generate a list of genes or other items that fall in your region of interest.

Applying Filters to DGV Variants in Gbrowse



If you were only interested in comparing your data to deletions found in DGV, you can filter the data displayed in the browser by selecting the option in the “Filter Variants” box at the top.

DGV Content

- With the release of the new DGV, there are a number of updates and changes to the content and representation of the data.
- Stable, long term accessions have been assigned and a controlled, and structured vocabulary of terms used to describe the data have been developed.
- The following slides will help introduce DGV users to some of these changes.

New DGV Accessions

- Each study from DGV has been accessioned by one of the two groups; dbVAR have assigned nsv/nssv accessions, while DGVa has assigned esv/essv accessions. An nsv is an NCBI structural variant, and an nssv is an NCBI supporting structural variant. An esv is an EBI structural variant, and an essv is an EBI supporting structural variant.
- Supporting structural variants ("ssv") are typically sample level variants, where each ssv represents the variant called in a single sample/individual. In a few studies the ssv represents the variant called by a single algorithm. If multiple algorithms were used, overlapping ssv's from the same individual would be combined to generate a sample level sv. If there are many samples analysed in a study and if there are many samples which have the same variant, there will be multiple ssv's with the same start and end coordinates. These sample level variants are then merged and combined to form a representative variant that highlights the common variant found in that study. This is called a structural variant ("sv") record.
- DGV has always provided this type of summary/merged variant and we have continued to do so in cases where there are a number of overlapping variants that are almost identical, but may be slightly different due to the inherent variability between experiments. If there are clusters of variants within a single study that share at least 70% reciprocal overlap in size/location, we will merge these together and provide an accession record that has our internal "dgv" prefixed identifier. The dgv merged identifier has been updated to help improve the consistency and stability across updates of the database. The format of the dgv accession is now "dgv + variant number + study accession. As an example, the first merged variant from the Shaikh et al 2009 study (study accession=nstd21) would be dgv1n21. The second merged variant would be dgv2n21 and so forth.

Examples of DGV Variant Types

Example 1

DGV merged variant
-merge of nsv/esv variant regions which overlap

dgv333n21 = nsv111+ nsv222
Samples:NA15510+NA10851+NA12291+NA18291

nsv/esv = variant regions
-merge of nssv/essv variant calls which overlap

nsv111 = nssv123+ nssv321
Samples:NA15510+NA10851

nsv222 = nssv567+ nssv765
Samples:NA12291+NA18291

nssv/essv = variant calls
-represent the supporting or sample level calls

nssv123 : Sample=NA15510

nssv321 : Sample=NA10851

nssv567 : Sample=NA12291

nssv765 : Sample=NA18291

Examples of DGV Variant Types

Example 2

DGV merged variant
-merge of nsv/esv variant regions which overlap

No overlapping nsv/esv variants
No DGV merged variant is created

nsv/esv = variant regions
-merge of nssv/essv variant calls which overlap

nsv482 = nssv164+ nssv955
Samples:NA15581+NA12761

nssv/essv = variant calls
-represent the supporting or sample level calls

nssv164 : Sample=NA15581

nssv955 : Sample=NA12761

Examples of DGV Variant Types

Example 3 (variant called in only 1 sample)

DGV merged variant
-merge of nsv/esv variant regions which overlap

No overlapping nsv/esv variants
No DGV merged variant is created

---

nsv/esv = variant regions
-merge of nssv/essv variant calls which overlap.
-there are no overlapping variants therefore the variant region is the same as the variant call.

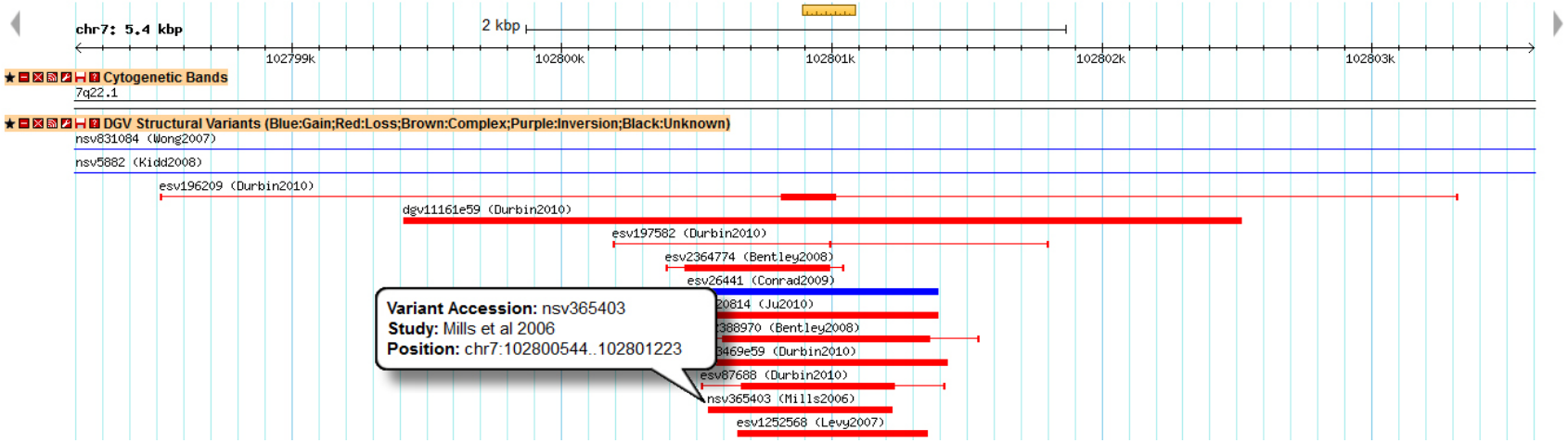
nsv372 = nssv862
Samples:NA13781

---

nssv/essv = variant calls
-represent the supporting or sample level calls.
-in this case, the variant was only detected in one sample.

nssv862 : Sample=NA13781

DGV data



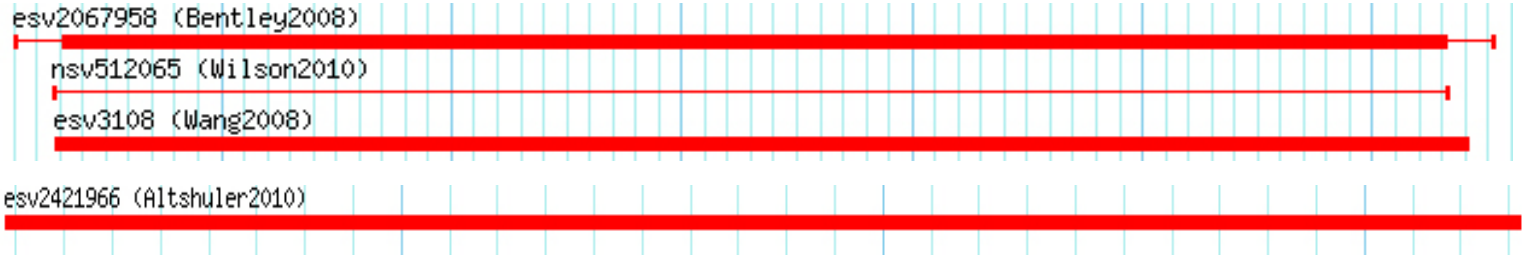
Internal ID	117790			
Landmark				
Location Information	**Type**	**Coordinates**	**Assembly**	**Other Links**
		chr12:109632315..109632415	hg19	UCSC Ensembl
		chr12:108116698..108116798	hg18	UCSC Ensembl
Cytoband	12q24.11			
Allele length	**Assembly**	**Allele length**		
	hg19	101		
	hg18	101		
Variant Type	CNV Deletion			
Copy Number				
Allele State	Homozygous			
Allele Origin	Not tested			
Probe Count				
Merged Status	S			
Merged Variants	esv1456600			
Supporting Variants				
Samples	HuRef			
Known Genes	ACACB			
Method	SNP_array			
Analysis	analysis_type=Sequence Alignment			
Comments				
Reference	Levy et al 2007			
Pubmed ID	17803354			
Accession Number(s)	essv3652338			
Frequency	**Sample Size**	2		
	Observed Gain	0		
	Observed Loss	1		
	Observed Complex	n/a		
	Frequency	n/a		

Users can mouse over the variations to obtain summary information, or can click on the image to go to the variants detail page (below).

Additional details such as the allele length, number of gains/losses and the allele state (heterozygous/homozygous) have been added (when available).

DGV data

There are a few different types if images used to display structural variants in the genome browser. The different types are shown below with a corresponding description.

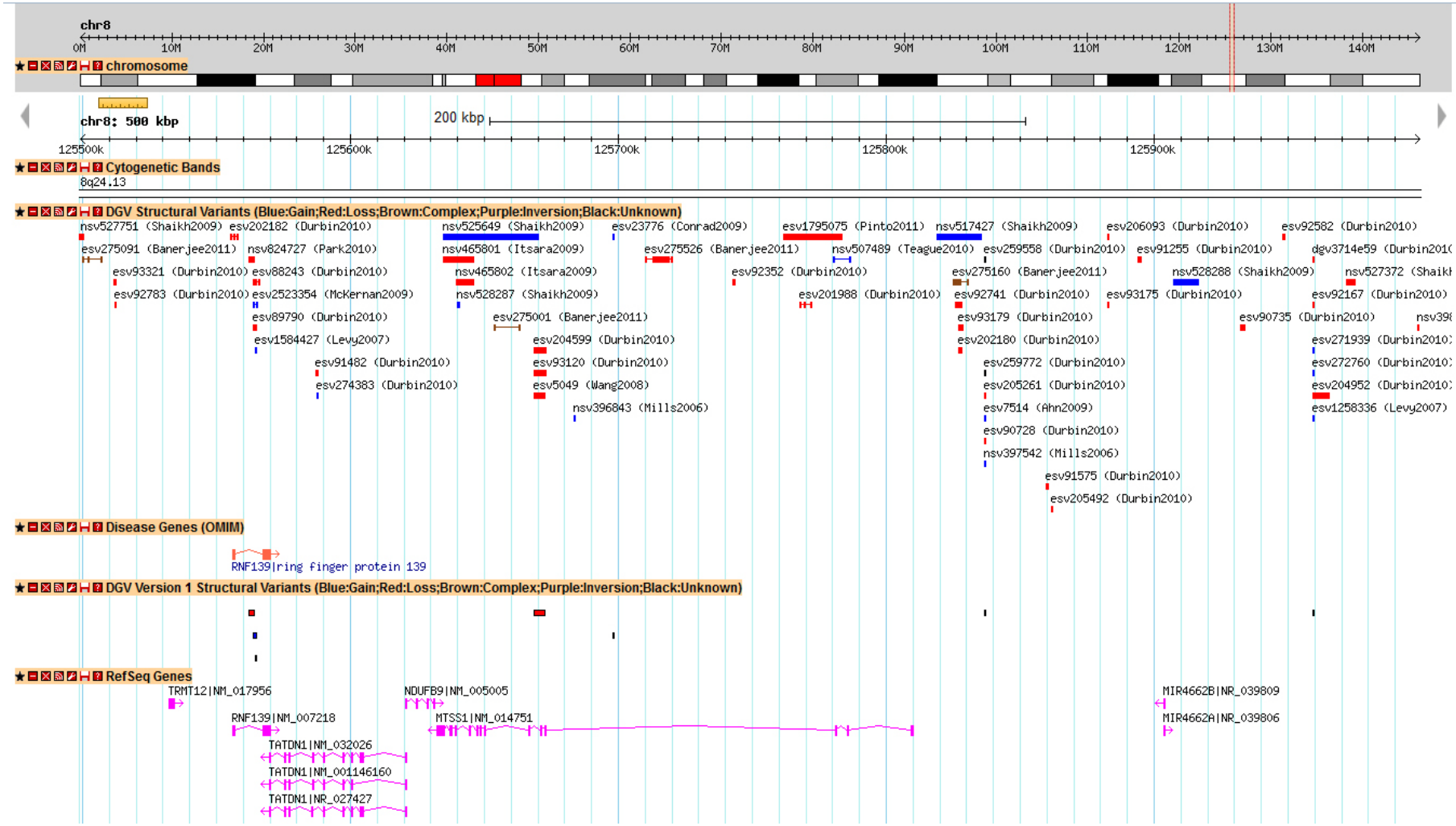


To accurately reflect the inherent differences in the resolution of different approaches, the assignment of boundaries for the structural variants has been updated and the display has been updated.

1. Variant boundaries may be assigned a start and stop position. This will be common for sequencing based studies where the actual breakpoints are known (example = esv3108; Wang2008).
2. An outer start and outer stop coordinate will be assigned for studies that use a mapping based strategy (paired-end, optical mapping) or BAC clone approach where the variant boundaries are likely overestimated, and the maximum or outer boundaries are known, but the actual variant likely resides somewhere within this region. (example = nsv512065; Wilson2010, Paired-end Mapping)
3. The inner start and inner stop coordinates are used for studies where the boundaries are likely underestimated and may include oligo (probe) based CGH experiments. The actual boundary of the variant would likely reside somewhere between the last positive probe on the array and the next neighbouring negative probe (example = esv2421966; Altshuler2010).
4. For some studies, a combination of outer start-outer stop and inner start-inner stop coordinates are described when information is available on the boundary regions (example = esv2067958, Bentley2008).

DGV data

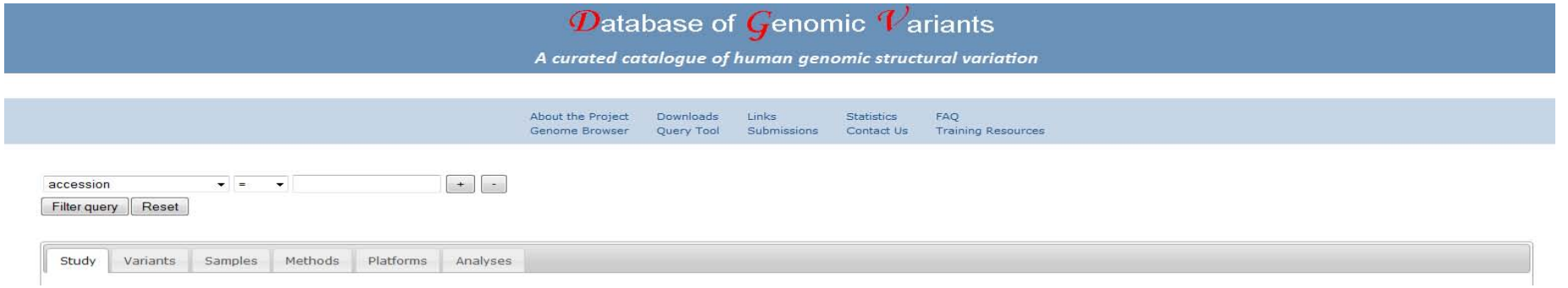
The original (DGV Version 1) and new DGV structural variation data are displayed for comparison.



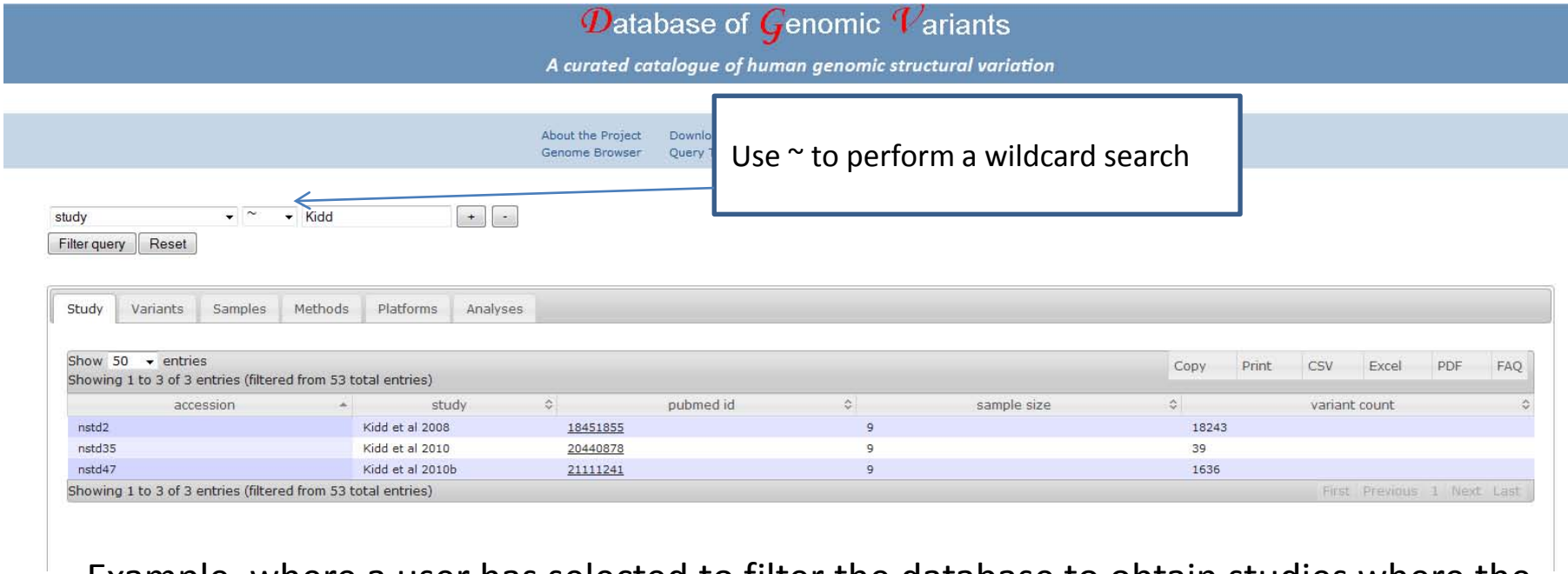
Query Tool

- The query tool is a set of inter-related tables containing all the data from the studies included in DGV. Options to search and filter the data across studies has been developed.
- Each table represents a specific category of data for all of the studies in the database.
  - The Study table contains general information about each study, including the number of samples and variants detected, with links to the published report in PubMed.
  - The Variant Table contains a list of all the variants reported in all the studies and their respective attributes.
  - The Sample Table contains information about the samples that were analysed in each study, including gender, ethnicity and source (where available).
  - The Method Table contains summary information on the approaches used, including details on the specific methods employed.
  - The Platform Table has specific information on the experimental approaches taken in each study used to identify and validate variants.
  - The Analysis Table has detailed information about the tools and algorithms used to analyse the data when generating the set of structural variants
- Options to select, filter and manage the data have been developed, providing the option to customize the output based on a number of terms and attributes.
- Searches and filters can be applied to view information from only a single study, or can be applied across studies to find all information related to a specific term. For example you could retrieve all variants from all studies identified in a single sample. Similarly, if you would like to obtain all the variants identified in a specific population (HapMap Yoruba), or derived from a single approach this would be an option as well.

Query Tool



Users can filter data based on the options presented in the tabs above.

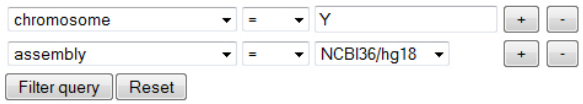


Example, where a user has selected to filter the database to obtain studies where the primary author is “Kidd”.

Use the “+” button to add additional terms
Use the “-” button to remove search terms

Query Tool

Users can save, copy or print the output using these options.



Study | Variants | Samples | Methods | Platforms | Analyses

Show 50 entries
Showing 1 to 50 of 3,507 entries (filtered from 6,407,415 total entries)

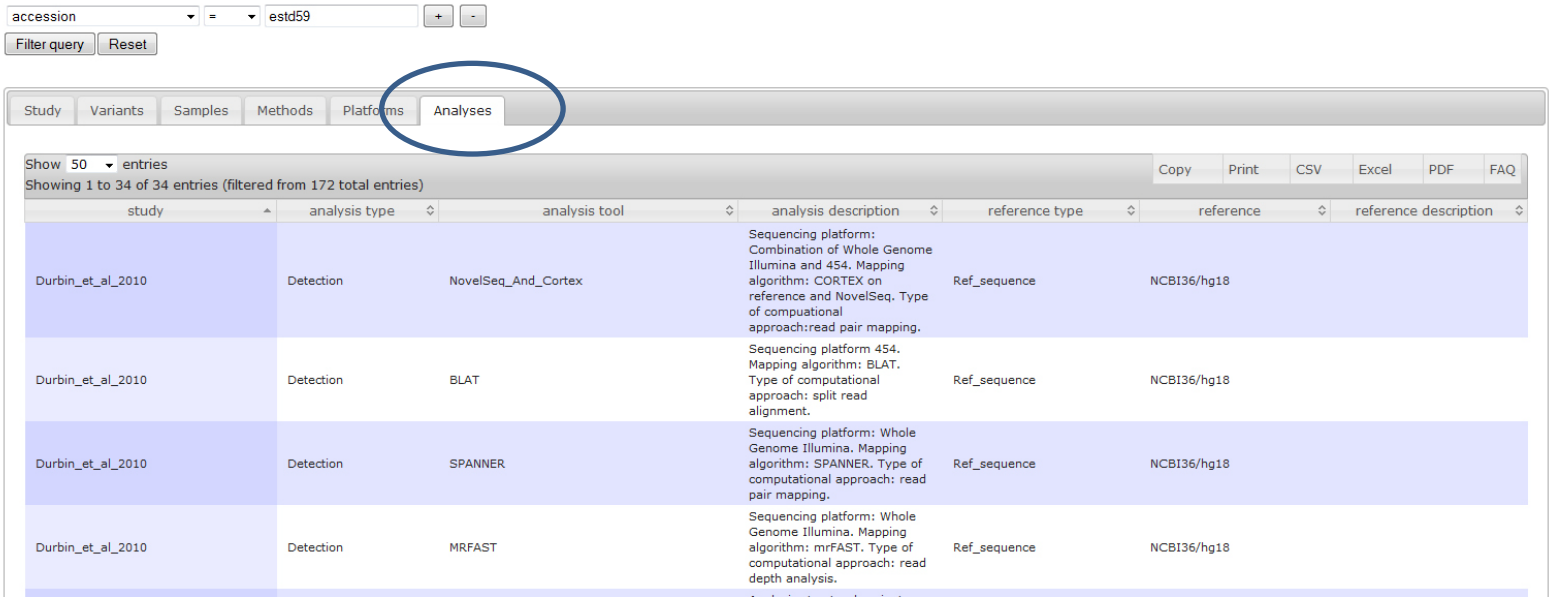
Copy | Print | CSV | Excel | PDF | FAQ

chromosome	start	stop	study	variant accession	M=Merged, S=Sample Call	ethnicity	gender	assembly	variant type	variant subtype
Y	12101151	12209550	Durbin_et_al_2010	essv287618	S	YRI	Male	NCBI36/hg18	CNV	Loss
Y	10519992	10654530	Durbin_et_al_2010	essv287075	S	YRI	Female	NCBI36/hg18	CNV	Loss
Y	11926654	11943340	Ahn et al 2009	essv31307	S	KOREAN	Male	NCBI36/hg18	CNV	Loss
Y	57393261	57399506	Durbin_et_al_2010	essv287096	S	CEU	Female	NCBI36/hg18	CNV	Loss
Y	57229393	57236795	Durbin_et_al_2010	essv2301422	S	CEU	Male	NCBI36/hg18	CNV	Loss
Y	2391693	2402131	Durbin_et_al_2010	essv287759	S	CEU	Male	NCBI36/hg18	CNV	Loss
Y	22973854	23049013	Teague et al 2010	nsv508822	S			NCBI36/hg18	CNV	Insertion
Y	5103369	5104505	Durbin_et_al_2010	essv287769	S	CEU	Female	NCBI36/hg18	CNV	Loss
Y	12111044	12115461	Ahn et al 2009	essv28495	S	KOREAN	Male	NCBI36/hg18	CNV	Loss
Y	10602703	10604281	Durbin_et_al_2010	esv252015	M			NCBI36/hg18	CNV	Loss
Y	20837530	20837530	Levy et al 2007	essv4016907	S	Caucasian	Male	NCBI36/hg18	CNV	Insertion
Y	57404391	57404705	Durbin_et_al_2010	essv287111	S	CEU	Female	NCBI36/hg18	CNV	Loss
Y	21745263	21777102	Durbin_et_al_2010	esv151872	S			NCBI36/hg18	CNV	Loss
Y	11921334	11944992	Ahn et al 2009	essv32152	S	KOREAN	Male	NCBI36/hg18	CNV	Loss
Y	57264599	57265327	Durbin_et_al_2010	essv2301601	S	CHB	Male	NCBI36/hg18	CNV	Loss
Y	57394520	57395721	Durbin_et_al_2010	dgv12705e59	M			NCBI36/hg18	CNV	Loss
Y	57402930	57403514	Durbin_et_al_2010	essv287473	S	CEU	Female	NCBI36/hg18	CNV	Loss
Y	57398842	57403535	Durbin_et_al_2010	esv152121	S			NCBI36/hg18	CNV	Loss
Y	11924997	11942375	Ahn et al 2009	esv8763	S	KOREAN	Male	NCBI36/hg18	CNV	Loss
Y	57263602	57264731	Durbin_et_al_2010	essv2301329	S	YRI	Male	NCBI36/hg18	CNV	Loss
Y	10630453	10630629	Durbin_et_al_2010	essv287220	S	CEU	Female	NCBI36/hg18	CNV	Loss
Y	20849751	20852550	Durbin_et_al_2010	esv151835	S			NCBI36/hg18	CNV	Loss

Another example where the user has selected variants on chromosome Y, mapped to assembly version hg18. To filter across all tables, select the “Filter query” button. If you then select the “Study” tab, only studies which have identified variants on the Y chromosome are included in the results.

Query Tool

Additional details on the platforms used, and the specific types of analyses performed are available as shown for the 1,000 genomes pilot dataset (Durbin2010).



Help and Support

http://dgv.tcag.ca/dgv/app/contacts

If you have any questions while using the DGV beta site, or if you notice any errors or bugs in the database, please contact the DGV Team at your earliest convenience.

Email: dgv-contact@sickkids.ca.

If you would like to receive updates and notifications about DGV, please sign up for our newsletter

Thank you for taking the time to test and use the database, your assistance and feedback is greatly appreciated.

Sincerely,
The DGV Team

DGV Inter-Operates With Other Genome Databases

- **Other sites displaying data from DGV:**

DECIPHER

Ensembl

UCSC

HapMap

GeneCards