The Database of Genomic Variants

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

FGED Toronto 2012
Jeff MacDonald

BETA

Database of Genomic Variants

BETA

A curated catalogue of human genomic structural variation

Keyword, Landmark or Region Search: Search NCBI36/hg18

Examples: RP11-34P13; CFTR, 7q11.21; chr7:71890181-72690180

Find DGV Variants

by Study | by Sample
by Method | by Variant
by Platform | by Chromosome

Summary Statistics

Stat	Merged-level	Sample-level
CNVs:	179450	326620
Inversions:	459	1330

Number of Studies: 37

News: April 2012 Update and Newsletter has been issued

Hosted by The Centre for Applied Genomics
Grant support for DGV

Large-Scale Copy Number Polymorphism in the Human Genome

Jonathan Sebat,[1] B. Lakshmi,[1] Jennifer Troge,[1] Joan Alexander,[1] Janet Young,[2] Pär Lundin,[3] Susanne Månér,[3] Hillary Massa,[2] Megan Walker,[2] Maoyen Chi,[1] Nicholas Navin,[1] Robert Lucito,[1] John Healy,[1] James Hicks,[1] Kenny Ye,[4] Andrew Reiner,[1] T. Conrad Gilliam,[5] Barbara Trask,[2] Nick Patterson,[6] Anders Zetterberg,[3] Michael Wigler[1]*

www.sciencemag.org SCIENCE VOL 305 23 JULY 2004

BRIEF COMMUNICATIONS

nature genetics

com/naturegenetics

Detection of large-scale variation in the human genome

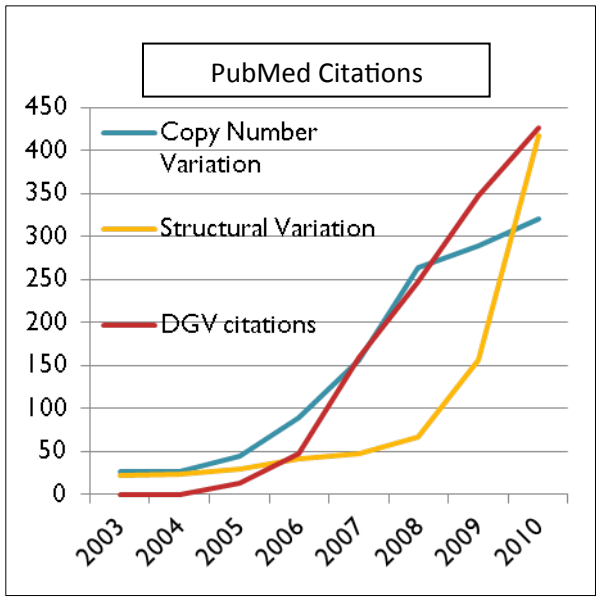
A John Iafrate[1,2], Lars Feuk[3], Miguel N Rivera[1,2], Marc L Listewnik[1], Patricia K Donahoe[2,4], Ying Qi[3], Stephen W Scherer[3,5] & Charles Lee[1,2,5]

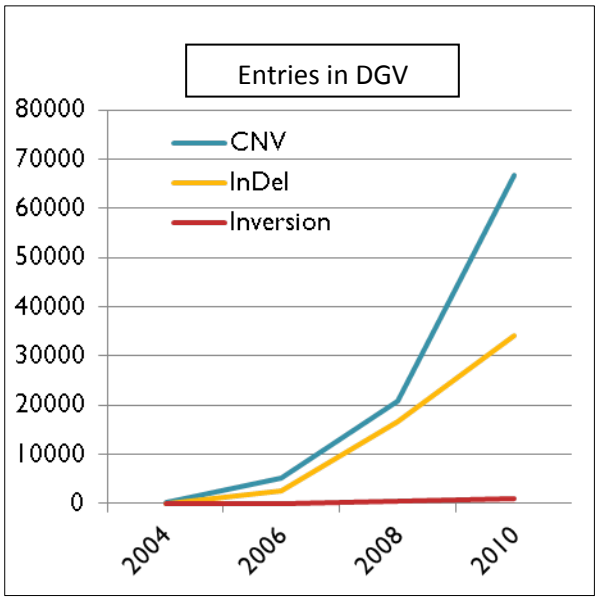
identified LCVs may result from gains or losses involving as much as 2 Mb of DNA sequence (1 Mb to each flanking clone). We identified 102 LCVs (41%) that occurred in more than one individual and 24 LCVs that were present in >10% of the individuals studied. The remaining 153 clones may represent LCVs that occur at lower frequencies. The genomic regions that we identified probably do not represent false positives, because control self-versus-self hybridization experiments

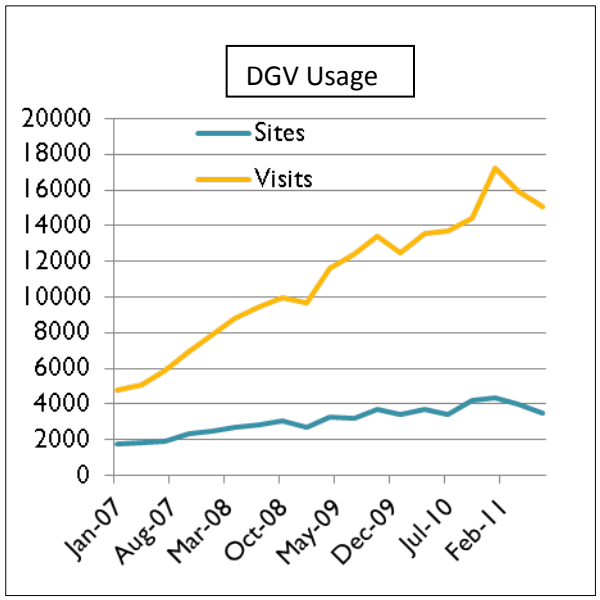
The Database of Genomic Variants

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

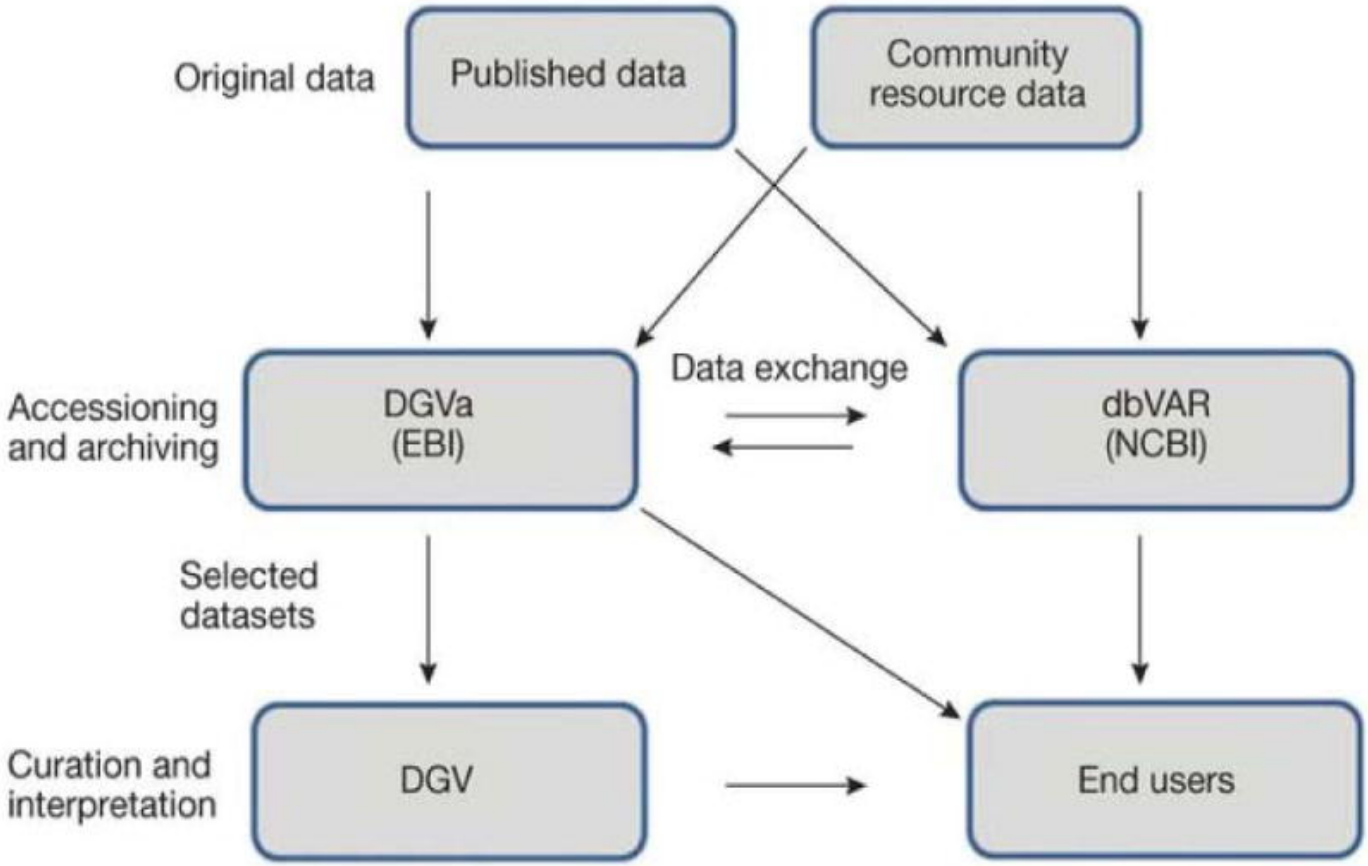
- The Database of Genomic Variants (DGV) has been working in partnership with the new database archives (DGVa and dbVar).
- In 2008, a collaboration between TCAG and EMBL-EBI, was established to collect, organize and curate genome-wide information on copy number variation. The goal was to expand the data curation capabilities to accommodate higher resolution array and sequencing data.





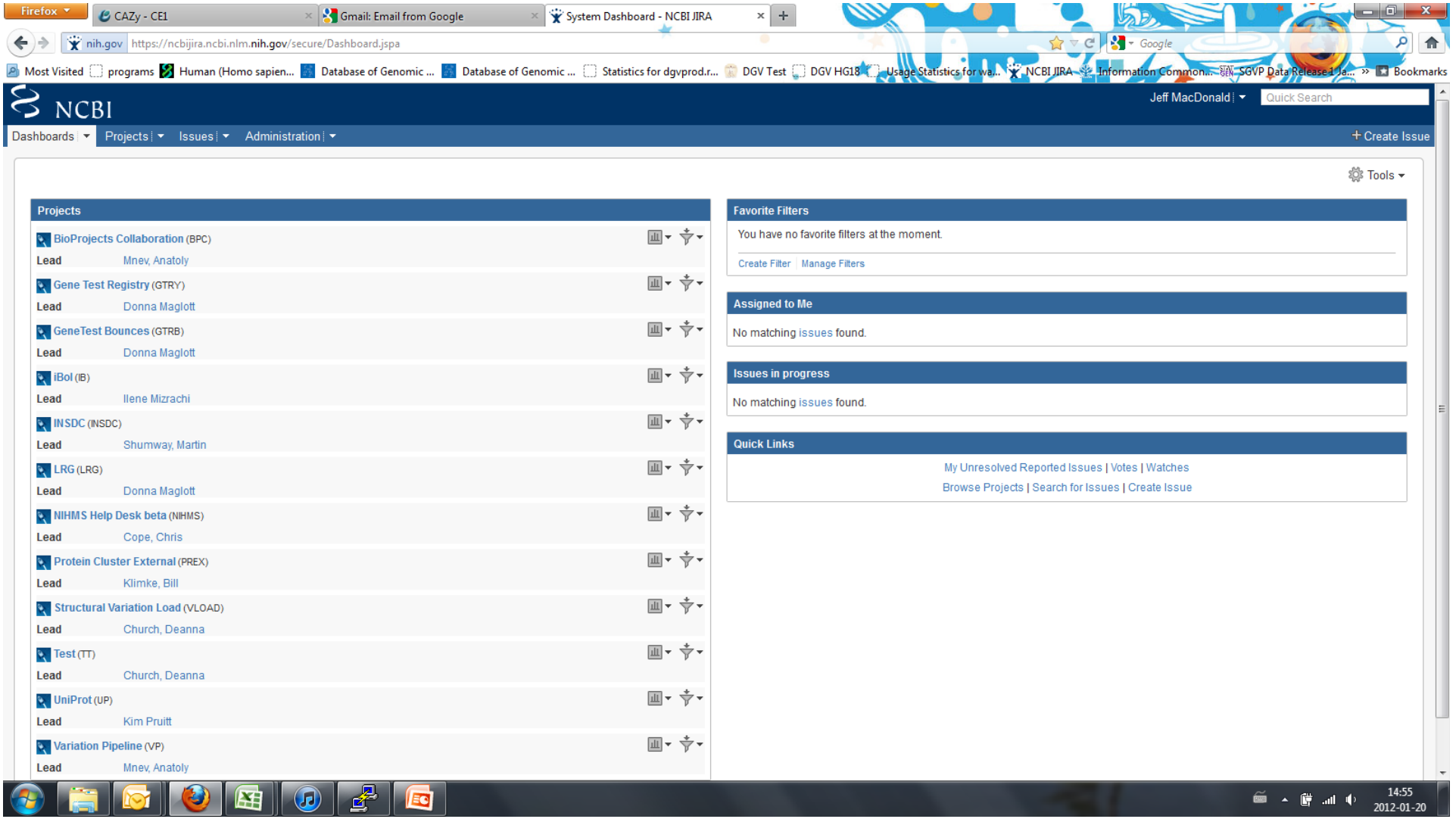


Public data archives for genomic structural variation

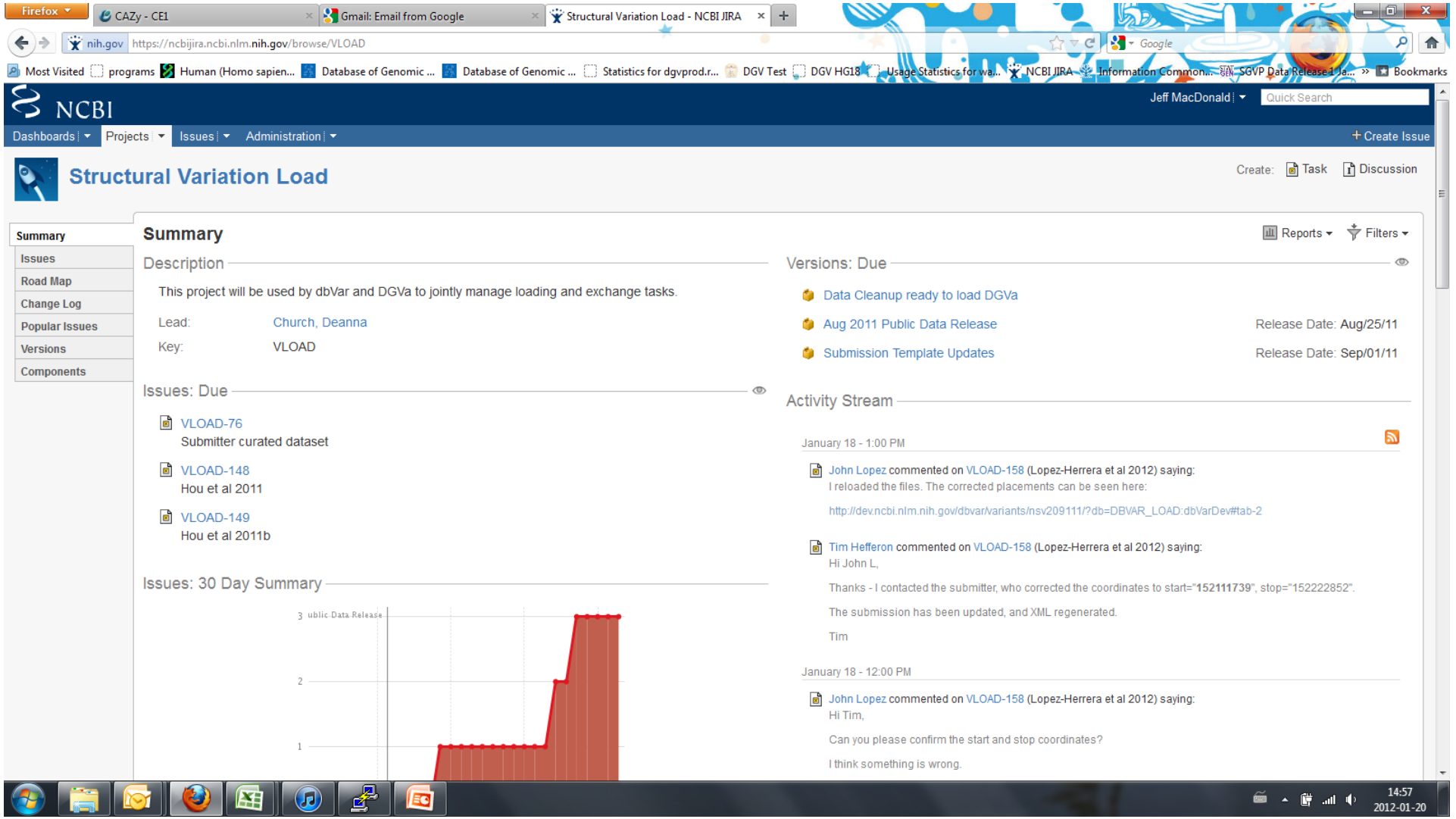


Church DM et. al. Nat Genet. 2010 Oct;42(10):813-4

Data Management: NCBI JIRA Portal



Study Tracking: NCBI JIRA Portal



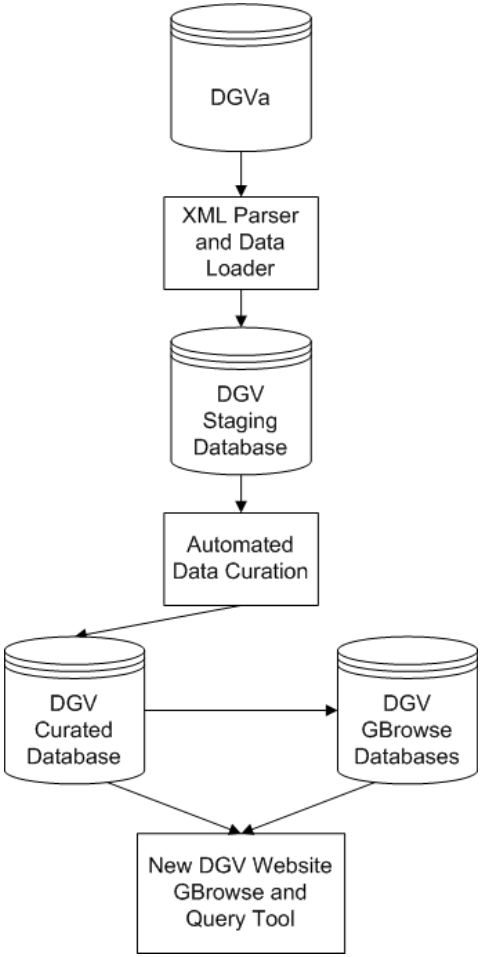
Automation, Standardization and Integration

-Developed tools and pipelines to import, format and load studies into DGV

-Shared data structure, vocabulary and terminology between all three sites

-Interactive links between each database (DGV, DGVa and dbVAR)

Data Flow Diagram



Standardized Submission Template

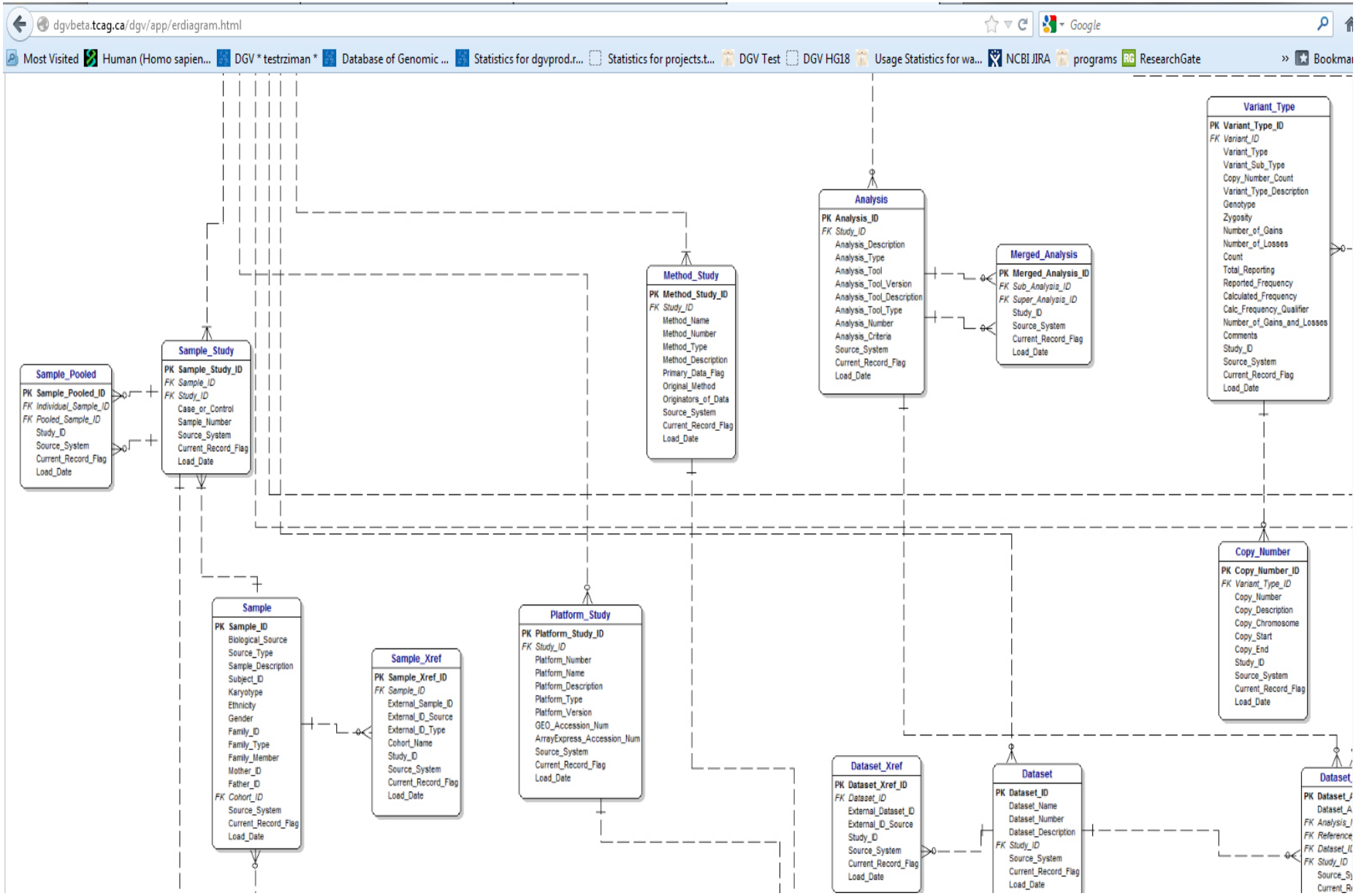
	A	B	C	D
1	**STUDY**			
2	**FIELD**	**TO BE COMPLETED BY SUBMITTER**	**NOTES**	
3	**study_id**		A study identifier in the format **AuthorYear** where **Author** is the last name of the first author and **Year** is the year (or anticipated year) of publication, e.g. **Smith2012**.	
4	**study_description**		A description (<700 words) of the study. If the study has been published, you may use the abstract.	
5	**study_alias**		A comma-delimited list of alternative names by which the study will be indexed, e.g. 1000 Genomes, Thousand Genomes.	
6	**study_type**		Choose a term from the list: Case Control = comparison between healthy and diseased samples, Case Set = disease-associated variation, Control Set = variation within normal samples, Curated = variation compiled from the literature or online resources, Tumour vs. Matched Normal = variation between tumour and healthy samples from the same individual.	
7	**taxonomy_id**		The NCBI taxonomy identifier(s), as a comma-delimited list, for the species studied. To find a taxonomy ID, go to http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy, click the link for the species and look for "Taxonomy ID=".	
8	**project_id**		The European Bioinformatics Institute project identifier.	
9	**pubmed_id**		** REQUIRED if the study has been published. The NCBI Pubmed ID(s), as a comma-delimited list.	
10	**study_url**		If the study has a website, provide the URL.	
11	**EGA_id**		If the study has been submitted to the European Genome-phenome Archive (EGA), provide the EGA study accession number.	
12	**hold_date**		If you wish to delay public release of the study data, enter a date for public release, in the format YYYY-MM-DD.	
13				
14				
15				
16				
17				

PLEASE READ FIRST | CONTACTS | **STUDY** | SAMPLESETS | Info-SAMPLES | SAMPLES | Info-EXPERIMENTS | EXPERIMENTS | Info-VARIANT CALLS | VARIANT_CALLS

Standardized Terminology

	A	B	C	D	E	F	G	H	I
1	**Info- Controlled Vocabulary**		Certain **fields** on some **WORKSHEETS** require the use of Controlled Vocabulary. This worksheet lists the allowable values for these fields.						
2	**CONTACTS**	**STUDY**	**SAMPLESETS**	**SAMPLESETS**	**SAMPLES**	**SAMPLES**	**EXPERIMENTS**	**EXPERIMENTS**	**EXPERIMENTS**
3	**contact_role**	**study_type**	**sampleset_sex**	**sampleset_type**	**subject_collection**	**subject_sex**	**experiment_type**	**method_type**	**analysis_type**
4	Submitter	Case Control	Female	Case	1000 Genomes	Female	Discovery	BAC aCGH	Curated
5	First author	Case-Set	Male	Control	AUTISM	Male	Validation	Curated	*De novo* and Local Sequence Assembly
6	Senior author	Control Set	Unknown		CEPH	Unknown		Digital array	*De novo* Sequence Assembly
7	Curator	Curated			HapMap			FISH	Local Sequence Assembly
8	Other	Tumor vs. Matched-Normal			HGDP			Gene expression array	Manual observation
9					NINDS			Karyotyping	Merging
10					PDR			MAPH	Multiple Complete Digestion
11								MassSpec	One End Anchored Assembly
12								MCD Analysis	Optical Mapping
13								Merging	Other
14								MLPA	Paired-end Mapping
15								Oligo aCGH	Probe Signal Intensity
16								Optical mapping	Read Depth
17								PCR	Sequence Alignment
18								qPCR	SNP Genotyping Analysis
19								ROMA	Split Read Mapping
20								RT-PCR	
21								Sequencing	
22								SNP Array	
23									

DGV Data Model



Data Access; download and archive

BETA **Database of Genomic Variants** BETA

A curated catalogue of human genomic structural variation

About the Project | Downloads | Links | Statistics | FAQ
Genome Browser | Query Tool | Submissions | Contact Us | Training Resources

Download Data

The contents of the Database of Genomic Variants can be downloaded as tab delimited text files. The first line of each file is the column description. Each variant as reported in the original study is given a VariationID and this represents a non-stable identifier which will change each time the database is updated. Each variant is also provided with an accession which is a stable identifier and will remain constant. Studies where CNV coordinates are based on the NCBI 36 (hg18) assembly are not mapped back to older builds. The Build 35 download file will therefore contain less data than the Build 36 download file.

Release Versions

Release Date	Build 35 (hg17)	Build 36 (hg18)	GRCh 37 (hg19)
2012-03-29	link	link	link

Filtered Variants

Release Date	GFF3 File
2012-03-29	link

Studies included in DGV2 and links to the original data

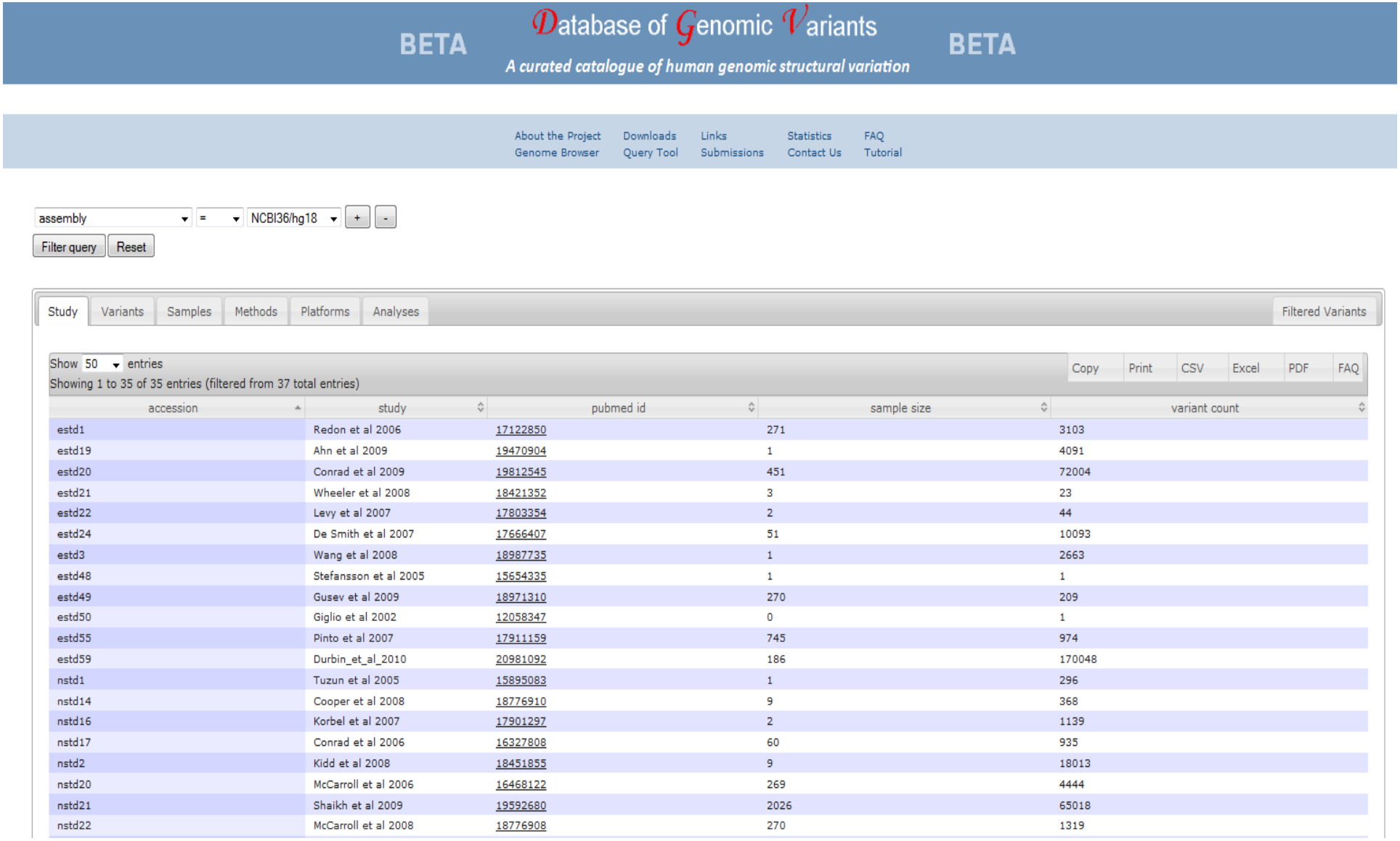
The table below outlines the studies that have been used as sources for the data and the number of regions included from each study. Here you can also download the original dataset for studies with more than ten CNV entries.

dgvbeta.tcag.ca/dgv/docs/NCBI36_hg18_2012-03-29.txt

Most Visited | Human (Homo sapien... | DGV * testrziman * | Database of Genomic ... | Statistics for dgvprod.r... | Statistics for projects.t... | DGV Test | DGV HG18 | Usage Statistics for wa... | NCBI JIRA | programs | ResearchGate | Bookmarks

variantid		chr	start	end	varianttype	variantsubtype	reference	pubmedid	method	platform	mergeid	frequency	samplesize	cohortdescription	genes
4322	20	206901	209000	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several populations that	
4323	20	60781151	60784750	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4324	20	1500351	1505850	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II	6764			6	Samples from several populations that	
4325	20	54607001	54608950	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4326	20	51764151	51767850	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4327	20	48979501	48982600	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4328	20	56643601	56645700	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4329	20	5736451	5740500	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II	6659			6	Samples from several populations that	
4330	20	15088051	15091650	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4331	20	25988951	25993150	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4332	20	54288401	54291450	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4333	20	47617276	47619400	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4334	20	40562751	40566750	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4335	20	26205451	26219350	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II			2699971	6	Samples from several	
4336	20	22370651	22373550	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4337	20	62194501	62198300	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4338	20	53239451	53242950	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4339	20	11422701	11424750	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	
4340	20	637701	640400	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several populations that	
4341	20	59846651	59849650	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II			2700003	6	Samples from several	
4342	20	25737951	25774950	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II			6615	6	Samples from several	
4343	20	26012151	26020750	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II			2699966	6	Samples from several	
4344	20	1693801	1696000	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several populations that	
4345	20	19024251	19026550	CNV	Loss	Durbin_et_al_2010	20981092	Sequencing	Illumina Genome Analyzer II				6	Samples from several	

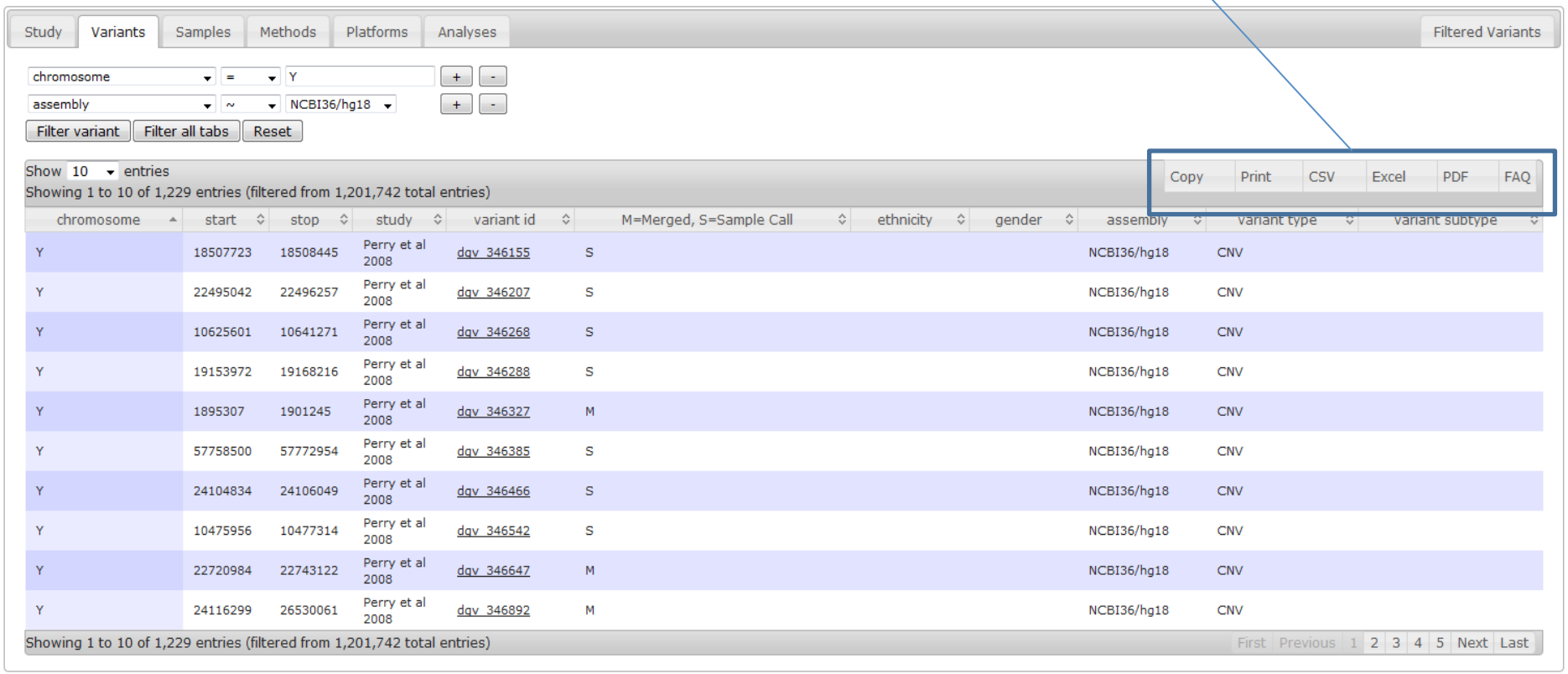
Interactive Access to Data: Query Tool



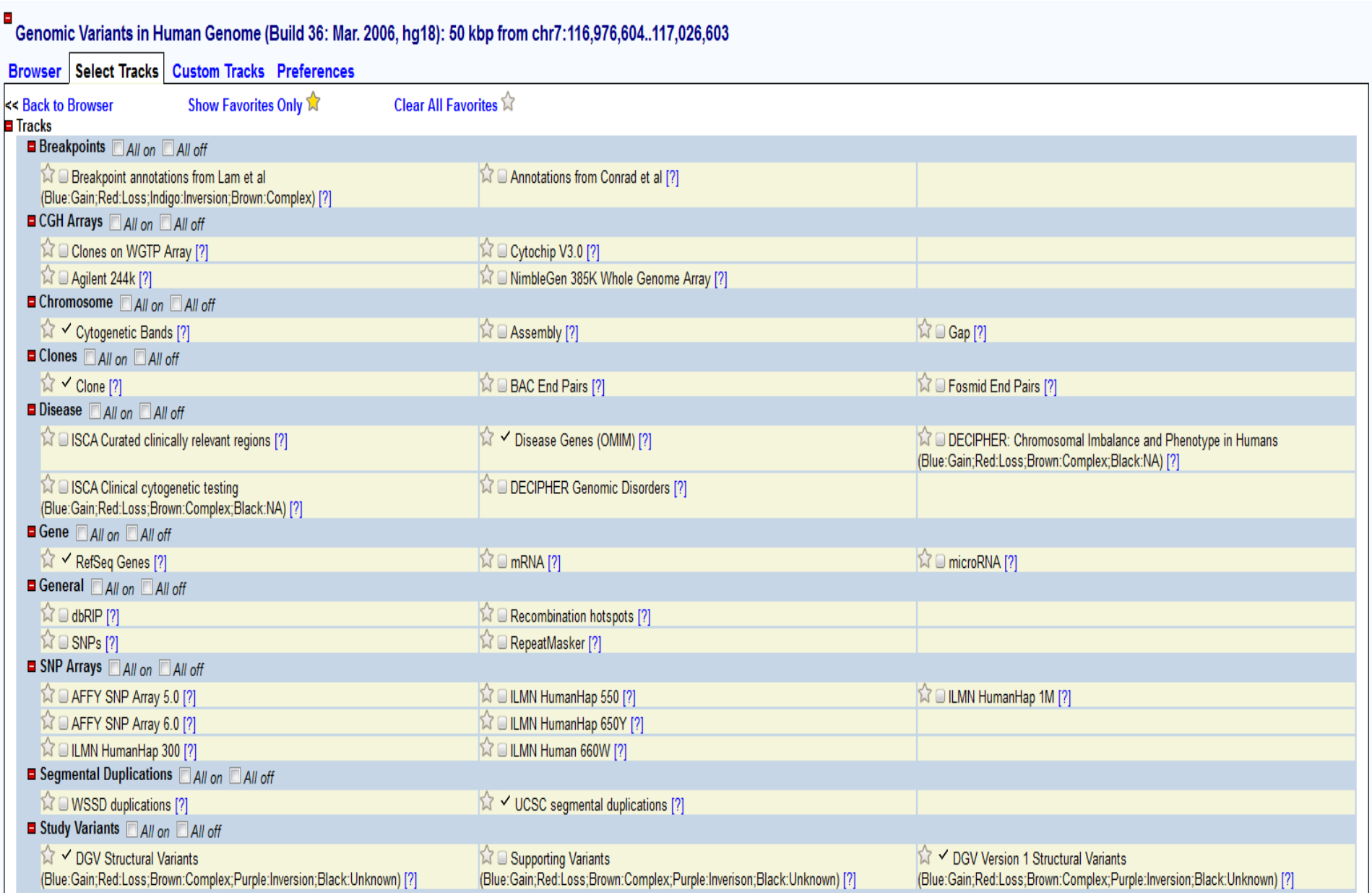
accession	study	pubmed id	sample size	variant count
estd1	Redon et al 2006	17122850	271	3103
estd19	Ahn et al 2009	19470904	1	4091
estd20	Conrad et al 2009	19812545	451	72004
estd21	Wheeler et al 2008	18421352	3	23
estd22	Levy et al 2007	17803354	2	44
estd24	De Smith et al 2007	17666407	51	10093
estd3	Wang et al 2008	18987735	1	2663
estd48	Stefansson et al 2005	15654335	1	1
estd49	Gusev et al 2009	18971310	270	209
estd50	Giglio et al 2002	12058347	0	1
estd55	Pinto et al 2007	17911159	745	974
estd59	Durbin_et_al_2010	20981092	186	170048
nstd1	Tuzun et al 2005	15895083	1	296
nstd14	Cooper et al 2008	18776910	9	368
nstd16	Korbel et al 2007	17901297	2	1139
nstd17	Conrad et al 2006	16327808	60	935
nstd2	Kidd et al 2008	18451855	9	18013
nstd20	McCarroll et al 2006	16468122	269	4444
nstd21	Shaikh et al 2009	19592680	2026	65018
nstd22	McCarroll et al 2008	18776908	270	1319

Data Extract: Query Tool

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



Genome Browser Track Options



Future Directions and Challenges

-How to improve submission process to increase number of entries
    -work/interact with journals to push for mandatory submission

-Should we expand and include cases/patient data in DGV
    -privacy/confidentiality issues

-How to handle/display a large number of future personal genomes data.
    -what data do we want to include, which are better suited for dbSNP(?)
    -need to optimize data acquisition and processing pipelines

-Generating a DGV curated Gold Standard Set
    -based on feedback from community, each individual may have different requirements; is there a one size fits all solution?