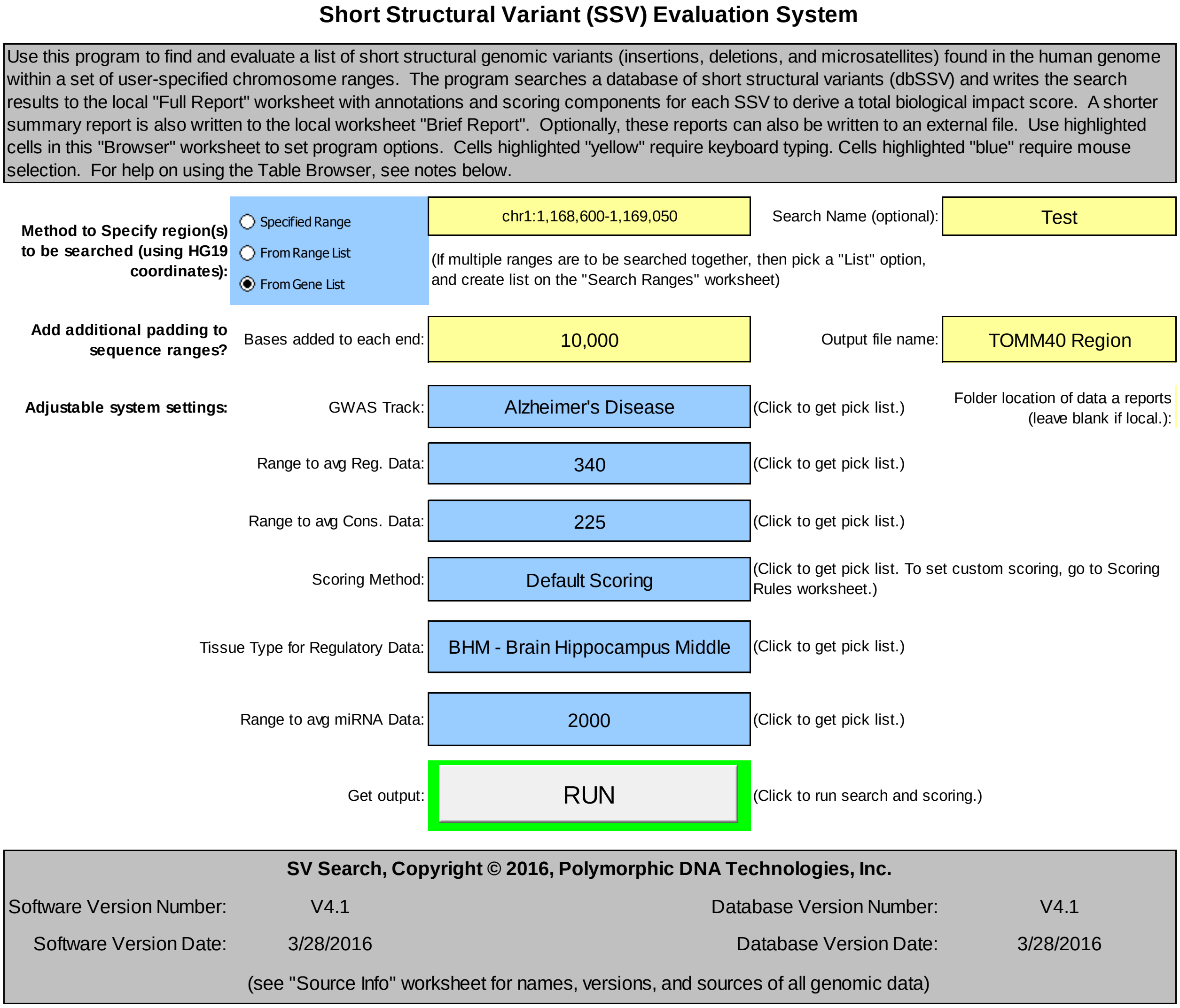


Short Structural Variant (SSV) Evaluation System

Use this program to find and evaluate a list of short structural genomic variants (insertions, deletions, and microsatellites) found in the human genome within a set of user-specified chromosome ranges. The program searches a database of short structural variants (dbSSV) and writes the search results to the local "Full Report" worksheet with annotations and scoring components for each SSV to derive a total biological impact score. A shorter summary report is also written to the local worksheet "Brief Report". Optionally, these reports can also be written to an external file. Use highlighted cells in this "Browser" worksheet to set program options. Cells highlighted "yellow" require keyboard typing. Cells highlighted "blue" require mouse selection. For help on using the Table Browser, see notes below.

Method to Specify region(s) to be searched (using HG19 coordinates):
○ Specified Range
○ From Range List
● From Gene List

chr1:1,168,600-1,169,050

Search Name (optional): Test

(If multiple ranges are to be searched together, then pick a "List" option, and create list on the "Search Ranges" worksheet)

Add additional padding to sequence ranges? Bases added to each end: 10,000

Output file name: TOMM40 Region

Adjustable system settings: GWAS Track: Alzheimer's Disease (Click to get pick list.)

Folder location of data a reports (leave blank if local.):

Range to avg Reg. Data: 340 (Click to get pick list.)

Range to avg Cons. Data: 225 (Click to get pick list.)

Scoring Method: Default Scoring (Click to get pick list. To set custom scoring, go to Scoring Rules worksheet.)

Tissue Type for Regulatory Data: BHM - Brain Hippocampus Middle (Click to get pick list.)

Range to avg miRNA Data: 2000 (Click to get pick list.)

Get output: RUN (Click to run search and scoring.)

SV Search, Copyright © 2016, Polymorphic DNA Technologies, Inc.

Software Version Number:	V4.1	Database Version Number:	V4.1
Software Version Date:	3/28/2016	Database Version Date:	3/28/2016

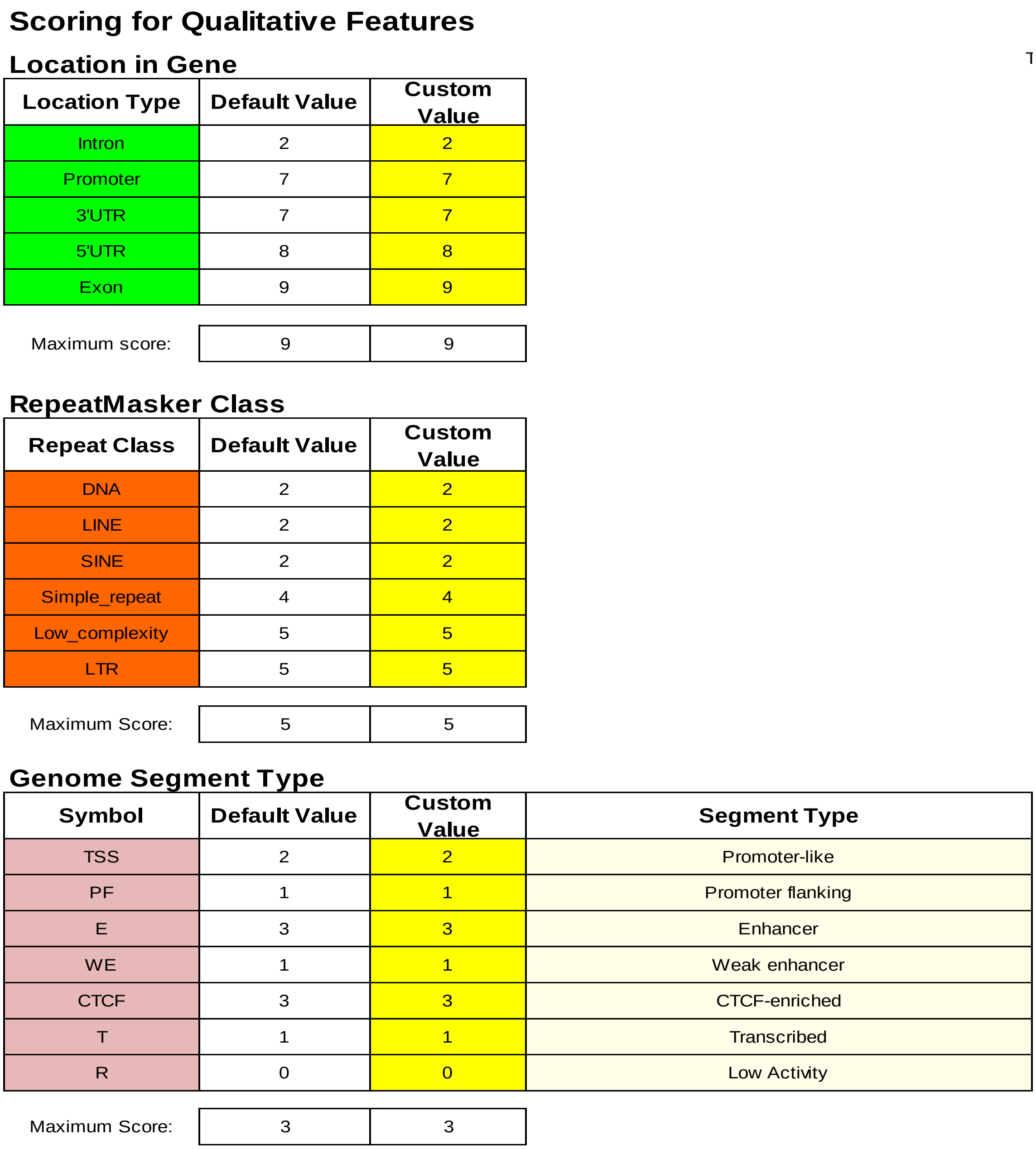
(see "Source Info" worksheet for names, versions, and sources of all genomic data)

Supp. Figure S1. Image of the main "Browser" page for the SSV Search program. Several "option button" controls are used to choose whether to specify the search range(s) directly with a chromosome range, or by using a separate list of either ranges or gene names. The inclusion of outside padding of each range can be achieved by specifying the number of base-pairs to add. There are also five "list box" controls that are used to specify key parameters and selections. For example, there is a "GWAS" list box, which permits selection of a specific genome-wide track of GWAS signals, and a tissue type list box, which is used to specify the particular tissue-type source for regulatory signals. A "RUN" button on the Browser page runs the macro to perform the search.

Supp. Figure S2A

Scoring for Numerical Properties						Default Settings		Custom Settings	
Field Name	Property Measured	Source Track	Derivation	Average Value	Maximum value	Weight Factor	Maximum Score	Weight Factor	Maximum Score
No. of Variants	Variability	dbSSV	Obtained by scanning dbSNP	1.1	9	0.6	5	0.6	5
Size Range of Variation	Variability	dbSSV	Obtained by scanning dbSNP	2.7	48	0.35	10	0.35	10
SSR Slippage	Variability	Local worksheet "Slippage Index"	Propritary table derived from PCR slippage experiments on 200 SSRs	1.9	40	0.15	5	0.15	5
Cluster Index	Synergy of consecutive variants	Calculated during reporting	Program assigns a value of "1" to the a neighboring variant if within 4 nucleotides. Uses value of "3" if an SSR.	0.3	6	1.00	6	1.00	6
GWAS Log(1-Log(p))	Association with a trait	Custom dbGWAS	GWAS data sets were converted to continuous tracks of signal, S = Log(1-Log(p)).	0.2	2.4	15.0	25	15.0	25
H3K4me1 Value	Nearby regulatory element	Tissue-specific dbREG	Downloaded from NIH Epigenomcs Roadmap, for the assay H3K4me1, for various tissues and donors. Local program smoothens signal over user-defined window.	2.6	48.0	0.16	5	0.16	5
H3K4me3 Value	Nearby regulatory element	Tissue-specific dbREG	Downloaded from NIH Epigenomcs Roadmap, for the assay H3K4me3, for various tissues and donors. Local program smoothens signal over user-defined window.	2.8	100.0	0.16	5	0.16	5
H3K9ac Value	Nearby regulatory element	Tissue-specific dbREG	Downloaded from NIH Epigenomcs Roadmap, for the assay H3K9ac, for various tissues and donors. Local program smoothens signal over user-defined window.	2.7	85.0	0.16	5	0.16	5
H3K27ac Value	Nearby regulatory element	Tissue-specific dbREG	Downloaded from NIH Epigenomcs Roadmap, for the assay H3K27ac, for various tissues and donors. Local program smoothens signal over user-defined window.	1	52	0.16	5	0.16	5
Dnase Value	Nearby regulatory element	Tissue-specific dbREG	Downloaded from NIH Epigenomcs Roadmap, for the assay DNase, for various tissues and donors. Local program smoothens signal over user-defined window.	2.00	800	0.02	5	0.02	5
RRBS Value	Nearby regulatory element	Tissue-specific dbREG	Downloaded from NIH Epigenomcs Roadmap, for the assay RRBS, for various tissues and donors. Local program smoothens signal over user-defined window.	0.20	10	2.00	5	2.00	5
TS miRNA Value	Nearby regulatory element	dbOREG	Downloaded from TargetScan's track for micro RNA. Peak signals were extended and attenuated over ranges of either 60 bp, 400 bp, or 2,000 bp, as chosen by user.	80	160	0.04	5	0.04	5
TFBS ChIP Value	Signal for TF Binding Sites	dbTFBS	Downloaded from wgEncodeRegTfbsClusteredV3. Multiple signals combined by summing squares and taking square root.	700	4000	0.0030	5	0.0030	5
TFBS in silico Value	Signal for TF Binding Sites	dbTFBS	Downloaded from tfbsConsSites. Signals are extended over 200 bp window. Multiple signals combined by summing squares and taking square root.	700	4000	0.0030	5	0.0030	5
Mammal Cons. Value	Region conserved among mammals	Custom dbCONS	Track "phastCons46wayPlacental" , smoothened over 25 bp (S = short)	0.06	1.00	0.0	0	0.0	0
Mammal Cons. Value, large window	Region conserved among mammals	Custom dbCONS	Mammal Cons. Track, further smoothened over specified range (L = long)	0.06	1.00	0.0	0	0.0	0
Mammal Cons. Value Difference	Local drop in conservation value	calculated	Difference of "L - S" for Mammals	0.04	1.00	25.0	10	25.0	10
Primate Cons. Value	Region conserved among Primates	Custom dbCONS	Track "phastCons46wayPrimates" , smoothened over 25 bp (S = short)	0.06	1.00	0.0	0	0.0	0
Primate Cons. Value, large window	Region conserved among Primates	Custom dbCONS	Primate Cons. Track, further smoothened over specified range (L = long)	0.06	1.00	0.0	0	0.0	0
Primate Cons. Value Difference	Local drop in conservation value	calculated	Difference of "L - S" for Primates	0.04	1.00	0.0	0	0.0	0
Intron Size	Intron Size	Custom dbGEN	Feature size from RefSeq track	2000	over 100,000	5.0	3	5.0	3

Supp. Figure S2B



Scoring for Qualitative Features

Location in Gene

Location Type	Default Value	Custom Value
Intron	2	2
Promoter	7	7
3'UTR	7	7
5'UTR	8	8
Exon	9	9
Maximum score:	9	9

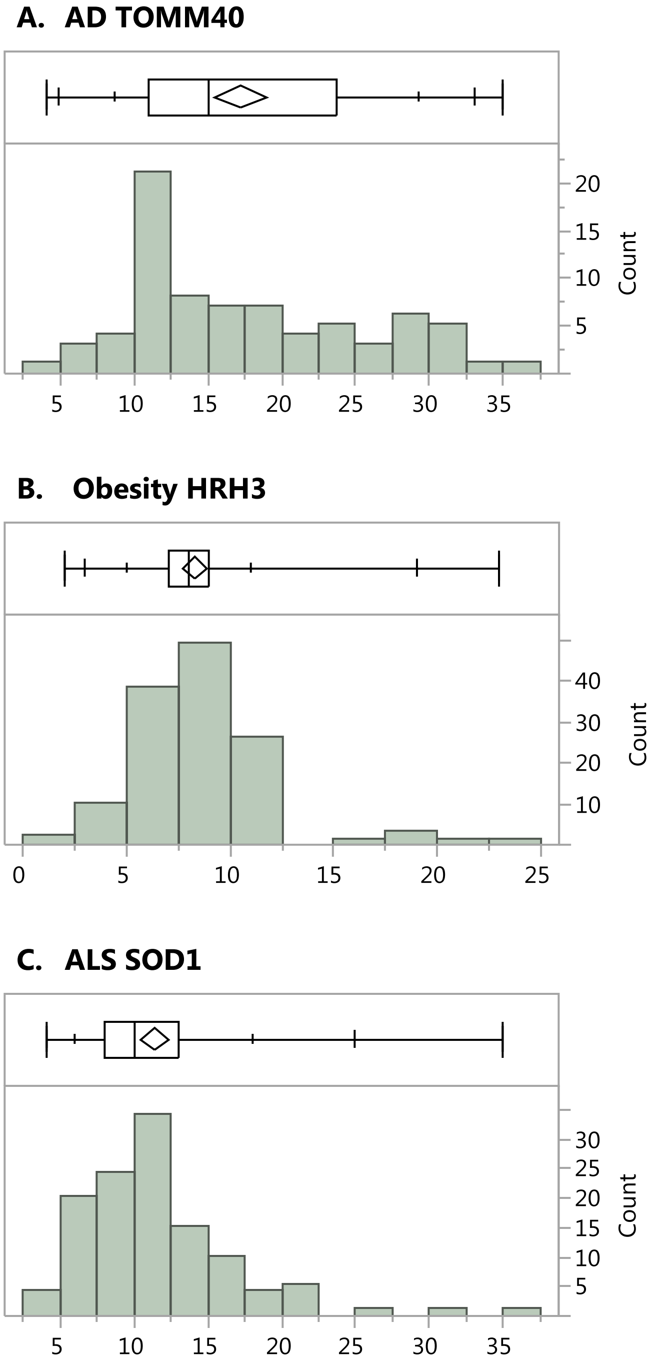
RepeatMasker Class

Repeat Class	Default Value	Custom Value
DNA	2	2
LINE	2	2
SINE	2	2
Simple_repeat	4	4
Low_complexity	5	5
LTR	5	5
Maximum Score:	5	5

Genome Segment Type

Symbol	Default Value	Custom Value	Segment Type
TSS	2	2	Promoter-like
PF	1	1	Promoter flanking
E	3	3	Enhancer
WE	1	1	Weak enhancer
CTCF	3	3	CTCF-enriched
T	1	1	Transcribed
R	0	0	Low Activity
Maximum Score:	3	3	

Supp. Figure S2. The ‘Scoring Rules’ worksheet in the SV Search program file. This worksheet contains our current preferred parameters, which are listed as default settings. There are also cells available to the user for custom scoring. **(A)** Shows the part of this worksheet used for scoring numerical parameters. For these, each partial score is calculated by multiplying the data value by the weighting factor. **(B)** Shows the part of the worksheet used for certain qualitative fields. For each text description, the appropriate look-up value is used for the partial score. The program combines all partial scores into category scores for use in the Brief Report and then combines all category scores to produce the final “Total Impact Score”.



Supp. Figure S3. The distribution of total potential impact scores for three examples of genetic association studies. Distributions are shown for **(A)** AD and the TOMM40 gene, **(B)** Obesity and the HRH3 gene and **(C)** Amyotrophic Lateral Sclerosis (ALS) ALS and the SOD1 gene. In this search the GWAS track was not used. Literature reports that support the association of these genes with the specific diseases are given in Roses et al. for AD and TOMM40, Yoshimoto et al. for obesity, and HRH3 and Tafuri et al. for ALS and SOD1.

The quantile box plot, located above the histogram distribution is a simple, graphical depiction of the quantiles of the distribution: the 25% and 75% quartiles are defined by the rectangle with the median as the line in the middle; the diamond shows the mean and 95% confidence interval for the mean; the short, horizontal lines on either side of the rectangle define quantiles: 0.5%, 2.5%, 10%, 90%, 97.5% and 99.5%.

Supp. References

Roses AD, Lutz MW, Amrine-Madsen H, Saunders AM, Crenshaw DG, Sundseth SS, et al. A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. Pharmacogenomics J. 2010;10(5):375-84. PMCID: PMC2946560.

Tafuri F, Ronchi D, Magri F, Comi GP, Corti S. SOD1 misplacing and mitochondrial dysfunction in amyotrophic lateral sclerosis pathogenesis. Front Cell Neurosci. 2015;9:336. PMCID: PMC4548205.

Yoshimoto R, Miyamoto Y, Shimamura K, Ishihara A, Takahashi K, Kotani H, et al. Therapeutic potential of histamine H3 receptor agonist for the treatment of obesity and diabetes mellitus. Proc Natl Acad Sci U S A. 2006;103(37):13866-71. PMCID: PMC1560086.

Supp. Table S1. Sources for Data Sets Used in dbSSV

Description	Track Name	Version Date	Source
Human Reference Sequence	GRCh37/hg19	Feb-09	hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/
Polymorphism data from dbSNP	SNP142	Nov-14	genome.ucsc.edu/cgi-bin/hgTables
Recombination Rates from HapMap	Recombination	Jun-08	hapmap.ncbi.nlm.nih.gov/downloads/recombination/latest/rates/
RefSeq Genes	refGene	Oct-11	hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refgene.txt
Mammalian Conservation	phastCons46wayPlacental	Nov-09	hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/placentalMammals/
Primate Conservation	phastCons46wayPrimates	Nov-09	hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/primates/
GWAS Data Sets	GWAS Catalog	Mar-15	genome.gov/gwastudies
Repeats, RepeatMasker	rmsk	Apr-09	hgdownload.cse.ucsc.edu/goldenPath/hg19/database/rmsk.txt
Regulation, H3K4me1, BHM	EA9BHMH3K4me112Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K4me1, BITL	EA9BITLH3K4me112Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K4me1, BMFL	EA9BMFLH3K4me112Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K4me1, BSN	EA9BSNH3K4me112Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K4me3, BHM	EA9BHMH3K4me312Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K4me3, BITL	EA9BITLH3K4me312Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K4me3, BMFL	EA9BMFLH3K4me312Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K4me3, BSN	EA9BSNH3K4me312Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K9ac, BHM	EA9BHMH3K9ac 12Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K9ac, BITL	EA9BITLH3K9ac 12Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K9ac, BMFL	EA9BMFLH3K9ac 12Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K9ac, BSN	EA9BSNH3K9ac 12Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K27ac, BHM	EA9BHMH3K27ac 1249Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K27ac, BITL	EA9BITLH3K27ac 1253Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K27ac, BMFL	EA9BMFLH3K27ac 1251Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, H3K27ac, BSN	EA9BSNH3K27ac 4947Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, Dnase I, FB	EA9FBDNase 1072Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, RRBS, BITL	EA9BITLRRBS4970Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, RRBS, BMFL	EA9BMFLRRBS4967Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, RRBS, BSN	EA9BSNRRBS1272Sig	May-15	abesapien.stamlab.org/cgi-bin/hgTables
Regulation, TS miRNA sites	targetScanS	Dec-10	genome.ucsc.edu/cgi-bin/hgTables
Regulation, Txn Factor ChiP	wgEncodeRegTfbsClusteredV3	Jul-13	genome.ucsc.edu/cgi-bin/hgTables
Regulation, TFBS Conserved	tfbsConsSites	Mar-11	genome.ucsc.edu/cgi-bin/hgTables
Genome Segmentations	hub_4607_GM12878_Combined_segmentation	Jan-11	genome.ucsc.edu/cgi-bin/hgTables

This table provides a list of the track names and websites used to download each type of data used in the dbSSV system**.**

Supp. Table S2. Derivation of dbSSV Fields

Name of Field	Property	Data Source	Derivation method
Column A. Chromosome Number	Integer value of chromosome number	GRCh37/hg19	Autosomal chromosomes numbered normally. ChrX = 23, ChrY = 24, ChrM not covered.
Column B. Chromosome Position	Integer value of chromosome number	GRCh37/hg19, SNP142	Position from dbSNP. For variants with a position range, the lowest integer is used.
Column C. Variant Type	Type of variant (insertion, deletions, etc.)	SNP142	Type defined by dbSNP, except for simple sequence repeats which are called SSRs
Column D. Symbol	Symbolic description of variation	SNP142	Usually a pair of alleles separated by a "/", with the reference allele list first. E.g., an "A" insertion is "-/A".
Column E. Span in Reference Sequence (bp)	Range size for reference allele	SNP142	For deletions, this is the size deleted. For insertions, value is zero.
Column F. Representative Variant Name	Reference SNP (refSNP) ID from dbSNP, if any	SNP142	When multiple IDs describe the same variant, the version with the lowest chromosome position is used.
Column G. No. of Known Variants in dbSNP	No. of named variants overlapping with an SSR span	SNP142	Obtained by scanning dbSNP
Column H. Nucleotide Size Range in dbSNP	Size difference between shortest and longest allele	SNP142	Obtained by scanning dbSNP
Column I. SSR Slippage Index	Estimated mutability of an SSR	Proprietary table	Table derived from PCR slippage experiments on 200 different SSRs
Column J. Repeat Name	RepeatMasker Repeat Name	Repeatmasker	Directly from RepeatMasker
Column K. Repeat Class	RepeatMasker Repeat Class	Repeatmasker	Directly from RepeatMasker
Column L. Repeat Family	RepeatMasker Repeat Family	Repeatmasker	Directly from RepeatMasker
Column M. Clustering Index	Synergy of consecutive variants	Calculated	Neighboring variants are given scores based on nearness and variant type.
Column N. Associated Gene	Gene name, if variant is within or close to that gene	refGene	Directly from refGene
Column O. RefSeq Accession Number	Name of gene expression track	refGene	Directly from refGene
Column P. Gene Feature	Coding Exon, Intron, 5'-UTR, 3'UTR, or Promoter	refGene	Directly from refGene
Column Q. Strand	Transcription dir. of gene relative to chrom. sequence	refGene	Directly from refGene
Column R. Feature Size	Size of gene featurein bp	refGene	From coordinates of feature in refGene
Column S. Position in Feature	Position of variant within feature	refGene	From coordinates of feature in refGene
Column T. Mammal, 25 bp window	Conservation of region in mammals	phastCons46wayPlacental	phastCons46wayPlacental values, averaged over 25 bp
Column U. Mammal, 75 bp window	Conservation of region in mammals	phastCons46wayPlacental	phastCons46wayPlacental values, averaged over 75 bp
Column V. Mammal, 125 bp window	Conservation of region in mammals	phastCons46wayPlacental	phastCons46wayPlacental values, averaged over 125 bp
Column W. Mammal, 225 bp window	Conservation of region in mammals	phastCons46wayPlacental	phastCons46wayPlacental values, averaged over 225 bp
Column X. Primate, 25 bp window	Conservation of region in primates	phastCons46wayPrimates	phastCons46wayPrimates values, averaged over 25 bp
Column Y. Primate, 75 bp window	Conservation of region in primates	phastCons46wayPrimates	phastCons46wayPrimates values, averaged over 75 bp
Column Z. Primate, 125 bp window	Conservation of region in primates	phastCons46wayPrimates	phastCons46wayPrimates values, averaged over 125 bp
Column AA. Primate, 225 bp window	Conservation of region in primates	phastCons46wayPrimates	phastCons46wayPrimates values, averaged over 225 bp
Columns AB through AT. GWAS_001 through GWAS_019	Signal derived from Genome Wide Association Studies for each of 19 different phenotypes	GWAS Catalog	P-values for associated SNPs from the GWAS catalog are converted to a signal defined as Log(1-Log(p)). A continuous track of this signal is calculated by using the recombination-rate track to determine the reduction over distance of this value in both direactions.
Column BF LD Block Name	Name assigned consecutively to an extended chromosomal range with a low recombination rate.	Recombination Rates from HapMap	LD Block is defined here as a chromosomal region that has no sites with a recombination rate greater than 3.0 units.
Column BG LD Block start position	Chromosome position of beginning of LD block	Recombination Rates from HapMap	LD Block is defined here as a chromosomal region that has no sites with a recombination rate greater than 3.0 units.
Column BH LD Block end position	Chromosome position of end of LD block	Recombination Rates from HapMap	LD Block is defined here as a chromosomal region that has no sites with a recombination rate greater than 3.0 units.
Columns BI through GD. Regulatory data for various tissues from NIH Roadmap project.	Regulatory signals for H3K4me1, H3K4me3, H3K9ac, H3K27ac, Dnase, and RRBS, for various tisue tytpes	NIH Roadmap Epigenetics Project	Various regulatory tracks are averaged over windows of 100 bp, 220 bp, and 340 bp.
Column GE. miRNA, 60 bp window	Signal for miRNA sites from Target Scan miRNA	targetScanS	Peak signal value extended over a 60 bp window.
Column GF. miRNA, 400 bp window	Signal for miRNA sites from Target Scan miRNA	targetScanS	Peak signal value, linearly attenuated over +/- 200 bp flanking.
Column GF. miRNA, 2000 bp window	Signal for miRNA sites from Target Scan miRNA	targetScanS	Peak signal value, linearly attenuated over +/- 1000 bp flanking.
Column GK, Txn Factor binding site name	Name of Transcription binding site, from ChiP	wgEncodeRegTfbsClusteredV3	Directly from wgEncodeRegTfbsClusteredV3. Multiple names are concatenated.
Column GL, Txn Factor binding site signal	Signal from Transcription binding site, from ChiP	wgEncodeRegTfbsClusteredV3	Directly from wgEncodeRegTfbsClusteredV3. Multiple signals combined by summing squares and taking square root
Column GM, Txn Factor binding site name	Name of Transcription binding site, in silico derived	tfbsConsSites	From tfbsConsSites adding +/-100 bp of flanking. Multiple names are concatenated.
Column GN, Txn Factor binding site signal	Signal from Transcription binding site, in silico derived	tfbsConsSites	From tfbsConsSites, extended over 200 bp window. Multiple signals combined by summing squares and taking square root.
Column GQ, Genome Segmentation	Predicted regulatory status of regions	hub_4607_GM12878_Combined_segmentation	TSS = "Promoter-like", PF = Promoter flanking, E = Enhancer, WE = Weak enhancer, CTCF = CTCF-enriched, T = Transcribed, R = Low Activity

This table describes the property and derivation of each field found in dbSSV.