Genome variation - part 1

Dr Jason Wong

UNSW AUSTRALIA

Never Stand Still | Medicine | Prince of Wales Clinical School

Introductory bioinformatics for human genomics workshop, UNSW

Day 2 – Friday 21th January 2016

UNSW AUSTRALIA

ADULT CANCER PROGRAM

LOWY CANCER RESEARCH CENTRE

tcrn bridging the gap translational cancer research network

NSW GOVERNMENT

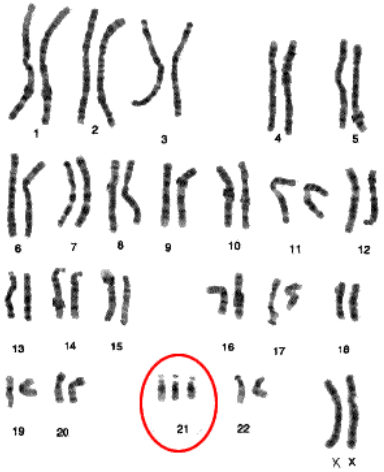
cancer institute NSW

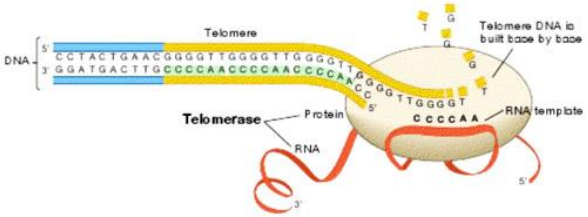
Aims of the session

- Introduce major human genome variation databases.
  - dbSNP
  - 1000genomes
- Basic variant annotation using UCSC.
- We will look at ExAC/gnomAD in the next session.

Types of variation

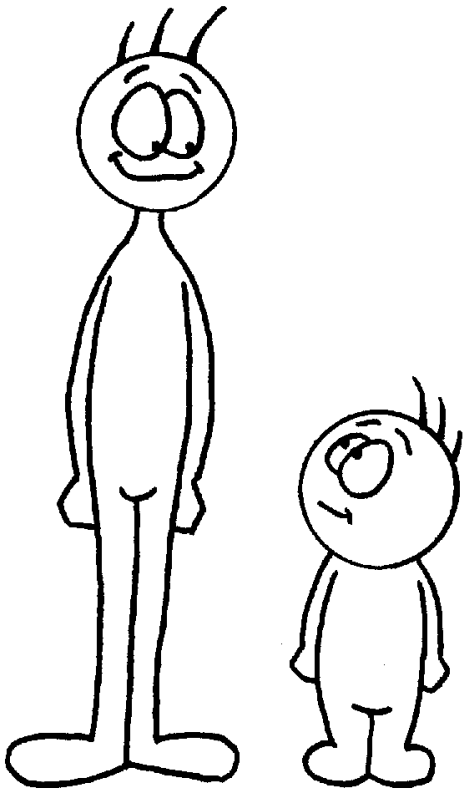
- **Cytological level:**
  - Chromosome numbers
  - Segmental duplications, rearrangements, and deletions
- **Sub-chromosomal level:**
  - Transposable Elements
  - Short Deletions/Insertions, Tandem Repeats
- **Sequence level:**
  - Single Nucleotide Polymorphisms (SNPs)
  - Small Nucleotide Insertions and Deletions (Indels)

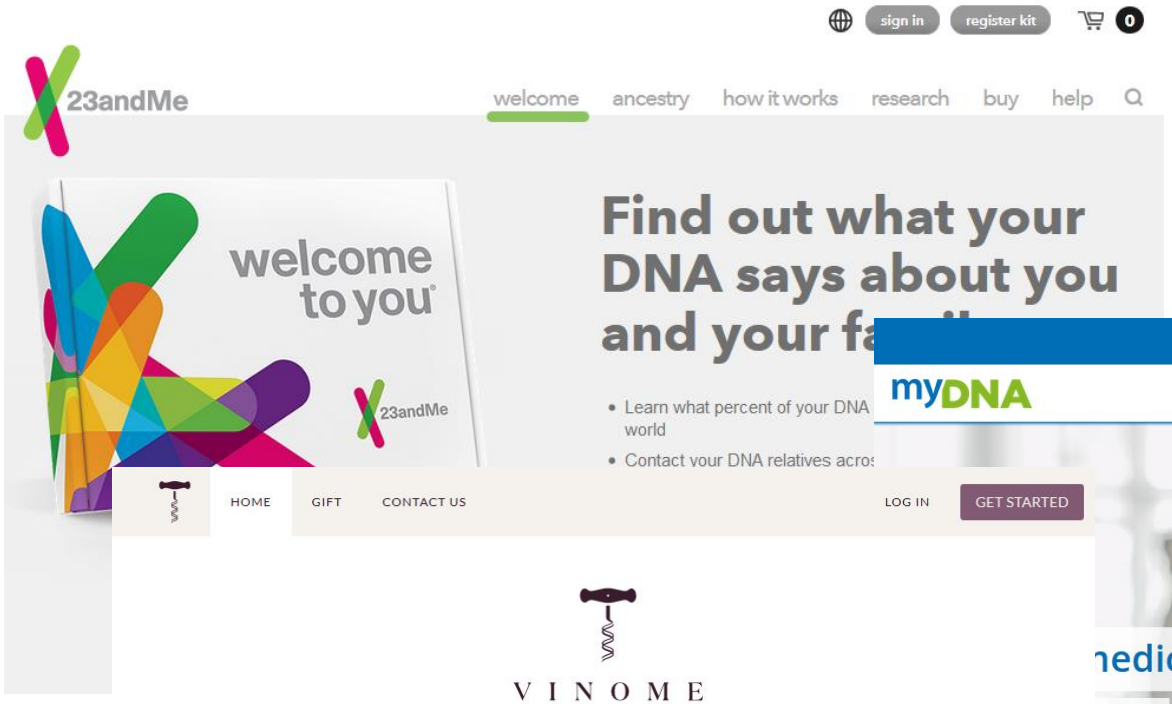


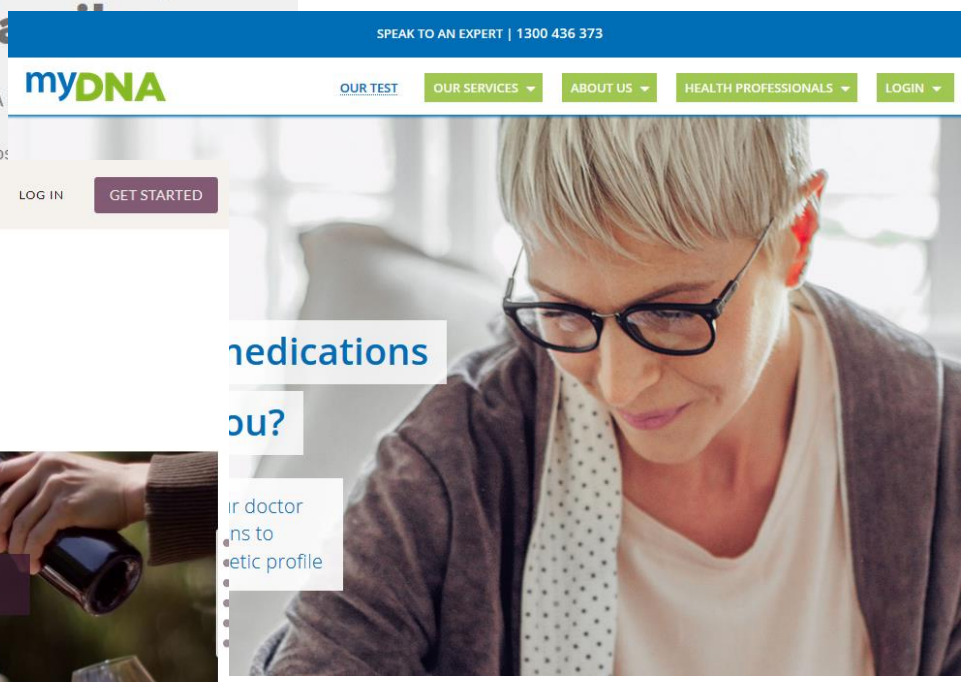


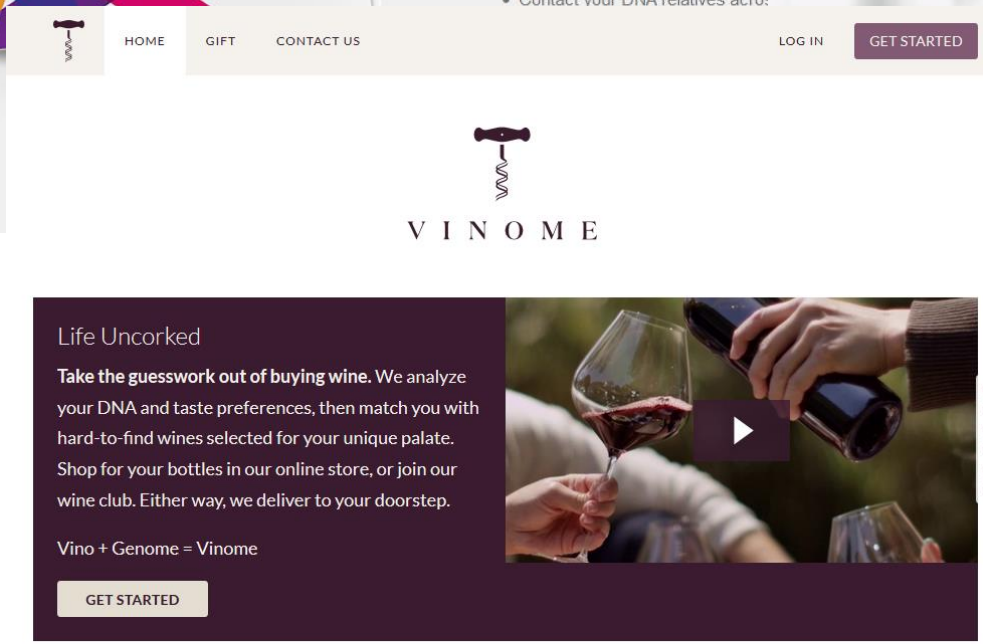
Why study sequence variation?

- Rare disease
- Determine disease risk
- Response to therapy
- Forensics
- Evolution









What if we told you there was a better way to experience wine?

Uncorking a better wine experience has never been simpler

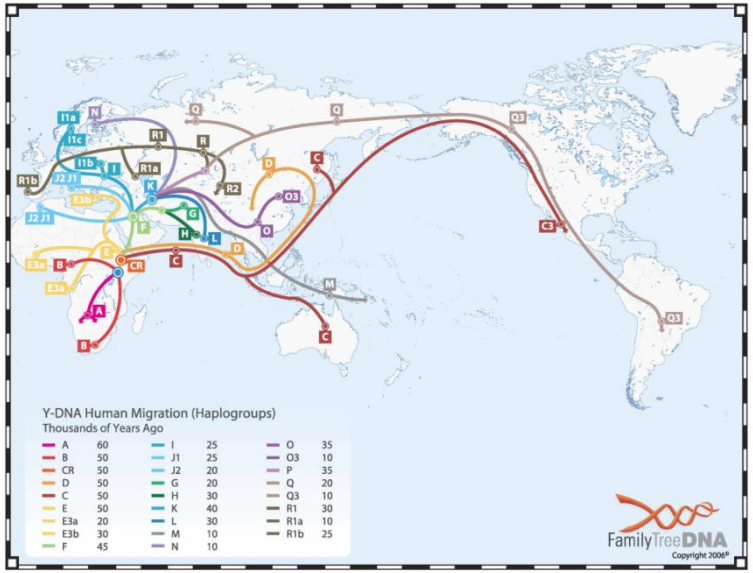
Single nucleotide polymorpisms (SNP)

- Typically refers to single bases substitution.

GAATTC
GAACTC

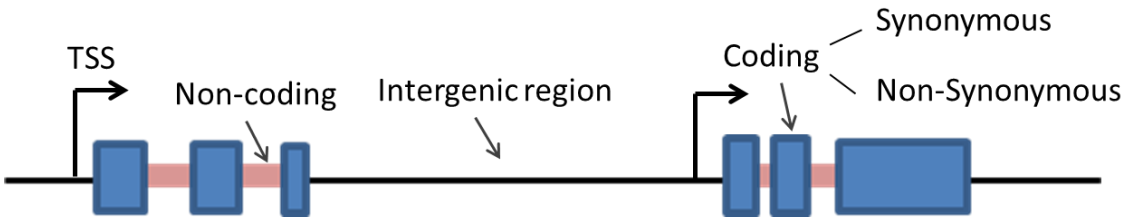
- There are ~40 M common SNPs in human population.

- A given individual would expect to differ from reference genome by 1% (i.e. 3 million SNPs)



Types of SNPs

- Genic, coding SNPs
  - Frameshift
  - Splice site
  - Non-synonymous (missense, nonsense)
  - Synonymous (splice enhancer/suppressor?)
- Genic, non-coding SNPs
  - Untranslated region
  - Regulatory SNPs
  - Intronic SNPs
- Intergenic



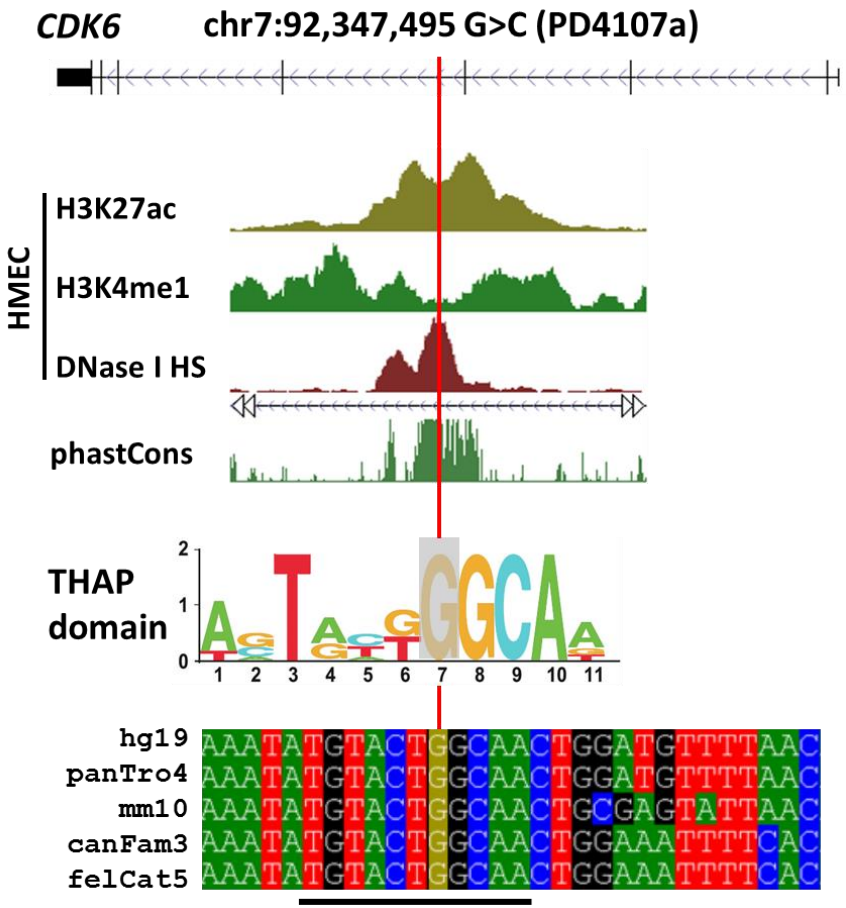
Predicting effect of coding SNPs

- Functional importance of SNPs usually based on:
  - Sequence conservation.
  - Frequency in population.
  - Alter protein 2D/3D structure.
  - Within protein motifs.

- Many tools are now available for coding SNP function prediction – BUT is still far from perfect.

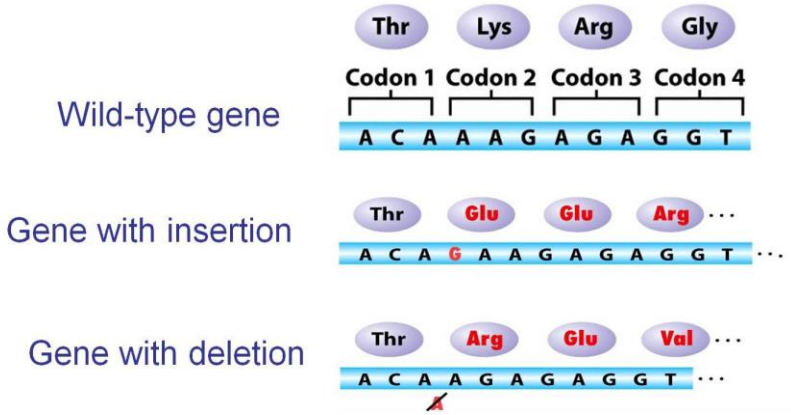
Non-coding SNPs

- Traditionally more difficult to annotate as >98% of the genome is non-coding.
- Want to find SNPs that is associated with gene expression (eQTLs).
- With the ENCODE/Epigenome project, it is easier (but still very difficult) to find potential functional non-coding SNPs.



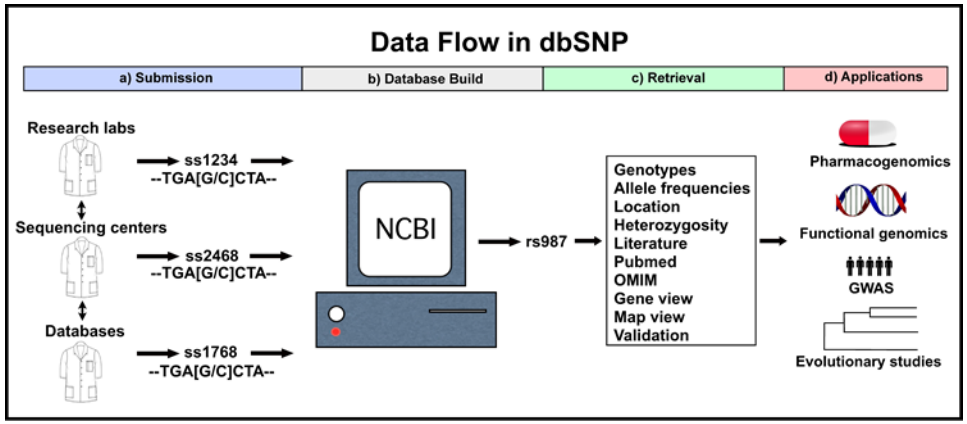
Insertion/**del**etion (Indels)

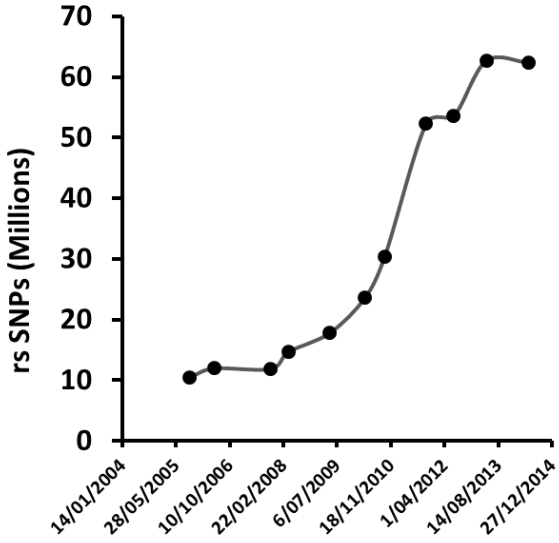
- Typically defined as gain or loss of 1-50 bps
- Less frequent than SNPs (~10% of all sequence variation).
- But if in coding sequence can additionally cause frameshift mutations.



dbSNP

- Online database from NCBI for cataloguing all SNPs submitted by the scientific community.





www.ncbi.nlm.nih.gov/SNP/

UNSW THE UNIVERSITY OF NEW SOUTH WALES

Reliability of dbSNP

- Problem with dbSNP is that anyone can upload variants and therefore it is claimed that there is a false positive rate of perhaps > 10%.

- Furthermore, some somatic mutations have also found their way into dbSNP.

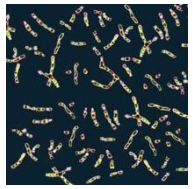
- Therefore, use dbSNP – BUT ideally only SNPs from 1000genomes project.

1000 Genomes project

- Goal of the project is to find virtually all genetic variants with frequency of at least 1% in the human population.

- Ultimate aims was to sequence ~2500 humans at 4x whole-genome coverage

www.1000genomes.org



1000 Genomes samples

Major populations	Total samples
East Asian (ASN)	523
South Asian (SAN)	494
African (AFR)	691
European (EUR)	514
Americas (AMR)	355
Total	2,577

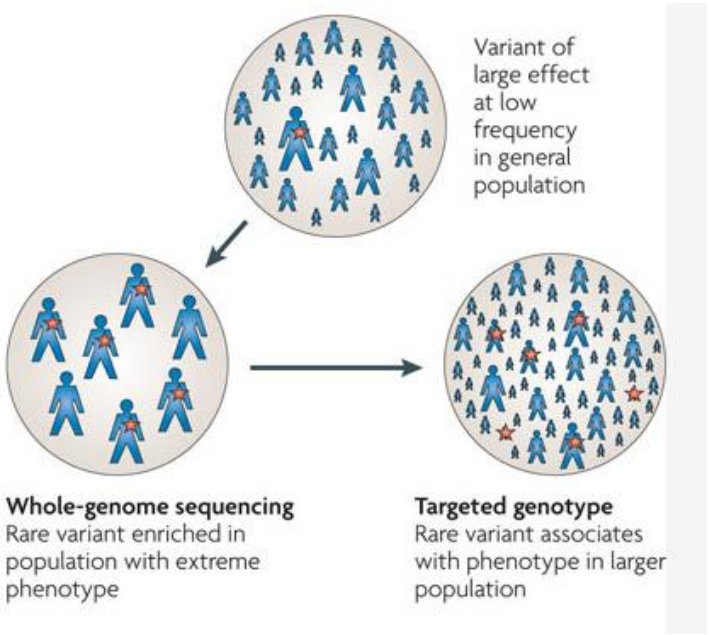
26 sub-populations

Population Code	Population Description	Super Population
CHB	Han Chinese in Bejing, China	ASN
JPT	Japanese in Tokyo, Japan	ASN
CHS	Southern Han Chinese	ASN
CDX	Chinese Dai in Xishuangbanna, China	ASN
KHV	Kinh in Ho Chi Minh City, Vietnam	ASN
CEU	Utah Residents (CEPH) with Northern and Western European ancestry	EUR
TSI	Toscani in Italia	EUR
FIN	Finnish in Finland	EUR
GBR	British in England and Scotland	EUR
IBS	Iberian population in Spain	EUR
YRI	Yoruba in Ibadan, Nigera	AFR
LWK	Luhya in Webuye, Kenya	AFR
GWD	Gambian in Western Divisons in The Gambia	AFR
MSL	Mende in Sierra Leone	AFR
ESN	Esan in Nigera	AFR
ASW	Americans of African Ancestry in SW USA	AFR
ACB	African Carribbeans in Barbados	AFR
MXL	Mexican Ancestry from Los Angeles USA	AMR
PUR	Puerto Ricans from Puerto Rico	AMR
CLM	Colombians from Medellin, Colombia	AMR
PEL	Peruvians from Lima, Peru	AMR
GIH	Gujarati Indian from Houston, Texas	SAN
PJL	Punjabi from Lahore, Pakistan	SAN
BEB	Bengali from Bangladesh	SAN
STU	Sri Lankan Tamil from the UK	SAN
ITU	Indian Telugu from the UK	SAN

Usage of variant data

- GWAS studies have typically only been able to find associations of variants of frequency of > 5%.

- 1000 Genomes project enables screening of variants discovered in sequencing of patients with genetic disorders and in cancer genomes.



Per individual variant load (*Nature* 491, 56-65)

Variant type	Number of derived variant sites per individual			Excess rare deleterious	Excess low-frequency deleterious
	Derived allele frequency across sample				
	<0.5%	0.5–5%	>5%		
All sites	30–150 K	120–680 K	3.6–3.9 M	ND	ND
Synonymous*	29–120	82–420	1.3–1.4 K	ND	ND
Non-synonymous*	130–400	240–910	2.3–2.7 K	76–190†	77-130†
Stop-gain*	3.9–10	5.3–19	24–28	3.4–7.5†	3.8–11†
Stop-loss	1.0–1.2	1.0–1.9	2.1–2.8	0.81–1.1†	0.80–1.0†
HGMD-DM*	2.5–5.1	4.8–17	11–18	1.6–4.7†	3.8–12†
COSMIC*	1.3–2.0	1.8–5.1	5.2–10	0.93–1.6†	1.3–2.0†
Indel frameshift	1.0–1.3	11–24	60–66	ND§	3.2–11†
Indel non-frameshift	2.1–2.3	9.5–24	67–71	ND§	0–0.73†
Splice site donor	1.7–3.6	2.4–7.2	2.6–5.2	1.6–3.3†	3.1–6.2†
Splice site acceptor	1.5–2.9	1.5–4.0	2.1–4.6	1.4–2.6†	1.2–3.3†
UTR*	120–430	300–1,400	3.5–4.0 K	0–350‡	0–1.2 K‡
Non-coding RNA*	3.9–17	14–70	180–200	0.62–2.6‡	3.4–13‡
Motif gain in TF peak*	4.7–14	23–59	170–180	0–2.6‡	3.8–15‡
Motif loss in TF peak*	18–69	71–300	580–650	7.7–22‡	37–110‡
Other conserved*	2.0–9.9 K	7.1–39 K	120–130 K	ND	ND
Total conserved	2.3–11 K	7.7–42 K	130–150 K	150–510	250–1.3 K

Only sites in which ancestral state can be assigned with high confidence are reported. The ranges reported are across populations. COSMIC, Catalogue of Somatic Mutations in Cancer; HGMD-DM, Human Gene Mutation Database (HGMD) disease-causing mutations; TF, transcription factor; ND, not determined.

* Sites with GERP >2

† Using synonymous sites as a baseline.

‡ Using 'other conserved' as a baseline.

§ Rare indels were filtered in phase I.

State of 1000 genomes project

- Phase 1
  - Completed in 2012
  - 1,092 humans.
  - 14 populations
  - 36.7 M SNPs
  - 1.38 M Indels
- Phase 3
  - Completed in 2014
  - 2,535 humans.
  - 26 populations.
  - 78.1 M SNPs
  - 3.1 M Indels

1.9 M SNPS are not shared

- Some samples not shared.
- Different sequencing platforms.
- Change in variant calling pipeline.

Accessing 1000 genomes data

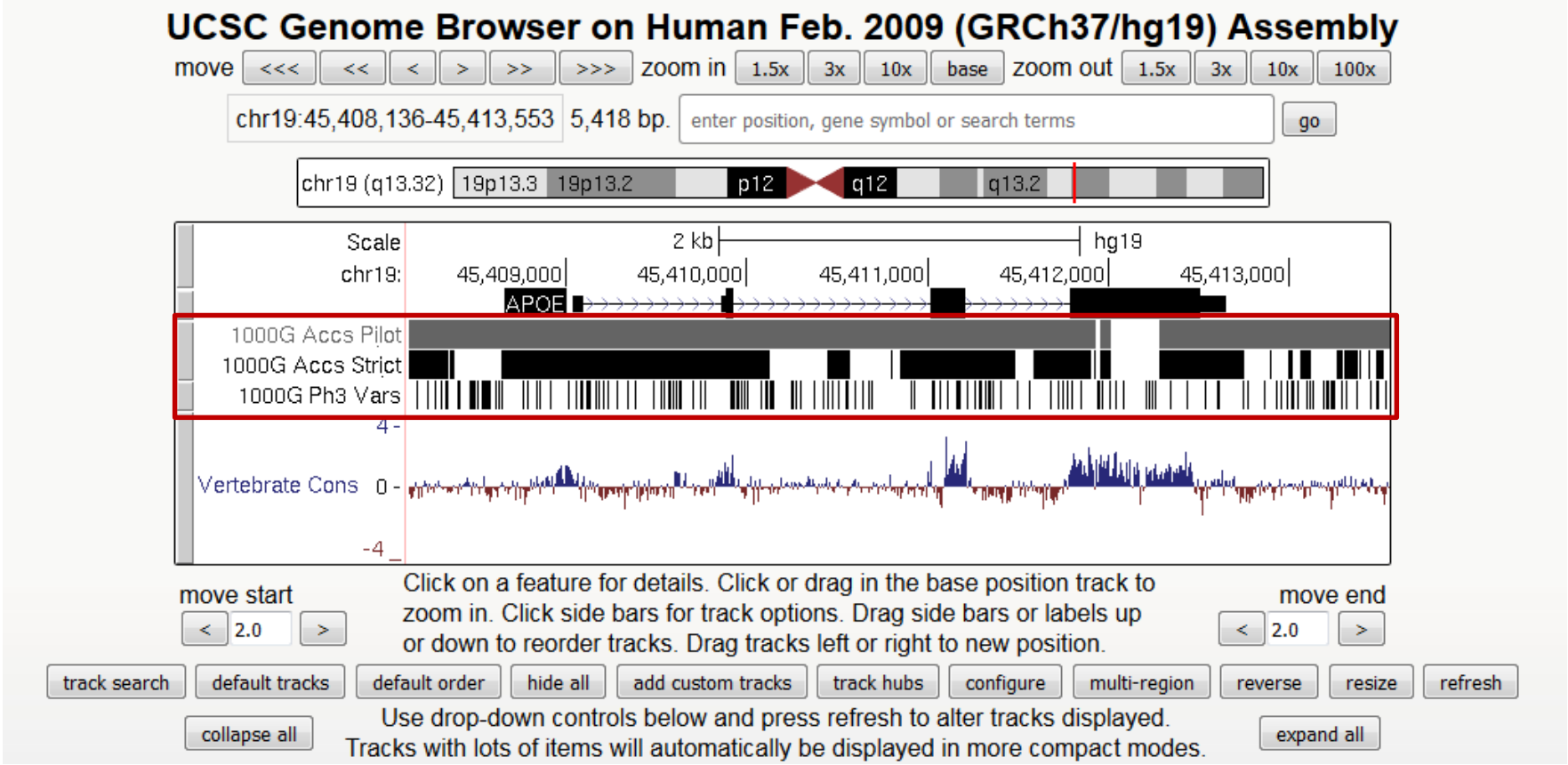
- www.1000genomes.org contains variant calls (VCF format), aligned BAM files and RAW files.

- Almost all (if not all) SNPs from 1000 genomes also catalogued in dbSNP.

- UCSC has a track containing all 1000 genome Phase 1 and 3 data.

Accessing 1000genomes from UCSC

Aim: Visualise all coding 1000 genome SNPs over the *APOE* gene and highlighting non-synonymous SNPs



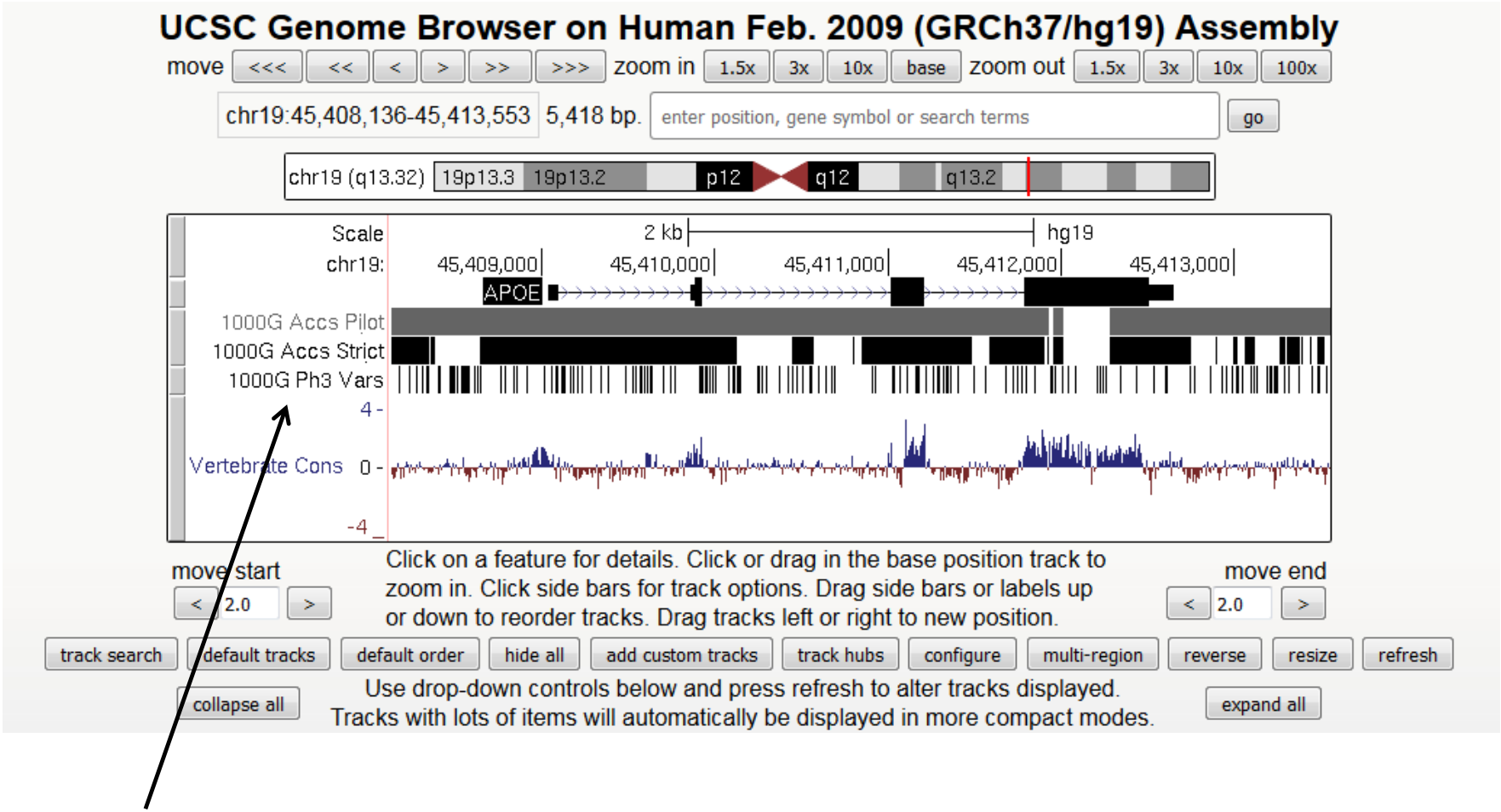
To load session, user: jasewong session name: bioinf_workshop_SNP_2016

UNSW THE UNIVERSITY OF NEW SOUTH WALES

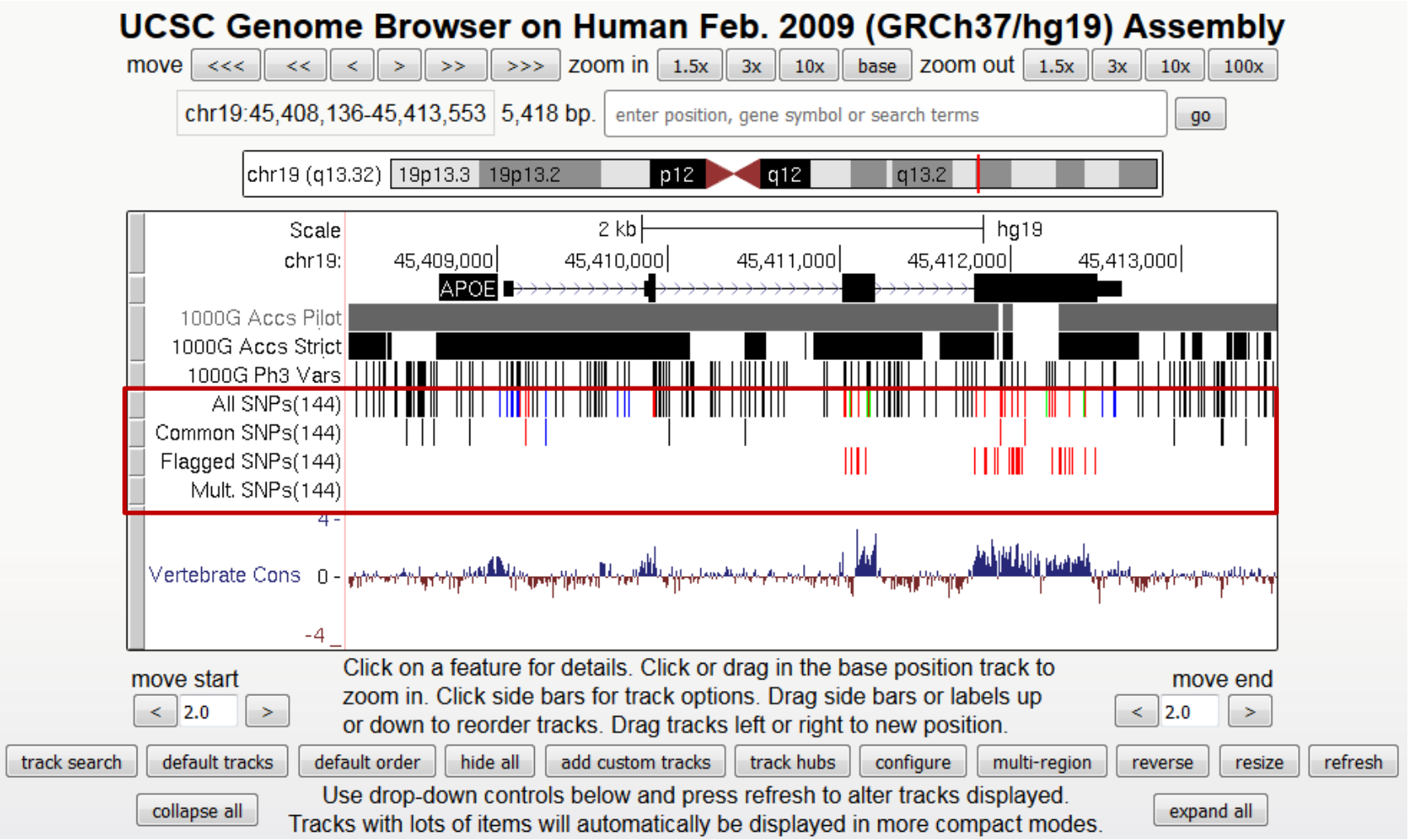
Accessibility tracks

- Shows regions of the genome where variant calls can be reliably made using whole genome NGS data.
- Note that some regions still have variant calls because 1000 genomes didn't just use whole gnome NGS.
- Useful for raising caution if your variants come from these regions.

https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=tgpPhase3Accessibility



Good for downloading SNPs, but not good for visualisation – for this bring up dbSNP tracks



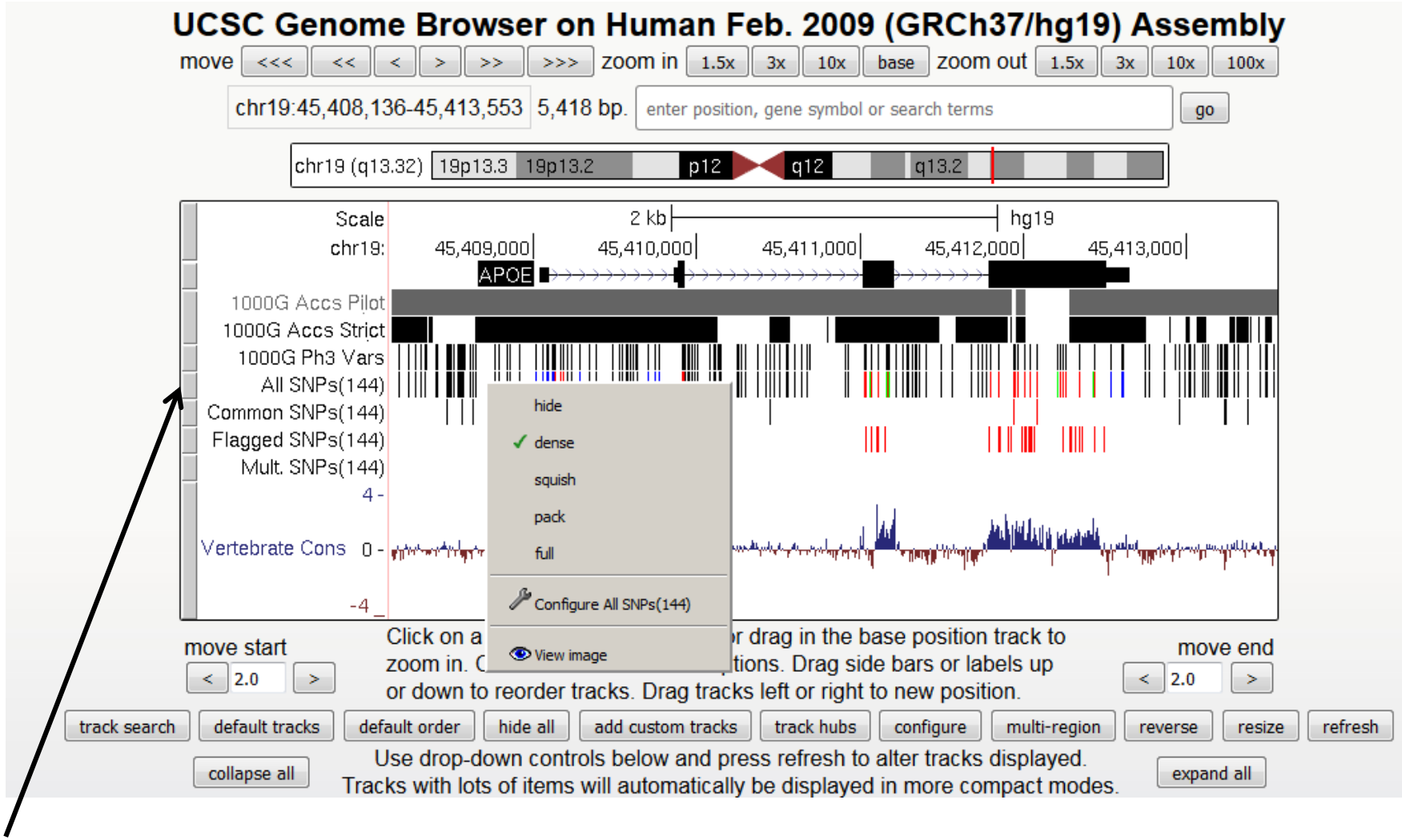
Bring up dbSNP 144 All SNPs, Common SNPs, Flagged SNPs and Multi. SNP from the “Variation” section of tracks. Select “dense” for each one.

SNP definitions

- Common SNPs - SNPs with >= 1% minor allele frequency (MAF), mapping only once to reference assembly.
- Flagged SNPs - SNPs < 1% minor allele frequency (MAF) (or unknown), mapping only once to reference assembly, flagged in dbSNP as "clinically associated" -- ***not necessarily a risk allele!*** (These are rare SNPs that with known clinical function).
- Mult. SNPs - SNPs mapping in more than one place on reference assembly.
- All SNPs - all SNPs from dbSNP mapping to reference assembly.

https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=snp144

Need to configure track to only show 1000 genome coding SNPs



Right-click anywhere on the “All SNPs” track and select “Configure All SNPs(144)

Display mode: dense Submit

Include Chimp state and observed human alleles in name:
(If enabled, chimp allele is displayed first, then '>', then human alleles).

Use Gene Tracks for Functional Annotation

Filtering Options

Minimum Average Heterozygosity: 0
Maximum Weight: 1 *Range: 1, 2 or 3; SNPs with higher weights are less reliable*
Minimum number of distinct Submitters: 0
Minor Allele Frequency range: 0 to 0.5 *Range: 0.0 - 0.5*
Minimum chromosome sample count (2N) for Allele Frequency data: 0

Filter by attribute:
Check the boxes below to include SNPs with those attributes. In order to be displayed, a SNP must pass the filter for each category. Some assemblies may not contain any SNPs that have some of the listed attributes.

Class: Set all Clear all
Unknown ☑ Single Nucleotide Polymorphism In/Del Heterozygous
Microsatellite Named Mnp Insertion
Deletion

Validation: Set all Clear all
Unknown By Cluster By Frequency By Submitter
By 2 Hit / 2 Allele By HapMap ☑ By 1000 Genomes Project

Function: Set all Clear all
Unknown ☑ synonymous_variant intron_variant ☑ downstream_gene_variant
☑ upstream_gene_variant ☑ nc_transcript_variant ☑ stop_gained ☑ missense_variant
☑ stop_lost ☑ frameshift_variant ☑ inframe_indel ☑ 3_prime_UTR_variant
☑ 5_prime_UTR_variant ☑ splice_acceptor_variant ☑ splice_donor_variant

Molecule Type: Set all Clear all
☑ Unknown ☑ Genomic ☑ cDNA

Unusual Conditions (UCSC): Set all Clear all
☑ None ☑ RefAlleleMismatch ☑ RefAlleleRevComp ☑ DuplicateObserved
☑ MixedObserved ☑ FlankMismatchGenomeLonger ☑ FlankMismatchGenomeEqual ☑ FlankMismatchGenomeShorter
☑ NamedDeletionZeroSpan ☑ NamedInsertionNonzeroSpan ☑ SingleClassLongerSpan ☑ SingleClassZeroSpan
☑ SingleClassTriAllelic ☑ SingleClassQuadAllelic ☑ ObservedWrongFormat ☑ ObservedTooLong
☑ ObservedContainslupac ☑ ObservedMismatch ☑ MultipleAlignments ☑ NonIntegerChromCount
☑ AlleleFreqSumNot1 ☑ SingleAlleleFreq ☑ InconsistentAlleles

Miscellaneous Attributes (dbSNP): Set all Clear all
☑ None ☑ Clinically Associated ☑ MAF >= 5% in Some Population ☑ MAF >= 5% in All Populations
☑ Appears in OMIM/OMIA ☑ Has Microattribution/Third-Party Annotation ☑ Submitted by Locus-Specific Database ☑ Genotype Conflict
☑ Ref SNP Cluster has Nonoverlapping Alleles ☑ Some Assembly's Allele Does Not Match Observed

Coloring Options

SNP Feature for Color Specification: Function Set defaults

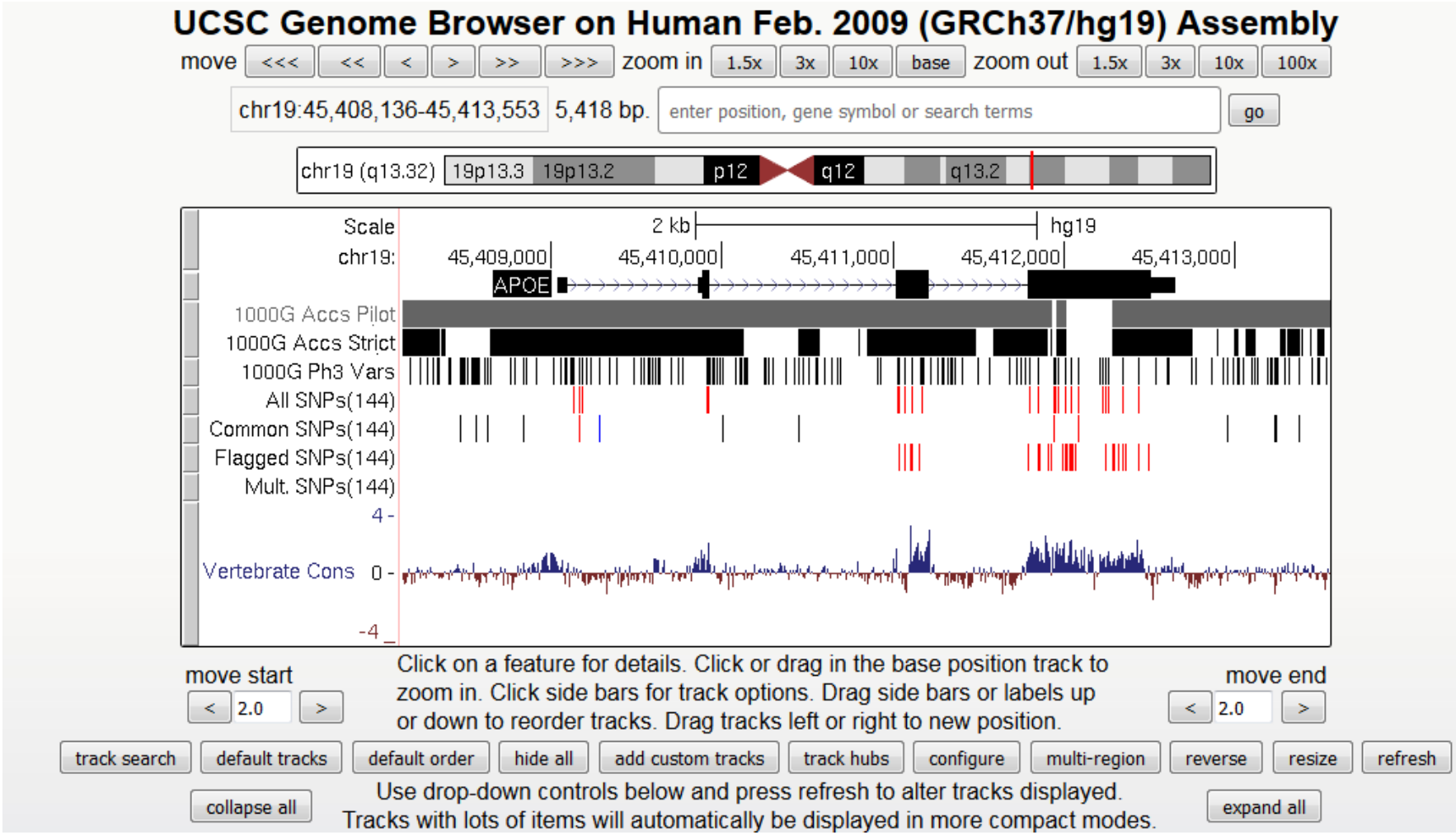
The selected "Feature for Color Specification" above has the selection of colors below for each attribute. Only the color options for the feature selected above will be used to color items; color options for other features will to weakest, is red, green, blue, gray, and black.

Unknown black Locus black Coding - Synonymous green Coding - Non-Synonymous red
Untranslated blue Intron black Splice Site gray

1. Set Single Nucleotide Polymorphism only

2. Set 1000 Genomes Project only

3. Remove all except missense variant



Display mode: dense Submit

Include Chimp state and observed human alleles in name:
(If enabled, chimp allele is displayed first, then '>', then human alleles).

Use Gene Tracks for Functional Annotation

Filtering Options

Minimum Average Heterozygosity: 0
Maximum Weight: 1 *Range: 1, 2 or 3; SNPs with higher weights are less reliable*
Minimum number of distinct Submitters: 0
Minor Allele Frequency range: 0 to 0.5 *Range: 0.0 - 0.5*
Minimum chromosome sample count (2N) for Allele Frequency data: 0

Filter by attribute:
Check the boxes below to include SNPs with those attributes. In order to be displayed, a SNP must pass the filter for each category. Some assemblies may not contain any SNPs that have some of the listed attributes.

Class: Set all Clear all
Unknown, Single Nucleotide Polymorphism, In/Del, Heterozygous, Microsatellite, Named, Mnp, Insertion, Deletion

Validation: Set all Clear all
Unknown, By Cluster, By Frequency, By Submitter, By 2 Hit / 2 Allele, By HapMap, By 1000 Genomes Project

Function: Set all Clear all
Unknown, synonymous_variant, intron_variant, downstream_gene_variant, upstream_gene_variant, nc_transcript_variant, stop_gained, missense_variant, stop_lost, frameshift_variant, inframe_indel, 3_prime_UTR_variant, 5_prime_UTR_variant, splice_acceptor_variant, splice_donor_variant

Molecule Type: Set all Clear all
Unknown, Genomic, cDNA

Unusual Conditions (UCSC): Set all Clear all
None, RefAlleleMismatch, RefAlleleRevComp, DuplicateObserved, MixedObserved, FlankMismatchGenomeLonger, FlankMismatchGenomeEqual, FlankMismatchGenomeShorter, NamedDeletionZeroSpan, NamedInsertionNonzeroSpan, SingleClassLongerSpan, SingleClassZeroSpan, SingleClassTriAllelic, SingleClassQuadAllelic, ObservedWrongFormat, ObservedTooLong, ObservedContainslupac, ObservedMismatch, MultipleAlignments, NonIntegerChromCount, AlleleFreqSumNot1, SingleAlleleFreq, InconsistentAlleles

Miscellaneous Attributes (dbSNP): Set all Clear all
None, Clinically Associated, MAF >= 5% in Some Population, MAF >= 5% in All Populations, Appears in OMIM/OMIA, Has Microattribution/Third-Party Annotation, Submitted by Locus-Specific Database, Genotype Conflict, Ref SNP Cluster has Nonoverlapping Alleles, Some Assembly's Allele Does Not Match Observed

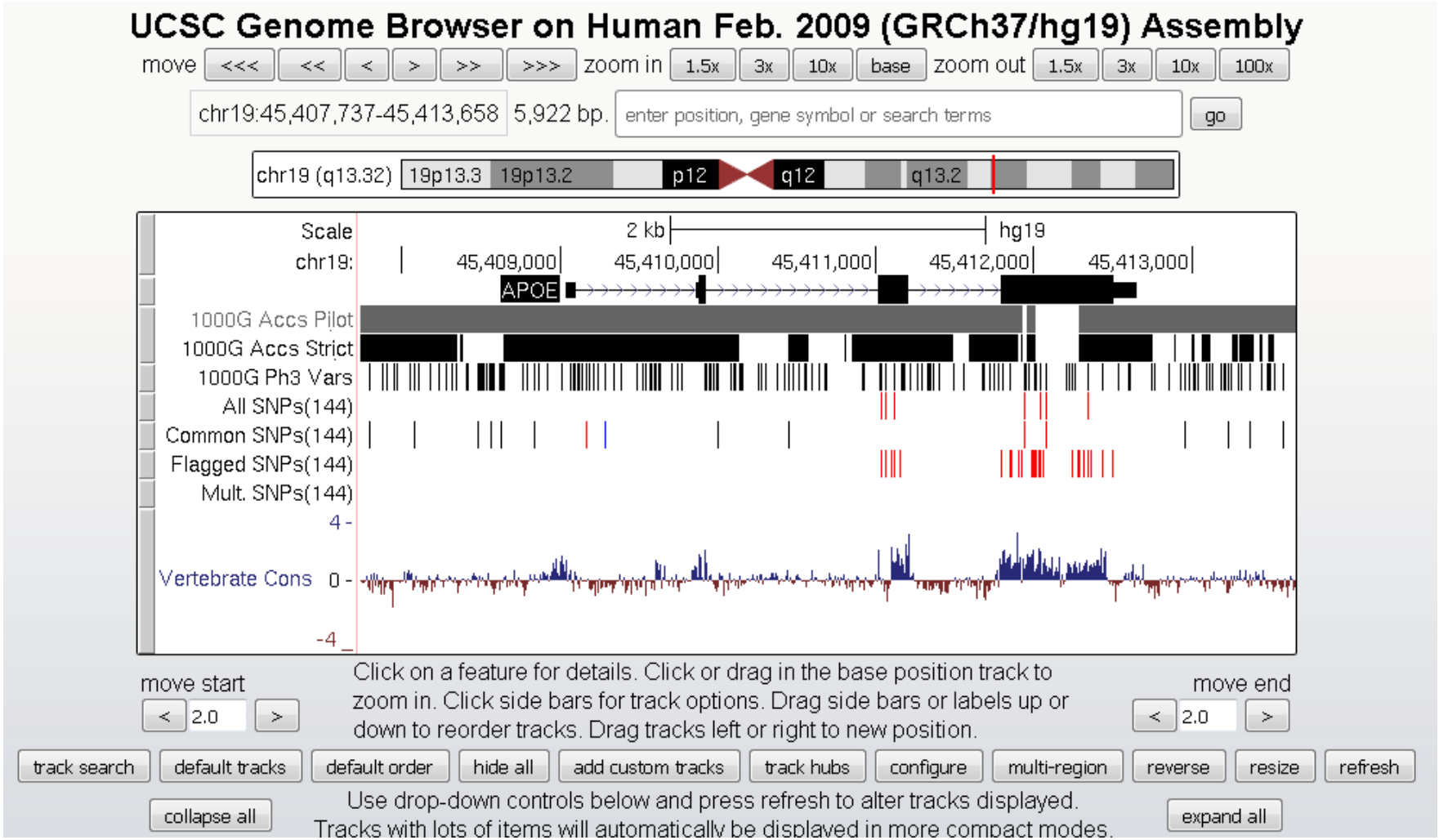
Coloring Options

SNP Feature for Color Specification: Function Set defaults

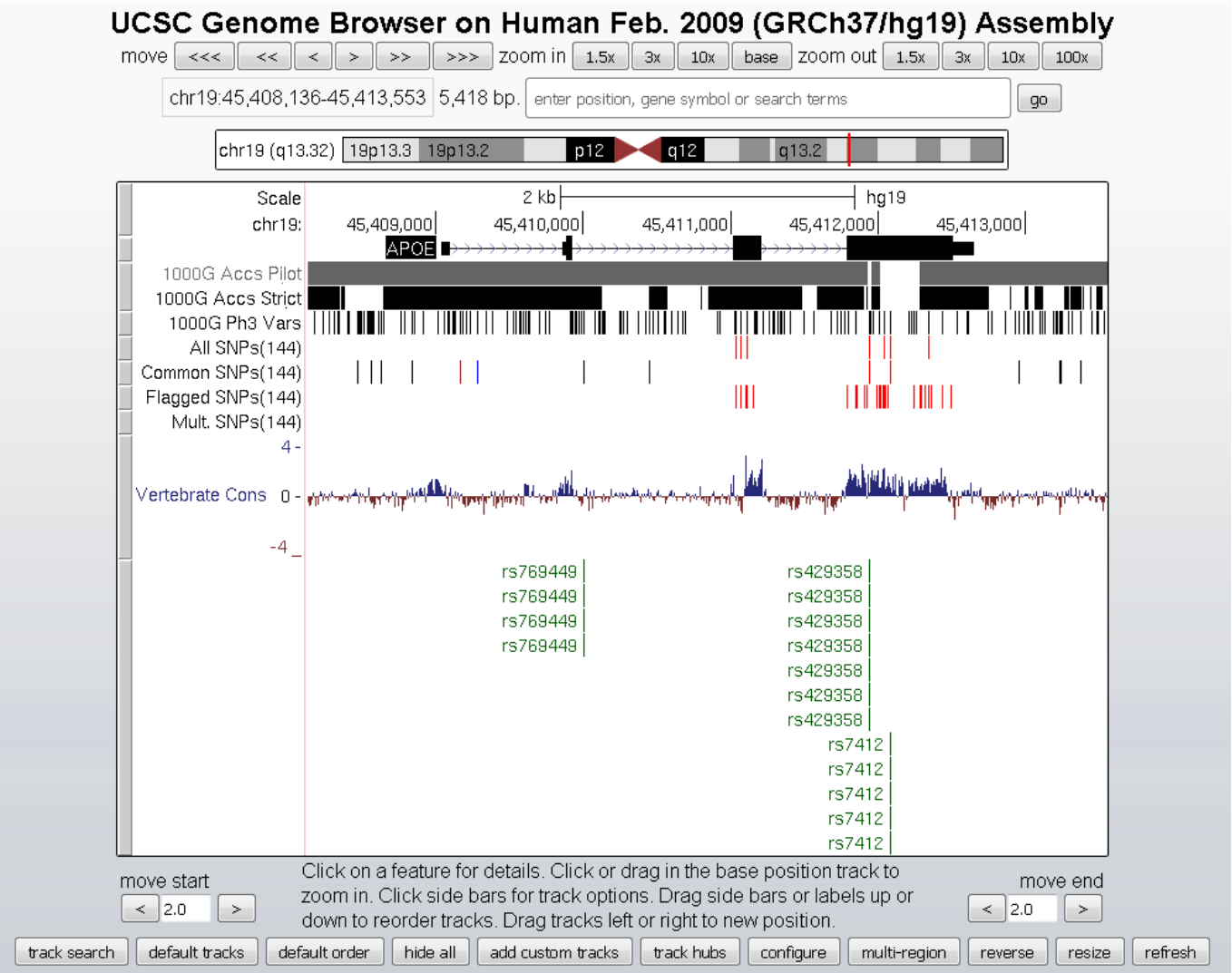
The selected "Feature for Color Specification" above has the selection of colors below for each attribute. Only the color options for the feature selected above will be used to color items; color options for other features will to weakest, is red, green, blue, gray, and black.

Unknown black, Locus black, Coding - Synonymous green, Coding - Non-Synonymous red
Untranslated blue, Intron black, Splice Site gray

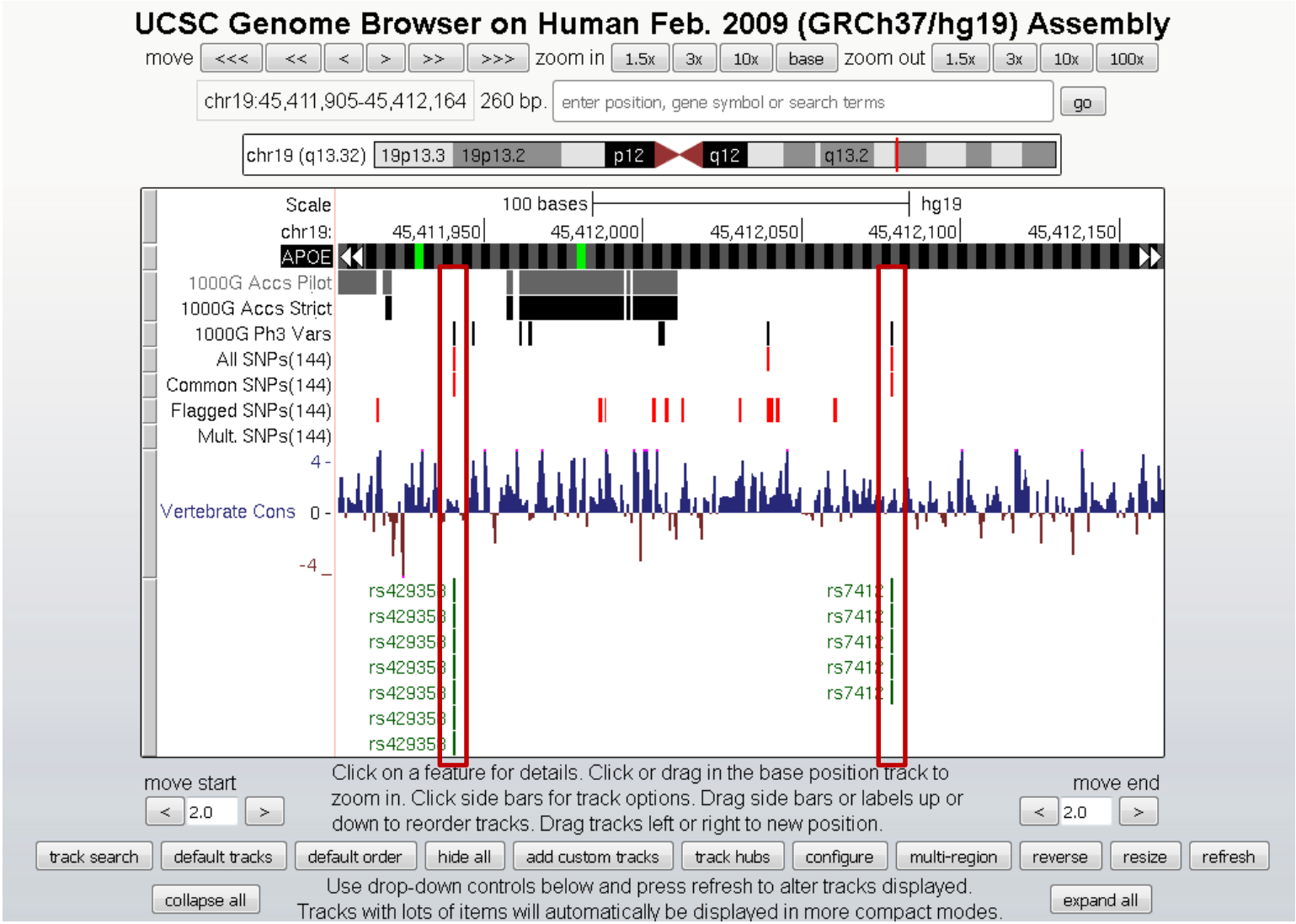
Select only clinical associated variants



Why are there less clinically associated, missensed, 1000 genome SNPs than Flagged SNPs?

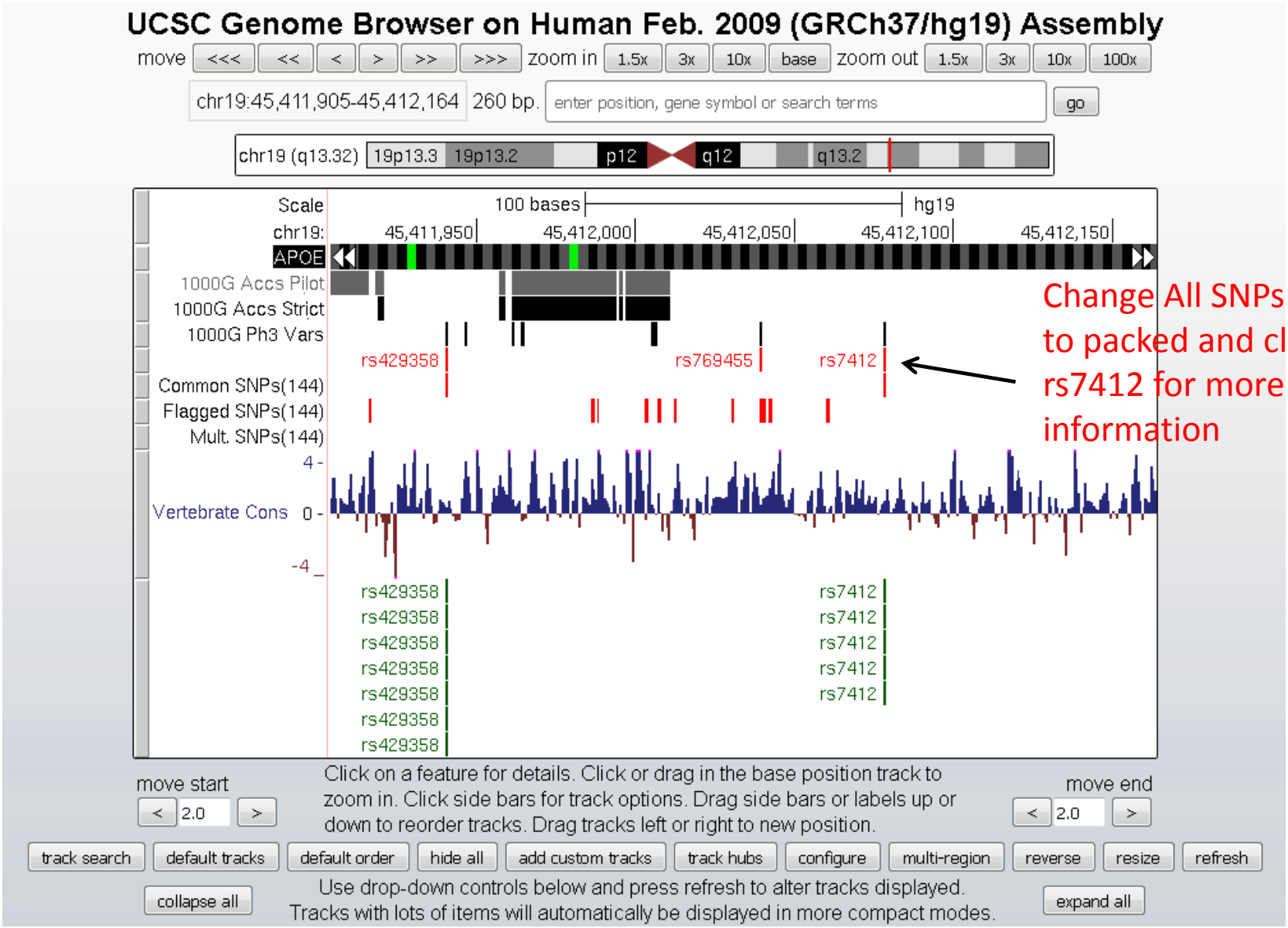


APOE contains two well known Alzhimer's disease risk associated SNPs rs429358 and rs7412
Bring up GWAS Catalog track (under Phenotype and Literature) to find where these are.



Note:

1. The two SNPs are NOT in the Flagged SNPs track because their MAF >= 1%
2. The two SNPs lie in regions inaccessible to short read NGS.


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr19:45,411,905-45,412,164 260 bp.
enter position, gene symbol or search terms
go
chr19 (q13.32) 19p13.3 19p13.2 p12 q12 q13.2
Scale
100 bases
hg19
chr19:
45,411,950
45,412,000
45,412,050
45,412,100
45,412,150
APOE
1000G Accs Pilot
1000G Accs Strict
1000G Ph3 Vars
rs429358
rs769455
rs7412
Common SNPs(144)
Flagged SNPs(144)
Mult. SNPs(144)
4 -
Vertebrate Cons
0 -
-4 _
rs429358
rs7412
Change All SNPs track to packed and click on rs7412 for more information
move start
2.0
move end
2.0
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
track search
default tracks
default order
hide all
add custom tracks
track hubs
configure
multi-region
reverse
resize
refresh
collapse all
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.
expand all

Simple Nucleotide Polymorphisms (dbSNP 144)

dbSNP build 144 rs7412

dbSNP: rs7412
Position: chr19:45412079-45412079
Band: 19q13.32
Genomic Size: 1
View DNA for this feature (hg19/Human)

Click to link out to dbSNP

Summary: C>C/T (chimp allele displayed first, then '>', then human alleles)
Strand: +
Observed: C/T
Reference allele: C

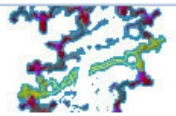
Chimp allele:	C	**Chimp strand:**	+	**Chimp position:**	chr19:50098660-50098660
Orangutan allele:	C	**Orangutan strand:**	+	**Orangutan position:**	chr19:46149262-46149262
Macaque allele:	C	**Macaque strand:**	+	**Macaque position:**	chr19:50949311-50949311

Class	single
Validation	by-cluster,by-frequency,by-2hit-2allele,by-hapmap,by-1000genomes
Function	missense_variant
Molecule Type	genomic
Average Heterozygosity	0.139 +/- 0.224
Weight	1
Submitter Handles	1000GENOMES, BUSHMAN, CGAP-GAI, COMPLETE_GENOMICS, CUORCGL, DEBNICK, EVA-GONL, EVA_DECODE, EVA_EXAC, EVA_GENOME_DK, EVA_SVP, EVA_UK10K_ALSPAC, EVA_UK10K_TWINSUK, EXOME_CHIP, GMI, ILLUMINA, JMKIDD_LAB, KRIBB_YJKIM, LEE, NCBI-CURATED-RECORDS, NHLBI-ESP, OMICIA, OMIM-CURATED-RECORDS, PAGE_STUDY, RSG_UW, SC_SNP, SSMP
Allele Frequencies	C: 92.492% (4632 / 5008); T: 7.508% (376 / 5008)

Miscellaneous properties annotated by dbSNP:
SNP is in OMIM/OMIA and/or at least one submitter is a Locus-Specific Database ("clinically associated")
SNP is in OMIM/OMIA
SNP has a microattribution or third-party annotation
SNP was submitted by Locus-Specific Database
Minor Allele Frequency is at least 5% in at least one population assayed

NCBI

dbSNP
Short Genetic Variations



dbVar | ClinVar | GaP | PubMed | Nucleotide | Protein

Search small variations in dbSNP or large structural variations in dbVar

Search Entrez dbSNP Go

Have a question about dbSNP? Try searching the SNP FAQ Archive!

Go

GENERAL
RSS Feed
Contact Us
Site Map
dbSNP Homepage
NCBI Variation Resources
Announcements
dbSNP Summary
FTP Download
SNP SUBMISSION
DOCUMENTATION
SEARCH
RELATED SITES

Reference SNP (refSNP) Cluster Report: rs7412 ** With Pathogenic allele **

RefSNP	
Organism:	human (Homo sapiens)
Molecule Type:	Genomic
Created/Updated in build:	52/146
Map to Genome Build:	107/Weight
Validation Status:	
Citation:	PubMed
Association:	NHGRI GWAS PheGenI

Allele	
Variation Class:	SNV: single nucleotide variation
RefSNP Alleles:	C/T (FWD)
Allele Origin:	C:germline T:germline
Ancestral Allele:	C
Variation Viewer:	VarView OMIM
Clinical Significance:	With Pathogenic allele [ClinVar]
MAF/MinorAlleleCount:	T=0.0751/376
MAF Source:	1000 Genomes

HGVS Names
NC_000019.10:g.44908822C>T
NC_000019.9:g.45412079C>T
NG_007084.2:g.8041C>T
NM_000041.3:c.526C>T
NM_001302688.1:c.604C>T
NM_001302689.1:c.526C>T
NM_001302690.1:c.526C>T
NM_001302691.1:c.526C>T
NP_000032.1:p.Arg176Cys

...more

SNP Details are organized in the following sections:

GeneView | Map | Submission | Fasta | Resource | Diversity | Validation

Integrated Maps (Hint: click on 'Chr Pos' to see variant in the new NCBI variation viewer)

Assembly	Annotation Release	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Neighbor SNP	Map Method
GRCh38.p2	107	19	44908822	NT_011109.17	17667948	Fwd	C	Fwd	view	mapup
GRCh37.p13	105	19	45412079	NT_011109.16	17680297	Fwd	C	Fwd	view	blast

GeneView

GeneView via analysis of contig annotation: APOE *apolipoprotein E*

View more variation on this gene (click to hide).

Clinical Source: in gene region cSNP has frequency double hit Go

Primary Assembly Mapping

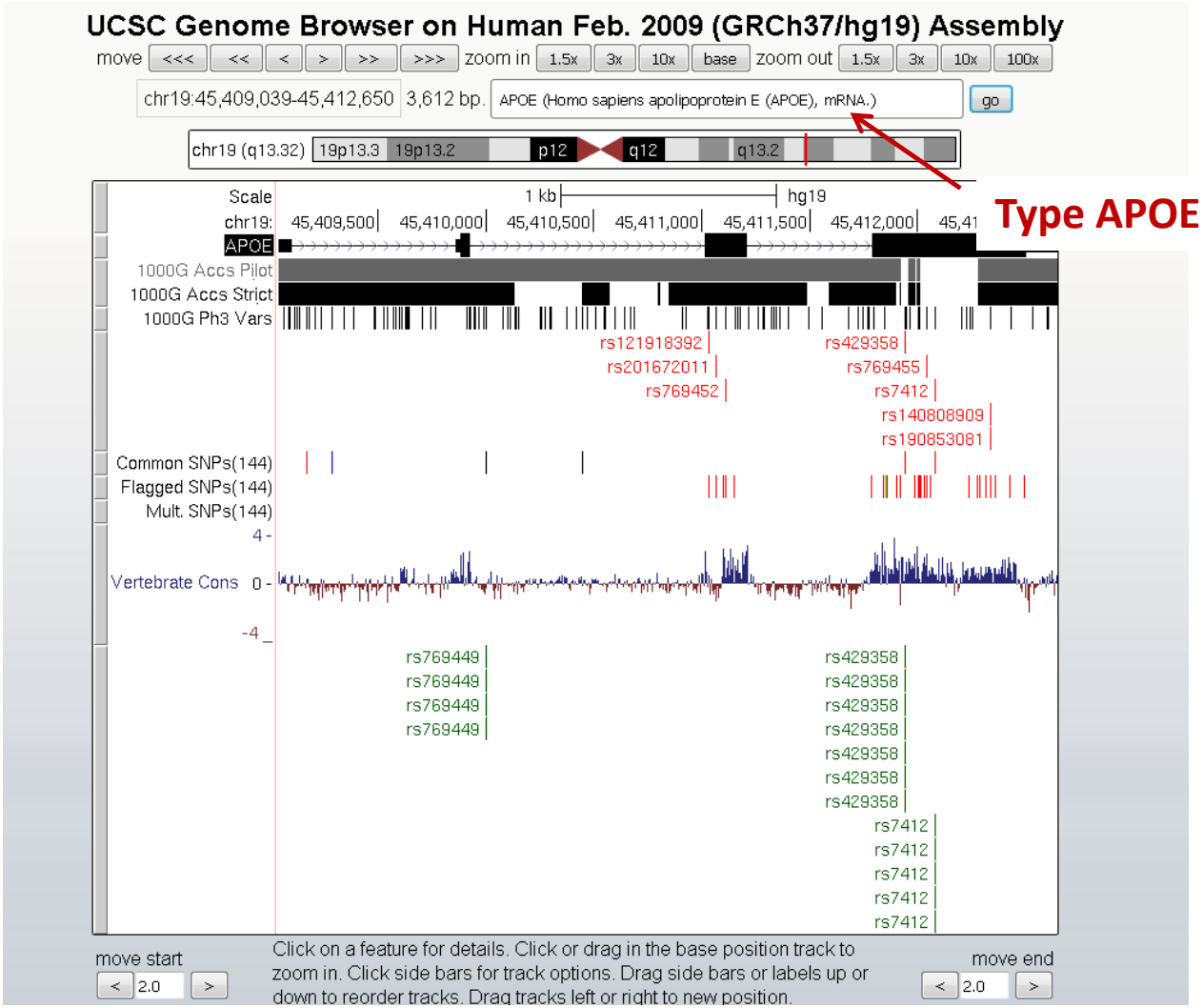
Assembly	SNP to Chr	Chr	Chr position	Contig	Contig position	Allele
GRCh38.p2	Fwd	19	44908822	NT_011109.17	17667948	C

RefSeqGene Mapping

RefSeqGene	Gene (ID)	SNP to RefSeqGene	Position	Allele
NG_007084.2	APOE (348)	Fwd	8041	C

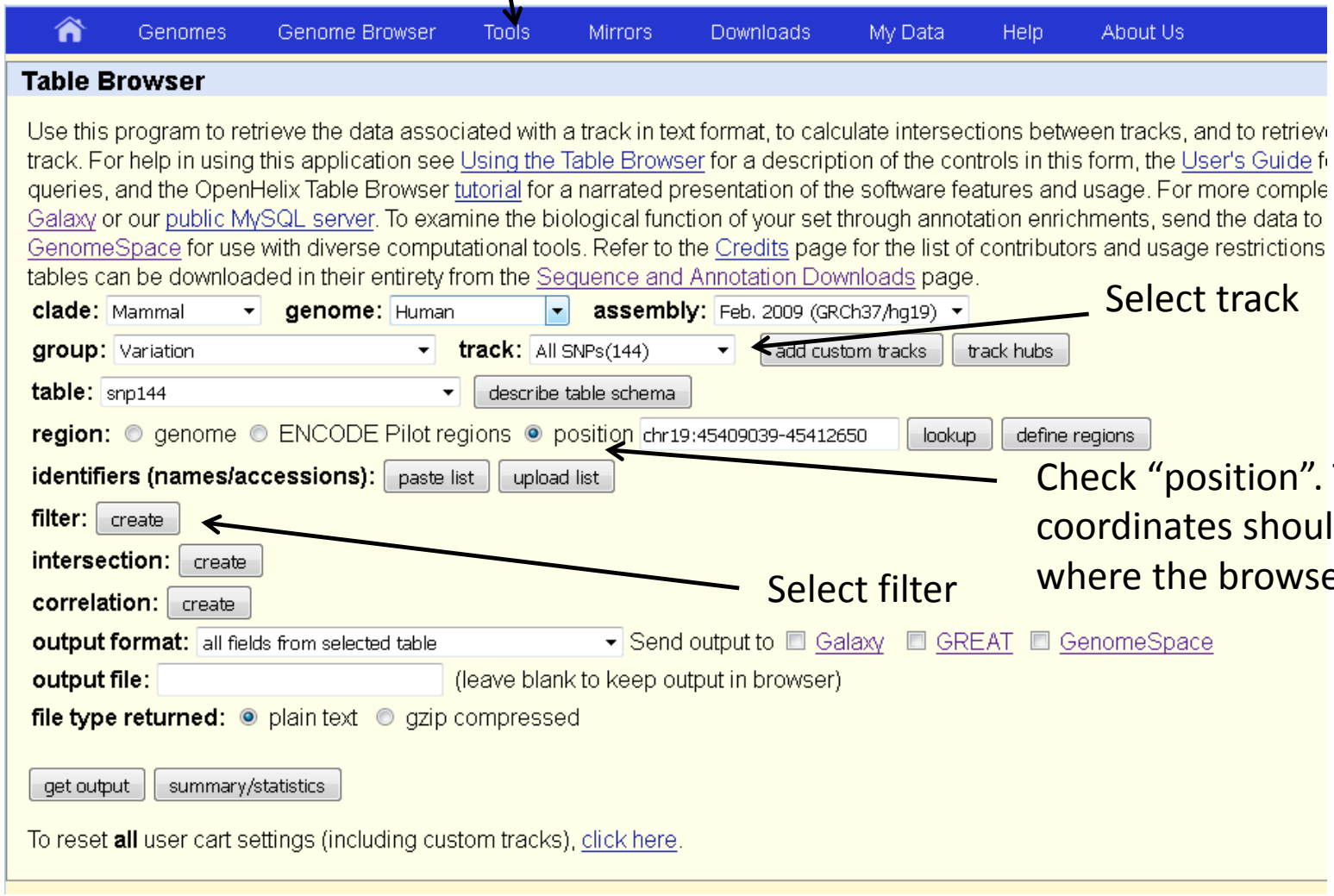
Exporting specific SNPs using table browser

Aim is to download all 1000genome clinically associated missense SNPs over *APOE*



Type APOE and click “go”.

Select “Table browser” from “Tools” menu”



Select track

Check “position”. The coordinates should be where the browser last was.

Select filter

Filter

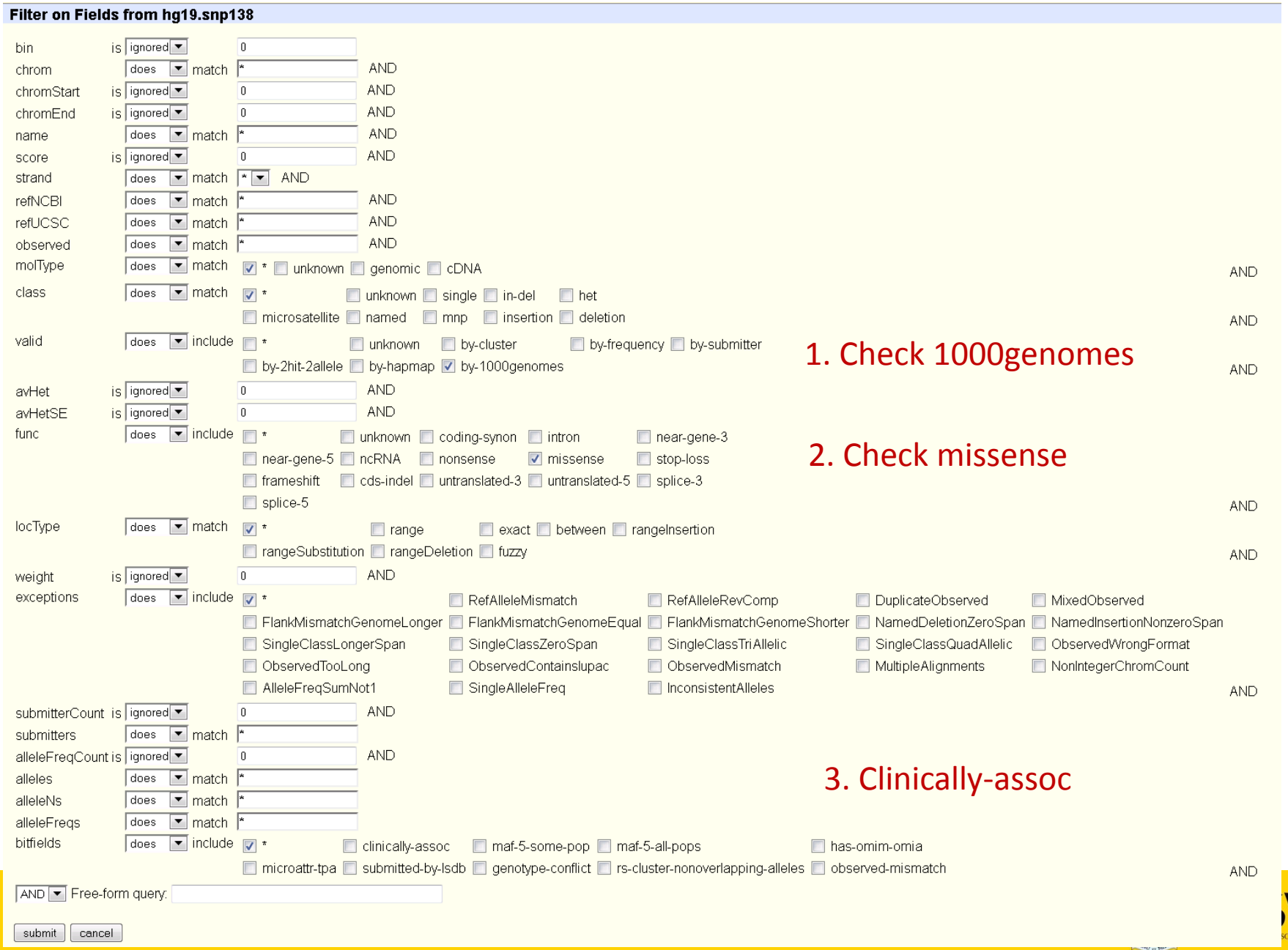

Filter on Fields from hg19.snp138
bin is ignored 0
chrom does match * AND
chromStart is ignored 0 AND
chromEnd is ignored 0 AND
name does match * AND
score is ignored 0 AND
strand does match * AND
refNCBI does match * AND
refUCSC does match * AND
observed does match * AND
molType does match * unknown genomic cDNA AND
class does match * unknown single in-del het
microsatellite named mnp insertion deletion AND
valid does include * unknown by-cluster by-frequency by-submitter
by-2hit-2allele by-hapmap by-1000genomes AND
avHet is ignored 0 AND
avHetSE is ignored 0 AND
func does include * unknown coding-synon intron near-gene-3
near-gene-5 ncRNA nonsense missense stop-loss
frameshift cds-indel untranslated-3 untranslated-5 splice-3
splice-5 AND
locType does match * range exact between rangeInsertion
rangeSubstitution rangeDeletion fuzzy AND
weight is ignored 0 AND
exceptions does include * RefAlleleMismatch RefAlleleRevComp DuplicateObserved MixedObserved
FlankMismatchGenomeLonger FlankMismatchGenomeEqual FlankMismatchGenomeShorter NamedDeletionZeroSpan NamedInsertionNonzeroSpan
SingleClassLongerSpan SingleClassZeroSpan SingleClassTriAllelic SingleClassQuadAllelic ObservedWrongFormat
ObservedTooLong ObservedContainslupac ObservedMismatch MultipleAlignments NonIntegerChromCount
AlleleFreqSumNot1 SingleAlleleFreq InconsistentAlleles AND
submitterCount is ignored 0 AND
submitters does match *
alleleFreqCount is ignored 0 AND
alleles does match *
alleleNs does match *
alleleFreqs does match *
bitfields does include * clinically-assoc maf-5-some-pop maf-5-all-pops has-omim-omia
microattr-tpa submitted-by-lsdb genotype-conflict rs-cluster-nonoverlapping-alleles observed-mismatch AND
AND Free-form query:
submit cancel
1. Check 1000genomes
2. Check missense
3. Clinically-assoc

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

group: Variation **track:** All SNPs(144) add custom tracks track hubs

table: snp144 describe table schema

region: genome ENCODE Pilot regions position chr19:45409039-45412650 lookup define regions

identifiers (names/accessions): paste list upload list

filter: edit clear

intersection: create

correlation: create

output format: BED - browser extensible data Send output to Galaxy GREAT GenomeSpace

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Select BED as the output format
Optionally type in a name for the output file to download the file.

Output snp144 as BED

Include custom track header:

name= tb_snp144

description= table browser query on snp144

visibility= pack

url=

Create one BED record per:

- Whole Gene
- Upstream by 200 bases
- Downstream by 200 bases

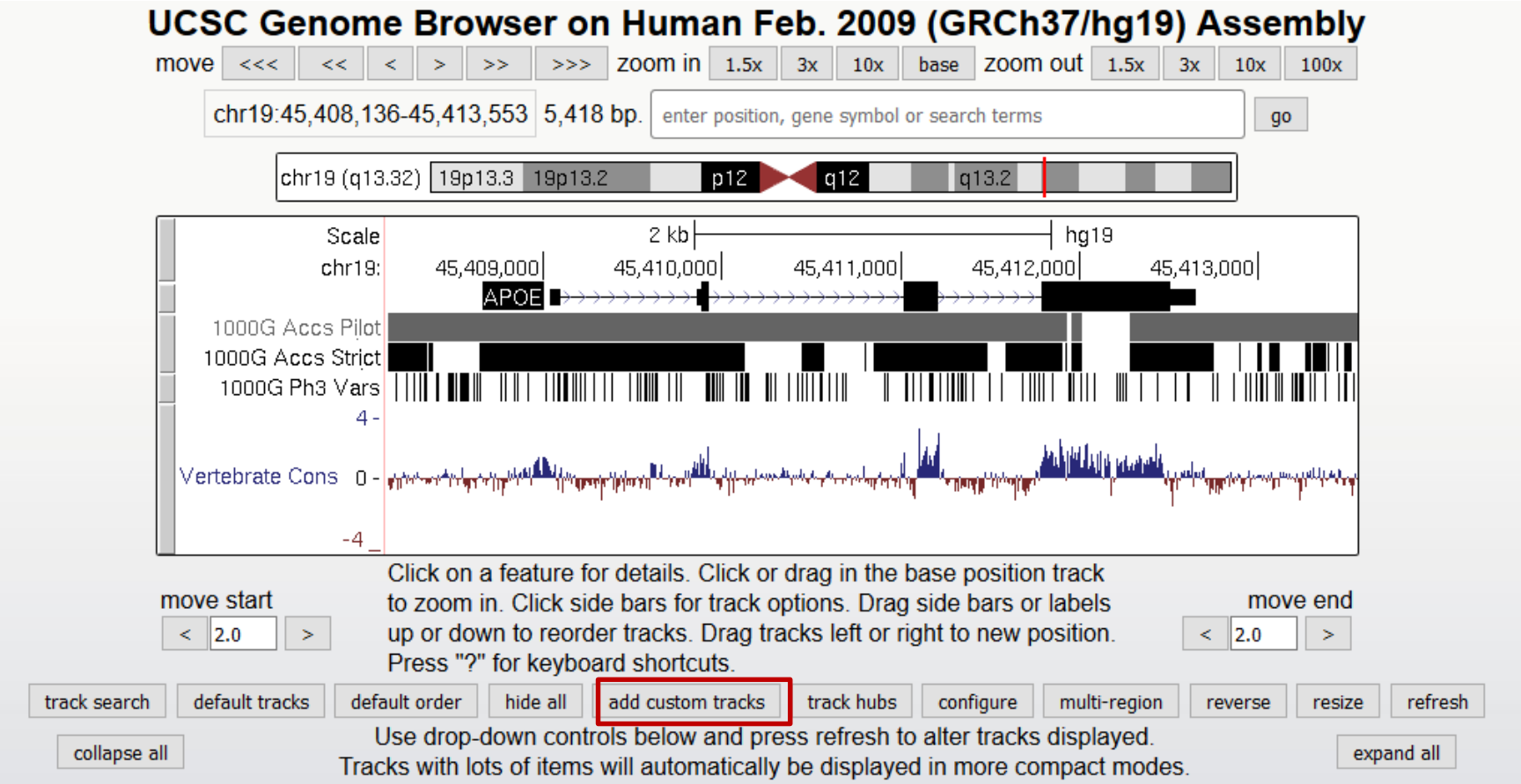
Note: if a feature is close to the beginning or end of a chromosome and upstream/downst
the edge of the chromosome.

get BED cancel

BED file from UCSC – should be 8 SNPs in total

```
chr19   45411033        45411034        rs121918392     0       +
chr19   45411063        45411064        rs201672011     0       +
chr19   45411109        45411110        rs769452        0       +
chr19   45411940        45411941        rs429358        0       +
chr19   45412039        45412040        rs769455        0       +
chr19   45412078        45412079        rs7412  0       +
chr19   45412336        45412337        rs140808909     0       +
chr19   45412339        45412340        rs190853081     0       +
```

Annotating variants in UCSC



The VCF file format

- Most common format to store variant/mutation information.
- In text format, but difficult to “view” in text editor/Excel.
- Header contains useful information.

```
##fileformat=VCFv4.0
##filedat=20101112
##datarelease=20100804
##samples=629
##description="Where BI calls are present, genotypes and alleles are from BI.  In there absence, UM genotypes are used.  If neither are available, no genotype information
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth (only filtered reads used for calling)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Log-scaled likelihoods for AA,AB,BB genotypes where A=ref and B=alt; not applicable if site is not biallelic">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GD,Number=1,Type=Float,Description="Genotype dosage.  Expected count of non-ref alleles [0,2]">
##FORMAT=<ID=OG,Number=1,Type=String,Description="Original Genotype input to Beagle">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=CB,Number=.,Type=String,Description="List of centres that called, UM (University of Michigan), BI (Broad Institute), BC (Boston College), NCBI">
##INFO=<ID=EUR_R2,Number=1,Type=Float,Description="R2 From Beagle based on European Samples">
##INFO=<ID=AFR_R2,Number=1,Type=Float,Description="R2 From Beagle based on AFRICAN Samples">
##INFO=<ID=ASN_R2,Number=1,Type=Float,Description="R2 From Beagle based on Asian Samples">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00098 HG00100 HG00106 HG00112 HG00114 HG00116 HG00117 HG00118 HG00119 HG00120 HG00122 HG00123 HG0
19      45409113        rs9282609       C       T       .       PASS    DP=1539;AF=0.008;CB=UM,BI,BC,NCBI;AFR_R2=0.799  GT:AD:DP:GD:GL:GQ:OG    ./.:.:.:.:.,.,.:.:./.   ./.
19      45409167        rs440446        C       G       .       PASS    DP=1412;AF=0.628;CB=UM,BI,BC,NCBI;EUR_R2=0.892;AFR_R2=0.828     GT:AD:DP:GD:GL:GQ:OG    0|0:.:.:.:.
19      45409283        rs877973        C       A       .       PASS    DP=1678;AF=0.009;CB=UM,BI,BC,NCBI;EUR_R2=0.574;AFR_R2=0.65      GT:AD:DP:GD:GL:GQ:OG    0|0:.:.:.:.
19      45409482        .       A       G       .       PASS    DP=1943;AF=0.002;CB=UM,BI,BC,NCBI       GT:AD:DP:GD:GL:GQ:OG    ./.:.:.:.:.,.,.:.:./.   ./.:.:.:.:.,.,.:.:.
19      45409488        .       C       T       .       PASS    DP=1923;AF=0.001;CB=UM,NCBI     GT:AD:DP:GD:GL:GQ:OG    0|0:.:3:0.000:0,3,26:29:./.     0|0:.:0:0.000:0,0,0
19      45409579        rs769448        C       T       .       PASS    DP=430;AF=0.000;CB=BI,BC,NCBI;EUR_R2=0.578      GT:AD:DP:GD:GL:GQ:OG    0|0:.:.:.:.:.,.,.:18.36:./.
19      45409595        .       G       T       .       PASS    DP=913;AF=0.014;CB=UM,NCBI      GT:AD:DP:GD:GL:GQ:OG    ./.:.:.:.:.,.,.:.:./.   ./.:.:.:.:.,.,.:.:./.   ./.
19      45409929        .       G       A       .       PASS    DP=1691;AF=0.001;CB=UM,BC,NCBI  GT:AD:DP:GD:GL:GQ:OG    0|0:.:0:0.000:0,0,0:26:./.      0|0:.:0:0.000:0,0,0
```

https://samtools.github.io/hts-specs/VCFv4.2.pdf

Add Custom Tracks

clade Mammal genome Human assembly Feb. 2009 (GRCh37/hg19)

Display your own data as custom annotation tracks in the browser. Data must be formatted in bigBed, bigChain, bigGenePred, bigMaf, bigPsl, bigWig, BAM, VCF, BED, BED detail, bedGraph, broadPeak, CRAM, GFF, GTF, MAF, narrowPeak, Personal Genome SNP, PSL, or WIG formats. To configure the display, set track and browser line attributes as described in the User's Guide. Data in the bigBed, bigWig, bigGenePred, BAM and VCF formats can be provided via only a URL or embedded in a track line in the box below. Examples are here.

Paste URLs or data: Or upload: Browse... No file selected. Submit

Clear

Upload APOE_1000genomes.vcf

Optional track documentation: Or upload: Browse... No file selected.

Clear

Click here for an HTML document template that may be used for Genome Browser track descriptions.

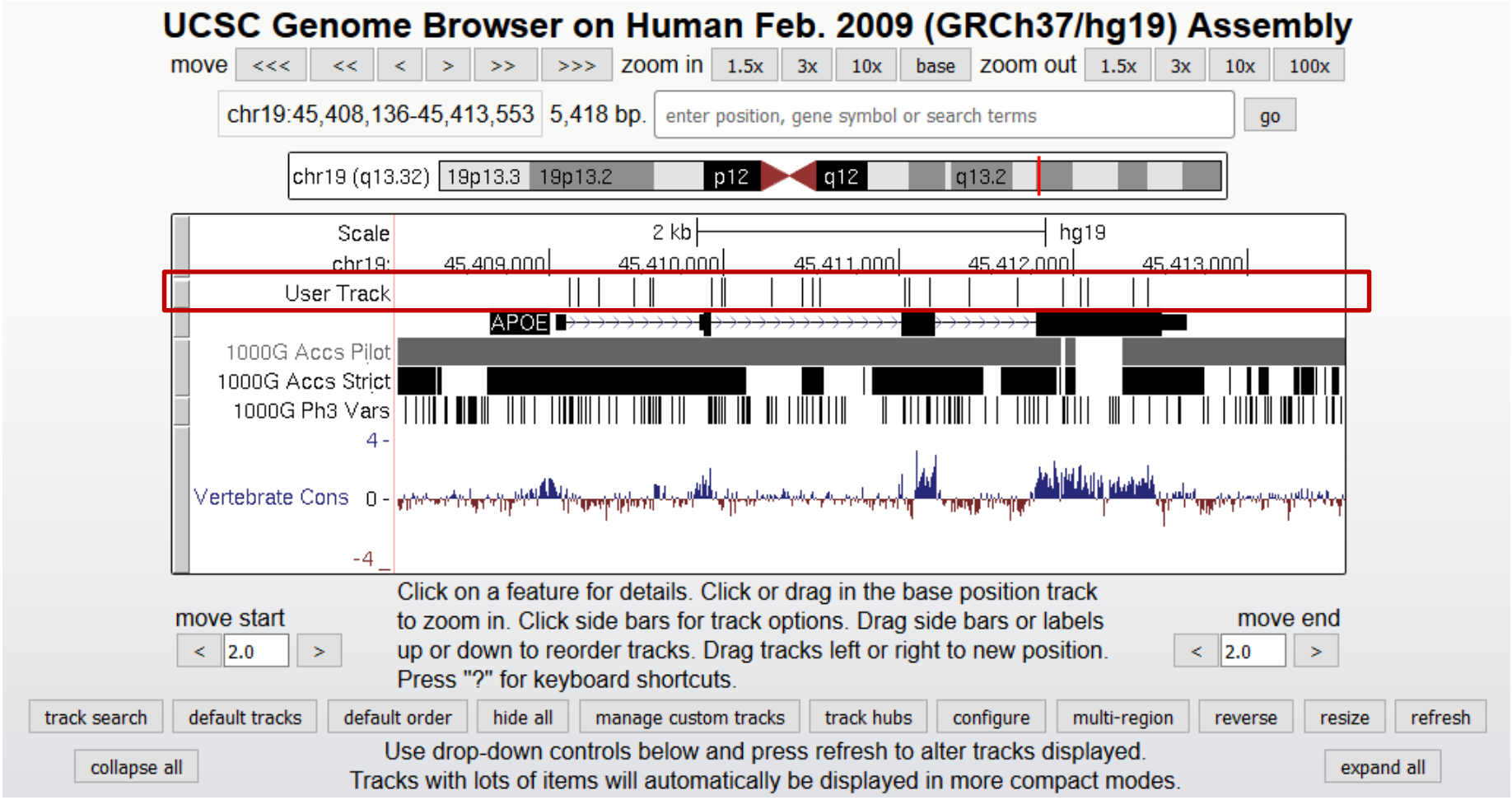
Manage Custom Tracks

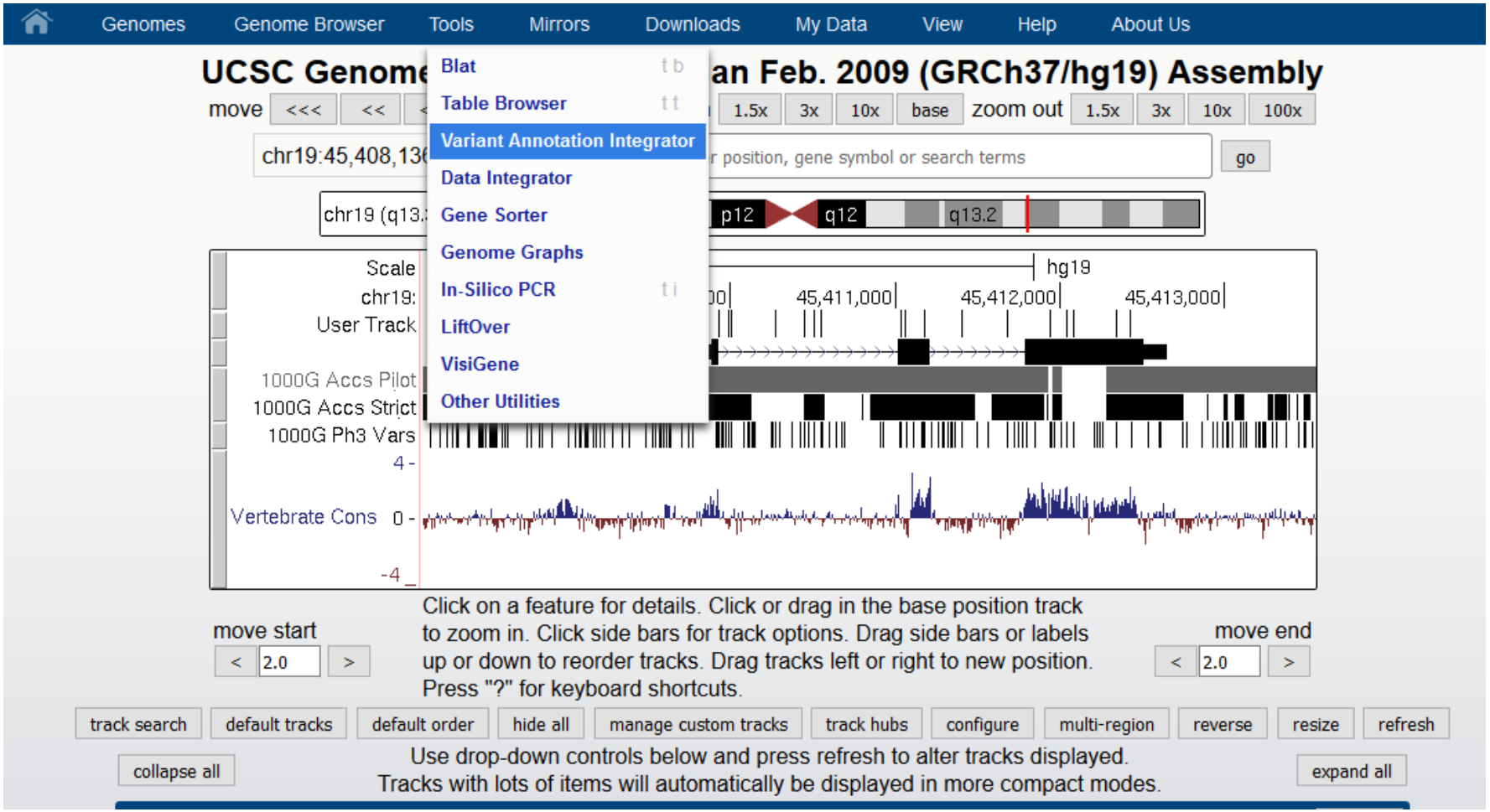
genome: Human **assembly:** Feb. 2009 (GRCh37/hg19) [hg19]

Name	Description	Type	Doc	delete
User Track	User Supplied Track	vcf		☐

view in Genome Browser go

add custom tracks




Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
Blat t b
Table Browser t t
Variant Annotation Integrator
Data Integrator
Gene Sorter
Genome Graphs
In-Silico PCR t i
LiftOver
VisiGene
Other Utilities
an Feb. 2009 (GRCh37/hg19) Assembly
move
zoom out
1.5x
3x
10x
base
100x
go
p12
q12
q13.2
Scale
chr19:
User Track
hg19
45,411,000
45,412,000
45,413,000
1000G Accs Pilot
1000G Accs Strict
1000G Ph3 Vars
Vertebrate Cons
move start
move end
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.
track search
default tracks
default order
hide all
manage custom tracks
track hubs
configure
multi-region
reverse
resize
refresh
collapse all
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.
expand all

Variant Annotation Integrator

Select Genome Assembly and Region

clade: Mammal | genome: Human | assembly: Feb. 2009 (GRCh37/hg19)

region to annotate: genome

Select Variants

variants: User Supplied Track

maximum number of variants to be processed: 10,000

manage custom tracks | track hubs | To reset **all** user cart settings (including custom tracks), click here.

Select variants here (since we only have one uploaded this will be the default/only option.

Select Genes

The gene predictions selected here will be used to determine the effect of each variant on genes, for example intronic, missense, splice site, intergenic etc.

UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)

Select Regulatory Annotations

The annotations in this section provide predicted regulatory regions based on various experimental data. When a variant overlaps an annotation selected here, the consequence term regulatory_region_variant will be assigned. Follow the links to description pages that explain how each dataset was constructed. Some datasets cover a significant portion of the genome and it may be desirable to filter these annotations by cell type and/or score in order to avoid an overabundance of hits.

- ☐ DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
  - + **filter items**
- ☐ Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
  - + **filter items**

Select More Annotations (optional)

- − **Database of Non-synonymous Functional Predictions (dbNSFP)**

dbNSFP (Liu *et al.* 2013) release 2.0 provides pre-computed scores and predictions of functional significance from a variety of tools. Every possible coding change to transcripts in Gencode release 9 (Ensembl 64, Dec. 2011) gene predictions has been evaluated. *Note: This may not encompass all transcripts in your selected gene set.*

Set all | Clear all

- ☑ SIFT (D = damaging, T = tolerated)
- ☑ PolyPhen-2 with HumDiv training set (D = probably damaging, P = possibly damaging, B = benign)
- ☑ PolyPhen-2 with HumVar training set (D = probably damaging, P = possibly damaging, B = benign)
- ☐ MutationTaster (A = disease causing automatic, D = disease causing, N = polymorphism, P = polymorphism automatic)
- ☐ MutationAssessor (high or medium: predicted functional; low or neutral: predicted non-functional)
- ☐ Likelihood ratio test (LRT) (D = deleterious, N = Neutral, U = unknown)

Set all | Clear all

- [x] SIFT (D = damaging, T = tolerated)
- [x] PolyPhen-2 with HumDiv training set (D = probably damaging, P = possibly damaging, B = benign)
- [x] PolyPhen-2 with HumVar training set (D = probably damaging, P = possibly damaging, B = benign)
- [ ] MutationTaster (A = disease causing automatic, D = disease causing, N = polymorphism, P = polymorphism automatic)
- [ ] MutationAssessor (high or medium: predicted functional; low or neutral: predicted non-functional)
- [ ] Likelihood ratio test (LRT) (D = deleterious, N = Neutral, U = unknown)
- [ ] InterPro protein domains
- [ ] GERP++ Rejected Substitutions (RS)
- [ ] GERP++ Neutral Rate (NR)

+ Transcript status

− Known variation

- [x] Include dbSNP rs# ID if one exists

+ COSMIC

+ Conserved elements

+ Conservation scores

Define Filters

+ Functional role

+ Known variation

+ Conservation

Configure Output

output format: Variant Effect Predictor (HTML)

Change to HTML for easier viewing

output file: (leave blank to keep output in browser)

file type returned: (•) plain text ( ) gzip compressed (ignored if output file is blank)

Get results

UCSC Genome Bioinformatics

Annotated Variants in VEP/HTML format

Variants: User Supplied Track
Transcripts: UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics) (hg19.knownGene)
dbSNP: Simple Nucleotide Polymorphisms (dbSNP 147) (/gbdb/hg19/vai/snp147.bed4.bb)
Keys for Extra column items:
SIFT: SIFT (D = damaging, T = tolerated)
PP2HVAR: PolyPhen-2 with HumVar training set (D = probably damaging, P = possibly damaging, B = benign)
PP2HDIV: PolyPhen-2 with HumDiv training set (D = probably damaging, P = possibly damaging, B = benign)

Uploaded Variation	Location	Allele	Gene	Feature	Feature type	Consequence	Position in cDNA	Position in CDS	Position in protein	Amino acid change	Codon change	Co-located Variation	Extra
rs9282609	chr19:45409113	T	TOMM40	uc002ozz.3	Transcript	downstream_gene_variant	-	-	-	-	-	rs9282609	DISTANCE=2176
rs9282609	chr19:45409113	T	TOMM40	uc002ozx.4	Transcript	downstream_gene_variant	-	-	-	-	-	rs9282609	DISTANCE=2167
rs9282609	chr19:45409113	T	TOMM40	uc002ozy.4	Transcript	downstream_gene_variant	-	-	-	-	-	rs9282609	DISTANCE=2167
rs9282609	chr19:45409113	T	TOMM40	uc002paa.4	Transcript	downstream_gene_variant	-	-	-	-	-	rs9282609	DISTANCE=2167
rs9282609	chr19:45409113	T	APOE	uc002pab.3	Transcript	intron_variant	-	-	-	-	-	rs9282609	INTRON=1/3
rs440446	chr19:45409167	G	TOMM40	uc002ozz.3	Transcript	downstream_gene_variant	-	-	-	-	-	rs440446	DISTANCE=2230
rs440446	chr19:45409167	G	TOMM40	uc002ozx.4	Transcript	downstream_gene_variant	-	-	-	-	-	rs440446	DISTANCE=2221
rs440446	chr19:45409167	G	TOMM40	uc002ozy.4	Transcript	downstream_gene_variant	-	-	-	-	-	rs440446	DISTANCE=2221
rs440446	chr19:45409167	G	TOMM40	uc002paa.4	Transcript	downstream_gene_variant	-	-	-	-	-	rs440446	DISTANCE=2221
rs440446	chr19:45409167	G	APOE	uc002pab.3	Transcript	intron_variant	-	-	-	-	-	rs440446	INTRON=1/3
rs877973	chr19:45409283	A	TOMM40	uc002ozz.3	Transcript	downstream_gene_variant	-	-	-	-	-	rs877973	DISTANCE=2346
rs877973	chr19:45409283	A	TOMM40	uc002ozx.4	Transcript	downstream_gene_variant	-	-	-	-	-	rs877973	DISTANCE=2337
rs877973	chr19:45409283	A	TOMM40	uc002ozy.4	Transcript	downstream_gene_variant	-	-	-	-	-	rs877973	DISTANCE=2337
rs877973	chr19:45409283	A	TOMM40	uc002paa.4	Transcript	downstream_gene_variant	-	-	-	-	-	rs877973	DISTANCE=2337

...

rs7412	chr19:45412079	T	APOE	uc002pab.3	Transcript	missense_variant	609	526	176	R/C	Cgc/Tgc	rs7412	SIFT=D(0.020000); PP2HVAR=D(1.0); PP2HDIV=D(1.0); EXON=4/4

Further reading/resources

- 1000 genomes project (*www.**1000genomes**.org/)*
  - Phase 1 paper (*www.ncbi.nlm.nih.gov/pubmed/23128226*)
  - Phase 3 paper (*www.ncbi.nlm.nih.gov/pubmed/2643224*)