Recent work on Human Variation tracks

Genecats, April 6, 2011

Angie Hinrichs <angie@soe.ucsc.edu>

Overview

- Genetic variation in 1 slide
- dbSNP: background
- dbSNP: more & better
- VCF (1000 Genomes, …)
- GVF (dbVar/ISCA)

Genetic variation

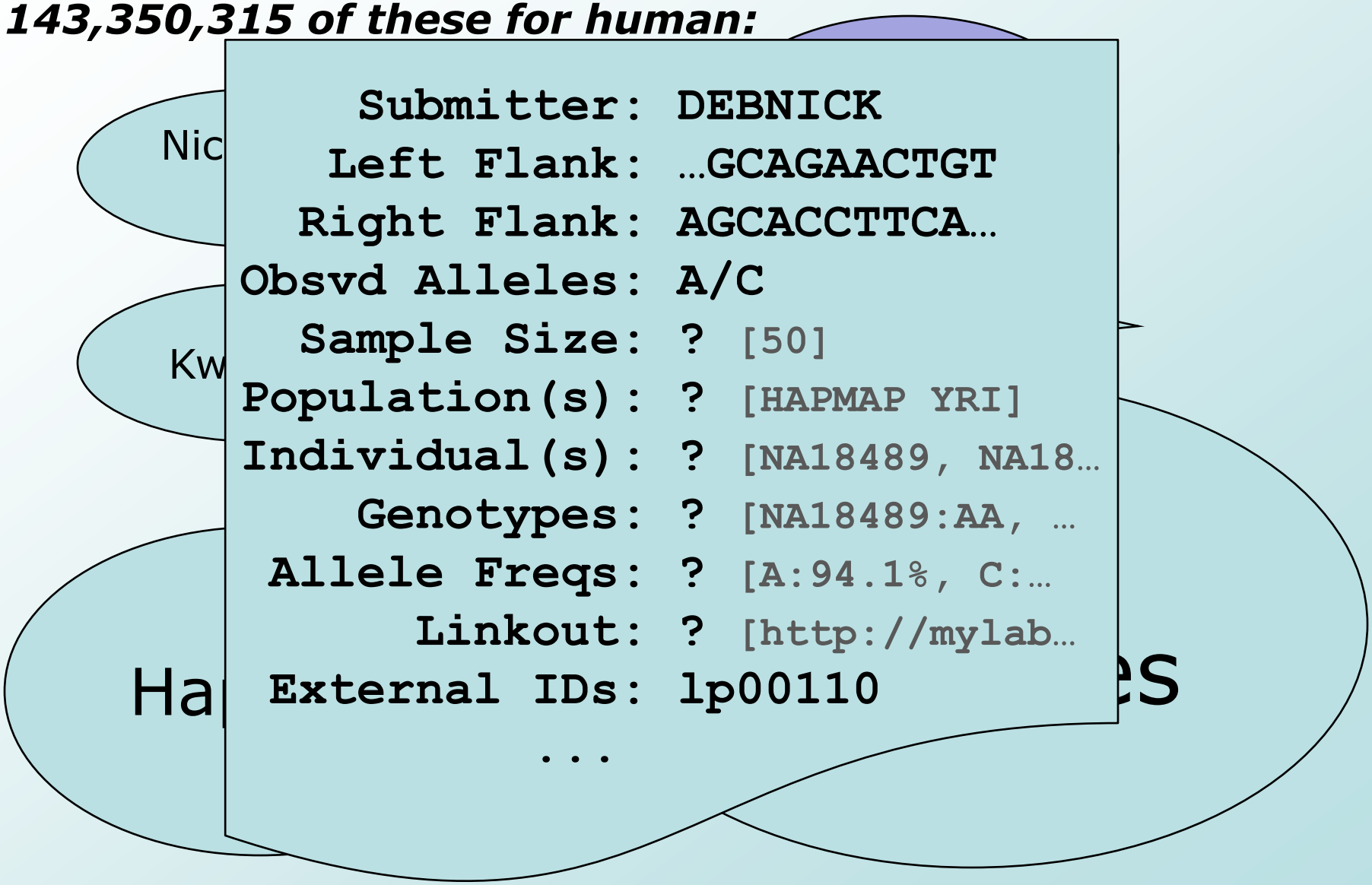
- Differences in genomic sequence
  - substitutions, insertions, deletions, copy number variation, …
  - Allele: observed version of variant
- We each have 2 sets of chrom's
  - Haplotype: one of the 2 sets
  - Genotype: both sets
  - Phased genotype: from Mom, from Dad
- Polymorphism: variant fixed in a population (not just any old variant)

dbSNP

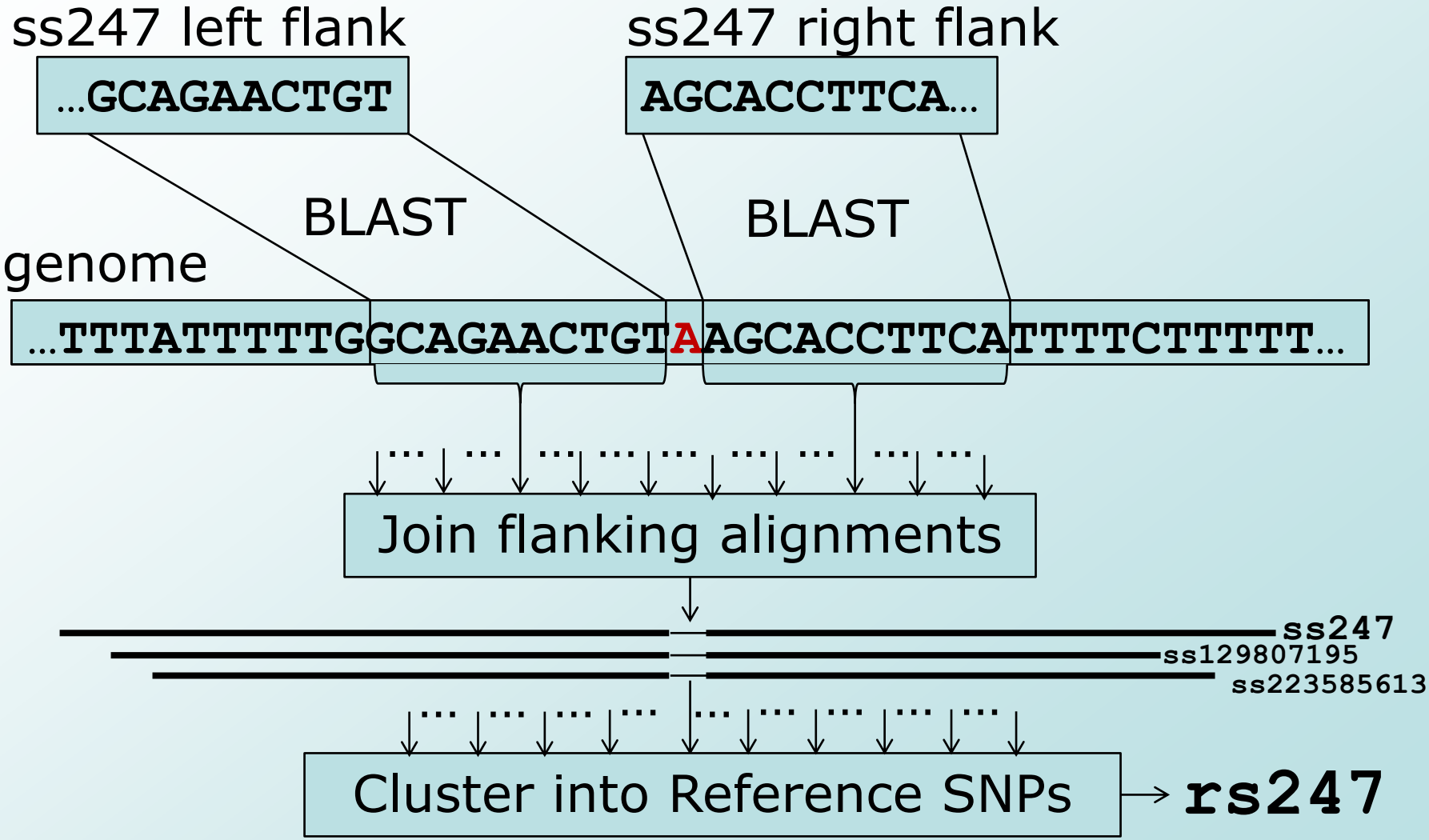
- Warehouse for discovered variants
- First submissions: late 1990s
- Human build 132: 143M submissions
- Provides stable IDs for variants
- “SNP” somewhat of a misnomer
- Garbage in, garbage out

dbSNP's monumental task

143,350,315 of these for human:



dbSNP's mapping process



UCSC's little slice of dbSNP

- Only Reference SNP clusters (rs#), not Submitted SNPs (ss#)
- Only SNPs mapped to reference genome (not Celera, Venter etc)
- BED6 + ten attributes out of dozens

UCSC's extra sauce

- 18 anomalous conditions: "Exceptions"
- Noted on details page
- Used internally to filter SNPs e.g. before cross-species liftOver

Mary-Claire King's kick in the A

- dbSNP's "clinically associated" SNPs are indistinguishable in our track
- Need subset for filtering out boring variation

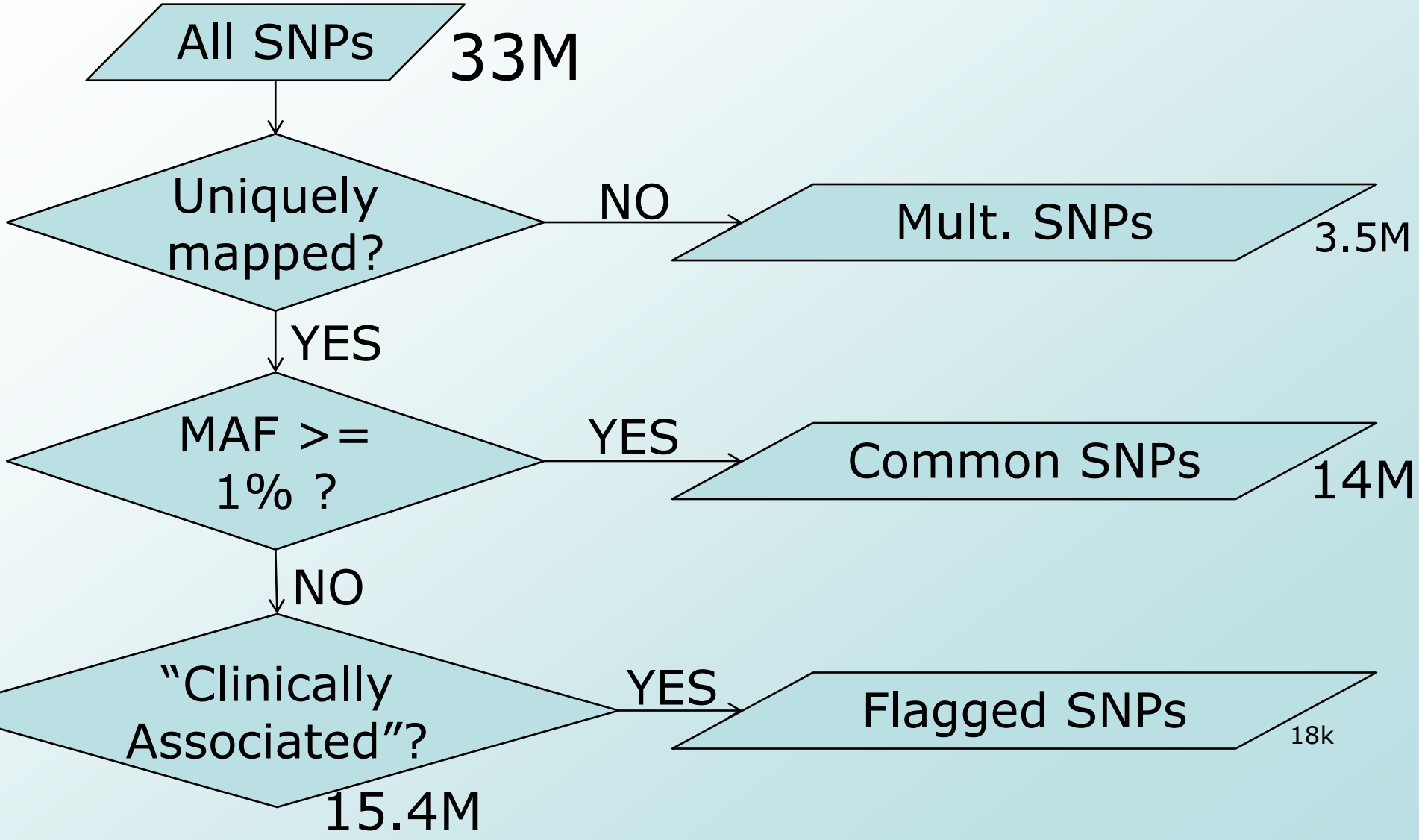
Souping up our SNPs track

- More info for discerning users
- Categorization -> subset tracks
- Genome Browser: colors, filters, details

New data columns

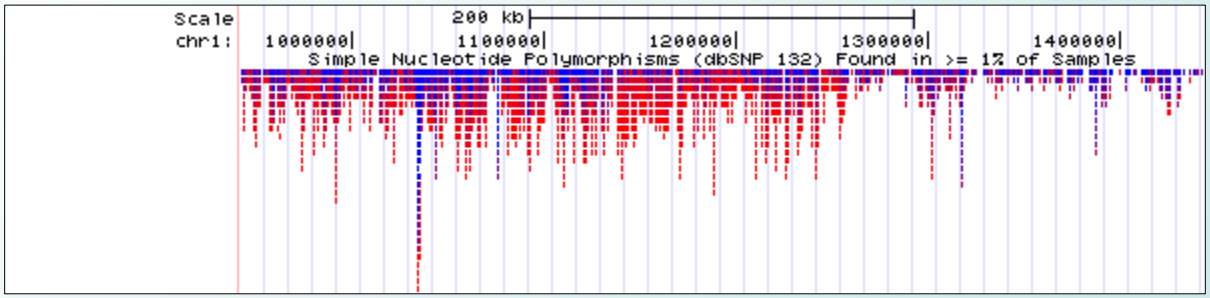
- Submitter handles
- Allele Frequencies
- dbSNP flags: “clinically associated” and a few others
- Separate table of exceptions pulled into main table as mysql set column

SNP subsets:
Common, Flagged, Multi-Mapped

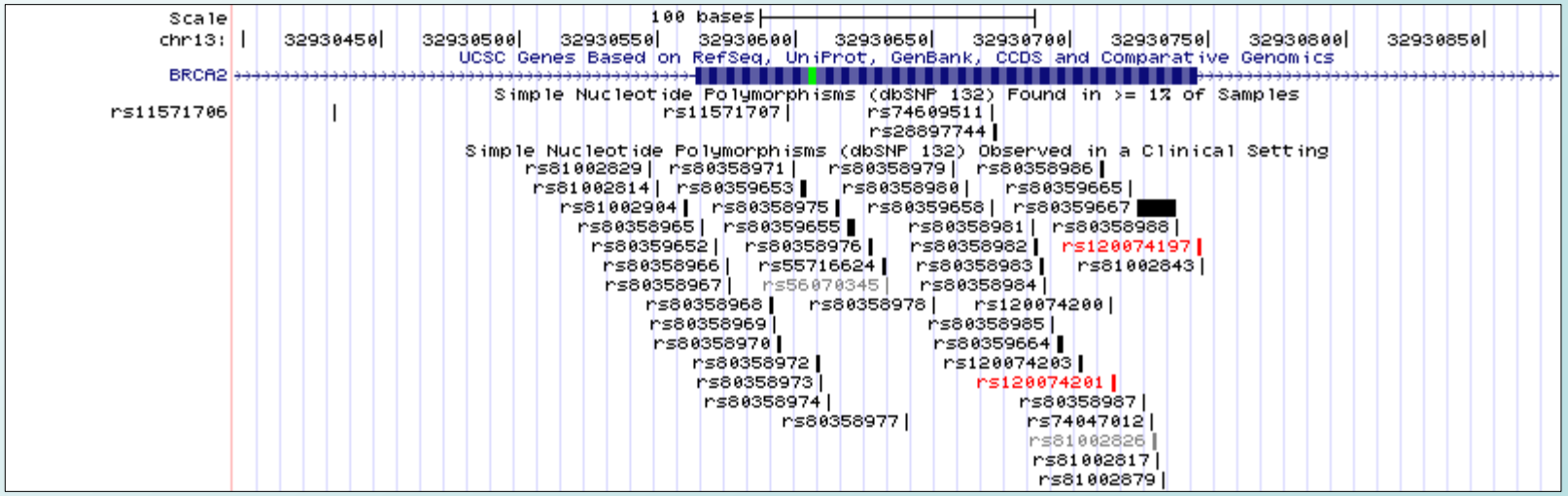


Genome Browser SNPs track:
new coloring and filtering options

Color by allele frequency (red=rare, blue=common):



Color by exception (anomaly), if any:



Genome Browser SNPs track: fun with new details

Single-submitter SNPs often have some issues:

Class	single
Validation	by-cluster,by-frequency
Function	unknown
Molecule Type	genomic
Average Heterozygosity	0.500 +/- 0.000
Weight	1
Submitter Handles	HUMANGENOME_JCVI
Allele Frequencies	G: 50.000% (2 / 4); A: 50.000% (2 / 4)

?

old

new

Miscellaneous properties annotated by dbSNP:

?

Minor Allele Frequency is at least 5% in all populations assayed

Genome Browser SNPs track: fun with new details

4 submitters, 1060 chromosomes:

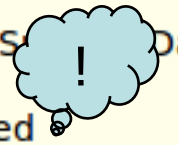
Class	single
Validation	by-cluster
Function	missense
Molecule Type	genomic
Weight	1
Submitter Handles	BIC_BRODY, ICRCG, ILLUMINA, PERLEGEN
Allele Frequencies	T: 1.038% (11 / 1060); C: 98.962% (1049 / 1060)

Miscellaneous properties annotated by dbSNP:

SNP is in OMIM/OMIA and/or at least one submitter is a Locus-S Database ("clinically associated")

SNP was submitted by Locus-Specific Database

Minor Allele Frequency is at least 5% in all populations assayed



New data sources and formats

- tabix: make any old text format fast like bigBed!
- 1000 Genomes: VCF
- dbVar / ISCA: GVF

tabix: compress & index any old line-based text format

- bgzip (block-based gzip) compression, like BAM
- Tell tabix which columns have sequence name and start offset
- Voila! BAM-like binary index file

Variant Call Format (VCF)

- Developed for 1000 Genomes project
- Goals:
  - Describe diverse types of variation
  - Support user-defined structured data
  - Not bloated
- Data definitions in header

VCF:
define data fields in header

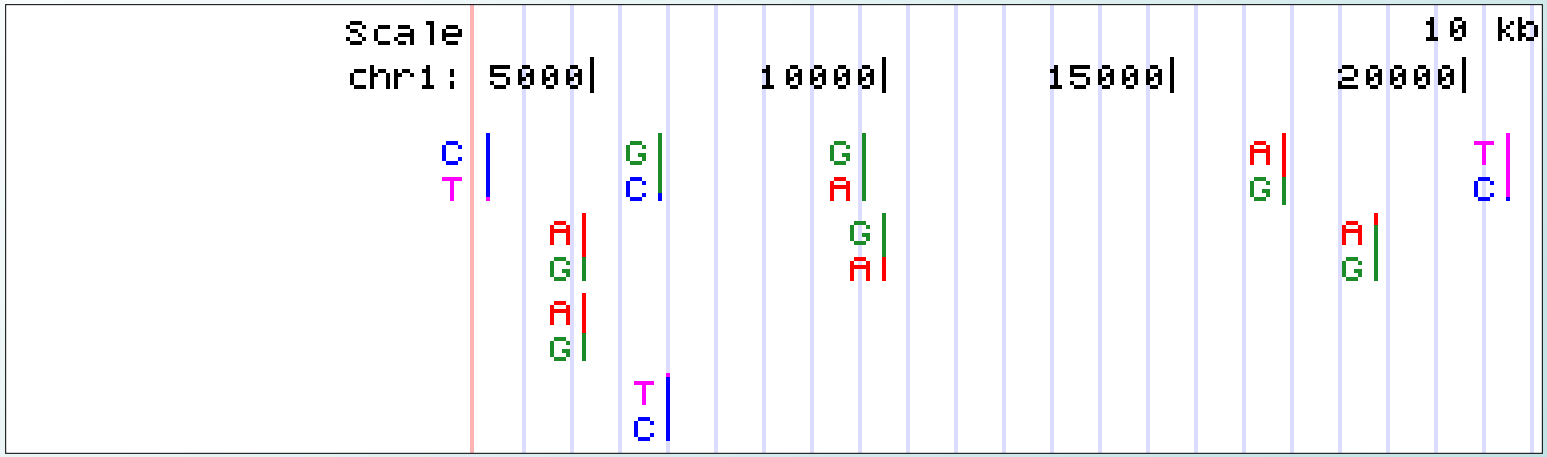
```
##fileformat=VCFv4.0
##INFO=<ID=DP,Number=1,Type=Integer,
Description="Total Depth">

##INFO=<ID=HM2,Number=0,Type=Flag,
Description="HapMap2 membership">

##INFO=<ID=AA,Number=1,Type=String,
Description="Ancestral Allele,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/ancestral_alignments/README">
```

VCF+tabix = custom track

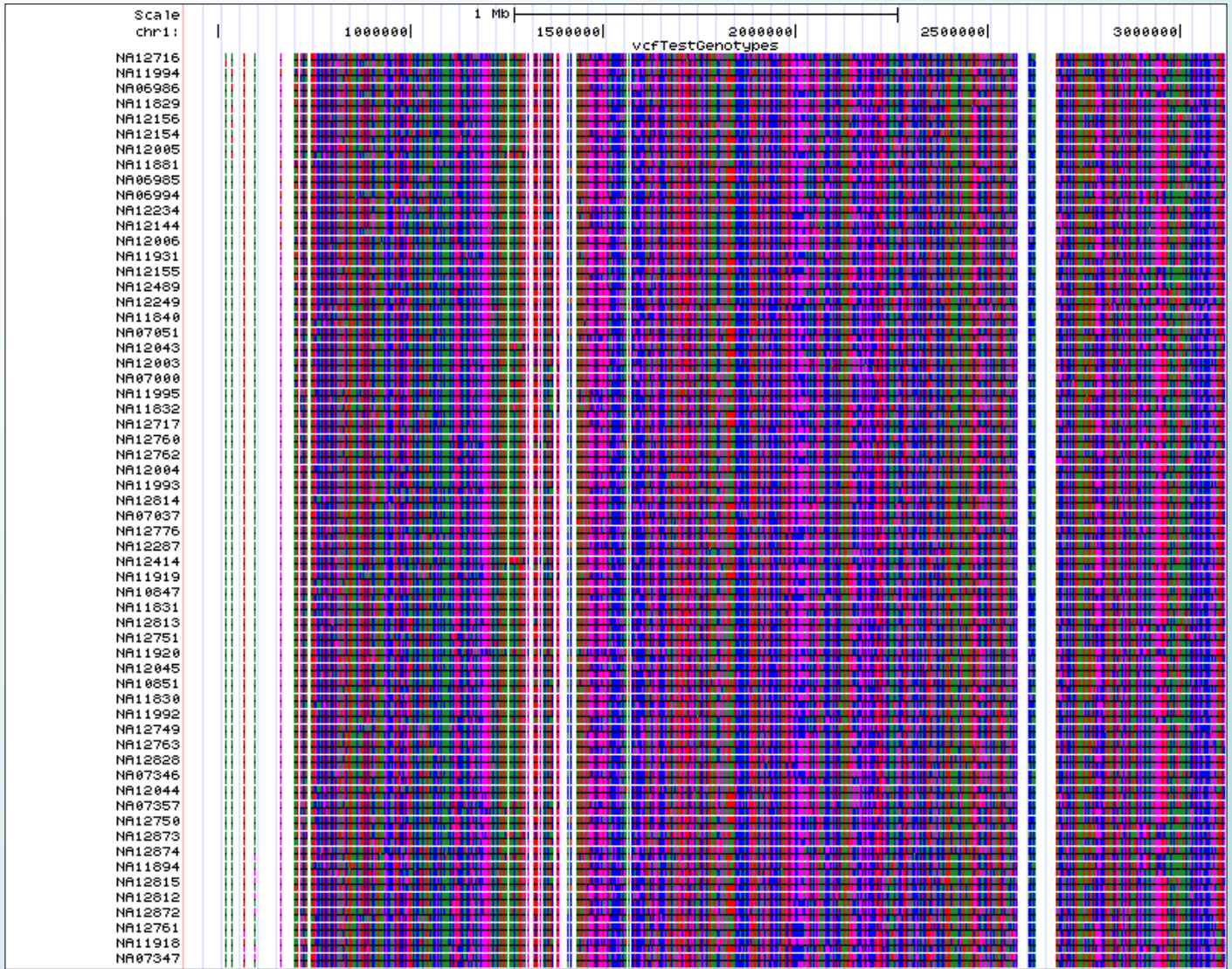
```
track
  name=CEU_lowCov_pilot_genotypes
  type=vcfTabix
  bigDataUrl=ftp://ftp-
trace.ncbi.nih.gov/1000genomes/ftp/pilo
t_data/release/2010_07/low_coverage/snp
s/CEU.low_coverage.2010_07.genotypes.vc
f.gz
  visibility=pack
  db=hg18
```

VCF example: SNPs with allele counts

Convert to pgSnp items for display:



VCF example: SNPs with genotypes

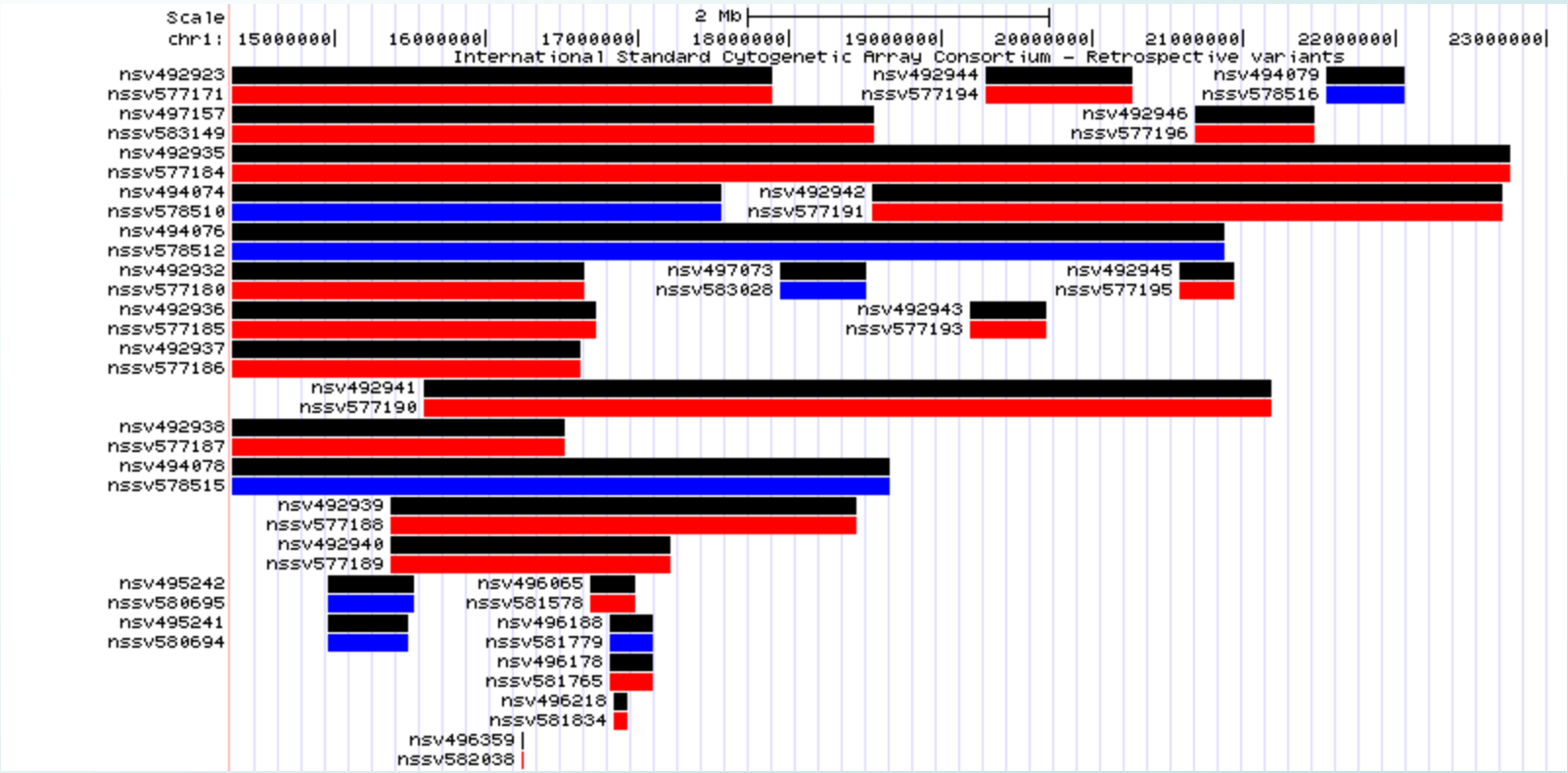


Genome Variation Format (GVF)

- Based on GFF3; additional attributes

```
chrY    dbVar   Gain    2714852
57440868        .       .       .
ID=nssv579294;Name=nssv579294(Gain);
Parent=nsv494707;var_type=Gain;
Start_range=.,2714852;End_range=57440
868,.;clinical_int=Uncertain
Significance: likely
pathogenic;samples=ISCA_ret_id_2795;
phenotype=Developmental Disabilities
```

ISCA Retrospective (GVF from dbVar)



ISCA Retrospective - details

Name: nssv577191(Loss)
Parent: nsv492942
Variant type: Loss
Start range: outer start unknown, inner start 18546902
End range: inner end 22711974, outer end unknown
Clinical interpretation: Uncertain Significance: likely pathogenic
Samples: ISCA_ret_id_1257
Phenotype: Developmental Disabilities

Thanks!