The 1000 Genomes Browser

Graphical visualization of genotype data from the 1000 Genomes Project

https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Scope

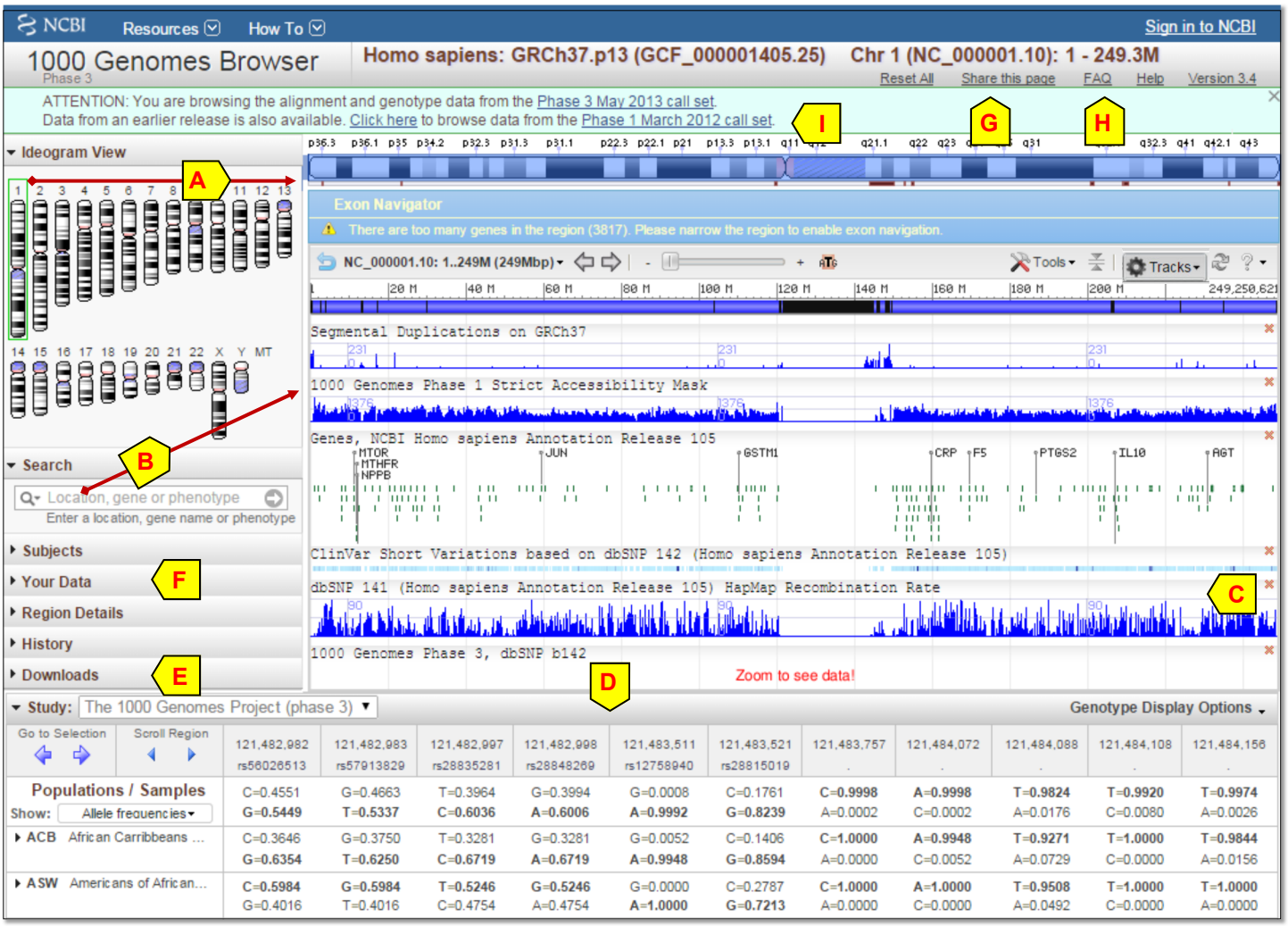
The 1000 Genomes Project aims to produce an extensive public catalog of human genetic variation, including SNPs, structural variants, and their haplotype contexts [1]. The genomes of more than 2500 de-identified people from 25 populations around the world are being sequenced using next-generation sequencing technologies with results of the study freely accessible to researchers worldwide. The data files from this project are mirrored at NCBI through FTP, Aspera, and Amazon Cloud:

ftp.ncbi.nlm.nih.gov/1000genomes/
www.ncbi.nlm.nih.gov/public/1000genomes/
s3.amazonaws.com/1000genomes

The **1000 Genomes Browser,** developed from NCBI's Graphical Sequence Viewer (SV) [2], addresses the challenge posed by the influx of this large set of data. This tool provides a convenient display for users to visualize the 1000 Genomes Project data and analyses by mapping the variation data to the reference genome assembly and allowing the retrieval and interactive examination of genotype data for a specific gene or genomic region. Through comparison with existing genome annotations, users can correlate the genotype and population prevalence with gene function.

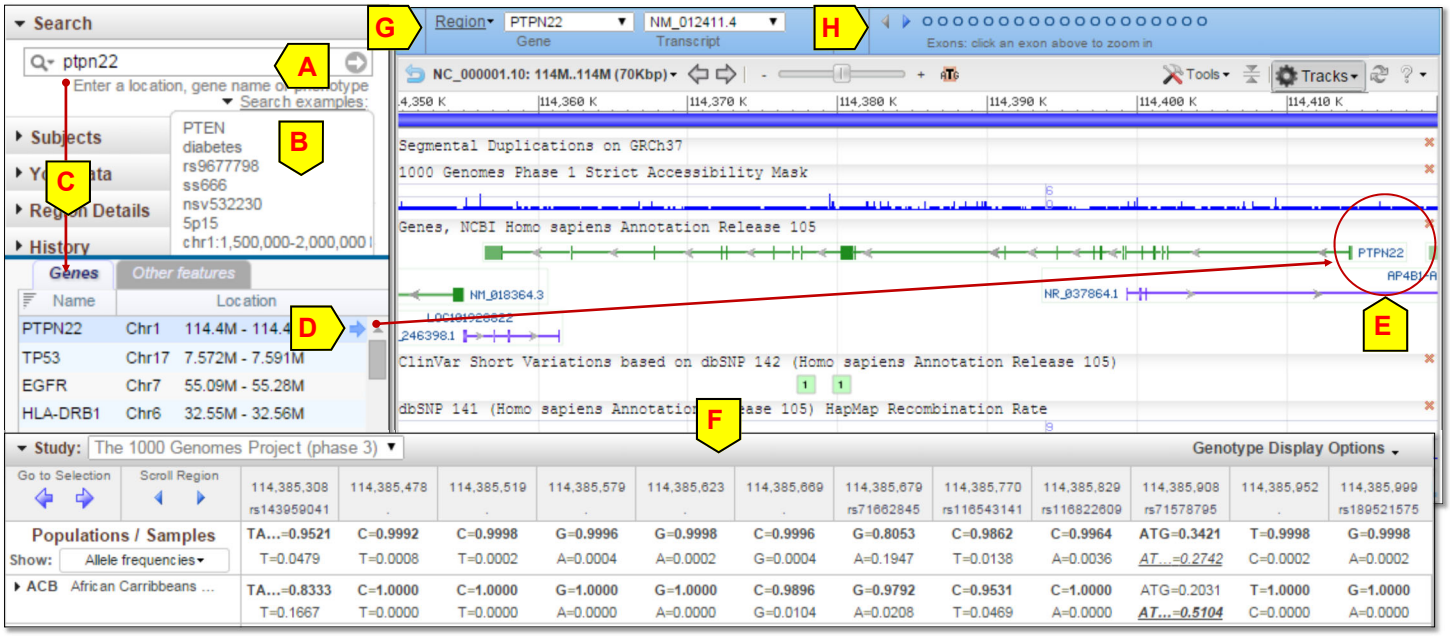
Interface for data access and display

The 1000 Genomes Browser (www.ncbi.nlm.nih.gov/variation/tools/1000genomes/) enables the selection and display of a selected chromosome (**A**). Keyword searching (**B**) zooms the display to the specific location, showing annotation and variation tracks (**C**), and listing detailed variations with underlying genotype in the table below (**D**). The "Downloads" widget (**E**) exports the genotype and alignment data for the selected genomic region. With the "Your data" widget (**F**), users can upload custom data for private display in the graphical panel. The "Share this page" (**G**) creates a unique URL for linking to the customized display. Links to FAQ and Help (**H**) provide additional details about the function of this tool. The browser displays data from the Phase 3 call set. A link at the top (**I**) provides access to the Phase 1 data.



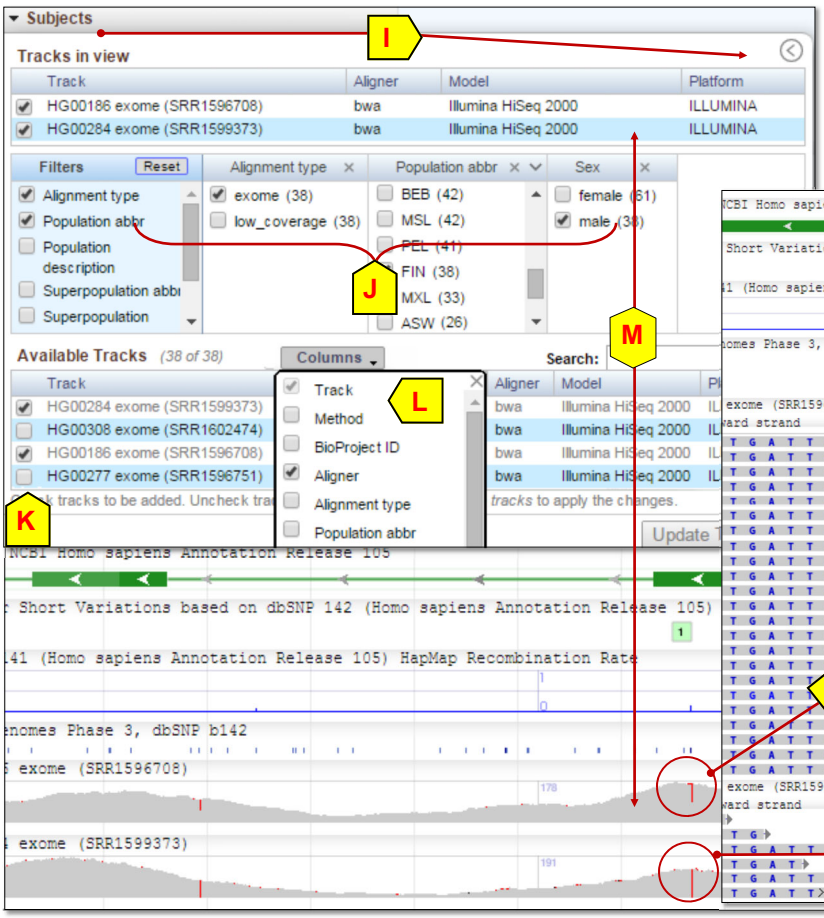
Searching with custom terms

The search box at the left (**A**) provides a way to quickly navigate to the targeted genomic region of interest. Clicking the "Search examples" (**B**, activated upon hovering) expands the section to display example query formats. Entering gene symbol "ptpn22" and pressing the return key retrieves a list of matches (**C**). Hover over an entry to activate the navigation arrow (**D**), which zooms the viewer to the target region (**E**) and simultaneously updates the genotype table (**F**). The gene region displayed is shown at the top (**G**). The exon navigator (**H**) allows zooming in to specific exons through a single radio button click, or flipping through exons using the arrows.

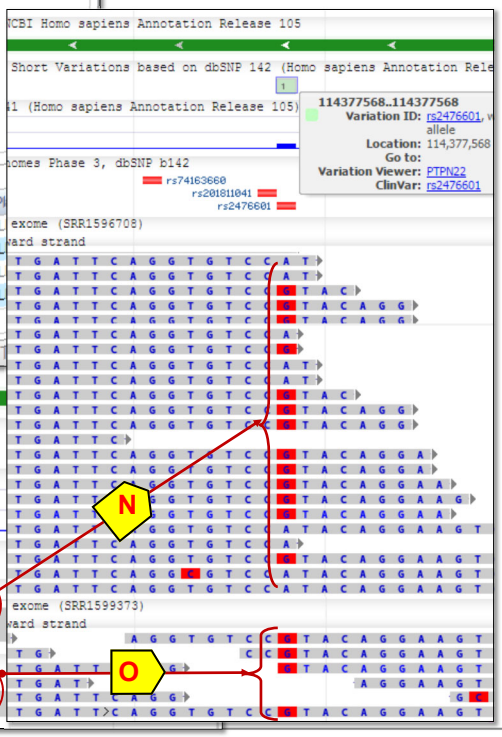


Examining detailed genotypes for specific subjects

Clicking the "Subjects" arrow and the "Tracks in view" arrow (**I**) expands the section to provide access to read alignment data supporting the genotype calls of specific subjects (individuals). This section can be resized by dragging on its sides.

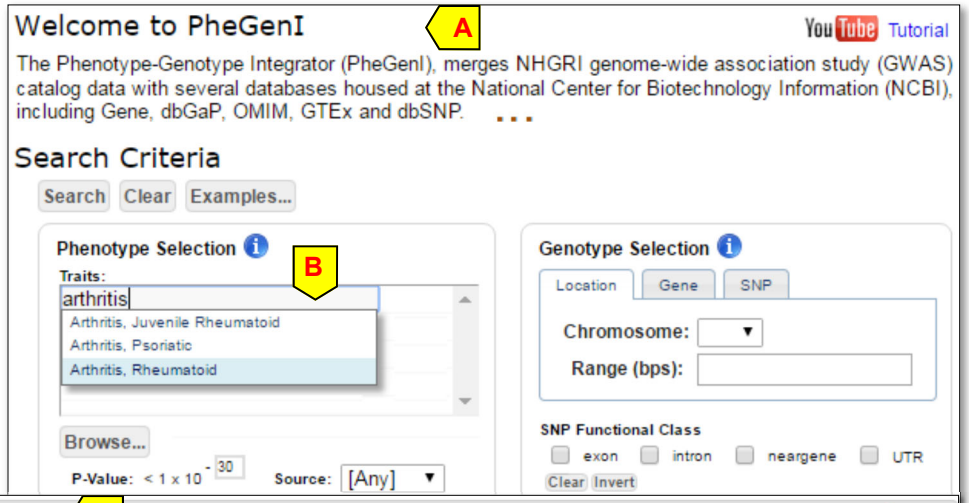


Alignment data can be filtered using available criteria (**J**), free-text searches, or their combination. Filtered search results are shown in the "Available Tracks" table and selectable by clicking the checkboxes (**K**). Additional track details can be added using the "Column" menu (**L**). Clicking the "Update Tracks" button (**M**) adds the selected alignment tracks to the display and updates the "Tracks in view" table at the top. Zooming to the sequence level in the display provides alignment details with the variation calls clearly marked (**N**, **O**).



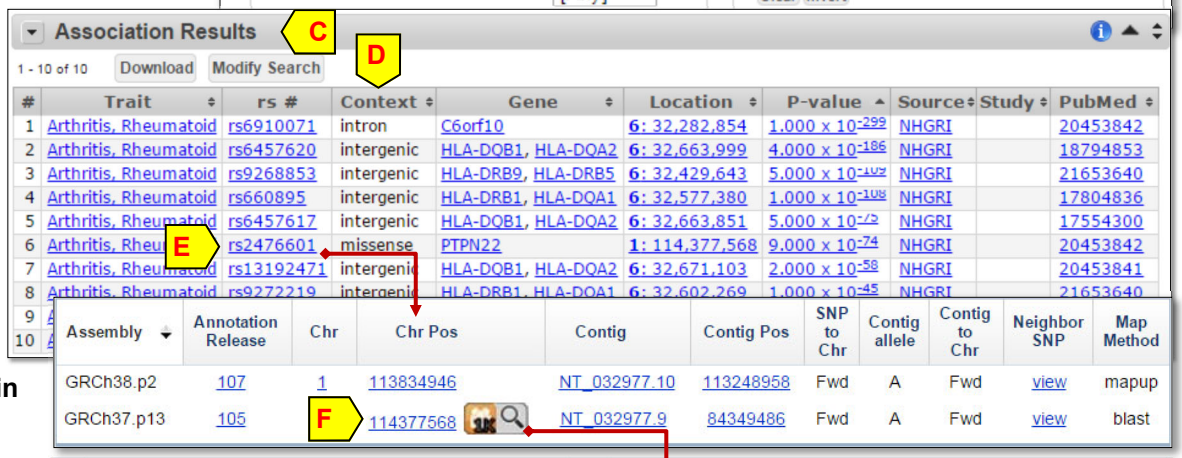
An example usage case

Data from many publicly funded genome-wide association studies are available in the dbGaP database at NCBI. The Phenotype and Genotype-Integrator tool (PheGenI, A) [3] provides a convenient way to get the genes and SNPs associated with studied phenotypes or disorders. The following example begins using PheGenI to identify candidate genes and SNPs associated with rheumatoid arthritis (relevant PubMed abstracts: http://1.usa.gov/K7h6Vi). The retrieved variations are used to examine the genotypes in different populations in data generated by the 1000 Genomes Project.



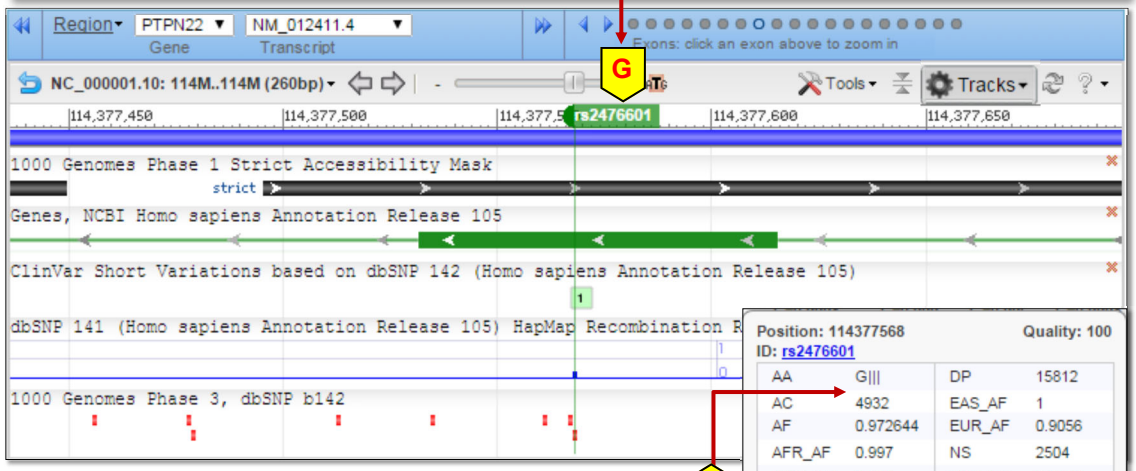
Searching in PheGenI

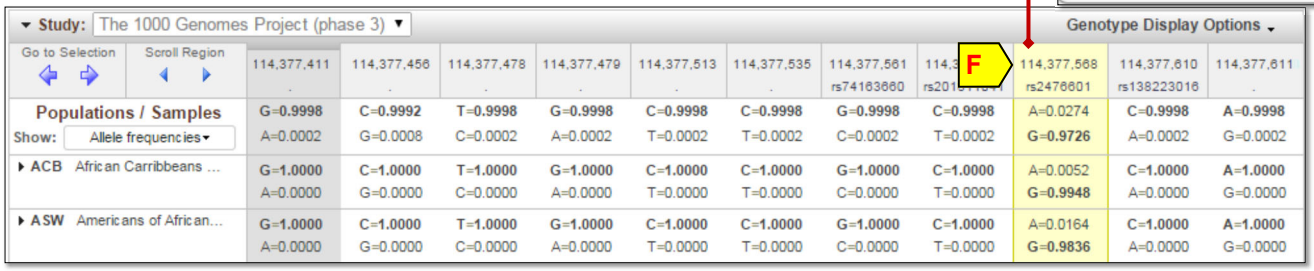
Selecting the "Arthritis, Rheumatoid" trait with a p-value cutoff at 10^{-30} (B) retrieves a list of genes and SNPs strongly associated with the disorder. The "Association Results" section (C) lists these SNPs (rs#) and genes in descending order of significance. The nature of listed SNPs is available in the "Context" column (D). Among this group, rs2476601 from PTPN22 is the only missense SNP. The rsID (E) is hyperlinked to the SNP record, where a direct link to 1000 Genomes Browser (F) allows the interactive examination.



Examining the SNP in the 1000 Genomes Browser

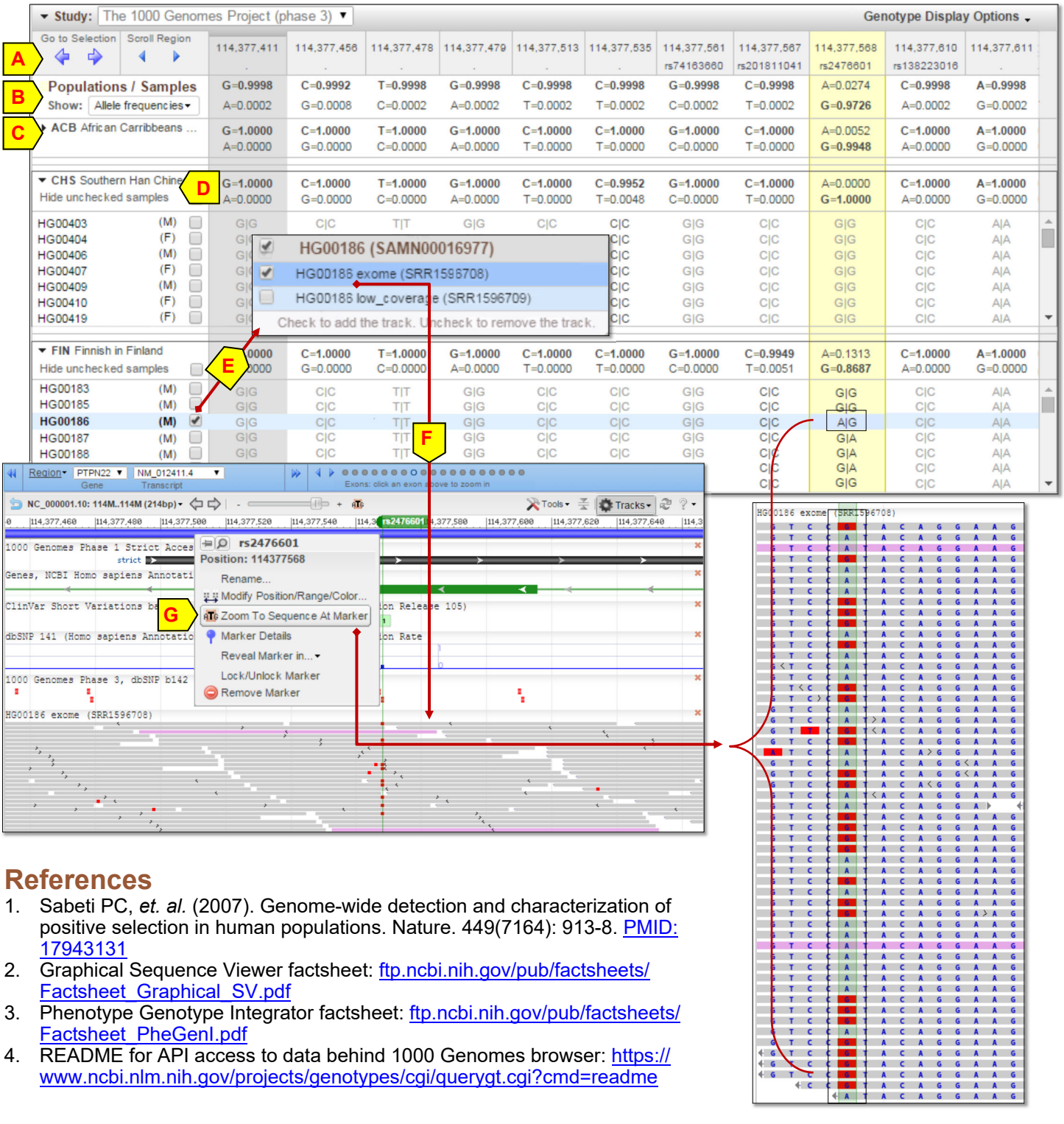
Clicking the magnifying glass icon (F) in the SNP record displays a zoomed-in view in the 1000 Genomes Browser, which highlights the target SNP with an added marker (G). The genotype table also highlights this SNP by coloring the column in yellow (H). SNPs outside the graphic display have their columns colored grey (I). Genotyping statistics for a SNP are in the popup (I) activated upon hovering over the column header.





An example usage case (cont.)

The "Genotype" table (shown below) provides the detailed genotype data. Its top row (**A**) displays the variation calls and their genomic coordinates. The "Populations/Samples" row (**B**) lists the global allele frequencies (click the arrow to adjust to counts), while rows below (**C**) list the allele frequency for individual populations. Clicking a population cell (**D**) toggles open the individual genotypes section for that population. Clicking the checkbox (**E**) next to an individual triggers the read selector popup from which alignments can be selected for display in the graphical view, allowing for a more detailed examination (**F**). Upon "Zoom to Sequence at Marker" menu selection (**G**), aligned reads from the selected individual provide supporting evidence for genotype calls at this position. The prevalence of rheumatoid arthritis reported for different populations from published studies is generally consistent with data from the 1000 Genomes study. This further supports the conclusion of the reported GWAS studies. Two such published reports are available from PubMed (http://1.usa.gov/K7h6Vi).



References

1. Sabeti PC, *et. al.* (2007). Genome-wide detection and characterization of positive selection in human populations. Nature. 449(7164): 913-8. PMID: 17943131
2. Graphical Sequence Viewer factsheet: ftp.ncbi.nih.gov/pub/factsheets/Factsheet_Graphical_SV.pdf
3. Phenotype Genotype Integrator factsheet: ftp.ncbi.nih.gov/pub/factsheets/Factsheet_PheGenI.pdf
4. README for API access to data behind 1000 Genomes browser: https://www.ncbi.nlm.nih.gov/projects/genotypes/cgi/querygt.cgi?cmd=readme