The NHGRI Sample Repository for Human Genetic Research: biospecimens and a new genomic data search tool

A. Liautaud, D. Kusic, G. Calendo, J. Madzo, S. Sampson, J. Schulz, L. Willis, E. Horton, M. Mitchell, N. Gharani, L. Scheinfeldt.

Coriell Institute for Medical Research, 403 Haddon Avenue, Camden, NJ 08103

Background

The NHGRI Sample Repository for Human Genetic Research (NHGRI Repository) housed at the Coriell Institute for Medical Research, facilitates studies of human genetic and genomic variation by establishing, characterizing and distributing a large (over 3,700) and diverse publicly available collection of renewable biospecimens obtained from human populations living around the world, including biospecimens associated with the HapMap and 1000 Genomes Projects as well as the Human Pangenome Reference Consortium. Participants that have generously donated to the NHGRI Repository consented to their biospecimens and associated data being used for a wide range of general research and to broad data sharing of largescale genomic data. Through the 1000 Genomes Project, the majority of this collection has been characterized with publicly available whole genome sequencing data and other large-scale genomic data. We recently developed two user-friendly and integrated search tools that allows catalog users to dynamically query 1000 Genomes Project whole genome sequencing data (30x)[1]. A set of up to 100 variants at a time can be searched by individual variant (rsid) using the multi-SNP search (**Figure 1**) which returns an interactive table of each subject id, population descriptor, genotype for each SNP included in the search, and sex that can be filtered or exported to an excel file. This tool is intended to support researchers interested in identifying biospecimens with particular multi-variant profiles such as those used in pharmacogenetics (PGx) and polygenic risk assessments. For example, a researcher interested in identifying cells or DNA that carry one or two copies of the CYP2C19*2 allele can simultaneously search the three haplotype defining SNPs (rs12769205, rs4244285, rs3758581) according to PharmVar. Alternatively, researchers can search star allele annotations directly using Star Allele Search tool (**Figure 2**). In addition, researchers interested in identifying cells or DNA that carry high or low numbers of polygenic risk alleles can search up to 100 variants at a time; for example, 23 of the 34 variants that define the ROOT PRS model for breast cancer are included in the WGS dataset and can be searched together in a single batch. These tools add to the growing number of web-based genomic search tools affiliated with the NHGRI Repository[2,3], including a Gene search tool and a Gene Expression search tool. These tools are intended to enable researchers with a fast and simple way to identify biospecimens with the most appropriate variation profiles for their research goals. Our Genomic Data Search tools can be accessed at https://www.coriell.org/1/Browse/Genomic-Data-Search, and more information about the NHGRI Repository can be found at https://catalog.coriell.org/NHGRI.

[1] Byrska-Bishop, Cell 185: 3426–3440 (2022).

[2] Calendo et al., bioRxiv, https://doi.org/10.1101/2023.07.24.550372 (2023).

[3] Gharani et al., bioRxiv, https://doi.org/10.1101/2023.07.27.550362 (2023).

Table 1: NHGRI Repository Population Samples

Population	# Unique Subjects	# Unique Subjects with WGS[1]	Population	# Unique Subjects	# Unique Subjects with WGS[1]
African American People living in St. Louis, Missouri	9	–	Indian Telugu People living in the UK	118	107
African American People living in Southwest, USA	107	74	Japanese People living in Tokyo, Japan	131	104
African Caribbean People living in Barbados	120	116	Kinh People living in Ho Chi Minh City, Vietnam	124	122
Bengali People living in Bangladesh	144	131	Luhya People living in Webuye, Kenya	122	99
British people living in England and Scotland	100	91	Maasai People living in Kinyawa, Kenya	205	--
Chinese Dai People living in Xishuangbanna, China	102	93	Mende People living in Sierra Leone	128	99
Chinese People living in Metropolitan Denver, CO, USA	129	--	Mexican American People living in Los Angeles, CA, USA	104	97
Colombian People living in Medellín, Colombia	136	132	Peruvian People living in Lima, Peru	122	122
Esan People living in Nigeria	173	149	Puerto Rican People living in Puerto Rico	139	139
Finnish People living in Finland	103	99	Punjabi People living in Lahore, Pakistan	158	146
Gambian People living in Western Division, Mandinka	179	178	Sri Lankan Tamil People living in the UK	128	114
Gujarati People living in Houston, TX, USA	117	103	Toscani People living in Italia	117	107
Han Chinese People living in Beijing, China	162	103	Yoruba People living in Ibadan, Nigeria	229	178
Southern Han Chinese People living in China	163	163	CEPH Collection *	186	179
Iberian People living in Spain	157	157	* Samples available from the NIGMS Repository		

Multi-SNP Search

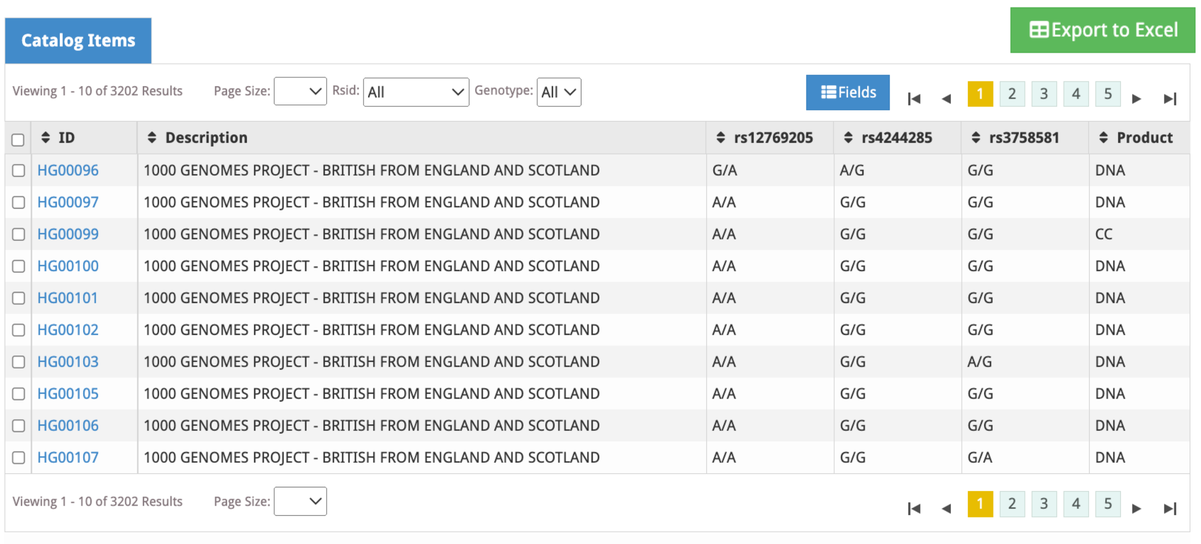

Figure 1. Search up to 100 rsids to return an interactive table of each sample, subject population affiliation, and genotype that can be filtered or exported to an Excel file.

Star Allele Search

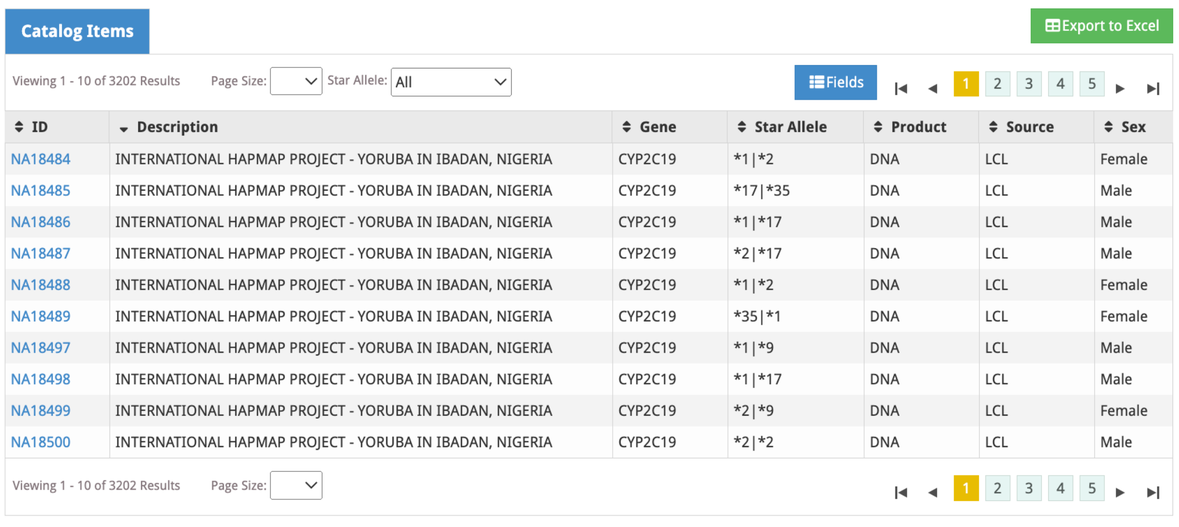

Figure 2. Search by pharmcogene to to return an interactive table of each sample, star allele diplotype, subject population affiliation, and genotype that can be filtered or exported to an Excel file.

The NHGRI Sample Repository for Human Genetic Research is supported by the National Human Genome Research Institute (U24HG008736).