ALFA: allele frequency data for variant analysis and interpretation

- Closed captioning: www.captionedtext.com & enter 4376574
- The **recording of this webinar** will be on our YouTube channel in the Webinars playlist in a few days: <youtube>/user/NCBINLM/playlists
- Presentation files are in the **Handouts Pod** on the webinar side-panel.
- Type in the **Questions Pod** on the side-panel to ask questions when you think of them. *Don't wait until the end.*
- Answers to all questions will be available after the webinar with a link on our Webinars page: <ncbi>/home/coursesandwebinars.shtml
- Slides, Supplemental Materials, Q&A: https://go.usa.gov/xvbCp

<ncbi> = www.ncbi.nlm.nih.gov
<ftp> = ftp.ncbi.nlm.nih.gov
<YouTube> = www.youtube.com

U.S. National Library of Medicine
National Center for Biotechnology Information

NCBI Webinars

1

NCBI's ALFA: allele frequency data for variant analysis and interpretation

Lon Phan – NCBI dbSNP resource lead

U.S. National Library of Medicine
National Center for Biotechnology Information

2

Outline

- ALFA Introduction
- Data Release Summary
- Data Access Tutorials
  - Web
  - API
  - FTP
- Conclusion and Q&A

U.S. National Library of Medicine
National Center for Biotechnology Information

3

3

Allele Frequency Aggregator (ALFA)

https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/

- The goal is to provide allele frequency from more than 1 million subjects dbGaP subjects
- Aggregating heterogenous data from dbGaP studies include chip array, exome, and genomic sequencing
- Variant, population, and frequency data normalized and provided in standard exchange formats.
- The data will be integrated with dbSNP regular build release with assigned RS accessions for variants and available for access by web, FTP, and API.

U.S. National Library of Medicine
National Center for Biotechnology Information

4

4

ALFA Process Overview

Preprocessing → **Computation** → **Product Release**

Preprocessing
- 'Approved' Study selection
- Study QA/QC
- GRAF-Pop ancestry calculation
- Data conversion to VCF

Computation
- Normalization of variants
  - SPDI[1] conversion
  - Remapping
  - VOCA[1]
  - RefSNP assignment
- Aggregation
  - Individuals and genotype
  - Populations
- Compute frequency

Product Release
- RefSNP Web page
- API
- FTP

U.S. National Library of Medicine
National Center for Biotechnology Information

[1]Holmes et al. https://pubmed.ncbi.nlm.nih.gov/31738401/

5

5

ALFA Release Summary (March 10,2020)

Inputs	
Studies	42
Subjects	98,494
Genotypes	551,345,630,054

Outputs	
Unique Variants	446,909,594
• Exist in dbSNP	442,539,572
• Novel	4,370,022

U.S. National Library of Medicine
National Center for Biotechnology Information

6

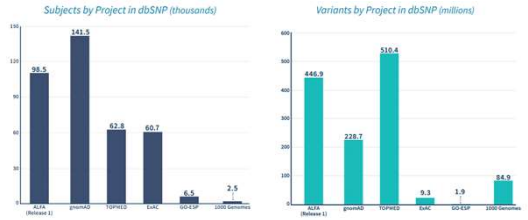
6

Slide 5

P[4 Improved precision and accurarcy of RS assignment; reducing duplicates.

Phan, Lon (NIH/NLM/NCBI) [E], 4/21/2020

dbSNP Allele Frequency

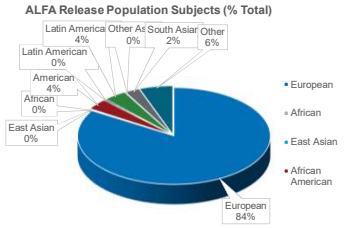


ALFA (release 1) has ~10% of dbGaP 1M potential input subjects with genotype analyzed by Graf-Pop

U.S. National Library of Medicine
National Center for Biotechnology Information

7

Population Summary



Population	Subjects
European	82,475
African[1]	114
East Asian	153
African American	3,555
Latin American 1	354
Latin American 2	3,801
Other Asian	84
South Asian	2,459
Other	5,499
ALFA Total	**98,494**
dbGaP Total	**2,491,605**
dbGaP Total (Unique)	**1,397,069**

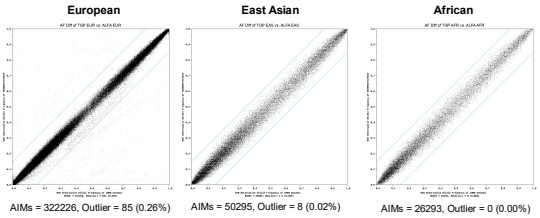
[1] Included both African and African-Americans

U.S. National Library of Medicine
National Center for Biotechnology Information

https://www.ncbi.nlm.nih.gov/snp/docs/gsr/data_inclusion/#population

8

Concordance between AFLA and 1000 Genomes Populations



AIMs = 322226, Outlier = 85 (0.26%)

AIMs = 50295, Outlier = 8 (0.02%)

AIMs = 26293, Outlier = 0 (0.00%)

- > 95% concordance with GnomAD and TopMed

U.S. National Library of Medicine
National Center for Biotechnology Information

9

9

P[5

ALFA Variants of Interest

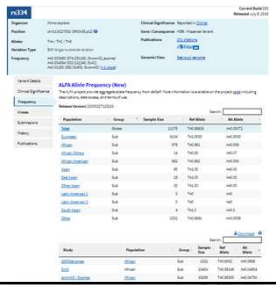
Variant Source	ALFA RS Count	Percent(%) of Total	Total RS
ClinVar	229963	59	387427
GTR	366	77	474
GWAS Catalog (p-value < 10^-5)	21494	98	21972
Pubmed Cited	188809	82	231502

U.S. National Library of Medicine
National Center for Biotechnology Information

10

10

ALFA frequency reported on RefSNP Page



U.S. National Library of Medicine
National Center for Biotechnology Information

11

11

Search dbSNP Web

https://www.ncbi.nlm.nih.gov/snp/docs/entrez_help/

- RS number
- Position (GRCh37 and GRCh38)
- Gene
  - Symbol
  - ID
- And other attributes
  - Frequency
  - ClinVar clinical significance
  - Many others...

U.S. National Library of Medicine
National Center for Biotechnology Information

12

12

Slide 10

P[5 Variants of high interest

Phan, Lon (NIH/NLM/NCBI) [E], 4/21/2020

Sample Variant Search

https://www.ncbi.nlm.nih.gov/snp/docs/entrez_help/

- ClinVar Variant
- Conflicting interpretations of pathogenicity
- Cited in PubMed
- Missense variant
- Cluster Validation (observed by multiple submitters)
- NO frequency from other projects (1000G, GnomAD, etc.)

U.S. National Library of Medicine
National Center for Biotechnology Information

13

13

Generate Query Term

(((conflicting interpretations of pathogenicity[Clinical_Significance] AND snp_pubmed_cited[Filter] AND missense variant[Function_Class] AND by cluster[Validation]))) NOT "by frequency"[Validation Status]

- Using Filters
- Using Advance Search

U.S. National Library of Medicine
National Center for Biotechnology Information

14

14

Search dbSNP using eUtils

https://www.ncbi.nlm.nih.gov/snp/docs/entrez_help/

Use same web query syntax

```
eShandle = Entrez.esearch(db="snp",  # search dbSNP
                          #complex query for missense and pathogenic variants in LPL gene with global MAF betweeen 0 and 0.01.
                          term='(((conflicting interpretations of pathogenicity[Clinical_Significance] AND snp_pubmed_cited[Filt
er] AND missense variant[Function_Class] AND by cluster[Validation]))) NOT "by frequency"[Validation Status]',
                          usehistory="y", #cache result on server for download in batches
                          retmax=20 # return 20 RSID max
                         )
```

https://github.com/ncbi/dbsnp/blob/master/tutorials/eUtils.ipynb

U.S. National Library of Medicine
National Center for Biotechnology Information

15

15

NCBI Variation Service API

www.ncbi.nlm.nih.gov/variation/v0/

GET	/metadata/frequency	Retrieve frequency metadata
GET	/interval/{seq_id}:{position}:{length}/overlapping_frequency_records	Retrieve frequencies for variants overlapping a position range
GET	/refsnp/{rsid}/frequency	Lookup frequency data for specified RSID

U.S. National Library of Medicine
National Center for Biotechnology Information

16

16

NCBI Variation Service API Tutorials

https://github.com/ncbi/dbsnp/tree/master/tutorials/Variation%20Services/Jupyter_Notebook

- Launch `by_gene.ipynb` notebook on Binder: launch binder

Retrieve frequency data by gene using eUtils and Variation Service

- Launch `by_rsid.ipynb` notebook on Binder: launch binder

Retrieve frequency data by rsid in JSON format

- Launch `frequencies_for_vcf.ipynb` notebook on Binder: launch binder

Compare frequencies for two populations for variations from a VCF file

- Launch `metadata_as_hash.ipynb` notebook on Binder: launch binder

Retrieve and transform ALFA project and population meta data

- Launch `plot.ipynb` notebook on Binder: launch binder

Plot population minor allele frequencies (MAFs) for variants in a gene location

- Launch `querying_subsets_ftp.ipynb` notebook on Google Colab

Retrieving subsets of VCF data from remote NCBI FTP site

U.S. National Library of Medicine
National Center for Biotechnology Information

17

17

Conclusions

- New source of frequency data from dbGaP growing catalog of studies
- Quarterly updates with new dbSNP and dbGaP data
  - New unrestricted studies
  - Novel variants
  - Updated population frequency
- Normalized data with improved quality and precision reported in standard formats
- A beacon to data of interest in dbGaP

U.S. National Library of Medicine
National Center for Biotechnology Information

18

18

- URLs & Places to Learn More -

- NCBI Insights Blog: ncbiinsights.ncbi.nlm.nih.gov
- Learn: www.ncbi.nlm.nih.gov/home/learn.shtml
- Factsheets: ftp.ncbi.nih.gov/pub/factsheets/
- NCBI YouTube Channel: www.youtube.com/ncbinlm
- NCBI Helpdesk: info@ncbi.nlm.nih.gov

Questions?

Please take the evaluation following the webinar!

U.S. National Library of Medicine
National Center for Biotechnology Information

NCBI Webinars

19