Open access to dbGaP new aggregated allele frequency for variant interpretation

Lon Phan, Ph.D.

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

dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. dbGaP provides two levels of access - open and controlled.

Content	Count
Studies	1,274
Phenotypes	7,481
Variables	328,482
Subjects	2,178,532
Samples	2,475,763

2007
Genomic Summary Results (GSR) does not include individual level information was open-access

2008
Homer et al. publication

2008-2018
GSR in controlled-access

Updated Genomic Data Sharing Policy November 2018

Author:
Eric Green, M.D., Ph.D.

Share:

NEWS | NOVEMBER 01, 2018

Today, we are updating the way we manage data related to the NIH Genomic Data Sharing Policy to again allow unrestricted access to genomic summary results for most of the studies we support in order to advance health or further research purposes. These summary results are importantly distinct

Today, w
Data Sh
results f
research
genomic
information. Instead, these results come from analyzing pooled genomic data from multiple individuals together to generate a statistical result for the entire dataset. Such information can be a powerful tool for helping researchers determine which genomic variants potentially contribute to a disease or disorder.

https://go.usa.gov/xVM2d

Allele Frequency Aggregator (ALFA)

- A new source of aggregated frequency data from dbGaP
- Allow users access to pooled allele frequency data from dbSNP and dbGaP for variant interpretation and analysis

Data CoLab:
Friday, October 18
10:00AM - 10:30AM
Exhibit Hall, CoLab Theater 1 #345

Outline

- Pipeline overview
- Data Normalization
- Pipeline output summary
- Data Release
- Conclusion and Q&A

ALFA Process Overview

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

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

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

- RefSNP Web page
- API
- FTP

[1]Holmes et al. https://www.biorxiv.org/content/10.1101/537449v3

Project Challenges

Large volume of data

- Thousands of studies
- Millions of subjects and variants
- Trillions of genotype

Harmonized subjects and populations across studies

- No ancestry reported
- Study-reported ancestry is inconsistent with genotype
- Population definition and grouping not consistent

Normalize variant and genotype

- Different reporting (format, assembly, etc.)
- Mixed quality

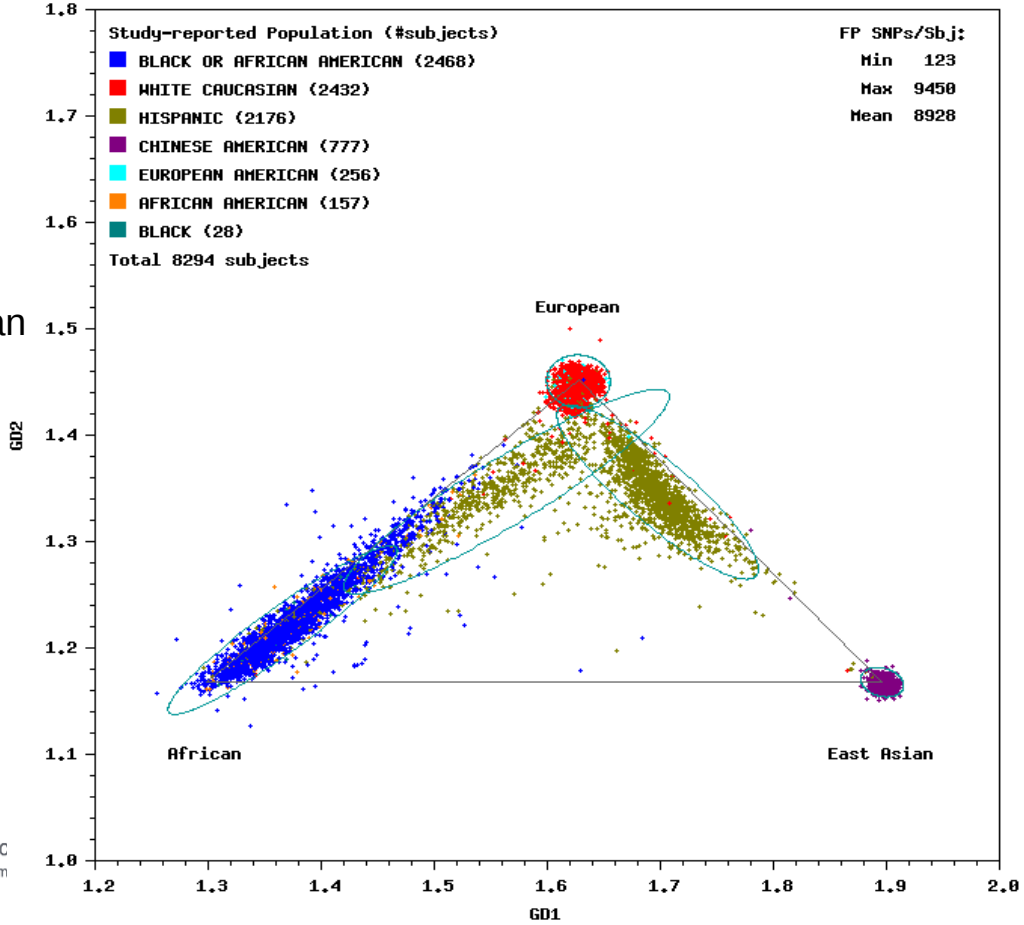
GRAF-pop

GRAF-pop: A Fast Distance-Based Method To Infer Subject Ancestry from Multiple Genotype Datasets Without Principal Components Analysis
Jin Y et al. (2019), https://www.ncbi.nlm.nih.gov/pubmed/31151998

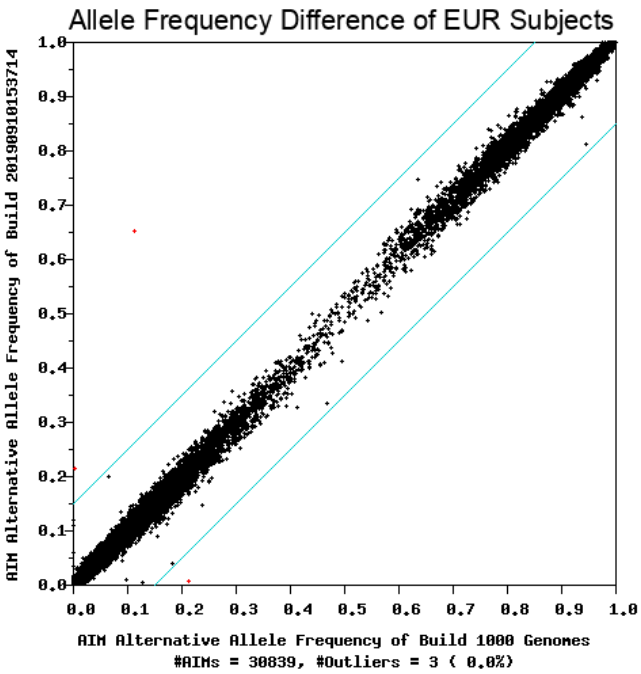
- Uses Ancestry Informative Markers (AIMS)
- Validate sample ancestry
- Remove duplicate samples
- Assign computed populations

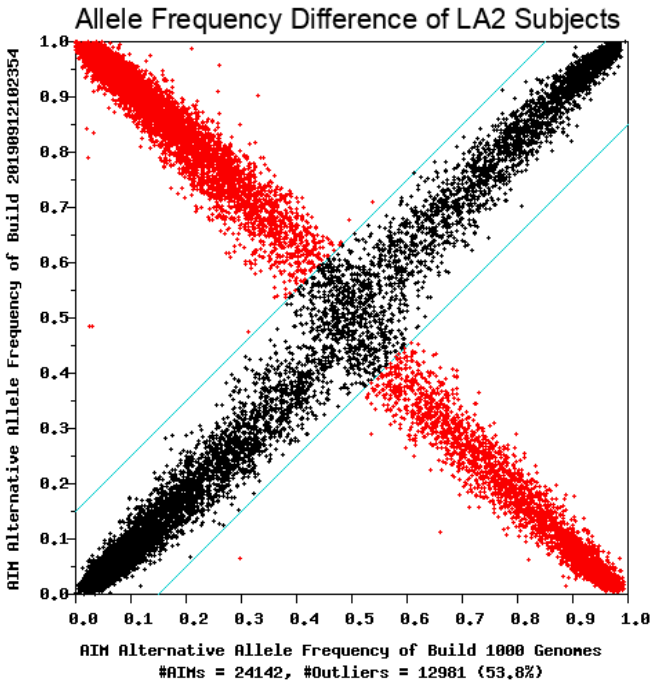
GRAF-pop Computed Populations

ALFA populations
Global
European
Hispanic 1
Hispanic 2
African or African American
African
African American
Asian
South Asian
Asian or Pacific Islander
East Asian
Other Asian
Other



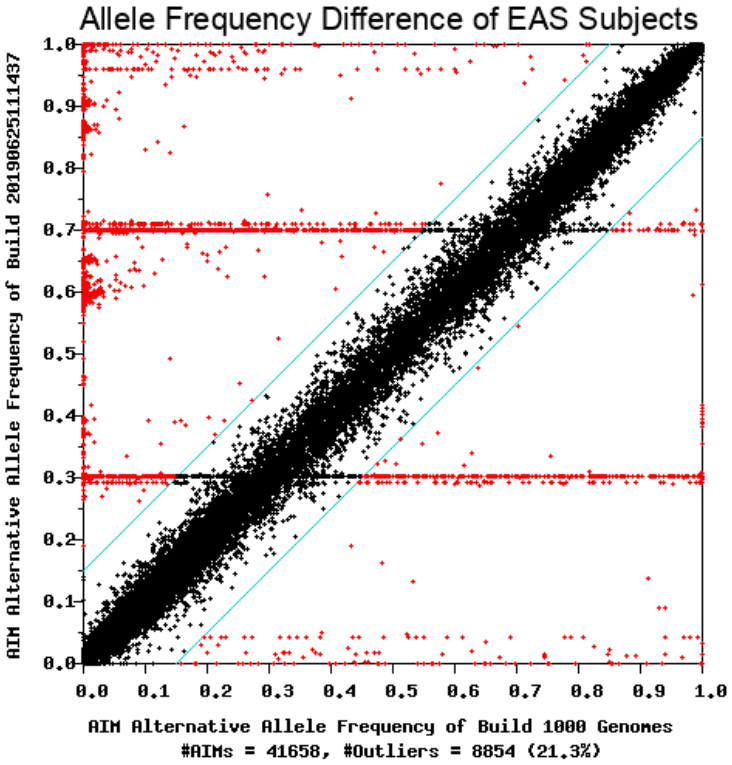
Ancestry Informative Marker (AIM) Analysis



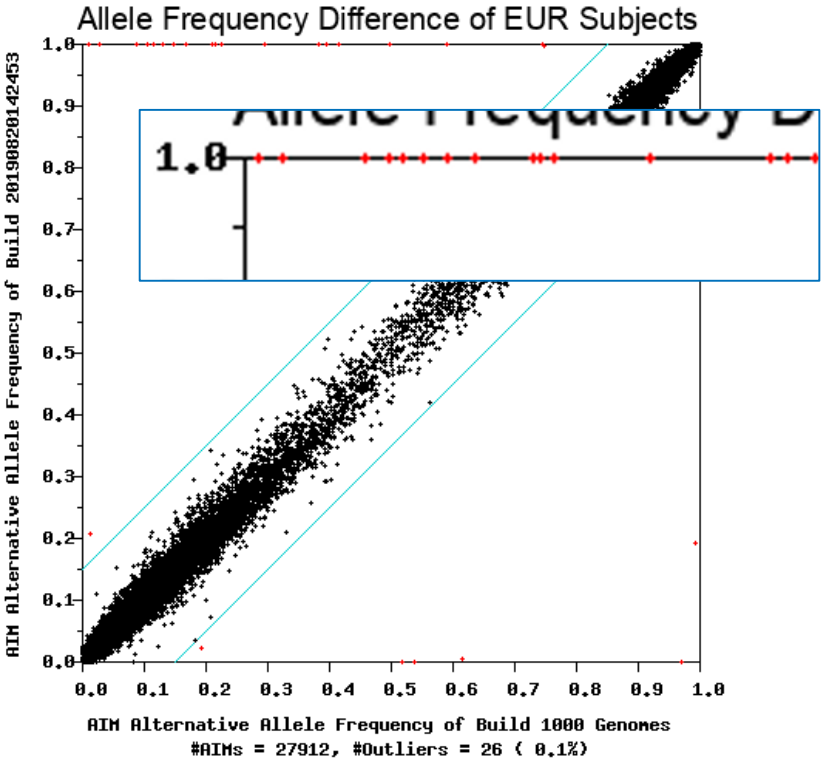


Study Reporting Issues

Software Bug



Assay Error



ALFA Process Overview

Preprocessing

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

Computation

- Variant Normalization
  - SPDI1 conversion
  - Remapping
  - VOCA[1]
  - RefSNP assignment
- Aggregation
  - Individuals and genotype
  - Populations
- Compute frequency

Product Release

- RefSNP Web page
- API
- FTP

[1]Holmes et al. https://www.biorxiv.org/content/10.1101/537449v3

Variant Normalization

Validated VCF

Conversion

Sequence **P**osition **D**eletion **I**nsertion (SPDI)

NC_000008.10:19810825:G:A
NC_000008.11:19953314:G:A

Remap

VOCA normalization

NC_000008.11:19953314:G:A
NC_000008.11:19953314:G:A

Cluster

Assign RS
Compute Frequency

rs328

[1]**Holmes et al.**
https://www.biorxiv.org/content/10.1101/537449v3

ALFA Build Summary

Inputs

Studies	53
Subjects	142,032
Genotypes	696,289,573,125
Genotypes Excluded	791,461,091 (0.1%)

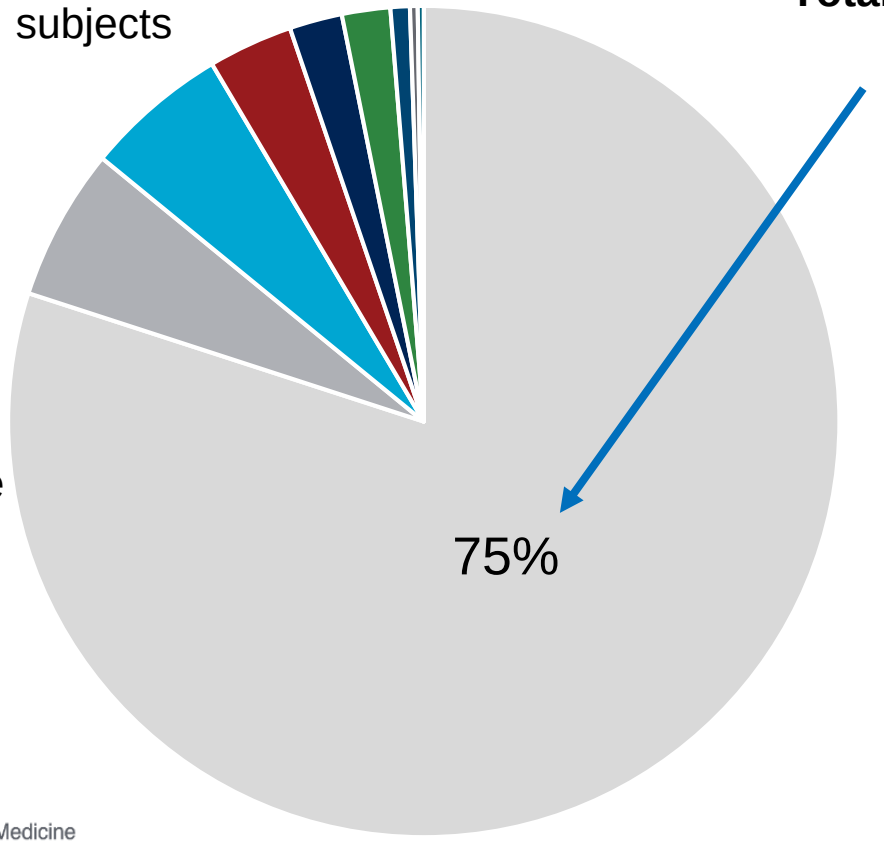
Outputs

RefSNPs	531,167,487
• Exist in dbSNP	512,589,631
• Novel	18,577,856

Population Build Summary

~10% of dbGaP 1.5M subjects

- Increase under represented populations
- New cohorts and populations
- Quarterly release in dbSNP (Coming Soon!!!)



Total (142,032)

- European (113622)
- African American (8412)
- Other (7909)
- Latin American 2 (4680)
- South Asian (2914)
- East Asian (2671)
- Latin American 1 (1069)
- Other Asian (425)
- African (330)

ALFA frequency reported on RefSNP Page

- Variant Details
- Clinical Significance
- **Frequency**
- Aliases
- Submissions
- History
- Publications

dbGaP Population Frequency Project — **Release Version:** 20190529232800

Population	Group	Sample Size	Ref Allele	Alt Allele
Global	Global	173172	T=0.43013	C=0.56987
Europe	Sub	145214	T=0.42688	C=0.57312
All African Ancestry	Sub	8664	T=0.479	C=0.521
95% Exclusive African Ancestry	Sub	304	T=0.48	C=0.52
African American	Sub	8360	T=0.479	C=0.521
Asian	Sub	4546	T=0.506	C=0.494
95% East Asian Ancestry	Sub	4264	T=0.507	C=0.493
South East Asian and Pacific Islanders	Sub	282	T=0.48	C=0.52
Latin American 1	Sub	1032	T=0.390	C=0.610
Latin American 2	Sub	2134	T=0.344	C=0.656
South Asian	Sub	5020	T=0.438	C=0.562
Other	Sub	6562	T=0.413	C=0.587

Filter: Download

Study	Population	Group	Sample Size	Ref Allele	Alt Allele
TopMed	Global	Study-wide	125568	T=0.43702	C=0.56298
The PAGE Study	Global	Study-wide	78702	T=0.4254	C=0.5746
The PAGE Study	AfricanAmerican	Sub	32516	T=0.4690	C=0.5310
The PAGE Study	Mexican	Sub	10810	T=0.3455	C=0.6545

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

FTP VCF

```
shaod@iebdev22 snapshot$ tabix freq.vcf.gz NC_000001.11:12753827-13675453 | head -40
NC_000001.11	12753838	rs886794292	G	A	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753845	rs143511628	G	C	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753847	rs1354810082	T	C	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753858	rs1466919850	A	G	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753863	rs1005120514	A	G	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753869	rs969571672	A	G	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753877	rs1016557243	C	T	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753878	rs1319084848	G	A	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753890	rs898155383	A	G	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753892	rs1232348986	T	G	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753894	rs768999328	ATC	A	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753894	rs1303007114	A	C	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753898	rs1199035339	C	A	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753900	rs1245610285	C	CT	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753901	rs980984054	TC	T	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753902	rs988263928	CT	C,CTT	.	.	.	AN:AC	192:0,0	0:0,0	0:0,0	0:0,0	0:0,0	0:0,0	0:0,0	0:0,0	2:0,0	0:0,0	0:0,0	194:0,0
NC_000001.11	12753902	rs11801468	C	T	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753903	rs1465910853	T	TC	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753904	rs1033393986	T	C	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753911	rs959390179	T	A	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753912	rs1013342431	A	T	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753919	rs946708467	T	TA	.	.	.	AN:AC	192:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0	2:0	0:0	0:0	194:0
NC_000001.11	12753936	rs186910793	A	C,G	.	.	.	AN:AC	192:0,0	0:0,0	0:0,0	0:0,0	0:0,0	0:0,0	0:0,0	0:0,0	2:0,0	0:0,0	0:0,0	194:0,0
NC_000001.11	12753938	rs191780515	C	A,T	.	.	.	AN:AC	192:0,0	0:0,0	0:0,0	0:0,0	0:0,0	0:0,0	0:0,0	0:0,0	2:0,0	0:0,0	0:0,0	194:0,0
```

Conclusions

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

Acknowledgements

dbSNP

Lon Phan
Eugene Shekhtman
Qiang Wang
Hua Zhang
Eric Moyer
David Shao
Dmitry Revoe
Evgeny Ivanchenko
Ricardo Villamarin

MGV Web Team

Nora Husain

dbGaP

Mike Feolo
Yumi Jin
Mas Kimura
Jack Wang
Luning Hao
Nataliya Sharopova
Monica Bihan

CSD

Guangfeng Song
Dany Floisand

Project Manager

Vitaliy Lyoshin

Product Manager

Kuljeet Kaur

Leadership

Brad Holmes
Brandi Kattman

Reminder:
Data CoLab #345
Friday, October 18
10:00AM - 10:30AM

NCBI Booth: #214