Gene-by-gene diversity exploration in large genome data sets

Sofia Hauck

D Phil (PhD) student

UNIVERSITY OF OXFORD

Department of Zoology

NGS: what is it good for?

- High-throughput WGS is already here
  - Proven use in surveillance
  - Extending the benefit of existing data

- Focus **on genes**, not isolates
  - DNA mutates randomly
  - So parts of the genomes with too much or too little diversity are being kept that way by evolution
  - With “big data”, signals stand out from the noise

Questions and **goals**

- How can we use existing data to learn about the biology of pathogens?

- Methods should be…
  - Simple and open (ideally free)
  - Make biological sense
  - Few assumptions
  - Robust to troublesome data

Starting point

- Table of isolates and genes, with allele IDs
  - Plus any categories for loci, if known

id	isolate	species	MYCO000001	MYCO000005	MYCO000006	MYCO000009	MYCO000011	MYCO000013	MYCO000015
1	ATCC 19977	Mycobacterium abscessus		64	119	20	18		
10	CDC1551	Mycobacterium tuberculosis	1	1	2	1	1	1	1
11	F11	Mycobacterium tuberculosis	1	2	2	1	1	2	1
12	H37Ra	Mycobacterium tuberculosis	1	1	1	1	1	1	1
13	H37Rv	Mycobacterium tuberculosis	1	1	1	1	1	1	1

- Access to the FASTA files, either…
  - A folder with the files in your computer
  - Or the name of an open BIGSdb database

First phase: **processing**

- Using Perl & MAFFT
  - Both free to install and use

- Script with options
  - Can consider all alleles or remove singletons
  - Uses default MAFFT parameters or ones given
  - Creates aligned ClustalW and translated FASTA
  - Outputs a table with summary of results

Second phase: **visualisation**

- R with packages ggplot2 and Shiny
  - Again all free to install and use
  - But there is no need to install!
  - Try the app at our website:

https://apps-maidenlab.zoo.ox.ac.uk/GbGDiv/

Measuring **diversity**

$$\frac{\textit{count of unique nucleotide sequences}}{\textit{average length of those sequences}}$$

- Calculated per locus
  - Count of all unique alleles in nucleotide format
  - Length is average of all alleles found
  - **Analogous to substitutions per site**
  - Called “AllelicDiv” in GbGDiv

Measuring **selection**

$$\frac{\textit{count of unique amino acid sequences}}{\textit{count of unique nucleotide sequences}}$$

- Calculated per locus
  - Since have sequences, can translate
  - Then just check for duplicate amino acid sequences
  - **Analogous to dN/dS ratio**
  - Called “RatioCount” in GbGDiv

How **robust** is it?

- No assumption about recombination
  - The unit is the coding sequence
  - Caveat: be careful with interpreting the results
- Missing data doesn't detract from results
  - Can add even pretty poorly assembled genomes
  - Caveat: if gaps are not random, might skew results
- Have the option to ignore singletons (etc.)
  - Removes sequencing errors, lab-grown mutations
  - Caveat: removes most of the data

Closing thoughts

- If you already have allele-based data analysis, explore it from the perspective of the genes
  - Get a grasp on the diversity and selection forces across your loci; plus paralogous, missing, etc.

- Will be publishing soon, please try it out and send feedback so it can be improved!

Acknowledgements

- Supervisor: Martin Maiden
  - James Bray
  - Keith Jolley
- Funding

http://maidenlab.zoo.ox.ac.uk/

GbGDiv Stats & Graphs

Upload your GbGDiv table

For examples and help, click here.

Choose ResultsTable file

Choose File no file selected

Max 'missing' allele tag percentage per locus:

0 20 100

0 10 20 30 40 50 60 70 80 90 100

Please upload a table on the left

Choose variables to plot

Which variable do you want to explore?

AllelicDiv

Which variable do you want in the point colours?

None (labels only)

Optional parameters

Use z-scores for y-axis values

Reset labelled points

Download Plot

Scatter Plot | Excluding points | Data Table

Please upload a table on the left

GbGDiv Stats & Graphs

Upload your GbGDiv table

For examples and help, click here.

Choose ResultsTable file

Choose File RaMi-ResultsTable

Upload complete

Max 'missing' allele tag percentage per locus:

0 20 100

0 10 20 30 40 50 60 70 80 90 100

Isolates reviewed: 7670
Loci included: 3902
85 out of possible 3987 loci were removed. Cutoff of 20 % maximum isolates with missing allele designation filtered out loci in less than 1534 isolates.

Choose variables to plot

Which variable do you want to explore?

AllelicDiv

Which variable do you want in the point colours?

None (labels only)

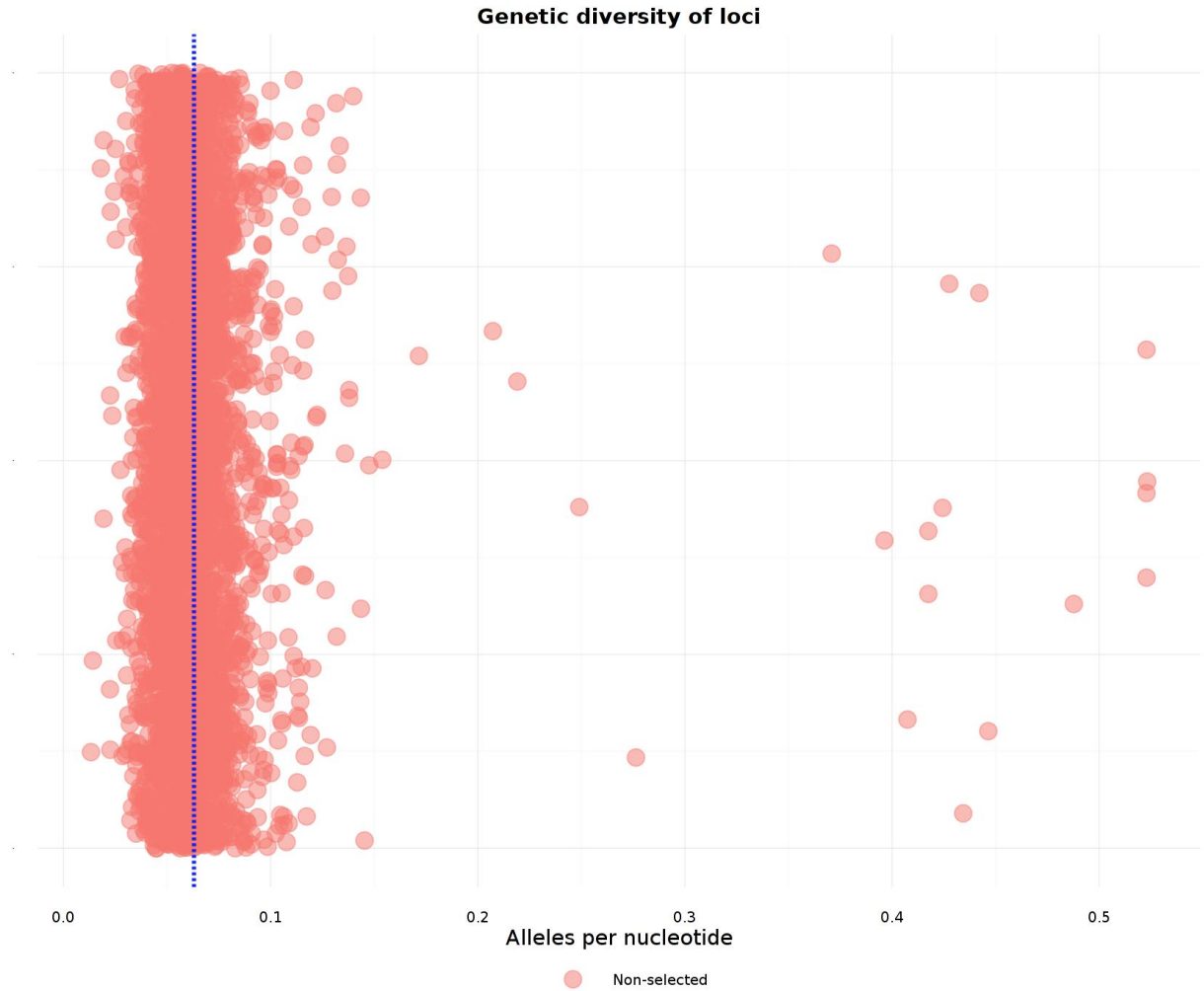
Optional parameters

- [ ] Use z-scores for y-axis values
- [ ] Use categories

Reset labelled points

Download Plot

Scatter Plot | Excluding points | Data Table



GbGDiv Stats & Graphs

Upload your GbGDiv table

For examples and help, click here.

Choose ResultsTable file

Choose File RaMi-ResultsTable

Upload complete

Max 'missing' allele tag percentage per locus:

0 20 100

0 10 20 30 40 50 60 70 80 90 100

Isolates reviewed: 7670
Loci included: 3902
85 out of possible 3987 loci were removed. Cutoff of 20 % maximum isolates with missing allele designation filtered out loci in less than 1534 isolates.

Choose variables to plot

Which variable do you want to explore?

AllelicDiv

Which variable do you want in the point colours?

None (labels only)

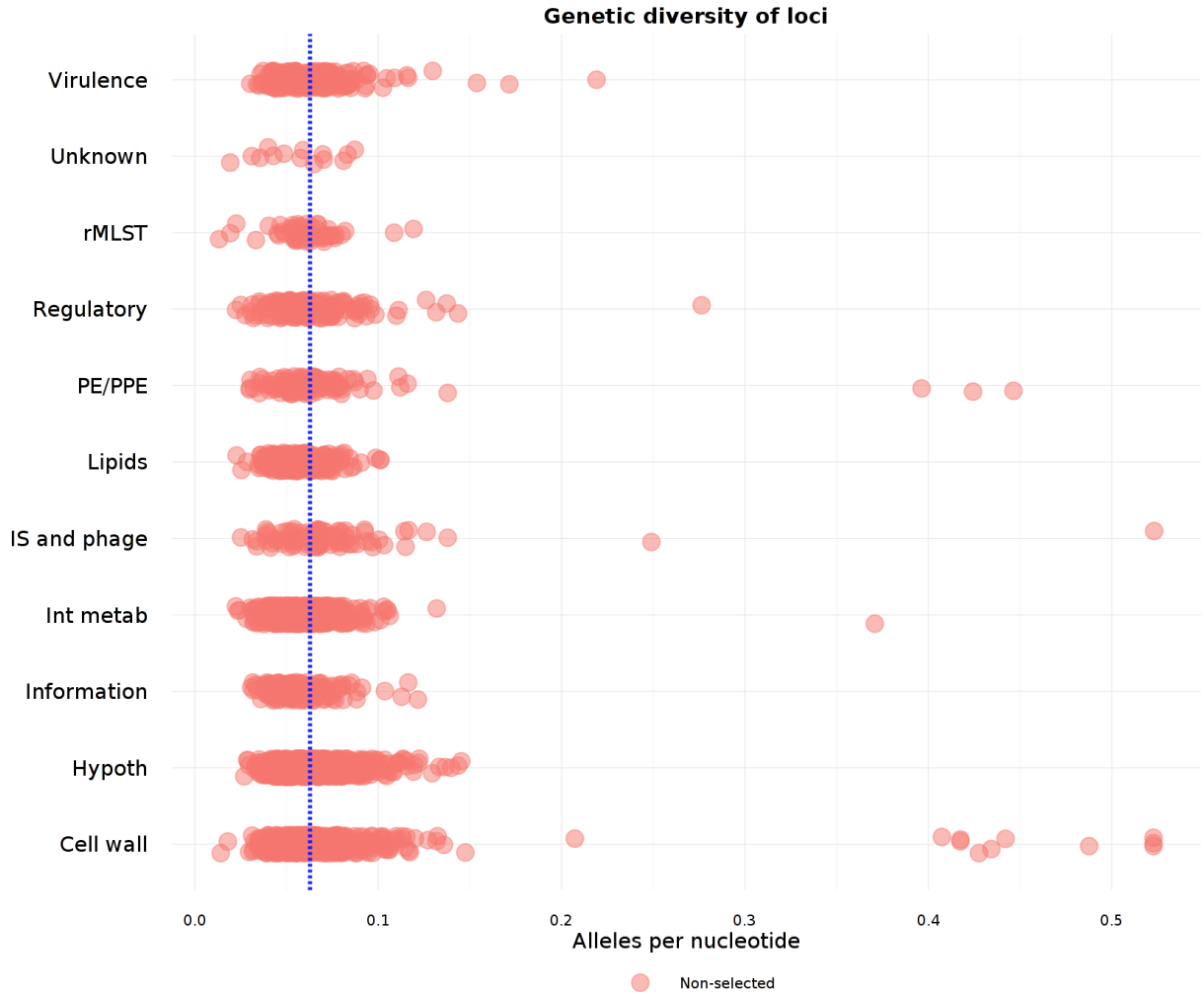
Optional parameters

- [ ] Use z-scores for y-axis values
- [x] Use categories

Reset labelled points

Download Plot

Scatter Plot | Excluding points | Data Table



GbGDiv Stats & Graphs

Upload your GbGDiv table

For examples and help, click here.

Choose ResultsTable file

Choose File RaMi-ResultsTable

Upload complete

Max 'missing' allele tag percentage per locus:

0 20 100

0 10 20 30 40 50 60 70 80 90 100

Isolates reviewed: 7670
Loci included: 3902
85 out of possible 3987 loci were removed. Cutoff of 20 % maximum isolates with missing allele designation filtered out loci in less than 1534 isolates.

Choose variables to plot

Which variable do you want to explore?

AllelicDiv

Which variable do you want in the point colours?

None (labels only)

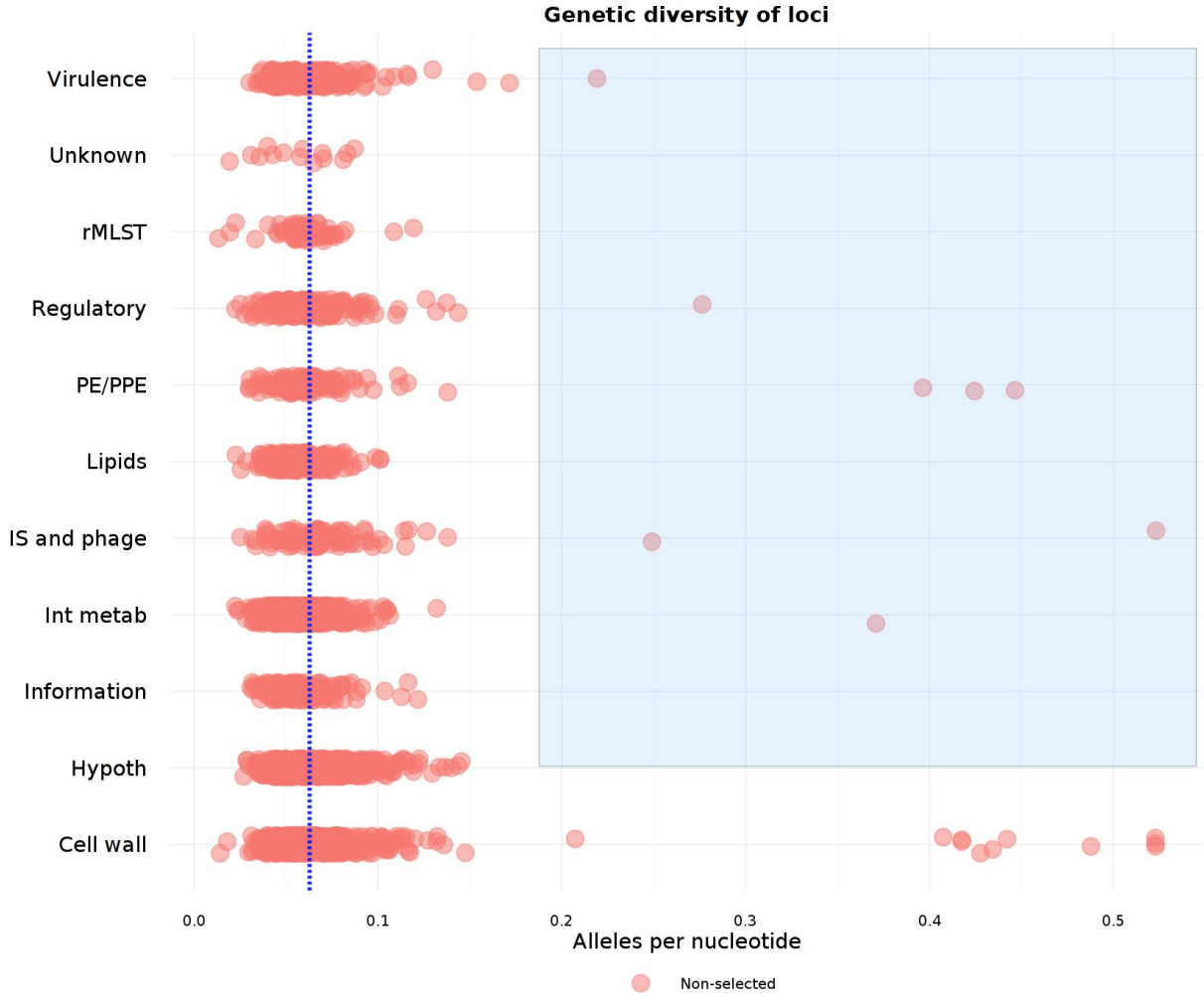
Optional parameters

- [ ] Use z-scores for y-axis values
- [x] Use categories

Reset labelled points

Download Plot

Scatter Plot | Excluding points | Data Table



GbGDiv Stats & Graphs

Upload your GbGDiv table

For examples and help, click here.

Choose ResultsTable file

Choose File RaMi-ResultsTable

Upload complete

Max 'missing' allele tag percentage per locus:

0 20 100

0 10 20 30 40 50 60 70 80 90 100

Isolates reviewed: 7670
Loci included: 3902
85 out of possible 3987 loci were removed. Cutoff of 20 % maximum isolates with missing allele designation filtered out loci in less than 1534 isolates.

Choose variables to plot

Which variable do you want to explore?

AllelicDiv

Which variable do you want in the point colours?

None (labels only)

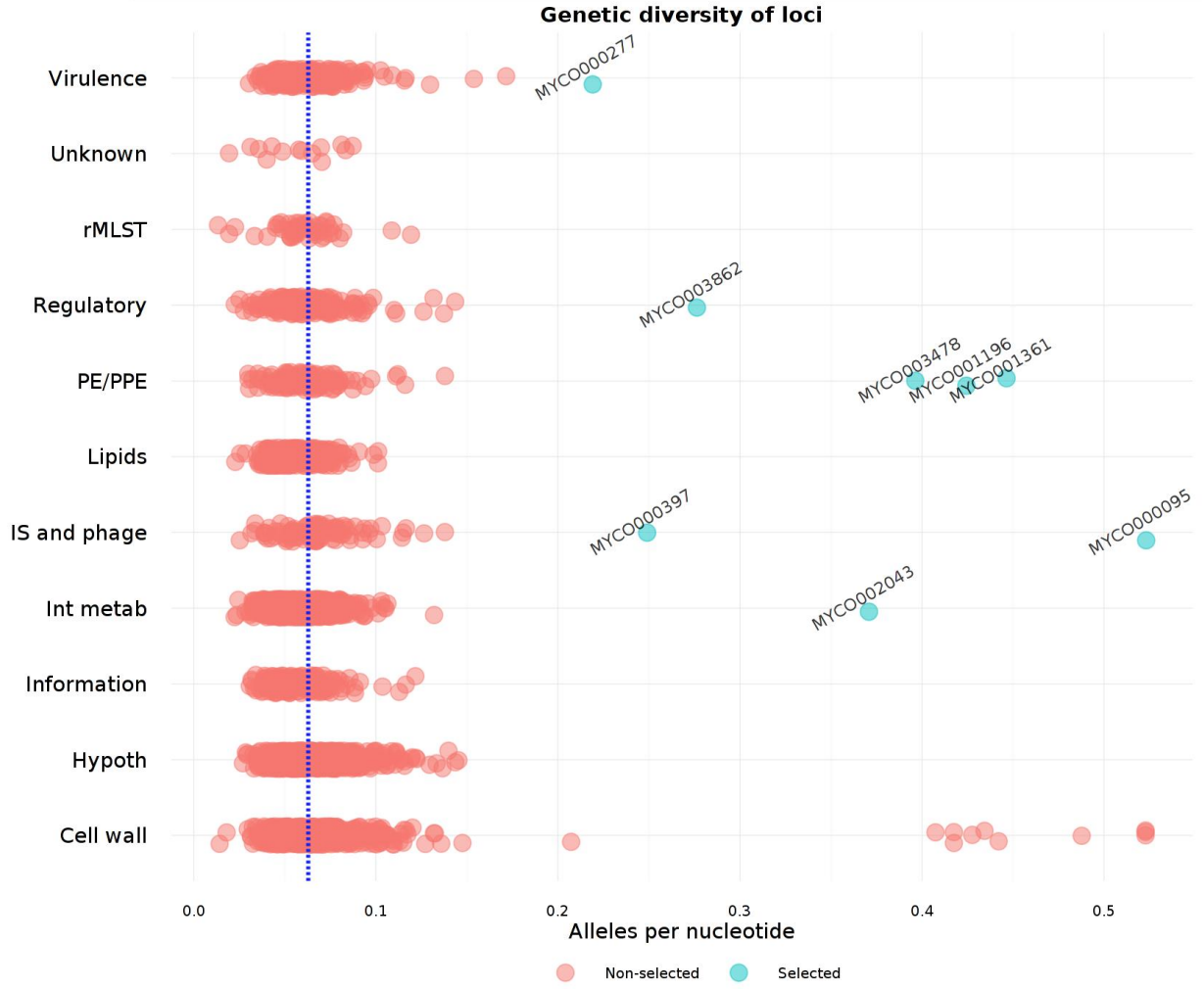
Optional parameters

- [ ] Use z-scores for y-axis values
- [x] Use categories

Reset labelled points

Download Plot

Scatter Plot | Excluding points | Data Table



GbGDiv Stats & Graphs

Upload your GbGDiv table

For examples and help, click here.

Choose ResultsTable file

Choose File RaMi-ResultsTable

Upload complete

Max 'missing' allele tag percentage per locus:

0 20 100

0 10 20 30 40 50 60 70 80 90 100

Isolates reviewed: 7670
Loci included: 3902
85 out of possible 3987 loci were removed. Cutoff of 20 % maximum isolates with missing allele designation filtered out loci in less than 1534 isolates.

Choose variables to plot

Which variable do you want to explore?

RatioCount

Which variable do you want in the point colours?

None (labels only)

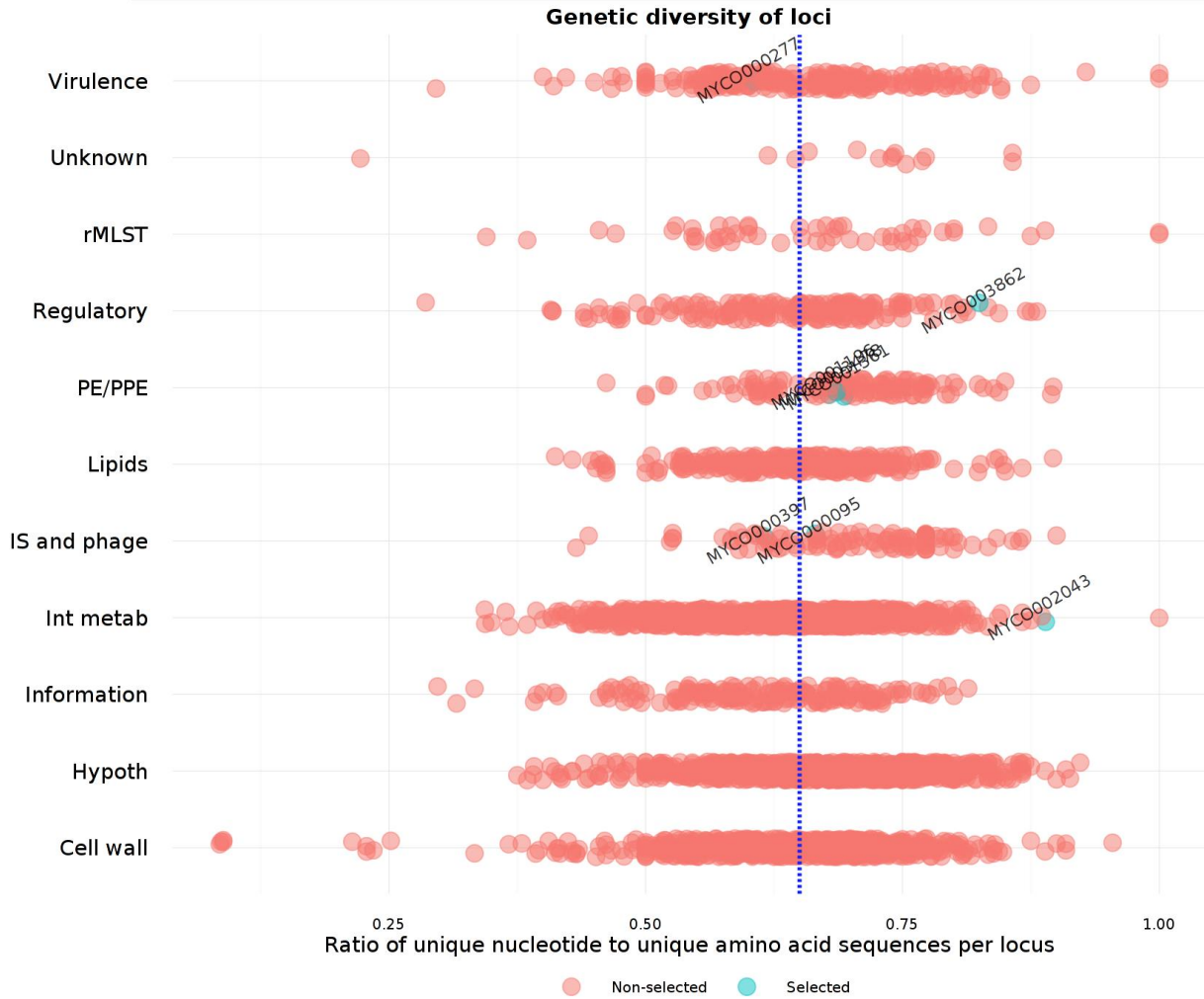
Optional parameters

- [ ] Use z-scores for y-axis values
- [x] Use categories

Reset labelled points

Download Plot

Scatter Plot | Excluding points | Data Table



GbGDiv Stats & Graphs

Upload your GbGDiv table

For examples and help, click here.

Choose ResultsTable file

Choose File RaMi-ResultsTable

Upload complete

Max 'missing' allele tag percentage per locus:

0 20 100

0 10 20 30 40 50 60 70 80 90 100

Isolates reviewed: 7670
Loci included: 3902
85 out of possible 3987 loci were removed. Cutoff of 20 % maximum isolates with missing allele designation filtered out loci in less than 1534 isolates.

Choose variables to plot

Which variable do you want to explore?

RatioCount

Which variable do you want in the point colours?

None (labels only)

Optional parameters

- [ ] Use z-scores for y-axis values
- [x] Use categories

Reset labelled points

Scatter Plot | Excluding points | Data Table

Show 25 entries Search:

Locus	Category	Missing	Paralogous	CountNuc	CountAA	MinLength	MaxLength
MYCO000095	IS and phage	3282	0	215	143	408	411
MYCO000277	Virulence	245	3229	94	57	429	429
MYCO000397	IS and phage	1168	2567	91	56	357	375
MYCO001196	PE/PPE	3800	558	501	340	1170	1191
MYCO001361	PE/PPE	3688	575	528	366	1170	1194
MYCO002043	Int metab	297	0	208	185	546	570
MYCO003478	PE/PPE	3805	414	468	321	1161	1191
MYCO003862	Regulatory	119	0	97	80	351	351
BACT000001	rMLST	21	0	103	72	1443	1446
BACT000002	rMLST	218	0	39	30	861	867
BACT000003	rMLST	742	1	46	30	654	825
BACT000004	rMLST	40	1	34	18	603	606
BACT000005	rMLST	125	1	38	26	501	675
BACT000006	rMLST	7	1	20	15	291	297
BACT000007	rMLST	16	1	22	10	471	471
BACT000008	rMLST	30	1	29	10	399	399
BACT000009	rMLST	33	1	33	22	393	456
BACT000010	rMLST	3	1	15	9	306	309
BACT000011	rMLST	62	1	28	16	390	420

GbGDiv Stats & Graphs

Upload your GbGDiv table

For examples and help, click here.

Choose ResultsTable file

Choose File RaMi-ResultsTable

Upload complete

Max 'missing' allele tag percentage per locus:

0 20 100

0 10 20 30 40 50 60 70 80 90 100

Isolates reviewed: 7670
Loci included: 3902
85 out of possible 3987 loci were removed. Cutoff of 20 % maximum isolates with missing allele designation filtered out loci in less than 1534 isolates.

Choose variables to plot

Which variable do you want to explore?

AllelicDiv

Which variable do you want in the point colours?

Missing

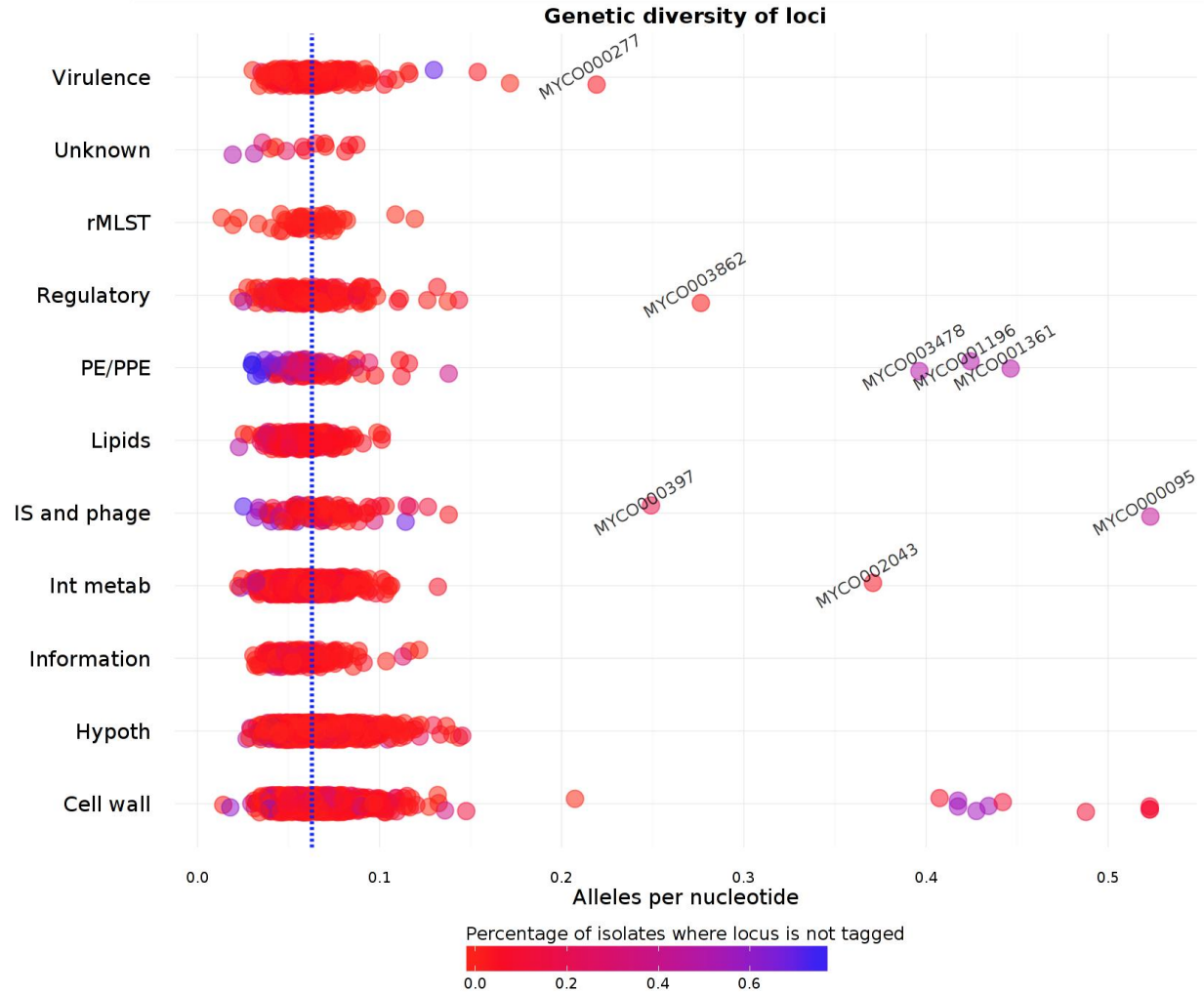
Optional parameters

- [ ] Use z-scores for y-axis values
- [x] Use categories

Reset labelled points

Download Plot

Scatter Plot | Excluding points | Data Table



BIGS DB two-database system

Genomes database

Isolate #141:

(N/A, *F11*, South Africa, epidemic…)

Isolate #142:

(1934, *H37Rv*, China, lab strain…)

Loci database

Locus #12:

Locus #13:

Locus #14:

Locus #15:

BIGS DB two-database system

Genomes database

Isolate #141:
(N/A, *F11*, South Africa, epidemic…)
Isolate #142:
(1934, *H37Rv*, China, lab strain…)

Loci database

Locus #12:
Locus #13:
Locus #14:
Locus #15:

scanning & tagging

partial match

BIGS DB two-database system

Genomes database

Isolate #141:
(N/A, *F11*, South Africa, epidemic…)
Isolate #142:
(1934, *H37Rv*, China, lab strain…)

Loci database

Locus #12:
Locus #13:
Locus #14:
Locus #15:

new allele is defined