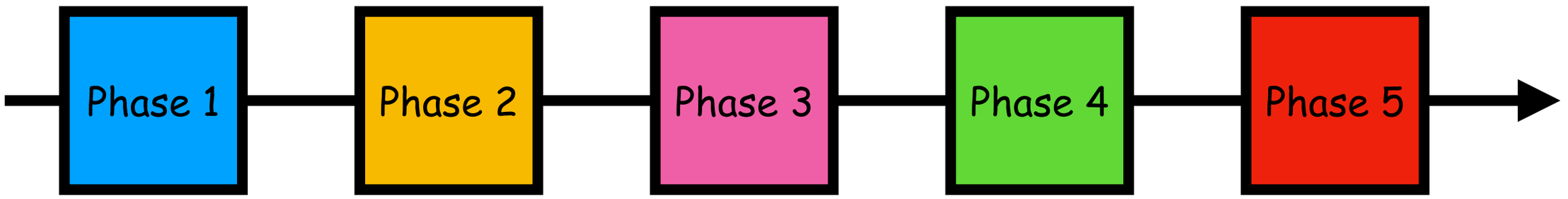
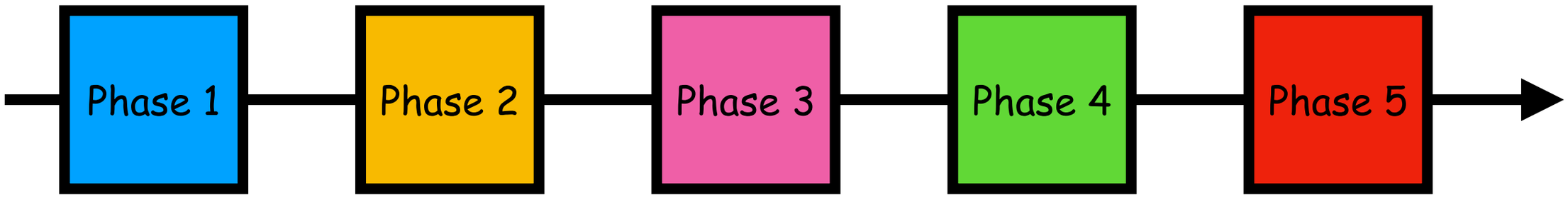
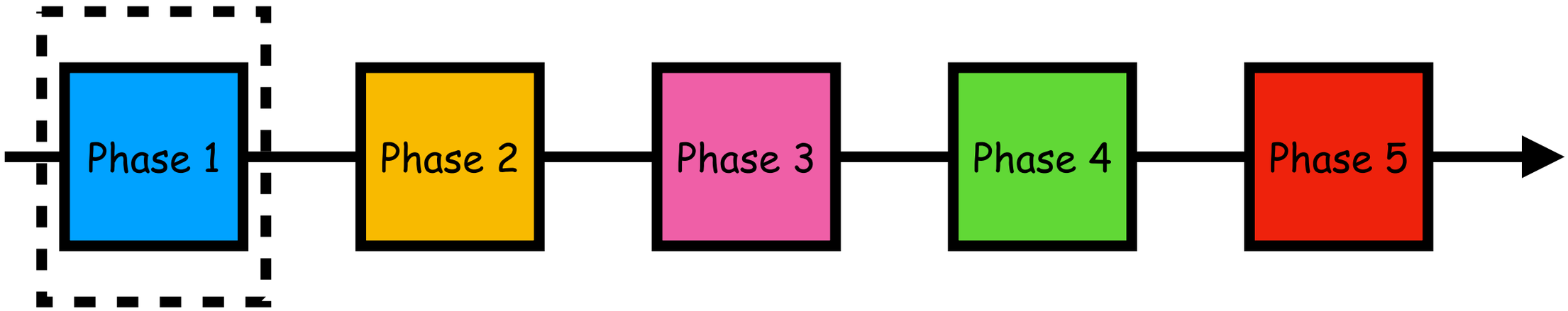
Convergence Analysis

Tomz notes

We have identified 5 phases to the analysis and have a rough idea of what each one will look like



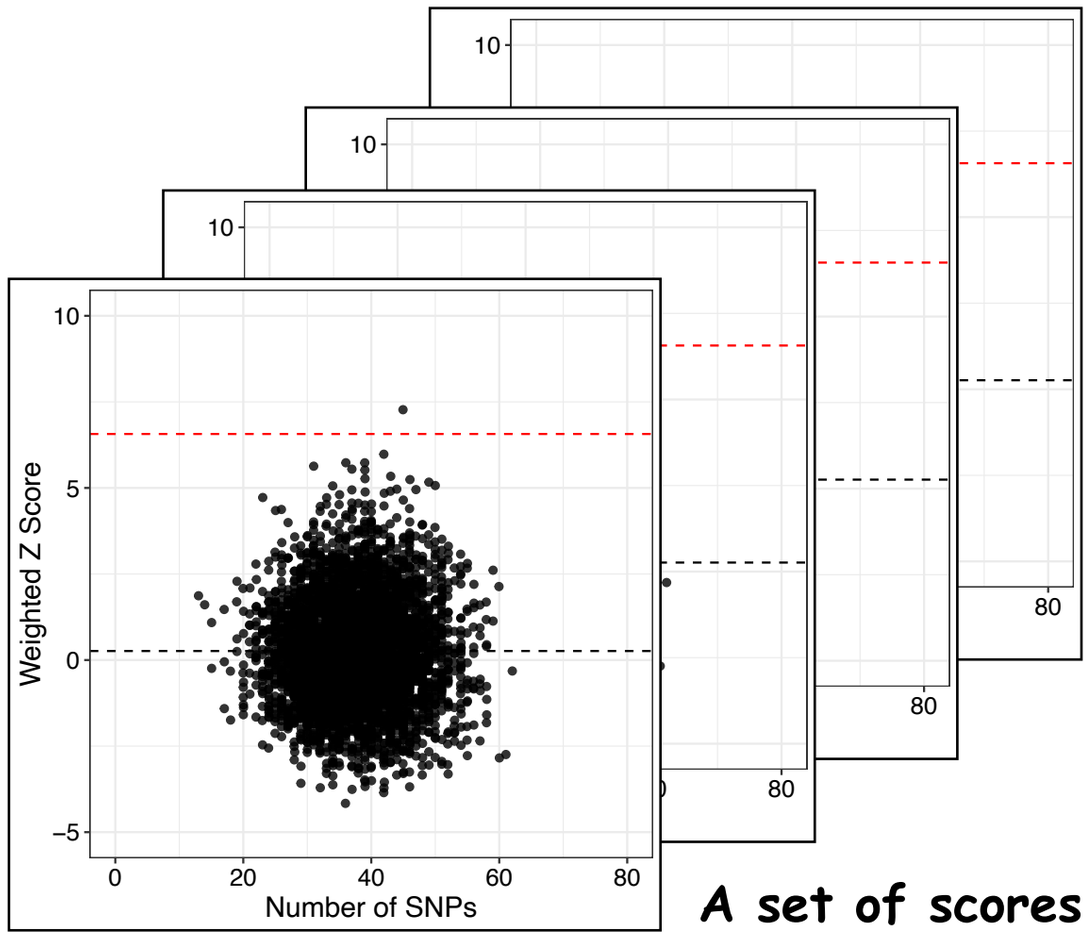

Phase 1
Phase 2
Phase 3
Phase 4
Phase 5
A specific score for each gene/ window in the genome
We've devised the WZA for this
Within a species, combine information across orthogroups
Correcting for multiple comparisons
Across all species, compare the evidence that a particular orthogroup is "interesting"
We've devised the pMax for this
Combine information across all comparisons to get a single score for each orthogroup present in the clade
A post-hoc comparison of genes to ask question like, is there evidence that the top hits often exhibit high LD?



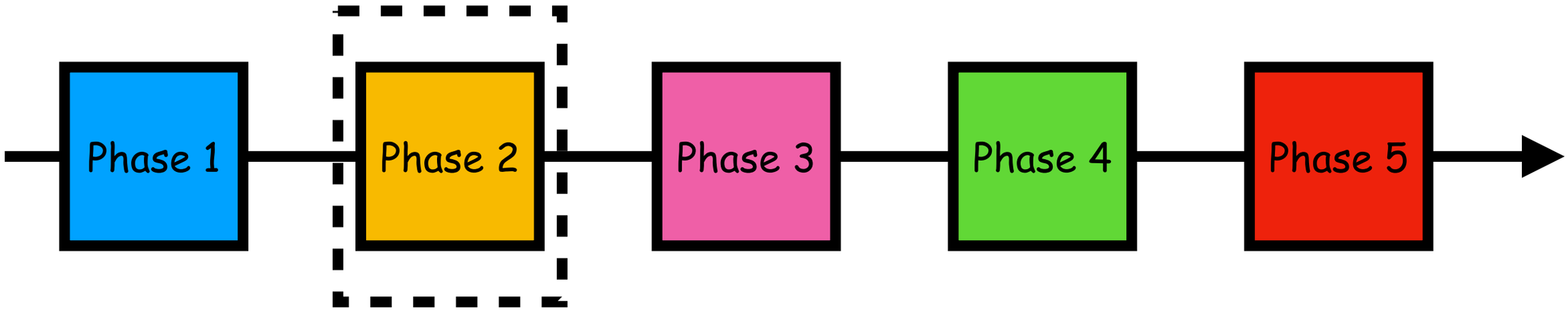
A specific score for each gene/window in the genome

We've devised the WZA for this

$$Z_{w,k} = \frac{\sum_{i=1}^{n} \overline{p_i q_i} z_i}{\sqrt{\sum_{i=1}^{n} (\overline{p_i q_i})^2}},$$



A set of scores for each species being examined



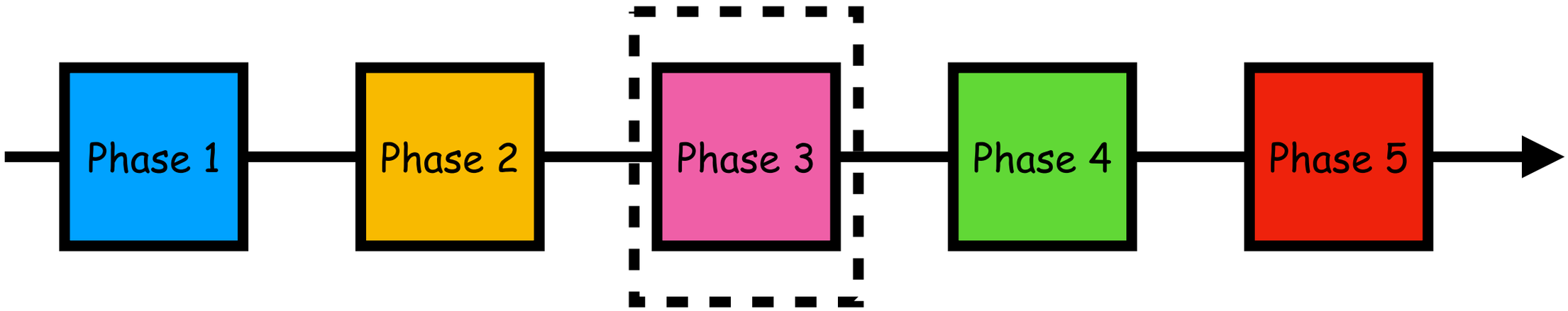
Within a species, combine information across orthogroups
Correcting for multiple comparisons

For each species we have a table that looks something like this:

Gene	Orthogroup	Zw score
1	1	0.128
2	1	0.261
3	2	0.655
4	3	0.085
5	3	0.157
6	3	0.824
7	4	0.503
8	5	0.112

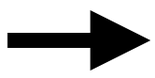
We need a table that looks like this:

Orthogroup	Empirical p-value
1	0.200
2	0.400
3	0.600
4	0.800
5	1.000

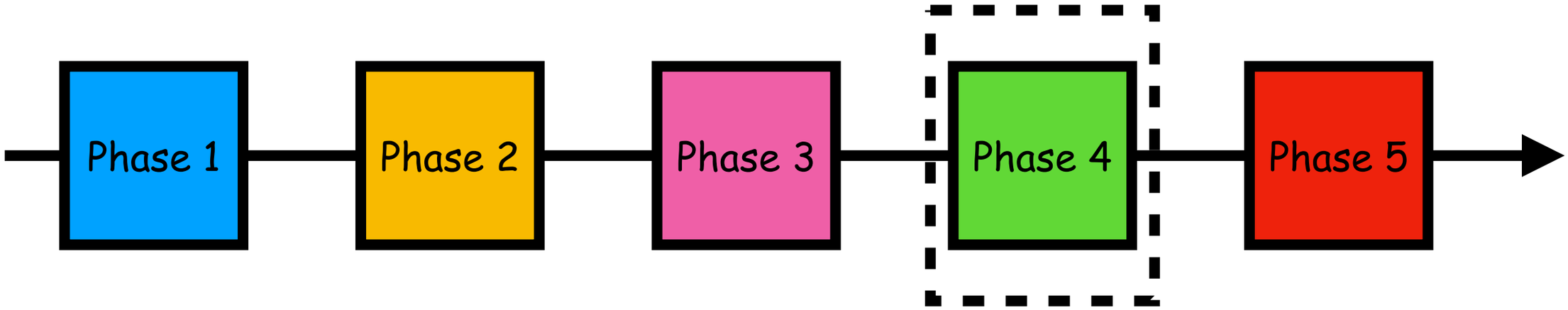


Across all species, compare the evidence that a particular orthogroup is "interesting"
The pMax was devised for this

Orthogroup	Empirical p-value Species 1	Empirical p-value Species 2	Empirical p-value Species 3	Empirical p-value Species 4
1	0.200	0.400	0.100	0.050
2	0.400	0.300	0.500	0.200
3	0.600	N/A	0.800	0.400
4	0.800	0.800	0.001	0.001
5	1.000	0.900	N/A	N/A



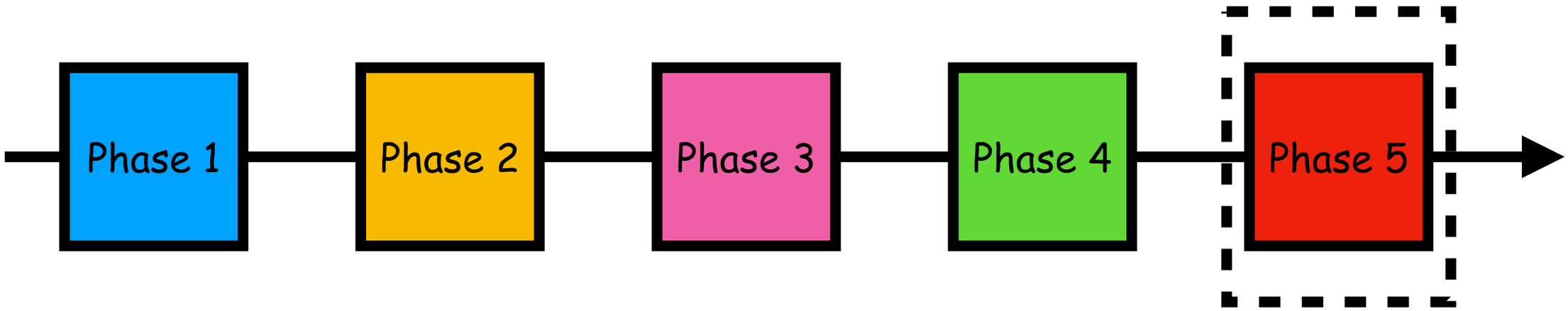
Orthogroup	pMax 2 Species	pMax 3 Species	pMax 4 Species
1	0.050	0.012	0.500
2	0.123	0.213	0.321
3	0.900	0.500	N/A
4	0.001	0.300	0.900
5	0.900	N/A	N/A



Combine information across all comparisons to get a single score for each orthogroup present in the clade

Orthogroup	pMax 2 Species	pMax 3 Species	pMax 4 Species
1	0.050	0.012	0.500
2	0.123	0.213	0.321
3	0.900	0.500	N/A
4	0.001	0.300	0.900
5	0.900	N/A	N/A

Orthogroup	p-value of convergence (or something)
1	0.050
2	0.123
3	0.900
4	0.001
5	0.900



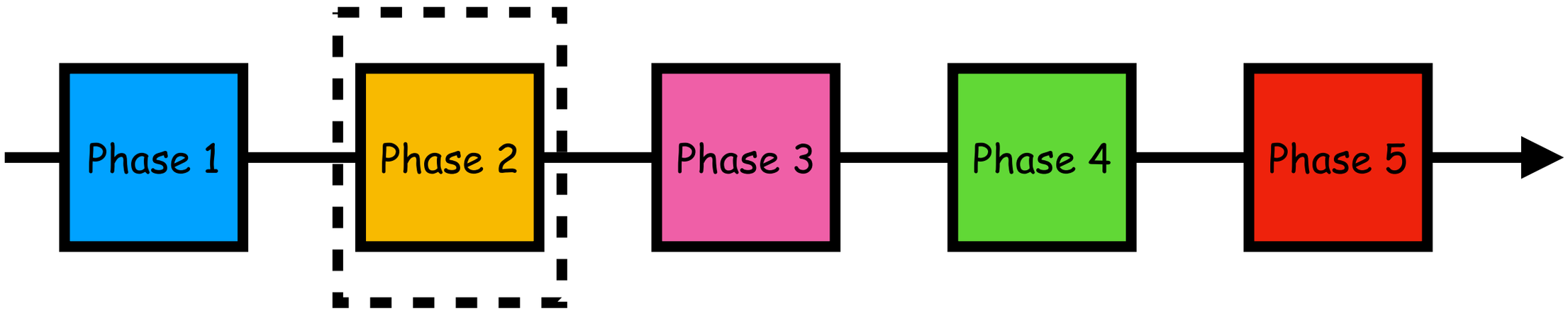
A post-hoc comparison of genes to ask question like, is there evidence that the top hits often exhibit high LD?

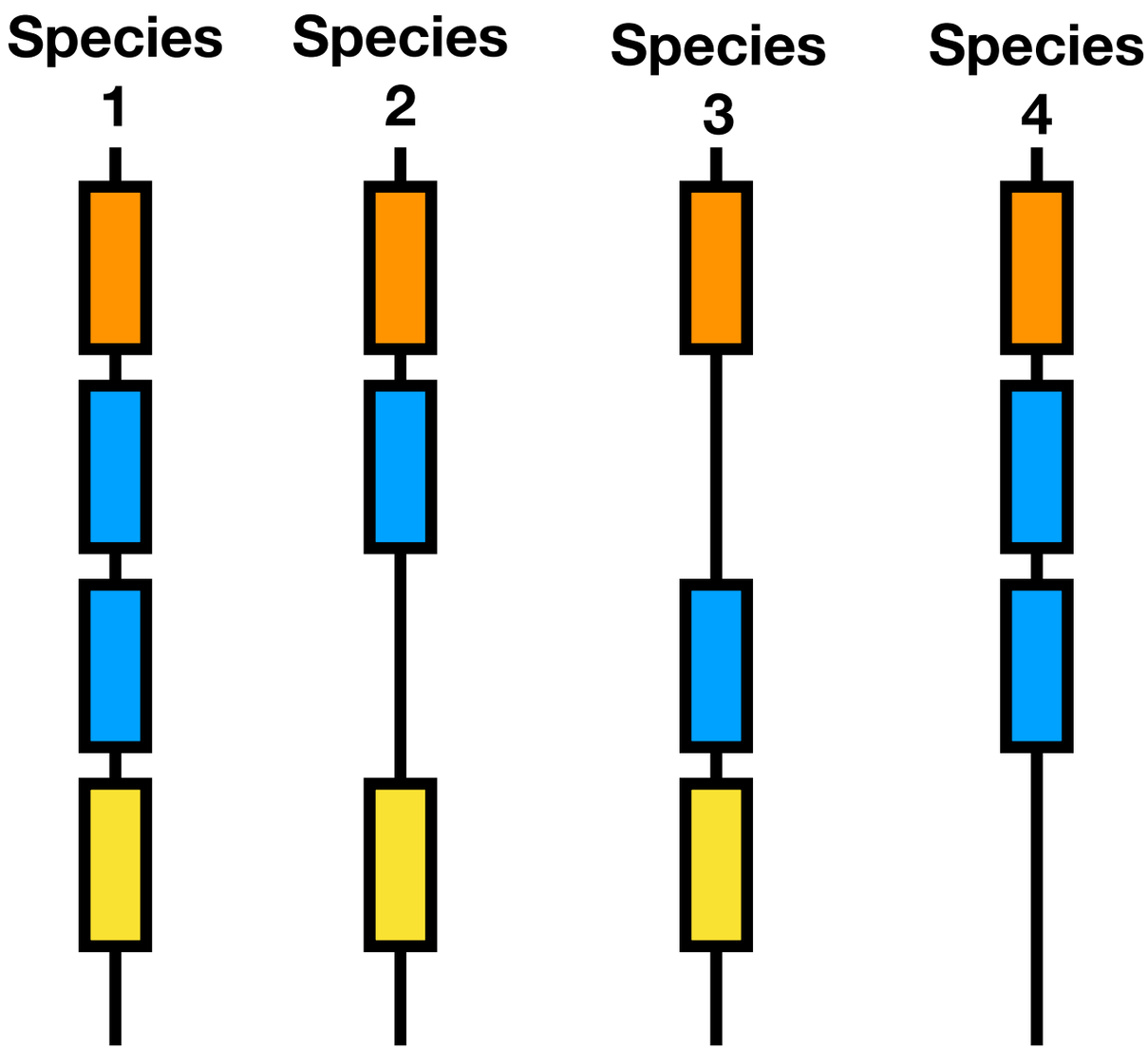
Orthogroup	p-value of convergence (or something)
1	0.050
2	0.123
3	0.900
4	0.001
5	0.900

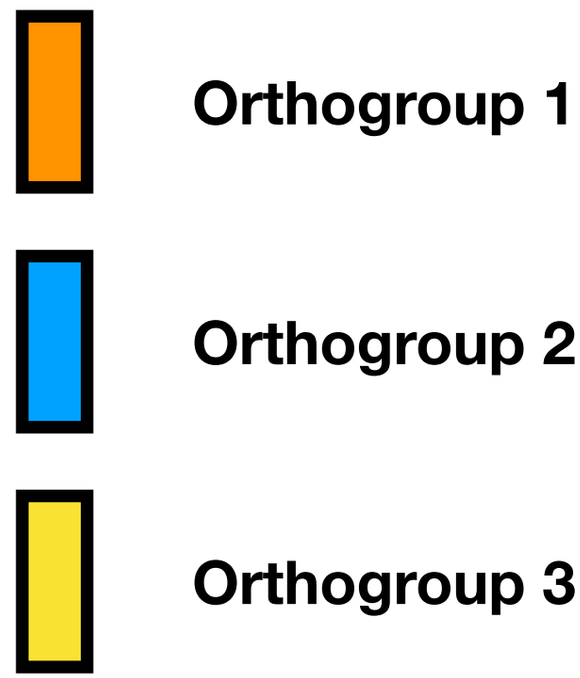
Perhaps there is very high LD between orthogroups 1 and 2

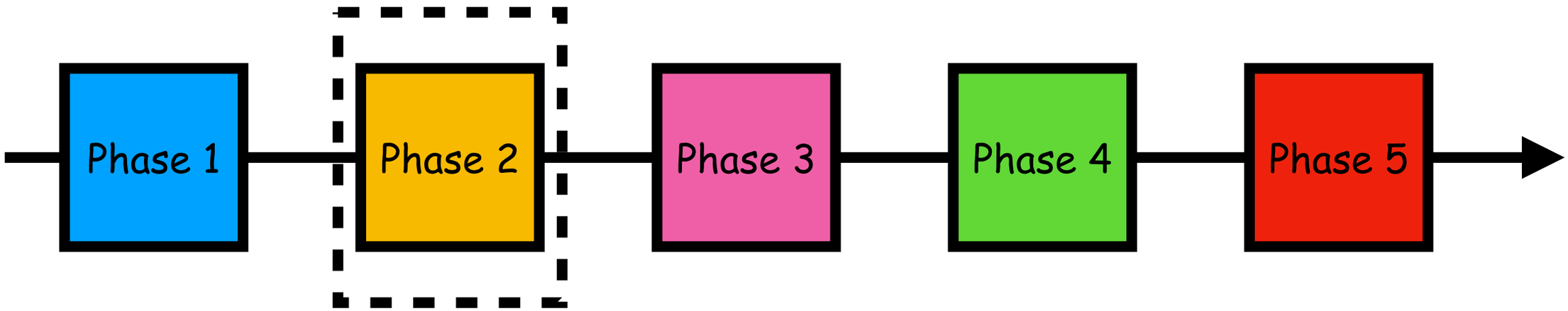
How do we deal with that?

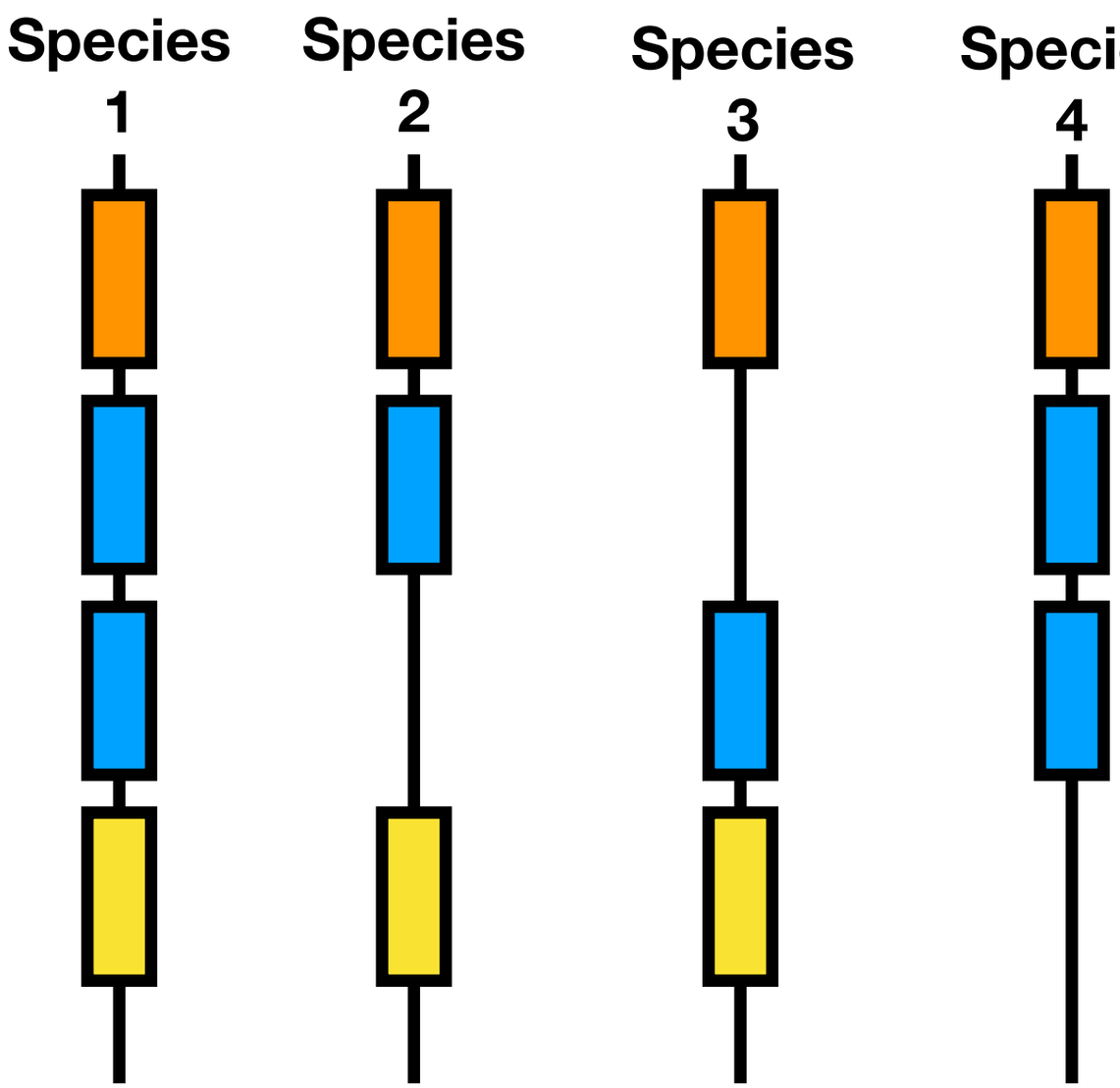
I don't know


Phase 1
Phase 2
Phase 3
Phase 4
Phase 5


Species 1
Species 2
Species 3
Species 4


Orthogroup 1
Orthogroup 2
Orthogroup 3

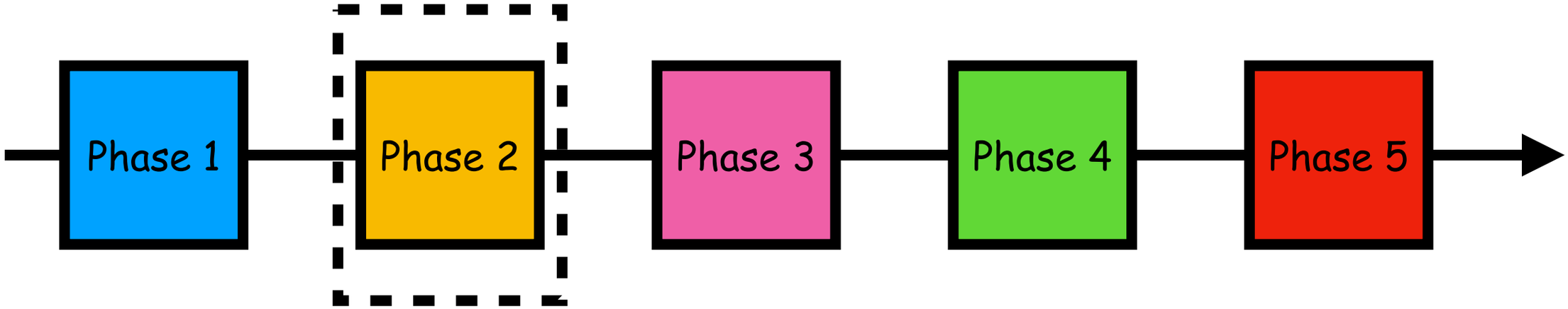


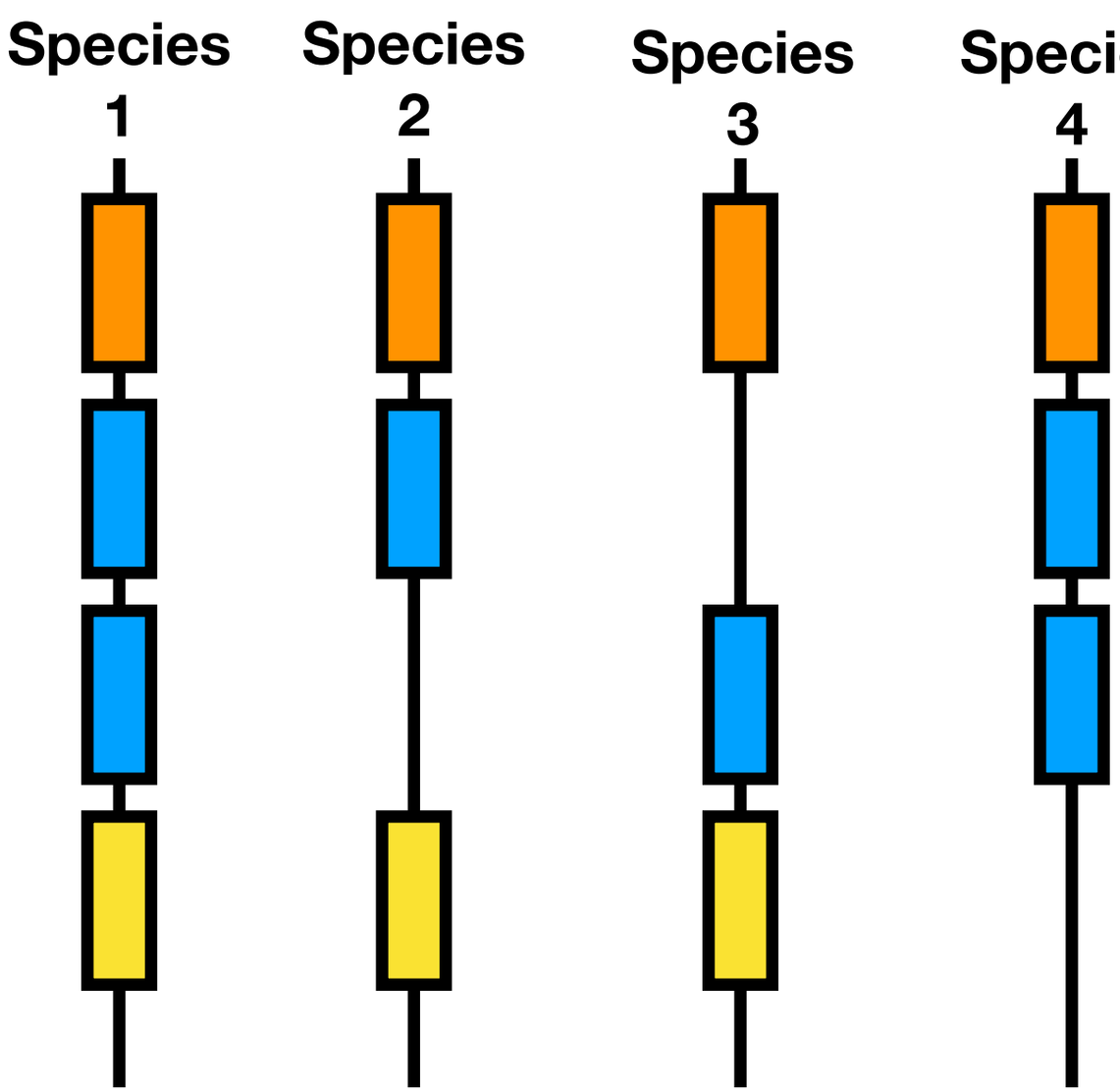


It seems highly likely that we would not be able to determine which member of an orthogroup corresponds to all others in the different species

Besides, it seems like the question of convergence is (to my understanding) whether a particular orthogroup is used in local adaptation in a particular lineage

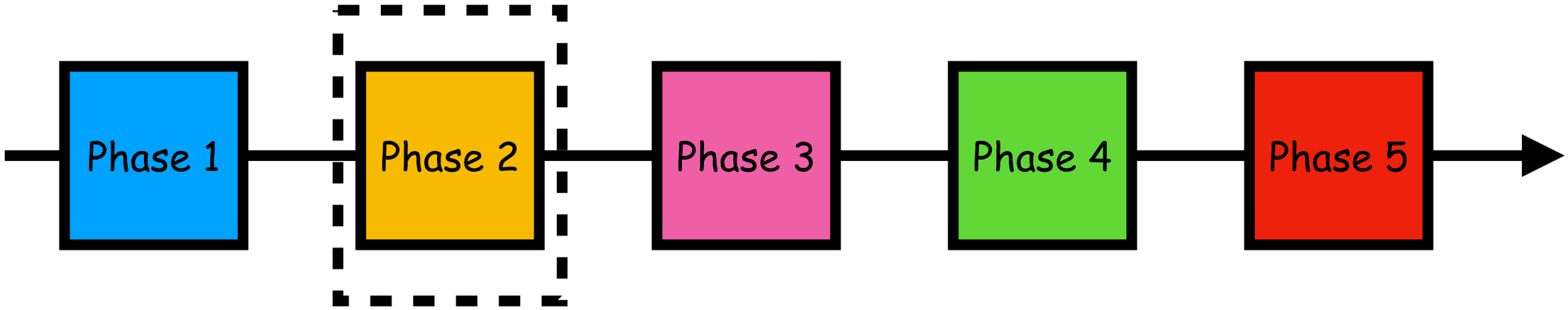
With that in mind, we can reframe the question as “is there evidence that a particular orthogroup is used in local adaptation?”





We have a table like this for each species

Gene	Ortho-group	Species 1 Zw
1	1	0.955
2	2	0.588
3	2	0.959
4	3	0.650



So let's take a table like this for each species and get a single score for each orthogroup

Gene	Ortho-group	Species 1 Zw
1	1	0.691
2	2	0.313
3	2	0.120
4	3	0.727

With this orthogroup way of framing it, these two Zw scores become a case of multiple tests of the same hypothesis

Converting these to a single score per gene could be done by taking the minimum p-value, after correcting for multiple comparisons

That was our first idea, so I'll look into that first