1262022_analysis

Tanvi Ingle

1/27/2022

Objective

Here I will analyze the results from the promoter optimization run on 1-26-2022. For this optimization attempt, I ran 5 scenarios with unique distributions for promoters A, B, and D. Each scenario was simmulated 10 times. Each simulation was run for 300 generations.

Here are the details of each run. Each value represents the exponent on e, ie if uA = 11, then the mean strength of promoter A is eˆ11.

Starting promoter distributions

Dive into Scenario A

What do the reports for each simulation look like? For Scenario A, I want to explore the parameter space (ie promoter strengths) below what I found manually. First, I look at how the RMSE changes over 300 generations in each of the 10 simulations. Here are my observations:

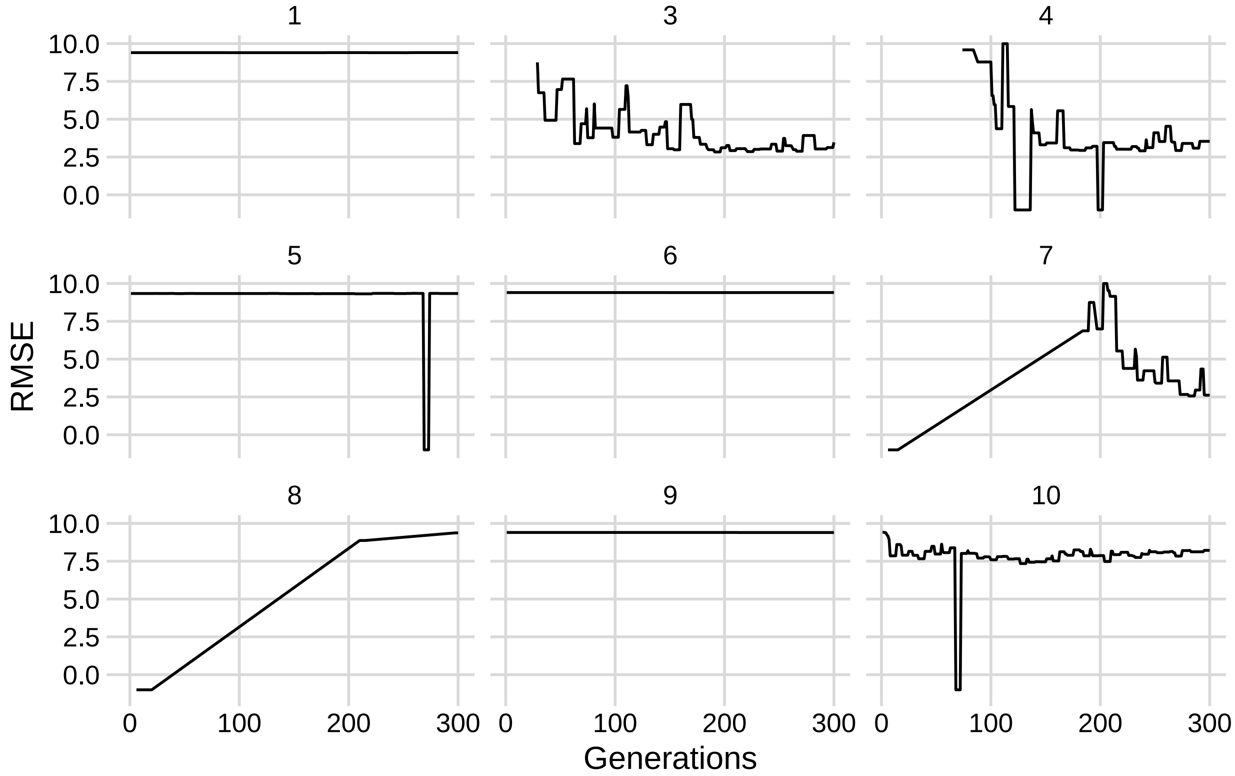
1. Simulation 2 report is missing. For some reason it quit the simulation at generation XXX. Next time I'll use 'nohub' with the disown command so that it will save the output & I can trace the error.
2. Simulations 3, 4, 7, and 10 have the lowest errors. Simulations 1, 5, 6, and 9 look pretty similar, with an error around 9ish. Simulation 8 was the only one which (strangely) increased in error over time. From this it seems like the optimization techniques works ok – **should I run it for more generations?**

   Scenario A - Change in Error over Generations

scenario	uA	oA	uB	oB	uD	oD
A	11.00	10	12.80	10	12.30	11
B	11.80	10	13.50	12	12.80	11
C	12.21	10	14.33	12	14.11	11
D	12.80	10	14.80	12	14.50	11
E	13.20	10	15.40	12	14.80	11

gen	pA	pB	pD	error	sim
300	871392.5	15033897.1	176176733	3.47082	3
300	816566.8	594394.9	3175453602	3.53530	4
300	586754.0	12450376.6	57183432641	2.61829	7
300	919414.3	595647.6	1540429	8.21875	10

Figure 1: Scenario A



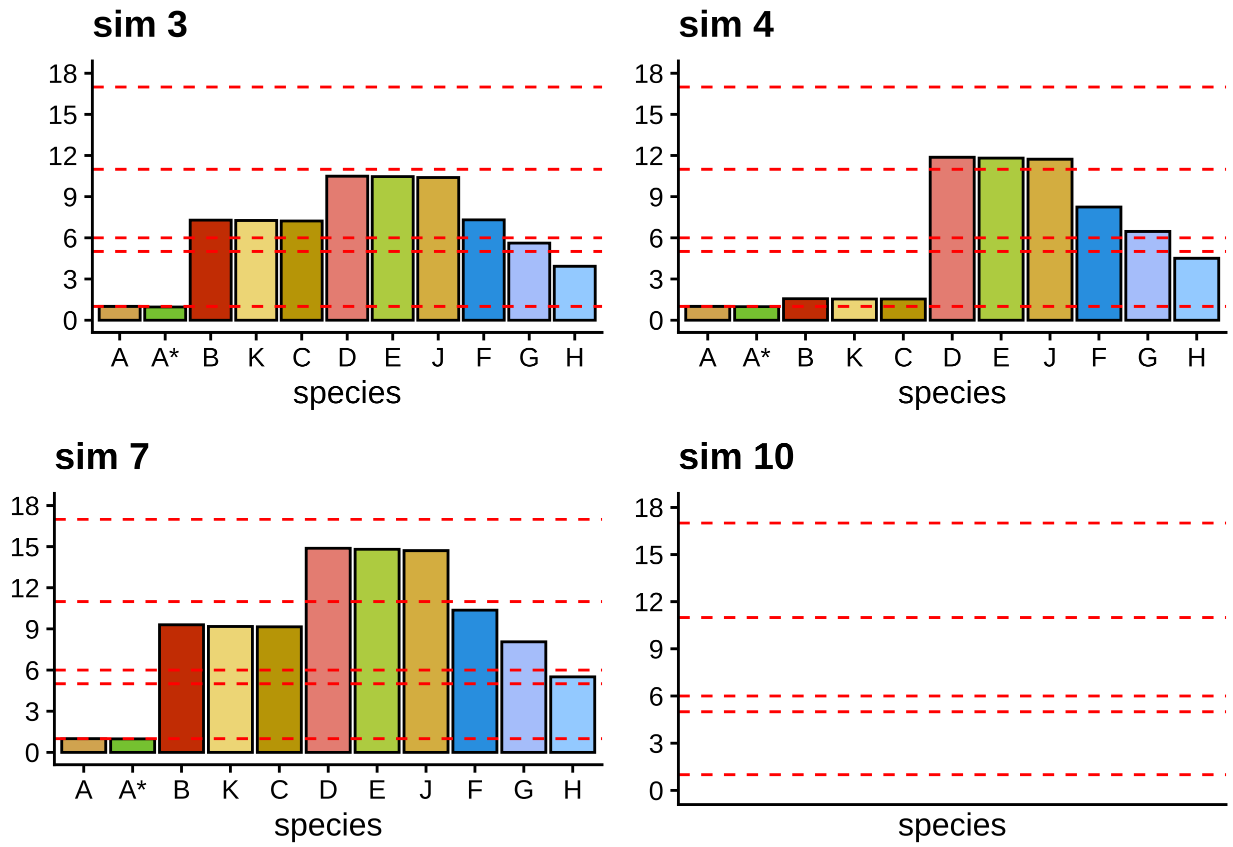
What do the distributions look like for the lowest error simulations?

Scenario A - Associated Errors for the best simulations

For Scenario A, simulations 3, 4, 7, and 10 had the lowest errors at the 300th generation.

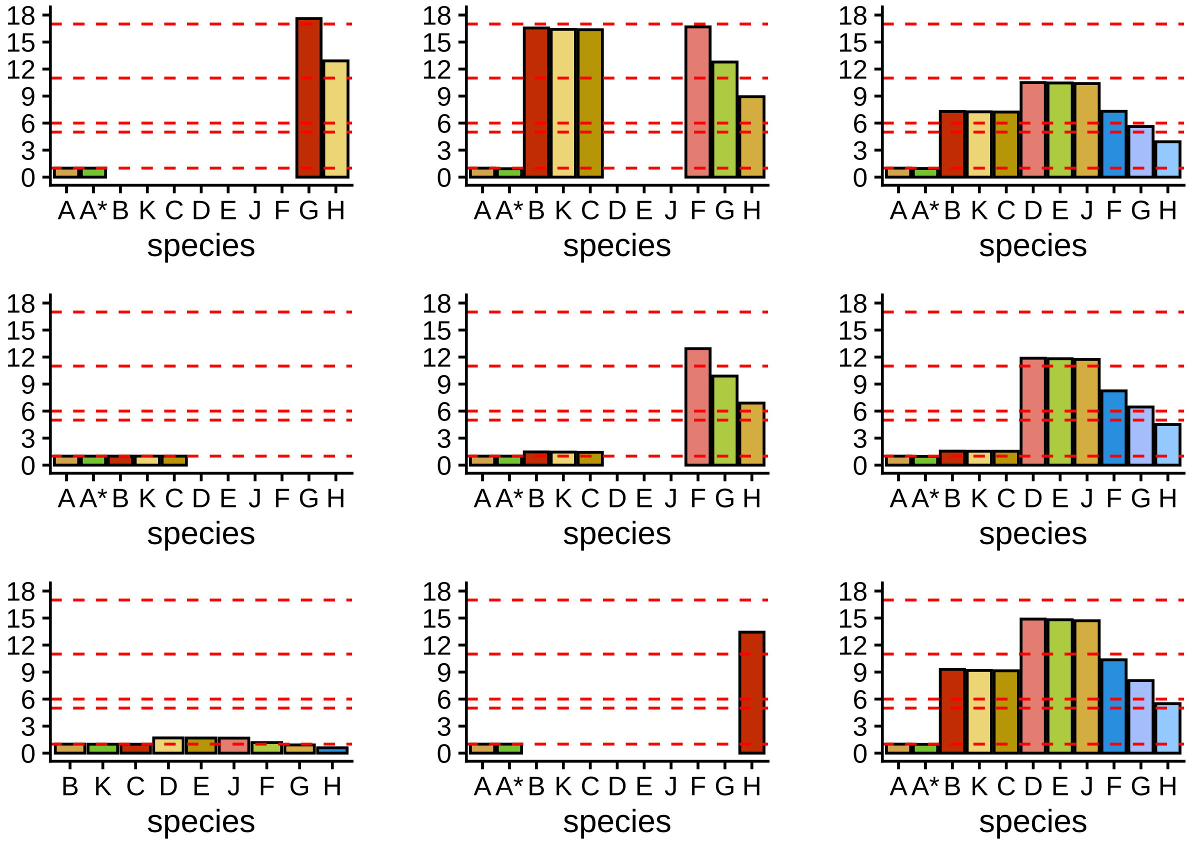
For Scenario A, the distributions for sims 3, 4, and 7 are qualitatively the best. Simulation 10 had no expression, revealing that RMSE score alone is not enough to find the best promoter values. **Should I use another error metric? cosine error?** Visually & quantitatively, sim 7 is the best of the Scenario A batch.

Final Transcript Abundances for the best of Scenario A



Here is how the sims 3, 4, and 7 change over time (gen 1, gen 150, gen 300). TOP = sim 3, MIDDLE = sim 4, and BOTTOM = sim 7

Evolution of transcript abundances over generations (Sim 3, 4, 7) - Scenario A



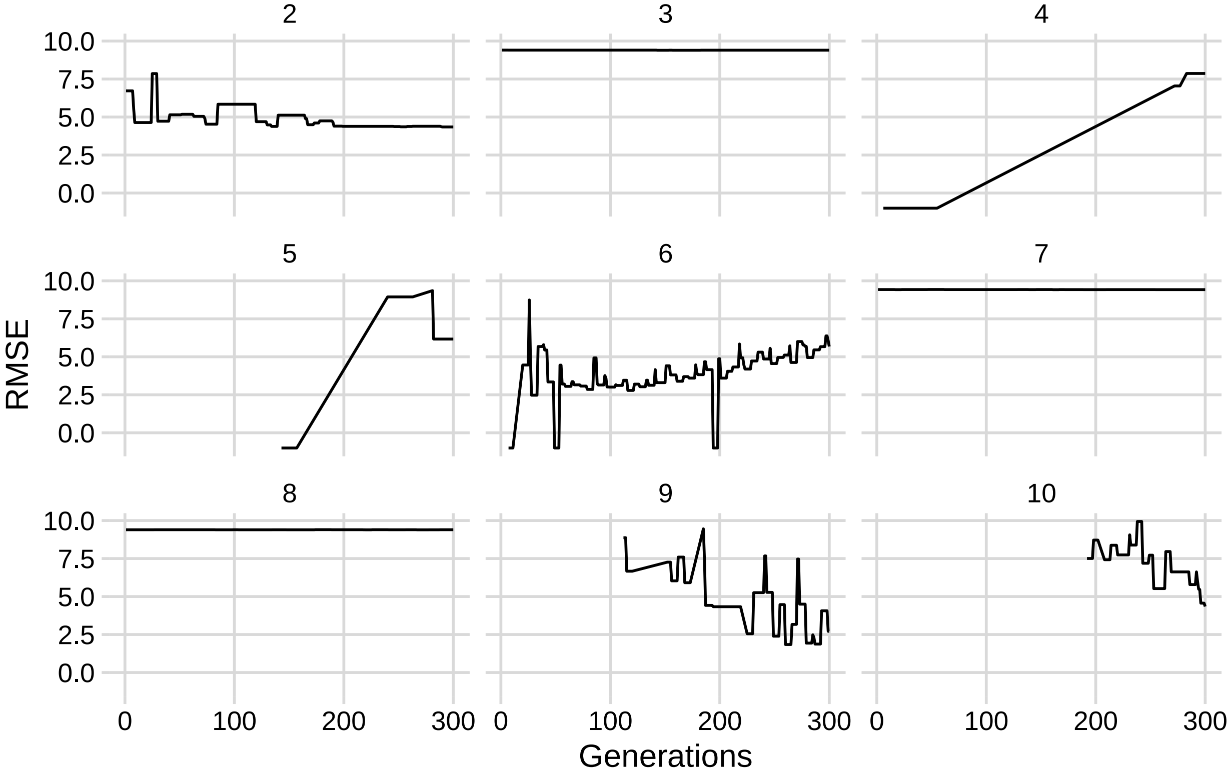
Dive into Scenario B

What do the reports for each simulation look like? For Scenario B, I want to explore the parameter space (ie promoter strengths) below what I found manually. First, I look at how the RMSE changes over 300 generations in each of the 10 simulations. Here are my observations:

1. Simulation 1 report is missing. For some reason it quit the simulation at generation XXX. Next time I'll use 'nohub' with the disown command so that it will save the output & I can trace the error.
2. Simulations 2, 4, 5, 6, 9, and 10 have the lowest errors. Simulations 3, 7, and 8 look pretty similar, with an error around 9ish. Simulations 4, 5, and 6 interestingly increased over time but still had relatively lower errors.

gen	pA	pB	pD	error	sim
300	975647.1	2.421673e+09	1.269744e+06	4.34676	2
300	511066.2	1.238675e+12	1.271191e+06	7.86485	4
300	471171.6	2.875833e+18	1.735204e+07	6.16749	5
300	868910.2	6.437813e+06	3.502768e+06	5.67668	6
300	476217.1	4.661934e+06	1.192904e+12	2.71397	9
300	686470.6	4.734017e+08	1.281719e+06	4.34402	10

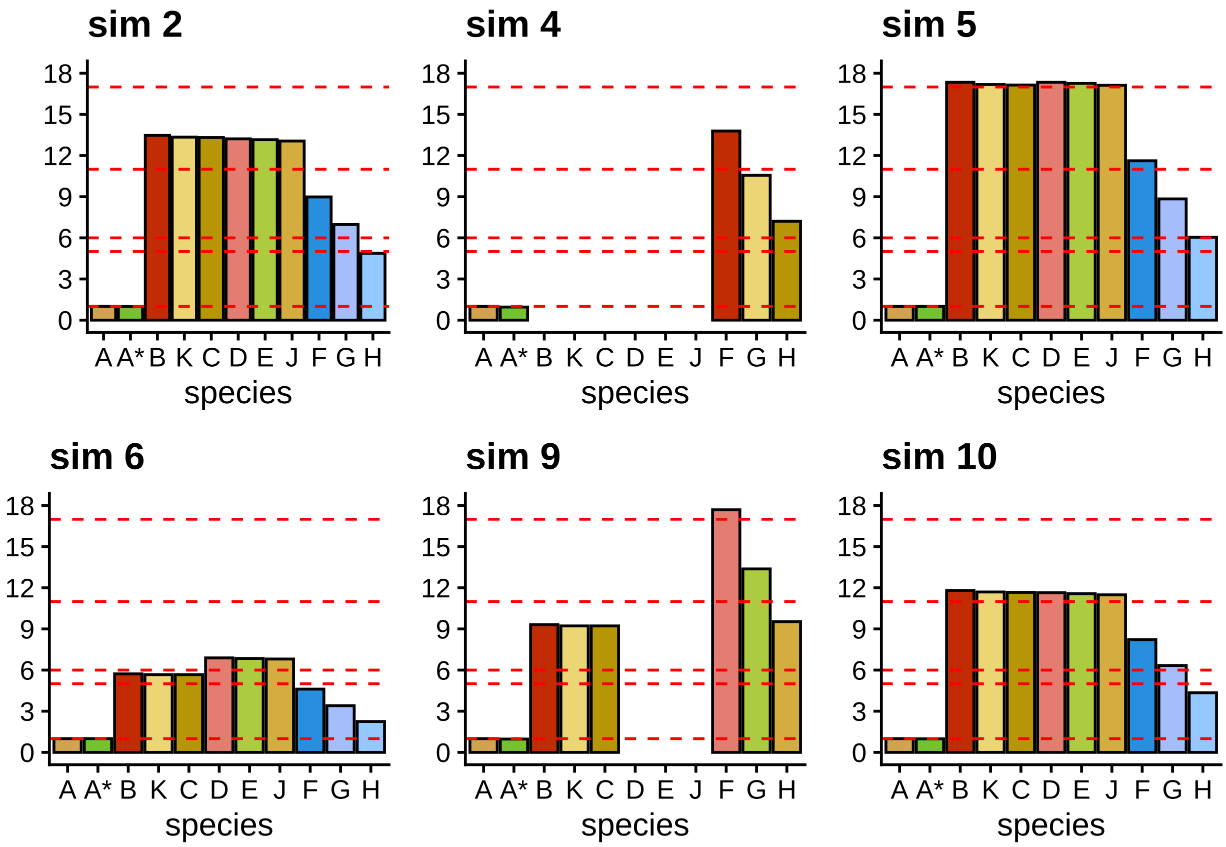
Figure 2: Scenario B

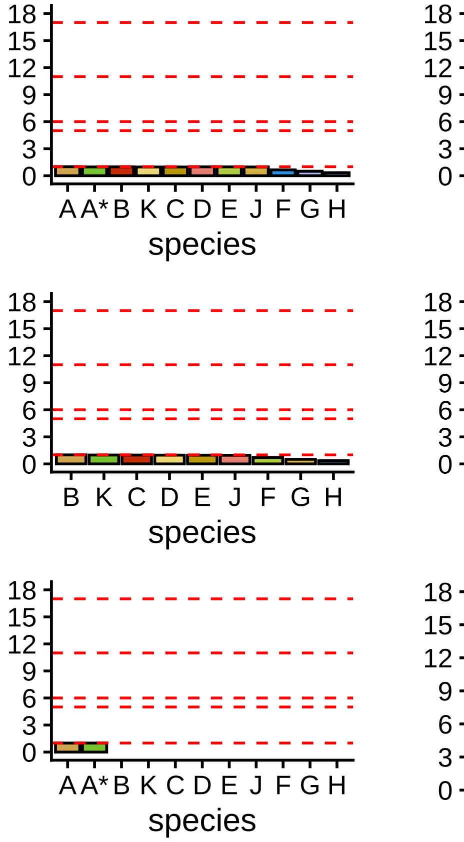


What do the distributions look like for the lowest error simulations? For Scenario B, simulations 2, 4, 5, 6, 9, and 10 have the lowest errors at the 300th generation.

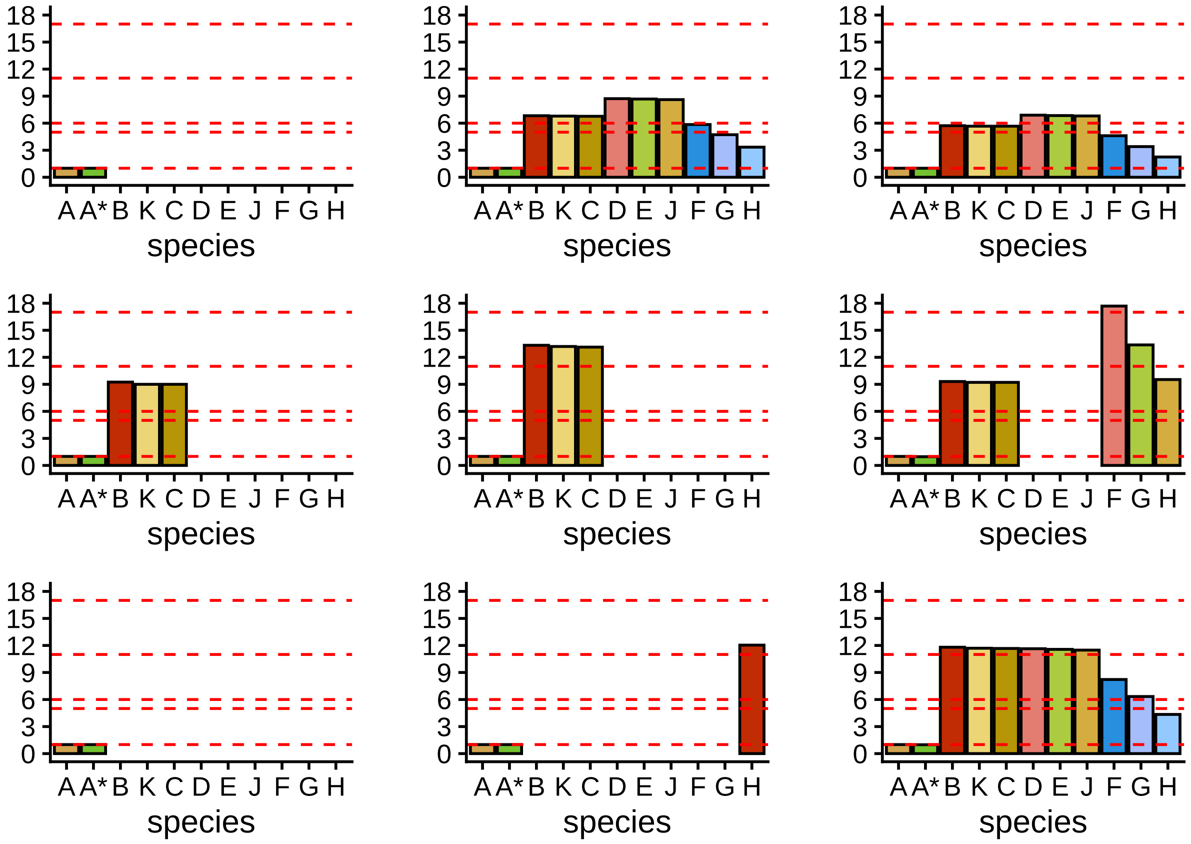
For Scenario B, the none of the distributions look particularly good – sim 6 looks the best because it captures the step between genes BKC and DEJ. Also, sims 4 and 9 still reported a low error but are missing gene expressions.

Visually & quantitatively, sim 6 is the best of the Scenario B batch.


sim 2
sim 4
sim 5
sim 6
sim 9
sim 10
0
3
6
9
12
15
18
A A* B K C D E J F G H
species



Here is how the sims 2, 4, 5, 6, 9, and 10 change over time (gen 1, gen 150, gen 300).



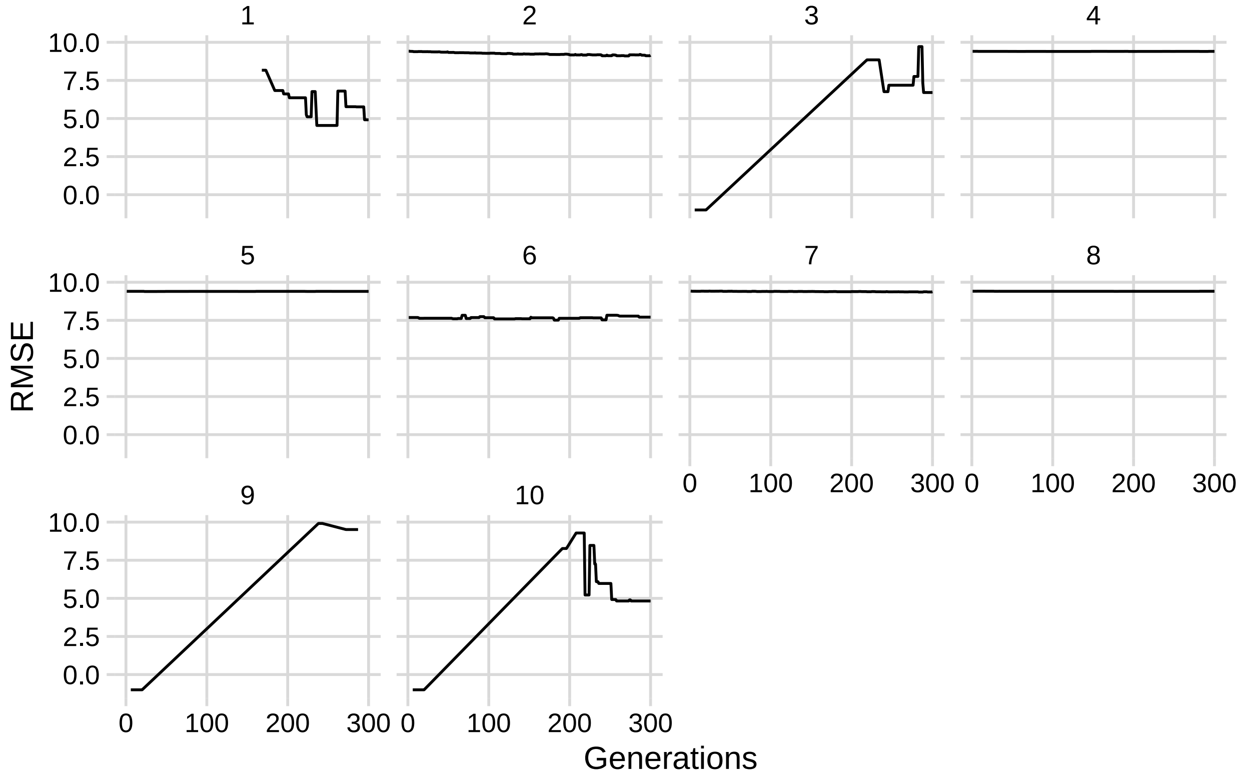
Dive into Scenario C

What do the reports for each simulation look like? For Scenario C, I want to explore the parameter space (ie promoter strengths) below what I found manually. First, I look at how the RMSE changes over 300 generations in each of the 10 simulations. Here are my observations:

1. Simulations 1, 3, 6, and 10 have the lowest errors. Simulations 3, 9, and 10 all increase in error for most generations but then decrease towards the end.

gen	pA	pB	pD	error	sim
300	669392.3	5.801758e+13	1253509	4.91734	1
300	584646.7	9.065407e+13	1491457	6.70474	3
300	4417646.2	2.647037e+13	20620091	7.70935	6
300	648827.0	1.435701e+11	1594533	4.82738	10

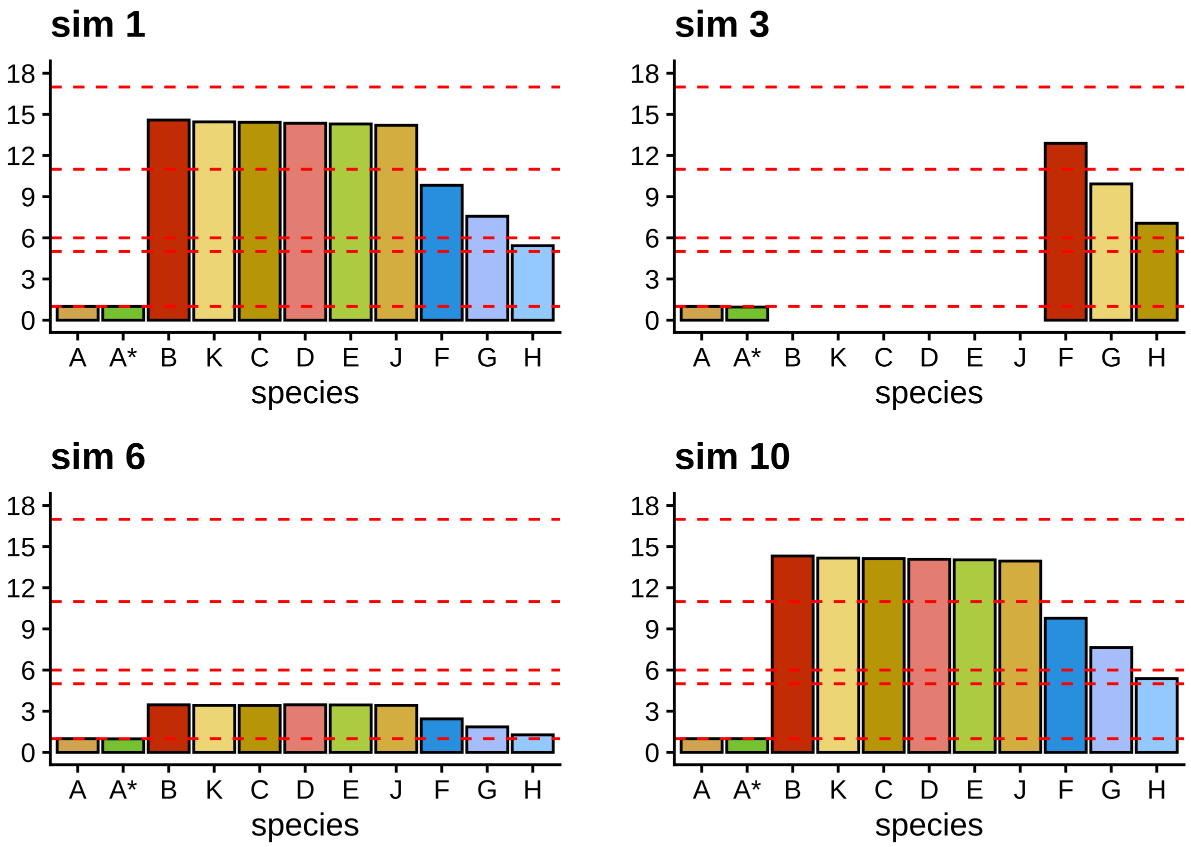
Figure 3: Scenario C

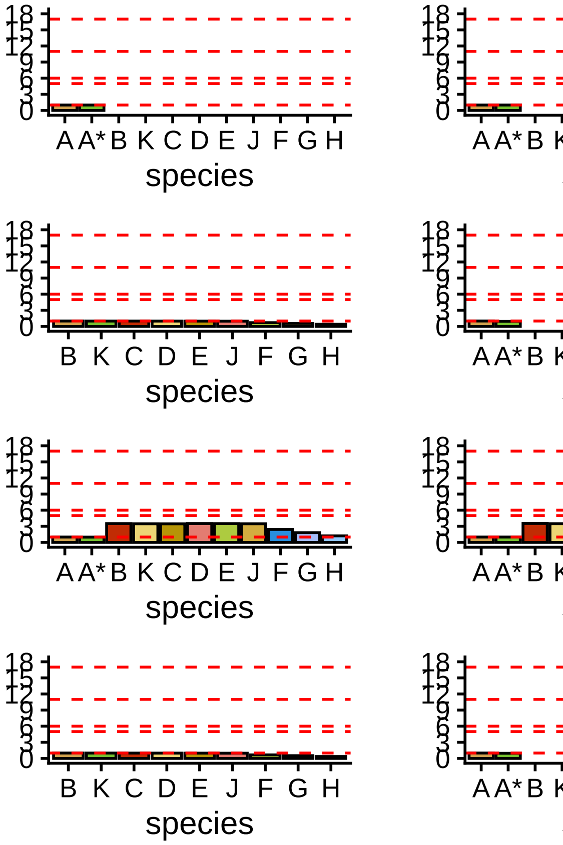


What do the distributions look like for the lowest error simulations? For Scenario C, simulations 1, 3, 6, and 10 have the lowest errors at the 300th generation.

For Scenario C, the none of the distributions look particularly good – sim 1 and sim 10 are the best, however they do not have the step whise increase in gene expression for genes BKC to DEJ. Sim 6 has too low expression levels while sim 3 is missing gene expression for genes BKCDEJ.

Visually & quantitatively, sim 1 and 10 are the best of the Scenario C batch.


sim 1
18
15
12
9
6
3
0
A
A*
B
K
C
D
E
J
F
G
H
species
sim 3
18
15
12
9
6
3
0
A
A*
B
K
C
D
E
J
F
G
H
species
sim 6
18
15
12
9
6
3
0
A
A*
B
K
C
D
E
J
F
G
H
species
sim 10
18
15
12
9
6
3
0
A
A*
B
K
C
D
E
J
F
G
H
species



Here is how the sims 1, 3, 6, 10 change over time (gen 1, gen 150, gen 300).

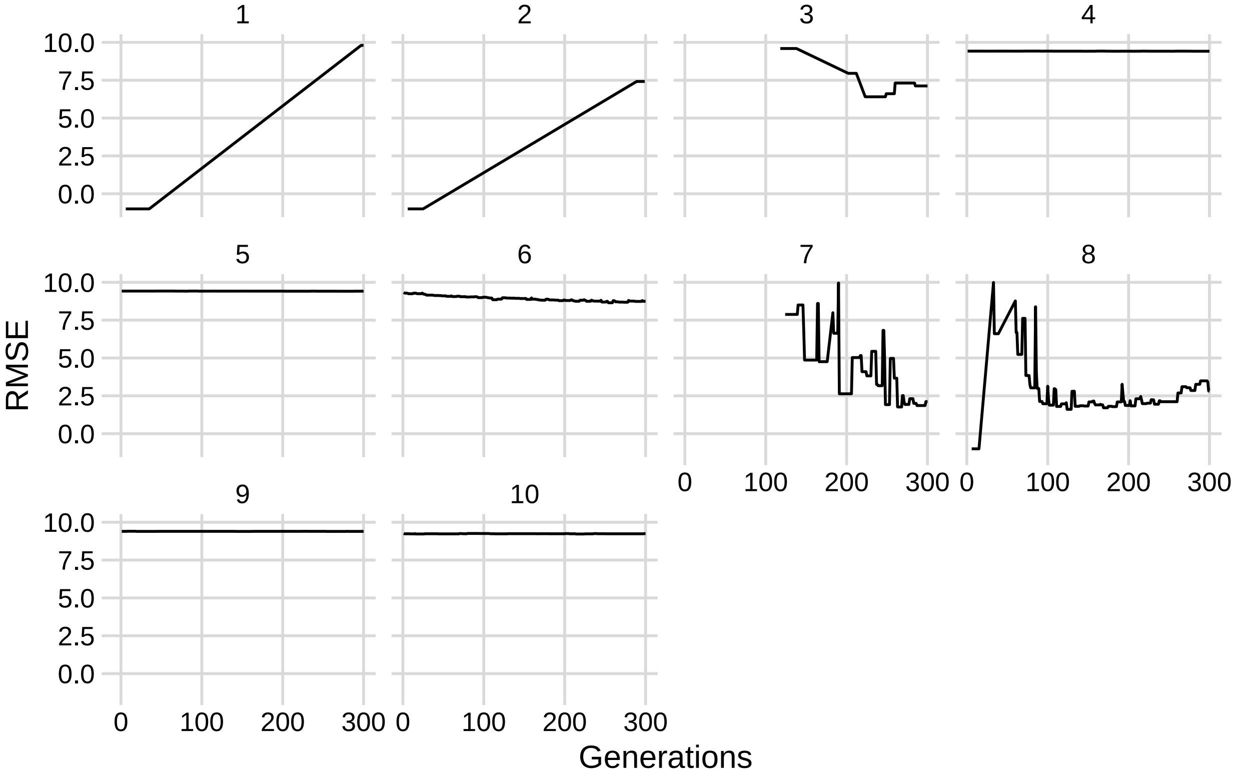
Dive into Scenario D

What do the reports for each simulation look like? For Scenario D, I want to explore the parameter space (ie promoter strengths) below what I found manually. First, I look at how the RMSE changes over 300 generations in each of the 10 simulations. Here are my observations:

1. Simulations 3, 7, and 8 have the lowest errors. Simulations 1 and 2 increase in error but never decrease. Scenarios 4, 5, 6, 9, and 10 have an RSME of 9ish.

gen	pA	pB	pD	error	sim
300	1489920806	4038221	1540743	9.41526	4
300	293305450	4774980	1370299	9.41426	5
300	3911531	6430642	2281340	8.74533	6
300	6965340817	3319093	337441741	9.39487	9
300	34297220	1681505563	108097765	9.24606	10

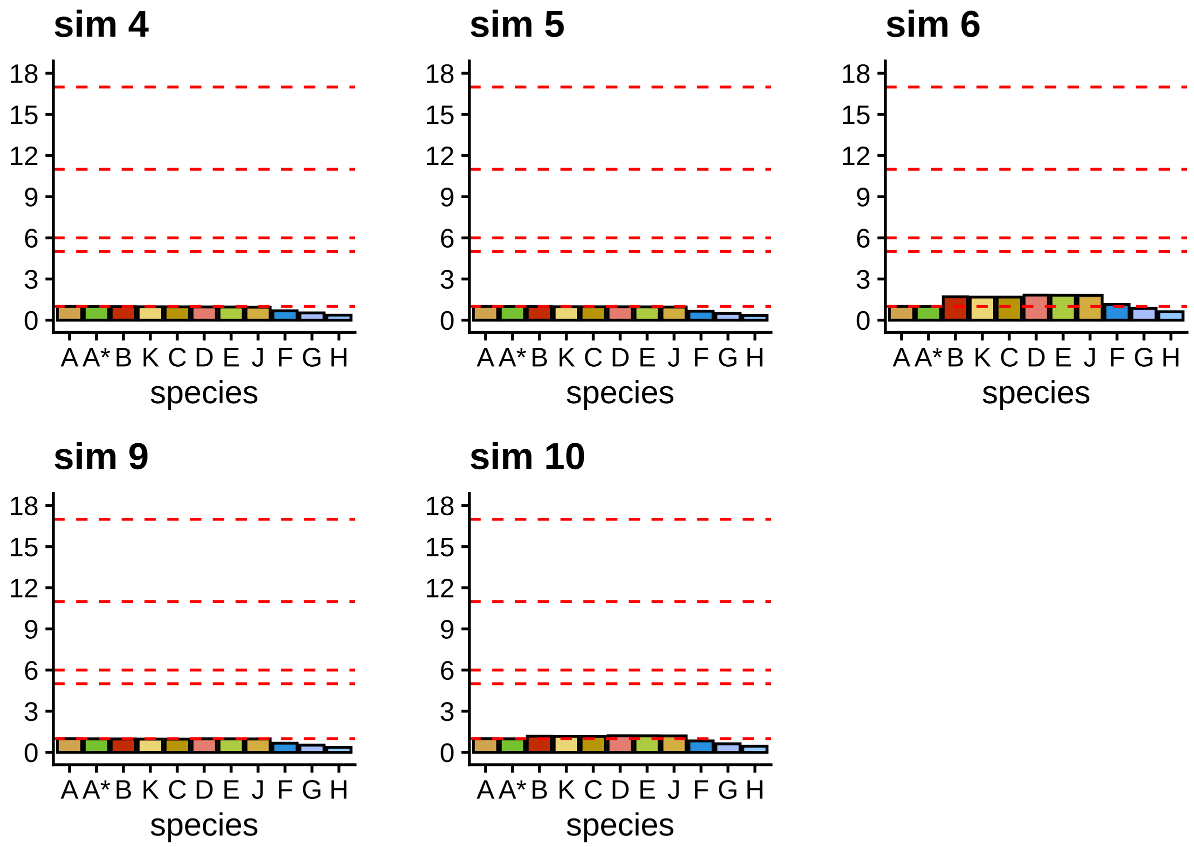
Figure 4: Scenario D

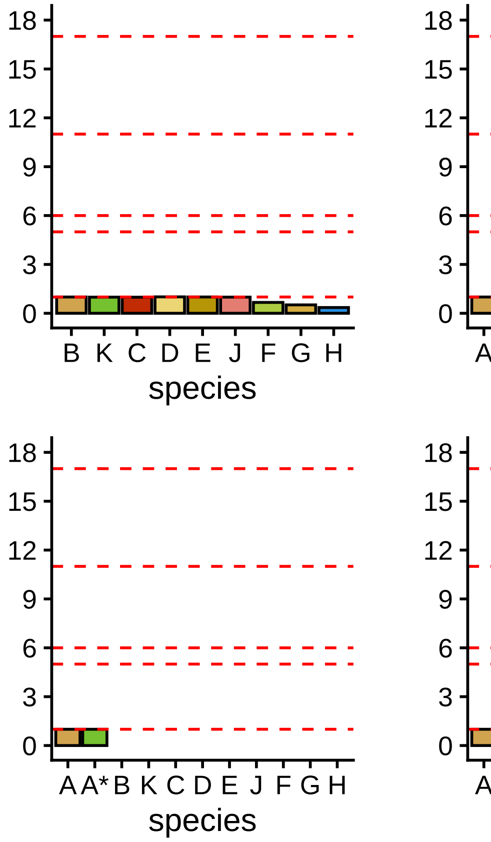


What do the distributions look like for the lowest error simulations? For Scenario D, simulations 4, 5, 6, 9, and 10 have the lowest errors at the 300th generation.

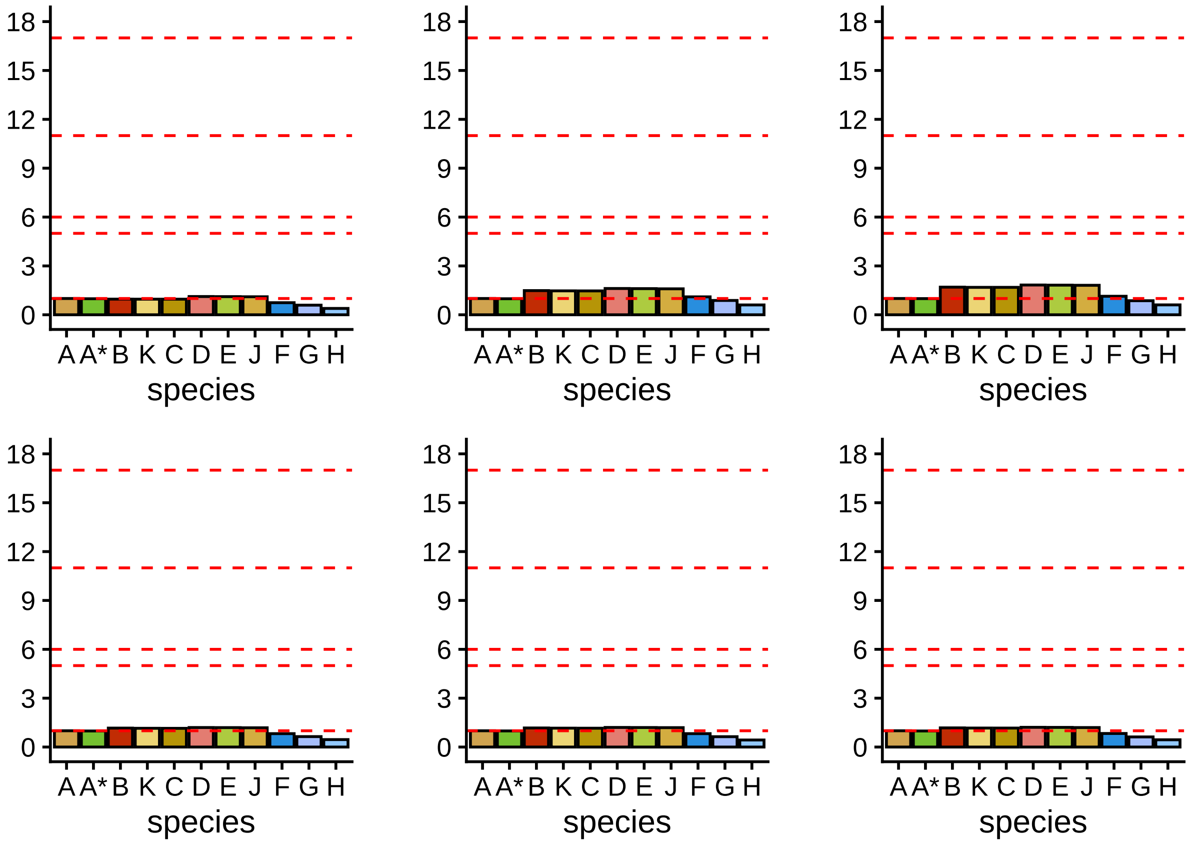
For Scenario D, the none of the distributions look particularly good – at least all of the simulations had all genes expressed. **Why could increasing promoter values result in low transcript abundance levels?**

Visually & quantitatively, none are particularly good.


sim 4
18
15
12
9
6
3
0
A A* B K C D E J F G H
species
sim 5
18
15
12
9
6
3
0
A A* B K C D E J F G H
species
sim 6
18
15
12
9
6
3
0
A A* B K C D E J F G H
species
sim 9
18
15
12
9
6
3
0
A A* B K C D E J F G H
species
sim 10
18
15
12
9
6
3
0
A A* B K C D E J F G H
species



Here is how the sims 4, 5, 6, 9, and 10 change over time (gen 1, gen 150, gen 300).



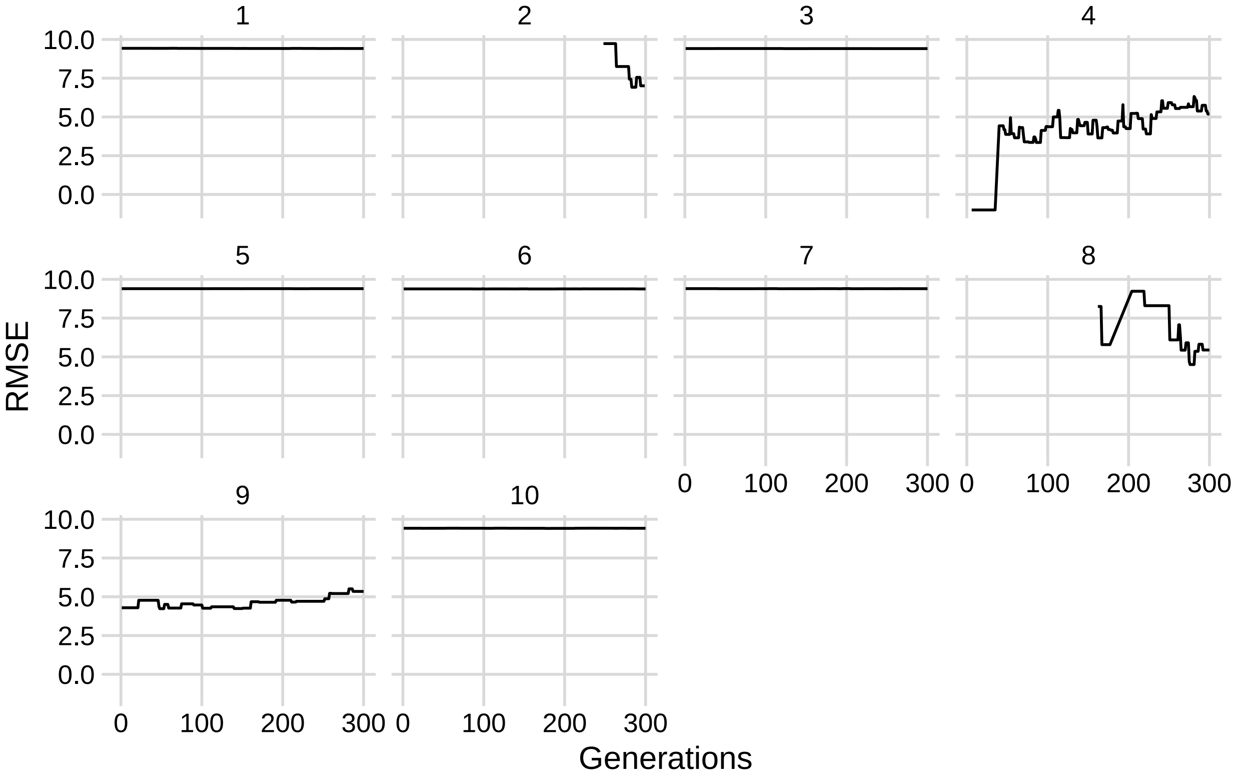
Dive into Scenario E

What do the reports for each simulation look like? For Scenario E, I want to explore the parameter space (ie promoter strengths) below what I found manually. First, I look at how the RMSE changes over 300 generations in each of the 10 simulations. Here are my observations:

1. Simulations 2, 4, 8, and 9 have the lowest error. The others all converge around an RMSE of 9ish.

gen	pA	pB	pD	error	sim
300	473477.7	696865525786	1108670	11.07673	2
300	660137.1	4329271	1531706	5.18862	4
300	738001.9	59627427238	1342909	5.44287	8
300	1643886.0	2990155571	171398511761	5.34631	9

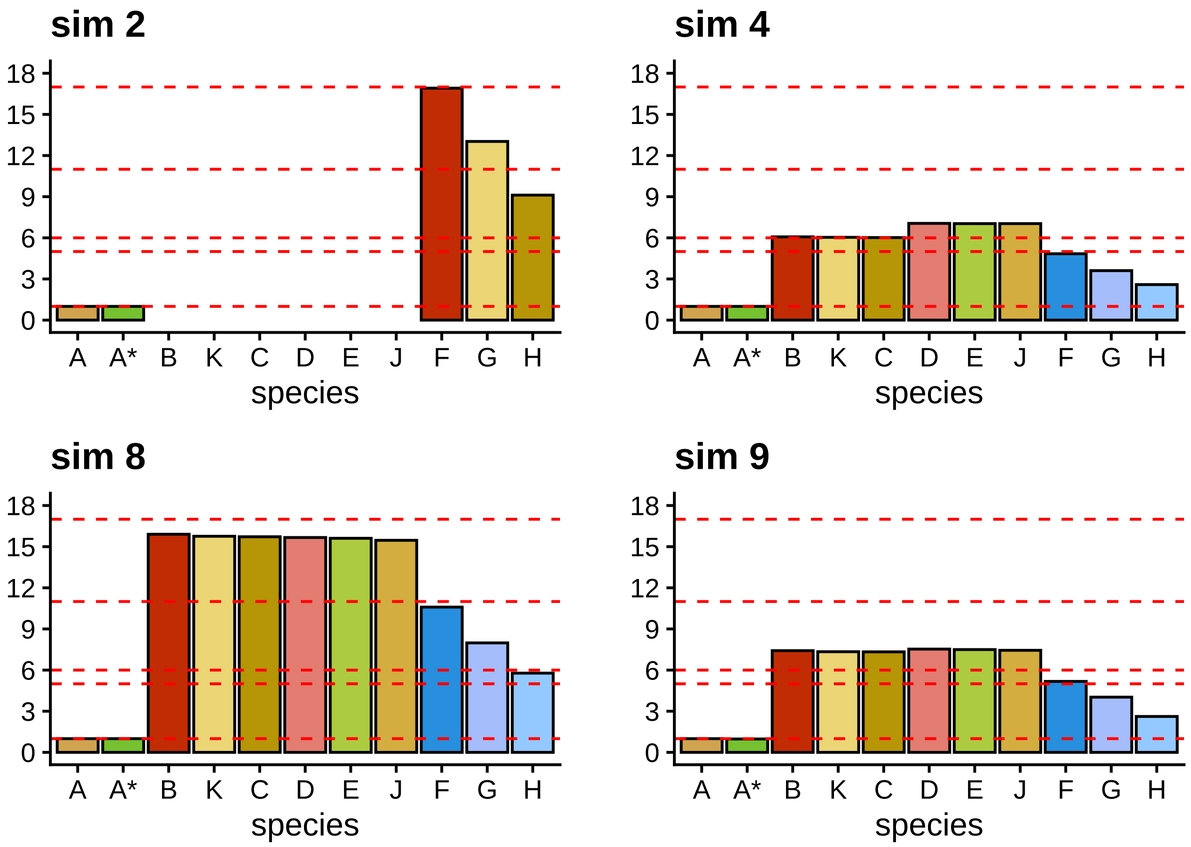
Figure 5: Scenario E

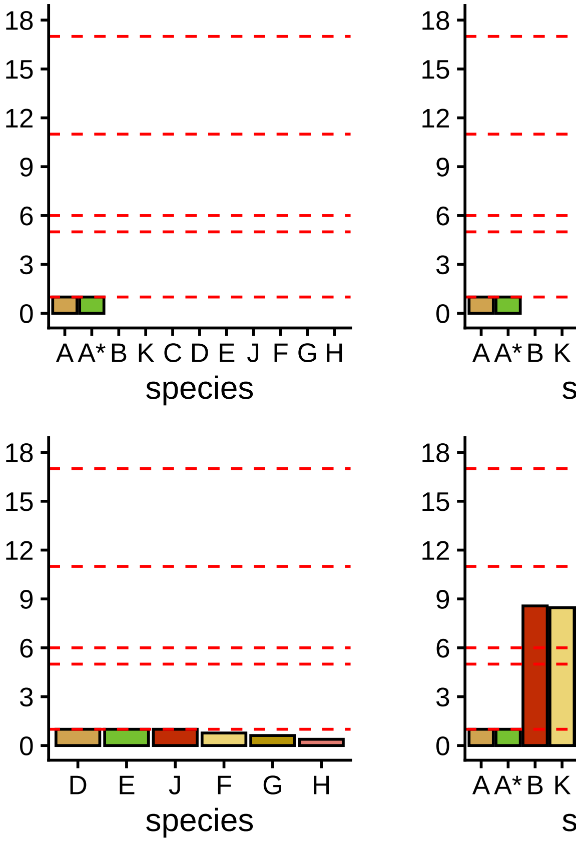


What do the distributions look like for the lowest error simulations? For Scenario E, simulations 2, 4, 8, and 9 have the lowest errors at the 300th generation.

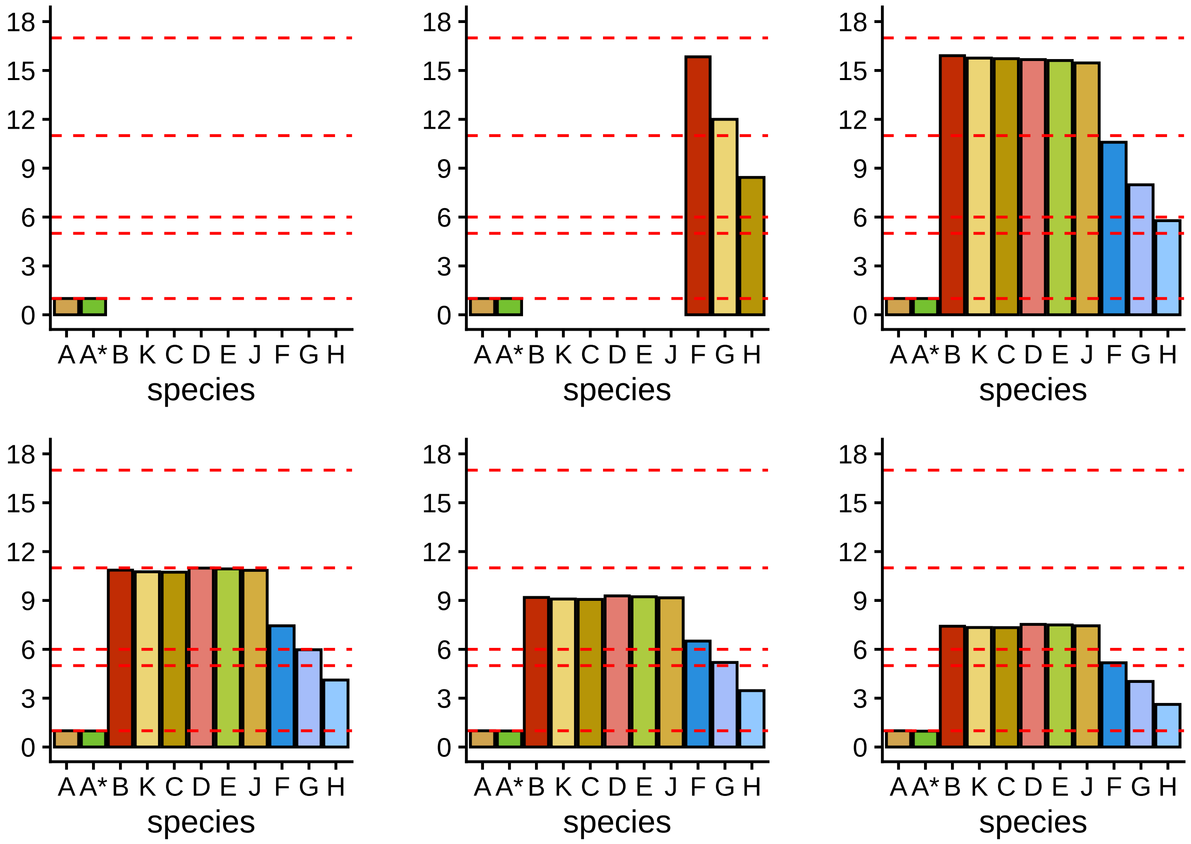
For Scenario E, simulation 4 looks good because it captures the step wise increase from BKC to DEJ, yet the abundances for DEJ aren't high enough. Simulation 8 also looks good since it is high enough, however there is no step wise increase. Simulation 2 is missing expression in genes BKCDEJ and simulation 9 is meh.

Visually & quantitatively, simulations 4 and 8 are particularly good in Scenario E.


sim 2
sim 4
sim 8
sim 9
0
3
6
9
12
15
18
A
A*
B
K
C
D
E
J
F
G
H
species



Here is how the sims 2, 4, 8, 9 change over time (gen 1, gen 150, gen 300).


18
15
12
9
6
3
0
A
A*
B
K
C
D
E
J
F
G
H
species
18
15
12
9
6
3
0
A
A*
B
K
C
D
E
J
F
G
H
species
18
15
12
9
6
3
0
A
A*
B
K
C
D
E
J
F
G
H
species
18
15
12
9
6
3
0
A
A*
B
K
C
D
E
J
F
G
H
species
18
15
12
9
6
3
0
A
A*
B
K
C
D
E
J
F
G
H
species
18
15
12
9
6
3
0
A
A*
B
K
C
D
E
J
F
G
H
species