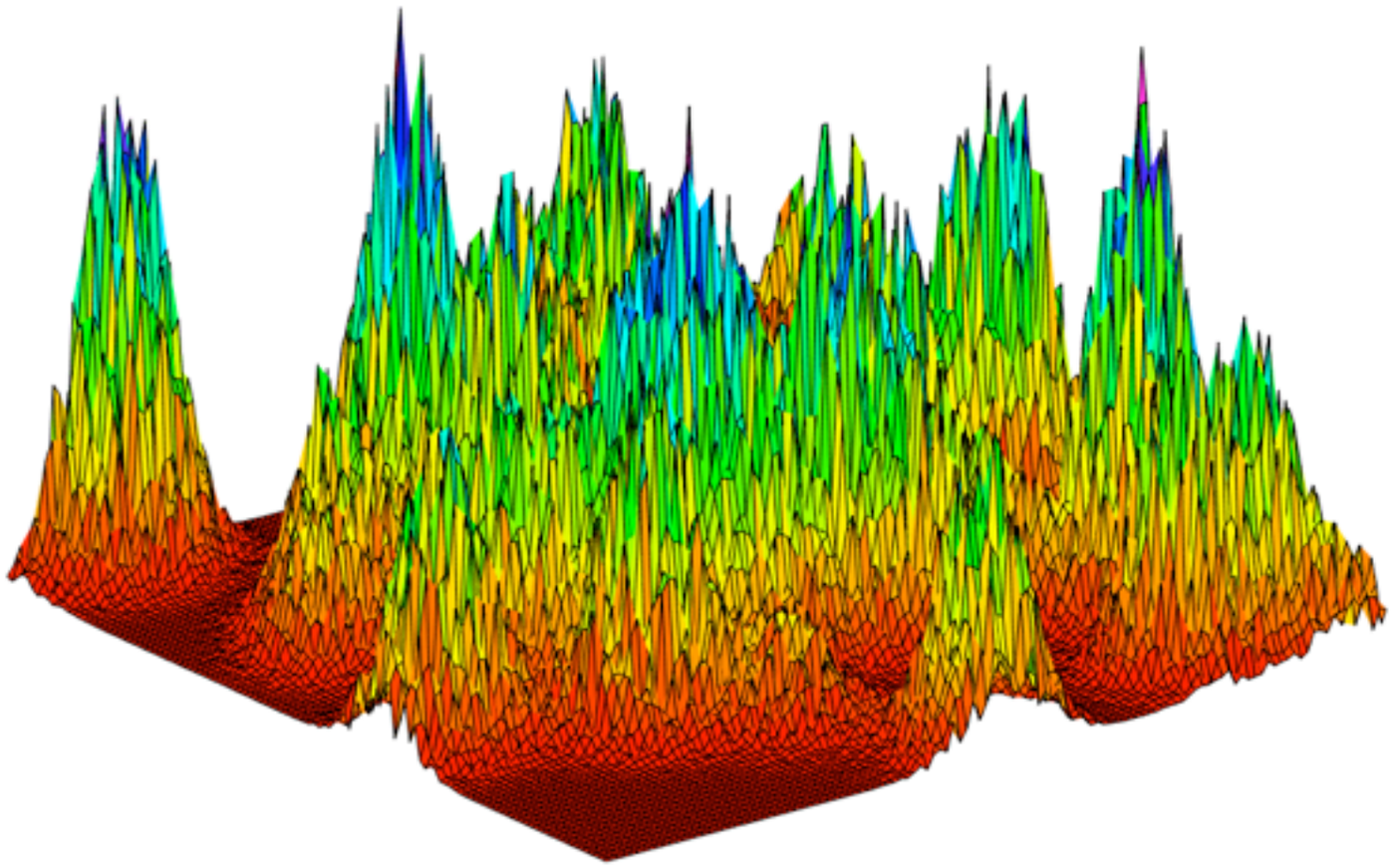
Exploring Protein Functional Landscapes using Multiplexed Gene Synthesis and Characterization



Calin Plesa

August 30, 2017

Kosuri Lab @ UCLA

Exploring Fitness Landscapes at Multiple Scales

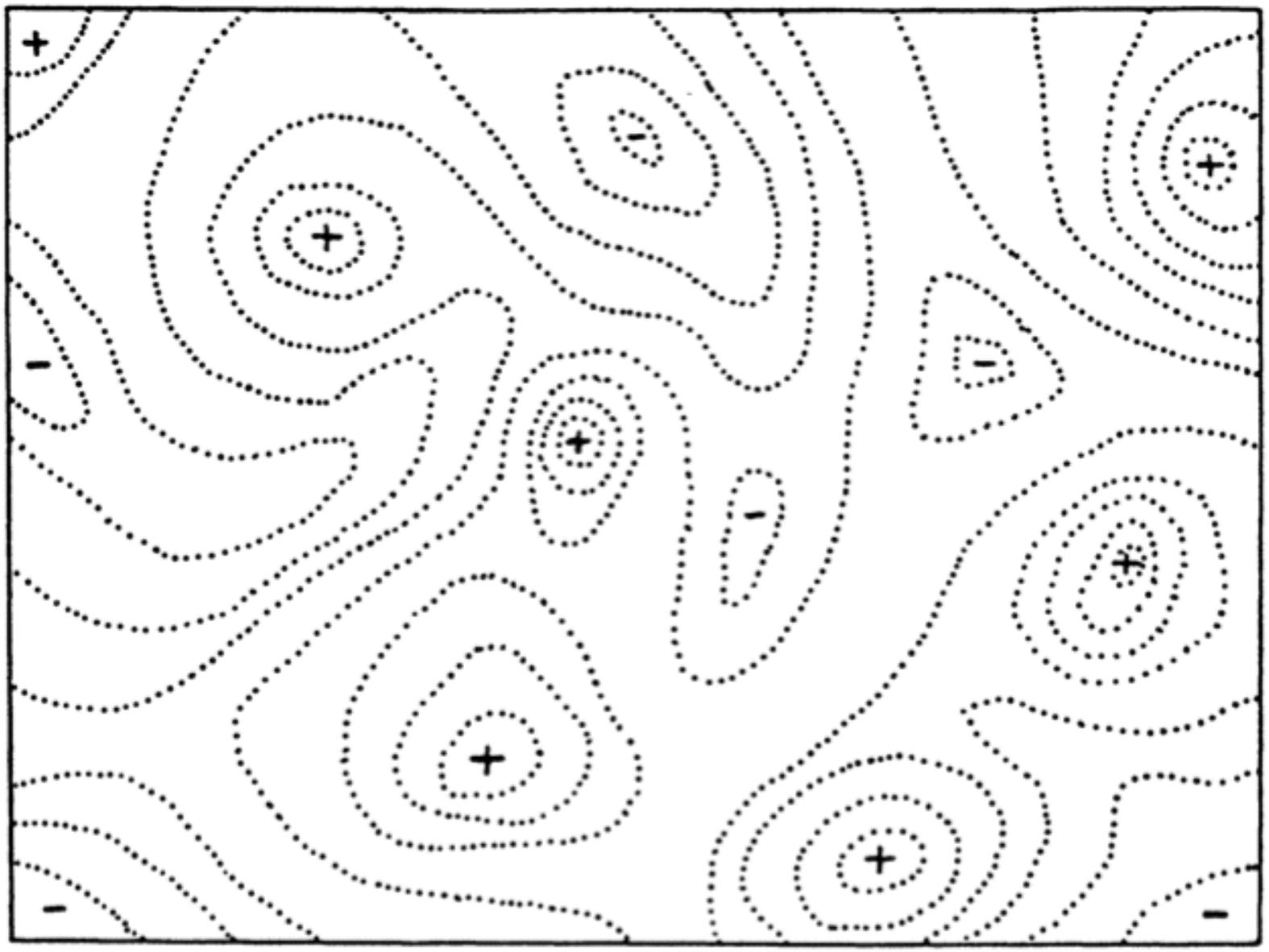


FIG. 2.—Diagrammatic representation of the field of gene combinations in two dimensions instead of many thousands. Dotted lines represent contours with respect to adaptiveness.

Wright 1932

Sequence space

Atoms in universe
10^{82}

<

Potential sequences of 64 aa protein
20^{64}

Median Bacterial protein length
267 a.a.

Median H. sapiens protein length
361 a.a.

Sequence space
20^{361}

Sequence space

Function space

Function space

All sequences capable of carrying out some function

Function space

Evolutionary space

Evolutionary space

The sequence space that's been explored by nature.

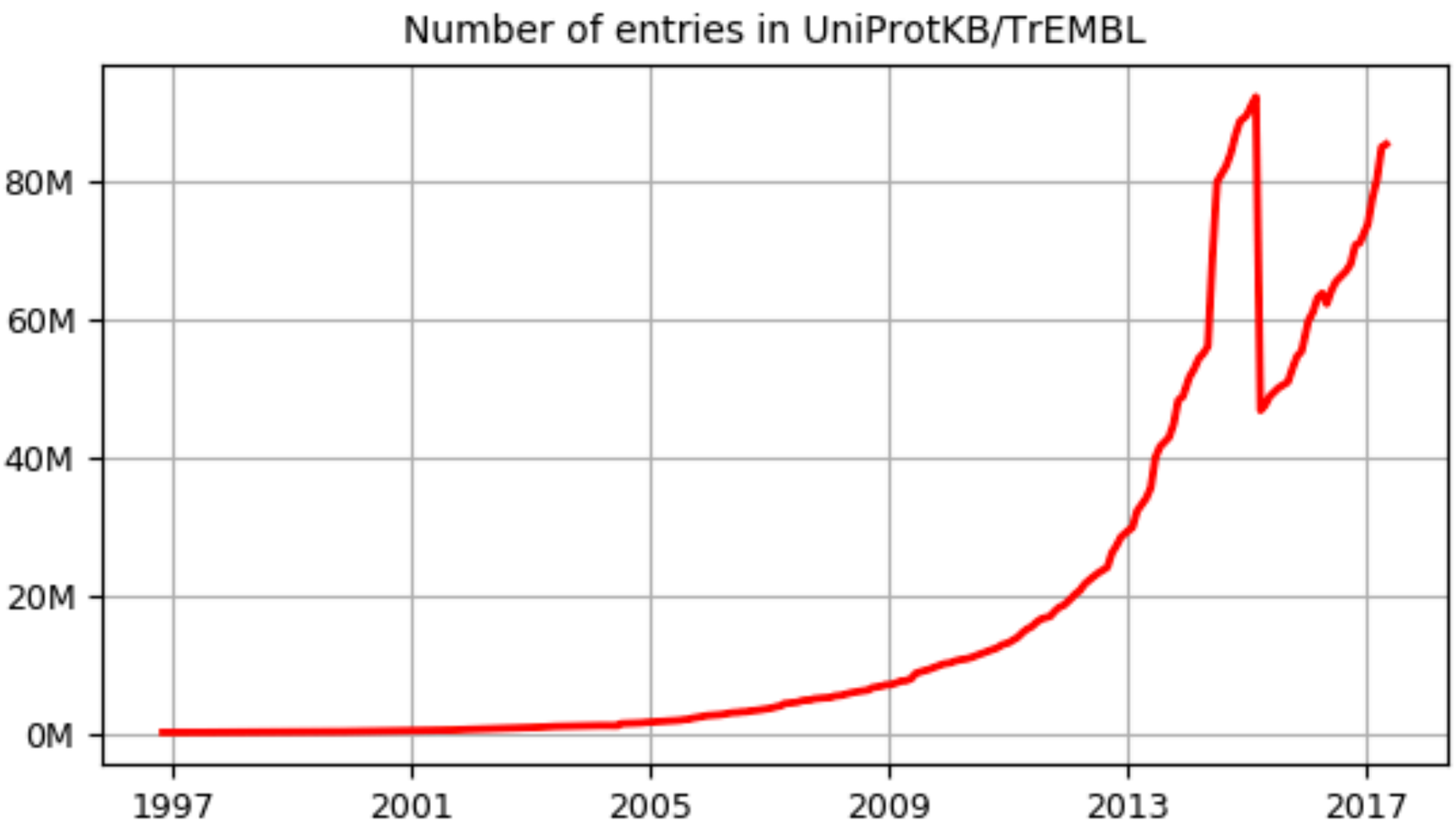
Evolutionary space

Observed space

Observed space

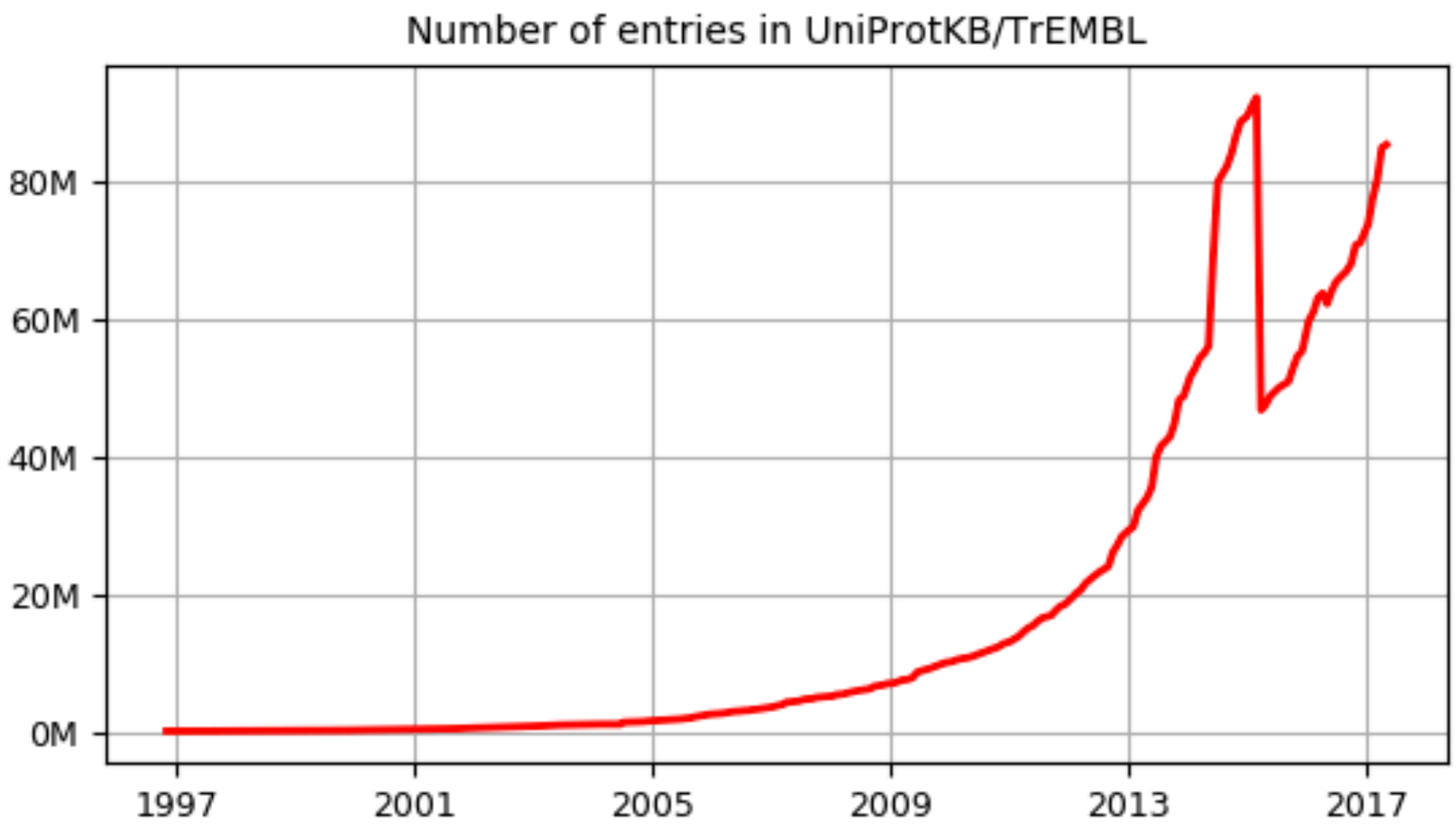
The sequence space we know about due to metagenomic sequencing.

Observed sequence space is growing exponentially



85,272,789 total sequences

Corresponding functional info is not growing exponentially

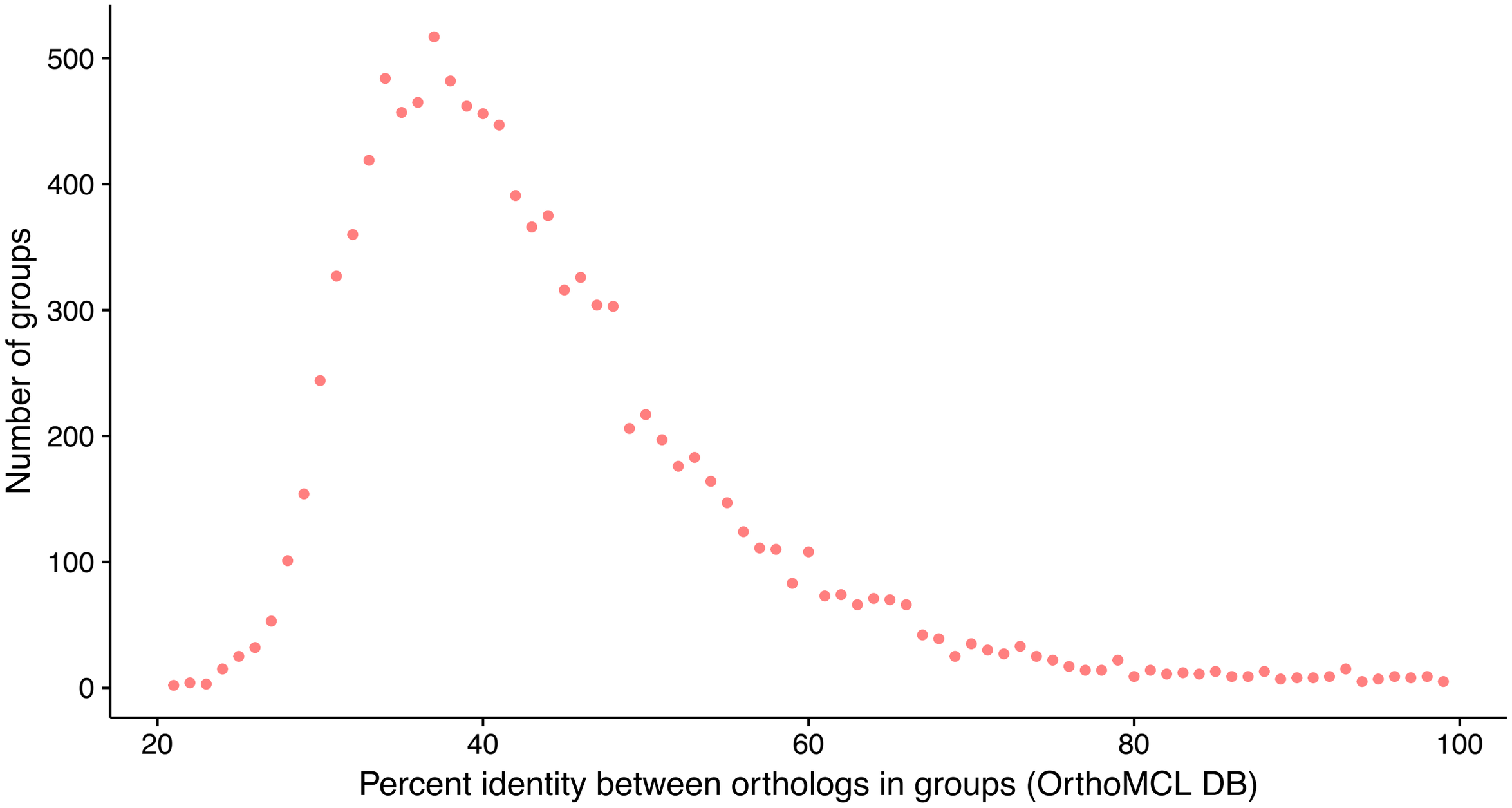


85,272,789 total sequences
693,956 with GO Experimental annotation (0.8%)

Gene Ontology Consortium 2017

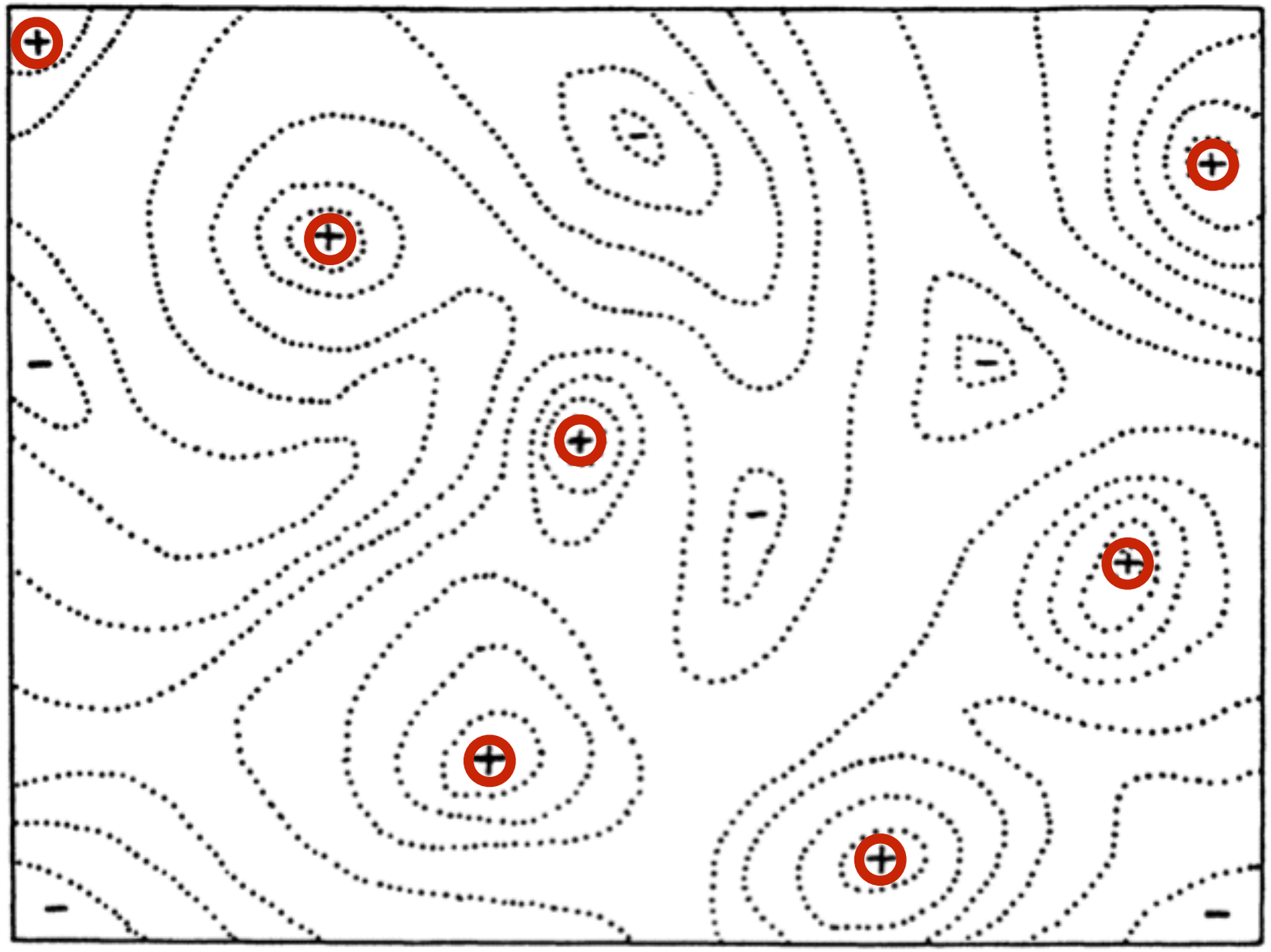
Protein existence (PE):	entries	%
1: Evidence at protein level	128838	0.15%
2: Evidence at transcript level	1082426	1 .27%
3: Inferred from homology	20603690	24.16%
4: Predicted	63457835	74.42%

Sequences of identical function are highly divergent



10,672 total groups with at least 20 seqs.

Exploring Functional Space



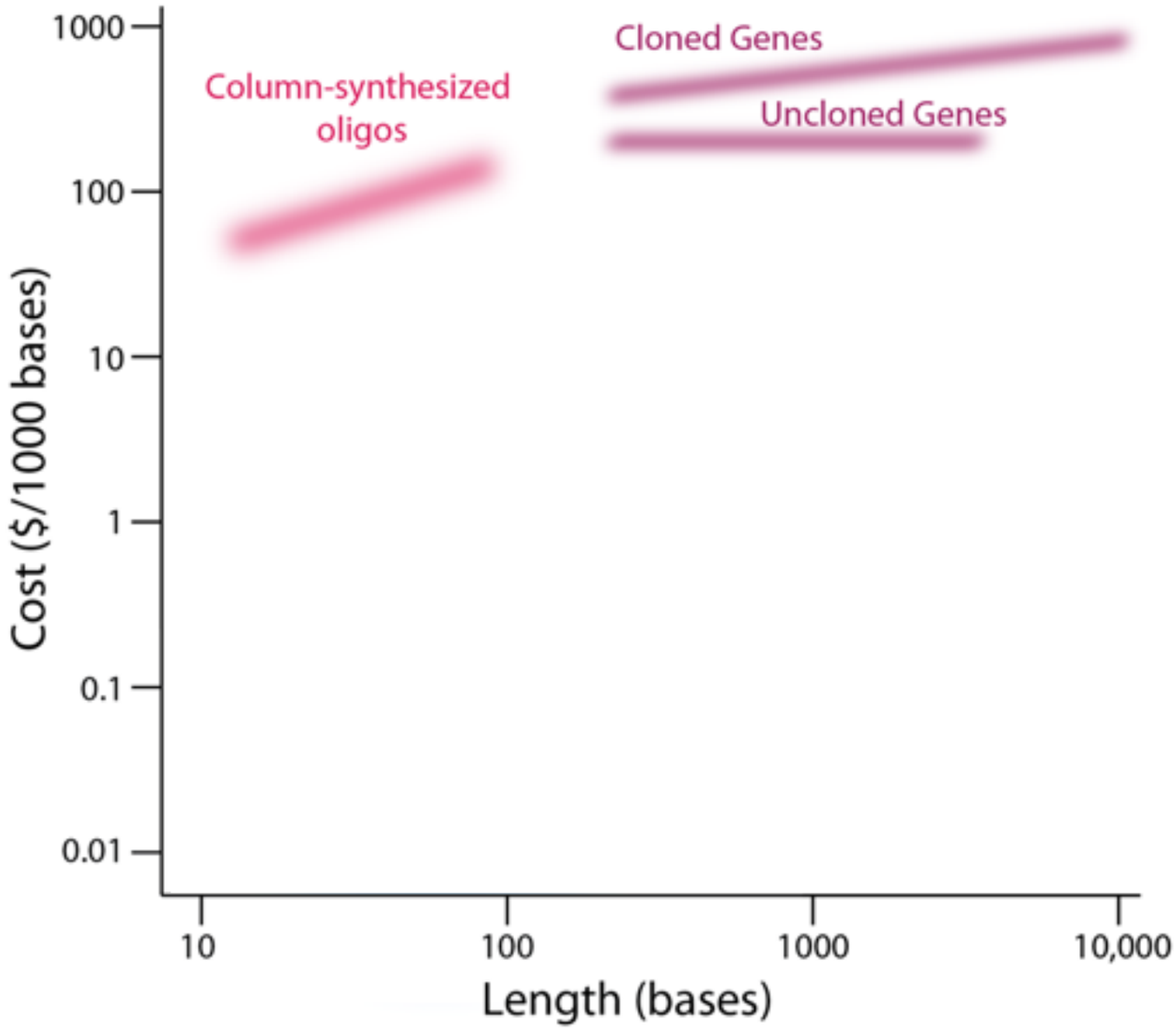
Wright 1932

FIG. 2.—Diagrammatic representation of the field of gene combinations in two dimensions instead of many thousands. Dotted lines represent contours with respect to adaptiveness.

Use homologs as an educated guess for fitness peaks at large distances

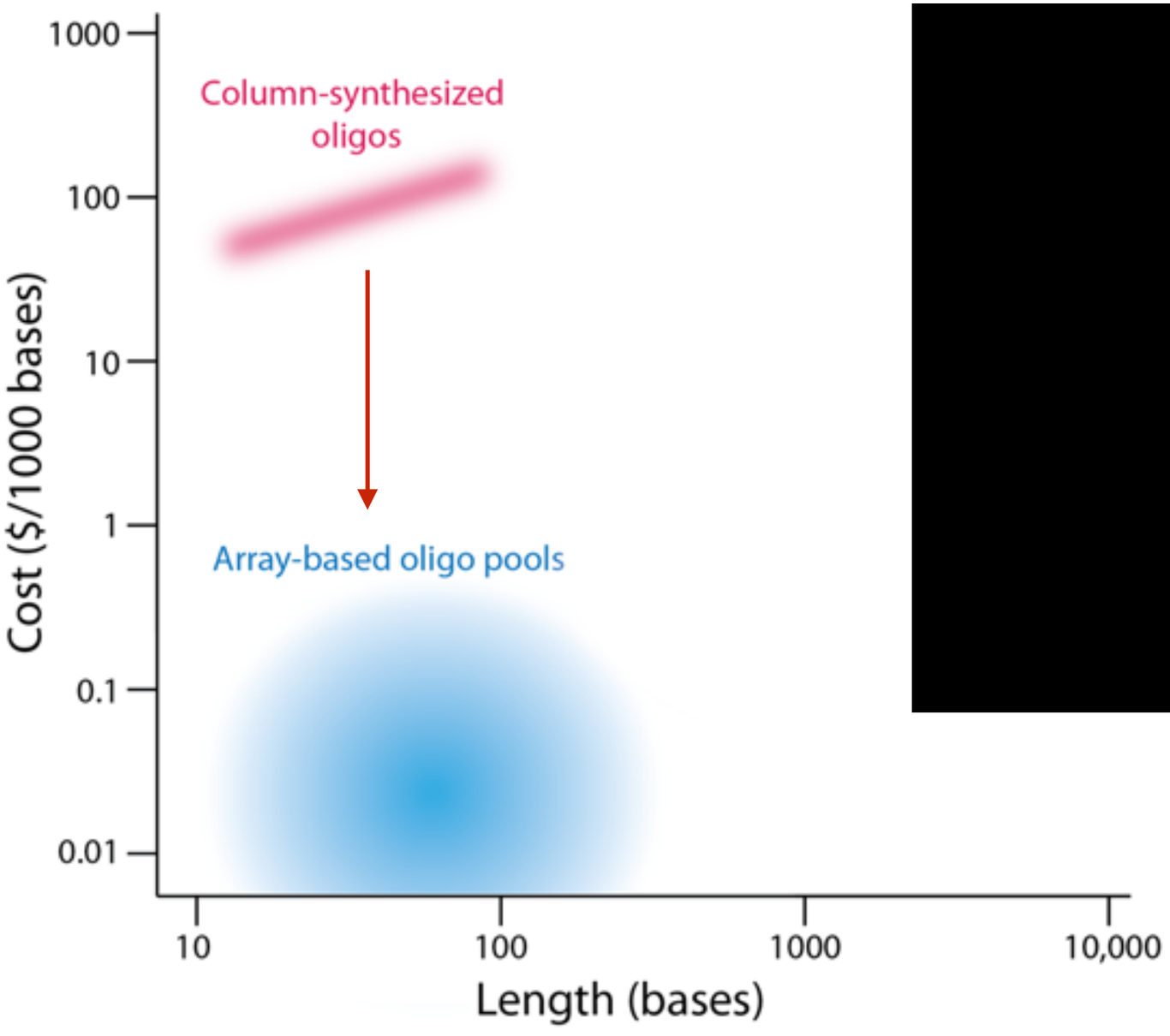
How do we get the physical DNA sequence needed for testing?

- PCR impractical - most source organisms inaccessible
- Gene synthesis - too expensive beyond a few dozen
- Oligo synthesis - short lengths



Kosuri & Church. Nature Methods 11, 499–507 (2014)

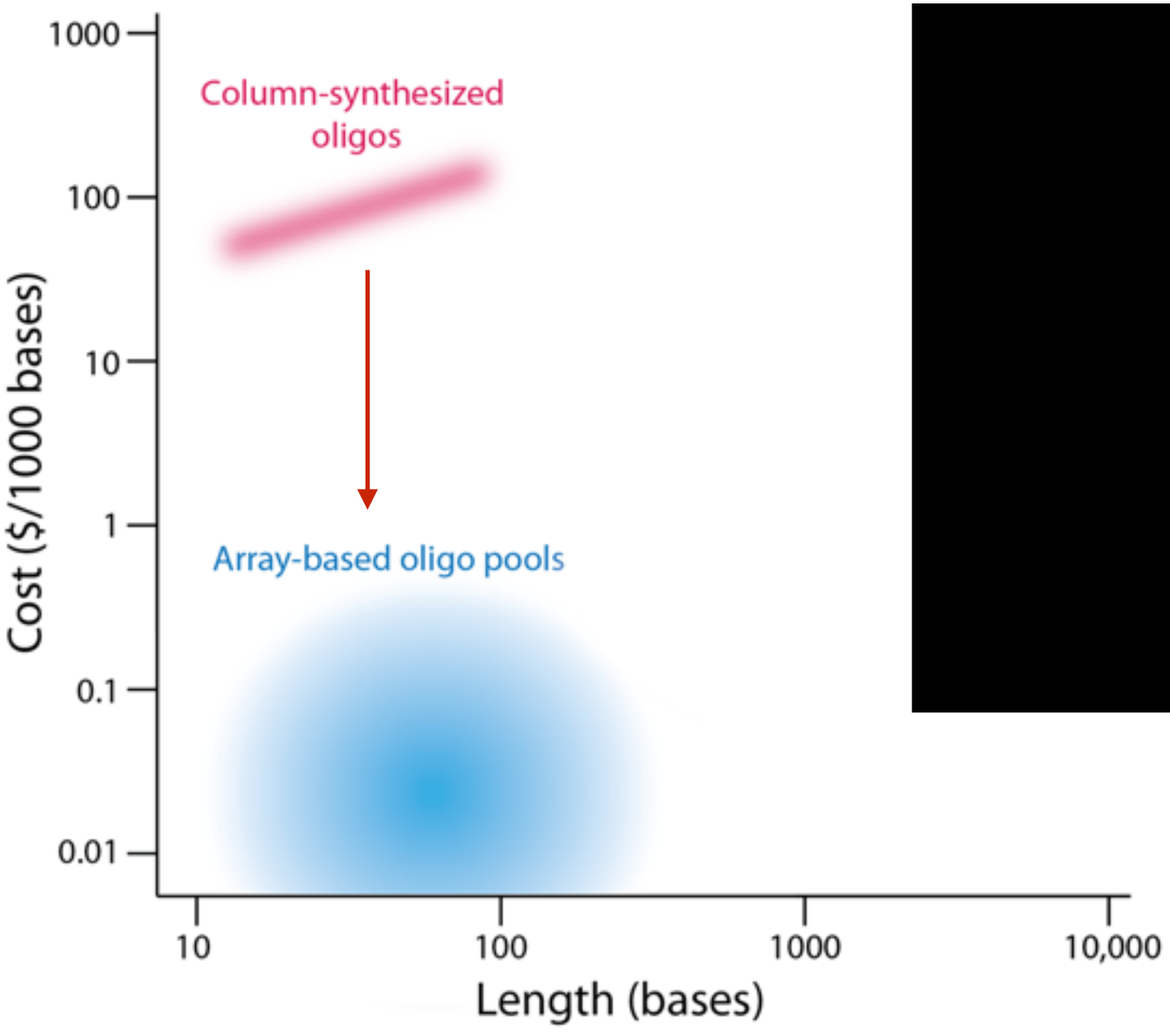
Reducing gene synthesis costs



Agilent

Kosuri & Church. Nature Methods 11, 499–507 (2014)

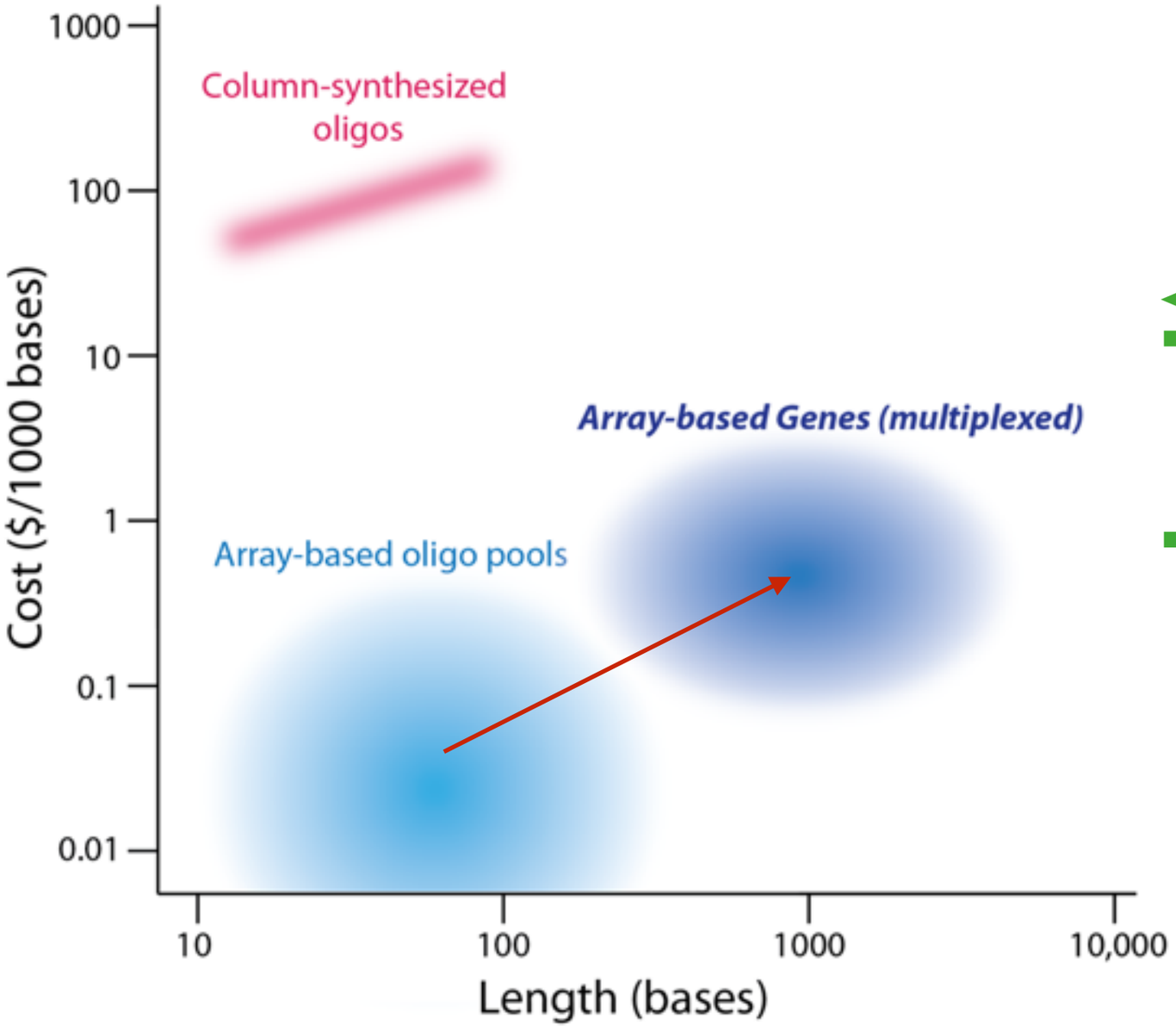
Reducing gene synthesis costs

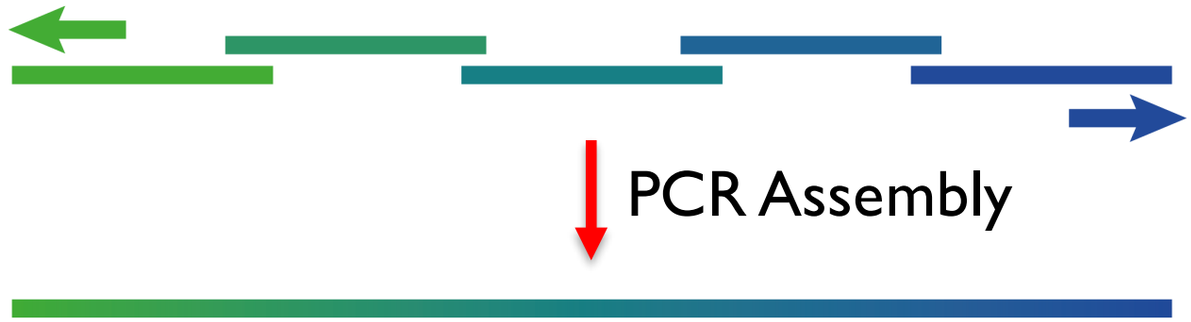


Agilent

Kosuri & Church. Nature Methods 11, 499–507 (2014)

Reducing gene synthesis costs

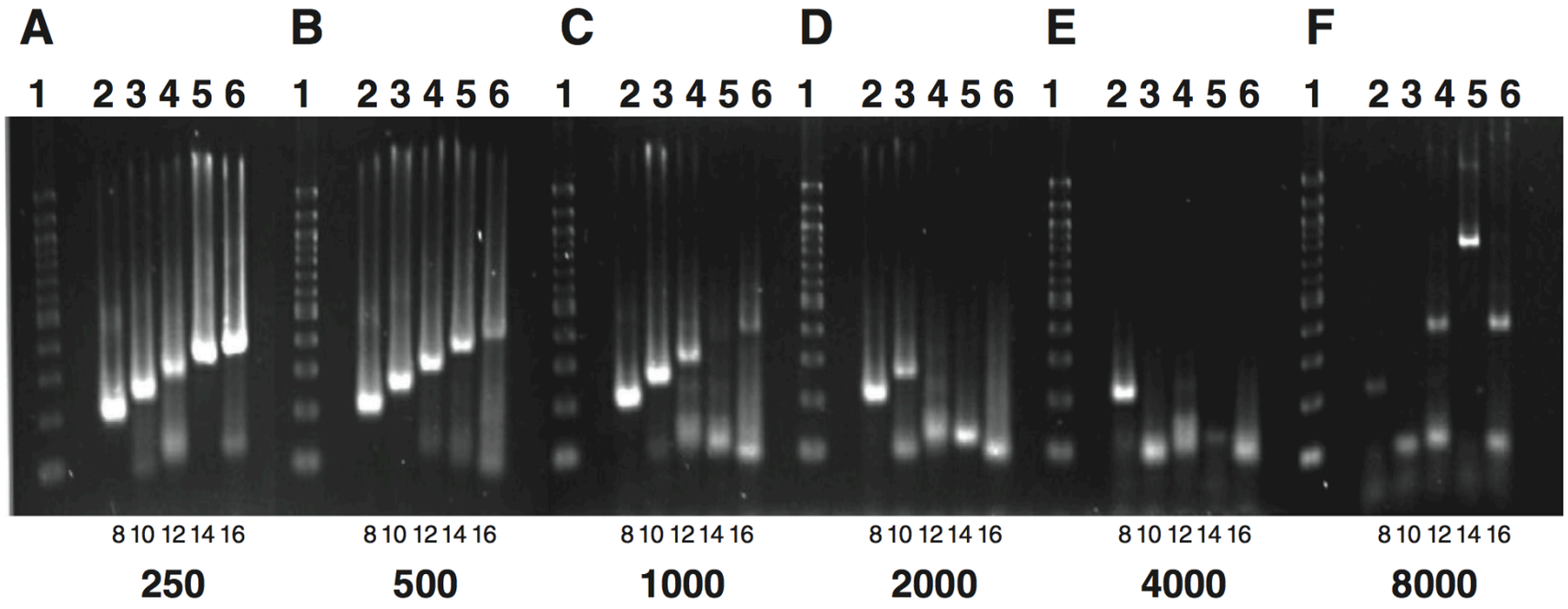




Kosuri & Church. Nature Methods 11, 499–507 (2014)

Pool based gene synthesis

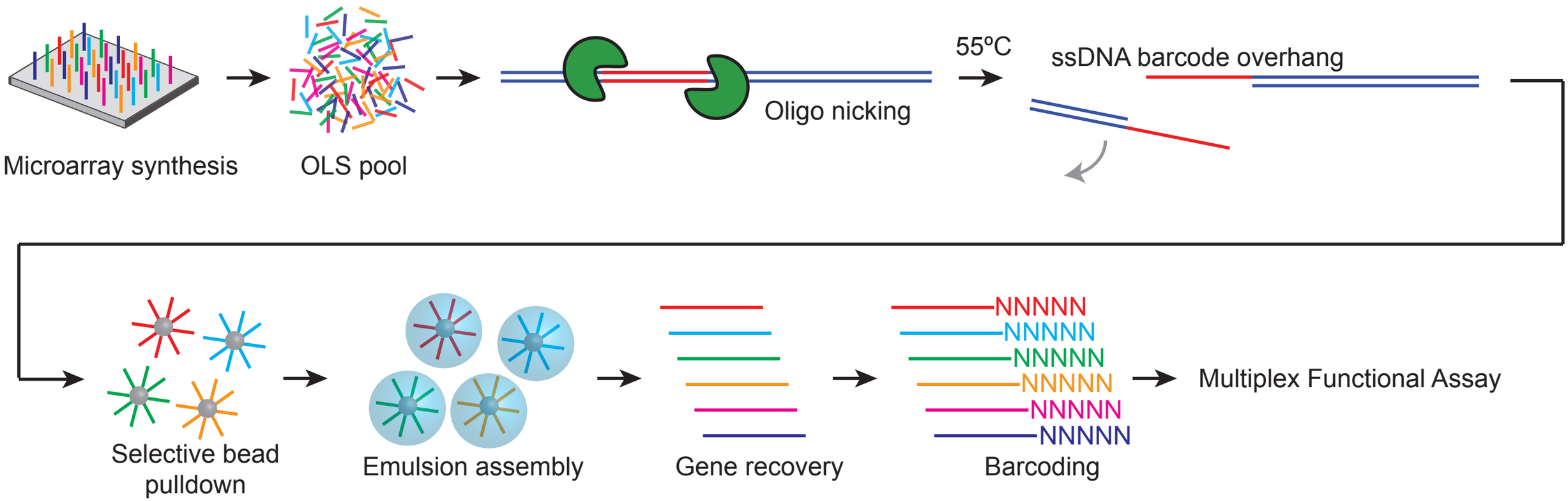
OLS Pool



- Higher background complexity leads to:
  - Larger search space
  - Higher probability of off-target hybridization
- Higher error rates (-> local landscape)

Borovkov et al. Nucleic Acids Res. Oct; 38(19): e180. 2010

DropSynth



Cost <$2 per gene

Oligo design

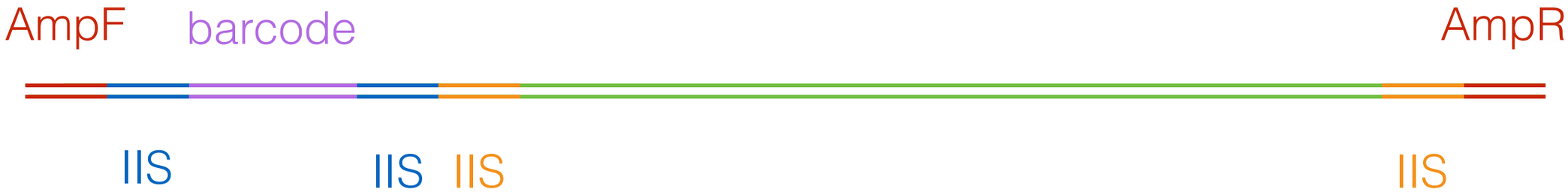
Oligo design

AmpF

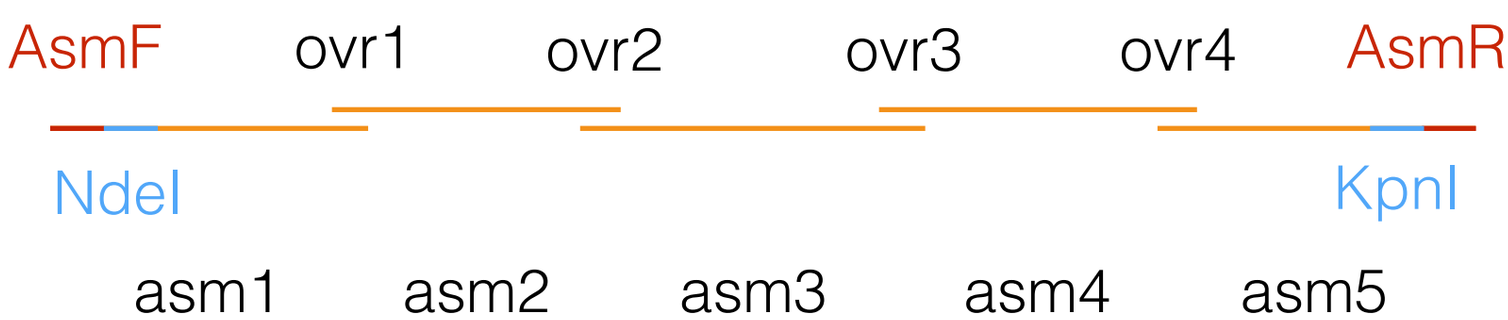
AmpR

gene

Oligo design

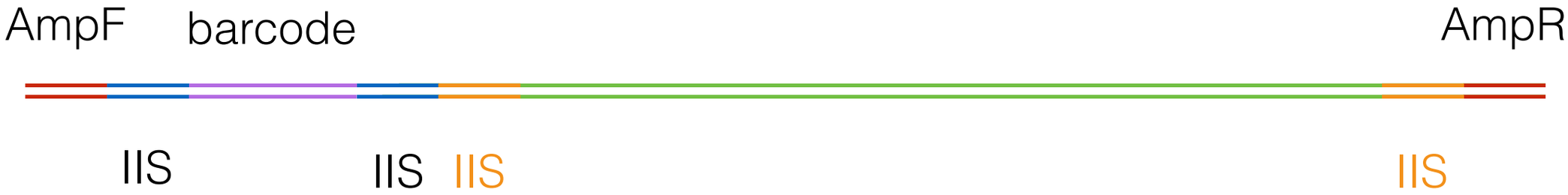


gene

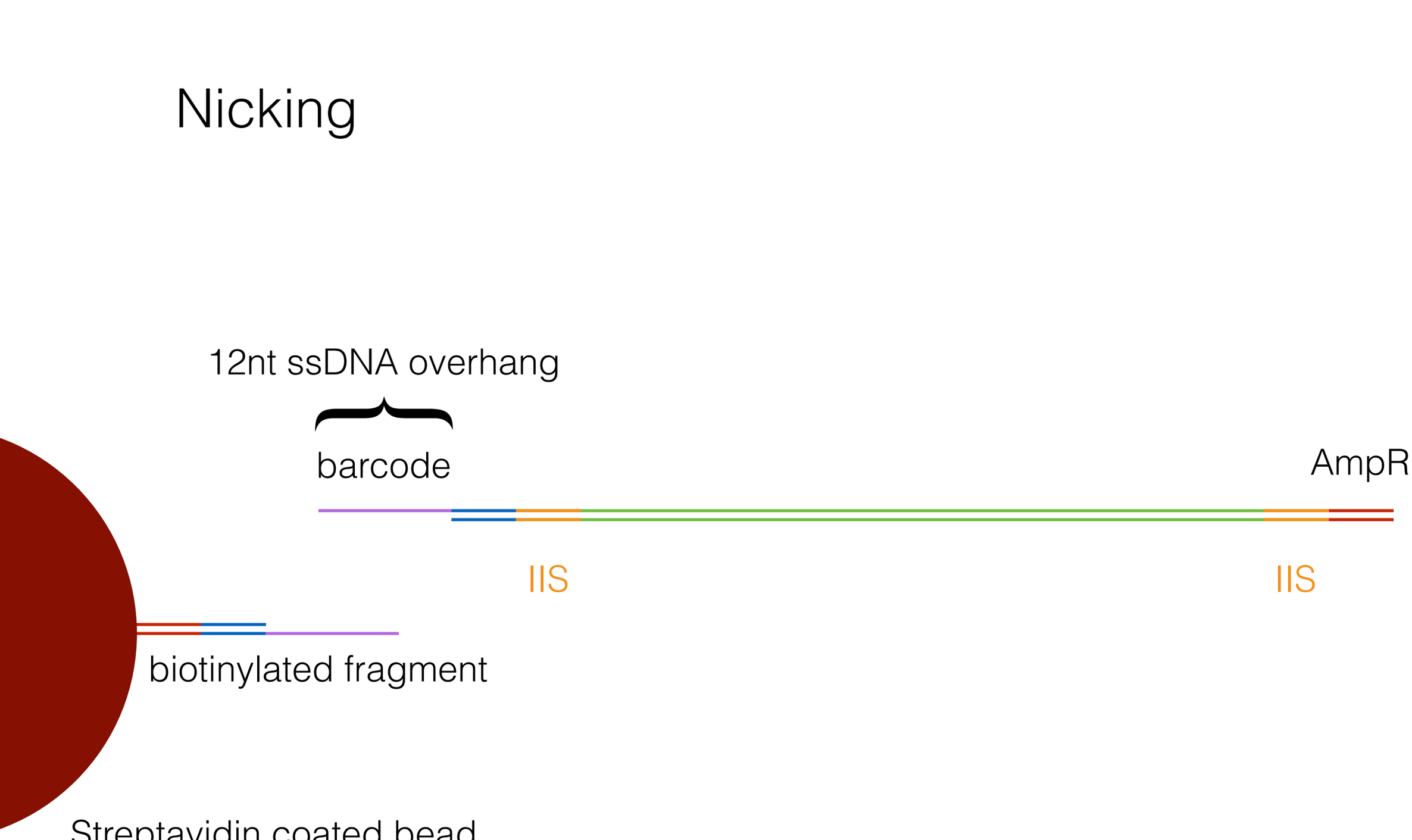


Oligo design

Oligo design

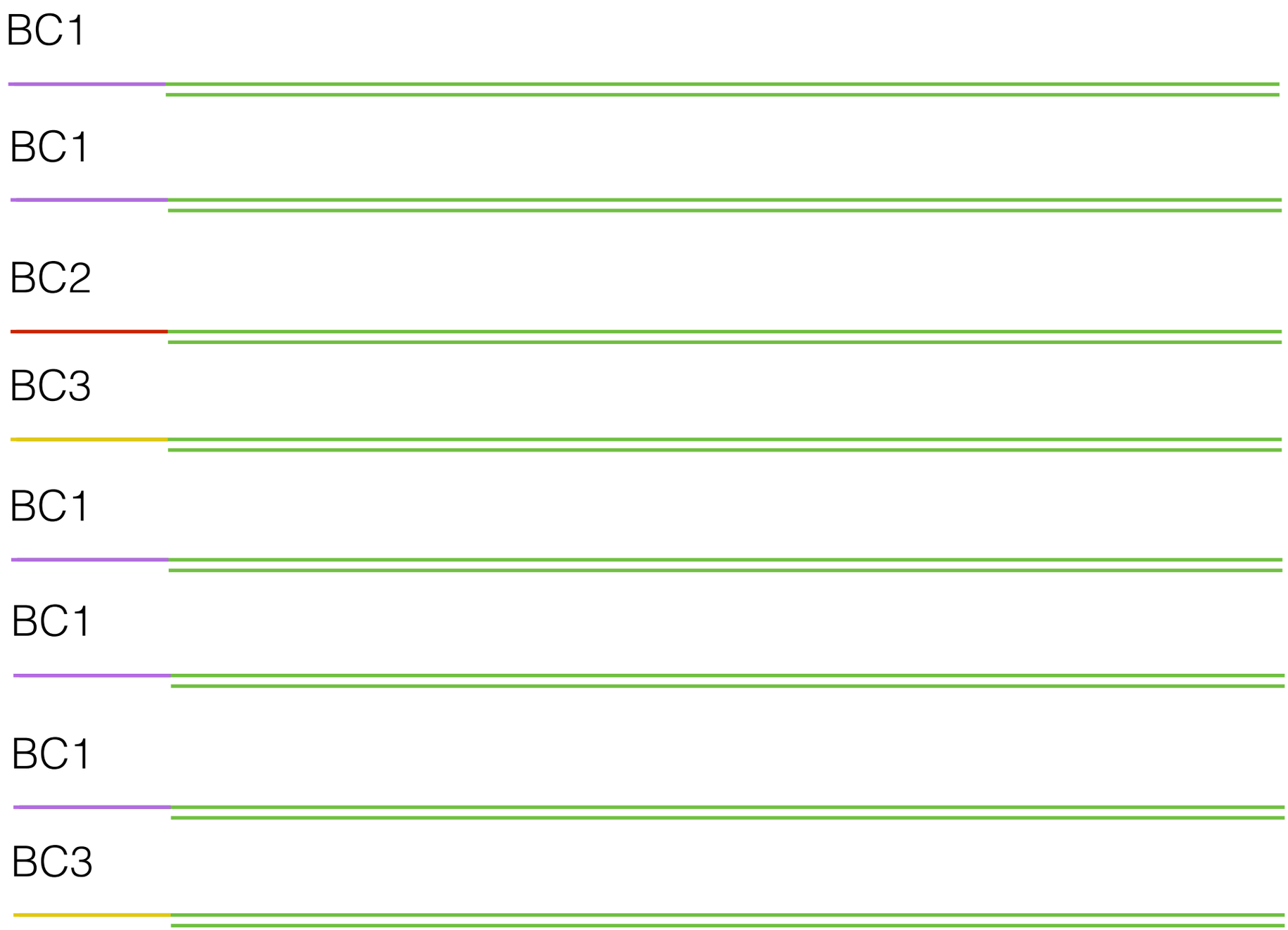


Nicking

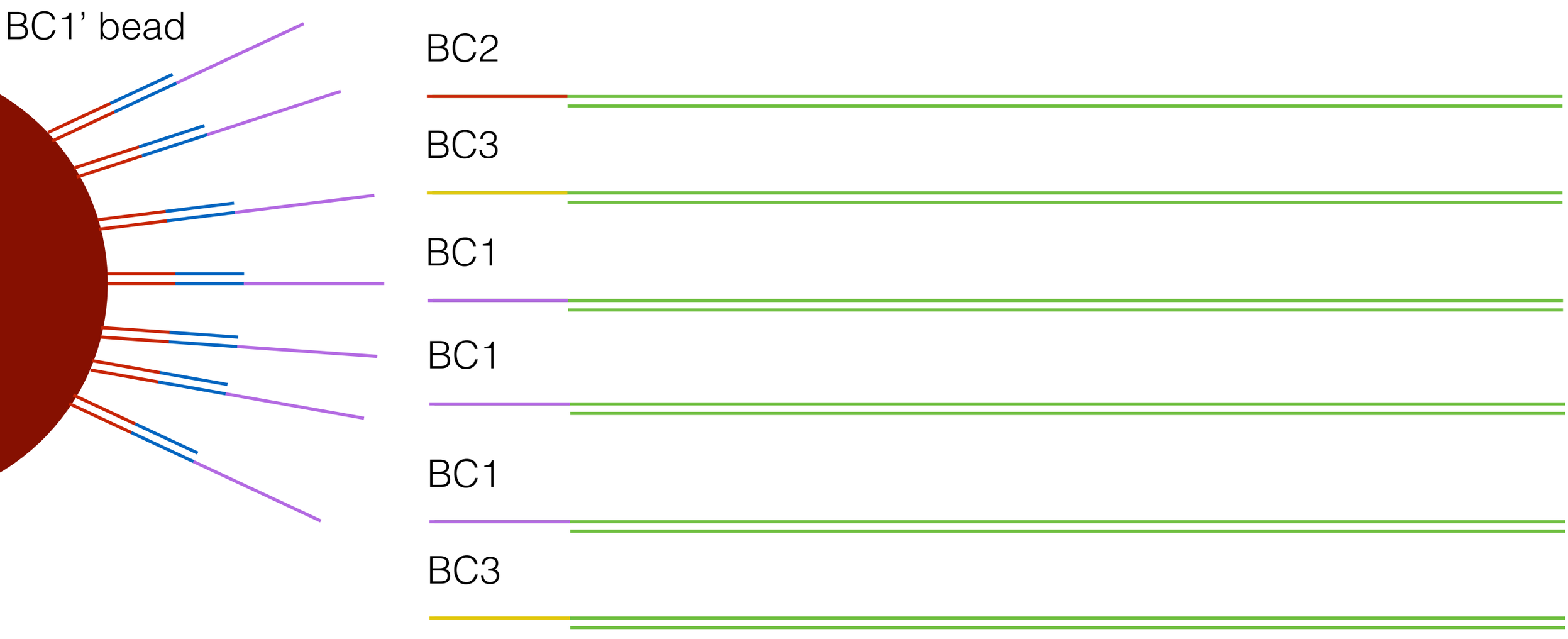

12nt ssDNA overhang
barcode
AmpR
IIS
IIS
biotinylated fragment
Streptavidin coated bead

Processed oligo

Processed oligo

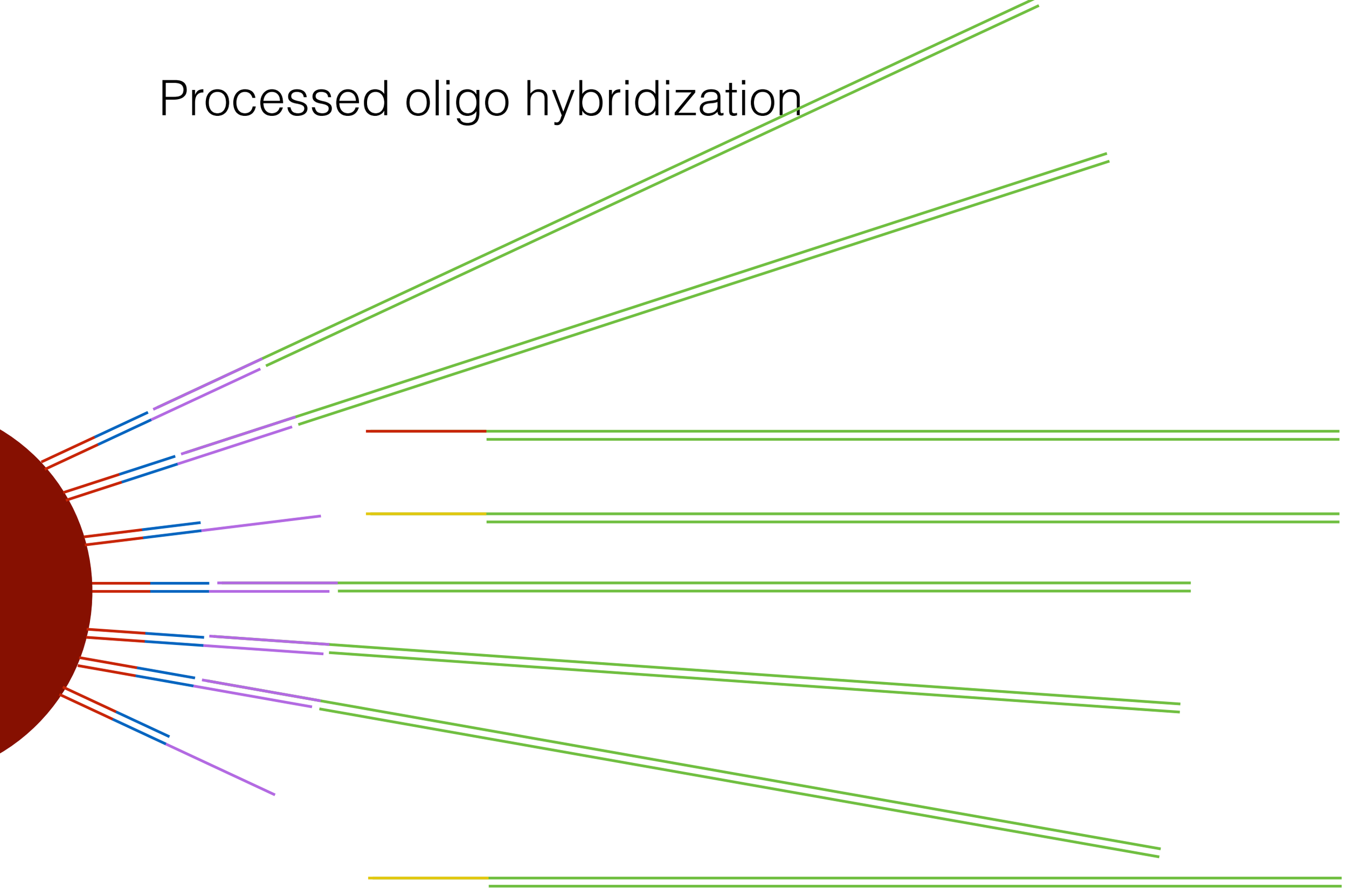


Processed oligo hybridization

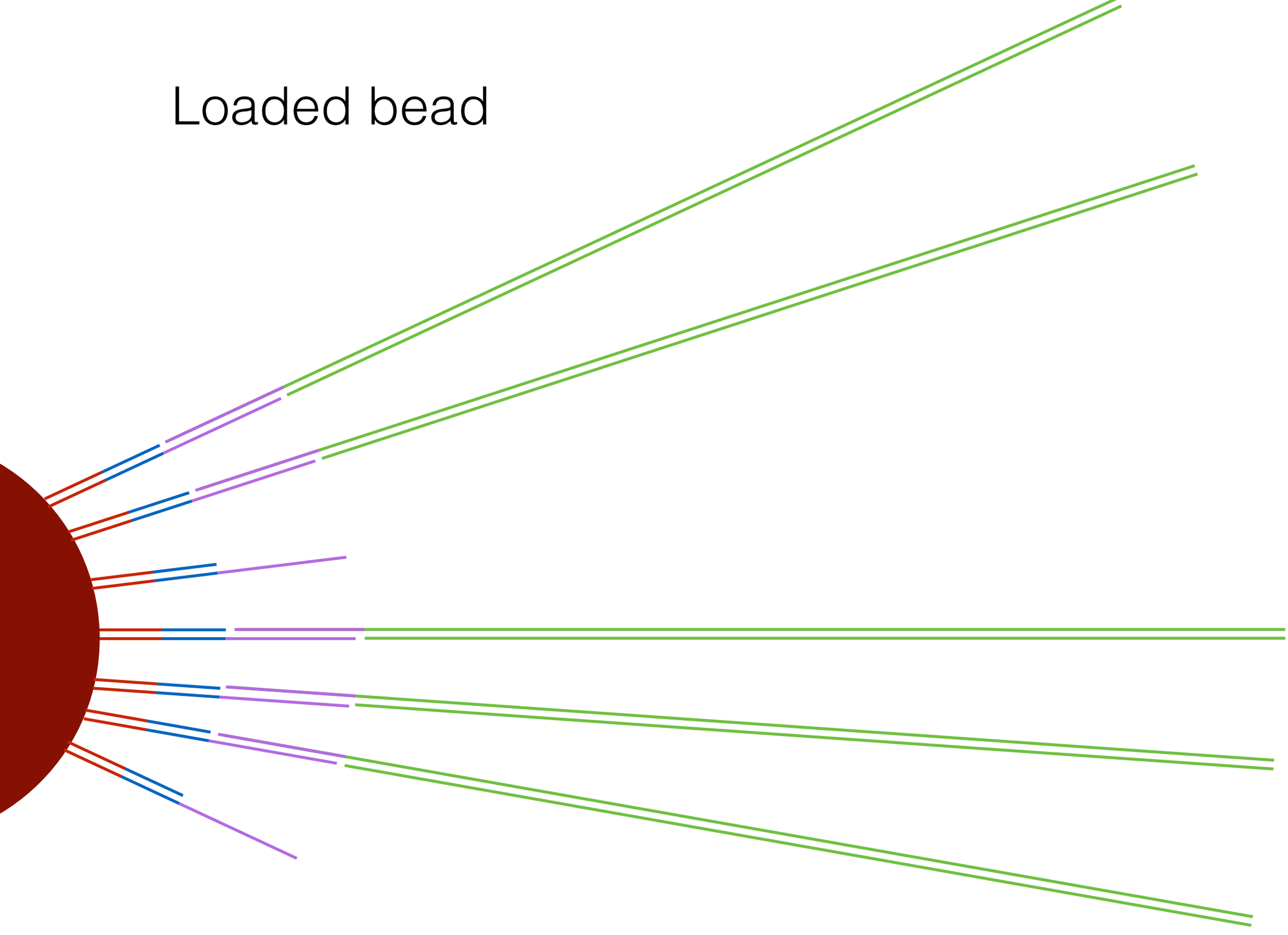


Processed oligo hybridization

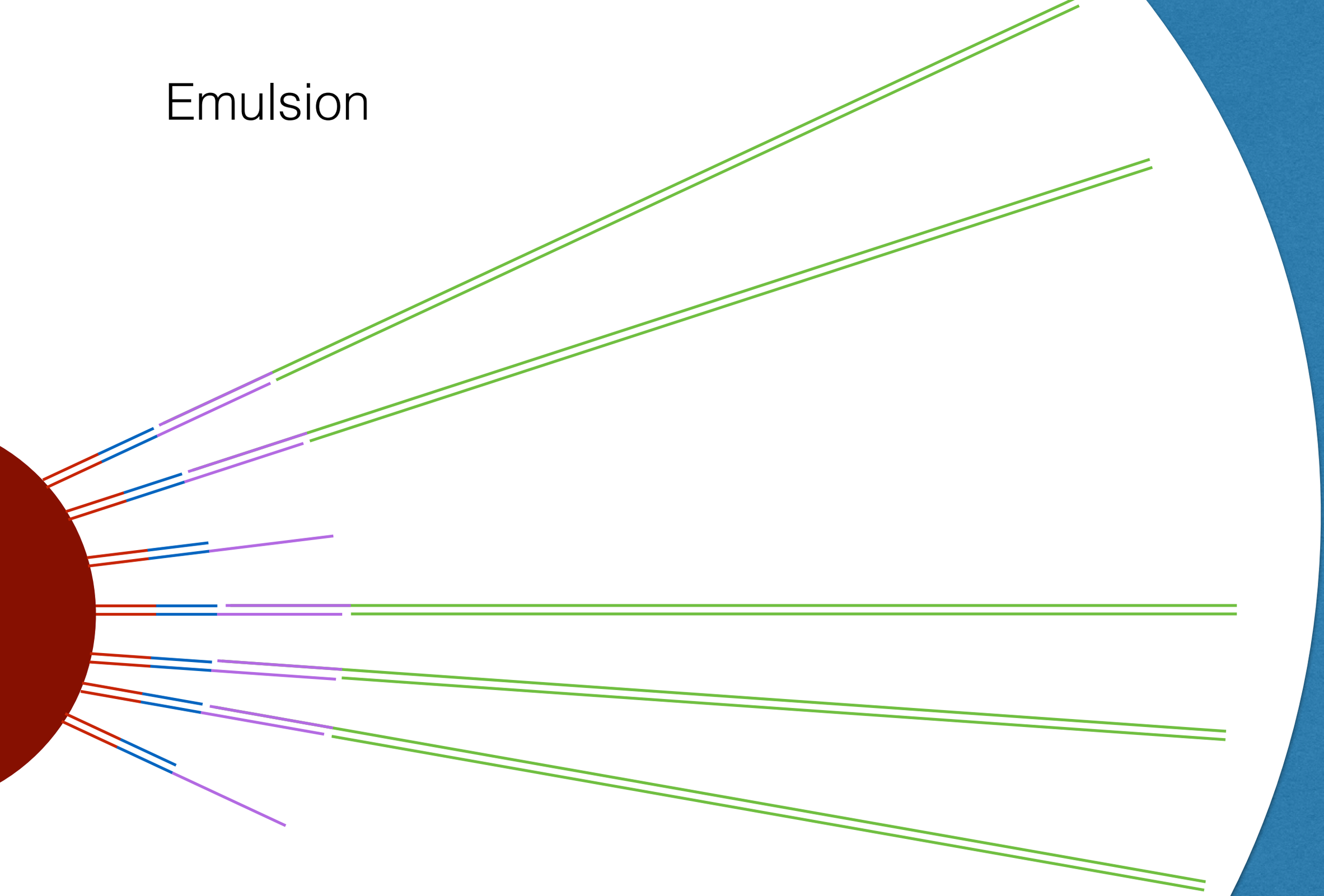
Processed oligo hybridization



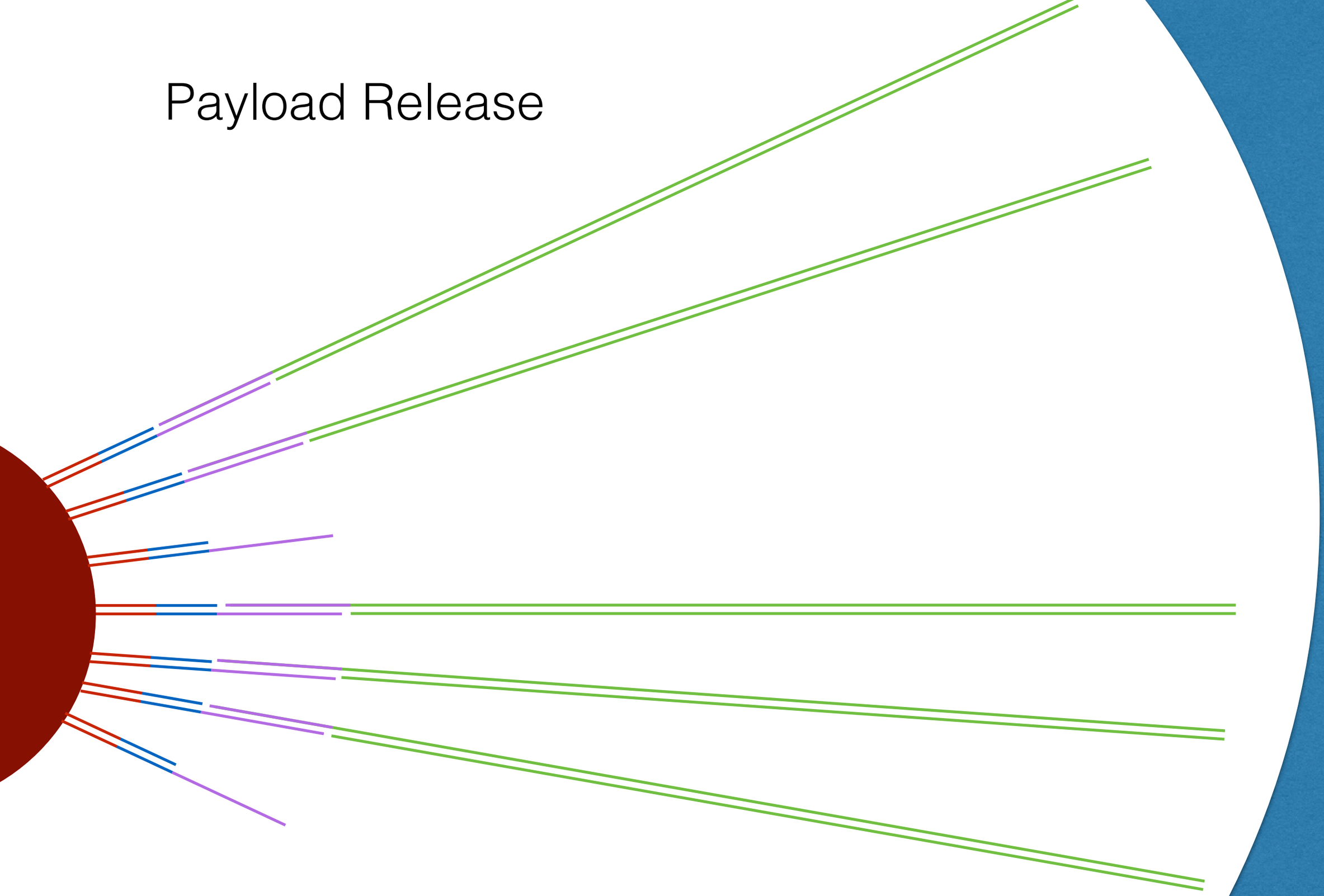
Loaded bead



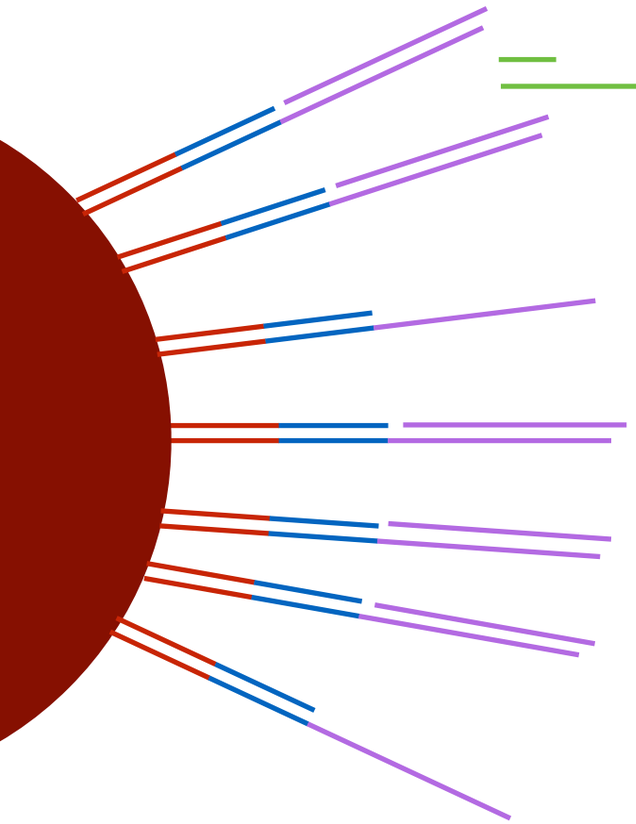
Emulsion



Payload Release



Gene Assembly PCA



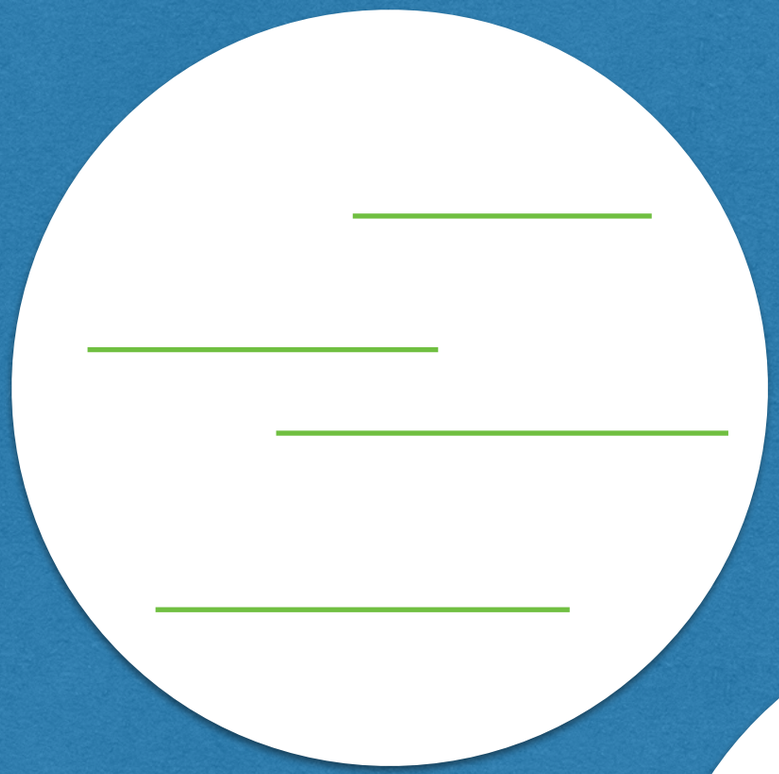
Gene Assembly PCA

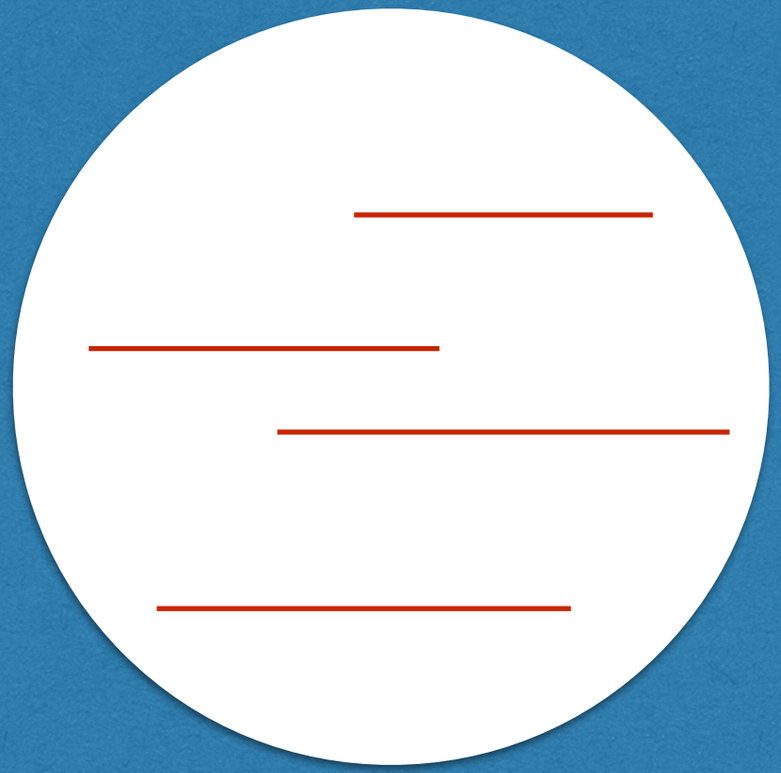
Gene Assembly PCA

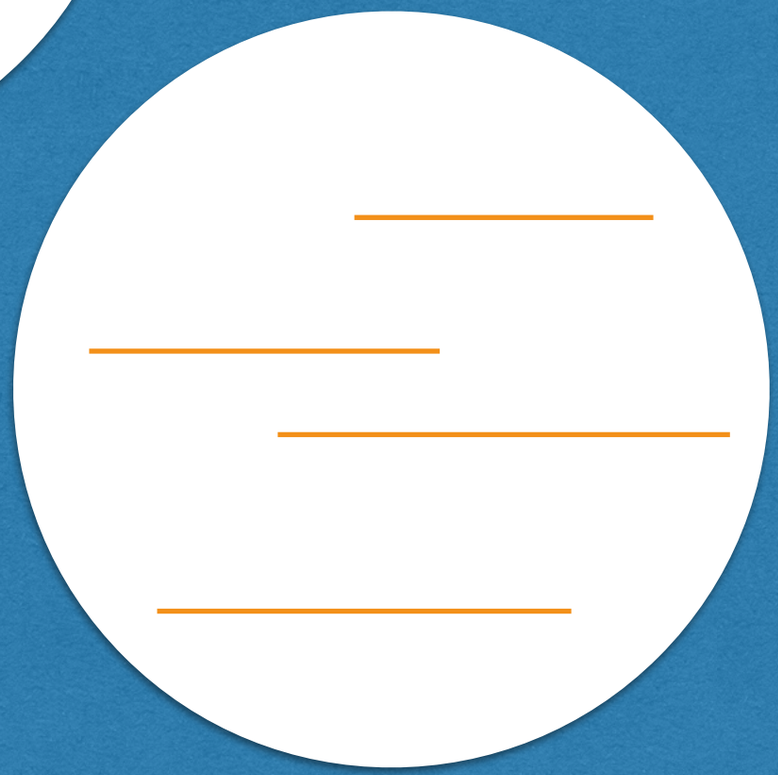
Gene Assembly PCA

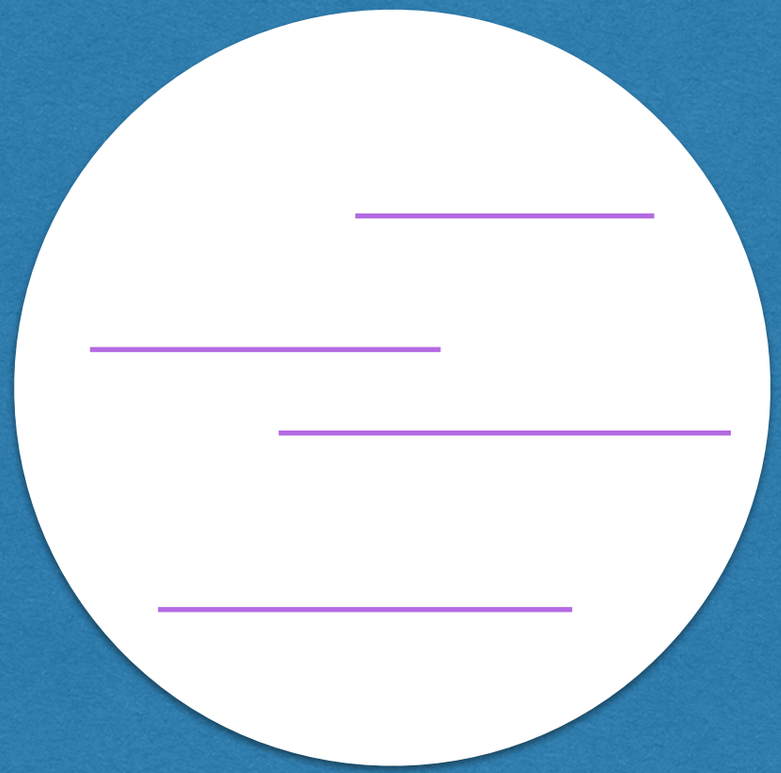
Gene Assembly PCA

Gene Assembly PCA



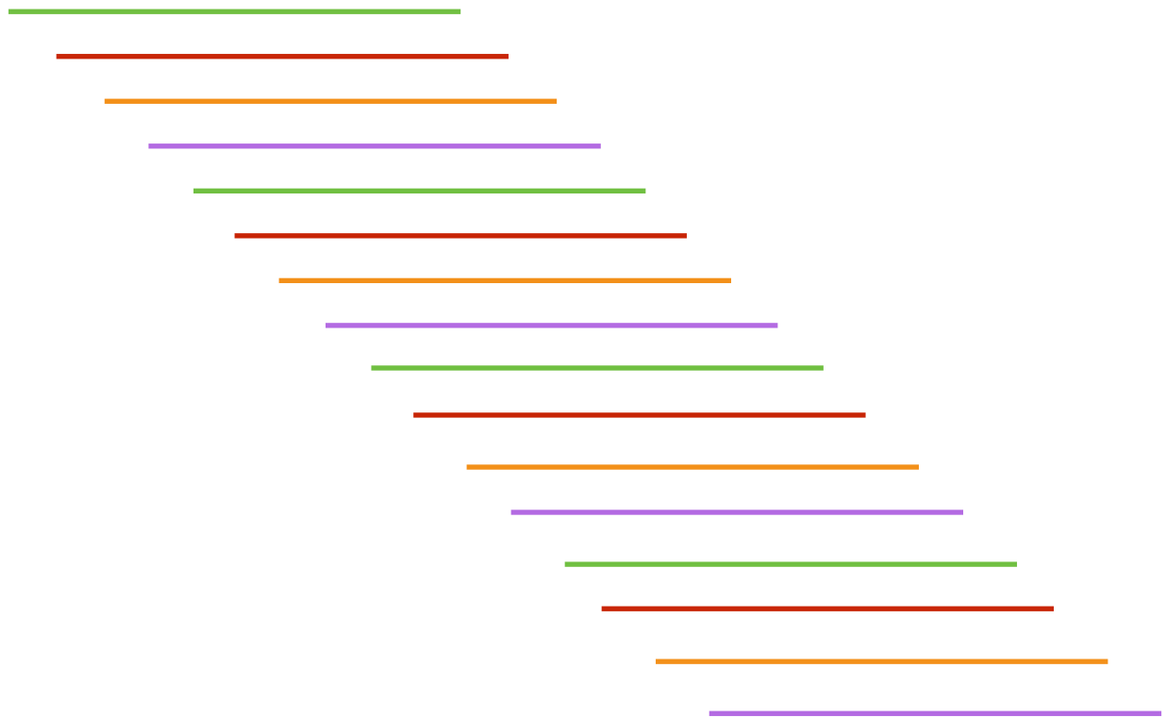




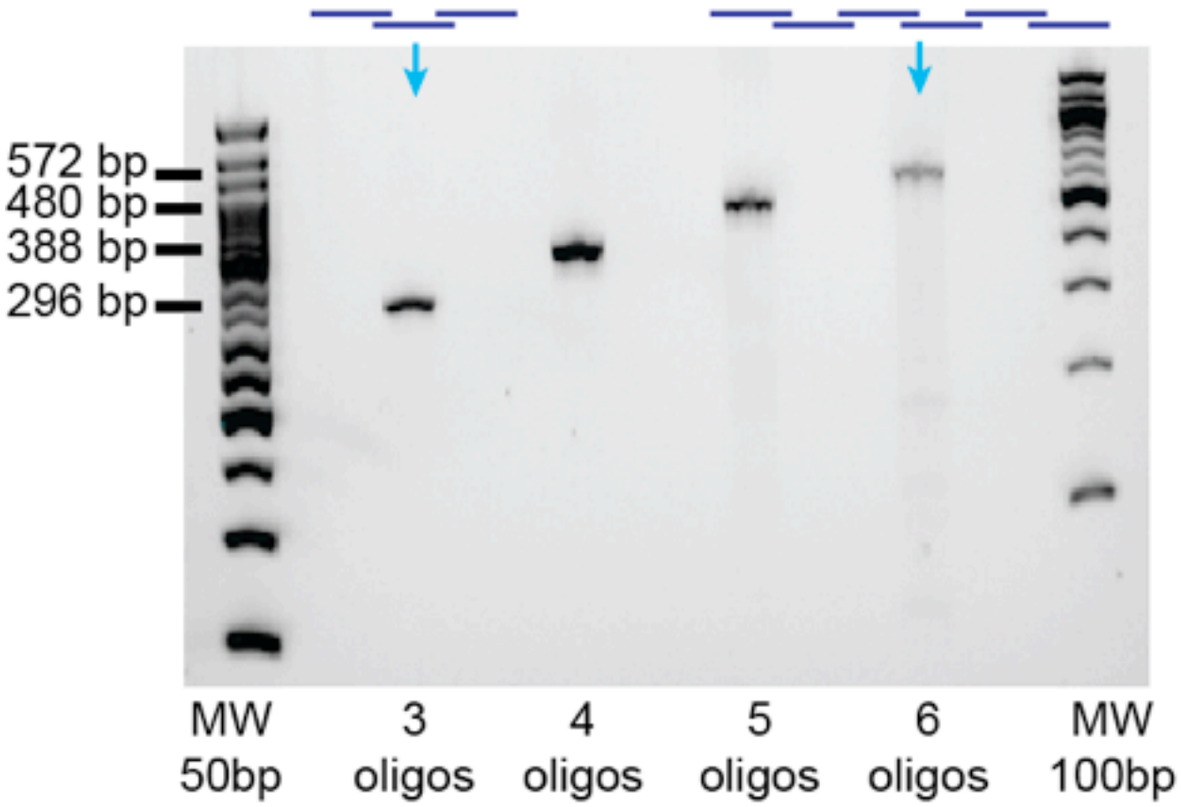


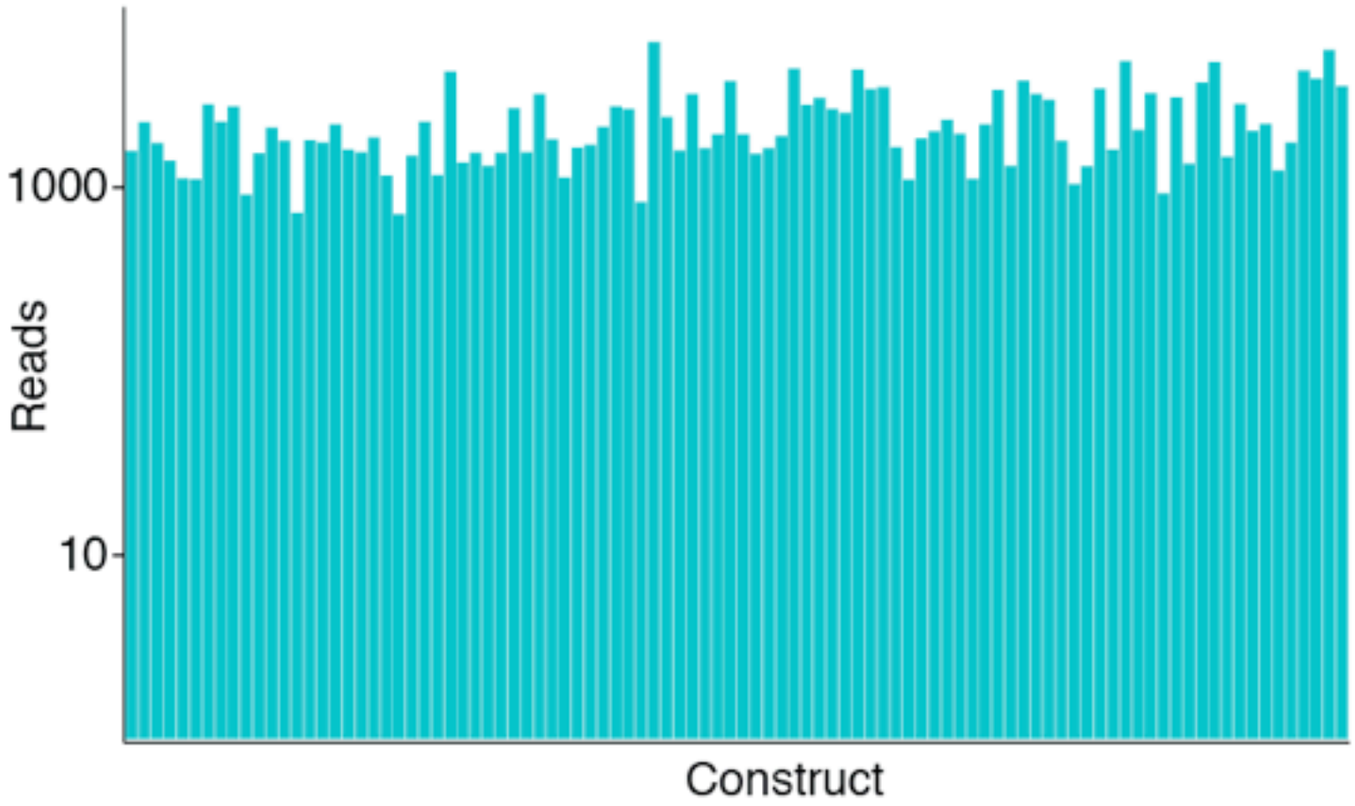
Break Emulsion

Assembled Gene Library



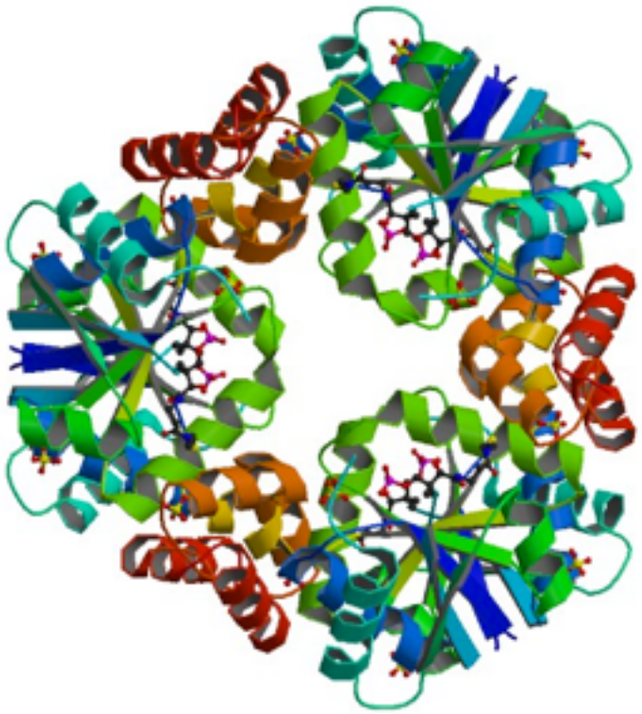
Optimization

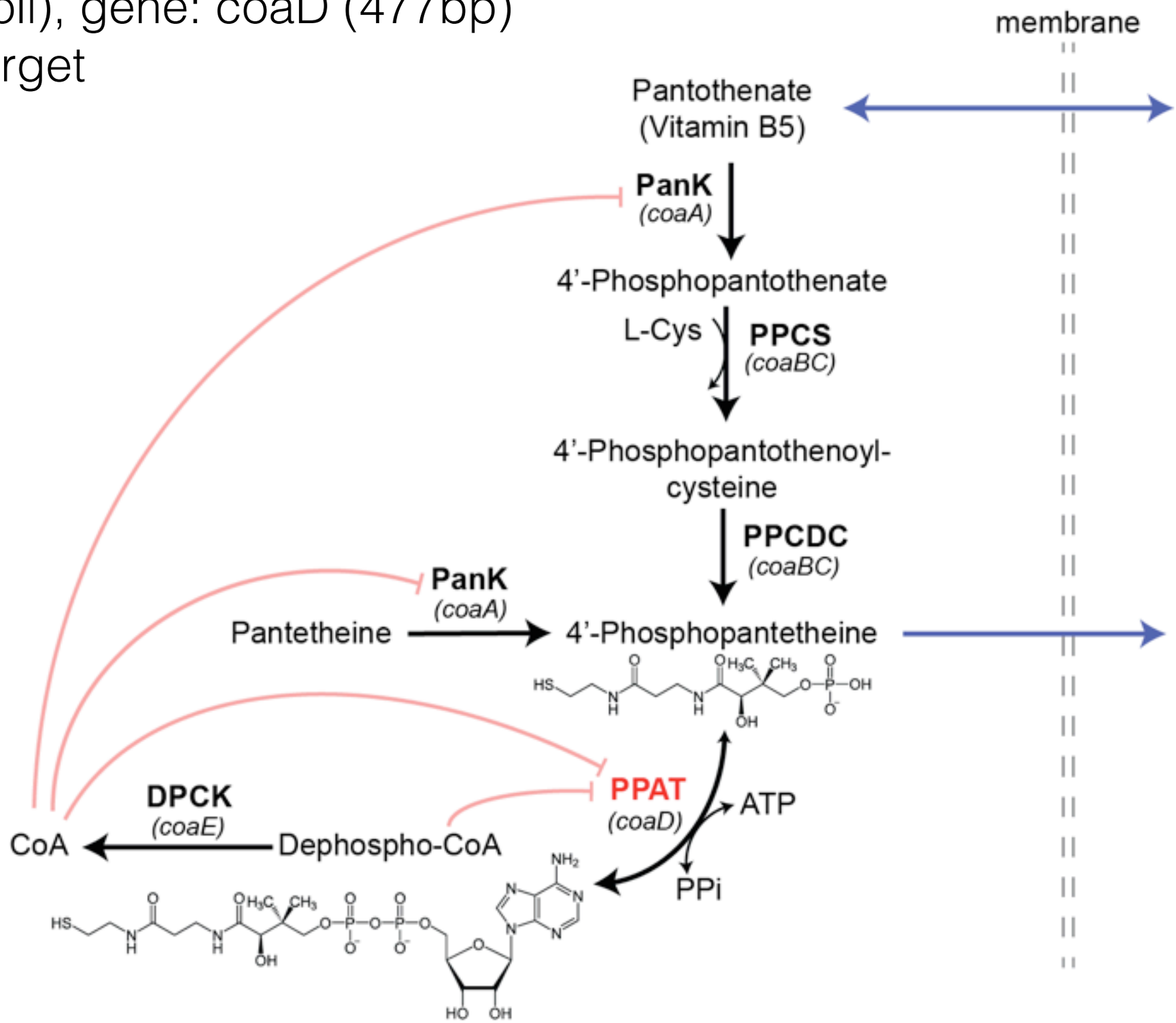




Phosphopantetheine adenylyltransferase (PPAT)

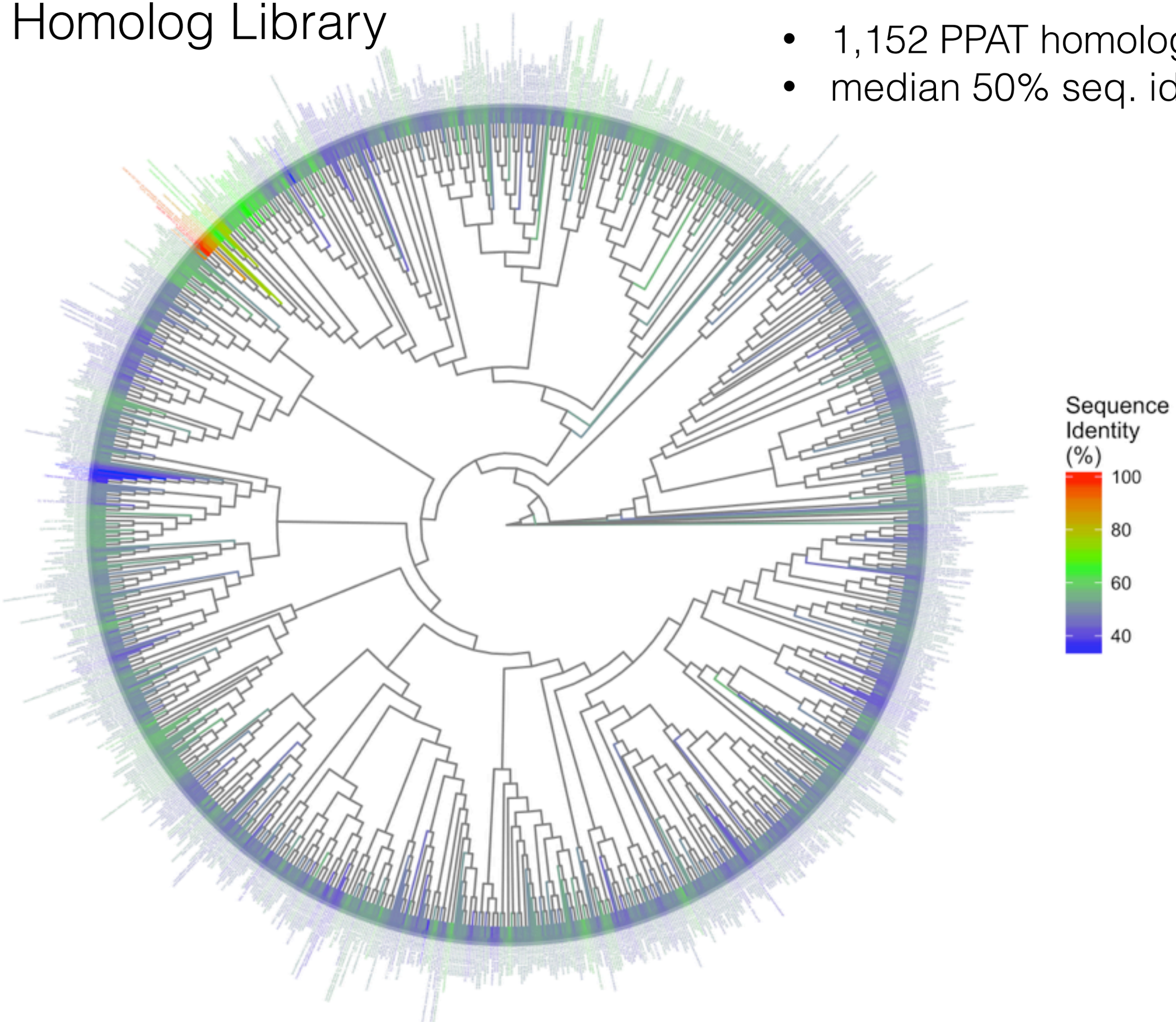
- essential enzyme in coenzyme A biosynthesis
- protein: 159 aa (E. coli), gene: coaD (477bp)
- potential antibiotic target



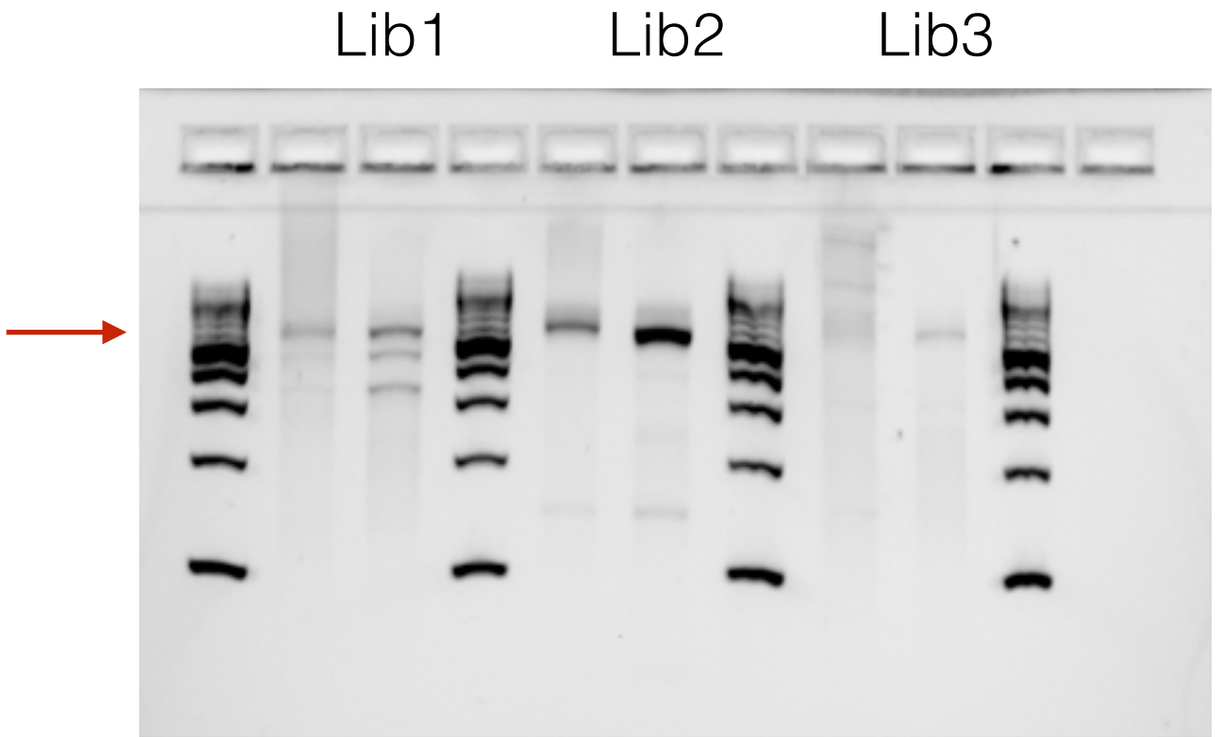


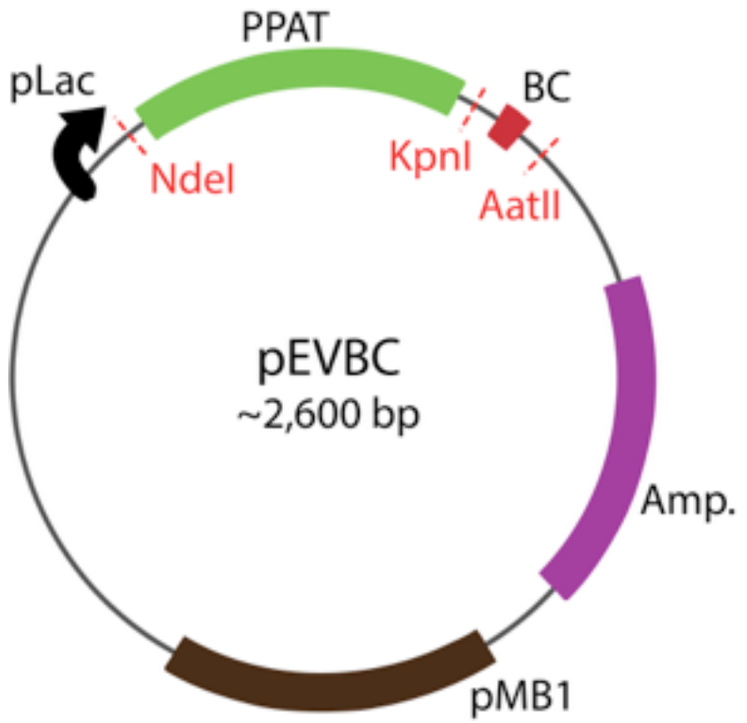
PPAT Homolog Library

- 1,152 PPAT homologs
- median 50% seq. identity

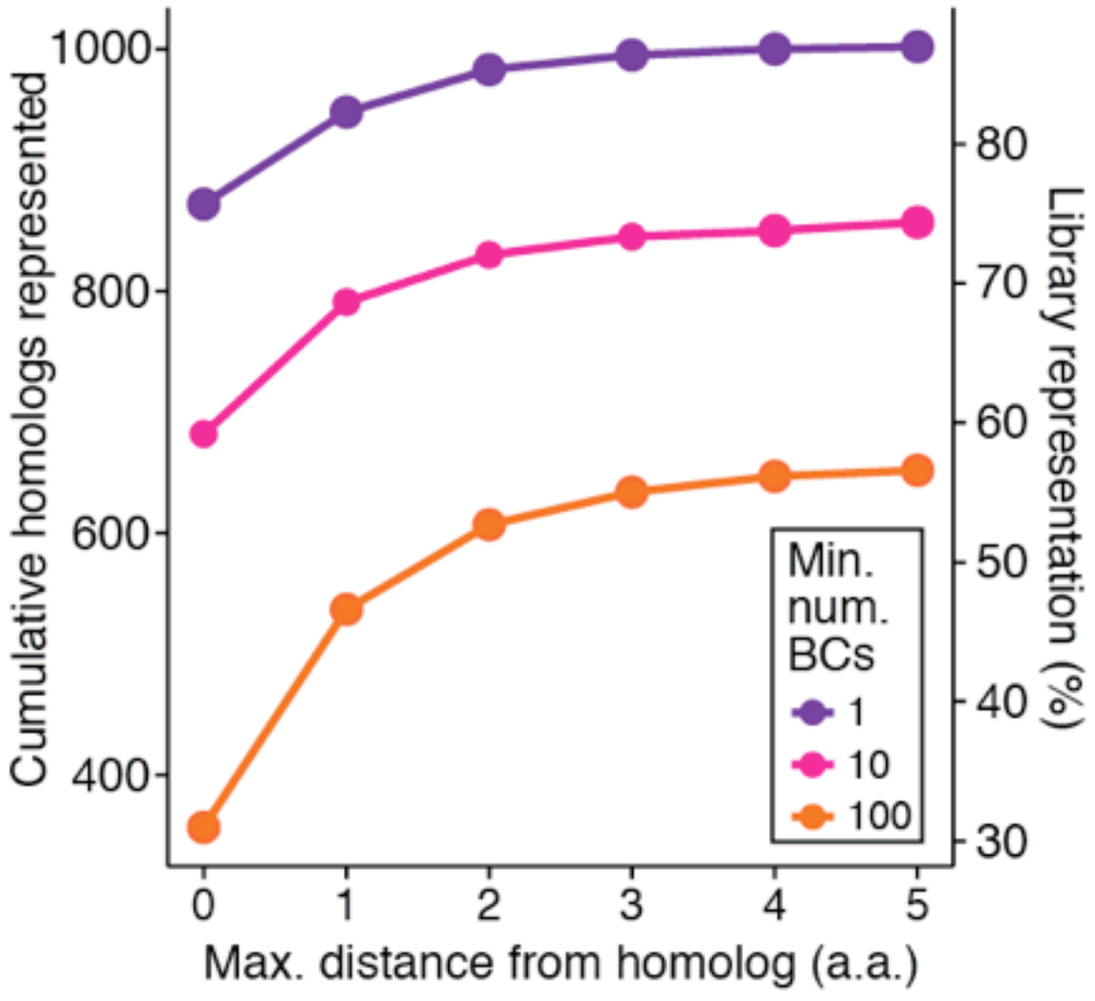


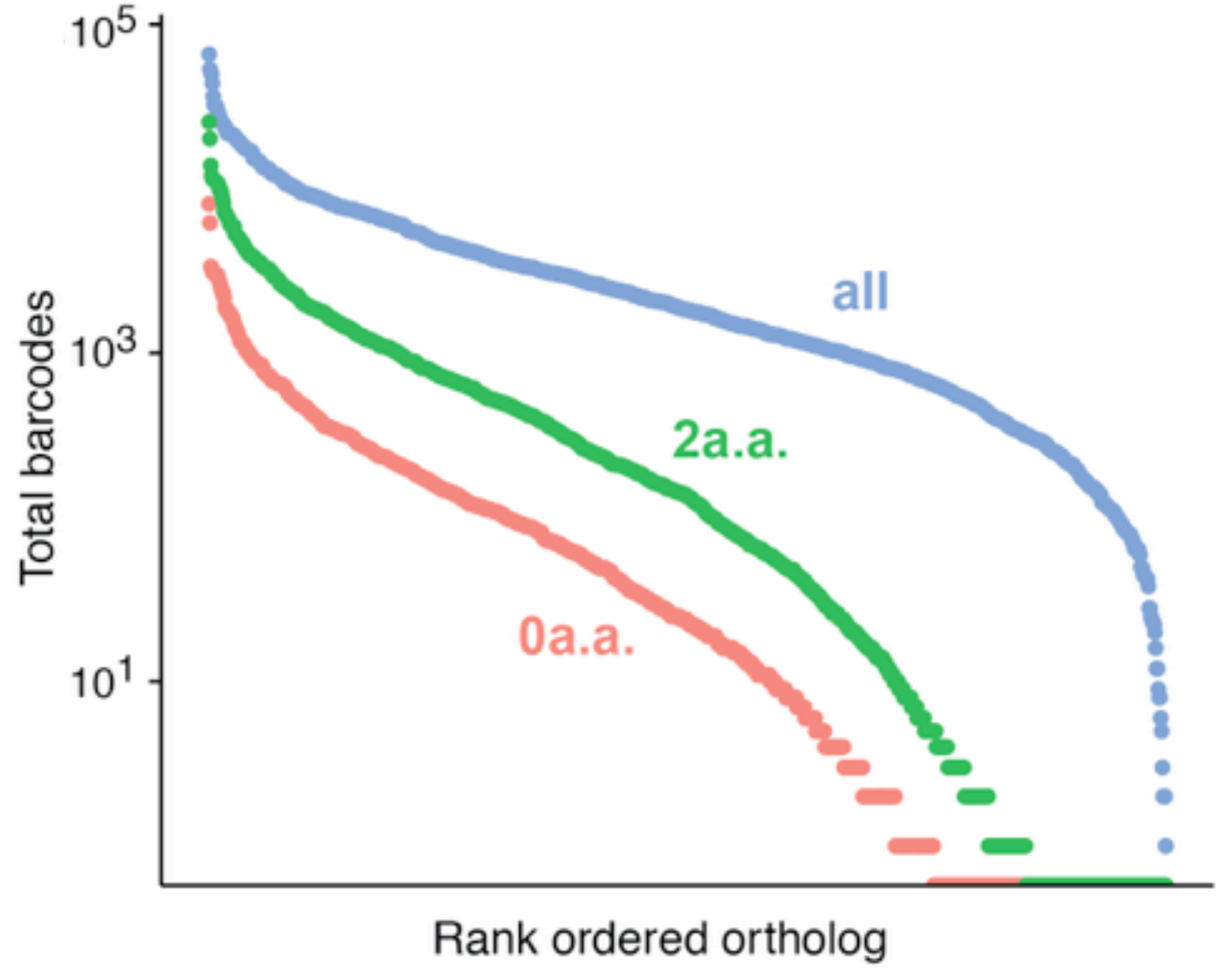
Assembly


Lib1
Lib2
Lib3

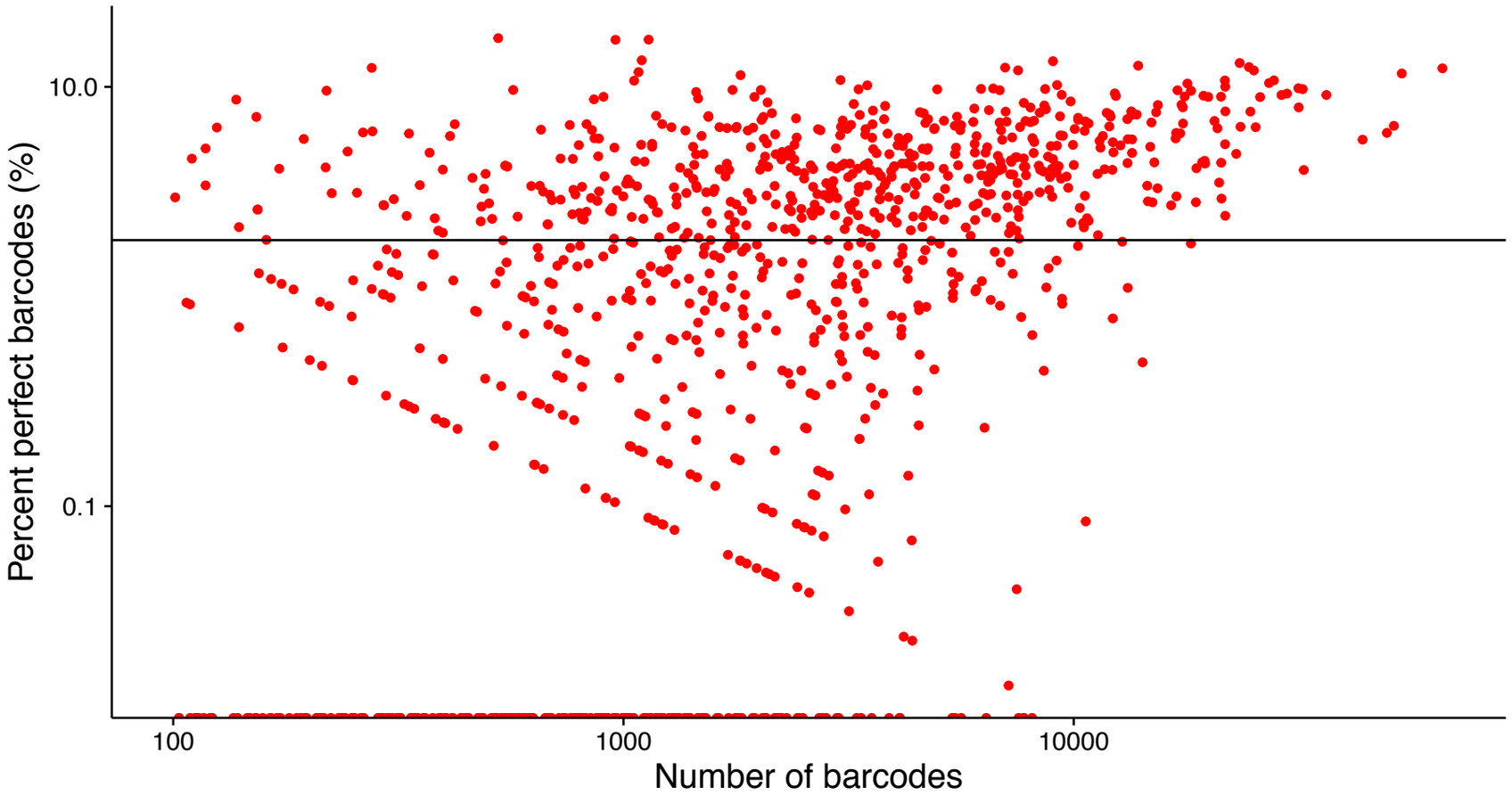

PPAT
pLac
BC
NdeI
KpnI
AatII
pEVBC
~2,600 bp
Amp.
pMB1

Library Coverage and Uniformity



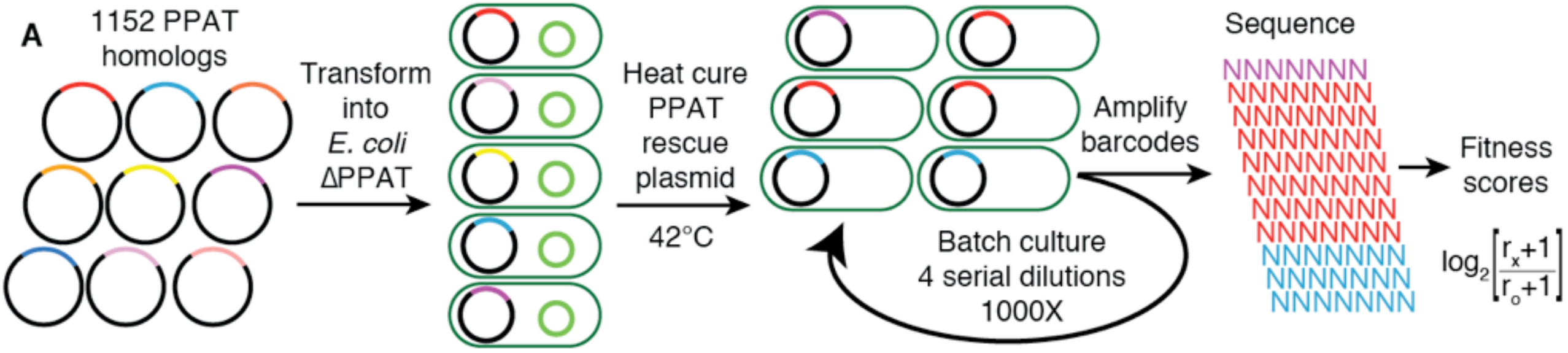


Perfect Assemblies

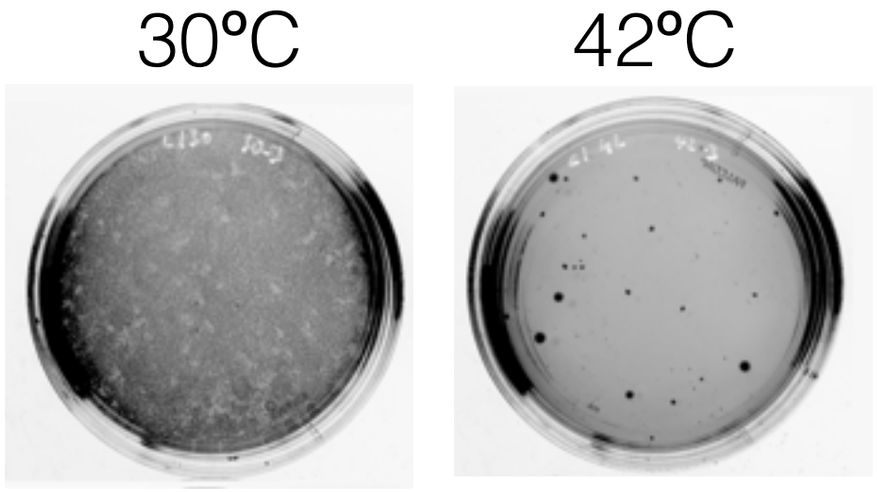


For homologs with at least 100 BCs
Median = 1.8%

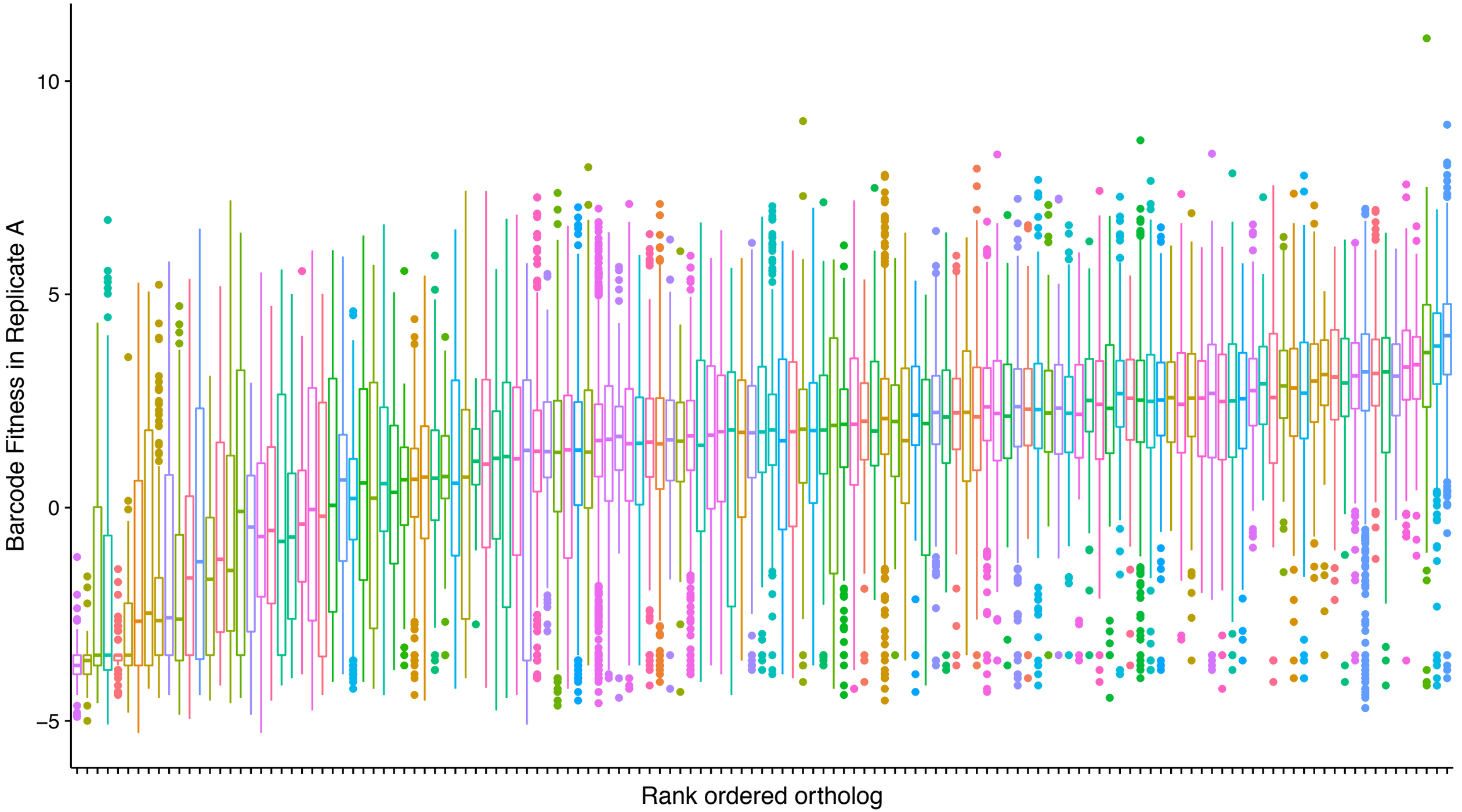
Pooled Complementation Screen



Rescue plasmid escape frequency:
1 in 20,000

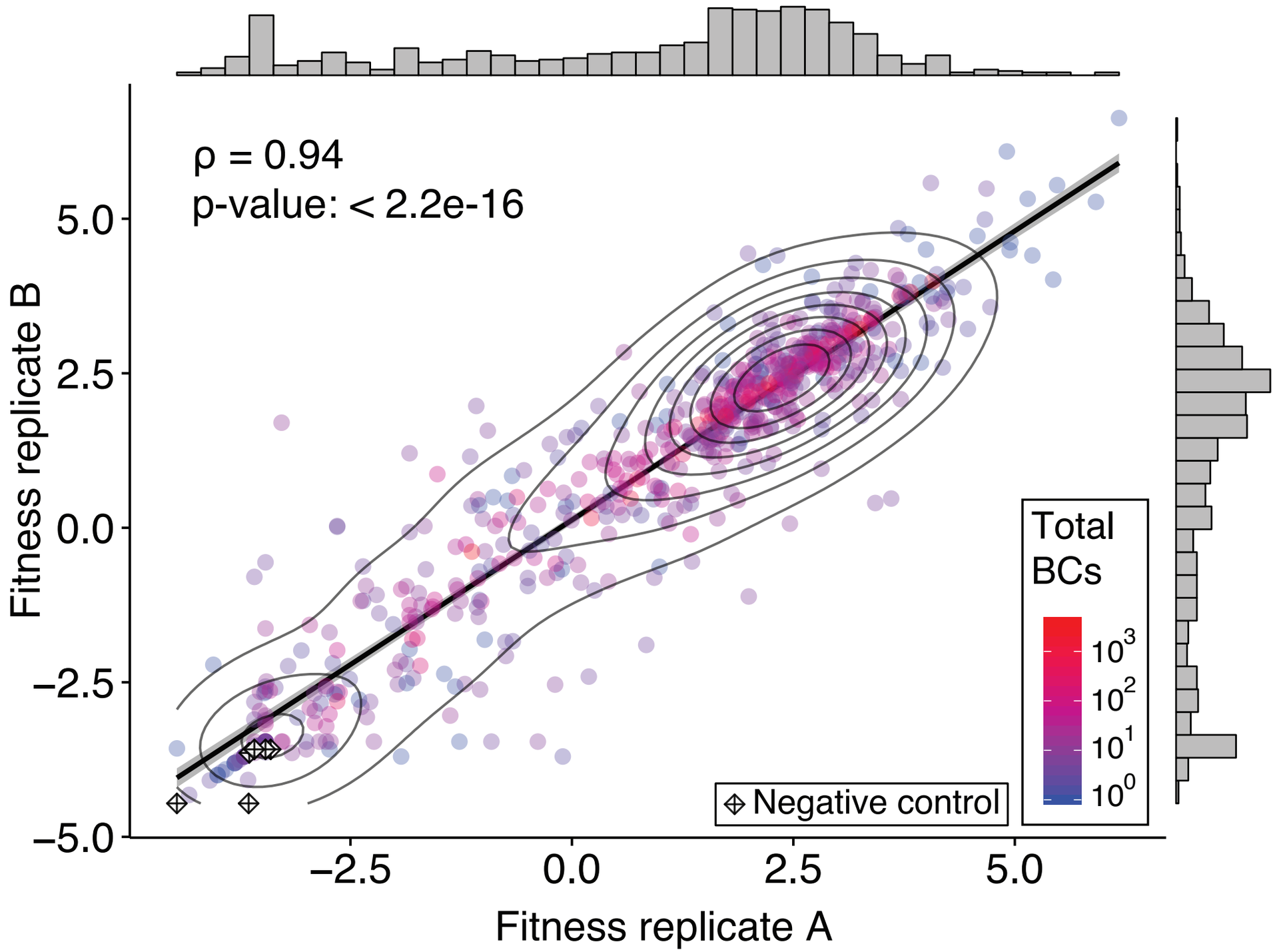


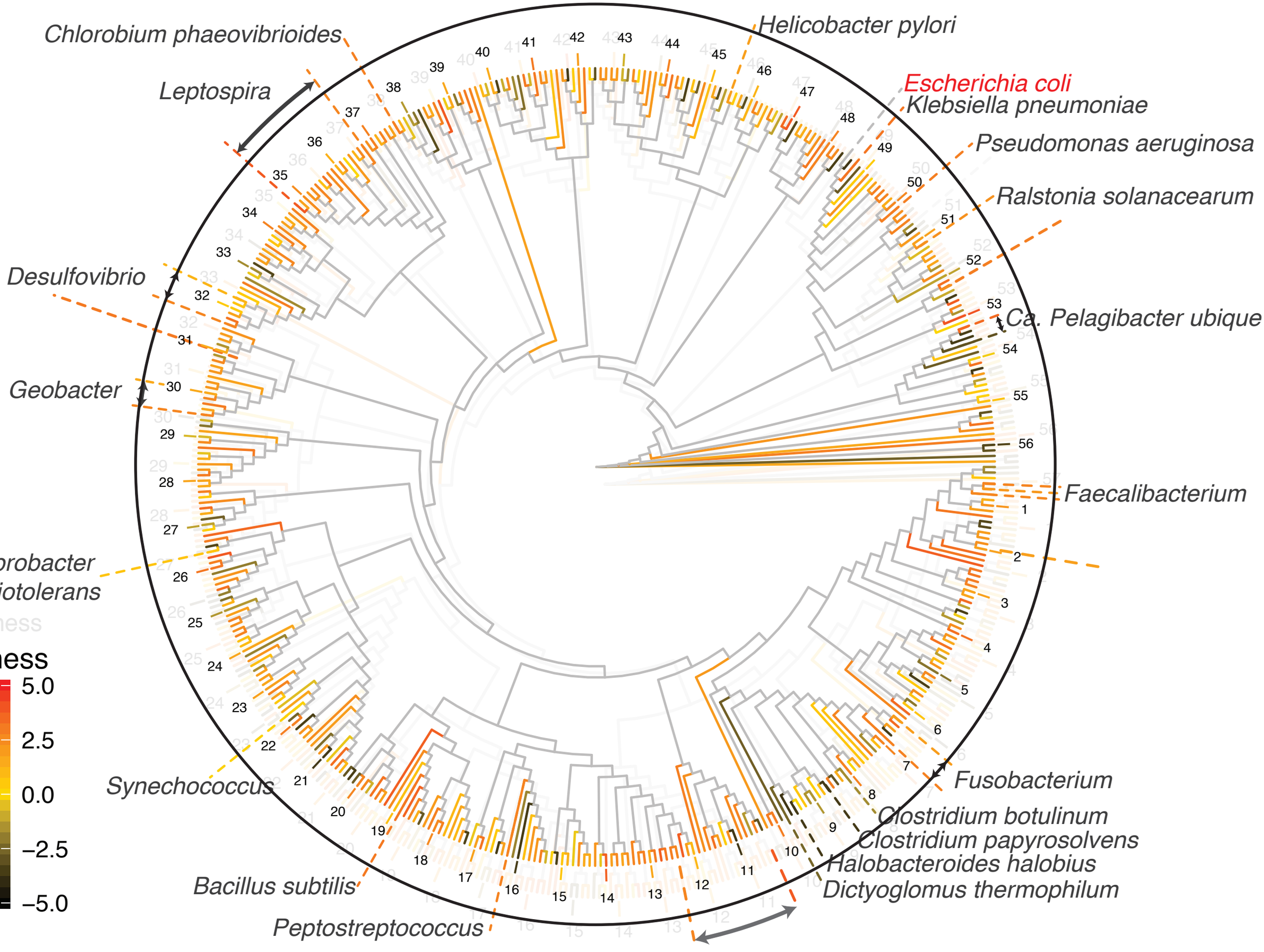
Noise in Barcode Fitness

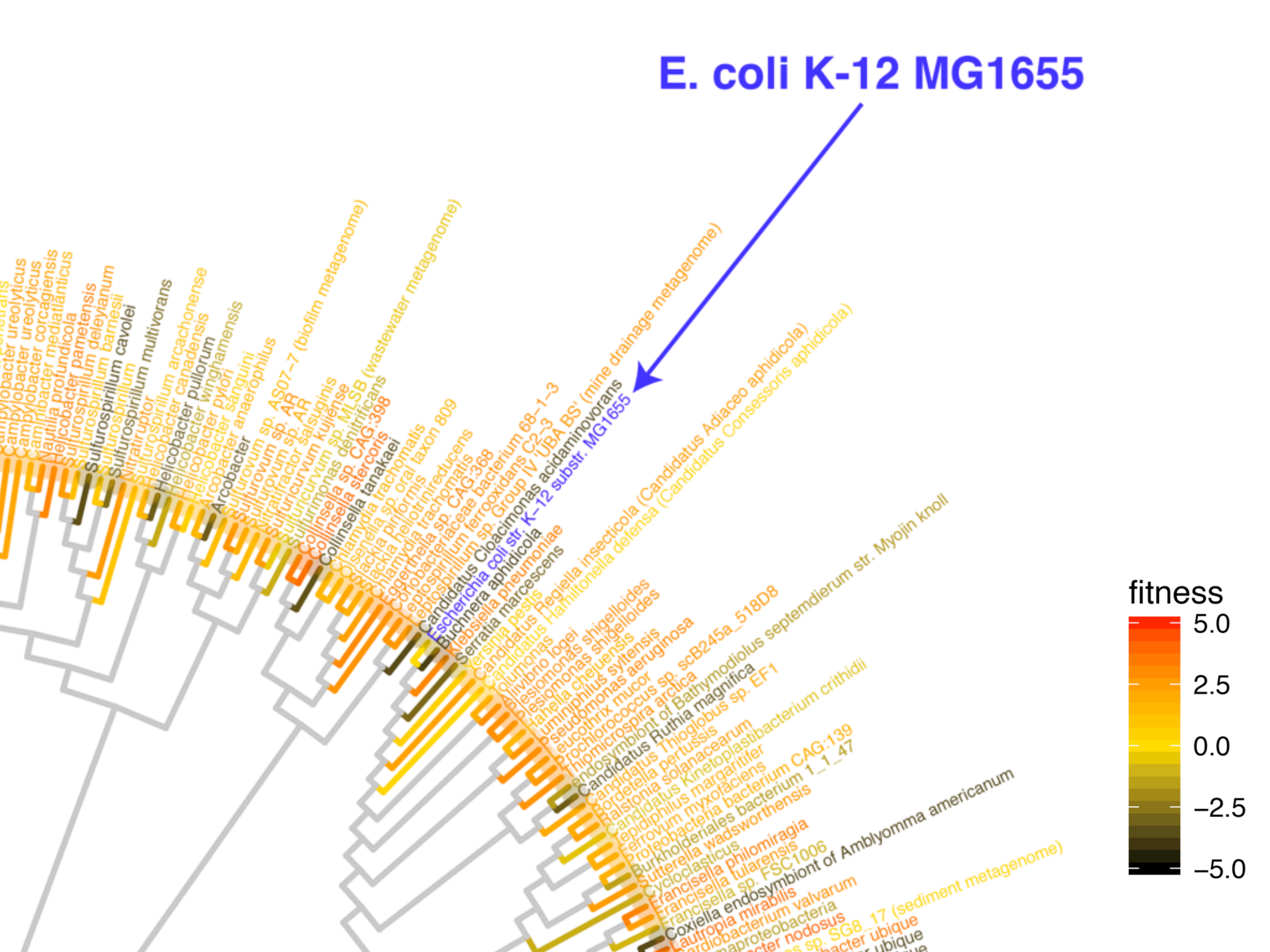


For orthologs with at least 50 BCs

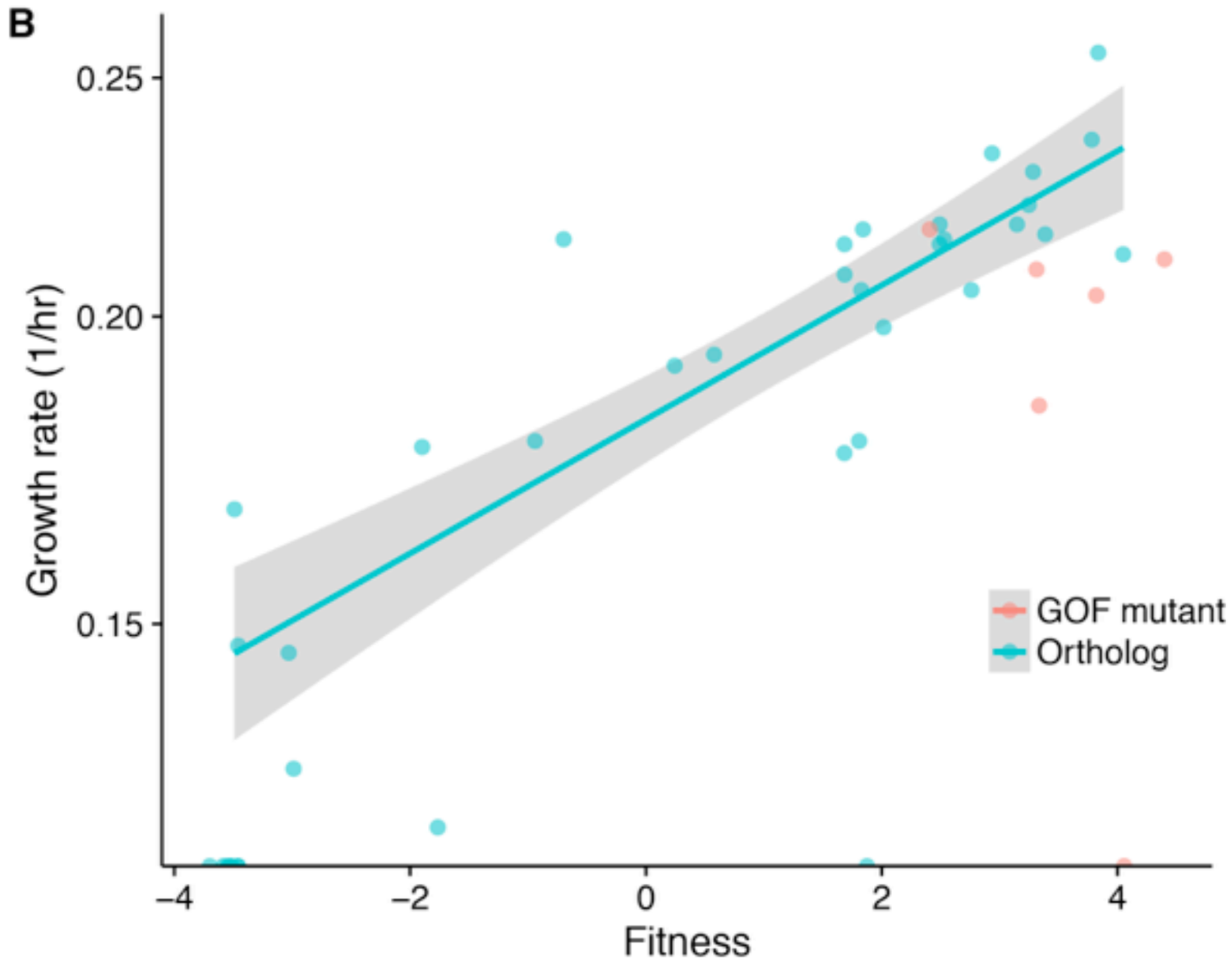
Biological Replicates (just homologs)



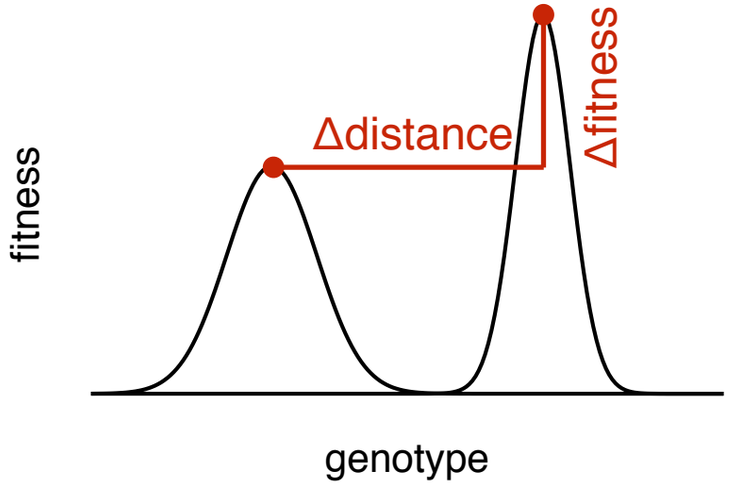

Chlorobium phaeovibrioides
Leptospira
Helicobacter pylori
Escherichia coli
Klebsiella pneumoniae
Pseudomonas aeruginosa
Ralstonia solanacearum
Ca. Pelagibacter ubique
Faecalibacterium
Fusobacterium
Clostridium botulinum
Clostridium papyrosolvens
Halobacteroides halobius
Dictyoglomus thermophilum
Peptostreptococcus
Bacillus subtilis
Synechococcus
Rubrobacter radiotolerans
Geobacter
Desulfovibrio
fitness
fitness
5.0
2.5
0.0
−2.5
−5.0

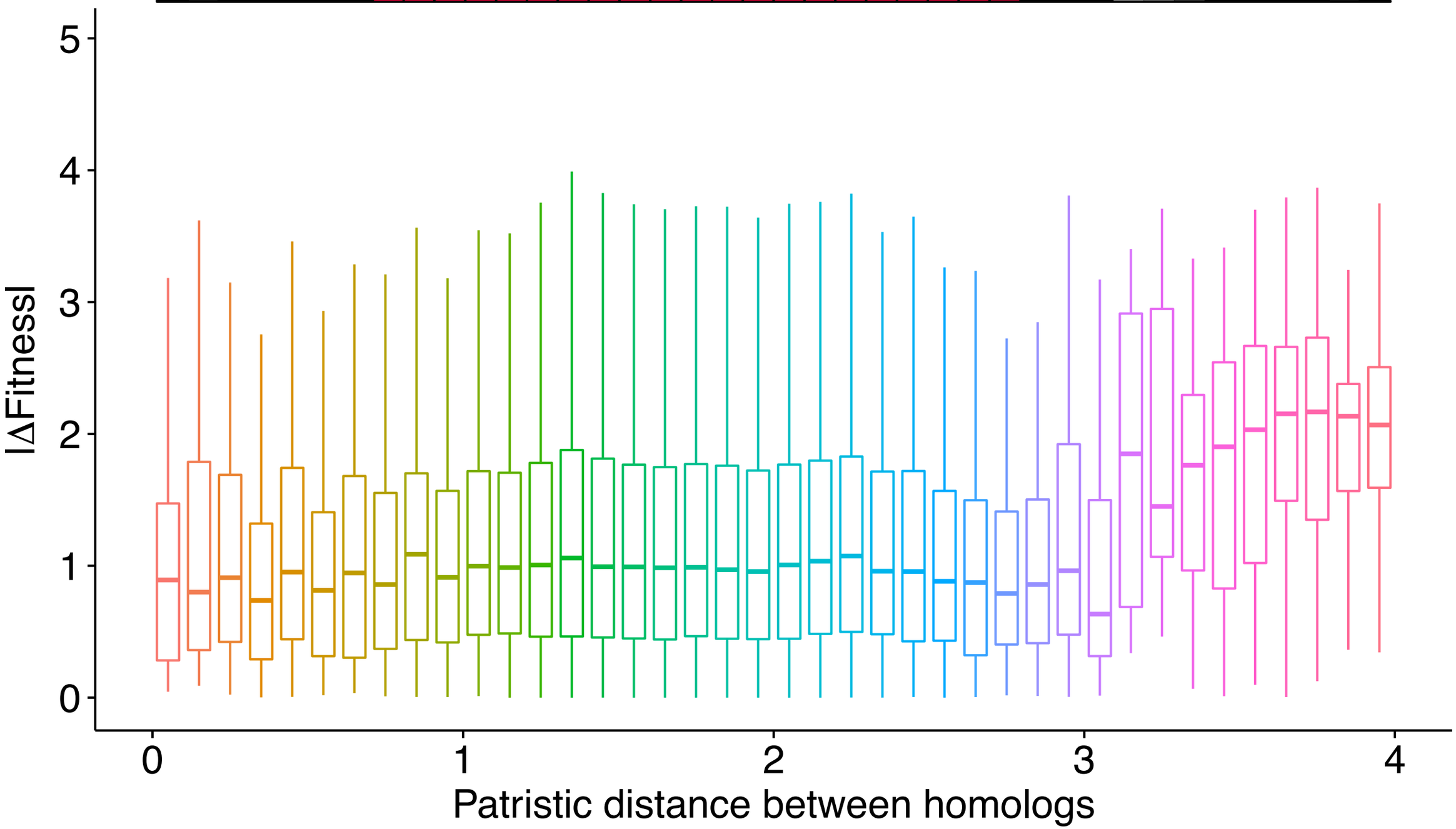

E. coli K-12 MG1655
fitness
5.0
2.5
0.0
−2.5
−5.0
Escherichia coli str. K-12 substr. MG1655
Candidatus Cloacimonas acidaminovorans
Buchnera aphidicola
Klebsiella pneumoniae
Serratia marcescens
Yersinia pestis
Candidatus Regiella insecticola (Candidatus Adiaceo aphidicola)
Candidatus Hamiltonella defensa (Candidatus Consessoris aphidicola)
endosymbiont of Bathymodiolus septemdierum str. Myojin knoll
Candidatus Ruthia magnifica
Pseudomonas aeruginosa
Bordetella pertussis
Ralstonia solanacearum
Sutterella wadsworthensis
Francisella tularensis
Coxiella endosymbiont of Amblyomma americanum
Collinsella tanakaei
Sulfurospirillum multivorans
Sulfurospirillum cavolei
Helicobacter pullorum
Arcobacter

Individual Fitness Testing

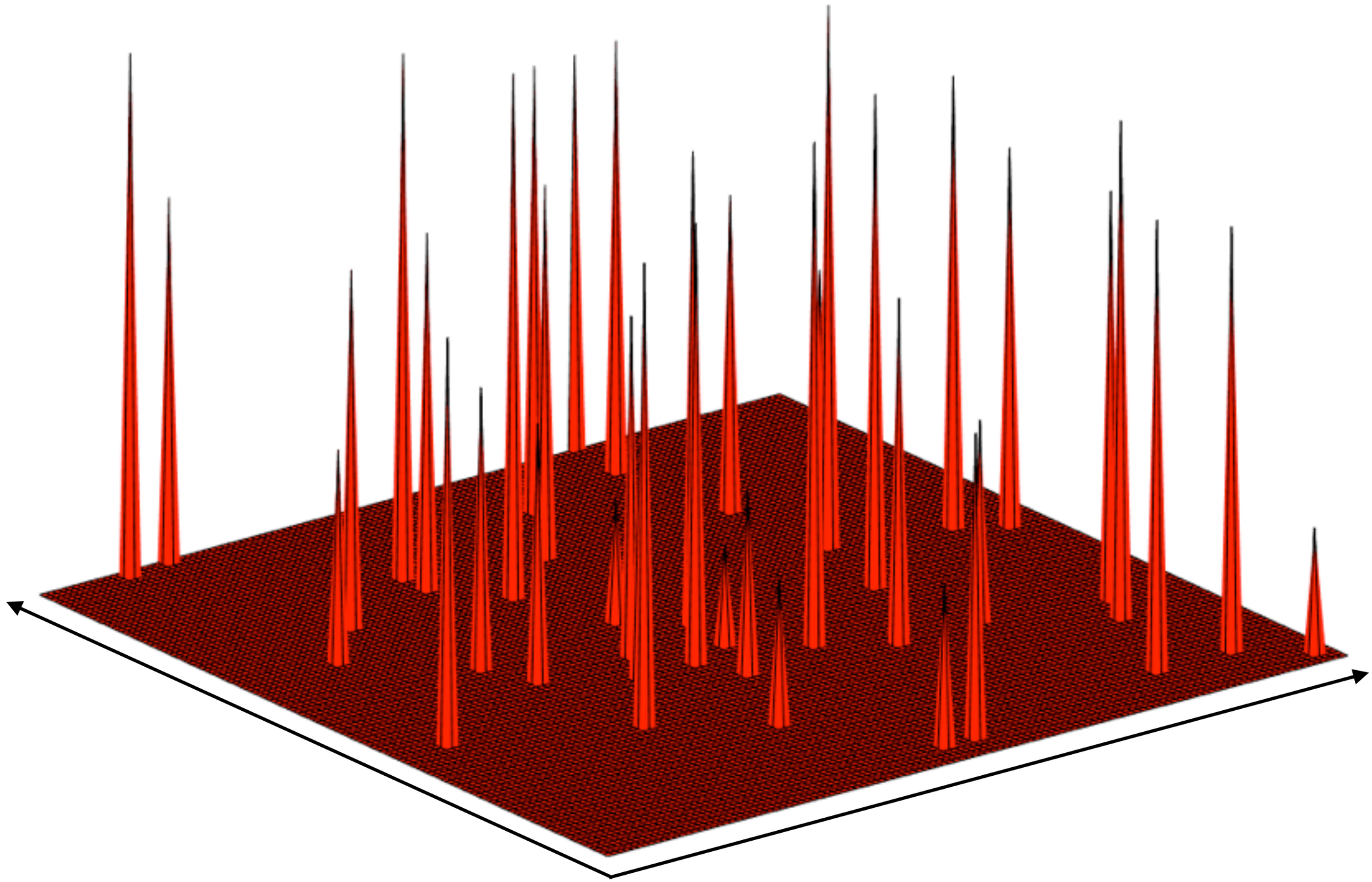


Topography of the fitness landscape



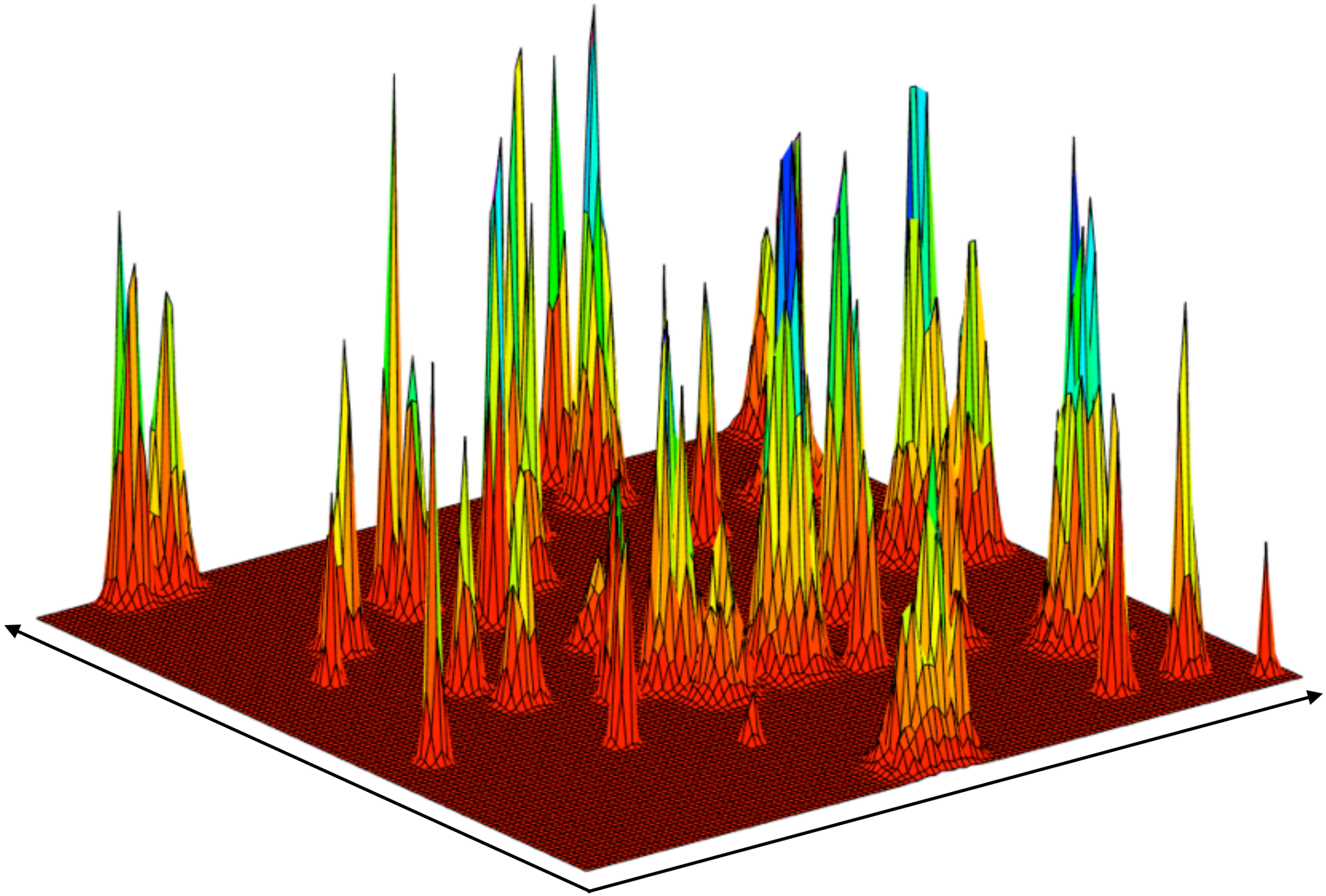


What about the local landscape?



sequence space

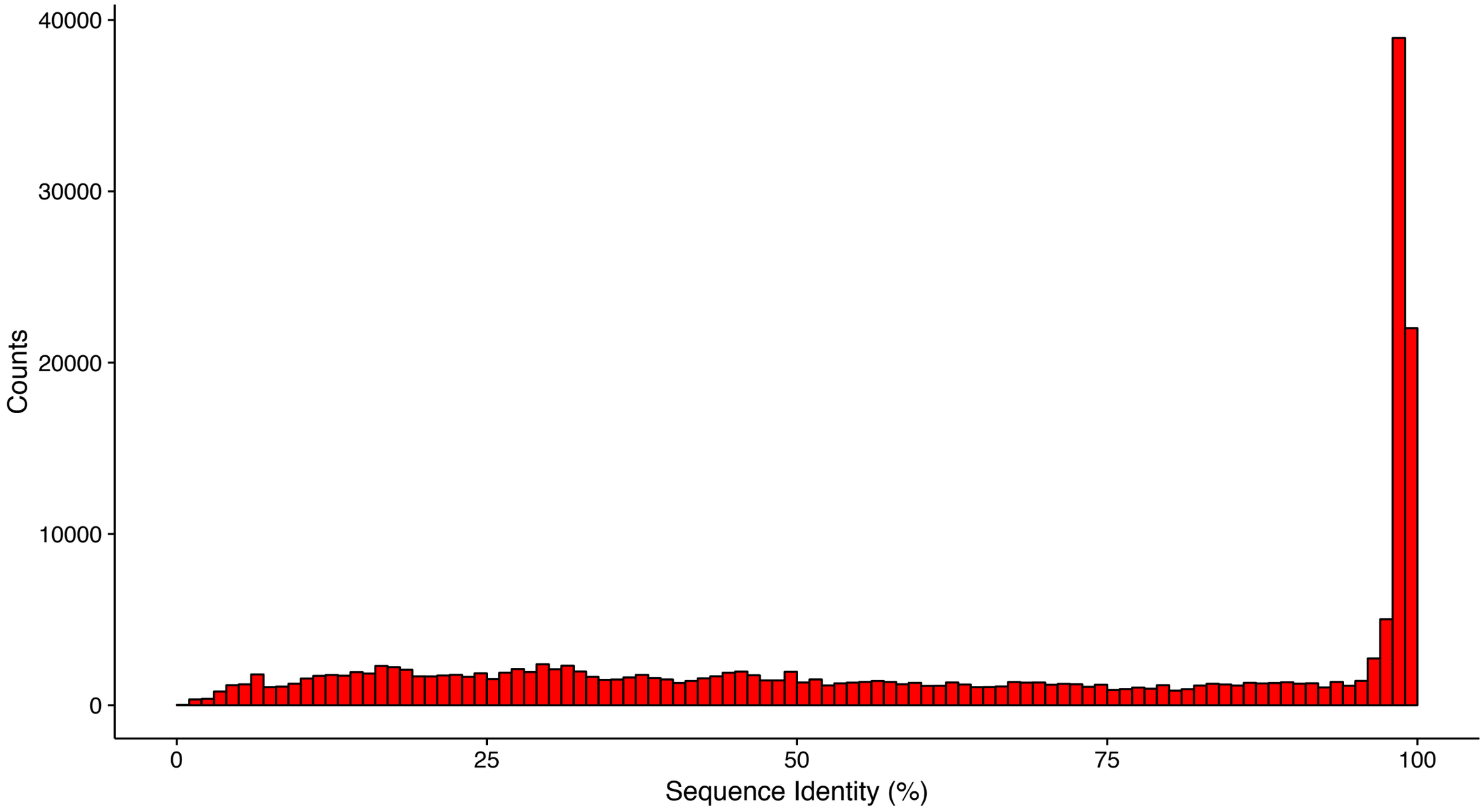
What about the local landscape?



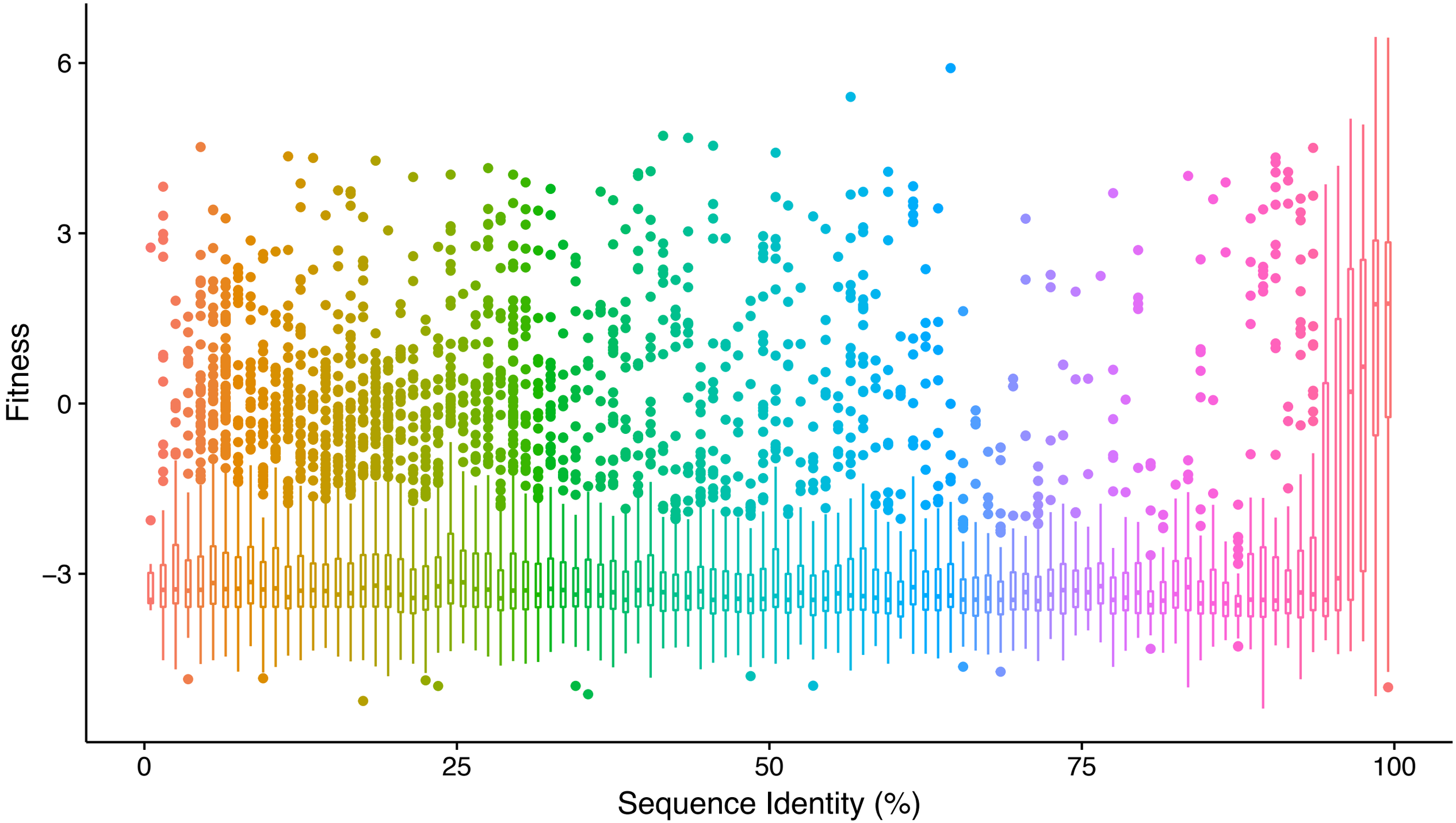
sequence space

Mutant Distribution

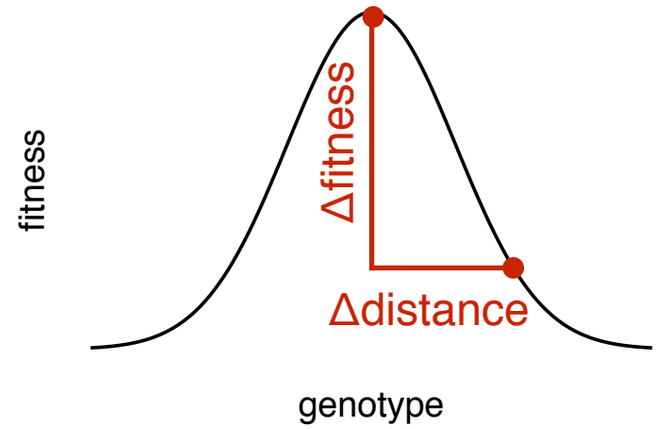
>75,000 mutants within distance 20 a.a.
>1,000,000 mutants below read threshold

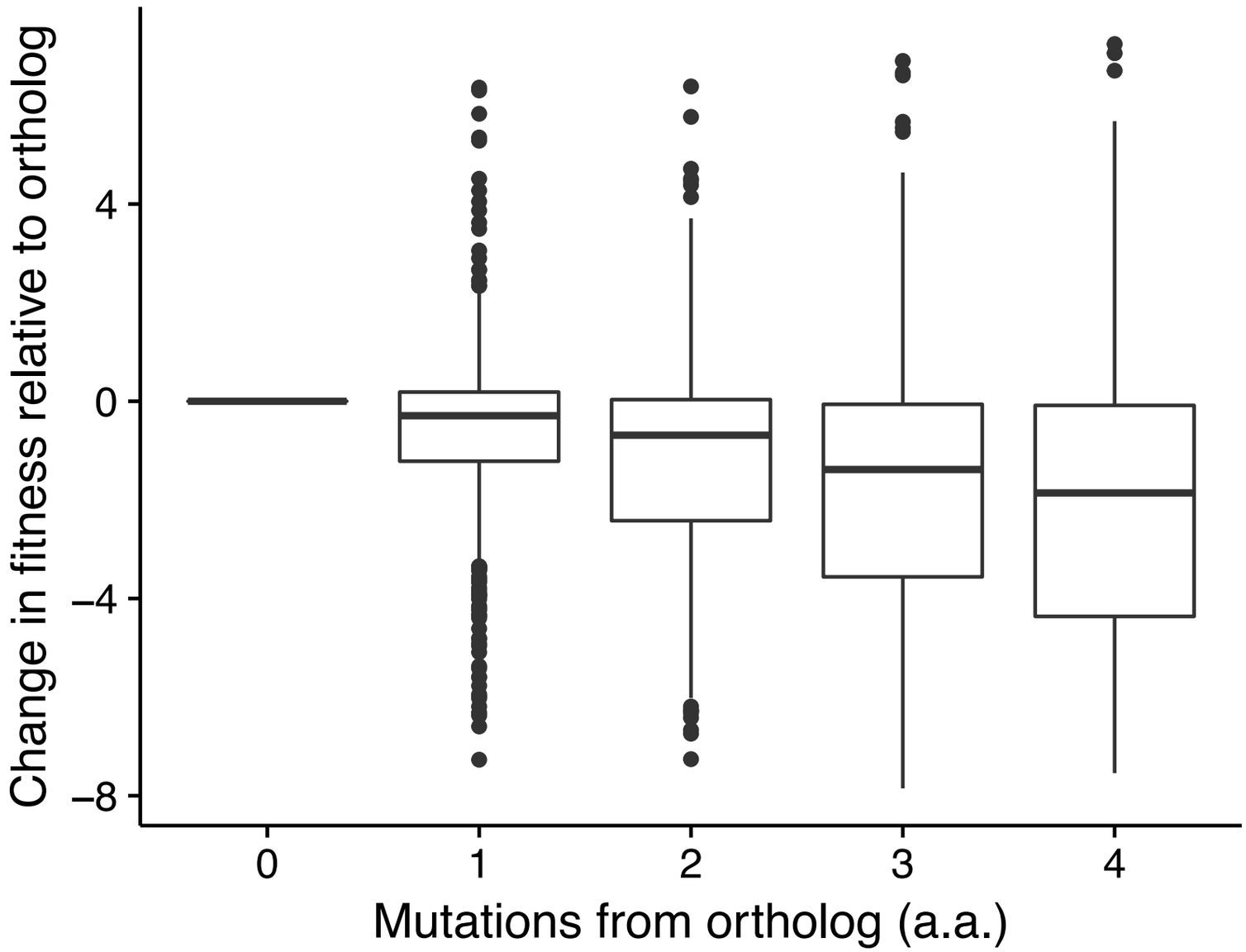


Mutant Fitness



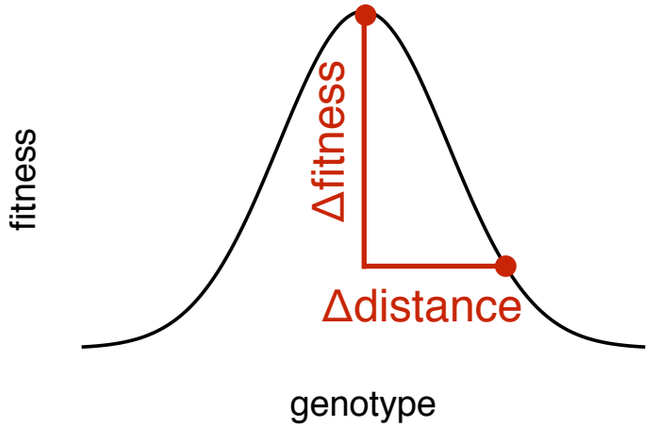
Mutant Fitness

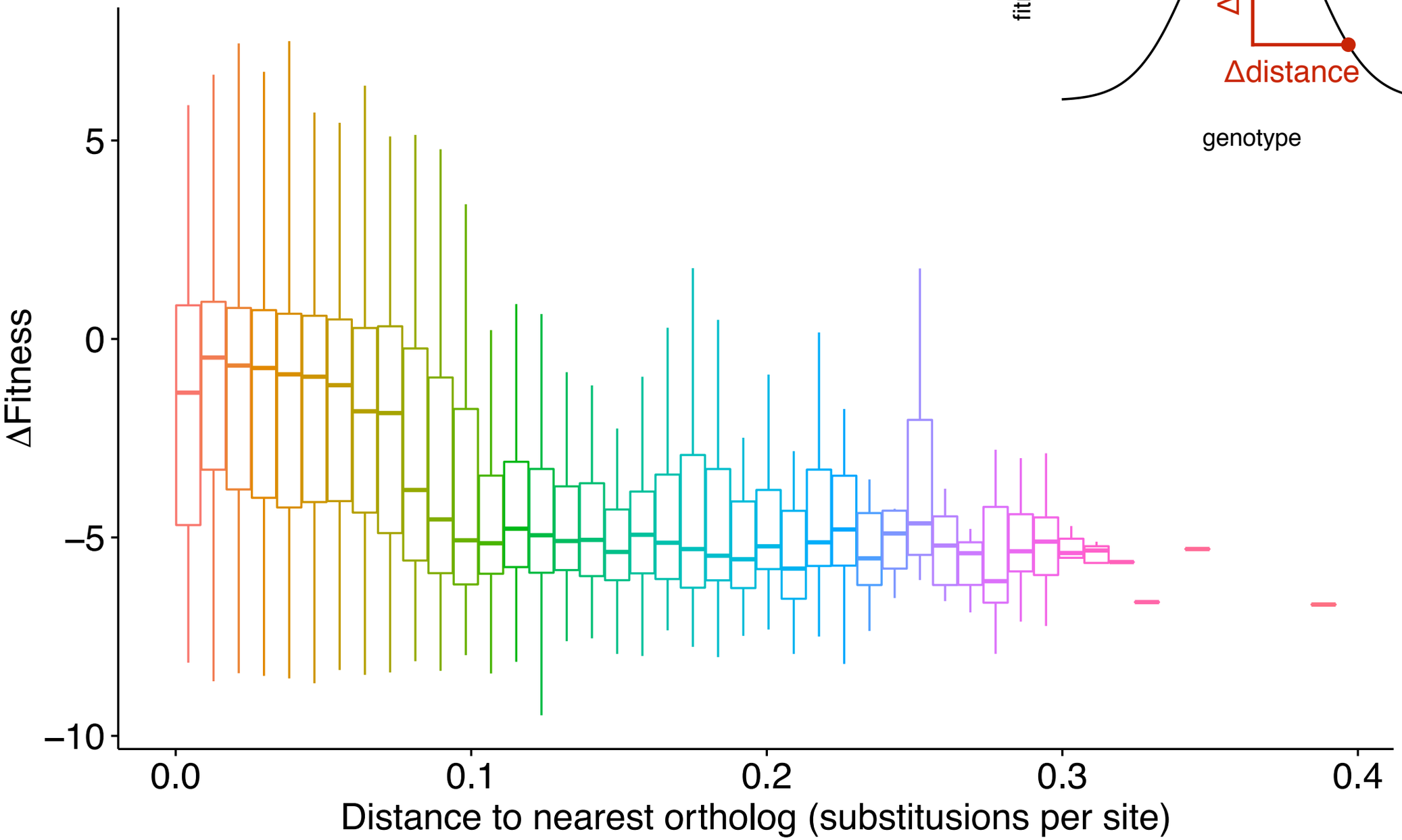




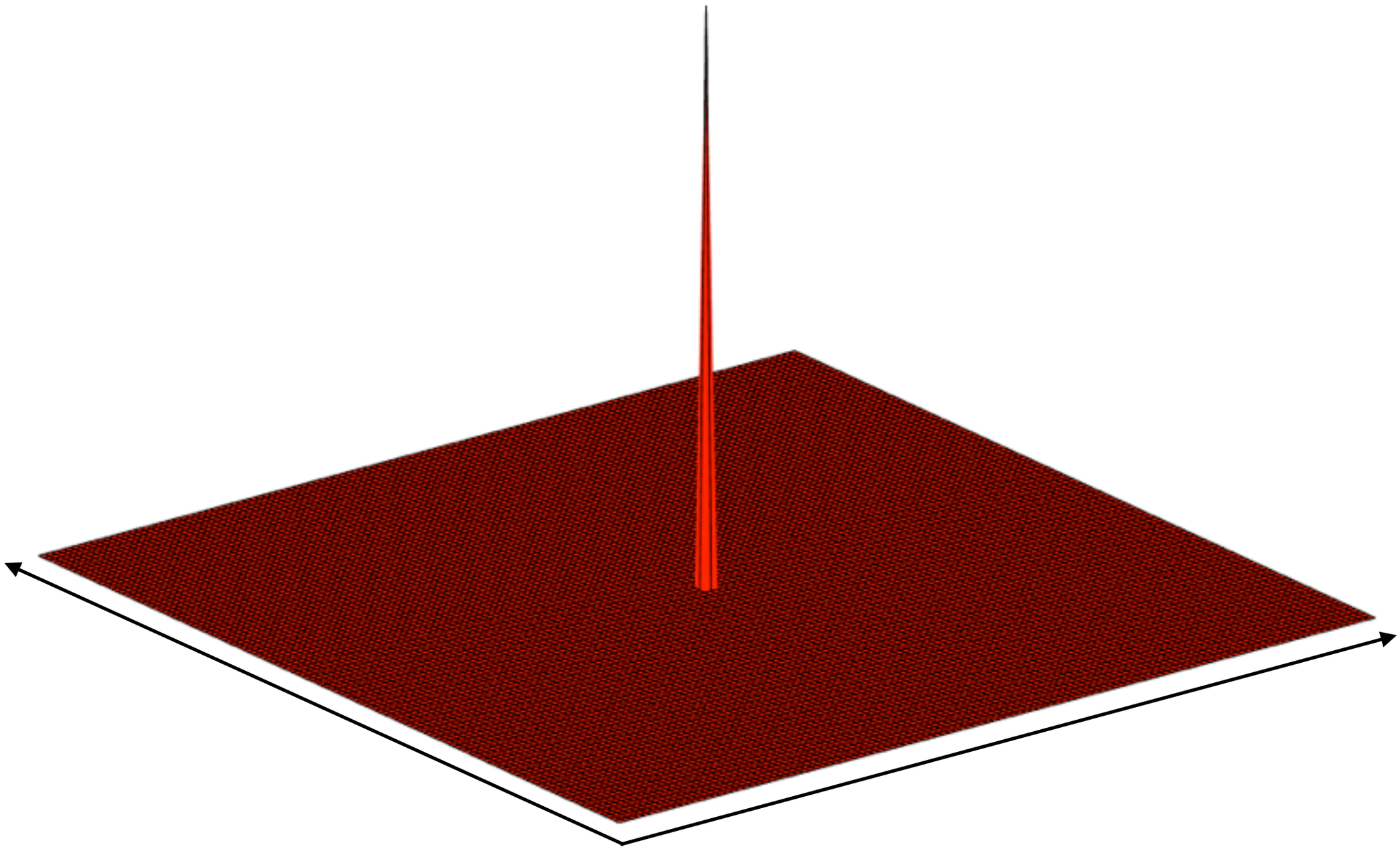
Local Peak Landscape

around 360 peaks

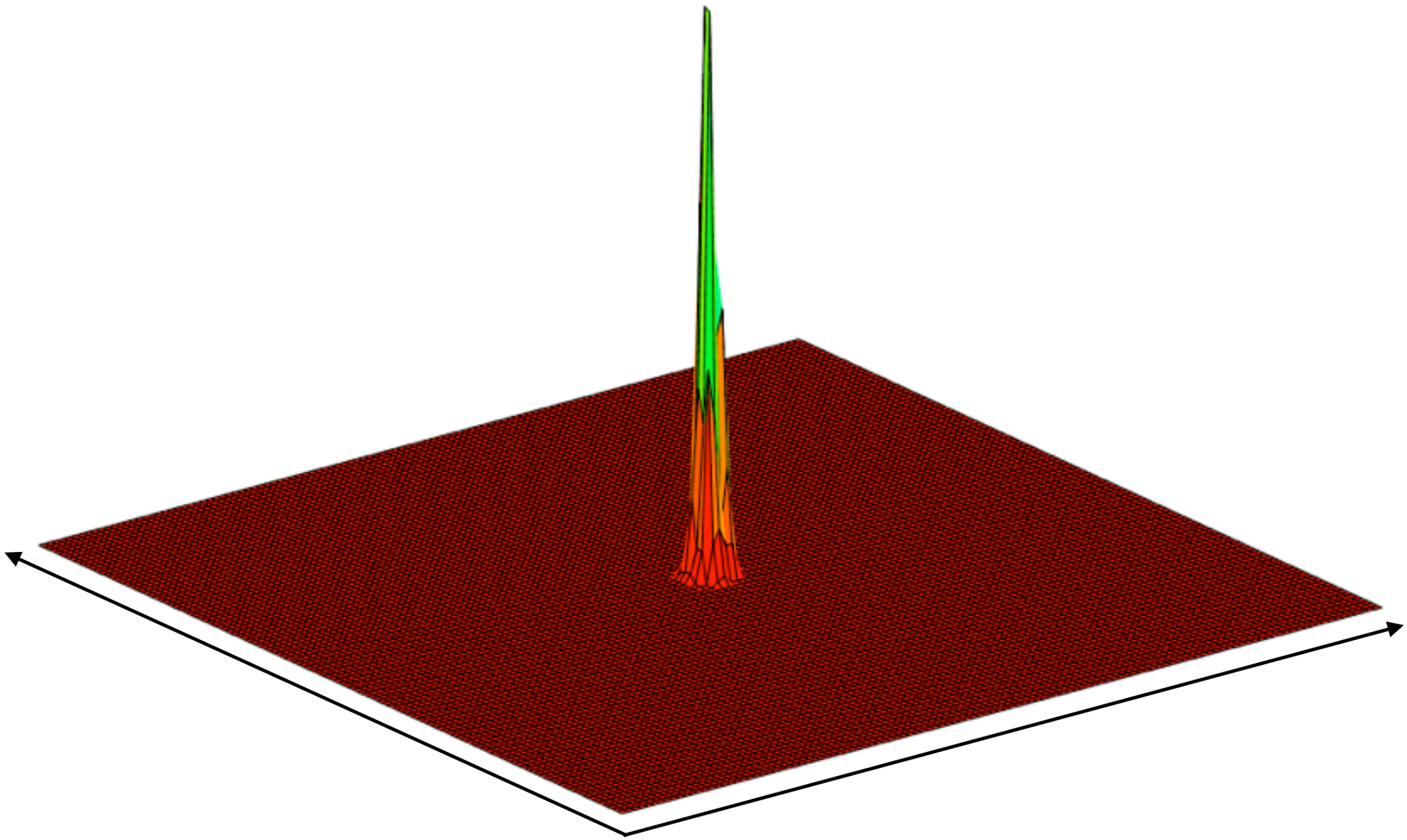




Single sequence



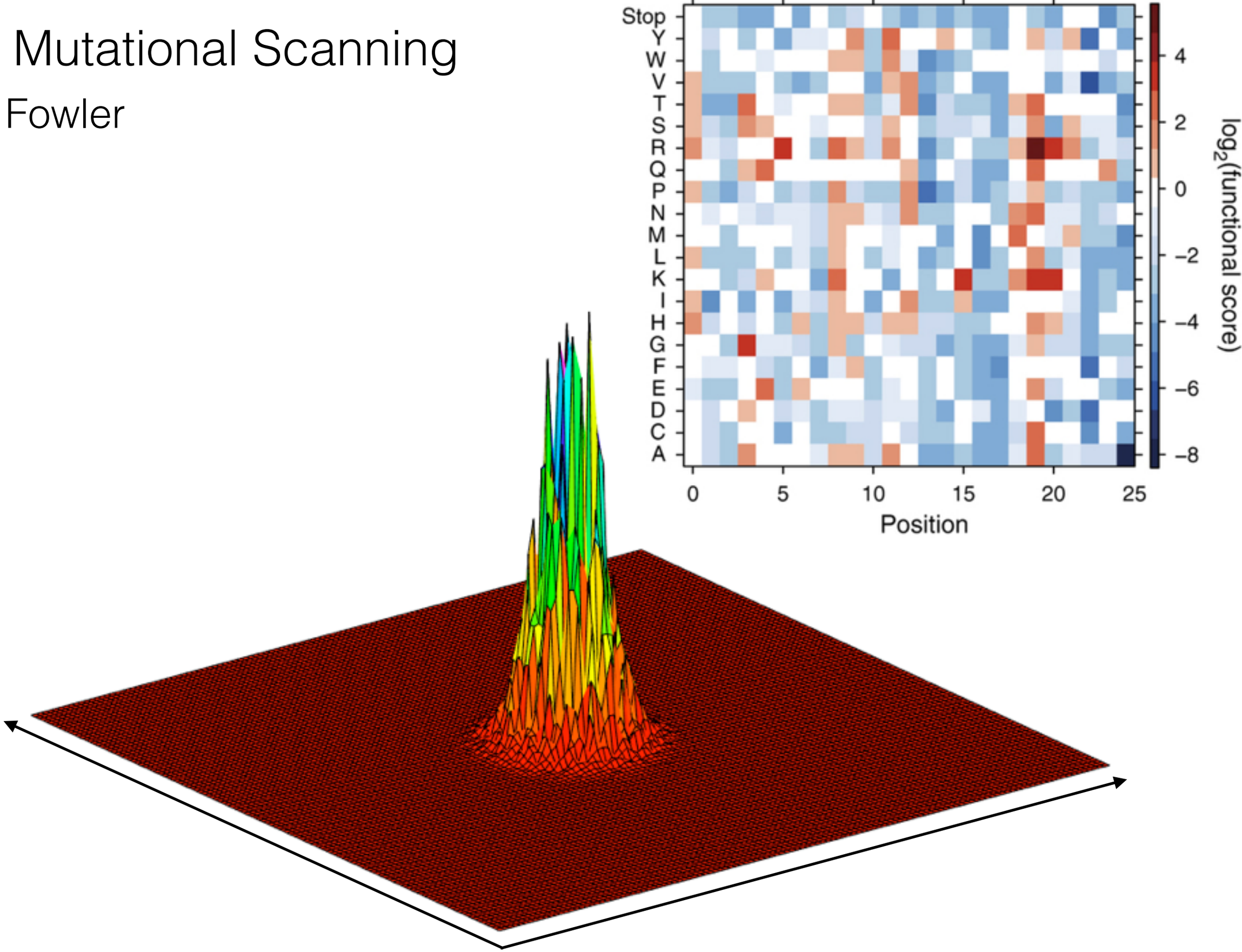
Mutagenesis



sequence space

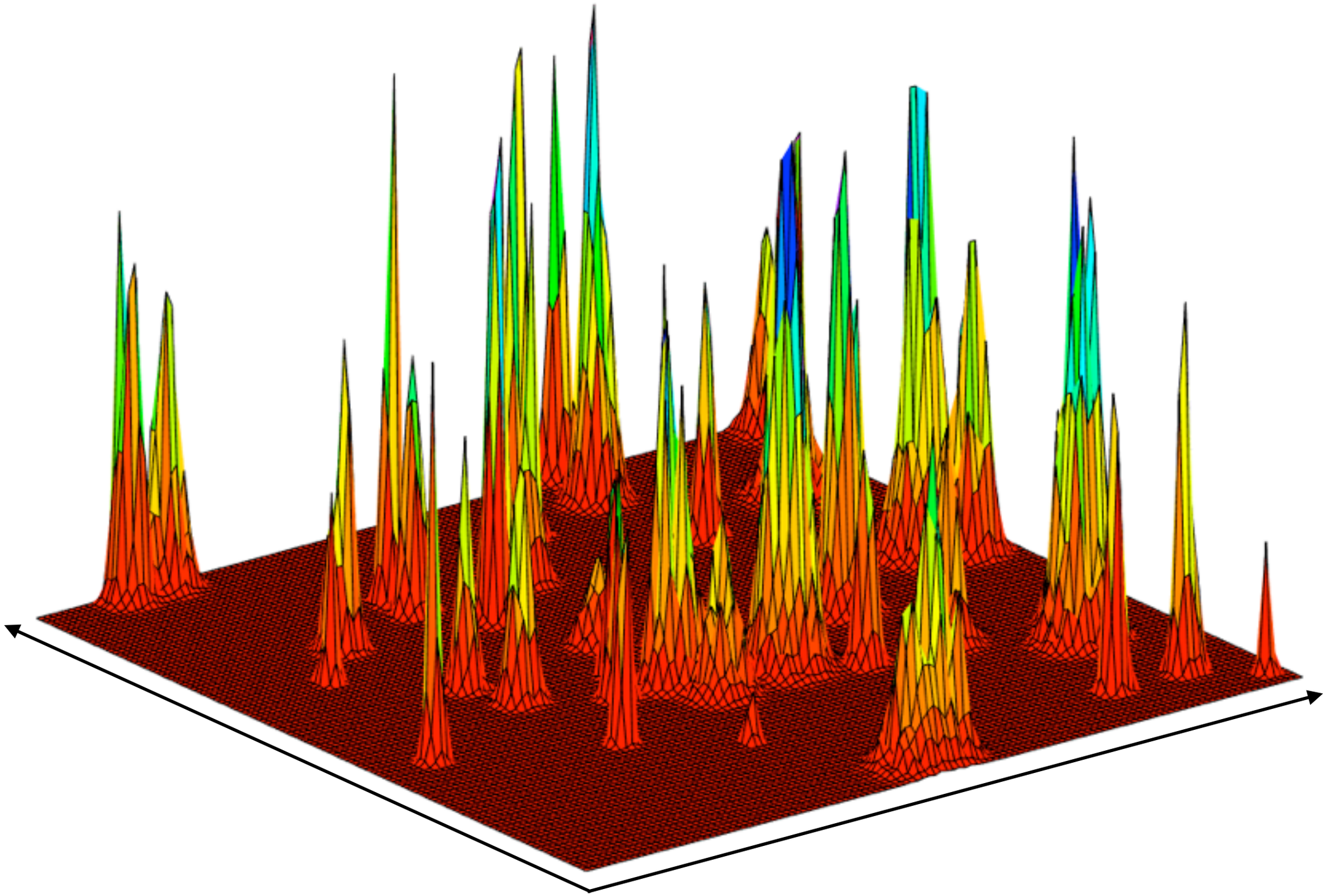
Deep Mutational Scanning

2010 - Fowler



Fowler & Fields, Nature Methods 11, 801–807 (2014)

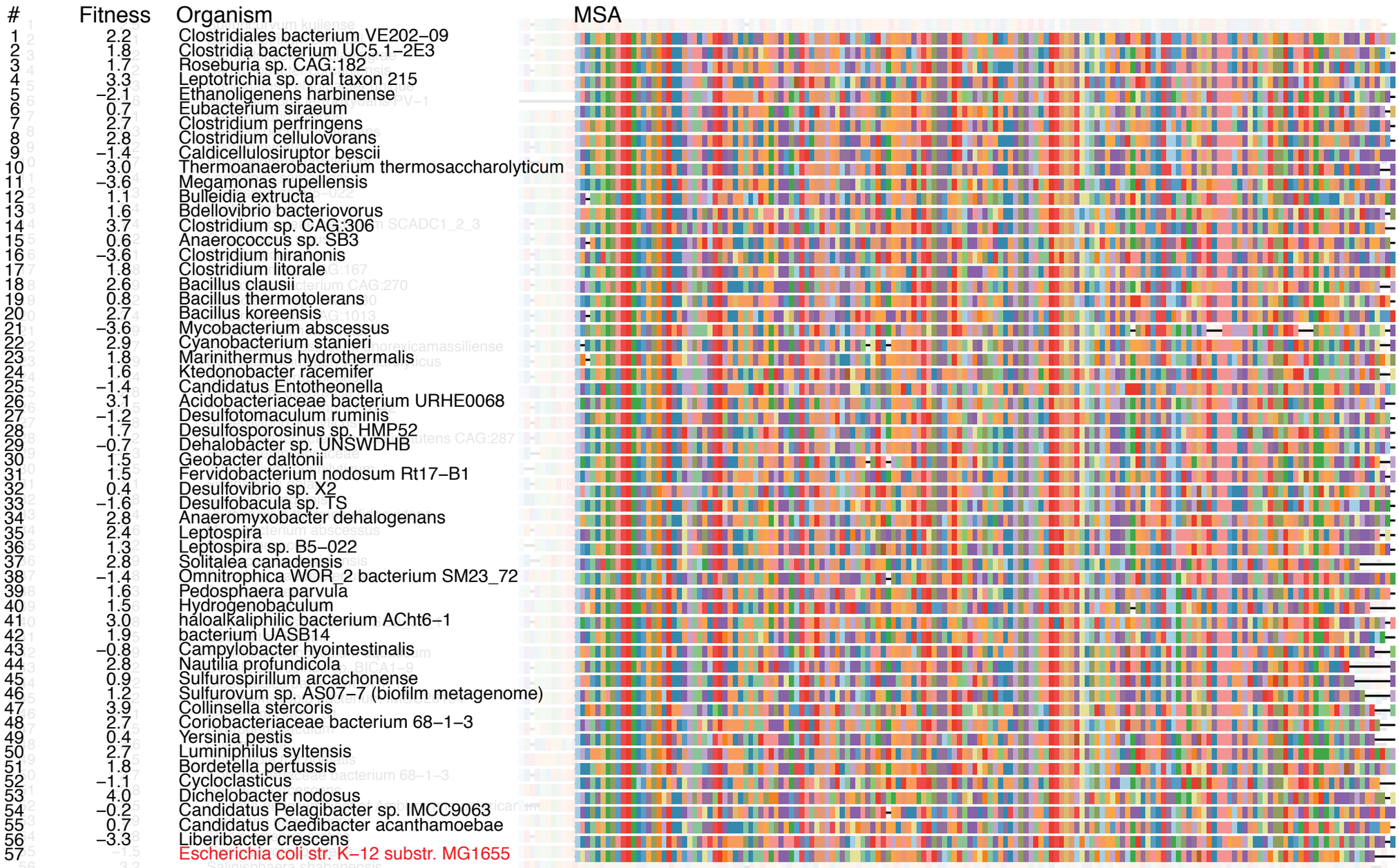
Broad Mutational Scanning



sequence space

Broad Mutational Scanning

1) Select 497 complementing homologs (and their 71,061 mutants)
2) Multiple sequence alignment
3) Collapse fitness for orthologs and their mutants onto E.coli reference sequence and determine mean fitness at each position & a.a. combination.



EVmut.
Cons.
Rel. Sol. Acc.
Sec. Struct.
Del.
Avg. Fit.

β1 α1 β2 α2 β3 α3 β4 α4 β5 α5 α6 α7 α8

Amino acid

hydrophobic
hydrophilic
arom.
aliphatic
polar
neg.
pos.

W F Y P M I L V A G C S T Q N D E H R K

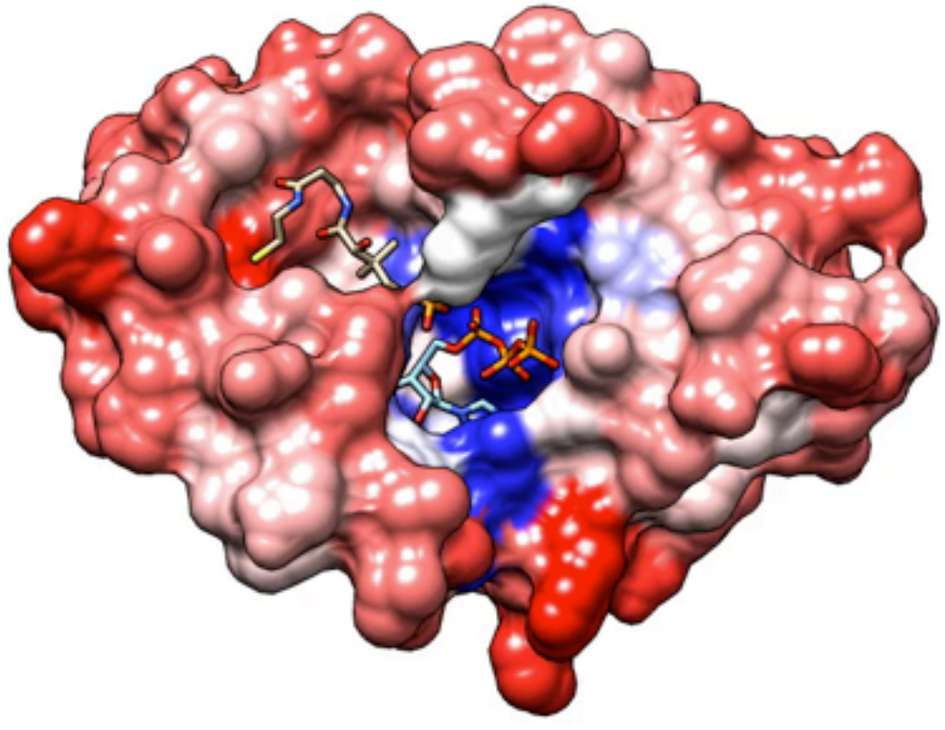
0 30 60 90 120 150

ATP
TxGH motif
4'-phosphopantetheine
4'-phosphopantetheine
ATP

Position (aa)

6 3 0 –3 Fitness

WT *E.coli* aa
No data



2.7 0 -2.5 Avg. Fitness

EVmut.
Cons.
Rel. Sol. Acc.
Sec. Struct.
Del.
Avg. Fit.

β1 α1 β2 α2 β3 α3 β4 α4 β5 α5 α6 α7 α8

Amino acid

hydrophobic
hydrophilic
arom.
aliphatic
polar
neg.
pos.

W F Y P M I L V A G C S T Q N D E H R K

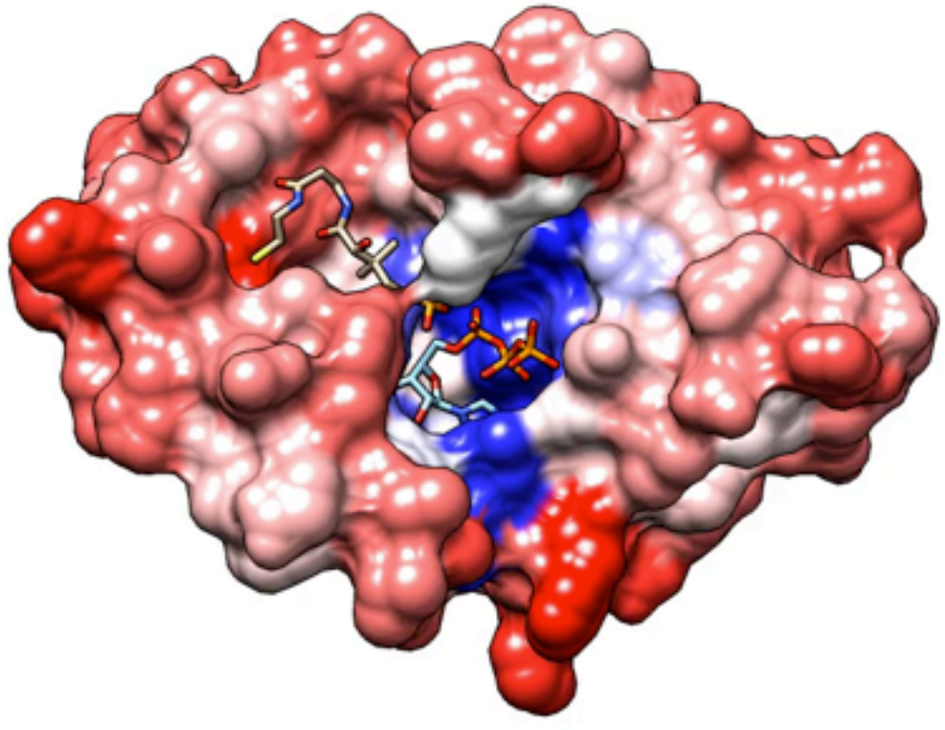
0 30 60 90 120 150

ATP
TxGH motif
4'-phosphopantetheine
4'-phosphopantetheine
ATP

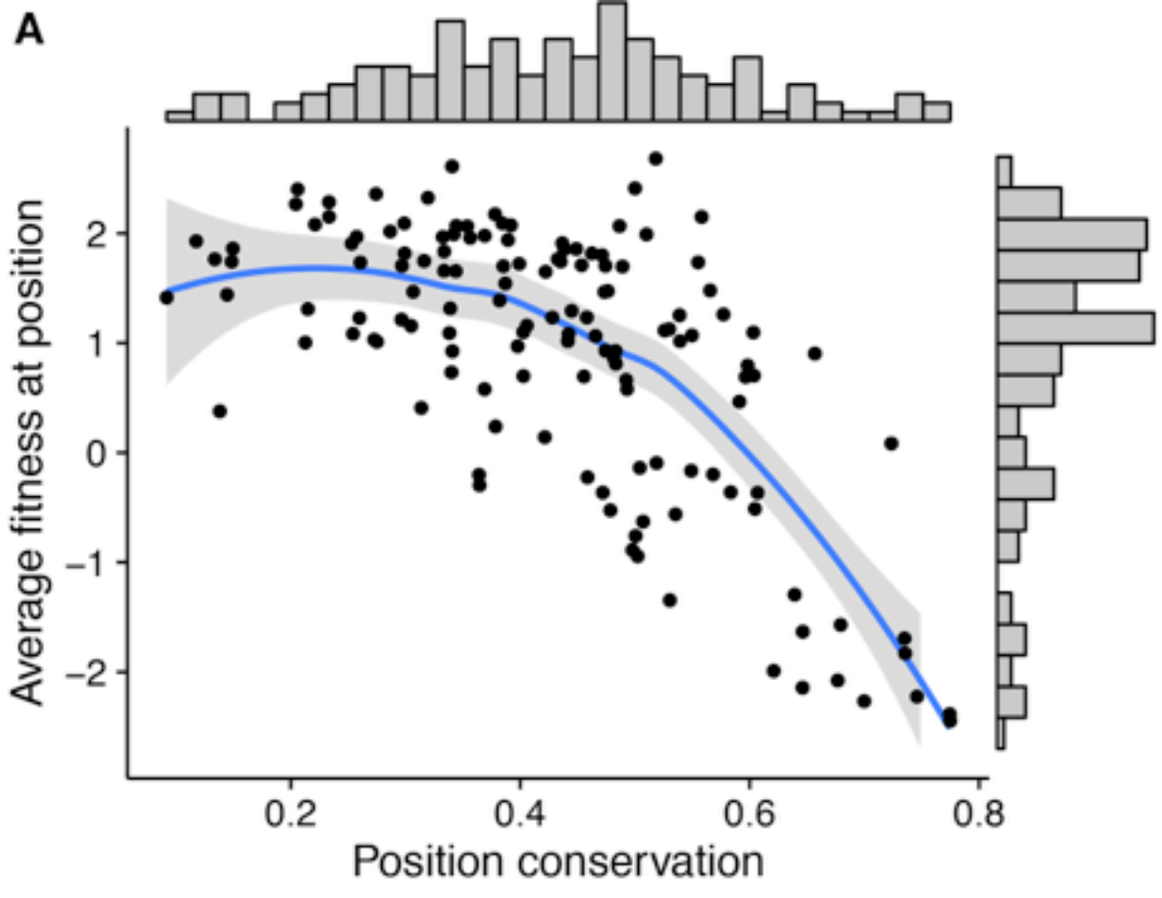
Position (aa)

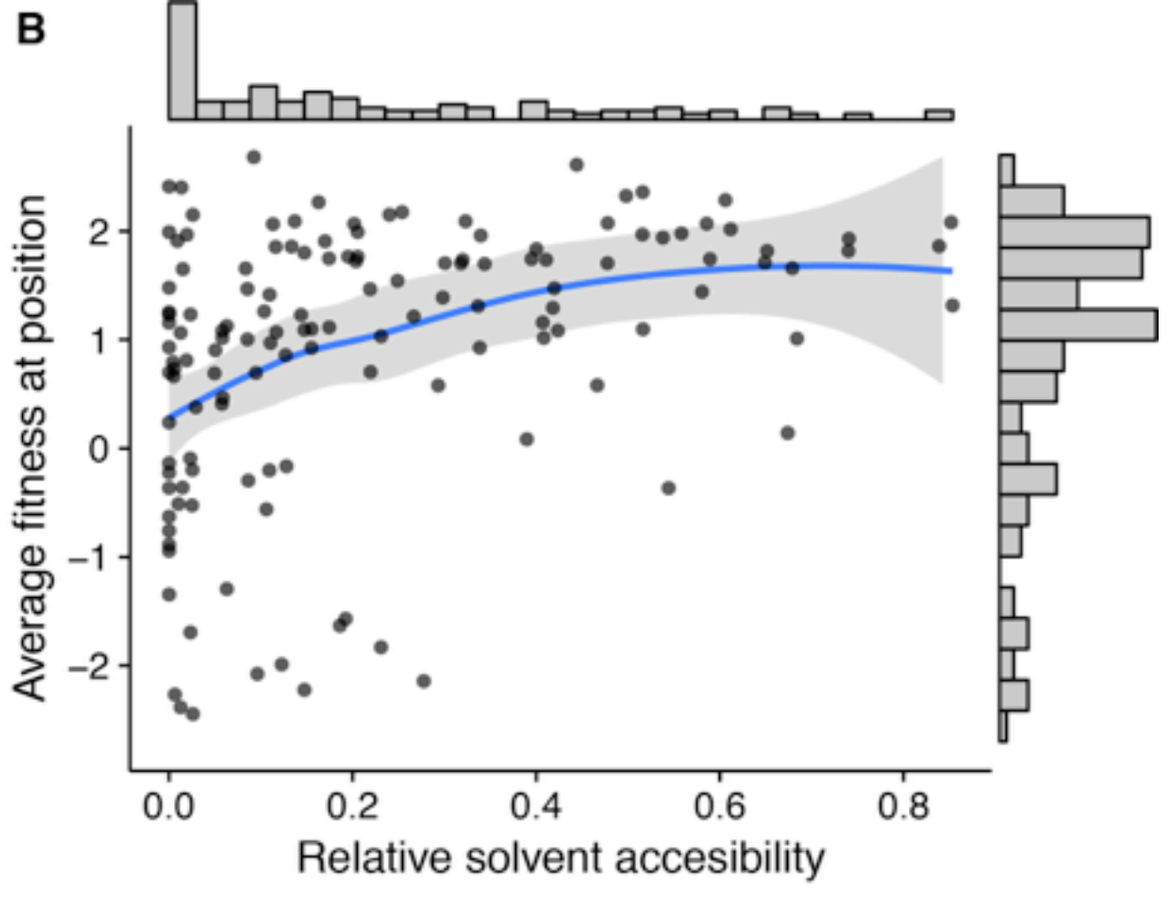
6 3 0 –3 Fitness

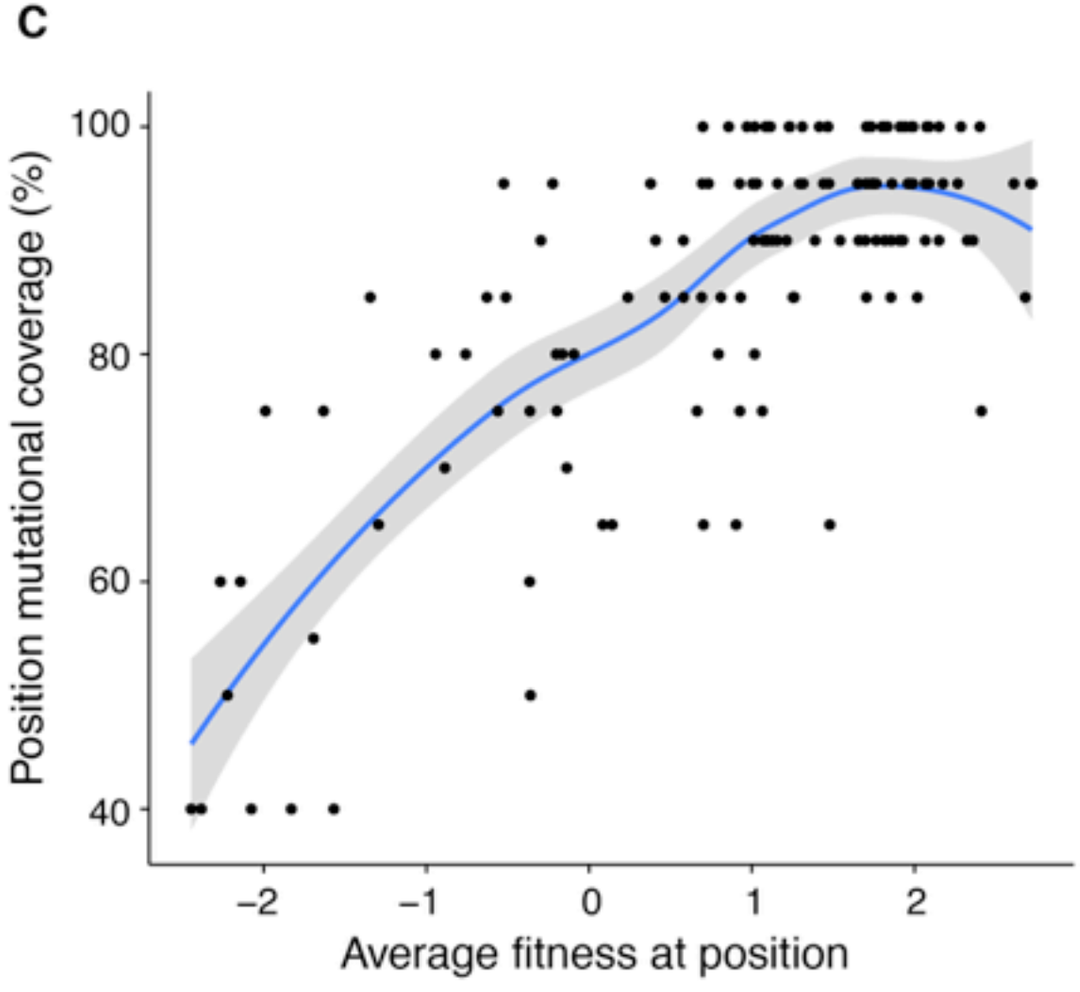
WT *E.coli* aa
No data

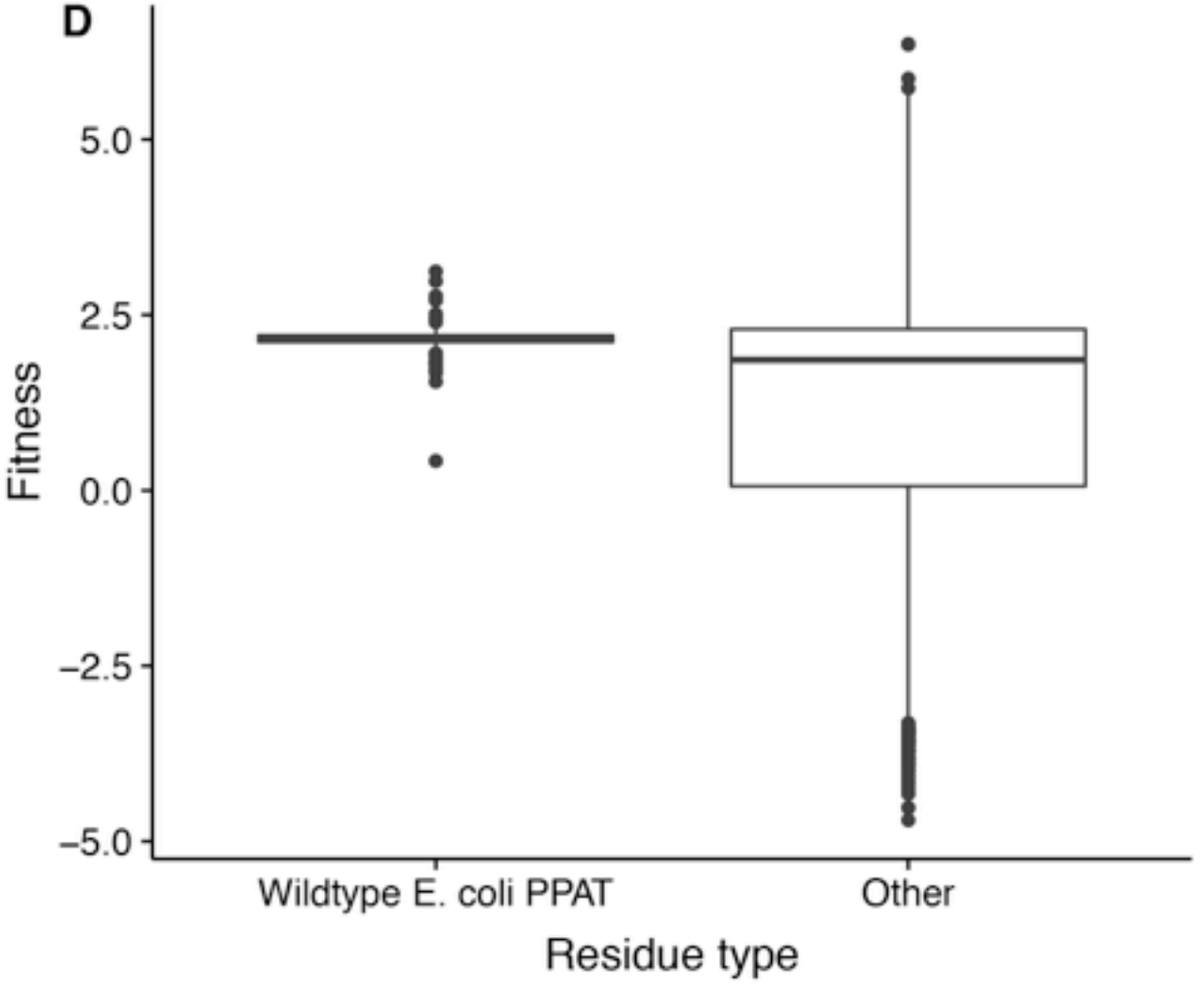


2.7 0 -2.5 Avg. Fitness

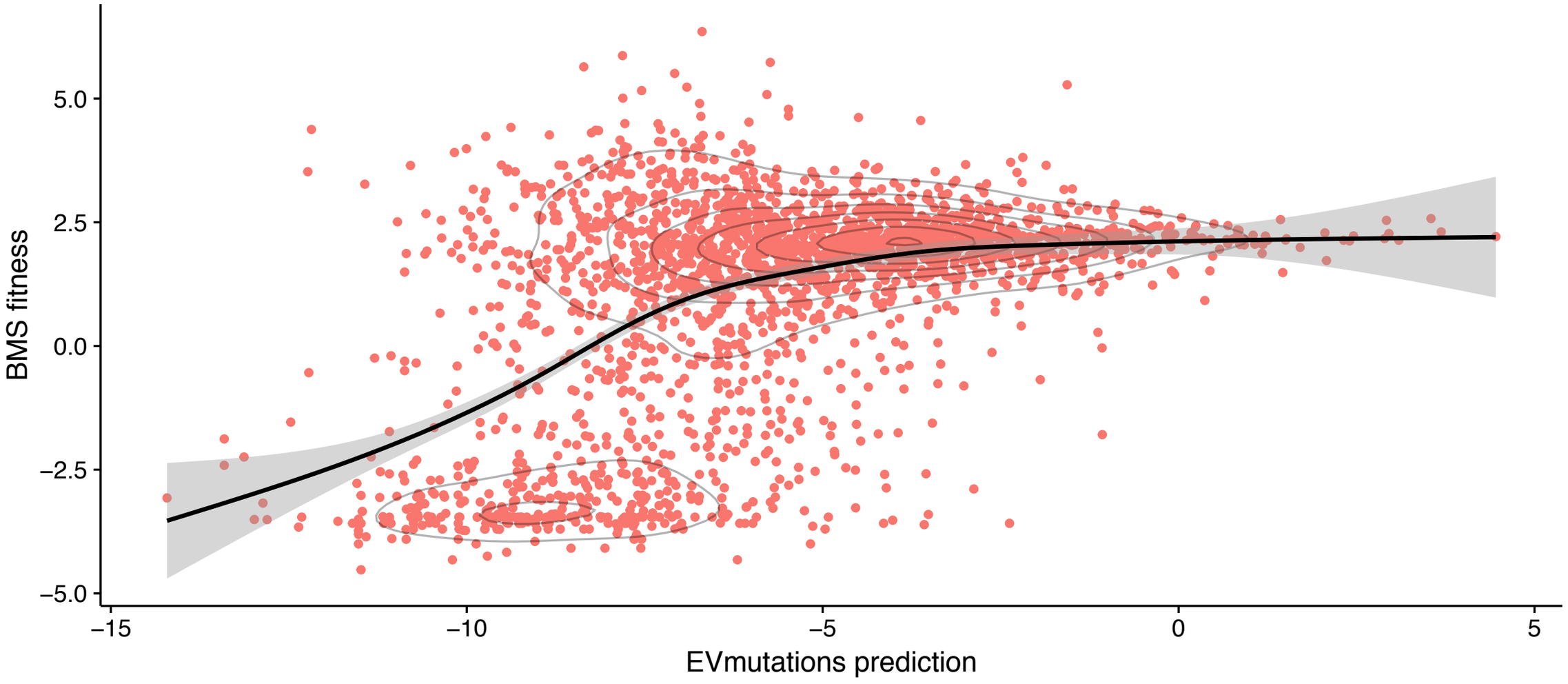

A
Average fitness at position
2
1
0
−1
−2
0.2
0.4
0.6
0.8
Position conservation

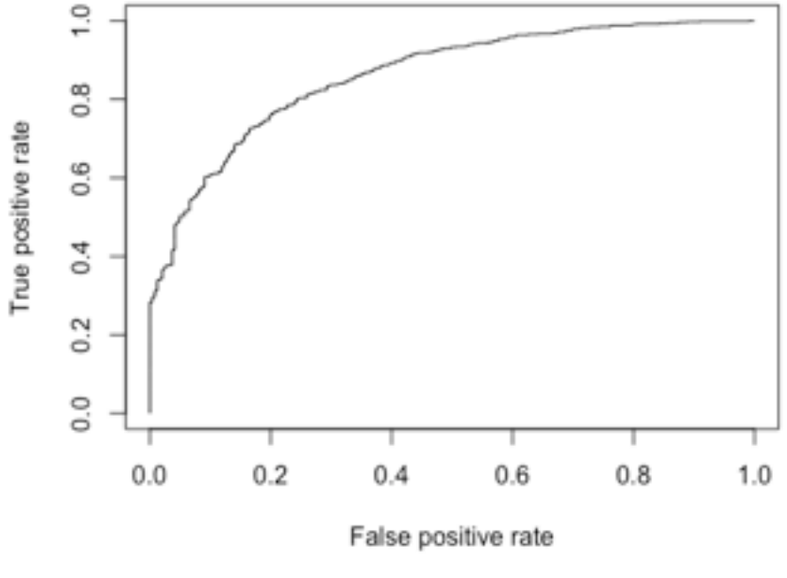

B
Average fitness at position
2
1
0
−1
−2
0.0
0.2
0.4
0.6
0.8
Relative solvent accesibility


C
Position mutational coverage (%)
100
80
60
40
−2
−1
0
1
2
Average fitness at position

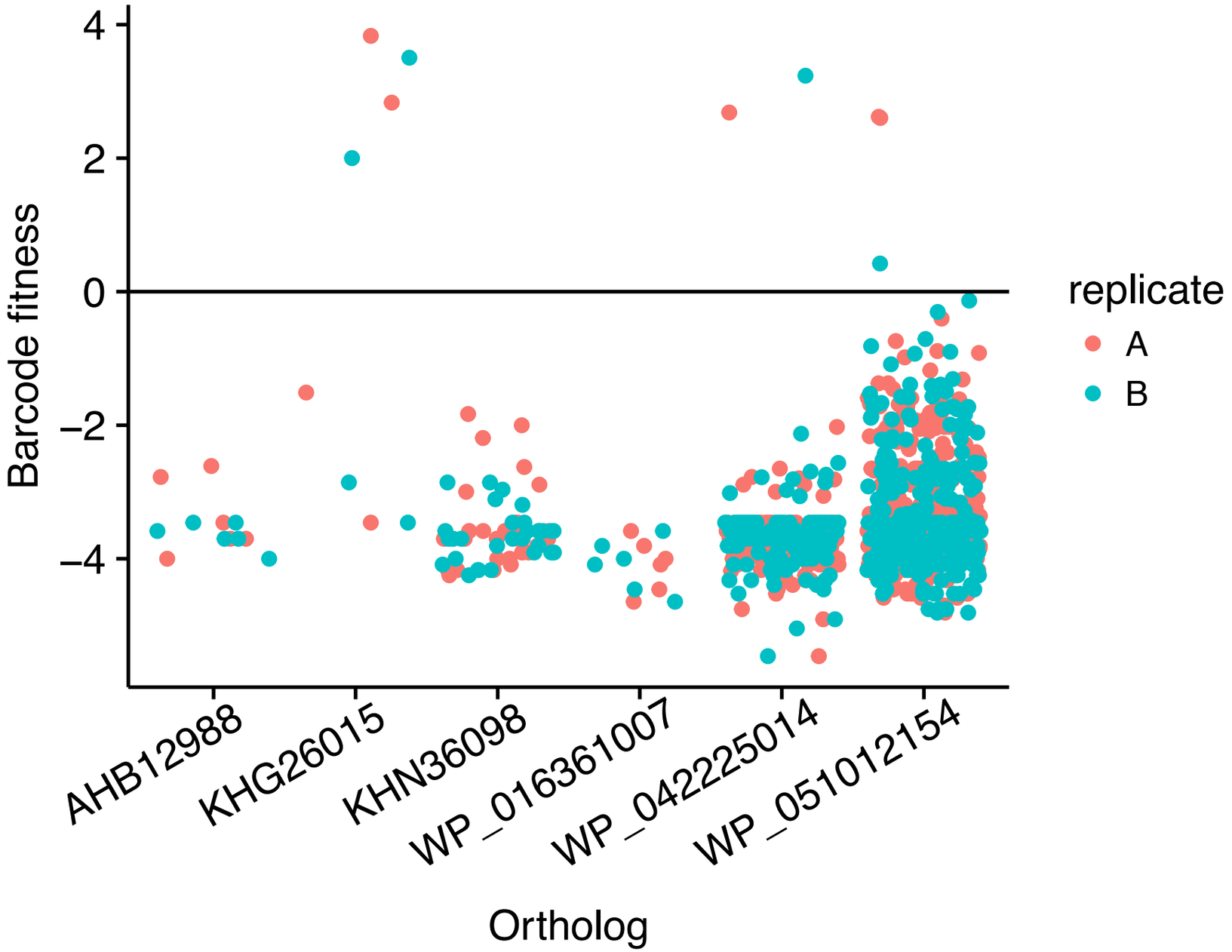

D
Fitness
5.0
2.5
0.0
−2.5
−5.0
Wildtype E. coli PPAT
Other
Residue type

BMS compared to EVmutations prediction





Individual Barcode Noise in Negative Controls

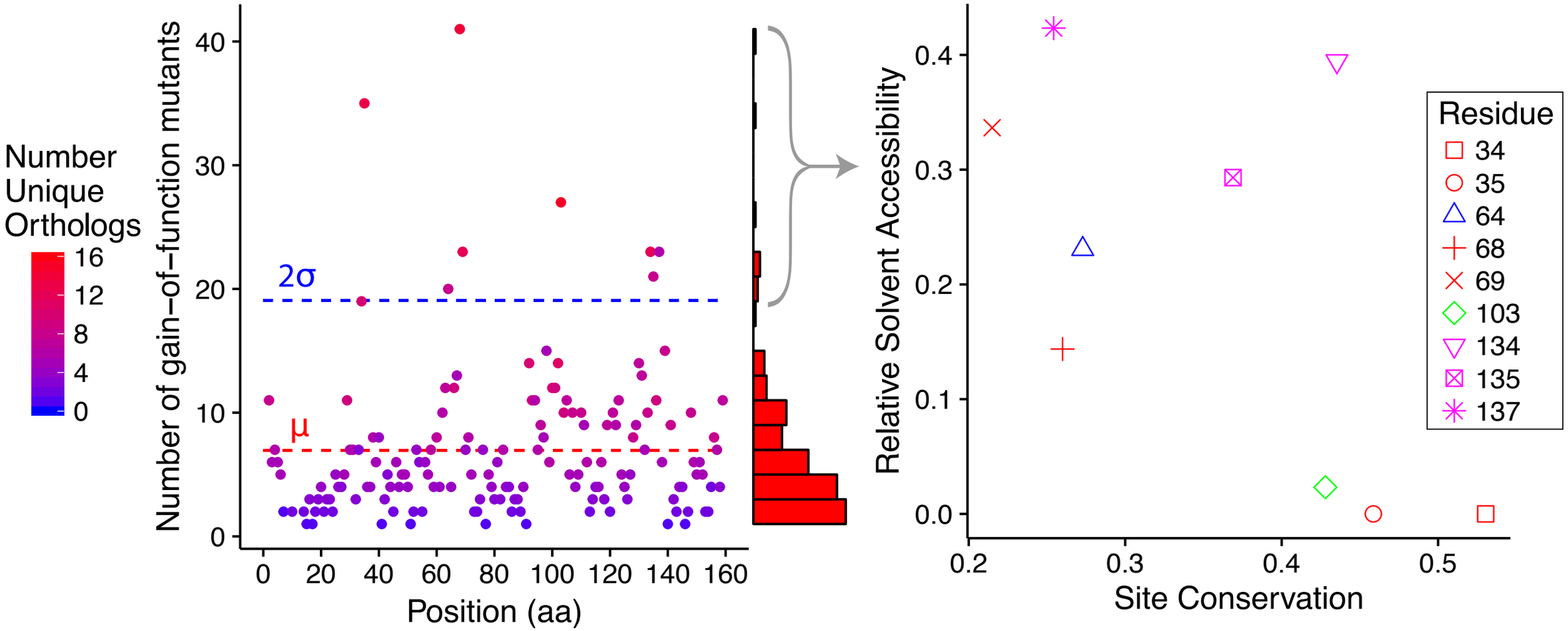


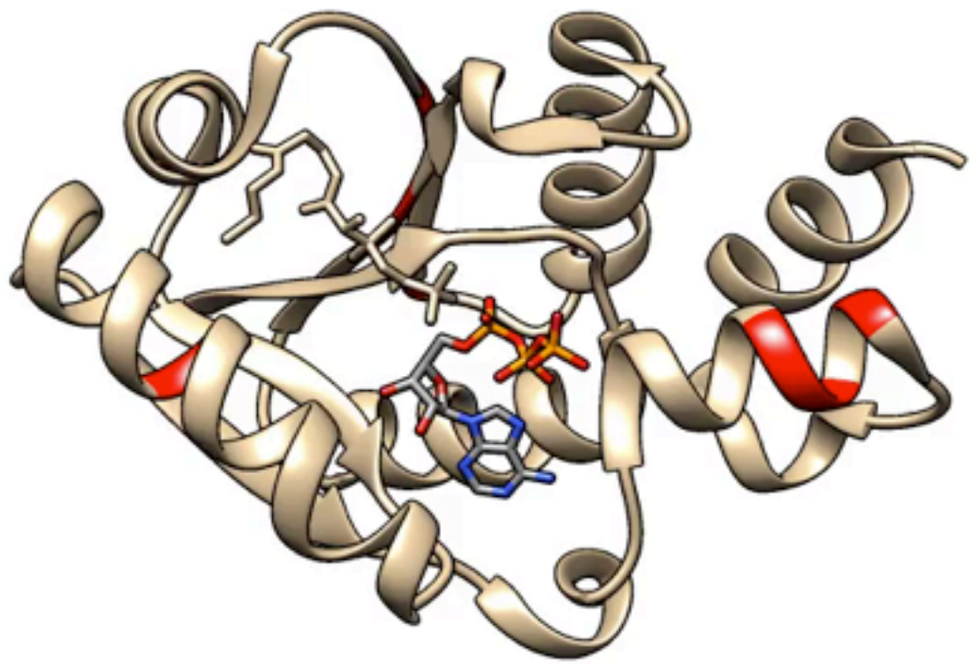
9 BCs with positive fitness out of 994 total.
False positive rate of 0.9%

Gain of Function Mutations for Low-Fitness Homologs

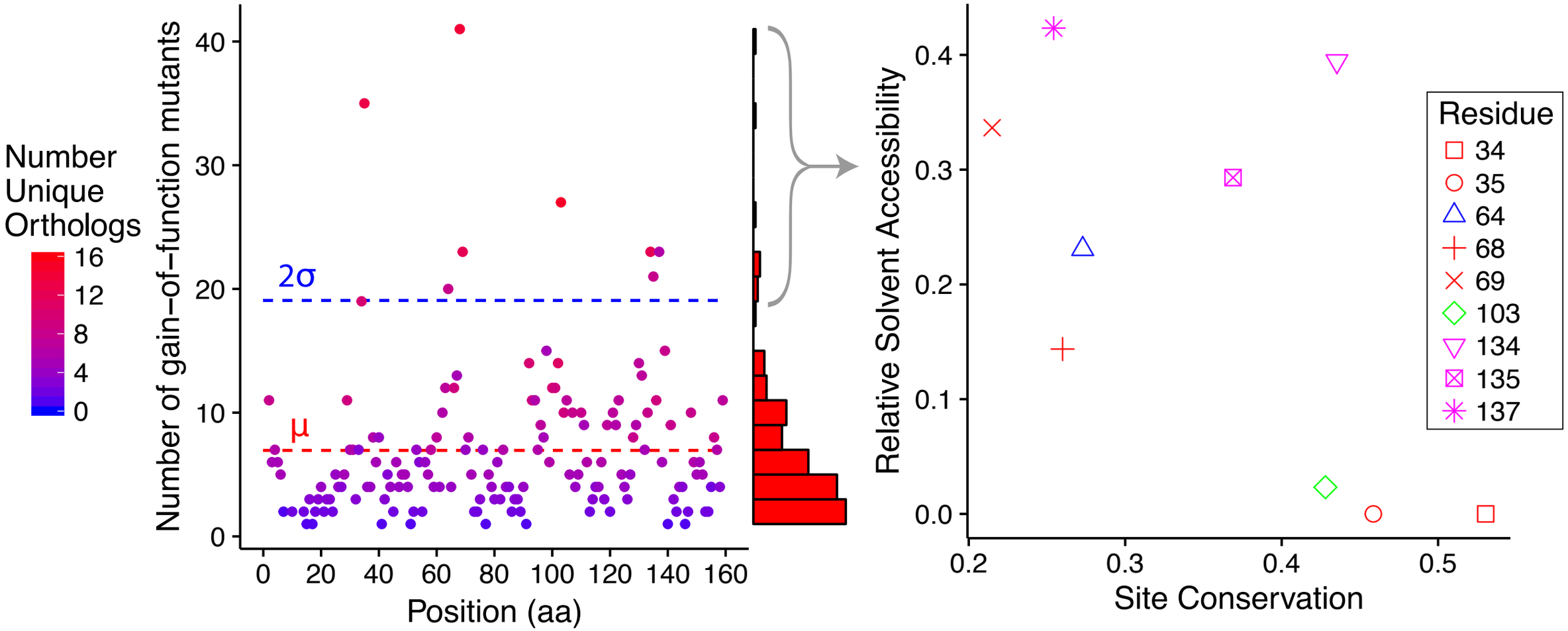
1) Select 129 low-fitness orthologs (fitness < -2.5)
2) Select mutants within 5 a.a. with positive fitness (GoF)
3) Found 569 GoF mutants (out of 4,658) across 72 dropout orthologs
4) Multiple sequence alignment
5) Collapse fitness for GoF mutants onto E.coli reference sequence

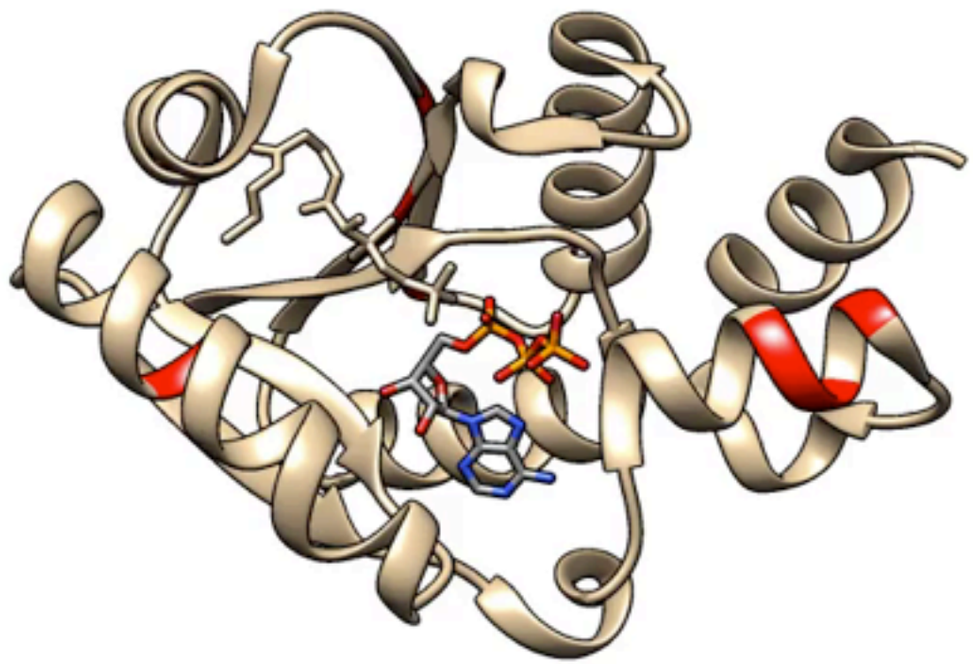
Gain of Function Mutations for Low-Fitness Homologs



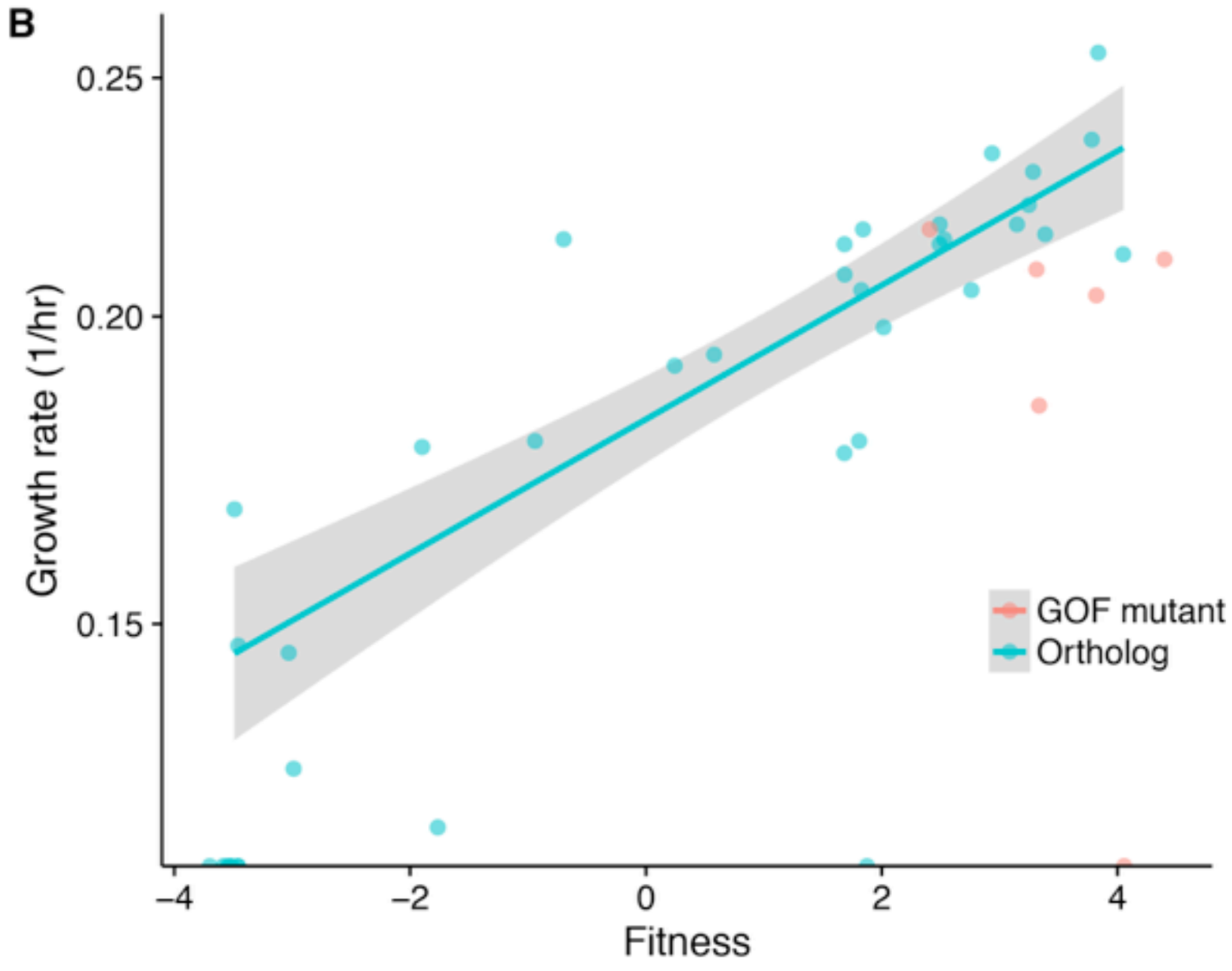


Gain of Function Mutations for Low-Fitness Homologs

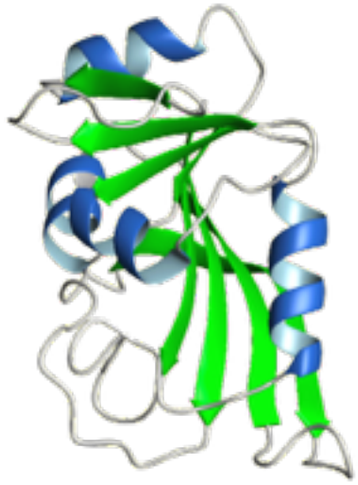




Individual Fitness Testing

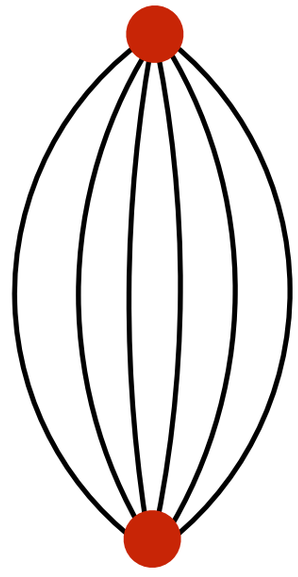


Up next



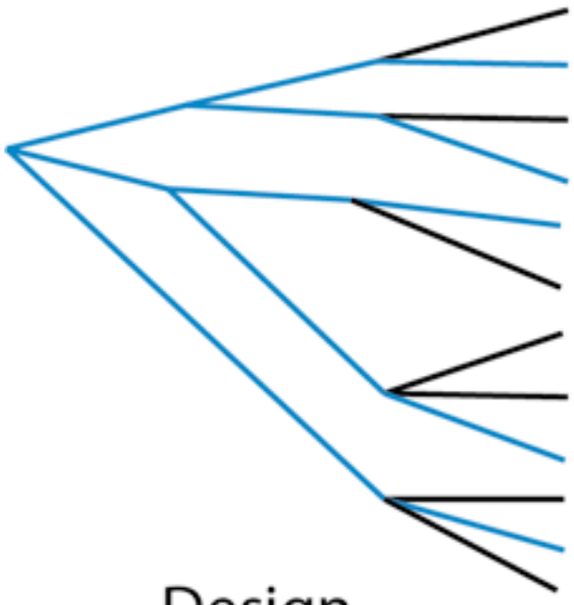
- >5,000 DHFR homologs
  - resistome mapping
  - emergence of antibiotic resistance

- evolutionary accessibility of mutational pathways

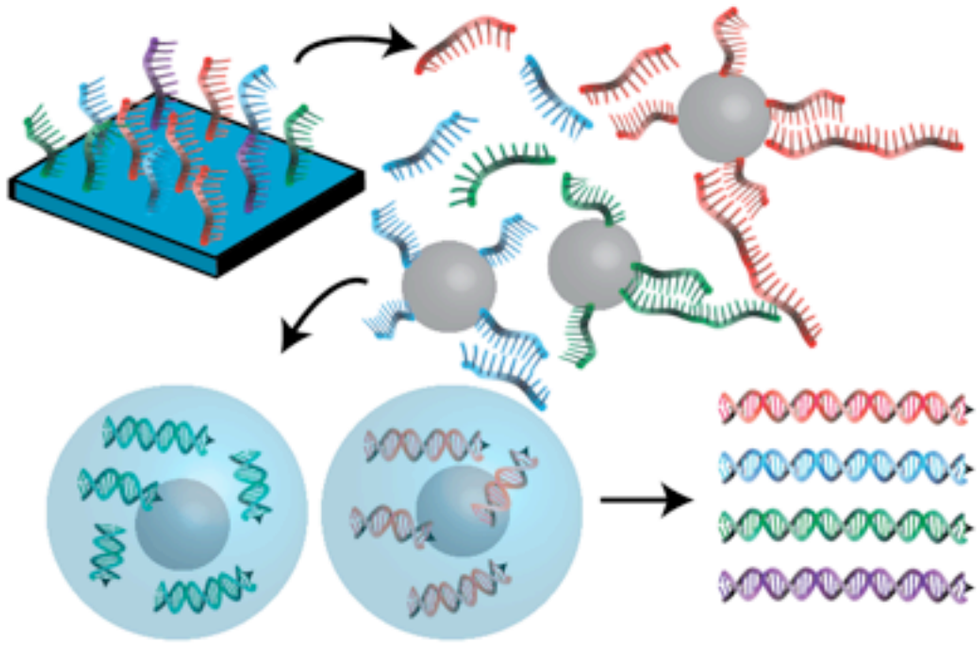


?

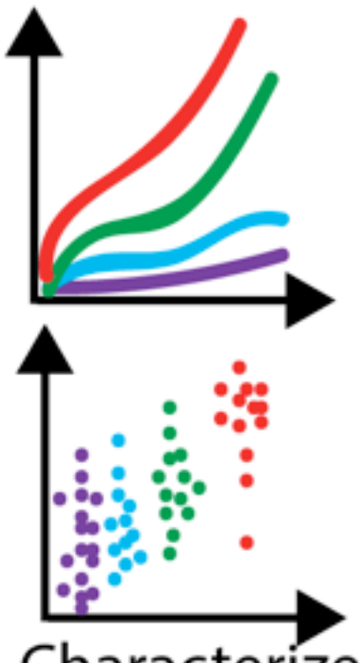
- ancestral sequence reconstruction



Design



Assemble



Characterize

Acknowledgements

- Sri Kosuri
- Angus Sidore
- Nathan Lubock
- Kosuri lab members
- Di Zhang
- George Church

UCLA CHEMISTRY & BIOCHEMISTRY

UCLA-DOE INSTITUTE

SEARLE SCHOLARS PROGRAM
FUNDING EXCEPTIONAL YOUNG SCIENTISTS

NIH DIRECTOR'S NEW INNOVATOR AWARD

NWO
Netherlands Organisation for Scientific Research

HFSP
HUMAN FRONTIER SCIENCE PROGRAM
FUNDING FRONTIER RESEARCH INTO COMPLEX BIOLOGICAL SYSTEMS

Questions?

Preprint: https://doi.org/10.1101/163550

Interested in using DropSynth?
Ideas?

plesa@ucla.edu
or
sri@ucla.edu