Economics of Functional Components in Biological and Technological Systems

Sergei Maslov

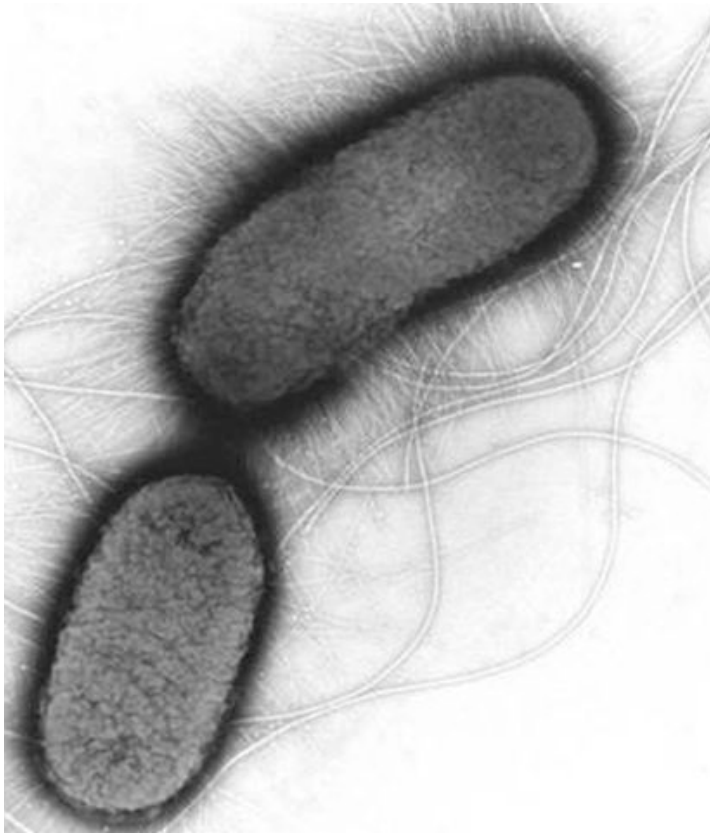
Department of Biosciences

Brookhaven National Laboratory, NY

Laufer Center for Quantitative Biology

Stony Brook University, NY

Tool-centric view of biological and technological systems

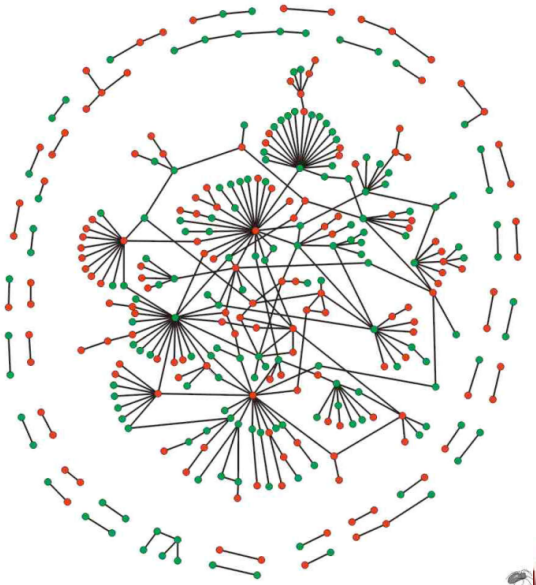


Tools: genes encoded in bacterial genomes

Tools: software installed on computers

Genes as interacting/evolving tools

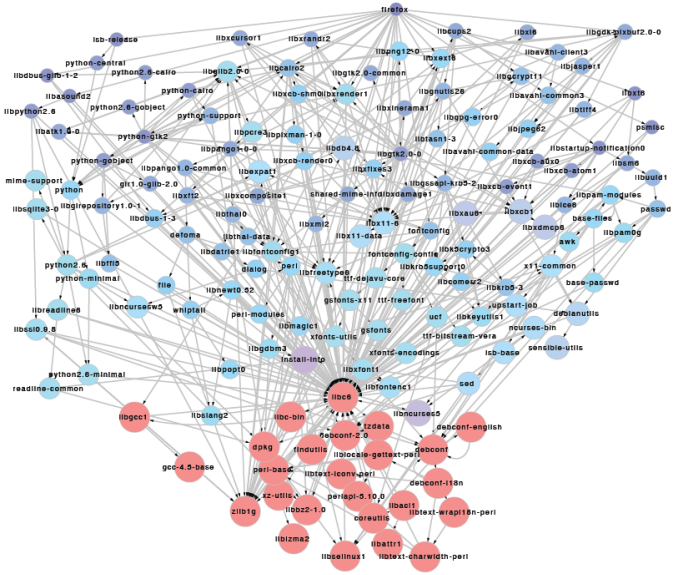
- Genes interact via biomolecular networks
- Genes evolve new functions (slow)
- Genes are exchanged between organisms via Horizontal Gene Transfer (fast)



S. Maslov, TY Pang, K. Sneppen, S. Krishna, PNAS (2009)

Software as interacting/evolving tools

- Software packages interact via dependency networks



- Software engineers write new functions (slow)

- Software packages are exchanged between computers downloaded via internet (fast)

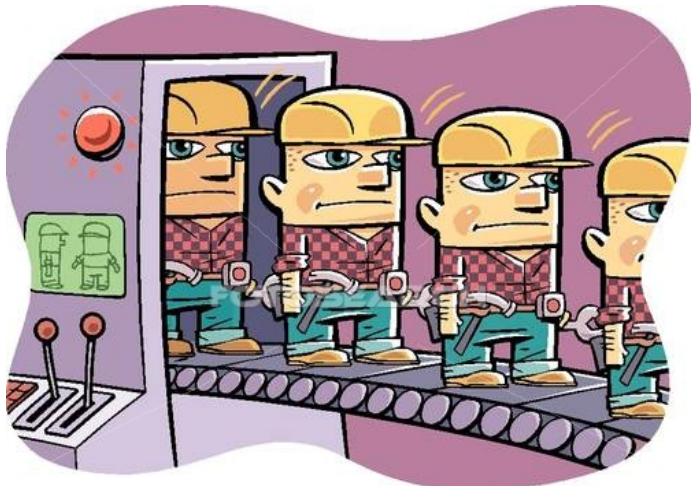
Bacterial genomes ~= Linux packages. TY Pang, **S. Maslov**, PNAS (2013)

Toolbox model of evolution of prokaryotic genomes by Horizontal Gene Transfer

- **S. Maslov**, TY Pang, K. Sneppen, S. Krishna, PNAS (2009)
- TY Pang, **S. Maslov**, PLoS Comp Bio (2011)
- J Grilli, B Bassetti, **S Maslov**, M Cosentino Lagomarsino, NAR (2012)
- TY Pang, **S. Maslov**, PNAS (2013)

How many bureaucrats does a bacterium need?

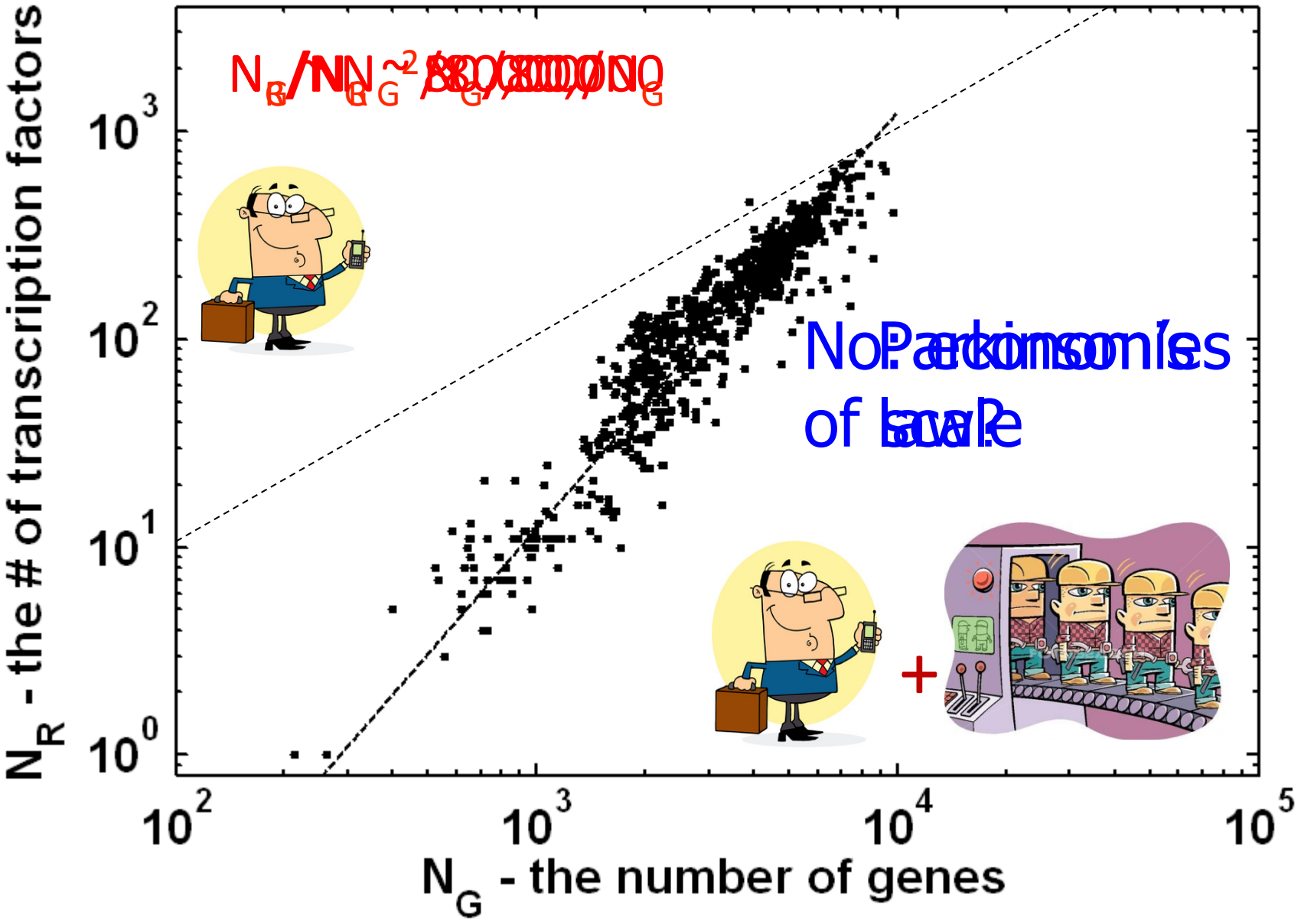
Regulator genes

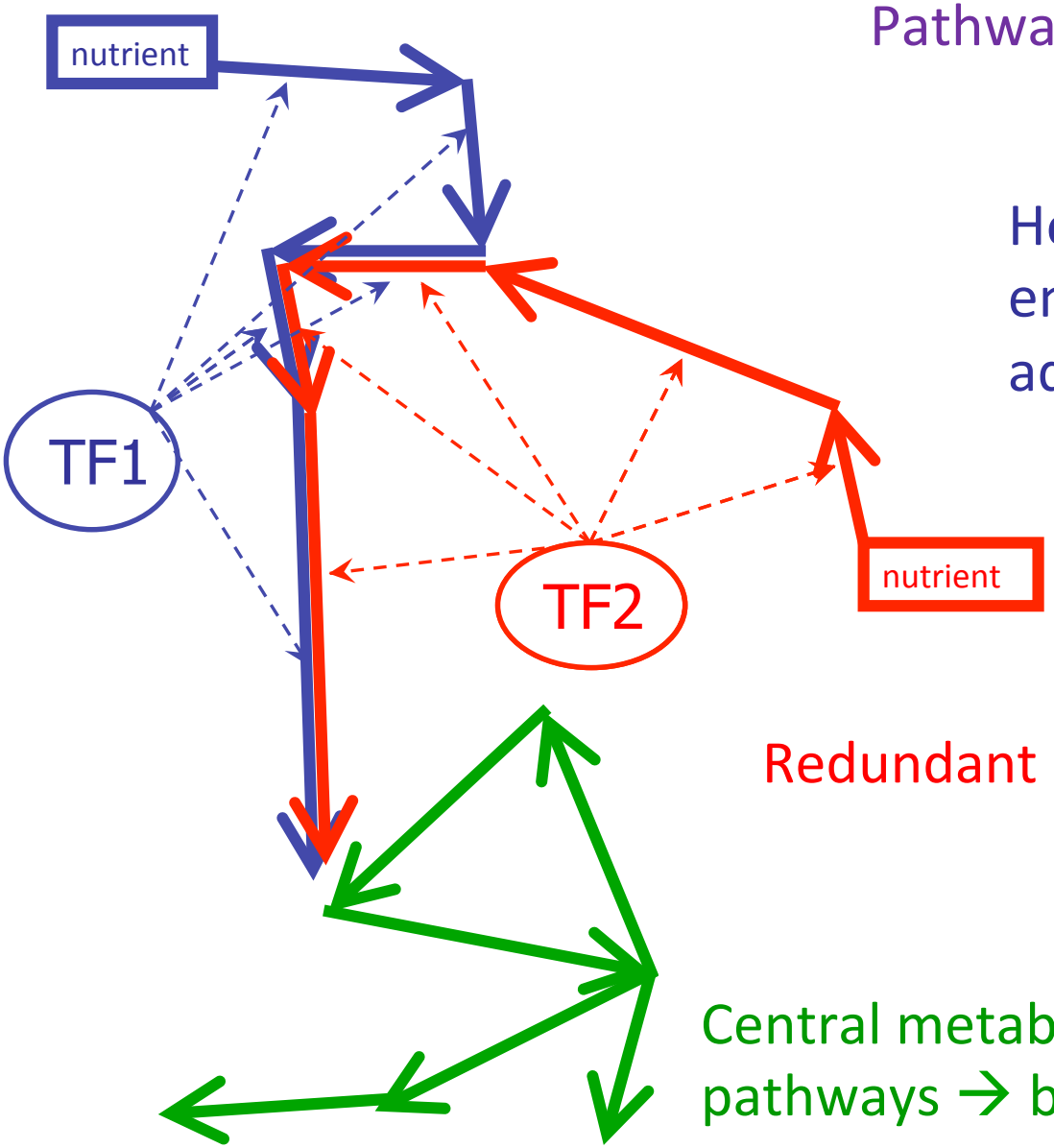


Worker genes

E. Van Nimwegen, Trends in Genetics, 2003

Figure adapted from **S. Maslov**, TY Pang, K. Sneppen, S. Krishna, PNAS (2009)





Pathways can be deleted too

Horizontal gene transfer: entire pathways could be added in one step

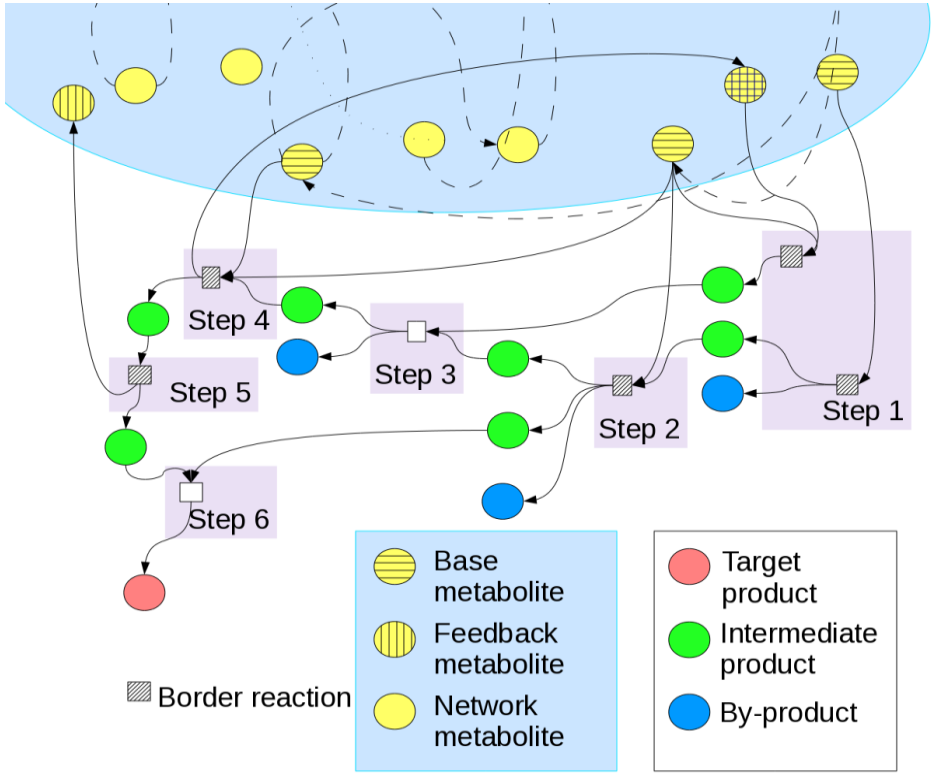
Redundant enzymes are removed

Central metabolic core → anabolic pathways → biomass production

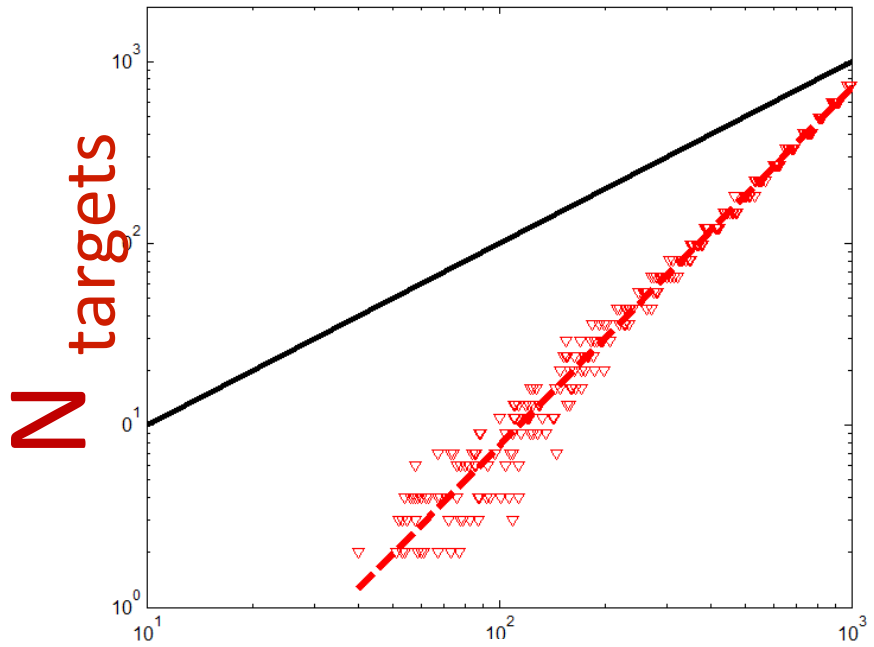
Horizontal Gene Transfer
→Franken-Pathways

Thanks to Chris Marx for idea of this slide

Synthetic biology flavor: find the minimal pathway synthesizing the target metabolite from all reactions in KEGG database



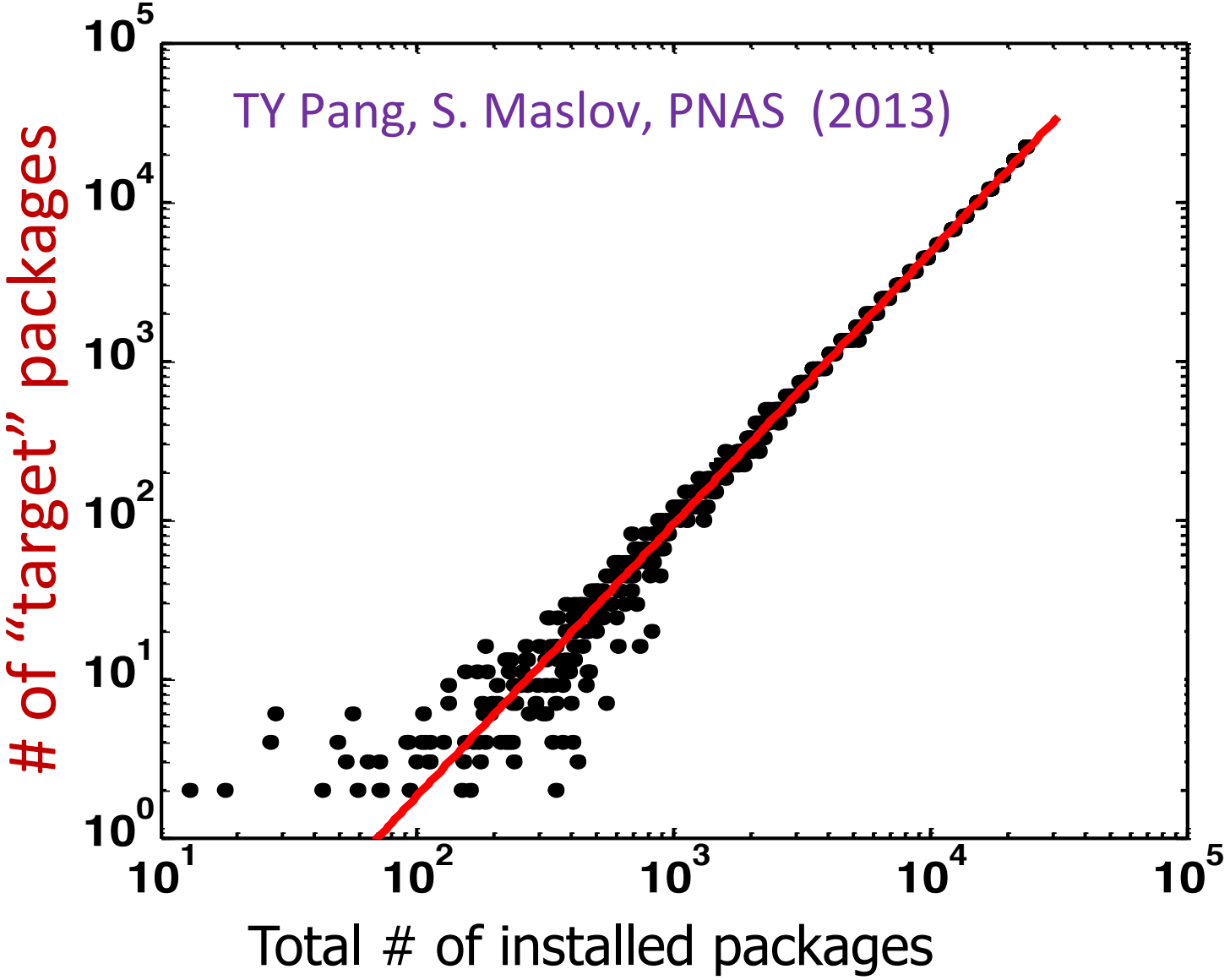
Inspired by "scope-expansion" algorithm of T. Handorf, O. Ebenhoh, R. Heinrich, JME 2005

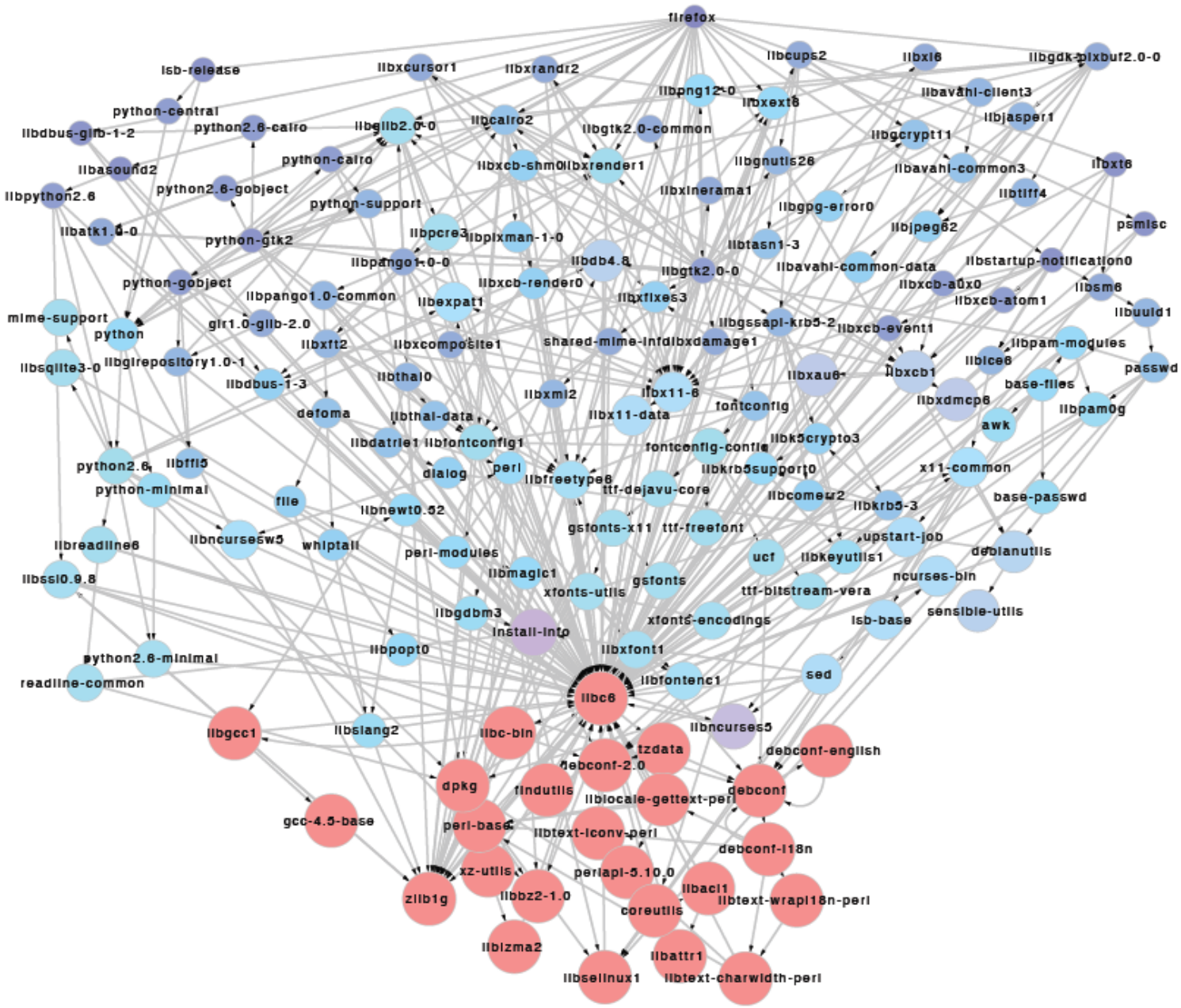


TY Pang, S. Maslov, PLoS Comp Bio (2011)

Metabolic networks	Software programs
Reactions catalyzed by Enzymes	Subroutines or Installable Packages
Metabolites	Variables
AND function on inputs to reaction nodes:	AND function on inputs to subroutine nodes:
OR function on inputs to metabolite nodes:	OR function on inputs to variable nodes:

Software packages for Linux





TY Pang, S. Maslov, PNAS (2013)

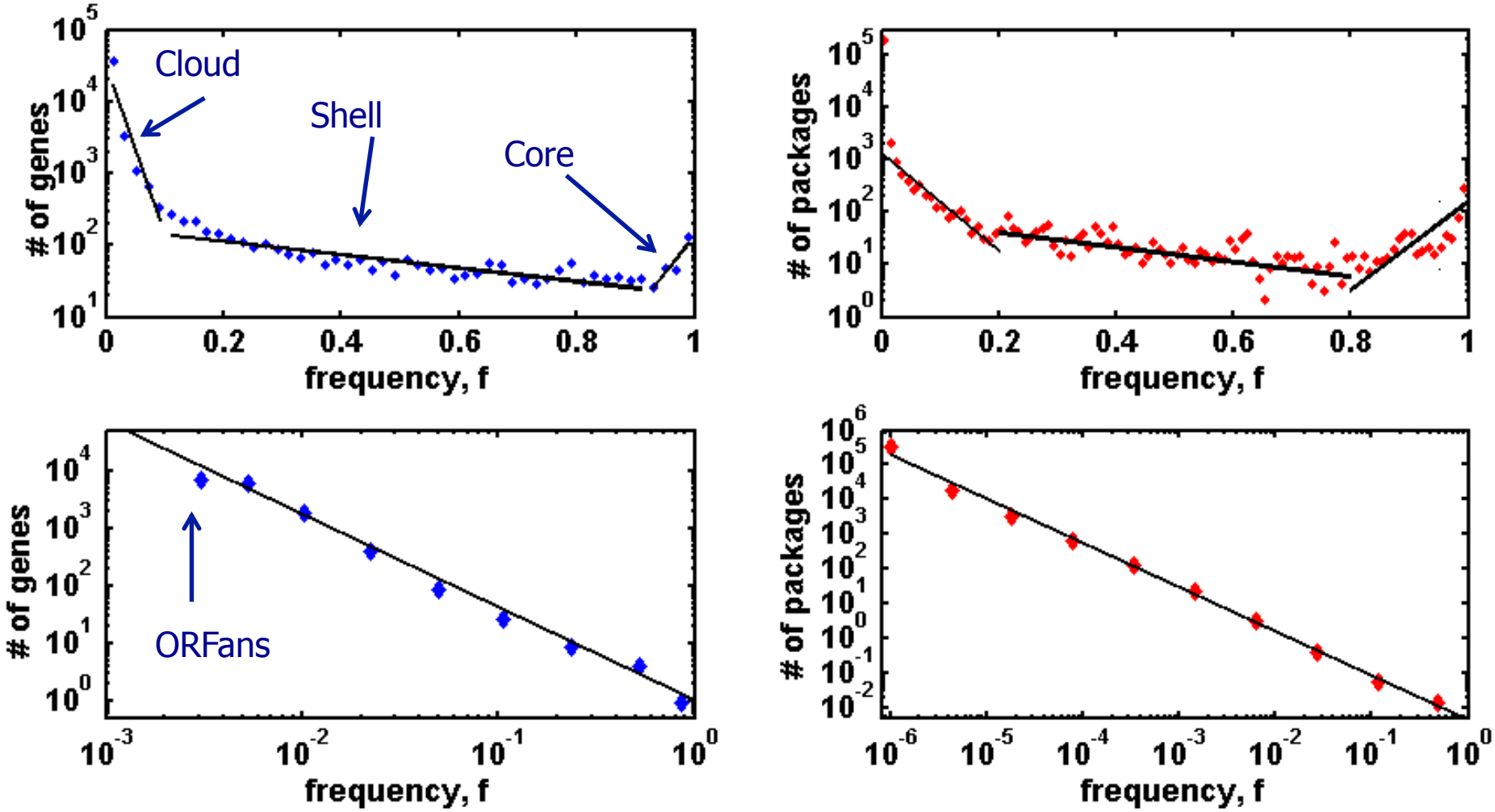
Frequency distribution of genes and software packages

TY Pang, **S. Maslov**, PNAS (2013)

Empirical data on component frequencies and dependency networks

- Bacterial genomes (eggnog.embl.de):
  - 500 sequenced prokaryotic genomes
  - 44,000 Orthologous Gene families
  - Dependency network:
    Upstream-downstream position in pathways
- Linux packages (popcon.ubuntu.com):
  - 200,000 Linux packages installed on
  - 2,000,000 individual computers
  - Dependency network is at Ubuntu project website

Frequency distributions



$P(f) \sim f^{-1.5}$ except the top $\sqrt{N}$ "universal" components with $f \sim 1$

TY Pang, S. Maslov, PNAS (2013)

What determines the frequency of installation/use of a gene/package?

- Popularity:
  - Frequency ~ self-amplifying popularity
    AKA preferential attachment
  - Relevant for social systems: WWW links, facebook friendships, scientific citations
- Functional role:
  - Frequency ~ breadth or importance
    of the functional role
  - Relevant for biological and technological systems
    where selection adjusts undeserved popularity

Gene/Package frequency is correlated with position in dependency network

- Frequency ~ dependency degree, K_{dep}
  - K_{dep} (i)= the total number of components that **directly or indirectly depend** on package i
- Empirical correlation coefficient is 0.4
- By combining popularity & importance correlation coefficient → 0.8

Model of dependency network evolution

- N components are added gradually over evolutionary time
- New component directly depends on *D* previously existing components selected randomly
- $N^{(D-1)/D}$ universal components
- Other components have frequencies $P(f) \sim f^{-(1+1/D)}$

Eukaryotes run windows

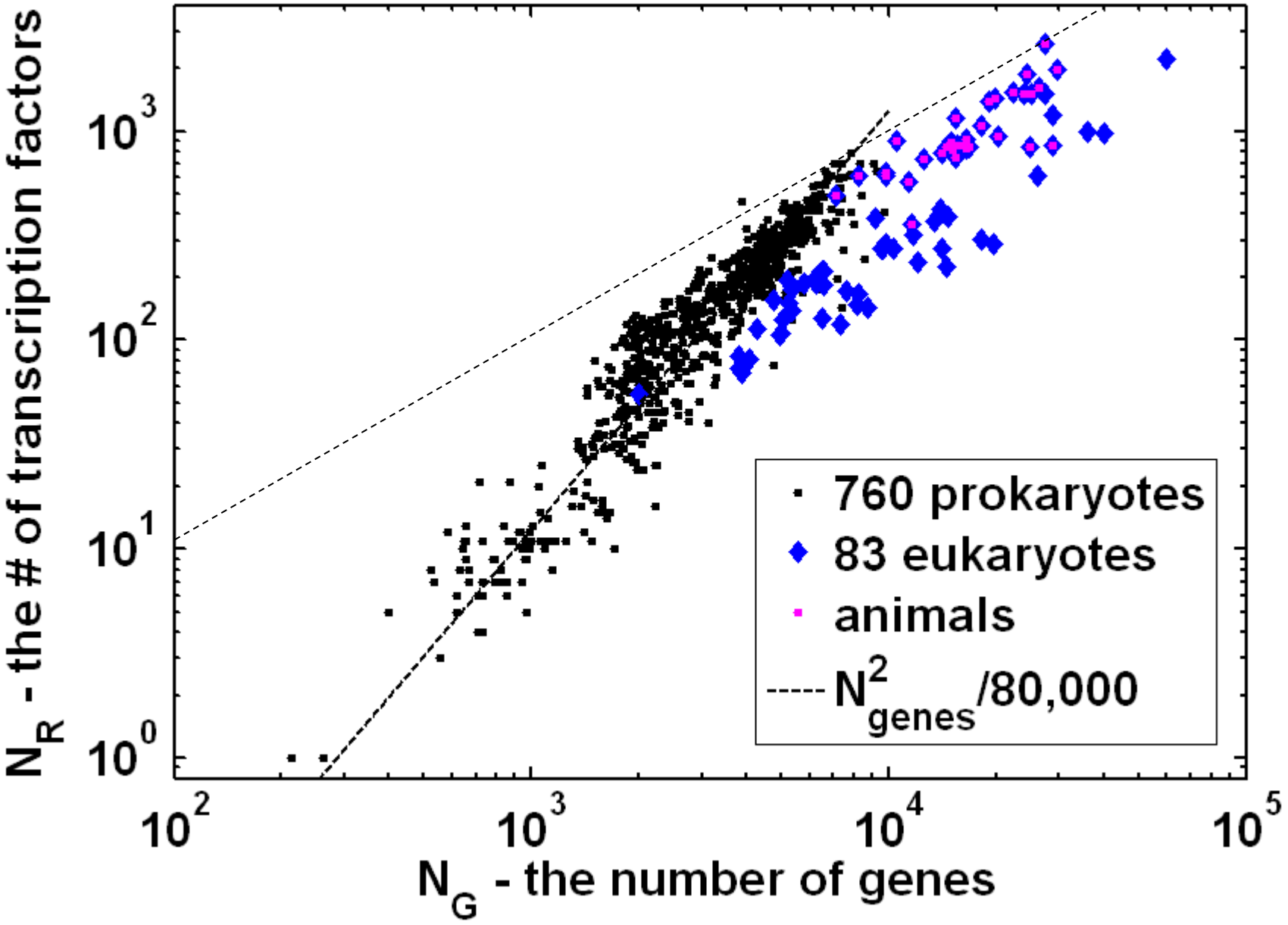


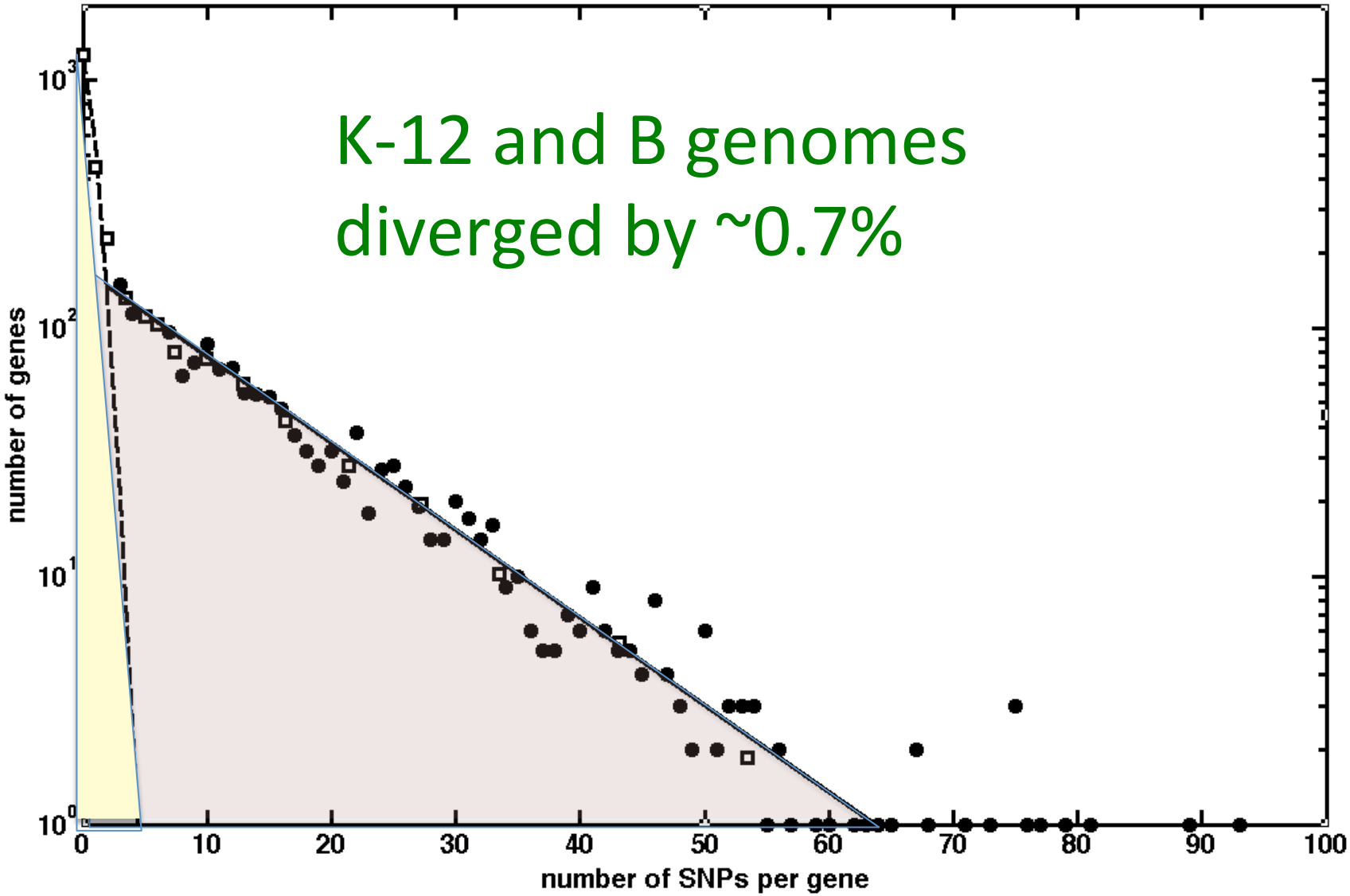
Fig S3 from S. Maslov, TY Pang, K. Sneppen, S. Krishna, PNAS (2009)

Evolution of bacterial strains

species → strains
"installed" genes → "updated" genes

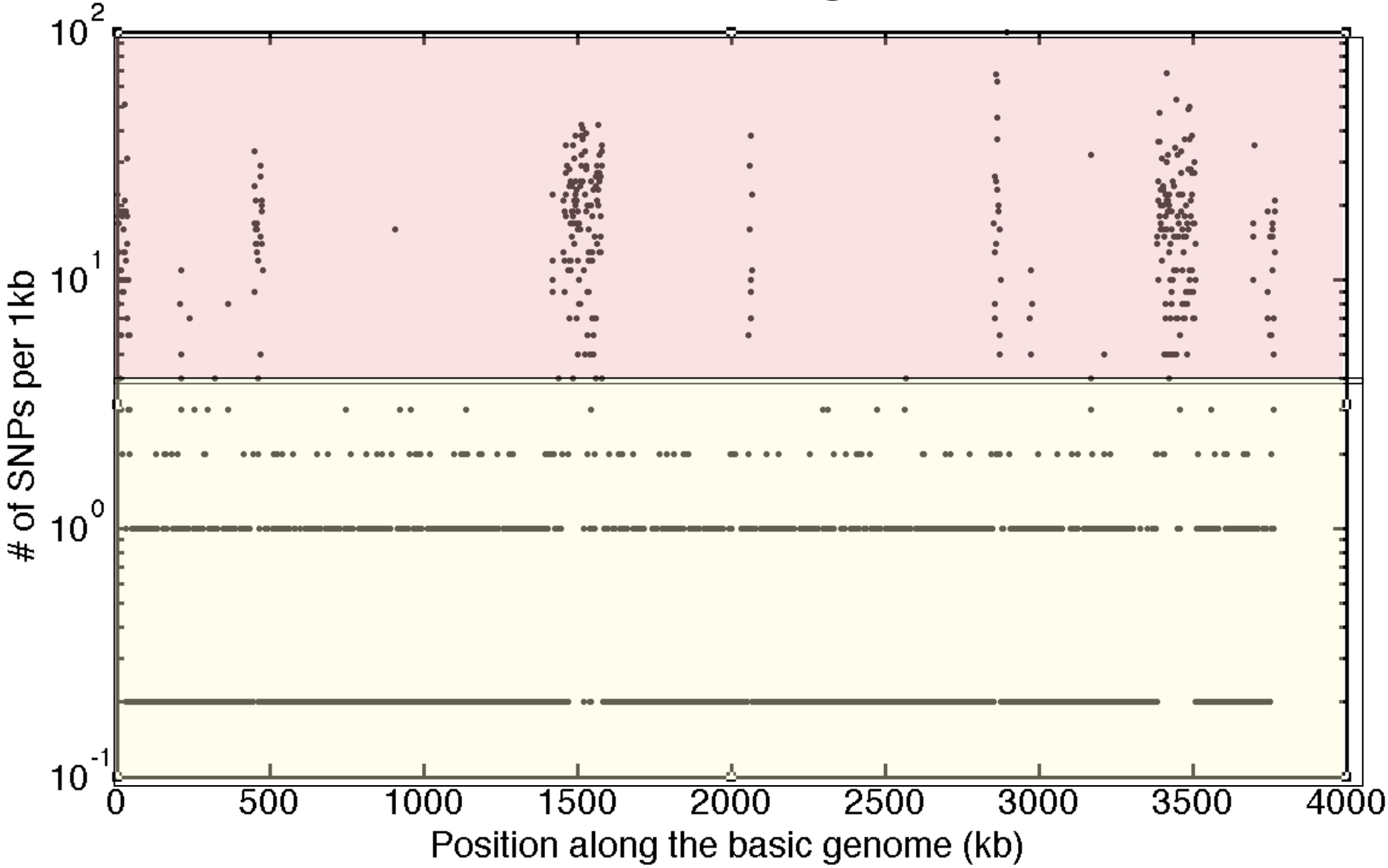
- Studier FW, Daegelen P, Lenski RE**, Maslov S,** Kim JF, J. Mol Biol. (2009)
- Dixit P, Pang TY, Studier FW, **Maslov S**, submitted (2014); arXiv:1405.2548

SNPs in K-12 vs B strains of E. coli

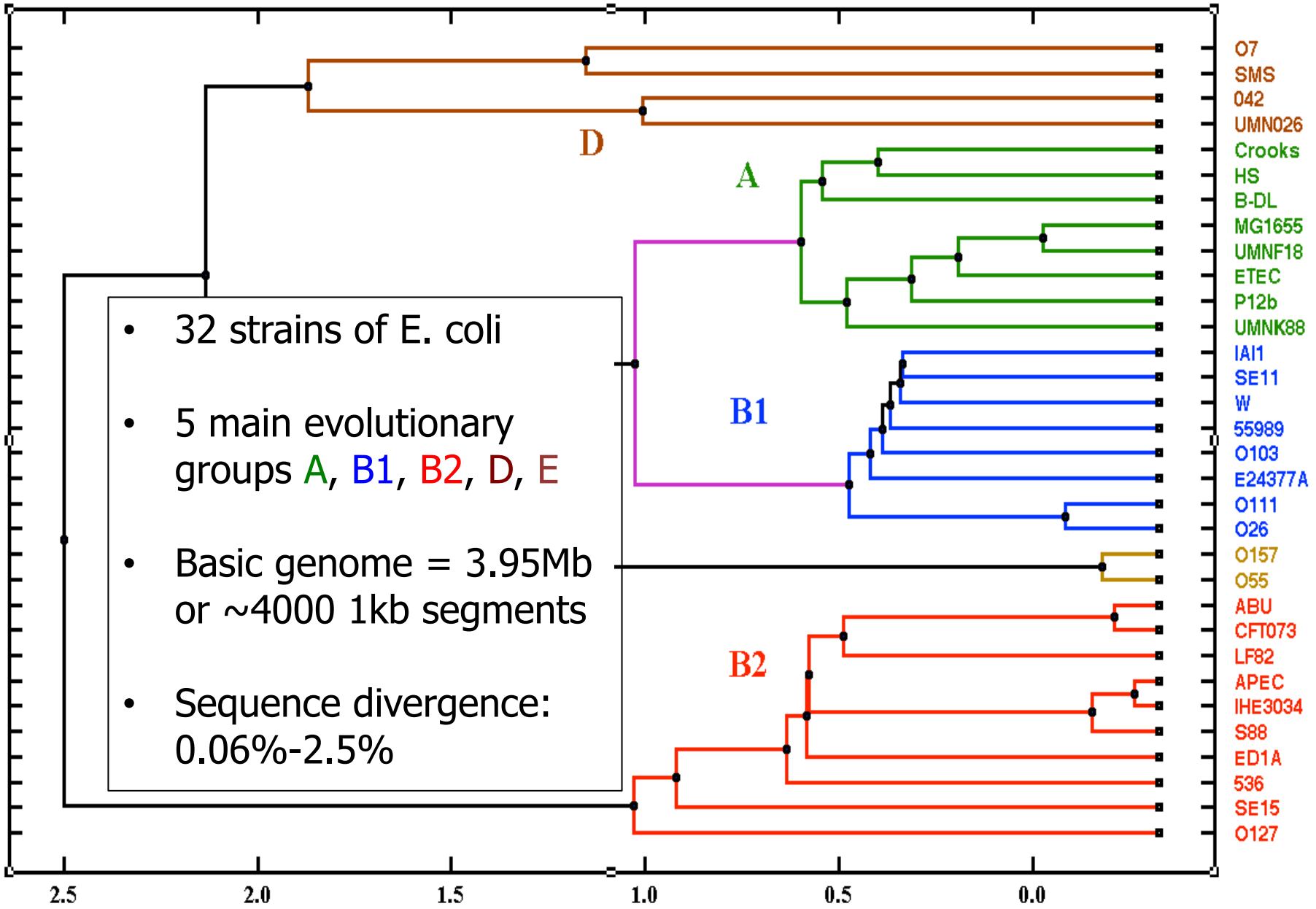


Studier FW, Daegelen P, Lenski RE, **Maslov S**, Kim JF, J. Mol Biol. (2009)

SNPS are clustered along chromosome

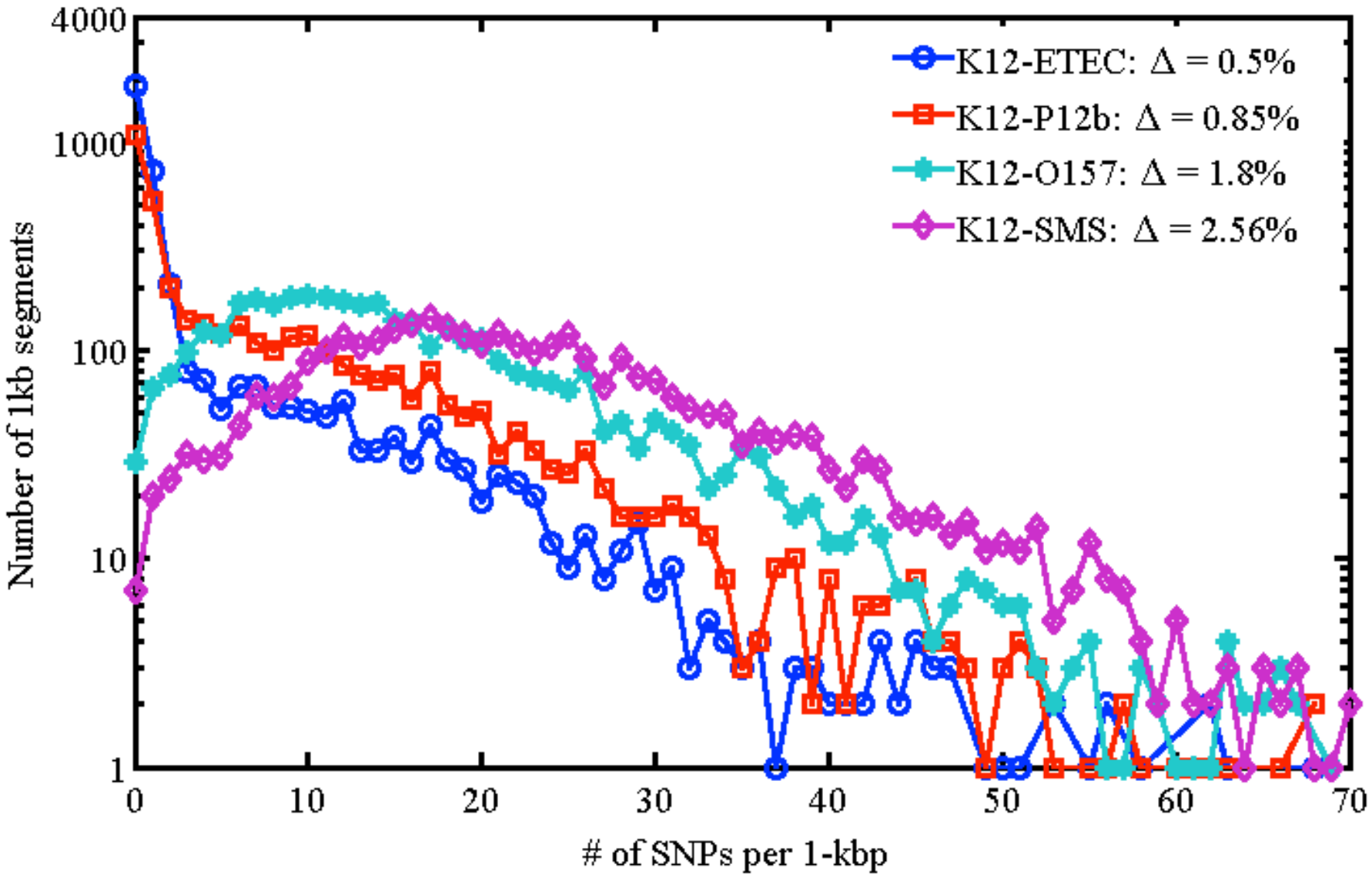


K-12 vs UMNF18 diverged by ~0.18%



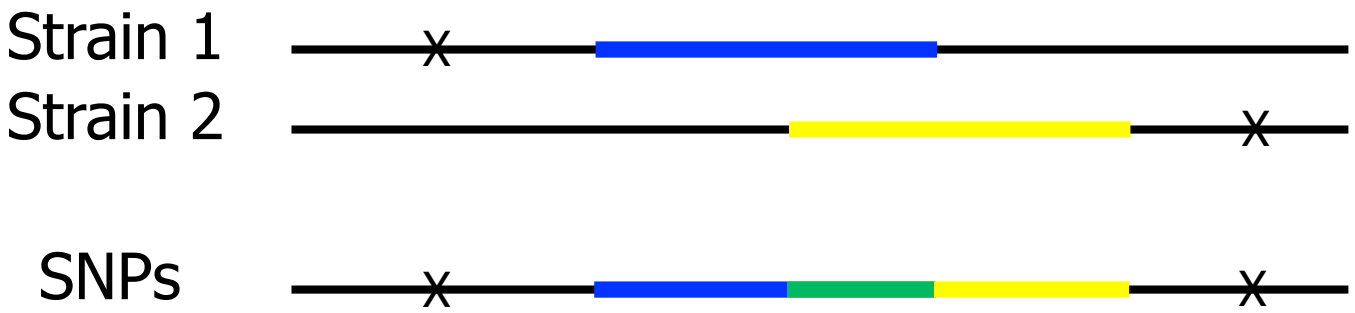
- 32 strains of E. coli
- 5 main evolutionary groups A, B1, B2, D, E
- Basic genome = 3.95Mb or ~4000 1kb segments
- Sequence divergence: 0.06%-2.5%

Dixit P, Pang TY, Studier FW, **Maslov S,** submitted (2014); arXiv:1405.2548

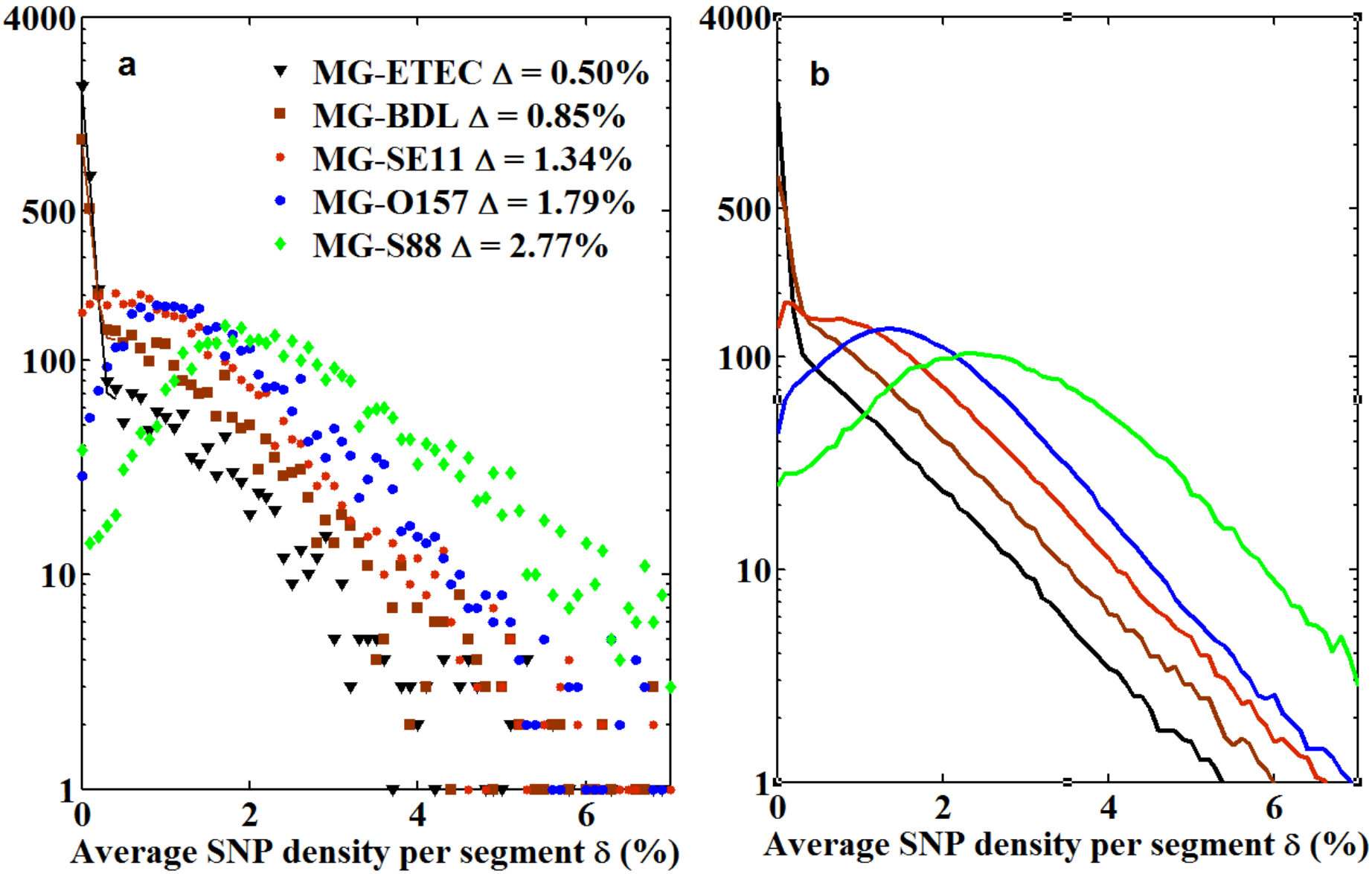

4000
1000
100
10
1
0
10
20
30
40
50
60
70
Number of 1kb segments
# of SNPs per 1-kbp
K12-ETEC: Δ = 0.5%
K12-P12b: Δ = 0.85%
K12-O157: Δ = 1.8%
K12-SMS: Δ = 2.56%

Neutral evolutionary model

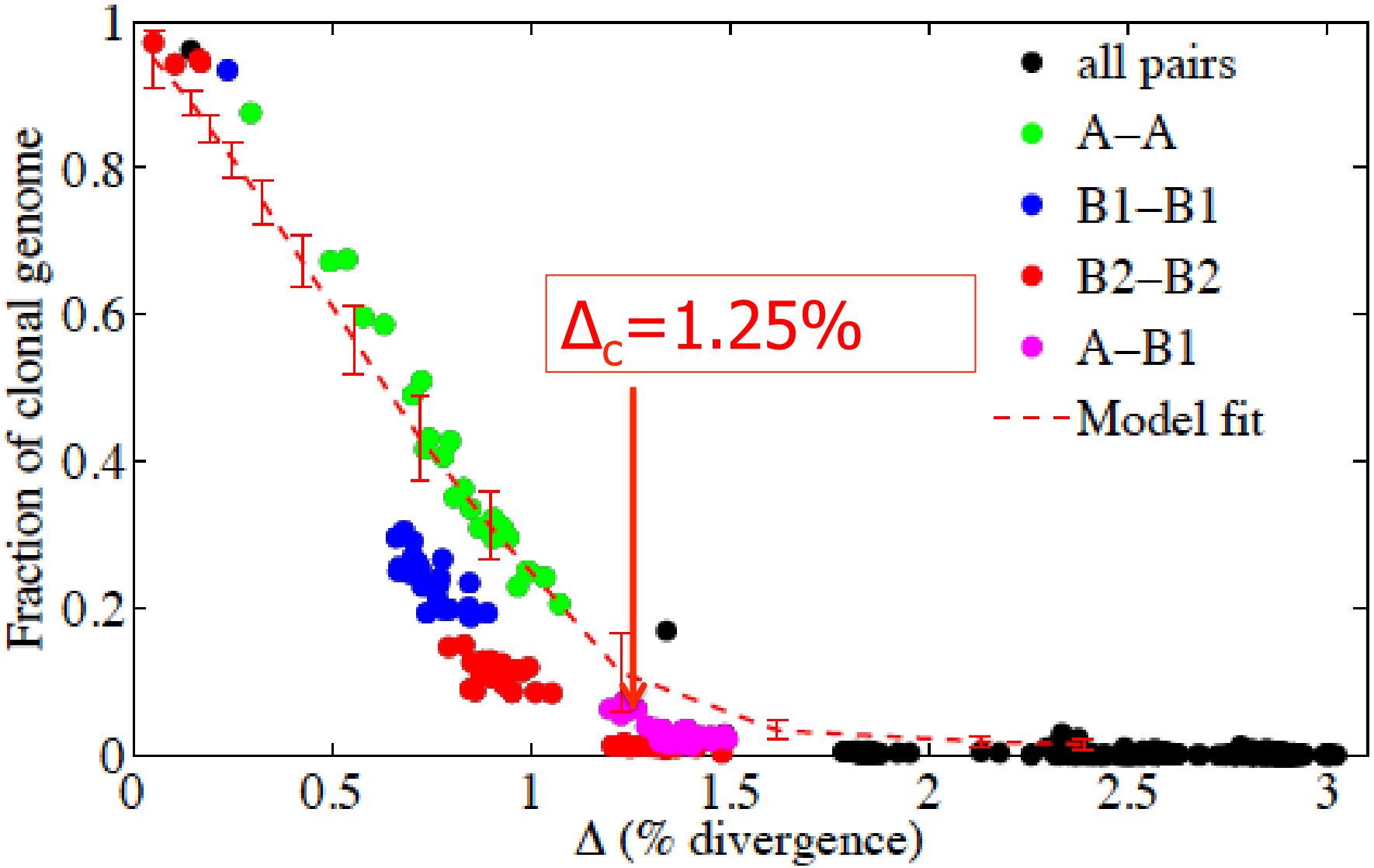
- Follow divergence of 2 strains
- Mutation rate μ =6.3 x 10^{-10} /bp/generation
- Recombination rate ρ =1.7 x 10^{-10}/bp/generation
- l_R=10kb - average length of recombined segments
- δ_{qc}=1% defining probability of successful transfer + recombination: $p_{recomb} \sim \exp(-\delta/\delta_{qc})$

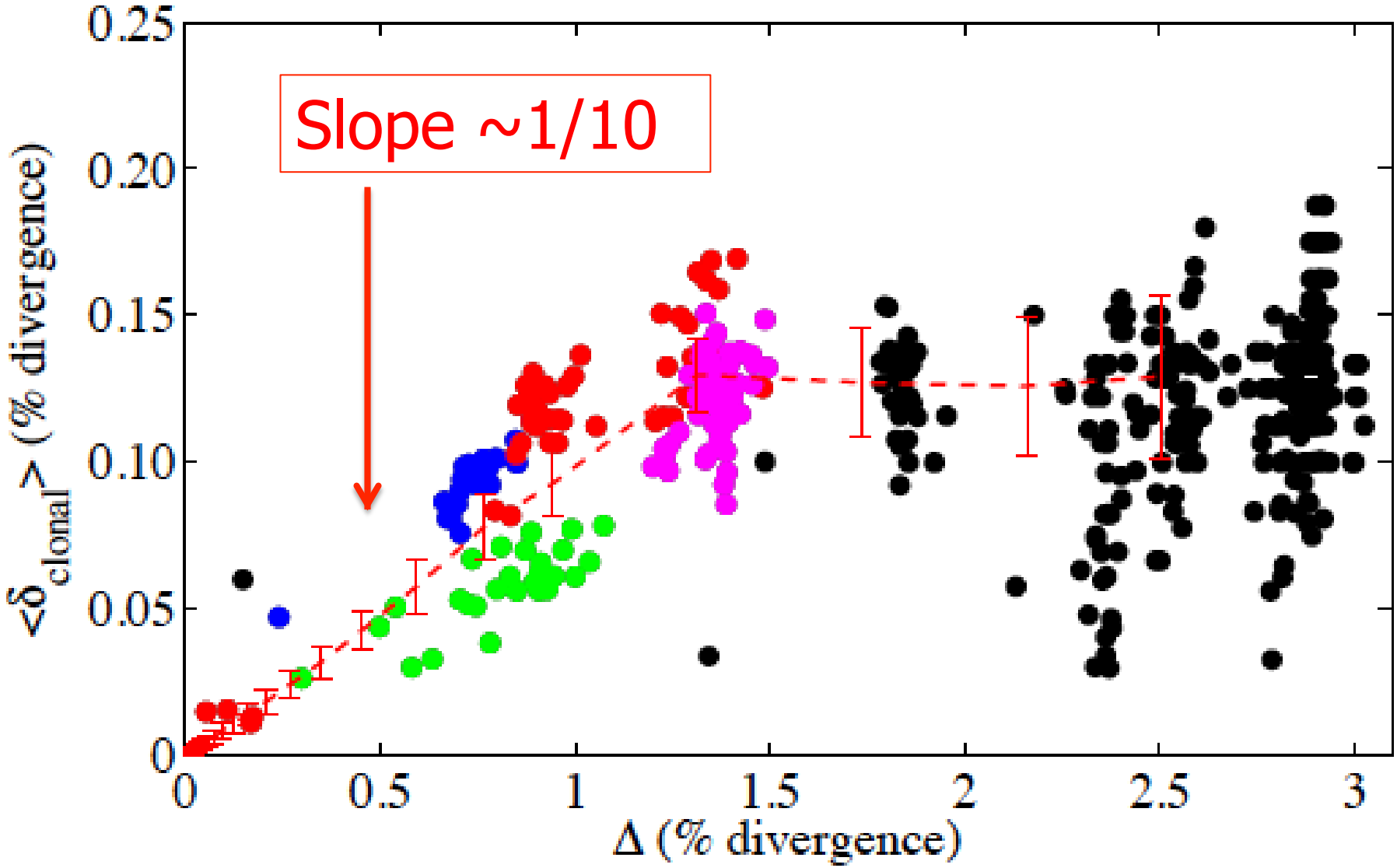


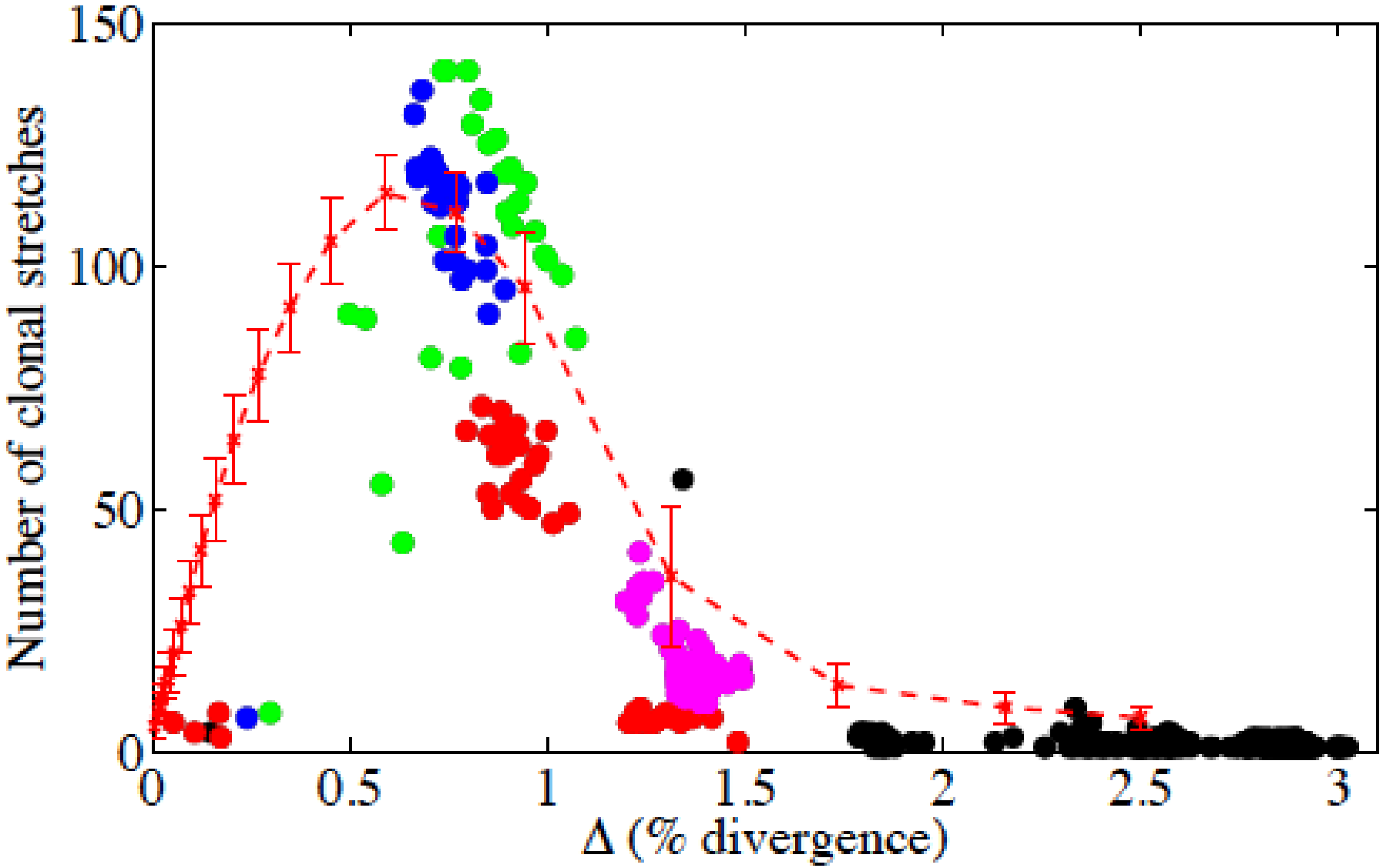
Dixit P, Pang TY, Studier FW, **Maslov S,** submitted (2014); arXiv:1405.2548

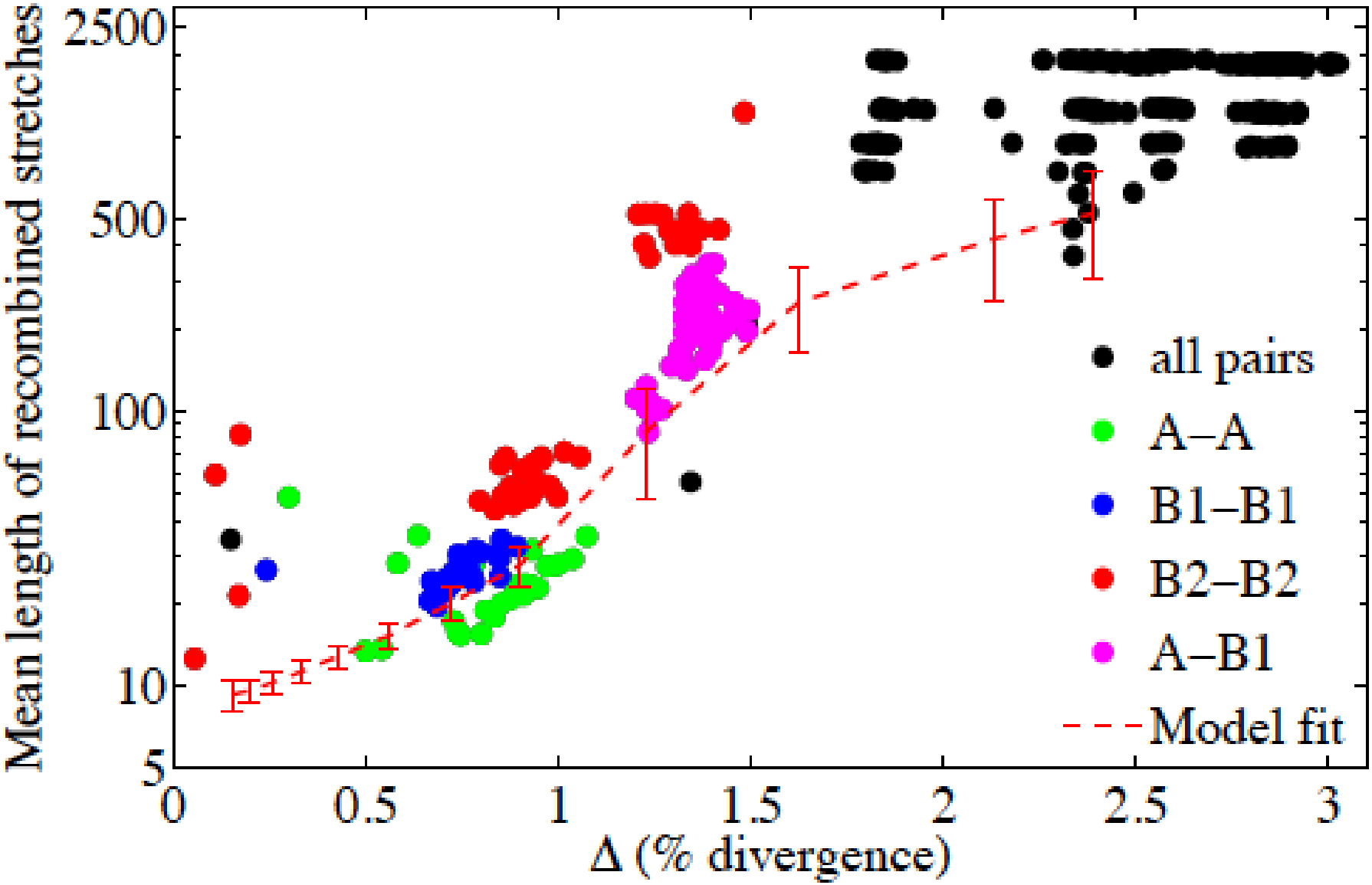


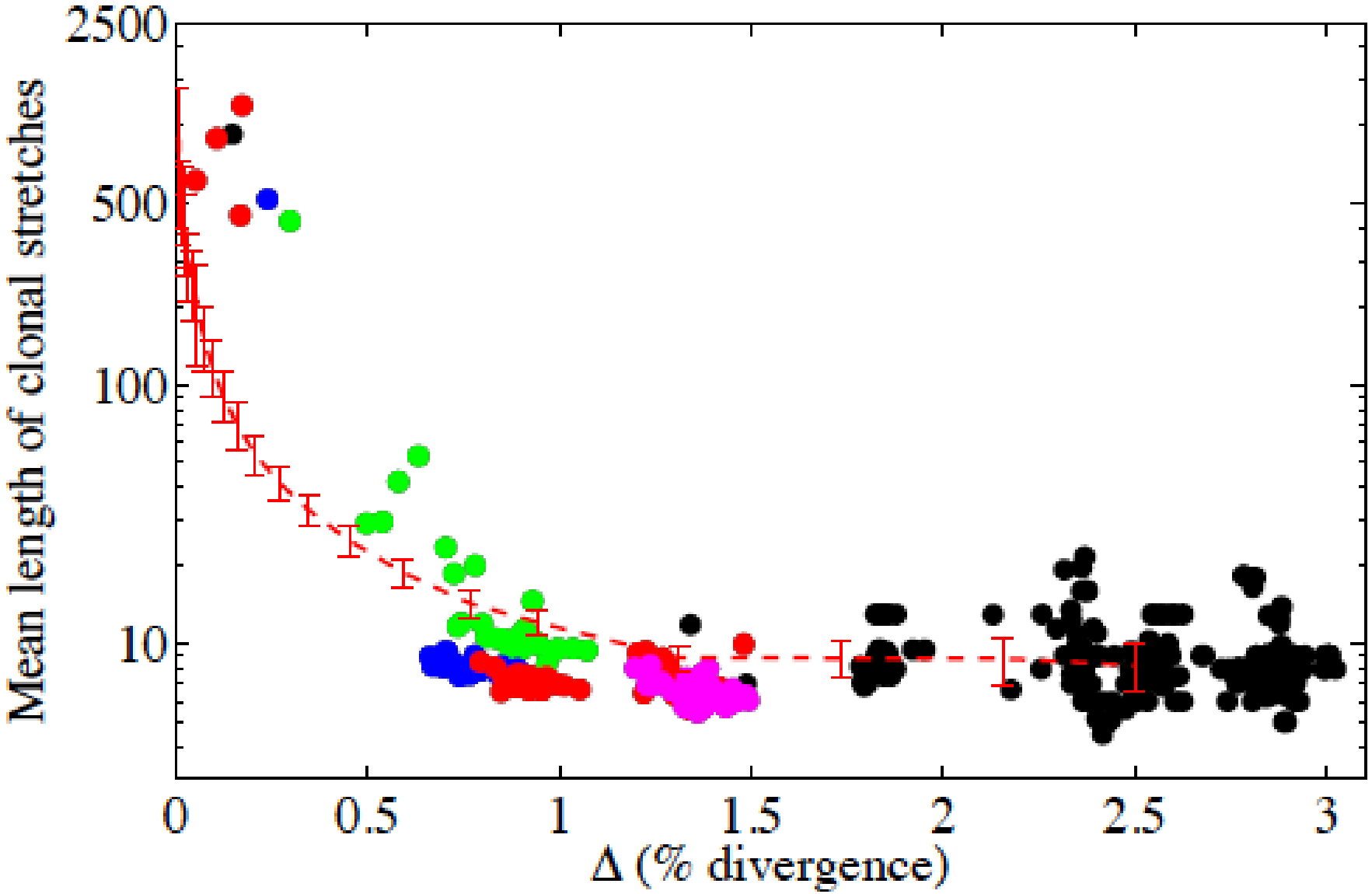
Dixit P, Pang TY, Studier FW, **Maslov S,** submitted (2014); arXiv:1405.2548


Fraction of clonal genome
Δ (% divergence)
Δc=1.25%
all pairs
A–A
B1–B1
B2–B2
A–B1
Model fit
0
0.2
0.4
0.6
0.8
1
0
0.5
1
1.5
2
2.5
3


Slope ~1/10
0.25
0.20
0.15
0.10
0.05
0
0
0.5
1
1.5
2
2.5
3
<δclonal> (% divergence)
Δ (% divergence)


150
100
50
0
Number of clonal stretches
0
0.5
1
1.5
2
2.5
3
Δ (% divergence)


Mean length of recombined stretches
2500
500
100
10
5
0
0.5
1
1.5
2
2.5
3
Δ (% divergence)
all pairs
A–A
B1–B1
B2–B2
A–B1
Model fit


2500
500
100
10
Mean length of clonal stretches
0
0.5
1
1.5
2
2.5
3
Δ (% divergence)

Phages are likely responsible for HT

- In 6 closest pairs of strains HT segments are 42kb to 115 kb long – a good match for phage capsids. Above it ~10kbp likely due to RM system.
- What defines bacterial species?
  - NOT common clonal ancestor (disappears at ~1%)
  - Same as in animals/plants it is SEX: the ability to exchange genetic material with each other
  - Biophysics of homologous recombination: "Quality control"
  - Phage-Bacteria Infection Networks: biological "BitTorrent"

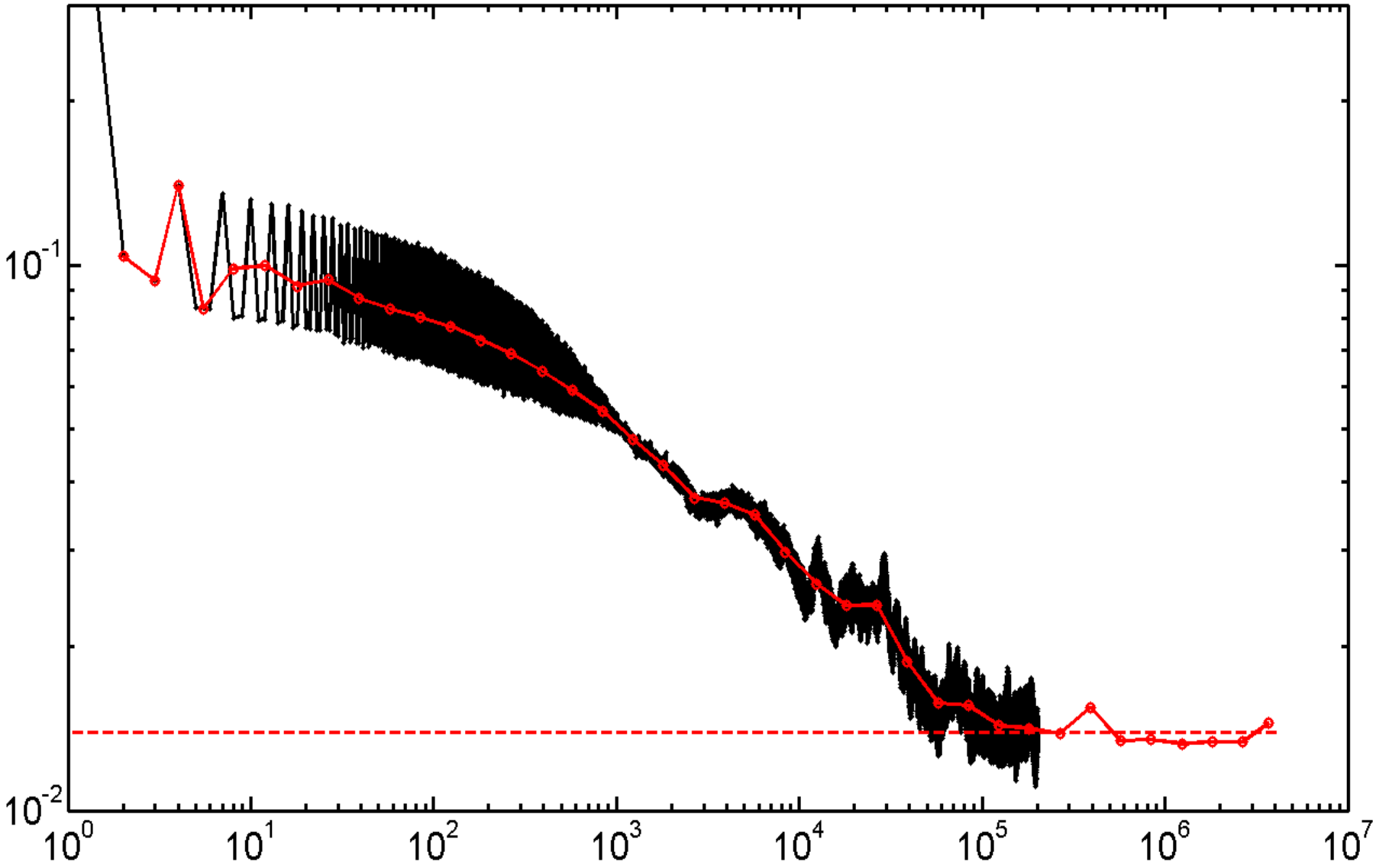
Take home messages for E. coli strain evolution

- HGT + Recombination brings 10 times more variation than clonal mutations
- Clonal segments disappear when genomes diverge above ~1%
- Recombined segments are long: 100 kb -10kb and likely carried by phages
- Genome divergence brought by HGT is exponentially distributed

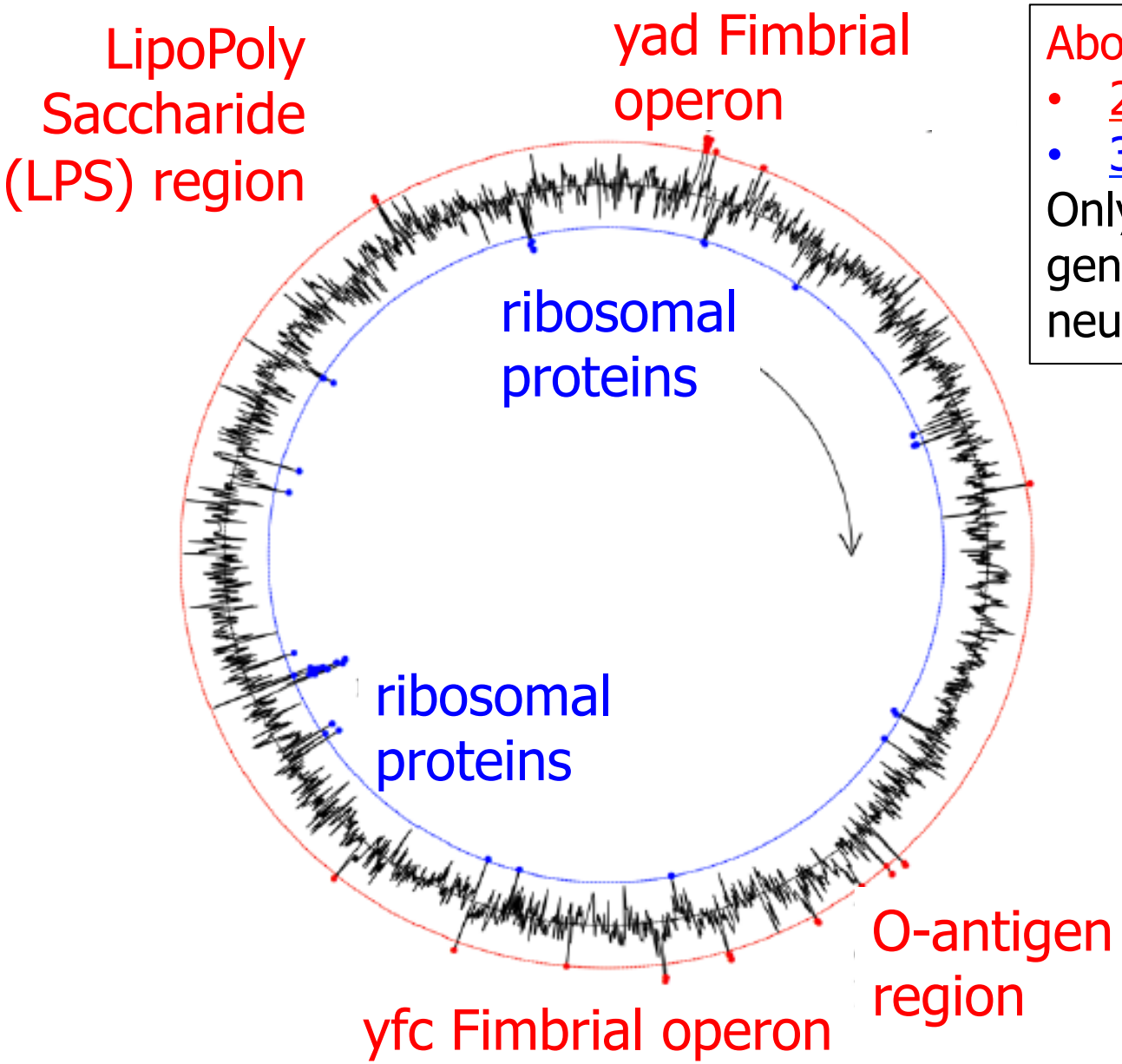
Why exponential tail?

- Time to coalescence:
  $Prob(t)= 1/N_e\ (1-1/N_e)^{t-1}=exp(-t/N_e)$
- $\delta= 2\mu t \rightarrow Prob(\delta)=exp(-\delta/2\mu N_e)$
  $\rightarrow$ exponential slope $=1/2\mu N_e$ or $1/\theta$
- $\theta=60$ or population size $N_e=1 \times 10^9$
  consistent with earlier estimates
- Alternative explanation: biophysics of homologous recombination.
- $Prob(\delta)=(1-\delta)^{2L}=exp(-\delta \cdot 2L)$. $L=30$

SNP-SNP correlations in *B. subtilis* sp168 vs BEST195



Hypervariable and conserved genes

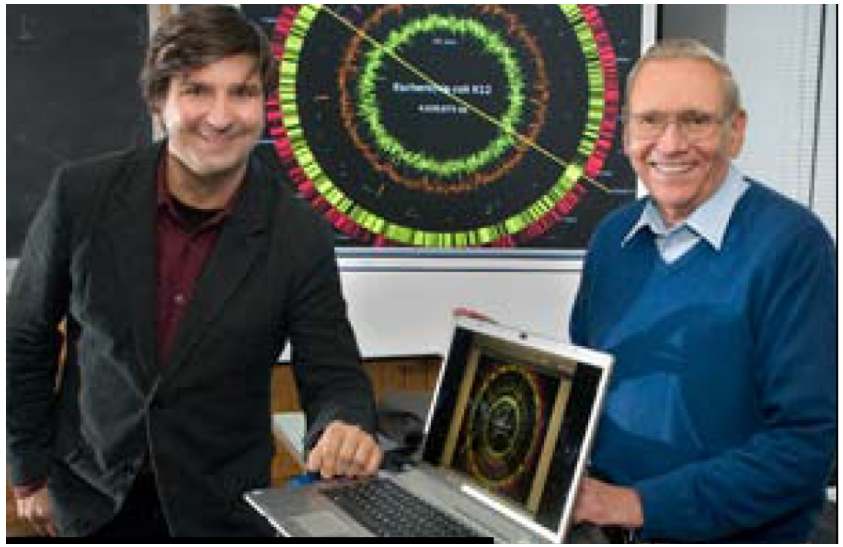


Above/Below 3 std

- 27 hypervariable
- 36 conserved

Only 1.5% of all genes disagree with neutral model

Collaborators & Funding



- **Tin Yau Pang (Stony Brook)**
- **Purushottam Dixit (BNL)**
- **Bill Studier (BNL)**
- Rich Lenski (Michigan State)
- Patrick Daegelen (France)
- Jinhyun Kim (Korea)

DOE Office of Biological and Environmental Research

DOE Systems Biology Knowledgebase (KBase)

Studier FW, Daegelen P, Lenski RE**, Maslov S,** Kim JF, J. Mol Biol. (2009)

Dixit P, Pang TY, Studier FW, **Maslov S,** submitted (2014); arXiv:1405.2548