Lecture Three

Symmetries in the Genome

HC Lee
National Central University

Symmetry in Nature

- Symmetry has been considered as an aspect of beauty in mathematics, physics, chemistry, evolution, human appearance, and psychology
- The cause of symmetry is usually subtle*, and the pursue of it often leads to a deep understanding of the possessor of the symmetry
- **"Subtle is the Lord, but malicious He is not."*

  - Albert Einstein

Symmetry in Genomes

- Symmetry: key to structure and dynamical principles
- Chargaff's parity rule (1951)
  - In DNA, A=T and C=G
  - Watson and Crick's double helix (1953)
- Chargaff's second parity rule (1968)
  - A~T and C~G in SINGLE strand of DNA

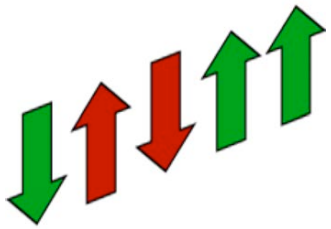
Question: Is CPR2 part of a general phenomenon?
If so, what is it? What is the source of the symmetry?
What can it tell us?

Kong. et al. PLoS ONE (2009) 4, e7553

Reverse, complement, & Inverse symmetries

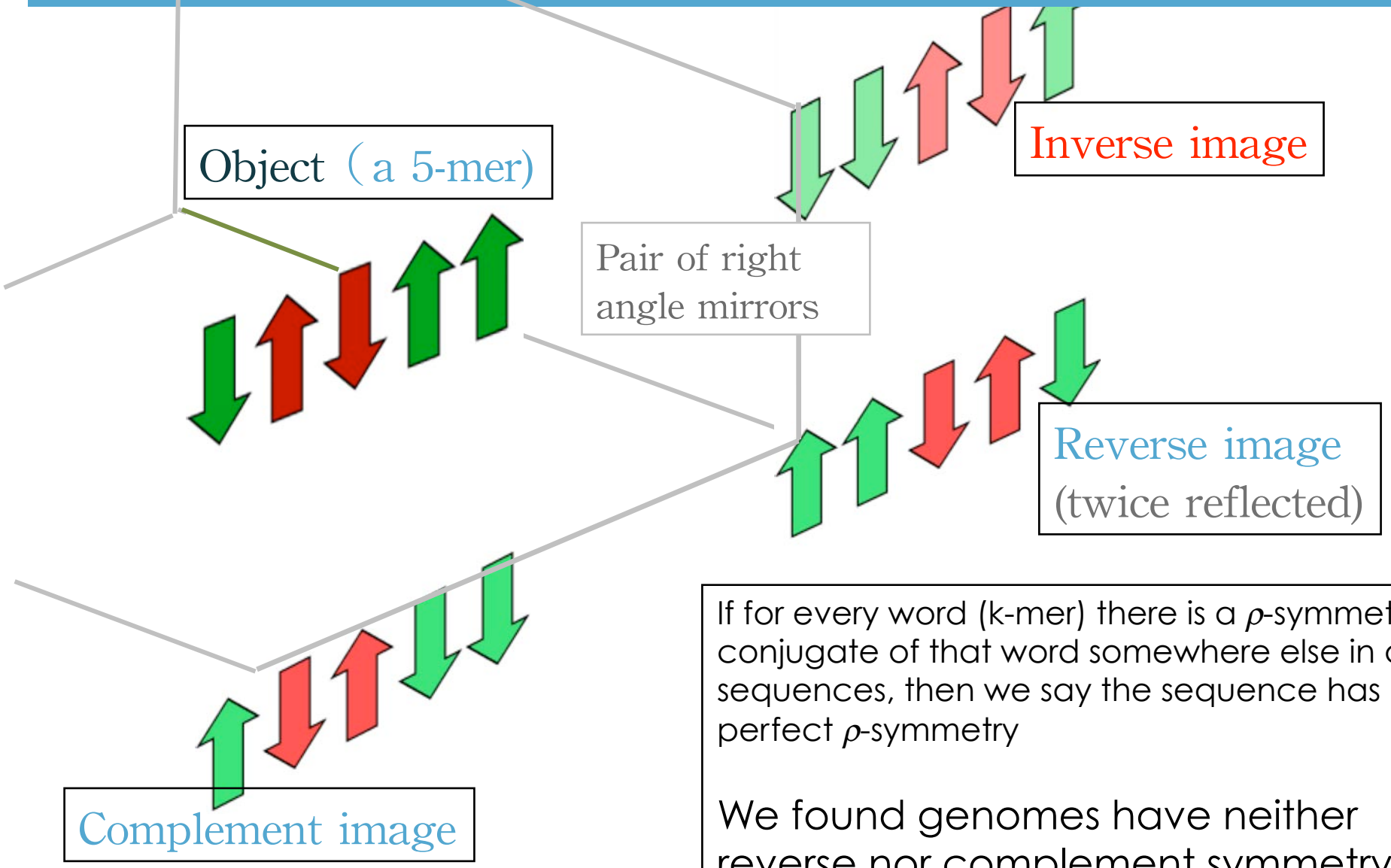
- Conjugation (example) of AAGTC
  - Reverse: CTGAA
  - Complement: TTCAG
  - Inverse (reverse-complement) GACTT
- Symmetry - measure of balance of word frequencies of conjugate pairs

Inverse symmetry and the (Yang-Lee) parity

If we represent the four nucleotides by two pairs of spins, red (C/G) and green (A/T) (for example,  is TCGAA),

then inverse symmetry is exactly the same as (the Yang-Lee) parity in physics.

Reverse/complement/inverse images



If for every word (k-mer) there is a ρ-symmetry conjugate of that word somewhere else in a sequences, then we say the sequence has perfect ρ-symmetry

We found genomes have neither reverse nor complement symmetry but have excellent inverse symmetry

Symmetry index χ

$$\chi_\rho^2 = \frac{1}{2N_\rho} \sum_{(u,u^\dagger)\in P_\rho} \left(\frac{f_u - f_{u^\dagger}}{\sigma_{m_u}} \right)^2, \rho = r, c, \text{or } i, \qquad (1)$$

Index determined by relative, not absolute, frequency difference

u: a k-letter word; u^+, its ρ-conjugate

f_u: occurrence frequency of u in a sequence

P_ρ: all ρ-conjugate pairs

σ_{mu}: the standard deviation of complete set (of words) to which u and u^+ belong

$\chi_\rho \sim 1$, no ρ- symmetry; $\chi_\rho = 0$, perfect ρ-symmetry

χ *vs.* $S^1 = 1 - \left(\sum_{u \in \mathcal{S}} |f_u - f_{u^\dagger}|\right) / \left(\sum_{u \in \mathcal{S}} f_u + f_{u^\dagger}\right)$

Randomized

k	Chr.	S_c^1	$S_{c,R}^1$	χ_c	$\chi_{c,R}$	S_i^1	$S_{i,R}^1$	χ_i	$\chi_{i,R}$
2	*E. coli*	0.9280	0.9992	0.9563	1.1359	0.9974	0.9991	0.0345	1.1925
	HS1	0.8866	0.9998	1.1006	1.1927	0.9992	0.9996	0.0093	1.4425
3	*E. coli*	0.8863	0.9983	0.9607	1.0558	0.9965	0.9982	0.0255	1.0602
	HS1	0.8509	0.9996	0.8821	1.1241	0.9992	0.9996	0.0061	1.1587
4	*E. coli*	0.8323	0.9960	0.9465	1.0086	0.9943	0.9963	0.0307	0.9497
	HS1	0.8001	0.9993	0.8703	1.0733	0.9989	0.9993	0.0065	1.1097
5	*E. coli*	0.7765	0.9918	0.9320	1.0177	0.9905	0.9921	0.0399	0.9706
	HS1	0.7590	0.9988	0.8792	1.0511	0.9984	0.9988	0.0066	1.0207
6	*E. coli*	0.7328	0.9839	0.9188	1.0110	0.9824	0.9846	0.0611	0.9671
	HS1	0.7159	0.9976	0.8888	1.0191	0.9973	0.9976	0.0091	1.0082

L^1-distance (S^1) is not a good measure of symmetry

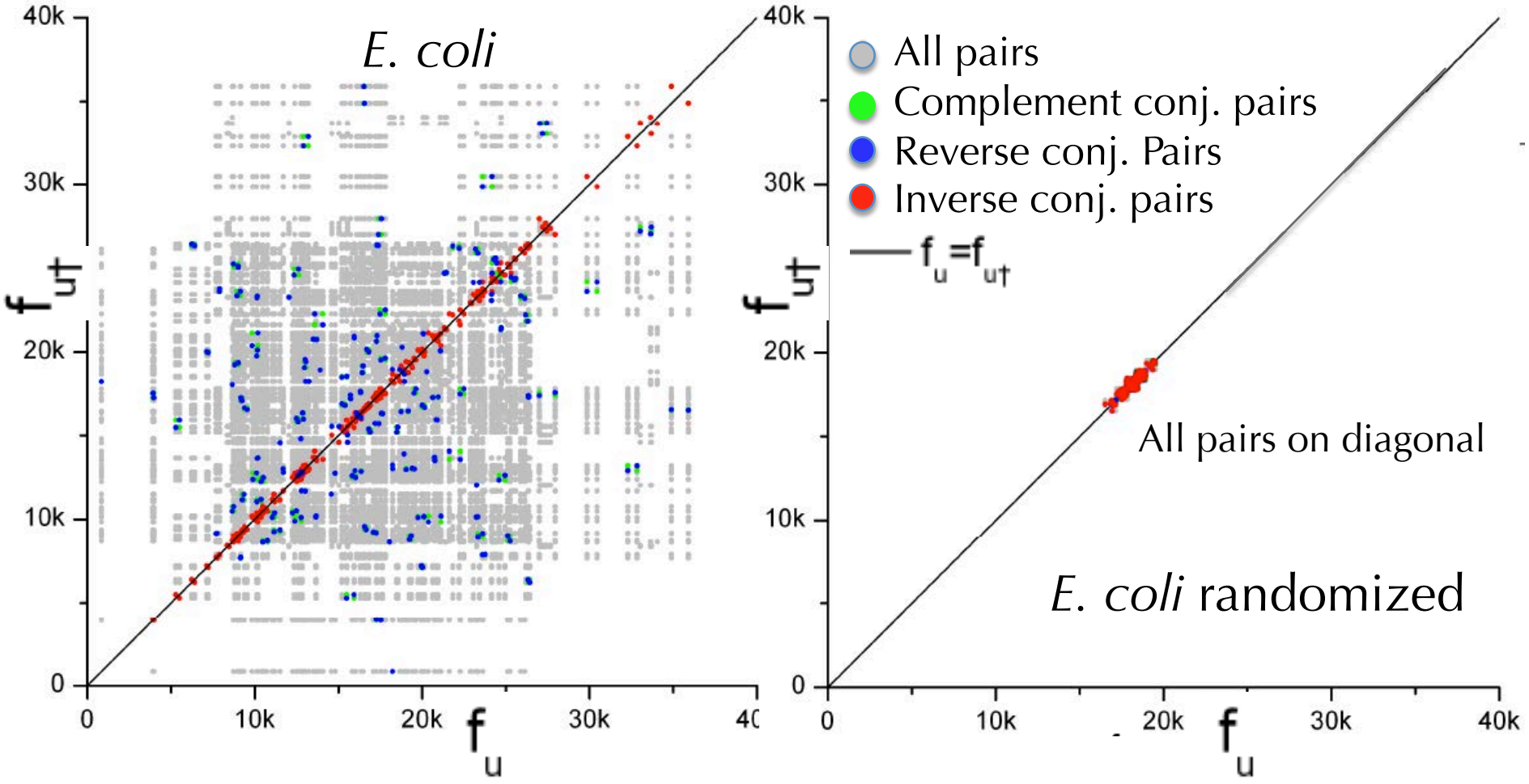
χ *vs.* $S^1 = 1 - \left(\sum_{u \in \mathcal{S}} |f_u - f_{u^\dagger}|\right) / \left(\sum_{u \in \mathcal{S}} f_u + f_{u^\dagger}\right)$

Randomized

k	Chr.	S_c^1	$S_{c,R}^1$	χ_c	$\chi_{c,R}$	S_i^1	$S_{i,R}^1$	χ_i	$\chi_{i,R}$
2	*E. coli*	0.9280	0.9992	0.9563	1.1359	0.9974	0.9991	0.0345	1.1925
	HS1	0.8866	0.9998	1.1006	1.1927	0.9992	0.9996	0.0093	1.4425
3	*E. coli*	0.8863	0.9983	0.9607	1.0558	0.9965	0.9982	0.0255	1.0602
	HS1	0.8509	0.9996	0.8821	1.1241	0.9992	0.9996	0.0061	1.1587
4	*E. coli*	0.8323	0.9960	0.9465	1.0086	0.9943	0.9963	0.0307	0.9497
	HS1	0.8001	0.9993	0.8703	1.0733	0.9989	0.9993	0.0065	1.1097
5	*E. coli*	0.7765	0.9918	0.9320	1.0177	0.9905	0.9921	0.0399	0.9706
	HS1	0.7590	0.9988	0.8792	1.0511	0.9984	0.9988	0.0066	1.0207
6	*E. coli*	0.7328	0.9839	0.9188	1.0110	0.9824	0.9846	0.0611	0.9671
	HS1	0.7159	0.9976	0.8888	1.0191	0.9973	0.9976	0.0091	1.0082

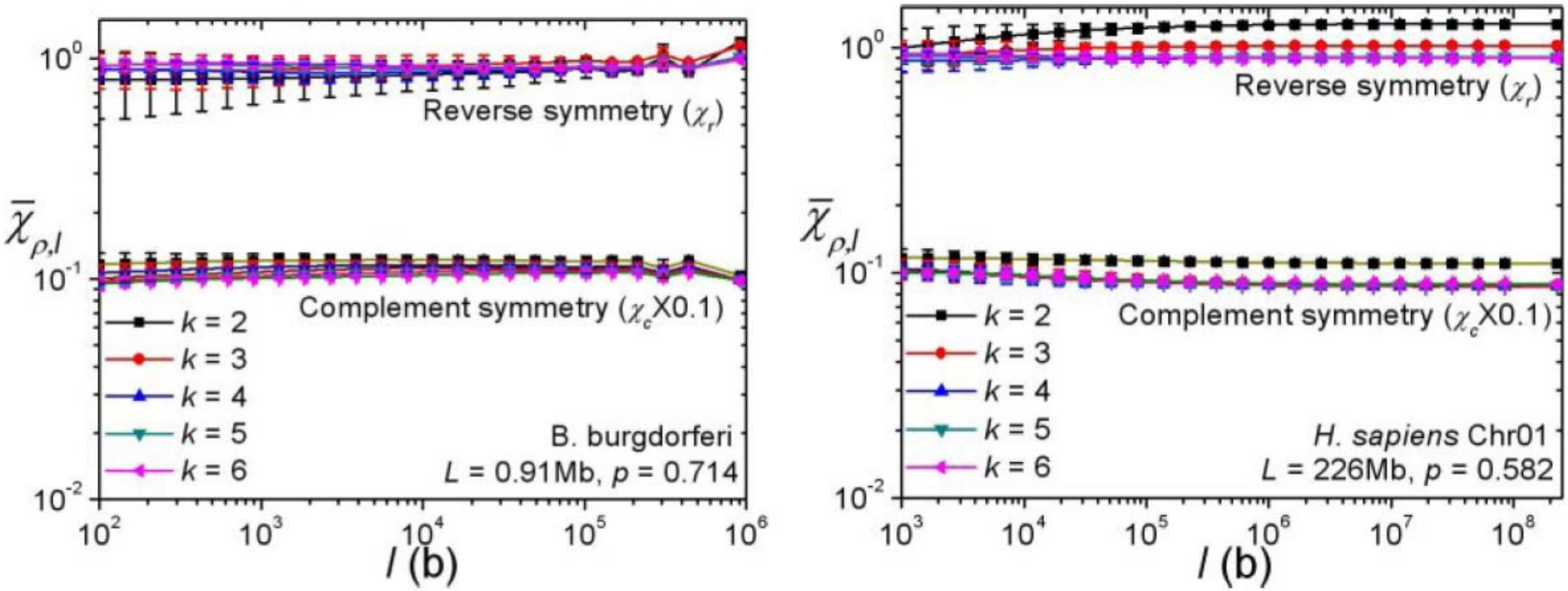
L^1-distance (S^1) is not a good measure of symmetry

f_u *vs.* f_{u+} plots are good visuals, but can be misleading



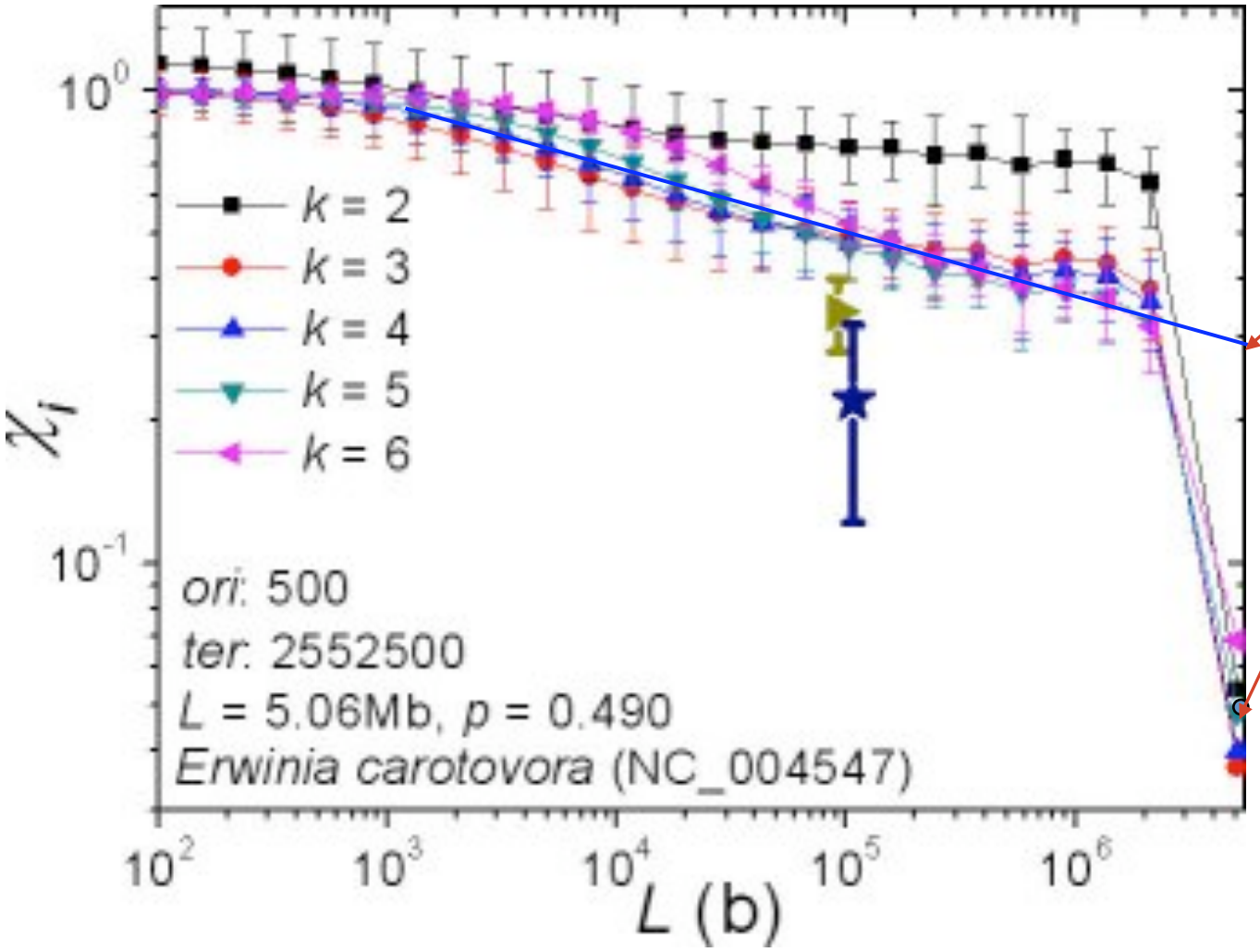
Reverse and complement symmetry absent in all genomes on all scales

χ_l : χ of segment of length l



segment length

Structure in segmental $\chi_{inverse}$



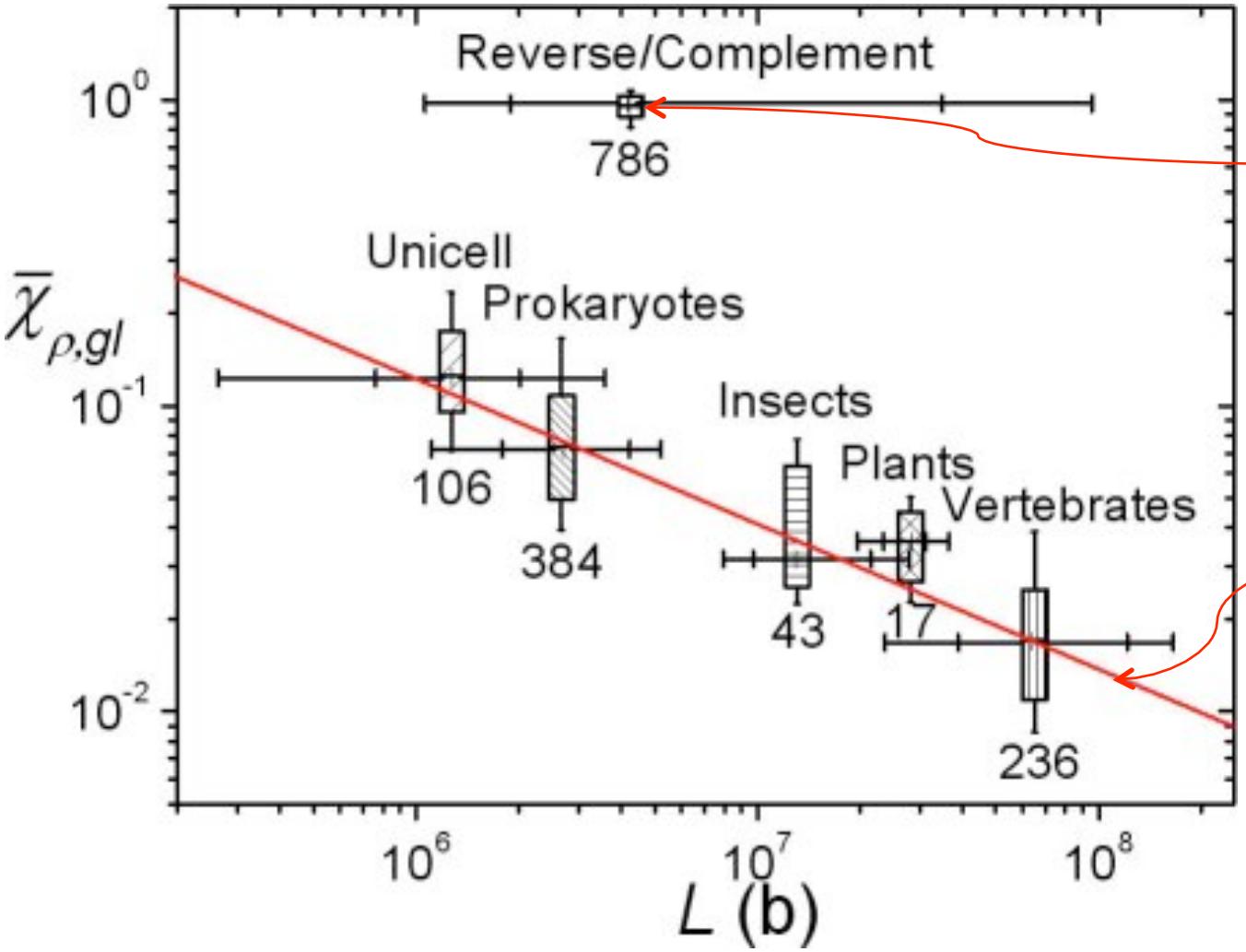
$\chi_{background}$

χ_{global}

$r_\chi = \chi_{background} (\chi_{global})^{-1}$

Global and local symmetry not the same

No reverse/complement symmetry but strong global inverse symmetry

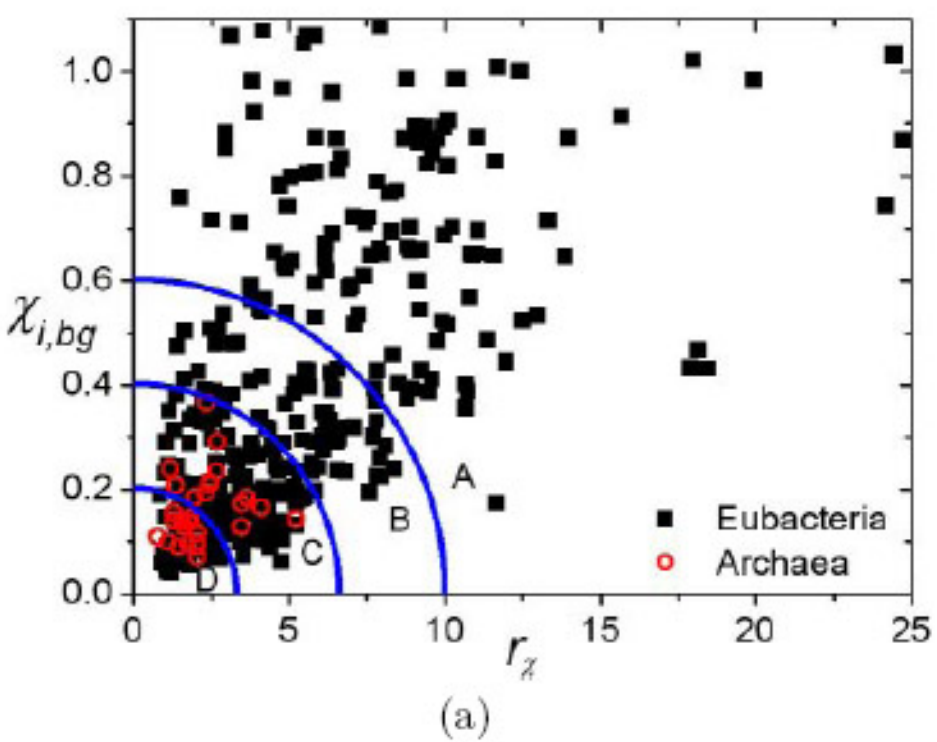


Reverse & Complement: $\chi \sim 1$ for all genomes

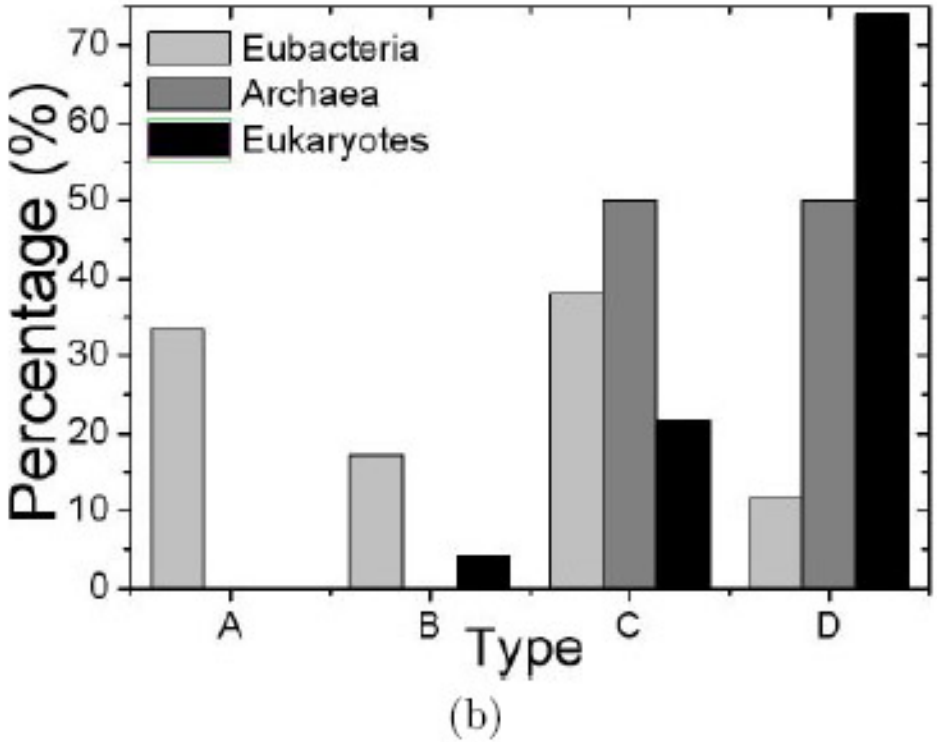
Inverse: $\chi << 1$ for all genomes

Global $\chi_{inverse}$ scales with chromosome length

Classification by $\chi_{inverse}$



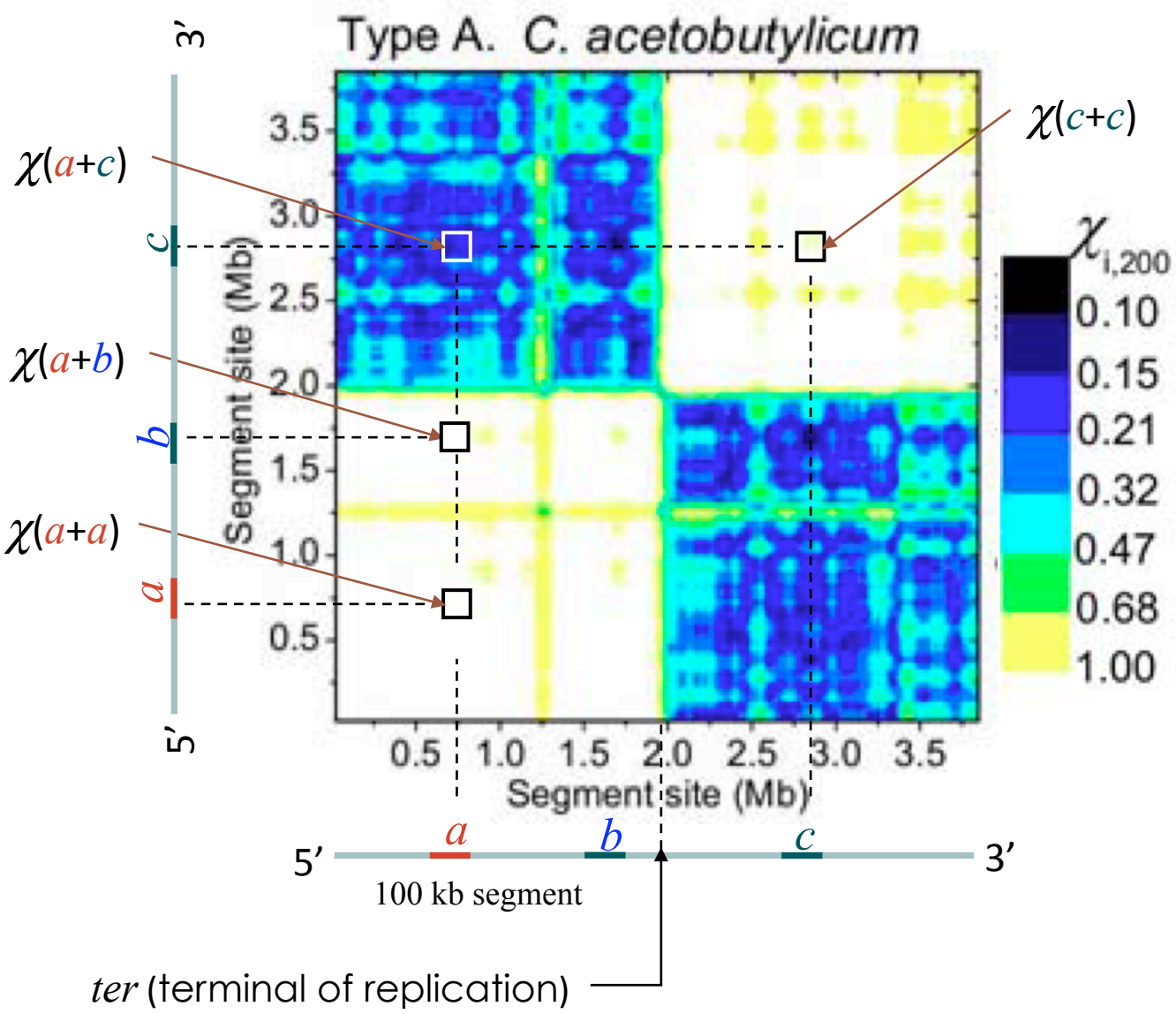
(a)



(b)

- 50% eubacteria types A & B
- Archaea all types C & D
- Eukaryotes: multi-cells type D; all types B & C are unicells

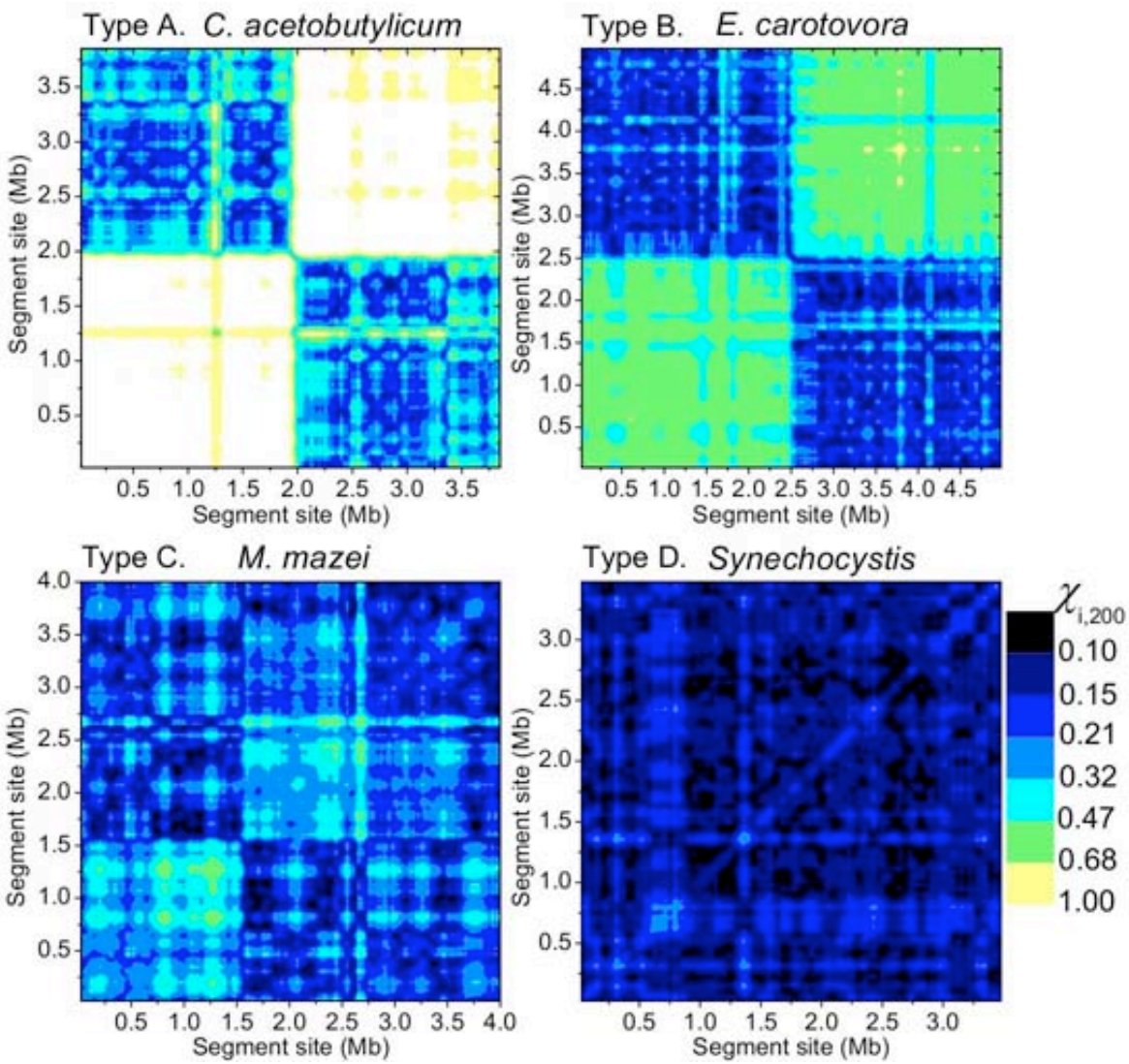
The χ-plot



- $\chi(a+a) \sim \chi(c+c) \sim 1$
  no local IS
- $\chi(a+b) \sim 1$
  no mutual IS if both segments from same half
- $\chi(a+c) \sim 0.2$
  high mutual IS if segments from different half
- Inference:
  - No local IS anywhere
  - Fore and aft of chromosome have high mutual IS

Four types of chromosomes have characteristic χ-plots

Type A: No local IS, bisected by *ter*/*ori* sites into two mutual high IS halves

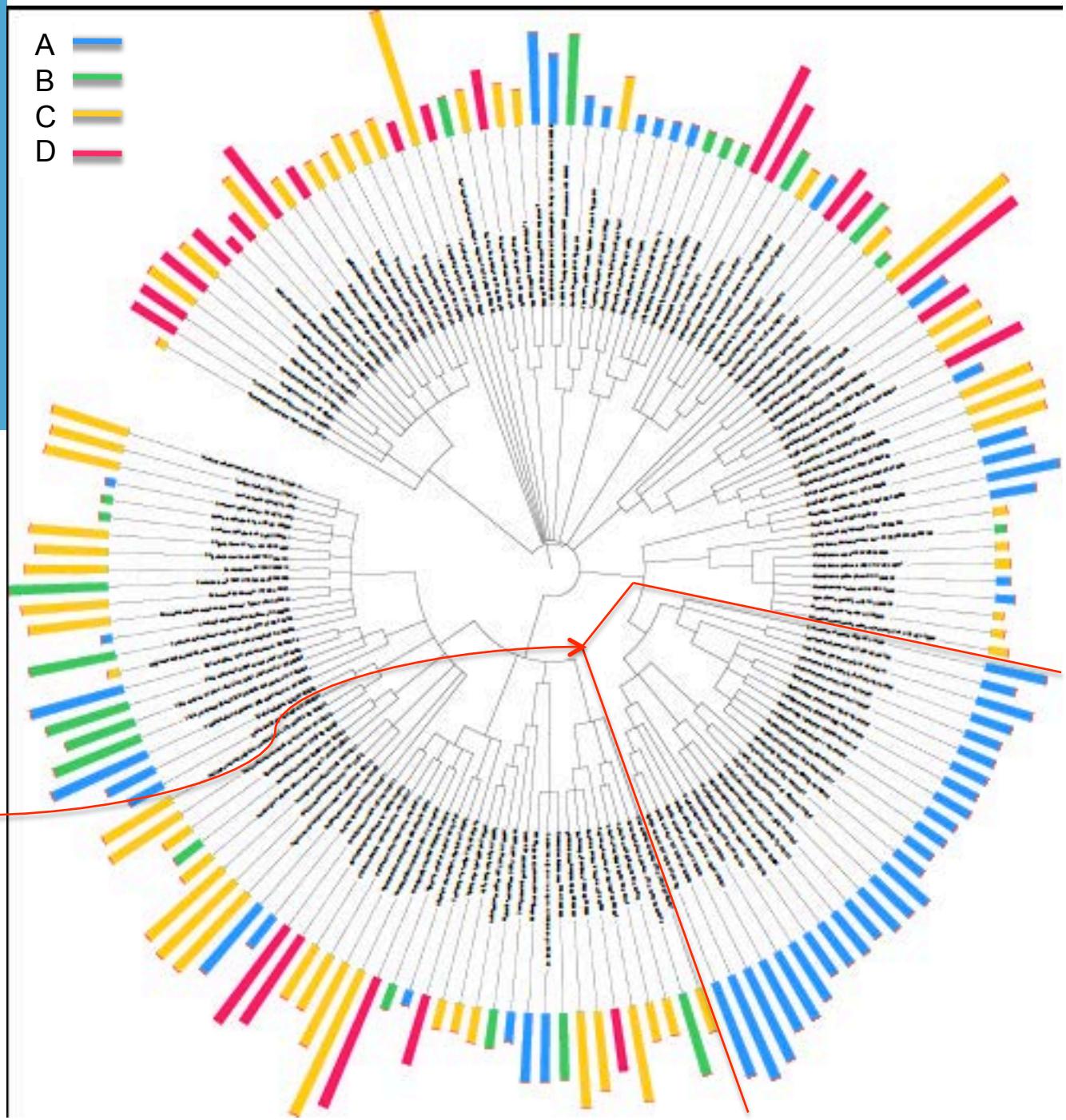


Type B: like type A, but higher local IS background

Type C: Hybrid of Band D

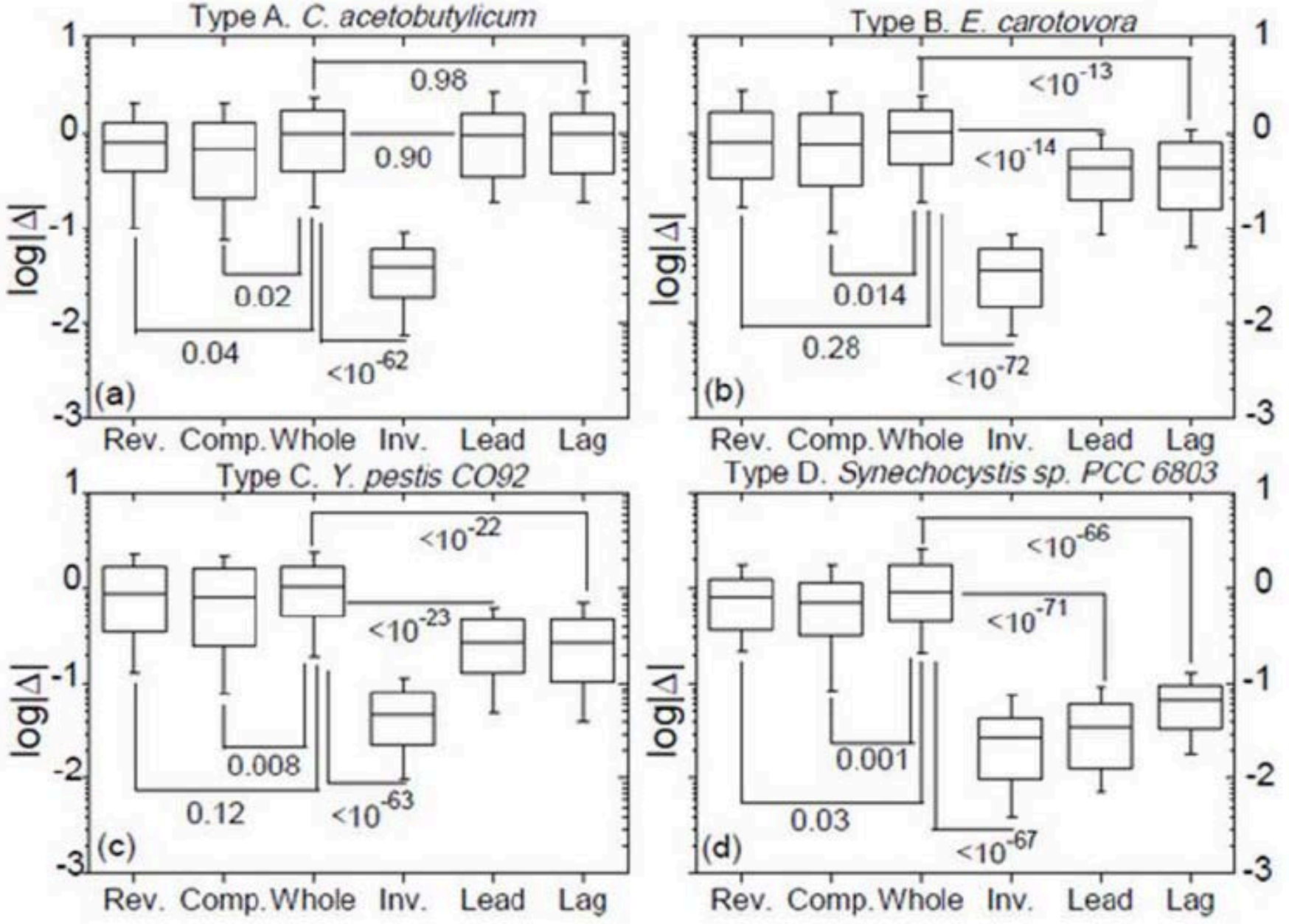
Type D: Homo-geneous high local & mutual IS

IS type correlated with taxonomy

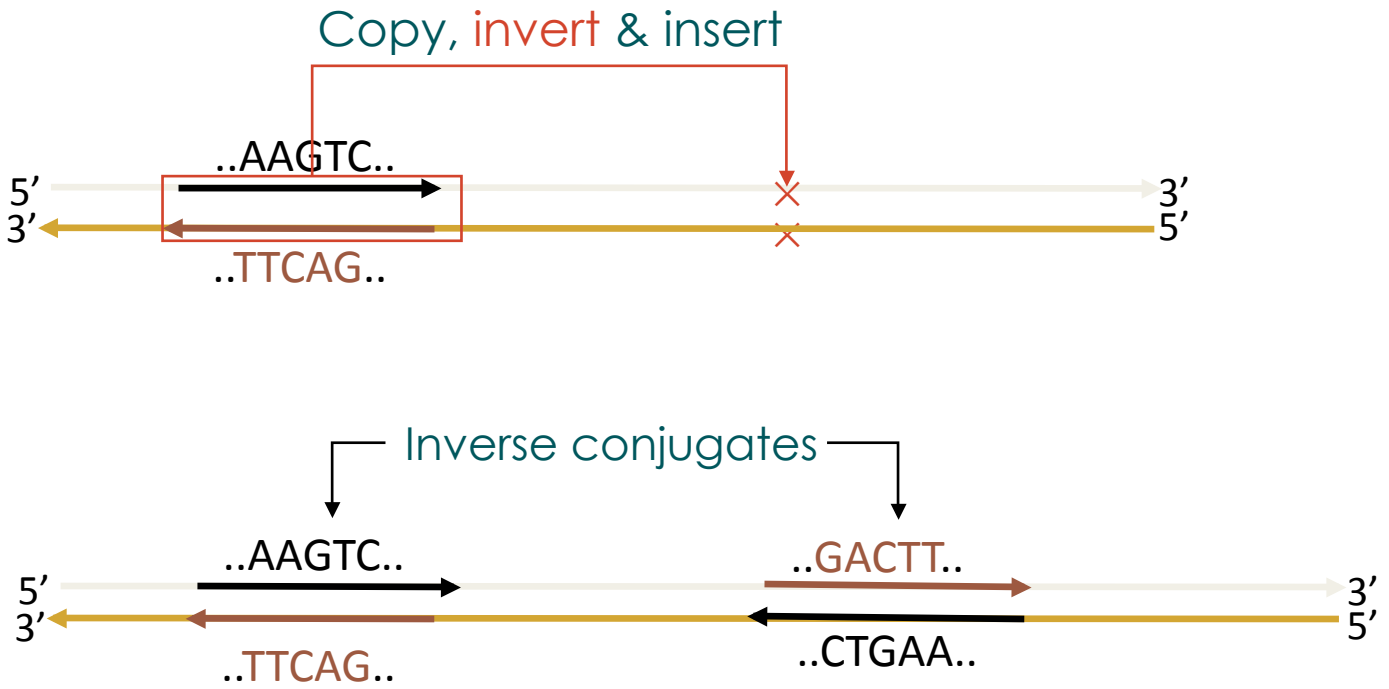


Phylum Firmicutes: all are type A

P-values for distributions of $\Delta_u = |f_u - f_{u^\dagger}| / \sigma_{m_u}$

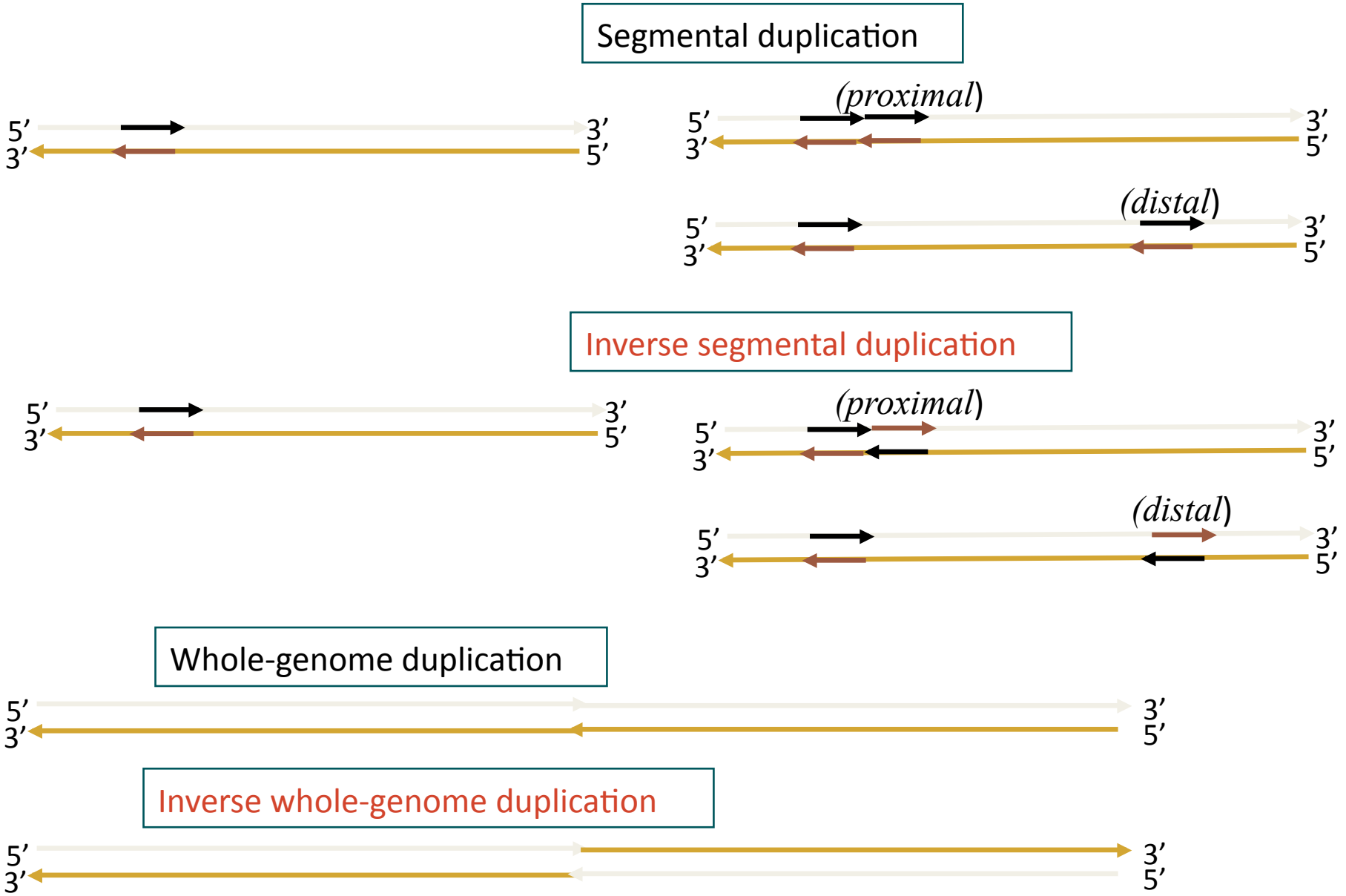


Inverse segmental duplication (ISD) generates IS



Absence of similar mechanism for generating reverse/ complement symmetries may explain their absence

Types of segmental duplications

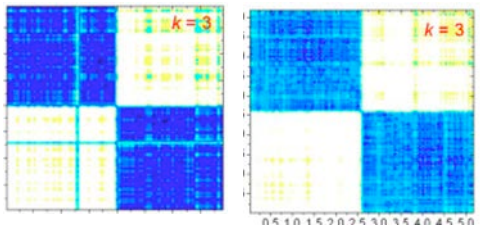


Proximal and *distal* ISDs generate different effects

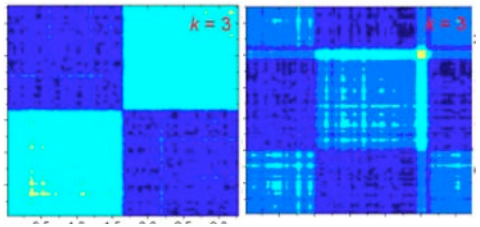
Distal ISD enhance global symmetry but not local symmetry

Proximal ISD enhance local symmetry and global symmetry

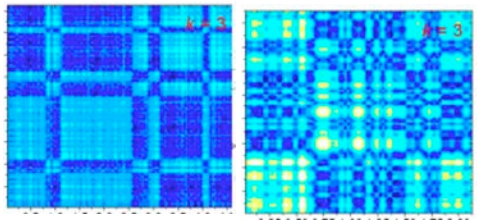
Inferences from χ-plots on chromosome evolution



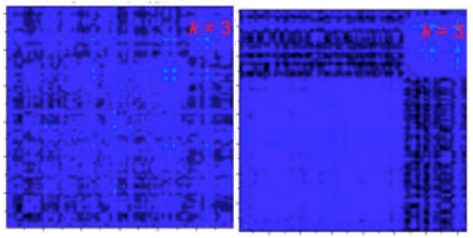
- Whole-chromosome ISD (WISD)
- Few *dist*-SDs (but possibly many *prox*-SDs)
- Very few *prox*-ISDs



- WISD
- Few *dist*-SDs
- Low to medium level of *proximal* SDs

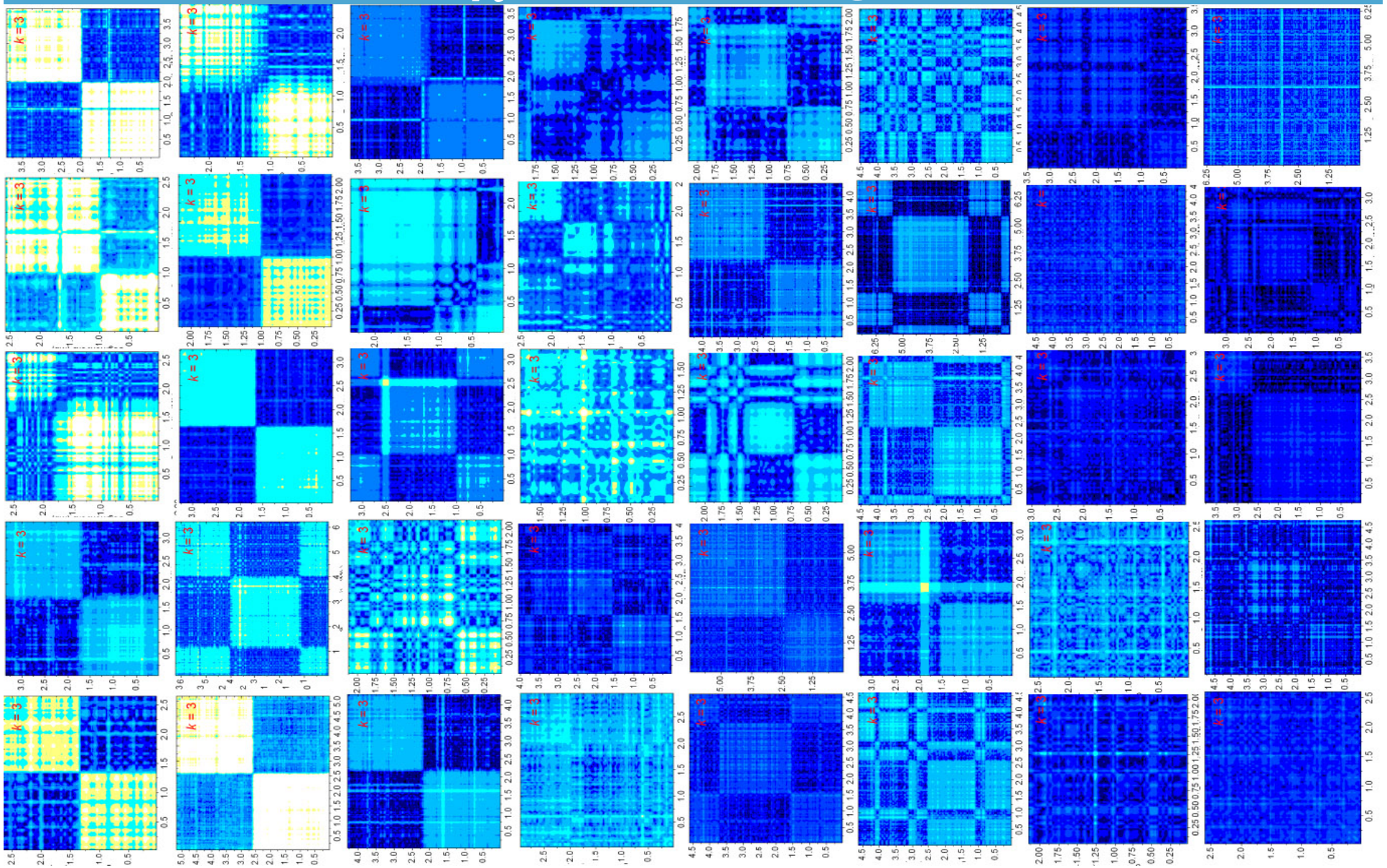


- WISD
- Some *dist*-SDs (or chrom. Re-arrangements)
- Various level of *prox*-ISDs

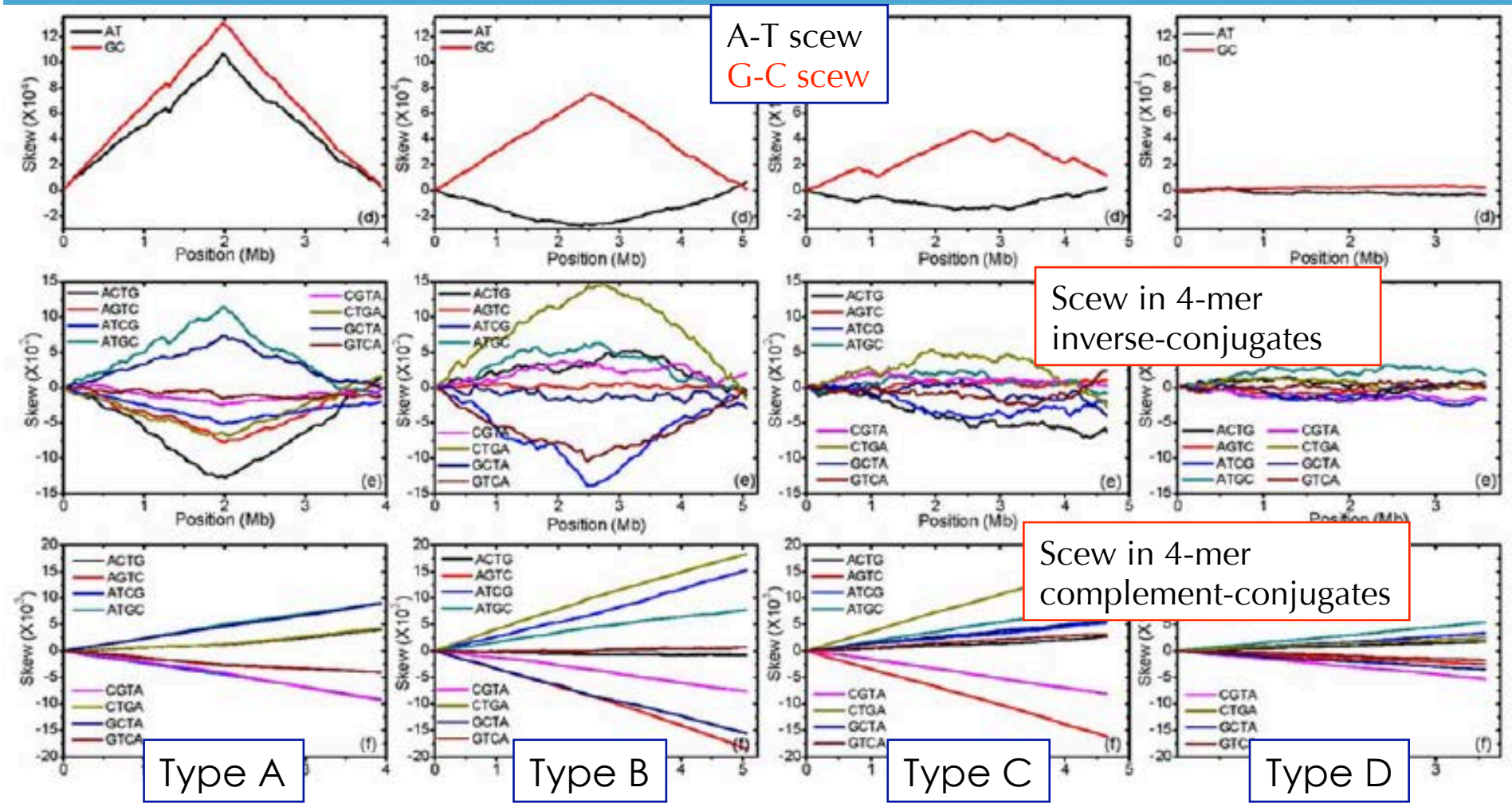


- With or w/o WISD
- Unconstrained SDs
- Saturating amount of *prox*-ISDs

Mosaic of prokaryotic χ-plots invites "χ-archeology"



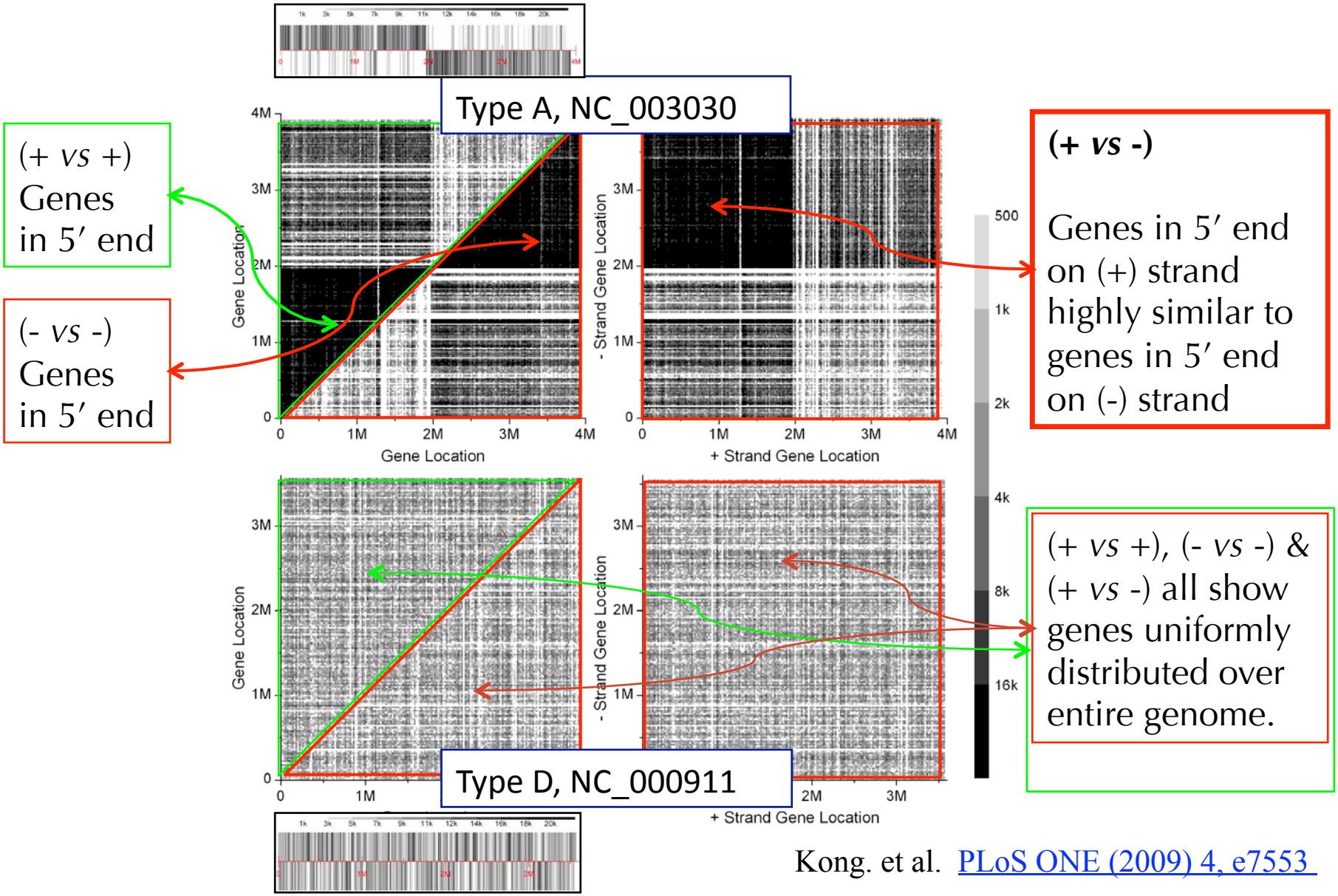
k-mer skews



$$|f_u - f_{u^\dagger}| \approx |f'_u - f'_{u^\dagger}| \approx \chi_{i,bg}\sqrt{2}\sigma_{m_u}, \qquad (6)$$

From universality $\sqrt{2}\sigma_m \approx \sqrt{2}\bar{\sigma} \approx 0.20L(3.2)^{-k}$.

BLASTing genes on (+ vs +), (- vs -), and (+ vs -) strands



Kong. et al. PLoS ONE (2009) 4, e7553

Mean-field approximation

Let $\langle \Delta f \rangle = \sqrt{2}\bar{\sigma}$ be the average frequency difference of an unrelated pair (of k-mers). This means a typical pair of k-mers have respective frequencies $f_- = \bar{f} - \langle \Delta f \rangle/2$ and $f_+ = \bar{f} + \langle \Delta f \rangle/2$, so that on average the fraction of unrelated k-mers that are "paired-up" is

$$2f_-/\bar{f} = 2v_0 = (\bar{f} - \langle \Delta f \rangle/2)/\bar{f} = 1 - \sqrt{2}\bar{\sigma}/2\bar{f}.$$

Similarly, if $\langle \Delta f \rangle_\rho$ is the average frequency difference in a ρ-conjugate pair, then the fraction of k-mers paired with their respective ρ-conjugate partners is

$$2v_\rho = \left(\bar{f} - \langle \Delta f \rangle_\rho/2\right)/\bar{f} = 1 - \langle \Delta f \rangle_\rho/2\bar{f}. \qquad (3)$$

The mean field value of χ_ρ is, by definition

$$\langle \chi_\rho \rangle = \langle \Delta f \rangle_\rho / \sqrt{2}\bar{\sigma} = (\langle \Delta f \rangle_\rho / 2\bar{f})(\sqrt{2}\bar{\sigma}/2\bar{f})^{-1} = (1-2v_\rho)/(1-2v_0).$$

Mean-field interpretation of χ_{inv}

- In mean-field approximation, fraction of a sequence that causes it to manifest inverse-inverse symmetry is

$$2\,\delta v_{\text{inv}} \sim (1\text{-}2v_0)(1\text{-}<\chi_{\text{inv}}>)$$

$(1\text{-}2v_0)$ is the fraction of unrelated k-mers that are paired-up. With $<\chi_{\text{inv}}> \sim 0.073 +/- 0.066$

$$2\delta v_{inv} \sim 0.25 +/- 0.15$$

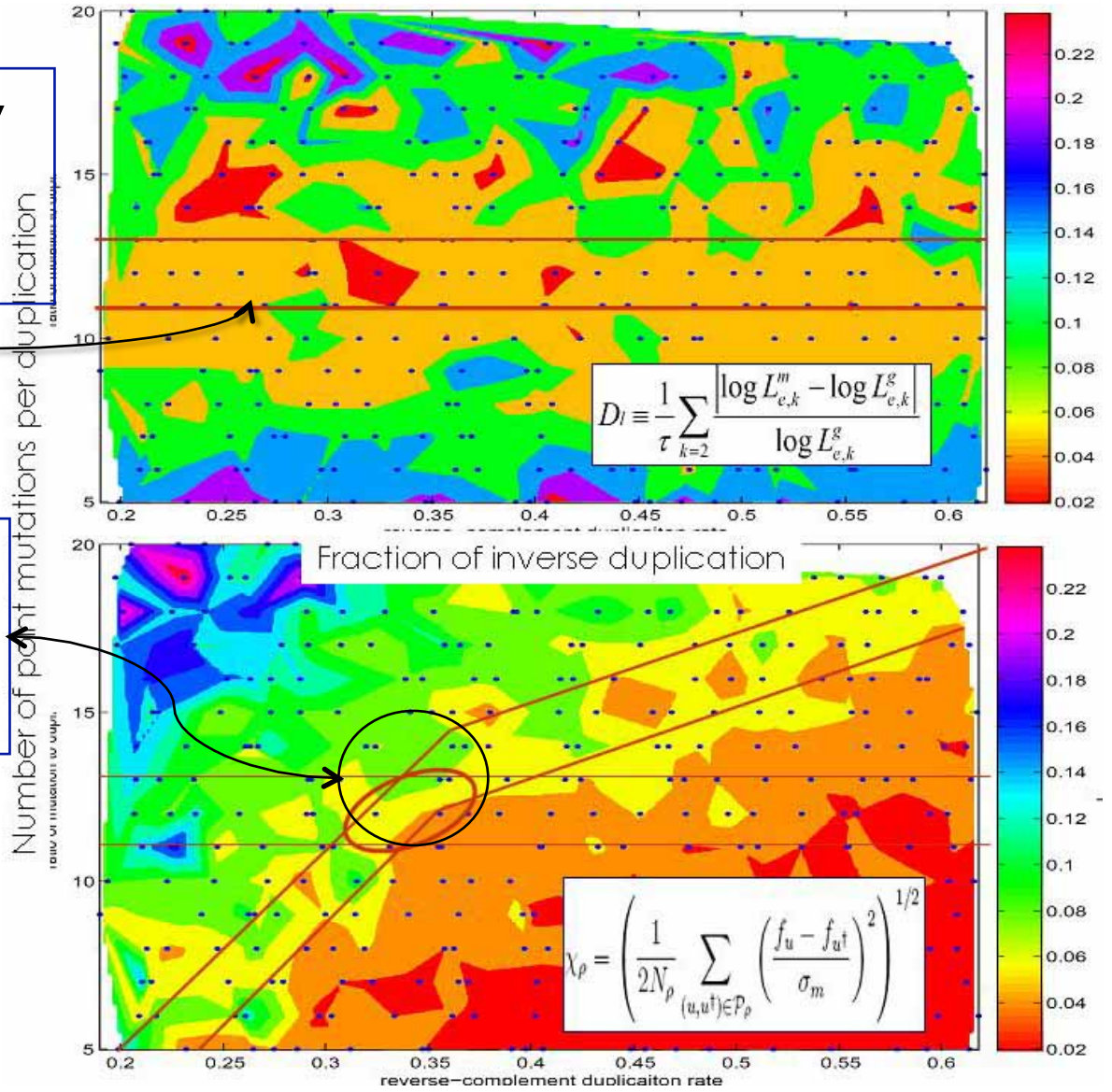
That is, about one quarter of chromosomes is composed of ISD generated segments.

Genome growth model based on random segmental duplication

Band fits universality class of equivalent length

Fraction of inverse duplication

~ $2\delta v_{inv}$ ~ 0.34+/ - 0.05



Minimal genome growth model

Hsieh et al. PRL (2003) 018101-104.
Chen et al. PRL (2005) 178103-106.
Kong et al. PRE (2009) 061911 (1-11).

Summary I

- All chromosome have strong global IS
- Local and global reverse symmetry (RS)and complement symmetry (CS) are absent in genomes
- Chargaff's second parity rule is a reduction of global IS, not of CS
- (Large scale) IS is likely generate by ISD
- There is no comparable mechanism that can generate RS or CS

Summary II

- IS of chromosomes have types
  - Type A: No local IS; whole genome inverse duplication (WGID)
  - Type D: Only local IS; saturating amount of *prox*-ISDs
- Skews in single base AND *k*-mers conjugates can be quantitatively accounted for in terms of IS
- Origin & terminal sites play prominent roles in IS
- Model of genome growth based on SD and ISD can explain genomic patterns IS

Issues remaining

- IS-type classification has clear correlation with taxonomy, and suggests an ancient origin of WGID. Attempt to determining time of occurrence under way.
- Not sure why global IS is so strong. Does promote equal utilization of both strands, but A and D types represent different strategies

End of Lecture Three

Related papers

- Inverse symmetry in genomes and whole-genome inverse duplication. SG Kong, et al. PLoS ONE (2009) 4(11): e7553.
- Quantitative measure of randomness and order for complete genomes. SG Kong et al. Phys. Rev. E 79 (2009) 061911.
- Universal Global Imprints of Genome Growth and Evolution -- Equivalent Length and Cumulative Mutation Density. HD Chen et al. PLoS ONE (2010) 5(4): e9844.
- Divergence and Shannon information in genomes
  HD Chen et al. Phys. Rev. Lett. 94 (2005) 178103.
- Minimal model for genome evolution and growth
  LC Hsieh et al. Phys. Rev. Lett. 90 (2003) 018101-104.

- People
  - Sing-Guan Kong, Physics, Systems Biology, NCU
  - Hong-Da Chen, Physics, Systems Biology, NCU
  - Wen-Lang Fan, Physics, NCU; Genome Center, Sinica
  - Zi-Ting Hsu, Systems Biology, NCU
  - Neng-Ji Zhou, Inst. Mod. Phys., Zhejiang U.
  - Prof. B. Zheng, Inst. Mod. Phys., Zhejiang U.

- Financial support from
  - National Science Council
  - Ministry of Education
  - National Center for Theoretical Sciences
  - National Central University