Is inversion symmetry of chromosomes a law of nature?

David Horn

TAU Safra bioinformatics retreat, 28/6/2018

Lecture based on

Inversion symmetry of DNA k-mer counts: validity and deviations.

Shporer S, Chor B, Rosset S, Horn D

BMC Genomics. 2016 Aug 31

and D Horn, Atlas of Science, April 14, 2017.

Laws of Nature

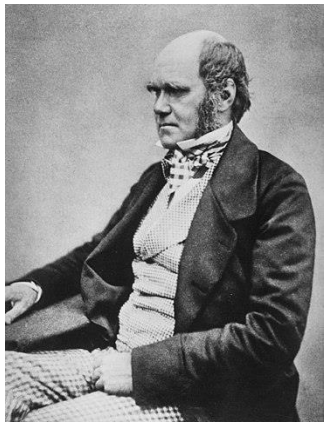
- Physics: Boyle's law of gases, Newton’s Laws of Motion, Maxwell Laws of Electromagnetism, Energy Conservation, etc.

- Biology: Darwin’s Natural Selection.

From The Origin of Species, 1859:

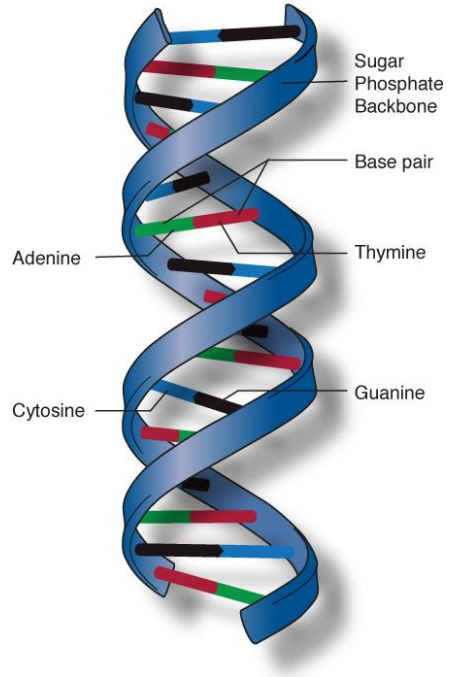
if variations useful to any organic being do occur, assuredly individuals thus characterised will have the best chance of being preserved in the struggle for life; and from the strong principle of inheritance they will tend to produce offspring similarly characterised.

This principle of preservation, I have called, for the sake of brevity, Natural Selection.



Erwin Chargaff has made, in 1950, the important observation that the numbers of nucleotides in DNA satisfy
#A = #T and #G = #C.

This played an important role in understanding the double-helix structure of DNA.



- Chargaff E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia. 1950;6(6):201–9.
- Chargaff E. Structure and function of nucleic acids as cell constituents. Federal Proc. 1951;10:654–9.
- Crick F, Watson JD. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171:737–8.

Second Chargaff rule (SCR), in 1968, states that #A = #T and #G = #C **holds for each string separately.**

- Rudner R, Karkas JD, Chargaff E. Separation of B. subtilis DNA into complementary strands. III. Direct Analysis. Proc Natl Acad Sci U S A. 1968;60:921–2.
- Mitchell D, Bridge R. A test of Chargaff's second rule. Biochem Biophys Res Commun. 2006;340(1):90–4.

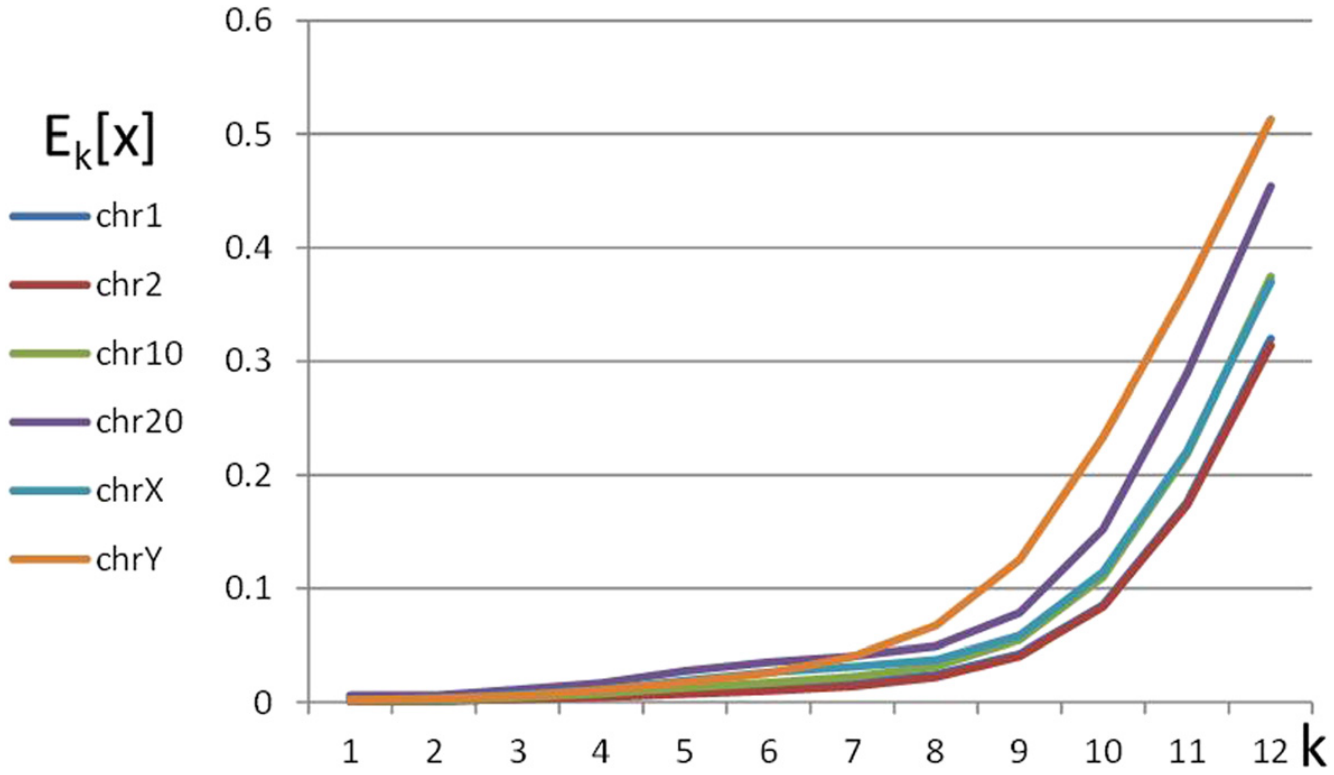
Inversion Symmetry (IS): the counts of a k-mer of nucleotides on a chromosomal strand are almost equal to those of its inverse (reverse-complement) string.

$$X = |N(S)-N(S^*)|/(N(S) + N(S^*)) \to 0$$

Reverse
CGA->AGC

Complement
CGA->GCT

Inverse
CGA->TCG



$$E\_k[X] = \Sigma\, X(S,S^*)/M\_k$$

where M_k is the number of different k-mers encountered empirically (roughly 4**k).

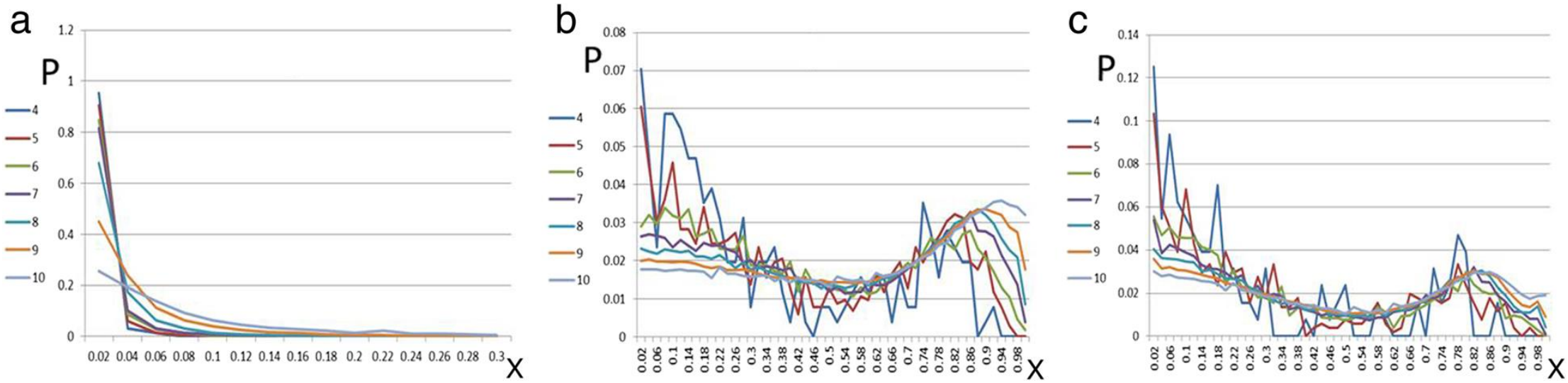
Reverse
CGA->AGC

Complement
CGA->GCT

Inverse
CGA->TCG

Only inverse works

HG38 chr1: Histogram (probability distribution in bins of Δx = 0.02) of relative occurrences of k-mer pairs vs x for different values of k (4 to 10).
a inverse pairs; plotted range is x < 0.3, above which the histogram values are negligibly small.
b random pairs for full x range;
c Reverse pairs for full x range



Statistical Analysis

Three stochastic variables

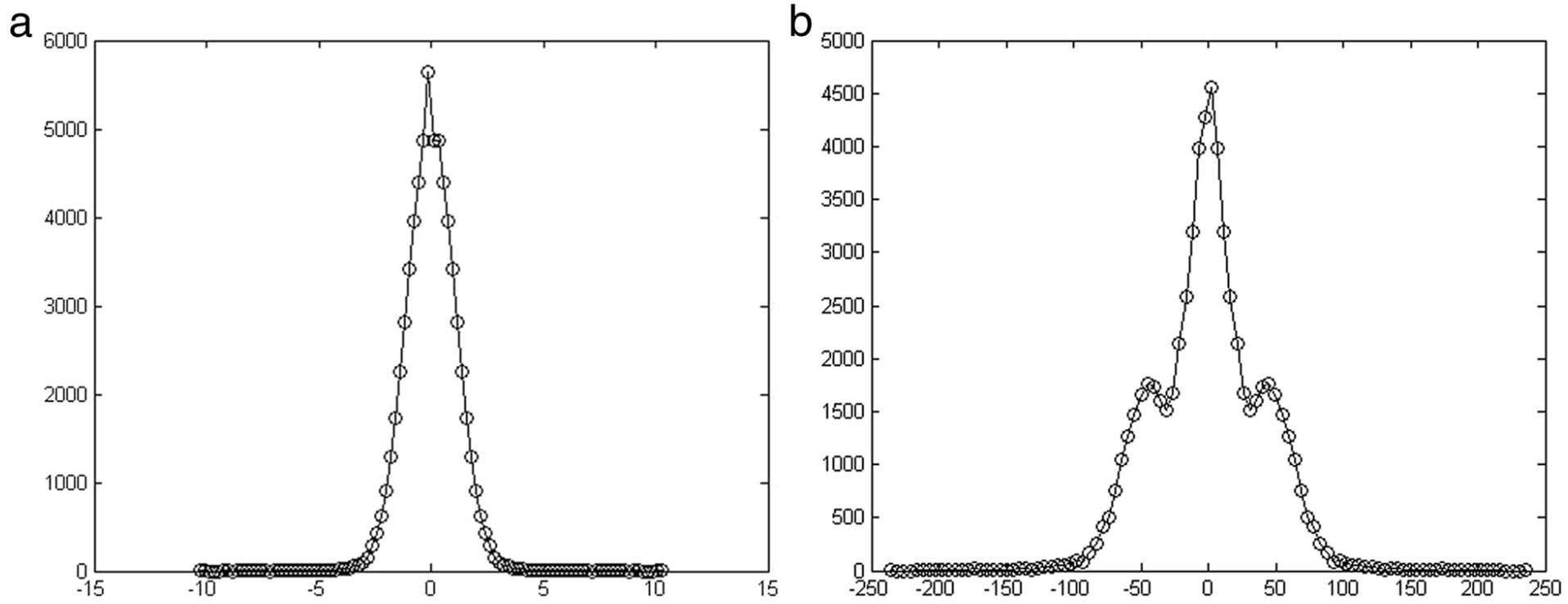
$$X = \frac{|N(S) - N(S*)|}{N(S) + N(S*)}$$

$$Y = \frac{N(S) - N(S*)}{N(S) + N(S*)}$$

$$Z = \frac{N(S) - N(S*)}{\sqrt{(N(S) + N(S*))}}$$

If both S and S* follow the same Poisson distribution along the chromosome, the **statistic Z will be standard normal** (Gaussian with mean=0 and variance=1).

A: Z histogram for inverse pairs, k=8.
B: Z histogram for reverse pairs. Peak due to palindroms. k=8



K	1	2	3	4	5	6	7	8	9	10	
E(|Z|)	4.2	2.6	1.7	1.4	1.2	1.1	1.0	0.93	0.88	0.84	Validity
E(X)	0.0004	0.0005	0.0007	0.0013	0.0025	0.0047	0.0094	0.019	0.040	0.084	Accuracy

Results of analysis of human chr 1

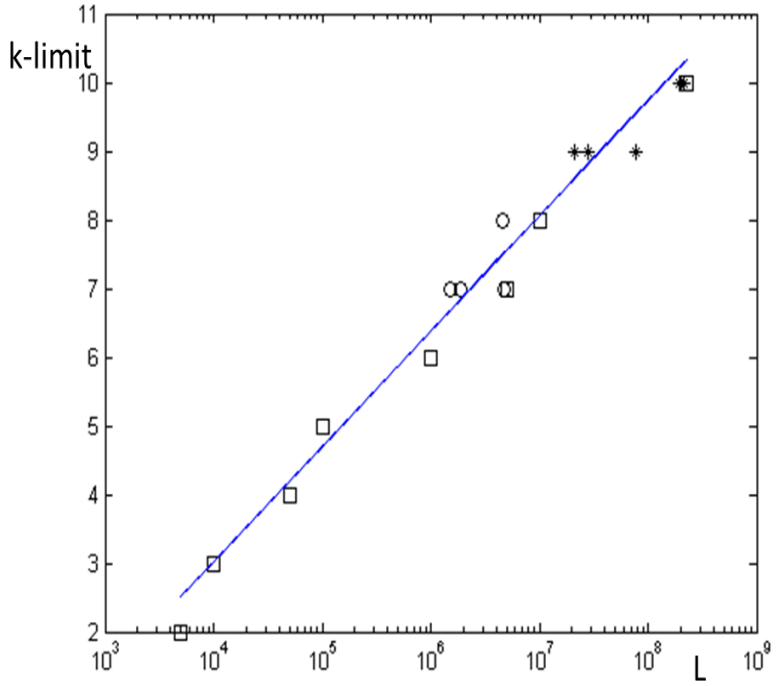
IS-Poisson predicts E(|Z|)=0.8, with standard deviation of 0.6.

- For low k, statistical equality is invalid, yet accuracy of E(X) =0 is high
- For high k, statistical equality cannot be refuted, but accuracy is low

Conclusion: IS and SCR are broken at the level of 0.001.

Using IS-Poisson one can prove that the k-limit (for which E_k[X] =0.1) obeys KL = lnL/ln4 + const = 0.72lnL + const.

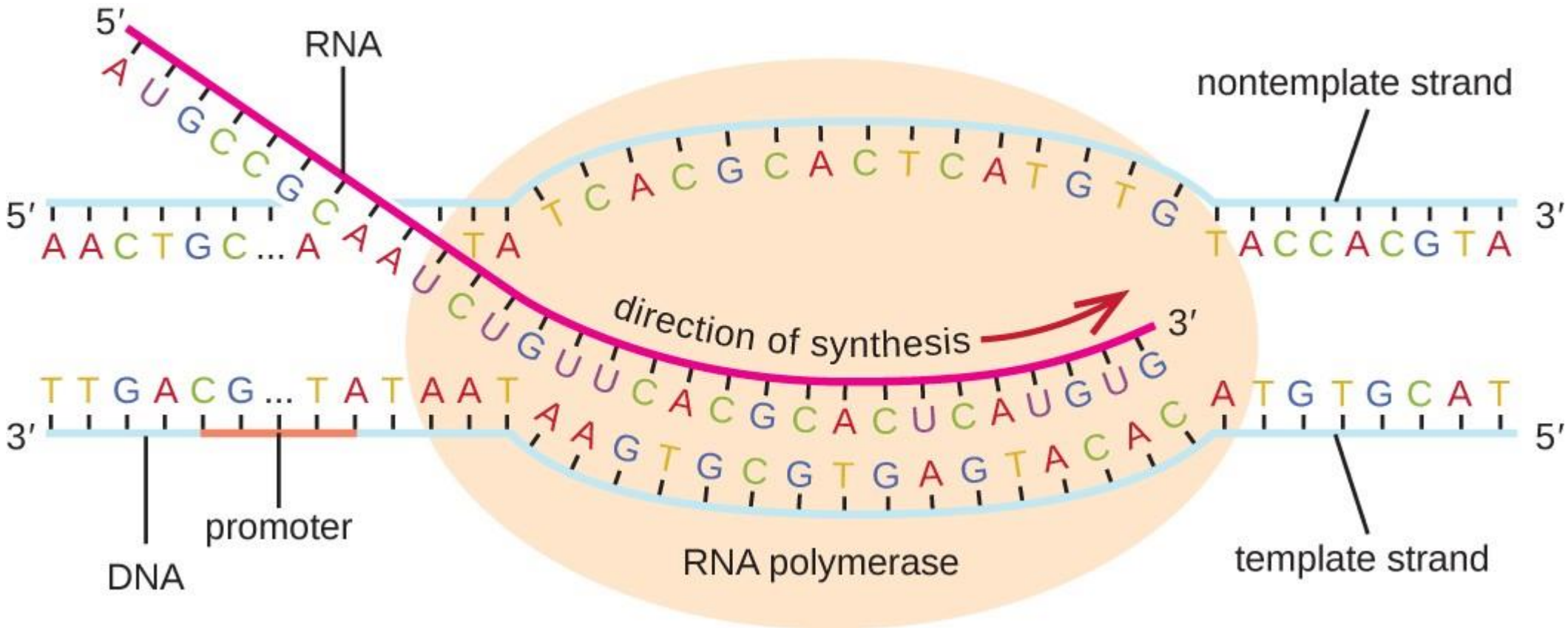
species	length	k-limit
HG38.chr1	230479627	10
HG18.chr1	224999368	10
chimp.panTro2.chr1	217189828	10
mouse.mm10.chr1	191908761	10
HG18.chrX	151058618	9
zebrafish.danRer6.chr7	76727960	9
melanogaster.dm3.chr3R	27905045	9
elegans.ce10.chrV	20924149	9
HG18.chrY	25652849	8
human section of 10M	10000000	8
Escherichia_coli_K_12_substr__W3110	4646325	8
Bacillus_subtilis_uid76	4215599	8
human section of 5M	5000000	7
Mycobacterium_avium_paratuberculosis	4829775	7
Pyrococcus_furiosus_uid287	1908250	7
Thermotoga_maritima_uid111	1860719	7
cerevisiae.sacSer3.chrIV	1531933	7
human section of 1M	1000000	6
human section of 100K	100000	5
human section of 50K	50000	4
human section of 10K	10000	3
human section of 5K	5000	2



k-limits *vs* chromosomal length, display universal logarithmic behavior. Boxes are human data, stars denote other eukaryotes, and circles represent prokaryotes. The shown fit to this set of data is 0.73*ln(length), and should serve as an indication of the observed logarithmic increase of the k-limits.

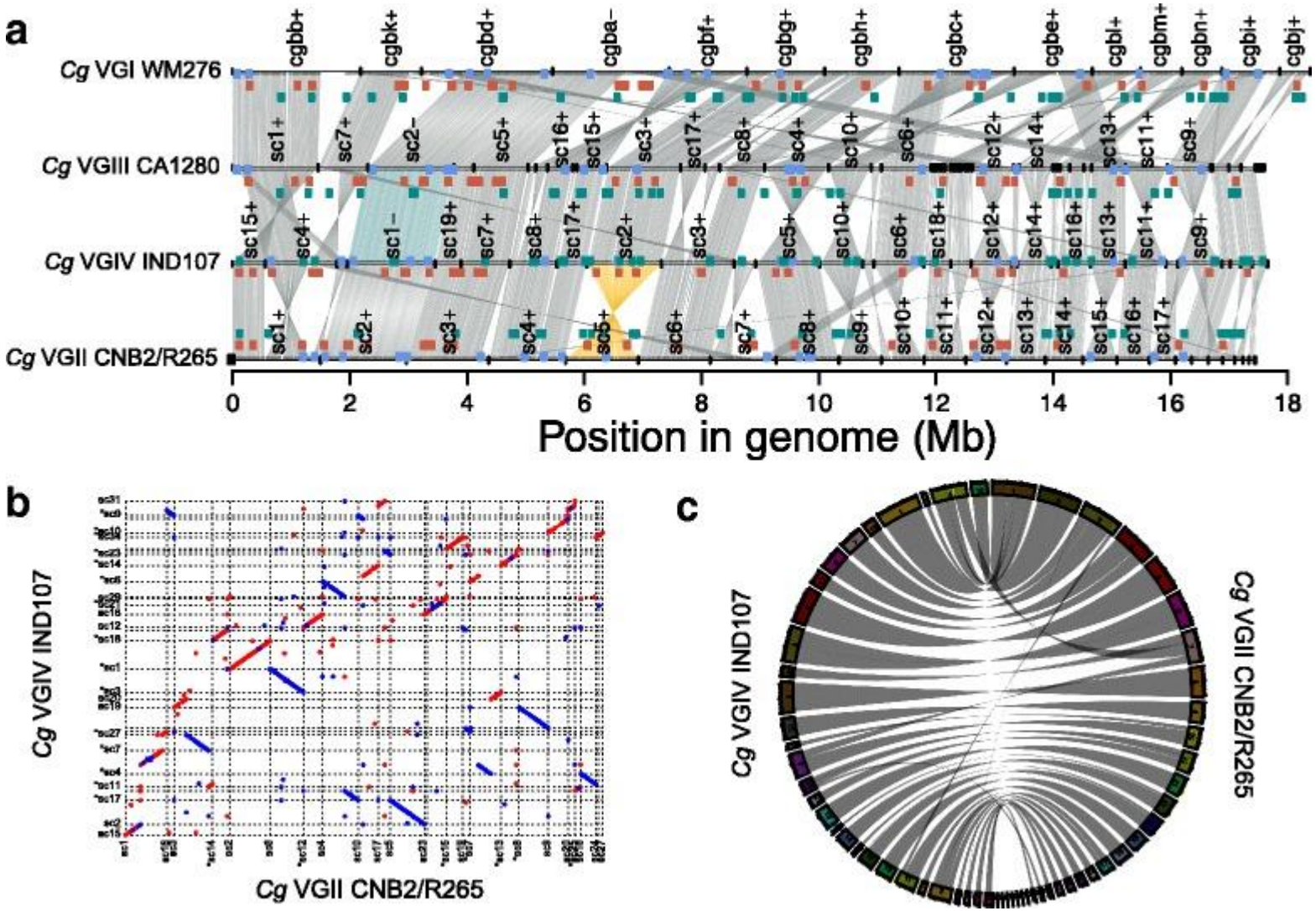
Further Questions:

- How did Inversion Symmetry come about?
- Is there a biological meaning to its breaking?

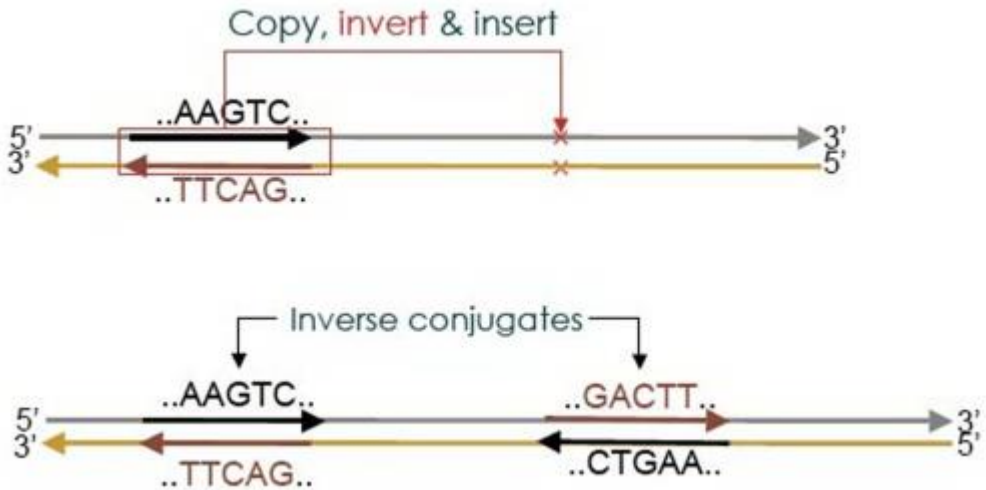

5′
RNA
A U G C C G C A A U C U G U U C A C G C A C U C A U G U G C
nontemplate strand
5′
A A C T G C ... A T A
T C A C G C A C T C A T G T G
T A C C A C G T A
3′
direction of synthesis
3′
T T G A C G ... T A T A A T
A A G T G C G T G A G T A C A
A T G T G C A T
3′
5′
promoter
DNA
RNA polymerase
template strand

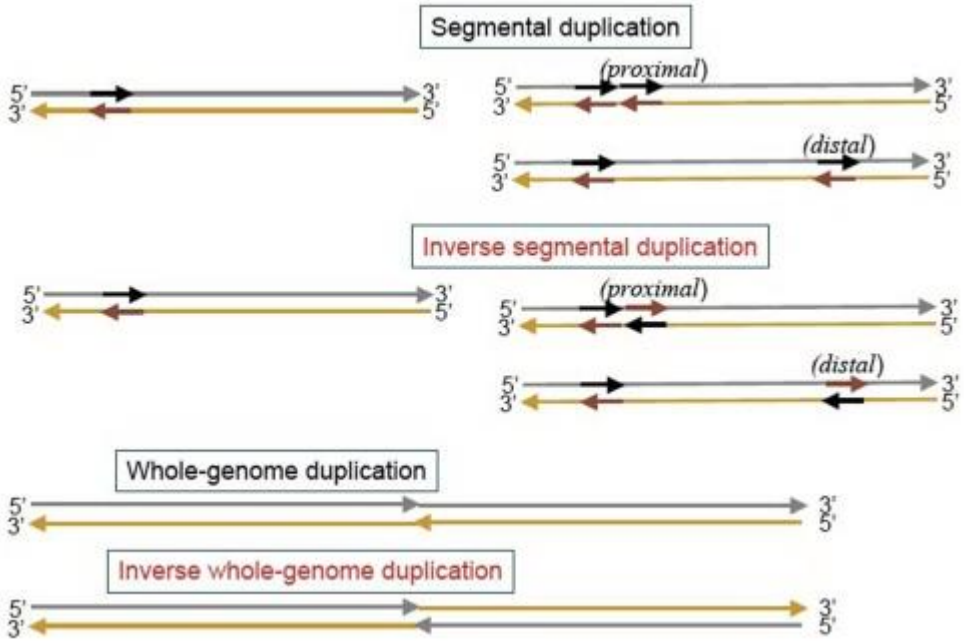
Synteny imaging tool
R A Farrer BMC Bioinformatics. 2017; 18: 507.

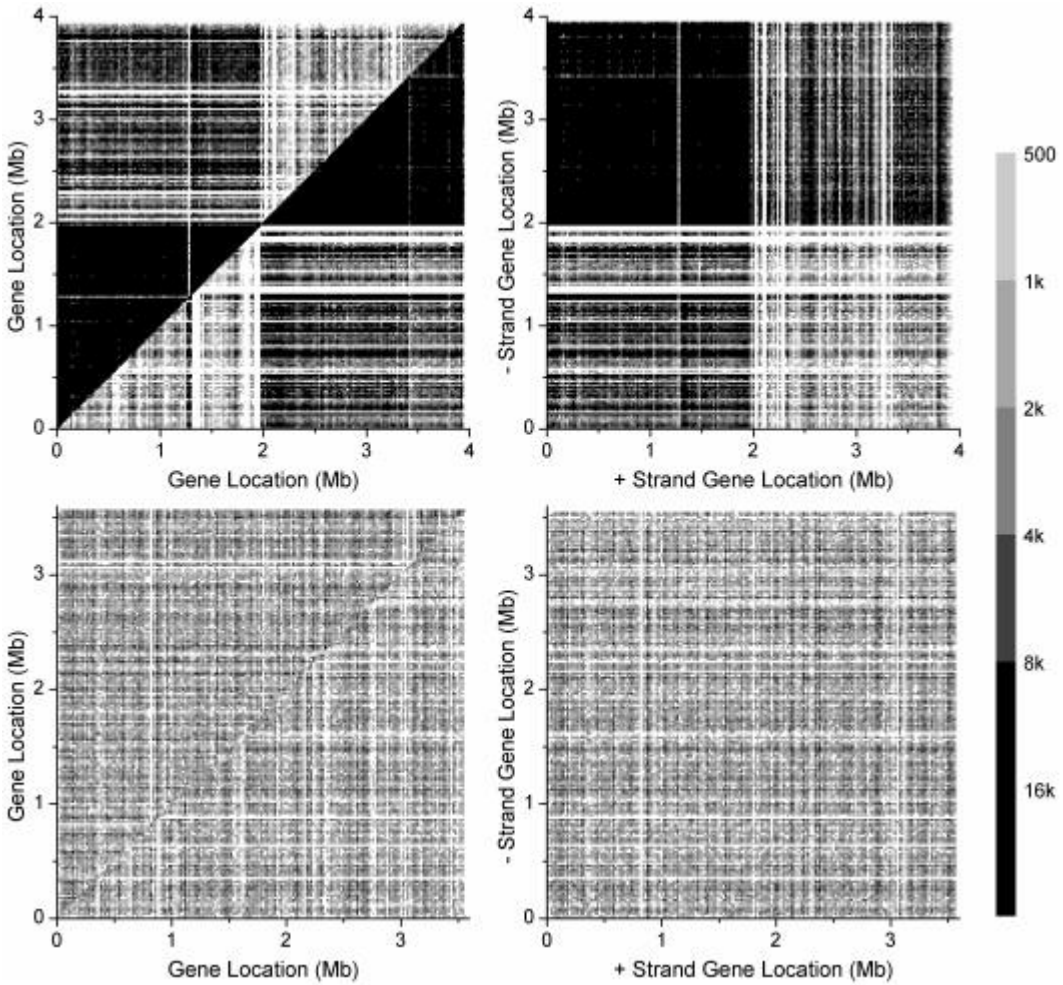
Synteny is shown for four genomes representing each of the four lineages of the pathogenic fungus *Cryptococcus gattii*



Kong S-G, Fan W-L, Chen H-D, Hsu Z-T, Zhou N, Zheng B, and Lee H-C (2009). Inverse symmetry in complete genomes and whole-genome inverse duplication. PlosOne 4, e7553.



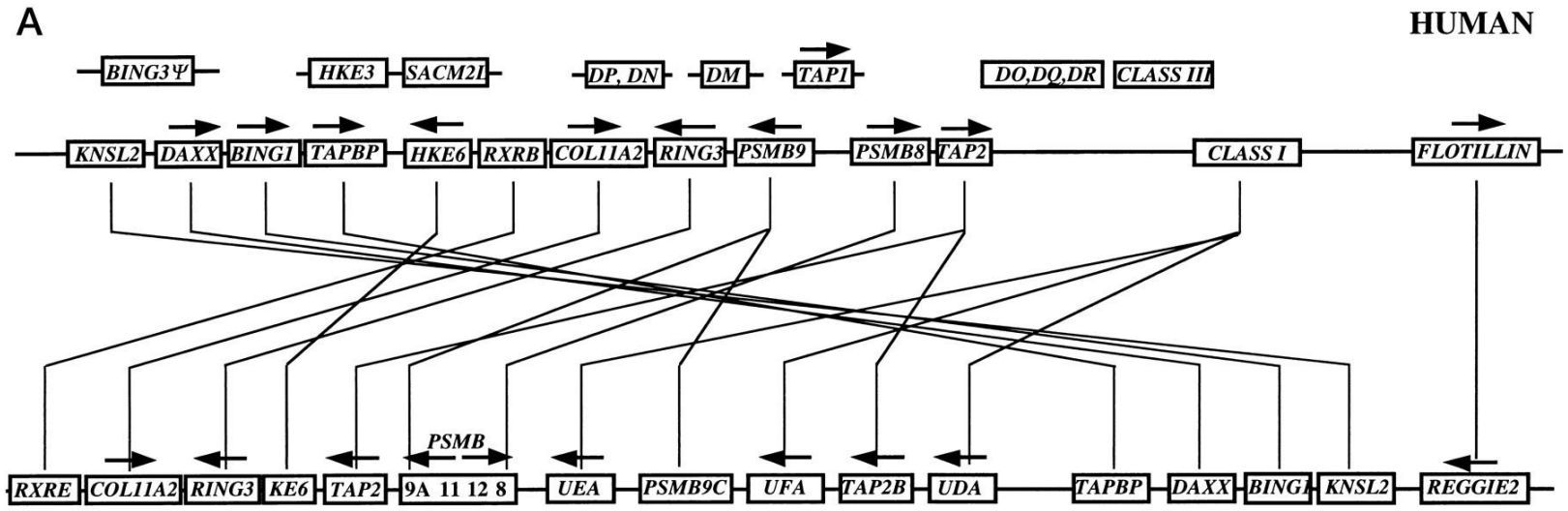


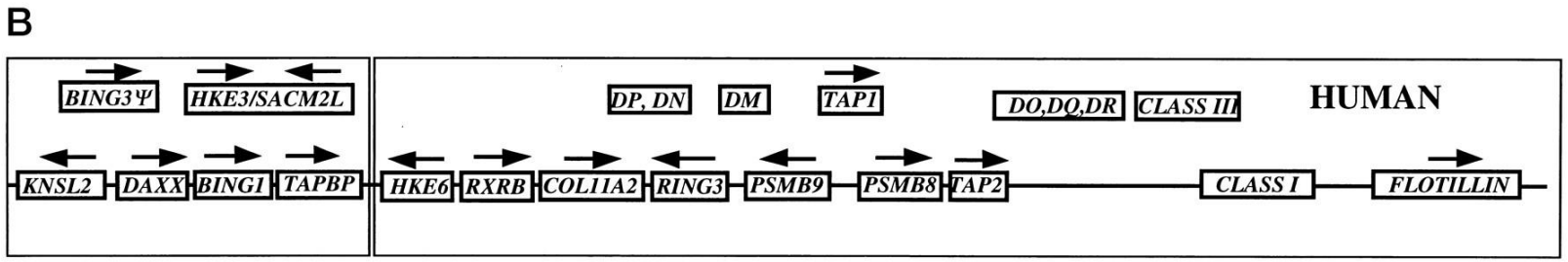


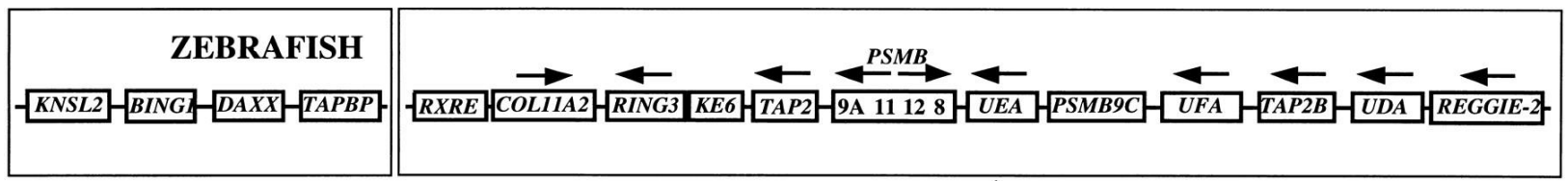
BLAST plots of homologs in *C. acetobutylicum* and *Synechocystis*.The top pair of plots are for *C. acetobutylicum* and the bottom plots pair are for *Synechocystis*. In each plot, coordinates are sites of homologs on the chromosome.

Plots on left: top-left (bottom-right) triangle gives BLAST scores for intra-strand homologs on the positive (negative) strand; pixels on the diagonals, which include very high scores from same-gene BLASTs, are removed.

Plots on right: BLAST scores for inter-strand-homologs; x-axis (y-axis) gives sites on the positive (negative) strand. The bottom plots suggest a relatively low level of homology in the type-D *Synechocystis* for both inter-strand and inter-strand pairs.


A
HUMAN
BING3Ψ
HKE3
SACM2L
DP, DN
DM
TAP1
DO,DQ,DR
CLASS III
KNSL2
DAXX
BING1
TAPBP
HKE6
RXRB
COL11A2
RING3
PSMB9
PSMB8
TAP2
CLASS I
FLOTILLIN
PSMB
RXRE
COL11A2
RING3
KE6
TAP2
9A 11 12 8
UEA
PSMB9C
UFA
TAP2B
UDA
TAPBP
DAXX
BING1
KNSL2
REGGIE2
ZEBRAFISH


B
BING3Ψ
HKE3/SACM2L
DP, DN
DM
TAP1
DO,DQ,DR
CLASS III
HUMAN
KNSL2
DAXX
BING1
TAPBP
HKE6
RXRB
COL11A2
RING3
PSMB9
PSMB8
TAP2
CLASS I
FLOTILLIN


ZEBRAFISH
KNSL2
BING1
DAXX
TAPBP
PSMB
RXRE
COL11A2
RING3
KE6
TAP2
9A 11 12 8
UEA
PSMB9C
UFA
TAP2B
UDA
REGGIE-2

Inversion

Further Questions:

- How did Inversion Symmetry come about?

Through many inversions in the evolutionary process of chromsomes.

- Is there a biological meaning to its breaking?

It is known that there exists an excess of #G > #C and #T > #A on the coding strand within most genes. Could IS breaking be connected to an asymmetry between numbers of genes on the two strands?

Violations of the 2nd Chargaff rule on HG38. Columns contain the values of #T/#A, #G/#C on different chromosomes, as well as their Y and Z values. The latter reflect the significance of the inequality

	T/A	G/C	Y(T,A)	Y(G,C)	Z(T,A)	Z(G,C)
chr1	1.002593	1.001175	0.001295	0.000587	15	5.76
chr2	1.00274	1.002747	0.001368	0.001372	16.41	13.49
chr3	1.002416	1.002824	0.001207	0.00141	13.19	12.5
chr4	1.001062	1.002595	0.000531	0.001296	5.75	11.04
chr5	1.004679	1.004144	0.002334	0.002068	24.44	17.5
chr6	1.000537	1.001981	0.000268	0.000989	2.72	8.12
chr7	1.003332	1.001884	0.001663	0.000941	16.15	7.57
chr8	0.999241	1.002536	0.00038–	0.001266	3.53–	9.65
chr9	1.001327	1.002823	0.000663	0.001409	5.61	9.99
chr10	1.0039	1.002911	0.001946	0.001454	17.18	10.82
chr11	1.001915	1.002815	0.000956	0.001405	8.48	10.51
chr12	1.003102	1.003317	0.001548	0.001656	13.75	12.2
chr13	1.003831	1.005012	0.001912	0.002499	14.83	15.36
chr14	1.008943	1.007342	0.004451	0.003658	32.58	22.24
chr15	1.001842	1.00411	0.00092	0.002051	6.44	12.23
chr16	1.009601	1.007001	0.004778	0.003488	32.17	21.07
chr17	1.002905	1.006812	0.00145	0.003395	9.77	20.81
chr18	1.005494	1.016917	0.00274	0.008388	19.03	47.34
chr19	1.009276	1.007636	0.004617	0.003803	25.46	20.13
chr20	1.011147	1.012815	0.005542	0.006367	33.22	33.7
chr21	1.003017	1.005026	0.001506	0.002507	7.33	10.15
chr22	0.998893	1.009337	0.00055–	0.004647	2.52–	19.94
chrX	1.003463	1.005699	0.001728	0.002842	16.73	22.23
chrY	1.008873	1.000209	0.004417	0.000105	17.58	0.34

Gene occurrences on the plus (#P) and minus (#M) strands of HG38 display abundance of the former

Three of the results are insignificant (highlighted ***p* > 0.05**, q > 0.044 using FDR corrections). Four chromosomes have opposite preferences, set in italics for *P* < *M* and *T* < *A*. For all significant results we find 16 chromosomes displaying both *P* > M, T > A, and G > C. Chr 22 has both *P* < M and T < A. **Last column indicates significant correlations of T>A and G>C with gene counts (positive by v and negative by x)**

chr	P	M	Y(P,M)	Z(P,M)	p values	Z(T,A)	Z(G,C)	corr
1	4488	4291	0.022	2.103	0.018	15.00	5.76	v
2	4106	3367	0.099	8.549	0	16.41	13.49	v
3	2938	2516	0.077	5.714	5.65E-09	13.19	12.50	v
4	2542	1792	0.173	11.392	0	5.75	11.04	v
5	2777	2186	0.119	8.389	0	24.44	17.50	v
6	4840	3563	0.152	13.931	0	2.72	8.12	v
7	3024	2402	0.115	8.444	0	16.15	7.57	v
8	2135	2032	0.025	1.596	**0.055**	*3.53*–	9.65	
9	3032	2180	0.163	11.802	0	5.61	9.99	v
10	2532	2156	0.080	5.492	2.01E-08	17.18	10.82	v
11	2879	4047	0.169–	*14.035*–	0	8.48	10.51	x
12	3003	2771	0.040	3.053	0.0011	13.75	12.20	x
13	1261	1227	0.014	0.682	**0.25**	14.83	15.36	
14	2092	1906	0.047	2.942	0.0016	32.58	22.24	v
15	4226	3547	0.087	7.702	6.77E-15	6.44	12.23	v
16	2529	1875	0.149	9.855	0	32.17	21.07	v
17	3582	2902	0.105	8.445	0	9.77	20.81	v
18	1182	1490	0.115–	*5.958*–	1.26E-09	19.03	47.34	x
19	3287	3036	0.040	3.157	0.00079	25.46	20.13	v
20	1258	1193	0.027	1.313	**0.09500**	33.22	33.70	
21	670	779	0.075–	*2.863*–	0.00212	7.33	10.15	x
22	1429	1793	0.113–	*6.413*–	7.28E-11	*2.52*–	19.94	?
X	1927	1572	0.101	6.001	9.87E-10	16.73	22.23	v
Y	491	184	0.455	11.816	0.00E + 00	17.58	**0.34**	
				P < M	***p* > 0.05**	*T < A*	***p* > 0.05**	

In summary, both SCR and its generalization into Inversion Symmetry (IS), are valid biological rules.

SCR (and IS) suffers from small violations, which correlate with a small asymmetry of gene occurrences on the two strands.

The IS rules may be viewed as emergent phenomena, caused by the tinkering of evolution with chromosomal sections, rearranging them randomly in either a direct or inverted fashion into novel DNA molecules.