Quantitative Genome Evolution

HC Lee
National Central University
Graduate Institute of Systems Biology
National Central University
Zhongli/Jhongli/Chungli, Taiwan

Enriched analysis of microarrays
Three questions I like to ask

• **WHAT** is the phenomenon?
  – What is important/unusual/interesting?

• **HOW** did it happen?
  – (Physics)

• **WHY** did it happen?
  – (Biology)
What we plan to discuss

How did life evolve so rapidly
- random sequence
- statistical properties of the genome
- uniformity
- homogeneity and universality
- order of genome
- genome the information carrier
- genome at the “edge of chaos”
- symmetries in the genome
- coding versus non-coding regions
- how genome grew
- segmental & whole genome duplication
- genome the blind self-plagiarizer
- genome in a state of fixed-point
- the nearly universal cumulative point mutation density
- age of the genome
- the mystery of the “missing” mutations
- the phyletic gradualism versus punctuated equilibria debate
Motivation: Understand large scale growth and evolution of genome; how did it evolve so rapidly

Materials: Complete genome of prokaryotes and eukaryote

Method: Variety of statistical analysis of frequency of oligonucleotides (k-mers)
Sequence alignment versus statistical analysis

<table>
<thead>
<tr>
<th>Feature</th>
<th>Seq alignment</th>
<th>Stats analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nature of quantity measured</strong></td>
<td>relative</td>
<td>absolute</td>
</tr>
<tr>
<td><strong>Sequence length</strong></td>
<td>&lt; 10k</td>
<td>&gt; 5k</td>
</tr>
<tr>
<td><strong>Sequence comparison</strong></td>
<td>excellent</td>
<td>poor</td>
</tr>
<tr>
<td>relative word in order</td>
<td>poor</td>
<td>moderate</td>
</tr>
<tr>
<td>word content</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Positional information</strong></td>
<td>excellent</td>
<td>poor</td>
</tr>
<tr>
<td>local, relative</td>
<td>poor</td>
<td>good</td>
</tr>
<tr>
<td>large scale structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td>good</td>
<td>poor</td>
</tr>
<tr>
<td>short range</td>
<td>poor</td>
<td>good</td>
</tr>
<tr>
<td>long range</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>(Absolute) Order of sequence</strong></td>
<td>---</td>
<td>excellent</td>
</tr>
<tr>
<td><strong>Evolution time scale</strong></td>
<td>good</td>
<td>---</td>
</tr>
<tr>
<td>&lt; 100 Mya (divergence time)</td>
<td></td>
<td>good</td>
</tr>
<tr>
<td>&gt;&gt; 100 Mya (growth time)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mode of genome growth</strong></td>
<td>---</td>
<td>good</td>
</tr>
</tbody>
</table>
Methods

• Treat genome as a 4-letter pseudo-text
• Count frequency of occurrence of “words”
  – A k-letter word is called a “k-mer”
• Number of possible k-mers is $4^k$
• Use a sliding window of width k and slide 1
• Analyze set of k-mer frequencies

\[
N(GTTACCC) = N(GTTACCC) + 1
\]

After a single sweep have, for each k, a $4^k$-component vector \{f_i | i=4^k\} whose element $f_i$ is the frequency of the $i$th k-mer.
Given freq. set \{f_i\}, define

**k-spectrum**
\{n_f | f=1,2,...\}
\[ \Sigma_i f_i = \Sigma_n f n_f \]

Relative spectral width
\[ \sigma = \text{std dev}/\langle f \rangle \]

**Example:**
6-spectrum of *B. subtilis*

**Width** (2x Std. Deviation)
**Mean frequency**
• All sequences downloaded from GenBank (Nov 2006)
• 422 complete prokaryotic ("bacterial") genomes
  – Lengths 0.5 to 9 M base pairs (Mb)
• 19 complete eukaryotic genomes totaling 328 complete chromosomes
  – Lengths 0.2 to 250 Mb
• Total data set: $2.2 \times 10^{10}$ base pairs
Complete Genomes are diverse

Fractional (A+T) content $p$

Sequence length $L$ (bases)

PF: *Plasmodium falciparum* (A eukaryotic Malaria causing parasite)
Lecture One. Genome: A Large System with Small System Statistics

Lecture Two. Genome: Information Carrier at the “Edge of Chaos”

Lecture Three. Symmetries in Genomes

Lecture Four. Genome the Blind Self-Plagiarizer and the Mystery of the “Missing” Mutations
Lecture One

Genome: A large System with Small-System Statistics

HC Lee
National Central University
Life is highly diverse and complex

Tree of Life

W.F. Doolittle

We are here
And it took a long time to get here

![Divergence of species diagram](image)

- **Bacteria**
- **Eukarya**
- **Archaea**

- Cyanobacteria
- α-proteobacteria
- Mitochondria
- Archezoa

**4 billion yrs ago**

**Now**
• Typical mutation rate is 1/site/Byr
• Life in the form of microbe existed less than 1 Byr after cooling of earth
• Suppose we were “given” a genome 1000 bp long at dawn of life
• After 1 Byr every site mutated once
• Size of sequence space is $4^{1000} \approx 10^{600}$
• Evolution could not have been driven by point mutation
Evolution of Genomes and the Second Law of Thermodynamics

- **Genomes grew & evolved stochastically**
  - modulated by natural selection
  - Bigger genomes carry more information than smaller ones

- **The second law of thermodynamics**:
  - the entropy of closed system can never decrease
  - a system that grows stochastically tends to acquire entropy
  - Increased randomness → more entropy

- **Shannon information**
  - Information decreases with increasing entropy
Long-range variation in GC content is evident not just from extreme outliers, but throughout the genome...the standard deviation barely decreases as window size increases by successive factors of four - 5.9%, 5.2%, 4.9% and 4.6% for windows of size 5, 20, 80 and 320 kb.
CG-content in Human genome has long-range variation

Window size increased by factor of $10^4$, but SD decreased by only 2.6
For GC-content histograms: sample size = window size

Variation of CG-content in Human genome does not obey central limit theorem
Power law is universal for complete eukaryotic genomes: $\gamma = -0.06 \text{ to } -0.42$
Consider $\tau$ equally probable events occurring a total of $L$ times.

Distribution of occurrence frequency characterized by

- mean frequency: $f_{\text{ave}} = L/\tau$
- SD (standard deviation) $\Delta$; or
  
  $\text{CV (coefficient of variation)} = \Delta/f_{\text{ave}}$
- Higher moments of distribution
Random events

• Random events given by Poisson distribution
  – $\Delta^2 = f_{\text{ave}}$, or, $(CV)^2 = 1/f_{\text{ave}}$
  – That is, $(CV)^2 = \tau/L$

• For fixed $\tau$, $(CV)^2 \sim 1/L$
  – Large $L$ limit (thermodynamic limit): $L \sim \infty$, $CV \sim 0$

• For given $\tau$, if $CV$ is known, then
  – $L \sim \tau/(CV)^2$
Frequency set, “k-spectrum” & relative spectral width

Given freq. set \{f_i\}, define

\textbf{k-spectrum} \{n_f | f=1,2,...\}
\Sigma_i f_i = \Sigma_n f n_f

-----

Relative spectral width
\sigma = \text{std dev}/\langle f \rangle

Example: 6-spectrum of \textit{B. subtilis}

Width (2x Std. Deviation)
Mean frequency
Huge difference between genomes and random sequences

Black: genome of *E. coli*
Green: matching random sequence
(Red: model sequence)
2D “portrait” of *E. coli* genome

Frequencies of 10-mers in *E. coli*  

$2^{10} \times 2^{10} = 4^{10} = 1048576$ pixels. Each pixel shows the color-coded frequency of a 10-mer.

Random implies uniformity …
Random implies uniformity; but sometimes other factors intervene.

5-spectra of “genomes” with different base compositions:

- **Green** – random
- **Black** – genome
- **Orange** – model

(A) 50/50
(B) 60/40
(C) 70/30
"k-spectrum is composed of \((k+1)\) sub-distribution of "m-sets"
Partition of $k$-mers into $m$-sets

- Consider genome with fractional AT-content $p$ (then fractional GC-content $q=1-p$)
- There will be more AT-rich words than GC-rich words
- Partition $k$-mers into sets, called $m$-sets, $m=0,1,\ldots,k$; $k$-mers in an $m$-set all have $m$ ATs.
- Total number of:
  - kinds of $k$-mers: $\tau=4^k$
  - $k$-mers: $L$ (sequence length); mean frequency is $L/\tau$
  - kinds of $k$-mers in m-set: $\tau_m$
  - $k$-mers in an m-set: $L_m$; mean frequency is $f_m = L_m/\tau_m$

\[
\tau_m = \binom{k}{m} 2^k, \quad L_{\{\infty}\} = L \binom{k}{m} p^m q^{k-m},
\]
\[
\bar{f} = L/\tau, \quad \tau = 4^k \quad \bar{f}_{m}\{\infty}\} = L_{\{\infty}\}/\tau_m
\]
$f_m$ well approximated by its large-$L$ limit

$$k = 5$$

<table>
<thead>
<tr>
<th>Sequence</th>
<th>$f_m$</th>
<th>$(m = )$ 0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 0.492$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E. coli$</td>
<td>2509</td>
<td>2245</td>
<td>1877</td>
<td>1760</td>
<td>1944</td>
<td>2656</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>2101</td>
<td>2044</td>
<td>1987</td>
<td>1922</td>
<td>1857</td>
<td>1795</td>
<td></td>
</tr>
<tr>
<td>$\lim_{L \to \infty}$ Random*</td>
<td>2114</td>
<td>2048</td>
<td>1983</td>
<td>1920</td>
<td>1860</td>
<td>1801</td>
<td></td>
</tr>
<tr>
<td>$p = 0.691$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$C. acetobutylicum$</td>
<td>154</td>
<td>397</td>
<td>918</td>
<td>1951</td>
<td>4272</td>
<td>10300</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>176</td>
<td>394</td>
<td>882</td>
<td>1970</td>
<td>4400</td>
<td>9832</td>
<td></td>
</tr>
<tr>
<td>$\lim_{L \to \infty}$ Random*</td>
<td>176</td>
<td>393</td>
<td>880</td>
<td>1968</td>
<td>4402</td>
<td>9845</td>
<td></td>
</tr>
</tbody>
</table>

All sequences normalized to a length of 2 Mb; $\bar{f} = 2 \times 10^6/4^5 = 1953$. Random means matching random sequence, or sequence obtained by scrambling the genome. *Values of $\bar{f}_m^{(\infty)}$ given by Eq. (6).

doi:10.1371/journal.pone.0009844.t006
Decomposition of standard deviation into “non-fluctuating” and “fluctuating” parts

\[
\sigma^2 = \tau^{-1} \sum_{u \in S} (f_u - \bar{f})^2 = \tau^{-1} \sum_{m=0}^{k} \sum_{u \in S_m} (f_u - \bar{f}_m + \bar{f}_m - \bar{f})^2
\]

\[
= \tau^{-1} \sum_{m=0}^{k} \left( \tau_m (\bar{f}_m - \bar{f})^2 + 2(\bar{f}_m - \bar{f}) \sum_{u \in S_m} (f_u - \bar{f}_m) + \sum_{u \in S_m} (f_u - \bar{f}_m)^2 \right).
\]

\[
\sigma^2 \equiv \sigma_{nf}^2 + \sigma_{fl}^2,
\]

\[
\sigma_{nf}^2 = \sum_{m=0}^{k} \frac{\tau_m}{\tau} (\bar{f}_m - \bar{f})^2,
\]

(7) (generalization of parallel axis theorem)

(9) (depends only on averages)

\[
\sigma_{fl}^2 = \sum_{m=0}^{k} \sum_{u \in S_m} \frac{(f_u - \bar{f}_m)^2}{\tau} \equiv \sum_{m=0}^{k} \frac{\tau_m}{\tau} \sigma_{m,fl}^2.
\]

(10)
Large-L limit of CV

Large-L limit for non-fluctuating part is known (for all sequences)

\[
(CV^{\{\infty}\})^2 \equiv \lim_{L \to \infty} CV_{nf}^2 = \sum_{m=0}^{k} 2^{-k} \binom{k}{m} (2^k p^m q^{k-m} - 1)^2
\]

\[
= 2^k (p^2 + q^2)^k - 1,
\]

(12)

So is large-L limit for fluctuating part for random sequences

\[
\lim_{L \to \infty} CV_{fl}^2 = \frac{1}{f^2} \lim_{L \to \infty} \sigma_{fl}^2 = \frac{1}{f^2} \sum_{m} \frac{\tau_m}{\tau} \bar{f}_m
\]

\[
= \frac{1}{f^2} \frac{\tau}{L} = \frac{\tau}{L} \quad (\text{random sequence}),
\]

(13)

For genome-size random sequence, \( CV_{nf} >> CV_{fl} \)
Genome and random sequ’s differ only in $CV_{fl}$
Equivalent length of a sequence

Because \( CV_{fl} \sim \tau/L \), we define for any sequence an equivalent length \( l_e \).

\[
 l_e = b_k \tau / \left( CV_{fl} \right)^2 \quad \text{(for } k \geq 2) \tag{14}
\]

\( b_k = 1 - 1/2^{k-1}; =1/2 \text{ for } k=2 \text{ and approaches } 1 \text{ quickly for larger } k' \text{'s} \)

(note: for random sequence \( L = l_e \)). The \( l_e \) of a sequence is the length of the random sequence whose \( CV_{fl} \) is the same as that of the sequence.

Notation: use \( l_e \) for a segment, and \( L_e \) for a whole genome/chromosome
Genomic (E. coli, worm, mustard and human) $l_e$ does not grow with sequence length

$L_e$ of coding and non-coding parts not much different (I)
\( L_e \) of coding and non-coding parts not much different (II)

<table>
<thead>
<tr>
<th>Category</th>
<th>((k =)) 2</th>
<th>5</th>
<th>7</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.359 ± 0.172</td>
<td>4.56 ± 2.01</td>
<td>33.7 ± 15.9</td>
<td>388 ± 223</td>
</tr>
<tr>
<td>( gn ) (41.8%)</td>
<td>0.317 ± 0.141</td>
<td>4.21 ± 1.67</td>
<td>31.2 ± 13.4</td>
<td>337 ± 186</td>
</tr>
<tr>
<td>( ig ) (59.6%)</td>
<td>0.462 ± 0.302</td>
<td>4.99 ± 2.36</td>
<td>31.6 ± 14.5</td>
<td>213 ± 170</td>
</tr>
<tr>
<td>( ex ) (3.3%)</td>
<td>0.292 ± 0.122</td>
<td>4.40 ± 1.62</td>
<td>35.3 ± 13.1</td>
<td>620 ± 298</td>
</tr>
<tr>
<td>( in ) (31.8%)</td>
<td>0.348 ± 0.230</td>
<td>3.65 ± 1.50</td>
<td>23.5 ± 8.7</td>
<td>213 ± 105</td>
</tr>
<tr>
<td>( L_e^{{uc}} ) ((p = 0.5))</td>
<td>0.310 ± 0.150</td>
<td>4.90 ± 2.24</td>
<td>30.1 ± 13.8</td>
<td>487 ± 235</td>
</tr>
<tr>
<td>RSD model</td>
<td>0.597 ± 0.351</td>
<td>4.79 ± 0.70</td>
<td>32.0 ± 5.8</td>
<td>510 ± 149</td>
</tr>
</tbody>
</table>

\( L_e(k), k = 2, 5, 7 \) and 10, averaged over 865 chromosomes. Total sequences length is about \( 2.2 \times 10^{10} \) bases. Abbreviations: All, complete chromosome; \( gn \), genes; \( ig \), intergenic; \( ex \), exons; \( in \), introns. Percentage given indicates portion of complete sequence. \( L_e^{\{uc\}} \) is defined in Eq. (1) and RSD results are averaged over 200 model sequences. See Table S4 for \( L_e(k) \) of other \( k \) values. doi:10.1371/journal.pone.0009844.t002
Difference in $L_e$ of coding and non-coding parts mostly, but not all, caused by difference in $p$.
Genomes have "universal" equivalent lengths

Equivalent ($L_e$) has only weak dependence on base composition ($p$), sequence length ($L$) or species.

Genomes form a $k$-dependent "universal class" in $L_e$

$$L_e^{\{\text{uc}\}}(k; p) = L_e^2 \exp((-k+2)a(p)); \ (2 \leq k \leq 10) \quad (1)$$

$$a(p) = \frac{a_0}{1 + \epsilon \tan \left(\frac{(p^2 + (1-p)^2 - 0.5)\pi}{2}\right)} \quad (2)$$

where $a_0 = 0.92$, $L_e^2 = 310^{+290}_{-150} \text{ b}$, and $\epsilon = 0.50 \pm 0.05$. 
Universality class strengthens with increasing genome length (or statistics)

<table>
<thead>
<tr>
<th>Fraction of k-mers whose P-value is less than P₃, P₆, or P₈</th>
</tr>
</thead>
<tbody>
<tr>
<td>k = 2 (Lₑ = 310 b)</td>
</tr>
<tr>
<td>-------------------------------------</td>
</tr>
<tr>
<td><strong>Length (Mb)</strong></td>
</tr>
<tr>
<td>0.8</td>
</tr>
<tr>
<td>4.6</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>226</td>
</tr>
</tbody>
</table>

P-values for k-mer distribution given by Eq. (1) (at p = 0.5). Null theory assumes genomes are random sequences. The P-values P₃ = 2.7 × 10⁻³, P₆ = 2.0 × 10⁻⁹, and P₈ = 1.3 × 10⁻¹⁵ correspond to z-values of three, six and eight, respectively.

doi:10.1371/journal.pone.0009844.t004
Universality Class of genomic equivalent length

$$\log L_e(k) \sim a \ k + B$$

$a = 0.40$

$B = 1.69 \pm 0.28$

Mild exception: Plasmodium
Is universality class a consequence of similarity in genomes?

Similarity index and similarity matrix

Given a pair of equal-length sequences $\alpha$ and $\beta$, the similarity index $\eta_{sim}(\alpha, \beta)$ for the pair is defined as

$$\eta_{sim}^2(\alpha, \beta) = \frac{1}{k+1} \sum_m \frac{1}{2\tau_m} \sum_{u \in S_m} \frac{(f_u^\alpha - f_u^\beta)^2}{\sigma_m^\alpha \sigma_m^\beta}$$  \hspace{1cm} (15)

where $S_m$ is an $m$-set and $\sigma_m^2$ is the variance of the frequency of the $k$-mers in $S_m$. The pair are similar (in $k$-mer-content) when $\eta_{sim} \ll 1$, are (considered to be) identical when $\eta_{sim} = 0$, and are highly dissimilar when $\eta_{sim} \gg 1$. 

Homogeneity of Chromosome on a scale of 10 kb

Figure 5. Intra-chromosomes similarity plots. Plots are for $k = 2$ (Methods). Sliding window has width 25 kb and slide 10 kb; pixel size is 10 kb by 10 kb. In each plot, the coordinates for the upper-left triangle are sites along the chromosome (chr), and those for the lower-right triangle are along a concatenate composed of gene (gn, left side) and intergene (ig, right side) parts. In effect, the upper-left triangle shows chr-chr similarity, and the lower-right triangle shows gn-gn (lower-left sub-triangle), ig-ig (upper-right sub-triangle), and gn-ig (rectangular) similarities in three separate regions. The lengths of the gn and ig parts are given in Table 3.

doi:10.1371/journal.pone.0009844.g005
Universality class not due to similarity in genomes

Figure 6. Intra-*E. coli* and inter-chromosome similarity plots. The plots are those of *E. coli* chromosome vs. the chromosomes of, left to right and top to bottom, *E. coli*, *E. coli UT189*, *Salmonella*, the delta-proteobacteria *S. aciditrophicus*, the cyanobacteria *Synechocystis*, the archaea *P. aerophilum*, chromosome 5 of the fungus *A. fumigatus*, and the first 4.5 Mb segment from chromosome 1 of *H. sapiens*. Coordinates are sites along the sequence. Sliding window width is 100 kb and slide is 25 kb, pixel size is 25 kb by 25 kb.
doi:10.1371/journal.pone.0009844.g006
Universality class – for fixed word length $k$, $L_e$ is (approximately) the same for all genomes
- $\log L_e(k) = ak + B$; $a, B$ universal constants (for ~900 complete sequence)

For $k<8$, $L_e$ is much shorter than true genome length; $L_e(k=2) \sim 300$ b

$L_e$ for coding and non-coding parts about the same

Universality not due sequence similarity
Small $L_e$ is a signature of segmental duplication

- Recall: $L_e$ is the length of random sequence having genomic CV

- Take a random sequence of length $L_e$ and replicate it $n$ times, then sequence length is $nL_e$ but equivalent length is still $L_e$ (for all $k << L_e$)

- Hint: small genomic $L_e$ caused by segmental duplication
Fun with $l_e$ and concatenates

\[
l_e = \frac{\tau}{(CV)^2} = \frac{\tau}{(\sigma/f)^2} = L^2 \frac{\tau^3}{\sigma^2} \quad (\tau = 4k)
\]

Let $X$ be genomic, and $R$ random, sequence of same length;
RX be concatenate of $R$ and $X$.
Then $\sigma(X) >> \sigma(R)$, $l_e(X) << l_e(R) = L(R) = L(X)$.

\[
l_e(RX) = L(RX)^2 \frac{\tau^3}{\sigma(RX)^2} \sim (L(R) + L(X))^2 \frac{\tau^3}{\sigma(X)^2}
= 4L(X)^2 \frac{\tau^3}{\sigma(X)^2} = 4 \quad l_e(X)
\]

Similarly,

\[
l_e(RXX) \sim (3/2)^2 l_e(X); \quad l_e(RRX) \sim 9 \quad l_e(X)
\]
Similar (notation \(\sim\)): similar in word content. Suppose \(X\) and \(X'\) are same length sequences from the same genome, \(Y\) from a different genome, and \(R\) is a random sequence. \(XY\) denote concatenate of \(X\) and \(Y\).

- \(l_e(X) \sim l_e(X') \sim l_e(XX')\)
- If \(X \sim Y\), then \(l_e(X) \sim l_e(Y)\), \(l_e(XY) \sim l_e(X)\),
- If \(X\) not\(\sim\) \(Y\), then \(l_e(XY) > \min(l_e(x), l_e(Y))\)

- \(l_e(RX) \approx 4 l_e(X)\)
- \(l_e(RXY) \approx 2.3 l_e(XY)\)
- \(l_e(RR'X) \approx 9 l_e(X)\)

<table>
<thead>
<tr>
<th></th>
<th>(k = 2)</th>
<th></th>
<th>(k = 6)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(l = 50)</td>
<td>(l = 200)</td>
<td>(l = 50)</td>
<td>(l = 200)</td>
</tr>
<tr>
<td>R</td>
<td>47.5 ± 28.2</td>
<td>154 ± 126</td>
<td>48.6 ± 1.5</td>
<td>192 ± 5</td>
</tr>
<tr>
<td>RR'</td>
<td>37.0 ± 16.2</td>
<td>124 ± 46</td>
<td>48.2 ± 1.2</td>
<td>197 ± 5</td>
</tr>
<tr>
<td>A</td>
<td>348 ± 0.37</td>
<td>360 ± 0.33</td>
<td>9.55 ± 0.69</td>
<td>11.7 ± 7</td>
</tr>
<tr>
<td>AA'</td>
<td>357 ± 0.46</td>
<td>352 ± 0.23</td>
<td>9.88 ± 1.07</td>
<td>11.1 ± 7</td>
</tr>
<tr>
<td>AC</td>
<td>351 ± 0.61</td>
<td>361 ± 0.21</td>
<td>9.37 ± 1.01</td>
<td>11.5 ± 6</td>
</tr>
<tr>
<td>AD</td>
<td>355 ± 0.43</td>
<td>384 ± 0.84</td>
<td>9.18 ± 0.83</td>
<td>11.6 ± 9</td>
</tr>
<tr>
<td>AC</td>
<td>359 ± 0.51</td>
<td>371 ± 0.34</td>
<td>11.0 ± 0.9</td>
<td>14.2 ± 1.5</td>
</tr>
<tr>
<td>AD</td>
<td>411 ± 0.44</td>
<td>423 ± 0.24</td>
<td>11.8 ± 0.9</td>
<td>14.3 ± 6</td>
</tr>
<tr>
<td>AD</td>
<td>.942 ± 0.275</td>
<td>1.05 ± 0.09</td>
<td>14.9 ± 1.4</td>
<td>20.4 ± 1.1</td>
</tr>
<tr>
<td>AD</td>
<td>.598 ± 0.104</td>
<td>.613 ± 0.052</td>
<td>17.9 ± 1.6</td>
<td>24.0 ± 1.6</td>
</tr>
<tr>
<td>AD</td>
<td>.324 ± 0.052</td>
<td>.383 ± 0.055</td>
<td>11.2 ± 1.9</td>
<td>16.9 ± 1.9</td>
</tr>
<tr>
<td>B</td>
<td>.124 ± 0.029</td>
<td>.166 ± 0.099</td>
<td>5.17 ± 0.68</td>
<td>6.54 ± 2.00</td>
</tr>
<tr>
<td>BB'</td>
<td>.232 ± 0.155</td>
<td>.258 ± 0.183</td>
<td>6.16 ± 1.94</td>
<td>7.54 ± 2.30</td>
</tr>
<tr>
<td>AB</td>
<td>.463 ± 0.241</td>
<td>.502 ± 0.263</td>
<td>11.2 ± 1.9</td>
<td>15.2 ± 3.5</td>
</tr>
<tr>
<td>RA</td>
<td>1.19 ± 0.09</td>
<td>1.34 ± 0.20</td>
<td>22.6 ± 1.2</td>
<td>38.5 ± 3.0</td>
</tr>
<tr>
<td>RB</td>
<td>.575 ± 0.321</td>
<td>.754 ± 0.637</td>
<td>15.6 ± 4.2</td>
<td>23.3 ± 8.5</td>
</tr>
<tr>
<td>RAB</td>
<td>873 ± 424</td>
<td>1.10 ± 0.49</td>
<td>18.4 ± 3.2</td>
<td>31.3 ± 6.0</td>
</tr>
<tr>
<td>RR'A</td>
<td>2.63 ± 0.66</td>
<td>3.16 ± 0.30</td>
<td>31.5 ± 2.1</td>
<td>72.2 ± 6.8</td>
</tr>
<tr>
<td>RR'B</td>
<td>1.03 ± 0.62</td>
<td>1.37 ± 0.70</td>
<td>22.9 ± 4.5</td>
<td>44.7 ± 14.3</td>
</tr>
</tbody>
</table>
Equivalent length ($l_e$) of a sequence after an $n$-fold increase in length via three basic modes of growth. Initial sequence length is $l_0 >> 1$, final sequence length is $L = nl_0$.

<table>
<thead>
<tr>
<th>Sequence type</th>
<th>Initial $l_e$</th>
<th>Mode of growth</th>
<th>Final $l_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>$l_0$</td>
<td>Random base-by-base growth</td>
<td>$L (=nl_0)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Whole-sequence replication ($n-1$ times)</td>
<td>$\approx l_0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Segmental duplication</td>
<td>$l_0 &lt; l_e &lt; &lt; nl_0$</td>
</tr>
<tr>
<td>Non-random</td>
<td>$l_{e0}$ ($&lt;&lt; l_0$)</td>
<td>Random base-by-base growth</td>
<td>$\approx \min(n^2l_{e0}, L)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Whole-sequence replication ($n-1$ times)</td>
<td>$\approx l_{e0}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Segmental duplication</td>
<td>$l_{e0} &lt; l_e &lt; &lt; \min(n^2l_{e0}, L)$</td>
</tr>
</tbody>
</table>
RSD model with three universal parameters generates artificial genomes with universal $L_e$

Note: Deep consequences in understanding speed of evolution (Lecture 4)

End of Lecture One