Inevitability and containment of replication errors for eukaryotic genome lengths spanning Megabase to Gigabase

Al Mamun, Mohammed; Albergante, Luca; Moreno, Alberto; Carrington, Jamie T.; Blow, John; Newman, Timothy J.

Published in:
Proceedings of the National Academy of Sciences

DOI:
10.1073/pnas.1603241113

Publication date:
2016

Citation for published version (APA):
SI Text

Mathematical Derivations

Probability of a specific number of DFSs

The previous approach can be extended to calculate the probability of an arbitrary number of double stalls. The probability of exactly one double fork stall, which will be called Prob(1DS), can be calculated directly as

\[ \text{Prob}(1\text{DS}) = \sum_k p_k \cdot \prod_{k_1 \neq k} (1 - p_{k_1}) \]  

(M12)

Combining Eq. (M12) with Eq. (M7) we obtain

\[ \text{Prob}(1\text{DS}) = \text{Prob}(\text{NDS}) \cdot \sum_k \frac{p_k}{(1 - p_k)} \]

Therefore

\[ \frac{\text{Prob}(1\text{DS})}{\text{Prob}(\text{NDS})} = \sum_k \frac{p_k}{(1 - p_k)} \]

To simplify the next steps, we introduce the following definitions

\[ S_1 = \sum_k \frac{p_k}{1 - p_k} \]

\[ S_2 = \sum_k \left( \frac{p_k}{1 - p_k} \right)^2 \]

\[ \vdots \]

\[ S_m = \sum_k \left( \frac{p_k}{1 - p_k} \right)^m \]

Additionally, let Prob(mDS) be the probability of \( m \) double fork stalls, the following conventions will be used

\[ R_1 = \frac{\text{Prob}(1\text{DS})}{\text{Prob}(\text{NDS})} \]

\[ R_2 = \frac{\text{Prob}(2\text{DS})}{\text{Prob}(\text{NDS})} \]
\[ R_m = \frac{\text{Prob}(m \text{DS})}{\text{Prob}(N \text{DS})} \quad (M13) \]

Hence,

\[ R_1 = \sum_k \frac{P_k}{1 - P_k} = S_1 \quad (M14) \]

And

\[ R_2 = \frac{1}{2!} \sum_{k_1} \sum_{k_2 \neq k_1} \frac{P_{k_1}}{1 - P_{k_1}} \frac{P_{k_2}}{1 - P_{k_2}} \]

which can be rewritten as

\[ R_2 = \frac{1}{2!} \left[ \left( \sum_k \frac{P_k}{1 - P_k} \right)^2 - \sum_k \left( \frac{P_k}{1 - P_k} \right)^2 \right] = \frac{1}{2!} (S_1^2 - S_2) \]

Similarly

\[ R_3 = \frac{1}{3!} \sum_{k_1} \sum_{k_2 \neq k_1} \sum_{k_3 \neq k_1, k_2} \frac{P_{k_1}}{1 - P_{k_1}} \frac{P_{k_2}}{1 - P_{k_2}} \frac{P_{k_3}}{1 - P_{k_3}} \]

\[ = \frac{1}{3!} (S_1^3 - 3 S_1 S_2 + 2 S_3) \]

Iterating the same approach it is possible to show that

\[ R_4 = \frac{1}{4!} (S_1^4 - 6 S_1^2 S_2 + 8 S_1 S_3 + 3 S_2^2 - 6 S_4) \]

\[ R_5 = \frac{1}{5!} (S_1^5 - 10 S_1^3 S_2 + 15 S_1^2 S_3 S_2 + 20 S_1 S_2^2 - 20 S_1 S_3 - 30 S_1 S_4 + 24 S_5) \]

\[ R_6 = \frac{1}{6!} (S_1^6 - 15 S_1^4 S_2 + 45 S_1^2 S_2^2 - 15 S_2^2 S_3 - 15 S_3^2 + 40 S_1^3 S_3 S_2 + 120 S_1 S_2 S_3 S_4 - 90 S_2^2 S_4 + 90 S_2 S_4 + 144 S_1 S_5 - 120 S_6) \]

Finally, combining \( R_1, R_2, R_3, R_4, R_5 \) and \( R_6 \) with Eq. (M13), we can obtain the probability of one to six double fork stalls as follows

\[ \text{Prob}(1 \text{DS}) = \text{Prob}(N \text{DS}) \cdot R_1 \]
\[ \text{Prob}(2 \text{DS}) = \text{Prob}(N \text{DS}) \cdot R_2 \]
\[ \text{Prob}(3 \text{DS}) = \text{Prob}(N \text{DS}) \cdot R_3 \]
\[ \text{Prob}(4 \text{DS}) = \text{Prob}(N \text{DS}) \cdot R_4 \]
\[ \text{Prob}(5 \text{DS}) = \text{Prob}(N \text{DS}) \cdot R_5 \]
\[ \text{Prob}(6 \text{DS}) = \text{Prob}(N \text{DS}) \cdot R_6 \]
As shown in the SI Table 2, direct calculations on IMR90 cell line suggest that only the leading power is playing a significant role for the $R_i$. Therefore, we write:

\[
R_2 \approx \frac{1}{2!} (S_1^2)
\]
\[
\vdots
\]
\[
R_k \approx \frac{1}{k!} (S_1^k)
\]

and hence

\[
\text{Prob}(kDS) = \text{Prob}(NDS) \cdot R_k \approx \text{Prob}(NDS) \frac{1}{k!} (S_1^k)
\]

The probability density function of a Poisson distribution is

\[
\text{Prob}(n) = \exp(-\lambda) \frac{\lambda^n}{n!}
\]

Therefore, for our distribution to follow a Poisson we have to show that

\[
\text{Prob}(NDS) = \exp(-\lambda)
\]

This implies, $S_1 = \lambda$.

Now, from (M7) and (M15) we have

\[
\lambda = -\log\left(\prod_k (1 - P_k)\right) = -\sum_k \log(1 - P_k) = \sum_k \log\left(\frac{1}{1 - P_k}\right) = \sum_k \log\left(1 + \frac{P_k}{1 - P_k}\right)
\]

The value $P_k/(1 - P_k)$ is very small and we can use a Taylor expansion to obtain

\[
\lambda = \sum_k \left(\frac{P_k}{1 - P_k} + O(P_k^2)\right) \approx \sum_k \left(\frac{P_k}{1 - P_k}\right) = S_1
\]

Since the $P_k$ are very small, this approximation is generally very good.
Frequency of replicons of particular size

We have shown in Equation (M2) that the probability of a DFS in the region of DNA between a pair of adjacent ROs separated by \( N \) nucleotides is,

\[
\text{Prob}(N) = 1 - (1 + Nq)(1 - q)^N \quad (M16)
\]

Now, we calculate the probability of DFS in a cohort of \( M \) replicons whose size is “in the vicinity of” \( N \). The probability of no error occurring from this cohort would be the following product,

\[
\prod_k (1 - P(N_k)).
\]

Where the product is restricted to those replicons within the cohort. This probability will be very close to one, and we denote it by \( \theta \). Substituing (M16) into this expression, and recognising that all of the \( N_k \) are close to \( N \), enables us to rewrite the probability of no error from the cohort as:

\[
(1 + Nq)^M \cdot (1 - q)^{MN} = \theta
\]

Now, taking the natural logarithm,

\[
M \cdot \log(1 + Nq) + MN \cdot \log(1 - q) = \log(\theta) \quad (M17)
\]

Since \( q \ll 1 \), \( \log(1 - q) \approx -q \), and thus we write

\[
M \cdot \log(1 + Nq) - MNq = \log(\theta),
\]

Hence,

\[
M = \frac{\log(\theta)}{[\log(1 + Nq) - Nq]} \quad (M18)
\]

For \( Nq \ll 1 \) it is straightforward to show from expanding the denominator that \( M \sim 1/N^2 \).
**SI Figure Legends**

**SI Figure 1:** Inter-RO distances from Besnard et al. (B) and Picard et al. (P) datasets are plotted. Due to the difference in resolution of detection, the minimum inter-RO distance in Picard et al. data is 4001 bp while in Besnard et al. data it is 240 bp. The overlapping bar charts show the two datasets are compatible. More detail of the compatibility of the two datasets is discussed in reference 18 of the main article.

**SI Figure 2:** Data in the left and right column is from HeLa human dataset B and P respectively. a & b) Frequency of replicons in each cohort; defined according to the following size ranges, \(<10^3\) bp = XS, \(10^3–10^4\) bp = S, \(10^4–10^5\) bp = M, \(10^5–10^6\) bp = L, \(>10^6\) bp = XL. c & d) Probability of DFS in each cohort of the replicons. e & f) Higher resolution plot of probability of DFS at the transition from “medium (M)” to “large (L)” gap cohorts, contributing most towards the P(DFS); red bars show the bins with maximum P(DFS) in respective datasets. g & h) Theoretical frequency distribution of replicons inferred from the plots e & f are presented in blue; grey shows the actual frequency distribution in those bins in the data and red highlights the red bins in e & f.

**SI Figure 3:** Data in the left and right column is from hESC and K562 in human dataset B and P respectively. a & b) Frequency of replicons in each cohort; defined according to the following size ranges, \(<10^3\) bp = XS, \(10^3–10^4\) bp = S, \(10^4–10^5\) bp = M, \(10^5–10^6\) bp = L, \(>10^6\) bp = XL. c & d) Probability of DFS in each cohort of the replicons. e & f) Higher resolution plot of probability of DFS at the transition from “medium (M)” to “large (L)” gap cohorts, contributing most towards the P(DFS); red bars show the bins with maximum P(DFS) in respective datasets. g & h) Theoretical frequency distribution of replicons inferred from the plots e & f are presented in blue; grey shows the actual frequency distribution in those bins in the data and red highlights the red bins in e & f.

**SI Figure 4:** Data is from iPSC human dataset in B. a) Frequency of replicons in each cohort; defined according to the following size ranges, \(<10^3\) bp = XS, \(10^3–10^4\) bp = S, \(10^4–10^5\) bp = M, \(10^5–10^6\) bp = L, \(>10^6\) bp = XL. b) Probability of DFS in each cohort of the replicons. c) Higher resolution plot of probability of DFS at the transition from “medium (M)” to “large (L)” gap cohorts, contributing most towards the P(DFS); red bars show the bins with maximum P(DFS) in respective datasets. d) Theoretical frequency distribution of replicons inferred from the plots e & f are presented in blue; grey shows the actual frequency distribution in those bins in the data and red highlights the red bins in c.