The complexity of mutagenesis: beyond the molecular clock

Kelley Harris
University of Washington

Workshop on Population and Speciation Genomics
January 24, 2020
Forces that shape genomic diversity

- Mutation
- Genetic Drift
- Selection

Image co-artist: Natalie Telis
Mutations as a molecular clock
When the clock breaks down (runs out of batteries?)

• Almost every population genetic method assumes that mutations accumulate at a constant rate per year within populations

• This assumption works fine until it doesn’t

• The mutation process has complex features that can trip you up if you aren’t looking out for them

  • and are also interesting phenotypes to study in their own right

• Estimates of the mutation rate per year and generation time are needed to calibrate output of PSMC and other demographic inference methods
Molecular clock 101

- Mutagenesis is more clock-like over short timescales compared to long time scales.

- A simple branch length test can reveal whether mutagenesis is clock-ish in your data:

  ![Diagram](image)

  $$D(A,O) = D(B,O)$$

Data can fail this test due to mutation rate variation, selection, or introgression.
Violation of molecular clock over very long timescales

Drake 1991

The error threshold

- A simple model by Eigen & Schuster (1979) justifies Drake’s rule

- Consider a “master” virus with fitness $1+s$ and genome length $L$

- All mutant viruses have fitness 1

- The master sequence will die out due to Muller’s rachet/“error catastrophe” if and only if the mutation rate $\mu$ is below a threshold:

  - $\mu < \log(s)/L$
Stable quasispecies vs error catastrophe

Lauring and Andino 2010

$\mu < \text{error threshold}$

$\mu > \text{error threshold}$

Lauring and Andino 2010
How can we gather mutation rate data to test these theories?
Measuring mutation rates with mutation accumulation (MA) lines

Keightly and Charlesworth 2005
MA with a reporter gene

Inactive/ broken promoter

Point mutations can restore promoter function

Reporter gene (e.g. encoding GFP or luciferase)

DNA

mRNA

A reporter protein Amount is easily measured (e.g. GFP by fluorescence)
Mutation rate estimates vary enormously in quality

• PSMC results, divergence time estimates, etc. depend heavily upon a mutation rate estimate. Where does that number come from?

• Calculation from phylogenetic divergence data (substitutions / estimated divergence time)

• MA experiment + whole genome sequencing ($$-$$$$$$)

• MA experiment + reporter gene sequencing (cheap today, only reasonable direct estimate 10 years ago)

• Whole-genome trio sequencing ($$$$$$$$$$$$$)
Drake’s rule driven mostly by viruses and bacteria


Mesoplasma florum
Paramecium tetraurelia
Chlamydomonas reinhardtii
Saccharomyces cerevisiae
Neurospora crassa
Trypanosoma brucei
Saccharomyces pombe
Plasmodium falciparum
The effective population size is an important parameter in understanding the genetic drift and mutation rates in different species. The figure shows a relationship between the base-substitutional mutation rate/site/cell division and the effective population size. In this study, the mutation rates on an absolute time scale in two other prokaryotes were 0.0082 (0.0013) and 0.0077 (0.0006) for base substitutions alone (with only minor additional contributions from insertion/deletions). For eukaryotes, the data presented here are inconsistent with the 0.003 expected under Drake's Rule. There is no compelling evidence that mutation accumulates in any species, and would lead to a predicted equilibrium genomic A/T content of 1/3. However, the drift-barrier hypothesis predicts that selective opposition to the accumulation of A/T bases in nature would lead to a predicted equilibrium genomic A/T content of 1/3. The mechanisms responsible for the discontinuity in scaling of mutation rate with genome size remain unclear. There is no evidence that the mutation rate estimates for Archaebacteria fall below the general trend. The data for multicellular eukaryotes are summarized in Table S8. The prokaryote reported in this study, with genome sizes of ∼10 Mb, is the genome size in megabases. The mechanisms responsible for the discontinuity in scaling of mutation rate with genome size remain unclear. There is no evidence that the mutation rate estimates for Archaebacteria fall below the general trend. The data for multicellular eukaryotes are summarized in Table S8. The prokaryote reported in this study, with genome sizes of ∼10 Mb, is the genome size in megabases.
Why should effective population size affect mutation rate?

Why is the mutation rate what it is?
1. The Cost-of-Fidelity Model

2. The Drift-Barrier Hypothesis

- Biophysics Barrier Mutation Rate
- Drift Barrier Mutation Rate

Mutation Rate vs. Cost of Fidelity

Excess Mutation Load \( \sim 1/Ne \)

Lynch *Trends in Genetics* 2010  
Sung, *et al.* *PNAS* 2012
Mutators can be favored in asexual organisms

- Expected extra load of deleterious mutations must not exceed the expected benefit of beneficial mutations
- Robustness to environmental change
- Stress-induced mutagenesis?
Elevated Mutagenesis Does Not Explain the Increased Frequency of Antibiotic Resistant Mutants in Starved Aging Colonies

Sophia Katz, Ruth Hershberg

Published: November 14, 2013 • https://doi.org/10.1371/journal.pgen.1003968
Selection against mutator alleles is weak in sexual organisms.
Other factors affecting the mutation rate

Environmental Mutagens

Life history
Male mutation bias

Hurst and Ellegren 1998
Paternal age effect (the classical model)

Amster and Sella 2016
Branch length ~ number of substitutions
Label = Estimate of 
(male mutation rate) 
/(female mutation rate)
Two additional \textit{de novo} mutations per year of paternal age

A small but significant maternal age effect (0.5 muts/year)

If spermatocyte replication causes the paternal age effect, the fraction of paternal mutations should increase with parental age.
Human trio data now contradict this prediction

Overlooked roles of DNA damage and maternal age in generating human germline mutations

Ziyue Gao\textsuperscript{a,b,1}, Priya Moorjani\textsuperscript{c,d}, Thomas A. Sasani\textsuperscript{e}, Brent S. Pedersen\textsuperscript{e}, Aaron R. Quinlan\textsuperscript{e,f}, Lynn B. Jorde\textsuperscript{e}, Guy Amster\textsuperscript{g,2}, and Molly Przeworski\textsuperscript{g,h,1,2}
Maternal age causes C>G mutation accumulation in localized regions of chromosomes 5, 7, and 16

Jonsson, et al. 2017
Maternal age causes C>G mutation accumulation in localized regions of chromosomes 5, 7, and 16


1000 Genomes Variant Density

Position on Chromosome 16

Large CEPH families reveal variability in paternal age effect between families

Sasani, et al. eLife 2018
Other causes of mutation rate variation along the genome

- Replication timing
- Transcription-associated-mutagenesis (TAM) and transcription-coupled-repair (TCR)
- Non-B-DNA structures and other DNA repeats
- Chromatin state
Replication timing


Francioli, et al. 2015
Replication and transcription induce strand asymmetry

Excess of G+T over A+C on coding strand of most genes

Green, et al. 2003
Measuring the human mutation rate

Human

Chimpanzee

Nachman and Crowell 2001

Parent-child trios

1.0e-8 mutations per site per gen

1000 Genomes Consortium 2010

2.5e-8 mutations per site per gen
What is the real human mutation rate?

Has the mutation rate slowed down during recent human history?
The Hominoid Mutation Rate Slowdown

Too large to explain by paternal age effect alone (Scally and Durbin 2012)

Have genetics and environment played a role?

Adapted from http://www.bio.indiana.edu/graduate/multidisciplinary/GCMS/trainees/thomas_gregg.php

“The” mutation rate encompasses a menagerie of mutation types:

**Point Mutations**

- Transitions

**Multinucleotide Mutations**

- Other Transversions

**Small indels**

- CC → TT

**Large Copy Number Changes**

- GC-conservative Transversions

Diagram of nucleotide bases (A, T, C, G) illustrating the different types of mutations.
CpG Mutations

- Many species (incl humans, not incl *Drosophila*) methylate C when it’s next to G (C-phosphate-G)

- CpG methylation regulates gene expression
CpG sites are hypermutable

- On average, CpG sites have a 30-fold higher mutation rate than other C’s in the human genome

- 70-80% of CpGs are methylated in mammals; most unmethylated CpGs are part of CpG islands

- Fewer than 1% of dinucleotides in the human genome are CpGs, although the expected frequency is $0.21 \times 0.21 = 4.41\%$
CpG transitions are somewhat more clocklike than other mutations

In a tree of 19 mammals, CpG mutations yield a more clocklike tree than mutations occurring in other contexts

Hwang and Green 2004

CpG mutations also appear more clocklike than other mutations in great ape tree

Limits to clock-like behavior of CpGs

Bulk sequencing of tumor samples

CpG mutation count

Age of patient at diagnosis

Fast renewing tissue

Slowly renewing tissue

Single-cell sequencing of normal neurons

Number of somatic mutations

A (15yr, F)

B (17yr, M)

C (42yr, F)

Others

C>T

Highly efficient repair

Embryonic development

Time between two consecutive cell divisions
“Mutational signatures” of types of DNA damage in cancer

<table>
<thead>
<tr>
<th>C</th>
<th>A</th>
<th>Tobacco exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>AA</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TCT</th>
<th>TAT</th>
<th>Error-prone Polymerase ε activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>ACA</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TCT</th>
<th>TTT</th>
<th>Off-target DNA editing by APOBEC enzymes</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCA</td>
<td>TTA</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TCT</th>
<th>TGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCA</td>
<td>TGA</td>
</tr>
</tbody>
</table>
APOBEC / AID deaminases

• APOBEC attacks RNA viruses, mutating TCA and TCT by deamination

• Its homologue AID hypermutates T cell receptors for proper immune function

• Both cause off-target germline mutations, especially in endogenous retroviral sequences

• APOBEC is erroneously switched on in many cancers (esp cervical), associated with poorer outcomes
Mutations that involve the pyrimidine-to-purine classes of nucleotides; that is, purine-to-pyrimidine transitions (typically >1 large-scale genomic changes). We emphasize how different mutational processes could be used as prognostic indicators, as predictors of therapeutic sensitivity or as targets of disease control.

Ongoing mutational processes reflect active biological processes in the cancer that represent a DNA repair pathway that is awry. Bursts of APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide)-induced deamination, and Signature D is ongoing through life. Signature B can be matched up with the signatures of tobacco smoking, Signature C can represent historical mutational processes are no longer active. Signature A represents deamination of methylated cytosines, which have been active in the entire lifetime. Ongoing mutational processes reflect active biological processes in the cancer that have produced by a mutational process. The final mutational portrait is the sum of all of the different mutational processes (A–D) that have been active in the cancer genome.

Figure 1 | Active mutational processes over the course of cancer development.

Metastasis


Mutational processes over time:

Mutational Process 1:
- C>A (blue)
- C>G (black)
- C>T (green)
- T>A (red)
- T>C (gray)
- T>G (pink)

Signature of Mutational Process 1: Weaker constant endogenous mutational process.

Mutational Process 2:
- C>A (blue)
- C>G (black)
- C>T (green)
- T>A (red)
- T>C (gray)
- T>G (pink)

Signature of Mutational Process 2: Stronger intermittent mutational process.

Mutational Process 3:
- C>A (blue)
- C>G (black)
- C>T (green)
- T>A (red)
- T>C (gray)
- T>G (pink)

Signature of Mutational Process 3: Moderate intermittent mutational process.

Mutational Process 4:
- C>A (blue)
- C>G (black)
- C>T (green)
- T>A (red)
- T>C (gray)
- T>G (pink)

Signature of Mutational Process 4: Moderate mutational process activates at different times.

Mutational spectrum of the final cancer genome:

C>A: Blue
C>G: Black
C>T: Green
T>A: Red
T>C: Gray
T>G: Pink

Current Opinion in Genetics & Development

Alexandrov and Stratton 2014
Mutation signature analysis

Mutation counts in 96 triplet contexts across cancers

Nonnegative matrix factorization

Breast
- Signature 1B
- Signature 2
- Signature 3
- Signature 6
- Signature 8
- Signature 13
- Other

Cervix
- Signature 1B
- Signature 2

Colorectum
- Signature 1B
- Signature 6
- Signature 10
- Signature 12
- Other

CLL
- Signature 1B
- Signature 2
- Signature 9

Lung adenocarcinoma
- Signature 1B
- Signature 2
- Signature 4
- Signature 5
- Other

Melanoma
- Signature 1B
- Signature 7
- Signature 11
- Other

Sample
Effect of BRCA germline mutations on breast cancer mutation distribution
Mutational signatures in the germline?
Hypothesis: different germline signatures have different evolutionary histories.
Private European SNPs are enriched for a mutational signature of unknown origin.
A signature of elevated mutagenesis in the European germline

Harris *PNAS* 2015
Harris and Pritchard *eLife* 2017
Genes mirror geography within Europe

John Novembre¹,², Toby Johnson⁴,⁵,⁶, Katarzyna Bryc⁷, Zoltán Kutalik⁴,⁶, Adam R. Boyko⁷, Adam Auton⁷, Amit Indap⁷, Karen S. King⁸, Sven Bergmann⁴,⁶, Matthew R. Nelson⁸, Matthew Stephens²,³ & Carlos D. Bustamante⁷
Hypothetical Signature of a TCC-to-TTC mutation rate increase

- **TCC-to-TTC mutation rate**
  - Present: High
  - Past: Low

- **TCC-to-TTC mutation fraction**
  - Time: Present to Past
  - Allele frequency: Low to High
Pulse replicates in the UK10K data
A pulse of TCC-to-TTC mutations in Europe and South Asia?
Minor components of the pulse
Expected TCC fraction as a function of allele frequency

- Partition time into discrete intervals

- $A(k,i) = \text{the total branch length subtending } k \text{ lineages between times } T_i \text{ and } T_{i-1}$

- $r_i = \text{the rate of TCC mutations between } T_i \text{ and } T_{i-1}$

Expected TCC fraction as a function of allele frequency is

$$E[f(k)] \sim \frac{\sum_i A(k,i) \cdot r_i}{\sum_i A(k,i)}$$
Inference of a mutation pulse lasting from 15,000 to 2,000 years ago
Similar simultaneous mutation pulses in Europeans, South Asians, and...a dog STD??

Canine transmissible venereal tumors (CTVTs) all descend from an ancestral tumor in a dog who lived 4000 to 8500 years ago.

CTVTs experienced a high load of GTCCA>GTTCA mutations that ceased ~1,000 years ago.

Same timeframe as the European mutation pulse and similar (though not identical) sequence bias.

“A recent study (37) detected evidence for an excess of C>T mutations at TCC contexts, the mutation type most prevalent in signature A, accumulating in the human germ line between 15,000 and 2000 years ago. If this human mutation pulse is due to signature A, it could indicate a shared environmental exposure that was once widespread but has now disappeared.”
Future direction: are mutation pulses the relics of lost mutator alleles?

Fixed differences between great ape spectra

European TCC mutation pulse

Japanese mutation pulse
How mutator alleles could promote rapid mutation spectrum turnover

Sawyer and Malik *PNAS* 2006
Positive selection in DNA repair genes and other housekeeping genes

- BRCA1 & BRCA2 are under positive selection in primates
- 5 Nonhomologous end joining genes experienced positive selection during primate evolution, incl XRCC4 which has been under selection in Europeans
- Iron-uptake receptor TfR1 evolves under positive selection to avoid facilitating viral entry

Demogines, et al. 2010
Demogines, et al. 2013
MuSHI: Mutation Spectrum History Inference

$k$-SFS:

\[
\begin{bmatrix}
    1 & 2 & 1 \\
    1 & 0 & 0 \\
    2 & 0 & 0
\end{bmatrix}
\]

sample frequency

Dewitt, Harris, and Harris, in prep
MuSHI estimates demographic history jointly with the mutation spectrum history (mush)

Simulated data

1000 Genomes Continental Groups

1000 Genomes Europeans

Dewitt, Harris, and Harris, in prep
A simulated example of pulse recovery

96 MUTATION TYPES WITH LATENT PULSE SIGNATURE AFFECTING 5

$\ell_2$-smooth pulse:

$\ell_1$-smooth pulse:

Dewitt, Harris, and Harris, in prep
Automatic mutational signature extraction from Europeans (CEU)
UMAP visualization of mutation spectrum divergence over time

Dewitt, Harris, and Harris, in prep
Great ape species display greater mutation spectrum drift than human populations do.

Mutation spectrum of human #1: (% AA > ACA, % AA > AGA, …)

- Human
- Gorilla
- Bornean Orangutan
- Chimpanzee
- Bonobo
- Western chimps
- Other chimps

PC1 (48.2%) and PC2 (21.5%) variance explained

Goldberg and Harris, bioRxiv preprint
Ape mutation spectra cluster by phylogeny, pointing to fixation of genetic mutators (not environmental mutagens)

Euclidean distances, NCNR compartment, 100x

Goldberg and Harris, bioRxiv preprint
A case study of a mutational process that complicates population genetics

Multinucleotide mutations (MNM) are nearby SNPs that appear in the same generation

\[
\begin{align*}
AAAGTTTAGCCGACAC
\end{align*}
\]
\[
\downarrow
\]
\[
AAAGATAAACCAGACAC
\]

Schrider, et al. 2011

Harris and Nielsen 2014
Effect of MNMs in the distribution of tracts of identity by state

![Graph showing the distribution of L-base IBS tracts by IBS tract length (L). The graph includes data for Europe, Africa, Europe vs. Africa, and a theoretical prediction. The y-axis represents the frequency of L-base IBS tracts on a log-scale, ranging from $10^{-16}$ to $10^{-5}$. The x-axis represents the IBS tract length (L) on a log-scale, ranging from $10^0$ to $10^6$. The graph highlights the difference between ancient and recent demographic history.]
Direct evidence for MNMs

- Most methods assume that all SNPs arise from rare, independent mutation events

- MA experiments and trio sequences show that *de novo* mutations are too clustered for this to be true

Independent mutation hypothesis

Observed: Excess correlation between *de novo* mutations

“Mutator” yeast strains: Some abnormal polymerases generate clustered mutations at a higher rate
MNMs could accelerate evolution across fitness valleys
Multinucleotide mutation should create pairs of SNPs in perfect linkage disequilibrium (LD) (derived alleles occur in the same set of individuals)

```
multinucleotide mutation

Perfect LD

Not Perfect LD
```
Independent mutations at neighboring sites can also create SNPs in perfect LD.

One MNM

Two independent mutations
Compared to theoretical predictions, the 1000 Genomes Phase I data (1,092 humans from Africa, Europe, Asia, and the Americas) has excess close-together SNPs in perfect LD.
SNPs in perfect LD are enriched for transversions

- 66% of human mutations are transitions (A>G, G>A, C>T, T>C)
- Pairs of SNPs in perfect LD are enriched for transversions, suggesting a different balance of mutational signatures
Transversion-enrichment as a function of the distance between linked SNPs
A candidate mechanism: error-prone translesion synthesis

Northam et al., Nucleic Acids Res. 2014
Matching mutational signatures between human variation and laboratory yeast

Environmental and Molecular Mutagenesis 53:777–786 (2012)

Research Article

DNA Polymerase zeta Generates Clustered Mutations During Bypass of Endogenous DNA Lesions in Saccharomyces cerevisiae

Jana E. Stone, Scott A. Lujan, and Thomas A. Kunkel*
Laboratory of Molecular Genetics and Laboratory of Structural Biology, National Institute of Environmental Health Sciences, NIH, DHHS, North Carolina

• Stone, et al. created yeast deficient in nucleotide excision repair machinery and observed a high MNM rate

• Mechanism: increased translesion synthesis by Pol Zeta
A matching dinucleotide mutational signature

- Frequency of $A_1D_2 \rightarrow D_1D_2$

- Graph showing the frequency of SNP pairs compared to Stone et al. 2012 and 1000 Genomes.

- SNP pairs include: GAGCGCGAGAT, AGCGACCACGC, CCCAATACA, and others not observed in Stone et al. 2012.
Further characterization of the Pol zeta mutational signature

- GC>AA mutations are concentrated in late-replicating regions of the genome

- Usually occur in GCG context, triggered by CpG deamination followed by polymerase stalling

- CpG deamination is triggered by transcription; usually occurs on transcribed strand

- Some genes contain GC>TT mutation hotspots, including HRAS where the mutation causes the Mendelian disorder Costello Syndrome

Seplyarskiy, et al. 2015
Costello Syndrome is caused by selection within the aging testis

- A high penetrance Mendelian disease caused by a nonsynonymous point mutation in the HRAS oncogene
- Causes developmental delay and early childhood tumors
- Most commonly caused by a GC>TT mutation with a mutation rate of $10^{-5}$ per generation (normal mutation rate is $10^{-8}$ per site per generation)
- Biggest risk factor is paternal age
HRAS mutations experience selfish selection within the testis

Goriely and Wilkie 2012