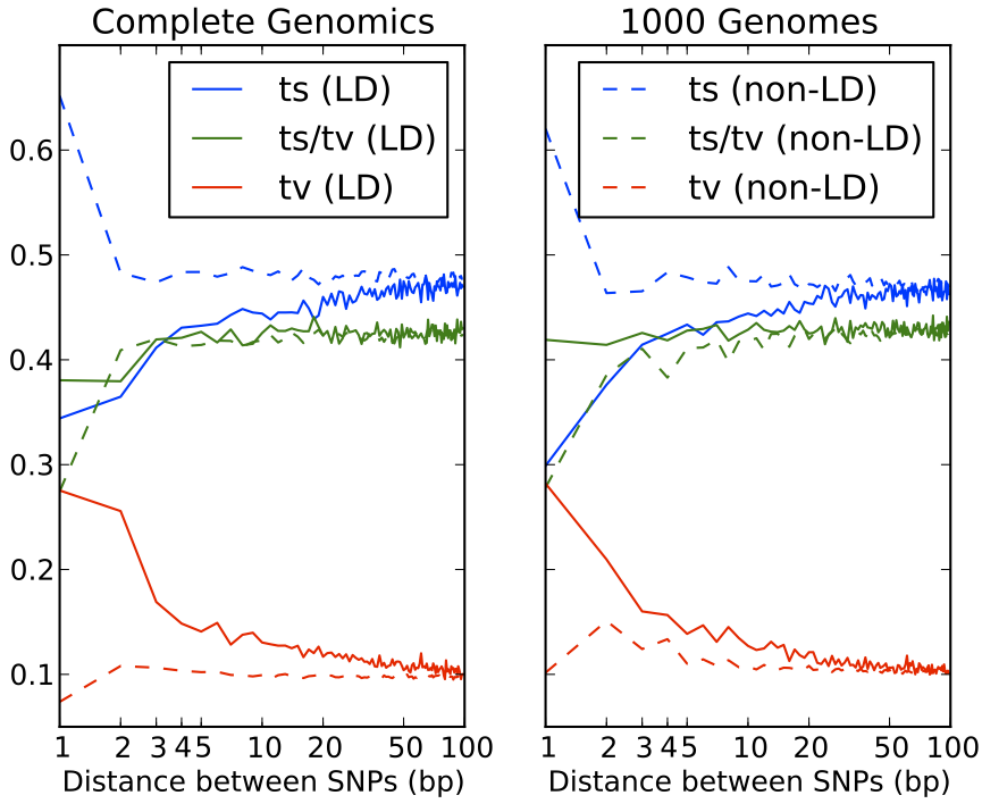
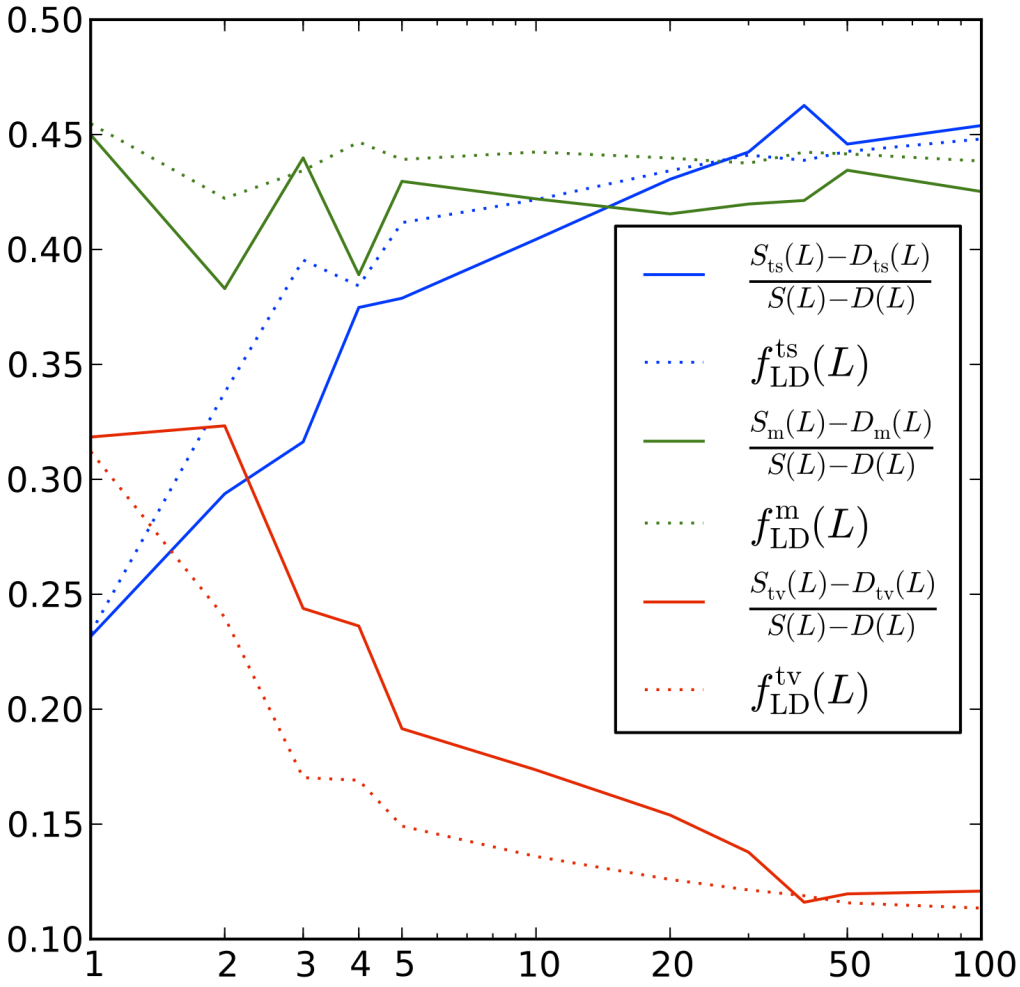
Supplemental Information: Error-prone polymerase activity causes multinucleotide mutations in humans

Kelley Harris and Rasmus Nielsen



Supplementary Figure S1. Consistency of the transition: transversion ratio across linked SNPs from different sequencing platforms. This figure shows that excess transversions in perfect LD are not an artifact of Illumina sequencing or the 1000 Genomes pipeline, but are also present in a set of 54 human genomes sequenced by Complete Genomics (CG). To make this comparison, we subsampled 54 genomes from the 1000 Genomes Phase I dataset that had approximately the same population breakdown as the 54 CG individuals. Because the CG data are unphased, we ignore all 1000 Genomes phasing information, classifying each SNP pair as being in perfect LD if it is in perfect LD with respect to at least one possible haplotype phasing. We ignore all CG SNPs at which more than 10% of the samples have a missing genotype. Note that most pairs of nearby SNPs in the CG data are not annotated as "SNPs" in the MasterVarBeta files that are publicly availably, but as "complex" substitutions where a string of two or more bases is regarded as one polymorphic unit. We ignored all complex substitutions that included indels, but extracted SNPs from each substitution multi nucleotide substitution where all variant alleles had the same length and a one-to-one mapping between sites was possible.

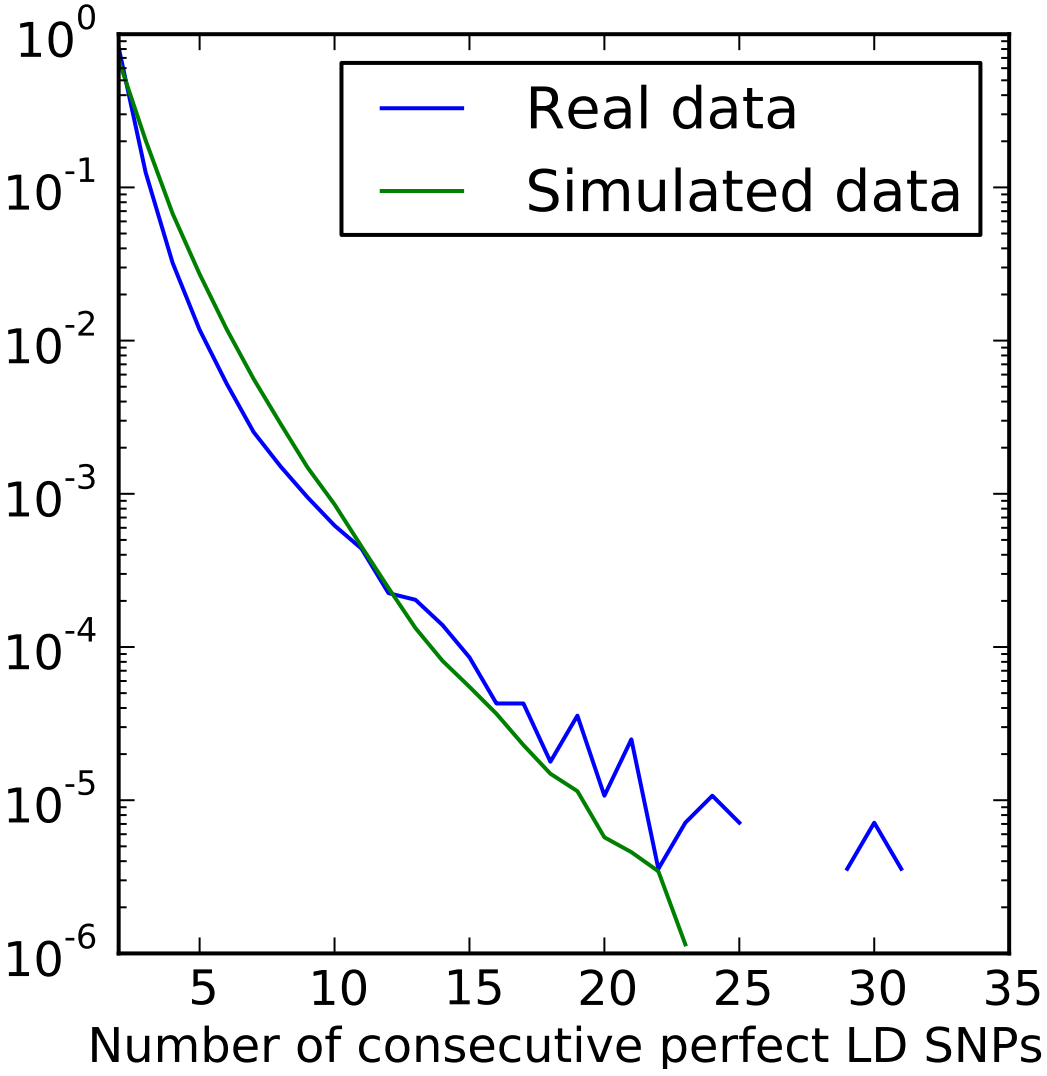


Supplementary Figure S2. Quantifying simultaneous transitions vs. transversions. This figure plots the relative abundances of transitions, transversions, and mixed pairs as fractions of the quantity $S(L) - D(L)$. The transversion fraction $(S_{\rm tv}(L) - D_{\rm tv}(L))/(S(L) - D(L))$ is slightly higher than $f_{\rm LD}^{\rm tv}(L)$, especially for small L where MNMs are the most apparent.

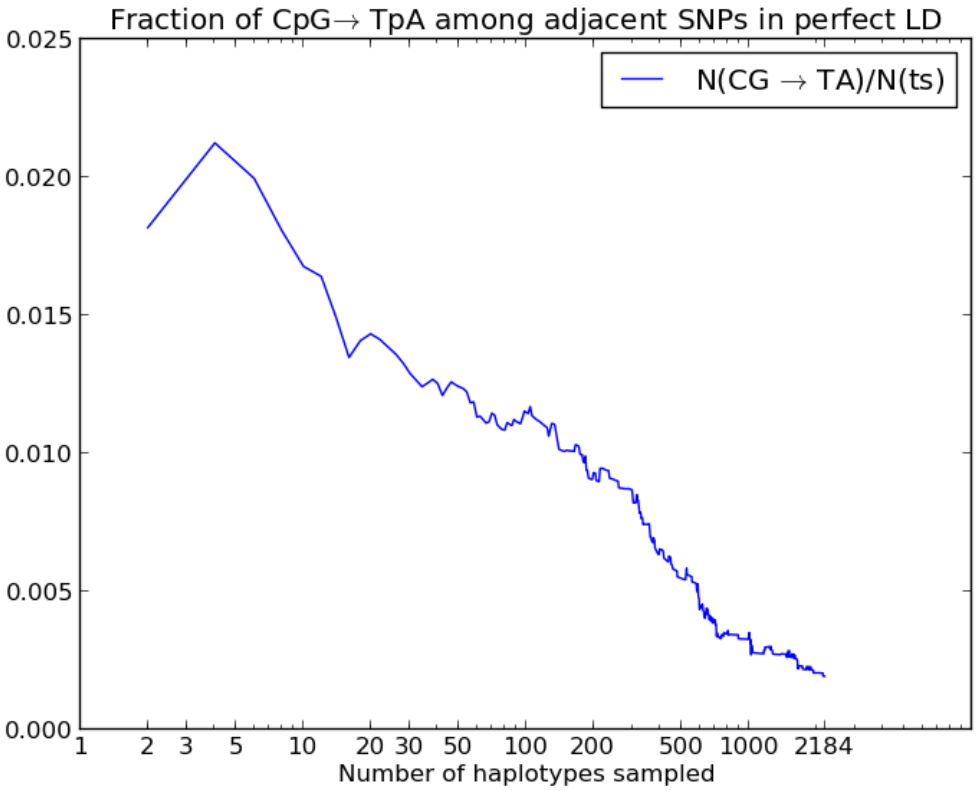
Supplementary Table S1. Quantifying MNMs spanning > 100 bp

L	$S_{\mathrm{ts}}^{(\mathrm{LD})}(L)$	$m_{\mathrm{ts}}(L)$	$S_{\mathrm{m}}^{(\mathrm{LD})}(L)$	$m_{\mathrm{m}}(L)$	$S_{\mathrm{tv}}^{(\mathrm{LD})}(L)$	$m_{\mathrm{tv}}(L)$
100	4.507487e+10	0.370434497853	4.414249e+10	0.354296958936	1.142233e+10	0.389186681775
200	4.217310e+10	0.379874606701	4.100119e+10	0.372376712439	1.054684e+10	0.37707655264
300	3.961781e+10	0.395451459315	3.880264e+10	0.377681881936	9.884130e+09	0.371605021077
400	3.807976e+10	0.399251605723	3.726721e+10	0.395012183525	9.351356e+09	0.38983897119
1000	2.094000e+10	0.426248637331	2.096024e+10	0.423787575103	5.275246e+09	0.397717995479
3000	1.287797e+10	0.443566715331	1.270760e+10	0.418373624368	3.234083e+09	0.382000444317
5000	9.147849e+09	0.403365299761	9.088510e+09	0.392177623168	2.377678e+09	0.38602646942
10000	5.147443e+09	0.290639194365	5.139715e+09	0.280151830098	1.330389e+09	0.267209538362

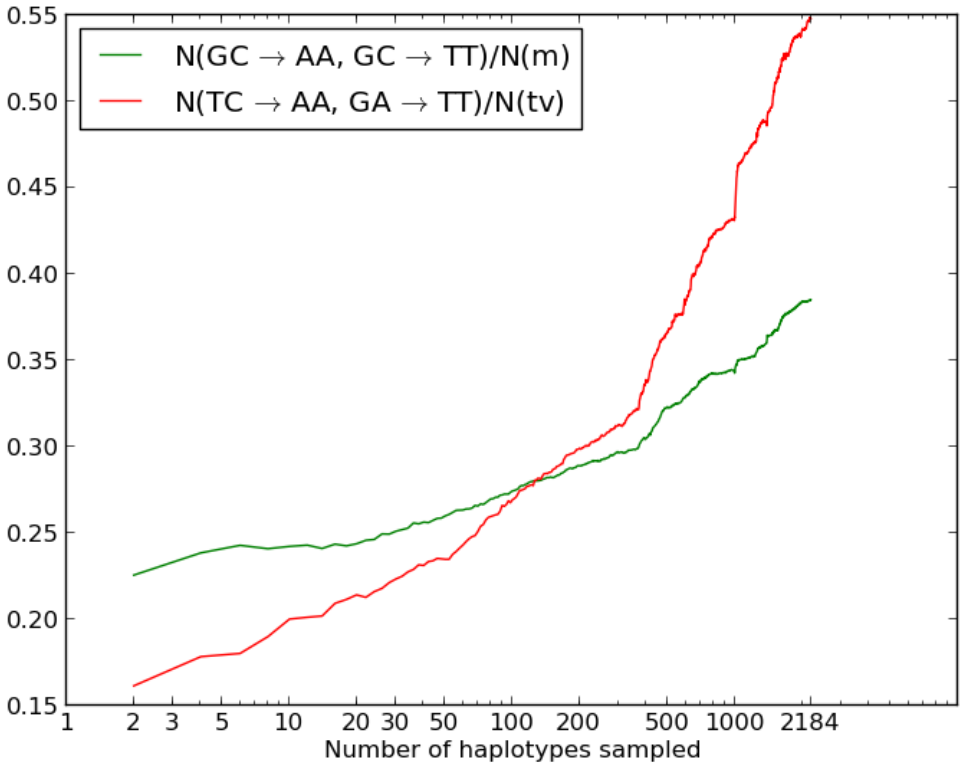
For each distance L listed above, we counted all SNPs in perfect LD between L and $L + 100$ bp apart. For each pair type t, we subsampled haplotype pairs in order to calculate $S_t(L), D_t(L), S_t^{(\mathrm{LD})}(L)$, and $m_t(L)$ aggregated over this 100 bp window. The results suggest that the ratio of MNMs to perfect LD SNPs achieves its minimum value around $L = 100$ and then stops decreasing with the distance between SNPs.



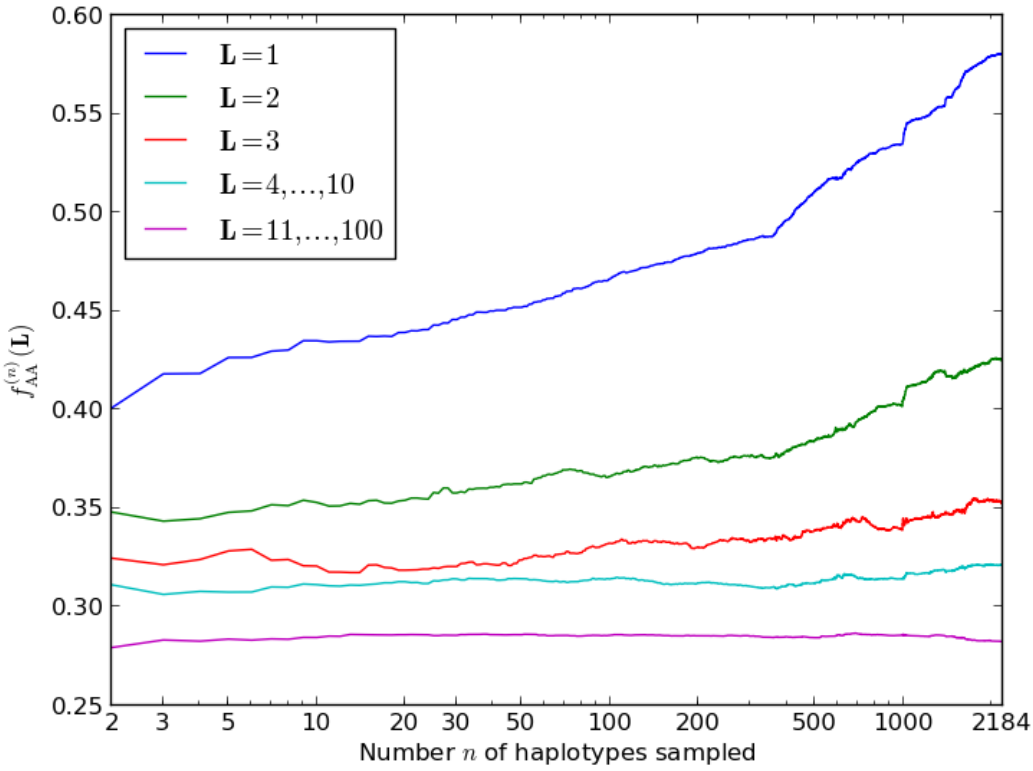
Supplementary Figure S3. Clusters of 2 or more perfect LD SNPs. In the 1000 Genomes data, we found all clusters of 2 or more perfect LD SNPs with fewer than 1 kb between adjacent pairs. We plot the resulting distribution of cluster sizes and compare it to the distribution of cluster sizes in data simulated under the Harris and Nielsen (2013) model using `ms`. The cluster sizes from real data are slightly more dispersed toward very small and very large clusters. It is possible that the longest clusters formed by error-prone replication of single-stranded DNA following double-strand breakage as proposed by Roberts et al. (2012).



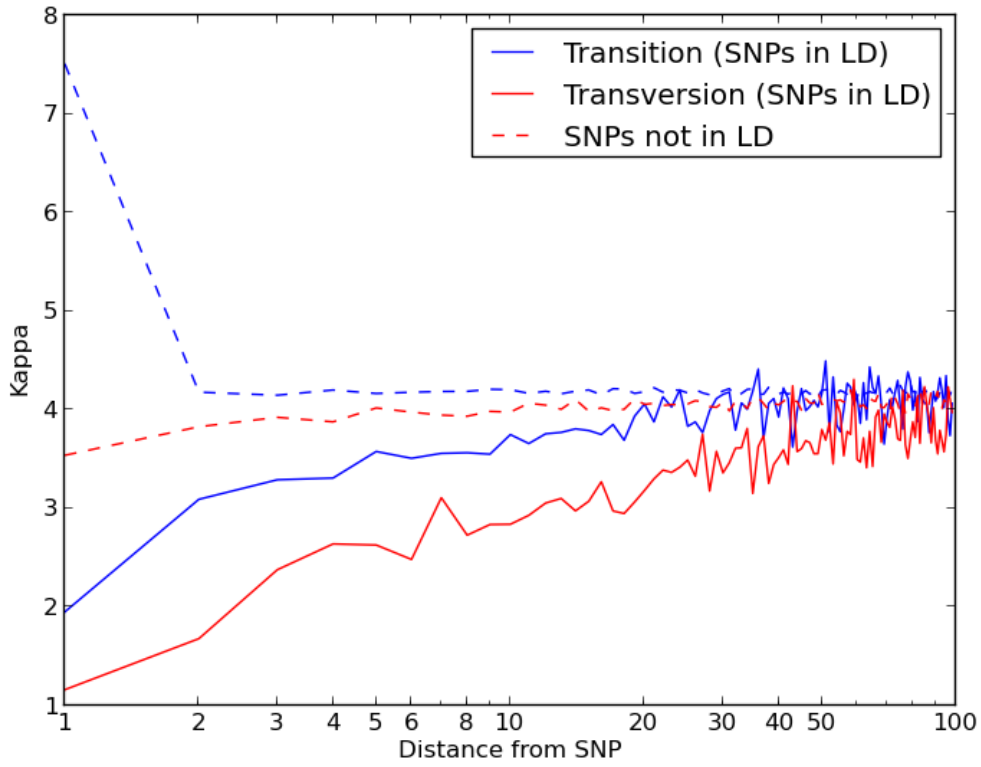
Supplementary Figure S4. CpG double mutations in perfect LD. It is well known that CpG dinucleotides undergo a high rate of C→T mutation related to deamination of the methylated cytosine. This process should elevate the rate at which CpG mutations undergo double transitions into the TpA dinucleotide. When the CG → TA double mutation is mediated by cytosine deamination, the two mutations should occur non-simultaneously in general, and are more likely to be found in perfect LD in small samples than large samples. We counted dinucleotide mutations in subsets of the 1000 Genomes data ranging in size from 2 to 2,184 haplotypes and tabulated the fraction of perfect LD transition pairs that were of allelic type CG → TA. As shown here, this fraction declines almost 10-fold as the sample size increases from 2 to 2,184.



Supplementary Figure S5. Pol ζ mutations in subsets of the 1000 Genomes data. We counted all four types of Pol ζ-associated mutations in subsamples of 2–2184 1000 Genomes haplotypes and pooled counts of the reverse complements TC→AA, GA→TT and GC→AA, GC→TT. We divided these counts, respectively, by the total counts of adjacent transversions and adjacent mixed pairs in perfect LD. The fraction of each Pol ζ-associated mutation increases with sample size as expected for MNMs.



Supplementary Figure S6. Linked derived AA/TT allele pairs. To calculate $f_{\mathrm{AA}}^{(n)}(\mathbf{L})$ for a vector $\mathbf{L}$ of lengths $L_1, \ldots, L_k$ measured in base pairs, we sampled n haplotypes from the 1000 Genomes data and counted to total number of perfect LD SNPs separated by a distanced falling in the range $L_1, \ldots, L_k$. We then counted the number of these with AA or TT as the pair of derived alleles and computed the ratio of AA/TT pairs to total pairs. As shown, the three curves $f_{\mathrm{AA}}^{(n)}(1)$, $f_{\mathrm{AA}}^{(n)}(2)$, and $f_{\mathrm{AA}}^{(n)}(3)$ are all increasing functions of n, the proportion of AA/TT derived alleles increasing as more haplotypes are sampled. However the same pattern does not hold for vectors $\mathbf{L}$ of longer distances. This suggests that derived AA/TT's are only overrepresented among MNMs separated by very short distances of 1 to 3 base pairs.



Supplementary Figure S7. Explaining the variation of κ in the neighborhood of a SNP. It is common to summarize the transition: transversion ratio with a parameter κ for which $f_{\text{transversion}} = 2/(\kappa + 2)$. On average, $\kappa = 4$ in the human genome, meaning that each type of transition is 4 times as abundant as a particular type of transversion. Looking only at SNPs in LD, κ appears significantly depressed in the neighborhood of a segregating SNP, suggesting that a negative positional correlation between mutation rate and κ. However, this effect is not apparent among SNPs that are not in LD. κ is elevated, on average, at the site next to a transition because of adjacent transitions generated by mutations at both positions of a CpG site. In addition, κ appears slightly reduced among unlinked SNPs in the neighborhood of a transversion, suggesting that there is some regional variation in κ but not as much as appears to be the case if simultaneous mutations are regarded as independent.

1 Exponential decay of LD over short genomic distances

In data simulated under the standard coalescent with recombination using `ms` (Hudson, 2002), we saw that the count $N_{\mathrm{LD}}(L)$ of SNPs in perfect LD L bp apart decays approximately exponentially for L between 1 and 100 bp. Here, we give a heuristic argument why this should be true in the asymptotic limit $L \ll 1/\rho$, where ρ is the population-scaled recombination rate.

Let $T_1, \ldots, T_L$ be the sequence of n-leaf coalescence trees that occur at the sites of a sequence of length L that has been evolving with mutation and recombination parameters θ and ρ. For simplicity, we assume a constant effective population size N. The rates θ and ρ are population-scaled such that $\mu = \theta/(4N)$ is the mutation rate per site per generation and $r = \rho/(4N)$ is the recombination rate per site per generation. Given any of these trees T_i, let $P(T_i)$ be the set of points $(x, y) \in T_i$ with the property that x and y lie on the same branch of T_i. The sequential coalescent yields a natural map from points on T_i to points on T_{i+1}, though not every point on T_i necessarily maps to a point on T_{i+1} if a recombination has occurred between the sites. Let $\epsilon_{x,y}(T_i)$ be defined such that $1-\epsilon_{x,y}(T_i)$ is the probability that x and y both map to T_{i+1}, $(x, y) \in P(T_{i+1})$, and the branch containing (x, y) subtends the same set of lineages in both T_i and T_{i+1}.

A pair of points $(x, y) \in P(T_i)$ can give rise to a pair of SNPs in perfect LD at sites i and j if the following events occur: E1) a mutation occurs at position x on tree T_i, E2) x and y map to a single branch of each tree between T_i and T_j that subtends the same set of lineages, and E3) A mutation occurs at position y on tree T_j. Not every pair of SNPs in perfect LD must correspond to a pair of points x, y satisfying E1–E3; for example, the integrity of the clade by the branch containing x and y could be broken up and re-formed by two separate recombinations occurring between sites i and j. If the sample size n is relatively large, however, it will be combinatorially unlikely for any clade to re-form after it has been broken up by recombination, particularly within a very short genomic window. Motivated by this, we will estimate $N_{\mathrm{LD}}(L)$ assuming that all linked SNP pairs arise at pairs of points (x, y) that satisfy E1–E3 for some T_i and T_j.

Integrating over x, y and $T_i, \ldots, T_{i+L}$, we compute that the probability of observing a pair of SNPs in perfect LD at sites i and $i + L$ is the following:

$$\begin{aligned} N_{\mathrm{LD}}(i, i+L) &= \theta^2 \int_{T_i,\ldots,T_{i+L}} \int_{(x,y)\in P(T_i)} (1-\epsilon_{x,y}(T_i)) \cdots (1-\epsilon_{x,y}(T_{i+L})) d_{(x,y)} d_{(T_i,\ldots,T_{i+L})} \qquad (1) \\ &= \theta^2 + \theta^2 \sum_{k=1}^{L} (-1)^k \int_{T_i,\ldots,T_{i+L}} \int_{(x,y)\in P(T_i)} \sum_{i \le j_1 < \cdots < j_k \le i+L} \epsilon_{x,y}(T_{j_1}) \cdots \epsilon_{x,y}(T_{j_k}) d_{(x,y)} d_{(T_i,\ldots,T_{i+L})}. \end{aligned}$$

Let $\ell(T)$ denote the total branch length of tree T. Since any alteration of tree structure requires a recombination event, $\epsilon_{x,y}(T) \le \rho \cdot \ell(T)$. This implies that

$$\sum_{i \le j_1 < \cdots < j_k \le i+L} \epsilon_{x,y}(T_{j_1}) \cdots \epsilon_{x,y}(T_{j_k}) \le (\epsilon_{x,y}(T_i) + \cdots + \epsilon_{x,y}(T_{i+L}))^k \le (L\rho(\ell(T_i) + \cdots + \ell(T_{i+L})))^k \quad (2)$$

for every k. Letting $T^{(2)}$ denote the sum of squares of the branch lengths of a coalescent tree T, this implies that

$$\begin{aligned} N_{\mathrm{LD}}(i, i+L) &= \theta^2 - \theta^2 \int_{T_i,\ldots,T_{i+L}} \int_{(x,y)\in P(T_i)} \rho(\epsilon_{x,y}(T_i) + \cdots + \epsilon_{x,y}(T_{i+L})) d_{(x,y)} d_{T_i,\ldots,T_{i+L}} + O((\rho L)^2) \\ &= \theta^2 \mathbb{E}(T^{(2)})(1 - \rho L \cdot \mathbb{E}(\epsilon_{x,y}(T_i))) + O((\rho L)^2). \qquad (3) \end{aligned}$$

In human-like data where $N = 10,000$ and $\rho = 0.0004$, we can see that $\rho L \le 0.04 \ll 1$ when $L < 100$. Therefore, the first-order linear decay rate of $N_{\mathrm{LD}}(i, i + L)$ is small compared to L. In addition, we can

see from equation (2) that the $O((\rho L)^2)$ term of the Taylor expansion will be positive, meaning that $N_{\mathrm{LD}}(i, i+L)$ has concave upward shape. This makes it reasonable, for our purposes, to approximate $N_{\mathrm{LD}}(i, i+L)$ by an exponential function.

2 Enrichment of MNMs in large datasets

As a consequence of the argument in Section S1, we saw that the abundance of linked independent mutations in a sample of n lineages is proportional to the expected sum of squared branch lengths in an n-leaf coalescence tree. This is a simple consequence of the fact that two mutations must affect a single branch to create SNPs in perfect LD. In contrast, the abundance of MNMs should be proportional to the total tree length, just as the total number of segregating sites is proportional to the expected tree length.

It is a standard result in population genetics that the expected total tree length $\mathbb{E}(T_{\mathrm{total}})$ equals the harmonic number $\sum_{i=1}^{n-1} 1/i$ (Watterson, 1975). To show this, let T_i be the length of time that the a random genealogy has exactly i lineages, which has distribution function $f_i(t) = \binom{i}{2}\exp(-t\binom{i}{2})$. It follows that

$$\mathbb{E}(T_{\mathrm{total}}) = \mathbb{E}\left(\sum_{i=2}^{n} iT_i\right) = \sum_{i=2}^{n} i\mathbb{E}(T_i) = \sum_{i=2}^{n} i \cdot \frac{2}{i(i-1)} = \sum_{i=1}^{n-1} \frac{1}{i} \approx \log(n-1) \tag{4}$$

Therefore, if μ_{MNM} is the rate of MNMs per coalescent time unit, the expected number of MNMs approaches infinity with increasing n at the asymptotic rate $\mu_{\mathrm{MNM}}\log(n)$.

In contrast, if μ is the rate of ordinary point mutations, linked independent mutations appear at the rate $\mu^2\mathbb{E}(T^{(2)}_{\mathrm{total}})$, where $T^{(2)}_{\mathrm{total}}$ is the sum of squares of the coalescent tree branch lengths. We can show that $\mathbb{E}(T^{(2)}_{\mathrm{total}})$ approaches a constant as $n \to \infty$. To proceed, we let $\ell_1, \ldots, \ell_n$ denote the lengths of the n leaves of the tree and $b_{n-1}, \ldots, b_2$ denote the lengths of the $n-2$ internal branches, indexed such that the more recent endpoint of branch i is the first time when the tree has i lineages:

$$\mathbb{E}(T^{(2)}_{\mathrm{total}}) = n\mathbb{E}(\ell_n^2) + \sum_{i=2}^{n-1} \mathbb{E}(b_i^2). \tag{5}$$

Given that a branch is present when the tree has i lineages, the probability that the branch is ended by the next coalescence event is $(i-1)/\binom{i}{2} = 2/i$. Therefore, given $j < i$, the probability that $b_i = T_i + \cdots + T_j$ is

$$\mathbb{P}(b_i = T_i + \cdots + T_j) = \left(1 - \frac{2}{i}\right)\cdots\left(1 - \frac{2}{j+1}\right)\cdot\frac{2}{j} = \frac{(i-2)\cdots(j-1)\cdot 2}{i\cdots(j+1)\cdot j} = \frac{2(j-1)}{i(i-1)}. \tag{6}$$

It follows that

$$\begin{aligned}
\mathbb{E}(b_i^2) &= \sum_{j=2}^{i} \mathbb{P}(b_i = T_i + \cdots + T_j) \cdot \mathbb{E}((T_i + \cdots + T_j)^2) && (7)\\
&= \sum_{j=2}^{i} \frac{2(j-1)}{i(i-1)} \left(\sum_{k=j}^{i} \mathbb{E}(T_k^2) + 2 \sum_{j \le k < \ell \le i} \mathbb{E}(T_k)\mathbb{E}(T_\ell) \right) && (8)\\
&= \sum_{j=2}^{i} \frac{2(j-1)}{i(i-1)} \left(\sum_{k=j}^{i} \frac{8}{k^2(k-1)^2} + \sum_{j \le k < \ell \le i} \frac{8}{k(k-1)\ell(\ell-1)} \right) && (9)\\
&= \sum_{j=2}^{i} \frac{2(j-1)}{i(i-1)} \left(\sum_{k=j}^{i} \frac{4}{k^2(k-1)^2} + \left(\sum_{k=j}^{i} \frac{2}{k(k-1)} \right)^2 \right). && (10)\\
&= \sum_{j=2}^{i} \frac{2(j-1)}{i(i-1)} \left(\frac{4}{3} \left(\frac{1}{j^3} - \frac{1}{i^3} \right) + \left(\frac{2}{j-1} - \frac{2}{i} \right)^2 + O\left(\frac{1}{j^4} + \frac{1}{i^4} \right) \right). && (11)\\
&= \frac{2}{i(i-1)} \left(4\log(i-1) - 3/2 \right) + O(i^{-3}). && (12)
\end{aligned}$$

This implies that

$$\begin{aligned}
\mathbb{E}(T_{\text{total}}^{(2)}) &= n\mathbb{E}(b_n^2) + \sum_{i=2}^{n-1} \mathbb{E}(b_i^2) && (13)\\
&= \frac{8\log(n-1)}{n-1} + \sum_{i=2}^{n-1} \frac{8\log(i-1)}{i(i-1)} + O(1/n) && (14)\\
&= \frac{1}{2}(\log(2)+1) + \frac{7\log(n-1)}{n-1} + O(1/n), && (15)
\end{aligned}$$

which decreases asymptotically to the limit $(\log(2)+1)/2$ as n approaches infinity.

It may seem counterintuitive that $\mathbb{E}(T_{\text{total}}^{(2)})$ decreases as more lineages are sampled and $\mathbb{E}(T_{\text{total}})$ increases unboundedly, but in both simulated and real data we observe fewer SNPs in perfect LD in a sample of 2,184 haplotypes than in a subset of e.g. 1,000 haplotypes. To explain why, we note that the total tree length grows at rate $\log(n)$ as more lineages are sampled, but the tree length is subdivided among distinct branches at the faster rate $O(1/n)$. Because branch subdivision occurs faster than the growth rate of the total tree length, the sum of squared branch lengths decreases with increasing sample size, reducing the prevalence of independent linked SNPs and enhancing the signature of MNMs.

3 Close LD SNPs in a single diploid genome

In their paper titled "Pervasive multinucleotide mutational events in eukaryotes," Schrider, *et al.* used a chimpanzee outgroup to polarize SNPs found in human trio data (Schrider et al., 2011). For each pair of SNPs fewer than 20 bp apart, they recorded whether the derived alleles lay on the same phased haplotype or on different haplotypes. Figure 3 of their main text records the number of nearby derived alleles that they found on the same lineage ($N_S(L)$) or on different lineages ($N_D(L)$) as a function of the distance L between the sites. By the reasoning in their paper, $N_D(L)$ should equal the number of derived allele pairs found on the same haplotype L bp apart that were produced by independent mutations. $N_S(L) - N_D(L)$

should equal the number that were created by multinucleotide mutation. From visual inspection of their Figure 3, it appears that

$$\frac{N_S(1) - N_D(1)}{N_S(1)} \approx \frac{1400 - 2000}{1400} \approx 0.86,$$

meaning that about 86% of adjacent same-lineage SNPs were caused by MNM. In contrast, in appears that

$$\frac{N_S(20) - N_D(20)}{N_S(20)} < \frac{100}{2000} = 0.05,$$

meaning that fewer than 5% of same-lineage SNPs 20 bp apart were caused by MNM.

Schrider, *et al.* defined a multinucleotide polymorphism (MNP) to be a pair of SNPs within 20 bp of each other where both derived alleles lie on the same haplotype of a diploid genome that has been phased using trio information. For each possible set of two ancestral and two derived alleles, they tabulated the abundance of that pair as a fraction of all MNPs in their dataset. In particular, they reported that GC→TT plus its reverse complement GC→AA made up 2.3% of the total, more than any other mixed transition/transversion mutation pair but fewer than the 16% of adjacent perfect LD mutations reported in our paper. Similarly, TC→AA and GA→TT made up 1.5% of their total, compared to 10% of ours. Derived AA/TT allele pairs made up 15.8% of their total MNPs.

We believe that two factors explain the differences in these results. One is that Schrider, *et al*, pooled together all MNPs separated by 1–20 bp instead of focusing on adjacent polymorphisms. The other is that they were only able to sample pairs in perfect LD in a sample of size two, which should contain more linked independent mutations than perfect LD pairs in a larger sample of haplotypes. To verify this, we replicated the Schrider, *et al.* analysis on our data, tabulating all pairs of SNPs less than 20 bp apart where the derived alleles lay on the same haplotype of at least one 1000 Genomes individual. We obtained that GC→TT plus GC→AA made up 3.0% of the total, TC→AA and GA→TT made up 12.8%, and XY→AA plus XY→TT made up 19.0%. These numbers are similar enough that the two datasets do not appear to be qualitatively different; it is only the difference in sampling scheme that reveals more Pol ζ-associated mutations in our analysis of the 1000 Genomes data. Figure S5 confirms that Pol ζ-associated tandem mutations increase in frequency as more lineages are sampled, suggesting that these mutations are enriched among MNMs. Figure S6 demonstrates a similar result for the frequency $f_{\mathrm{AA}}(L)$ of derived AA/TT allele pairs in perfect LD at distances of L =1–3 bp apart. Although $f_{\mathrm{AA}}(L)$ decreases as L increases from 4 to 1,000, it does not appear to depend on the number of haplotypes sampled. A possible explanation is that MNMs only create excess derived AA/TTs that lie 1–3 bp apart but that transcriptional strand bias creates excess AA/TTs among independent mutations that occur at nearby sites (Green et al., 2003).

4 Disruption of MNMs by recombination

We have seen that MNM is not the only source of perfect LD SNPs in genetic data. Conversely, not all MNMs need give rise to perfect LD SNPs; it is possible for recombination to decouple two derived alleles that have been created by a single complex mutation. To assess how many MNMs are likely to be broken down in this way, we used `ms` to simulate many short ancestral recombination graphs on 2,184 haplotypes. To mimic the composition of the 1000 Genomes data, we sampled 492 African haplotypes and 1,692 non-African haplotypes according to the model of human history proposed in Harris and Nielsen (2013). For each length L listed in the table below (L = 1, 10, 100, 1000, 10000), we simulated LOOKUP sequences L bp long assuming a constant recombination rate of 1.0×10^{-8} recombinations per site per generation. Given the trees at the left and right ends of each simulated sequence s_i, we calculated the probability p_i that a point placed on the left tree uniformly at random would be ancestral to a subtree that had not recombined between the left tree and the right tree. Letting ℓ_i be the total branch length of

the leftmost tree, we obtain the following estimate $\hat{\rho}_{\mathrm{MNM}}(L)$ for the probability that an MNM spanning L bp will be broken up by recombination:

$$\hat{\rho}_{\mathrm{MNM}}(L) = \frac{\sum_{i=1}^{k}(1-p_i)\cdot \ell_i}{\sum_{i=1}^{k} \ell_i} \tag{16}$$

The resulting values of $\hat{\rho}_{\mathrm{MNM}}(L)$ are recorded in Table 2. Even for $L = 10000$, we estimate that only 22% of MNMs will recombine out of perfect LD. Although 10,000 bp is about ten times the mean recombination distance for a sample of two lineages, the majority of MNMs are expected to occur at low frequency on the sample of 2,184 haplotypes and persist in LD for relatively long genomic distances.

Supplementary Table S2. Estimates of $\hat{\rho}_{\mathrm{MNM}}(L)$

L	$\hat{\rho}_{\mathrm{MNM}}(L)$
1	6.60×10^{-5}
10	3.20×10^{-4}
100	3.64×10^{-3}
1000	4.70×10^{-2}
10000	0.224

5 Simulating data with a realistic MNM distribution

We argue that MNM affects many features of genetic data including SNP density, the local transition/transverion ratio, and linkage disequilibrium. To capture these effects, it may be useful for readers to incorporate MNM into simulations of human-like SNP data. Tables S3 and S4 provide a framework for doing this. For each SNP pair type $t \in \{\mathrm{ts}, \mathrm{m}, \mathrm{tv}\}$ and each distance L between 1 and 100 bp, the table entry (L, t) provides the probability $P_{\mathrm{MNM}}(L, t)$ that a given mutation should be an MNM of type t and spacing L. The entries of the two tables add up to 0.018, indicating that MNMs should account for 1.8% SNPs. To simulate a dataset with θ total SNPs, one should first use a program such as `ms` to generate a dataset with $\theta \times (1 - 0.009)$ total SNPs. After this, one should select a fraction $P(L, t)$ of SNPs uniformly at random to be MNMs of type t and spacing L. For each selected SNP, a new SNP should be introduced in perfect LD exactly L bp to the left.

References

Green P, Ewing B, Miller W, Thomas P, and Green E. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genetics* **33**: 514–517.

Harris K and Nielsen R. 2013. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics* **9**: e1003521.

Hudson R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337–338.

Roberts S, Sterling J, Thompson C, Harris S, Mav D, Shah R, Klimczak L, Kryukov G, Malc E, Mieczkowski P, et al.. 2012. Clustered mutations in yeast and in human cancers can arise from damaged long single-stranded DNA regions. *Mol Cell* **46**: 424–435.

Schrider D, Hourmozdi J, and Hahn M. 2011. Pervasive multinucleotide mutational events in eukaryotes. *Curr Biol* **21**: 1051–1054.

Watterson G. 1975. On the number of segregating sites in genetical models without recombination. *Theor Pop Biol* **7**: 256–276.

Supplementary Table S3. Table of Parameters for Simulating Data with Realistic MNM (Part 1 of 2)

L	ts	m	tv
1	0.000689075099285	0.00139395307314	0.0011649961804
2	0.000197253881722	0.000264153610395	0.000248926345734
3	9.24985667775e-05	0.00014275347624	9.31164214363e-05
4	0.000114899830562	0.000122700855385	8.44464230223e-05
5	9.3567275539e-05	0.000114868148911	5.64888107995e-05
6	8.36026776582e-05	9.12884447692e-05	4.88456557619e-05
7	7.58303670667e-05	8.55603852686e-05	4.15000412291e-05
8	7.5266572516e-05	8.47592580657e-05	4.42345400698e-05
9	7.59109091453e-05	8.5119765307e-05	4.01596006209e-05
10	7.11589265039e-05	7.91113112853e-05	3.86046895154e-05
11	7.1486593824e-05	7.26941774735e-05	3.21955518884e-05
12	7.18145139791e-05	7.38269722864e-05	3.14879573414e-05
13	6.48462106821e-05	7.07801448587e-05	3.13363299385e-05
14	6.77155120397e-05	7.02332783973e-05	3.01738531827e-05
15	6.52151208566e-05	6.85536171231e-05	2.86070366858e-05
16	6.83303623306e-05	6.734269853e-05	2.8152154477e-05
17	6.53380909148e-05	7.12488875399e-05	2.9415716168e-05
18	6.28786897511e-05	7.09363924191e-05	2.97695134415e-05
19	6.72646218263e-05	6.64833369478e-05	2.62820831742e-05
20	6.77155120397e-05	6.56239753657e-05	2.74951023976e-05
21	6.31325919102e-05	6.26737000383e-05	2.31857258277e-05
22	6.11743849651e-05	6.08755305214e-05	2.15897676445e-05
23	6.07827435761e-05	6.29359330929e-05	2.1678431988e-05
24	6.13310415207e-05	6.08380686565e-05	2.07917885529e-05
25	5.98428042425e-05	6.06132974668e-05	2.20774215338e-05
26	5.9607819409e-05	6.27860856331e-05	2.20774215338e-05
27	5.99211325203e-05	6.26737000383e-05	2.07917885529e-05
28	5.9646983548e-05	6.1100301711e-05	2.14567711293e-05
29	6.15660263542e-05	5.95269033837e-05	2.03041346636e-05
30	5.91378497422e-05	5.92272084642e-05	1.97278164308e-05
31	6.44333909867e-05	5.74757746569e-05	1.85649923763e-05
32	6.63962625002e-05	6.4106112176e-05	1.74589928305e-05
33	6.53721556235e-05	5.87093258232e-05	1.7182492944e-05
34	6.69509870583e-05	5.95573922501e-05	1.77354927169e-05
35	6.23851772334e-05	6.09836857862e-05	1.76169927656e-05
36	6.53294845037e-05	5.74372261829e-05	1.57209935442e-05
37	6.21718216341e-05	5.77070655006e-05	1.74984928143e-05
38	6.10623725178e-05	5.92490044585e-05	1.83674924574e-05
39	6.20438082746e-05	5.53170601158e-05	1.67084931387e-05
40	6.1019701398e-05	5.7360129235e-05	1.71429929603e-05
41	6.30261621992e-05	6.22107504952e-05	1.78347642869e-05
42	6.02395698522e-05	5.98087910166e-05	1.65308074641e-05
43	6.15509074273e-05	6.00489869645e-05	1.82974586434e-05
44	5.97478182616e-05	6.0849640124e-05	1.89704686165e-05
45	6.00756526554e-05	5.89280725412e-05	1.68673124506e-05
46	5.79857083952e-05	6.11699013878e-05	1.74561961771e-05
47	6.17967832226e-05	6.05694115182e-05	1.76244486703e-05
48	6.07313214429e-05	6.04493135443e-05	1.88863423699e-05
49	5.90921494741e-05	6.06094441762e-05	1.65728705874e-05
50	6.16328660257e-05	5.76870601439e-05	1.72038174371e-05

Supplementary Table S4. Table of Parameters for Simulating Data with Realistic MNM (Part 1 of 2)

51	6.57879112884e-05	6.09082009503e-05	1.91108170597e-05
52	6.11257758428e-05	6.17346351559e-05	1.98371188957e-05
53	6.45360415855e-05	6.04536621373e-05	1.82029397647e-05
54	6.03487532685e-05	6.13214180531e-05	1.89746354654e-05
55	5.97875702982e-05	6.30569298848e-05	1.74766379287e-05
56	6.41906982191e-05	6.0412340427e-05	1.87476661417e-05
57	6.44497057439e-05	6.11974529223e-05	1.74766379287e-05
58	6.31115001993e-05	6.05776272681e-05	1.74766379287e-05
59	6.1082607922e-05	6.0412340427e-05	1.67957299574e-05
60	6.30683322785e-05	5.73132121561e-05	1.94285741129e-05
61	5.92263873279e-05	5.90074022775e-05	1.87476661417e-05
62	6.06940966349e-05	5.80983246514e-05	1.72042747402e-05
63	6.05214249517e-05	6.07842358195e-05	1.81121520352e-05
64	6.24208134666e-05	5.88421154364e-05	1.81121520352e-05
65	6.2334477625e-05	5.95445845111e-05	1.74312440639e-05
66	6.07804324765e-05	6.086687924e-05	1.78397888467e-05
67	6.32841718825e-05	6.09082009503e-05	1.62510035804e-05
68	5.97012344566e-05	6.20238871278e-05	1.74766379287e-05
69	5.9010547724e-05	6.01644101653e-05	1.68865176869e-05
70	6.01329136646e-05	5.83875766233e-05	1.90654231949e-05
71	6.22049738626e-05	5.91313674084e-05	1.85206968179e-05
72	5.97012344566e-05	6.06189489784e-05	1.63871851747e-05
73	5.95285627735e-05	5.84288983336e-05	1.72042747402e-05
74	6.14711192092e-05	5.99991233242e-05	1.74766379287e-05
75	6.11689437636e-05	5.68999950533e-05	1.87022722769e-05
76	5.99170740606e-05	6.19412437073e-05	1.72042747402e-05
77	6.11257758428e-05	5.77264292589e-05	1.69773054164e-05
78	5.88810439616e-05	5.87594720158e-05	1.89746354654e-05
79	6.20754701003e-05	5.66520647916e-05	1.72950624697e-05
80	6.04350891101e-05	5.80570029411e-05	1.65233667689e-05
81	6.16006229715e-05	5.73958555767e-05	1.83845152237e-05
82	5.90537156447e-05	5.97511930625e-05	1.65687606337e-05
83	6.28093247537e-05	5.92140108289e-05	1.77943949819e-05
84	6.16869588131e-05	5.63628128197e-05	1.70226992812e-05
85	5.6290968714e-05	5.83875766233e-05	1.72950624697e-05
86	5.72406629714e-05	6.02057318756e-05	1.77490011172e-05
87	6.01760815854e-05	6.0123088455e-05	1.81575458999e-05
88	6.12552796052e-05	5.98751581934e-05	1.61148219862e-05
89	6.06940966349e-05	5.87181503056e-05	1.72496686049e-05
90	5.97444023774e-05	5.76024641281e-05	1.72042747402e-05
91	6.11689437636e-05	5.75198207075e-05	1.77943949819e-05
92	5.78881817833e-05	5.69413167636e-05	1.64325790394e-05
93	6.03487532685e-05	5.63628128197e-05	1.67503360927e-05
94	5.63341366348e-05	5.78503943897e-05	1.57062772034e-05
95	6.1298447526e-05	5.67347082122e-05	1.58878526624e-05
96	6.03055853477e-05	5.95859062214e-05	1.65233667689e-05
97	5.84925326744e-05	5.70239601842e-05	1.81575458999e-05
98	5.96148986151e-05	5.7147925315e-05	1.57062772034e-05
99	5.93990590111e-05	5.71066036047e-05	1.66595483632e-05
100	6.23776455458e-05	5.73132121561e-05	1.55247017444e-05