Identity by Descent Genome Segmentation Based on Single Nucleotide Polymorphism Distributions

Thomas W. Blackwell, Eric Rouchka and David J. States

Institute for Biomedical Computing, Washington University in St. Louis, 700 South Euclid Ave., St. Louis, MO 63110, {blackwel, ecr, states}@ibc.wustl.edu

Abstract

In the course of our efforts to build extended regions of human genomic sequence by assembling individual BAC sequences, we have encountered several instances where a region of the genome has been sequenced independently using reagents derived from two different individuals. Comparing these sequences allows us to analyze the frequency and distribution of single nucleotide polymorphisms (SNPs) in the human genome. The observed transition/-transversion frequencies are consistent with a biological origin for the sequence discrepancies, and this suggests that the data produced by large sequencing centers are accurate enough to be used as the basis for SNP analysis. The observed distribution of single nucleotide polymorphisms in the human genome is not uniform.

An apparent duplication in the human genome extending over more than 130 kb between chromosomes 1p34 and 16p13 is reported. Independently derived sequences covering these regions are more than 99.9% identical, indicating that this duplication event must have occurred quite recently. FISH mapping results reported by the relevant laboratories indicate that the human population may be polymorphic for this duplication.

We present a population genetic theory for the expected distribution of SNPs and derive an algorithm for probabilistically segmenting genomic sequence into regions that are identical by descent (IBD) between two individuals based on this theory and the observed locations of polymorphisms. Based on these methods and a random mating model for the human population, estimates are made for the mutation rate in the human genome.

Introduction

High throughput genome sequence analysis has been independently performed more than once on several regions of the human genome. Comparing these independently derived sequences demonstrates that the distribution of SNPs in the human genome is highly non-uniform. This finding is striking because SNPs are thought to arise by a mutation process that occurs with approximately uniform rates across the genome.

In comparing two chromosomes, segments sharing a common ancestor are said to be identical by descent (IBD). The age of the last common ancestor for a segment is defined as the coalescence time for that segment. Genetic recombination results in the interchange of segments between homologous chromosomes, so a pair of contemporary chromosomes will be composed of a patchwork of IBD segments varying in age. The identity of the last common ancestor for each particular segment will vary as well.

SNPs arise as substitution mutations occurring along one or other lineage derived from the last common ancestor. Segments with a comparatively recent common ancestor will have had little time in which to accumulate mutations, so we expect that SNPs in these segments will be sparse. Conversely, segments with an ancient common ancestor will have had more time during which to accumulate mutations and we expect SNPs to be dense in these regions. The expected size of an IBD segment also depends on its age. For a segment with a recent common ancestor, few generations will have occurred during which recombination events could disrupt the region of identity by descent which it represents, so the segment is likely to be large. Loci with ancient common ancestors are likely to have had nearby recombination breakpoints and thus are likely to be found in relatively short IBD segments. As a result, the distribution of the number of SNPs per IBD segment is independent of the age of the segment. (Technically, two chromosomes which differ by a SNP are not strictly identical in that region. However, we will refer to corresponding segments of genomic sequence as 'IBD' because the majority of the sequence remains identical by descent.)

To obtain data on SNP distributions in the human genome, we need to compare sequence data derived independently from different individuals. Due to physical limitations of current sequencing and cloning techniques, the genome must be broken down into smaller portions in the range of 20 kilobases (kb) for plasmid clones to 250 kb for bacterial artificial chromosomes (BACs) and yeast artificial chromosomes (YACs) [Lodish et al., 1995]. As the sequence data for each of these shorter regions becomes available, it would be helpful to connect them with adjacent overlapping regions previously sequenced. It is possible that overlapping sequences could originate from different sequencing centers. Since a complete sequence of each human chromosome is desired, a method to assemble these smaller sequences into larger contiguous regions (contigs) is constructed.

Methods

GenBank is used as the reference database for human genomic DNA used in building the contigs. The results are based upon release 110.0, which includes sequences submitted to GenBank up until December 5, 1998 [Benson et al., 1998]. The GenBank primate division is used in order to create stable human contigs. In release 110.0, this is divided into gbpri1, gbpri2, and gbpri3. Table 1 shows a breakdown of the sequences in the primate divisions by sequence size.

Sequence length (nucleotides)	**Number of GenBank entries**
> 200,000	72
150,000-199,999	324
100,000-149,999	681
75,000-99,999	340
50,000-74,999	169
25,000 -49,999	1058
TOTAL > 25,000	2,644

Table 1: Length of primate GenBank entries. This table indicates the number of sequences in the primate divisions (gbpri1, gbpri2, and gbpri3) of GenBank release 110.0.

Some of the genome sequencing centers incorporate neighboring clone information into their GenBank entries. Table 2 shows some examples of how this data is entered into the comments section. Use of this information could help in the creation of genome contigs. However, as Table 2 indicates, this data is not standardized among the sequencing centers. The data is entered by hand in a manner that is easy for a human to read, but not easily parsed by a computer. The overlap between two clones, if given, is present only in a positional manner. An alignment between two overlapping clones is not given.

We create most of the contigs using an automated procedure. The first step is to retrieve human sequences from GenBank which are greater than 25 kb in length. After these sequences are retrieved their ends are searched against the primate division of GenBank for overlapping regions at least 70 base pairs (bp) long and at least 98% identical. These searches are performed using wu2blastn version 2.0 [Gish, 1994-1997], the Washington University version of BLAST [Altschul et al., 1990] with gaps for nucleic acid sequences.

When overlapping clones are found, they are merged together into a contig based on the blast alignment. Discrepancies in the alignment resulting from gaps and mismatches are marked by the character N in the contig. After a set of contigs has been assembled, they are compared against contigs found at the NCBI [http://www.ncbi.nlm.nih.gov/genome/seq/] and ORNL [http://compbio.ornl.gov/tools/channel/] web sites. Any differences are looked at in more detail. As a result of heuristics used in BLAST, in some cases the reported highest scoring pairs do not correspond to an optimal pairwise alignment. In some cases, the heuristic search and assembly restrictions need to be relaxed for automatic assembly to occur. Other contigs need to be assembled by hand in order to create the overlapping region. Since the volume of sequencing data is growing exponentially, these steps are largely automated using PERL scripts.

Sequencing center [sample GenBank accession number]	**Overlap information in COMMENT section**
Sanger Centre [Z99715]	The true right end of clone 1114G22 is at 104. The true left end of clone 262D12 is at 51983.
University of Washington Genome Sequencing Center [AC004398]	Overlapping Sequences: 5': UWGC: g1248a010 (Accession: AC004107) 3': UWGC: g1248a139
Whitehead Institute for Biomedical Research [AC005303]	Only 90.0 kilobases from the middle of this clone are being submitted. The remainder overlaps either accession AC003664 (WICGR project L281) or accession AC005277 (WICGR project L351).
Washington University Genome Sequencing Center [AC002378]	NEIGHBORING SEQUENCE INFORMATION: The clone being sequenced to the left is BK085E05; the clone being sequenced to the right is DJ102K02. Actual start of this clone is at base position 1 of DJ438O4.
Baylor College of Medicine [AC002523]	Begining of sequence overlaps with AF007262, end of sequence overlaps with AF011889. (*Note that Beginning is misspelled here*)

Table 2: Overlapping clone information. The right hand column contains examples of overlapping clone information from the COMMENT sections of the GenBank entries identified in the left column. The overlapping clone information is typical for the sequencing centers shown in the left column.

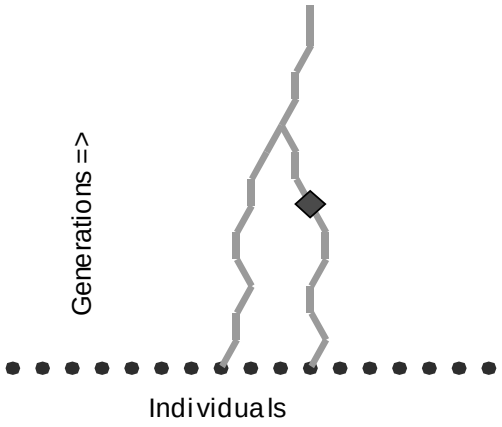
Difficulties

There are several difficulties with trying to find overlapping end segments. One problem is that clones may not overlap with 100% identity due to sequencing errors and polymorphisms. The PERL scripts are written in such a manner as to allow overlapping sequences greater than 98% identical. This allows the possibility that some overlaps might be missed. Most overlapping segments should be detected, however, since polymorphisms occur in the population at a rate of 7/10,000 [Taillon-Miller et al., 1998], and acceptable sequencing error rates are 1/10,000 [Collins et al., 1998].

Another difficulty is that the end of a sequence may contain repetitive elements. Prime examples of this are Alus and LINEs. In these cases, blast will produce multiple hits to otherwise unrelated sequences. It becomes hard to determine whether or not two sequences should be assembled into a contig when the overlap between them occurs in these repeat regions. Examples of such sequences are GenBank accession AC004021, AC004202, and AC004186.

The length of the overlap also varies greatly. Some sequencing centers such as Washington University Genome Sequencing Center (WUGSC) and Sanger Centre have a relatively constant sequence overlap length for known overlapping sequences. (In the case for WUGSC it is 200 bp; for Sanger Centre it is 100 bp.) For the assembled contigs, the size ranges from 0 base pair overlaps from the Japan Science and Technology Corporation efforts on chromosome 21 to a 82,766 base pair overlap between GenBank accession HS326L12 and HS232G24 from the Sanger Centre on chromosome X. Sequences with less than a 70 base pair overlap were hand assembled. The GenBank entries for these sequences have been used to aid in the detection and assembly of these contigs. For the shorter overlapping segments, running blast to find the alignment between two sequences takes a matter of seconds, but for larger regions, the time spent to find the alignment can take hours.

A theory for SNP distribution in the genome



Shown in Figure 1 is a schematic representation for population genetics. Each point along the horizontal line represents an individual chromosome from the current population. The jagged line traces the lineage of a particular locus through preceding generations. The diamond indicates a substitution mutation which results in a SNP in the current population.

We adopt a neutral Wright-Fisher model for population genetics. This consists of a random mating population of constant size with N_e diploid individuals and discrete, non-overlapping generations and no selection. Time into the past and distance along the chromosome are both measured as continuous variables, with units of generations and nucleotides respectively. (These two continuous approximations are known as a diffusion time scale and an infinite sites model.) Locations of recombination breakpoints and of point mutations are represented by two independent Poisson processes for each chromosome in each generation, with rates ρ and μ for recombination and mutation respectively. Each rate gives the expected number of events per nucleotide, per chromosome, per generation.

Consider $n = 2$ chromosomes chosen at random from the current day population. It is a standard result (reviewed in Tavare, 1984; Donnelly and Tavare, 1995) that at any single locus, the time T to the most recent common ancestor for these two chromosomes has an exponential distribution with mean $2N_e$ generations and density

$$f(T \mid N_e) = \frac{1}{2N_e} \exp\left(-\frac{T}{2N_e}\right).$$

We define the "IBD segment" containing a given locus to be a maximal region around the locus, within which no recombination has occurred in the lineage leading to either of the two sampled chromosomes, in any generation since their most recent common ancestor. This definition is specific to the pair of chromosomes in question. By construction, every site within this segment has the same coalescence time. If any single nucleotide polymorphisms are found between the two current day chromosomes within this segment, they must have arisen by a point mutation, in one lineage or the other, in some generation since the time of the most recent common ancestor.

Conditional on the coalescence time T, the length of an IBD segment has an exponential distribution with mean $1/2\rho T$. Marginal over T, this length L has density

$$f(L \mid N_e, \rho) = \left[\frac{2\rho}{(1+4N_e\rho L)} + \frac{4N_e\rho}{(1+4N_e\rho L)^2}\right] \exp(-2\rho L).$$

(This excludes recombinations which might occur during the current generation.)

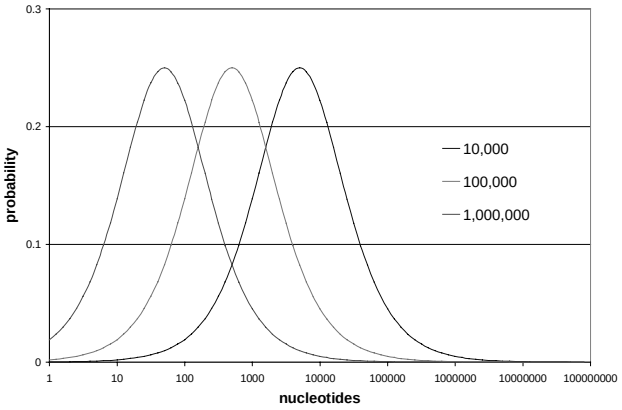
Marginal over the segment length, the number of SNPs occurring in an IBD segment has a geometric distribution with mean μ/ρ,

$$\Pr(N \mid \rho, \mu) = \left(\frac{\mu}{\rho+\mu}\right)^N \left(\frac{\rho}{\rho+\mu}\right),$$

and is independent of the segment's coalescence time. (This is a consequence, rather than an assumption of the model.) However, conditional on a segment's length, the number of SNPs found in it has a modified negative binomial distribution,

$$\Pr(N \mid L, N_e, \rho, \mu) = \frac{\Pr[X \le N+1]}{\Pr[Y \le 1]} \cdot \left(\frac{4N_e\mu L}{1+4N_e(\rho+\mu)L}\right)^N \left(\frac{1+4N_e\rho L}{1+4N_e(\rho+\mu)L}\right)^2 (N+1).$$

Here, X and Y are independent Poisson-distributed random variables with means $2(\rho+\mu)L+\frac{1}{2N_e}$ and $2\rho L+\frac{1}{2N_e}$ respectively. Their Poisson cumulative distributions simply give a convenient representation for values of the incomplete gamma function. The locations of the *N* observed SNPs are uniformly distributed within the boundaries of the segment, conditional on *L* and *N*.



Shown in Figure 2 is the probability distribution for the log of the IBD segment length for random mating populations of 1 million, 100,000 or 10,000 individuals. The mean IBD segment length does depend on the population size, and for all population sizes, the distribution of IBD segment lengths is extremely broad.

Given a region of the genome containing a number of SNPs, and some hypothetical segmentation *S* into intervals that are IBD, the likelihood for *S* is calculated as the product over all segments marginal over all possible segment coalescence times of the probability that an IBD segment of that age would have the hypothesized length *L* and number of SNPs *N* at the *N* observed locations. Thus the overall segmentation likelihood is just

$$L(S)=\prod_{segments}\int_0^\infty \frac{1}{L^N}\Pr(N\mid L,T)f(L\mid T)f(T)\,dT\,.$$

Since we do not have a way of knowing which of the many possible IBD segmentations describing a region corresponds to the true genetic history, we marginalize over all possible IBD segmentations for the region. This is accomplished using a dynamic programming approach [Lawrence and Reilly, 1985]. Given a set of all possible segmentations for a region of length *R*, to segment a region of length *R+1* we must either extend a terminal segment or start a new segment. The likelihood of each segment depends only on its length and SNP content so the problem is partitionable and dynamic programming can be applied. Since there are *R* possible terminal segments ranging in length from 1 to the full length of the region, the calculation requires linear storage and $O(R^2)$ time.

When *R* is large (~10^5), we introduce an approximation to further accelerate the calculation. Let *r* be the resolution of a segmentation. By this we mean that we will only consider segment boundaries placed at an integer multiple of *r* across the region. In this case, only *R/r* segment boundaries need be considered and for each of these, only *R/r* possible terminal segments need be considered. The calculation time is then $O((R/r)^2)$. Thus, by limiting the resolution to 10 nucleotides, a factor of 100 improvement in run time is achieved. In practice, no signficant difference in results was obtained for resolutions of 10, 20, 50 or 100 nucleotides.

Results

Two overlapping clones from different chromosomes

An interesting region occurs between two overlapping clones originating from two separate chromosomes. The first entry is GenBank accession AL021921 and the second entry is GenBank accession U95738. The 135 kb AL021921 is sequenced by Sanger Centre and is annotated as 1p36.13. The 171 kb entry U95738 is sequenced by The Institute for Genome Research (TIGR) and is annotated as 16p13.11. According to the blast hits, AL021921 lies completely within U95738 with 100 mismatches, 74 of which are transitions (A<->G; C<->T) and 26 are transversions. There are also 22 gaps composed of 123 indel events. The ratio of transitions to transversions is consistent with a biological origin for these sequence discrepancies. On average substitution mutation is expected to produce twice as many transitions as transversions [Li et al., 1996] while a random error process (such as sequencing error) would be expected to produce a two-fold excess of transversions. We observe three times as many transitions as transversions.

Interestingly, although these two sequences are 99.9 percent identical, they now appear to have been derived from different chromosomal locations. The Sanger center confirms the FISH localization of their clone on chromosome 1p. The source clone for the longer TIGR sequence was mapped by FISH at the California Institute of Technology. That data shows FISH signals on both chromosomes 1p34 and 16p13 (Figure 3). Together, these FISH data are consistent with the presence of a duplication event between chromosomes 1 and 16 occurring so recently that the human population may be polymorphic for this duplication. Other possible explanations include probe contamination or a chimeric BAC. There is no reason to suspect the former and we are not aware of any documented examples of chimeric BACs. Further, Pieter de Jong [personal communication] has surveyed over 400 BACs from the RPCI-11 library looking specifically for evidence of chimerism and found none.

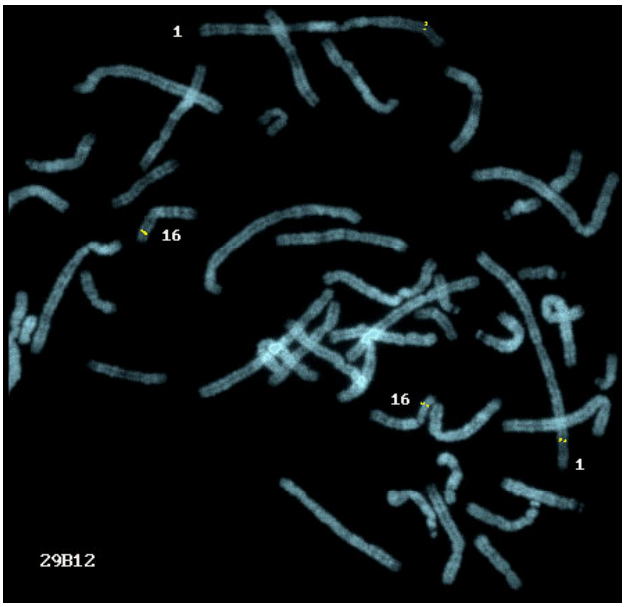
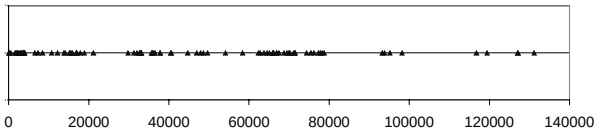


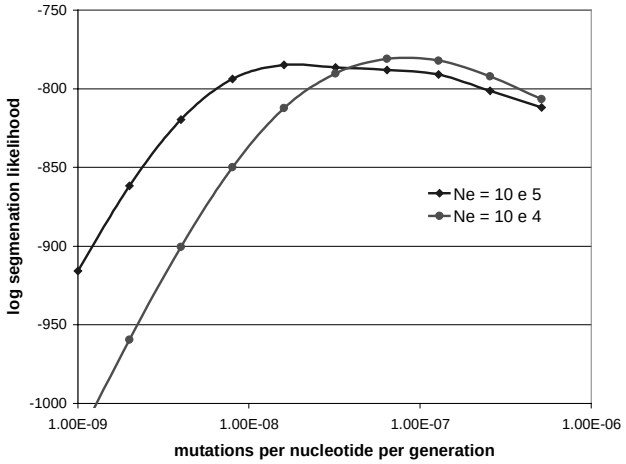
Figure 3: FISH results for GenBank accession U95738. This FISH image indicates a homology between chromosomes 1 and 16. (Image courtesy of California Institute of Technology) (http://www.tree.caltech.edu/pictures/fish-29B12.jpg).

The distribution of single base polymorphisms across this 134 Kb interval is shown in Figure 4. The distribution is highly nonuniform with some intervals spanning 10 Kb with no polymorphisms at all and other intervals of 100 nucleotides containing multiple SNPs.



Shown in Figure 4 is the distribution of SNPs along the overlap region between sequences U95738 and AL021921. Overall this overlap spans 134 kb with 100 sequence discrepancies (74 transitions and 26 transversions).

The dynamic programming algorithm described above was used to calculate IBD segmentation likelihoods for all possible segmentations at a variety of population sizes and single point substitution rates. In all cases a uniform population with random mating was assumed. For the purpose of this analysis, the recombination rate was fixed at one recombination per 10^8 nucleotides per generation (one centimorgan equals one megabase). While it is known that the relationship of genetic to physical distance varies across the genome and in some locations is even sex specific, these variations are relatively modest in comparison to the assumptions we have made about population size and mating behavior. Results are shown in Figure 5 below. Calculations were performed with a resolution (see above) of 10, 20, 50, and 100 nucleotides. No significant variations with resolution were observed.



Shown in Figure 5 is the calculated likelihood for all possible IBD segmentations of the U95738/AL021921 overlap as a function of point substitution mutation rates given a recombination rate of one per 10^8 per nucleotide per generation (one centimorgan equals one megabase) and a homogeneous population of 100,000 or 10,000 individuals. The best fit is obtained with a mutation rate moderately in excess of the recombination rate and a population size of roughly 10,000.

Discussion

In this paper we present experimental evidence for non-uniform distribution of SNPs across the human genome and derive theory for the expected distribution of SNPs in the genome. Interestingly, the observed non-uniform distribution can be accounted for by the interplay of two uniform and random processes, single point mutation and recombination. A striking finding is the very broad distribution of expected IBD segment lengths and ages. Another interesting result is that the random number of SNPs per segment is independent of the IBD segment's age.

Given the small number of SNPs that occur, it is not possible to determine which of the many possible IBD segmentations corresponds to the true genetic history of the interval being analyzed. Instead we apply a dynamic programming approach to calculate the likelihood marginalized over all possible segmentations of the region. Using this approach we obtain an approximate estimate for the point substitution rate in the human genome over the time to coalescence, roughly a few hundred thousand years. This estimate depends on knowledge of the recombination rate, which is well-established from pedigree genetics, and assumptions about population structure and history for the human species (reviewed in Jorde, Bamshad and Rogers, 1998). For assumed population sizes differing by an order of magnitude, the derived mutation rate differs by a factor of four.

The single point substitution mutation rate estimated here is in agreement with estimates derived through

pseudogene analysis. For example, Li et al. have examined point substitution rates for a number of pseudogenes identified in primate species. Comparing human to old world monkey species, they find substitution rates varying between 0.053 and 0.098 [Li et al., 1996]. These species are thought to have diverged roughly 25 million years or 1.25×10^{6} generations ago, assuming a generation time or 20 years. The corresponding point substitution rates are $4\text{-}8 \times 10^{-8}$ substitutions per nucleotide per generation.

Our estimate for the human point substitution mutation rate is independent of many of the assumptions made in traditional molecular clock calculations. First, we are calculating the mutation rate with respect to the recombination rate and therefore do not need to make assumptions about the average generation time for the species. Second, the time scale is inferred from the SNP distribution observed in the genome, and we do not need to make reference to the fossil record.

Our results suggest that SNPs found in isolation (no other SNPs in close proximity on the genome) are likely to have been derived from a relatively recent mutation event while SNPs found in clusters are more likely to have been derived from a comparatively ancient ancestor. If there has been stratification of the human population, the older SNPs are more likely to be present in all contemporary branches of the population and thus might be preferred for use as genetic or diagnostic markers.

Acknowledgments

We wish to thank Brendan Loftus of the TIGR Center and Andrew King of the Sanger Centre for assistance in reviewing clone origins and FISH data and the Caltech Genome Center for permission to reproduce Figure 3. This work was supported by the Department of Energy under grant DE-FG02-94ER61910 and by the National Institutes of Health under grant R01 HG01391.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403-410.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., (1998) GenBank. *Nucleic Acids Research*, 26(1):1-7.

Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L., (1998) New Goals for the U.S. Human Genome Project: 1998-2003. *Science*, 282(5389): 682-689.

Donnelly, P. and Tavare, S., (1995) Coalescents and Genealogical Structure Under Neutrality. *Annual Review of Genetics*, 29:401-421.

Gish, W., (1994-1997). unpublished.

Jorde, L.B., Bamshad, M. and Rogers, A.R., (1998) Using Mitochondrial and Nuclear DNA Markers to Reconstruct Human Evolution. *BioEssays*, 20(2):126-136.

Lawrence, C.E., Reilly, A.A. (1985) Maximum Likelihood Estimation of Subsequence Conservation. *Journal of Theoretical Biology*, 13(3):425-39.

Li, W.H., Ellsworth, D.L., Krushkal, J., Change, B.H.J., and Hewett-Emmett, D. (1996) Rates of Nucleotide Substitutions in Primates and Rodents and the Generation-Time Effect Hypothesis. *Molecular Phylogenetics and Evolution*, 5:182-7.

Lodish, H., Baltimore, D., Berk, A., Zipursky, S.L., Matsudaira, P., Darnell, J. (1995). *Molecular Cell Biology*. New York: Scientific American Books.

Taillon-Miller P., Gu Z., Li Q., Hillier L., Kwok P.Y., (1998) Overlapping Genomic Sequences: a Treasure Trove of Single-nucleotide Polymorphisms. *Genome Research*, 8(7):748-54.

Tavare, S. (1984) Line-of-descent and Genealogical Processes, and their Applications in Population Genetics Models. *Theoretical Population Biology*, 46:119-164.