Quantifying Ascertainment Bias and Determining Proxy Ancestral Alleles in Human Genome-Wide Polymorphic Data for Use in the Determination of Human Demographic History

Author: Damien Croteau-Chonka

Quantifying Ascertainment Bias and Determining Proxy Ancestral Alleles in Human Genome-Wide Polymorphic Data for Use in the Determination of Human Demographic History

Senior Honors Project by Damien Croteau-Chonka

Advisors: Dr. Gabor Marth, Dr. Eric Tsung

Boston College – Biology Department of

Submitted: May 4, 2007

Preface

My senior project for the Boston College Honors Program is a continuation of work done previously as a member of Dr. Gabor Marth's research lab in the Biology Department, which focuses on population genetics[1] and bioinformatics[2], particularly the role of single nucleotide polymorphisms (SNPs)[3] in genetic variation. I have worked closely under the direction of Dr. Eric Tsung, a post-doctoral fellow in the lab.

More directly, my work (in addition to Dr. Tsung's) is part of a lab effort to extend the work of Dr. Marth's 2004 *Genetics* paper "The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations" by applying its methods to new datasets (1). My contribution toward this end has been to create computer code (in Perl and Bash) to quantify ascertainment bias and determine proxy ancestral alleles in human genome-wide polymorphic data for Dr. Tsung's use in the determination of human demographic history.

The final results of my efforts will be part of a poster by Dr. Tsung (with myself as a second author) displayed at the 2007 Biology of Genomes Symposium at Cold Spring Harbor Laboratory in Cold Spring Harbor, New York. Our goal is to turn that poster into a paper (on which I will be an author) for submission for publication in a major scientific research periodical and which will also be available in the future at http://bioinformatics.bc.edu/marthlab/ascertainment-ancestral/.

Acknowledgments

I would like to take this opportunity to thank some people who have so helpful during my time here at Boston College: my family for their tremendous ongoing support and understanding of all kinds; Prof. Mark O'Connor of the Honors Program for being a taskmaster, cheerleader, and mentor; Dr. Gabor Marth for giving me the incredible opportunity to gain three years of hands-on bioinformatics research experience in his lab; and Dr. Eric Tsung for his day-to-day insights in our ongoing collaborative efforts.

Thanks also to Dr. Tsung for the contributed section "*Comparison of Models and Data in HapMap*" found in *Results* which help further illustrate and illuminate the concepts brought up in this thesis and to Dr. Arlene Wyman for her insightful comments on an early draft.

[1] Population genetics is the science of studying the frequency of genetic alleles that control certain traits in a group of more than one individual.
[2] Bioinformatics is the mathematical and computer science of organizing and analyzing large amounts of biological information.
[3] SNPs are sites in the human genome where individuals often differ in their DNA sequence by a single base.

Experimental Motivation

There are two competing models for the origin of genetic diversity in humanity: 1) the *Garden of Eden* model (GOE), which posits that humanity developed from a single population on the African continent and eventually settled the entire world only a hundred thousand years ago; 2) the *multiregional evolution* model (MRE), which posits that humanity has for the last two million years been spread across the world and the various regions of settlement account for known genetic diversity (2). Marth *et al.* 2004 supports the GOE model because its estimated population bottleneck and expansion times are consistent with the timing of the "Out of Africa" human migration event suggested by mitochondrial DNA (2). Other pieces of archeological and genetic evidence, however, cast doubt on the validity of the GOE theory (4).

Inferences about population demographic histories from SNPs are made by modeling special histograms known as allele frequency spectra (AFS) (see Figure 1 in *Results*). These spectra are a collection of allele frequencies for a given population. Allele frequency in a population of individuals refers to the frequency of an allele at a particular genetic locus for that population. Here, we are concerned with frequencies of bi-allelic polymorphisms. This type of polymorphism consists of sites in the chromosome where every individual person in a given population carries one of only two alleles, one of which is termed "ancestral" and the other "mutant" (5). The ancestral allele is the allele of the last common ancestor to that population (6). We base the study of our AFSs on the ancestral allele frequencies.

In the course of making inferences about human origins, it is impossible to know for certain the ancestral allele. One method that is agnostic to this issue is to use the frequency of the rarer form of the allele—referred to as the "minor allele"—as the basis for study (2). This is a fair assumption since more recent mutations tend to result in smaller allele frequencies. However, this is a condition-dependent definition of the minor allele frequency; what is the "minor" allele in one population may be the "major" allele in another.

Another method is to use a proxy for the ancestral human allele. Chimpanzees, apes, and now Neanderthal sequence data can all be used for this purpose (7, 8, 9). This approximation is dependent on the infinite-site model, which assumes that, given a sufficiently long genomic sequence and a low frequency of polymorphism, each genetic locus can be subject to only one mutation, thus simplifying analysis (10). The major caveat is that certain important sites of mutation may be shared by humans and other primates. Using a novel method based on the coalescent modeling method, our ultimate goal is—by quantifying the accuracy of other primate genomic datasets as a proxy for the human ancestral allele—to discover which SNPs are unique to humans. Currently, we are exploring the chimpanzee dataset.

Coalescent modeling can be used to predict the expected shape of an AFS under specific demographic model structures and parameters (5). Coalescence refers to the idea that, looking backwards in time, genetic alleles merge together at points of common ancestry (10). Along this principle, a "coalescent tree" structure can be constructed to estimate the genetic descent of a set of alleles (much like a "family tree" shows familial genealogy). It is important to note that these trees are only guesses at such relationships. The coalescent model used in Marth *et al.* 2004 assumes that genes do not undergo recombination and represents genetic drift using a stochastic Markov-chain Monte Carlo technique.

Fitting the resulting coalescent models to the curve shapes observed in experimentally collected allele frequency data—our AFSs—gives numerical parameters. Those models (and corresponding numerical parameters) that would produce similar shapes to the observed data are identified; the mathematically best-fitting model can then be considered the best estimate for the demographic history that actually produced the AFS data (5).

Multiple figures in Marth *et al.* 2004 demonstrate visually that the specific shape of the AFS for a given population is profoundly influenced by its long-term demography. For example, a recent population expansion increases the number of rare alleles; conversely, a recent population collapse decreases the number of rare alleles and increases that of common alleles (5).

It must be emphasized that while there are genetic differences among human populations, all humans are notably very similar to each other. The estimated effective population size of humans—which is a representation of genetic diversity (10, 11)—is approximately 10,000 (12, 13). The effective population sizes of other primate species are generally larger in comparison: bonobos are estimated at 12,400 and chimpanzees are at 20,900 (14).

Another important consideration in inferring demographic history is that the shape of an AFS is also greatly influenced by the nature of the SNPs that are chosen to be included in it. Furthermore, the manner in which SNPs are ascertained biases their discovery and thus their inclusion. A large variety of discovery protocols (ascertainment schemes) have been used to create publicly available SNP data (15). SNPs found in the public database dbSNP—made available by the National Center for Biotechnology Information (NCBI)—were found to be biased toward more common alleles in the ten Encyclopedia of DNA Elements (ENCODE) regions (16). The two-stage method of SNP discovery is another example of this: the first stage involves sequencing the entire genome of just a few individuals; the second stage involves genotyping in many individuals specific locations identified during the sequencing stage (15). Sequencing records every nucleic acid within a region; there is no ascertainment bias because all positions of that region are determined and all SNPs discovered have been randomly ascertained. Genotyping records information about a specific site of a genome or region; this

method does have ascertainment bias because only specific chosen positions are determined.

The effect of "ascertainment bias" has been theorized or documented on a case by case basis, but Nielsen 2004 suggests that few attempts have been made in the scientific literature to quantify this bias on different types of sequence or genotype data.[4] Bias can lead from over- or under-representation of rare alleles ultimately to misleading conclusions about demographic history (15). A polymorphism is termed "rare" if the mutant allele is infrequently observed in a sample, and it is "common" if both ancestral and mutant alleles are observed in similar frequency in a sample (5). Understanding the nature of how a SNP was discovered (its ascertainment scheme) allows the researcher to adjust for that bias.

Methods developed by Marth, Tsung, and Croteau-Chonka based on Nielsen and Signorovitch 2003 attempt to perform this task of quantitatively characterizing ascertainment bias on actual sequence data as well as to develop further methods that will be able to better describe this bias (17). Using Perl programs written by Croteau-Chonka, two datasets were gathered and processed into a desired format for the aforementioned purpose of SNP ascertainment bias study. Technical descriptions of the programs follow in the *Methods* section.

The first, a genotyping-derived dataset, comes from the International HapMap Project (HapMap). The purpose of the HapMap Project is to develop a haplotype[5] map of the human genome (18). The 270 individuals sampled in HapMap comprised four human populations: U.S. residents from Utah with northern and western European ancestry (CEU), Chinese from Beijing, China (CHB), Japanese from Tokyo, Japan (JPT), and Yorubans from Ibadan, Nigeria (YRI) (18). We used the CEU population as a proxy for Europeans in general, the CHB and JPT together as a proxy for East Asians, and the YRI as a proxy for Africans (18). The ascertainment conditions for SNP discovery were derived from Venter *et al.* 2001 (19) and provided by Jim Mullikin at the National Institutes of Health.

The second dataset, a sequencing-derived dataset, comes from the ENCODE consortium—which was launched by the National Human Genome Research Institute (NHGRI) in order "to carry out a project to identify all functional elements in the human genome sequence" (16). The UCSC Genome Bioinformatics Group "manages the official repository of sequence-related data for the ENCODE consortium and supports the coordination of data submission, storage, retrieval, and visualization" (20). The group also provides a complete human-chimp genome alignment data relevant to this project. Each ENCODE region was sequenced in approximately 1 kb overlapping segments called amplicons (18). This overlap allows for the discovery of SNPs inside sequencing

[4] The article specifically states that "[w]ith the exception of population growth parameters, the effect of the ascertainment bias on inferences regarding demographic parameters has not been extensively analysed in the literature."
[5] A haplotype is a set of associated SNP alleles in a region of a chromosome.

primers. Quinlan and Marth 2007 discovered that many of these primers contained SNPs, which in turn introduces another bias into the data (21).

These are the questions that we hope to answer with the two datasets studied for this senior project: With different ascertainment conditions, how does HapMap genotype data skew in allele frequency spectra? What ascertainment bias is there when using only SNPs that are located in regions aligned between the human and chimp genomes? How well does the chimp genome work as a proxy for human ancestral alleles?

Results

Determination of Allele Frequency Spectra from International HapMap Project Data

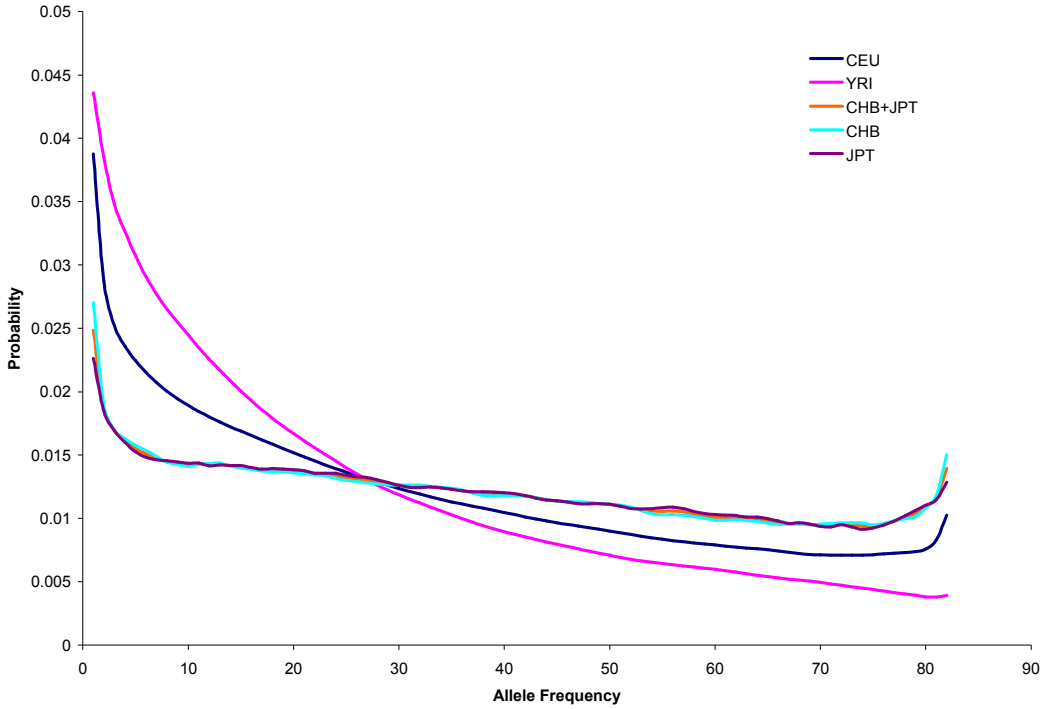

Figure 1 – Allele frequency spectra (unfolded, reduced to an *m* value of 83 chromosomes, and normalized) for four world population sets (plus CHB+JPT) from HapMap. This figure represents a combination of all ascertainment conditions.

Composite Allele Frequency Spectra

Figure 1 shows that the allele frequency spectrum is dependent on population. Each plot represents SNP allele frequencies from all human chromosomes aggregated from the four HapMap populations (CEU, CHB and JPT combined, CHB and JPT separated, and YRI, respectively) regardless of ascertainment conditions (18). Because of the closeness of the CHB and JPT allele frequencies, further analysis will combine these two populations together (known henceforth as CHB+JPT). The unfolded spectrum was constructed by using the chimp allele as a proxy for the human ancestral allele. Information on the chimp ancestral allele received from Jim Mullikin.[6]

The lower prevalences of rare SNPs in the European (CEU) and in the combined Chinese and Japanese (CHB+JPT) populations as compared to the Yorubans (YRI) are indicative of population bottlenecks (1). In addition, the Chinese and

[6] ftp://kronos.nhgri.nih.gov/pub/outgoing/mullikin/SNPs/SNPdiscoveryInfo.b121.tar

Japanese populations show an even narrower bottleneck than the European population.

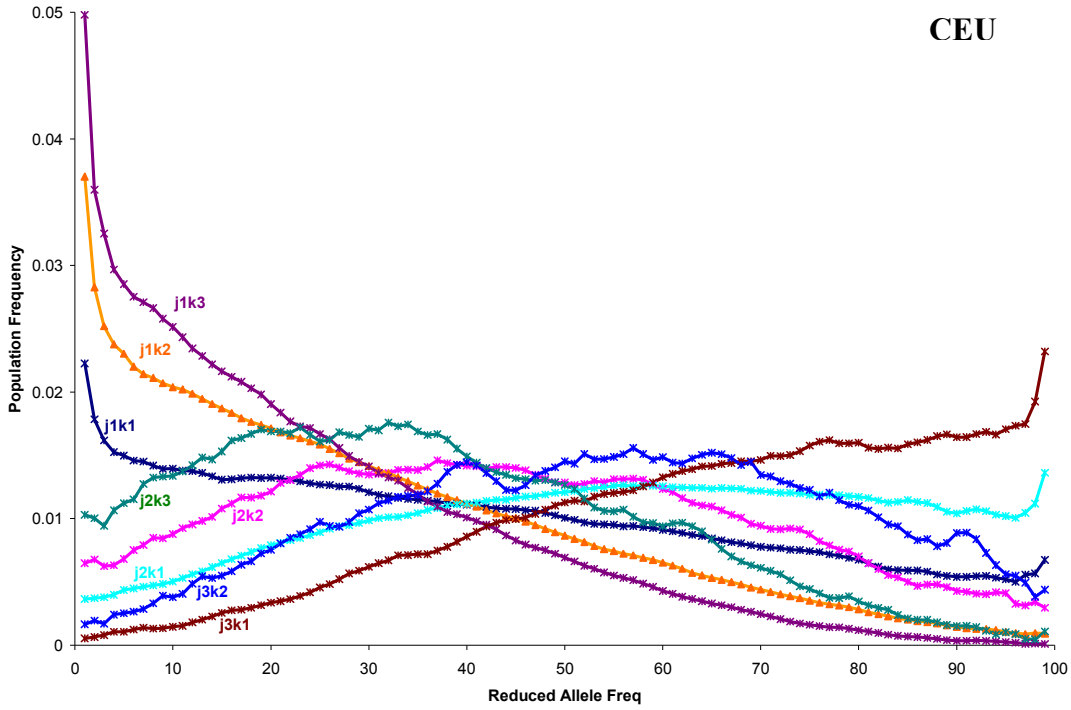

Figure 2 – Allele frequency spectra (reduced to m = 100 and normalized) for the CEU HapMap population where each plot represents a different ascertainment condition. These conditions (1-1, 1-2, 2-1, 2-2, 3-1, 1-3, 3-2, and 2-3) are represented by the two parameters j and k, which are counts of alleles, non-ancestral and ancestral, respectively, in the HapMap SNP discovery set.

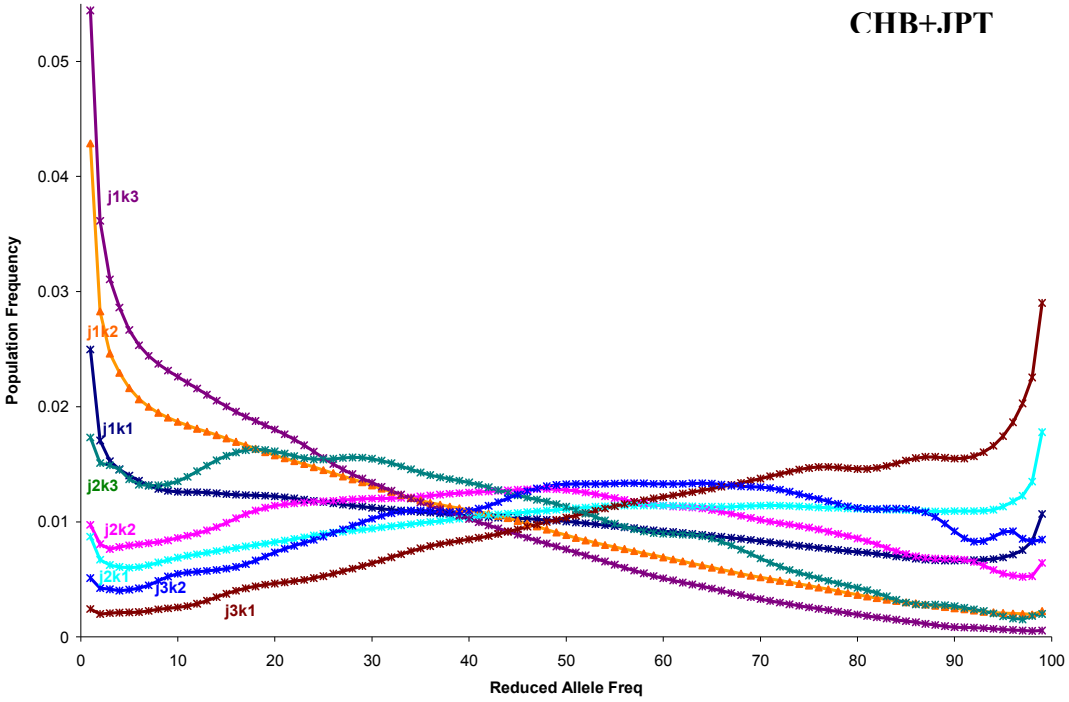

Figure 3 – Allele frequency spectra for the CHB+JPT HapMap combined population (same as Figure 2).

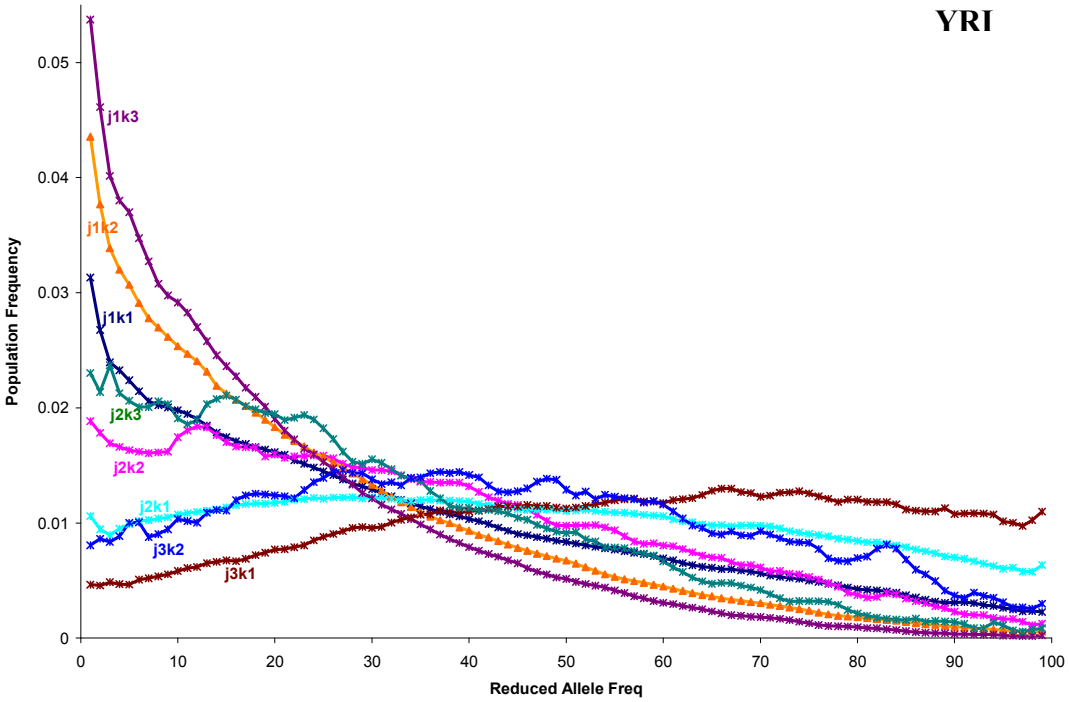

Figure 4 – Allele frequency spectra for the YRI HapMap population (same as Figure 2).

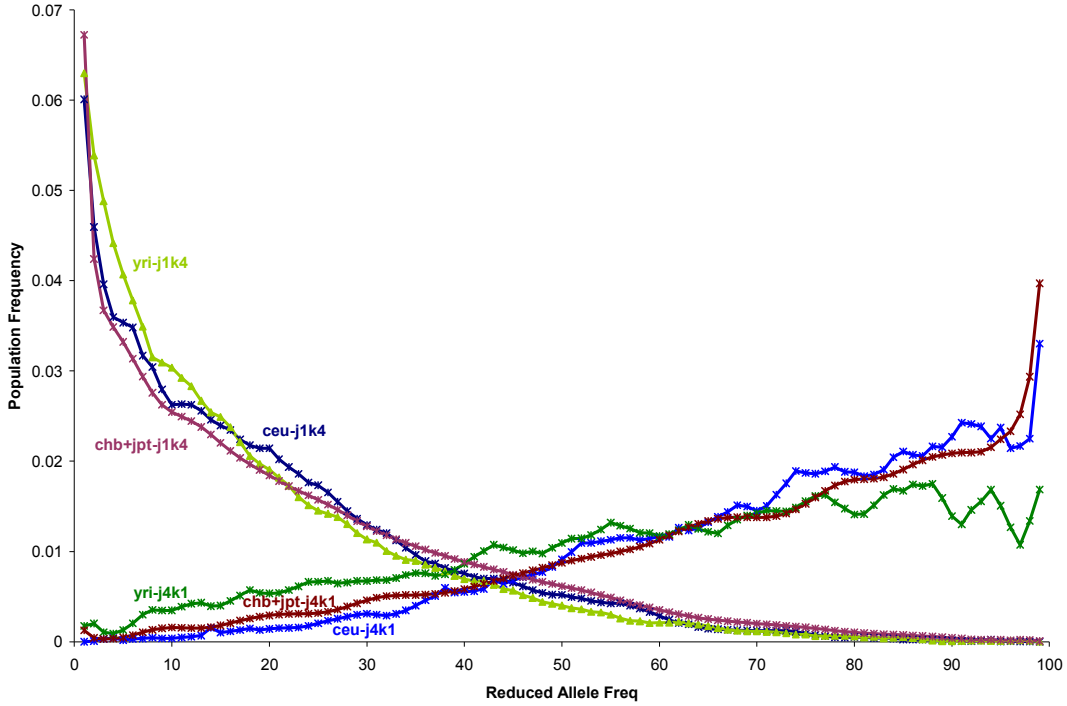

Figure 5 – Allele frequency spectra (reduced to m = 100 and normalized) for all HapMap populations (CEU, CHB+JPT, YRI) under ascertainment conditions j and k are 1-4 and 4-1, respectively.

Effects of Ascertainment Bias on HapMap AFS

Figure 2, Figure 3, and Figure 4 (CEU, CHB+JPT, and YRI, respectively) show that SNP ascertainment conditions have a significant effect on the shapes of allele frequency spectra. For two ascertainment conditions, 1-4 and 4-1, Figure 5 shows a direct comparison of allele frequency spectra across these same HapMap populations. In all, these groupings of spectra represent SNP allele frequencies from autosomal chromosomes aggregated under varying ascertainment conditions as labeled (18). The sex chromosomes were excluded from this dataset because the subsequent modeling is valid only for diploid chromosomes. Again, the unfolded spectrum was constructed by using the chimp allele as a proxy.

The discovery count parameters (j and k) used to filter SNPs by their ascertainment conditions during this collection process were the allele frequencies of the non-ancestral and ancestral alleles derived from the allele frequencies of five human genomic subjects in the Venter sequencing effort: two males and three females—one African-American, one Asian-Chinese, one Hispanic-Mexican, and two Caucasians (19). The maximum sum of j and k was 5 and many ascertainment conditions (e.g., $j = 1$, $k = 2$) totaled less than this maximum. Table 1 shows the total counts of SNPs for each ascertainment condition, which are consistent across the HapMap populations.

For any given SNP in these spectra, at most, only one of the two alleles from each individual subject was recorded in the data we used. Also, not all of the individuals were sequenced at all positions in the genome. One possible explanation for this incompleteness is the nature of coverage in the Venter sequencing effort—for the goal was not to fully genotype each individual separately but to obtain a single complete composite reference sequence (19).

Table 1 – Total counts of SNPs in each population under various ascertainment conditions. "Any 0" refers to any ascertainment condition where either *j* or *k* is 0.

	Total count of SNPs		
Ascertainment conditions (*j* and *k*)	**CEU**	**CHB+JPT**	**YRI**
Any 0	1,097,800	1,032,504	1,097,938
1-1	177,904	171,996	178,880
1-2	202,807	191,410	211,329
1-3	59,486	55,759	62,821
1-4	8,496	7,920	9,002
2-1	122,556	119,551	122,261
2-2	19,206	18,937	18,462
2-3	6,598	6,464	6,325
3-1	26,276	25,569	26,888
3-2	4,718	4,672	10,773
4-1	2,919	2,831	3,069

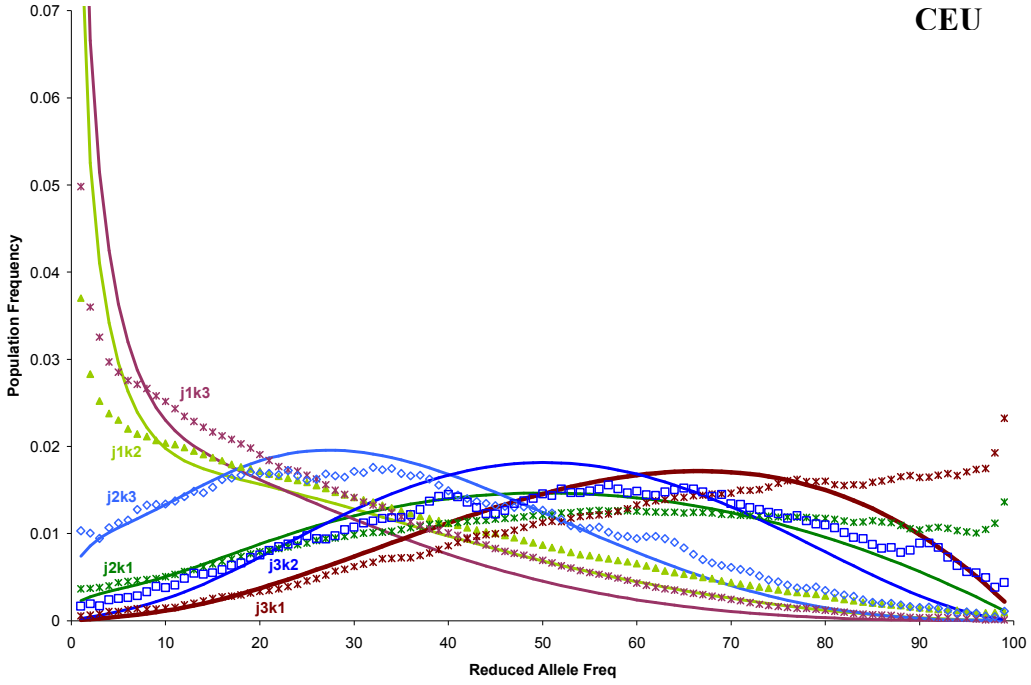

Figure 6 – Comparing ascertainment-adjusted coalescent models to allele frequency spectra (unfolded, reduced to m = 100, and normalized) from CEU HapMap population under various ascertainment conditions (1-2, 2-1, 2-3, 3-2, 1-3, and 3-1). The models are generated under the assumption that the chimpanzee allele is a perfect ancestral proxy.

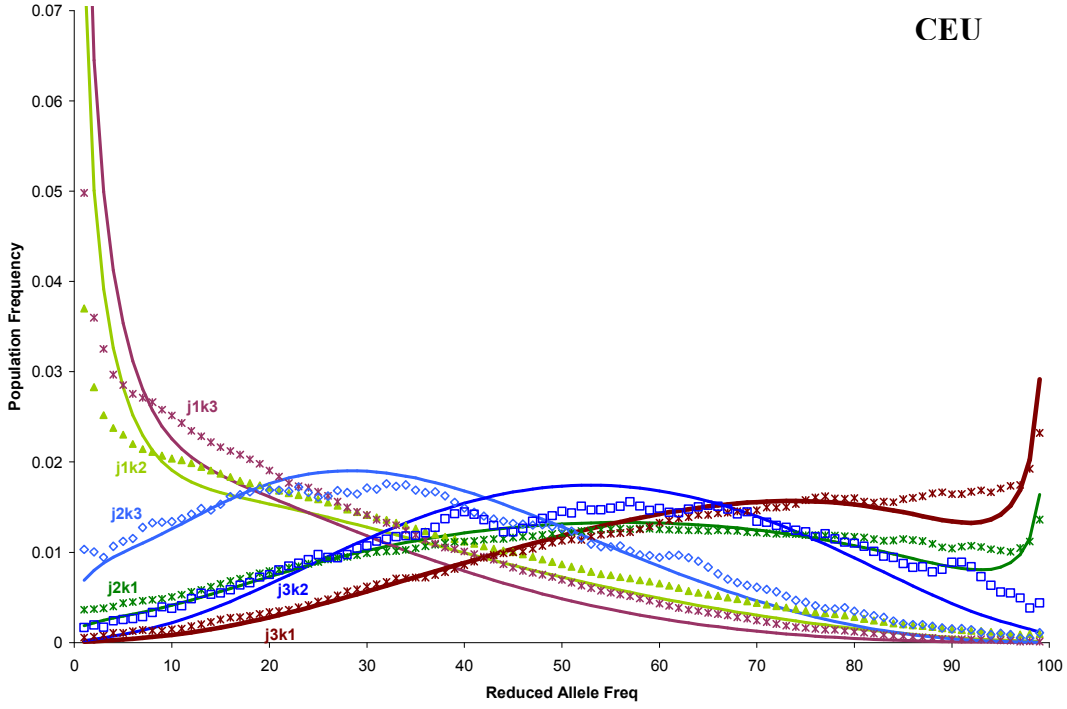

Figure 7 – Comparing ascertainment-adjusted coalescent models to allele frequency spectra from CEU HapMap population (same as Figure 6). The models are generated under the assumption that 90% of ancestral alleles are correct.

Comparison of Models and Data in HapMap

With the coalescent model, employing population demographic parameters from Marth *et al.* 2004 (1) and adjusting for ascertainment bias, we derived multiple AFS that can be compared to the HapMap data in previous figures. The bias adjustment methods presented here were adapted by Tsung and Marth from Marth (1) and Nielsen (15). Previously, Marth had developed Perl code for adjusting only for the ascertainment conditions $j = 1$ and $k = 1$.

Comparing models to actual data (Figure 6, Figure 8A), we see a consistent downward deviation at high allele frequencies. One possible explanation of this pattern is that the chimp allele is an imperfect proxy for the human ancestral allele. In making the assumption that only 90% of the chimp alleles are indeed ancestral, this deviation is diminished in the model (Figure 7).

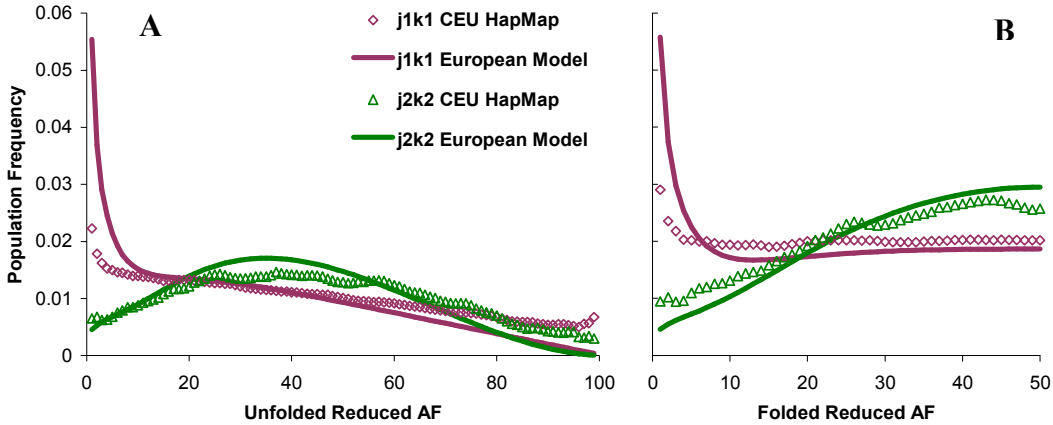

Figure 8 – Comparing coalescent models to unfolded (A) and folded (B) allele frequency spectra (reduced to m = 100 and normalized) from CEU HapMap population under ascertainment conditions 1-1 and 2-2.

When ascertainment conditions are equal to each other ($j = k$) and the spectrum is folded, any deviation due to the imprecision of the proxy ancestral allele should be eliminated. However, Figure 8B still shows systematic deviation from the model.

One potential cause is that in a great majority of cases, not all five individuals were sequenced for each SNP in the Venter sequencing effort (see Table 1). Our models for ascertainment assume that the SNPs all come from the same set of individuals; however, the data itself could be any subset of the five individuals sequenced. Another potential cause is that the Venter sequencing effort was not the only mechanism used to determine HapMap SNPs, thus the model assumption of a two-stage method of SNP discovery is not entirely accurate. The discrepancy between the model and this data suggests that further study is warranted with a more well-defined dataset. We believe that the data from the ENCODE sequencing effort may suit this purpose better.

Determination of Chimp Ancestral Allele and Mapping of HapMap Genotype Frequencies for ENCODE Regions SNP Data

Initial SNP Data Statistics from Determining Chimp Ancestral Allele

There were 27 SNP positions from a contiguous stretch of genome in the ENr113 ENCODE region (chromosome 4, positions 119,116,529-119,119,785) where the chimp ancestral allele is unknown. There is no coverage in the UCSC Genome Browser's "Chimp Reciprocal Net" track dataset (rBestNetPanTro1) for this small genomic region and thus no coverage in the "Simple Differences [Between Human and Chimpanzee] in Regions of High Quality Sequence" track dataset (chimpSimpleDiff). A high-coverage alignment such as rBestNetPanTro1 enables the use of the chimp as a proxy for human ancestral alleles.

Where there is coverage in rBestNetPanTro1 and the chimp ancestral allele is known, out of a 15,498 total ENCODE SNPs (received from Quinlan and Marth), 2662 SNPs were present in the chimpSimpleDiff dataset. Thus, at those positions, the chimp ancestral allele differed from the human ancestral allele.

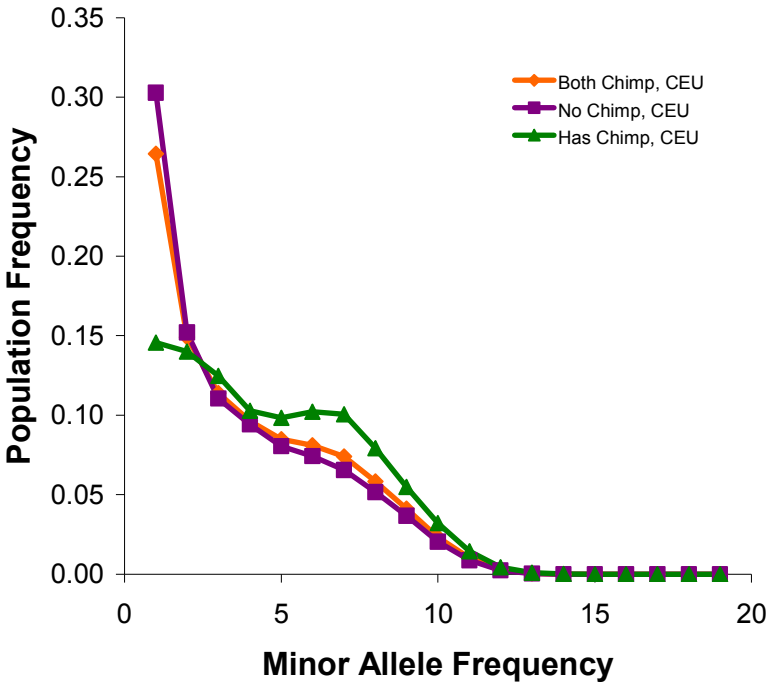

Figure 9 – Minor allele frequency spectra (reduced to m = 20 and normalized) using ENCODE SNPs from CEU population. These spectra represent all SNPs discovered in the five ENCODE regions under study: ENm013, ENm014, ENr112, ENr113, and ENr131. "No Chimp" refers to SNPs located in regions where there was no human-chimp genomic alignment; "Has Chimp" refers to SNPs located in regions where there was human-chimp genomic alignment; and "Both Chimp" refers to these two aforementioned datasets combined together.

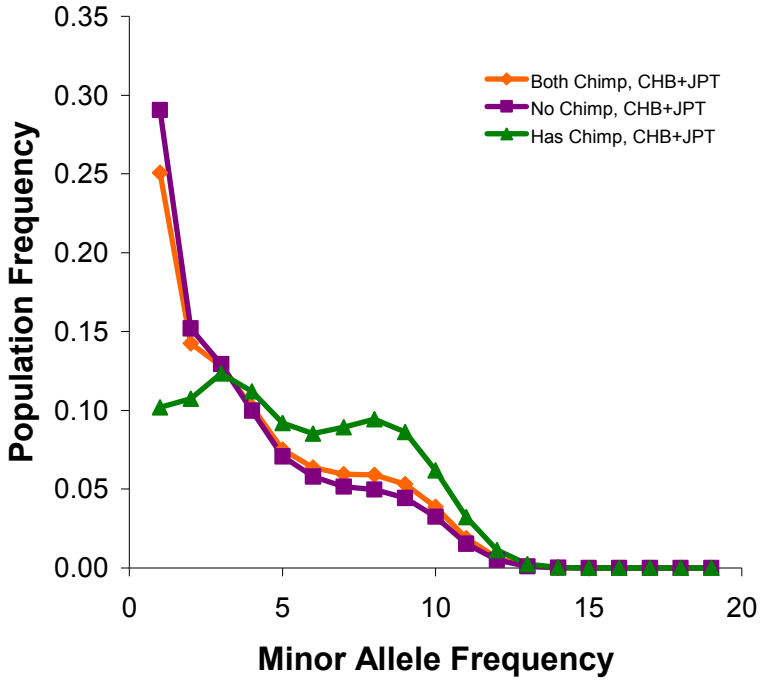

Figure 10 – Minor allele frequency spectra (reduced to m = 20 and normalized) using ENCODE SNPs from CHB+JPT population (same as Figure 9).

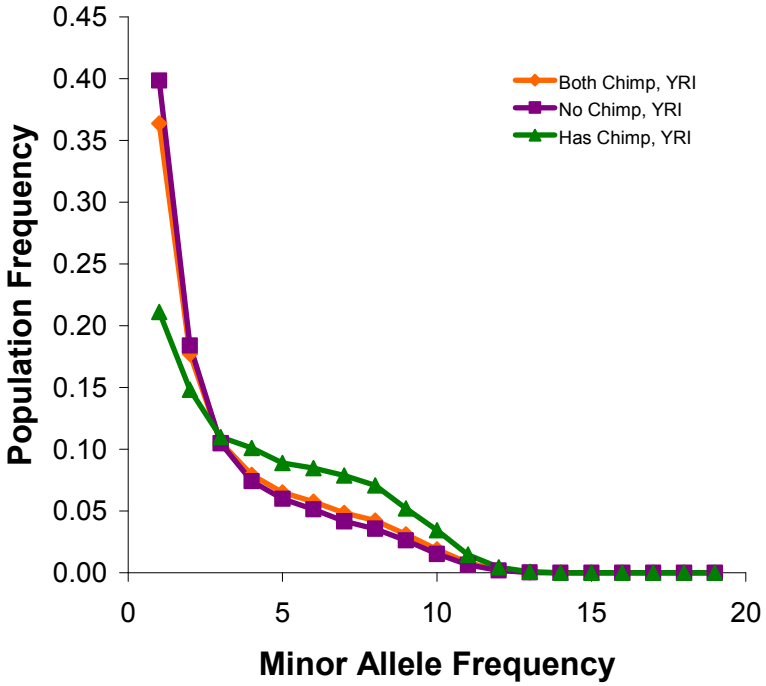

Figure 11 – Minor allele frequency spectra (reduced to m = 20 and normalized) using ENCODE SNPs from YRI population (same as Figure 9).

Comparison of Allele Frequencies between SNPs in Aligned and Non-Aligned Genomic Regions

Eliminating the SNP primer bias effect (21), Figure 9, Figure 10, and Figure 11 show that for those portions of the ENCODE regions (ENm014, ENr112, ENr113, and ENr131) in which the human and chimp genomes align there is a significant reduction in the number of SNPs at lower minor allele frequencies.

Because the ENCODE regions were designed to represent the entire human genome, we believe that this effect would be observed genome-wide. Thus, this bias will have consequences for full-spectrum demographic modeling when using the chimp genome as a proxy for the human ancestral allele.

Methods

Determination of Allele Frequency Spectra from International HapMap Project Data

Data Gathering Overview

This set of programs written in Perl and Bash script query local copies of databases downloaded from the International HapMap Project and generates AFSs based on parameters like population (European, African, and Asian) and discovery count numbers.

The main program that drives the data gathering process is `afsHapMapRun05.sh` (where the number "05" can be replaced with the particular run number in the name of the file and a variable inside it which is used to help name other files).

`afsHapMapRun05.sh` makes a separate directory for each population and then copies all of the Bash scripts, Perl scripts, and configuration files into each one. (The purpose of this is so that the files that were used to generate data are kept together with that data; the originals are available for subsequent tweaking.) The program then steps into each resulting directory and runs `afsRun00-pop.sh` where it is renamed according to the current run number and population.

`afsRun00-pop.sh` does most of the heavy lifting. First, it runs the main Perl program `afs-db12.pl` (where "12" is the current version) with various pairs of ancestral and non-ancestral discovery count numbers as parameters. This produces an allele frequency spectrum (AFS) and a discovery count info file for each parameter set. (A description of the algorithm behind this program is in the next section.) `afsRun00-pop.sh` then creates two new files, one combines all of the AFSs and the other combines all of the discovery count info files. The AFSs are added together using a special program called spectrumAdd.pl while the discovery count info files are simply concatenated together. A Perl program called stripAFS.pl removes the list of SNPs from the end of each line in the `allancestbias.txt` file.

The individual AFSs are now processed by a Bash script called `processAFS.sh`. It produces two resulting files. The first one comes from each individual AFS being run through `stripAFS.pl`, `spectrumFold.pl`, `spectrumReduce.pl`, and then `curveNormalize.pl`. The second one goes through the same sequence of Perl scripts except for `spectrumFold.pl`. These sets of folded and unfolded reduced and normalized AFSs are then graphed into .PNG files by `plotAFS.sh`.

In the future, it will be at this point that models will be generated using `modelAFS.sh` to compare to the AFSs created from HapMap data by using `curveFit.pl`. Currently, this functionality is not in place.

Now `afsRun00-pop.sh` does some check up and clean up of its work. It does some renaming to make all of the population acronyms in lower-case and to ensure that the graph files end in the correct extension. Then `check.sh` is run on all of the individual AFSs and discovery count info files to help validate that data by calling `dbcheck.pl`. Finally, files are organized and moved into named directories. The resulting tree in each population folder is as such:

`biasinfo/`	Contains the discovery count info files
`graphs/`	Contains the .PNG graphs of the processed AFSs
`logs/`	Contains all of the logs from `afs-db12.pl` and `dbcheck.pl`
`normdata/`	Contains all of the processed AFSs
`composite/`	Contains composite data files like `allancestbias.txt` and `allafs.txt`
`rawdata/`	Contains all of the unprocessed AFSs.
`scripts/`	Contains all of the Bash scripts, Perl scripts, and configuration files

How afs-db12.pl Works

Inside `afsRun00-pop.sh`, the following code calls `afs-db12.pl`:

```
perl ./afs-db12.pl --ascCol HuAA,HuCC,HuDD,HuFF
      --logSnpID --ancestralTableName LocusInfo.dbsnp_mullikin
      --outputLog --checkFreqRsId --discNonAnces $1 --discAnces
      $2
      --AncesBiasInfoToggleName AncesBiasInfo > afs-db11c-
      run$run-pop-j$1k$2.txt
```

`afs-db12.pl` can take in a number of parameters but the ones in the above code segment are most important and are the ones used in the actual runs. The `ascCol` parameters are the columns in the HapMap database that make up the discovery set. The parameter `logSnpID` toggles the inclusion of information on the ancestral nature of each SNP into the program's logfile. The parameter `ancestralTableName` provides the name of the HapMap database and table containing the ancestral information for each SNP under consideration. The parameter `outputLog` toggles the output of a log of SNPs included in the resulting AFS files. The parameter `checkFreqRsId` toggles the inclusion of a SNP list for every particular line in the resulting AFS files. The parameters `discNonAnces` and `discAnces` cause the program to only consider SNPs whose total discovery counts equal those numbers (either REF and VAR or VAR and REF, respectively). The parameter `AncesBiasInfoToggleName` toggles the output of a discovery count file associated with each AFS and uses the parameter string "`AncesBiasInfo`" as the root of the discovery count file's name. The standard output of `afs-db12.pl` is an AFS and it is piped to a file that indicates the program data run, population defined in `db.cfg`, and the discovery count parameters.

First, `afs-db12.pl` constructs a primary MySQL based on the supplied parameters. The exact query is recorded in the main logfile. An example:

```
SELECT rsId AS snp_id, totalcount*2, refhom_count*2+het_count,
otherhom_count*2+het_count, allele1, allele2,
IF((M.HuAA_ref+M.HuCC_ref+M.HuDD_ref+M.HuFF_ref+1) = 0 AND
(M.HuAA_var+M.HuCC_var+M.HuDD_var+M.HuFF_var) = 5, 1, 0) AS
refAncesAssump
FROM Hapmap_v1p0.Hapmap_2005_10_snp_info H,
LocusInfo.dbsnp_mullikin M
WHERE (((M.HuAA_ref+M.HuCC_ref+M.HuDD_ref+M.HuFF_ref+1) = 5 AND
M.HuAA_var+M.HuCC_var+M.HuDD_var+M.HuFF_var = 0) OR
(M.HuAA_var+M.HuCC_var+M.HuDD_var+M.HuFF_var = 5 AND
(M.HuAA_ref+M.HuCC_ref+M.HuDD_ref+M.HuFF_ref+1) = 0)) AND
H.rsId = M.snp_id AND panelLSID IN
('urn:lsid:dcc.hapmap.org:Panel:CEPH-30-trios:1') AND
refhom_count*2+het_count!=0 AND otherhom_count*2+het_count!=0
```

The query is asking for a SNP's ID, its total number of chromosomes, its number of "reference" chromosomes, its number of "other" chromosomes, the allele letters associated with each of the previous two numbers, and a Boolean representing whether or not the reference allele can truly be assume to be the ancestral one. This is determined by two conditions: whether or not the total "reference" numbers from the HapMap human discovery set plus one equal the parameter `discAnces` and whether or not the total "other" numbers from the same discovery set equal the parameter `discNonAnces`. The particular discovery set to be used is defined as column names in the command line parameters and brought together by string parsing and manipulation in `afs-db12.pl`. The resulting Boolean having a value of one means that the "reference" allele is assumed to be ancestral based on the discovery count information; the logic returning a zero means that the "other" allele is assumed to be ancestral. In the example above, `discNonAnces` is equal to five and `discAnces` is equal to zero. Since discovery counts are always positive and their total plus one can therefore never equal zero in the case of the "reference" set, the condition fails and the "reference" allele of the SNP is not assumed to be the ancestral one.

The query looks for information in databases and tables specified in `db.cfg` and through the parameter `ancestralTableName`. The query is limited to SNPs whose total discovery counts (REF and VAR) equal are in the set of `discAnces` and `discNonAnces` or `discNonAnces` and `discAnces` and to a population specified in `db.cfg`. Finally, the query uses some simple logic to eliminate any "marginal" SNPs, that is, any SNPs whose number of either "reference" or "other" chromosomes is equal to zero.

After the MySQL query has done a first pass of filtering SNPs for consideration, the "AFS generation" section of `afs-db12.pl` sets to work on a second, more specific filtering of the primary query results. The program steps through this list and individually examines each SNP. The first if-then statement hinges on whether the parameter `ancestralTableName` is defined, which is what is

currently used. Untested is the program's ability to otherwise use the minor allele frequency; the program tries to choose the non-ancestral allele for SNPs where that information can be determined from CHIMP data. If ancestral information is missing for either allele or both alleles "claim" to be ancestral, the program will only include the SNP if the toggle `useMinorWhenMissAnces` is on.

Otherwise, the two choices are that either the "reference" allele is ancestral or the "other" allele is ancestral. For instance, even if the "reference" allele is supposed to be ancestral, before the SNP is included, the value `refAncesAssump` of must also agree with this assumption. If the two pieces of information disagree about whether or not the allele is ancestral, the SNP is not included in the AFS. Once a SNP passes this test, there is some additional logic in the code that keeps track of SNP lists for the AFS and puts together a hash table of information for the discovery count file; whether or not these actions take place are contingent on command line toggle parameters initially fed to the program. A side note: the logfile code uses the same logic to ensure that SNPs included in the resulting AFS are correctly logged.

Data Validation

After the data is generated, it is important for it to be validated. This is accomplished by the `dbcheck.pl` program, which can validate both AFSs and discovery count files (command line parameters `fa` and `fd`, respectively). It is important to note that `dbcheck.pl` requires the datafiles input to it to have SNP lists in order to work and that this particular data validation program only works for files representing a single population like the Yorubans. The program opens up files and reads each line, splitting its content into an array; the list of SNPs is put into its own separate array. The program then iterates through the list of SNPs from the current line and queries the HapMap database for information using the SNP ID as a key.

`dbcheck.pl` uses two MySQL queries to check the datafiles it is reading in. The first one, the "ancestral" query, is:

```
SELECT IF(CHIMP_ref=1 AND CHIMP_var=0, 'REF',
IF(CHIMP_ref=0 AND CHIMP_var=1, 'VAR', 'NONE')) FROM
LocusInfo.dbsnp_mullikin M WHERE snp_id=?
```

This query checks a hard-coded ancestral information table for information on the SNP's values of `CHIMP_ref` and `CHIMP_var`. which should be either zero or one. Using these values, the program determines whether the chimp data says that the SNP's "reference" allele is ancestral or whether it says that the "other" allele is ancestral. `dbcheck.pl` does this with a short bit of nested logic in the MySQL statement itself. If `CHIMP_ref` is equal to one and `CHIMP_var` is equal to zero, then the "reference" allele is ancestral. If this is not the case, the logic checks to see if the reverse is true, that the data indicate that the "other" allele is ancestral. If `CHIMP_ref` and `CHIMP_var` are the same as each other or have a value of any

other number than zero or one, the logic returns "NONE", meaning that the data is either missing ancestral information or it is corrupted.

The second query, the "AFS" query, is:

```
SELECT rsId, totalcount*2, refhom_count*2+het_count,
otherhom_count*2+het_count FROM
Hapmap_v1p0.Hapmap_$hapMapVer H WHERE panelLSID IN
($popHash{$population}) AND rsId=? ORDER BY totalcount*2
```

For a given SNP in a population defined in `db.cfg`, the query returns its total number of chromosomes, its number of "reference" chromosomes, and its number of "other" chromosomes, where the data is ordered by the total number of chromosomes.

The main logic of `dbcheck.pl` is fairly simple. It compares the results of the two aforementioned MySQL queries directly to the SNP's corresponding data from the file stored in an array and outputs an error message if there is disagreement. The program specifies which particular column appears to be incorrect.

Determination of Chimp Ancestral Allele and Mapping of HapMap Genotype Frequencies for ENCODE Regions SNP Data

A dataset of ENCODE SNPs was obtained following a procedure derived from Quinlan and Marth 2007 and from personal communiqué with those authors (21). The five ENCODE regions under investigation are ENm013, ENm014, ENr112, ENr113, and ENr131. First, genomic traces from the NCBI Trace Archive[7] were matched to their particular amplicon[8] representing a part of an ENCODE region. Then, the traces were assembled using Marth's bioinformatics program Polybayes (22). The resulting genomic assembly was analyzed by Polyphred to discover SNPs. Finally, SNPs in regions sequenced using primers that have SNPs themselves were filtered out, resulting in a collection of SNPs that are assumed not to have an ascertainment bias (21).

These ENCODE SNP files (one for each of the five regions) are "distilled" using the Perl program `getUniqueSnps.pl`, which takes in a list of target text files as a command-line argument. The program produces a text file that contains only a single instance (line) of each unique SNP from the ENCODE SNP file with the following columns of information from the original file:

```
snpId   rs      chr     chrPos      region      regionPos
snpGenotypedByHapMap    snpDiscoveredByPolyPhred
snpPolymorphicInHapMapGenotypes
snpPolymorphicInPolyPhredGenotypes
```

[7] http://www.ncbi.nlm.nih.gov/Traces/

[8] An amplicon is a piece of DNA that has been synthesized between two primer pairs using amplification techniques like polymerase chain reaction (PCR).

Then, the distilled ENCODE SNP files are annotated with the Perl program `getSnpChimpData.pl`, which takes in a list of target text files. The end result is an annotated file that has three more columns added to the end of each line in the distilled file: `humanAllele`, `chimpAllele`, and `alignmentLevel`. These columns refer to information gathered from a local copy of a database (hg16, NCBI Build 34.3) downloaded from the UCSC Genome Browser about the reference allele for a given SNP in humans and in chimps. The underlying question is whether or not there is a difference between the two species' reference allele.

The first MySQL query in `getSnpChimpData.pl` is below:

```
SELECT distinct S.chromEnd, S.tSeq, S.qSeq FROM
hg16.chimpSimpleDiff S, hg16.rBestNetPanTro1 M WHERE
S.chrom = ? AND S.chromEnd = ? AND M.tName = S.chrom AND
S.chromEnd BETWEEN M.tStart + 1 and M.tEnd;
```

The first source of information for determining each reference allele is the database table `hg16.chimpSimpleDiff`. This table shows simple differences between chimp alignments and the human assembly within regions of high quality chimp sequence. The chimp data was obtained from the 13 Nov. 2003 Arachne assembly. [9] This information is joined with information from `hg16.rBestNetPanTro1`. This table (`rBestNetPanTro1`) shows the "reciprocal best" human/chimpanzee alignment net; it is useful for finding orthologous regions and for studying genome rearrangement.[10]

The program reads in a line of the distilled ENCODE SNP file and provides the MySQL query a chromosome number (`chrom`) and position (`chromEnd`) which together represent a single SNP. The query returns an instance of the SNP in the `chimpSimpleDiff` table and also makes sure the position lies within the genomic coverage of the human/chimpanzee alignment net. If there is a "simple difference", the human and chimp alleles (`tSeq` and `qSeq`) are printed into a new annotation text file and the next SNP is examined.

If there is no simple difference, the program determines if this is because there is no difference between the two species' alleles or if there is just no information about the SNP's position in the alignment net.

```
SELECT R.level, R.strand FROM hg16.rBestNetPanTro1 R WHERE
R.tName = ? AND R.tStart <= ? AND R.tEnd >= ? ORDER BY
R.level;
```

This MySQL query is provided the chromosome number (`tName`) and start and end positions (`tStart`, `tEnd`, which are the same number) of the SNP under

[9] http://genome.ucsc.edu/goldenPath/help/trackDescriptions.html
[10] http://genome.ucsc.edu/goldenPath/help/trackDescriptions.html

consideration. If there are no returned results, it is recorded in the annotation file that there is no data for the given SNP with the abbreviation "ND". If there are any results, they are sorted in descending order by the level of the alignment chain (reference website for hierarchy explanation), and it is assumed that the human and chimp alleles are the same. (The columns level and strand are returned just to act as a Boolean for the presence of database query results.)

Since the two species' alleles are the same, the exact allele is determined by locating the SNP's position in a local copy of human reference genomic sequence (NCBI Build 34.3). `Genome.genomeRefSeq_b34_3` contains equally-sized chunks of the sequence that can be accessed by position. This database is also stored locally.

```
SELECT G.endPos, G.seq FROM Genome.genomeRefSeq_b34_3 G
WHERE isAlt = 'N' AND G.chrom = ? AND G.startPos <= ? AND
G.endPos >= ?;
```

The query is provided the SNP's chromosome number and start/end position and returns the end position and the sequence chunk to which it corresponds. This information is used by `getSnpChimpData.pl` to parse the sequence string and return the allele letter to be recorded in the annotation file.

After the data annotation, a modified version of `afs-db12.pl` was used to produce the ENCODE region allele frequency spectra. The primary MySQL query used was:

```
SELECT proprietary_snpid AS snp_id, pb_totalcount*2,
pb_refhom_counts*2+pb_het_counts,
pb_other_counts*2+pb_het_counts FROM EncodeSNPAnalysis.
EncodeSNPInfo E WHERE hm_population = ? AND
pb_refhom_counts*2+pb_het_counts != 0 AND
pb_other_counts*2+pb_het_counts != 0 AND E.chrom !=
'X' AND E.chrom!='Y' AND numSNPs_in_all_primers = 0
```

In the process of studying the effects of the human-chimp genomic alignment on the AFS, the following line was appended to the primary MySQL query:

```
AND chimpHum_alignLevel != ?
```

References

1. Marth, G. T., Czabarka, E., Murvai, J., & Sherry, S. T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372 (2004).
2. Harpending, H. & Rogers, A. Genetic Perspectives on Human Origins and Differentiation. *Annu. Rev. Genomics Hum. Genet*. **1,** 361–85 (2000).
3. Cann, R. L., Stoneking, M., & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325,** 31–36 (1987).
4. Thorne, A. G., & Wolpoff, M. H. The Multiregional Evolution of Humans. *Scientific American* **266,** 76–83 (1992).
5. Marth, G. T. Coalescent modeling. BI820: Seminar in Quantitative and Computational Problems in Genomics. *Boston College* (Fall 2005). Available: http://clavius.bc.edu/~marth/BI820/
6. Rogers, A. R. *et al.* Ancestral Alleles and Population Origins: Inferences Depend on Mutation Rate. *Mol. Biol. Evol.* **24,** 990–997 (2007).
7. Crouau-Roy, B., Service, S., Slatkin, M., & Freimer, N. A fine-scale comparison of the human and chimpanzee genomes: linkage, linkage disequilibrium and sequence analysis. *Hum. Mol. Gen.* **5,** 1131–1137 (1996).
8. Noonan, J. P. *et al.* Sequencing and Analysis of Neanderthal Genomic DNA. *Science* **314,** 1113–1118 (2006).
9. Green, R. E. *et al.* Analysis of one million base pairs of Neanderthal DNA. *Nature* **444,** 330–336 (2006).
10. Hartl, D. L., Clark, A. G. *Principles of Population Genetics, 3rd ed.* (Sinauer Associates, Inc., Sunderland, Massachusetts, 1997).
11. Frankham, R. Relationship of Genetic Variation to Population Size in Wildlife. *Conserv. Biol.* **10,** 1500–1508 (1996).
12. Takahata, N. Allelic genealogy and human evolution. *Mol. Biol. Evol.* **10,** 2–22 (1993).
13. Zhao, Z. *et al.* Worldwide DNA sequence variation in a 10 kilobase noncoding region on chromosome 22. *Proc. Natl. Acad. Sci. USA* **97,** 11354–11358 (2000).
14. Yu, N. *et al.* Low Nucleotide Diversity in Chimpanzee and Bonobos. *Genetics* **164,** 1511–1518 (2003).
15. Nielsen, R. Population genetic analysis of ascertained SNP data. *Human Genomics* **1,** 218–224 (2004).
16. National Human Genome Research Institute. The ENCODE Project: ENCyclopedia Of DNA Elements. *National Institutes of Health*, (2007). Available: http://www.genome.gov/10005107/
17. Nielsen, R. and Signorovitch, J. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theor. Popul. Biol.* **63,** 245–55 (2003).
18. The International HapMap Consortium. A haplotype of the human genome. *Nature* **437,** 1299-1320 (2005).

19. Venter *et al.* The Sequence of the Human Genome. *Science* **291,** 1304–1351 (2001).
20. ENCODE Project. ENCODE Project at UCSC. *University of California at Santa Clara* (2007). Available: http://genome.ucsc.edu/ENCODE/
21. Quinlan, A. R. & Marth, G. T. Primer-site SNPs mask mutations. *Nat. Methods* **4,** 192 (2007).
22. Marth, G. T. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23,** 452–456 (1999).