HHS Public Access

Author manuscript

Bioessays. Author manuscript; available in PMC 2014 September 01.

Published in final edited form as:

Bioessays. 2013 September ; 35(9): 780–786. doi:10.1002/bies.201300014.

SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it

Joseph Lachance[1,*] and **Sarah A. Tishkoff**[1]

[1]Departments of Biology and Genetics, University of Pennsylvania, Philadelphia, PA 19104 USA

Summary

Whole genome sequencing and SNP genotyping arrays can paint strikingly different pictures of demographic history and natural selection. This is because genotyping arrays contain biased sets of pre-ascertained SNPs. In this short review, we use comparisons between high-coverage whole genome sequences of African hunter-gatherers and data from genotyping arrays to highlight how SNP ascertainment bias distorts population genetic inferences. Sample sizes and the populations in which SNPs are discovered affect the characteristics of observed variants. We find that SNPs on genotyping arrays tend to be older and present in multiple populations. In addition, genotyping arrays cause allele frequency distributions to be shifted towards intermediate frequency alleles, and estimates of linkage disequilibrium are modified. Since population genetic analyses depend on allele frequencies it is imperative that researchers are aware of the effects of SNP ascertainment bias. With this in mind we describe multiple ways to correct for SNP ascertainment bias.

Keywords

African hunter-gatherers; human genetics; population genetics; SNP ascertainment bias; whole genome sequencing

African hunter-gatherers and the power of whole genome sequencing

Due to technological advances and increases in computational power the cost of genotyping has plummeted over the past few years. Because of this, it is now feasible to conduct population genetic analyses of whole genome sequencing data. One advantage of whole genome sequencing is that SNP ascertainment bias is reduced compared to alternative genotyping technologies. This lack of SNP ascertainment bias is critical for accurate population genetic analyses where allele frequency distributions are used to infer demographic history and scan for past targets of natural selection. Using the technology of Complete Genomics [1], we recently sequenced the whole genomes of 15 African hunter-gatherers at >60x coverage [2]. Sequenced individuals included five Pygmies from Cameroon, five Hadza from Tanzania, and five Sandawe from Tanzania. The genomes of these diverse African hunter-gathers contain millions of previously unknown variants [2]. Because these variants were largely free of ascertainment bias found on genotyping arrays,

*Corresponding author (lachance.joseph@gmail.com).

by using whole genome sequencing we were able to make much more accurate inferences of demographic history, ancient admixture, and local adaptation.

By comparing whole genome sequence data of African hunter-gatherers to SNP variation observed using an Illumna-1M Duo BeadChip array, we highlight how SNP ascertainment bias can have a major impact on population genetic analyses. The Illumina-1M Duo array contains approximately 1.2 million markers, the majority of which are tag SNPs for use in genome-wide association studies. This array is enriched for SNPs in the MHC region of chromosome 6 and for coding SNPs that were discovered by the 1000 genomes project. We chose the Illumina-1M Duo array for comparisons with whole genome sequence data because it is representative of a genotyping technology that uses a genome-wide set of pre-ascertained SNPs. An additional set of SNP genotyping arrays are made by Affymetrix. Population genetic analysis of Human Genome Diversity Panel (HGDP) data using Illumina 650K and Affymetrix 500K SNP arrays gave largely similar results for both genotyping platforms [3]. However, heterozygosity estimates were smaller for the Affymetrix 500k array than the Illumina 650K array [3]. Other more specialized genotyping arrays, such as the Metabochip [4], are expected to present additional challenges with respect to ascertainment bias.

SNP ascertainment bias arises from many sources

SNP ascertainment bias is the systematic deviation of population genetic statistics from theoretical expectations, and it can be caused by sampling a nonrandom set of individuals or by biased SNP discovery protocols. Unless the whole genome of every individual in a population is sequenced there will always be some form of SNP ascertainment bias. This is because a small sample size is more likely to “catch” common alleles than rare alleles [5]. An additional issue is that ascertainment schemes are not always explicitly known and SNPs on genotyping arrays are often ascertained in a non-uniform manner. For example, a disproportionate number of SNPs have been identified by sequencing European individuals [6]. Previous comparisons between HapMap and Perlegen datasets indicate that population genetic analyses based on ascertained SNP data yield inaccurate results [7]. This is because classical models of theoretical population genetics do not explicitly take into account SNP ascertainment bias [8]. By contrast, the problem of ascertainment bias is less of an issue for individual identification, paternity analyses, and assigning individuals to different populations [9,10].

Because genetic diversity is unequally distributed across populations, the populations in which SNPs are discovered contribute to SNP ascertainment bias. For example, human populations from Africa contain greater genetic diversity than populations from Europe, Asia, Oceania, or the Americas. This pattern arises from serial bottlenecks and founder effects as modern humans expanded from Africa to colonize other continents. Even among African populations the number of SNPs detected can vary substantially: we observed a total of 8.9 million variants in five Pygmy genomes, 7.3 million variants in five Hadza genomes, and 8.2 million variants in five Sandawe genomes [2]. When SNPs are ascertained in one population and used to genotype other populations erroneous conclusions can result [11]. This is particularly important for SNPs originally discovered in less diverse (non-African)

populations. For example, when the heterozygosity of variants ascertained in European populations are assessed it can falsely lead to the conclusion that European populations harbor a greater amount of variation than African populations [12] and to mis-estimates of effective population sizes [13].

An additional form of bias is that genotyping arrays are enriched for SNPs in some genomic regions and deficient for SNPs in other genomic regions. For each African hunter-gatherer population, we compared the number of SNPs per 100kb window from whole genome sequencing and the number of SNPs present on the Illumna-1M Duo genotyping array (Fig. 1). As expected, SNP density was much higher for whole genome sequencing. Overall, genomic regions with high numbers of SNPs from whole genome sequencing also had a high number of SNPs on the Illumina-1M Duo array (Pearson's r : 0.62457 for Pygmies, 0.62404 for Hadza, 0.63124 and for Sandawe). However, not all 100kb windows are represented equally on the Illumina-1M Duo array (Fig. 1), and a moderately high correlation between the number of variants per windows does not mean that the population genetic properties of SNPs are the same for SNPs assayed using different genotyping platforms.

Compared to SNPs, ascertainment bias is stronger for indels and weaker for microsatellites [14]. The high amount of bias for indels may be due to ascertainment schemes that are enriched for large allele frequency differences between European and African populations [14]. Microsatellites are relatively buffered from ascertainment bias because these variants have high mutation rates and are more likely to be heterozygous (microsatellites tend to be polymorphic in multiple populations) [15]. However, even microsatellites are not entirely free from ascertainment bias [16].

Ascertained SNPs lead to a biased view of evolutionary history

Small sample sizes bias allele frequency distributions towards common SNPs. This occurs even if data arise from whole genome sequencing. The probability that an autosomal variant with an allele frequency of p is polymorphic in a sample of n diploid genomes is given by: $P(polymorphic|\ n, p) = 1 - p^{2n} - (1 - p)^{2n}$ [17]. This expression indicates that polymorphisms are more likely to be detected if allele frequencies are intermediate ($p = 0.5$) and sample sizes are large (Fig. 2A). Under the neutral theory of evolution the probability density of alleles is inversely proportional to the derived allele frequency [18], meaning that the majority of polymorphic sites contain low frequency derived alleles. The theoretical probability density of neutral derived alleles can be weighted by the probability of detecting a polymorphism in a small sample (Fig. 2B). Here, we see that smaller sample sizes result in derived allele frequency distributions with fewer rare alleles (Fig. 2B).

Pre-ascertained SNPs found on genotyping arrays result in skewed allele frequency distributions. SNPs identified from whole genome sequencing tend to be uncommon (Fig. 3A–C). By contrast, SNPs on the lllumina-1M Duo array are biased toward intermediate frequency alleles that are found in multiple populations (Fig. 3D–F). This rightward shift in allele frequency distributions towards intermediate frequency alleles has also been observed in comparisons between ascertained HapMap SNPs and Perlegen data from low-coverage

whole genome sequencing [7]. Modern computational methods use joint allele frequency distributions from multiple populations to infer divergence times, ancestral population sizes, and migration rates [19,20]. Because genotyping arrays modify allele frequency distributions (Fig. 3) demographic inference from ascertained SNPs will yield flawed results (including misestimated divergence times, effective population sizes, and migration rates [21]). Additionally, the mean derived allele frequency (DAF) is greater for ascertained SNPs than SNPs identified from whole genome sequencing of African hunter-gatherers (Fig. 4A). Because population bottlenecks also result in increased derived allele frequencies, SNP ascertainment bias can make it appear that populations have shrunk in size [11,22]. Also, the relative increase of derived allele frequency due to SNP ascertainment bias varies by population (32% for Pygmies, 41% for Hadza, and 27% for Sandawe).

Compared to variants observed from whole genome sequencing, pre-ascertained SNPs are biased toward older SNPs. Using derived allele frequency distributions and an equation that connects allele frequency to SNP age (Equation 4 in [23]), we estimated the ages of SNPs found via whole genome sequencing vs. the ages of SNPs on the Illumina-1M Duo array. Ascertained SNPs were 13–18% older than SNPs from whole genome sequencing of African hunter-gatherers (Fig. 4B). Older SNPs are less likely to be population-specific, and this can cause fine-scale patterns of diversity to be missed. However, analyzing large numbers of SNPs can overcome some of these difficulties [24,25].

Measures of population differentiation, such as F_{ST}, are also affected by SNP ascertainment bias. Older SNPs, such as ascertained SNPs found on genotyping arrays, can drift to very different allele frequencies in divergent populations. Furthermore, the magnitude of F_{ST} depends upon minor allele frequencies in each population [26]. Comparisons between whole genome sequences and SNP array data for African hunter-gatherers reveals that ascertained SNPs tend to have higher values of F_{ST} (Fig. 4C). This means that ascertained SNPs will tend to overestimate the amount of population differentiation. Interestingly, STRUCTURE [27], which assigns individual ancestry to a finite number of population clusters, is relatively robust to SNP ascertainment bias [28]. This occurs because STRUCTURE uses large multi-locus datasets and the most informative SNPs for ancestry inference are variants with large frequency differences across populations [29].

Ascertainment bias can also affect scans of selection. Signatures of natural selection include extended haplotype homozygosity [30,31], large locus-specific branch lengths [32], skewed allele frequency distributions [33], and high values of F_{ST} [34]. Demographic history shapes variation across the whole genome while the effects of selection tend to be locus-specific. Because of this, outlier loci are strong candidates for natural selection. However, outlier approaches can lead to many false positives and negatives [35,36], particularly when an ascertained subset of SNPs is analyzed. In our analysis of high-coverage whole genome sequences of African hunter-gatherers we identified many putative targets of natural selection [2]. The majority of these putative targets of selection were population-specific. Genotyping arrays with ascertained SNPs have also been used to detect signatures of natural selection in Pygmy, Hadza, and Sandawe populations [37,38]. However, with a few notable exceptions, there was minimal overlap between putatively selected genomic regions identified using different genotyping platforms. Some of this lack of overlap is due to

methodological differences across studies, but it illustrates that ascertainment bias can potentially affect which genomic regions are identified from scans of selection.

SNP ascertainment bias can also cause linkage disequilibrium (LD) to be misestimated. The ability of marker SNPs to tag genomic regions depends on the amount of LD, and because LD decays faster in African populations [39], greater SNP densities are required for association studies that use African samples. Although less bias is expected for multi-locus haplotype statistics than single-locus statistics [40], there is evidence from theoretical population genetics that ascertainment bias modifies LD decay curves (including elevated r^2 values and reduced |D'| values) [41]. LD decay curves from ascertained SNP data have been used to estimate the effective sizes of HapMap populations [42], divergence times and population size changes for 17 global populations [43], and demographic inference of Yoruba and French populations from the HGDP panel [44]. To directly assess the effects of ascertainment bias, we generated LD decay curves for Pygmy, Hadza, and Sandawe populations using data from whole genome sequencing and the Illumina1M-Duo SNP array (Fig. 5). Here, 10,000 randomly chosen pairs of linked SNPs were chosen for each population and genotyping platform, SNP pairs were binned into 1kb intervals, and mean r^2 values were calculated for each bin. Qualitatively similar patterns of LD were found for each platform: values of r^2 were largest for the Hadza, intermediate for Sandawe, and smallest for Pygmy samples (Fig. 5). However, estimates of r^2 were consistently larger for ascertained SNPs than SNPs obtained from whole genome sequence data (Fig. 5). This underscores the need to correct for SNP ascertainment bias when population genetic inferences are made from LD decay curves.

How can researchers cope with SNP ascertainment bias?

Ideally, whole genome sequencing would be used to avoid SNP ascertainment bias. However, because allele frequency distributions are shaped by small sample sizes one can never be completely free of SNP ascertainment bias. Furthermore, the cost of high-coverage whole genome sequencing ($3000–$5000 per human genome) is high enough that few labs will be able to afford large sample sizes. Low-coverage sequencing, such as that of the 1000 genomes project [45], runs the risk of biasing allele frequency distributions because singletons are harder to call. As an alternative, exome sequencing [46] and other sequence capture approaches [47] can be used. Under these methods an affordable subset of the genome is sequenced using next-generation sequencing technologies. However, this can be problematic because coding regions of the genome have different population genetic properties than non-coding regions [2], and care must be taken to capture enough independent regions of the genome.

Barring the use of whole genome sequences, ascertainment bias can be reduced by using haplotype data. Partly because they are more likely to be polymorphic in multiple populations, haplotype statistics are less affected by ascertainment bias than single locus statistics [40,48,49]. However, use of haplotype statistics requires that phase information is known and accurate phase information requires large sample sizes and/or pedigree data.

One way to correct for SNP ascertainment bias is to modify raw genotype data. Corrections can be made by modeling SNP discovery (including depth of coverage) and then incorporating this information into either a maximum likelihood [8,11] or Bayesian [50] framework to estimate allele frequency distributions and other demographic parameters. For example, Albrechtsen et al. used this approach to correct for ascertainment bias on the Affymetrix 500K SNP chip [51]. They used maximum likelihood and Celera sequence data to reverse-engineer ascertainment schemes and modify allele frequency distributions from the Affymetrix chip. However, because resequencing data are not always available it is not always possible to correct for ascertainment bias as per Albrechtsen et al. [51].

An alternative method is to explicitly incorporate SNP ascertainment bias into population genetics models. Here, empirical datasets can then be compared to modified expectations from theory. Using this approach Nielsen and Signorovitch demonstrated that modeling SNP ascertainment bias leads to more accurate estimates of linkage disequilibrium [41]. Similarly, by incorporating ascertainment bias into theoretical models Wakeley et al. were able to improve estimates of gene flow and population size changes [21]. Software packages like SIMCOAL2 are also able to simulate some forms of ascertainment bias through the use of a minimum allele frequency filter [52]. However, one limitation of explicitly incorporating bias into theoretical models is that it is not always possible to know the ascertainment scheme that was actually used. One way around this is to use SNPs ascertained in a single individual, such as SNPs found on the Human Origins Array [53], but this approach is most useful when it is restricted to the population of original ascertainment.

Conclusions and outlook

The value of population genetic data is maximized when researchers are aware of existing biases. Using whole genome sequence data from African hunter-gatherers we have illustrated how SNP ascertainment bias can distort inferences about demographic history and natural selection of populations. It is important that published studies describe any ascertainment schemes used to generate data. Because whole genome sequencing reduces the amount of ascertainment bias, we encourage researchers to use high-coverage sequencing data whenever it is feasible. Regardless of the genotyping platform used, it is important to correct for ascertainment bias (either by modifying raw data or incorporating ascertainment bias into theoretical models of population genetics).

Acknowledgements

This work was supported by an NIH postdoctoral fellowship (F32HG006648-02) to JL and an NIH Pioneer Award (DP1 ES022577-04) to ST. The authors would like to thank three anonymous reviewers, Andrew Moore, and the editorial staff at BioEssays for helpful comments and suggestions.

Abbreviations

LD	linkage disequilibrium
MHC	major histocompatibility complex
SNP	single-nucleotide polymorphism

References

1. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010; 327:78–81. [PubMed: 19892942]
2. Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell. 2012; 150:457–69. [PubMed: 22840920]
3. Lopez Herraez D, Bauchet M, Tang K, Theunert C, Pugach I, et al. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PLoS One. 2009; 4:e7888. [PubMed: 19924308]
4. Buyske S, Wu Y, Carty CL, Cheng I, Assimes TL, et al. Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: the PAGE study. PLoS One. 2012; 7:e35651. [PubMed: 22539988]
5. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, et al. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A. 2011; 108:11983–8. [PubMed: 21730125]
6. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]
7. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005; 15:1496–502. [PubMed: 16251459]
8. Nielsen R, Hubisz MJ, Clark AG. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics. 2004; 168:2373–82. [PubMed: 15371362]
9. Morin PA, Luikart G, Wayne RK, group tSw. SNPs in ecology, evolution and conservation. Trends in Ecology and Evolution. 2004; 19:208–16.
10. Bradbury IR, Hubert S, Higgins B, Bowman S, Paterson IG, et al. Evaluating SNP ascertainment bias and its impact on population assignment in Atlantic cod, Gadus morhua. Mol Ecol Resour. 2011; 11(Suppl 1):218–25. [PubMed: 21429176]
11. Nielsen R. Population genetic analysis of ascertained SNP data. Hum Genomics. 2004; 1:218–24. [PubMed: 15588481]
12. Eller E. Effects of ascertainment bias on recovering human demographic history. Hum Biol. 2001; 73:411–27. [PubMed: 11459422]
13. Rogers AR, Jorde LB. Ascertainment bias in estimates of average heterozygosity. Am J Hum Genet. 1996; 58:1033–41. [PubMed: 8651264]
14. Romero IG, Manica A, Goudet J, Handley LL, Balloux F. How accurate is the current picture of human genetic variation? Heredity (Edinb). 2009; 102:120–6. [PubMed: 18766200]
15. Harpending H, Rogers A. Genetic perspectives on human origins and differentiation. Annu Rev Genomics Hum Genet. 2000; 1:361–85. [PubMed: 11701634]
16. Eriksson A, Manica A. Detecting and Removing Ascertainment Bias in Microsatellites from the HGDP-CEPH Panel. G3 (Bethesda). 2011; 1:479–88. [PubMed: 22384358]
17. Sethupathy P, Hannenhalli S. A tutorial of the poisson random field model in population genetics. Adv Bioinformatics. 2008:257864. [PubMed: 19920987]
18. Lachance J. Disease-associated alleles in genome-wide association studies are enriched for derived low frequency alleles relative to HapMap and neutral expectations. BMC Med Genomics. 2010; 3:57. [PubMed: 21143973]
19. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009; 5:e1000695. [PubMed: 19851460]
20. Lukic S, Hey J, Chen K. Non-equilibrium allele frequency spectra via spectral methods. Theor Popul Biol. 2011; 79:203–19. [PubMed: 21376069]
21. Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K. The discovery of single-nucleotide polymorphisms--and inferences about human demographic history. Am J Hum Genet. 2001; 69:1332–47. [PubMed: 11704929]

22. Polanski A, Kimmel M. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics. 2003; 165:427–36. [PubMed: 14504247]
23. Slatkin M, Rannala B. Estimating allele age. Annu Rev Genomics Hum Genet. 2000; 1:225–49. [PubMed: 11701630]
24. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. Genes mirror geography within Europe. Nature. 2008; 456:98–101. [PubMed: 18758442]
25. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, et al. Correlation between genetic and geographic structure in Europe. Curr Biol. 2008; 18:1241–8. [PubMed: 18691889]
26. Jakobsson M, Edge MD, Rosenberg NA. The Relationship Between FST and the Frequency of the Most Frequent Allele. Genetics. 2012
27. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–59. [PubMed: 10835412]
28. Haasl RJ, Payseur BA. Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. Heredity (Edinb). 2011; 106:158–71. [PubMed: 20332809]
29. Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet. 2003; 73:1402–22. [PubMed: 14631557]
30. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006; 4:e72. [PubMed: 16494531]
31. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002; 419:832–7. [PubMed: 12397357]
32. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, et al. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics. 2004; 1:274–86. [PubMed: 15588487]
33. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123:585–95. [PubMed: 2513255]
34. McDonald, JH. Detecting natural selection by comparing geographic variation in protein and DNA polymorphisms. In: Golding, B., editor. Non-neutral evolution: theories and molecular data. Chapman & Hall; New York: 1994. p. 88-100.
35. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. 2006; 16:980–9. [PubMed: 16825663]
36. Thornton KR, Jensen JD. Controlling the false-positive rate in multilocus genome scans for selection. Genetics. 2007; 175:737–50. [PubMed: 17110489]
37. Jarvis JP, Scheinfeldt LB, Soi S, Lambert C, Omberg L, et al. Patterns of Ancestry, Signatures of Natural Selection, and Genetic Association with Stature in Western African Pygmies. PLoS Genet. 2012; 8:e1002641. [PubMed: 22570615]
38. Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, et al. Limited evidence for classic selective sweeps in African populations. Genetics. 2012; 192:1049–64. [PubMed: 22960214]
39. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 2008; 451:998–1003. [PubMed: 18288195]
40. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet. 2006; 38:1251–60. [PubMed: 17057719]
41. Nielsen R, Signorovitch J. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theor Popul Biol. 2003; 63:245–55. [PubMed: 12689795]
42. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, et al. Recent human effective population size estimated from linkage disequilibrium. Genome Res. 2007; 17:520–6. [PubMed: 17351134]
43. McEvoy BP, Powell JE, Goddard ME, Visscher PM. Human population dispersal “Out of Africa” estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Res. 2011; 21:821–9. [PubMed: 21518737]

44. Theunert C, Tang K, Lachmann M, Hu S, Stoneking M. Inferring the history of population size change from genome-wide SNP data. Mol Biol Evol. 2012; 29:3653–67. [PubMed: 22787284]

45. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–73. [PubMed: 20981092]

46. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–6. [PubMed: 19684571]

47. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010; 7:111–8. [PubMed: 20111037]

48. Lohmueller KE, Bustamante CD, Clark AG. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. Genetics. 2009; 182:217–31. [PubMed: 19255370]

49. Novembre J, Ramachandran S. Perspectives on human population structure at the cusp of the sequencing era. Annu Rev Genomics Hum Genet. 2011; 12:245–74. [PubMed: 21801023]

50. Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, et al. Demographic history of Oceania inferred from genome-wide data. Current biology. 2010; 20:1983–92. [PubMed: 21074440]

51. Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol. 2010; 27:2534–47. [PubMed: 20558595]

52. Laval G, Excoffier L. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics. 2004; 20:2485–7. [PubMed: 15117750]

53. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, et al. Ancient admixture in human history. Genetics. 2012; 192:1065–93. [PubMed: 22960212]

54. Weir, BS. Genetic data analysis II : methods for discrete population genetic data. Vol. xii. Sinauer Associates; Sunderland, Mass: 1996. p. 445

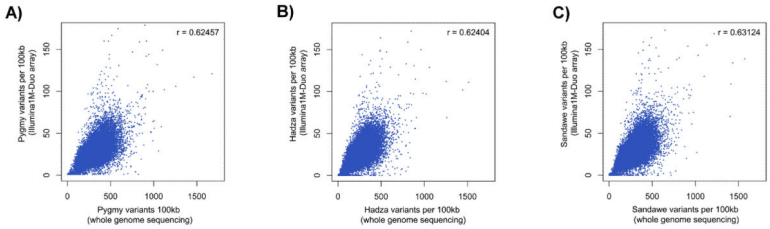

Figure 1.

SNP density using two different genotyping platforms. Numbers of fully called autosomal SNPs per 100kb window are plotted for the llumina1M-Duo array and whole genome sequencing by Complete Genomics. In each panel the genomes of five African hunter-gatherers were analyzed (**A:** Pygmy, **B:** Hadza, and **C:** Sandawe).

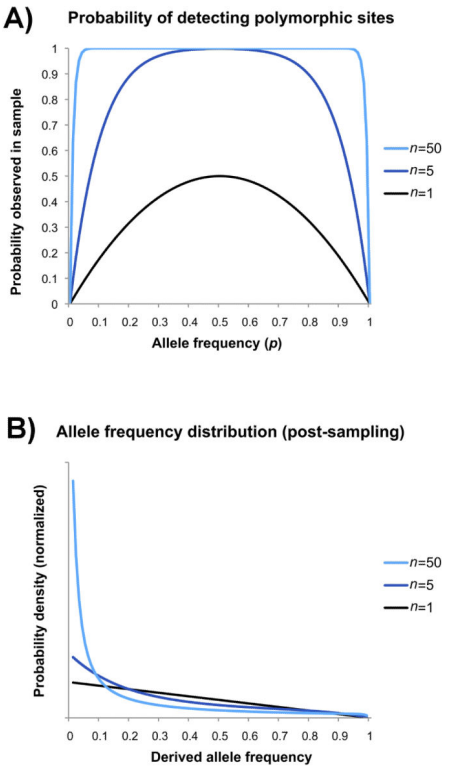

Figure 2.

Small sample sizes modify allele frequency distributions. Allele frequency distributions shown are for source populations, not samples. Neutrality and constant population sizes are assumed. **A:** Probability that a site is polymorphic in a sample of *n* diploid individuals as a function of derived allele frequency. **B:** Theoretical allele frequency distributions given a SNP discovery panel of *n* diploid individuals.

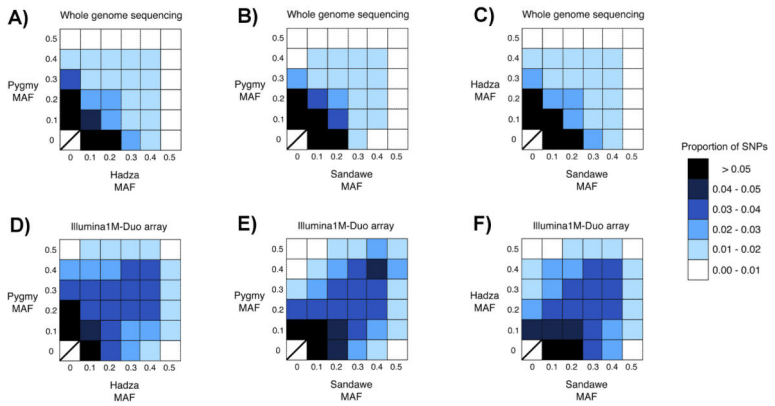

Figure 3.

Joint allele frequency distributions of African hunter-gatherers. Analyzed SNPs analyzed are autosomal, called in all samples, and polymorphic in at least one population. Minor allele frequencies for pairs of populations are shown in each panel. Top panels show allele frequencies from whole genome sequencing (Complete Genomics). **A:** Pygmy and Hadza, **B:** Pygmy and Sandawe, **C:** Hadza and Sandawe. Bottom panels show allele frequencies from the Illumina1M-Duo genotyping array. **D:** Pygmy and Hadza, **E:** Pygmy and Sandawe, **F:** Hadza and Sandawe.

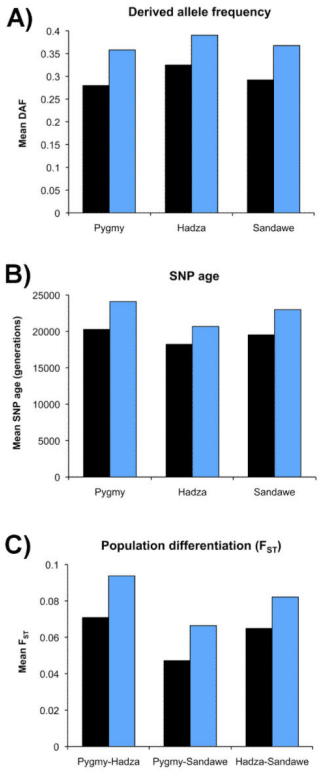

Figure 4.

Population genetic statistics from two different genotyping platforms. Fully called autosomal SNPs were analyzed, and hypermutable CpG sites were omitted. Whole genome sequencing data is in black and SNP data from the Illumina1M-Duo genotyping array is in blue. **A:** Derived allele frequencies. **B:** Mean age of SNPs, using Equation 4 from [23] and effective population sizes from [2]. **C:** Population differentiation, as measured by mean F_{ST}.

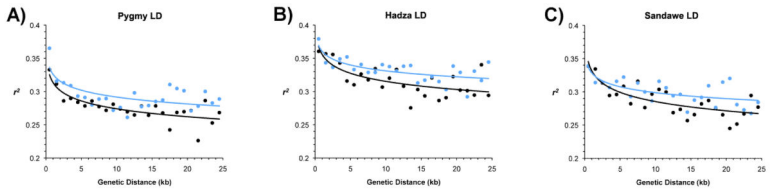

Figure 5.

Linkage disequilibrium (LD) decay curves. Fully called autosomal SNPs were analyzed and gamete frequencies were calculated as per the EM algorithm in [54]. Five individuals per population were analyzed. Whole genome sequencing data is in black and SNP data from the Illumina1M-Duo genotyping array is in blue. Solid lines indicate the logarithmic least squares fit for each genotyping platform. Each panel shows LD decay curves for a different hunter-gatherer population (**A:** Pygmy, **B:** Hadza, and **C:** Sandawe).