Low-Pass Genome-Wide Sequencing and Variant Inference Using Identity-by-Descent in an Isolated Human Population

A. Gusev,^{*,†,1} M. J. Shah,^{*,‡} E. E. Kenny,^{*,§} A. Ramachandran,^{*,†} J. K. Lowe,^{§,**,††} J. Salit,^{§,††} C. C. Lee,[‡] E. C. Levandowsky,[‡] T. N. Weaver,[‡] Q. C. Doan,[‡] H. E. Peckham,[‡] S. F. McLaughlin,[‡] M. R. Lyons,[‡] V. N. Sheth,[‡] M. Stoffel,^{§§} F. M. De La Vega,[‡] J. M. Friedman,[§] J. L. Breslow,[§] and I. Pe'er^{*,†}

^{*}Department of Computer Science and [†]Center of Computational Biology and Bioinformatics, Columbia University, New York, New York 10027, [‡]Life Technologies, Beverly, Massachusetts 01915, [§]The Rockefeller University, New York, New York 10065, ^{**}Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, ^{††}Program in Medical and Population Genetics, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, ^{§§}School of Medicine, Cornell University, New York, New York 10065, and ^{§§}Institute of Molecular Systems Biology, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland

ABSTRACT Whole-genome sequencing in an isolated population with few founders directly ascertains variants from the population bottleneck that may be rare elsewhere. In such populations, shared haplotypes allow imputation of variants in unsequenced samples without resorting to complex statistical methods as in studies of outbred cohorts. We focus on an isolated population cohort from the Pacific Island of Kosrae, Micronesia, where we previously collected SNP array and rich phenotype data for the majority of the population. We report identification of long regions with haplotypes co-inherited between pairs of individuals and methodology to leverage such shared genetic content for imputation. Our estimates show that sequencing as few as 40 personal genomes allows for inference in up to 60% of the 3000-person cohort at the average locus. We ascertained a pilot data set of whole-genome sequences from seven Kosraean individuals, with average 5× coverage. This assay identified 5,735,306 unique sites of which 1,212,831 were previously unknown. Additionally, these variants are unusually enriched for alleles that are rare in other populations when compared to geographic neighbors (published Korean genome SJK). We used the presence of shared haplotypes between the seven Kosraean individuals to estimate expected imputation accuracy of known and novel homozygous variants at 99.6% and 97.3%, respectively. This study presents whole-genome analysis of a homogenous isolate population with emphasis on optimal rare variant inference.

FOUNDER populations play significant roles in population genetics and trait mapping due to the effects of bottlenecks and drift on their genetic variation (Hirschhorn and Daly 2005). Such populations are singularly useful in identifying rare disease variants that often appear in the isolated cohort at a higher frequency or within a more clearly discernible haplotype structure (Peltonen *et al.* 2000) than in out-bred populations. Additionally, identified variants are still valuable beyond the isolated group as their effect replicates in more outbred populations (Newman *et al.* 2004;

Lowe *et al.* 2009) and can implicate new functionally important genes. Next-generation sequencing of personal genomes has revealed a multitude of rare variants. Presently, high-quality personal genomes exist from representatives of the major continental groups (Levy *et al.* 2007; Bentley *et al.* 2008; Wang *et al.* 2008; Wheeler *et al.* 2008; Ahn *et al.* 2009; Kim *et al.* 2009; McKernan *et al.* 2009), but only genotyping and lower-throughput sequencing data are available for isolated populations (Shifman *et al.* 2008; Sabatti *et al.* 2009). This article reports low-pass whole-genome sequencing and analysis of seven individuals from an isolated Pacific population, chosen specifically for the insight they might provide into the larger cohort as well as the presence and functional importance of rare variants they carry.

The cost of whole-genome sequencing is not trivial and the best strategy for identification of rare causative variants

Copyright © 2012 by the Genetics Society of America
doi: 10.1534/genetics.111.134874

Manuscript received September 21, 2011; accepted for publication November 14, 2011
Supporting information is available online at <http://www.genetics.org/content/suppl/2011/12/01/genetics.111.134874.DC1>.

¹Corresponding author: Department of Computer Science, Columbia University, 500 W. 120th St., New York, NY 10027. E-mail: gusev@cs.columbia.edu

must balance the number of genomes sequenced with the insights gained that are applicable to different populations and multiple traits. For common traits, one may sequence a reference panel to statistically impute variants in populations represented by such a panel (Marchini *et al.* 2007). However, this requires sequencing high numbers of genomes and is still severely underpowered in populations or variants that are underrepresented in such data sets [e.g., isolated populations (Huang *et al.* 2009) and rare variants (Marchini and Howie 2010)]. For Mendelian diseases a successful strategy has been whole-exome capture in a small number of individuals (Ng *et al.* 2009; Ng *et al.* 2010). However, such studies are limited to extremely penetrant phenotypes, inherently ignore noncoding regions, and do not yet scale to population-based analysis (Galvan *et al.* 2010). Another alternative strategy has been targeted resequencing of candidate loci detected in a genome-wide association study (GWAS) across many individuals. Nevertheless, pursuing such a strategy genome-wide is still resource intensive despite a considerable drop in sequencing costs and scales poorly for multiple traits across a large number of loci in each.

We set out to leverage the opportunities and address the challenges of sequencing-based mapping in a multitrait GWAS cohort from an isolated population. Ongoing work by large sequencing consortia, such as the 1000 Genomes Project (2010) (Durbin *et al.* 2011), has shown that analyzing multiple individuals, even at low coverage, improves quality and completeness of detecting and calling novel variants. Moreover, information from a small subsample of sequenced individuals combined with relatively inexpensively acquired SNP array platforms can be used to impute much of the missing variation with high accuracy (Marchini *et al.* 2007; Li *et al.* 2009). In the current study we used this knowledge to develop a sequencing-based framework that leverages the inherent potential of a sizeable phenotyped cohort with a small founder population. We applied it to the Kosraen data set in which we previously found an abundance of long stretches of the genome identical-by-descent even between reportedly unrelated pairs of individuals (Gusev *et al.* 2009).

We sequenced a pilot group of seven individuals and performed multisample calling and imputation to quantify the informativeness of this cohort. The detected variants were validated and compared with those observed in other published whole-genome sequencing efforts from different populations. Internally, we analyzed the distribution of all variation as well as individual functional classes. Finally, we estimated the effectiveness of identical-by-descent (IBD) segments detected from SNPs in predicting the underlying untyped variants.

Materials and Methods

We have been studying genetic determinants for a multitude of traits in a cohort of 2906 individuals (the majority of adults) from the Micronesian island of Kosrae. This cohort has been previously described (Shmulewitz *et al.* 2006) and genotyped on the Affymetrix 500k SNP array platform to

detect positive GWAS results for seven phenotypes (Lowe *et al.* 2009). Subsequently, we reported a GWAS in which 27 traits were analyzed under family-based models (Kenny *et al.* 2011) and quantified the abundance of IBD segments within the cohort (Gusev *et al.* 2009).

Here, we utilized the autosomal SNP genotype data to estimate pervasiveness of IBD in genomic regions between arbitrary pairs of samples and, as a consequence, the potential for imputation based on identity-by-descent in this population. Pairwise, identical-by-descent regions were discovered using GERMLINE, a tool for efficient whole-genome IBD segment detection from partially phased data (Gusev *et al.* 2009). For the purpose of imputation, we conservatively examined only IBD segments >5 cM, where GERMLINE has been demonstrated to have 100% specificity in simulation (Browning and Browning 2010; Gusev *et al.* 2009). We found that for an average individual, such regions span a substantial 10.8% of all genotypes in the remaining cohort. We then sought to estimate the utility of these IBD segments for imputing genomic data within the population from a sequenced subgroup. We developed a method for optimizing the selection of highly representative individuals to sequence and quantifying the amount of data that can be inferred from their genomes. We next introduce precise notation and detail the algorithm.

Terminology and notation

IBD haplotypes: A pair of descendants from the same ancestor is identical-by-descent where they share loci that have been transmitted along the respective lineages leading to them. A continuous run of such loci with no recombination along the lineages is then an IBD haplotype. The shared haplotypes lie on homologous chromosomes of different individuals or of the same individual, where in the latter case the individual has related parents. Let $\mathbf{P} = \{1, 2, \dots, n\}$ be the set of all individuals belonging to the population under study. We denote by $\mathbf{R}(i, j)$ the collection of shared regions between individuals $i, j \in \mathbf{P}$. A shared region in $\mathbf{R}(i, j)$ is identified by a tuple (l, r) , where l is its left endpoint and r its right endpoint.

Total information content: Our aim is to sequence only a subset of individuals to infer information about the unsequenced population. Total information content (TIC) of a subset of individuals $\mathbf{Q}$ is the fraction of the cohort members' genomes that we directly obtain or indirectly can infer by sequencing the individuals in $\mathbf{Q}$. Formally, if we define an indicator function $I(i, k, \mathbf{Q})$ for individual i specifying the existence of an IBD segment with any individual in $\mathbf{Q}$ at locus k ,

$$I(i, k, \mathbf{Q}) = \begin{cases} 1 & \exists q \in \mathbf{Q} (\exists (l, r) \in \mathbf{R}(i, q) (l < k \wedge r > k)) \\ 0 & \text{otherwise.} \end{cases}$$

Then, the amount that can be inferred for an individual i not in $\mathbf{Q}$ is simply the sum of these indicators over the genome G :

$$L(i, \mathbf{Q}) = \sum_{g \in G} I(i, g, \mathbf{Q}).$$

And the total information content for cohort $\mathbf{P}$, subset $\mathbf{Q}$ is

$$\text{TIC}(\mathbf{P}, \mathbf{Q}) = \frac{|\mathbf{Q}|G + \sum_{i \in \mathbf{P} \setminus \mathbf{Q}} L(i, \mathbf{Q})}{|\mathbf{P}|G}.$$

Utility of sequencing an individual: Given a set of already sequenced individuals $\mathbf{Q}$, we associate each individual $i \in \mathbf{P} \setminus \mathbf{Q}$ with it a quantity $U(i, \mathbf{Q})$ that corresponds to the utility of sequencing i at this stage. $U(i, \mathbf{Q})$ is the total fraction of uninferred regions that i shares with all unsequenced individuals across all chromosomes, calculated as $\text{TIC}(\mathbf{P}, \{\mathbf{Q}, i\}) - \text{TIC}(\mathbf{P}, \mathbf{Q})$.

Given a fixed sequencing budget b , our proposed methodology optimizes the selection of individuals in $\mathbf{Q}$ of size b to maximize TIC. This problem is reducible to the classic NP-hard maximal coverage problem. At each locus, every individual has a set that contains elements corresponding to any other individual with whom it maintains an IBD segment at that locus. The problem then becomes that of picking a limited number of individuals such that their corresponding sets cover a maximal number of elements across all loci. We propose a greedy approach, selecting individuals one at a time and gradually admitting samples into the set $\mathbf{Q}$. Table 1 shows the pseudocode for this approach.

Formal algorithm outline

A naive implementation of this greedy approach would be computationally intractable due to maintaining and searching lists of shared regions for each pair of individuals. As the algorithm progresses, such regions keep getting shattered by the interval exclusion operation in step 7 of Table 1. Efficient implementation that maintains these intervals requires a special data structure for quickly calculating overlapping segments. We use a structure known as an “interval tree” for this purpose. Each node in the tree contains an interval representing a shared region along with a pointer to the node in the tree of the other individual with whom the region is shared (INFOSTIP Algorithm, [Supporting Information, File S2](#)). The first step is to calculate $U(i, \mathbf{Q})$, for each individual $i \in \mathbf{P} \setminus \mathbf{Q}$. Our greedy approach now selects the individual j with the highest value of $U(i, \mathbf{Q})$. Upon making the choice, we need to exclude regions that have been imputed by picking j from subsequent selection. These are complete regions r in $R(i, j)$, $i \in \mathbf{P} \setminus \mathbf{Q}$ that we will impute directly by sequencing individual j . Additionally, we will indirectly impute parts of regions r' in $R(i, k)$, $k \in \mathbf{P} \setminus \mathbf{Q}$ that overlap with r in $R(i, j)$. We can then repeat the procedure for selecting and eliminating the next individual. This continues until we have picked individuals up to our sequencing budget b . To better understand the procedure, consider a simple example with three individuals A , B , and C (Figure 1).

Suppose individuals A and B share region (5, 20) (green), B and C share region (15, 25) (red), and A and C share region (30, 50) (blue). We calculate that $U(A, \mathbf{Q}) = 35$, $U(B, \mathbf{Q}) = 25$, and $U(C, \mathbf{Q}) = 30$. The greedy algorithm will first pick A and add it to $\mathbf{Q}$. Next, we exclude directly imputed regions (5, 20) of $R(A, B)$ and (30, 50) of $R(A, C)$. Also, the region (15, 20) in $R(B, C)$ has been indirectly imputed by picking A . The recalculated $U(B, \mathbf{Q}) = U(C, \mathbf{Q}) = 5$ and the algorithm iterates.

The algorithm was implemented in C++ and is available for download at <http://www1.cs.columbia.edu/~itsik/INFOSTIP/readme.html>. All experiments were conducted on a Linux-based cluster controlled by a Sun Grid Engine on a node with 16 GB memory.

Imputation benchmarks

To estimate the practical utility of guided sample selection, we simulated the process of variant imputation by bootstrapping markers genotyped on the array. Hiding a set of markers and imputing them from reference panels selected under different strategies allow us to assess their relative effectiveness. Specifically, using chromosomes 18–22 (~10% of the autosomal genome) of the Kosraen genotype, we randomly selected 5% of polymorphic SNPs to be our target variants for imputation. To ensure that they were amply representative of the minor allele frequency spectrum in the data, variants were selected independently from each window of the spectrum in steps of 1%. We then constructed 11 different subsets of individuals to be used as reference samples for respective imputation simulations and hid the selected markers from the remaining individuals in each simulation. One of the subsets was selected in a guided fashion on the basis of the greedy INFOSTIP priority from the five chromosomes and the remaining 10 subsets were each selected randomly and independently. Imputation was performed using the BEAGLE framework with default parameters, maintaining consistency with the data phasing. Finally, accuracy was measured for each reference set as the average allelic squared correlation (r^2) between the inferred variant call and the true variant call in the population; markers that could not be imputed (or imputed as monomorphic) were counted as having r^2 of zero.

Sequencing benchmarks

We sequenced a discovery panel of seven low-pass personal genomes, four of which (K1955, K2033, K5866, and K1674) were selected according to the aforementioned procedure with the remainder (K6169, K6494, and K5675) chosen on the basis of being phenotype extremes for multiple traits (not shown). For each of the seven individuals, 10–30 mg of genomic DNA was used to generate a library following Life Technologies’ long mate-pair protocol. The libraries were sequenced using the SOLiD system, with 8,239,389,322 raw 50-bp mate-paired reads and an additional 740,209,937 raw 35-bp mate-paired reads, generating a total of 438 Gb. The raw reads were aligned and paired to the

Table 1 Greedy algorithm for selecting sequenced samples

Given: $\mathbf{P}$ population set, $\mathbf{R}$ segment set, b budget

```

1 while  $|\mathbf{Q}| \leq b$ :
2   Find individual  $j \in \mathbf{P} \setminus \mathbf{Q}$  such that  $U(j, \mathbf{Q}) = \arg\max_{i \in \mathbf{P} \setminus \mathbf{Q}} U(i, \mathbf{Q})$ 
3    $\mathbf{Q} \leftarrow \{\mathbf{Q}, j\}$ 
4   for  $i \in \mathbf{P} \setminus \mathbf{Q}$ :
5     Remove any  $r \in \mathbf{R}(i, j)$  from  $\mathbf{R}$ 
6   for  $k \in \mathbf{P} \setminus \{\mathbf{Q}, j\}$ :
7     Remove part or whole of any  $r' \in \mathbf{R}(i, k)$  that overlaps
       with any  $r \in \mathbf{R}(i, j)$ 
8   return  $\mathbf{Q}$ 

```

reference human genome (hg18), using the AB SOLiD Corona Lite pipeline (<http://solidsoftwaretools.com>). Up to three mismatches were allowed for 35-bp reads and up to five mismatches were allowed for 50-bp reads. This generated 158 Gb that map to the genome as uniquely placed normal mate pairs within the expected distance (1.5-kb insert size), order, and orientation. A total of 96.6 Gb of these uniquely placed normal mates are nonredundant, which represents a $>30\times$ coverage of the “Kosraen genome”. On average, 3–6 $\times$ sequence coverage of nonredundant normal uniquely placed pairs was achieved for each individual (Table S1).

Variant calling

Following the structure of the 1000 Genomes Project low-pass pilot, we performed variant calling on all seven samples together, using the Genome Analysis Toolkit (McKenna *et al.* 2010) as well as several steps of imputation. All analyses followed the best practices and parameter suggested by the GATK documentation (File S4). In summary, we performed local realignment and quality score recalibration of the reads from each individual separately; variants in all samples were then called together using an iterative Bayesian algorithm that attempts to infer allele frequency in the population in support of individual genotype calls; for previously known variants, we used a strict call quality threshold to minimize false positives; for novel variants, we performed an additional variant quality score recalibration procedure to minimize expected false positives; finally, we performed internal imputation using the BEAGLE framework (Browning and Browning 2009) and external imputation to the 1000 Genomes pilot haplotypes using MaCH (Li *et al.* 2009). Multisample calling and imputation allows us to leverage the presence of a confidently observed variant in higher-coverage samples to recover calls in lower-coverage samples that would not have been called individually. In recent work, a similar low-pass/multisample strategy has been shown to be effective for maximizing the amount of observed rare variation (Li *et al.* 2011). However, due to the isolated nature of the population and the relative underrepresentation of rare variants in the 1000 Genomes pilot, we expect a number of true variants to remain unseen and quantify this expectation in the following section.

Figure 1 Illustration of INFOTIP-guided sample selection. Sample selection in a single 60-locus chromosome population of three individuals (A, B, and C) is shown. (A) Initial step with pairwise IBD segments represented as colored lines and total utility listed for each individual, with sample A having highest increase. (B) IBD segments after sample A has been selected, with all associated IBD segments grayed out. Samples B and C are tied with utility of 5 shared by a single remaining segment across loci (20–25).

Variant quality control

We performed rigorous quality control on the set of called variants using the available array-based genotypes and additional novel genotyping as validation data (Table 2). Of the previously known sequence calls, 2,958,772 overlapped with the genotyped variants and were used for validation, allowing evaluation of specificity of detecting nonreference sites, as well as of calling each genotype class. We measured the specificity of calling SNPs from all seven samples together on nonreference calls at these known sites to be 98.2%. We independently computed similar levels of specificity, 98.9%, that would be expected on the basis of the observed transition/transversion ratio of 2.07 across all of the previously known variant sites (File S6). Per-call specificity of SNP calling from the samples together varied by sample, with an average of 94.1% for heterozygous calls and 92.5% for homozygous calls (Table S2). Low coverage often caused heterozygous SNPs to be identified as

Table 2 Experimental validation of known and novel variants

Call type	Unique sites	Observed <i>Ti/Tv</i> ratio	Expected <i>Ti/Tv</i> ratio ^a	% calculated variant specificity ^b	Total calls validated	% experimental variant specificity ^c	% sensitivity
Known	4,522,474	2.07	2.10	98.9	2,958,772	98.2 ^d	92.8 ^d
Novel	1,212,831	1.74	2.07	88.6	64	87.5	—

^a From 1000 Genomes high-quality low-pass estimates.

^b Calculated from difference in expected and observed *Ti/Tv* ratio (File S6).

^c Percentage of nonreference calls experimentally validated as nonreference.

^d Sensitivity/specificity of calling SNPs of all seven samples together using GATK Unified Genotyper and imputation on combined low-coverage samples.

homozygous. The low coverage from this pilot, our strict quality thresholds, and the pooling of all seven samples led to a sensitivity to detect nonreference sites of 92.8%. For variants that were not previously known, we observed an overall transition/transversion ratio of 1.74, corresponding to expected specificity of 88.9%. We validated a total of 64 called novel sites using Sequenom genotyping (Table S3) and found the empirical specificity of nonreference calls to be 87.5%, in line with our overall estimates. Additionally, we detailed the array-based validation results at each step of the calling pipeline and found the largest increase in accuracy to come from calling all samples together rather than individually and from internal imputation (Table S2) as previously reported (Bansal *et al.* 2010; Durbin *et al.* 2011; Shen *et al.* 2011).

Results

Expected information content

We estimated total information content expected by sequencing a small reference panel (see *Materials and Methods*) as a function of the sequencing budget (Figure 2). We observe that sequencing 50 randomly chosen individuals (1.7% of the cohort) would give us the potential to impute both alleles of variants in 59.5% of the cohort genome, but choosing individuals in an optimized fashion using INFOSTIP decreases the sequenced sample size needed for the same benchmark by 0.76-fold to 38 individuals (1.3% of the cohort). Remarkably, sequencing only 7 individuals (0.24% of the cohort) still provides imputation capacity of 24% of the cohort genome. For comparison, we conducted the same analysis within a cohort of 1200 Ashkenazi individuals (Barrett *et al.* 2008), a population known to be isolated but less densely related. In this case we found that utilizing our optimal selection method and sequencing 38 individuals gave us the potential to impute variants in only 16% of the cohort genome, whereas sequencing 7 individuals allowed us to infer only 4% of the cohort genome (see Figure 2B, Figure S1, and File S3 for additional analysis). This type of imputation is agnostic of allele frequency as long as a relevant IBD segment is available.

Imputation accuracy

We evaluated imputation accuracy (see *Materials and Methods*) across a range of reference sample sizes from 0 to 600 individuals (Figure 3). In all instances >50 samples, we see

that INFOSTIP selection produces more accurate imputed variants, with an average increase in r^2 of 1.3-fold across reference sizes >100 samples. This quotient makes intuitive sense: r^2 has the traditional interpretation as the ratio of effective to actual sample size (Pritchard and Przeworski 2001), and the increase in r^2 is consistent with our previous theoretical observation that INFOSTIP selection increases the effective sample size by 1.3-fold over the same reference sizes (File S1). This increased accuracy is persistent across all allele frequency windows (Figure S2) with a slightly higher increase in accuracy for low-frequency variants (1.6-fold for variants <10% minor allele frequency). Although this relative accuracy supports the utility of guided sample selection, we stress that the increase is observed using a standard imputation analysis; we expect a thoroughly IBD-sensitive imputation to offer absolute accuracy closer to the expected total information potential.

Variation discovered

We now focus specifically on variants identified in the autosomes, as these are directly applicable to our IBD-based analysis. After quality control, our final set of single nucleotide variants (SNVs) contained 22,221,159 nonreference calls across all seven discovery samples for a total 5,735,305 unique sites of which 1,212,831 (21%) were previously unknown (not in dbSNP v130). The total number of nonreference calls ranges across individual samples from 3.1 million to 3.4 million (Table S4). We expect this to be an incomplete estimate, representing the limitation of low-pass sequencing in calling variants at low-coverage sites due to undersampling of the variant allele. For a fair comparison to other genomes, we extrapolate the total number of variants in the mappable genome on the basis of the error rates described in the previous section (see *Materials and Methods*). Thus, we estimate an average Kosraen sample to contain 3,241,030 total autosomal variants ($\pm 66,996$ SD). Comparing the genome-wide estimates to a variety of published genome sequences (Figure S3), we find the overall number of variants is nearly identical to the 3.25 million observed in average East Asian autosomes (Wang *et al.* 2008; Ahn *et al.* 2009; Kim *et al.* 2009). Comparing the called variants to markers previously annotated in dbSNP v130, we estimate 10.2% (0.90% SD) of the variants in the average Kosraen to be novel, not significantly different from the 9.4% average (3.5% SD) in other East Asian samples (Wang *et al.* 2008; Ahn *et al.* 2009; Kim

Figure 2 Imputation strategy and information capacity in founder population. (A) Schematic outline of the strategy. IBD shared haplotypes (color-coded) are identified from genotype data (gray circles) in a cohort; a small panel of individuals with abundance of IBD is sequenced (top); and sequenced variants within shared haplotypes are inferred to the rest of the cohort within shared regions (bottom). (B) Percentage of genomic content of the cohort that is inferred (total information potential) as a function of sequenced reference panel size, calculated from autosomal genotype data. Rapid growth of total information potential in the highly related Kosraen population (2906 samples, green) is compared to slower growth in the less-related Ashkenazi Jewish population [1200 Crohn's disease case-control samples (Barrett *et al.* 2008), brown].

et al. 2009). Due to the long history of isolation in this cohort, we suspect many of the observed novel variants to be mutations private to the island.

Within an average Kosraen sequence, we find 50.7% of nonreference sites to be homozygous ($\pm 2.2\%$ SD). We caution that this estimate is likely to reflect under-called heterozygotes and expected to drop with deeper sequencing (Wheeler *et al.* 2008). However, it is significantly higher than the observed homozygosity rate in the other personal genomes, even those with similar coverage [next highest, 41.7% in an anonymous Asian (Wang *et al.* 2008)]. Historically, the population experienced a series of severe bottlenecks, which would have resulted in many variants drifting to higher frequency and becoming homozygous.

We estimate the unique novel variation that an arbitrary sequenced individual would contribute by averaging over all 5040 permutations of the seven samples (Figure S4). We find that the first Kosraen sample to be considered contributes approximately one-third of the novel sites that were called in all samples (352,162 of 1,014,310; Table S5), with decreasing contribution of subsequent samples. This decrease, initially by 40% of novel variants that are overlapping between an average pair of samples, becomes more gentle for additional samples, reflecting the enrichment of rare variants in this set, eventually reaching $\sim 50,000$ variants contributed by the last sample only—the average number of single-carrier variants.

Analysis across sequenced populations

We analyzed the population specificity of the Kosraen variants by examining their respective allele frequency in the reference populations sequenced as part of 1000 Genomes Pilot I. Within each of the pilot cohorts of Yoruban

(YRI), European (CEU), and East Asian (JPTCHB) origin, we measured the allele frequency of homozygous variants from the Kosraen (KOS), Korean (SJK), European (JCV), and YRI sequenced genomes. The proportion of variants that fall into each allele-frequency window of a reference cohort is shown in Figure 4. Focusing on the differences between the Kosraens and their closest analyzed neighbor, the SJK Korean genome, we observe that the average Kosraen is relatively enriched for rare variants in all three populations. Specifically, the percentage of alleles in an average Kosraen that were uncommon in JPTCHB ($<10\%$ frequency) was 4.4-fold higher than that of SJK (Figure 4A). In the other cohorts, we see more subtle but consistent enrichment of 1.42-fold and 1.24-fold in uncommon CEU and YRI alleles, respectively (Figure 4, B and C). This trend suggests that lower-frequency alleles in other populations that are present in Kosrae have drifted to higher frequency within the cohort, compared to the Korean genome.

Analysis within the isolate population

We annotated the called variants according to their functionality and analyzed the carrier frequencies of sites that have coding or splicing implications. From the overall frequency distributions (Figure S5), we observe a significant increase in novel coding variants appearing as either singletons or fixed nonreference alleles when compared to novel noncoding variants ($P = 8.7 \times 10^{-10}$ from a χ^2 -test). Focusing on specific coding subclasses in Figure 5, we examine variants called in all samples and annotated as (Figure 5A) splice junction, all coding, synonymous, missense, and nonsense, as well as (Figure 5B) showing all known and novel variants (full data in Table S5, Figure S5). As reported in previous studies (Ng *et al.* 2009; Fujimoto *et al.* 2010; Li

Figure 3 Bootstrapped imputation accuracy for guided and random sample selection. Mean imputation accuracy (allelic r^2 correlation between inferred and true variant) is shown as a function of reference sample size for samples selected randomly ("Random", dotted line) and based on information content ("INFOSTIP", solid line). Error bars for random series show maximum and minimum accuracy from 10 trials of sampled individuals (with replacement). In general, INFOSTIP nearly doubles the effective reference sample size for a given accuracy.

et al. 2010) we see significant overabundance of singleton nonsense mutations compared to other singleton variant classes ($P = 5.4 \times 10^{-4}$ or 1.4×10^{-4} compared, respectively, to coding or all variants). This overabundance is consistent with the effects of purifying selection negatively affecting the frequency of functionally important variants. All of the detected nonsynonymous mutations were significantly enriched for genes with the gene ontology term "olfactory receptor activity" [$P = 4.4 \times 10^{-7}$ after Bonferroni correction (Eden *et al.* 2009); increased enrichment when compared to synonymous mutations], evidence of a continued process of pseudogenization in this family.

Structural variation

We identified short insertions and deletions, using the Dindel algorithm (Albers *et al.* 2011) in all samples together. Briefly, Dindel identifies candidate indels within the read data and then attempts to align them to haplotypes that represent alternative sequences to the reference (detailed protocol in *Materials and Methods*). Figure S6 details the distribution of novel and previously known indels across the seven sequenced individuals. Overall, we observe a steep decrease of indel carrier rate, with 46% of all indels present in a single individual. As with SNVs, the novel indels tend to be enriched for singleton and fixed indels in this cohort when compared to previously known sites.

We identified structural variants >10 kb, using the SOLiD software Tools, which combines depth coverage,

Figure 4 Population-specific genome-wide allele frequency spectrum. Variants previously observed in the 1000 Genomes Project pilot are plotted according to abundance in each sequenced genome (y-axis) as a function of allele frequency in the reference cohort (x-axis). The Kosraen ("KOS") genome is compared to sequenced Korean ("SJK"), Yoruban ("YRI"), and European ("JCV") genomes. Allele frequency spectrum is measured in (top) East Asian (JPTCHB), (middle) African (YRI), and (bottom) European (CEU) origin reference cohorts.

Figure 5 Frequency spectrum of putatively functional variants: histogram of variants annotated by their coding class and novelty. Each bar represents the percentage of variant sites (y-axis) at the respective allele frequency (x-axis) of all observed in the respective variant class. (A) distribution of variants by functional class, showing significant enrichment for singleton nonsense mutations; (B) distribution of previously known (white) and novel (yellow) variants across the entire genome. Only sites where a call could be made in all samples are considered (total counts in Table S1).

predicted mappability, and GC content, within a hidden Markov model framework to make copy number variation (CNV) calls. Overall, an average individual contained 77.1 Mb of copy-variable regions end-to-end, with the longest variant being a 7.6-Mb heterozygous deletion on chromosome 19p13. We analyzed the lengths of the CNVs found in all the samples by variant type and length (Figure S7). In particular, CNVs of size <100 kb constitute 66.9% of the calls, with most being heterozygous deletions. We also looked at the number of shared and private CNVs among the Kosraen individuals, with a CNV being considered shared between two individuals if the overlap between the two called regions was at least 80%. For an average individual, ~20% of CNVs are shared by the entire population.

IBD analysis

Assessing the IBD-based motivation for this pilot, we focused on the 1522 shared segments predicted between the sequenced individuals, ranging in length from 330 kb to 74 Mb. Unlike the conservative INFOSTIP analysis, which examined fewer but higher-quality IBD segments, these segments were detected using GERMLINE's default parameters with no adjustment (3 cM segment length minimum), allowing us to estimate IBD accuracy under practical conditions. We evaluated the accuracy and utility of the IBD-based approach by examining variant concordance within these regions. Specifically, two samples that are IBD across a region should not have sites with homozygous calls for opposite alleles in that region. For a pair of such samples, we examine all sites in the IBD region that are mutually homozygous with at least one sample being nonreference and report concordance as the percentage of these sites that are not homozygous for opposite alleles. Lack of such concordance is indicative of either falsely called IBD or poor genotype calls due to undersampling of sequence reads (true heterozygous sites miscalled as homozygous). Aside from some effects on multisample calling, the concordance rate can be treated as a measure of baseline homozygous variant imputation accuracy when one of the individuals has not been sequenced. Figure 6 shows concordance across IBD segments, separated into previously known (Figure 6A) and

novel (Figure 6B) variants. For comparison, we measured the background distribution of such concordance across 30 random selections of same-sized regions, shown in black points. We observe the vast majority of IBD segments having nearly 100% concordance, with only 1.4% and 10.6% of segments <90% concordance for known and novel variants, respectively. If we take a weighted average across all segments, the aggregate concordance is 99.6% (known) and 97.3% (novel) in IBD segments, providing encouraging estimates for accuracy of IBD-based imputation. This is compared to a background concordance statistic averaging 82.9% (known) and 31.0% (novel) in permuted segments. We attribute the difference between known and novel concordance in IBD regions to be an artifact of lower sensitivity to novel variants and the overall deviation from full concordance to be indicative of inaccurate detection of IBD regions or their exact boundaries.

Of particular interest to the IBD community is the minimum length at which stretches of SNPs identical-by-state (IBS) are still predictive of identity at untyped variants (Powell *et al.* 2010). To estimate this, we omit the minimum segment length restriction for IBD detection, resulting in a tally of all runs of at least 128 IBS SNP-array sites in the sequenced samples, rather than the set of putative IBD regions we considered thus far. We measured concordance in length windows of 1 cM from 0 to ≥ 10 cM. Figure S8 shows this concordance distribution for known and novel variants, as well as the number of segments measured within each window. As previously documented (Gusev *et al.* 2009), we see a direct correlation of concordance with segment length, as longer IBS segments are more likely to represent true recent IBD. However, we observe only a slow decrease in concordance from high-quality 10-cM segments down to 2–3 cM, indicating either a small number of false-positive segments or overcalled false IBD primarily around the boundaries of true IBD segments. Even within the 0- to 1-cM length window (median physical length 815 kb) we see 98.9% (known) and 91.6% (novel) concordance, significantly above the average in non-IBD regions. These initial findings suggest that even very short IBS segments can be useful for variant inference.

Figure 6 Concordance of known and novel variants in IBD and non-IBD regions. We examined concordance of called variants in previously predicted pairwise IBD regions. For all sites that are called homozygous in both samples with at least one being nonreference, we measure concordance (x-axis) as the percentage where both are nonreference. We expect 100% concordance in truly co-inherited regions with no sequence error. The y-axis shows percentage of IBD segments at a given concordance level. Concordance for previously known variants (A, white bar) and novel variants (B, yellow bar) is shown in comparison to randomly placed non-IBD regions of an equal length distribution (both, black points). On average, IBD segments maintained 99.6% (known) and 97.3% (novel) concordance compared to 82.9% (known) and 31.0% (novel) in a background distribution of non-IBD segments.

Discussion

While the population genetics of isolated groups have been of interest for decades, the contribution of such groups to understanding heritable traits is strongly dependent on the research methodology employed. In the context of inbred populations, linkage analysis of Mendelian traits using microsatellite scans has mapped many mutations that are rare in the general population. In contrast, association analysis with SNP arrays relies on linkage disequilibrium in populations and, by primarily targeting common, ancient variation has been mainly applied to outbred peoples. High-throughput sequencing now makes possible discovery of rare variants in the general population but, as shown in this article, with the proper strategy can be applied to the study of isolated communities efficiently and to great effect.

Different strategies for high-throughput sequencing offer various trade-offs of investment and potential for discovery. Whole-genome sequencing at high coverage is the gold

standard, but is still expensive to pursue with substantial sample sizes. Focusing on a captured target, either around a genomic area of interest (Gnirke *et al.* 2009) or considering all exonic regions (Turner *et al.* 2009), sacrifices potential information from most of the genome for high-quality data regarding the most promising parts. Low-pass sequencing offers a different trade-off, considering the entire genome, but accepting lower-quality data. Indeed, the first reference and multiple personal genomes (Lander *et al.* 2001; Levy *et al.* 2007; Wheeler *et al.* 2008) are all low pass, with meaningful insights regarding technology (Shendure and Ji 2008), population genetics (Pool *et al.* 2010), and mutation detection (Mardis *et al.* 2009; Ng *et al.* 2009). This work follows suit and provides population-based sequencing of Pacific Islanders.

With an emphasis on accurate variant detection, we ascertained the effectiveness of low-pass sequencing in conjunction with multisample calling, achieving overall nonreference specificity and sensitivity >90%. In particular, some of the lower-coverage samples netted two- to threefold increases in accuracy when compared to independent calling. Overall, this strategy allowed us to uncover 1,212,831 previously unknown variants with high accuracy.

Examining the spectrum of variation, we explored characteristics unique to this cohort, which had undergone a series of severe bottleneck events. As expected from such an extreme founder population, the qualitative variant statistics reveal an abundance of novel variation and overall homozygosity. Moreover, those sites that have been observed in other sequenced populations still exhibit enrichment for alleles that are rare outside of Kosrae. Demonstrating the effects of purifying selection, we observe a significant abundance of rare coding variants and singleton nonsense mutations compared to all variants and synonymous mutations, respectively.

Leveraging the wealth of relatedness and haplotype sharing in the population, we find 97.3% concordance of novel variants within segments shared IBD by the sequenced samples, demonstrating the potential for inferring such variants in other untyped but IBD individuals. With a high rate of concordance even in very short putative IBD segments, we expect a full panel of 40 sequenced individuals to infer at least 60% of the overall population genome. We caution that such high-inference potential is due largely to the decreased levels of variation in this unique cohort and may not be generalizable to other more diverse isolated populations. Indeed, assessing inference in an isolated cohort of Ashkenazi Jewish origin, we find that the same number of sequenced individuals yields a lower potential at 23% of the population genome. In generalizing these results to a specific cohort, we urge researchers to take demography and location into consideration.

Our work highlights the manageability of population sequencing for isolated populations. While infrastructure efforts by large consortia such as the 1000 genomes lay foundations for comprehensive catalogs of variants in

outbred populations, we demonstrate sequencing at the scale of an individual laboratory as a means to make genetics of such populations fully tractable. As sequencing studies expand geographically to capture the bulk of common variation, isolated populations can help broaden our understanding of rare alleles. While this effort sequenced only a handful of individuals and the sequence coverage of each of them is low, their relation to one another and with many other islanders facilitates both reliable variant calling and powered association analysis to variants detected by full sequencing. This approach avoids the ascertainment bias of previous SNP-based studies and suggests a strategy to leverage SNP array data in large samples, where sequencing is still expensive.

Literature Cited

- 1000 Genomes Project, 2010 Available at <http://www.1000genomes.org>. Accessed April, 2011.
- Ahn, S. M., T. H. Kim, S. Lee, D. Kim, H. Ghang *et al.*, 2009 The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* 19: 1622–1629.
- Albers, C. A., G. Lunter, D. G. Macarthur, G. McVean, W. H. Ouwehand *et al.*, 2011 Dindel: accurate indel calls from short-read data. *Genome Res.* 21(6): 961–73.
- Bansal, V., O. Harismendy, R. Tewhey, S. S. Murray, N. J. Schork *et al.*, 2010 Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res.* 20(4): 537–545.
- Barrett, J. C., S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr *et al.*, 2008 Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* 40: 955–962.
- Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton *et al.*, 2008 Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53–59.
- Browning, B. L., and S. R. Browning, 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84: 210–223.
- Browning, S. R., and B. L. Browning, 2010 High-resolution detection of identity by descent in unrelated individuals. *Am. J. Hum. Genet.* 86: 526–539.
- Durbin, R. M., G. R. Abecasis, D. L. Altshuler, A. Auton, L. D. Brooks *et al.*, 2011 A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.
- Eden, E., R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, 2009 GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10: 48.
- Fujimoto, A., H. Nakagawa, N. Hosono, K. Nakano, T. Abe *et al.*, 2010 Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nat. Genet.* 42: 931–936.
- Galvan, A., J. P. Ioannidis, and T. A. Dragani, 2010 Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends Genet.* 26: 132–141.
- Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust *et al.*, 2009 Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 27: 182–189.
- Gusev, A., J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler *et al.*, 2009 Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* 19: 318–326.
- Hirschhorn, J. N., and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6: 95–108.
- Huang, L., Y. Li, A. B. Singleton, J. A. Hardy, G. Abecasis *et al.*, 2009 Genotype-imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.* 84: 235–250.
- Kenny, E., M. Kim, A. Gusev, J. Lowe, J. Salit *et al.*, 2011 Increased power of mixed-models facilitates association mapping of 10 loci for metabolic traits in an isolated population. *Hum. Mol. Genet.* 20(4): 827–839.
- Kim, J. I., Y. S. Ju, H. Park, S. Kim, S. Lee *et al.*, 2009 A highly annotated whole-genome sequence of a Korean individual. *Nature* 460: 1011–1015.
- Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
- Levy, S., G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern *et al.*, 2007 The diploid genome sequence of an individual human. *PLoS Biol.* 5: e254.
- Li, Y., C. Willer, S. Sanna, and G. Abecasis, 2009 Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* 10: 387–406.
- Li, Y., N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, T. Jiang *et al.*, 2010 Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* 42: 969–972.
- Li, Y., C. Sidore, H. M. Kang, M. Boehnke, and G. R. Abecasis, 2011 Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* 21: 940–951.
- Lowe, J. K., J. B. Maller, I. Pe'er, B. M. Neale, J. Salit *et al.*, 2009 Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. *PLoS Genet.* 5: e1000365.
- Marchini, J., and B. Howie, 2010 Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11: 499–511.
- Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly, 2007 A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39: 906–913.
- Mardis, E. R., L. Ding, D. J. Dooling, D. E. Larson, M. D. McLellan *et al.*, 2009 Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* 361: 1058–1066.
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297–1303.
- McKernan, K. J., H. E. Peckham, G. L. Costa, S. F. McLaughlin, Y. Fu *et al.*, 2009 Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* 19: 1527–1541.
- Newman, D. L., S. Hoffjan, C. Bourgain, M. Abney, R. I. Nicolae *et al.*, 2004 Are common disease susceptibility alleles the same in outbred and founder populations? *Eur. J. Hum. Genet.* 12: 584–590.
- Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham *et al.*, 2009 Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461: 272–276.
- Ng, S. B., K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor *et al.*, 2010 Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.* 42: 30–35.
- Peltonen, L., A. Palotie, and K. Lange, 2000 Use of population isolates for mapping complex traits. *Nat. Rev. Genet.* 1: 182–190.
- Pool, J. E., I. Hellmann, J. D. Jensen, and R. Nielsen, 2010 Population genetic inference from genomic sequence variation. *Genome Res.* 20: 291–300.
- Powell, J. E., P. M. Visscher, and M. E. Goddard, 2010 Reconciling the analysis of IBD and IBS in complex trait studies. *Nat. Rev. Genet.* 11: 800–805.

- Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69: 1–14.
- Sabatti, C., S. K. Service, A. L. Hartikainen, A. Pouta, S. Ripatti *et al.*, 2009 Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* 41: 35–46.
- Shen, Y., Z. Wan, C. Coarfa, R. Drabek, L. Chen *et al.*, 2010 A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res.* 20: 273–280.
- Shendure, J., and H. Ji, 2008 Next-generation DNA sequencing. *Nat. Biotechnol.* 26: 1135–1145.
- Shifman, S., M. Johannesson, M. Bronstein, S. X. Chen, D. A. Collier *et al.*, 2008 Genome-wide association identifies a common variant in the *reelin* gene that increases the risk of schizophrenia only in women. *PLoS Genet.* 4: e28.
- Shmulewitz, D., S. C. Heath, M. L. Blundell, Z. Han, R. Sharma *et al.*, 2006 Linkage analysis of quantitative traits for obesity, diabetes, hypertension, and dyslipidemia on the island of Kosrae, Federated States of Micronesia. *Proc. Natl. Acad. Sci. USA* 103: 3502–3509.
- Turner, E. H., C. Lee, S. B. Ng, D. A. Nickerson, and J. Shendure, 2009 Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* 6: 315–316.
- Wang, J., W. Wang, R. Li, Y. Li, G. Tian *et al.*, 2008 The diploid genome sequence of an Asian individual. *Nature* 456: 60–65.
- Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen *et al.*, 2008 The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452: 872–876.

Communicating editor: N. A. Rosenberg

GENETICS

Supporting Information

<http://www.genetics.org/content/suppl/2011/12/01/genetics.111.134874.DC1>

Low-Pass Genome-Wide Sequencing and Variant Inference Using Identity-by-Descent in an Isolated Human Population

A. Gusev, M. J. Shah, E. E. Kenny, A. Ramachandran, J. K. Lowe, J. Salit, C. C. Lee, E. C. Levandowsky, T. N. Weaver, Q. C. Doan, H. E. Peckham, S. F. McLaughlin, M. R. Lyons, V. N. Sheth, M. Stoffel, F. M. De La Vega, J. M. Friedman, J. L. Breslow, and I. Pe'er

Supporting Note

Figure 1:

There are several metrics of relative improvement from guided sample selection which we will define here, with example measurements illustrated in Figure 1.

The left panel of Figure 1 plots the expected distribution of Total Information Potential (TIP) as a function of the sequenced sample size under guided and random sample selection. Δ_{TIP} measures the ratio of TIP at a given sequenced sample size, corresponding to the increase in proportion of the cohort-genome that is inferred by using INFOTIP. Because INFOTIP chooses the sample with highest inference potential at each iteration, this ratio depends primarily on the non-randomness of genetic relatedness within the cohort; for example, in a homogeneously related or completely unrelated cohort we would expect the ratio would approach 1. This is consistent with our empirical findings that the initially selected samples are of the oldest generation in the pedigree and from each of the distinct island villages. Δ_{TIP} averages 1.16 (s.d. 0.06) across the first 50 selected samples (shown). Along the x-axis, Δ_{seq} measures the ratio of sequenced individuals necessary to achieve a specified TIP between guided and random selection schemes. This metric corresponds to the increase in effective sequenced sample size offered by using INFOTIP. The value is a convolution of the Δ_{TIP} (i.e. heterogeneity of relatedness) and the increase in TIP resulting from each additional sequence sample (i.e. sample uniqueness).

The right panel of Figure 1 plots the empirical distribution of imputation accuracy (r^2) for a subset of hidden variants as a function of reference panel size under guided and random sample selection. The relative accuracy between guided and random reference panels given a reference panel sample size is measured as Δ_{r^2} . Because r^2 follows the ratio of effective sample size over actual sample size, Δ_{r^2} represents the relative increase in effective reference panel size. This metric is therefore analogous to expected Δ_{seq} as defined previously. Although the scale of the distributions is not directly comparable because of differing imputation assumptions, over the sample/reference size 100-400 (in steps of 100) where data is available for both tests and both metrics show enrichment, Δ_{r^2} averages 1.30 (s.d 0.05) and Δ_{seq} averages 1.30 (s.d 0.09) supporting their underlying relationship. Lastly, we measure Δ_{ref} , the ratio of reference panel size required to achieve a specified accuracy under the two sample selection schemes. Δ_{ref} averages to 1.83x in the range of 100-300 samples (where it can be measured), suggesting that INFOTIP selection nearly doubles the effective reference panel size even for current imputation algorithms in this population.

INFOSTIP algorithm implementation

Within the nearly 3,000 sample Kosraen cohort, an average individual shared on an average thirteen million regions per chromosome. The magnitude of the dataset demands efficient management of these shared regions, particularly for operations such as insertion, deletion, querying for overlaps (for finding intersection with imputed regions) and modification. Thus, a data structure was required that scales efficiently with the number of regions. One of the possible choices was a linked list. The linked list provides speed in terms of construction ($O(n)$ time), insertion and deletion ($O(1)$ time) of segments with linear and constant running times respectively. But the bottleneck was querying for overlaps ($O(n)$ time for linked list) and subsequent modification of the regions, where majority of the running time was spent. In some cases the entire linked list may have to be traversed before an overlap is found and the worst case running time is linear in the number of shared regions. Thus, the linked list is not the best solution for this implementation. The interval tree data structure provides a better alternative. An interval tree is an ordered and self-balancing tree data structure that efficiently identifies all intervals overlapping a given point (FIGURE S9). Since a region is represented with a start and end point, it was suitable to be modeled as an interval or node of the tree. Moreover the worst case time to query for an overlap required $O(\log n + m)$ time, where the n refers to the total number of intervals in the tree and m refers to the number of overlapping intervals in the query. This improvement in running time is due to the balanced nature of the interval tree that made it possible to search in only a section of the tree to find the overlapping intervals.

To determine the overall complexity of the algorithm we need to take into account the two main operations; construction of the interval tree and querying for overlap. Construction of an interval tree requires $O(n \log n)$ time, where n represents the average number of shared regions per individual. Querying for overlap requires $O(\log n + m)$ time, with n being the total number of intervals in the interval tree and m being the number of overlapping intervals. Thus, the total complexity can be given as $O(n \log n + \log n + m)$. In terms of space complexity the interval tree requires $O(n)$ space.

IBD segment analysis and Total Information Potential

The pedigree of 2,906 Kosraen individuals was divided into three groups without replacement: two parents and a single child (trio), a single parent and a single child (duo), and single samples (unrelated). Using the BEAGLE framework (BROWNING and BROWNING 2009), the individuals were phased and missing data inferred taking into consideration their respective group structure. The phased genotype data was processed with GERMLINE under default parameters and with genetic distance annotation data corresponding to the Affymetrix 500k chip to generate the genotype-based IBD shared segments. The same data was additionally processed with GERMLINE under the phase-specific haplotype-extension parameters, which explicitly treats each homolog separately in generating matches.

The INFOSTIP analysis was performed on both genotype and haplotype oriented IBD segments. For haplotype data, INFOSTIP executed upon each homolog as if it were an independent set of shared segments, but in choosing a sample for the sequence panel excluded all of the matches originating from that individual on either homolog. As such, a site must be either autozygous, or contained within an IBD segment of two differing sequenced individuals to be fully inferred. For genotype data, INFOSTIP ran with no modification and hence, a site is considered fully inferred if either homolog is in IBD with a sequenced individual. Because the imputed regions were SNP-chip oriented, the total cohort genome length was calculated as the individual end-to-end length of the genome that contained SNPs, multiplied by the number of samples (for genotype data) or twice the number of samples (for haplotype data). FIGURE S1 shows a comparison of the two inference techniques (haplotype, genotype) as well as the two selection methodologies (greedy, random). Due to the non-negligible presence of some autozygosity within the cohort, these two distributions represent an upper and lower bound on the imputation capacity.

Combined SNV calling

We followed the protocol for best practice variant detection detailed in the GATK v2 documentation, with individual parameters tuned for low-pass data. The specific analysis steps are as follows:

1. Reads from all lanes for each sample were merged and duplicate molecules flagged.
2. For each sample, reads were locally realigned around small suspicious intervals (generally indels). For single nucleotide variants, dbSNP v130 and the SOLiD single-sample calling were used; for indels, dbSNP v130 and 1000 Genomes Project Pilot (07/2010) calls were used. Finally, mate pair reads were synchronized.
3. For each sample, base quality scores were recalibrated using per-base covariates: machine cycle for base; di-nucleotide combination for base; number of consecutive previous bases matching this base (accounting for homopolymers); position of the base in the read; the primer round for this base (SOLiD specific).
4. All samples were called together using the UnifiedGenotyper module, which uses an iterative Bayesian likelihood model to estimate allele frequency in the population and genotype calls. We allowed an emitted quality value of 10 and a minimum quality value of 30 for confident calls. Similarly, indels supported by at least 2 reads and at least 60% of all reads were also called for the purpose of masking. Any calls that were low confidence, overlapped a detected indel, or consisted of 3 or more SNVs within 10bp were excluded at this stage.
5. To assess novel variants as accurately as possible, we performed variant quality score re-calibration in three stages. We isolated variants at sites that were expected to be known using the HapMap calls (release 27), 1000 Genomes Project Pilot low-coverage calls (07/2010), and dbSNP (v129, downgraded to minimize new, poorer quality SNPs). We identified clusters within these calls based on quality, allelic balance, strand-bias, and homopolymer run. We then classified all variants according to their expected False Discovery Rate (FDR) given the established cluster boundaries and kept only those novel variants that had expected FDR below 20%.
6. Finally, we performed imputation of un-typed variants in two stages. First, we imputed variants within the cohort using the BEAGLE framework and incorporated any sites with a minimum r^2 cutoff of 0.50. Next, we imputed variants from the 1,000 Genomes (07/2010) haplotypes using MaCH. Processing each chromosome separately, we estimated model parameters from all seven samples and then used these estimates to perform a greedy imputation with default parameters, keeping any previously un-typed sites with a minimum r^2 cutoff of 0.50.

Our final set of calls consisted of high-specificity/low-sensitivity novel variants and aggressively called and imputed known variants. Variants not observed in dbSNP v130 were annotated as novel.

Combined InDel calling

Indels were called using the Dindel v1.01, a program for calling small indels from short-read sequence data. While Dindel does not yet explicitly model multiple independently sequenced samples, we performed the analysis in several rounds and shared the reference information across all samples. As per the user manual, we first generated a list of candidate indels and mate-pair distance distributions for each sample separately using the GATK realigned and recalibrated reads. We then pooled all of the candidate indels into a single reference library and mapped each individual against the pooled library to identify the final set of indels. We classified indels as previously known if they overlapped with any insertion/deletion site in dbSNP v130.

Variant extrapolation

We perform error estimates and variant extrapolation independently for each sample as well as separately for known and novel variants in accordance with the following protocol:

- For previously known variants we measure specificity and sensitivity based on the set of calls overlapping with the genotype array, taken as ground truth. Sensitivity is measured as the percentage of non-reference genotype sites that are called as non-reference in the sequence; specificity is measured as the percentage sequence sites called non-reference that are also called non-reference by the array.
- For novel variants, accurately measuring sensitivity is particularly difficult in low-pass data, and so we conservatively assume sensitivity to be the same as for known variants. For measuring specificity, we assume that the totality of calls is a mixture of true-positive variants with an expected transition bias and false-positive variants occurring randomly and exhibiting no transition bias (WHEELER *et al.* 2008). Formally, given an expected transition rate of δ_{ex} , an observed transition rate of δ_{ob} , a true-positive percentage ϕ_{TP} , and false-positive percentage ϕ_{FP} we establish the following system: $\phi_{TP} + \phi_{FP} = 1$ and $\delta_{ex} \phi_{TP} + \frac{1}{3} \phi_{FP} = \delta_{ob}$ and can solve for specificity as

$$\phi_{TP} = \frac{\frac{1}{3} - \delta_{ob}}{\frac{1}{3} - \delta_{ex}}.$$

We observe that this estimate is very consistent with the empirical specificity in known variants and novel variants

based on experimental validation (TABLE 1).

- We take expected ratios of 2.10 for known sites and 2.07 for novel sites from the GATK variant detection best-practices (<http://www.broadinstitute.org/gsa>), calculated as weighted averages across the 1000 Genomes CEU and YRI trios.

For both variant types we then extrapolate the total expected number of variants in the standard way as expected = (observed) x (specificity) / (sensitivity).

Figure S1 Imputation capacity for genotype versus haplotype schemes. Detailed breakdown of Total Information Potential as a function of sequencing budget for genotype-based (light blue) and haplotype-based (dark blue) inference; categorized according to random selection of samples (dashed line) and greedy optimization of sample selection (solid line).

Figure S2 Bootstrapped imputation accuracy dependent on reference size and allele frequency. Mean imputation accuracy (allelic r^2 correlation between inferred and true variant) as a function of true variant allele frequency is shown for reference samples selected randomly (colored points) and based on information content (solid lines/points). Results are shown for six reference panel sizes from 100 to 600 individuals, with colored concentration bands showing maximum and minimum accuracy from 10 trials of sampled individuals (with replacement). In all cases, sample selection based on information content increases imputed marker accuracy.

Figure S3 Comparison of variants detected in published genomes. Distribution of detected autosomal variants in previously published genomes shown compared to average Kosraen individual (“KOS (avg)” at left). Previously known variants (in dbSNP 130) shown in white bar and novel variants shown in yellow (extrapolation detailed in Supporting Information).

Figure S4 Novel variants attained from arbitrary sequenced individual. Expected number of additional unique novel variant sites gained (y-axis) from each newly sequenced genome (x-axis). Counts were measured across variant sites observed in any sample (grey bar) and variant sites that could be called in all samples (yellow bar). The distribution represents an expectation across all unique permutations of the seven individuals.

Figure S5 Overall distribution of coding and general variants. Each bar represents the percentage of variant sites (y-axis) at the respective allele frequency (x-axis) out of all sites observed in each variant class: all known (white), known coding (gray), all novel (yellow), novel coding (dark yellow). Only sites where a call could be made in all samples are considered (total counts in Table S1).

Figure S6 Distribution of detected short insertions and deletions. Histogram of novel (yellow) and previously known (white) indels across all seven individuals. Number of indels at each carrier level shown at cap of each bar.

Figure S7 Distribution of copy number variants detected. Number and type of CNV identified in four sequenced individuals. Total number of discrete regions called (y-axis) as function of CNV length (x-axis).

Figure S8 Sequence concordance as function of IBS segment length. Concordance of sequenced homozygous variants in array-based IBS regions of increasing length. Previously known (black line) and novel (yellow line) variants shown on left y-axis as a function of segment length on x-axis. Number of segments at each IBD length window shown in grey bars on right y-axis.

Figure S9 Representation of a sample Interval Tree. Each node represents a shared region. The highlighted intervals are returned in a query for intervals overlapping (66,78).

Table S1 Read placement and coverage per sample

ID	Read type	Raw reads	Aligned reads	Aligned mates	Non-redundant mates	Genomic Coverage
K1955	2x50	1,048,656,097	504,557,476	182,789,788	67,174,561	
	2x35	740,209,937	498,459,246	193,481,169	76,709,567	
	<i>sum</i>	1,788,866,034	1,003,016,722	376,270,957	143,884,128	4.32
K2033	2x50	1,027,824,174	523,272,510	197,496,500	149,412,266	5.34
K5866	2x50	1,012,198,246	558,930,954	220,603,222	123,969,777	4.43
K1674	2x50	959,926,988	513,403,455	194,480,787	135,734,837	4.85
K6169	2x50	1,786,661,548	671,728,313	225,801,169	169,854,108	6.07
K6494	2x50	1,727,132,405	626,281,340	207,118,404	178,149,982	6.36
K5675	2x50	676,989,864	298,728,860	215,139,732	87,855,892	3.14

Table S2 Array-based quality control in filtered sequenced variants

		Single sample		Raw multi-sample		Filtered multi-sample		BEAGLE internal		MaCH: 1000 Genomes	
	Sample	Specificity	Sensitivity ¹	Specificity	Sensitivity ¹	Specificity	Sensitivity ¹	Specificity	Sensitivity ¹	Specificity	Sensitivity ¹
Reference	K6169	80.2%	99.91%	90.1%	97.90%	89.8%	97.95%	91.3%	98.00%	92.8%	97.08%
	K2033	91.4%	99.98%	90.8%	99.86%	90.5%	99.87%	91.9%	99.79%	93.1%	98.96%
	K5675	62.8%	99.99%	76.5%	99.91%	75.9%	99.93%	79.1%	99.02%	81.3%	97.81%
	K1955	70.7%	99.99%	84.9%	99.89%	84.5%	99.90%	86.3%	99.42%	88.0%	98.43%
	K1674	80.8%	99.98%	89.2%	99.85%	88.9%	99.87%	90.5%	99.77%	92.0%	98.88%
	K5866	79.4%	99.98%	86.7%	99.85%	86.4%	99.86%	88.6%	99.66%	90.2%	98.73%
	K6494	81.0%	99.98%	91.1%	99.88%	90.7%	99.89%	92.0%	99.84%	93.3%	99.01%
	Sample	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
Heterozygous	K6169	99.1%	49.8%	88.5%	82.3%	88.5%	81.1%	89.2%	83.8%	87.9%	87.4%
	K2033	99.9%	72.5%	98.2%	79.1%	98.2%	77.6%	98.2%	81.1%	96.3%	85.1%
	K5675	99.8%	7.4%	98.3%	36.1%	98.5%	35.0%	94.8%	53.0%	91.5%	57.5%
	K1955	99.9%	21.4%	98.4%	60.1%	98.4%	58.5%	97.0%	68.6%	94.7%	73.3%
	K1674	99.9%	44.8%	98.1%	77.4%	98.1%	75.7%	98.0%	79.8%	96.1%	84.0%
	K5866	99.9%	38.2%	98.4%	70.3%	98.4%	68.5%	97.8%	74.7%	95.8%	79.5%
	K6494	99.9%	54.4%	97.8%	82.4%	97.8%	81.3%	98.0%	84.2%	96.2%	87.4%
	Sample	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
Homozygous Variant	K6169	88.0%	83.2%	95.0%	91.5%	95.0%	91.3%	95.9%	93.1%	95.6%	93.2%
	K2033	92.8%	96.8%	93.9%	95.8%	93.9%	95.6%	95.0%	97.5%	94.7%	97.7%
	K5675	77.8%	47.7%	80.9%	80.1%	81.0%	79.9%	85.2%	94.0%	84.7%	94.4%
	K1955	80.6%	68.0%	87.9%	88.9%	88.0%	88.6%	90.2%	96.2%	89.8%	96.4%
	K1674	85.6%	86.5%	93.2%	95.6%	93.2%	95.3%	94.6%	97.3%	94.3%	97.5%
	K5866	83.2%	86.2%	90.4%	94.6%	90.5%	94.3%	92.5%	97.1%	92.2%	97.3%
	K6494	90.1%	82.2%	95.4%	95.5%	95.4%	95.3%	96.3%	97.2%	96.0%	97.3%

1: Sensitivity of calling SNPs of all seven samples together measured as percentage of genotyped reference alleles called as non-reference or no-call

Table S3 Experimental validation of called novel variants

	Total	Validated Ref	Validated Het	Validated Var	Call specificity	Non reference specificity
Called Het.	28	4	24	0	86%	87.5%
Called Hom.	36	4	14	18	50%	

Table S4 Called and extrapolated variants in seven Kosraen samples

Sample	Total Called	Known hom	Known het	Novel hom	Novel het	Total expected ¹
K1674	3,241,805	1,587,554	1,227,402	88,395	338,454	3,282,447
K1955	3,082,651	1,628,324	1,073,233	87,563	293,531	3,214,836
K2033	3,213,192	1,606,658	1,187,234	90,701	328,599	3,232,382
K5675	2,814,993	1,638,929	841,377	91,277	243,410	3,106,023
K5866	3,197,269	1,612,003	1,177,740	91,720	315,806	3,266,299
K6169	3,389,746	1,500,652	1,393,928	80,226	414,940	3,304,190
K6494	3,281,503	1,566,269	1,243,493	87,240	384,501	3,281,031

¹Extrapolated from known/novel variant error rates (Supplementary Methods)

Table S5 Putatively functional variants detected, called in all samples, and variant in a single sample

Variant type:	Unique sites	Called in all	Singletons
All	5,735,305	5,035,424	15.2%
Annotated	4,522,474	4,021,114	11.3%
Novel	1,212,831	1,014,310	30.6%
All coding	31,229	26,281	17.2%
Synonymous	14,659	12,219	17.0%
Missense	15,843	13,447	17.1%
Nonsense	751	615	21.0%
Splice variants	2,785	2,356	16.3%