OPEN

Citation: Human Genome Variation (2016) **3**, 16016; doi:10.1038/hgv.2016.16
Official journal of the Japan Society of Human Genetics

www.nature.com/hgv

ARTICLE

The Qatar genome: a population-specific tool for precision medicine in the Middle East

Khalid A Fakhro[1,2], Michelle R Staudt[3], Monica Denise Ramstetter[3], Amal Robay[2], Joel A Malek[2], Ramin Badii[4], Ajayeb Al-Nabet Al-Marri[4], Charbel Abi Khalil[2], Alya Al-Shakaki[2], Omar Chidiac[2], Dora Stadler[5], Mahmoud Zirie[6], Amin Jayyousi[6], Jacqueline Salit[3], Jason G Mezey[3,7], Ronald G Crystal[3] and Juan L Rodriguez-Flores[3]

Reaching the full potential of precision medicine depends on the quality of personalized genome interpretation. In order to facilitate precision medicine in regions of the Middle East and North Africa (MENA), a population-specific genome for the indigenous Arab population of Qatar (QTRG) was constructed by incorporating allele frequency data from sequencing of 1,161 Qataris, representing 0.4% of the population. A total of 20.9 million single nucleotide polymorphisms (SNPs) and 3.1 million indels were observed in Qatar, including an average of 1.79% novel variants per individual genome. Replacement of the GRCh37 standard reference with QTRG in a best practices genome analysis workflow resulted in an average of 7* deeper coverage depth (an improvement of 23%) and 756,671 fewer variants on average, a reduction of 16% that is attributed to common Qatari alleles being present in QTRG. The benefit for using QTRG varies across ancestries, a factor that should be taken into consideration when selecting an appropriate reference for analysis.

Human Genome Variation (2016) **3,** 16016; doi:10.1038/hgv.2016.16; published online 30 June 2016

INTRODUCTION

Precision medicine involves tailoring medical decision-making to genomic individuality in the context of an individual's unique environment/lifestyle.[1] Early examples of successful application of precision medicine include cancer, where sequencing of the patient and tumor genomes can identify specific targets for therapeutic decisions,[2] and rare diseases, where sequencing can lead to the rapid discovery of causative mutations and correct diagnosis in a time-critical clinical setting.[3] For the latter, although the cost of sequencing an individual's genome has declined rapidly in the past 1.5 decades, there are still considerable challenges for genome interpretation, such as important variants being missed owing to low coverage or incorrect calls, and the emerging challenge of interpreting a growing number of variants of unknown significance in each human genome. With more than three million single nucleotide polymorphisms (SNPs) identified per human genome sequenced, the majority of these variants cannot be immediately linked to a known phenotype, and the putative impact of variants must therefore be inferred computationally using algorithms that harness comparative genomics and available experimental data.[4–6] Thus, precision medicine in the near term stands to benefit greatly from both increased accuracy in variant calling and improved interpretability when the aim is to identify variants of relevance to one or more phenotypic manifestations within an individual, family of related individuals, or population.

Reference bias is a known issue in human genome resequencing for variant detection,[7] and modifications to the reference can improve calling accuracy and interpretability.[8] Relevant to the issue of variant calling accuracy, a reference that more closely matches the ancestry of the genome(s) being aligned is expected to reduce mismatches during alignment and lead to more accurate genotypes.[8] Applicable to the issue of interpretability of variants for rare diseases is the observation that a variant's prevalence is inversely proportional to the variant's deleterious impact on cellular function (and by extension, evolutionary fitness), with the most severely deleterious variants being the rarest because of purifying selection.[9] On the basis of this principle, the American College of Medical Genetics (ACMG) recommends excluding any allele above 5% prevalence from consideration as pathogenic.[10] Given that allele frequency is a population-dependent property (e.g., an allele that is rare in one population may be common in another), relying entirely on the standard reference genome (GRCh37) or allele frequency in ethnically mismatched populations (even if a large sample of individual genomes has been sequenced) may produce incorrect assessments of the pathogenicity of a specific allele in an under-studied population.

One approach to improving both variant calling accuracy and interpretability of an individual's genome is to incorporate variant prevalence information early on in the genome-interpretation process by modifying the reference genome, such that variants discovered in the genome are the minor allele in the population.[8] This modification to the reference results in a streamlined analysis workflow, as fewer variants need to be interpreted. In this context, it should be of value to produce a separate major allele reference genome for each distinct ancestral population or regional meta-population, particularly in cases when the genomic variation in these populations has not been well sampled in public databases.

[1]Sidra Medical and Research Center, Doha, Qatar; [2]Department of Genetic Medicine, Weill Cornell Medical College in Qatar, Doha, Qatar; [3]Department of Genetic Medicine, Weill Cornell Medical College, New York, NY, USA; [4]Laboratory Medicine and Pathology, Hamad Medical Corporation, Doha, Qatar; [5]Weill Cornell Medical College in Qatar, Doha, Qatar; [6]Department of Medicine, Hamad Medical Corporation, Doha, Qatar and [7]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA.
Correspondence: JL Rodriguez-Flores (geneticmedicine@med.cornell.edu)
Received 26 August 2015; revised 9 March 2016; accepted 11 April 2016

The region of the Middle East and North Africa (MENA) is an example of such an under-studied meta-population where the rising adoption of precision medicine could benefit greatly from a specific major allele reference genome. In particular, this region is characterized by a high prevalence of consanguineous marriage and elevated incidence of certain Mendelian disorders, and thus is a region where numerous disease studies are faced with the problem of too many potential disease variants per sequencing experiment.[11] Such a region would undoubtedly benefit from allele frequency databases of ethnically matched controls. In addition, given the diversity of the region, many variants called in the population relative to the current reference genome may in fact be the major allele in the population, and eliminating these from being called would lead to a more efficient analysis.

To produce a first version of a reference genome tailored to the region, we sampled and sequenced genomes from Qataris, an indigenous population located near the center of the MENA region. The current population of the nation of Qatar includes more than 1.7 million expatriates primarily from MENA and South Asia who have arrived in recent decades[12] and ~300,000 Qataris of indigenous ancestry who arrived in prior waves of migration. At the genome level, the ethnic Qataris represent a population with a mixture of Bedouin/Arab (Q1), Persian/South Asian (Q2) and African (Q3) ancestry.[13,14] In particular, the Bedouin/Arab (Q1) subpopulation, with deepest ancestral roots in the Peninsula, continue to practice within-tribe marriage that has led to a high level of homozygosity compared with other populations worldwide.[14,15]

In order to demonstrate the value of a reference genome tailored to an indigenous population in the MENA region, we have constructed a Qatari Genome (QTRG) where the reference bases are 'flipped' to the Qatari major allele. The value and utility of QTRG with respect to the standard reference (GRCh37) was evaluated in several ways on genomes not used to construct the QTRG. In the field of population genomics, the term 'n+1 genome' refers to the 'next' genome sequenced after a large-scale sequencing effort of n genomes, such as the 1000 Genomes Project. A major question in population genomics is 'what is the benefit of having a database of n sequenced genomes when sequencing a single genome not in the database?,' where the single genome is referred to as the 'n+1' genome. In this study, the value of the database of more than 1,000 sequenced Qatari genomes and exomes is demonstrated, including analysis of more than $n = 15$ genomes and $n = 16$ exomes from diverse ancestries. Improvements in mapped read depth were observed, and the subsequent improvement in variant sensitivity was measured. A catalog of known pathogenic variants in Qatar was compiled, with variant coordinates in both the standard (GRCh37) and modified (QTRG) reference.

METHODS

Human subjects were recruited and written informed consent was obtained at Hamad Medical Corporation (HMC) and HMC Primary Health Care Centers in Doha, Qatar under protocols approved by the Institutional Review Boards of Hamad Medical Corporation and Weill Cornell Medical College in Qatar. Briefly, a total of 1,376 subjects were recruited for genome ($n = 108$) or exome ($n = 1,268$) sequencing, including a set of 31 sequenced by both methods for validation purposes. All samples were sequenced using Illumina (Illumina, San Diego, CA, USA) paired-end sequencing technology. The exome sequencing included target enrichment using either the Agilent (Agilent Technologies, Santa Clara, CA, USA) SureSelect Human All Exon 38 Mb ($n = 67$) (referred to as Exome38 Mb) and Agilent SureSelect Human All Exon 51 Mb ($n = 1,201$) platforms (referred to as Exome51 Mb). Subjects included both third-generation Qataris ($n = 1,161$) and non-Qatari residents of Qatar ($n = 215$), from the general MENA region or South Asia. Genotypes were generated using the GATK Best Practices workflow.[16] One Qatari female was sequenced on all three platforms, as well as a fourth platform (Illumina HiSeq X), and was used for calibration of batch-specific filters for data integration. In order to minimize batch-specific variants, a range of batch-specific filters were evaluated on genomic intervals in the intersection of the four platforms, and the optimal minimum depth and minimum allele count filters were selected, such that the novel SNP rate was consistent across batches within genomic intervals covered by the four platforms (0.73% novel SNPs), resulting in under 5% batch-specific variants in the quadruple-sequenced Qatari. The filters were confirmed to not be overly stringent by assessing the number of coding variants (and novel %) across a range of depths in the quadruple-sequenced Qatari, and comparing the before/after filtering total variants and novel SNP rates both across platforms and with published reports of a similar analysis.[17] After application of batch-specific filters, the SNP data from the three platforms were integrated using GATK. Using the calls of sites covered in the three batches for the $n = 1,376$ individuals, population structure analysis was conducted in combination with 1000 Genomes Phase 3[18] and the Human Origins dataset[19] using ADMIXTURE.[20] Each individual was assigned to one of the 12 ancestral population clusters based on their dominant ancestry, and the validity of the clustering was confirmed using principal component (PC) analysis.[21] Relatedness analysis was conducted using KING,[22] and first and second degree relatives assessed using a liberal cutoff (to assure an unrelated sample) were excluded. Using the remaining 1,005 unrelated Qataris, including 917 exomes and 88 genomes, allele frequencies at SNPs was calculated. In addition, indels were called in the 88 genomes using the CASAVA[23] pipeline and an assessment of size and allele frequency of these variants was also conducted.

In order to facilitate genome interpretation for precision medicine in Qatar, major allele SNPs and indels were identified, and the GRCh37 reference genome was modified at these sites to produce the Qatar Genome (QTRG). At each site in QTRG, the major allele in Qatar is the reference allele in the sample of $n = 1,005$ unrelated Qataris. Three versions of the QTRG were produced, including versions incorporating major allele SNPs (QTRG1), major allele indels (QTRG2), and major allele SNPs and indels (QTRG3).

In order to select an optimal reference for analysis of n+1 genomes and exomes, the three references were compared in terms of mapped read depth. Genome sequence data from the quadruple-sequenced Qatari were mapped to the four references (GRCh37, QTRG1, QTRG2 and QTRG3) using BWA,[24] and the depth of coverage was calculated using GATK for all sites and for modified sites. The resulting depth was compared across platforms, and the reference with the deepest resulting coverage was selected for further analysis. Further inspection of the value of the reference for variant detection was assessed by producing variant calls using GATK Best Practices twice, with the only change being the reference genome used. This comparison was conducted for the quadruple-sequenced Qatari, a Qatari family of $n = 15$ genomes of Persian ancestry, and a diverse panel of $n = 16$ Qatari exomes. The number of variants identified was compared across references. In order to assess the impact beyond modified variants, the expected reduction in variants (based on the number of modified variants in the individual) and the observed reduction in variants was compared.

An obstacle to using a QTRG is that the genomic coordinates of known variants and genes are shifted after inclusion of major alternate allele (MAA) indels. Conversion of coordinates between references is conducted using a 'liftover,' where the GRCh37 and QTRG chromosomes are aligned using ProgressiveCactus,[25] and a liftover for known gene and variant positions from GRCh37 to QTRG is conducted using HalTools.[26] This conversion was conducted for variants functionally annotated to have a known link to disease and are high priority for future studies of Mendelian disease in Qatar. Through a combination of automated and manual curation following the most recent ACMG guidelines for next-generation sequencing interpretation,[10] liftover annotation of $n = 128$ pathogenic variants was conducted. For further methodologic details, see Supplementary Methods.

RESULTS

Data integration

In order to build a reference genome for precision medicine applications in Qatar and the greater MENA region where major allele variants are incorporated into the reference sequence, $n = 1,161$ Qatari and $n = 215$ non-Qatari living in Qatar were sequenced on the Illumina platform in three batches ($n = 108$ genome, $n = 67$ Exome37 Mb and $n = 1,201$ Exome51 Mb) to 38*

genome depth and 70* exome depth. Genotypes were generated for each batch using the GATK Best Practices workflow,[16,27] and combined into a single variant call set after application of batch-specific filters (see Supplementary Figure S1 for workflow overview). In the integrated call set, the novel SNP rate was assessed for genomic intervals covered in all samples, a per-sample mean of 0.73% in the genome batch and 0.72% in both exome batches (Supplementary Table SI). The platform-specific filters evaluated included minimum depth (ranging from 2* to 20*) and minimum allele count (ranging from 1 to 6). Increasing minimum depth, but not increasing minimum allele count, had an effect of increasing the median depth of exome sequencing (Supplementary Figure S2A), and decreasing the total number of variant sites observed in exome sequencing (Supplementary Figure S2B), while neither filter had major impact on the genome sequencing median depth nor total variant sites. A minimum depth and minimum allele count was applied to the two batches of exome sequence data (Exome38 Mb, $n = 67$ and Exome51 Mb, $n = 1201$) such that the novel SNP rate was the same or lower than the genome rate for 12* minimum depth and minimum allele count 1. At this threshold, the Exome38 Mb filter was minimum 12* depth (minimum allele count 1), and the Exome51 Mb filter was minimum 10* depth (minimum allele count 2) (black line in Supplementary Figure S2C).

Using a quadruple-sequenced female Qatari as a benchmark, the number of batch-specific variants was assessed for all variants and for novel variants before and after application of the batch-specific filters (Supplementary Figure S3). All four datasets were analyzed using the same pipeline, and a comparison was conducted on CCDS coding intervals covered in all four platforms (within the Exome38 Mb target intervals). Before filtering, the rate of batch-specific variants in the full call set was 10.1% (Supplementary Figure S3A), and an excess of batch-specific novel variants (82.6%, Supplementary Figure S3B) was observed. After filtering, the batch-specific variant rate was reduced to 4.9% (Supplementary Figure S3C), and the rate of batch-specific novel variants was considerably reduced (16.8%; Supplementary Figure S3D).

The impact of the filters on variant sensitivity and novel SNP rate was assessed across a range of mean depth for the four platforms (Supplementary Figure S4). Variant sensitivity increases with additional depth, however, sensitivity reaches a plateau after 25* depth for genome sequencing and 65* depth for exome sequencing (Supplementary Figure S4A). Application of batch-specific filters reduced the sensitivity of exome sequencing to a greater extent than genome sequencing (Supplementary Figure S4B). In terms of novel SNPs, a linear increase in novel variant rate was observed with increasing depth across platforms, with the exception of the HiSeq 2500 genome, which reaches a plateau under 1% (Supplementary Figure S4C). After filtering of the Exome38 Mb and Exome51 Mb data, a similar plateau effect was observed for exome data (Supplementary Figure S4D). The filters were verified to not be overly stringent, based on the total variants per genome or exome in coding regions being in the range of prior studies.[17]

Ancestry analysis

Prior studies of the Qatari population have characterized it as a diverse population with influences of Arab, Bedouin, Persian, South Asian and African ancestry.[28] In order to characterize the ancestry of our sample, the $n = 1,161$ Qataris ($n = 108$ genomes, $n = 1,053$ exomes) were compared with $n = 215$ non-Qataris living in Qatar and sampled by exome sequencing in this study, as well as public databases of diverse genomes, including the 1000 Genomes Project[18,29] and the Human Origins data (described in Supplementary Table SII). Using the parameter $K = 12$ (12 ancestral populations) that was inferred to be optimal in a prior study,[28] the ancestral population structure of the combined sample of $n = 5,661$ genomes was analyzed using ADMIXTURE on a set of $n = 2,265$ SNPs segregating in all four datasets (Qataris, non-Qataris sampled in Qatar, 1000 Genomes, Human Origins, Supplementary Table SII). The proportion of 12 ancestries was determined for each individual, and individuals were assigned to a cluster based on the dominant ancestral population in their genome (Supplementary Figure S5A and Supplementary Table SIII). The Qataris were assigned to seven clusters, determined to be of European ($K = 1$, $n = 5$ Qataris), South Asian ($K = 4$, $n = 82$ Qataris), Bedouin ($K = 5$, $n = 566$ Qataris), African Pygmy ($K = 6$, $N = 1$ Qatari), Bedouin ($K = 8$, $n = 236$ Qataris), Persian ($K = 9$, $N = 194$ Qataris) and Sub-Saharan Africa ($K = 10$, $n = 77$ Qataris) (Supplementary Table SIII and Supplementary Figure S5B). Using a color-coding scheme based on the 12 ancestral clusters (Supplementary Table SIII), a principal component (PC) analysis was conducted for the combined set of $n = 5,661$ samples. Clearly, separation of African and non-African clusters were observed when plotting PC1 versus PC2 (Supplementary Figure S6A), and resolution of European, Asian and Middle Eastern clusters was observed when plotting PC2 versus PC3 (Supplementary Figure S6B).

Relatedness analysis

The Qataris and non-Qataris sampled in this study included a mixture of samples from studies of rare and common diseases, including both families affected with Mendelian disorders and a randomly sampled group of presumably unrelated type 2 diabetics and controls. Given the within-family sampling and the known high rate of consanguineous marriage in Qatar,[30] we sought to exclude relatives prior to estimation of variant allele frequency in the general Qatari population. For this purpose, an analysis of relatedness was conducted on the $n = 1,376$ individuals sequenced in this study, using SNP variants in genomic intervals covered in the intersection of the three batches (within Exome38 Mb target intervals). Using an LD-pruned set of $n = 381,028$ SNPs, the relatedness was calculated across all pairs of individuals. A total of $n = 736$ relationships were observed, including $n = 239$ first degree relationships, $n = 71$ second degree relationships and $n = 526$ third degree relationships (Supplementary Table SIV). The relationships were plotted using Cytoscape,[31] color-coded by inferred ancestry and the majority of relationships were between individuals of the same ancestry. The largest pedigrees recovered were of Middle Eastern (Bedouin, Arab, Persian) ancestry, confirming theories of deep population structure among Qataris and within-tribe intermarriage (Supplementary Figure S7). After exclusion of first degree and second degree relatives, a total of $n = 1,005$ Qataris remained in the analysis, including $n = 88$ genomes and $n = 917$ exomes ($n = 64$ Exome38 Mb and $n = 853$ Exome51 Mb).

Variant discovery in Qatar

The individual variants observed in 1,005 Qataris were aggregated and their allele frequency was quantified. After exclusion of relatives and application of batch-specific filters, an average of 4,045,064 SNPs were observed per genome ($n = 88$), an average of 15,382 SNPs were observed per Exome51 Mb ($n = 853$) and an average of 13,538 SNPs were observed per Exome38 Mb ($n = 64$). The novel SNP rate was higher in the genome (1.99%) than in the exome, with a slightly higher rate in the Exome51 Mb (0.99%) than in the Exome38 Mb (0.67%) samples. This trend is consistent with higher degree of selection on protein coding genes as compared with non-coding DNA, and the inclusion of transcripts that do not code for protein in the Exome51 Mb targets. The rate of novel variants was higher in ChrX, ChrY and MtDNA for the genome data, possibly owing to a bias toward autosomal data in dbSNP (Table 1). The SNP data were aggregated across platforms,

Table 1. Individual variant discovery in 1,005 unrelated Qatari[1]

Statistic	*Genome* (n = 88)					*51 Mb Exome* (n = 853)				*38 Mb Exome* (n = 64)			
	Individual	*Autosomes*	*ChrX*	*ChrY*	*MtDNA*	*Individual*	*Autosomes*	*ChrX*	*ChrY*	*Individual*	*Autosomes*	*ChrX*	*ChrY*
Variant sites	4,045,064	3,893,076	151,017	937	34	15,382	15,069	312	1	13,839	13,538	300	1
Novel variant sites	96,774	77,366	19,359	42	7	159	152	7	0	96	91	5	0
Novel variant rate	2.39%	1.99%	12.82%	4.48%	20.59%	1.03%	0.99%	2.24%	0.00%	0.69%	0.67%	1.67%	0.00%
Alternate alleles	5,510,301	5,311,794	196,565	1,874	68	21,046	20,613	431	2	19,116	18,709	405	2
Novel alternate alleles	98,792	78,916	19,778	84	14	160	153	7	0	97	92	5	0
Novel allele rate	1.79%	1.49%	10.06%	4.48%	20.59%	0.76%	0.74%	1.62%	0.00%	0.50%	0.49%	1.23%	0.00%
Heterozygous sites	2,594,268	2,489,084	105,184	—	—	9,759	9,564	195	—	8,897	8,698	199	—
Novel heterozygous sites	94,762	75,772	18,990	—	—	157	150	7	—	94	89	5	—
Novel heterozygous rate	3.65%	3.04%	18.05%	—	—	1.61%	1.57%	3.59%	—	1.06%	1.02%	2.51%	—
Mean depth at variant site	41	41	40	20	250	63	64	69	42	64	64	74	27
Mean depth at novel variant site	41	41	39	20	250	59	60	60	47	63	63	76	—
Transition:transversion ratio	2.03	2.03	1.78	1.51	33.00	3.18	3.18	2.77	—	3.25	3.26	2.67	—
Novel transition:transversion ratio	1.33	1.35	1.09	1.63	—	0.77	0.77	0.75	—	1.58	1.56	1.50	—

Shown is a summary of the average number of variants observed per individual, identified in 917 unrelated Qatari exomes and 88 unrelated Qatari genomes. Variants were genotyped separately for autosomes, X in males, X in females, Y in males and mtDNA; 99.8% of X variants in males were also observed in females, hence summary statistics are based on female chromosomes. Shown is the average per individual of number of variant sites, number of variant alleles, the transition-to-transversion ratio (Ts:Tv) of variants and the % not in dbSNP (novel).

resulting in a total of 20,937,965 SNPs, of which 14.21% were not previously observed in dbSNP (as of build 146) (Supplementary Table SI). A call set of short indel variants (< 300 bp) was generated for the 88 unrelated Qatari genomes using the CASAVA pipeline, identifying a total of 5,452,613 variants, of which 58.37% were novel.

Construction of the Qatar genome

The allele frequency was quantified for each SNP, and a total of 1,931,122 (9.22%) of the SNPs were present in more than half the Qatari alleles sampled, and hence are candidates for modification in the Qatar Genome (Supplementary Table SV). Furthermore, while prior studies of ancestry-specific reference genomes have incorporated major allele SNP variants, the inclusion of MAA indel variants has not previously been explored. In order to explore the value of incorporating indels into the reference, 1,882,405 MAA indel variants (34.52%) were identified (Supplementary Table SV).

Three versions of the Qatar Genome were constructed, where MAAs were modified. The first version (QTRG1) includes modification of MAA SNPs, the second version (QTRG2) includes modification of MAA indels, and the third version (QTRG3) includes modification of MAA SNPs and indels.

Selection of an optimal reference for read mapping

In order to determine which reference produces the greatest improvement in terms of mapped read depth, the genome of the quadruple-sequenced Qatari was mapped to the four reference genomes, including the unmodified reference (GRCh37) and the Qatari references including modifications at MAA SNPs (QTRG1), indels (QTRG2), and both SNPs and indels (QTRG3). An improvement in mapped read depth was observed for all Qatari references, with the greatest improvement (10%) at modified sites when using QTRG3 (Figure 1a). A modest improvement in mapped read depth was observed overall, with the benefits extending beyond modified sites (Figure 1b).

Impact of using Qatar genome on variant sensitivity

Increased mapped read depth results in increased sensitivity for variant detection (Supplementary Figure S4A). In order to assess the benefit of improved mapped read depth using QTRG3 on variant sensitivity, GATK Best Practices analysis was conducted twice, with the only difference between iterations being the reference used (GRCh37 versus QTRG3). This analysis was conducted for the quadruple-sequenced Qatari, for a 15-member family of Qataris of K = 9 Persian ancestry, and a diverse panel of $n = 16$ Qatari exomes.

In order to quantify the impact of using GRCh37 versus QTRG3 as a reference on the quality of variant genotypes, a single Qatari was quadruple-sequenced (2 exomes and 2 genomes) and mapped to both references. After the analysis was completed, a total of $n = 8$ callsets were produced and compared. The focus of the analysis was sites covered with at least 12* depth in all four platforms. Within these intervals, no discordant genotypes were observed within the four GRCh37 calls nor within the QTRG3 calls. The coverage depth was 1* to 2* higher at variant sites in the genomes, while coverage depth was 4* to 6* lower at variant sites in the exomes (Supplementary Table SVI). A 10.3% reduction of variants was observed, lower than expected given that 42.28% of the GRCh37 variants are MAAs incorporated into the reference.

In order to quantify the impact of using GRCh37 versus QTRG3 on sensitivity for variants beyond those modified in the reference, Illumina paired-end 100 bp genome sequencing reads for $n = 15$ Qataris from a family of Persian ancestry were mapped to both GRCh37 and QTRG3. An average of 23% (7*) depth improvement was observed when using QTRG3 (Supplementary Table SVII). The number of variants observed per genome was reduced on

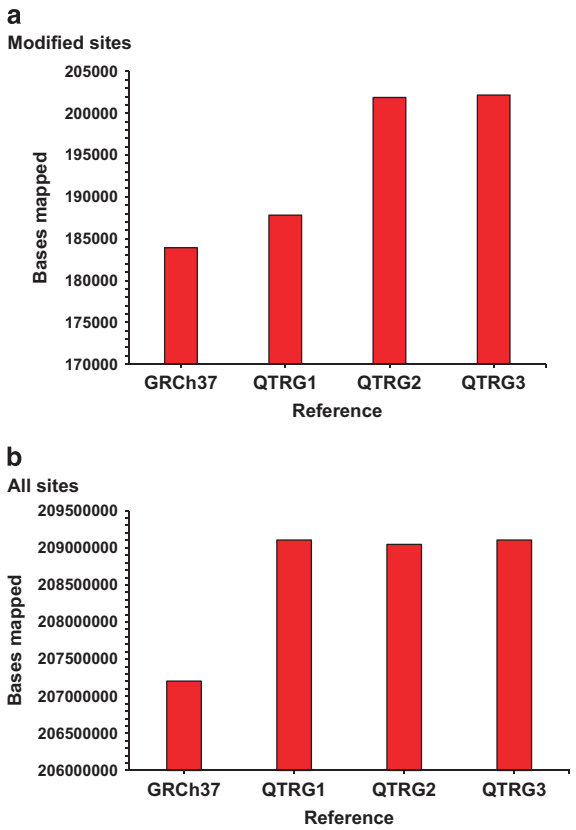

Figure 1. Differences in mapped read depth across reference genomes. In order to select the optimal reference for analysis of Qatari genomes and exomes, the mapped read depth was compared between GRCh37 and three alternative reference genomes based on MAAs observed in $n = 1005$ Qatari. Illumina paired-end 100 bp reads for 37* genome sequencing of a female Qatari were mapped using BWA to GRCh37, QTRG1, QTRG2 and QTRG3 reference genomes. The differences between the three Qatari references is that QTRG1 incorporates MAA SNPs, QTRG2 incorporates MAA indels, and QTRG3 incorporates both MAA SNPs and MAA indels. The depth of coverage was measured at (**a**) across the genome and (**b**) at MAA sites modified in the QTRG. MAAs, major alternate alleles; SNP, single nucleotide polymorphism.

average by $n = 756{,}671$ (16%), however, this was on average 25% lower than expected, based on an average of 41% modified sites in each genome (Supplementary Table SVII).

Given the diversity of the Qatari population, use of QTRG may not provide the same benefit for all ancestries. In order to quantify ancestry-specific differences in depth and variant sensitivity, genome analysis using both GRCh37 and QTRG3 was conducted for a diverse panel of $n = 16$ Qatari exomes. In contrast to the genome, across all ancestries, a reduction of variants (up to 76%) was observed to be in excess of the expected (up to 42%, based on modified sites). A significant difference in the reduction was observed between Qataris of Sub-Saharan African ancestry and Qataris of Bedouin or Arab ancestry (one-tailed *t*-test P value < 0.01) (Supplementary Table SVIII).

Liftover of Mendelian disease variants in Qataris

A major challenge for use of QTRG3 that incorporates MAA SNPs and indels is the migration of variant positions due to indels. The position of known variants and of genes is different in QTRG3 than in GRCh37. Hence, a major obstacle to use of QTRG3 in precision medicine studies is the lack of genome interpretation databases on QTRG3 coordinates. A similar issue arises in genomics when a novel assembly of the human reference genome is produced, the

Table 2. Variants in Qatar, stratified by allele frequency and potential for pathogenicity[1]

Category	*Major reference allele (MRA)*						*Major alternate allele (MAA)*					
	All variant alleles (GRCh37)		*Rare alternate allele (< 5% alternate allele frequency)*		*Common alternate allele (5% to 50% alternate allele frequency)*		*Common reference allele (50% to 95% alternate allele frequency)*		*Rare reference allele (95% to 100% alternate allele frequency)*		*Unobserved reference allele (100% alternate allele frequency)*	
	n	%	n	%	n	%	n	%	n	%	n	%
All SNPs	20,864,277	100.00	12,948,368	62.06	5,938,490	28.46	1,693,649	8.12	195,466	0.94	88,303	0.42
3–Potentially pathogenic	155,571	0.75	124,947	0.60	23,282	0.11	5,957	0.03	956	< 0.01	428	< 0.01
2–In gene linked to phenotype	50,757	0.24	40,894	0.20	7,445	0.04	2,002	0.01	290	< 0.01	125	< 0.01
1–Variant with known link	2,152	0.01	999	< 0.01	876	< 0.01	253	< 0.01	21	< 0.01	2	< 0.01

The major allele variants are modified in the QTRG genome, such that all reported variants are the minor allele. A total of 230,395 potentially deleterious SNPs in the 917 exomes and 88 genomes were computationally categorized with respect to allele frequency and databases of genes and variants with reported links to a phenotype. Variants were assigned to genes and their function was predicted with respect to ENSEMBL[34] gene models using SNPEFF,[32] and potentially deleterious coding SNPs (nonsynonymous, splice donor site, splice acceptor site, stop gain, start loss) variants were extracted for further analysis. A database that combines OMIM,[35] HGMD,[36] GWAS,[37] PharmGKB,[38] Human Phenotype Ontology[39] and ClinVar[40] was compiled, where these annotations were used to divide the potentially deleterious variants into three categories, variant and gene linked to a phenotype (Category 1), gene but not variant linked to a phenotype (Category 2) and neither variant nor gene linked to a phenotype (Category 3). The totals for each category are shown in the left-most columns, including number and percentage. These variants were then sub-classified into two major (major reference allele, major alternate allele) and five minor categories based on variant allele frequency in Qatar rare alternate allele (up to 5% variant allele frequency), common alternate allele (between 5 and 50% allele frequency), common reference allele (from 50 to 95% alternate allele frequency), rare reference allele (from 95 to 100% alternate allele frequency), unobserved reference allele (100% allele frequency). The major alternate alleles (MAA) are modified in the Qatar Genome (QTRG).
Abbreviation: SNP, single nucleotide polymorphism.

translation of coordinates from one assembly to another[26] is known as a 'liftover.' There is an established method to accomplish this task, involving pairwise sequence alignment of pairs of chromosomes, such as GRCh37 Chr20 and QTRG3 Chr20, and then using the alignment, the coordinates for a list of sites on GRCh37 Chr20 are 'lifted over' to QTRG3 Chr20 coordinates.

Variants were annotated using SNPEFF[32] (Supplementary Table SIX), and potentially deleterious SNPs in coding genes were then grouped into three categories. Out of $n = 20{,}864{,}277$ SNPs, a total of $n = 155{,}571$ are potentially pathogenic protein coding SNPs (Category 3), including $n = 50{,}757$ in genes linked to a phenotype (Category 2) and $n = 2{,}152$ of these that are variants with known links to a phenotype (Category 1; Table 2). On the basis of ACMG recommendations,[10] only the $n = 1{,}020$ (999+21) Category 1 at minor allele frequency below 5% (rare) are retained for further consideration (Table 2). The rare Category 1 variants were further filtered to exclude $n = 200$ singletons (single alleles in the population), which are enriched for false positives (Supplementary Figure S3). The literature for the remaining variants was reviewed, and $n = 122$ variants linked to a phenotype that can clearly be defined as a Mendelian (dominant, recessive, x-linked) disease are presented in Supplementary Table SX. By using the liftover method, the QTRG3 coordinates for these pathogenic variants was ascertained (Supplementary Table SX).

DISCUSSION

This study presents a set of publicly available bioinformatics tools and resources for genome interpretation studies in Qatar and closely related MENA populations. More than 1,000 Qataris were sequenced to produce this resource, effectively sampling nearly 0.4% of the indigenous population of Qatar. Of the 26 million SNPs and indels observed in the autosomes, sex chromosomes and mitochondrial DNA, more than 9% were in fact the major allele in Qatar. Using the complete catalog of variants, a reference genome custom tailored to disease research in the Qatari population was constructed, named here the Qatari Genome (QTRG).

The value of utilizing the QTRG *vis-a-vis* the standard reference genome GRCh37 was demonstrated through improved read depth and variant sensitivity. Use of this reference in Qatar will lead to higher quality interpretation for both individual genomes and Mendelian disease studies. In order to facilitate Mendelian disease research using the QTRG, a catalog of known pathogenic mutations in Qatar was compiled, with genomic coordinates on both the GRCh37 and QTRG included.

Although prior studies have constructed reference genomes tailored to distinct ancestries,[8] this is the first such study that incorporates both SNPs and indels into the reference, and solves the crisis of genome interpretation that would ensue without liftover of genomic coordinates for known variants. Although in the near future, *de novo* assembly of personal genomes could become routine,[33] the reference genome is expected to remain useful for comparisons across a large sample of genomes, and for comparison with public databases. As expressed by the version information for QTRG (version 1, 2 and 3), future versions of the Qatari Genome are therefore planned for release, based on inclusion of major alleles for a broader spectrum of genetic variation (such as copy number variants, genome rearrangements and short tandem repeats), and ongoing sampling and sequencing of a larger representation both within Qatar and in the MENA region.

DATA ACCESS

Sequence data mapped to GRCh37 in BAM format and variant calls in VCF format for 1376 Qatari, including 4 independent sequencing experiments for a single Qatari female are available for download from the NCBI Sequence Read Archive (SRA accessions SRP060765, SRP061943 and SRP061463, accessible online at http://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP060765%2CSRP061943%2CSRP061463&go=go). (SRA accession SRP061943). The Qatar Genome (QTRG) sequence, the database of annotated variants in Qatar, and bioinformatics tools for analysis of genomes using QTRG are available at our website (http://geneticmedicine.weill.cornell.edu/genome).

ACKNOWLEDGEMENTS

We thank 1000 Genomes Project collaborators, as well as colleagues in the American Society of Human Genetics and the Society for Molecular Biology and Evolution for valuable discussion on genome analysis methods; we thank N. Mohamed for editorial support. These studies were supported, in part, by the Qatar Foundation and Weill Cornell Medical College in Qatar; and NIH T32 HL09428; Qatar National Research Fund NPRP 5-436-3-116 and NPRP 7-1425-3-370.

COMPETING INTERESTS

The authors declare no conflict of interest.

REFERENCES

1 Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015; **372**: 793–795.
2 de Bono JS, Ashworth A. Translating cancer research into targeted therapeutics. *Nature* 2010; **467**: 543–549.
3 Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, Newsham I *et al.* Whole-genome sequencing for optimized patient management. *Sci Transl Med* 2011; **3**: 87re3.
4 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014; **46**: 310–315.
5 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; **4**: 1073–1081.
6 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.
7 Li B, Chen W, Zhan X, Busonero F, Sanna S, Sidore C *et al.* A likelihood-based framework for variant calling and *de novo* mutation detection in families. *PLoS Genet* 2012; **8**: e1002944.
8 Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ *et al.* Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet* 2011; **7**: e1002280.
9 Henn BM, Botigue LR, Bustamante CD, Clark AG, Gravel S. Estimating the mutation load in human genomes. *Nat Rev Genet* 2015; **16**: 333–343.
10 Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015; **17**: 405–424.
11 Lucassen A, Houlston RS. The challenges of genome analysis in the health care setting. *Genes (Basel)* 2014; **5**: 576–585.
12 Qatar Statistics Authority. Results of the 2010 Census of Population, Housing and Establishments (2010). http://www.qsa.gov.qa/QatarCensus/Census_Results.aspx, Accessed 30 August 2012.
13 Rodriguez-Flores JL, Fakhro K, Hackett NR, Salit J, Fuller J, gosto-Perez F *et al.* Exome sequencing identifies potential risk variants for Mendelian disorders at high prevalence in Qatar. *Hum Mutat* 2014; **35**: 105–116.
14 Hunter-Zinck H, Musharoff S, Salit J, Al-Ali KA, Chouchane L, Gohar A *et al.* Population genetic structure of the people of Qatar. *Am J Hum Genet* 2010; **87**: 17–25.
15 Rodriguez-Flores JL, Fakhro K, Agosto-Perez F, Ramstetter MD, Arbiza L, Vincent TL *et al.* Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. *Genome Res.* 2016; **26**: 151–162.
16 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011; **43**: 491–498.
17 Chilamakuri CS, Lorenz S, Madoui MA, Vodak D, Sun J, Hovig E *et al.* Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* 2014; **15**: 449.
18 Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO *et al.* A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.

19 Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 2014; **513**: 409–413.
20 Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009; **19**: 1655–1664.
21 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904–909.
22 Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010; **26**: 2867–2873.
23 Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008; **456**: 53–59.
24 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–1760.
25 Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. Cactus: Algorithms for genome multiple sequence alignment. *Genome Res* 2011; **21**: 1512–1528.
26 Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* 2013; **29**: 1341–1342.
27 GATK development team. GATK Best Practices. https://www.broadinstitute.org/gatk/guide/best-practices, 2015 (last accessed 3/7/16).
28 Rodriguez-Flores JL, Fakhro K, gosto-Perez F, Ramstetter MD, Arbiza L, Vincent TL *et al.* Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. *Genome Res* 2016; **26**: 151–162.
29 Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.
30 Bener A, Hussain R, Teebi AS. Consanguineous marriages and their effects on common adult diseases: studies from an endogamous population. *Med Princ Pract* 2007; **16**: 262–267.
31 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; **13**: 2498–2504.
32 Cingolani P, Platts A, Wang lL, Coon M, Nguyen T, Wang L *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012; **6**: 80–92.
33 Cao H, Wu H, Luo R, Huang S, Sun Y, Tong X *et al. De novo* assembly of a haplotype-resolved human genome. *Nat Biotechnol* 2015; **33**: 617–622.
34 Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S *et al.* Ensembl 2015. *Nucleic Acids Res* 2015; **43**: D662–D669.
35 Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* 2015; **43**: D789–D798.
36 Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med* 2009; **1**: 13.
37 Yu W, Yesupriya A, Wulf A, Hindorff LA, Dowling N, Khoury MJ *et al.* GWAS Integrator: a bioinformatics tool to explore human genetic associations reported in published genome-wide association studies. *Eur J Hum Genet* 2011; **19**: 1095–1099.
38 Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB *et al.* PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res* 2002; **30**: 163–165.
39 Groza T, Kohler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T *et al.* The Human Phenotype Ontology: semantic unification of common and rare disease. *Am J Hum Genet* 2015; **97**: 111–124.
40 Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014; **42**: D980–D985.

Supplementary Information for this article can be found on the *Human Genome Variation* website (http://www.nature.com/hgv)