Genomics Data 3 (2015) 116–127

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Kuwaiti population subgroup of nomadic Bedouin ancestry—Whole genome sequence and analysis

Sumi Elsa John [1], Gaurav Thareja [1], Prashantha Hebbar, Kazem Behbehani, Thangavel Alphonse Thanaraj *,[2], Osama Alsmadi **,[2]

Dasman Diabetes Institute, P.O. Box 1180, Dasman 15462, Kuwait

ARTICLE INFO

Article history:
Received 30 July 2014
Received in revised form 27 November 2014
Accepted 28 November 2014
Available online 19 December 2014

Keywords:
Whole genome sequence
Arabian Peninsula
Nomadic Bedouin ancestry
Kuwaiti population
Intergenome distances
"Tent-dwelling" Bedouins

ABSTRACT

Kuwaiti native population comprises three distinct genetic subgroups of Persian, "city-dwelling" Saudi Arabian tribe, and nomadic "tent-dwelling" Bedouin ancestry. Bedouin subgroup is characterized by presence of 17% African ancestry; it owes it origin to nomadic tribes of the deserts of Arabian Peninsula and North Africa. By sequencing whole genome of a Kuwaiti male from this subgroup at 41X coverage, we report 3,752,878 SNPs, 411,839 indels, and 8451 structural variations. Neighbor-joining tree, based on shared variant positions carrying disease-risk alleles between the Bedouin and other continental genomes, places Bedouin genome at the nexus of African, Asian, and European genomes in concordance with geographical location of Kuwait and Peninsula. In congruence with participant's medical history for morbid obesity and bronchial asthma, risk alleles are seen at deleterious SNPs associated with obesity and asthma. Many of the observed deleterious 'novel' variants lie in genes associated with autosomal recessive disorders characteristic of the region.

Introduction

Population of Kuwait comprises early settlers that include tribes from Arabian and Persian countries, and nomadic Bedouins of the desert [1]. By way of analyzing genome-wide genotypes from 273 Kuwaiti natives, we recently demonstrated three distinct genetic subgroups in Kuwaiti population [2]: Kuwait P (KWP) of Persian ancestry; Kuwait S (KWS) of "city-dwelling" Saudi Arabian tribe ancestry, and Kuwait B (KWB) that includes most of the "tent-dwelling" Bedouin participants (recruited to provide samples for genotyping). The KWB is distinguished from the other two groups by a characteristic presence of 17% African ancestry (ranging from 11.7% to 39.4%); Arabian ancestry is seen more in the Saudi Arabian tribe ancestry subgroup (at 69%) than in the Bedouin group (at 40%). Populations from other states of the Arabian Peninsula also display such a characteristic presence of African ancestry: (i) analysis of mitochondrial DNA variation in Saudi Arabian samples reveals that the Saudi Arabian population harbors as much as 20% genetic contribution from Africa [3]; (ii) analysis of Saudi Arabian Y-chromosome data indicates that around 14% of the Saudi Arabian Y-chromosome pool is typical of African biogeography ancestry [4]; (iii) analysis of mitochondrial DNA variation in populations from Near East and Africa identifies a very high frequency of African lineages (specifically sub-Saharan) in the Yemen Hadramawt [5]; and (iv) analysis of genome-wide genotypes in individuals from Qatar identifies three clear clusters of genotypes with the third cluster comprising individuals with high African admixture [6].

Bedouins are "tent-dwelling" nomads who roamed the deserts of Middle East; they epitomize the best adaptation of human life to desert conditions [7]. In much of the Middle East and North Africa, the term Bedouin is used to descriptively differentiate between those (*bedu*) whose livelihood is based on raising livestock by mainly natural graze and those (*hadar*) who have an agricultural or urban base [8]. Bedouins are originally desert-dwelling tribes of the Arabian Peninsula and are particularly descendants of (i) those settled in the southwestern Arabia, in the mountains of Yemen; and (ii) those settled in North-Central Arabia. Bedouins started to spread out to surrounding deserts of Middle East (particularly Arabian and Syrian deserts) and North Africa (particularly Sinai Peninsula of Egypt and the Sahara Desert of North Africa) due to repeated droughts, growing population and tribal wars. While the "pure" urban-dwelling Arabian tribes formed the leadership class and owned vast amounts of lands, the nomadic Bedouins often worked in the lands of the Arab tribes or tended sheep and camels and moved from one location to another in search of grazing grounds. The Bedouins, as tradition dictated, often married cousins. Marrying within the family helped strengthen bonds among extended families struggling to survive the desert. This centuries-old custom of intermarriage has had devastating genetic effects [9].

* Corresponding author. Tel.: +965 2224 2999x3320; fax: +965 2249 2436.
** Corresponding author. Tel.: +965 2224 2999x4343(work); fax: +965 2249 2406.
E-mail addresses: Alphonse.Thangavel@dasmaninstitute.org (T.A. Thanaraj), osama.alsmadi@dasmaninstitute.org (O. Alsmadi).

[1] These two authors have contributed equally and may be considered as joint first authors.
[2] These two authors have contributed equally.

http://dx.doi.org/10.1016/j.gdata.2014.11.016

The Kuwait Genome Project (KGP) aims to sequence genomes from the three different ethnic subgroups inhabiting Kuwait. In this paper, we report, for the first time, genome sequence resource from the Bedouin subgroup by sequencing a whole genome at 40.96× coverage. The participant that provided the sample is of Yemeni Bedouin ancestry from Kuwait. We catalog 3,752,878 SNPs, 411,839 short indels and 8451 structural variations. We further present neighbor-joining trees that depict intergenome comparisons between the genomes of nomadic Bedouins, "city-dwelling" Arabian tribes, and other continental populations.

Results

We sequenced whole genome of a 20 year old male (of Yemeni origin) from the Kuwaiti Bedouin subgroup using Illumina HiSeq 2000. We generated 1273.08 million paired-end reads of length 101 bps that were aligned to the human reference genome hg19. 95.57% of the reads were mapped to the reference genome, resulting in coverage of 41× (**Supplementary Table S1**).

Ancestry estimation and haplogroup analysis

Examination of genetic clusters derived using principal component analysis (PCA) for Kuwait population (Fig. 1) reveals that this sample is located deep in the Bedouin cluster (and not at boundaries of the clusters or in regions that overlap among the three clusters). The surname lineage classification identified the participant as a desert-dwelling Bedouin tribe. Further, the ancestry composition of the genetic makeup of the KWB individual is seen as: European (French_Basque)—11%; Arab (Negev Bedouin)—44.7%; sub-Saharan African (Biaka_Pygmies)—17.3%; and West Asia (Druze, Brahui)—24.8%. This is consistent with the observed compositions in the Bedouin substructure of Kuwaiti population [2]: European (French_Basque)—11%; Arab (Negev Bedouin)—45.0%; sub-Saharan African (Biaka_Pygmies)—17.0%; and West Asia (Druze, Brahui)—25%. As stated in the Introduction, the nomadic Bedouin subgroup (KWB) is distinguished from the other two groups by a characteristic presence of 17% African ancestry. Arabian ancestry is seen more in the Saudi Arabian tribe ancestry subgroup (KWS) (at 69%) than in the Bedouin group (at 40%); and West Asian ancestry is seen more in the Persian subgroup (KWP) (at 56%) than in any of the other two groups. In order to further illustrate that the ancestry admixture composition of the Bedouin individual sequenced in this study is typical of the KWB group, we present ancestry compositions of 15 samples from the KWS subgroup [10] and one sample from the KWP subgroup [11] (**Supplementary Table S2**). While the sequenced Bedouin sample shows presence of 17.3% African ancestry, the 16 samples from the other two subgroups show African ancestry to the extent of only 0.1% to 5.1%; while the sequenced Bedouin sample shows Arabian ancestry at 44.7%; the 15 samples from the KWS subgroup shows Arabian ancestry to a large extent of 66.1% to 86.5%; and while the sequenced Bedouin sample shows only 24.8% West Asian ancestry, the sample from the KWP subgroup shows as high as 64.5%.

The KWB sample is observed to have J1e [J-P58] Y-chromosome haplogroup which is seen in the Arabian Peninsula. The overall estimated time of expansion of J1e haplogroup is around 10,000 years and the ancestors of J1e haplogroups are observed in the Caucasus and eastern Anatolian populations [12]. The frequencies of J1e in populations from the region are [12]: Sudan: (74.2%; $n = 35$); Yemen (67.7%; $n = 62$); Negev Bedouin (64.4%; $n = 28$); Ismaili Damascus (58.8%; $n = 51$); Qatar (56.9%; $n = 72$); Jordan (48.7%; $n = 76$); Sunni Hama (44.4%; $n = 36$); Oman (37.2%; $n = 121$); and UAE (34.8%; $n = 164$). The observed high value of 67.7% for the frequency of J1e in Yemeni population is consistent with the self-reported Yemeni ancestry by the participant.

The mitochondrial haplogroup (indication of maternal ancestry) of the Bedouin participant is determined as L3d1a1a [L3d], that is predominantly seen in West-Central Africa—among the Fulani [13], Chadians [13], Ethiopians [14], Akan people [15], Mozambique [14], and Yemen [14]. Kivisild et al. [14] analyzed mitochondrial DNA variations in 115 volunteer Yemeni donors in Kuwait (who claimed that their maternal origin was in Yemen) and found that the L macro-haplogroup (the most ancestral mitochondrial lineage) is seen in 47% of the 115 Yemeni individuals; they further found that 20 (17.4%) of the 115

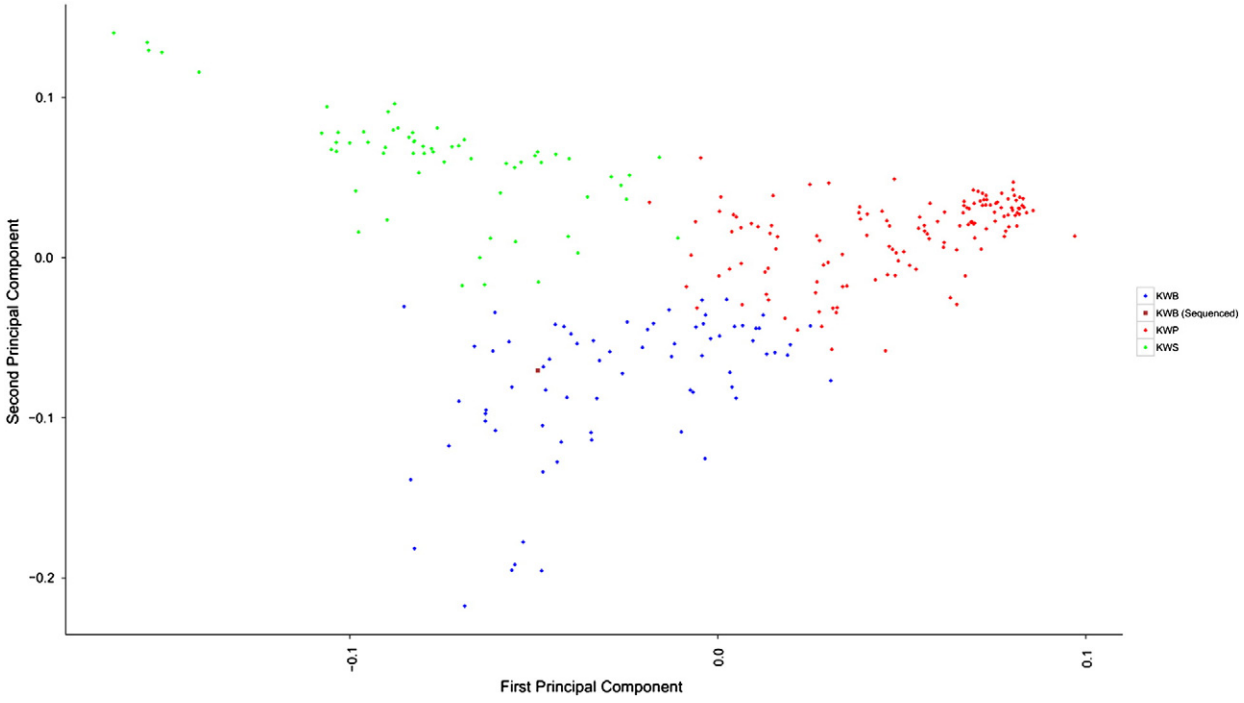

Fig. 1. Scatter plot representing the first two principal components of merged data sets of the three Kuwaiti subgroups. The nomadic Bedouin sample considered for whole genome sequencing in this study is color-coded.

Yemeni participants has the L3 mitochondrial haplogroup (that are most frequently found in sub-Saharan Africa); of these 20 participants, 6 (5.21% of 115 participants) displayed the L3d1 subclade that we observe for the individual sequenced in this study. Thus, the observation of L3d1 haplogroup for the participant in our study is consistent with Yemeni maternal origin. In order to further illustrate that the above observed L3d1 mitochondrial haplogroup is characteristic of the Bedouin sample sequenced in this study, we examined the mitochondrial haplogroups that we identified for a control group of 16 individuals from the other two subgroups of Kuwaiti population (see Supplementary Table S2); none of these 16 samples exhibit the clades of the L macrohaplogroup. Kivisild et al. [14] further compared haplotype diversity seen in Yemeni participants with those reported for Ethiopian population (East Africa); their results highlight the complexity of Ethiopian and Yemeni genetic heritage and are consistent with the introduction of maternal lineages into the South Arabian gene pool from different source populations of East Africa. Horn of Africa (a peninsula in the eastern region of the African sub-continent, enclosing Ethiopia, Somalia, Djibouti and Eritrea) is separated from the south Arabian Peninsula (particularly Yemen) by a short distance of only ~10 miles at the strait of Bab-el-Mandeb (the Gate of Tears); the distance across is only ~20 miles from Ras Menheli in Yemen to Ras Siyyan in Djibouti. Outside of Africa, L3d is mainly found in African Americans; approximately 6% of all African Americans are descendents of the L3d family line [16].

Identification of SNPs and indels

We compared the Bedouin genome with the reference human genome (hg19) for the identification of variants (SNPs and indels). We identified 4,164,717 variants, of which 3,752,878 are SNPs and 411,839 indels. We characterized the variants as 'known' and 'novel' (see Materials and methods) based on dbSNP 138 [17] annotation (which includes variants reported in 1000 Genomes Project phase I release). We find 1.94% (72,881 of 3,752,878) of the SNPs and 7.94% (32,686 of 411,839) of the indels as "novel".

Transition-to-transversion ratio

The genome-wide transition-to-transversion (Ti:Tv) ratio, that is often used to measure specificity for SNP discovery, is 2.11 (in the case of known SNPs) and 1.98 (in the case of novel SNPs). These values are consistent with those reported in literature for whole genomes in 1000 Genomes Project and in other studies [18]—these studies observe 2–2.1 for known SNPs and 1.90–2.1 for novel SNPs.

Validation of SNP calls

We confirmed the validity of the SNP calls by utilizing the genotype data from the same sample derived using the Illumina HumanOmniExpress BeadChip (Illumina Inc, USA). The discordance in SNP calls is seen in a small number of cases (392 out of 312,694) leading to a concordance rate of SNP calls between deep sequencing experiments and genome-wide genotyping at 99.87%. The observed concordance rate is in agreement with that reported in literature—Kenna et al. [19] report a genotype concordance rate of 98.9% upon comparing genotypes for 85 variants inferred across 567 samples using Illumina highthroughput sequencing platforms with genotypes ascertained using Illumina BeadChips. Upon defining the SNPs as homozygous or heterozygous based on BeadChip calls, we find that the disagreements in the SNP calls are more often with homozygous SNPs (206 out of 392) than with heterozygous SNPs (186 out of 392) (**Supplementary Table S3**). As is the practice [20], we choose not to remove the inconsistent calls.

Classification of SNPs and indels based on genome annotation

Based on the locations of the variants relative to annotated genes in the genome, we classified the variants into broad classes such as intergenic, intronic, and coding variants (**Supplementary Table S4**). Most of the variants lie in intergenic regions (59 % of the total variants), followed by those that lie in introns, 3′ UTR and coding regions. The number of intronic SNPs (1,441,241) is around 24 times the number of exonic SNPs (61,643 comprising coding, non-coding exonic, and UTR SNPs); this is consistent with the notion that 1.1 to 1.5% of the human genome codes for exons while 26 to 30% codes for introns. We observe that SNPs from UTRs (30,043) outnumber those from coding exons of transcripts (21,616); this is consistent with observations made in studies using Illumina technology for whole-genome sequencing—e.g. Wong et al. [21] find that 0.86% of the SNPs identified through whole-genome sequencing of 100 southeast Asian Malays lie in coding exons while 1.06% lie in UTRs. Further we classified the variants from coding regions based on their effect on the encoded protein sequences (**Supplementary Table S5**). We identified 58 Stopgain and 12 Stoploss variants among known coding SNPs and 3 Stopgain variants among novel coding SNPs; we further identified 5 Stopgain variants among known indels and one Stopgain variant among novel indels; such Stopgain and Stoploss variants can truncate or elongate the coded peptide sequence. The 70 known SNPs that bring about loss or gain of stop codons are mapped to 74 genes. Of these 74 genes, 29 are found to be annotated in OMIM database and are associated with diseases such as Cohen syndrome, Schizophrenia, and Sepsis (**Supplementary Table S6**); 2 out of 3 Stopgain variants among novel SNPs, 2 out of 5 Stopgain variants from known indels and one Stopgain variant from novel indels are annotated in OMIM. 73 coding SNPs and 52 coding indels are found to disrupt splicing. We also observe that 80 of the known indels and 16 of the novel indels bring about frameshift changes in the encoded proteins.

We examined the observed 9893 nonsynonymous variants from the list of coding SNPs, using SIFT [22] and PolyPhen2 [23], to identify "potentially deleterious" variants (see Materials and methods). In this manner, we identified 2166 known and 105 novel potentially deleterious SNPs from 1841 genes. On checking functional categorization of these genes using Gene Ontology [24], we observed that the significant GO terms all point to sensory perception (such as that of olfaction and cognition) processes, neurological system process, and GPCR signaling pathways & plasma membrane (**Supplementary Table S7**).

In order to identify which of these potentially deleterious SNPs have been previously associated with (or implicated as causal variants for) diseases and phenotype traits, we examined (i) the NHRI GWAS Catalog, a curated resource of SNP-trait association [25], and (ii) the OMIM, a curated catalog of human genes and genetic disorders and traits with particular emphasis on molecular relationship between genetic variation and phenotypic expression [26]. A set of 48 deleterious SNPs are seen annotated for association with diseases in OMIM database and/or in GWAS Catalog (Table 1); the risk alleles at these SNPs are derived using GWAS Catalog or ClinVar [27]. Of these 48 deleterious variants associated with diseases, particularly interesting are those that are in conformity with the phenotype characteristics of the KWB participant, as detailed below:

(a) Morbid obesity (at BMI of 45.5 kg/m^2): The deleterious variants rs2043112 (RICTOR) [28], rs2275848 (NINJ1) [29], rs11042023 (RPL27A) [30] are associated with obesity and related traits. All these three variants carry the risk allele.
(b) Abnormal waist circumference (134 cm): The deleterious variant rs1919128 (C2orf16) is associated with the bivariate trait of waist circumference—triglycerides (WC-TG) [31], and rs1545 (MKKS) is associated with the metabolic syndrome of abdominal obesity [32].

Table 1
Deleterious SNPs annotated for association with diseases in OMIM database and/or in GWAS Catalog.

SNPs	Geno-type	Strongest SNP-risk allele (GWAS Catalog, ClinVar, OMIM)	Mapped gene (OMIM ID)	Disease/trait; (phenotype MIM #); [inheritance][c]
OMIM annotated variants				
rs2297950 [1:g:203194186C > T][Gly102Ser]	het	T[b]	CHIT1(600031)	Chitotriosidase deficiency; (#614122); [?]
rs1056827 [2:g:38302177C > A][Ala119Ser] [a]	hom	?	CYP1B1(601771)	Glaucoma 3A, primary open angle, congenital, juvenile, or adult onset; (#231300,#604229); [AR]
rs34231037 [4:g:55972946T > C][Cys482Arg]	het	G[b]	KDR(191306)	Hemangioma, capillary infantile, susceptibility to; (#602089); [AD]
rs1573496 [4:g:100349669C > G][Gly92Ala]	het	?	ADH7(103720)	Aerodigestive tract cancer, squamous cell, alcohol-related, protection against; (#103780); [MF]
rs1801394 [5:g:7870973A > G][Ile22Met]	het	G[b]	MTRR(602568)	Neural tube defects, folate-sensitive, susceptibility todown syndrome, susceptibility to, included; (#601634); [AR]
rs351855 [5:g:176520243G > A][Gly388Arg]	hom	A[b]	FGFR4(134935)	Cancer progression and tumor cell motility; (no OMIM Id); [?]
rs3807153 [7:g:138417791A > G][Met580Thr]	het	G[b]	ATP6V0A4(605239)	Renal tubular acidosis, distal; (#602722); [AR]
rs1801968 [9:g:132580901C > G][Asp216His]	het	G[b]	TOR1A(605204)	Dystonia-1, modifier of Dystonia-1, torsion; (#128100); [AD]
rs1800450 [10:g:54531235C > T][Gly54Asp]	het	T[b]	MBL2(154545)	Chronic infections, due to MBL deficiency; (#614372); [?]
rs3135506 [11:g:116662407G > C][Ser19Trp]	het	C[b]	APOA5(606368)	Hypertriglyceridemia, susceptibility to; (#145750); [AD}
rs7308720 [12:g: 40657700C > G][Asn551Lys]	het	G[b]	LRRK2(609007)	Parkinson disease 8; (#607060); [AD]
rs2232387 [12:g:52827608C > T][Ala12Thr]	het	T[b]	KRT75(609025)	Pseudofolliculitis barbae, susceptibility to; (#612318); [?]
rs10151259 [14:g:21790040G > T][Ala547Ser] [a]	het	T[b]	RPGRIP1(605446)	Cone-rod dystrophy 13; (#608194); [AR]
rs3743930 [16:g:3304626C > G][Glu148Gln]	het	G[b]	MEFV(608107)	Familial Mediterranean fever, AD, familial Mediterranean fever, AR; (#134610,#249100); [AD; AR]
rs4673 [16:g:88713236A > G][Tyr72His]	het	G[b]	CYBA(608508)	Chronic granulomatous disease, (#233690); [AR]
rs6504649 [17:g:48437456C > G][Thr801Arg]	het	G[b]	XYLT2(608125)	Pseudoxanthoma elasticum, modifier of severity of; (#264800); [AR]
rs1545 [20:g:10386013C > A][Gly532Val] [a]	het	?	MKKS(605552)	Abdominal obesity—metabolic syndrome; (%605552); [?]
rs1801265 [1:g:98348885G > A][Arg29Cys]	hom	A[b]	DPYD(612779)	Dihydropyrimidine dehydrogenase deficiency; (#274270); [AR]
rs486907 [1:g:182554557C > T][Arg462Gln]	hom	T[b]	RNASEL(180435)	Prostate cancer 1; (#601518); [AD]
rs2286963 [2:g:211060050T > G][Lys333Gln]	het	G[b]	ACADL	Metabolite levels; [?]
rs6180 [5:g:42719239A > C][Ile526Leu] [a]	het	C[b]	GHR(600946)	Hypercholesterolemia, familial, modification of; (#143890); [AD]
rs1051931 [6:g:46672943A > G][Val379Ala] [a]	hom	G[b]	PLA2G7(601690)	Asthma, susceptibility to, Atopy;(#600807, #147050); [AD, MF]
rs7133914 [12:g:40702911G > A][Arg1398His]	het	A[b]	LRRK2(609007)	Parkinson disease 8; (#607060); [AD]
rs10246939 [7:g:141672604T > C][Ile296Val]	hom	C[b]	TAS2R38(607751)	Phenylthiocarbamide tasting; (#171200); [AD]
rs61751507 [10:g: 101829514C > T][Gly178Asp]	Het	T[b]	CPN1(603103)	Carboxypeptidase N deficiency; (#212070); [AR]
rs3827103 [20:g:54824029G > A][Val81Ile]	Het	?	MC3R(155540)	Mycobacterium tuberculosis, protection against; (%612929); [?]
GWAS annotated variants				
rs3811444 [1:g:248039451C > T][Thr374Met]	hom	A[b]	CERS2	Platelet counts, red blood cell traits; [?]
rs676210 [2:g:21231524G > A][Pro2739Leu] [a]	het	G	APOB	LDL (oxidized), lipid metabolism phenotypes; [?]
rs6756629 [2:g:44065090G > A][Arg50Cys] [a]	het	G	ABCG5	Cholesterol total, LDL cholesterol (protective effect?); [?]
rs2043112 [5:g:38955796G > A][Ser837Phe] [a]	het	A[b]	RICTOR	Obesity-related traits; [?]
rs240768 [6:g:100957344T > C][Tyr2176Cys]	het	T	ASCC3	Economic and political preferences (immigration/crime); [?]
rs11042023 [11:g:8662516T > C][His324Arg] [a]	hom	C[b]	RPL27A	Obesity; [?]
rs11820589 [11:g:116633862G > A][Pro148Leu] [a]	het	A[b]	BUD13	Metabolic syndrome (bivariate traits); [?]
rs3213764 [12:g:14587301A > G][Lys530Arg]	het	G[b]	ATF7IP	Prostate-specific antigen levels; [?]
rs2297067 [14:g:103566785C > T][Arg77Trp]	het	T[b]	EXOC3L4	Platelet counts; [?]

(continued on next page)

Table 1 (*continued*)

SNPs	Geno-type	Strongest SNP-risk allele (GWAS Catalog, ClinVar, OMIM)	Mapped gene (OMIM ID)	Disease/trait; (phenotype MIM #); [inheritance][c]
rs2303759 [19:g:49869051T > G][Met34Arg]	het	C[b]	DKKL1	Multiple sclerosis; [?]
rs267738 [1:g:150940625T > G][Glu115Ala] [a]	het	A	CERS2	Rhegmatogenous retinal detachment; [?]
rs1919128 [2:g:27801759A > G][Ile774Val] [a]	het	A	C2orf16	Waist circumference—triglycerides (WC-TG); [?]
rs2275848 [9:g:95887320G > T][Ala110Asp] [a]	hom	T[b]	NINJ1	Obesity (early onset extreme); [?]
rs874628 [19:g:18304700A > G][Met72Val]	het	A	MPV17L2	Multiple sclerosis; [?]
rs2239785 [22:g:36661330G > A][Glu166Lys]	het	G	APOL1	Glomerulosclerosis; [?]
Variants annotated in both OMIM and GWAS				
rs11887534 [2:g:44066247G > C][Asp19His]	het	C[b]	ABCG89(605460)	Gallstones, gallbladder disease 4; (#611465); [?]
rs2227564 [10:g: 75673101T > C][Leu141Pro]	hom	C[b]	ABCG8(191840, 605526)	Inflammatory bowel disease; Alzheimer disease, late-onset, susceptibility to;(#104300); [AD]
rs1799853 [10:g:96702047C > T][Arg144Cys]	het	?	CYP2C9(601130)	Warfarin maintenance dose, warfarin sensitivity; (#122700); [AD]
rs4149056 [12:g:21331549T > C][Val174Ala]	het	T,C	SLCO1B1(604843)	Sex hormone-binding globulin levels, response to statin therapy; rotor type hyperbilirubinemia,; (#601816, #237450); [DR]
rs1801272 [19:g:41354533A > T][Leu160His] [a]	het	T[b]	CYP2A6(122720)	Smoking behavior, coumarin resistance (#122700); [AD]
rs1799990 [20:g:4680251A > G][Met129Val]	het	A	PRNP(176640)	Prion diseases, Creutzfeldt–Jakob disease (#606688); [AD]
rs738409 [22:g:44324727C > G][Ile148Met]	het	G[b]	PNPLA3(609567)	Nonalcoholic fatty liver disease; (%613282); [MF]

Abbreviations:
[a] Variants for which the associated phenotype traits are seen with the participant (or his family) that provided sample for genome sequencing.
[b] The alternate allele seen in the Bedouin genome corresponds to the risk allele.
[c] AD: Autosomal dominant; AR: Autosomal recessive; MF: Multi-factorial; DR: digenic recessive;?: not known or multi-factorial.

(c) Bronchial asthma: The deleterious variant rs1051931 (PLA2G7) is associated with susceptibility to asthma [33]. The participant carries homozygous risk allele for the trait.
(d) Family history of retinopathy: The deleterious variant rs267738 (CERS2) is associated with rhegmatogenous retinal detachment [34], and rs10151259 (RPGRIP1) is associated with Cone-rod dystrophy 13 [35]. The participant carries risk allele at the second marker. Presence of these two markers could be indicative of genetic factor for retinopathy seen in the patient's family history.
(e) Smoking: The deleterious variant rs1801272 (CYP2A6) associated with smoking behavior [36].
(f) Prehypertensive (at SBP/DBP of 134/73 Hg/mm^2) and family history of high cholesterol: The following deleterious variants are associated with metabolic syndromes: rs676210 (APOB) associated with LDL [37], rs6756629 (ABCG5) associated with LDL [38], and rs11820589 associated with bivariate traits of TG and HDL [31]—the participant carries the risk allele at the second and third variants).

Each of the above discussed phenotypes is seen with the Bedouin participant and the sequenced genome contains the risk alleles at one or more SNPs that are associated with each of the phenotypes. Though these genotype–phenotype associations have been demonstrated in literature and annotated in OMIM database and/or GWAS Catalog, it is imperative to mention that these individual genotype variants alone are not necessarily sufficient to account for the disorders with the Bedouin participant (for reasons mentioned below):

(i) Each of the discussed phenotypes is influenced by multiple loci and multiple genetic factors, and it is often the case that a component gene can have multiple genetic variants associated with the disorder. For example, the cone-rod dystrophy (CRD) is associated with several genes (including the RPGRIP1 discussed in this work) [35]—the autosomal dominant form of CRD is associated with mutations in the peripherin/RDS, CRX, and RetGC-I genes, and the autosomal recessive form is associated with mutations in the ABCR gene. It is possible that a single locus has only a modest effect on the disease susceptibility and few or all of the reported loci may collectively participate to account for the disorder. Disease might occur only if a particular combination (pattern) of genotypes is present at different susceptibility loci [39].
(ii) Data on genotype-phenotype associations, discussed in this work, come mostly from GWAS studies (**Supplementary Table S8**). Though GWAS studies lead to identification of associated loci, the variants that are identified need not necessarily be the 'causal' variants [40].
(iii) The reported associations in the databases for these disorders are not necessarily demonstrated in the population of Arabian Peninsula and are very often demonstrated in European populations (see Supplementary Table S8). Such associations may not necessarily hold in ethnic populations (that are underrepresented in the global genome-wide surveys) as some genetic variation is private to populations with particular continental ancestry.

We examined the genotypes at SNPs, identified as associated with the phenotypes of the Bedouin sample, in the genomes/exomes of 16 participants from the other two subgroups of Saudi Arabian tribe ancestry and Persian ancestry (see Supplementary Table S8). It is seen that for each of the studied phenotypes, the risk allele is seen in at least one another individual (from the control group) having the same phenotype. As an example: in the case of cone-rod dystrophy, 8 (out of 16) participants from the Persian and Saudi Arabian tribe ancestry subgroups have the phenotype; and 4 of these 8 patients have the alternate allele at rs10151259 (RPGRIP1) as seen in the Bedouin sample; the remaining

four patients show reference allele (so are the remaining 8 participants that form the control group of unaffected participants). In these 4 patients, other mutations associated with cone-rod dystrophy might be present—efforts to identify such other mutations are out of scope for this study. We further observe that except in the cases of cone-rod dystrophy and TG-HDL phenotypes, at least one unaffected individual exhibit the risk allele—this is in concordance with the concerns listed above, particularly the concern that disease might occur only if a particular combination (pattern) of genotypes is present at different susceptibility loci [39].

Further, 9 of the annotated disorders associated with the 48 potentially deleterious SNPs are autosomal recessive. Of the 48 deleterious variants, 10 occur in homozygous form (see Table 1); the participant carries homozygous causal variants for the following two recessive disorders—(rs1056827, Glaucoma 3A [41]) and (rs1801265, dihydropyrimidine dehydrogenase deficiency [42]). Due to the practice of consanguineous marriages and inbreeding, autosomal recessive disorders are prevalent in the region.

Upon considering the 239 genes that harbor the 218 novel nonsynonymous variants, we find that 73 genes are annotated in OMIM database (**Supplementary Table S9**). The annotated diseases mostly include rare genetic disorders such as Myasthenia, limb-girdle, familial—autosomal recessive and congenital; Charcot–Marie–Tooth disease; Hermansky–Pudlak syndrome 3—autosomal recessive; microcephaly—autosomal recessive; mental retardation—autosomal dominant; spondyloepiphyseal dysplasia—Kimberley type; brittle cornea syndrome—autosomal recessive; deafness—autosomal recessive and congenital; Watson syndrome; congenital cataracts; nephrosis-1—congenital and Finnish type; and Mucolipidosis III gamma.

Upon examining the 9,549 noncoding SNPs for annotation in NHGRI GWAS Catalog, we find phenotype association for 26 variants (**Supplementary Table S10**). Many of these 26 variants are associated with phenotypes relating to diabetes, obesity and metabolite levels.

Annotation of the genome for structural variants

We identify 8451 variations consisting of 2893 deletions, 2472 duplications, 1580 insertions, 114 inversions, 470 intrachromosomal translocations, and 922 interchromosomal translocations (**Supplementary Table S11**). Of the total 8451 structural variations, 7672 (90.78%) are "known" structural variations, annotated in DGV (Database of Genomic Variations, a curated catalog of human genomic structural variations) [43]. Further, we see that 6696 (79.23%) of the total deletion variants lie in repeat-rich regions containing SINE (which include ALU), LINE and LTR repeat elements.

Comparison with other individual genomes

In order to assess the extent of variability that the genome of Kuwaiti subgroup of tent-dwelling Bedouin ancestry exhibits with genomes of other populations, we compare the KWB genome with two representative genomes of Kuwaiti subgroup of Saudi Arabian tribe ancestry (KWS) [10] and ten representative genomes (see Materials and methods) from four continents namely Africa (3 genomes), America (3 whites), Europe (2 whites) and Asia (1 Chinese and 1 Korean). As these 13 genomes have been sequenced using six different technologies (see Materials and methods) that have different genome coverage, the genomes cannot be directly compared to evaluate the extent of shared variants. In order to evaluate the intergenome distances among these 13 genomes, we adopt the method of Moore et al. [44] that takes care of variability across the platforms by calculating the extent of shared variant locations chromosome-by-chromosome. The consensus neighbor-joining tree derived by using this method for the 13 genomes is presented in Fig. 2. The three African sequences are closely neighbored, so are the two Asians, and the five Europeans. The two KWS genomes are clustered together and are separated from the KWB genome; all these three Kuwaiti genomes are placed amidst the five Europeans. We then examined the intersection of each of the (KWB, two Kuwaiti genomes of "city-dwelling" Saudi Arabian tribe ancestry, five Europeans, two Asians and three Africans) genome's variants with known disease-causing/predisposing alleles as cataloged in OMIM. The neighbor-joining tree based on the number of shared variant positions carrying the OMIM disease alleles is presented in Fig. 3. The OMIM variant-based tree depicts the European genomes next to one another, the Asian genomes next to one another, and the African genomes next to one another; and more importantly, the two KWS samples and the KWB genomes are near neighbors to one another and are now placed between clusters of African, and clusters of Asian (and European) genomes in congruence with the geographical location of Kuwait and the Peninsula. Of the three genomes (KWS1, KWS2, and KWB) from Kuwait, the Bedouin genome is placed closer to the African cluster in agreement with the earlier observation that the KWB is distinguished from the KWS group by a characteristic presence of 17% African ancestry.

In both the trees (depicting shared genome-wide variants and shared OMIM variants), the two KWS genomes (KWS1 and KWS2) and the KWB genome are near-neighbors to one another; as these three genomes are sequenced using the same Illumina technology, it should be possible to perform direct comparisons on extent of shared variants between these two subgroups. We compared the SNPs from KWS1, KWS2 and KWB genomes; all the three genomes share a high percent of common variants with one another—KWS1 and KWB share 44.7% common variants; KWS2 and KWB share 43.1% common variants; and KWS1 and KWS2 share 45.5% common variants.

Genome view of the variants

Fig. 4 provides a high-level view of the contents of the draft genome sequence for KWB subgroup in terms of density of known and novel variants (SNPs, short and long indels) as observed from the whole genome sequence, density of duplications and the extent of chromosomal translocations. We have also created a genome browser (see the section on Data availability). The browser lets the users to view an annotated display of the identified variants and structural variations in the context of sequence and annotation tracks from other genome resources.

Discussion

Bedouins are the nomadic Arabs of the desert who live on the fringes of the Arabian Peninsula which includes parts of Kuwait, Saudi Arabia, Qatar, United Arab Emirates, Oman, Iraq, Jordan and Syria as well as Negev and Sinai desert [45]. Our earlier work [2] with genome-wide genotype data from Kuwaiti participants has shown that the Kuwaiti population is composed of three distinct genetic clusters—the first group (KWP) is largely of West Asian ancestry, representing Persians with European admixture; the second group (KWS) is predominantly of city-dwelling Saudi Arabian tribe ancestry, and the third group (KWB) includes most of the tent-dwelling Bedouin participants (recruited to provide samples for genotyping) and is characterized by the presence of 17% African ancestry; Arabian ancestry is seen more in the Saudi Arabian tribe ancestry subgroup (at 69%) than in the Bedouin group (at 40%). In this study, we consider an individual of Yemeni ancestry settled in Kuwait; the ancestry composition of the genetic makeup and the surname lineage classification of the individual are typical of the Kuwait B group. The principal component analysis places the individual near the centroid of the Kuwaiti B group. Both the Y-chromosome and the mitochondrial haplogroups are consistent with Yemeni ancestry; the observed mitochondrial haplogroup of L3d1a1a [L3d] is predominantly seen in West-Central Africa; and the J1e [J-P58] Y-chromosome haplogroup is seen in high frequencies in states from the Arabian Peninsula.

The whole genome of the Bedouin individual is sequenced at a high-depth coverage of >40×. Validation of identified SNPs led to a concordance rate of 98.9% between sequencing and BeadChip array genotyping results. Up to 96% of the identified SNPs and indels are validated in the dbSNP 138 database. We believe the remaining 72,881 novel SNPs and 32,686 novel indels add to the repertoire of observed human genome variations. Further, of the identified 8451 structural variations, 779 are novel (i.e. not annotated in DGV, Database of Genomic Variations). We believe that functional level analysis of such population-dependent genomic variations may further shed light on disease mechanisms. Neighbor-joining tree constructed using intergenome distances, calculated based on shared disease-causing variants, between the Bedouin genome and continental genomes places the Bedouin genome (along with the two Kuwaiti genomes of Saudi Arabian tribe ancestry) at the juncture of the clusters of African, Asian, and European genomes; this is in congruence with the geographical location of the Arabian Peninsula. The Peninsula is at the nexus of Africa, Europe and Asia and has been implicated as part of early human migration route out of Africa [46,47] and of early intercontinental trade routes [48]. The tree further illustrates that the global distribution of known disease-causing and predisposing variants within every genome is influenced by the individual's ethnicity. Through this study, we report analysis of a personal genome sequence from the contexts of both population genetics and genotype-phenotype associations. The reported reference data set of genome variants from the individual of “tent-dwelling” Bedouin ancestry from the Peninsula helps to enrich understanding of

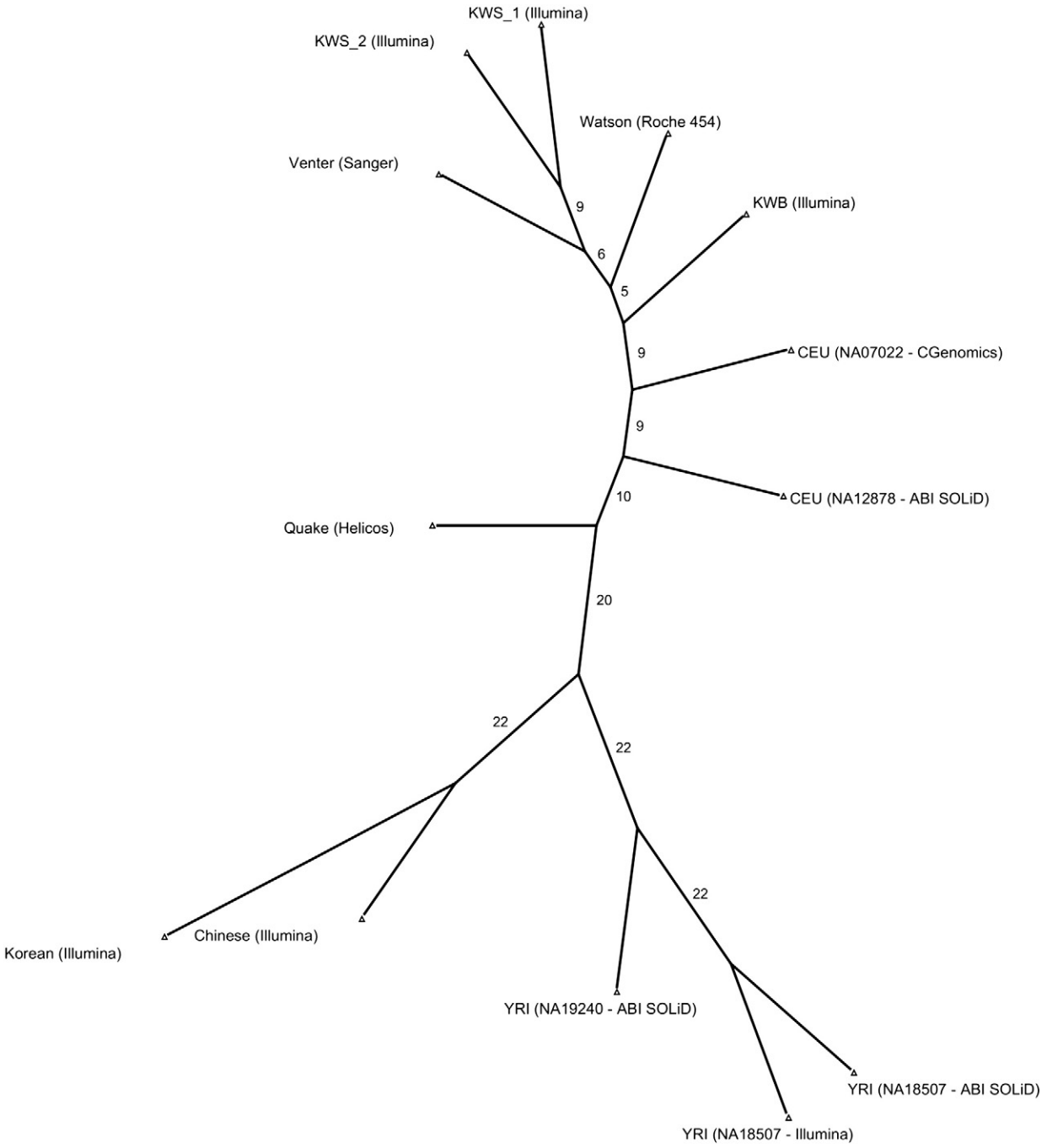

Fig. 2. Neighbor-joining tree based on intergenome distances calculated using genome-wide variant positions shared between the KWB genome, KWS genomes, and representative genomes from intercontinental populations.

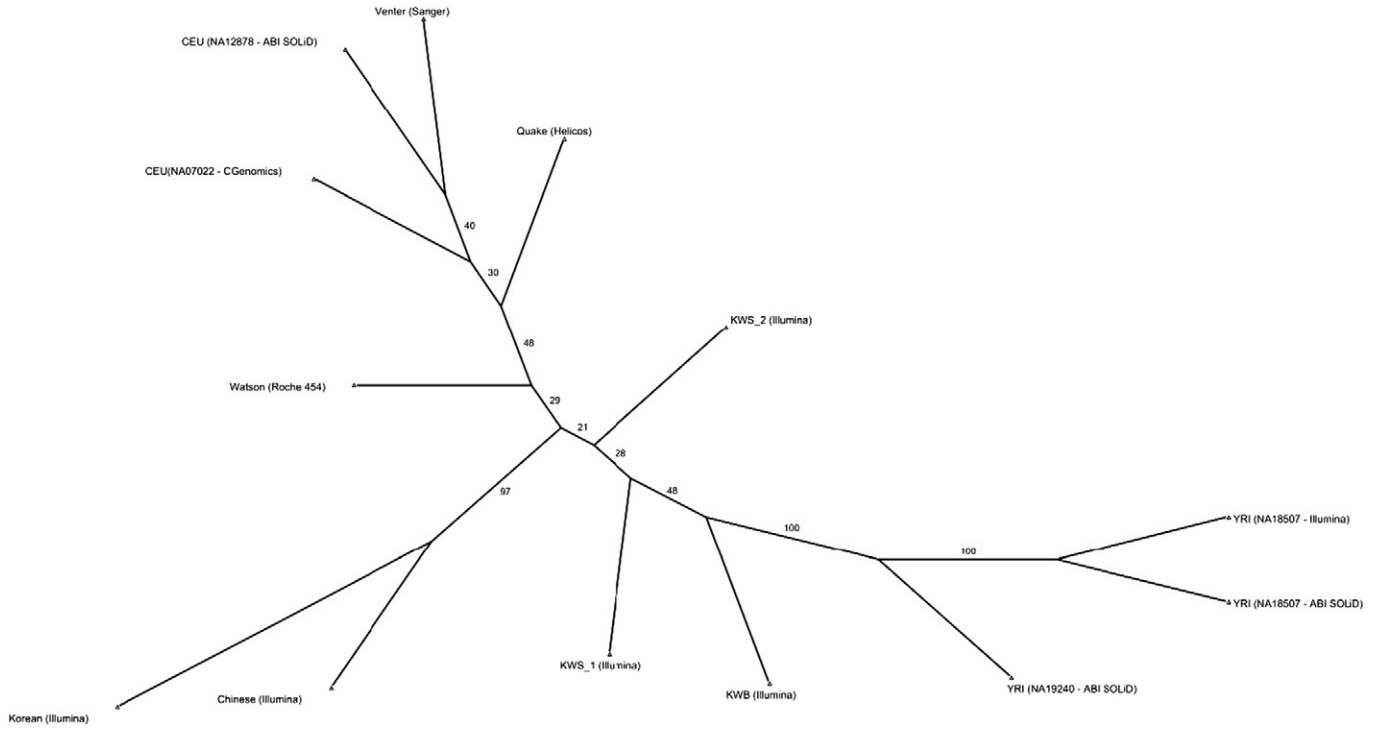

Fig. 3. Neighbor-joining tree based on intergenome distances calculated using variant positions associated with OMIM disease genes and shared between the KWB genome, KWS genomes, and representative genomes from intercontinental populations.

human genome variation across diverse populations. This is particularly important given that the large-scale global sequencing projects have not so far considered populations from Arabian Peninsula.

The rate of consanguineous mating in Kuwaiti population can be as high as 54.3% with higher rates noted among Bedouin tribes [45,49,50]. Frequency of intermarriage with other communities has been particularly low and this has resulted in sustained isolation particularly for the Bedouins and wealthy families. As a result of the extreme inbreeding, consanguinity, and isolation over many centuries, Bedouins (& other Arab tribes) exhibit a high incidence of genetic disorders [45], particularly autosomal recessive disorders. Studies with Bedouins in Negev region of Israel also have shown that they exhibit a high rate of genetically-determined neurological, skeletal, eye, cardiac, gastrointestinal, skin and eye diseases [9]. Thus the Bedouin population is of considerable interest to the medical genetics community that strives to understand the pathophysiology of genetic disorders. Therefore, the reported full-length reference genome sequence for the "tent-dwelling" Bedouins is significantly important to the medical genetics community. Moreover, we find a large number of potentially deleterious missense variants (both known and novel), annotated for diseases in OMIM or GWAS Central catalog, which could be causal variants for a number of autosomal recessive disorders.

Examination of potentially deleterious missense SNPs from the reported genome for disease annotation (in OMIM and GWAS Catalog) leads to deciphering the relationship between the genotype and phenotype characteristics of the individual. (a) The participant is known to suffer from bronchial asthma. A potentially deleterious variant is seen at rs1051931 T = > C (A379V) in PLA2G7 gene; the A379V change has been shown to contribute to increased risk for asthma and atopy [33]. Asthma and related allergic diseases are complex conditions caused by a combination of genetic and environmental factors—the environmental factors differ from population to population. Kuwait, like other states of the Arabian Peninsula, has an arid climate with very hot dry summers and mild winters. Sandstorms are a regular climatic feature occurring most frequently in summer. Rapid urbanization in the post-oil era further contributes to increasing prevalence of asthma in Kuwait and other states. (b) The participant is morbid obese and has abnormal waist circumference. Potentially deleterious variants seen at the following markers have been associated with obesity and related traits (such as abdominal obesity): rs3733418 (C4orf39), rs2043112 (RICTOR), rs2275848 (NINJ1), rs11042023 (RPL27A), and rs1545 (MKKS) all carrying risk allele in the genome of the participant. Both adult obesity and childhood obesity are prevalent in Kuwait. (c) The participant has a family history of retinopathy. The individual carries risk allele at rs10151259 G = > T (A547S) (RPGRIP1) that is associated with Cone-rod dystrophy 13. (d) The participant is a smoker. Potentially deleterious variant rs1801272 A = > T (Leu160His) (CYP2A6) that is associated with smoking behavior is seen in the genome of the participant. Further, the study illustrates that the genome contains risk alleles for many autosomal disorders which are prevalent in the region. Examples include (a) Familial Mediterranean Fever (that can occur in both autosomal dominant and autosomal recessive forms)—marked by rs3743930 G→ C (E148Q) (MEFV gene) in the sequenced Bedouin genome—affects predominantly populations living in the Mediterranean region, especially North African Jews, Armenians, Turks, and Arabs [51]; and (b) Parkinson disease (autosomal dominant) is marked by rs7133914 G = > A(R1398H) (LRRK2 gene) and rs7308720 C = > G (N551K) (again LRRK2 gene) in the sequenced Bedouin genome; while the LRRK2 mutations account for only about 1–2% of sporadic Parkinson's disease cases worldwide, in genetically isolated populations, such as Ashkenazi Jews and North African Arabs, the mutations can account for upwards of 30–40% of sporadic and familial PD cases [52].

Neighbor-joining trees, depicting comparisons between the genome of nomadic Bedouin (KWB) individual (along with two genomes from the KWS subgroup of Saudi Arabian tribe ancestry) and genomes from four continents, give different information depending on whether all the shared genome-wide SNPs or only those shared disease-associated SNPs (as cataloged in OMIM database) are used. While the neighbor-joining tree based on genome-wide SNPs clusters the KWB and KWS genomes amidst the European genomes, that based on only the OMIM SNPs places the KWB and KWS genomes between the three clusters of African, Asian and European genomes (in concordance with the

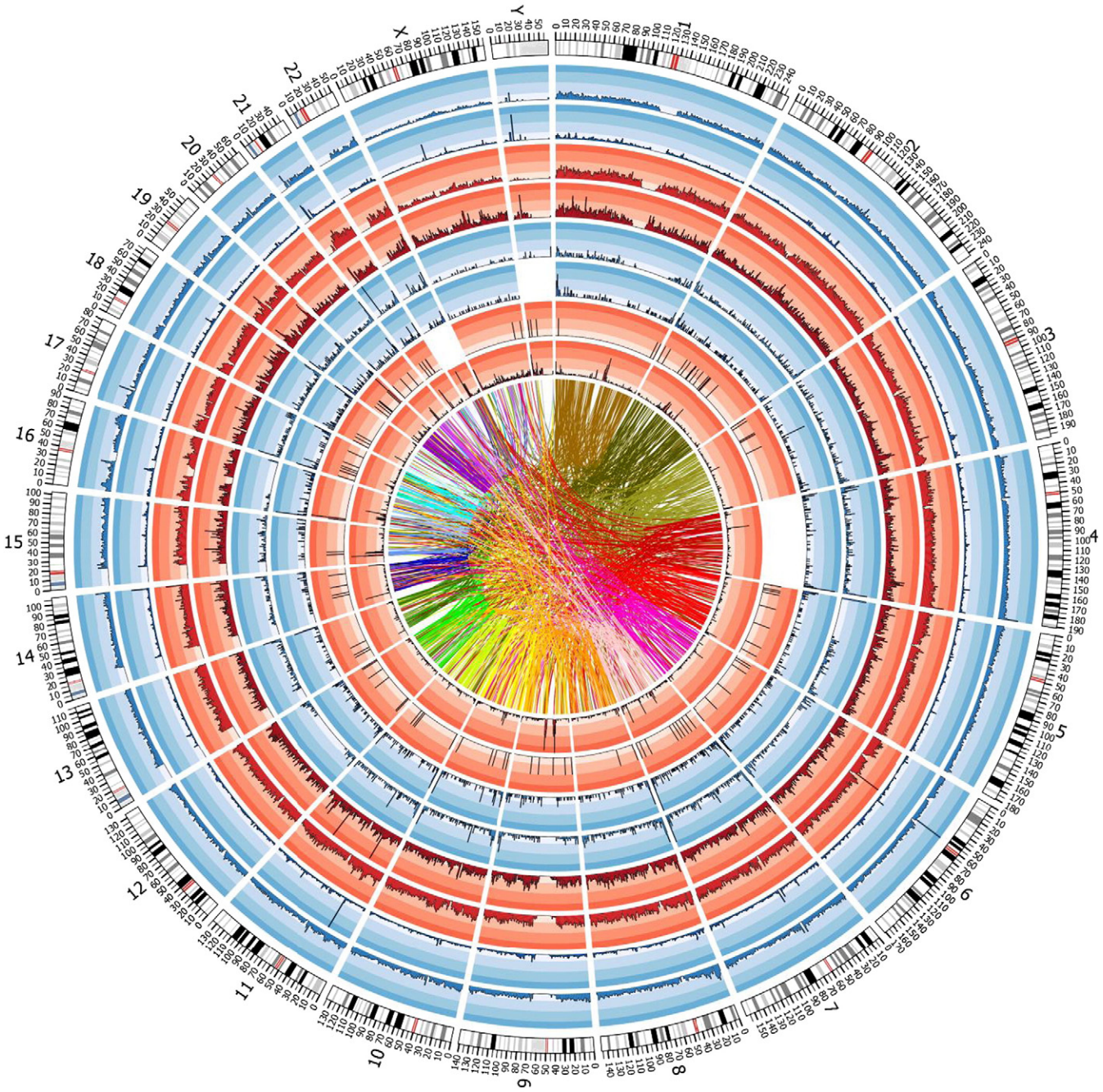

Fig. 4. Summary of analysis of genomes from Kuwait subgroup of Bedouin ancestry. Tracks (from outer to inner): karyotype of human genome; density (in every window of 1 Mb) of 'known' SNPs (i.e. annotated in dbSNP 138); density of 'novel' SNPs (i.e. not annotated in dbSNP138); Density of 'known' indels; density of 'novel' indels; density of long deletions; density of long insertions; density of inversions; density of duplications; links representing intra- and interchromosomal translocations.

geographical location of the origin of the sample at the nexus of Africa, Asia and Europe). This illustrates that the disease profile of individual populations can be different, irrespective of their overall shared origin, and that ethnicity acts as the dominant trend structuring disease-associated SNP locations. This is in agreement with reports that increased levels of population differentiation are detected in disease associated genes when compared to genome-wide base levels [53]. Placement of Kuwaiti genomes amidst the European genomes in the tree based on genome-wide SNPs is in concordance with the following reports: It has been recently suggested, based on analysis of ancient European genomes, that one of the three groups to which the present-day Europeans trace their ancestry is Middle Eastern farmers [54,55]; the three groups are hunter–gatherers who arrived from Africa more than 40,000 years ago, Middle Eastern farmers who migrated to the west much more recently, and those that probably spanned between northern Europe and Siberia. The ancestry admixture due to Middle Eastern farmers in European ancestry may account, at least partially, for the affinity that we see between Europeans and the KWB & KWS participants in the tree based on genome-wide variants.

In conclusion, this is the first study to report a reference genome resource for the population of nomadic "tent-dwelling" Bedouin ancestry. We report novel genome variants that include SNPs, indels and structural variations that enlarge the current repertoire of human genome variation. Neighbor-joining tree built using shared disease-causing variants between the Bedouin genome and other continental genomes positions the Bedouin genome between the clusters of African and

clusters of Asian and European genomes; this is in concordance with the geographical location of the origin of the sample at the nexus of Africa, Asia and Europe. Apart from the findings from population-context, the study illustrates that the medical history of the participant for morbid obesity and bronchial asthma as well as the medical history of the participant's family for retinopathy is accounted by the presence of a large number of genome variants that are known to be associated with these traits. Further, the study illustrates that the genome contains risk alleles for autosomal disorders that are prevalent in the region. The presented genome data provides a starting point for designing large-scale genetic studies in population subgroup of Bedouin ancestry in Kuwait and other states of the Middle East and North Africa.

Data availability

The reported whole genome sequence and all the identified variants (known and novel) are available on the ftp site (ftp://dgr.dasmaninstitute.org). The data can be visualized using genome browser with other annotations tracks from UCSC at http://dgr.dasmaninstitute.org/DGR/gb.html. Proper functionality of the web server requires Firefox version 6 (or later versions) or Internet Explorer version 10 (or later versions).

Materials and methods

Ethics statement

The study was approved by the Scientific Advisory Board and the Ethics Advisory Committee at Dasman Diabetes Institute, Kuwait. Written informed consent for the study was obtained from participant before blood samples were collected.

Detailed methodologies

Details on the methodologies and the tools used to process sample, to sequence the whole genome, and to analyze the genome and variants are presented in **Supplementary Data—Appendix A**. We present below only the essential information on methodologies.

Participant recruitment and sample collection

A 20 year old male participant, clustering with the genetically distinct Bedouin subgroup (KWB) of Kuwaiti population [2], was considered for whole genome sequencing. Blood sample was collected by a trained nurse. Ancestry estimates for the sequenced sample are as extracted from our previous study [2]. For purposes of illustrating the placement of this sample in the Bedouin genetic cluster, the principal component analysis (PCA) plot derived from our previous work [2] is used.

Whole genome sequencing

Processing of blood sample and preparation of libraries for whole genome sequencing were performed as per standard procedures. Paired-end sequencing was performed using Illumina HiSeq 2000.

Identification of genome variants (SNP and Indel) and validation of SNPs

Sequenced paired-end reads were aligned to human reference genome hg19 (UCSC) [56]. The aligned reads were processed to identify SNPs and indels using SAMtools [57] and GATK [58,59] workflows; in order to reduce the likelihood of false discoveries due to the choice of the variant caller, we only utilized the consensus set of variants identified by both the tools. The validity of the SNP calls was confirmed by utilizing genome-wide genotype data from the same sample.

Annotation of variants (SNPs and indels)

A variant is denoted as "novel" if either the variant is not annotated in dbSNP 138 [17] database or the alternate allele seen in the variant in the sample is not a subset of alleles reported in dbSNP. SIFT [22] and PolyPhen2 [23] were used to annotate non-synonymous variants as "deleterious variants" depending on the predicted impact of the amino acid substitution on the protein functionality. The databases of OMIM [26], GWAS Catalog [25], and Ensembl Variation database v72 [60] were used to annotate variants for disease associations.

Detecting structural variations

We used HugeSeq [61] pipeline that implements four different algorithms, to detect structural variations from paired-end reads data. Deletions were annotated using Annovar [62]. A detected deletion is defined to be 'known' if at least 50% of the detected deletion overlaps with annotated deletions in the Database of Genomic Variants [43]; otherwise, the deletion is considered to be "novel".

Identifying Y-chromosome and mitochondrial haplogroup

The Y-chromosome variants were used to call haplogroups using AMY-tree software [63], which uses data from ISOGG (International Society of Genetic Genealogy). The paired-end reads aligned to hg19 mitochondrial sequence were realigned to rCRS (Revised Cambridge Reference Sequence [64]) and then the variants were used to call haplogroups using HaploGrep software [65].

Neighbor-joining trees based on intergenome distances between the genomes of Bedouin, Saudi Arabian tribes, and continental populations

We consider a total of 10 genomes (downloaded from the sites of 10Gen [44]) covering diverse ethnicities from other continents, and 2 Kuwaiti genomes of Saudi Arabian tribe ancestry [10] for comparing the intergenome similarities with the genome of Bedouin ancestry sequenced in this study. The data set of 10 genomes from other continents includes three African (Yoruba) genomes (NA19240 [44,66] on ABI SOLiD, NA18507 [67] on Illumina, NA18507 [68]) on ABI SOLiD), two Asian genomes (Chinese [69] on Illumina, Korean genome [70] on Illumina), five genomes of European descent (Venter [71] on Sanger sequencing, Watson [72] on Roche 454, NA07022 [73] on CGenomics, NA12878 [44,66] on ABI SOLiD, Quake [74] on Helicos). The two Kuwaiti genomes of Saudi Arabian tribe ancestry and the third Kuwaiti genome of nomadic Bedouin ancestry are sequenced using Illumina sequencing technology. We adopt the methods used by Moore et al. [44] to calculate intergenome distances based on information relating to shared variant locations between genomes, and to create a consensus neighbor-joining tree (using PHYLIP [75]) depicting intergenome similarities. The method to calculate the distance is robust with respect to the depth of coverage and hence works well with genomes even when they are sequenced using different sequencing technologies. We built two trees: (i) genome-wide variant tree, that is based on intergenome distances calculated using genome-wide shared variant locations; (ii) OMIM variant tree, that is based on intergenome distances calculated using only those shared variant locations where at least one of the genomes contains an OMIM allele.

Visualization of the content of sequenced genome

The software tool, Circos [76] is used to create the high-level view of the contents (such as density of duplications) of the draft genome sequence. We have built a genome browser, using JBrowse (version 1.8.1) [77] to visualize the sequenced genome sequence and the variants.

Competing interests

The authors declare that they do not have any competing interests.

Author contributions

The study design was performed by OA, TAT and KB. OA directed sample collection & sequencing experiments and contributed to writing the manuscript. TAT directed the data analysis, and developed the manuscript. KB participated in discussions and approved the manuscript. GT and SEJ performed all the data analysis and contributed substantially to writing the manuscript. PH developed the data dissemination protocols and the web sites.

Acknowledgments

The authors thank Philip Beales and Mike Hubank (University College of London Genomics, London) for their advice and suggestions. The authors thank Antony Brooks (University College of London Genomics, London) for help with preparing sequencing libraries. The authors thank Dr Bahareh Azizi for her support and encouragement. The authors thank Daisy Thomas, Motasem K Melhem, Maisa Mahmoud and Ghazi Alghanim for help with recruiting participants. The Ethical Committee and the Scientific Advisory Board at Dasman Diabetes Institute are acknowledged for approving the study. The Kuwait Foundation for the Advancement of Sciences (KFAS) is acknowledged for funding the activities at our institute.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.11.016.

References

[1] M.S. Casey, F.W. Thackeray, J.E. Findling, The History of Kuwait. Greenwood Press, Westport, Conn, 2007.
[2] O. Alsmadi, G. Thareja, F. Alkayal, R. Rajagopalan, S.E. John, P. Hebbar, K. Behbehani, T.A. Thanaraj, Genetic substructure of Kuwaiti population reveals migration history. PLoS One 8 (9) (2013) e74913.
[3] K.K. Abu-Amero, J.M. Larruga, V.M. Cabrera, A.M. Gonzalez, Mitochondrial DNA structure in the Arabian Peninsula. BMC Evol. Biol. 8 (2008) 45.
[4] K.K. Abu-Amero, A. Hellani, A.M. Gonzalez, J.M. Larruga, V.M. Cabrera, P.A. Underhill, Saudi Arabian Y-Chromosome diversity and its relationship with nearby regions. BMC Genet. 10 (2009) 59.
[5] M. Richards, C. Rengo, F. Cruciani, F. Gratrix, J.F. Wilson, R. Scozzari, V. Macaulay, A. Torroni, Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. Am. J. Hum. Genet. 72 (4) (2003) 1058–1064.
[6] H. Hunter-Zinck, S. Musharoff, J. Salit, K.A. Al-Ali, L. Chouchane, A. Gohar, R. Matthews, M.W. Butler, et al., Population genetic structure of the people of Qatar. Am. J. Hum. Genet. 87 (1) (2010) 17–25.
[7] P.K. Hitti, The Arabs: A Short History. Regnery Publishing, Washington, D.C., 1996
[8] D. Chatty, Nomadic societies in the Middle East and North Africa: entering the 21st century. Brill. Netherlands, Boston, 2006.
[9] B. Markus, I. Alshafee, O.S. Birk, Deciphering the fine-structure of tribal admixture in the Bedouin population using genomic data. Heredity (Edinburgh) 112 (2) (2014) 182–189.
[10] O. Alsmadi, S.E. John, G. Thareja, P. Hebbar, D. Antony, K. Behbehani, T.A. Thanaraj, Genome at juncture of early human migration: a systematic analysis of two whole genomes and thirteen exomes from Kuwaiti population subgroup of inferred Saudi Arabian tribe ancestry. PLoS One 9 (6) (2014) e99069.
[11] G. Thareja, S.E. John, P. Hebbar, K. Behbehani, T.A. Thanaraj, O. Alsmadi, Comprehensive analysis of a personal genome of Persian ancestry from Kuwait. BMC Genomics (2015) (in press).
[12] J. Chiaroni, R.J. King, N.M. Myres, B.M. Henn, A. Ducourneau, M.J. Mitchell, G. Boetsch, I. Sheikha, et al., The emergence of Y-chromosome haplogroup J1e among Arabic-speaking populations. Eur. J. Hum. Genet. 18 (3) (2010) 348–353.
[13] D.M. Behar, R. Villems, H. Soodyall, J. Blue-Smith, L. Pereira, E. Metspalu, R. Scozzari, H. Makkan, et al., The dawn of human matrilineal diversity. Am. J. Hum. Genet. 82 (5) (2008) 1130–1140.
[14] T. Kivisild, M. Reidla, E. Metspalu, A. Rosa, A. Brehm, E. Pennarun, J. Parik, T. Geberhiwot, et al., Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am. J. Hum. Genet. 75 (5) (2004) 752–770.
[15] L. Fendt, A. Rock, B. Zimmermann, M. Bodner, T. Thye, F. Tschentscher, E. Owusu-Dabo, T.M. Gobel, et al., MtDNA diversity of Ghana: a forensic and phylogeographic view. Forensic Sci. Int. Genet. 6 (2) (2012) 244–249.
[16] M.W. Allard, D. Polanskey, K. Miller, M.R. Wilson, K.L. Monson, B. Budowle, Characterization of human control region sequences of the African American SWGDAM forensic mtDNA data set. Forensic Sci. Int. 148 (2–3) (2005) 169–179.
[17] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29 (1) (2001) 308–311.
[18] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. del Angel, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43 (5) (2011) 491–498.
[19] K.P. Kenna, R.L. McLaughlin, S. Byrne, M. Elamin, M. Heverin, E.M. Kenny, P. Cormican, D.W. Morris, et al., Delineating the genetic heterogeneity of ALS using targeted high-throughput sequencing. J. Med. Genet. 50 (11) (2013) 776–783.
[20] J.L. Rodriguez-Flores, J. Fuller, N.R. Hackett, J. Salit, J.A. Malek, E. Al-Dous, L. Chouchane, M. Zirie, et al., Exome sequencing of only seven Qataris identifies potentially deleterious variants in the Qatari population. PLoS One 7 (11) (2012) e47614.
[21] L.P. Wong, R.T. Ong, W.T. Poh, X. Liu, P. Chen, R. Li, K.K. Lam, N.E. Pillai, et al., Deep whole-genome sequencing of 100 southeast Asian Malays. Am. J. Hum. Genet. 92 (1) (2013) 52–66.
[22] P. Kumar, S. Henikoff, P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4 (7) (2009) 1073–1081.
[23] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, S.R. Sunyaev, A method and server for predicting damaging missense mutations. Nat. Methods 7 (4) (2010) 248–249.
[24] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25 (1) (2000) 25–29.
[25] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, et al., The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42 (Database issue) (2014) D1001–D1006.
[26] V.A. McKusick, Mendelian Inheritance in Man and its online version, OMIM. Am. J. Hum. Genet. 80 (4) (2007) 588–604.
[27] M.J. Landrum, J.M. Lee, G.R. Riley, W. Jang, W.S. Rubinstein, D.M. Church, D.R. Maglott, ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42 (Database issue) (2014) D980–D985.
[28] A.G. Comuzzie, S.A. Cole, S.L. Laston, V.S. Voruganti, K. Haack, R.A. Gibbs, N.F. Butte, Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. PLoS One 7 (12) (2012) e51954.
[29] E. Wheeler, N. Huang, E.G. Bochukova, J.M. Keogh, S. Lindsay, S. Garg, E. Henning, H. Blackburn, et al., Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nat. Genet. 45 (5) (2013) 513–517.
[30] S.I. Berndt, S. Gustafsson, R. Magi, A. Ganna, E. Wheeler, M.F. Feitosa, A.E. Justice, K.L. Monda, et al., Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat. Genet. 45 (5) (2013) 501–512.
[31] A.T. Kraja, D. Vaidya, J.S. Pankow, M.O. Goodarzi, T.L. Assimes, I.J. Kullo, U. Sovio, R.A. Mathias, et al., A bivariate genome-wide approach to metabolic syndrome: STAMPEED consortium. Diabetes 60 (4) (2011) 1329–1339.
[32] K. Hotta, T. Nakamura, J. Takasaki, H. Takahashi, A. Takahashi, Y. Nakata, S. Kamohara, K. Kotani, et al., Screening of 336 single-nucleotide polymorphisms in 85 obesity-related genes revealed McKusick–Kaufman syndrome gene variants are associated with metabolic syndrome. J. Hum. Genet. 54 (4) (2009) 230–235.
[33] S. Kruse, X.Q. Mao, A. Heinzmann, S. Blattmann, M.H. Roberts, S. Braun, P.S. Gao, J. Forster, et al., The Ile198Thr and Ala379Val variants of plasmatic PAF-acetylhydrolase impair catalytical activities and are associated with atopy and asthma. Am. J. Hum. Genet. 66 (5) (2000) 1522–1530.
[34] M. Kirin, A. Chandra, D.G. Charteris, C. Hayward, S. Campbell, I. Celap, G. Bencic, Z. Vatavuk, et al., Genome-wide association study identifies genetic risk underlying primary rhegmatogenous retinal detachment. Hum. Mol. Genet. 22 (15) (2013) 3174–3185.
[35] A. Hameed, A. Abid, A. Aziz, M. Ismail, S.Q. Mehdi, S. Khaliq, Evidence of RPGRIP1 gene mutations associated with recessive cone-rod dystrophy. J. Med. Genet. 40 (8) (2003) 616–619.
[36] T.E. Thorgeirsson, D.F. Gudbjartsson, I. Surakka, J.M. Vink, N. Amin, F. Geller, P. Sulem, T. Rafnar, et al., Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat. Genet. 42 (5) (2010) 448–453.
[37] K.M. Makela, I. Seppala, J.A. Hernesniemi, L.P. Lyytikainen, N. Oksala, M.E. Kleber, H. Scharnagl, T.B. Grammer, et al., Genome-wide association study pinpoints a new functional apolipoprotein B variant influencing oxidized low-density lipoprotein levels but not cardiovascular events: AtheroRemo Consortium. Circ. Cardiovasc. Genet. 6 (1) (2013) 73–81.
[38] R. Goodloe, K. Brown-Gentry, N.B. Gillani, H. Jin, P. Mayo, M. Allen, B. McClellan Jr., J. Boston, et al., Lipid trait-associated genetic variation is associated with gallstone disease in the diverse Third National Health and Nutrition Examination Survey (NHANES III). BMC Med. Genet. 14 (2013) 120.
[39] J. Hoh, J. Ott, Mathematical multi-locus approaches to localizing complex human trait genes. Nat. Rev. Genet. 4 (9) (2003) 701–709.
[40] M.I. McCarthy, J.N. Hirschhorn, Genome-wide association studies: potential next steps on a genetic journey. Hum. Mol. Genet. 17 (R2) (2008) R156–R165.
[41] G. Chavarria-Soley, H. Sticht, E. Aklillu, M. Ingelman-Sundberg, F. Pasutto, A. Reis, B. Rautenstrauss, Mutations in CYP1B1 cause primary congenital glaucoma by reduction of either activity or abundance of the enzyme. Hum. Mutat. 29 (9) (2008) 1147–1153.

[42] P. Vreken, A.B. Van Kuilenburg, R. Meinsma, A.H. van Gennip, Identification of novel point mutations in the dihydropyrimidine dehydrogenase gene. J. Inherit. Metab. Dis. 20 (3) (1997) 335–338.
[43] J.R. MacDonald, R. Ziman, R.K. Yuen, L. Feuk, S.W. Scherer, The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 42 (Database issue) (2014) D986–D992.
[44] B. Moore, H. Hu, M. Singleton, F.M. De La Vega, M.G. Reese, M. Yandell, Global analysis of disease-related DNA sequence variation in 10 healthy individuals: implications for whole genome-based clinical diagnostics. Genet. Med. 13 (3) (2011) 210–217.
[45] A.S. Teebi, Autosomal recessive disorders among Arabs: an overview from Kuwait. J. Med. Genet. 31 (3) (1994) 224–233.
[46] V. Cabrera, K. Abu-Amero, J. Larruga, A. González, The Arabian Peninsula: Gate for Human Migrations Out of Africa or Cul-de-Sac? A Mitochondrial DNA Phylogeographic Perspective. in: M.D. Petraglia, J.I. Rose (Eds.), The Evolution of Human Populations in Arabia, Springer, Netherlands, 2010, pp. 79–87.
[47] J. Rose, M. Petraglia, Tracking the Origin and Evolution of Human Populations in Arabia. in: M.D. Petraglia, J.I. Rose (Eds.), The Evolution of Human Populations in Arabia, Springer, Netherlands, 2010, pp. 1–12.
[48] B. Slot, Kuwait: The Growth of a Historic Identity. Arabian Publishing, London, 2003.
[49] S.A. Al-Awadi, M.A. Moussa, K.K. Naguib, T.I. Farag, A.S. Teebi, M. el-Khalifa, L. el-Dossary, Consanguinity among the Kuwaiti population. Clin. Genet. 27 (5) (1985) 483–486.
[50] K.E. Al-Nassar, C.L. Kelly, A. EL-Kazimi, Patterns of consanguinity in the population of Kuwait. Am. J. Hum. Genet. 45 (Suppl. 4) (1989) 0915A.
[51] M. Shohat, G.J. Halpern, Familial Mediterranean fever—a review. Genet. Med. 13 (6) (2011) 487–498.
[52] B.R. Haas, T.H. Stewart, J. Zhang, Premotor biomarkers for Parkinson's disease—a promising direction of research. Transl. Neurodegener. 1 (1) (2012) 11–23.
[53] R. Amato, M. Pinelli, A. Monticelli, D. Marino, G. Miele, S. Cocozza, Genome-wide scan for signatures of human population differentiation and their relationship with natural selection, functional pathways and diseases. PLoS One 4 (11) (2009) e7927.
[54] Callaway, E., *Ancient European genomes reveal jumbled ancestry: Nature News & Comment.* DOI: citeulike-article-id:12904863.
[55] I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P.H. Sudmant, J.G. Schraiber, et al., Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513 (7518) (2014) 409–413.
[56] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, et al., Initial sequencing and analysis of the human genome. Nature 409 (6822) (2001) 860–921.
[57] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, et al., The Sequence Alignment/Map format and SAMtools. Bioinformatics 25 (16) (2009) 2078–2079.
[58] G.A. Van der Auwera, M. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, et al., From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics 43 (2013) 11.10.1-11.10.33.
[59] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20 (9) (2010) 1297–1303.
[60] P. Flicek, I. Ahmed, M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, et al., Ensembl 2013. Nucleic Acids Res. 41 (Database issue) (2013) D48–D55.
[61] H.Y. Lam, C. Pan, M.J. Clark, P. Lacroute, R. Chen, R. Haraksingh, M. O'Huallachain, M.B. Gerstein, et al., Detecting and annotating genetic variations using the HugeSeq pipeline. Nat. Biotechnol. 30 (3) (2012) 226–229.
[62] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38 (16) (2010) e164.
[63] A. Van Geystelen, R. Decorte, M.H. Larmuseau, AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. BMC Genomics 14 (2013) 101.
[64] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. 23 (2) (1999) 147.
[65] A. Kloss-Brandstatter, D. Pacher, S. Schonherr, H. Weissensteiner, R. Binna, G. Specht, F. Kronenberg, HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum. Mutat. 32 (1) (2011) 25–32.
[66] F.M. De La Vega, F.C.L. Hyland, S. McLaughlin, A.R. MacBride, E.F. Tsung, H. Peckham, C. Scafe, C. Lee, et al., Functional analysis of the genetic variation within the genomes of three HapMap individuals obtained by whole-genome, second-generation sequencing. 2009. (https://tools.lifetechnologies.com/content/sfs/posters/cms_065553.pdf).
[67] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, et al., Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456 (7218) (2008) 53–59.
[68] K.J. McKernan, H.E. Peckham, G.L. Costa, S.F. McLaughlin, Y. Fu, E.F. Tsung, C.R. Clouser, C. Duncan, et al., Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 19 (9) (2009) 1527–1541.
[69] J. Wang, W. Wang, R. Li, Y. Li, G. Tian, L. Goodman, W. Fan, J. Zhang, et al., The diploid genome sequence of an Asian individual. Nature 456 (7218) (2008) 60–65.
[70] S.M. Ahn, T.H. Kim, S. Lee, D. Kim, H. Ghang, D.S. Kim, B.C. Kim, S.Y. Kim, et al., The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. 19 (9) (2009) 1622–1629.
[71] S. Levy, G. Sutton, P.C. Ng, L. Feuk, A.L. Halpern, B.P. Walenz, N. Axelrod, J. Huang, et al., The diploid genome sequence of an individual human. PLoS Biol. 5 (10) (2007) e254.
[72] D.A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.J. Chen, et al., The complete genome of an individual by massively parallel DNA sequencing. Nature 452 (7189) (2008) 872–876.
[73] R. Drmanac, A.B. Sparks, M.J. Callow, A.L. Halpern, N.L. Burns, B.G. Kermani, P. Carnevali, I. Nazarenko, et al., Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327 (5961) (2010) 78–81.
[74] D. Pushkarev, N.F. Neff, S.R. Quake, Single-molecule sequencing of an individual human genome. Nat. Biotechnol. 27 (9) (2009) 847–850.
[75] J. Felsenstein, PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 5 (1989) 164–166 (DOI: citeulike-article-id:2344765.
[76] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S.J. Jones, M.A. Marra, Circos: an information aesthetic for comparative genomics. Genome Res. 19 (9) (2009) 1639–1645.
[77] O. Westesson, M. Skinner, I. Holmes, Visualizing next-generation sequencing data with JBrowse. Brief. Bioinform. 14 (2) (2013) 172–177.