Identification of single nucleotide variants in the Moroccan population by whole-genome sequencing

Lucy Crooks
The University of Sheffield
Johnathan Cooper-Knock (j.cooper-knock@sheffield.ac.uk)
University of Sheffield https://orcid.org/0000-0002-0873-8689
Paul R. Heath
The University of Sheffield
Ahmed Bouhouche
Universite Mohammed V de Rabat - Souissi
Elmostafa El Fahime
Centre National de la Recherche Scientifique et Technologique
Mimoun Azzouz
The University of Sheffield
Youssef Bakri
Universite Mohammed V de Rabat Faculte des Sciences
Mohammed Adnaoui
Universite Mohammed V de Rabat Faculte de Medecine et de Pharmacie Rabat
Azeddine Ibrahimi
Universite Mohammed V de Rabat Faculte de Medecine et de Pharmacie Rabat
Rachid Tazi-Ahnini
The University of Sheffield

Research article

Keywords: Whole genome sequencing, DNA, population genetics.

Posted Date: April 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-21199/v1

Version of Record: A version of this preprint was published on September 21st, 2020. See the published version at https://doi.org/10.1186/s12863-020-00917-4.

Abstract

Background Large-scale human sequencing projects have described around a hundred-million single nucleotide variants (SNVs), which have predominately focused on individuals with European ancestry despite the fact that genetic diversity is expected to be highest in Africa where Homo sapiens evolved and has maintained a large population for the longest time. The more recent African Genome Variation Project examined several African populations but these were all located south of the Sahara. Morocco is on the northwest coast of Africa and mostly lies north of the Sahara, which makes it very attractive for studying genetic diversity. Recent genomic data of Taforalt individuals in Eastern Morocco revealed 15,000-year-old modern humans, showed that North Africa individuals are expected to show genetic differences from previously studied African populations.

Results We present single nucleotide variant (SNV) results from whole genome sequencing (WGS) of three Moroccans. From a total of 5.9 million SNVs detected, over 200,000 were not identified by 1000G. We provide a summary of the SNVs by genomic position, gene context and effect on protein coding. Comparison of genome-wide information of the Moroccan individuals to individuals from 1000G by principal component analysis revealed a substantial genomic distinction between the Moroccan population and sub-Saharan African populations.

Conclusions We conclude that Moroccan samples lie in the middle of the previously observed cline between populations of European and African ancestry. WGS of Moroccan individuals can identify a large number of new SNVs and aid in functional characterisation of the genome.

Background

Africa represents the birthplace of humanity (Loosdrecht et al. 2018) and still retains a greater genetic diversity than any other region (Botigue et al 2013). However, the majority of genome sequencing studies have focussed their attention upon populations from many parts of the world but largely missed Northern Africa (Schlebush and Jkobsson 2018). Understanding this missing diversity can help to illuminate fundamental function within the human genome.

Genetic analysis of mtDNA and Y-chromosomes showed that North African populations are mixtures of autochthonous and populations from the Middle East, Europe and sub-Saharan Africa (Bekada et al. 2015). The human history of North Africa is characterised by multiple migration movements, admixtures and founder effects. The first appearance of modern life in North Africa was dated to ~ 160,000 years ago (Camp et al. 1994; Smith et al. 2007). Recent genomic sequence of individuals from North African supports the theory of back-to-Africa migration in pre-Holocene times, where Berbers migrated from the Middle East back-to-Africa (Henn et al. 2012). This was followed by many population movements resulting from invasions and migrations by such groups as the Phoenicians, Romans, Vandals, Byzantines, Jew, Arab and more recently Spanish and French (Brett & Fentress. 1996).

For many centuries, North African populations have been mistakenly viewed as part of the sub-Saharan populations simply because they are from the same continent. However, Arredi et al. in 2004 showed that North African populations have predominantly Neolithic origin using Y-chromosome DNA markers (Arredi et al. 2014). Analysis of SNPs from 2,099 individuals in 43 populations showed that North African populations are closer to European and Near East than they are to Sub-Saharan populations (Botigue et al. 2013). More recently, gene flow from North Africa enriches genetic diversity in southern Europe, which is not the case for Sub-Saharan populations. In fact, between 4% and 20% of Southwestern European genomes were assigned to North African ancestral cluster whereas only Sub-Saharan ancestry was detected a < 1% in Europe (Botigue et al 2013).

The North African population, which exceeds 160 million people, has not been included in any of the international genome sequence consortiums including The 1,000 genomes and The African Genome variation Project. To further characterise the North African genome and establish its phylogenic relationship with genomes from other ethnical groups we decided to perform whole genome sequencing (WGS) of three individuals from North Africa.

Methods

Volunteers from different Moroccan regions in the far West of Africa, were recruited in Specialised Hospital of Rabat. All participants were informed regarding the study and written consent was obtained from them according to research protocol presented and approved by local research committee of the Faculty of Medicine and Pharmacy at the University Mohammed V in Rabat. Blood samples were coded and DNA was extracted according to standard protocol. Then three samples were randomly chosen for sequencing. The samples were sequenced using Illumina HiSeq 2500 Rapid Runs. Fastq files were produced with bcl2fastq (version 1.8.4) with adapter trimming and allowing a single mismatch in the index sequence. Reads were mapped by lane to the hg19 human reference genome using bwa aligner (version 0.7.5a). Duplicate reads were marked by Picard (version 1.101) and realignment around indels was performed with GATK (version 2.6-5-gba531bd). The lane-level bam files were then merged with Picard, followed by marking duplicates and indel realignment on the full bam file. Variants were called by GATK's HaplotypeCaller with a minimum calling quality of 1 and allowing for a 10 alternative haplotypes. GATK was used to obtain coverage at every position for reads with a minimum mapping quality of 20 and minimum base quality of 10. The mean and standard deviation across the autosomes and sex chromosomes was calculated with a custom script. A maximum depth of the mean plus 6 standard deviations from the mean was chosen as a filtering cut-off.

Variant quality control was also performed using GATK. Variant sites were filtered out if they had variant quality (QUAL) < 30, genotype quality (GQ) < 20, depth (DP) < 6 or quality by depth (QD) < 2. Sites with SNVs were also filtered out if Root Mean Square Mapping Quality (MQ) < 40, Mapping Quality Rank Sum Test (MQRankSum)<-12.5, Read Position Rank Sum Test (ReadPosRankSum) < − 8, Fisher Strand (FS) > 60. Sites with indels were also filtered out if ReadPosRankSum<-20 or FS > 200. Additionally, sites with two alternative alleles were split so that each allele was on separate line and the representation was

altered if necessary so that the variant was left aligned. Variant alleles present in Phase 3 of the 1000 Genomes data were also filtered out.

Results

Quality control

The mean depth of sequencing read for the autosomes was 16-30X and 88–92% was covered to at least 20X. The large number of novel variants identified necessitated custom quality thresholds. To determine optimum thresholds SNVs were assigned as known or novel based on whether they matched alleles identified in HapMap phase 3 or the Omni 2.5M SNP chip. Based on the assumption that known SNVs are likely to be correct, a strategy was developed to remove as many novel SNVs as possible whilst retaining nearly all known SNVs. The thresholds chosen for keeping SNVs were − 2.5 < MQRankSum < 1 and QD > 7 or QD > 0.5 and − 0.01 < MQRankSum < 0.01, and MQ > 50 or FS = 0 and MQ > 40. This removed 13.3% of novel sites and retained 99.4% of known sites. Variant Quality Score Recalibration (VQSR) was then run on the remaining SNVs with depth (DP), Strand Odds Ratio (SOR), ReadPosRankSum and FS as input metrics. The graphs generated showed clear clusters and separation of positive and negative training sets. A VQSLOD (variant quality score log-odds) threshold of -0.5 for keeping SNVs was chosen. After both filtering steps, 15.4% of novel SNVs were removed and 99.2% of known SNVs were retained. There were 5,880,537 biallelic SNVs and 5,359 multiallelic SNVs remaining.

A high proportion of Moroccan SNVs are novel and predicted to be functional

210,000 of observed SNVs were novel defined as absent from the 1000G (1000 Genomes) project. There was a very high number of novel SNVs between 30–40 Mb on chr 6 and concentration of novel SNVs on chromosome 5 (Fig. 1). A small number, 0.6%, of the SNV alleles were located in highly conserved exons or splice sites; interestingly novel SNVs were enriched in these regions (Fig. 2). Moreover, 1.5% of novel exon SNVs were nonsense variants which is 4-fold higher the value for reference and novel alleles combined (Fig. 3). Overall this is consistent with significant numbers of high impact novel variants within Moroccan individuals.

Moroccan genomes form a distinct population

We then analysed the entire Moroccan genomes by principal components analysis to determine similarity to 1000G populations. Data projected onto PC1 and PC2 calculated from 1000G placed the Moroccans between European and African 1000G populations (Fig. 4)

Discussion

Sequencing new populations is important to understand the diversity and ultimately function of the human genome. The Moroccan population has been neglected in part because of assumptions that it is

not distinct and offers little new information. Our study demonstrates that this is not the case. In fact, Moroccans harbour a large number of novel SNVs and are genetically intermediate between European and African populations.

The study of pathogenic genetic variation rests upon successful filtering of likely benign variants. Large scale population studies such as gnomAD (Karczewski et al 2019) have significantly advanced this effort by distinguishing common from rare variants; rare variants are more likely to be under negative selection and therefore pathogenic. The amount of new information per genome is maximised when individuals are drawn from different populations. Our study demonstrates that Moroccans carry a large number of high impact genomic variation in conserved coding regions. None of the individuals in this study suffered a genetic disease at the age of sampling and therefore it is likely that much of the genetic variation we have described is benign and should be excluded in future attempts to define the genetic architecture of diseases.

A high number of novels SNVs were found in a region of chromosome 6 and on chromosome 5, which might indicate a specific environmental selection. Large scale WGS of additional Moroccans may further define this effect and further our understanding of the gene-environment interaction.

Conclusions

WGS in three Moroccan individuals without significant pathology identified a relatively large number of novel SNVs particularly within conserved coding regions. We conclude that inclusion of Moroccan individuals in future population-level WGS studies is a cost-effective method to maximise the discovery and characterisation of human genetic variation.

Abbreviations

SNV: Single nucleotide variants

WGS: whole genome sequencing

1000G: 1000 Genomes datset

GATK annotations:

MQRankSum: Mapping Quality Rank Sum Test

QD: Quality by Depth

MQ: Root Mean Square Mapping Quality

FS: Fisher Strand

ReadPosRankSum: Read Position Rank Sum Test

VQSR: Variant Quality Score Recalibration

VQSLOD: variant quality score log-odds

SOR: Strand Odds Ratio

Declarations

Ethics approval and consent to participate: informed regarding the study and written consent was obtained from them according to research protocol presented and approved by local research committee of the Faculty of Medicine and Pharmacy at the University Mohammed V in Rabat.

Consent for publication: Not applicable.

Availability of data and methods: Data will be made available through the University of Sheffield Data Repository (https://www.sheffield.ac.uk/library/rdm/orda)

Competing interests: The authors declare that they have no competing interests

Funding: This work has support RTA University account X/006461

Author contributions:

LC, JCK, PRH, AB, EEF, MA, YB, MA, AI and RTA were responsible for the conception and design of the study. AB, EEF, YB, MA, AI and RTA were responsible for data acquisition. LC, JCK, PRH and RTA were responsible for analysis of data. LC, JCK, PRH, AB, EEF, MA, YB, MA, AI and RTA were responsible for interpretation of data. All authors were responsible for revising the manuscript and approving the final version for publication. All authors are responsible for the accuracy and integrity of the work. All authors meet the four ICMJE authorship criteria.

Acknowledgements: We thank blood donors who kindly volunteered to participate in this study.

References

1. Smith TM, Tafforeau P, Reid DJ, Grün R, Eggins S, Boutakiout M, et al. Earliest evidence of modern human life history in North African early *Homo sapiens*. Proc Natl Acad Sci U S A. 2007;104(15): 6128–33.
2. Camps G. Les civilisations préhistoriques de l'Afrique du Nord et du Sahara. Paris: Doi; 1974. p. 374.
3. Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet. 2012;8(1): e1002397.
4. Brett M, Fentress E. The Berbers. Oxford: Bl; 1996.
5. Coudray C, Olivieri A, Achilli A, Pala M, Melhaoui M, Cherkaoui M, et al. The complex and diversified mitochondrial gene pool of Berber populations. Ann Hum Genet. 2009;73(2): 196–214

6. Bekada A, Arauna LR[2], Deba T, Calafell F, Benhamamouch S, Comas D. Genetic Heterogeneity in Algerian Human Populations. PLoS One. 2015 Sep 24;10(9):e0138453
7. Arredi B, Poloni ES, Paracchini S, Zerjal T, Fathallah DM, Makrelouf M, et al. A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. Am J Hum Genet. 2004;75(2): 338–45
8. Botigué LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, Atzmon G, Burns E, Ostrer H, Flores C, Bertranpetit J, Comas D, Bustamante CD. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc Natl Acad Sci U S A. 2013 Jul 16;110(29):11791-6
9. Schlebush CM, Jakobsson M. Tales of human migration, admixture and selection in Africa. Ann Rev Gen and Hum Gen. 2018. 19:405-428
10. Bosch E, Calafell F, Perez-Lezaun A, Clarimon J, Comas D, Mateu E, Martinez-Arias R, Brakez Z, Akhayat O, Sefiani A, Hariti G, Cambon-Thomsen A, and Bertranpetit J. Genetic structure of north-west Africa revealed by STR analysis. European Journal of Human Genetic. 2000. 8. 360-366.
11. Bentayebi K, Abada F, Ihzmad H, and Amzazi S. Genetic ancestry of a Moroccan population as inferred from autosomal STRs. Meta Gene. 2014. 2. 427-438.

Figures

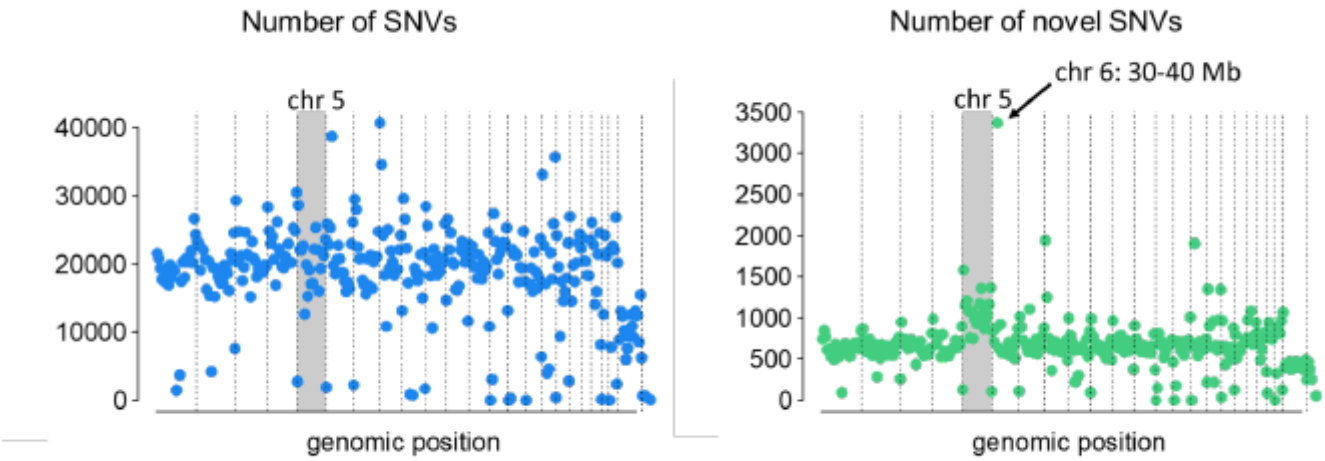

Figure 1

Number of SNV sites in 10 Mb blocks across the genome. Left graph shows all sites, right shows sites with novel SNVs (not in 1000G). Chr 5 is shaded grey. Lines indicate chromosomes.

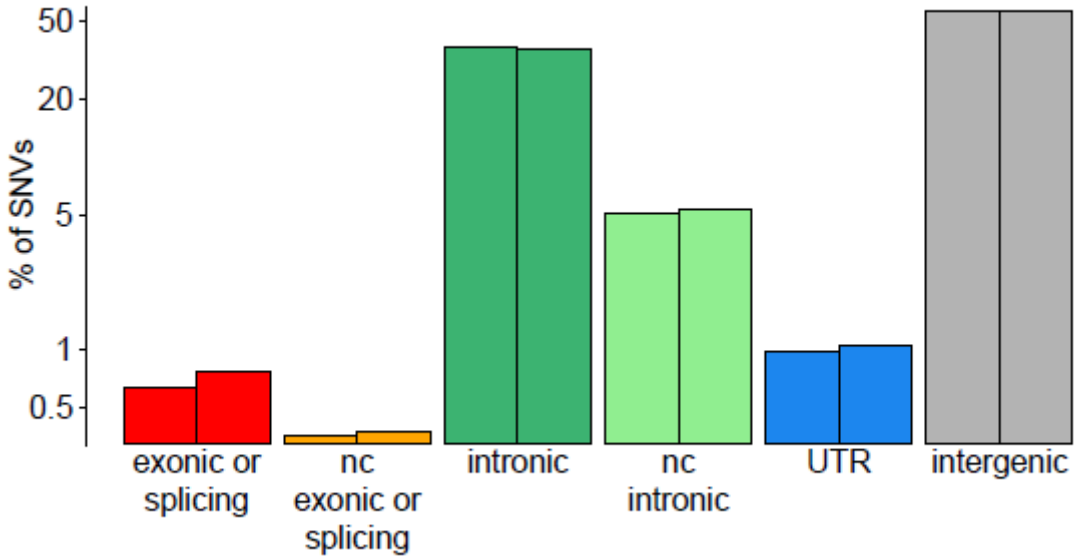

Figure 2

Location of SNV alleles in relation to genes. Y axis is log10 scaled. nc= in non-coding transcript. Structural annotation was performed in ANNOVAR v 16/04/18 using the RefSeq gene model. Genomic regions are arranged left to right; within eaxch region all alleles are plotted (left column) compared to novel alleles only (right column).

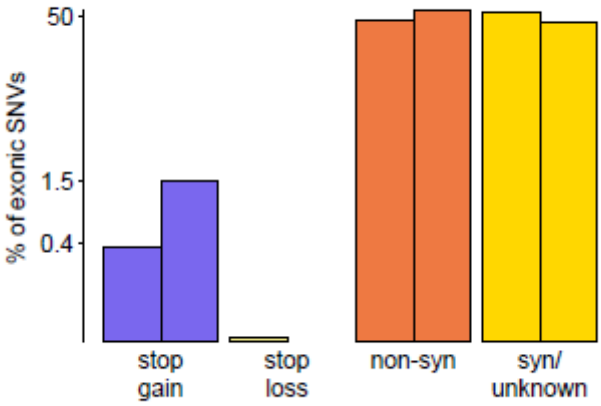

Figure 3

Effect on coding sequence of exonic SNV alleles. Y axis is log10 scaled. Consequences were annotated with ANNOVAR v 16/04/18 using the RefSeq gene model. Consequences are arranged left to right; for each consequence all alleles are plotted (left column) compared to novel alleles only (right column).

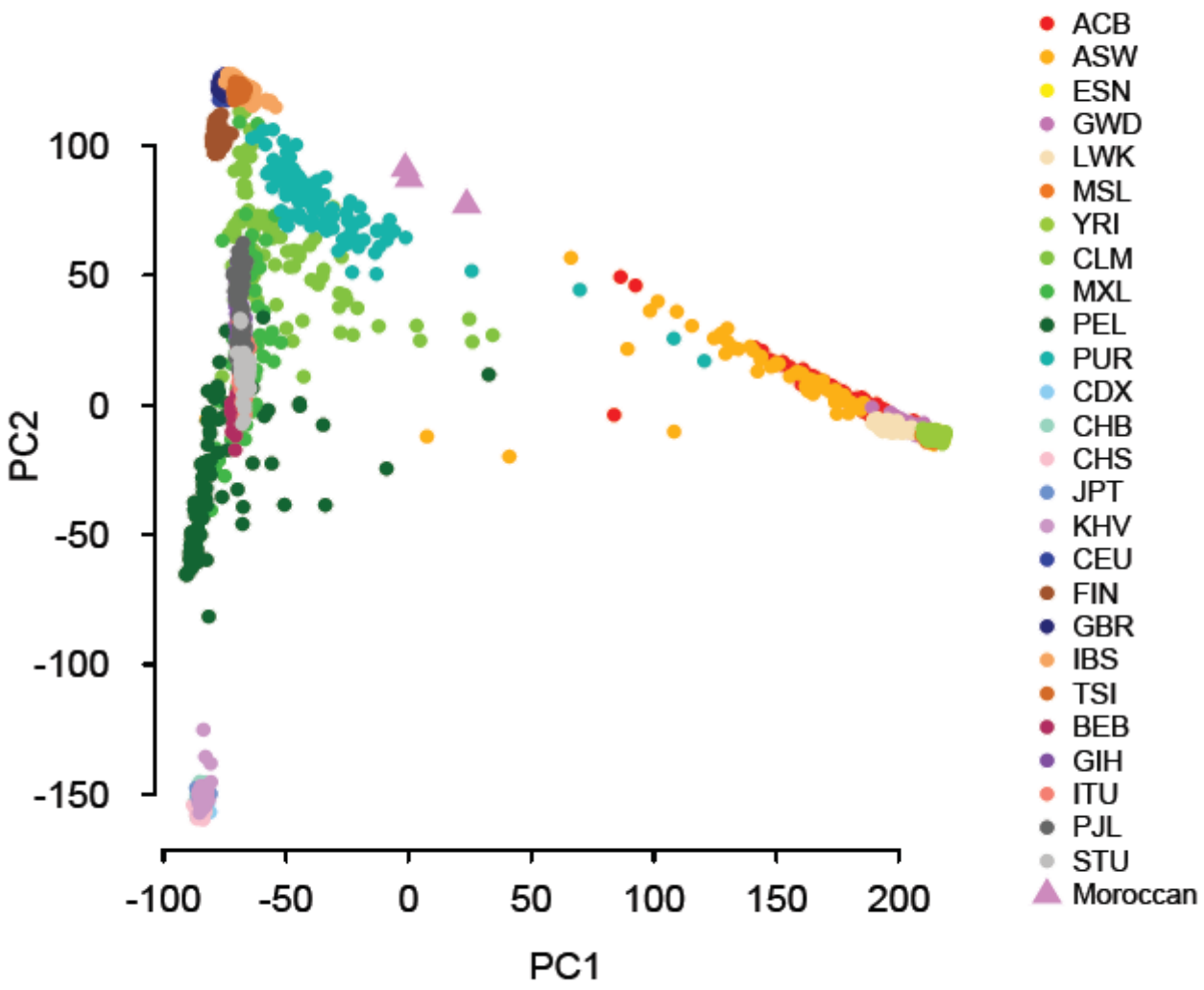

Figure 4

Moroccan data projected onto PC 1 and 2 calculated from the 1000G genotypes. Stringent quality control was applied to the 1000G data. Genotypes at 1000G sites that were not called in the Moroccan individuals were generated with the –L and—include Non Variant Sites options of GenotypeGVCF. Only sites that had DP≥ 10 and GQ ≥ 20 for all three samples and QUAL ≥ 30 if they were invariant, were used. LD pruning was performed on the 1000G genotypes, leaving about 300,000 sites. PCA was carried out in R v3.3.1 with the prcomp function.