Gene 663 (2018) 157–164

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Design of Arab Diabetes Gene-Centric Array (ADGCA) in population with an epidemic of Type 2 Diabetes: A population specific SNP evaluation

Khalid Al-Rubeaan[a,*], Mohthash Musambil[b], Salma Majid Wakil[c], Haya Al-Saud[c], Amr T.M. Saeb[b], Sara Al-Qasim[b], Dhekra Al-Naqeb[a]

[a] University Diabetes Center, College of Medicine, King Saud University, Riyadh, Saudi Arabia
[b] Department of Genetics, Strategic Center for Diabetes Research, King Saud University, Riyadh, PO, Box 245, Saudi Arabia
[c] Department of Genetics, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia

ARTICLE INFO

Keywords:
Arab ethnicity
Custom-designed array
Diabetes-array
Diabetes-chip
High through put
Type 2 diabetes

ABSTRACT

In the case of diabetes and other complex diseases, the challenge has always been to find genetic markers that explain the excess risk associated with development of the disease. In the last 12 years, advances in genotyping technology provided substantial development in the discovery of loci contributing to Type 2 diabetes (T2D) susceptibility. Therefore, the aim of this study is to custom design, for the first time in Arab world, an "Arab Diabetes Gene Centric Array" (ADGCA) that assays 643, 745 SNP markers including 50,617 diabetes associated SNPs. The array content was designed after comprehensive literature search prioritizing Diabetes associated SNPs. PCA was performed to evaluate the relationship between world populations and the Saudi population in building the backbone for the array. A genotype data matrix for PCA analysis was produced by including the genotypes of the 270 HapMap samples including JPT, CHB, YRI and CEU to genotypes of the 1457 Saudi samples. Imputation was executed using IMPUTE2 software and the 1000GP Phase III reference panel. All markers incorporated to ADGCA were validated. Quality checks and evaluation of its capacity and performance as a platform for genetic screening for T2D was performed using the latest stastical tools available. We were successful in designing ADGCA as a custom made chip array designed with a motive to capture genetic variation in loci known or reported to be associated with the development of T2D. However, implementation of ADGCA is currently being performed by our research group using 2000 DNA samples respectively from diabetic and non diabetic individuals which could further validate the use of ADGSA in genetic screening of T2D.

1. Introduction

Diabetes and its complications have become one of the leading causes of deaths worldwide. While type 2 Diabetes (T2D) remains the most prevalent form, it has increased parallelly with the cultural and societal changes around the world. The latest data on the prevalence of diabetes provides confirmation of about 425 million people worldwide being diabetic in 2017 (International Diabetes Federation, 2017). It is also estimated by the International Diabetes Federation IDF (2017) that one in two adults still remain undiagnosed (212 million) and are therefore more likely at risk of developing diabetic complications. Furthermore, 1 in 6 births is affected by hyperglycaemia in pregnancy. Both of these conditions are associated with an increased risk of developing T2D in later part of their lives (International Diabetes Federation, 2017). In 2013, Saudi Arabia was ranked seventh among the top 10 countries known for high diabetes prevalence, mainly T2D, and it was expected to reach the sixth position by 2035 (International Diabetes Federation, 2013). Abnormal glucose metabolism has reached an epidemic state in the Saudi population, wherein, one-quarter (25.4%) of the population are affected and almost half (40.3%) of its diabetic patients are unaware of their disease (Al-Rubeaan et al., 2015). This could be a reflection of major lifestyle changes and a strong genetic contribution secondary to high consanguinity rates (56%) as seen in this community (El Mouzan et al., 2008). This has made this population

Abbreviations: ADGCA, Arab Diabetes Gene Centric Array; AFR, African; ASW, African Ancestry in SW USA; Axiom_GW_Hu_SNP array, Axiom Genome-Wide Human SNP array; CEU, Europeans in Utah; CHB, Han Chinese; CNV, Copy number variation; EMBL-EBI, The European Bioinformatics Institute Part of the European Molecular Biology Laboratory; EUR, European; ExPASy, SIB Bioinformatics Resource Portal; FIN, Finnish; GBR, British; GWAS, Genome wide association studies; IBS, Iberian in Spain; IDF, International Diabetes Federation; JPT, Japanese in Tokyo; KEGG, Kyoto Encyclopedia of Genes and Genomes; LWK, Luhya in Kenya; MAF, Minor allele frequency; PCA, Principle component analysis; SAU, Saudi Arabian; SNPs, Single nucleotide polymorphisms; T2D, Type 2 Diabetes; TSI, Toscani in Italy; YRI, Yoruba in Nigeria

* Corresponding author.
E-mail address: krubeaan@dsrcenter.org (K. Al-Rubeaan).

https://doi.org/10.1016/j.gene.2018.04.019
Received 21 November 2017; Received in revised form 5 April 2018; Accepted 9 April 2018
Available online 22 April 2018
0378-1119/

particularly suitable for genetic studies that could help in identifying individuals at risk using the latest genetic tools.

There have been extensive efforts to uncover genetic variants of complex diseases using single nucleotide polymorphisms (SNPs), which were a product of the International HapMap Project and carried out as an extension of the Human Genome Project providing information about more than one million SNPs in the human genome (International HapMap Consortium, 2005). The advancement of genotyping technologies resulted in unveiling more information about the SNPs and their association with a particular type of disease. The last 10 years have witnessed tremendous advances in genetic studies associated with T2D which included Genome wide association studies (GWAS) which uses ever larger samples and denser genotyping arrays supplemented by imputation against more complete reference panels and richer ethnic diversity. These studies have yielded > 80 robust association signals (Fuchsberger et al., 2016). Even though studies have been successful in providing the first insights into the genetic architecture and pathophysiological basis of the disease, future fine-mapping efforts including advancement in re-sequencing technologies, large-scale targeted or whole-genome studies in large sample sizes are very much necessary in finding rest of the unknown genetic associations (Morris, 2014).

Previous studies focussed on elucidating the genetic architecture of diabetes and were only limited to linkage analysis and candidate gene studies. Later on, GWAS was used to uncover the common variants involved in common complex disorders such as T2D. GWAS involves genotyping thousands of SNPs and structural variants across the genome and their haplotype contexts. GWAS assays have been made possible by projects such as the HapMap project (International HapMap Consortium, 2005) and the 1000 genome project (1000 Genomes Project Consortium et al., 2010). Until now, several GWAS studies for T2D have been conducted across various ethnicities (Morris, 2014) and the major part of them have always pointed out the inconsistency in their results. This indicated a large discrepancy in the prevalence of T2D among different ethnic groups and clearly showing the role of ethnic differences contributing to insulin action and secretion (Bennet et al., 2014; Morris, 2014). Although GWAS produce a large amount of data demonstrating hundreds of genetic loci and their association to complex traits in humans, the question of identifying additional genetic loci by studying the most significantly associated variants and a further investigation via genetic fine-mapping makes investigators choose custom-designed arrays. Custom-designed arrays are a powerful and cost-effective approach to follow-up large-scale genotyping and sequencing studies and to advance our understanding of the genetic basis of complex human diseases and traits (Voight et al., 2012).

To the best of our knowledge, there is no array explicitly designed and focussed on T2D considering the Arab population. In this study, we chose a Saudi population that is known to have an epidemic of T2D in order to develop Arab Diabetes Gene Centric Array (ADGCA) for T2D. The ADGCA contains 643,745 SNPs including 50,617 T2D related SNPs corresponding approximately to 150 genes which have been designed to perform deep replication of established disease signals and fine mapping of GWAS loci. The ADGCA has been designed to cover all T2D established variants and also variants present on genes associated with it. This will allow for more robust risk prediction, therefore giving patients a better chance for identification and prevention which will as a result, reflect in reducing the disease prevalence.

2. Methods

2.1. The array design

2.1.1. The process of loci selection

The basis for this study which would look into the design of ADGCA using population known to have an epidemic of Type 2 Diabetes, was built on a meta-analysis study that was carried out to investigate the SNPs associated with T2D in Arab and Caucasian ethnicities (Al-Rubeaan et al., 2013). To expand on the information obtained from the meta-analysis, online databases PubMed, PubMed Central, Google Scholar, Entrez Gene and OMIM were used to identify the additional candidate SNPs. The genetic search criteria comprised of SNP, gene, allele, variant, polymorphism, insertion, deletion, and duplication. The diabetes focussed search terms were diabetes, T2D, impaired glucose tolerance, gestational diabetes, insulin resistance, hyperglycemia, obesity, diabetic nephropathy, diabetic neuropathy, diabetic retinopathy, and vasculopathy. The scope of literature search used in this study included candidate gene studies, linkage analysis, association studies and molecular studies. Over 3037 articles published between 1997 and 2014 were collected systematically, analyzed and scrutinized focussing on the sample size, quality and strength of the study. Further details on SNPs collected from the analyzed publications were obtained from NCBI's db SNP (http://www.ncbi.nlm.nih.gov/SNP/) in addition to SNPedia (http://www.snpedia.com/), SNP nexus (http://snp-nexus.org/) and ENSEMBL genome browser (http://www.ensembl.org/index.html). Annotations for the SNPs were found through Gene Ontology (http://geneontology.org/), KEGG (Kyoto Encyclopedia of Genes and Genomes)(http://www.genome.jp/kegg/), EMBL-EBI (The European Bioinformatics Institute Part of the European Molecular Biology Laboratory) (http://www.ebi.ac.uk/) and the protein functions were obtained through ExPASy (SIB Bioinformatics Resource Portal) (http://www.expasy.org/genomics).

Filtration of loci was based on the following search criteria: organism & *Homo sapiens*, also special attention was given in classifying them based on their class, variation, clinical significance, and pathogenesis. SNP selection also depended upon the location of the SNPs and their functional classification which included those involved in splice sites, UTRs, frame shift SNPs, intronic SNPs, missense SNPs, nonsense SNPs and stop-gain SNPs. Validation status for all SNPs was checked and SNPs were validated from the 1000 genomes project (http://www.1000genomes.org) and selected for further analysis.

2.2. The array backbone

Due to the lack of publicly available genome wide SNP discovery data in the Saudi (SAU) population or in other Arab populations, the chip backbone for the Arab population was created using data from Axiom Genome-Wide Human SNP array (Axiom_GW_Hu_SNP array) genotypes for 1457 normal SAU individuals in order to enrich marker selection for the genome wide coverage of this population. This data was employed to find out the closest world population to the SAU population in order to select 1000GP phase 1 markers to maximize imputation coverage of the world population most closely resembling the SAU population. After this, the marker set was enhanced with a group of variants with a frequency increase in the SAU population.

To assess the relationship between SAU population and other world populations, genetic data from the 270 Phase I HapMap samples, which included; 45 Japanese in Tokyo (JPT), 45 Han Chinese (CHB), 90 Yoruba in Nigeria (YRI) and 90 Europeans in Utah (CEU), were genotyped on the Axiom_GW_Hu_SNP array. Principle component analysis (PCA) was carried out for the genotypes of the 270 Phase I HapMap individuals (Phase I HapMap - International HapMap Consortium, 2005) and the 1457 SAU individuals for 40,000 SNPs from the Axiom_GW_Hu_SNP array. These SNPs from autosomes were randomly selected from 399,780 SNPs that were of high quality and polymorphic SNPs for the SAU individuals (See supplementary material for details of PCA analysis performed).

2.3. Diabetes centric probeset selection

Probesets were selected to interrogate 50,617 unique SNPs that comprised the array custom content and from which, 221 SNPs (Fig. 1 and Supplementary Table 1) were flagged as very important for diabetes associations. All markers were interrogated with a validated