Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome.

Henry Richard Johnston, Emory University
Yijuan Hu, Emory University
Jingjing Gao, AbbVie
Timothy D. O’Connor, University of Maryland
Gonçalo R. Abecasis, University of Michigan
Genevieve L. Wojcik, Stanford University
Christopher Gignoux, Stanford University
Pierre-Antoine Gourraud, University of California
Antoine Lizée, University of California
Mark Hansen, Illumina, Inc.

Only first 10 authors above; see publication for full author list.

Journal Title: Scientific Reports
Volume: Volume 7
Publisher: Nature Publishing Group: Open Access Journals - Option C | 2017-04-21, Pages 46398-46398
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1038/srep46398
Permanent URL: https://pid.emory.edu/ark:/25593/s2q66

Final published version: http://dx.doi.org/10.1038/srep46398

Copyright information:
© 2017, The Author(s)
This is an Open Access work distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Accessed September 13, 2019 4:44 PM EDT
Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome


A primary goal of The Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) is to develop an ‘African Diaspora Power Chip’ (ADPC), a genotyping array consisting of tagging SNPs, useful in comprehensively identifying African specific genetic variation. This array is designed based on the novel variation identified in 642 CAAPA samples of African ancestry with high coverage whole genome sequence data (~30 × depth). This novel variation extends the pattern of variation catalogued in the 1000 Genomes and Exome Sequencing Projects to a spectrum of populations representing the wide range of West African genomic diversity. These individuals from CAAPA also comprise a large swath of the African Diaspora population and incorporate historical genetic diversity covering nearly the entire Atlantic coast of the Americas. Here we show the results of designing and producing such a microchip array. This novel array covers African specific variation far better than other commercially available arrays, and will enable better GWAS analyses for researchers with individuals of African descent in their study populations. A recent study cataloging variation in continental African populations suggests this type of African-specific genotyping array is both necessary and valuable for facilitating large-scale GWAS in populations of African ancestry.

The design of the African Diaspora Power Chip (ADPC) was a primary goal as part of the NIH-supported Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA). Because of the overall poor coverage for African specific-variants on commercially available GWAS arrays, amongst other difficulties, relatively few GWAS have been performed in populations of African descent, in part because they were underpowered to identify association with genes controlling risk for complex disease. Previous GWAS studies in populations of African descent may have missed critical association signals because the single nucleotide polymorphisms (SNPs) genotyped on existing commercial arrays were selected for being informative among individuals of European ancestry, and generally do a poor job of tagging haplotypes and variants in individuals of non-European ancestry. This stems from the fact that the frequency of SNPs on currently available arrays is not well matched to the frequency of untagged variants in non-European populations. Essentially, the variant spectrum on current SNP arrays is flat, ensuring common genetic variants are well tagged, but making it difficult to tag low frequency SNPs (even though they may be highly polymorphic in non-European populations). This missing genetic variation in non-Europeans, however, consists largely of low frequency and rare variants, which will always be poorly covered by tagging SNPs with higher minor allele frequency (MAF). By building a large

1Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA. 2Data and Statistical Sciences, AbbVie, North Chicago, IL, USA. 3Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA. 4Program in Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. 5Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. 6Department of Biostatistics, SPH, University of Michigan, Ann Arbor, MI, USA. 7Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. 8Department of Neurology, University of California, San Francisco, San Francisco, CA, USA. 9Illumina, Inc., San Diego, CA, USA. 10Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. 11Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. 12Department of Medicine, Johns Hopkins University, Baltimore, MD, USA. †A comprehensive list of consortium members appears at the end of the paper. Correspondence and requests for materials should be addressed to K.C.B. (email: kathleen.barnes@ucdenver.edu)
Table 1. Illumina projected coverage of African variants on several commercially available GWAS arrays.

<table>
<thead>
<tr>
<th>Array Type</th>
<th>&lt;0.010 (non-singletons)</th>
<th>&gt;0.010</th>
<th>&gt;0.025</th>
<th>&gt;0.050</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanExomev1-2</td>
<td>0.031</td>
<td>0.032</td>
<td>0.034</td>
<td>0.035</td>
</tr>
<tr>
<td>HumanCore-12v1-0</td>
<td>0.127</td>
<td>0.152</td>
<td>0.203</td>
<td>0.256</td>
</tr>
<tr>
<td>HumanCoreExome-12v1-0</td>
<td>0.148</td>
<td>0.172</td>
<td>0.221</td>
<td>0.271</td>
</tr>
<tr>
<td>OmniExpress-12v1-1</td>
<td>0.210</td>
<td>0.249</td>
<td>0.326</td>
<td>0.395</td>
</tr>
<tr>
<td>OmniExpressExome-8v1-1</td>
<td>0.226</td>
<td>0.263</td>
<td>0.337</td>
<td>0.403</td>
</tr>
</tbody>
</table>

In the remaining set of SNPs. These LD estimates were used by FESTA, a TagSNP selection program, to select SNPs using a rather strict r^2 threshold of 0.8. A total of 1,004,268 TagSNPs were selected. Among them, 4,000 were removed because they were too similar in their probe design to function well on the array. Limited by the capacity of the array, only TagSNPs with an MAF >= 1% were retained for inclusion on the array. Raising the MAF of 1% was chosen as the preliminary cutoff to limit the initial pool of variants to be tagged. In addition, using 1% as the MAF cut-off eliminated concerns about potential false positive variant calls for rare variants derived from the sequence data. The Illumina HumanExomev1-2 still only covers 20% of known YRI SNPs based on their analysis. In contrast, variants of European (CEU) ancestry are 21% covered by the OmniExpress and 44% covered by the Omni 2.5. These are large differences in genomic coverage, and are supported by other studies with similarly pessimistic estimates of effective coverage among non-European populations. Even the primary manufacturer of the Omni series genotyping chips, Illumina, refers to low coverage levels for their currently available commercial GWAS arrays in non-European populations. Table 1 It is important to note that there is no standardized definition for efficient 'coverage'; as each method uses a different set of SNPs to assess coverage levels. Regardless of the method used, however, contemporary commercially available arrays do a poor job of tagging common haplotypes or 'covering' all genetic variation in non-European or admixed populations.

Usage of the most recent imputation panels can significantly improve the coverage of African variants, but this practice is still hamstrung by the lack of low-frequency variants on genotyping arrays (MAF < 5%). Imputation of low-frequency variants is most efficient and accurate when the SNP to be imputed has a similar minor allele frequency as the genotyped SNP, so a relative lack of low frequency variants on an array can render imputation of similar frequency variants difficult. To address this shortcoming, the ADPC was designed using the whole-genome sequencing results on 642 CAAPA samples, including 328 African Americans, 125 African Caribbean subjects, 164 African ancestry individuals with some Latino ancestry, and 25 individuals from Nigeria. The whole genomes of these individuals were sequenced using the Illumina HiSeq 2000. A total of 47.9 million biallelic SNPs were identified in these CAAPA samples. Of those, 15.6 million variants have a MAF greater than or equal to 1%.

To create an affordable array for large-scale chip-based studies, 700,000 variants was the maximum size of the array. A MAF of 1% was chosen as the preliminary cutoff to limit the initial pool of variants to be tagged. In addition, using 1% as the MAF cut-off eliminated concerns about potential false positive variant calls for rare variants derived from the sequence data. Additionally, the ADPC is designed to be used in conjunction with the OmniExpress array, a low-cost GWAS array popular among researchers, leveraging the high MAF coverage available from OmniExpress and freeing the ADPC to focus on low frequency variants.

To narrow the pool further, SNPs with poor Illumina design scores (proprietary scores which attempt to give a numeric representation of the likelihood that a marker will work properly on the array) were removed from consideration as potential tag SNPs. MaCH and minimac were used to determine which CAAPA variants were well-imputable based on the 1000 Genomes Phase I African Reference Panel and variants from the OmniExpress array. Those SNPs that could be imputed well (r^2 > 0.8) were removed from the pool of SNPs needing to be tagged.

Fugue was then used to determine the pairwise linkage disequilibrium (LD) between each pair of SNPs in the remaining set of SNPs. These LD estimates were used by FESTA, a TagSNP selection program, to select SNPs using a rather strict r^2 threshold of 0.8. A total of 1,004,268 TagSNPs were selected. Among them, 4,000 were removed because they were too similar in their probe design to function well on the array. Limited by the capacity of the array, only TagSNPs with an MAF >= 1% were retained for inclusion on the array. Raising the MAF floor was deemed to be the best approach to thinning the TagSNP pool as it was unbiased. Additional content was then added including both SNPs previously found to be associated with African-specific diseases but not previously selected as TagSNPs and approximately 600 additional SNPs in the human leukocyte antigen (HLA) region. HLA SNPs added to the array are relevant for HLA imputation and analyses of diseases or phenotypes related to immunity. These SNPs represent a pruned selection of SNPs with high tagging power, directly through LD, as well as SNPs preferentially selected by existing SNP-based HLA imputation algorithms. Finally, a “GWAS fingerprint” that includes 274 markers, identical to that used on the Illumina HumanExome array, was added to enable researchers to ensure accurate sample labeling and analysis, while facilitating sample tracking across experiments. In final count, 627,998 variants were included on the ADPC array Fig. 1.
Results
Based on the combination of OmniExpress and the ADPC, coverage is exceptionally good in both the 1000 Genomes African and admixed African populations. Figure 2 The average $r^2$ for all variants is greater than 0.8 at $\geq1\%$ MAF. Coverage is slightly better amongst admixed African populations than continental African populations, which is useful for the study of African Americans in particular. It is also not surprising, given that we had only one continental African population, compared to 15 African-admixed populations. In all populations, this represents a much better coverage level than with previous commercially available arrays, and represents an important step forward for studies of individuals of African descent.

Additional analysis of SNP coverage at $\geq1\%$ MAF in the CAAPA population was conducted to ensure our array will enable researchers to have sufficient power to identify novel associations between disease phenotypes and low frequency variants specific to African populations. Despite raising the MAF threshold for TagSNPs to 1.6%, we report coverage of all variants with MAF greater than or equal to 1% to give a full picture of the low frequency coverage the ADPC can provide. Genome-wide, the OmniExpress array is estimated to tag 20% of CAAPA variants at $r^2=0.9$, 26% at $r^2=0.8$ and 39% at $r^2=0.5$. All selected variants of the ADPC, alone, are estimated to tag 12% of known variants at $r^2=0.9$, 16% at $r^2=0.8$, and 31% at $r^2=0.5$. The combination of these two arrays is estimated to tag 29% of all CAAPA variants at $r^2=0.9$, 37% at $r^2=0.8$, and 56% at $r^2=0.5$, an improvement of about 50% more variants tagged over the OmniExpress array across the three thresholds Table 2.

While we consider these coverage statistics strong, they only refer to the coverage for variants identified through whole-genome sequencing in CAAPA. This is the most difficult possible test set, since coverage of more
common variants in the 1000 Genomes data is not included here. We use this information to give researchers an accurate view of the coverage available for the wealth of novel, low frequency genetic variation identified by CAAPAs whole-genome sequencing.

Using ~12,000 samples from the CAAPA consortium for the initial run of the ADPC array, ~495,000 out of 700,000 variants passed Illumina’s QC thresholds. This relatively high marker failure rate is not unexpected, however, because the array is comprised of nearly 100% novel markers, never before manufactured. The 494,094 markers successfully manufactured performed excellently, with missing genotype rate averaging only 0.3%.

Figure 2. Estimated imputation coverage of variants tagged by the ADPC content as part of the MEGA array. (a) Coverage in 1000 Genomes African populations is $\geq 0.8 \ r^2$ down to 1% MAF. (b) Coverage in 1000 Genomes admixed African populations is $\geq 0.8 \ r^2$ down to 1% MAF.
An important and unique feature of the ADPC is the significantly skewed MAF spectrum of variants on the array (Fig. 3). Compared to OmniExpress, the ADPC contains vastly more low frequency variants (Figure S2). This was not a conscious decision in the design process. Instead, it is the result of trying to tag a set of variants not previously tagged by commercially available genome-wide marker arrays. As a result, the combination of the ADPC and OmniExpress is an efficient pairing that increases coverage of the full MAF spectrum for novel variants and dramatically improves the imputation power for low frequency variants.

The ADPC content is currently available as part of the MEGA array through Illumina. That array combines the ADPC content described here plus a GWAS backbone based on OmniExpress, plus additional content from around the world to provide a single use array for researchers. Through the use of this array, researchers will have greater statistical power to find associations with complex diseases in populations of African ancestry, which has several practical benefits. This array will provide additional value from its tremendous improvement in the quality of imputed genotypes across the genome. At the same time, new researchers will be able to determine power before starting a study in populations of African ancestry, using the combination of the ADPC and OmniExpress on the MEGA array, leading to smaller sample sizes needed, and more studies being possible.

### Discussion

In this paper, we present the African Diaspora Power Chip, an affordable genotyping array that dramatically increases the coverage of genetic variants specific to African populations (and their descendants). Through the use of this array, researchers can now be better powered to detect disease associations in populations of African ancestry. For the CAAPA consortium, this means using the ADPC to genotype >13,000 Asthma cases and controls from 9 populations across the Americas.

Although this array was designed to meet the specific needs and timeline of the CAAPA consortium, and may not represent the ideal for a strictly African-based SNP array, its content is available to researchers now as part of the MEGA array and has demonstrably increased coverage in variants of West African ancestry. This is, of course, the most common African admixed population in African Americans or other members of the African diaspora, so it is especially well positioned to be useful in studies of admixed African Americans. This population, which has been under-covered by previously existing arrays should benefit greatly from the SNP content now available. Furthermore, the CAAPA consortium has released an imputation reference panel, based on the results of our whole-genome sequencing experiment, and it is available through the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html), and maximizes coverage provided by the ADPC content.

Our immediate plans are to assess the coverage provided by the ADPC content in populations of African descent not originating in West Africa. Of specific interest to geneticists are populations in East and North Africa. In the future, through the use of the MEGA array, the number of meaningful results from GWAS studies...
conducted in populations of African descent should increase significantly, providing a more accurate picture of causal disease variants in this group. This will enable personalized medicine techniques to be applied to a much larger subset of Americans than is currently feasible.

Methods

All study participants in the whole genome sequencing study provided written informed consent for the use of their DNA in genetic studies. A careful review was conducted to verify that the consents, study methods, and experimental protocols were consistent with the activities of this study. All methods were performed in accordance with the relevant guidelines and regulations. Institutional review board approval was obtained at Johns Hopkins University (GRAAD, BAGS, BIAS, HONDAS, PGCA), Howard University (GRAAD), Columbia University (REACH), Wake Forest University (SARP), Morehouse School of Medicine (COPDGene), Henry Ford Health System (SAPPHIRE), the University of California, San Francisco (coordinator center for SAGE II and GALA II), the Western Institutional Review Board for the recruitment in Puerto Rico (GALA II Puerto Ricans), Baylor College of Medicine from Texas, Albert Einstein College of Medicine Yeshiva University, Jacobi Medical Center, the North Central Bronx Hospital from New York (GALA II Dominicans), Children's Hospital and Research Center Oakland and Kaiser Permanente-Vallejo Medical Center (SAGE II), Vanderbilt University (BREATHE, VALID), the University of Chicago (CAG, AEGS), University of the West Indies, Mona campus (JAS) and Cave Hill Campus, Barbados (BAGS), The University of Cartagena (PGCA), the Universidad Católica de Honduras in San Pedro Sula (HONDAS), the Federal University of Bahia and endorsed by the National Commission for Ethics in Human Research in Brazil (BIAS, SCAALA), and The University of Ibadan, Nigeria (AEGS).

The 642 individuals in the data freeze were sequenced using Illumina's Hi-Seq 2000 equipment and the reads were 100bp, paired-end runs. Assembly was performed by the Consensus Assessment of Sequence and Variation (CASAVA) package, which is the Illumina in-house assembly and variant calling technology. The SNP-caller implemented in CASAVA uses a probabilistic model to ultimately generate probability distributions over all diploid genotypes for each site in each genome. A set of MAXGT quality scores is thus generated for each genomic site, corresponding to the ‘consensus quality’ in the SamTools SNP calling method. These quality scores are then parsed based on a set of consortium-wide rules in order to determine the likely set of variants.

Data processing to generate a 691-sample VCF file for each chromosome from the Illumina MAXGT single-sample SNP VCF files provided in Illumina's standard deliverable package was performed at Knome, Inc. (Cambridge, MA, USA). The individual VCF files only contained calls for variants, not ref/ref homozygotes. To generate a multi-sample VCF file, these individual VCF files were merged using VCFTools16 (v0.1.11), then using custom scripts, a multi-sample VCF file was backfilled to include homozygous reference genotypes and depth of coverage from the sites.txt files. Custom QC scripts confirmed the multi-sample VCFs and the single-sample VCFs had the same number of heterozygous and homozygous alternate genotypes. VCFTools was used to confirm all subjects were included in each multi-sample VCF. The multi-sample VCF was generated including the 48 samples from the SCAALA (Salvador, Brazil) group, but these samples were subsequently dropped from all analyses, leaving 642 individuals, and variants unique to SCAALA were removed from the variant pool. Mathias et al.4.

To pare down the list of variants needing to be tagged, several exclusion sets were created, starting with design score analysis. The segment extending 60 base pairs up and down stream from each variant position were sur - a 0–1 scale, representing the estimated success rate for the variant. Any variant scoring below 0.5 was removed. Any variant scoring below 0.5 was removed.

References

The authors gratefully acknowledge the contributions of Paul Levett, Anselm Hennis, P. Michele Lashley, Raana Naidu, Malcolm Howitt, and Timothy Roach (BAGS), Audrey Grant, Eduardo Viera Ponte, Alvaro A. Cruz, and Edgar Carvalho (BIAS), Susan Balcer-Whaley, Maria Stockton-Porter, and Mao Yang (GRAAD), Mario Meraz, Jaime Nuñez, Eileen Fabiani Herrera Mejía (HONDAS), Deanna Ashley (JAS), Silvia Jimenez, Nathalie Acevedo, Dilia Mercado (PGCA), Ann Jedlicka (REACH), Addison K. May, Caroline Gilmore, Patricia Minton (Vanderbilt University), Qun Niu, (University of Chicago), Adeyinka Falusi, Aboyami Odetunde (University of Ibadan, Nigeria). The authors also acknowledge the support of John Jay Shannon (Cook County Health Systems) and Kevin Weiss (Northwestern University) and the Indians Zemunes guards (San Basilio de Palenque, Bolivar, Colombia), Ulysses Ateba Ngoa (Leiden University), and Charles Rotimi, Ademayo Adebowale, Floyd J Malveaux, and Elena Reece (Howard University). We thank the numerous health care providers and community clinics and co-investigators who assisted in the phenotyping and collection of DNA samples, and the families and patients for generously donating DNA samples to BAGS, BIAS, BREATHE, CAG, GRAAD, HONDAS, REACH, SAGE II, VALID, SAPPHIRE, SARP, COPDGene, JAAS, GALLA II, PGCA, and AEGS. Special thanks to community leaders, teachers, doctors and personnel from health centers at the Garifuna communities for organizing the medical brigades and to the medical students at Universidad Católica de Honduras, Campus San Pedro y San Pablo for their participation in the fieldwork related to HONDAS; study coordinator Sandra Salazar, and the recruiters in SAGE and GALLA: Dunny Alva, MD, Gaby Ayala-Rodriguez, Ulysses Burley, Lisa Caine, Elizabeth Castellanos, Jaime Colon, Denise DeJesus, Iliana Flexas, Blanca Lopez, Brenda Lopez, MD, Luis Martos, Vivian Medina, Juana Olivo, Mario Peralta, Esther Pomares, MD, Jian Quaishi, Johanna Rodriguez, Shahdad Saeedi, Dean Soto, Ana Tavers, Emmanuel Viera, Dr. Michael LeNoir, Dr. Kelley Meade, Mindy Jensen, and Adam Davis; and health liaisons and public health officers of the main Conde office, Adalidades Concepción, Luciana Quintela, Ivanice Santos, Anali Lima, Benivaldo Valber Oliveira Silva, and Iraça Santos Araujo, and students from the Federal University of Bahia who assisted in data collection in BIAS: Rafael Santana, Roberta Barbosa, Ana Paula Santana, Carlton Barros, Marcele Brandão, Ludmila Almeida, Thiago Cardoso and Daniela Costa. We are grateful for the support from the international state governments and universities from Honduras, Colombia, Brazil, Gabon, Nigeria, Netherlands, Jamaica, Barbados and the United States who made this work possible. We also thank Robert Genuario for invaluable assistance in the whole genome sequencing at Illumina, Inc., Gonzalo Abecasis, William Cookson, Miriam Moffatt, for helpful discussions, Pat Oldewurtel and Murali Bopparaju for technical support, Shuai Yuan for assistance in the whole genome sequencing at Illumina, Inc., Gonzalo Abecasis, William Cookson, Miriam Moffatt, for helpful discussions, Pat Oldewurtel and Murali Bopparaju for technical support, Shuai Yuan for software support, and Kit Rees and Cate Kiefe for artistic contributions. We thank Steven Salzberg and Alex Szalay for computing and data storage resources available on the Data-Scope instrument at the Institute for Data Intensive Science (IDIES), Johns Hopkins University. The authors also acknowledge the support from James Kiley, Susan Banks-Schlegel, and Weiniu Gan at the National Heart, Lung, and Blood Institute. Funding for this study was provided by National Institutes of Health (NIH) R01HL104608. Additional NIH funding includes NC1, R21CA178706 (RDH), U01CA161032, P50CA125183 (OO). NCRR, G12RR030048 (GMD), R024975 (TH), NHGRI, R01HG007644, R21HG007233 (RDH), R21HG004751 (HRJ, JG, ZSQ), T32HG000044 (CRG). NHLBI, R01HL087699 (KCB), R01HL118267 (LKW), R01HL117004, R01HL088133, R01HL044646 (EGB), HL081332, HL112656 (LBW), R01HL69167, U01HL109164 (EB, DM), RC2HL101651, R2IHL101543, U01HL49596, R01HL072414 (CO), R01HL089897, R01HL089856, R01HL092601 (MGP), R01HL51492, R01HL/AI67905 (GHSF), HHSN26820130004C, HS28NH268201300047C, HHSN268201300048C, HHS28NH268201300049C, HHSN268201300050C (JGW). NIAID, R01AI07582 (TH), R01AI079139 (LKW), U19AI095230 (CO). NIHES, R01ES015794 (EGB). NIGMS, S06GM08016 (MUF), T23GM07157 (CRG). NIHMD, P06MD006902 (EGB), R01US44MD007588, R02MD0066881 (MGP). NSFC, #11344247 (RT). Additional sources of funding include: American Asthma Foundation (LKW and EGB), American Lung Association Clinical Research Grant (TH), Colombian Government (Colciencias) 331–2004 and 680–2009 (LC), EDCETP:CT:2011.40200.025 (AA), EU-IDEA HEALTH-F3-2009-241642 and EU-TheSchistoVac HEALTH-F3-2009-242107 (MY), Ernest Bazley Fund (PCA, RK, LG, RS), and the Fund for Henry Ford Hospital (LKW). The Jamaica 1986 Birth Cohort Study was supported by grants from the Caribbean Health Research Council, Caribbean Cardiac Society, National Health Fund (Jamaica) and Culture Health Arts Sports and Education Fund (Jamaica). Study nurses were supported by the University Hospital of the West Indies (TF, JKM), Ralph and Marion Falk Medical Trust (COO, OO, OO, GA), UCSF Dissertation Year Fellowship (CRG), Universidad Católica de Honduras, San Pedro Sula (EHP), University of Cartagena (JM), Wellcome Trust 072405/Z/03/Z, 088662/Z/09/Z (PJC). The Jackson Heart Study is supported by contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C from the NHLBI and the NIMHD. EGB was funded by Flight Attendant Medical Research Institute, RWJF Amos Medical Faculty Development Award and the Sandler Foundation; and the Sloan Foundation to RDH; CRG was supported in part by the UCSC Chancellor’s Research Fellowship and Dissertation Year.
Fellowship. KCB was supported in part by the Mary Beryl Patch Turnbull Scholar Program. RAM was supported in part by the MOSAIC Initiative Awards from Johns Hopkins University. MP-Y was funded by a Postdoctoral Fellowship from Fundación Ramón Areces. MIA is an investigator supported by National Council for Scientific and Technological Development (CNPq). TVH was supported in part by K24 AI 77930, UL1 TR00445, and U19 AI95227. RO was funded by NHLBI Diversity Supplement R01HL104608. Funding for the cohorts was provided by the following: AEGS, BAGS, BIAS, BREATHE (K08AI001582, RR24975), CAG, COPDGene, GALA II, GRAAD, HONDAS, JAAS (The Jamaica 1986 Birth Cohort Study was supported by grants from the Caribbean Health Research Council, Caribbean Cardiac Society, National Health Fund (Jamaica) and Culture Health Arts Sports and Education Fund (Jamaica). The study nurses were supported by the University Hospital of the West Indies.), PGCA (University of Cartagena and Colciencias Contracts 183–2002, 680–2009), REACH, SAGE II, SAPHIRE, SARP, SCAALA, VALID. Individual-level sequence data presented in this study are currently being prepared for deposition into dbGAP. Currently, these data are available through a data access agreement to respect the privacy of the participants for the transfer of genetic data, by contacting R.A.M. and K.C.B. and CAAPA (http://www.caapaproject.org).

Author Contributions
H.R.J. designed and performed experiments, analyzed data and wrote the manuscript. J.G., Y.J.H., Z.S.Q., R.A.M., T.H.B. and K.C.B. conceived the experiment and critically reviewed the manuscript. M.H., D.B., R.G. and C.L. provided technical assistance. P.A.G. and A.L. provided HLA expertise. T.O., G.A., E.K., G.W., C.G. and C.B. performed experiments, generated figures and assisted with the writing and review of the manuscript. CAAPA Consortium assisted in final interpretation of results and final manuscript preparation.

Additional Information
Accession codes: The whole genome sequence data that support the findings of this study have been deposited in dbGAP with the accession code phs001123.v1.p1.

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing Interests: Mark Hansen, Rob Genuario, Dave Bullis and Cindy Lawley are (or were) Illumina employees.

How to cite this article: Johnston, H. R. et al. Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome. Sci. Rep. 7, 46398; doi: 10.1038/srep46398 (2017).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2017