


RESEARCH ARTICLE

Open Access

# Copy number variation in human genomes from three major ethno-linguistic groups in Africa



Oscar A. Nyangiri<sup>1,2</sup>, Harry Noyes<sup>3</sup>, Julius Mulindwa<sup>1</sup>, Hamidou Ilboudo<sup>4</sup>, Justin Windingoudi Kabore<sup>5</sup>, Bernardin Ahouty<sup>6</sup>, Mathurin Koffi<sup>7</sup>, Olivier Fataki Asina<sup>8</sup>, Dieudonne Mumba<sup>8</sup>, Elvis Ofon<sup>9</sup>, Gustave Simo<sup>9</sup>, Magambo Phillip Kimuda<sup>1</sup>, John Enyaru<sup>10</sup>, Vincent Pius Alibu<sup>10</sup>, Kelita Kamoto<sup>11</sup>, John Chisi<sup>11</sup>, Martin Simuunza<sup>12</sup>, Mamadou Camara<sup>13</sup>, Issa Sidibe<sup>5</sup>, Annette MacLeod<sup>14</sup>, Bruno Bucheton<sup>13,15</sup>, Neil Hall<sup>3,16</sup>, Christiane Hertz-Fowler<sup>3</sup>, Enock Matovu<sup>1\*</sup>  and for the TrypanoGEN Research Group, as members of The H3Africa Consortium

## Abstract

**Background:** Copy number variation is an important class of genomic variation that has been reported in 75% of the human genome. However, it is underreported in African populations. Copy number variants (CNVs) could have important impacts on disease susceptibility and environmental adaptation. To describe CNVs and their possible impacts in Africans, we sequenced genomes of 232 individuals from three major African ethno-linguistic groups: (1) Niger Congo A from Guinea and Côte d'Ivoire, (2) Niger Congo B from Uganda and the Democratic Republic of Congo and (3) Nilo-Saharan from Uganda. We used GenomeSTRiP and cn.MOPS to identify copy number variant regions (CNVRs).

**Results:** We detected 7608 CNVRs, of which 2172 were only deletions, 2384 were only insertions and 3052 had both. We detected 224 previously un-described CNVRs. The majority of novel CNVRs were present at low frequency and were not shared between populations. We tested for evidence of selection associated with CNVs and also for population structure. Signatures of selection identified previously, using SNPs from the same populations, were overrepresented in CNVRs. When CNVs were tagged with SNP haplotypes to identify SNPs that could predict the presence of CNVs, we identified haplotypes tagging 3096 CNVRs, 372 CNVRs had SNPs with evidence of selection ( $iHS > 3$ ) and 222 CNVRs had both. This was more than expected ( $p < 0.0001$ ) and included loci where CNVs have previously been associated with HIV, Rhesus D and preeclampsia. When integrated with 1000 Genomes CNV data, we replicated their observation of population stratification by continent but no clustering by populations within Africa, despite inclusion of Nilo-Saharan and Niger-Congo populations within our dataset.

**Conclusions:** Novel CNVRs in the current study increase representation of African diversity in the database of genomic variants. Over-representation of CNVRs in SNP signatures of selection and an excess of SNPs that both tag CNVs and are subject to selection show that CNVs may be the actual targets of selection at some loci. However, unlike SNPs, CNVs alone do not resolve African ethno-linguistic groups. Tag haplotypes for CNVs identified may be useful in predicting African CNVs in future studies where only SNP data is available.

**Keywords:** CNV, Structural variation, Niger Congo A, Niger Congo B, Nilo-Saharan, Signatures of selection, Adaptation, Tag haplotypes

\* Correspondence: [matovue04@yahoo.com](mailto:matovue04@yahoo.com)

<sup>1</sup>College of Veterinary Medicine, Animal Resources and Biosecurity, Makerere University, P. O. Box 7062, Kampala, Uganda

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Copy number variants are defined as duplications or deletions of genomic segments greater than 1 kb in length [1]. While most genomic studies focus on single nucleotide variants (SNV), reports of larger genomic variants such as copy number variants (CNVs) are more limited [2]. However, given their size, CNVs cover more bases than SNV [2] and may have greater influence on gene expression and structure [3, 4]. These variations can also be associated with disease or adaptations to changing environments [5–7]. In addition, CNVs can be the functional variant underlying quantitative trait loci (QTL) found by genome wide association studies (GWAS).

African populations have the highest genomic diversity globally [8]. The four major ethno-linguistic groups in Africa are the Afro-Asiatic, Nilo-Saharan, Khoisan and Niger Congo, the latter of which consists of two major subdivisions; Niger-Congo-A and Niger-Congo-B [9]. These populations occupy diverse environments, have different cultures and ancestry and show stratification at genomic level [9]. Such genomic differences between groups may be associated with differences in susceptibility to infectious diseases such as malaria, tuberculosis and HIV [10] or environmental adaptations such as increases in copies of amylase genes associated with increased carbohydrate consumption [5, 11]. Studies of genomic variation such as CNVs in Africans may therefore help explain adaptation, population stratification and disease susceptibility.

African populations are under-represented in genomic studies [12], but are likely to harbour a large number of unique CNVs given their higher genomic diversity than European, American and Asian populations [8]. Here, we analyse whole genome sequence (WGS) data for CNVs in populations from Nilo-Saharan, Niger Congo A and Niger Congo B ethno-linguistic groups. Niger Congo A and Niger Congo B are the two largest linguistic groups in Africa. Niger Congo B is comprised of the Bantu languages and is a subgroup of Niger Congo A and therefore these two groups are a single lineage. We included the Nilo-Saharan Lugbara as an out group to make it possible to contrast diversity within the Niger-Congo populations with diversity between major linguistic groups.

The populations surveyed and their respective countries were: Ugandan Nilo-Saharans of Lugbara ethnicity (UNL,  $n = 50$ ); Niger-Congo-B speaking populations from Uganda (UBB,  $n = 33$ ) and the Democratic Republic of Congo (DRC,  $n = 50$ ); and Niger-Congo A speaking populations from Côte d'Ivoire (CIV,  $n = 50$ ) and Guinea (GAS,  $n = 49$ ). We aimed to discover novel CNV region (CNVR) variants, investigate population differences associated with CNVs and identify SNP haplotypes which tag CNVs and may predict such CNVs in future genome wide association studies (GWAS). The CNVs identified may also be important in understanding African CNV diversity and allowing inference of CNVs from population specific SNP-chip data.

## Results

### Participant characteristics

The countries of origin and ethnicities of participants are shown in Table 1 and a full list of the 232 samples is shown in Additional file 1. We used about 50 samples per population except for 33 from the Ugandan UBB population (Table 1). 50 samples provide a 95% chance of discovering CNVRs that have a frequency greater than 7%, while 232 samples give a 95% chance of detecting CNV with greater than 2% frequency.

### Identification of CNVs

To examine the distribution and extent of CNVs in human African populations, we selected 232 individuals from four countries (Table 1), representing Ugandan Nilo-Saharan population of Lugbara ethnicity (UNL); Niger-Congo B-speaking populations from Uganda (UBB) and Democratic Republic of Congo (DRC); Niger Congo A speakers from Côte d'Ivoire (CIV) and Guinea (GAS). Mean depth of sequence coverage was 10X and we used autosomal data only. We used two programs adapted for population scale data for CNV discovery: cn.MOPS and GenomeSTRiP, which have been benchmarked previously (see Materials and Methods). cn.MOPS calls CNVs based on read depth alone, whereas GenomeSTRiP combines read pairs, split reads, and read depth to generate CNV calls [14].

**Table 1** Ethnicity and origin of individuals analysed for CNV

Pop	Country	District	Ethno-linguistic group (ethnologue code, n)
UNL	Uganda	Maracha	Lugbara (IGG, 50)
UBB	Uganda	Iganga	Basoga (XOG, 33)
DRC	Democratic Republic of Congo	Bandundu	Kingongo (NOQ, 30) Kimbala (MDP, 20)
GAS	Guinea	Forecariah Boffa, Dubreka	Soussou (SUS, 49)
CIV	Côte d'Ivoire	Bonon Sinfra	Baoule (BCI, 11) Gouro (GOA 21) Moore (MOS, 12) Senoufo (SEF, 4) Malinke (LOI, 1) Koyaka (KGA, 1)

Ethnologue codes are derived from the ethnic languages of the world resource [13]

**Comparison of cn.MOPS and GenomeSTRiP**

Figure 1 summarizes the analysis workflow and Table 2 shows descriptive statistics for the CNVs predicted by the two methods. Additional file 2 and Figs S1 A & B give further details on comparison of CNV called by both methods. GenomeSTRiP detected 16,149 CNVRs compared to 9213 detected by cn.MOPS. The CNVR were filtered by removing 37 samples that appeared to be outliers on a multiple dimensional scaling plot (MDS) (Additional file 2: Fig S2). These outlier samples all had exceptionally high numbers of CNVRs, mean of outliers = 2718 compared with mean of retained = 548,  $p = 6.4e-09$  and also had higher inbreeding co-efficient (F) [15],  $F = 0.13$  for outliers compared with  $F = 0.04$  for non-outliers,  $p = 7.8e-05$ .

After removing the outliers, predicted CNVR retained for further analysis were 11,725 from GenomeSTRiP and 2115 from cn.MOPS. We defined as high confidence CNVRs those called by both GenomeSTRiP and cn.MOPS. This identified 7608 GenomeSTRiP CNVR that overlapped or were within cn.MOPS loci (Additional file 3). No CNVRs were predicted in a single sample only.

**Characteristics of CNVRs identified by GenomeSTRiP and cn.MOPS**

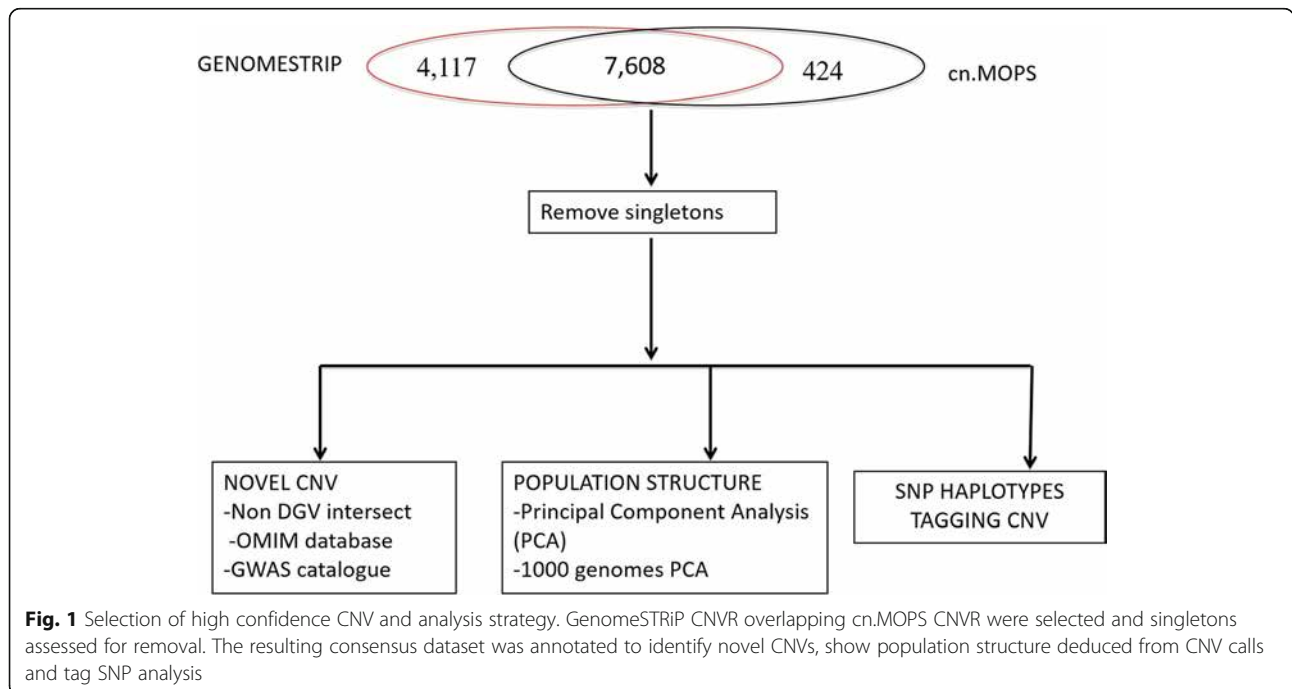
The CNVRs discovered by GenomeSTRiP (median length 5.2 kb) were much shorter than those discovered by cn.MOPS (median length 32 kb) (Table 2) and were more similar in length to those in the database of genomic variants (DGV; release date 2016-05-15) (median length 3.3 kb for CNVR > 1 kb) [16, 17].

GenomeSTRiP called more CNVRs (7608) than cn.MOPS (1691) and there were multiple GenomeSTRiP CNVRs within each cn.MOPS CNVR. The total lengths of CNVRs were 108 Mb and 1145 Mb in GenomeSTRiP and cn.MOPS, respectively. We found that 81 Mb (75%) of the GenomeSTRiP CNVRs were within cn.MOPS CNVRs, almost twice as much as the 43 Mb (40%) that was expected from random placement of the GenomeSTRiP CNVRs by simulation. Given that the GenomeSTRiP CNVRs conformed most closely in size to those described in DGV we used the GenomeSTRiP CNVRs for subsequent analysis. Amongst the 7608 CNVRs, there were 2172 CNVRs with only deletions, 2384 with only insertions and 3052 with both insertions and deletions. Counts of each class of CNV for each population are shown in Additional file 4.

24% of CNVRs were common to all three major linguistic groups represented in the data, 55% were unique to single linguistic groups and 21% were shared between pairs of major populations (Fig. 2a). Frequencies of shared CNVs were most correlated between Niger-Congo A and Niger-Congo ( $r^2 = 0.38$ ), and least correlated between Niger-Congo and Nilo-Saharan ( $r^2 = 0.17$ ). Individuals of Nilo-Saharan origin had the lowest proportion of private CNVRs (20%) whilst the Niger-Congo A and Niger-Congo B populations shared more with each other than with the Nilo-Saharans, consistent with their closer linguistic relationship.

**Genomic distribution of CNVR**

The density of CNVRs varied by about two-fold (1.43–2.41 CNVRs Mb<sup>-1</sup>) between the five populations



**Table 2** CNV statistics using GenomeSTRiP and cn.MOPS algorithms

Parameter	GenomeSTRiP	cn.MOPS	GenomeSTRiP that overlap cn.MOPS
Raw CNV regions (CNVR)	16,149	9213	
CNVR after QC	11,275	2115	7608
Total CNV scored	127,699	37,679	106,922
Deletion CNV	65,588	26,008	61,025
Gain CNV	62,111	11,671	45,897
Mean CNV count per CNVR	11.3	17.8	14.0
Mean CNVR per individual	654	193	548
Count of overlapping CNVRs <sup>a</sup>	7608	1691	7608
Mean Length of CNVR (kb)	9.5	541.7	10.7
SD length of CNVR (kb)	13.2	1287.6	14.1
Median Length of CNVR (kb)	5.3	32.4	6
Total Length of CNVR (Mb)	108.1	1145.8	81.2
Observed Length CNV present in both methods (Mb) (Simulated $\pm$ SD) <sup>b</sup>	81.2 (43.4 $\pm$ 1.0)		

Descriptive statistics of CNVR found using GenomeSTRiP and cn.MOPS. Note that: GenomeSTRiP has about 5.3 times the number of CNVs compared with cn.MOPS (11,275 cf. 2115); GenomeSTRiP CNVRs were shorter (median length 5.3 kb) than cn.MOPS (median length 32.4 kb); Total length of cn.MOPS CNVRs was about 10.6 times greater (1146 Mb cf. 108 Mb) than GenomeSTRiP CNVRs. CNVR = CNV region; a genomic location with chromosome, start and end base pair positions that has overlapping CNVs; CNVRs after QC = The CNVRs left after some CNVRs were dropped because they were only found in samples that were outliers in principal component analysis (PCA) plots of raw data. CNV count per CNVR = Number of samples with a CNV at each CNV region = Total CNVs count/ Total CNVRs; Mean CNVRs per sample = Count of CNV divided by number of samples; Mean, Standard deviation, Median, Total length, Observed length: Calculated per CNV not CNVR

<sup>a</sup>Count of any overlap (minimum 1 bp) between GenomeSTRiP and cn.MOPS CNVR

<sup>b</sup>The expected length of CNVs that would be found by both methods was obtained by 100 simulations using all the observed lengths of CNVs allocated to random places in the genome

(Additional file 2: Fig S3). The density of CNVRs also varied between chromosomes in both our data and 1000 Genomes data (Fig. 3) with the mean densities per chromosome correlated between both datasets ( $r^2 = 0.71$ ) (Fig. 4). The density of CNVs also varied across chromosomes (Additional file 2: Fig S3). The CNVRs per Mb ranged from a minimum of 5 in chromosome 18 to a maximum of 15 in chromosome 21. This trend was similar in counts of CNV calls per Mb with chromosome 18 displaying a minimum of 12 calls and 150 CNVs per Mb predicted on chromosome 21. We tested the 1000 genomes data for CNVR density by chromosome to confirm that variation in CNVR density is common in other datasets. The same phenomenon was observed with chromosomes 19 and 22 having high ( $\sim 24$  CNVRs  $\text{Mb}^{-1}$ ) numbers of CNVRs per Mb compared with other chromosomes ( $\sim 14$  CNVRs  $\text{Mb}^{-1}$ ) (Fig. 3).

#### Functional annotation of CNVR

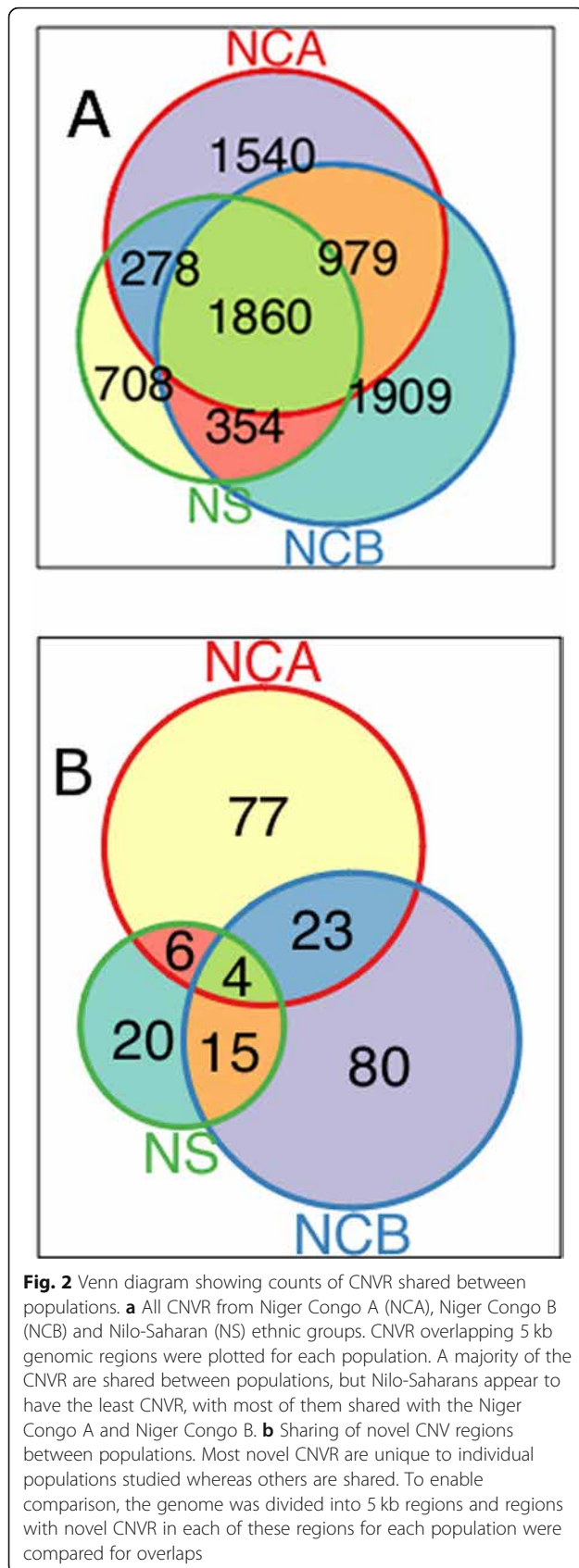
CNVRs were annotated with the classes of genomic features which they intersected. The most common annotations were coding and open chromatin regions (Additional file 2: Fig S4).

#### Novel CNV loci

We found 7384 of the 7608 final CNVRs analysis set overlapped known CNVRs in the human DGV and 224 (2.9%) had not been previously reported, and were defined as novel CNVRs. Unique CNVR boundaries in the

DGV cover 75% of the genome and much of the rest could be repeat regions where reads cannot be mapped with certainty and therefore CNVRs cannot be detected. CNVs in novel CNVRs were 10 times less frequently observed compared with CNV in known CNVR (mean frequency of novel CNVs was 0.74% compared with 7.4% for known CNVs). The novel CNVs were annotated using BEDTools intersect [18] against the list of Ensembl genes and regulatory regions (Additional file 5 and Fig S3B). We sought to clarify the frequency, likely functional roles and sharing of CNVRs between populations. Novel CNVRs were distributed throughout the genome at low frequencies (Fig. 5a). They intersected 293 unique genes or regulatory regions, with no specific function enriched and were not generally shared between the populations (Fig. 2b). When novel CNVRs intersecting protein coding genes were annotated in PANTHER [19] using gene ontology (GO) terms, 27% (30/109) of the novel CNVRs overlapped genes encoding binding function (GO: 0005488) and 20% (22/109) overlapped genes involved in catalytic activity (GO: 0003824). The novel CNVRs also overlap SNPs associated with traits in the genome wide association study catalogue (Additional file 2: Fig S5 and Additional file 6). Using BEDTools intersect; we found that both the known and novel CNVR overlapped Mendelian inheritance disease associated genes (Additional file 7).





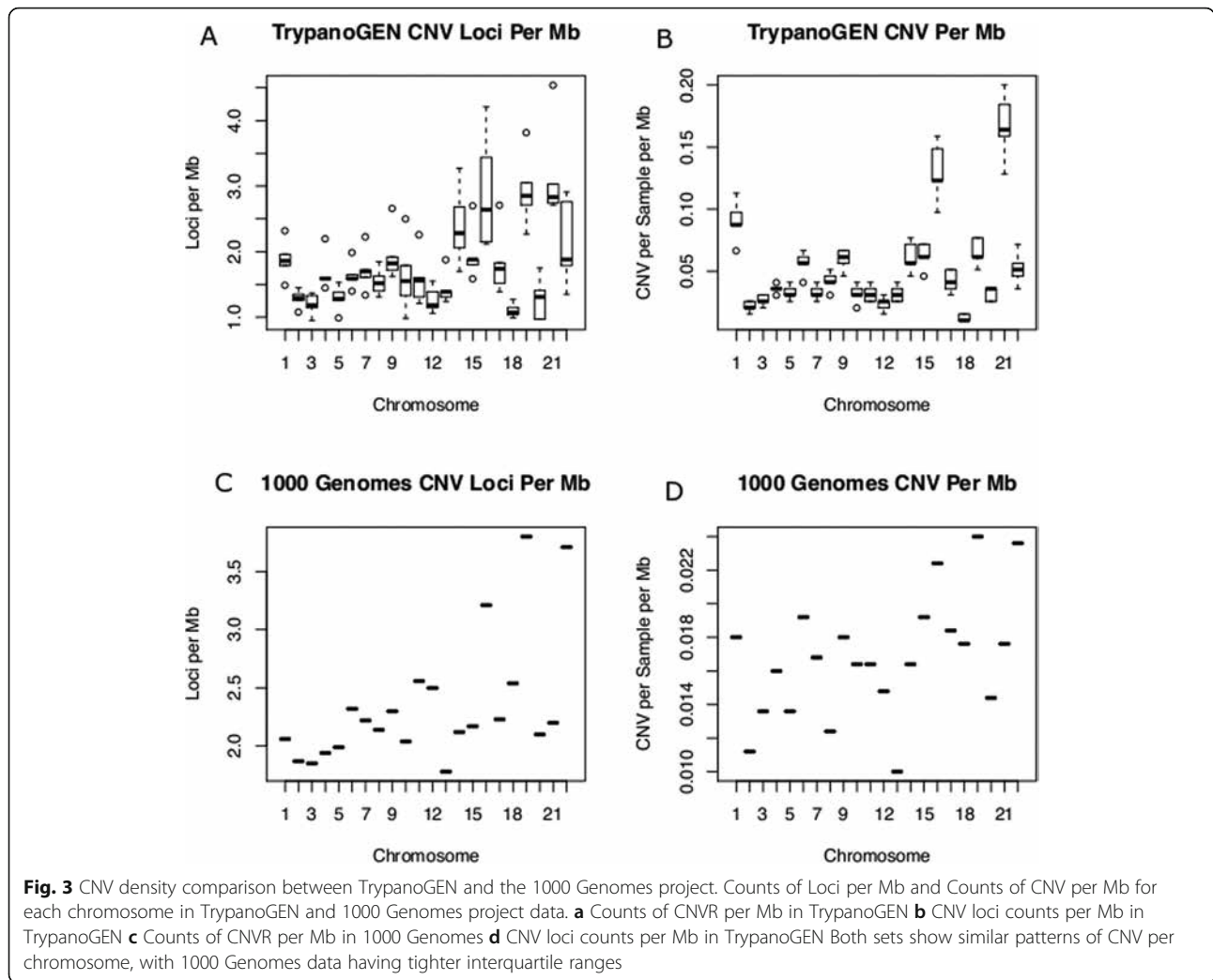
**Identification of haplotypes tagging CNVR**

SNP haplotypes that tag CNVRs in our populations were identified to assist the interpretation of SNP based GWAS studies. We assumed that if a haplotype is associated with a CNV then the number of alleles (0, 1, 2) of that haplotype will be correlated with the observed number of copies reported in samples in the dataset. Therefore, copy number is plotted against haplotype count for each sample and the value of  $r^2$  is calculated for the regression line and also the  $p$  value that the slope is zero. Haplotype blocks were defined using linkage disequilibrium ( $r^2 > 0.8$ ), which has been shown to tag shorter haplotypes in African American genomes compared to West Eurasians [20]. Alleles of 6942 haplotypes were associated with 3096 (41%) CNVRs as shown in Additional file 8. The mean count of CNVs at tagged CNVRs was 27.1 (CNV frequency = 12%) compared with 15.9 (7%) at untagged loci. The proportion of CNVRs that were tagged increased with frequency; less than 36% of CNVRs with CNV frequencies less than 10% were tagged but 64% of CNVRs with frequencies > 10% were tagged (Additional file 2: Fig S6). There was no difference between populations in the proportion tagged. Shorter (< 10 kb) CNVRs were less likely to be tagged (40% tagged) than longer (> 10 kb) CNVRs (49% tagged), reflecting the larger number of haplotypes found in longer CNVRs; there were a mean of 19 haplotypes in CNVRs < 10 kb and 37 haplotypes in CNVRs > 10 kb. Haplotypes that tag the CNVR detected in each of the five populations tested are shown in Additional file 8. The numbers of haplotype tagged CNVRs in each population were; 1286 (38.1%) in the CIV, 1540 (36.6%) in the DRC, 1261 (36.9%) in the GAS, 1169 (40.3%) in the UBB and 3200 (39.0%) in the UNL.

**CNVRs are overrepresented at loci under selection**

In order to identify CNVs with potentially functional effects we tested for association between CNVRs and loci that have been identified as under selection, with integrated haplotype score (iHS > 3.0) in the UNL population in a separate study of the same data [21]. There were 12,278 SNPs with evidence of selection ( $-\log_{10} iHS p > 3.0$ ), of these 1805 were within CNVRs, more than twice as many as would be expected by chance ( $\chi = 1822, p < 10^{-10}$ ) (Table 3), indicating a positive bias of selection on human CNVRs as shown in a previous study [22].

556 of the 1805 SNP with significant iHS scores were within 548 genes (+/- 5 kb flanks), including 146 protein coding genes (Additional file 9). The genes were classified by Ensembl Gene Type and the observed numbers of each gene type were compared with expected numbers from Ensembl (Table 4).



Immunoglobulin heavy chain variable and constant region genes were particularly overrepresented with 16 and 57 times as many genes in these classes as would be expected by chance. However, since these genes are found in tight clusters, the counts in CNVRs are not independent and this observation needs interpreting with some caution. Protein coding genes were underrepresented with 75% of the expected number.

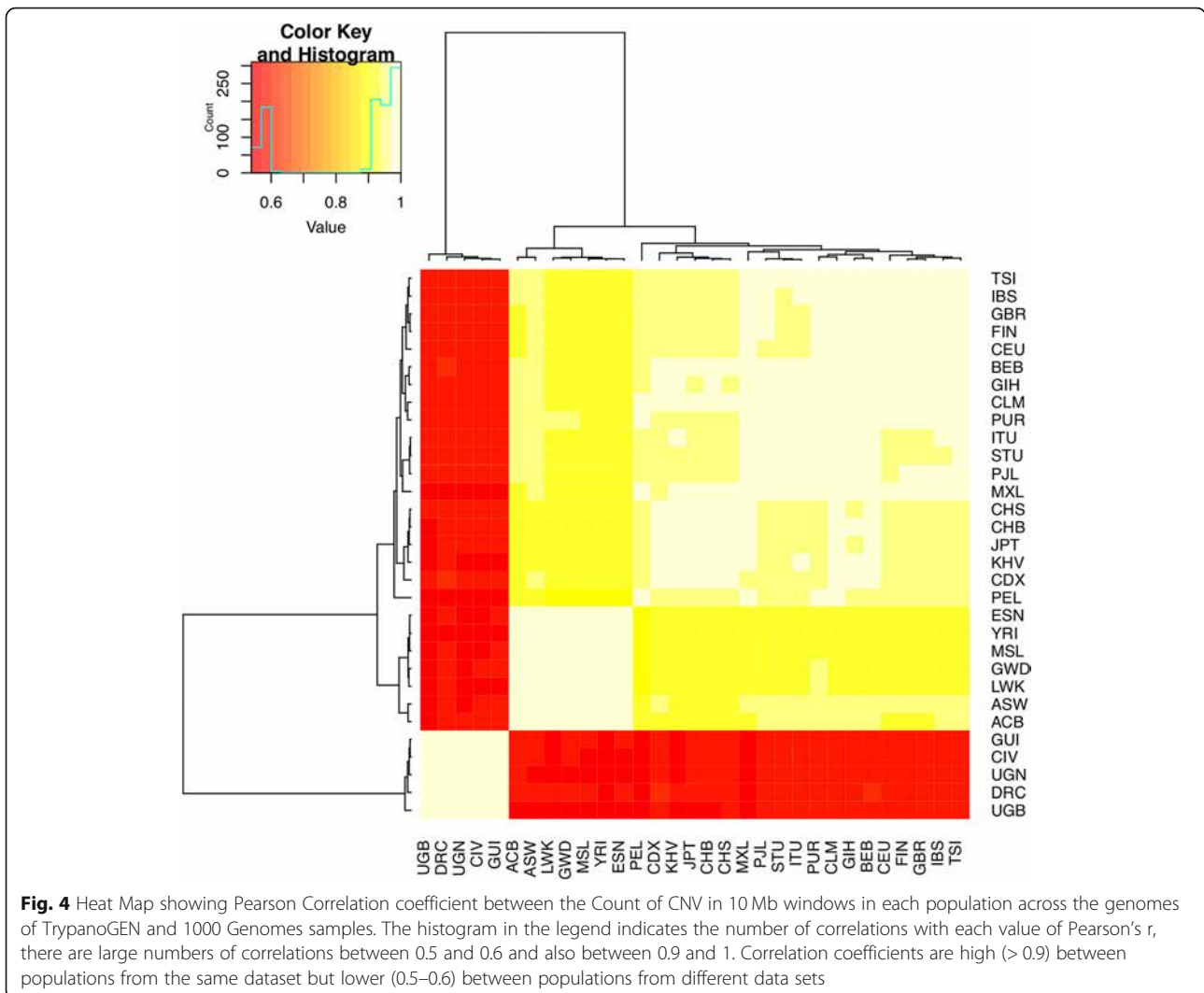
The mean frequency of CNVs in the CNVRs with SNPs under selection (19%) was twice that of CNVRs without SNPs under selection (8.5%) ( $\chi^2 = 11,673$ ;  $p < 10^{-10}$ , Table 5). CNVs may have been driven to higher frequency by selection in these populations.

There were 2693 CNVRs with SNPs that tag haplotypes in the UNL population and 372 CNVRs with SNPs with evidence of selection. Given that there was a total of 7608 CNVRs, 132 CNVRs would be expected to have both tag SNPs and SNPs with evidence of selection. However, 222 CNVRs

were observed with both tag SNPs and SNPs with evidence of selection, more than 50% as many as expected ( $p = 2.8^{-15}$ ) (Additional file 9 and Additional file 10). There was also a 32% excess of individual SNPs that both tagged CNVRs and had evidence of selection (16 expected; 22 observed) but this was not significant ( $p = 0.09$ ).

**Population structure and differentiation**

Principal Component Analysis (PCA) of combined 1000 Genomes and TrypanoGEN populations showed population structure at the continental level (East Asians, South Asians, Caucasians, Americans, Africans) Fig. 6a. However, there was no evidence of structure within most continental populations including Africans (Fig. 6a, b, c). Considering bi-allelic deletions only, the populations in our study here coincided with the 1000 Genomes African populations (Fig. 6b), but bi-allelic duplications revealed no population structure within Africa.



**Fig. 4** Heat Map showing Pearson Correlation coefficient between the Count of CNV in 10 Mb windows in each population across the genomes of TrypanoGEN and 1000 Genomes samples. The histogram in the legend indicates the number of correlations with each value of Pearson's  $r$ , there are large numbers of correlations between 0.5 and 0.6 and also between 0.9 and 1. Correlation coefficients are high (>0.9) between populations from the same dataset but lower (0.5–0.6) between populations from different data sets

$F_{ST}$  analyses of CNVs showed little difference ( $F_{ST} < 0.05$ ) between populations (Table 6). The Nilo-Saharan Lugbara from Uganda (UNL) were the most distinctive,  $F_{ST}$  between UNL and Niger-Congo populations were approximately double those amongst Niger-Congo populations.

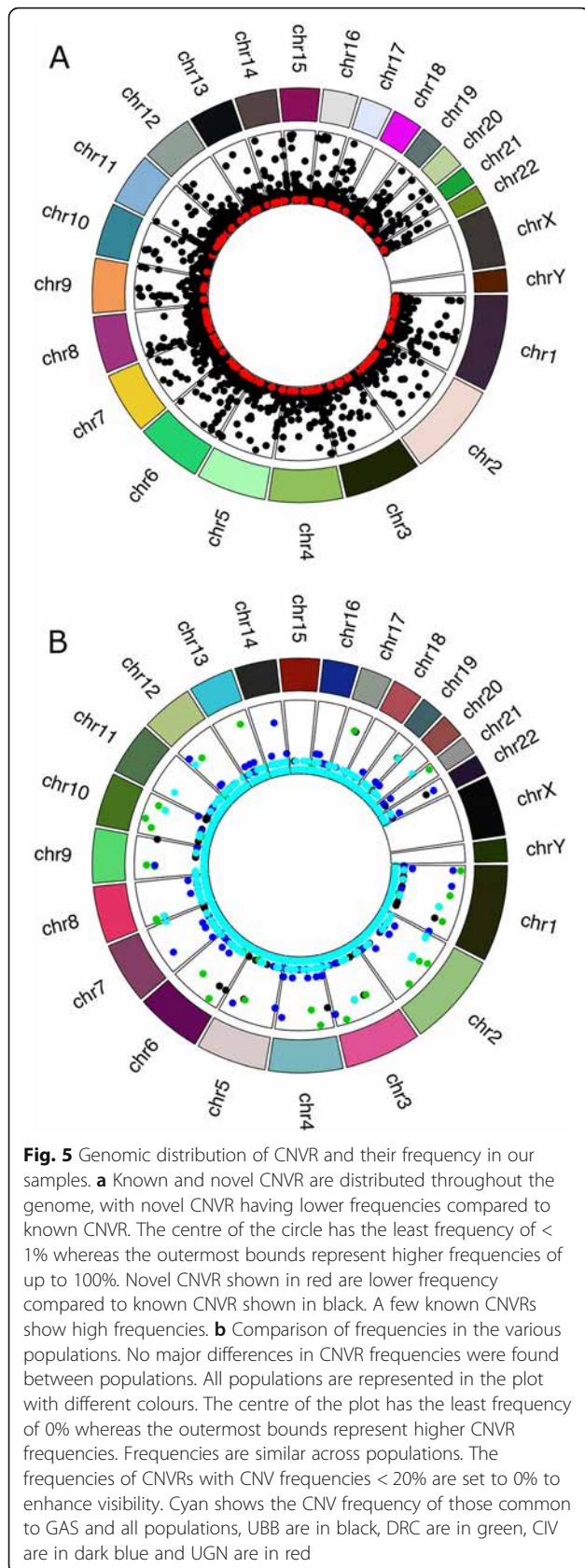
Although the mean  $F_{ST}$  across all CNVRs could not distinguish between populations 486 CNVRs show high  $F_{ST}$  (>3 standard deviations from the mean  $F_{ST}$ ) between populations. High  $F_{ST}$  loci (>3sd) intersected selected loci (iHS >3) within our data. CNVR regions with the highest  $F_{ST}$  difference between populations are annotated in Additional file 11. They overlap genes which have been associated with such disease; such as *UGT2B17* (UDP Glucuronosyltransferase Family 2 Member B17) associated with the bone mineral density quantitative trait locus and *IRGM* (Immunity-related GTPase family M protein) associated with inflammatory bowel disease 19.

## Discussion

### CNVR description and novel CNVRs

We identified 7608 consensus CNVs, using GenomeSTRiP and cnMOPS in five African populations. We only retained CNVRs that were called in more than one sample and were identified both by cn.MOPS and GenomeSTRiP. The cn.MOPS CNVRs were much larger, with a mean of 4.5 GenomeSTRiP CNVRs overlapping each cn.MOPS CNVR (Table 2). Given the better match of GenomeSTRiP CNVR size to the DGV CNVR size we interpreted this as evidence that cn.MOPS did not correctly identify CNVR breakpoints and had merged multiple independent CNVRs. cn.MOPS only uses read depth while GenomeSTRiP combines read pairs, split reads, and read depth to generate CNV calls [14]. It is known that the identification of breakpoints is more difficult with read depth dependent methods [24], but the large size difference suggests that cn.MOPS may have been missing breakpoints altogether and concatenating





**Table 3** Counts of SNPs inside and outside CNVRs with significant ( $-\log_{10} p > 3$ ) and non-significant  $p$  values

UNL CNV + 5 kb flanks	$-\log_{10} p > 3$	$-\log_{10} p < 3$
SNP in CNVR	1805	493,241
SNP not in CNVR	10,473	8,114,213

CNVRs were defined as the boundaries identified by GenomeSTRiP plus 5 kb upstream and downstream flanks to maintain consistency with the Tag SNP analysis

adjacent CNVRs. The differences in CNVs detected by GenomeSTRiP and cn.MOPS are consistent with reports observing that different algorithms for detecting CNVs from whole genome sequencing data show major differences in the CNVs detected [25]. Therefore, to minimise the risk of identifying CNVs at CNVRs that were an artefact of a particular algorithm we used the conservative approach of only identifying CNVs at CNVRs detected by both algorithms.

224 of the 7608 CNVR were defined as novel since they have not been previously submitted to the DGV. All the novel loci had low frequencies of < 10% (Fig. 5a). The locations of CNVR breakpoints are rarely identified very precisely making it difficult to distinguish homologous CNVR from merely overlapping ones [24]. We have taken a conservative approach to defining novel CNVRs by including only those that do not overlap known ones. Given that DGV CNVRs span 75% of the genome and that the remainder includes centromeres and telomeres and repeat regions the low proportion of CNVs in novel regions is not surprising. None of the novel CNVs in our data were common and less than 2% were shared between populations.

**Genomic distribution of CNVR**

There was a threefold variation in CNV and CNVR frequency per Mb between chromosomes in our dataset and a nearly twofold variation in the 1000 Genomes data, even after correction for chromosome length (Fig. 3). The density of CNVRs per Mb for each chromosome was correlated in the 1000 Genomes and our datasets ( $r^2 = 0.71$ ), suggesting that CNVR density may be an intrinsic property of chromosomes. The shorter chromosomes tended to have the higher densities of CNVRs in both our data and the 1000 Genomes data and a probe based study of CNVR distribution also found relatively high CNVR density on shorter chromosomes (15,16 and 22) [26]. Although the different studies found different chromosomes with maximum CNVR density, in all cases the highest densities were on the shorter chromosomes, and it is possible that these are more sensitive to structural variation or that shorter chromosomes have higher variance on these parameters. A cross species comparison would be required to test this hypothesis.



**Table 4** Classification of Genes in CNVR with evidence of selection

Type	Observed Count	Count in Ensembl	Ratio Observed: Expected
pseudogene	259	14,975	1.5
protein_coding	184	21,817	0.7
lincRNA	89	7177	1.1
IG_V_gene	25	138	16.1
IG_V_pseudogene	22	187	10.5
antisense	20	5339	0.3
miRNA	19	3243	0.5
snRNA	11	2001	0.5
processed_transcript	9	799	1.0
IG_C_gene	9	14	57.2
misc_RNA	8	2127	0.3

SNP with evidence of selection were annotated with a gene name if they were within 5 kb of the gene start or end. Counts of gene types were based on Ensembl annotation and the Count in Ensembl was the total number of each type recorded in Ensembl Biomart

### Tagging haplotypes to CNV

CNVs may be the functional variant underlying some QTLs discovered by genome wide association studies using SNP chips. In order to identify SNPs that could predict the presence of CNVs at a locus, we discovered haplotypes with alleles that were associated with CNVs at a locus. The haplotypes we predicted are not only associated with the presence or absence of the CNV locus but also the likely copy number. Previous studies have shown correlations between SNPs or haplotypes and CNVRs. High copy numbers of CNVs at the *HPR* (Haptoglobin related protein) locus have been tagged by haplotypes [27]. There is also strong correlation between alleles of a SNP and CNVs in the *CCL4* (Cysteine-Cysteine Ligand 4) chemokine gene [28].

In the current study, SNP haplotypes tagged 41% of CNVRs. The mean frequency of CNVs at tagged CNVRs (12%) was nearly twice that of untagged loci (7%). This may reflect a lower power to detect associations with rarer CNVs. Longer CNVRs tended to contain more haplotypes and a higher proportion of these were associated with copy number. The relationship between SNP haplotypes and CNVs could be confounded by the same CNV recurring on different haplotypes or by clustering overlapping non-homologous CNV into a single CNVR. The weak association between CNV genotype and population structure in the PCA analysis was consistent with both these hypotheses. Therefore, the number of CNVRs associated with SNP haplotypes may be an indicator of the proportion of stable, non-recurrent homologous, high frequency CNVRs.

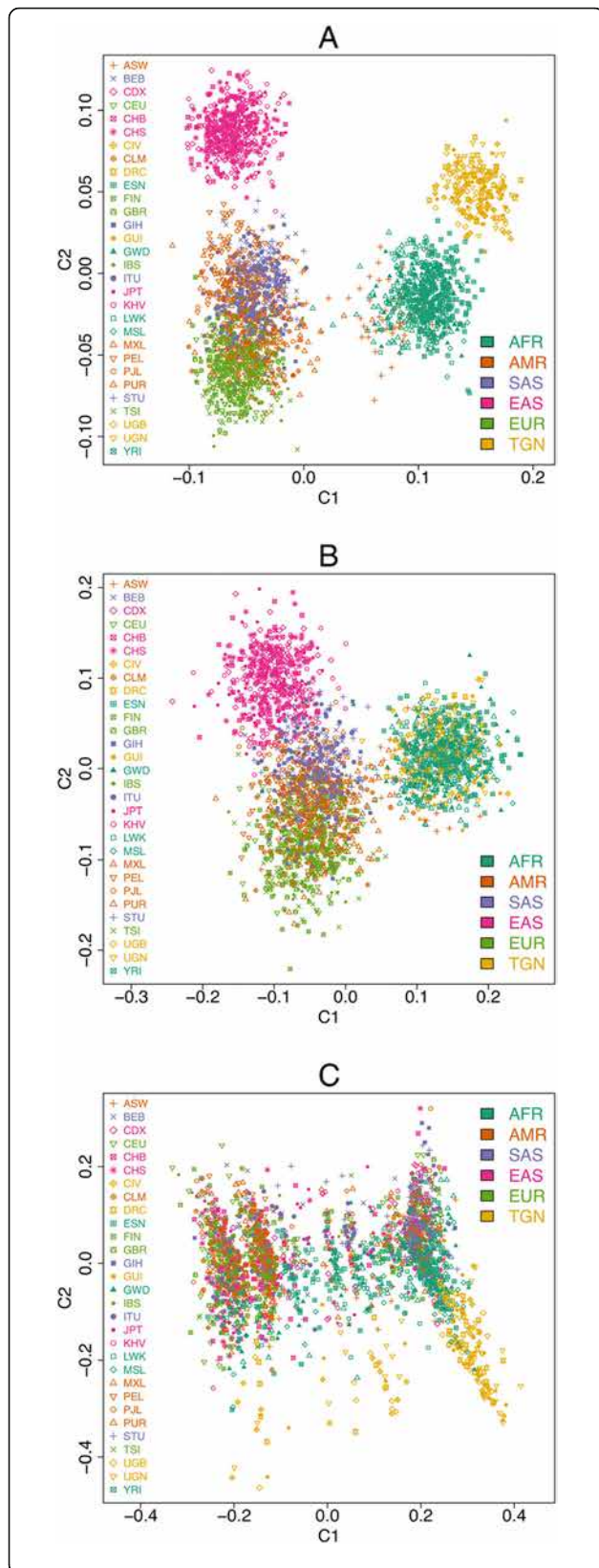
**Table 5** Counts of CNV at CNVR with and without SNP under selection

	Deletions	Wild Type	Insertions
CNVR with Selected SNP	2779	39,811	6534
CNVR without Selected SNP	83,003	1,566,194	63,553

### CNVs may be driving selection at some loci

The excess of CNVRs with both tagged haplotypes and SNPs with signatures of selection ( $-\log_{10} p_{iHS} > 3$ ) suggests that the CNV may be the genomic feature that is under selection at these loci. CNVs have been found to be the structure under selection by other methods [29] but this is the first time that we are aware that the combination of SNP signatures of selection combined with SNP haplotype tags have been used. This strategy makes it possible to use the extended haplotype homozygosity test to identify CNVRs under selection, which is more powerful than the previous methods based on  $F_{ST}$ . However, it should be noted that although there was a highly significant excess of CNVRs with SNPs that tagged CNVs and SNPs that had evidence of selection, the 32% excess of SNPs that were both Tag SNP and had evidence of selection was not significant. Therefore, it is possible that the tag SNP and the SNP with evidence of selection may both be correlated with some third factor other than CNV.

Among the loci that had CNVR overlapping selected loci were Rhesus D (RhD), *C1orf63* (Chromosome 1 Open Reading Frame 63), Human Leukocyte Antigen (*HLA*), Killer-cell Immunoglobulin-like Receptor (*KIR*). The complete deletion of the RhD gene is the commonest cause of Rhesus negative status. Given the severe consequences of the interaction between Rhesus negative ( $Rh^{-ve}$ ) mothers with Rhesus positive ( $Rh^{+ve}$ ) foetus it has been assumed that the null allele might be maintained by some yet unknown selective advantage. Genetic studies have found evidence for heterozygous advantage at the RhD locus in an ecological regression study [30] and an analysis of Rh blood group genes shows that they have experienced positive selection [31]. However, an evolutionary genetics study of the RhD genetics found no evidence for positive natural selection



**Fig. 6** PCA plot showing CNV population structure in our data compared to 1000 Genomes. The PCA distinguishes major continental populations from each other, but is not able to resolve specific populations within the continental populations. Africans in the 1000 Genomes (AFR) are closer to our data (TGN). Conventions for major continental populations are described by the 1000 genomes project [8, 23]. **b** PCA plot showing population structure for bi-allelic deletion CNV. Phase information is non-ambiguous for bi-allelic deletions. The Africans in the 1000 Genomes overlay the TrypanoGEN African samples, indicating similar CNV in the datasets. **c** PCA plot showing population structure due to bi-allelic insertion CNV. There was no specific pattern observed as fewer bi-allelic insertions were available in the data

affecting the frequency of the RhD selection [32]. These studies have mostly been conducted in European populations, our study has found evidence of both deletions and insertions at this locus so it is not clear which allele might be under selection.

Variants in the Human Leucocyte Antigen, class II, DQ beta 1 (*HLADQB1*) has been associated with pre-eclampsia in Iranian women [33]. Interestingly, the *HLA* locus interacts with the locus for *KIR* which has been associated with preeclampsia in Ugandan Bantu women [34]. *KIR3DL1* is also associated with risk of HIV [35]. Given the association of *HLADQB1* and *KIR* in pre-eclampsia and infectious disease which may impact infant birth and survival, they may be the actual targets of positive selection, resulting in the signatures of selection which have been seen in these loci. These observations generate useful hypothesis for testing. If QTL are discovered in these regions, then CNV should be high priority candidates for the functional variant.

**Population structure**

***WGS CNVs resolve continental populations but not intra-continental populations***

We found that CNV distinguish major continental populations, when we included Asians, South Asians, Americans, Europeans and Africans from the 1000 Genomes in the same PCA plot. Similarly, the 1000 Genomes project found CNV data resolve population clustering at

**Table 6**  $F_{ST}$  for CNVs computed from numbers of deletions per locus

	UNL	DRC	GAS	UBB	CIV
<b>UNL</b>	0				
<b>DRC</b>	0.004	0			
<b>GAS</b>	0.008	0.004	0		
<b>UBB</b>	0.004	0.003	0.004	0	
<b>CIV</b>	0.008	0.004	0.001	0.004	0

$F_{ST}$  were calculated in PLINK using only bi-allelic deletions since phase of these is known

continental scales [2]. Consequently, it appears that African populations may not be resolved using CNVR data, although the current study was limited to individuals of Nilo-Saharan, Niger-Congo A and Niger-Congo B origins and did not have access to the Afro-Asiatic or Khoisan populations. To confirm that the inability to resolve populations within Africa was not an artefact of the dataset, we combined our data with 1000 Genomes Project data and found that samples clustered by continent of origin but not at any finer scale. In contrast to these observations Nilo-Saharans, Niger Congo A and Niger Congo B have been shown to cluster separately in SNP based PCA [21]. CNV data therefore have low resolution in distinguishing intra-continental populations despite genomic CNV accounting for at least seven times more genomic base variation than SNP [36]. Several factors may account for the poor resolution of population structure analyses using CNV data including: 1) overlapping but distinct CNVs being coded as from a single CNVRs; 2) samples with different phase being coded the same, e.g. samples with 3:1 chromosomal copies being coded as 2:2; 3) recurrent CNVR at the same locus that do not correlate with population history. The potential importance of phase was illustrated by the better resolution obtained in both the PCA and the  $F_{ST}$  analysis using only CNVRs with complete deletions, which means that phase is known. Recurrent CNV at specific CNVR have been identified in different colonies of the same mouse strain and have been associated with disease in humans such as bronchopulmonary dysplasia (BPD) among premature infants [37, 38]. In a study of parent child trios up to 7% of variant loci in the child could not be associated with variants in the parents, which is indicative of novel or recurrent variants or alternatively, problems in variant genotyping [39]. This rate of recurrent mutation could quickly disrupt associations between variant genotypes and populations.

#### **Bias of selection in CNVR despite low $F_{ST}$**

Consistent with the PCA analysis, global  $F_{ST}$  showed that African CNV frequencies were similar across populations. However, there were 3–5-fold more CNVRs than expected by chance in regions where previous studies have found evidence for selection ( $p < 0.00001$ ). The higher than expected number of CNVRs in regions under selection suggests either the CNVRs are under selection or that selection signatures are more likely to arise in CNVRs regions. Consistent with the hypothesis that selection drives the enrichment of CNVRs in particular populations, the alpha thalassaemia deletion reveals signatures of selection [40] and is selected to high frequencies in

malaria endemic areas because it confers protection against infection by the malaria-causing parasite *Plasmodium* [41] and in our data the *KIR* locus had both SNP signatures of selection and SNP haplotypes tagging CNV. This suggests that CNVs are the polymorphism under selection in at-least some of these regions.

#### **Conclusion**

We have presented a CNVR landscape of populations representing the Niger-Congo A, Niger-Congo B and Nilo-Saharan African ethnic groups. These include known CNVRs that have been described in the DGV, and novel ones (3%), that are not reported in the DGV, reflecting the diverse nature of these African populations. Some of the CNVRs described may have medical significance as they occur in Mendelian disease-causing genes and overlap SNPs significantly associated with various traits in the GWAS catalogue. We have used haplotypes to tag CNVRs. Haplotypes tagging CNVRs are useful in imputing CNVs from SNP genotyping data in future studies, especially in African populations known to have low linkage disequilibrium. We found overrepresentation of CNVRs in regions showing signatures of selection in SNP based studies, and an excess of CNVRs with both haplotypes tagging CNVs and SNP haplotypes with signatures of selection, suggesting a possible role of CNVR in selection and adaptation. Finally, we show that CNV distinguish between continental populations but do not stratify within the continent, such as the Africans in the current study.

#### **Materials and methods**

##### **Population description**

The study was conducted in the context of the TrypanoGEN project [42], which aims to determine host genetic susceptibility to Human African Trypanosomiasis. Samples were selected from the TrypanoGEN bio-bank [42]. The populations included were from East, Central and West Africa. East African populations were the Ugandan Nilo-Saharan language speakers (Lugbara) ( $n = 50$ ) from Northern Uganda and Ugandan Niger-Congo B speakers (Basoga) ( $n = 33$ ) whereas Central African populations were Niger-Congo B speakers ( $n = 50$ ) from the Democratic Republic of the Congo. West African populations were Niger-Congo A speakers ( $n = 49$ ) from Guinea and Cote D'Ivoire ( $n = 50$ ). The samples in the current study are a subset of those described in the TrypanoGEN bio-bank [42]. Sample collection, ethical considerations and approvals have been previously described [42]. The summary of the population, linguistic group and sampling foci are in Table 1.

### Sequencing and SNP calling

We used the Illumina Truseq polymerase chain reaction (PCR) free kit to prepare WGS libraries. Sequencing was done at the Centre for Genomic Research at the University of Liverpool using the Illumina HiSeq2500 system at 10X coverage. We used Burrow Wheeler Alignment (BWA) to map sequenced reads onto the 1000 Genomes project human\_g1k\_v37\_decoy reference genome. The Genome analysis tool kit (GATK v3.4) was used for SNP calling following GATK best practice guidelines. Quality control measures of SNPs included filtering by a) removing loci with > 10% missing SNP, b) removing individuals with > 10% missing SNP loci and c) removing loci with Hardy Weinberg  $P$ -value < 0.01.

### CNV description and functional analysis

We use CNVR to refer to a locus at which one or more samples may have a CNV; the overlapping CNVs at a CNVR may each have different boundaries. To select methods to identify CNVR we reviewed the literature and found 4 major approaches to CNV discovery: 1) Paired end methods (PE) estimate insert size between the paired ends but is limited by the size of the fragment sequenced; 2) Split Read methods (SR) are focused on identifying exact break points in reads that do not map but have a mate pair that maps perfectly, it works well with deletions but can only identify small (< 50 bp) insertions; 3) Read Depth (RD) methods count the number of reads at each genomic location and 4) De Novo assembly methods create new genome assemblies and compare them with the reference and are best suited to high coverage with long read data [14]. Read Depth methods are most widely used for CNV discovery as they can estimate numbers of copies whereas PE and SR methods detect the presence of a variant but cannot quantify it. RD methods can be further subdivided into those that use a single sample, paired samples (e.g. parent child) or population samples. Of the six methods benchmarked recently by Trost and colleagues [43], only cn.MOPS and GenomeSTRiP use population scale data. Ideally, the performance of these algorithms should be evaluated against 'known' CNVR. This reference of 'known' CNVR is made when several algorithms are used to come up with consensus CNVR. To evaluate performance of algorithms, the results of each algorithm are compared to the consensus 'known' CNVR by calculating sensitivity (proportion of CNV in benchmark which an algorithm identifies) and false discovery rate (proportion of CNVs discovered by the algorithm that are not in the benchmark). Due to limited African CNV datasets, we referenced an evaluation of CNVR detection algorithms for sensitivity and false discovery rate against CNVR in the HuRef CNV Benchmark [43]. From these evaluations, two algorithms (GenomeSTRiP and cn.MOPS) integrated data from multiple samples

concurrently to discover CNVR and showed reasonable sensitivity and false discovery rates. GenomeSTRiP had sensitivity of 0.68; and a false discovery rate of 0.49, whereas cn.MOPS had sensitivity of 0.38; and a false discovery rate of 0.33. We therefore used GenomeSTRiP [27] and cn.MOPS [44] to detect CNVs in binary alignment map (BAM) files of our data. GenomeSTRiP has previously been used to detect CNVs in the 1000 Genomes project of human populations [27]. To validate detected CNVs we tested for overlap with published CNVs in the public Database of Genomic Variants (DGV; release date 2016-05-15) using BEDTools [18]. For GenomeSTRiP we used parameters recommended for 10X sequence data, whereas for cn.MOPS, we tested various parameters (Additional file 12). We annotated CNV overlaps with gene names from the UCSC genome browser [45, 46] and Ensembl Biomart [47] for Genome version hg19/GRCh37 using BEDTools [18].

### Population clustering (PCA)

We sought evidence for population structure due to CNVs using PCA in PLINK [48]. CNV data were first converted into multi-allelic genotype format and represented as 1 1 (0 copies), 1 2 (1 copy), 2 2 (normal copy number 2), 2 3 (3 copies), 3 3 (4 copies) up to a maximum of 4 4 for six copies of more. Since phase was not known we assumed that alleles were as equally distributed between chromosomes as possible. To merge with 1000 genomes data, common loci in both datasets were used. PLINK was used for population clustering as described in the documentation. We used the PLINK cluster command, which relies on identity by state values and Hamming distance to perform complete linkage clustering. R was used to visualise the resulting principal components.

### Population differentiation: $F_{ST}$ analysis

We investigated population differentiation by comparing  $F_{ST}$  between CNVs in the different populations. We used multi-allelic data format as described above for population differentiation ( $F_{ST}$ ) analysis.  $F_{ST}$  were calculated using PLINK v 1.9 [48] using only bi-allelic deletions since phase of these is known.

### Tag haplotypes for CNV

We used the method described by Handsaker et al. [13], implemented with a custom Java programme [49]. This method assumes that if a haplotype is associated with a CNV then the number of alleles (0,1,2) of that haplotype will be correlated with the observed number of copies reported for each sample. Therefore, copy number is plotted against haplotype count and the value of  $r^2$  is calculated for the regression line and the  $p$  value for the null hypothesis that the slope of the regression line is zero. We used a custom Java program; TagCNV available from Github [49] to find SNP haplotypes and test



their association with CNVs. To build haplotypes VCFtools was used to obtain the correlation ( $r^2$ ) between alleles in 50 kb windows. Sets of alleles within 5 kb of CNV boundaries and with  $r^2 > 0.8$  with at least one other SNP in the region were assembled into haplotypes. The counts of each allele were plotted against the fractional copy number for the CNV for each sample and the correlation ( $r^2$ ) between haplotype count and copy number was obtained using R and the association was considered significant if  $p < 0.05$  for the null hypothesis that the slope of the regression line is zero. A Bonferroni correction was applied for the number of haplotypes at the CNVR that were tested for association. Haplotypes which were only present in samples with identical copy numbers were considered uninformative and excluded from all calculations.

### Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-020-6669-y>.

**Additional file 1: Table S1.** List of samples in the study.

**Additional file 2.** Correlation of GenomeSTRIP and cn.MOPS and supplementary figures.

**Additional file 3: Table S2.** GenomeSTRIP CNVR that intersect cn.MOPS CNVR after QC.

**Additional file 4: Table S3.** Counts of CNV types detected by both algorithms.

**Additional file 5: Table S4.** Genes intersected by novel CNV.

**Additional file 6: Table S5.** Novel CNVR intersecting the GWAS catalogue SNPs.

**Additional file 7: Table S6A.** All CNVR intersecting OMIM genes B. Novel CNVR intersecting OMIM genes.

**Additional file 8: Table S7.** Haplotypes that tag CNVR in each of the populations.

**Additional file 9: Table S8.** Counts of CNVR and SNP with tagged SNP and SNP with Signatures of Selection (iHS > 3).

**Additional file 10: Table S9.** CNVR tagged by SNP haplotypes and also containing SNP signatures of selection.

**Additional file 11: Table S10.** CNVR with high  $F_{ST}$  between populations.

**Additional file 12: Table S11.** Effect of Prior Impact and Minimum width on concordance with DGV in cn.MOPS.

### Abbreviations

BAM: Binary Alignment Map; BCI: Baoule ethnologue code; BWA: Burrow Wheeler Alignment; *C1orf63*: Chromosome 1 Open Reading Frame 63; *CCL4*: Cysteine-Cysteine Ligand 4 chemokine gene; CIV: Côte d'Ivoire Niger Congo A speakers; CM: cn.MOPS; cn.MOPS: Copy number mixture of Poissons; CNV(s): Copy number variant(s); CNVR(s): Copy number variant region(s); DGV: Database of genomic variants; DRC: Democratic Republic of Congo;  $F_{ST}$ : F statistics; GAS: Guinea Niger Congo A speakers; GATK: Genome Analysis Tool Kit; GenomeSTRIP: Genome Structure in Populations; GS: GenomeSTRIP; GO: Gene ontology; GOA: Gouro ethnologue code; GWAS: Genome wide Association study; HIV: Human Immunodeficiency Virus; H3Africa: Human Heredity and Health in Africa; *HLADQB1*: Human Leucocyte Antigen, class II, DQ beta 1; *HPR*: Haptoglobin related protein; IGG: Lugbara ethnologue code; iHS: Integrated haplotype score; *IRGM*: Immunity-related GTPase family M protein; KEMRI: Kenya Medical Research Institute; KGA: Koyaka ethnologue code; *KIR*: Killer-cell Immunoglobulin like receptor; *KIR3DL1*: Killer cell immunoglobulin-like

receptor 3DL1; LOI: Malinke ethnologue code; MDP: Kingongo ethnologue code; MDS: Multidimensional scaling; MOA: More ethnologue code; NCA: Niger Congo A; NCB: Niger Congo B; NOQ: Kimbala ethnologue code; NS: Nilo Saharans; PCR: Polymerase chain reaction; PE: Paired end CNV detection methods; QC: Quality control; QTL: Quantitative Trait Loci; RD: Read depth; RhD: Rhesus D; SEF: Senoufo ethnologue code; SNP(s): Single Nucleotide Polymorphism(s); SR: Split read CNV detection methods; SUS: Soussou ethnologue code; UBB: Uganda Bantu Basoga; UCSC: University of California Santa Cruz; *UGT2B17*: UDP Glucuronosyltransferase Family 2 Member B17; UNL: Uganda Nilotic Lugbara; WGS: whole genome sequence; XOG: Basoga ethnologue code

### Acknowledgements

We are grateful to study participants for sample donation, field workers and hospital staff from the participating countries for their dedication in collecting and processing these specimens.

Membership of the TrypanoGEN network is at [www.trypanogen.net](http://www.trypanogen.net)

### Authors' contributions

EM, AM, IS, BB, CHF and NH conceived the study; EM, HI, MK, DM, GS, JE, MC, JC, MS, VPA, supervised sample collection; MPK, KK, BA, OFA, EO, JWK, collected samples and the related metadata; ON, JM and HN analysed the data, ON, HN, CHF and AM wrote the manuscript. All authors read and approved the final manuscript.

### Funding

This study constitutes output from the TrypanoGEN project; ID 099310/Z/12/Z funded by the Wellcome Trust as part of the H3Africa consortium [33]. Oscar Nyangiri was co-funded through a Wellcome Trust Strategic Award to the KEMRI-Wellcome Trust Research Programme (grant number 084538). This work was supported through the DELTAS Africa Initiative (Grant no. 107743). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS), Alliance for Accelerating Excellence in Science in Africa (AESA), and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant no. 107743) and the UK Government. The authors wish to acknowledge the use of The Uganda Medical Informatics Centre (UMIC) compute cluster. Computational support from Uganda Medical Informatics Centre (UMIC) was made possible through funding from the Medical Research Council (MRC) [MC\_EX\_MR/L016273/1]. The funders had no role in design, collection, analysis, interpretation or writing of this manuscript.

### Availability of data and materials

The sequences dataset used in this study is available from the European Genome-phenome Archive (EGA): EGAS00001002602. The program for tagging haplotypes to CNVR, TagHap.jar is available from Github [49]. GenomeSTRIP was downloaded from the GenomeSTRIP website [50]. cn.MOPS is also available from bio-conductor [51].

### Ethics approval and consent to participate

All study participants gave written informed consent to participate in the study. Ethical review committees of countries participating in TrypanoGEN approved the study. Respective Institutional Review Board approval numbers are as follows: Democratic Republic of Congo (No 1/2013), Guinea (1–22/04/2013), Côte d'Ivoire (2014/No 38/MSLS/CNER-dkn) and Uganda (HS 1344).

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>College of Veterinary Medicine, Animal Resources and Biosecurity, Makerere University, P. O. Box 7062, Kampala, Uganda. <sup>2</sup>Epidemiology and Demography Department, Kenya Medical Research Institute (KEMRI)/Wellcome Trust Research Programme, P.O. Box 230, Kilifi, Kenya. <sup>3</sup>Centre for Genomic Research, University of Liverpool, Liverpool L69 7ZB, UK. <sup>4</sup>Institut de Recherche en Sciences de la Santé (IRSS) - Unité de Recherche Clinique de Nanoro (URCN), Nanoro, Burkina Faso. <sup>5</sup>Centre International de Recherche-Développement sur l'Élevage en zones Subhumides (CIRDES), Unité des Maladies à Vecteurs et Biodiversité (UMaVeB), 01 BP 454,

Bobo-Dioulasso 01, Burkina Faso. <sup>6</sup>Felix Houphouët Boigny University (UFHB), Cocody, Abidjan, Côte d'Ivoire. <sup>7</sup>Université Jean Lorougnon Guédé (UJLoG) de Daloa, Daloa, Côte d'Ivoire. <sup>8</sup>Institut National de Recherche Biomedicale, Avenue de la Démocratie, Kinshasa Gombe, P. O. Box 1197, Kinshasa, Democratic Republic of Congo. <sup>9</sup>Faculty of Science, University of Dschang, P. O. Box 67, Dschang, Cameroon. <sup>10</sup>College of Natural Sciences, Makerere University, P. O. Box 7062, Kampala, Uganda. <sup>11</sup>College of Medicine, Department of Basic Medical Sciences, University of Malawi, Private Bag 360, Chichiri, Blantyre 3, Malawi. <sup>12</sup>Department of Disease Control, School of Veterinary Medicine, University of Zambia, P. O. Box 32379, Lusaka, Zambia. <sup>13</sup>Programme National de Lutte contre la Trypanosomose Humaine Africaine, BP 851, Conakry, Guinea. <sup>14</sup>Wellcome Centre for Molecular Parasitology, Institute of Biodiversity, Animal Health and Comparative Medicine, Garscube Estate, Glasgow G61 1QH, UK. <sup>15</sup>Institut de Recherche pour le Développement (IRD), IRD-CIRAD 177, TA A-17/G, Campus International de Baillarguet, F-34398 Montpellier, France. <sup>16</sup>Present address: Earlham Institute Norwich Research Park Innovation Centre, Colney Ln, Norwich NR4 7UZ, UK.

Received: 1 August 2019 Accepted: 12 March 2020

Published online: 10 April 2020

## References

- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. *Nature*. 2006;444:444–54.
- Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. *Science*. 2015;349:aab3761.
- Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526:75–81.
- Gamazon ER, Stranger BE. The impact of human copy number variation on gene expression. *Brief Funct Genomics*. 2015;14:352–7.
- Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. *Nat Genet*. 2007;39:1256–60.
- Hollox EJ, Hoh B-P. Human gene copy number variation and infectious disease. *Hum Genet*. 2014;133:1217–33.
- Lee C, Scherer SW. The clinical context of copy number variation in the human genome. *Expert Rev Mol Med*. 2010;12:e8.
- 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
- Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. *Science*. 2009;324:1035–44.
- Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African genome variation project shapes medical genetics in Africa. *Nature*. 2014;517:327–32.
- Inchley CE, Larbey CDA, Shwan NAA, Pagani L, Saag L, Antão T, et al. Selective sweep on human amylase genes postdates the split with Neanderthals. *Sci Rep*. 2016;6:37198.
- The H3Africa Consortium, Matovu E, Bucheton B, Chisi J, Enyaru J, Hertz-Fowler C, et al. Enabling the genomic revolution in Africa. *Science*. 2014;344:1346–8.
- Eberhard DM, Gary FS, Charles DF, (eds). *Ethnologue: Languages of the World*. Twentythird edition. 2020. <https://www.ethnologue.com/>. Accessed 20 Mar 2020.
- Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 2013;14 Suppl 11:S1.
- Wright S. Coefficients of inbreeding and relationship. *Am Nat*. 1922;56:330–8.
- MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42(Database issue):D986–92.
- DGV. Database of Genomic Variants. 2017. <http://dgv.tcag.ca/dgv/docs/InclusiveGain+Loss.hg19.2015-02-03.txt>. Accessed 5 Jul 2017.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics Oxf Engl*. 2010;26:841–2.
- PANTHER - Gene List Analysis. <http://www.pantherdb.org/>. Accessed 5 Jul 2019.
- Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A. Linkage disequilibrium patterns of the human genome across populations. *Hum Mol Genet*. 2003;12:771–6.
- Mulindwa J, Noyes HA, Ilboudo H, Nyangiri O, Koffi M, Mumba D, et al. Evidence of population specific selection inferred from 289 genome sequences of Nilo-Saharan and Niger-Congo linguistic groups in Africa. *bioRxiv*. 2017. <https://doi.org/10.1101/186700>.
- Nguyen D-Q, Webber C, Ponting CP. Bias of selection on human copy-number variants. *PLoS Genet*. 2006;2:e20.
- Population | 1000 Genomes. <https://www.internationalgenome.org/category/population/>. Accessed 27 Feb 2020.
- Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 2011;21:974–84.
- Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*. 2014;15:256–78.
- Kato M, Kawaguchi T, Ishikawa S, Umeda T, Nakamichi R, Shapero MH, et al. Population-genetic nature of copy number variations in the human genome. *Hum Mol Genet*. 2010;19:761–73.
- Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic copy number variations in humans. *Nat Genet*. 2015;47:296–303.
- Colobran R, Comas D, Faner R, Pedrosa E, Anglada R, Pujol-Borrell R, et al. Population structure in copy number variation and SNPs in the CCL4L chemokine gene. *Genes Immun*. 2008;9:279–88.
- Iskrow RC, Gokcumen O, Lee C. Exploring the role of copy number variants in human adaptation. *Trends Genet TIG*. 2012;28:245–57.
- Flegr J. Heterozygote advantage probably maintains rhesus factor blood group polymorphism: ecological regression study. *PLoS One*. 2016;11:e0147955.
- Kitano T, Saitou N. Evolution of Rh blood group genes have experienced gene conversions and positive selection. *J Mol Evol*. 1999;49:615–26.
- Perry GH, Xue Y, Smith RS, Meyer WK, Caliskan M, Yanez-Cuna O, et al. Evolutionary genetics of the human Rh blood group system. *Hum Genet*. 2012;131:1205–16.
- Mohammadi M, Farazmandfar T, Shahbazi M. Relationship between human leukocyte antigen (HLA)-DQA1\*0102/HLA-DQB1\*0602 polymorphism and preeclampsia. *Int J Reprod Biomed Yazd Iran*. 2017;15:569–74.
- Nakimuli A, Chazara O, Hiby SE, Farrell L, Tukwasibwe S, Jayaraman J, et al. A KIR B centromeric region present in Africans but not Europeans protects pregnant women from pre-eclampsia. *Proc Natl Acad Sci U S A*. 2015;112:845–50.
- Pelak K, Need AC, Fellay J, Shianna KV, Feng S, Urban TJ, et al. Copy number variation of KIR genes influences HIV-1 control. *PLoS Biol*. 2011;9:e1001208.
- Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. *Nature*. 2010;464:704–12.
- Egan CM, Sridhar S, Wigler M, Hall IM. Recurrent DNA copy number variation in the laboratory mouse. *Nat Genet*. 2007;39:1384–9.
- Ahmad A, Bhattacharya S, Sridhar A, Iqbal AM, Mariani TJ. Recurrent copy number variants associated with bronchopulmonary dysplasia. *Pediatr Res*. 2016;79:940–5.
- Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun*. 2019;10:1784.
- Qiu Q-W, Wu D-D, Yu L-H, Yan T-Z, Zhang W, Li Z-T, et al. Evidence of recent natural selection on the southeast Asian deletion (–(SEA)) causing  $\alpha$ -thalassaemia in South China. *BMC Evol Biol*. 2013;13:63.
- Flint et al. High frequencies of alpha-thalassaemia are the result of natural selection by malaria. - PubMed - NCBI. <https://www.ncbi.nlm.nih.gov/pubmed/3713863>. Accessed 27 Mar 2019.
- Ilboudo H, Noyes H, Mulindwa J, Kimuda MP, Koffi M, Kaboré JW, et al. Introducing the TrypanoGEN biobank: a valuable resource for the elimination of human African trypanosomiasis. *PLoS Negl Trop Dis*. 2017;11:e0005438.
- Trost B, Walker S, Wang Z, Thiruvahindrapuram B, MacDonald JR, Sung WWL, et al. A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. *Am J Hum Genet*. 2018;102:142–55.
- Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res*. 2012;40:e69.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res*. 2002;12:996–1006.
- Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, et al. The UCSC genome browser database: 2017 update. *Nucleic Acids Res*. 2017;45:D626–34.

47. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. *Nucleic Acids Res.* 2019;47:D745–51.
48. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559–75.
49. Noyes H. Tag Copy Number Variations (CNV) with SNP haplotypes. 2018. <https://github.com/LiverpoolHarry/TagCNV>. Accessed 2 May 2018.
50. Genome STRiP | GenomeSTRiP. <http://software.broadinstitute.org/software/genomestrip/>. Accessed 5 Jul 2019.
51. cn.mops. Bioconductor. <http://bioconductor.org/packages/cn.mops/>. Accessed 5 Jul 2019.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

