Genet. Res., Camb. (2015), vol. 97, e18. © Cambridge University Press 2015
doi:10.1017/S0016672315000191

Global patterns of large copy number variations in the human genome reveal complexity in chromosome organization

AVINASH M. VEERAPPA[1], RAVIRAJ V. SURESH[1]†, SANGEETHA VISHWESWARAIAH[1]†, KUSUMA LINGAIAH[1]†, MEGHA MURTHY[1]†, DINESH S. MANJEGOWDA[2], PRAKASH PADAKANNAYA[3] AND NALLUR B. RAMACHANDRA[1]*

[1]*Genetics and Genomics Lab, Department of Studies in Genetics & Genomics, University of Mysore, Manasagangotri, Mysore-06, Karnataka, India*
[2]*Nitte University Centre for Science Education & Research, K. S. Hegde Medical Academy, Nitte, University, Deralakatte, Mangalore-18, Karnataka, India*
[3]*Department of Studies in Psychology, University of Mysore, Manasagangotri, Mysore-06, Karnataka, India*

(*Received 18 November 2014; revised 27 May 2015; accepted 16 August 2015*)

Summary

Global patterns of copy number variations (CNVs) in chromosomes are required to understand the dynamics of genome organization and complexity. For this study, analysis was performed using the Affymetrix Genome-Wide Human SNP Array 6.0 chip and CytoScan High-Density arrays. We identified a total of 44 109 CNVs from 1715 genomes with a mean of 25 CNVs in an individual, which established the first drafts of population-specific CNV maps providing a rationale for prioritizing chromosomal regions. About 19 905 ancient CNVs were identified across all chromosomes and populations at varying frequencies. CNV count, and sometimes CNV size, contributed to the bulk CNV size of the chromosome. Population specific lengthening and shortening of chromosomal length was observed. Sex bias for CNV presence was largely dependent on ethnicity. Lower CNV inheritance rate was observed for India, compared to YRI and CEU. A total of 33 candidate CNV hotspots from 5382 copy number (CN) variable region (CNVR) clusters were identified. Population specific CNV distribution patterns in p and q arms disturbed the assumption that CNV counts in the p arm are less common compared to long arms, and the CNV occurrence and distribution in chromosomes is length independent. This study unraveled the force of independent evolutionary dynamics on genome organization and complexity across chromosomes and populations.

1. Introduction

The migration of humans into diverse environments has subjected the human genome to the force of various adaptations. Indexing the type and frequency of variations will help determine the contributions of the prevailing adaptive nature of the human genome towards phenotypic diversity. One such variation is the copy number variation (CNV), which forms an informative tool for assessing the alterations in the genome, since they leave CN signatures behind, which help to track and understand the effects of such variations on the genome (Kang *et al.*, 2008; Simonson *et al.*, 2010; Lou *et al.*, 2011; Zhang *et al.*, 2012; Veerappa *et al.*, 2013 *b*). The size and location of the variation determines the impact on the phenotype (Veerappa *et al.*, 2014). CNVs, especially through gene duplication and exon shuffling, drive gene and genome evolution, which can be seen through the frequent recurrence of CNVs in regions of the genome that bear various multigene families (Niimura and Nei, 2003; Go and Niimura, 2008; Sudmant *et al.*, 2010; Kim *et al.*, 2012; Veerappa *et al.*, 2013 *c*). However, CNVs that are rare and benign variants are usually more recent and may have been *de novo* in origin (Xu *et al.*, 2008), and have been observed in patients with mental retardation, developmental delay, dyslexia, schizophrenia and autism (de Vries *et al.*, 2005; Sharp *et al.*, 2006; Sebat *et al.*, 2007; Walsh *et al.*, 2008; Veerappa *et al.*, 2013 *b*; Veerappa *et al.*, 2013 *d*).

* Corresponding author: Nallur B. Ramachandra, Genetics and Genomics Lab, Department of Studies in Genetics & Genomics, University of Mysore, Manasagangotri, Mysore-06, Karnataka, India. Tel: +91-821-2419781/2419888. Fax: +91-821-2516056. E-mail: nallurbr@gmail.com
† These authors contributed equally to this work.

Both low and high resolution CNV studies have been performed across control population cohorts' samples since 2003, majorly covering Africa, America, Europe, China, Tibet, Taiwan, India, Germany and Finland (Lin *et al.*, 2008; McElroy *et al.*, 2009; Chen *et al.*, 2011; Lou *et al.*, 2011; Gautam *et al.*, 2012; Zhang *et al.*, 2012; Kanduri *et al.*, 2013; Liu *et al.*, 2013). In view of this, the rationale for conducting this study are as follows: (i) to identify the number, size and frequency of the CNVs in the genome; (ii) to construct a high resolution, regional as well as a global CNV map; (iii) to characterize CNVs in the imperative chromosomal structures, pseudoautosomal regions (PARs) and telomeres; (iv) to detect CNV hotspot regions in the chromosomes; (v) to identify the presence or absence of sex bias in distribution of CNVs; and (vi) to calculate the CNV inheritance and *de novo* mutation rate in family trios.

2. Materials and methods

For this study, a total of 1767 individuals, including 43 normal members from 12 randomly selected families residing in Karnataka, India, with family members of different age groups ranging from 13–73 years, 270 HapMap samples covering CEPH collection (CEU), Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT) and Yoruba in Ibadan, Nigeria (YRI) populations, 31 Tibetan samples, 155 Chinese samples, 471 Ashkenazi Jews from replicate 1, 482 Ashkenazi Jews from replicate 2, 204 individuals from Taiwan, 64 from Australia and 47 from New World populations (Totonacs and Bolivians), were selected for the CNV analysis in the genome. A total of 5 ml EDTA blood was collected from each member of the Indian study group and genomic DNA was extracted using the Promega Wizard® Genomic DNA purification kit. The isolated DNA was quantified by Bio-photometer and gel electrophoresis. This research was approved by the University of Mysore Institutional Human Ethics review committee (IHEC). Written informed consent was obtained from all sample donors and the IHEC approved the sample consent procedure. Written informed consent was obtained from parents/guardians in the cases of participants being minors. Data for the 270 individuals from the four HapMap populations was obtained from the International HapMap Consortium (The International HapMap Consortium, 2003). The samples from HapMap populations come from a total of 270 people: 30 both-parent-and-adult-child trios from the YRI, 45 unrelated JPT, 45 unrelated CHB, and 30 both-parent-and-adult-child trios from CEU. The raw, unprocessed data from the Affymetrix Genome-Wide SNP 6.0 array for the following datasets were obtained from the ArrayExpress archive with the accession numbers E-GEOD-21661, E-GEOD-29851, E-GEOD-30481, E-GEOD-15826, E-GEOD-23636, E-GEOD-23201, E-GEOD-33355 and E-GEOD-33356. The data have been made publicly accessible through the University of Mysore Genome Centre Database (http://umgc.uni-mysore.ac.in/index.php/search/cnv). The CNV data from these populations had not been analysed by the original authors and had remained largely unexplored as these were the control subsets of the experiment. The CNVs identified in this study are highly consistent, because of the higher stringency adopted in both the selection and validation of CNVs using multiple algorithms. Although these samples are well characterized, no medical information (except for HapMap) was obtained, meaning that structural variation ascertained is not necessarily benign or neutral. The Ashkenazi Jews II dataset works as a negative control for the comparative analysis showing very similar data points with negligible deviations. Similarity of data points between the Ashkenazi Jews datasets validates the homogeneity in the sampling process of the ethnic group.

(i) *Genotyping*

The human genome build hg18 was used as a reference genome. Genome-wide genotyping was performed using an Affymetrix Genome-wide Human SNP Array 6.0 chip and Affymetrix CytoScan® High-Density (HD) array having 1·8 million and 2·6 million combined SNP and CNV markers with the median inter-marker distance of 500–600 bases. These chips provide maximum panel power and the highest physical coverage of the genome (Affymetrix, Inc., 2005; Affymetrix, Inc., 2007; Affymetrix, Inc., 2008 *a*; Affymetrix, Inc., 2009). Genotyping quality was assessed using Affymetrix Genotyping Console Software. Copy number analysis method offers two types of segmenting methods, univariate and multivariate. These methods are based on the same algorithm, but use different criteria for determining cut-points denoting CNV boundaries. CNVs >100 kb were only included in the analysis, ruling out the 1–99 kb range CNVs, as we believe including these will only create more background noise with just too many false positives.

(ii) *Algorithms for CN state calling*

(a) *BirdSuite (v2)*

BirdSuite (BirdSuite Algorithm, 2010) is a suite originally developed to detect known common copy number polymorphisms (CNPs) based on prior knowledge as well as to discover rare CNVs from Affymetrix SNP 6.0 array data. To do this, it incorporates two main methods; the 'Birdsuite' algorithms and the 'Canary' (Affymetrix, Inc., 2008 *b*). The

https://doi.org/10.1017/S0016672315000191 Published online by Cambridge University Press

Birdsuite algorithm uses a hidden Markov model (HMM) approach to find regions of variable CN in a sample. For the HMM, the hidden state is the true CN of the individual's genome and the observed states are the normalized intensity measurements of each array probe. CNV calls from the Canary and Birdsuite algorithms were collated for each sample and kept as long as they met the following criteria: (i) Birdsuite calls with a log10 of odds (LOD) score (odds ratio) greater than or equal to ten (corresponding to an approximate false discovery rate of ~5%); (ii) Birdsuite calls with CN states other than two were retained; and (iii) Canary CNP calls with CN states different from the population mode were retained.

(b) *Canary*

CNP analysis was performed using the Canary algorithm. Canary was developed by the Broad Institute for making CN state calls in genomic regions with CNPs. The Canary algorithm computes a single intensity summary statistic using a subset of manually selected probes within the CNP region. The intensity summaries are compared in aggregate across all samples to intensity summaries previously observed in training data to assign a CN state call.

(c) *CNVFinder*

CNVFinder developed at the Wellcome Trust Sanger Institute uses a dynamic, multiple-threshold based approach to allow robust classification of CN changes in data of varying qualities. This algorithm makes two main assumptions: (i) that the majority of data points are normally distributed around a log2 ratio of zero and (ii) that data points falling outside of the centralized log2 ratio distribution are representative of a difference in CN between test and reference genomes.

(iii) *Genotyping console*

After processing CEL files and the Birdseed to call genotypes, we used the Genotyping Console (GTC v.3.0.2) to detect CNVs from the Affymetrix 6·0 array for samples that passed initial quality controls (QCs). The default parameters of >1 Kb size and >5 probes in this algorithm were used.

(iv) *Data analysis*

A genome-wide CNV study was carried out using SVS Golden Helix Version 7.2 (Bozeman, 2010) and Affymetrix Genotyping Console software as prescribed in their manuals (Affymetrix, Inc., 2005; Affymetrix, Inc., 2007; Affymetrix, Inc., 2008 *a*). The Eigenstrat method was used to avoid the possibility of spurious associations resulting from population stratification. Bonferroni correction was employed for multiple testing and the corrected data were then used for CNV testing. Bonferroni methods for population data genotyped on the Affymetrix 6·0 platform was $\alpha = 0{\cdot}05$ thresholds between 1×10^{-7} and 7×10^{-8}. We adopted higher stringency screening methods for CNVs to overcome possible unequal sample size effects.

Analysing the collated data from the arrays that have CNP calls with a LOD score greater than or equal to ten (corresponding to a false discovery rate of ~5%) was the criteria selected for the present investigation. CNVs that were >100 Kb in size and picked by >5 probes were used as the parameters for CNV calling in several algorithms. This approach ensured CNVs with only a greater degree of confidence are identified, although there is a concomitant loss of power to detect some novel CNVs. All SNPs that were called using the Birdseed v2 algorithm had a QC call rate of >97% across individuals. All the subjects and members with SNPs that passed SNP QC procedures were entered into the CNV analysis. Filters were set for ID call rates for the overall SNPs to identify IDs with poor quality DNA, if any. The CNV calls were generated using the Canary algorithm. In the Affymetrix Genotyping Console Software, contrast QC has to be >0·4 to be included in the CNV analyses. In this study, observed contrast QC was >2·5 across all samples showing a robust strength. To control for the possibility of spurious or artifact CNVs, we used the Eigenstrat approach of Price *et al.* (2006). This method derives the principal components of the correlations among gene variants and corrects for those correlations in the testing. We removed 55 individuals from the study group because they were extreme outliers on one or more significant Eigenstrat axes and dropped a further 543 CNVs from the members selected for the study for not meeting the required QC measures. CNVs were considered validated when there was a reciprocal overlap of 50% or greater with the reference set. Though the Jaccard statistic is sensitive to the number of CNVs called by each algorithm (ideally each of two algorithms would detect a similar number of CNV calls), the relative values between the different comparisons of algorithms/platform/site are very informative. All the overlap analyses performed have handled losses and gains separately except when otherwise stated, and were conducted hierarchically. The calls from the algorithms that were called in both were not considered; instead, they were collated so that the relative values between the different comparisons of algorithms/platform/site are still very informative.

(v) *Chromosome length assessment*

The length of both duplication and deletion CNVs for all chromosomes were calculated for all the

https://doi.org/10.1017/S0016672315000191 Published online by Cambridge University Press

populations and the difference obtained was correlated with the standard length of that particular chromosome.

(vi) *HD-CNV*

In order to compare and identify CNVs between samples of the same and different populations as hotspots and rare, and to also correlate their abeyant effects on a wide variety of biological contexts, Hotspot Detector for CNVs (HD-CNV) (Butler *et al.*, 2013) was used to analyse and detect recurrent CNV regions by finding cliques in an interval graph generated from the input. HD-CNV requires CNV calls as an input to detect recurrent regions based on percentage overlap. Hotspot Detector imports pre-formatted CSV files containing detected CNV events in the study. CNV events are treated as nodes in an interval graph and are used to represent regions (intervals) on a real line, and edges are added where intervals overlap as postulated by Lekkerkerker *et al.* (1962). Based on this, Butler *et al.* (2013) modified and added edges between nodes that share the base pair overlap required to consider two CNV events part of a merged region (default 40%) and the overlap required for a family (region with highly similar CNV events, default 99%). Merged regions, therefore, contain a collection of CNV events where each CNV overlaps all others in the merged region by the minimum overlap specified, indicating the genomic location where those groups of overlapping CNVs are found (Butler *et al.*, 2013). The output graph generated by the HD-CNV was then visualized using Gephi graph creation software.

(vii) *Breakpoint validations*

In order to validate the presence of the CNVs, PCR amplification was performed on four recurring CNV breakpoints on 500 randomly chosen individuals from India. Chimeric primers flanking normal and deleted/duplicated sequences were designed to bind only to the CNV breakpoint regions. Samples that do not contain these specific CNVs would fail to amplify. The primers used for amplifying the breakpoint regions are: AGGTCTGTTATGTGGCTGAGCCG CA on 3q29 for breakpoints 195276060–195446910 bp; ACTCTAGCCAACACATCCTCTGCGC on 15q14 for 34695310–34857998 bp; GAGTAAAGAAAC AAAGGCCATCT on 21q11.2 for 14594223–15101046 bp; and AGGGATCCACCCCCTGGCTG TGGA on 16p13.11 for 16377650–16635603 bp at annealing temperatures of 64, 73, 62 and 71 °C, respectively, using the DreamTaq polymerase in the Kyratec PCR System (Kyratec, Australia). All PCR products were analysed on 1% agarose gels and documented with a Vilber Lourmat Imaging system (Vilber Lourmat, France).

3. Results

Whole-genome CNV analysis in 1715 individuals from 12 different ethnic populations identified a total of 44 109 CNVs across the genome (Fig. 1 and Supplementary Table 1). This study represents the first draft of population-specific maps as well as a cross-population map, comprising both older and newer CNVs (Fig. 1 and Supplementary Table 1). Duplication CNVs (74·71%) were significantly higher than deletion CNVs (25·28%). All of the 1715 genomes analysed here showed CNVs and the majority were identified in the range of five to >50 CNVs for each individual. The populations showed a varying degree of CNVs ranging from higher (>40), to moderate (>25–40), to lower (>10–25). New World and Australian populations showed the highest CN events, while Taiwan, Tibet, India and Ashkenazi Jewish populations showed moderate CNV events, whereas JPT, CEU and YRI showed a comparatively much lower number of CNV events.

(i) *CNV burden on chromosomes*

The 100–250 kb and 250–500 kb sized duplications as well as deletion CNVs were abundantly scattered across all chromosomes with the same frequencies. However, 500 kb–1 Mb CNVs showed slightly fewer duplications and a scant distribution of deletions. The 1–2 Mb variations were rich in chromosomes 2, 5, 9, 14 and 15. The 2–3 Mb size load was only found in chromosomes 2, 9, 11, 15 and Y, whereas only chromosome 15 showed deletions. The 3–4 Mb, 4–5 Mb and >5 Mb variations were less and limited only to 9, 21, X and Y chromosomes (Supplementary Fig. 1).

CNVs were observed across all chromosomes and were more concentrated in chromosomes 14, 8, 2 and 15 (~8%), whereas chromosomes 13, 20 and 18 showed remarkably fewer CNV counts (~1%) (Fig. 2). CNV count load of all chromosomes for all populations was similar. Chromosomes 2, 7, 10, 12, 13, 14, 18, 21, 22 and Y showed higher distribution of duplication CNVs; however, equal distribution of CNVs were observed for 8, 6 and 19, whereas chromosomes 20 and Y showed an almost negligible count of deletion CNVs (Supplementary Fig. 2 (*a*) and (*b*)).

CNV size distribution in chromosomes was distinguishable with chromosomes 15, 14, 2, 9 and 8 bearing a heavy CNV size burden whereas, chromosomes 20, 13, 18, 21 and 6 bore the lowest CNV size (Supplementary Fig. 3 (*a*)). A total of nine chromosomes showed heavy CNV size burden on the p arm, while q arm burden was found in seven chromosomes (Supplementary Fig. 3 (*b*)). HapMap populations

https://doi.org/10.1017/S0016672315000191 Published online by Cambridge University Press

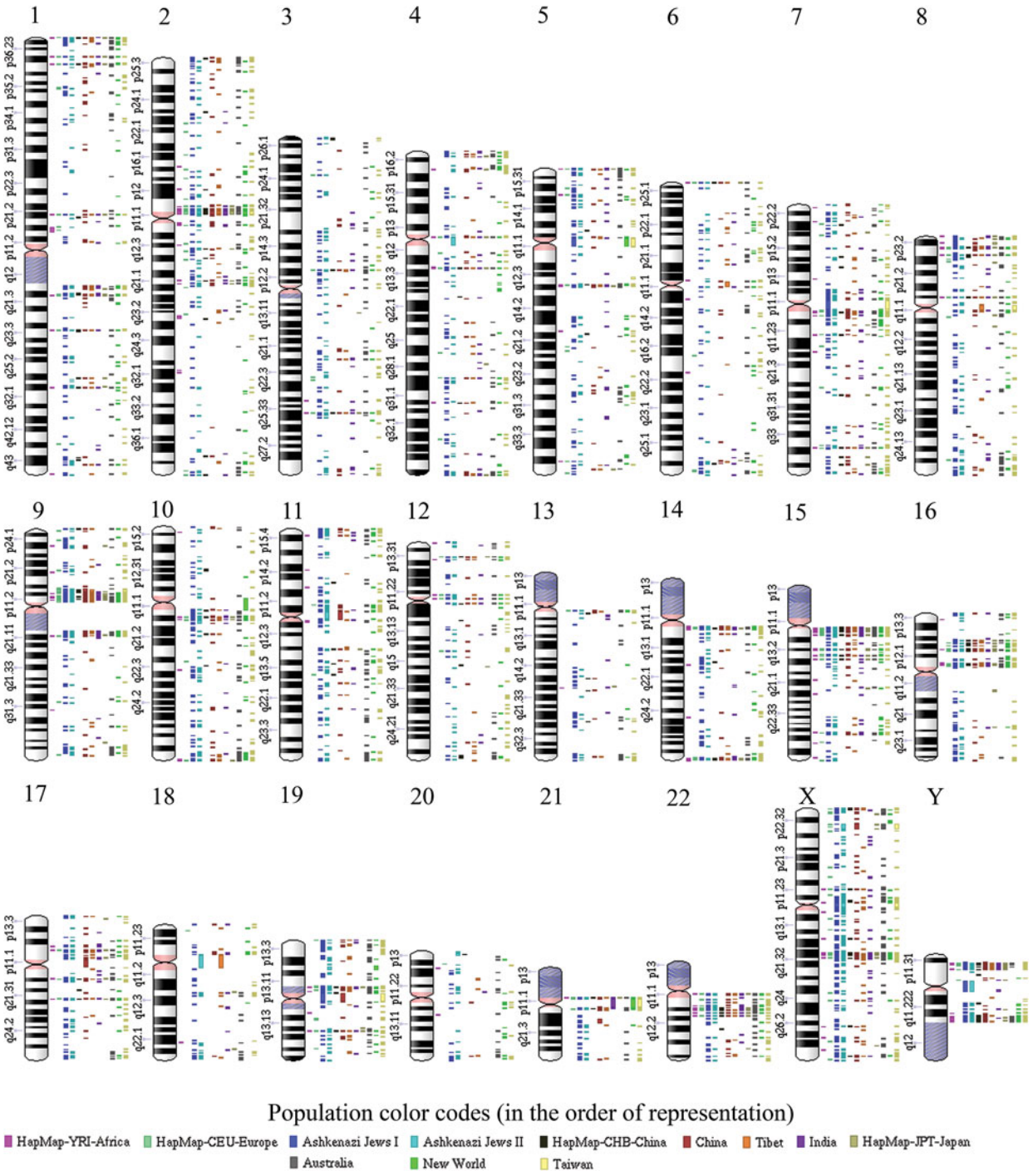

Fig. 1. CNV map of chromosomes of all 12 populations. Each vertical dedicated lane represents the genomic positions of CNVs of a particular population.

showed the highest CNV size (>21%) in chromosome 15, while JPT, India, CEU and China showed very few CNVs on chromosome 20 (Supplementary Fig. 2 and 3).

CNVs were found to be distributed across both p and q arms in nearly all the chromosomes except a few. The highest deletion CNVs were found in the p arm of chromosome 16 (~9·6%), followed by duplications in the q arm of chromosome 14 (9·2%), 22 (~8·6%) and X (~8·5%). Interestingly, the p arm of the Y chromosome showed only duplications, and the q arm of chromosome 14 was under severe CNV burden across all populations. However, population specific dominance across p and q arms of chromosomes were also observed (Supplementary Fig. 4 (*a*)). JPT peaked in the deletion CNVs of the 4q region, whereas Australia, showed higher concentration of duplication CNVs in 3p, 6q, 8q, 10q, 11q, 12q and

https://doi.org/10.1017/S0016672315000191 Published online by Cambridge University Press

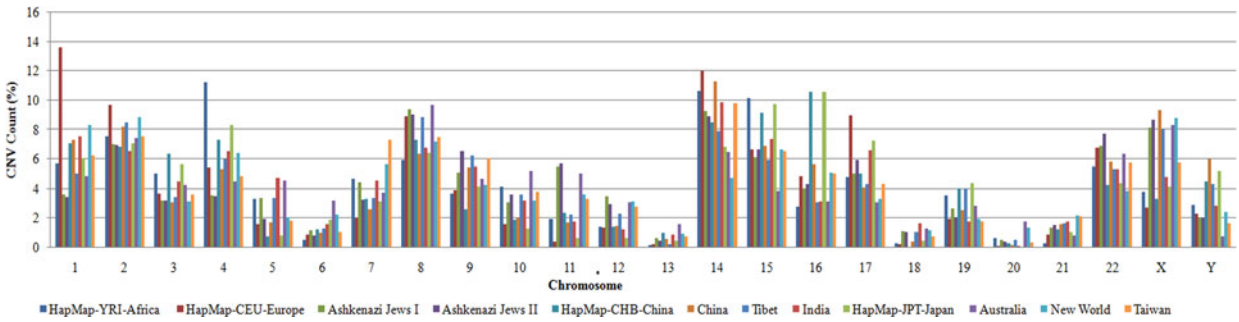

Fig. 2. Distribution of CNV counts across populations in all chromosomes, where each cluster represents the 12 populations in different colors. The 14th chromosome shows consistently high CNV counts in all populations. The clusters containing chromosomes 1–4, chromosomes 14–17 and chromosomes 22 and X show cluster specific count load, whereas the remaining chromosomes display gradient stages of CNV count.

13q chromosomes. India dominated CNV events in 3p and 5q duplications, Tibet in 9p duplications and Taiwan in 9p deletions. Whereas Ashkenazi Jews showed high specificity for the duplications in 11p and New World showed the highest in the duplications of 1q and 7p. However, high CNV burden was observed universally for chromosomes 14, 22 and X across all populations (Supplementary Fig. 4 (*b*)).

CNVs in telomeres showed a scattered distribution in both p and q arms across all populations. CHB, Tibet, India, JPT, Australia, New World and Taiwan showed a high number of duplication CNVs in telomeres compared to China, where more deletion CNVs were found. Tibet, India, Ashkenazi Jews and Australia showed both p and q arm telomere duplications, whereas varying patterns were seen in JPT, New World and Taiwan (Supplementary Fig. 5(*a*)).

CNVs in PARs were very few compared to telomere regions, with both duplications and deletions observed more in the p arm than the q arm. YRI showed no CNVs in the PAR vicinity, whereas CEU and China showed duplication CNVs in both p and q arms. On the contrary, Ashkenazi Jews and Taiwan showed both duplications and deletions in the p arms only. No one population showed both p and q arm duplications and deletions, as the CNVs were found to be recurring alternatively.

(ii) *CNVs impact on chromosome length*

Chromosome 1 showed shortening of the length in the Old World populations (−0·66 to −28·15μm) and increased length in the New World populations (+90·13 to +135·93 μm) (Fig. 3). Similarly, 2nd and 3rd chromosomes showed population specific lengthening and shortening with Tibet losing parts of both chromosomes, whereas New World populations were gaining in their length. Chromosomes 12, 20 and Y showed only gain of length across all populations (Supplementary Table 2). Chromosomes 8, 9, 14, 15, 22 and X showed increased chromosome length across all populations, but with few losses in Tibet and Europe. However, chromosomes 6, 10, 11, 13, 17 and 18 showed less impact of CNVs on the length, which varied among Asia and Australia but showed shortening in New World populations (Fig. 3).

(iii) *Status of CNV inheritance and mutation rate*

CNV rate analysis on trio families of YRI, CEU and India, revealed three clusters of CNVs based on inheritance. CNVs with precise breakpoints inherited from one generation to the next, of the exact same size and present in the same location were regarded as 'inherited' CNVs. CNVs that were not identified in the first generation but appeared in the second, were regarded as *de novo* CNVs. The third cluster contained CNVs that showed at least one breakpoint (either start or end point) being precisely inherited, while the other non-inherited, extended breakpoint was considered as *de novo* and was collectively referred to as 'inherited + *de novo*'. Of the 3103 CNV calls examined, 175 were assigned as inherited, 668 were *de novo* CNVs and the remaining 298 were the unusual 'inherited + *de novo*' CNV events for which a single parental origin could not be assigned. A total of 1160 precise CNV breakpoints that were identified in the probands were used to study the CNV inheritance and *de novo* mutation rate. YRI and CEU showed an almost similar rate (~18–20%) of the CNVs being inherited, while India showed a frequency of 9·8%; however, *de novo* mutation rate was ~50–56% for YRI and CEU. On the contrary, India showed 65%. However, the third cluster showed unequal rates with YRI being the highest (32%), followed by India (26%) and CEU (23·5%).

In Indian trios, mother to proband CNV transmission was observed at 7%, with higher duplications than deletion CNVs. On the contrary, father to proband transmission was lesser but with equal ratios of both duplication and deletion CNVs. *De novo* mutation rate showed higher duplication CNVs than deletion CNVs. No CNV bias was observed in the

https://doi.org/10.1017/S0016672315000191 Published online by Cambridge University Press

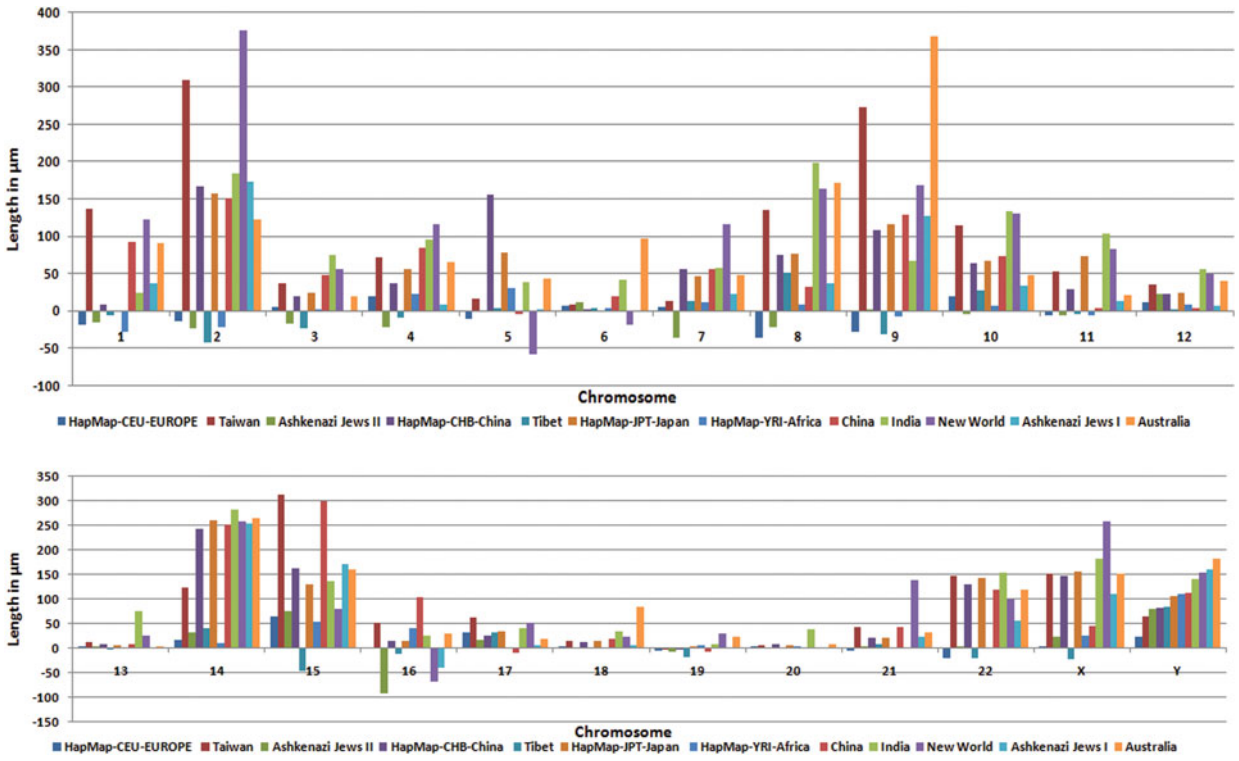

Fig. 3. Impact of CNVs on chromosome length. Shortening or lengthening of chromosomes is observed in different populations. Lengthening is represented in the positive quadrant and shortening in the negative quadrant.

paternal and maternal genome on the CNVs of the 'inherited + *de novo*' state. YRI showed increased transmission of duplication CNVs (6·15%) from father to proband. *De novo* origin as well as 'inherited + *de novo*' cluster showed higher duplications than deletion CNVs. CEU showed even distribution of CNV (4–6%) transmission across duplication and deletions from both paternal and maternal genomes. *De novo* mutation rate resulted in significant deletion CNVs compared to duplication CNVs.

(iv) *CNV hotspots*

All 44 109 CNV calls were used as an input to detect recurrent and unique CN regions based on percentage overlap. A total of 33 CNV hotspots (from 5382 CNVR hotspots), along with ~5435 intermediate events were identified across all chromosomes. Chromosomes 1, 3, 5, 6, 8, 11, 12, 14, 15, 18, 20–22 and Y are significantly enriched with a higher number of CN hotspots compared to the other chromosomes, which showed more unique events. Unique and intermediate events were distinctly found across all chromosomes of the populations (Fig. 4).

(v) *Sex bias in CNVs*

The male to female sex ratio across all populations is almost 1:1. CNV presence was observed to be biased in male and female genomes in several populations (Supplementary Fig. 5 (*b*)). China and JPT showed larger duplication size CNVs in females (~6–7 Mb) compared to males (~1·5 Mb), while Tibet and Australia showed larger duplication size CNVs in males (~8 Mb) compared to females (~1·8 Mb). The remaining populations showed balanced CNV size in duplications and deletions in both males and females. New World populations did not contain female samples, but showed the largest duplication CNV in males (~10 Mb) across populations, whereas Chinese males showed the least duplication CNVs (~1 Mb) (Supplementary Fig. 5 (*b*)).

(vi) *Breakpoint validations*

Successful amplification of four recurring CNV breakpoints was performed on 500 randomly chosen individuals from India validating the presence of the CNVs. Varying frequency and amplification status was observed for breakpoints. About 48–54% of individuals showed the presence of these CNV breakpoints at varying frequencies in the selected Indian samples.

4. Discussion

CNVs can be used as a tool to understand various aspects of genome organization such as distribution of CNVs in the miRNA and coding regions (Veerappa *et al.*, 2013 *a*, Veerappa *et al.*, 2013 *b*;

https://doi.org/10.1017/S0016672315000191 Published online by Cambridge University Press

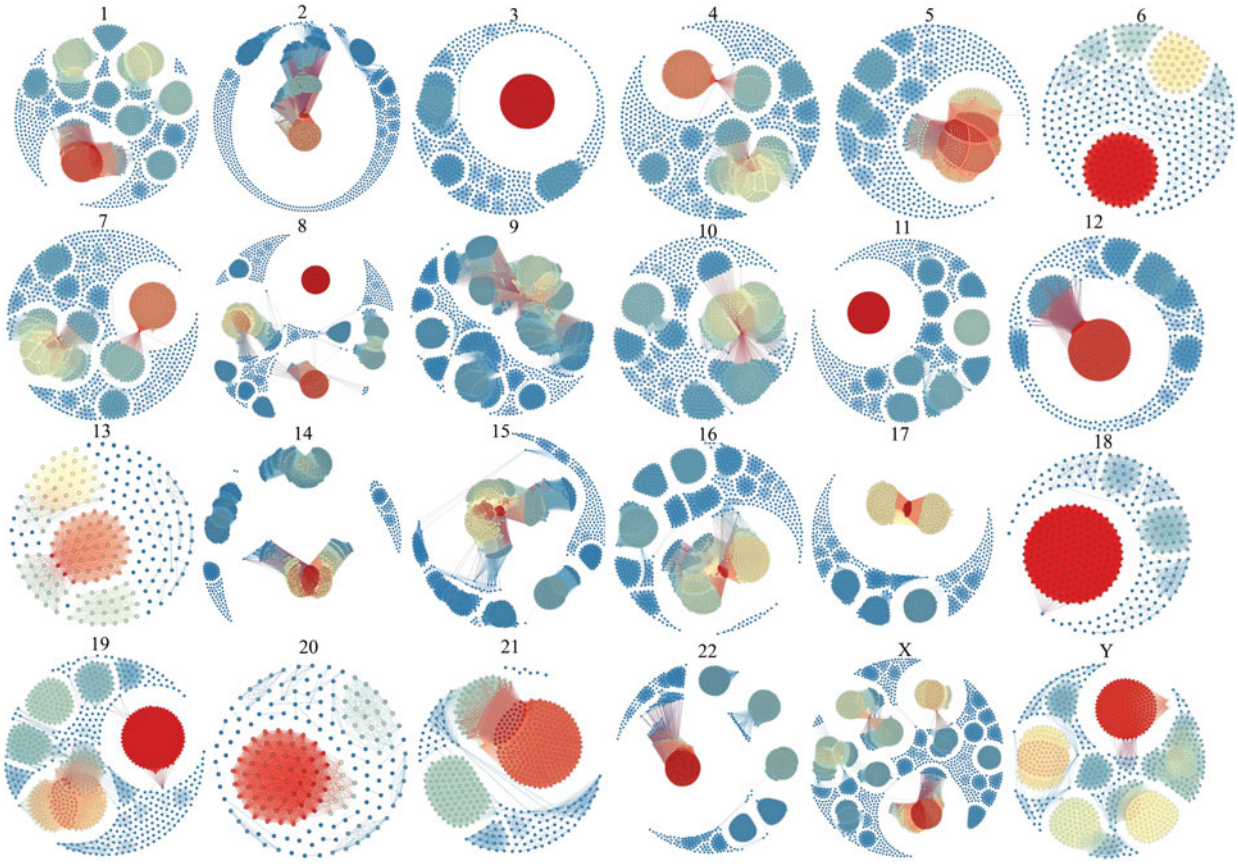

Fig. 4. Hotspot detection on CNVs was identified using HD-CNV software, which generated output files containing overlapping CNV regions, seen as clusters. Red indicates CNV hotspots; blue indicates rare CNV spots and other colors indicate intermediate CNV events. A total of 33 CNV hotspots (from 5382 CNVR hotspots), along with ~5435 intermediate events were identified across all chromosomes.

Veerappa *et al.*, 2013 *c*; Veerappa *et al.*, 2013 *d*; Veerappa *et al.*, 2014). The CNV map we generated provides a rationale for prioritizing chromosomal regions, population specific genome organization and uncovered elements of the genome that had largely been unexplored.

Since the 1–100 kb CNVs showed maximum signal to noise ratio, we restricted our investigations to a CNV size limit of only >100 kb. The majority of the discovered CNVs were found within the 100–500 kb size and the count of the remaining CNVs gradually declined with increase in size. We observed only 13 CNV events near the 4–5 Mb range, which were found localized near the heterochromatic gene poor regions. Furthermore, we did not detect any CNVs >5·5 Mb. This size limitation is probably to avoid larger sized CNVs attaining fixation near the gene rich regions, which would drastically influence gene expression and regulation.

(i) *CNVs on chromosomes*

CNVs were well distributed across all chromosomes and did not show any sort of bias on the lengths of the chromosomes. This CNV load remained consistent across all populations, but a few chromosomes showed higher CNV load when compared to others. The reason for this distribution pattern is probably due to the genome sequence topology and architecture. Although the Y chromosome is the smallest in size, the CNV load was on a par with that of chromosomes 10, 11 and 12, indicating that the CNV occurrence is not influenced by chromosome length. China and Taiwan showed almost similar CNV count distribution across all chromosomes, except in chromosomes 7 and Y, where either of them showed higher load, but never both, and China showed the highest CNV load in the sex chromosomes. Domination of CNV count of the populations on chromosomes varied, with Australia leading in seven chromosomes. It was the CNV count that largely contributed to the represented bulk CNV size of the chromosome; however, sometimes it was the nature of large CNV size segments that contributed to the bulk of size on that chromosome. For instance, chromosome 14 showed a high number of CNV counts compared to chromosome 15, but with respect to CNV size, chromosome 15 showed large size

https://doi.org/10.1017/S0016672315000191 Published online by Cambridge University Press

segments compared to chromosome 14. The same pattern was observed between many chromosomes, for example, between 8 and 9, and also 9 and 2, and was not limited to only these examples. Some chromosomes showed CNV counts directly proportional to CNV size load and the former also varied when the latter deviated.

The CNV distribution pattern in p and q chromosomal arms was analysed in order to provide an arm-level quantization and correlation. CNVs in the p arm of chromosome 16 and q arm of 14 was high. This finding disturbs the assumption that CNV counts in short arms are less common compared to long arms. Almost equal ratios of CNV counts were observed in both arms, suggesting that the CNV occurrence and distribution is length independent. Populations showed specific CNV distribution pattern in arms: Australia showed specific CNV gains mostly in q arms, whereas JPT, CHB, India and Ashkenazi Jews showed this in both p and q arms, while Tibet and Taiwan showed contradicting CNV types in the p arm of the 9th chromosome. Although these findings are new, correlations between CNVs at the arm level have been previously been established for some clinical features (Broad Institute TCGA Genome Data Analysis Centre, 2013). These arm specific CNVs in populations provide evidence for genetic diversity and reveal new insights into involvement in local adaptation.

We found CNVs close to the telomere and centromere regions, but the calls in the latter were filtered out and were not included in the present study. CNVs in the telomere region were found in both p and q arms and did not show any definite pattern of distribution in the populations. We also did not find any single factor that might explain this pattern of CNV distribution; however, CNVs were significantly overrepresented in number within 1–2 Mb of telomeres. Earlier studies have reported contradictory findings regarding representation of CNVs in telomere regions; however, the present study indicates an inclination towards the higher representation (Sharp *et al.*, 2005; Nguyen *et al.*, 2006).

The recombining regions of the tips of the X and Y chromosomes are PAR1 and PAR2, whereas a new pseudoautosomal region named PAR3, located in the Xq21.3 and Yp11.2, exhibits similar recombination frequency as PAR2 (Veerappa *et al.*, 2013 *a*). Few populations showed CNV bias on the occurrence of CNVs in the PARs, with Ashkenazi Jews and Taiwan showing CNVs only in p arms, whereas CEU and China showed only duplication CNVs in both p and q arms. PAR variations have been associated with mental and stature disorders while other studies have reported duplications and deletions in both PAR1 and PAR2 in infertile men (Gabriel-Robez *et al.*, 1990; Mohandas *et al.*, 1992; Jorgez *et al.*, 2011) suggesting that the presence of such variations in *AZF* and *SHOX* genes of the PARs are more complex and pathogenic. However, we did not observe any CNVs to be present under or near the fertility genes in PARs, indicating the neutral effect of CNVs on the phenotype. CNVs on PAR3 were frequent at the SD region of Xq21.3 near *PCDH11X* and *TGIF2LX* in the X chromosome, and *PCDH11Y* and *TGIF2LY* in the Y chromosome. Though these CNVs sometimes disrupted the coding structure entirely or partially, only the variations in *PCDH11X* have been associated with developmental dyslexia, whereas the impact of the other variations in the remaining genes on phenotype change is yet to be associated (Veerappa *et al.*, 2013 *d*). An assessment on the role of CNVs in normal individuals is necessary to delineate pathogenic PAR CNVs from neutral CNVs.

Chromosomes 15, 14, 2, 9 and 8 were seen to be selectively under the burden of large CNVs across all populations. Though previous studies (Girirajan *et al.*, 2011) have implicated large CNVs in various neurodevelopmental phenotypes, CNVs in these chromosomes seem to enrich or deplete the gene copies and it remains to be seen whether the size loads contributes to possible functional change. Several thousand disease phenotypes have been associated with the CNVs on these chromosomes, and the distribution of larger CNVs in the general population has remained largely unexplored (Girirajan *et al.*, 2011). We find that a significant fraction of CNVs are common and that both CNV size and gene density strongly correlate. Such substantial diversity in size exhibits a diverse array of phenotypes imparted functionally by way of majorly reorganizing the landscape of functional elements in the regions affected.

Chromosome arm lengths are critical elements of the human genome, which previous studies have failed to assess. The relative size of present day chromosome lengths influenced by the recurring occurrence of CNVs over time has been measured here. Variations >100 kb, which are more prevalent in the human genome, have gradually over time altered the length of the chromosomes due to CNVs attaining fixation. Old World populations showed shortening for few chromosomes while New World populations were gaining in length. Population specific lengthening and shortening was also observed for few chromosomes and, contrarily, New World populations were only gaining in their length. Chromosomal arm lengths ensure that homologous chromosomes cross over during meiosis and are suggested to be important for size-dependent control in meiotic reciprocal recombination in the yeast *Saccharomyces cerevisiae* (Kaback *et al.*, 1999). Altered chromosomal lengths are directly linked to recombination rates and we believe the observed present day altered chromosomal lengths are probably the reason behind varying recombination rates for chromosomes across populations.

https://doi.org/10.1017/S0016672315000191 Published online by Cambridge University Press

(ii) *Sex bias*

Gender plays a pivotal role in human genetic identity and is also manifested in many genetic disorders. CNV presence showed sex bias in several populations, largely dependent on ethnicity and was observed in China, Tibet, Japan and Australia. There was a significant difference only in the duplication size of the CNVs in both the male and female genome, but the deletion size of the CNVs across both sex and population were stable. These findings imply a more recent evolutionary role for gender. Many human genetic phenotypes, including those related to olfaction, developmental delay and intellectual disabilities, have shown similar sex bias and along with recombination rates are known to be different in both males and females (Girirajan *et al.*, 2011; Shadravan, 2013).

(iii) *CNV inheritance status*

Based on the inheritance status, YRI and CEU showed a fold higher inheritance rate compared to India, whereas rate of *de novo* CNV occurrences were strikingly similar. However, the 'inherited + *de novo*' cluster showed unequal rates with YRI being the highest, followed by CEU and India. The reasoning and the mechanism of the origin of 'inherited + *de novo*' CNVs is yet to be clearly understood. A probable mechanism could be that an environmental exposure is found to influence a CNV, found to be a part of an initial alteration in the germline epigenome and is found to promote genetic changes such as induced CNVs in later stages as described by Guerrero-Bosagna (2010) and Livide *et al.* (2012). Previous studies have indicated hypermethylation status and CN changes on several oncosuppressor genes as causing retinoblastoma (Livide *et al.*, 2012); however, the present findings highlight the probable role of epigenetic changes on the genes encoding for recombinatory machinery in creating the 'inherited + *de novo*' cluster. This study indicates the probable association of epigenetic changes with CNV inheritance and mutation rate. All the CNVs clustered based on the point and type of origin showed a diverse array of contributions from both parental origins. This study also confirms the diverse array and biased contributions of CNVs from mother compared to father, but no such bias was observed for the CNVs of the 'inherited + *de novo*' state. Earlier studies reported diverse CNV transmission and *de novo* event rates in probands as shown in several neurodevelopmental phenotype and monozygotic twin studies (Sebat *et al.*, 2007; Marshall *et al.*, 2008; Levy *et al.*, 2011; Ehli *et al.*, 2012). These varying rates of transmission and mutation are believed to be related to the types of CNV discovery tools and ascertainment biases employed in the studies. CNV frequency bias was observed on the CNV transmissions from both maternal and paternal genomes showing major contributions from the duplications rather than the deletion CNVs. A higher number of maternal transmissions were observed for the probands of the Indian population, and on the contrary, YRI showed higher paternal transmissions.

(iii) *Hotspot detection*

Investigation on CNV hotspot rates is also critical for understanding the CNV mutation rate and genomic instability in the aetiology of the CNV-related traits. Hotspot analysis was performed using HD-CNV to identify the hotspots and unique CNVs between samples of the same and different populations, and also to correlate their abeyant effects in a wide variety of biological contexts. The HD-CNV program identified the hotspot, intermediate and rare CNVs between samples of the same and different populations, and showed 1–4 hotspots bearing extensive hotspot CNV clusters. Unique CNV events were considerably higher compared to hotspots and intermediate events, and almost all events were found to overlap with at least one other event. These events are very commonly recurring events that have similar boundaries and not much distinction can be observed in the groups. Chromosomes 8, 14, 22 and Y show very highly conserved patterns, as there are large groupings of overlapping CNV calls, and very few individual calls. The number of events on chromosome 6 identified in our study is relatively low, contrary to the previous studies on HapMap data (Butler *et al.*, 2013). The locus-specific mutation rates for CNV have been observed to be ~100 to 10 000 times higher than those for nucleotide substitution rates, which not only highlights the instability of CNV regions but also suggests large variation in CNV mutation rate (Fu *et al.*, 2010). Data from the 1000 Genomes Project suggests that the distribution of CNVs in the genome is biased towards multiple hotspots, including segmental duplications, and away from genes encoding protein complexes and other dosage sensitive genes (1000 Genome Project Consortium, 2010). This hotspot analysis elucidates the fragility of the genome, which contains recurrent CNV regions bearing a large concentration of repeats in the flanking sequences. Most of the intermediate subgroups are seen to be shared with both hotspot and rare clusters, with equal chances of being converted into either cluster. These subgroups are in a dynamic state and tend to shift either towards hotspots or rare clusters over a continuous evolutionary CNV burden. However, this state seems to be inert in nature currently, with little or no scope for state conversion, but may tend to change over a period of evolutionary CNV burden indicating that the intermediate subgroup is under selective pressure.

https://doi.org/10.1017/S0016672315000191 Published online by Cambridge University Press

This comprehensive whole-genome study identified that larger CNVs were fewer and limited only to certain chromosomes. CNV count and sometimes CNV size contributed to the bulk CNV size of the chromosome. Population specific CNV distribution pattern in p and q arms disturbed the assumption that CNV counts in the p arm are less common compared to long arms, and the CNV occurrence and distribution in chromosomes is length independent. Population specific lengthening and shortening of chromosomal length was observed. Asians showed adequate loss of genome compared to YRI. Sex bias for CNV presence was largely dependent on ethnicity. Lower CNV inheritance rate was observed for India, compared to YRI and CEU. Around 19 905 ancient CNVs were identified across all chromosomes, and populations at varying frequencies clearly indicate the complex organization of the human genome.

We thank the funding agency, Department of Science and Technology-Health Science (SR/SO/HS-103/2007), Government of India, New Delhi; Yenepoya University Seed grant (YU/Seed Grant/2011-011), Mangalore; the subjects and their families for participating in this study; The Chairman, DOS in Zoology, University of Mysore; Prof. H. A. Ranganath and Prof. K. S. Rangappa, University of Mysore, for their help and encouragement; and also the University of Mysore for providing the facilities to conduct this work; CSIR, New Delhi, for awarding Fellowship to K.L. (CSIR Order No.9/119(0196) 2K13-EMR-I Dated: 19.03.2013); DST-INSPIRE, New Delhi, for awarding Fellowship to M.M. (IF120351, DST sanction letter No. DST/INSPIRE FELLOWSHIP/2012/327); E. M. Locke of Western University, London, for inputs on HD-CNV; and P. N. Radhika and A. N. Somanna, University of Mysore, for their support.

Declaration of interest

None.

Supplementary material

To view Supplementary material for this article, Please visit http://dx.doi.org/10.1017/S0016672315000191

References

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. & McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073.

Affymetrix, Inc. (2009). Data Sheet: Genome Wide Human SNP Array 6·0. Available at http://media.affymetrix.com/support/technical/datasheets/genomewide_snp6_datasheet.pdf (accessed 30 August 2015).

Affymetrix, Inc. (2005). Technical Note: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation. Available at http://media.affymetrix.com/support/technical/technotes/plier_technote.pdf (accessed on 30 August 2015).

Affymetrix, Inc. (2008 *a*). User Manual: Genotyping Console™ Software 2.1. Available at http://array.mc.vanderbilt.edu/microarray/dna/GTC_Manual.pdf (accessed 30 August 2015).

Affymetrix, Inc. (2008 *b*). White Paper: Affymetrix® Canary Algorithm Version 1.0., 1–7. Available at http://media.affymetrix.com/support/technical/whitepapers/canary_algorithm_whitepaper.pdf (accessed 30 August 2015).

Affymetrix, Inc. (2007). White Paper: BRLMM-P: A Genotype Calling Method for the SNP Array 5.0. Available at http://media.affymetrix.com/support/technical/whitepapers/brlmmp_whitepaper.pdf (accessed 30 August 2015).

Birdsuite Algorithm (2010). Available at http://www.broad.mit.edu/mpg/birdsuite/birdseed.html (accessed 13 February 2013).

Bozeman, M. T. (2010). Golden Helix, Inc. SNP & Variation Suite (Version 7.x) (Software). Available at http://www.goldenhelix.com (accessed 13 January 2013).

Broad Institute TCGA Genome Data Analysis Center (2013). Prostate Adenocarcinoma (Primary solid tumor cohort) – 21 April 2013: Correlation between CN variations of arm-level result and selected clinical features. *Broad Institute of MIT and Harvard* doi:10.7908/C1JD4TRG. Available at http://gdac.broadinstitute.org/runs/analyses__2015_04_02/reports/cancer/PRAD-TP/index.html (accessed 30 August 2015).

Butler, J. L., Osborne Locke, M. E., Hill, K. A. & Daley, M. (2013). HD-CNV: hotspot detector for copy number variants. *Bioinformatics* **29**, 262–263.

Chen, W., Hayward, C., Wright, A. F., Hicks, A. A., Vitart, V., Knott, S. & Porteous, D. J. (2011). Copy number variation across European populations. *PLoS One* **6**, e23087.

de Vries, B. B., Pfundt, R., Leisink, M., Koolen, D. A., Vissers, L. E., Janssen, I. M. & Veltman, J. A. (2005). Diagnostic genome profiling in mental retardation. *The American Journal of Human Genetics* **77**, 606–616.

Ehli, E. A., Abdellaoui, A., Hu, Y., Hottenga, J. J., Kattenberg, M., van Beijsterveldt, T. & Davies, G. E. (2012). *De novo* and inherited CNVs in MZ twin pairs selected for discordance and concordance on attention problems. *European Journal of Human Genetics* **20**, 1037–1043.

Fu, W., Zhang, F., Wang, Y., Gu, X. & Jin, L. (2010). Identification of copy number variation hotspots in human populations. *The American Journal of Human Genetics* **87**, 494–504.

Gabriel-Robez, O., Rumpler, Y., Ratomponirina, C., Petit, C., Levilliers, J., Croquette, M. F. & Couturier, J. (1990). Deletion of the pseudoautosomal region and lack of sex-chromosome pairing at pachytene in two infertile men carrying an X;Y translocation. *Cytogenetic and Genome Research* **54**, 38–42.

Gautam, P., Jha, P., Kumar, D., Tyagi, S., Varma, B., Dash, D. & Indian Genome Variation Consortium (2012). Spectrum of large copy number variations in 26 diverse Indian populations: potential involvement in phenotypic diversity. *Human Genetics* **131**, 131–143.

Girirajan, S., Brkanac, Z., Coe, B. P., Baker, C., Vives, L., Vu, T. H. & Eichler, E. E. (2011). Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genetics* **7**, e1002334.

Go, Y. & Niimura, Y. (2008). Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. *Molecular Biology and Evolution* **25**, 1897–1907.

Guerrero-Bosagna, C., Settles, M., Lucker, B. & Skinner, M. K. (2010). Epigenetic transgenerational actions of vinclozolin on promoter regions of the sperm epigenome. *PLoS One* **5**, e13100.

Jorgez, C. J., Weedin, J. W., Sahin, A., Tannour-Louet, M., Han, S., Bournat, J. C., Mielnik, A., Cheung, S. W., Nangia, A. K., Schlegel, P. N., Lipshultz, L. I. & Lamb, D. J. (2011). Aberrations in pseudoautosomal regions (PARs) found in infertile men with Y-chromosome microdeletions. *Journal of Clinical Endocrinology & Metabolism* **96**, E674–E679.

Kaback, D. B., Barber, D., Mahon, J., Lamb, J. & You, J. (1999). Chromosome size-dependent control of meiotic reciprocal recombination in *Saccharomyces cerevisiae*: the role of crossover interference. *Genetics* **152**, 1475–1486.

Kanduri, C., Ukkola-Vuoti, L., Oikkonen, J., Buck, G., Blancher, C., Raijas, P. & Järvelä, I. (2013). The genome-wide landscape of copy number variations in the MUSGEN study provides evidence for a founder effect in the isolated Finnish population. *European Journal of Human Genetics* **21**, 1411–1416.

Kang, T. W., Jeon, Y. J., Jang, E., Kim, H. J., Kim, J. H., Park, J. L. & Kim, S. Y. (2008). Copy number variations (CNVs) identified in Korean individuals. *BMC Genomics* **9**, 492.

Kim, H. L., Iwase, M., Igawa, T., Nishioka, T., Kaneko, S., Katsura, Y. & Satta, Y. (2012). Genomic structure and evolution of multigene families: "flowers" on the human genome. *International Journal of Evolutionary Biology* **2012**, 917678.

Lekkerkerker, C. G. & Boland, J. C. (1962). Representation of a finite graph by a set of intervals on the real line. *Fundamental Mathematics* **51**, 45–64.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y. H., Leotta, A., Kendall, J. & Wigler, M. (2011). Rare *de novo* and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886–897.

Lin, C. H., Li, L. H., Ho, S. F., Chuang, T. P., Wu, J. Y., Chen, Y. T. & Fann, C. S. (2008). A large-scale survey of genetic copy number variations among Han Chinese residing in Taiwan. *BMC Genetics* **9**, 92.

Liu, X., Cheng, R., Ye, X., Verbitsky, M., Kisselev, S., Mejia-Santana, H., Louis, E., Cote, L., Andrews, H., Waters, C., Ford, B., Fahn, S., Marder, K., Lee, J. & Clark, L. (2013). Increased rate of sporadic and recurrent rare genic copy number variants in Parkinson's disease among Ashkenazi Jews. *Molecular Genetics & Genomic Medicine* **1**, 142–154.

Livide, G., Epistolato, M. C., Amenduni, M., Disciglio, V., Marozza, A., Mencarelli, M. A. & Ariani, F. (2012). Epigenetic and copy number variation analysis in retinoblastoma by MS-MLPA. *Pathology & Oncology Research* **18**, 703–712.

Lou, H., Li, S., Yang, Y., Kang, L., Zhang, X., Jin, W. & Xu, S. (2011). A map of copy number variations in Chinese populations. *PLoS One* **6**, e27341.

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J. & Scherer, S. W. (2008). Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics* **82**, 477–488.

McElroy, J. P., Nelson, M. R., Caillier, S. J. & Oksenberg, J. R. (2009). Copy number variation in African Americans. *BMC Genetics* **10**, 15.

Mohandas, T. K., Speed, R. M., Passage, M. B., Yen, P. H., Chandley, A. C. & Shapiro, L. J. (1992). Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: meiotic studies in a man with a deletion of distal Xp. *The American Journal of Human Genetics* **51**, 526–533.

Nguyen, D.-Q., Webber, C. & Ponting, C. P. (2006). Bias of selection on human copy-number variants. *PLoS Genetics* **2**, e20.

Niimura, Y. & Nei, M. (2003). Evolution of olfactory receptor genes in the human genome. *Proceedings of the National Academy of Sciences USA* **100**, 12235–12240.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T. & Wigler, M. (2007). Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449.

Shadravan, F. (2013). Sex bias in copy number variation of olfactory receptor gene family depends on ethnicity. *Frontiers in Genetics* **4**, 32.

Sharp, A. J., Cheng, Z. & Eichler, E. E. (2006). Structural variation of the human genome. *Annual Review of Genomics and Human Genetics* **7**, 407–442.

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U. & Eichler, E. E. (2005). Segmental duplications and copy-number variation in the human genome. *The American Journal of Human Genetics* **77**, 78–88.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J. & Ge, R. (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**, 72–75.

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A. & Eichler, E. E. (2010). Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646.

The International HapMap Consortium (2003). The International HapMap Project. *Nature* **426**, 789–796.

Veerappa, A. M., Murthy, M. N., Vishweswaraiah, S., Lingaiah, K., Suresh, R. V., Nachappa, S. A. & Ramachandra, N. B. (2014). Copy number variations burden on miRNA genes reveals layers of complexities involved in the regulation of pathways and phenotypic expression. *PLoS One* **9**, e90391.

Veerappa, A. M., Padakannaya, P. & Ramachandra, N. B. (2013 *a*). Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. *Functional and Integrative Genomics* **13**, 285–293.

Veerappa, A. M., Saldanha, M., Padakannaya, P. & Ramachandra, N. B. (2013 *b*). Family-based genome-wide copy number scan identifies five new genes of dyslexia involved in dendritic spinal plasticity. *Journal of Human Genetics* **58**, 539–547.

Veerappa, A. M., Vishweswaraiah, S., Lingaiah, K., Murthy, M., Manjegowda, D. S., Nayaka, R. & Ramachandra, N. B. (2013 *c*). Unravelling the complexity of human olfactory receptor repertoire by copy number analysis across population using high resolution arrays. *PLoS One* **8**, e66843.

Veerappa, A. M., Saldanha, M., Padakannaya, P. & Ramachandra, N. B. (2013 *d*). Genome-wide copy number scan identifies disruption of PCDH11X in developmental dyslexia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **162**, 889–897.

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M. & Sebat, J. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543.

Xu, B., Roos, J. L., Levy, S., Van Rensburg, E. J., Gogos, J. A. & Karayiorgou, M. (2008). Strong association of *de novo* copy number mutations with sporadic schizophrenia. *Nature Genetics* **40**, 880–885.

Zhang, Y. B., Li, X., Zhang, F., Wang, D. M. & Yu, J. (2012). A preliminary study of copy number variation in Tibetans. *PLoS One* **7**, e41768.