Impact of copy number variations burden on coding genome in humans using integrated high resolution arrays

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

[1] *Genetics and Genomics Lab, Department of Studies in Zoology, University of Mysore, Manasagangotri, Mysore-06, Karnataka, India*
[2] *NUCSER, K. S. Hegde Medical Academy, Nitte University, Deralakatte, Mangalore-575 018, Karnataka, India*

(*Received 29 September 2014; revised 12 November 2014; accepted 13 November 2014*)

Summary

Copy number variations (CNVs) alter the transcriptional and translational levels of genes by disrupting the coding structure and this burden of CNVs seems to be a significant contributor to phenotypic variations. Therefore it was necessary to assess the complexities of CNV burden on the coding genome. A total of 1715 individuals from 12 populations were used for CNV analysis in the present investigation. Analysis was performed using Affymetrix Genome-Wide Human SNP Array 6·0 chip and CytoScan High-Density arrays. CNVs were more frequently observed in the coding region than in the non-coding region. CNVs were observed vastly more frequently in the coding region than the non-coding region. CNVs were found to be enriched in the regions containing functional genes (83–96%) compared with the regions containing pseudogenes (4–17%). CNVs across the genome of an individual showed multiple hits across many genes, whose proteins interact physically and function under the same pathway. We identified varying numbers of proteins and degrees of interactions within protein complexes of single individual genomes. This study represents the first draft of a population-specific CNV genes map as well as a cross-populational map. The complex relationship of CNVs on genes and their physically interacting partners unravels many complexities involved in phenotype expression. This study identifies four mechanisms contributing to the complexities caused by the presence of multiple CNVs across many genes in the coding part of the genome.

Introduction

Copy number variations (CNVs) are a form of structural variation in the genome, ranging from 1 kb to several megabases in size, which disturb the normal biological balance of the bi-allelic segmental state (Cheung *et al.*, 2003). These CNVs include duplications, deletions and insertions, and are found in all humans. Widespread presence of CNVs in normal individuals seems to be a significant contributor to phenotypic variation (Beckmann *et al.*, 2007) and affects more nucleotides per genome than SNPs. Analysis of individual as well as population genomes has revealed unexpected numbers and frequencies of CNVs in human populations indicating the existence of diversity in CNV distribution among different populations (Armengol *et al.*, 2009).

Copy numbers of a segment of recurrent recombining coding or non-coding DNA hotspots attain fixation in the genomes of a normal population when bred over a long period of time and it is during these times that various molecular mechanisms, including gene dosage, gene disruption, gene fusion and position effects, create Mendelian or sporadic traits, or become associated with complex diseases (Zhang *et al.*, 2009). CNVs, especially through gene duplication and exon shuffling, drive gene and genome evolution, which can be seen through the recurrent frequenting of CNVs in regions of the genome that bear several multigene families (Niimura *et al.*, 2003; Go & Niimura, 2008; Sudmant *et al.*, 2010; Kim *et al.*, 2012; Veerappa *et al.*, 2013 *a*). However,

* Corresponding author: Nallur B. Ramachandra, Genetics and Genomics Lab, Department of Studies in Zoology, 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.

CNVs, which are rare and malignant variants usually, are more recent and may have been *de novo* in origin, or they may have been passed down for only a few generations within a family (Girirajan & Eichler, 2010). These variants have been observed in patients with mental retardation, developmental delay, dyslexia, schizophrenia and autism. This growing body of evidence across the genome suggests that CNVs are a significant cause in contributing towards varying human phenotypes and disorders (Redon *et al.*, 2006; Sharp *et al.*, 2006; Sebat *et al.*, 2007; Walsh *et al.*, 2008; Veerappa *et al.*, 2013 *b*). Both low- and high-resolution CNV studies have been performed across control population cohorts since 2003, majorly covering Africa, America, Europe, China, Tibet, Taiwan, India, Germany and Finland (The International HapMap Consortium, 2003; Lin *et al.*, 2008; McElroy *et al.*, 2009; Simonson *et al.*, 2010; Chen *et al.*, 2011; Lou *et al.*, 2011; Gautam *et al.*, 2012; Zhang *et al.*, 2012; Kanduri *et al.*, 2013; Liu *et al.*, 2013), but many of these studies have failed to explain the genotypic complexities caused by the CNVs on chromosomes as well as in the coding regions.

The burden of CNV on chromosomes was also analysed, which will be published elsewhere (A. M. Veerappa *et al.*, unpublished observations). In brief, collated CNV data from two different arrays and from those meeting the copy number polymorphism (CNP) calls with a log10 of odds score ⩾10 criteria were selected for the investigation. CNVs that were chosen by >5 probes and those >100 kb in size were used for CNV calling in multiple algorithms ensuring CNVs with only a greater degree of confidence are identified. CNPs are those that are common and occur in the general population with an overall frequency of greater than 1%, whereas CNVs are less common and appear to be relatively rare variants when compared to CNPs. The whole-genome CNV study identified a total of 44 109 CNVs from 1715 genomes across 12 populations with a mean size of 7000 ± 3000 kb CNV size. Duplication CNVs were significantly higher than deletion CNVs and a significant portion was within 500 kb, also the CNV count considerably declined as the size of the variation increased. We established the first drafts of population-specific CNV maps providing a rationale for prioritizing chromosomal regions. New World populations showed the highest number of CNVs across all populations. Larger CNVs were fewer and limited only to chromosomes 9, 21, X and Y. CNV count, and sometimes CNV size, contributed to the bulk CNV size of the chromosome. Population-specific CNV distribution pattern in p and q chromosomal arms as well as lengthening and shortening of chromosomes was observed. Almost equal ratios of CNV counts were observed across all lengths of chromosomes suggesting that CNV occurrence and distribution is length independent. Asian populations showed adequate loss of genome compared to others, and sex bias was observed for the CNV presence in several populations. Lower CNV inheritance rate was observed for India, compared to Yoruba in Ibadan, Nigeria (YRI) and Utah residents with ancestry from northern and western Europe (CEU) and a total of 33 candidate CNV hotspots were identified across many chromosomes. Around 19 905 ancient CNVs were identified across all chromosomes and populations, at varying frequencies.

In this study, we have made an attempt to identify the CNV burden on genes in the coding regions across all populations and to establish a stress centralized protein interaction network and its comparative biological and molecular pathways using the CNV genes to elucidate the possible regions of stress/disconnection in the pathways. We also provide several layers of complexities contributed by CNVs on the coding genome, which is caused by the presence of multiple CNVs across many genes in a genome.

Materials and methods

For this study, a total of 1715 individuals involving 43 normal members from 12 randomly selected families residing in Karnataka, India, with different age group members ranging from 13–73 years old, 270 HapMap samples (The International HapMap Consortium, 2003) covering CEU (CEPH collection), Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT) and YRI populations, 31 Tibetan samples, 155 Chinese samples, 472 of Ashkenazi Jews (AJ) replicate I, 480 of AJ replicate II, 204 individuals from Taiwan, 55 from Australia and 64 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 a 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 the participants in accordance with IHEC requirements. The raw, unprocessed data from the Affymetrix Genome Wide SNP 6·0 array for the 31 individuals from the Tibet population, submitted by Simonson *et al.* (2010), and the remaining populations except India were obtained from the ArrayExpress Archive of the European Bioinformatics Institute. The genotyping was performed on all subjects across all study populations using multiple algorithms. Details of genotyping, data analysis, breakpoint validations and other programs used in this study have been described in detail in the supplementary material (available online).

https://doi.org/10.1017/S0016672314000202 Published online by Cambridge University Press

Table 1. *Distribution of annotated and singleton genes across 12 populations*

Population	Individuals assessed	Total annotated genes			Singleton genes		
		No.	Mean	%	No.	Mean	%
HapMap-YRI-Africa	90	3992	44·35	3·16	525	5·83	3·45
HapMap-CEU-Europe	90	3841	85·35	3·04	522	5·80	3·43
Ashkenazi Jews I	464	34807	75·01	27·58	2678	5·77	17·63
Ashkenazi Jews II	480	36167	75·34	28·66	2537	5·28	16·70
HapMap-CHB-China	44	1788	40·63	1·41	372	8·45	2·44
China	155	9184	59·25	7·27	1020	6·58	6·71
Tibet	31	2586	83·41	2·04	812	26·19	5·34
India	38	2854	75·10	2·26	834	21·94	5·49
HapMap-JPT-Japan	45	1826	40·57	1·44	389	8·64	2·56
Australia	53	5952	114·46	4·71	1761	27·51	11·59
New World	41	4950	105·31	3·92	1224	26·04	8·06
Taiwan	184	18243	117·69	14·45	2511	13·64	16·53

CEU, Utah residents with ancestry from northern and western Europe; CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo, Japan; YRI, Yoruba in Ibadan, Nigeria.

Results

We identified 15 185 singleton genes from a total of 126 190 genes from 44 109 CNVs (Supplementary Table 1) across 12 populations with a mean of 1329 genes per population. Australian genomes bore the heaviest burden from CNVs with the highest mean of 27·51 genes per individual and the lowest burden was observed for AJII (5·28 genes) (Table 1).

CNV burden on genes

CNVs were observed across all regions of the genomes and we did not see any part of the genome that was spared from the CNV burden. About 70% of CNVs were observed in coding regions compared with about 30% of CNVs observed for non-coding regions. An excess of duplications compared to deletions was observed in both coding and non-coding regions, except in a few instances for populations such as CHB and JPT where deletions were observed more in the non-coding region. India, Tibet, Australia, New World, Taiwan and AJ showed a surplus of duplications in both coding and non-coding regions, whereas, CHB and JPT populations showed more deletions than duplications in both coding and non-coding regions; however, Africa offered a balanced level of duplications and deletions against all other populations (Fig. 1(a)).

Chromosomes 8, 15, 17, 1, 14, 22 and 16 show a large concentration of genes (~10–15%) that are under CNV influence, and this remains consistent across all populations; however, there are considerable differences within the populations, for instance, CHB and CEU in chromosome 8 crosses the 35 and 20% threshold, and also CEU and JPT show varying gene concentration crossing 18 and 15% respectively. Chromosomes 20, 18, 13, 12, 21 and 6 show a significantly lower number of genes and the pattern remains consistent with all populations. Sex chromosomes show 3–4% gene content and few populations do not show any gene content, the same is also true for autosomes where a few populations do not show the presence of any genes under the CNVs in a few chromosomes (Fig. 1(b)). CNV burden on the amount of genes was not biased as even the smaller chromosomes sometimes showed >5% of gene content compared to the bigger ones, also it is quite evident that the CNV count and size factors played a relatively significant role in the concentration of the genes.

CNVs were found to disrupt the gene structure both entirely and partially, and a significantly higher number of CNVs were found to disrupt the entire coding structure and these truncations were largely in the form of duplications as opposed to deletions. Taiwan, AJ, Tibet, Australia, India and China show (55–70%) higher duplications in the intact regions compared with other populations. Deletion CNVs were considerably balanced and the frequency varied from 5–10% with AJ and Australia sharing almost equal number of deletion disruptions followed by the remaining populations (~5%). Further, partial gene disruptions were also found to be more often caused by duplication CNVs and in only a few instances were equal numbers of duplications to deletions observed for the disruptions, for instance, New World, JPT and YRI were observed with a near equal number of duplications to deletions ratio, whereas, Australia, AJ, Tibet, India and Taiwan showed elevated duplication CNVs causing partial disruptions of the gene structure (Supplementary Fig. 1(a)).

CNVs were found to be enriched in the regions containing functional genes (83–96%) compared with regions containing pseudogenes (4–17%) (Supplementary Table 2). Australia and AJI showed the highest

https://doi.org/10.1017/S0016672314000202 Published online by Cambridge University Press

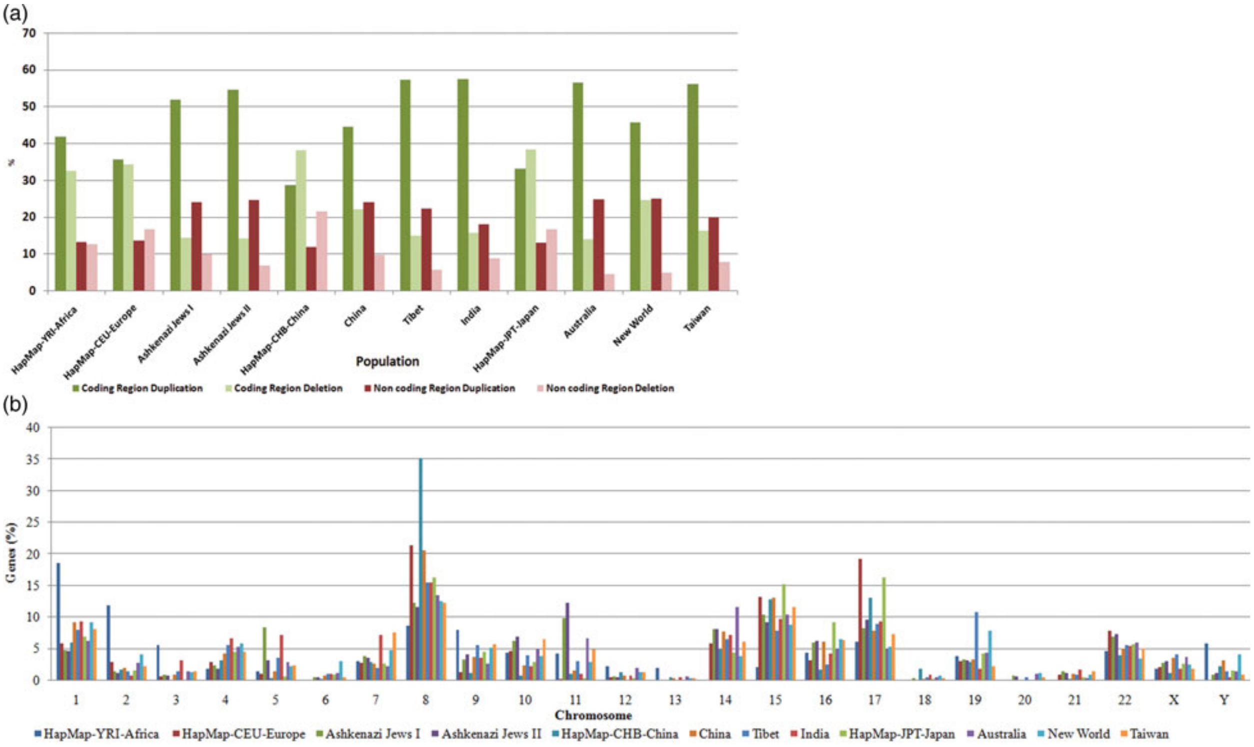

Fig. 1. Distribution of genes across populations. (a) Coding and non-coding regions across chromosomes for all populations. Each cluster consists of four bars representing coding region duplication, coding region deletion, non-coding region duplication and non-coding region deletion respectively. (b) Percentage of genes under copy number variations across chromosomes for all populations, where each cluster represents the 12 populations in different colours. Chromosomes 8, 15, 17, 1, 14, 16 and 22 show large concentration of genes (~10–15%) that are under the copy number variation influence. Chromosomes 20, 18, 13, 12, 21 and 6 show significantly less number of genes and sex chromosomes show 3–4% gene content. CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo, Japan; YRI, Yoruba in Ibadan, Nigeria.

burden (94–96%) in functional genes, while in pseudogenes JPT and CHB bore the heavy burden (16–17%). CNVs partially disrupting genes were found considerably less often than entire overlapping CNVs. Partial gene CNVs were found at a frequency of 4–15% involving both partial duplications and deletions and as anticipated, duplications were higher than deletions, with Australia showing the highest partial disruptions (~15%) followed by Tibet and AJ. YRI and CHB (5–6%) showed equal ratios of less partial duplication to deletions (Supplementary Fig. 1(a)).

Gene concentration in duplication and deletion regions

Gene content was found to be over-represented in duplication regions of 15 chromosomes, and deletion regions of six chromosomes and absolute equal ratios of gene presences were observed in both duplication and deletion regions in three chromosomes (Supplementary Fig. 1(*b*)). The highest gene content under deletion CNVs was observed for CHB (24·2%) followed by YRI (14·5%), CEU (12·9%) and AJII (12·37%) populations, while the highest gene content under duplication CNVs was observed at 10–12% across all populations. There were many chromosomes that showed no gene content under both duplication and deletion CNV events, 42 such deletion CNV events in 13 chromosomes and 14 deletion CNV events were identified from ten chromosomes.

Gene categories

About 5901 genes from the total 15 185 singleton genes identified in the study were divided into three major categories of biological and functional processes and on location using the WebGestalt tool. In biological processes, a majority of the genes were identified under metabolic process and biological regulation, while the lowest burden was observed for genes involved in growth, proliferation, reproduction and death (177–410 genes), and ~2400 genes remained unclassified under any categories (Supplementary Fig. 2(*a*)).

Under molecular function category, genes encoding protein binding, ion binding and DNA binding (nucleic acid and nucleotide binding) were significantly high, followed by genes encoding hydrolase, transducer and transferase activity, which were intermediary, while those performing anti-oxidant activity and oxygen binding were under less burden and 2260 genes remained unclassified (Supplementary Fig. 2 (*b*)). The majority of genes identified coded for proteins that were located in the regions of the membrane and nucleus, while ~300 genes were located in various cell organelles and ~300 were located in extracellular regions and ~2000 proteins remained unassigned (Supplementary Fig. 2(*c*)). We analyzed the functional enrichment of the genes contained in the CNVs from all populations using the Panther classification system (Supplementary Table 3). The CNVs were significantly enriched with genes involved in olfactory receptor activity, retinoic acid receptor binding and signalling, transcription factor binding and immunoglobulin binding.

Phenotypic abnormality of CNV genes

Although these samples are normal and well characterized, no medical information (except for HapMap) was obtained, meaning that structural variation ascertained from them is not necessarily benign or neutral. Normal is referred to as the typical or wild-type condition with no known health defects during the time of sampling procedure. An effort was made to identify disease-associated CNV genes by employing a statistical analysis led prediction model involving 15 185 genes, which are visualized based on organ system as they are broken down into fine dissected narrow units. The resulting network map shows a distinct grouping of genes in the abnormality of the endocrine system with 32 genes involved in the abnormality of the thyroid gland. Abnormalities in skin pigmentation were found for 20 genes resulting in the hypo-pigmentation of the skin. Abnormalities of prenatal development, by way of decreased fetal movement and abnormality of hair through hirsutism were found to be contained in 13 genes each. Abnormality of the urinary system, particularly of the bladder, consisted of seven genes involved in the urinary urgency phenotype. Complete duplication of phalanx, arthritis and anomalous trichromacy phenotypes were found under fewer burdens (Supplementary Fig. 3).

CNV genes interaction network

Discerning the components of protein interactors and networks is a fundamental step towards understanding the protein dynamics in a cell. Proteins identified by CNV genes in the protein interaction network represent a mixture of both direct and indirect interactors. This analysis led to the creation of 8365 high-confidence interactions (0·05% false-discovery rate) involving 14 297 interactions, which are visualized as a network (Fig. 2), the p-values have been corrected for multiple testing. The network shows a distinct grouping of >1500 proteins as the giant subcomplexes of the network encompassing >800 complexes with an elevated degree of interconnectedness. A following group of >50 sparsely connected complexes defined by the network are not connected to other major complexes. The clustering coefficient was found to be 0·226 with a similar centralization value of 0·217, and the mean neighbours for a node was ~18·5 with a network heterogeneity of >2. There were no isolated nodes, self loops nor even multi-edge node pairs in the network. Betweenness centrality, which measures the frequency of a node appearing in a short path between nodes, was found concentrated at 0·16–0·25 with 1–800 nodes appearing in regions with <0·05. The closeness centrality, which measures the mean distance from a starting node to all other nodes, were found aggregated (>10–300 neighbours) from 0·15–0·35.

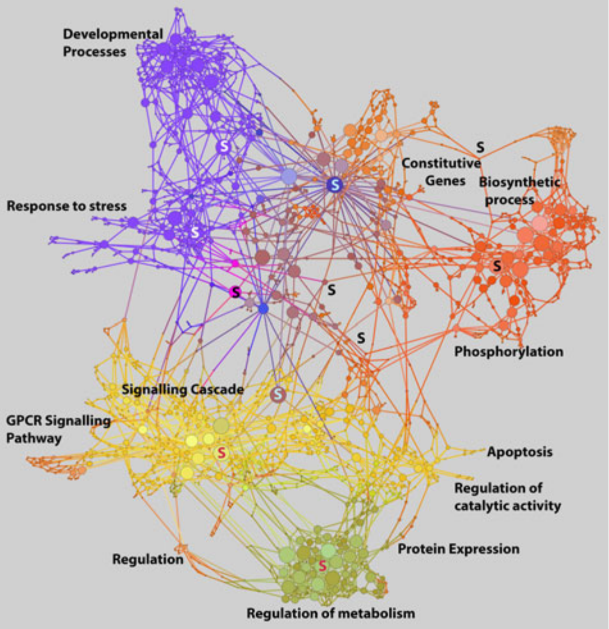

Fig. 2. Copy number variation genes protein interaction map. Network of protein interaction clusters of genes under copy number variations identified across populations labeled with gene ontology accession terms. S indicates stress regions. The network generated has a clustering coefficient of 0·226 and the network heterogeneity is >2. Each gene has an average of ~18·5 neighbours. Distinct clusters showed intense stress on certain major centralized protein components, which further show dense connectivity to sub-nodes with similar functional groups. A total of >100 clusters in the map show significant enrichment for gene ontology terms, and functional pathways ($p < 0{\cdot}01$).

Distinct clusters of protein complexes were observed in the CNV genes network, which showed intense stress on certain centralized protein components. These major protein components further show dense connectivity to sub-nodes with similar functional groups. A total of >100 clusters in the map show significant enrichment for gene ontology (GO)

https://doi.org/10.1017/S0016672314000202 Published online by Cambridge University Press

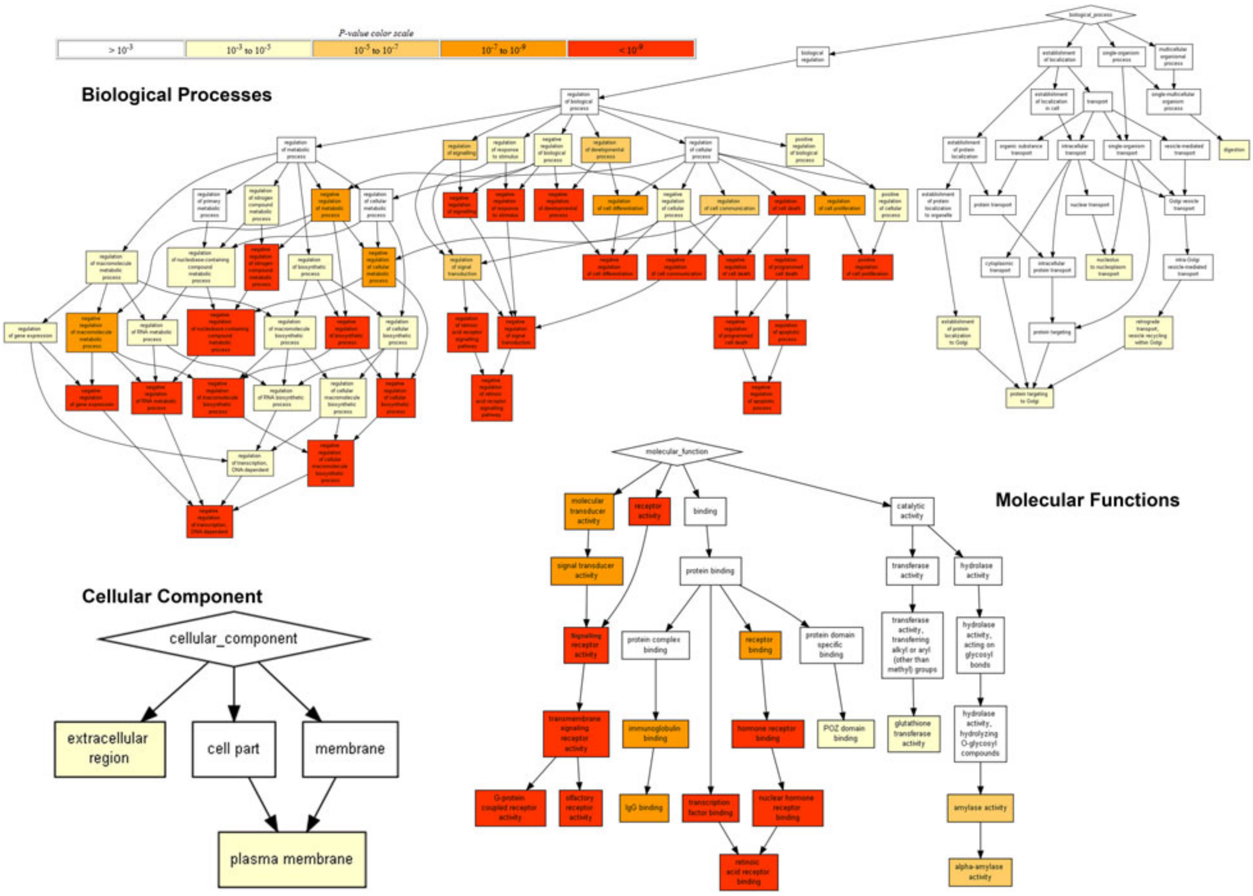
P-value color scale
> 10^-3
10^-3 to 10^-5
10^-5 to 10^-7
10^-7 to 10^-9
< 10^-9
Biological Processes
biological_process
biological regulation
regulation of biological process
regulation of metabolic process
regulation of signalling
regulation of response to stimulus
negative regulation of biological process
regulation of developmental process
regulation of cellular process
positive regulation of biological process
regulation of primary metabolic process
regulation of nitrogen compound metabolic process
negative regulation of metabolic process
regulation of cellular metabolic process
negative regulation of signalling
negative regulation of response to stimulus
negative regulation of developmental process
regulation of cell differentiation
negative regulation of cellular process
regulation of cell communication
regulation of cell death
regulation of cell proliferation
positive regulation of cellular process
regulation of macromolecule metabolic process
regulation of nucleobase-containing compound metabolic process
negative regulation of nitrogen compound metabolic process
regulation of biosynthetic process
negative regulation of cellular metabolic process
regulation of signal transduction
negative regulation of cell differentiation
negative regulation of cell communication
negative regulation of cell death
regulation of programmed cell death
positive regulation of cell proliferation
regulation of gene expression
negative regulation of macromolecule metabolic process
regulation of RNA metabolic process
negative regulation of nucleobase-containing compound metabolic process
regulation of macromolecule biosynthetic process
negative regulation of biosynthetic process
regulation of cellular biosynthetic process
regulation of retinoic acid receptor signalling pathway
negative regulation of signal transduction
negative regulation of programmed cell death
regulation of apoptotic process
negative regulation of gene expression
negative regulation of RNA metabolic process
negative regulation of macromolecule biosynthetic process
regulation of RNA biosynthetic process
regulation of cellular macromolecule biosynthetic process
negative regulation of cellular biosynthetic process
negative regulation of retinoic acid receptor signalling pathway
negative regulation of apoptotic process
regulation of transcription, DNA dependent
negative regulation of cellular macromolecule biosynthetic process
negative regulation of transcription, DNA dependent
establishment of localization
single-organism process
multicellular organismal process
establishment of localization in cell
transport
single-multicellular organism process
establishment of protein localization
organic substance transport
intracellular transport
single-organism transport
vesicle-mediated transport
digestion
establishment of protein localization to organelle
protein transport
nuclear transport
Golgi vesicle transport
cytoplasmic transport
intracellular protein transport
nucleolus to nucleoplasm transport
intra Golgi vesicle-mediated transport
establishment of protein localization to Golgi
protein targeting
retrograde transport, vesicle recycling within Golgi
protein targeting to Golgi
Molecular Functions
molecular_function
molecular transducer activity
receptor activity
binding
catalytic activity
signal transducer activity
protein binding
transferase activity
hydrolase activity
signalling receptor activity
protein complex binding
receptor binding
protein domain specific binding
transferase activity, transferring alkyl or aryl (other than methyl) groups
hydrolase activity, acting on glycosyl bonds
transmembrane signalling receptor activity
immunoglobulin binding
hormone receptor binding
POZ domain binding
glutathione transferase activity
hydrolase activity, hydrolyzing O-glycosyl compounds
G-protein coupled receptor activity
olfactory receptor activity
IgG binding
transcription factor binding
nuclear hormone receptor binding
amylase activity
retinoic acid receptor binding
alpha-amylase activity
Cellular Component
cellular_component
extracellular region
cell part
membrane
plasma membrane

Fig. 3. Enrichment of genes under copy number variations in molecular, cellular and biological pathways. A gradient p-value colour scale from 10^{-3}–10^{-9} has been allotted for the various processes.

terms, and functional pathways ($p < 0{\cdot}01$) indicating that members share common biological or functional attributes. These >100 annotation-enriched clusters include >2200 proteins that lacked any annotation. Assessment of protein complexes provides insights into specific as well as general functional aspects of the map.

CNV gene enrichment pathways

Genes of the similar functional group are expected to be enriched for GO (Panther Classification and GO enRIchment anaLysis and visuaLizAtion tool (GORILLA)) annotations, share the same pathways (WikiPathways, KEGG and Pathway Commons), enabling the integrated use of these systems to establish biological, molecular and cellular pathways based on the genes enriched in CNVs. The statistical analysis to score gene enrichment on the pathway map revealed varying enrichment along both biological and molecular pathways (Supplementary Fig. 4 and Supplementary Fig. 5). The biological pathway begins from several thousands of genes, which are under CNVs, and continues to dissect until the precise point of significant enrichment is reached. The pathway analyses revealed an exhaustive sum of genes to be under the broad marquee of regulation of metabolic processes, signalling and development (Fig. 3). Genes for negative regulation of metabolic processes were enriched (p-value: 10^{-7}–10^{-9}) compared with positive regulation of primary and cellular metabolic process (including nitrogen compounds) (p-value: 10^{-3}–10^{-5}). Negative regulation continued to be enriched along the downstream of the pathway with few instances of positive regulation. Gene enrichment (p-value: 10^{-7}–10^{-9}) for negative regulation was specifically observed for nucleo-base containing compounds, macromolecules and other bio-synthetic processes compared with genes enriched for regulating gene expression, RNA metabolism and for other cellular bio-synthetic processes (p-value: 10^{-3}–10^{-5}); however, the former processes also showed negative regulation further downstream and the enrichment of genes reached the crucial and definitive point for DNA dependent transcription regulation (p-value: 10^{-9}).

On the enrichment of genes in the regulation of signalling and development front, the regulation process continued to be under burden, with regulation of both types being balanced. Genes for positive regulation were found to be enriched for cell proliferation while

https://doi.org/10.1017/S0016672314000202 Published online by Cambridge University Press

genes for negative regulation were over-represented for cell communication/signalling, especially for retinoic acid receptor signalling (p-value: 10^{-7}). Genes for cell death were significantly enriched throughout the pathway with no deviations, and genes for regulation of apoptotic processes followed by negative regulation of programmed cell death enhanced the pathway leading to significant negative regulation of apoptotic process (p-value: 10^{-9}). Genes were slightly under-represented in pathways involving golgi organelles (protein localization and tagging, and vesicle recycling) and digestion processes (p-value: 10^{-3}).

Genes in molecular functions were enriched in signalling mechanisms preceded by receptor binding events, while those involved in catalytic activity were less represented. Immunoglobulin binding, G-protein coupled olfaction signalling and hormone-receptor binding were significantly enriched with genes (p-value: 10^{-9}), but the hydrolysing functions of alpha-amylase were decently represented (p-value: 10^{-5}). Component wise genes enrichment was observed for proteins located in the plasma membrane and extracellular region (p-value: $>10^{-3}$) compared to all other regions (Fig. 3 and Supplementary Fig. 6).

Complexities of multiple hits in personal genomes

On further investigation of the complexities of CNVs, we found a unique CNV influence phenomenon on the genomes of the study subjects. CNVs across the genome of an individual showed multiple hits/presences across many genes, whose proteins interact physically and function under the same pathway (Fig. 4(*a–h*)). We identified varying number of proteins (1–15) and degrees (1–10 interactions) within protein complexes of single individual genomes. These multiple CNV hit protein complexes were found to contain proteins whose genes were either deleted or duplicated and sometimes even both. The intact or partial disruptions of genes were computed along with copy number (CN) states for each CNV event within multiple CNV hit protein complexes. These complications arising from the CNVs were not limited to enrichment of one or multiple pathways, but instead showed diverse occurrences across multiple pathways and were largely determined by the number of the CN states for each gene of the protein complex that showed haplo-insufficiency for a certain amount of genes, while complete knock out of the genes were observed resulting in no protein.

Breakpoint validations

Successful amplification of four recurring CNV breakpoints was performed on 400 randomly chosen individuals from India validating the presence of the CNVs. Varying frequency and amplification status was observed for breakpoints. About 48–54% frequency for the chosen breakpoints was observed for the chosen group.

Discussion

Identification of CNVs across diverse populations helps in the understanding of the organization and distribution pattern of CNVs, evolutionary dynamics of the human genome and accounts for differences in the expression of genes (Nguyen *et al.*, 2006). There have been increasing numbers of CNV survey studies using different ethnic backgrounds; however, there have not been many that comparatively include populations across all continents to study notable variations on the genome, particularly on genes. This study represents the first draft of a population-specific CNV genes map as well as a cross-populational map. Here we present a comprehensive global CNV gene spectrum by identifying 15 185 singleton genes across 12 populations. The genes and CNVs were synergistic and conferred tremendous burden on the coding genome. Genetic diversity in humans affects both disease and normal phenotypic variation. Presence of CNVs alters the transcriptional and translational levels of overlapping or nearby genes by disrupting the coding structure or by altering gene dosage thereby conferring differential susceptibility to complex diseases (Gonzalez *et al.*, 2005; Aitman *et al.*, 2006; Fellermann *et al.*, 2006; Park *et al.*, 2006; Fanciulli *et al.*, 2007; Yang *et al.*, 2007; Hollox *et al.*, 2008).

CNVs were observed in both coding and non-coding parts of the genome and interestingly, all populations showed a consistent rate of CNVs in coding as well as non-coding regions, but with the former region showing more number of CNVs than the latter. CNVs across the human genome have been found to be associated with normal genetic heterogeneity as well as for a number of diseases and disorders. Previous studies have identified CNVs more in non-coding regions than in coding regions hence their role and correlations is often not clear (Feuk *et al.*, 2006; Redon *et al.*, 2006). However, it can be reasoned that non-coding regions may play an important regulatory role for the distant genes by acting like enhancers or suppressors. Whereas, for CNVs that occur in coding regions, knowledge of individual genes underlying these regions, followed by correlating this knowledge with the biological pathways may help explain the role of CNVs on disease phenotypes.

CNV counts and its size distribution affect the genes underlying them. Since distribution of genes on chromosomes is not even, therefore, the CNV count and size will vary from one CNV to the other. CNV burden on genes was found highest in the 8th chromosome across all populations with CHB dominating the group. Though the CNV count was

https://doi.org/10.1017/S0016672314000202 Published online by Cambridge University Press

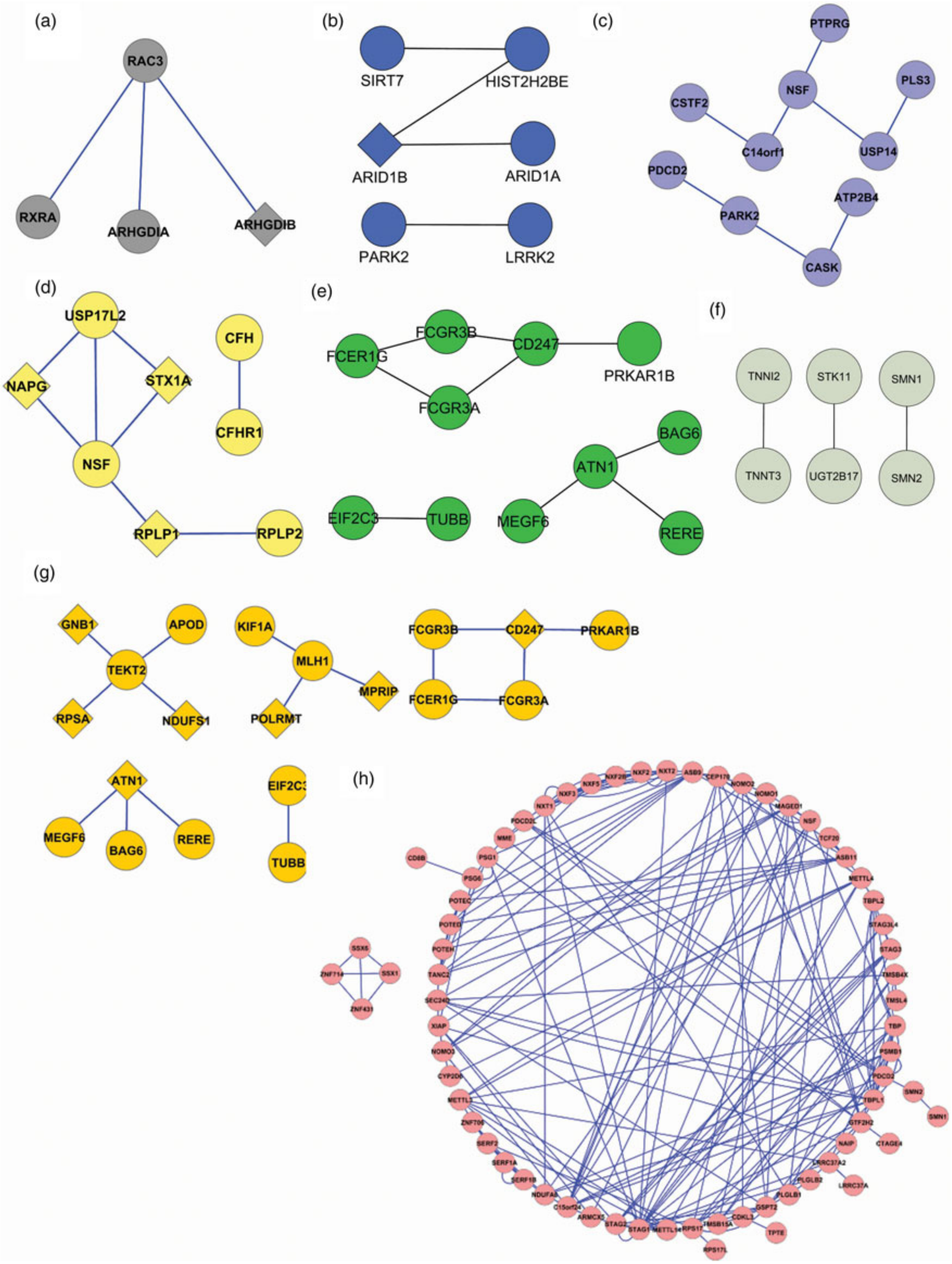

Fig. 4. Complexities of copy number variation effect on genes. Representation of the different levels of complexities (a–h) present in the genome as a result of copy number variation influence.

observed more in chromosome 14 than in chromosome 8, the CNV burden on genes was observed high on the 8th chromosome but intermediate on the 14th chromosome. Although there is not much difference in the gene count between chromosomes 8 and 14, the CNV presence in chromosome 8 was found enriched with genes, and on the contrary, the 15th chromosome showed an abundance of both CNV and gene counts. Besides, the least number of genes was found in chromosomes 13, 18, 20 and 21, with a minimum of two populations failing to show any count on any of the chromosomes. On the whole,

https://doi.org/10.1017/S0016672314000202 Published online by Cambridge University Press

CNV count was found to be directly proportional with the gene concentration and the majority of the CNV count was found in the gene poor regions except in the instance of chromosome 8. Previous studies reported CNVs to be abundant in the gene poor regions; however, the present study indicates the CNV presence in both gene rich and poor regions. CNVs were observed to be abundant in coding regions of the genome compared with non-coding regions, with CEU, CHB and JPT populations showing bias towards more deletions in both coding and non-coding regions. The higher proportion of CNV overlaps in the coding regions could be interpreted to reflect selection favouring the fixation of gene-containing structural changes and we believe it is these changes that drive gene and genome evolution.

CNVs overlapping entire and partial genes

Many CNVs were found to be overlapping genes in the genome, sometimes entirely or partially. Partial overlaps of genes truncated either in the exon or intron regions of the gene, whereas complete overlap of genes truncated from a little upstream involving the regulatory region or sometimes from the coding region. All populations showed >45% of entire gene duplication, whereas entire gene deletion ranged from 12–45%. The data points for the entire gene duplications and deletions plot were a reflection of a mirror image (when halved), with populations showing duplications on the higher side, and lower number of deletions on the diametrically opposite side; for instance, Taiwan showed higher number of entire gene duplications (~72%) in regards to the least number of entire gene deletion (12%), and this pattern remained consistent for all populations.

Of the 15 185 singleton genes identified in the study, the majority of genes were identified under metabolic processes and biological regulation, while lowest burden was observed for genes involved in growth, proliferation, reproduction and death. We believe the extent of mutational burden on genes involved in the metabolic pathway demonstrates the tolerable limit the human system can bear since metabolism is involved only in the performance of an individual unlike other sensitive organ systems, which would be fatal with such a burden. This elasticity enables the sustainment of the variation burden in metabolic pathways.

Genes encoding protein binding, ion binding and DNA binding were found to be significantly high under CNVs and these genes are typically the constitutive genes that are required for the maintenance of basic cellular function, and are expressed in almost all cells in both normal and pathological conditions. Some of these genes usually express at relatively constant levels, however, presence of CNVs alters the expression levels by either decreasing or increasing them, and the possible effects of such changes on the genome are yet to be seen.

CNVs were found to be significantly enriched with genes involved in olfactory receptor activity, retinoic acid receptor binding and signalling, transcription factor binding and immunoglobulin binding across all populations. Some of these pathways were also found enriched in the variations reported from a recent study (Azim *et al.*, 2013). Genes under previously known CNVs were mainly related to processes such as development, cell adhesion and regulation of gene expression (Krzywinski *et al.*, 2009). However, our results indicate deviation towards the over-representation of genes related to growth, development and signalling, and slightly overlaps with previous studies in processes such as secretion, cell adhesion and immunity-related proteins in CNV burdened genes.

CNV genes based pathways indicated an enrichment of genes involved in the abnormalities of thyroid, skin (hypo-pigmentation) and bone (arthritis), these medical conditions impair the general population by frequenting in 2–8% of the population. Abnormality of thyroid is the non-functioning of the thyroid resulting from the low/high activity or even due to lack of thyroid gland, and is also found to be associated with increased stress. A study by Abu-Helalah *et al.*, (2010) concluded that about 8% of women >50 years old and men >65 years old in UK have a medical condition from an under active thyroid. Hypo-pigmentation is another condition, found enriched in CNVs, which results in the loss of skin colour. It is caused by melanocyte or melanin depletion, or due to a decrease in tyrosine. The CNVs identified here can be described as frequent variations in the population and argues for an involvement of CNVs in the etiology of these conditions. Though the present study focuses on the significantly enriched pathways, singling out specific genes to identify up/deregulated pathways will further help in the understanding of the frequent burden of CNVs on the genome.

CNV genes interaction network

Determining the consequence on complexes of functional protein networks and their interactions is an essential step towards the understanding of the altered protein dynamics caused due to the presence of CNVs on the genome (Guruharsha *et al.*, 2011). The protein interaction network was constructed in an effort to identify the interconnected stress regions of the functional pathways caused by the presence of CNVs near the genes. The network showed high-confidence interactions involving >14 000 connections and revealed several stress bearing regions involved in pathways with an elevated degree of interconnectedness. Distinct clusters of protein complexes were

https://doi.org/10.1017/S0016672314000202 Published online by Cambridge University Press

observed in the CNV genes network, which showed intense stress on certain centralized protein components. These major protein components showed further interconnectedness to sub-nodes of similar functional groups. The regions of the network with various functional groups converging and interlinking initiated by few protein representatives from each functional pathway were more strained, these regions further provided insight into the sensitiveness of functional pathways and on genes/complexes that are under such tension.

Minor allele frequency

We identified a total of 51 singleton genes from 15 185 genes across 12 populations with a frequency of >50%, these 51 genes were found to be under tremendous influence of CNVs and were found to participate in diverse pathways (Table 1 and Supplementary Fig. 7). Singleton refers to a representative gene taken from a set of repetitive occurrences of that gene, under CNVs in multiple individuals. Our analysis of different populations allowed an assessment of population-specific genes (Krzywinski *et al.*, 2009) (Supplementary Fig. 8). We did not find any single factor that might explain the reason behind the force of CNVs on these genes, but it seems highly plausible that these genes have originated due to a high rate of multiplicity. Though some among these 51 genes confer a population-specific phenotype, most of them are found across all populations at similar frequencies indicating a universal effect of CNVs on these gene-bearing regions.

Genotypic complexities provided by CNVs

CNVs in genes influence expression by either increasing or reducing it. Studies on the implications of CNVs on disease genes have been extensively studied; however, the layers of complexities remained unidentified. Although there have been numerous investigations to find the influence of CNVs on gene-expression phenotypes, only a few studies have effectively been able to associate them with diseases (Girirajan *et al.*, 2012), while for the most part CNVs still need to be explored in order to understand phenotypic differences. In the present investigation, several layers of complexities created due to the presence of multiple CNVs across many gene regions, which physically interact among themselves within a genome, was extensively studied to unravel the complex mechanism of CNVs towards the phenotype. This was performed by identifying the protein interactions of genes, which were disrupted by the presence of CNVs. The following CNV complexities were identified in the genomes of individuals: (i) CNVs disrupted the coding structure of 5–25 genes either by duplications or deletion of the entire or partial genes; (ii) the presence of CNVs in the genomic regions of the immediate interacting protein partners were also identified in the same individual; (iii) identification of knock out of multiple genes or protein complexes due to CNVs; and (iv) correlation between duplication and deletion events with gene overlaps/non-overlaps along with CN state of all the genes involved in the protein complexes (Fig. 4(*a–h*)). For instance, CNVs were identified to be disrupting the genes *GTF2I* (-), *CCR1* (+), *CCL3L1* (-), *EIF2AK2* (+), *HISTH2AC* (+) and *CDY1* (-) in an individual, and all the protein forms of these genes were found to physically interacting among each other through the *STAT3* (-) intermediary. The '+' and '-' sign beside each gene indicates the CNV type of duplication and deletion in an individual. Instances such as these were found in the genomes of many individuals. The layers of CNV complexities are amplified with the increase of interacting partners among the disrupted genes.

This complex effect of CNVs on genes and their immediate interacting protein partners, brought about by the burden of CNVs, unravels the complexities involved in gene function and phenotypic expression. We propose four different types of mechanisms that create the complexity in gene function, indicating that multiple CNV–genes–protein partners are potential functional variants and these three proposed mechanisms should be considered as criteria while performing genotype–phenotype association and gene-regulation studies. This complex effect of CNVs on genes and their physical interacting partners due to the presence of CNVs, unravels the many complexities involved in phenotype expression.

We thank the subjects and their families for participating in this study; The Chairman, DOS in Zoology, University of Mysore; Prof. H. A. Ranganath, Prof. K. S. Rangappa and Prof. P. Prakash, University of Mysore for their help and encouragement; and also University of Mysore for providing the facilities to conduct this work; CSIR, New Delhi, for awarding Fellowship to K. Lingaiah (CSIR Order No.9/119 (0196)2K13-EMR-I Dated: 19.03.2013); DST-INSPIRE, New Delhi, for awarding Fellowship to M. N. Murthy (IF120351, DST sanction letter No. DST/INSPIRE FELLOWSHIP/2012/327); E. M. Locke of Western University, London for their inputs on HD-CNV; P. N. Radhika and A. N. Somanna, University of Mysore for their support.

Declaration of interests

None.

Supplementary material

The online supplementary material can be found available at http://journals.cambridge.org/GRH

https://doi.org/10.1017/S0016672314000202 Published online by Cambridge University Press

References

Abu-Helalah, M., Law, M. R., Bestwick, J. P., Monson, J. P. & Wald, N. J. (2010). A randomized double-blind crossover trial to investigate the efficacy of screening for adult hypothyroidism. *Journal of Medical Screening* **17**, 164–169.

Aitman, T. J., Dong, R., Vyse, T. J., Norsworthy, P. J., Johnson, M. D., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A. J., Petretto, E., Hodges, M. D., Bhangal, G., Patel, S. G., Sheehan-Rooney, K., Duda, M., Cook, P. R., Evans, D. J., Domin, J., Flint, J., Boyle, J. J., Pusey, C. D. & Cook, H. T. (2006). Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855.

Armengol, L., Villatoro, S., González, J. R., Pantano, L., García-Aragonés, M., Rabionet, R., Cáceres, M. & Estivill, X. (2009). Identification of copy number variants defining genomic differences among major human groups. *PLoS One* **4**, e7230.

Azim, M. K., Yang, C., Yan, Z., Choudhary, M. I., Khan, A., Sun, X., Li, R., Asif, H., Sharif, S. & Zhang, Y. (2013). Complete genome sequencing and variant analysis of a Pakistani individual. *Journal of Human Genetics* **58**, 622–626.

Beckmann, J. S., Estivill, X. & Antonarakis, S. E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Reviews Genetics* **8**, 639–646.

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

Cheung, J., Estivill, X., Khaja, R., MacDonald, J. R., Lau, K., Tsui, L. C. & Scherer, S. W. (2003). Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biology* **4**, R25.

Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J. M., Gough, S. C., de Smith, A., Blakemore, A. I., Froguel, P., Owen, C. J., Pearce, S. H., Teixeira, L., Guillevin, L., Graham, D. S., Pusey, C. D., Cook, H. T., Vyse, T. J. & Aitman, T. J. (2007). *FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature Genetics* **39**, 721–723.

Fellermann, K., Stange, D. E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C. L., Reinisch, W., Teml, A., Schwab, M., Lichter, P., Radlwimmer, B. & Stange, E. F. (2006). Chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *American Journal of Human Genetics* **79**, 439–448.

Feuk, L., Carson, A. R. & Scherer, S. W. (2006). Structural variation in the human genome *Nature Reviews Genetics* **7**, 85–97.

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

Girirajan, S. & Eichler, E. E. (2010). Phenotypic variability and genetic susceptibility to genomic disorders. *Human Molecular Genetics* **15**, R176–R187.

Girirajan, S., Rosenfeld, J. A., Coe, B. P., Parikh, S., Friedman, N., Goldstein, A., Filipink, R. A., McConnell, J. S., Angle, B., Meschino, W. S., Nezarati, M. M., Asamoah, A., Jackson, K. E., Gowans, G. C., Martin, J. A., Carmany, E. P., Stockton, D. W., Schnur, R. E., Penney, L. S., Martin, D. M., Raskin, S., Leppig, K., Thiese, H., Smith, R., Aberg, E., Niyazov, D. M., Escobar, L. F., El-Khechen, D., Johnson, K. D., Lebel, R. R., Siefkas, K., Ball, S., Shur, N., McGuire, M., Brasington, C. K., Spence, J. E., Martin, L. S., Clericuzio, C., Ballif, B. C., Shaffer, L. G. & Eichler, E. E. (2012). Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *New England Journal of Medicine* **367**, 1321–1331.

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.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J., Murthy, K. K., Rovin, B. H., Bradley, W., Clark, R. A., Anderson, S. A., O'Connell, R. J., Agan, B. K., Ahuja, S. S., Bologna, R., Sen, L., Dolan, M. J. & Ahuja, S. K. (2005). The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440.

Guruharsha, K. G., Rual, J. F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D. Y., Cenaj, O., McKillip, E., Shah, S., Stapleton, M., Wan, K. H., Yu, C., Parsa, B., Carlson, J. W., Chen, X., Kapadia, B., VijayRaghavan, K., Gygi, S. P., Celniker, S. E., Obar, R. A. & Artavanis-Tsakonas, S. (2011). A protein complex network of *Drosophila melanogaster*. *Cell* **147**, 690–703.

Hollox, E. J., Huffmeier, U., Zeeuwen, P. L., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof., P. C., Traupe, H., de Jongh, G., den Heijer, M., Reis, A., Armour, J. A. & Schalkwijk, J. (2008). Psoriasis is associated with increased beta-defensin genomic copy number. *Nature Genetics* **40**, 23–25.

Kanduri, C., Ukkola-Vuoti, L., Oikkonen, J., Buck, G., Blancher, C., Raijas, P., Karma, K., Lähdesmäki, H. & 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.

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

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, J. & Horsman, D. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research* **19**, 1639–1645.

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 and Genomic Medicine* **1**, 142–154.

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

https://doi.org/10.1017/S0016672314000202 Published online by Cambridge University Press

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

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 of the United States of America* **100**, 12235–12240.

Park, J., Chen, L., Ratnashinge, L., Sellers, T. A., Tanner, J. P., Lee, J. H., Dossett, N., Lang, N., Kadlubar, F. F., Ambrosone, C. B., Zachariah, B., Heysek, R. V., Patterson, S. & Pow-Sang, J. (2006). Deletion polymorphism of UDP-glucuronosyltransferase 2B17 and risk of prostate cancer in African American and Caucasian men. *Cancer Epidemiology, Biomarkers and Prevention* **15**, 1473–1478.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W. & Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature* **444**, 444–454.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y. H., Hicks, J., Spence, S. J., Lee, A. T., Puura, K., Lehtimäki, T., Ledbetter, D., Gregersen, P. K., Bregman, J., Sutcliffe, J. S., Jobanputra, V., Chung, W., Warburton, D., King, M. C., Skuse, D., Geschwind, D. H., Gilliam, T. C., Ye, K. & Wigler, M. (2007). Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449.

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.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., Prchal, J. T. & 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., Sampas, N., Bruhn, L., Shendure, J., 1000 Genomes Project & 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., Vishweswaraiah, S., Lingaiah, K., Murthy, M., Manjegowda, D. S., Nayaka, R. & Ramachandra, N. B. (2013 *a*). 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*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.

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., Nord, A. S., Kusenda, M., Malhotra, D., Bhandari, A., Stray, S. M., Rippey, C. F., Roccanova, P., Makarov, V., Lakshmi, B., Findling, R. L., Sikich, L., Stromberg, T., Merriman, B., Gogtay, N., Butler, P., Eckstrand, K., Noory, L., Gochman, P., Long, R., Chen, Z., Davis, S., Baker, C., Eichler, E. E., Meltzer, P. S., Nelson, S. F., Singleton, A. B., Lee, M. K., Rapoport, J. L., King, M. C. & Sebat, J. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543.

Yang, Y., Chung, E. K., Wu, Y. L., Savelli, S. L., Nagaraja, H. N., Zhou, B., Hebert, M., Jones, K. N., Shu, Y., Kitzmiller, K., Blanchong, C. A., McBride, K. L., Higgins, G. C., Rennebohm, R. M., Rice, R. R., Hackshaw, K. V., Roubey, R. A., Grossman, J. M., Tsao, B. P., Birmingham, D. J., Rovin, B. H., Hebert, L. A. & Yu, C. Y. (2007). Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *American Journal of Human Genetics* **80**, 1037–1054.

Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics* **10**, 451–481.

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.

https://doi.org/10.1017/S0016672314000202 Published online by Cambridge University Press