Recent Advances in Studying of Copy Number Variation and Gene Expression

George E. Liu¹, Lingyang Xu¹,² and Kevin S. Huang¹,³

¹Bovine Functional Genomics Laboratory, USDA, ARS, Beltsville, MD, USA. ²Department of Animal and Avian Sciences, University of Maryland, College Park, MD, USA. ³Centennial High School, 4300 Centennial Lane, Ellicott City, MD, USA.

ABSTRACT: Copy number variations (CNVs) are gains and losses of genomic sequence between two individuals of a species. CNV can have a significant impact on the expression of genes mapped within, nearby and far away from the CNV event itself. Earlier studies have revealed that CNV effects can be caused by affecting gene dosage, through position effect, altering downstream pathways and regulatory networks, or modifying the chromosome structure and position within the nucleus. We will first review recently published results based on high-resolution CNV and transcriptome data. We will then discuss new insights brought by emerging new technologies of next generation sequencing and single-cell sequencing.

KEYWORDS: copy number variation, gene expression, mechanism, next generation sequencing, single-cell sequencing

INTRODUCTION

One major theme of modern genetics is to understand the relationship between genotype and phenotype. Variation of gene expression (narrowly defined here as the transcription of DNA into mRNA) may serve as a proxy for phenotypic variation.⁵ Gene expression has been studied in many species in order to understand its genetic basis of variation and its regulatory networks. These studies provide multiple models to explain the causes of phenotypic variation.

Along with single nucleotide polymorphisms (SNPs), recent studies dealing with larger, more complex forms of genetic variation, structural variations, have attracted much attention.⁴ Among the largest of the genetic variant types, copy number variations (CNVs or unbalanced structural variations) are deletions and insertions of genomic sequence between two individuals of a species.⁵ Additionally, balanced structural variants include chromosomal fusions, translocations, and inversions. In this review, we generally refer all of them as CNV unless we state otherwise. Phenotypic effects of CNVs are caused by changes in gene expression, directly affecting gene dosage, indirectly through position effect, or downstream pathways and regulatory networks. One extreme example of CNV is chromosome aneuploidy (ie imbalances in chromosome number), which has been shown to globally influence gene expression on the variant chromosome and is usually lethal to the host.⁶,⁷ Early studies of CNV and gene expression and their potential mechanisms have been reviewed before.⁸,⁹ In this review, we will focus on newly published CNV studies based on high-resolution CNV and transcriptome data. We will then discuss influences of emerging new technologies of next generation sequencing (NGS) and single-cell sequencing on the future studies.
CNVs in Human and Model Organisms

CNVs have been intensively studied in humans and several model organisms, including mouse, rat, and fruit fly. Besides a number of plants, CNV reports also appeared in other animals, such as chimpanzee, macaque, dog, and cattle. The newest version (July 2013) of Database of Genomic Variants (DGV) collected 55 human CNV studies (http://dgv.tcg.ca/dgv/app/home), containing 230,4349 human CNVs and 109,863 CNV regions. The Database of Genomic Structural Variation (dGVaR) hosts 106 published studies, including 85 human, 12 mouse, and several other animal datasets (http://www.ncbi.nlm.nih.gov/dgv/).

Different CNV detection platforms (hybridization or sequence based) and calling algorithms resulted in different CNV datasets in varying resolution. Several major CNV formation mechanisms have been proposed, including NAHR (Non-Allelic Homologous Recombination), NHEJ (Non-Homologous End Joining), MEI (Mobile Element Insertion), and FoSTeS/MMBIR (Fork Stalling and Template Switching and Microhomology-Mediated Break-Induced Replication). Specifically, large CNVs are mainly formed by NAHR mediated by repetitive sequences, including segmental duplications (SDs). Although tandem duplications seem the most likely archetypical organization as shown in mouse, rat, dog, cattle, and fruit fly, human and great ape species show a preponderance of interspersed duplications. As a result, large CNVs are associated with SDs in mammals and enriched in centromeres and telomeres.

Dozens of CNVs have been shown to be important in both normal phenotypic variability and disease susceptibility in human. Although analyses of a subset of CNVs provided evidence of linkage disequilibrium with flanking SNPs, a significant portion of CNVs fell in genomic regions not well covered by SNP arrays, such as SD regions, and thus were not genotyped. Combining CNV and SNP data in human genome-wide association studies has associated CNVs with diseases such as intellectual disability, autism, schizophrenia, neuroblastoma, Crohn’s disease, and severe early-onset obesity.

Early Studies of CNVs and Gene Expression

By year 2011, several reports have reported associations of CNVs with gene expression variation in human, rodents, and fruit fly based on early low-resolution CNV and gene expression data. For example, a comprehensive survey by Stranger et al reported that SNPs and CNVs captured about 2% of the total detected genetic variation in gene expression, respectively. Chaingnat et al studied brain and liver tissues of mouse embryos at different stages and observed the spatial and temporal influence of CNVs on tissue transcriptomes throughout development. Notably, they detected that CNVs are significantly enriched within differentially expressed genes in both adulthood and embryonic stage among mouse strains. Additionally, they found that only in brain, some expressed genes within CNVs appear to be under compensatory loops only at specific time points, suggesting that the effect of CNVs on these genes in brain is modulated during development. These studies and potential mechanisms have been reviewed before. We will briefly summarize molecular mechanisms below. All of them may play a role, either individually or in combination, highlighting complex relationships between CNV and gene expression.

**Dosage sensitive.** Only a weak positive correlation was found between CNV and gene expression, dependent on tissue type. This correlation is driven by a small fraction of genes (5–18% in rodents). A dosage reversal is used to describe the physical dissociation of the gene and its regulatory cis elements. Position effect has been reported for the expression of genes of up to 1.5 Mb from the CNV event itself.

**Dosage insensitive and dosage compensation.** In over 70% of the fruit fly genes, CNV had no effect on gene expression, suggesting either dosage compensation mechanisms or the incomplete inclusion of regulatory elements in the CNV event. In addition, mechanisms similar to genetic imprinting may also inhibit the expression at a CNV locus.

**Dosage reversed.** For 2–15% of the genes, gene expression levels were significantly inversely correlated with CNV. This observation can be explained by either a negative feedback loop, which reduces the expression of the CNV gene or a steric hindrance of the extra copies of a gene, which impairs their access to a specific transcription factory. For example, deletion of a transcriptional repressor could serve to elevate gene expression.

**Position effect.** Position effect or cis-rupture is used to describe the physical dissociation of the gene and its regulatory cis elements. Position effect has been reported for the expression of genes of up to 1.5 Mb from the CNV event itself.

**Chromosome structure and positioning.** There are several hypotheses explaining disruption of CNV of long-range trans regulations: CNVs can result in modification of transcriptional control through alteration of chromosome structure and modification of the positioning of chromosome within the nucleus and/or within a chromosome territory of a genomic region.

**Other effects.** Deletion of CNVs can alter a phenotype by unmasking recessive mutations. CNVs may also influence gene expression through perturbation of transcript structure.

Altogether, this first batch of studies only evaluated the effects of large CNVs at a low resolution, ascertained with the early versions of microarray platforms with low probe densities. Indeed, array comparative genomic hybridization (aCGH) can only assay the sequence present in the reference assembly. It cannot distinguish between tandem and non-tandem duplications. Another caveat is the reliability of CNV detection, particularly in defining CNV breakpoints with confidence when aCGH data are used. Unreliable boundary calling can lead to genes being erroneously included or excluded from CNVs, which can artificially skew gene expression profiles and affect correlation results.
Recent Studies of CNVs and Gene Expression

Technological advances in microarray and NGS have recently led to "second-generation" CNV maps with increased resolution and broadened CNV size range.⁵,¹⁴ The 1000 Genomes Project reported CNVs of 50 bp to a 1 Mb in size (median size 730 bp) in more than 150 individuals, with over 50% of CNV were mapped to nucleotide resolution.⁵ The count of detected CNVs has increased from below a hundred to several thousand per individual.⁵,¹⁴ Improved CNV genotyping technology further facilitated associating CNVs with phenotypic data.⁵,¹⁴,⁴¹,⁷²,⁷³

Recent human studies. By relating second-generation CNV genotypes to transcriptome sequencing (RNA-Seq) data, Schlattl et al evaluated the impact of CNV on gene expression focusing on proximal effects less than 200 kb.⁷⁴ Totally, they identified CNVs associated with the expression of 110 genes. They observed an enrichment of large CNVs, including large intergenic CNVs, relative to the entire set of expression-associated CNVs. While a minor part of gene expression may be affected by CNV, their results suggest that dosage compensation in human cells may be less pronounced than in Drosophila melanogaster.²⁰ Furthermore, 48% of the genes they identified in the context of CNV-associated expression QTLs were also associated with SNPs before.⁷⁵,⁷⁶ This portion is higher than the fraction of genes (18%) associated with CNVs by Stranger et al, which also displayed a significant SNP association.⁵⁵ They attribute this higher overlap to the increased resolution and accuracy of the second-generation CNV data. Altogether, their results suggest that association studies can gain in resolution and power by including fine-scale CNV information. It is noted that this study was bias against duplications/insertions as most of associated CNVs were deletions. Also, a low count of associated genes (110) may be an underestimation of the impact of CNVs on gene expression, because CNVs involving long-range cis- and trans-regulatory elements were not investigated in this study.

Hulse and Cai recently reported a genome-wide screening for expression variability quantitative trait loci (evQTL) in humans and found that most cis-acting evQTL are located in CNV regions.⁷⁷ On the one hand, their results indicate that evQTL account for the mixture of positive and negative correlations between CNV and gene expression. On the other hand, their results could suggest that control of gene copy number is associated with reducing the variability in gene expression.⁷⁸ In human tumor cells, copy number changes of oncogenes and tumor suppressor genes may disrupt gene expression and stoichiometric relationships in cell metabolism and physiology, leading to cancer development and progression. It is estimated that ~60% of the genes show differential expression concordant to their copy number status, suggesting that the global correlation between CNV and gene expression is consistent across cancer studies.⁷⁹

Recent studies in nonhuman primates. CNV synteny sequences have been found in primates with 22% and 25% of chimpanzee and macaque CNVs overlapping human CNVs, suggesting the existence of CNV hotspots in primates.²¹,²² Iskow et al reported that regulatory element copy number differences can shape gene expression profiles in primates.⁸⁰ They discovered 964 copy number differences of conserved sequences across human, chimpanzee, and macaque and detected their impacts on species-specific gene expression profiles. Individual samples with copy number different genes had significantly different expression than samples with neutral copy number. Transcription factor genes and other regulatory elements (ultraconserved elements and long intergenic noncoding RNAs) differed in copy number and were associated with significant expression differences. Using NGS-based CNV maps, Gokcen et al showed that the lineage-specific formation rates of NAHR and Alu repeats resulted in markedly different CNV landscapes in chimpanzee, orangutan, and macaque.⁸¹ They further described several gene duplications, which led to evolutionary innovation through the gain of gene expression in new tissues.

Future Directions

Single-cell sequencing is generating new biological insights, due to the improved protocols for isolation of single cells and amplification of their genomes (DNA) or transcriptomes (RNA).⁸²–⁸⁴ Traditionally, pooling large quantity of cells from a tissue obscures the heterogeneity of complex systems. Single-cell sequencing helped to discover clonal mutations, cryptic cell types, or transcriptional features that would be diluted or averaged out in bulk tissue.⁸⁵–⁸⁷

To provide the proof of concept, genome-wide SNPs and CNVs were successfully detected by using a single human cell.⁸⁸ Using single-cell sequencing, McConnell et al further showed that mosaic CNV is abundant in human neurons.⁸⁹ Neurons taken from postmortem human frontal cortex tissue and neurons derived from induced pluripotent stem cell differentiation in vitro showed surprising diversity in individual cell genomes: 13–41% of neurons had at least one megabase-scale de novo CNV with twice more deletions than duplications. On the other hand, single neuron sequencing performed by Evrony et al suggested that LINE-1 retrotransposition is not a major generator of neuronal diversity in brain tissue.⁹⁰ It will be interesting to see how these CNV results are correlated with the strictly regulated expression of CNV genes in the brain as shown by the mouse embryo study.⁹¹ Therefore, more detailed RNA-Seq studies at single neuron level are warranted. Actually, single-cell RNA-Seq already revealed dynamic, random monoallelic gene expression in mouse cells, which increases the heterogeneity among cells and likely contributes to the phenotypic variance among individuals of identical genotype.⁹² Single-cell DNA methylome landscapes of mouse embryonic stem cells and early embryos were also analyzed using reduced representation bisulfite sequencing.⁹²
As single-cell amplification matures and sequencing cost drops, this approach will become a standard tool for understanding CNV and gene expression relationship at high resolution. Many of the mechanisms underlying the CNV-associated gene expression changes remain to be elucidated. Detailed investigations at single-cell level are warranted to shed light on how CNVs impact the expression of genes. When NGS cost reduces, more paired datasets will provide information on CNV and gene expression. This will remove probe bias and limitations of the microarray platforms and provide additional information on alternative splicing and miRNA expression. Whole genome sequence with NGS will provide additional information like SNPs, and balanced structural variations such as translocation and inversion. Furthermore, additional paired information such as epigenetic DNA methylation and histone modification will provide more detailed insights on gene expression.

Author Contributions
Conceived the concept: GEL. Analyzed the data: GEL. Wrote the first draft of the manuscript: GEL. Contributed to the writing of the manuscript: GEL, LX, KSH. Agree with manuscript results and conclusions: GEL, LX, KSH. Jointly developed the structure and arguments for the paper: GEL, LX, KSH. Made critical revisions and approved final version: GEL, LX, KSH. All authors reviewed and approved of the final manuscript.

REFERENCES
69. Heard E, Bickmore W. The ins and outs of gene regulation and chromosome
68. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for
67. Lee JA, Lupski JR. Genomic rearrangements and gene copy-number alterations
65. Gabellini D, Green MR, Tupler R. Inappropriate gene activation in FSHD: a
64. Bhatia S, Bengani H, Fish M, et al. Disruption of autoregulatory feedback by a
61. Schuster-Bockler B, Conrad D, Bateman A. Dosage sensitivity shapes the evolu
60. Beckmann JS, Estivill X, Antonarakis SE. Copy number variants and genetic
59. Beckmann JS, Estivill X, Antonarakis SE. Copy number variants and genetic
57. Chaignat E, Yahya-Graison EA, Henrichsen CN, et al. Copy number variation
55. Henrichsen CN, Vinckenbosch N, Zollner S, et al. Segmental copy number
49. Cook EH Jr, Scherer SW. Copy-number variations associated with neuropsychi
47. Cook EH Jr, Scherer SW. Copy-number variations associated with neuropsychi
45. Shi J, Levinson DF, Duan J, et al. Common variants on chromosome 6p22.1 are
44. Bochkukova EG, Huang N, Keogh J, et al. Large, rare chromosomal deletions
43. Strange BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and
copy number variation on gene expression phenotypes. Science. 2007;315(5813):
41. Henrichsen CN, Vinckenbosch N, Zollner S, et al. Segmental copy number
gene expression and metabolic traits in mice. Hum Mol Genet. 2007;16(18):
traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev
38. Schuster-Bockler B, Conrad D, Bateman A. Dosage sensitivity shapes the evolu
37. Chaignat E, Yahya-Graison EA, Henrichsen CN, et al. Copy number variation
35. Henrichsen CN, Vinckenbosch N, Zollner S, et al. Segmental copy number
34. Bochukova EG, Huang N, Keogh J, et al. Large, rare chromosomal deletions
33. Stefansson H, Ophoff RA, Steinberg S, et al. Common variants conferring risk
29. Cook EH Jr, Scherer SW. Copy-number variations associated with neuropsychi