Chapter 10

Population Genetic Nature of Copy Number Variation

Per Sjödin and Mattias Jakobsson

Abstract

Copy number variation has recently received considerable attention, and copy number variants (CNVs) have been shown to be both common in mammalian genomes and important for understanding genetic and phenotypic variation. As empirical knowledge and detection methods are quickly advancing, evolutionary theories about CNVs are rapidly updated and often revised. Here, we review recent progress on understanding CNVs, and we discuss some key issues for future research. In essence, we discuss four major forces in population genetics, recombination, mutation, selection, and demography, in relation to CNVs.

Key words: Copy number variation, Recombination, Mutation, Selection, Demography

1. Introduction

Large numbers of duplicated gene regions have been identified in genome sequences of several different organisms (1–5) and large numbers of differences have been detected in gene content between related species (6–10). This observation indicates that beneficial copy number variants (CNVs, see Table 1 for definitions) sometimes arise and increase in frequency, potentially via positive selection (12–17). However, CNVs also have the potential to disrupt genes, which suggests that many CNVs may be deleterious, preventing them to rise to high frequency (18–21). Recent studies have found that CNVs are polymorphic within the human population and that CNVs are widespread in the human genome (22–25). CNVs have also been associated with genetic diseases (24, 26–31) suggesting that CNVs may be causal for some common human diseases. Furthermore, it has been postulated that CNVs may represent a major genetic component underlying phenotypic variation (29), in addition to being a source of genetic variation among individuals and among human groups of different ethnic origin.

Lars Feuk (ed.), *Genomic Structural Variants: Methods and Protocols*, Methods in Molecular Biology, vol. 838,
DOI 10.1007/978-1-61779-507-7_10,

Table 1
Definitions adapted from Scherer et al. (11)

Structural variant	"… the umbrella term to encompass a group of genomic alterations involving segments of DNA typically larger than 1 kb…" "… may be quantitative (copy number variants comprising deletions, insertions, and duplications) and/or positional (translocations) or orientational (inversions)"
SD	"A segment of DNA >1 kb in size that occurs in two or more copies per haploid genome, with the different copies sharing >90% sequence identity"
Indel	"… collective abbreviation to describe relative gain or loss of a segment of one or more nucleotides …"
CNV	"… at least 1 kb in size …" "… genomic copy number gains (insertions or duplications) or losses (deletions or null genotypes) relative to a designated reference genome sequence"
CNP	"… a CNV that occurs in more than 1% of the population"
CNVR or a CNV locus	A CNV but also used to refer to "… a multiplex arrangement of variant units in close proximity, forming a CNV region"

Since CNVs are likely to be functionally important, they are also likely to be of evolutionary importance. In order to understand evolutionary changes, or the contribution of genetic variation to phenotypic variation, we need to consider CNVs in the light of population genetics. Here, we focus on four major population genetic forces: recombination, mutation, selection, and demography. It is well-known that these forces are not mutually independent. As an example, consider a new mutation in a particular region of a genome. This mutation would, according to some empirical studies, lower the probability of recombination occurring in heterozygote individuals. The recombination rate in the region decreases and selection is less efficient in this part of the genome due to the Hill–Robertson effect. These correlations are probably quite weak for mutations involving single nucleotides, but might be much stronger for mutations involving large stretches of DNA, such as CNVs.

2. Recombination

Since recombination events occur in pairs of chromosomes, it is particularly important to understand how a CNV that is heterozygote in an individual affects recombination. A heterozygote CNV will create a considerable length difference between the chromosomes, which in itself may have a large impact on the recombination process. Large sequence differences inhibit the homologous recombination process (32, 33) and it is therefore not surprising that

homologous recombination is inhibited by a heterozygous length difference (34–38). A specific case of homologous recombination, nonallelic homologous recombination (NAHR), may however be *more* frequent in heterozygote individuals than in homozygote individuals (39). A likely explanation for this effect is that the unpaired DNA loop in heterozygote individuals is free to pair with nonallelic loci during meiosis.

The presence of a CNV may also affect the relative probability that a recombination event results in a crossing-over event or a gene conversion event. In fact, a recent study (40) suggests that gene conversion may be less inhibited by heterozygous sequence differences than recombination resulting in crossing-over events. If this result extrapolates to heterozygous length differences, the effect of gene conversion on CNVs may be appreciable. In particular, similar to how the bias for G/C nucleotides in gene conversion tracts mimic positive selection for increased GC content (for a recent review, see ref. 41), a bias to retain either the short or the long allele would mimic selection for a change in genome size. The effect would be double: if the short allele is preferred, changes that make the genome smaller (deletions) would be promoted, while changes that enlarge the genome (insertions) would be disfavored (and vice versa if the bias is for the long allele). However, the support for gene conversion favoring the long, or short, allele in a heterozygote has not been unanimous: an early study showed a preference for the long allele (42), whereas a later study implied a more complicated scenario where length and structure of the indel also matter (43).

3. Mutation Process

Based on a rough estimate that each CNV affects at most 0.008% of the genome (1,447 CNVRs covering 12% of the human genome (24)), the infinite sites model (where every mutation hits a new place in the genome) could be a reasonable mutation model for CNVs. However, CNVs are frequently clustered into hotspots with high mutation rates and these CNV hotspots can potentially be hit by recurrent mutations (44, 45). As an illustration, by comparing CNVs in chimpanzees to known CNVs in humans, it was found that only 24 out of 355 CNVs (6.8%) were specific to chimpanzees (46, 47). In other words, 93.2% of CNVs found in chimpanzee were also polymorphic in humans—a much higher fraction of shared polymorphisms than expected between these species. This observation is best explained by recurrent mutations hitting specific regions in the primate genome, i.e., CNV hotspots. If CNVs are characterized by being confined to hypermutable hotspots, population genetic modeling of CNVs will need to incorporate these properties. For instance, it may lead to an inflation of derived

allele frequencies, which in turn can cause false inference of positive selection. In general, both the infinite allele model and infinite site model may be inappropriate for CNVs that may be better modeled by finite allele models such as the K-allele model (48) or the stepwise mutation model (49).

Several studies have found a striking enrichment of segmental duplications (SDs, Table 1) in CNV regions (24). The same genomic region may sometimes be identified as both a SD and a CNV, but because these two types of structural variation are commonly detected using different technologies and methods, it is often difficult to sort out their relationship, and the apparent enrichment of SDs in CNVs can be an artifact (24, 50). However, CNVs are also overrepresented in older SDs (51), indicating a more causal role for SDs in creating CNVs. It has been suggested that SDs predispose to NAHR events (24, 45, 47). NAHR events typically generate novel structural variation, and this would account for both CNV hotspots and the commonly observed genomic overlap between CNVs and SDs. Although SDs and CNVs are closely related concepts—SDs are basically fixed CNVs—SDs have been a major research focus in their own right and are known to be associated with interspecific and recurrent synteny breaks (see ref. 52 for a recent review). Interestingly, the chromosomal distribution of SDs in primates (including humans) seems to be the outlier among mammals: while SDs are typically organized in tandem (<1 Mb apart) in all investigated mammals (dog, cow, and mouse), the distance between SD copies in primates is much longer and more varied (53). The burst of Alu activity in the primate ancestor some 40 million years ago (54) is a strong candidate explanation as the genomic distribution of old SDs is strongly correlated with Alus while young SDs and CNVs are not (55) implying that SD/CNV formation is a dynamic process that changes over time.

If a distinction is made between the CNVs that overlap with SDs and those that do not, two distinct classes of CNVs emerge. The CNVs that do not overlap SDs typically do not belong to the copy number polymorphism (CNP) class (i.e., they segregate at frequencies below 1%, see Table 1), and they seem to be much less affected by recurrent mutations than CNVs that overlap SDs (Table 2). It appears that one class represents CNVs in hotspots, while the other class represents CNVs with low frequencies that behave like normal biallelic markers (51). As mentioned above, an individual that is heterozygote for a CNV may have an increased NAHR rate (39). This connection could provide a unifying mechanism for the two classes of CNVs since a CNV destined to become fixed, thereby creating a SD, is expected to be in a heterozygote state half of the time until fixation (56). As a consequence, a positive feedback loop will be initiated since the increased rate of NAHR will lead to an increase in the mutation rate for novel structural variation, which in turn acts as a stimulant for the NAHR

Table 2
Contrast between CNVs that do not overlap SDs and those that do

	CNVs outside SDs	CNVs overlapping SDs	References
Population frequency	–	+	(51)
Length	–	+	(21)
Gene-rich regions	(+)	+	(51)
Gene-poor regions	+	–	(51)
Synteny breaks	–	+	(21)
dN/dS	–	+	(21)
Environmental response genes	(+)	+	(51)
Disease genes	+	–	(21)
Alu	–	+	(55)

rate. A prediction, admittedly not easily tested, is that most hotspot CNVs should be initiated by the fixation of a duplication—not a deletion—since duplications necessarily provide the substrate for NAHR events which deletions do not.

The relative rate of NAHR-mediated duplications and deletions is also an important factor for understanding these hotspots. Relevant empirical data is however scarce as most studies report the number of duplications and deletions observed in the genome, which depends on both mutation rate and selection (see below). Although limited to male-specific NAHR and to a few CNV hotspots, Turner et al. (45) studied the de novo mutation rate directly and found that NAHR more often generated deletions than insertions.

Besides NAHR, nonhomologous end joining (NHEJ) has been suggested as a recombination-based mechanism for CNV formation. However, while recurrent CNV events are strongly associated with NAHR, CNVs due to NHEJ are rarely recurrent (50). Consequently, while the de novo mutation rate of CNVs due to NAHR has been estimated to be almost on the same order of magnitude as for microsatellites, CNV formations due to NHEJ probably have a much lower mutation rate similar to the rate estimated for SNPs (50). Retrotransposition is another major mechanism for CNV formation but in contrast to NAHR and NHEJ, it does not give rise to deletions, only to insertions of transcribed sequence. Finally, a novel replication-based mechanism, fork stalling and template switching (FoSTeS), has been proposed to account for CNVs that are difficult to explain by NAHR, NHEJ, or retrotranspositions. Together, these four mechanisms are believed to be responsible for the majority of CNVs (see ref. 31 for a review).

4. Selection

While the detrimental effect of an indel is likely to increase with its length (57), there is also accumulating evidence that deletions are more deleterious than insertions (58–62). For instance, CNV duplications are more common among CNVs with high frequency and CNV deletions are more common in CNVs with low frequency (19). Furthermore, a significantly lower proportion of CNV deletions than CNV duplications overlap with genes (24). This difference in selective constraint between deletions and insertions may be explained by the fact that deletions have two cut-points while insertions only have one: the probability that an insertion disrupts an important sequence motif does not depend on the length of the insertion, but the probability of disruption increases with the length of a deletion (58).

Genes involved in environmental response have been found to be overrepresented in CNVs in several organisms, including humans (23, 24, 63), mice (5), dogs (64), cows (53), and possibly also in fruit flies (16). A more detailed picture emerges when we separate CNVs that overlap SDs and those that do not. CNVs that do not overlap SDs show no (21), or considerably weaker (51), enrichment of environmental genes. These CNVs are instead overrepresented in gene-poor regions in contrast to CNVs that overlap SDs (51). This trend has been explained by positive selection for CNV changes in genes involved in environmental response (65, 66). Indeed, many of the differences listed in Table 10.2 between CNVs outside of SDs and those that overlap SDs are potentially explained by positive selection for CNVs/SDs in regions with a high density of genes involved in environmental response. However, some of these differences, such as the difference between frequency and length, could also be explained if CNVs that overlap SDs are more affected by recurrent events (this would inflate the frequency and also the detected length if several overlapping events are counted as one). Other differences could be secondary effects. CNVs may, for instance, be common in gene-rich regions as an effect of certain sequence motifs (e.g., non-B_DNA forming sequence), which are enriched in gene regions at the same time as they increase the rate of CNV formations (67). More conservative explanations are now being considered for the connection between genes involved in environmental response and CNVs: these genes may be enriched in CNVs not as a result of positive selection, but instead due to relaxation of selective constraint (21, 68).

Positive selection for CNVs is nonetheless likely to have played a significant role in at least some cases and one such example is described in a study by Perry et al. (15). In this study, the authors

showed that positive selection, driven by new starch-rich diets, increased the number of copies of the salivary amylase gene (*AMY1*) in some human populations. Finally, many arguments for why gene duplications may be favorable are also relevant for CNV duplications (69).

5. Demography

On the one hand, demographic history has a strong effect on patterns of genetic variation and CNVs are not likely to be an exception. On the other hand, the influence of demography on polymorphic CNVs (>1%, Table 1) is potentially overshadowed if CNVs are typically hypermutable and often affected by recurrent mutations (Table 2). However, comparisons of inferences of human population structure based on microsatellites and SNPs—two types of polymorphisms with very different mutational models and mutation rates—show great similarities between types of polymorphisms (70, 71). Despite an observed skew of CNVs toward rare alleles, potentially caused by purifying selection (19, 72, 73), it has been of interest to assess whether patterns of copy number variation across populations match the corresponding patterns for other types of loci, such as SNPs (24, 73, 74). In the extreme case of strong purifying selection, CNVs would have such low frequencies that they would not be shared even among closely related populations. Alternatively, if CNVs are generally unaffected by selection, they would display similar distributions across populations to the alleles of other types of neutral polymorphic loci. Discordance in inferences of human population structure based on CNVs and those based on SNPs would therefore suggest that most CNVs are sufficiently transient that they have not followed the same pattern of events (e.g., divergences, migrations, and admixture) in human history that typical neutral genetic variants have experienced. In contrast, agreement in CNV-based and SNP-based inferences would suggest that some CNVs are old enough to have migrated out of Africa with human founding populations, and that many other CNVs are old enough to be affected by recent human history of various divergences and migration events.

The greatest amount of genetic variation (based on mtDNA, microsatellites, and SNPs) has consistently been identified within African populations and variation outside of Africa has been shown to be a subset of the African diversity (18, 70, 75, 76). Most genome-wide population level investigations of CNVs have been conducted in the four HapMap populations, Yoruba from West Africa, European Americans, Han Chinese, and Japanese (18, 19,

24, 67, 68, 77–83), or in populations with similar ethnicities (84). Although the results for the HapMap populations are not always consistent, CNVs show patterns of variation similar to SNPs: higher diversity and more unique CNVs in African populations than in the European or East Asian populations (18, 79, 82, 84). Redon et al. (24) also showed that individuals can be correctly assigned to populations based on CNVs, demonstrating that CNV patterns of variation are at least partly shaped by human demographic history. However, some studies of CNVs that investigate samples from multiple non-HapMap populations failed to detect population stratification (72, 85) and conclude that there is limited evidence for population stratification of CNVs in geographically distinct human populations.

The CNV patterns across populations in a different and much larger data set, the HGDP panel of individuals (more than a thousand individuals representing more than fifty populations across the globe (86)) was recently investigated (73, 74, 87). By examining patterns of variation in CNVs and SNPs in the same individuals from the same populations, Jakobsson et al. (73) found that inferences of population structure based on CNVs largely accorded with those based on SNPs, but it was unclear how much ascertainment bias—exclusion of troublesome SNPs (potentially due to being located in a CNV region) from commercial SNP-typing arrays—affect the results by reducing CNV calls in the ascertained populations (73, 88). Examining the same populations, Itsara et al. (87) suggested that variation in genotyping intensity (due to variation in the experimental material) across the genome can produce false CNV calls, which may affect inferences of population stratification, and argue that there is limited evidence for stratification of CNVs in geographically distinct populations. However, in a recent study, Wang et al. (74) reanalyzed the CNV data of Jakobsson et al. (73) after excluding high-variance (in genotyping intensity) individuals from the analysis, and found that the similarity of SNP-based and CNV-based inference of population structure increases, especially when accounting for the smaller number of CNV loci studied compared to the number of SNPs. By using the data from the stringent CNV calls in Wang et al. (74), we compute the percentage of CNVs that are private to geographic regions. After correcting for sample size differences across geographic regions (89), we find a greater percentage of private CNVs in Africa, followed by Eurasia, and similar levels for East Asia, Oceania, and the Americas (Fig. 1). In summary, several studies (including the computation above) support the view that neither the rate of (recurrent) mutations nor the amount of purifying selection against CNVs has been great enough to erase the underlying signature of past human migrations from patterns of copy number variation across populations.

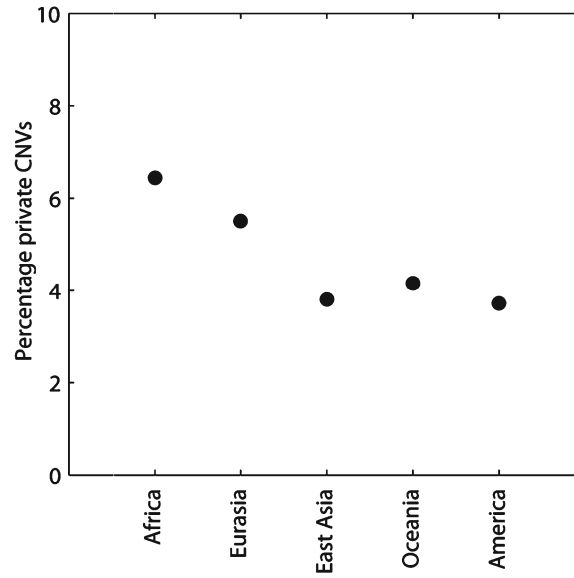

Fig. 1. Percentage of autosomal nonsingleton CNVs that is private to a particular geographic region. The remaining 76.36% of the CNVs are found in more than one region. The percentages correspond to a sample of ten individuals from each region and the CNV calls were performed using the PennCNV algorithm (79), and only including individuals with "standard deviation of the log *R* ratio" less than 0.22 (see ref. 74 for details).

6. Linkage disequilibrium and CNVs

Several studies report a severe lack of LD between CNVs and flanking SNPs (19, 24, 83), implying that searching for CNV-associated diseases may be a waste of effort (unless the disease causing CNV is included as a marker in the study). Redon et al. (24) considered three possible explanations for this lack of LD. First, some CNV duplications might represent transposition events that would generate linkage disequilibrium around the (unknown) acceptor locus but not the donor locus. Second, some CNVs might undergo recurrent mutations or reversions. Third, CNVs might occur preferentially in genomic regions with lower densities of SNP markers. These three hypotheses have all received subsequent support although it is clear that at least part of the explanation is the lack of SNPs in regions associated with CNVs and SDs (88). Redon et al. (24) argued that under their second explanation, CNV duplications should be in lower linkage disequilibrium with flanking SNPs than CNV deletions. Interestingly, although Redon et al. failed to detect a difference in LD between CNV duplications and CNV deletions, two subsequent studies showed that LD is indeed lower for duplications than for deletions (67, 90). In contrast, Schrider and Hahn (90) do not interpret this finding as evidence for the lack

of LD being caused by recurrent mutations, but instead propose that because the location of the duplicated copy is (sometimes) unknown, the relevant flanking SNPs are also unknown.

7. Future Directions

Many issues concerning CNVs will be resolved with upcoming advances in sequencing technology. At present, chromosomes are reconstructed from millions of short sequence reads but the upcoming "third generation" sequencing platforms will allow reading large parts of chromosomes in a single read (91, 92). This will resolve the problem of telling individual CNVs and SDs apart as well as determining the exact breakpoints for a CNV. It will also lead to a new perspective where the study of general structural variation will be more straightforward since inversions and translocations will be much easier to characterize. Finally, as small-scale "indels" and larger "structural variation" apparently share many features—such as the excess of de novo deletions compared to insertions, and deletions being more deleterious than insertions—both small-scale and large-scale structural mutations might be governed by the same population genetic properties. As hinted at in Conrad et al. (67), the different categories of structural variation should perhaps be studied as variation of a common theme, and for this purpose the upcoming sequencing technology is particularly promising.

Acknowledgments

We thank M. Lascoux for helpful comment on the manuscript. This work was supported by grants from Carl Trygger's foundation and by the Swedish Research Council Formas.

References

1. Lynch, M., and Conery, J.S. (2000) The evolutionary fate and consequences of duplicated genes *Science* **290**, 1151–5.
2. Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., Cherry, J.M., Henikoff, S., Skupski, M.P., Misra, S., Ashburner, M., Birney, E., Boguski, M.S., Brody, T., Brokstein, P., Celniker, S.E., Chervitz, S.A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R.F., Gelbart, W.M., George, R.A., Goldstein, L.S., Gong, F., Guan, P., Harris, N.L., Hay, B.A., Hoskins, R.A., Li, J., Li, Z., Hynes, R.O., Jones, S.J., Kuehl, P.M., Lemaitre, B., Littleton, J.T., Morrison, D.K., Mungall, C., O'Farrell, P.H., Pickeral, O.K., Shue, C., Vosshall, L.B., Zhang, J., Zhao, Q., Zheng, X.H., and Lewis, S. (2000) Comparative genomics of the eukaryotes *Science* **287**, 2204–15.
3. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome *Nature* **409**, 860–921.

4. Zhang, J. (2003) Evolution by gene duplication: an update *Trends Ecol Evol* **18**, 292–8.
5. She, X., Cheng, Z., Zöllner, S., Church, D.M., and Eichler, E.E. (2008) Mouse segmental duplication and copy number variation *Nat Genet* **40**, 909–14.
6. Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J.M., Beyne, E., Bleykasten, C., Boisramé, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J.M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G.F., Straub, M.L., Suleau, A., Swennen, D., Tekaia, F., Wésolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P., and Souciet, J.L. (2004) Genome evolution in yeasts *Nature* **430**, 35–44.
7. Parkinson, J., Mitreva, M., Whitton, C., Thomson, M., Daub, J., Martin, J., Schmid, R., Hall, N., Barrell, B., Waterston, R.H., McCarter, J.P., and Blaxter, M.L. (2004) A transcriptomic analysis of the phylum Nematoda *Nat Genet* **36**, 1259–67.
8. Demuth, J.P., De Bie, T., Stajich, J.E., Cristianini, N., and Hahn, M.W. (2006) The evolution of mammalian gene families *PLoS One* **1**, e85.
9. Hahn, M.W., Hanm M.V., and Han, S.G. (2007) Gene family evolution across 12 Drosophila genomes *PLoS Genet* **3**, 2135–46.
10. Opazo, J.C., Hoffmann, F.G., and Storz, J.F. (2008) Differential loss of embryonic globin genes during the radiation of placental mammals *Proc Natl Acad Sci USA* **105**, 12950–5.
11. Scherer, S.W., Lee, C., Birney, E., Altshuler, D.M., Eichler, E.E., Carter, N.P., Hurles, M.E., and Feuk, L. (2007) Challenges and standards in integrating surveys of structural variation *Nat Genet* **39**, S7–15.
12. Johnson, M.E., Viggiano, L., Bailey, J.A., Abdul-Rauf, M., Goodwin, G., Rocchi, M., and Eichler, E.E. (2001) Positive selection of a gene family during the emergence of humans and African apes *Nature* **413**, 514–19.
13. Nguyen, D.Q., Webber, C., and Ponting, C.P. (2006) Bias of selection on human copy-number variants *PLoS Genet* **2**, e20.
14. Heger, A., and Ponting, C. P. (2007) Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes *Genome Res* **17**, 1837–49.
15. Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., Carter, N.P., Lee, C., and Stone, A.C. (2007) Diet and the evolution of human amylase gene copy number variation *Nat Genet* **39**, 1256–60.
16. Emerson, J.J., Cardoso-Moreira, M., Borevitz, J.O., and Long, M. (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster *Science* **320**, 1629–31.
17. Xue, Y., Sun, D., Daly, A., Yang, F., Zhou, X., Zhao, M., Huang, N., Zerjal, T., Lee, C., Carter, N.P., Hurles, M.E., and Tyler-Smith, C. (2008) Adaptive evolution of UGT2B17 copy-number variation *Am J Hum Genet* **83**, 337–46.
18. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. (2006) A high-resolution survey of deletion polymorphism in the human genome *Nat Genet* **38**, 75–81.
19. Locke, D.P., Sharp, A.J., McCarroll, S.A., McGrath, S.D., Newman, T.L., Cheng, Z., Schwartz, S., Albertson, D.G., Pinkel, D., Altshuler, D.M, and Eichler, E.E. (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome *Am J Hum Genet* **79**, 275–90.
20. Dopman, E.B., and Hartl, D.L. (2007) A portrait of copy-number polymorphism in Drosophila melanogaster *Proc Natl Acad Sci USA* **104**, 19920–5.
21. Nguyen, D.Q., Webber, C., Hehir-Kwa, J., Pfundt, R., Veltman, J., and Ponting, C.P. (2008) Reduced purifying selection prevails over positive selection in human copy number variant evolution *Genome Res* **18**, 1711–23.
22. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004) Detection of large-scale variation in the human genome *Nat Genet* **36**, 949–51.
23. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T.C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004) Large-scale copy number polymorphism in the human genome *Science* **305**, 525–8.
24. 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., and Hurles, M.E. (2006) Global variation in copy number in the human genome *Nature* **444**, 444–54.

25. Wong, K.K., deLeeuw, R.J., Dosanjh, N.S., Kimm, L.R., Cheng, Z., Horsman, D.E., MacAulay, C., Ng, R.T., Brown, C.J., Eichler, E.E., and Lam, W.L. (2007) A comprehensive analysis of common copy-number variations in the human genome *Am J Hum Genet* **80**, 91–104.
26. Lupski, J.R., de Oca-Luna, R.M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B.J., Saucedo-Cardenas, O., Barker, D.F., Killian, J.M., Garcia, C.A., Chakravarti, A., and Patel, P.I. (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A *Cell* **66**, 219–32.
27. Singleton, A.B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R., Lincoln, S., Crawley, A., Hanson, M., Maraganore, D., Adler, C., Cookson, M.R., Muenter, M., Baptista, M., Miller, D., Blancato, J., Hardy, J., and Gwinn-Hardy, K. (2003) α-synuclein locus triplication causes Parkinson's disease *Science* **302**, 841.
28. 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., and Ahuja, S.K. (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility *Science* **307**, 1434–40.
29. Beckmann, J.S., Estivill, X., and Antonarakis, S.E. (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability *Nat Rev Genet* **8**, 639–46.
30. Mefford, H.C., Sharp, A.J., Baker, C., Itsara, A., Jiang, Z., Buysse, K., Huang, S., Maloney, V.K., Crolla, J.A., Baralle, D., Collins, A., Mercer, C., Norga, K., de Ravel, T., Devriendt, K., Bongers, E.M., de Leeuw, N., Reardon, W., Gimelli, S., Bena, F., Hennekam, R.C., Male, A., Gaunt, L., Clayton-Smith, J., Simonic, I., Park, S.M., Mehta, S.G., Nik-Zainal, S., Woods, C.G., Firth, H.V., Parkin, G., Fichera, M., Reitano, S., Lo Giudice, M., Li, K.E., Casuga, I., Broomer, A., Conrad, B., Schwerzmann, M., Räber, L., Gallati, S., Striano, P., Coppola, A., Tolmie, J.L., Tobias, E.S., Lilley, C., Armengol, L., Spysschaert, Y., Verloo, P., De Coene, A., Goossens, L., Mortier, G., Speleman, F., van Binsbergen, E., Nelen, M.R., Hochstenbach, R., Poot, M., Gallagher, L., Gill, M., McClellan, J., King, M.C., Regan, R., Skinner, C., Stevenson, R.E., Antonarakis, S.E., Chen, C., Estivill, X., Menten, B., Gimelli, G., Gribble, S., Schwartz, S., Sutcliffe, J.S., Walsh, T., Knight, S.J., Sebat, J., Romano, C., Schwartz, C.E., Veltman, J.A., de Vries, B.B., Vermeesch, J.R., Barber, J.C., Willatt, L., Tassabehji, M., and Eichler, E.E. (2008) Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes *N Engl J Med* **359**, 1685–99.
31. Zhang, F., Gu, W., Hurles, M.E., and Lupski, J.R. (2009) Copy number variation in human health, disease, and evolution *Annu Rev Genomics Hum Genet* **10**, 451–81.
32. McKim, K.S., Peters, K., and Rose, A.M. (1993) Two types of sites required for meiotic chromosome pairing in Caenorhabditis elegans *Genetics* **134**, 749–68.
33. Hammarlund, M., Davis, M.W., Nguyen, H., Dayton, D., and Jorgensen, E.M. (2005) Heterozygous insertions alter crossover distribution but allow crossover interference in Caenorhabditis elegans *Genetics* **171**, 1047–56.
34. Navarro, A., Betrán, E., Barbadilla, A., and Ruiz, A. (1997) Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes *Genetics* **146**, 695–709.
35. Shaw, C.J., and Lupski, J.R. (2004) Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease *Hum Mol Genet* **13**, R57–64.
36. Lupski, J.R., and Stankiewicz, P., (2005) Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes *PLoS Genet* **1**, e49.
37. Erdogan, F., Chen, W., Kirchhoff, M., Kalscheuer, V.M., Hultschig, C., Müller, I., Schulz, R., Menzel, C., Bryndorf, T., Ropers, H.H., and Ullmann, R. (2006) Impact of low copy repeats on the generation of balanced and unbalanced chromosomal aberrations in mental retardation *Cytogenet Genome Res* **115**, 247–53.
38. Lindsay, S.J., Khajavi, M., Lupski, J.R., and Hurles, M.E. (2006) A chromosomal rearrangement hotspot can be identified from

population genetic variation and is coincident with a hotspot for allelic recombination *Am J Hum Genet* **79**, 890–902.

39. Sun, X., Zhang, Y., Yang, S., Chen, J.Q., Hohn, B., and Tian, D. (2008) Insertion DNA Promotes Ectopic Recombination during Meiosis in Arabidopsis *Mol Biol Evol* **25**, 2079–83.
40. Welz-Voegele, C., and Jinks-Robertson, S. (2008) Sequence divergence impedes crossover more than noncrossover events during mitotic gap repair in yeast *Genetics* **179**, 1251–62.
41. Duret, L., and Galtier, N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes *Annu Rev Genomics Hum Genet* **10**, 285–311.
42. Lamb, B.C. (1985) The effects of mispair and nonpair correction in hybrid DNA on base ratios (G + C content) and total amounts of DNA *Mol Biol Evol* **2**, 175–88.
43. Bill, C.A., Taghian, D.G., Duran, W.A., and Nickoloff, J.A. (2001) Repair bias of large loop mismatches during recombination in mammalian cells depends on loop length and structure *Mutat Res* **485**, 255–65.
44. White, S.J., Vissers, L.E., Geurts van Kessel, A., de Menezes, R.X., Kalay, E., Lehesjoki, A.E., Giordano, P.C., van de Vosse, E., Breuning, M.H., Brunner, H.G., den Dunnen, J.T., and Veltman, J.A.(2007) Variation of CNV distribution in five different ethnic populations *Cytogenet Genome Res* **118**, 19–30.
45. Turner, D.J., Miretti, M., Rajan, D., Fiegler, H., Carter, N.P., Blayney, M.L., Beck, S., and Hurles, M.E. (2008) Germline rates of de novo meiotic deletions and duplications causing several genomic disorders *Nat Genet* **40**, 90–5.
46. Kehrer-Sawatski, H., and Cooper, D.N. (2008) Comparative analysis of copy number variation in primate genomes *Cytogenet Genome Res* **123**, 288–96.
47. Perry, G.H., Tchinda, J., McGrath, S.D., Zhang, J., Picker, S.R., Cáceres, A.M., Iafrate, A.J., Tyler-Smith, C., Scherer, S.W., Eichler, E.E., Stone, A.C., and Lee, C. (2006) Hotspots for copy number variation in chimpanzees and humans *Proc Natl Acad Sci USA* **103**, 8006–11.
48. Ewens, W. J. (2004) Mathematical Population Genetics. Second Revised Edition. Springer-Verlag, New York.
49. Ohta, T., and Kimura, M. (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population *Genet Res* **22**, 201–4.
50. Conrad, D.F., and Hurles, M.E. (2007) The population genetics of structural variation *Nat Genet* **39**, S30–6.
51. Cooper, G.M., Nickerson, D.A., and Eichler, E.E. (2007) Mutational and selective effects on copy-number variants in the human genome *Nat Genet* **39**, S22–9.
52. Marques-Bonet, T., Girirajan, S., and Eichler, E.E. (2009) The origins and impact of primate segmental duplications *Trends Genet* **25**, 443–54.
53. Liu, G.E., Ventura, M., Cellamare, A., Chen, L., Cheng, Z., Zhu, B., Li, C., Song, J., and Eichler, E.E. (2009) Analysis of recent segmental duplications in the bovine genome *BMC Genomics* **10**, 571.
54. Bailey, J.A., Liu, G.E., and Eichler, E.E. (2003) An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications *Am J Hum Genet* **73**, 823–34.
55. Kim, P.M., Lam, H.Y., Urban, A.E., Korbel, J.O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., and Gerstein, M.B. (2008) Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history *Genome Res* **18**, 1865–74.
56. Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R., Bergelson, J., and Chen, J.Q. (2008) Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes *Nature* **455**, 105–8.
57. Petrov, D.A., and Hartl, D.L. (2000) Pseudogene evolution and natural selection for a compact genome *J Heredity* **91**, 221–7.
58. Petrov, D.A. (2002) Mutational Equilibrium Model of Genome Size Evolution *Theor Pop Biol* **61**, 533–46.
59. Taylor, M.S., Ponting, C.P., and Copley, R.R. (2004) Occurrence and Consequences of Coding Sequence Insertions and Deletions in Mammalian Genomes *Genome Res* **14**, 555–66.
60. Taylor, M.S., Kai, C., Kawai, J., Carninci, P., Hayashizaki Y, and Semple, C.A. (2006) Heterotachy in mammalian promoter evolution *PLoS Genet* **2**, e30.
61. Kim, J., He, X., and Sinha, S. (2009) Evolution of Regulatory Sequences in 12 Drosophila Species *PLoS Genet* **5**, e1000330.
62. Sjödin, P., Bataillon, T., and Schierup, M.H. (2010) Insertion and deletion processes in recent human history *PLoS One* **5**, e8650.
63. Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D.,

Olson, M.V., and Eichler, E.E. (2005) Fine-scale structural variation of the human genome *Nat Genet* **37**, 727–32.

64. Nicholas, T.J., Cheng, Z., Ventura, M., Mealey, K., Eichler, E.E., and Akey, J.M. (2009) The genomic architecture of segmental duplications and associated copy number variants in dogs *Genome Res* **19**, 491–9.
65. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome *Nature* **420**, 520–62.
66. Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution *Nature* **428**, 493–521.
67. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C.H., Kristiansson, K., Macarthur, D.G., Macdonald, J.R., Onyiah, I., Pang, A.W., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J.; The Wellcome Trust Case Control Consortium, Tyler-Smith, C., Carter, N.P., Lee, C., Scherer, S.W., and Hurles, M.E. (2010) Origins and functional impact of copy number variation in the human genome *Nature* **464**, 704–12.
68. Perry, G.H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A.S., Hyland, C., Stone, A.C., Hurles, M.E., Tyler-Smith, C., Eichler, E.E., Carter, N.P., Lee, C., and Redon, R. (2008) Copy number variation and evolution in humans and chimpanzees *Genome Res* **18**, 1698–710.
69. Korbel, J.O., Kim, P.M., Chen, X., Urban, A.E., Weissman, S., Snyder, M., and Gerstein, M.B. (2008) The current excitement about copy-number variation: how it relates to gene duplications and protein families *Curr Opin Struct Biol* **18**, 366–74.
70. Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K., and Feldman, M.W. (2005) Clines, clusters, and the effect of study design on the inference of human population structure *PLoS Genet* **1**, e70.
71. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008) Worldwide human relationships inferred from genome-wide patterns of variation *Science* **319**, 1100–4.
72. Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., Oseroff, V.V., Albertson, D.G., Pinkel, D., and Eichler, E.E. (2005) Segmental duplications and copy-number variation in the human genome *Am J Hum Genet* **77**, 78–88.
73. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., Bras, J.M., Schymick, J.C., Hernandez, D.G., Traynor, B.J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H.M., Hardy, J.A., Rosenberg, N.A., and Singleton, A.B. (2008) Genotype, haplotype, and copy-number variation in worldwide human populations *Nature* **451**, 998–1003.
74. Wang, C., Szpiech, Z.A., Degnan, J.H., Jakobsson, M., Pemberton, T.J., Hardy, J.A., Singleton, A. B., and Rosenberg, N.A. (2010) Comparing Spatial Maps of Human Population-Genetic Variation Using Procrustes Analysis *Stat Appl in Genet and Mol Biol* **9**, Article 13.
75. Cann, H.M., Cohen, D., and Dausset, J. (1987) Diagnosis of genetic disease by linkage analysis *Birth Defects Orig Artic Ser* **23**, 33–60.
76. Garrigan, D., and Hammer, M.F. (2006) Reconstructing human origins in the genomic era *Nat Rev Genet* **7**, 669–80.
77. The International HapMap Consortium (2005) A haplotype map of the human genome *Nature* **437**, 1299–320.
78. McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J., Altshuler, D.M., and International HapMap Consortium (2006) Common deletion polymorphisms in the human genome *Nat Genet* **38**, 86–92.
79. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H., and Bucan, M. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data *Genome Res* **17**, 1665–74.
80. Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N.A., Tsang, P., Newman, T.L., Tüzün, E., Cheng, Z., Ebling, H.M., Tusneem, N., David, R., Gillett, W., Phelps, K.A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J.D., Korn, J.M., McCarroll, S.A., Altshuler, D.A., Peiffer, D.A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D.A., Mullikin, J.C., Wilson, R.K., Bruhn, L., Olson, M.V., Kaul, R., Smith, D.R., and Eichler, E.E. (2008) Mapping and sequencing of structural variation from eight human genomes *Nature* **453**, 56–64.

81. Armengol, L., Villatoro, S., González, J.R., Pantano, L., García-Aragonés, M., Rabionet, R., Cáceres, M., and Estivill, X. (2009) Identification of copy number variants defining genomic differences among major human groups *PLoS One* **4**, e7230.
82. Takahashi, N., Satoh, Y., Kodaira, M., and Katayama, H. (2008) Large-scale copy number variants (CNVs) detected in different ethnic human populations *Cytogenet Genome Res* **123**, 224–33.
83. Kato, M., Kawaguchi, T., Ishikawa, S., Umeda, T., Nakamichi, R., Shapero, M.H., Jones, K.W., Nakamura, Y., Aburatani, H., and Tsunoda, T. (2010) Population-genetic nature of copy number variations in the human genome *Hum Mol Genet* **19**, 761–73.
84. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X., and Frazer, K.A. (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome *Nat Genet* **38**, 82–5.
85. de Ståhl, T.D., Sandgren, J., Piotrowski, A., Nord, H., Andersson, R., Menzel, U., Bogdan, A., Thuresson, A.C., Poplawski, A., von Tell, D., Hansson, C.M., Elshafie, A.I., Elghazali, G., Imreh, S., Nordenskjöld, M., Upadhyaya, M., Komorowski, J., Bruder, C.E., and Dumanski, J.P. (2008) Profiling of copy number variations (CNVs) in healthy individuals from three ethnic groups using a human genome 32 K BAC-clone-based array *Hum Mutat* **29**, 398–408.
86. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G.B., Friedlaender, J.S., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R.J., Huang, X., Kidd, J., Kidd, K.K., Langaney, A., Lin, A.A., Mehdi, S.Q., Parham, P., Piazza, A., Pistillo, M.P., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J.L., Greely, H.T., Feldman, M.W., Thomas, G., Dausset, J., and Cavalli-Sforza LL. (2002) A human genome diversity cell line panel *Science* **296**, 261–2.
87. Itsara, A., Cooper, G.M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R.M., Myers, R.M., Ridker, P.M., Chasman, D.I., Mefford, H., Ying, P., Nickerson, D.A., and Eichler, E.E. (2009) Population analysis of large copy number variants and hotspots of human genetic disease *Am J Hum Genet* **84**, 148–61.
88. McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A., Elliott, A.L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P.J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K.W., Rava, R., Daly, M.J., Gabriel, S.B., and Altshuler, D. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation *Nat Genet* **40**, 1166–74.
89. Kalinowski, S.T. (2004) Counting alleles with rarefaction: private alleles and hierarchical sampling designs *Conserv Genet* **5**, 539–43.
90. Schrider, D.R., and Hahn, M.W. (2010) Lower linkage disequilibrium at CNVs is due to both recurrent mutation and transposing duplications *Mol Biol Evol* **27**, 103–11.
91. Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S.B., Krstic, P.S., Lindsay, S., Ling, X.S., Mastrangelo, C.H., Meller, A., Oliver, J.S., Pershin, Y.V., Ramsey, J.M., Riehn, R., Soni, G.V., Tabard-Cossa, V., Wanunu, M., Wiggin, M., and Schloss, J.A. (2008) The potential and challenges of nanopore sequencing *Nat Biotechnol* **26**, 1146–53.
92. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009) Real-time DNA sequencing from single polymerase molecules *Science* **323**, 133–8.