Chapter 11

Detection and Interpretation of Genomic Structural Variation in Mammals

Ira M. Hall and Aaron R. Quinlan

Abstract

Structural variation (SV) encompasses diverse types of genomic variants including deletions, duplications, inversions, transpositions, translocations, and complex rearrangements, and is now recognized to be an abundant class of genetic variation in mammals. Different individuals, or strains, of a given species can differ by thousands of variants. However, despite a large number of studies over the past decade and impressive progress on many fronts, there remain significant gaps in our knowledge, particularly in species other than human. Arguably the most relevant among these are genetically tractable models such as mouse, rat, and dog. The emergence of efficient and affordable DNA sequencing technologies presents an opportunity to make rapid progress toward understanding the nature, origin, and function of SV in these, and other, domesticated species. Here, we summarize the current state of knowledge of SV in mammals, with a focus on the similarities and differences between domesticated species and human. We then present methods to identify SV breakpoints from next-generation sequence (NGS) data by paired-end mapping, split-read mapping, and local assembly, and discuss challenges that arise when interpreting these data in lineages with complex breeding histories and incomplete reference genomes. We further describe technical modifications that allow for identification of variants involving repetitive DNA elements such as transposons and segmental duplications. Finally, we explore a few of the key biological insights that can be gained by applying NGS methods to model organisms.

Key words: Structural variation, Copy number variation, Mammals, Model systems, Paired-end mapping, Split-read mapping, Breakpoint assembly, Mutation mechanism, Next-generation sequencing, Genomic rearrangements

1. Introduction

Genomic differences underlie the vast majority of heritable phenotypic differences and provide the raw material for evolution. They come in a broad range of shapes and sizes, from single-nucleotide polymorphisms (SNPs) to chromosomal rearrangements involving many megabases of DNA. As a rule, our appreciation for the different

Lars Feuk (ed.), *Genomic Structural Variants: Methods and Protocols*, Methods in Molecular Biology, vol. 838,
DOI 10.1007/978-1-61779-507-7_11, © Springer Science+Business Media, LLC 2012

classes of genetic variation has been directly linked to our ability to detect them, and most conceptual advances have had their roots in technological advances. For most of the twentieth century geneticists focused on large-scale genomic rearrangements because these were the only variants that were visible by early cytogenetic methods. However, given their large size (>3 Mb) and frequent association with either cancer, sporadic disease, or long periods of genome evolution, genomic rearrangements were generally thought to be rare within species. The discovery of the structure of DNA and the elaboration of the genetic code spawned great interest in the role of small-scale variants such as SNPs, and the development of molecular cloning, DNA sequencing, and PCR during the 1970s and 1980s allowed for researchers to directly ascertain these variants in an increasingly high-throughput manner. By the turn of the century, as thousands of automated DNA sequencers were churning out data for the Human Genome Project, conventional wisdom held that the overall structure of a genome was relatively static, and that most natural variation could be explained by small-scale sequence differences. This is best exemplified by an oft-repeated statement of this era that the phenotypic differences between two humans could be explained by the presence of 1 SNP in every 1,000 bp, and the differences between a human and chimpanzee by 1 in 100.

The first convincing evidence that genome architecture was more dynamic came from the work of E. Eichler and colleagues who, upon aligning the draft human genome to itself, discovered that roughly 5% of DNA is contained within recent segmental duplications (SDs) (1). SDs are defined as multicopy segments larger than 1 kb that share greater than 90% pairwise nucleotide identity. The prevalence of highly similar duplications indicated that the human genome had undergone substantial large-scale structural mutation over recent evolutionary time. Similar levels of segmental duplication have been reported in all mammalian genomes sequenced thus far including various primates, mouse, rat, dog, and cow (2–6).

The availability of a reference genome sequence enabled the construction of genome-wide microarrays containing bacterial artificial chromosomes (BACs) or oligonucleotide probes, and comparative genomic hybridization to these arrays [array comparative genome hybridization (aCGH)] reveals DNA copy number variation (CNV) between two genomes. The first two studies to apply this technology to "normal" human individuals discovered surprising levels of CNV (7, 8), and this observation led to a profound conceptual shift in our view of genome variation (9). While the first studies detected merely a handful of CNVs in pairwise comparisons, over the next 6 years a large number of genome-wide mapping studies using progressively higher resolution oligonucleotide microarrays, fosmid end-sequencing, and massively parallel paired-end sequencing have reported several hundred to several thousand variants in pairwise comparisons between humans, depending on the resolution and scope of the methods employed. Thus, structural variation (SV) is an abundant class of genetic variation in mammals. We operationally define SV as differences in the copy number, orientation, or location of genomic segments exceeding 100 bp in size. This definition encompasses diverse types of genomic variants including deletions, duplications, inversions, transpositions, translocations, and complex rearrangements. SVs affect many genes and by some measures a larger fraction of the genome than SNPs, and the extent to which these differences underlie phenotypic variation and disease is currently an active line of investigation in many laboratories around the world.

There have been numerous studies and many important findings in this field over the past decade. There are also many unresolved questions, particularly in species other than human. Arguably the greatest challenge in answering these questions is methodological; accurate and unbiased interpretations depend upon accurate and unbiased SV detection, and this technical goal has not yet been achieved. Now, next-generation sequencing technologies offer the unprecedented opportunity to, at least in theory, map virtually all classes of SV at extremely high resolution and at reasonable cost. It should be possible to make rapid progress toward a better understanding of SV in all mammals. Unfortunately, interpretation of next-generation sequence (NGS) data in complex and repetitive genomes presents a number of nontrivial computational difficulties, and these difficulties can be exacerbated in experiments involving organisms that are not as well characterized as human.

In this review, we focus on detection and interpretation of SV in model mammalian species such as mouse, rat, and dog. To provide context, we first summarize the current state of knowledge of SV in mammals, with a focus on the similarities and differences between domesticated species and human. We then present methods to identify SV breakpoints from NGS data, and discuss specific challenges that can arise when interpreting these data in lineages with complex breeding histories and incomplete reference genomes. Finally, we explore a few of the biological insights that can be gained by applying NGS methods to model systems.

1.1. Abundance of Genomic SV in Domesticated Mammals

Over the past 6 years, there have been 11 SV mapping studies in mouse (3, 10–19), one in rat (20), two in dog (6, 21) and one in cow (22). All of these except two mouse studies (17, 18) have used aCGH, and thus were only able to detect relatively large CNVs. One common theme emerging from these studies is that all mammalian species examined thus far display abundant SV in their genomes, and appear to have roughly similar overall levels. For example, the first aCGH studies in mouse were published shortly after the first human studies (10, 11), and these studies discovered roughly similar numbers of CNVs. Subsequent studies using progressively higher resolution arrays and genetically diverse panels of inbred mouse strains (3, 12–16) identified many thousands of CNVs. In one study alone, Cutler et al. identified 2,094 CNVs among 41 inbred strains (13). The single highest resolution aCGH study to date (15) identified ~300 CNVs between two "classical" inbred strains, which is consistent with the most comprehensive aCGH study in human (23) if resolution differences are taken into account. Genome-wide experiments using a common platform (~385,000-probe NimbleGen arrays) in mouse (14), rat (20), dog (21), and cow (22) discovered an average of 10–20 CNVs per sample. Similar to previous findings in human (24, 25), aCGH experiments in mouse (3) and dog (6) using microarrays targeting segmental duplications discovered remarkable levels of CNV in these dynamic regions of genome.

The only comprehensive NGS-based study in mouse (our own) discovered 7,196 SVs in a single strain comparison (18), which is a remarkable level of variation and substantially more than has been reported in human studies if resolution differences are taken into account (26–30). However, this high level of SV is mostly accounted for by transposable element variants (TEVs), which comprise ~70% of all SV. This result is not entirely surprising given that retroelements are known to be very active in mouse (17, 31). Independent of the abundance of transposons, a direct comparison of SV levels measured by NGS is complicated by the fact that vastly different SV detection algorithms have been employed by different studies; however, at a first approximation the levels of non-TE variation appear to be similar.

While these experiments demonstrate that overall levels of SV are roughly similar, there are a few important caveats. First, direct comparisons between species are complicated by methodological differences. It remains a possibility that important differences in SV levels will be found once reference genome assemblies are completed to an equal level of accuracy and all species are examined on a common SV discovery platform, such as genome-wide sequencing using a single analysis pipeline. Second, in contrast to humans, domesticated species have been subjected to artificial selection and directed breeding, and as a consequence different strains, or breeds (hereafter referred to as strains), within a species may exhibit vastly different levels of genetic relatedness. For example, comparison of two mouse strains that are closely related by their breeding history will yield far fewer SVs than a comparison between strains that were generated by different founding stock. In addition, experiments involving wild populations or geographically isolated subspecies will detect substantially more variation than those involving inbred lines. Finally, the relative abundance of different SV classes may vary between different species, and thus the relative levels of SV that are detected may differ depending on the experimental platform that is used. For example, as mentioned above, retrotransposons are especially active in rodents and thus genome-wide sequencing experiments that are able to detects variable TE insertions are likely to find substantially more SV in rodents than similar experiments in human or dog.

2. Genomic Distribution of SV

A true comparison of the genomic distribution of SV is confounded by the fact that studies have rarely mapped the physical location of variable genomic segments, and thus we must rely upon the location of the affected segment in the reference genome and upon analyses of recent segmental duplications (which are often either SVs themselves or hotspots for SV). Nevertheless, these are useful proxies and much has been learned.

2.1. Correlation with Segmental Duplications

The first notable observation is that in all species examined thus far structural variants are highly enriched at sites of segmental duplication in the reference genome. Typical estimates range from four- to tenfold enrichment relative to a random model (14, 32–34). The breakpoints of large-scale rearrangements that occurred during mammalian evolution are also highly correlated with SDs (35–37). This enrichment has generally been explained by the propensity of duplicated sequences to promote nonallelic homologous recombination (NAHR), and there are many clear examples of this among the de novo SVs that cause human genomic disorders (38). However, there are suggestions from the literature (39–45) and direct evidence from our own work (18) that homology-independent mechanisms also preferentially occur in SDs. The cause of this is not presently clear and there are various possibilities, but regardless of mechanism the preferential localization of SVs in segmentally duplicated genomic regions is a common theme. Since segmental duplications have a markedly nonuniform genomic distribution (46), patterns of structural variation are also nonuniform and differ substantially from other classes of variation such as SNPs.

2.2. Distribution of Segmental Duplications and Associated SV

An important consideration is that patterns of SD are very different between primates and other mammalian species. In all species, most SDs are intrachromosomal, but in primates the majority of SDs are interspersed, with only ~30% being present in a “tandem” configuration (<1 mb apart) (2). In contrast, SDs in other mammalian genomes are primarily tandem. For example, in the highest-quality nonprimate reference genome, the mouse, ~90% of SDs are present in local clusters of tandem duplications (3, 47). The observation that a tandem configuration predominates in the genomes of diverse mammals including mouse (3), rat (5), dog (6), and cow (4) indicates that this is the ancestral state and that the primate lineage underwent a marked shift (2). It is not known whether the different patterns are due to differences in mutational mechanism or selection, but these differences inevitably affect the distribution of SV. Moreover, SDs present in a tandem configuration may have higher rates of spontaneous mutation, since the frequency of NAHR depends upon the proximity of participating duplicons. As suggested by She et al. (3), this effect might partially explain the extremely high mutation rates documented at certain SD loci in our own previous study of spontaneous SV among closely related mouse strains (19).

2.3. Transposable Elements

Another important difference is the relative activity of different transposon classes. Overall transposon activity is much greater in rodents relative to human and dog (48) and thus TEV comprises a greater fraction of overall SV. For example, while aCGH experiments suggest similar levels of large-scale CNV (see above), DNA sequence-based studies suggest that there are at least 5,000 TEVs between two inbred mouse strains (such as DBA/2J and C57BL/6J) (18), and less than 1,000 between two humans (49). Moreover, there are differences in the relative activity of different element classes. Whereas LINE and LTR transposons comprise the vast majority of TEV in mouse, human TEV is dominated by Alu-SINE elements (49). Differences in transposon activity do not merely affect TEV, but can also lead to other forms of SV. For example, LINE element machinery can cause retrotransposition of host transcripts leading to retrogenes, and indeed there are hundreds of variable retrogenes between inbred mouse strains (18). LINE elements can also cause double-stranded breaks even at sites where insertion does not occur (50), and inaccurate repair of DNA breaks can contribute to new SV. Because LINE elements are much more active in rodents than human or dog, these mechanisms may potentially lead to distinct patterns of SV among species. Moreover, while all TE classes can generate new SV through NAHR, the genomic distribution of various TE classes may differ. For example, Alu-SINE elements are enriched at boundaries of segmental duplications in the human genome (51), suggesting that NAHR between these elements has been a major force driving duplications. In contrast, there is an enrichment of LTR and LINE elements, but not SINEs, at mouse SDs (3). Given that SINEs preferentially insert into GC-rich regions (which are gene-rich) and LINEs and LTRs preferentially insert into AT-rich genomic regions (which are gene-poor), the relative frequency of NAHR between different TE classes can lead to significant differences in the genomic distribution of SV (3).

2.4. Complex Variants

An unresolved question concerns the prevalence of complex variants, defined as SVs that appear to have arisen from a single mutational event yet contain multiple adjacent breakpoints in proximity (e.g., <1 kb apart). Studies of complex disease-causing rearrangements at several loci in the human genome led J. Lupski, P. Hastings, and colleagues to propose a new SV formation mechanism, termed FoSTeS (52) or MMBIR (53), that involves template switching during DNA replication and/or repair. We hereafter refer to this mechanism by the more general term "template switching." We have shown that complex rearrangements involving multiple adjacent breakpoints are common in the mouse genome and account for ~16% of all SVs. It remains to be seen whether genome-wide studies in human (or other species) will find a similar level of complex variation, or whether the mouse genome is particularly prone to this mechanism. We expect that ongoing analysis of 1000 Genomes Project data (http://www.1000genomes.org/) will help resolve this question. Complex variants can have a strong and somewhat misleading effect on the genomic distribution of SV since complex variants manifest as local clusters of SV calls, and these can only be disentangled into a single variant call in datasets with subkilobase resolution.

3. Breeding Effects

Patterns of variation are strongly influenced by breeding history and selection, and in this respect model genetic systems differ substantially from human. Artificial selection has fixed genetic variants underlying desired traits (and linked variants via hitchhiking), and inbreeding has "flushed" strongly deleterious variants from inbred lineages. The extent to which these effects may be apparent is dependent on the genetic composition of the founding stock, the precise breeding history, and the nature and strength of artificial selection. These variables may differ considerably among different species and strains. For example, laboratory mouse strains have their origin in "fancy" mice, which were derived at least several hundred years ago by interbreeding of divergent subspecies. After hundreds of years of selective breeding, scientists at the turn of the century derived a relatively large number of inbred strains from a relatively small pool of genetically diverse progenitor fancy mice (54). This breeding history gives the mouse genome a unique composition, such that each genome is a mosaic of segments with different subspecific origins. Thus, in pairwise comparisons between strains whose genomes have different mosaic patterns, regions of the genome with a different subspecific origin show very high levels of variation and regions of the genome that are identical by descent (IBD) show very low levels of variation (55). Similar patterns exist in rat (56) and to a lesser extent in dog (48). As we discuss later, these patterns can complicate SV detection. In addition to these effects, regions of the genome harboring genes that have been selected for during domestication may show little or no SV among all strains of a given species.

4. SV Discovery by Sequencing

While the studies conducted to date illustrate that, in general, the landscape of SV in domesticated mammalian species is similar to that in primates, fundamental questions regarding SV frequency, size, genomic distribution, mechanistic causes, and phenotypic impact remain unanswered. In large part, these questions persist because of the inherent limitations of aCGH for studying this class of variation. The resolution of aCGH methods is typically limited to 10–100 kb, and aCGH is blind to balanced rearrangements such as inversions and reciprocal translocations. Moreover, aCGH has limited sensitivity to detect lesions arising from repetitive sequences such as segmental duplications and transposable elements. Given the relatively high activity of TEs in rodent genomes (17, 18, 31), and the fact that segmental duplications are hotspots for SV, the inability to screen for mutations in duplicated sequence precludes the detection of a substantial and functionally relevant portion of structural variation (18, 29).

The recent proliferation of accurate, high-throughput DNA sequencing techniques eliminates many of the biases inherent to microarrays and provides a powerful and economical approach for genome-wide SV characterization. Current sequencing techniques can localize SV breakpoints much more precisely and, given sufficient sequence depth and/or read length, allow one to infer the causal mechanism of SV formation by characterizing the nucleotide sequences flanking SV breakpoints (discussed in more detail below). The specific sequencing methods used by the different available technologies are diverse, yet with respect to the detection of SV they can be broadly classified into two categories: shorter, paired-end sequences and longer, contiguous sequences. In the following two sections, we discuss the merits and weaknesses of the two molecular approaches in the context of SV discovery and characterization.

4.1. SV Discovery with Paired-End Mapping

Recently, substantial focus has been placed on the development of computational methods to exploit so-called paired-end sequences for the discovery of diverse classes of SV. The fundamental principles of this sequencing approach have been described in detail elsewhere (57–59); the basic premise is that the two respective ends (hereafter referred to as "matepairs," or "pairs") of millions of larger DNA molecules are sequenced from an experimental (or "test") genome and compared to a reference genome. Prior to sequencing, DNA fragments from the test genome are carefully restricted to a predictable size range (e.g., 500 bp). Paired-end mapping (PEM) approaches proceed by aligning the sequenced matepairs to a reference genome, and use the expected size distribution and orientation of the pairs to infer whether the structure of the test genome agrees with that of the reference. So-called concordant matepairs align with the expected distance and orientation and indicate that the structure of the test genome agrees with that of the reference genome. The corollary is that "discordant" matepairs, which align with an unexpected distance and/or orientation, suggest possible structural variation between the test and reference genomes.

Each class of structural variation (e.g., deletion, insertion, inversion, etc.) has a characteristic discordant mapping "signature" (Fig. 1). PEM approaches must carefully detect and exclude concordant matepairs so that putative SV can be confidently identified from the remaining discordant pairs. In order to rule out remaining alignment artifacts and chimeric molecules, most SV discovery algorithms screen for multiple discordant matepairs (typically two or more) that have the same signature and support

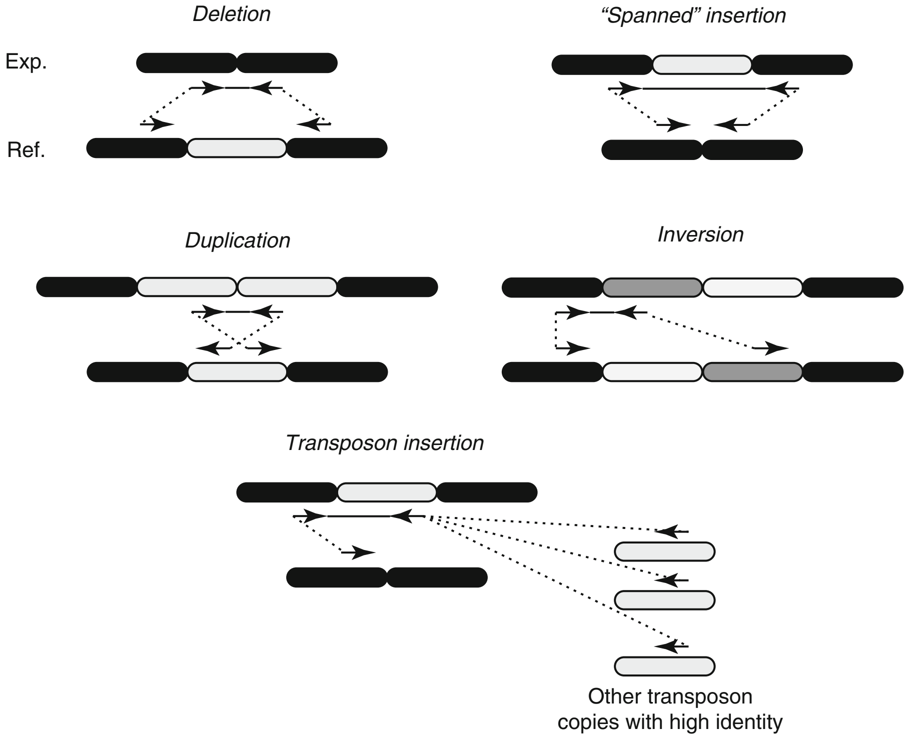

Fig. 1. Paired-end mapping signatures. Shown are five different classes of structural variation and the matepair mapping patterns that result. In each case, the experimental (Exp.) genome is shown on top and the reference genome (Ref.) below. The genomic segment affected by the SV is shown in the reference genome in *gray* and unaffected segments in *black*. Each matepair derived from the experimental genome has a known size (e.g., 500 bp) and the two respective reads are in opposing orientation (*black arrows*), as shown beneath the experimental genome. Structural variation is apparent when matepairs map to the reference genome with an unpredicted size and/or orientation. *Dotted lines* connect the actual matepair sequences, obtained from the experimental genome, to the reference genome. The orientation of the alignment in the reference genome is indicated by the direction of the *arrows*. Note that each SV class gives a distinctive pattern. Note also that the read mapping to the transposon insertion in the bottom-most example will also map to all other similar copies of that transposon class (three copies are shown).

the same SV breakpoint. A complementary approach to PEM is to use the depth of sequence coverage (DOC) from concordant matepairs to detect duplications and deletions in the test genome (18, 58, 60–62). DOC approaches are analogous to aCGH and are simplified in species such as mouse where the genome of a given strain is largely homozygous.

4.2. Detection of Repeats

Mammalian genomes are highly repetitive and a significant fraction of SV involves multicopy elements such as transposons and segmental duplications. Accurate detection of SV at repetitive loci requires certain technical modifications. For example, when a rearrangement occurs within a segmental duplication, the discordant matepairs that indicate the mutation will align to the locus where the mutation occurred, as well as to all other similar copies of the repeated sequence. Similarly, when a mobile element transposes to an otherwise nonrepetitive location in a test genome, one end of the discordant matepairs will align uniquely to the region flanking the insertion site while the end sequenced from the mobile element insertion will align to all other similar mobile elements in the reference genome (Fig. 1). The identification of such variants necessitates sensitive read alignment such that many mappings (i.e., hundreds or thousands) are reported for matepairs derived from repetitive elements. In addition, SV discovery algorithms must be able to cluster discordant mappings such that only a single mapping for a given matepair is included in a single variant call. At the time of writing two published algorithms are capable of this (18, 29); however, we expect that other extant algorithms will be extended accordingly.

4.3. Technical Challenges Affecting Accurate SV Discovery by PEM

In our research, we have found that false-positive variant calls primarily arise from three sources: (1) insufficiently sensitive sequence alignment; (2) incomplete removal of sequencing artifacts; and (3) matepairs originating from poorly assembled genomic regions.

4.3.1. Sequence Alignment

PEM using NGS data generally relies upon relatively short reads (<100 bp), and the error rate increases steadily as longer sequences are obtained. As a result, the already difficult task of accurately aligning millions of short sequences to a repetitive genome is exacerbated by the need to account for errors and polymorphism in the aligned sequences. The most typical misalignment artifact arises when the paired sequences have sufficient polymorphisms and/or errors to prevent the aligner from detecting the proper alignment(s) which would indicate that the matepair is concordant with the reference genome. Instead, the only alignment(s) detected will erroneously suggest that the matepair is discordant. When this error occurs in a systematic fashion at specific genomic loci (as in our experience it often does), multiple erroneously aligned matepairs will cluster at each of these loci and be identified by the PEM-detection algorithm, resulting in a substantially elevated false-positive SV discovery rate.

This problem is especially pernicious given the breeding history domesticated mammals. Because of this history, both SNPs and SVs are nonrandomly distributed throughout the genome, and the variation observed between any two lines will largely mimic pairwise differences in haplotype structure. In the laboratory mouse, for example, haplotypes of different subspecific origin can harbor extremely high rates of polymorphism (e.g., ~1 per 200 bp), while haplotypes with the same subspecific origin generally have extremely low rates (e.g., ~1 per 10 kb) (55, 63). Between two classical inbred strains divergent haplotypes comprise roughly one-third of the genome (63). Thus, genomic regions where the test genome and the reference genome have different subspecific origins will be greatly enriched in problematic alignments. These regional effects can confound data interpretation because they lead to very different false discovery rates in different parts of the genome, and this effect varies depending on the strains that are compared. Since the rates of polymorphism may be as much as fivefold higher between divergent haplotypes in domesticated species than among human individuals, more sensitive sequence alignment is required in these species to achieve similar levels of accuracy.

We have found that an effective approach to mitigating alignment artifacts is to use a tiered, increasingly sensitive alignment scheme (18). Such an approach begins with a fast, yet less sensitive aligner such as BWA (64) to quickly identify the majority of the easily identified concordant matepairs. The remaining discordant matepairs are iteratively scrutinized with successively more sensitive aligners (e.g., Novoalign (65) or Mosaik (66)) and settings until one is confident that only truly discordant matepairs remain.

4.3.2. Sequencing Artifacts

Additional complications to PEM approaches are caused by experimental artifacts that arise during DNA library construction and sequencing. The most common and problematic of these artifacts are the "duplicate" matepairs; that is, a single matepair that is sequenced multiple times solely owing to artifacts in the library construction and/or sequencing processes. Duplicate molecules lead to spurious positive SV calls by falsely creating clusters of seemingly independent matepairs that suggest the same SV breakpoint. Duplicate matepairs arise either by PCR amplification of insufficiently complex DNA libraries, or in the case of the Illumina/Solexa platform, when the base-calling software incorrectly calls multiple sequences from a single cluster. Such duplicates can be identified after sequence alignment by screening for matepairs that have identical alignment coordinates. Software packages such as SAMTOOLS (67) and PICARD (68) provide utilities for removing duplicates sequences; however they do not allow duplicates to be detected from matepairs that have approximately the same alignment coordinates. We find this to be a necessary consideration as sequencing errors at the beginning or end of reads can cause bona fide duplicates to have alignment coordinates that differ by 1 or 2 bp. Moreover, neither SAMTOOLS nor PICARD have utilities for removing duplicates from datasets that contain multiple mappings for matepairs that align to nonunique sequence. As discussed above, including these mappings is a necessary requirement for identifying variants that involve segmental duplications and transposons. Removing duplicates in such datasets is complicated because, depending on alignment sensitivity and the number of mappings recorded, only a subset of mappings may be shared between duplicate reads. It is therefore necessary to examine *all* mappings for any that might be duplicates, and then to remove *all* of the mappings for all but one of the duplicate matepairs.

4.3.3. Reference Genome Effects

An insidious source of false positives arises from the incomplete nature of current reference genome assemblies. Repetitive genomic loci are notoriously difficult to assemble accurately. Such loci may be assembled improperly or incompletely, or may even be entirely missing from the reference genome. This is a major technical issue for SV detection for two reasons. First, improper or incomplete assembly yields a PEM signature that is indistinguishable from true SV. Second, matepairs arising from sequences that are not present in the reference genome—namely, centromeres, telomeres, and assembly gaps—can often be aligned to other genomic locations. Insofar as these erroneous alignments occur in a systematic fashion (as we are convinced they often do) false-positive SV calls will result. Artifacts caused by the reference genome can be difficult to identify and disregard. For example, in humans such effects might only be apparent after analyzing numerous genomes and noticing that certain variants were called in all samples (so-called monomorphic variants). Indeed, two sequencing-based studies that analyzed multiple humans reported an abnormal number of monomorphic variants (24, 61). Highly inbred species such as mouse and rat offer an important advantage in this respect since it is possible to re-sequence an individual from the reference strain as a control, and there should be very few new genetic differences between closely related inbred individuals (19). Our recent study represents the first to use this control, and we identified 405 high-confidence “variants” between our C57BL/6J individual and the reference (18). Only 10–20 of these appear to be real variants, and most of the remainder represent artifacts caused by the reference genome itself. While these confounding effects may be relatively mild in human and mouse, they will present significant obstacles for applying PEM in domesticated species with less complete reference genomes including rat, dog, cow, cat, pig, and others. This argues for a continued effort to improve the quality of reference genome assemblies in diverse organisms.

In some respects, accurate PEM can also depend on genome annotations. For example, some strategies (such as our own) remove reads that map to simple sequence repeats (SSRs), and others may attempt to identify transposon insertions by aligning matepairs directly to annotated TE sequences. Moreover, even if SV detection is entirely independent of genome annotations, the manner in which variants are classified and interpreted is inherently dependent on them. Genome annotations are not as comprehensive nor as accurate in genomes other than human, and in this respect SV discovery and/or interpretation can be more difficult.

4.4. SV Discovery with Split-Read Mapping

A more powerful approach to SV discovery is the use of longer (e.g., >200 bp) DNA sequences to characterize SV breakpoints at single base-pair resolution. DNA sequences from a test genome that span the site of an SV breakpoint will align to the reference genome in "split" fashion (69). That is, distinct segments of the DNA sequence will align to different loci in the reference genome, and the distance and orientation of these alignments indicate the type of rearrangement that occurred in the test genome (Fig. 2). The fundamental advantage of this approach is that a single "split" read can identify the exact nucleotide at which the breakpoint occurred. Multiple "split" reads corroborating the same breakpoint can be assembled with programs such as PHRAP (70) to generate a consensus sequence describing the breakpoint locus. Importantly, by aligning the consensus sequence to the reference genome, one can infer the causal mechanism based on the sequence homology in the regions flanking the breakpoint (Fig. 2). For example, extensive (e.g., >50 bp) sequence homology is a hallmark of NAHR, while breakpoints exhibiting microhomology or no homology suggest nonhomologous end joining (NHEJ) or template switching (52, 53). Complex variants that contain multiple breakpoints in close proximity or that have accumulated insertions of DNA directly into the breakpoint itself, most likely arose via template switching (71). These studies are crucial because there is substantial uncertainty about the relative role of different SV formation mechanisms (18, 24, 72), and obtaining a more coherent understanding of these molecular forces is a necessary prerequisite for understanding the etiology of human diseases that are caused by de novo structural variation, namely, genomic disorders and cancer. Thus, future research should be focused on characterizing as many SV breakpoints as possible from diverse germline and somatic genomes. In this regard, the use of long DNA sequences for split-read detection in model organisms is a very attractive approach.

4.4.1. Technical Caveats

A technical consideration for this approach is that longer sequences often overlap or contain repetitive DNA such as transposons, segmental duplications, or SSRs. Thus, sequence reads whose alignment(s) do not meet the criteria for "concordance" with the reference genome (e.g., >90% identity and >90% length) could

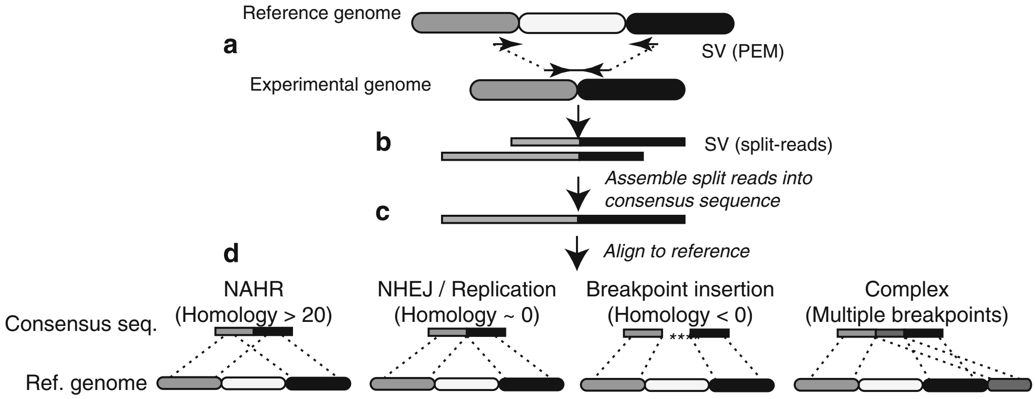
Reference genome
a
SV (PEM)
Experimental genome
b
SV (split-reads)
Assemble split reads into consensus sequence
c
d
Align to reference
NAHR (Homology > 20)
NHEJ / Replication (Homology ~ 0)
Breakpoint insertion (Homology < 0)
Complex (Multiple breakpoints)
Consensus seq.
Ref. genome

Fig. 2. Breakpoint isolation. At top is shown the reference genome and beneath that the experimental genome. (**a**) Breakpoints are localized to a genomic interval by paired-end mapping, where the reads are shown as *black arrows*, with the direction of the *arrow* indicating read orientation, and the alignments to the reference are shown as *dotted lines*. Note that the matepair maps the reference genome with a larger span than predicted, indicating a deletion in the experimental genome corresponding to the *light-gray* segment in the reference. (**b**) Long-reads that map to the predicted breakpoint region in split fashion (as shown in panel **d**) identify the breakpoint at single-nucleotide resolution. (**c**) Assembly of the long-reads produces a consensus sequence that describes the SV breakpoint. (**d**) Alignment of the consensus breakpoint sequence to the reference genome can reveal the molecular mechanism that generated the SV. From left, when the breakpoint contains significant homology to the regions flanking the deletion, the likely cause is NAHR. When the breakpoint contains little or no homology, the likely cause is NHEJ or replication-based template switching. When DNA has been inserted directly into the breakpoint, this manifests as a gap in the alignment to the reference. If the inserted DNA segment was generated by template-independent DNA synthesis, the likely cause is NHEJ. If the inserted DNA originated from elsewhere in the genome, the likely cause is template switching. At right, complex variants are apparent by the presence of multiple adjacent, often intertwined breakpoints. Most of these are likely due to template switching.

have smaller subsequences that align to hundreds or thousands of locations with similar identity. For this reason, one must define strict criteria for determining which alignments are retained for each distinct segment of a read when screening for putative split-read alignments. We also note that many of the technical challenges presented by short-read PEM, as discussed in the section above, are true for split-read mapping (SRM) as well. Indeed, while accurate alignments are much easier to obtain for long-reads, the difficulties presented by mapping SVs in repetitive elements are very similar, and the artifacts caused by duplicate reads and reference genome quality are likewise an issue.

4.4.2. Current Limitations of SRM

The current limitation to this approach is the cost of obtaining a sufficient number of long DNA sequences from a test genome. At the time of writing, traditional capillary sequencing and Roche/454 pyrosequencing are the primary means for generating longer reads. However, the low throughput and high cost of these technologies prohibit genome-wide SV breakpoint discovery in large genomes. Sequencing technologies from Pacific Biosciences, Life Technologies, and Ion Torrent promise to reduce the cost of generating sufficiently long DNA sequences, but whether the throughput of these systems will be on the scale necessary for genome-wide SV discovery in mammalian genomes is not clear. In the interim, a hybrid approach seems to be the most economical. In one possible hybrid strategy, SV breakpoints would localized to small genomic intervals using genome-wide PEM. Then, DNA capture techniques (73) would be used to isolate breakpoint-containing genomic regions, and breakpoints would be sequenced with a long-read technology. Breakpoints could then be characterized and/or genotyped on a large scale among many individuals or strains using the aforementioned "split-read" approach (see Fig. 2).

5. Application of NGS Methods to Model Mammals

Given the decreasing costs of DNA sequencing, it seems likely that all major lines of mouse, rat, and dog will be fully sequenced in the next few years using next-generation technologies. Indeed, the Mouse Genomes Project is already sequencing 17 inbred strains (http://www.sanger.ac.uk/resources/mouse/genomes/). These NGS data will not be sufficient for de novo whole-genome assembly, but they will be sufficient to generate genome-wide variation maps. To the extent that these maps are comprehensive and accurate they should resolve most outstanding questions regarding the prevalence and genomic distribution of SV in model mammals. What else might we learn about structural variation in the coming years?

5.1. Mechanisms of SV Formation

One unresolved question is the mechanistic origins of SV. This question has not yet been fully addressed in any species because mapping breakpoints to single-nucleotide resolution has historically been a laborious process. The large number of breakpoints that will be characterized over the next few years should allow for a direct comparison of the relative role of NAHR, NHEJ, and template switching among different species. Given their close evolutionary relationship one might expect that the contribution of different mechanisms would be conserved among all mammals; however, differences in the genomic distribution of segmental duplications indicates that there may be mechanistic differences. This important question should be resolved in the next few years.

A far more difficult question to address is how genetic factors or environmental conditions affect SV genesis. This subject has profound implications for our understanding of evolution and disease. It is possible, perhaps even likely, that genetic variation in genes that affect genome stability causes certain individuals to be more or less susceptible to genomic rearrangements. This could cause an increased risk of sporadic disease in these individuals and their progeny. Similar effects during evolution might lead to very different rates and/or patterns of genome evolution in specific lineages, as has been well-documented for gibbon (2). Furthermore, it was recently proposed that DNA replication-based mutations may be promoted by cellular stress (53, 71), and this raises the intriguing question of whether certain environmental conditions may affect rates of SV formation. While there is scant evidence for either of the above hypotheses, there is clearly strong historical precedent from cancer research that both genes and environment affect susceptibility to a disease marked by genomic rearrangements.

In our view, the only way to adequately address these questions is through development of a high-throughput screening method that is capable of measuring the effects of many different genetic backgrounds and environmental compounds in an unbiased manner. This is best accomplished in the laboratory mouse. While the screening methods for such studies do not yet exist and are difficult to envision with current technologies, we are hopeful that ongoing development of long-read single molecule sequencing technologies will increase the sensitivity of SV detection to a sufficient extent that the frequency of *rare* SVs can be measured *within* somatic and germline cell populations by SRM. This or a similar method could serve as a quantitative genome-wide measure of structural mutation rates, which would allow for unbiased identification of specific factors that modulate genome stability.

5.2. Somatic Variation

Another interesting question that could readily be addressed in mice is the prevalence of somatically acquired SV. There are intriguing suggestions from studies in human and mouse that individuals are composed of genetically variable somatic cell populations (74–78), and the extent to which this is true has important implications for diverse fields of biology including sporadic disease, cancer, aging, and stem cell therapy. Mice offer the obvious advantage of allowing many different tissues to be examined for SV in an inbred background. However, the major obstacle for studying somatic variation is obtaining pure samples of a given lineage. Crude tissue samples are generally composed of many different cell-types with diverse developmental histories, and current genome-wide methods cannot detect variants that are rare within a population of cells (as most somatic mutations are expected to be). Recent advances in stem cell technology provide a means to copy the genome of individual somatic cells through induced expression of 3–4 genes (79). By generating transgenic mice in which all somatic cells contain drug-inducible versions of these genes (80), it is possible to clone single somatic cells from diverse lineages (81). Application of sequence-based SV discovery methods to stem cell lines generated by this technology could answer a number of unresolved questions. For example, how prevalent is somatic variation? Are genomic patterns of somatic SV different from germline SV? Are somatic variants generated by the same mechanisms as germline variants, or do they more closely resemble the aberrations found in cancers? Are different developmental lineages more or less susceptible to new mutations?

5.3. The Genetics of Gene Expression

There is great interest in how genetic variation affects heritable variability in gene expression. The typical approach is to treat gene expression as a quantitative trait and to use conventional mapping methods to identify associated genetic variation (so-called eQTLs). Notably, aCGH-based experiments in human (82) and mouse (15, 16) have shown that CNVs make a significant contribution to gene expression and underlie as much as 20% of heritable differences. However, this line of investigation is limited in human due to the difficulty of obtaining a sufficient number of samples from interesting tissues and/or cell-types. In rodents, gene expression can be assessed in virtually any tissue in a stable and renewable inbred background. The availability of recombinant inbred lines and segregating populations make these organisms ideal for investigating this topic. Some interesting questions that could be addressed include the following. What types of variants have the greatest contribution to gene expression? Do these operate in *cis* or *trans*? What sorts of genetic factors are involved? At a systems level, what is the genetic architecture of gene expression control? One interesting line of investigation is the role of transposons in epigenetic gene control. TEs, in particular, LTRs, are known to silence genes and to serve as alternative promoters through epigenetic processes such as DNA methylation and RNA-interference. There are roughly ~5,000 variable TE insertions between two classical inbred mouse strains, many of which lie within or near genes (18), and likely many more TEVs among wild-derived strains. Therefore, assessing the functional effects of TE-mediated gene control is a tractable problem that can be addressed with current genomic methods.

5.4. Mapping Phenotypic Variation

The relationship between genetic and phenotypic variation is a fundamental question in biology, and at present the contribution of SV is unclear. Model mammals offer significant practical advantages for addressing this question. Most domesticated species display extremely high levels of phenotypic diversity between strains yet little within strains, and due to breeding history and artificial selection their genomes generally have simpler patterns of variation. These features allow trait mapping to be accomplished with a much smaller number of individuals. Moreover, artificial selection favors penetrant alleles with large phenotypic effects, which are the easiest to map, and due to inbreeding model mammals suffer from many of the same genetic diseases and susceptibilities that plague humans.

Comprehensive SV maps will be immediately useful for identifying functional variants in genomic regions identified by genome-wide association studies (GWAS). GWAS has proven to be a powerful mapping strategy in humans and model species, and using this approach many genomic regions have been identified that show a significant association with traits. There has recently been impressive success using GWAS in dogs (83), and current efforts to generate more diverse mapping populations in mouse through the "Collaborative Cross" project (84) promise to greatly increase the power of this approach. One limitation of GWAS is that pinning down the causal variant(s) embedded within a large associated region can be difficult. This is especially true in domesticated species since, due to breeding bottlenecks, LD can extend over large genomic distances, often several megabases (48, 56, 63). Whole-genome variation maps will make it far easier to identify and test candidate causal variants, and this will help to reveal the role of common SVs in phenotypic variation.

However, it is increasingly clear that for many (if not most) complex traits common genetic variation explains a minor fraction of the phenotypic variance. What, then, is the genetic basis for this missing heritability? Most current hypotheses revolve around a role for rare variants (i.e., recent mutations) with large phenotypic effects (85), and in this respect SVs are prime candidates: they are often large, can affect gene dosage and/or structure, and rates of SV formation are relatively high compared to SNPs, particularly at hotspots (38). How might model species help to resolve this issue?

The only reliable way to assess the role of recent mutations is by direct detection. Ideally, this would be done by whole-genome sequencing of all the individuals in a mapping population, but this will remain prohibitively expensive for at least 5 years. In the interim, a potential alternative is to perform direct genotyping on the most informative set of the genetic differences obtained by sequencing widely used strains. This approach is powerful in domesticated species because most relevant genetic variation can be captured by sequencing a small number of individuals. This includes recent mutations that are not well-tagged by SNPs. In this context, the high-resolution SV detection methods that we describe above are perfectly suited for assessing the functional impact of both ancient and recent SV. One potential approach is to develop high-throughput PCR assays to directly genotype SV breakpoints. Coupled with conventional SNP genotyping, this method could be useful as a more direct form of trait mapping. Alternatively, one could imagine a more powerful genotyping system based upon sequence capture of all SV breakpoints, all putative functional SNPs, as well as an adequate number of haplotype-tagging SNPs, followed by DNA sequence-based genotyping with NGS technologies.

Using such methods we expect that significant breakthroughs will be obtained in coming years and that a coherent understanding of the causes and consequences of genomic structural variation will emerge.

References

1. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., and Eichler, E. E. (2001) Segmental duplications: organization and impact within the current human genome project assembly, *Genome Res 11*, 1005–1017.
2. Marques-Bonet, T., Girirajan, S., and Eichler, E. E. (2009) The origins and impact of primate segmental duplications, *Trends Genet 25*, 443–454.
3. She, X., Cheng, Z., Zollner, S., Church, D. M., and Eichler, E. E. (2008) Mouse segmental duplication and copy number variation, *Nat Genet 40*, 909–914.
4. 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.
5. Tuzun, E., Bailey, J. A., and Eichler, E. E. (2004) Recent segmental duplications in the working draft assembly of the brown Norway rat, *Genome Res 14*, 493–506.
6. 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*, 491–499.
7. 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*, 949–951.
8. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, 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*, 525–528.
9. Feuk, L., Carson, A. R., and Scherer, S. W. (2006) Structural variation in the human genome, *Nature Reviews Genetics 7*, 85–97.
10. Li, J., Jiang, T., Mao, J. H., Balmain, A., Peterson, L., Harris, C., Rao, P. H., Havlak, P., Gibbs, R., and Cai, W. W. (2004) Genomic segmental polymorphisms in inbred mouse strains, *Nat Genet 36*, 952–954.
11. Snijders, A. M., Nowak, N. J., Huey, B., Fridlyand, J., Law, S., Conroy, J., Tokuyasu, T., Demir, K., Chiu, R., Mao, J. H., Jain, A. N., Jones, S. J., Balmain, A., Pinkel, D., and Albertson, D. G. (2005) Mapping segmental and sequence variations among laboratory mice using BAC array CGH, *Genome Res 15*, 302–311.
12. Adams, D. J., Dermitzakis, E. T., Cox, T., Smith, J., Davies, R., Banerjee, R., Bonfield, J., Mullikin, J. C., Chung, Y. J., Rogers, J., and Bradley, A. (2005) Complex haplotypes, copy number polymorphisms and coding variation in two recently divergent mouse strains, *Nat Genet 37*, 532–536.
13. Cutler, G., Marshall, L. A., Chin, N., Baribault, H., and Kassner, P. D. (2007) Significant gene content variation characterizes the genomes of inbred mouse strains, *Genome Res 17*, 1743–1754.
14. Graubert, T. A., Cahan, P., Edwin, D., Selzer, R. R., Richmond, T. A., Eis, P. S., Shannon, W. D., Li, X., McLeod, H. L., Cheverud, J. M., and Ley, T. J. (2007) A high-resolution map of segmental DNA copy number variation in the mouse genome, *PLoS Genet 3*, e3.
15. Cahan, P., Li, Y., Izumi, M., and Graubert, T. A. (2009) The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells, *Nat Genet 41*, 430–437.
16. Henrichsen, C. N., Vinckenbosch, N., Zollner, S., Chaignat, E., Pradervand, S., Schutz, F., Ruedi, M., Kaessmann, H., and Reymond, A. (2009) Segmental copy number variation shapes tissue transcriptomes, *Nat Genet 41*, 424–429.
17. Akagi, K., Li, J., Stephens, R. M., Volfovsky, N., and Symer, D. E. (2008) Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition, *Genome Res 18*, 869–880.
18. Quinlan, A. R., Clark, R. A., Sokolova, S., Leibowitz, M. L., Zhang, Y., Hurles, M. E., and Hall, I. M. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome, *Genome Research In Press.*
19. Egan, C. M., Sridhar, S., Wigler, M., and Hall, I. M. (2007) Recurrent DNA copy number variation in the laboratory mouse, *Nat Genet 39*, 1384–1389.
20. Guryev, V., Saar, K., Adamovic, T., Verheul, M., van Heesch, S. A., Cook, S., Pravenec, M., Aitman, T., Jacob, H., Shull, J. D., Hubner, N., and Cuppen, E. (2008) Distribution and functional impact of DNA copy number variation in the rat, *Nat Genet 40*, 538–545.
21. Chen, W. K., Swartz, J. D., Rush, L. J., and Alvarez, C. E. (2009) Mapping DNA structural variation in dogs, *Genome Res 19*, 500–509.
22. Liu, G. E., Hou, Y., Zhu, B., Cardone, M. F., Jiang, L., Cellamare, A., Mitra, A., Alexander, L. J., Coutinho, L. L., Dell'aquila, M. E.,

Gasbarre, L. C., Lacalandra, G., Li, R. W., Matukumalli, L. K., Nonneman, D., Regitano, L. C., Smith, T. P., Song, J., Sonstegard, T. S., Van Tassell, C. P., Ventura, M., Eichler, E. E., McDaneld, T. G., and Keele, J. W. Analysis of copy number variations among diverse cattle breeds, *Genome Res.*

23. 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., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., and Hurles, M. E. (2009) Origins and functional impact of copy number variation in the human genome, *Nature.*

24. 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., Tuzun, 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., Stama-toyannopoulos, 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.

25. 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., Gonzalez, J. R., Gratacos, 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., 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*, 444–454.

26. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara, E. C. M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008) Accurate whole human genome sequencing using reversible terminator chemistry, *Nature 456*, 53–59.

27. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J.,

Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G. K., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., and Yang, H. (2008) The diploid genome sequence of an Asian individual, *Nature 456*, 60–65.

28. Ahn, S. M., Kim, T. H., Lee, S., Kim, D., Ghang, H., Kim, B. C., Kim, S. Y., Kim, W. Y., Kim, C., Park, D., Lee, Y. S., Kim, S., Reja, R., Jho, S., Kim, C. G., Cha, J. Y., Kim, K. H., Lee, B., Bhak, J., and Kim, S. J. (2009) The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group, *Genome Res.*

29. Hormozdiari, F., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes, *Genome Res 19*, 1270–1278.

30. McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek, J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D., Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., Macbride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., De La Vega, F. M., and Blanchard, A. P. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding, *Genome Res.*

31. Kazazian, H. H., Jr. (2004) Mobile elements: drivers of genome evolution, *Science 303*, 1626–1632.

32. 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, *American Journal of Human Genetics 77*, 78–88.

33. 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, *Nature Genetics 37*, 727–732.

34. Perry, G. H., Tchinda, J., McGrath, S. D., Zhang, J., Picker, S. R., Caceres, 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 U S A 103*, 8006–8011.

35. Bourque, G., Pevzner, P. A., and Tesler, G. (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes, *Genome Res 14*, 507–516.

36. Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D., and Eichler, E. E. (2004) Hotspots of mammalian chromosomal evolution, *Genome Biol 5*, R23.

37. Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L., Hitte, C., Meyers, S. N., Milan, D., Ostrander, E. A., Pape, G., Parker, H. G., Raudsepp, T., Rogatcheva, M. B., Schook, L. B., Skow, L. C., Welge, M., Womack, J. E., O'Brien S, J., Pevzner, P. A., and Lewin, H. A. (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps, *Science 309*, 613–617.

38. Lupski, J. R. (2007) Genomic rearrangements and sporadic disease, *Nat Genet 39*, S43-47.

39. Stankiewicz, P., Shaw, C. J., Dapper, J. D., Wakui, K., Shaffer, L. G., Withers, M., Elizondo, L., Park, S. S., and Lupski, J. R. (2003) Genome architecture catalyzes nonrecurrent chromosomal rearrangements, *Am J Hum Genet 72*, 1101–1116.

40. Lee, J. A., Inoue, K., Cheung, S. W., Shaw, C. A., Stankiewicz, P., and Lupski, J. R. (2006) Role of genomic architecture in PLP1 duplication causing Pelizaeus-Merzbacher disease, *Hum Mol Genet 15*, 2250–2265.

41. Bauters, M., Van Esch, H., Friez, M. J., Boespflug-Tanguy, O., Zenker, M., Vianna-Morgante, A. M., Rosenberg, C., Ignatius, J., Raynaud, M., Hollanders, K., Govaerts, K., Vandenreijt, K., Niel, F., Blanc, P., Stevenson, R. E., Fryns, J. P., Marynen, P., Schwartz, C. E., and Froyen, G. (2008) Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair, *Genome Res 18*, 847–858.

42. Carvalho, C. M., Zhang, F., Liu, P., Patel, A., Sahoo, T., Bacino, C. A., Shaw, C., Peacock, S., Pursley, A., Tavyev, Y. J., Ramocki, M. B., Nawara, M., Obersztyn, E., Vianna-Morgante,

A. M., Stankiewicz, P., Zoghbi, H. Y., Cheung, S. W., and Lupski, J. R. (2009) Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching, *Hum Mol Genet 18*, 2188–2203.

43. 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*, 1865–1874.
44. Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M., and Eichler, E. E. (2004) Analysis of segmental duplications and genome assembly in the mouse, *Genome Res 14*, 789–801.
45. Hampton, O. A., Den Hollander, P., Miller, C. A., Delgado, D. A., Li, J., Coarfa, C., Harris, R. A., Richards, S., Scherer, S. E., Muzny, D. M., Gibbs, R. A., Lee, A. V., and Milosavljevic, A. (2009) A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome, *Genome Res 19*, 167–177.
46. Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. (2002) Recent segmental duplications in the human genome, *Science 297*, 1003–1007.
47. Church, D. M., Goodstadt, L., Hillier, L. W., Zody, M. C., Goldstein, S., She, X., Bult, C. J., Agarwala, R., Cherry, J. L., DiCuccio, M., Hlavina, W., Kapustin, Y., Meric, P., Maglott, D., Birtle, Z., Marques, A. C., Graves, T., Zhou, S., Teague, B., Potamousis, K., Churas, C., Place, M., Herschleb, J., Runnheim, R., Forrest, D., Amos-Landgraf, J., Schwartz, D. C., Cheng, Z., Lindblad-Toh, K., Eichler, E. E., and Ponting, C. P. (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse, *PLoS Biol 7*, e1000112.
48. Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., 3rd, Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., DeJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C. W., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K. P., Parker, H. G., Pollinger, J. P., Searle, S. M., Sutter, N. B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-Zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., Foley, C., Franke, A., Friedrich, D., Gage, D., Garber, M., Gearin, G., Giannoukos, G., Goode, T., Goyette, A., Graham, J., Grandbois, E., Gyaltsen, K., Hafez, N., Hagopian, D., Hagos, B., Hall, J., Healy, C., Hegarty, R., Honan, T., Horn, A., Houde, N., Hughes, L., Hunnicutt, L., Husby, M., Jester, B., Jones, C., Kamat, A., Kanga, B., Kells, C., Khazanovich, D., Kieu, A. C., Kisner, P., Kumar, M., Lance, K., Landers, T., Lara, M., Lee, W., Leger, J. P., Lennon, N., Leuper, L., LeVine, S., Liu, J., Liu, X., Lokyitsang, Y., Lokyitsang, T., Lui, A., Macdonald, J., Major, J., Marabella, R., Maru, K., Matthews, C., McDonough, S., Mehta, T., Meldrim, J., Melnikov, A., Meneus, L., Mihalev, A., Mihova, T., Miller, K., Mittelman, R., Mlenga, V., Mulrain, L., Munson, G., Navidi, A., Naylor, J., Nguyen, T., Nguyen, N., Nguyen, C., Nicol, R., Norbu, N., Norbu, C., Novod, N., Nyima, T., Olandt, P., O'Neill, B., O'Neill, K., Osman, S., Oyono, L., Patti, C., Perrin, D., Phunkhang, P., Pierre, F., Priest, M., Rachupka, A., Raghuraman, S., Rameau, R., Ray, V., Raymond, C., Rege, F., Rise, C., Rogers, J., Rogov, P., Sahalie, J., Settipalli, S., Sharpe, T., Shea, T., Sheehan, M., Sherpa, N., Shi, J., Shih, D., Sloan, J., Smith, C., Sparrow, T., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Stone, S., Sykes, S., Tchuinga, P., Tenzing, P., Tesfaye, S., Thoulutsang, D., Thoulutsang, Y., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Venkataraman, V., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Yang, S., Yang, X., Young, G., Yu, Q., Zainoun, J., Zembek, L., Zimmer, A., and Lander, E. S. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog, *Nature 438*, 803–819.
49. Xing, J., Zhang, Y., Han, K., Salem, A. H., Sen, S. K., Huff, C. D., Zhou, Q., Kirkness, E. F., Levy, S., Batzer, M. A., and Jorde, L. B. (2009) Mobile elements create structural variation: analysis of a complete human genome, *Genome Res 19*, 1516–1526.

50. Cordaux, R., and Batzer, M. A. (2009) The impact of retrotransposons on human genome evolution, *Nat Rev Genet 10*, 691–703.
51. Bailey, J. A., Liu, G., and Eichler, E. E. (2003) An Alu transposition model for the origin and expansion of human segmental duplications, *Am J Hum Genet 73*, 823–834.
52. Lee, J. A., Carvalho, C. M., and Lupski, J. R. (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders, *Cell 131*, 1235–1247.
53. Hastings, P. J., Ira, G., and Lupski, J. R. (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation, *PLoS Genet 5*, e1000327.
54. Wade, C. M., and Daly, M. J. (2005) Genetic variation in laboratory mice, *Nat Genet 37*, 1175–1180.
55. Wade, C. M., Kulbokas, E. J., 3rd, Kirby, A. W., Zody, M. C., Mullikin, J. C., Lander, E. S., Lindblad-Toh, K., and Daly, M. J. (2002) The mosaic structure of variation in the laboratory mouse genome, *Nature 420*, 574–578.
56. Saar, K., Beck, A., Bihoreau, M. T., Birney, E., Brocklebank, D., Chen, Y., Cuppen, E., Demonchy, S., Dopazo, J., Flicek, P., Foglio, M., Fujiyama, A., Gut, I. G., Gauguier, D., Guigo, R., Guryev, V., Heinig, M., Hummel, O., Jahn, N., Klages, S., Kren, V., Kube, M., Kuhl, H., Kuramoto, T., Kuroki, Y., Lechner, D., Lee, Y. A., Lopez-Bigas, N., Lathrop, G. M., Mashimo, T., Medina, I., Mott, R., Patone, G., Perrier-Cornet, J. A., Platzer, M., Pravenec, M., Reinhardt, R., Sakaki, Y., Schilhabel, M., Schulz, H., Serikawa, T., Shikhagaie, M., Tatsumoto, S., Taudien, S., Toyoda, A., Voigt, B., Zelenika, D., Zimdahl, H., and Hubner, N. (2008) SNP and haplotype mapping for genetic analysis in the rat, *Nat Genet 40*, 560–566.
57. Medvedev, P., Stanciu, M., and Brudno, M. (2009) Computational methods for discovering structural variation with next-generation sequencing, *Nat Methods 6*, S13-S20.
58. Du, J., Bjornson, R. D., Zhang, Z. D., Kong, Y., Snyder, M., and Gerstein, M. B. (2009) Integrating sequencing technologies in personal genomics: optimal low cost reconstruction of structural variants, *PLoS Comput Biol 5*, e1000432.
59. Bashir, A., Volik, S., Collins, C., Bafna, V., and Raphael, B. J. (2008) Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer, *PLoS Comput Biol 4*, e1000051.
60. Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J. O., Baker, C., Malig, M., Mutlu, O., Sahinalp, S. C., Gibbs, R. A., and Eichler, E. E. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing, *Nat Genet 41*, 1061–1067.
61. Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage, *Genome Res.*
62. Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. S. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing, *Nat Methods 6*, 99–103.
63. Frazer, K. A., Eskin, E., Kang, H. M., Bogue, M. A., Hinds, D. A., Beilharz, E. J., Gupta, R. V., Montgomery, J., Morenzoni, M. M., Nilsen, G. B., Pethiyagoda, C. L., Stuve, L. L., Johnson, F. M., Daly, M. J., Wade, C. M., and Cox, D. R. (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains, *Nature 448*, 1050–1053.
64. Li, H., and Durbin, R. (2009) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform, *Bioinformatics.*
65. Novoalign. (www. novocraft.com).
66. Mosiak. (http://code.google.com/p/mosaik-aligner/).
67. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics 25*, 2078–2079.
68. Picard. (http://picard.sourceforge.net/).
69. Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome, *Genome Res 16*, 1182–1190.
70. Green, P. (unpublished) http://www.phrap.org/phredphrapconsed.html.
71. Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira, G. (2009) Mechanisms of change in gene copy number, *Nat Rev Genet 10*, 551–564.
72. Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M., and Snyder, M. (2007) Paired-end mapping reveals extensive structural variation in the human genome, *Science 318*, 420–426.
73. Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., and Turner, D. J. Target-

enrichment strategies for next-generation sequencing, *Nat Methods 7*, 111–118.

74. Liang, Q., Conte, N., Skarnes, W. C., and Bradley, A. (2008) Extensive genomic copy number variation in embryonic stem cells, *Proc Natl Acad Sci U S A 105*, 17453–17456.
75. Bruder, C. E., Piotrowski, A., Gijsbers, A. A., Andersson, R., Erickson, S., de Stahl, T. D., Menzel, U., Sandgren, J., von Tell, D., Poplawski, A., Crowley, M., Crasto, C., Partridge, E. C., Tiwari, H., Allison, D. B., Komorowski, J., van Ommen, G. J., Boomsma, D. I., Pedersen, N. L., den Dunnen, J. T., Wirdefeldt, K., and Dumanski, J. P. (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles, *Am J Hum Genet 82*, 763–771.
76. Piotrowski, A., Bruder, C. E., Andersson, R., de Stahl, T. D., Menzel, U., Sandgren, J., Poplawski, A., von Tell, D., Crasto, C., Bogdan, A., Bartoszewski, R., Bebok, Z., Krzyzanowski, M., Jankowski, Z., Partridge, E. C., Komorowski, J., and Dumanski, J. P. (2008) Somatic mosaicism for copy number variation in differentiated human tissues, *Hum Mutat 29*, 1118–1124.
77. Lam, K. W., and Jeffreys, A. J. (2007) Processes of de novo duplication of human alpha-globin genes, *Proc Natl Acad Sci U S A 104*, 10950–10955.
78. Flores, M., Morales, L., Gonzaga-Jauregui, C., Dominguez-Vidana, R., Zepeda, C., Yanez, O., Gutierrez, M., Lemus, T., Valle, D., Avila, M. C., Blanco, D., Medina-Ruiz, S., Meza, K., Ayala, E., Garcia, D., Bustos, P., Gonzalez, V., Girard, L., Tusie-Luna, T., Davila, G., and Palacios, R. (2007) Recurrent DNA inversion rearrangements in the human genome, *Proc Natl Acad Sci U S A 104*, 6099–6106.
79. Takahashi, K., and Yamanaka, S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors, *Cell 126*, 663–676.
80. Boland, M. J., Hazen, J. L., Nazor, K. L., Rodriguez, A. R., Gifford, W., Martin, G., Kupriyanov, S., and Baldwin, K. K. (2009) Adult mice generated from induced pluripotent stem cells, *Nature 461*, 91–94.
81. Wernig, M., Lengner, C. J., Hanna, J., Lodato, M. A., Steine, E., Foreman, R., Staerk, J., Markoulaki, S., and Jaenisch, R. (2008) A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types, *Nat Biotechnol 26*, 916–924.
82. Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavare, S., Deloukas, P., Hurles, M. E., and Dermitzakis, E. T. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes, *Science 315*, 848–853.
83. Shearin, A. L., and Ostrander, E. A. Leading the way: canine models of genomics and disease, *Dis Model Mech 3*, 27–34.
84. Iraqi, F. A., Churchill, G., and Mott, R. (2008) The Collaborative Cross, developing a resource for mammalian systems genetics: a status report of the Wellcome Trust cohort, *Mamm Genome 19*, 379–381.
85. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A., and Visscher, P. M. (2009) Finding the missing heritability of complex diseases, *Nature 461*, 747–753.