Comparison of constitutional and replication stress-induced genome structural variation by SNP array and mate-pair sequencing


*Department of Human Genetics
§Department of Pathology
**Department of Biological Chemistry and University of Michigan DNA Sequencing Core

University of Michigan Medical School
Ann Arbor MI, 48109
Running title: Replication stress-induced copy number variation

Keywords:

human genome

copy number variant (CNV)

replication stress

genome sequencing

SNP microarray

Corresponding author:

Thomas E. Wilson

Department of Pathology

University of Michigan Medical School

2065 BSRB

109 Zina Pitcher Place

Ann Arbor MI, 48109-2200

wilsonte@umich.edu

Phone: 734-764-2212
Abstract

Copy number variants (CNVs) are a major source of genetic variation in human health and disease. Previous studies have implicated replication stress as a causative factor in CNV formation. However, existing data are technically limited in the quality of comparisons that can be made between human CNVs and experimentally induced variants. Here, we used two high resolution strategies – single nucleotide polymorphism (SNP) arrays and mate-pair sequencing – to compare CNVs that occur constitutionally to those that arise following aphidicolin-induced DNA replication stress in the same human cells. Although the optimized methods provided complementary information, sequencing was more sensitive to small variants and provided superior structural descriptions. The majority of constitutional and all aphidicolin-induced CNVs appear to be formed via homology-independent mechanisms, while aphidicolin-induced CNVs were of a larger median size than constitutional events even when mate-pair data were considered. Aphidicolin thus appears to stimulate formation of CNVs that closely resemble human pathogenic CNVs and the subset of larger nonhomologous constitutional CNVs.
**Introduction**

In recent years, submicroscopic structural variants (SVs) have been found to be widely distributed throughout the human genome where they represent an important component of genetic variation and phenotypic diversity (Conrad et al. 2010b; Iafrate et al. 2004; Sebat et al. 2004; Sharp et al. 2005). These include deletions, duplications, insertions and inversions, with the majority being copy number variations (CNVs) discovered in systematic studies using microarrays (Conrad et al. 2010b; Park et al. 2010). More than 10,000 CNVs have now been described in healthy individuals that represent gains or losses of approximately one Kb to over a Mb. CNVs can alter gene expression in affected regions, confer redundancy, and provide substrates for evolution. Spontaneous CNVs are also known to be a major cause of genetic and developmental disorders, including mental retardation, autism, schizophrenia, epilepsy, skeletal defects and many others (Cook and Scherer 2008; Kirov et al. 2009; Miller et al. 2010; Stankiewicz and Beaudet 2007; Tam et al. 2009; Zhang et al. 2009). Systematic studies of human population CNVs have provided further correlation to human conditions including Crohn's disease, rheumatoid arthritis and diabetes (Craddock et al. 2010). Related systematic efforts have finally also revealed a high degree of submicroscopic chromosomal structural alterations in cancer (Bignell et al. 2010; Stratton et al. 2009).

Despite their importance, there is limited understanding of how SVs arise (Hastings et al. 2009b; Stankiewicz and Lupski 2010). The exceptions are local genome rearrangements that occur by unequal recombination between neighboring low-copy repeated sequences or segmental duplications, a process known as non-allelic homologous recombination (NAHR) (Sasaki et al. 2010). Such events are well described and underlie the specific recurrent alterations responsible for a variety of human microdeletion syndromes (Sasaki et al. 2010;
However, the majority of both non-recurrent pathogenic CNVs and those observed in the normal population do not appear to proceed by NAHR but instead show, at most, limited microhomology at the breakpoint junctions (Conrad et al. 2010a; Vissers et al. 2009). Multiple pathways might catalyze the formation of such junctions, including the best described nonhomologous end-joining (NHEJ) pathway of DNA double-strand break repair (Lieber 2010; Lieber and Wilson 2010), alternative end joining pathways recently implicated in chromosomal translocations (Booila et al. 2010; McVey and Lee 2008; Simsek and Jasin 2010), and entirely distinct pathways in which stalled replication structures are processed by mechanisms variably known as template switching or microhomology-mediated break-induced replication (MMBIR) (Hastings et al. 2009a; Lee et al. 2007).

To date, these mechanisms have largely been inferred by examination of human CNV breakpoint sequences (Conrad et al. 2010a; Korbel et al. 2007; Vissers et al. 2009). To begin to explore CNV mechanisms experimentally, we recently reported a system in which normal human fibroblasts were treated with the replication inhibitor aphidicolin (Arlt et al. 2009). Treatment was associated with a substantially increased frequency of new CNVs in subclones, as detected by array comparative genome hybridization (aCGH). The observed CNVs were generally consistent with many normal and pathogenic human CNVs and suggested either template switching or nonhomologous repair formation mechanisms (Hastings et al. 2009a; Lee et al. 2007; Lieber 2010; Lieber and Wilson 2010). However, the resolution of the aCGH method used left uncertainty as to the full spectrum of CNVs that are induced by aphidicolin as compared to those observed in the human germline.

To address these issues, we have explored and optimized various methods for detecting CNVs and other SVs with a focus on those with sufficient power and low enough cost for routine
experimentation. We report an in-depth analysis of two complementary technologies, high
density SNP arrays and whole genome mate-pair sequencing, and use them to compare
aphidicolin-induced CNVs to the baseline constitutional CNVs in the same experimental
samples. Our software platform, VAMP (BIRKELAND et al. 2010), was expanded to support the
many bioinformatics aspects of the study. We found a surprisingly low correspondence between
array and sequencing methods in the detection of constitutional SVs and accordingly identified
more than 600 SVs by mate-pair analysis that were not previously known from systematic array-
based studies (CONRAD et al. 2010b; PARK et al. 2010). A much higher method correspondence
was observed for aphidicolin-induced CNVs mainly because these events were consistently
larger than most constitutional CNVs even when the higher resolution of mate-pair sequencing
was brought to bear in the analysis.
Materials and Methods

Human cell lines. All experiments were performed with normal human fibroblast cell line HGMDFN090 (090), which was obtained from the Progeria Research Foundation Cell & Tissue Bank (Peabody, MA). The source individual is a female of European descent with a normal 46,XX karyotype who does not carry the mutation for Hutchinson-Gilford Progeria. Two aphidicolin-treated subclones of 090 that contain novel induced CNVs, called A3A2 and A1A1, have also been described that were obtained prior to immortalization of 090 (ARLT et al. 2009). More recently, 090 was immortalized by stable transfection with vector pBABE-Hygro-hTERT (COUNTER et al. 1998). A hygromycin-resistant clone was isolated, expanded and called 090D2. Parental (i.e. not aphidicolin-treated) SNP array and mate-pair analyses were performed with 090D2 when specified. Genomic DNA was prepared from cell lines using the Blood & Cell Culture DNA Mini Kit (Qiagen).

SNP microarrays. Microarrays were the Illumina HumanOmni1-Quad BeadChip, which has both SNP and non-SNP probes selected by the vendor to optimize the detection of human CNVs. One μg of genomic DNA was submitted to the University of Michigan DNA Sequencing Core for labeling, array hybridization and scanning according to the manufacturer’s instructions. X, Y, Log R Ratio and B Allele Frequency values were obtained using Illumina BeadStudio.

Mate-pair sequencing. Twenty to forty μg of genomic DNA were used to construct mate-pair libraries using the Illumina Mate Pair Library Prep Kit followed by paired end sequencing by the University of Michigan DNA Sequencing Core according to the manufacturer’s instructions. Image analysis and base-calling were performed using Illumina programs Firecrest and Bustard, respectively.
**Data analysis.** All further data analysis was performed using an expanded version of our VAMP software platform (BIRKELAND et al. 2010), which is available for download at http://tewlab.path.med.umich.edu/vamp.html. See Supplementary Methods and Supplementary Figures 1-8 for a description of the platform, logic, and parameters. Human genome Build 36 (hg18) served as the reference genome.

SNP microarray analysis was performed using moving average windows of 5, 10, 20 and 50 probes and a threshold of 5 standard deviations from the array mean. Candidate genome regions identified with these parameters were subjected to further filtering during visualization that required the best segment call within a region to have either (i) a change in the log₂ of the intensity ratio (log₂R) of at least 0.15 or (ii) a change in the B allele frequency of informative probes of at least 0.083, as well as (iii) a Z statistic of at least 7. The Z statistic is the deviation of the average value of a contiguous segment of probes relative to the average value over all probes in the array, expressed as the number of standard errors of the array mean. Thus, Z is influenced by the absolute deviation of a segment, the number of probes within it, and the noise level of the array. All passing genome regions were individually examined and CNVs manually adjusted and committed. Constitutional CNV analysis was performed on 090D2 array data alone. A3A2 was analyzed using 090D2 as the normalization reference.

For mate-pair sequencing, mapping filters allowed up to 5 mismatches relative to hg18, including indels, and up to 10 initial genome map positions per read. The combined 090 data, used to detect constitutional CNVs and which acted as the reference for detecting induced CNVs in A3A2 and A1A1, merged four sequencing lanes, two derived from 090 libraries, two from 090D2 libraries. Candidate genome regions were identified by seeking sets of anomalous fragments as described in BIRKELAND et al. 2010 and Supplementary Methods. Sets were
subjected to filtering during visualization that required them to have (i) no more than 40% promiscuously mapped fragments, (ii) a fractional overlap of no more than 90%, (iii) an average fragment size deviation, $\Delta$, of no more than 3 population SD, (iv) no more than 40% of fragments where $\Delta$ exceeded 2 SD, and (iv) no more than 10% of fragments in the region contributed by the reference sample (comparative studies only). For insertions, an additional filter required that the set contain at least 5 fragments. All passing sets were individually examined and manually committed.

CNV segments predicted by SNP arrays and mate-pair sequencing were finally compared to each other, to an analysis of 090 performed using PennCNV (WANG et al. 2007), and to CNVs from published compendia (CONRAD et al. 2010b; PARK et al. 2010). Two events were declared as matching if the overlap of the two spans was at least 5% of the larger of the two spans. Conclusions were not substantially different when calculated at a match threshold of 33%, given that most events showed either no match or a strong match of >50% (Supplementary Figures 9 and 10).

**Breakpoint analysis.** For a subset of CNVs detected by mate-pairs, a single PCR primer pair was designed that flanked the anomalous junction predicted by the analysis in Supplementary Figure 3. Occasionally, the first primer pair failed to give a product, but in most such cases products were obtained by moving the primers to different positions nonetheless consistent with the same junction. All products were then subjected to standard sequencing.

**Data availability.** Array data have been deposited in the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession GSE26121, mate-pair sequence data have been deposited in the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra) under study number SRP003289, and called SVs have
been deposited in NCBI dbVar

Results

Detection of aphidicolin-induced CNVs. We first sought to compare the ability of SNP microarrays and mate-pair sequencing (Table 1) to detect induced CNVs in A3A2 and A1A1, two aphidicolin-treated subclones of the normal human fibroblast cell line 090, including CNVs previously identified by aCGH (ARLT et al. 2009) as well as potentially unknown events. All CNVs shown in Table 2 were detected by at least one of the three methods and subsequently validated, in two cases by confirming loss of heterozygosity of informative SNPs in the deleted region (A3A2 Chr7 and A1A1 Chr13) (ARLT et al. 2009) and in all other cases by flanking PCR and sequencing of the junction (Supplementary Table 1). Although concordance was high, no one technique identified all novel CNVs in the samples. As a trend, microarrays were least robust at detecting the smallest events because discovery algorithms are strongly influenced by the probe count. Not limited by probe density, mate-pair sequencing detected three new events not called by either the SNP array or aCGH (A3A2 Chr4, A1A1 Chr1 and A1A1 Chr16; Table 2), even though these could be appreciated in array data once attention was directed to the region. All were copy number gains, which can be difficult to detect by microarray due to their smaller signal deviation, a limitation not applicable to mate-pair sequencing as it detects novel junctions directly. Mate-pair sequencing in turn failed to detect a pericentromeric deletion on A3A2 Chr7 that because of repeat content had too many artifactual mappings for reliable event calling. The A1A1 deletion on Chr13 was also missed, almost certainly as a result of the lower sequencing coverage of this sample (Table 1).

When using optimized filtering parameters (see Experimental Methods), the list of candidate new CNVs returned from A3A2 SNP array data did not include any false positives. A larger number of false-positive CNVs were returned as candidates from mate-pair data,
approximately 20 per sample. These false events all failed to show the expected change in coverage of the normal ~3kb fragments within the putative CNV and were not corroborated by array data, in contrast to real events (Figure 1 and Supplementary Figure 11). They also consistently had only two crossing mate-pairs and thus likely correspond to stochastic artifactual fragments or mappings. These artifacts were easy to dismiss with experience, but visual inspection of all candidate events was necessary and CNV calling was most reliable when parallel array and mate-pair data could be compared.

**Detection of constitutional CNVs.** We next sought to explore the content of constitutional CNVs in 090, for comparison to induced CNVs and to better understand the capabilities of the methods. Table 3 shows the number of copy number gains and losses called from SNP array and mate-pair sequence data, grouped according to their correspondence to the other detection method (see Supplementary Table 2 for a complete list). Consistent with other human individuals (CONRAD et al. 2010b; KORBEL et al. 2007; PANG et al. 2010; PARK et al. 2010; WHEELER et al. 2008), more than $10^3$ CNVs were discovered, with more losses than gains (1513 vs. 146, respectively). Unlike induced CNVs, a surprisingly large number of constitutional CNVs were called by only one of the two detection methods (1329, 80% of all CNVs; Table 3, rows A-C; see Figure 2 for examples). A small number of events were judged to be falsely negative calls (Table 3, row D). Other failed detections were attributable to technical limitations, in particular having too few array probes (Table 3, last column) or too small an event size for mate-pair detection. However, even after these factors were considered, many CNVs that might have been detected by both methods were not.

We entertained many explanations for the low concordance between SNP arrays and mate-pairs. First, VAMP might give poor detection of CNVs from SNP array data. Use of other
CNV calling algorithms such as PennCNV (WANG et al. 2007) did not improve the concordance, however (Supplementary Figure 9). Moreover, the array data in genomic regions called as mate-pair-specific CNVs were not statistically different from the array average (Table 3, rows H and I and Figure 3). This dichotomy could not be attributed to a specific array run (Figure 3). An alternative would be that VAMP substantially over-called CNVs from mate-pair data. Many observations suggest this is not the case. The three categories of mate-pair CNVs showed little or no quality difference in the number of associated mate-pair mappings or rate of base mismatches in associated reads (Table 3, row J and K and Supplementary Figure 12). Also, when mate-pair-specific CNVs had adequate array probe coverage, the fraction that matched a known human CNV remained very high (Table 3, row E).

Interestingly, the category of mate-pair-specific SVs having too few SNP probes for array detection (Table 3, last column, row G) was noticeably different in having a low rate of correspondence to human CNVs detected by ultra-high density aCGH (Table 3, row E) (CONRAD et al. 2010b; PARK et al. 2010). We looked for systematic factors that might make these events amenable to mate-pair as opposed to array detection. Most CNVs created by underlying inversions were not well sampled by the SNP array (66, 94% of all inversion CNVs; Table 3, row L), perhaps because they were relatively small (median size 1.6 kb, range 0.1 to 20 kb). A majority (67%) of mate-pair events with insufficient array probe coverage contained short tandem repeat elements (BENSON 1999) whose contraction might lead to size loss by a non-CNV mechanism (Table 3, row M). Some mate-pair-specific events also likely represent simple deletions of mobile repetitive elements, such as LINEs and HERVs, for which array probes often cannot be meaningfully designed (e.g. Figure 2D).
Finally, we examined the smaller number of CNVs identified by SNP arrays but not mate-pairs. The array statistics for these events were consistently robust (Table 3, rows H and I and Figure 3) and they once again showed a high concordance with known human CNVs (Table 3, row E).

In summary, most constitutional event calls could be validated either internally by virtue of detection by both methods or externally by correspondence to known human CNVs, despite the low concordance between SNP arrays and mate-pairs. Mate-pair-specific SVs were the largest unvalidated group, so we randomly selected 35 such events and attempted to amplify the anomalous junctions by PCR (Supplementary Table 1). These SVs encompassed most observed event types, including CNVs with and without sufficient probes for array detection. Only two had been previously described by systematic array studies (CONRAD et al. 2010b; PARK et al. 2010). Nearly all showed a PCR product consistent with the predicted SV (31/35, 89%, Supplementary Table 1), confirming the high reliability of mate-pair calls.

**Fine structure of aphidicolin-induced CNVs.** Mate-pair sequencing was very robust at describing SV structures. With regard to aphidicolin-induced CNVs, this included recognition that one breakpoint of a deletion on Chr15 fell 0.4 kb from a known human population CNV, as well as the unambiguous characterization of the two alleles underlying a homozygous deletion on Chr3 (Figure 1A and B). Most strikingly, mate-pairs described the precise structure of a complex inversion that created two duplication segments on Chr4, an event difficult to detect and impossible to describe from array data alone (Figure 1C). Similar observations held true for constitutional events, including the identification of CNVs associated with 38 inversions with a median inverted segment size of 2.3 kb (range 0.1 to 28 kb, Supplementary Table 2).
Aphidicolin-induced CNVs are large compared to many constitutional CNVs. Our last and most important goal was to compare constitutional and aphidicolin-induced SVs to explore the hypothesis that replication stress induces events typical of human polymorphic CNVs and whether the detection method had any effect on this conclusion. As expected, the size distribution of constitutional 090 SVs matched the pattern described for other human individuals and the human population (Conrad et al. 2010b; Korbel et al. 2007; Pang et al. 2010; Park et al. 2010; Wheeler et al. 2008), with an inverse correlation between size and frequency down to a lower detection limit of roughly 1 kb (Figure 4). Array CNVs matched many known human CNVs particularly closely, as expected since the SNP array was designed to target these events. Mate-pair events showed a slightly different pattern in being very sensitive for small gains and losses in regions not as easily sampled by arrays, but overall the methods gave a similar result that most constitutional CNVs (~94%) were smaller than 10 kb.

In marked contrast to the constitutional CNVs, 93% of aphidicolin-induced CNVs were larger than 10 kb (median 148 kb; Figure 4), a data set that includes all aphidicolin-induced CNVs we have described previously (Arlt et al. 2009), in the current study, and in ongoing unpublished analyses using Illumina 1M SNP arrays. We noted this different size distribution previously but did not know whether it might simply reflect a bias of arrays against detecting small induced events (Arlt et al. 2009). In contrast to the deliberate over-sampling of constitutional CNV regions, most induced CNVs are sampled at the average density of 0.37 probes per kb, so that ~14 kb are required to cross the five probes needed for CNV calling. Mate-pairs are not subject to this limit and indeed were readily able to detect constitutional changes as small as 1 kb (Figure 4). Despite this, mate-pairs did not reveal a new and more frequent class of small aphidicolin-induced CNVs in two different subclones. Some new
aphidicolin-induced CNVs were discovered by mate-pairs that were indeed smaller than those previously known from aCGH, but these were nonetheless larger than 10 kb (Table 2).

**Aphidicolin-induced and constitutional CNVs show similar nonhomologous junctions.** To assess the mechanisms likely to underlie formation of the observed CNVs, we sequenced a subset of breakpoint junctions (Supplementary Table 1). Extending previous observations (ARLT et al. 2009), all seven aphidicolin-induced events showed microhomologies, blunt joints and insertions and thus were inconsistent with NAHR. Constitutional CNV junctions revealed a mixture of apparent mechanisms typical of previous reports (CONRAD et al. 2010a; KORBEL et al. 2007; VISSERS et al. 2009), with 2 of 14 (14%) showing extended homology indicative of NAHR and 12 of 14 (86%) being homology-independent. We further observed that only 337 of 1351 090 mate-pair deletions and duplications (25%) had a homology segment (at least 75 bp and 80% identity within 500 bp of the predicted junction) available to support NAHR (24 of 80 (30%) and 313 of 1271 (25%) for events larger and smaller than 10 kb, respectively). Thus, nonhomologous mechanisms must also account for most observed constitutional CNVs regardless of size.

Despite the different size distribution, examining the exact breakpoint structure of all available sequenced aphidicolin-induced and human constitutional homology-independent CNVS did not reveal an obvious difference when comparing either aphidicolin-induced CNVs to constitutional CNVs or large CNVs to small CNVs (Supplementary Table 1 and not shown). Thus, aphidicolin-induced CNVs appear typical of all larger nonhomologous CNVs.
Much recent work has been devoted to the description of CNVs and other SVs in the human genome. We approached this problem from the specific perspective of optimizing an experimental cell system being used to probe the environmental and genetic influences on SV formation (ARLT et al. 2009). The combined data provide a strong basis for comparing the properties of induced and constitutional CNVs as well as their methods of detection.

Constitutional compared to replication stress-induced CNVs. We have described the constitutional SVs in a female of European descent who is not one of the commonly studied individuals. We nonetheless observed a very similar overall pattern of genetic changes as previous studies (CONRAD et al. 2010b; KORBEL et al. 2007; PANG et al. 2010; PARK et al. 2010; WHEELER et al. 2008), including the number of SVs observed and the relatively small ~2.5 kb median event size (Table 3, Figure 4, Supplementary Table 2). Similar to human population studies (CONRAD et al. 2010a), we infer that the majority (~75%) of SV junctions were formed by homology-independent mechanisms regardless of event size and type, a fact confirmed in the 14 new junction sequences reported (Supplementary Table 1). The potential impact of these changes on shaping inter-individual variation is evident in the 180 distinct genes having at least one exon affected by a SV (Supplementary Table 3). Importantly, the event list here contained >600 previously undocumented SVs. Although some of these may have been missed in population studies due to methodological considerations below, some likely represent low frequency population polymorphisms or private alleles in our individual.

Relative to the constitutional CNVs, aphidicolin-induced CNVs showed a much larger median size of 148 kb and exclusive utilization of homology-independent mechanisms (Figure 4 and Supplementary Table 1). Importantly, although most constitutional CNVs are smaller than
10 kb, larger homology-independent events are still readily observed in most individuals
(CONRAD et al. 2010a; CONRAD et al. 2010b; KORBEL et al. 2007; PANG et al. 2010; PARK et al. 2010; WHEELER et al. 2008). Thus, aphidicolin-induced CNVs are best described as mainly correlating to the subset of germline SVs that are both larger and homology-independent (Figure 4). Several factors might contribute to this pattern. First, we have analyzed only two aphidicolin-treated subclones and eleven observed CNVs by the higher resolution mate-pair method. Non-technical factors include that constitutional events in any human individual have been subjected to a much greater negative selection pressure than what we observe in cell culture. This almost certainly skews population polymorphisms toward smaller events inherently less likely to disturb gene function. An interesting corollary observation is that most human pathogenic CNVs are large, including nonhomologous CNVs (VISSERS et al. 2009), but once again these have mostly been discovered using low resolution arrays (COOK and SCHERER 2008; KIROV et al. 2009; MILLER et al. 2010; STANKIEWICZ and BEAUDET 2007; TAM et al. 2009; ZHANG et al. 2009).

A non-exclusive possibility is that the mechanisms contributing to formation of constitutional CNVs are more diverse than those stimulated by aphidicolin. This is certainly true for the subset of small SVs manifest as changes in VNTRs or mobile genetic elements. Beyond these special cases, we found no obvious structural signature that could distinguish the nonhomologous junctions that characterize most aphidicolin-induced and constitutive CNVs of any size, although sequence information is scant for constitutional CNVs above 30 kb. Regardless of size or source, most CNVs are variably characterized by microhomologies, blunt ends and/or short inserted sequences, features that might result from many mechanisms including NHEJ, alternative end joining, MMBIR and template switching (HASTINGS et al. 2009a; LEE et
It cannot be judged from current data whether aphidicolin induces just a subset of these mechanisms that are more diversely utilized in constitutional events, or whether some unknown feature causes the same mechanism(s) to be used for joint formation but with a tendency toward larger segment jumps when inhibited replication is the underlying stimulus.

**SNP arrays compared to mate-pair sequencing.** Overall, mate-pair sequencing had the best power for describing SVs as compared to 1M feature SNP arrays. This was apparent not only in the increased detection of *bona fide* induced and constitutional events, but also in the markedly superior descriptions of their underlying structure (Tables 2 and 3 and Figures 1 and 2). Array analysis nonetheless had its advantages, among them a much greater simplicity and lower cost. Moreover, because arrays use an entirely different basis of detection they often helped to clarify mate-pair data and uniquely detected a number of events. Indeed, the best descriptions of SVs were undoubtedly obtained when array and mate-pair data could be compared.

The failed correlations of arrays and mate-pair sequencing highlight the limitations of each technique. An overriding issue was the strong dependence on the genome locations sampled by the array design. Human population CNVs are often small (~1 kb) but could be detected by the 1M feature SNP array because these regions were deliberately over-sampled. This obviously limits arrays for detection of unknown small events sampled at the average array density, such as might be induced in our cell system or underlie a pathology of interest. Genomic regions that are difficult to sample will be further underrepresented. Indeed, the largest single category of method discrepancies was events detected by mate-pairs in regions that had too few probes for array detection (Table 3). A less obvious problem is the potential negative
impact of over-sampling on probe quality, including the use of non-SNP probes and the increased frequency of probes placed in repetitive elements. This evident in the large number of known human CNVs unambiguously detected by mate-pairs in 090 for which the SNP array data were not statistically deviant from the array average (Figure 3).

For mate-pair sequencing, a main limitation was genome coverage, easily appreciated by comparing the A3A2 and A1A1 samples (Tables 1 and 2). However, present technology is substantially advanced over that used to obtain much of our data so that a single Illumina sequencing lane now provides sufficient coverage for three or more crossing fragments per junction. More challenging is mapping mate-pairs and making event calls. False negative calls are the most likely errors as a result of CNVs contained entirely within highly repetitive genome regions where accurate mapping is all but impossible. A special class of false negative calls might occur when the reference genome itself carries a duplication common to the studied sample. Here, mate-pairs would not detect an anomaly but array methods would still reveal the increased copy number. This phenomenon might help account for the bias of array-specific calls toward copy number gains (Table 3 and Figure 3).

False positive mate-pair calls are possible but much less likely when multiple independent fragments predict an event. In this context, many factors likely contribute to the seemingly large number of mate-pair specific calls. First, all calls were relative to Build 36/hg18 and need not represent true SVs. Indeed, at least once sequenced SV mapped correctly to an alternative genome assembly. Copy number neutrality might also be consistent with a mate-pair event if there were corresponding gains and losses on different alleles for which one allele escaped mate-pair detection. Most importantly, mate-pair analysis detects all manner of SVs that
are invisible to arrays, including tandem repeat expansions and gain or loss of mobile genetic elements.

Recent descriptions of data from the 1000 Genomes Project and other high coverage human genome sequences highlight final limitations of both SNP arrays and low-depth mate-pair sequencing (Durbin et al. 2010; Pang et al. 2010; Sudmant et al. 2010). First, it is now clear that the frequency of human population CNVs increases continuously with decreasing event size and does not show the rapid drop-off below ~1 kb as observed here (Figure 4), which for SNP arrays and mate-pair sequencing reflect the limitations of probe density and fragment size distribution, respectively. Further, the new ability to compare deep sequencing of multiple human individuals has established the high frequency of human CNVs in repetitive genome segments such as segmental duplications (Sudmant et al. 2010), regions inherently difficult to study by either microarray or low depth mate-pair sequencing. Nonetheless, results here demonstrate the robust detection of most induced CNVs in an experimental setting by either SNP arrays or low-depth mate-pair sequencing.
Acknowledgements

This work was supported by grants NIH RCI-ES018672 to TEW and TWG and a research grant from the March of Dimes Foundation to TWG. We thank Drs. Jun Li and Steve Qin for many helpful discussions regarding mate-pair sequencing in the early phases of this project.
References


COUNTER, C. M., W. C. HAHN, W. WEI, S. D. CADDLE, R. L. BEIJERSBERGEN et al., 1998

CRADDOCK, N., M. E. HURLES, N. CARDIN, R. D. PEARSON, V. PLAGNOL et al., 2010 Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464: 713-720.


PANG, A. W., J. R. MACDONALD, D. PINTO, J. WEI, M. A. RAFIQ et al., 2010 Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11: R52.


SEBAT, J., B. LAKSHMI, J. TROGE, J. ALEXANDER, J. YOUNG et al., 2004 Large-scale copy number polymorphism in the human genome. Science 305: 525-528.


TAM, G. W., R. REDON, N. P. CARTER and S. G. GRANT, 2009 The role of DNA copy number variation in schizophrenia. Biol Psychiatry 66: 1005-1012.


WANG, K., M. LI, D. HADLEY, R. LIU, J. GLESSNER et al., 2007 PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.

Table 1. Mate-pair sample summary

<table>
<thead>
<tr>
<th></th>
<th>Untreated</th>
<th>aphidicolin treated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>090</td>
<td>A3A2</td>
</tr>
<tr>
<td>lanes</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>libraries</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>read length</td>
<td>35-39</td>
<td>35-36</td>
</tr>
<tr>
<td>mean fragment length (kb)</td>
<td>2.9-4.4</td>
<td>2.8-3.4</td>
</tr>
<tr>
<td>SD fragment length (kb)</td>
<td>0.22-0.42</td>
<td>0.24-0.55</td>
</tr>
<tr>
<td>mate-pairs</td>
<td>5.9 x 10^7</td>
<td>2.9 x 10^7</td>
</tr>
<tr>
<td>unique mate-pairs</td>
<td>4.4 x 10^7 (74%)</td>
<td>2.1 x 10^7 (74%)</td>
</tr>
<tr>
<td>fraction mapped reads</td>
<td>84%</td>
<td>81%</td>
</tr>
<tr>
<td>~3 kb fragments</td>
<td>2.4 x 10^7 (56%)</td>
<td>8.3 x 10^6 (39%)</td>
</tr>
<tr>
<td>haploid fragment coverage</td>
<td>24.9</td>
<td>8.7</td>
</tr>
<tr>
<td>diploid fragment coverage</td>
<td>12.4</td>
<td>4.3</td>
</tr>
<tr>
<td>haploid read coverage</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>diploid read coverage</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Table 2. Aphidicolin-induced CNVs

<table>
<thead>
<tr>
<th>Sample/Type</th>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Size  (kb)</th>
<th>aCGH log2R</th>
<th>SNP array log2R</th>
<th>B frequency</th>
<th>Mate-pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3A2 Deletion</td>
<td>7</td>
<td>61,379,057</td>
<td>61,440,507</td>
<td>61</td>
<td>-0.28 (26) **</td>
<td>0.04 (11)</td>
<td>(0)</td>
<td>0</td>
</tr>
<tr>
<td>A3A2 Deletion</td>
<td>11</td>
<td>6,476,303</td>
<td>6,564,800</td>
<td>88</td>
<td>-0.51 (73) ***</td>
<td>-0.33 (51) ****</td>
<td>1.00 (1)</td>
<td>6</td>
</tr>
<tr>
<td>A3A2 Deletion</td>
<td>15</td>
<td>69,469,199</td>
<td>69,499,819</td>
<td>31</td>
<td>-0.30 (23) ***</td>
<td>-0.33 (11) ****</td>
<td>1.00 (4) ***</td>
<td>6</td>
</tr>
<tr>
<td>A3A2 Deletion</td>
<td>18</td>
<td>7,630,270</td>
<td>7,864,089</td>
<td>234</td>
<td>-0.31 (183) ****</td>
<td>-0.38 (85) ****</td>
<td>1.00 (22) ****</td>
<td>5</td>
</tr>
<tr>
<td>Bi-allelic deletion</td>
<td>3</td>
<td>117,988,930</td>
<td>118,751,078</td>
<td>762</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Bi-allelic deletion</td>
<td>118,056,525</td>
<td>118,673,931</td>
<td>617</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>A1A1 Deletion</td>
<td>13</td>
<td>93,393,314</td>
<td>93,545,246</td>
<td>152</td>
<td>-0.24 (129) ****</td>
<td>nd</td>
<td>nd</td>
<td>0</td>
</tr>
<tr>
<td>Duplication</td>
<td>1</td>
<td>172,806,440</td>
<td>172,861,505</td>
<td>55</td>
<td>0.27 (23) **</td>
<td>0.24 (7)</td>
<td>(0)</td>
<td>2</td>
</tr>
<tr>
<td>Duplication</td>
<td>9</td>
<td>130,303,541</td>
<td>130,699,329</td>
<td>396</td>
<td>0.20 (232) ****</td>
<td>0.16 (126) ****</td>
<td>0.62 (43) ****</td>
<td>5</td>
</tr>
<tr>
<td>Duplication</td>
<td>10</td>
<td>111,820,163</td>
<td>114,253,039</td>
<td>2,433</td>
<td>0.29 (1971) ****</td>
<td>0.19 (879) ****</td>
<td>0.64 (303) ****</td>
<td>9</td>
</tr>
<tr>
<td>Inversion</td>
<td>4</td>
<td>123,973,228</td>
<td>124,044,943</td>
<td>72</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Inversion</td>
<td>123,999,958</td>
<td>124,068,805</td>
<td>69</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>Inversion</td>
<td>123,973,228</td>
<td>123,999,958</td>
<td>27</td>
<td>0.26 (24) ****</td>
<td>0.15 (14) *</td>
<td>0.62 (5) ****</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Inversion</td>
<td>124,044,943</td>
<td>124,068,805</td>
<td>24</td>
<td>0.28 (20) ***</td>
<td>0.15 (14) **</td>
<td>0.63 (2)</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A1A1 Duplication</td>
<td>1</td>
<td>29,268,937</td>
<td>29,423,288</td>
<td>154</td>
<td>0.12 (101) ****</td>
<td>nd</td>
<td>nd</td>
<td>2</td>
</tr>
<tr>
<td>A1A1 Duplication</td>
<td>81,847,144</td>
<td>81,882,508</td>
<td>35</td>
<td>0.15 (28) ***</td>
<td>nd</td>
<td>nd</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Columns log2R and B frequency show the average log2 of the sample to reference ratio or the average B allele frequency, respectively, followed by the number of probes in parentheses, followed by a significance value calculated using the t-test comparing all event probes to the mean value for the array: ****, p < 0.00001; ***, p < 0.0001; **, p < 0.001; *, p < 0.01.

Boldface type indicates that the event was called by the detection algorithm. For the complex A3A2 bi-allelic deletion (Chr3) and inversion with flanking duplications (Chr4), top rows give the anomalous junctions and crossing mate-pairs, while rows below give the CNV spans created by these junctions. ‘nd’, not done; ‘-’, not applicable.
Table 3. Constitutional CNVs

<table>
<thead>
<tr>
<th>detection method detected by other method</th>
<th>SNP array</th>
<th>Mate-pair</th>
<th>&lt;5 probes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>A gains</td>
<td>4</td>
<td>98</td>
<td>4</td>
</tr>
<tr>
<td>B losses</td>
<td>326</td>
<td>120</td>
<td>322</td>
</tr>
<tr>
<td>C total events</td>
<td>330</td>
<td>218</td>
<td>326</td>
</tr>
<tr>
<td>D false negative calls</td>
<td>-</td>
<td>1 (0.5%)</td>
<td>-</td>
</tr>
<tr>
<td>E matches known CNV</td>
<td>326 (99%)</td>
<td>202 (93%)</td>
<td>319 (98%)</td>
</tr>
<tr>
<td>F median size (kb)</td>
<td>1.8</td>
<td>1.9</td>
<td>3.1</td>
</tr>
<tr>
<td>G median fold probe density vs. array</td>
<td>20.8</td>
<td>18.4</td>
<td>12.2</td>
</tr>
<tr>
<td>H median [log2R]</td>
<td>0.48</td>
<td>0.33</td>
<td>0.45</td>
</tr>
<tr>
<td>I median [Z]</td>
<td>13.4</td>
<td>8.9</td>
<td>13.1</td>
</tr>
<tr>
<td>J median mate-pairs</td>
<td>-</td>
<td>-</td>
<td>13</td>
</tr>
<tr>
<td>K median fraction mismatched bases</td>
<td>-</td>
<td>-</td>
<td>0.007</td>
</tr>
<tr>
<td>L created by inversion</td>
<td>1 (0.3%)</td>
<td>?</td>
<td>1 (0.3%)</td>
</tr>
<tr>
<td>M tandem repeat (losses only)</td>
<td>112 (34%)</td>
<td>82 (68%)</td>
<td>170 (53%)</td>
</tr>
</tbody>
</table>
Figure Legends

**Figure 1. Examples of aphidicolin-induced CNVs.** Intensity data (log2R) from SNP arrays are at the top of each panel with mate-pairs of the indicated type below. Array data are averaged over a five probe window. Individual mate-pairs are drawn as bars connecting the two reads as mapped to the reference genome. Blue color denotes the reference 090 sample, magenta the aphidicolin-treated A3A2 sample. Chromosome coordinates are in Mb. (A) An induced deletion on Chr15 that overlaps a constitutional CNV. The orange box denotes the position of a known human population CNV. (B) A bi-allelic deletion on Chr3. (C) A complex inversion on Chr4 that creates two associated copy number gains. One potential underlying allele is drawn as an event chain similar to Supplementary Figure 7.

**Figure 2. Examples of constitutional CNVs.** Major classes of constitutional CNVs observed in 090 are depicted similarly to Figure 1. Orange boxes highlight known human CNVs. (A) An 8.2 kb deletion on Chr2 evident in both array and mate-pair data. (B) A 6.5 kb gain on Chr18 evident in array but not mate-pair data. (C) A 6.5 kb deletion on Chr2 evident in mate-pair but not array data. (D) A 10.2 kb deletion on Chr12 evident in mate-pair data in a region with no array probes. A trace shows the relative density of normal ~3 kb mate-pair mappings, and a cyan box highlights a HERVK element in the reference genome.

**Figure 3. Array run correlation by CNV category.** Scatter plots show the correlation of two different SNP array runs of 090, where the top panel shows all data and the bottom panel is magnified to the region near the origin, corresponding to predicted heterozygous events. Green circles are the array log2R values of CNV regions detected by both array and mate-pairs,
blue circles are CNVs detected only by array, and red circles are CNVs detected only by mate-pairs for which array probes were present.

**Figure 4. Size distribution of SVs.** Histograms show the fraction of events by event size within each of four groups. “Const. array” are constitutional CNVs observed in 090D2 by SNP array (blue circles). “Const. mate-pair” are constitutional SVs observed in 090 and/or 090D2 by mate-pair sequencing (red circles). “APH array” are new CNVs observed in a series of 23 independent aphidicolin-treated subclones of 090 or 090D2 by SNP array and/or aCGH (open circles). “APH mate-pair” are new CNVs observed in aphidicolin-treated subclones A3A2 and A1A1 by mate-pair sequencing (black circles). Also shown is a trace of the CNVs detected by microarray from two published compendia (green circles) (CONRAD et al. 2010b; PARK et al. 2010). Numbers in parentheses indicate the total number of events in each group. Array traces combine both gains and losses, while mate-pair traces combine deletion and duplication sets.
Figure 3

- Blue dots: array only
- Green dots: array + mate-pair
- Red dots: mate-pair only

The graphs show the log2R values for two different arrays: array #1 and array #2. The scatter plots compare the values for each condition.
Figure 4

- Const. array (548)
- Const. mate-pair (1351)
- APH array (93)
- APH mate-pair (9)
- Human CNVs (32,748)