SUPPLEMENTARY METHODS

Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants

Dalila Pinto[1], Katayoon Darvishi[2], Xinghua Shi [2], Diana Rajan[3], Diane Rigler[3], Tom Fitzgerald[3], Anath C. Lionel[1], Bhooma Thiruvahindrapduram[1], Jeffrey R. MacDonald[1], Ryan Mills[2], Aparna Prasad[1], Kristin Noonan[2,4], Susan Gribble[3], Elena Prigmore[3], Patricia K. Donahoe[4], Richard S. Smith [2], Ji Hyeon Park[2,7], Matthew E. Hurles[3], Nigel P. Carter[3], Charles Lee[2], Stephen W. Scherer[1,5], Lars Feuk[6]

[1]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada. [2]Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [3]Wellcome Trust, Sanger Institute, Hinxton, Cambridge, UK. [4]Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [5]McLaughlin Centre and Department of Molecular Genetics, University of Toronto, Toronto, Canada. [6]Department of Immunology, Genetics and Pathology, SciLifeLab Uppsala, Rudbeck Laboratory, Uppsala University, Sweden. [7]Present address: Department of Obstetrics and Gynecology, Pochon CHA University College of Medicine, Seoul, South Korea.

March 17, 2011

Nature Biotechnology: doi:10.1038/nbt.1852

1. DNA samples

All DNA used for the experiments was purchased from Coriell Cell Repositories (Camden, New Jersey), and originated from lymphoblastoid cell lines. For samples processed at The Centre for Applied Genomics at the Hospital for Sick Children in Toronto (TCAG) and at Harvard Medical School (HMS), the same lot numbers of DNA were used for all experiments, while lot numbers differed for a subset of the samples processed at the Wellcome Trust Sanger Institute (WTSI). The samples were NA10851 (46, XY; Lot 6), NA12239 (46, XX; Lot A1), NA18517 (46, XX; Lot B1), NA18576 (46, XX; TCAG and HMS Lot 1, WTSI Lot 3) and NA18980 (46, XX; Lot 1) from the HapMap collection, and NA15510 (46, XX; TCAG and HMS Lot 4, WTSI Lot 7) from the Polymorphism Discovery Resource.

2. Microarrays

In total, eleven different microarray platforms were used in this study, representing both CGH arrays and SNP arrays. The CGH arrays included in the study (laboratories where experiments were performed listed in parenthesis) are the Human Whole-Genome TilePath (WGTP) BAC array (WTSI), Agilent 244K CGH Array (TCAG), Agilent 2x244K CGH Array (TCAG and WTSI), NimbleGen CGH Array 3x720K (WTSI) and NimbleGen HD2 CGH Array 2.1M (HMS and WTSI). The SNP arrays included in the study were the Affymetrix 500K GeneChip array set (TCAG), Affymetrix 6.0 SNP Array (TCAG and WTSI), Illumina HumanHap 650Y BeadChip (TCAG), Illumina 1M BeadChip (HMS and TCAG), Illumina 660W-Quad BeadChip (HMS and TCAG), and Illumina Omni (HMS and TCAG). For all CGH platforms sample NA10851 was used as reference DNA in competitive hybridization experiments. For these platforms, one set of triplicates represented self-self hybridizations with the NA10851 sample.

CGH arrays

For Agilent and NimbleGen platforms the first two replicate experiments were run with the test sample labeled with Cy5 and the reference labeled with Cy3. The third triplicate was run with the template labeled with Cy3 and the reference DNA labeled with Cy5 (dye-swap strategy). This was done to enable an evaluation of the dye-swap strategy compared to single hybridization. For the BAC array, each replicate represents the average of two hybridizations between target and reference sample (see below). All CGH experiments (five female HapMap samples NA18576, NA18980, NA15510, NA12239 and NA18517) were run as dual color competitive hybridizations using the male sample NA10851 as reference DNA. Details of the protocol used for each CGH platform are outlined below.

BAC array

The BAC array used was the human Whole-Genome TilePath (WGTP) array[1, 2]. After addition of new clones to fill coverage gaps, the array includes 29,043 large-insert clones, covering >97% of the euchromatic portion of the human genome reference sequence. Microarray hybridizations were performed as previously described[1]. All experiments were performed in duplicate, with each replicate representing the average of two hybridizations. Array images were acquired using a 5-µm resolution Agilent Technologies G2505A laser scanner. Fluorescence intensities and log2 ratio values were extracted using BlueFuse software (BlueGnome Ltd). Fusion of duplicate experiments and subsequent analyses were performed using custom Perl scripts as previously described [3], with one additional step: After image quantification, log2 ratio calculation and block median normalization, we have introduced a G+C correction, which consists of normalizing the log2 ratios of each clone using the content in G+C percent of that clone. This correction was performed by linear regression using the module LineFit in Perl (http://search.cpan.org/~randerson/Statistics-LineFit-0.07) and applied on each individual profile before the fusion of duplicate results. Copy number variable segments were detected using the CNVfinder algorithm[3].

Agilent arrays

Two different Agilent arrays were included in the study, the standard 244K CGH array (AMADID G4423A) and the 2x244K array set (AMADID 018897 and 018898). The 244K array consists of 236,381 distinct 60-mer oligonucleotide probes, plus 1,000 triplicates (i.e. two additional copies of 1,000 probes) and an additional 5,045 quality control features, resulting in an 8.9 kb overall median spacing (**Supplementary Fig. 1 and Tables 1 and 2**). The 2x244K array set contains oligonucleotide probes (60-mer) printed on two slides containing 244K probes each. Probes for chromosomes 1-8 are assigned to array 018897, and probes for chromosomes 9-22, X and Y are assigned to array 018898 (NCBI build v.36, hg18), where 10,000 probes are repeated on both arrays for normalization purposes[4]. There are a total of 476,609 probes (i.e.

non-expanded content), where 462,607 are unique probes plus ~14,000 replicates comprised of copies of 10,000 unique probes (**Supplementary Fig. 1 and Tables 1 and 2**). Both arrays have non-unique probes, whose sequences map to multiple locations in the genome (e.g. probes that match segmental duplication regions[5]), which can be used by the DNA Analytics to help identifying contiguous aberrant regions. Such probes are not used by third-party software that makes no use of the Agilent design files. In contrast when using the Agilent DNA analytics software, the data from the two arrays is fused/merged and treated as a single dataset, where the log2 ratios of each measured probe in the non-expanded content are assigned to all perfect genomic matches of this sequence in the NCBI v.36 build of the reference genome resulting in probes assigned to a total of 500,974 genomic locations (i.e. expanded array content).

For each sample processed at TCAG, 1.9µg of DNA was used, whereas 300ng of starting DNA was used at WTSI. At both sites, DNA was labeled using the Invitrogen BioPrime CGH labeling kit (cat. no. 18094011, Invitrogen, Carlsbad, CA) according to the manufacturer’s protocol. Onto each slide, the amount of combined total DNA hybridized was approximately 5.5ug. Purification, hybridization and washing of samples was carried out according to the 40-hour manual protocol as described by the manufacturer (Agilent Technologies, Santa Clara, CA). Arrays were scanned at a 5 micron resolution in an Agilent scanner and processed with Agilent Feature Extraction v9.5.1.1 (Agilent Technologies, Santa Clara, CA). Samples were required to show derivative Log Ratio spread (DLRS) values <0.3, an indicator of noise for a given sample that corresponds to the spread of log ratio differences between consecutive probes along all chromosomes. CNV detection was performed using two different algorithms; the Agilent’s recommended ADM-2 as part of the DNA Analytics package, as well as the Nexus rank segmentation software.

NimbleGen arrays

A total of 4ug of the 6 samples were analyzed on NimbleGen HD2 arrays in triplicate. 1 ug reactions of each sample were labeled in quadruplicate according to manufacturer’s instructions using cy3- and cy5-Random Nonamers. Samples were purified, hybridized, and washed according to manufacturer’s instructions. Arrays were scanned on a DNA Microarray Scanner with Surescan High-Resolution Technology (G2565CA, Agilent Technologies, Santa Clara, CA) at 3µm resolution. Hybridization was carried out on the Maui Hybridization System (BioMicro

Systems Inc., Salt Lake City, UT). Images were aligned using Nimblescan v2.4 (Roche NimbleGen, Madison, WI). Normalization was performed using the segMNT algorithm at a 1X window averaging and minimum segment difference score of 0.3 (Nimblescan, Roche NimbleGen, Madison, WI). The Nexus rank segmentation software was used for CNV discovery.

SNP array platforms

For all SNP arrays, each batch of six samples were processed as independent replicate experiments. These experiments thus differ from the CGH experiments described above in that no reference sample is used and single color absolute intensity data is used for result interpretation.

Affymetrix arrays

DNA was independently labelled, and hybridized to the Affymetrix arrays according to standard protocols as provided by the manufacturer. The Affymetrix 500K (250K NspI and 250K StyI) GeneChip array set was processed only at the TCAG site according to the manufacturer's protocol, and have been described previously, including QC procedures (contrast QC > 0.4 and inter-quartile range (IQR) < 0.4)[6]. All 250K-StyI arrays failed QC[6], therefore further analysis for the 500K array set was restricted to the 250K-NspI arrays (**Supplementary Tables 1 and 2**). CNV analysis used dCHIP[7] as well as the CN4 algorithm in the Affymetrix Genotyping Console and Partek Genomics Suite[8].

For the Affymetrix 6.0 arrays, two sets of triplicates were genotyped in two different genotyping centres, the TCAG and WTSI respectively, according to standard protocols as provided by the manufacturer and described elsewhere[9-11]. At TCAG in Toronto, the main processing protocol was used, with one exception in the PCR purification step, in which samples were purified individually using YM30 filter columns (Millipore) instead of using plates and bead purification. Arrays were washed on the Affymetrix fluidics stations and scanned using the Gene Chip Scanner 3000 7G. At WTSI, the Affymetrix 6.0 arrays were hybridized following the low throughput assay developed for cytogenetics labs for processing of less than 10 samples. All steps were performed as outlined in the manufacturer's protocol - Affymetrix Cytogenetics Copy Number Assay User Guide.

For array quality control, CEL files were processed using modules from the Affymetrix power tools (APT v.1.12), and genotypes called using Birdseed v.2.0[10]. Samples passing the recommended values for contrast QC > 0.4 and median of the absolute intensity values of all pairwise differences (MAPD) QC < 0.3, were further analyzed using five different CNV calling methods; the recommended CN5 algorithm in the Affymetrix Genotyping Console plus four additional methods, Birdsuite[10], dCHIP[7], iPattern[12, 13] and Partek Genomics Suite[8]. At the TCAG site, a batch of 132 samples genotyped together with the 18 study samples in the same time period was used as an internal reference for all these programs, in order to minimize batch effects. No internal reference library was available at the WTSI site, and the Affymetrix supplied reference library based on HapMap samples was therefore used. We also explored whether the samples genotyped at TCAG could work as a reference for the samples genotyped at WTSI by computing pair-wise Pearson correlations of median normalized intensities between sets of samples genotyped in different labs and/or same lab at different times (**Supplementary Fig. 2**). We found that the TCAG samples did not correlate well with the samples genotyped at WTSI, while the Affymetrix supplied reference library showed borderline acceptable QC values. Besides sample-specific variability, we observed that systematic effects between a sample and the reference can greatly inflate per-chip variability estimates, and consequently the ability to make reliable loss or gain calls. Overall, we observed that MAPD values < 0.3 corresponded to approximately Pearson correlations > 0.88.

Illumina arrays

The six samples were analyzed in the same reagent batch and experiments were performed in triplicate. Samples were processed using the Illumina manufacturer's recommended protocol with no modifications for both the Infinium II and Infinium HD assays (Illumina, San Diego, CA). All assay protocols featured single tube PCR-free amplification, all of which were processed in a semi-automated production environment. Briefly, 750ng of DNA (for the single-sample Hap650Y and Hap1M arrays) or 200ng DNA (for the multi-sample format 660W-Quad and Omni-Quad arrays) was denatured and neutralized before amplification. The denatured DNA was isothermally amplified in an overnight step, and the amplified product was then enzymatically fragmented (i.e. end-point fragmentation), precipitated with isopropanol, collected by centrifugation at 4ºC, and resuspended in hybridization buffer. All beadchips were prepared for hybridization in a capillary flow-through chamber. Samples were applied to beadchips and the loaded beadchips were incubated overnight in the Illumina Hybridization Oven. Unhybridized and non-specifically hybridized DNA was washed away, and beadchips were prepared for staining and extension. The beadchips were scanned on the Illumina BeadArray Reader using default settings, and intra-chip normalization was performed using the Illumina's GenomeStudio software v.1.0.1 with a GenCall cutoff of 0.1 and call rate cutoff of 98%. Built-in controls, both sample independent (including staining controls, extension controls, target removal controls, and hybridization controls) and sample-dependent (including stringency controls, non-specific binding controls, and non-polymorphic controls), were inspected to assess the quality of the experiment. For CNV detection using different algorithms, various measurements were exported directly from GenomeStudio: the intra-chip normalized X and Y intensity values from the A and B allele-specific probes respectively, and two other measurements derived from X and Y after normalization to the reference population, i.e. the Log R ratios (LRR) and B allele frequency (BAF) values. For the algorithms that used LRR, representing the total signal intensity, and BAF, representing the allelic balance, the Illumina's cluster file made of >120 HapMap samples was used to generate intensities and genotypes using GenomeStudio. For much larger sample sets that cannot be hybridized in one batch, we recommend to use a cluster file made of samples internal to the project, in order to minimize batch effects. For the purpose of CNV discovery, a series of QC measures were applied to all the assayed Illumina array types as previously described[13]. Specifically, arrays were removed if their call rate was < 98%, standard deviation (SD) of autosomal LRR values was >0.27, SD for BAF values was >0.13 (i.e., allelic ratios within the 0.25 to 0.75 ranges), or cross sample-batch normalized ratio standard deviation >0.27 (i.e. sample-batch level QC). Samples passing QC were further analyzed using the Illumina's CNV Partition plus three additional CNV methods, iPattern[12, 13], QuantiSNP[14] and PennCNV[15].

3. Analytical tools and statistical methods

For each array platform, we have used the manufacturer's recommended software and at least one additional known CNV detection algorithm when possible. CNV discovery was restricted to autosomes due to difficulties normalizing the LRR and BAF in sample collections with both males and females and because some CNV detection methods (such as Nexus) cannot handle sex chromosomes correctly. All CNV analyses were performed using the original array coordinates based on the human genome assembly NCBI v.36 (hg18).

For CNV detection, samples were required to pass initial platform-specific quality control (QC) as described above, as well as additional CNV-specific QC filters that were systematically applied to all CNV datasets generated by each algorithm as previously described[13]. We excluded CNVs if they resided in regions of extreme GC content (>70%) or if they were within centromere proximal cytobands. We removed samples that were outliers with respect to: (1) excessive number of CNVs detected as defined as exceeding the mean number of CNVs across all samples plus 3 standard deviations; (2) excessive aggregate length of CNVs as these likely correspond to large karyotypic chromosome abnormalities, or cell line artifacts[16]. We note that no samples were excluded by this last criterion. All CNVs identified by any algorithm with sizes larger than 1 Mb were inspected manually. Manual curation was used to exclude potential false positives due to whole chromosome aneuploidies, potential cell line mosaicism and artifacts. For all platforms, we kept CNVs passing quality control with ≥ 1 Kb length and spanning at least 5 probes, except for CNVs detected with BAC arrays where all CNVs with at least one clone were considered.

For CNV discovery using SNP arrays or CNV-SNP arrays there are various tools that differ by varying degrees in the statistical models employed, input data (absolute intensity values vs. log R ratios), use of genotype information, use of a reference baseline, assessment of input data variability, as well as reporting of CNV boundaries. The platforms for which the CNV detection methods were originally tailored may also differ. For instance, dCHIP[7], the Affymetrix CN4 and CN5 and Birdsuite[10] were originally developed for Affymetrix data, and PennCNV for Illumina data, though ongoing work promises that at least some of these tools could also analyze Illumina or Affymetrix intensity data, respectively. iPattern was developed for Illumina and Affymetrix arrays[12, 13]. In aCGH there are no allele-specific probes, thus the data consists of a one-dimensional series of intensity measurements that are analogous to the LRR measurements, and consequently the existing algorithms search for segments or partitions of the genome that show homogeneous consecutive deflections in mean intensity signals. Therefore, any method developed for aCGH should in principle equally work for Agilent or Nimblegen platforms.

A description of the CNV methods used in this study is given below.

CNV algorithms

ADM-2 (DNA Analytics v4.0, Agilent Technologies)

Array data from 244K and 2x244 K Agilent was analyzed with DNA Analytics v.4.0.85 (Agilent Technologies, Santa Clara, CA) using the built-in ”Aberration Detection Method-2” (ADM-2) algorithm. This algorithm incorporates quality information about each probe measurement while it searches for intervals in which a Z-score based on the average weighted log ratio of the sample and reference channels exceeds a user-specified threshold[4, 17]. CNV analysis for the 2x244K array set was performed using the same parameters as described before[4] with an ADM-2 threshold of 5.0 for sensitivity, a minimum absolute average log2 ratio in called intervals of 0.25, and a minimum of 5 consecutive probes per region. The two design files of the 2x244K array were combined together into one larger array using the fused option and the non-unique probes were expanded to help identify contiguous copy number variable regions (i.e. replicate probes were combined to increase the confidence of detected regions). A nested filter of 2 was used, with subsequent filtering of child-overlapping calls using a custom-script, which maintained the call with the maximum outside boundaries.

BirdSuite (v.1.5.3)

BirdSuite is a suite of methods originally developed to detect known common copy number polymorphism (CNP) based on prior knowledge[11], as well as to discover rare CNVs, from Affymetrix SNP 6.0 array data. To do this, it incorporates two main methods; the ”Canary” and the ”Birdseye” algorithms. The ”Canary” algorithm assigns copy number across regions of known common CNPs, obtained from a reference set with 1,292 autosomal CNPs with a minor allele frequency >1% created on the basis of 263 HapMap samples genotyped with Affy6[11]. The Birdseye algorithm uses a hidden Markov model (HMM) approach to find regions of variable copy number in a sample. For the HMM, the hidden state is the true copy number of the individual’s genome and the observed states are the normalized intensity measurements (means with their estimated variance) of each array probe. CNV calls from the Canary and Birdseye algorithms were collated for each sample, and kept as long as they met the following criteria: i) Birdseye calls with a $\log_{10}$ of odds (LOD) score (Odds Ratio) greater than or equal to 10 (corresponding to an approximate FDR of ~5%)[10], ii) Birdseye calls with CN states other than 2 were retained; iii) Canary CNP calls with CN states different from the population mode were retained. Since the original population mode was defined based in only 263 samples[11], here we have redefined the population mode at each of the 1,292 autosomal CNPs based on a larger sample set of 2,357 control individuals genotyped with Affy6 (Lionel A, *in preparation*)[18]. Specifically, we run Birdsuite on two separate large Caucasian control cohorts, with 1,234 and 1,123 samples respectively. For each control set, Birdsuite generates automatically a matrix CNP id x sample id (row x column). Then for each CNP id, we chose as mode the most frequent CN state occurring across all samples from both cohorts (ie. N=2,357). A final matrix with CNP states was generated after redefining the CNP mode in such a way that a given CNP call for a sample was retained only if its copy number state was other than the new population mode. CNP calls with confidence score < 0.1 in the 2,357 control samples were excluded.

CNVFinder

CNVFinder was originally developed for the whole genome tile path BAC array created at the Welcome Trust Sanger Institute[3]. It uses a dynamic, multiple-threshold based approach to allow robust classification of copy number changes in data of varying qualities. This algorithm makes two main assumptions i) that the majority of data points are normally distributed around a log2 ratio of zero, and ii) that data points falling outside of the centralized log2 ratio distribution are representative of a difference in copy number between test and reference genome. The thresholds used in CNVFinder are dynamically adjusted using a noise parameter they termed SDe. This parameter is equal to the 68th percentile of the absolute log2 ratio distribution and is calculated on a chromosome-by-chromosome basis during copy number detection. The SDe is an estimation of the standard deviation that is relatively insensitive to outliers and thus, as long as the two main assumptions hold true, can be used as a reliable measure of experimental variability.

CNV Partition (v2.3.4, Illumina, Inc.)

Illumina's built-in CNV segmentation algorithm uses a recursive partitioning approach that is compatible with GenomeStudio, and was used with default parameters.

dCHIP

Affy CEL files were normalized using the built-in invariant set probe selection method and running the median smoothing method. It is a model-based method that uses an HMM to examine the summarized intensities to identify duplications and deletions where the observations are the normalized intensity ratios that are assumed to be distributed according to a Student *t* distribution[7].

Genotyping Console (CN4 for 500K array set, CN5 for Affy6, Affymetrix Inc)

After using the APT routines to process CEL files and the Birdseed to call genotypes as described above, we used the Genotyping Console (GTC v.3.0.2) to detect CNVs from either the Affy 500K array set or Affy6 array for samples that passed initial QCs. The default parameters in CN4 or CN5 algorithms were used (ie. filtering out CNVs with < 1Kb size and < 5 probes). The CN4 algorithm in GTC typically combines the calls detected for the two 250K arrays and only reports those calls supported by both arrays for each sample. However, because the 250K-Sty arrays failed initial QC and were excluded from the analyses, we only report on the CNV calls detected with the 250K-Nsp array.

iPattern

iPattern implements a non-parametric, density-based, clustering model that integrates intensity data across samples to assign individual samples to distinct copy number states, and it is applicable to multiple array platforms (originally developed for Affy6 as well as Illumina 1M)[12, 13]. iPattern data pre-processing produces a single one-dimensional summary of the relative intensity of test to internal reference samples. Specifically, data pre-processing evaluates the background signal-to-noise ratio for each batch of tested samples, and outliers from the standard deviation of the sample batch are removed. Normalization of chromosome X probes is performed separately for males and females before CNV calling of this chromosome. A two-stage analytical framework is then used to identify CNV regions, with a moving window-based approach followed by secondary boundary refinement. The largest cluster of unrelated samples is dynamically chosen as reference, and samples with higher or lower intensities are assigned as relative CNV gains or losses. Simulation studies were carried out on synthetic data derived from various X chromosome copies, and clustering parameters were chosen to maximize sensitivity while setting a genome-wide false discovery rate (FDR) to 5%. CNV lengths are calculated based on the distance between the first and last array probes internal to the variant.

Nexus Copy Number software

Nexus Copy Number software (v.4.1, BioDiscovery, Inc., El Segundo, CA) uses a rank segmentation algorithm for analyzing output files of all array CGH platforms. The rank segmentation algorithm is similar to the circular binary segmentation (CBS) algorithm[26], with the major difference that rank segmentation uses the probe's log-ratio rank as opposed to the actual log-ratio value (personal communication by Soheil Shams, BioDiscovery). The Nexus default settings were used for all aCGH, consisting of a significance threshold of $1x10^{-6}$ and a minimum number of probes per segment of 5, except for the BAC arrays where the significance threshold was set to 0.05 and a minimum number of probes per segment of 1. Nexus generates quality control (QC) scores for experimental results based on the statistical variance of the probe-to-probe log ratios. A small number (3%) of the outliers are excluded from this calculation to remove changes due to true CNVs. This QC value can be indicative of the quality of the sample and experiment, with lower QC scores indicating better quality results. A QC score less than 0.15 is considered the cut-off for best quality results for these arrays. A score between 0.15-0.25 is considered borderline. We observed that Nexus Rank overestimated the CNV size by reporting half of the distance to the next probe instead of mapping the CNV start and end coordinates to probe positions, and used a custom script to adjust the reported array boundaries. Specifically, for the BAC array we used a similar approach to CNV finder to report CNV boundaries[3] - for example, if a CNV call is supported by three BAC clones, we reported the start position of the first clone and the last position of the third clone. For all the other arrays, we reported the probe positions that corresponded to the centre of the probe.

Partek Genomics Suite (Partek, Inc.)

The hidden Markov model (HMM) region detection method from Partek Genomics Suite version 6.4 was used to obtain CNV calls from the log intensities of the arrays. Default parameters were used for the HMM: Max probability = 0.98, Sigma = 1, Genomic decay = 0, CNV states to detect deletions (CN state equal to 0.1 or 1) and duplications (CN equal to 3, 4 or 5).

PennCNV

The PennCNV algorithm uses combined information from LRR, corresponding to the total signal intensity, and BAF that corresponds to the allelic intensity ratio at each SNP marker, and an HMM approach to infer CNVs (for details see Wang *et al.*[15]). It has been originally developed for Illumina arrays, though a tool for converting the raw CEL Affymetrix intensity data in LRR and BAF has been made available recently, which are the typical input data for PennCNV. Default settings were used in the analysis.

QuantiSNP

QuantiSNP uses an objective Bayes (OB) hidden Markov model (HMM) approach for CNV calling in which OB measures are used to set hyperparameters (false positive rates) and the copy number state is inferred with an HMM that analyses each chromosome of each sample separately (for details see Colella *et al.*[14]). It also uses a fixed rate of heterozygosity for each SNP. CNVs were filtered for log Bayes factor of 15 corresponding to an FDR of 5%[14].

Receiver operating characteristic (ROC) curves and calculation of the area under the curve (AUC)

Using a similar strategy as in Matsuzaki et al.[19], we examined how well experiments on a platform can be applied to detect changes in copy number by comparing the intensity ratios of chromosome X and chromosome 2 for male and female samples (**Supplementary Fig. 3**). For each experimental condition, a ROC curve was generated to determine how well separated are the distribution of intensities for chromosome 2 and for chromosome X. In other words, it evaluates how well each platform is able to detect single-copy losses, and the AUC value is calculated to reflect the “goodness” of separation. The workflow of this analysis for each experiment was as follows: 1) Experimental data for a male sample (NA10851) and a female sample (NA15510) were obtained and signal intensities of all probes on chromosome X and chromosome 2 were extracted from the data for the male sample and the female sample respectively; 2) For each of the two chromosomes in each sample, a subset with 10% of the probes was randomly selected; 3) For each probe in the random subset of probes, a log2 ratio was obtained by calculating the median signal of the male sample and the female sample, dividing the median male signal by the median female signal and finally taking the log2 of this

Nature Biotechnology: doi:10.1038/nbt.1852

ratio; 4) The range of log2 ratios for probes on chromosome X were split into 100 equal size-bins, and we then counted how many of chr2 log2 ratios fit into those chrX bins. The values of these bins were considered as false positive rates; 5) For each bin, the number of probes whose log2 ratios were less than the values of the bin was divided by the total number of probes in the subset to get the true positive rate for the corresponding bin; 6) The false positive rates were plotted on the X-axis and the true positive rates were plotted on the Y-axis to generate a ROC curve; 7) The AUC values were determined using Wilcoxon Rank Sum Test.

Evaluation of concordance among CNV calls from triplicate experiments for a same sample

The concordance among replicates was evaluated based on the proportion of calls present in two out of three replicates. Two CNV calls were considered the same event when the reciprocal overlap in length between the two CNVs was greater than 80%. The reciprocal overlap was calculated by taking the ratio of the common overlap between two calls by the total length of the call. Thus, CNVs detected in two or more replicates of the same sample that overlapped 80% or more of their lengths based on reciprocal overlap were considered high-confidence calls and their sizes correspond to the minimal shared boundaries between two or three replicates (**Supplementary Figs. 4 and 5, and main Figure 1**). A list of all CNVs, including the high-confidence calls, can be found in **Supplementary Table 3**. All downstream analyses were performed using the high-confidence set of CNV calls for all algorithms.

We further inspected the calling reproducibility across the different size ranges for the different combinations of platforms and algorithm by considering five size bins: 1-10Kb, 10-50Kb, 50-200Kb, >200Kb and >50Kb (**Supplementary Table 4**). We find that for most algorithms and array platforms, the reproducibility was fairly constant across the different size bins, although the reliability of such measurements for large size bins can be influenced by fewer number of CNV calls (ie. we therefore also looked at the size bin >50Kb). In addition, we also examined the degree of call fragmentation for large calls, which happens when a single CNV is detected as multiple smaller variants, by lowering the minimum CNV overlap required for a call to be considered replicated from 80% to any overlap, and found that the reproducibility of large calls increases (**Supplementary Table 4**). Finally, we observed that genomic complexity, or specifically the overlap of segmental duplications, could affect the reproducibility for large calls more than for small calls (**Supplementary Table 5**).

We next investigated to what extent the different platforms detect calls >50kb by comparing the results of each platform at the sample level, one platform at a time, to all variants >50kb that were identified by at least two other platforms (**Supplementary Table 6**). Specifically, we prepared a sample level ”gold-standard” dataset by taking all CNV calls > 50Kb detected by at least two high resolution array platforms for 4 samples (NA12239, NA18516, NA18576, NA18980), where the array type to be evaluated was not part of the ”gold-standard”. For this analysis we excluded lower resolution arrays (BAC, Affy250K, AG244K and Illmn650Y) and, for each sample, CNV calls detected by two or more platforms (any algorithm) were considered to be the same event. This means that the ”gold-standard” was built for each platform by taking calls from other platforms. For each platform, we calculated the proportion of detected CNV calls that were also found by at least two other array platforms (ie. named as overlapping calls) (**Supplementary Table 6**). In addition, we determined the number of ”gold-standard” regions that were missed by each array (ie. named as missed ”gold-regions”). These data allow us to estimate a false negative rate for large calls >50kb, which ranges from 15-77% for the different arrays (**Supplementary Table 6**). The differences between platforms can to some extent be explained by overlap with segmental duplications, as the Agilent 2x244K dataset has a larger fraction of calls overlapping segmental duplications, compared to the Illumina SNP arrays. In fact when examining such missing “gold-calls” in SNP arrays e.g. Illmn Omni, we find that 80-85% of the missing “gold-calls” are overlapped by SegDups at least 50% of their length (depending on the SegDups definition, respectively; see **Supplementary Tables 1 or 5** for the two definitions), of which a great proportion has poor probe coverage (i.e. low number of probes with an uneven distribution), and 95% have some overlap with SegDups, thus explaining the lack of calling accuracy in these regions for SNP arrays in general.

Evaluation of CNV calling reproducibility by comparison to Reference datasets

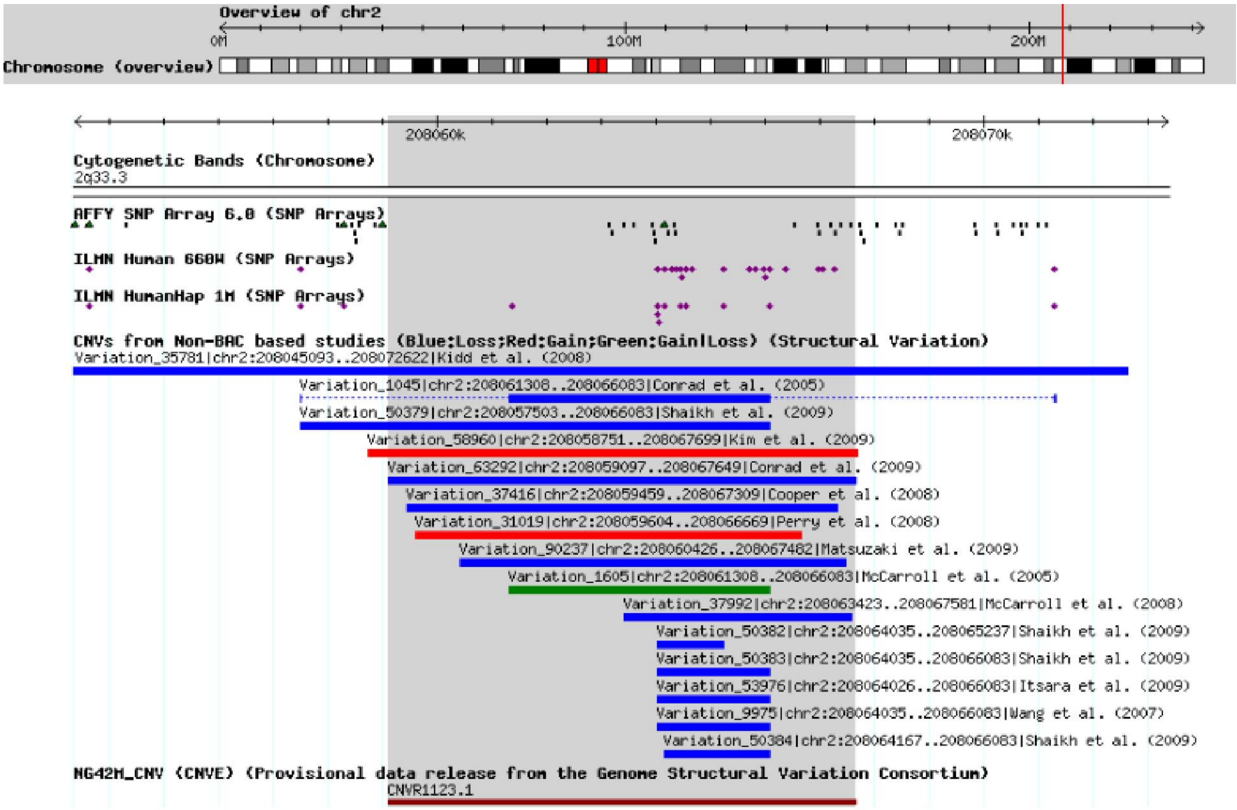
To evaluate CNV calling reproducibility in the main **Figure 2C**, we prepared five different reference datasets (**Supplementary Fig. 10**):

1. Database of Genomic Variants (DGV) based datasets. We used the studies listed in DGV version (variation.hg18.v9.mar.2010) after excluding BAC and Affymetrix 500K studies as well as studies that were not genome-wide (i.e. FISH, MLPA, PCR), and discarded indels or copy number variable regions with sizes of 100bp-1Kb. CNV regions detected using Mendelian inconsistencies and genotypes detected as ”null” were treated together with losses. After this initial curation, three DGV datasets were prepared: a DGV all, a high-resolution array-based subset of DGV and a sequencing-based DGV subset (**Supplementary Fig. 10**). We compared all DGV events against all sample level calls (unmerged data) using a reciprocal overlap of 50%;

2. An ultra-high resolution aCGH-based set of 8,599 validated CNVs[20] and a CNV genotype data for 4,978 variants of the same study[20] (**Supplementary Fig. 10**);

3. A set of 1,157 deletions derived from paired-end mapping based on fosmid-end sequencing[21] (**Supplementary Fig. 10**).

Though we initially explored three cutoffs (any-overlap, 50% and 70% reciprocal overlaps), we chose 50% as a more suitable criteria to apply to a wide variety of resolutions and probe distributions. Nevertheless, even 50% overlap sometimes does not capture all variants. An example of such an exception where a 50% reciprocal overlap is not able to capture a Illumina 660W-variant is shown below. Note that the 660W probe coverage is not sufficient compared to the size of the variant detected using an ultra-high resolution aCGH-based 42M array[20] (grey area).


Overview of chr2
0M
100M
200M
Chromosome (overview)
208060k
208070k
Cytogenetic Bands (Chromosome)
2q33.3
AFFY SNP Array 6.0 (SNP Arrays)
ILMN Human 660W (SNP Arrays)
ILMN HumanHap 1M (SNP Arrays)
CNVs from Non-BAC based studies (Blue:Loss;Red:Gain;Green:Gain|Loss) (Structural Variation)
Variation_35781|chr2:208045093..208072622|Kidd et al. (2008)
Variation_1045|chr2:208061308..208066083|Conrad et al. (2005)
Variation_50379|chr2:208057503..208066083|Shaikh et al. (2009)
Variation_58960|chr2:208058751..208067699|Kim et al. (2009)
Variation_63292|chr2:208059097..208067649|Conrad et al. (2009)
Variation_37416|chr2:208059459..208067309|Cooper et al. (2008)
Variation_31019|chr2:208059604..208066669|Perry et al. (2008)
Variation_90237|chr2:208060426..208067482|Matsuzaki et al. (2009)
Variation_1605|chr2:208061308..208066083|McCarroll et al. (2005)
Variation_37992|chr2:208063423..208067581|McCarroll et al. (2008)
Variation_50382|chr2:208064035..208065237|Shaikh et al. (2009)
Variation_50383|chr2:208064035..208066083|Shaikh et al. (2009)
Variation_53976|chr2:208064026..208066083|Itsara et al. (2009)
Variation_9975|chr2:208064035..208066083|Wang et al. (2007)
Variation_50384|chr2:208064167..208066083|Shaikh et al. (2009)
NG42M_CNV (CNVE) (Provisional data release from the Genome Structural Variation Consortium)
CNVR1123.1

Overlap analyses

All the overlap analyses performed have handled losses and gains separately except when otherwise stated, and were conducted hierarchically as follows. Once a subset of overlapping calls was found, we computed pairwise reciprocal overlaps for all calls against all other calls in the subset, ranked the overlaps, and progressively merged calls together into clusters starting from maximal overlapping pairs, i.e. CNVs were added to the cluster only if the degree of overlap met the overlap cutoff criterion. Finally, we merged clusters together if the overlap between every pair in the cluster was greater than the cutoff criterion.

Measuring similarity between each two algorithm/platform/site combinations

The Jaccard similarity coefficient was used to compare between CNV datasets for every algorithm/platform/site combination. This measure of similarity or overlap between two algorithms (or platforms) was computed as the size of the intersection divided by the total number of calls in the union set of the two algorithms (or platforms):

$$\text{Jaccard similarity coefficient: } J\,(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

A series of pairwise comparisons were performed at the sample level, and results shown for two main comparisons:

i) sample-level overlaps considering the total amount of calls made for four samples NA12239, NA18980, NA18576 and NA18516, where the total amount of non-redu ndant CNV calls was considered in the denominator. We calculated the percentage of all CNVs from 4 samples (union) found by both algorithms (**Supplementary Fig. 9A)** for sizes below (**Supplementary Fig. 9B)** and above 50Kb (**Supplementary Fig. 9C)**.

ii) sample-level overlaps for various combinations of algorithm/platform/site for a single sample NA12239 (**Supplementary Fig. 15A)** followed by comparison to a reference dataset of validated NA12239 CNV calls derived from a ultra-high resolution aCGH CNV discovery study[20]. We calculated the percentage of validated NA12239 CNV calls out of all shared calls between algorithms for all comparison of any two algorithms/platform/site (**Supplementary Fig. 15B**), as well as the percentage of NA12239 CNV calls detected by only one of the two algorithms tested, and that were found to be validated when compared to the reference set (**Supplementary Fig.**

15C). CNVs were considered validated when there was a reciprocal overlap of 50% or greater with the reference set. Though the Jaccard statistic is sensitive to the number of CNVs called by each algorithm (ideally each two algorithms would detect similar number of CNV calls), the relative values between the different comparisons of algorithms/platform/site are still very informative.

Reproducibility of CNV breakpoints (variability between triplicates for a sample)

We took all CNV calls detected in any of the 3 replicates and merged overlapping calls using a 1% threshold to account for possible fragmentation of calls in one of the replicates. Then, we calculated the distances (or start-start and end-end breakpoint differences) between two or three replicates for every array/algorithm combination, divided the distances into size bins for each platform, and plotted the proportion of CNVs in each bin for the start and end coordinates respectively (main **Figure 3**). We also computed the distances using the number of probes between the max and min start (and max and min end), and obtained similar results.

Accuracy of CNV breakpoints

To evaluate breakpoint precision, two nucleotide-resolution breakpoint datasets were used: i) a set of 862 non-redundant deletion breakpoints compiled from two published sources, a library of breakpoints collated from personal sequencing projects (Lam et al. 2010)[22] and a set of breakpoints derived from targeted hybridization-based DNA capture and 454 sequencing of all array-based CNVs detected in three unrelated individuals (Conrad et al. 2010)[23] (**Supplementary Fig. 12**), and ii) a set of sample-level deletion breakpoints derived from four samples sequenced in the 1000 Genomes Project (http://www.1000genomes.org)[24, 25].

We prepared the set of 862 non-redundant deletion breakpoints compiled from two published studies, Lam et al and Conrad et al. that were used in the main **Figure 4A** as well as **Supplementary Fig. 13**, as follows. We took the Conrad et al. set as the starting point and added the Lam et al. set of breakpoints as long as they did not overlap with Conrad et al. This ensured that each breakpoint was unique. We kept only autosomal breakpoints (a total of 925) that did not overlap segmental duplications (UCSC track, NCBI v.36, hg18) to ensure that both left and right breakpoints were free of segmental duplications, leaving a total of 781 breakpoint deletions for analysis (**Supplementary Fig. 12**). Our array-based deletions and the reference-based deletions were considered to represent the same event as long as they had a reciprocal overlap of at least 50%.

We used a set of sample-level deletions for four samples (NA18517, NA18576, NA12239 and NA10851) sequenced as part of the 1000 GP for which a breakpoint assembly was obtained for those samples. Thus, there are other deletion predictions with no known assembled breakpoints but these were not included. We used only unique high-confidence breakpoint predictions, and compared datasets using a 50% reciprocal overlap criterion. Note that some CGH gains might correspond to deletions in the reference, though there were very few calls in this category.

4. References

1. Fiegler, H., Redon, R. & Carter, N.P. Construction and use of spotted large-insert clone DNA microarrays for the detection of genomic copy number changes. *Nat Protoc* **2**, 577-587 (2007).
2. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-454 (2006).
3. Fiegler, H. et al. Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res* **16**, 1566-1574 (2006).
4. Whitby, H. et al. Benign copy number changes in clinical cytogenetic diagnostics by array CGH. *Cytogenet Genome Res* **123**, 94-101 (2008).
5. Bailey, J.A. et al. Recent segmental duplications in the human genome. *Science* **297**, 1003-1007 (2002).
6. Marshall, C.R. et al. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* **82**, 477-488 (2008).
7. Zhao, X. et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res* **64**, 3060-3071 (2004).
8. Partek® Genomics SuiteTM, v., build 6.08.0110. Copyright © 2008 Partek Inc., St. Louis, MO, USA.
9. Kennedy, G.C. et al. Large-scale genotyping of complex DNA. *Nat Biotechnol* **21**, 1233-1237 (2003).
10. Korn, J.M. et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253-1260 (2008).
11. McCarroll, S.A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166-1174 (2008).
12. Zhang, J. et al. iPattern: a cross-sample copy number variation discovery method for multiple array platforms. *Submitted* (2010).
13. Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-372 (2010).
14. Colella, S. et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* **35**, 2013-2025 (2007).
15. Wang, K. et al. 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 (2007).
16. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-454 (2006).

17. Lipson, D., Aumann, Y., Ben-Dor, A., Linial, N. & Yakhini, Z. Efficient calculation of interval scores for DNA copy number data analysis. *J Comput Biol* **13**, 215-228 (2006).
18. Lionel, A.C. et al. Cross-neuropsychiatric disorder comparisons of rare copy number variation identify risk genes for attention deficit and hyperactivity. *in preparation* (2010).
19. Matsuzaki, H., Wang, P.H., Hu, J., Rava, R. & Fu, G.K. High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. *Genome Biol* **10**, R125 (2009).
20. Conrad, D.F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712 (2010).
21. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
22. Lam, H.Y. et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* **28**, 47-55 (2010).
23. Conrad, D.F. et al. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet* **42**, 385-391 (2010).
24. 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
25. Mills, R.E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65 (2011).

Nature Biotechnology: doi:10.1038/nbt.1852

Supplementary Table 1. Counts of probes (SNP and CN probes) per platform, including the proportion of probes in genes and segmental duplications. The proportion of genes and transcripts without probes is also given.

	Probe coverage							No Probe Coverage	
	Genome-wide			Gene(2)	Gene +/- 10Kb	SegDups (>90%,>1Kb)(3)	SegDups (>95%,>10Kb) (3)	Transcripts (31,276)(2)	Genes (20,648)(2)
Platform	Total #Probes	#SNP probes	#CN probes(1)	#Probes (%)	#Probes (%)	#Probes (%)	#Probes (%)	n (%)	n (%)
BAC array	29,043	0	29,043	19,988 (68.8)	20,400 (70.2)	8,283 (28.5)	1,892 (6.5)	475 (1.5)	351 (1.7)
Agilent 244K	236,381 *	0	236,381 *	127,397 (53.4)	154,270 (64.7)	6,717 (2.8)	4,137 (1.7)	4,522 (14.5)	3,335 (16.2)
Agilent 2x244K	462,609 **	0	462,609 **	189,098 (39.7)	236,356 (49.6)	57,653 (12.1)	41,348 (8.7)	7,547 (24.1)	5,606 (27.2)
NimbleGen 720K	720,412	0	720,412	299,434 (41.6)	363,669 (50.5)	20,034 (2.8)	15,702 (2.2)	4,429 (14.2)	3,393 (16.4)
NimbleGen 2.1M	2,161,679	0	2,161,679	883,057 (40.9)	1071,499 (49.6)	83,315 (3.9)	62,447 (2.9)	1,356 (4.3)	940 (4.6)
Affymetrix 500K***	500,568	500,568	0	196,009 (39.2)	233,785 (46.7)	7,525 (1.5)	3,078 (0.6)	9,004 (28.8)	6,468 (31.3)
Affymetrix 250K NspI	262,454	262,454	0	100,195 (38.2)	117,977 (45.0)	3,926 (1.5)	1,611 (0.61)	13,761 (44.0)	9,357 (45.3)
Affymetrix 250K StyI***	238,684	238,684	0	95,814 (40.1)	115,808 (48.5)	3,599 (1.5)	1,467 (0.61)	11,809 (37.8)	8,153 (39.5)
Affymetrix 6.0	1,879,489	933,683	945,806	734,551 (39.1)	883,699 (47.0)	45,353 (2.4)	25,838 (1.4)	5,082 (16.2)	3,561 (17.2)
Illumina 650Y	655,352	655,352	0	267,653 (40.8)	318,795 (48.6)	6,059 (0.9)	1,953 (0.3)	6,327 (20.2)	4,542 (22.0)
Illumina 1M-single	1,072,820	1,049,008	23,812	498,758 (46.5)	604,476 (56.3)	34,903 (3.3)	13,062 (1.2)	1,862 (6.0)	1,283 (6.2)
Illumina 660W	657,366	561,490	95,876	272,079 (41.4)	330,199 (50.2)	27,178 (4.1)	18,370 (2.8)	5,659 (18.1)	4,312 (20.9)
Illumina Omni	1,140,419	1,016,423	123,996	496,862 (43.6)	616,912 (54.1)	50,633 (4.4)	27,973 (2.5)	1,841 (5.9)	1,225 (5.9)

(1) CN probes, copy number or intensity-only probes; * 236,381 unique probes plus 1,000 triplicates (i.e. two additional copies of 1,000 probes) and an additional 5,045 quality control features; ** 462,609 unique probes plus ~14,000 replicates comprised of copies of 10,000 unique probes (i.e. non-expanded content). In the expanded array content, the log2 ratios of all probes are assigned to a total of 500,974 genomic locations (see Supplementary Information); *** Only the 250K-NspI array was included in the final analysis.

(2) NCBI v36, hg18, RefSeq downloaded from UCSC in March 2010;

(3) SegDups, segmental duplications or large recent duplications (blocks of non-RepeatMasked sequence ≥ 1 kb and ≥ 90% identity) as defined in Bailey et al. 2002 [5]. Regions were downloaded from UCSC, and two groups were considered:
- SegDups (>90%,>1Kb): at least 1 Kb of the total sequence (containing at least 500 bp of non-RepeatMasked sequence) had to align and a sequence identity of at least 90% was required;
- SegDups (>95%,>10Kb): blocks of at least >10Kb in size and 95% identity.

Supplementary Table 2. Proportion of probes with positions mapping in genes, arranged by chromosome and platform

A. CGH arrays

	BAC array		AG 244K		AG 2x244K		NG 720K		NG 2.1M	
Chr	total #probes	% probes in genes	total #probes	% probes in genes	total #probes	% probes in genes	total #probes	% probes in genes	total #probes	% probes in genes
1	2,231	77.1	19,364	49.8	35,795	44.1	57,267	45.8	171,757	45.4
2	2,241	66.3	18,635	50.1	36,851	37.1	61,100	39.4	183,249	39.7
3	2,042	67.8	15,951	60.2	28,617	42.4	48,846	45.2	146,506	43.7
4	1,987	57.6	13,472	43.3	32,245	29.6	46,577	35.5	139,731	34.6
5	1,976	59.2	13,358	54.1	26,994	34.3	44,383	36.8	133,149	35.4
6	1,799	67.0	13,383	55.7	26,657	39.9	42,242	40.8	126,695	40.2
7	1,703	72.1	13,335	48.9	29,704	40.9	39,507	45.7	118,517	44.4
8	1,436	66.4	11,064	56.4	24,908	39.5	36,192	38.1	108,547	38.4
9	994	68.9	9,477	46.4	22,031	37.1	30,407	41.0	91,213	39.6
10	1,212	74.8	11,127	60.9	23,537	42.4	34,025	45.4	102,066	45.1
11	1,413	75.2	11,538	57.2	21,654	41.0	32,820	44.8	98,426	43.8
12	1,448	72.7	11,253	58.1	21,168	42.1	32,608	43.4	97,812	43.5
13	856	54.6	7,722	52.6	14,783	36.2	24,766	34.3	74,283	33.1
14	825	66.3	8,258	55.4	13,437	32.3	22,550	40.1	67,627	39.4
15	805	80.6	8,080	60.3	14,806	44.3	20,496	46.3	61,486	46.5
16	759	73.6	6,881	61.6	15,242	48.4	19,871	45.1	59,590	45.0
17	809	86.2	7,728	65.2	15,262	51.3	19,951	53.4	59,833	53.3
18	740	61.2	5,816	53.7	12,442	35.0	19,343	36.2	58,024	35.7
19	524	90.5	6,043	59.6	13,317	51.8	13,748	49.7	41,242	50.5
20	629	75.2	5,385	56.8	11,220	42.6	15,329	42.2	45,971	42.8
21	349	66.5	3,409	50.1	6,100	37.3	8,716	38.1	26,142	36.1
22	443	72.7	4,109	59.6	8,468	49.1	8,913	53.0	26,718	52.9
X	1,637	62.4	11,036	38.1	18,034	31.9	35,316	35.0	106,458	31.6
Y	185	52.4	1,295	23.6	3,335	14.6	5,439	17.9	16,637	17.4

B. SNP arrays

	Affy 500K		Affy 250K-NspI		Affy 250K-StyI		Affy 6.0		Illmn 650Y*		Illmn 1M**		Illmn 660W		Illmn Omni	
Chr	total #probes	% probes in genes	total #probes	% probes in genes	total #probes	% probes in genes	total #probes	% probes in genes	total #probes	% probes in genes	total #probes	% probes in genes	total #probes	% probes in genes	total #probes	% probes in genes
1	40,261	42.8	19,886	42.3	20,375	43.3	146,524	43.4	49,639	43.0	87,592	51.5	50,524	44.2	94,459	48.9
2	41,419	37.7	22,228	37.9	19,191	37.3	153,732	37.5	53,765	38.0	84,508	44.3	52,352	39.1	86,694	41.2
3	33,820	43.2	18,384	41.6	15,436	45.1	127,815	43.0	44,564	44.3	70,542	49.7	42,548	45.4	69,665	47.1
4	32,365	32.9	19,079	32.0	13,286	34.3	120,360	32.8	39,942	33.7	61,452	39.1	38,931	34.8	66,438	36.7
5	32,078	34.4	17,179	33.4	14,899	35.5	115,731	34.0	40,976	35.8	63,513	41.4	39,871	36.8	64,147	38.3
6	31,481	38.1	17,151	38.2	14,330	38.0	112,895	38.7	43,239	40.0	69,458	44.0	41,546	40.7	84,930	39.9
7	25,839	43.0	13,951	42.3	11,888	43.8	101,093	42.4	35,507	44.2	56,815	49.0	35,299	45.0	60,297	47.1
8	27,471	37.3	14,842	37.5	12,629	37.0	98,306	37.7	37,282	40.0	54,261	44.1	36,158	40.5	58,720	40.0
9	22,875	39.9	11,941	39.2	10,934	40.6	82,225	39.9	31,192	39.9	45,643	45.5	31,746	39.3	53,512	40.2
10	28,511	41.9	14,282	42.2	14,229	41.5	93,655	43.1	34,493	43.9	52,187	48.8	33,880	44.8	58,897	46.0
11	26,277	40.0	13,309	38.1	12,968	41.8	89,615	41.2	32,005	44.5	52,304	49.2	31,370	45.8	55,160	46.5
12	24,964	39.4	13,064	38.6	11,900	40.3	87,372	40.5	31,873	43.2	51,636	49.2	31,502	44.3	53,878	45.4
13	19,210	31.9	11,120	31.6	8,090	32.4	66,106	32.3	25,191	31.5	36,590	36.8	23,979	33.2	38,213	36.1
14	15,744	36.5	8,181	35.1	7,563	38.1	57,121	37.2	21,450	36.6	33,906	43.7	21,401	35.7	33,797	40.3
15	14,378	44.5	7,028	43.6	7,350	45.4	53,595	44.4	19,594	45.5	32,141	51.1	20,570	45.4	34,862	46.4
16	15,309	44.8	7,024	43.5	8,285	46.0	54,215	44.1	19,727	45.6	33,915	53.0	20,632	46.1	37,610	51.2
17	11,286	49.2	4,854	47.5	6,432	50.5	46,678	50.6	16,629	49.4	32,498	58.7	18,978	50.5	34,611	54.9
18	14,882	34.5	8,150	34.1	6,732	35.0	52,109	34.7	20,165	35.2	29,191	39.3	19,167	35.8	30,964	35.8
19	6,400	44.7	2,693	41.0	3,707	47.3	30,362	45.3	10,739	51.3	24,783	58.8	13,323	51.5	27,667	57.4
20	12,406	38.4	5,839	38.4	6,567	38.3	43,648	39.7	16,911	39.3	26,387	46.7	16,158	40.6	31,334	44.4
21	7,126	35.2	3,937	33.6	3,189	37.2	25,129	34.8	9,645	37.3	13,777	41.1	9,484	38.4	15,872	42.4
22	6,210	49.8	2,521	47.3	3,689	51.6	24,513	50.4	9,730	49.9	16,492	55.6	11,244	47.7	17,694	50.9
X	10,555	28.8	5,715	27.3	4,840	30.5	87,204	28.7	16,472	29.8	40,097	39.2	16,509	29.8	27,493	35.0
Y	271	30.3	96	35.4	175	27.4	9,486	14.0	10	60.0	2,283	20.6	44	11.4	2,322	23.0

* 650Y v.3
** 1M-single v1

Nature Biotechnology: doi:10.1038/nbt.1852

Supplementary Table 3. List of all CNVs that passed QC
Data provided separately, as an Excel workbook file.

Supplementary Table 4. CNV call concordance between triplicate experiments for different size ranges[1]

A. By algorithm

	80% reciprocal overlap						*any overlap**					
Size range (kb)	**all**	**1-10**	**10-50**	**50-200**	**>200**	**>50**	**all**	**1-10**	**10-50**	**50-200**	**>200**	**>50**
Birdsuite (2 sites, 1 array type)	**66.8%**	**63.6%**	**68.0%**	**69.4%**	**74.4%**	**72.1%**	**71.6%**	**65.0%**	**71.2%**	**75.4%**	**88.8%**	**82.1%**
TCAG	75.9%	77.7%	75.4%	67.3%	75.5%	72.2%	82.4%	79.6%	80.4%	74.5%	89.1%	82.2%
WTSI	57.7%	49.5%	60.6%	71.4%	73.3%	72.1%	60.8%	50.4%	62.0%	76.4%	88.5%	81.9%
CNVFinder (1 site, 1 array type)	**60.0%**	**na**	**66.7%**	**63.6%**	**57.1%**	**59.9%**	**73.6%**	**na**	**66.7%**	**64.3%**	**73.6%**	**73.7%**
CNVPart (2 sites, 4 array types)	**43.0%**	**42.1%**	**34.2%**	**38.6%**	**42.3%**	**40.1%**	**62.2%**	**51.2%**	**45.3%**	**49.4%**	**55.6%**	**54.4%**
TCAG	45.2%	45.7%	36.5%	38.1%	43.6%	40.3%	64.5%	54.8%	48.3%	49.5%	53.8%	53.2%
HMS	40.0%	37.2%	31.2%	39.3%	40.5%	39.9%	59.2%	46.5%	41.4%	49.2%	58.1%	56.0%
dChip (2 sites, 1 array type)#	**35.4%**	**18.3%**	**36.7%**	**27.6%**	**51.2%**	**36.3%**	**59.5%**	**38.6%**	**52.5%**	**43.0%**	**68.3%**	**61.0%**
TCAG	36.3%	16.7%	32.5%	31.4%	50.0%	39.3%	58.1%	27.3%	60.0%	40.0%	64.3%	58.2%
WTSI	34.4%	20.0%	40.9%	23.9%	52.4%	33.3%	60.9%	50.0%	45.0%	45.9%	72.2%	63.8%
ADM-2 (2 sites, 2 array types)	**66.4%**	**67.3%**	**66.4%**	**61.3%**	**66.2%**	**62.9%**	**79.4%**	**73.5%**	**72.9%**	**70.4%**	**75.6%**	**76.1%**
TCAG	68.6%	70.2%	71.0%	63.1%	65.2%	64.3%	82.6%	76.6%	76.7%	72.7%	74.1%	78.4%
WTSI	61.9%	64.4%	57.2%	57.8%	68.1%	60.0%	73.1%	70.4%	65.4%	65.8%	78.4%	71.4%
GTC-CN5 (2 sites, 1 array type)	**33.1%**	**32.8%**	**31.4%**	**39.6%**	**37.3%**	**39.0%**	**49.3%**	**39.6%**	**42.2%**	**58.3%**	**60.1%**	**66.0%**
TCAG	49.7%	50.3%	50.0%	50.4%	41.8%	47.6%	74.5%	57.9%	68.2%	75.0%	67.5%	85.0%
WTSI	16.5%	15.3%	12.7%	28.7%	32.9%	30.4%	24.2%	21.2%	16.3%	41.5%	52.7%	47.0%
iPattern (2 sites, 4 array types)	**80.8%**	**86.2%**	**70.3%**	**65.9%**	**66.0%**	**65.6%**	**87.5%**	**88.7%**	**73.5%**	**70.0%**	**75.8%**	**72.7%**
TCAG	82.9%	86.2%	74.7%	71.8%	69.5%	70.9%	89.6%	88.7%	77.7%	76.3%	79.4%	78.5%
HMS	77.9%	86.2%	64.4%	58.1%	61.4%	58.5%	84.6%	88.7%	67.9%	61.6%	70.9%	64.9%
Nexus (2 sites, 5 array types)	**45.6%**	**40.0%**	**41.5%**	**40.6%**	**42.7%**	**42.3%**	**61.0%**	**48.3%**	**54.9%**	**51.3%**	**56.1%**	**59.9%**
TCAG	53.2%	55.6%	54.5%	55.1%	49.1%	53.6%	73.1%	69.7%	80.8%	66.7%	58.9%	69.9%
WTSI	41.8%	32.2%	42.6%	40.1%	45.1%	43.2%	63.0%	37.6%	51.1%	50.6%	60.1%	62.1%
Partek (2 sites, 1 array type)#	**52.1%**	**41.7%**	**50.3%**	**64.8%**	**48.2%**	**57.4%**	**76.8%**	**54.8%**	**68.8%**	**81.4%**	**70.4%**	**79.5%**
TCAG	51.0%	39.3%	42.3%	70.0%	52.5%	63.0%	75.4%	52.8%	64.1%	86.3%	71.9%	84.8%
WTSI	53.2%	44.1%	58.3%	59.5%	43.9%	51.8%	78.2%	56.9%	73.5%	76.5%	69.0%	74.2%
PCNV (2 sites, 4 array types)	**39.2%**	**45.8%**	**30.1%**	**31.8%**	**30.7%**	**32.1%**	**54.2%**	**55.2%**	**38.3%**	**40.8%**	**41.0%**	**44.4%**
TCAG	36.7%	44.5%	27.2%	30.1%	28.9%	30.3%	51.2%	54.4%	34.4%	37.6%	40.1%	40.5%
HMS	42.6%	47.1%	33.9%	34.1%	33.0%	34.4%	58.1%	55.9%	43.4%	45.1%	42.1%	49.5%
QSNP (2 sites, 4 array types)	**47.2%**	**58.1%**	**39.7%**	**37.3%**	**36.9%**	**36.6%**	**65.6%**	**66.2%**	**49.9%**	**49.0%**	**54.2%**	**56.3%**
TCAG	48.7%	65.5%	40.4%	38.4%	34.5%	36.8%	66.8%	71.4%	49.3%	49.4%	52.7%	56.5%
HMS	45.2%	48.2%	38.8%	35.8%	40.1%	36.4%	64.0%	59.3%	50.8%	48.6%	56.2%	55.9%

B. By array platform

	80% reciprocal overlap						*any overlap**					
Size range (kb)	**all**	**1-10**	**10-50**	**50-200**	**>200**	**>50**	**all**	**1-10**	**10-50**	**50-200**	**>200**	**>50**
AG2X244K (2 sites, 2 algorithms)	**57.9%**	**60.2%**	**53.1%**	**56.6%**	**57.0%**	**57.1%**	**74.5%**	**69.5%**	**63.0%**	**65.6%**	**67.5%**	**70.7%**
TCAG	55.8%	57.5%	51.6%	53.6%	59.9%	55.3%	72.3%	65.8%	61.3%	62.0%	72.3%	69.0%
HMS	60.1%	62.9%	54.6%	59.5%	54.1%	58.8%	76.7%	73.2%	64.7%	69.1%	62.8%	72.3%
Illmn1M (2 sites, 4 algorithms)	**48.5%**	**50.0%**	**46.8%**	**47.7%**	**48.0%**	**47.8%**	**63.2%**	**54.8%**	**55.3%**	**57.0%**	**58.8%**	**63.2%**
TCAG	49.0%	52.4%	48.8%	46.9%	47.9%	47.3%	61.4%	56.5%	55.9%	54.3%	56.2%	58.8%
HMS	48.0%	47.6%	44.8%	48.6%	48.1%	48.4%	65.1%	53.2%	54.7%	59.7%	61.5%	67.7%
Illmn660W (2 sites, 4 algorithms)	**51.9%**	**58.4%**	**38.5%**	**32.3%**	**31.8%**	**32.4%**	**66.7%**	**66.8%**	**46.0%**	**40.7%**	**45.0%**	**44.6%**
TCAG	56.0%	62.2%	41.2%	36.0%	37.5%	36.5%	71.3%	70.8%	48.6%	46.6%	52.7%	51.1%
HMS	47.8%	54.5%	35.7%	28.5%	26.2%	28.3%	62.0%	62.7%	43.5%	34.9%	37.3%	38.1%
IllmnOMNI (2 sites, 4 algorithms)	**56.0%**	**60.3%**	**42.1%**	**44.9%**	**47.6%**	**45.6%**	**71.1%**	**71.3%**	**50.9%**	**53.8%**	**59.5%**	**57.3%**
TCAG	53.6%	58.7%	38.7%	41.4%	38.2%	41.0%	69.8%	70.8%	47.4%	48.8%	47.2%	50.6%
HMS	58.5%	61.9%	45.6%	48.4%	56.9%	50.2%	72.3%	71.9%	54.4%	58.8%	71.7%	64.1%
Affy6 (2 sites, 5 algorithms)	**50.7%**	**43.6%**	**50.7%**	**53.3%**	**54.9%**	**54.1%**	**67.1%**	**53.0%**	**62.1%**	**66.8%**	**74.3%**	**74.3%**
TCAG	58.8%	52.7%	56.7%	59.2%	58.4%	59.8%	76.0%	59.6%	72.4%	72.1%	77.3%	80.4%
WTSI	40.4%	32.2%	43.1%	45.9%	50.6%	46.9%	56.0%	44.6%	49.2%	60.1%	70.6%	66.7%
Affy250K (1 site, 2 algorithms)	**23.5%**	**na**	**50.0%**	**13.3%**	**27.4%**	**21.7%**	**39.0%**	**na**	**50.0%**	**25.3%**	**38.3%**	**37.4%**
BAC (1 site, 2 algorithms)	**52.3%**	**na**	**66.7%**	**51.5%**	**51.6%**	**52.2%**	**68.1%**	**na**	**66.7%**	**53.6%**	**68.9%**	**68.0%**
NG2.1M (2 sites, 1 algorithm)	**21.5%**	**14.0%**	**17.3%**	**23.3%**	**25.0%**	**24.2%**	**39.9%**	**14.0%**	**29.1%**	**35.8%**	**39.4%**	**42.0%**
Illmn650Y (1 site, 3 algorithms)	**46.0%**	**50.0% †**	**39.1%**	**46.5%**	**46.5%**	**45.9%**	**62.8%**	**50.0%†**	**47.3%**	**55.8%**	**62.0%**	**60.4%**
AG244K (1 site, 2 algorithms)	**61.7%**	**na**	**70.9%**	**58.7%**	**60.2%**	**59.1%**	**79.0%**	**na**	**92.9%**	**70.3%**	**70.2%**	**76.0%**

[1] The reproducibility across different size bins for each dataset (1-10kb, 10-50kb, 50-200kb, >200kb and >50kb) is given as percentage. Values were averaged across array types for an algorithm (**A**) or across algorithms for a same array platform (**B**). Values are also given for each algorithm/platform combination when CNV data is available from two sites (**C**). Two CNV calls were considered the same event when the reciprocal overlap in length between the two CNVs was greater than 80%;
* To account for possible call fragmentation, we also lowered the minimum CNV overlap required for a call to be considered replicated from 80% (the default used for all other analyzes) to any overlap; 'na': indicates that no CNV calls in that size range were detected for the particular platform. # no Partek or dChip results have been considered in this table for the Affy250K-Nsp array, because of the lower number of concordant calls made for any of the size bins. For example, Partek detected less than 5 calls in all samples and only one was concordant. † The average number of calls per sample in this bin is <5.

C. By site

		80% reciprocal overlap						*any overlap**					
Size range (kb)		**all**	**1-10**	**10-50**	**50-200**	**>200**	**>50**	**all**	**1-10**	**10-50**	**50-200**	**>200**	**>50**
Ilmn1M/	HMS	38.7%	24.1%	32.9%	48.1%	48.5%	48.3%	57.7%	24.1%	44.3%	63.6%	73.1%	71.6%
cnvPart	TCAG	38.9%	35.7%	39.2%	35.9%	47.7%	39.7%	50.0%	41.0%	48.8%	42.7%	53.7%	47.1%
Ilmn1M/	HMS	79.1%	90.9%	73.1%	71.0%	70.6%	70.9%	86.7%	90.9%	79.5%	76.3%	75.0%	80.3%
iPattern	TCAG	86.1%	90.6%	85.7%	87.5%	67.9%	81.5%	89.8%	90.6%	85.7%	90.3%	70.4%	87.4%
Ilmn1M/	HMS	39.4%	39.0%	40.8%	39.3%	34.5%	37.7%	65.3%	55.6%	55.4%	50.0%	53.2%	60.7%
QSNP	TCAG	38.7%	44.4%	42.2%	32.0%	35.2%	33.9%	58.5%	49.2%	54.6%	39.8%	57.8%	51.0%
Ilmn1M/	HMS	34.9%	36.3%	32.6%	35.9%	38.7%	36.8%	50.9%	42.1%	39.9%	48.9%	44.8%	57.9%
PCNV	TCAG	32.3%	39.0%	27.9%	32.3%	40.9%	34.0%	47.1%	45.3%	34.2%	44.4%	42.9%	49.7%
Ilmn660W/	HMS	37.8%	42.1%	24.9%	30.3%	23.1%	29.6%	56.3%	54.2%	34.1%	36.3%	38.7%	41.4%
cnvPart	TCAG	46.9%	52.1%	33.5%	33.6%	38.1%	34.9%	68.1%	64.3%	42.4%	48.0%	55.9%	52.8%
Ilmn660W/	HMS	68.2%	76.3%	53.7%	37.8%	43.5%	38.4%	76.3%	80.2%	55.8%	38.6%	60.0%	42.1%
iPattern	TCAG	79.3%	86.1%	62.6%	48.9%	63.0%	51.3%	87.8%	91.1%	65.0%	54.3%	68.0%	57.7%
Ilmn660W/	HMS	40.8%	46.5%	30.5%	25.2%	21.6%	24.7%	55.4%	53.3%	40.1%	34.9%	30.3%	38.3%
QSNP	TCAG	57.7%	65.6%	42.7%	33.6%	30.6%	32.7%	72.9%	72.6%	50.9%	45.3%	50.0%	52.0%
Ilmn660W/	HMS	44.6%	53.1%	33.8%	20.8%	16.7%	20.3%	60.0%	63.1%	43.8%	29.9%	20.0%	30.5%
PCNV	TCAG	40.3%	45.2%	26.0%	27.9%	18.2%	27.2%	56.4%	55.3%	36.1%	38.7%	36.8%	41.7%
IlmnOMNI/	HMS	43.5%	45.2%	35.8%	39.4%	50.0%	41.7%	63.5%	61.0%	45.8%	47.8%	62.5%	55.0%
cnvPart	TCAG	41.0%	45.1%	28.9%	33.1%	25.0%	31.4%	66.2%	63.9%	42.1%	44.9%	35.5%	44.7%
IlmnOMNI/	HMS	86.4%	91.4%	66.4%	65.4%	70.0%	66.1%	90.9%	95.0%	68.3%	70.0%	77.8%	72.4%
iPattern	TCAG	84.9%	88.5%	67.1%	73.8%	75.0%	74.0%	91.1%	92.6%	70.6%	75.6%	85.7%	77.1%
IlmnOMNI/	HMS	55.6%	59.0%	45.0%	43.0%	64.0%	46.8%	71.3%	68.9%	57.0%	60.9%	85.0%	68.8%
QSNP	TCAG	46.6%	52.1%	32.1%	33.0%	34.7%	33.5%	64.5%	63.9%	41.6%	45.7%	41.5%	50.7%
IlmnOMNI/	HMS	48.4%	52.0%	35.2%	45.7%	43.8%	46.2%	63.6%	62.5%	46.6%	56.5%	61.5%	60.0%
PCNV	TCAG	42.0%	49.2%	26.5%	25.8%	18.2%	25.0%	57.5%	62.6%	35.5%	29.1%	26.3%	29.7%
Affy6/	TCAG	75.9%	77.7%	75.4%	67.3%	75.5%	72.2%	82.4%	79.6%	80.4%	74.5%	89.1%	82.2%
Birdsuite	WTSI	57.7%	49.5%	60.6%	71.4%	73.3%	72.1%	60.8%	50.4%	62.0%	76.4%	88.5%	81.9%
Affy6 /	TCAG	49.7%	50.3%	50.0%	50.4%	41.8%	47.6%	74.5%	57.9%	68.2%	75.0%	67.5%	85.0%
GTC-CN5	WTSI	16.5%	15.3%	12.7%	28.7%	32.9%	30.4%	24.2%	21.2%	16.3%	41.5%	52.7%	47.0%
Affy6/	TCAG	51.0%	39.3%	42.3%	70.0%	52.5%	63.0%	75.4%	52.8%	64.1%	86.3%	71.9%	84.8%
Partek	WTSI	53.2%	44.1%	58.3%	59.5%	43.9%	51.8%	78.2%	56.9%	73.5%	76.5%	69.0%	74.2%
Affy6/	TCAG	36.3%	16.7%	32.5%	31.4%	50.0%	39.3%	58.1%	27.3%	60.0%	40.0%	64.3%	58.2%
dChip	WTSI	34.4%	20.0%	40.9%	23.9%	52.4%	33.3%	60.9%	50.0%	45.0%	45.9%	72.2%	63.8%
AG2x244K/	TCAG	66.7%	70.2%	60.1%	63.8%	60.0%	63.0%	78.2%	76.6%	67.7%	70.8%	67.0%	74.3%
ADM-2	WTSI	61.9%	64.4%	57.2%	57.8%	68.1%	60.0%	73.1%	70.4%	65.4%	65.8%	78.4%	71.4%
AG2x244K/	TCAG	53.4%	55.6%	49.1%	55.2%	48.3%	54.6%	75.1%	69.7%	61.6%	67.5%	58.7%	70.3%
Nexus	WTSI	49.8%	50.6%	46.0%	49.4%	51.7%	50.7%	71.5%	61.3%	57.2%	58.3%	66.1%	66.7%
NG2.1M/	HMS	13.8%	no-calls	10.7%	13.8%	19.9%	15.7%	28.7%	no-calls	18.0%	23.1%	34.5%	30.8%
Nexus	WTSI	29.3%	13.9%	23.8%	32.7%	30.1%	32.7%	51.1%	13.9%	40.2%	48.4%	44.4%	53.1%

Supplementary Table 5. Overlap of non-concordant calls with segmental duplications (SegDup, %)[1]

	Non-concordant calls	
Size bins	**SegDups[2] (>90%, >1Kb)**	**SegDups[2] (>95%, >10Kb)**
1-10 kb	10%	4%
> 50 kb	51%	45%
> 200 kb	60%	55%

[1] Overlap was considered when a CNV call was overlapped by a segdup block across ≥ 50% of its length;

[2] SegDups, segmental duplications or large recent duplications (blocks of non-RepeatMasked sequence ≥ 1 kb and ≥ 90% identity) as defined in Bailey et al. 2002 [5]. Regions were downloaded from UCSC, and two groups were considered:

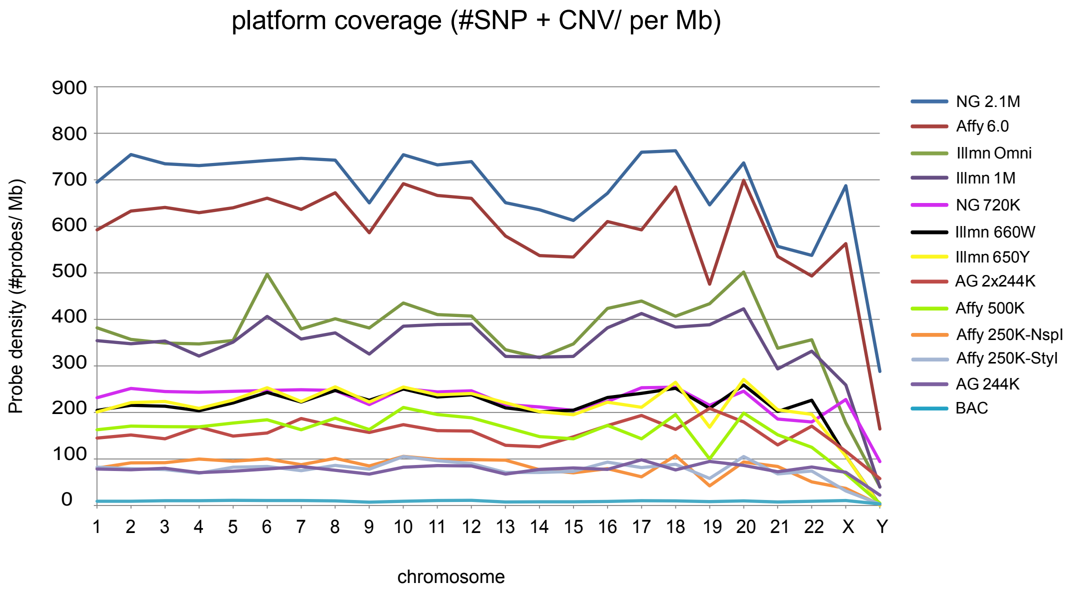
- SegDups (>90%,>1Kb): at least 1 Kb of the total sequence (containing at least 500 bp of non-RepeatMasked sequence) had to align and a sequence identity of at least 90% was required;
- SegDups (>95%,>10Kb): blocks of at least >10Kb in size and 95% identity.

Supplementary Table 6. Evaluation of the ability of each array platform to detect variants with sizes >50kb [1]. The table lists the proportion of overlapping and missing CNV calls for each platform by comparison to a ”gold-standard” made of CNV calls with sizes > 50Kb detected by at least two array platforms.

		#Missed ‘gold-standard’ calls (%)	#Overlapping calls (%)
NA18576	**AG2X244K**	7 (14)	161 (62)
NA18576	**AFFY6.0**	40 (63)	86 (90)
NA18576	**NG2.1M**	14 (25)	68 (79)
NA18576	**ILMNOMNI**	48 (76)	57 (86)
NA18576	**ILMN1M**	49 (77)	68 (91)
NA18576	**ILMN660W**	50 (79)	53 (73)
NA18576	**NG720K**	28 (47)	36 (88)
NA18517	**AG2X244K**	5 (9)	174 (58)
NA18517	**AFFY6.0**	48 (71)	93 (78)
NA18517	**NG2.1M**	19 (35)	74 (74)
NA18517	**ILMNOMNI**	55 (77)	64 (89)
NA18517	**ILMN1M**	57 (79)	67 (93)
NA18517	**ILMN660W**	49 (74)	43 (72)
NA18517	**NG720K**	41 (57)	38 (86)
NA12239	**AG2X244K**	10 (23)	154 (58)
NA12239	**AFFY6.0**	34 (61)	104 (85)
NA12239	**NG2.1M**	21 (45)	40 (66)
NA12239	**ILMNOMNI**	36 (65)	99 (93)
NA12239	**ILMN1M**	37 (67)	91 (75)
NA12239	**ILMN660W**	31 (57)	101 (74)
NA12239	**NG720K**	28 (53)	29 (78)
NA18980	**AG2X244K**	7 (14)	170 (51)
NA18980	**AFFY6.0**	48 (73)	75 (83)
NA18980	**NG2.1M**	18 (32)	56 (75)
NA18980	**ILMNOMNI**	53 (80)	49 (82)
NA18980	**ILMN1M**	58 (85)	33 (79)
NA18980	**ILMN660W**	56 (81)	40 (62)
NA18980	**NG720K**	28 (44)	40 (85)

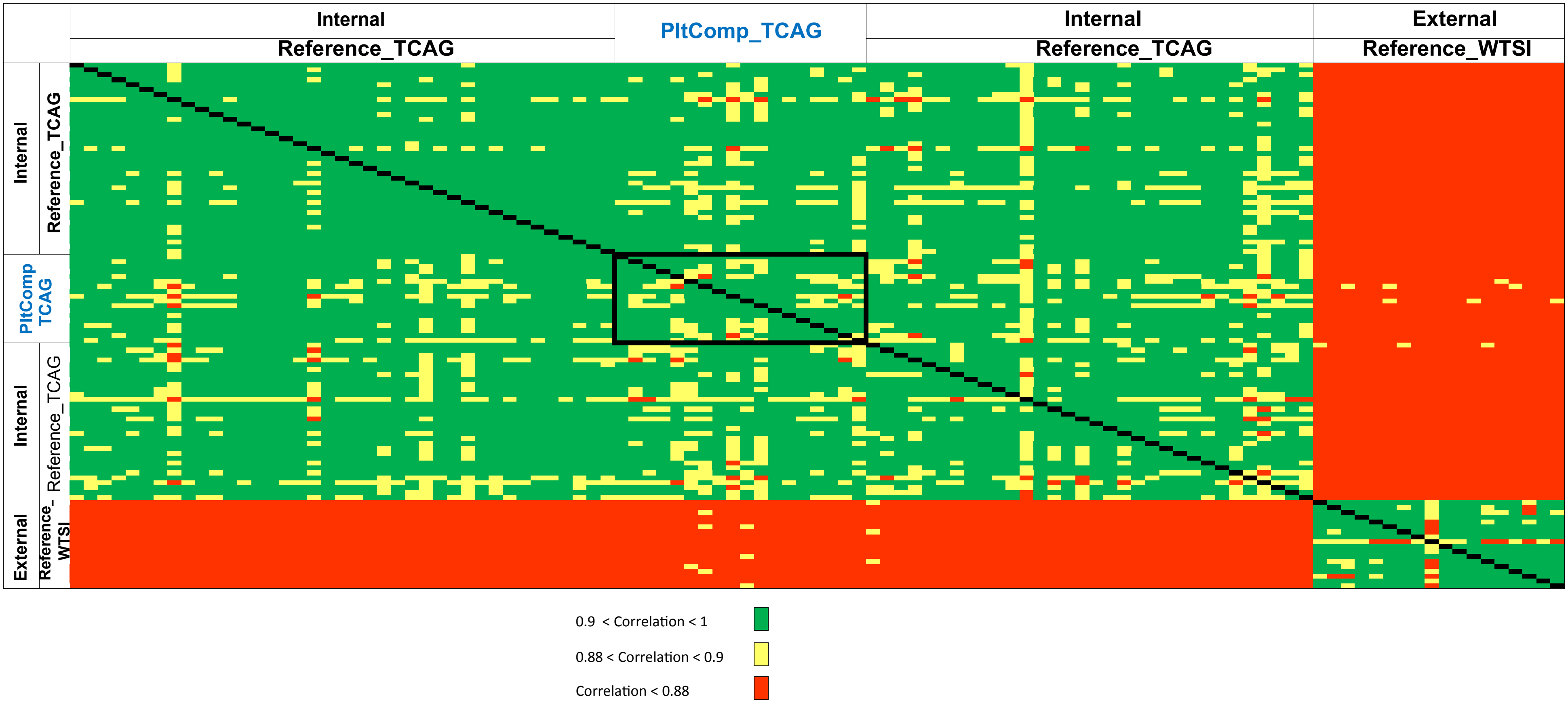
[1] Lower resolution arrays (BAC, Affy250K, AG244K and Illmn650Y) were not used in this analysis. These data allow us to estimate a false negative rate for large calls >50kb, which ranges from 15-77% (averaged across samples) for the different arrays.

Supplementary Fig. 1. Probe coverage per platform. Coverage per chromosome is measured as the #probes per Mb

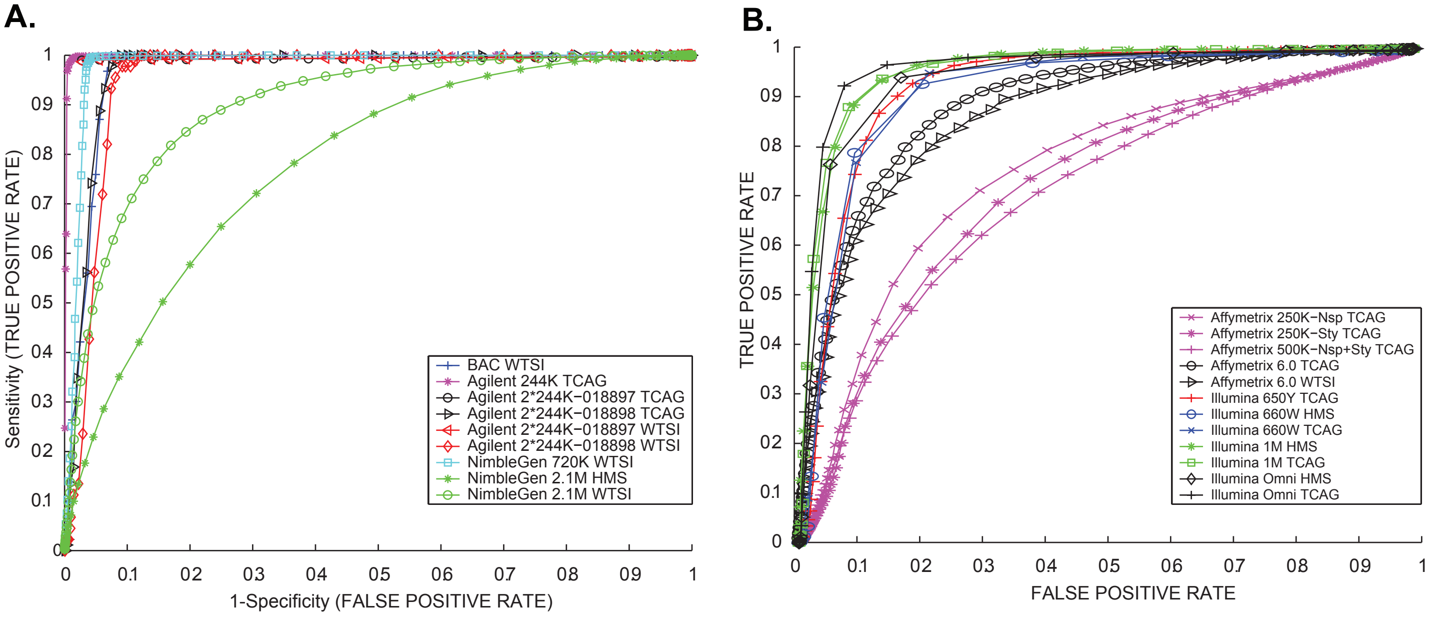


Nature Biotechnology: doi:10.1038/nbt.1852

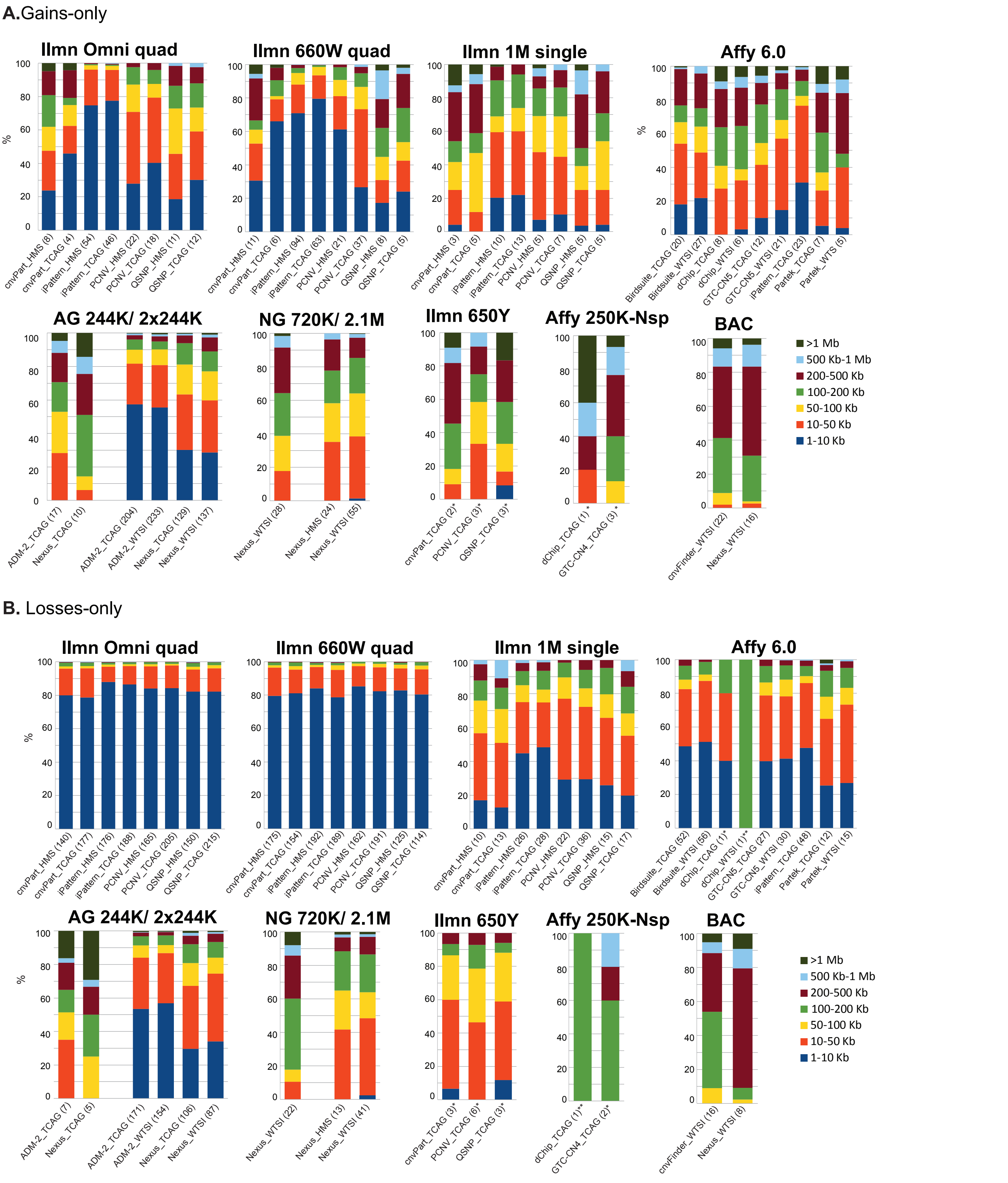
Supplementary Fig. 2. Selection of a suitable reference batch for CNV analysis. This figure shows the clustering of Affymetrix 6.0 arrays according to Person correlation pairwise coefficients of median normalized intensities. Upper left corner corresponds to arrays genotyped at TCAG - inner black box represents 18 platform comparison arrays, which show good correlation with the other TCAG samples from the same genotyping batch (same site or internal reference). Lower right corner contains platform comparison arrays processed at WTSI (external site or external reference) which exhibit poor correlation with those genotyped at TCAG. Correlation value of 0.88 corresponds to the recommended Affymetrix Median Absolute Pairwise Difference (MAPD) metric value of 0.3.



Supplementary Fig. 3. Receiver operating characteristic (ROC) curves based on the comparison between NA15510 vs. NA10851 using probes on chromosome 2 vs. chromosome X. **A.** ROC curves for CGH arrays. **B.** ROC curves for SNP arrays.

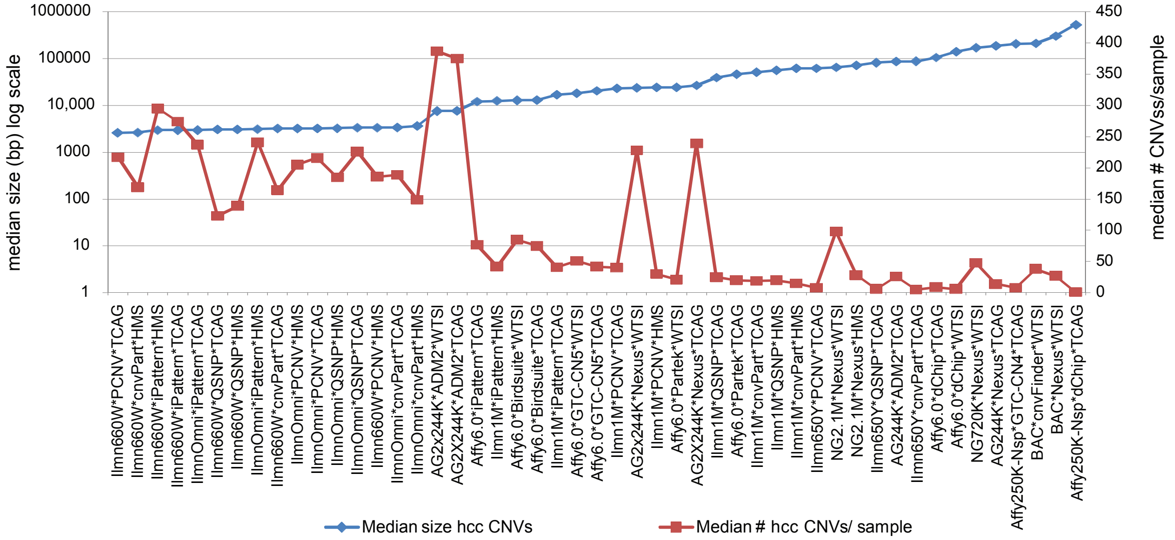


Supplementary Fig. 4. Size distribution as in Figure 1 for **A.** Gains-only and **B.** Losses-only. (*) Note that for some platform/algorithm combinations, not all samples have CNVs for all size bins, the size distribution is therefore not representative of a sample, instead it represents the sizes for CNVs found in a total of six samples. Results are shown for all genotyping sites.



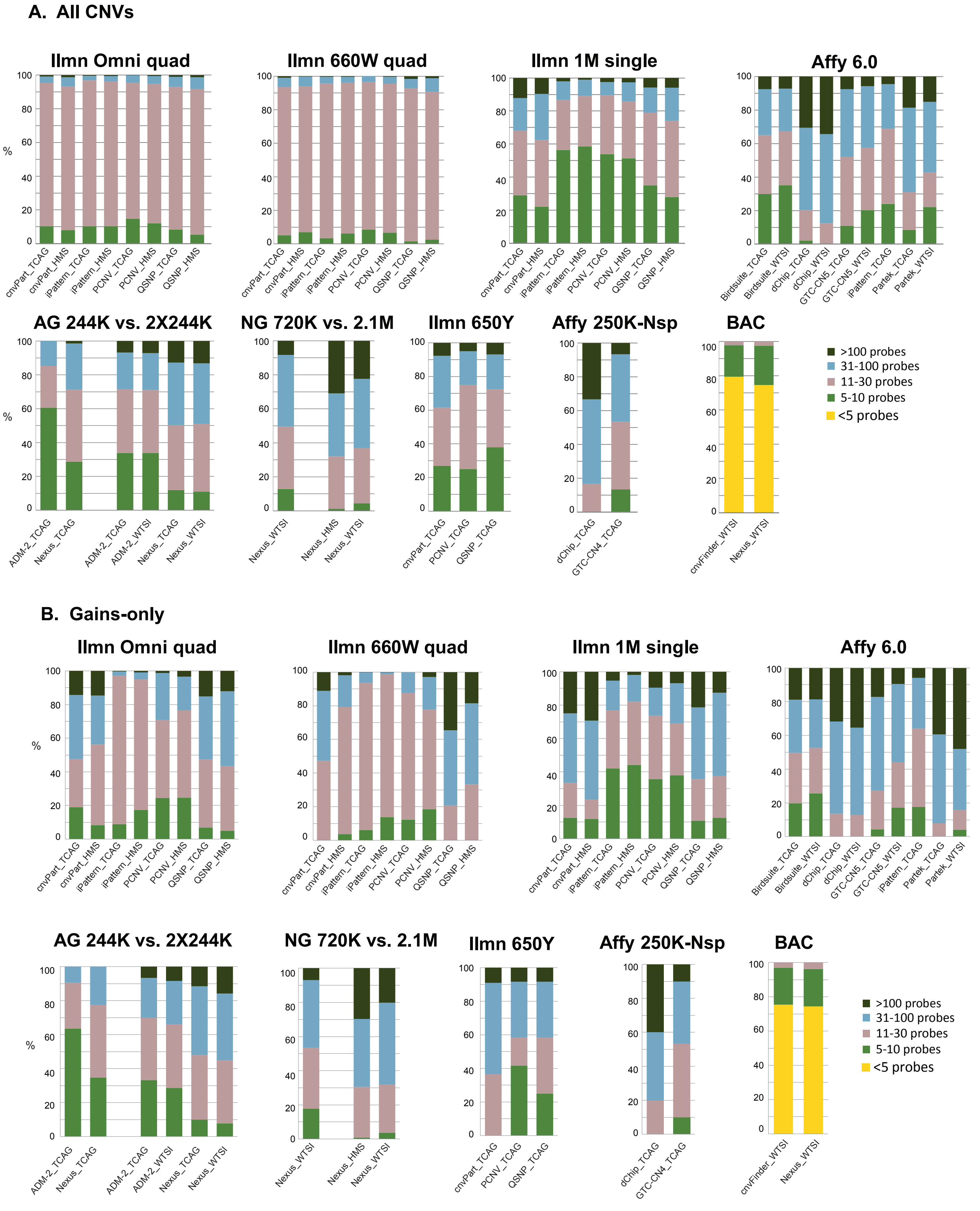
(**) Affy250K-Nsp: dChip detected one loss in sample NA15510 only; Affy6: dChip detected one loss in sample NA12239 only.

Supplementary Fig. 4C. Higher-resolution arrays show a tendency to detect higher number of CNV calls and concomitant smaller CNV sizes



Nature Biotechnology: doi:10.1038/nbt.1852

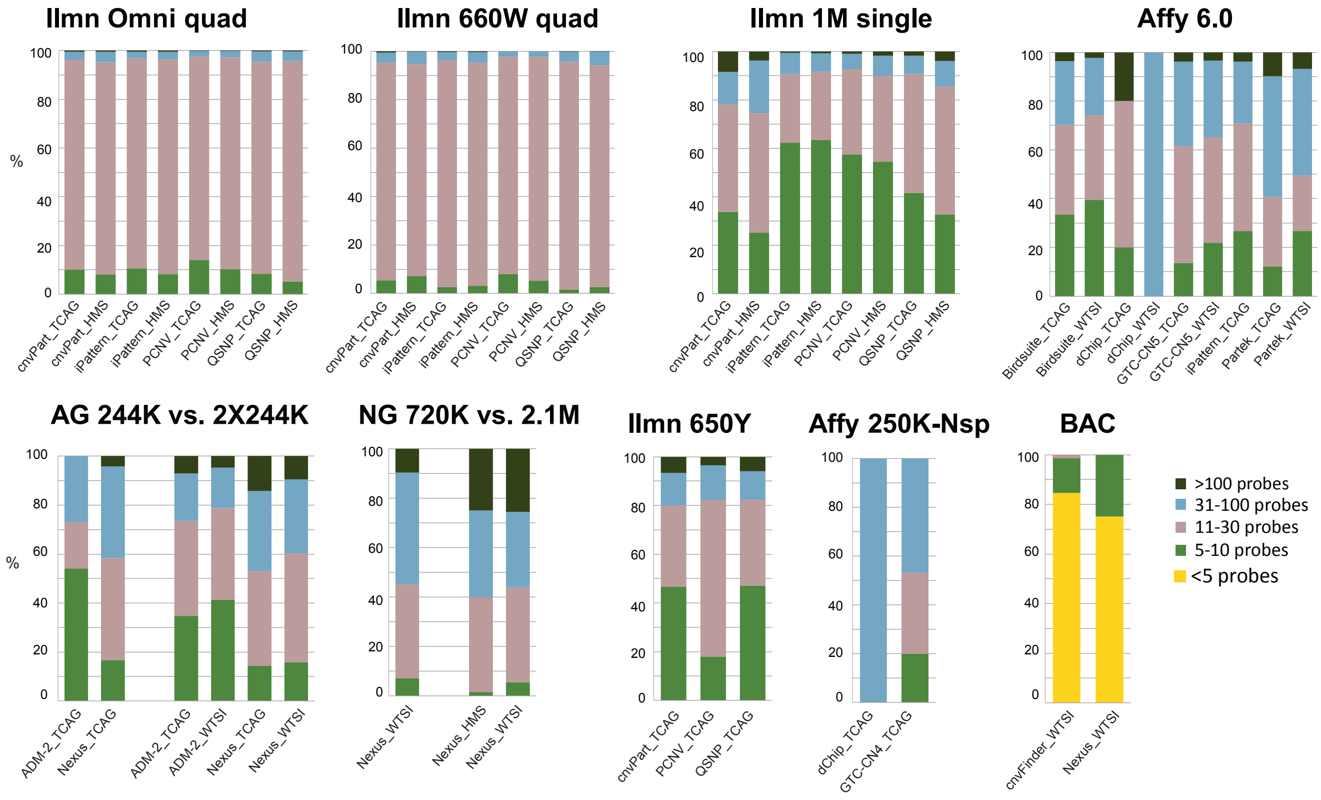
Supplementary Fig. 5. Distribution of the number of probes per CNV for all array/algorithm/site combinations. Each bin depicts the proportion of CNVs with a given number of probes. Five probe bins have been considered. Note that a minimum of 5 probes was required for CNV calling for all platforms except for the BAC arrays where a minimum of one probe (ie. one clone) was considered because of its length. Results are shown for all genotyping sites.


A. All CNVs

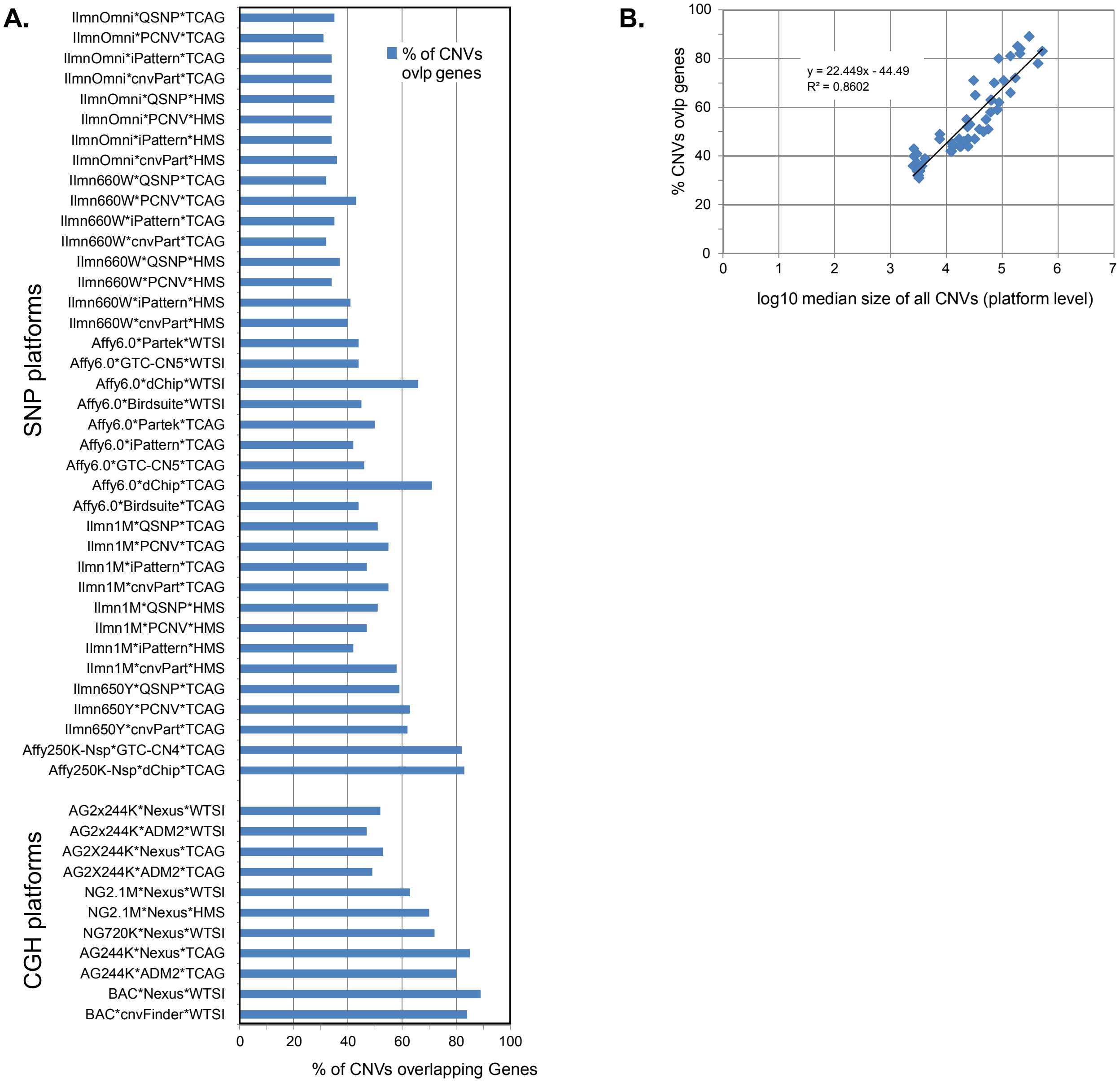
B. Gains-only

Supplementary Fig. 5C. Distribution of the number of probes per CNV for all array/algorithm/site combinations. Each bin depicts the proportion of CNVs with a given number of probes. Five probe bins have been considered. Note that a minimum of 5 probes was required for CNV calling for all platforms except for the BAC arrays where a minimum of one probe (ie. one clone) was considered because of its length. Results are shown for all genotyping sites.

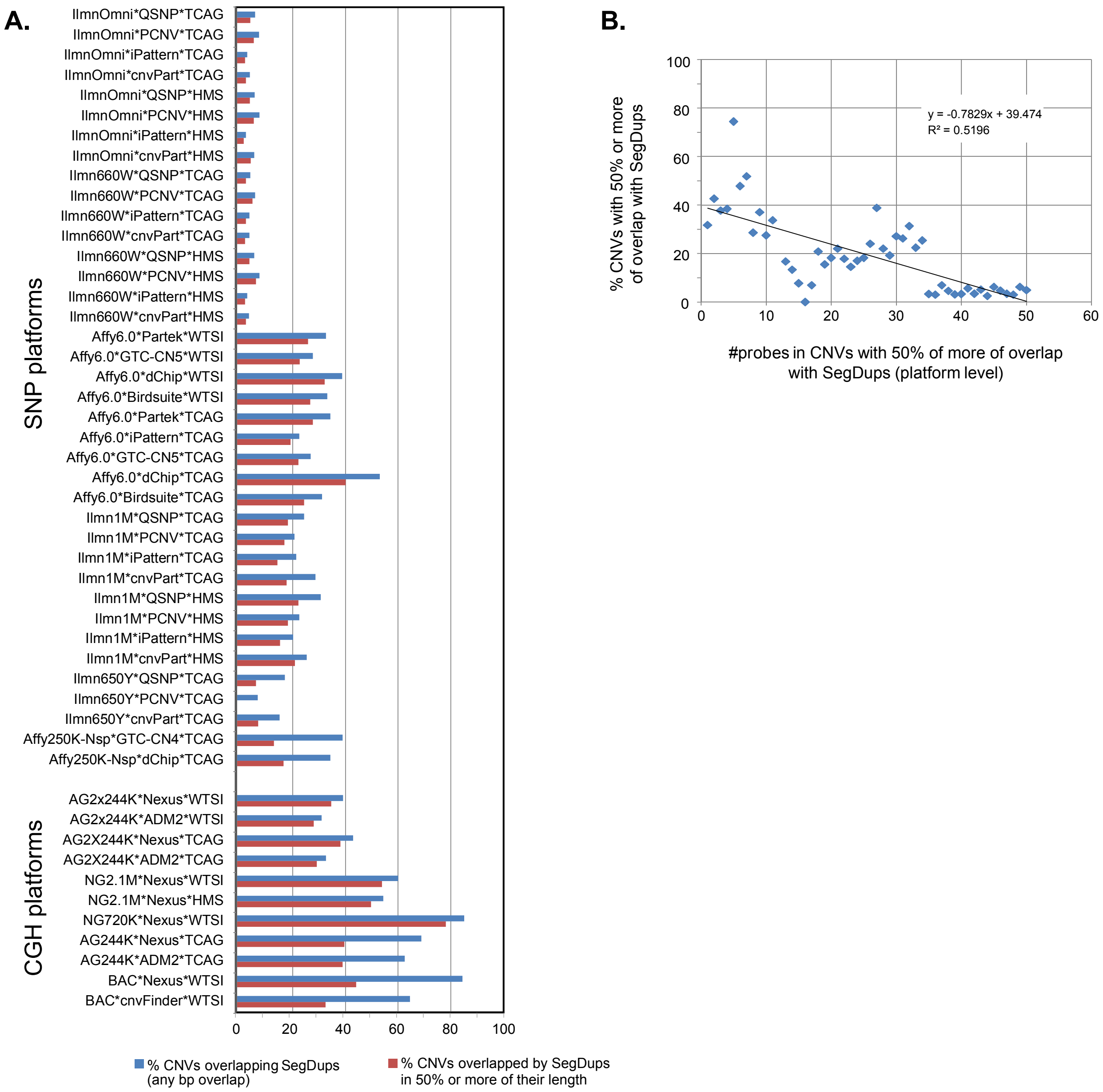
C. Losses-only



Supplementary Fig. 6. Proportion of high-concordant CNVs (hcc) overlapping genes (%)



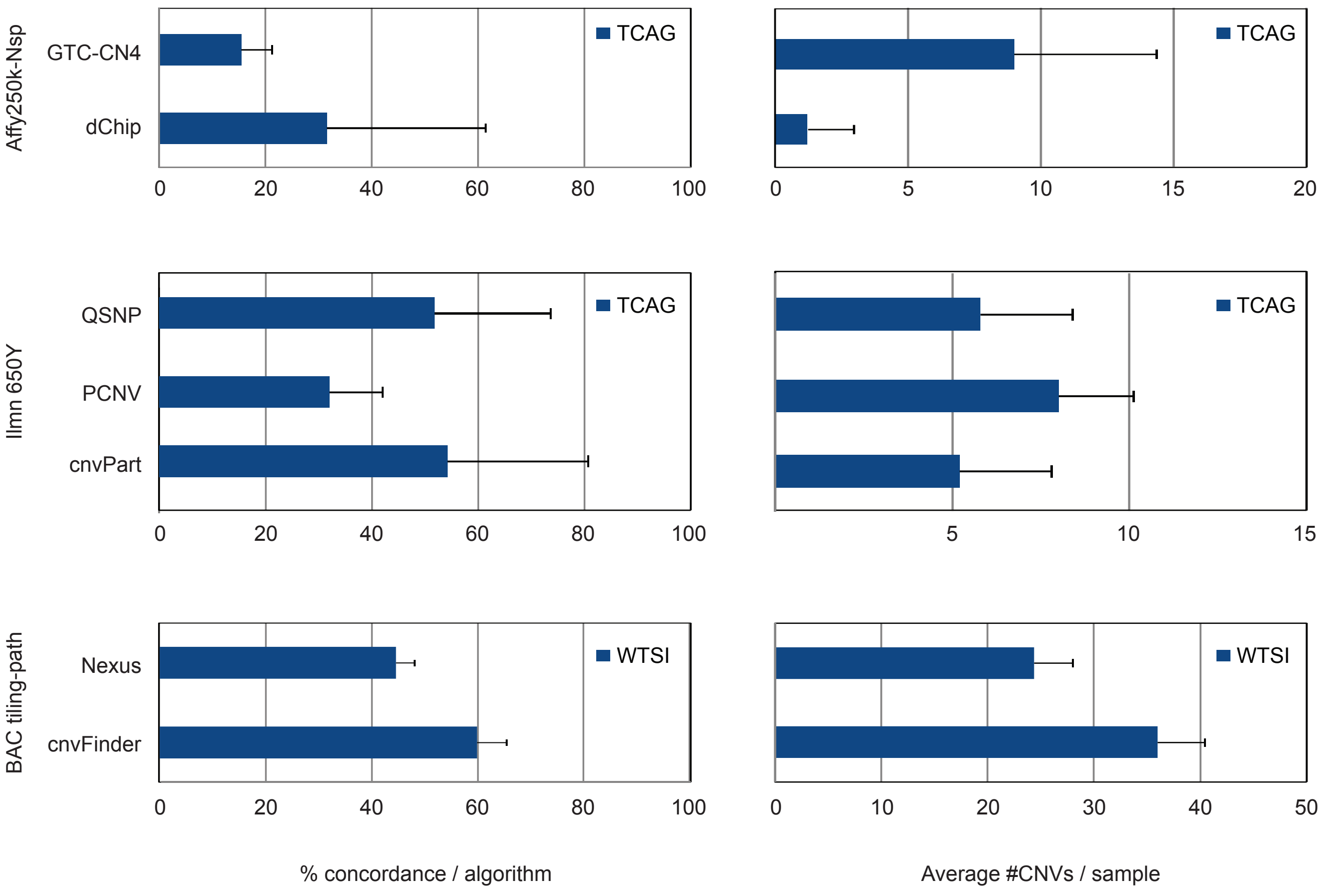
Supplementary Fig. 7. Proportion of high-concordant CNVs (hcc) overlapping segmental duplications (%)



Nature Biotechnology: doi:10.1038/nbt.1852

Supplementary Fig. 8. Call reproducibility.

A) Call concordance between replicates for lower-resolution platforms, where the results for higher-resolution arrays can be found in the main **Figure 2A**. The percentage of concordant CNV calls is shown on the left side, for each combination of array and algorithm, and the corresponding average number of CNVs per sample is displayed to the right side.

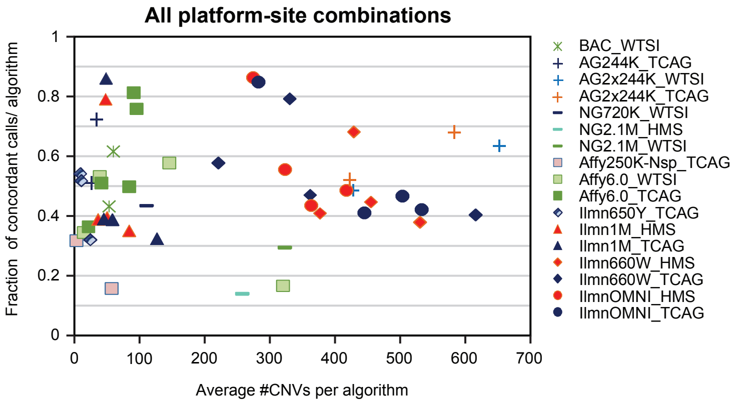


A. Between-replicates concordance for lower-resolution platforms

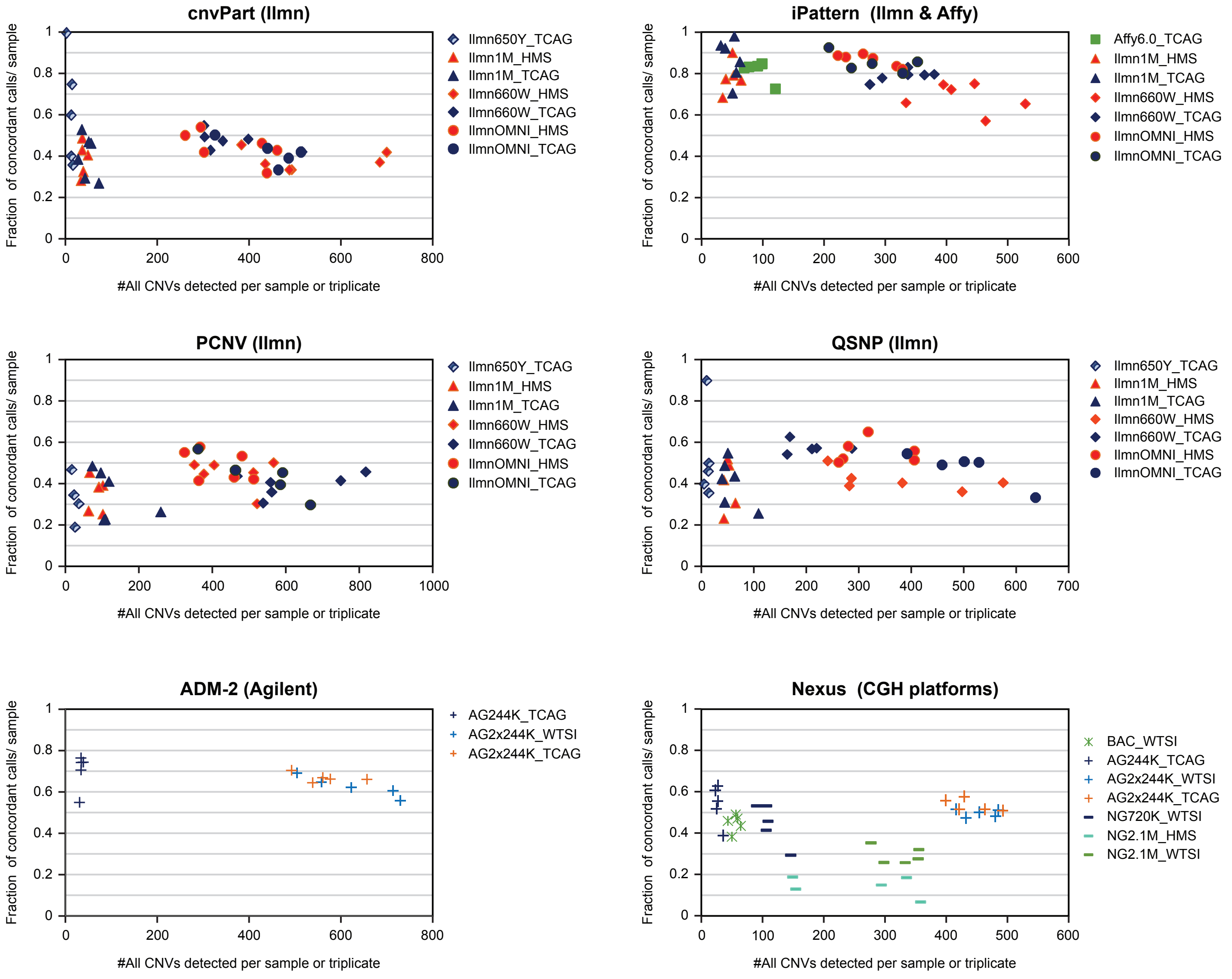
Nature Biotechnology: doi:10.1038/nbt.1852

To complement the main **Figure 2A** and **Supplementary Fig. 8A**, the fraction of concordant calls between replicates is also shown **B)** at the algorithm level, where each symbol represents an algorithm (ie. results were averaged across all six samples); and **C)** at the sample level, where each symbol represents a sample and results are shown for algorithms that can handle more than one platform type.

B. Algorithm level concordance



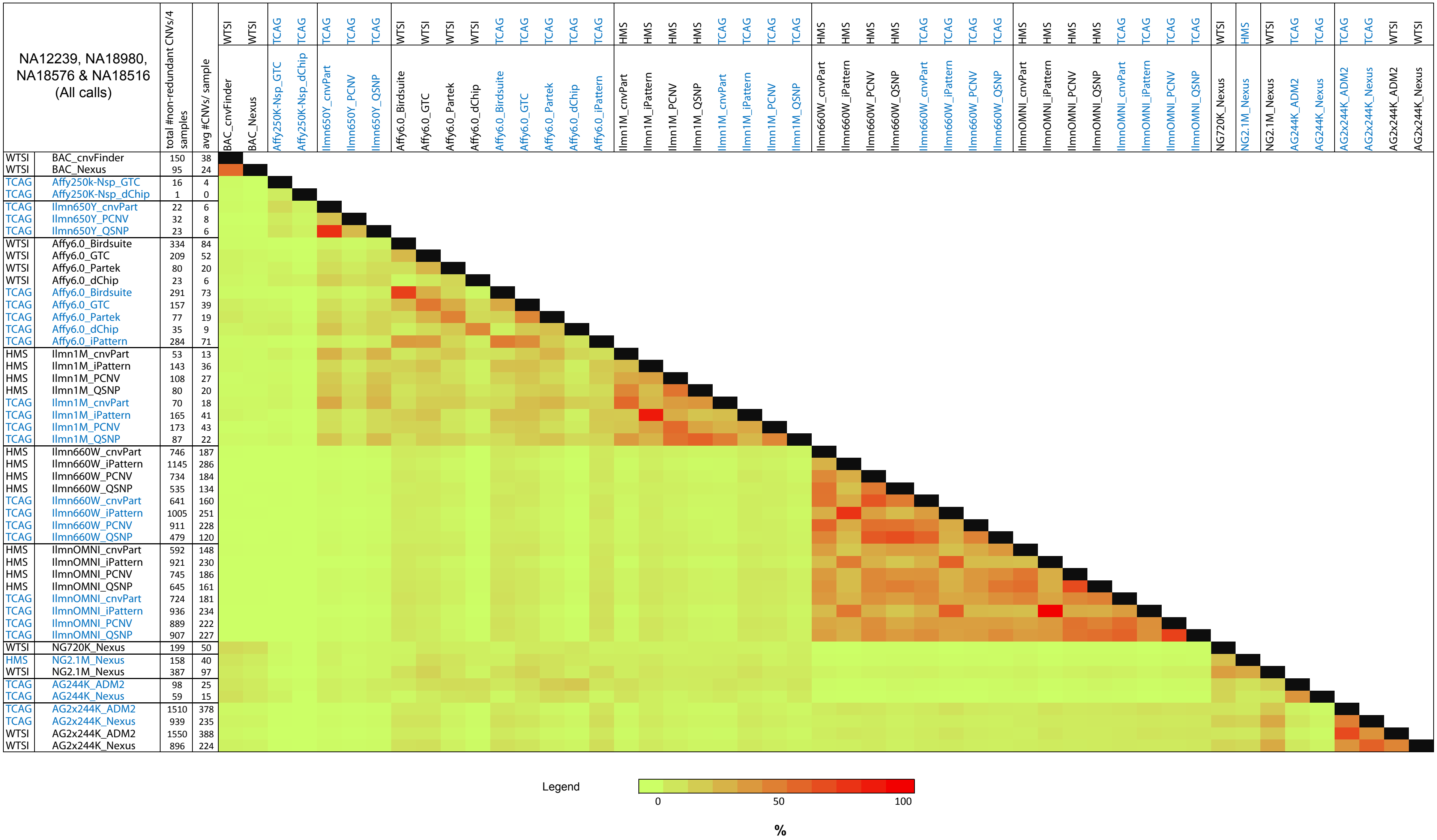
C. Sample level concordance



Nature Biotechnology: doi:10.1038/nbt.1852

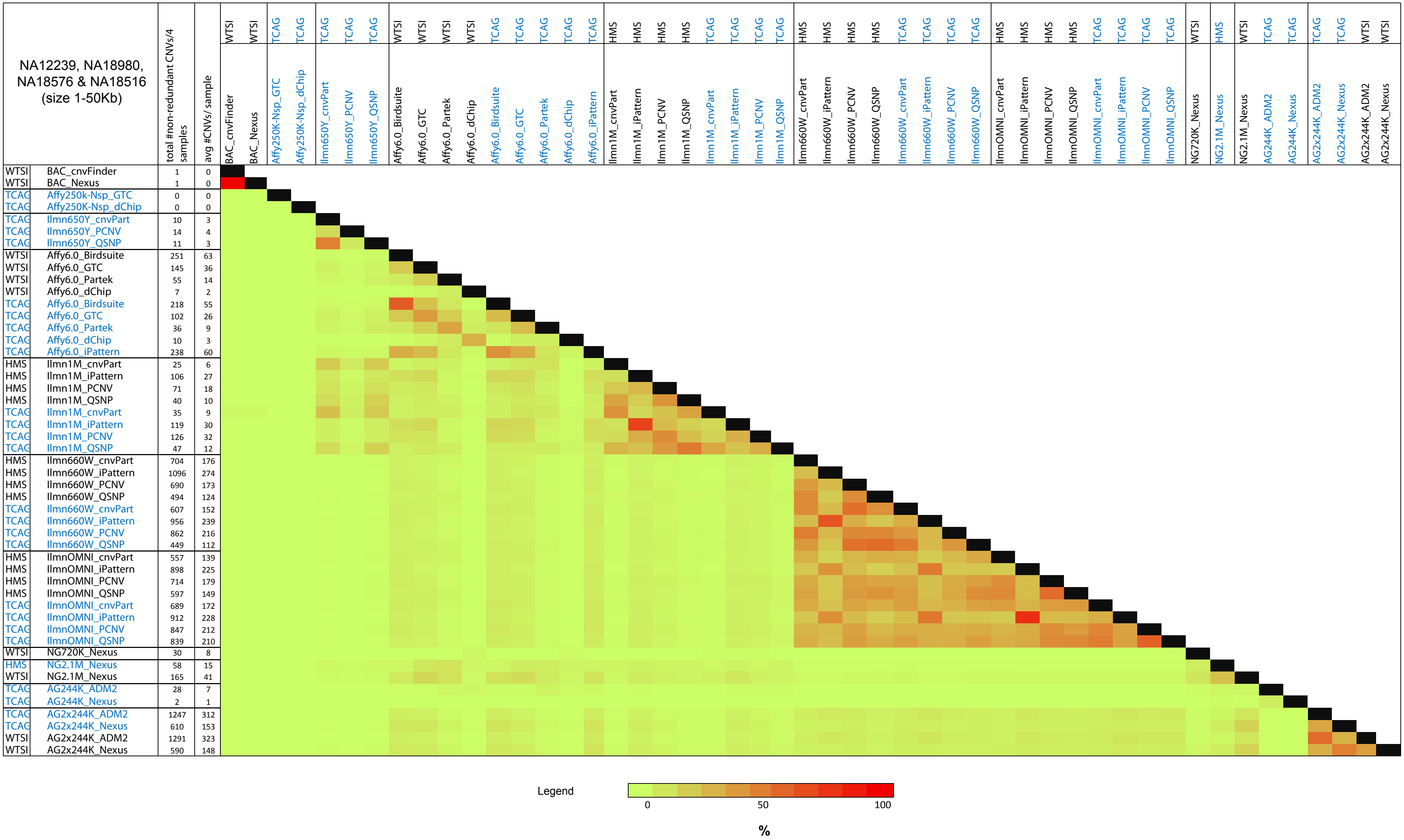
Supplementary Fig. 9. CNV overlap between different combinations of algorithm/platform/site for 4 samples (NA12239, NA18516, NA18576, NA18980) measured using the Jaccard similarity coefficient (1). A. Percentage of all CNVs from 4 samples (union) found by both algorithms (ie. intersection). All calls were considered independently of size or # probes. Results are consistent with **Supplementary Fig.15**. The similarity increases substantially with increase of CNV size (see **B** and **C**).

A.



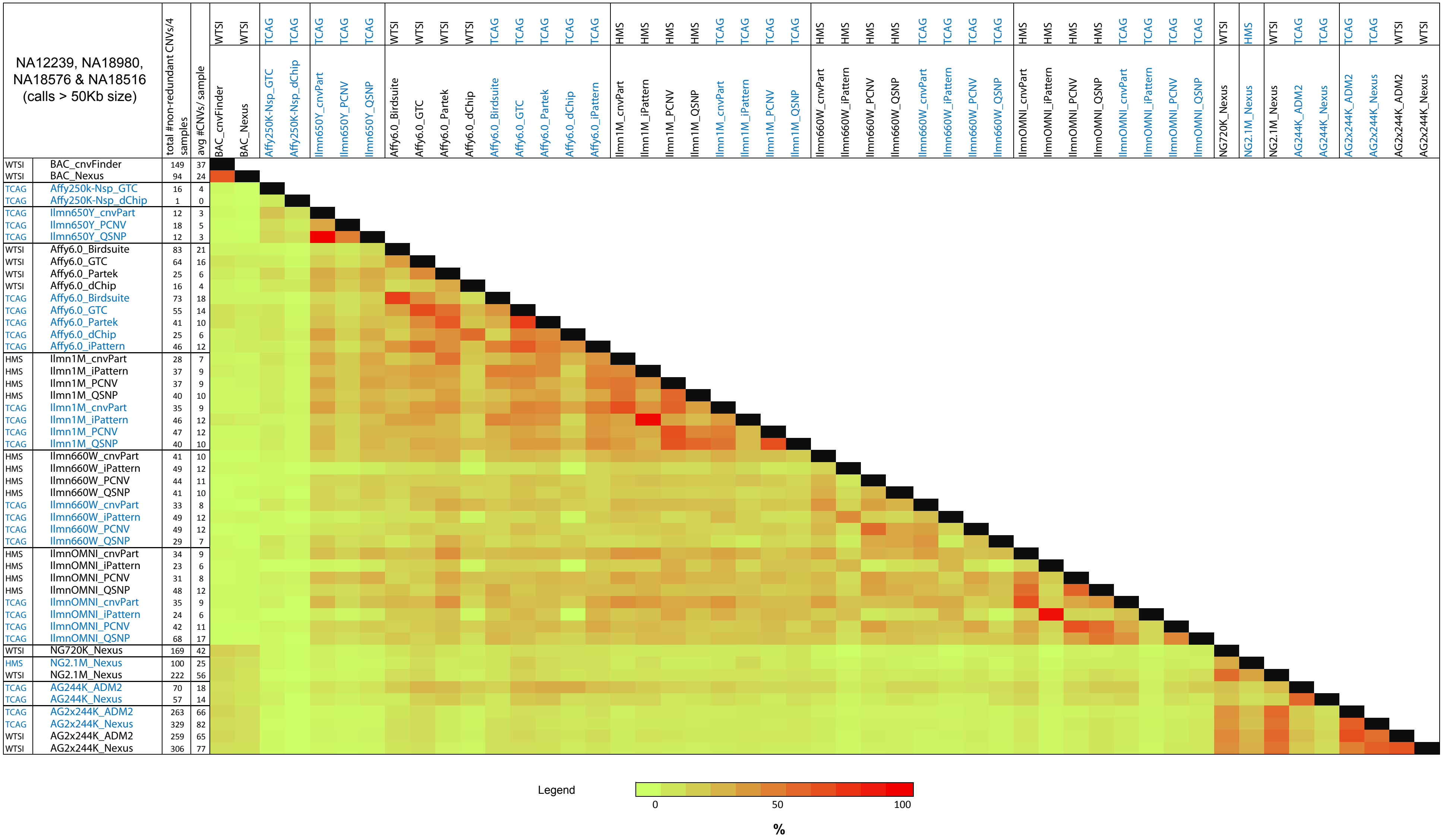
NA12239, NA18980, NA18576 & NA18516 (All calls)		total #non-redundant CNVs/4 samples	avg #CNVs/ sample
WTSI	BAC_cnvFinder	150	38
WTSI	BAC_Nexus	95	24
TCAG	Affy250k-Nsp_GTC	16	4
TCAG	Affy250K-Nsp_dChip	1	0
TCAG	Ilmn650Y_cnvPart	22	6
TCAG	Ilmn650Y_PCNV	32	8
TCAG	Ilmn650Y_QSNP	23	6
WTSI	Affy6.0_Birdsuite	334	84
WTSI	Affy6.0_GTC	209	52
WTSI	Affy6.0_Partek	80	20
WTSI	Affy6.0_dChip	23	6
TCAG	Affy6.0_Birdsuite	291	73
TCAG	Affy6.0_GTC	157	39
TCAG	Affy6.0_Partek	77	19
TCAG	Affy6.0_dChip	35	9
TCAG	Affy6.0_iPattern	284	71
HMS	Ilmn1M_cnvPart	53	13
HMS	Ilmn1M_iPattern	143	36
HMS	Ilmn1M_PCNV	108	27
HMS	Ilmn1M_QSNP	80	20
TCAG	Ilmn1M_cnvPart	70	18
TCAG	Ilmn1M_iPattern	165	41
TCAG	Ilmn1M_PCNV	173	43
TCAG	Ilmn1M_QSNP	87	22
HMS	Ilmn660W_cnvPart	746	187
HMS	Ilmn660W_iPattern	1145	286
HMS	Ilmn660W_PCNV	734	184
HMS	Ilmn660W_QSNP	535	134
TCAG	Ilmn660W_cnvPart	641	160
TCAG	Ilmn660W_iPattern	1005	251
TCAG	Ilmn660W_PCNV	911	228
TCAG	Ilmn660W_QSNP	479	120
HMS	IlmnOMNI_cnvPart	592	148
HMS	IlmnOMNI_iPattern	921	230
HMS	IlmnOMNI_PCNV	745	186
HMS	IlmnOMNI_QSNP	645	161
TCAG	IlmnOMNI_cnvPart	724	181
TCAG	IlmnOMNI_iPattern	936	234
TCAG	IlmnOMNI_PCNV	889	222
TCAG	IlmnOMNI_QSNP	907	227
WTSI	NG720K_Nexus	199	50
HMS	NG2.1M_Nexus	158	40
WTSI	NG2.1M_Nexus	387	97
TCAG	AG244K_ADM2	98	25
TCAG	AG244K_Nexus	59	15
TCAG	AG2x244K_ADM2	1510	378
TCAG	AG2x244K_Nexus	939	235
WTSI	AG2x244K_ADM2	1550	388
WTSI	AG2x244K_Nexus	896	224

B. Percentage of all CNVs from 4 samples (union) found by both algorithms (ie. intersection), for CNV sizes < 50 Kb



NA12239, NA18980, NA18576 & NA18516 (size 1-50Kb)		total #non-redundant CNVs/4 samples	avg #CNVs/ sample
WTSI	BAC_cnvFinder	1	0
WTSI	BAC_Nexus	1	0
TCAG	Affy250k-Nsp_GTC	0	0
TCAG	Affy250K-Nsp_dChip	0	0
TCAG	Ilmn650Y_cnvPart	10	3
TCAG	Ilmn650Y_PCNV	14	4
TCAG	Ilmn650Y_QSNP	11	3
WTSI	Affy6.0_Birdsuite	251	63
WTSI	Affy6.0_GTC	145	36
WTSI	Affy6.0_Partek	55	14
WTSI	Affy6.0_dChip	7	2
TCAG	Affy6.0_Birdsuite	218	55
TCAG	Affy6.0_GTC	102	26
TCAG	Affy6.0_Partek	36	9
TCAG	Affy6.0_dChip	10	3
TCAG	Affy6.0_iPattern	238	60
HMS	Ilmn1M_cnvPart	25	6
HMS	Ilmn1M_iPattern	106	27
HMS	Ilmn1M_PCNV	71	18
HMS	Ilmn1M_QSNP	40	10
TCAG	Ilmn1M_cnvPart	35	9
TCAG	Ilmn1M_iPattern	119	30
TCAG	Ilmn1M_PCNV	126	32
TCAG	Ilmn1M_QSNP	47	12
HMS	Ilmn660W_cnvPart	704	176
HMS	Ilmn660W_iPattern	1096	274
HMS	Ilmn660W_PCNV	690	173
HMS	Ilmn660W_QSNP	494	124
TCAG	Ilmn660W_cnvPart	607	152
TCAG	Ilmn660W_iPattern	956	239
TCAG	Ilmn660W_PCNV	862	216
TCAG	Ilmn660W_QSNP	449	112
HMS	IlmnOMNI_cnvPart	557	139
HMS	IlmnOMNI_iPattern	898	225
HMS	IlmnOMNI_PCNV	714	179
HMS	IlmnOMNI_QSNP	597	149
TCAG	IlmnOMNI_cnvPart	689	172
TCAG	IlmnOMNI_iPattern	912	228
TCAG	IlmnOMNI_PCNV	847	212
TCAG	IlmnOMNI_QSNP	839	210
WTSI	NG720K_Nexus	30	8
HMS	NG2.1M_Nexus	58	15
WTSI	NG2.1M_Nexus	165	41
TCAG	AG244K_ADM2	28	7
TCAG	AG244K_Nexus	2	1
TCAG	AG2x244K_ADM2	1247	312
TCAG	AG2x244K_Nexus	610	153
WTSI	AG2x244K_ADM2	1291	323
WTSI	AG2x244K_Nexus	590	148

C. Percentage of all CNVs from 4 samples (union) found by both algorithms (ie. intersection), for CNV sizes >50 Kb.

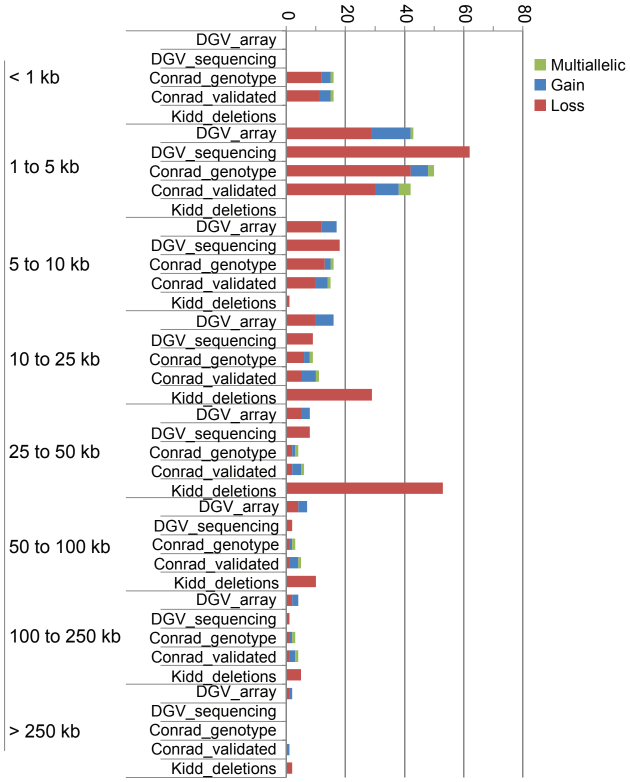


NA12239, NA18980, NA18576 & NA18516 (calls > 50Kb size)

		total #non-redundant CNVs/4 samples	avg #CNVs/ sample
WTSI	BAC_cnvFinder	149	37
WTSI	BAC_Nexus	94	24
TCAG	Affy250k-Nsp_GTC	16	4
TCAG	Affy250K-Nsp_dChip	1	0
TCAG	Ilmn650Y_cnvPart	12	3
TCAG	Ilmn650Y_PCNV	18	5
TCAG	Ilmn650Y_QSNP	12	3
WTSI	Affy6.0_Birdsuite	83	21
WTSI	Affy6.0_GTC	64	16
WTSI	Affy6.0_Partek	25	6
WTSI	Affy6.0_dChip	16	4
TCAG	Affy6.0_Birdsuite	73	18
TCAG	Affy6.0_GTC	55	14
TCAG	Affy6.0_Partek	41	10
TCAG	Affy6.0_dChip	25	6
TCAG	Affy6.0_iPattern	46	12
HMS	Ilmn1M_cnvPart	28	7
HMS	Ilmn1M_iPattern	37	9
HMS	Ilmn1M_PCNV	37	9
HMS	Ilmn1M_QSNP	40	10
TCAG	Ilmn1M_cnvPart	35	9
TCAG	Ilmn1M_iPattern	46	12
TCAG	Ilmn1M_PCNV	47	12
TCAG	Ilmn1M_QSNP	40	10
HMS	Ilmn660W_cnvPart	41	10
HMS	Ilmn660W_iPattern	49	12
HMS	Ilmn660W_PCNV	44	11
HMS	Ilmn660W_QSNP	41	10
TCAG	Ilmn660W_cnvPart	33	8
TCAG	Ilmn660W_iPattern	49	12
TCAG	Ilmn660W_PCNV	49	12
TCAG	Ilmn660W_QSNP	29	7
HMS	IlmnOMNI_cnvPart	34	9
HMS	IlmnOMNI_iPattern	23	6
HMS	IlmnOMNI_PCNV	31	8
HMS	IlmnOMNI_QSNP	48	12
TCAG	IlmnOMNI_cnvPart	35	9
TCAG	IlmnOMNI_iPattern	24	6
TCAG	IlmnOMNI_PCNV	42	11
TCAG	IlmnOMNI_QSNP	68	17
WTSI	NG720K_Nexus	169	42
HMS	NG2.1M_Nexus	100	25
WTSI	NG2.1M_Nexus	222	56
TCAG	AG244K_ADM2	70	18
TCAG	AG244K_Nexus	57	14
TCAG	AG2x244K_ADM2	263	66
TCAG	AG2x244K_Nexus	329	82
WTSI	AG2x244K_ADM2	259	65
WTSI	AG2x244K_Nexus	306	77

Nature Biotechnology: doi:10.1038/nbt.1852

Supplementary Fig. 10. Proportion of all variants (%) for five reference datasets used in the main figure 2C, binned by size and type.



	MinSize	MaxSize	Q1	MedianSize	Q3	AverageSize	Total counts
DGV_array-based*	1,000	4,564,802	2,599	6,517	22,100	33,283	35,030
DGV_sequencing-based*	1,000	1,185,626	1,577	2,978	7,700	11,215	10,843
Conrad_validated	444	1,102,849	1,382	3,539	11,479	20,099	8,599
Conrad_genotyped	447	1,102,849	1,415	3,068	7,126	14,476	4,978
Kidd_deletions	6,856	930,251	23,039	33,270	45,148	48,216	1,157

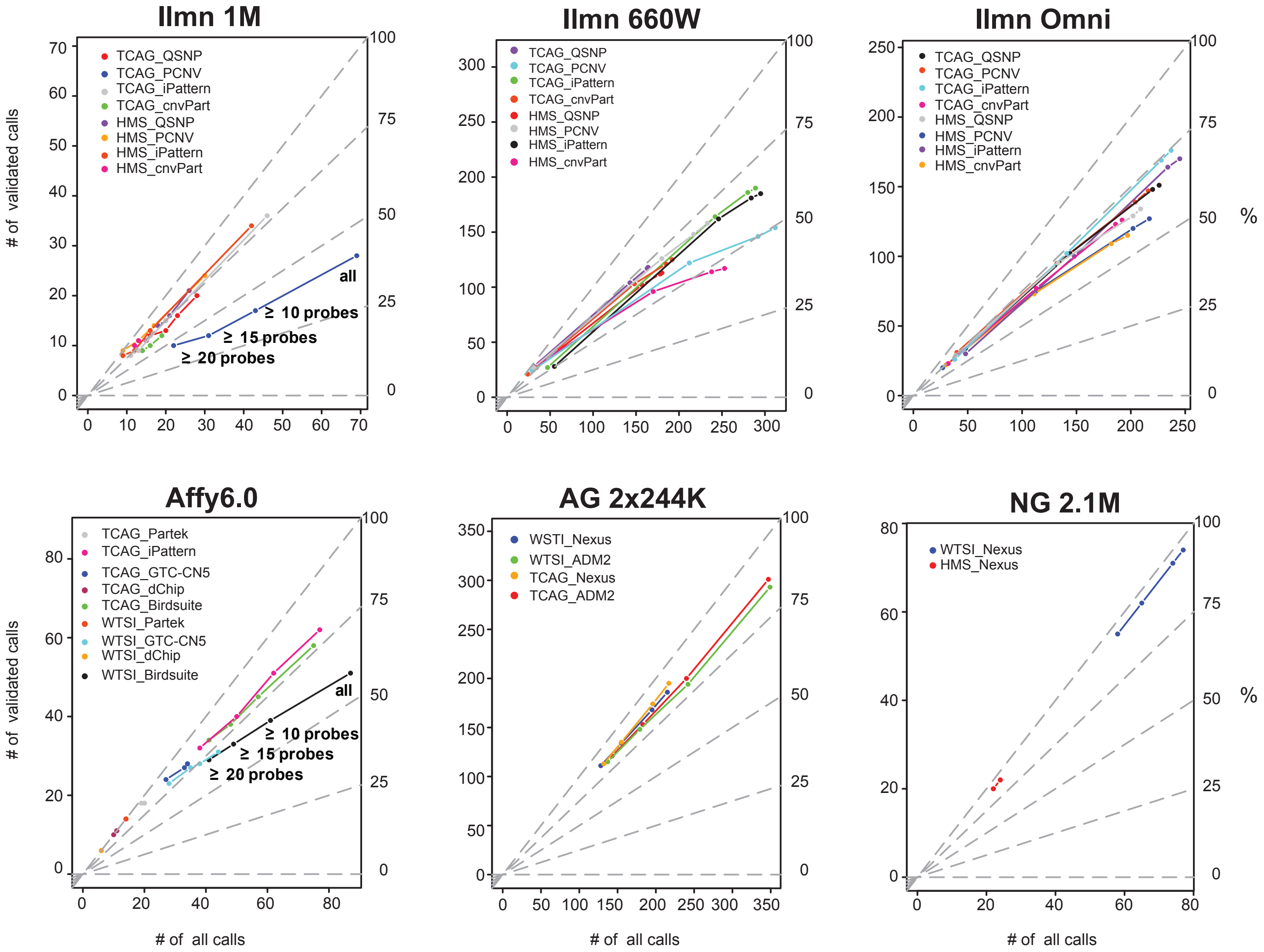
* List of studies listed in the DGV (variation.hg18.v9.mar.2010) that have been excluded:
i) BAC studies
ii) Affy500K studies
iii) ROMA (average size ~400Kb; Sebat et al 2004)
iv) studies that are not genome-wide (ie. FISH, MLPA, PCR)

Specifically, data from the following 11 studies were excluded:
Locke et al. (2006)
Wong et al. (2007)
Sharp et al. (2005)
Iafrate et al. (2004)
Zogopoulos et al. (2007)
Redon et al. (2006)
Pinto et al. (2007)
Giglio et al. (2002)
Young et al. (2008)
Feuk et al. (2005)
Sebat et al. (2004)

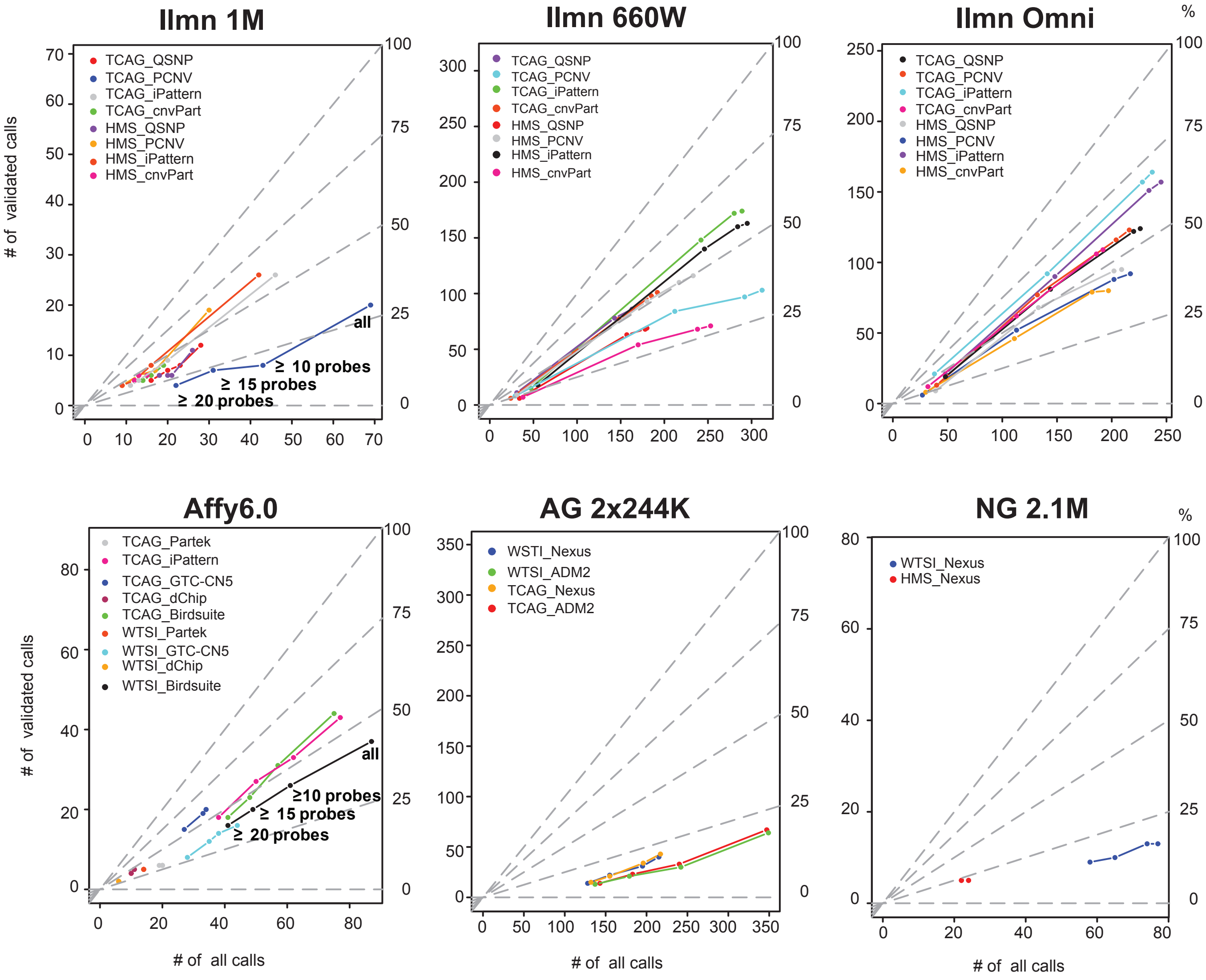
Variants on chromosomes X and Y, and those overlapping pericentromeric or telomeric regions were not considered.

Supplementary Fig. 11. Sample level comparisons for NA12239 with two versions of the reference dataset Conrad et al. 2010: **A.** discovery (all) CNVs; **B.** genotyped-only. CNVs were considered validated when there was a reciprocal overlap of 50% or greater with the reference 11A or 11B. For each comparison, the number (and %) of validated calls are shown, with further breakdown by their number of probes (≥20 probes, ≥15 probes, ≥10 probes and 'all').

A. discovery (all), any overlap

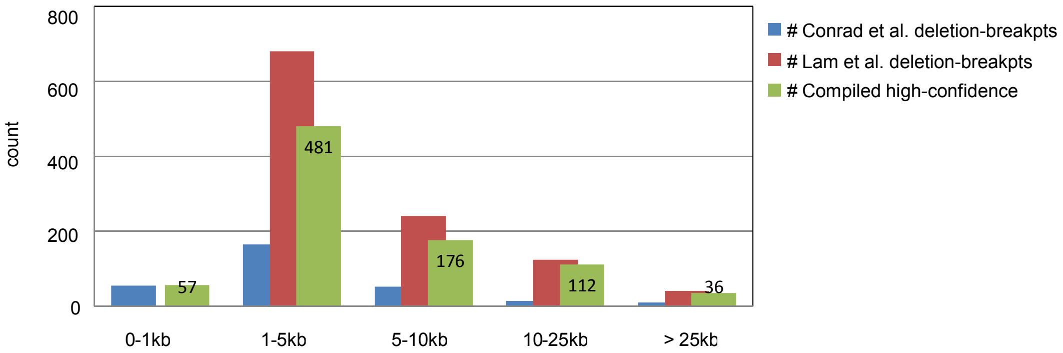


B. genotyped-only, any overlap



Nature Biotechnology: doi:10.1038/nbt.1852

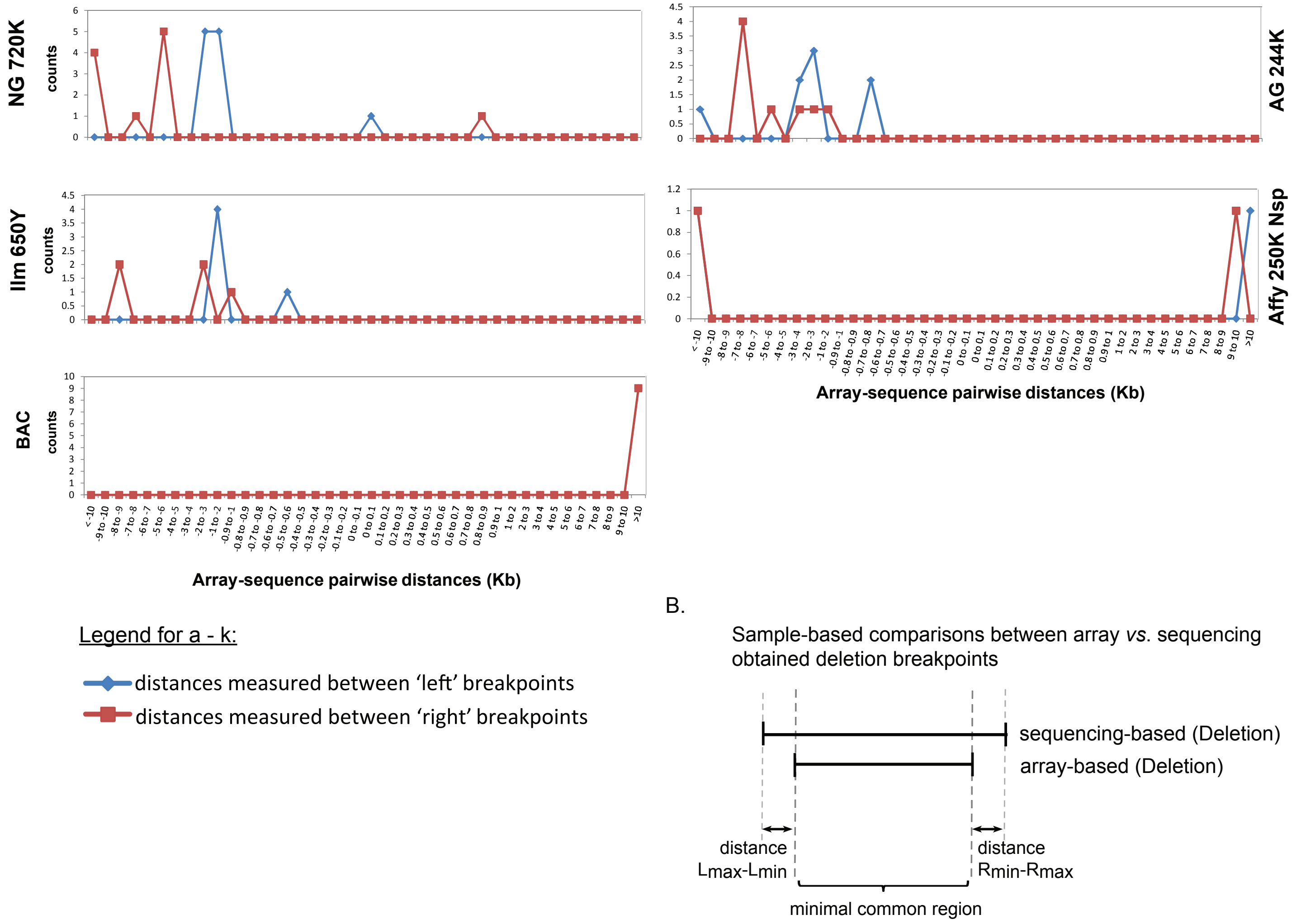
Supplementary Fig. 12. Size distribution for the reference set deletion breakpoints assembled from various published sequencing studies, Conrad et al 2010 and Lam et al. 2010, and used in the main Figure 4A and Supplementary Fig. 13.



Nature Biotechnology: doi:10.1038/nbt.1852

Supplementary Fig. 13. Evaluation of array CNV breakpoint accuracy by comparing to a reference set of deletions detected using sequencing methods, and compiled from various studies (Conrad et al. 2010 and Lam et al. 2010). These results are complementary to the main Figure 4A - here results are shown for lower-resolution arrays only.

A. Lower resolution arrays

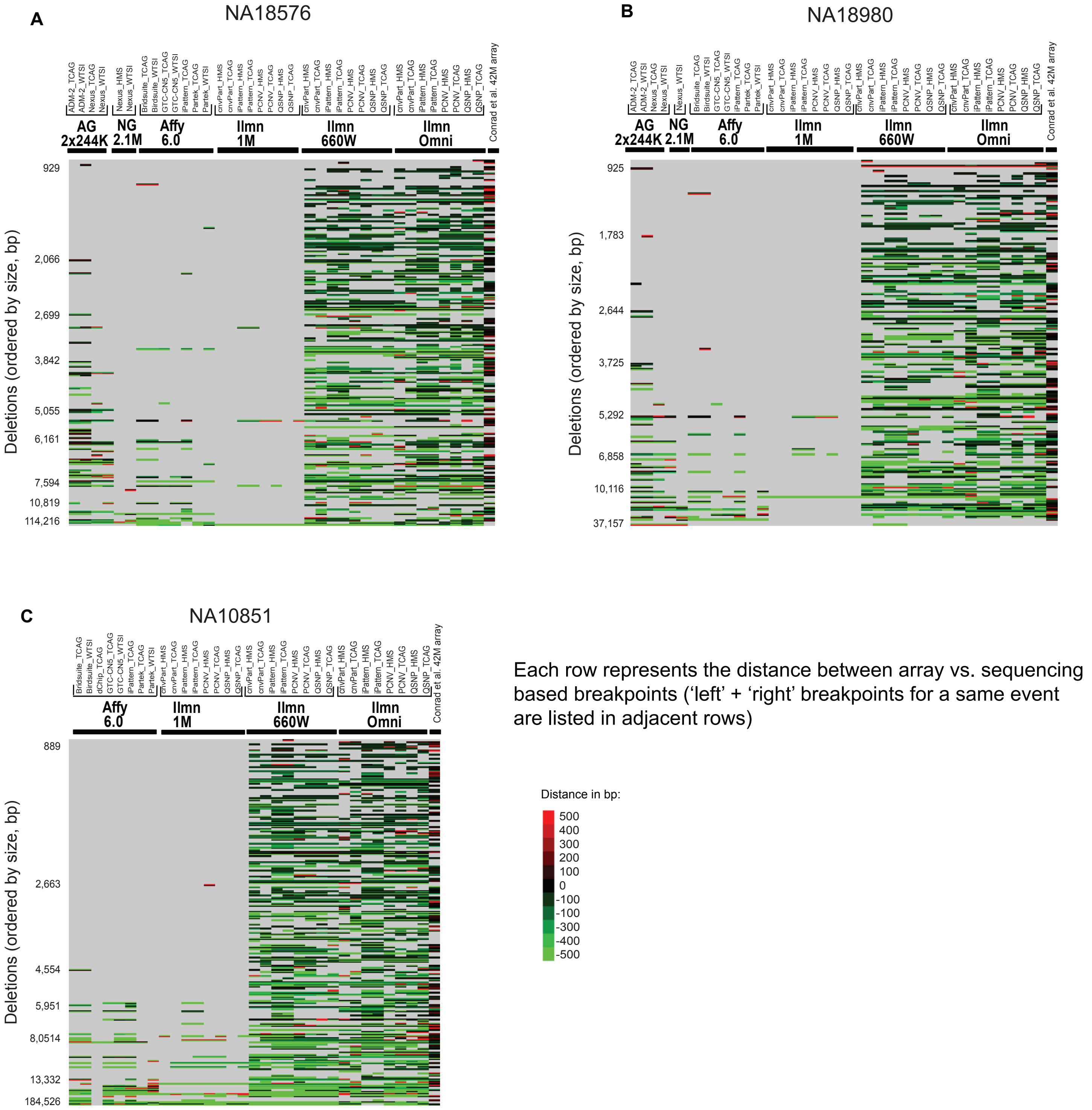


B.

Sample-based comparisons between array *vs.* sequencing obtained deletion breakpoints

Nature Biotechnology: doi:10.1038/nbt.1852

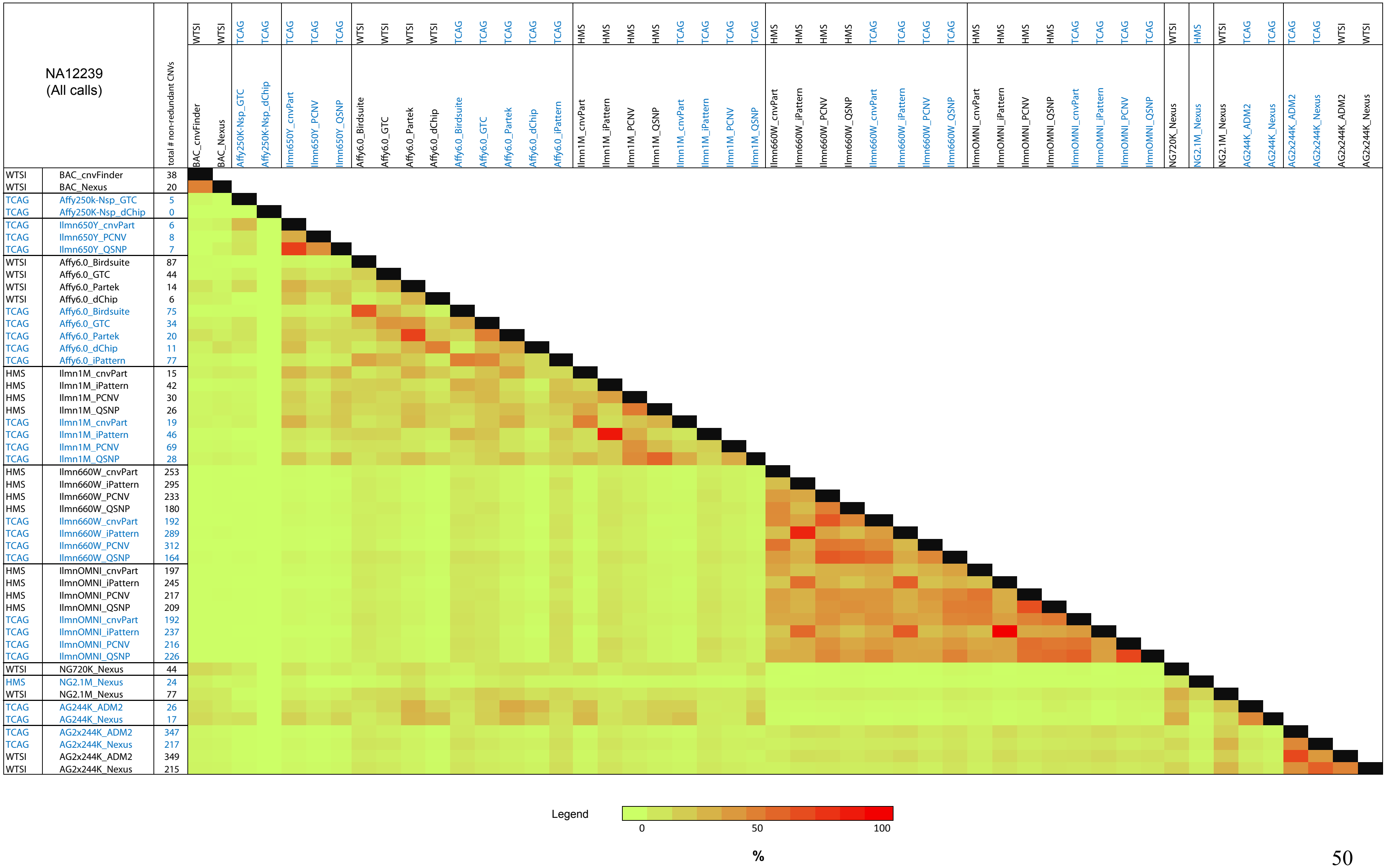
Supplementary Fig. 14. Sample level comparison array-based vs. 1000G deletions for **A.** NA18576; **B.** NA18990, and **C.** NA10851



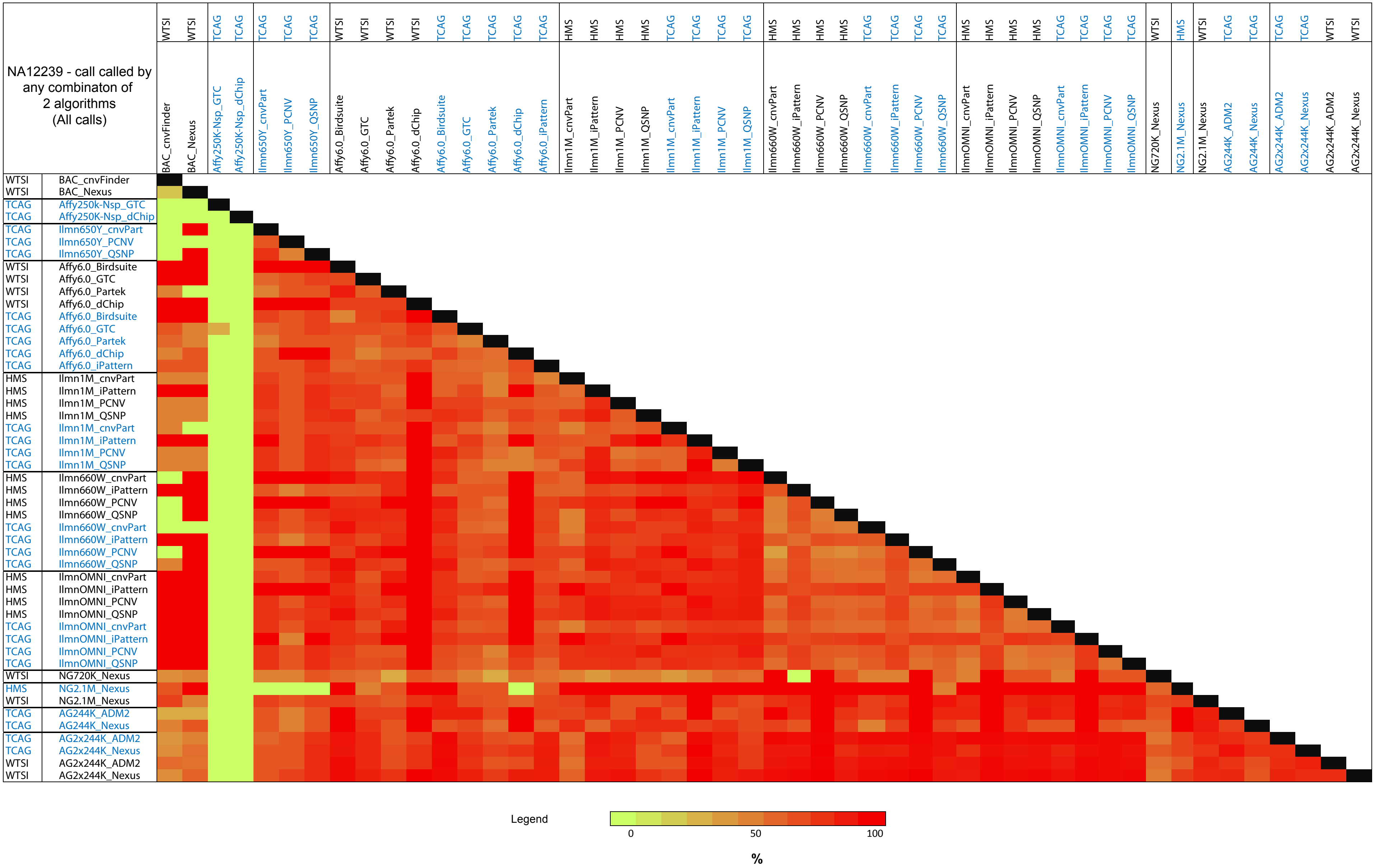
Each row represents the distance between array vs. sequencing based breakpoints (‘left’ + ‘right’ breakpoints for a same event are listed in adjacent rows)

Supplementary Fig. 15. Sample-level overlaps for various combinations of algorithm/platform/site for a single sample NA12239 measured using the Jaccard similarity coefficient. **A.** Percentage of CNVs found by both algorithms (ie. intersection) out of all CNVs detected by each of the two algorithms (ie. union). All calls were considered independently of size or # probes. The variability between algorithms is higher than the variability between sites. Within a platform, there is typically 25-50% overlap between datasets. Across platforms, this number drops substantially. Variability between platforms depends on CNV size (**Supplementary Fig. 9B and 9C**).

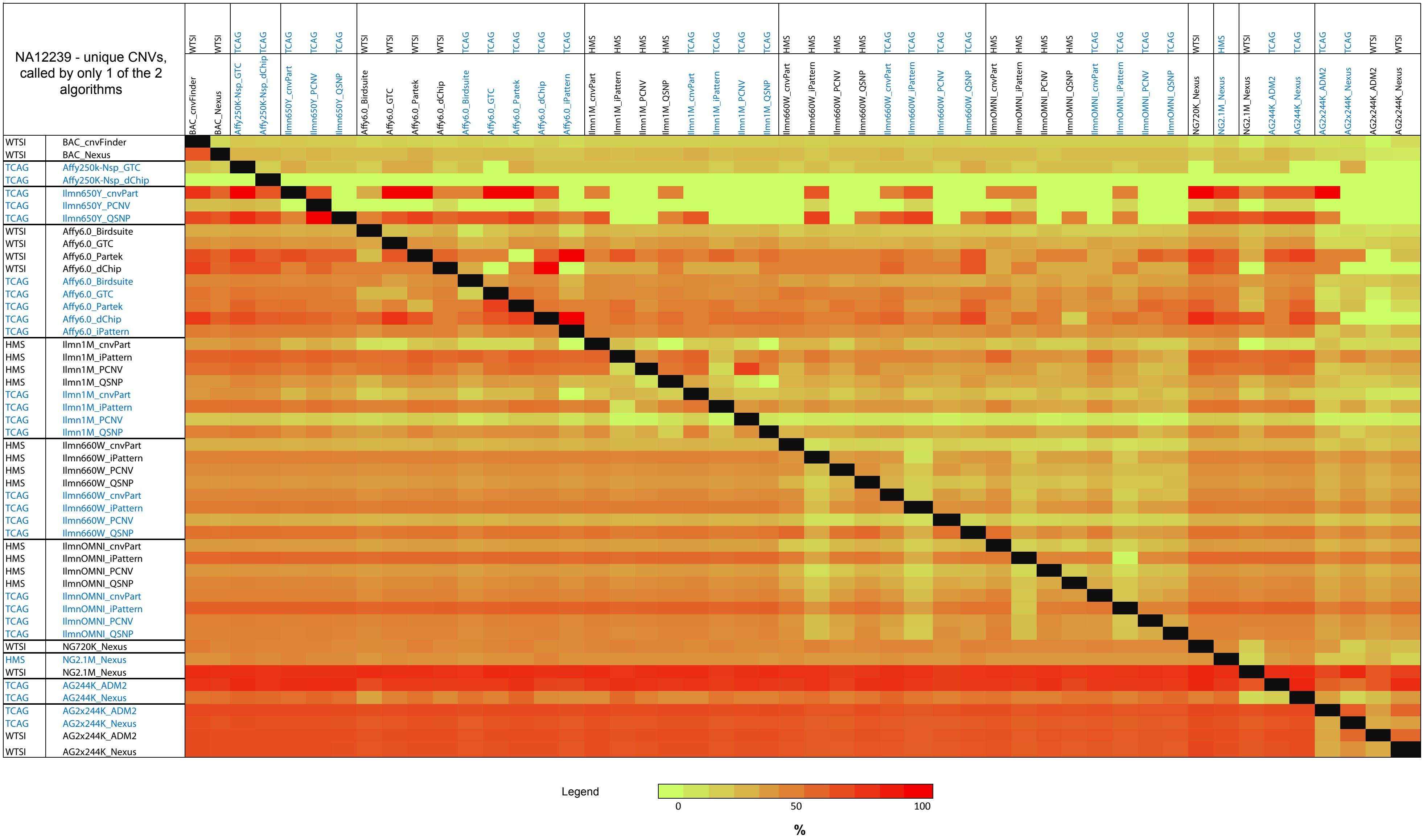
A.



B. Percentage of **validated** NA12239 CNV calls out of all shared calls between **two algorithms**. Calls were considered independently of their size or # of probes. Results are shown for any combination of algorithms/platforms/site. CNVs were considered validated when there was a reciprocal overlap of 50% or greater with the reference dataset of “validated NA12239-CNV calls in Conrad et al. 2010”.



C. Percentage of NA12239 CNV calls detected by **only one of 2 algorithms** tested and that are found to be **validated** when compared to a reference dataset. All NA12239 calls were considered independently of their size or # of probes. Results are shown for any combination of algorithms/platforms/site. CNVs were considered validated when there was a reciprocal overlap of 50% or greater with the reference dataset of “validated NA12239-CNV calls in Conrad et al. 2010”.


NA12239 - unique CNVs, called by only 1 of the 2 algorithms
WTSI BAC_cnvFinder
WTSI BAC_Nexus
TCAG Affy250k-Nsp_GTC
TCAG Affy250K-Nsp_dChip
TCAG Illmn650Y_cnvPart
TCAG Illmn650Y_PCNV
TCAG Illmn650Y_QSNP
WTSI Affy6.0_Birdsuite
WTSI Affy6.0_GTC
WTSI Affy6.0_Partek
WTSI Affy6.0_dChip
TCAG Affy6.0_Birdsuite
TCAG Affy6.0_GTC
TCAG Affy6.0_Partek
TCAG Affy6.0_dChip
TCAG Affy6.0_iPattern
HMS Illmn1M_cnvPart
HMS Illmn1M_iPattern
HMS Illmn1M_PCNV
HMS Illmn1M_QSNP
TCAG Illmn1M_cnvPart
TCAG Illmn1M_iPattern
TCAG Illmn1M_PCNV
TCAG Illmn1M_QSNP
HMS Illmn660W_cnvPart
HMS Illmn660W_iPattern
HMS Illmn660W_PCNV
HMS Illmn660W_QSNP
TCAG Illmn660W_cnvPart
TCAG Illmn660W_iPattern
TCAG Illmn660W_PCNV
TCAG Illmn660W_QSNP
HMS IllmnOMNI_cnvPart
HMS IllmnOMNI_iPattern
HMS IllmnOMNI_PCNV
HMS IllmnOMNI_QSNP
TCAG IllmnOMNI_cnvPart
TCAG IllmnOMNI_iPattern
TCAG IllmnOMNI_PCNV
TCAG IllmnOMNI_QSNP
WTSI NG720K_Nexus
HMS NG2.1M_Nexus
WTSI NG2.1M_Nexus
TCAG AG244K_ADM2
TCAG AG244K_Nexus
TCAG AG2x244K_ADM2
TCAG AG2x244K_Nexus
WTSI AG2x244K_ADM2
WTSI AG2x244K_Nexus
WTSI BAC_cnvFinder
WTSI BAC_Nexus
TCAG Affy250K-Nsp_GTC
TCAG Affy250K-Nsp_dChip
TCAG Illmn650Y_cnvPart
TCAG Illmn650Y_PCNV
TCAG Illmn650Y_QSNP
WTSI Affy6.0_Birdsuite
WTSI Affy6.0_GTC
WTSI Affy6.0_Partek
WTSI Affy6.0_dChip
TCAG Affy6.0_Birdsuite
TCAG Affy6.0_GTC
TCAG Affy6.0_Partek
TCAG Affy6.0_dChip
TCAG Affy6.0_iPattern
HMS Illmn1M_cnvPart
HMS Illmn1M_iPattern
HMS Illmn1M_PCNV
HMS Illmn1M_QSNP
TCAG Illmn1M_cnvPart
TCAG Illmn1M_iPattern
TCAG Illmn1M_PCNV
TCAG Illmn1M_QSNP
HMS Illmn660W_cnvPart
HMS Illmn660W_iPattern
HMS Illmn660W_PCNV
HMS Illmn660W_QSNP
TCAG Illmn660W_cnvPart
TCAG Illmn660W_iPattern
TCAG Illmn660W_PCNV
TCAG Illmn660W_QSNP
HMS IllmnOMNI_cnvPart
HMS IllmnOMNI_iPattern
HMS IllmnOMNI_PCNV
HMS IllmnOMNI_QSNP
TCAG IllmnOMNI_cnvPart
TCAG IllmnOMNI_iPattern
TCAG IllmnOMNI_PCNV
TCAG IllmnOMNI_QSNP
WTSI NG720K_Nexus
HMS NG2.1M_Nexus
WTSI NG2.1M_Nexus
TCAG AG244K_ADM2
TCAG AG244K_Nexus
TCAG AG2x244K_ADM2
TCAG AG2x244K_Nexus
WTSI AG2x244K_ADM2
WTSI AG2x244K_Nexus
Legend