Published in final edited form as:

Cancer Genet. 2015 ; 208(0): 408–413. doi:10.1016/j.cancergen.2015.04.010.

Concordance of copy number alterations using a common analytic pipeline for genome-wide analysis of Illumina and Affymetrix genotyping data: A report from the Children's Oncology Group

Marijana Vujkovic^a, Edward F. Attiyeh^a, Rhonda E. Ries^b, Michelle Horn^a, Elizabeth K. Goodman^a, Yang Ding^a, Marko Kavcic^c, Todd A. Alonzo^d, Robert B. Gerbing^e, Betsy Hirsch^f, Susana Raimondi^a, Alan S. Gamis^g, Soheil Meshinchi^{b,*,**}, and Richard Aplenc^{a,*,**}

^aDivision of Oncology, The Children's Hospital of Philadelphia, Philadelphia, PA

^bClinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA

^cUniversity Medical Centre, Ljubljana, Slovenia

^dKeck School of Medicine, University of Southern California, Los Angeles, CA

^eChildren's Oncology Group, Monrovia, CA

^fUniversity of Minnesota, Minneapolis, MN

^gDivision of Hematology/Oncology/Bone Marrow Transplantation, Children's Mercy Hospitals & Clinics, Kansas City, MO

Abstract

Copy number alterations (CNAs) are a hallmark of pediatric cancer genomes. An increasing number of research groups use multiple platforms and software packages to detect and analyze CNAs. However, different platforms have experimental and analysis-specific biases that may yield different results. We sought to estimate the concordance of CNAs in children with *de novo* acute myeloid leukemia between two experimental platforms: Affymetrix SNP 6.0 array and Illumina OmniQuad 2.5 BeadChip. Forty-five paired tumor-remission samples were genotyped on the two platforms and CNAs were estimated from total signal intensity and allelic contrast values using the allele-specific copy number analysis of tumors (ASCAT) algorithm. Both platforms were comparable in detection of CNAs, missing only 2 segments from a total of 42 CNAs (4.6%). Overall, there was an inter-platform agreement of 96% for allele-specific tumor profiles. However, poor quality samples with low signal-to-noise ratios showed a high rate of false positive segments independent of the genotyping platform. These results demonstrate that a common analytic pipeline can be utilized for SNP array data from these two platforms. The customized

*Corresponding Author: vujkovicm@email.chop.edu.

**Both authors contributed equally.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

programming template for the preprocessing, data integration, and analysis is publicly available at <https://github.com/AplenCHOP/affyLumCNA>.

Keywords

Pediatrics; acute myeloid leukemia; SNP array; copy number alterations

Introduction

Changes in DNA copy number occur frequently in tumor genomes (1, 2, 3) and can be measured with SNP array technologies, such as Illumina and Affymetrix. These platforms differ in experimental methodology, with the Affymetrix SNP array using oligonucleotide hybridization on fixed glass slides, while Illumina technology employs *in situ* single nucleotide extension reactions on bead arrays (4). The comparability of output from these two arrays has been reviewed regarding data quality, CNA calling, breakpoint accuracy, reproducibility, and concordance (4, 5). These reviews reported an inter-platform concordance of less than 50%. Data analysis is additionally complicated by technology-specific output formats (6). Gunnarsson *et al* described that tools for detecting CNAs from Illumina and Affymetrix raw data from identical samples typically do not yield identical CNA calls and insufficiently discriminate CNA artifacts from true CNA calls (7). Furthermore, CNA detection tools for Illumina and Affymetrix often yielded different CNA calls with Affymetrix detecting a higher degree of CNAs, while Illumina detecting higher rates of loss-of-heterozygosity (LOH) (4, 5). However, as SNP analytic packages improve, data variability may be overcome using appropriate data pre-processing (6). Currently numerous tools exist for array analysis of cancer samples on multiple platforms (e.g. ASCAT, OncoSNP, GPHMM, GAP, and PennCNV-Tumor) (8, 9, 10). We aimed to integrate and analyze CNA data from the Affymetrix and Illumina platforms and test the concordance between these two platforms in a set of matched tumor-remission DNA samples from pediatric AML patients genotyped on the Affymetrix SNP 6.0 array and Illumina OmniQuad 2.5 BeadChip.

Materials and Methods

A total of 90 matched tumor-remission samples from 45 children with newly diagnosed AML were obtained from the COG trials AAML0531 (11), AAML03P1 (12), and CCG-2961 (13). Cytogenetic abnormalities were determined by standard chromosome-banding analysis. Paired tumor-remission samples were genotyped on the Affymetrix SNP 6.0 chip at the Hudson Alpha Institute, Birmingham, AL, and on the Illumina 2.5M Omni Quad chip at the Children's Hospital of Philadelphia, Philadelphia, PA. The Affymetrix 6.0 SNP array contains 946K polymorphic and non-polymorphic copy number probes with a 2.18 KB median inter-marking spacing. The Illumina 2.5M BeadChip contains 2,390K SNPs with a median 0.64 KB probe spacing. The BeadChip does not contain any non-polymorphic CNV loci by design. To calculate Log R Ratio (LRR) and B Allele Frequency (BAF) values, raw Affymetrix CEL files were converted to allele-specific signals and genotype calls using Affymetrix Power Tools and BirdSeed algorithm. All markers were

updated to hg19, GRCh37 using LiftOver. Total signal intensity (LRR) was obtained by summing allele-specific intensities and was normalized with HapMap3 data. BAF, or allelic contrast, was calculated as the normalized ratio of the quantity of the variant allele to the total quantity of both alleles. For the Illumina samples, BAFs and LRRs were exported directly from Illumina GenomeStudio and signal intensity values corrected for patterns of genomic wave using PennCNV (14).

A matched (tumor vs. remission) allele-specific copy number analysis of tumors ASCAT was performed using ASCAT 2.2 in R (15). ASCAT was run with default parameters. All copy number profiles were manually reviewed (BG or MV). For the manual review, the *in silico* identified CNA segments by ASCAT algorithm were reviewed by examining the LRR and BAF visualizations. Illumina GenomeStudio GenomeViewer and Partek Genomics Suite for Affymetrix data were used as necessary. In a subsequent manual review, one study member (MV) was blinded to the ASCAT copy number output and visually identified CNAs from LRR and BAF plots. False negatives were manually annotated and the final list of manually reviewed CNA calls was defined as the gold standard. Platform-specific sensitivity and specificity were estimated by comparing the number of CNAs from the gold standard to the ASCAT output generated for Illumina and Affymetrix data. All altered segments were classified as CNA (+) and normal diploid segments were classified as CNA (-). Sensitivity was the number of *in silico* identified CNAs divided by the total number of true CNA as identified by visual inspection. Specificity was defined as the number of samples with no computationally identified CNAs divided by the total number of samples where no CNA was visually detected. The inter-observer Cohen's kappa statistic was used to test the platform agreement in CNA calls.

Results

Forty-three of 45 patients were successfully genotyped on both platforms. Two samples failed quality control and were excluded from further analysis. Table 1 summarizes the observed CNAs. Forty two CNAs were observed consisting of 18 losses, 6 regions of copy-neutral LOH, and 18 gains. In three patients with a normal karyotype, the SNP array detected an additional region of CN-LOH. Eight patients had gain of chromosome 8, observed with both cytogenetics and genotyping, except for one patient whose cytogenetics was unknown. In four patients the +8 was the sole abnormality identified by Gbanding, but the SNP array detected an additional CNA in two: a deletion of 7q and 9q. Other cytogenetically-identified chromosomal trisomies and monosomies were detected with SNP arrays as well. One CN-LOH on chromosome 13 was visually identified based on BAF values, but not detected with Affymetrix nor Illumina array because of low allelic contrast. Cytogenetics detected nine balanced inversions or translocations, which, as expected, were not detected by SNP arrays. However in two of these cases, both involving inversion 16 (*CBFB-MYH11* fusion), the SNP array detected two cryptic abnormalities: an associated deletion of the MYH11 gene in one and a secondary deletion of 7q36.1-7q36.2 in the other. In total, the SNP array identified four chromosomal losses, not visible by G-banding, and seven regions of CN-LOH that were not detectable by cytogenetics. In contrast, cytogenetics identified one gained region, i.e. add(2)(q35) in patient 11 that was not detected by array. The arrays further characterized some of the cytogenetically visible abnormalities. For

example, in patient 14, the array showed that the additional material translocated to 7q31 (the add 7q31) was an extra copy of the distal long arm of chromosome 1. Of the remaining 23 patients without observable CNA's, 7 had a normal karyotype and others had either inversions and/or translocations. Two patients had an unknown karyotype with no insertions or deletions.

Using the manually reviewed CNAs as the gold standard, ASCAT correctly called the majority of true CNAs independent of the platform. The smallest CNA detected was a 358 KB loss on chromosome 10 and was detected on both platforms. Overall, Illumina had a sensitivity of 95.3% (95% CI: 82.9 – 99.2%) with ASCAT and a specificity of 33.6% (95% CI: 31.9 – 35.3%). Affymetrix had a similar sensitivity 95.3% (95% CI: 82.9 – 99.2%) and specificity of 33.2% (95% CI: 31.6 – 34.9%). The inter-platform agreement was high with a Kappa statistic of 96%. Figure 1 shows representative examples of CNA calling on both platforms.

Discussion

These results demonstrate that platform-independent copy number analysis in pediatric AML can be robustly performed with a customized analytic pipeline using ASCAT. In addition, our data demonstrate that Affymetrix and Illumina SNP arrays have high concordance in calling CNAs and that more than 95% of the CNAs occurring in the AML tumor genome could be identified by either technology. This pipeline involves several key steps, namely the conversion of raw image output files to LRR and BAF files, followed by genomic wave adjustment and a matched allele-specific copy number analysis of tumors. Our customized programming template is publicly accessible at <https://github.com/AplenCHOP/affyLumCNA>. Although standard cytogenetics provide essential diagnostic and prognostic information in pediatric AML, small alterations are incidentally missed due to limited karyotype resolution and failure (16). In our study two focal gains, five focal losses, and all regions of CN-LOH were missed by cytogenetics. However, genotyping arrays failed to detect balanced rearrangements that were observed by cytogenetics and a few unbalanced rearrangements identified by G-banding (8). Thus, the combination of cytogenetics and array provided the most complete characterization of the abnormal clones.

The current literature lacks formally evaluated procedures for combining genotype data from different platforms. One abstract reported the simultaneous implementation of the ASCAT algorithm on Affymetrix and Illumina SNP array cancer samples (17). The authors state that the ASCAT algorithm adjusted for stromal cell contamination which resulted in more accurate calls in the CNAs than without ASCAT processing. Another study reports ASCAT to be the most appropriate software for analyzing Illumina SNP array data (18). However, a complete analytic pipeline for data integration and analysis was absent in both reports. Furthermore, most studies process data output in platform-specific pipelines and merged the CNA segments *post hoc* (5). This approach may introduce attenuation in CNA estimates due to differences in array coverage and density, sets of reference markers, and algorithms used for downstream CNA calling. BioDiscovery's Nexus Copy Number software is a commercial alternative for platform-independent analysis and it can integrate and process data from various array platforms including Affymetrix and Illumina. Such commercially-

licensed software have the advantages of extensive documentation and product support. However, such commercial packages may offer less flexibility in analyses.

We selected ASCAT because it provides platform independent CNV analysis and accounts for tumor ploidy and stromal contamination. Other programs that detect genomic aberrations do not typically address all issues. Unlike many other segmentation algorithms, ASCAT does not use a fixed probe-number for partitioning individual probes into segments. The algorithm consists of an allele-specific piecewise constant fitting regression model, initially starting with 40 predefined major segments followed by a search for change points that optimize the model to the observed copy number and allele ratio. The high accuracy and concordance of breakpoint estimation observed in our study may also be attributable to increased probe density and improved individual probe performance. A previous study showed that current genotyping arrays outperform older versions in terms of inferring calls and reproducibility (5).

We consider ASCAT complementary to manual CNA inspection. Clinically, ASCAT use may be useful for visualisation in contaminated samples as ASCAT eliminates noise introduced from the stromal tissue sample. Other than the visually observable CN-LOH on chromosome 11 in patient #3 there were no aberrations missed in the current cohort of AML patients by *in silico* detection.

In summary, these data demonstrate that it is possible to directly integrate datasets from Illumina and Affymetrix SNP arrays and perform a uniform CNA analysis. Despite the differences between the two technologies, there is a high concordance in CNA segment calls between the platforms, and no platform-specific biases were observed. Manual review remains important in the validation of *in silico* identified CNAs, particularly in samples with poor quality DNA. Although this analytic pipeline was developed for research use, the analysis process could be applied to clinical sample analysis after appropriate validation studies were performed. (17). This may facilitate harmonizing data analysis across data sets and platforms. However, use of this algorithm in clinical settings would require further validation studies.

Acknowledgments

Funding:

R01 CA133881

U10 CA098543

U10 CA98413-08

References

1. Ciriello G, Miller ML, Arman Aksoy B, et al. Emerging landscape of oncogenic signatures across human cancers. *Nat Gen.* 2013; 45:1127–1133.
2. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. *Nat Rev Cancer.* 2004; 4:177–183. [PubMed: 14993899]
3. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. *Nature.* 2010; 464:704–12. [PubMed: 19812545]

4. Curtis C, Lynch AG, Dunning MJ, et al. The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics*. 2009; 10:588. [PubMed: 19995423]
5. Pinto D, Darvishi K, Shi X, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol*. 2011; 29:512–20. [PubMed: 21552272]
6. Scherer SW, Lee C, Birney E, et al. Challenges and standards in integrating surveys of structural variation. *Nat Gen*. 2007; 39:S7–S15.
7. Gunnarsson R, Staaf J, Jansson M, et al. Screening for copy-number alterations and loss of heterozygosity in chronic lymphocytic leukemia -- a comparative study of four differently designed, high resolution microarray platforms. *Genes Chrom Cancer*. 2008; 47:697–711. [PubMed: 18484635]
8. Mosén-Ansorena D, Aransay AM, Rodríguez-Ezpeleta N. Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. *BMC Bioinf*. 2012; 13:192.
9. Mosén-Ansorena D, Aransay AM. Bivariate segmentation of SNP-array data for allele-specific copy number analysis in tumour samples. *BMC Bioinf*. 2013; 14:84.
10. Oros KK, Arcand SL, Bayani J, et al. Analysis of genomic abnormalities in tumors: a review of available methods for Illumina two-color SNP genotyping and evaluation of performance. *Cancer Genet*. 2013; 206:103–15. [PubMed: 23623180]
11. Sung L, Aplenc R, Alonzo TA, et al. Effectiveness of supportive care measures to reduce infections in pediatric AML: a report from the Children's Oncology Group. *Blood*. 2013; 121:3573–7. [PubMed: 23471307]
12. Cooper TM, Franklin J, Gerbing RB, et al. AAML03P1, a pilot study of the safety of gemtuzumab ozogamicin in combination with chemotherapy for newly diagnosed childhood acute myeloid leukemia: A report from the Children's Oncology Group. *Cancer*. 2011; 118:761–9. [PubMed: 21766293]
13. Lange BJ, Smith FO, Feusner J, et al. Outcomes in CCG-2961, a Children's Oncology Group phase 3 trial for untreated pediatric acute myeloid leukemia: a report from Children's Oncology Group. *Blood*. 2008; 111:1044–53. [PubMed: 18000167]
14. Diskin SJ, Li M, Hou C, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res*. 2008; 36:e126. [PubMed: 18784189]
15. Van Loo P, Nordgard SH, Lingjærde OC, et al. Allele-specific copy number analysis of tumors. *PNAS*. 2010; 107:16910–16915. [PubMed: 20837533]
16. Maciejewski JP, Tiu RV, O'Keefe C. Application of array-based whole genome scanning technologies as a cytogenetic tool in haematological malignancies. *Br J Haematol*. 2009; 146:479–88. [PubMed: 19563474]
17. Keshavan R, Verma S, Shams S. Dealing with sample aneuploidy and mosaicism using the ASCAT algorithm on different SNP array platforms. *Cancer Genet*. 2011; 204:472.
18. Dougherty MJ, Tooke LS, Sullivan LM, et al. Clinical utilization of high-resolution single nucleotide polymorphism based oligonucleotide arrays in diagnostic studies of pediatric patients with solid tumors. *Cancer Genet*. 2012; 205:42–54. [PubMed: 22429597]

Figure 1.

Examples of CNAs in 3 Illumina and Affymetrix samples visualized with ASCAT. The ASCAT profile contains the allele-specific copy-number of all assayed loci, with BAF or LRR value shown on the Y-axis and genomic loci on the x-axis. BAF plots only display germline heterozygous calls.

Abbreviations: CNA, copy number alteration; BAF, b-allele frequency, LRR, log R ratio; ASCAT, allele-specific copy number analysis of tumors.

Cytogenetic and Copy Number Alterations of 43 successfully genotyped patients on Illumina and Affymetrix.

Table 1

Pt	Sex	Chr	Start (Mb)	End (Mb)	CNA Type	Affymetrix	Illumina	Karyotype
1	M	19	4,616,7901	59,097,854	CN-LOH	Yes	Yes	46,XY[55]
2	F	13	20,425,911	115,103,529	CN-LOH	No	Yes	46,XX[20]
3	F	11	198,572	49,580,690	CN-LOH	No	No	46,XX[30]
4	F	8	46,246	146,301,427	Gain	Yes	Yes	47,XX,+8[12]
		9	65,629,772	106,056,348	Loss	Yes	Yes	
5	M	8	46,246	146,301,427	Gain	Yes	Yes	47,XY,+8[12]
		13	20,223,976	115,103,529	CN-LOH	Yes	Yes	
6	F	8	46,246	146,301,427	Gain	Yes	Yes	47,XX,+8[20]
7	F	7	147,223,640	158,933,239	Loss	Yes	Yes	47,XX,+8[19]/46,XX[1]
		8	46,246	146,301,427	Gain	Yes	Yes	
8	F	8	46,246	146,301,427	Gain	Yes	Yes	46,X,-X,+8,t(8;21)(q22;q22)[20]
9	F	5	15,532	171,051,005	Gain	Yes	Yes	48,XX,t(5;9;11;9)(q11.2;p22;q23;b22),+der(5)t(5;9;11;9),+8[20]
		8	46,246	146,301,427	Gain	Yes	Yes	
		9	107,457,275	141,091,394	Gain	Yes	Yes	
10	M	2	224,877,511	243,068,869	CN-LOH	Yes	Yes	47,XY,+8,t(10;11)(p12;q23)[10]
		8	46,246	146,301,427	Gain	Yes	Yes	
11	M	8	46,246	146,301,427	Gain	Yes	Yes	46,XY,add(2)(q35)[5]/46,s,del(9)(q12;q32)[3]/47,sdl1,+8[12]
		9	65,616,971	106,074,356	Loss	Yes	Yes	
12	M	8	46,246	146,301,427	Gain	Yes	Yes	Unknown
13	F	7	126,392,026	160,927,726	Loss	Yes	Yes	46,XX,t(8;21)(q22;q22)[2]/46,idem,del(7)(q32)[23]
14	M	1	177,751,810	249,224,388	Gain	Yes	Yes	46,XY,t(8;21)(q22;q22)[10]/46,idem,add(7)(q31)[10]
		7	110,110,566	160,927,726	Loss	Yes	Yes	

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Pt	Sex	Chr	Start (Mb)	End (Mb)	CNA Type	Affymetrix	Illumina	Karyotype
15	M	2	203,171,672	243,068,869	Loss	Yes	Yes	45,X,-Y,t(8;21)(q22;q22),add(2)(q34)[20]
		15	45,493,321	102,520,751	Gain	Yes	Yes	
16	M	16	15,818,141	16,362,362	Loss	Yes	No	46,XY,inv(16)(p13.1q22)[20]
17	F	7	151,906,053	154,818,571	Loss	Yes	Yes	46,XX,inv(16)(p13.1q22)[20]
18	F	9	46,587	36,246,949	CN-LOH	Yes	Yes	46,XX,t(10;11)(p11.2;q23)[20]
19	M	1	61,735	16,970,178	Loss	Yes	Yes	46,XY,t(9;11)(p22;q23)[8]/48,idem,+2mar[12]
		17	58,296,494	81,048,659	Gain	Yes	Yes	
20	M	1	230,029,167	234,797,002	Loss	Yes	Yes	48,XY,add(1)(q32),+4,t(10;11)(p13;q21),-11,+12,der(12)t(11;12)(q13;p11.2)x2,13,+2mar[4]/49,idem,+6[14]
		4	12,281	191,027,923	Gain	Yes	Yes	
		6	149,661	171,051,005	Gain	Yes	Yes	
		10	23,046,141	23,404,754	Loss	Yes	Yes	
		11	62,890,494	66,288,076	Loss	Yes	Yes	
		11	73,386,388	134,944,770	Gain	Yes	Yes	
		12	9,009,042	15,437,997	Loss	Yes	Yes	
		12	15,438,032	133,778,189	Gain	Yes	Yes	
		13	39,837,566	58,045,789	Loss	Yes	Yes	
		13	60,856,044	61,310,470	Loss	Yes	Yes	
		13	61,768,626	62,992,697	Loss	Yes	Yes	
		13	68,663,212	71,926,813	Loss	Yes	Yes	
		14	22,288,841	22,853,518	Loss	Yes	Yes	

Abbreviations: CNA, copy number alteration; CN-LOH, Copy-Neutral Loss-Of-Heterozygosity