Scharpf, Robert B.; Ruczinski, Ingo; Carvalho, Benilton; Doan, Betty; Chakravarti, Aravinda

A multilevel model to address batch effects in copy number estimation using SNP arrays. (English) Zbl 1437.62601

Biostatistics 12, No. 1, 33-50 (2011).

Summary: Submicroscopic changes in chromosomal DNA copy number dosage are common and have been implicated in many heritable diseases and cancers. Recent high-throughput technologies have a resolution that permits the detection of segmental changes in DNA copy number that span thousands of base pairs in the genome. Genomewide association studies (GWAS) may simultaneously screen for copy number phenotype and single nucleotide polymorphism (SNP) phenotype associations as part of the analytic strategy. However, genomewide array analyses are particularly susceptible to batch effects as the logistics of preparing DNA and processing thousands of arrays often involves multiple laboratories and technicians, or changes over calendar time to the reagents and laboratory equipment. Failure to adjust for batch effects can lead to incorrect inference and requires inefficient post hoc quality control procedures to exclude regions that are associated with batch. Our work extends previous model-based approaches for copy number estimation by explicitly modeling batch and using shrinkage to improve locus-specific estimates of copy number uncertainty. Key features of this approach include the use of biallelic genotype calls from experimental data to estimate batch-specific and locus-specific parameters of background and signal without the requirement of training data. We illustrate these ideas using a study of bipolar disease and a study of chromosome 21 trisomy. The former has batch effects that dominate much of the observed variation in the quantile-normalized intensities, while the latter illustrates the robustness of our approach to a data set in which approximately 27% of the samples have altered copy number. Locus-specific estimates of copy number can be plotted on the copy number scale to investigate mosaicism and guide the choice of appropriate downstream approaches for smoothing the copy number as a function of physical position. The software is open source and implemented in the R package crlmm at Bioconductor (http://www.bioconductor.org).

MSC:

62P10 Applications of statistics to biology and medical sciences; meta analysis

Software:

CARAT ; crlmm; PLASQ; R

Full Text: DOI

References:

[1] “Mapping autism risk loci using genetic linkage and chromosomal rearrangements.”, 39, 319-328 (2007)

[2] Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma, 104, 20007-20012 (2007)

[3] Increased MET gene copy number negatively affects survival of surgically resected non-small-cell lung cancer patients, 27, 1667-1674 (2009)

[4] “Quantifying uncertainty in genotype calls.”, 26, 242-249 (2010)

[5] Determination of cancer risk associated with germ line BRCA1 missense variants by functional analysis, 67, 1494-1501 (2007)

[6] Copy Number Variation Analysis with SVS 7 (2009)

[7] Mechanisms for human genomic rearrangements, 1, 4 (2008)

[8] CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays, 7, 83 (2006)

[9] Mapping and sequencing of structural variation from eight human genomes, 453, 56-64 (2008)

[10] Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs, 40, 1253-1260 (2008)

[11] PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data, 8, 323-336 (2007) · Zbl 1144.62098

[12] Replicated microarray data, 12, 31 (2001) · Zbl 1004.62086

[13] Genomic disorders ten years on, 1, 42 (2009)

[14] MMP13, Birc2 (cIAP1), and Birc3 (cIAP2), amplified on chromosome 9, collaborate with p53 deficiency in mouse osteosarcoma progression, 69, 2559-2567 (2009)

[15] Structural variation of chromosomes in autism spectrum disorder, 82, 477-488 (2008)

[16] Integrated detection and population-genetic analysis of SNPs and copy number variation, 40, 1166-1174 (2008)

[17] Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis, 67, 409-413 (2008)

[18] A genome-wide investigation of SNPs and CNVs in schizophrenia, 5 (2009)

[19] Global variation in copy number in the human genome, 444, 444-454 (2006)

[20] Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays, 2, 687-713 (2008) · Zbl 1400.62285

[21] Gene copy number variation in schizophrenia, 96, 93-99 (2007)

[22] Estimating genome-wide copy number using allele specific mixture models, 15, 857-866 (2008)

[23] Identification of potential driver genes in human liver carcinoma by genomewide screening, 69, 4059-4066 (2009)

[24] A statistical framework for the analysis of microarray probe-level data, 1, 333-357 (2007) · Zbl 1126.62111

[25] Singleton deletions throughout the genome increase risk of bipolar disorder, 14, 376-380 (2008)

This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.