Title: A computational approach to identify whole genome homozygosity mapping across multiple SNP mapping experiments

Corresponding author: Roberta Spinelli, Institute of Biomedical Technologies, Segrate, Milan (ITB-CNR), *roberta.spinelli@itb.cnr.it*

Other authors:

A. Gessi[2], M.C. Proverbio[2], E. Mangano[2], F. Ferrari[3], I. Cifola[1], M. Bardini[4], G. Cazzaniga[4], A. Salvatoni[5], C. Battaglia[2]

1 Institute of Biomedical Technologies, Segrate, Milan (ITB-CNR)

2 Department of Science and Biomedical Technologies (DiSTeB) and PhD School of molecular medicine, University of Milan;

3 Department of Biology, University of Padua, via U.Bassi 58/B, Padova;

4 Centro Ricerca Tettamanti, Clinica Pediatrica Univ. Milano-Bicocca, Monza, Italy;

5 Department of Clinical and Biological Science (DSCB) , Pediatric Clinic, University of Insubria, Varese

Abstract:

The recent development of microarray platforms, capable to genotype more than thousands of single nucleotide polymorphisms (SNPs) in individuals, had provided an opportunity to rapidly identify susceptibility loci for complex phenotypes. High density SNP mapping arrays have been widely applied to association studies, to copy number (CN) analysis in cancers and recently to investigate the role of homozygosity extended regions in individuals. Long stretches of CN neutral and homozygous SNPs, defined as runs of homozygosity (ROHs) can be found either in a single individual or shared across samples. The identification of ROHs among affected individuals of the same family or among unrelated ones with same disease, can underline loci potentially implicated in the genetic basis of the disease under study. Therefore the identification of ROHs in affected individuals or pathological datasets gives a chance to identify disease associated loci and new causative mutations. In order to identify ROHs pattern across Affymetrix SNP mapping datasets, we developed a computational strategy including several computational steps: 1) loss of heterozygosity analysis by dChip2007 software; 2) a within-subject step allowing the identification of ROHs in a single sample; 3) an across-subject step extracting the ROH fingerprint of the dataset and 4) the identification of a common ROHs pattern based on frequency across the dataset under study, varying the number of individuals carrying common ROHs; 5) the annotation step allowing the association of genes to selected ROHs. In order to obtain an effective ROHs visualization, we use dChip software for the entire samples dataset. We assess our strategy to two SNP mapping datasets including 100K leukemia and 250K congenital recessive diseases. The procedure all owed the identification of a unique genetic ROH fingerprint of clinical datasets potentially important to discover new diseases associated loci suitable for further investigations.