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

Consiglio Nazionale delle Ricerche
ITB - Istituto Tecnologie Biomediche

Roberta Spinelli[1], Alessandra Gessi[2], Maria Carla Proverbio[2], Eleonora Mangano[2], Francesco Ferrari[3], Ingrid Cifola[1], Michela Bardini[4], Gianni Cazzaniga[4], Alessandro Salvatoni[5], Cristina 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, 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

INTRODUCTION

The recent development of microarray platforms, capable to genotype thousands of single nucleotide polymorphisms (SNPs) in individuals, has 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 and to investigate the role of homozygosity extended regions in individuals(1). Long stretches of CN neutral and homozygous SNPs, defined as runs of homozygosity (ROH) can be found either in a single individual or shared across samples (2). The identification of ROH 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 ROH in affected individuals or pathological datasets gives a chance to identify disease associated loci and new causative mutations. In order to identify ROH pattern across Affymetrix SNP mapping datasets, we developed a computational strategy including several computational steps using dChip2007 software, the R language and UCSC Genome Browser. We applied our strategy to two SNP mapping datasets including 100K SNP Mapping leukemia patients and 250K SNP Mapping congenital recessive diseases patients (CRD) for a total of 166 individuals. The procedure allowed the identification of unique clinical ROH patterns and revealed genomic region potentially important to discover new diseases associated genes suitable for further investigations.

METHODS

The computational strategy entails several computational steps including: SNP preprocessing analysis, within-subject and across-subjects ROH identification analysis, ROH genes annotation. At each step, the R language was used to code the procedure (http://www.r-project.org/) and the genomic visualization of resulting ROH was carried out using dChip (http://biosun1.harvard.edu/complab/dchip/).

STEP1: SNP preprocessing analysis

a. Loss of Heterozygosity (LOH) and Copy Number (CN) analyses (*Figure 1*) were performed to determine the CN and LOH profiling for each sample by dChip2007 software. We performed unpaired LOH analysis using HMM considering haplotype correction (HC/LD-HMM) by comparing our patients dataset to CEPH 60 HapMap parents as reference samples, as suggested by the dChip manual (http://biosun1.harvard.edu/complab/dchip/manual.htm). The LOH analysis assigned to each SNP an inferred SNP LOH probability defined as P(LOH) and ranging from 0 to 1. CN analysis was performed using median smoothing and trimmed analysis, according to the software. In parallel, all samples were also analyzed with CNAG2.0 (www.genome.umin.jp) by using normal references used for dChip analysis. A matrix of SNP LOH probability data was exported from dChip2007 structured with samples in columns and SNP in rows (e.g. ***Table 1***).

b. Annotation of SNP LOH probability data matrix. Inferred SNP LOH probability data matrix were combined with additional genomic information (e.g. physical position, chromosome, cytoband, allele and genotype frequencies) included in mapping array specific annotation file. Then, annotated data matrix was sorted according the SNP physical position along the genome and used to ROH identification analysis at single sample or across samples levels.

STEP2: Within-subject ROHs analysis (single sample data)

We used single sample data and extracted stretches of consecutive events of LOH within-subject setting the P(LOH)>=0.5 defined as runs of homozygosity (ROH). The procedure produced a table listing the ROH of each patient (sample). Each ROH was described by start and end genomic SNP position (bp) and cytoband, length (bp), number of SNPs, the maximum and minimum value of P(LOH) associated to SNP belonging to ROH (***Table 2***).

STEP3: Across-subjects ROHs analysis (multiple samples data)

a. Fingerprint Analysis. We identified the ROH pattern across the entire dataset by setting the P(LOH)>=0.5, defined as "fingerprint analysis". The fingerprint was determined by the presence of same ROH among at least two samples in dataset. The fingerprint ROH data were described by start and end genomic SNP position (bp) and cytoband, length of ROH, number of SNPs in the region, the maximum and minimum number of patients (frequency) sharing the same ROH (***Table 3 and Figure 2***).

b. Shared ROHs Analysis. Fixing the number of individuals carrying same ROH namely "shared ROH", we then extracted common regions of ROH across all the subjects under study. Data tables containing a list of shared ROH can be calculated using 2, 3 or more samples. Usually, this step reduces the length of shared ROH (see yellow lines in ***Table 3 and 4***) and permits an enhancement of genomic regions clinically important for the disease under study (***Figure 3***).

STEP4: Annotation of genes to ROHs

The annotation step allowed the association of genes to selected ROH. The list of genes associated to ROH was obtained using the UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables), by querying hg17 database and three tables such as KnowGenes, RefGene and RefFlat. Genes list was used to investigate for new diseases candidate genes.

RESULTS

The entire homozygosity mapping procedure was applied to determine the ROHs pattern of 46 leukemia (L) patients profiled by 100K SNP Mapping arrays and of 35 congenital recessive diseases (CRD) patients analyzed by 250K SNP mapping arrays.

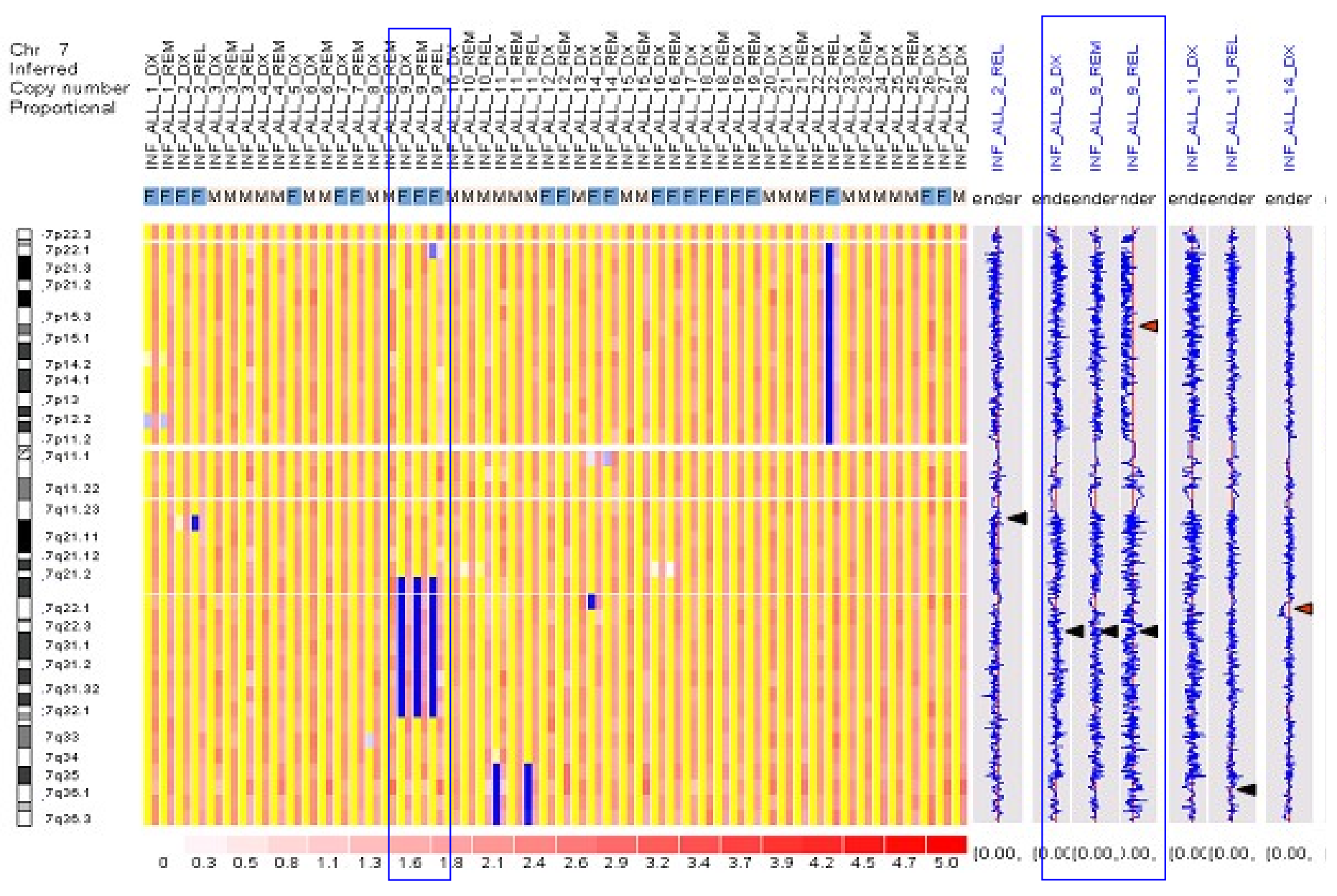

***Figure 1*. Example of LOH and CN profiling in a representative chromosome in leukemia 100K patients.** The inferred LOH probability is displayed from yellow (0) to white (0.5) to blue (1). Copy number is displayed in an increasing color scale from withe (0 copies) to red (5 copies). The blue line on the right showed the smoothed CN profile according to diploid red line for each sample.

To note in the figure, a large region of ROH on arm 7q in Pt.9 at diagnosis, remission and relapse characterized by genomic tract of homozygous genotypes and copy number neutral profiling (blue box and black arrows); a small deletions is visible in Pt.14 (red arrow).

Data type: Inferred LOH call													
Marker	dbSNP	Chromosome	Position	Genetic Dist(cM)	Score	S1	S2	S3	S4	S5	S6	S7
SNP_A-1972092	rs1909520	3	25038705		0.03	0	1	0	0	0	0	0
SNP_A-4209367	rs9310769	3	25067771		0.04	0	1	0	0.14	0	0.13	0
SNP_A-1895961	rs4858694	3	25072740		0.04	0	1	0	0.15	0	0.15	0
SNP_A-2032668	---	3	25083807		0.04	0	1	0	0.17	0	0.17	0
SNP_A-2300582	rs4858697	3	25091586		0.05	0	1	0	0.19	0	0.18	0
SNP_A-1952768	rs6775433	3	25091862		0.05	0	1	0	0.19	0	0.18	0
SNP_A-2097783	rs4858698	3	25093457		0.05	0	1	0	0.19	0	0.19	0
SNP_A-2031508	rs7430038	3	25099388		0.05	0	1	0	0.19	0	0.19	0
SNP_A-1843366	---	3	25130875		0.04	0	1	0	0.21	0	0.21	0
SNP_A-4227209	rs11129180	3	25131812		0.04	0	1	0	0.21	0	0.21	0
SNP_A-1822258	rs17576085	3	25139629		0.04	0	1	0	0.21	0	0.21	0
SNP_A-2193566	rs17516853	3	25139810		0.04	0	1	0	0.21	0	0.21	0
SNP_A-2148250	---	3	25139888		0.04	0	1	0	0.21	0	0.21	0
SNP_A-1972098	rs17517019	3	25146657		0.04	0	1	0	0.21	0	0.21	0
SNP_A-2080992	rs11129184	3	25147604		0.04	0	1	0	0.21	0	0.21	0
SNP_A-1886628	rs11129187	3	25147790		0.04	0	1	0	0.21	0	0.21	0
SNP_A-2289494	rs6804502	3	25148353		0.04	0	1	0	0.21	0	0.21	0
SNP_A-4240965	rs7616818	3	25152717		0.04	0	1	0	0.21	0	0.21	0
SNP_A-4222227	rs9812604	3	25157766		0.04	0	1	0	0.21	0	0.21	0
SNP_A-4225484	rs12495790	3	25157785		0.04	0	1	0	0.21	0	0.21	0
SNP_A-1828349	rs13060347	3	25169305		0.04	0	1	0	0.21	0	0.21	0
SNP_A-2082200	rs2363595	3	25169973		0.04	0	1	0	0.21	0	0.21	0

***Table 1*. Example of inferred LOH probability obtained by dChip2007**. SNP annotations, score and sample SNP LOH probability data are listed for each patients (S). The sample LOH probability was calculated according to dChip manual and LOH score measuring the prevalence of LOH at a marker across the samples, and is computed as the average probability of LOH.

Chr	Start Cluster (bp)	End Cluster (bp)	Start Cytoband	End Cytoband	#SNPs	max P(LOH)	min P(LOH)	Length Cluster (bp)	ID
3	66819911	67105010	p14.1	p14.1	50	0,55	0,55	285100	S1
3	113946481	114468156	q13.2	q13.2	50	0,89	0,85	521676	S1
4	20661407	20944165	p15.31	p15.31	50	0,72	0,72	282759	S1
4	169735288	170004183	q32.3	q32.3	50	0,74	0,73	268896	S1
6	122405198	122845811	q22.31	q22.31	50	0,58	0,58	440614	S1
4	7705495	8407068	p16.1	p16.1	50	0,59	0,59	701574	S2
5	36641766	37541744	p13.2	p13.2	50	0,96	0,96	899979	S2
1	4952639	5278962	p36.32	p36.32	50	0,98	0,98	326324	S3
2	28322606	29084934	p23.2	p23.2	50	0,9	0,9	762329	S3
2	106948310	107263107	q12.3	q12.3	50	0,65	0,64	314798	S3
2	185273737	187123635	q32.1	q32.1	150	1	0,95	1849899	S3

***Table 2*. Single sample ROHs data.** ROH events are extracted for each sample. ROHs are described by chromosome, physical position (start-end), cytoband (start-end), number of SNPs markers, the maximum and minimum value of SNP LOH probability within the region, length (bp) and sample label (S).

In the case of 35 CRD patients, we found a total of 765 events of ROH that were further evaluated by CNAG software. Each ROH was associated to annotated genes and bioinformatics analysis has been carried out.

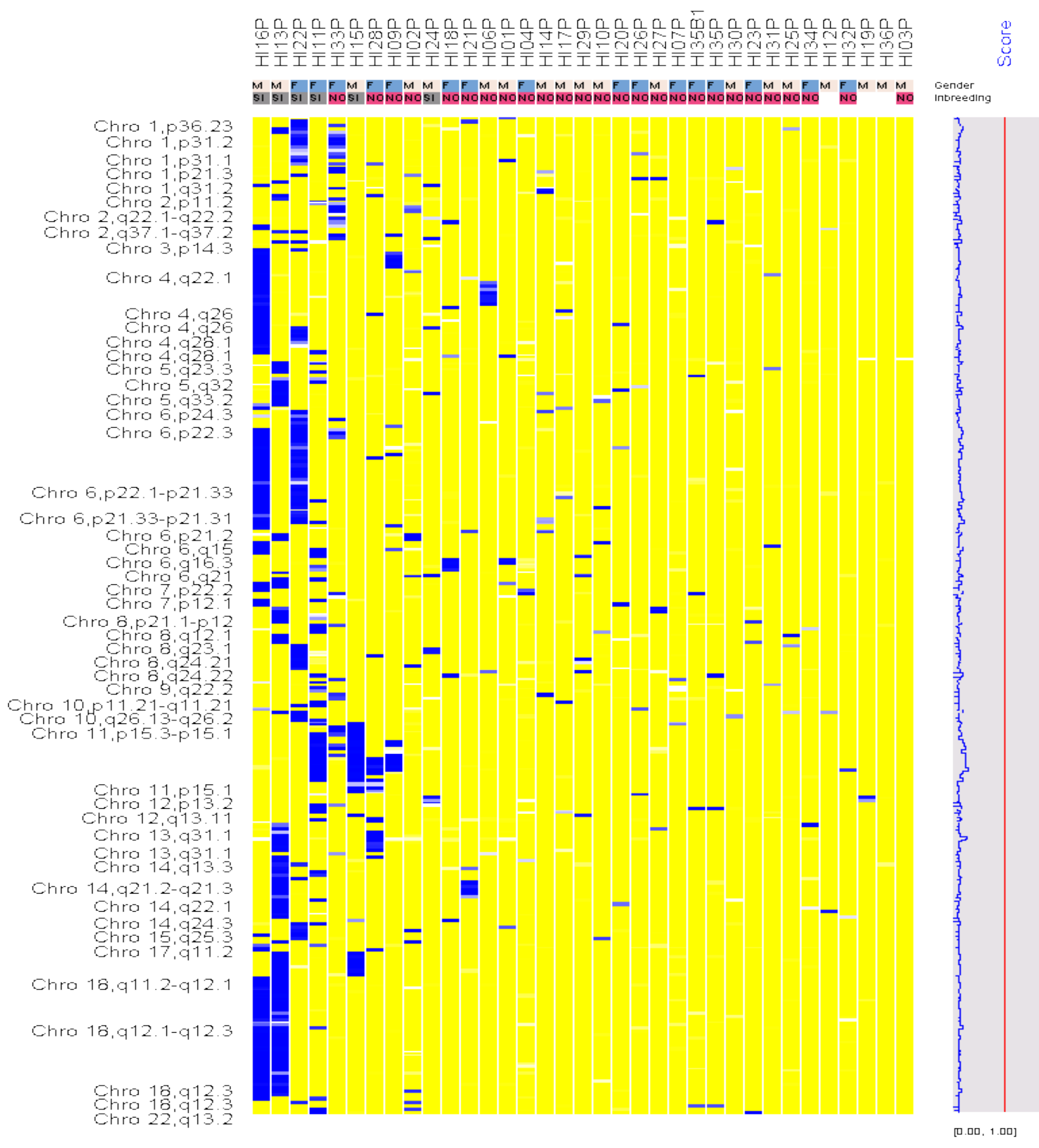

***Figure 2*. Fingerprint dataset view**. The plot shows the fingerprint of 35 CRD patients composed by 164 ROH clusters (blue boxes). Rows represent chromosome ROH regions and columns represent samples.

We noticed that the number and the length of ROH events were mainly associated to patients derived from consanguineous families (inbreeding).

Cluster#	Chr	Start Cluster (bp)	End Cluster (bp)	Start Cytoband	End Cytoband	#SNPs	freq max	freq min	Length Cluster (bp)
Cluster1	1	7767568	8034406	p36.23	p36.23	29	2	2	266839
Cluster2	1	54671315	55176764	p32.3	p32.3	50	2	2	505450
Cluster3	1	64101570	64493815	p31.3	p31.3	50	2	2	392246
Cluster4	1	65856897	66497951	p31.2	p31.2	100	3	3	641055
Cluster5	1	66795692	67566623	p31.2	p31.2	100	2	2	770932
Cluster6	1	68028825	68528821	p31.2	p31.2	50	2	2	499997
Cluster7	1	71714715	72767289	p31.1	p31.1	100	2	2	1052575
Cluster8	**1**	**73701875**	**74847962**	**p31.1**	**p31.1**	**100**	**3**	**2**	**1146088**
Cluster9	1	75905253	76192135	p31.1	p31.1	50	2	2	286883
Cluster10	1	77554882	78408842	p31.1	p31.1	50	3	3	853961
Cluster11	1	96195413	96568739	p21.3	p21.3	50	2	2	373327

Table 3*. Fingerprint data.** From annotated *Table 1* in the case of 35 CRD patients we extract a total of 164 ROH. The ROH regions were visualized by dChip giving the fingerprint of CRD patients (Figure 2***).

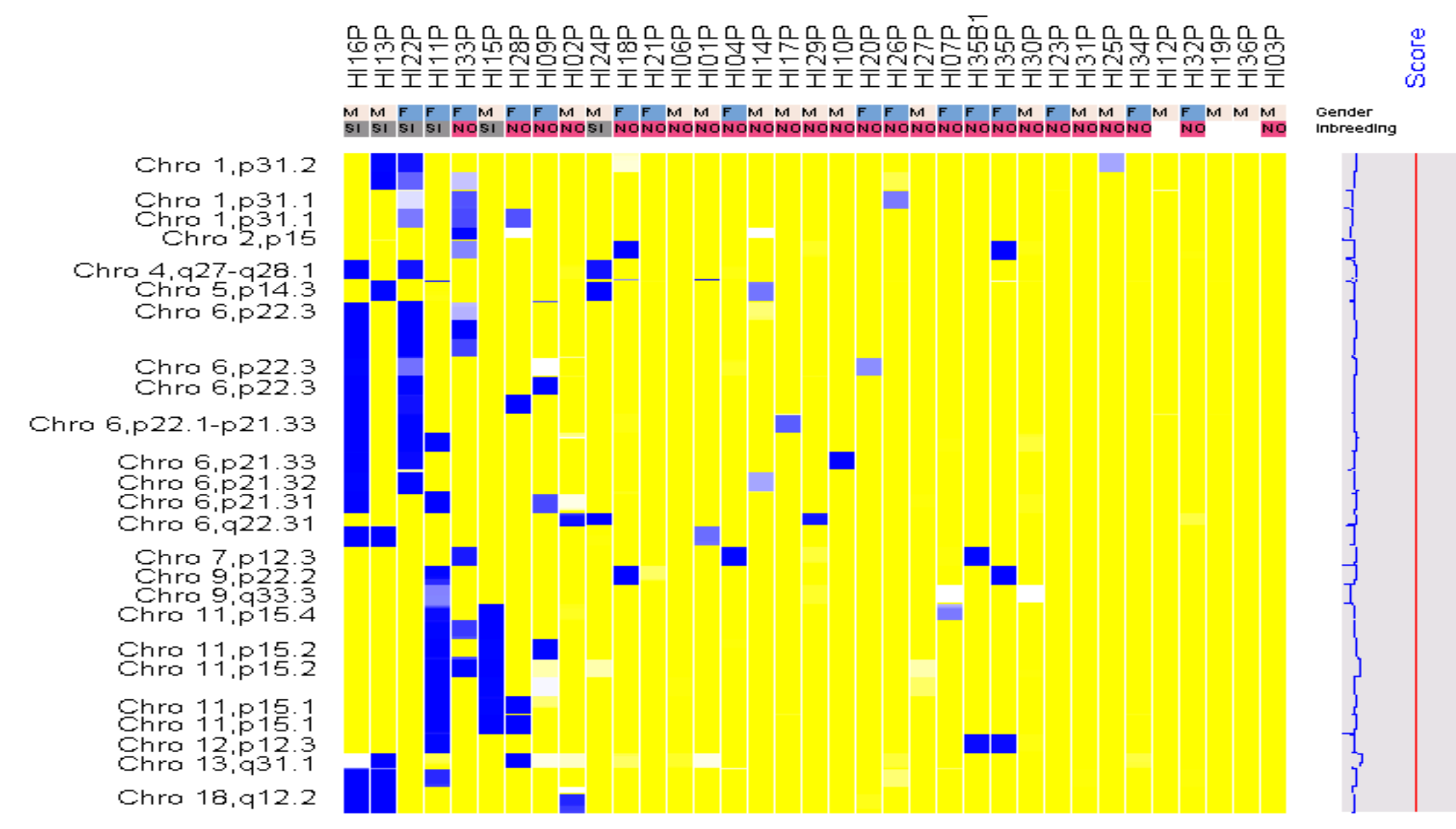

***Figure 3*. Three patients shared ROHs view**. The plot shows 35 shared ROH pattern (blue boxes) obtained setting the number of CRD patients equal to 3.

Cluster#	Chr	Start Cluster (bp)	End Cluster (bp)	Start Cytoband	End Cytoband	#SNPs	freq max	freq min	Length Cluster (bp)
Cluster1	1	65856897	66497951	p31.2	p31.2	100	3	3	641055
Cluster2	1	73701875	74190336	p31.1	p31.1	50	3	3	488462
Cluster3	1	77554882	78408842	p31.1	p31.1	50	3	3	853961
Cluster1	2	61288114	61898960	p15	p15	33	3	3	610847
Cluster2	2	185273737	186042074	q32.1	q32.1	50	3	3	768338
Cluster1	4	123882940	124448184	q27	q28.1	50	3	3	565245
Cluster1	5	19269375	19269562	p14.3	p14.3	2	3	3	188
Cluster2	5	128226162	128228401	q23.3	q23.3	3	3	3	2240
Cluster3	5	152939962	153345621	q33.2	q33.2	50	3	3	405660
Cluster1	6	20121370	21504377	p22.3	p22.3	150	3	3	1383008
Cluster2	6	22516941	22722709	p22.3	p22.3	50	3	3	205769

Table 4*. Three patients shared ROH data.** From *Table 1* we identified ROH based on fixed frequency across the dataset. Using three patients as fixed frequency, we found a total of 35 clusters. The ROH regions were visualized by dChip giving the view of common ROH pattern of CRD patients (Figure 3***).

For the CRD dataset we found several patterns of shared ROH. In particular considering 2 CRD patients we identified 169 ROH clusters containing 829 annotated genes, using 3 samples we found 35 clusters associated to 139 genes and increasing the number of patients to 4 we found 4 ROH regions containing 42 genes.

CONCLUSIONS

We developed a computational strategy allowing the identification of ROH in a single sample and ROH patterns of two clinical datasets (leukemia and the congenital recessive diseases). We found genomic regions containing interesting genes potentially implicated in our diseases and suitable for further investigations. The procedure could be applied to any genome wide SNP genotyping data matrix and can be extended to large number of individuals. We observed that the presence of many ROH events within subject is often associated to inbreeding and could be represent a genetic risk factor (2).

References:
1. Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Simon-Sanchez J. et al. Human Molecular Genetics 2007 16(1):1-14; 2. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Lencz T. et al. PNAS 2007 104(50):19942-19947.

ACKNOWLEDGMENTS:This work was supported by grants from the Italian Ministry of University and Research (MIUR-FIRB n°RBLA03ER38 and PRIN 2006).

Contacts: Roberta Spinelli (roberta.spinelli@itb.cnr.it)