ELSEVIER

Cancer Genetics ■ (2014) ■

Cancer Genetics

BRIEF COMMUNICATION

Combinations of genetic data in a study of neuroblastoma risk genotypes

Mario Capasso [a,b], Francesco Maria Calabrese [b], Achille Iolascon [a,b], Erling Mellerup [c,*]

[a] *Department of Molecular Medicine and Medical Biotechnologies, University of Napoli Federico II, Naples, Italy;* [b] *CEINGE (Centro Ingegneria Genetica) Advanced Biotechnologies, Naples, Italy;* [c] *Laboratory of Neuropsychiatry, Department of Neuroscience and Pharmacology, Faculty of Health, University of Copenhagen, Copenhagen, Denmark*

Analysis of combinations of genetic changes that occur exclusively in patients may be a supplementary strategy to the single-locus strategy used in many genetic studies. The genotypes of 16 SNPs within susceptibility loci for neuroblastoma (NB) were analyzed in a previous study. In the present study, combinations of these genotypes have been analyzed. The theoretical number of combinations of 3 SNP genotypes taken from 16 SNPs is 15,120. Of these, 14,307 were found in 370 patients and 803 controls; 12,772 combinations were common to both patients and controls; 1,213 were found in controls only; and 322 combinations were found in patients only. Among the latter, a cluster of 24 combinations was found to be significantly associated with NB ($P < 0.00001$).

Keywords Neuroblastoma, combinations, SNPs, data mining

The majority of heritable genetic risk factors identified by genome-wide association studies for most common diseases remain elusive (1), indicating a need for supplementary methods for elucidating the genetics of complex diseases. A specific combination of genetic changes is the genetic basis for a polygenic disease, and the combination will be present in all patients. If the disease shows genetic heterogeneity, several combinations of genetic changes can be the basis for the disease. These combinations can be found in subgroups of patients but never in control subjects who are genetically unrelated to these patients. In order to identify genetic changes related to diseases, most genetic studies use a single-locus strategy to look for changes that are significantly different with respect to their distribution in patients and controls; however, even the changes with the lowest *P*-values in statistical analyses are always found in both patients and controls. This single-locus approach can be supplemented by methods that look at combinations of genetic changes found in patients but never in controls. In a study of bipolar disorder, combinations of 3 SNP genotypes taken from 803 SNPs resulted in almost 2 billion combinations—60 million of which were found in patients only, and 4 clusters of these patient-specific combinations were significantly associated with the disorder (2).

In a previous study of neuroblastoma (NB), 16 SNPs within the genes *LINC00340*, *BARD1*, *LMO1*, *DUSP12*, *HSD17B12*, *DDX4*, *IL31RA*, *HACE1*, and *LIN28B* were analyzed in 370 cases and 809 controls (3). The aim of the present study was to analyze combinations of these SNP genotypes.

Materials and methods

SNP genotype data

Computational analysis was performed on data pertaining to 16 SNPs associated with NB risk. The SNPs were genotyped in an Italian population of 370 patients and 809 controls as described previously (3). Briefly, the case subjects were defined as children with a diagnosis of NB or ganglioneuroblastoma; their data were collected by the Italian Neuroblastoma Group. All control subjects were recruited from Italian blood donor centers. Eligibility criteria for control subjects were as follows: Italian origin, availability of DNA, and no serious underlying medical disorder, including

M.C., F.M.C., and A.I.:declare no conflict of interest. E.M. owns stock in the company Genokey.

Received October 31, 2013; received in revised form February 9, 2014; accepted February 10, 2014.

* Corresponding author.
E-mail address: mellerup@hotmail.com

2210-7762/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cancergen.2014.02.004

cancer. The Italian DNA samples were genotyped using the SNP Genotyping Assay on a 7900HT Real-Time PCR System (Applied Biosystems, Monza, Italy) and validated by Sanger sequencing (3730 DNA Analyzer, Applied Biosystems).

Combinations and statistics

Combinations of SNP genotypes were analyzed by the bioinformatics company, GenoKey (Hoersholm, Denmark), by means of computer programs using array-based logic (4–6), which enabled the counting of combinations that occurred in the subjects. No statistical analyses were performed for combinations common to both controls and patients or for combinations found in controls only. Combinations that occurred exclusively in patients were screened for combinations and clusters of combinations that were significantly associated with NB. Screening was performed with permutation tests (100,000 permutations per test) followed by Bonferroni correction of the *P*-values.

Results

The theoretical number of combinations of 2 SNP genotypes taken from 16 SNPs is $16!/(16-2)!2! \times 3^2 = 1{,}080$. The subjects contained 1,074 of these combinations: 1,052 were common to both patients and controls; 16 were found in controls only; and 6 combinations occurred in patients only. The theoretical number of combinations of 3 SNP genotypes taken from 16 SNPs is $16!/(16-3)!3! \times 3^3 = 15{,}120$. Of these, 14,307 were found in the subjects: 12,772 combinations were common to both patients and controls; 1,213 were found in controls only; and 322 combinations were found in patients only. The theoretical number of combinations of 4 SNP genotypes is $16!/(16-4)!4! \times 3^4 = 147{,}420$. A total of 119,489 of these combinations were found: 87,284 were common to both patients and controls; 25,154 were found in controls only, and 7,051 combinations occurred in patients only. Permutation tests showed that the 6, the 322, and the 7,051 patient-specific combinations might be random findings; however, among the 322 patient-specific combinations of 3 SNP genotypes, 24 combinations, which all contained the rs6435862_*BARD1* genotype 2, formed a cluster with 32 patients. This cluster was significantly associated with NB ($P < 0.00001$). Another cluster with 20 patients and 25 combinations, which all contained the rs110419_*LMO1* genotype 2, was also significantly associated with NB ($P = 0.032$).

Among the 7,051 patient-specific combinations of 4 SNP genotypes, 20 combinations, which all contained the rs6435862_*BARD1* genotype 2, formed a cluster with 22 patients. This cluster was significantly associated with NB ($P = 0.013$). Another cluster with 19 patients and 24 combinations, which all contained the rs3768716_*BARD1* genotype 2, was also significantly associated with NB ($P = 0.028$).

After Bonferroni correction for multiple statistical tests, only the first of the four clusters remained significantly associated with NB. Some patients were represented in this cluster by more than one combination, thereby contributing to the cluster with more than three genotypes. Table 1 shows the SNP genotypes in each of the patients in this cluster.

Of the 370, 158 were high-risk NB cases (3), whereas of the 32 patients in the cluster, 20 were high-risk cases. Thus, high-risk NB cases were enriched in the cluster (chi-square test, $P = 0.0403$).

Analyses of combinations containing five, six, seven, or eight genotypes did not result in clusters of combinations significantly associated with NB.

The cluster in Table 1 contained 32 patients and 24 combinations of 3 SNP genotypes. Due to marked overlap between genotypes in the 24 combinations, the cluster contained only 21 genotypes, which are shown in the first row of the table. The suffix indicates the genotype (0 homozygous for the major allele, 1 heterozygous, and 2 homozygous for the minor allele). The first column contains dummy numbers for the individual patients in the cluster; italics indicate high-risk patients. The last column lists the number of SNP genotypes for each patient in the cluster. The last row lists the number of each SNP genotype in the cluster.

Discussion

Analyses of combinations of genetic data related to diseases are uncommon, probably because of the computational challenges created by the large number of possible combinations. This challenge has resulted in a search for methods of calculating disease-related combinations of SNPs (7–9), as well as services from bioinformatics companies. Handling of a large number of combinations may be facilitated when only combinations that occur exclusively in patients are studied, because the majority of combinations are common to both patients and controls. Furthermore, most of the combinations found in patients may be random findings, leaving few, if any, combinations significantly associated with the disease. Thus, in a study of bipolar disorder, combinations of 3 SNP genotypes taken from 803 SNPs resulted in almost 2 billion combinations; however, less than 60 million of these were found exclusively in the patients, and only about 160 of these were significantly associated with the disorder (2).

In the study by Capasso and coworkers (3), the most significant SNP genotype was the rs6435862_*BARD1* genotype 2 ($P = 8.4 \times 10^{-15}$), which was found in 69 of 344 patients and in 49 of 784 controls. This result highlights the importance of this genotype as a contributor to NB risk, but it also indicates that the genotype alone is not a sufficient genetic basis for NB. An analysis of combinations independently identified rs6435862_*BARD1* genotype 2 as the defining genotype for a cluster of combinations significantly associated with NB. Although each of the 24 combinations of 3 SNP genotypes in the cluster was found in patients only, and never in controls, it was the cluster as such that reached statistical significance. This statistical significance is supported clinically by the finding of relatively more high-risk NB cases in the cluster than in the whole group.

Several patients in the cluster contributed more than one combination, allowing us to identify combinations of up to 11 SNP genotypes in a patient. Very few patients in the cluster had the same pattern of SNP genotypes (see Table 1),

Table 1 SNP genotypes for each patient in the cluster

Patientno.	rs1048108_*BARD1*$_0$	rs6435862_*BARD1*$_2$	rs3768716_*BARD1*$_0$	rs3768716_*BARD1*$_2$	rs4758051_*LMO1*$_1$	rs4758051_*LMO1*$_2$	rs2229571_*BARD1*$_0$	rs2070094_*BARD1*$_1$	rs110419_*LMO1*$_0$	rs110419_*LMO1*$_2$	rs1027702_*DUSP12*$_0$	rs1027702_*DUSP12*$_1$	rs4336470_*HACE1*$_1$	rs4336470_*HACE1*$_2$	rs11037575_*HSD17B12*$_1$	rs7585356_*BARD1*$_1$	rs6939340_*LINC00340*$_1$	rs6939340_*LINC00340*$_2$	rs4712653_*LINC00340*$_1$	rs4712653_*LINC00340*$_2$	rs17065417_*LIN28B*$_1$	Total SNP genotypes
1553	x	x		x			x	x		x		x	x		x			x		x		11
1485	x	x		x			x	x				x	x		x		x			x		10
1545	x	x						x	x		x		x		x	x	x		x			10
1504	x	x		x			x	x			x		x				x		x			8
1549		x	x			x				x				x				x		x	x	8
1058		x	x							x								x		x	x	6
2394		x				x				x	x			x							x	6
2449		x							x						x	x		x	x			6
1193	x	x						x				x	x									5
2188		x								x						x		x		x		5
2388		x			x					x				x							x	5
2467		x		x			x	x				x										5
2497		x													x	x		x	x			5
2823		x							x							x	x			x		5
1086		x				x					x			x								4
1362	x	x						x			x											4
1493		x			x					x				x								4
1503		x	x		x					x												4
1525		x								x								x		x		4
2266		x								x								x		x		4
2392		x			x					x				x								4
2484		x													x	x	x					4
2503		x													x	x	x					4
1120		x				x								x								3
1214		x			x					x												3
1419		x													x	x						3
1460		x																x	x			3
1480		x																x	x			3
2212		x															x			x		3
2586		x														x				x		3
2825		x									x			x								3
2859		x									x			x								3
Totals	6	32	3	4	5	4	4	7	3	12	7	4	4	9	8	9	7	10	6	10	4	

indicating genetic heterogeneity; however, the various combinations were very similar, with at least one common genotype, and the others were taken from a narrow set of SNP genotypes, suggesting that accumulation of key genotypes from a pool of genotypes may be a strong risk factor. In a previous study of bipolar disorder, four clusters of combinations were identified; in this previous study, the patients also had personal patterns of SNP genotypes, but again within a narrow set of genotypes (2).

The collection of combinations in Table 1 is significantly associated with NB, indicating that these combinations may be portions of the complete set of genetic changes that are the genetic basis for NB. The single SNP genotypes in the cluster were all found in control subjects, but the combinations of these genotypes shown in Table 1 were not found in any of the control subjects. Once most of the predisposing variants to NB and other complex diseases are identified by ongoing genome-wide association studies, combinatorial methods might be useful in elucidating the genetic picture of the disease.

Analysis of relatively large combinations of genetic data is still uncommon (2,10,11), but fast data mining methods make it possible to supplement many of the existing studies of single genetic changes with combinations of these changes. In addition to allowing researchers to extract more information from the studies, these methods could also identify individual patterns of genetic changes, which could be potentially interesting with regard to personalized medicine.

Because studies of combinations of genetic data are uncommon, the direct findings in the present study cannot be compared with similar studies. Therefore, these findings need to be replicated in an independent cohort of NB patients. In addition, the methods used to combine genetic data need to be tested in more data sets. The interpretation of the cluster as a pool of key SNP genotypes that to a varying degree can accumulate in the genome, and thereby increase the risk of disease, is a suggestion that needs supporting evidence from other studies of combinations of genetic data.

Acknowledgments

The study was funded by grants from the Associazione Italiana per la Ricerca sul Cancro (10537), MIUR- FIRB Ricerca in Futuro (RBFR08DWQ3 to M.C.); Fondazione Italiana per la Lotta al Neuroblastoma (to M.C. and F.M.C.); and Associazione Oncologia Pediatrica e Neuroblastoma (to M.C.).

References

1. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747–753.
2. Koefoed P, Andreassen OA, Bennike B, et al. Combinations of SNPs related to signal transduction in bipolar disorder. PLoS One 2011;6:e23812.
3. Capasso M, Diskin SJ, Totaro F, et al. Replication of GWAS-identified neuroblastoma risk loci strengthens the role of BARD1 and affirms the cumulative effect of genetic variations on disease susceptibility. Carcinogenesis 2013;34:605–611.
4. More T. Axioms and theorems for a theory of arrays. IBM J Res Develop 1973;17:135–175.
5. Franksen OI. Invariance under nesting - an aspect of array-based logic with relation to Grassmann and Peirce. Boston Stud Philos Sci 1994;187:303–335.
6. Møller GL. On the Technology of Array-based Logic. Ph.D. thesis. Technical University of Denmark 1995.
7. Lin HY, Chen YA, Tsai YY, et al. TRM: a powerful two-stage machine learning approach for identifying SNP-SNP interactions. Ann Hum Genet 2012;76:53–62.
8. Wu J, Devlin B, Ringquist S, et al. Screen and clean: a tool for identifying interactions in genome-wide associations studies. Genet Epidemiol 2010;34:275–285.
9. Zhang Y, Jiang B, Zhu J, et al. Bayesian models for detecting epistatic interactions from genetic data. Ann Hum Genet 2011; 75:183–193.
10. Xie Q, Ratnasinghe LD, Hong H, et al. Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer; a novel method. BMC Bioinformatics 2005; 6(Suppl 2):S4.
11. Mellerup E, Andreassen OA, Bennike B, et al. Connection between Genetic and Clinical Data in Bipolar Disorder. PLoS One 2012;7:e44623.