Gene 512 (2013) 161–165

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Short Communication

Investigation of secreted protein transcripts as early biomarkers for type 1 diabetes in the mouse model

Evie Melanitou [a,*], Fredj Tekaia [b], Edouard Yeramian [c]

[a] Institut Pasteur, Laboratoire Immunophysiologie et Parasitisme, 25, rue du Dr Roux, 75725 Paris, Cedex 15, France
[b] Institut Pasteur, Unité de Génétique Moléculaire des Levures (URA 2171 CNRS and UFR927 Univ. P.M. Curie), 25, rue du Dr Roux, 75725 Paris, Cedex 15, France
[c] Institut Pasteur, Unité de Bioinformatique Structurale, (URA 2185 CNRS), 25, rue du Dr Roux, 75725 Paris, Cedex 15, France

ARTICLE INFO

Article history:
Accepted 14 September 2012
Available online 29 September 2012

Keywords:
Diabetes
Secretome
Biomarkers
Autoimmunity
Correspondence analysis

ABSTRACT

Type 1 diabetes (T1D) represents a serious health burden in the world, complicated by the fact that disease onset can be preceded by a long time period without evident clinical signs. It would be then of critical importance to detect the disease in its early stages. In this direction, we seek here to identify early preinflammatory markers for autoimmune diabetes, mining our previously reported transcriptome data relevant to distinct early sub-phenotypes in the NOD mouse, associated with early insulin autoantibodies (E-IAA). More specifically we focus on secreted or transmembrane protein transcripts, identifying in this category 71 differentially expressed transcripts which are regulated at the early preinflammatory stages of T1D in the pancreatic lymph nodes (PLN). Following the expression patterns of these 71 transcripts, correspondence analysis (a multivariate analysis method) reveals a clear-cut segregation of the individual samples according to the early subphenotype used. Thus the 71 transcripts coding for secreted proteins constitute a candidate-set of predictive biomarkers for the development of autoimmune damage of the β cells of the pancreas. The majority of these genes have human orthologs and accordingly they represent potential candidate biomarkers for the human disease. In addition, for predictive purposes, the analysis reveals the possibility to reduce significantly the size of the candidate-set in practice, with various genes displaying identical expression profiles.

1. Introduction

Type 1 diabetes is characterized by a long asymptomatic period preceding the disease onset in humans and in the NOD mice. For diagnostic purposes, as well as on fundamental grounds, it is then highly desirable to grasp the important changes (environmental, cellular and gene expression) occurring in this critical period of time and leading ultimately to the disease onset. Such a global comprehensive scheme is still elusive at present, but it seems possible, with the increasingly available genomic data and information, to develop appropriate candidate biomarkers for risk assessment.

In this direction, autoantibodies are useful in detecting the disease prior to glycemia in the NOD mice and in humans. There are however some limitations for their use notably, the development of transient autoantibodies can be followed by their disappearance months or even years later without the development of diabetes (Abiru et al., 2001; Barker et al., 2004). There also exists the possibility of transplacental transfer of autoantibodies to infants of T1D mothers (Greeley et al., 2002). Despite the identification of several risk factors for type 1 diabetes, early prediction is still missing due to insufficient predictive power of the individual risk factors (Purohit and She, 2008). Accordingly we seek here an alternative solution for the development of candidate biomarkers for early stages of autoimmune diabetes, mining our previously published trancriptome data (Regnault et al., 2009). More specifically we focus on the class of transcripts coding for secreted proteins in the NOD mouse and we demonstrate indeed the potentiality to use this specific category of genes as biomarkers for autoimmune predisposition.

The rationale of the focus here on secreted protein transcripts is twofold: (1) On functional grounds, secreted proteins are known to play key roles in various important biological processes such as morphogenesis (Thomas et al., 2011; Tremble et al., 1993), angiogenesis (Onuffer and Horuk, 2002), cellular differentiation (Nalbant et al., 2005; Rosenow et al., 2010), apoptosis (Danielsen and Maihle, 2002) and modulation of immune response (Flavell, 2002; Grandvaux et al., 2002). These proteins are also implicated in disease processes, such as cancer progression (Welsh et al., 2003). Accordingly, this class of proteins has attracted much interest in recent years, with notably the initiation of a concerted effort called "The Secreted Protein Discovery Initiative" (SPDI), aiming to the identification of human secreted and

Abbreviations: T1D, Type 1 diabetes; NOD, Non obese diabetic; PLN, Pancreatic lymph node; CA, Correspondence analysis; SPG, Secreted protein genes; GEO, Gene expression omnibus.

* Corresponding author at: Immunophysiology and Parasitism laboratory, Department of Parasitology and Mycology, Institut Pasteur, 25-28 rue du Dr Roux, 75015 Paris, France. Tel.: +33 1 40 68 3299; fax: +33 1 45 68 8332.
E-mail addresses: evie.melanitou@pasteur.fr (E. Melanitou), fredj.tekaia@pasteur.fr (F. Tekaia), edouard.yeramian@pasteur.fr (E. Yeramian).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.09.055

transmembrane proteins with bioinformatic approaches (Clark et al., 2003). (2) A diagnostic potential exists when secreted proteins are found in the serum. Indeed such proteins may be targeted by specific antibodies or small molecules in body fluids and thus constitute potential biomarkers for disease detection.

In this general background, we develop here proof of concept for the possibility of early biomarkers for T1D, in the NOD mouse model, based on "secretome" data. More precisely, mining our previously published transcriptome data in this model (Regnault et al., 2009), we identify 71 genes coding for proteins located in the extracellular space. We then exploit the global properties of the expression signatures of these secreted protein genes (SPGs) with multivariate analysis, resorting to correspondence analysis (CA). Such analysis revealed a clear-cut segregation of the individual samples with the various subphenotypes. Examination of the genes involved in the analysis shows that the segregation is not following simple functional characteristics, further highlighting the interest in the global multivariate approach. In addition, such examination also reveals that a significant fraction of the secreted genes share identical expression profiles, allowing the perspective of a significant reduction of the candidate-set size in practice. Finally, with most of the genes in the set having human orthologs, it is expected that it will be possible to extend the developed system towards an efficient test for establishing the association of the presence of the corresponding proteins in the peripheral blood with the early detection of T1D in humans.

2. Materials and methods

2.1. Animals and analysis of microarray data

The settings for experimental data collection and microarray analysis have been described previously (Regnault et al., 2009). In brief, NOD/*Tac* mice were used and monitored for the presence of insulin autoantibodies by radioimmunoassay at one week intervals after weaning (starting at 3 weeks of age) (Melanitou et al., 2004; Yu et al., 2000). Animals were sacrificed at 5 weeks of age as previously described (Melanitou et al., 2004). Two groups of animals were selected (E-IAA$_{pos}$ and E-IAA$_{neg}$) and high quality T-RNAs were prepared (Qiagen) from PLN tissues. T-RNA samples (4.5 μg) were monitored for cDNA preparation and hybridized on the Affymetrix MG_U74A_version 2 arrays (Santa Clara, Ca) containing 12 486 probe sets. The resulting data (in Affymetrix CEL format) were subjected to RMA (Robust Multichip Average) normalization (Bolstad et al., 2003) and transcripts with "Absent" Detection Call in both conditions were removed from further analysis. The RMA-normalized data set was used to identify modulated gene expression by the LPE (Local Pooled Error) statistical test (Jain et al., 2003), following adjustment for multiple comparisons by the Hochberg and Benjamini test (Hochberg and Benjamini, 1990). Box plots metrics of the RMA-normalized data illustrated the robustness of the data for both sets of samples (see supplementary data in Regnault et al., 2009).

The data used for the correspondence analyses concerned nine individual mice (PLN samples) and were extracted from the initial data set, deposited in the NCBI's Gene Expression Omnibus (Edgar et al., 2002), GEO: GSE15582. The samples were in two groups, relative to early insulin autoantibodies (E-IAA) subphenotypes (Table S1): 5 negative samples (E-IAA$_{neg}$) and 4 positive ones (E-IAA$_{pos}$).

2.2. Multivariate analysis

The expression patterns of the transcripts for secreted-protein genes were analyzed with correspondence analysis (CA). For this analysis, the expression levels of the various genes were represented by their "expression flags" (A: absent; M: marginal and P: present), after corrections for background noise (as previously described (Regnault et al., 2009)) in the nine individual samples used for the arrays. As such, the correspondence analysis concerned 3×9 samples, with each mouse individual An (either E-IAA$_{neg}$ or E-IAA$_{pos}$; n indexing variable) associated with three specific expression profiles An.pf(A), An.pf(M) and An.pf(P), according to the expression flags (A, M or P) for the set of genes in the study. Accordingly, with such coding, we are led to a table (not shown) containing 27 samples (9×3, corresponding to the various An.pf(X)) *versus* 71 SPGs. The fold change values for the expression levels (between negative and positive subphenotypes) were not used in the CA analysis.

3. Results

3.1. Identification of secreted protein-genes (SPGs)

For type 1 diabetes disease pathogenesis in the NOD mouse, the early pre-inflammatory stages generally occur at the age of 4–6 weeks, after weaning and early priming of T cells reactive to islet β cell antigens takes place in the PLN (Turley et al., 2003). Accordingly, we performed the analyses in this tissue, in mice at 5 weeks of age. In such conditions, considering the E-IAA subphenotype, we identified by microarray analysis 165 transcripts regulated in the PLN of E-IAA$_{pos}$ mice (Melanitou et al., 2004; Regnault et al., 2009).

Out of the 165 differentially expressed genes identified in the E-IAA$_{pos}$ *vs* E-IAA$_{neg}$ samples, 71 genes (43%), code for secreted proteins, as assigned by Gene Ontology annotations (http://www.geneontology.org/) (Huang et al., 2007) for cellular localization ($p < 2.7e-28$). Such percentage corresponds to an over-representation of this class of genes, as compared to the overall percentage of genes coding for secreted proteins (a mere 16%) in the Affymetrix arrays (MGU74Av2).

Expression differences of the 71 transcripts in the E-IAA$_{pos}$ *vs* E-IAA$_{neg}$ samples varied between 90 and 1.5 fold, with the majority of the genes (66 genes) being up-regulated in the E-IAA$_{pos}$ PLN (Fig. 1). Only five transcripts were down-regulated, all of them coding for immunoglobulin-related proteins. With the exception of 3 ESTs, the remaining 68 genes code for known proteins and the majority of them have human orthologs (Table S1).

3.2. Correspondence analysis of expression patterns of SPGs

In order to gain a global picture of the -complex- dependencies between expression patterns of the SPGs and the E-IAA subphenotype we resorted to correspondence analysis (CA) a multivariate statistical method (Tekaia and Yeramian, 2006; Tekaia et al., 2002). Our aim was to assess the usage of secreted proteins as biomarkers for autoimmune predisposition in the NOD mouse.

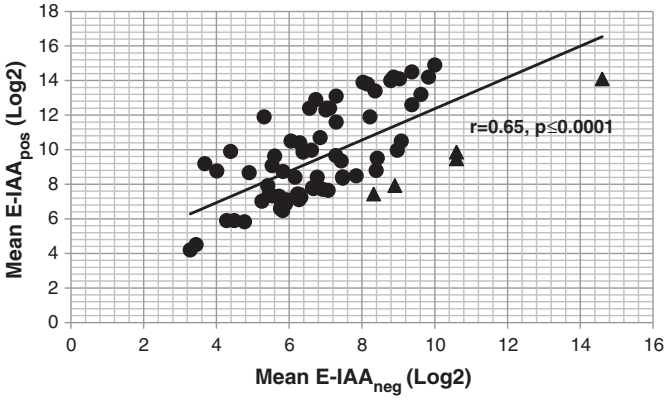

Fig. 1. Relative expression patterns of secreted protein genes (SPGs) in the PLN of E-IAA$_{pos}$ and E-IAA$_{neg}$. Linear regression line is drawn (correlation coefficient $r = 0.65$ with $p \leq 0.0001$). Down-regulated genes in the E-IAA$_{pos}$ samples are represented as triangles.