Integration of genetic and molecular data paves the way to precision medicine in multiple sclerosis

STEFFAN DANIEL BOS (MSC, PHD)[1, 2], ANTOINE LIZEE[3, 4], HANNE FLINSTAD HARBO[1, 2], PIERRE ANTOINE GOURRAUD[3, 4, 5]

[1]*Institute of Clinical Medicine, Nevrologisk avdeling DMII. Oslo (Norway).*
[2]*Department of Neurology. Oslo University Hospital. Oslo (Norway).*
[3]*Department of Neurology. School of Medicine University of California. San Francisco, CA (USA).*
[4]*UMR Inserm 1064. Nantes University. Nantes (France).*
[5]*Department of Public Health Nantes University Hospitals. Nantes (France).*

Correspondence: Institute of Clinical Medicine, Nevrologisk avdeling DMII, rom L-265
Gaustadalleen 340372 OSLO (Norway) – Tel: 4796016326

ABSTRACT. In less than a decade, genetic screens have revealed hundreds of loci associated to complex human diseases, including well over 100 of such loci for multiple sclerosis (MS). Most of these genetic associations provide very little or no direct insights into the functional mechanisms driving the associations. Therefore, focus of research is now shifting towards identification of a biological function for the genetic variants. In spite of the lack of knowledge on the direct role of genetic variants, the accumulated value of these MS loci provide some insights. Here, we discuss how such accumulated genetic information may be combined with additional data sources in order to characterize individual patients. With increasing understanding of the biological processes involved, such patient level data may provide additional input for clinical decision-making. The post GWAS era of MS Genetics is at the vanguard of precision medicine, if not for its clinical utility, it may be for the collaborative network data organization and heavy statistical modeling required by approaches that link robust population-based discovery and heterogeneous individual risk assessment.

Key words: multiple sclerosis, precision medicine, genetics, data integration, contextualization.

RESUMEN. En menos de una década, los barridos genéticos han revelado cientos de loci asociados con las enfermedades humanas complejas, incluyendo más de 100 loci en la esclerosis múltiple (EM). La mayoría de estas asociaciones genéticas dan poca o ninguna luz acerca de los mecanismos por los que ocurre la asociación. Sin embargo, el foco de la investigación se está dirigiendo ahora hacia la identificación de la función biológica de las variantes genéticas. A pesar de la falta de conocimiento acerca del papel directo de las variantes genéticas, el valor acumulado acerca de esos loci en la EM comienza a facilitar una cierta comprensión. En este trabajo, discutimos cómo la información genética acumulada puede ser combinada con fuentes adicionales de datos con el fin de caracterizar a los pacientes individuales. Con el aumento de la comprensión de los procesos biológicos involucrados, tal nivel de datos acerca de un paciente determinado, puede mejorar la toma de decisiones en la clínica. La era pos GWAS y posgenética de la EM está en la vanguardia de la medicina de precisión, si no por su utilidad clínica, puede ser por la organización de redes colaborativas y el modelado de estadísticas complejas, requeridas para las aproximaciones que unen los descubrimientos robustos basados en las poblaciones y la evaluación del heterogéneo riesgo individual.

Palabras clave: esclerosis múltiple, medicina de precisión, genética, integración de datos, contextualización.

❑ Introduction

Multiple Sclerosis (MS) is a primary leading non traumatic cause of neurological disability for young adults[1]. To date, treatment is largely aimed at slowing down or stopping the disease process, however clinicians are in need or reliable markers and tools to guide the choice of therapy. The established risk factors for developing MS include genetic as well as environmental factors, although the increase of risk for the presence of each individual factor is generally modest to low. The expected interplay between these risk factors adds to the complexity of the disease[2]. The early genetic risk factors for MS were identified in the Major Histocompatibility Complex (MHC) region in the early days of serological study of the Human Leukocyte Antigen (HLA) region in the 1970's[3]. More recently, large-scale studies led by the international MS genetics consortium (IMSGC) analyzed of vast numbers of single nucleotide polymorphisms (SNPs) and reported 110 non-MHC loci associated with MS[4-6]. Similar to other complex diseases, the associations have modest to small effect sizes (OR ~1.1) and are frequent in the general population (10-50% minor allele frequency). The overall risk of an individual is driven by the total number of carried risk alleles, the environmental factors and interplay between these. Efforts to summarize the genetic risk for MS have reliably shown that an individual's risk increases with the number of risk alleles carried[7, 8].

Genes that are important for T helper cell differentiation show an overrepresentation amongst the associated MS risk SNPs, confirming the central role of the immune system in MS[6]. Further strengthening the involvement of the immune system is the efficacy of drugs affecting T- and B-cell function in the treatment of MS[9]. Involvement of CD4+ and CD8+ T cells in MS pathology has been shown during the past decade[10-12], and more recently Th17 cells, a subtype of the CD4+ T cell-lineage, was pushed forward as one of the driving cell types behind the pathological processes in MS[13]. Although the immune system is considered a major contributor in MS etiology, knowledge on the functional mechanisms that underlie the associations of specific gene variants with MS is sparse. The publications of the IMSGC led to an increase in studies into the molecular mechanisms which may underlie MS associations of specific SNPs. Furthermore, better insights in the underlying pathways are obtained through complex pathway analyses of all the implicated SNPs for MS[14]. To better facilitate clinical use of these markers there is a need for better insights in how risk factors contribute to the disease risk and progression of disease, both at population as well as per-subject level.

❑ Challenges in interpretation of genetic risk factors

Genetic screens allow an unbiased approach to identify the genetic components of diseases and can provide unforeseen insights into disease etiology, as well as confirm known or suspected disease pathways[6]. However, the polymorphic sites with MS association are likely to represent larger genetic regions due to local linkage disequilibrium (LD). LD results from non-random recombination events during meiosis in which several polymorphisms are typically transmitted together on stretches of DNA. Such LD blocks may span across multiple genes and may contain many more genetic variants, further complicating interpretation of the association signal[15]. An association signal may therefore also be observed for non-causal SNPs that display full, very strong or even moderate LD. On the other hand, the LD also allows the imputation of additional SNPs[16] when one or multiple of the flanking polymorphisms are actually genotyped and the LD structure of the population is known[17]. Imputation is a cost effective method to harmonize several genetic studies and has facilitated meta-analyses at world-wide scale for complex diseases including MS[4].

Further advances in genetics research may be expected from the recent advent of next-generation sequencing. Increasingly low sequencing costs enable researchers to sequence an entire human genome, specific candidate regions or the full gene expression fingerprint of a specific cell type (the transcriptome). A decade ago the cost of such experimental designs inhibited such studies, however in the near future more of such studies are expected to provide highly detailed insights in the genetic architecture of MS. Currently, studies into gene expression and rare genetic variants are emerging for less complex diseases such as specific cancer types and with increasing availability the more complex diseases are expected to follow. Genome sequencing may contribute with essential insights into rare genetic variation, including insertions and deletions at the DNA level. This technique may to some extent resolve the problem regarding LD structure encountered when using genotyping arrays since the per-base resolution will have more power to identify the most associated variant in a region of moderate to strong LD. Moreover, algorithms taking into account all known regulatory and molecular aspects of variants in their genetic context can rank SNPs in LD blocks based on the potential for having biological implications[18, 19]. Another approach consists of integrating previous knowledge from GWAS studies, expression profiles in different tissues and interaction between protein coding genes in a heterogeneous network to predict probability of gene-disease association. Such prioritization may expedite the identification of the true functional variants through molecular biology approaches. A functional follow upon the identification of associated genetic regions is elementary for the understanding of disease etiology. Nevertheless, the ‘markers’ of a genetic region that modulate risk of developing MS may be used to characterize patients without the knowledge of the underlying molecular mechanisms. It is worth noticing that the very nature of the genetic screens’ results are challenging for the reductionist approach of modern molecular and cellular biology. While genetics screens demonstrate that hundreds of SNPs are involved, molecular techniques usually test variant one at a time, thus neglecting the complex background suggested by genetics screens. Network based models are under development to tackle this issue and embrace the complexity of data with systemic approach[20].

❑ Summarizing genetic risk in scores

Complex disease genetics is characterized by the identification of common variants that do not cause disease onset by their mere presence. These variants are observed in individuals that will never develop disease and only contribute to a small increase in the risk of developing disease. It can be reasoned that with a larger number of such variants present by chance, the risk for an individual to actually develop disease will increase. Based on this hypothesis a score summarizing per-person these presence of all known genetic risk factors can be calculated: these scores can be conceived as a convenient tool to study the susceptibility in familial segregation or in association with other traits such as the Multiple Sclerosis Genetic Burden (MSGB)[7,21] or the score can be studied alone or in combination for its (limited) predictive properties such as the weighted Genetic Risk Score for MS (wGRS)[8]. Indeed, these studies have shown that MS patients display higher scores in comparison to control groups. Extending from this hypothesis, associations of these scores to (para-) clinical parameters of the disease such as age at onset, gender distribution and the presence of oligoclonal bands have been shown by multiple of studies[21-23].

The predictive power of these scores at a per-person level is limited. Additional genetic studies will continue to fuel these summary scores, potentially increasing the use for these scores to profile patient populations or subgroups of patients meeting specific criteria. The use of this score for individual patients is currently limited by lack of insights in the components these scores are composed of, however the expected increasing insights in the molecular mechanisms behind MS and the MS-associations of genetic variants may allow for use of refined scores in specific disease characteristics such as progression and/or response to treatment.

❑ Studies of detailed MS phenotypes

Phenotypic subdividing of patients in complex diseases can assist in the identification of relevant factors for specific disease manifestations, thereby providing new disease insights. For MS, patients may be subdivided into groups that show a relapsing remitting disease course (RRMS) progressing at a later point to the secondary progressive course (SPMS), and the less common primary progressive disease phenotype (PPMS). More specific disease sub phenotypes can be clinical or para-clinical characteristics such as presence or absence of oligoclonal bands in the CSF[24, 25] or quantification of the lesion load as determined though MRI[26]. Detailed molecular studies as well as translational studies that relate the molecular findings back to clinical presentations of disease require carefully characterized patients and where relevant healthy controls. Data from these detailed studies may further fuel the overall characterization of patients at a population level as well as the individual patient level.

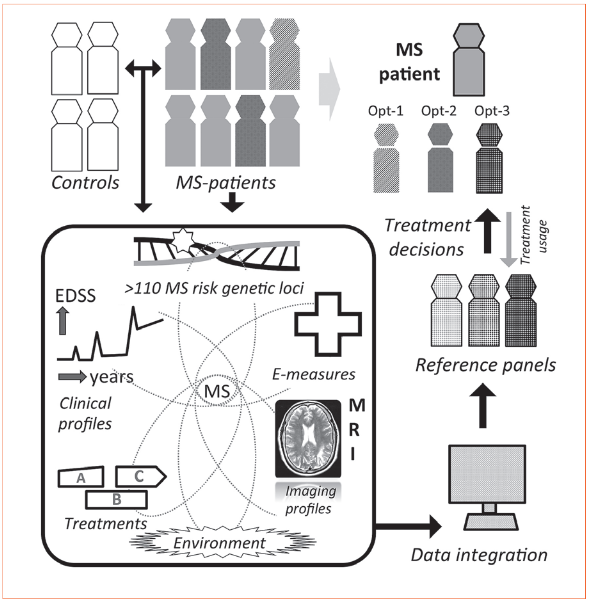

Figure 1 *PRISMA diagram demonstrating the search strategy and its results. The figure illustrates the growing role of multiple type of data obtained in reference longitudinal datasets in order to sustain with data the presentation of multiple therapeutic options for a particular decision making.*

For this reason, careful clinical and para-clinical phenotyping of patients that are already included in -omics studies is relevant for creating datasets that allow contextualization of individual patient findings. Large cohorts of well-characterized patients with long period of follow-up remain of extreme importance in the continuously evolving MS treatment landscape. Complex diseases are in need of data-driven evaluation of criteria like the "no evidence of disease activity" (NEDA)[27, 28]. Precision medicine will see the emergence of new ways to use data. As shown in Figure 1, groups of well-characterized patients for clinical, para-clinical and genetic parameters will continuously serve a reference panel for new individual patients and may give clinicians better grip on their individual disease profiles.

❑ The advent of precision medicine in complex diseases

The complexity of MS genetics has been insightful for an initiative led by Prof. Hauser's research group; the MS Bioscreen[29]. The development of this application has deep roots in MS Genetics complexity. The MSGB scores allowed the computation of an individual percentile reflecting a normality assessment of the genetic risk factors load in cases as well as in controls. The quantified nature of the genetic score and its translation into a percentile compared to reference population can directly be applied to any complex metrics such as the long-established EDSS or scores computed from MRI such as lesion load or parenchymal brain annual atrophy. The addition of these data in larger databases allows the individual patient to be put in context of a reference distribution for any elaborated metrics. This then allows for a better interpretation of the per-patient characteristics by clinicians and patients alike. The prototypic iOS tablet application called the "MS Bioscreen" adopted this "contextualization" principle at its core in order to develop first generation of precision medicine tools for complex traits[29]. Its contextualization engine leveraged the availability of large research cohorts as reference datasets to expand from the normality assessment of single points and offer a dynamic context to whole patient trajectories and thus inform decision-making in an actionable, natural way.

❑ The potential of epigenetic and transcriptomic profiles for precision medicine

New methods are allowing additional datasets to be generated, which may be integrated as reference panels in order to characterize individual patients. Here, we focus on emerging genetic data that may contribute to such deep phenotypes of individuals.

Similar to other diseases, MS heritability is far from fully explained. In order to better profile individual patients, it is important that research continues to identify additional heritable factors for MS. At least a part of the unexplained heritability could be ascribable to common genetic variations with only minimal effect sizes and large effect-size genetic variations with very low frequencies in the population. Furthermore, gene-gene interactions and gene-environment interactions[30] may be responsible for considerable proportion of heritability. More recently, research into inherited epigenetic marks such as DNA methylation[31] has become more affordable due to the development of of chip-based techniques that characterise large numbers of DNA methylation sites[32]. Given that DNA methylation is a strong regulator of gene activity, it is not remarkable that DNA methylation profiles have shown associations to complex diseases such as Type 1 diabetes and MS[33-36]. DNA methylation is, similar to gene expression, cell-specific and requires collection of samples with high cellular homogeneity. Larger studies applying upcoming tools which have a higher coverage of the DNA methylation sites in the human genome will increase our insight in the epigenetic mechanisms involved in disease development and potentially allows for this type of data to contribute to precision medicine.

Similar to epigenetic research, data on gene expression of relevant cell types in MS may directly reflect activity of pathological processes in MS patients. Using the RNA expression profiles of whole blood, Parnell *et al*. reported differences in expression of the transcription factors EOMES and TXB in MS[37]. Given that whole blood is a rather heterogeneous sample including diverse cell types, the isolation of specific cell types may further increase the knowledge on involved genes in MS. Indeed, a study on the T cell receptor β (TCR-β) of $CD8^+$ T cells isolated from cerebrospinal fluids (CSF) and blood of MS patients illustrates the value of emerging techniques such as RNA sequencing. Lossius *et al*. determined whether specific clones of $CD8^+$ T cells expanded in the CSF or in the blood compartiment[38]. Decreasing cost, in addition to improving techniques to purify specific cell types from heterogeneous collections will further potentiate DNA and RNA sequencing studies.

The specific data that results from epigenetic and gene expression studies may play an important role in determining an individual patients' potential for disease progression. The availability of well-characterized reference panels is also for these types of data of high importance. Although epigenetic and gene expression data is more complex, meta-analyses of several smaller studies may provide enough statistical power to identify new genetic and genomic leads in these data.

❑ A future for precision medicine in MS

The following years will see an expanding knowledge of the molecular mechanisms behind currently known genetic associations in MS. Studies are ongoing and planned to identify additional gene variants associated with MS in general as well as genetic associations within subgroups of MS patients. In addition, it is likely that other approaches such as epigenetic studies, transcriptomics and functional studies will provide additional important knowledge of disease mechanisms and MS endophenotypes. The ongoing collaborative efforts both in local and worldwide settings have made it possible to identify enough MS patients to attain the required power. These efforts may be prophetic of emerging collaborative research work, data-sharing infrastructure and open-source algorithms development. A continuous increased understanding of the molecular mechanisms that impact on MS disease risk and progression will hopefully lead to improved treatments for this disease.

At the intersection of Science, Technology and Data lies a new, modern form of precision medicine that is bound to add tremendous value in overall clinical care for complex diseases and MS in particular. This precision medicine stems from the recent advances in the treatment and management of genomic, proteomic, imaging and clinical data but aims at directly informing care for individual patients. To this aim, it integrates ever-growing sources of clinical and para-clinical data into a framework that translates the most recent research advances and population structure into actionable tools that (i) informs clinical decision making and (ii) improves communication around clinical decisions between the different stakeholders (patient, clinician, caregiver, provider). While this approach is still in a prototyping phase, the consensus around the need, usefulness and feasibility of such precision medicine tools warrants their development and implementation in the near future.

Bibliografía

1.- Compston A & Coles A. Multiple sclerosis. Lancet 2008;372:1502-17.

2.- Gourraud PA, Harbo, HF, Hauser SL & Baranzini SE. The genetics of multiple sclerosis: an up-to-date review. Immunol 2012;248:87-103.

3.- Naito S, Namerow N, Mickey MR & Terasaki PI. Multiple sclerosis: association with HL-A3. Tissue Antigens 1972; 2:1-4.

4.- International Multiple Sclerosis Genetics, C. et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet 2013;45:1353-60.

5.- International Multiple Sclerosis Genetics Consortium, et al. Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 2007;357:851-62.

6.- International Multiple Sclerosis Genetics Consortium, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 2011;476:214-9.

7.- Gourraud PA, et al. Aggregation of Multiple Sclerosis Genetic Risk Variants in Multiple and Single Case Families. Annals of Neurology 2011;69:65-74.

8.- De Jager PL, et al. Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. Lancet Neurology 2009;8:1111-19.

9.- Linker RA, Kieseier BC & Gold R. Identification and development of new therapeutics for multiple sclerosis. Trends in Pharmacological Sciences 2008;29:558-65.

10.- Chitnis T. The role of CD4 T cells in the pathogenesis of multiple sclerosis. Int Rev Neurobiol 2007;79:43-72.

11.- Huseby ES, Huseby PG, Shah S, Smith R & Stadinski BD. Pathogenic CD8 T cells in multiple sclerosis and its experimental models. Front Immunol 2012;3:64.

12.- Broux B, Stinissen P & Hellings N. Which immune cells matter? The immunopathogenesis of multiple sclerosis. Crit Rev Immunol 2013;33:283-306.

13.- Kleinewietfeld M & Hafler DA. Regulatory T cells in autoimmune neuroinflammation. Immunol Rev 2014;259:231-44.

14.- International Multiple Sclerosis Genetics Consortium. Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls. Am J Hum Genet 2013;92:854-65.

15.- Slatkin, M. Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 2008;9:477-85.

16.- Howie B, Fuchsberger C, Stephens M, Marchini J & Abecasis GR. Fast and accurate genotype imputation in genomewide association studies through pre-phasing. Nature Genetics 2012;44:955-9.

17.- Genomes Project Consortium, et al. A global reference for human genetic variation. Nature 2015;526:68-74.

18.- Farh KK, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 2015;518:337-43.

19.- Bailey P, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature 2016.

20.- Wang L, Himmelstein DS, Santaniello A, Parvin M & Baranzini SE. iCTNet2: integrating heterogeneous biological interactions to understand complex traits. F1000Res 2015;4:485.

21.- Harbo HF, et al. Oligoclonal bands and age at onset correlate with genetic risk score in multiple sclerosis. Mult Scler 2014;20:660-8.

22.- Hilven K, Patsopoulos NA, Dubois B & Goris A. Burden of risk variants correlates with phenotype of multiple sclerosis. Multiple Sclerosis Journal 2015;21:1670-80.

23.- Barizzone N, et al. The burden of multiple sclerosis variants in continental Italians and Sardinians. Multiple Sclerosis Journal 2015;21:1385-95.

24.- Goris A, et al. Genetic variants are major determinants of CSF antibody levels in multiple sclerosis. Brain 2015;138:632-43.

25.- Mero IL, et al. Genetic differences relating to oligoclonal band status in multiple sclerosis. Multiple Sclerosis Journal 2012;18:17-18.

26.- Berg-Hansen P, Moen SM, Harbo HF & Celius EG. High prevalence and no latitude gradient of multiple sclerosis in Norway. Mult Scler 2014;20:1780-2.

27.- Nygaard GO, et al. A Longitudinal Study of Disability, Cognition and Gray Matter Atrophy in Early Multiple Sclerosis Patients According to Evidence of Disease Activity. PLoS One 2015;10, e0135974.

28.- Havrdova E, et al. Effect of natalizumab on clinical and radiological disease activity in multiple sclerosis: a retrospective analysis of the Natalizumab Safety and Efficacy in Relapsing-Remitting Multiple Sclerosis (AFFIRM) study. Lancet Neurol 2009;8:254-60.

29.- Gourraud PA, et al. Precision Medicine in Chronic Disease Management: The Multiple Sclerosis BioScreen. Annals of Neurology 2014;76:633-42.

30.- Buil A, et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. Nat Genet 2015;47:88-91.

31.- Cortijo S, et al. Mapping the epigenetic basis of complex traits. Science 2014;343;1145-8.

32.- Bibikova M, et al. High density DNA methylation array with single CpG site resolution. Genomics 2011;98:288-95.

33.- Dang MN, Buzzetti R & Pozzilli P. Epigenetics in autoimmune diseases with focus on type 1 diabetes. Diabetes Metab Res Rev 2013;29:8-18.

34.- Bos SD, et al. Genome-wide DNA methylation profiles indicate CD8+ T cell hypermethylation in multiple sclerosis. PLoS One 2015;10, e0117403.

35.- Graves M, et al. Methylation differences at the HLA-DRB1 locus in CD4+ T-Cells are associated with multiple sclerosis. Mult Scler 2013;20:1033-41.

36.- Maltby VE, et al. Genome-wide DNA methylation profiling of CD8+ T cells shows a distinct epigenetic signature to CD4+ T cells in multiple sclerosis patients. Clin Epigenetics 2015;7:118.

37.- Parnell GP, et al. The autoimmune disease-associated transcription factors EOMES and TBX21 are dysregulated in multiple sclerosis and define a molecular subtype of disease. Clin Immunol 2014;151:16-24.

38.- Lossius A, et al. High-throughput sequencing of TCR repertoires in multiple sclerosis reveals intrathecal enrichment of EBV-reactive CD8+ T cells. Eur J Immunol 2014;44:3439-52.