Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Analysis of Genetic Diversity in Patients with Major Psychiatric Disorders versus Healthy Controls: A molecular-genetic study of 1,698 subjects genotyped for 100 candidate genes (549 SNPs)

Hans Stassen (k454910@bli.uzh.ch)
Psychiatric University Hospital (KPPP)
Silke Bachmann (silke.bachmann@gmail.com)
René Bridler (R.Bridler@sanatorium-kilchberg.ch)
Katja Cattapan (Katja.Cattapan@sanatorium-kilchberg.ch)
Annette Hartmann (annette.hartmann@meduniwien.ac.at)
Dan Rujescu (Dan.Rujescu@meduniwien.ac.at)
Medical University of Vienna https://orcid.org/0000-0002-1432-313X
Erich Seifritz (erich.seifritz@bli.uzh.ch)
Matthias Weisbrod (matthias.weisbrod@srh.de)
Christian Scharfetter (c.scharfetter@ifrg.ch)

Article

Keywords:

DOI: https://doi.org/

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors have declared there is NO conflict of interest to disclose

source: https://doi.org/10.48350/176187 | downloaded: 26.4.2024

Analysis of Genetic Diversity in Patients with Major Psychiatric Disorders versus Healthy Controls

A molecular-genetic study of 1,698 subjects genotyped for 100 candidate genes (549 SNPs)

H.H. Stassen [1], S. Bachmann [2], R. Bridler [3], K. Cattapan [3,4], A.M. Hartmann [5], D. Rujescu [5], E. Seifritz [6], M. Weisbrod [7] and Chr. Scharfetter [1†]

[1] Institute for Response-Genetics, Department of Psychiatry, Psychotherapy and Psychosomatics, Psychiatric University Hospital, CH-8032 Zurich, Switzerland
[2] Department of Psychiatry, Psychotherapy, and Psychosomatics, University of Halle, D-06112 Halle, Germany & Clienia AG, Psychiatric Hospital, CH-9573 Littenheid, Switzerland
[3] Sanatorium Kilchberg, CH-8802 Kilchberg, Switzerland
[4] University Hospital of Psychiatry and Psychotherapy, University of Bern, Switzerland
[5] Medical University of Vienna, Clinical Division of General Psychiatry, A-1090 Wien, Austria
[6] Department of Psychiatry, Psychotherapy and Psychosomatics, Psychiatric University Hospital, CH-8032 Zurich, Switzerland
[7] Department of General Psychiatry, Center of Psychosocial Medicine, University of Heidelberg, D-69115 Heidelberg, Germany, SRH Klinikum Karlsbad-Langensteinbach

Institute for Response-Genetics
Psychiatric University Hospital (KPPP), University of Zurich
CH-8032 Zurich, Switzerland
Websites: "https://www.ifrg.ch", "https://www.blix.uzh.ch"
Email: "k454910@bli.uzh.ch"

† This work is dedicated to Professor Dr. Christian Scharfetter, a colleague and good friend over decades, with whom I planned and realized this project. Inconceivably for all of us, he passed away much too early after a short, serious illness. *Hans H. Stassen*

silke.bachmann@gmail.com
R.Bridler@sanatorium-kilchberg.ch
Katja.Cattapan@sanatorium-kilchberg.ch
annette.hartmann@meduniwien.ac.at
dan.rujescu@meduniwien.ac.at
erich.seifritz@bli.uzh.ch
Matthias.Weisbrod@srh.de
c.scharfetter@ifrg.ch

[Translational Psychiatry]

Abstract

In this study we addressed the question of the extent to which irregularities in genetic diversity might separate patients with major psychiatric disorders from healthy controls. Genetic diversity was quantified per gene through multidimensional "gene vectors" assembled from 4-8 polymorphic SNPs located within each of 100 candidate genes. The number of different genotypic patterns observed per gene was called the gene's "diversity index". Our sample was comprised of 1,698 subjects from Central Europe (1,431 psychiatric patients, 267 healthy controls), all genotyped for 549 specifically selected SNPs.

The evaluation of the diversity indices of the 100 candidate genes resulted in a mean value of 109.4±82.8, ranging from 18 to 476. Highly significant deviations from "normal" diversity values were detected for (1) major depression (n=596): a significant reduction ($p<0.0001$); (2) Alzheimer's disease (n=75): a significant reduction ($p<0.0001$); and (3) schizoaffective disorders (n=64): a significant increase ($p<0.0001$). Almost one third of the genes were correlated with each other, with correlations ranging from 0.0303 to 0.7245.

The central finding of this study was the discovery of "singular genes" characterized by distinctive genotypic patterns that appeared exclusively in patients but not in healthy controls. In each of the diagnostic subgroups under study, there were no less than 45%-55% of patients who exhibited genotypic patterns of singular genes that did not at all show up in the healthy controls. Neural Net (NN) analyses enabled the construction of nonlinear classifiers that correctly identified up to 90% of patients in comparisons with healthy controls at false-positive error rates of zero percent. The NN analyses revealed considerable overlaps on the genotype level between the various clinically defined diagnostic subgroups, suggesting that diagnosis-crossing, unspecific vulnerabilities are likely involved in the pathogenesis of major psychiatric disorders.

Clinical applications of the proposed method are immediately possible and will facilitate the early detection of latent psychiatric disorders among risk cases, so that early interventions can be started before clinically relevant symptoms develop. A larger number of hospitalizations could be prevented in this way.

Keywords: Vulnerability, resilience, gene vectors, singular genes, neural nets, artificial intelligence, classifiers, schizophrenia, depression, bipolar illness, schizoaffective disorders, Alzheimer's disease

Background

There is little proven knowledge about etiology and pathogenesis of psychiatric disorders[1]. Even after 50 years of modern psychiatry, (1) there are no causal treatment options; (2) it is not possible to reliably predict if and when a particular patient will respond to a particular treatment; and (3) in individual cases it is hardly possible to make any reliable prognosis.

As to the genetically predisposed factors postulated to be involved in the pathogenesis of psychiatric disorders, evidence clearly speaks against single causes as psychiatric disorders aggregate in families, but do not segregate. That is, psychiatric disorders do not follow simple Mendelian modes of inheritance. No homotypic diagnostic patterns are observed in families with multiple affected subjects: typically, the clinical diagnoses of first and second degree relatives appear to be largely independent of the index case's primary diagnosis. This raises doubts about the usefulness of psychiatric diagnoses as the main source of genetic studies. Syndrome-oriented approaches might be more appropriate when investigating, for example, the nature of depressive symptoms among patients suffering from schizophrenic disorders in contrast to major depression.

Our studies of monozygotic (mz) twins discordant for schizophrenic disorders, who share identical genomes, have made it clear that genetically predisposed factors are not a sufficient condition for the development of psychiatric disorders [**1**]. Rather, genetics in psychiatry does not act in the sense of a definitive, unalterable fate in terms of a "biogenic deficit" or a combination of "biogenic imbalances". Susceptible subjects, like the unaffected co-twins of mz twins with schizophrenic disorders, can function perfectly well in daily life if they take the necessary precautions.

Another crucial point regarding psychiatric disorders is that we cannot assume etiological entities. Rather, etiological heterogeneity appears to reflect clinical reality more convincingly, suggesting that multiple pathways can lead to the same clinical picture. *Eugen Bleuler*, the renowned father of "schizophrenia" and former director of our hospital, already spoke of the "group of schizophrenias" in order to emphasize the etiological heterogeneity of schizophrenic disorders [**2**].

Taking all available information together, the most plausible and most likely etiological scenario is a complex interplay between multiple, genetically predisposed endogenous

[1] We use the term "psychiatric disorders" to emphasize that we are working here with the practical realization of the theoretical concept of "mental disorders" on the basis of quantitative/qualitative symptoms and signs.

111 factors and multiple exogenous factors, which may induce the development of latent
112 disorders. In this scenario, exogenous factors ultimately trigger the manifestation of clinically
113 relevant symptoms. Among the exogenous factors, lifestyle, diet, consumption behavior, and
114 physical activity play a prominent role. Inflammation is another major exogenous constituent
115 explaining some 15-25% of the observed phenotypic variance [**3**-**4**].

116 In this project, we did not aim to elucidate the genetic background of major psychiatric
117 disorders by means of standard genotype-to-phenotype association methods that use
118 "psychiatric diagnosis" as phenotype (e.g., *GWA*s: genome-wide associations) [**5-9**]. Rather,
119 we addressed the question of the extent to which irregularities in genetic diversity might
120 separate patients with major psychiatric disorders from healthy controls, with "genetic
121 diversity" denoting the multitude of genotypic patterns observed with each gene. In
122 particular, we were interested in vulnerability and resilience[2] genes that might be specific to
123 psychiatric disorders.

124 Inevitably, analyses of genetic diversity bring up the question of biological ethnicity
125 ("population stratification") [**10-12**], as it may well be that any differences between patients
126 and healthy controls are due to population stratification rather than to the disorders under
127 investigation. We tackled this problem in two different ways: (1) we aimed to recruit half of
128 the healthy controls from the patients' unaffected first-degree relatives; and (2) with
129 "unsupervised learning" methods[3] and 73 SNPs located within the *CLOCK* gene we aimed to
130 develop a "natural" model of biological ethnicity. The *CLOCK* gene was chosen because it
131 was deemed to contain distinctive adaptations of typical North-South and West-East
132 specifics. Both methodological approaches allowed us to estimate the amount of variance
133 that is explainable by population stratification.

134 Using 100 candidate genes reported in the literature as likely to be involved in the
135 pathogenesis of psychiatric disorders, and whose genotypic patterns were assessed through
136 549 SNPs (the genes' distinctive "fingerprints"), we searched for psychiatry-related
137 configurations of vulnerability and resilience genes by means of methods of Artificial
138 Intelligence (AI) in combination with multi-layer Neural Nets (NNs) ("supervised learning").
139 Specifically, we addressed the following questions:

---

[2] The term "resilience" is used here as a broad concept, encompassing all those endogenous mechanisms that support and maintain health, thereby enabling patients to cope with challenging situations.

[3] "Unsupervised learning" detects "natural" structures in empirical data using metric/non-metric distance or similarity measures in connection, for example, with "nearest neighbor" methods and random seeds.

(1) How to quantify genetic diversity at high resolution in a reproducible way?
(2) Are there genes for which genetic diversity is reduced in male schizophrenic patients, given the fact that some 80% of male patients have no offspring?
(3) Are there psychiatry-specific vulnerability and resilience genes, or combinations thereof, whose genotypic patterns discriminate between psychiatric patients and healthy controls, or between psychiatric diagnoses?
(4) To what extent do vulnerability and resilience genes correlate with each other, i.e. are there genotypic patterns that show up more than randomly with each other?

Methods: Data Material

Data from patients and controls from five of our previous studies were (1) pooled, (2) coded in a standardized way, and (3) analyzed together. The study details can be found elsewhere [**13**-**18**]. Totally 1,698 subjects were genotyped for 100 genes[4] and 549 specifically selected SNPs at a missing data rate < 5% (96 autosomal genes; 1,431 psychiatric patients; 267 healthy controls, of which 141 (52.8%) were unaffected 1st degree relatives of the patients).

The patients had been recruited from the daily admissions at three university hospitals in Switzerland and Germany, and from the daily admissions at two private mental health treatment centers in Switzerland. Selection criterion had been a suspected ICD-10 diagnosis of one of the following disorders: F20 (schizophrenia), F25 (schizoaffective disorders), F31 (bipolar illness), and F32 or F33 (major depression). All patients had been informed about the goals of this research project and that they can discontinue participation at any time without giving reasons and without facing any disadvantages from this. Finally, the patients had signed a written informed consent before entering the studies.

The patients' psychopathology had been assessed by specifically trained interviewers. The study protocol included (1) assessments of previous history and overall social functioning through the syndrome-oriented 63-item SADS Syndrome Check List SSCL-16 and the 83-item SADS-Supplement SSCL-SUPP (lifetime versions) [**19**]; (2) assessments of the 30-item Positive and Negative Syndrome Scale PANSS [**20**] and/or the 17/21-item Hamilton Depression Scale HAM-D [**21**] over 5 weeks; (3) assessments of medication and unwanted side effects through the 46-item Medication and Side Effects Inventory MEDIS [**22**]; and (4) the collection of blood samples for serum extraction and DNA isolation.

[4] For details see supplementary Tables 1, 2.

A minimum baseline score of at least 21 on the general psychopathology PANSS-G Scale (primary "F2x.x" diagnoses), or of at least 15 on the HAM D17 Scale (primary "F3x.x" diagnoses), was required at entry into study. The definitive diagnoses for this project were decided by consensus of two experienced senior psychiatrists, with unclear cases being assigned to the residual group "other diagnoses".

The healthy control subjects had been recruited either through advertising or from the patients' unaffected first-degree relatives. All control subjects had filled out the 63-item Zurich Health Questionnaire "ZHQ" [**23**]. On the basis of ZHQ data, we assigned subjects with a significant history regarding «consumption behavior», «psychosomatic disturbances», or «impaired mental health», to the residual diagnostic subgroup "other diagnoses".

The Alzheimer's disease patients came from the NIHM (DNA and DSM-4 diagnoses). Details on the final sample composition are given in Table 1.

>>>>> insert Table 1 about here <<<<<

Methods: Quantifying Genetic Diversity

Our approach to estimating the genetic diversity associated with a catalog of 100 genes relied on "gene vectors" which were assembled per gene from the genotypes of 4-8 polymorphic SNPs located within each gene. As a SNP can exhibit three different expressions regardless of allele definition, a base-3 system[5] was used to construct gene vectors:

"gene vector": $v_i^{(j)} = \sum_{k=1}^{m(j)} s_{ik}^{(j)} \, 3^{k-1}$

$i=1,2, \ldots N$	subjects
$J=1,2, \ldots M$	genes
$s_{ik}^{(j)} \in \{0,1,2\}$	SNPs
$m(j)$	number of SNPs in the *j-th* gene

With *m* SNPs, a total of 3^m different genotypic patterns would be theoretically possible for a gene. However, no more than half of them were actually found in this project's population of 1,698 subjects, due to the correlation of SNPs within genes. As a rule of thumb, one can expect an average of 100 different genotypic patterns for a 10-dimensional gene vector of five SNPs, thus implying a pretty high resolution regarding the envisaged quantification of genetic diversity, as plenty of "variation" means plenty of "information". The number of different genotypic patterns of a gene was referred to as the gene's "diversity index".

[5] A base-4 system would make the genotypic patterns much easier for people to read, but at the cost of 25% more memory and a 25% higher computational load.

As the observable genetic diversity critically depends on sample size, we generated calibration data by drawing 32 random samples of equal size from the total sample (n=1,698) for each gene, and for 24 sample sizes in steps of 50 between 50 and 1,200. By averaging across the 32 random samples, we obtained 100*24 normative distribution functions for the 100 candidate genes and the various sample sizes given in the diagnostic subgroups.

As an estimate of the correlation between two genes j_1 and j_2, we used the maximum frequency among the combinations of genotypic patterns of gene j_1 with gene j_2, divided by the sample size.

Methods: Neural Nets and Artificial Intelligence

Nonlinear NN models connect the "neurons" of the input layer (the subjects' gene vectors) with the "neurons" of the output layer (the subjects' psychiatric diagnoses) via "hidden" layers. Our goal was to construct NN models that classified all 1,698 subjects in terms of psychiatric diagnoses through their gene vectors as correctly as possible (Figure. 1).

>>>>> insert Figure 1 about here <<<<<

NN connections are realized by (1) weight matrices; and (2) model fitting algorithms minimizing an error function in the weight space ("goodness of fit"). All outputs are computed using sigmoid thresholding of the scalar product of the corresponding weight and input vectors. Outputs at stage "***s***" are connected to each input of stage "***s***+1". The most popular model fitting strategy, the backpropagation algorithm [**24**], looks for the minimum of the error function using the method of gradient descent:

(i)	**Output**:	$s_i = \sigma\left[\sum_j w_{ij} s_j\right]$	s_i: y_i observed	(i = 1,2,... N_i)
(j)	**Hidden layers**:	$s_j = \sigma\left[\sum_k w_{jk} s_k\right]$		(j = 1,2,... N_j)
(k)	**Input**:	$s_k = x_k$	x_k observed	(k = 1,2,... N_k)
	Improvements:	$\Delta w_{ij} = \alpha \cdot \varepsilon_i^{\nu} \cdot s_j \cdot s_i(1-s_i) \quad \varepsilon_i^{\nu} = y_i^{\nu} - s_i^{\nu}$		(ν = 1,2,.. p)
		$\Delta w_{jk} = \alpha \cdot \sum_{i=1}^{N_i} \varepsilon_i^{\nu} \cdot s_k \cdot s_i(1-s_i) \cdot w_{ij} \cdot s_j(1-s_j)$		

Here $\boldsymbol{x}_k$ denote observed stimuli, $\boldsymbol{y}_j$ observed responses, $\boldsymbol{\sigma}$ the activation function of sigmoid-type: **R→(0,1)**, $\boldsymbol{\alpha}$ the learning rate, and $\boldsymbol{p}$ the number of subjects. The achievable precision of

the model essentially depends on the number of intermediate layers implemented to model nonlinear interactions. The computational load, on the other hand, increases exponentially with the number of layers.

Additionally, we relied on methods of Artificial Intelligence (AI) and searched for (1) illness-specific genes for which genotypic patterns showed up exclusively in patients, but not in healthy controls; and (2) genotypic patterns that clustered (>70%) in one diagnostic subgroup, while being rare (<5%) in at least one of the other diagnostic subgroups. The genes and genotypic patterns identified this way were assigned special weights that were further iteratively optimized and used as a-priori knowledge in the NN analyses.

Methods: Quantifying Biological Ethnicity

To construct a "natural" model of biological ethnicity, we relied on 73 polymorphic SNPs located within the *CLOCK* gene. We quantified the subjects' biological ethnicity through the five gene vectors derived by subdividing the gene into five segments, each with 15 SNPs (the 5^{th} segment held 13 SNPs). As "unsupervised learning" methods, we used six different cluster analyses (SAS/STAT 9.4 PROCs: *ACECLUS*, *CLUSTER*, *FASTCLUS*, *MODECLUS*, and *VARCLUS*) to detect "natural" subgroups that constituted population stratification. A principal component analysis was carried out prior to the cluster analyses which eliminated the correlations between the five gene vectors (SAS/STAT 9.4 PROCs: *PRINCOMP*, *FACTOR*).

Methods: Statistical Analyses

We used the *Statistical Analysis Software SAS/STAT 9.4* by SAS Institute Inc. and *PROC HPNEURAL* from *SAS Enterprise Miner 15.1* for Neural Net analyses, complemented by NN and AI programs developed at our institute.

Ethics

The studies were approved by the local ethics committees of the Canton of Zurich, the Canton of Thurgau, the University of Heidelberg and the University of Munich. All participants signed the written informed consent.

Results: Diversity Index

In this Central European population of 1,698 subjects, the evaluation of the diversity indices of the 100 candidate genes resulted in a mean value as high as 109.4±82.8, ranging from 18 (*CYP2C19*) to 476 (*GPR39*). The diversity indices depended primarily on the genes and only to

a minor extent on the number of SNPs making up the gene vectors. The distribution of the diversity indices exhibited two peaks (diversity indices around 70 and 170), along with seven genes exhibiting a diversity index above 250 (Figure 2). It is expected that genes with a higher diversity index will show a higher discriminating power when it comes to resolving subtle between-population differences.

>>>>> insert Figure 2 about here <<<<<

The 100*24 normative calibration curves, covering all 100 genes and population sizes of this project, displayed a very robust behavior with respect to scattering and, when regarded as a function of sample size, with respect to continuity. Therefore, simple linear interpolation between sampling points was sufficient to calculate diversity indices for intermediate sample sizes. Even extrapolations beyond the total sample size of 1,698 subjects appeared to work quite well (Table 2).

>>>>> insert Table 2 about here <<<<<

This robustness became evident, for example, through Figure 3 which shows the diversity indices of the two genes *CYP2J2* and *SCL6A6* for sample sizes between 50 and 1,700 in steps of 50. Noticeable differences between the two curves in terms of shape and steepness likely indicate different gene types, as *CYP2J2* belongs to the left gene group in Figure 2 (distribution peak around 70), and *SLC6A6* to the middle gene group (distribution peak around 170).

>>>>> insert Figures 3 about here <<<<<

The validity of the normative calibration curves was verified by comparing males (n=742) with females (n=956) regarding the diversity indices of 96 autosomal genes taken as an entity. Virtually no differences showed up after correction for sample size (p=0.9459). None of the genes made an exception in this respect.

Next, we took the distribution of the diversity indices of the total sample (n=1,698) as reference and carried out comparisons with the diagnostic subgroups with respect to the diversity indices of 96 autosomal genes taken as an entity. After correction for sample size, the analysis yielded several highly significant differences: (1) a significant reduction in genetic diversity ($p<0.0001$) for patients suffering from major depression (n=596); (2) a significant reduction in genetic diversity ($p<0.0001$) for patients suffering from Alzheimer's disease (n=75); and (3) a significant increase in genetic diversity ($p<0.0001$) for patients suffering from schizoaffective disorders (n=64). The deviations were related to a small

number of genes, while the vast majority of genes showed no such differences. Contrary to expectations, the hypothesis of a reduction in genetic diversity among male patients for schizophrenia-specific genes could not be confirmed (p=0.0693).

Finally, we analyzed the extent to which genes were correlated with each other. It turned out that almost one third of the genes under investigation showed such correlations. For example, we found for the subgroup of patients suffering from schizophrenic disorders (n=363) correlation coefficients ranging from r=0.0303 (*GRIK3*/*TNF*) to r=0.7245 (*CYP3A5*/*CYP3A7*), with a mean correlation of 0.1027±0.1025. The differences to the subgroup of patients diagnosed with major depression (n=596) were marginal, with a maximum correlation of 0.7248 (*CYP3A5*/*CYP3A7*) and a mean correlation of 0.1069±0.1020. The same was true for the subgroup of healthy controls, with a mean correlation of 0.1069±0.1020 and a correlation of 0.6854 between *CYP3A5* and *CYP3A7*.

Results: Singular Genes

The distributions of the genotypic patterns of the genes under study showed no substantial differences between healthy controls and the patients of the 5 diagnostic subgroups (Fig. 4A, B, D), with the only exception of a few genes among the Alzheimer's patients (Fig. 4C). Although comparisons of single genotypic patterns occasionally reached statistical significance outlasting Bonferroni corrections, the phenotypic variance explained by this remained very small and was not additive, comparable to the situation with single SNPs.

>>>>> insert Figure 4 about here <<<<<

By contrast, detailed analyses revealed genes that appeared to be illness-specific, as they exhibited genotypic patterns that were found exclusively in patients but not in healthy controls. For example, 33.9% of schizophrenic patients showed distinctive genotypic patterns inherent in gene *GPR39* which were completely absent in healthy controls. Similarly, 33.0% of depressed patients showed distinctive genotypic patterns inherent in gene *GRIA1*; 21.8% of bipolar patients showed distinctive genotypic patterns inherent in gene *STAT1*; 25.8% of schizoaffective patients showed distinctive genotypic patterns inherent in gene *ABCB1*; and 18.7% of Alzheimer's patients showed distinctive genotypic patterns inherent in gene *SCL6A1*, with all of those genotypic patterns being completely absent in healthy controls. Because of their distinctive characteristics, these genes were termed “singular genes”.

For each diagnostic subgroup, we found some 13-30 singular genes whose genotypic

patterns appeared exclusively in at least 10% of patients but not in healthy controls. As one would have expected, most of the singular genes had higher than average diversity indices (as "variation" means "information"). The number of singular genes did not depend on sample size: (1) a total of 29 singular genes were found in the subgroup of schizophrenic patients (n=363), virtually identical with the 28 singular genes observed in the subgroup of bipolar patients (n=134); whereas (2) just 24 singular genes showed up in the subgroup of depressive patients (n=596), compared to the 33 singular genes found in the much smaller subgroup of schizoaffective patients (n=62) (Table 3).

>>>>> insert Table 3 about here <<<<<

Singular genes were found to be inter-correlated within diagnostic groups, thus leading to overlaps between the patients identified by these genes (indicating "non-additivity of singular genes"). In consequence, pooling the patients typically covered about 45%-55% per diagnostic subgroup. The singular genes differed from diagnostic subgroup to diagnostic subgroup regarding genotypic patterns as well as intrinsic weights. Therefore, it was even possible to identify a set of singular genes specific to the differences between schizophrenic and depressed patients. By contrast, we have not been successful in finding health-specific "resilience genes", i.e. genes with genotypic patterns observed in significant numbers among healthy controls but not in patients.

To verify the reproducibility of the results, we weakened the rigorous definition of "healthy" for the control group ("Controls"; n=267) by extending it with those 201 cases who did not meet the criteria of major psychiatric disorders at entry into study ("Controls(+)"; n=468). But this left the results essentially unchanged. Only the number of singular genes reaching significance dropped somewhat in each diagnostic subgroup (Table 3).

It is unlikely that the existence of singular genes in our data was for the most part the result of population stratification, as half of the healthy controls were unaffected 1st-degree relatives of the study patients. Furthermore, no interrelation to the status of affectedness or to the patients' clinical diagnoses was found for the "biological ethnicity" groupings revealed by the cluster analyses[6] on the basis of 73 polymorphic SNPs located within the CLOCK gene.

All this underlined that the patients of this study possessed true illness-specific irregularities in genetic diversity, expressed by singular genes that exhibited a variety of genotypic patterns not found in healthy controls.

[6] Details on the results of Cluster and Principal Component analyses are available on request.

Results: Neural Net Analyses

Augmented by the pre-structured a priori knowledge of singular genes, the NN analyses achieved satisfactory to good steady-state results when comparing, for example, diagnostic subgroups with healthy controls at the clinically desirable false-positive error rate of 0%. Most notably, for the diagnostic subgroups of patients with schizophrenic disorders, depression, bipolar illness, and schizoaffective disorders, the NN algorithm yielded in each case a rate of about 90% correctly classified patients along with a 10% subgroup labeled as "unknown" when corrected for sample size (Table 4). The only exceptions were (1) the subgroup of patients suffering from Alzheimer's disease which performed with 80% correctly classified subjects slightly worse; and (2) the conglomerate subgroup of patients with "other diagnoses" where the optimization terminated with almost 40% of "unknowns" (39.8% false-negative error rate).

>>>>> insert Table 4 about here <<<<<

The NN method was somewhat less successful in the construction of classifiers that separated patients diagnosed with schizophrenic disorders from patients with (1) bipolar illness; (2) depression; or (3) schizoaffective disorders. The performance of the resulting classifiers was with false-negative error rates of almost 20% less efficient compared to what we saw in the comparisons between the diagnostic subgroups and healthy controls. In particular, the NN constraint of a clinically desirable false-positive error rate of 0% could not be upheld and had to be raised to 5% to achieve useful results. All this was due to considerable overlaps between the diagnostic subgroups under comparison.

The classifiers identified through the NN analyses were composed of 6-10 genes: 4-5 core genes that were common to all classifiers, plus 2-5 accessory genes that depended on the target population (Table 5). The classifiers turned out to be non-unique. It was readily possible to exclude 1-2 genes (up to 3 genes) of an optimized classifier and re-run the NN analyses. This replaced the eliminated genes by other compatible genes and adjusted the weight matrices accordingly, so that the modified classifiers achieved similar performances.

>>>>> insert Table 5 about here <<<<<

This apparent redundancy was due to the fact that the genes, especially the singular genes, were not independent of each other but inter-correlated. For example, in the diagnostic subgroup of schizophrenic disorders (n=363), gene *STAT1* was correlated with genes *CYP3A5*, *CYP3A7*, *CYP3A4*, *CYP1A1*, *CYP1A2*, *CYP2B6*, and *CYP2D7*, with correlation

coefficients between 0.1377 and 0.2287. And gene *STAT4* was correlated with genes *CYP3A5*, *CYP3A7*, and *CYP2B6*, with correlation coefficients ranging from 0.1240 to 0.1405, while gene *CYP27A1* was correlated with genes *CYP3A5*, *CYP3A7*, *CYP3A4*, and *SLC4A3*, with correlation coefficients between 0.3636 and 0.5840. The results of the other diagnostic subgroups were similar. The virtually ubiquitous interconnectedness of genes was found to be very complex and could not be broken down in a straightforward manner.

Given this redundancy, it seems unlikely that there is a direct causal link between singular genes and psychiatric disorders since then several genes would have to overlap in their causal effects. Consequently, it must be assumed that singular genes with their illness-specific characteristics are secondary effects of a largely unspecific, genetically predisposed vulnerability, for example, in the sense of an elevated fragility in small genomic sections.

All classifiers were translated into SAS macros so that their performance could be successfully verified under SAS 9.4.

Results: Mental Health

Though most theoretical concepts of "mental health" can easily be grasped intuitively, their operationalization for NN analyses was difficult, because mental health cannot be modeled independently from somatic health, consumption behavior, and personality traits, amongst others. In particular, the straightforward approach that simply contrasts healthy controls from patients in a categorized way turned out to be an inadequate basis for modeling "healthiness" on the molecular-genetic level. On the other hand, with the controls (n=267) as target population and the patients (n=1,431) as control population, the NN analysis did indeed come up with genotypic patterns that exhibited unique characteristics in the target population, but the contribution of each significant gene to discrimination was quite small. Even with 18 genes, no more than 46% of subjects were correctly classified, while 54% were labeled as "unknown". No major contributor was identified, so that this "healthiness" model was not really promising in view of strengthening resilience among patients and controls.

Discussion

Unlike standard genotype-to-phenotype association methods with "psychiatric diagnosis" as phenotype [**25**, **26**], we focused our interest on "genetic diversity", that is, on the multitude of genotypic patterns observed with each gene in a given population. The basic assumption was that the genetic component underlying psychiatric disorders leaves distinct traces in the

patients' genotypic patterns, thus providing clues about the pathogenesis of these disorders [**27**, **28**]. Key elements of the proposed method were (1) the "gene vectors" assembled from 4-8 polymorphic SNPs located within genes and representing the genes' distinctive "fingerprints" in terms of the underlying genotypic patterns; (2) the genes' diversity indices defined through the number of different genotypic patterns observed with each gene; and (3) the quantification of correlations between genes. The method was found to offer a reliable framework for investigations into genetically complex population structures that emerge from the variation of genotypic patterns in genes and from the correlations between genes [**29**-**32**].

The evaluation of the diversity indices of the 100 specifically selected candidate genes resulted in a mean value as high as 109.4±82.8 for the studied Central European population, ranging from 18 to 476. Similarly unexpected was the finding that almost one-third of the candidate genes were correlated with each other, across diagnostic subgroups. Most intriguingly, highly significant deviations from "normal" diversity indices ($p<0.0001$) were detected for three diagnostic subgroups: (1) major depression (n=596), a significant decrease; (2) Alzheimer's disease (n=75), a significant decrease; and (3) schizoaffective disorders (n=64), a significant increase. These deviations were related to a small number of genes, while the majority of genes showed no such differences, thus suggesting that psychiatric disorders may indeed be related to irregularities in genetic diversity [e.g., **33**].

Detailed investigations into the observed irregularities revealed the existence of singular genes, that is, illness-specific genes for which certain genotypic patterns inherent in these genes showed up exclusively in patients, but not in healthy controls. For each of the diagnostic subgroups, we found between 13 and 30 singular genes, the respective numbers being independent of the sample sizes under investigation. When singular genes were combined, about 45% to 55% of patients in each diagnostic subgroup could be identified through distinctive genotypic patterns that did not appear in the healthy controls.

It is very unlikely that the detection of singular genes with their illness-specific characteristics was a purely random phenomenon, entirely due to methodological artifacts. It is equally unlikely that the singular genes were for the most part the result of population stratification, since half of the healthy controls were unaffected 1st-degree relatives of the study patients. Consequently, results apparently suggest that the patients of our study were characterized by an irregular, illness-specific genetic diversity, manifest in singular genes that

exhibited a variety of genotypic patterns not found in healthy controls.

Given the highly distinctive characteristics of singular genes, it is not really surprising that subsequent NN analyses, under consideration of the a priori information provided the singular genes, achieved a steady-state result of 80%-90% correctly classified subjects when comparing diagnostic subgroups with healthy controls. For the subgroups of patients with schizophrenic disorders, major depression, bipolar illness, and schizoaffective disorders, configurations of 6-10 genes separated from controls at a rate of about 90% correctly classified subjects: 4-5 core genes that were common to all classifiers, plus 2-5 accessory genes that depended on the target populations. The only exception with a 20% false-negative error rate was the subgroup of Alzheimer's disease patients. Evidently, genes of critical relevance to the Alzheimer's disease subgroup were missing.

It is unlikely that the successful separation between patients and healthy controls by means of NN classifiers was for the most part due to hidden, ethnicity-related population stratification, since half of the healthy controls were unaffected 1st-degree relatives of the study patients. We therefore believe that the findings of this study were diligently validated and deserve to be made known to the research community so that in-depth scientific discussions can follow, along with replications through independent patient samples.

Because of inter-correlatedness of genes, the classifiers constructed by the NN analyses were not unique. That is, considerable redundancy was involved. Because of this redundancy, a direct causal link between psychiatric disorders and the irregularities in genetic diversity is quite unlikely, since then several genes would exhibit the same causal effects, at least to a certain extent.

But what would be a possible interpretation if the findings are not mere methodological or data-inherent artifacts? In our eyes, the illness-specific and in singular genes manifested irregularities could be signs of a latent cross-diagnosis vulnerability that makes it easier for exogenous factors to trigger the onset of major psychiatric disorders, as well as to weaken the resilience of those affected. Here, resilience is understood as the counterpart to vulnerability, making it possible to succeed in daily life despite latent vulnerabilities.

The results clearly supported such a diagnosis-crossing, largely unspecific vulnerability, as the NN analyses revealed considerable overlaps on the genotype level between the various clinically defined diagnostic subgroups. All of this underlined the crucial role that diagnosis-crossing vulnerabilities and resilience factors may play in the context of psychiatric disorders.

Moreover, the overlap between the diagnostic subgroups may indicate that the clinically defined diagnoses do not necessarily represent biological entities.

Although the exact mechanisms behind this diagnosis-crossing vulnerability are still unknown, clinical applications are nevertheless immediately possible should the findings of this study be replicated. In fact, the said irregularities can be analyzed very easily, for example, by means of the SAS macros which we provide free of charge on request. This type of clinical application would undoubtedly facilitate the early detection of developing psychiatric disorders, since it may contribute with good reliability to the timely identification of at-risk cases. Thus, an early treatment can be started before clinically relevant symptoms develop.

Given the robustness of the results, our analyses can surely be replicated by independent patient samples with at most minor reductions in performance. Existing GWAs of psychiatric patients appear to be a good choice to evaluate genetic diversity without much effort in the proposed way. However, GWAs typically have relatively high error rates along with high percentages of missing data which might become an unmanageable obstacle for multidimensional methods [**34**]. Another problem could arise from the fact that the SNPs within genes cannot be freely selected, so that sufficiently high information contents for the resulting SNP combinations (genotypic patterns) are not necessarily given.

Conclusions

Our approach to quantifying genetic diversity through multidimensional gene vectors and diversity indices provided a powerful framework for investigations into genetically complex population structures, which emerge from the variation of genotypic patterns in genes and from the correlations between genes. Indeed, the proposed method of approach has the potential to make a significant contribution to the progress in psychiatry research.

The central finding of this study was the discovery of singular genes that, while not establishing a direct causal link between genotype level and psychiatric disorders, were quite amazing in their ability to separate patients from healthy controls. In each of the diagnostic subgroups under study, there were no less than 45%-55% of patients who exhibited genotypic patterns of such singular genes that did not at all show up in the healthy controls.

In the case of confirmation by independent research groups, clinical applications are readily possible and will facilitate the early detection of latent psychiatric disorders among

risk cases, so that early interventions can be started before clinically relevant symptoms develop. A larger number of hospitalizations could be prevented in this way.

Limitations

The vast majority of patients and controls were from Central Europe, so that the variation in biological ethnicity could only be expected to be modest. One must also assume that the classifiers constructed here will not necessarily show the same good performance with ethnically different populations. Another limiting factor is that some diagnostic subgroups were relatively small, which may affect the performance of the respective classifiers.

Acknowledgements

Funded in part by a grant of the Swiss National Science Foundation (32B0-119885).

Conflicts of Interest

The authors do not have any competing financial interests.

Supplementary information is available at MP's website.

References

1. Braun S, Bridler R, Müller N, Schwarz MJ, Seifritz E, Weisbrod M, et al. Inflammatory Processes and Schizophrenia: Two Independent Lines of Evidence from a Study of Twins Discordant and Concordant for Schizophrenic Disorders. Eur Arch Psychiatry Clin Neurosci. 2017; 267: 377-389.
2. Bleuler E. Dementia praecox or the group of schizophrenias. Translated from German by Joseph Zinkin, International Universities Press Inc., 8th Edition, New York 1969.
3. Stassen HH, Bachmann S, Bridler R, Cattapan K, Herzig D, Schneeberger A, et al. Inflammatory Processes linked to Major Depression & Schizophrenic Disorders and the Effects of Polypharmacy in Psychiatry: Evidence from a longitudinal Study of 279 Patients under Therapy. Eur Arch Psychiatry Clin Neurosci. 2021; 271(3): 507-520.
4. Wang Y, Wei J, Chen T, Yang X, Zhao L, Wang M, et al. A Whole Transcriptome Analysis in Peripheral Blood Suggests That Energy Metabolism and Inflammation Are Involved in Major Depressive Disorder. Front Psychiatry. 2022; 13: 907034.
5. Dennison CA, Legge SE, Pardiñas AF, Walters JTR. Genome-wide association studies in schizophrenia: Recent advances, challenges and future perspective. Schizophr Res. 2020; 217: 4-12.
6. Legge SE, Santoro ML, Periyasamy S, Okewole A, Arsalan A, Kowalec K. Genetic architecture of schizophrenia: a review of major advancements. Psychol Med. 2021; 51(13): 2168-2177.
7. Levey DF, Stein MB, Wendt FR, Pathak GA, Zhou H, Aslan M, et al. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. Nat Neurosci. 2021; 24(7): 954-963.

8. Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019; 22(3): 343-352.
9. Gordovez FJA, McMahon FJ. The genetics of bipolar disorder. Mol Psychiatry. 2020; 25(3): 544-559.
10. Berger M, Stassen HH, Köhler K, Krane V, Mönks D, Wanner C, et al. Hidden population substructures in an apparently homogeneous population bias association studies. Eur J Hum Genetics. 2006; 14: 236-244.
11. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010; 11(7): 459–463.
12. Shi H, Gazal S, Kanai M, Koch EM, Schoech AP, Siewert KM, et al. Population-specific causal disease effect sizes in functionally important regions impacted by selection. Nat Commun. 2021; 12(1): 1098.
13. Stassen HH, Bridler R, Hägele S, Hergersberg M, Mehmann B, Schinzel A, et al. Schizophrenia and smoking: evidence for a common neurobiological basis? Am J Med Genetics. 2000; 96: 173-177.
14. Weisbrod M, Hill H, Sauer H, Niethammer R, Guggenbühl S, Stassen HH. Nongenetic pathologic developments of brain-wave patterns in monozygotic twins discordant and concordant for schizophrenia. Am J Med Genetics. 2004; 125: 1-9.
15. Stassen HH, Scharfetter C, Angst J. Functional Psychoses —Molecular-genetic Evidence for a Continuum. In: A. Marneros and H.S. Akiskal (eds) The overlap of affective and schizophrenic spectra. Cambridge University Press 2006; pp. 55-78.
16. Stassen HH, Angst J, Hell D, Scharfetter C, Szegedi A. Is there a common resilience mechanism underlying antidepressant drug response? Evidence from 2848 patients. J Clin Psychiatry. 2007; 68(8): 1195-1205.
17. Stassen HH, Anghelescu IG, Angst J, Böker H, Lötscher K, Rujescu D, et al. Predicting Response to Psychopharmacological Treatment. Survey of Recent Results. Pharmacopsychiatry. 2011; 44: 263-272.
18. Stassen HH, Bachmann S, Bridler R, Cattapan K, Herzig D, Schneeberger A, et al. Detailing the Effects of Polypharmacy in Psychiatry: Longitudinal Study of 320 Patients Hospitalized for Depression or Schizophrenia. Eur Arch Psychiatry Clin Neurosci. 2022; 272(4): 603-619.
19. Endicott J, Spitzer RL. A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia (SADS). Arch Gen Psychiatry. 1978; 35: 837-844.
20. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Symptom Scale (PANSS) for Schizophrenia. Schiz Bull. 1987; 13: 261-276.
21. Hamilton M. A Rating Scale for Depression (HAM-D). J Neurosurg Psychiat. 1960; 23: 56-62.
22. Kuny S. Medication and Side Effects (MEDIS) among Psychiatric Patients under Treatment. Psychiatric University Hospital Zurich, 1988: available on request.
23. Kuny S, Stassen HH. The Zurich Health Questionnaire (ZQH). Quantifying “regular exercises”, “consumption behavior”, “impaired physical health”, “psychosomatic disturbances” and “mental health” in the general population through 63 items. Psychiatric University Hospital Zurich, 1988: available in 6 major languages on request.
24. Hecht-Nielsen, R. Theory of Backpropagation Neural Network. In: Proceedings of the International Joint Conference on Neural Networks. IEEE 1989; 1: 593-611.
25. Horwitz T, Lam K, Chen Y, Xia Y, Liu C. A decade in psychiatric GWAS research. Mol Psychiatry. 2019; 24(3): 378-389.

584 26. Unal-Aydin P, Aydin O, Arslan A. Genetic Architecture of Depression: Where Do We
585 Stand Now? Adv Exp Med Biol. 2021; 1305: 203-230.
586 27. Kendler KS. A joint history of the nature of genetic variation and the nature of
587 schizophrenia. Mol Psychiatry. 2015; 20(1): 77-83.
588 28. Smeland OB, Frei O, Dale AM, Andreassen OA. The polygenic architecture of
589 schizophrenia - rethinking pathogenesis and nosology. Nat Rev Neurol. 2020; 16(7): 366-
590 379.
591 29. McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine Learning for Detecting Gene-
592 Gene Interactions. Appl Bioinformatics. 2006; 5(2): 77–88.
593 30. Boucher B, Jenna S. Genetic interaction networks: better understand to better predict.
594 Front Genet. 2013; 4: 290.
595 31. Moore JH, Mackay TFC, Williams SM. Testing the assumptions of parametric linear
596 models: the need for biological data mining in disciplines such as human genetics.
597 BioData Min. 2019; 12: 6.
598 32. Chimusa ER, Defo J. Dissecting Meta-Analysis in GWAS Era: Bayesian Framework for
599 Gene/Subnetwork-Specific Meta-Analysis. Front Genet. 2022; 13: 838518.
600 33. Talarico F, Costa GO, Ota VK, Santoro ML, Noto C, Gadelha A, et al. Systems-Level
601 Analysis of Genetic Variants Reveals Functional and Spatiotemporal Context in
602 Treatment-resistant Schizophrenia. Mol Neurobiol. 2022; 59(5): 3170-3182.
603 34. Dattani S, Howard DM, Lewis CM, Sham PC. Clarifying the causes of consistent and
604 inconsistent findings in genetics. Genet Epidemiol. 2022; [online ahead of print]
605

606 **Table 1**: The «Zurich Molecular-Genetic Study of Psychiatric Vulnerability» encompasses 2,008 patients
607 hospitalized for major psychiatric disorders along with 464 healthy controls. For this project, 1,698 subjects
608 were genotyped for 100 specifically selected genes and 549 polymorphic SNPs located within these genes.

609 **Table 2:** Expected values regarding diversity indices for 10 genes and sample sizes ranging from 100 to 1,000.
610 Due to the well-behaved characteristics of the underlying calibration curves, simple linear interpolation
611 between the sampling points is sufficient to calculate indices for intermediate sample sizes.

612 **Table 3:** «Singular genes» denote illness-specific genes for which genotypic patterns inherent in these genes
613 show up exclusively in patients, but not in healthy controls. For each diagnostic subgroup, we found some 13-
614 30 singular genes with frequencies between 10.0% and 36.4%. Weakening the clear-cut definition of
615 "healthiness" for the control population (n=267) by extending it with the 201 patients of our sample without
616 severe psychiatric diagnoses (n=468) left the results essentially unchanged. Only the number of singular genes
617 reaching significance dropped somewhat in each diagnostic subgroup.

618 **Table 4:** For four target populations, we found in comparisons with health controls a rate of about 90%
619 correctly classified patients along with a 10% subgroup labeled as "unknown". The only exception was the
620 subgroup of patients with "Alzheimer's disease" where apparently one or more genes of relevance were
621 missing in the selection of candidate genes.

622 **Table 5:** Classifier genes have been identified by the NN algorithm as contributing to the separation between
623 the diagnostic subgroups and healthy controls. All genetic analyses relied on a genetic-physical map derived
624 from *Ensembl* Build 105 of September 25, 2021.

625 **Figure 1**: Principal schema of a neural net model where multiple genes and clinical diagnosis (affectedness) are
626 connected to each other by complex interactions.

627 **Figure 2**: Distribution of the diversity indices of 100 genes as observed in 1,698 Central European subjects
628 (including a small number of U.S. Americans). The diversity index ranged from 18 (*CYP2C19*) to 476 (*GPR39*)
629 with a mean value of 109.4 ± 82.8. The distribution revealed two peaks (diversity indices around 70 and 170),
630 along with 7 genes exhibiting a diversity index above 250. These results may indicate different types of genes.

631 **Figure 3**: Diversity index as a function of sample size, with sample sizes ranging from 50 to 1,700. Upper half:
632 gene *CYP2J2* on chromosome 1 with **diversity index=69**. *CYP2J2* belongs to the left group of genes in Figure 2.
633 Lower half: gene SLC6A6 on chromosome 3 with diversity index=182. SLC6A6 belongs to the middle group of
634 genes in Figure 2. All genetic analyses relied on a genetic-physical map derived from *Ensembl* Build 105 of
635 September 25, 2021.

636 **Figure 4**: The distributions of the genotypic patterns of the genes under study showed no substantial
637 differences between healthy controls (Distribution "D") and the patients of the 5 diagnostic subgroups. For
638 example, distribution "A" relates to the diagnostic subgroup "Schizophrenia" and distribution "B" to the
639 diagnostic subgroup "Depression". By contrast, distribution of the Alzheimer's subgroup ("C") exhibited
640 significant deviations from the other ones ("A", "B"," D").

641

Figures

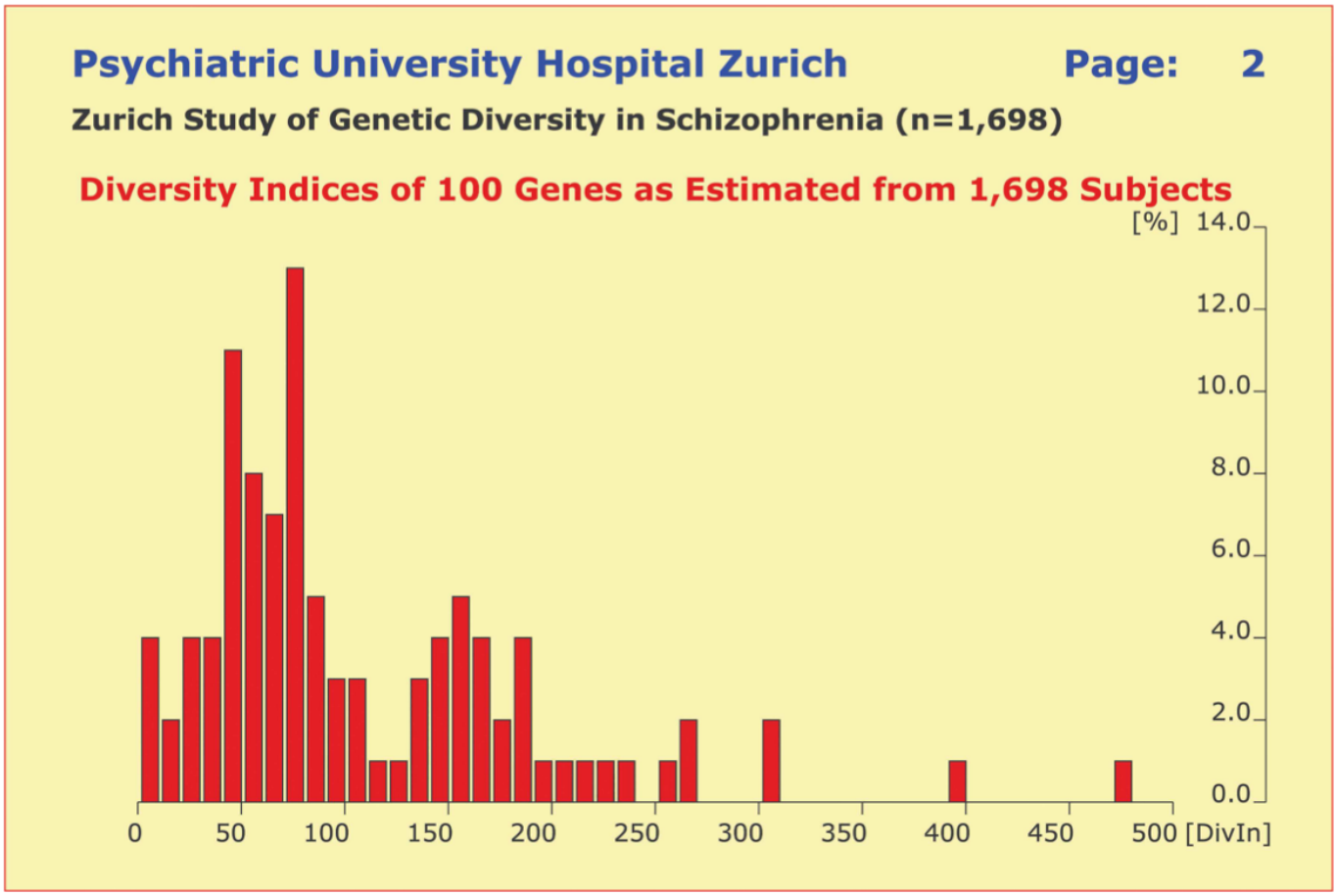

Figure 1

Figure 2

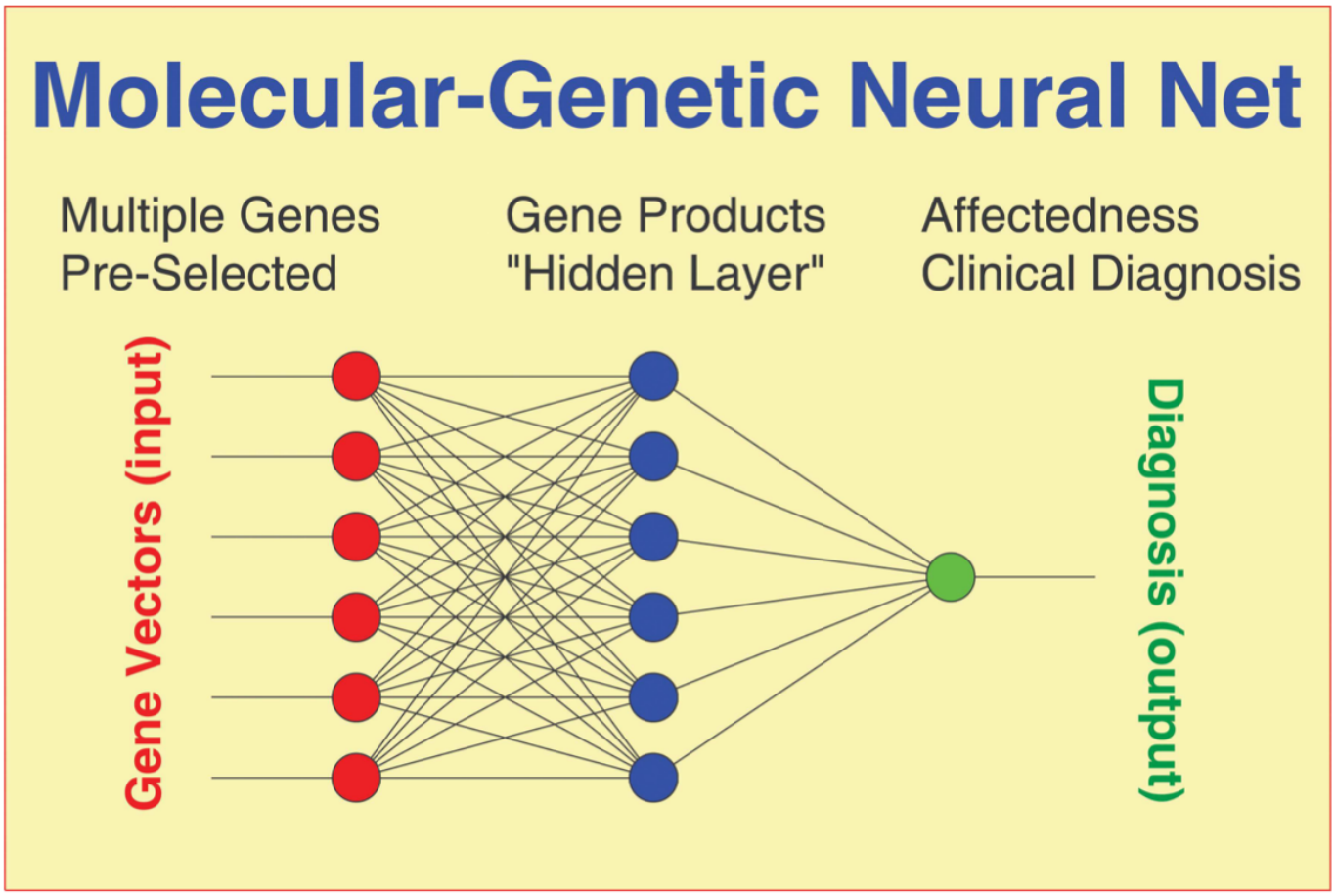

Figure 2

Figure 1

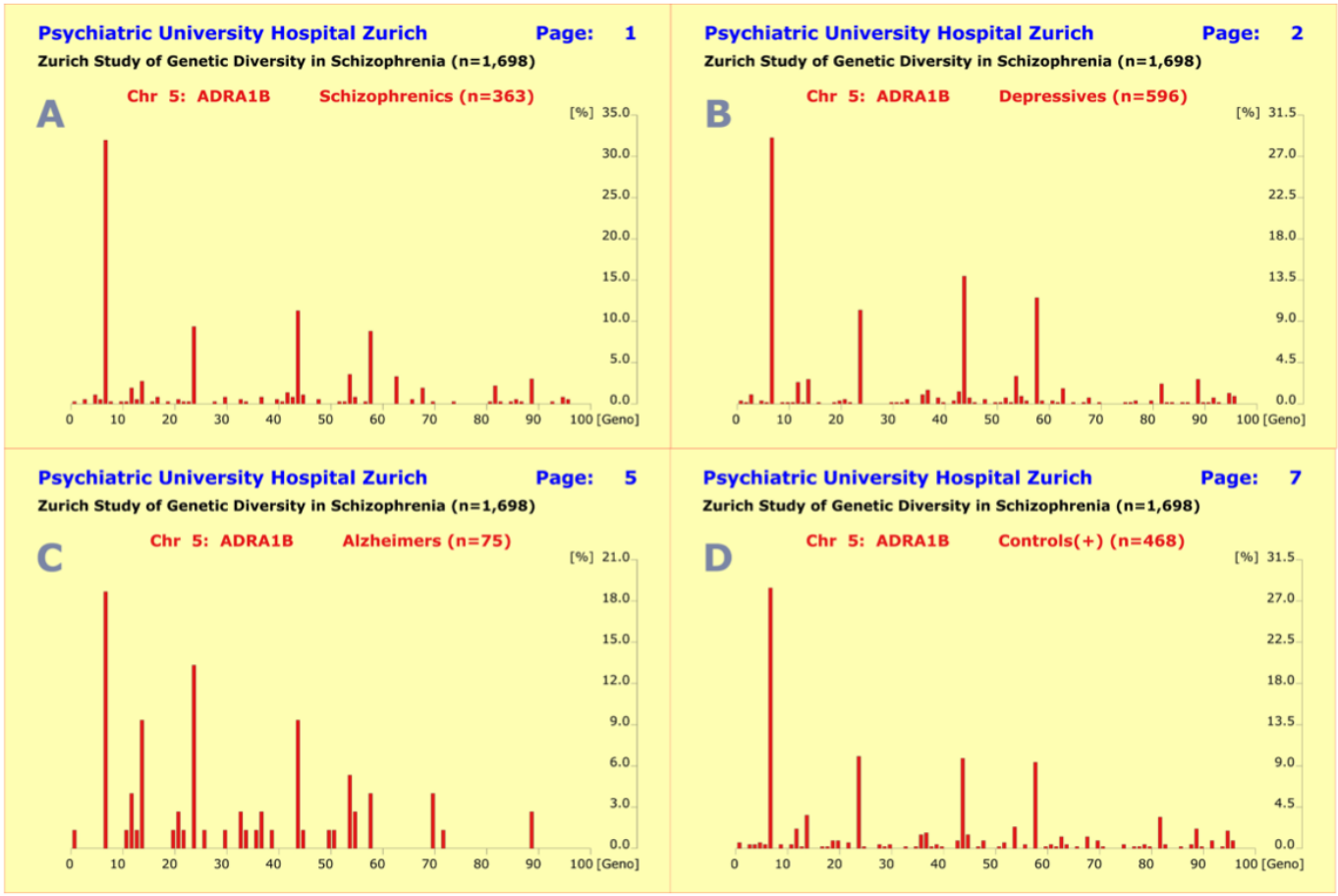

Figure 3

Figure 4

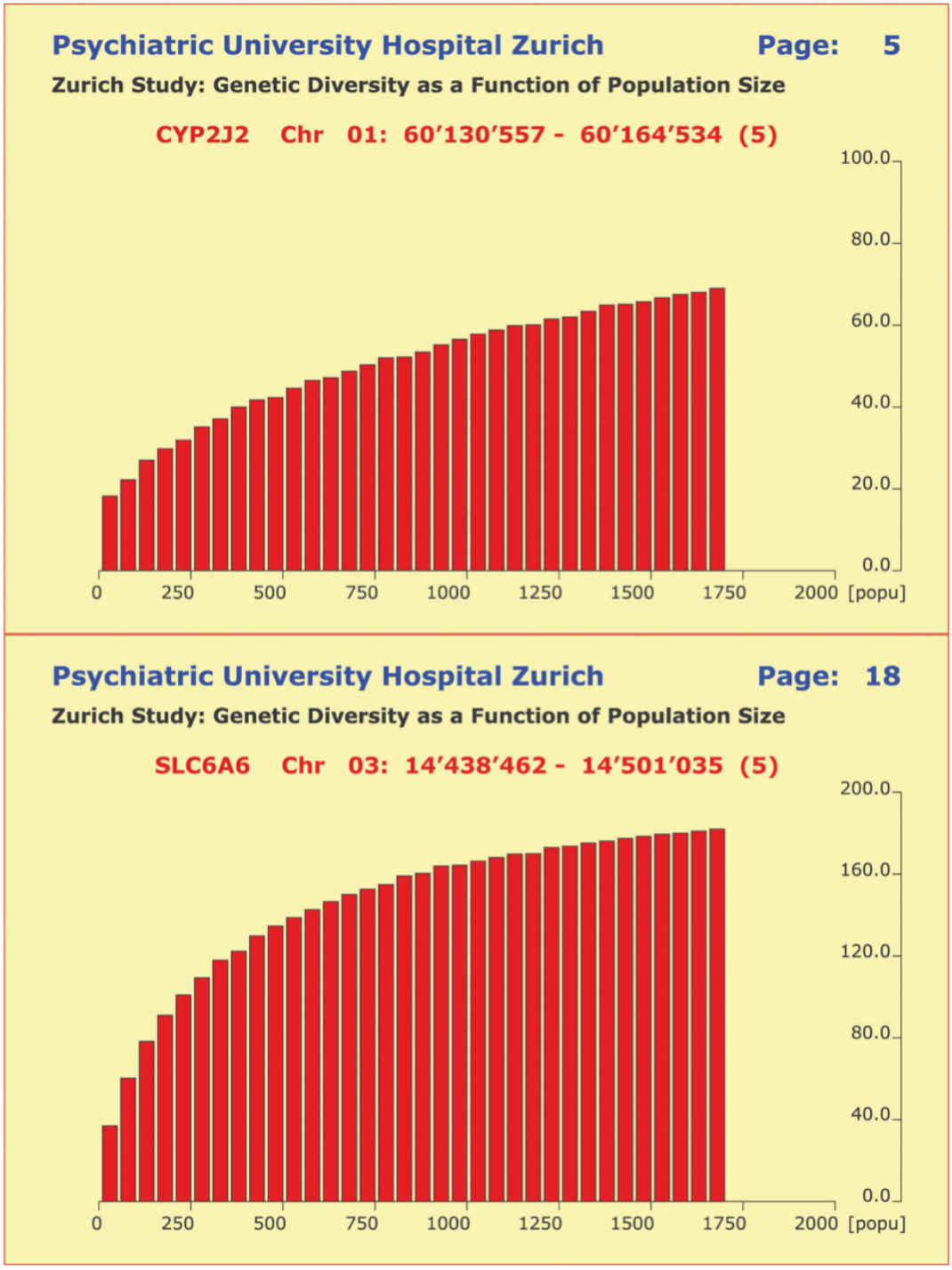

Figure 4

Figure 3

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table2.jpg
- Table4.jpg
- Table3.jpg
- SupplementaryTable6a.jpg
- Table1.jpg
- SupplementaryTable6b.jpg
- Table5.jpg
- SupplementaryTable6a.jpg
- SupplementaryTable6b.jpg