Check for updates

ARTICLE   OPEN

Metabolic signature of the pathogenic 22q11.2 deletion identifies carriers and provides insight into systemic dysregulation

Julie Courraud[1,2,3,4,11], Francesco Russo[1,11], Gonçalo Espregueira Themudo[2,5,6,7], Susan Svane Laursen[1,2], Andrés Ingason[2,8], David M. Hougaard[1,2], Arieh S. Cohen[1,2], Thomas Werge[2,5,9,10✉] and Madeleine Ernst[1,2✉]

Large deletions at chromosome 22q11.2 are known to cause severe clinical conditions collectively known as 22q11.2 deletion syndrome. Notwithstanding the pathogenicity of these deletions, affected individuals are typically diagnosed in late childhood or early adolescence, and little is known of the molecular signaling cascades and biological consequences immediately downstream of the deleted genes. Here, we used targeted metabolomics to compare neonatal dried blood spot samples from 203 individuals clinically identified as carriers of a deletion at chromosome 22q11.2 with 203 unaffected individuals. A total of 173 metabolites were successfully identified and used to inform on systemic dysregulation caused by the genomic lesion and to discriminate carriers from non-carriers. We found 84 metabolites to be differentially abundant between carriers and non-carriers of the 22q11.2 deletion. A predictive model based on all 173 metabolites achieved high Accuracy (89%), Area Under the Curve (93%), F1 (88%), Positive Predictive Value (94%), and Negative Predictive Value (84%) with tyrosine and proline having the highest individual contributions to the model as well as the highest interaction strength. Targeted metabolomics provides insight into the molecular consequences possibly contributing to the pathology underlying the clinical manifestations of the 22q11 deletion and is an easily applicable approach to first-pass screening for carrier status of the 22q11 to prompt subsequent verification of the genomic diagnosis.

INTRODUCTION

22q11.2 deletion syndrome (22q11.2DS) was first reported in 1965 by Dr Angelo DiGeorge and correspondingly named DiGeorge syndrome [1]. It is the most common chromosomal microdeletion syndrome, with an estimated incidence of 1 over 3672 live births [2]. Deletions of various sizes are gathered under the 22q11.2DS name, but the most common is a 3MB deletion in between two low copy repeats (LCRs) zones (LCR22A and LCR22D).

Irrespective of deletion size, 22q11.2DS comes with heterogeneous clinical presentation, including several very severe conditions [3]. None of the clinical presentations are specific to 22q11.2DS, and far from all carriers are clinically ascertained. In fact, a considerable proportion of 22q11.2DS (~10%) is inherited, typically from clinically un-affected or non-recognized parents carrying the deletion [3], and has been reported to be more severely affected than de-novo cases. Among the clinical presentations detected in utero or at birth, congenital heart

defects are the most frequent, followed by velopharyngeal insufficiency, cleft palate, or dysmorphic craniofacial features. Many disorders will only be detectable or develop later in life, such as hypocalcemia due to primary hypoparathyroidism [4], as well as developmental disabilities, mental disorders [5, 6], and life-threatening severe chronic immune deficiency [7]. While the syndrome is not curable, many of the clinical manifestations can be improved if treated in a timely manner.

On average, 22q11.2DS is diagnosed at 9–13 years, and children with 22q11.2DS often go through a diagnostic odyssey before receiving diagnosis and appropriate care [3, 6]. In fact, due to the heterogeneous clinical presentation of 22q11.2DS, patients meet, on average, seven experts, including psychiatrists, pediatricians, surgeons, general practitioners, psychologists, and clinical geneticists, before being oriented toward molecular diagnostics [8].

As the diagnosis of 22q11.2DS can only be confirmed genetically, developing techniques for early clinical diagnosis of

Table 1. Characteristics of the study cohort.

	Cases (n = 203)			Controls (n = 203)		
	median	range	NA	median	range	NA
Age at sampling (days)	5	[2–40]	19	5	[2–30]	19
Birth weight (g)	3130	[1200–5550]	9	3400	[830–5115]	7
Gestational age (weeks)	39	[32–42]	106	39	[33–42]	96
Month of birth*	7	[1–12]	0	7	[1–12]	0
Age of mothers at birth (years)	29.6	[18.0–42.4]		29.5	[16.6–41.3]	

NA not available data.
*January = 1 and December = 12.

22q11.2DS that can be integrated into existing screening programs is essential. While it has been suggested that 22q11.2DS could be detected during neonatal metabolic screening, which is standard in most countries [9], such assays are not yet available. Consequently, finding biomarkers specific to this syndrome that are reliably detectable in neonatal dried blood spots (DBS) could greatly help the diagnosis process, leading to earlier appropriate care. In fact, *PRODH*, the gene encoding for proline dehydrogenase and among the genes deleted in 22q11.2DS, has been reported to result in hyperprolinaemia [10–12], suggesting that patients with hyperprolinaemia should be screened for 22q11.2DS [11].

In this work, we investigate for the first time a large cohort of individuals with the 22q11.2 deletion using targeted metabolomics of neonatal DBS to inform on the molecular consequence of a large genomic lesion and to probe metabolomics screening as a diagnostic tool for the early identification of individuals carrying the 22q11.2 deletion.

MATERIALS AND METHODS

Materials
Methanol (MeOH), acetonitrile (ACN), isopropanol (IPA), water (H$_2$O), and formic acid (FA) were of Optima™ LCMS-grade and were purchased from Thermo Fisher Scientific (Waltham, MA, USA).

Study cohort
We acquired metabolic profiles of residual dried blood spots (DBS) from 406 children born between 1983 and 2012 (median 2000) collected at a median age of 5 days after birth (range [2–40] days) and stored at the Danish National Biobank at −20 °C [13]. Cases were identified through the Danish Cytogenetic Central Registry (DCCR) as children later diagnosed with 22q11.2DS (102 females, 101 males). Controls were selected at random from the collection of population-wide neonatal dried blood spots stored at the Danish National Biobank and matched based on sex and date of birth. Most individuals were carriers of the 3MB deletion (n = 166) versus 11 and 5 individuals being carriers of the 1.5MB and 1MB deletions, respectively. An overview of the study cohort is shown in Table 1.

Data acquisition
We measured absolute concentrations of a total of 408 compounds in the DBS using the AbsoluteIDQ p400 kit (Biocrates Life Science Ag., Innsbruck, Austria) following the manufacturer's instructions. Metabolites from DBS were extracted according to the standard operation procedure specific to DBS provided by Biocrates. All instrumentation and preprocessing software were from Thermo Scientific (Waltham, MA, USA). Our liquid chromatography-tandem mass spectrometry (LC–MS/MS) platform consisted of an autosampler CTC Combi PAL HTS TMO (two injectors), an LC system (LX-2 with two Ultimate 3000 Dionex RS pumps and a Transcend II Valve Interface Module) and a Q-Exactive Orbitrap mass spectrometer with a HESI-II probe. Case and control sample pairs were randomized into six 96-well analytical plates, and data was acquired from May to June 2018 with no more than 2–5 days between each plate. Each sample was injected four times as instructed: two times for LC–MS and two times for flow injection analysis (FIA) data acquisition. LC–MS and FIA-based data were acquired on

a dedicated injector and pump to limit technical variation. Each plate included a solvent blank, three paper blanks (blank filter paper on which dried blood spots are collected), a 7-point calibration curve (injected for LC–MS data acquisition only), and quality control (QC) samples provided by Biocrates (three concentration levels QC1–3). QC1 and QC3 were injected once, while QC2 was injected five times across the plate among the experimental samples. Four case–control pairs had to be replicated on the last plate as technical difficulties occurred during extraction; we, therefore, kept only the results from the second analysis on plate 6.

Quality control and data processing
LC–MS-based data were preprocessed as instructed by Biocrates using Xcalibur (v4.1.31.9) for peak integration and after thorough optimization and manual check. The exported feature table was then imported into MetIDQ Carbon-2793 (Biocrates) along with FIA-based raw data for plate validation and concentration calculation. Concentrations were normalized based on QC2 target values in MetIDQ to reduce the batch effect. All measurements were then exported along with their individual status computed by MetIDQ (e.g., "Valid"). For further quality control and preprocessing, we used the MeTaQuaC R package v0.1.32.9001 [14]. All measurements with a MetIDQ status other than "Valid" were replaced by missing values, and only compounds detected in at least 30% of the Biocrates' QC level 2 samples were kept in the analysis. Furthermore, compounds with 100% missing values in experimental samples and compounds that had >80% missing values in both case and control samples were removed. Finally, we normalized the absolute concentrations of the batch through centering by subtracting the column means (omitting NAs) of each batch and scaling by the standard deviation. The final data table consisted of 173 compounds (33 out of 42 and 140 out of 366 for the LC and FIA injections, respectively, Supplementary Fig. 1) and 406 samples (203 cases and controls, respectively).

Statistical analyses
To initially contrast quantities of the analyzed metabolites between 22q11.2 carriers and controls, we performed a differential abundance analysis by applying a paired Wilcoxon signed-rank test using the R function Wilcox.test from the stats R package [15]. Subsequently, we used the function p.adjust (from the stats R package) to calculate the False Discovery Rate (FDR) adjusted p-values. Metabolites having FDR-adjusted p-value < 0.05 were considered differentially abundant.

Predictive models and feature importance were built by training an eXtreme Gradient Boosting (XGBoost) model using the R packages xgboost (version 1.4.1.1) [16] and caret (version 6.0–88) [17]. We divided the dataset into train and test sets (80% train and 20% test). Firstly, we performed an analysis including all the 173 metabolites, without tuning and using the following recommended parameters: nrounds = 500, max_depth = 6, colsample_bytree = 1, eta = 0.3, subsample = 1, gamma = 0, min_child_-weight = 1. Secondly, we improved the performance of the model with all the 173 metabolites by applying a tenfold cross-validation. The final model fitted the following tuned parameters on the full training set: nrounds = 1400, max_depth = 4, eta = 0.025, gamma = 0.5, colsample_bytree = 0.8, min_child_weight = 1, subsample = 0.5. Prediction on the test set and evaluation of performance was performed, and the R package pROC (version 1.18.0) [18] was used to calculate the area under the curve (AUC) and plot the receiver operating characteristic (ROC) curve. In addition to AUC and accuracy, we provide F1, positive, and negative predictive values calculated using the caret function confusionMatrix.

To discover which metabolites contributed most to the prediction of 22q11.2DS, we computed feature importance using the caret function varImp. To further validate the importance of features, we performed additional analyses. We measured how important each feature was for the predictions by applying the R function FeatureImp (iml package, version 0.10.1 [19]), which implements the method described by Fisher and coworkers [20]. The method works by shuffling each feature and measuring how much the performance drops according to the prediction loss/error (in our case, the classification error). The prediction error is measured before and after shuffling the values of the feature and the larger the increase of the error is, the more important the feature. This process was repeated 1000 times (n.repetitions parameter set to 1000), since the higher the number of repetitions, the more stable and accurate the results become.

We then explored how strongly features interact with each other in the prediction model by applying the R function Interaction from the iml package, which measures interactions through Friedman's H-statistic [21]. The interaction measure shows how much of the variance of each feature is explained by the interaction, with values ranging from 0 (no interaction) to 1.

To evaluate the predictive value of the most important selected metabolites, we compared the results of four XGBoost models without tuning the models' parameters (using default/recommended settings for all models). The first model included the top two most important metabolites (tyrosine and proline), the second model only the first most important metabolite (tyrosine), and the third model only the second most important metabolite (proline).

All statistical analyses were performed in R [15] (v4.1.1), and scripts and Jupyter notebooks are publicly available at: https://github.com/SSI-Metabolomics/22q11_SupplementaryMaterial/.

RESULTS

To investigate the potential of targeted metabolomics as a diagnostic tool for 22q11.2DS in neonatal screening, we performed broad metabolic profiling using a validated commercial kit (see "Methods" for details) of neonatal blood spots from 203 individuals identified as carriers of a deletion at chr. 22q11.2 and 203 age- and sex-matched individuals without the deletion. After quality control, data on 173 compounds were obtained and used for subsequent analyses (see Methods for details). As our focus was on assessing whether a metabolic marker of 22q11.2DS could be found in a potential clinical setting, we only included compounds that passed stringent quality control measures as

defined by Biocrates. Many acylcarnitines, but also some cholesterol esters glycerides, glycerophospholipids, and sphingolipids, fell below the lower limit of quantification. However, despite these limitations, we still achieved reasonable coverage of most metabolite classes (Supplementary Fig. 1).

To initially contrast quantities of the analyzed metabolites between 22q11.2 carriers and controls, we performed a differential abundance analysis of the 173 metabolites that passed QC (see "Methods" for details) by applying a paired Wilcoxon signed-rank test. In total, 84 metabolites were differentially abundant (FDR-adjusted p-value < 0.05). Differentially abundant metabolites were found in all measured compound classes, most predominantly across amino acids, sphingolipids, biogenic amines, and glycerophospholipids (Supplementary Fig. 1). A majority of metabolites were decreased in 22q11.2 carriers when compared to controls (Supplementary Fig. 2). All cholesterol esters and glycerides were decreased in 22q11.2 carriers and also a majority of amino acids, glycerophospholipids, sphingolipids, and biogenic amines, with the exception of one amino acid (proline), one biogenic amine (creatinine), five glycerophospholipids (LPC(16:0), PC(32:1), PC-O(36:2), PC-O(34:1), PC-O(34:0)) and four sphingolipids (SM(40:1), Cer(42:2), Cer(42:1), Cer(40:1)). Interestingly all measured acylcarnitines (hexadecanoylcarnitine, AC(16:0); octadecanoylcarnitine, AC(18:0)) were increased in 22q11.2 carriers when compared to controls (Supplementary Fig. 2). Tyrosine was the most significant differentially abundant metabolite (FDR-adjusted p-value = 7.53e −15; Fig. 1) with a highly significant lower abundance in cases versus controls. Also, proline showed a significant difference in abundance (FDR-adjusted p-value = 2.25e−03) with a slightly higher abundance in cases compared to controls. The full list of significant differentially abundant metabolites can be found in Supplementary Data 1 and stratified by compound class in Supplementary Fig. 2.

To further investigate the metabolic signature of individuals carrying the 22q11.2 deletion, we first determined the combined ability of the 173 analyzed metabolites to classify the 406 22q11.2 case–control individuals in a predictive model and next determined the individual contributions of each metabolite to the resulting model. As shown in Fig. 2 and Table 2, the model achieved high Accuracy (89%), AUC (93%), F1 (88%), Positive predictive value (94%) and Negative predictive value (84%). The
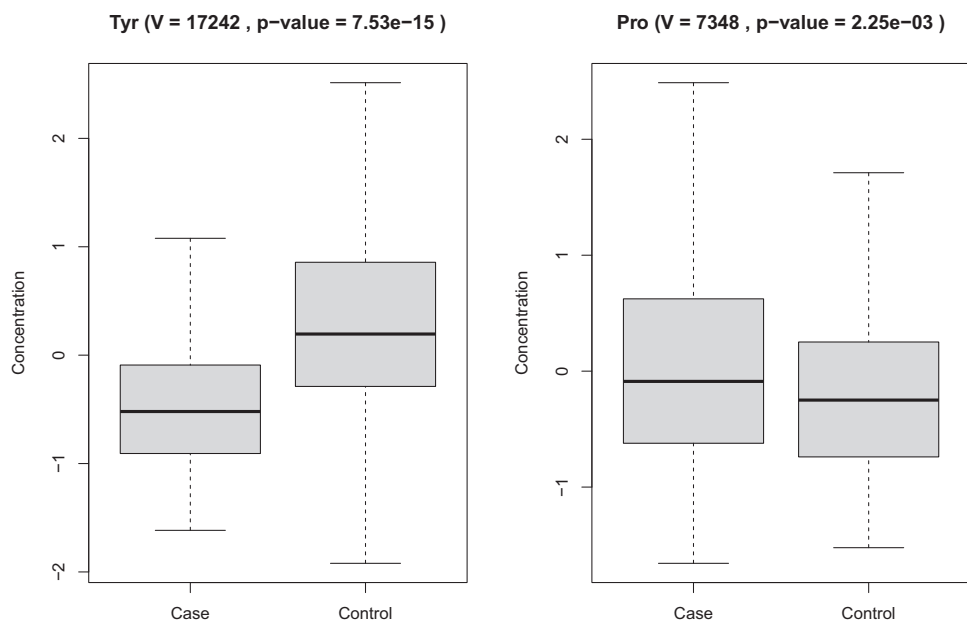

Fig. 1 Differential abundance of tyrosine and proline. Using the Wilcoxon signed-rank test showing the comparison of cases versus controls for tyrosine (left) and proline (right). Box plots including V-statistics (V value) and FDR-adjusted p-value.

4

individual contributions to the final model of the top 20 metabolites are shown in Fig. 3, with tyrosine and proline ranked as first and second. This finding was confirmed in a subsequent analysis, estimating the loss of performance, i.e., the classification error, as a function of shuffling measurements between metabolites [22].

Next, we examined overall interactions between the analyzed metabolites and found tyrosine and proline to be the two metabolites with the highest interaction with other metabolites based on Friedman's *H*-statistic [21]. Furthermore, tyrosine and proline were among the strongest interactors of each other, as shown in the interaction network of proline (Fig. 4).

To contrast the individual and combined predictive values of proline and tyrosine with all the other metabolites combined, we compared the results of four XGBoost models (see "Methods" for details), considering (1) all 173 metabolites, (2) proline and tyrosine, (3) only tyrosine, and (4) only proline. As detailed in Table 2, these analyses show that (i) tyrosine outperforms proline on all five metrics relative to the prediction of 22q11.2 status, (ii) combining tyrosine and proline improves prediction somewhat relative to tyrosine alone, while (iii) all metabolites combined increases prediction even further (Table 2).
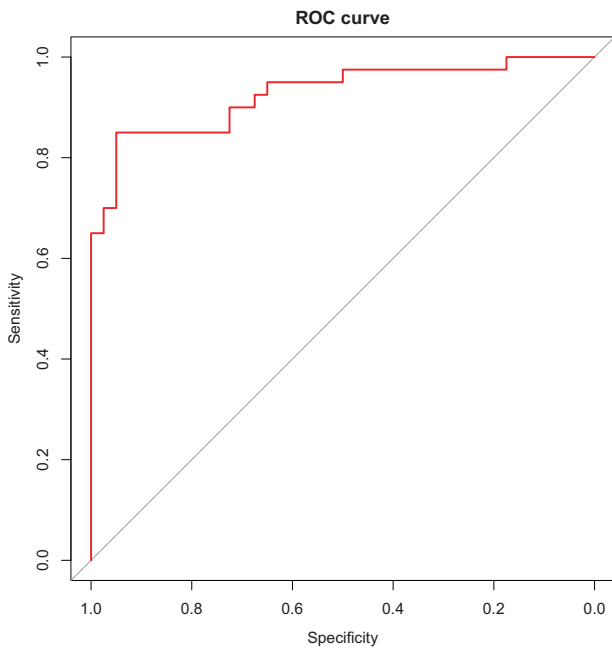
Fig. 2 ROC curve for the predictive model. Receiver operating characteristic (ROC) curve shows the performance of the final classification model, including a total of 173 metabolites (tenfold CV).

DISCUSSION

In this work, we explored the potential of targeted metabolomics to inform on molecular consequences of the pathogenic deletion at chromosome 22q11.2 and to serve as a neonatal diagnostic screening tool to identify individuals carrying the copy number variant. We found that 22q11.2 deletion was associated with significant alterations in levels of metabolites in whole blood at the time of birth, which may reflect metabolic alterations in other organ systems, such as the brain. Furthermore, we used machine learning techniques and differential abundance analysis to document that the metabolite profiles discriminate 22q11.2 deletion carriers from non-carriers and identify the amino acids tyrosine and proline as the most dysregulated metabolites.

The increased levels of proline observed in this study is consistent with decreased proline-to-glutamate conversion due to hemizygosity of proline dehydrogenase 1 (*PRODH*) in 22q11.2 deletion carriers and generalized case reports on individuals with the 22q11.2 deletion [10–12, 23, 24]. Notably, glutamate, along with another 14 amino acids, was found significantly decreased in 22q11.2 deletion carriers (Supplementary Fig. 2). Increased levels of proline in carriers of the 22q11.2 deletion have also been reported in previous metabolomics studies in plasma and dried blood spots of older children carrying the 22q11.2 deletion as well as studies reporting on hyperprolineamia in patients with 22q11.2DS [10, 11, 23, 24]. In contrast, the observation of reduced levels of tyrosine is novel and may derive from feedback inhibition of phenylalanine hydroxylase mediated conversion of dietary phenylalanine into tyrosine. Such feedback inhibition could be due to accumulating levels of catecholamines resulting from hemizygosity of the catechol-O-methyltransferase encoding *COMT*-gene in 22q11.2 deletion carriers [25].

Our results largely corroborate with findings from two previous studies describing significantly altered metabolomic profiles in plasma and dried blood spots of children carrying the 22q11.2 deletion versus controls [23, 24]. Our study includes the largest sample size up to date and is the first study to report significantly altered metabolomic profiles in prediagnostic samples collected a few days after birth. We found differentially abundant metabolites across all measured compound classes, most predominantly across amino acids, sphingolipids, biogenic amines, and glycerophospholipids (Supplementary Fig. 1). Sphingolipids have previously been reported to be significantly altered in brain tissue from Df(16)A ± mice, a model of the 22q11.2 deletion syndrome [26] and acylcarnitines and glycerides were previously found among metabolites that distinguish children with 22q11.2DS from controls [24]. Five amino acids (proline, histidine, tryptophan, threonine, and serine) and one biogenic amine (methionine sulfoxide) were previously found differentially abundant across plasma from children (age 8–15 years) carrying the 22q11.2 deletion versus age- and sex-matched typically developing controls [23]. In contrast to our findings, however, Napoli and collaborators [23] found that histidine, tryptophan, threonine,

Table 2. Performance comparison across models.

	All metabolites No tuning	Tyrosine and Proline No tuning	Tyrosine No tuning	Proline No tuning	All metabolites Parameter tuning
Positive predictive value	0.865	0.707	0.727	0.429	0.943
Negative predictive value	0.814	0.718	0.66	0.444	0.844
F1	0.831	0.716	0.657	0.400	0.853
Accuracy	0.837	0.712	0.687	0.437	0.887
AUC	0.897	0.753	0.732	0.509	0.927

We built five models with (1) all 173 metabolites, (2) only tyrosine and proline, (3) only tyrosine, and (4) only proline. Additionally, we show the results of the tuned parameters (tenfold CV).
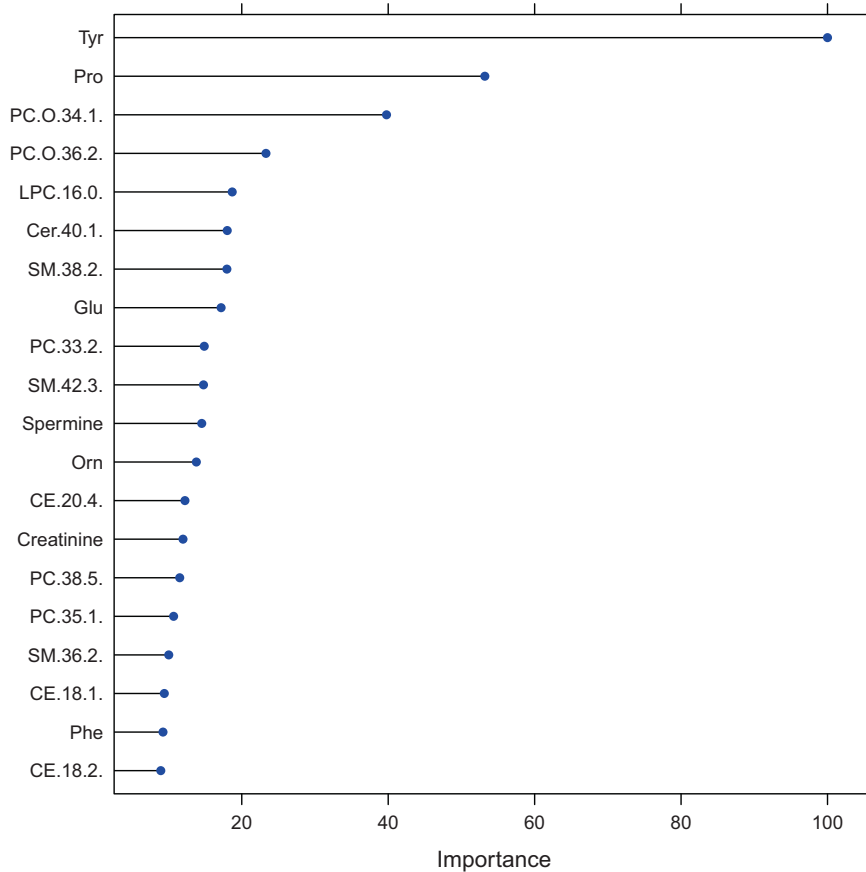
Fig. 3 Top-20 metabolites ranked by their importance. The individual contributions to the final model of the top-20 metabolites based on the caret R function varImp, reporting a scaled measure of importance. The measure is equivalent to the Gain value (the improvement in accuracy brought by a feature to the branches it is on). The higher the value is, the more important the feature is.

serine, and methionine sulfoxide were more highly abundant in carriers of the deletion versus controls.

As demonstrated in this study, the capacity of targeted metabolomics to discriminate between individuals carrying the 22q11.2 deletion from non-carriers points to great potential for the clinical applicability of this approach in neonatal screening, which would ensure early identification and corresponding timely and proper healthcare provision and socio-educational support. Recent findings based on population-representative studies in Denmark documented that a considerable fraction of 22q11.2 deletion carriers go unnoticed in the healthcare system, even when this is public and egalitarian [2], a shortcoming that also affects a significant proportion of carriers with other pathogenic copy-number-variants [27].

The potential benefits and shortcomings of neonatal screening for 22q11.2 deletions have been considered in the past primarily in relation to prenatal genetic testing [28]. Advantages include early intervention to cardiac defects, hypocalcemia-induced seizures, and severe immune deficiency, while concern of causing vulnerable child syndrome through national 22q11.2 screening programs has been raised given the less-than-full penetrance of the genomic aberration [28]. Importantly, these concerns contrast with ambitions to eliminate the 'diagnostic odyssey' and intervene against failure to thrive and early, prodromal manifestations of mental disorders; complications that are often not considered when deciding on clinical screening programs.

STRENGTHS AND LIMITATIONS
The strength of this study stems from the nationwide biobank from where the samples were obtained, is representative of the
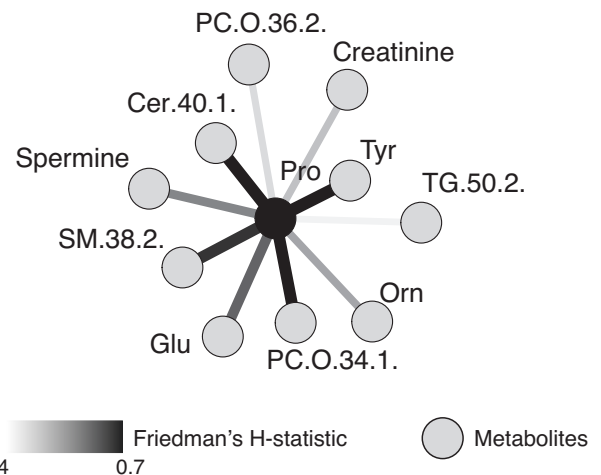
Fig. 4 Pairwise interactions for proline. Top ten metabolites interacting most strongly with proline are shown. Edges are colored according to Friedman's *H*-statistic, with the strongest interaction in black and the weakest interaction in light grey. The black node indicates proline, while the grey nodes are its interactors. Tyrosine shows the strongest interaction.

population as a whole, and rests on decades of national neonatal screening for severe congenital disorders originally prompted by analyses for phenylketonuria. Importantly, the neonatal blood spots used in our study conveniently provide genomic DNA for

efficient follow-up genotype- or sequence-based verification to eliminate false positive findings, which in the case of national routine screening will be a concern. In light of the ease of targeted genomic follow-up analyses, future refinement of the metabolic analyses to reduce false negative findings should be expected. Our study exhibits a number of limitations warranting further research. Although we identified metabolites dysregulated in 22q11.2 with relatively high predictive power, our study cohort is small. Results can, therefore, not be directly applied in the clinic. Instead, our study provides the first evidence that metabolic markers at birth may be used to identify carriers of the 22q11.2 deletion. This finding needs to be confirmed in a larger cohort, which should also include information on phenotypes and 22q11.2 deletion sizes to assess metabolic differences among carriers. In addition, such a study should also include a comparison to other diseases, such as hyperprolinaemia, to assess the specificity of the diagnostic method. As most individuals of our study cohort were carriers of the 3MB deletion we cannot conclude on metabolic subtypes within 22q11.2DS.

CONCLUSION
In conclusion, we identified metabolites dysregulated in 22q11.2, in particular tyrosine and proline, and document that careful metabolic profiling leveraging existing clinical screening programs could allow for the identification of individuals carrying the 22q11.2 deletion, although we emphasize that clinical performance and applicability awaits further studies.

DATA AVAILABILITY
The data underlying this study are not publicly available due to the Danish Data Protection Act and European Regulation 2016/679 of the European Parliament and of the Council (GDPR) that prohibit the distribution of personal data. The data are available from the corresponding authors upon reasonable request and under a data transfer and collaboration agreement.

REFERENCES
1. Same Name Campaign—22q.org. https://web.archive.org/web/20170610065340/http://www.22q.org/awareness-events/awareness/same-name-campaign/. Published 10 June 2017. Accessed 11 June 2021.
2. Olsen L, Sparsø T, Weinsheimer SM, Dos Santos MBQ, Mazin W, Rosengren A. et al. Prevalence of rearrangements in the 22q11.2 region and population-based risk of neuropsychiatric and developmental disorders in a Danish population: a case-cohort study. Lancet Psychiatry. 2018;5:573–80. https://doi.org/10.1016/S2215-0366(18)30168-8.
3. Palmer LD, Butcher NJ, Boot E, Hodgkinson KA, Heung T, Chow EWC. et al. Elucidating the diagnostic odyssey of 22q11.2 deletion syndrome. Am J Med Genet A. 2018;176:936–44. https://doi.org/10.1002/ajmg.a.38645.
4. Cabrer M, Serra G, Gogorza MS, Pereg V. Hypocalcemia due to 22q11.2 deletion syndrome diagnosed in adulthood. Endocrinol Diabetes Metab Case Rep. 2018;2018:17–0140. https://doi.org/10.1530/EDM-17-0140.
5. Schneider M, Debbané M, Bassett AS, Chow EWC, Fung WLA, van den Bree M. et al. Psychiatric disorders from childhood to adulthood in 22q11.2 deletion syndrome: results from the international consortium on brain and behavior in 22q11.2 deletion syndrome. Am J Psychiatry. 2014;171:627–39. https://doi.org/10.1176/appi.ajp.2013.13070864.
6. Hoeffding LK, Trabjerg BB, Olsen L, Mazin W, Sparsø T, Vangkilde A. et al. Risk of psychiatric disorders among individuals with the 22q11.2 deletion or duplication: a Danish nationwide, register-based study. JAMA Psychiatry. 2017;74:282–90. https://doi.org/10.1001/jamapsychiatry.2016.3939.
7. Perez E, Sullivan KE. Chromosome 22q11.2 deletion syndrome (DiGeorge and velocardiofacial syndromes). Curr Opin Pediatr. 2002;14:678–83. https://doi.org/10.1097/00008480-200212000-00005.
8. McDonald-McGinn DM, Sullivan KE, Marino B, Philip N, Swillen A, Vorstman JAS. et al. 22q11.2 deletion syndrome. Nat Rev Dis Prim. 2015;1:15071. https://doi.org/10.1038/nrdp.2015.71.
9. Martin-Nalda A, Cueto-González AM, Argudo-Ramírez A, Marin-Soria JL, Martinez-Gallo M, Colobran R. et al. Identification of 22q11.2 deletion syndrome via newborn screening for severe combined immunodeficiency. Two years'

experience in Catalonia (Spain). Mol Genet Genom Med. 2019;7:e1016. https://doi.org/10.1002/mgg3.1016.
10. Goodman BK, Rutberg J, Lin WW, Pulver AE, Thomas GH. Hyperprolinaemia in patients with deletion (22)(q11.2) syndrome. J Inherit Metab Dis. 2000;23:847–8. https://doi.org/10.1023/a:1026773005303.
11. Raux G, Bumsel E, Hecketsweiler B, van Amelsvoort T, Zinkstok J, Manouvrier-Hanu S. et al. Involvement of hyperprolinemia in cognitive and psychiatric features of the 22q11 deletion syndrome. Hum Mol Genet. 2007;16:83–91. https://doi.org/10.1093/hmg/ddl443.
12. Vorstman JAS, Turetsky BI, Sijmens-Morcus MEJ, de Sain MG, Dorland B, Sprong M. et al. Proline affects brain function in 22q11DS children with the low activity COMT 158 allele. Neuropsychopharmacology. 2009;34:739–46. https://doi.org/10.1038/npp.2008.132.
13. Nørgaard-Pedersen B, Hougaard DM. Storage policies and use of the Danish Newborn Screening Biobank. J Inherit Metab Dis. 2007;30:530–6. https://doi.org/10.1007/s10545-007-0631-x.
14. Kuhring M, Eisenberger A, Schmidt V, Kränkel N, Leistner D, Kirwan J. et al. Concepts and software package for efficient quality control in targeted metabolomics studies: MeTaQuaC. Anal Chem. 2020;92:10241–5. https://doi.org/10.1021/acs.analchem.0c00136.
15. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. https://www.R-project.org/ Published 2021. Accessed November 21.
16. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 KDD '16. ACM; 785–94. https://doi.org/10.1145/2939672.2939785.
17. Kuhn, M. Building Predictive Models in R Using the Caret Package. J Stat Softw. 2008;**28**. https://doi.org/10.18637/jss.v028.i05.
18. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinforma. 2011;12:77. https://doi.org/10.1186/1471-2105-12-77.
19. Molnar C. iml: An R package for Interpretable Machine Learning. J Open Source Softw. 2018;3:786. https://doi.org/10.21105/joss.00786.
20. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. 2021. http://arxiv.org/abs/1801.01489. Published online 23 December 2019. Accessed 9 September.
21. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat. 2008;**2**. https://doi.org/10.1214/07-AOAS148.
22. Hui W, Gel YR, Gastwirth JL. lawstat: an R package for law, public policy and biostatistics. J Stat Softw. 2008;**28**. https://doi.org/10.18637/jss.v028.i03.
23. Napoli E, Tassone F, Wong S, Angkustsiri K, Simon TJ, Song G. et al. Mitochondrial citrate transporter-dependent metabolic signature in the 22q11.2 deletion syndrome. J Biol Chem. 2015;290:23240–53. https://doi.org/10.1074/jbc.M115.672360.
24. Korteling D, Boks MP, Fiksinski AM, van Hoek IN, Vorstman JAS, Verhoeven-Duif NM. et al. Untargeted metabolic analysis in dried blood spots reveals metabolic signature in the 22q11.2 deletion syndrome. Transl Psychiatry. 2022;12:97. https://doi.org/10.1038/s41398-022-01859-4.
25. Hufton SE, Jennings IG, Cotton RG. Structure and function of the aromatic amino acid hydroxylases. Biochem J. 1995;311:353–66. https://doi.org/10.1042/bj3110353.
26. Wesseling H, Xu B, Want EJ, Holmes E, Guest PC, Karayiorgou M. et al. System-based proteomic and metabonomic analysis of the Df(16)A+/- mouse identifies potential miR-185 targets and molecular pathway alterations. Mol Psychiatry. 2017;22:384–95. https://doi.org/10.1038/mp.2016.27.
27. Calle Sanchez X, Montalbano S, Vaez M, Krebs MD, Bygerg-Grauholm J, Mortensen PB, et al. Sex chromosome aneuploidies are underdiagnosed and associated with increased risk of mental disorders. SSRN Electron J. https://doi.org/10.2139/ssrn.4165610. Published online 2022.
28. Bales AM, Zaleski CA, McPherson EW. Newborn screening programs: should 22q11 deletion syndrome be added?. Genet Med. 2010;12:135–44. https://doi.org/10.1097/GIM.0b013e3181cdeb9a.

AUTHOR CONTRIBUTIONS
Study design: TW and JC. Data acquisition: SSL and JC. Data preprocessing and quality control: JC and ME. Statistics: ME, FR, GT, JC and AI. Drafting paper: JC, TW, FR,

and ME. Supervision: TW and ME. TW, ASC, and DH conceptualized the study, obtained funding, and took full responsibility for compliance with data-sharing policies and ethical approval of the study. All authors critically revised the manuscript for important intellectual content.

COMPETING INTERESTS
The authors declare no competing interests.

ETHICAL APPROVAL
The scientific board of the Danish Cytogenetic Central Register approved the study and provided access to the genetic data that provided information on individuals with a 22q11.2 deletion and was obtained from the Danish Cytogenetic Central Registry (DCCR), which was established in 1968 and contains data on every karyotype obtained by clinical departments performing chromosomal analyses in Denmark. The project was approved by the scientific ethics committees (Videnskabsetiske Komitéer, Reg. nr: H-B-2009-026) and the Danish Data Protection Agency (Datatilsynet, Reg. nr: 2007-58-0015).

ADDITIONAL INFORMATION
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41398-023-02697-8.

Correspondence and requests for materials should be addressed to Thomas Werge or Madeleine Ernst.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.