Instituto_Nacional de Saúde Doutor Ricardo Jorge
Observações_ Boletim Epidemiológico
2018 número 22
www.insa.pt
artigos breves_ n. 9
_Doenças não transmissíveis

_Associação entre variantes genéticas e perfil clínico multidimensional de doentes com perturbação do espetro do autismo: uma abordagem integrativa

Complex associations between genetic variants and clinical profiles in autism spectrum disorder patients: an integrative systems biology approach

Muhammad Asif[1], Hugo Martiniano[1], Francisco Couto[2], Astrid M. Vicente[1]

astrid.vicente@insa.min-saude.pt

(1) Departamento de Promoção da Saúde e Prevenção de Doenças Não Transmissíveis, Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa, Portugal.
(2) Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal.

_Resumo

A complexidade genética e clínica que caracterizam a perturbação do espetro do autismo (PEA) têm limitado o desenvolvimento de biomarcadores que permitam um diagnóstico precoce e um prognóstico fiável, assim como uma abordagem personalizada para a intervenção terapêutica. Neste estudo pretendeu-se desenvolver uma abordagem integrativa para predição da apresentação clínica baseada em informação de variantes genéticas (*Copy Number Variants*, CNVs), com aplicação clínica no diagnóstico e prognóstico na PEA. Para tal, técnicas de aprendizagem automática (*machine learning*) foram aplicadas a dados clínicos e genéticos de 2446 doentes com PEA, recrutados no âmbito do consórcio *Autism Genome Project*. Análise de *clustering* de dados clínicos multidimensionais definiu, nesta população, dois subgrupos de pacientes com perfis clínicos diferindo significativamente em termos de capacidade verbal, nível cognitivo, gravidade da doença e comportamento adaptativo. A análise dos CNVs que afetam especificamente genes do cérebro, nos mesmos indivíduos, identificou 15 processos biológicos enriquecidos em genes alterados. A aplicação de um algoritmo de *machine learning* para classificação dos doentes com apresentação clínica mais disfuncional, com base nos processos biológicos alterados, mostrou que correlações entre fenótipo clínico e biologia subjacente são possíveis na PEA e que, para grupos populacionais com dados informativos, existe um poder preditivo razoável. Para implementação deste conceito na prática clínica serão necessários estudos mais alargados com dados clínicos e genómicos mais completos.

_*Abstract*

The genetic and clinical complexity that characterize Autism Spectrum Disorder (ASD) has hindered the development of biomarkers for early diagnosis and reliable prognosis, as well as a personalized to therapeutic intervention. This study aimed to develop an integrative approach for clinical presentation prediction based on Copy Number Variants (CNVs), with clinical application for diagnosis and prognosis of ASD. For this purpose, machine learning techniques were applied to a dataset of 2446 patients with ASD, recruited by the Autism Genome Project. Clustering analysis of multidimensional clinical data allowed the definition of two patient subgroups in this population, with clinical profiles differing significantly in verbal ability, cognitive level, disease severity and adaptive behavior. In the same subjects, analysis of CNVs specifically affecting brain-expressed genes identified 15 biological processes enriched for the disrupted genes. A machine learning algorithm was trained and tested to classify patients with more dysfunctional clinical presentation based on altered biological processes. The results showed that correlations between clinical phenotype and underlying biology can be established in ASD and that, for datasets with sufficiently informative data, there is a reasonable predictive power. Further studies with more complete clinical and genomic data are needed to implement this concept in clinical practice.

_Introdução

A perturbação do espetro do autismo (PEA) é uma patologia do neurodesenvolvimento com apresentação clínica muito heterogénea, variando a gravidade, o nível cognitivo, as alterações de linguagem, o comportamento adaptativo e a presença de co-morbilidades como a epilepsia ou o défice intelectual (1). A PEA tem subjacente uma arquitetura genética complexa e pouco esclarecida, que envolve mais de uma centena de genes diferentes (2). Em cerca de 20% dos casos é possível um diagnóstico etiológico, consistindo na sua grande maioria de deleções ou duplicações de segmentos do genoma designadas como *Copy Number Variants* (CNVs). Cada CNV associado à PEA é individualmente raro mas, dada a grande diversidade encontrada em pessoas com esta patologia, no seu conjunto os CNVs explicam uma fração substancial do risco genético. Um CNV pode abranger um ou mais genes implicados na PEA, perturbando processos biológicos específicos e resultando em apresentações clínicas variáveis (2).

_Objetivo

Dada a ausência de biomarcadores, o diagnóstico da PEA é inteiramente comportamental, dificultando um diagnóstico precoce e um prognóstico preciso.

O objetivo do presente estudo foi o desenvolvimento de um método preditivo para a apresentação clínica da PEA, a partir da identificação de variantes genéticas, e consequentes alterações em processos biológicos, como biomarcadores para diagnóstico e prognóstico.

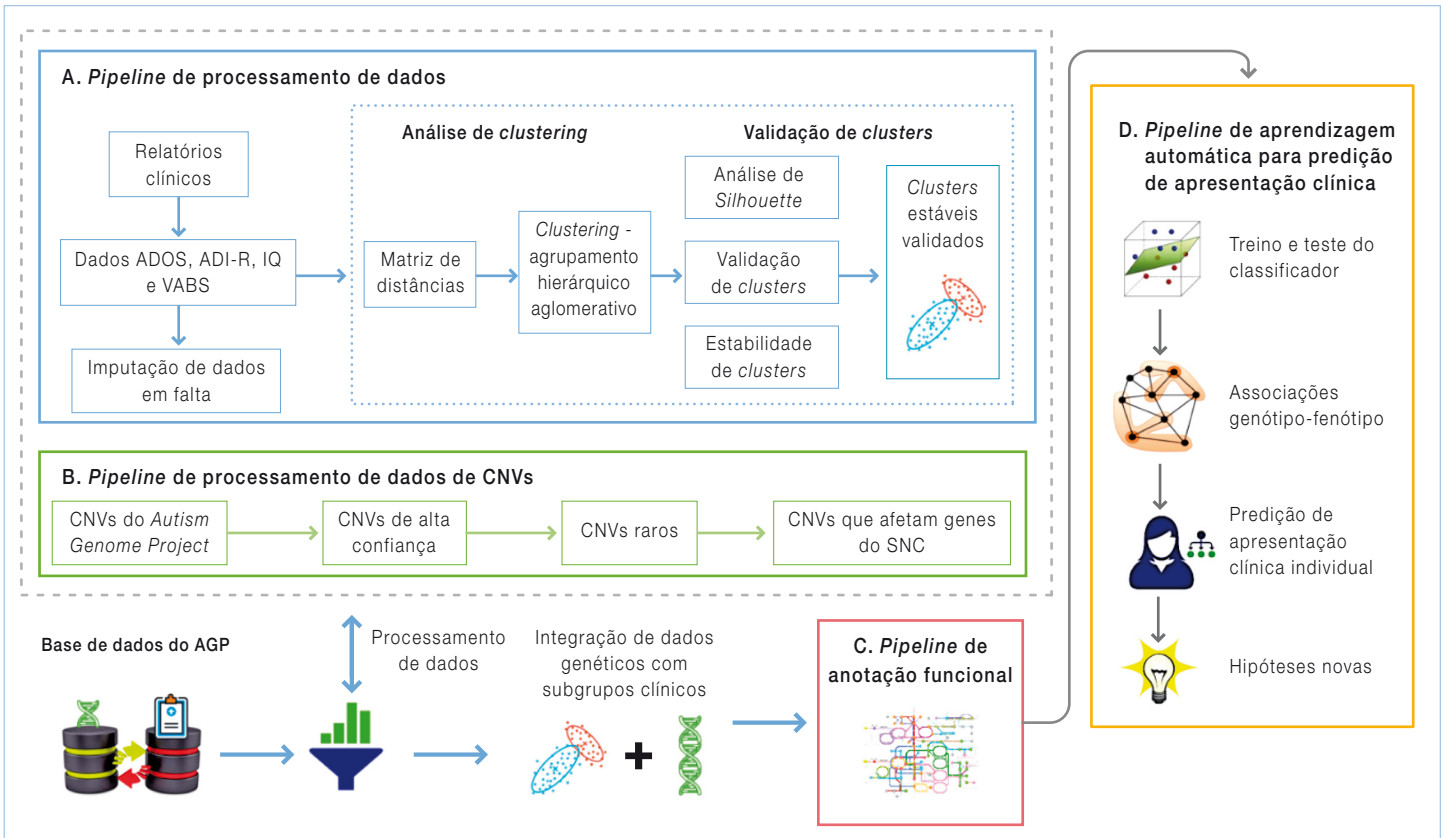
_Método

Foi definida uma abordagem integrativa, baseada em técnicas sofisticadas de aprendizagem automática (*machine learning*) supervisionadas e não supervisionadas [3-5], para predição da apresentação clínica associada a alterações em processos biológicos devidas a deleções ou duplicações de genes expressos no sistema nervoso central (SNC). Esta abordagem está esquematicamente representada na figura 1.

_Resultados

Numa primeira fase pretendeu-se agregar doentes com apresentações clínicas variáveis em subgrupos mais homogéneos. Com este fim, foram analisados os relatórios clínicos, anonimizados, de 2446 casos de PEA recrutados no âmbito do consórcio internacional *Autism Genome Project* (AGP) [6]. Os relatórios clínicos disponibilizam múltiplas medidas clinicamente relevantes, e neste estudo foram focadas essencialmente a gravidade da PEA (definida através de um instrumento estruturado de diagnóstico, o *Autism Diagnostic Observation Schedule*, ADOS), a função adaptati-

Figura 1: Representação esquemática da abordagem integrativa para identificação de associações entre variantes genéticas, processos biológicos correspondentes e apresentação clínica.

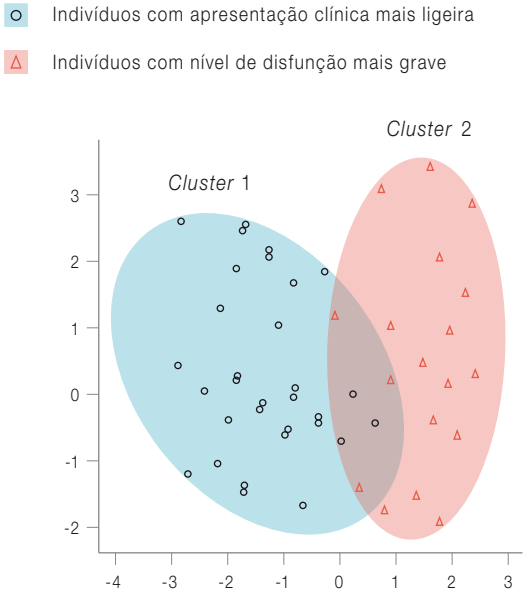


A. Identificação de subgrupos (*clusters*) a partir de dados clínicos multidimensionais; **B**. Processamento de dados de *Copy Number Variants* (CNVs) para identificação de variantes que afetam genes expressos no cérebro, potencialmente relevantes para a PEA; **C**. Anotação funcional e análise de enriquecimento para definição de processos biológicos afetados pelas variantes genéticas; **D**. *Pipeline* para aplicação de um algoritmo de *Machine Learning* para predição do subgrupo clínico a partir de processos biológicos definidos pelos CNVs.

va (avaliada utilizando a *Vineland Adaptive Behavior Scale*, VABS), a competência cognitiva (avaliada utilizando várias metodologias) e a verbalidade (a partir da *Autism Diagnostic Interview*, ADI). Partindo destes dados clínicos multidimensionais, uma análise empregando agrupamento hierárquico aglomerativo (análise de *clusters*) [3] identificou dois subgrupos de doentes clinicamente mais homogéneos (figura 2). Ambos os subgrupos (*clusters*) identificados são coesos (métrica de *Silhouette value* [7] médio para os dois *clusters* de 0,571) e estáveis (estabilidade de 0,998 e 0,996 para os *Clusters* 1 e 2, respetivamente). Em termos clínicos, os dois subgrupos diferem significativamente nos múltiplos parâmetros avaliados (tabela 1), sendo o *Cluster* 1 constituído por uma proporção maior de indivíduos com uma apresentação clínica mais ligeira e nível cognitivo normal, enquanto o *Cluster* 2 apresenta uma elevada proporção de casos de maior gravidade, não-verbais, com baixo nível cognitivo e comportamento adaptativo disfuncional.

Para os casos de PEA recrutados pelo AGP são conhecidas também as variantes genéticas raras, nomeadamente os CNVs. Neste estudo foram avaliados especificamente CNVs que deletam ou duplicam genes expressos no SNC e, como tal, de potencial relevância para o autismo [8].

Figura 2: *Clusters* de pacientes com perturbação do espetro do autismo.



Métricas de validação dos *clusters*	*Cluster* 1	*Cluster* 2
Tamanho do *cluster* (n)	903	494
Distância média entre dois indivíduos	0.235	0.231
Silhouette value	0.579	0.579
Silhouette media dos dois *clusters*	0.571	
Estabilidade do *cluster*	0.998	0.996

Tabela 1: Caracterização clínica dos *clusters*.

Parâmetro clínico	Categorias clínicas	*Cluster* 1 n (%)	*Cluster* 2 n (%)	*P-value*
ADIR estatuto verbal	ADI-R-não-verbal	0 (0)	494 (100)	<0.00001 [a]
	ADI-R-verbal	903 (100)	0 (0)	
ADOS nível de gravidade	ADOS Nível de gravidade Autismo (cotação 6-10)	714 (79,07)	392 (79,35)	<0.00001 [b]
	ADOS Nível de gravidade PEA (cotação 4-5)	64 (7,09)	102 (20,65)	
	ADOS Nível de gravidade *Non-spectrum* (cotação 1-3)	125 (13,84)	0 (0)	
VABS comunicação	VABS comunicação disfuncional (cotação ≤ 70)	307 (34)	493 (99,8)	<0.00001 [a]
	VABS comunicação normal (cotação > 70)	596 (66)	1 (0,2)	
VABS autonomia	VABS autonomia disfuncional (cotação ≤ 70)	478 (52,94)	484 (97,98)	<0.00001 [b]
	VABS autonomia normal (cotação > 70)	425 (47,07)	10 (2,02)	
VABS socialização	VABS socialização disfuncional (cotação ≤ 70)	497 (55,04)	490 (99,19)	<0.00001 [a]
	VABS socialização normal (cotação > 70)	406 (44,96)	4 (0,81)	
QI de realização	Incapacidade grave (cotação < 50)	2 (0,22)	218 (44,13)	<0.00001 [b]
	Incapacidade moderada (cotação ≥ 50 and ≤ 70)	31 (3,43)	125 (25,3)	
	Capacidade normal (cotação > 70)	870 (96,35)	151 (30,57)	
Género	Masculino	830 (91,92)	417 (84,41)	0.000015 [b]
	Feminino	73 (8,08)	77 (15,59)	

[a] *Fisher Exact Test* [b] *Chi-Square Test*

Para identificação dos processos biológicos potencialmente afetados por estes CNVs, efetuou-se uma análise de enriquecimento funcional de genes [9]. Esta análise, baseada em anotações funcionais disponíveis em bases de dados e literatura, avalia conjuntos de genes que partilham a mesma função biológica, e permite identificar os processos biológicos "enriquecidos" em genes do SNC afetados pelos CNVs. Entre os 15 processos biológicos identificados nesta análise como significativamente enriquecidos incluem-se "desenvolvimento do sistema nervoso", "cognição" e "poli-ubiquitinação de proteínas", reforçando múltiplos estudos prévios sobre a biologia subjacente à PEA (tabela 2). Os 15 processos biológicos identificados contribuem todos positivamente para a classificação dos doentes nos *clusters* definidos.

Para atingir o objetivo final, nomeadamente a predição da apresentação clínica a partir dos processos biológicos definidos pelas alterações genéticas identificadas nos pacientes, utilizou-se o método de aprendizagem automática supervisionada de Naive-Bayes [5]. Este método usa um algoritmo de classificação baseado no Teorema de Bayes para treino e teste de modelos preditivos. O classificador foi treinado assumindo que indivíduos com uma apresentação clínica mais disfuncional, agrupados no *Cluster* 2, apresentariam um padrão de processos biológicos alterados diferente dos doentes associados ao *Cluster* 1, caracterizados por uma disfunção mais ligeira. Num total de 1300 doentes com dados clínicos de *clustering* e informação sobre processos biológicos alterados, o classificador não teve um bom desempenho preditivo, apresentando baixos níveis de exatidão. Para compreender melhor os resultados obtidos, foi calculado um índice de Conteúdo de Informação (CI) para os processos biológicos identificados em cada participante. Para os 325 indivíduos com maior índice de CI, correspondentes ao 4º quartil, o desempenho do classificador de Naive-Bayes melhorou significativamente, apresentando valores de precisão de 0,82 e especificidade de 0,70, embora uma sensibilidade de 0,39 (tabela 3).

Tabela 2: Processos biológicos enriquecidos em genes expressos no SNC afetados por CNVs.

Parâmetro clínico	Genes (n)	FDR *p-value*
Homophilic cell adhesion via plasma membrane adhesion molecules	53	6.30E-09
Cell-cell adhesion via plasma-membrane adhesion molecules	66	1.70E-07
Cellular component organization or biogenesis	944	5.70E-05
Cellular component organization	915	7.00E-05
Cellular component biogenesis	475	0.00066
Cellular component assembly	434	0.00177
Nervous system development	363	0.00215
Organelle organization	562	0.00475
Protein polyubiquitination	64	0.00592
Cell projection organization	231	0.00836
Cellular localization	418	0.0091
Single-organism behavior	83	0.0196
Regulation of cellular component organization	364	0.0257
Plasma membrane bounded cell projection organization	223	0.0282
Cognition	56	0.0364
Single-organism organelle organization	263	0.044

FDR – *False Discovery Rate*

Tabela 3: Desempenho do algoritmo de Naive-Bayes na predição da apresentação clínica, de acordo com o conteúdo informativo dos processos biológicos.

Dados usados para a classificação	n	Precisão	Sensibilidade	Especificidade	FDR *p-value*
Todos os casos de PEA	1300	0.221	0.379	0.655	0.279
Casos de PEA nos últimos 3 quantis de CI	974	0.29	0.389	0.672	0.329
Casos de PEA no 3º e 4º quantil de CI	649	0.23	0.384	0.65	0.284
Casos de PEA no 4º quantil, com CI mais elevado	325	0.816	0.389	0.699	0.526

PEA – perturbação do espetro do autismo CI – Conteúdo de Informação

_Conclusões

O estudo presente indica que, para a PEA, é possível estabelecer correlações entre a apresentação clínica e uma arquitetura genética/base biológica complexas na PEA. A heterogeneidade fenotípica típica do autismo pode ser racionalizada identificando subgrupos de pacientes clinicamente mais homogéneos a partir de dados clínicos multidimensionais. Por outro lado, verificou-se que processos biológicos específicos, alterados por CNVs contendo genes do SNC, estão associados a apresentações clínicas distintas em termos de gravidade, tipo e nível de disfunção. Esta observação sugere que subgrupos com perfil clínico semelhante terão subjacentes os mesmos mecanismos biológicos. De facto, na presença de dados altamente informativos sobre processos biológicos, foi possível obter um nível razoável de confiança na predição da apresentação clínica, embora com baixa sensibilidade. Para se obter um poder preditivo razoável para a generalidade dos pacientes é, no entanto, necessário melhorar o conteúdo informativo dos processos biológicos do sistema nervosos central, nomeadamente sobre as vias fisiológicas do cérebro e sobre as consequências funcionais de variantes genéticas.

No seu global, esta abordagem serve como prova de conceito de que correlações entre variantes genéticas e perfis clínicos multidimensionais podem ser estabelecidas para a PEA, e de que o conhecimento dos processos biológicos afetados por CNVs pode predizer a apresentação clínica. A identificação destas correlações será importante para a descoberta de alvos fisiológicos para terapêutica farmacológica que sejam mais eficientes em subgrupos de doentes. Estas observações têm ainda um potencial para aplicação clínica, no diagnóstico precoce e no prognóstico do doente mediante a identificação de variantes genéticas, apoiando assim uma abordagem personalizada para intervenção terapêutica precoce. Estudos adicionais em amostras populacionais de maior dimensão, com dados clínicos e genómicos mais completos, são agora necessários para explorar este conceito e permitir a sua implementação na prática clínica.

Referências bibliográficas:

(1) American Psychiatric Association. Cautionary Statement for Forensic Use of DSM-5. IN: Diagnostic And Statistical Manual Of Mental Disorders, 5th ed. Washington, DC: American Psychiatric Association Publishing, 2013.

(2) Geschwind DH, State MW. Gene hunting in autism spectrum disorder: on the path to precision medicine. Lancet Neurol. 2015;14(11):1109-20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4694565/

(3) Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J Classif. 2014;31(3):274-95.

(4) Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.

(5) Kuncheva LI. On the optimality of Naïve Bayes with dependent binary features. Pattern Recognit Lett. 2006;27(7):830-7.

(6) Pinto D, Pagnamenta AT, Klei L, et al. Functional impact of global rare copy number variation in autism spectrum disorder. Nature. 2010;466(7304):368-72. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3021798/

(7) Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20(C):53-65.

(8) Parikshak NN, Luo R, Zhang A, et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell. 2013;155(5):1008-21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3934107/

(9) Reimand J, Arak T, Adler P, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016;44(W1):W83-9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4987867/