VINICIUS DAGUANO GASTALDI

**Análise composicional de sequenciamento completo
do exoma de probandos do transtorno do espectro
autista**

Dissertação apresentada à Faculdade de
Medicina da Universidade de São Paulo para
obtenção do título de Mestre em Ciências

Programa de Psiquiatria

Orientadora: Profa. Dra. Helena Paula Brentani

(VERSÃO CORRIGIDA. RESOLUÇÃO COPGR 6018/11, DE 1 DE NOVEMBRO DE 2011. A
VERSÃO ORIGINAL ESTÁ DISPONÍVEL NA BIBLIOTECA DA FMUSP)

SÃO PAULO
2019

VINICIUS DAGUANO GASTALDI

**Análise composicional de sequenciamento completo
do exoma de probandos do transtorno do espectro
autista**

Dissertação apresentada à Faculdade de
Medicina da Universidade de São Paulo para
obtenção do título de Mestre em Ciências

Programa de Psiquiatria

Orientadora: Profa. Dra. Helena Paula Brentani

(Versão corrigida. Resolução CoPGr 6018/11, de 1 de novembro de 2011. A versão
original está disponível na Biblioteca da FMUSP)

SÃO PAULO

2019

Dados Internacionais de Catalogação na Publicação (CIP)

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

©reprodução autorizada pelo autor

Gastaldi, Vinicius Daguano

Análise composicional de sequenciamento completo do exoma de probandos do transtorno do espectro autista / Vinicius Daguano Gastaldi. -- São Paulo, 2019.

Dissertação (mestrado)--Faculdade de Medicina da Universidade de São Paulo.

Programa de Psiquiatria.

Orientadora: Helena Paula Brentani.

Descritores: 1.Sequenciamento completo do exoma
2.Genética 3.Psiquiatria 4.Transtorno do espectro autista 5.Análise por conglomerados 6.Biologia computacional

USP/FM/DBD-271/19

Responsável: Erinalva da Conceição Batista, CRB-8 6755

À meus pais, Cristiani e Mário,
À meus irmãos, Thiago e Matheus,
À Renata,
Obrigado.

AGRADECIMENTOS

À minha orientadora, Professora Helena Brentani, por ter me recebido em seu grupo de pesquisa e ensinado tanto a respeito de genética e pesquisa. À minha coorientadora, Ana Carolina Tahira, por toda a paciência para me ensinar bioinformática. À Viviane Neri de Souza Reis por toda sua ajuda desde o momento da escrita do meu primeiro projeto. Ao Professor Ricardo Vêncio pela colaboração e auxílio nas análises finais.

À todo o grupo PsysBio e IPq: André, Arthur, Bianca, Caroline, Cleiton, Gisele, Joana, Leonardo, Marta, Maurício, Mirian, Raphael, Rayssa, Verônica, Victor. Muito obrigado pelas risadas, conversas e apoio.

À Vanessa de Jesus Rodrigues de Paula por ser a melhor companhia de laboratório que eu poderia pedir. Por toda a ajuda, conselhos, companhia, almoços e risadas. Obrigado por tudo.

À Eliza e Isabel por todo o suporte desde antes da minha matrícula até o último momento do mestrado.

Aos meus pais, Cristiani e Mário, por sempre acreditarem em mim e me darem todo o apoio do mundo. Aos meus irmãos, Thiago e Matheus, por sempre estarem presentes.

À Renata, que esteve ao meu lado durante toda essa etapa e desde muito antes disso. Obrigado por todo o companheirismo e por ser minha parceira para todas as horas. Sem você eu nunca teria chegado até aqui.

À Simons Foundation Autism Research Initiative, SFARI, por disponibilizar os dados utilizados para realização desse projeto. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"Before the truth will set you free, it'll piss you off..."

MANTRA, Bring Me The Horizon

Sumário

Lista de abreviaturas	VIII
Lista de gráficos	X
Lista de figuras	XI
Lista de tabelas	XII
Resumo.....	XIII
Abstract.....	XV
1 Introdução	1
1.1 Heterogeneidade fenotípica do Transtorno do Espectro Autista	1
1.2 Arquitetura Genética Complexa: transtorno poligênico e multifatorial	2
1.3 Estudos de associação genômica completa em Transtorno Espectro Autista - variantes comuns	5
1.4 Estudos de associação de copy number variations e sequenciamento completo do exoma em TEA - variantes de novo e muito raras	9
1.5 Classificações propostas para o Transtorno do Espectro Autista	17
1.6 Análise composicional	20
1.7. Hipótese do estudo	21
2 Objetivos	23
2.1 Objetivo Principal	23
2.2 Objetivos Específicos	23
3 Métodos	24
3.1. Amostra.....	24
3.2 Processamento de dados de sequenciamento completo do exoma	24
3.3 Processamento dos VCFs anotados	26
3.4 Análise composicional	28
3.5 Análise de agrupamentos.....	28
3.6 Avaliação fenotípica dos agrupamentos estratificados	29
3.7 Aprovação do comitê de ética	29
4 Resultados	30
4.1 Processamento de dados e tamanho da amostra.....	30
4.2 Abordagem Bimodal - Análise de composição, análise de agrupamentos e comparação de fenótipos	31
4.3 - Abordagem com priori uniforme - Análise de composição, análise de agrupamentos e comparação de fenótipos.....	35
5 Discussão.....	42
6 Conclusão	47

7 Anexos	48
8 Bibliografia	66

Lista de abreviaturas

DSM-5	Manual Diagnóstico e Estatístico de Transtornos mentais, Versão 5
TEA	Transtorno do Espectro Autista
DI	Deficiência Intelectual
CNVs	do inglês, <i>Copy Number Variations</i>
MAF	do inglês, <i>Minor Allele Frequency</i>
GWAS	do inglês, <i>Genome-Wide Association Studies</i>
SCE	Sequenciamento Completo do Exoma
SNPs	do inglês, <i>Single Nucleotide Polymorphisms</i>
SSC	<i>Simons Simplex Collection</i>
QTL	do inglês, <i>Quantitative Trait Locus</i>
DNA	do inglês, <i>deoxyribonucleic acid</i>
LGD	do inglês, <i>likely gene-disruptive</i>
PPI	do inglês, <i>protein protein interaction</i>
FPE	do inglês, <i>female protective effect</i>

SFARI	<i>Simons Foundation for Autism Research Initiative</i>
CBCL	<i>Child Behavior Checklist</i>
GATK	do inglês, <i>Genome Analysis ToolKit</i>
VCFs	do inglês, <i>Variant Call Format files</i>
CLR	do inglês, <i>centered log ratio</i>
AU	do inglês, <i>Approximately Unbiased</i>
ADI-R	<i>Autism Diagnostic Interview - Revised</i>
RBS-R	<i>Repetitive Behavior Scale-Revised</i>
ABC	<i>Autism Behavior Checklist</i>
cdf	do inglês, <i>cumulative distribution function</i>

Lista de gráficos

Gráfico 1 – Boxplot com beeswarm dos agrupamentos (priori bimodal).....	34
Gráfico 2 – Comparação da distribuição cumulativa do agrupamento A2 para a subescala de Sociabilidade da Vineland II.....	34
Gráfico 3 – Boxplot com beeswarm dos agrupamentos (priori uniforme).....	39
Gráfico Anexo C – Comparação da distribuição cumulativa para subescalas da ABC.....	52
Gráfico Anexo D – Comparação da distribuição cumulativa para ADI-R A Total.....	53
Gráfico Anexo E – Comparação da distribuição cumulativa para comportamentos externalizantes e internalizantes da CBCL.....	54
Gráfico Anexo F – Comparação da distribuição cumulativa para a medida de QI Verbal.....	55
Gráfico Anexo G – Comparação da distribuição cumulativa para subescalas da RBS-R.....	56
Gráfico Anexo H – Comparação da distribuição cumulativa para subescalas da ABC.....	57
Gráfico Anexo I – Comparação da distribuição cumulativa para subescalas da ABC.....	58
Gráfico Anexo J – Comparação da distribuição cumulativa para subescalas da ADI-R.....	59
Gráfico Anexo K – Comparação da distribuição cumulativa para comportamentos externalizantes e internalizantes da CBCL.....	60
Gráfico Anexo L – Comparação da distribuição cumulativa para medidas de QI fornecidas pela SSC.....	61
Gráfico Anexo M – Comparação da distribuição cumulativa para medidas de QI fornecidas pela SSC.....	62
Gráfico Anexo N – Comparação da distribuição cumulativa para subescalas da RBS-R.....	63
Gráfico Anexo O – Comparação da distribuição cumulativa para subescalas da RBS-R.....	64
Gráfico Anexo P – Comparação da distribuição cumulativa para subescalas da Vineland II.....	65

Lista de figuras

Figura 1 – Agrupamento hierárquico (priori bimodal).....	31
Figura 2 – Agrupamento hierárquico (priori uniforme).....	37

Lista de tabelas

Tabela 1 – Detalhes das vias genéticas utilizadas no trabalho.....	27
Tabela 2 – Resultado das filtragens aplicadas aos arquivos VCF.....	30
Tabela 3 – Descrição da quantidade de variantes encontradas nos genes das vias selecionadas por Luo et al., 2018.....	30
Tabela Digital 1 – Valores de composição e transformação CLR obtidos na abordagem com priori bimodal para todos os probandos.....	LINK ABAIXO
Tabela 4 – Detalhes dos agrupamentos encontrados com o uso da priori bimodal.....	32
Tabela Digital 1 – Valores de composição e agrupamento a qual pertence para ambas as abordagens..... para todos os probandos.....	LINK ABAIXO
Tabela Digital 3 – Valores de composição e transformação CLR obtidos na abordagem com priori uniforme para todos os probandos.....	LINK ABAIXO
Tabela 5 – Resultados da comparação das distribuições cumulativas (priori bimodal).....	36
Tabela 6 – Detalhes dos agrupamentos encontrados com o uso da priori uniforme.....	38/39
Tabela 7 – Resultados da comparação das distribuições cumulativas (priori uniforme).....	40/41
Tabela Anexa A – Descrição da amostra do SSC (2313 probandos).....	48/49
Tabela Anexa B – Detalhes das subescalas fenotípicas utilizadas no trabalho.....	50/51

Link para acesso as **Tabela Digitais 1, 2, 3**: http://bit.ly/dissertacao_vinicius_gastaldi

Resumo

Gastaldi VD. Análise composicional de sequenciamento completo do exoma de probandos do transtorno do espectro autista [dissertação]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2019.

O sequenciamento completo do exoma (SCE) cobre a porção codificadora de proteínas do genoma, a qual representa apenas 1% do mesmo, mas que estima-se conter 85% das mutações causadoras de doenças em distúrbios mendelianos. Nos últimos anos, o SCE tem contribuído para melhor caracterizar a complexa arquitetura genética de alguns distúrbios do neurodesenvolvimento através da detecção de variantes raras, ajudando a descrever as características da variação genética que contribuem para a variabilidade fenotípica hereditária. Neste estudo, hipotetizamos que uma composição de variantes raras poderia agrupar probandos com Transtorno do Espectro Autista (TEA) refletindo sua variabilidade fenotípica, incluindo o QI (coeficiente de inteligência) e outras medidas relacionadas ao TEA. Para testar essa hipótese, usamos análise composicional em conjunto com análise de agrupamentos hierárquica para estratificar probandos através dos seguintes tipos de variantes: *de novo missense*, *de novo* Likely Gene Disrupting (LGD - provavelmente gene disruptivas), *missense* herdada e LGD herdada. Utilizamos como amostra um conjunto de dados de 2313 probandos com TEA provenientes da *Simons Simplex Collection*, um dos projetos principais da Iniciativa de Pesquisa em Autismo da Fundação Simons. A análise composicional acoplada a inferência bayesiana tradicional com uma priori bimodal resultou em seis agrupamentos onde o principal fator de agrupamento foi a falta de tipos de variação, mas sem diferença estatisticamente significativa no QI entre os agrupamentos. Uma comparação da distribuição cumulativa dos escores dos fenótipos usando um teste Kolmogorov-Smirnov de dois lados de amostra única revelou diferenças significativas para quatro agrupamentos em várias subescalas das escalas ABC, ADI-R, CBCL e Vineland II. Esses resultados indicam que ter uma proporção maior das variações mais prejudiciais não se correlaciona diretamente com os piores fenótipos. A presença ou ausência de tipos de variação não explica os fenótipos, o que pode indicar que o achatamento de dados resultante da abordagem bimodal pode esconder relações importantes entre composição e fenótipos. Por essa razão, usamos uma segunda abordagem com a adição de uma priori uniforme antes da análise composicional. Os 29 agrupamentos estratificados nessa abordagem apresentaram uma visão

composicional mais fina e com maior variabilidade. Observamos que indivíduos com *de novo* LGD e *de novo missense* apresentam tendência de QI mais baixo, contudo, a associação entre a composição de variantes e o QI não é direta. Isso também foi observado na comparação da distribuição cumulativa dos escores fenotípicos. Existem diferenças significativas para todas as escalas e para 88% das subescalas para a maior parte dos agrupamentos, no entanto, isto não é suficiente para identificar uma associação clara entre composição e fenótipos. Nossos resultados mostram que a análise composicional dos dados de SCE é capaz de estratificar aglomerados estáveis de probandos com TEA com diferentes composições de variantes e de fenótipos. Concluimos que a análise composicional estratificou agrupamentos promissores quanto à relação composição-fenótipo e, ao aprimorar essa abordagem, pode ser possível entender melhor a arquitetura genética do TEA e de outras doenças complexas.

Descritores: Sequenciamento completo do exoma; Genética; Psiquiatria; Transtorno do espectro autista; Análise por conglomerados; Biologia computacional.

Abstract

Gastaldi VD. Compositional analysis of whole-exome sequencing data of autism spectrum disorder probands [dissertation]. São Paulo: “Faculdade de Medicina, Universidade de São Paulo”; 2019.

Whole-exome sequencing (WES) covers the protein coding portion of the genome, which accounts for only 1% of its length, but is estimated to contain 85% of the disease-causing mutations in Mendelian disorders. In recent years, WES has been contributing to better characterize the complex genetic architecture of some neurodevelopmental disorders through the detection of rare variants, helping to describe the characteristics of genetic variation that contribute to heritable phenotypic variability. In this study, we hypothesized that a composition of rare variants could group Autism Spectrum Disorder (ASD) probands reflecting their phenotypic variability, including IQ (intelligence quotient) and other ASD related measurements. To test this hypothesis, we used compositional analysis and hierarchical clustering to stratify probands using the following variant types: *de novo missense*, *de novo* Likely Gene Disrupting (LGD), inherited *missense*, and inherited LGD. Our sample consisted of a dataset containing WES and phenotype data from over 2300 ASD probands from the Simons Simplex Collection, a core project and source of the Simons Foundation Autism Research Initiative. Compositional analysis coupled to traditional Bayesian inference with a bimodal prior resulted in six clusters where the main grouping factor was the lack of variation types, but with no statistically significant difference in IQ between the clusters. A comparison of the cumulative distribution of the scores of the phenotypes using a one-sample two-sided Kolmogorov-Smirnov test revealed significant differences for four clusters in a number of subscales of the ABC, ADI-R, CBCL and Vineland II scales. These results indicate that having a higher proportion of the most damaging variations does not directly correlates to having the worst phenotypes. The presence or absence of types of variation does not explain the phenotypes, what might indicate that the data flattening resulting from the bimodal approach might hide important relationships between composition and phenotypes. For this reason, we used a second approach with the addition of a uniform prior before the compositional analysis. The 29 stratified clusters in this approach had a finer

compositional view with more variability. We observed that individuals with *de novo* LGD and *de novo missense* tend to have lower IQ, however, the variant composition and IQ association is not straightforward. This was also observed in the comparison of cumulative distribution of phenotype scores. Even though there are significant differences for all scales and 88% of the subscales for a majority of the clusters, they were not sufficient to identify a clear association between composition and phenotypes. Our results show that compositional analysis of WES data is able to stratify stable clusters of ASD probands with different variant compositions and phenotypes. We conclude that the compositional analysis produced promising clusters regarding composition-phenotypes relationship and, by improving this approach, it may be possible to better understand the architecture of ASD and of other complex diseases.

Descriptors: Whole exome sequencing; Genetics; Psychiatry; Autism spectrum disorder; Cluster analysis; Computational biology.

1 Introdução

1.1 Heterogeneidade fenotípica do Transtorno do Espectro Autista

Os transtornos do neurodesenvolvimento são caracterizados por início precoce durante o desenvolvimento do indivíduo, na maioria das vezes antes da criança atingir idade escolar. São definidos por prejuízos no desenvolvimento, afetando o funcionamento do indivíduo em diversas esferas, como pessoal, social e acadêmica¹.

De acordo com o Manual Diagnóstico e Estatístico de Transtornos¹ da Associação Americana de Psiquiatria, duas dimensões são necessárias para o diagnóstico do Transtorno do Espectro Autista (TEA): alterações da sociabilidade e presença de padrões restritivos e repetitivos de comportamento, interesses ou atividades. Segundo a *Autism and Developmental Disabilities Monitoring Network*², a prevalência do autismo em crianças com oito anos de idade é de 1 em 68, com 1 caso a cada 42 meninos e 1 caso a cada 189 meninas. A média de idade para o primeiro diagnóstico do TEA é de 53 meses. É interessante notar que a proporção de casos registrados citada por Devlin e Scherer³ e também por Duchan e Patel⁴ é de quatro meninos para cada menina diagnosticada, valor similar ao encontrado pela *Autism and Developmental Disabilities Monitoring Network*².

O diagnóstico do TEA é clínico, feito com base nas características individuais e com o uso de especificadores para uma melhor caracterização. Estes especificadores relacionam-se a presença ou ausência de prejuízos intelectuais, comorbidades clínicas e psiquiátricas, prejuízos de linguagem, outras condições médicas/genéticas ou ambientais/adquiridas, além de informações quanto a idade dos primeiros sintomas, se houve regressão das habilidades adquiridas e a severidade¹.

A interação social do indivíduo é marcada por problemas no funcionamento social, desde inadequação aos diferentes contextos sociais até falta de interesse por outras

pessoas. O compartilhamento de interesses é reduzido e o indivíduo apresenta dificuldades para iniciar ou responder a interações sociais⁵. Também estão incluídos aqui os déficits de comunicação marcados por problemas no uso da comunicação não-verbal, variando de contato ocular e linguagem corporal anormal, até total ausência de expressões faciais. No domínio dos comportamentos há uso de linguagem idiossincrática, além de linguagem e movimentos motores repetitivos e estereotipados. O indivíduo insiste na manutenção da rotina, com grande angústia associada a mudanças como as de trajeto e alimentação. Apresenta um grau incomum de interesse e fixação por determinados assuntos e também hipossensibilidade ou hipersensibilidade a estímulos sensoriais, com possível desenvolvimento de interesse incomum em um determinado¹.

Cerca de 70% dos pacientes com TEA apresentam alguma comorbidade psiquiátrica e 48% duas ou mais⁶. Simonoff et al.⁶ encontraram deficiência intelectual (DI) em 70% dos participantes com TEA. Duchan e Patel⁴ relatam que 87,3% das crianças diagnosticadas com TEA na faixa etária entre 6 e 17 anos também apresentam transtorno de déficit de atenção com hiperatividade (TDAH), depressão, ansiedade e problemas de conduta ou comportamento. Além das comorbidades citadas anteriormente, são também associadas aos TEA: epilepsia, transtornos do sono, transtornos de linguagem, enurese, síndromes genéticas⁶⁻¹⁰.

1.2 Arquitetura Genética Complexa: transtorno poligênico e multifatorial

A observação da relação do autismo com síndromes raras e desordens cromossômicas, assim como estudos com gêmeos¹¹⁻¹³ e famílias^{14,15}, mostraram a contribuição da genética para o TEA¹⁶. Diversos tipos de estudos genéticos foram feitos para TEA sugerindo herança autossômica dominante assim como recessiva. Nos estudos de ligação, Piggot et al.¹⁷ cita que foram encontradas regiões de interesse associadas aos

TEA nos cromossomos 1p, 2q, 3q, 6q, 7q, 13, 16p, 17q, 19p e Xq. Freitag et al.¹⁸ em sua revisão mostram as regiões 2q21-33, 3q25-27, 3p25, 4q32, 6q14-21, 7q22, 7q31-36, 11p12-13 e 17q11-21, como regiões encontradas em pelo menos dois estudos independentes. No entanto, Abrahams e Geschwind¹⁶ ressaltam que apenas loci em 17q11-17q21 e 7q foram replicados em níveis que poderiam ser considerados significativos para o genoma todo, mas mesmo assim validando uma contribuição poligênica.

Em torno de 10% dos indivíduos com TEA também apresentam uma doença Mendeliana ou síndrome genética¹⁶. As mais comumente citadas segundo os autores são a síndrome do X frágil ($\approx$ 1-2% dos casos de TEA), esclerose tuberosa ($\approx$ 1%), síndrome de Rett ($\approx$ 0,5%) e neurofibromatose (tipo 1, $<$ 1%)^{3,19}. Casos de anormalidades cromossômicas também foram verificados em aproximadamente 5% dos casos de TEA^{3,19}. A anormalidade cromossômica mais comum encontrada nos estudos citogenéticos é a duplicação do alelo materno 15q11-15q13 da região da síndrome de Prader-Willi/Angelman, presente 1-3% dos casos^{3,19}. No cromossomo 7, as principais regiões de interesse são 7q22 e 7q31. Benvenuto et al.¹⁹ destaca o gene *RELN* na região 7q22, envolvido na migração de neurônios, e o gene *MET* da região 7q31, envolvido no desenvolvimento do córtex cerebral e do cerebelo; ambos fatores citados anteriormente. Outras regiões destacadas por Benvenuto et al.¹⁹, são as regiões 16p11.2, 2q37 e 2p25.2. Deleções na região 16p11.2 foram associadas com TEA²⁰, obesidade severa^{21,22}, epilepsia e tamanho anormal da cabeça²³.

Graças a novas técnicas de análise estima-se que entre 10 e 20% dos pacientes com TEA possuem etiologia genética confirmada, sendo que mais de 103 genes e 44 loci já foram identificados como fatores genéticos que podem levar a esse quadro²⁴. Devlin e Scherer³ citam anormalidades cromossômicas raras como presentes em até 7,8% dos

casos de TEA, enquanto *copy number variations* (CNVs) como presentes em até 10% dos casos, incluindo CNVs herdadas raras e *de novo*. Assim, tanto variantes estruturais quanto de base única raras e comuns podem contribuir para o TEA. As variantes muito comuns são aquelas que possuem frequência alélica mínima (MAF, do inglês *minor allele frequency*) entre 5% e 50%, enquanto as menos comuns possuem MAF entre 1% e 5%. As variantes raras são aquelas com MAF abaixo de 1%, mas ainda assim presentes em uma das populações humanas principais²⁵. As variantes comuns são o objeto de trabalho dos estudos de associação genômica completa (GWAS, do inglês *Genome-wide association studies*) e as variantes raras dos estudos de CNV e sequenciamento completo do exoma (SCE), os quais serão abordados nas próximas seções.

Por outro lado, existem vários fatores de risco de exposição ambiental associados com TEA. A idade dos pais no momento da concepção pode estar associada com um maior risco de desenvolvimento de TEA, com aumento do risco de acordo com aumento da idade^{4,26}. Nascimento prematuro^{4,27}, baixo peso gestacional^{4,26,28}, hipóxia intraparto²⁶, corioamnionite e hemorragia intraparto aguda²⁸ aparecem como condições obstétricas associadas.

Hoje podemos dizer que TEA é uma desordem complexa, multifatorial e poligênica^{3,17,29}. Gaugler et al.³⁰ calcularam que a herdabilidade em sentido restrito do TEA apresenta valor de aproximadamente 52,4%, confirmando a presença de um forte componente genético e ambiental. A amostra utilizada incluía 1.6 milhões de famílias suecas parte do estudo “Estudo Genético e Ambiental de Autismo baseado em População” (no inglês, *Population-Based Autism Genetics and Environment Study*)³⁰.

1.3 Estudos de associação genômica completa em Transtorno Espectro Autista - variantes comuns

Nos últimos anos GWAS tornou-se uma abordagem popular para o trabalho com fenótipos complexos³¹, revelando diversos genes com implicações em TEA³².

Wang et al.³³, Weiss et al.³⁴ e Ma et al.³⁵ realizaram estudos de GWAS em TEA para tentar identificar variantes comuns associadas dentro deste transtorno. O primeiro grupo de autores realizou o estudo com 780 famílias onde 6 SNPs (do inglês, *Single Nucleotide Polymorphisms*) foram associados com TEA, sendo o rs4307059 da região 5p14.1 o único significativo ($P < 5 \times 10^{-8}$)³³. Weiss et al.³⁴ realizou um estudo de linkage genético e mapeamento de associações com 1.031 famílias usando 500 mil SNPs do genoma todo. Neste trabalho os SNPs rs10513025, rs10513026 e rs16883317 na região 5p15 foram associados com os TEA. O estudo de Ma et al.³⁵ envolveu 438 famílias caucasianas e identificou 96 SNPs na região 5p14.1 com forte sugestão de associação com TEA, porém nenhum deles apresentou associações significativas.

Anney et al.³⁶ realizaram um estudo de GWAS com mais de 1.369 famílias. Apenas um SNP, o rs4141463 na região intrônica de 300 kb do gene *MACROD2* (20p12.1), foi identificado como significante. Os autores destacam o fato de que o número de amostras limitou a capacidade de detecção de associações e a replicabilidade das mesmas³⁶. Estes foram apenas os primeiros estudos com GWAS feitos para TEA, os quais continuaram sendo realizados³².

O SNP rs4307059 identificado por Wang et al.³³ foi estudado por St. Pourcain et al.³⁷ em 7.131 crianças. Esse SNP reside na região entre dois genes codificadores de moléculas de adesão de células neuronais, *CDH9* e *CDH10*. O foco do estudo foi verificar se o SNP rs4307059 é um QTL (*quantitative trait locus*), como foco especial no domínio

comunicação em TEA. St. Pourcain et al.³⁷ sugeriram que há associação entre o rs4307059 e fenótipos de TEA no espectro da dimensão comunicação social.

No estudo de GWAS de Salyakina et al.³⁸ foram utilizadas 124 famílias com diagnóstico de síndrome de Asperger no conjunto de dados de descoberta e outras 110 famílias no conjunto de dados de validação. A estratificação em um grupo mais homogêneo foi utilizada para aumentar a probabilidade de encontrar loci de susceptibilidade reais. As regiões 5q21.1 e 15q22.1-q22.2 foram os achados mais significativos na combinação dos dois conjuntos de dados. Foram também identificadas 3 regiões de ligação com sobreposição a estudos anteriores para síndrome de Asperger, além de oito regiões de associação sobrepostas a áreas de ligação de TEA. Os autores sugerem que a síndrome de Asperger possui fatores genéticos comum com TEA, mas que há fatores de risco genético únicos para Asperger³⁸.

O estudo de Cho et al.³⁹ contou com 42 pacientes coreanos do sexo masculino diagnosticados com TEA e atrasos no desenvolvimento da linguagem. Os SNPs rs7125479 e rs11212733 da região 11q22.3 foram detectados como marcadores de atraso no desenvolvimento da linguagem em TEA.

Anney et al.⁴⁰ não encontraram nenhuma associação significativa entre os mais de um milhão de SNPs testados por eles. Entretanto, o SNP com o menor valor-p, o rs1718101 da região 7q35, o qual faz parte do gene *CNTNAP2*, cuja implicação com TEA foi sugerida por Poot et al.⁴¹. Connolly, Glessner e Hakonarson⁴² em um estudo com 2.165 participantes associaram variantes em 8 genes com sinais e sintomas presentes em TEA. O gene *KCND2* foi associado com “expressões faciais excessivamente sérias”, *NOS2A* com perda de habilidades motoras, *SDKI* com tensão em contextos sociais, *C8ORFK32* com a tendência a focar apenas nas partes, e *NELL1* com desmaios.

O estudo do *Cross-Disorder Group of the Psychiatric Genomics Consortium*⁴³ analisou diversas desordens psiquiátricas em 33.332 casos e 27.888 de descendência europeia, relacionando mutações no gene *CACNA1C* com a síndrome de Timothy, cuja variação de fenótipo inclui TEA. O trabalho de Xia et al.⁴⁴ identificou através de meta-análise os SNPs rs936938, rs6537835 e rs1877455, além dos haplótipos *AMPD1-NRAS-CSDE1*, *TRIM33* e *TRIM33-BCAS2*, todos na região 1p13.2.

St. Pourcain et al.⁴⁵ realizou um estudo com crianças cujos pais reportaram terem déficits de comunicação social. A coorte inicial possuía 5.584 indivíduos e uma segunda coorte de 1.364 indivíduos também foi utilizada, ambas com descendência europeia. Os SNPs rs9257616 da região cromossômica 6p22.1 e o rs2352908 da região 14q22.1 foram replicados na segunda coorte, ainda que com níveis de associação menores que os normalmente utilizados. A região 14q22.1 é depletada de genes, enquanto a região 6p22.1 faz parte de um complexo principal de histocompatibilidade, no qual o locus candidato mais forte é o *TRIM27*.

O estudo de Jones et al.⁴⁶ com 965 jovens adultos não teve nenhum SNP com associação *genome-wide* significativa ($p < 5.0 \times 10^{-8}$). No entanto, associações observadas nominais ($p < 1.0 \times 10^{-5}$) forneceram suporte para dois candidatos previamente identificados, *PRKCB1* e *CBLN1*.

O *The Autism Spectrum Disorders Working Group* do *The Psychiatric Genomics Consortium*⁴⁷ conduziu uma meta-análise com mais de 16 mil indivíduos afetados. Foram utilizadas 14 coortes independentes, com a meta-análise primária feita com 7.387 indivíduos com diagnóstico de TEA e 8.567 controles. Com uma meta-análise combinada com as amostras dos dados de GWAS em TEA do *Psychiatric Genomics Consortium*⁴⁸ e da *The Lundbeck Foundation Initiative for Integrative Psychiatric Research*, uma única associação por GWAS foi identificada para rs1409313-T, um marcador localizado na

região cromossômica 10q24.32 dentro do intron 1 do *CUEDC2*. Essa região já foi previamente associada com habilidades sociais em uma coorte populacional independente⁴⁶. Outros 12 novos GWS loci foram identificados na análise conjunta com os dados de GWAS de esquizofrenia do *Psychiatric Genomics Consortium*, incluindo loci anteriormente implicados com TEA, como *FOXP1* na região 3p13, *ATP2B2* na região 3p25.3, e um “hub de neurodesenvolvimento” na região cromossômica 8p11.23⁴⁷.

As análises de GWAS para características complexas têm mostrado que são necessários grandes números amostrais para detecção de variantes comuns, muitas vezes precisando de um número amostral acima de 10.000³¹. Além disso, os SNPs encontrados por estudos de GWAS muitas vezes não são replicados em estudos posteriores. Torricio et al.⁴⁹ buscaram replicar achados de GWAS para os genes *SEMA5A* (rs1051325), *MACROD2* (r4141463) e *MSNPI* (rs4307059) utilizando pelo menos cinco populações europeias. Os resultados não mostraram nenhuma associação significativa para nenhuma das cinco populações e nem para a amostra combinada utilizando o modelo genético aditivo. Entre as possíveis razões para a falha na replicação os autores destacam o tamanho da amostra e possíveis diferenças no pool genético e padrões de desequilíbrio de ligação.

Com a ideia de que vários SNPs de pequeno efeito agindo em conjunto seriam mais importantes para o risco do TEA do que um SNP único, o olhar para GWAS modificou-se, buscando escores de risco poligênico. Os escores de risco poligênico têm sido utilizados para avaliar o efeito dos marcadores que não atingem significância estatística de maneira individual⁵⁰. Os marcadores são selecionados a partir de uma amostra inicial e utilizados para construir um escore formado pela soma ponderada dos alelos associados para cada indivíduo. A associação entre uma variável e esse escore implica que um sinal genético está presente nos marcadores selecionados, podendo então

ser utilizado para a predição de valores de variáveis individuais⁵⁰. O *The International Schizophrenia Consortium*⁵¹ calculou um primeiro escore poligênico para esquizofrenia, o qual serviu como base para trabalhos com TEA^{52,53}.

Klei et al.⁵² mostrou que variantes genéticas comuns apresentam pequeno efeito quando consideradas de forma individual, mas agem de forma aditiva, sendo responsáveis por uma grande porção do risco de TEA. Clarke et al.⁵³ calculou um risco poligênico para TEA e TDAH com base em estudos de GWAS conduzidos pelo *Psychiatric Genetics Consortium*. Para TEA, os autores encontraram uma correlação positiva do risco poligênico com habilidade cognitiva geral, memória lógica e inteligência verbal.

1.4 Estudos de associação de copy number variations e sequenciamento completo do exoma em TEA - variantes de novo e muito raras

Como destacado por Cirulli e Goldstein²⁵, variantes raras são aquelas com MAF menor ou igual a 1%. As variantes raras aparecem enriquecidas para variantes deletérias e, portanto, funcionais³¹. Essas variantes raras são detectadas por estudos que analisam SCE e de microarranjos de DNA³¹ (do inglês, *deoxyribonucleic acid*). Os estudos utilizando arrays são sensíveis a grandes CNVs (CNVs entre 25 e 50 kbp possuem quase 100% de validação experimental quando testados com array contendo 2.7 milhões de sondas), enquanto estudos de SCE permitem maior especificidade e resolução⁵⁴.

Sebat et al.⁵⁵ testaram a hipótese de que CNVs *de novo* estavam associados com TEA. Através de hibridização genômica comparativa os autores confirmaram que CNVs *de novo* estão significativamente associados ao TEA (p 0,0005). A maioria deles eram CNVs pequenos, afetando diferentes regiões genômicas e em alguns casos um único gene. O trabalho de Sebat et al.⁵⁵ mostrou que mutações *de novo* eram um fator de risco mais importante do que previamente imaginado, impulsionando estudos do tipo para TEA.

Bucan et al.⁵⁶ realizaram genotipagem de alta densidade em 912 famílias multiplex e compararam os resultados com 1.488 controles. Foram encontrados mais de 150 loci com variantes raras em múltiplos probandos não relacionados, enquanto ausentes nos controles. Desses 150 loci, 27 foram confirmadas na análise de uma coorte de replicação independente com 859 casos e 1.051 controles. Os autores destacaram deleções exônicas no gene *NRXNI* e duplicações em genes da região 15q11-q13, incluindo *UBE3A*. O trabalho detectou CNVs adicionais nas regiões 16p11.2 e 22q11.21. Wiśniowiecka-Kowalik et al.⁵⁷ também encontraram CNVs em *NRXNI* em pacientes com diagnóstico de TEA ou que apresentavam características do espectro.

Glessner et al.⁵⁸ analisaram CNVs no genoma todo com uma coorte de 859 casos de TEA e 1409 crianças saudáveis de descendência europeia. Os achados positivos foram confirmados com uma coorte independente com 1.336 casos e 1.110 controles. Foi encontrada uma maior proporção de CNVs em genes como *NRXNI*, *CNTN4*, *NLGN1* e *ASTN2* nos casos ($p = 9.5 \times 10^{-3}$). CNVs em genes da via da ubiquitina como *UBE3A*, *PARK2*, *RFWD2* e *FBXO40* foram identificados apenas nos casos ($p = 3.3 \times 10^{-3}$). Esses CNVs afetam genes que participam na adesão celular neuronal e na degradação da ubiquitina, podendo contribuir para susceptibilidade ao TEA.

Pinto et al.⁵⁹ através de um estudo de microarray analisou 996 indivíduos com TEA de descendência europeia que foram comparados com 1.287 controles. Os autores descreveram uma maior carga de CNVs raras nos probandos com TEA (1.19 maior, $p = 0.012$), especialmente para loci previamente associados com TEA (1.69 maior, $p = 3.4 \times 10^{-4}$). Os CNVs detectados eram tanto *de novo* quanto herdados, sendo alguns dos genes afetados *SHANK2*, *SYNGAP1*, *DLGAP2* e no locus *DDX53-PTCHD1*. As vias afetadas envolviam proliferação celular, projeção e motilidade, além da sinalização da GTPase/Ras.

Willemsen et al.⁶⁰ avaliou quatro probandos do sexo masculino microdeleções *de novo* na região 16q24.3. Foi encontrada uma sobreposição de 90 kb para esses pacientes, afetando os genes *ANKRD11* e *ZNF778*. Esses pacientes apresentaram sinais característicos de TEA, assim como grau variável de déficit cognitivo, dimorfismos faciais e anormalidades cerebrais.

O estudo de O’Roak et al.⁶¹ foi o primeiro estudo de SCE em TEA. O SCE de 20 trios de casos esporádicos de TEA foram sequenciados, onde 21 mutações *de novo* foram encontradas, das quais 11 alteravam proteínas. Foram encontrados eventos de mutação *de novo* do tipo *missense* para TGM3, LAMC3, GPR139, SCN1A, SYNE1, SLC30A5, XIRP1, TLK2 e RBMS3; 3’ UTR para MYO1A, TAF1L, DCTN5; sinônima para PLCD1, IL1R2, AFF4, EPHB6, ARHGAP15, CHST5, TTN; 3’ splice para GRIN2B; e um *indel frameshift* para FOXP1. Esse trabalho mostrou o potencial do SCE para a identificação *de novos* genes candidatos para TEA, além de sugerir que as mutações *de novo* contribuem significativamente para a etiologia genética dos TEA.

Sanders et al.⁶² realizaram uma análise de CNVs raros em 1.124 famílias com TEA, sendo 872 quartetos com um probando afetado e um irmão não afetado e 252 trios. Foram encontradas associações significativas de TEA com duplicações *de novo* na região cromossômica 7q11.23, região na qual deleções equivalentes causam a síndrome de Williams-Beuren, caracterizada por uma grande sociabilidade. Também foram encontrados CNVs nas regiões 1q21.1 (duas duplicações), 15q13.2-q13.3 (uma deleção e uma duplicação), 16p13.2 (duas duplicações) e uma deleção de 5Mb com sobreposição de uma deleção exônica 34 kb interrompendo o gene *CDH13*.

Levy et al.⁶³ utilizaram os dados do Simons Simplex Collection⁶⁴ (SSC), que na época contava com aproximadamente 1000 famílias, para uma análise de CNVs. Foram identificados 61 CNVs *de novo* em genes nos probandos afetados, enquanto apenas nove

CNVs *de novo* estavam em genes nos irmãos não afetados. Foi vista também uma diferença significativa entre os sexos ($p = 0,05$), onde a mediana do número de genes afetados por eventos *de novo* (*single nucleotide variants* - SNVs e CNVs) em meninas foi de 15,5 e apenas 2,0 em meninos. Assim como Sanders et al.⁶² os autores também observaram eventos de duplicação na posição 7q11.23. Outros CNVs destacados estão nas posições 16p11.2, 16p13.2, 15q11.2 e 16p13.11.

Sanders et al.⁶⁵ utilizou o SCE de 928 indivíduos, sendo que 200 desses eram pares de irmãos com fenótipos distintos. Foram identificados 279 SNVs *de novo*, sendo que dois SNVs independentes (variante *nonsense*) afetaram o gene *SCN2A* em probandos não relacionados. Os autores demonstraram através das taxas de mutação em indivíduos não afetados que múltiplas mutações *de novo* independentes em um mesmo gene em diferentes probandos é uma maneira confiável de indicar alelos de risco.

Chahrour et al.⁶⁶ analisou o SCE de 16 probandos escolhidos pelo fato de seu genoma apresentar a maior proporção de homozigose (entre 5.431 indivíduos do grupo do *Autism Genetic Research Exchange*), e encontrou quatro genes candidatos (*UBE3B*, *CLTCL1*, *NCKAP5L* e *ZNF18*) para TEA. Esses genes codificam proteínas envolvidas na proteólise, sinalização mediada pela GTPase, organização do citoesqueleto, além de potenciais efeitos em neurônios, já que a despolarização neuronal regula a transcrição dos mesmos.

Iossifov et al.⁶⁷ utilizou o SCE de 343 famílias, cada uma com um filho com TEA e pelo menos um filho não afetado. Foram identificados um número de *indels de novo* e substituições pontuais, a maioria vindas da linhagem paterna e relacionadas com a idade. Não foram observadas grandes diferenças significativas no número de mutações *missense de novo* ao comparar crianças com TEA com as não afetadas, mas variantes do tipo *likely gene-disruptive* (LGD) são duas vezes mais frequentes, 59 nos casos e nos 28 controles.

Os autores estimaram entre 350 e 400 genes de susceptibilidade ao autismo, muitos deles relacionados com a síndrome do X-frágil, o que reforça a ligação do autismo com a plasticidade sináptica.

Neale et al.⁶⁸ sequenciou o SCE de 175 probandos com TEA e seus pais, com 161 mutações pontuais detectadas em regiões codificadoras. Através da análise de redes PPI (do inglês, *protein protein interaction*) das variantes *de novo* encontradas e 31 genes relacionados com sinapse previamente associados com TEA, demonstrou-se que esses genes são biologicamente relacionados uns com os outros e também com genes previamente identificados candidatos para TEA e/ou DI.

O’Roak et al.⁶⁹ utilizou o SCE de 677 indivíduos e encontrou 126 mutações *de novo*, em sua maioria de origem paterna (4:1) e positivamente correlacionadas com a idade paterna. Os autores mapearam 39% das mutações *de novo* mais severas para uma rede de proteínas de remodelamento Beta-catenina/cromatina altamente interconectada. Essa rede é ranqueada significativamente para genes candidatos do TEA.

Yu et al.⁷⁰ utilizou o SCE de três famílias parte do grupo do *Homozygosity Mapping Collaborative for Autism*, o qual tem como objetivo identificar famílias geneticamente informativas com TEA que sejam consanguíneas e/ou tenham vários indivíduos afetados. Foram encontradas mutações bialélicas nos genes AMT, PEX7, SYNE1, VPS13B, PAH e POMGNT1 relacionadas com TEA. Lim et al.⁷¹ utilizaram o SCE de 933 probandos e 869 controles e conseguiram identificar um aumento de duas vezes no número de *knockouts* completo em genes autossômicos com baixas taxas de variação *loss of function* (frequência menor ou igual a 5%). A contribuição desses achados para TEA foi estimada em 3%, sendo confirmada por um conjunto independente de 563 probandos e 4605 controles. Os resultados são evidência de que *knockouts* completos de

genes autossômicos e do cromossomo X são fatores de risco herdados importantes para TEA.

Poultney et al.⁷² utilizaram o programa *eXome Hidden Markov Model* para detectar CNVs exônicos raros pequenos (com tamanho entre 1 e 30 kb). Com uma coorte final de 559 indivíduos (299 casos de TEA e 260 controles) foram identificados 1.386 CNVs (803 nos casos e 583 nos controles). Os indivíduos com TEA apresentaram um número maior de CNVs do que os controles, ligados principalmente a um enriquecimento de deleções. Os autores destacaram os genes *ACTG1*, *MYH4*, *VILL*, *MYOM2*, *RPL27*, *RPL8*, *KRT6A*, *KRT6B*, *KRT5*, *KRT3*, *ACTB*, os quais estão envolvidos nas categorias de função molecular de atividade molecular estrutural (GO:0005198) e de componente estrutural do citoesqueleto (GO:0005200).

De Rubeis et al.⁷³ utilizou SCE em um estudo com 3.871 casos de TEA e 9.937 controles parentais ou com ascendência compartilhada, implicando 22 genes autossômicos com $FDR < 0,05$. Muitos desses genes codificam proteínas envolvidas em vias de formações sinápticas, regulação transcricional e remodelação da cromatina.

Iossifov et al.⁷⁴ em um estudo com a coorte do SSC mostraram através da comparação dos probandos afetados com irmãos não afetados que 13% das mutações *de novo missense* e 43% das mutações *de novo* provavelmente disruptivas contribuem para 12% e 9% dos diagnósticos, respectivamente. Em conjunto com CNVs, essa contribuição sobe para 30% de todos os casos simplex e 45% de todos os diagnósticos para meninas. Mutações provavelmente disruptivas em 400 genes contribuem para o grupo de meninas e meninos com baixo QI afetados, com uma sobreposição e número similar de genes vulneráveis a mutações *missense*.

Coe et al.⁷⁵ criaram um mapa de morbidade para CNVs utilizando 29.085 crianças com atraso de desenvolvimento, TEA ou DI, e 19.584 controles saudáveis, sendo

identificados 70 CNVs significativos. Após uma integração com dados de SNVs os autores chegaram a uma lista de 10 genes enriquecidos para mutações do tipo *loss of function*, incluindo ZMYND11 em indivíduos com autismo.

O estudo de Correia, Oliveira e Vicente⁷⁶ mostrou que as proteínas associadas ao TEA possuem uma interação direta maior do que o esperado por processos aleatórios. Através da intersecção de duas redes geradas com conjuntos de dados de TEA e seis de outras doenças não relacionadas, foram identificados 14 genes exclusivos das redes de TEA. Esses genes estão envolvidos em processos biológicos como adesão celular, organização do citoesqueleto e orientação axônica, processos esses relacionados com TEA⁷⁶. Chiocchetti et al.⁷⁷ através de uma análise de redes PPI indicou que as vias celulares responsáveis pela plasticidade sináptica no cérebro possuíam diferenças na expressão de componentes sensíveis a oxirredução em pacientes com TEA.

Toma et al.⁷⁸ realizaram um estudo de SCE com 10 famílias multiplex, sendo nove delas com dois probandos afetados e uma com três. As variantes selecionadas pelo estudo estão enriquecidas para genes envolvidos em funções neuronais ou previamente relacionados com doenças psiquiátricas. Os autores sugerem que variantes truncantes heterozigotas possuem um papel predominante na etiologia do TEA, sendo que o número dessas mutações que foram transmitidas é duas vezes maior do que as que não foram transmitidas. Através de análises de PPI os autores determinaram o gene *YWHAZ* como o nó mais conectado da rede, indicando ele e o gene *DRP2* como possíveis genes candidatos de TEA.

Krumm et al.⁷⁹ gerou um conjunto de SNVs exônicas e CNVs de 2.377 afetadas por TEA. Os probando apresentaram um enriquecimento de SNVs truncantes herdados em genes conservados quando comparados com irmãos não afetados. A análise detectou uma segunda classe de genes candidatos para TEA. Essa classe provavelmente não é

completamente penetrante, mas as mutações transmitidas em genes como *RIMS1*, *CUL7* e *LZTR1* podem predispor um indivíduo ao TEA, requerendo fatores genéticos ou não-genéticos adicionais para apresentar a doença.

Egawa et al.⁸⁰ sequenciou o exoma de seis indivíduos de duas famílias japonesas. Foram encontrados 6 SNVs, sendo que a variante *CLN8* R24H, gene relacionado a ceroidolipofuscinose neuronal, se destacou por ter sido herdada por três filhos com TEA de um pai com um transtorno global do desenvolvimento sem outra especificação. Os autores sugerem que *CLN8* R24H tem participação na etiologia genética do TEA, pelo menos para um subconjunto de pacientes.

O estudo de AlAyadhi et al.⁸¹, com 43 indivíduos, detectou 17 CNVs, sendo que 16 deles já estavam catalogados em algum banco de dados (Autism Database – AutDB, Autism Chromosome Rearrangement Database – ACRD, Database of Genomic Variants – DGV). Um CNV novo foi detectado na região 3q23 em 4 casos de TEA. Esse CNV causa a deleção do intron 2 do gene *CLSTN2*, o qual é exclusivamente expresso no cérebro.

Li et al.⁸² utilizou 17104 mutações *de novo* de 3.555 trios de quatro doenças psiquiátricas (TEA, encefalopatia epilética, DI e esquizofrenia), junto com irmãos não afetados como controle, vindos de 36 diferentes estudos de SCE e genoma completo. Foi encontrada uma maior frequência de mutações *de novo* nos probandos, as quais em conjunto com análise de Gene Ontology, alvos de microRNA e de fatores de transcrição revelou processos biológicos e elementos regulatórios não codificantes comuns para genes candidatos nestas doenças neuropsiquiátricas. Foram identificados 53 genes candidatos associados com mais de uma doença, sugerindo uma etiologia genética compartilhada. O gene *SCN2A* foi frequentemente encontrado nas quatro doenças. Os

autores elaboraram também um banco de dados, o NPdenovo, que contém uma lista de mutações *de novo* identificadas em doenças psiquiátricas.

Levando em consideração os trabalhos acima e demais na literatura, fica evidente que CNVs e SNVs ocorrem em conjunto no TEA^{75,83-85}. Apesar das variantes específicas não serem recorrentes, os genes afetados pelas mesmas e, principalmente, as vias das quais estes fazem parte, apresentam um nível significativo de recorrência⁸⁶. Luo et al.⁸⁶ avaliou cerca de 8 mil mutações vindas de dois datasets distintos e propôs um modelo “gene + pathway dual hit”. De acordo com este modelo as variantes afetam as funções gênicas relacionadas às vias e, quando ocorrem no mesmo gene, tendem a atingir resíduos próximos e no mesmo domínio, atingindo principalmente vias relacionadas com função sináptica, morfologia e plasticidade⁸⁶. As vias apontadas foram: Degradação de lisina (Lysine degradation - hsa00310), via de sinalização da MAPK (MAPK signaling pathway - hsa04010), via de sinalização da Wnt (Wnt signaling pathway - hsa04310), junções ocludentes (Tight junction - hsa04530), arrastamento circadiano (Circadian entrainment - hsa04713), sinapse glutamatérgica (Glutamatergic synapse - hsa04724), sinapse GABAérgica (GABAergic synapse - hsa04727), regulação do citoesqueleto de actina (Regulation of actin cytoskeleton - hsa04810), e digestão e absorção de proteínas (Protein digestion and absorption - hsa04974). Essas vias refletem aquelas exploradas previamente na literatura e destacadas como relevantes^{73,87-93}.

1.5 Classificações propostas para o Transtorno do Espectro Autista

Diante da heterogeneidade clínica e genética do TEA, há uma busca por marcadores biológicos de prognóstico e resposta ao tratamento buscando subgrupos de TEA. Essas classificações consideram aspectos como função cognitiva, trajetória de

desenvolvimento, severidade da apresentação clínica e outras propostas para o agrupamento de fenótipos do TEA⁹⁴.

Uma das classificações utilizada é a de TEA sindrômico (secundário) e TEA não-sindrômico (idiopático), uma distinção baseada exclusivamente em critérios clínicos⁹⁵. A classificação “não-sindrômico” refere-se a casos onde o TEA é o diagnóstico clínico primário e não parte de uma desordem mais complexa. O TEA sindrômico indica que este é parte de uma síndrome genética, onde uma parte significativa dos afetados desenvolvem sintomas característicos do TEA^{94,96}. Nesse caso há uma causa genética bem definida para a síndrome, mas a natureza dessa comorbidade não é precisa⁹⁶. Assim, os casos sindrômicos seriam quase “monogênicos” ou com variantes muito raras de grande efeito, enquanto os não-sindrômicos apresentariam um componente poligênico, ou seja, variações comuns de pequeno efeito individual, mas que agregadas trazem a manifestação clínica⁹⁶.

Outra tentativa neste sentido foi a ligação do fenótipo sexo com variantes genéticas específicas. Há na literatura suporte para uma diferença genética do TEA em relação aos sexos. O TEA, assim como outras desordens do neurodesenvolvimento, é mais prevalente em indivíduos do sexo masculino^{97,98}, mas os mecanismos genéticos responsáveis por essa diferença não são completamente entendidos^{94,99}.

Jacquemont et al.⁹⁷ com uma coorte de 762 famílias com TEA encontraram um aumento de três vezes no número de CNVs autossômicos deletérios ($p = 7 \times 10^{-4}$) e um excesso SNVs deletérios ($OR = 1.43$, $p = 0,03$) em probandos do sexo feminino comparados com o sexo masculino. Os probandos do sexo feminino apresentaram resultados menores nos testes cognitivos, mas mesmo após correção a carga mutacional excessiva foi mantida⁹⁷. Os autores sugerem que o cérebro masculino precisaria de alterações mais leves para apresentar o TEA. Isso poderia ser explicado pelo “efeito

protetor feminino” (do inglês, *female protective effect* - FPE). De acordo com essa teoria, o probando médio do sexo feminino, quando comparado com um do sexo masculino, teria um quadro clínico mais severo, maior número de variantes e uma maior quantidade de fatores de risco para o TEA, necessário para superar o FPE^{97,100}, o que já havia sido sugerido por outros autores^{62,73,74}. No entanto, estudos epidemiológicos mostram que dado que uma família tem um primeiro filho com TEA, a chance de ter um segundo filho afetado é maior se o primeiro for do sexo masculino¹⁰¹.

Outra visão foi apresentada por Werling, Parikshak e Geschwind¹⁰². Os autores testaram se genes associados a TEA estavam enriquecidos em genes com expressão diferencial entre os sexos. Não foi identificado um enriquecimento para esses genes, mas observaram nos padrões masculinos de expressão um viés para genes regulados positivamente em cérebros de probandos com TEA em análises *post-mortem*, como genes de microglia e astrócitos. Dessa forma, a responsável pelo viés na prevalência não seria a regulação diferencial do sexo nos genes de risco do TEA, mas sim processos dimórficos sexuais que ocorrem naturalmente, potencialmente envolvendo interações neurônio-glia, que modulam o impacto das variantes de risco¹⁰². Assim apesar do avanço no entendimento da heterogeneidade genética, ainda não foi possível usar esta informação na busca de subtipos fenotípicos que façam sentido quanto ao prognóstico e ou resposta terapêutica.

Mas a classificação mais recorrente é a de TEA simplex e TEA multiplex, também conhecidos como TEA esporádico e TEA familiar, respectivamente. O TEA simplex é definido quando há apenas um indivíduo afetado na família, enquanto o multiplex engloba famílias com dois ou mais indivíduos afetados¹⁰³. De acordo com a literatura, inicialmente acreditava-se que as contribuições das variantes comuns e raras, assim como o efeito do ambiente, seria diferente entre esses dois tipos^{52,103}. Na parte fenotípica os pais e irmãos

de probandos com TEA simplex apresentariam melhores habilidades sociais e melhor comunicação não-verbal do que suas contrapartes de probandos com TEA multiplex^{52,104}.

Inicialmente acreditava-se que o TEA simplex era devido a um número de variantes raras de grande efeito, enquanto o TEA multiplex seria consequência da presença de diversos alelos comuns, os quais individualmente não teriam grandes efeitos, mas agiriam de forma aditiva¹⁰³. Entretanto, essa justificativa é questionada. Polimorfismos comuns também contribuem para o TEA simplex⁵², além da importância de ambos os tipos de variantes terem sido ressaltadas na literatura quanto ao papel na susceptibilidade ao TEA^{30,73,105,106}.

1.6 Análise composicional

Além da contribuição vinda em conjunto das variantes raras e comuns, um fator importante é o tipo de impacto funcional causado pela mesma. Shohat, Ben-David e Shifman¹⁰⁷ exploraram variações *de novo missense*, LGD e sinônimas em uma comparação entre TEA, DI e esquizofrenia. Os autores apontam que tanto variantes LGD quanto variantes *missense* são importantes, sendo que as conclusões de diversos estudos seriam devido ao tipo de variante explorado. Vias diferentes teriam tolerância diferentes a cada tipo de mutação, portanto apenas considerando todos os tipos seria possível ter uma visão completa dos processos afetados e sua relação com o fenótipo¹⁰⁷.

Uma maneira de explorar esses dados é através da análise composicional. A análise composicional é caracterizada por um formato específico de dados, onde seus componentes são não negativos e fazem parte de um inteiro, formando o espaço simplex. Essa restrição aplicada aos dados exige metodologia apropriada para sua análise ou a adequação de demais métodos¹⁰⁸. Uma das maneiras de trabalhar com dados composicionais com ferramentas padrão é a transformação dos mesmo através da razão

logarítmica centralizada^{108,109}, no entanto perde-se a visão composicional. A ausência de ferramentas apropriadas ou uso dos dados de maneira inadequada pode levar a inferências errôneas¹⁰⁸.

Em muitos dos problemas nos quais análise composicional é aplicada, a composição é a entidade relevante. Ela torna possível a comparação entre objetos fundamentalmente similares, mas em escalas diferentes¹⁰⁸. Um exemplo é a comparação de dados provenientes de SCE, onde a profundidade é essencial para comparação. Através da transformação dos dados em uma composição é possível comparar indivíduos com profundidades de análise diferentes, uma vez que qualquer propriedade escolhida será analisada tendo em mente sua proporção em relação ao total de cada indivíduo.

Problemas de diversas áreas já foram explorados através da análise composicional¹⁰⁸: agricultura, biologia do desenvolvimento, economia, geologia, paleontologia, entre outros. No entanto, seu uso em pesquisa na saúde é limitado. Fernandes et al.¹¹⁰ usou análise composicional para combinar datasets de RNA-seq e sequenciamento de rRNA 16S do *Human Microbiome Project*. Carson et al.¹¹¹, Dumuid et al.¹¹², Fairclough et al.¹¹³ e Foley et al.¹¹⁴ utilizaram análise composicional para analisar padrões de comportamento e uso de tempo em relação à atividade física. Tsilimigras e Fodor¹⁰⁹ exploraram a análise composicional como ferramenta aplicada ao estudo do microbioma humano. No entanto, a abordagem composicional ainda não foi aplicada a dados de SCE.

1.7. Hipótese do estudo

Os dois principais fatores ressaltados pela literatura recente do TEA são o maior grau de recorrência de variações nas vias associadas e a importância de levar em consideração os diferentes impactos funcionais das variantes de maneira combinada.

Diante destes dois fatores, nossa hipótese é que a utilização da análise composicional para explorar dados de SCE na busca de agrupamentos através da composição genética dos probandos poderá expandir o conhecimento entre a relação de fenótipos e composição genética. Para testar essa hipótese, avaliamos as variantes genéticas *de novo missense*, *de novo* LGD, *missense* herdadas raras e LGD herdadas raras nos genes das vias selecionadas por Luo et al.⁸⁶. Após a contagem os dados foram transformados em composições e explorados através de análise de agrupamentos hierárquica. Os agrupamentos estratificados foram avaliados quanto a diferenças fenotípicas para QI e outras medidas relevantes para o TEA.

2 Objetivos

2.1 Objetivo Principal

Estabelecer se há relação entre composição genética e fenótipo para probandos do Transtorno do Espectro Autista.

2.2 Objetivos Específicos

- a) Verificar se é possível estratificar grupos estáveis de probandos através da composição genética.
- b) Explorar se existe relação entre gravidade da apresentação e composição genética para probandos do Transtorno do Espectro Autista.

3 Métodos

3.1. Amostra

Os SCE utilizados na nossa análise são provenientes do SSC do SFARI Base (base de dados da Simons Foundation for Autism Research Initiative). A amostra SSC consiste em dados fenotípicos de 2.858 probandos e dados de SCE de 2.389 probandos, sendo que 2.382 possuíam ambos os dados. O próximo passo foi mantermos apenas os probandos com identificação de sexo e faixa etária. Os dados não possuem identificação de idade, mas estão divididos em duas faixas etárias, de 2 a 5 anos e de 6 a 18 anos, para aplicação da escala *Child Behavior Checklist*¹¹⁵ (CBCL). Após filtragem por estes dois critérios nossa amostra é composta por 2.313 probandos, 321 do sexo feminino e 1.992 do sexo masculino, com 1.762 probandos na faixa de 6 a 18 anos e 535 na faixa de 2 a 5 anos. A caracterização da amostra está descrita no **Anexo A**.

3.2 Processamento de dados de sequenciamento completo do exoma

Os arquivos de chamada de variante utilizados foram disponibilizados através do sistema Globus pela equipe do SSC na versão de montagem do genoma hg19. Esses arquivos foram gerados com o uso do programa Genome Analysis Toolkit¹¹⁶ (GATK) versão 3.8. Foram aplicados os seguintes critérios adicionais de filtragem para manutenção apenas das variantes de alta qualidade:

1. Filtro (Filter): Foram filtradas as variantes que apresentavam qualquer anotação além de “PASS”.
2. Cobertura (DP): Foram filtradas as variantes com cobertura menor do que 20 vezes, isto é, todas aquelas com menos de 20 reads cobrindo o alelo. A taxa de cobertura é similar ou maior do que as obtidas em outros estudos^{65,67,68}.

3. Qualidade por Profundidade (QD): Foram filtradas as variantes com QD abaixo de 2. O QD é obtido através da divisão da confiabilidade da variante (QUAL) pela profundidade não filtrada de amostras não homozigóticas para a referência. Essa anotação é utilizada para normalizar a qualidade das variantes, evitando um possível excesso quando há excesso de cobertura.
4. Fisher Strand (FS): Foram filtradas as variantes com FS acima de 60. FS indica a probabilidade em escala de Phred de *strand bias* em uma determinada posição. Strand bias é um viés de sequenciamento onde uma das fitas de DNA é favorecida em relação a outra, o que proporciona uma avaliação incorreta na quantidade de alelos.
5. Valor Eficaz da Qualidade do Mapeamento (MQ): Foram filtradas as variantes com MQ abaixo de 40. MQ é a qualidade média quadrada do mapeamento de todos os *reads* daquela posição.
6. Rank Sum Test da Qualidade de Mapeamento (MQRankSum): Foram filtradas as variantes com MQRankSum menor do que - 12,5. O MQRankSum compara a qualidade de mapeamento dos *reads* do alelo referência com os do alelo alternativo. Um valor próximo a zero indica que há pouca diferença entre as qualidades de mapeamento.
7. Rank Sum Test da Posição do Read (ReadPosRankSum): Foram filtradas as variantes com ReadPosRankSum menor do que -8. O ReadPosRankSum compara se as posições dos alelos de referência e alternativa são diferentes nos *reads*. Um valor próximo de zero é melhor porque indica que há pouca diferença entre as posições dos alelos de referência e alternados nas leituras.

Uma profundidade mínima de 20 leituras de base foi utilizada para cada membro do trio. A chamada de variantes *de novo* foi feita através da ferramenta PossibleDeNovo do GATK¹¹⁶. Foram consideradas como *de novo* somente aquelas variantes apontadas pelo PossibleDeNovo como de alta confiabilidade e que os genótipos seguiram os seguintes critérios:

- Os pais foram considerados homozigotos verdadeiros para a variante de interesse quando o alelo com a menor contagem não excedeu 5% da contagem total de leituras do loci. Caso contrário serão considerados heterozigotos.
- Para probandos o mesmo critério foi aplicado quando avaliando homozigotos. No entanto, se a possível variante *de novo* fosse heterozigota, o alelo com a menor contagem deveria corresponder a pelo menos 20% da contagem total de leituras do loci.

Após a filtragem e chamada de variantes *de novo*, os VCFs foram anotados com o uso do software InterVar¹¹⁷. O InterVar realiza uma classificação semiautomática de variantes através das diretrizes da ACMG-AMP¹¹⁸ e de informações fornecidas pelo usuário. A classificação de variantes pelas diretrizes e padrões da ACMG/AMP/CAP é feita a partir da combinação de uma lista de critérios divididos em diferentes níveis¹¹⁸. O InterVar permite uma classificação rápida utilizando diversas fontes, desde bancos de dados como o ClinVar¹¹⁹, algoritmos computacionais como SIFT¹²⁰ e PolyPhen-2¹²¹, e informações curadas manualmente pelos autores¹¹⁷.

3.3 Processamento dos VCFs anotados

Os VCFs anotados foram processados utilizando o R¹²² (versão 3.4.4). Cada probando teve suas variantes filtradas, mantendo apenas aquelas nos genes dentro das 9

vias selecionadas para a amostra do SSC por Luo et al.⁸⁶. O pacote biomaRt^{123,124} (versão 2.32.1) foi utilizado para obter os genes das vias, totalizando 908 genes. Os detalhes específicos de cada via estão descritos na **Tabela 1**.

Tabela 1 – Nome, descritor hsa e número de genes obtidos através do pacote biomaRt para cada uma das vias indicadas por Luo et al.⁸⁶.

Nome	Descritor hsa	Número de genes
Degradação de lisina	hsa00310	44
Via de sinalização da MAPK	hsa04010	268
Via de sinalização da Wnt	hsa04310	150
Junções ocludentes	hsa04530	132
Arrastamento circadiano	hsa04713	96
Sinapse glutamatérgica	hsa04724	114
Sinapse GABAérgica	hsa04727	89
Regulação do citoesqueleto de actina	hsa04810	213
Digestão e absorção de proteínas	hsa04974	81

Após essa seleção mantivemos apenas as variantes em regiões anotadas como *exonic*, *exonic;splicing* ou *splicing*. No passo seguinte mantivemos apenas as variantes classificadas como *missense* (SNVs não sinônimo, inserções não-*frameshift* e deleções não-*frameshift*) e LGD (deleções *frameshift*, inserções *frameshift*, perda de códon de parada, ganho de códon de parada, *splicing*) de acordo com anotação de impacto funcional. O último passo foi a filtragem de todas as variantes com MAF maior do que 0,01 para qualquer um dos bancos populacionais usados pelo InterVar¹²⁵ (gnomAD), 1000 Genomes¹²⁶ e ESP6500¹²⁷. Também foi adicionada uma coluna para indicar se a variante foi apontada como *de novo* pelo GATK e sobreviveu nosso critério de filtragem.

3.4 Análise composicional

Usando os VCFs processados, criamos uma tabela com a contagem de variantes para cada probando. As variantes foram atribuídas a uma das quatro classes: *de novo* LGD, *de novo missense*, LGD e *missense*. Os dados observados são, portanto, o número de ocorrências de cada classe de variação, um vetor quadridimensional. O perfil do probando (X) é modelado como uma distribuição de probabilidade Multinomial 4-dimensional com proporções de cada classe mutacional sendo o parâmetro de interesse: $\pi = (p_1, p_2, p_3, p_4)^{108}$. Na notação estatística padrão: $X_j \sim M(\pi_j | N_j)$ onde N_j é a quantidade total de mutações do j-ésimo probando. A distribuição não-informativa de densidade de probabilidade uniforme é usada como priori e procedimentos padrões de Inferência Bayesiana levam a uma distribuição posterior de Dirichlet. Os procedimentos de aprendizagem de máquina não supervisionados são então realizados com base nas estimativas da probabilidade máxima a posteriori (em inglês, *maximum a posteriori probability* - MAP). Isto é, as proporções de quatro partes p apresentam o maior valor de densidade de Dirichlet. Estimativas MAP produzem objetos não nulos normalizados que são instâncias particulares de objetos de espaço simplex¹⁰⁸. As distâncias entre os probandos são calculadas de acordo com a estrutura de análise de dados de composição da Aitchison, adequada para vetores simplex após a transformação de razão logarítmica centralizada (CLR)¹⁰⁸.

3.5 Análise de agrupamentos

A análise de agrupamentos foi realizada na tabela de composição transformada utilizando CLR através de uma análise de agrupamento hierárquica com o critério do método de Ward aplicado. Para consolidar os agrupamentos, calculamos as estimativas de nível de evidência baseadas em bootstrap usando uma abordagem de reamostragem

multiescala implementada no pacote pvclust (versão 2.0)¹²⁸. Os níveis de evidência de coesão do agrupamento são compilados em um valor denominado AU (do inglês, *Approximately Unbiased*). Um valor AU acima de 0,99 em 10.000 rodadas de reamostragem é considerado como suporte à validade do agrupamento.

3.6 Avaliação fenotípica dos agrupamentos estratificados

Parte da literatura indica que o QI mais baixo no TEA está correlacionado com o aumento da gravidade dos sintomas comportamentais e um aumento na taxa de comorbidade epiléptica e outros sintomas neurológicos^{7,129–131}. Robinson et al.¹³² compararam pacientes com QI alto e baixo, mostrando uma associação entre QI e diferenças na arquitetura genética. Assim, iniciamos nossa análise comparativa dos agrupamentos estratificados com a medição de QI disponibilizada pelo SSC, mas também incluímos outras medidas fenotípicas importantes para caracterizar o TEA, como *Autism Behavior Checklist*¹³³ (ABC), *Autism Diagnostic Interview-Revised*¹³⁴ (ADI-R), CBCL¹¹⁵, *Repetitive Behavior Scale-Revised*¹³⁵ (RBS-R) e Vineland II¹³⁶. Detalhes das subescalas utilizadas e informações quanto a interpretação de cada uma estão disponíveis no **Anexo B**.

3.7 Aprovação do comitê de ética

Este estudo foi aprovado pelo comitê de ética em pesquisa do Hospital das Clínicas da Faculdade de Medicina (CAPPesq) através do parecer número 2.647.063. É um subprojeto do projeto “Estudo da dosagem gênica na contribuição em fenótipos de transtornos do espectro do autismo”, número CAAE 57067016.2.0000.0068, parecer número 1.637.312.

4 Resultados

4.1 Processamento de dados e tamanho da amostra

Na **Tabela 2** descrevemos o número total de variantes antes e depois do nosso processo de filtragem.

Tabela 2 - Medições gerais detalhando os resultados dos critérios de filtragem aplicados aos arquivos VCF. Os números são relativos às contagens de variantes.

	Média (DP)	Mediana	Máximo	Mínimo
Antes da filtragem	48.415,03 (6.173,17)	49.043	65.109	28.010
Após a filtragem	27.766,43 (4.531,74)	28.251	45.125	10.486

Após a anotação com o PossibleDeNovo e filtragem de acordo com os nossos critérios, 16 probandos tinham mais de 111 variantes *de novo*, ou até mais, se considerarmos apenas os critérios de alta confiança do GATK. Estes probandos foram removidos da nossa análise, uma vez que o número de variantes *de novo* excedeu o que é esperado com base na literatura⁵⁴. A **Tabela 3** detalha nossas descobertas em relação às variantes após a manutenção apenas daquelas, após a filtragem e o processo de seleção detalhado na seção **3.2**.

Tabela 3 - Informações gerais sobre as variantes encontradas após a filtragem e processamento dos VCFs.

	Média (DP)	Mediana	Primeiro quartil	Terceiro quartil
<i>de novo</i> LGD	0,008 ($\pm$ 0,09)	0	0	0
<i>de novo</i> missense	0,036 ($\pm$ 0,19)	0	0	0
LGD	1,026 ($\pm$ 1,02)	1	0	1
<i>missense</i>	19,14 ($\pm$ 7,10)	19	14	23

Conforme indicado na **Tabela 3**, os valores para as variantes *de novo* são pequenos quando comparados com os outros tipos de variantes. Isto é devido ao baixo número de variantes *de novo* na amostra após a filtragem. Apenas 81 probandos tinham

variantes *de novo missense*, enquanto 20 probandos tinham variantes *de novo* LGD e apenas um probando tinha ambos os tipos de variantes.

4.2 Abordagem Bimodal - Análise de composição, análise de agrupamentos e comparação de fenótipos

Dados composicionais possuem uma limitação na forma com que lidam com componentes ausentes. A primeira abordagem que utilizamos para lidar com essa limitação foi utilizar uma priori bimodal através da adição de valor infinitesimal (10^{-20}) a todos os valores antes de transformá-los em uma composição. Os valores completos para todos os probandos estão disponíveis na **Tabela Digital 1**. Para explorar os agrupamentos de pacientes de acordo com sua composição de variantes, realizamos a análise de agrupamentos hierárquica com o método Ward D. A tabela de composição foi transformada utilizando CLR e analisada com o pacote pvclust. Usando um valor AU de 0,99 obtivemos 6 agrupamentos estáveis (**Figura 1** - disponível em alta resolução em http://bit.ly/dissertacao_vinicius_gastaldi).

Figura 1 - Gráfico de agrupamento hierárquico mostrando os agrupamentos obtidos com a abordagem bimodal e um valor AU mínimo de 0,99. As linhas vermelhas indicam os agrupamentos estáveis.

Para calcular os valores de composição para cada agrupamento, somamos as contagens de variantes de cada probando que era membro do grupo e calculamos a

composição. Estas composições refletem os valores reais, uma vez que eles foram calculados sem a adição do valor infinitesimal. Os seis agrupamentos estão resumidos na **Tabela 4**. Os agrupamentos foram nomeados de acordo com o impacto variante assumido proposto por He et al.¹³⁷: *de novo* LGD > *de novo* missense > LGD transmitido > *missense* transmitido. Os valores de composição para cada probando individual e a informação a qual agrupamento pertencem está disponível na **Tabela Digital 2**.

Tabela 4 - Resumo dos agrupamentos estáveis (AU> = 0,99) estratificados usando agrupamento hierárquico nos dados de composição transformados pelo CLR (priori bimodal).

Agrupamento	<i>de novo</i> LGD	<i>de novo</i> missense	LGD	<i>missense</i>	Número de probandos
A1	0,044	0,007	0	0,949	6
A2	0,043	0	0,052	0,905	14
A3	0	0,049	0	0,951	27
A4	0	0,043	0,064	0,893	53
A5	0	0	0,069	0,931	1520
A6	0	0	0	1	677

Usando os seis agrupamentos, fizemos um boxplot com combinado com o pacote beeswarm¹³⁸ (versão 0.2.3) usando os valores de QI em escala total fornecidos pelo SSC (**Gráfico 1**). Embora não tenham sido observadas diferenças estatisticamente significativas na comparação entre os seis agrupamentos, a ordem dos mesmos está próxima do que seria esperado quando consideramos o impacto assumido das variantes na codificação de proteínas. As exceções são o agrupamento A2 com um QI mediano menor que A1 e o agrupamento A6 menor que A5 e igual a A4.

Para melhor explorar a relação entre composição e fenótipo, comparamos a distribuição cumulativa dos escores dos probandos para QI, ADI-R, Vineland, RBS, CBCL, ABC, de acordo com os seis agrupamentos. As diferenças para as distribuições

foram testadas usando um teste Kolmogorov-Smirnov de dois lados de uma amostra ($p \leq 0,05$) como implementado no pacote KSgeneral¹³⁹ (versão 0.1.1) onde $F(x)$ é a função de distribuição cumulativa discreta (cdf) calculada usando toda a amostra menos o agrupamento usado na comparação (por exemplo, agrupamento A1 vs cdf(A2 + A3 + A4 + A5 + A6)). Através desta abordagem, encontramos uma diferença significativa para o QI Verbal quando comparamos o agrupamento A5 com a distribuição dos demais probandos ($p = 0,028$; **Gráfico 2**). As distribuições são muito próximas, mas os valores para o agrupamento A5 estão um pouco mais a direita, sendo tecnicamente melhores. Observamos também diferenças estatisticamente significativas ($p \leq 0,05$) para as seguintes escalas e medidas: ABC (Subescalas I e V), ADI-R (A Total), CBCL (Comportamentos externalizantes e internalizantes), RBS (subescalas I, II, III, V) e Vineland II (sociabilidade). Os detalhes das comparações estão resumidos na **Tabela 5** e os gráficos para as demais medidas nos **Anexos C, D, E, F e G**.

Os agrupamentos A1 e A4 não apresentaram diferenças significativas quando comparados com o restante da amostra para os fenótipos avaliados, apesar de o agrupamento A1 ter supostamente os fenótipos com maior comprometimento. Os demais agrupamentos não seguiram um padrão definido entre composição e fenótipo esperado, com exceção do agrupamento A2 que apresentou resultados inferiores ao restante da amostra. No entanto, A2 apresentou resultado significativo para apenas uma das subescalas.

Gráfico 1 - Boxplot com beeswarm para os valores de QI em escala total dos seis agrupamentos identificados usando agrupamento hierárquico nos valores de composição transformados pelo CLR da abordagem bimodal. Ordenado pela mediana do QI, do menor para o maior: A2 - 71,5; A1 - 73,5; A3 - 83,5; A4 - 85; A6 - 85; A5 - 86.

Gráfico 2 - Distribuição cumulativa do agrupamento A5 (ciano) e do restante da amostra (amarelo) para a subescala de Sociabilidade da Vineland II. p do Teste de Kolmogorov-Smirnov de dois lados de uma amostra igual a 0,028.

Levando esses resultados em consideração, parece que apresentar maior quantidade de variações ditas prejudiciais não se correlaciona diretamente com os piores fenótipos. A presença ou ausência de tipos de variação não explica os fenótipos, o que pode indicar que o achatamento de dados resultante da abordagem bimodal pode esconder relações importantes entre composição e fenótipos. Por essa razão, usamos uma segunda abordagem com uma priori uniforme para a análise composicional.

4.3 - Abordagem com priori uniforme - Análise de composição, análise de agrupamentos e comparação de fenótipos

A segunda abordagem foi o uso de um priori uniforme (com a adição de um valor 1) em conjunto a análise composicional. Os valores completos para todos os probandos estão disponíveis na **Tabela Digital 3**. A tabela transformada CLR foi submetida à análise de agrupamentos hierárquicos usando o pacote. Usando um valor AU de 0,99 obtivemos 29 agrupamentos estáveis (**Figura 2** – disponível em alta resolução em http://bit.ly/dissertacao_vinicius_gastaldi). No entanto, dois probandos não faziam parte de nenhum agrupamento estável e foram removidos da análise.

Os 29 agrupamentos estão resumidos na **Tabela 6**, ordenada de acordo com o impacto esperado da carga de variação, conforme descrito por He et al.¹³⁷. Assim como na **Tabela 4**, os valores de composição para cada agrupamento foram calculados através da soma das contagens de variantes de cada probando que era membro do grupo e transformados em composição, sem adição da priori uniforme. As cores da **Tabela 6** indicam a qual agrupamento produzido com a priori bimodal cada um dos novos agrupamentos pertence. As informações sobre a associação de cada probando dentro dos agrupamentos estratificados e seus valores de composição individual estão disponíveis na **Tabela Digital 2**.

Tabela 5 - Comparações entre agrupamentos individuais obtidos com uso da priori bimodal e amostra total (amostra total refere-se ao total menos o agrupamento em questão). A interpretação de “melhor” e “pior” utiliza os critérios descritos no **Anexo B** e foi baseada nas observações feitas nos gráficos da **Gráfico 2** e dos **Anexos C, D, E, F e G**. Todos os agrupamentos na tabela apresentam $p \leq 0,05$.

Subescala	Melhor em relação ao restante da amostra	Pior em relação ao restante da amostra
QI - Verbal	A5	-
ABC - Subescala I - Irritabilidade	A5	A3
ABC - Subescala V - Fala Inapropriada	-	A5
ADI-R - A Total	A5	A6
CBCL - Problemas Externalizantes	A5	A3
CBCL - Problemas Internalizantes	-	A3
RBS-R - Subescala I - Comportamento Estereotipado	-	A5
RBS-R - Subescala II - Comportamento Auto Agressivo	A6	A5
RBS-R - Subescala III - Comportamento Compulsivo	A5	A3
RBS-R - Subescala V - Comportamento de Insistência na Mesmice	-	A3
Vineland II - Sociabilidade	-	A2

Usando esses 29 agrupamentos, fizemos um boxplot com o beeswarm para a distribuição dos valores de QI em escala total fornecidos pelo SSC (**Gráfico 3**). Os agrupamentos foram ordenados na mesma ordem do **Gráfico 1**, isto é, conforme a mediana do QI dos 6 agrupamentos da priori bimodal: A2, A1, A3, A4, A6 e A5. Dentro destes os agrupamentos foram ordenados pela sua mediana de QI.

Figura 2 - Gráfico de agrupamento hierárquico mostrando os agrupamentos obtidos com a abordagem com priori uniforme e um valor AU mínimo de 0,99. As linhas vermelhas indicam os agrupamentos estáveis.

Nesta análise, podemos observar que uma visão composicional mais precisa e com maior variabilidade entre os agrupamentos (**Anexos H, I, J, K, L, M, N, O e P**). Organizando os 29 agrupamentos pelos seis agrupamentos originais, as medianas de QI refletiram a ordem dos seis agrupamentos da priori bimodal. Podemos observar que indivíduos com LGD *de novo* e *missense de novo* tendem a ter QI mais baixo, mas a composição da variante e a associação de QI não é direta.

Assim como feito para a abordagem bimodal, comparamos a distribuição cumulativa dos escores dos probandos para as medidas fenotípicas de QI, ADI-R, Vineland, RBS, CBCL, ABC. As diferenças para as distribuições foram testadas usando um teste Kolmogorov-Smirnov de dois lados de uma amostra ($p \leq 0,05$) como implementado no pacote KSgeneral onde $F(x)$ é a função de distribuição cumulativa discreta (cdf) calculada usando toda a amostra menos o agrupamento usado na comparação (por exemplo, agrupamento B1 vs $\text{cdf}(B2 + B3 + \dots + B29)$).

Tabela 6 - Resumo dos agrupamentos estáveis ($AU > = 0,99$) estratificados usando agrupamento hierárquico nos dados de composição transformados pelo CLR (priori uniforme).

Agrupamento	<i>de novo</i> LGD	<i>de novo</i> missense	LGD	<i>missense</i>	Mediana QI	Agrupamento original	Número de membros
B1	0,04	0	0,05	0,91	71,5	A2	14
B2	0,04	0	0	0,96	73,5	A1	4
B3	0	0,09	0	0,91	101	A3	6
B4	0	0,06	0,06	0,88	90,5	A4	12
B5	0	0,04	0,09	0,86	85	A4	21
B6	0	0,04	0,04	0,92	81,5	A4	20
B7	0	0,04	0	0,96	82	A3	21
B8	0	0	0,52	0,48	93	A5	33
B9	0	0	0,46	0,54	84,5	A5	2
B10	0	0	0,40	0,60	108	A5	2
B11	0	0	0,35	0,65	34	A5	5
B12	0	0	0,18	0,82	89	A5	7
B13	0	0	0,17	0,83	87	A5	37
B14	0	0	0,16	0,84	67,5	A5	6
B15	0	0	0,13	0,87	85	A5	35
B16	0	0	0,13	0,88	92	A5	29
B17	0	0	0,11	0,89	84	A5	21
B18	0	0	0,11	0,89	87	A5	63
B19	0	0	0,10	0,90	77	A5	13
B20	0	0	0,08	0,917	85	A5	150
B21	0	0	0,08	0,924	86	A5	254
B22	0	0	0,06	0,94	87	A5	334
B23	0	0	0,04	0,96	84	A5	529

continua

B24	0	0	0	1	78	A6	117
B25	0	0	0	1	80	A6	169
B26	0	0	0	1	85	A6	137
B27	0	0	0	1	87	A6	68
B28	0	0	0	1	89	A6	116
B29	0	0	0	1	97	A6	70

Gráfico 3 - Boxplot com beeswarm para os valores de QI em escala total dos 29 agrupamentos identificados usando agrupamento hierárquico nos valores de composição transformados pelo CLR da abordagem com uniforme priori.

Na análise utilizando os agrupamentos da priori uniforme observamos diferenças significativas ($p \leq 0,05$) para todas as subescalas, exceto para as subescalas IV (escore de comportamento ritualístico) e VI (escore de comportamento restrito) da RBS-R e o escore padrão da soma da Vineland II. Os resultados estão descritos na **Tabela 7**. Na sequência de cada agrupamento que apresentou uma diferença significativa, indicaremos

sua participação em relação aos agrupamentos identificados na abordagem da priori bimodal dentro de parênteses

Tabela 7 - Comparações entre agrupamentos individuais obtidos com uso da priori uniforme e amostra total (amostra total refere-se ao total menos o agrupamento em questão). A interpretação de “melhor” e “pior” utiliza os critérios descritos no **Anexo B** e foi baseada nas observações feitas nos gráficos dos **Anexos H, I, J, K, L, M, N, O e P**. Todos os agrupamentos na tabela apresentam $p \leq 0,05$.

Subescala	Melhor em relação ao restante da amostra	Pior em relação ao restante da amostra
QI - Total	B13 (A5), B29 (A6)	B11 (A5), B24 (A6)
QI - Não Verbal	B8 (A5), B29 (A6)	B11 (A5), B24 (A6)
QI - Verbal	B13 (A5), B27 (A6)	B11 (A5), B24 (A6)
ABC - Subescala I - Irritabilidade	-	B7 (A3)
ABC - Subescala II - Letargia	B23 (A5), B27 (A6)	-
ABC - Subescala III - Estereotipia	B27 (A6)	B22 (A5), B24 (A6)
ABC - Subescala IV - Hiperatividade	B5 (A4)	-
ABC - Subescala V - Fala Inapropriada	-	B20 (A5)
ADI-R - A Total	-	B27 (A6)
ADI-R - B Não Verbal Total	B13 (A5)	-
ADI-R - B Verbal Total	B13 (A5)	-
ADI-R - C Total	-	B14 (A5)
CBCL - Comportamentos Externalizantes	B15 (A5)	B4 (A4), B7 (A3)
CBCL - Comportamentos Internalizantes	-	B7 (A3), B18 (A5)
RBS-R - Subescala I - Comportamento Estereotipado	B27 (A6)	B17 (A5)
RBS-R - Subescala II - Comportamento Auto Agressivo	B7 (A3), B23 (A5), B25 (A6)	B11 (A5)

continua

RBS-R - Subescala III - Comportamento Compulsivo	-	B3 (A3), B24 (A6)
RBS-R - Subescala V - Comportamento de Insistência na Mesmice	-	B7 (A3)
RBS-R - Escore Total	B12 (A5)	-
Vineland II - Autonomia Diária	B24 (A6)	-
Vineland II - Comunicação	B11 (A5), B24 (A6)	-
Vineland II - Sociabilidade	B1 (A2), B24 (A6)	-

A maioria dos agrupamentos (68,9%) apresentou diferença significativa para algum dos comportamentos avaliados. No entanto, uma associação clara entre composição e fenótipos não é possível. O agrupamento B2, originado do agrupamento A1 e com ocorrência de *de novo* LGD, não apresentou diferenças significativas em nenhuma das comparações, sendo que inicialmente esperávamos que fosse o pior agrupamento na questão fenotípica. O agrupamento B1, outro agrupamento com *de novo* LGD, apresentou apenas uma diferença significativa, uma melhor funcionalidade no quesito sociabilidade de acordo com a escala Vineland II. O agrupamento B4, único representante do agrupamento original A4, apresentou apenas uma diferença significativa (CBCL - Problemas externalizantes), com escores piores que o restante da amostra na comparação. Os demais agrupamentos apresentaram múltiplas diferenças, mas sem um padrão definido.

5 Discussão

Uma vez que os distúrbios do neurodesenvolvimento são poligênicos e multifatoriais, o avanço das tecnologias de sequenciamento contribuiu para um melhor entendimento de sua arquitetura genética com a descoberta de inúmeras novas variantes^{3,31,140}. Especificamente para TEA, tornou-se possível estabelecer o importante papel das variações comuns e raras, herdadas e *de novo*^{61,65,67–70,73,74,79,80,141}, com vários estudos mostrando um enriquecimento de variantes estruturais e de base única em casos de TEA comparados a indivíduos não afetados^{63,107,142,143}. Considerando que a contribuição de variantes de base única para o risco de TEA depende do impacto que essas variantes geram devido à troca de aminoácidos na proteína codificada^{132,137}, variantes do tipo LGD causariam mais danos do que as variantes de falta de sentido. Nesse sentido, foi proposta uma associação entre variantes de alto impacto e a gravidade da apresentação clínica^{74,132,144,145}.

No entanto, ao procurar por essas variantes, ficou claro que além das variantes de alto impacto ocorrerem em baixas quantidades⁵⁴, a probabilidade de dois pacientes apresentarem exatamente a mesma variante é muito pequena⁸⁶. Essas descobertas levaram a modelos que consideraram o impacto no gene^{146,147} e modelos posteriores que adicionaram vias genéticas⁸⁶. Levando em conta que um indivíduo apresenta vários tipos de variantes que afetam as vias gênicas, nós hipotetizamos que a composição dos tipos variantes poderia estar relacionada à apresentação fenotípica e que eles seriam capazes de estratificar grupos mais homogêneos de pacientes. Nosso estudo utilizou apenas variantes hereditárias *de novo* e raras, dadas as evidências que corroboram o importante papel destas na apresentação fenotípica^{107,137} e também apenas avaliou variantes em genes de vias ligadas ao TEA⁸⁶.

A abordagem composicional foi capaz de estratificar os pacientes de acordo com suas composições de forma bastante estável, mostrando que pode ser uma ferramenta promissora na busca de agrupamentos de probandos. Como esperado, os agrupamentos com composições que incluíam variantes *de novo* representavam apenas uma pequena porção de nossa amostra. Estudos prévios que mostraram a associação entre tipos de variantes e fenótipos foram feitos com medições de QI^{132,145}, então essa foi a nossa primeira abordagem. As análises incluindo a priori bimodal mostram que, embora não tenham sido observadas diferenças estatisticamente significativas comparando os seis agrupamentos, a ordem dos mesmos é quase como prevista com base na presença de variações que impactam na codificação da proteína¹³⁷. Considerando que o agrupamento A2 tem um QI mediano menor que o agrupamento A1, parece que uma composição com LGD *de novo* e LGD herdada poderia ter um impacto maior do que LGD *de novo* e *missense de novo*, ao contrário do que foi descrito na literatura^{132,137}. No entanto, a associação entre a composição e o fenótipo de QI não é perfeita, pois não esperávamos que o agrupamento A6 tivesse um QI mediano inferior ao agrupamento A5 e igual ao do agrupamento A4.

O uso da priori bimodal gerou agrupamentos com base no fato de terem ou não cada tipo de variação. A análise composicional com o uso da priori uniforme, estratificou 29 agrupamentos, mas que dessa vez refletiram as diferenças nos valores das composições. Como esperado, os 29 agrupamentos eram divisões quase perfeitas dos seis agrupamentos originais, sendo a única diferença dois probandos que não eram estáveis em nenhum dos grupos e foram removidos pelo próprio pacote pvclust. Os 29 agrupamentos refletiram a mediana do QI dos grupos originais, mas dentro de cada grupo original é evidente que não há relação direta entre a composição e a gravidade da apresentação do QI.

A comparação das distribuições cumulativas mostrou um número de diferenças significativas para quase todos os fenótipos testados. Essas diferenças mostram que os agrupamentos estratificados possuem fenótipos diferentes de acordo com suas composições, mas sem uma correlação direta^{132,137}. Esperávamos que o agrupamento A1 tivesse diferenças significativas quando comparado com o resto da amostra devido à presença de variantes LGD *de novo*, mas em conjunto com o agrupamento A4 não houve diferenças. Com a priori uniforme, 88% das subescalas apresentaram diferenças significativas para ao menos um agrupamento, mas não observamos um padrão definido. Agrupamentos pertencentes ao mesmo grupo apresentaram resultados mais graves e menos graves para as mesmas subescalas, o que indica que a composição formada por estes tipos de variantes nos genes selecionados não é suficiente para explicar a variabilidade fenotípica.

Nossos resultados refletem em parte os encontrados por Nishiyama et al.¹⁴⁸. Nishiyama et al.¹⁴⁸ exploraram a correlação genética entre traços autísticos e QI utilizando uma amostra de 45 pares de gêmeos do sexo masculino. Através de um modelo bivariado, os autores sugeriram que os fatores genéticos do TEA são substancialmente similares aos que influenciam o QI.

Os resultados aqui apresentados foram obtidos utilizando os genes das vias identificadas por Luo et al.⁸⁶, que usaram o mesmo conjunto de amostras que nós, mas relataram um número maior de variantes *de novo*. Isso indica que os critérios aplicados no presente trabalho para chamadas *de novo* foram mais rigorosos. A identificação de novas variantes *de novo* alteraria os agrupamentos e poderia ter produzido uma perspectiva diferente em relação à associação entre composição e fenótipos. Alterações na seleção inicial dos genes também mudaria parte dos nossos resultados.

Em relação ao nosso método de seleção na análise composicional, não estamos mais interessados nas quantidades absolutas apresentadas, mas nas proporções entre as partes¹⁰⁸. Dados composicionais consistem em vetores cujos componentes são a proporção ou porcentagem de uma unidade. Quando alguém diz que um problema é de composição, reconhece-se que os tamanhos de nossos espécimes são irrelevantes. Os dados de composição têm dois princípios importantes: invariância de escala e coerência subcomposicional. A coerência subcomposicional significa que eles fornecem informações apenas sobre as magnitudes relativas dos componentes de composição e interpretações envolvendo valores absolutos não podem ser justificadas¹⁰⁸. Essas propriedades beneficiam a presente análise dos dados de SCE, uma vez que, como estamos falando de uma comparação das proporções entre as partes, a profundidade da biblioteca de cada probando não tem a mesma interferência. Além disso, a composição inicial permanece válido mesmo com a adição *de novos* componentes, o que permitiria adicionar novas classes mutacionais sem invalidar nossa análise.

Duas possibilidades para melhorar o método proposto passam justamente por este princípio. A adição de classes específicas que identificam variações em genes intolerantes à variação¹⁴⁵ e variantes classificadas como patogênicas usando InterVar^{117,118}. Genes tolerantes à variação, mesmo aqueles em vias de interesse, podem não contribuir com informações relevantes para a classificação dos probandos ou podem ter sua importância superestimada¹⁴⁵. As variações não patogênicas, mesmo as mutações de LGD, podem não ter um impacto significativo no processo de codificação da proteína (por exemplo, variação no final de um exon ou fora de domínio de ação, troca de um aminoácido por outro com características físico-químicas semelhantes), sendo essencialmente inofensivas^{117,118}. No entanto, essas classes provavelmente estariam representadas em pequenas quantidades e em um número reduzido de probandos, o que poderia

desestabilizar os agrupamentos. Outras possibilidades passam desde a integração com dados provenientes de outras tecnologias até novos achados da literatura, como o possível impacto de variantes sinônimas e não codificadoras em diversos transtornos^{84,149,150}.

6 Conclusão

Através de quatro tipos de variantes genéticas e um conjunto de genes relevantes obtidos na literatura, aplicamos uma metodologia de análise composicional em conjunto com análise de agrupamentos hierárquicos com duas abordagens distintas. Ao utilizar uma abordagem com priori bimodal, verificamos que a simples presença e ausência de tipos de variantes genéticas não é capaz de explicar a totalidade de fenótipos relacionados com o TEA, diferente do que parte da literatura apresenta. Com o uso de uma abordagem com uma priori uniforme, introduzimos uma nova maneira de estratificar grupos de probandos com TEA sem a necessidade de conhecimento prévio quanto a sexo e informações familiares do probando. As abordagens empregadas fornecem visões da composição genética dos probandos, com parte dos agrupamentos estratificados sendo fenotipicamente diferentes. No entanto, essas diferenças fenotípicas não puderam ser diretamente relacionadas com a composição neste modelo.

7 Anexos

Anexo A

Tabela Anexa A - Caracterização da amostra proveniente do SSC. Os valores foram calculados para os 2313 probandos iniciais.

Subescala	Média (DP)	Mediana	Q1	Q3	Máximo	Mínimo
QI - Total	81,63 (27,8)	86	64	101	167	7
QI - Não Verbal	84,91 (26)	89	69	103	161	9
QI - Verbal	78,53 (31,07)	84	59	102	167	5
ABC - Subescala I - Irritabilidade	11,42 (8,77)	10	4	17	45	0
ABC - Subescala II - Letargia	9,81 (7,17)	8	4	14	40	0
ABC - Subescala III - Estereotipia	4,85 (4,27)	4	1	7	21	0
ABC - Subescala IV - Hiperatividade	16,64 (10,5)	15	8	24	48	0
ABC - Subescala V - Fala Inapropriada	3,58 (2,92)	3	1	5	12	0
ADI-R - A Total	20,4 (5,71)	21	16	25	30	8
ADI-R - B Não Verbal Total	16,51 (4,28)	17	13	20	26	6
ADI-R - B Verbal Total	9,26 (3,45)	9	7	12	14	0
ADI-R - C Total	6,55 (2,5)	6	5	8	12	0
CBCL - Comportamentos Externalizantes	56,59 (10,65)	57	48	64	97	32
CBCL - Comportamentos Internalizantes	60,41 (9,59)	61	54	67	90	33
RBS-R - Subescala I - Comportamento Estereotipado	4,45 (3,35)	4	2	7	18	0
RBS-R - Subescala II - Comportamento Auto Agressivo	2,10 (2,85)	1	0	3	21	0

continua

RBS-R - Subescala III - Comportamento Compulsivo	4,11 (3,94)	3	1	6	24	0
RBS-R - Subescala IV - Comportamento Ritualístico	5,04 (3,97)	4	2	7	18	0
RBS-R - Subescala V - Comportamento de Insistência na Mesmice	7,74 (5,96)	6	3	11	33	0
RBS-R - Subescala VI - Comportamento Restrito	3,68 (2,83)	3	1	5	12	0
RBS-R - Escore Geral	27,13 (17,39)	24	15	36	105	0
Vineland II - Autonomia Diária	76,58 (13,94)	76	68	85	127	25
Vineland II - Comunicação	77,22 (14,45)	77	69	86	132	30
Vineland II - Sociabilidade	71,02 (12,53)	71	63	80	117	34
Vineland II - Soma dos Escores Gerais	252,76 (60,38)	241	210	293	445	91

Anexo B

Tabela Anexa B - Detalhes das subescalas utilizadas neste trabalho. As descrições, com exceção das da escala CBCL, foram retiradas do dicionário de dados fornecido pela SSC. As descrições da CBCL foram retiradas do trabalho de Achenbach e Ruffle (2000).

Escala	Subescala	Descrição	Interpretação
QI	Total	Produzido atrás da medição verbal e não verbal vindas de um mesmo teste	Maior é melhor
QI	Não Verbal	Medição não verbal de um teste selecionado com base na idade e nível de habilidade da criança.	Maior é melhor
QI	Verbal	Medição verbal de um teste selecionado com base na idade e nível de habilidade da criança.	Maior é melhor
ABC	Subescala I - Irritabilidade	Soma dos escores dos itens de irritabilidade	Menor é melhor
ABC	Subescala II - Letargia	Soma dos escores dos itens de letargia	Menor é melhor
ABC	Subescala III - Estereotipia	Soma dos escores dos itens de estereotipia	Menor é melhor
ABC	Subescala IV - Hiperatividade	Soma dos escores dos itens de hiperatividade	Menor é melhor
ABC	Subescala V - Fala Inapropriada	Soma dos escores dos itens de fala inapropriada	Menor é melhor
ADI-R	A Total	Baseado na soma das maiores pontuações ou mais anormais de itens selecionados para o domínio da Interação Social Recíproca.	Menor é melhor
ADI-R	B Não Verbal Total	Baseado na soma das maiores pontuações ou mais anormais de itens selecionados do domínio de Comunicação. Todos os pacientes recebem essa pontuação.	Menor é melhor
ADI-R	B Verbal Total	Baseado na soma das maiores pontuações ou mais anormais de itens selecionados do domínio de Comunicação. Apenas os pacientes verbais recebem essa pontuação.	Menor é melhor
ADI-R	C Total	Os totais são calculados somando as maiores pontuações dos itens selecionados do domínio de padrões restritos, repetitivos e estereotipados de comportamento.	Menor é melhor
CBCL	Comportamentos Externalizantes	Pontuação relativa a comportamentos expressos diretamente no ambiente (comportamento agressivo, violação das regras).	Menor é melhor

continua

CBCL	Comportamentos Internalizantes	Pontuação relativa a comportamentos interiorizados, tais como retraimento, depressão, ansiedade, fobias e queixas somáticas.	Menor é melhor
RBS-R	Subescala I - Comportamento Estereotipado	Pontuação total do comportamento estereotipado que considera o número e a gravidade dos comportamentos.	Menor é melhor
RBS-R	Subescala II - Comportamento Auto Agressivo	Pontuação total do comportamento auto agressivo que leva em conta tanto o número quanto a gravidade dos comportamentos.	Menor é melhor
RBS-R	Subescala III - Comportamento Compulsivo	Pontuação total do comportamento compulsivo que leva em conta tanto o número quanto a gravidade dos comportamentos.	Menor é melhor
RBS-R	Subescala IV - Comportamento Ritualístico	Pontuação total do comportamento ritualístico que leva em conta tanto o número quanto a gravidade dos comportamentos.	Menor é melhor
RBS-R	Subescala V - Comportamento de Insistência na Mesmice	Pontuação total do comportamento de insistência na mesmice que leva em conta tanto o número quanto a gravidade dos comportamentos.	Menor é melhor
RBS-R	Subescala VI - Comportamento Restrito	Pontuação total do comportamento restritivo que leva em conta tanto o número quanto a gravidade dos comportamentos.	Menor é melhor
RBS-R	Escore Geral	Pontuação total que leva em conta tanto o número quanto a gravidade dos comportamentos de todos os domínios.	Menor é melhor
Vineland II	Autonomia Diária	Pontuação padrão que diz respeito ao nível geral de autonomia diária da criança. Pode ser comparada entre crianças de diferentes idades.	Maior é melhor
Vineland II	Comunicação	Pontuação padrão que diz respeito ao nível geral de comunicação da criança. Pode ser comparada entre crianças de diferentes idades.	Maior é melhor
Vineland II	Sociabilidade	Pontuação padrão que diz respeito ao nível geral de sociabilidade da criança. Pode ser comparada entre crianças de diferentes idades.	Maior é melhor
Vineland II	Soma dos Escores Gerais	Para crianças menores de 7 anos, esta é a soma dos escores padrão de domínio em Comunicação, Autonomia Diária, Sociabilidade e Habilidades Motoras. Para crianças de 7 anos ou mais, essa é a soma das pontuações padrão de domínio em Comunicação, Autonomia Diária e Sociabilidade.	Maior é melhor

Anexo C

Gráfico Anexo C - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as subescalas da ABC. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo D

Gráfico Anexo D - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as subescalas da ADI-R. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo E

Gráfico Anexo E - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as subescalas da CBCL. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo F

Gráfico Anexo F - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para a medida de QI verbal fornecida pelo SSC. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo G

Gráfico Anexo G - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as subescalas da RBS-R. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo H

Gráfico Anexo H - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as subescalas da ABC. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo I

Gráfico Anexo I - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as subescalas da ABC. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo J

Gráfico Anexo J - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as subescalas da ADI-R. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo K

Gráfico Anexo K - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para os comportamentos externalizantes e internalizantes da CBCL. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo L

Gráfico Anexo L - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as medidas de QI fornecidas pela SSC. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo M

Gráfico Anexo M - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as medidas de QI fornecidas pela SSC. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo N

Gráfico Anexo N - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as subescalas da RBS-R. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo O

Gráfico Anexo O - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as subescalas da RBS-R. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

Anexo P

Gráfico Anexo P - Distribuição cumulativa dos agrupamentos estratificados na abordagem com uso da priori bimodal com diferença significativa no Teste de Kolmogorov-Smirnov de dois lados de uma amostra para as subescalas da Vineland II. Os agrupamentos estão indicados no título e na legenda. Os círculos representam o agrupamento sendo testado, enquanto os triângulos representam o restante da amostra.

8 Bibliografia

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th Editio. Arlington, VA: American Psychiatric Association; 2013.
2. Autism and Developmental Disabilities Monitoring Network. Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 2010. Morb Mortal Wkly report [Internet]. 2014;63 Suppl 2(2):1–21. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24670961>
3. Devlin B, Scherer SW. Genetic architecture in autism spectrum disorder. Curr Opin Genet Dev [Internet]. 2012;22(3):229–37. Available from: <http://dx.doi.org/10.1016/j.gde.2012.03.002>
4. Duchan E, Patel DR. Epidemiology of autism spectrum disorders. Pediatr Clin North Am [Internet]. 2012;59(1):27–43. Available from: <http://dx.doi.org/10.1016/j.pcl.2011.10.003>
5. Achenbach TM, Dumenci L. Advances in empirically based assessment: Revised cross-informant syndromes and new DSM-oriented scales for the CBCL, YSR, and TRF: Comment on Lengua, Sadowksi, Friedrich, and Fisher (2001). J Consult Clin Psychol [Internet]. 2001;69(4):699–702. Available from: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-006X.69.4.699>
6. Simonoff E, Pickles A, Charman T, Chandler S, Loucas T, Baird G. Psychiatric Disorders in Children With Autism Spectrum Disorders: Prevalence, Comorbidity, and Associated Factors in a Population-Derived Sample. J Am Acad Child Adolesc Psychiatry [Internet]. 2008;47(8):921–9. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0890856708600593>
7. Maski KP, Jeste SS, Spence SJ. Common neurological co-morbidities in autism spectrum disorders. Curr Opin Pediatr [Internet]. 2011 Dec;23(6):609–15. Available from: <http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00008480-201112000-00008>
8. Tureck K, Matson JL, May A, Davis TE, Whiting SE. Investigation of the Rates of Comorbid Symptoms in Children with ADHD Compared to Children with ASD. J Dev Phys Disabil. 2013;25(4):405–17.
9. Matson JL, Goldin RL. Comorbidity and autism: Trends, topics and future directions. Res Autism Spectr Disord [Internet]. 2013;7(10):1228–33. Available from: <http://dx.doi.org/10.1016/j.rasd.2013.07.003>
10. Matson JL, Cervantes PE. Commonly studied comorbid psychopathologies among persons with autism spectrum disorder. Res Dev Disabil [Internet]. 2014;35(5):952–62. Available from: <http://dx.doi.org/10.1016/j.ridd.2014.02.012>
11. Folstein S, Rutter M. INFANTILE AUTISM: A GENETIC STUDY OF 21 TWIN PAIRS. J Child Psychol Psychiatry [Internet]. 1977 Sep;18(4):297–321. Available from: <http://doi.wiley.com/10.1111/j.1469-7610.1977.tb00443.x>
12. Steffenburg S, Gillberg C, Hellgren L, Andersson L, Gillberg IC, Jakobsson G, et al. A Twin Study of Autism in Denmark, Finland, Iceland, Norway and Sweden. J Child Psychol Psychiatry [Internet]. 1989 May;30(3):405–16. Available from: <http://doi.wiley.com/10.1111/j.1469-7610.1989.tb00254.x>
13. Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, Yuzda E, et al. Autism as a strongly genetic disorder: Evidence from a British twin study. Psychol Med. 1995;25(01):63.
14. Bartak L, Rutter M, Cox A. A Comparative Study of Infantile Autism and Specific Developmental Receptive Language Disorder: I. The Children. Br J Psychiatry [Internet]. 1975;126(2):127–45. Available from: <http://bjp.rcpsych.org/cgi/doi/10.1192/bjp.126.2.127%5Cnhttp://booksc.org/book/49841415>
15. PIVEN J, CHASE GA, LANDA R, WZOREK M, GAYLE J, CLOUD D, et al. Psychiatric Disorders in the Parents of Autistic Individuals. J Am Acad Child Adolesc Psychiatry [Internet]. 1991;30(3):471–8. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0890856709645688>
16. Abrahams BS, Geschwind DH. Advances in autism genetics: on the threshold of a new neurobiology. Nat Rev Genet [Internet]. 2008;9(6):493–493. Available from: <http://www.nature.com/doi/10.1038/nrg2380>
17. Piggot J, Shirinyan D, Shemmashian S, Vazirian S, Alarcón M. Neural systems approaches to the neurogenetics of autism spectrum disorders. Neuroscience [Internet]. 2009;164(1):247–56. Available from: <http://dx.doi.org/10.1016/j.neuroscience.2009.05.054>
18. Freitag CM, Staal W, Klauck SM, Duketis E, Waltes R. Genetics of autistic disorders: review and clinical implications. Eur Child Adolesc Psychiatry [Internet]. 2010;19(3):169–78. Available

- from: <http://www.ncbi.nlm.nih.gov/pubmed/19941018>
19. Benvenuto A, Moavero R, Alessandrelli R, Manzi B, Curatolo P. Syndromic autism: Causes and pathogenetic pathways. *World J Pediatr.* 2009;5(3):169–76.
 20. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, et al. Association between Microdeletion and Microduplication at 16p11.2 and Autism. *N Engl J Med* [Internet]. 2008;358(7):667–75. Available from: <http://www.nejm.org/doi/abs/10.1056/NEJMoa075974>
 21. Bochukova EG, Huang N, Keogh J, Henning E, Purmann C, Blaszczyk K, et al. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* [Internet]. 2010;463(7281):666–70. Available from: <http://www.nature.com/doi/abs/10.1038/nature08689>
 22. Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* [Internet]. 2010;463(7281):671–5. Available from: <http://www.nature.com/doi/abs/10.1038/nature08727>
 23. Shinawi M, Liu P, Kang SHL, Shen J, Belmont JW, Scott DA, et al. Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *J Med Genet.* 2010;47(5):332–41.
 24. Betancur C. Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. *Brain Res* [Internet]. 2011;1380:42–77. Available from: <http://dx.doi.org/10.1016/j.brainres.2010.11.078>
 25. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* [Internet]. 2010;11(6):415–25. Available from: <http://www.nature.com/doi/abs/10.1038/nrg2779>
 26. Kolevzon A, Gross R, Reichenberg A. Prenatal and Perinatal Risk Factors for Autism. *Arch Pediatr Adolesc Med.* 2007;161:326–33.
 27. Larsson HJ, Eaton WW, Madsen KM, Vestergaard M, Olesen AV, Agerbo E, et al. Risk factors for autism: Perinatal factors, parental psychiatric history, and socioeconomic status. *Am J Epidemiol.* 2005;161(10):916–25.
 28. Limperopoulos C, Bassan H, Sullivan NR, Soul JS, Robertson RL, Moore M, et al. Positive Screening for Autism in Ex-preterm Infants: Prevalence and Risk Factors. *Pediatrics* [Internet]. 2008;121(4):758–65. Available from: <http://pediatrics.aappublications.org/cgi/doi/10.1542/peds.2007-2158>
 29. Rutter M. Aetiology of autism: Findings and questions. *J Intellect Disabil Res.* 2005;49(4):231–8.
 30. Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, et al. Most genetic risk for autism resides with common variation. *Nat Genet* [Internet]. 2014;46(8):881–5. Available from: <http://www.nature.com/doi/abs/10.1038/ng.3039>
 31. Kiezun A, Garimella K, Do R, Stitzel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. *Nat Genet* [Internet]. 2012;44(6):623–30. Available from: <http://www.nature.com/doi/abs/10.1038/ng.2303>
 32. Glessner JT, Connolly JJ, Hakonarson H. Genome-Wide Association Studies of Autism. *Curr Behav Neurosci Reports* [Internet]. 2014;1(4):234–41. Available from: <http://dx.doi.org/10.1007/s40473-014-0023-0>
 33. Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS, et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* [Internet]. 2009;459(7246):528–33. Available from: <http://www.nature.com/doi/abs/10.1038/nature07999>
 34. Weiss LA, Arking DE, Daly MJ, Chakravarti A, Arking DE, Brune CW, et al. A genome-wide linkage and association scan reveals novel loci for autism. *Nature* [Internet]. 2009;461(7265):802–8. Available from: <http://www.nature.com/doi/abs/10.1038/nature08490>
 35. Ma D, Salyakina D, Jaworski JM, Konidari I, Whitehead PL, Andersen AN, et al. A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann Hum Genet.* 2009;73(3):263–73.
 36. Anney R, Klei L, Pinto D, Regan R, Conroy J, Magalhaes TR, et al. A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet.* 2010;19(20):4072–82.
 37. St. Pourcain B, Wang K, Glessner JT, Golding J, Steer C, Ring SM, et al. Association between a high-risk autism locus on 5p14 and social communication spectrum phenotypes in the general population. *Am J Psychiatry.* 2010;167(11):1364–72.
 38. Salyakina D, Ma DQ, Jaworski JM, Konidari I, Whitehead PL, Henson R, et al. Variants in several genomic regions associated with asperger disorder. *Autism Res.* 2010;3(6):303–10.
 39. Cho SC, Yoo HJ, Park M, Cho IH, Kim BN, Kim JW, et al. Genome-wide association scan of Korean autism spectrum disorders with language delay: A preliminary study. *Psychiatry Investig.* 2011;8(1):61–6.

40. Anney R, Klei L, Pinto D, Almeida J, Bacchelli E, Baird G, et al. Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum Mol Genet*. 2012;21(21):4781–92.
41. Poot M, Beyer V, Schwaab I, Damatova N, Van'T Slot R, Prothero J, et al. Disruption of CNTNAP2 and additional structural genome changes in a boy with speech delay and autism spectrum disorder. *Neurogenetics*. 2010;11(1):81–9.
42. Connolly JJ, Glessner JT, Hakonarson H. A Genome-Wide Association Study of Autism Incorporating Autism Diagnostic Interview-Revised, Autism Diagnostic Observation Schedule, and Social Responsiveness Scale. *Child Dev*. 2013;84(1):17–33.
43. Group C, Consortium PG. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* [Internet]. 2013;381(9875):1371–9. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0140673612621291>
44. Xia K, Guo H, Hu Z, Xun G, Zuo L, Peng Y, et al. Common genetic variants on 1p13.2 associate with risk of autism. *Mol Psychiatry* [Internet]. 2014;19(11):1212–9. Available from: http://dx.doi.org/10.1038/mp.2013.146%5Cn10.1038/mp.2013.146%5Cnhttps://www.researchgate.net/publication/258281271_Common_genetic_variants_on_1p132_associate_with_risk_of_autism
45. St Pourcain B, Whitehouse AJO, Ang WQ, Warrington NM, Glessner JT, Wang K, et al. Common variation contributes to the genetic architecture of social communication traits. *Mol Autism* [Internet]. 2013;4(1):34. Available from: <http://molecularautism.biomedcentral.com/articles/10.1186/2040-2392-4-34>
46. Jones RM, Cadby G, Melton PE, Abraham LJ, Whitehouse AJ, Moses EK. Genome-Wide Association Study of Autistic-Like Traits in a General Population Study of Young Adults. *Front Hum Neurosci* [Internet]. 2013;7(October):1–10. Available from: <http://journal.frontiersin.org/article/10.3389/fnhum.2013.00658/abstract>
47. Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol Autism* [Internet]. 2017 Dec 22;8(1):21. Available from: <http://molecularautism.biomedcentral.com/articles/10.1186/s13229-017-0137-9>
48. Sullivan PF. The Psychiatric GWAS Consortium: Big science comes to psychiatry. *Neuron* [Internet]. 2010;68(2):182–6. Available from: <http://dx.doi.org/10.1016/j.neuron.2010.10.003>
49. Torricio B, Chiochetti AG, Bacchelli E, Trabetti E, Hervás A, Franke B, et al. Lack of replication of previous autism spectrum disorder GWAS hits in European populations. *Autism Res*. 2017;10(2):202–11.
50. Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet*. 2013;9(3).
51. Consortium TIS. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* [Internet]. 2009 Jul 1;460:748. Available from: <http://dx.doi.org/10.1038/nature08185>
52. Klei L, Sanders SJ, Murtha MT, Hus V, Lowe JK, Willsey AJ, et al. Common genetic variants, acting additively, are a major source of risk for autism. *Mol Autism*. 2012;3(1):1–13.
53. Clarke T-K, Lupton MK, Fernandez-Pujals AM, Starr J, Davies G, Cox S, et al. Common polygenic risk for autism spectrum disorder (ASD) is associated with cognitive ability in the general population. *Mol Psychiatry* [Internet]. 2016;21(3):419–25. Available from: <http://www.nature.com/doifinder/10.1038/mp.2015.12>
54. Wilfert AB, Sulovari A, Turner TN, Coe BP, Eichler EE. Recurrent de novo mutations in neurodevelopmental disorders: Properties and clinical implications. *Genome Med*. 2017;9(1):1–16.
55. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. *Science* (80-). 2007;316(5823):445–9.
56. Bucan M, Abrahams BS, Wang K, Glessner JT, Herman EI, Sonnenblick LI, et al. Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet*. 2009;5(6).
57. Wiśniowiecka-Kowalik B, Nesteruk M, Peters SU, Xia Z, Cooper ML, Savage S, et al. Intragenic rearrangements in NRXN1 in three families with autism spectrum disorder, developmental delay, and speech delay. *Am J Med Genet Part B Neuropsychiatr Genet* [Internet]. 2010;9999B(5):n/a-n/a. Available from: <http://doi.wiley.com/10.1002/ajmg.b.31064>
58. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* [Internet]. 2009 May

- 28;459(7246):569–73. Available from: <http://dx.doi.org/10.1038/nature07953>
59. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* [Internet]. 2010;466(7304):368–72. Available from: <http://www.nature.com/doifinder/10.1038/nature09146>
60. Willemsen MH, Fernandez BA, Bacino CA, Gerkes E, de Brouwer AP, Pfundt R, et al. Identification of ANKRD11 and ZNF778 as candidate genes for autism and variable cognitive impairment in the novel 16q24.3 microdeletion syndrome. *Eur J Hum Genet* [Internet]. 2010 Apr 18;18(4):429–35. Available from: <http://dx.doi.org/10.1038/ejhg.2009.192>
61. O’Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* [Internet]. 2011;43(6):585–9. Available from: <http://www.nature.com/doifinder/10.1038/ng.835>
62. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, et al. Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. *Neuron*. 2011;70(5):863–85.
63. Levy D, Ronemus M, Yamrom B, Lee Y ha, Leotta A, Kendall J, et al. Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders. *Neuron* [Internet]. 2011;70(5):886–97. Available from: <http://dx.doi.org/10.1016/j.neuron.2011.05.015>
64. Fischbach GD, Lord C. The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. *Neuron* [Internet]. 2010 Oct;68(2):192–5. Available from: <http://dx.doi.org/10.1016/j.neuron.2010.10.006>
65. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* [Internet]. 2012;485(7397):237–41. Available from: <http://www.nature.com/doifinder/10.1038/nature10945>
66. Chahrour MH, Yu TW, Lim ET, Ataman B, Coulter ME, Hill RS, et al. Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. *PLoS Genet*. 2012;8(4).
67. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, et al. De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* [Internet]. 2012;74(2):285–99. Available from: <http://dx.doi.org/10.1016/j.neuron.2012.04.009>
68. Neale BM, Kou Y, Liu L, Ma’ayan A, Samocha KE, Sabo A, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* [Internet]. 2012;485(7397):242–5. Available from: <http://www.nature.com/doifinder/10.1038/nature11011>
69. O’Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* [Internet]. 2012;485(7397):246–50. Available from: <http://www.nature.com/doifinder/10.1038/nature10989>
70. Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, Ataman B, et al. Using Whole-Exome Sequencing to Identify Inherited Causes of Autism. *Neuron* [Internet]. 2013;77(2):259–73. Available from: <http://dx.doi.org/10.1016/j.neuron.2012.11.002>
71. Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, MacArthur DG, et al. Rare Complete Knockouts in Humans: Population Distribution and Significant Role in Autism Spectrum Disorders. *Neuron* [Internet]. 2013;77(2):235–42. Available from: <http://dx.doi.org/10.1016/j.neuron.2012.12.029>
72. Poultney CS, Goldberg AP, Drapeau E, Kou Y, Harony-Nicolas H, Kajiwarra Y, et al. Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. *Am J Hum Genet*. 2013;93(4):607–19.
73. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* [Internet]. 2014;515(7526):209–15. Available from: <http://dx.doi.org/10.1038/nature13772>
74. Iossifov I, O’Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* [Internet]. 2014;515(7526):216–21. Available from: <http://www.nature.com/doifinder/10.1038/nature13908>
75. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* [Internet]. 2014;46(10):1063–71. Available from: <http://www.nature.com/doifinder/10.1038/ng.3092>
76. Correia C, Oliveira G, Vicente AM. Protein interaction networks reveal novel autism risk genes within GWAS statistical noise. *PLoS One*. 2014;9(11):1–11.
77. Chiocchetti AG, Haslinger D, Boesch M, Karl T, Wiemann S, Freitag CM, et al. Protein

- signatures of oxidative stress response in a patient specific cell line model for autism. *Mol Autism* [Internet]. 2014;5(1):10. Available from: <http://molecularautism.biomedcentral.com/articles/10.1186/2040-2392-5-10>
78. Toma C, Torrico B, Hervás A, Valdés-Mas R, Tristán-Noguero A, Padillo V, et al. Exome sequencing in multiplex autism families suggests a major role for heterozygous truncating mutations. *Mol Psychiatry*. 2014;19(7):784–90.
 79. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, et al. Excess of rare, inherited truncating mutations in autism. *Nat Genet* [Internet]. 2015;47(6):582–8. Available from: <http://www.nature.com/doi/10.1038/ng.3303>
 80. Egawa J, Watanabe Y, Wang C, Inoue E, Sugimoto A, Sugiyama T, et al. Novel rare missense variations and risk of autism spectrum disorder: Whole-exome sequencing in two families with affected siblings and a two-stage follow-up study in a Japanese population. *PLoS One*. 2015;10(3):1–9.
 81. AlAyadhi LY, Hashmi JA, Iqbal M, Albalawi AM, Samman MI, Elamin NE, et al. High-resolution SNP genotyping platform identified recurrent and novel CNVs in autism multiplex families. *Neuroscience* [Internet]. 2016;339:561–70. Available from: <http://dx.doi.org/10.1016/j.neuroscience.2016.10.030>
 82. Li J, Cai T, Jiang Y, Chen H, He X, Chen C, et al. Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol Psychiatry* [Internet]. 2016 Feb 7;21(2):290–7. Available from: <http://www.nature.com/articles/mp201540>
 83. Codina-Solà M, Rodríguez-Santiago B, Homs A, Santoyo J, Rigau M, Aznar-Laín G, et al. Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders. *Mol Autism* [Internet]. 2015;6(1):1–16. Available from: ???
 84. Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, et al. Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am J Hum Genet* [Internet]. 2016;98(1):58–74. Available from: <http://dx.doi.org/10.1016/j.ajhg.2015.11.023>
 85. De Souza Reis VN, Kitajima JP, Tahira AC, Feio-Dos-santos AC, Fock RAS, Lisboa BCG, et al. Integrative variation analysis reveals that a complex genotype may specify phenotype in siblings with syndromic autism spectrum disorder. *PLoS One*. 2017;12(1).
 86. Luo W, Zhang C, Jiang YH, Brouwer CR. Systematic reconstruction of autism biology from massive genetic mutation profiles. *Sci Adv*. 2018;4(4):1–14.
 87. Ben-David E, Shifman S. Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. *Mol Psychiatry* [Internet]. 2013 Oct 13;18(10):1054–6. Available from: <http://www.nature.com/articles/mp2012148>
 88. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, et al. Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell* [Internet]. 2013 Nov;155(5):1008–21. Available from: <http://dx.doi.org/10.1016/j.cell.2013.10.031>
 89. Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, Muhle RA, et al. Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. *Cell* [Internet]. 2013 Nov;155(5):997–1007. Available from: <http://dx.doi.org/10.1016/j.cell.2013.10.020>
 90. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* [Internet]. 2014;506(7487):179–84. Available from: <http://dx.doi.org/10.1038/nature12929>
 91. Hamdan FF, Daoud H, Rochefort D, Piton A, Gauthier J, Langlois M, et al. De Novo Mutations in FOXP1 in Cases with Intellectual Disability, Autism, and Language Impairment. *Am J Hum Genet* [Internet]. 2010 Nov;87(5):671–8. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S000292971000515X>
 92. McCarthy SE, Gillis J, Kramer M, Lihm J, Yoon S, Bernstein Y, et al. De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Mol Psychiatry*. 2014;19(6):652–8.
 93. Hormozdiari F, Penn O, Borenstein E, Eichler EE. The discovery of integrated gene networks for autism and related disorders. *Genome Res*. 2015;25:142–54.
 94. Yin J, Schaaf CP. Autism genetics – an overview. *Prenat Diagn*. 2017;37(1):14–30.
 95. Sztainberg Y, Zoghbi HY. Lessons learned from studying syndromic autism spectrum disorders. *Nat Neurosci*. 2016;19(11):1408–18.
 96. Glennon JM, Karmiloff-Smith A, Thomas MSC. Syndromic Autism: Progressing Beyond Current

- Levels of Description. *Rev J Autism Dev Disord*. 2017;4(4):321–7.
97. Jacquemont S, Coe BP, Hersch M, Duyzend MH, Krumm N, Bergmann S, et al. A higher mutational burden in females supports a “female protective model” in neurodevelopmental disorders. *Am J Hum Genet*. 2014;94(3):415–25.
98. Ferri SL, Abel T, Brodtkin ES. Sex Differences in Autism Spectrum Disorder: a Review. *Curr Psychiatry Rep* [Internet]. 2018;20(2):9. Available from: <http://link.springer.com/10.1007/s11920-018-0874-2>
99. Werling DM, Geschwind DH. Recurrence rates provide evidence for sex-differential, familial genetic liability for autism spectrum disorders in multiplex families and twins. *Mol Autism* [Internet]. 2015 Dec 13;6(1):27. Available from: <http://www.molecularautism.com/content/6/1/27>
100. Halladay AK, Bishop S, Constantino JN, Daniels AM, Koenig K, Palmer K, et al. Sex and gender differences in autism spectrum disorder: Summarizing evidence gaps and identifying emerging areas of priority. *Mol Autism* [Internet]. 2015;6(1):1–5. Available from: <http://dx.doi.org/10.1186/s13229-015-0019-y>
101. Waye MMY, Cheng HY. Genetics and Epigenetics of Autism (Review). *Psychiatry Clin Neurosci* [Internet]. 2017;1–17. Available from: <http://doi.wiley.com/10.1111/pcn.12606>
102. Werling DM, Parikshak NN, Geschwind DH. Gene expression in human brain implicates sexually dimorphic pathways in autism spectrum disorders. *Nat Commun* [Internet]. 2016;7:1–11. Available from: <http://dx.doi.org/10.1038/ncomms10717>
103. Virkud Y V., Todd RD, Abbacchi AM, Zhang Y, Constantino JN. Familial aggregation of quantitative autistic traits in multiplex versus simplex autism. *Am J Med Genet Part B Neuropsychiatr Genet*. 2009;150(3):328–34.
104. Gerdts JA, Bernier R, Dawson G, Estes A. The broader autism phenotype in simplex and multiplex families. *J Autism Dev Disord*. 2013;43(7):1597–605.
105. Ben-David E, Shifman S. Networks of neuronal genes affected by common and rare variants in autism spectrum disorders. *PLoS Genet*. 2012;8(3).
106. Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat Genet* [Internet]. 2017;49(7):978–85. Available from: <http://www.nature.com/doi/10.1038/ng.3863>
107. Shohat S, Ben-David E, Shifman S. Varying Intolerance of Gene Pathways to Mutational Classes Explain Genetic Convergence across Neuropsychiatric Disorders. *Cell Rep* [Internet]. 2017;18(9):2217–27. Available from: <http://dx.doi.org/10.1016/j.celrep.2017.02.007>
108. Aitchison J. A Concise Guide to Compositional Data Analysis. In: 2nd Compositional Data Analysis Workshop. Girona; 2003.
109. Tsilimigras MCB, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol* [Internet]. 2016;26(5):330–5. Available from: <http://dx.doi.org/10.1016/j.annepidem.2016.03.002>
110. Fernandes, D. A, Reid, J., Macklaim, M. J, et al. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* [Internet]. 2014;2(15):1–13. Available from: ***
111. Carson V, Tremblay MS, Chaput J-P, Chastin SFM. Associations between sleep duration, sedentary time, physical activity, and health indicators among Canadian children and youth using compositional analyses. *Appl Physiol Nutr Metab* [Internet]. 2016;41(6 (Suppl. 3)):S294–302. Available from: <http://www.nrcresearchpress.com/doi/10.1139/apnm-2016-0026>
112. Dumuid D, Olds T, Lewis LK, Martin-Fernández JA, Barreira T, Broyles S, et al. The adiposity of children is associated with their lifestyle behaviours: a cluster analysis of school-aged children from 12 nations. *Pediatr Obes*. 2018;13(2):111–9.
113. Fairclough SJ, Dumuid D, Taylor S, Curry W, McGrane B, Stratton G, et al. Fitness, fatness and the reallocation of time between children’s daily movement behaviours: An analysis of compositional data. *Int J Behav Nutr Phys Act*. 2017;14(1):1–12.
114. Foley L, Dumuid D, Atkin AJ, Olds T, Ogilvie D. Patterns of health behaviour associated with active travel: a compositional data analysis. *Int J Behav Nutr Phys Act* [Internet]. 2018 Dec 21;15(1):26. Available from: <http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L621323157%0Ahttp://dx.doi.org/10.1186/s12966-018-0662-8>
115. Achenbach TM, Ruffle TM. The Child Behavior Checklist and Related Forms for Assessing Behavioral/Emotional Problems and Competencies. *Pediatr Rev* [Internet]. 2000;21(8):265–71.

- Available from: <http://pedsinreview.aappublications.org/cgi/doi/10.1542/pir.21-8-265>
116. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–303.
 117. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet* [Internet]. 2017;100(2):267–80. Available from: <http://dx.doi.org/10.1016/j.ajhg.2017.01.004>
 118. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* [Internet]. 2015;17(5):405–23. Available from: <http://www.nature.com/doi/10.1038/gim.2015.30>
 119. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014;42(D1):980–5.
 120. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003;31(13):3812–4.
 121. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods* [Internet]. 2010;7(4):248–9. Available from: <http://dx.doi.org/10.1038/nmeth0410-248>
 122. Team RC. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: <https://www.r-project.org/>
 123. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics.* 2005;21:3439–40.
 124. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc.* 2009;4:1184–91.
 125. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* [Internet]. 2019;531210. Available from: <https://www.biorxiv.org/content/10.1101/531210v1>
 126. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. *Nature* [Internet]. 2015 Oct 30;526(7571):68–74. Available from: <http://www.nature.com/articles/nature15393>
 127. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* [Internet]. 2013 Jan 28;493(7431):216–20. Available from: <http://www.nature.com/articles/nature11690>
 128. Suzuki R, Shimodaira H. Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* [Internet]. 2006 Jun 15;22(12):1540–2. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl117>
 129. Mayes SD, Calhoun SL, Murray MJ, Zahid J. Variables Associated with Anxiety and Depression in Children with Autism. *J Dev Phys Disabil* [Internet]. 2011 Aug 23;23(4):325–37. Available from: <http://link.springer.com/10.1007/s10882-011-9231-7>
 130. Amr M, Raddad D, El-Mehesh F, Bakr A, Sallam K, Amin T. Comorbid psychiatric disorders in Arab children with Autism spectrum disorders. *Res Autism Spectr Disord* [Internet]. 2012;6(1):240–8. Available from: <http://dx.doi.org/10.1016/j.rasd.2011.05.005>
 131. Tureck K, Matson JL, Cervantes P, Konst MJ. An examination of the relationship between autism spectrum disorder, intellectual functioning, and comorbid symptoms in children. *Res Dev Disabil* [Internet]. 2014;35(7):1766–72. Available from: <http://dx.doi.org/10.1016/j.ridd.2014.02.013>
 132. Robinson EB, Samocha KE, Kosmicki JA, McGrath L, Neale BM, Perlis RH, et al. Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proc Natl Acad Sci* [Internet]. 2014;111(42):15161–5. Available from: <http://www.pnas.org/lookup/doi/10.1073/pnas.1409204111>
 133. Aman MG, Singh NN, Stewart AW, Field CJ. The aberrant behavior checklist: A behavior rating scale for the assessment of treatment effects. Vol. 89, *American Journal of Mental Deficiency*. US: American Assn on Mental Retardation; 1985. p. 485–91.
 134. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* [Internet]. 1994 Oct;24(5):659–85. Available from:

- <http://www.ncbi.nlm.nih.gov/pubmed/7814313>
135. Bodfish JW, Symons FJ, Parker DE, Lewis MH. Varieties of repetitive behaviour in autism: comparison to mental retardation. *J Autism Dev Disord*. 2000;30(3):237–43.
 136. Sparrow SS, Balla DA, Cicchetti D V. Vineland II: Vineland adaptive behavior scales. American Guidance Service; 2005.
 137. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, et al. Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. *PLoS Genet*. 2013;9(8).
 138. Eklund A. beeswarm: The Bee Swarm Plot, an Alternative to Stripchart [Internet]. 2016. Available from: <https://cran.r-project.org/package=beeswarm>
 139. Dimitrova DS, Kaishev VK, Tan S. Computing the Kolmogorov-Smirnov Distribution when the Underlying cdf is Purely Discrete, Mixed or Continuous. 2017;
 140. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nat Rev Genet* [Internet]. 2018;19(2):110–24. Available from: <http://dx.doi.org/10.1038/nrg.2017.101>
 141. O’Roak BJ, Stessman HA, Boyle EA, Witherspoon KT, Martin B, Lee C, et al. Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nat Commun*. 2014;5:1–6.
 142. Gratten J, Wray NR, Keller MC, Visscher PM. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat Neurosci* [Internet]. 2014;17(6):782–90. Available from: <http://dx.doi.org/10.1038/nn.3708>
 143. Gonzalez-Mantilla AJ, Moreno-de-luca A, Ledbetter DH, Martin CL. A Cross-Disorder Method to Identify Novel Candidate Genes for Developmental Brain Disorders. *JAMA psychiatry*. 2017;73(3):275–83.
 144. Hamdan FF, Daoud H, Rochefort D, Piton A, Gauthier J, Langlois M, et al. De novo mutations in FOXP1 in cases with intellectual disability, autism, and language impairment. *Am J Hum Genet*. 2010;87(5):671–8.
 145. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. *Nat Genet* [Internet]. 2014;46(9):944–50. Available from: <http://www.nature.com/doi/10.1038/ng.3050>
 146. Zhao X, Leotta A, Kustanovich V, Lajonchere C, Geschwind DH, Law K, et al. A unified genetic theory for sporadic and inherited autism. *Proc Natl Acad Sci* [Internet]. 2007;104(31):12831–6. Available from: <http://www.pnas.org/cgi/doi/10.1073/pnas.0705803104>
 147. Ronemus M, Iossifov I, Levy D, Wigler M. The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet* [Internet]. 2014;15(2):133–41. Available from: <http://www.nature.com/doi/10.1038/nrg3585>
 148. Nishiyama T, Taniai H, Miyachi T, Ozaki K, Tomita M, Sumi S. Genetic correlation between autistic traits and IQ in a population-based sample of twins with autism spectrum disorders (ASDs). *J Hum Genet*. 2009;54(1):56–61.
 149. Takata A, Ionita-Laza I, Gogos JA, Xu B, Karayiorgou M. De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia. *Neuron* [Internet]. 2016;89(5):940–7. Available from: <http://dx.doi.org/10.1016/j.neuron.2016.02.024>
 150. McRae JF, Clayton S, Fitzgerald TW, Kaplanis J, Prigmore E, Rajan D, et al. Prevalence and architecture of de novo mutations in developmental disorders. *Nature*. 2017;542(7642):433–8.