UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO "
Campus de Botucatu

Aprendizado de Máquina e Biologia de Sistemas aplicada ao estudo da Síndrome de Microdeleção 22q11

CAMILA CRISTINA DE OLIVEIRA ALVES

BOTUCATU-SP
2019

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO "
Campus de Botucatu

Aprendizado de Máquina e Biologia de Sistemas aplicada ao estudo da Síndrome de Microdeleção 22q11

CAMILA CRISTINA DE OLIVEIRA ALVES

Orientadora: Profa. Dra. Lucilene Arilho Ribeiro Bicudo

Co-orientador: Prof. Dr. Guilherme Targino Valente

Dissertação apresentada ao Instituto de Biociências, Câmpus de Botucatu, UNESP, para obtenção do título de Mestre no Programa de Pós-Graduação em Ciências Biológicas (Genética).

BOTUCATU-SP

2019

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Alves, Camila Cristina de Oliveira.
Aprendizado de máquina e biologia de sistemas aplicada ao estudo da Síndrome de Microdeleção 22q11 / Camila Cristina de Oliveira Alves. - Botucatu, 2019

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Lucilene Arilho Ribeiro Bicudo
Coorientador: Guilherme Targino Valente
Capes: 20205007

1. Aprendizado de máquina. 2. Interação proteína-proteína. 3. Cromossomos - Distúrbios. 4. Cromossomos humanos par 22.

Palavras-chave: 22q11DS; Aprendizado de máquinas; Rede de interação proteína-proteína; Síndrome DiGeorge.

Aos meus pais Eliete e Neto, e a toda minha família e amigos pelo auxilio e constante incentivo.

AGRADECIMENTOS

A minha orientadora Profa. Dra. Lucilene Arilho Ribeiro Bicudo pela oportunidade, pelo aprdendizado profissional e pessoal e pela confiança depositada em mim.

Ao meu co-orientador Prof. Dr. Guilherme Targino Valente por me receber em seu laboratório e sua equipe. Obrigada pela colaboração, por todo ensinamento, incentivo, paciência e acolhimento. Sou muito grata por todo esforço e empenho durante esse tempo, tenho uma enorme admiração pela pessoa e profissional que você é.

Ao Me. Ivan Rodrigo Wolf por compartilhar seus conhecimentos e pelo auxilio desde a escrita do projeto, a colaboração em cada fase do trabalho até as sugestões e correções durante a elaboração da dissertação. Obrigada por me receber no laboratório, pela paciência e amizade.

Ao Dr. Bruno Faulin Gamba por todo o apoio e parceria. Muito obrigada pelos conselhos, orientações e toda ajuda na interpretação dos resultados e na elaboração dessa dissertação!

Aos meus amigos Lucas Farinazzo, Luiz Card, Lucas Lazari, Guilherme Luz, Eric Kawagoe e Giovanna Rock, que fizeram parte dessa jornada. Muito obrigada pelo acolhimento no laboratório, pelo apoio, orientações, risadas e por tantos momentos compartilhados. Guardarei boas lembranças.

A Camila Vaz Souza e Talita Aleixo que me acolheram em Botucatu da melhor forma e me deram todo o apoio durante o mestrado. Aprendi muito com vocês! Obrigada por todo o companheirismo.

A todos os outros amigos e colegas da Pós-graduação em Genética que contribuíram com esse trabalho tanto de forma direta quanto indireta.

Agradeço aos meus pais, Neto e Eliete por todo o apoio, conselhos, companheirismo, carinho, paciência e por estarem comigo em cada conquista. Vocês são meus exemplos de força e perseverança, amo vocêis!

A CAPES pela bolsa de estudos concedida!

RESUMO

ALVES, CCO. **Aprendizado de Máquina e Biologia de Sistemas aplicada ao estudo da Síndrome de Microdeleção 22q11.** 2019, 117 p. Dissertação de mestrado – Instituto de Biociências de Botucatu, Univeeresidade Estadual Paulista "Júlio de Mesquita Filho".

A Síndrome de Microdeleção 22q11 (SD22q11), causada por uma deleção de aproximadamente 3Mb na região 22q11, apresenta uma frequencia média de 1 em 4000 a 9800 nascidos vivos sendo considera a síndrome de microdeleção mais frequente e a segunda causa mais comum de atraso no desenvolvimento e de doença congênita grave, após a síndrome de Down. De acordo com o tamanho e a localização da deleção, diferentes genes podem ser afetados e o principal gene considerado como responsável pelos sinais clássicos da síndrome é o *TBX1*. A SD22q11 caracteriza-se por um espectro fenotípico bastante amplo, com efeitos pleiotrópicos que resultam no acometimento de praticamente todos os órgãos e/ou sistemas, altamente variáveis com mais de 180 sinais clínicos já descritos, tanto físicos como comportamentais. Nesse trabalho aplicamos ferramentas de bioinformática com o intuito de descobrir padrões clínicos e sistêmicos da deleção 22q11, classificando casos sindrômicos em típicos e atípicos e estudando o impacto da deleção em redes de interação proteína-proteína (PPI). Para avaliação dos sinais clínicos que pudessem diferenciar pacientes sindrômicos foi aplicado uma metodologia baseada em aprendizado de máquina para classificar os casos em típico e atípico de acordo com os sinais clínicos através do algoritmo J48 (um algoritmo de árvore de decisão). As árvores de decisão selecionadas foram altamente precisas. Sinais clínicos como fissura oral, insuficiência velofaríngea, atraso no desenvolvimento de fala e linguagem, incapacidade de aprendizagem específica, anormalidade comportamental e atraso de crescimento foram indicativos para classificação dos casos. Já a avaliação do impacto da deleção da região 22q11 foi realizada através de estudos envolvendo redes biológicas. Assim, os genes codificadores de proteínas envolvidos na deleção foram removidos da rede PPI humana para simular a deleção. Diferentes análises topológicas foram utilizadas para comparar a rede global (GN) com a rede paciente (PN). Além disso foi verificado as comunidades de ambas as redes e realizou-se uma análise de enriquecimento de ontologia. Os resultados mostraram que não há diferença significativa ao comparar GN e PN, porém observamos que há diferença entre as comunidades dessas redes. Além disso, foi possível analisar diferentes genes que estavam presentes em regiões enriquecidas com termos ontológicos semelhantes. Dessa forma, podemos concluir que estudos envolvendo Aprendizado de Máquina e Redes Biológicas podem apontar novas hipoteses no estudo da SD22q11 além de ter potencial para esclarecer diversos aspectos de diferentes patologias que não são prontamente acessíveis pela biologia molecular convencional ou abordagens genéticas.

Palavras-chaves: 22q11SD; Síndrome DiGeorge; Aprendizado de máquinas; Rede de interação proteína-proteína.

ABSTRACT

ALVES, CCO. **Machine Learning and Systems Biology applied to the study of the 22q11 Microdeletion Syndrome.** 2019, 117 p. Master's degree in Science – Instituto de Biociências de Botucatu, Universidade Estadual Paulista "Júlio de Mesquita Filho".

The 22q11 Microdeletion Syndrome (22q11DS), caused by a deletion of approximately 3Mb in the 22q11 region, has an average frequency of 1 in 4000 to 9800 live births and is considered the most frequent microdeletion syndrome and the second most common cause of developmental delay and severe congenital disease after Down syndrome. According to the size and location of the deletion, different genes may be affected and the main gene considered to be responsible for the classic signs of the syndrome is *TBX1*. 22q11DS is characterized by a very broad phenotypic spectrum with pleiotropic effects that result in the involvement of variable organs and/or systems with more than 180 clinical signs already described, both physical and behavioral. In this work, we applied bioinformatics tools to detect clinical and systemic patterns of 22q11 deletion, classifying typical and atypical syndromic cases, and studying the impact of deletion on protein-protein interaction (PPI) networks. To evaluate clinical signs that could differentiate syndromic patients, a machine-learning based methodology was used to classify the cases into typical and atypical according to the clinical signs through the algorithm J48 (a decision tree algorithm). The selected decision trees were highly accurate. Clinical signs such as oral fissure, velopharyngeal insufficiency, speech and language development delay, specific learning disability, behavioral abnormality and growth delay were indicative for case classification. The evaluation of the impact of the 22q11 region deletion was performed through studies involving biological networks. To achieve this goal, the protein coding genes involved in the deletion were removed from the human PPI network to mimic the deletion. Different topological analyzes were used to compare the global network (GN) with the patient network (PN). In addition, the communities of both networks were verified and an ontology enrichment analysis was performed. The results showed that there is no significant difference when comparing GN and PN, but we observed that there is difference between the communities of these networks. In addition, it was possible to analyze different genes that were present in regions enriched with similar ontological terms. Thus, we can conclude that studies involving Machine Learning and Biological Networks may point out new hypotheses in the study of 22q11DS and have the potential to clarify several aspects of different pathologies that are not readily accessible by conventional molecular biology or genetic approaches.

Keywords: 22q11DS; DiGeorge syndrome; Machine learning; Protein-protein interaction network.

SUMÁRIO

INTRODUÇÃO E REVISÃO DA LITERATURA

INTRODUÇÃO E REVISÃO DA LITERATURA

1. ESTRUTURA DO GENOMA

Distúrbios genômicos são doenças resultantes da perda ou ganho de material cromossômico. Os rearranjos cromossômicos em humanos são diversos, frequentes e geralmente resultam em anormalidades fenotípicas, defeitos de nascimento e letalidade embrionária (Shaffer and Lupski, 2000). As desordens genômicas mais comuns e delineadas são divididas em duas categorias principais: as que resultam da perda do número de cópias (deleções) e do ganho de número de cópias (duplicações) (Picchi, 1997). Além disso, alguns rearranjos cromossômicos ocorrem em segmentos menores, assim, deleções cromossômicas muito pequenas que não são detectadas pela microscopia, utilizando métodos citogenéticos tradicionais, são denominadas microdeleções (Shaffer and Lupski, 2000).

Ao longo do genoma, várias cópias de Regiões de repetição de pequeno número de cópias (LCRs - *Low Copy Repeats*) podem ser encontrados (Cardoso *et al.*, 2016). Os LCRs são sequencias homologas, com comprimento maior ou igual a 1Kb, que foram gerados através de eventos de duplicação (Cardoso *et al.*, 2016; Harel and Lupski, 2018). Os LCRs com sequencias de alta homologia podem promover uma recombinação homóloga não alélica (NAHR) (Shaffer and Lupski, 2000; Shaikh, Kurahashi and Emanuel, 2001; Burnside, 2015).

Dois tipos de NAHR podem ocorrer entre os LCRs: eventos Intercromossômicos entre LCRs parálogos ou eventos Intracromossômicos (Figura 1). Assim, LCRs levam a ocorrência de NAHR que resultam em variações no número de cópias (CNVs). O tamanho das CNVs é de mais de 50pb e pode resultar em 1,2% de diferença em relação ao genoma humano de referência (Zarrei *et al.*, 2015; Nowakowska, 2017). Além disso, um continuo espectro de fenótipos é resultante de recorrentes CNVs (Zarrei *et al.*, 2015; Harel and Lupski, 2018).

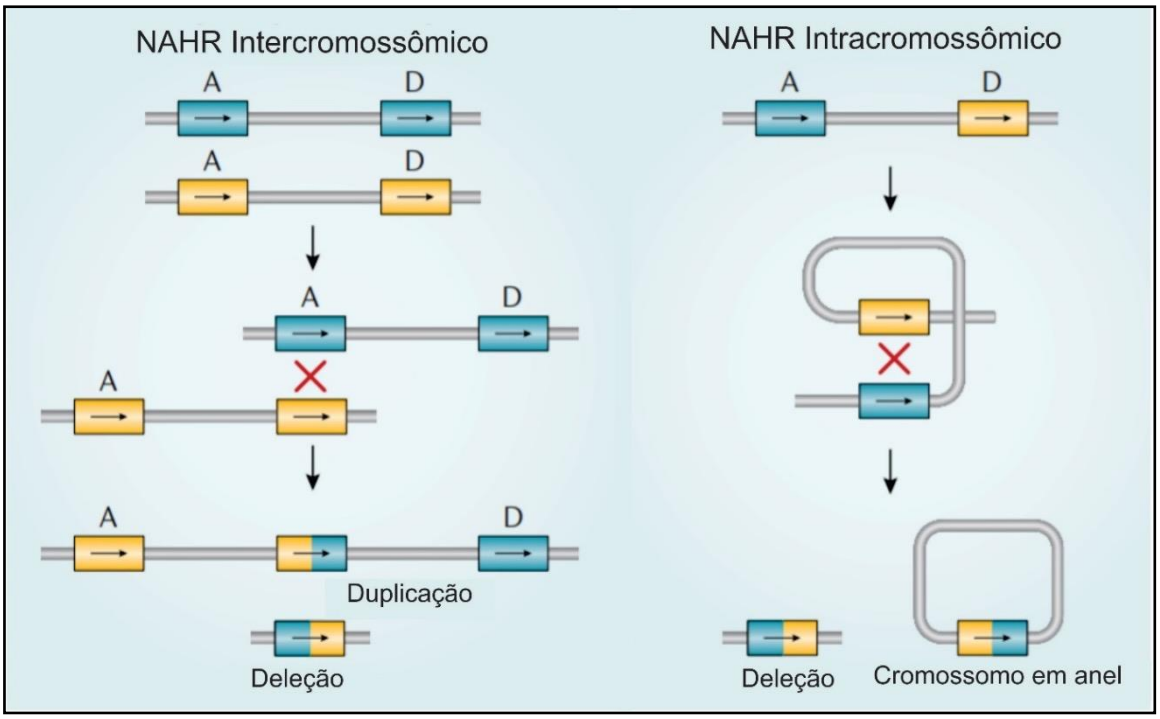


Figura 1. Diagrama dos dois tipos diferentes NAHR (Recombinação Homóloga Não Alélica) que podem ocorrer entre LCRs (Repetições de Pequeno Número de Cópias). No lado esquerdo, observa-se um rearranjo intercromossômico entre dois LCRs, indicados como A e D respectivamente. Esse processo resulta em uma duplicação ou deleção de genes nos gametas resultantes (O "X" mostra o cruzamento dos dois cromossomos). Já no lado direito, está esquematizado uma recombinação intracromossômica que ocorre devido cruzamento (indicado por "X") dentro de um alelo, resultando em um deleção ou um cromossomo em anel (não viável). Imagem adaptada de McDonald-McGinn *et al.*, 2015.

2. SÍNDROME DE MICRODELEÇÃO 22q11

Distúrbios genômicos, resultantes de recorrentes CNVs, foram descritos nos cromossomos 2, 7, 15, 16, 17 e 22 (Lupski, 1998; Shaffer and Lupski, 2000). Dentre estes, destaca-se a região q11 do cromossomo 22, uma área rica em genes que apresenta um conjunto de regiões de LCRs as quais predispõe à deleção ou duplicação dessa região (Guo *et al.*, 2011). Dessa forma, a deleção (que geralmente possui 3Mb) na região 1 banda 1 do braço longo (q) do cromossomo 22 é considerada a etiologia da Síndrome de Microdeleção 22q11 (SD22q11) (Figura 2).

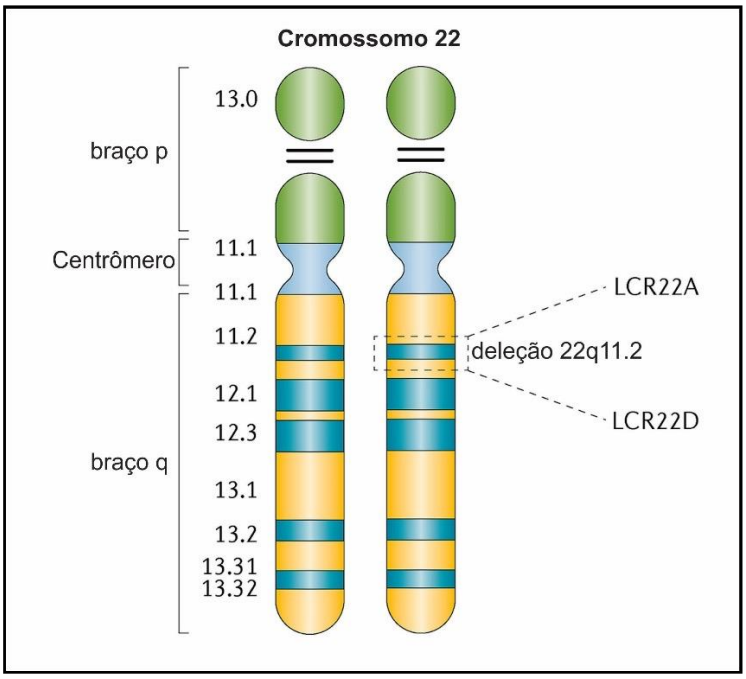


Figura 2. Representação do cromossomo 22 mostrando os braços curtos (p), braços longos (q) e o centrômero. A deleção na região 22q11 ocorre no braço longo de um dos dois cromossomos (representado pelas linhas tracejadas) devido a recombinação homóloga não alélica (NAHR) entre as regiões de repetição de pequeno número de copias (LCRs) A e D (Modificado de McDonald-McGinn *et al.*, 2015).

2.1. Nomenclatura

A Síndrome de microdeleção 22q11 foi descrita em várias partes do mundo de diferentes formas e em diferentes momentos. Consequentemente diferentes nomes foram atrelados a essa condição dando a entender que existiam diferentes desordens relacionadas a deleção 22q11 (Robin and Shprintzen, 2005). Wulfsberg e colaboradores (1996) compararam a evolução do estudo referente a deleção 22q11 com a fábula indiana que descreve um grupo de homens cegos tentando descrever um elefante cada um examinando uma parte separada (McDonald-McGinn, Zackai and Low, 1997).

A primeira descrição publicada foi o relato de um grupo de pacientes com voz anasalada e diminuição da mímica facial realizada por Sedlačkova em 1955 (Rosa et al., 2009). No decorrer do anos, outros clínicos descreveram diferentes pacientes que apresentavam anomalias do arco aórtico associado a outros sinais como dismorfia facial, deficit cognitivo ou deficiência imune. Porém, a primeira descrição clínica formal foi publicada por DiGeorge que descreveu que os pacientes sindrômicos apresentavam: hipopartireoidismo, defeitos cardíacos conotruncais, dismorfismo facial e imunodeficiência (DIGEORGE, 1968). Assim, ela ficou conhecida na época como Síndrome de DiGeorge (SDG; OMIM#188400).

Em 1976, Kinouchi e colaboradores relataram uma síndrome a qual se caracterizava por cardiopatia congênita e aparência facial típica que ele denominou como Síndrome da anomalia facial conotruncal (CTAF). Na mesma época, Shprintzen relatou um quadro de cardiopatia congênita, voz anasalada com anomalias de palato, aparência facial característica e dificuldades de aprendizagem em diferentes pacientes e usou o termo Síndrome Velocardiofacial (SVCF; OMIM#192430) para caracterizar o quadro, que também ficou conhecido como Síndrome de Shprintzen (Robin and Shprintzen, 2005).

No início de 1990 diversos grupos associaram uma microdeleção 22q11.2 em pacientes com a sequencia DiGeorge mas que possuíam diferentes manifestações clínicas. Essa foi a base para associar os diferentes casos já relatados a uma mesma região cromossômica (Goodship *et al.*, 1998; Swillen *et al.*, 2000). Dessa forma ficou claro que não havia diferentes síndromes e sim diferentes manisfestações clínicas para uma deleção na mesma região cromossômica (Robin and Shprintzen, 2005).

Com o intuito de unificar as diferentes nomenclaturas relacionados à deleção 22q11, Bassett e colaborados sugeriram o termo Síndrome de deleção 22q11, o qual é utilizado até hoje (Rosa *et al.*, 2009). Assim, depois da ampla utilização da técnica de FISH, as síndromes anteriormente conhecidas como SDG, SVCF e a CTAF passaram a ser referidas por sua etiologia cromossômica como SD22q11 (Bassett *et al.*, 2011).

2.2. Frequência

Estima-se que a frequência média da ocorrência da SD22q11 é de 1 em 4000 a 9800 nascidos vivos, porém ainda não há estudos que confirmem essa incidência ao nascimento (Burnside, 2015; Panamonta *et al.*, 2016; Dugoff, Mennuti and McDonald-McGinn, 2017). Porém, a variável nomenclatura, a variabilidade fenotípica e, consequentemente, a dificuldade no diagnóstico tornam o cálculo da frequência subestimado (Miller *et al.*, 2010; Rosenfeld *et al.*, 2013).

Apesar da situação observada para a SD22q11, essa é considera a síndrome de microdeleção mais frequente e a segunda causa mais comum de atraso no desenvolvimento e de doença congênita grave, após a síndrome de Down (Burnside, 2015). Além disso, é responsável por aproximadamente 2,4% dos indivíduos com deficiência no desenvolvimento e cerca de 10% a 15% dos pacientes com tetralogia de Fallot (Bassett *et al.*, 2011).

Do ponto de vista hereditário, a SD22q11 é herdada de forma autossômica dominante, em até 10% dos indivíduos, e observa-se que indivíduos do sexo masculino e feminino são igualmente afetados (Lindsay, 2001). Poŕem, a maioria das deleções do cromossomo 22q11

(cerca de 90%) são esporádicas (deleção *de novo*) (Bassett *et al.*, 2011). A identificação de uma deleção esporádica implica em baixo risco de recorrência (1 a 3%), enquanto que em uma deleção herdada, há um risco de 50% de transmissão da deleção (McDonald-McGinn *et al.*, 1997).

2.3. Etiologia

A área pericentromérica do cromossomo 22 tem uma complexa estrutura que contém oito LCRs distintos (LCR22A-H) com alta homologia entre si levando a ocorrência de NAHR (Sullivan, 2019). Dessa forma, as deleções observadas são causadas por um evento de recombinação homóloga não alélica durante a meiose (Scambler, 2000; Bittel *et al.*, 2009; Rosa *et al.*, 2009). Na maioria das vezes, a deleção é secundária a um erro de pareamento das sequências de DNA entre dois cromossomos 22 (intercromossômica), de forma que a LCR proximal de um deles reconhece a distal do outro (Scambler, 2000; Bittel *et al.*, 2009; Rosa *et al.*, 2009).

De acordo com a sua posição em relação ao centrômero, as deleções envolvendo as regiões LCR22 são designadas como: deleção na região proximal (A–B, A–D, A–E, A–F), central (B–D, C–D), distal tipo 1 (C–E, D–E, D–F), tipo 2 (E–F) e tipo 3 (inclui o gene SMARCB1) (Lindsay, 2001; Burnside, 2015) (Figura 3). Como pode ser observado, as extensões de deleções 22q11 são variáveis, mas estudos indicam que aproximadamente 90% das deleções que ocorrem na SD22q11 se estendam da LCR22-A a LCR22-D, compreendendo 3Mb. Devido a frequência observada, essa extensão é conhecida como Região Tipicamente Deletada (*Typically Deleted Region* - TDR) e abrange um número estimado de 90 genes conhecidos ou preditos incluindo 46 genes codificadores de proteínas e microRNAs, 10 RNAs não codificantes e 27 pseudogenes (Dugoff, Mennuti and McDonald-McGinn, 2017). Em aproximadamente 8% dos pacientes ocorre uma deleção menor de 1.5Mb, que compreende em torno de 24 genes, e a minoria dos pacientes apresentam deleções atípicas, com um número variável de genes afetados (Edelmann, Pandita and Morrow, 1999; Yamagishi and Srivastava, 2003; Rosa *et al.*, 2009; McDonald-McGinn *et al.*, 2015).

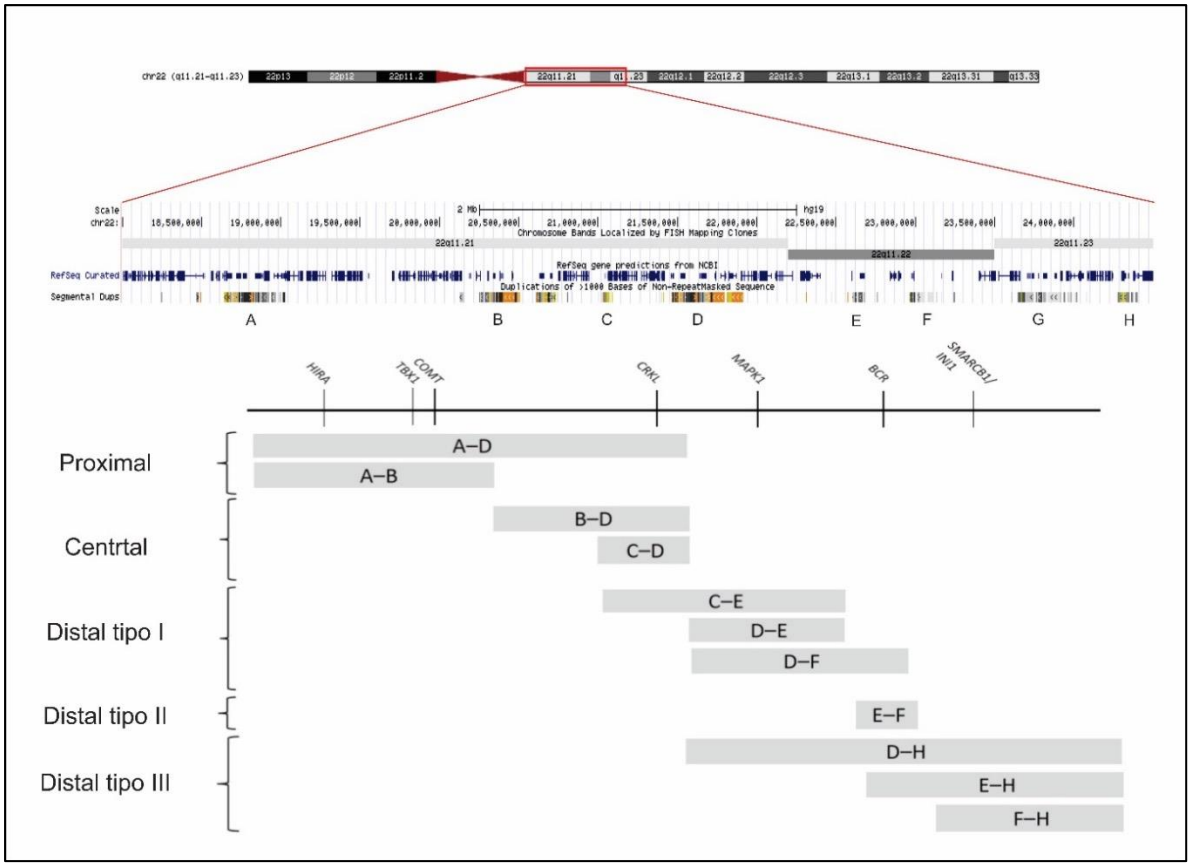


Figura 3. Região proximal 22q11 de acordo com UCSC Genome Browser. A classificação das deleções envolvendo as regiões LCR22 são esquematizadas. Deleção na região proximal (A–B, A–D, A–E, A–F), central (B–D, C–D), distal tipo 1 (C–E, D–E, D–F), tipo 2 (E–F) e tipo 3(inclui o gene SMARCB1). Além disso, genes chaves presentes na região podem ser observados. Imagem modificada de Burnside., 2015.

Sandrin-Garcia (Sandrin-Garcia *et al.*, 2007) destaca que pacientes que apresentam a deleção de 1.5Mb, compartilham sinais clínicos característicos da síndrome com pacientes que possuem a deleção de 3Mb. Dessa forma, o aspecto fenotípico típico da SD22q11 normalmente ocorre independentemente do tamanho da deleção da região 22q11 e, por isso, ainda não foi possível realizar uma correlação do tamanho da deleção com o fenótipo. Além disso, é importante salientar que indivíduos afetados possuem a deleção em apenas um dos cromossomos 22, presume-se, portanto, que essa é uma síndrome que envolve genes haploinsuficientes (Bassett *et al.*, 2011).

De acordo com o tamanho e a localização da deleção, diferentes genes podem ser afetados e alguns desses genes têm sido associados ao fenótipo da SD22q11. O principal gene considerado como responsável pelos sinais clássicos da síndrome é o *TBX1,* o qual foi identificado em modelos murinos na região proximal da deleção, entre os LCRsA-B (Dugoff, Mennuti and McDonald-McGinn, 2017). É um membro da família gênica T-box, um grupo

com fatores de transcrição evolutivamente conservados que compartilham domínios de ligação ao DNA, chamados T-box (Bollag *et al.*, 1994).

Foi demonstrado que a mutação no gene *TBX1* produz um amplo espectro fenótipo, incluindo artérias do arco aórtico anormais, comumente associadas a Síndrome de microdeleção 22q11(Jerome and Papaioannou, 2001). Gao e colaboradores (Gao *et al.*, 2015) utilizaram camundongos com o gene *TBX1* silenciado para determinar a base molecular dos defeitos dentários observados em pacientes com a síndrome e determinaram que esse gene é essencial para o desenvolvimento embrionário.

Estudos em modelos animais também demostraram que a haploinsuficiência de *TBX1* provoca uma remodelação e crescimento anormal da faringe e estruturas relacionadas a ela, o que explica vários achados clínicos da síndrome, incluindo dismorfia facial, defeitos no palato, hipoplasia das glândulas paratireoides e timo, problemas odontológicos, de alimentação e deglutição (Scambler, 2000; Jerome and Papaioannou, 2001; Gao, Li and Amendt, 2013). O gene *TBX1* também vem sendo associado a outros órgãos e sistemas, como o estudo de Chen e colaboradores (Chen *et al.*, 2016) que demonstra o papel desse gene no desenvolvimento e funções do ouvido.

Além do gene *TBX1*, outros genes são considerados críticos na manifestação das características clínicas principais da síndrome, como os genes *HIRA* e *COMT* (Burnside, 2015). O gene *HIRA* (*histone cell cycle regulator*) está localizado na região TDR e foi demostrado que ele atua, em modelos animais, na crista neural e nos tecidos neurais durante o desenvolvimento embrionário e desempenha um papel essencial na formação do coração (Ju *et al.*, 2016). Estudos de Jae-Hyun Yang e colaborados (Yang *et al.*, 2016) demonstram o papel do *HIRA* na expressão de genes miogênicos, o que corrobora com a ideia de que a deleção desse gene pode afetar o desenvolvimento cardíaco e, por conseguinte, causar defeitos cardíacos congênitos como os observados na SD22q11.

Outro gene localizado na região TDR é o *COMT* que codifica a enzima *Catechol-O-Methyltranferase,* essa enzima está envolvida na decomposição de neurotransmissores incluindo a dopamina, epinefrina e norepinefrina (Radoeva *et al.*, 2014). Estudos indicam sua atuação na degradação de catecolaminas e, por isso, ele é considerado como um dos candidatos para explicar os efeitos neurológicos apresentado por pacientes com SD22q11 (Zeitz *et al.*, 2013).

Entretanto, o locus do *TBX1* nem sempre está incluso na deleção 22q11, deleções centrais envolvendo os LCRB-D e C-D por exemplo não incluem o gene *TBX1*, nem o *HIRA* (Burnside, 2015). Por isso, é possível considerar que outros genes possam influenciar a

expressão das características clínicas observadas ou desempenhar papéis importantes na etiologia da síndrome (Jerome and Papaioannou, 2001). O gene *CRKL* é considerado como candidato nas deleções centrais e o gene *MAPK1* para a deleção distal (Burnside, 2015; Racedo *et al.*, 2015). Ainda não está claro os efeitos dos genes deletados na região da deleção distal tipo 2, mas para a deleção distal tipo 3 o gene *SMARCB1* é considerado como um gene crítico devido a alta taxa de tumores rabdóides malignos em indivíduos com essa deleção (Burnside, 2015).

Na região das deleções centrais, o gene *CRKL* é considerado como crítico para a fisiopatologia da síndrome. O *CrK-like* ou *CRKL* é um gene codificador de proteínas envolvidas na cascata de sinalização de IL-2 e de interferon tipo I. Estudos indicam que o *CRKL* pode ter um papel funcional na deficiência de células T em pacientes com SD22q11 (Giacomelli *et al.*, 2016). Além disso, acredita-se que a haploinsuficiencia desse gene pode influenciar no desenvolvimento de anomalias cardíacas em indivíduos que apresentam uma deleção distal envolvendo os LCR22B-D e LCR22C-D (Dugoff, Mennuti and McDonald-McGinn, 2017).

O gene *MAPK1* ou *ERK2* localiza-se na região da deleção distal tipo 1 e está associado a sinalização intracelular. Samuels e colaboradores (Samuels *et al.*, 2008) examinaram o papel desse gene no desenvolvimento neurológico em modelos murinos e observaram que a perda de *ERK2* resulta na redução da espessura cortical, além da formação diminuída de neurônios. Por fim, estudos mostram uma associação do gene supressor de tumor *SMARCB1* com tumores rabdóides em pacientes com a deleção distal 22q11 (Hacıhamdioğlu, Hacihamdioglu and Delil, 2015).

2.4. Caracteristicas clínicas

A SD22q11 caracteriza-se por um espectro fenotípico bastante amplo, com efeitos pleiotrópicos que resultam no acometimento de praticamente todos os órgãos e/ou sistemas, altamente variáveis com mais de 180 sinais clínicos já descritos, tanto físicos como comportamentais (Carlson *et al.*, 1997; Robin and Shprintzen, 2005; Hay, 2007). Porém não há nenhum sinal clínico ou conjunto de sinais que ocorrem em todos os indivíduos com a deleção 22q11, indicando a inexistência de manifestações obrigatórias para a síndrome (Hay, 2007; Shprintzen, 2008).

Apesar da diversidade de características clínicas associadas a SD22q11 pode-se citar alguns sinais clássicos como: doença cardíaca congênita, fissura de palato, insuficiência velofaríngea, aspectos faciais característicos, dificuldade de aprendizado e deficiência

imunológica (Dugoff, Mennuti and McDonald-McGinn, 2017). Achados adicionais incluem a hipocalcemia, problemas para se alimentar, anomalias renais, perda da audição, deficiência do hormônio de crescimento, desordens autoimunes, convulsões, anormalidades esqueléticas, doenças psiquiatricas, entre outras (McDonald-McGinn *et al.*, 2015). Dessa forma os principais aspectos clínicos da SD22q11 são:

- **Cardiopatias congênitas:**

As cardiopatias congênitas são características chave da SD22q11 e são compostas por alterações estruturais e funcionais do coração presentes ao nascer, além de serem considerados o maior fator de mortalidade da síndrome (Bales, Zaleski and McPherson, 2010). Observa-se que os defeitos cardíacos mais frequentes envolvem as vias de saída do coração (conotruncais), dentre eles podemos citar: anomalias do arco aórtico como interrupção do arco aórtico do tipo B (IAA-B), *truncus arteriosus*, tetralogia de Fallot, defeito no septo ventricular, atresia pulmonar, estenose pulmonar e arco aórtico à direita (Hay, 2007; Dugoff, Mennuti and McDonald-McGinn, 2017). A prevalência das malformações cardíacas, de acordo com Habel e colaboradores (Habel *et al.*, 2014) varia de 80% a 92% na infância mas essa estimativa varia de acordo com o estudo. Além disso, a deleção do cromossomo 22q11 parece ser a segunda causa mais comum de doença cardíaca congênita, depois da síndrome de Down (Robin and Shprintzen, 2005).

- **Anomalias palatinas:**

Anomalias palatinas são reportadas em torno de 49% a 69% de pacientes com SD22q11 sendo as mais frequentes: Insuficiência velofaringea(IVF), fissura de palato e úvula bífida (Monteiro *et al.*, 2013). A IVF é definida como uma alteração estrutural do mecanismo velofaríngeo, cujo sintoma mais característico é a hipernasalidade, associado à emissão nasal de ar e a fraca pressão intra-oral e os distúrbios articulatórios compensatórios (Fukushiro, 2007). Essa é anomalia de palato mais frequente, podendo ser observada em 29% a 50% dos casos e pode ser consequência de alterações funcionais ou estruturais (Monteiro *et al.*, 2013). Outras manifestações observadas na cavidade oral são: atraso de erupção dentária, hipoplasia ou hipomielinização do esmalte, alterações da morfologia dentária, cáries e hipodontia (Cummings, McCauley and Baylis, 2015; McDonald-McGinn *et al.*, 2015).

- **Aspectos faciais:**

Caracteristicas faciais “tipicas” são descritas, dentre elas podemos citar: aumento do comprimento vertical da face, blefaroptose ou *hooded eyelids* (devido à configuração da estrutura óssea, as pálpebras ficam parcialmente cobertas por pele quando os olhos estão abertos), fendas palpebrais estreitas, base nasal larga, hipoplasia alar, hipertelorismo, anormalidades estruturais na orelha e retrognatia (Hay, 2007; Monteiro *et al.*, 2013). Porém, a face denominada “típica” não está sempre presente já que outros dismorfismos são relatadas, ou seja, os aspectos faciais são variáves (Figura 4). Além disso essas características podem não ser evidentes em recém-nascidos, crianças, ou devido a etnia (Hay, 2007).

Figura 4. Fotografias de individuos diagnosticados com SD22q11 demostrando como o aspecto fenotipico é variavel para essa síndrome (Imagens adaptadas de Ben-Shachar et al., 2008; Nogueira et al., 2008; Digilio et al., 2009; Garavelli et al., 2011; Michaelovsky et al., 2012; Rump et al., 2014; Bengoa-Alonso et al., 2016).

- **Dificuldade de aprendizagem:**

Problemas educacionais e de desenvolvimento são frequentemente reportados na SD22q11, como atenção seletiva, dificuldade de aprendizagem, dificuldades na visão espacial

e dificuldades na habilidade fonêmica (Habel *et al.*, 2014). O Transtorno do déficit de atenção com hiperatividade (ADHD) também é relatado porém é difícil atribuir esse sinal a síndrome ou a um achado comum em crianças com deficiências de desenvolvimento (Hay, 2007).

- **Deficiências imunológicas:**

As principais imunodeficiências observadas são: defeito no timo ou funções de células T, como diminuição no número ou função dessas células, e defeitos em anticorpos (Habel *et al.*, 2014). Doenças autoimunes também são reportadas, incluindo hipotireoidismo, hipertireoidismo, anemia hemolitica autoimune, monoartrite, reumatoide juvenil, artrites, vitiligo, neutropenia autoimune, anemia aplasica e doenças celíacas (Hay, 2007).

- **Aspectos psicológicos:**

Transtornos comportamentais e psiquiátricos são relatados em diversos pacientes, onde os sinais variam de ansiedade e depressão à psicose e esquizofrenia (Norkett *et al.*, 2018). Estudos indicam que pacientes com a deleção 22q11 apresentam um risco aumentado em desenvolver esquizofrenia, uma doença psicologia séria que geralmente ocorre na fase da adolescência ou fase adulta inicial (Hay, 2007; Bassett *et al.*, 2017).

2.5. Diagnóstico

O diagnóstico é realizado baseando-se nas características clínicas que chamam a atenção dos pais, da família e do médico, as quais podem variar dependendo da idade do paciente e das características de cada caso. No entanto, há alguns sinais clássicos que são utilizados para o diagnóstico, tais como: deficiência no desenvolvimento e/ou deficiência de aprendizado, cardiopatias congênitas, defeitos no palato, regurgitação nasal, problemas de comportamento, doenças psiquiátricas; imunodeficiência, hipocalcemia, e traços faciais característicos (Bassett *et al.*, 2011).

Após a análise do caso e suspeita da SD22q11, é necessário confirmar o diagnóstico realizando análises genéticas, por meio de técnicas como FISH (*fluorescence in situ hybridization*), MLPA (*Multiplex Ligation-dependent Probe Amplification*) e/ou arrayCGH (*microarray comparative genome hybridization*). Apesar das técnicas de diagnósticos serem precisas e relativamente fáceis de serem implementadas, a dificuldade no reconhecimento da condição e/ou familiaridade com os métodos de testes genéticos em conjunto com a grande variabilidade fenotípica da síndrome dificultam o diagnóstico precoce. O atraso no diagnóstico

leva a uma intervenção tardia e, dessa forma, o prognostico pode ser afetado. Em contraste, o diagnóstico e intervenção precoce pode ajudar no acompanhamento e tratamento dos pacientes (Dugoff, Mennuti and McDonald-McGinn, 2017).

A maioria dos pacientes apresentam uma deleção pequena de 3Mb na região 22q11, detectável por FISH, uma técnica que integra a utilização da citogenética clássica com a genética molecular, por meio do uso de sondas de DNA marcadas com material fluorescente que identificam regiões específicas do genoma. Dessa forma, historicamente, o teste diagnóstico de FISH é o mais utilizado na detecção de deleção na região 22q11.2, onde se utilizam sondas comerciais como N25 ou TUPLE para mapear a região de LCR22-A a LCR22-B (McDonald-McGinn *et al.*, 2015; Dugoff, Mennuti and McDonald-McGinn, 2017; Morrow *et al.*, 2018). Porém a técnica de FISH possui limitações por ser um teste direcionado, ou seja, exige uma suspeita clínica suficiente para a análise da região correta (Kuo, Signer and Saitta, 2018). Outra limitação é que o teste pode não detectar deleções e duplicações que estão fora da região coberta pelas sondas usadas no FISH. Assim, pacientes estudados apenas com este método podem passar despercebidos (Bassett *et al.*, 2011; Kuo, Signer and Saitta, 2018).

Para superar essas dificuldades e detectar deleções fora dos LCRsA-B, algumas técnicas moleculares mais atuais podem ser utilizadas como o arrayCGH e o MLPA (Dugoff, Mennuti and McDonald-McGinn, 2017). O MLPA, técnica desenvolvida por Schouten et al. (2002), registrada pela marca comercial MRC-Holland®, é baseada na reação de PCR (Reação em cadeia da polimerase) multiplex quantitativo para a determinação do número de cópias relativa de uma sequência alvo. Mostrou-se bem sucedida no diagnóstico da SD22q11 já que utiliza sondas que se ligam ao longo de toda região 22q11 e com isso pode detectar deleções típicas e atípicas (Bassett *et al.*, 2017; Morrow *et al.*, 2018). Assim, por ser um método rápido e efetivo não somente para a detecção, mas, também, para a determinação do tamanho das deleções recorrentes e duplicações na região proximal 22q11 o número de pacientes diagnosticados com deleção 22q11 está aumentando (Bassett *et al.*, 2017).

A técnica de arrayCGH é baseada em microarrays de DNA que detecta mudanças no número de cópias de uma sequência de DNA (Hupé *et al.*, 2004). Esta metodologia tem sido amplamente utilizada já que o arrayCGH fornece diversas vantagens como uma melhor resolução e cobertura (Albertson and Pinkel, 2003). Geralmente essa técnica é utilizada para diagnóstico quando não foi possível realizar um diagnóstico definitivo baseado na avaliação clínica ou quando o MLPA não está clinicamente disponível (Morrow *et al.*, 2018).

2.6. Tratamento

Devido ao amplo número de manifestações clínicas associadas a SD22q11, o manejo clínico dos pacientes é, de certa forma, complexo (Shprintzen, 2008). Logo, o tratamento deve ser direcionado para melhor atender cada indivíduo, dependendo de sua idade ou estágio de desenvolvimento, e seus sinais clínicos particularmente associados à gravidade e necessidade de tratamento (Hay, 2007).

Todo o acompanhamento e tratamento deve ser realizado por uma equipe multiprofissional, como médicos, psicólogos, fonoaudiólogos, fisioterapeutas e enfermeiros para atender todos os aspectos resultantes da SD22q11 (Bassett, 2011). O acompanhamento regular do crescimento, estado endócrino, hematológico e função imunológica são de extrema importância para permitir uma intervenção precoce e o auxílio na manutenção da saúde (Habel *et al.*, 2014).

O aconselhamento genético para SD22q11 inclui uma análise sobre a prevalência, etiologia, detecção, variabilidade, intervenções e opções pré-natais e de preconcepção. No aconselhamento deve-se incluir informações atualizadas sobre as condições comumente associadas à síndrome, além de esclarecer como é o desenvolvimento nas suas diferentes fases. Além disso, informações sobre estratégias de cuidado, recursos locais, e apoios devem ser fornecidas aos pacientes, suas famílias e profissionais envolvidos (Bassett, 2011).

3. APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina (*Machine learning*- ML), é uma área da Inteligência Artificial que utiliza máquinas e computadores para otimizar um critério de desempenho utilizando exemplos de dados e experiências de aprendizagem (Alpaydin, 2010). Tem como um dos principais objetivos reconhecer padrões complexos e tomar decisões inteligentes com base em dados (Mitchell, 1997; Alpaydin, 2010). Diferentes métodos são utilizados, os quais podemos destacar:

- Aprendizado Supervisionado: método utilizado com o objetivo de descobrir a relação entre os atributos de entrada (*input*) e uma classe, a relação descoberta pode ser utilizada como um modelo (Maimon and Rokach, 2015). Assim, os algoritmos desta categoria deduzem uma função a partir dos dados de treinamento, onde o objetivo é que essa função seja capaz de predizer a saída para qualquer entrada válida, após ter visto um número suficiente de exemplos de treinamento. Para atingir este objetivo, o algoritmo de classificação deve ter capacidade de generalização para que possa prever, de maneira aceitável, a classe para dados ainda não vistos (Breve, 2010).

- Aprendizado Não Supervisionado: nesse método, os dados de treinamento compreendem somente dados de entrada (*input*), sem rótulos ou valores de saída. Os algoritmos desta categoria buscam determinar como os dados estão organizados afim de encontrar padrões (Alpaydin, 2010; Breve, 2010).

Dentre os modelos de aprendizado de máquinas supervisionado destacamos o modelo de Classificação que tem como objetivo mapear os dados de entrada em classes predefinidas, sendo exemplos de classificadores: *support vector machines*, árvores de decisão, *probabilistic summaries*, *algebraic function* (Maimon and Rokach, 2015). O modelo de classificação do tipo árvore de decisão foi utilizado neste trabalho e, por isso, alguns aspectos desse tema serão abordados nos tópicos a seguir.

3.1. Dados de entrada ou *Input*

Os dados de entrada para o aprendizado de máquinas é um conjunto de instâncias. As instâncias (os exemplos para o ML) são os objetos que serão classificados, associados ou agrupados. De forma geral, cada instância é um exemplo independente, individual do conceito que será estudado onde são caracterizadas por um conjunto pré-determinado de atributos (Witten, I. H. , Frank, E., & Hall, 2011). De acordo com cada conjunto de dados, observa-se diferentes tipos de atributos, porém em mineração de dados tipicamente se trabalha com valores numéricos, nominais e/ou categóricos (Witten, I. H. , Frank, E., & Hall, 2011).

Dessa forma, o dado que é utilizado para realizar o treinamento é normalmente representado em forma de tabela e é denominado Conjunto de treinamento ou *Training set*. Cada linha representa uma única instância e cada coluna corresponde a um atributo que caracteriza as instâncias. Além disso, em tarefas de classificação, uma das colunas corresponde ao atributo de destino (classe) que se tenta prever (Maimon and Rokach, 2015).

Apresentamos aqui o exemplo do conjunto de dados meteorológicos de Witten, I. H. , Frank, E., & Hall, 2011 (Figura 5). Nesse exemplo, avalia-se as condições de tempo adequadas para jogar algum tipo de jogo. Assim, cada exemplo (linhas da tabela) são as instâncias e os atributos (colunas da tabela) compreendem: *outlook* (aspecto), *temperature* (temperatura), *humidity* (humidade) e *windy* (vento). A classe ou consequência seria *play*, ou seja, a decisão de jogar ou não jogar.

No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Nominal	Nominal	Nominal	Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Figura 5. Conjundo de dados metereologicos. Fonte: Witten, I. H., Frank, E., & Hall, 2011.

Ao começar a trabalhar em um problema de mineração de dados, é necessário primeiro reunir todos os dados em um conjunto de instâncias. Realizar a integração dos dados de diferentes fontes pode ser um desafio, pois cada fonte ou banco de dados pode utilizar diferentes registros, convenções, graus de agregação de dados e, diferentes tipos de erros podem existir (Witten, I. H. , Frank, E., & Hall, 2011). Assim, os dados devem ser montados, integrados e padronizados para sua utilização.

Em vários algoritmos de aprendizado de máquina, o tamanho do conjunto de treinamento e o desempenho preditivo estão correlacionados positivamente. Ou seja, de acordo com os recursos computacionais disponíveis, é preferível utilizar o maior conjunto de treinamento possível (Maimon and Rokach, 2015).

3.2. Árvores de decisão

Árvores de decisão são um dos exemplos de classificadores citados anteriormente e é o algoritmo utilizado neste trabalho. Compreende uma técnica eficiente em predizer e explicar a relação entre medidas sobre determinado alvo. Utiliza-se uma abordagem conhecida como *divide-and-conquer* que consiste em recursivamente desmembrar o problema em dois ou mais subproblemas (Witten, I. H. , Frank, E., & Hall, 2011).

De forma simplificada, uma Árvore de Decisão é uma lista de perguntas (ramos da árvore) com suas respostas, que pode ser do tipo “sim” ou “não”, hierarquicamente arranjadas, que levam a uma decisão (Souto *et al.*, 2003; Landrum *et al.*, 2018). Dessa forma, as árvores de decisão são utilizadas para classificar um objeto ou instância em um conjunto predefinido de classes baseado em seus atributos (Maimon and Rokach, 2015).

A estrutura de uma árvore de decisão é constituída basicamente por “nós”. Esses nós podem ser classificados em: “nó raiz” que não possui arestas de entrada; “nós internos” que representam o teste realizado em um atributo; “nó terminal” ou “folhas” representado pelo valor da variável de decisão ou a classe em que os atributos são classificados. Além disso, define-se como “ramos” as conexões entre os nós que contêm os valores dos atributos de cada variável decisória (Figura 6) (Witten, I. H. , Frank, E., & Hall, 2011; Maimon and Rokach, 2015). Para “ler” a árvore de decisão basta começar pelo nó raiz, seguindo cada teste até que uma folha seja alcançada.

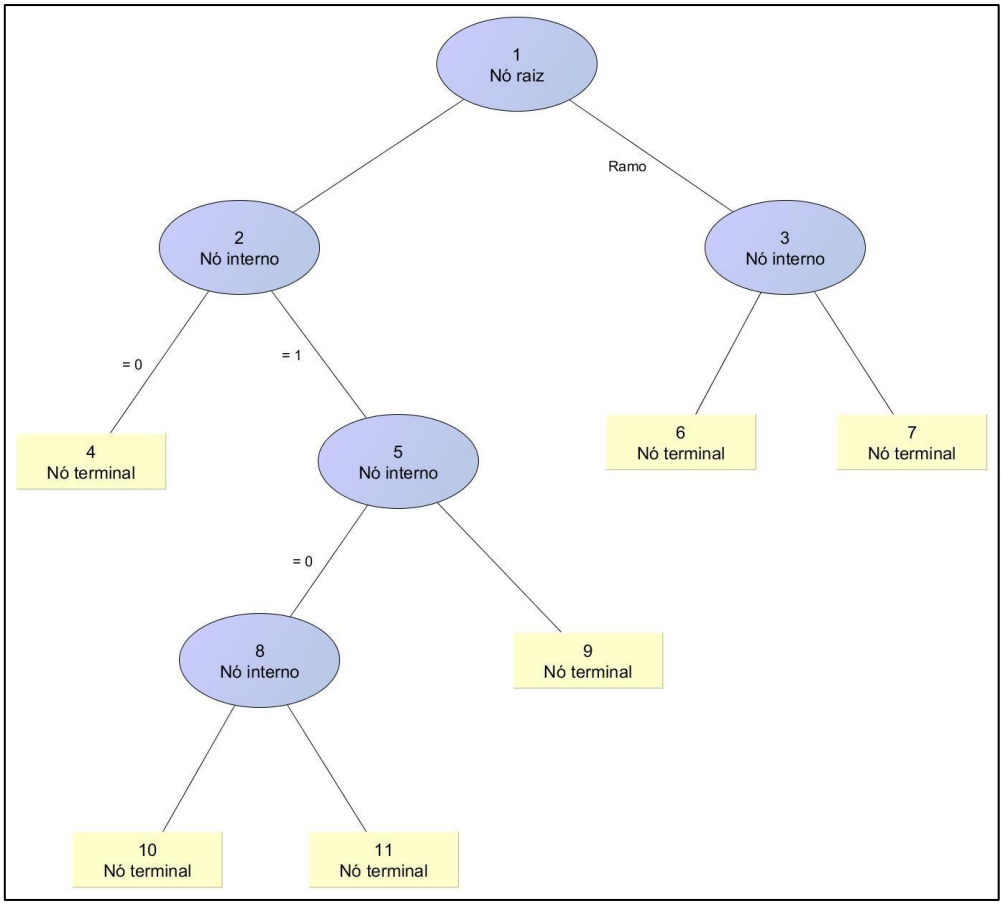


Figura 6. Estrutura geral de uma árvore de decisão. O nó raiz ou principal está representado pelo número 1. Os números 2, 3, 5 e 8 representam os nós internos. As folhas ou nós terminais (com o valor da variável de decisão) são os números 4, 6, 7, 9, 10 e 11. Ramos são as conexões entre os nós que contêm os valores dos atributos de cada variável decisória.

Os atributos observados em uma árvore de decisão são, na maioria das vezes, um subconjunto de todos os atributos dentro de um conjunto de dados (Witten, I. H. , Frank, E., & Hall, 2011). Para selecionar qual atributo será utilizado como nó da árvore, ou seja, qual será o nó raiz e assim por diante, utiliza-se uma medida de pureza. Essa medida é a quantidade de "informação" que cada atributo possui e é quantificado em unidades chamadas *bits*. Ao contrário dos bits na memória do computador, a quantidade esperada de informações geralmente envolve frações de um *bit* - e geralmente é menor que 1 (Witten, I. H. , Frank, E., & Hall, 2011).

Os *bits* são calculados de acordo com o número de instâncias de determinada classe nas folhas. Calcula-se o ganho de informação para cada atributo e o atributo com maior informação é selecionado. Em seguida, o conjunto de dados é dividido de acordo com esse atributo, seleciona-se dentre os restantes outro atributo com maior informação e assim recursivamente. O processo de seleção do atributo como nó da arvore termina, idealmente, quando todos os nós folha são puros - isto é, quando eles contêm instâncias que possuem a mesma classificação. No

entanto, pode não ser possível alcançar essa situação, dessa forma, o processo termina quando os dados não podem ser mais divididos (Witten, I. H. , Frank, E., & Hall, 2011).

Para demostrar como as árvores de decisão são construídas vamos trabalhar com o exemplo do conjunto de dados meteorológicos apresentado anteriormente. No conjunto de dados meteorológicos há quatro atributos que poderiam ser o nó raiz e é preciso verificar o grau de pureza, os bits, para definir qual atributo seria o melhor. Assim o atributo que produz a folha mais pura seria a mais adequada (Figura 7) (Witten, I. H. , Frank, E., & Hall, 2011).

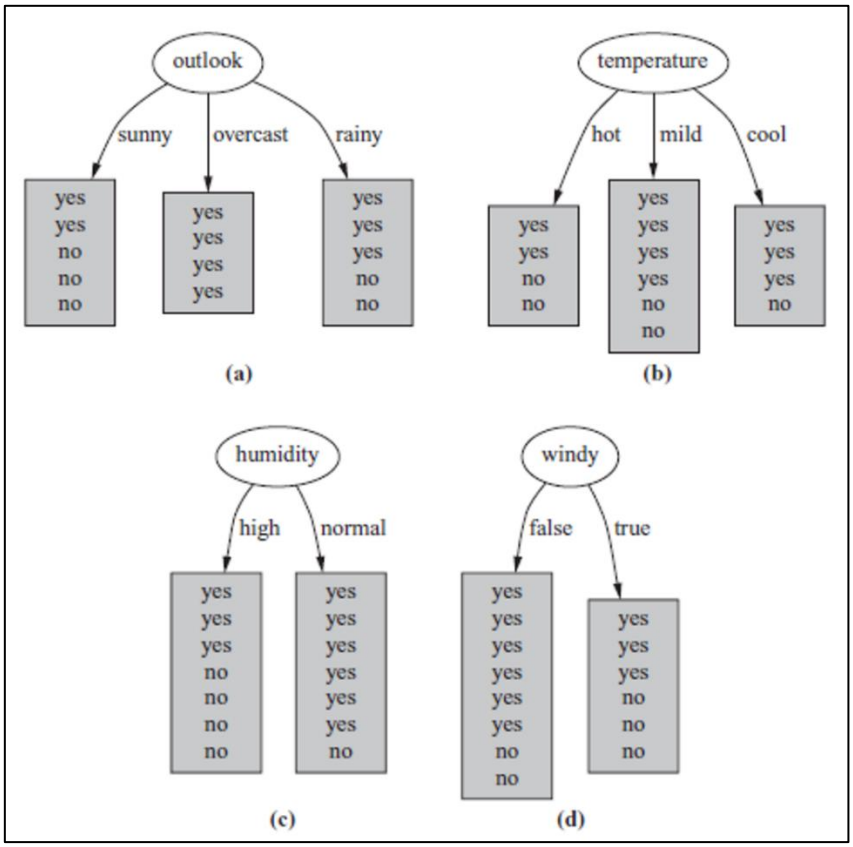


Figura 7. Possíveis árvores para o conjunto de dados metereológicos: (a) *Outlook* como nó raiz. (b) *Temperature* como nó raiz. (c) *Humidity* como nó raiz. (d) *Windy* como nó raiz. Fonte: Witten, I. H., Frank, E., & Hall, 2011.

Analisando a primeira árvore da figura 7(a) observamos que o número de classes “yes” e “no” das folhas são [2,3], [4,0] e [3,2], respectivamente. A quantidade de informação desses nós são:

- *info([2, 3]) = 0.971 bits*
- *info([4, 0]) = 0.0 bits*
- *info([3, 2]) = 0.971 bits*

A fórmula para calcular os bits de cada folha não será descrita nesse trabalho, mas pode ser encontrado no livro “Data mining: Pratical machine learning tools and techniques” de Witten, I. H. , Frank, E., & Hall, 2011. Em seguida calcula-se a informação média dessas folhas, levando em consideração o número de instâncias que chegou em cada folha (5 instâncias na primeira, 4 instâncias na segunda folha e 5 instâncias na terceira folha).

$$Info([2,3],[4,0][3,2]) = (5/14) \times 0.971 + (4/14) \times 0 + (5/14)x0.971$$
$$= 0,693 \text{ bits}$$

Essa média representa a quantidade de informação que esperamos que seja necessária para especificar a classe de uma nova instância, de acordo com a estrutura da árvore em questão (figura 7a). Se observamos a Figura 5, verificamos que o os exemplos de treinamento compreendem nove “yes” e cinco “no”, correspondendo a um valor de informação de:

$$info([9, 5]) = 0.940 \text{ bits}$$

Dessa forma, a árvore da figura 7a possui um ganho de informação de:

$$gain(outlook) = info([9, 5]) - info([2, 3], [4, 0], [3, 2]) = 0.940 - 0.693$$
$$= 0.247 \text{ bits}$$

Esse valor pode ser interpretado como o valor informativo da criação de uma ramificação no atributo do *outlook*. Assim, calcula-se o ganho de informação de cada atributo e será escolhido como nó o que tiver mais informações. No caso das árvores da figura 7, o ganho de informações de cada possível nó raiz é:

- *gain(outlook) = 0.247 bits*
- *gain(temperature) = 0.029 bits*
- *gain(humidity) = 0.152 bits*
- *gain(windy) = 0.048 bits*

Como o atributo *outlook* possui o maior ganho de informação, ele é utilizado como nó raiz. Esse processo continua recursivamente e a figura 8 mostra as possibilidades de mais uma ramificação a partir do ramo *sunny*.

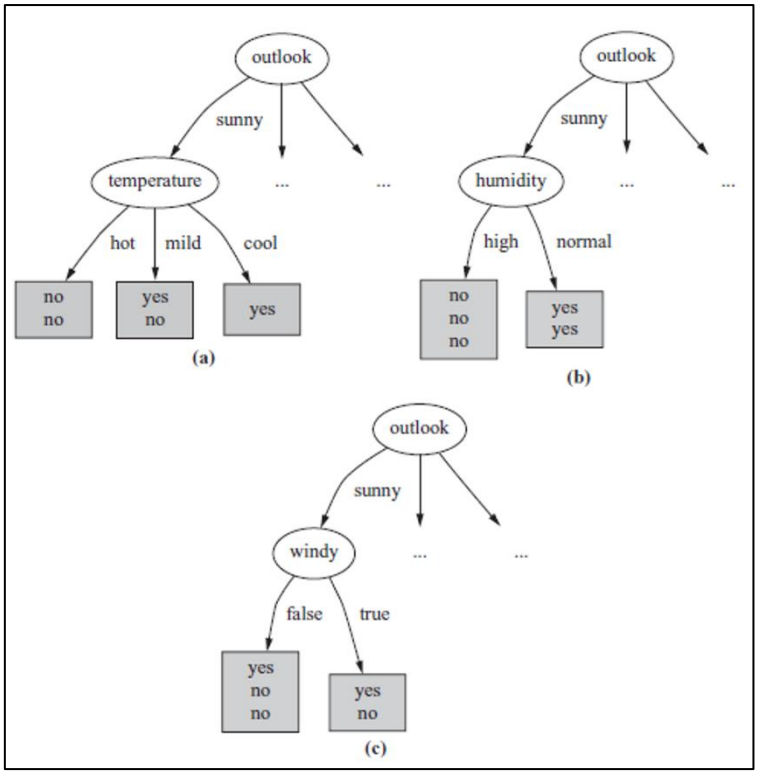


Figura 8. Possíveis árvores para o conjunto de dados meteorológico: (a) *Temperature* como nó do ramo *sunny*. (b) *Humidity* como nó do ramo *sunny*. (d) *Windy* como nó do ramo *sunny*. Fonte: Witten, I. H., Frank, E., & Hall, 2011.

O ganho de informação para cada nó é:

- *gain(temperature) = 0.571 bits*
- *gain(humidity) = 0.971 bits*
- *gain(windy) = 0.020 bits*

O atributo *humidity* é escolhido e como as folhas resultantes são puras, não há necessidade de dividir esses nós ainda mais. A aplicação desse processo nos outros ramos leva à árvore de decisão da Figura 9 para o conjunto de dados meteorológicos.

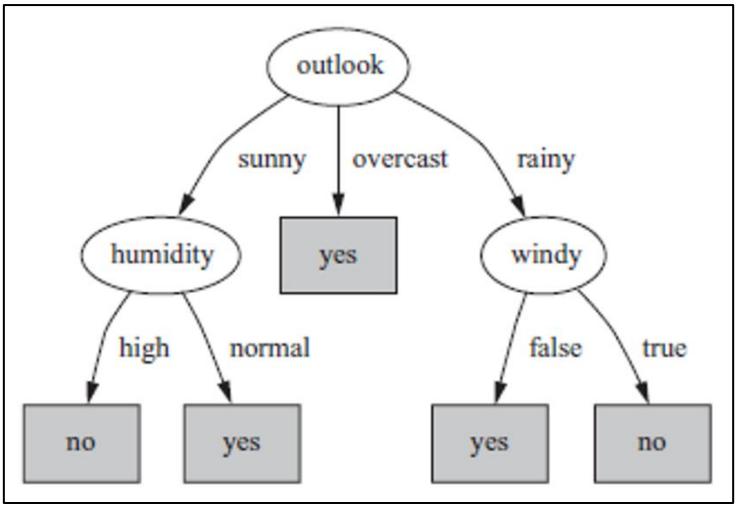


Figura 9. Árvore de decisão construída de acordo com o conjunto de dados meteorológico. Fonte: Witten, I. H., Frank, E., & Hall, 2011.

Para se obter uma árvore de decisão mais compreensível é preferível, naturalmente, uma árvore de decisão que não seja complexa. A complexidade de uma árvore normalmente é medida de acordo com o número total de nós, o número de folhas, o tamanho da árvore e o número de atributos utilizados como nós (Maimon and Rokach, 2015).

Além disso, uma vantagem das árvores de decisão é que elas costumam ser autoexplicativas, ou seja, não há a necessidade de profundo conhecimento em ML para se compreender a ordem de uma árvore de decisão (Maimon and Rokach, 2015). Esta característica torna-se significativa já que facilita o processo de análise das árvores de decisão e deliberação se o modelo aprendido é plausível, dadas as restrições do mundo real (Souto *et al.*, 2003).

3.3. Algoritmo de classificação J48

Algoritmos consistem em instruções seguidas por um computador para completar uma tarefa específica, como achar um certo padrão em um conjunto de dados (Deo, 2015). Algoritmos para construção de árvores de decisão são bem conhecidos, além de serem amplamente utilizados. Dentre esses algoritmos o ID3 e seu sucessor C4.5, criado por Ross Quinlan (Quinlan, 1992), estão entre os mais populares na comunidade de Aprendizado de máquinas onde sua função é descobrir padrões em um conjunto de dados e gerar um classificador em forma de árvore de decisão (Salzberg, 1994; Witten, I. H. , Frank, E., & Hall, 2011). O algoritmo de classificação utilizado nesse trabalho é conhecido como J48 e é uma implementação do algoritmo C4.5 na plataforma WEKA (*Waikato Environment for Knowledge Analysis*) (Hall *et al.*, 2009; Witten, I. H., Frank, E., & Hall, 2011).

O *software* WEKA foi desenvolvido na Universidade de Waikato, Nova Zelândia, e surgiu da necessidade de uma ferramenta unificada que permitiria aos pesquisadores fácil acesso a técnicas utilizadas em aprendizado de máquina (Hall *et al.*, 2009; Maimon and Rokach, 2015). A plataforma é escrita em Java e distribuída sob os termos da Licença Pública Geral GNU. Fornece uma interface para diversos algoritmos de aprendizado e métodos para pré e pós-processamento dos dados. Atualmente, o WEKA é reconhecido como um sistema de referência em mineração de dados e aprendizado de máquina (Hall *et al.*, 2009; Maimon and Rokach, 2015).

3.4. Medidas de desempenho

A matriz de confusão é uma ferramenta padrão para descrição de modelos estatísticos e é utilizada como uma indicação das propriedades de uma regra de classificação. Compõe essa matriz o número de elementos que foram classificados correta ou incorretamente para cada classe. Na diagonal (do lado esquerdo superior para o inferior direito) da matriz pode-se observar o número de observações que foram corretamente classificadas para cada classe e os elementos fora da diagonal representam o número de observações que foram incorretamente classificadas (Tabela 1) (Maimon and Rokach, 2015).

Tabela 1. Tabela representando uma Matriz de confusão onde observa-se os valores classificados corretamente ou incorretamente para cada classe.

		Valor previsto	
		Negativo	**Positivo**
Valor verdadeiro	**Negativo**	A	B
	Positivos	C	D

Baseado nos valores presentes na matriz de confusão é possível calcular as seguintes métricas (Witten, I. H. , Frank, E., & Hall, 2011; Reis, 2014; Maimon and Rokach, 2015):

- Taxa de verdadeiros positivos (VP): porcentagem de VP (D / (C + D))
- Taxa de falsos positivos (FP): taxa FP (B / (A + B))
- Acurácia: calculado utilizando a formula (A+ D) / (A + B + C + D)

- Precisão: a precisão de um algoritmo preditor é dada pelo número de instâncias classificadas corretamente dividido pelo número total de instâncias classificadas como positivas (D / (B + D))

Outra medida de desempenho utilizada é a *receiver operating characteristic curve* (curva ROC), uma medida que pode ser utilizada para análise entre a taxa de verdadeiro positivo e falso positivo. Dessa forma, a curva ROC descreve o desempenho de um classificador sem considerar a distribuição de classes ou os custos de erro. Para resumir as curvas ROC em uma única quantidade, às vezes usa-se o valor de *area under the curve* (AUC), já que, geralmente, quanto maior a área melhor o modelo (Witten, I. H. , Frank, E., & Hall, 2011).

Assim, o gráfico da curva ROC apresenta a taxa de VP no eixo vertical e a taxa de FP no eixo horizontal. O ponto ideal dentro do gráfico da curva ROC seria de FP = 0 e VP = 1, ou seja, todos os exemplos positivos são classificados corretamente (VP = 1) e nenhum exemplo negativo é classificado erroneamente como positivo (FP = 0) (Figura 10). Assim a curva ROC fornece um bom resumo do desempenho do modelo classificador (Reis, 2014; Maimon and Rokach, 2015).

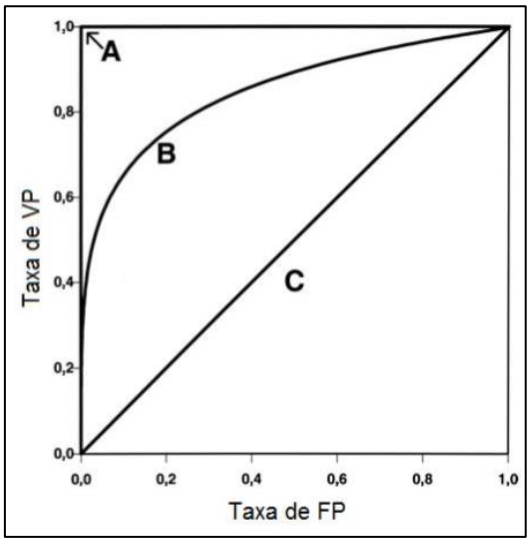


Figura 10. Exemplos de curva ROC em que o eixo X representa uma taxa de falso positivo e o eixo Y a taxa de verdadeiro positivo. VP = verdadeiro positivo; FP = falso positivo; (A) modelo ideal; (B) modelo real; (C) modelo aleatório (Imagem retirada de Reis, 2014).

4. REDE DE INTERAÇÃO PROTEINA-PROTEINA

Um dos campos da biologia de sistemas é a aplicação de redes para avaliar processos biológicos com uma visão holística. Exemplos de redes biológicas são as redes de regulação gênica, redes de transdução de sinal, redes de interação proteína-proteína (PPI) e redes metabólicas (Junker and Schreiber, 2008).

Nesse trabalho foi utilizado redes PPI para o estudo da SD22q11, nesse caso, os principais genes codificantes envolvidos foram representados pelos nós e as interações pelas ligações em redes PPI. Desse modo, um breve revisão dos conceitos envolvendo o estudo de redes biológicos é descrito a seguir.

4.1. Biologia de sistemas

É coerente cogitar que desde sempre se pensa em contexto de sistema, porém considera-se que a origem da área de estudo "Biologia de sistemas" ocorreu quando Ludwig Von Bertalanffy descreveu sua teoria de sistemas em 1969 (Junker and Schreiber, 2008). O interesse por essa área cresce cada vez mais, onde os sistemas mais comumente discutidos são redes gênicas ou proteicas, contudo não há escala fixa na qual a biologia dos sistemas opera (Hillmer, 2015). Diante desse cenário, qual é a definição de Biologia de sistemas?

Não há uma definição clara do termo Biologia de sistemas, isto é, diversos estudiosos descrevem a Biologia de Sistemas de uma forma diferente. Breitling (Breitling, 2010), por exemplo, define a biologia de sistemas como o esforço de pesquisa que fornece a base científica para o sucesso da biologia sintética. Diz ainda que é uma área que se baseia em estudos abrangentes da diversidade molecular dos sistemas vivos, naturais e sintéticos, e na integração do conhecimento biológico em modelos complexos que caracterizam a vida (Breitling, 2010). Já Kitano define a biologia de sistemas como um novo campo na biologia que visa a compreensão dos sistemas biológicos em nível de sistema (Kitano and Kitano, 2002).

Em geral, pode-se dizer que a Biologia de Sistema tem uma visão holística ao invés da visão reducionista. Ou seja, o objetivo da biologia de sistemas é entender sistemas biológicos por inteiro, elucidando, modelando e prevendo o comportamento de todos os componentes e interações. Assim, a biologia de sistema é um campo de estudo onde se pode estudar interações complexas que podem ser retratadas em forma de redes, por isso, podemos dizer que a biologia de sistema é baseada na teoria dos grafos (Junker and Schreiber, 2008).

4.2. Teoria dos grafos

Uma rede pode ser descrita como uma série de nós conectados uns aos outros por links ou arestas, onde cada link representa as interações entre dois componentes (Barabási and Oltvai, 2004a; Chan and Loscalzo, 2012). Desta forma, os nós e links formam uma rede ou, em linguagem matemática, um grafo (Barabási and Oltvai, 2004a; Pržulj, Wigle and Jurisica, 2004).

A teoria dos grafos começou com Leonard Euler e seu "problema da ponte de Königsberg" em 1736. O problema era: na cidade Königsberg (Prússia), um rio atravessa a cidade e sete pontes foram construídas sobre ele, Euler queria saber se era possível encontrar um caminho que passasse por toda a cidade atravessando cada ponte somente uma vez. Euler foi o primeiro a organizar seu "problema" em forma de um grafo e ao analisar a estrutura desse grafo, como mostra a Figura 11, ele provou que isso não é possível (Junker and Schreiber, 2008).

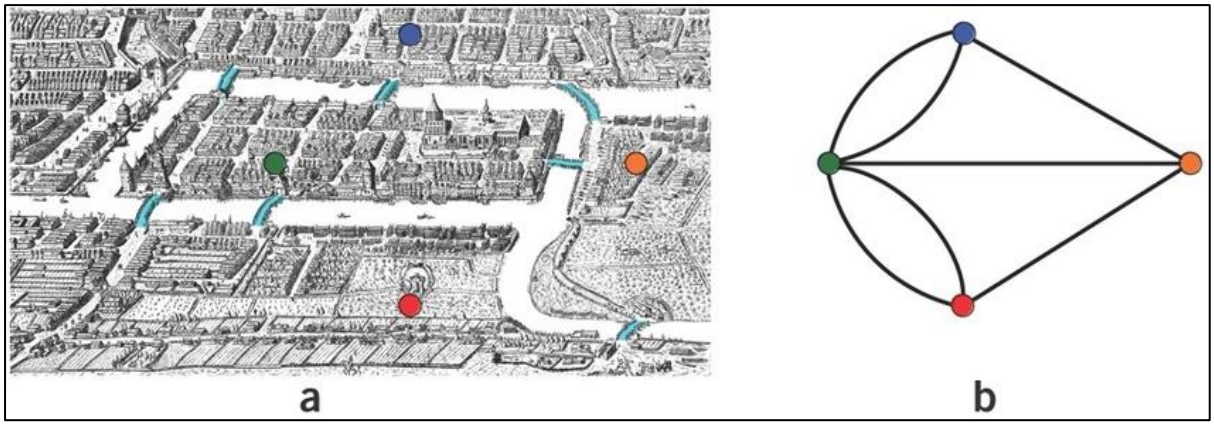


Figure 11. (a) Representação da cidade Königsberg onde se encontra as sete pontes e cada parte da cidade indicada por pontos com cores distintas. **(b)** O modelo de grafo correspondente a representação da cidade. Fonte: Chatterjee, 2015.

Um grafo pode ser representado matematicamente por $G = (V, E)$, onde V consiste em um conjunto de vértices (também chamados de nós ou pontos) e E um conjunto de arestas. Uma aresta e conectando dois vértices A e B pode ser representado por: $e = \{A,B\}$. A maneira mais comum de se visualizar um grafo é desenhar um ponto para cada vértice e uma linha para cada aresta que conecta os pontos correspondentes de seus vértices como demostrado na Figura 12 (Junker and Schreiber, 2008).

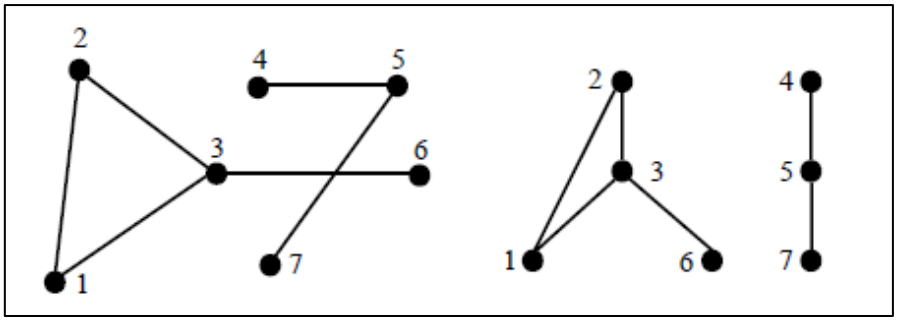


Figura 12. Representação de dois grafos *G = (V, E)* com um conjunto de vértices *V = {1,2,3,4,5,6,7}* e um conjunto de arestas *E = {{1,2},{2,3},{1,3},{3,6},{4,5},{5,7}}*. Fonte: Junker and Schreiber, 2008.

Além do grafo também podemos definir o subgrafo, ou seja, um subconjunto dos vértices e arestas de um grafo. O subgrafo pode ser representado por: *G'= (V',E')* do grafo *G = (V, E)* onde *V'* é um subconjunto de *V* e *E'* é um subconjunto de *E*. Assim, se o grafo *G'* é um subgrafo do grafo *G* e o conjunto de arestas *E'* contém todas as arestas de E que conectam vértices de *V'*, o subgrafo é chamado de subgrafo induzido de *G*. Na figura 13 pode-se observar um exemplo de subgrafo e subgrafo induzido (Junker and Schreiber, 2008).

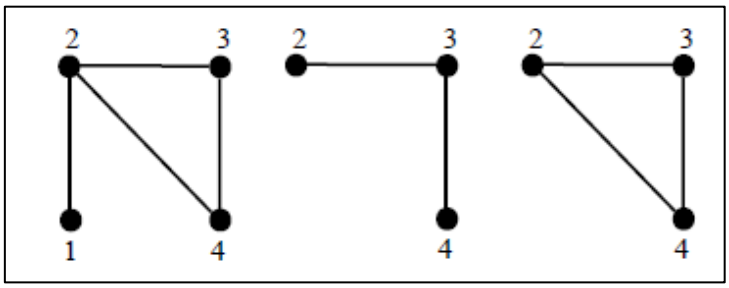


Figura 13. Da esquerda para direta: Grafo *G*, subgrafo *G'*, e o subgrafo induzido de *G*. Fonte: Junker and Schreiber, 2008.

Considerando a sequência *(V0, E1, V1, E2, V2, ... , Vk−1, Ek, Vk)* de vértices e arestas, define-se como *Path*, ou caminho, o percurso do vértice *V0* a *Vk* onde todas as arestas são distintas. Define-se como *Simple path* ou caminho simples, caso todos os vértices sejam distintos. E caracteriza-se como Ciclo quando o vértice inicial e final do grafo for o mesmo. Já o *Shortest path* traduzido como caminho mais curto é definido como o comprimento mínimo entre dois vértices, onde pode haver diferentes possibilidades de caminho mais curto entre dois vértices de um grafo. Por fim, o comprimento do caminho é dado pelo seu número de arestas (Junker and Schreiber, 2008).

De acordo com o tipo de interação entre os vértices, um grafo pode ser classificado como direto, indireto ou misto (Figura 14) (Junker and Schreiber, 2008):

- **Grafo não direcionado:** a aresta entre os vértices *u* e *v* é representada pelo par de vértices não ordenados *{u, v}*, é uma interação mútua onde o nó *u* interage com o nó *v* da mesma forma que *v* interage com *u*. Exemplos de grafo não direcionados são redes de interação de proteínas, redes filogenéticas e redes de correlação (Barabási and Oltvai, 2004b).

- **Grafo direcionado:** a aresta entre os vértices *u* e *v* é representada pelo par de vértices ordenados *(u, v)*. Geralmente visualiza-se a direção de uma aresta em um grafo através da direção de uma seta. Exemplos de redes biológicas modeladas por gráficos direcionados são redes metabólicas, redes de regulação de genes e redes alimentares (Junker and Schreiber, 2008).

- **Grafo misto:** há tanto interações direcionadas como não direcionadas. Um exemplo são as redes de proteínas onde algumas interações não são direcionadas (por exemplo, obtidas por experimentos de dois híbridos) e outras são direcionadas representando ativação, fosforilação e outras interações direcionadas (Junker and Schreiber, 2008).

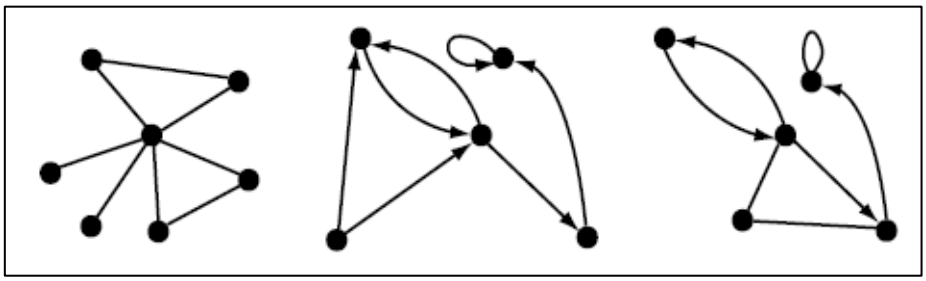


Figure 14. Da direita para a esquerda observa-se um grafo não direcionado, direcionado e um grafo misto. Fonte: Junker and Schreiber, 2008.

4.3. Propriedades gerais das redes

- **Distância:** comprimento do caminho mais curto entre dois vértices quaisquer. O caminho mais curto entre dois vértices não precisa ser único, muitas vezes existem vários caminhos alternativos com a mesma distância (Junker and Schreiber, 2008).

- **Diâmetro:** é o valor da maior distância entre dois vértices em um grafo (Kolaczyk and Csárdi, 2014). O diâmetro ou comprimento médio de um grafo é definido como a distância média entre todos os pares de vértices (Junker and Schreiber, 2008).

- ***Degree*****:** o *degree (k)* corresponde ao número de arestas de um vértice. O grafo direcionado possui o *in-degree*, que equivale ao número de links que chegam a esse nó, e o *out-degree* referente ao número de links que sai do nó (Barabási and Oltvai, 2004b; Sahinalp *et al.*, 2009; Raman, 2010). Por exemplo, no grafo não direcionado da figura 15a, o nó *A* possui um *degree k=5*, já no caso da figura 15b o nó *A* tem um $K_{in} = 4$ e $k_{out} = 1$ (Barabási and Oltvai, 2004a).

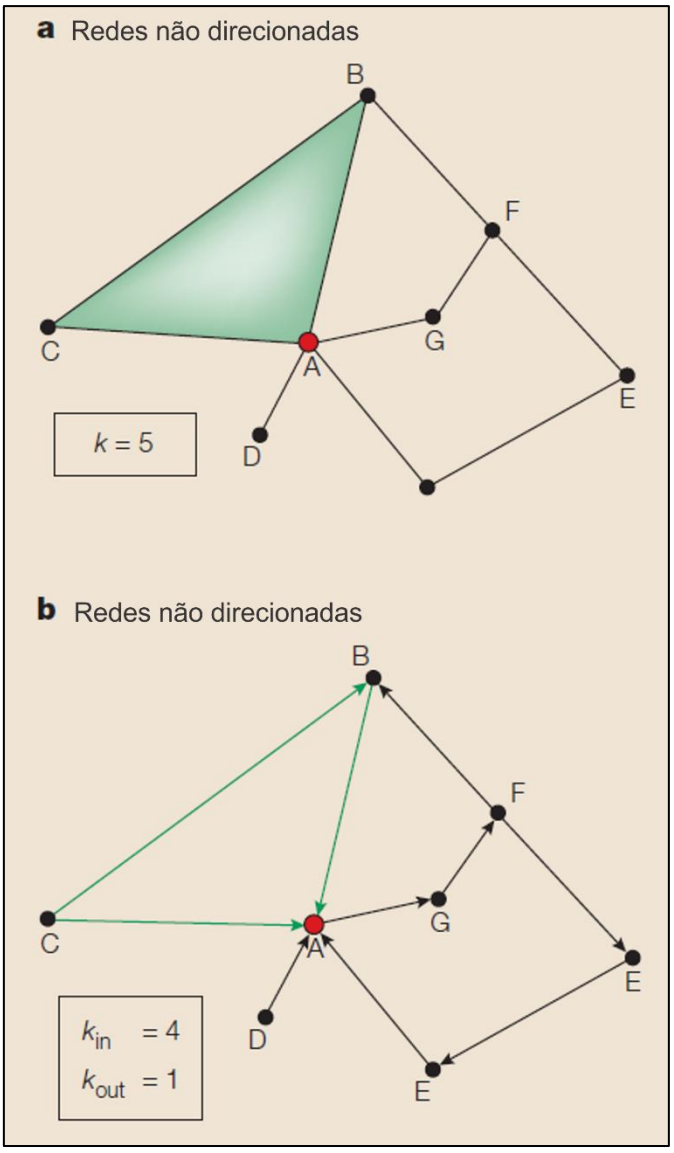


Figura 15. Exemplo de redes não direcionada e direcionada. **(a)** Rede não direcionada onde o nó *A* interage com outros cinco nós, ou seja o nó *A* possui um degree (*k*) igual a 5. **(b)** Rede direcionada, onde o nó *A* possui o *in-degree* (k_{in}) igual a 4 e o *out-degree* (k_{out}) igual a 1. Figura adaptada de Barabási and Oltvai, 2004[a].

- **Assortividade:** Medida que varia de -1 a 1, onde -1 é uma rede totalmente dissortativa e 1 representa uma rede totalmente assortativa. Uma rede dissortativa, por exemplo, apresenta vértices com alto *degree* conectando preferencialmente a vértices com baixo *degree*. O inverso também é verdadeiro, uma rede assortativa apresenta vértices com alto *degree* se conectando preferencialmente a outros vértices com alto *degree* (Junker and Schreiber, 2008).

- ***Betweenness:*** o *betweenness* é uma medida de centralidade de um vértice dentro de um grafo. Representa o número de caminhos mais curtos entre qualquer par de nós passando por um nó (Han, 2008; Sahinalp *et al.*, 2009). Nós com valores mais altos de *betweenness* estão posicionados em mais *shortest paths* em um grafo (Raman, 2010).

- **Módulos ou Comunidades:** são frequentemente definidos como um subconjunto de vértices que são densamente conectados entre si, mas são pouco conectados a outros vértices fora da comunidade. No contexto da análise de rede complexa, similaridade de vértices pode ser definida de diferentes maneiras, por exemplo, com relação ao caminho mais curto entre dois vértices, o número total de caminhos entre vértices, entre várias outras possibilidades (Junker and Schreiber, 2008).

4.4. Modelos de redes biológicas

Modelos de redes são utilizados para moldar nossa compreensão de redes complexas além de serem importantes para explicar as características das redes formadas. Aqui descrevemos o modelo de Erdõs-Rényi e o de Barabási-Albert já que esses são exemplos que influenciaram e auxiliaram o entendimento das redes biológicas (Junker and Schreiber, 2008).

- **Modelo de Erdõs-Rényi:** modelo aleatório onde cada par de nós está conectado com uma probabilidade igual (Figura16Aa) (Jeong *et al.*, 2000). O *degree* dos nós desse modelo segue uma distribuição de Poisson (Figura 16Ab), em que a maioria dos nós possuem um *degree* próximo ao valor do *degree* médio (<k>) (Barabási and Oltvai, 2004a). As principais limitações para uma comparação direta das propriedades desse modelo com as redes empíricas são sua distribuição homogênea do *degree*, a ausência de estrutura local e a falta de correlações dos *degrees* (Junker and Schreiber, 2008).

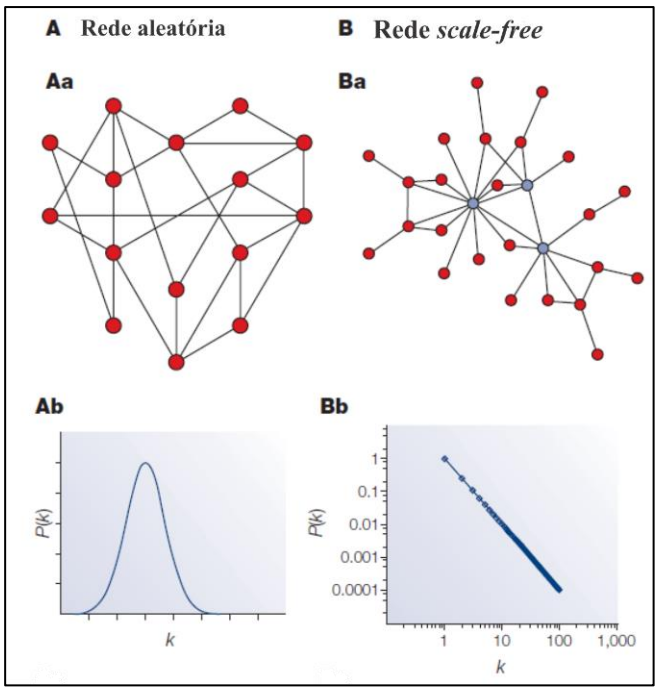


Figura 16. Exemplo de uma rede aleatória no lado esquerdo (Modelo de Erdõs-Rényi) e no lado direito uma rede *scale-free (*Modelo de Barabási-Albert). **(Aa)** Grafo aleatório onde pode ser observado que não há uma preferência de conexões entre os nós. **(Ab)** Distribuição de Poisson onde observar-se que a maioria dos nós possuem um valor médio de *degree* (*k*). **(Ba)** Exemplo de uma rede *scale-free* onde um menor número de nós, os *hubs*, são altamente conectados. **(Bb)** Distribuição do tipo lei de potência, demostrando que muitos nós possuem um baixo *degree* (*k*) enquanto poucos nós apresentam um alto *degree* (*k*). Figura adaptada de Barabási and Oltvai, 2004[a].

- **Modelo de Barabási-Albert (BA):** esse modelo assume a não aleatoriedade das conexões que deu origem a rede chamada *scale-free* (Figura B). O modelo de BA primeiro assume que vértices são adicionados, fazendo com que as redes cresçam em função do tempo. O modelo também assume o princípio de Conexão preferencial onde novas arestas não são introduzidas aleatoriamente. Os novos nós vão preferir se conectar com os nós da rede que possuem uma maior quantidade de conexões, ou seja, quanto maior o *degree* de um nó maior sua probabilidade de receber novas conexões (Junker and Schreiber, 2008). Assim, o *degree* do modelo BA segue uma distribuição do tipo lei de potência (*power-law*) (Figura 16Bb). Aqui, a probabilidade de um nó ser altamente conectado é estatisticamente mais significativa do que em um gráfico aleatório. Além disso, essa rede apresenta um pequeno número de nós altamente conectados, conhecidos como hubs (Figura 16Ba, nós azuis) (Barabási and Oltvai, 2004a).

OBJETIVOS

OBJETIVOS

1. OBJETIVO GERAL

Utilizar os métodos de aprendizado de máquinas e as redes de interação proteína-proteína (PPI) para investigação da relação genótipo-fenótipo observada na SD22q11.

2. OBJETIVOS ESPECÍFICOS

- Realizar um levantamento clínico e genético dos casos de SD22q11 encontrados na literatura para construção de um conjunto de dados.

- Classificar os casos clínicos em típico ou atípico, a partir dos dados obtidos, através de um algoritmo de aprendizado de máquina supervisionado.

- Listar os genes codificadores de proteínas afetados na deleção de 3Mb da SD22q11 e identificar os mesmos nas redes PPI obtidas de repositórios públicos.

- Verificar as alterações topológicas predominantes nas redes de cada grupo.

- Determinar os genes centrais das redes de cada grupo por meio de suas características topológicas.

REFERÊNCIAS

REFERÊNCIAS

Albertson, D. G. and Pinkel, D. (2003) 'Genomic microarrays in human genetic disease and cancer', *Human Molecular Genetics*, 12(suppl 2), pp. R145–R152. doi: 10.1093/hmg/ddg261.

Alpaydin, E. (2010) *Introduction to Machine Learning*. 2nd ed. The MIT Press. doi: 10.1016/j.neuroimage.2010.11.004.

Bales, A. M., Zaleski, C. A. and McPherson, E. W. (2010) 'Newborn screening programs: Should 22q11 deletion syndrome be added?', *Genetics in Medicine*, 12(3), pp. 135–144. doi: 10.1097/GIM.0b013e3181cdeb9a.

Barabási, A. L. and Oltvai, Z. N. (2004a) 'Network biology: Understanding the cell's functional organization', *Nature Reviews Genetics*, 5(2), pp. 101–113. doi: 10.1038/nrg1272.

Barabási, A. L. and Oltvai, Z. N. (2004b) 'Network biology: Understanding the cell's functional organization', *Nature Reviews Genetics*, 5(2), pp. 101–113. doi: 10.1038/nrg1272.

Bassett, A. (2011) 'Practical guidelines for managing patients with 22q11. 2 deletion syndrome', *J Pediatr.*, 17(2), pp. 281–294. doi: 10.1016/j.jpeds.2011.02.039.Practical.

Bassett, A. S. *et al.* (2011) 'Practical guidelines for managing patients with 22q11.2 deletion syndrome', *Journal of Pediatrics*. Mosby, Inc., 159(2), p. 332–339.e1. doi: 10.1016/j.jpeds.2011.02.039.

Bassett, A. S. *et al.* (2017) 'Rare genome-wide copy number variation and expression of schizophrenia in 22q11.2 deletion syndrome', *American Journal of Psychiatry*, 174(11), pp. 1054–1063. doi: 10.1176/appi.ajp.2017.16121417.

Ben-Shachar, S. *et al.* (2008) '22q11.2 Distal Deletion: A Recurrent Genomic Disorder Distinct from DiGeorge Syndrome and Velocardiofacial Syndrome', *American Journal of Human Genetics*, 82(1), pp. 214–221. doi: 10.1016/j.ajhg.2007.09.014.

Bengoa-Alonso, A. *et al.* (2016) 'Delineation of a recognizable phenotype for the recurrent LCR22-C to D/E atypical 22q11.2 deletion', *American Journal of Medical Genetics, Part A*, 170(6), pp. 1485–1494. doi: 10.1002/ajmg.a.37614.

Bittel, D. C. *et al.* (2009) 'Refining the 22q11.2 deletion breakpoints in DiGeorge syndrome by aCGH', *Cytogenetic and Genome Research*, 124(2), pp. 113–120. doi: 10.1159/000207515.

Bollag, R. J. *et al.* (1994) 'An ancient family of embryonically expressed mouse genes sharing a conserved protein motif with the T locus', *Nature Genetics*, 7(3), pp. 383–389. doi: 10.1038/ng0794-383.

Breitling, R. (2010) 'What is systems biology?', *Frontiers in Physiology*, 1 MAY(May), pp. 1–5. doi: 10.3389/fphys.2010.00009.

Breve, F. A. (2010) *Aprendizado de máquina em redes complexas*. Instituto de Ciências Matemáticas e de Computação - ICMC-USP.

Burnside, R. D. (2015) '22q11.21 deletion syndromes: A review of proximal, central, and distal deletions and their associated features', *Cytogenetic and Genome Research*, 146(2), pp. 89–99. doi: 10.1159/000438708.

Cardoso, A. R. *et al.* (2016) 'Major influence of repetitive elements on disease-associated copy number variants (CNVs)', *Human Genomics*. Human Genomics, 10(1), pp. 6–11. doi: 10.1186/s40246-016-0088-9.

Carlson, C. *et al.* (1997) 'Molecular Definition of 22q11 Deletions in 151 Velo-Cardio-Facial Syndrome Patients', *The American Journal of Human Genetics*, 61(3), pp. 620–629. doi: 10.1086/515508.

Chan, S. Y. and Loscalzo, J. (2012) 'The emerging paradigm of network medicine in the study of human disease', *Circulation Research*, 111(3), pp. 359–374. doi: 10.1161/CIRCRESAHA.111.258541.

Chatterjee, A. (2015) 'Studies on the Structure and Dynamics of Urban Bus Networks in Indian Cities', (December 2015). Available at: http://arxiv.org/abs/1512.05909.

Chen, J. *et al.* (2016) 'Identification of a Novel ENU-Induced Mutation in Mouse Tbx1 Linked to Human DiGeorge Syndrome', *Neural Plasticity*. Hindawi Publishing Corporation, 2016. doi: 10.1155/2016/5836143.

Cummings, C., McCauley, R. and Baylis, A. (2015) 'The Effect of Loudness Variation on Velopharyngeal Function in Children with 22q11.2 Deletion Syndrome: A Pilot Study', *Folia Phoniatrica et Logopaedica*, 67(2), pp. 76–82. doi: 10.1159/000438670.

Deo, R. C. (2015) 'Machine learning in medicine', *Circulation*, 132(20), pp. 1920–1930. doi: 10.1161/CIRCULATIONAHA.115.001593.

Digilio, M. C. *et al.* (2009) 'Three patients with oculo-auriculo-vertebral spectrum and microdeletion 22q11.2', *American Journal of Medical Genetics, Part A*, 149(12), pp. 2860–2864. doi: 10.1002/ajmg.a.33034.

Dugoff, L., Mennuti, M. T. and McDonald-McGinn, D. M. (2017) 'The benefits and limitations of cell-free DNA screening for 22q11.2 deletion syndrome', *Prenatal Diagnosis*, 37(1), pp. 53–60. doi: 10.1002/pd.4864.

Edelmann, L., Pandita, R. K. and Morrow, B. E. (1999) 'Low-Copy Repeats Mediate the Common 3-Mb Deletion in Patients with Velo-cardio-facial Syndrome', *The American Journal of Human Genetics*, 64(4), pp. 1076–1086. doi: 10.1086/302343.

Empke, S. L. L. (2015) '*Caracterização fenotípica em indivíduos com microarranjos na região cromossômica 22q11'*, Universidade de São Paulo.

Fukushiro A. P. (2007) '*Análise perceptiva, nasométrica e aerodinâmica da fala de indivíduos submetidos à cirurgia do retalho faríngeo para correção da insuficiência velofaríngea [tese]',* Bauru: Hospital de Reabilitação de Anomalias Craniofaciais, Universidade de São Paulo.

Gao, S. *et al.* (2015) 'TBX1 protein interactions and microRNA-96-5p regulation controls cell proliferation during craniofacial and dental development: Implications for 22q11.2 deletion syndrome', *Human Molecular Genetics*, 24(8), pp. 2330–2348. doi: 10.1093/hmg/ddu750.

Gao, S., Li, X. and Amendt, B. A. (2013) 'Understanding the role of Tbx1 as a candidate gene for 22q11.2 deletion syndrome', *Current Allergy and Asthma Reports*, 13(6), pp. 613–621. doi: 10.1007/s11882-013-0384-6.

Garavelli, L. *et al.* (2011) '22q11.2 distal deletion syndrome: Description of a new case with truncus arteriosus type 2 and review', *Molecular Syndromology*, 2(1), pp. 35–44. doi: 10.1159/000334262.

Giacomelli, M. *et al.* (2016) 'Reduction of CRKL expression in patients with partial DiGeorge syndrome is associated with impairment of T-cell functions', *Journal of Allergy and Clinical Immunology*. Elsevier Ltd, 138(1), p. 229–240.e3. doi: 10.1016/j.jaci.2015.10.051.

Goodship, J. *et al.* (1998) 'A population study of chromosome 22q11 deletions in infancy', *Archives of Disease in Childhood*, 79(4), pp. 348–351. doi: 10.1136/adc.79.4.348.

Guo, X. *et al.* (2011) 'Characterization of the past and current duplication activities in the human 22q11.2 region', *BMC Genomics*. BioMed Central Ltd, 12(1), p. 71. doi: 10.1186/1471-2164-12-71.

Habel, A. *et al.* (2014) 'Towards a safety net for management of 22q11.2 deletion syndrome: Guidelines for our times', *European Journal of Pediatrics*, 173(6), pp. 757–765. doi: 10.1007/s00431-013-2240-z.

Hacıhamdioğlu, B., Hacihamdioglu, D. O. and Delil, K. (2015) '22Q11 Deletion Syndrome: Current Perspective', *The Application of Clinical Genetics*, p. 123. doi: 10.2147/TACG.S82105.

Hall, M. *et al.* (2009) 'The WEKA data mining software', *SIGKDD Explorations Newsletter*, 11(1), p. 10. doi: 10.1145/1656274.1656278.

Han, J. D. J. (2008) 'Understanding biological functions through molecular networks', *Cell Research*, 18(2), pp. 224–237. doi: 10.1038/cr.2008.16.

Harel, T. and Lupski, J. R. (2018) 'Genomic disorders 20 years on—mechanisms for clinical manifestations', *Clinical Genetics*, 93(3), pp. 439–449. doi: 10.1111/cge.13146.

Hay, B. N. (2007) 'Deletion 22q11: Spectrum of Associated Disorders', *Seminars in Pediatric Neurology*, 14(3), pp. 136–139. doi: 10.1016/j.spen.2007.07.005.

Hillmer, R. A. (2015) 'Systems Biology for Biologists', *PLOS Pathogens*, 11(5), p. e1004786. doi: 10.1371/journal.ppat.1004786.

Hupé, P. *et al.* (2004) 'Analysis of array CGH data: From signal ratio to gain and loss of DNA regions', *Bioinformatics*, 20(18), pp. 3413–3422. doi: 10.1093/bioinformatics/bth418.

Jeong, H. *et al.* (2000) 'The large-scale organization of metabolic networks', *Nature*, 407(6804), pp. 651–654. doi: 10.1038/35036627.

Jerome, L. A. and Papaioannou, V. E. (2001) 'DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1', *Nature Genetics*, 27(3), pp. 286–291. doi: 10.1038/85845.

Ju, Z. R. *et al.* (2016) 'HIRA gene is lower expressed in the myocardium of patients with tetralogy of Fallot', *Chinese Medical Journal*, 129(20), pp. 2403–2408. doi: 10.4103/0366-6999.191745.

Junker, B. and Schreiber, F. (2008) *Analysis of biological networks*. Available at: http://books.google.com/books?hl=en&lr=&id=2DloLXaXSNgC&oi=fnd&pg=PR5&dq=Analysis+of+biological+networks&ots=2rpTdQ0a5M&sig=Jutk-5cXZZNgIdbX72yJxVuw5yc.

Kitano, H. and Kitano, H. (2002) 'Systems biology: A brief overview', *Science (New York, NY)*, 295(5560), pp. 1662–1664. Available at: papers3://publication/uuid/9C499668-2F87-4114-A79F-534B45ADD24F.

Kobrynski, L. J. and Sullivan, K. E. (2007) 'Velocardiofacial syndrome, DiGeorge syndrome: the chromosome 22q11.2 deletion syndromes', *Lancet*, 370(9596), pp. 1443–1452. doi: 10.1016/S0140-6736(07)61601-8.

Kolaczyk, E. D. and Csárdi, G. (2014) *Statistical Analysis of Network Data with R*, *International Statistical Review*. New York, NY: Springer New York (Use R!). doi: 10.1007/978-1-4939-0983-4.

Kuo, C. Y., Signer, R. and Saitta, S. C. (2018) 'Immune and Genetic Features of the Chromosome 22q11.2 Deletion (DiGeorge Syndrome)', *Current Allergy and Asthma Reports*. Current Allergy and Asthma Reports, 18(12), p. 75. doi: 10.1007/s11882-018-0823-5.

Landrum, M. J. *et al.* (2018) 'ClinVar: Improving access to variant interpretations and supporting evidence', *Nucleic Acids Research*. Oxford University Press, 46(D1), pp. D1062–D1067. doi: 10.1093/nar/gkx1153.

Lindsay, E. A. (2001) 'Chromosomal microdeletions: Dissecting DEL22Q11 syndrome', *Nature Reviews Genetics*, 2(11), pp. 858–868. doi: 10.1038/35098574.

Lupski, J. R. (1998) 'Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits', *Trends in Genetics*, 14(10), pp. 417–422. doi: 10.1016/S0168-9525(98)01555-8.

Maimon, O. and Rokach, L. (2015) 'Data mining with decision trees: theory and applications'. Ben-Gurion University of the Negev, Israel: World Scientific Publishing Co. Pte. Ltd.

McDonald-McGinn, D. M. *et al.* (2015) '22Q11.2 Deletion Syndrome', *Nature Reviews Disease Primers*, 1(November). doi: 10.1038/nrdp.2015.71.

McDONALD-McGINN, D. M. *et al.* (1997) 'The 22q11.2 Deletion: Screening, Diagnostic Workup, and Outcome of Results; Report on 181 Patients', *Genetic Testing*, 1(2), pp. 99–108. doi: 10.1089/gte.1997.1.99.

McDonald-McGinn, D. M., Zackai, E. H. and Low, D. (1997) 'What's in a name? The 22q11.2 deletion', *American Journal of Medical Genetics*, 72(2), pp. 247–247. doi: 10.1002/(SICI)1096-8628(19971017)72:2<247::AID-AJMG25>3.0.CO;2-M.

Michaelovsky, E. *et al.* (2012) 'Genotype-phenotype correlation in 22q11.2 deletion syndrome', *BMC Medical Genetics*. BMC Medical Genetics, 13(1), p. 1. doi: 10.1186/1471-2350-13-122.

Miller, D. T. *et al.* (2010) 'Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies', *American Journal of Human Genetics*. The American Society of Human Genetics, 86(5), pp. 749–764. doi: 10.1016/j.ajhg.2010.04.006.

Mitchell, T. M. (1997) *Machine Learning*, *Boston : WCB/*. McGraw-Hill.

Monteiro, F. P. *et al.* (2013) 'Defining new guidelines for screening the 22q11.2 deletion based on a clinical and dysmorphologic evaluation of 194 individuals and review of the literature', *European Journal of Pediatrics*, 172(7), pp. 927–945. doi: 10.1007/s00431-013-1964-0.

Morrow, B. E. *et al.* (2018) 'Molecular genetics of 22q11.2 deletion syndrome', *American Journal of Medical Genetics Part A*, 176(10), pp. 2070–2081. doi: 10.1002/ajmg.a.40504.

Nogueira, S. I. *et al.* (2008) 'Atypical 22q11.2 deletion in a patient with DGS/VCFS spectrum', *European Journal of Medical Genetics*, 51(3), pp. 226–230. doi: 10.1016/j.ejmg.2008.02.001.

Norkett, E. M. *et al.* (2018) 'Social cognitive impairment in 22q11 deletion syndrome : A review', *Psychiatry Research*. Elsevier Ireland Ltd, 253(January 2017), pp. 99–106. doi: 10.1016/j.psychres.2017.01.103.

Nowakowska, B. (2017) 'Clinical interpretation of copy number variants in the human genome', *Journal of Applied Genetics*. Journal of Applied Genetics, 58(4), pp. 449–457. doi: 10.1007/s13353-017-0407-4.

Quinlan, J.R. (1992) '*C4.5 Programs for Machine Learning'*, San Mateo, CA: Morgan Kaufmann.

Panamonta, V. *et al.* (2016) 'Birth Prevalence of Chromosome 22q11.2 Deletion Syndrome: A Systematic Review of Population-Based Studies.', *Journal of the Medical Association of Thailand = Chotmaihet thangphaet*, 99 Suppl 5(18), pp. S187-93. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29906080.

Picchi, G. F. A. (1997) 'Síndromes relacionadas a microdeleções: revisão da literatura'.

Pržulj, N., Wigle, D. A. and Jurisica, I. (2004) 'Functional topology in a network of protein interactions', *Bioinformatics*, 20(3), pp. 340–348. doi: 10.1093/bioinformatics/btg415.

Racedo, S. E. *et al.* (2015) 'Mouse and human CRKL is dosage sensitive for cardiac outflow tract formation', *American Journal of Human Genetics*. The American Society of Human Genetics, 96(2), pp. 235–244. doi: 10.1016/j.ajhg.2014.12.025.

Radoeva, P. *et al.* (2014) 'Association between autism spectrum disorder in individuals with velocardiofacial (22q11.2 deletion) syndrome and PRODH and COMT genotypes', *Psychiatric Genetics*, 24(6), pp. 269–272. doi: 10.1097/YPG.0000000000000062.

Raman, K. (2010) 'Construction and analysis of protein–protein interaction networks', *Automated Experimentation*, 2(1), p. 2. doi: 10.1186/1759-4499-2-2.

Reis, E. C. dos (2014) *Predição de fenótipos de Escherichia coli através de redes biológicas e aprendizado de máquina*. Universidade Estadual Paulista "Júlio de Mesquita Filho".

Robin, N. and Shprintzen, R. (2005) 'Defining the clinical spectrum of deletion 22q11. 2', *The Journal of pediatrics*, 147, pp. 90–96. doi: 10.1016/j.jpeds.2005.03.007.

Rosa, R. F. M. *et al.* (2009) 'Síndrome de deleção 22q11.2: compreendendo o CATCH22', *Revista Paulista de Pediatria*, 27(2), pp. 211–220. doi: 10.1590/S0103-05822009000200015.

Rosenfeld, J. A. *et al.* (2013) 'Estimates of penetrance for recurrent pathogenic copy-number variations', *Genetics in Medicine*, 15(6), pp. 478–481. doi: 10.1038/gim.2012.164.

Rump, P. *et al.* (2014) 'Central 22q11.2 deletions', *American Journal of Medical Genetics, Part A*, 164(11), pp. 2707–2723. doi: 10.1002/ajmg.a.36711.

Sahinalp, S. C. *et al.* (2009) 'The Effect of Insertions and Deletions on Wirings in Protein-Protein Interaction Networks: A Large-Scale Study', *Journal of Computational Biology*, 16(2), pp. 159–167. doi: 10.1089/cmb.2008.03tt.

Salzberg, S. L. (1994) 'C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993', *Kluwer Academic Publishers*, 16(3), p. pp 235–240. doi: https://doi.org/10.1007/BF00993309.

Samuels, I. S. *et al.* (2008) 'Deletion of ERK2 Mitogen-Activated Protein Kinase Identifies Its Key Roles in Cortical Neurogenesis and Cognitive Function', *Journal of Neuroscience*, 28(27), pp. 6983–6995. doi: 10.1523/JNEUROSCI.0679-08.2008.

Sandrin-Garcia, P. *et al.* (2007) 'Typical phenotypic spectrum of velocardiofacial syndrome occurs independently of deletion size in chromosome 22q11.2', *Molecular and Cellular Biochemistry*, 303(1–2), pp. 9–17. doi: 10.1007/s11010-007-9450-5.

Scambler, P. J. (2000) 'The 22q11 deletion syndromes', *Human Molecular Genetics*, 9(16), pp. 2421–2426. doi: 10.1093/hmg/9.16.2421.

Shaffer, L. G. and Lupski, J. R. (2000) 'Chromosomal Rearrangements in H Umans', *Annual review of genetics*, 34, pp. 297–329.

Shaikh, T. H., Kurahashi, H. and Emanuel, B. S. (2001) 'Evolutionarily conserved low copy repeats (LCRs) in 22q11 mediate deletions, duplications, translocations, and genomic instability: An update and literature review', *Genetics in Medicine*, 3(1), pp. 6–13. doi: 10.1097/00125817-200101000-00003.

Shprintzen, R. J. (2008) 'Velo-Cardio-Facial Syndrome: 30 Years of Study', *Developmental Disabilies Research Reviews*, 14(1), pp. 3–10. doi: 10.1002/ddrr.2.Velo-Cardio-Facial.

Souto, M. C. P. *et al.* (2003) 'Técnicas de aprendizado de máquina para problemas de biologia molecular', *Sociedade Brasileira de Computaçao*, (October). Available at: http://www.cin.ufpe.br/~mcps/ENIA2003/jaia2003-14-08.pdf.

Sullivan, K. E. (2019) 'Chromosome 22q11.2 deletion syndrome and DiGeorge syndrome', *Immunological Reviews*, 287(1), pp. 186–201. doi: 10.1111/imr.12701.

Swillen, A. *et al.* (2000) 'Chromosome 22q11 deletion syndrome: Update and review of the clinical features, cognitive-behavioral spectrum, and psychiatric complications.', *American Journal of Medical Genetics*, 97, pp. 128–135. doi: 10.1002/1096-8628(200022)97:2<128::AID-AJMG4>3.0.CO;2-Z.

Witten, I. H. , Frank, E., & Hall, M. A. (2011) *Data Mining : Practical Machine Learning Tools and Techniques*, *Morgan Kaufmann Publishers*. doi: 10.1016/C2009-0-19715-5.

Yamagishi, H. and Srivastava, D. (2003) 'Unraveling the genetic and developmental mysteries of 22q11 deletion syndrome', *Trends in Molecular Medicine*, 9(9), pp. 383–389. doi: 10.1016/S1471-4914(03)00141-2.

Yang, J.-H. *et al.* (2016) 'Differential regulation of the histone chaperone HIRA during muscle cell differentiation by a phosphorylation switch', *Experimental & Molecular Medicine*. Nature Publishing Group, 48(8), pp. e252–e252. doi: 10.1038/emm.2016.68.

Zarrei, M. *et al.* (2015) 'A copy number variation map of the human genome', *Nature Reviews Genetics*. Nature Publishing Group, 16(3), pp. 172–183. doi: 10.1038/nrg3871.

Zeitz, M. J. *et al.* (2013) 'Implications of COMT long-range interactions on the phenotypic variability of 22q11. 2 deletion syndrome', *Nucleus*, 4(6), pp. 6–7. doi: 10.4161/nucl.27364.

Conforme estabelecido pelo Programa de Pós-Graduação em Ciências Biológicas(Genética) do IBB, UNESP, os resultados obtidos durante a execução deste projeto de mestrado foram reunidos em dois artigos científicos para publicação.

CAPÍTULO 1 - MACHINE LEARNING-BASED METHODOLOGY FOR THE CLINICAL STRATIFICATION OF 22Q11 DELETION SYNDROME.

CAPÍTULO 2 - A SYSTEM BIOLOGY APPROACH IN THE STUDY OF THE 22Q11 DELETION SYNDROME.

CAPÍTULO 1

CAPÍTULO 1

MACHINE LEARNING-BASED METHODOLOGY FOR THE CLINICAL STRATIFICATION OF 22q11 DELETION SYNDROME

Camila C. Alves[a*], Bruno F. Gamba[b], Ivan R. Wolf[c], Lucilene Ribeiro-Bicudo[b], Guilherme T. Valente[c]

[a] São Paulo State University (UNESP), Botucatu, São Paulo, Brazil
[b] Biological science institute, Federal University of Goiás (UFG), Goiânia, Goiás, Brazil.
[c] School of Agronomic Sciences, São Paulo State University (UNESP), Botucatu, São Paulo, Brazil
*cris_camila@yahoo.com.br

Abstract

The 22q11 Deletion Syndrome (22q11DS) involves deletion of approximately 0.7 to 3 Mb and results in a broad phenotypic spectrum. In addition is considered as the most common microdeletion syndrome, with a prevalence of one case per 4.000 to 9.800 live births. Clinical features that can differentiate patients with typical or atypical deletion of the 22q11 region could be interesting as physicians could prescribe specific genetic tests for faster diagnosis of 22q11DS cases. The present work uses a machine learning-based methodology to classify cases diagnosed with 22q11DS in typical or atypical according to their clinical features. A bibliographic survey was performed to obtain the clinical and genetic data of cases diagnosed with 22q11DS for the construction of a dataset. The decision trees were made with the classification algorithm known as J48 in WEKA platform. As a result, we had a dataset of 43 clinical features as attributes, 95 cases, 46 were classified as typical, and 49 are atypical. Four trees were selected, which had highly accurate at 83-91% and CCI at 83-91%. Attributes used as nodes of the tree-like oral cleft, velopharyngeal insufficiency, delayed speech, and language development, specific learning disability, behavioral abnormality and growth delay. Cardiac defects sign was not used as a node in any of the selected classification trees, demonstrating that this clinical sign does not significantly assist in the classification of typical and atypical deletions. In conclusion, we can say that the machine-learning method accomplishes the goal proposed in this work. However, we do not aim to create a classificatory model; the machine learning-based methods were used here in order to aid in the interpretability of the results.

Keywords: 22q11DS, DiGeorge Syndrome, Machine learning-based methodology, algorithm J48.

1. Introduction

The 22q11 Deletion Syndrome (22q11DS) (Online Mendelian Inheritance in Man - OMIM #192430) is commonly known as DiGeorge Syndrome (OMIM #188400), Velocardiofacial Syndrome (OMIM #192430) and Conotruncal Anomaly Face Syndrome. It involves the deletion of approximately 0.7-3 Mb and is considered the most common microdeletion syndrome, with a prevalence of one case per 4.000 to 9.800 live births (Burnside, 2015; McDonald-McGinn *et al.*, 2015; Panamonta *et al.*, 2016; Dugoff, Mennuti and McDonald-McGinn, 2017).

22q11DS is characterized by a vast phenotypic spectrum with pleiotropic effects involving variables organs and/or systems with more than 180 clinical signs already described, both physical and behavioral (Carlson *et al.*, 1997; Robin and Shprintzen, 2005; Hay, 2007). However, there is no clinical sign present in all individuals with the 22q11 deletion, indicating the absence of mandatory manifestations for the syndrome (Hay, 2007; Shprintzen, 2008).

The pericentromeric region of chromosome 22 harbor eight distinct low copy repeats (LCRs) with high homology to each other that can lead to non-allelic homologous recombination (NAHR) resulting in a deletion within the 22q11 region (McDonald-McGinn *et al.*, 2015). According to their position related to the centromere, deletions involving the LCRs22 regions can be designated as: proximal deletions (A-B, A-D, A-E, A-F), central deletions (B-D, C-D) and distal deletion type I (C-E, D-E, D-F), type II (E-F) and type III (D-H, E-H, F-H) (Burnside, 2015). Around 90% of the 22q11DS cases have a 3Mb deletion, which encompasses 45 know protein-coding genes, seven microRNAs and ten non-coding RNAs, and that is considered the Typically Deletion Region (TDR) (Yamagishi and Srivastava, 2003; McDonald-McGinn *et al.*, 2015; Morrow *et al.*, 2018). Therefore, 8% of cases have a 1.5Mb deletion which encompasses 24 genes, and a minority has atypical deletions of the 22q11 region, involving different LCRs22, overlapping and not overlapping (Lindsay, 2001; Burnside, 2015).

Despite the diversity of clinical features associated with 22q11DS, some main signs such as congenital heart disease, cleft palate, velopharyngeal insufficiency, craniofacial dysmorphism, thymic aplasia or hypoplasia, learning disability and immune deficiency are found (Dugoff, Mennuti and McDonald-McGinn, 2017). Additional findings include hypocalcemia, eating disorders, renal abnormalities, hearing loss, growth hormone deficiency, autoimmune disorders, seizures, skeletal abnormalities and psychiatric disorders (McDonald-McGinn *et al.*, 2015). Consequently, clinical features that can differentiate patients with typical or atypical deletion of the 22q11 region could be interesting as physicians could prescribe

specific genetic tests and more patients could be diagnosed. To achieve the goal of understanding the association of 22q11 deletion with the disease, the present work uses a machine learning-based approach (ML) to classify diagnosed cases of 22q11DS in typical or atypical.

As far as we know, ML methodology was not used to study typical and atypical cases of 22q11DS. Here we present the use of a machine learning-based methodology to classify cases diagnosed with 22q11DS in typical or atypical according to their clinical features using the J48 algorithm (a decision tree algorithm). This way, the algorithm found a patter for each condition and the results of the generated decision trees may lead to reflection on current clinical practice in the context of SD22q11.

2. Methods

2.1. Data collection

A bibliographic survey was performed to obtain the clinical and genetic data of cases diagnosed with 22q11DS for the construction of a dataset in the form of a matrix. Firstly, 45 cases were obtained through the work of Empke (Empke, 2015), in which cases were confirmed by MLPA, and all of then presented the typical deletion of 3Mb (Supplementary Table 1).

Mining PubMed was done to obtain reports about diagnosed 22q11DS cases, especially atypical deletions. Each article was evaluated in order to verify whether clinical cases had a detailed report of clinical signs, generating 50 cases (49 atypical and 1 typical). The cases were searched up in order to obtain a similar number in both typical and atypical cases, generating a total of 95 cases, hereafter referred to as instances (Supplementary Table 1).

2.2. Data preparation

The name of all clinical features was standardized according to The Human Phenotype Ontology (HPO) (Köhler *et al.*, 2017), and clinical and genetic data were here used as attributes (it means features). A total of 90 clinical features were obtained from all cases, and each feature was assigned as 1 or 0, indicating the presence or absence of each clinical sign, respectively. For the supervised learning purpose, each instance was defined into one of two classes (the last column of the matrix) such as: the instance that possessed as attributes a deletion of 3Mb involving the LCRs22 of A-D were classified as “typical” while the “atypical” class encompasses deletions of different sizes involving different LCRs22.

After, the most relevant attributes were selected through the *InfoGainAttributeEval* attribute evaluator available in the WEKA (Waikato Environment for Knowledge Analysis) (Hall *et al.*, 2009; Witten, I. H. , Frank, E., & Hall, 2011). This evaluator considers the value of an attribute by measuring the information gain in comparison to the assigned class and performs attributes ranking. After sorting all the ranked attributes, they were split into one of two clusters using the SimpleKMeans algorithm implemented in WEKA (Hall *et al.*, 2009; Witten, I. H. , Frank, E., & Hall, 2011). This way, the attributes encompass the cluster with greater InfoGain were selected in all instances for further analysis.

2.3. Decision-tree Modeling

The decision trees were performed using the J48 algorithm, the WEKA's implementation of the C4.5 algorithm (Hall *et al.*, 2009; Witten, I. H. , Frank, E., & Hall, 2011). The 10-fold cross-validation was used to evaluate the model, and the values of precision, accuracy and the receiver operating characteristics were used to evaluate the learning performance. Figure 1 resumes the methodology used in this work. Thus, the first decision tree was performed under unpruned mode and default parameters, and all selected attributes were modeled used as tree's nodes. The supervised learning step was redo under manual pruning in order to check whether other secondary attributes maintain the classification accuracy, to aid results interpretability, and to avoid overfitting (Quinlan, 1986). Thus, the root node (it reflects the most critical attribute on data) identified in the first learning was removed in the data, followed by new supervised learning using the same parameters aforementioned; the pruning process was redone until no attribute was available anymore. After the construction of all the trees, the percentage of correctly classified instances (CCI) was used to select the trees with the best classification performances.

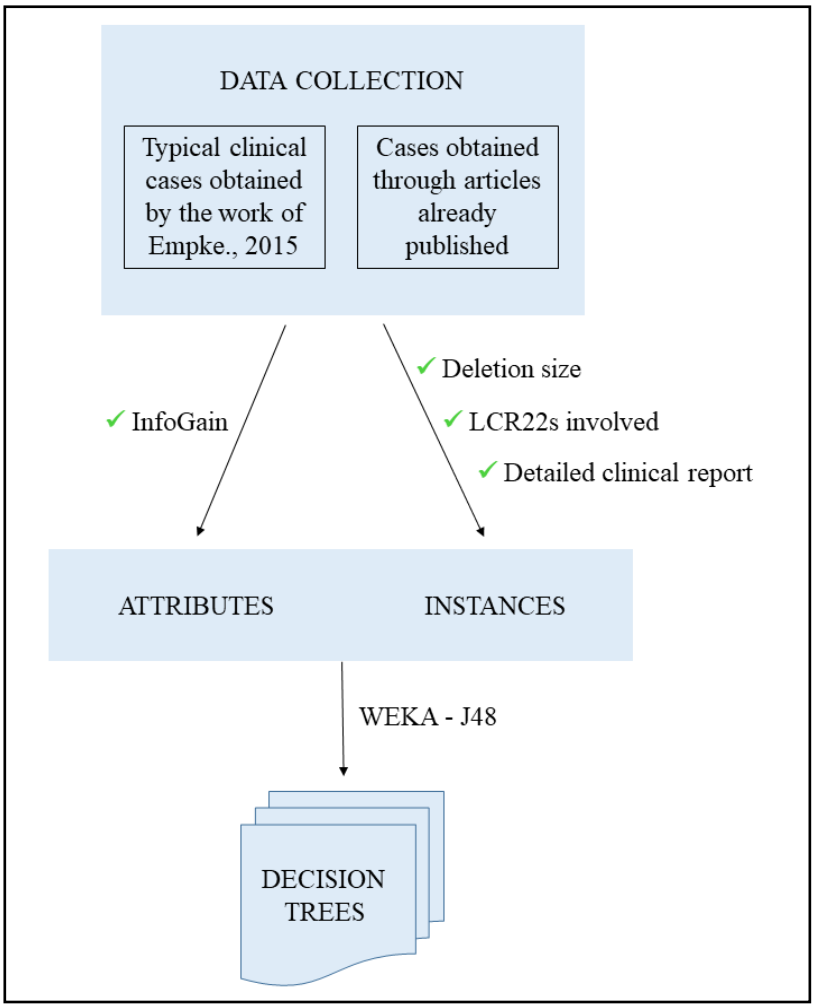


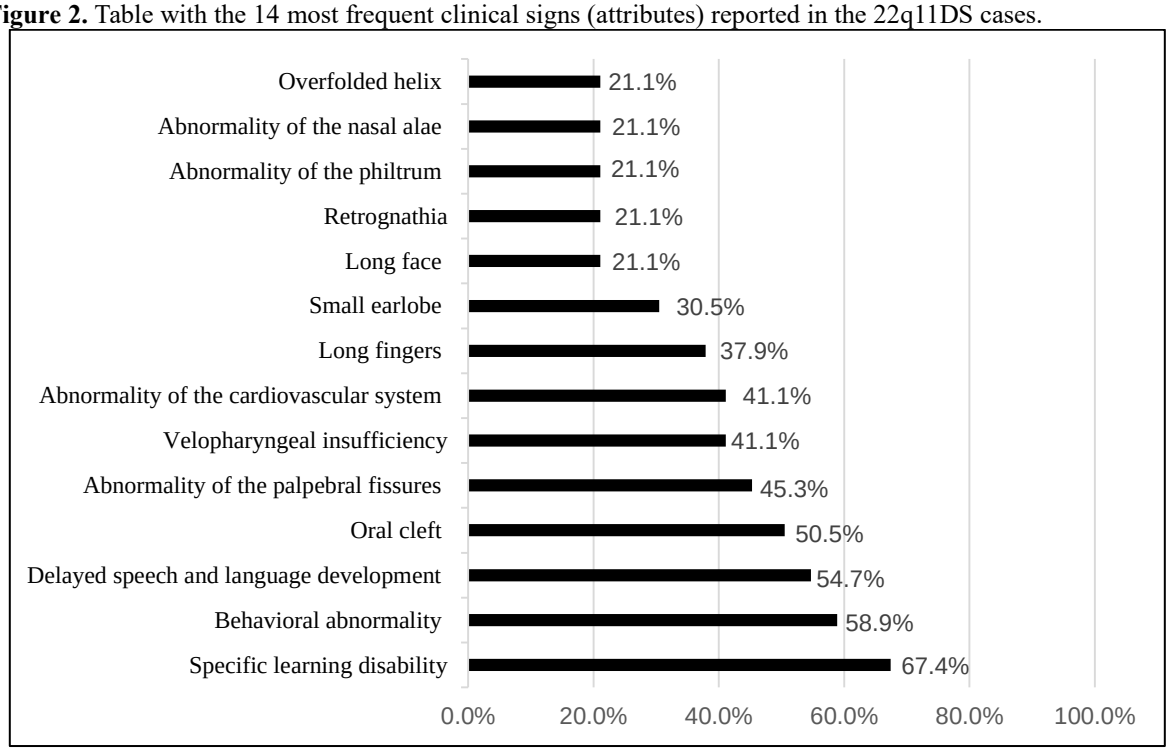
Figure 1. Procedure for data collection and preparation.

3. Results

3.1. Data collation and preparation

The final dataset had 95 clinical patients with diagnosis confirmed by MLPA or CGH. Of these cases, 46 and 49 were classified as typical and atypical, respectively. A total of 50.5% of the subjects were male, and 49.5% were female. In total, 90 clinical features were organized as attributes and the most frequent clinical features can be observed in figure 2. Supplementary table 2 shows the frequency of all clinical features and the percentage of typical and atypical cases for each clinical characteristics found.

Figure 2. Table with the 14 most frequent clinical signs (attributes) reported in the 22q11DS cases.



After assessing the attributes with the InfoGainAttributeEval and divide them into two clusters it was observed that the attributes with InfoGain greater than 0.0223753 (the red dots in figure 3) were consisted with the attribuss from cluster 2. Thus, the 43 attributes from cluster 2 were selected since these attributes, according to their InfoGain, have a higher classification power. Therefore, the final dataset, used in supervised learning, had a total of 95 instances and 43 attributes.

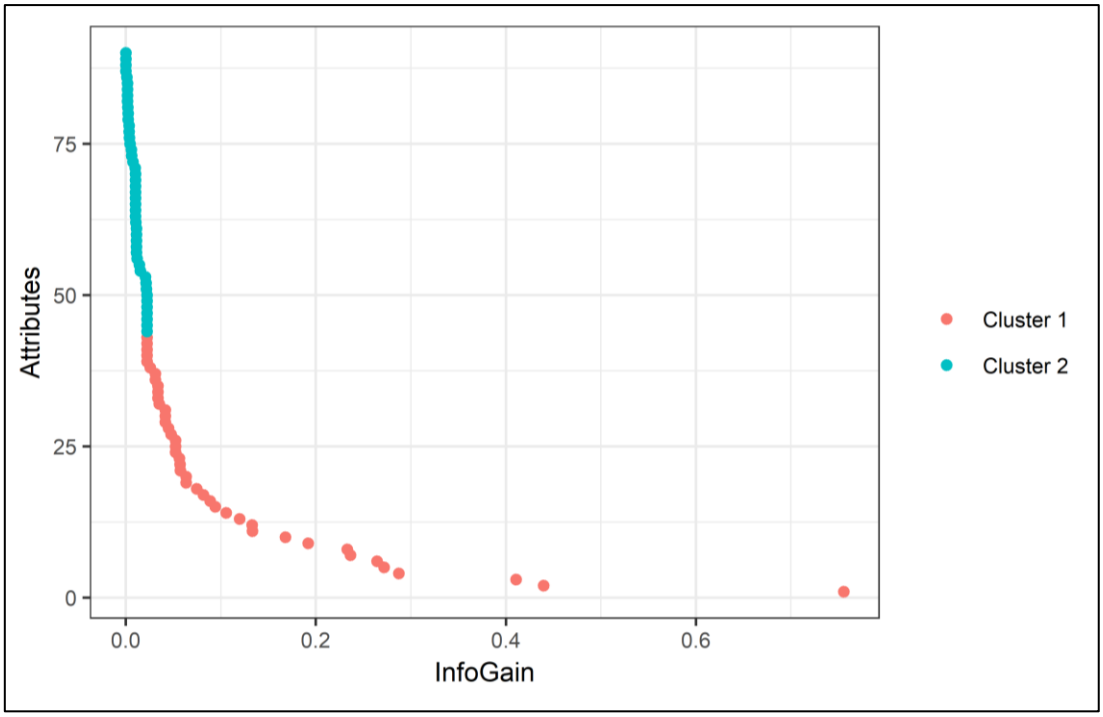


Figure 3. Distributions of attributes according to their InfoGain, where the x-axis represents the Information Gain while the y-axis represents the attributes. The attributes were divided into two clusters: cluster 1 (in pink) comprises the attributes with greater InfoGain and cluster 2 (in blue) comprises the attributes with smallest InfoGain. Attributes from cluster 1 were the selected.

3.2. Decision tree model

The first decision tree (hereafter referred to as T1, the unpruned tree) was made with a dataset that consisted of 95 instances (22q11DS cases) and 43 attributes (the clinical features with InfoGain higher than 0.0223753). The tree's nodes were the oral cleft, specific learning disability, and nasal base attributes, with 91% of cases being correctly classified with ROC area of 0.902.

The iterative manual pruning generated a total of 43 decision trees, but only the trees T1-T3 and T5 had feasible predictive performances (CCI >80%) (Figure 4). Concerning the most important attributes of the best trees, the decision tree T2 (CCI of 85%) ranked the velopharyngeal insufficiency as the most important attribute, the T3 (90% of CCI) rooted in the delayed speech and language development and the T5 (CCI of 83%) had as root abnormality of the shape of the midface (Figure 5).

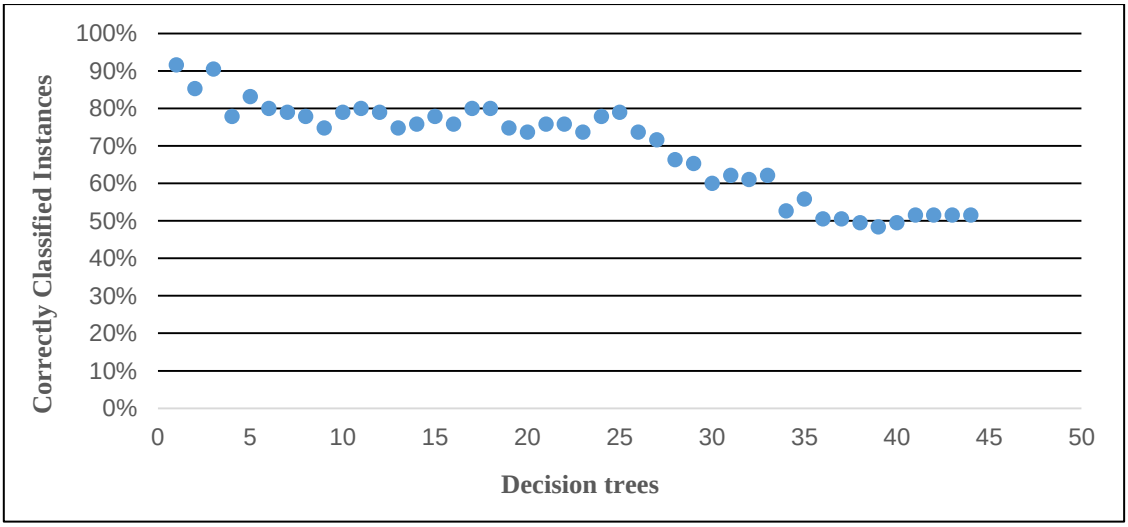


Figure 4. Distribution of CCI from the pruning process. T1-T43 = generated trees.

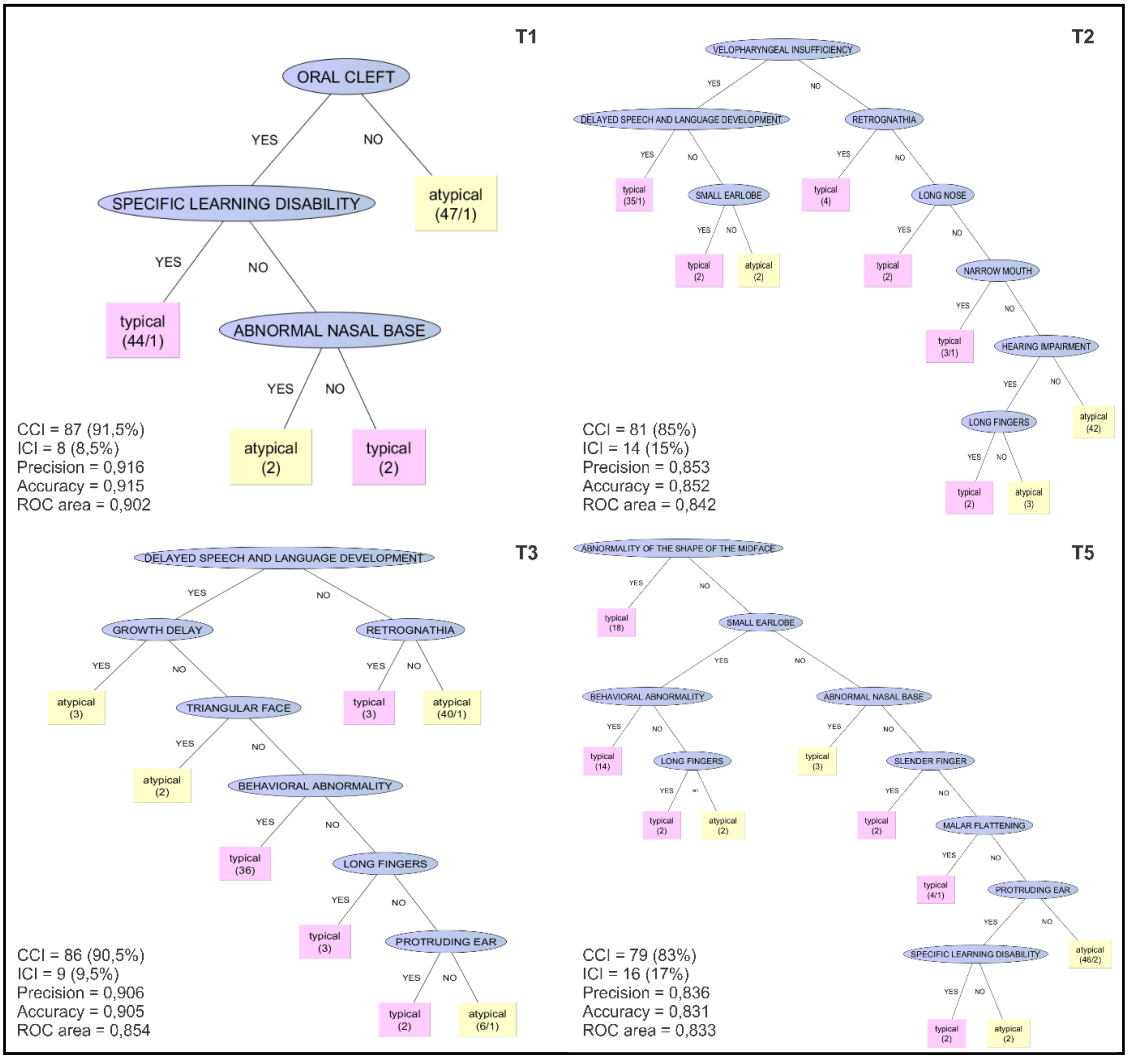


Figure 5. Decision trees T1, T2, T3 and T5 and each evaluated parameters (CCI, precision, accuracy, ROC area, and the confusion matrix). The first number of the parenthesis is equal to the weight of instances that reached the leaf; the right number is the weight of those instances that are misclassified. CCI = correctly classified instances. ICI = Incorrectly Classified Instances.

In table 2, it is possible to view all clinical features used in at least one of the selected trees while the frequency of all attributes can be seen in supplementary table 1. One clinical sign that is usually frequent and considered an important feature of the syndrome is the congenital heart disease (here called abnormality of the cardiovascular system). However, this characteristic was not included in any selected classification tree.

Table 1. Attributes that are part of the selected classification trees, where it can be observed the frequency of each attribute considering the total cases, in typical cases and atypical cases.

Clinical feature	Total of cases (95)	Typical cases (46)	Atypical cases (49)
Specific learning disability	67.40%	69%	31%
Behavioral abnormality	58.90%	71%	29%
Delayed speech and language development	54.70%	79%	21%
Oral Cleft	50.50%	94%	6%
Velopharyngeal insufficiency	41.10%	92%	8%
Long fingers	37.90%	81%	19%
Small earlobe	30.50%	93%	7%
Retrognathia	21.10%	100%	0%
Abnormality of the shape of the midface	18,9%	100%	0%
Protuding ear	15,8%	80%	20%
Hearing impairment	14,7%	79%	21%
Growth delay	12.6%	0%	100%
Long nose	11,6%	100%	0%
Abnormal nasal base	10,5%	100%	0%
Narrow mouth	8,4%	88%	13%
Slender finger	8,4%	100%	0%
Malar flattening	7,4%	86%	14%
Triangular face	7,4%	79%	21%

4. Discussion

With the application of the machine learning technique, it is possible to find new associations not yet discovered by clinical observations or univariate analyzes techniques. In addition, the decision tree structure gives us an insight into which attributes are most significant for classification (Dugan *et al.*, 2015). Here, decision trees were constructed with the purpose

of classifying the cases of 22q11DS in typical or atypical, the objective of interpreting the use of clinical signs to differentiate these cases.

Some characteristics used in the chosen trees such as an oral cleft, velopharyngeal insufficiency, delayed speech and language development, specific learning disability, behavioral abnormality and growth delay are amongst the most common and cited clinical signs for 22q11DS. The frequency found of this features is equivalent to the literature (Hay, 2007; Kobrynski and Sullivan, 2007; Turan *et al.*, 2008; Maggadottir and Sullivan, 2013; Kruszka *et al.*, 2017).

The oral cleft feature was used as the root node of the T1 tree and its prevalence, according to this study, is 94% in patients with typical deletion. Here, we generalized the types of oral cleft in one attribute, but about 11% of the pediatric patients present an evident cleft palate, and around 65% of patients present milder manifestations such as a submucosal cleft, bifid uvula and velopharyngeal dysfunction (McDonald-McGinn *et al.*, 2015). Among the palatine anomalies, submucosal cleft palate is one of the most common forms reported in 22q11DS (Richieri-costa, 2008; Shprintzen, 2008), but there are no reports indicating the frequency of oral cleft in patients with different deletions in the 22q11 region.

Velopharyngeal insufficiency was the attribute used as the root node of the T2 tree, which may indicate that this clinical sign is relevant during the clinical evaluation of the patient. This feature was considered one of the most evident clinical features in SD22q11, with a prevalence varying from 27% to 92% (Hay, 2007; Kobrynski and Sullivan, 2007; Maggadottir and Sullivan, 2013; McDonald-McGinn *et al.*, 2015). The etiology in patients with 22q11DS is related to structural abnormalities, such as palatine abnormalities (Spruijt *et al.*, 2012), but there are no studies evaluating the incidence in patients with typical and atypical deletion.

The delayed speech and language development delay were used as the root node of the decision tree T3. However, is important to notice that this clinical sign could be a consequence of others conditions like some palatine dysfunction, such as the cleft palate. Thus syndromic individuals may manifest language delay and phonation difficulties (HAY, 2007). Significant language delay occurs in about 70 to 84% of individuals with 22q11DS, but as in previous cases there are no studies that indicate a difference of this signal in typical and atypical cases (Hay, 2007; Kobrynski and Sullivan, 2007; Maggadottir and Sullivan, 2013; McDonald-McGinn *et al.*, 2015).

In addition to language difficulties, learning difficulties are often reported in patients with 22q11DS where studies indicate an incidence of 70-90% of cases (Hay, 2007; Kobrynski and Sullivan, 2007; Maggadottir and Sullivan, 2013; McDonald-McGinn *et al.*, 2015). This

clinical sign was also used in the T1 tree, and when the patients had the oral cleft signal together with a specific learning disability, the case was classified as typical. This pattern found shows that these two clinical signs are usually found together which may mean that there is a connection between them.

The behavioral abnormality clinical sign is the second most frequent among the attributes used as a node of the selected trees, but it was only used as one of the nodes of the tree T5. This feature has a frequency of 71% in patients with typical deletion and 29% in atypical deletions. Different types of behavioral abnormality have already been described for 22q11DS among them we can cite autism spectrum disorder, attention-deficit / hyperactivity disorder (ADHD), and schizophrenia (Tang *et al.*, 2015; Clements *et al.*, 2017). Interestingly, an increased rate of autism spectrum disorder was observed among individuals with a deletion in region LCRA-B, compared to individuals whose nested deletions did not involve that region (Clements *et al.*, 2017).

Growth retardation was used as the node of the T3 tree, and when present with the attribute delayed speech and language development three cases were classified as atypical. This clinical signal is considered a significant problem, especially in infancy and in the preschool years of the individuals (Bassett, 2011; Habel *et al.*, 2012). In addition to the direct effect of deletion on patient development, other problems associated with the syndrome may cause or potentiate growth retardation are cardiac defects requiring surgery, immunodeficiency with recurrent infections, hypoparathyroidism, hypothyroidism, growth hormone insufficiency (Bassett, 2011; Habel *et al.*, 2012). Among the literature, was reported 27 patients who had a central deletion 22q11.2 involving LCR22-B or LCR22-C, and LCR-D and was observed that signs of growth restriction, low stature, and microcephaly were more frequent in patients with CD deletion (RUMP e colab., 2014).

However, the prevalence of some characteristics observed on syndromic patient as long fingers, retrognathia, abnormality of the shape of the midface, narrow mouth, slender finger and malar flattening is not available in the literature. The features protruding ear (15,8%) and small earlobe (30.50%) were dysmorphologies chosen as tree nodes and are features found especially in cases with typical deletion. The frequency of this specifically characteristics could not be found in literature; however a global prevalence of 59% for ear anomalies was reported (KRUSZKA., 2017). The presence of these clinical features as tree nodes can demonstrate that this dysmorphism may be necessary for the diagnosis of the syndrome, or to differentiate typical and atypical cases.

When the typical and atypical cases are compared, it is possible to notice that the percentage of every clinical feature used as node was more significant in typical cases. The frequency of the features for each case seems to reflect on the classification tree since most cases classified as atypical do not have the presence of a specific clinical feature.

In T1, 46 cases were classified as atypical when they did not present an oral cleft. In T2, 42 cases were classified as atypical, and all of them did not present velopharyngeal insufficiency, retrognathia, long nose, narrow mouth or hearing impairment. The same pattern can be observed in tree T3, and T5 as most of the atypical cases were classified when they did not present one or more clinical feature used as a node of the tree.

Cardiac defects present a relatively high frequency among the clinical findings of 22q11DS, with approximately 59 to 83% of the cases (Hay, 2007; Kobrynski and Sullivan, 2007; Maggadottir and Sullivan, 2013; McDonald-McGinn *et al.*, 2015). The suspicion or diagnosis of the syndrome usually results in the subsequent identification of a cardiac defect and, although 20% of the patients do not present with the cardiac anomaly, this condition remains the most common reason for patient medical referral (Maggadottir and Sullivan, 2013). However, this clinical sign was not used, as a node in any of the selected classification trees, since the frequency of this attribute among the cases used in this study was 41.1%, where 51% of these cases were typical, and 49% had an atypical deletion. Thus, the clinical sign cardiac anomaly has a similar frequency in both cases, which means that it is not a useful attribute to differentiate typical from atypical cases.

Burnside, 2015 reported the presence of cardiac defects in patients with deletions involving different CSFs22. The study of Burnside.,2015 reports that subjects with proximal deletions involving the 22 A-B CSFs should present cardiac defects as often as those with central deletions and less frequently than those with proximal deletions LCRs22 A-D (Burnside, 2015). However, there is insufficient clinical information to compare the frequency of congenital heart defects in the different types of 22q11 region deletion.

Here we demonstrate that the clinical sign abnormality of the cardiovascular system does not significantly assist in the classification of typical and atypical deletions. In addition, cardiovascular anomalies are often identified in the prenatal or neonatal period leading to diagnosis (Sullivan, 2019). Therefore, it is suggested that patients that have some abnormality of the cardiovascular system must be investigated throughout the 22q11 region in order to account for the presence of both typical and atypical deletions.

5. Conclusion

It was possible to use the J48 algorithm to classify and interpret the clinical signs that differentiate the patients into typical and atypical cases. The selected trees were highly accurate at 83-91% and CCI at 83-91%. Attributes used as nodes of the tree-like oral cleft, velopharyngeal insufficiency, delayed speech and language development, specific learning disability, behavioral abnormality, and growth delay have been studied and validated in the literature as being the classical signs of the 22q11DS.

It was seen that typical cases of 22q11DS were classified when they had the presence of combined clinical signs, especially oral cleft with a specific learning disability; delayed speech and language development with velopharyngeal insufficiency or behavioral abnormality. This way, according to our results, patients who present this pattern of clinical signal could be diagnosed with FISH technique. In contrast, most of the atypical cases did not present the typical clinical signs of the syndrome, implying that atypical cases require further clinical investigations to be diagnosed.

Furthermore, the results of the generated decision trees may lead to reflection on current clinical practice in the context of 22q11DS. Noteworthy, it was not possible to classify the patients in typical and atypical according to the presence of heart disease. Thus, the use of diagnostic methods that explore the entire 22q11 region is relevant when the patient has an abnormality of the cardiovascular system.

We can say that the machine-learning method accomplishes the goal proposed in this work. However, we do not aim to create a classificatory model to be applied on the clinical routine. The machine learning-based methods, especially the use of the decision tree, was used here in order to aid in the interpretability of the results. The disadvantage of the use of decision tree is this kind of analysis can be influenced by its dataset. In other words, due to differences in clinical assessment from various clinics and locations, could be a distinction between decision trees made based in other datasets. To overcome this difficulty we suggest that more cases should be added for future analysis and recommend that the same clinical evaluation should be made for each patient.

6. Reference

Alpaydin, E. (2010) *Introduction to Machine Learning*. 2nd ed. The MIT Press. doi: 10.1016/j.neuroimage.2010.11.004.

Bassett, A. (2011) ‘Practical guidelines for managing patients with 22q11. 2 deletion

syndrome', *J Pediatr.*, 17(2), pp. 281–294. doi: 10.1016/j.jpeds.2011.02.039.Practical.

Burnside, R. D. (2015) '22q11.21 deletion syndromes: A review of proximal, central, and distal deletions and their associated features', *Cytogenetic and Genome Research*, 146(2), pp. 89–99. doi: 10.1159/000438708.

Carlson, C. *et al.* (1997) 'Molecular Definition of 22q11 Deletions in 151 Velo-Cardio-Facial Syndrome Patients', *The American Journal of Human Genetics*, 61(3), pp. 620–629. doi: 10.1086/515508.

Clements, C. C. *et al.* (2017) 'Critical region within 22q11 . 2 linked to higher rate of autism spectrum disorder'. Molecular Autism, pp. 1–17. doi: 10.1186/s13229-017-0171-7.

Dugan, T. M. *et al.* (2015) 'Machine Learning Techniques for Prediction of Early Childhood Obesity', *Applied Clinical Informatics*, 6(3), pp. 506–520. doi: 10.4338/ACI-2015-03-RA-0036.

Dugoff, L., Mennuti, M. T. and McDonald-McGinn, D. M. (2017) 'The benefits and limitations of cell-free DNA screening for 22q11.2 deletion syndrome', *Prenatal Diagnosis*, 37(1), pp. 53–60. doi: 10.1002/pd.4864.

EMPKE, S. L. L. (2015) *Caracterização fenotípica em indivíduos com microarranjos na região cromossômica 22q11*. Universidade de São Paulo.

Habel, A. *et al.* (2012) 'Syndrome-specific growth charts for 22q11.2 deletion syndrome in Caucasian children', *American Journal of Medical Genetics, Part A*, 158 A(11), pp. 2665–2671. doi: 10.1002/ajmg.a.35426.

Habibi, S., Ahmadi, M. and Alizadeh, S. (2015) 'Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining', *Global Journal of Health Science*, 7(5), pp. 304–310. doi: 10.5539/gjhs.v7n5p304.

Hall, M. *et al.* (2009) 'The WEKA data mining software', *SIGKDD Explorations Newsletter*, 11(1), p. 10. doi: 10.1145/1656274.1656278.

Hay, B. N. (2007) 'Deletion 22q11: Spectrum of Associated Disorders', *Seminars in Pediatric Neurology*, 14(3), pp. 136–139. doi: 10.1016/j.spen.2007.07.005.

Kobrynski, L. J. and Sullivan, K. E. (2007) 'Velocardiofacial syndrome, DiGeorge syndrome: the chromosome 22q11.2 deletion syndromes', *Lancet*, 370(9596), pp. 1443–1452. doi: 10.1016/S0140-6736(07)61601-8.

Köhler, S. *et al.* (2017) 'The human phenotype ontology in 2017', *Nucleic Acids Research*, 45(D1), pp. D865–D876. doi: 10.1093/nar/gkw1039.

Kruszka, P. *et al.* (2017) '22q11.2 deletion syndrome in diverse populations', *American Journal of Medical Genetics Part A*, 173(4), pp. 879–888. doi: 10.1002/ajmg.a.38199.

Landrum, M. J. *et al.* (2018) 'ClinVar: Improving access to variant interpretations and supporting evidence', *Nucleic Acids Research*. Oxford University Press, 46(D1), pp. D1062–

D1067. doi: 10.1093/nar/gkx1153.

Lindsay, E. A. (2001) 'Chromosomal microdeletions: Dissecting DEL22Q11 syndrome', *Nature Reviews Genetics*, 2(11), pp. 858–868. doi: 10.1038/35098574.

Maggadottir, S. M. and Sullivan, K. E. (2013) 'The Diverse clinical features of chromosome 22q11.2 deletion syndrome (DiGeorge Syndrome)', *Journal of Allergy and Clinical Immunology: In Practice*. Elsevier Inc, 1(6), pp. 589–594. doi: 10.1016/j.jaip.2013.08.003.

Maimon, O. and Rokach, L. (2015) 'Data mining with decision trees: theory and applications'. Ben-Gurion University of the Negev, Israel: World Scientific Publishing Co. Pte. Ltd.

McDonald-McGinn, D. M. *et al.* (2015) '22Q11.2 Deletion Syndrome', *Nature Reviews Disease Primers*, 1(November). doi: 10.1038/nrdp.2015.71.

Morrow, B. E. *et al.* (2018) 'Molecular genetics of 22q11.2 deletion syndrome', *American Journal of Medical Genetics Part A*, 176(10), pp. 2070–2081. doi: 10.1002/ajmg.a.40504.

Panamonta, V. *et al.* (2016) 'Birth Prevalence of Chromosome 22q11.2 Deletion Syndrome: A Systematic Review of Population-Based Studies.', *Journal of the Medical Association of Thailand = Chotmaihet thangphaet*, 99 Suppl 5(18), pp. S187-93. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29906080.

Quinlan, J. R. (1986) 'Induction of Decision Trees', *Machine Learning*, 1(1), pp. 81–106. doi: 10.1023/A:1022643204877.

Richieri-costa, A. (2008) 'Craniofacial Morphology in Patients With Velocardiofacial Syndrome'. doi: 10.1597/08-278.1.

Robin, N. and Shprintzen, R. (2005) 'Defining the clinical spectrum of deletion 22q11. 2', *The Journal of pediatrics*, 147, pp. 90–96. doi: 10.1016/j.jpeds.2005.03.007.

Rump, P. *et al.* (2014) 'Central 22q11.2 deletions', *American Journal of Medical Genetics, Part A*, 164(11), pp. 2707–2723. doi: 10.1002/ajmg.a.36711.

Shprintzen, R. J. (2008) 'Velo-Cardio-Facial Syndrome: 30 Years of Study', *Developmental Disabilies Research Reviews*, 14(1), pp. 3–10. doi: 10.1002/ddrr.2.Velo-Cardio-Facial.

Souto, M. C. P. *et al.* (2003) 'Técnicas de aprendizado de máquina para problemas de biologia molecular', *Sociedade Brasileira de Computaçao*, (October). Available at: http://www.cin.ufpe.br/~mcps/ENIA2003/jaia2003-14-08.pdf.

Spruijt, N. E. *et al.* (2012) 'Velopharyngeal Dysfunction and 22q11.2 Deletion Syndrome: A Longitudinal Study of Functional Outcome and Preoperative Prognostic Factors', *The Cleft Palate-Craniofacial Journal*, 49(4), pp. 447–455. doi: 10.1597/10-049.

Sullivan, K. E. (2019) 'Chromosome 22q11.2 deletion syndrome and DiGeorge syndrome', *Immunological Reviews*, 287(1), pp. 186–201. doi: 10.1111/imr.12701.

Tang, K. L. *et al.* (2015) 'Behavioral and Psychiatric Phenotypes in 22q11.2 Deletion

Syndrome', *Journal of Developmental & Behavioral Pediatrics*, 36(8), pp. 639–650. doi: 10.1097/DBP.0000000000000210.

Turan, S. *et al.* (2008) 'Constitutional Growth Delay Pattern of Growth in Velo-Cardio-Facial Syndrome : Longitudinal', 1(1), pp. 43–48. doi: 10.4008/jcrpe.v1i1.13.

Witten, I. H. , Frank, E., & Hall, M. A. (2011) *Data Mining : Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers*. doi: 10.1016/C2009-0-19715-5.

Yamagishi, H. and Srivastava, D. (2003) 'Unraveling the genetic and developmental mysteries of 22q11 deletion syndrome', *Trends in Molecular Medicine*, 9(9), pp. 383–389. doi: 10.1016/S1471-4914(03)00141-2.

7. Supplementary material

7.1. Supplementary table 1 - Articles used to obtain clinical cases of DS22q11.

Case reports	Number of cases
BENGOA-ALONSO e colab., 2016	3
EMPKE, 2015	45
OGILVIE e colab., 2009	1
BRECKPOT e colab., 2012	1
GARAVELLI e colab., 2011	1
RUMP e colab., 2014	27
TAN, Tiong Yang e colab., 2011	3
RØDNINGEN e colab., 2008	2
RAKONJAC e colab., 2016	2
BEN-SHACHAR e colab., 2008	6
NOGUEIRA e colab., 2008	1
MONTEIRO e colab., 2013	3
TOTAL	**95**

7.2. Supplementary table 2 - Frequency of the 43 clinical features selected. It can be oserved the number of cases who presented each clinical sign as well as their frequency (%).

TOTAL	95					
Clinical characteristics found	***n***	Total (%)	Typical (n)	Typical (%)	Atypical (n)	Atypical (%)
Craniofacial/oral						
Long Face	20	21.1%	15	75%	5	25%
Retrognathia	20	21.1%	20	100%	0	0%
Abnormality of the shape of the midface	18	18.9%	18	100%	0	0%
Abnormal eyebrow morphology	12	12.6%	4	33%	8	67%
Decreased head circumference	11	11.6%	4	36%	7	64%
Facial asymmetry	8	8.4%	2	25%	6	75%
Micrognathia	7	7.4%	4	57%	3	43%
Malar flattening	7	7.4%	6	86%	1	14%
Triangular face	7	7.4%	0	0%	7	100%
Abnormality of the chin	6	6.3%	1	17%	5	83%
Narrow face	6	6.3%	5	83%	1	17%
High forehead	5	5.3%	0	0%	5	100%
Broad forehead	4	4.2%	0	0%	4	100%
Short neck	3	3.2%	3	100%	0	0%
Abnormality of the hairline	2	2.1%	2	100%	0	0%
Flat face	2	2.1%	1	50%	1	50%
Prominent forehead	2	2.1%	2	100%	0	0%
Square face	2	2.1%	0	0%	2	100%
Increased head circumference	2	2.1%	1	50%	1	50%
Brachycephaly	1	1.1%	0	0%	1	100%
Facial hypotonia	1	1.1%	0	0%	1	100%
Frontal bossing	1	1.1%	0	0%	1	100%
Hemifacial microsomia	1	1.1%	1	100%	0	0%
Small forehead	1	1.1%	0	0%	1	100%
Mouth						
Abnormality of the philtrum	20	21.1%	7	35%	13	65%

Thin vermilion border	17	17.9%	7	41%	10	59%
Narrow mouth	8	8.4%	7	88%	1	13%
Thick vermilion border	5	5.3%	3	60%	3	60%
Macroglossia	3	3.2%	3	100%	0	0%
Downturned corners of mouth	3	3.2%	1	33%	2	67%
Open mouth	3	3.2%	1	33%	2	67%
Ankyloglossia	1	1.1%	0	0%	1	100%
Nasal findings						
Abnormality of the nasal alae	20	21.1%	12	60%	8	40%
Abnormality of the nasal tip	14	14.7%	8	57%	6	43%
Abnormality of the nasal bridge	12	12.6%	5	42%	7	58%
Long nose	11	11.6%	11	100%	0	0%
Abnormal nasal base	10	10.5%	10	100%	0	0%
Eye findings						
Abnormality of the palpebral fissures	43	45.3%	32	74%	11	26%
Epicanthus	9	9.5%	5	56%	4	44%
Ptosis	6	6.3%	0	0%	6	100%
Hypertelorism	5	5.3%	2	40%	3	60%
Deeply set eye	5	5.3%	0	0%	5	100%
Strabismus	4	4.2%	0	0%	4	100%
Hypotelorism	3	3.2%	0	0%	3	100%
Microphthalmia	3	3.2%	3	100%	0	0%
Astigmatism	2	2.1%	2	100%	0	0%
Abnormally large globe	2	2.1%	2	100%	0	0%
Hypermetropia	2	2.1%	2	100%	0	0%
Stellate iris	1	1.1%	1	100%	0	0%
Ear/hearing findings						
Small earlobe	29	30.5%	*27*	93%	*2*	7%
Overfolded helix	20	21.1%	13	65%	7	35%
Protruding ears	15	15.8%	12	80%	3	20%
Hearing impairment	*14*	14.7%	11	79%	3	21%
Low set ears	6	6.3%	2	33%	4	67%
Auricular tag	6	6.3%	0	0%	6	100%

Asymmetry of the ears	4	4.2%	2	50%	2	50%
Pointed helix	2	2.1%	2	100%	0	0%
Cupped ear	2	2.1%	1	50%	1	50%
Bifid lobes	2	2.1%	2	100%	0	0%
Posteriorly rotated ears	1	1.1%	0	0%	1	100%
Limb Findings						
Long fingers	36	37.9%	29	81%	7	19%
Slender finger	8	8.4%	8	100%	0	0%
Syndactyly	5	5.3%	0	0%	5	100%
Fifth finger clinodactyly	4	4.2%	0	0%	4	100%
Short digit	4	4.2%	0	0%	4	100%
Polydactyly	2	2.1%	2	100%	0	0%
Short foot	2	2.1%	2	100%	0	0%
Deviation of the thumb	2	2.1%	2	100%	0	0%
Square digits	2	2.1%	2	100%	0	0%
Long arms	2	2.1%	2	100%	0	0%
Prominen interdigital folds	1	1.1%	1	100%	0	0%
Aplasia/hypoplasia of the palma creases	1	1.1%	1	100%	0	0%
Curved phalanges of the hand	1	1.1%	1	100%	0	0%
Narrow chest	1	1.1%	1	100%	0	0%
Other characteristics						
Specific learning disability	64	67.4%	44	69%	20	31%
Behavioral abnormality	56	58.9%	40	71%	16	29%
Delayed speech/ language development	52	54.7%	41	79%	11	21%
Oral Cleft	48	50.5%	45	94%	3	6%
Velopharyngeal insufficiency	39	41.1%	36	92%	3	8%
Abnormality of the cardiovascular system	*39*	41.1%	20	51%	19	49%
Abnormality of skeletal morphology	18	18.9%	2	11%	16	89%
Abnormality of the genitourinary system	13	13.7%	1	8%	12	92%
Growth delay	12	12.6%	0	0%	12	100%
Motor delay	10	10.5%	1	10%	9	90%

Hernia	10	10.5%	3	30%	7	70%
Muscular hypotonia	9	9.5%	5	56%	4	44%
Abnormality of the uvula	7	7.4%	7	100%	0	0%
Hypothyroidism	3	3.2%	2	67%	1	33%
Anorectal anomaly	3	3.2%	0	0%	3	100%
High-pitched voice	1	1.1%	0	0%	1	100%

CAPÍTULO 2

CAPÍTULO 2

A SYSTEM BIOLOGY APPROACH IN THE STUDY OF THE 22q11 DELETION SYNDROME

Camila C. Alves[a*], Ivan R. Wolf[c], Bruno F. Gamba[b], Lucilene Ribeiro-Bicudo[b], Guilherme T. Valente[c]

[a] São Paulo State University (UNESP), Botucatu, São Paulo, Brazil
[b] Biological science institute, Federal University of Goiás (UFG), Goiânia, Goiás, Brazil.
[c] School of Agronomic Sciences, São Paulo State University (UNESP), Botucatu, São Paulo, Brazil
*cris_camila@yahoo.com

Abstract

The 22q11 deletion syndrome (22q11.2DS) is a haploinsufficient disorder that involves microdeletion ranging from 0.7 to 3Mb on chromosome 22 which can affect different organs and systems. Although the noticeable variety of clinical features, approximately 90% of the cases have a 3Mb deletion involving LCR22A and LCR22D. Here, we hypothesize that the deletion of genes within the 22q11 region can disrupt the function and/or interaction of proteins outside this region. Thus, we aim to explore the consequences of the 3Mb 22q11 deletion from a system biology view using network biology. To achieve the goal, the protein-coding genes involved in the deletion were removed from human protein-protein interaction (PPI) network in order to simulate the deletion. Then, an analysis of the topological changes was used to obtain relevant biological interpretations and to infer the function and/or relevance of genes associated with SD22q11. PPI networks were obtained from public repositories. The unification of the PPI networks obtained was called Global Network (GN), then 48 nodes were removed from the GN to create the Patient Network (PN). Metrics from GN and PN were calculated, and each network was divided into communities. Gene Ontology Enrichment Analysis was performed within the communities. The GN had 21,492 proteins connected by 682,910 interactions while the PN has 21,444 proteins connected by 679,817 interactions. It was showed that there is no significant difference when comparing GN and PN. The analysis of neighboring proteins within the communities showed that there is a difference between the global and patient communities and was observed that the degree and betweenness of the neighboring protein increase, on average, in the patient's network. In addition, genes among the 3Mb deleted region in 22q11DS was enriched by SAFE analysis like *ZNF74*, *SNAP29*, *TRMT2A*, *LZTR1*, *TBX1*, *THAP7*, *DGCR6*, *CLDN5*, *SERPIND1*, *TSSK2*, *MED15*, *CRKL*, and *CLTCL1*. In conclusion, our results could found some genotype-phenotype relationship of 22q11DS with the genes analyzed through the PPI network. From the analyses of the neighboring proteins, we found that the gene *G6PT2* lost its betweenness in the PN and that the symptoms failure to thrive and short stature caused by this gene has also reported on cases of 22q11DS. Also, the results found through the GO enrichment analysis shows that GO terms linked to the genes agree with information from the literature which means that analysis using ontological terms can bring significant data for disease studies. Therefore, approaches using network biology has great potential for the discovery properties of different pathologies that are not readily accessible by conventional molecular biology or genetic approaches.

Keywords: 22q11DS; DiGeorge Syndrome; PPI network.

1. Introduction

The 22q11 deletion syndrome (22q11.2DS) (Online Mendelian Inheritance in Man - OMIM # 192430), also known as DiGeorge syndrome (OMIM # 188400) is a haploinsufficient disorder that involves microdeletion on chromosome 22 ranging from 0.7 to 3Mb, which results in a broad phenotypic spectrum (McDonald-McGinn *et al.*, 2015). The prevalence of the syndrome in question is estimated at one case per 4,000 to 9,800 live births, which makes the 22q11DS the most frequent chromosomal microdeletion (Burnside, 2015; McDonald-McGinn *et al.*, 2015; Panamonta *et al.*, 2016; Dugoff, Mennuti and McDonald-McGinn, 2017).

The affected individuals exhibit diverse phenotypes like congenital heart disease, palatal abnormalities, immunodeficiency, autoimmune disease, hypocalcemia, speech delay, cognitive deficits and neuropsychiatric illnesses, developmental delays, skeletal anomalies, thrombocytopenia, endocrine, genitourinary and gastrointestinal problems, and characteristic facial features (Yamagishi and Srivastava, 2003; Bassett *et al.*, 2011; McDonald-McGinn *et al.*, 2015; Bernice E Morrow *et al.*, 2018; Kathleen E Sullivan, 2019).

Despite the clinical features variety, approximately 90% of the cases have a 3Mb deletion involving LCR22A and LCR22D, also knows as a typical deleted region (TDR), where 45 known protein-coding genes, seven microRNAs and ten non-coding RNAs are present (Yamagishi and Srivastava, 2003; McDonald-McGinn *et al.*, 2015; Bernice E Morrow *et al.*, 2018). Besides that, 8% of patients have a smaller deletion of 1.5Mb, and the minority have atypical deletions with different sizes (Edelmann, Pandita, and Morrow, 1999; Yamagishi and Srivastava, 2003; Rosa *et al.*, 2009; McDonald-McGinn *et al.*, 2015).

The function elucidation and importance of each gene on chromosome 22q11 has been sought, and the most studied gene in this region is *TBX1* due to its role in the syndrome aspects (Baldini, Fulcoli and Illingworth, 2017). The *TBX1* encodes a T-box transcription factor, which is crucial in the development of the craniofacial region, thymus, and parathyroid glands, aortic arch as well as cardiac outflow tract (McDonald-McGinn *et al.*, 2015; Baldini, Fulcoli and Illingworth, 2017; Kathleen E. Sullivan, 2019). In addition, *TBX1* has also been associated with dental problems, feeding, and deglutition difficulties and in the development and function of the ear (Scambler, 2000; Jerome and Papaioannou, 2001; Gao, Li and Amendt, 2013; Chen *et al.*, 2016). In addition to the *TBX1*, other genes are considered critical in the manifestation of the main clinical features of the syndrome, such as the *HIRA*, *COMT* and *CRKL* (Zeitz *et al.*, 2013; Burnside, 2015; Racedo *et al.*, 2015; Ju *et al.*, 2016; Yang *et al.*, 2016).

Since the lack of common phenotypic spectrum considering all patients harboring the 22q11 deletion, it is possible to suppose that genetic variation outside the 22q11 region and

disturbance due to gene deletion could contribute to the syndrome development phenotype (Michaelovsky *et al.*, 2019; Santoro *et al.*, 2019). In fact, a study suggests that molecular events like CNVs and missense mutations outside the 22q11 can lead to disruptions in the neurodevelopmental process (Michaelovsky *et al.*, 2019).

The study of biological networks is a holistic approach and has been benefited from the application of high-throughput experiments (Schmith *et al.*, 2005; Raman, 2010). Thus, a biological network can be described as a series of nodes (e.g., proteins, genes, metabolites or miRNA) connected to each other by links/edges representing interactions between two components (Barabási and Oltvai, 2004a; Han, 2008; Chan and Loscalzo, 2012). Study biological networks considering interconnections among genes instead of considering genes function isolated, have been allowing new insights in health science (Rual *et al.*, 2005; Han, 2008; Chan and Loscalzo, 2012). Altogether, the network theory has been applied to human diseases studies like cancer, cardiac disease and diabetes (Ashley *et al.*, 2006; Sengupta *et al.*, 2009; Diez *et al.*, 2010; Jensen *et al.*, 2011; Wang, Gulbahce and Yu, 2011; Lui *et al.*, 2017).

Different types of biological interaction such as PPI, metabolic networks, signaling and transcription-regulatory networks have been published and updated (Barabási and Oltvai, 2004a). The PPI network can be characterized as an undirected network where a link represents a mutual binding between two proteins (Barabási and Oltvai, 2004a).

Previously PPI network and Gene Ontology (GO) terms already provided new insights in genes biological roles (Ashburner *et al.*, 2000; Li *et al.*, 2018; Gene and Consortium, 2019). Besides, genes associated with similar diseases tend to be in the same neighborhood within the molecular networks and form physical and functional modules (Wang, Gulbahce and Yu, 2011). Therefore, computational approaches can help to find candidate disease genes by predicting their functions and/or analyzing the gene neighborhood.

Since the 22q11DS phenotype could be influenced by genes outside the deletion region (Michaelovsky *et al.*, 2019; Santoro *et al.*, 2019), here, we hypothesize that the deletion of genes within the 22q11 region can disrupt the function and/or interaction among proteins coded by genes outside this region. Thus, we aim to explore the consequences of the 3Mb 22q11 deletion under the system biology view. To achieve the goal, the protein-coding genes involved in the deletion were removed from the human PPI network in order to simulate a network of a patient harboring the deletion. Then, topological features of networks were used to obtain relevant biological interpretations and to infer the function and/or relevance of genes associated with SD22q11. Hence, it was possible to report that the PPI network had robustness against the 22q11 deletion. However, the 22q11 deletion shifts the network structure at the community

level, changing the number of communities and the number of nodes of each community. Moreover, it was possible to infer the genotype-phenotype relationship with 22q11DS using network ontology enrichment.

2. Methods

2.1. Data sources

A PPI network of human was downloaded from public repositories (Biogrid, ComPPi, HumanNet, and HuRI (The Human Reference Protein Interactome Mapping Project)) (Table 1). First, it was selected, for each network, the protein interactions that were inferred from experimental methods. Then, the four networks were unified, and the protein identifiers were standardized to Uniprot nomenclature. Next, redundant interactions were removed from the unified network, which we call the Global Network (GN).

Table 1. Databases containing large-scale *Homo sapiens* PPIs.

Database	URL	Nº interactions	References
Biogrid	*https://thebiogrid.org/*	1,670,339	(Stark, 2005)
ComPPi	*http://comppi.linkgroup.hu/*	1,311,184	(Veres *et al.*, 2015)
HumanNet	*http://www.functionalnet.org/humannet/about.html*	18,714	(Blom *et al.*, 2011)
HuRI	*http://interactome.baderlab.org/*	14,921	(Rolland T. Tasan M., 2014)

2.2. Establish patient and global network

The unification of the PPI networks obtained in the previous step was used as a control network, called the global network (GN); this network was used to simulate someone without any deletion. To establish the PPI network simulating a patient harboring the 3Mb 22q11del (chr22: 18.168.234 – 21.206.711, according to Hg38) (Bertini *et al.*, 2017), the proteins and their interactions related to the deleted genes were removed from the GN. Thus, a total of 48 nodes were removed from the GN creating the PN (the patient network).

2.3. Network metrics

The R package igraph was used to network analysis (Kolaczyk and Csárdi, 2014) (Csárdi and Nepusz, 2006) in which it was calculated many topological features of both networks, allowing comparisons between GN and PN helping to better understanding how the 22q11 deletion can affect the *Homo sapiens* PPI network.

The topological metrics calculated were “Degree”, “Eigenvalue centrality”, “Diameter”, “Path length”, “Number of vertices”, “Number of edges”. In order to compare the metrics between GN and PN, first the Shapiro-Wilk test was used to verify the type of data distribution, and the Mann-Whitney U was used to compare the topological metrics between networks. Additionally, the first neighbors of deleted nodes (hereafter referred to as “neighboring proteins”) in both networks were analyzed and compared as described above.

2.4. Neighboring genes in the context of communities

Each network was split into communities through the algorithm *fastgreedy.community* implemented in Igraph. This algorithm is able to infer community structure from network topology which works by greedily optimizing the modularity (Clauset, Newman and Moore, 2004). After, the network topologies for neighboring proteins "within" the communities were recalculated. Also, the Pearson coefficient correlation between the metric degree and betweenness was calculated for GN and PN.

To select the proteins that suffered the most substantial variations from GN to PN, it was used the algorithm "SimpleKmeans" of WEKA (Hall *et al.*, 2009; Witten, I. H. , Frank, E., & Hall, 2011). Thus, the $\log_2$ fold-change of degree and betweenness were calculated, and the proteins were classified into one of three clusters. The cluster containing proteins with the most considerable variations were selected.

2.5. Gene Ontology Enrichment Analysis

Gene Ontology Enrichment Analysis was performed to the communities from each network using SAFE (Spatial analysis of functional enrichment) (Baryshnikova, 2016a). The Gene Ontology (GO) terms of biological process (BPs) were used as input for the enrichment of the characteristics of the network nodes (Ashburner *et al.*, 2000; Gene and Consortium, 2019).

3. Results

3.1 Analysis of the PPI networks of each established group

The GN PPI network has 21,492 proteins as nodes connected by 682,910 interactions (edges). Without the 48 nodes equivalent to the genes from the 3Mb deleted region of 22q11DS, the PN has 21,444 nodes and 679,817 edges. Network topological properties are reported in Table 2.

Table 2. Properties calculated for the global network (GN) and patient network (PN).

Topological properties	GN	PN
Average degree	63.5	63.40
Average betweenness	18693.76	18660.42
Average eigenvalue	0.031	0.031
Average diameter	8	8
Average path length	2.74	2.74
Number of edges	682,910	679,817
Number of vertices	21492	21444

The values of degree, betweenness and eigenvector have a non-normal distribution according to the Shapiro-Wilk test. Thus, the Mann-Whitney U test was used to compare the properties between networks. The P-values were 0.7669, 0.8471 and 0.9317 for degree, betweenness, and eigenvector, respectively, indicating no significant difference between both networks (Figure 1).

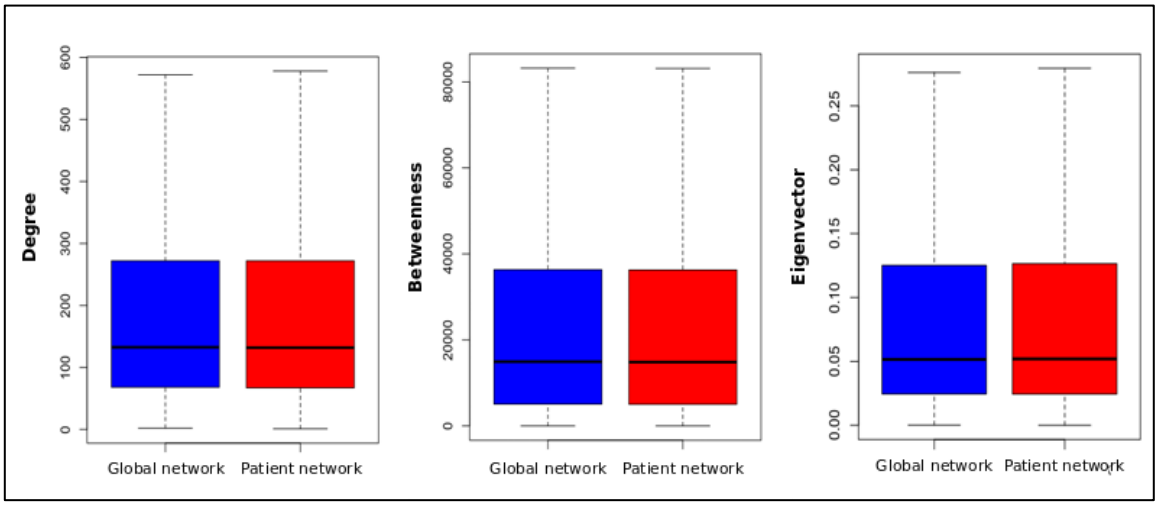


Figure 1. Boxplot of the degree, betweenness, and eigenvector of GN and PN.

3.2. Community context analysis of neighboring proteins

The GN and PN have a total of 13 and 19 communities, respectively (Table 3). A total of 2,404 neighboring proteins have been found to be distributed within communities 1 to 5 of the GN, while 2,389 neighboring proteins are distributed within communities 1 to 6 of the PN. The number of neighboring proteins from GN communities and PN communities had a difference of 15 proteins, which belongs to the 22q11 deleted region.

Table 3. A number of communities from GN and PN with the number of proteins, deleted proteins and neighboring proteins.

Community	Total proteins		Deleted protein	Neighboring proteins	
	GN	PN	GN	GN	PN
1	7,734	10,882	25	497	675
2	4,338	5,517	6	464	667
3	4,217	3,884	7	468	1025
4	2,819	343	7	699	18
5	1,630	741	4	274	3
6	723	19	-	2	1
7	9	10	-	-	-
8	6	8	-	-	-
9	7	9	-	-	-
10	3	3	-	-	-
11	2	6	-	-	-
12	2	7	-	-	-
13	2	3	-	-	-
14 a 19	-		-	-	-

GN = Global Network; PN = Patient Network

The degree and betweenness metrics of neighboring proteins were calculated for GN and PN communities, and the Shapiro-Wilk statistical test showed they are not normally distributed. Then, the Mann-Whitney U test showed there is a difference between the properties between the communities (P-value of $6.669E^{-16}$ and $1.701E^{-16}$ for the degree and betweenness, respectively). The Log2 fold-change of the degree and betweenness can be observed in figure 2B and figure 2D. The boxplot of these metrics can be seen in Figure 2A and Figure 2C.

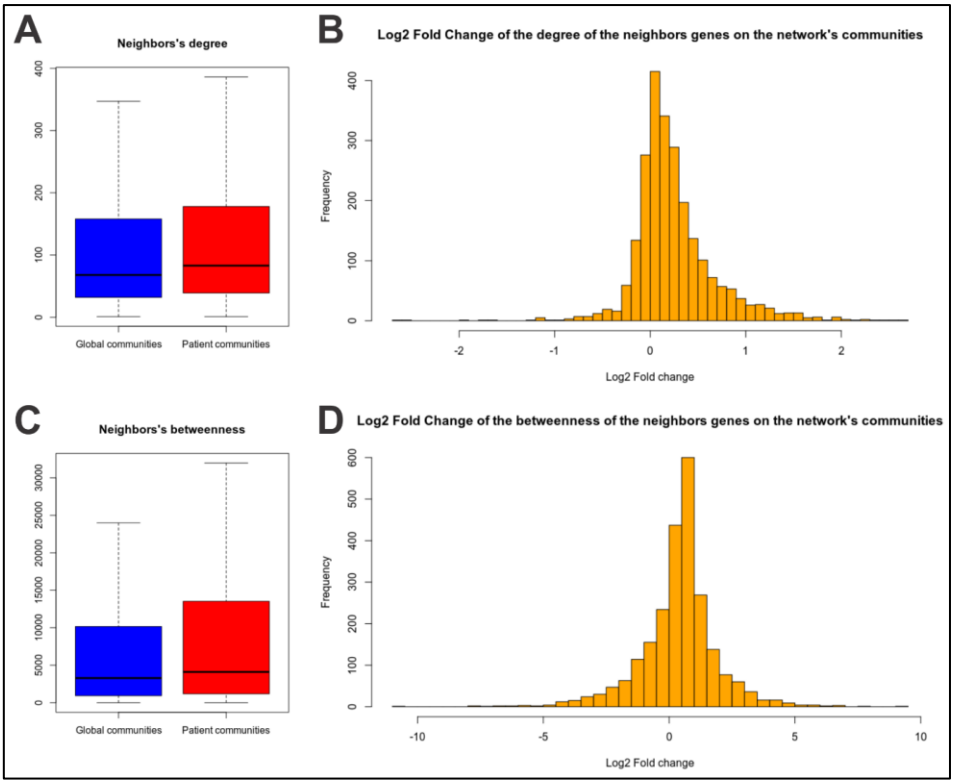


Figure 2. Parameters to evaluate the communities. **(A)** Boxplot comparing the degree of the neighboring proteins from the global communities and from patient communities. **(B)** Log2 Fold Change of the neighboring protein's degree from global and patient communities. **(C)** Boxplot comparing the betweenness of the neighboring proteins from the global communities and from patient communities. **(D)** Log2 Fold Change of the neighboring protein's betweenness from global and patient communities.

The Pearson coefficient correlation was calculated for the GN was 0.6737649 (P-value $<2.2E^{-16}$) and 0.7296686 (P-value $<2.2E^{-16}$) for the PN, showing there is a correlation between the metric degree and betweenness (Figure 3).

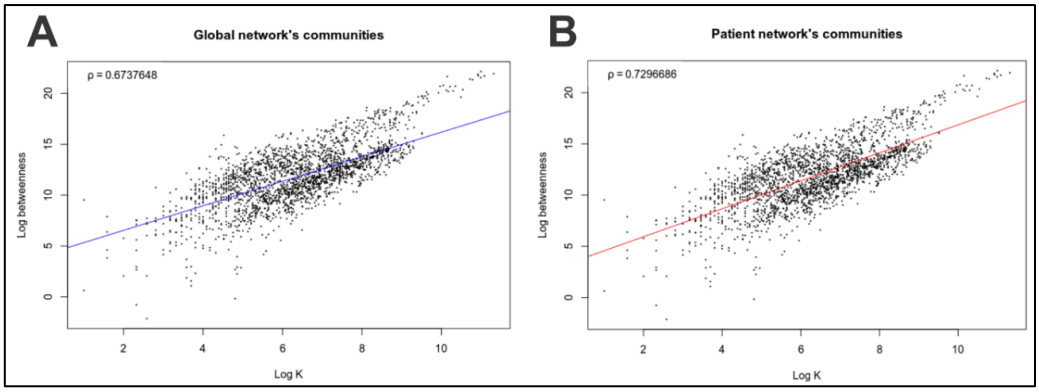


Figure 3. Histogram of the degree and betweenness of the global and patient communities. **(A)** Distribution of the degree and betweenness of the nodes from the global community and the p-value from the Pearson coefficient. **(B)** Distribution of the degree and betweenness of the nodes from the patient community and the p-value from the Pearson coefficient.

From the results from figure 6 and 7, it can be observed that the degree and betweenness of the neighboring protein increase, on average, in the patient's network (examples are presented in Figure 4).

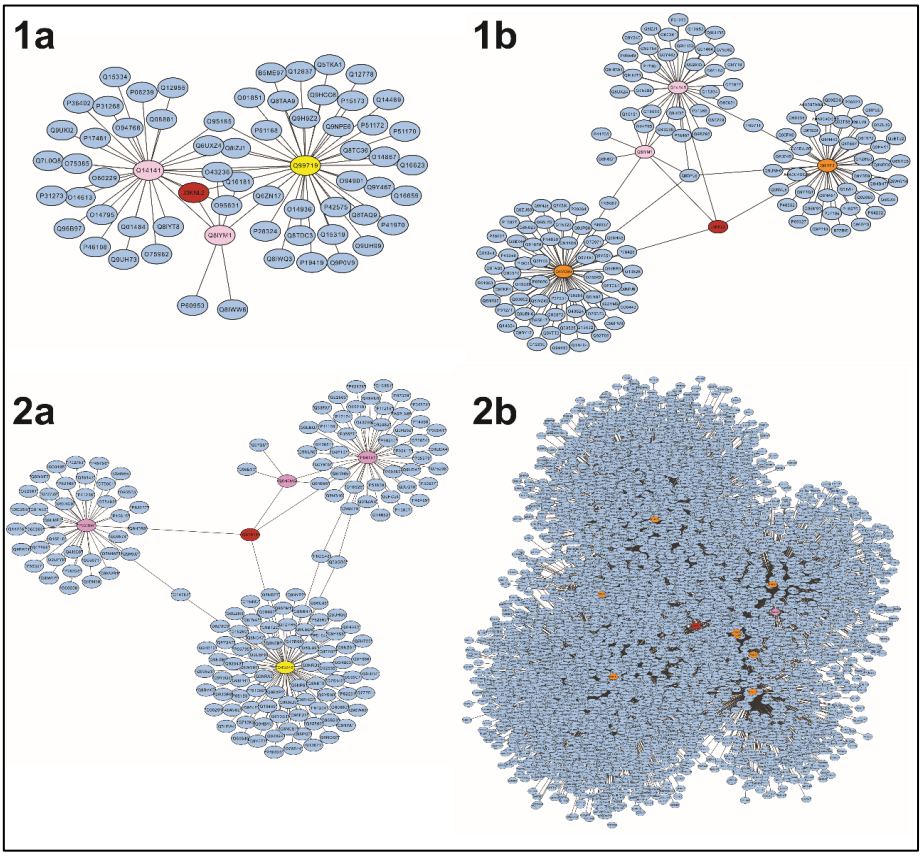


Figure 4. Examples of neighbor proteins that increase the degree in patient communities. **(1a)** Neighboring protein J3KNL2 and its interaction in the global community. **(1b)** Neighboring protein J3KNL2 and its interaction in the patient community. **(2a)** Neighboring protein Q92911 and its interaction in the global community. **(2b)** Neighboring protein Q92911 and its interaction in the patient community. **Yellow nodes**: a node that was deleted, equivalent from one of the proteins from the 22q11 region; **Red nodes**: an example of a neighboring protein where its degree is increased; **Pink nodes**: nodes that interact with the neighboring protein in question (red node); **Orange node**: new interactions of the neighboring protein

In order to select the proteins that passed by this shift, the clustering tool from WEKA software was used. Thus, the data was divided into 3 clusters (Figure 5) in order to sort the proteins that suffered the most changes in their metrics. The cluster with the highest fold-changes consists of 361 proteins (green dots on figure 5) and, therefore, these proteins were selected. Using the UniProt database, the functions of these proteins were determined, and the genes involved with a disease can be observed in the supplemental material 3. Furthermore, it was observed that 5 neighboring proteins, had the betweenness equals to 0 in which 3 of them (B5ME97, G6PT2, and SOWAHC) were 0 in the patient community while two (J3KNL2 and NUB1) were 0 in the GN. It shows that shifts are present after deletion (Table 4).

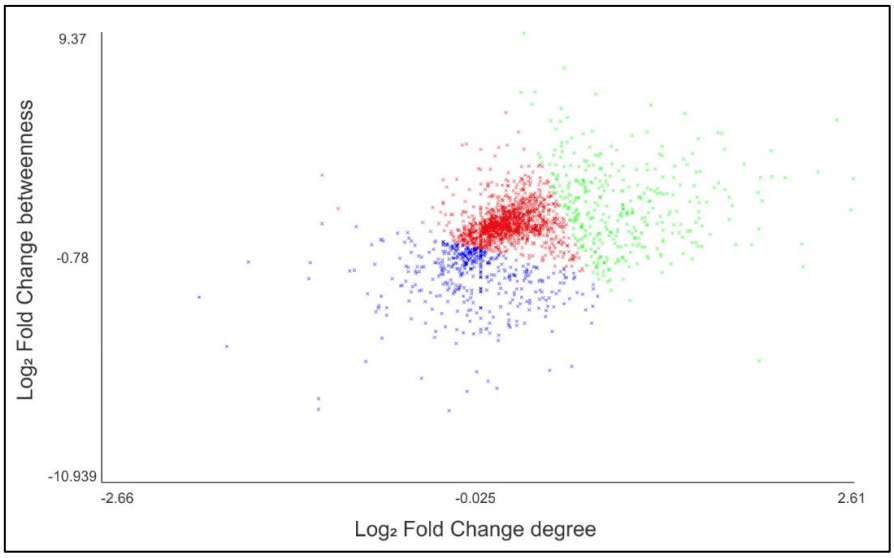


Figure 5. The three clusters, divided according to the degree and betweenness, formed using WEKA software. The cluster composed by green nodes were selected and is equitant with the nodes that suffered greater changes from GN to PN. The two other clusters composed of red and blue dots is equivalent to nodes that suffered few modifications.

Table 4. Proteins that lost or gained betweenness from global communities to patient communities.

Biogrid id/ Entry name	Global			Patient		
Protein	Community	Degree	Betweenness	Community	Degree	Betweenness
B5ME97	3	2	33.6	1	1	0
J3KNL2	3	3	0	1	4	19.5
P57057 / G6PT2	2	2	95.9	1	2	0
Q53LP3 / SOWAHC	1	6	234.5	4	1	0
Q9Y5A7 / NUB1	6	1	0	1	11	386.5

3.3. Gene Ontology Enrichment Analysis

The ontological terms were used to enrich the networks referring to the 1 to 6 GN and PN communities. The SAFE program (Baryshnikova, 2016a) was able to enrich communities 1, 2, 3, 4 and 5 of the GN and 1, 2 and 3 of the PN (Figure 6) (Supplementary material 1 and 2 shows these networks with the GO terms). Among the 3Mb deleted region in 22q11DS the genes which were in network regions enriched for similar GO biological terms were: *ZNF74*, *SNAP29*, *TRMT2A*, *LZTR1*, *TBX1*, *THAP7*, *DGCR6*, *CLDN5*, *SERPIND1*, *TSSK2*, *MED15*, *CRKL*, and *CLTCL1*. The biological process that each gene participates according to the ontological terms can be observed in table 5 and supplementary material 1.

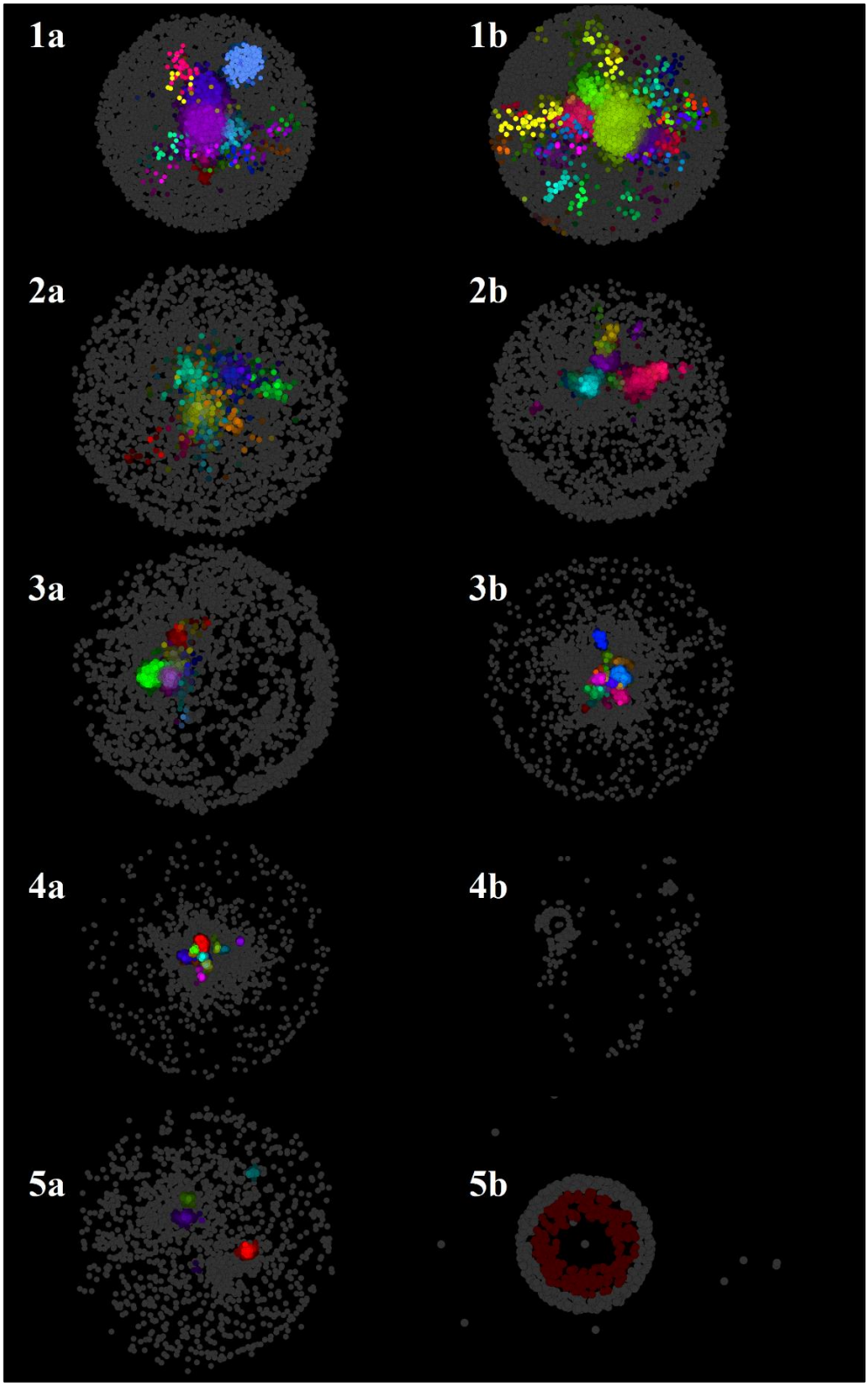


Figure 6. Global and patient communities that were enriched by Gene Ontology Enrichment Analysis using SAFE (Ashburner *et al.*, 2000; Baryshnikova, 2016b; Gene and Consortium, 2019). Network regions enriched for similar GO biological process terms are color-coded. **(1a)** Global community 1 annotated using SAFE. **(1b)** Patient community 1 annotated using SAFE. **(2a)** Global community 2 annotated using SAFE. **(2b)** Patient community 2 annotated using SAFE. **(3a)** Global community 3 annotated using SAFE. **(3b)** Patient community 3 annotated using SAFE. **(4a)** Global community 4 annotated using SAFE. **(4b)** Patient community 4 annotated using SAFE. **(5a)** Global community 5 annotated using SAFE. **(5b)** Patient community 5 annotated using SAFE.

Table 5. The 3Mb region deleted genes in 22q11DS from the nodes of the PN communities enriched by Gene Ontology Enrichment Analysis *(Ashburner* et al*., 2000; Gene and Consortium, 2019)*.

GENE NAME	GENE ONTOLOGY
ZNF74	RNA polymerase II transcription factor activity, sequence-specific DNA binding Metal ion binding
SNAP29 ***TRMT2A*** ***LZTR1***	Embryonic placenta development Cellular response to DNA damage stimulus Ubiquitin-dependent protein catabolic process Viral process Negative regulation of cell proliferation Regulation of hematopoietic stem cell differentiation Negative regulation of neuron death Negative regulation of apoptotic process Regulation of gene expression Regulation of transcription from RNA polymerase ii promoter
TBX1	Core promoter proximal region sequence-specific DNA binding RNA polymerase II transcription factor activity Protein dimerization activity
THAP7 ***DGCR6***	Cell division Protein localization to centrosome Non-motile cilium assembly Cell cycle Spindle assembly Centrosome duplication Regulation of centriole replication Ciliary basal body docking; cilium assembly
CLDN5	Calcium-independent cell-cell adhesion via plasma membrane Cell-adhesion molecules transmission of nerve impulse Regulation of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionate selective glutamate receptor activity
SERPIND1	Blood coagulation Post-translational protein modification Cellular protein metabolic process
TSSK2 ***MED15***	Establishment of cell polarity Cell division Regulation of mitotic cell cycle Cell cycle

	Intracellular signal transduction
	Protein autophosphorylation
	Protein phosphorylation
CRKL	Cellular response to cadmium ion
	Stress fiber assembly
	Peptidyl-tyrosine phosphorylation and autophosphorylation
	Apoptotic process
	Regulation of cell proliferation
	Cellular response to DNA damage stimulus
	Positive regulation of protein localization to plasma membrane
	Membrane organization
	Transmembrane receptor protein tyrosine kinase signaling pathway
	Cellular response to insulin stimulus
	Platelet activation
	Positive regulation of phosphatidylinositol 3-kinase activity
	Regulation of cell migration
	Fc-epsilon receptor signaling pathway
	T cell receptor signaling pathway
	Fc-gamma receptor signaling pathway involved in phagocytosis
	Stimulatory c-type lectin receptor signaling pathway
	Innate immune response
	Positive regulation of erk1 and erk2 cascade
	Positive regulation of protein phosphorylation
	Positive regulation of map kinase activity
	Axon guidance
	Cellular response to reactive oxygen species
	Ephrin receptor signaling pathway
	Vascular endothelial growth factor receptor signaling pathway
	Epidermal growth factor receptor signaling pathway
	Leukocyte migration
CLTCL1	Retrograde vesicle-mediated transport, Golgi to ER
	Protein transport
	Retrograde transport, endosome to Golgi

4. Discussion

The analyses of the 22q11DS with a biological network approach brought new information regarding some genes involved in the syndrome and a list of neighboring proteins that could have a potential role on the physiopathology of the syndrome. To simulate the PN it was removed 48 known protein-coding genes from the 3MB 22q11 region, that is 4 extra genes than the usually reported in the literature (Bernice E. Morrow *et al.*, 2018). However, we would like to point out it was used chr22: 18,168,234–21,206,711 coordinates (Hg38 assemble coordinates) to simulate a real patient deletion from the case report described by Bertini et al. 2017 (Bertini *et al.*, 2017).

We expected that removing nodes the result would be the disruption of a network (Barabási and Oltvai, 2004b). However, analyzing the global properties from the GN and PN, it was noticeable that gene deletion did not have a significant impact on the PPI network. The robustness against component failure can be explained by the topology of the biological PPI network, in which the degree distribution of both GN and PN is characterized by a power-law-degree distribution, in which most of the nodes have a few links and few nodes had a large number of links(Albert, Jeong, and Barabási, 2000; Barabási and Oltvai, 2004b).

Networks with a power degree distribution, such as the networks here studied, are called scale-free (Barabasi and Albert, 1999; Barabási and Oltvai, 2004b). Studies show that scale-free networks are very robust against random failures (Albert, Jeong, and Barabási, 2000; Barabási and Oltvai, 2004b; Friedel and Zimmer, 2007; Raman, 2010). This occurs because these failures mainly affect small degree nodes which do not alter the path structure of the remaining nodes, and thus do not disrupt the overall network integrity (Albert, Jeong, and Barabási, 2000; Barabási and Oltvai, 2004b; Raman, 2010). Altogether, we could conclude that both networks here studied seems to be real biological networks and the deletion is not able to disrupt overall biological features of PN network

On the other hand, the 22q11 deletion affected the network structure at the community level, which shifts on the number of communities and the number of nodes was observed. Moreover, statistical tests confirm that the properties like betweenness and degree significantly changed around neighboring genes. Taken together, besides lethal nodes can cause a disruption when removed (Pržulj, Wigle and Jurisica, 2004), here nodes with smaller degrees (mostly between 1 and 100) may have the potential to alter the form communities and probably the PPI interaction patterns. Besides, it has been reported that lethal nodes could be nodes highly connected or nodes that cause a disruption when removed (Pržulj, Wigle and Jurisica, 2004),

and in this way, it must be at least one protein derived from 22q11del that can cause a disruption.

Comparisons between the number of nodes from GN and PN communities, it can be noticed that PN 1, 2 and 3 communities had more significant number of nodes than GN. Hence, nodes present in GN 3, 4 and 5 communities became part of different communities after the 22q11 deletion. Thus, we suppose that communities rewiring may be occurring in an attempt to maintain the properties and interconnections of the network.

Besides the reorganization of the nodes within the communities, it was noticed that the value of betweenness for some genes was equal to zero, the meaning of the ones may have the potential to alter the community forms. From these nodes, 3 of them were equal to 0 in the PN community (B5ME97, G6PT2, and SOWAHC). The protein Septin 10 (B5ME97) and Ankyrin repeat domain-containing protein SOWAHC (*Q53LP3*/SOWAHC) had not been associated with the phenotype from this deletion. The *SLC37A4* seems to play an essential role in 22q11 phenotype development, once it encodes the G6PT2 protein and mutation on this gene can cause GSD-1b (Glycogen storage diseases; OMIM 232220), which is a defect in the glucose-6-phosphate translocase (transporter) enzyme (Choi *et al.*, 2017; Mameesh *et al.*, 2017). The deletion of G6PT2 can cause failure to thrive, doll-like facial appearance, short stature, protuberant abdomen, relatively thin extremities, and hepatomegaly (Mameesh *et al.*, 2017). Here, we can observe that the symptoms failure to thrive and short stature also was reported on 22q11 syndrome (Ryan *et al.*, 1997; Habel *et al.*, 2012; Bossi *et al.*, 2016).

The clinical signs short stature and failure to thrive has been associated with 22q11DS and studies shows that frequency of children with weight below the 50th centile was 83% and other study shows that 10% of them had short stature (Ryan *et al.*, 1997; Habel *et al.*, 2012; Bossi *et al.*, 2016). Interestingly, a case diagnosed with 22q11DS that was investigated due to progressive growth failure and none of the characteristic clinical features of the syndrome was already reported (Bossi *et al.*, 2016).

In contrast, the proteins J3KNL2 and NUB1 had a betweenness of 0 in the GN, but betweenness value increased in PN. The J3KNL2 (Septin-1) have unknown functions related to the 22q11DS. NUB1 (Nedd8 ultimate buster 1) is an adaptor protein which was associated with the function of specific down-regulator of the NEDD8 conjugation system and the recruitment of NEDD8, UBD, and their conjugates to the proteasome for degradation (Soubeyran *et al.*, 2017). This gene suppresses the formation of Lewy body-like inclusions and has been implicated in Huntington disease (Soubeyran *et al.*, 2017). Huntington disease has a progressive course and exhibits a combination of motor, cognitive and behavioral features

(Bates *et al.*, 2015). Clinical signs of motor, cognitive and behavioral impairment have been reported in cases of 22q11DS. However that is no studies proving the role of NUB1 in the onset of these features in either of this diseases.

Besides the genes previous discussed that presented a null value of betweenness, 361 proteins suffer great modification from global network to patient network. From the 361, 112 proteins had involvement on a disease according to UniProt (Consortium, 2019) and a list of these proteins can be found in Supplementary Material 3.

The genes *ZNF74*, *SNAP29*, *TRMT2A*, *LZTR1*, *TBX1*, *THAP7*, *GDCR6*, *CLDN5*, *SERPIND1*, *TSSK2*, *MED15*, *CRKL*, and *CLTCL1* were enriched with GO terms by SAFE analyses. The gene *ZNF74* (zinc finger protein 74) lie within LCR22B-LCR22C region, and its expression was detected in the pulmonary artery wall and in the aorta valve (Vago *et al.*, 2012; Woodward *et al.*, 2019). In addition, the *ZNF74* genes were enriched with the GO term RNA polymerase II transcription factor activity; the same term is linked with *TBX1* gene an essential gene for the 22q11DS (Baldini, Fulcoli and Illingworth, 2017). Thus, the association with the same GO term may indicate that *ZNF74* gene could have a potential role in the physiopathology of the 22q11DS.

The *SNAP29*, *TRMT2A*, *LZTR1* genes were enriched with similar GO biological process in which it can be highlighted the ontology terms of cell proliferation, regulation of hematopoietic stem cell differentiation, regulation of gene expression and regulation of transcription from RNA polymerase II promoter. The *SNAP29* (Synaptosomal-associated protein 29) gene has a role in the process of membrane fusion during intracellular trafficking process which is vital for the regulation of cell division (Vaccari *et al.*, 2019). There is not a lot of studies concerning *TRMT2A* gene (see citation Hicks *et al.*, 2010; Chang *et al.*, 2019), however, there is evidence that it has an inhibitory effect on cell proliferation and could be a cell cycle regulator (Chang *et al.*, 2019). The *LZTR1* gene is associated with the Noonan syndrome, schwannomatosis, and cancer, but there are no reports of its role in biological processes such as cell cycle regulation. Furthermore, *LZTR1* mutations do not seem to affect protein stability significantly (Motta et al., 2019). It was not possible to associate all the ontological terms of *SNAP29*, *TRMT2A*, *LZTR1* with reported functions already published, on the other hand, it does not mean that these genes do not have functions associated with the ontological terms or with the development of 22q11DS.

The gene *TBX1* encodes a T-box-containing transcription factor that belongs to a large family of transcription factors and is the most important gene for 22q11DS (Scambler, 2010; Gao, Li and Amendt, 2013). *TBX1* was enriched with the ontological terms Core promoter

proximal region sequence-specific DNA binding, RNA polymerase II transcription factor activity and Protein dimerization activity. The terms associated with *TBX1* gene is consistent with the literature, which it was shown that the DNA-binding motif of T-box proteins binds DNA in a sequence-specific manner. Furthermore, the Tbx1 interacts with Ash2l, a core component of a multimeric histone methyltransferase complex, which is required for methylation of histone lysine residues (Stoller *et al.*, 2011; Baldini, Fulcoli and Illingworth, 2017). It was also shown the ability of Tbx1 to interact at the chromatin level with the MLL3 complex and co-localization with the histone demethylase Lsd1 (Fulcoli *et al.*, 2016; Baldini, Fulcoli and Illingworth, 2017).

Studies show that loss of function of *TBX1* has broad consequences, especially in the development of the pharyngeal apparatus, which has an impact in the development of the heart, aortic arch and some great arteries, thymus, parathyroid and thyroid development (Fulcoli *et al.*, 2016). Thus, we can state that both literature and ontological terms show that the *TBX1* gene has an important regulatory role and that this gene is a dominant candidate gene for 22q11.2 deletion syndrome. Therefore, we suppose that genes dependent on the activation of *TBX1* may play a fundamental role in embryonic development.

THAP7 (THAP domain containing 7) gene encodes a chromatin-associated, histone tail binding protein (Ophoff *et al.*, 2016). Studies demonstrate that *THAP1* has zinc-dependent site-specific *in vitro* DNA-binding activity (Macfarlan *et al.*, 2005). Several ontological terms were associated with *THAP7* where it can be highlighted the cell division, cell cycle, and regulation of centriole replication. The previous ontological terms were also associated with *DGCR6*. The gene *DGCR6* (DiGeorge critical region gene 6) is expressed in neural crest cell migration, pharyngeal arch development, and in the regulation of other genes implicated in 22q11DS like *TBX1* (Hooper *et al.*, 2012). Although *DGCR6* exact function has not been clearly, due to the sites of expression, it is thought that this gene plays a role in the etiology of 22q11DS (Edelmann *et al.*, 2001). In addition, it was demonstrated that function pharyngeal arch development, a key in the process of 22q11DS, is associated with the *TBX1* gene. Therefore, *DGCR6* may also be responsible for the devolvement of conotruncal heart defects in cases with 22q11del (Gao *et al.*, 2015).

Another GO enriched gene was *CLDN5* (claudin-5). Claudin-5 is a small protein, and it is the principal claudin in brain endothelial cells and has an vital role in the modulation of the human blood-brain Barrier (BBB) (Du *et al.*, 2017; Ma *et al.*, 2017, 2018). The GO terms that *CLDN5* was enriched with were Calcium-independent cell-cell adhesion via plasma membrane and Cell-adhesion molecules transmission of nerve impulse. These terms are

consistent with the literature that report the importance of *CLDN5* for the organization of tight junctions and the maintenance of brain microvascular endothelial cells integrity (Du *et al.*, 2017; Ma *et al.*, 2017, 2018). Besides, *CLDN5* has a potential role in pathologies such as Alzheimer's disease, edema, hypoglycemia, inflammation, toxic damage, trauma, and tumors (Ma *et al.*, 2017).

Serpin peptidase inhibitor (*SERPIND1*) was enriched with the GO terms Blood coagulation, which is a function consistent with literature findings (Giuliani *et al.*, 2016; Li *et al.*, 2018; Woodward *et al.*, 2019). *SERPIND1*, also known as heparin cofactor II (HCII or HC2), is a member of the family of serine protease inhibitors that acts as a thrombin inhibitor through interactions with heparin and other endogenous glycosaminoglycans (Li *et al.*, 2018; Sisak *et al.*, 2018). Moreover, *SERPIND1* is a gene within the nested region between LCR22B and LCR22D which was found to be more expressed in the nervous system (Woodward *et al.*, 2019). This gene is associated with disseminated intravascular coagulation, inflammatory diseases and was proposed to promote angiogenesis in response to ischemia (Zhu *et al.*, 2016; Li *et al.*, 2018; Woodward *et al.*, 2019).

The GO terms Protein autophosphorylation and phosphorylation were linked with the genes *TSSK2* and *MED15*. The gene *TSSK2* (testis-specific serine/threonine kinase 2) is expressed in both human and mouse sperm, and studies described the ability of this protein to phosphorylates other proteins and to display robust autophosphorylation (Hawkinson *et al.*, 2017). In addition, polymorphisms of the *TSSK2* gene has been associated with spermatogenesis impairment which means that this gene plays an essential role in spermiogenesis and fertilization (Shetty *et al.*, 2016; Hawkinson *et al.*, 2017). The gene *MED15* has not been associated with spermiogenesis however is a gene that belongs to a multiprotein mediator complex necessary for the assembly of the RNA polymerase II complex, thereby regulating the Pol II-dependent transcription (Methylation *et al.*, 2014; Syring *et al.*, 2018; Weiten *et al.*, 2018).

Few studies have reported cases of 22q11DS associated with infertility. One study reported a gonosomal mosaic of chromosome 22q11 deletion in a patient with infertility without any severe clinical manifestations, except for oligoasthenozoospermia, mild bradycardia, and mild tricuspid regurgitation (Liu *et al.*, 2018). Another case reports a male patient with mild dysmorphic features, hypernasal voice, mental retardation, and azoospermia (Ozcan and Sahin, 2017). Moreover, a case of supernumerary inv dup(22)(q11.1) with infertility with hypogonadotropic hypogonadism has been reported (Mikelsaar, Lissitsina, and Bartsch, 2011). Since cases of 22q11DS present a broad phenotypic spectrum where many

patients have mild phenotypes (Ozcan and Sahin, 2017), an evaluation of a large group of cases with 22q11DS could be interesting to evaluate a possible association of male infertility, *TSSK2* gene, and 22q11DS.

The *CRKL* was the gene enriched with the most GO terms (27 terms). The gene *CRKL* together with *TBX1* and *MAPK1* is related to the etiopathogenesis of heart defects on the 22q11DS, and studies shows that these genes are involved in the regulation on heart outflow tract morphogenesis in mouse (Guris *et al.*, 2001; Lindsay, 2001; Moon *et al.*, 2006; Breckpot *et al.*, 2012; Bengoa-Alonso *et al.*, 2016). The ontological terms that can be associated with the pathologies previous cited are Vascular endothelial growth factor receptor signaling pathway. Furthermore, studies show that *CRKL* encodes an adapter protein that regulates intracellular signaling transduction from multiple growth factors (Kiryluk *et al.*, 2017) which could be associated with regulation of cell proliferation, regulation of cell migration and proliferation.

The gene *CLTCL1* was the last enriched by SAFE analyses, the GO terms linked with the genes are related to proteins transport, especially the retrograde transport involving the Golgi complex. The gene *CLTCL1*, also known as clathrin gene, encodes the minor clathrin heavy chain (CHC22) a protein that the literature believes to be involved in intracellular endosome trafficking (Nahorski *et al.*, 2015). However, an association of the *CLTCL1* function and the phenotype of 22q11DS has not yet been reported.

5. Conclusion

The results found in this work indicates that deletion of 48 nodes equivalent to the genes within the 22q11 region does not imply in drastic alterations in the human PPI network. However, when the neighboring proteins were analysed within the communities, we observed that the degree of the neighboring proteins is increased. Therefore, we can hypothesize two consequences of this change: (1) due to 22q11 deletion genes, the neighboring proteins of this network increased their interactions in order to maintain network functions; (2) the relevance of these nodes in the network may be increasing, indicating that mutation or deletion of the genes encoding these proteins may cause more relevant disturbances in the network structure.

Our results indicated some genotype-phenotype relationship of 22q11DS with the genes analyzed through the PPI network. From the analyses of the neighboring proteins, we found that the gene *G6PT2* lost its betweenness in the PN and that symptoms failure to thrive and short stature caused by this gene has also reported on cases of 22q11DS. Furthermore, the clinical signs of motor, cognitive and behavioral impairment have been reported in Huntington

disease and 22q11DS, which the gene *NUB1* has been implicated in both disorders. Thus, an analysis of the list of the 361 neighboring proteins could be used as targets for future studies as they can have a potential role in the physiopathology of the syndrome.

The results found through the GO enrichment analysis shows that GO terms linked to the genes agree with information from the literature. In this way, analysis using ontological terms can bring significant data to contribute to the construction of research projects to identify new potential disease genes or hypothesize possible functions of a gene. We highlight the association of *TSSK2* gene, which is in the 3Mb 22q11 deletion region, with infertility. Thus, we suggest an evaluation of a large group of cases with 22q11DS in an attempt to check the possible association of 22q11DS with male infertility.

Finally, the network biology approach has a great potential to discover properties of different pathologies that are not readily accessible by conventional molecular biology or genetic approaches. Moreover, analyses evaluating another issue of the deletion, such as studies with regulatory and/or gene expression networks, could complement our study and provide some insight into the broad phenotypic spectrum of 22q11DS.

6. References

Albert, R., Jeong, H. and Barabási, A.-L. (2000) 'Error and attack tolerance of complex networks. Nature, 406: 378–482, 2000', *Nature*, 406(6794), pp. 378–382. doi: 10.1038/35019019.

Ashburner, M. *et al.* (2000) 'Gene Ontology: tool for the unification of biology', *Nature Genetics*, 25(1), pp. 25–29. doi: 10.1038/75556.

Ashley, E. A. *et al.* (2006) 'Network analysis of human in-stent restenosis', *Circulation*, 114(24), pp. 2644–2654. doi: 10.1161/CIRCULATIONAHA.106.637025.

Baldini, A., Fulcoli, F. G. and Illingworth, E. (2017) *Tbx1: Transcriptional and Developmental Functions*. 1st edn, *Current Topics in Developmental Biology*. 1st edn. Elsevier Inc. doi: 10.1016/bs.ctdb.2016.08.002.

Barabási, A. L. and Oltvai, Z. N. (2004a) 'Network biology: Understanding the cell's functional organization', *Nature Reviews Genetics*, 5(2), pp. 101–113. doi: 10.1038/nrg1272.

Barabási, A. L. and Oltvai, Z. N. (2004b) 'Network biology: Understanding the cell's functional organization', *Nature Reviews Genetics*, 5(2), pp. 101–113. doi: 10.1038/nrg1272.

Barabasi and Albert (1999) 'Emergence of scaling in random networks', *Science (New York, N.Y.)*, 286(5439), pp. 509–12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10521342.

Baryshnikova, A. (2016a) 'Spatial Analysis of Functional Enrichment (SAFE) in Large

Biological Networks', *bioRxiv*, p. 094904. doi: 10.1101/094904.

Baryshnikova, A. (2016b) 'Systematic Functional Annotation and Visualization of Biological Networks', *Cell Systems*. The Author(s), 2(6), pp. 412–421. doi: 10.1016/j.cels.2016.04.014.

Bassett, A. S. *et al.* (2011) 'Practical guidelines for managing patients with 22q11.2 deletion syndrome', *Journal of Pediatrics*. Mosby, Inc., 159(2), p. 332–339.e1. doi: 10.1016/j.jpeds.2011.02.039.

Bates, G. P. *et al.* (2015) 'Huntington disease', *Nat Rev Dis Primers*, 1(April), pp. 1–21. doi: 10.1038/nrdp.2015.5.

Bengoa-Alonso, A. *et al.* (2016) 'Delineation of a recognizable phenotype for the recurrent LCR22-C to D/E atypical 22q11.2 deletion', *American Journal of Medical Genetics, Part A*, 170(6), pp. 1485–1494. doi: 10.1002/ajmg.a.37614.

Bertini, V. *et al.* (2017) 'Deletion Extents Are Not the Cause of Clinical Variability in 22q11 . 2 Deletion Syndrome : Does the Interaction between DGCR8 and miRNA-CNVs Play a Major Role ?', 8(May). doi: 10.3389/fgene.2017.00047.

Blom, U. M. *et al.* (2011) 'Prioritizing candidate disease genes by network-based boosting of genome-wide association data', *Genome Research*, 21(7), pp. 1109–1121. doi: 10.1101/gr.118992.110.

Bossi, G. *et al.* (2016) 'Failure to thrive as presentation in a patient with 22q11.2 microdeletion', *Italian Journal of Pediatrics*. Italian Journal of Pediatrics, 42(1), pp. 1–4. doi: 10.1186/s13052-016-0224-0.

Breckpot, J. *et al.* (2012) 'Congenital heart defects in a novel recurrent 22q11.2 deletion harboring the genes CRKL and MAPK1', *American Journal of Medical Genetics, Part A*, 158 A(3), pp. 574–580. doi: 10.1002/ajmg.a.35217.

Burnside, R. D. (2015) '22q11.21 deletion syndromes: A review of proximal, central, and distal deletions and their associated features', *Cytogenetic and Genome Research*, 146(2), pp. 89–99. doi: 10.1159/000438708.

Chan, S. Y. and Loscalzo, J. (2012) 'The Emerging Paradigm of Network Medicine in the Study of Human Disease', *Circulation Research*, 111(3), pp. 359–374. doi: 10.1161/CIRCRESAHA.111.258541.

Chang, Y. H. *et al.* (2019) 'TRMT2A is a novel cell cycle regulator that suppresses cell proliferation', *Biochemical and Biophysical Research Communications*. Elsevier Ltd, 508(2), pp. 410–415. doi: 10.1016/j.bbrc.2018.11.104.

Chen, J. *et al.* (2016) 'Identification of a Novel ENU-Induced Mutation in Mouse Tbx1 Linked to Human DiGeorge Syndrome', *Neural Plasticity*. Hindawi Publishing Corporation, 2016. doi: 10.1155/2016/5836143.

Clauset, A., Newman, M. E. J. and Moore, C. (2004) 'Finding community structure in very large networks', *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related*

Interdisciplinary Topics, 70(6), p. 6. doi: 10.1103/PhysRevE.70.066111.

Consortium, T. U. (2019) 'UniProt: a worldwide hub of protein knowledge', *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D506–D515. doi: 10.1093/nar/gky1049.

Csárdi, G. and Nepusz, T. (2006) 'The igraph software package for complex network research', *InterJournal Complex Systems*, 1695, pp. 1–9. doi: 10.3724/SP.J.1087.2009.02191.

Diez, D. *et al.* (2010) 'The use of network analyses for elucidating mechanisms in cardiovascular disease', *Molecular BioSystems*, 6(2), pp. 289–304. doi: 10.1039/b912078e.

Du, Y. *et al.* (2017) 'Increased cerebral expressions of MMPs, CLDN5, OCLN, ZO1 and AQPs are associated with brain edema following fatal heat stroke', *Scientific Reports*, 7(1), pp. 1–10. doi: 10.1038/s41598-017-01923-w.

Dugoff, L., Mennuti, M. T. and McDonald-McGinn, D. M. (2017) 'The benefits and limitations of cell-free DNA screening for 22q11.2 deletion syndrome', *Prenatal Diagnosis*, 37(1), pp. 53–60. doi: 10.1002/pd.4864.

Edelmann, L. *et al.* (2001) 'Two functional copies of the DGCR6 gene are present on human chromosome 22q11 due to a duplication of an ancestral locus', *Genome Research*, 11(2), pp. 208–217. doi: 10.1101/gr.GR-1431R.

Edelmann, L., Pandita, R. K. and Morrow, B. E. (1999) 'Low-Copy Repeats Mediate the Common 3-Mb Deletion in Patients with Velo-cardio-facial Syndrome', *The American Journal of Human Genetics*, 64(4), pp. 1076–1086. doi: 10.1086/302343.

Friedel, C. C. and Zimmer, R. (2007) 'Influence of degree correlations on network structure and stability in protein-protein interaction networks', *BMC Bioinformatics*, 8, pp. 1–10. doi: 10.1186/1471-2105-8-297.

Fulcoli, F. G. *et al.* (2016) 'Rebalancing gene haploinsufficiency in vivo by targeting chromatin', *Nature Communications*. Nature Publishing Group, 7, pp. 1–11. doi: 10.1038/ncomms11688.

Gao, S., Li, X. and Amendt, B. A. (2013) 'Understanding the role of Tbx1 as a candidate gene for 22q11.2 deletion syndrome', *Current Allergy and Asthma Reports*, 13(6), pp. 613–621. doi: 10.1007/s11882-013-0384-6.

Gao, W. *et al.* (2015) 'DGCR6 at the proximal part of the DiGeorge critical region is involved in conotruncal heart defects', *Human Genome Variation*. Nature Publishing Group, 2(1), pp. 1–7. doi: 10.1038/hgv.2015.4.

Gene, T. and Consortium, O. (2019) 'The Gene Ontology Resource: 20 years and still GOing strong', *Nucleic acids research*. Oxford University Press, 47(D1), pp. D330–D338. doi: 10.1093/nar/gky1055.

Giuliani, S. *et al.* (2016) 'Coagulation Gene Expression Profiling in Infants With Necrotizing Enterocolitis', *Journal of Pediatric Gastroenterology and Nutrition*, 63(6), pp. e169–e175. doi: 10.1097/MPG.0000000000001215.

Guris, D. L. *et al.* (2001) 'Mice lacking the homologue of the human 22q11.2 gene CRLK phenocopy neurocristopathies of DiGeorge syndrome', *Nature Genetics*, 27(3), pp. 293–298. doi: 10.1038/85855.

Habel, A. *et al.* (2012) 'Syndrome-specific growth charts for 22q11.2 deletion syndrome in Caucasian children', *American Journal of Medical Genetics, Part A*, 158 A(11), pp. 2665–2671. doi: 10.1002/ajmg.a.35426.

Hall, M. *et al.* (2009) 'The WEKA data mining software', *SIGKDD Explorations Newsletter*, 11(1), p. 10. doi: 10.1145/1656274.1656278.

Han, J. D. J. (2008) 'Understanding biological functions through molecular networks', *Cell Research*, 18(2), pp. 224–237. doi: 10.1038/cr.2008.16.

Hawkinson, J. E. *et al.* (2017) 'Potent Pyrimidine and Pyrrolopyrimidine Inhibitors of Testis-Specific Serine/Threonine Kinase 2 (TSSK2)', *ChemMedChem*, 12(22), pp. 1857–1865. doi: 10.1002/cmdc.201700503.

Hicks, D. G. *et al.* (2010) 'The expression of TRMT2A, a novel cell cycle regulated protein, identifies a subset of breast cancer patients with HER2 over-expression that are at an increased risk of recurrence', *BMC Cancer*, 10. doi: 10.1186/1471-2407-10-108.

Hooper, S. R. *et al.* (2012) 'Dysregulation of DGCR6 and DGCR6L: psychopathological outcomes in chromosome 22q11.2 deletion syndrome', *Translational Psychiatry*, 2(4), pp. e105–e105. doi: 10.1038/tp.2012.31.

Jensen, M. K. *et al.* (2011) 'Protein Interaction-Based Genome-Wide Analysis of Incident Coronary Heart Disease', *Circulation: Cardiovascular Genetics*, 4(5), pp. 549–556. doi: 10.1161/CIRCGENETICS.111.960393.

Jerome, L. A. and Papaioannou, V. E. (2001) 'DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1', *Nature Genetics*, 27(3), pp. 286–291. doi: 10.1038/85845.

Ju, Z. R. *et al.* (2016) 'HIRA gene is lower expressed in the myocardium of patients with tetralogy of Fallot', *Chinese Medical Journal*, 129(20), pp. 2403–2408. doi: 10.4103/0366-6999.191745.

Kiryluk, K. *et al.* (2017) 'Genetic Drivers of Kidney Defects in the DiGeorge Syndrome', *New England Journal of Medicine*, 376(8), pp. 742–754. doi: 10.1056/nejmoa1609009.

Kolaczyk, E. D. and Csárdi, G. (2014) *Statistical Analysis of Network Data with R*, *International Statistical Review*. New York, NY: Springer New York (Use R!). doi: 10.1007/978-1-4939-0983-4.

Li, C. *et al.* (2018) 'An analysis of plasma reveals proteins in the acute phase response pathway to be candidate diagnostic biomarkers for depression', *Psychiatry Research*, 272(November 2016), pp. 404–410. doi: 10.1016/j.psychres.2018.11.069.

Lindsay, E. A. (2001) 'Chromosomal microdeletions: Dissecting DEL22Q11 syndrome', *Nature Reviews Genetics*, 2(11), pp. 858–868. doi: 10.1038/35098574.

Liu, Y. *et al.* (2018) 'Infertility in a man with oligoasthenozoospermia associated with mosaic chromosome 22q11 deletion', *Molecular Genetics and Genomic Medicine*, 6(6), pp. 1249–1254. doi: 10.1002/mgg3.487.

Lui, L. T. *et al.* (2017) 'Characterization of the Molecular Mechanisms Underlying the Chronic Phase of Stroke in a Cynomolgus Monkey Model of Induced Cerebral Ischemia', *Journal of Proteome Research*, 16(3), pp. 1150–1166. doi: 10.1021/acs.jproteome.6b00651.

Ma, S. C. *et al.* (2017) 'Claudin-5 regulates blood-brain barrier permeability by modifying brain microvascular endothelial cell proliferation, migration, and adhesion to prevent lung cancer metastasis', *CNS Neuroscience and Therapeutics*, 23(12), pp. 947–960. doi: 10.1111/cns.12764.

Ma, S. C. *et al.* (2018) 'CLDN5 affects IncRNAs acting as ceRNA dynamics contributing to regulating blood-brain barrier permeability in tumor brain metastasis', *Oncology Reports*, 39(3), pp. 1441–1453. doi: 10.3892/or.2018.6208.

Macfarlan, T. *et al.* (2005) 'Human THAP7 is a chromatin-associated, histone tail-binding protein that represses transcription via recruitment of HDAC3 and nuclear hormone receptor corepressor', *Journal of Biological Chemistry*, 280(8), pp. 7346–7358. doi: 10.1074/jbc.M411675200.

McDonald-McGinn, D. M. *et al.* (2015) '22Q11.2 Deletion Syndrome', *Nature Reviews Disease Primers*, 1(November). doi: 10.1038/nrdp.2015.71.

Methylation, D. N. a *et al.* (2014) 'Biomarker Insights Gene as a Biomarker for Head and Neck Cancers', *Biomarker insights*, 9, pp. 53–60. doi: 10.4137/BMI.S16199.Received.

Michaelovsky, E. *et al.* (2019) 'Risk gene-set and pathways in 22q11.2 deletion-related schizophrenia: a genealogical molecular approach', *Translational Psychiatry*. Springer US, 9(1). doi: 10.1038/s41398-018-0354-9.

Mikelsaar, R., Lissitsina, J. and Bartsch, O. (2011) 'Small supernumerary marker chromosome (sSMC) derived from chromosome 22 in an infertile man with hypogonadotropic hypogonadism', *Journal of Applied Genetics*, 52(3), pp. 331–334. doi: 10.1007/s13353-011-0041-5.

Moon, A. M. *et al.* (2006) 'Crkl Deficiency Disrupts Fgf8 Signaling in a Mouse Model of 22q11 Deletion Syndromes', *Developmental Cell*, 10(1), pp. 71–80. doi: 10.1016/j.devcel.2005.12.003.

Morrow, B. E. *et al.* (2018) 'Molecular genetics of 22q11.2 deletion syndrome', *American Journal of Medical Genetics Part A*, 176(10), pp. 2070–2081. doi: 10.1002/ajmg.a.40504.

Morrow, B. E. *et al.* (2018) 'Molecular genetics of 22q11.2 deletion syndrome', *American Journal of Medical Genetics, Part A*, 176(10), pp. 2070–2081. doi: 10.1002/ajmg.a.40504.

Motta, M. *et al.* (2019) 'Dominant Noonan syndrome-causing LZTR1 mutations specifically affect the Kelch domain substrate-recognition surface and enhance RAS-MAPK signaling', *Human Molecular Genetics*, 28(6), pp. 1007–1022. doi: 10.1093/hmg/ddy412.

Nahorski, M. S. *et al.* (2015) 'A novel disorder reveals clathrin heavy chain-22 is essential for human pain and touch development', *Brain*, 138(8), pp. 2147–2160. doi: 10.1093/brain/awv149.

Ophoff, R. A. *et al.* (2016) 'Peripheral blood gene expression profiles linked to monoamine metabolite levels in cerebrospinal fluid', *Translational Psychiatry*. Nature Publishing Group, 6(12), pp. e983–e983. doi: 10.1038/tp.2016.245.

Ozcan, A. and Sahin, Y. (2017) 'DiGeorge Syndrome Associated with Azoospermia: First case in the literature', *Türk Üroloji Dergisi/Turkish Journal of Urology*, 43(3), pp. 390–392. doi: 10.5152/tud.2017.08555.

Panamonta, V. *et al.* (2016) 'Birth Prevalence of Chromosome 22q11.2 Deletion Syndrome: A Systematic Review of Population-Based Studies.', *Journal of the Medical Association of Thailand = Chotmaihet thangphaet*, 99 Suppl 5(18), pp. S187-93. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29906080.

Pržulj, N., Wigle, D. A. and Jurisica, I. (2004) 'Functional topology in a network of protein interactions', *Bioinformatics*, 20(3), pp. 340–348. doi: 10.1093/bioinformatics/btg415.

Racedo, S. E. *et al.* (2015) 'Mouse and human CRKL is dosage sensitive for cardiac outflow tract formation', *American Journal of Human Genetics*. The American Society of Human Genetics, 96(2), pp. 235–244. doi: 10.1016/j.ajhg.2014.12.025.

Raman, K. (2010) 'Construction and analysis of protein–protein interaction networks', *Automated Experimentation*, 2(1), p. 2. doi: 10.1186/1759-4499-2-2.

Rolland T. Tasan M., C. B. P. S. Z. Q. et al. (2014) 'A proteome-scale map of the human interactome network', *Cell*, 159(5), pp. 1213–1226. doi: 10.1016/j.cell.2014.10.050.A.

Rosa, R. F. M. *et al.* (2009) 'Síndrome de deleção 22q11.2: compreendendo o CATCH22', *Revista Paulista de Pediatria*, 27(2), pp. 211–220. doi: 10.1590/S0103-05822009000200015.

Rual, J. F. *et al.* (2005) 'Towards a proteome-scale map of the human protein-protein interaction network', *Nature*, 437(7062), pp. 1173–1178. doi: 10.1038/nature04209.

Ryan, A. K. *et al.* (1997) 'Spectrum of clinical features associated with interstitial chromosome 22q11 deletions: a European collaborative study.', *Journal of Medical Genetics*, 34(10), pp. 798–804. doi: 10.1136/jmg.34.10.798.

Santoro, M. L. *et al.* (2019) 'Downregulation of genes outside the deleted region in individuals with 22q11.2 deletion syndrome', *Human Genetics*. Springer Berlin Heidelberg, 138(1), pp. 93–103. doi: 10.1007/s00439-018-01967-6.

Scambler, P. J. (2000) 'The 22q11 deletion syndromes', *Human Molecular Genetics*, 9(16), pp. 2421–2426. doi: 10.1093/hmg/9.16.2421.

Scambler, P. J. (2010) '22q11 Deletion syndrome: A role for TBX1 in pharyngeal and cardiovascular development', *Pediatric Cardiology*, 31(3), pp. 378–390. doi: 10.1007/s00246-009-9613-0.

Schmith, J. *et al.* (2005) 'Damage, connectivity and essentiality in protein-protein interaction networks', *Physica A: Statistical Mechanics and its Applications*, 349(3–4), pp. 675–684. doi: 10.1016/j.physa.2004.10.038.

Sengupta, U. *et al.* (2009) 'Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications', *PLoS ONE*, 4(12). doi: 10.1371/journal.pone.0008100.

Shetty, J. *et al.* (2016) 'Recombinant production of enzymatically active male contraceptive drug target hTSSK2 - Localization of the TSKS domain phosphorylated by TSSK2', *Protein Expression and Purification*, 121(3), pp. 88–96. doi: 10.1016/j.pep.2016.01.009.

Sisak, F. *et al.* (2018) 'Protein expression in the liver and blood serum in chickens in response to Salmonella Enteritidis infection', *Veterinary Immunology and Immunopathology*. Elsevier B.V., 205, pp. 10–16. doi: 10.1016/j.vetimm.2018.10.006.

Soubeyran, P. *et al.* (2017) 'Regulation of NUB1 Activity through Non-Proteolytic Mdm2-Mediated Ubiquitination', *Plos One*, 12(1), p. e0169988. doi: 10.1371/journal.pone.0169988.

Stark, C. (2005) 'BioGRID: a general repository for interaction datasets', *Nucleic Acids Research*, 34(90001), pp. D535–D539. doi: 10.1093/nar/gkj109.

Stoller, J. Z. *et al.* (2011) 'Ash2l interacts with Tbx1 and is required during early embryogenesis', *Exp Biol Med*, 235(5), pp. 569–576. doi: 10.1258/ebm.2010.009318.

Sullivan, K. E. (2019) 'Chromosome 22q11.2 deletion syndrome and DiGeorge syndrome', *Immunological Reviews*, 287(1), pp. 186–201. doi: 10.1111/imr.12701.

Sullivan, K. E. (2019) 'Chromosome 22q11.2 deletion syndrome and DiGeorge syndrome', *Immunological Reviews*, 287(1), pp. 186–201. doi: 10.1111/imr.12701.

Syring, I. *et al.* (2018) 'The knockdown of the Mediator complex subunit MED15 restrains urothelial bladder cancer cells' malignancy', *Oncology Letters*, 16(3), pp. 3013–3021. doi: 10.3892/ol.2018.9014.

Vaccari, T. *et al.* (2019) 'A genetic model of CEDNIK syndrome in zebrafish highlights the role of the SNARE protein Snap29 in neuromotor and epidermal development', *Scientific Reports*, 9(1), pp. 1–13. doi: 10.1038/s41598-018-37780-4.

Vago, P. *et al.* (2012) 'An atypical 0.8 Mb inherited duplication of 22q11.2 associated with psychomotor impairment', *European Journal of Medical Genetics*. Elsevier Masson SAS, 55(11), pp. 650–655. doi: 10.1016/j.ejmg.2012.06.014.

Veres, D. V. *et al.* (2015) 'ComPPI: A cellular compartment-specific database for protein-protein interaction network analysis', *Nucleic Acids Research*, 43(D1), pp. D485–D493. doi: 10.1093/nar/gku1007.

Wang, X., Gulbahce, N. and Yu, H. (2011) 'Network-based methods for human disease gene prediction', *Briefings in Functional Genomics*, 10(5), pp. 280–293. doi: 10.1093/bfgp/elr024.

Weiten, R. *et al.* (2018) 'The Mediator complex subunit MED15, a promoter of tumour progression and metastatic spread in renal cell carcinoma', *Cancer Biomarkers*, 21(4), pp. 839–847. doi: 10.3233/CBM-170757.

Witten, I. H. , Frank, E., & Hall, M. A. (2011) *Data Mining : Practical Machine Learning Tools and Techniques*, *Morgan Kaufmann Publishers*. doi: 10.1016/C2009-0-19715-5.

Woodward, K. J. *et al.* (2019) 'Atypical nested 22q11.2 duplications between LCR22B and LCR22D are associated with neurodevelopmental phenotypes including autism spectrum disorder with incomplete penetrance', *Molecular Genetics and Genomic Medicine*, (September 2018), pp. 1–17. doi: 10.1002/mgg3.507.

Yamagishi, H. and Srivastava, D. (2003) 'Unraveling the genetic and developmental mysteries of 22q11 deletion syndrome', *Trends in Molecular Medicine*, 9(9), pp. 383–389. doi: 10.1016/S1471-4914(03)00141-2.

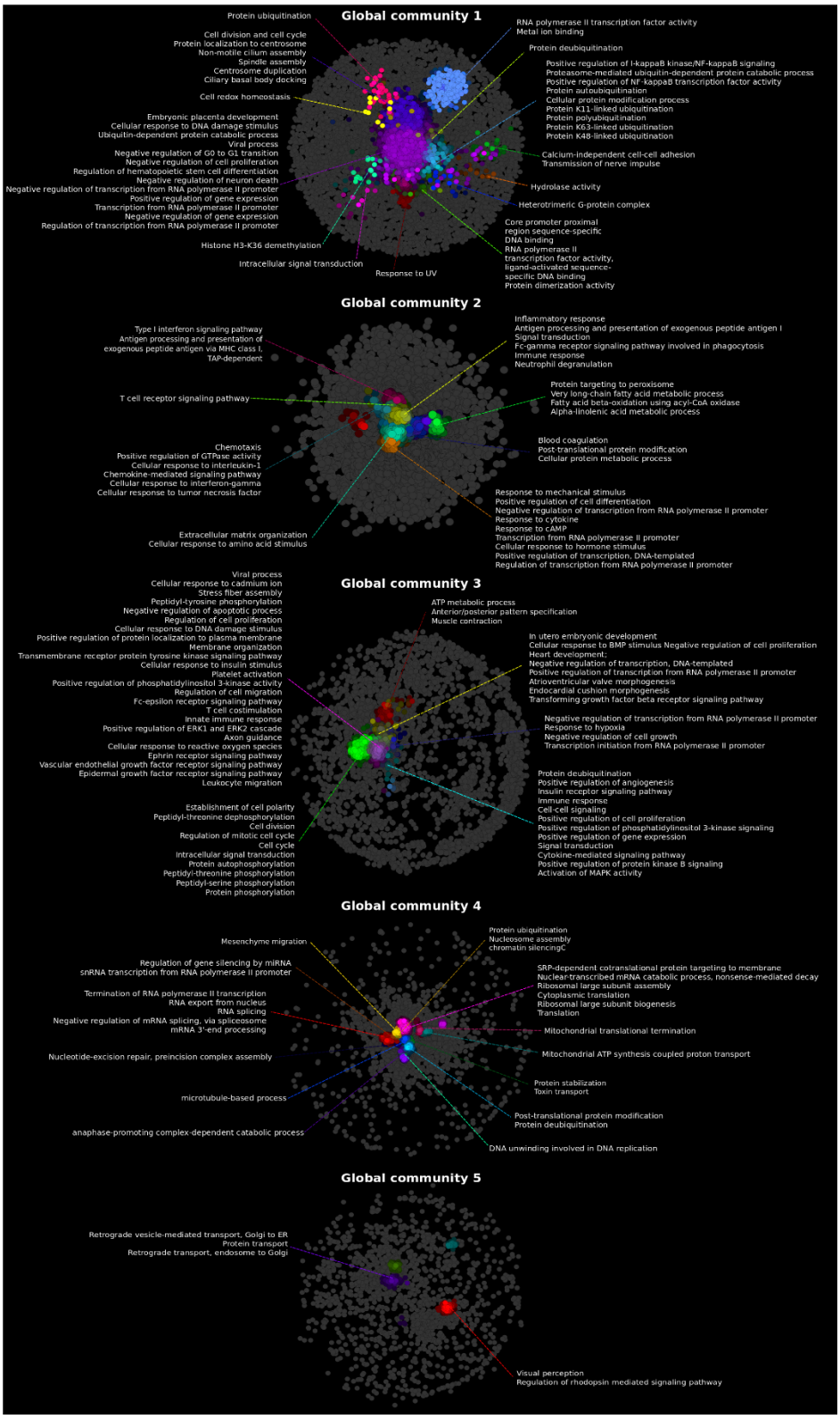
Yang, J.-H. *et al.* (2016) 'Differential regulation of the histone chaperone HIRA during muscle cell differentiation by a phosphorylation switch', *Experimental & Molecular Medicine*. Nature Publishing Group, 48(8), pp. e252–e252. doi: 10.1038/emm.2016.68.

Zeitz, M. J. *et al.* (2013) 'Implications of COMT long-range interactions on the phenotypic variability of 22q11. 2 deletion syndrome', *Nucleus*, 4(6), pp. 6–7. doi: 10.4161/nucl.27364.

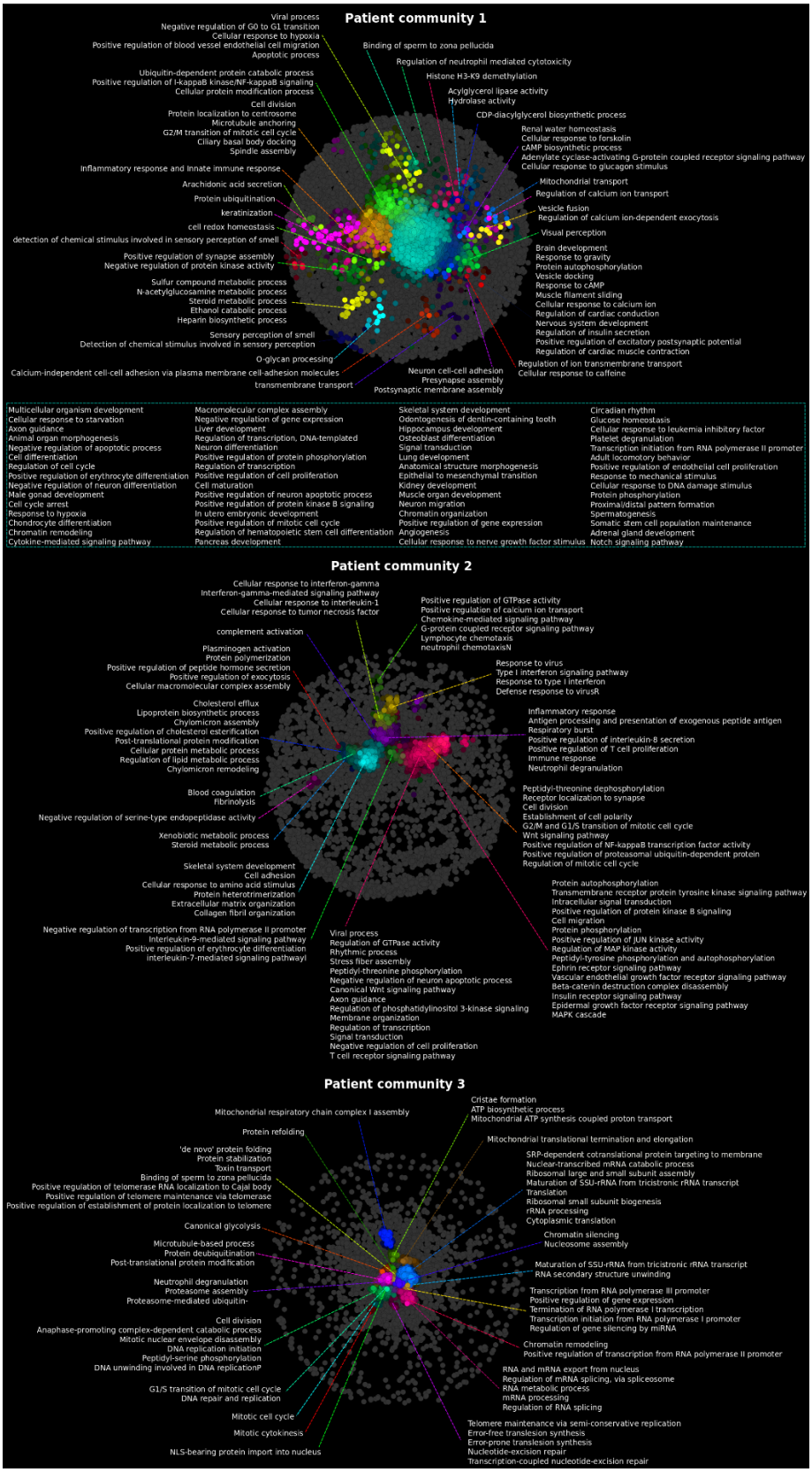
Zhu, L. *et al.* (2016) 'Analysis of the gene expression profile in response to human epididymis protein 4 in epithelial ovarian cancer cells', *Oncology Reports*, 36(3), pp. 1592–1604. doi: 10.3892/or.2016.4926.

7. Supplementary materials

7.1 Supplementary material 1 - Global communities that were enriched by Gene Ontology Enrichment Analysis using SAFE. Network regions enriched for similar GO biological process terms are color-coded, and the GO terms can be observed.



7.2 Supplementary material 2 - Patient communities that were enriched by Gene Ontology Enrichment Analysis using SAFE. Network regions enriched for similar GO biological process terms are color-coded, and the GO terms can be observed.



7. 3. Supplementary material 3 – Table with the neighboring proteins that has an association with diseases according to the UniProt database.

Entry name	Involvement in disease
2ABB	Spinocerebellar ataxia 12 (SCA12) [MIM:604326]: Spinocerebellar ataxia is a clinically and genetically heterogeneous group of cerebellar disorders. Patients show progressive incoordination of gait and often poor coordination of hands, speech and eye movements, due to degeneration of the cerebellum with variable involvement of the brainstem and spinal cord. SCA12 is an autosomal dominant cerebellar ataxia (ADCA).
ACTN1	Bleeding disorder, platelet-type 15 (BDPLT15) [MIM:615193]: An autosomal dominant form of macrothrombocytopenia. Affected individuals usually have no or only mild bleeding tendency, such as epistaxis. Laboratory studies show decreased numbers of large platelets and anisocytosis, but the platelets show no in vitro functional abnormalities.
ACTN2	Cardiomyopathy, familial hypertrophic 23, with or without left ventricular non-compaction (CMH23) [MIM:612158]: A hereditary heart disorder characterized by ventricular hypertrophy, which is usually asymmetric and often involves the interventricular septum. The symptoms include dyspnea, syncope, collapse, palpitations, and chest pain. They can be readily provoked by exercise. The disorder has inter- and intrafamilial variability ranging from benign to malignant forms with high risk of cardiac failure and sudden cardiac death.; Cardiomyopathy, dilated 1AA, with or without left ventricular non-compaction (CMD1AA) [MIM:612158]: A disorder characterized by ventricular dilation and impaired systolic function, resulting in congestive heart failure and arrhythmia. Patients are at risk of premature death.
ALG11	Congenital disorder of glycosylation 1P (CDG1P) [MIM:613661]: A form of congenital disorder of glycosylation, a multisystem disorder caused by a defect in glycoprotein biosynthesis and characterized by under-glycosylated serum glycoproteins. Congenital disorders of glycosylation result in a wide variety of clinical features, such as defects in the nervous system development, psychomotor retardation, dysmorphic features, hypotonia, coagulation disorders, and immunodeficiency. The broad spectrum of features reflects the critical role of N-glycoproteins during embryonic development, differentiation, and maintenance of cell functions.
ALG3	Congenital disorder of glycosylation 1D (CDG1D) [MIM:601110]: A form of congenital disorder of glycosylation, a multisystem disorder caused by a defect in glycoprotein biosynthesis and characterized by under-glycosylated serum glycoproteins. Congenital disorders of glycosylation result in a wide variety of clinical features, such as defects in the nervous system development, psychomotor retardation, dysmorphic features, hypotonia, coagulation disorders, and immunodeficiency. The broad spectrum of features reflects the critical role of N-glycoproteins during embryonic development, differentiation, and maintenance of cell functions.
AMMR1	Midface hypoplasia, hearing impairment, elliptocytosis, and nephrocalcinosis (MFHIEN) [MIM:300990]: An X-linked recessive disorder with onset in early childhood, characterized by midface hypoplasia, hearing impairment, elliptocytosis, and nephrocalcinosis. Variable clinical features include anemia, and mild early motor or speech delay.; Alport syndrome with mental retardation, midface hypoplasia and elliptocytosis (ATS-MR) [MIM:300194]: A X-linked contiguous gene deletion syndrome characterized by glomerulonephritis, sensorineural hearing loss, mental retardation, midface hypoplasia and elliptocytosis.
AP2S1	Hypocalciuric hypercalcemia, familial 3 (HHC3) [MIM:600740]: A form of hypocalciuric hypercalcemia, a disorder of mineral homeostasis that is transmitted as an autosomal dominant trait with a high degree of penetrance. It is characterized biochemically by lifelong elevation of serum calcium concentrations and is associated with inappropriately low urinary calcium excretion and a normal or mildly elevated circulating parathyroid hormone level. Hypermagnesemia is typically present. Affected individuals are usually asymptomatic and the disorder is considered benign. However, chondrocalcinosis and pancreatitis occur in some adults.
AT2C1	Hailey-Hailey disease (HHD) [MIM:169600]: Autosomal dominant disorder characterized by persistent blisters and suprabasal cell separation (acantholysis) of the epidermis, due to impaired keratinocyte adhesion. Patients lacking all isoforms except isoform 2 have HHD.
ATAD1	Hyperekplexia 4 (HKPX4) [MIM:618011]: An autosomal recessive severe neurologic disorder apparent from birth. HKPX4 is characterized by little if any development, hypertonia, early-onset refractory seizures in some patients, and respiratory failure resulting in early death, mostly in the first months of life.

ATL4	Ectopia lentis 2, isolated, autosomal recessive (ECTOL2) [MIM:225100]: An ocular abnormality characterized by partial or complete displacement of the lens from its space resulting from defective zonule formation.; Ectopia lentis et pupillae (ECTOLP) [MIM:225200]: An ocular abnormality characterized by displacement of the lenses and the pupils, associated with other ocular anomalies, but without systemic manifestations. The condition is usually bilateral, with the lenses and pupils displaced in opposite directions. Additional signs include enlarged corneal diameter, increased corneal astigmatism, increased anterior chamber depth, thinning and flattening of the iris with loss of crypts, angle malformation caused by enlarged iris processes, persistent pupillary membrane, loss of zonular fibers, tilted disk, and increased axial length. Secondary manifestations include refractive errors, glaucoma, early cataract development, and retinal detachment. Membrane formation on the posterior aspect of the iris has been observed both in histologic sections and on ultrasound biomicroscopy.
ATX1	Spinocerebellar ataxia 1 (SCA1) [MIM:164400]: Spinocerebellar ataxia is a clinically and genetically heterogeneous group of cerebellar disorders. Patients show progressive incoordination of gait and often poor coordination of hands, speech and eye movements, due to cerebellum degeneration with variable involvement of the brainstem and spinal cord. SCA1 belongs to the autosomal dominant cerebellar ataxias type I (ADCA I) which are characterized by cerebellar ataxia in combination with additional clinical features like optic atrophy, ophthalmoplegia, bulbar and extrapyramidal signs, peripheral neuropathy and dementia. SCA1 is caused by expansion of a CAG repeat in the coding region of ATXN1. Longer expansions result in earlier onset and more severe clinical manifestations of the disease. The disease is caused by expansion of the polyglutamine tract to about 40-83 repeats, causing accumulation in neurons and exerting toxicity.
BAT1	Cystinuria (CSNU) [MIM:220100]: An autosomal disorder characterized by impaired epithelial cell transport of cystine and dibasic amino acids (lysine, ornithine, and arginine) in the proximal renal tubule and gastrointestinal tract. The impaired renal reabsorption of cystine and its low solubility causes the formation of calculi in the urinary tract, resulting in obstructive uropathy, pyelonephritis, and, rarely, renal failure.
CADH1	Hereditary diffuse gastric cancer (HDGC) [MIM:137215]: A cancer predisposition syndrome with increased susceptibility to diffuse gastric cancer. Diffuse gastric cancer is a malignant disease characterized by poorly differentiated infiltrating lesions resulting in thickening of the stomach. Malignant tumors start in the stomach, can spread to the esophagus or the small intestine, and can extend through the stomach wall to nearby lymph nodes and organs. It also can metastasize to other parts of the body. Heterozygous CDH1 germline mutations are responsible for familial cases of diffuse gastric cancer. Somatic mutations have also been found in patients with sporadic diffuse gastric cancer and lobular breast cancer.; Endometrial cancer (ENDMC) [MIM:608089]: A malignancy of endometrium, the mucous lining of the uterus. Most endometrial cancers are adenocarcinomas, cancers that begin in cells that make and release mucus and other fluids. Note=Disease susceptibility is associated with variations affecting the gene represented in this entry.; Ovarian cancer (OC) [MIM:167000]: The term ovarian cancer defines malignancies originating from ovarian tissue. Although many histologic types of ovarian tumors have been described, ovarian epithelial carcinoma is the most common form. Ovarian cancers are often asymptomatic and the recognized signs and symptoms, even of late-stage disease, are vague. Consequently, most patients are diagnosed with advanced disease.; Breast cancer, lobular (LBC) [MIM:137215]: A type of breast cancer that begins in the milk-producing glands (lobules) of the breast.; Blepharocheilodontic syndrome 1 (BCDS1) [MIM:119580]: A form of blepharocheilodontic syndrome, a rare autosomal dominant disorder. It is characterized by lower eyelid ectropion, upper eyelid distichiasis, euryblepharon, bilateral cleft lip and palate, and features of ectodermal dysplasia, including hair anomalies, conical teeth, and tooth agenesis. An additional rare manifestation is an imperforate anus. There is considerable phenotypic variability among affected individuals.
CATD	Ceroid lipofuscinosis, neuronal, 10 (CLN10) [MIM:610127]: A form of neuronal ceroid lipofuscinosis with onset at birth or early childhood. Neuronal ceroid lipofuscinoses are progressive neurodegenerative, lysosomal storage diseases characterized by intracellular accumulation of autofluorescent liposomal material, and clinically by seizures, dementia, visual loss, and/or cerebral atrophy.
CDKN3	Hepatocellular carcinoma (HCC) [MIM:114550]: A primary malignant neoplasm of epithelial liver cells. The significant risk factors for HCC are chronic hepatitis B virus (HBV) infection, chronic hepatitis C virus (HCV) infection, prolonged dietary aflatoxin exposure, alcoholic cirrhosis, and cirrhosis due to other causes.
CEBPA	Leukemia, acute myelogenous (AML) [MIM:601626]: A subtype of acute leukemia, a cancer of the white blood cells. AML is a malignant disease of bone marrow characterized by maturational arrest of hematopoietic precursors at an early stage of development. Clonal expansion of myeloid blasts occurs in bone marrow, blood, and other tissue. Myelogenous leukemias develop from changes in cells that usually produce neutrophils, basophils, eosinophils, and monocytes.
CGL	Cystathioninuria (CSTNU) [MIM:219500]: Autosomal recessive phenotype characterized by abnormal accumulation of plasma cystathionine, leading to increased urinary excretion.
CHK2	Li-Fraumeni syndrome 2 (LFS2) [MIM:609265]: A highly penetrant familial cancer syndrome that in its classic form is defined by the existence of a proband affected by a sarcoma before 45 years with a first degree relative affected by any tumor before 45 years and another first degree relative with any tumor before 45 years or a sarcoma at any age. Other clinical definitions for LFS have been proposed (PubMed:8118819 and PubMed:8718514) and called Li-Fraumeni like syndrome (LFL). In these families affected relatives develop a diverse set of malignancies at unusually early

	ages. Four types of cancers account for 80% of tumors occurring in TP53 germline mutation carriers: breast cancers, soft tissue, and bone sarcomas, brain tumors (astrocytomas) and adrenocortical carcinomas. Less frequent tumors include choroid plexus carcinoma or papilloma before the age of 15, rhabdomyosarcoma before the age of 5, leukemia, Wilms tumor, malignant phyllodes tumor, colorectal and gastric cancers.; Prostate cancer (PC) [MIM:176807]: A malignancy originating in tissues of the prostate. Most prostate cancers are adenocarcinomas that develop in the acini of the prostatic ducts. Other rare histopathologic types of prostate cancer that occur in approximately 5% of patients include small cell carcinoma, mucinous carcinoma, prostatic ductal carcinoma, transitional cell carcinoma, squamous cell carcinoma, basal cell carcinoma, adenoid cystic carcinoma (basaloid), signet-ring cell carcinoma and neuroendocrine carcinoma.; Osteogenic sarcoma (OSRC) [MIM:259500]: A sarcoma originating in bone-forming cells, affecting the ends of long bones; Breast cancer (BC) [MIM:114480]: A common malignancy originating from breast epithelial tissue. Breast neoplasms can be distinguished by their histologic pattern. Invasive ductal carcinoma is by far the most common type. Breast cancer is etiologically and genetically heterogeneous. Important genetic factors have been indicated by familial occurrence and bilateral involvement. Mutations at more than one locus can be involved in different families or even in the same case.
COQ4	Coenzyme Q10 deficiency, primary, 7 (COQ10D7) [MIM:616276]: An autosomal recessive disorder resulting from mitochondrial dysfunction and characterized by decreased levels of coenzyme Q10, and severe cardiac or neurologic symptoms soon after birth, usually resulting in death. Rarely, symptoms may have later onset.
CREB1	Angiomatoid fibrous histiocytoma (AFH) [MIM:612160]: A distinct variant of malignant fibrous histiocytoma that typically occurs in children and adolescents and is manifest by nodular subcutaneous growth. Characteristic microscopic features include lobulated sheets of histiocyte-like cells intimately associated with areas of hemorrhage and cystic pseudovascular spaces, as well as a striking cuffing of inflammatory cells, mimicking a lymph node metastasis. Note=The gene represented in this entry may be involved in disease pathogenesis. A chromosomal aberration involving CREB1 is found in a patient with angiomatoid fibrous histiocytoma. Translocation t(2;22)(q33;q12) with CREB1 generates a EWSR1/CREB1 fusion gene that is the most common genetic abnormality in this tumor type.; Note=A CREB1 mutation has been found in a patient with multiple congenital anomalies consisting of agenesis of the corpus callosum, cerebellar hypoplasia, severe neonatal respiratory distress refractory to surfactant, thymus hypoplasia, and thyroid follicular hypoplasia.
CRY1	Delayed sleep phase syndrome (DSPS) [MIM:614163]: A circadian rhythm sleep disorder characterized by sleep-onset insomnia and difficulty in awakening at the desired time. Patients with DSPS have chronic difficulty in adjusting their sleep-onset and wake-up times to occupational, school, and social activities. An adenine-to-cytosine transversion within the 5'splice site following exon 11 has been found in multiple members of a DSPD family and segregates with the disorder with autosomal dominant inheritance pattern. This variant is predicted to cause exon 11 skipping and an in-frame deletion of 24 residues in the C-terminal region of CRY1. Functional studies show that the mutated protein acts as a more potent transcriptional repressor than wild-type causes reduced expression of key transcriptional targets and lengthens the period of circadian molecular rhythms.
DCPS	Al-Raqad syndrome (ARS) [MIM:616459]: A syndrome characterized by delayed psychomotor development, moderate to severe intellectual disability, weak or absent speech, microcephaly, congenital hypotonia, and severe growth delay.
DDIT3	Myxoid liposarcoma (MXLIPO) [MIM:613488]: A soft tissue tumor that tends to occur in the limbs (especially the thigh) of patients ranging in age from 35 to 55 years. It is defined by the presence of a hypocellular spindle cell proliferation set in a myxoid background, often with mucin pooling. Lipoblasts tend to be small and often monovacuolated and to cluster around vessels or at the periphery of the lesion. A chromosomal aberration involving DDIT3 has been found in a patient with malignant myxoid liposarcoma. Translocation t(12;16)(q13;p11) with FUS (PubMed:7503811).
DHB3	Male pseudohermaphroditism with gynecomastia (MPH) [MIM:264300]: An autosomal recessive disorder that manifests, in males, as undermasculinization characterized by hypoplastic-to-normal internal genitalia (epididymis, vas deferens, seminal vesicles, and ejaculatory ducts) but female external genitalia and the absence of a prostate. This phenotype is caused by inadequate testicular synthesis of testosterone, which, in turn, results in the insufficient formation of dihydrotestosterone in the anlage of the external genitalia and prostate during fetal development. At the expected time of puberty, there is a marked increase in plasma luteinizing hormone and, consequently, in the testicular secretion of androstenedione. Hence, a diagnostic hallmark of this disorder is a decreased plasma testosterone-to-androstenedione ratio. Significant amounts of the circulating androstenedione are, however, converted to testosterone, in peripheral tissues, thereby causing virilization.

DSPP	Deafness, autosomal dominant, 39, with dentinogenesis imperfecta 1 (DFNA39/DGI1) [MIM:605594]: A disorder characterized by the association of progressive sensorineural high-frequency hearing loss with dentinogenesis imperfecta; Dentinogenesis imperfecta, Shields type 2 (DGI2) [MIM:125490]: A form of dentinogenesis imperfecta, an autosomal dominant dentin disorder characterized by amber-brown, opalescent teeth that fracture and shed their enamel during mastication, thereby exposing the dentin to rapid wear. Radiographically, the crown appears bulbous, and pulpal obliteration is common. The pulp chambers are initially larger than normal prior and immediately after tooth eruption, and then progressively close down to become almost obliterated by abnormal dentin formation. Roots are short and thin. Both primary and permanent teeth are affected. DGI2 is not associated with osteogenesis imperfecta. DSPP defects causing dentin abnormalities to act in a dominant negative manner and include missense, splice-site, frameshift mutations. 5' frameshift mutations cause dentin dysplasia while frameshift mutations at the 3' end cause the more severe dentinogenesis imperfecta phenotype (PubMed:18521831 and PubMed:22392858).; Dentinogenesis imperfecta, Shields type 3 (DGI3) [MIM:125500]: A form of dentinogenesis imperfecta, an autosomal dominant dentin disorder characterized by amber-brown, opalescent teeth that fracture and shed their enamel during mastication, thereby exposing the dentin to rapid wear. Radiographically, the crown appears bulbous, and pulpal obliteration is common. The pulp chambers are initially larger than normal prior and immediately after tooth eruption, and then progressively close down to become almost obliterated by abnormal dentin formation. Roots are short and thin. Both primary and permanent teeth are affected. DGI3 teeth typically manifest multiple periapical radiolucencies. DGI3 is not associated with osteogenesis imperfecta. DSPP defects causing dentin abnormalities to act in a dominant negative manner and include missense, splice-site, frameshift mutations. 5' frameshift mutations cause dentin dysplasia while frameshift mutations at the 3' end cause the more severe dentinogenesis imperfecta phenotype (PubMed:18521831 and PubMed:22392858).; Dentin dysplasia 2 (DTDP2) [MIM:125420]: A dental defect in which the deciduous teeth are opalescent. The permanent teeth are of normal shape, form, and color in most cases. The root length is normal. On radiographs, the pulp chambers of permanent teeth are obliterated, have a thistle-tube deformity and contain pulp stones. {ECO:0000269|PubMed:12354781, ECO:0000269|PubMed:18521831}. Note=The disease is caused by mutations affecting the gene represented in this entry. DSPP defects causing dentin abnormalities to act in a dominant negative manner and include missense, splice-site, frameshift mutations. 5' frameshift mutations cause dentin dysplasia while frameshift mutations at the 3' end cause the more severe dentinogenesis imperfecta phenotype (PubMed:18521831, PubMed:22392858).
DUOX2	Thyroid dyshormonogenesis 6 (TDH6) [MIM:607200]: A disorder due to a defective conversion of accumulated iodide to organically bound iodine. The iodide organification defect can be partial or complete.; Note=Defects in DUOX2 may play a role in the pathogenesis of very early onset inflammatory bowel disease (VEOIBD), a chronic, relapsing inflammation of the gastrointestinal tract with a complex etiology diagnosed before 6 years of age. VEOIBD is subdivided into Crohn disease and ulcerative colitis phenotypes. Crohn disease may affect any part of the gastrointestinal tract from the mouth to the anus, but the phenotype of children with onset of Crohn disease occurring younger than the age of 10 is predominantly colonic, with a lower risk of ileal disease. Bowel inflammation is transmural and discontinuous; it may contain granulomas or be associated with intestinal or perianal fistulas. In contrast, in ulcerative colitis, the inflammation is continuous and limited to rectal and colonic mucosal layers; fistulas and granulomas are not observed. Both diseases include extraintestinal inflammation of the skin, eyes, or joints.
DYR	Megaloblastic anemia due to dihydrofolate reductase deficiency (DHFRD) [MIM:613839]: An inborn error of metabolism, characterized by megaloblastic anemia and/or pancytopenia, severe cerebral folate deficiency, and cerebral tetrahydrobiopterin deficiency. Clinical features include variable neurologic symptoms, ranging from severe developmental delay and generalized seizures in infancy to childhood absence epilepsy with learning difficulties, to lack of symptoms.
ELOV5	Spinocerebellar ataxia 38 (SCA38) [MIM:615957]: A form of spinocerebellar ataxia, a clinically and genetically heterogeneous group of cerebellar disorders. Patients show progressive incoordination of gait and often poor coordination of hands, speech and eye movements, due to degeneration of the cerebellum with variable involvement of the brainstem and spinal cord. SCA38 is an autosomal dominant form characterized by adult-onset of slowly progressive gait ataxia accompanied by nystagmus. Brain MRI shows cerebellar atrophy.
ELP2	Mental retardation, autosomal recessive 58 (MRT58) [MIM:617270]: A form of mental retardation, a disorder characterized by significantly below average general intellectual functioning associated with impairments in adaptive behavior and manifested during the developmental period. MRT58 transmission pattern is consistent with autosomal recessive inheritance.
ERCC8	Cockayne syndrome A (CSA) [MIM:216400]: A rare disorder characterized by cutaneous sensitivity to sunlight, abnormal and slow growth, cachectic dwarfism, progeroid appearance, progressive pigmentary retinopathy, and sensorineural deafness. There is delayed neural development and severe progressive neurologic degeneration resulting in mental retardation. Two clinical forms are recognized: in the classical form of Cockayne syndrome type 1, the symptoms are progressive and typically become apparent within the first few years or life; the less common Cockayne syndrome type 2 is characterized by more severe symptoms that manifest prenatally. Cockayne syndrome shows some overlap with certain forms of xeroderma pigmentosum. Unlike xeroderma pigmentosum, patients with Cockayne syndrome do not manifestly increased freckling and other pigmentation abnormalities in the skin and have no significant increase in skin cancer.; UV-sensitive syndrome 2 (UVSS2) [MIM:614621]: An autosomal recessive disorder characterized by cutaneous photosensitivity and mild freckling in the absence of neurological abnormalities or skin tumors.
ETFB	Glutaric aciduria 2B (GA2B) [MIM:231680]: An autosomal recessively inherited disorder of fatty acid, amino acid, and choline metabolism. It is characterized by multiple acyl-CoA dehydrogenase deficiencies resulting in large excretion not only of glutaric acid, but also of lactic, ethylmalonic, butyric, isobutyric, 2-methyl-butyric, and isovaleric acids.

ETFD	Glutaric aciduria 2C (GA2C) [MIM:231680]: An autosomal recessively inherited disorder of fatty acid, amino acid, and choline metabolism. It is characterized by multiple acyl-CoA dehydrogenase deficiencies resulting in large excretion not only of glutaric acid, but also of lactic, ethylmalonic, butyric, isobutyric, 2-methyl-butyric, and isovaleric acids.
FBLN4	Cutis laxa, autosomal recessive, 1B (ARCL1B) [MIM:614437]: A connective tissue disorder characterized by loose, hyperextensible skin with decreased resilience and elasticity leading to a prematurely aged appearance. Face, hands, feet, joints, and torso may be differentially affected. The clinical spectrum of autosomal recessive cutis laxa is highly heterogeneous with respect to organ involvement and severity. ARCL1B features include emphysema, lethal pulmonary artery occlusion, aortic aneurysm, cardiopulmonary insufficiency, birth fractures, arachnodactyly, and fragility of blood vessels.
FOXD3	Autoimmune disease 1 (AIS1) [MIM:607836]: An autoimmune disorder characterized by the association of vitiligo with autoimmune thyroiditis (Hashimoto thyroiditis).
FOXO1	Rhabdomyosarcoma 2 (RMS2) [MIM:268220]: A form of rhabdomyosarcoma, a highly malignant tumor of striated muscle derived from primitive mesenchymal cells and exhibiting differentiation along rhabdomyoblastic lines. Rhabdomyosarcoma is one of the most frequently occurring soft tissue sarcomas and the most common in children. It occurs in four forms: alveolar, pleomorphic, embryonal and botryoidal rhabdomyosarcomas. Note=The gene represented in this entry may be involved in disease pathogenesis. Chromosomal aberrations involving FOXO1 are found in rhabdomyosarcoma. Translocation (2;13)(q35;q14) with PAX3 and translocation t(1;13)(p36;q14) with PAX7. The resulting protein is a transcriptional activator.
FOXO4	Note=A chromosomal aberration involving FOXO4 is found in acute leukemias. Translocation t(X;11)(q13;q23) with KMT2A/MLL1. The result is a rogue activator protein.
FYB1	Thrombocytopenia 3 (THC3) [MIM:273900]: Thrombocytopenia is defined by a decrease in the number of platelets in circulating blood, resulting in the potential for increased bleeding and decreased the ability for clotting. THC3 is an autosomal recessive form characterized by onset in infancy.
FYV1	Corneal dystrophy, fleck (CFD) [MIM:121850]: A form of corneal stromal dystrophy characterized by numerous small white flecks scattered in all levels of the stroma, with configurations varying from semicircular to wreath-like, curvilinear, or punctate. Although CFD may occasionally cause mild photophobia, patients are typically asymptomatic and have normal vision.
GABT	GABA transaminase deficiency (GABATD) [MIM:613163]: An enzymatic deficiency resulting in psychomotor retardation, hypotonia, hyperreflexia, lethargy, refractory seizures, and EEG abnormalities.
GDF5	Acromesomelic chondrodysplasia, Grebe type (AMDG) [MIM:200700]: An autosomal recessive acromesomelic chondrodysplasia. Acromesomelic chondrodysplasias are rare hereditary skeletal disorders characterized by short stature, very short limbs and hand/foot malformations. The severity of limb abnormalities increases from proximal to distal with profoundly affected hands and feet showing brachydactyly and/or rudimentary fingers (knob-like fingers). AMDG is characterized by normal axial skeletons and missing or fused skeletal elements within the hands and feet; Acromesomelic chondrodysplasia, Hunter-Thompson type (AMDH) [MIM:201250]: An autosomal recessive form of dwarfism. Patients have limb abnormalities, with the middle and distal segments being most affected and the lower limbs more affected than the upper. AMDH is characterized by normal axial skeletons and missing or fused skeletal elements within the hands and feet. {ECO:0000269|PubMed:8589725}. Note=The disease is caused by mutations affecting the gene represented in this entry.; Brachydactyly C (BDC) [MIM:113100]: A form of brachydactyly. Brachydactyly defines a group of inherited malformations characterized by shortening of the digits due to abnormal development of the phalanges and/or the metacarpals. Brachydactyly type C is characterized by deformity of the middle and proximal phalanges of the second and third fingers, sometimes with hypersegmentation of the proximal phalanx. The ring finger may be essentially normal and project beyond the others. Some BDC patients with GDF5 mutations also manifest clinical features of ASPED angel-shaped phalango-epiphyseal dysplasia (ASPED), an autosomal dominant skeletal abnormality characterized by a typical angel-shaped phalanx, brachydactyly, specific radiological findings, abnormal dentition, hip dysplasia, and delayed bone age. This suggests that BDC and ASPED are part of the same clinical spectrum (PubMed:22828468).; Du Pan syndrome (DPS) [MIM:228900]: Rare autosomal recessive condition characterized by the absence of the fibulae and severe acromesomelic limb shortening with small, non-functional toes. Although milder, the phenotype resembles the autosomal recessive Hunter-Thompson and Grebe types of acromesomelic chondrodysplasia; Symphalangism, proximal 1B (SYM1B) [MIM:615298]: A disease characterized by the hereditary absence of the proximal interphalangeal joints. Distal interphalangeal joints are less frequently involved, and metacarpophalangeal joints are rarely affected whereas carpal bone malformation and fusion are common. In the lower extremities, the tarsal bone coalition is common. Conductive hearing loss is seen and is due to fusion of the stapes to the petrous part of the temporal bone.; Multiple synostoses syndrome 2 (SYNS2) [MIM:610017]: A bone disease characterized by multiple progressive joint fusions that commonly involve proximal interphalangeal, tarsal-carpal, humeroradial and cervical spine joints. Additional features can include progressive conductive deafness and facial dysmorphism.; Brachydactyly A2 (BDA2) [MIM:112600]: A form of brachydactyly. Brachydactyly defines a group of inherited malformations characterized by shortening of the digits due to abnormal development of the phalanges and/or the metacarpals. In brachydactyly type A2 shortening of the middle phalanges is confined to the index finger and the second toe, all other digits being more or less normal. Because of a rhomboid or triangular shape of the affected middle phalanx, the end of the second finger usually deviates radially.;

	Osteoarthritis 5 (OS5) [MIM:612400]: A degenerative disease of the joints characterized by degradation of the hyaline articular cartilage and remodeling of the subchondral bone with sclerosis. Clinical symptoms include pain and joint stiffness often leading to significant disability and joint replacement.; Brachydactyly A1, C (BDA1C) [MIM:615072]: A form of brachydactyly type A1. Brachydactyly defines a group of inherited malformations characterized by shortening of the digits due to abnormal development of the phalanges and/or the metacarpals. Brachydactyly type A1 is characterized by middle phalanges of all the digits rudimentary or fused with the terminal phalanges. The proximal phalanges of the thumbs and big toes are short. BDA1C inheritance can be autosomal dominant or autosomal recessive. Autosomal dominant BDA1C has a milder phenotype.
GLGB	Glycogen storage disease 4 (GSD4) [MIM:232500]: A metabolic disorder characterized by the accumulation of an amylopectin-like polysaccharide. The typical clinical manifestation is a liver disease of childhood, progressing to lethal hepatic cirrhosis. Most children with this condition die before two years of age. However, liver disease is not always progressive. No treatment apart from liver transplantation has been found to prevent progression of the disease. There is also a neuromuscular form of glycogen storage disease type 4 that varies in onset (perinatal, congenital, juvenile, or adult) and severity.; Note=Neuromuscular perinatal glycogen storage disease type 4 is associated with non-immune hydrops fetalis, a generalized edema of the fetus with fluid accumulation in the body cavities due to non-immune causes. Non-immune hydrops fetalis is not a diagnosis in itself but a symptom, a feature of many genetic disorders, and the end-stage of a wide variety of disorders.; Polyglucosan body neuropathy, adult form (APBN) [MIM:263570]: A late-onset, slowly progressive disorder affecting the central and peripheral nervous systems. Patients typically present after age 40 years with a variable combination of cognitive impairment, pyramidal tetraparesis, peripheral neuropathy, and neurogenic bladder. Other manifestations include cerebellar dysfunction and extrapyramidal signs. The pathologic hallmark of APBN is the widespread accumulation of round, intracellular polyglucosan bodies throughout the nervous system, which are confined to neuronal and astrocytic processes.
GP1BA	Non-arteritic anterior ischemic optic neuropathy (NAION) [MIM:258660]: An ocular disease due to ischemic injury to the optic nerve. It usually affects the optic disk and leads to visual loss and optic disk swelling of a pallid nature. Visual loss is usually sudden, or over a few days at most and is usually permanent, with some recovery possibly occurring within the first weeks or months. Patients with small disks having smaller or non-existent cups have an anatomical predisposition for non-arteritic anterior ischemic optic neuropathy. As an ischemic episode evolves, the swelling compromises circulation, with a spiral of ischemia resulting in further neuronal damage.; Bernard-Soulier syndrome (BSS) [MIM:231200]: A coagulation disorder characterized by a prolonged bleeding time, unusually large platelets, thrombocytopenia, and impaired prothrombin consumption..; Bernard-Soulier syndrome A2, autosomal dominant (BSSA2) [MIM:153670]: A coagulation disorder characterized by mild to moderate bleeding tendency, thrombocytopenia, and an increased mean platelet volume. Some individuals have no symptoms. Mild bleeding tendencies manifest as epistaxis, gingival bleeding, menorrhagia, easy bruising, or prolonged bleeding after dental surgery.; Pseudo-von Willebrand disease (VWDP) [MIM:177820]: A bleeding disorder characterized by abnormally enhanced binding of von Willebrand factor by the platelet glycoprotein Ib (GP Ib) receptor complex. The hemostatic function is impaired due to the removal of VWF multimers from the circulation.
GPDA	Hypertriglyceridemia, transient infantile (HTGTI) [MIM:614480]: An autosomal recessive disorder characterized by the onset of moderate to severe transient hypertriglyceridemia in infancy that normalizes with age. The hypertriglyceridemia is associated with hepatomegaly, moderately elevated transaminases, persistent fatty liver, and the development of hepatic fibrosis.
GPVI	Bleeding disorder, platelet-type 11 (BDPLT11) [MIM:614201]: A mild to the moderate bleeding disorder caused by defective platelet activation and aggregation in response to collagen.
GRHPR	Hyperoxaluria primary 2 (HP2) [MIM:260000]: A disorder characterized by elevated urinary excretion of oxalate and L-glycerate, progressive tissue accumulation of insoluble calcium oxalate, nephrolithiasis, nephrocalcinosis, and end-stage renal disease.
GTR9	Hypouricemia renal 2 (RHUC2) [MIM:612076]: A disorder characterized by impaired uric acid reabsorption at the apical membrane of proximal renal tubule cells, and high urinary urate excretion. Patients often appear asymptomatic but may be subject to exercise-induced acute renal failure, chronic renal dysfunction, and nephrolithiasis.
HCK	Note=Aberrant activation of HCK by HIV-1 protein Nef enhances HIV-1 replication and contributes to HIV-1 pathogenicity.; Note=Aberrant activation of HCK, e.g., by the BCR-ABL fusion protein, promotes cancer cell proliferation.
HDAC9	Note=A chromosomal aberration involving HDAC9 is found in a family with Peters anomaly. Translocation t(1;7)(q41;p21) with TGFB2 resulting in lack of HDAC9 protein.
HXK4	Maturity-onset diabetes of the young 2 (MODY2) [MIM:125851]: A form of diabetes that is characterized by an autosomal dominant mode of inheritance, onset in childhood or early adulthood (usually before 25 years of age), a primary defect in insulin secretion and frequent insulin-independence at the beginning of the disease; Familial hyperinsulinemic hypoglycemia 3 (HHF3) [MIM:602485]: Most common cause of persistent hypoglycemia in infancy. Unless early and aggressive intervention is undertaken, brain damage from recurrent episodes of hypoglycemia may occur; Diabetes mellitus, non-insulin-dependent (NIDDM) [MIM:125853]: A multifactorial disorder of glucose homeostasis caused by a lack of sensitivity to the body's own insulin. Affected

	individuals usually have an obese body habitus and manifestations of a metabolic syndrome characterized by diabetes, insulin resistance, hypertension, and hypertriglyceridemia. The disease results in long-term complications that affect the eyes, kidneys, nerves, and blood vessels.; Diabetes mellitus, permanent neonatal (PNDM) [MIM:606176]: A rare form of diabetes distinct from childhood-onset autoimmune diabetes mellitus type 1. It is characterized by insulin-requiring hyperglycemia that is diagnosed within the first months of life. Permanent neonatal diabetes requires lifelong therapy.
I10R1	Inflammatory bowel disease 28 (IBD28) [MIM:613148]: A chronic, relapsing inflammation of the gastrointestinal tract with a complex etiology. It is subdivided into Crohn disease and ulcerative colitis phenotypes. Crohn disease may affect any part of the gastrointestinal tract from the mouth to the anus, but most frequently it involves the terminal ileum and colon. Bowel inflammation is transmural and discontinuous; it may contain granulomas or be associated with intestinal or perianal fistulas. In contrast, in ulcerative colitis, the inflammation is continuous and limited to rectal and colonic mucosal layers; fistulas and granulomas are not observed. Both diseases include extraintestinal inflammation of the skin, eyes, or joints.
IFIH1	Diabetes mellitus, insulin-dependent, 19 (IDDM19) [MIM:610155]: A multifactorial disorder of glucose homeostasis that is characterized by susceptibility to ketoacidosis in the absence of insulin therapy. Clinical features are polydipsia, polyphagia, and polyuria which result from hyperglycemia-induced osmotic diuresis and secondary thirst. These derangements result in long-term complications that affect the eyes, kidneys, nerves, and blood vessels.; Note=IFIH1 is the CADM-140 autoantigen, involved in clinically amyopathic dermatomyositis (CADM). This is a chronic inflammatory disorder that shows typical skin manifestations of dermatomyositis but has no or little evidence of clinical myositis. Anti-CADM-140 antibodies appear to be specific to dermatomyositis, especially CADM. Patients with anti-CADM-140 antibodies frequently develop life-threatening acute progressive interstitial lung disease (ILD).; Aicardi-Goutieres syndrome 7 (AGS7) [MIM:615846]: A form of Aicardi-Goutieres syndrome, a genetically heterogeneous disease characterized by cerebral atrophy, leukoencephalopathy, intracranial calcifications, chronic cerebrospinal fluid (CSF) lymphocytosis, increased CSF alpha-interferon, and negative serologic investigations for common prenatal infection. Clinical features as thrombocytopenia, hepatosplenomegaly and elevated hepatic transaminases along with intermittent fever may erroneously suggest an infective process. Severe neurological dysfunctions manifest in infancy as progressive microcephaly, spasticity, dystonic posturing, and profound psychomotor retardation. Death often occurs in early childhood.; Singleton-Merten syndrome 1 (SGMRT1) [MIM:182250]: An autosomal dominant disorder with variable expression. Core features are marked aortic calcification, dental anomalies, osteopenia, acro-osteolysis, and to lesser extent glaucoma, psoriasis, muscle weakness, and joint laxity. Dental anomalies include delayed eruption and immature root formation of permanent anterior teeth, early loss of permanent teeth due to short roots, acute root resorption, high caries, and aggressive alveolar bone loss. Additional clinical manifestations include particular facial characteristics and abnormal joint and muscle ligaments.
KCND3	Spinocerebellar ataxia 19 (SCA19) [MIM:607346]: A form of spinocerebellar ataxia, a clinically and genetically heterogeneous group of cerebellar disorders. Patients show progressive incoordination of gait and often poor coordination of hands, speech and eye movements, due to degeneration of the cerebellum with variable involvement of the brainstem and spinal cord. SCA19 is a relatively mild, cerebellar ataxia syndrome with cognitive impairment, pyramidal tract involvement, tremor and peripheral neuropathy, and mild atrophy of the cerebellar hemispheres and vermis.; Brugada syndrome 9 (BRGDA9) [MIM:616399]: A tachyarrhythmia characterized by right bundle branch block and ST segment elevation on an electrocardiogram (ECG). It can cause the ventricles to beat so fast that the blood is prevented from circulating efficiently in the body. When this situation occurs, the individual will faint and may die in a few minutes if the heart is not reset.
KIF4A	Mental retardation, X-linked 100 (MRX100) [MIM:300923]: A disorder characterized by significantly below average general intellectual functioning associated with impairments in adaptive behavior and manifested during the developmental period. Intellectual deficiency is the only primary symptom of non-syndromic X-linked mental retardation, while syndromic mental retardation presents with associated physical, neurological and/or psychiatric manifestations. MRX100 clinical features include intellectual disability, epilepsy, microcephaly, and cortical malformations.
KLH40	Nemaline myopathy 8 (NEM8) [MIM:615348]: A severe form of nemaline myopathy. Nemaline myopathies are muscular disorders characterized by muscle weakness of varying severity and onset, and abnormal thread-like or rod-shaped structures in muscle fibers on histologic examination. NEM8 is characterized by fetal akinesia or hypokinesia, followed by contractures, fractures, respiratory failure, and swallowing difficulties apparent at birth. Most patients die in infancy. Skeletal muscle biopsy shows numerous small nemaline bodies, often with no normal myofibrils.
LYN	Note=Constitutively phosphorylated and activated in cells from a number of chronic myelogenous leukemia (CML) and acute myeloid leukemia (AML) patients. Mediates phosphorylation of the BCR-ABL fusion protein. Abnormally elevated expression levels or activation of LYN signaling may play a role in the survival and proliferation of some types of cancer cells.
MDR1	Inflammatory bowel disease 13 (IBD13) [MIM:612244]: A chronic, relapsing inflammation of the gastrointestinal tract with a complex etiology. It is subdivided into Crohn disease and ulcerative colitis phenotypes. Crohn disease may affect any part of the gastrointestinal tract from the mouth to the anus, but most frequently it involves the terminal ileum and colon. Bowel inflammation is transmural and discontinuous; it may contain granulomas or be associated with intestinal or perianal fistulas. In contrast, in ulcerative colitis, the inflammation is continuous and limited to rectal and colonic mucosal layers; fistulas and granulomas are not observed. Both diseases include extraintestinal inflammation of the skin, eyes, or joints.

MED25	Charcot-Marie-Tooth disease 2B2 (CMT2B2) [MIM:605589]: A recessive axonal form of Charcot-Marie-Tooth disease, a disorder of the peripheral nervous system, characterized by progressive weakness and atrophy, initially of the peroneal muscles and later of the distal muscles of the arms. Charcot-Marie-Tooth disease is classified in two main groups on the basis of electrophysiologic properties and histopathology: primary peripheral demyelinating neuropathies (designated CMT1 when they are dominantly inherited) and primary peripheral axonal neuropathies (CMT2). Neuropathies of the CMT2 group are characterized by signs of axonal degeneration in the absence of obvious myelin alterations, normal or slightly reduced nerve conduction velocities, and progressive distal muscle weakness and atrophy. {ECO:0000269|PubMed:19290556}. Note=The disease is caused by mutations affecting the gene represented in this entry.; Basel-Vanagaite-Smirin-Yosef syndrome (BVSYS) [MIM:616449]: An autosomal recessive syndrome characterized by eye, brain, cardiac and palatal abnormalities as well as growth retardation, microcephaly, and severe intellectual disability.
MEF2A	Coronary artery disease, autosomal dominant, 1 (ADCAD1) [MIM:608320]: A common heart disease characterized by reduced or absent blood flow in one or more of the arteries that encircle and supply the heart. Its most important complication is an acute myocardial infarction.
MEIS2	Cleft palate, cardiac defects, and mental retardation (CPCMR) [MIM:600987]: An autosomal dominant disease characterized by multiple congenital malformations, mild-to-severe intellectual disability with poor speech, and delayed psychomotor development. Congenital malformations include heart defects, cleft lip/palate, distally-placed thumbs and toes, and cutaneous syndactyly between the second and third toes.
MOT12	Cataract 47 (CTRCT47) [MIM:612018]: A form of cataract, an opacification of the crystalline lens of the eye that frequently results in visual impairment or blindness. Opacities vary in morphology, are often confined to a portion of the lens, and may be static or progressive. In general, the more posteriorly located and dense an opacity, the higher the impact on visual function. CTRCT47 is characterized by the association of cataract with microcornea and renal glucosuria. Microcornea is defined by a corneal diameter inferior to 10 mm in both meridians in an otherwise healthy eye. Renal glucosuria is defined by elevated glucose level in the urine without hyperglycemia and without evidence of renal morphological anomalies.
MSH2	Hereditary non-polyposis colorectal cancer 1 (HNPCC1) [MIM:120435]: An autosomal dominant disease associated with a marked increase in cancer susceptibility. It is characterized by a familial predisposition to early-onset colorectal carcinoma (CRC) and extra-colonic tumors of the gastrointestinal, urological and female reproductive tracts. HNPCC is reported to be the most common form of inherited colorectal cancer in the Western world. Clinically, HNPCC is often divided into two subgroups. Type I is characterized by hereditary predisposition to colorectal cancer, young age of onset, and carcinoma observed in the proximal colon. Type II is characterized by increased risk for cancers in specific tissues such as the uterus, ovary, breast, stomach, small intestine, skin, and larynx in addition to the colon. Diagnosis of classical HNPCC is based on the Amsterdam criteria: 3 or more relatives affected by colorectal cancer, one a first degree relative of the other two; 2 or more generation affected; 1 or more colorectal cancers presenting before 50 years of age; exclusion of hereditary polyposis syndromes. The term 'suspected HNPCC' or 'incomplete HNPCC' can be used to describe families who do not or only partially fulfill the Amsterdam criteria, but in whom a genetic basis for colon cancer is strongly suspected; Muir-Torre syndrome (MRTES) [MIM:158320]: Rare autosomal dominant disorder characterized by sebaceous neoplasms and visceral malignancy.; Endometrial cancer (ENDMC) [MIM:608089]: A malignancy of endometrium, the mucous lining of the uterus. Most endometrial cancers are adenocarcinomas, cancers that begin in cells that make and release mucus and other fluids.; Mismatch repair cancer syndrome (MMRCS) [MIM:276300]: An autosomal recessive, rare, childhood cancer predisposition syndrome encompassing a broad tumor spectrum. This includes hematological malignancies, central nervous system tumors, Lynch syndrome-associated malignancies such as colorectal tumors as well as multiple intestinal polyps, embryonic tumors, and rhabdomyosarcoma. Multiple cafe-au-lait macules, a feature reminiscent of neurofibromatosis type 1, are often found as the first manifestation of underlying cancer. Areas of skin hypopigmentation have also been reported in MMRCS patients; Colorectal cancer (CRC) [MIM:114500]: A complex disease characterized by malignant lesions arising from the inner wall of the large intestine (the colon) and the rectum. Genetic alterations are often associated with progression from premalignant lesion (adenoma) to invasive adenocarcinoma. Risk factors for cancer of the colon and rectum include colon polyps, long-standing ulcerative colitis, and genetic family history.

MSH6	Hereditary non-polyposis colorectal cancer 5 (HNPCC5) [MIM:614350]: An autosomal dominant disease associated with a marked increase in cancer susceptibility. It is characterized by a familial predisposition to early-onset colorectal carcinoma (CRC) and extra-colonic tumors of the gastrointestinal, urological and female reproductive tracts. HNPCC is reported to be the most common form of inherited colorectal cancer in the Western world. Clinically, HNPCC is often divided into two subgroups. Type I is characterized by hereditary predisposition to colorectal cancer, young age of onset, and carcinoma observed in the proximal colon. Type II is characterized by increased risk for cancers in specific tissues such as the uterus, ovary, breast, stomach, small intestine, skin, and larynx in addition to the colon. Diagnosis of classical HNPCC is based on the Amsterdam criteria: 3 or more relatives affected by colorectal cancer, one a first degree relative of the other two; 2 or more generation affected; 1 or more colorectal cancers presenting before 50 years of age; exclusion of hereditary polyposis syndromes. The term 'suspected HNPCC' or 'incomplete HNPCC' can be used to describe families who do not or only partially fulfill the Amsterdam criteria, but in whom a genetic basis for colon cancer is strongly suspected; Endometrial cancer (ENDMC) [MIM:608089]: A malignancy of endometrium, the mucous lining of the uterus. Most endometrial cancers are adenocarcinomas, cancers that begin in cells that make and release mucus and other fluids.; Mismatch repair cancer syndrome (MMRCS) [MIM:276300]: An autosomal recessive, rare, childhood cancer predisposition syndrome encompassing a broad tumor spectrum. This includes hematological malignancies, central nervous system tumors, Lynch syndrome-associated malignancies such as colorectal tumors as well as multiple intestinal polyps, embryonic tumors, and rhabdomyosarcoma. Multiple cafe-au-lait macules, a feature reminiscent of neurofibromatosis type 1, are often found as the first manifestation of underlying cancer. Areas of skin hypopigmentation have also been reported in MMRCS patients.; Colorectal cancer (CRC) [MIM:114500]: A complex disease characterized by malignant lesions arising from the inner wall of the large intestine (the colon) and the rectum. Genetic alterations are often associated with progression from premalignant lesion (adenoma) to invasive adenocarcinoma. Risk factors for cancer of the colon and rectum include colon polyps, long-standing ulcerative colitis, and genetic family history.
NCF1	Granulomatous disease, chronic, cytochrome-b-positive 1, autosomal recessive (CGD1) [MIM:233700]: A disorder characterized by the inability of neutrophils and phagocytes to kill microbes that they have ingested. Patients suffer from life-threatening bacterial/fungal infections.
NCOA4	Note=A chromosomal aberration involving NCOA4 is found in papillary thyroid carcinomas (PTCs). Inversion inv(10)(q11.2;q11.2) generates the RET/NCOA4 (PTC3) oncogene.
NCPR	Antley-Bixler syndrome, with genital anomalies and disordered steroidogenesis (ABS1) [MIM:201750]: A disease characterized by the association of Antley-Bixler syndrome with steroidogenesis defects and abnormal genitalia. Antley-Bixler syndrome is characterized by craniosynostosis, radiohumeral synostosis present from the perinatal period, midface hypoplasia, choanal stenosis or atresia, femoral bowing, and multiple joint contractures.; Disordered steroidogenesis due to cytochrome P450 oxidoreductase deficiency (DISPORD) [MIM:613571]: A disorder resulting in a rare variant of congenital adrenal hyperplasia, with apparent combined P450C17 and P450C21 deficiency and accumulation of steroid metabolites. Affected girls are born with ambiguous genitalia, but their circulating androgens are low, and virilization does not progress. Conversely, affected boys are sometimes born undermasculinized. Boys and girls can present with bone malformations, in some cases resembling the pattern seen in patients with Antley-Bixler syndrome.
NGBR	Congenital disorder of glycosylation 1AA (CDG1AA) [MIM:617082]: A form of congenital disorder of glycosylation, a multisystem disorder caused by a defect in glycoprotein biosynthesis and characterized by under-glycosylated serum glycoproteins. Congenital disorders of glycosylation result in a wide variety of clinical features, such as defects in the nervous system development, psychomotor retardation, dysmorphic features, hypotonia, coagulation disorders, and immunodeficiency. The broad spectrum of features reflects the critical role of N-glycoproteins during embryonic development, differentiation, and maintenance of cell functions. CDG1AA inheritance is autosomal recessive.; Mental retardation, autosomal dominant 55, with seizures (MRD55) [MIM:617831]: A form of mental retardation, a disorder characterized by significantly below average general intellectual functioning associated with impairments in adaptive behavior and manifested during the developmental period. MRD55 patients suffer from seizures appearing during the first years of life.
NNRE	Encephalopathy, progressive, early-onset, with brain edema and/or leukoencephalopathy (PEBEL) [MIM:617186]: An autosomal recessive severe neurometabolic disorder characterized by severe leukoencephalopathy usually associated with a minor febrile illness. Affected infants tend to show normal early development followed by acute psychomotor regression with ataxia, hypotonia, respiratory insufficiency, and seizures. Disease course is rapidly progressive, leading to coma, global brain atrophy, and death in the first years of life. Brain imaging shows multiple abnormalities, including brain edema and signal abnormalities in the cortical and subcortical regions.
NT2NA	Note=Defects in NOTCH2NLA may be a cause of chromosome 1q21.1 deletion/duplication syndrome (PubMed:29856954). Deletions of NOTCH2NL (NOTCH2NLA, NOTCH2NLB and/or NOTCH2NLC) are present in patients affected by microcephaly, whereas macrocephaly is observed in patients with NOTCH2NL duplications (PubMed:29856954).
OAT	Hyperornithinemia with gyrate atrophy of choroid and retina (HOGA) [MIM:258870]: A disorder clinically characterized by a triad of progressive chorioretinal degeneration, early cataract formation, and type II muscle fiber atrophy. Characteristic chorioretinal atrophy with progressive constriction of the visual fields leads to blindness at the latest during the sixth decade of life. Patients generally have average intelligence.

PAX2	Papillorenal syndrome (PAPRS) [MIM:120330]: An autosomal dominant disorder characterized by both ocular and renal anomalies, but may also include vesicoureteral reflux, high-frequency hearing loss, central nervous system anomalies, and/or genital anomalies. Eye anomalies in this disorder consist of a wide and sometimes excavated dysplastic optic disk with the emergence of the retinal vessels from the periphery of the disk, designated optic nerve coloboma or 'morning glory' anomaly. Associated findings may include a small corneal diameter, retinal coloboma, scleral staphyloma, optic nerve cyst, microphthalmia, and pigmentary macular dysplasia. The kidneys are small and abnormally formed (renal hypodysplasia), and have fewer than the normal number of glomeruli, which are enlarged (oligomeganephronia). These ocular and renal anomalies result in decreased visual acuity and retinal detachment, as well as hypertension, proteinuria, and renal insufficiency that frequently progresses to end-stage renal disease.; Focal segmental glomerulosclerosis 7 (FSGS7) [MIM:616002]: A renal pathology defined by the presence of segmental sclerosis in glomeruli and resulting in proteinuria, reduced glomerular filtration rate and progressive decline in renal function. Renal insufficiency often progresses to end-stage renal disease, a highly morbid state requiring either dialysis therapy or kidney transplantation.
PAX6	Aniridia 1 (AN1) [MIM:106210]: A congenital, bilateral, panocular disorder characterized by complete absence of the iris or extreme iris hypoplasia. Aniridia is not just an isolated defect in iris development, but it is associated with macular and optic nerve hypoplasia, cataract, corneal changes, nystagmus. Visual acuity is generally low but is unrelated to the degree of iris hypoplasia. Glaucoma is a secondary problem causing additional visual loss over time; Anterior segment dysgenesis 5 (ASGD5) [MIM:604229]: A form of anterior segment dysgenesis, a group of defects affecting anterior structures of the eye including the cornea, iris, lens, trabecular meshwork, and Schlemm canal. Anterior segment dysgeneses result from abnormal migration or differentiation of the neural crest-derived mesenchymal cells that give rise to components of the anterior chamber during eye development. Different anterior segment anomalies may exist alone or in combination, including iris hypoplasia, enlarged or reduced corneal diameter, corneal vascularization and opacity, posterior embryotoxon, corectopia, polycoria, abnormal iridocorneal angle, ectopia lentis, and anterior synechiae between the iris and posterior corneal surface. Clinical conditions falling within the phenotypic spectrum of anterior segment dysgeneses include aniridia, Axenfeld anomaly, Reiger anomaly/syndrome, Peters anomaly, and iridogoniodysgenesis.; Foveal hypoplasia 1 (FVH1) [MIM:136520]: An isolated form of foveal hypoplasia, a developmental defect of the eye defined as the lack of foveal depression with continuity of all neurosensory retinal layers in the presumed foveal area. Clinical features include the absence of a foveal pit on optical coherence tomography, the absence of foveal hyperpigmentation, the absence of foveal avascularity, the absence of foveal and macular reflexes, decreased visual acuity, and nystagmus. Anterior segment anomalies and cataract are observed in some FVH1 patients; Keratitis hereditary (KERH) [MIM:148190]: An ocular disorder characterized by corneal opacification, recurrent stromal keratitis and vascularization; Coloboma, ocular, autosomal dominant (COAD) [MIM:120200]: A set of malformations resulting from abnormal morphogenesis of the optic cup and stalk, and the fusion of the fetal fissure (optic fissure). The clinical presentation is variable. Some individuals may present with minimal defects in the anterior iris leaf without other ocular defects. More complex malformations create a combination of iris, uveoretinal and/or optic nerve defects without or with microphthalmia or even anophthalmia.; Coloboma of the optic nerve (COLON) [MIM:120430]: An ocular defect that is due to malclosure of the fetal intraocular fissure affecting the optic nerve head. In some affected individuals, it appears as enlargement of the physiologic cup with severely affected eyes showing large cavities at the site of the disk.; Bilateral optic nerve hypoplasia (BONH) [MIM:165550]: A congenital anomaly in which the optic disk appears abnormally small. It may be an isolated finding or part of a spectrum of anatomic and functional abnormalities that includes partial or complete agenesis of the septum pellucidum, other midline brain defects, cerebral anomalies, pituitary dysfunction, and structural abnormalities of the pituitary.; Aniridia 2 (AN2) [MIM:617141]: A form of aniridia, a congenital, bilateral, panocular disorder characterized by complete absence of the iris or extreme iris hypoplasia. Aniridia is not just an isolated defect in iris development, but it is associated with macular and optic nerve hypoplasia, cataract, corneal changes, nystagmus. Visual acuity is generally low but is unrelated to the degree of iris hypoplasia. Glaucoma is a secondary problem causing additional visual loss over time. A mutation in a PAX6 long-range cis-regulatory element, known as SIMO, affects PAX6 expression in the developing eye and has pathological consequences. The mutation is located in ELP4 intron 9, 150 kb downstream of PAX6.
PAX7	Rhabdomyosarcoma 2 (RMS2) [MIM:268220]: A form of rhabdomyosarcoma, a highly malignant tumor of striated muscle derived from primitive mesenchymal cells and exhibiting differentiation along rhabdomyoblastic lines. Rhabdomyosarcoma is one of the most frequently occurring soft tissue sarcomas and the most common in children. It occurs in four forms: alveolar, pleomorphic, embryonal and botryoidal rhabdomyosarcomas. A chromosomal aberration involving PAX7 is found in rhabdomyosarcoma. Translocation t(1;13)(p36;q14) with FOXO1. The resulting protein is a transcriptional activator.
PBX1	Congenital anomalies of kidney and urinary tract syndrome with or without hearing loss, abnormal ears, or developmental delay (CAKUTHED) [MIM:617641]: An autosomal dominant disorder characterized by variable congenital anomalies of the kidney and urinary tract, sometimes resulting in renal dysfunction or failure, dysmorphic facial features, and abnormalities of the outer ear. Most patients have hearing loss, and some may have a global developmental delay. {ECO:0000269|PubMed:28270404, ECO:0000269|PubMed:28566479}. Note=The disease is caused by mutations affecting the gene represented in this entry.; Note=A chromosomal aberration involving PBX1 is a cause of pre-B-cell acute lymphoblastic leukemia (B-ALL). Translocation t(1;19)(q23;p13.3) with TCF3. TCF3-PBX1 transforms cells by constitutively activating transcription of genes regulated by PBX1 or by other members of the PBX protein family.
PGFRA	Note=A chromosomal aberration involving PDGFRA is found in some cases of hypereosinophilic syndrome. Interstitial chromosomal deletion del(4)(q12q12) causes the fusion of FIP1L1 and PDGFRA (FIP1L1-PDGFRA). Mutations that cause overexpression and/or constitutive activation of PDGFRA may be a cause of the hypereosinophilic syndrome.; Gastrointestinal stromal tumor (GIST) [MIM:606764]: Common mesenchymal neoplasms arising in the gastrointestinal tract, most often in the stomach. They are histologically, immunohistochemically, and genetically different from typical leiomyomas, leiomyosarcomas, and schwannomas. Most GISTs are composed of a relatively uniform population of spindle-shaped cells. Some tumors are dominated by epithelioid

	cells or contain a mixture of spindle and epithelioid morphologies. Primary GISTs in the gastrointestinal tract commonly metastasize in the omentum and mesenteries, often as multiple nodules. However, primary tumors may also occur outside of the gastrointestinal tract, in other intra-abdominal locations, especially in the omentum and mesentery. Mutations causing PDGFRA constitutive activation have been found in gastrointestinal stromal tumors lacking KIT mutations (PubMed:12522257).
PGM1	Congenital disorder of glycosylation 1T (CDG1T) [MIM:614921]: A form of congenital disorder of glycosylation, a multisystem disorder caused by a defect in glycoprotein biosynthesis and characterized by under-glycosylated serum glycoproteins. Congenital disorders of glycosylation result in a wide variety of clinical features, such as defects in the nervous system development, psychomotor retardation, dysmorphic features, hypotonia, coagulation disorders, and immunodeficiency. The broad spectrum of features reflects the critical role of N-glycoproteins during embryonic development, differentiation, and maintenance of cell functions.
PLAP	Neurodevelopmental disorder with progressive microcephaly, spasticity, and brain anomalies (NDMSBA) [MIM:617527]: An autosomal recessive neurodevelopmental disorder characterized by progressive microcephaly, spastic quadriparesis, global developmental delay, profound mental retardation and severely impaired or absent motor function. More variable features include seizures and optic atrophy.
PPARG	Note=Defects in PPARG can lead to type 2 insulin-resistant diabetes and hypertension. PPARG mutations may be associated with colon cancer.; Obesity (OBESITY) [MIM:601665]: A condition characterized by an increase of body weight beyond the limitation of skeletal and physical requirements, as the result of excessive accumulation of body fat; Lipodystrophy, familial partial, 3 (FPLD3) [MIM:604367]: A form of lipodystrophy characterized by marked loss of subcutaneous fat from the extremities. Facial adipose tissue may be increased, decreased or normal. Affected individuals show an increased preponderance of insulin resistance, diabetes mellitus, and dyslipidemia.; Glioma 1 (GLM1) [MIM:137800]: Gliomas are benign or malignant central nervous system neoplasms derived from glial cells. They comprise astrocytomas and glioblastoma multiforme that are derived from astrocytes, oligodendrogliomas derived from oligodendrocytes and ependymomas derived from ependymocytes. Polymorphic PPARG alleles have been found to be significantly over-represented among a cohort of American patients with sporadic glioblastoma multiforme suggesting a possible contribution to disease susceptibility.
PRLR	Multiple fibroadenomas of the breast (MFAB) [MIM:615554]: A benign breast disease marked by lobuloalveolar growth with an abnormally high proliferation of the epithelium, and characterized by the presence of more than 3 fibroadenomas in one breast. Fibroadenomas are adenomas containing fibrous tissue.; Hyperprolactinemia (HPRL) [MIM:615555]: A disorder characterized by increased levels of prolactin in the blood not associated with gestation or the puerperium. HPRL may result in infertility, hypogonadism, and galactorrhea.
RAD51	Breast cancer (BC) [MIM:114480]: A common malignancy originating from breast epithelial tissue. Breast neoplasms can be distinguished by their histologic pattern. Invasive ductal carcinoma is by far the most common type. Breast cancer is etiologically and genetically heterogeneous. Important genetic factors have been indicated by familial occurrence and bilateral involvement. Mutations at more than one locus can be involved in different families or even in the same case.; Mirror movements 2 (MRMV2) [MIM:614508]: A disorder characterized by contralateral involuntary movements that mirror voluntary ones. While mirror movements are occasionally found in young children, persistence beyond the age of 10 is abnormal. Mirror movements occur more commonly in the upper extremities; Fanconi anemia, complementation group R (FANCR) [MIM:617244]: A disorder affecting all bone marrow elements and resulting in anemia, leukopenia, and thrombopenia. It is associated with cardiac, renal and limb malformations, dermal pigmentary changes, and a predisposition to the development of malignancies. At the cellular level, it is associated with hypersensitivity to DNA-damaging agents, chromosomal instability (increased chromosome breakage) and defective DNA repair.
RERE	Note=A chromosomal aberration involving RERE is found in the neuroblastoma cell line NGP. Translocation t(1;15)(p36.2;q24).; Neurodevelopmental disorder with or without anomalies of the brain, eye, or heart (NEDBEH) [MIM:616975]: An autosomal dominant syndrome characterized by developmental delay, intellectual disability, brain anomalies, and neurological abnormalities including seizures, hypotonia, and behavioral problems such as autism spectrum disorders. Brain anomalies include abnormalities and/or thinning of the corpus callosum, diminished white matter volume, abnormal cerebellar vermis, and ventriculomegaly. Congenital defects of the eye, heart and genitourinary system are present in half of the patients.
REST	Wilms tumor 6 (WT6) [MIM:616806]: A pediatric malignancy of kidney, and the most common childhood abdominal malignancy. It is caused by the uncontrolled multiplication of renal stem, stromal, and epithelial cells.; Fibromatosis, gingival, 5 (GINGF5) [MIM:617626]: An autosomal dominant form of hereditary gingival fibromatosis, a rare condition characterized by a slow, progressive overgrowth of the gingiva. The excess gingival tissue can cover part of or the entire crown and can result in diastemas, teeth displacement, or retention of primary or impacted teeth.

RET4	Retinal dystrophy, iris coloboma, and comedogenic acne syndrome (RDCCAS) [MIM:615147]: A disease characterized by retinal degeneration, ocular colobomas involving both the anterior and posterior segment, impaired night vision and loss of visual acuity. Additional characteristic features include developmental abnormalities and severe acne. Loss of functional RBP4 protein results in serum retinol deficiency. Lack of normal levels of retinol impairs the visual cycle leading to night blindness at early stages; prolonged deficiency may lead to retinal degeneration. Additionally, retinol deficiency may result in dry skin, increased susceptibility to infection and acne (PubMed:23189188).; Microphthalmia, isolated, with coloboma, 10 (MCOPCB10) [MIM:616428]: A disorder of eye formation, ranging from small size of a single eye to complete bilateral absence of ocular tissues. Ocular abnormalities like opacities of the cornea and lens, scaring of the retina and choroid, and other abnormalities may also be present. Ocular colobomas are a set of malformations resulting from abnormal morphogenesis of the optic cup and stalk and the fusion of the fetal fissure (optic fissure).
RT4I1	Optic atrophy 10 with or without ataxia, mental retardation, and seizures (OPA10) [MIM:616732]: An autosomal recessive disease characterized by a progressive visual loss in association with optic atrophy. Atrophy of the optic disk indicates a deficiency in the number of nerve fibers which arise in the retina and converge to form the optic disk, optic nerve, optic chiasm, and optic tracts. OPA10 patients may also manifest mild ataxia, mild mental retardation and, rarely, generalized seizures.
S12A1	Bartter syndrome 1, antenatal (BARTS1) [MIM:601678]: A form of Bartter syndrome, an autosomal recessive disorder characterized by impaired salt reabsorption in the thick ascending loop of Henle with pronounced salt wasting, hypokalemic metabolic alkalosis, and varying degrees of hypercalciuria. BARTS1 is a life-threatening condition beginning in utero, with marked fetal polyuria that leads to polyhydramnios and premature delivery. Another hallmark is a marked hypercalciuria and, as a secondary consequence, the development of nephrocalcinosis and osteopenia.
S12A5	Epileptic encephalopathy, early infantile, 34 (EIEE34) [MIM:616645]: A form of epileptic encephalopathy, a heterogeneous group of severe childhood-onset epilepsies characterized by refractory seizures, neurodevelopmental impairment, and poor prognosis. Development is normal prior to seizure onset, after which cognitive and motor delays become apparent. EIEE34 is characterized by the onset of refractory migrating focal seizures in infancy. Affected children show developmental regression and are severely impaired globally.; Epilepsy, idiopathic generalized 14 (EIG14) [MIM:616685]: An autosomal dominant form of idiopathic generalized epilepsy, a disorder characterized by recurring generalized seizures in the absence of detectable brain lesions and/or metabolic abnormalities. Generalized seizures arise diffusely and simultaneously from both hemispheres of the brain. Seizure types include juvenile myoclonic seizures, absence seizures, and generalized tonic-clonic seizures.
S12A8	Has been identified as a possible susceptibility gene for psoriasis mapped to chromosome 3q21 (PSORS5).
SALL2	Coloboma, ocular, autosomal recessive (COAR) [MIM:216820]: An ocular anomaly resulting from abnormal morphogenesis of the optic cup and stalk, and incomplete fusion of the fetal intra-ocular fissure during gestation. The clinical presentation is variable. Some individuals may present with minimal defects in the anterior iris leaf without other ocular defects. More complex malformations create a combination of iris, uveoretinal and/or optic nerve defects without or with microphthalmia or even anophthalmia.
SC23A	Craniolenticulosutural dysplasia (CLSD) [MIM:607812]: Autosomal recessive syndrome characterized by late-closing fontanels, sutural cataracts, facial dysmorphisms, and skeletal defects.
SC23B	Cowden syndrome 7 (CWS7) [MIM:616858]: A form of Cowden syndrome, a hamartomatous polyposis syndrome with age-related penetrance. Cowden syndrome is characterized by hamartomatous lesions affecting derivatives of ectodermal, mesodermal and endodermal layers, macrocephaly, facial trichilemmomas (benign tumors of the hair follicle infundibulum), acral keratoses, papillomatous papules, and elevated risk for development of several types of malignancy, particularly breast carcinoma in women and thyroid carcinoma in both men and women. Colon cancer and renal cell carcinoma have also been reported. Hamartomas can be found in virtually every organ, but most commonly in the skin, gastrointestinal tract, breast, and thyroid. CWS7 inheritance is autosomal dominant.; Anemia, congenital dyserythropoietic, 2 (CDAN2) [MIM:224100]: An autosomal recessive blood disorder characterized by morphological abnormalities of erythroblasts, ineffective erythropoiesis, normocytic anemia, iron overload, jaundice, and variable splenomegaly. Ultrastructural features include bi- or multinucleated erythroblasts in bone marrow, karyorrhexis, and the presence of Gaucher-like bone marrow histiocytes. The main biochemical feature of the disease is defective glycosylation of some red blood cells membrane proteins.
SC5A5	Thyroid dyshormonogenesis 1 (TDH1) [MIM:274400]: A disorder characterized by the inability of the thyroid to maintain a concentration difference of readily exchangeable iodine between the plasma and the thyroid gland, leading to congenital hypothyroidism.

SDHA	Mitochondrial complex II deficiency (MT-C2D) [MIM:252011]: A disorder of the mitochondrial respiratory chain with heterogeneous clinical manifestations. Clinical features include psychomotor regression in infants, poor growth with lack of speech development, severe spastic quadriplegia, dystonia, progressive leukoencephalopathy, muscle weakness, exercise intolerance, cardiomyopathy. Some patients manifest Leigh syndrome or Kearns-Sayre syndrome. Note=The disease is caused by mutations affecting the gene represented in this entry. ;Leigh syndrome (LS) [MIM:256000]: An early-onset progressive neurodegenerative disorder characterized by the presence of focal, bilateral lesions in one or more areas of the central nervous system including the brainstem, thalamus, basal ganglia, cerebellum, and spinal cord. Clinical features depend on which areas of the central nervous system are involved and include subacute onset of psychomotor retardation, hypotonia, ataxia, weakness, vision loss, eye movement abnormalities, seizures, and dysphagia.; Cardiomyopathy, dilated 1GG (CMD1GG) [MIM:613642]: A disorder characterized by ventricular dilation and impaired systolic function, resulting in congestive heart failure and arrhythmia. Patients are at risk of premature death; Paragangliomas 5 (PGL5) [MIM:614165]: A neural crest tumor usually derived from the chromoreceptor tissue of a paraganglion. Paragangliomas can develop at various body sites, including the head, neck, thorax, and abdomen. Most commonly, they are located in the head and neck region, specifically at the carotid bifurcation, the jugular foramen, the vagal nerve, and in the middle ear.
SDHB	Pheochromocytoma (PCC) [MIM:171300]: A catecholamine-producing tumor of chromaffin tissue of the adrenal medulla or sympathetic paraganglia. The cardinal symptom, reflecting the increased secretion of epinephrine and norepinephrine, is hypertension, which may be persistent or intermittent.; Paragangliomas 4 (PGL4) [MIM:115310]: A neural crest tumor usually derived from the chromoreceptor tissue of a paraganglion. Paragangliomas can develop at various body sites, including the head, neck, thorax, and abdomen. Most commonly, they are located in the head and neck region, precisely at the carotid bifurcation, the jugular foramen, the vagal nerve, and in the middle ear; Paraganglioma and gastric stromal sarcoma (PGGSS) [MIM:606864]: Gastrointestinal stromal tumors may be sporadic or inherited in an autosomal dominant manner, alone or as a component of a syndrome associated with other tumors, such as in the context of neurofibromatosis type 1 (NF1). Patients have both gastrointestinal stromal tumors and paragangliomas. Susceptibility to the tumors was inherited in an apparently autosomal dominant manner, with incomplete penetrance.
SEC63	Polycystic liver disease 2 with or without kidney cysts (PCLD2) [MIM:617004]: An autosomal dominant hepatobiliary disease characterized by overgrowth of biliary epithelium and supportive connective tissue, resulting in multiple liver cysts. A subset of patients may develop kidney cysts that usually do not result in clinically significant renal disease.
SPRE1	Neurofibromatosis 1-like syndrome (NFLS) [MIM:611431]: A disorder characterized mainly by cafe au lait macules without neurofibromas or other tumor manifestations of neurofibromatosis type 1, axillary freckling, and macrocephaly. Additional clinical manifestations include Noonan-like facial dysmorphism, lipomas, learning disabilities, and attention deficit-hyperactivity.
STX1A	Note=STX1A is located in the Williams-Beuren syndrome (WBS) critical region. WBS results from a hemizygous deletion of several genes on chromosome 7q11.23 thought to arise as a consequence of unequal crossing over between highly homologous low-copy repeat sequences flanking the deleted region.
STX1B	Generalized epilepsy with febrile seizures plus 9 (GEFSP9) [MIM:616172]: An autosomal dominant neurologic disorder characterized by febrile and/or afebrile seizures manifesting in early childhood. Seizure is variable and include generalized tonic-clonic, atonic, myoclonic, complex partial, and absence types. Most patients have remission of seizures later in childhood with no residual neurologic deficits. Rarely, patients may show mild developmental delay or mild intellectual disabilities.
STXB1	Epileptic encephalopathy, early infantile, 4 (EIEE4) [MIM:612164]: A severe form of epilepsy characterized by frequent tonic seizures or spasms beginning in infancy with a specific EEG finding of suppression-burst patterns, characterized by high-voltage bursts alternating with almost flat suppression phases. Affected individuals have the neonatal or infantile onset of seizures, profound mental retardation, and MRI evidence of brain hypomyelination.
SUN5	Spermatogenic failure 16 (SPGF16) [MIM:617187]: An infertility disorder caused by spermatogenesis defects and characterized by abnormally shaped spermatozoa in the semen of affected individuals. Most spermatozoa are made up of headless tails, while a small proportion has an abnormal head-tail junction. A few spermatozoa are made up of tailless heads.
SYAC	Charcot-Marie-Tooth disease 2N (CMT2N) [MIM:613287]: An axonal form of Charcot-Marie-Tooth disease, a disorder of the peripheral nervous system, characterized by progressive weakness and atrophy, initially of the peroneal muscles and later of the distal muscles of the arms. Charcot-Marie-Tooth disease is classified in two main groups on the basis of electrophysiologic properties and histopathology: primary peripheral demyelinating neuropathies (designated CMT1 when they are dominantly inherited) and primary peripheral axonal neuropathies (CMT2). Neuropathies of the CMT2 group are characterized by signs of axonal degeneration in the absence of apparent myelin alterations, normal or slightly reduced nerve conduction velocities, and progressive distal muscle weakness and atrophy.; Epileptic encephalopathy, early infantile, 29 (EIEE29) [MIM:616339]: A form of epileptic encephalopathy, a heterogeneous group of severe childhood-onset epilepsies characterized by refractory seizures, neurodevelopmental impairment, and poor prognosis. Development is normal prior to seizure onset, after which cognitive and motor delays become apparent. EIEE29 patients manifest severe infantile epileptic encephalopathy, clubfoot, absent deep tendon reflexes, extrapyramidal symptoms, and persistently deficient myelination.

TBX19	ACTH deficiency, isolated (IAD) [MIM:201400]: An autosomal recessive disorder that is characterized by adrenal insufficiency symptoms, such as weight loss, lack of appetite (anorexia), weakness, nausea, vomiting and low blood pressure (hypotension). The pituitary hormone ACTH is decreased or absent, and other cortisol and other steroid hormone levels in the blood are abnormally low.
TGIF1	Holoprosencephaly 4 (HPE4) [MIM:142946]: A structural anomaly of the brain, in which the developing forebrain fails to correctly separate into right and left hemispheres. Holoprosencephaly is genetically heterogeneous and associated with several distinct facies and phenotypic variability.
TMLH	Autism, X-linked 6 (AUTSX6) [MIM:300872]: A form of autism, a complex multifactorial, pervasive developmental disorder characterized by impairments in reciprocal social interaction and communication, restricted and stereotyped patterns of interests and activities, and the presence of developmental abnormalities by 3 years of age. Most individuals with autism also manifest moderate mental retardation. AUTSX6 patients may respond favorably to carnitine supplementation.
TNNI2	Arthrogryposis, distal, 2B (DA2B) [MIM:601680]: A form of distal arthrogryposis, a disease characterized by congenital joint contractures that mainly involve two or more distal parts of the limbs, in the absence of a primary neurological or muscle disease. DA2B is characterized by contractures of the hands and feet, and a distinctive face characterized by prominent nasolabial folds, small mouth, and downslanting palpebral fissures.
TPM2	Nemaline myopathy 4 (NEM4) [MIM:609285]: A form of nemaline myopathy. Nemaline myopathies are muscular disorders characterized by muscle weakness of varying severity and onset, and abnormal thread-like or rod-shaped structures in muscle fibers on histologic examination. Nemaline myopathy type 4 presents from infancy to childhood with hypotonia and moderate-to-severe proximal weakness with minimal or no progression. Major motor milestones are delayed, but independent ambulation is usually achieved, although a wheelchair may be needed in later life.; Arthrogryposis, distal, 1A (DA1A) [MIM:108120]: A form of distal arthrogryposis, a disease characterized by congenital joint contractures that mainly involve two or more distal parts of the limbs, in the absence of a primary neurological or muscle disease. Distal arthrogryposis type 1 is primarily characterized by camptodactyly and clubfoot. Hypoplasia and/or absence of some interphalangeal creases is common. The shoulders and hips are less frequently affected.; Cap myopathy 2 (CAPM2) [MIM:609285]: A rare congenital skeletal muscle disorder characterized by the presence of cap-like structures which are well demarcated and peripherally located under the sarcolemma and show abnormal accumulation of sarcomeric proteins. Clinical features are early onset of hypotonia and non-progressive or slowly progressive muscle weakness. Respiratory problems are common.; Arthrogryposis, distal, 2B (DA2B) [MIM:601680]: A form of distal arthrogryposis, a disease characterized by congenital joint contractures that mainly involve two or more distal parts of the limbs, in the absence of a primary neurological or muscle disease. DA2B is characterized by contractures of the hands and feet, and a distinctive face characterized by prominent nasolabial folds, small mouth, and downslanting palpebral fissures.
TPM3	Nemaline myopathy 1 (NEM1) [MIM:609284]: A form of nemaline myopathy with autosomal dominant or recessive inheritance. Nemaline myopathies are disorders characterized by muscle weakness of different onset and severity, and abnormal thread-like or rod-shaped structures in muscle fibers on histologic examination. Autosomal dominant NEM1 is characterized by a moderate phenotype with onset between birth and early second decade of life. Weakness is diffuse and symmetric with slow progression often with the need for a wheelchair in adulthood. The autosomal recessive form has onset at birth with moderate to severe hypotonia and diffuse weakness. In the most severe cases, death can occur before 2 years. Less severe cases have delayed major motor milestones, and these patients may walk, but often need a wheelchair before 10 years.; Note=A chromosomal aberration involving TPM3 is found in papillary thyroid carcinomas (PTCs). A rearrangement with NTRK1 generates the TRK fusion transcript by fusing the amino end of isoform 2 of TPM3 to the 3'-end of NTRK1.; Myopathy, congenital, with fiber-type disproportion (CFTD) [MIM:255310]: A genetically heterogeneous disorder in which there is relative hypotrophy of type 1 muscle fibers compared to type 2 fibers on skeletal muscle biopsy. However, these findings are not specific and can be found in many different myopathic and neuropathic conditions.; Cap myopathy 1 (CAPM1) [MIM:609284]: A rare congenital skeletal muscle disorder characterized by the presence of cap-like structures which are well demarcated and peripherally located under the sarcolemma and show abnormal accumulation of sarcomeric proteins. Clinical features are early onset of hypotonia and slowly progressive muscle weakness. Respiratory problems are common.
TREA	Trehalase deficiency (TREHD) [MIM:612119]: An autosomal recessive condition characterized by the inability to digest trehalose, a disaccharide found in mushrooms, products containing baker's yeast, and dried food. Individuals with trehalase deficiency suffer from abdominal pain, increased rectal flatulence, and diarrhea due to osmotic water flow into the colon.
VINC	Cardiomyopathy dilated 1W (CMD1W) [MIM:611407]: A disorder characterized by ventricular dilation and impaired systolic function, resulting in congestive heart failure and arrhythmia. Patients are at risk of premature death; Cardiomyopathy, familial hypertrophic 15 (CMH15) [MIM:613255]: A hereditary heart disorder characterized by ventricular hypertrophy, which is usually asymmetric and often involves the interventricular septum. The symptoms include dyspnea, syncope, collapse, palpitations, and chest pain. They can be readily provoked by exercise. The disorder has inter- and intrafamilial variability ranging from benign to malignant forms with high risk of cardiac failure and sudden cardiac death.

VP33A	Mucopolysaccharidosis-plus syndrome (MPSPS) [MIM:617303]: A form of mucopolysaccharidosis, a group of diseases characterized by excessive accumulation and secretion of oligomucopoloysaccharides. MPSPS is a multisystemic disorder characterized by coarse facial features, dysostosis multiplex, hepatosplenomegaly, respiratory difficulties, mental retardation, developmental delay, pyramidal signs, severe chronic anemia, renal involvement, and cardiac defects. Laboratory analyses show proteinuria with foamy glomerular cells, excess secretion of urinary glycosaminoglycans, and extremely high levels of plasma heparan sulfate. Disease onset is in infancy. Most patients die in the first years of life due to cardiorespiratory failure. MPSPS inheritance is autosomal recessive.
VP33B	Arthrogryposis, renal dysfunction and cholestasis syndrome 1 (ARCS1) [MIM:208085]: A multisystem disorder, characterized by neurogenic arthrogryposis multiplex congenita, renal tubular dysfunction and neonatal cholestasis with bile duct hypoplasia and low gamma glutamyl transpeptidase activity. Platelet dysfunction is common.
VPS35	Parkinson disease 17 (PARK17) [MIM:614203]: An autosomal dominant, adult-onset form of Parkinson disease. Parkinson disease is a complex neurodegenerative disorder characterized by bradykinesia, resting tremor, muscular rigidity, and postural instability, as well as by a clinically significant response to treatment with levodopa. The pathology involves the loss of dopaminergic neurons in the substantia nigra and the presence of Lewy bodies (intraneuronal accumulations of aggregated proteins), in surviving neurons in various areas of the brain.
VSX2	Microphthalmia, isolated, 2 (MCOP2) [MIM:610093]: A disorder of eye formation, ranging from small size of a single eye to complete bilateral absence of ocular tissues. Ocular abnormalities like opacities of the cornea and lens, scaring of the retina and choroid, and other abnormalities may also be present. {ECO:0000269|PubMed:15257456, ECO:0000269|PubMed:21976963}. Note=The disease is caused by mutations affecting the gene represented in this entry.; Microphthalmia with cataracts and iris abnormalities (MCOPCTI) [MIM:610092]: A disorder of eye formation, ranging from small size of a single eye to complete bilateral absence of ocular tissues. Ocular abnormalities like opacities of the cornea and lens, scaring of the retina and choroid, cataract and other abnormalities like cataract may also be present. {ECO:0000269|PubMed:10932181}. Note=The disease is caused by mutations affecting the gene represented in this entry.; Microphthalmia, isolated, with coloboma, 3 (MCOPCB3) [MIM:610092]: A disorder of eye formation, ranging from small size of a single eye to complete bilateral absence of ocular tissues. Ocular abnormalities like opacities of the cornea and lens, scaring of the retina and choroid, and other abnormalities may also be present. Ocular colobomas are a set of malformations resulting from abnormal morphogenesis of the optic cup and stalk and the fusion of the fetal fissure (optic fissure).
WIPF1	Wiskott-Aldrich syndrome 2 (WAS2) [MIM:614493]: An immunodeficiency disorder characterized by eczema, thrombocytopenia, recurrent infections, defective T-cell proliferation, and impaired natural killer cell function.