Makina Öğrenmesi ile Sinirsel Gelişim Hastalıkları için Gen Keşfi

A. Ercüment Çiçek

Yar. Doç. Dr. , Bilgisayar Mühendisliği Bolümü, Bilkent Üniversitesi

Adjunct Fakülte Üyesi, Computational Biology Department, School of Computer Science, Carnegie Mellon University

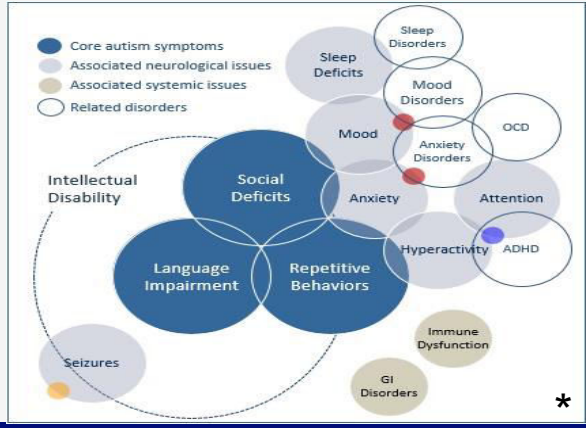
Yol Haritası

- Otizm ve sinirsel gelişim hastalıkları
- Gen keşfi problemi
- Literatürdeki algoritmalar
- ST-Steiner: Zaman-mekansal gen keşfi algoritması

Otizm

- Δ Semptomlar: Konuşma ve sosyal iletişim bozukluğu, tekrar eden hareketler
- Δ Geniş spektrum
- Δ ABD'de her 56 çocuktan 1'inde görülüyor
- Δ Genetik mimari:

Yüksek seviyede kalıtsal ve heterojen



*Resim: Autism Speaks

Gen Keşfi

Analysis of Rare, Exonic Variation amongst Subjects with Autism Spectrum Disorders and Population Controls

Li Liu, Aniko Sabo, Benjamin M. Neale, Uma Nagaswamy, Christine Stevens, Elaine Lim, Corneliu A. Bodea, Donna Muzny, Jeffrey G. Reid, Eric Banks, Hillary Coon, Mark DePristo, Huyen Dinh, [···], Kathryn Roeder [view all]

Published: April 11, 2013 • http://dx.doi.org/10.1371/journal.pgen.1003443

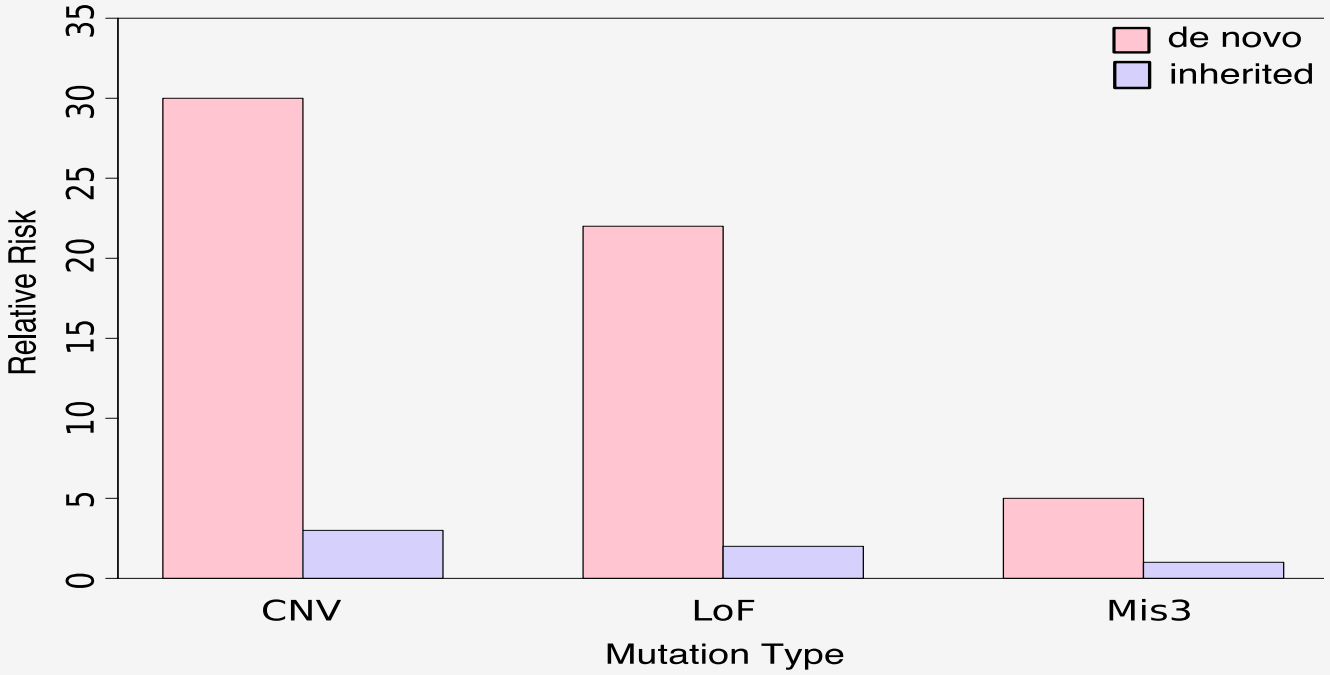
Abstract

We report on results from whole-exome sequencing (WES) of 1,039 subjects diagnosed with autism spectrum disorders (ASD) and 870 controls selected from the NIMH repository to be of similar ancestry to cases. The WES data came from two centers using different methods to produce sequence and to call variants from it. Therefore, an initial goal was to ensure the distribution of rare variation was similar for data from different centers. This proved straightforward by filtering called variants by fraction of missing data, read depth, and balance of alternative to reference reads. Results were evaluated using seven samples sequenced at both centers and by results from the association study. Next we addressed how the data and/or results from the centers should be combined. Gene-based analyses of association was an obvious choice, but should statistics for association be combined across centers (meta-analysis) or should data be combined and then analyzed (mega-analysis)? Because of the nature of many gene-based tests, we showed by theory and simulations that mega-analysis has better power than meta-analysis. Finally, before analyzing the data for association, we explored the impact of population structure on rare variant analysis in these data. Like other recent studies, we found evidence that population structure can confound case-control studies by the clustering of rare variants in ancestry space; yet, unlike some recent studies, for these data we found that principal component-based analyses were sufficient to control for ancestry and produce test statistics with appropriate distributions. After using a variety of gene-based tests and both meta- and mega-analysis, we found no new risk genes for ASD in this sample. Our results suggest that standard gene-based tests will require much larger samples of cases and controls before being effective for gene discovery, even for a disorder like ASD.

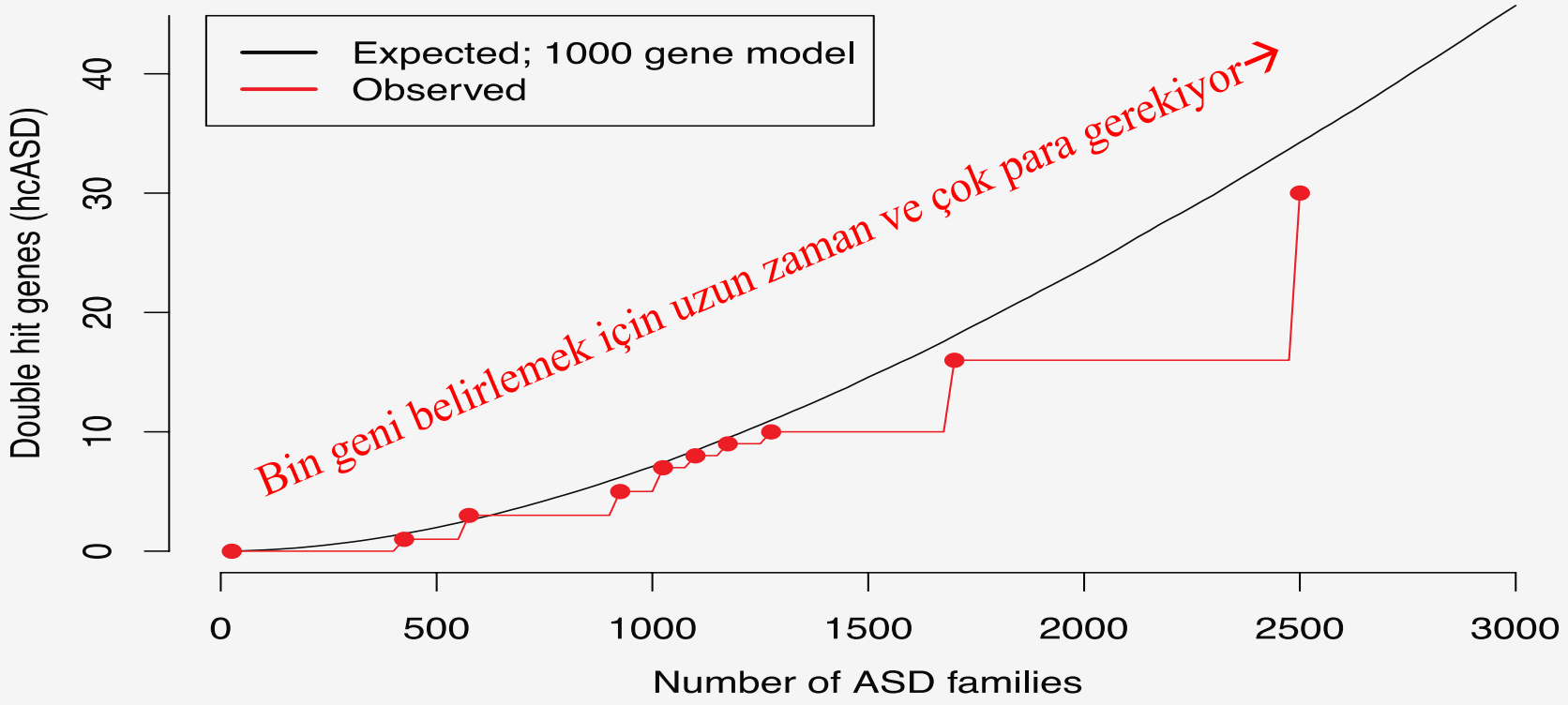
PLOS | GENETICS

de novo mutasyonların önemi

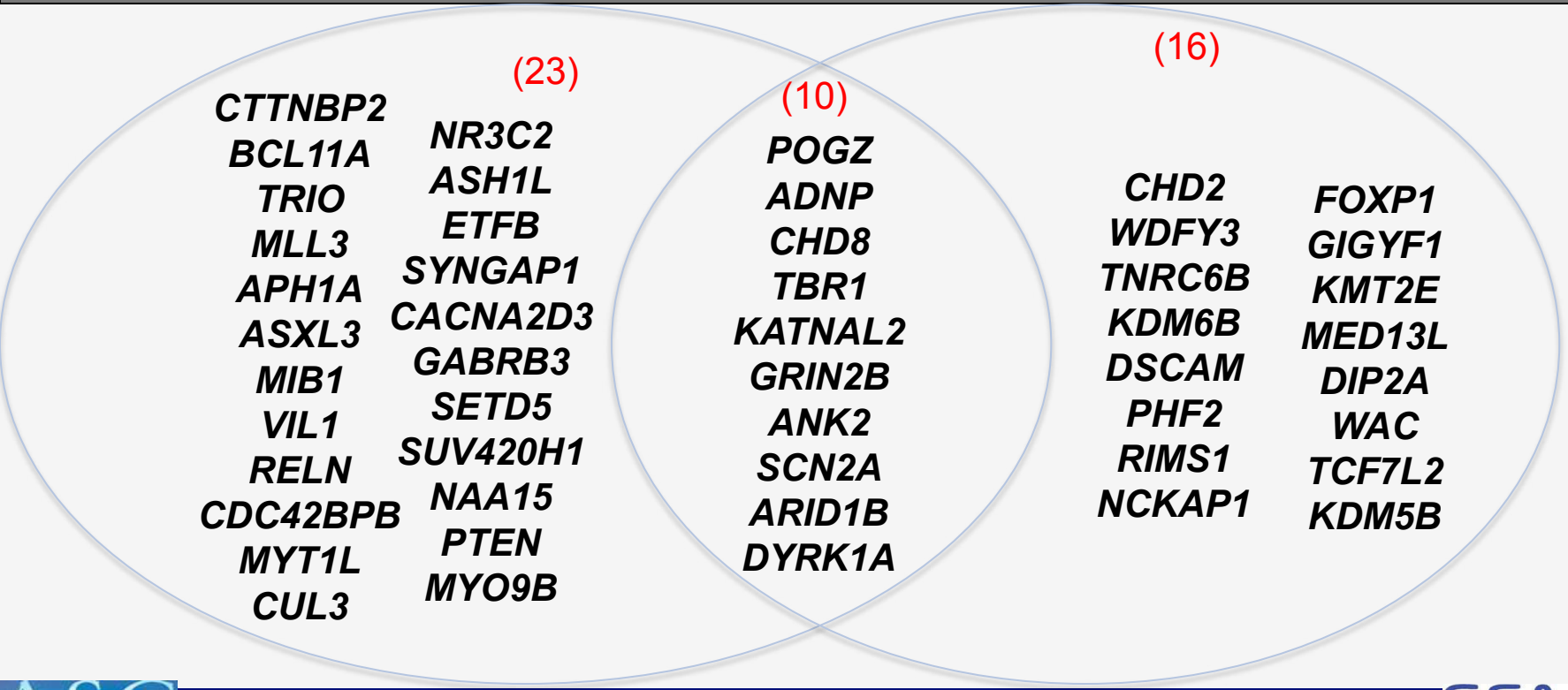
Mike Wigler Lab



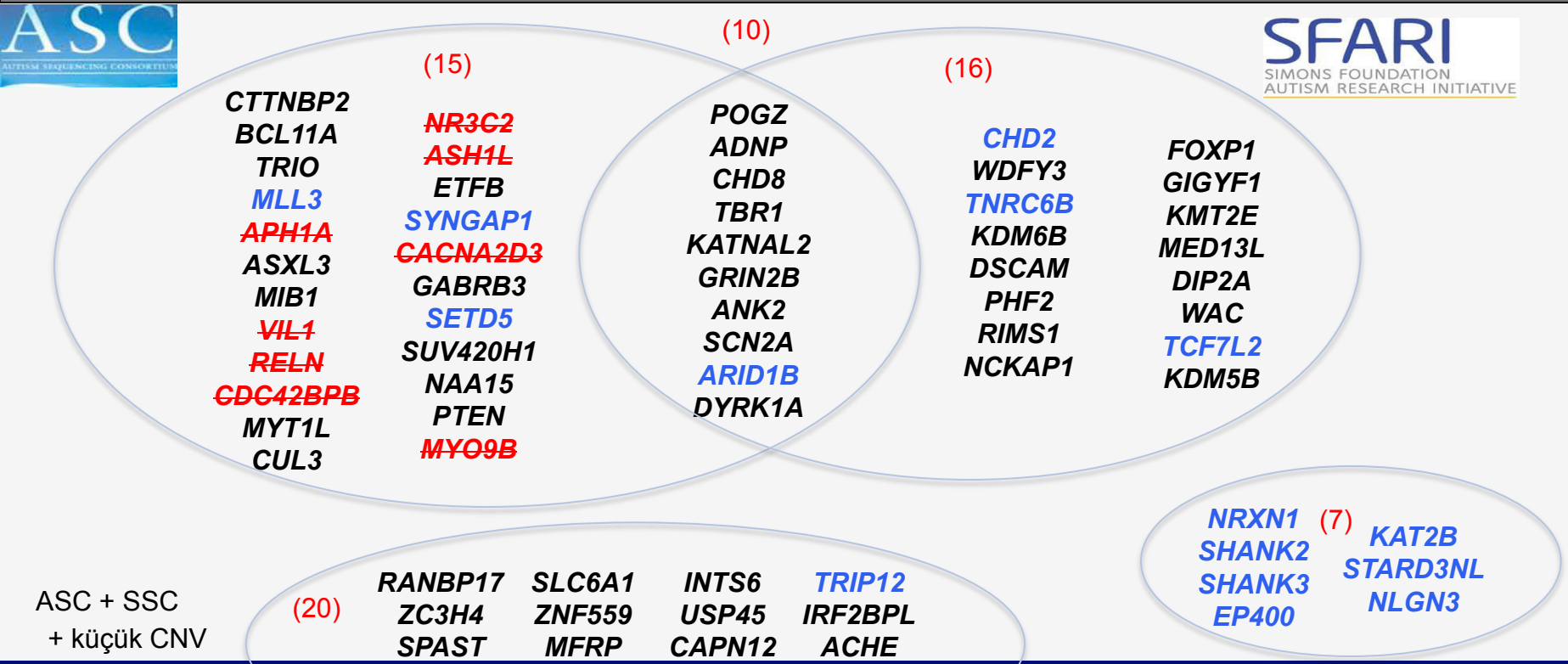
Gen Keşfi: *de novo* LoF mutasyonlar



Gen keşfi: *Son durum 2014*



Gen keşfi: *Son durum 2015*



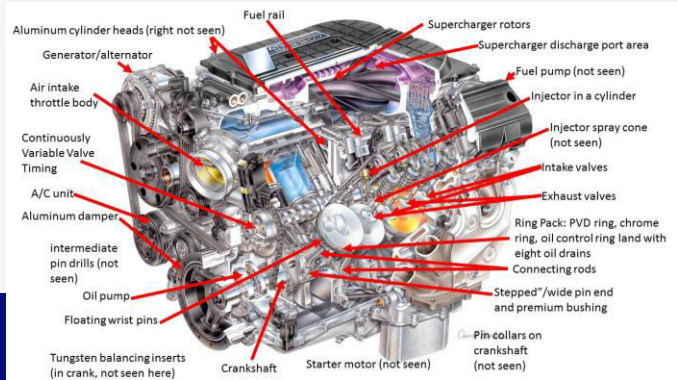
Slide by Stephan Sanders
Sanders et al., Neuron, 2015

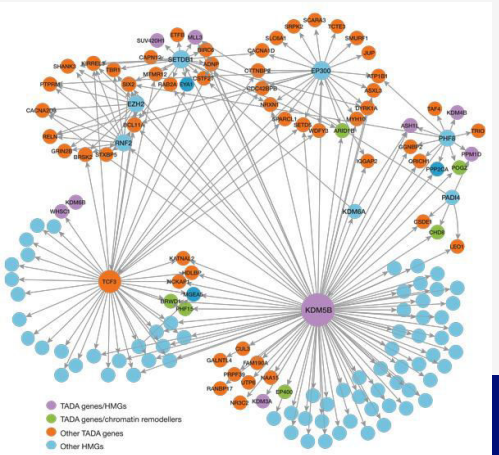
Gen Keşfi için algoritmalar

Kabullenim: Risk genleri biyolojik ağlarda fonksiyonel bir küme oluşturur.

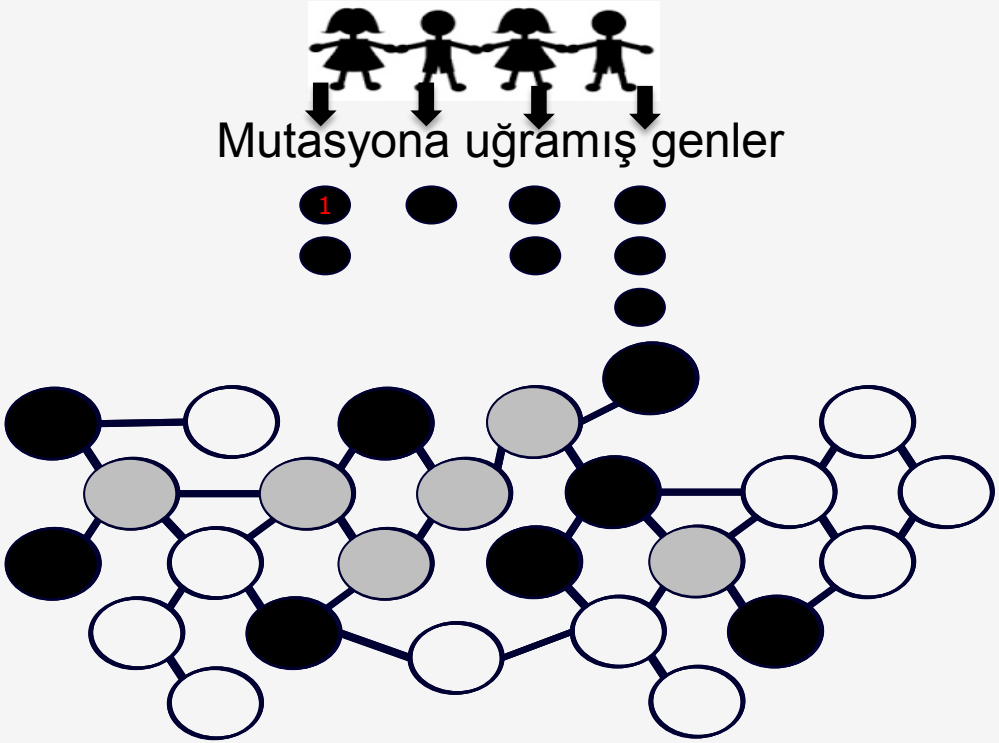
Guilt-by-association prensibi:

1. Bilinen ASD genlerini başlangıç noktası olarak kullan.
2. Sıkı ilişki içinde bir gen kümesi keşfet.





Gen Keşfi için algoritmalar



Gen Keşfi için algoritmalar

- **NETBAG**[1], NETwork-Based Analysis of Genetic Associations
- **DAWN**[2], Detecting Association With Networks
- **MAGI**[3], Merging Affected Genes into Integrated networks
- **ST-Steiner**[4], Spatio-Temporal Gene Discovery for ASD

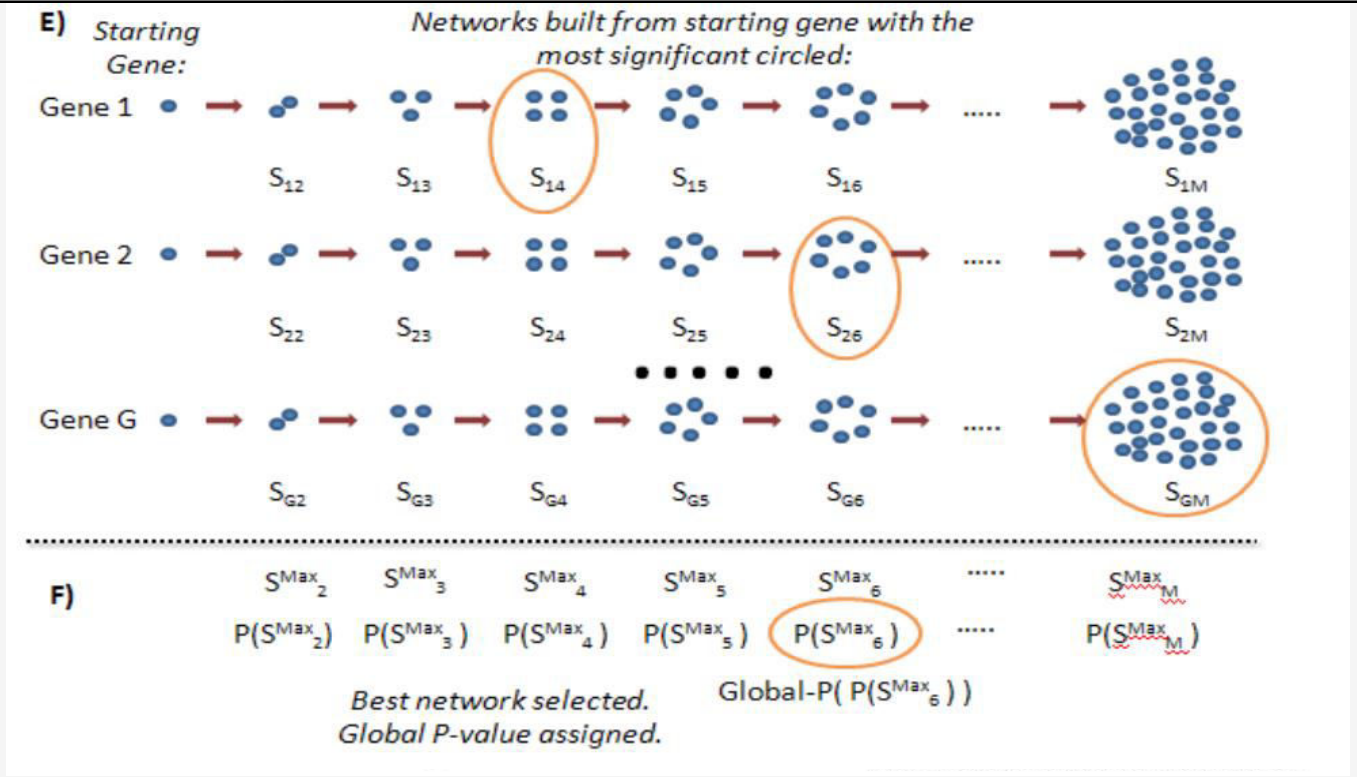
[1] Gilman et al. 2011, “Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses”. *Neuron*.

[2] Liu et al. 2014, “DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics”. *Molecular Autism*.

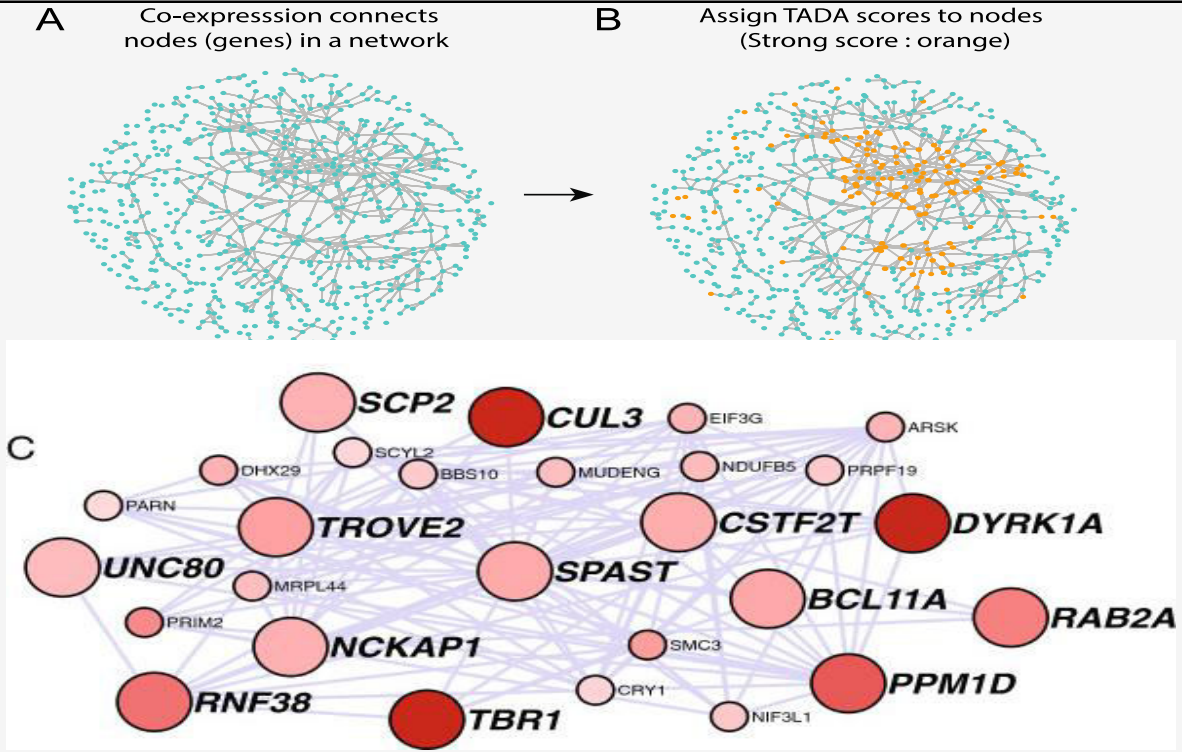
[3] Hormozdiari et al. 2015, “The discovery of integrated gene networks for autism and related disorders”. *Genome Research*.

[4] Norman and Cicek, 2018, Spatio-Temporal Gene Discovery for Autism Spectrum Disorder”. *bioRxiv 2018.*

NETBAG Algoritması

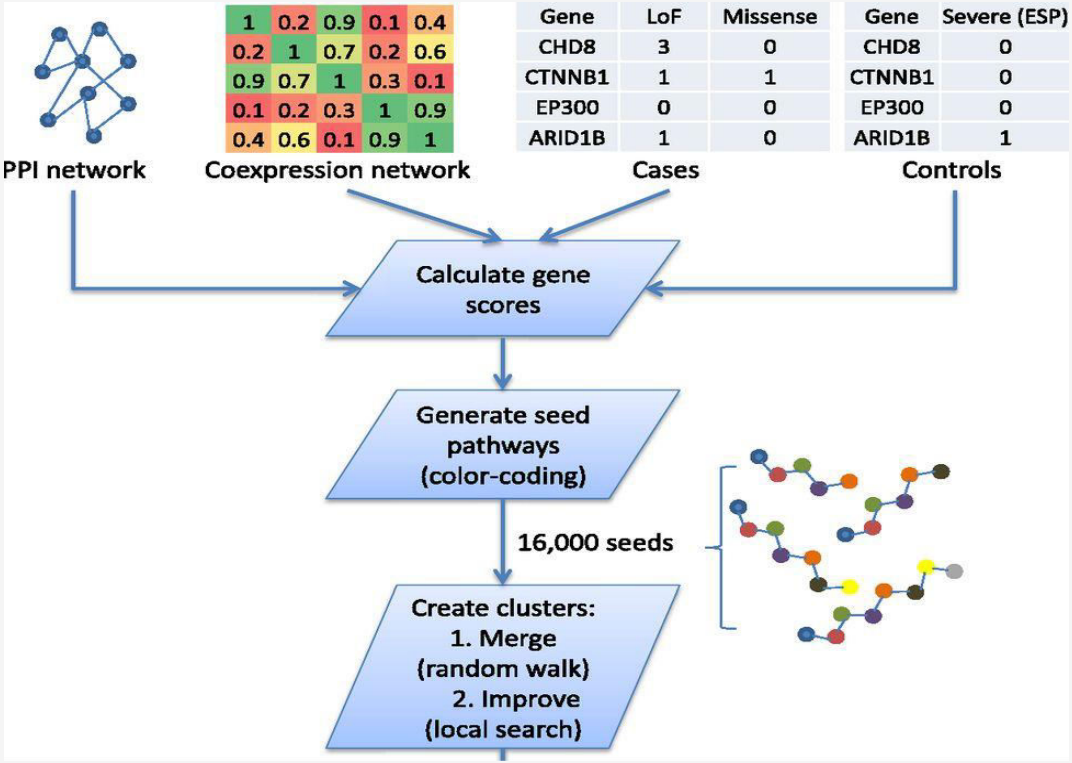


DAWN Algoritması



Liu et al. 2014, “DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics”. *Molecular Autism*.

MAGI Algoritması



1	0.2	0.9	0.1	0.4
0.2	1	0.7	0.2	0.6
0.9	0.7	1	0.3	0.1
0.1	0.2	0.3	1	0.9
0.4	0.6	0.1	0.9	1

Gene	LoF	Missense
CHD8	3	0
CTNNB1	1	1
EP300	0	0
ARID1B	1	0

Gene	Severe (ESP)
CHD8	0
CTNNB1	0
EP300	0
ARID1B	1

Figure from Hormozdiari et al. 2015, “The discovery of integrated gene networks for autism and related disorders”. *Genome Research*.

MAGI Algoritması

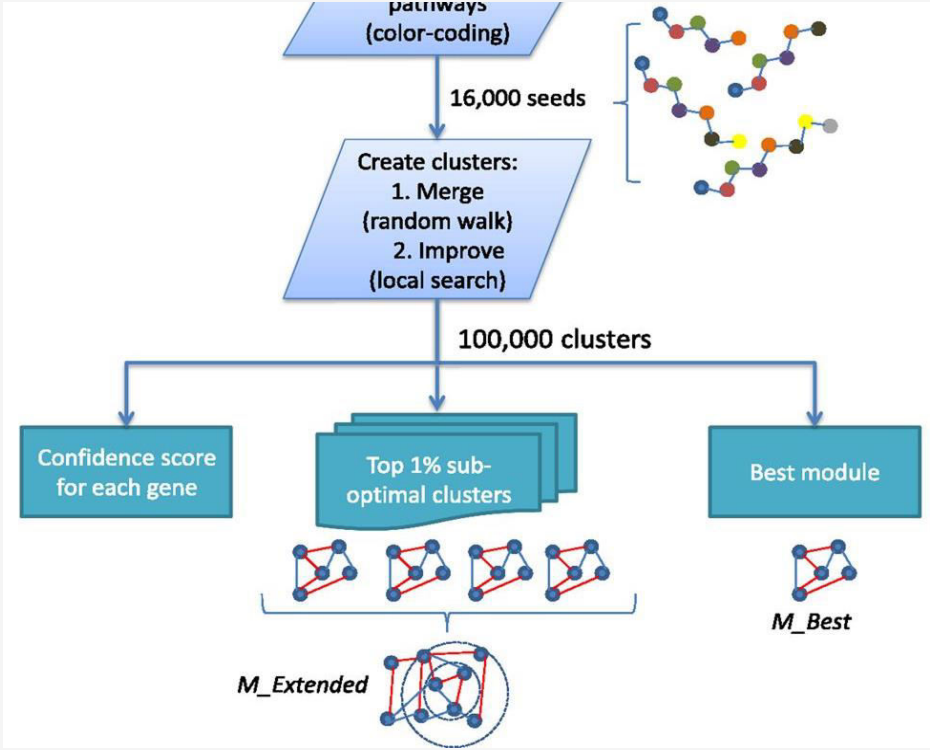


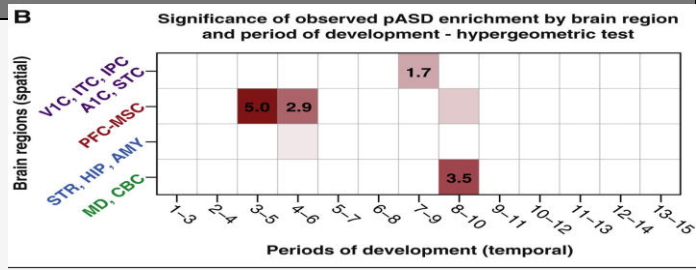
Figure from Hormozdiari et al. 2015, “The discovery of integrated gene networks for autism and related disorders”. *Genome Research*.

Eksiklikler

Gen etkileşimleri zamanla değişir

Farklı beyin gelişim pencereleri:

- Farklı topolojilere sahiptir,
- ASD gen kümelenmesi de zaman ve mekana göre değişir.



Willsey et. al., 2013, Cell 155.

Erken dönemlerde, biyolojik patikalardaki problemler ileriki dönemlere ardışık olarak etki eder.

- Erken donem patika hasarı ileri dönemleri etkiler ve 36 kat ASD risk artısına neden olur*
- Yetişkinlikteki problemler ise risk artışına neden olmaz*.

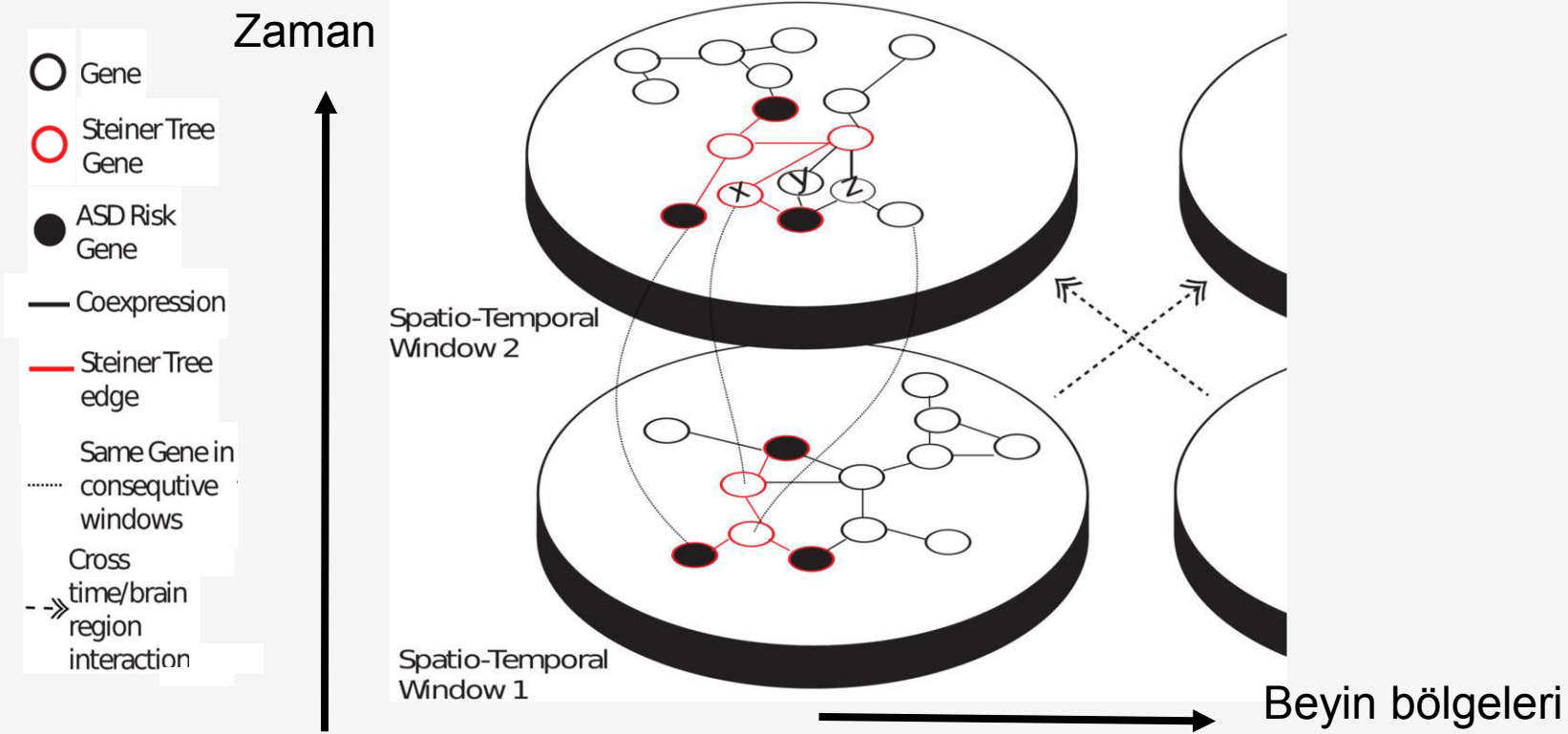
* Wang et. al., 2014, Neuron 83(3).

Eksiklikler

Bu nedenle statik biyolojik ağlar kullanan metotlar:
Dinamik beyin gelişimini modelleyemez
ve tahmin güçleri limitlidir.

Hipotezimiz:
“Gen kümelenmesi statik değil dinamiktir.”

ST-Steiner Algoritması



Methodlar

Statik biyolojik ağların yarattığı sorunları aşmayı hedefliyoruz:

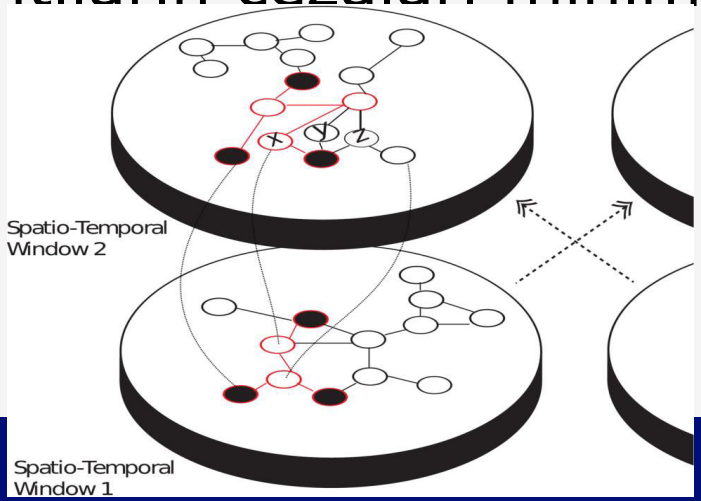
- Prize-collecting Steiner tree (PCST) problemi kullanacağız.
- Zaman ve mekan bilgisini de algoritmaya vereceğiz.

Orijinal Steiner tree (ağaç) problemi: Verilen bir ağda, ayrıcalıklı, tohum noktaları bağlayan minimum bir ağaç bulur.

Prize-Collecting Steiner Tree (PCST)

Prize-collecting Steiner tree problemi:
Öyle bir ağaç bul ki:

- Seçilen nodların ödülleri maksimum.
- Seçilen bağlantıların cezaları minimum olsun.



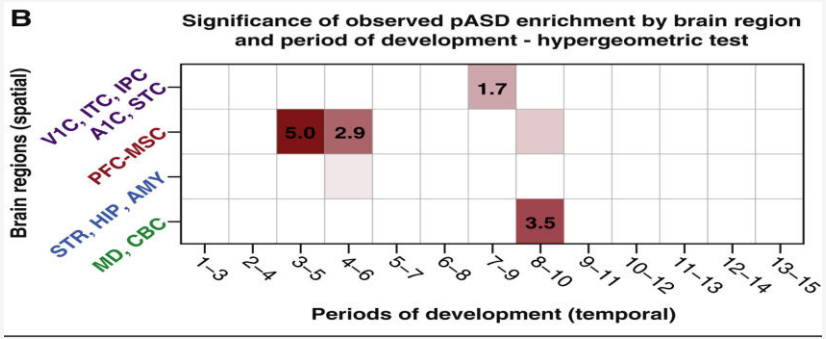
Prize-Collecting Steiner Tree (PCST)

$G(V, E)$, a nodları ve bağlantıları ağırlıklı bir ağ olsun
Nod seti V, bağlantı seti E, bağlantı cezası fonksiyonu $c(e) \geq 0$, nod ödülü fonksiyonu $p(v) \geq 0$

Aranan ağaç $T(V_T, E_T)$ bu fonksiyonu minimize eder:

$$o_T(T) = \sum_{e \in E_T} c(e) + \beta \sum_{v \notin V_T} p(v) \quad \beta \geq 0$$

Problem Formulasyonu



Willsey et. al., 2013, Cell 155.

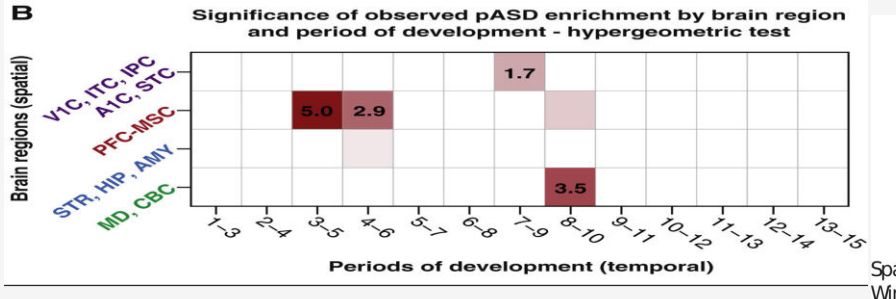
- Zaman-mekânsal sistem $G = (G_1, G_2, \cdots, G_T)$, T zaman penceresi
- i. pencere, $G_i = \{ \mathrm{G}_i^1, G_i^2 \cdots, G_i^n \}$ n tane beyin bölgesi içerir.
- Buradan bir alt sistem bulmak istiyoruz.

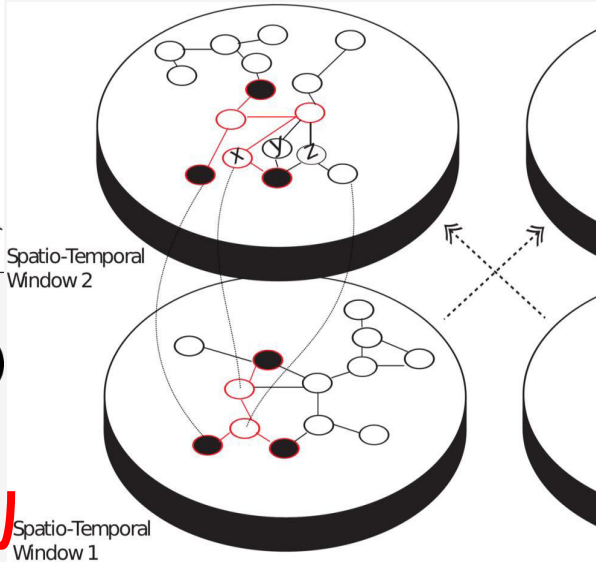
$$F = (F_1, F_2, \cdots, F_T)$$

$$F_i = \{ \mathrm{F}_i^1, F_i^2 \cdots F_i^n \}$$

$\mathrm{F}_i^j(V_{F_i}^j, E_{F_i}^j)$ alt kümesidir $\mathrm{G}_i^j(V_i^j, E_i^j)$

ST-Steiner



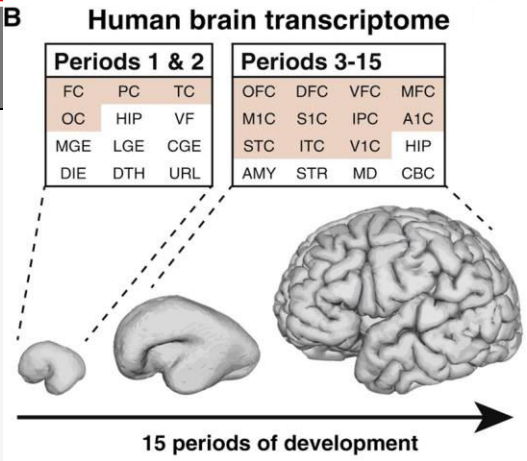


$$o(\boldsymbol{F}) = \sum_{i=1}^{T}\sum_{j=1}^{n} o_F(F_i^j) + \sum_{i=2}^{T}\sum_{j=1}^{n} \lambda_i^j \sum_{\substack{v \in V_i^j \\ v \notin V_{F_i^j}}} \phi(\alpha, v, p_i^j, \boldsymbol{F_{i-1}})$$

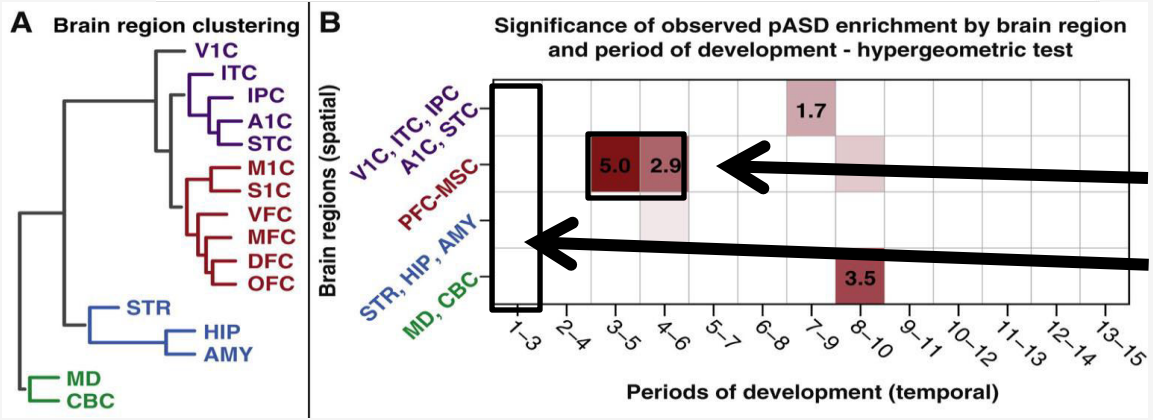
orijinal PCSF çözümleri

farklılıkları cezalandıran fonksiyon

Veri setleri



Period	Description	Age
1	Embryonic	4-8 PCW
2	Early fetal	8-10 PCW
3	Early fetal	10-13 PCW
4	Early mid-fetal	13-16 PCW
5	Early mid-fetal	16-19 PCW
6	Late mid-fetal	19-24 PCW
7	Late fetal	24-38 PCW
8	Neonatal & early infancy	0-6 M
9	Late infancy	6-12 M
10	Early childhood	1-6 Y
11	Middle and late childhood	6-12 Y
12	Adolescence	12-20 Y
13	Young adulthood	20-40 Y
14	Middle adulthood	40-60 Y
15	Late adulthood	60 Y+



Hedef

Önceki ağlar

Resim: Willsey et al. 2013, *Cell*.

Veri setleri

ASC WES[1] (Öğrenme)

- 16,098 DNA örneği
- 3,871 ASD örneği (trio)
- 9,937 kontrol örneği

SSC WES[2] (Test)

- 1,643 ek DNA örneği (trio)
- **251 vurulmuş gen**

Bu 251 geni tahmin edebilir miyiz?

[1] De Rubeis et al. 2014, “Synaptic, transcriptional and chromatin genes disrupted in autism”. *Nature*.

[2] Iossifov et al. 2014, “ The contribution of de novo coding mutations to autism spectrum disorder”. *Nature*.

Sonuçlar 1

gene set name	p-value	intersection / # genes predicted
ST-St. Every(1-3)+PFC(3-5)	**6.958e-12**	**21 / 234**
ST-St. Every(1-3)+PFC(4-6)	1.827e-09	19 / 256
NETBAG	2.476e-06	9 / 87
DAWN PFC(3-5)	3.842e-08	17 / 246
DAWN PFC(4-6)	9.017e-10	18 / 218
MAGI Best1	3.679e-05	6 / 47
MAGI Ext1	7.865e-05	8 / 104
MAGI Best2	2.628e-02	2 / 19
MAGI Ext2	4.396e-03	5 / 80

Alakalı Otizm Gen Listeleri

- SFARI Category 1 (24 gen)
- SFARI Category 2 (59 gen)
- FMRP Hedefleri (842 gen)
- RBFOX – peak (1048 gen)
- RBFOX – splice (587 gen)
- Histone Modifiers (152 gen)
- Sinaptik Genler (878 gen)

Sonuçlar 2

Predicted Gene Set S	SFARI Category 1		SFARI Category 2		FMRP Targets		RBFOX - peak		RBFOX - splice	
	p-value	$\lvert\cap\rvert/\lvert S\rvert$	p-value	$\lvert\cap\rvert/\lvert S\rvert$	p-value	$\lvert\cap\rvert/\lvert S\rvert$	p-value	$\lvert\cap\rvert/\lvert S\rvert$	p-value	$\lvert\cap\rvert/\lvert S\rvert$
ST-St. Every(1-3)+PFC(3-5)	**1.397e-25**	**16 / 234**	2.906e-22	19 / 234	3.198e-22	52 / 234	2.196e-09	38 / 234	**1.120e-11**	**31 / 234**
ST-St. Every(1-3)+PFC(4-6)	8.334e-23	15 / 256	5.890e-20	18 / 256	**8.684e-27**	**60 / 256**	**1.735e-11**	**44 / 256**	3.649e-08	27 / 256
NETBAG	5.062e-18	10 / 87	1.422e-13	10 / 87	1.750e-10	21 / 87	2.870e-04	14 / 87	8.199e-05	11 / 87
DAWN PFC(3-5)	5.479e-19	13 / 246	**4.242e-25**	**21 / 246**	3.985e-18	48 / 246	2.740e-09	39 / 246	3.774e-09	28 / 246
DAWN PFC(4-6)	7.067e-24	15 / 218	1.252e-19	17 / 218	7.762e-19	46 / 218	2.735e-10	38 / 218	2.231e-08	25 / 218
MAGI Best1	3.554e-07	4 / 47	9.615e-03	2 / 47	1.031e-08	14 / 47	3.943e-03	8 / 47	5.946e-04	7 / 47
MAGI Ext1	1.866e-07	5 / 104	3.220e-04	4 / 104	3.600e-06	17 / 104	1.500e-02	12 / 104	2.137e-05	13 / 104
MAGI Best2	2.652e-04	2 / 19	1.000e+00	0 / 19	2.321e-12	12 / 19	3.127e-03	5 / 19	2.512e-03	4 / 19
MAGI Ext2	3.064e-06	4 / 80	2.049e-03	3 / 80	1.165e-08	18 / 80	1.442e-03	12 / 80	3.735e-05	11 / 80

Sonuçlar 3

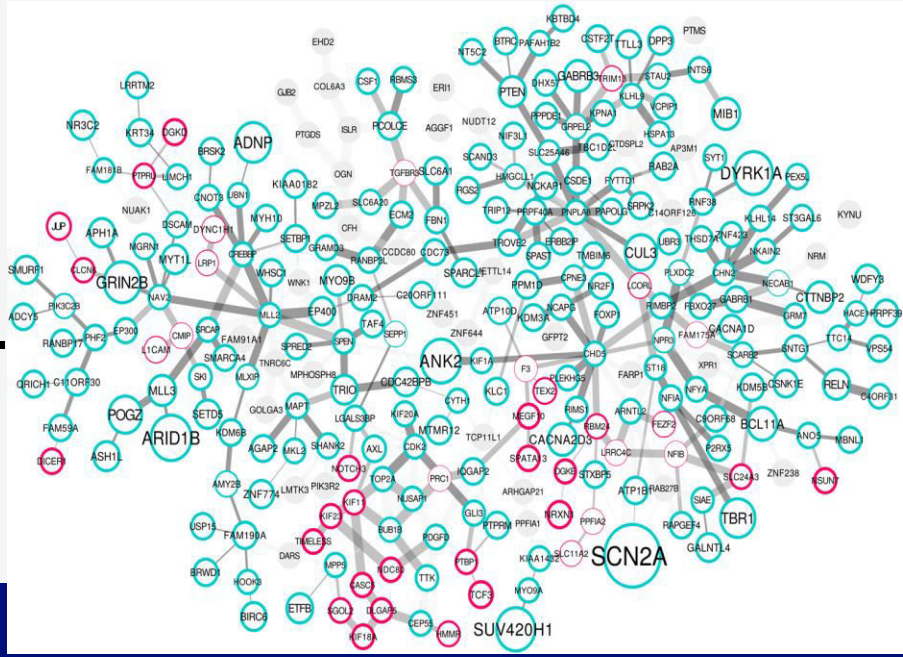
Predicted Gene Set S	Histone Modifiers		Synaptic Genes	
	p-value	$\|\cap\|/\|S\|$	p-value	$\|\cap\|/\|S\|$
ST-St. Every(1-3)+PFC(3-5)	**4.709e-07**	**12 / 234**	1.195e-04	25 / 234
ST-St. Every(1-3)+PFC(4-6)	7.890e-06	11 / 256	7.622e-05	27 / 256
NETBAG	7.103e-04	5 / 87	2.409e-03	11 / 87
DAWN PFC(3-5)	1.835e-04	9 / 246	3.274e-04	25 / 246
DAWN PFC(4-6)	7.347e-05	9 / 218	1.744e-03	21 / 218
MAGI Best1	2.100e-06	6 / 47	2.813e-01	4 / 47
MAGI Ext1	2.072e-06	8 / 104	9.394e-03	11 / 104
MAGI Best2	1.000e+00	0 / 19	**6.258e-08**	**9 / 19**
MAGI Ext2	1.000e+00	0 / 80	3.862e-06	15 / 80

Zaman bilgisi ile elde edilen yeni genler

- 5 Kinesin ailesi (KIF) genleri: "*transport cargo to dendritic spines undergoing synaptic plasticity over microtubules also play a role in organization of spindle microtubules during mitosis.*"

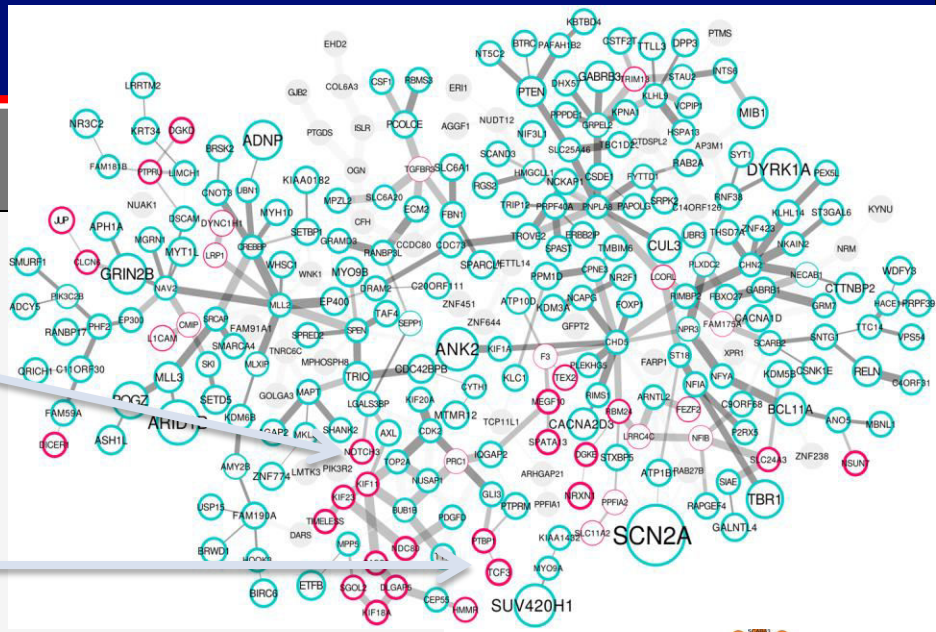
- NDC80 veSGOL2: kinetochore & microtubule attachment - cell division.

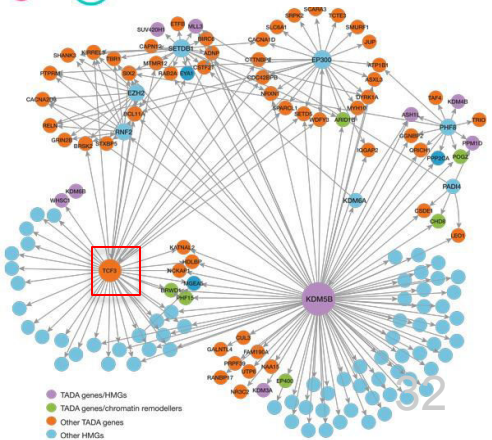
→Literatürde dile getirilmeyen genler.



Yorumlanabilir sonuçlar

- NOTCH3
  - Nöron başkalaşımı
- TCF3
  - Embriyonik kok hücre başkalaşımını başlatır
  - Nöron öncesi hücrelerde başkalaşımı durdurur
  - De Rubeis et. al., 2014, Nature yayınında merkez gen





Çıkarımlar

Otizmin genetik mimarisinin anlaşılması önemli bir problem.

Biyolojik ağ bazlı algoritmalar hangi genlerin ilgili olduğunu tespit etmede önemli katkıda bulunmuştur.

ST-Steiner algoritması bunları bir adım öteye götürerek zaman mekânsal değişiklikleri hesaplamalarına katmıştır.

ST-Steiner daha hassas sonuçlar vermektedir:

- Bağlantı için gerekli olmayan genleri dışarıda bırakır.
- Önceki donemde seçilen genleri hesaba katarak daha güvenli tercihler yapar.
- Sonuçlar kümelenmenin statik değil dinamik olduğunu desteklemektedir.

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY AFFILIATED CONFERENCE

HIBIT2018

Call for submissions opens	**May 15, 2018**
Submission deadline	**August 1, 2018**
Acceptance notification	**September 1, 2018**
Early registration discount ends	**September 15, 2018**
Conference Program	**October 25-27, 2018**
ISCB-RSG Turkey – Student Symposium	**October 27, 2018**

The 11th **HIBIT** Conference

International Symposium on Health Informatics and Bioinformatics

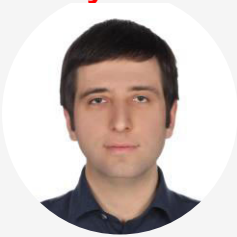
October 25-27, 2018 - Antalya

Medical
NMR
Metabolomics
Platform
of
Strasbourg

cicek@cs.bilkent.edu.tr

Teşekkürler!

Spatio-Temporal Gene Discovery for Autism Spectrum Disorder



Utku Norman

Bilkent Yüksek Lisans mezunu - EPFL doktora öğrencisi

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

A. Ercument Cicek, Ph.D.
ercumentcicek.com
@crashius

SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE