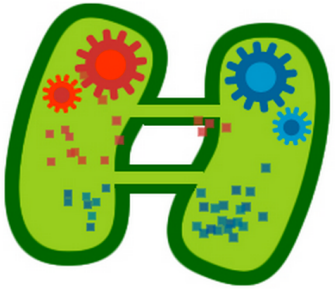
Variant analysis of Simplex Autism Families

Laura Jiménez
Lyon Lab

Instituto de Biotecnología
UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO



Bacillus booleanus

Acknowledgments

Gholson Lyon

Jason O'Rawe

Han Fang

Ivan Iossifov

Yiyang Wu

Autism

About 1 in 68 children has been identified with autism spectrum disorder (ASD) according to estimates from CDC's Autism and Developmental Disabilities Monitoring (ADDM) Network.

Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2010

Surveillance Summaries
March 28, 2014 / 63(SS02);1-21

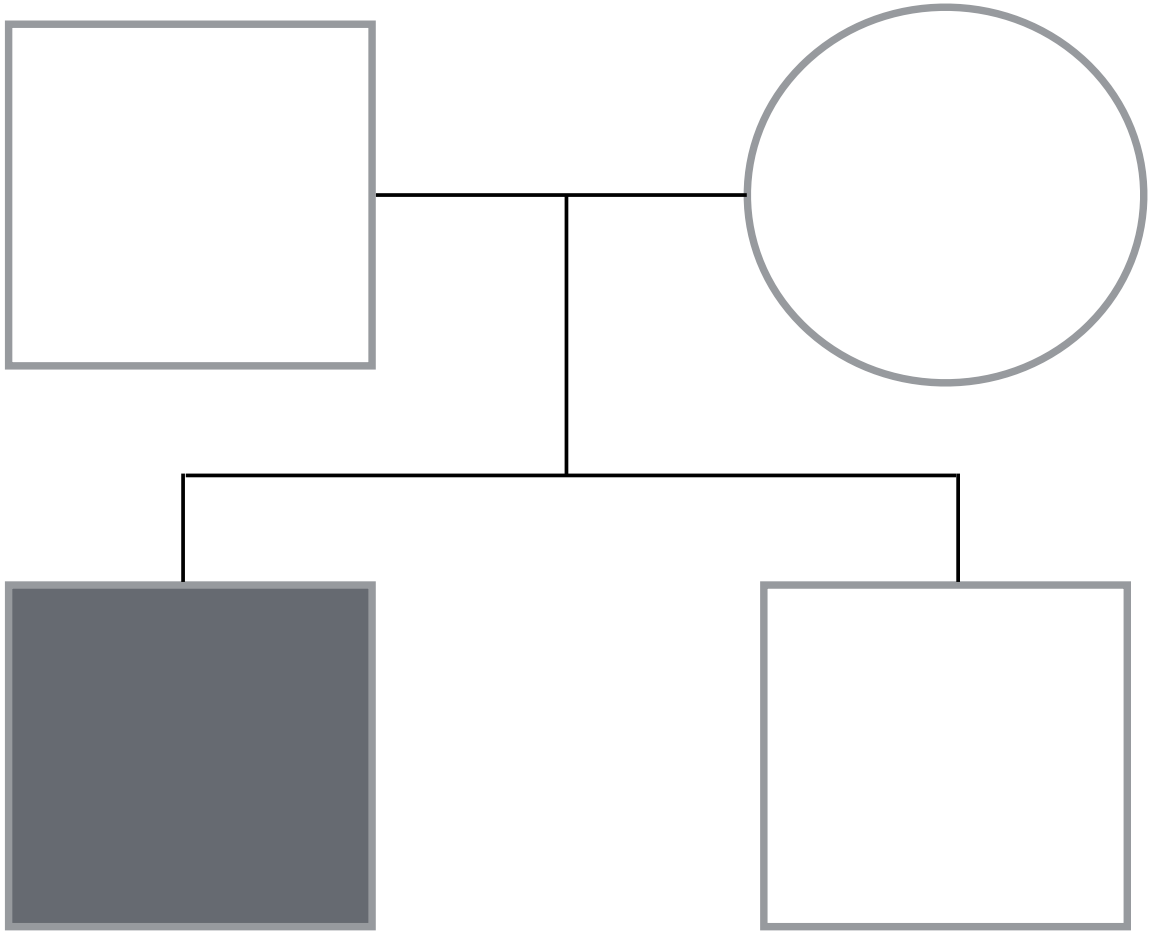
Autism and Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators

Corresponding author: Jon Baio, EdS, National Center on Birth Defects and Developmental Disabilities, CDC. Telephone: 404-498-3873; E-mail: jbaio@cdc.gov.

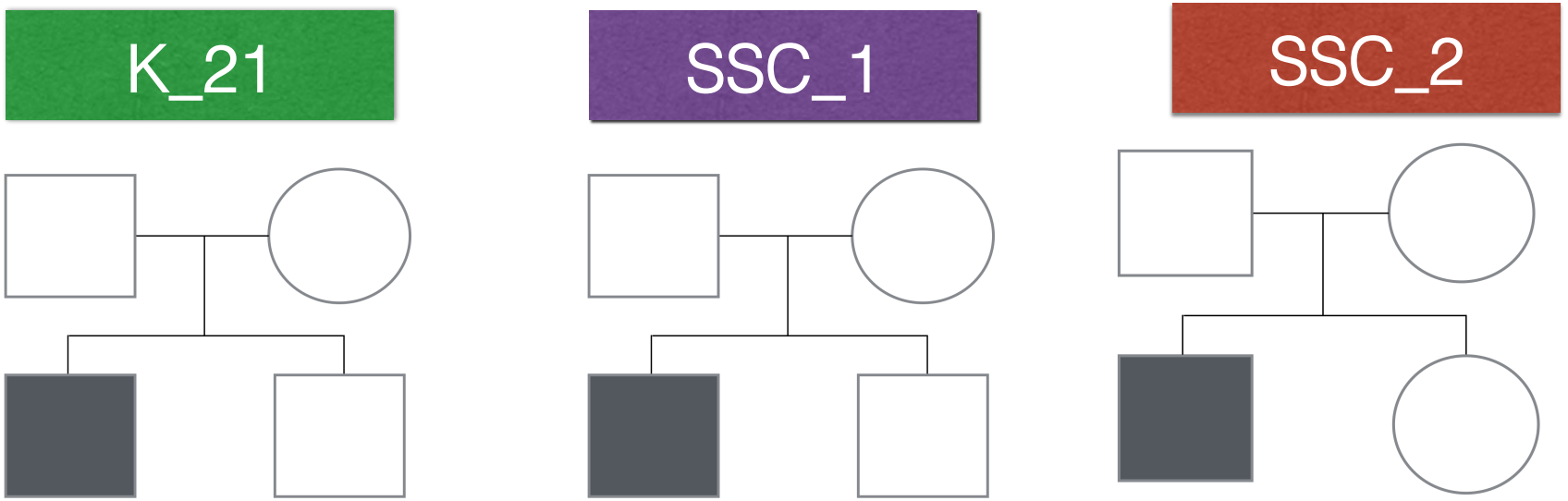
Previous Studies & Strategies

- Cytogenetic studies
- Linkage analysis, and candidate gene association analysis
- GWAS
- CNVs on Specific Regions i.e. VIP Project (16p11.2)
- WES & WGS on individuals, tríos, quads, multiplex pedigrees.

Simplex Autism Family



Analyzed Simplex Autism Families

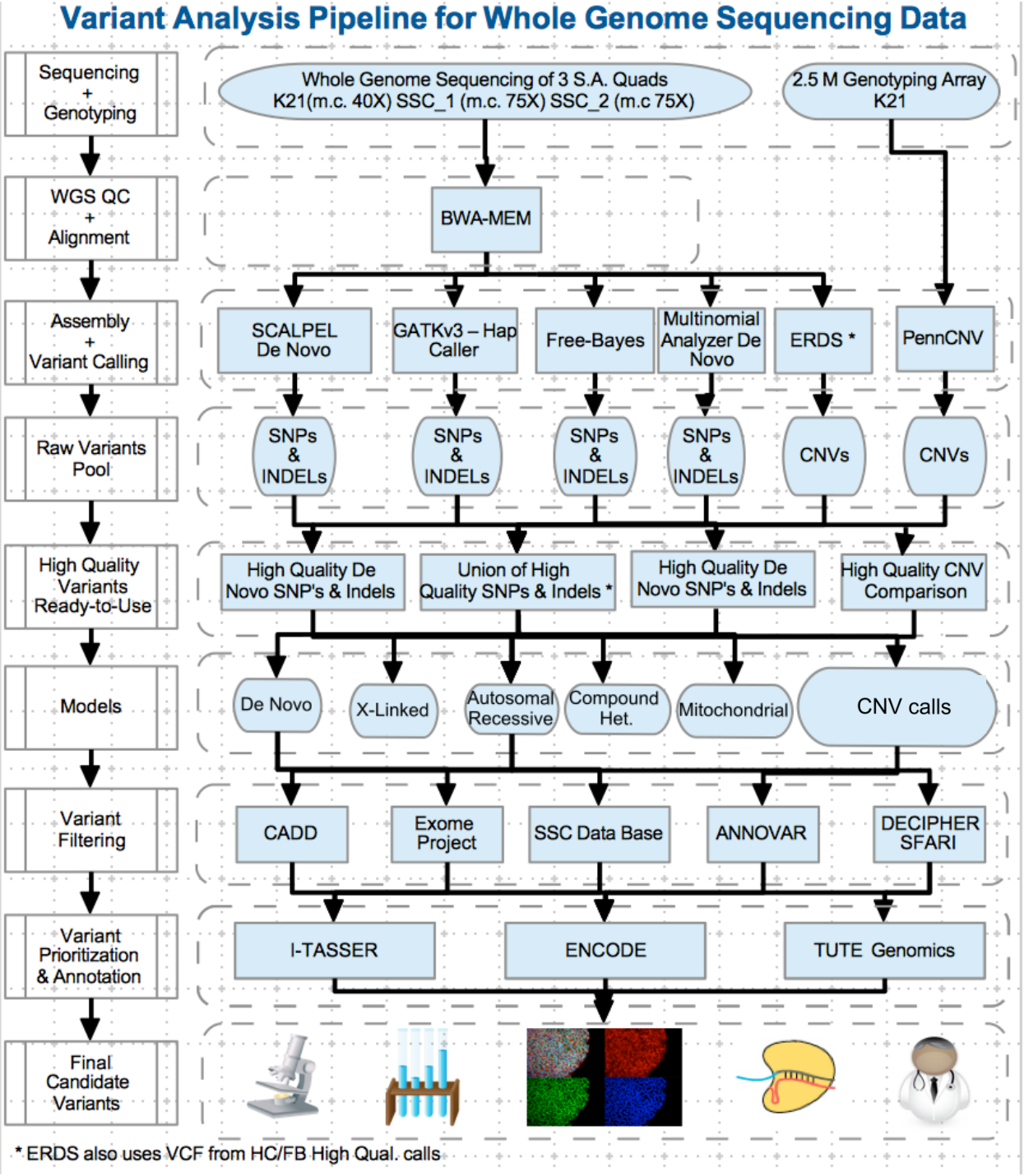


The Simons Simplex Collection

WGS ~ 40 X
Illumina HiSeq 2000
Genotyped 2.5 M
Illumina Omni2.5 array

WGS ~ 75 X
Illumina HiSeq 2000
WES ~40X
NimbleGen SeqCap EZ Exome v2.0

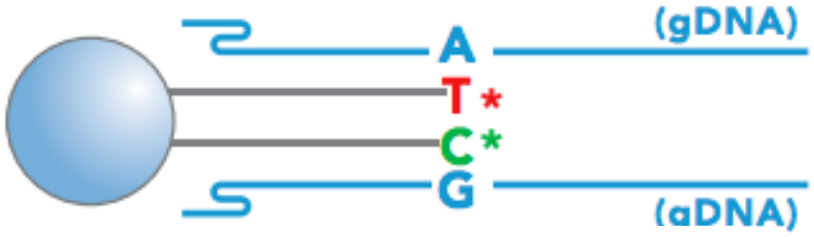
Variant Analysis Pipeline for Whole Genome Sequencing Data

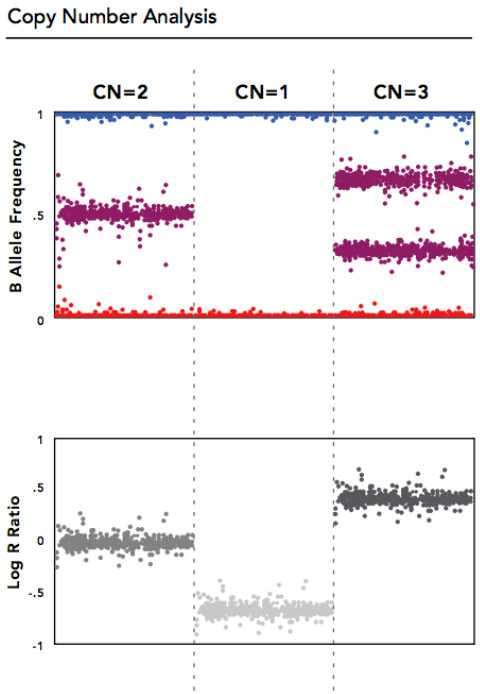


* ERDS also uses VCF from HC/FB High Qual. calls

Copy Number Variants

PennCNV calling for Microarray Data on K21





Log R Ratio = $\log_2(R_{observed}/R_{expected})$

R= (X+Y)

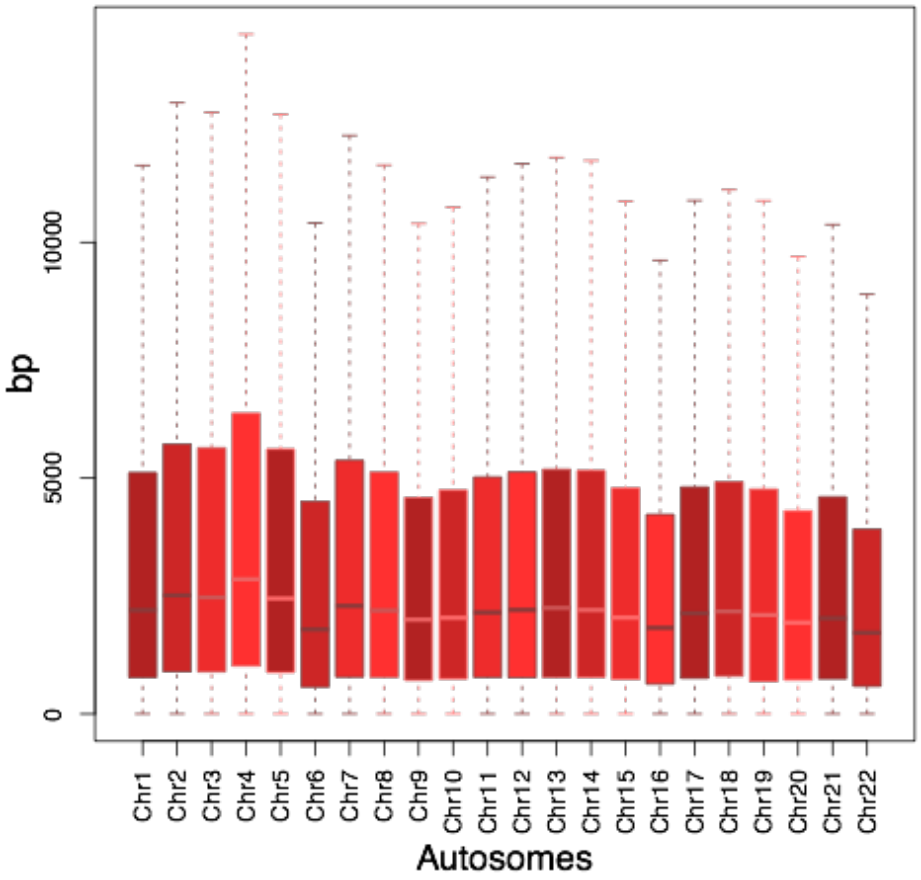
X=norm A allele Y=norm B allele

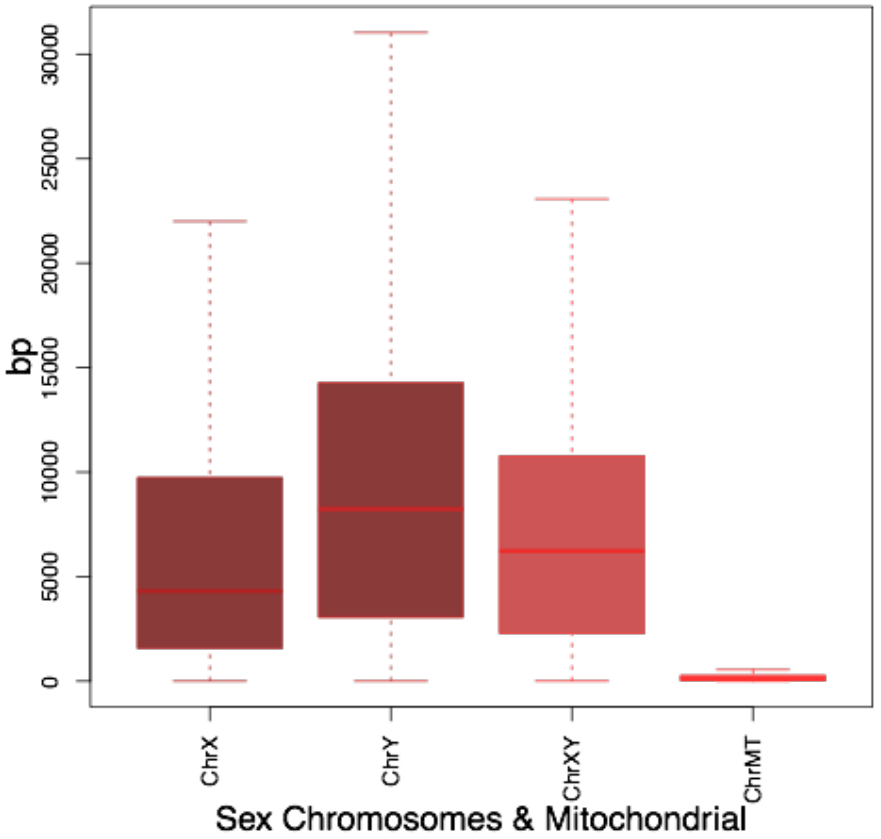
B Allele Freq. = normalized measure of relative signal intensity ratio of the B and A alleles

Population Frequency of B Allele File (generated from 600 controls)

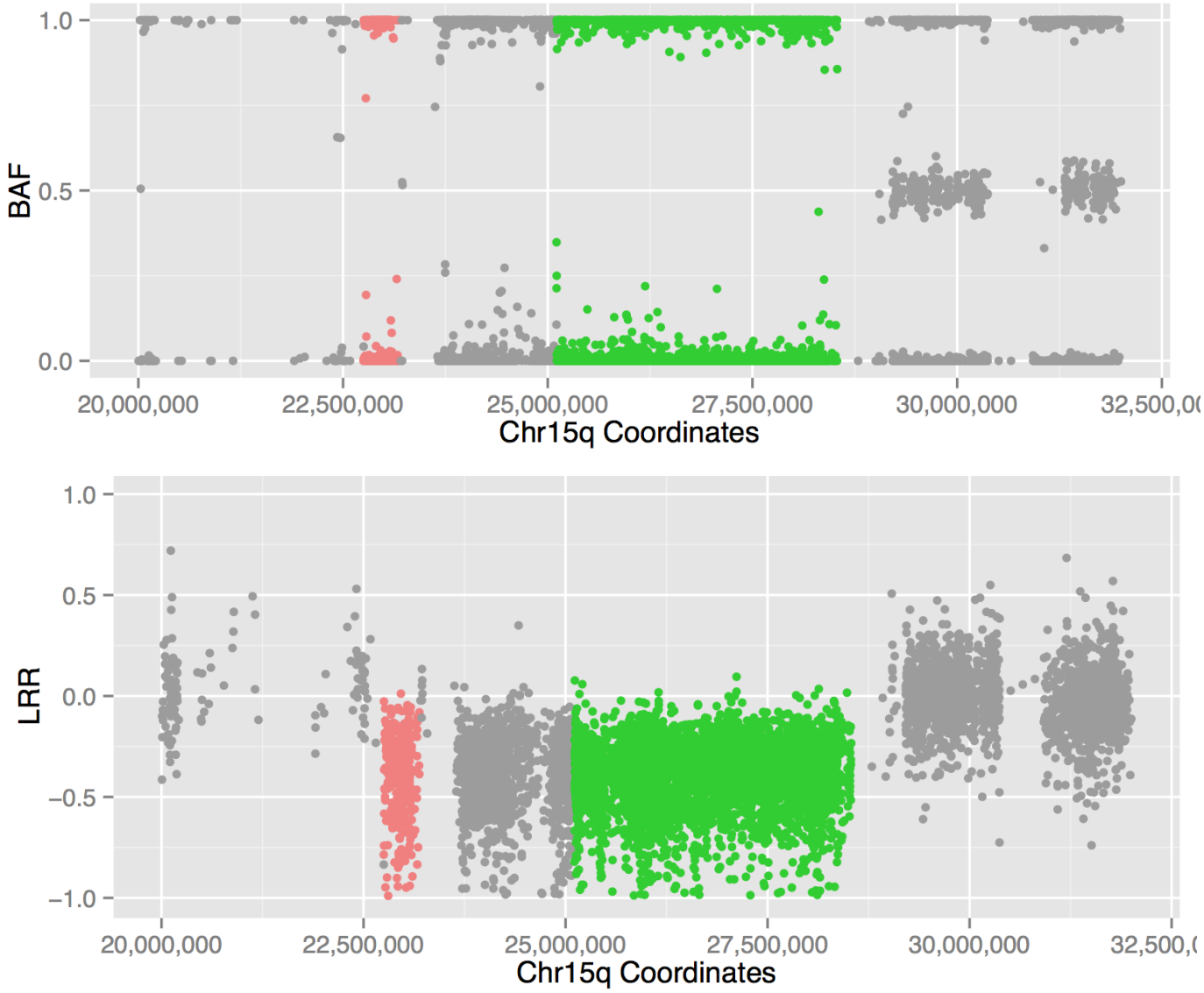
Joint Calling Algorithm

Intermarker Spacing Size by Chromosome





CNVs called by PennCNV on a Prader Willy Child



CNV calling from WGS with ERDS

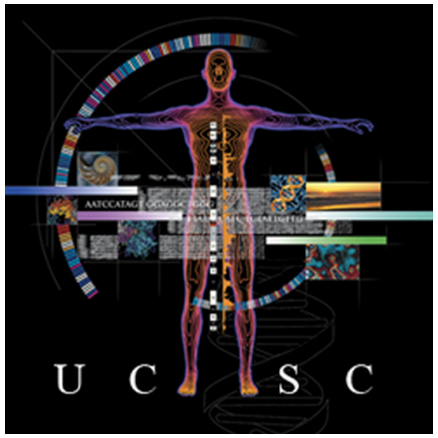
Ming Fu, et al.

Score > 300

Length > 10 Kb

Database of Genomic Variants

DECIPHER GRCh37



ANNOVAR

ClinVar

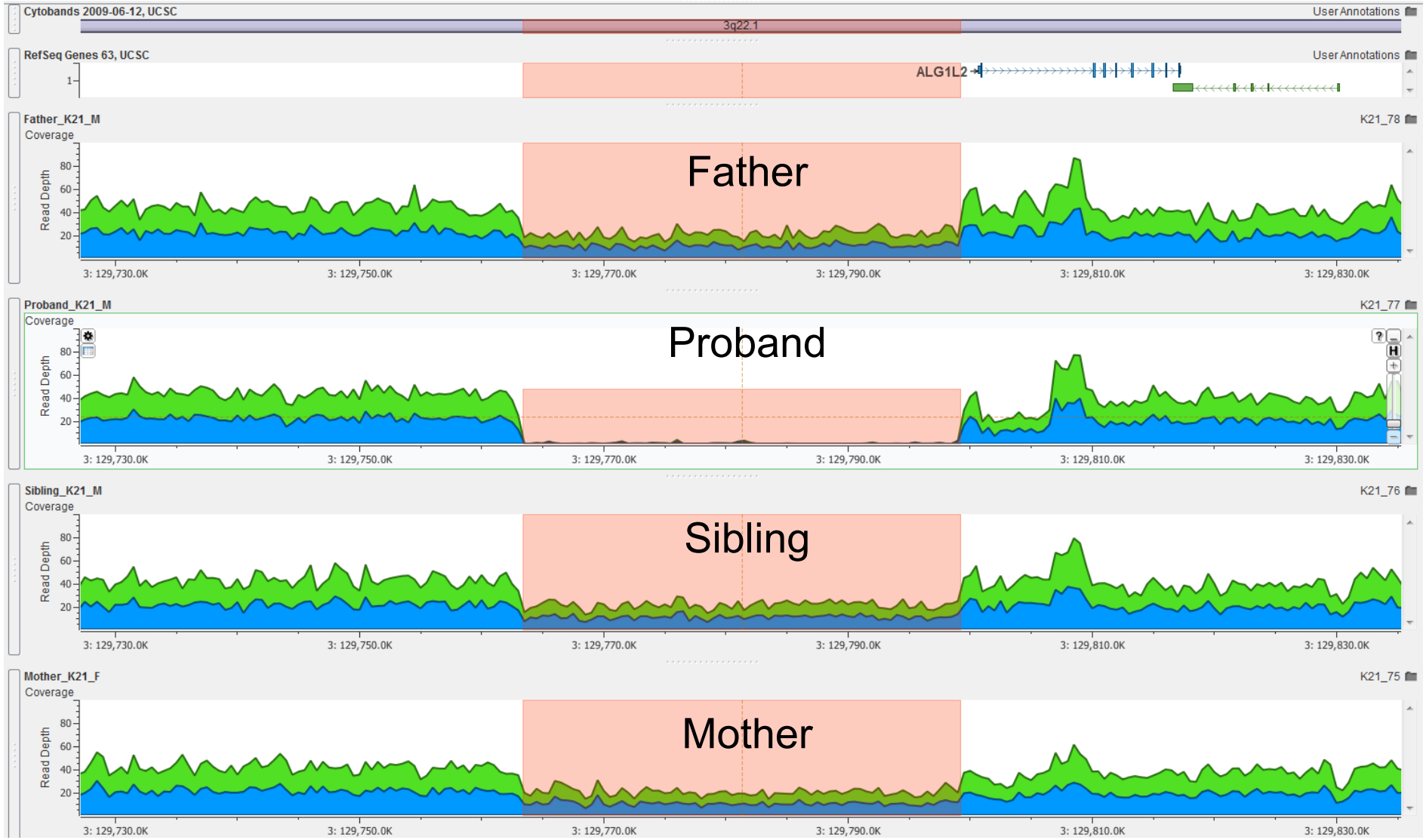
ENCODE

AutDB Autism Database

Region Alignment

K_21

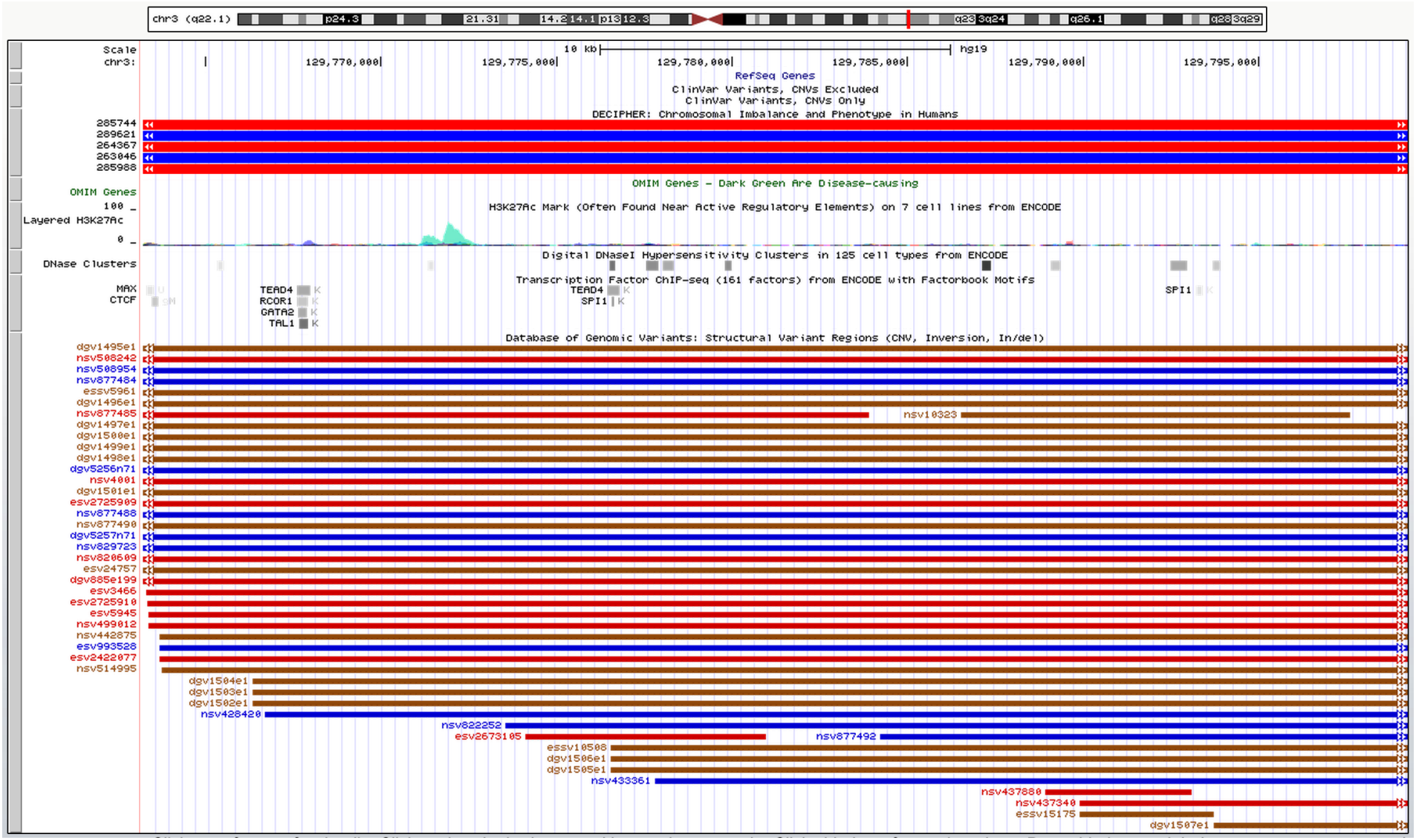
Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
3q22.1	~ 36 Kb.	intergenic	2/0	3026.58



Region Annotation

K_21

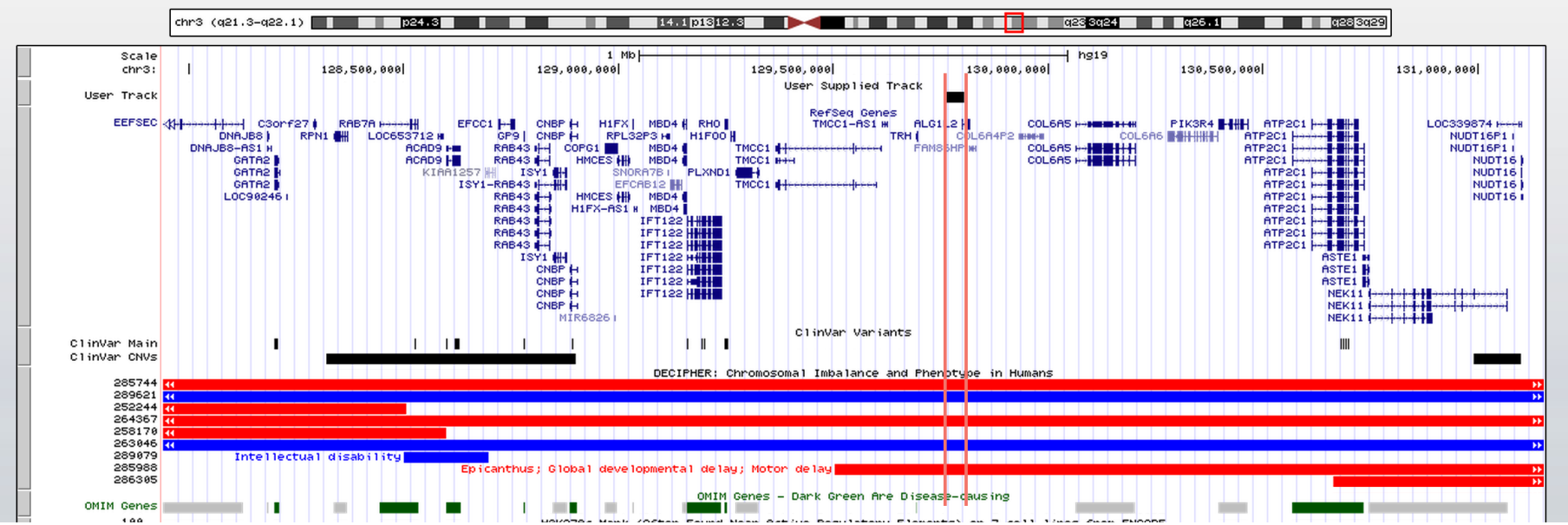
Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
3q22.1	~ 36 Kb.	intergenic	2/0	3026.58



Region Annotation Zoom Out

K_21

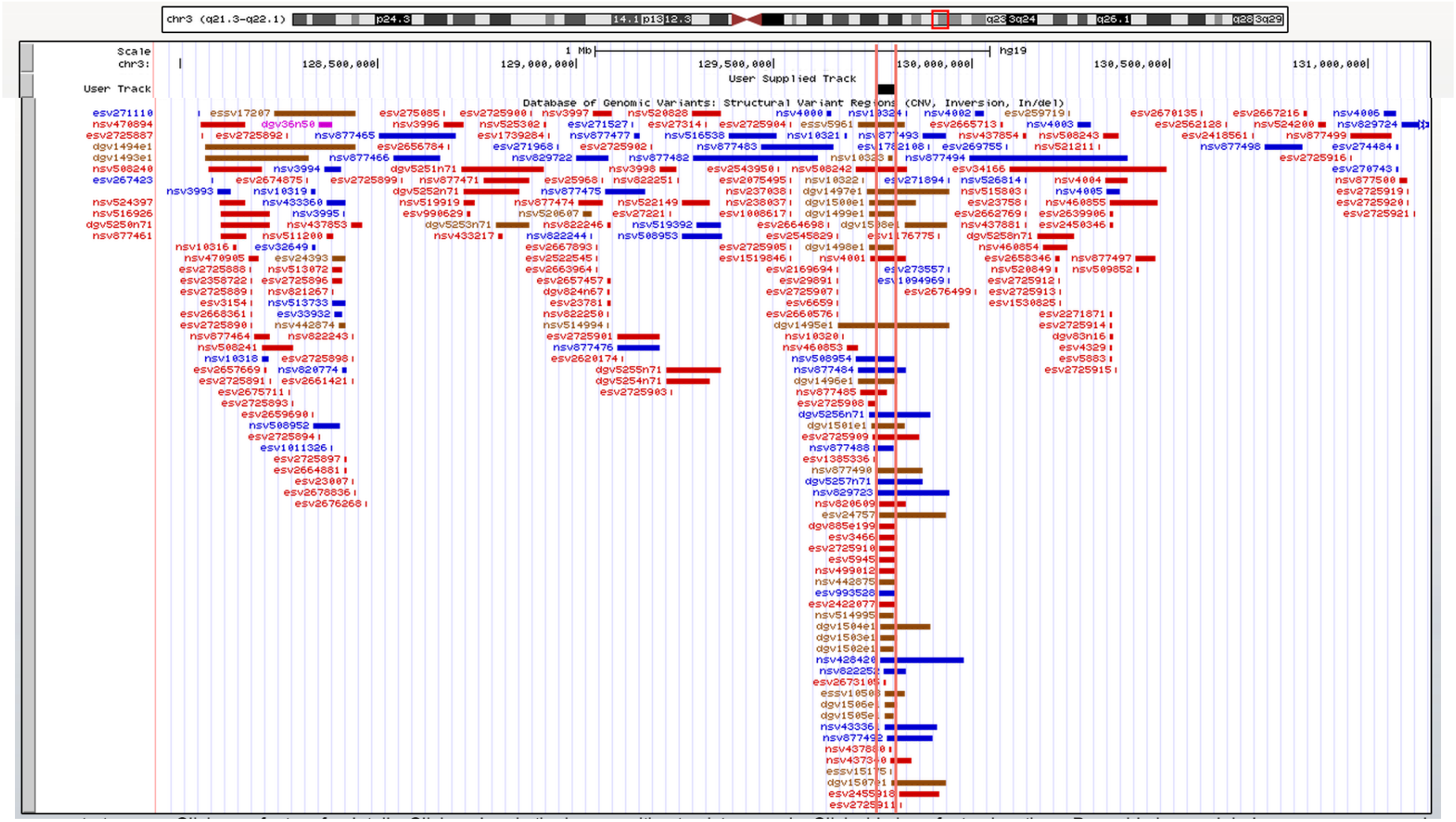
Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
3q22.1	~ 36 Kb.	intergenic	2/0	3026.58



Region Annotation Zoom Out

K_21

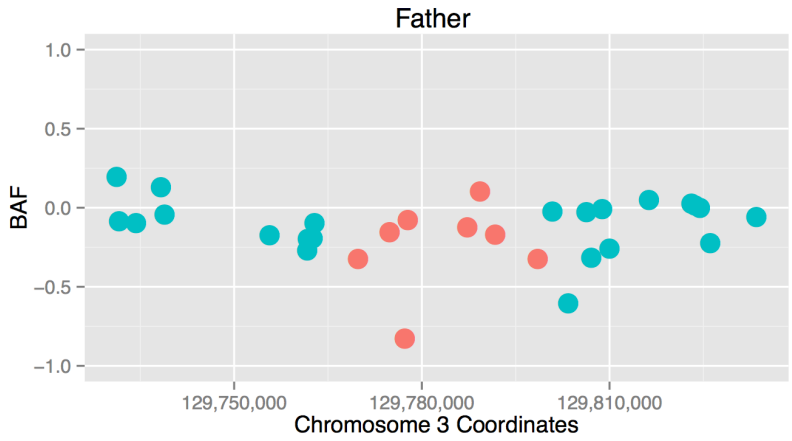
Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
3q22.1	~ 36 Kb.	intergenic	2/0	3026.58

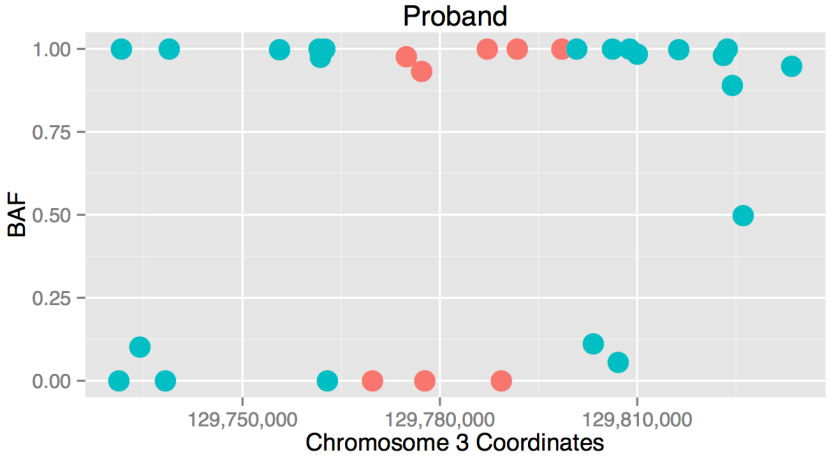


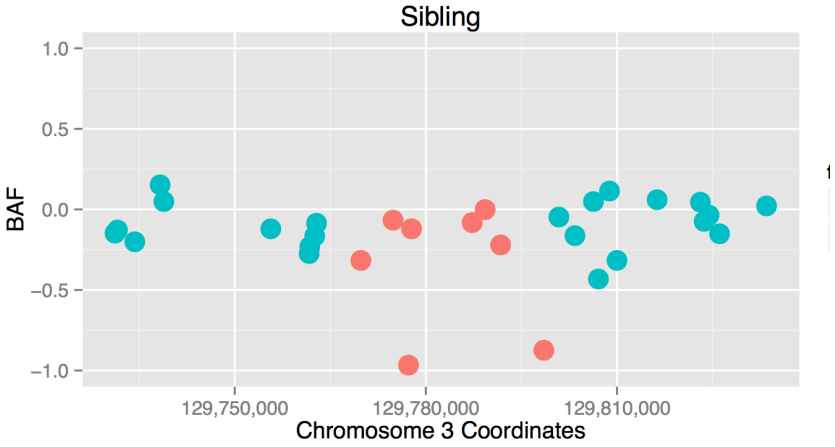
Microarray Data

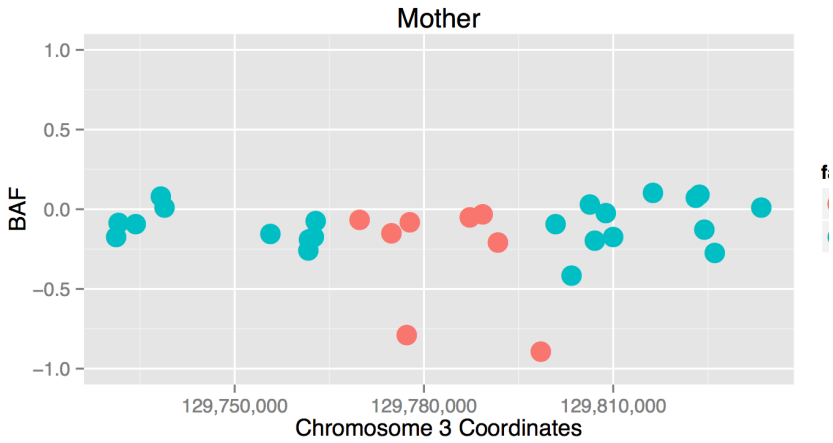
K_21

Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
3q22.1	~ 36 Kb.	intergenic	2/0	3026.58





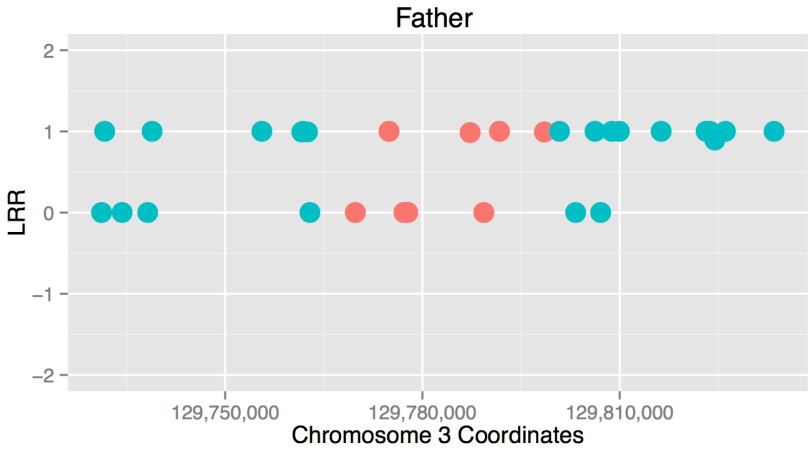


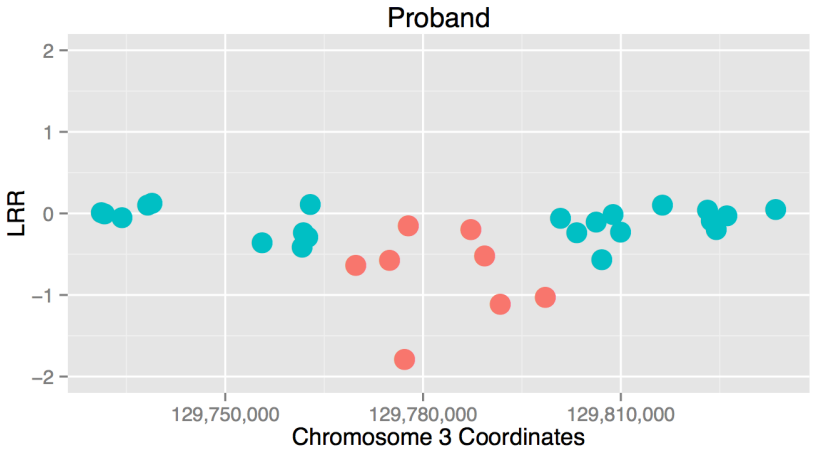


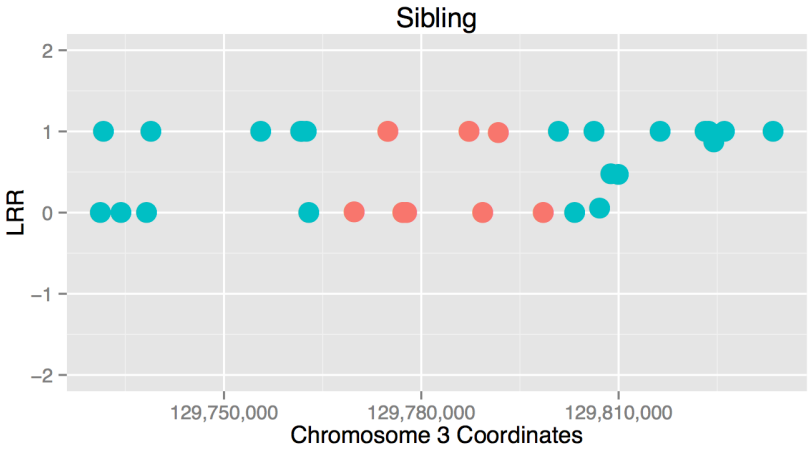
Microarray Data

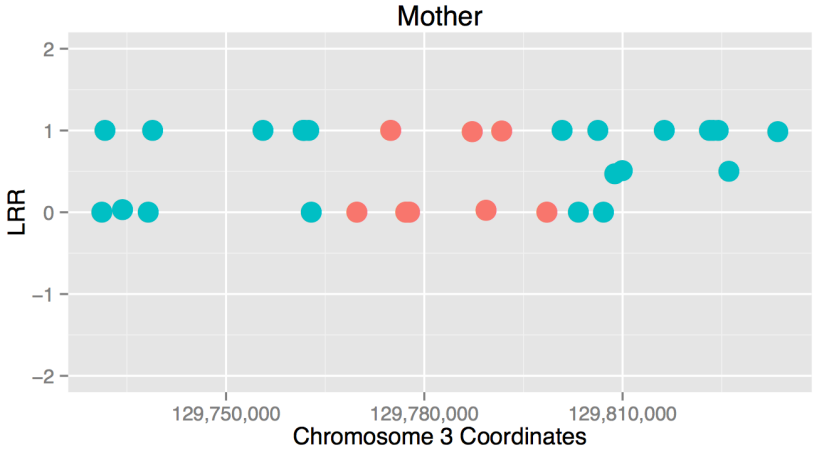
K_21

Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
3q22.1	~ 36 Kb.	intergenic	2/0	3026.58





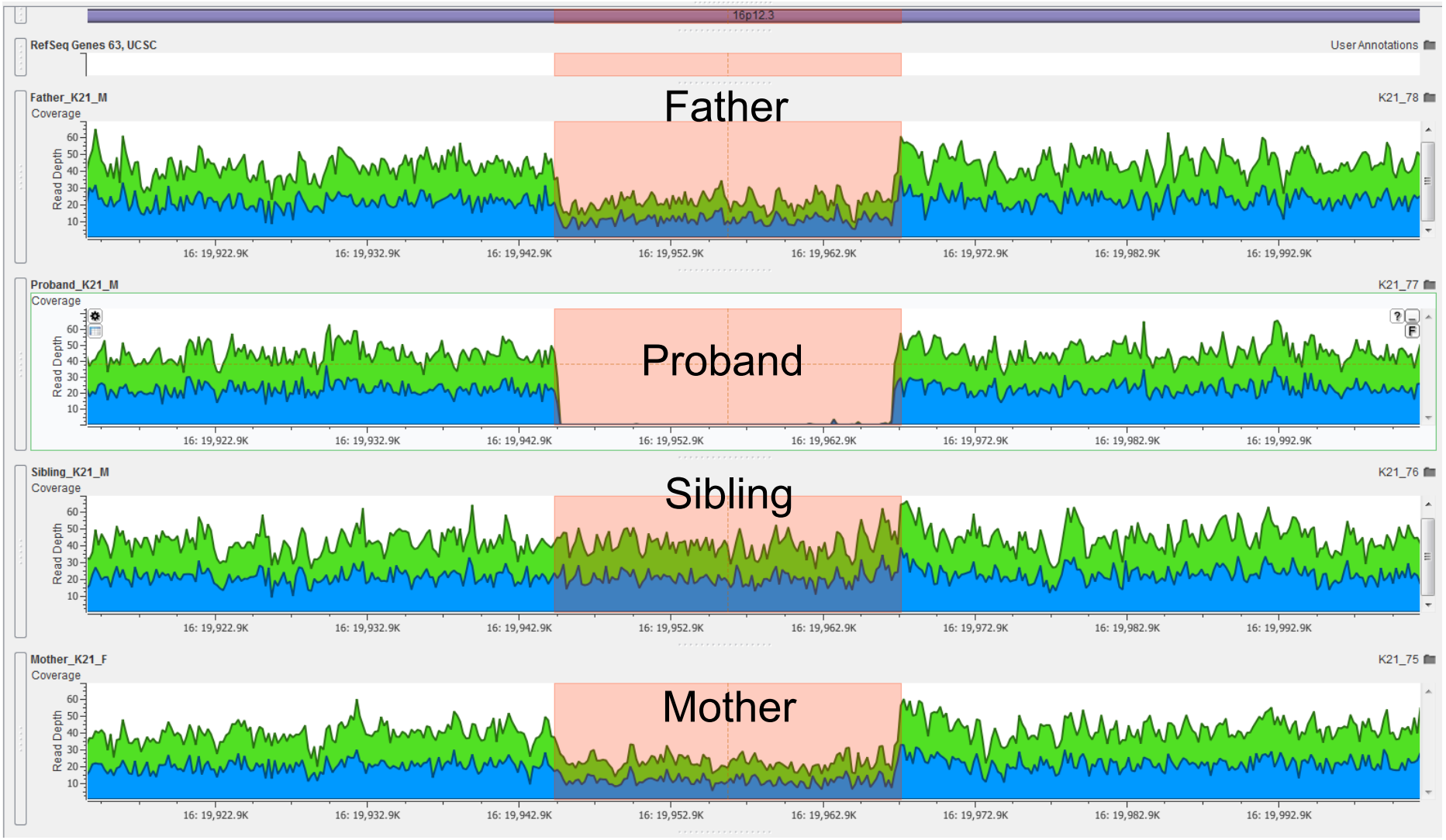




Region Alignment

K_21

Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
16p12.3	~ 22 Kb.	intergenic	2/0	2475.9



Region Annotation

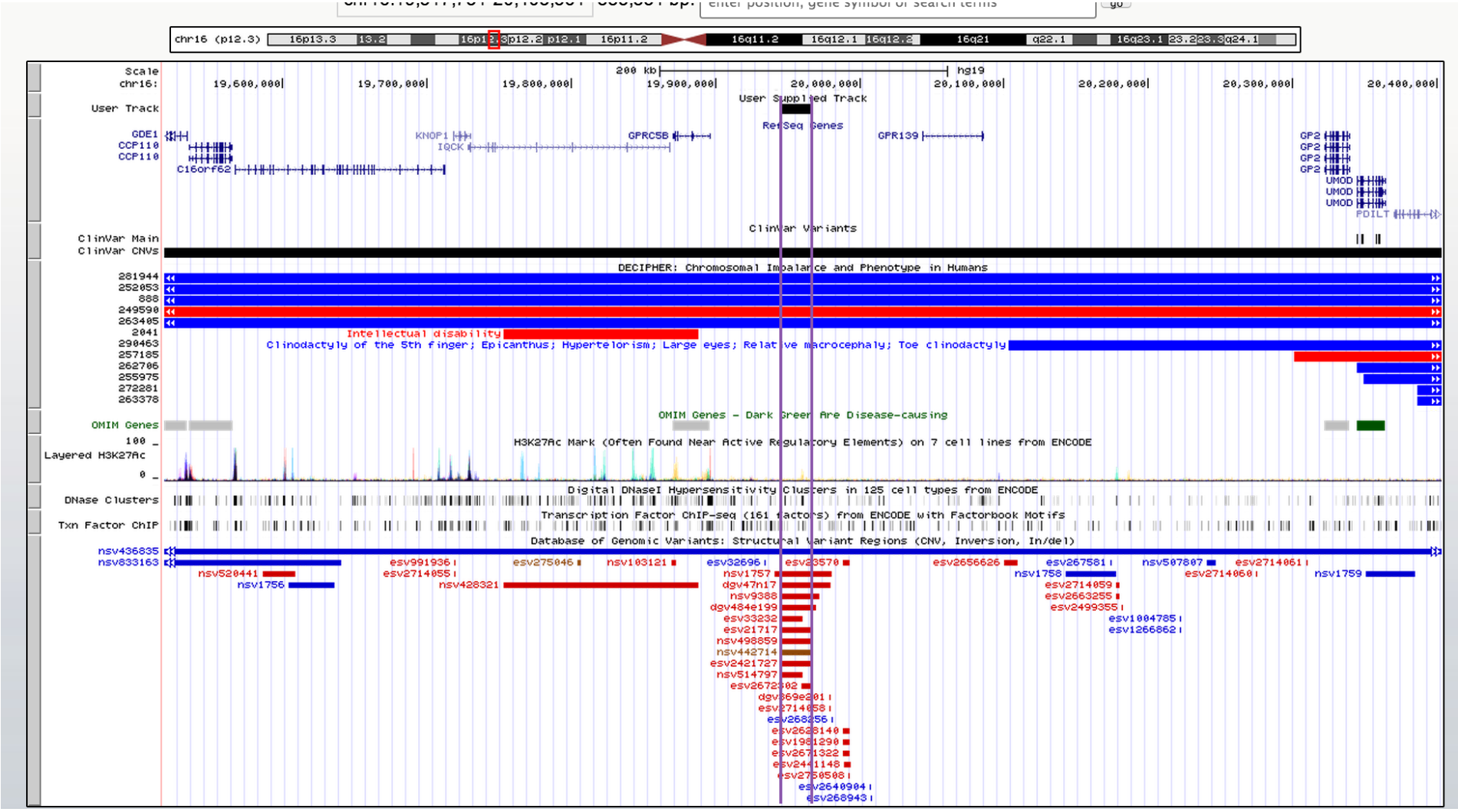
K_21

Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
16p12.3	~ 22 Kb.	intergenic	2/0	2475.9

Region Annotation Zoom Out

K_21

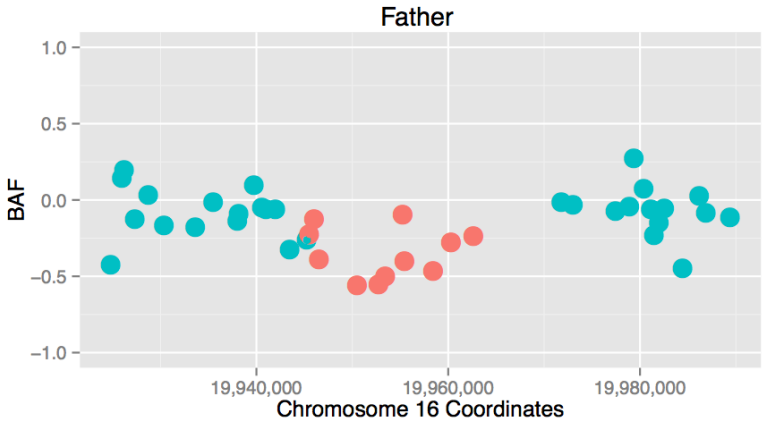
Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
16p12.3	~ 22 Kb.	intergenic	2/0	2475.9

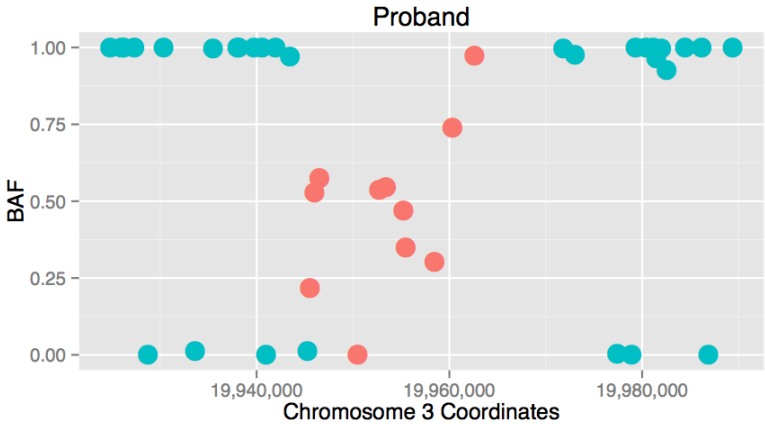


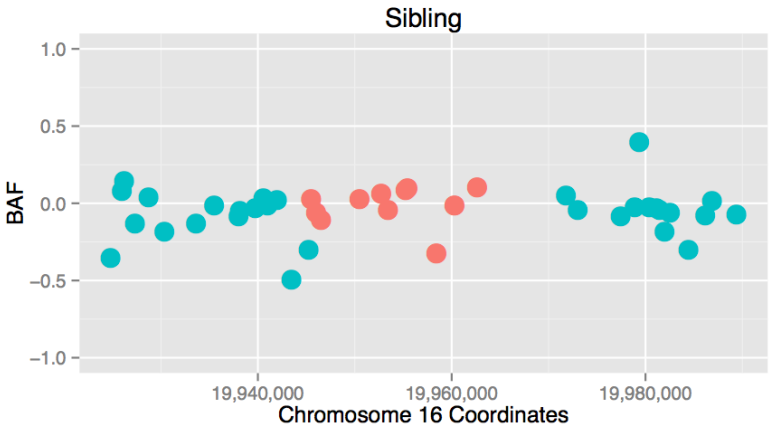
Microarray Data

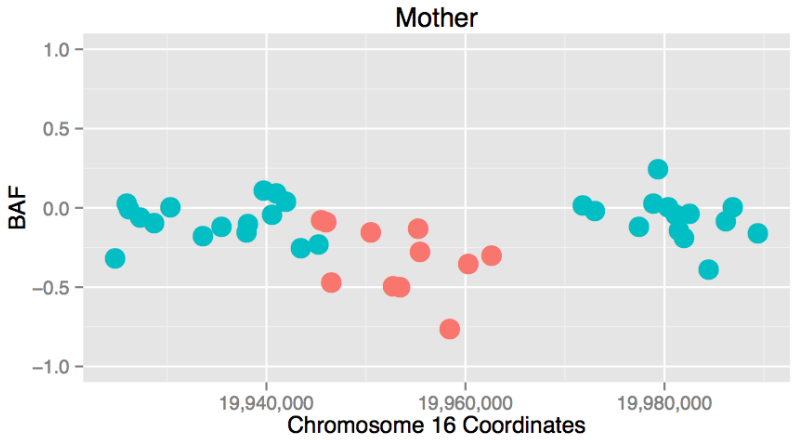
K_21

Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
16p12.3	~ 22 Kb.	intergenic	2/0	2475.9





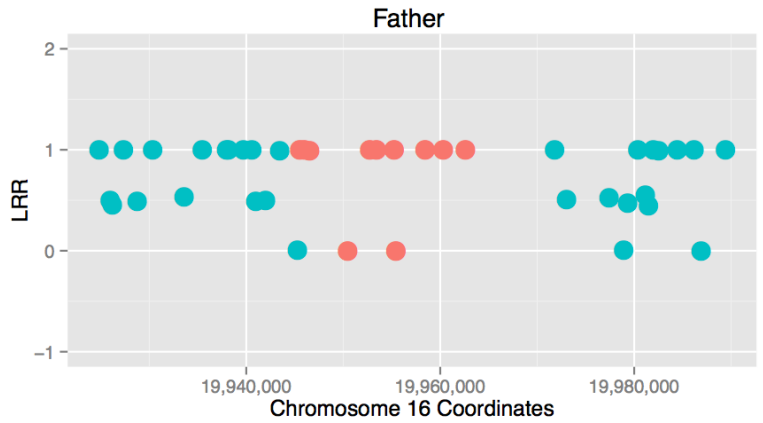


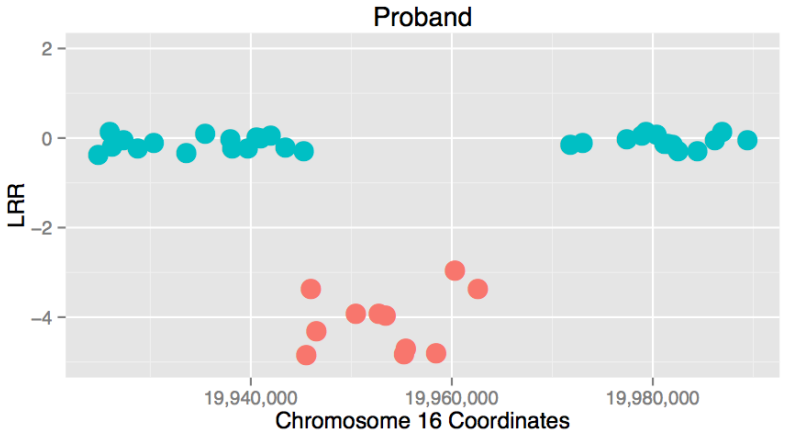


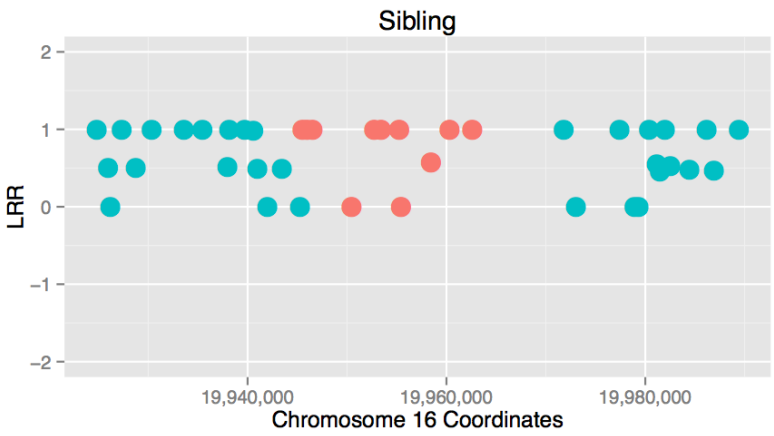
Microarray Data

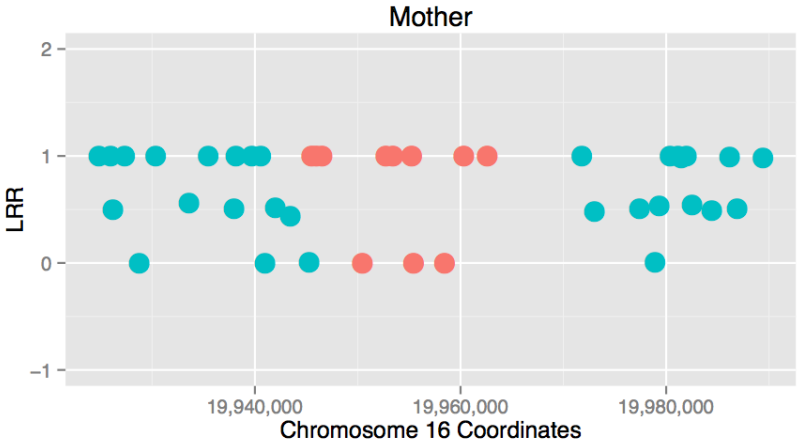
K_21

Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
16p12.3	~ 22 Kb.	intergenic	2/0	2475.9





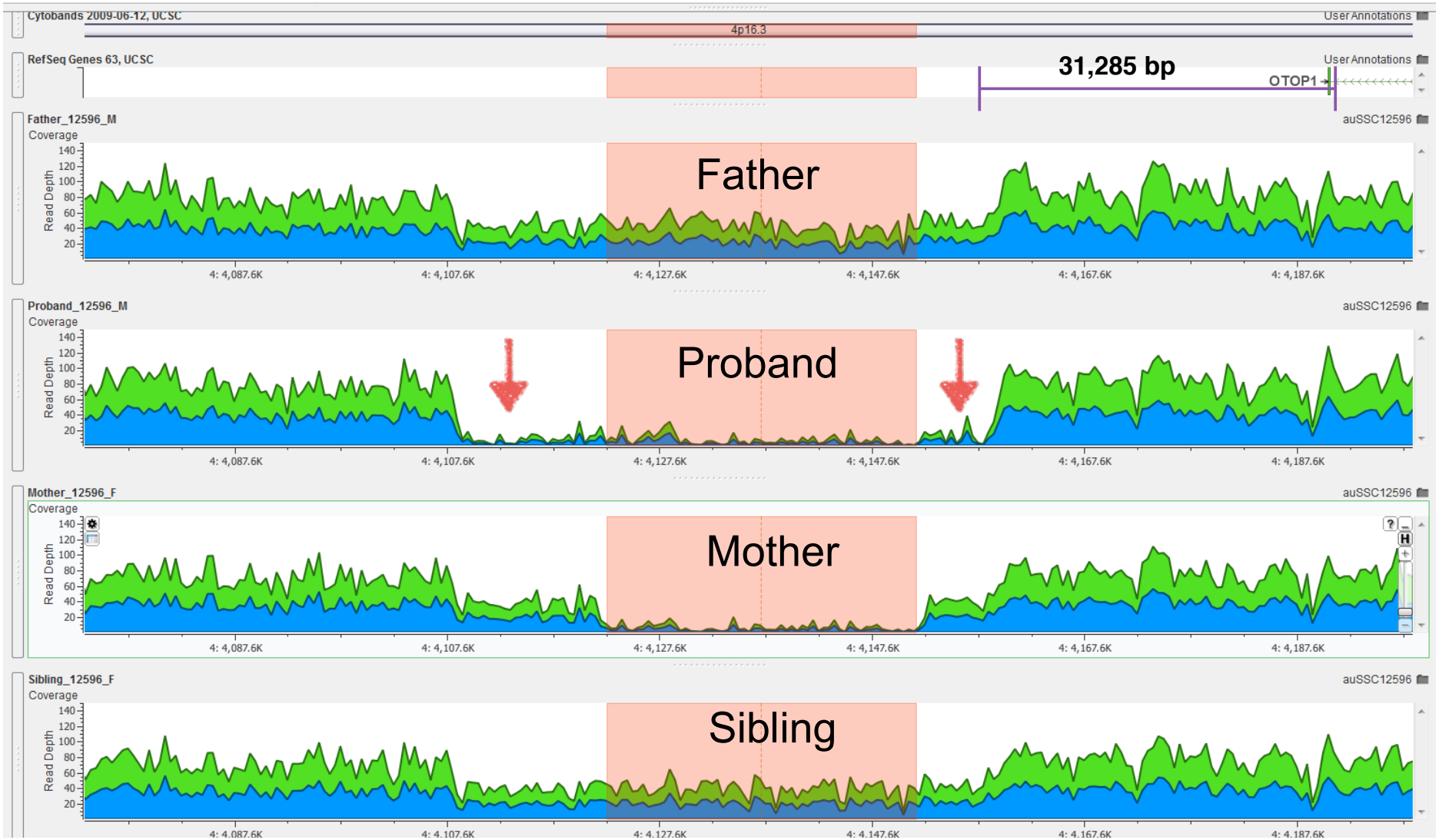




Microarray Data

SSC_2

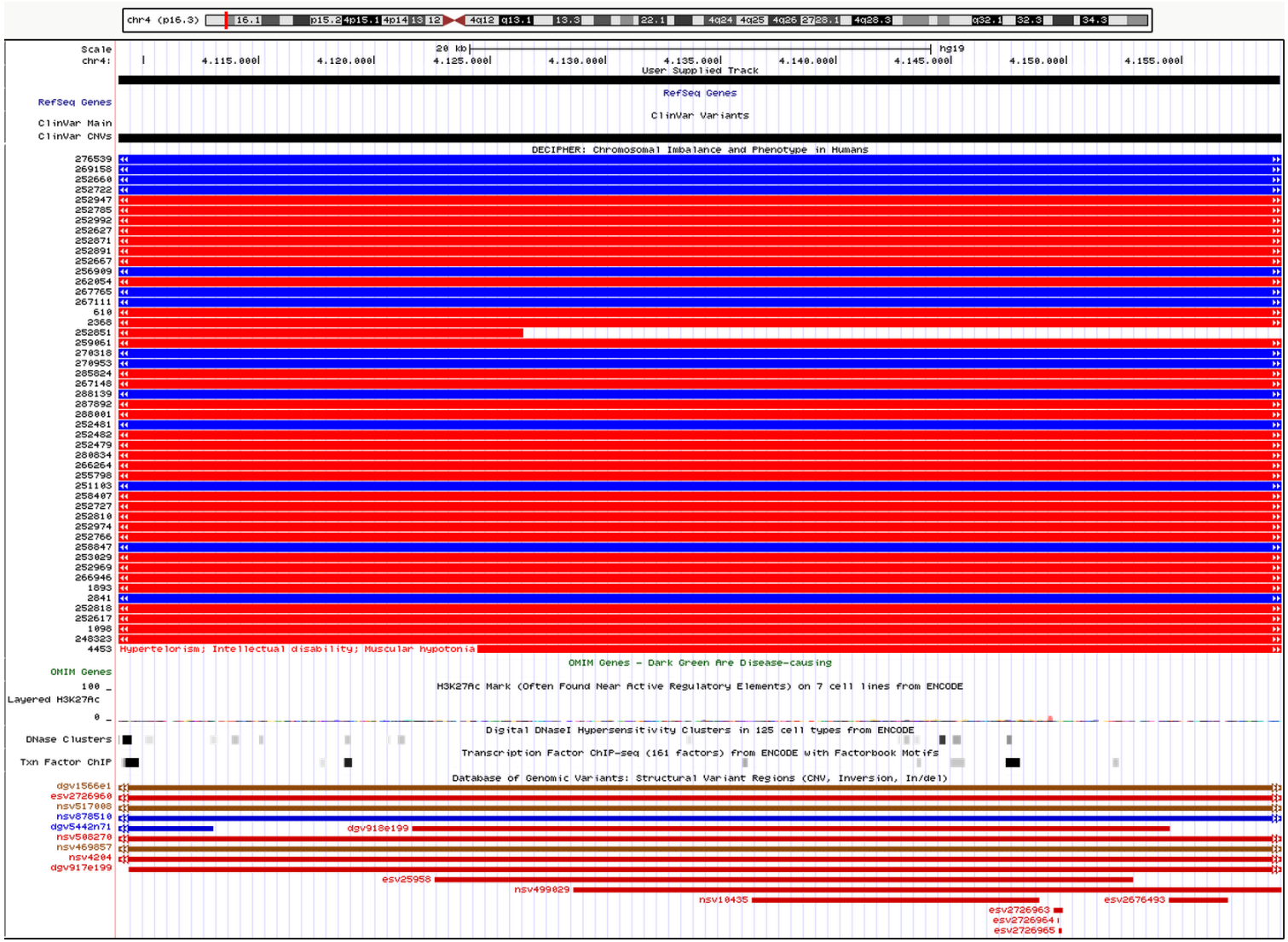
Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
4p16.3	~ 50 Kb.	intergenic	2/0	1525.08



Region Annotation

SSC_2

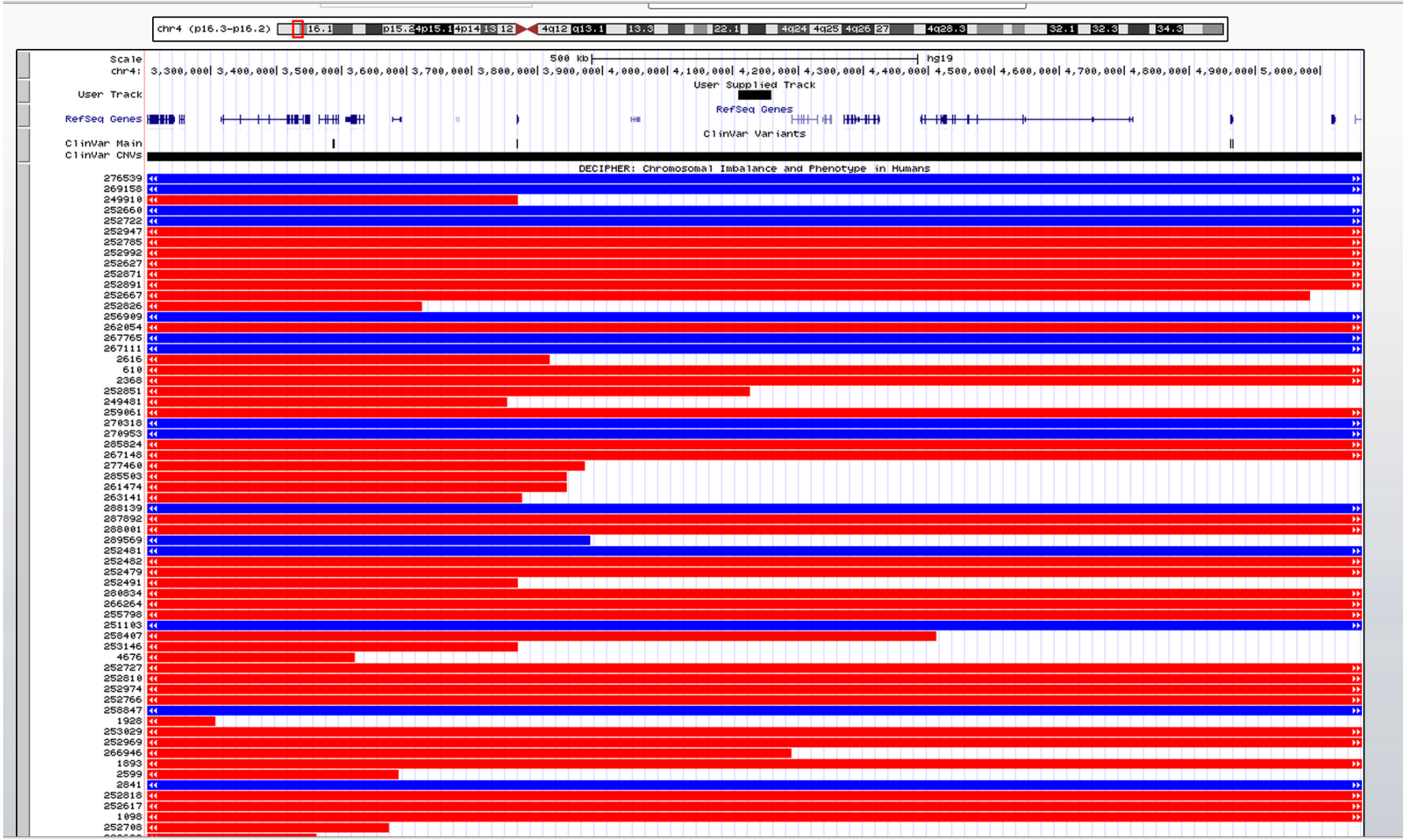
Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
4p16.3	~ 50 Kb.	intergenic	2/0	1525.08



Region Annotation

SSC_2

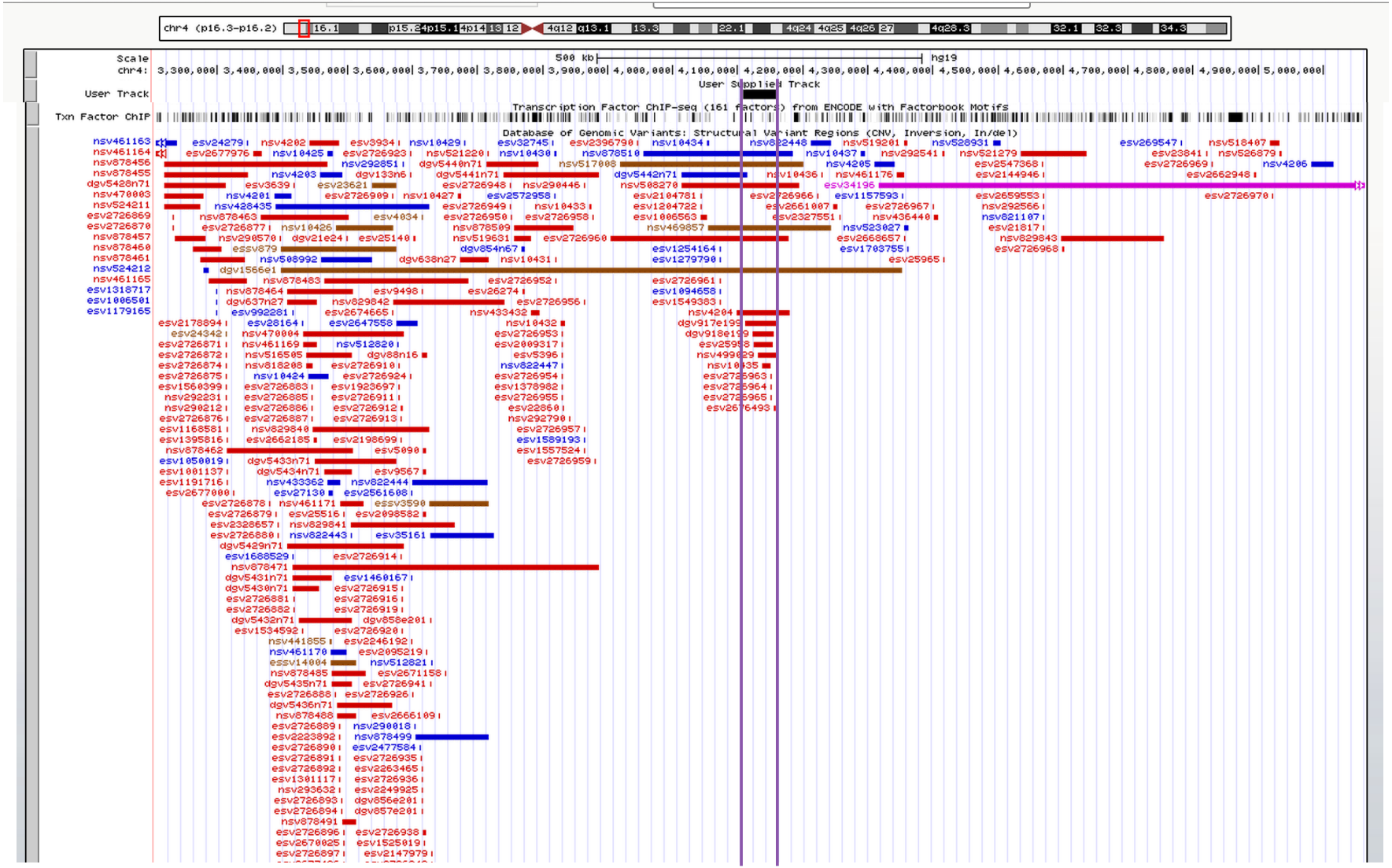
Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
4p16.3	~ 50 Kb.	intergenic	2/0	1525.08



Region Annotation Zoom Out

SSC_2

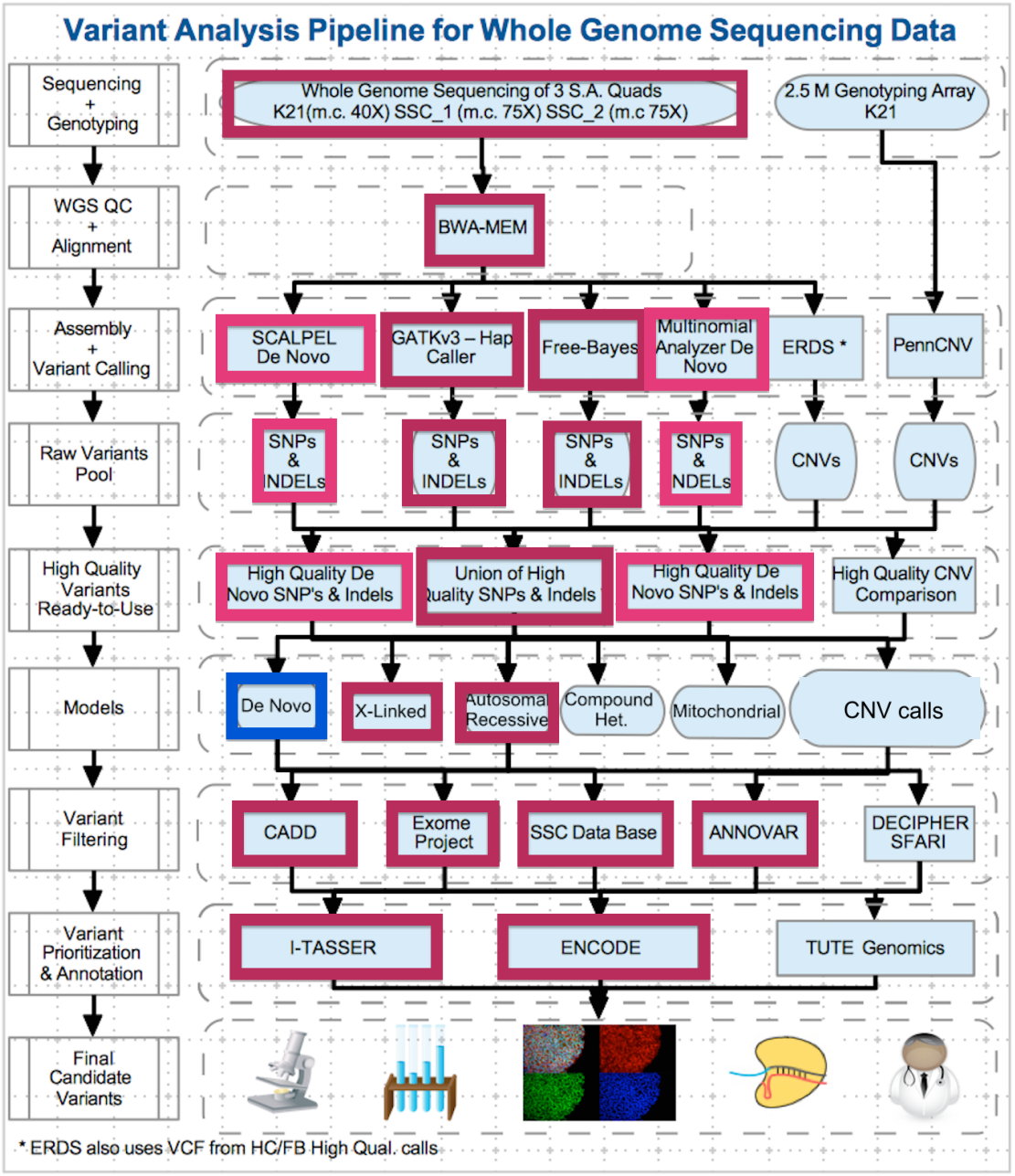
Position	Size	Gene(s)	Ref/Obs. # of Copies	ERDS Score
4p16.3	~ 50 Kb.	intergenic	2/0	1525.08



Key Points

- CNV Detection: Microarrays vs Sequencing
- Larger Amount of Samples
- Overlapping %
- Autosomal Recessive CNVs
- Regulatory Regions
- Functional Analysis
- Agregation of multiple variants

Smaller Variants



Algorithm	Parameters
BWA-MEM	Default Parameters
Picard Tools Mark Duplicates & Add Read Groups	Default Parameters
GATK Haplotype Caller 3.1-1	Default Parameters
GATK VQSR	Suggested parameters
Freebayes	Default Parameters
Filter FreeBayes	QUAL > 30
CADD	Score > 20
ANNOVAR	Ref genes / UCSC genes / DBGV / ENCODE / SSC / Exome Project

Neuron
Article

Multinomial Analyzer

Cell PRESS

De Novo Gene Disruptions in Children on the Autistic Spectrum

Ivan Iossifov,[1,6] Michael Ronemus,[1,6] Dan Levy,[1] Zihua Wang,[1] Inessa Hakker,[1] Julie Rosenbaum,[1] Boris Yamrom,[1] Yoon-ha Lee,[1] Giuseppe Narzisi,[1] Anthony Leotta,[1] Jude Kendall,[1] Ewa Grabowska,[1] Beicong Ma,[1] Steven Marks,[1] Linda Rodgers,[1] Asya Stepansky,[1] Jennifer Troge,[1] Peter Andrews,[1] Mitchell Bekritsky,[1] Kith Pradhan,[1] Elena Ghiban,[1] Melissa Kramer,[1] Jennifer Parla,[1] Ryan Demeter,[2] Lucinda L. Fulton,[2] Robert S. Fulton,[2] Vincent J. Magrini,[2] Kenny Ye,[3] Jennifer C. Darnell,[4] Robert B. Darnell,[4,5] Elaine R. Mardis,[2] Richard K. Wilson,[2] Michael C. Schatz,[1] W. Richard McCombie,[1] and Michael Wigler[1,*]
[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
[2]The Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA
[3]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA
[4]Laboratory of Molecular Neuro-oncology, Rockefeller University, New York, NY 10065, USA
[5]Howard Hughes Medical Institute, Rockefeller University, New York, NY 10065, USA
[6]These authors contributed equally to this work
*Correspondence: wigler@cshl.edu
DOI 10.1016/j.neuron.2012.04.009

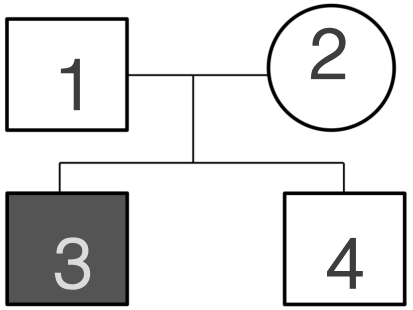
Scalpel

bioRχiv beta
THE PREPRINT SERVER FOR BIOLOGY

Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly

Giuseppe Narzisi, Jason A O'Rawe, Ivan Iossifov, et al.

Models



De Novo

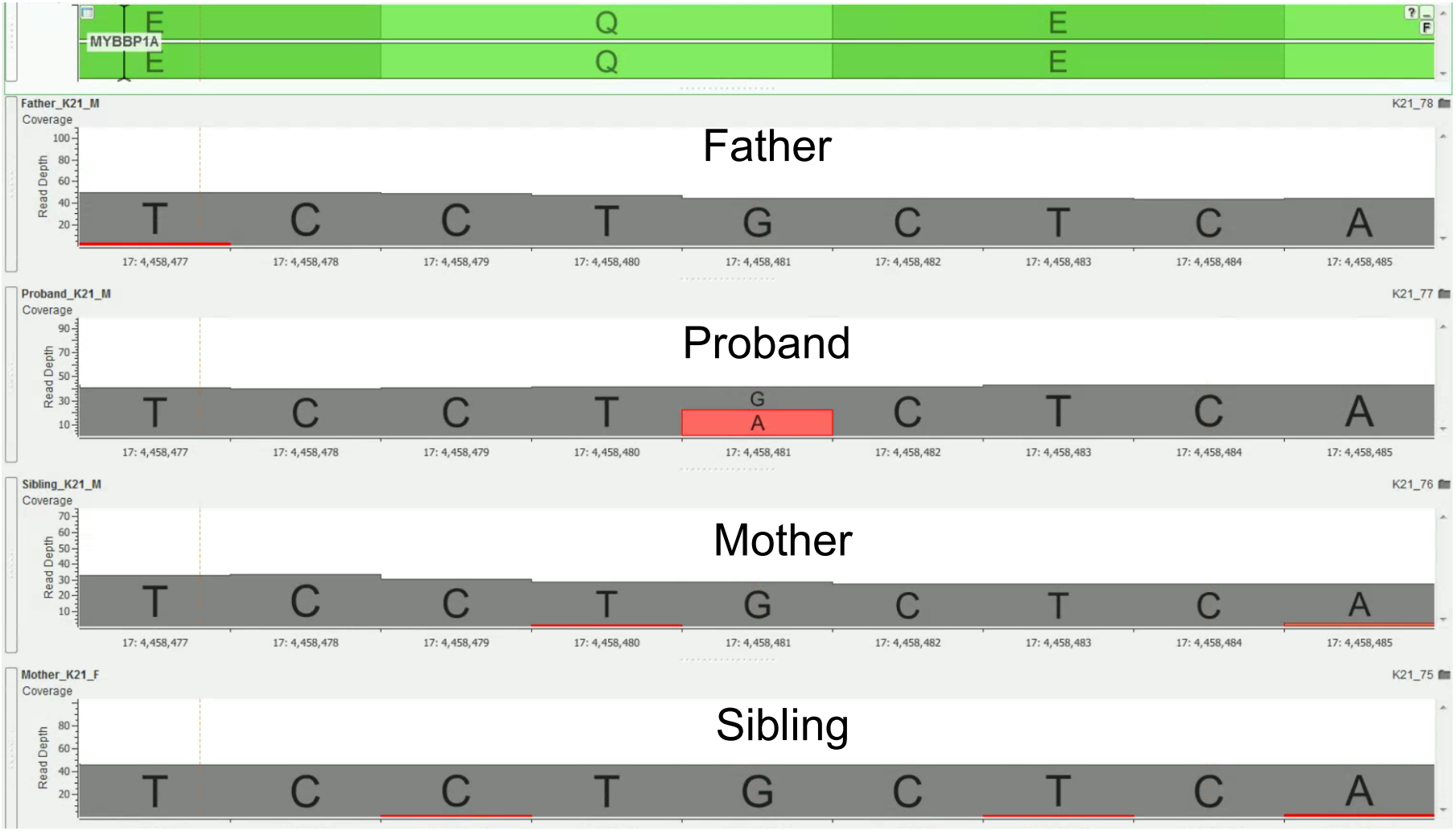
3 – (1 ∪ (2 ∪ 4))

Family	Position	CAAD score	Gene Mutation	Ref/Alt	Reads	Caller	Other DB
K_21	**17p13.2**	40	MYBBP1A: exon1:Stop	G/A	F44:0/ P19:22/ S28:0/ M45:0	FreeBayes/ HC/ Multinomial	SSC missense
SSC_1	**18q11.2**	36	intergenic	T/C	F8:1/ P9:11/ S9:0/ M7:0	FreeBayes/ HC/	NO

Region Alignment

K_21

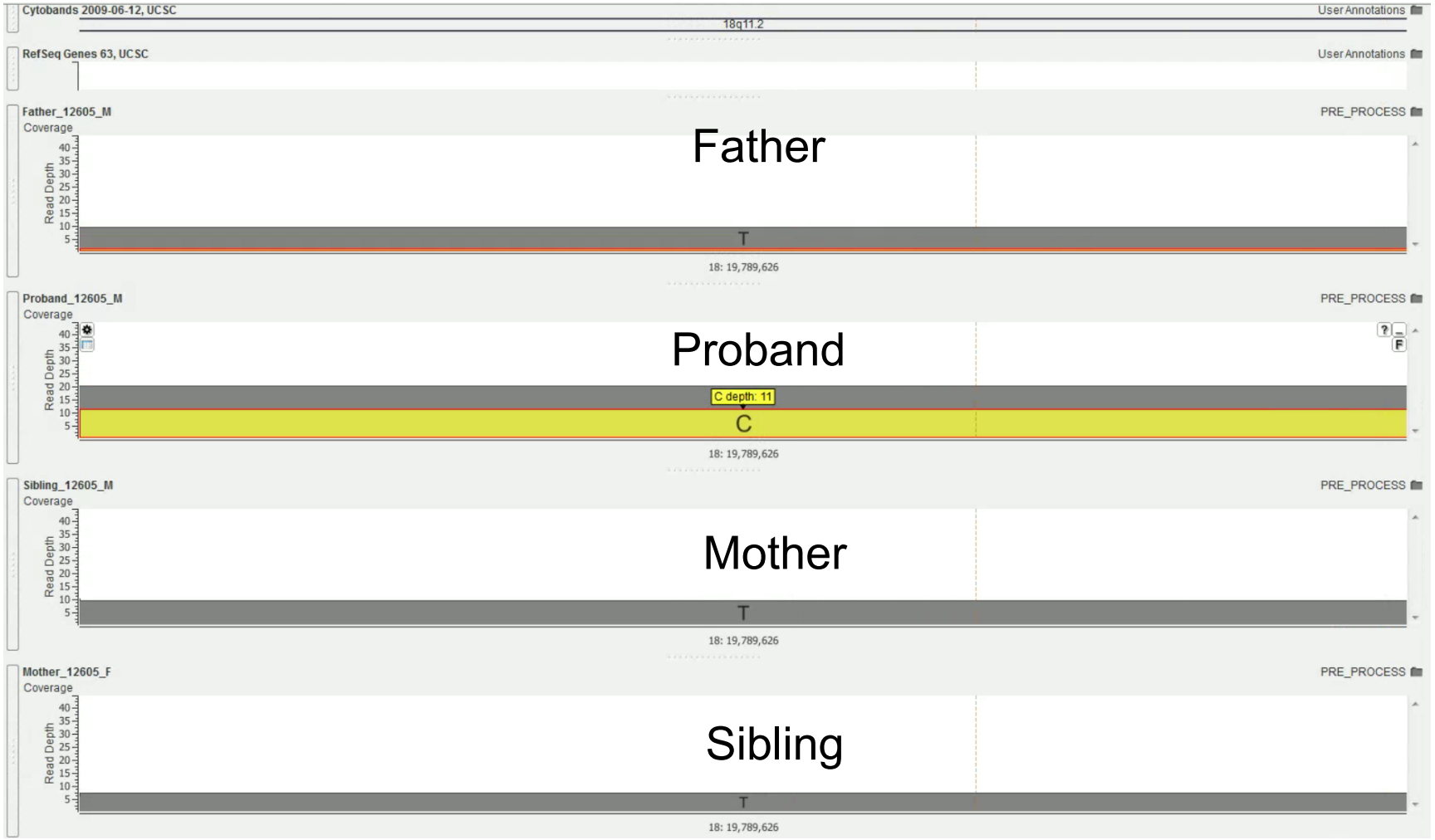
Position	CAAD score	Gene Mutation	Ref/Alt	Reads	Caller	Other DB
17p13.2	40	MYBBP1A: exon1:Stop	G/A	F44:0/ P19:22/ S28:0/ M45:0	FreeBayes/ HC/ Multinomial	SSC missense

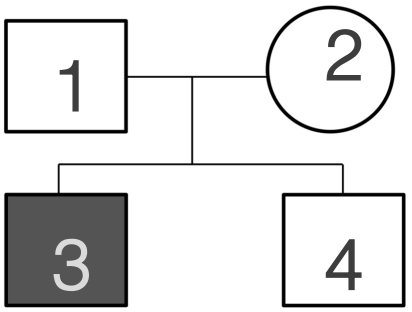


Region Alignment

SSC_1

Position	CAAD score	Gene Mutation	Ref/Alt	Reads	Caller	Other DB
18q11.2	36	intergenic	T/C	F8:1/ P9:11/ S9:0/ M7:0	FreeBayes/ HC/ Multinomial	NO





X - Linked

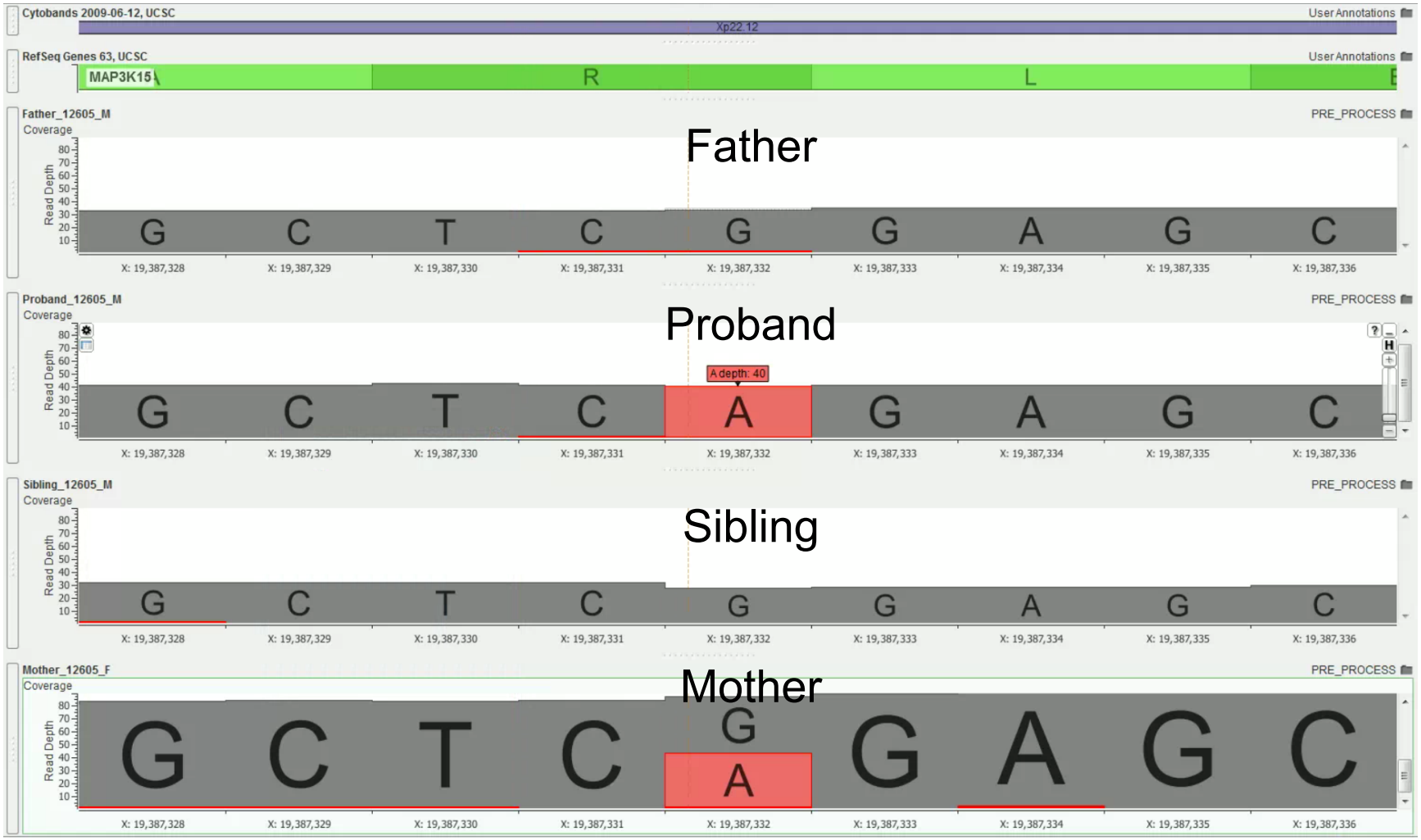
$$(1^{chrX} \cup 4^{chrX}) - (2^{chrX} \cap 3^{chrX})$$

Family	Position	CAAD score	Gene Mutation	Ref/Alt	Reads	Caller	Other DB
SSC_1	**Xp22.12**	39	MAP3K15 exon 25 of 26 Arg-Stop	G/A	F 32:0/ P 0:40 S 31:0/ M 44:42	FreeBayes/ GATK-HC/	NO

Region Alignment

SSC_1

Position	CAAD score	Gene Mutation	Ref/Alt	Reads	Caller	Other DB
Xp22.12	39	MAP3K15 exon 25 of 26 Arg-Stop	G/A	F 32:0/ P 0:40 S 31:0/ M 44:42	FreeBayes/ GATK-HC/	NO



Phyre²

SSC_1

3D Models

Position	CAAD score	Gene Mutation	Ref/Alt	Reads	Caller	Other DB
Xp22.12	39	MAP3K15 exon 25 of 26 Arg-Stop	G/A	F 32:0/ P 0:40 S 31:0/ M 44:42	FreeBayes/ GATK-HC/	NO

Complete Protein

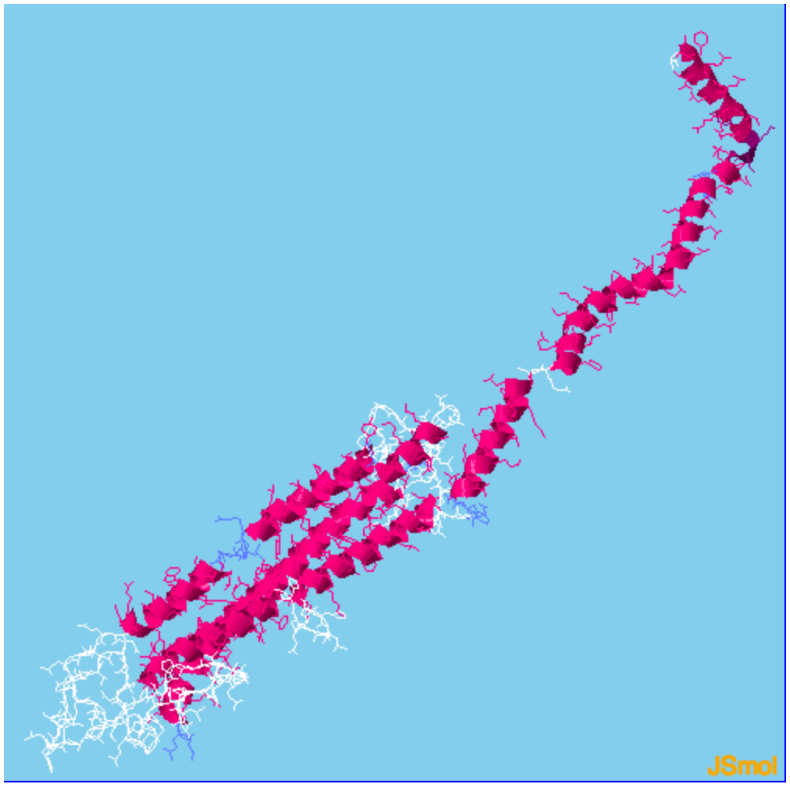
Incomplete Protein

3D Models

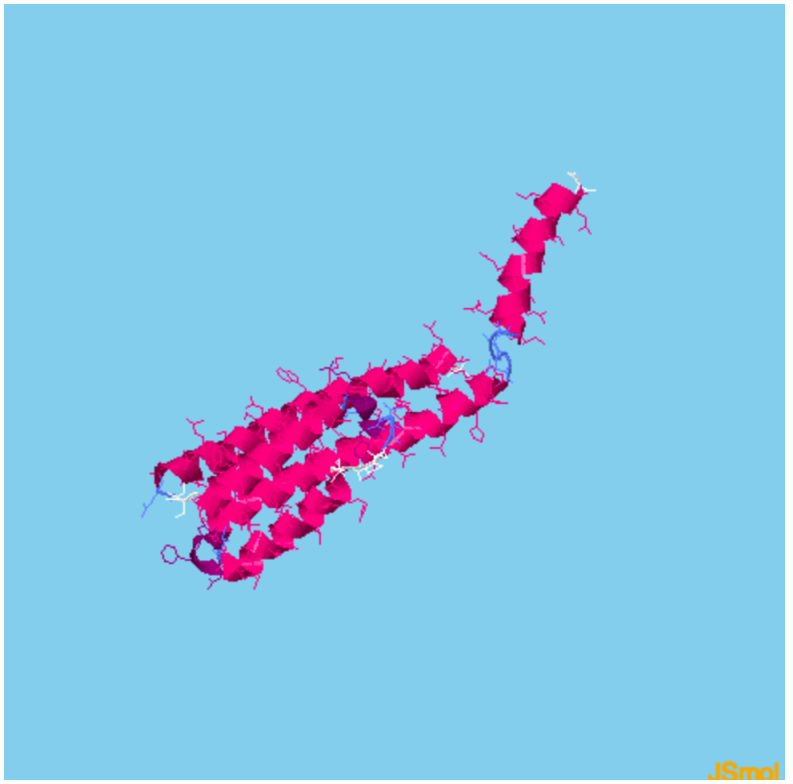
Position	CAAD score	Gene Mutation	Ref/Alt	Reads	Caller	Other DB
Xp22.12	39	MAP3K15 exon 25 of 26 Arg-Stop	G/A	F 32:0/ P 0:40 S 31:0/ M 44:42	FreeBayes/ GATK-HC/	NO

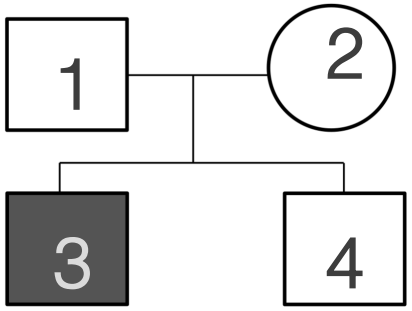
SSC_1

5th domain Complete Protein (1001-1313)



5th domain Incomplete Protein (104-1135)





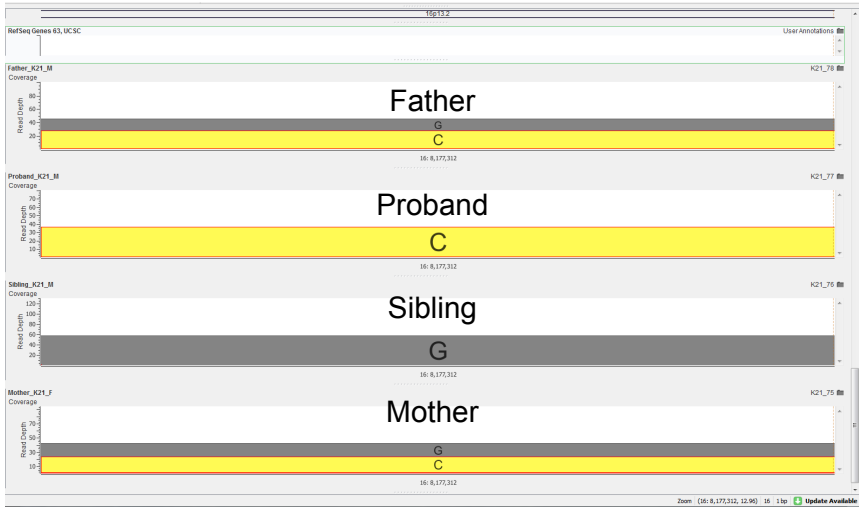
Autosomal Recessive

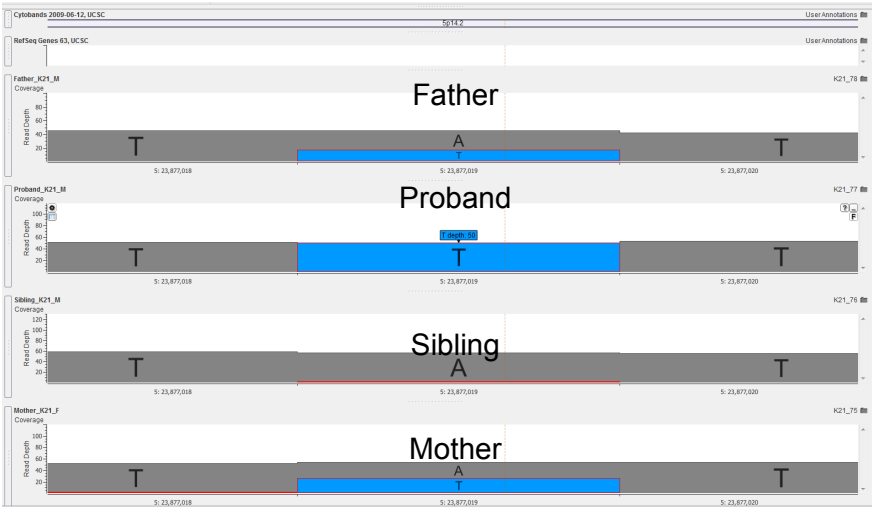
$$4 - (3_{h|hh} \cap (1_h \cap 2_h))$$

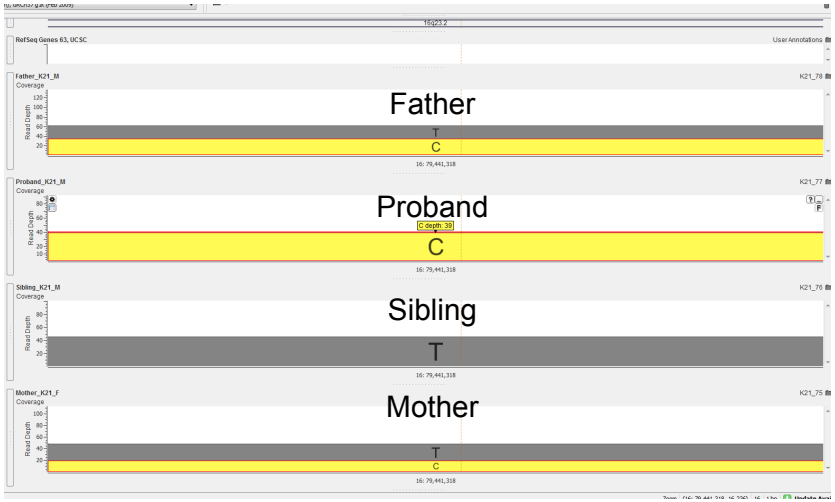
Family	Position	CAAD score	Gene Mutation	Ref/Alt	Reads	Caller	Other DB
K_21	16p13.2	30	intergenic	G/C	F 18:27 P0:36 S57:0 M19:22	FeeeBayes/ GATK HC	NO
K_21	16q36.2	37	intergenic	T/C	F27:34 P0:39 S45:0 M29:19	FeeeBayes/ GATK HC	NO
K_21	5p14.2	34	intergenic	A/T	F28:19 P0:50 S54:0 M27:26	FeeeBayes/ GATK HC	NO
K_21	9q31.1	33	intergenic	G/A	P19:16 P0:37 S38:0 M13:24	FeeeBayes/ GATK HC	NO

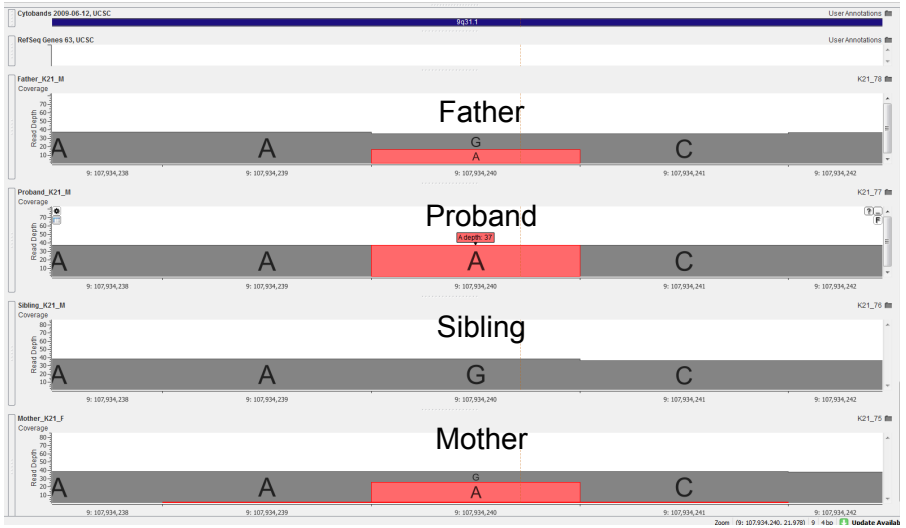
Autosomal Recessive

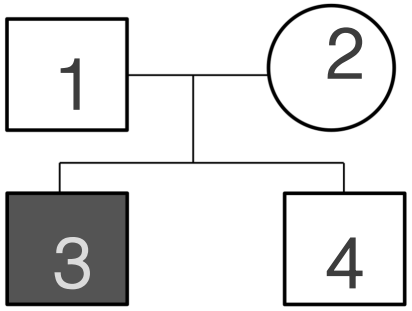
K_21











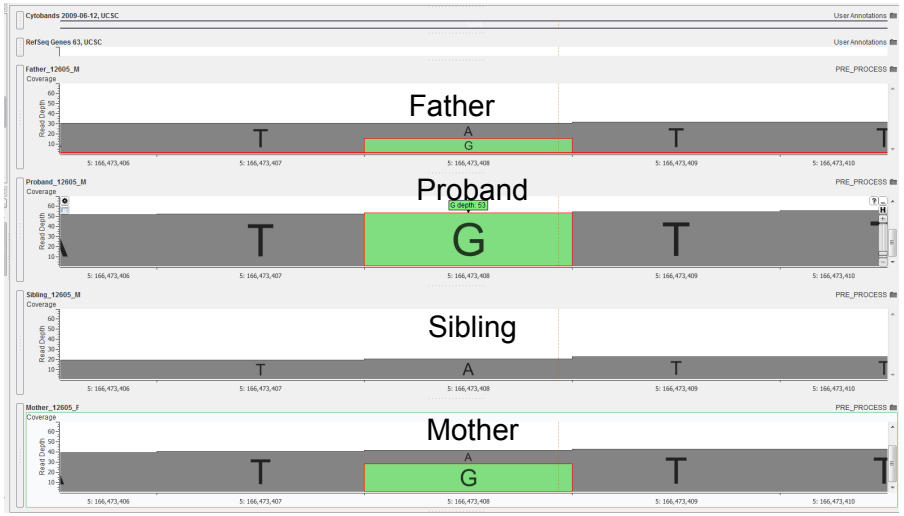
Autosomal Recessive

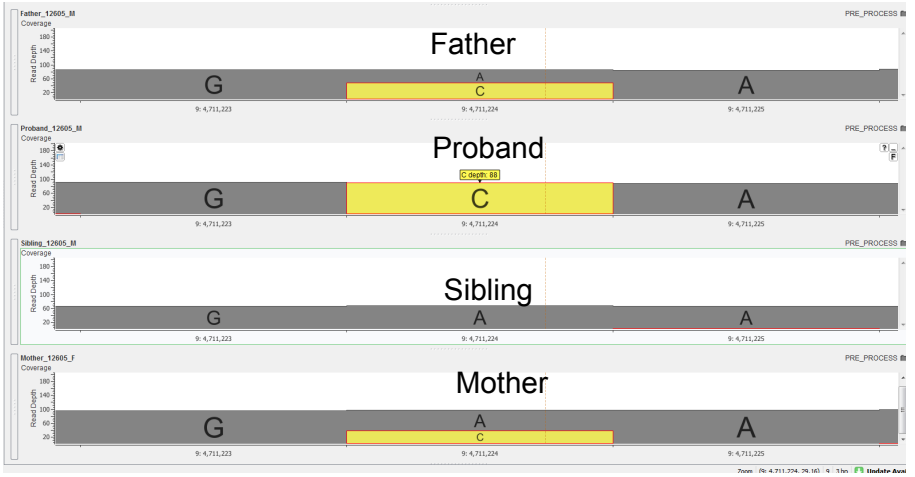
$$4 - (3_{h|hh} \cap (1_h \cap 2_h))$$

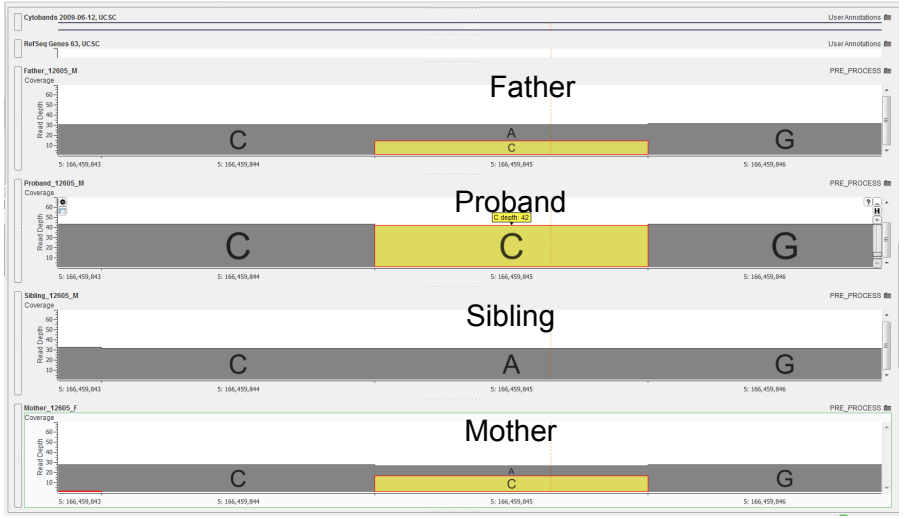
Family	Position	CAAD score	Gene Mutation	Ref/Alt	Reads	Caller	Other DB
SSC_1	5q34	32	intergenic	A/G	F15:14 P0:53 S20:0 M13:28	FeeeBayes/ GATK HC	NO
SSC_1	5q34	32	intergenic	A/C	F16:14 P0:42 S31:0 M10:16	FeeeBayes/ GATK HC	NO
SSC_1	15q26.2	32	intergenic	G/A	F26:34 P0:58 S44:0 M43:39	FeeeBayes/ GATK HC	NO
SSC_1	9p24.1	31	intergenic	A/C	F38:46 P0:88 S66:0 M59:37	FeeeBayes/ GATK HC	NO

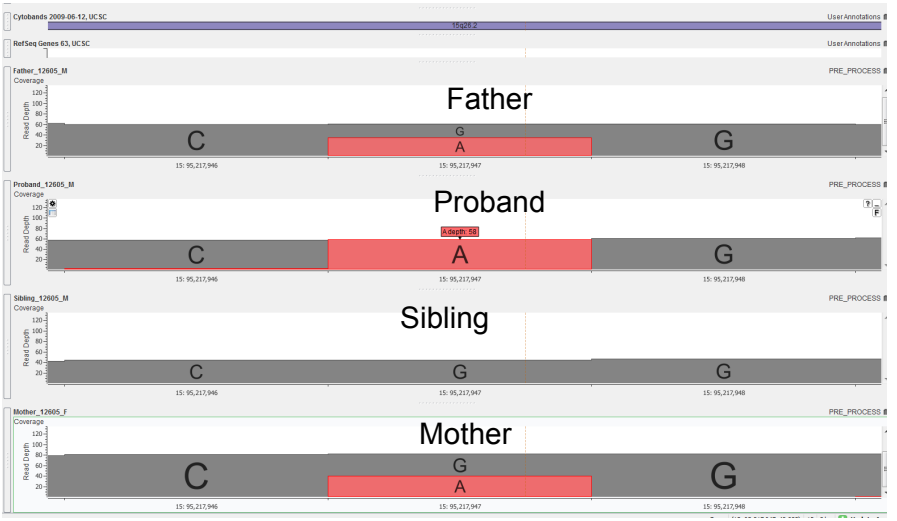
SSC_1

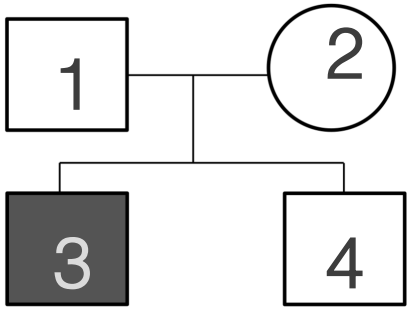
Autosomal Recessive











Compound Heterozygous

3 – (4 ∩ ((1 – 2) ∪ (2 – 1)))

Annotate all HQ variants

↓

Get variants sharing a gene

↓

Filter out exact match of combination of variants for a gene in controls

Conclussions

- Larger Amount of Samples for statistical power.
- No common pathogenic variants among different probands.
- Not only De Novo or Gene Disrupting Variants could be playing a role.
- Functional Analysis.

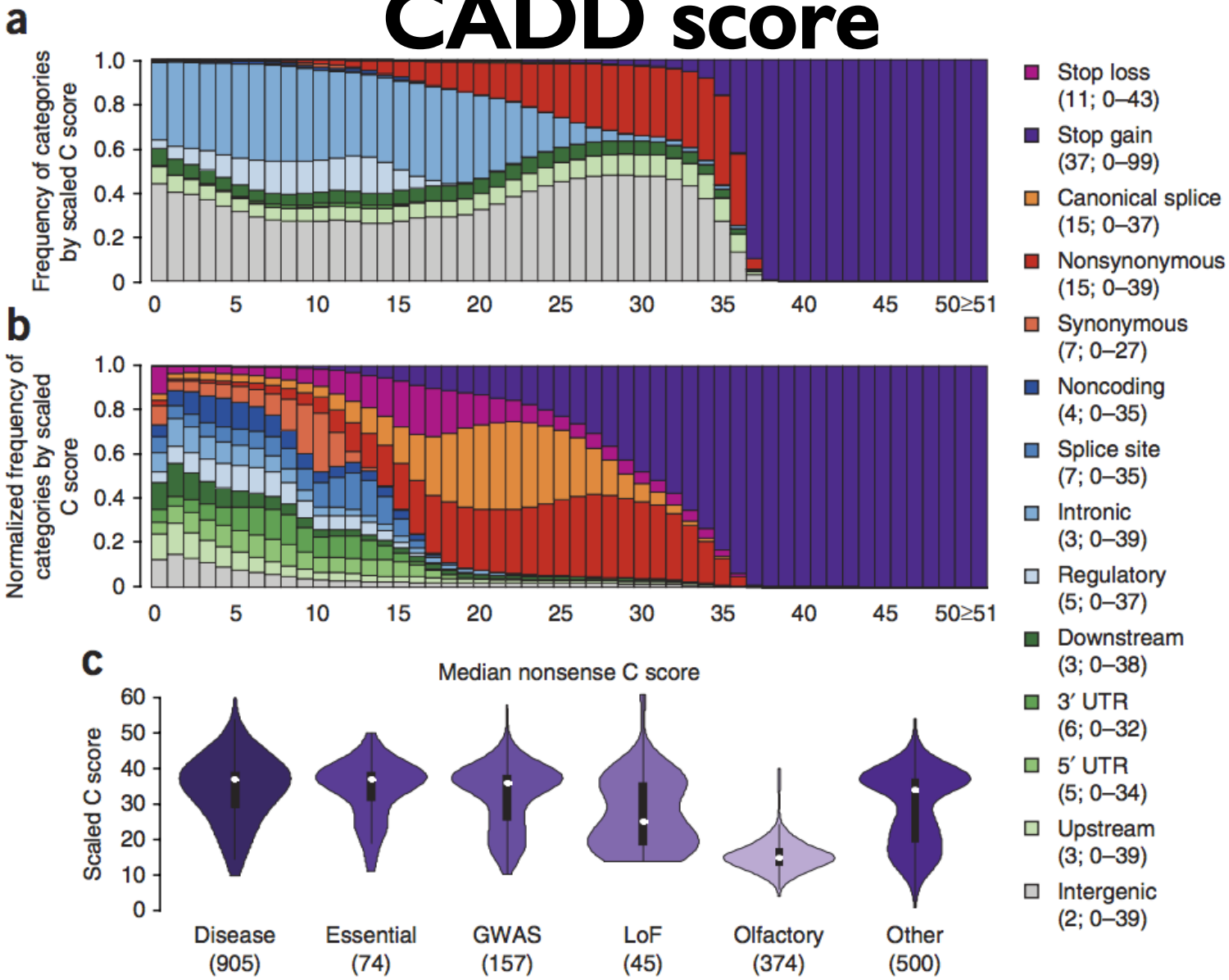
THANKS!

BAF

$$BAF = \begin{cases} 0, \text{ if } \theta < \theta_{AA} \\ 0.5(\theta - \theta_{AA})/(\theta_{AB} - \theta_{AA}), \text{ if } \theta \leq \theta < \theta_{AB} \\ 0.5 + 0.5(\theta - \theta_{AB})/(\theta_{BB} - \theta_{AB}), \text{ if } \theta_{AB} \leq \theta < \theta_{BB} \\ 1, \text{ if } \theta \geq \theta_{BB} \end{cases} \quad (1)$$

where θ_{AA}, θ_{AB}, and θ_{BB} are the θ values for three canonical genotype clusters generated from a large set of reference samples. The transformation from θ to BAF values adjusts for different chemical characteristics of each SNP so that values for different SNPs are more comparable to each other.

CADD score


a
Frequency of categories by scaled C score
1.0
0.8
0.6
0.4
0.2
0
0 5 10 15 20 25 30 35 40 45 50≥51
b
Normalized frequency of categories by scaled C score
1.0
0.8
0.6
0.4
0.2
0
0 5 10 15 20 25 30 35 40 45 50≥51
c
Median nonsense C score
Scaled C score
60
50
40
30
20
10
0
Disease (905)
Essential (74)
GWAS (157)
LoF (45)
Olfactory (374)
Other (500)
Stop loss (11; 0–43)
Stop gain (37; 0–99)
Canonical splice (15; 0–37)
Nonsynonymous (15; 0–39)
Synonymous (7; 0–27)
Noncoding (4; 0–35)
Splice site (7; 0–35)
Intronic (3; 0–39)
Regulatory (5; 0–37)
Downstream (3; 0–38)
3′ UTR (6; 0–32)
5′ UTR (5; 0–34)
Upstream (3; 0–39)
Intergenic (2; 0–39)