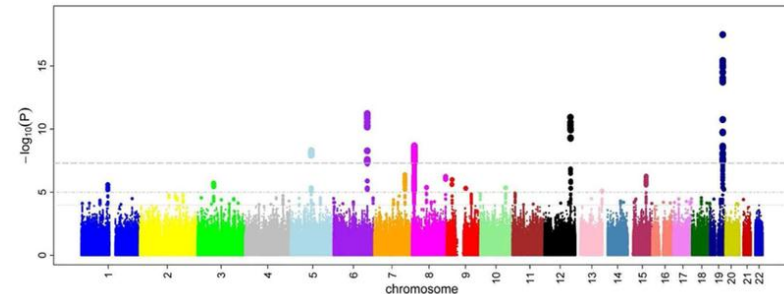




DEPTH: A Novel Algorithm for Feature Ranking with Application to Genome-Wide Association Studies

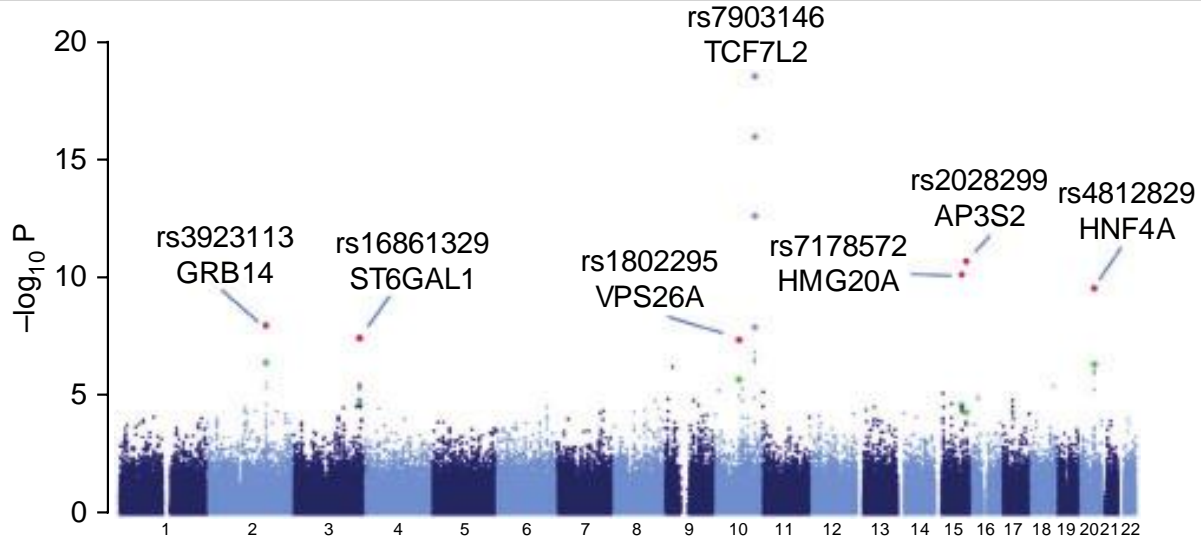
- Novel algorithm for feature ranking
- Ultra-high dimensional data sets
 - e.g., genome-wide association studies, epigenome-wide association studies
- Which features correspond to signal/noise?

- Genome-wide Association Studies
 - Case-control setup
 - N samples, p markers (SNPs)
 - Allele frequency in cases vs. controls
 - Report effect size as an odds ratio



- Conventional approach to analysis
 - Test each marker independently
 - Calculate frequentist p -value for each marker
 - Adjust for multiple testing
 - All p -values less than threshold considered true associations
 - e.g., Genome-wide significance threshold

Manhattan Plot and Skyline



- Difficult statistical problem
 - Large number of markers
 - Correlated data
 - Disease causing variants not measured
 - Conventional analysis minimises family-wise error rate
 - Highly conservative publications

- Our strategy: DEPTH
 - NHMRC Project grant
 - Designed to run in a parallel environment
 - Exploits data correlation structure
 - Examine all markers, or subset of markers
 - e.g., all markers in a gene or pathway of genes



chr6:144,578,898-163,835,234

19,256,337 bp.

enter position, gene symbol or search terms

go

Scale
chr6:

150,000,000

5 Mb

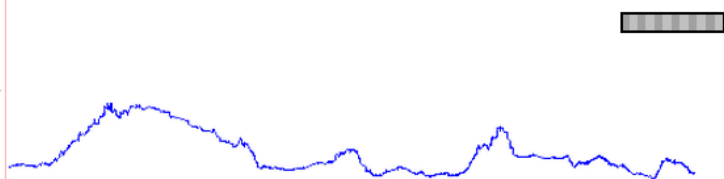
155,000,000

hg18

160,000,000

User Supplied Track 1

ER-ve 1

UTRN
UTRN

UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)



THE EVOLUTION STARTS HERE

chr6:144,578,898-163,835,234

19,256,337 bp.

enter position, gene symbol or search terms

go

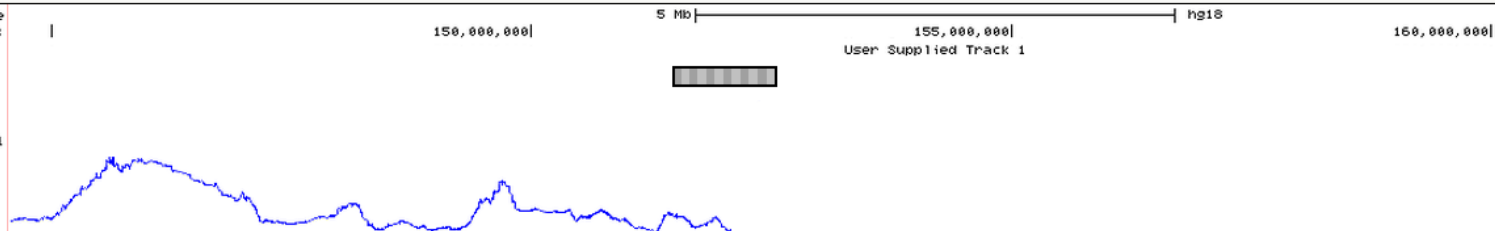


5 Mb

hg18

User Supplied Track 1

ER-ve 1



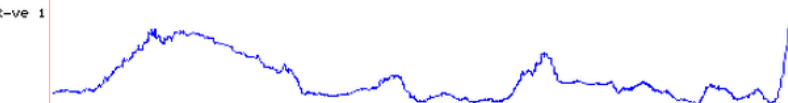
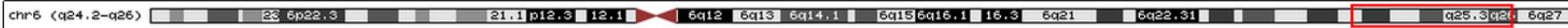
THE EVOLUTION STARTS HERE

chr6:144,578,898-163,835,234

19,256,337 bp.

enter position, gene symbol or search terms

go



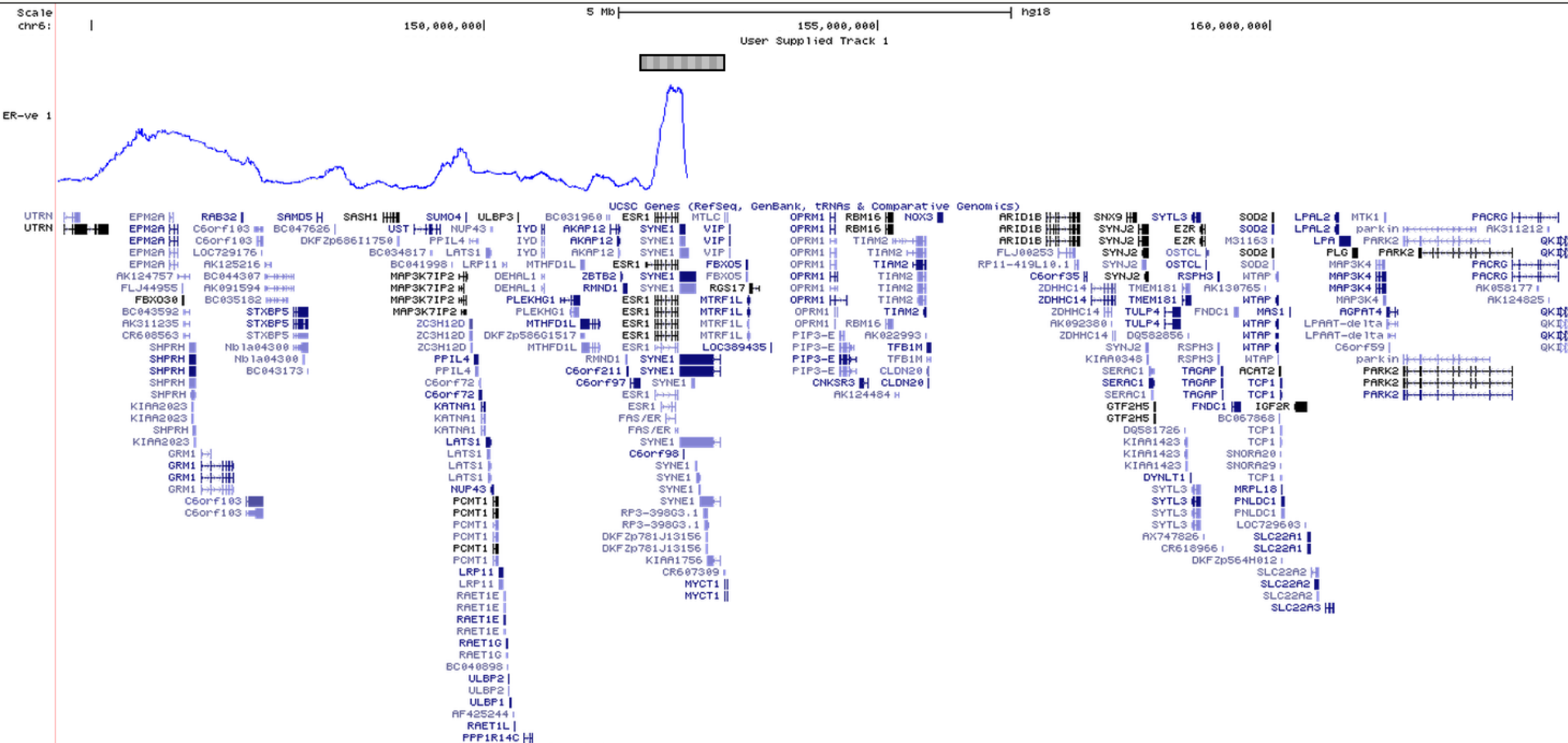
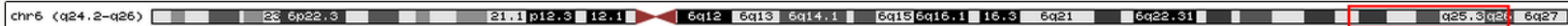
THE EVOLUTION STARTS HERE

chr6:144,578,898-163,835,234

19,256,337 bp.

enter position, gene symbol or search terms

go

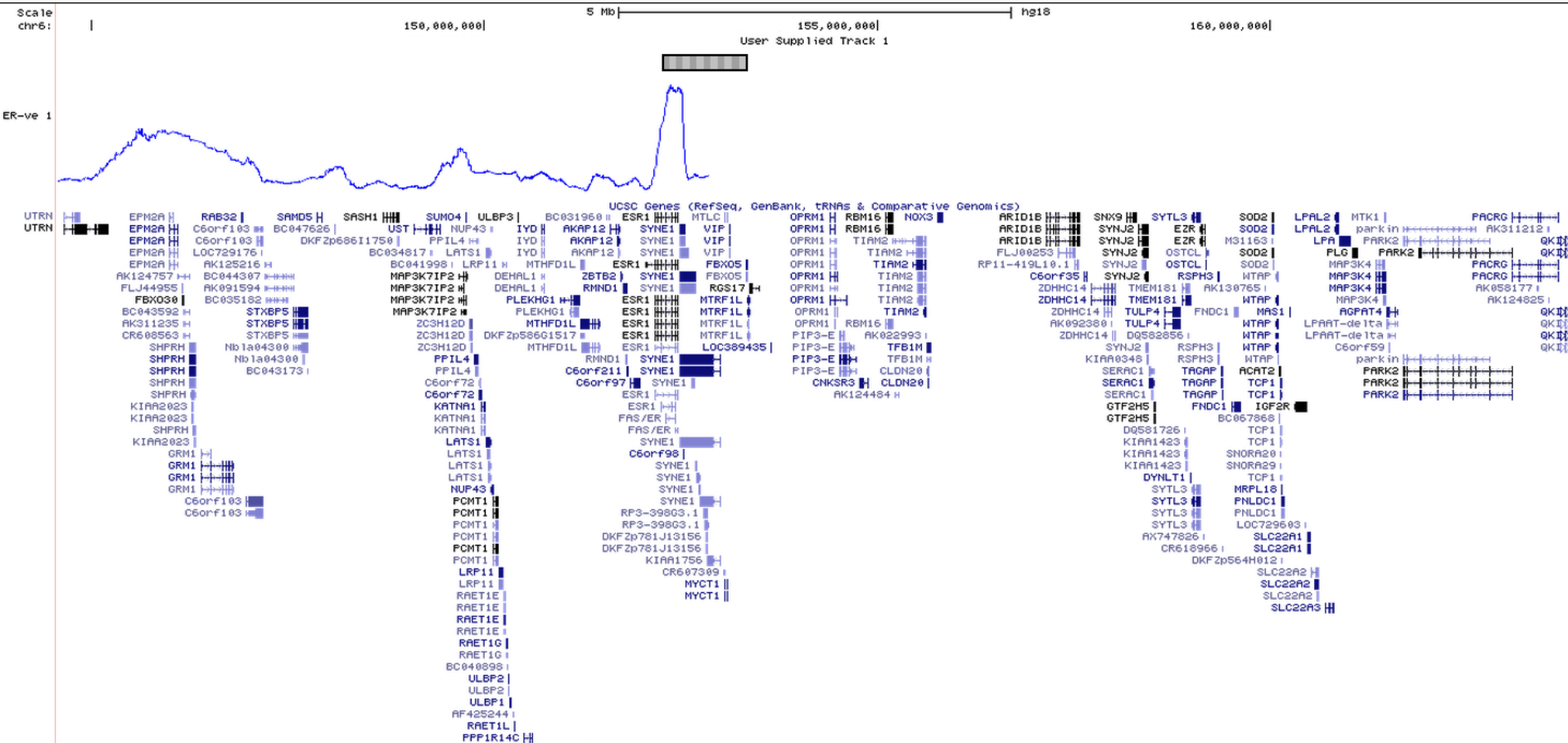
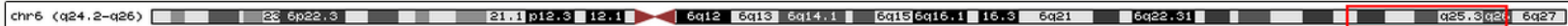


THE EVOLUTION STARTS HERE

chr6:144,578,898-163,835,234

19,256,337 bp. enter position, gene symbol or search terms

go



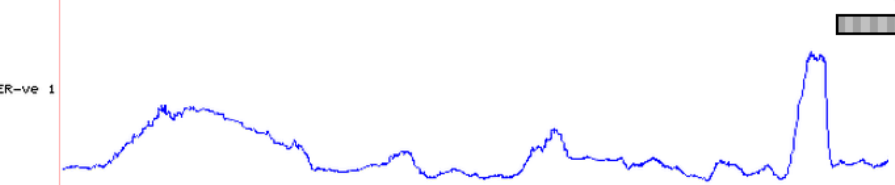
THE EVOLUTION STARTS HERE

chr6:144,578,898-163,835,234

19,256,337 bp.

enter position, gene symbol or search terms

go



UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)

EPN2A	RAB32	SAMD5	SASH1	SUMO4	ULBP3	BC031960	ESR1	MTLC	OPRM1	RBM15	NOX3	ARID1B	SNX9	SYTL3	SOD2	LPAL2	MTK1	PACRG
EPN2A	C6orf103	BC047626	UST	NUP43	IYD	AKAP12	SYNE1	VIP	OPRM1	RBM15		ARID1B	SYNJ2	EZR	SOD2	LPAL2	park in	AK311212
EPN2A	C6orf103	DKFZp68611750	PPIL4	IYD	AKAP12	SYNE1	VIP		OPRM1	TIAM2		ARID1B	SYNJ2	EZR	M31163	LPA	PARK2	QK1
EPN2A	LOC729176		BC034817	LATS1	IYD	AKAP12	SYNE1	VIP	OPRM1	TIAM2		FLJ06253	SYNJ2	OSTCL	SOD2	PLG	PARK2	QK1
EPN2A	AK125216		BC041998	LRP11	MTHFD1L	ESR1	FBXO5		OPRM1	TIAM2		RP11-419L10.1	SYNJ2	OSTCL	SOD2			
AK124757	BC044307		MAP3K7IP2	DEHAL1	ZBTB2	SYNE1	FBXO5		OPRM1	TIAM2		C6orf95	SYNJ2	RSPH3	NTAP	MAP3K4		PACRG
FLJ44955	AK091594		MAP3K7IP2	DEHAL1	RMND1	SYNE1	ROS17		OPRM1	TIAM2		ZDHHC14	TMEM181	AK130755	NTAP	MAP3K4		PACRG
FBXO30	BC035182		MAP3K7IP2	PLEKHG1	ESR1	MTRF1L			OPRM1	TIAM2		ZDHHC14	TMEM181	NTAP	NTAP	MAP3K4		AK124825
BC043592	STXBP5		MAP3K7IP2	M	ESR1	MTRF1L			OPRM1	TIAM2		ZDHHC14	TULP4	FNDC1	MAS1	AGPAT4		QK1
AK311235	STXBP5		2C3H12D	MTHFD1L	ESR1	MTRF1L			OPRM1	RBM15		AK092380	TULP4	NTAP	LPART-de lta			QK1
CR608563	STXBP5		2C3H12D	DKFZp586G1517	ESR1	MTRF1L			PIP3-E	AK022993		ZDHHC14	DG582856	NTAP	LPART-de lta			QK1
SHPRH	Nb1a04300		2C3H12D	MTHFD1L	ESR1	LOC399435			PIP3-E	TFB1M				NTAP	C6orf59			QK1
SHPRH	Nb1a04300		PPIL4		RMND1	SYNE1			PIP3-E	TFB1M				NTAP		park in		
SHPRH	BC043173		PPIL4		C6orf211	SYNE1			PIP3-E	CLDN20				NTAP		PARK2		
SHPRH			C6orf72		C6orf97	SYNE1			CNKSR3	CLDN20				NTAP		PARK2		
SHPRH			C6orf72		ESR1					AK124484				NTAP		PARK2		
KIAR2023			PPIL4		ESR1									NTAP		PARK2		
KIAR2023			PPIL4		FAS/ER									NTAP		PARK2		
SHPRH			KATNA1		FAS/ER									NTAP		PARK2		
KIAR2023			KATNA1											NTAP		PARK2		
GRM1			LATS1		SYNE1									NTAP		PARK2		
GRM1			LATS1		SYNE1									NTAP		PARK2		
GRM1			LATS1		SYNE1									NTAP		PARK2		
GRM1			LATS1		SYNE1									NTAP		PARK2		
C6orf103			NUP43		SYNE1									NTAP		PARK2		
C6orf103			PCMT1		SYNE1									NTAP		PARK2		
			PCMT1		RP3-398G3.1									NTAP		PARK2		
			PCMT1		RP3-398G3.1									NTAP		PARK2		
			PCMT1		DKFZp781J13156									NTAP		PARK2		
			PCMT1		DKFZp781J13156									NTAP		PARK2		
			PCMT1		DKFZp781J13156									NTAP		PARK2		
			PCMT1		KIAR1756									NTAP		PARK2		
			LRP11		CR607309									NTAP		PARK2		
			LRP11		MYCT1									NTAP		PARK2		
			RAET1E		MYCT1									NTAP		PARK2		
			RAET1E											NTAP		PARK2		
			RAET1E											NTAP		PARK2		
			RAET1E											NTAP		PARK2		
			RAET1G											NTAP		PARK2		
			BC040899											NTAP		PARK2		
			ULBP2											NTAP		PARK2		
			ULBP2											NTAP		PARK2		
			ULBP1											NTAP		PARK2		
			ULBP1											NTAP		PARK2		
			AF425244											NTAP		PARK2		
			RAET1L											NTAP		PARK2		
			PPP1R14C											NTAP		PARK2		

THE EVOLUTION STARTS HERE

chr6:144,578,898-163,835,234

19,256,337 bp.

enter position, gene symbol or search terms

go

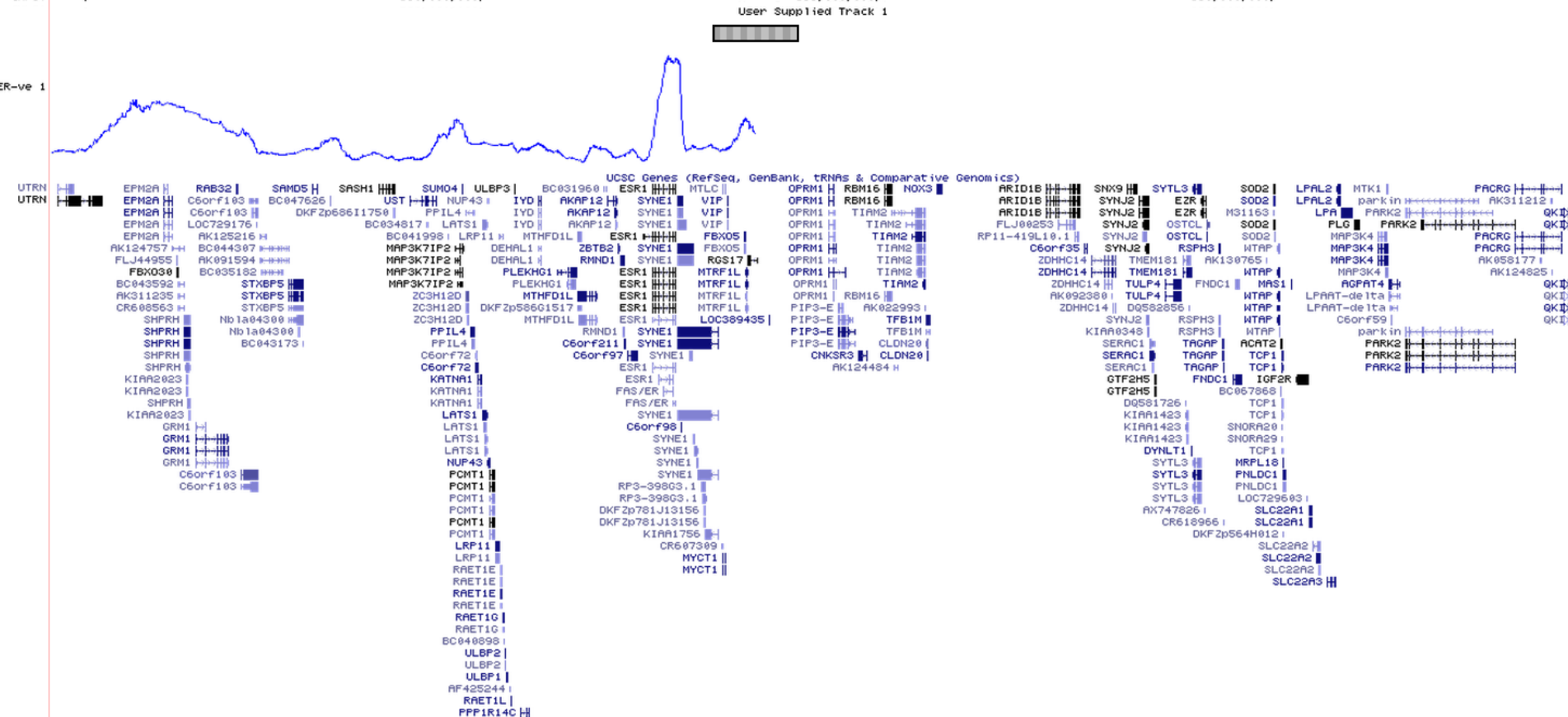


chr6:144,578,898-163,835,234

19,256,337 bp.

enter position, gene symbol or search terms

go



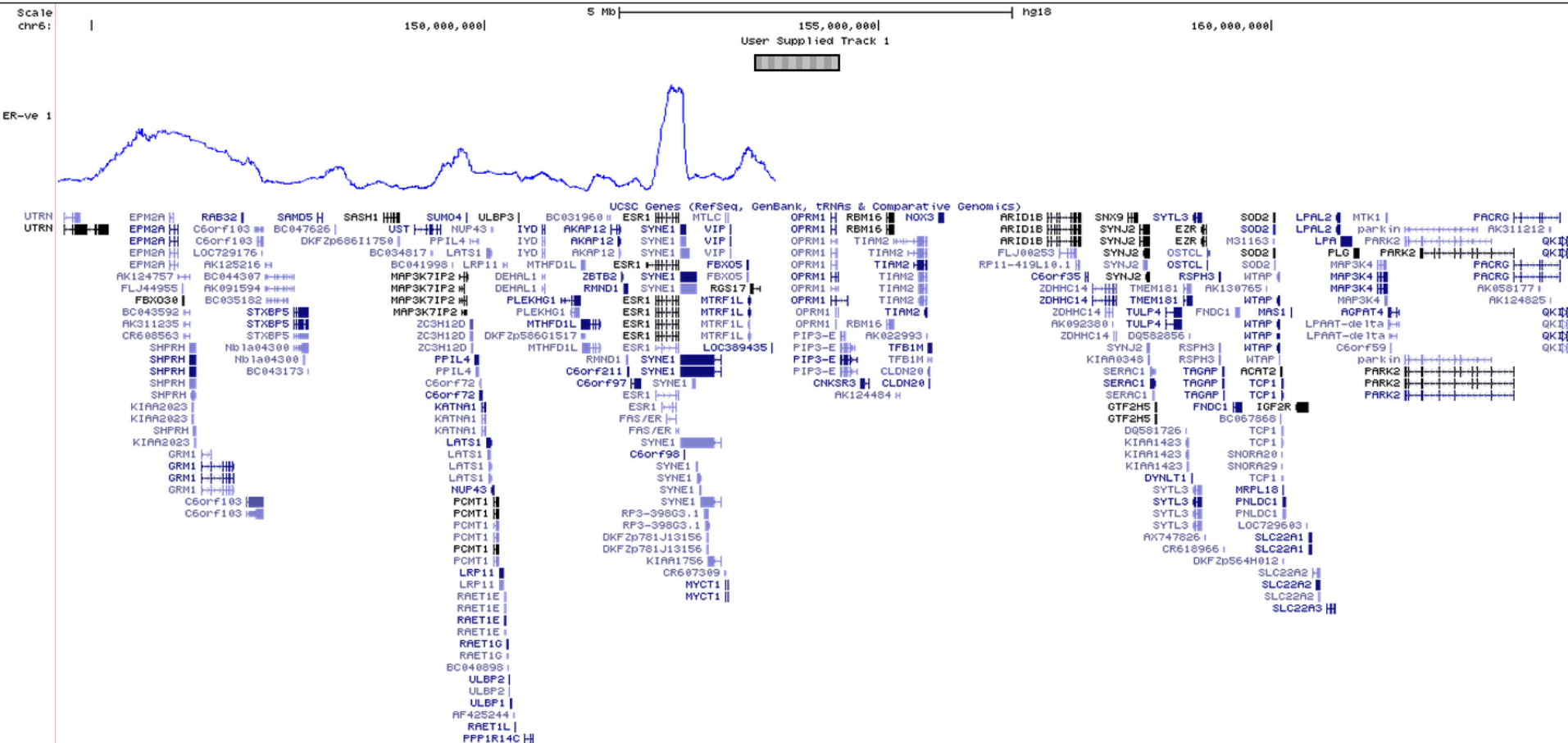
THE EVOLUTION STARTS HERE

chr6:144,578,898-163,835,234

19,256,337 bp.

enter position, gene symbol or search terms

go



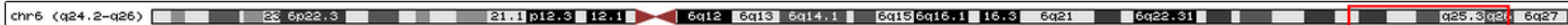
THE EVOLUTION STARTS HERE

chr6:144,578,898-163,835,234

19,256,337 bp.

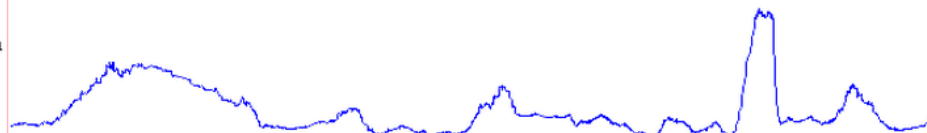
enter position, gene symbol or search terms

go



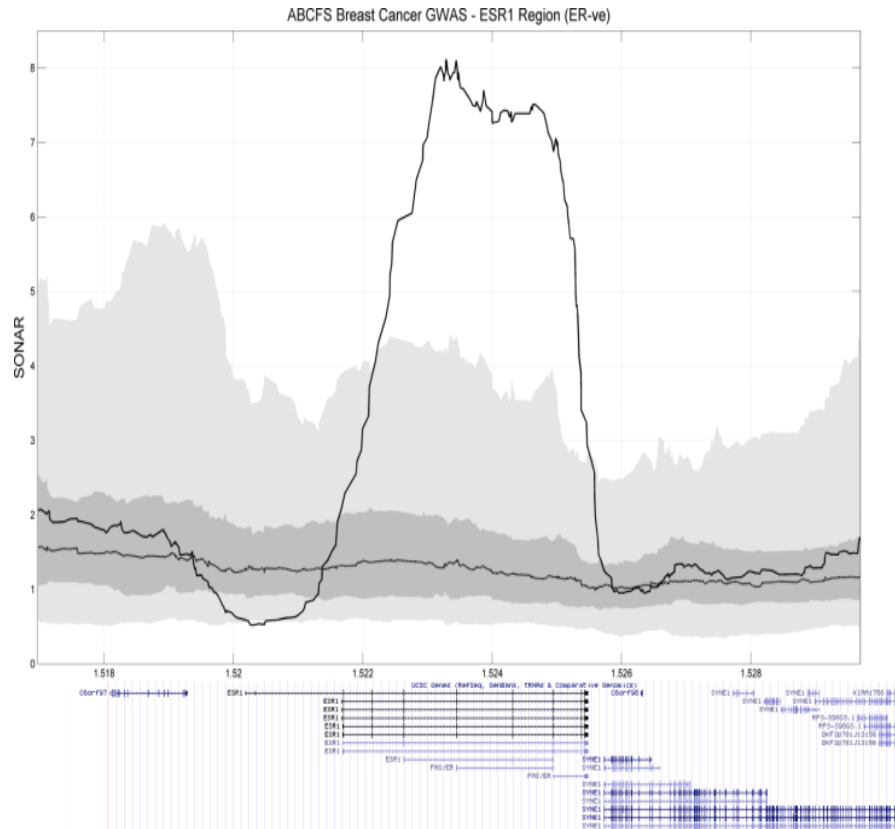
User Supplied Track 1

ER-ve 1

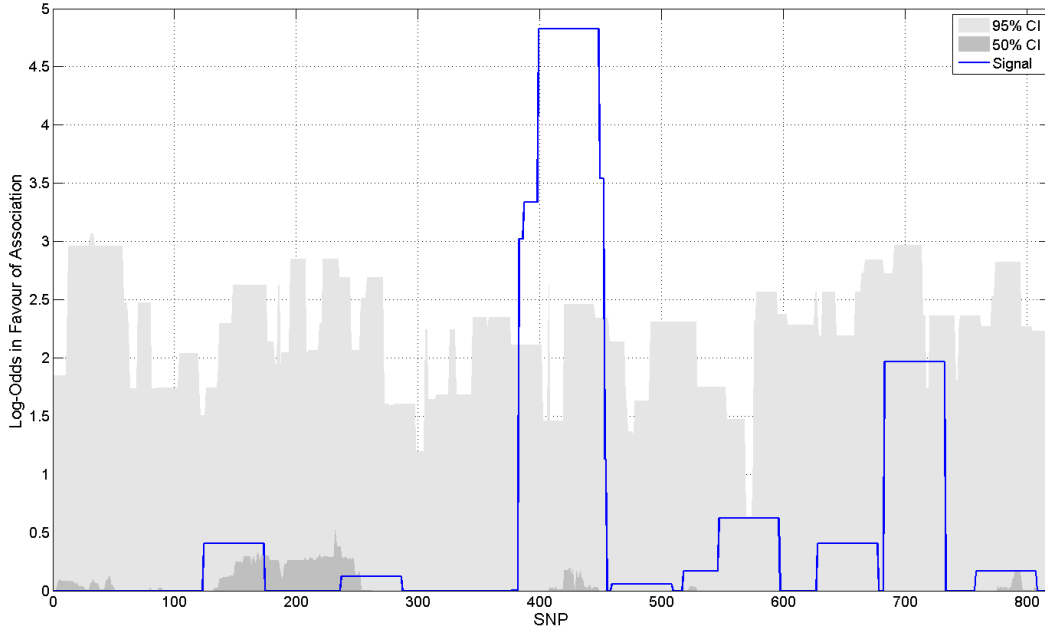


THE EVOLUTION STARTS HERE

Is this peak “statistically significant”?



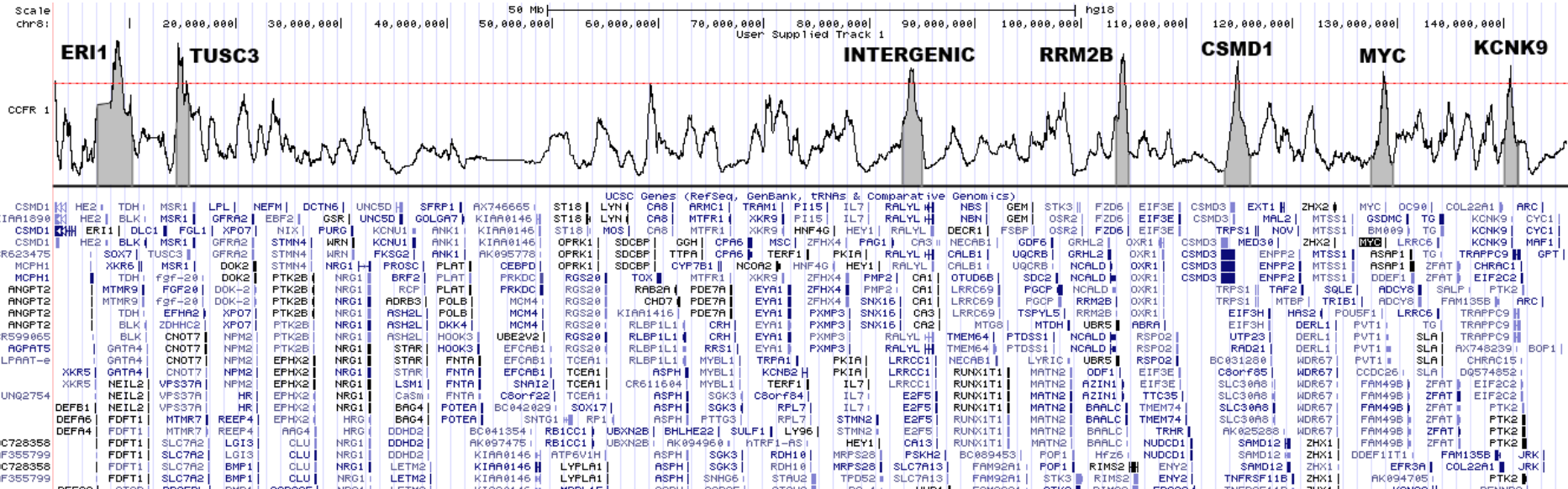
A non-parametric measure of signal



- CCFR Phase I GWAS data set
 - 1,179 cases and 998 controls
 - 2,121,264 markers
 - Standard QC
 - DEPTH genome-wide analysis (NHMRC Project Grant)

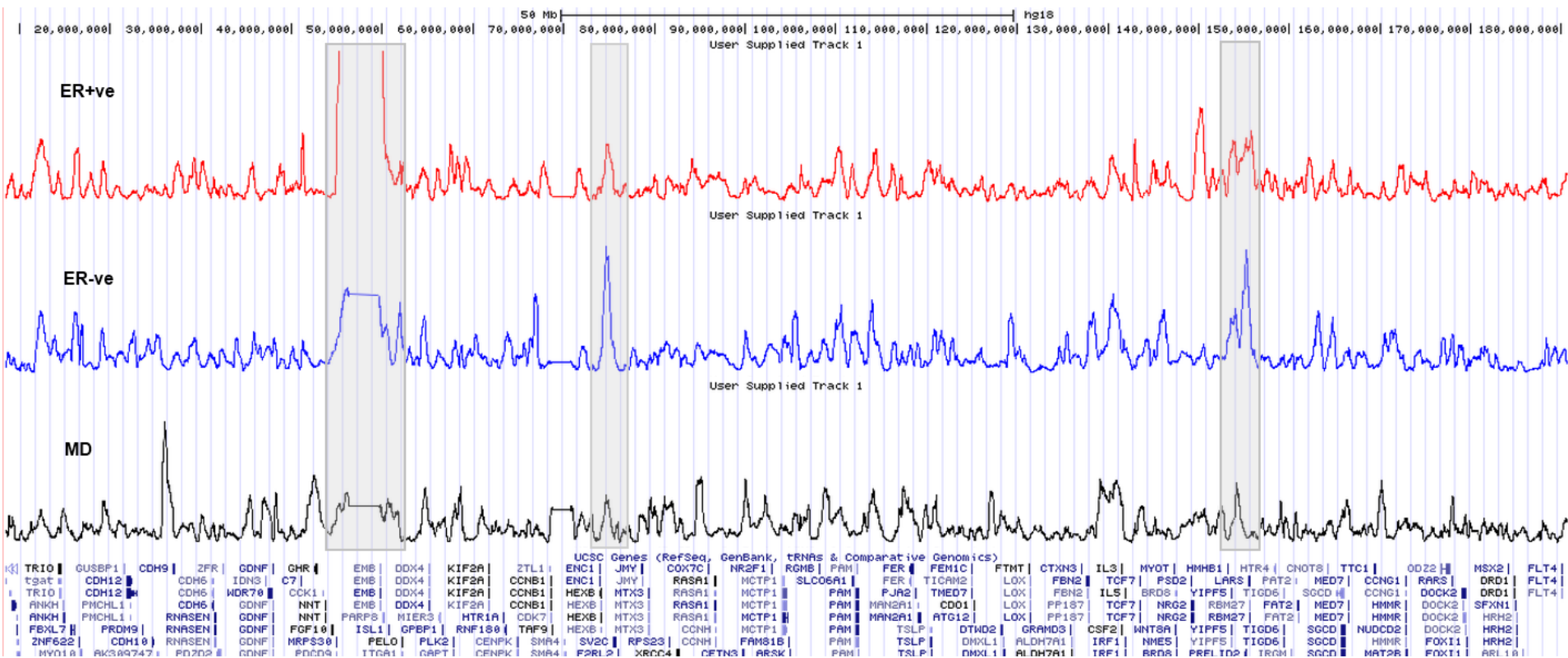
- DEPTH identified ~60 genomic regions associated with risk of CRC
- The genomic regions:
 - Some already known (e.g., *ERCC6*, *SMAD7*, *DCC*)
 - Many novel regions (e.g., *TUSC3*, *VIM*, *LIMA1*)

DEPTH and CCFR Colorectal Cancer GWAS data (3)



- DEPTH sub-analyses
 - Proximal colon versus distal colon and rectum
 - Mismatch repair deficient versus proficient tumours
 - Case-case analysis (stratified by age)
- DEPTH analysis of combined CRC GWAS and EWAS data

Future work – Replication (2)





Thank
You