Prioritization of *cis*-regulatory variants in cancer using whole-genome sequencing and integrative analysis of ChIP-seq and chromatin-state data

Hamid Bolouri
Div. Human Biology
Fred Hutchinson Cancer Research Center

http://labs.fhcrc.org/bolouri

National Cancer Institute

TARGET

Therapeutically Applicable Research to Generate Effective Treatments

http://target.cancer.gov/

NIH

Daniela Gerhardt

Tanja Davidson,…

JHMI (DNA Methylation)

Robert Arceci

Jason Farrar, …

FHCRC (pediatric AML)

Soheil Meshinchi

Rhonda Ries

Ranjani Ramamurthy

Kavita Garg (Tewari lab)

Phoenix Ho, …

Thanks to: Ali Shojaei (UW Biostats)

Paul Shannon & Martin Morgan (Bioconductor team)

CHILDREN'S ONCOLOGY GROUP

The world's childhood cancer experts

Todd Alonzo
Alan Gamis
Rob Gerbing

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

Current TARGET AML data sets:

2 x 138 whole genome sequences

(+ 66 relapse samples)

225+4 microarrays

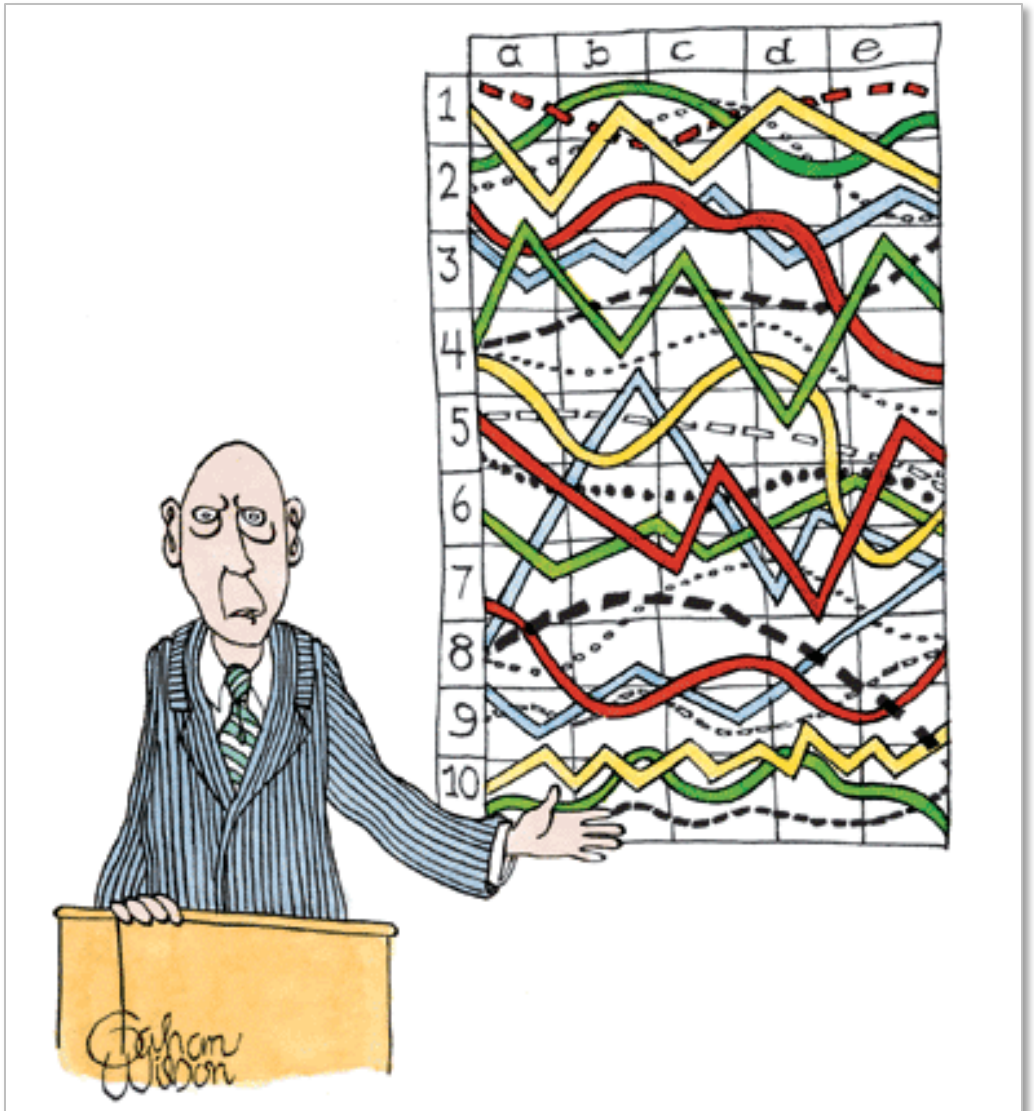
187 methylation arrays

182 miRNA-seqs (not discussed)

> 50 clinical data elements/sample

If a slide is confusing,

please interrupt & ask questions!



"I'll pause for a moment so you can let this information sink in."

cis-Regulatory Mutations Are a Genetic Cause of Human Limb Malformations

Julia E. VanderMeer and Nadav Ahituv

DEVELOPMENTAL DYNAMICS 240:920–930, 2011

TABLE 1. Enhancer Defects Known to Cause Limb Malformations in Human Patients

Mutation Name	Mutation	Location (hg19)	Phenotype	Reference
***BMP2* limb enhancer**				
Family 1, Dathe	duplication	~chr20:6,860,129-6,866,024	Brachydactyly type A2	Dathe et al., 2009
Family 2, Dathe	duplication	~chr20:6,860,477-6,866,024	Brachydactyly type A2	Dathe et al., 2009
***DLX5/6* BS1 enhancer (~chr7:96,357,368-96,357,92)**				
Patient, Kouwenhoven	deletion	~chr7:95,552,064-96,432,064	Split hand/foot malformation1	Kouwenhoven et al., 2010
***SHH* ZRS enhancer (~chr7:156,583,562-156,584,711)**				
739 A>G, Family A,C	SNP	chr7:156,583,831	Preaxial polydactyly & triphalangeal thumb	Gurnett et al., 2007
621 C>G, Family B	SNP	chr7:156,583,949	Preaxial polydactyly & triphalangeal thumb	Gurnett et al., 2007
463 T>G	SNP	chr7:156,584,107	Preaxial polydactyly & triphalangeal thumb	Farooq et al., 2010
404 G>C, Family 2	SNP	chr7:156,584,166	Werner mesomelic syndrome	Wieczorek et al., 2009
404 G>A, Family 1	SNP	chr7:156,584,166	Werner mesomelic syndrome	Wieczorek et al., 2009
404 G>A, Cuban	SNP	chr7:156,584,166	Preaxial polydactyly	Lettice et al., 2003
396 C>T, Turkish 1	SNP	chr7:156,584,174	Preaxial polydactyly & triphalangeal thumb	Semerci et al., 2009
334 T>G, French 2	SNP	chr7:156,584,236	Preaxial polydactyly	Albuisson et al., 2010
323 T>C, Belgian 2	SNP	chr7:156,584,241	Preaxial polydactyly	Lettice et al., 2003
30 5A>T, Belgian 1	SNP	chr7:156,584,266	Preaxial polydactyly	Lettice et al., 2003
297 G>A, French 1	SNP	chr7:156,584,273	Preaxial polydactyly	Albuisson et al., 2010
295 T>C	SNP	chr7:156,584,275	Triphalangeal thumb	Furniss et al., 2008
105 C>G, Dutch	SNP	chr7:156,584,465	Preaxial polydactyly	Lettice et al., 2003
Case, Lettice	translocation	t(5,7)(q11,q36)	Preaxial polydactyly & triphalangeal thumb	Lettice et al., 2002
Family, Klopocki	duplication	~chr7:156,143,386-156,732,204	Triphalangeal thumb-polysyndactyly	Klopocki et al., 2008
Family 6, Sun	duplication	~chr7:156,241,020-156,699,998	Triphalangeal thumb-polysyndactyly	Sun et al., 2008
Family 2, Sun	duplication	~chr7:156,241,020-156,677,759	Triphalangeal thumb-polysyndactyly	Sun et al., 2008
Family 5, Sun	duplication	~chr7:156,241,020-156,619,399	Syndactyly type IV	Sun et al., 2008
Family 4, Sun	duplication	~chr7:156,354,085-156,687,613	Triphalangeal thumb-polysyndactyly	Sun et al., 2008
Family 3, Sun	duplication	~chr7:156,354,085-156,619,399	Triphalangeal thumb-polysyndactyly	Sun et al., 2008
Family 3, Wieczorek	duplication	~chr7:156,368,541-156,661,877	Triphalangeal thumb-polysyndactyly	Wieczorek et al., 2009
Family 1, Sun	duplication	~chr7:156,539,605-156,699,998	Triphalangeal thumb-polysyndactyly	Sun et al., 2008
Family, Wu	duplication	~chr7:156,547,469-156,644,074	Syndactyly & tibial hypoplasia	Wu et al., 2009
Family 4, Wieczorek	duplication	~chr7:156,572,751-156,661,877	Triphalangeal thumb-polysyndactyly	Wieczorek et al., 2009
***SOX9* limb enhancer**				
Critical region	duplication	~chr17:65,642,665-66,847,686	Brachydactyly-anonychia	Kurth et al., 2009

Kleinjan & van Heyningen, Am. J. Hum. Genet., 2005, (76)8–32

Position-Effect Genes in Human Diseases

Gene	Gene Function	Domains/Motifs	Disease	Distance of Furthest Breakpoint[a] (kb)	3′ or 5′ Side	Reference
PAX6	TF	Paired box and homeodomain	Aniridia	125	3′	Kleinjan et al. 2001
TWIST	TF	…	Saethre-Chotzen syndrome	260	3′	Cai et al. 2003
POU3F4	TF	POU homeodomain	X-linked deafness	900	5′	de Kok et al. 1996
PITX2	TF	Homeodomain	Rieger syndrome	90	5′	Trembath et al. 2004
GLI3	TF	Zinc finger	Greig cephalopolysyndactyly syndrome	10	3′	Wild et al. 1997
MAF	TF	bZIP	Cataract, ocular anterior segment dysgenesis, and coloboma	1,000	5′	Jamieson et al. 2002
FOXC1	TF	Forkhead	Glaucoma/autosomal dominant iridogoniodysgenesis	25/1,200	5′	Davies et al. 1999
FOXC2	TF	Forkhead	Lymphedema distichiasis	120	3′	Fang et al. 2000
FOXL2	TF	Forkhead	Blepharophimosis-ptosis-epicanthus inversus syndrome	170	5′	Crisponi et al. 2004
SOX9	TF	HMG box	Campomelic dysplasia	850	5′	Bagheri-Fam et al. 2001; Pop et al. 2004
SRY	TF	HMG box	Sex reversal	3	5′/3′	McElreavy et al. 1992
SIX3	TF	Homeodomain	Holoprosencephaly (HPE2)	<200	5′	Wallis et al. 1999
SHH	Signaling	…	Holoprosencephaly (HPE3)	265	5′	Roessler et al. 1997
SHH	Signaling	…	Preaxial polydactyly	1,000	5′	Lettice et al. 2003
SHFM1	TF	DLX5/DLX6?	Split-hand/split-foot malformation	~450	5′/3′	Crackower et al. 1996
FSHD	??	…	Facioscapulohumeral dystrophy	100	3′	Gabellini et al. 2002; Jiang et al. 2003; Masny et al. 2004
HBB	Oxygen carrier	Globin	$\gamma\beta$-Thalassemia	50	5′	Kioussis et al. 1983
HBA	Oxygen carrier	Globin	α-Thalassemia	18	3′	Tufarelli et al. 2003
Hoxd complex	TF	Homeodomain	Mesomelic dysplasia and vertebral defects	60	3′	Spitz et al. 2002
LCT	Enzyme	Lactase	Lactase persistence	15/20	5′	Enattah et al. 2002

Gene	Disease	Location of rSNP	TF-binding site affected
HBB	β-thalassemia	Promoter	Several (TATA, CACCC, EKLF)
F9	Hemophilia B	Promoter	Several (HNF4, C/EBP)
LDLR	Familial hypercliolesterolemia	Promoter	Several (SPI, SRE repeat)
CollAl	Osteoporosis	Intron I (+2kb)	SPI (gain)
RET	Hirschprung	IntronI (+9.7 kb)	Unknown
HBA	α-thalassemia	Upstream (−13 kb)	GATAI (gain)
SHH	Preaxial polydactyly	Upstream (−1 Mb)	Unknown
SHH	Holoprosencephaly	Upstream (−470 kb)	Six3
SOX9	Pierre Robin Sequence	Upstream (−1.5 Mb)	MsxI
IRF6	Nonsyndromic cleft lip	Upstream (−14kb)	Ap2

Epstein, Briefings in Func. Genom. & Protemoics, 2009, 8(4)310-16

Epigenomic Enhancer Profiling Defines a Signature of Colon Cancer

Batool Akhtar-Zaidi,[1,2] Richard Cowper-Sal•lari,[3] Olivia Corradin,[1] Alina Saiakhova,[1] Cynthia F. Bartels,[1] Dheepa Balasubramanian,[1] Lois Myeroff,[4] James Lutterbaugh,[4] Awad Jarrar,[5] Matthew F. Kalady,[4,5,6] Joseph Willis,[4,7] Jason H. Moore,[3] Paul J. Tesar,[1,4] Thomas Laframboise,[1,4] Sanford Markowitz,[1,4,8] Mathieu Lupien,[3,9] Peter C. Scacheri[1,2,4*]

Cancer is characterized by gene expression aberrations. Studies have largely focused on coding sequences and promoters, even though distal regulatory elements play a central role in controlling transcription patterns. We used the histone mark H3K4me1 to analyze gain and loss of enhancer activity genome-wide in primary colon cancer lines relative to normal colon crypts. We identified thousands of variant enhancer loci (VELs) that comprise a signature that is robustly predictive of the in vivo colon cancer transcriptome. Furthermore, VELs are enriched in haplotype blocks containing colon cancer genetic risk variants, implicating these genomic regions in colon cancer pathogenesis. We propose that reproducible changes in the epigenome at enhancer elements drive a specific transcriptional program to promote colon carcinogenesis.

Science **336**, 736
11 MAY 2012

Mice Lacking a *Myc* Enhancer That Includes Human SNP rs6983267 Are Resistant to Intestinal Tumors

Inderpreet Kaur Sur,[1,2] Outi Hallikas,[3] Anna Vähärautio,[1,3] Jian Yan,[1] Mikko Turunen,[3] Martin Enge,[1] Minna Taipale,[1,3] Auli Karhu,[4] Lauri A. Aaltonen,[4] Jussi Taipale[1,3]*

7 DECEMBER 2012 VOL 338 **SCIENCE**

TERT Promoter Mutations in Familial and Sporadic Melanoma

Susanne Horn,[1,2] Adina Figl,[1,2] P. Sivaramakrishna Rachakonda,[1] Christine Fischer,[3] Antje Sucker,[2] Andreas Gast,[1,2] Stephanie Kadel,[1,2] Iris Moll,[2] Eduardo Nagore,[4] Kari Hemminki,[1,5] Dirk Schadendorf,[2]*† Rajiv Kumar[1]*†

SCIENCE VOL 339 22 FEBRUARY 2013

Highly Recurrent *TERT* Promoter Mutations in Human Melanoma

Franklin W. Huang,[1,2,3]* Eran Hodis,[1,3,4]* Mary Jue Xu,[1,3,4] Gregory V. Kryukov,[1] Lynda Chin,[5,6] Levi A. Garraway[1,2,3]†

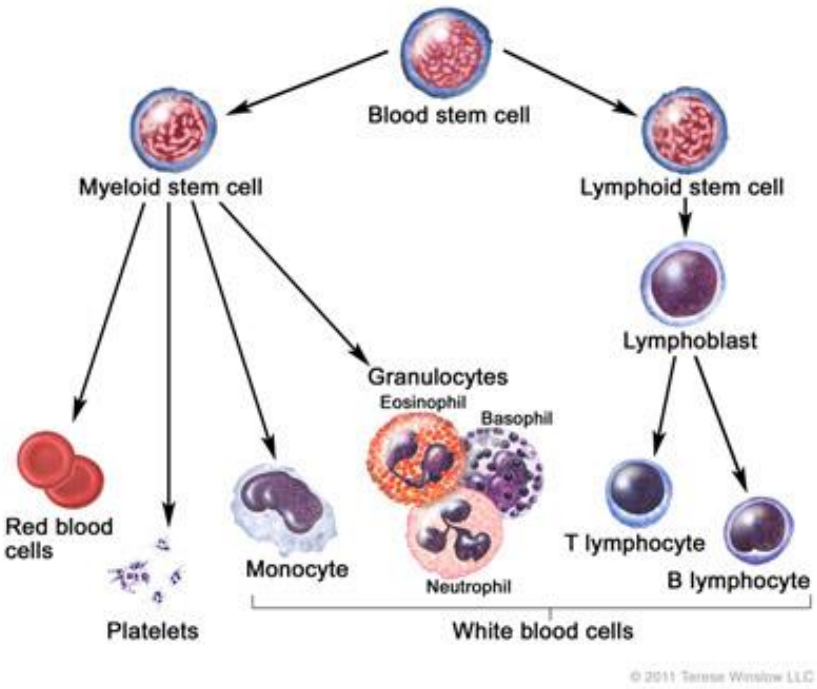
SCIENCE VOL 339 22 FEBRUARY 2013

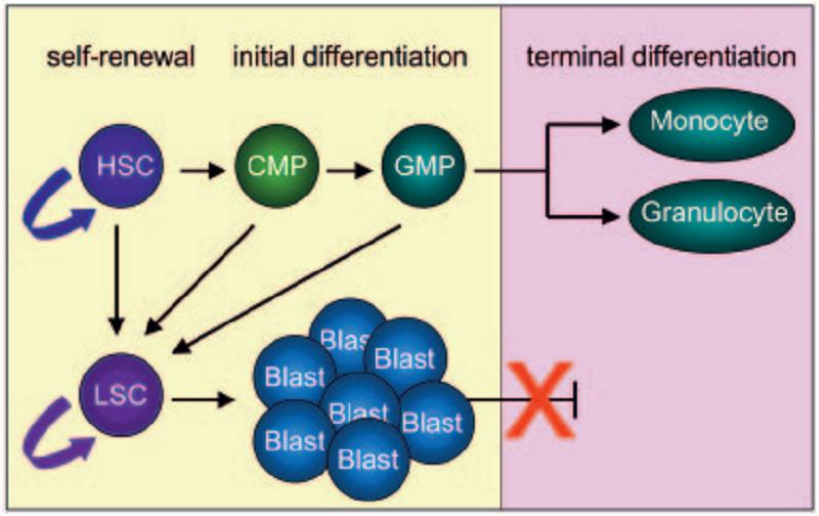
Pediatric Acute Myeloid Leukemia (AML)

Failure of a normal developmental process (block in HSC differentiation)

+

massive proliferation of immature white blood cells

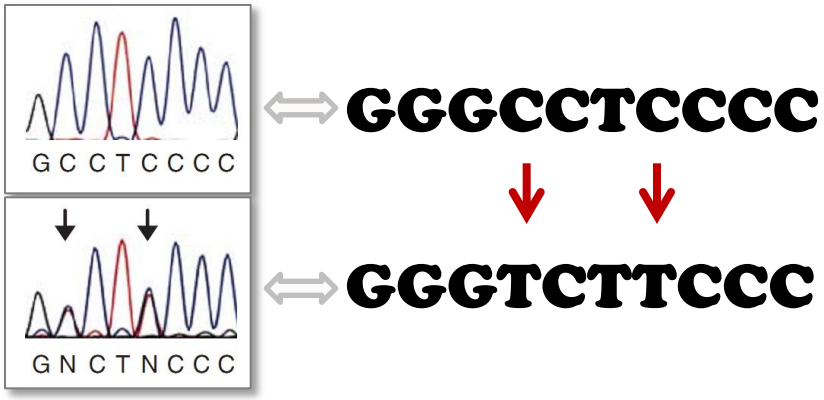




Blood, 2005, (106):1519-1524

An NF-κB binding-site variant in the SPI1 URE reduces PU.1 expression & is correlated with AML

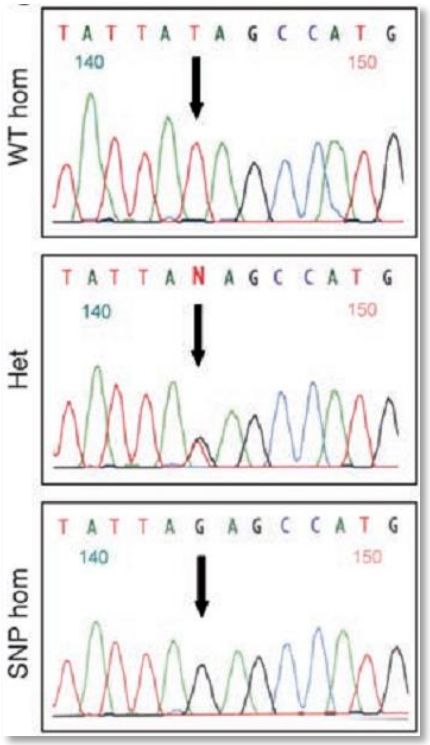
Bonadies et al, Oncogene, 2009, 29(7):1062-72.



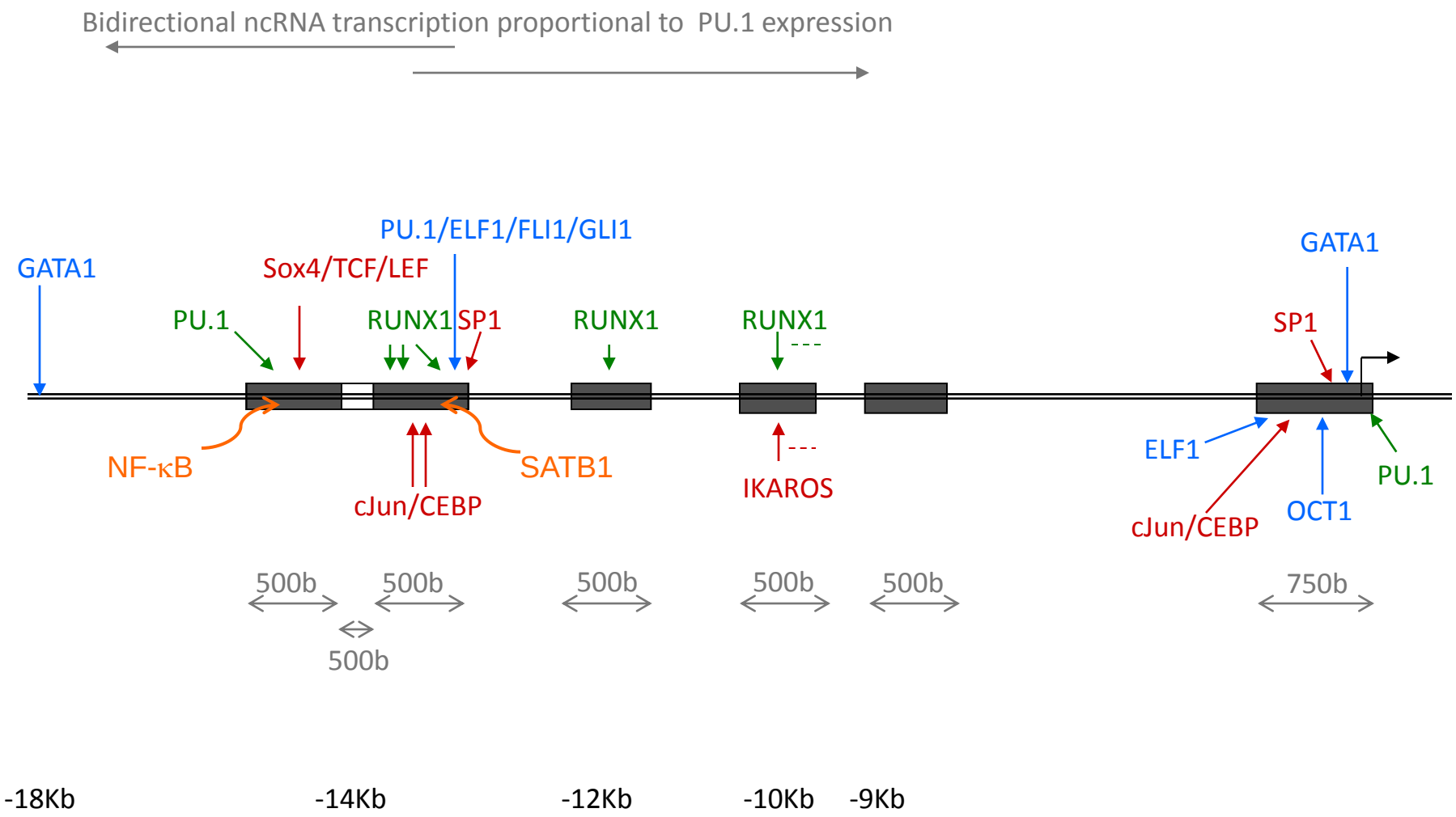
A distal single nucleotide polymorphism alters long-range regulation of the PU.1 gene in acute myeloid leukemia

Steidl et al, J Clin Invest. 2007, 117(9):2611-20.

SATB1 binding site



Regulation of SPi1 expression – part 2 (mouse coordinates)



Chou et al, Blood, 2009, 114: 983-994

Hoogenkamp et al, Molecular & Cellular Biology, 2007, 27(21):7425-7438

Zarnegar & Rothenberg, 2010, Mol. & cell Biol. 4922-4939

A historical perspective on Transcription Factor Binding Site (TFBS) identification

(1) Computational predictions:

"FUTILITY THEOREM — that essentially all predicted TFBSs will have no functional role."
Sandelin & Wasserman, Nature Reviews Genetics 2004; 5:276-287.

Solution: Limit computational motif mapping to experimentally-identified *cis*-regulatory regions.

(2) Data-driven approaches:

(A) Combinatorial histone marks identify active promoters and enhancers
Ernst et al , Nature 2011; 473(7345):43–49.
Predicted functional promoters & enhancers in **9** cell types cover ~9.8% of the genome.
Poor spatial resolution (~500-1000bp) results in high false positive rates.

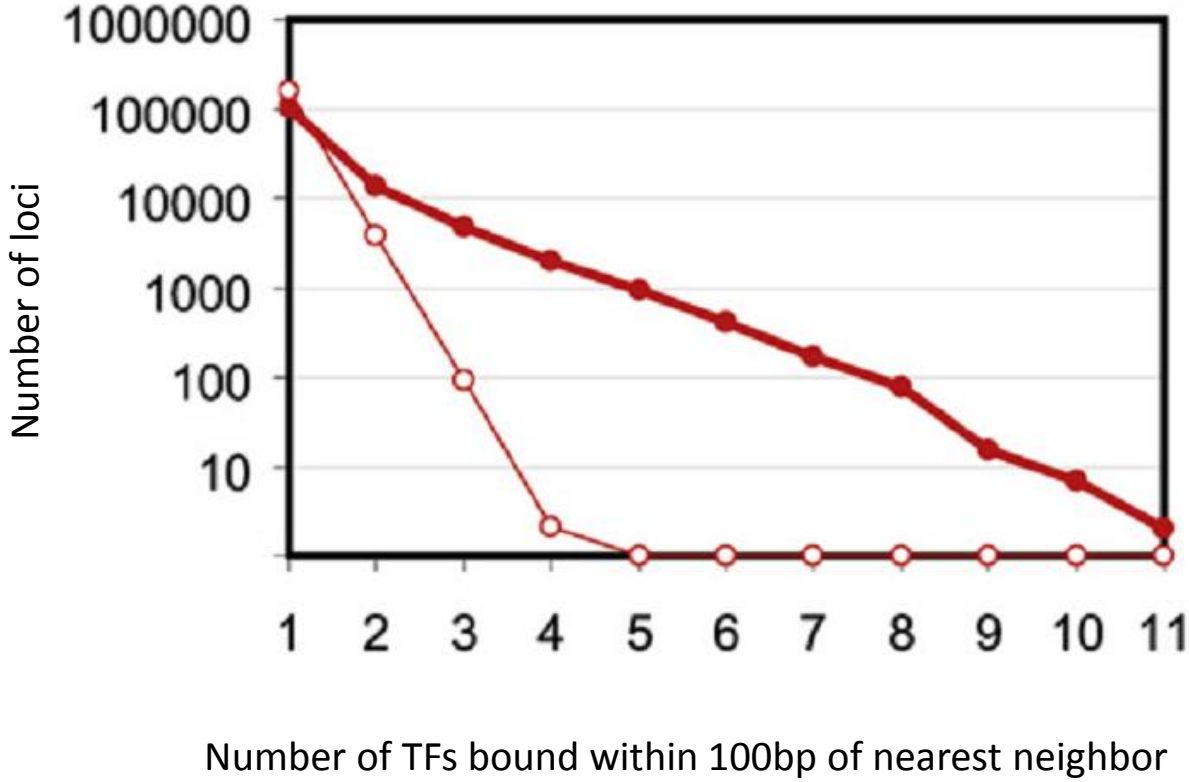
(B) DNAse1 hypersensitivity clusters mark *cis*-regulatory regions
Thurman et al (Stamatoyannopoulous lab, ENCODE project) Nature 2012; 489(7414):75-82.
150bp resolution. 2.9M peaks in **125** cell types → 436,970,762bp or ~14.6% of the genome.
As little as ~ 10% of the marked sequence may be functional TFBS.

(C) DNAse1 footprints directly delineate TFBS
Neph et al ((Stam lab, ENCODE project) , Nature 2012; 489(7414):83-90.
Costly but precise. 8.4M TFBS in **41** cell types → 164,010,758 bp or ~ 5.5% of the genome.
Will miss condition-specific TFBS in cells not assayed.

Our approach: TF ChIP-seq peak clusters with maximal DNase1 HS agreement

ChIP-seq of 13 sequence-specific TFs

Nanog, Oct4, STAT3, Smad1, Sox2, Zfx, c-Myc, n-Myc, Klf4, Esrrb, Tcfcp2l1, E2f1, and CTCF



Chen et al, Cell, 2008;133(6):1106-17

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 9, 2012

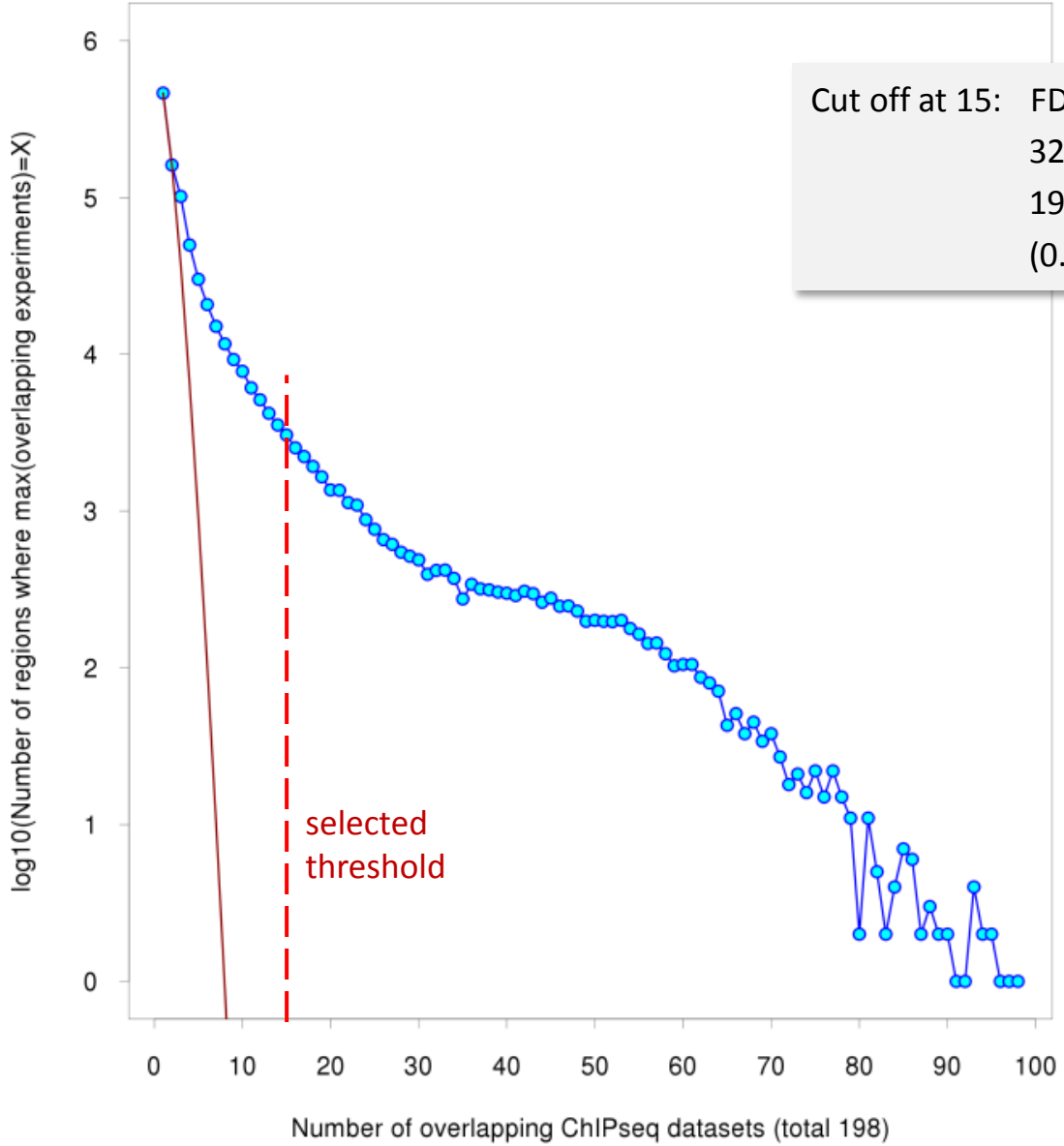
Pp. 1–9
DOI: 10.1089/cmb.2012.0100

Research Article

Integration of 198 ChIP-seq Datasets Reveals Human *cis*-Regulatory Regions

HAMID BOLOURI[1] and WALTER L. RUZZO[2,3,4]

Distribution of overlapping peaks for all 198 ChIPseq datasets combined



Cut off at 15: FDR << 1%
32,467 regions
19,992,599 bps
(0.67% genome)

Questions:

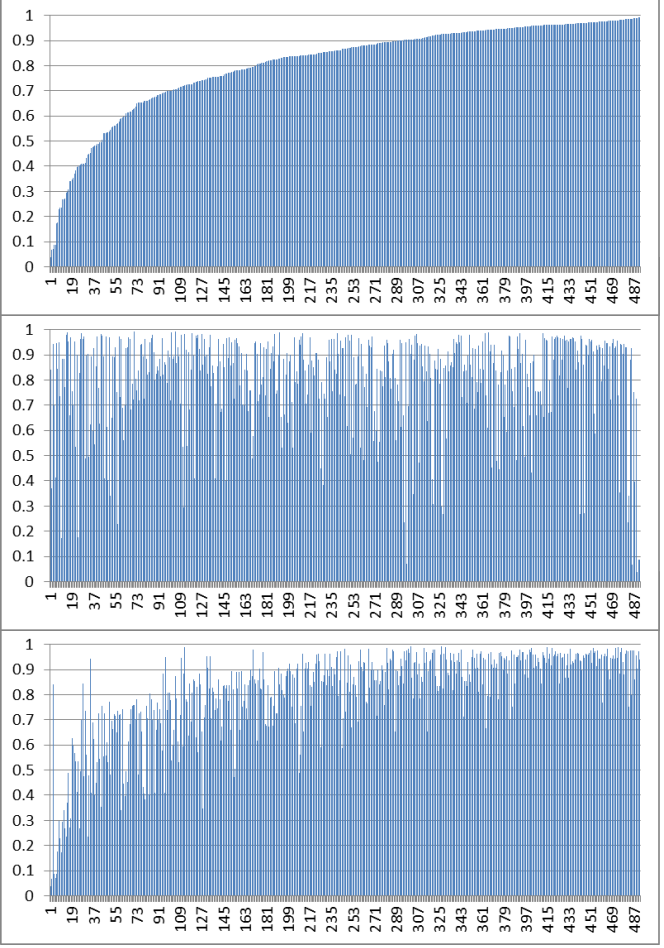
- Why remove unclustered peaks?
- Why a threshold of 15?
- Why not just use DNase1?

After Michael Waterman, Hearing Distant Echos: Using Extremal Statistics to Probe Evolutionary Biology, pp.90-113 in Lander & Waterman (Eds), Calculating the Secrets of Life - Applications of the Mathematical Sciences in Molecular Biology, National Academy Press, 1995.

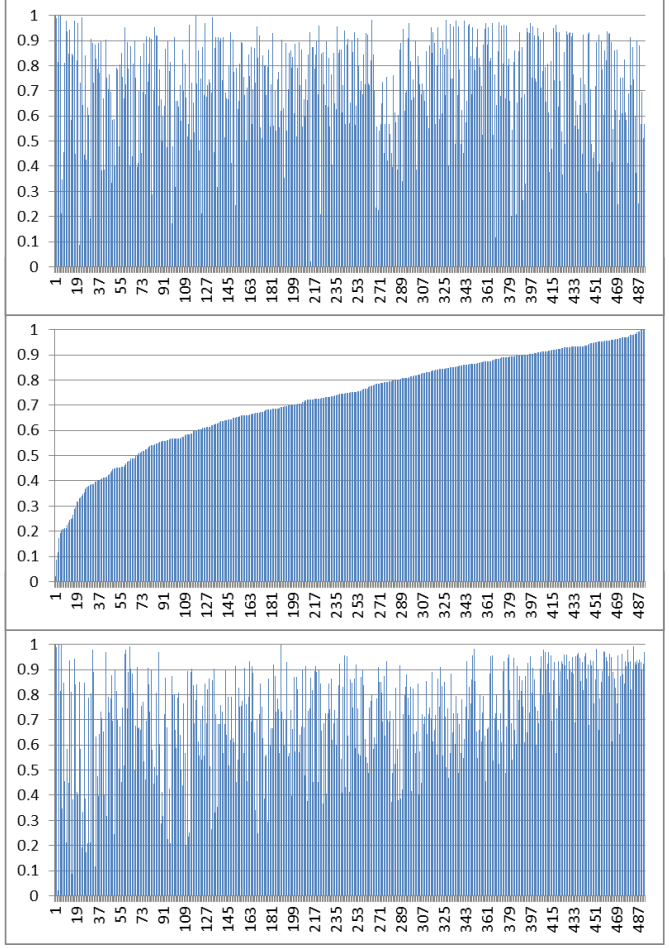
Comparing peaks called by peakSeq & SPP for 492 ENCODE ChIP-seq datasets

(optimized calls by Anshul Kundaje using FDR & the Irreproducible Discovery Rate method)

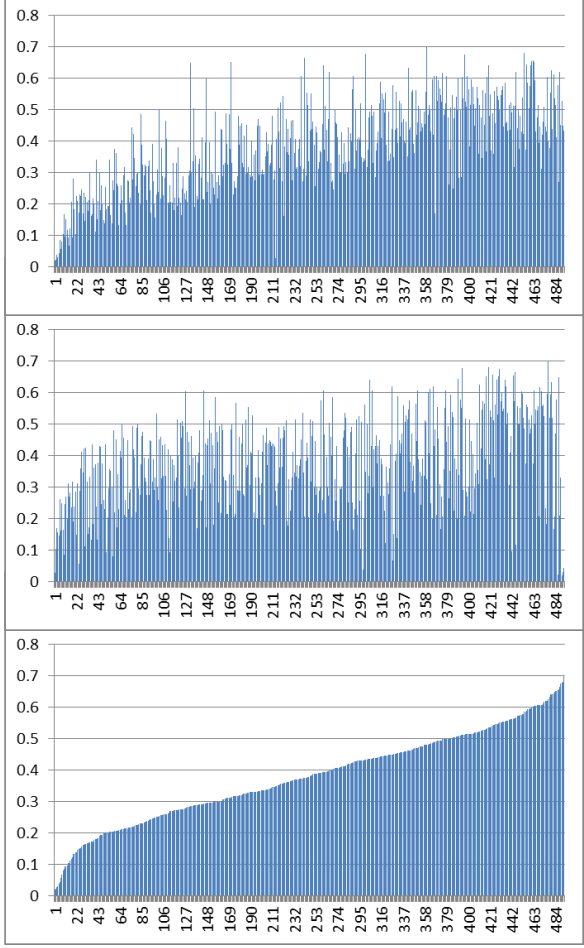
Fraction of PeakSeq peaks overlapping SPP peaks



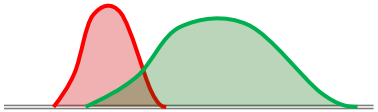
Fraction of SPP peaks overlapping peakSeq peaks



Overlapping base pairs as a fraction of total in peaks

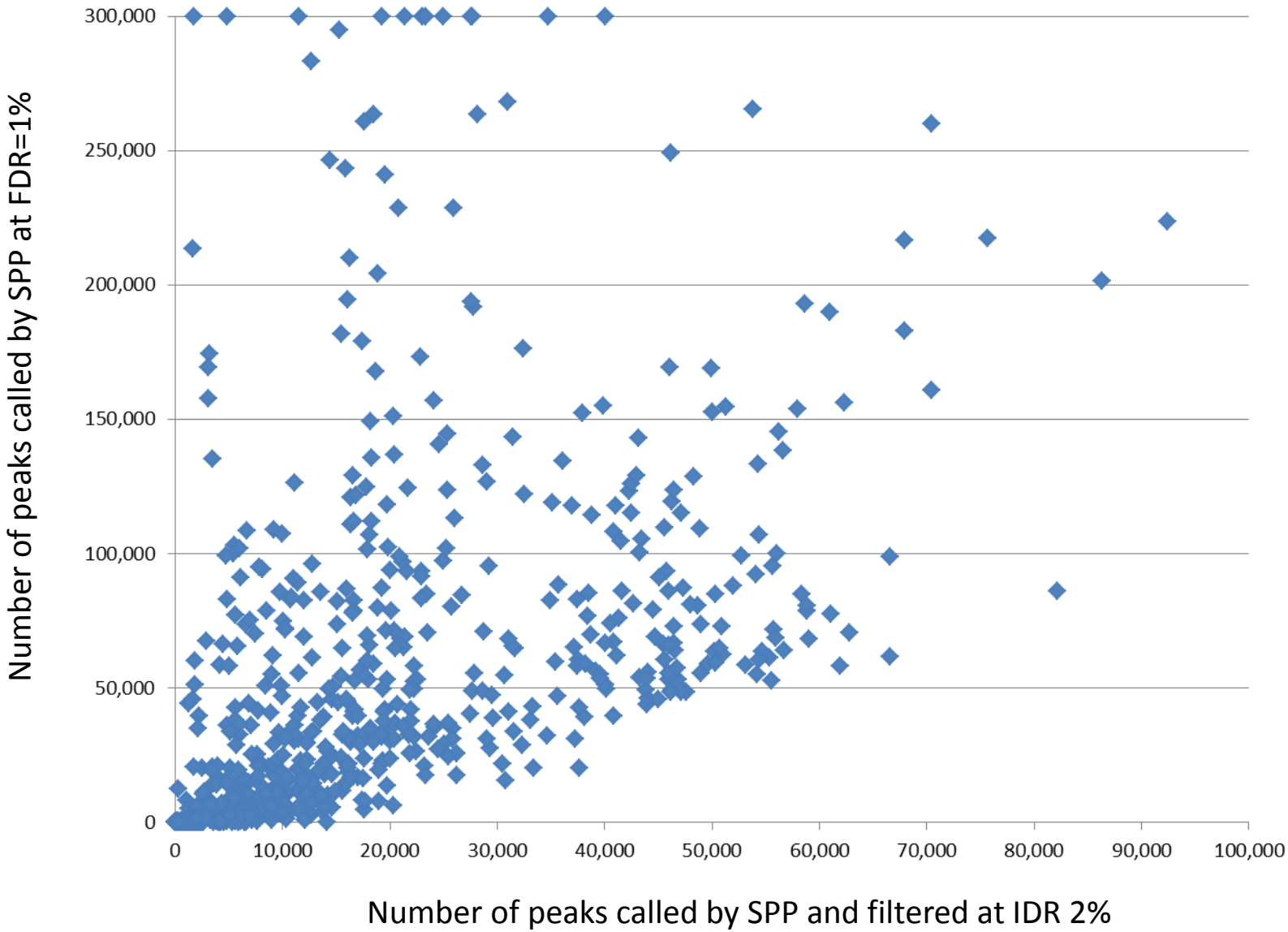


ordered samples

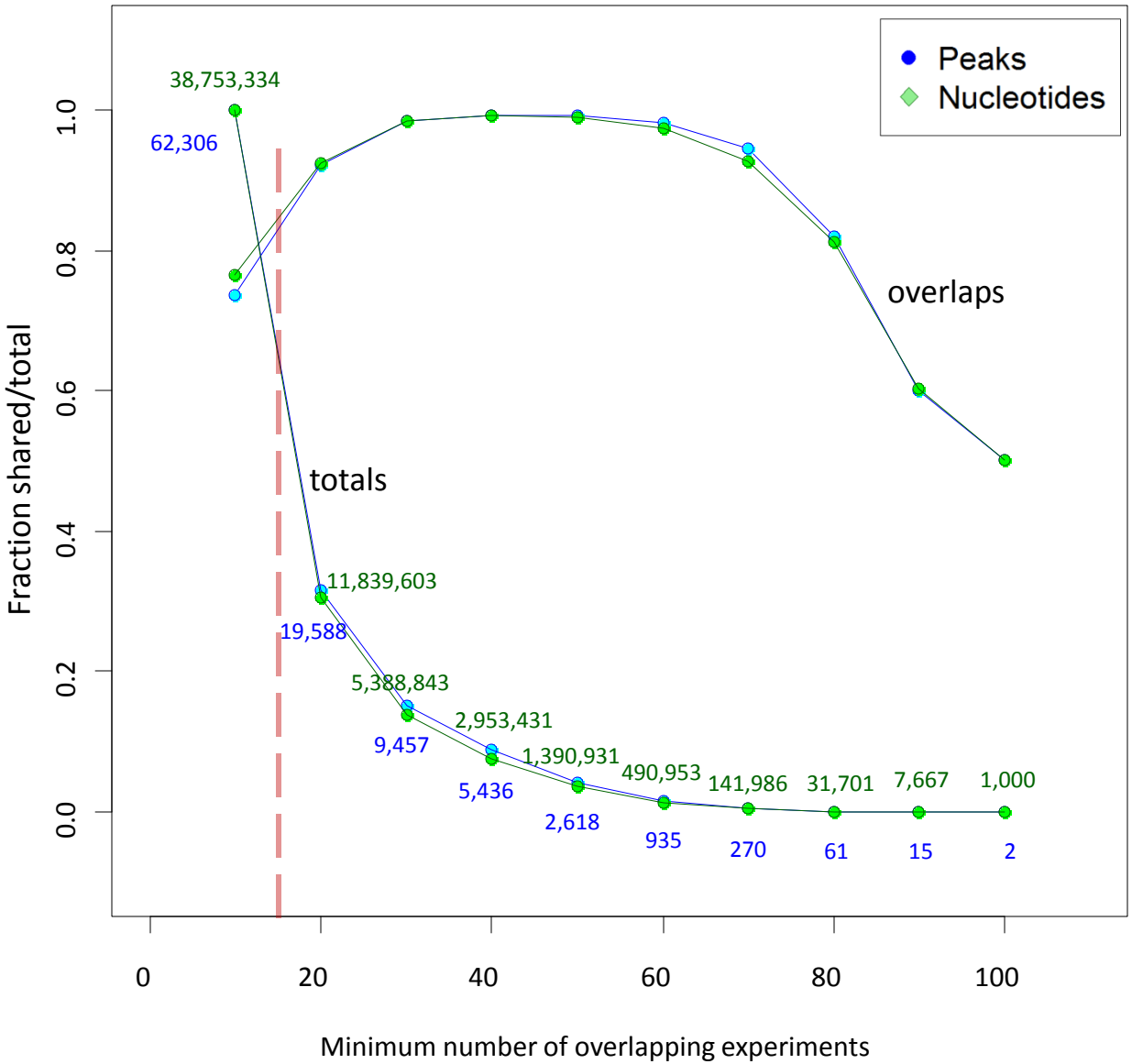


August 9, 2012 analysis of ENCODE ChIP-seq datasets by Anshul Kundaje

ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byFreeze/june2012/peaks/spp/README.txt



Effect of selection threshold on overlap with DNase1-marked binding regions.



High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells

Alan P. Boyle,[1] Lingyun Song,[1,2] Bum-Kyu Lee,[3] Darin London,[1] Damian Keefe,[4] Ewan Birney,[4] Vishwanath R. Iyer,[3] Gregory E. Crawford,[1,2,5] and Terrence S. Furey[1,5]

Genome Research 21:456–464 © 2011

(HeLaS3, HUVEC, K562, NHEK, H1hesc + 7 HapMap B-lymphoblastoid cell lines)

958,250 / 1,067,220 = 89.8% of DNase1 selected regions overlap histone marked regions

(total footprint of DNase1-selected-regions = 22,388,756 bps , ~ 0.75% of the genome)

442,295 / 1,067,404	= 41.4%	of DNase1Regions	overlap	CRR198
27,784 / 32,467	= 85.6%	of CRR198	overlap	DNase1Regions

84.2% of ChIPseq predicted CRMs are supported by both histone and DNase1-based predictions

ENCODE (Stam Lab, UW) DNASE1 Hyper Sensitive regions across **125 cell types**

2,890,742 regions

436,970,762 bp ~ **14.6%** of the genome

ENCODE (Stam Lab, UW) DNASE1 TF foot prints across **41 cell types**

6,447,639 regions

164,010,758 bp ~ **5.5%** of the genome

ENCODE (Stam Lab, UW) DNASE1 TF foot prints in mobilized **CD34+ cells**

164,049 HS regions at 1% FDR, of which

104,544 have signal p-value < 0.01

15,806,684 bp ~ **0.53%** of the genome

The need for filtering whole genome sequence variants

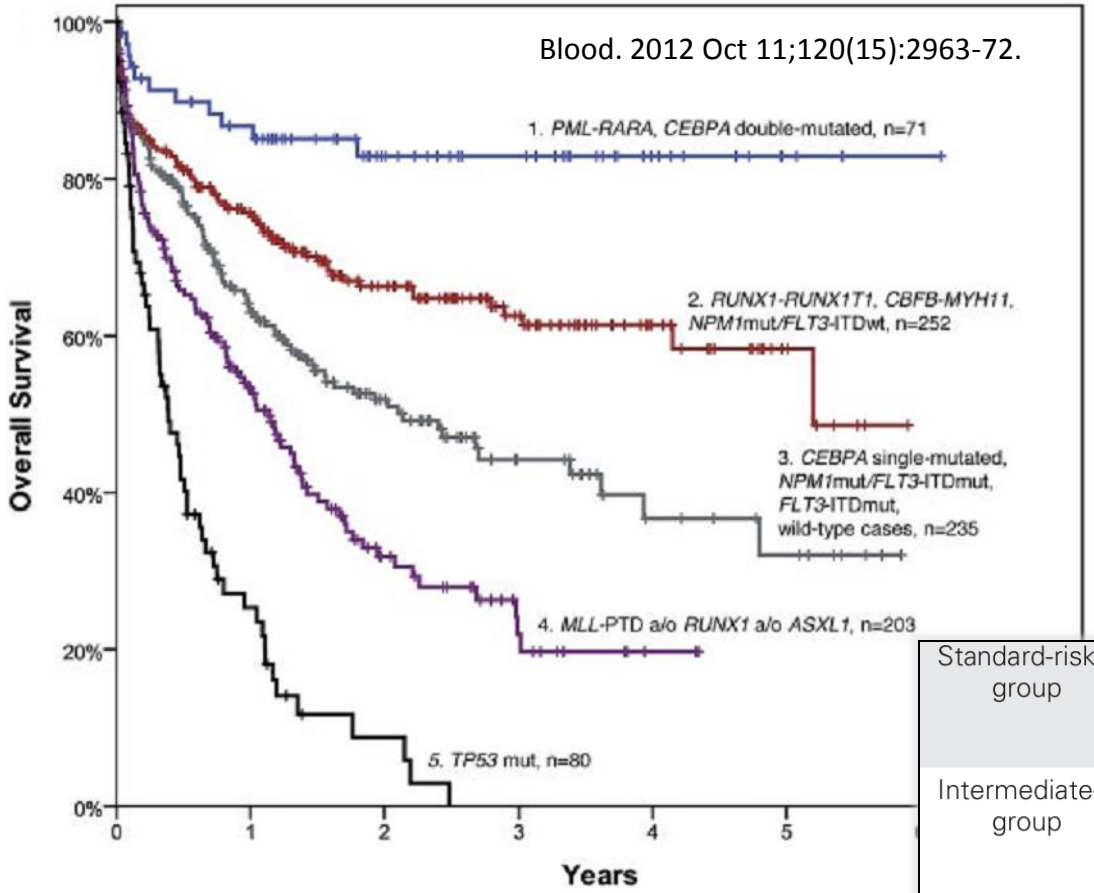
28,091,309	somatic variants in ***122*** samples	~ 230K	/ sample
13,752,804	are somatic (not LOH)	~ 112K	/ sample
1,438,103	have p-value < 0.05	~ 12K	/ sample
340,692	have p-value < 0.01	~ 2800	/ sample
83,308	have P-value < 0.01 & are SQHIGH	~ 683	/ sample
71,410	are SNVs (Single Nucleotide Variants)	~ 585	/ sample

read-count & allelic-ratio filters

ENSEMBL Variant Effect Predictor
(includes SIFT & PolyPhen2)

Number of SNVs in introns or 7.5Kbp upstream	~ 350	/sample
In DNAse1 footprints (41 cell types) & not in 54 CGI healthy genomes	~ 25	/sample
In recurrently impacted genes	~ 3.5	/sample

Genomic abnormality groups indicate outcome



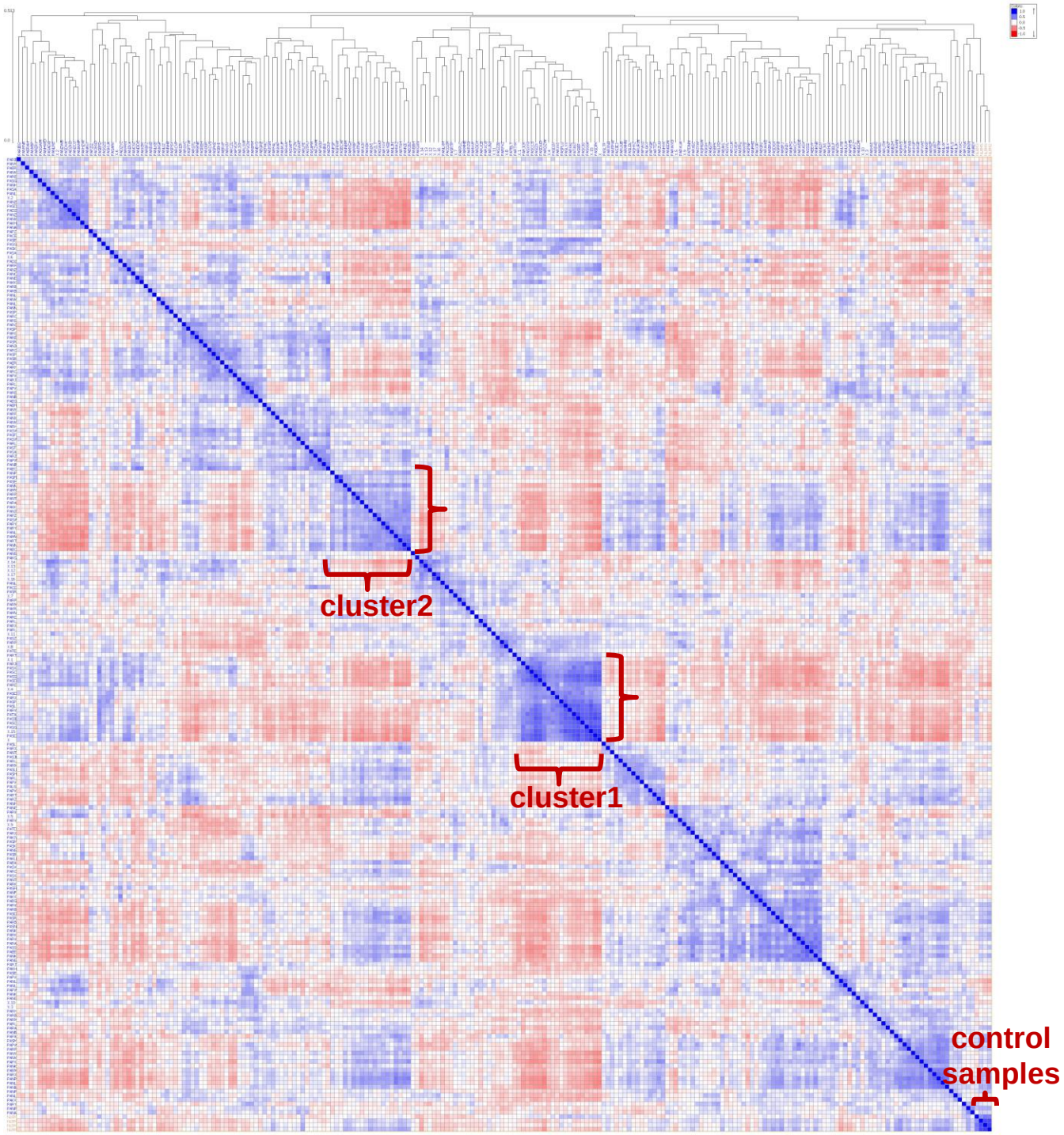
J Clin Oncol. 2010 Jun 1;28(16):2682-9.

Standard-risk group	Favorable cytogenetics (8;21)(q22;q22) or AML1-ETO t(15;17)(q22;q21) or PML-RARα Inversion (16)(p13;q22) or CBFβ/MYH11
Intermediate-risk group	Intermediate cytogenetics Aberration of 7q Trisomy 8 without favorable genetics Aberrations of chromosome 5 without favorable genetics *MLL* rearrangement t(9;11) without additional aberrations *MLL* rearrangement t(11;19) Other cytogenetic aberrations† Normal karyotype*
High-risk group	Unfavorable cytogenetics *MLL* rearrangement in t(9;11) with additional aberrations *MLL* rearrangements other than t(9;11) and t(11;19) Monosomy 7 Aberrations involving 12p without favorable genetics Complex karyotypes‡
Rare unfavorable cytogenetics	t(9;22)(q34;q11), t(8;16)(p11;p13); t(6;9)(p23;q34), t(7;11)(p15;p15), t(7;12)(q36;p13)§

Gene expression microarrays:

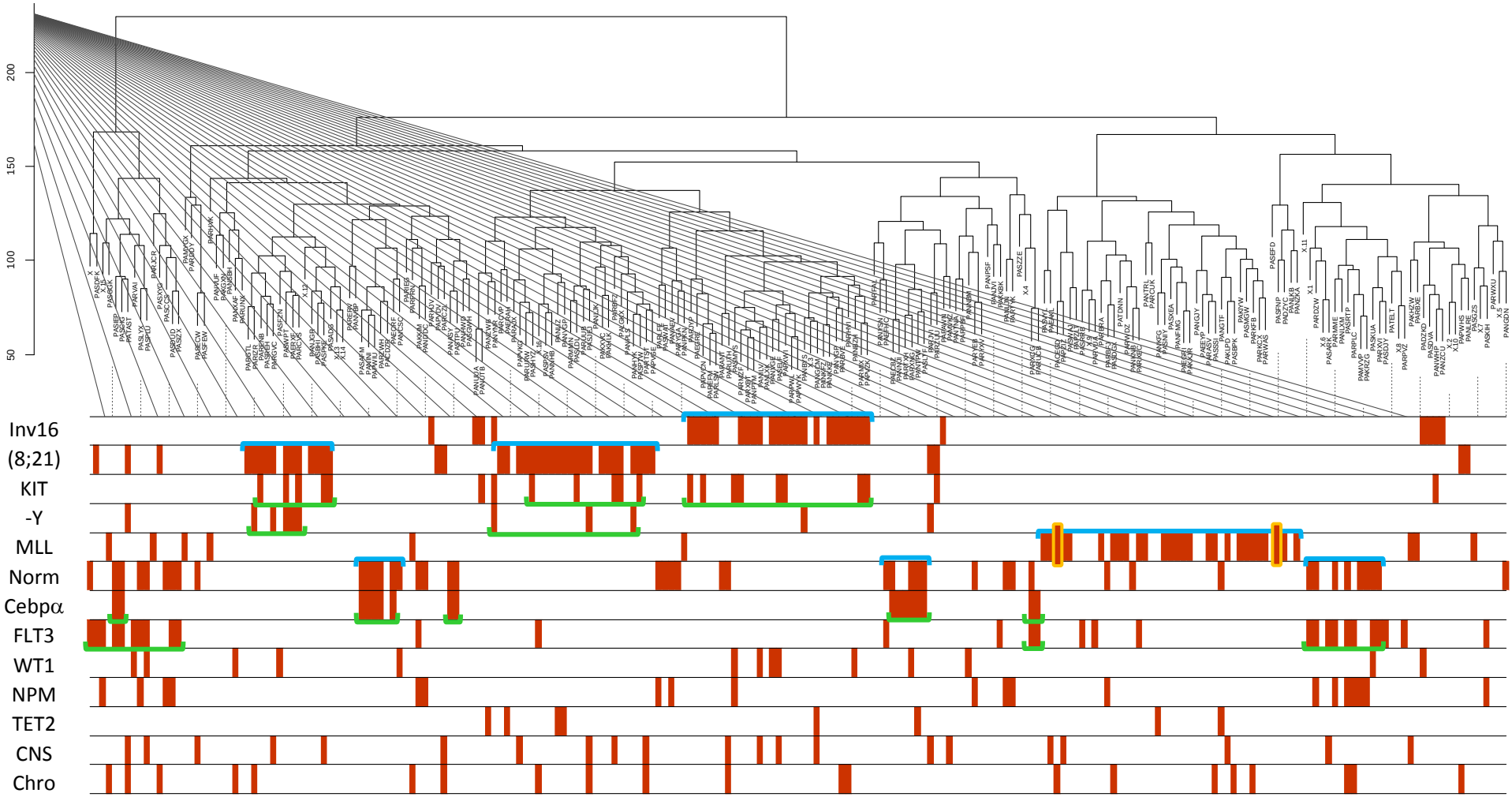
- 225 AML samples
- 4 control samples

Unsupervised clustering (Pearson correlation) confirms distinct patient groups



Expression data hierarchically clustered by ‘complete linkage’

(finds compact, spherical clusters)



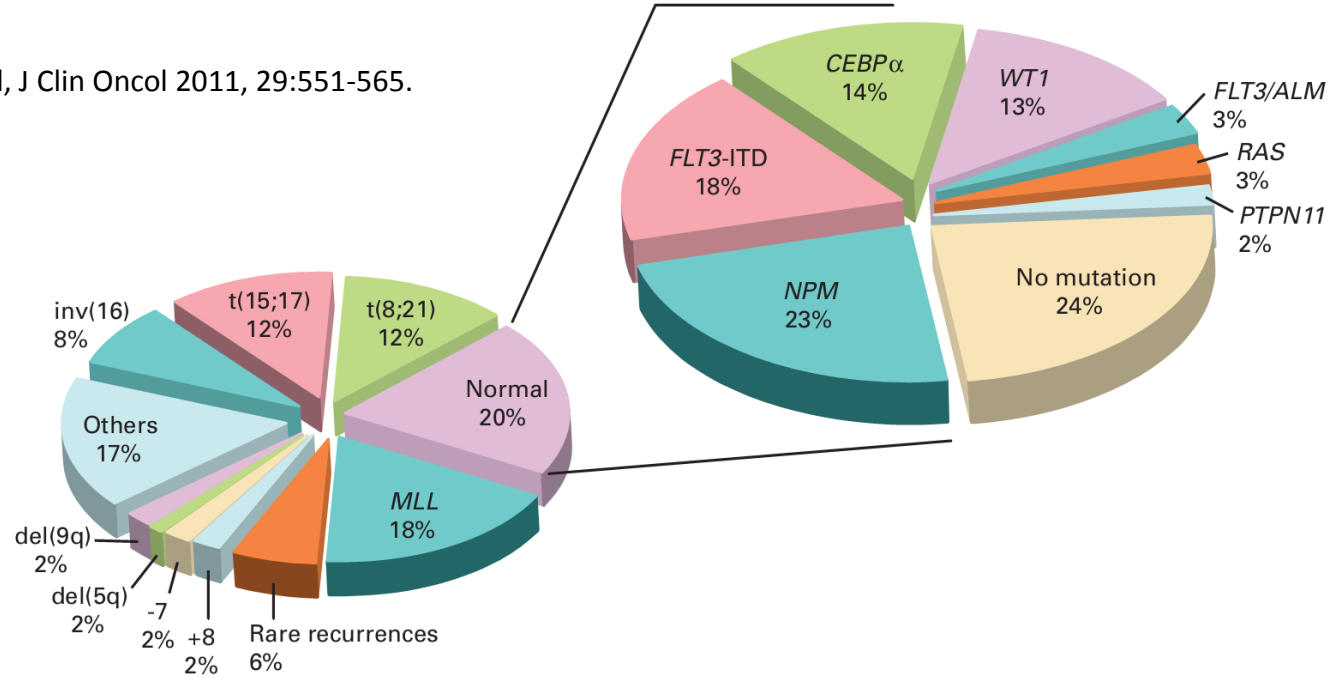
= Primary Cytogenetic Abnormality associated with expression cluster

= Secondary Abnormality co-occuring with expression cluster

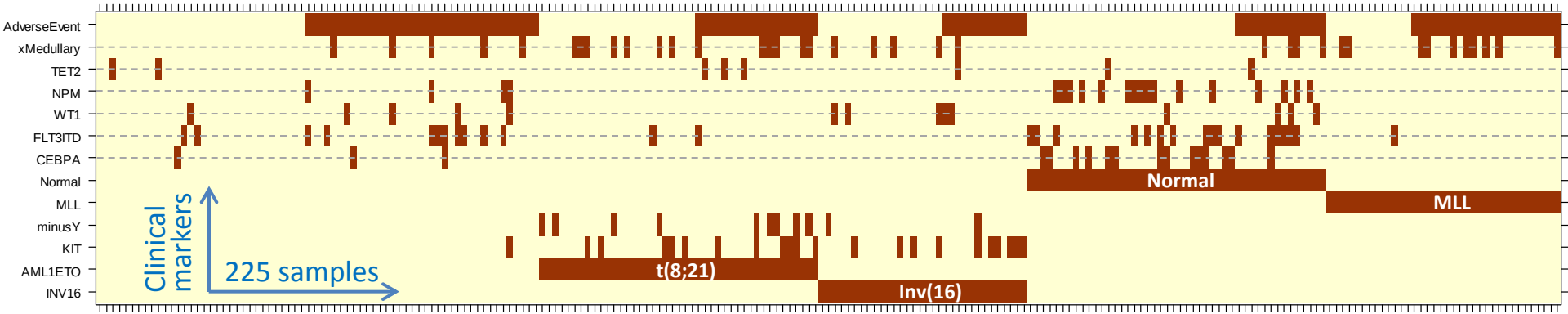
= MLL cases verified *after* initial clustering

> 95% of all children with AML have at least one known genomic abnormality

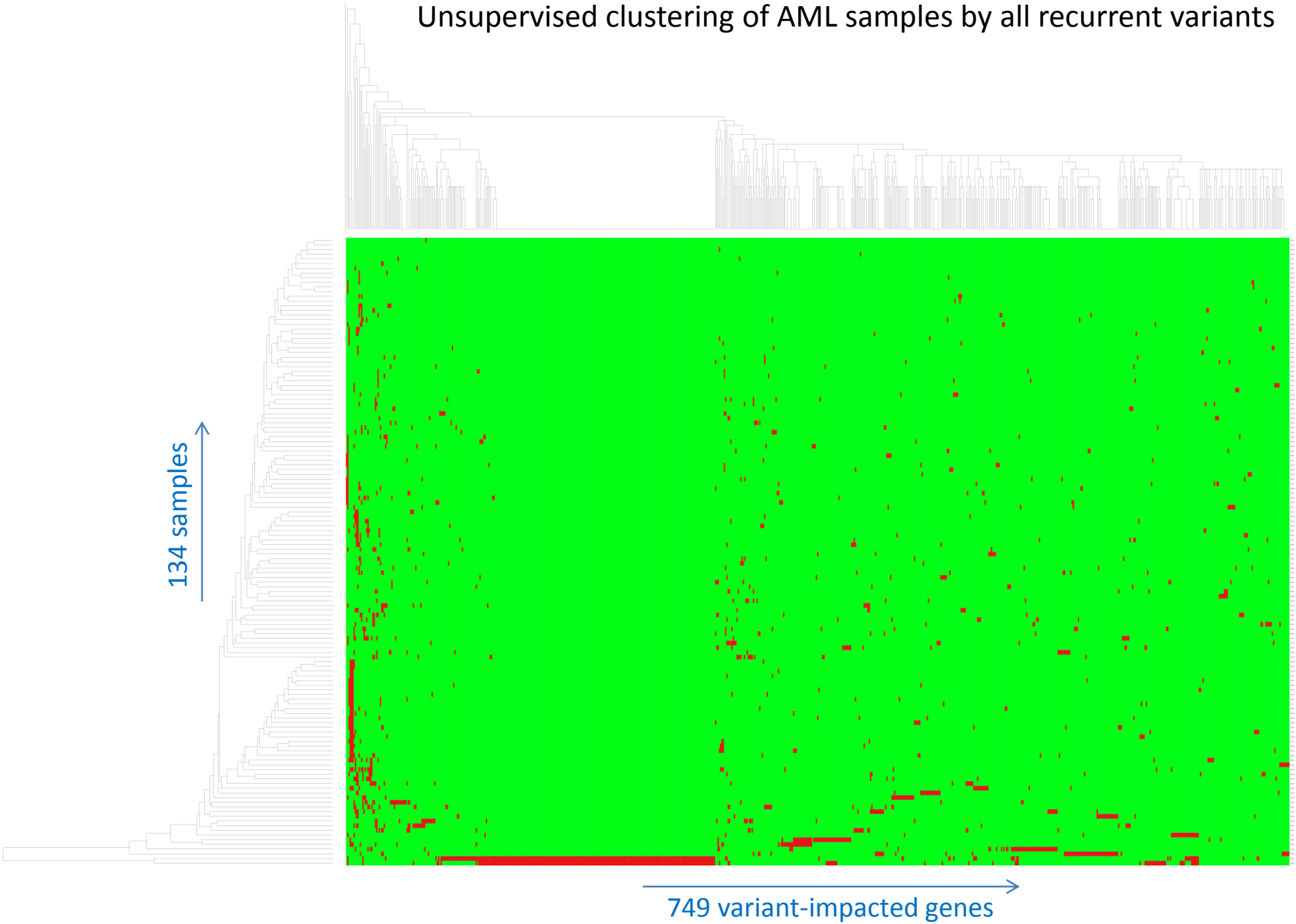
Pui et al, J Clin Oncol 2011, 29:551-565.



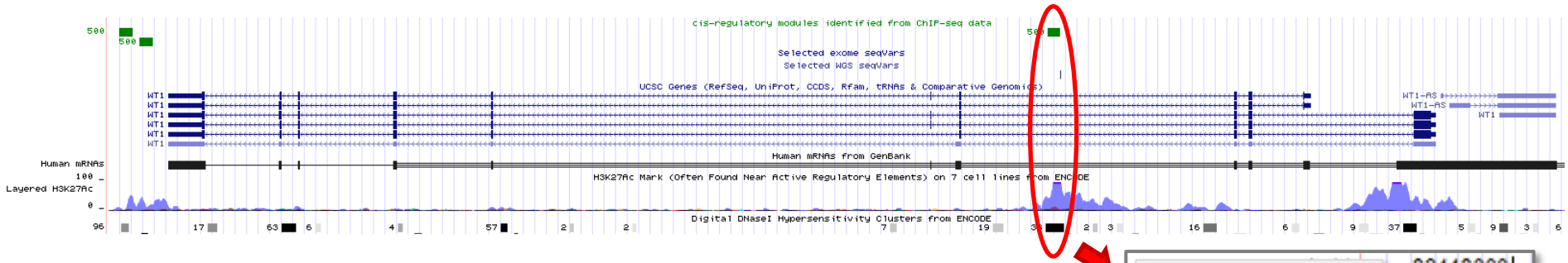
Pediatric AMLs cluster into cytogenetic groups with genetic sub-groups

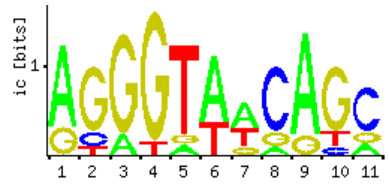


Unsupervised clustering of AML samples by all recurrent variants



Example potential regulatory SNV in intron3 of the Wilm's Tumor1 gene in AML





CGAGGTTGGCA

CTCTGCCTGT: NRF-2 [-]

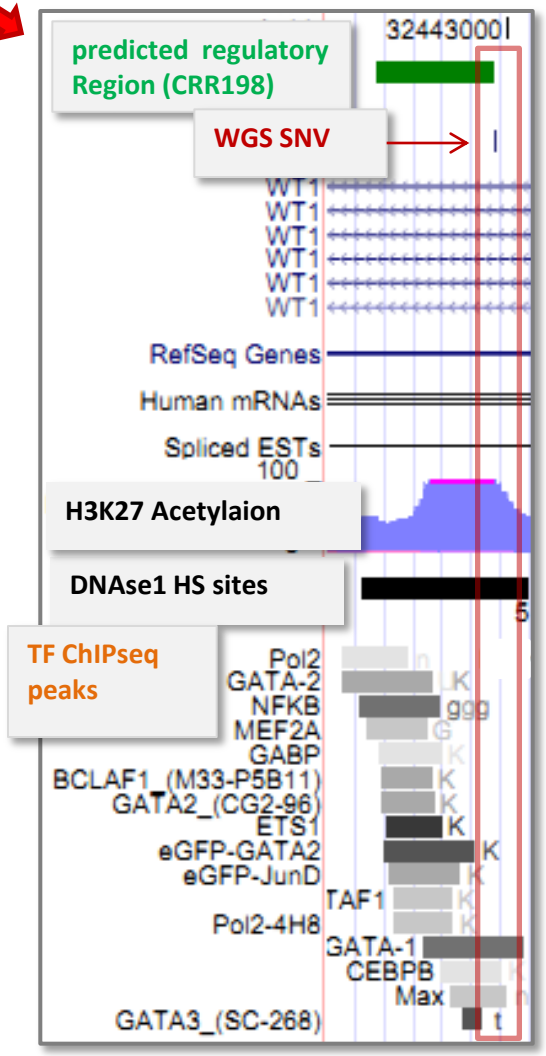
CTAACT: GATA-3 [-]

GCTCCAACCGT: Spz1 [-]

TTTCAGCTCCAACCGTCTAACTCTGCCTGT

10 T 20 30

SNV impacts an evolutionarily conserved Spz1 motif



bHLH-zip transcription factor Spz1 mediates mitogen-activated protein kinase cell proliferation, transformation, and tumorigenesis.

Hsu SH, et al. Cancer Res, 2005 May 15. PMID 15899793.

SNV selection criteria:

-- 0.4 < Allelic Ratio < 0.6

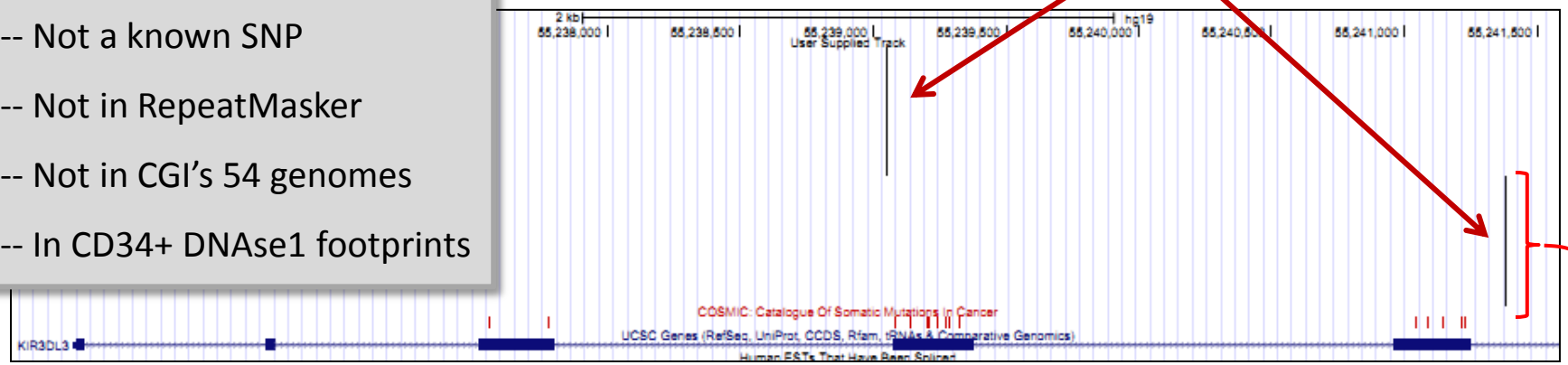
-- Not a known SNP

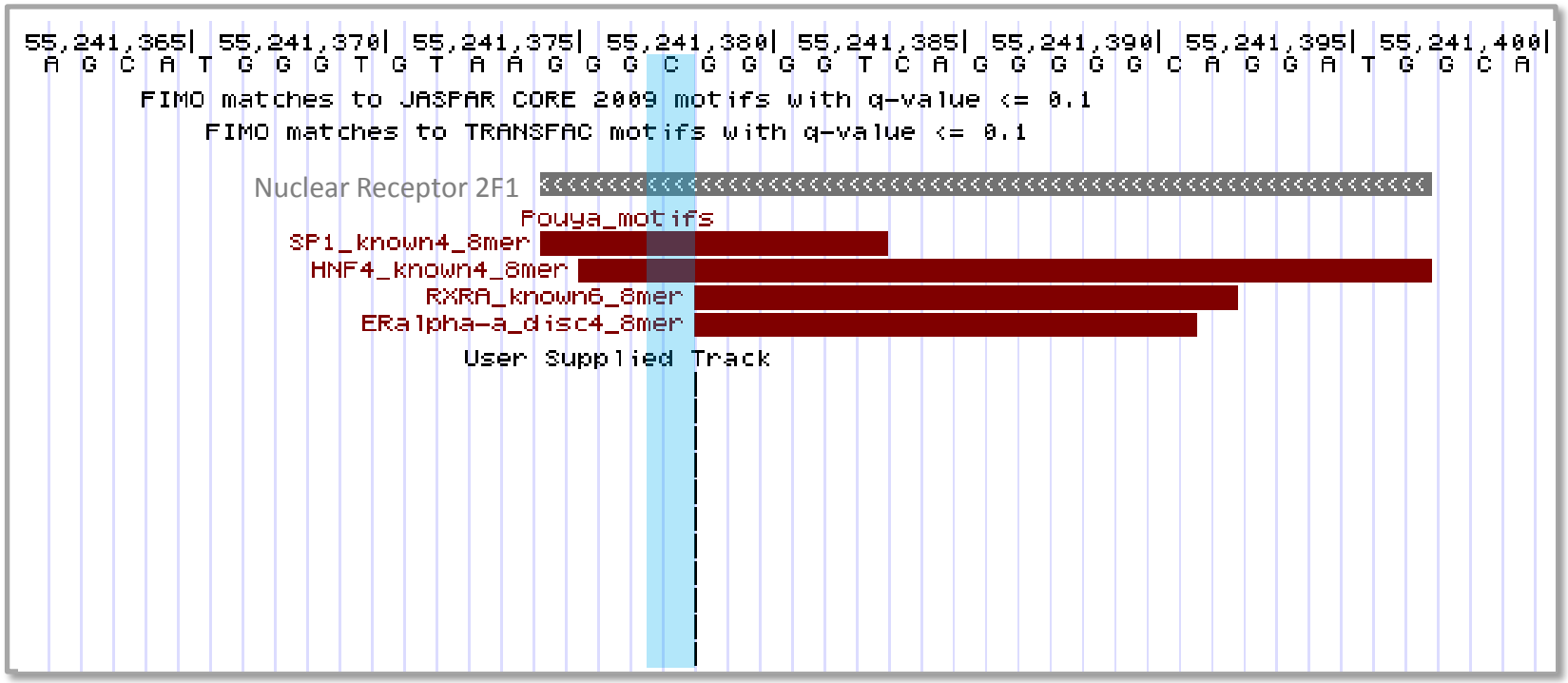
-- Not in RepeatMasker

-- Not in CGI's 54 genomes

-- In CD34+ DNAse1 footprints

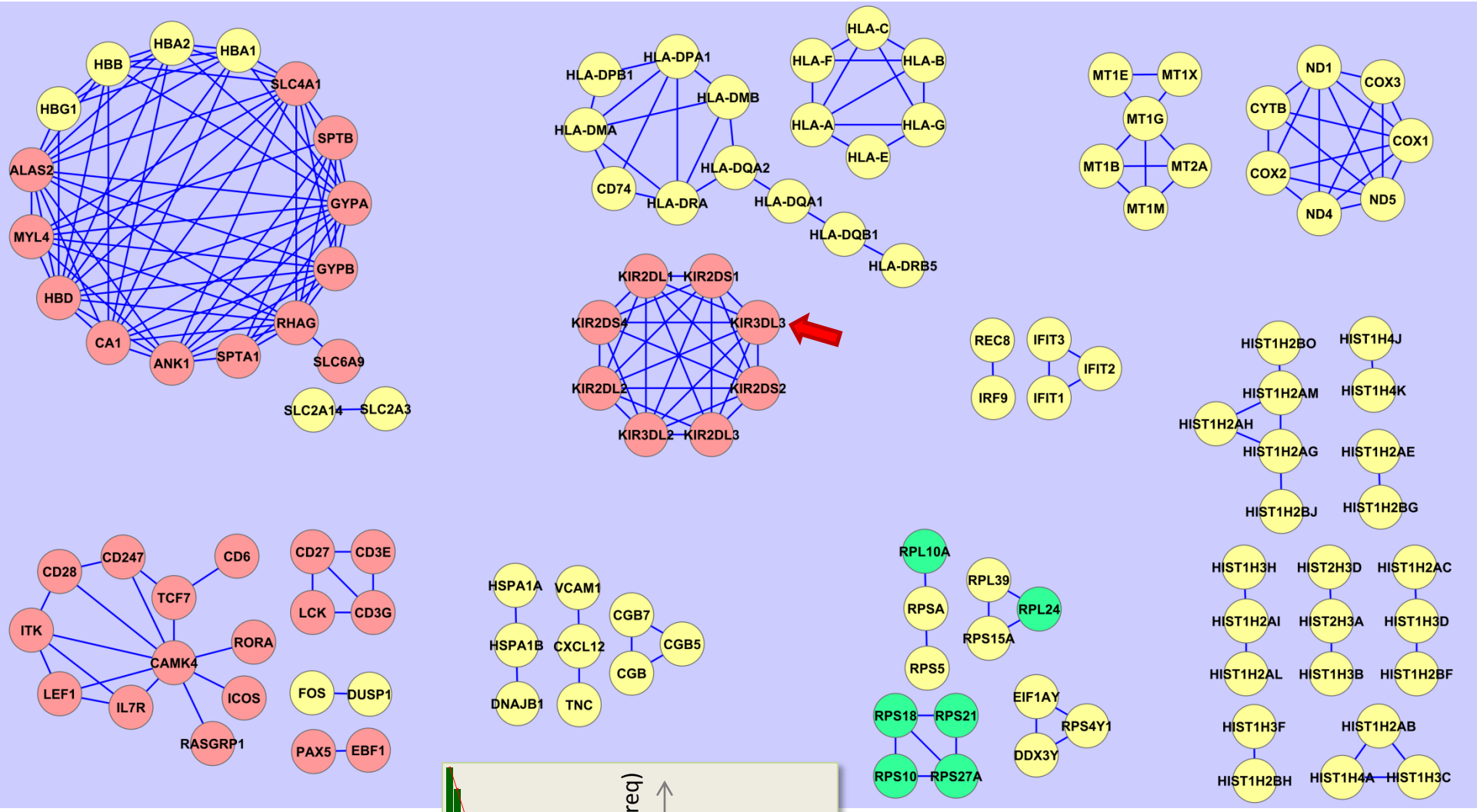
11 of 138 AML samples share ***germline*** SNVs at two KIR3DL3 sites





Interactions ***inferred*** from 225+4 expression arrays

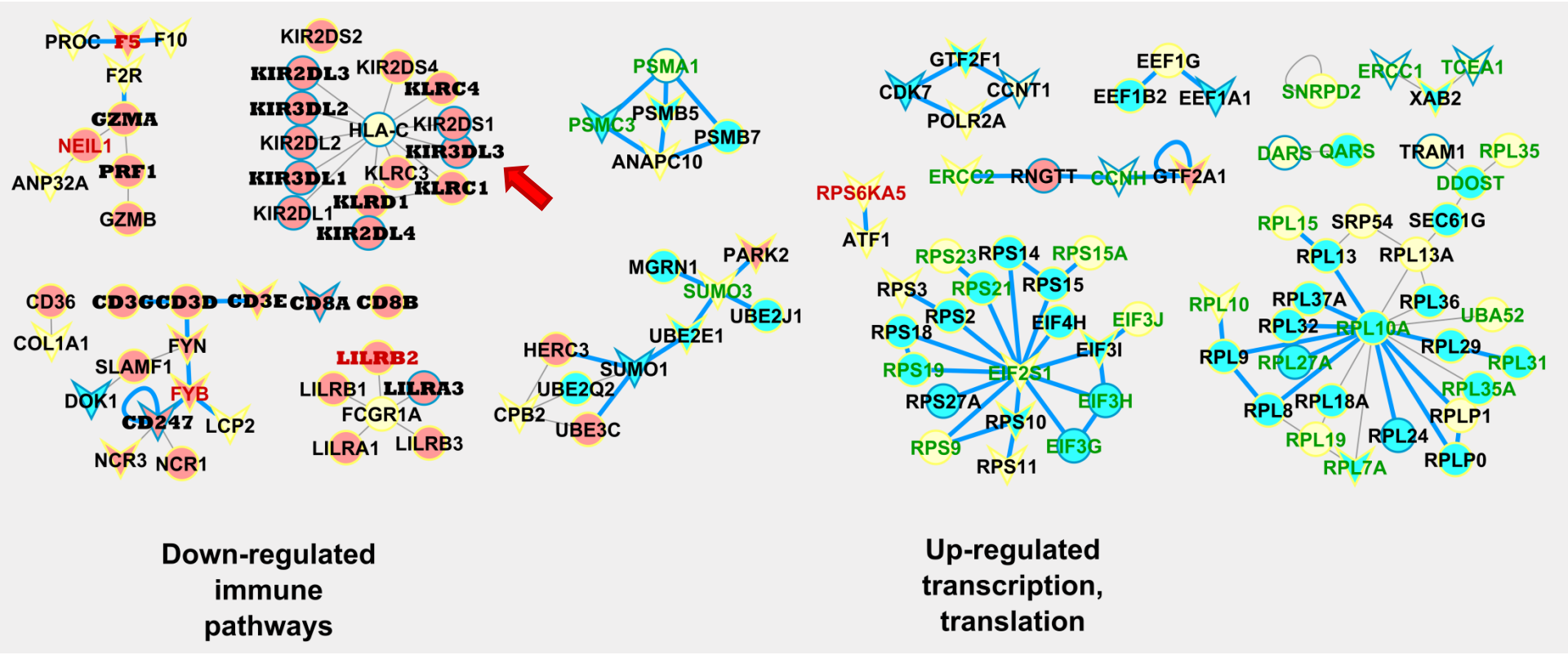
(Combining results from 4 algorithms: ARACNE, CLR, MRnet, & MRnetB)



Bioconductor package “Minet” & Cytoscape

Differential-expression enriched pathway interactions in 225+4 samples

(Using all pathways in Biocarta, KEGG, NCI PID, & Reactome)



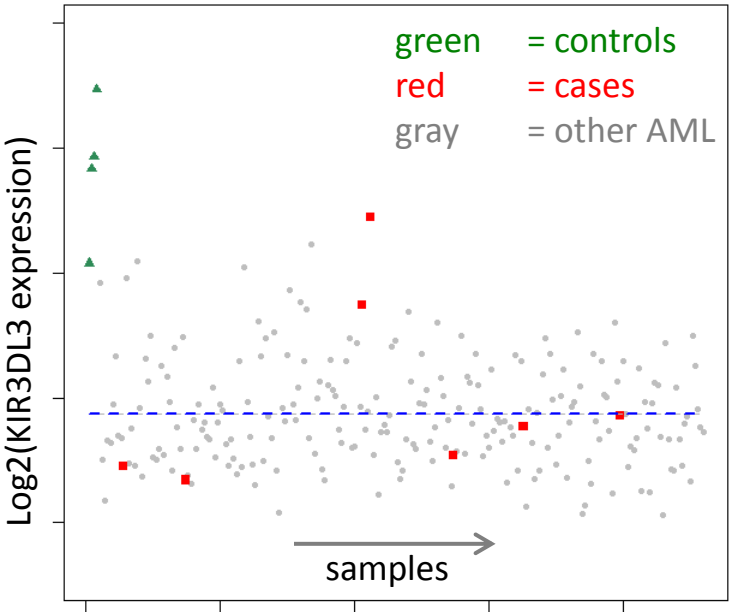
Down-regulated immune pathways

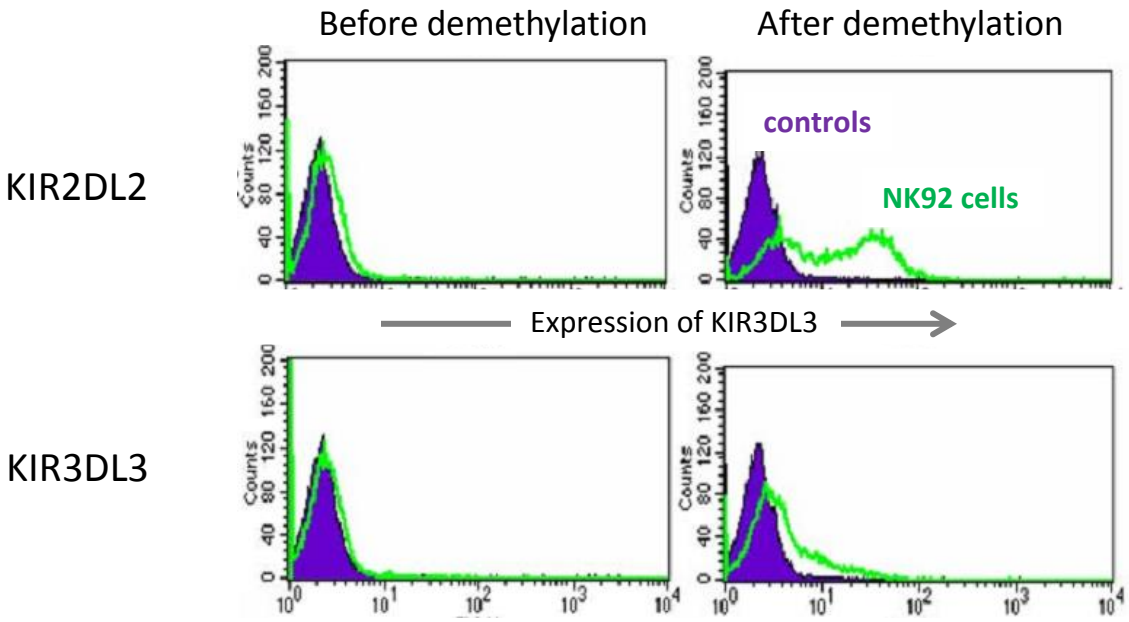
Up-regulated transcription, translation

Key:

- Cancer-associated gene (MSKCC list)
- Up-regulated in 225 AML samples
- Down-regulated in 225 AML samples
- **Bold** Gene differentially expressed in > half of samples
- Interaction also enriched in validation dataset J54
- Green label Up-regulated in validation dataset E23
- Red label Down-regulated in validation dataset E23

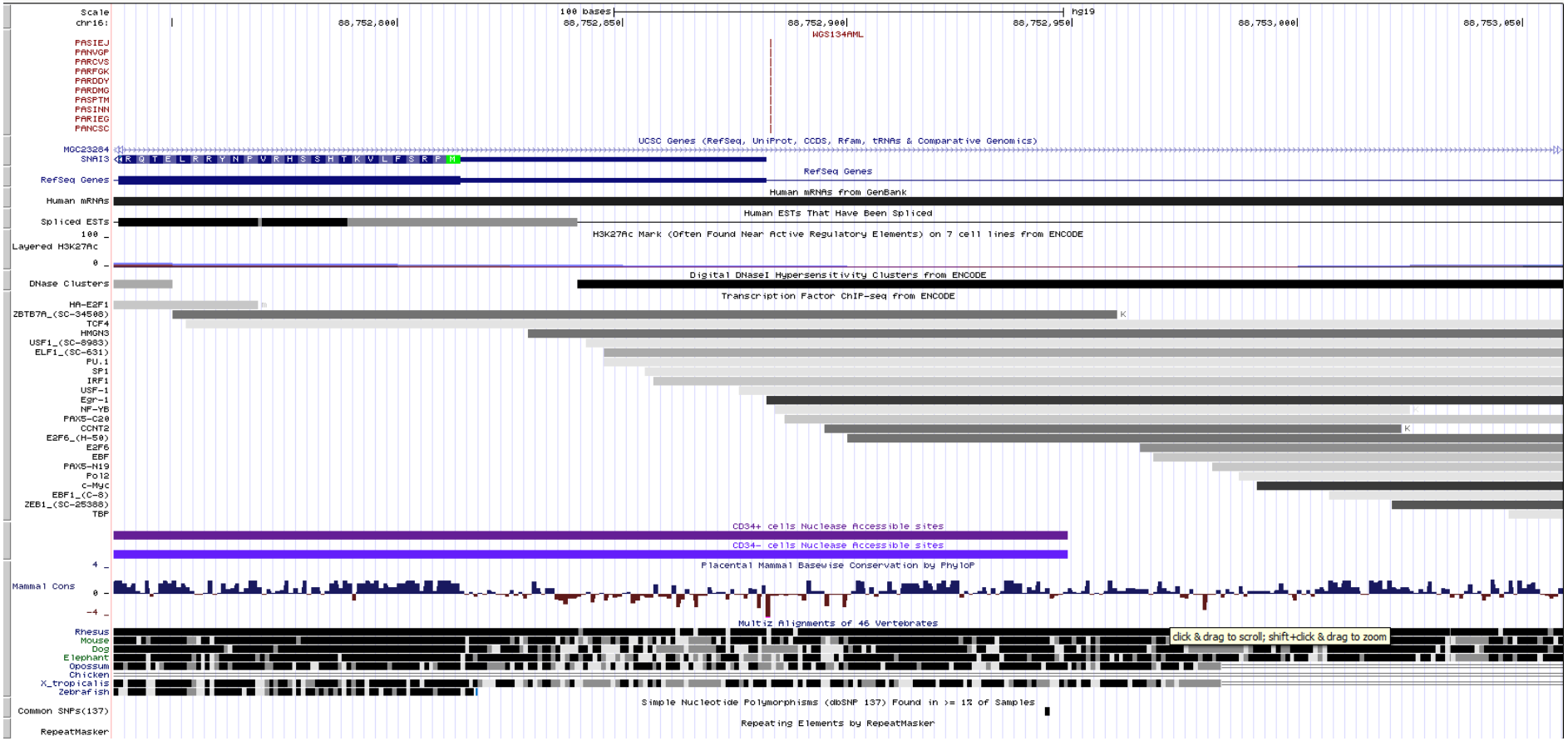
NCBI GEO, Bioconductor packages Graphite, DEGraph, Cytoscape

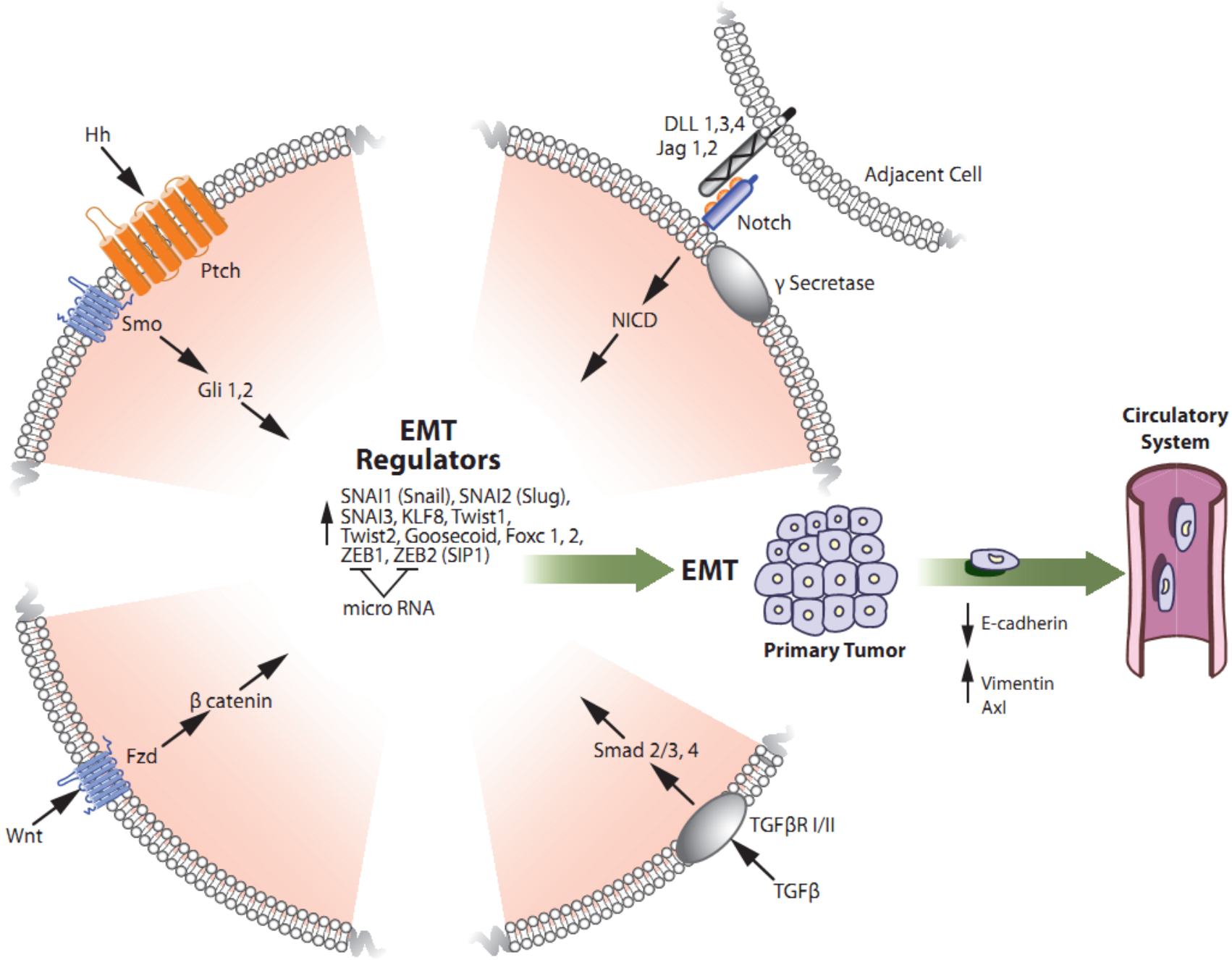




Immunogenetics (2006) 57: 904–916

A highly recurrent SNAIL3 upstream SNV in AML





Takebe et al, Breast Cancer Res. 2011; 13(3):211.

National Cancer Institute

TARGET

Therapeutically Applicable Research to Generate Effective Treatments

http://target.cancer.gov/

NIH

Daniela Gerhardt

Tanja Davidson,…

JHMI (DNA Methylation)

Robert Arceci

Jason Farrar, …

FHCRC (pediatric AML)

Soheil Meshinchi

Rhonda Ries

Ranjani Ramamurthy

Kavita Garg (Tewari lab)

Phoenix Ho, …

Thanks to: Ali Shojaei (UW Biostats)

Paul Shannon & Martin Morgan (Bioconductor team)

CHILDREN'S ONCOLOGY GROUP

The world's childhood cancer experts

Todd Alonzo
Alan Gamis
Rob Gerbing

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health