

Летняя школа по биоинформатике 2017

Multi-omics in biology: integration of omics techniques

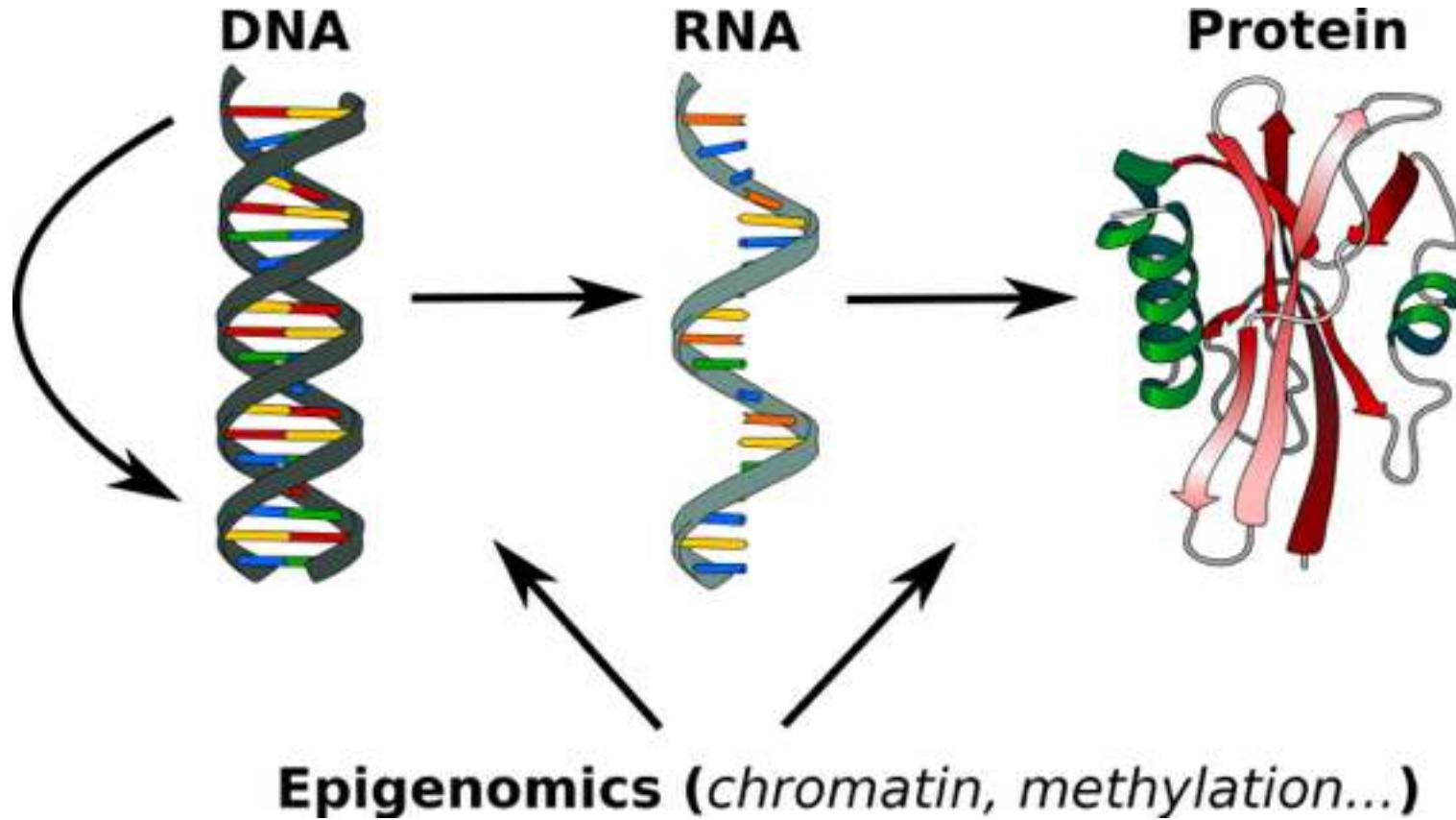
Konstantin Okonechnikov
Division of Pediatric Neurooncology B062
German Cancer Research Center (DKFZ)



Short prefix: use scientific language

- Manuscript standard
- Get used to it:
 - Next Generation Sequencing = Секвенирование нового поколения ~ НГС
 - Paired-end reads = парные прочтения ~ риды
 - Single nucleotide polymorphism = Однонуклеотидный полиморфизм ~ СНиП
 - ...
 - ∞
- Presentation slides are in English

Multi-omics origins



- **Main breakthrough in molecular biology:** ability to get signals from the majority of cell components

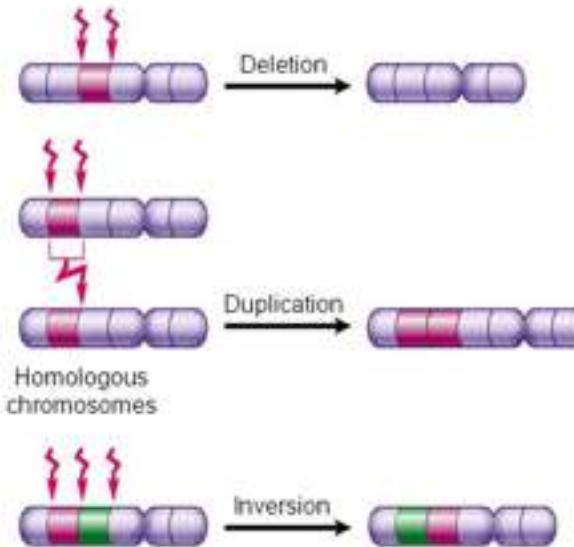
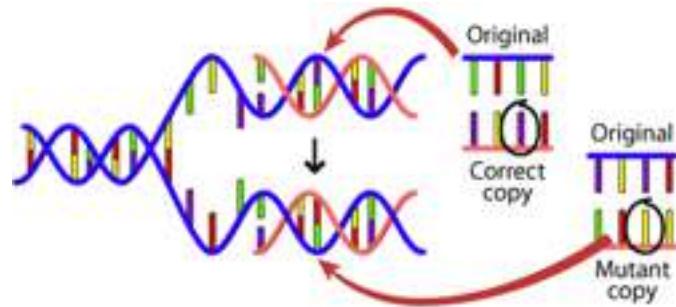
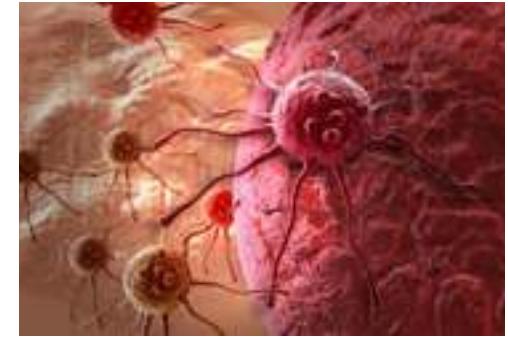
Available methods

- PCR -> DNA, RNA
- Sanger sequencing -> DNA, RNA
- Arrays -> DNA, RNA, methylation
- Spectrometry -> proteins
- ...
- High Throughput Sequencing -> large variety of cell areas



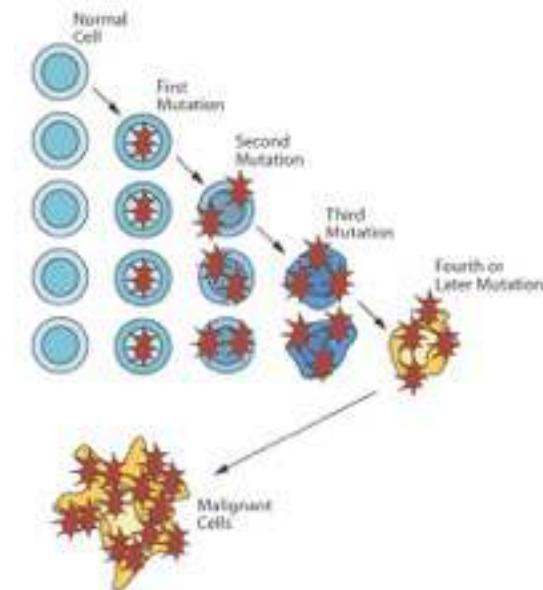
Complex process example: cancer

- Changes in functionality of a cell
- Cause: “unfixed” **DNA damage**
 - Mutations
 - Insertions, deletions
 - Translocations



Complex process example: cancer

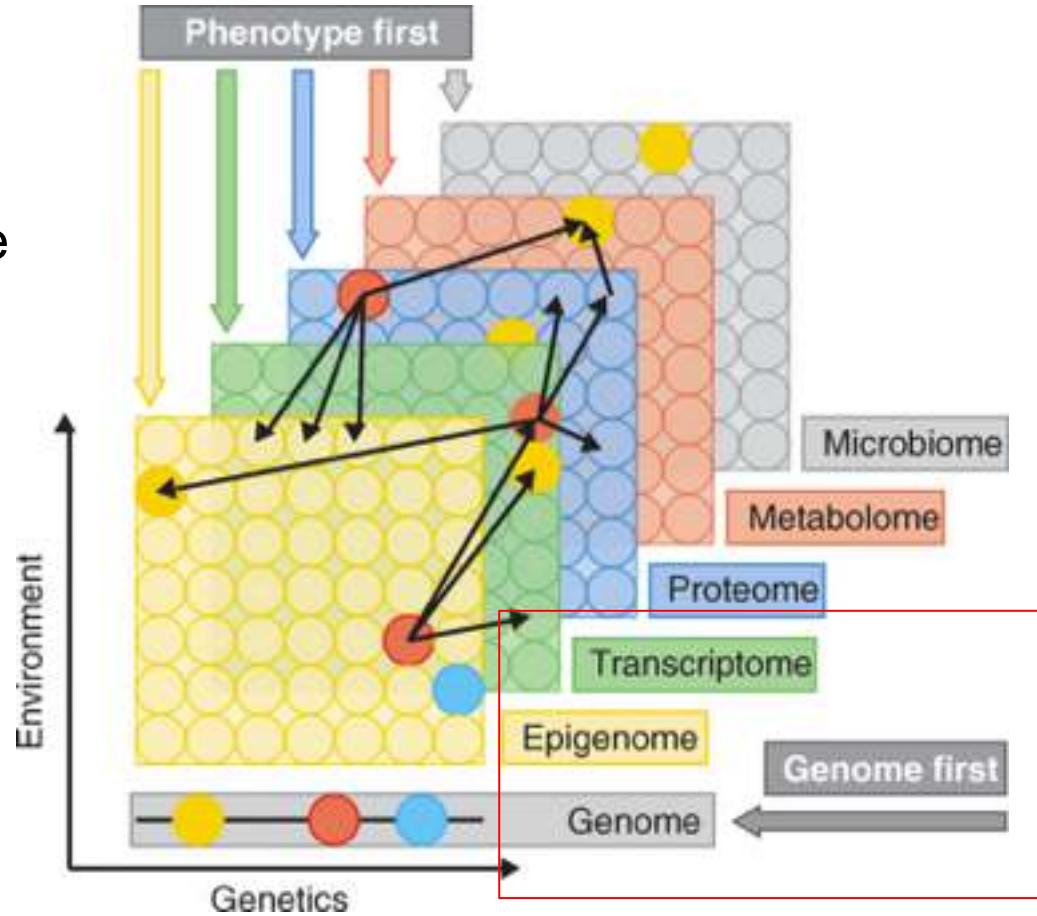
- Development from normal cell to malignant
 - Formed oncogenes
 - Broken repair and tumor suppressor genes
 - Changes in epigenomic activity



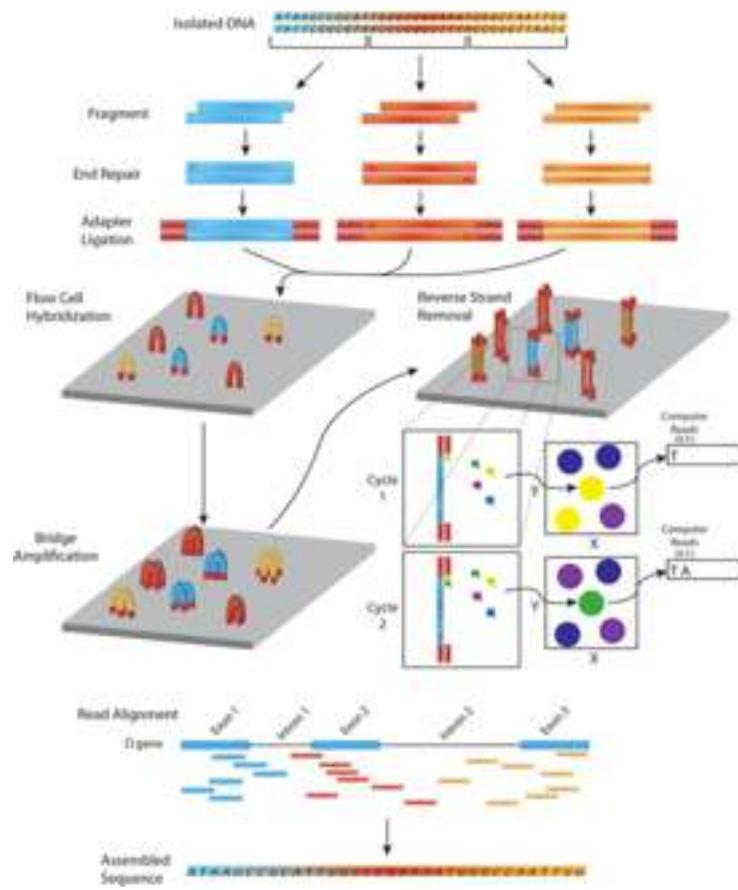
How to find the cause and understand the process?

Multi-omics approach

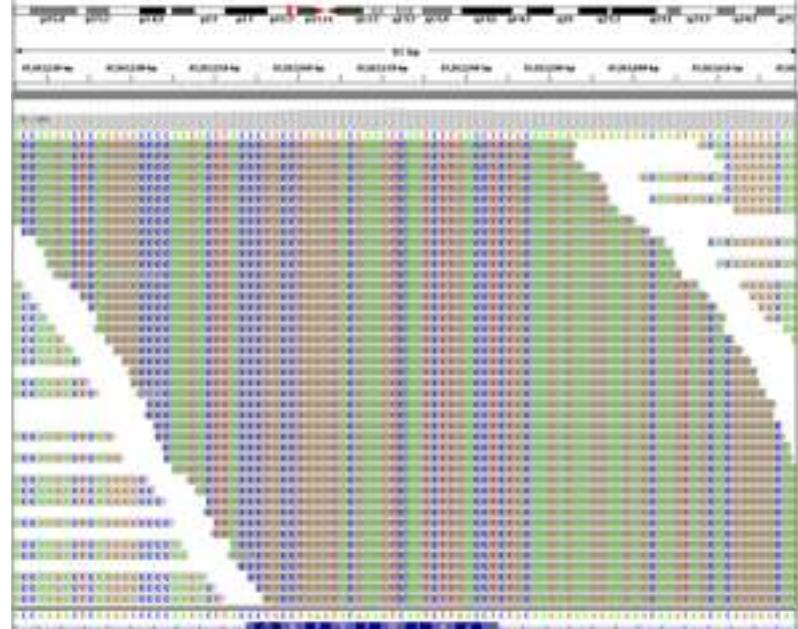
- Integration of different detected changes :
 - Genotype <-> phenotype
 - Cluster and form groups
 - Find possible causative factors
- **Focus on HTS:**
 - *Genomics*
 - *Transcriptomics*
 - *Epigenomics*



High Throughput Sequencing

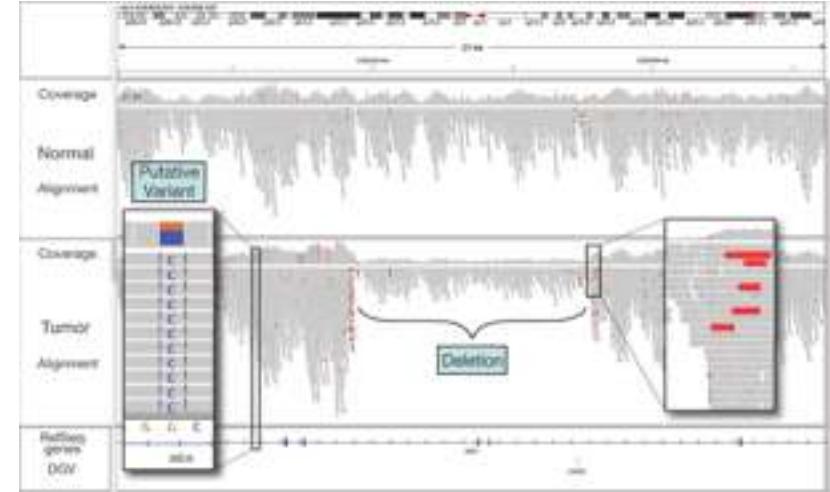
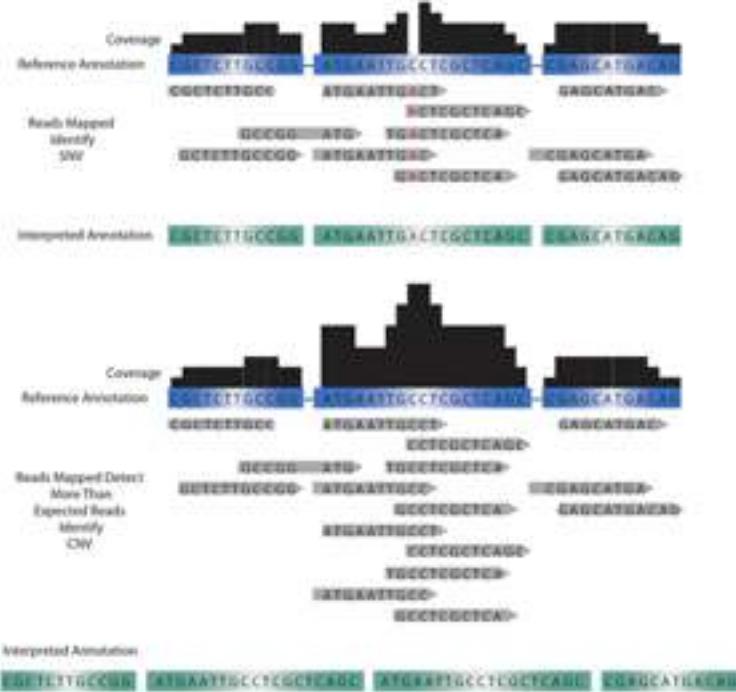


- Protocol properties
- Base analysis
 - Alignment
 - Assembly



Genomics

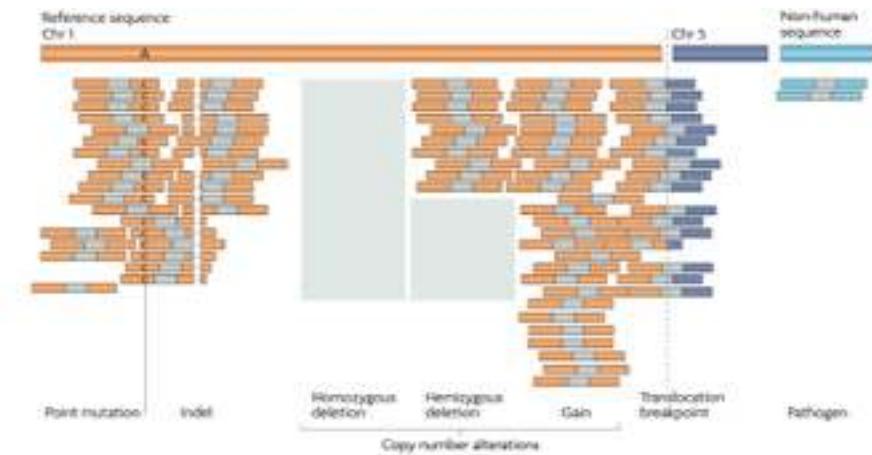
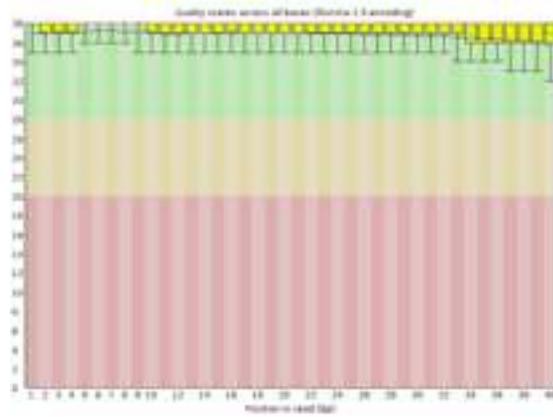
- **Whole genome/exome sequencing**
 - Single nucleotide polymorphisms/variants
 - Population control
 - Structural variations (InDels)
 - Copy number changes



Robinson, J. T., et al. *Nature biotechnology* 29.1 (2011): 24-26.

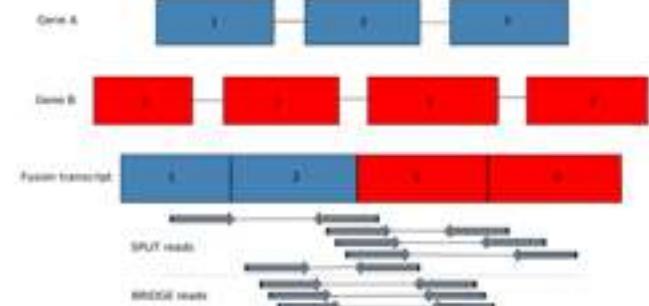
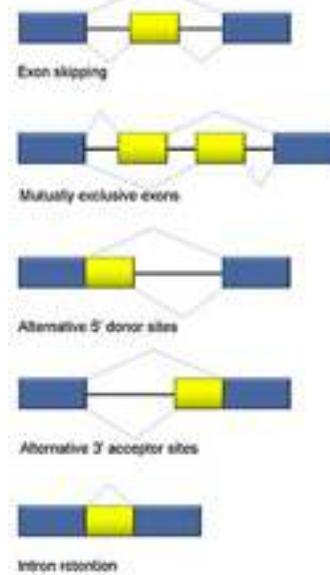
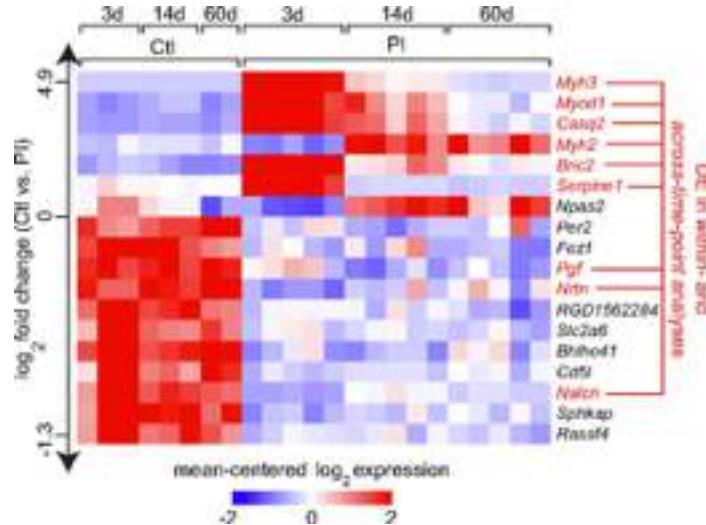
WGS/ES properties

- Procedure control (read length, single/paired, etc...)
- Coverage
 - Exome : 100-200x, genome 20-40x
- Results annotation
 - Genomic properties (gene type, known, etc)
 - Amino acid impact and conservation
 - Population frequency, clinical indication,



Transcriptomics

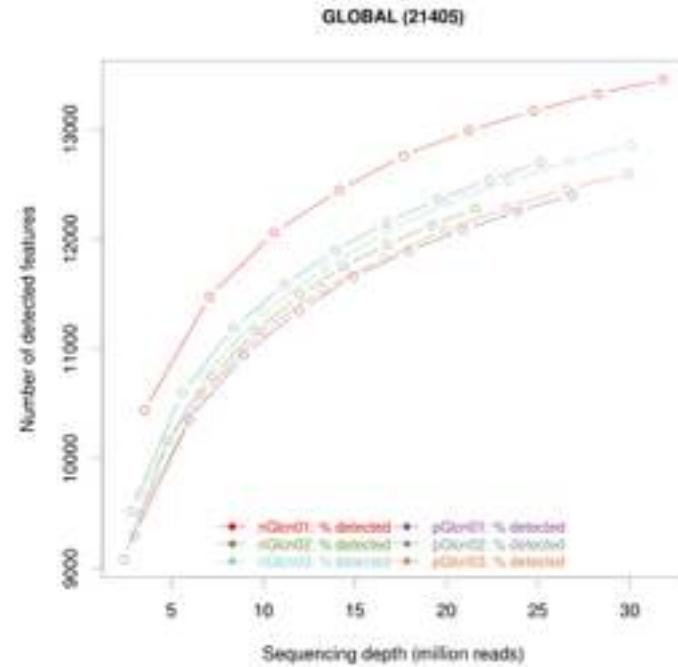
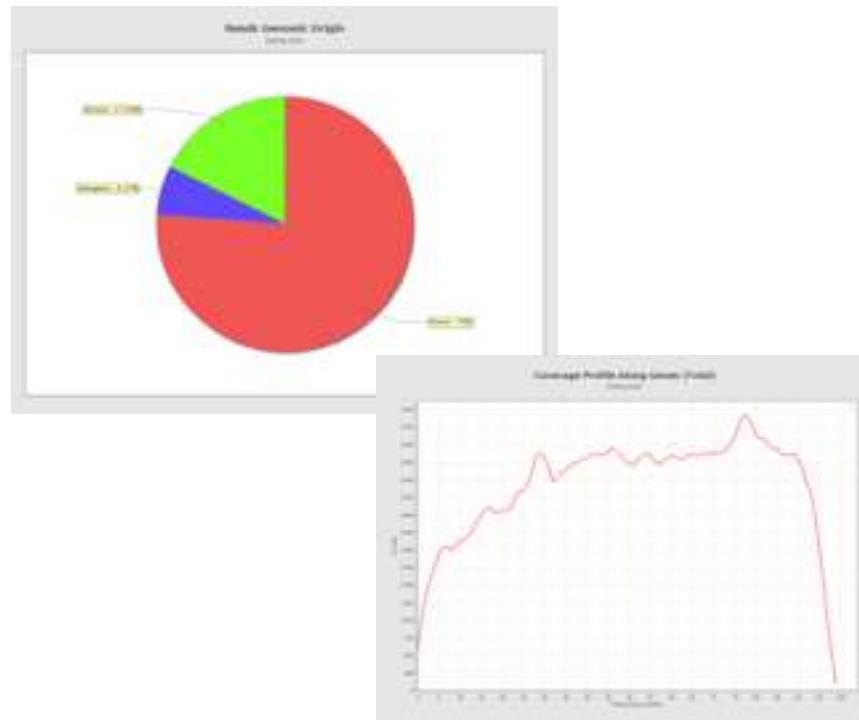
- **Microarrays, RNA-sequencing**
 - Gene expression analysis
 - Various types of RNA: small, non-coding,
 - Alternative splicing and isoforms
 - Fusion genes, chimeric transcripts



K. Okonechnikov et al 2016 PLOS One.

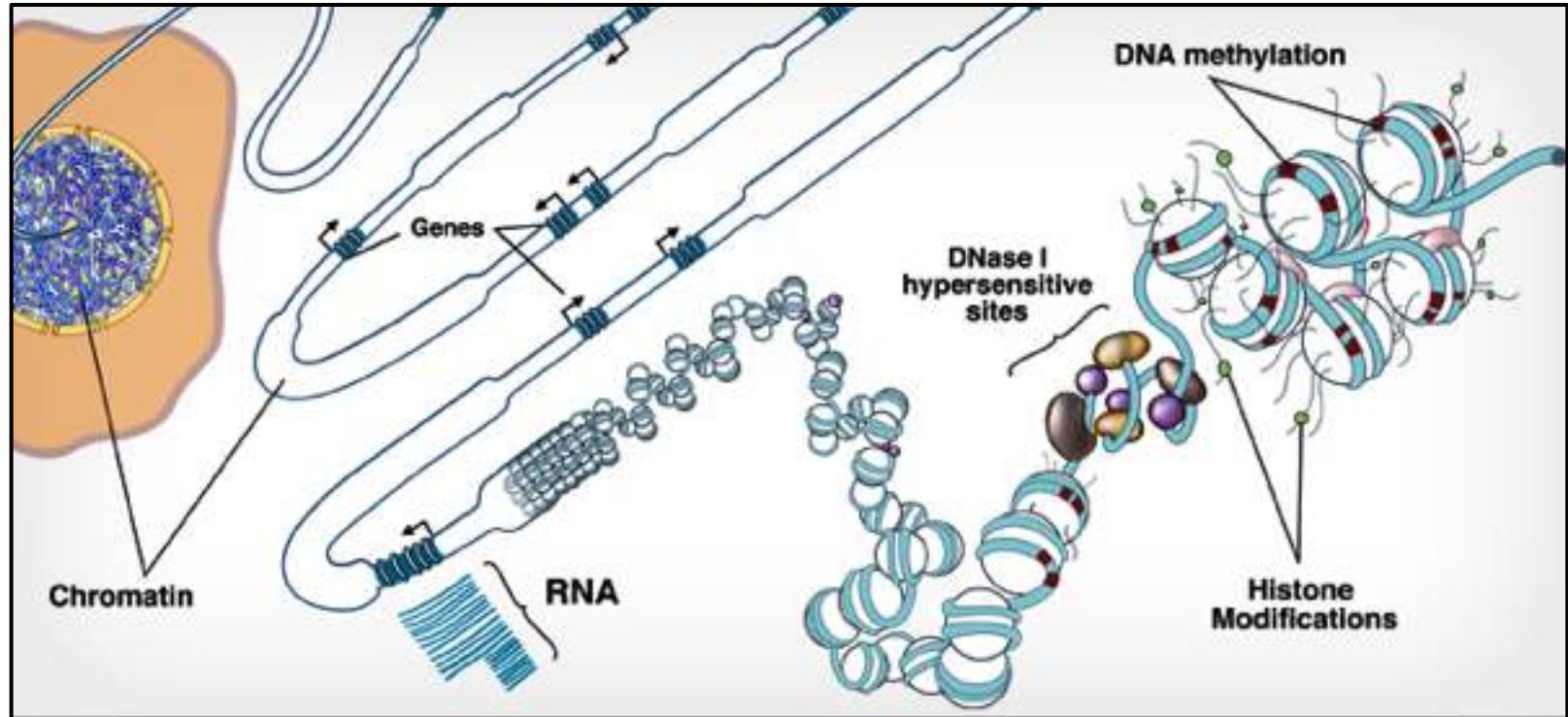
Gene expression profiling

- Specific quality control <- ***important for each type of data***
 - Covered genes, 5'-3' bias, etc..
- Normalization: RPKM, size factor, etc
- Batch effect: source of samples



Epigenomics

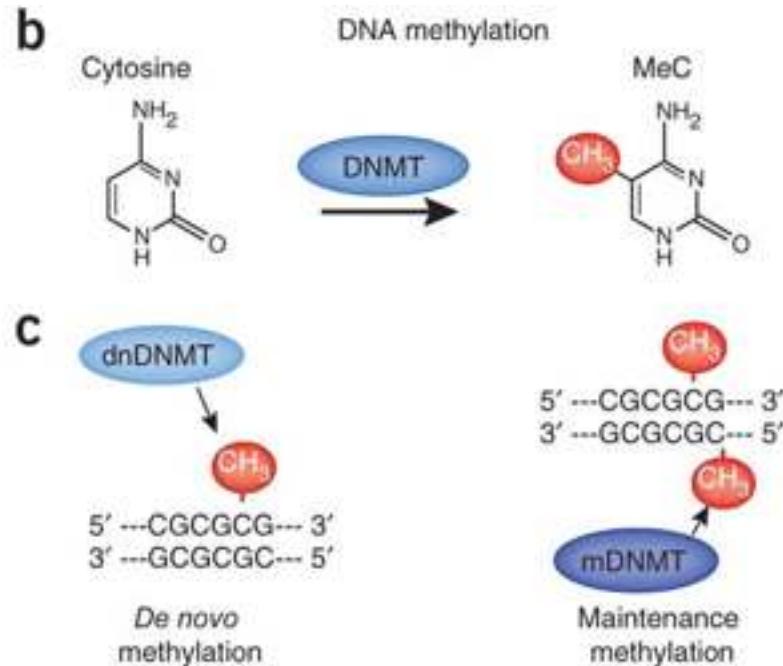
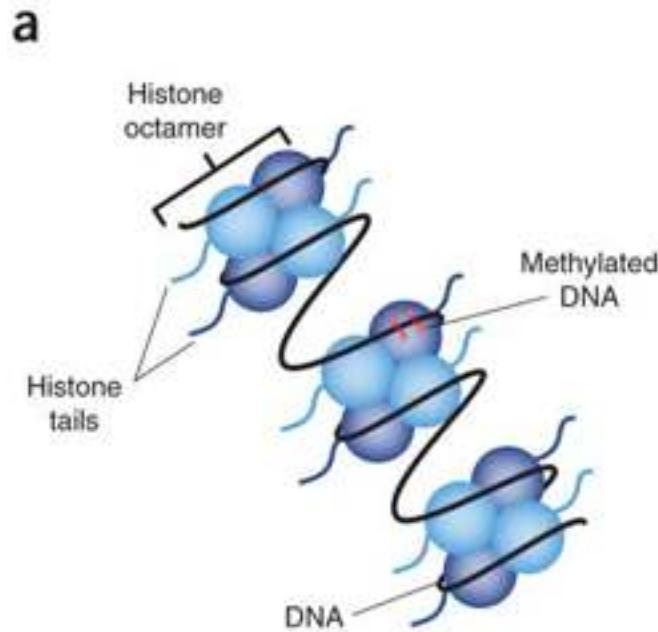
- Changes in gene expression without changes in genome



<http://www.roadmapepigenomics.org>

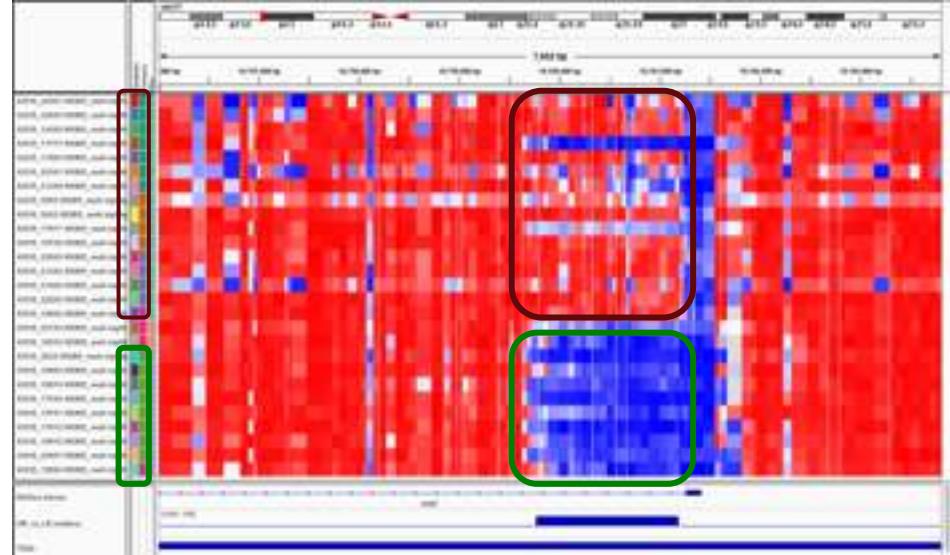
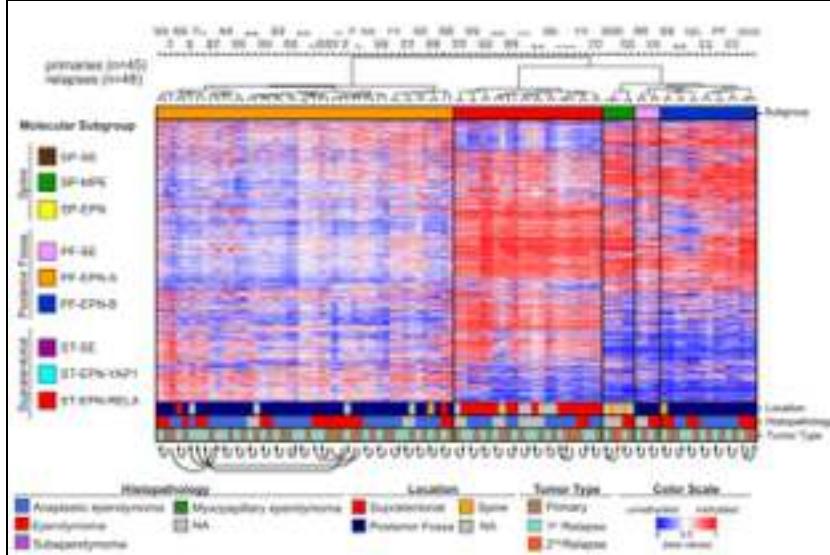
DNA methylation

- Epigenetic mark 5-methylcytosine (5mC), affects up to 80% of CpGs in genome
- Highly active in promoters
- Detection is performed on **CpG sites**



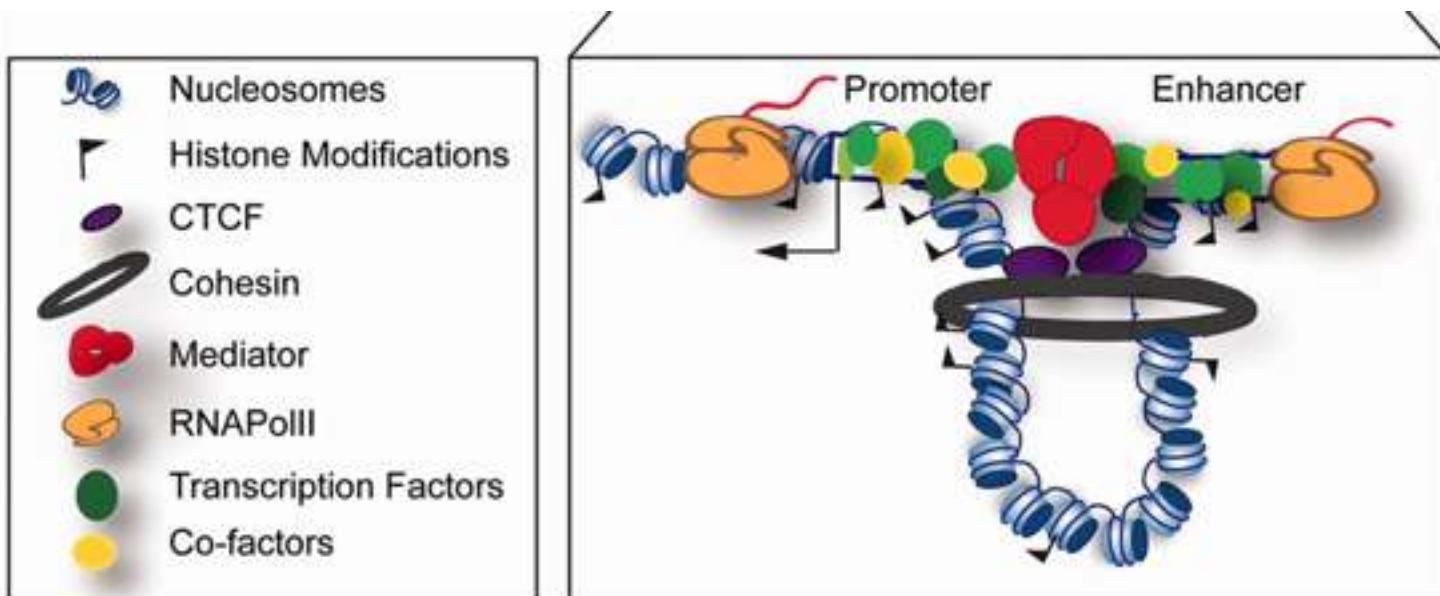
Methylation marks

- **450k arrays, whole genome bisulfite sequencing**
 - Genomic imprinting
 - Differentially methylated regions
 - DNA methylation valleys



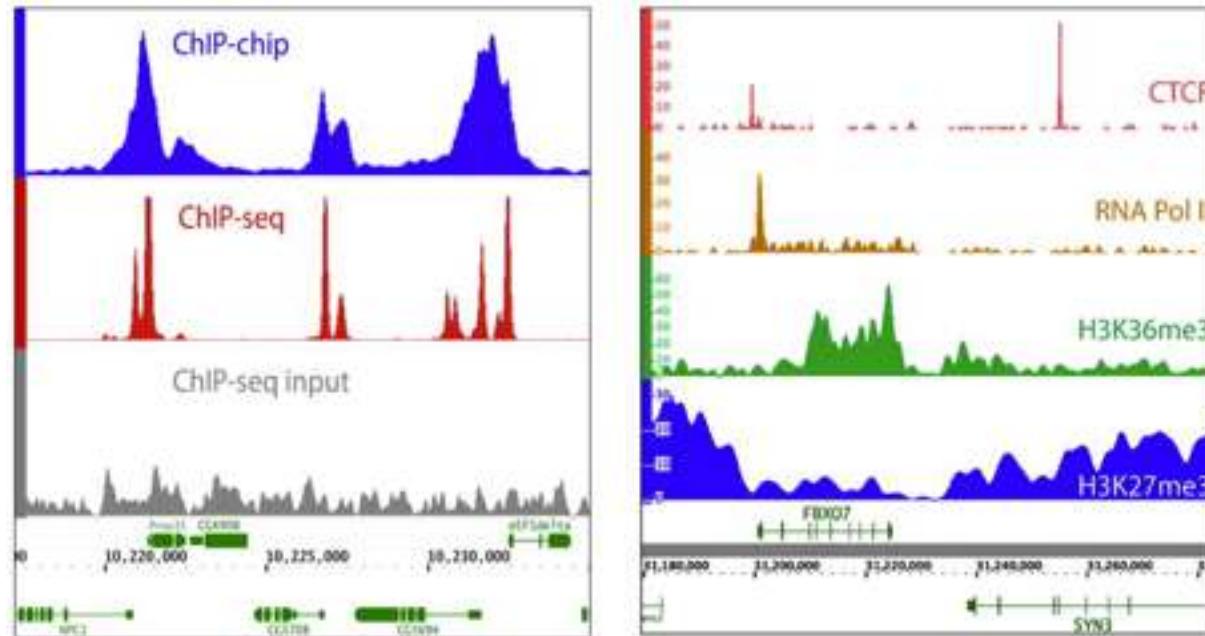
Histone modifications

- **Promoters:** DNA segments of TF binding sites
- **Enhancers:** transcriptional activation/deactivation of target genes
- **Core modifications:** H3K4me3 - active promoters, H3K4me1, H3K27ac – enhancers, ...



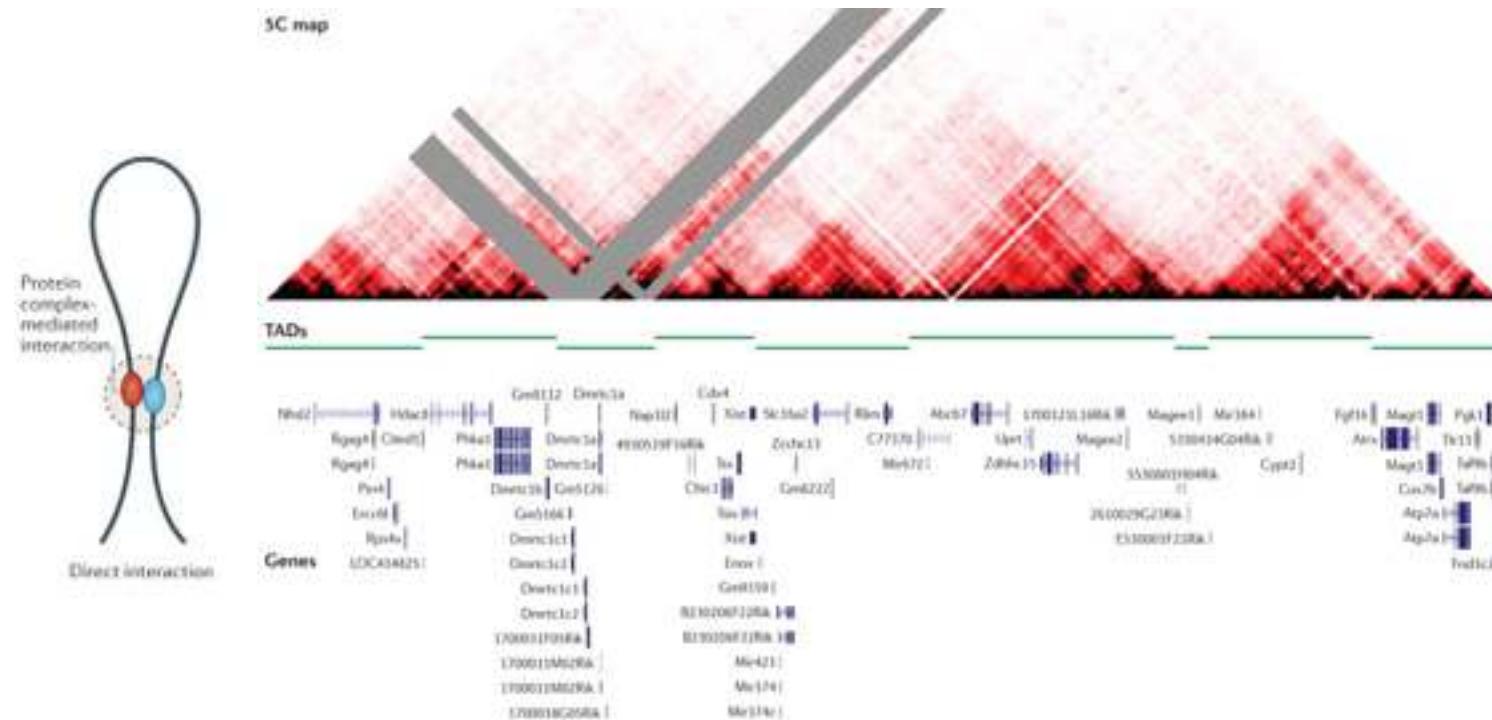
Histone activity variance

- **ChIP-sequencing (TFs, RNA Pol II, H3K27ac, ...):**
 - Detect active promoters and enhancers
 - Investigate the transcription factor activity
 - Find novel epigenetic patterns



Chromatin interactions

- 3C, 4C, HiC sequencing
 - Chromosome conformation capture
 - Spatial organization of genome: connection matrix
 - Topologically Associated Domains



Initial data analysis

- Main task: discovery of possible markers and functional points of biological process
- Standard pipelines:
 - Command line tools – **task specific** (*learn, discuss, check review manuscripts*)
 - Visualization (*IGV, UCSC genome browser, Circos ...*)
 - Workflow management (*Galaxy, Unipro UGENE, ..*)
- Public databases and online resources:
 - Ensembl, GENCODE - **stabilize for all data types**
 - ROADMAP
 - <depends on research focus...>

Summarizing results per sample

- Data availability
- Specific formats

#	Group	ID	WGS	WGBS	RNA-seq	smallRNA-seq	H3K27me
9	WT	AK_417	ICGC_GBM9	GBM9	ICGC_GBM9	ICGC_GBM9	DKFZ_0948
16	K27M	ICGC_GBM16	ICGC_GBM16		ICGC_GBM16	ICGC_GBM16	DKFZ_1072 DKFZ_1184
27	K27M	ICGC_GBM27	ICGC_GBM27	GBM27	ICGC_GBM27	ICGC_GBM27	DKFZ_1427
62	G34R	ICGC_GBM62	ICGC_GBM62	GBM62	ICGC_GBM62		
63	G34R	ICGC_GBM63	ICGC_GBM63	GBM63	ICGC_GBM63		
75	G34R	2071	ICGC_GBM75	GBM75			
76	G34R	AK178		PNET_178	ICGC_GBM76		DKFZ_0532
77	G34R	AK40	ICGC_GBM77	AK40	ICGC_GBM77		DKFZ_0392
78	WT	Glio_1.3	ICGC_GBM78	Glio1_3	ICGC_GBM78		DKFZ_0323

Samples overview

chr	start	end	Enrichment (nrpm)	Type
chr6	33986957	34116956	299,9438024	SSSEA
chr7	44254138	44335962	223,1921727	SSSEA
chr12	3311159	3389177	190,9766377	SSSEA
chr17	77404069	77486837	170,4074676	Enhancer
chr3	10484586	10560420	116,2546028	Enhancer
chr1	17019975	17045474	107,8777923	SSSEA

Enhancer signals per sample

Gene	S1	S2	S3	S4
TMPRSS2	0.234	1.2342	3.34	7.23
ERG	0.723	2.34	5.23	0.23
...

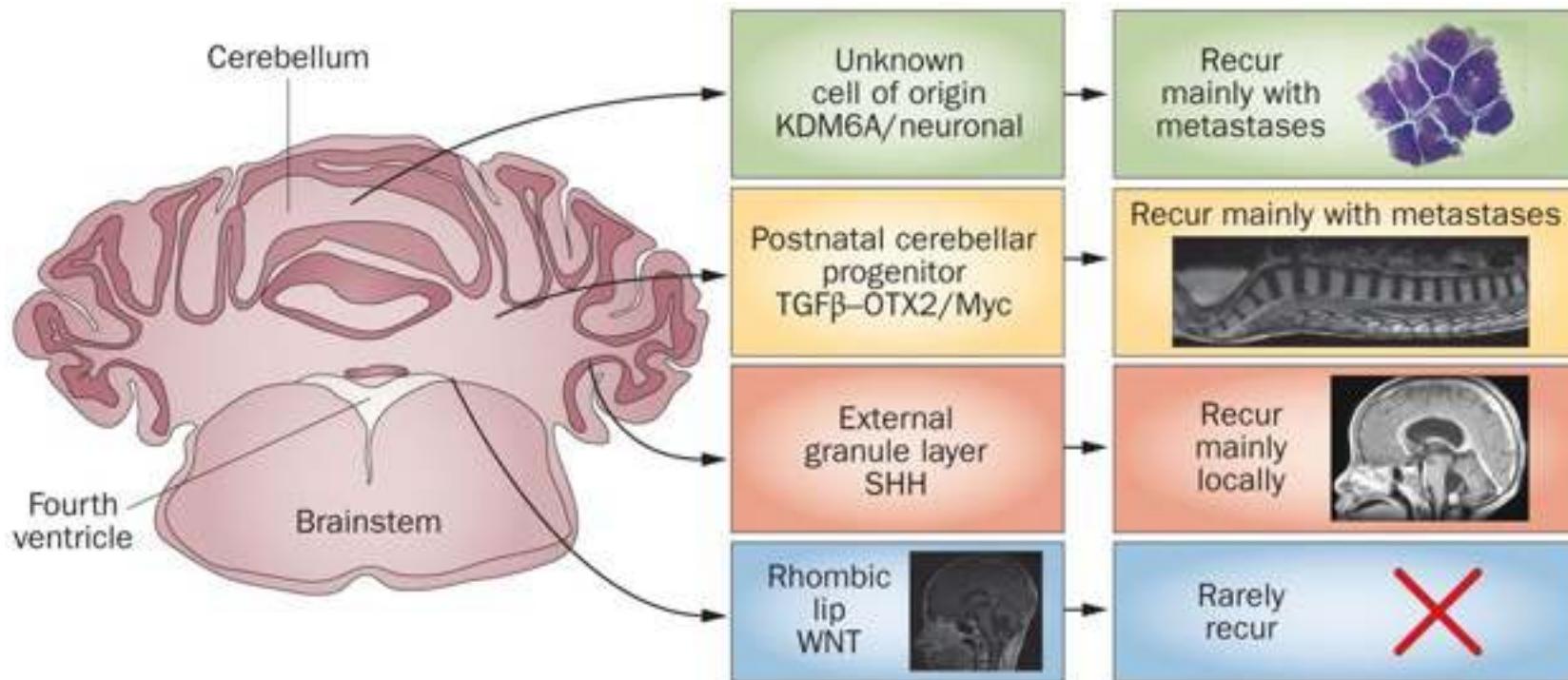
Gene expression per sample

Gene	AA	CHR	POS	REF	ALT	Expected Allele Frequency	Detected Allele Frequency (N=10)	Standard Deviation
EGFR	G719S	7	55241787	G	A	24.5	23.8	1.5
PMSCA	H1047R	3	178952085	A	G	17.5	17.5	1.3
KRAS	G13D	12	25308281	C	T	15	15	1.8
HRAS	Q61K	1	116250130	G	T	12.5	13.4	1.2
BRAF	V600E	7	140453136	A	T	10.5	9.9	0.3
KIT	D816V	4	55599021	A	T	19	10.3	1.1
PMSCA	E545K	3	178936091	G	A	9.0	8.5	1.1
KRAS	G12D	12	25308284	C	T	6.0	6.0	1.2
EGFR	L656R	7	55259615	T	G	3.0	2.7	0.5
EGFR	A674G	7	55242485	Del15bp		2.0	1.4	0.5
EGFR	A750		55242479					
EGFR	T790M	7	55248071	C	T	1.0	1.0	0.3

Sample mutations

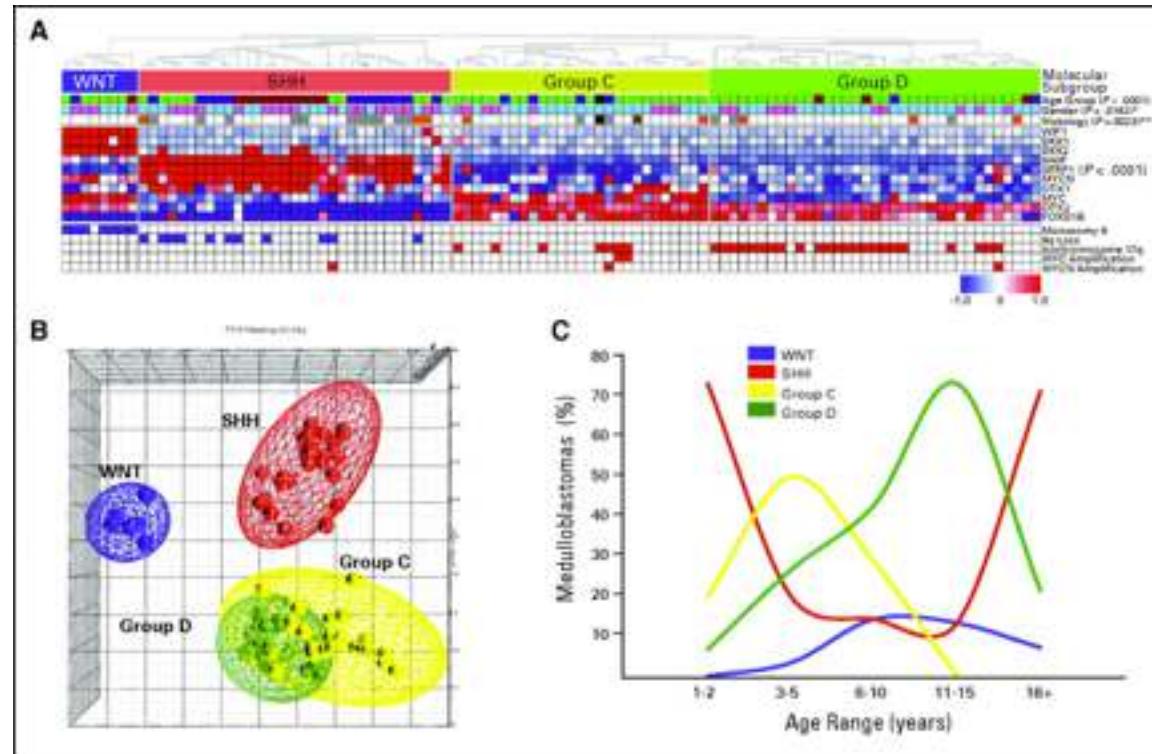
Multi-omics application: medulloblastoma

- Highly malignant tumour, mostly in young children
- Characterized by medical and genetic factors
- Extremely low mutation rate



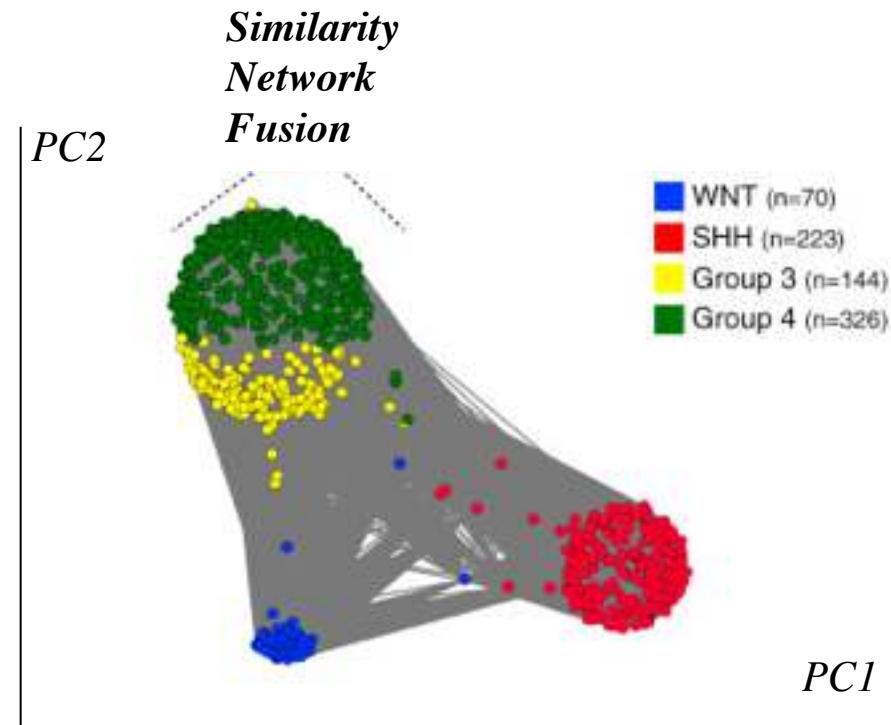
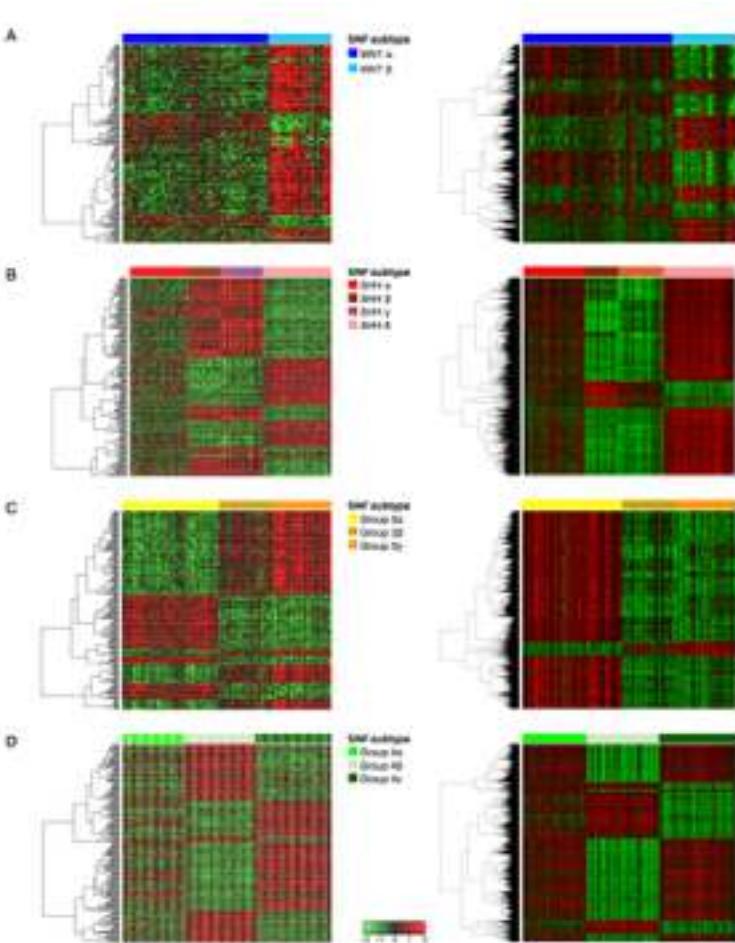
Clustering to find groups

- Methods:
 - Hierarchical clustering, Principal Component Analysis,...
- Resources:
 - RNA-seq, methylation (more stable)



Combined clustering

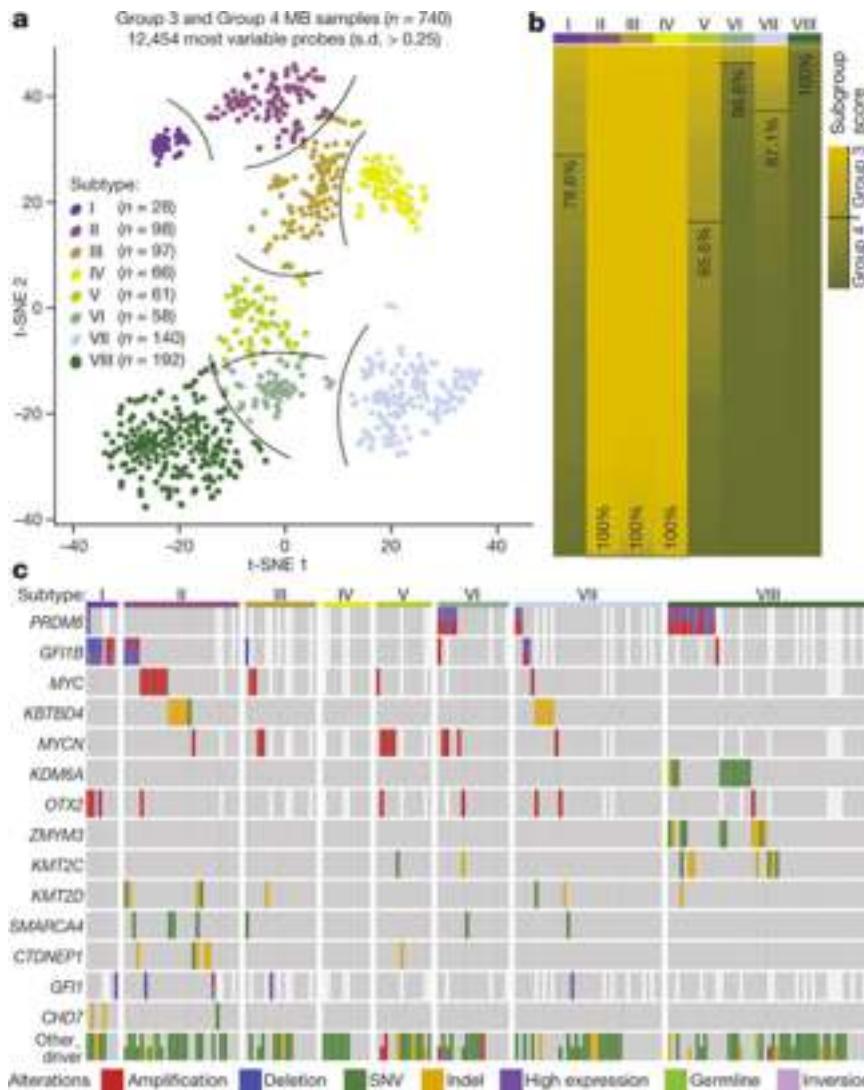
- Combination: RNA-seq + 450k methylation



Integrated Clustering results:

- Clear separation of G3 from G4
- Additional subgroups

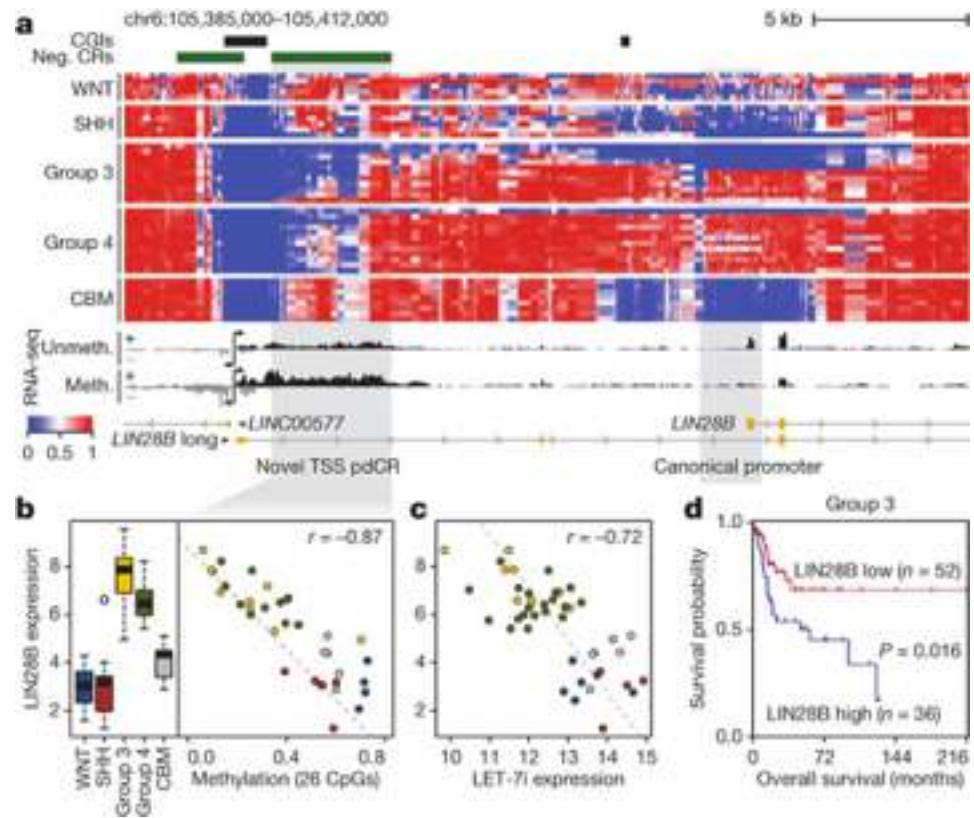
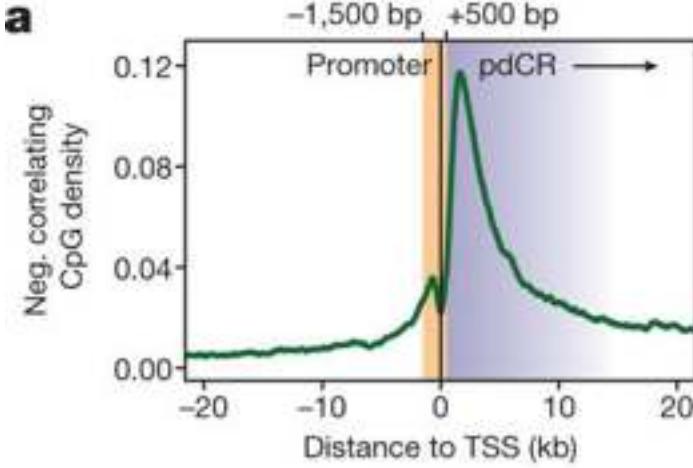
Cluster specificity based on genomics



- WGS:
 - Landscape of mutations, indels, etc specific for samples
- Methylation:
 - Clustering based on *t-Distributed Stochastic Neighbor Embedding*
 - Coordinates' "k-means" to set groups
- Clustering verification: genomic patterns to specify the groups

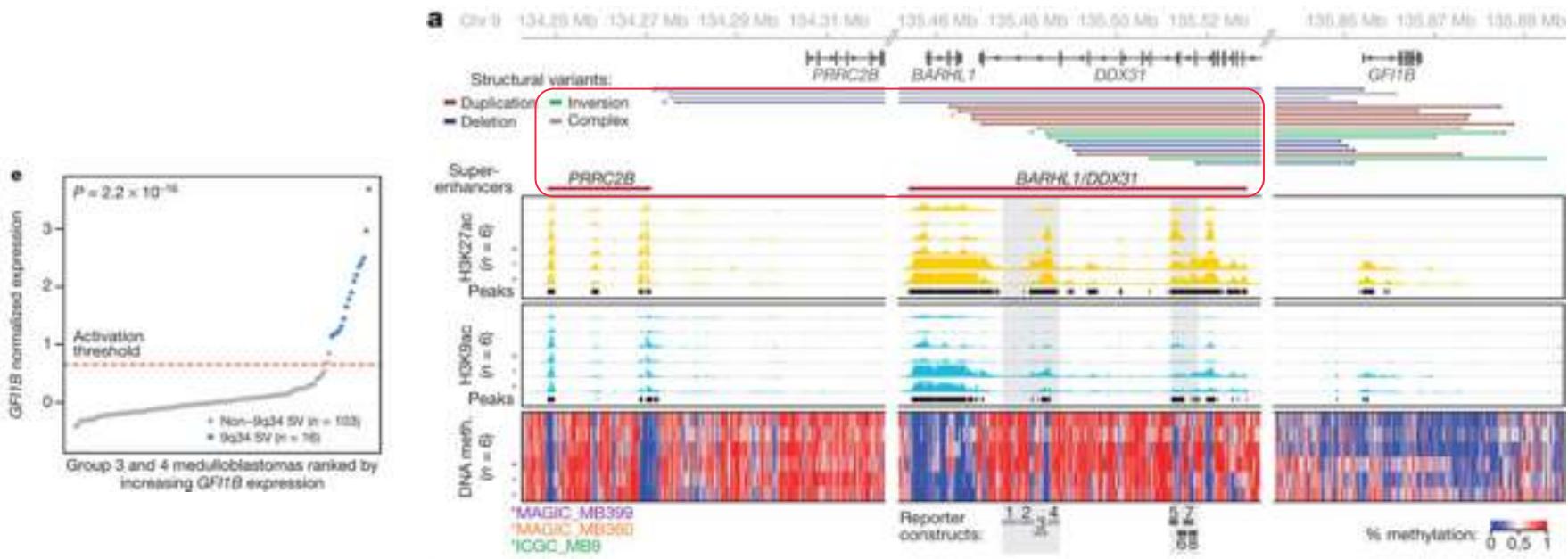
Methylation and expression

- Combination: WGBS + RNA-seq
- Correlation: regions of methylation with expression of overlapping genes



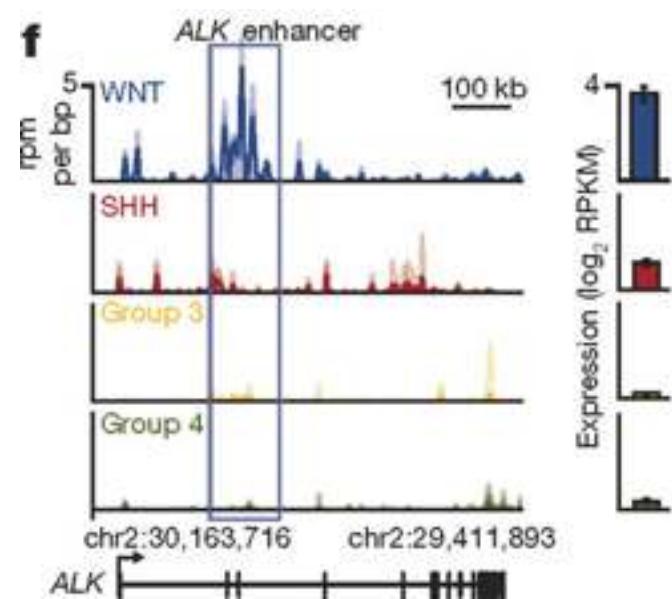
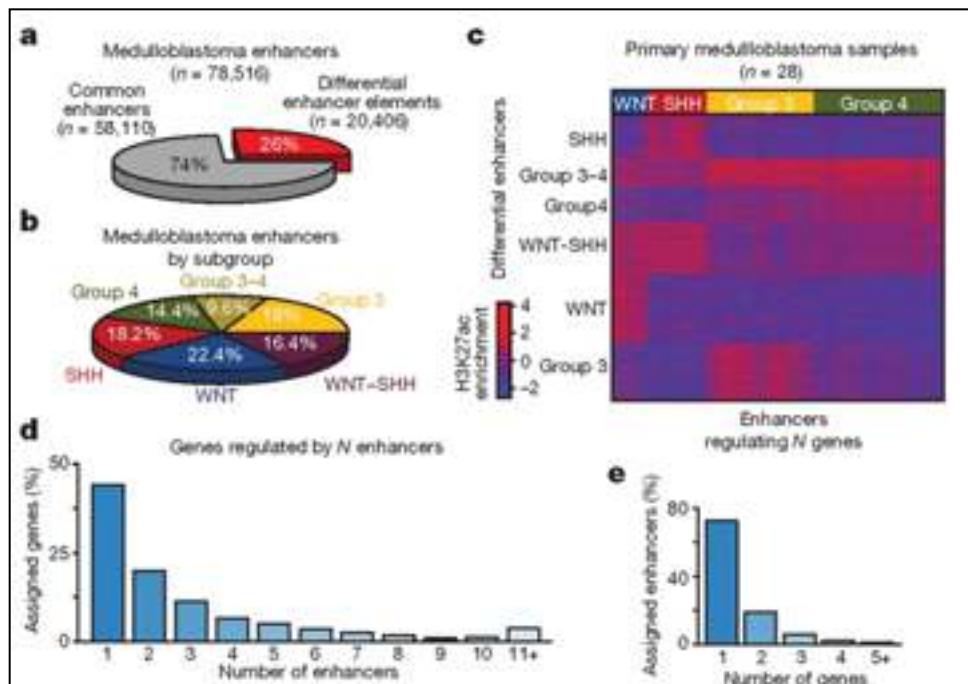
Enhancer hijacking

- Combination: WGS + RNA-seq + ChIP-seq
- Structural variation in genomic with enhancers -> outlier overexpression of a gene
- The variation affects enhancers connection
- Supported by mouse model



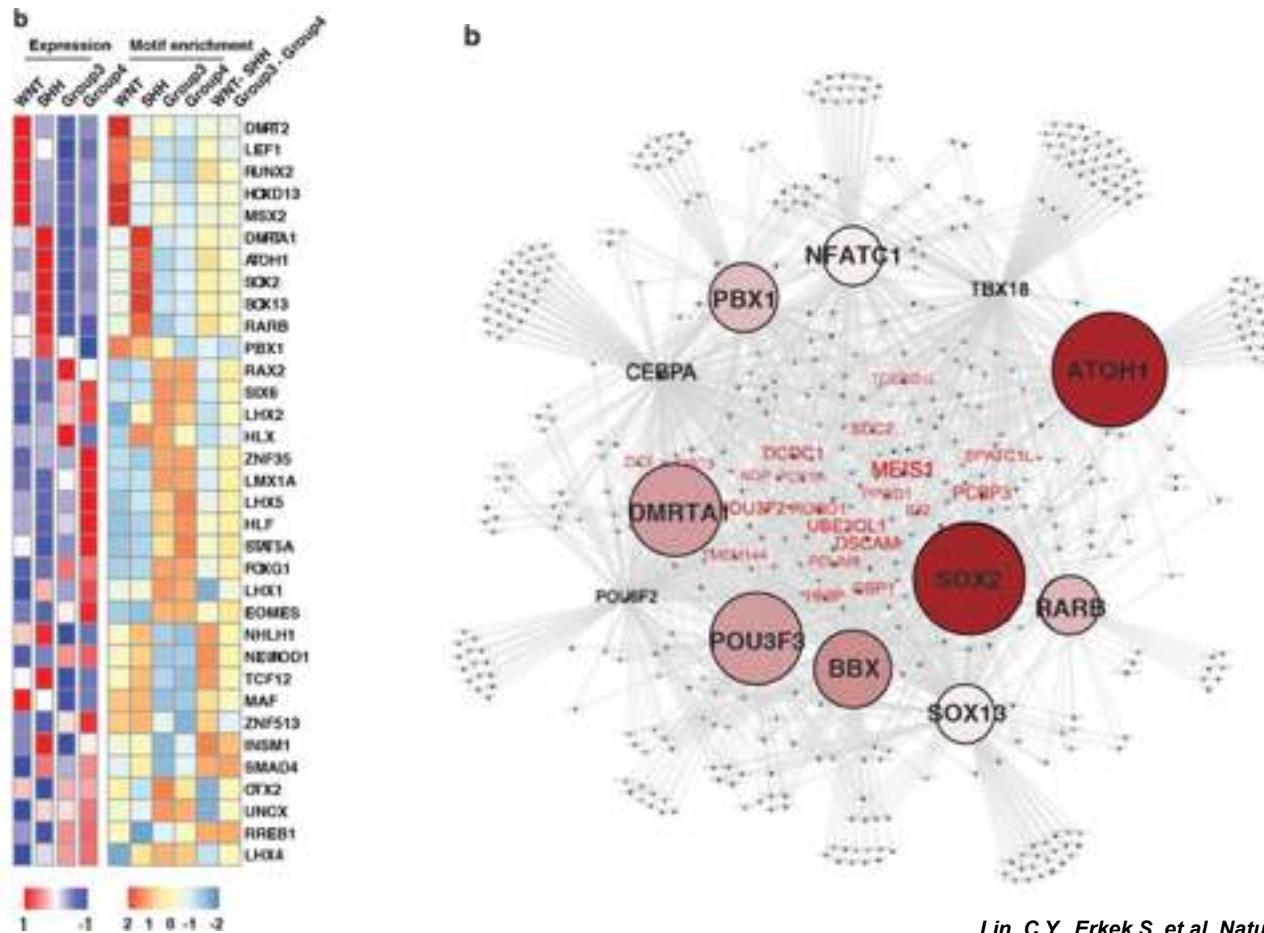
Enhancer associated genes

- Combination: H3K27ac ChIP-seq + RNA-seq + HiC
- Focus on Topologically Associated Domains (IMR90)
- The connection of genes and enhancers



Transcription factor network

- Key players: enhancer enriched and dominant in expression transcription factors



Summary

- Multiple experimental methods allow to obtain various signals from cell activity
- Good example: **high throughput sequencing**. It allows to obtain various genomics, transcriptomics and epigenomics signals
- Combination of multiple technique - **multi-omics** - allows to **confirm observations** and find possible **cause effects**
- **Various strategies** can be applied for this including multi-dimensional clustering, correlation, effects explanation etc

Спасибо за внимание!



Вопросы?