V11 Multi-variate analysis

Program for today:

- *Staphylococcus aureus* Africa project – analysis for confounding variables

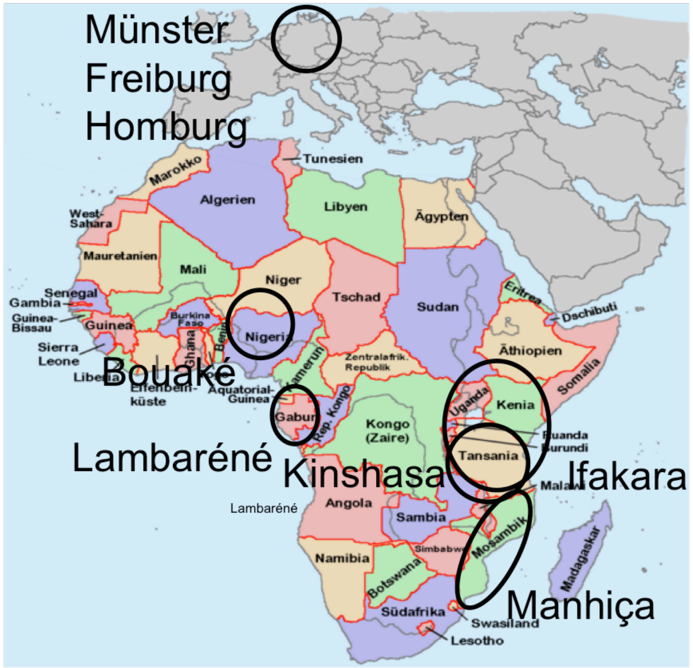
- Overview multivariate analysis for omics projects

- Case study: gene-regulatory network for breast cancer

- Case study: single cell methylation and expression data

Review*: S. aureus* in Germany vs. Africa: StaphNet

6 study sites each collected 100 isolates of healthy volunteers and 100 of blood culture or clinical infection sites.



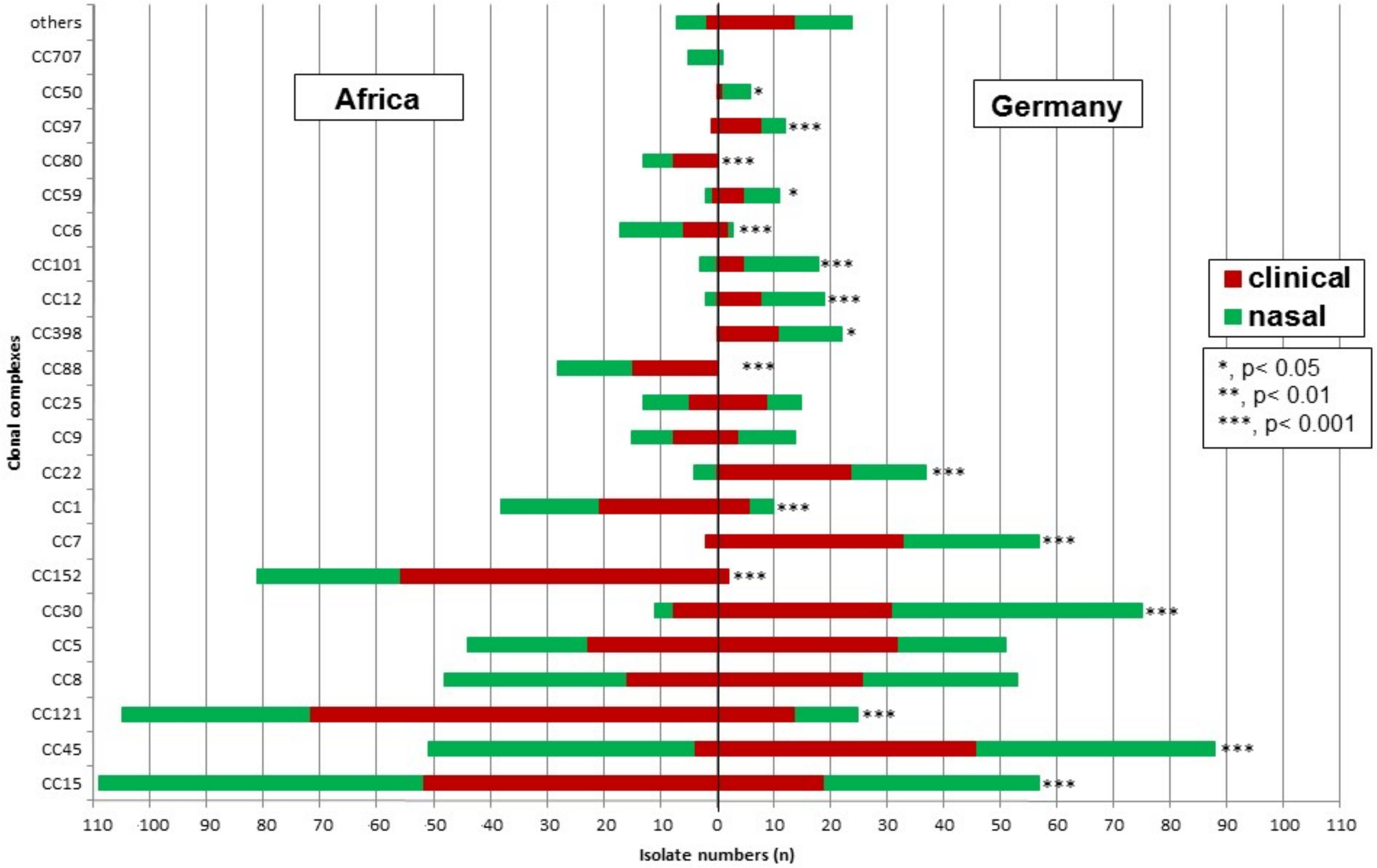
Aim

microbiological and molecular characterization of African *S. aureus* isolates

by DNA microarray analysis including clonal complex analysis

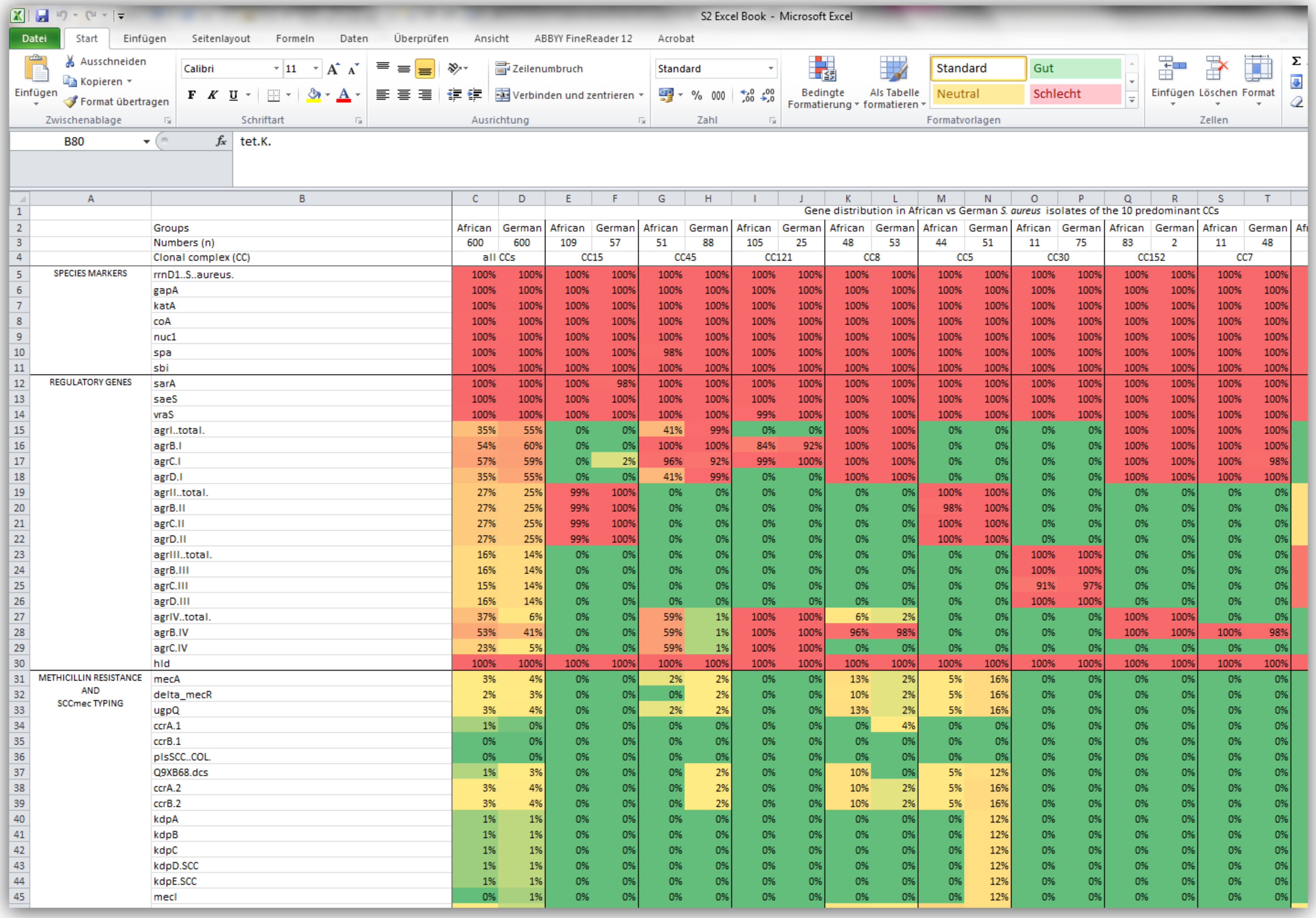
supplemented by Whole Genome Sequencing

Distribution of clonal complexes



Some clonal complexes more prevalent in Africa, others predominant in Germany.

Activitity of individual probes for CCs



B80 *fx* tet.K.

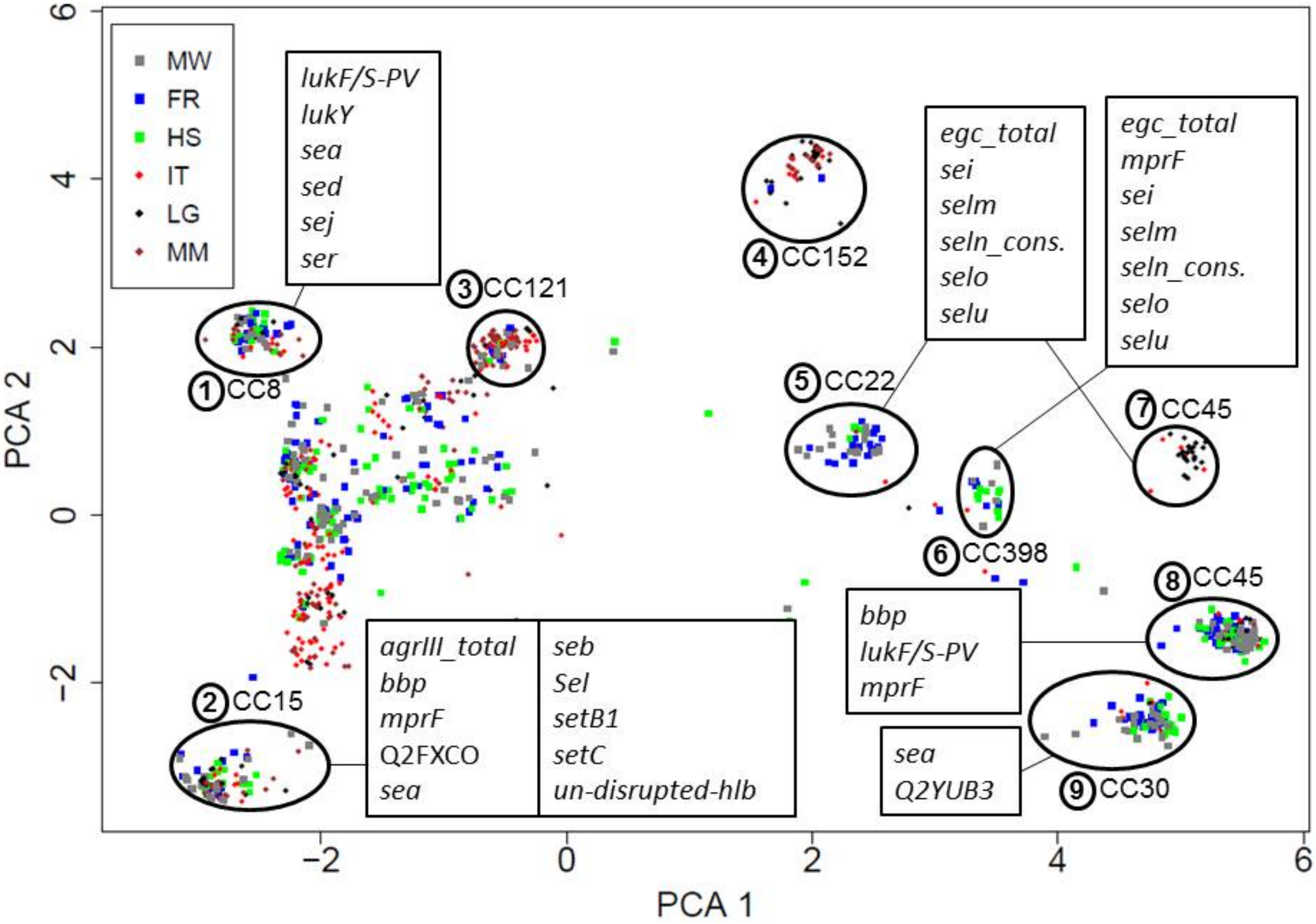
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1					Gene distribution in African vs German *S. aureus* isolates of the 10 predominant CCs																
2		Groups	African	German	African	German	African	German	African	German	African	German	African	German	African	German	African	German	African	German	Afr
3		Numbers (n)	600	600	109	57	51	88	105	25	48	53	44	51	11	75	83	2	11	48	
4		Clonal complex (CC)	all CCs		CC15		CC45		CC121		CC8		CC5		CC30		CC152		CC7		
5	SPECIES MARKERS	rrnD1..S..aureus.	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
6		gapA	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
7		katA	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
8		coA	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
9		nuc1	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
10		spa	100%	100%	100%	100%	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
11		sbi	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
12	REGULATORY GENES	sarA	100%	100%	100%	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
13		saeS	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
14		vraS	100%	100%	100%	100%	100%	100%	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
15		agrI..total.	35%	55%	0%	0%	41%	99%	0%	0%	100%	100%	0%	0%	0%	0%	100%	100%	100%	100%	
16		agrB.I	54%	60%	0%	0%	100%	100%	84%	92%	100%	100%	0%	0%	0%	0%	100%	100%	100%	100%	
17		agrC.I	57%	59%	0%	2%	96%	92%	99%	100%	100%	100%	0%	0%	0%	0%	100%	100%	100%	98%	
18		agrD.I	35%	55%	0%	0%	41%	99%	0%	0%	100%	100%	0%	0%	0%	0%	100%	100%	100%	100%	
19		agrII..total.	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	0%	
20		agrB.II	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	98%	100%	0%	0%	0%	0%	0%	0%	
21		agrC.II	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	0%	
22		agrD.II	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	0%	
23		agrIII..total.	16%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	
24		agrB.III	16%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	
25		agrC.III	15%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	91%	97%	0%	0%	0%	0%	
26		agrD.III	16%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	
27		agrIV..total.	37%	6%	0%	0%	59%	1%	100%	100%	6%	2%	0%	0%	0%	0%	100%	100%	0%	0%	
28		agrB.IV	53%	41%	0%	0%	59%	1%	100%	100%	96%	98%	0%	0%	0%	0%	100%	100%	100%	98%	
29		agrC.IV	23%	5%	0%	0%	59%	1%	100%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	
30		hld	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
31	METHICILLIN RESISTANCE AND SCCmec TYPING	mecA	3%	4%	0%	0%	2%	2%	0%	0%	13%	2%	5%	16%	0%	0%	0%	0%	0%	0%	
32		delta_mecR	2%	3%	0%	0%	0%	2%	0%	0%	10%	2%	5%	16%	0%	0%	0%	0%	0%	0%	
33		ugpQ	3%	4%	0%	0%	2%	2%	0%	0%	13%	2%	5%	16%	0%	0%	0%	0%	0%	0%	
34		ccrA.1	1%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	
35		ccrB.1	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	
36		plsSCC..COL.	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	
37		Q9XB68.dcs	1%	3%	0%	0%	0%	2%	0%	0%	10%	0%	5%	12%	0%	0%	0%	0%	0%	0%	
38		ccrA.2	3%	4%	0%	0%	0%	2%	0%	0%	10%	2%	5%	16%	0%	0%	0%	0%	0%	0%	
39		ccrB.2	3%	4%	0%	0%	0%	2%	0%	0%	10%	2%	5%	16%	0%	0%	0%	0%	0%	0%	
40		kdpA	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	
41		kdpB	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	
42		kdpC	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	
43		kdpD.SCC	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	
44		kdpE.SCC	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	
45		mecI	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	

Principle component analysis of 1200 strains

Input data: binary matrix of MA data; dimension 1200 x 334 probes

PCA identifies local clusters that are characteristic

for particular clonal complexes



Staphylococcus aureus data from Africa project (V1)

Age distribution is heavily skewed:

many small kids / babies in Africa – few seniors in Africa

very few small kids / babies in Germany – many seniors in Germany

Site	# of cases below 1 year	# of cases 1 to 5 years	# of cases 6 - 25 years	# of cases 26 – 65 years	# of cases above 66 years
Africa + Germany (clinical)	**88**	**109**	**90**	**225**	**88**
Africa + Germany (commensal)	19	34	363	175	9
Africa (clinical)	**86**	**106**	**53**	**54**	**1**
Africa (commensal)	17	**34**	156	89	**4**
Germany (clinical)	**2**	**3**	**37**	**171**	**87**
Germany (commensal)	2	**0**	207	86	**5**

Analyze whether age is a confounding variable

To test whether age is a confounding variable, one can compare the results from simple linear regression with those from multiple linear regression.

The principle difference between these two types of regression models is the number of explanatory variables:

(1) the simple linear regression (SLR) model uses only one dependent variable ***y*** and one explanatory variable ***x***: $y = a + b \cdot x$

In our case, ***y*** stands for the binary output from the Alere-chip experiment for a particular gene. ***y*** therefore has values of 0 or 1.

With the binary variable ***x*** we could encode the sites Africa/Germany.
a and ***b*** are weights estimated by the model.

Generally SLR tries to find such weights (values for ***a*** and ***b***) so that the difference between the estimated ***y*** and actual ***y*** will be the smallest.

Analyze whether age is a confounding variable

(2) the multiple linear regression model also has one dependent variable ***y*** but more than one explanatory variables

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_n \cdot x_n$$

As above, ***y*** will be the Alere-chip entry for a gene with value 0 or 1.

The site, clin/com and age categories will be used as explanatory variables .

Steps of testing age categories for confounding

(1) Estimate a linear regression model for the dependent variable and one or more explanatory variables.

(2) Repeat step 1 with age categories added as further explanatory variable.

(3) Compare the weights obtained in steps 1 and 2.

As a rule of thumb, if the weight (-s) (regression coefficient(-s)) from step 1 changes by more than 10%, then the variable (here: age) may be considered as a **confounder**.

By following these steps, one can test for every significant finding (for example, gene association) whether age is a confounder.

Reasons for this could be e.g. a significant imbalance in the distribution of age among samples.

Case study

Case study: test whether age categories are a confounding variable for the 2 genes lukS.PV and sdrC..total

Previously, we found that these 2 genes have different frequencies in African vs German sites as well as in clinical vs commensal samples.

Therefore we will now test age as a confounder in the association of those genes with the Africa/Germany and clinical/commensal categories.

Africa was encoded as 1 and Germany as 0.

Clinical samples were encoded as 1 and commensal with 0.

Age categories were encoded from 1 to 5.

Multiple linear regression model for the lukS.PV gene

The Alere result for this gene for different samples is the dependent variable and the site affiliation, clin/com values are explanatory variables.

The table lists the dependent (lukS.PV) and explanatory (Africa_value, clin_com_value) variables for 10 samples out of 1200 samples.

#	samples	lukS.PV	Africa_value	clin_com_value
1	FR-B001	0	0	1
2	FR-B003	0	0	1
3	FR-B004	0	0	1
4	FR-B005	0	0	1
5	FR-B007	0	0	1
6	FR-B008	0	0	1
7	FR-B009	0	0	1
8	FR-B010	0	0	1
9	FR-B011	0	0	1
10	FR-B012	0	0	1

Since all these samples are from a German site, the Africa_value = 0.
Also, all samples are clinical (clin_com_value = 1).

lukS.PV

Application of linear regression determines optimal weights w_1, w_2, w_3 so that we get for every sample

$$\text{lukS.PV} = w_1 + w_2 \cdot\text{Africa.value} + w_3 \cdot\text{clin com value} .$$

For the first sample FR-B001 the formula would be

$$0 = w_1 + w_2 \cdot 0 + w_3 \cdot 1 .$$

Results from multiple linear regression (coefficients marked in bold):

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.07250	0.01781	-4.070	5e-05 ***
Africa_value	**0.42833**	0.02057	20.825	<2e-16 ***
clin_com_value	**0.19500**	0.02057	9.481	<2e-16 ***

In other words, the following model is estimated:

$$\text{lukS.PV} = -0.07 + 0.42833 \cdot \text{Africa\_value} + 0.195 \cdot \text{clin\_com\_value}$$

lukS.PV

We then added a further variable “age category” with weight w_4 to the model.

lukS.PV = $w_1 + w_2 \cdot$ Africa.value + $w_3 \cdot$ clin com value + $w_4 \cdot$ age

	Estimate	Std. Error	t	value Pr(>|t|)
(Intercept)	0.06211	0.04559	1.362	0.17333
Africa_value	**0.39077**	0.02360	16.556	< 2e-16 ***
clin_com_value	**0.19470**	0.02049	9.503	< 2e-16 ***
age	**-0.03618**	0.01129	-3.206	0.00138 **

Residual standard error: 0.3549 on 1196 degrees of freedom

Multiple R-squared: 0.3102,

Adjusted R-squared: 0.3085

F-statistic: 179.3 on 3 and 1196 DF, p-value: < 2.2e-16

lukS.PV =

0.06211 + 0.39077 · Africa value + 0.19470 · clin com value - 0.03618 · age

lukS.PV

This result shows
(a) that the age category has a very small impact (its own weight is close to 0) and
(b) the two other weights (for the site and clin/com) did not change much.

E.g. the weight of the Africa_values changed in relative terms by :

$$\frac{(0.42833-0.39077)}{0.42833} \cdot 100\% = 8.8\%$$

The weight of clin_com_value changed by only **0.15%.**

Both values are smaller than 10% (rule of thumb).

Conclusion:

There is no statistical evidence that age acts as a confounding variable.

Same analysis for gene sdrC_total

Before adding age categories:

```
Coefficients:         Estimate Std. Error t value    Pr(>|t|)
(Intercept)           1.02083   0.0125     81.711     < 2e-16 ***
Africa_value          -0.12833  0.0144     -8.896     < 2e-16 ***
clin_com_value        -0.05833  0.0144     -4.044     5.6e-05 ***
Residual standard error: 0.2499 on 1197 degrees of freedom
Multiple R-squared: 0.07388, Adjusted R-squared: 0.07233
F-statistic: 47.75 on 2 and 1197 DoF, p-value: < 2.2e-16
```

After adding age categories:

```
Coefficients:         Estimate Std. Error  t value    Pr(>|t|)
(Intercept)           0.975445  0.0321     30.407     < 2e-16 ***
Africa_value          -0.115667 0.0166     -6.964     5.44e-12 ***
clin_com_value        -0.058232 0.0144     -4.039     5.71e-05 ***
age-category          0.012198  0.0079     1.536      0.125
Residual standard error: 0.2497 on 1196 degrees of freedom
Multiple R-squared: 0.0757, Adjusted R-squared: 0.07339
F-statistic: 32.65 on 3 and 1196 DF, p-value: < 2.2e-16
```

Weight of Africa_value changed by **9.87%**, weight of clin_com_value changed by **0.17%**

Conclusion

There is no evidence from our preliminary analysis for the genes lukS.PV and sdrC..total that age acts as a confounder in the associations of genes with invasiveness and site affiliation.

We wrote in our manuscript:
“The discrepancy in population age between the German and African cohort potentially biases the ‘true’ distribution of clones and genes between isolates from the different geographic regions …
[but] application of a multiple linear regression model for the detection rate of Panton-Valentine leucocidin genes failed to provide evidence that age acts as a confounding variable”

Ruffing et al. Sci. Rep. 7, 154 (2017)

Diabetes/HIV as confounding variables

Next, we tested using Fisher's exact test whether
(a) diabetes and HIV have similar frequencies in the total groups of African and German samples and
(b) whether diabetes and HIV have similar frequencies in selected groups of African and German individuals carrying particular clonal complexes.

The Fisher test considers the distribution provided in a 2 x 2 table.

	Africa	Germany	Row Total
HIV+	a	b	a + b
HIV-	c	d	c + d
Column Total	a + c	b + d	a + b + c+ d = n

The formula for the (exact) p-value calculation is :

$$p-value=\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

Explanation: these are the number of possible combinatoric combinations for these fields.

Analysis of HIV co-infection

First, we will test the null hypothesis that “HIV is equally distributed in African and German samples”.

(a) For all African samples and all German samples we obtain the following dependencies of HIV carriers (HIV+) and of individuals without approved HIV status (you may say non-carriers) (HIV-):

	Africa	**Germany**
HIV+	41	0
HIV-	315	586

The p-value obtained for this table can be interpreted as the sum of evidence provided by the observed data—or any more extreme table—for the null hypothesis that “there is no difference in the proportions of HIV carriers among the African and German individuals tested in our study”.

The smaller the value of p, the greater the evidence for rejecting the null hypothesis.

Analysis of HIV co-infection

For the data shown above,

$$p-value=\frac{(41+0)!(315+586)!(41+315)!(0+586)!}{41!0!315!586!942!}=1.03838\text{e-}18$$

Thus, there is very strong evidence from the observed frequencies that African and German individuals **are not equally likely to be HIV carriers**.

Analysis of diabetes co-infection

Similarly, we can obtain Fisher's exact p-value for the distribution of Diabetes among African and German samples.

	Africa	Germany
diab+	4	68
diab-	475	526

p-value = 3.73425e-14

Also, here, the null hypothesis of a similar distribution is **strongly rejected** suggesting the prevalence of Diabetes in individuals from Germany compared to individuals from Africa.

Of course, we can trace this imbalance back to the **difference in age categories** of the two groups.

HIV/diabetes in individuals with selected CCs

Next, we tested the distribution of HIV/Diabetes in individuals carrying *S. aureus* from selected clonal complexes (CC15, CC45, CC121, CC30 which showed significant imbalance in german/african samples).

These are the results (tables + p-values from Fisher's exact test)

RF_HIV

CC15	Africa	Germany
hiv+	4	0
hiv-	65	57
p-value	0.126	0.25 (after correction for false discovery rate (FDR))

CC45	Africa	Germany
hiv+	1	0
hiv-	40	87
p-value	0.320	0.42 (FDR-corrected)

HIV/diabetes in individuals with selected CCs

CC121	Africa	Germany
hiv+	11	0
hiv-	40	24
p-value	0.0132	0.05 (FDR-corrected)

CC30	Africa	Germany
hiv+	0	0
hiv-	7	75
p-value	1	1 (FDR-corrected)

HIV/diabetes in individuals with selected CCs

RF_CCSI_Diab_mel

CC15	Africa	Germany
diab+	0	1
diab-	88	56
p-value	0.393	0.52 (FDR-corrected)

CC45	Africa	Germany
diab+	0	12
diab-	47	75
p-value	0.0081	0.03 (FDR-corrected)

CC121	Africa	Germany
diab+	0	1
diab-	57	24
p-value	0.305	0.52 (FDR-corrected)

CC30	Africa	Germany
diab+	0	7
diab-	9	68
p-value	1	1 (FDR-corrected)

Interpretation

In most cases, there is no evidence based on our data to reject the null hypothesis of assuming a similar distribution of HIV and diabetes carriers among African and German samples belonging to **particular clonal complexes.**

The only exceptions to this are CC45 (diabetes – p=0.008/q=0.03)
and CC121 (HIV – borderline p=0.013/q=0.05).

Therefore, we concluded
“we observed statistically significant imbalances in the frequencies of all these clonal complexes XXX, YYY ... between African and Germany.
We tested based on Fisher's exact test that these imbalances were not due to an imbalance of HIV and diabetes carriers in both groups.
The only exceptions to this are CC45 (diabetes) and CC121 (HIV) where such associations cannot be ruled out.”

Interpretation

On the other hand, the FDR-corrected p-values for CC45 and CC121 are borderline (0.03 and 0.05).

Therefore, there only exists weak statistical evidence for a significant association between CC45 and diabetes or between CC121 and HIV.

integration of multi-omics data

Overview of methods for multivariate analysis:
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0857-9

Different types of high-throughput technologies allow us to collect information on the molecular components of biological systems
- e.g. nucleotide sequencing,
- DNA-chips measuring gene expression and
- protein mass spectrometry measuring protein abundances).

Therefore, in order to draw a more comprehensive view of biological processes, experimental data made on different layers have to be integrated and analyzed.

The development of methods for the integrative analysis of multi-layer datasets is one of the most relevant problems computational scientists are addressing nowadays.

Bersanelli et al. BMC Bioinformatics
(2016) **17(Suppl 2)**:S15

Graph-based integration of multi-omics data

Some group of approaches use graphs to model the interactions among variables.

These approaches, designated as “network-based” (NB), take into account currently known (e.g. protein-protein interactions) or predicted (e.g. from correlation analysis) relationships between biological variables.

Then, graph measures (e.g. degree, connectivity, centrality) and graph algorithms (e.g. sub-network identification) are used to identify valuable biological information.

Importantly, networks are used in the modeling of the cell’s intricate wiring diagram and suggest possible mechanisms of action at the basis of healthy and pathological phenotypes

Bersanelli et al. BMC Bioinformatics
(2016) **17(Suppl 2)**:S15

Bayesian integration of multi-omics data

The second criterion is whether the approach is Bayesian (BY).

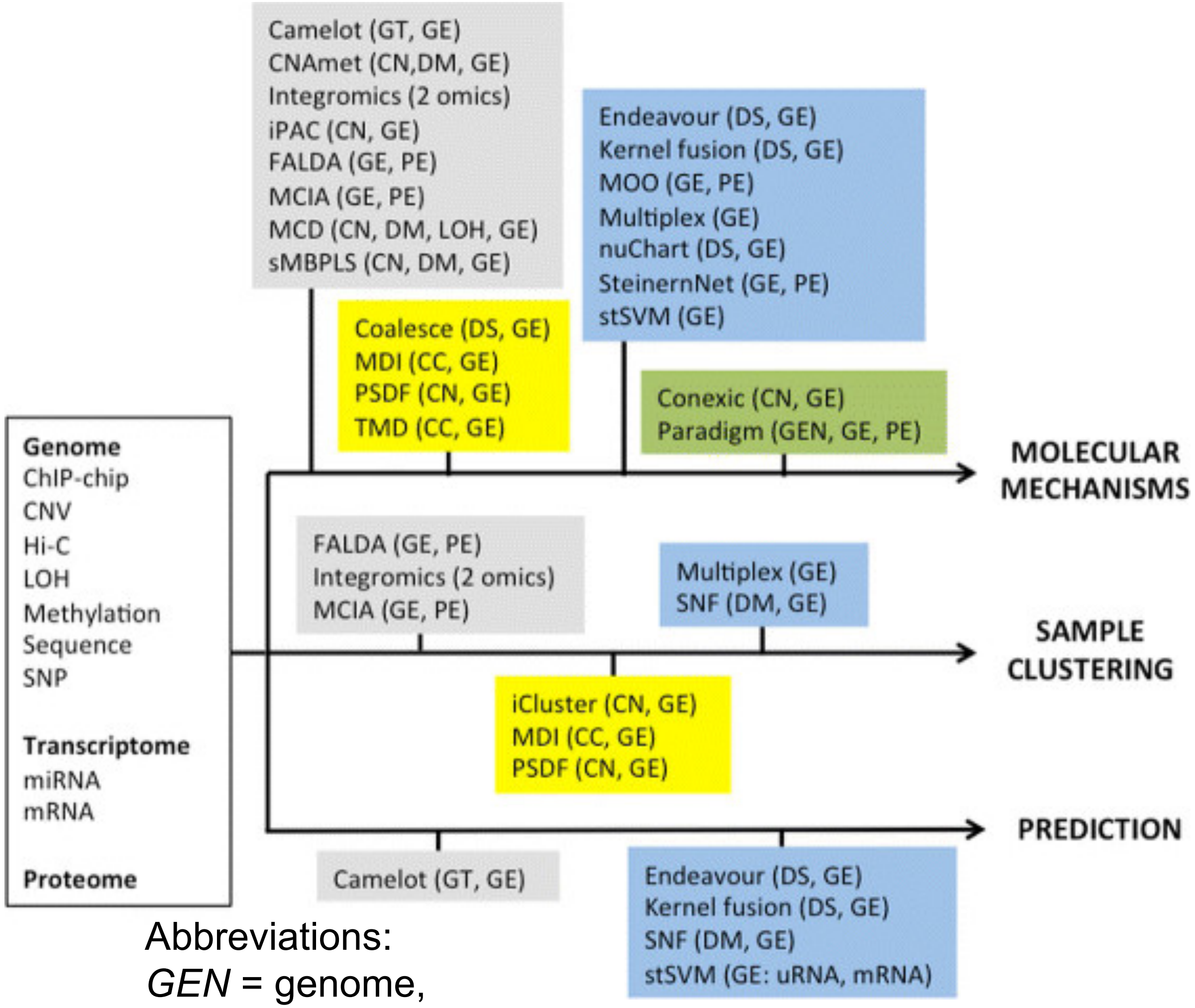
These approaches use a statistical model in which, starting from an *a priori* reasonable assumption about the data probability distribution (*parametric* or *non-parametric*)
it is possible to compute the updated posterior probability distribution making use of the Bayes' rule.

In the network-based area, Bayesian networks are another promising framework for the analysis multi-omics data.

4 classes of methods:
- network-free non-Bayesian (NF-NBY),
- network-free Bayesian (NF-BY),
- network-based non-Bayesian (NB-NBY) and
- network-based Bayesian (NB-BY) methods

Bersanelli et al. BMC Bioinformatics (2016) **17(Suppl 2)**:S15

Overview of multi-omics methods



Camelot (GT, GE)
CNAmet (CN,DM, GE)
Integromics (2 omics)
iPAC (CN, GE)
FALDA (GE, PE)
MCIA (GE, PE)
MCD (CN, DM, LOH, GE)
sMBPLS (CN, DM, GE)

Endeavour (DS, GE)
Kernel fusion (DS, GE)
MOO (GE, PE)
Multiplex (GE)
nuChart (DS, GE)
SteinernNet (GE, PE)
stSVM (GE)

Coalesce (DS, GE)
MDI (CC, GE)
PSDF (CN, GE)
TMD (CC, GE)

Conexic (CN, GE)
Paradigm (GEN, GE, PE)

MOLECULAR MECHANISMS

Genome
ChIP-chip
CNV
Hi-C
LOH
Methylation
Sequence
SNP

Transcriptome
miRNA
mRNA

Proteome

FALDA (GE, PE)
Integromics (2 omics)
MCIA (GE, PE)

Multiplex (GE)
SNF (DM, GE)

SAMPLE CLUSTERING

iCluster (CN, GE)
MDI (CC, GE)
PSDF (CN, GE)

PREDICTION

Camelot (GT, GE)

Endeavour (DS, GE)
Kernel fusion (DS, GE)
SNF (DM, GE)
stSVM (GE: uRNA, mRNA)

Grey: network-free, non-Bayesian methods;

yellow: network-free, Bayesian methods;

blue: network-based, non-Bayesian methods;

green: network-based Bayesian methods

Abbreviations:
GEN = genome,
CC = ChIP-chip,
CN = copy number variations, *DM* = DNA methylation,
DS = DNA sequence,
Hi-C = genome-wide data of chromosomal interactions,
LOH = loss of heterozigosity,
GT = genotype,
GE = gene expression,
PE = protein expression

Bersanelli et al. BMC Bioinformatics (2016) **17(Suppl 2)**:S15

Existing tools

Method	Multi-omics approach	Implementation
Camelot	Bivariate predictive regression model	NA
CNAmet	Multi-omics gene-wise scores	R
FALDA	FA + LDA of a joint matrix	NA
Integromics	Regularized CCA, sparse PLS	R
iPAC	Sequential	NA
MCD	Sequential	NA
MCIA	Multiple co-inertia analysis	R
sMBPLS	Sparse Multi-Block PLS regression	Matlab
Coalesce	Multi-omics probabilities	C ++
iCluster	Joint Gaussian latent variable models	R
MDI	DMA mixture models	Matlab
PSDF	Hierarchical DMA mixture models	Matlab
TMD	Hierarchical DMA mixture models	Matlab
Kernel Fusion	Integration of omics-specific kernels	Matlab
Endeavour	Integration of omics-specific ranks with order statistics	Webserver
MOO	Sub-network extraction on MWG	R
Multiplex	Joint analysis of multi-layered networks	NA
NuChart	Analysis of a MWG	R
SNF	Similarity network fusion	Matlab, R
SteinerNet	Sub-network extraction on MWG	Webserver
stSVM	MWG	R
Paradigm	Multi-omics bayesian factor graphs	C ++
Conexic	Sequential	Java

legend

MWG = multi-weighted graph;

FA = factor analysis;

LDA = linear discriminant analysis;

CCA = canonical correlation analysis;

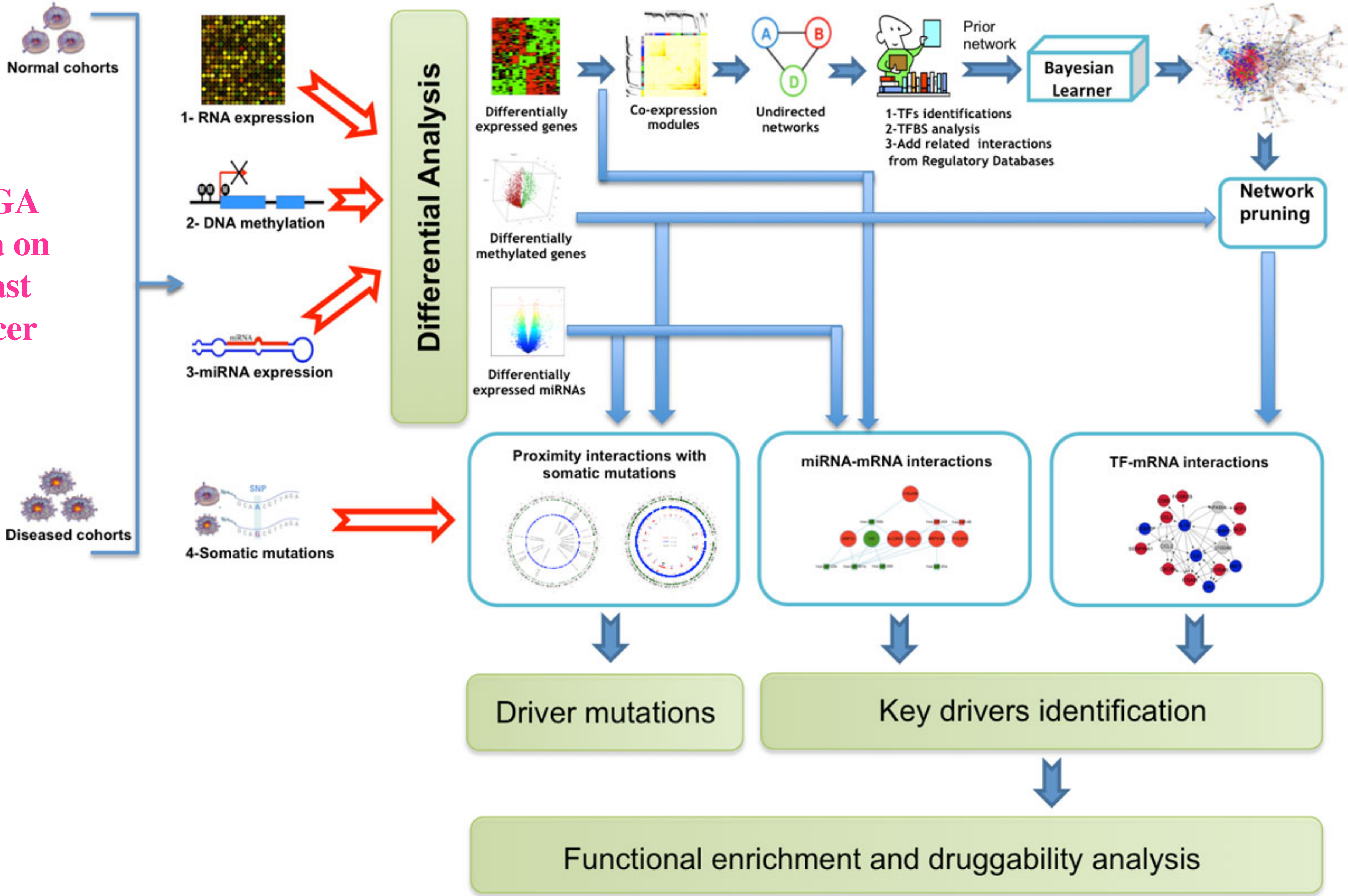
PLS = partial least squares;

DMA = Dirichelet multinomial allocation

Bersanelli et al. BMC Bioinformatics (2016) **17(Suppl 2)**:S15

Multi-omics analysis of breast cancer network

TCGA data on breast cancer



Hamed et al. BMC Genomics 16 (Suppl5):S2 (2015)

Breast cancer network from TCGA data

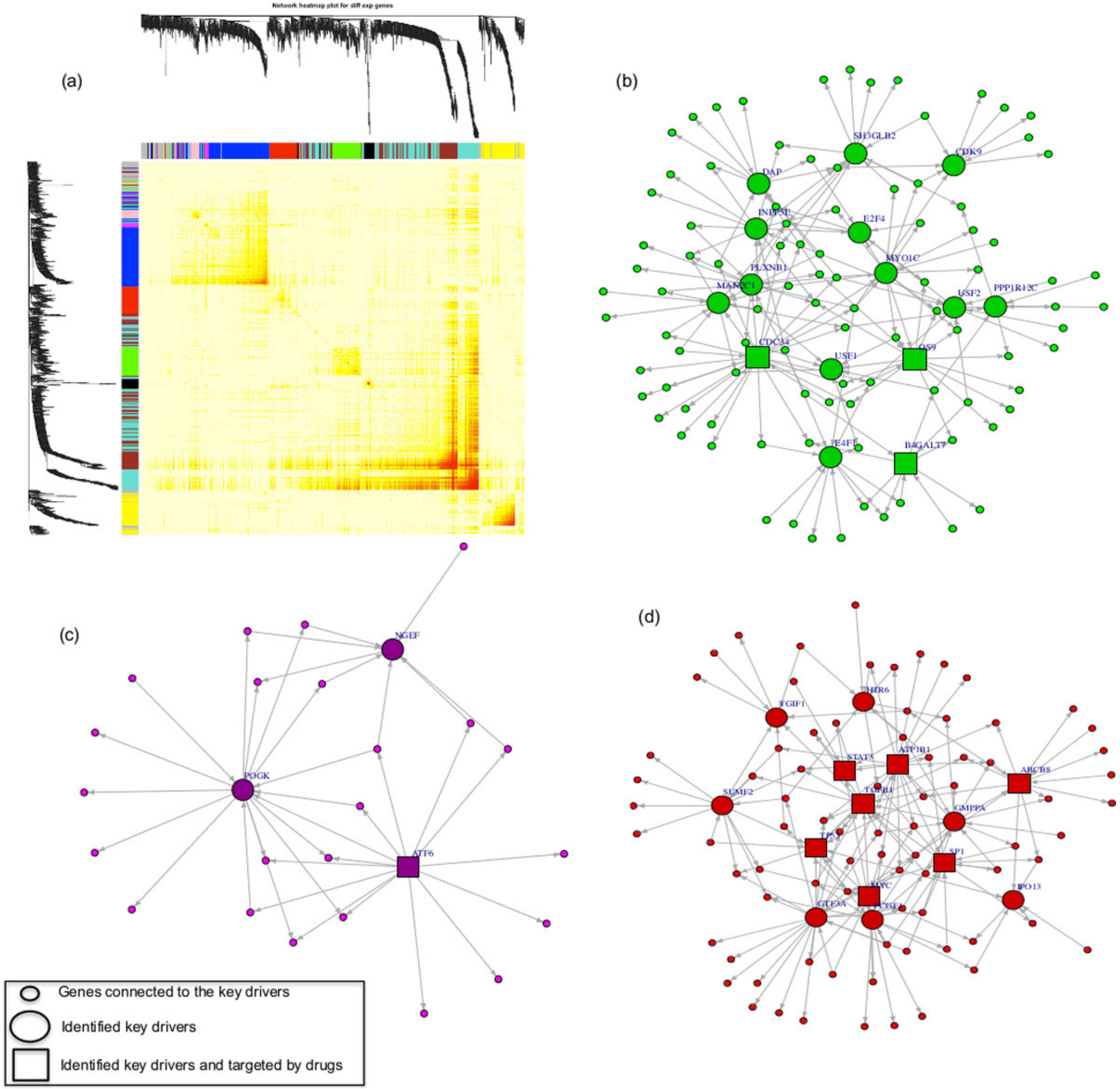
ca. 1300 differentially expressed genes.

Hierarchical clustering of co-expression network:
10 **modules** with
26 - 295 genes.

Regulatory info from databases Jaspar, Tred, MSigDB.

Shown are 3 modules.

Squares are known drug targets.



Hamed et al. BMC Genomics 16 (Suppl5):S2 (2015)

Drug Targets in breast cancer network

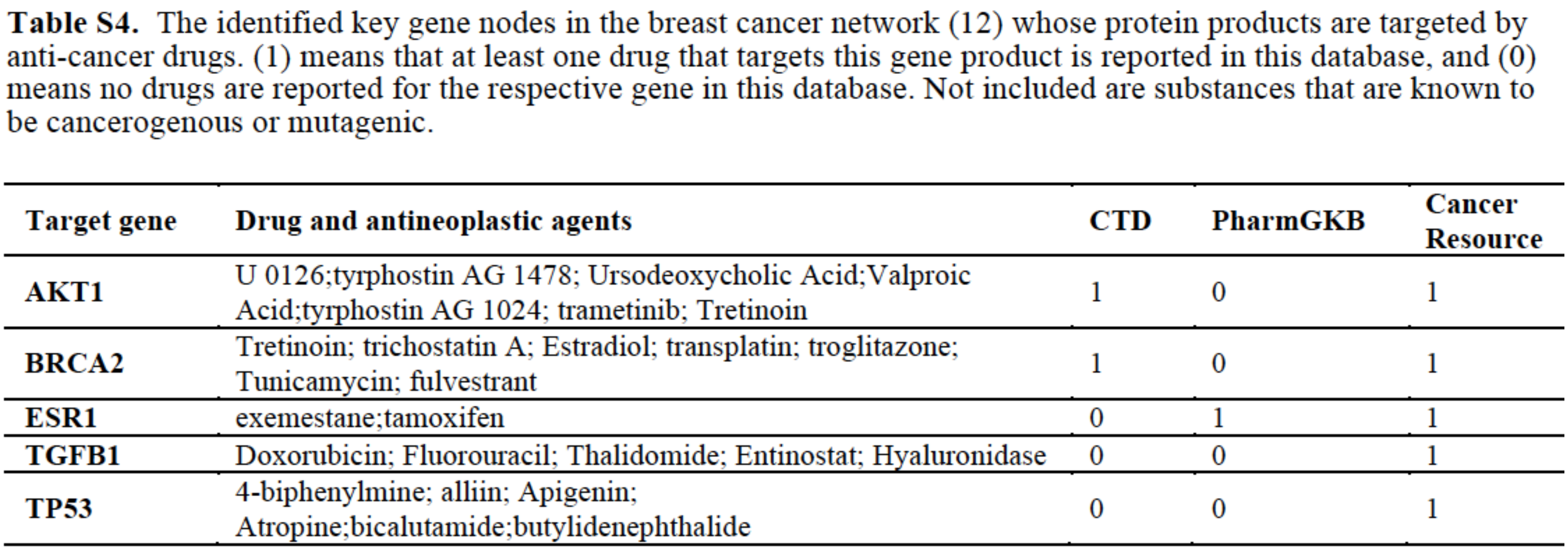


Table S4. The identified key gene nodes in the breast cancer network (12) whose protein products are targeted by anti-cancer drugs. (1) means that at least one drug that targets this gene product is reported in this database, and (0) means no drugs are reported for the respective gene in this database. Not included are substances that are known to be cancerogenous or mutagenic.

Target gene	Drug and antineoplastic agents	CTD	PharmGKB	Cancer Resource
AKT1	U 0126;tyrphostin AG 1478; Ursodeoxycholic Acid;Valproic Acid;tyrphostin AG 1024; trametinib; Tretinoin	1	0	1
BRCA2	Tretinoin; trichostatin A; Estradiol; transplatin; troglitazone; Tunicamycin; fulvestrant	1	0	1
ESR1	exemestane;tamoxifen	0	1	1
TGFB1	Doxorubicin; Fluorouracil; Thalidomide; Entinostat; Hyaluronidase	0	0	1
TP53	4-biphenylmine; alliin; Apigenin; Atropine;bicalutamide;butylidenephthalide	0	0	1

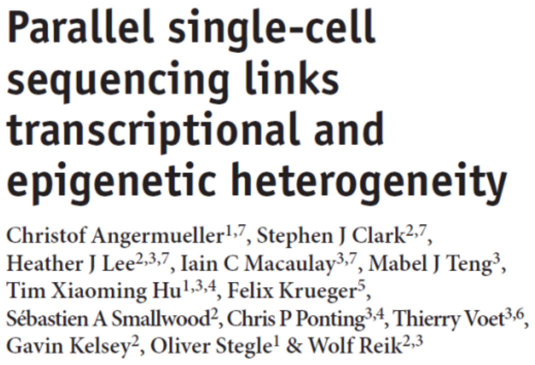
Some key genes are protein targets of known anti-cancer drugs,

→ relevance of key genes is validated

Hamed et al. *Nucl Acids Res* 43: W283-W288 (2015)

Case study: single-cell analytics

http://www.nature.com/nmeth/journal/v13/n3/full/nmeth.3728.html



Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity

Christof Angermueller[1,7], Stephen J Clark[2,7], Heather J Lee[2,3,7], Iain C Macaulay[3,7], Mabel J Teng[3], Tim Xiaoming Hu[1,3,4], Felix Krueger[5], Sébastien A Smallwood[2], Chris P Ponting[3,4], Thierry Voet[3,6], Gavin Kelsey[2], Oliver Stegle[1] & Wolf Reik[2,3]

In the presence of serum, mouse embryonic stem cells (ESCs) constitute a metastable population with stochastic switching between transcriptional states.

This **transcriptional heterogeneity** has been linked to the **differentiation potential** of ESCs. E.g. NANOG^{lo} cells have an increased propensity to differentiate and elevated expression of differentiation markers compared with NANOG^{hi} cells.

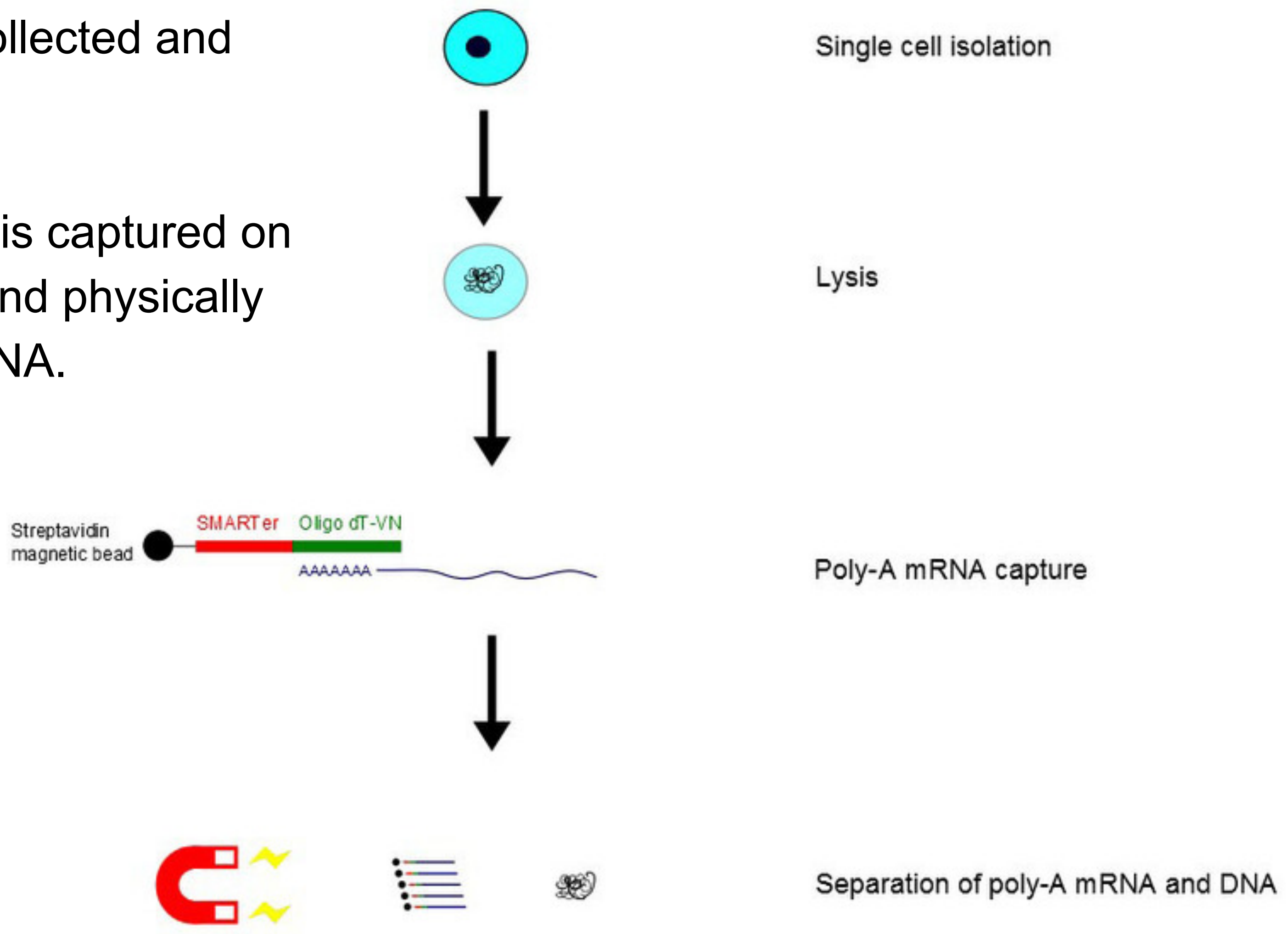
Sorted populations of cells show different levels of DNA methylation between transcriptional states, such as gains in DNA methylation in NANOG^{lo} and REX1/ZFP42^{lo} cells compared with, respectively, NANOG^{hi} and REX1^{hi} cells.

To investigate the link between epigenetic and transcriptional heterogeneity in ESCs, Reik et al. performed scM&T-seq on 76 individual serum ESCs.

scMT & T-seq protocol

Single cells are collected and lysed.

Then poly-A RNA is captured on magnetic beads and physically separated from DNA.

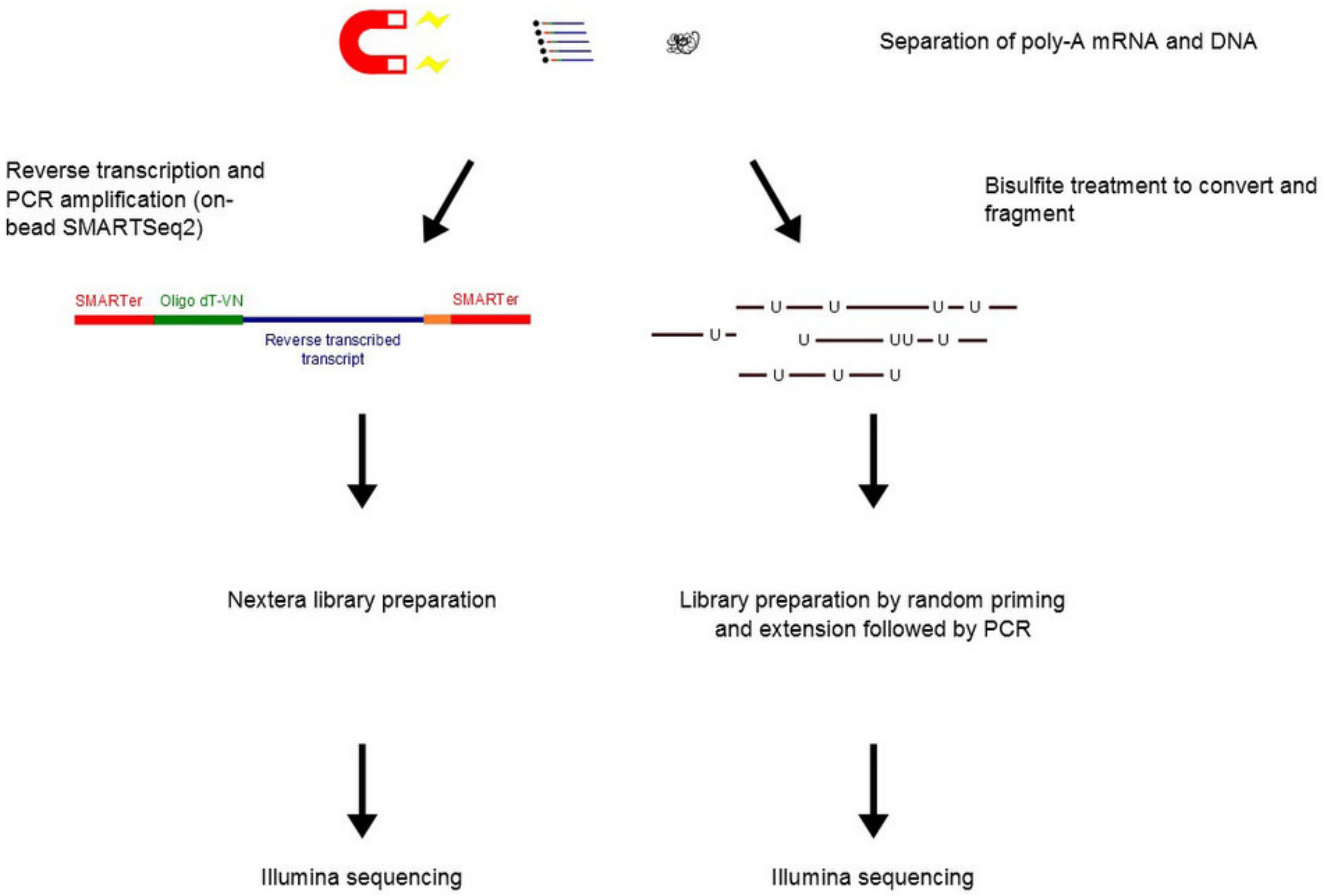


Angermüller et al.
Nature Methods (2016) 13, 229

scMT & T-seq protocol

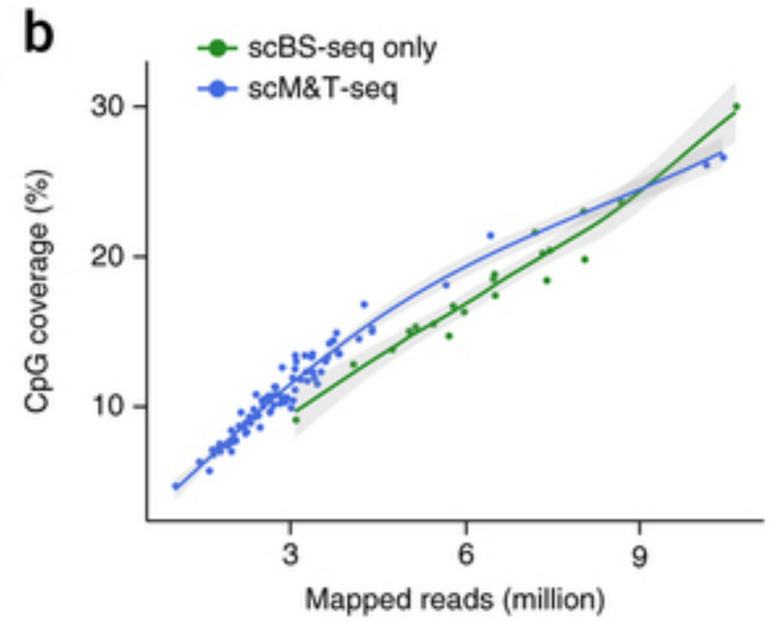
Amplified cDNA is generated from mRNA on beads.

DNA is bisulfite converted and Illumina sequencing libraries are prepared from both components in parallel.


Separation of poly-A mRNA and DNA
Reverse transcription and PCR amplification (on-bead SMARTSeq2)
Bisulfite treatment to convert and fragment
SMARTer
Oligo dT-VN
SMARTer
Reverse transcribed transcript
U
UU
Nextera library preparation
Library preparation by random priming and extension followed by PCR
Illumina sequencing
Illumina sequencing

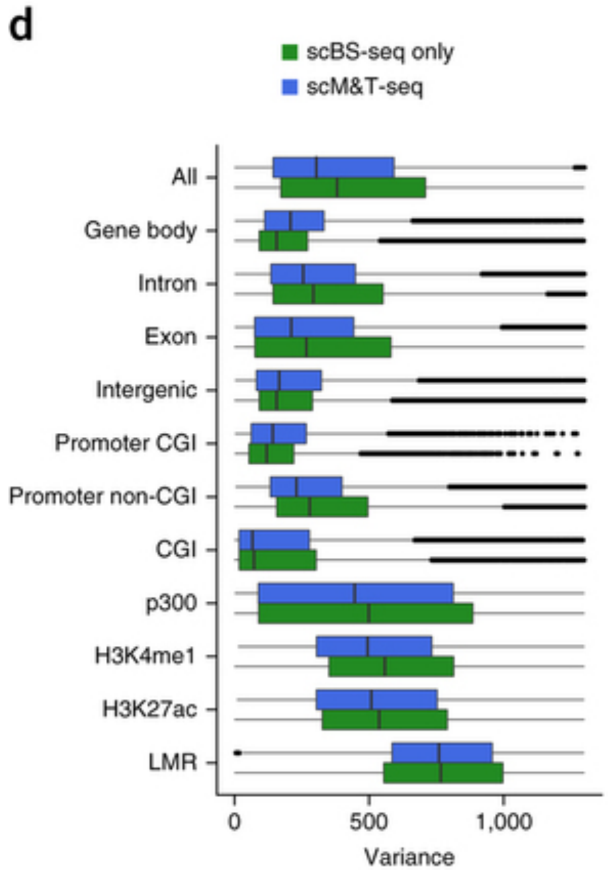
Angermüller et al.
Nature Methods (2016) 13, 229

How good is the protocol:
check against single cell bisulfite sequencing (scBS-seq)



Methylome coverage in scM&T-seq libraries was lower than that in scBS-seq libraries.

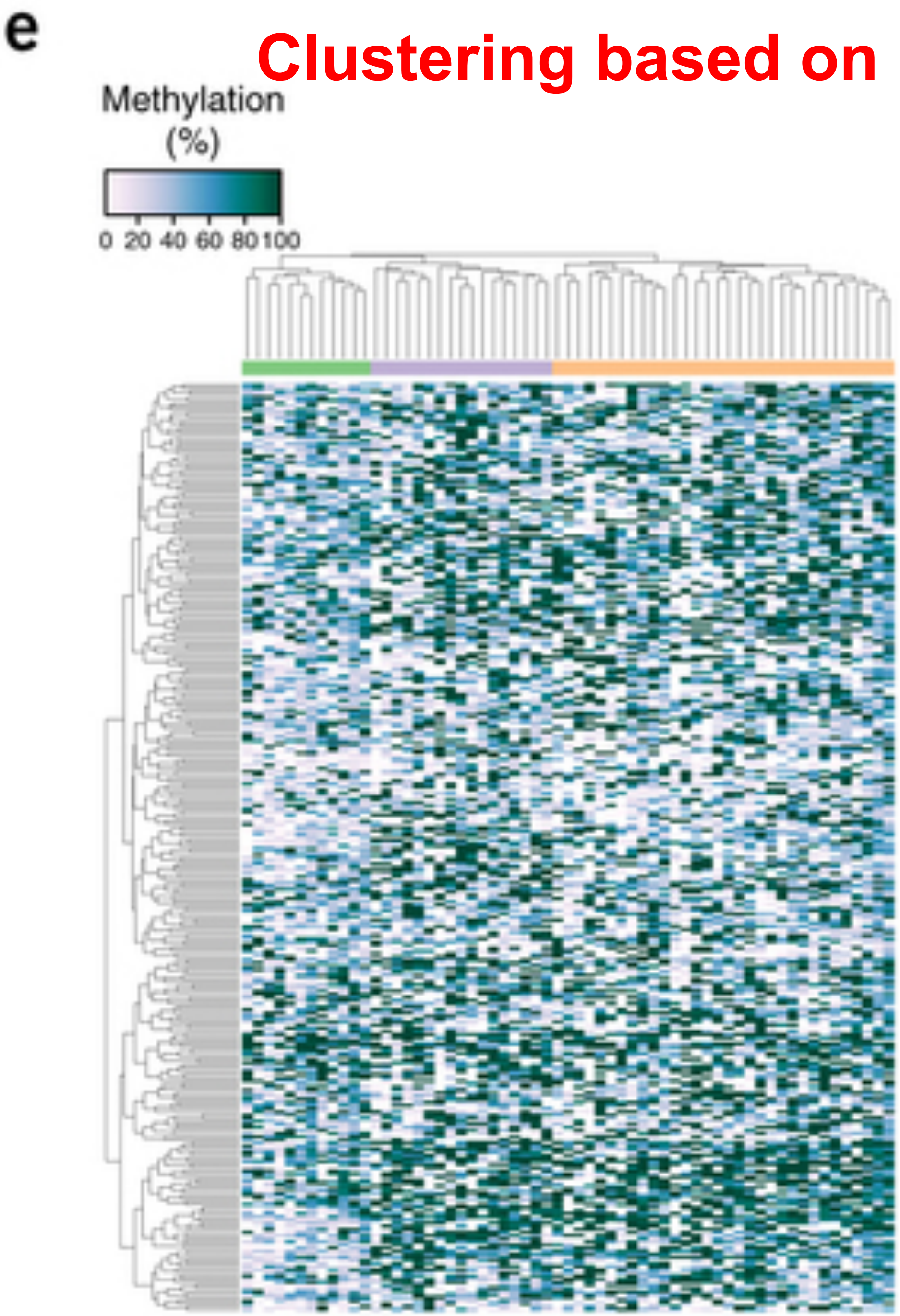
However, genome-wide CpG coverage at matched sequencing depth (c) and coverage of different regions (d) was consistent across protocols.



Angermüller et al.
Nature Methods (2016) 13, 229

V11

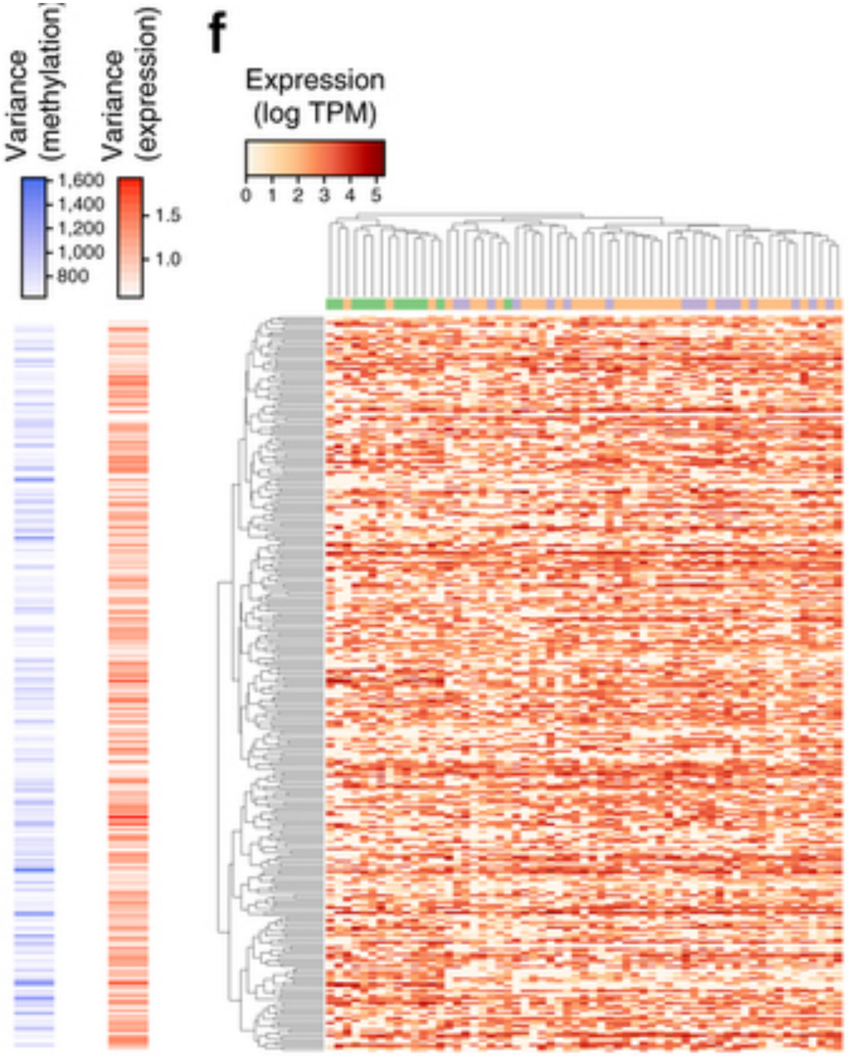
Clustering based on DNA methylation data



Shown is a hierarchical clustering analysis of gene-body methylation for the 300 most variable genes in terms of DNA methylation.

Angermüller et al.
Nature Methods (2016) 13, 229

Clustering based on expression data



Shown is a hierarchical clustering analysis of gene expression for the 300 most variable genes (on the basis of DNA-methylation variance)

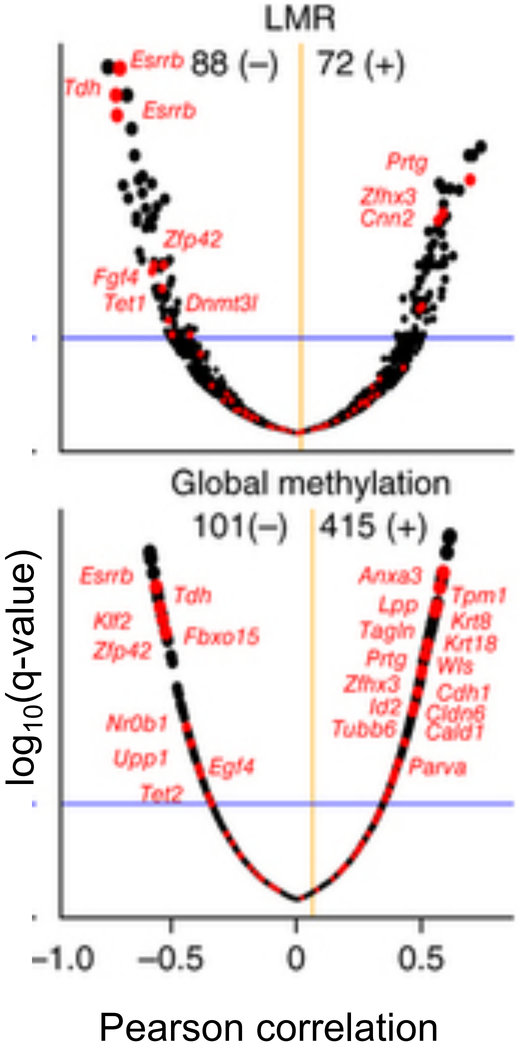
→ Both data yield distinct clustering of cells.

This suggests that global methylome and transcriptome profiles reveal complementary, but distinct, aspects of cell state.

This is also consistent with previous observations that the transcriptome and methylome are partially uncoupled in serum ESCs.

Angermüller et al.
Nature Methods (2016) 13, 229

Associations of expression and DNA-methylation variation



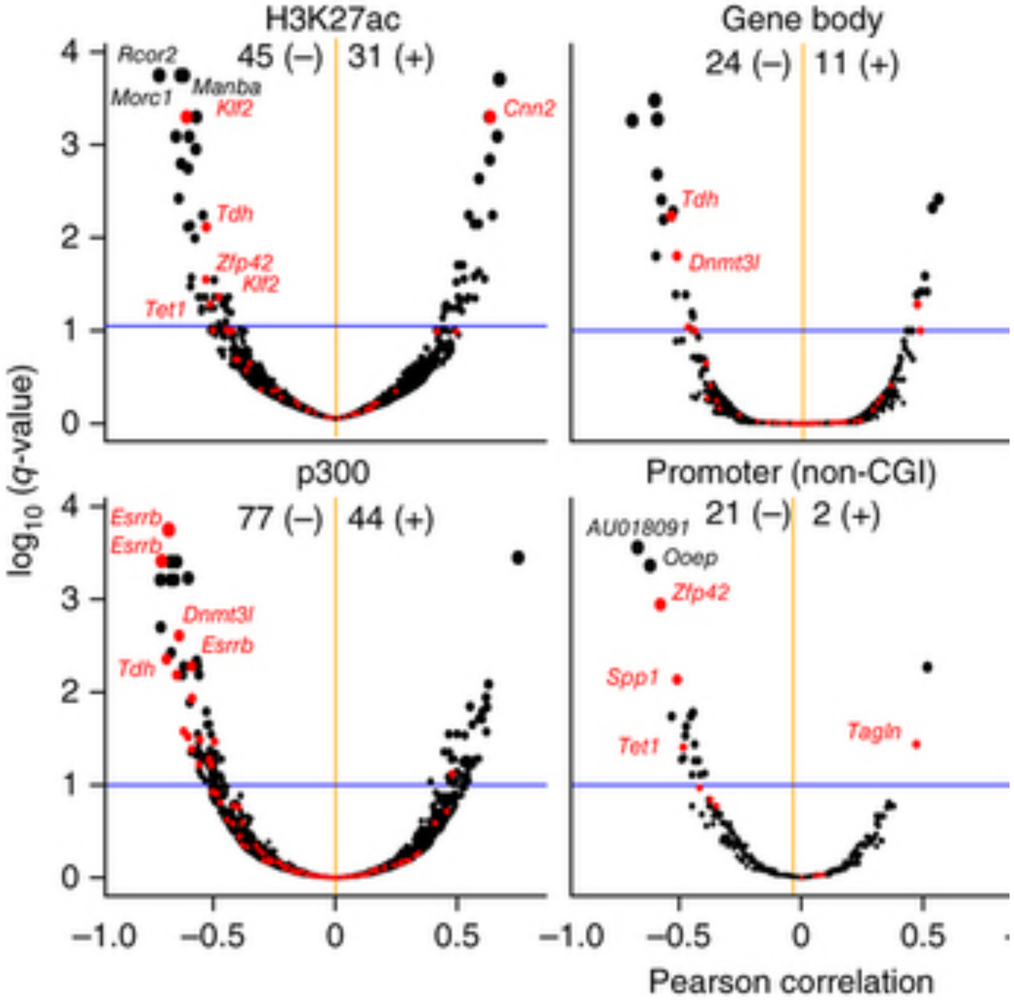
1,493 associations were found between the expression of individual genes and DNA-methylation variation in several genomic contexts (FDR) < 10%).

There exist **both positive and negative associations**, highlighting the complexity of interactions between the methylome and the transcriptome.

Also distal regulatory elements including low-methylation regions (LMRs) had a fair balance of positive and negative associations.

Angermüller et al.
Nature Methods (2016) 13, 229

Associations of expression and DNA-methylation variation



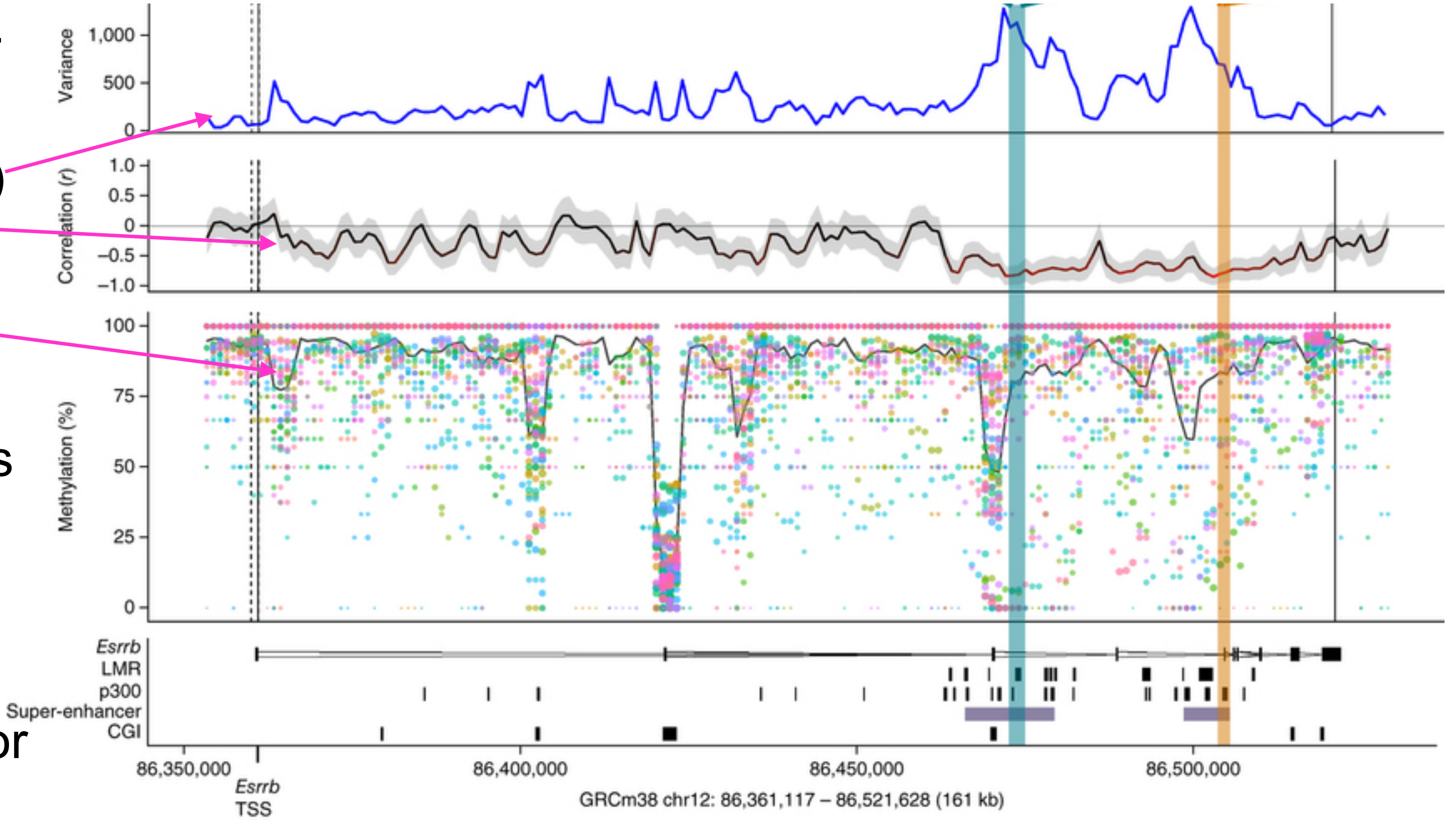
Negative correlations between DNA methylation and gene expression were predominant for non-CGI promoters.

Angermüller et al.
Nature Methods (2016) 13, 229

Zoomed-in analysis for *Esrrb*

Correlation of methylation and expression (black), variance (blue)

Solid black curve : weighted mean methylation rate across all cells



Estimated methylation rate of 3-kb windows for each cell

Dot size : CpG coverage,

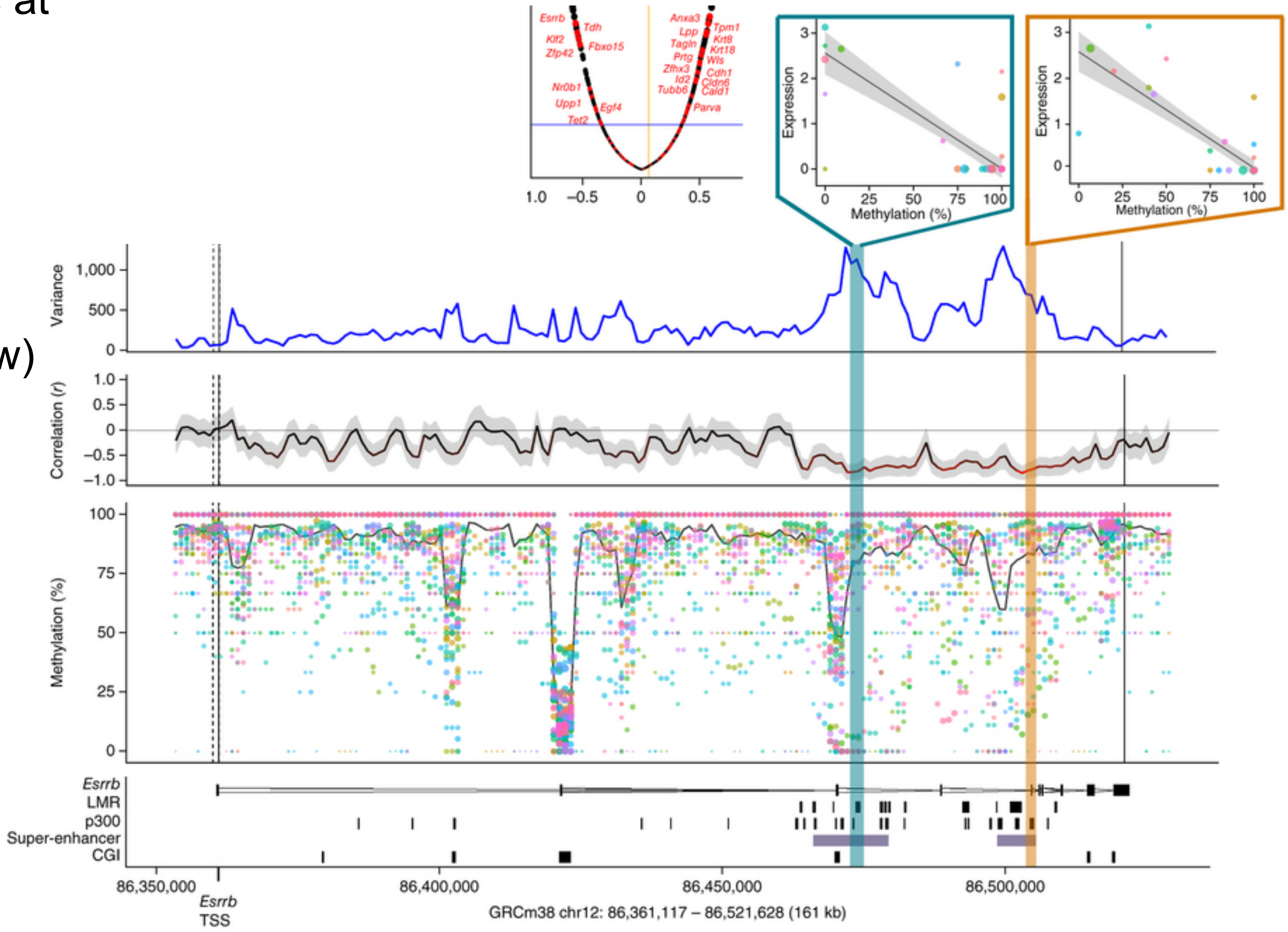
Dot colors correspond to single cells.

Esrrb is a known hub gene in pluripotency networks. Its expression negatively correlates with the methylation of several LMR and p300 sites overlapping 'superenhancers' in the genomic neighborhood.

Angermüller et al.
Nature Methods (2016) 13, 229

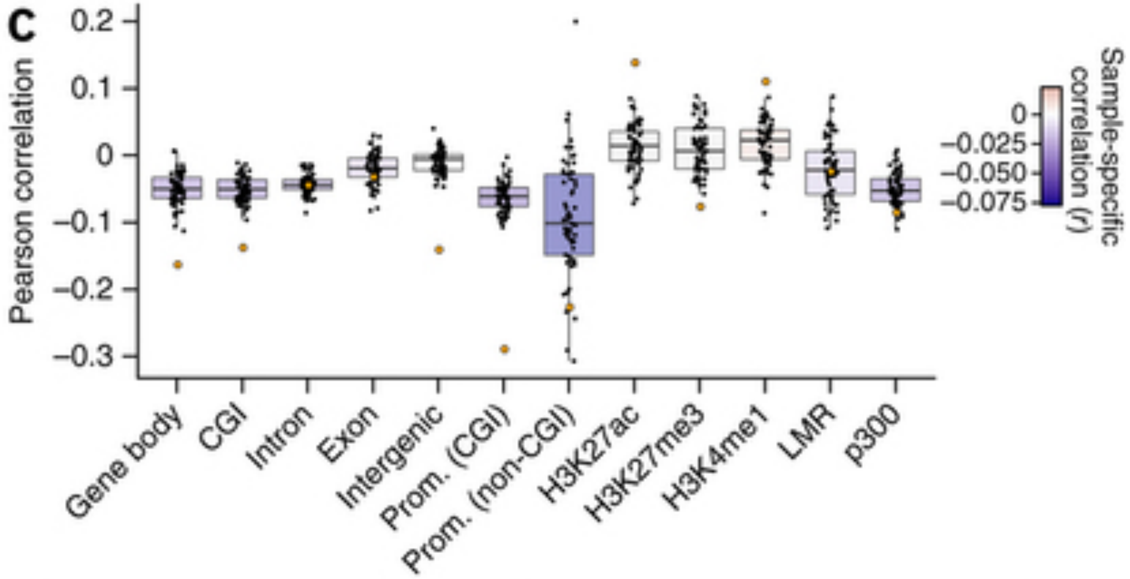
Zoomed-in analysis for *Esrrb*

Two scatter plots at the top right: association between DNA methylation at a p300 region (outlined in yellow) and at an LMR (outlined in blue) and *Esrrb* expression.



Angermüller et al.
Nature Methods (2016) 13, 229

Correlations of DNA methylation and expression



Gene-specific association analysis of correlations between DNA methylation in different genomic contexts and gene expression in individual cells.

Shown are methylation-expression correlations for all variable genes in single cells, for each annotation, with the correlation obtained from matched RNA-seq and BS-seq of a bulk cell population superimposed (orange circles).
Prom = promoter.

Angermüller et al.
Nature Methods (2016) 13, 229

Summary

- Multi-variate vs. single-variate analysis reveals possible confounding effects

- Multi-omics methods: graph-based and/or Bayesian methods for data integration

- Single cell analysis showed:

- non–CGI promoter methylation and transcription are negatively associated in single cells / both positive and negative associations at distal regulatory regions.

- expression levels of many pluripotency factors, such as *Esrrb* are negatively associated with DNA methylation → an important mechanistic component of fluctuating pluripotency in serum ESCs is epigenetic heterogeneity

- the strength of the connection between the methylome and the transcriptome can vary from cell to cell

- Q: is our understanding / data generation ready for multi-omics analysis?