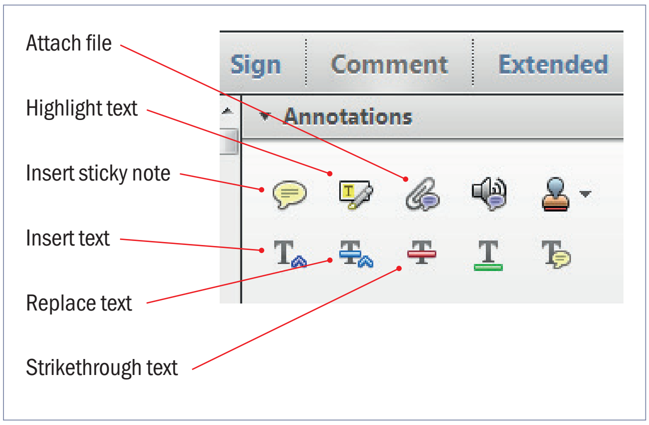
Making corrections to your proof

Please follow these instructions to mark changes or add notes to your proof. You can use Adobe Acrobat Reader (download the most recent version from **https://get.adobe.com**) or an open source PDF annotator.

For Adobe Reader, the tools you need to use are contained in **Annotations** in the **Comment** toolbar. You can also right-click on the text for several options. The most useful tools have been highlighted here. If you cannot make the desired change with the tools, please insert a sticky note describing the correction.

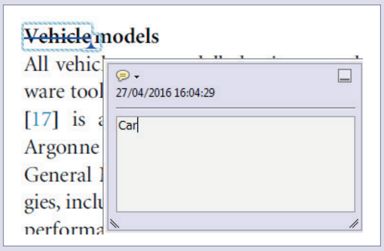
Please ensure all changes are visible via the 'Comments List' in the annotated PDF so that your corrections are not missed.



Do not attempt to directly edit the PDF file as changes will not be visible.

Replacing text

To replace text, highlight what you want to change then press the replace text icon, or right-click and press 'Add Note to Replace Text', then insert your text in the pop up box. Highlight the text and right click to style in bold, italic, superscript or subscript.



~~Vehicle~~ models

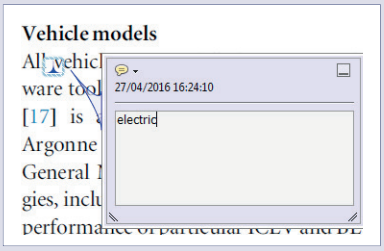
All vehic
ware tool
[17] is
Argonne
General
gies, inclu
performa

27/04/2016 16:04:29

Car

Inserting text

Place your cursor where you want to insert text, then press the insert text icon, or right-click and press 'Insert Text at Cursor', then insert your text in the pop up box. Highlight the text and right click to style in bold, italic, superscript or subscript.



Vehicle models

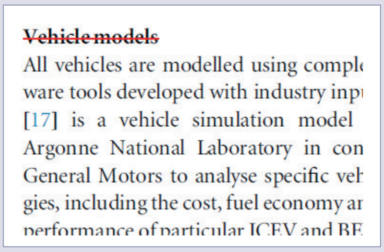
All vehic
ware tool
[17] is
Argonne
General
gies, inclu
performance of particular ICEV and BE

27/04/2016 16:24:10

electric

Deleting text

To delete text, highlight what you want to remove then press the strikethrough icon, or right-click and press 'Strikethrough Text'.

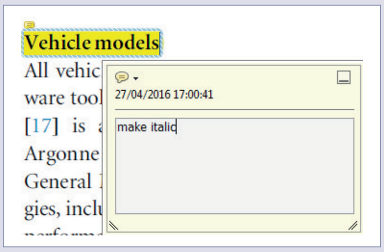


~~Vehicle models~~

All vehicles are modelled using compl
ware tools developed with industry inpu
[17] is a vehicle simulation model
Argonne National Laboratory in con
General Motors to analyse specific veh
gies, including the cost, fuel economy ar
performance of particular ICEV and BE

Highlighting text

To highlight text, with the cursor highlight the selected text then press the highlight text icon, or right-click and press 'Highlight text'. If you double click on this highlighted text you can add a comment.



Vehicle models

All vehic
ware tool
[17] is
Argonne
General
gies, inclu

27/04/2016 17:00:41

make italic

QUERIES

Page 1
AQ1
Please check that the **names of all authors as displayed in the proof are correct**, and that all **authors are linked to the correct affiliations**. Please also confirm that the correct corresponding author has been indicated. **Note that this is your last opportunity to review and amend this information before your article is published**.
Page 7
AQ2
Please check that the funding information below is correct for inclusion in the article metadata. FP7 Ideas: European Research Council: SIZEFFECTS.
Page 6
AQ3
Publisher location and name are required for book Font-Clos *et al* (2017). Please provide the missing information.
Page 7
AQ4
If an explicit acknowledgment of funding is required, please ensure that it is indicated in your article. If you already have an Acknowledgments section, please check that the information there is complete and correct.
AQ5
Please check the details for any journal references that do not have a link as they may contain some incorrect information. If any journal references do not have a link, please update with correct details and supply a Crossref DOI if available.
AQ6
Please provide the article title for all the references.
AQ7
Please update the publication details if appropriate in Muniandy *et al* (2017).

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

CrossMark

RECEIVED
12 January 2018

REVISED
14 March 2018

ACCEPTED FOR PUBLICATION
21 March 2018

PUBLISHED

AQ1

PAPER

Gene expression signature of obesity in monozygotic twins

Francesc Font-Clos[1], Stefano Zapperi[1,2,3] and Caterina A M La Porta[4]

[1] Department of Physics, Center for Complexity and Biosystems, Via Celoria 16, 20133 Milano, Italy
[2] CNR—Consiglio Nazionale delle Ricerche, Istituto di Chimica della Materia Condensata e di Tecnologie per l'Energia, Via R. Cozzi 53, 20125 Milano, Italy
[3] Department of Applied Physics, Aalto University, PO Box 11100, FIN-00076, Aalto, Finland
[4] Department of Environmental Science and Policy, Center for Complexity and Biosystems, University of Milano, via Celoria 26, 20133 Milano, Italy

E-mail: caterina.laporta@unimi.it

Keywords: obesity, gene expression, pathway deregulation score, twins

Abstract

Objective: Observational studies suggest that obesity might have a Mendelian origin, but it is not clear if gene expression patterns observed in obese subjects are secondary to genetic traits or not. *Approach*: Here we test a transcriptomic signature of obesity previously identified by our group on a large cohort of twin subjects (TwinsUK). *Main results*: The results show that the signature correlates strongly both with body mass index (BMI) and fat mass. Moreover, in paired transcriptomes of monozygotic twins, changes in signature correlate with changes in BMI and fat mass. We also identify a set of deregulated pathways involved in obesity, from inflammation to metabolism, and show that their pathway deregulation score is strongly correlated with BMI variations in pairs of identical twins. *Significance*: Taken together, our results strongly indicate that alterations in gene expression observed in obese subjects are not due to their genetic background, and should therefore primarily be associated with environment and lifestyle.

1. Introduction

Obesity has become a pandemic disease with an significant increase in children (Swinburn *et al* 2011), but the relevance of genetic background is still debated. Well-established cases of Mendelian forms of obesity approximately account for only 5% of the severely obese cases (Blakemore and Froguel 2010). In the case of common obesity, recent genome wide association studies (GWAS) have investigated possible relations between single nucleotide polymorphism (SNP) and body mass index (BMI) (Locke *et al* 2015). Despite the sheer amount of data and the effort devoted to this task, none of the resulting genetic loci have real predictive power. In particular, genetic contributions do not account for most BMI variations between subjects, which are thus likely to be due to lifestyle and environmental factors (Locke *et al* 2015). In a recent paper, an investigation of the gene expression profile in subcutaneous adipose tissue of BMI-discordant monozygotic twin pairs could not detect any molecular or clinical changes associated with subtypes of obesity (Muniandy *et al* 2017).

Here we tackle the fundamental question related to a possible involvement of the genetic background in the development of obesity by investigating if our gene expression signature of obesity recently identified has a Mendelian contribution (Font-Clos *et al* 2017). The genes strongly associated with obese subjects comprise genes involved in the interaction between cells and the extracellular matrix, inflammation and central nervous system (Font-Clos *et al* 2017). Moreover, this signature is able to capture the complexity of the pathology identifying features linked not only to inflammation and cancer but also to mood and reproductive disorders (Font-Clos *et al* 2017). This approach appears to be the best to capture a real snapshot of the obese subject and to identify underlying pathways that are usually impossible to find if few samples are studied. In this paper, we used the same framework described in Font-Clos *et al* (2017), analysing the gene expression data from a large cohort of twins (Buil *et al* 2015), including pairs of monozygotic twins. In particular, for this kind of study, where the samples available are few in number, the possibility to use an approach based on big data offers the advantage of reducing the noise by collecting and analysing a large set of data coming from different sources. However, since the dataset

Table 1. The 38 genes in the transcriptomic signature of obesity and their associated coefficients. Genes are ranked by the absolute value of their coefficient. Details of how these genes and coefficients were computed can be found in Font-Clos *et al* (2017).

Rank	Entrez ID	Gene symbol	Coefficient	Rank	Entrez ID	Gene symbol	Coefficient
1	1278	COL1A2	0.131	20	7045	TGFBI	0.0569
2	80763	SPX	−0.126	21	25878	MXRA5	0.0558
3	761	CA3	−0.0889	22	2982	GUCY1A3	0.0556
4	219 348	PLAC9	0.0742	23	2335	FN1	0.0555
5	25975	EGFL6	0.0731	24	7076	TIMP1	0.0553
6	2014	EMP3	0.0701	25	5396	PRRX1	0.0548
7	6696	SPP1	0.0690	26	4069	LYZ	0.0529
8	1397	CRIP2	0.0679	27	8076	MFAP5	0.0510
9	1490	CTGF	0.0674	28	3512	JCHAIN	0.0486
10	22822	PHLDA1	0.0667	29	10402	ST3GAL6	−0.0466
11	1880	GPR183	0.0659	30	3429	IFI27	0.0458
12	171 024	SYNPO2	0.0655	31	83442	SH3BGRL3	0.0457
13	1520	CTSS	0.0646	32	712	C1QA	0.0442
14	80114	BICC1	0.0638	33	474 344	GIMAP6	0.0441
15	115 207	KCTD12	0.0622	34	9457	FHL5	0.0438
16	151 887	CCDC80	0.0599	35	8470	SORBS2	0.0437
17	22918	CD93	0.0591	36	7037	TFRC	0.0431
18	389 136	VGLL3	0.0588	37	1291	COL6A1	0.0430
19	8542	APOL1	0.0581	38	57863	CADM3	0.0429

is nonhomogeneous the first well-known problem to solve is the batch effect. In fact, in the worse case, if batch effects are not detected and removed in the right way, they can lead to flawed results. According to Font-Clos *et al* (2017), we removed the batch effect by singular value decomposition, then we reduced the dimensionality using a pathway deregulation score; finally, we ranked the pathways more differentially expressed between paired twins with different BMI.

Our approach offers the possibility to use a pipeline to study biological problems tackling the complexity and interactions between organs and tissues. Interdisciplinary studies and the rapid development of complex network theory are the foundation of new and promising disciplines, such as network physiology and network medicine. Both these disciplines try to integrate and introduce new concepts and methods coming from modern statistical physics and network theory to biology and medicine, as discussed recently by Ivanov *et al* (2016). Our results cast serious doubt on the importance of a genetic background of gene expression patterns in obesity, while the role of the environment and lifestyle appears particularly critical. This result would have been impossible to achieve with a traditional approach, so that our findings show a general method that can be used to implement the principles of network medicine and physiology to a variety of cases.

2. Methods

2.1. Transcriptomic data

Transcriptomic data was obtained from TwinsUK (www.twinsuk.ac.uk/, (Buil *et al* 2015)). Samples without BMI information or with fat mass below 1% were discarded from the analysis, as well as those without a matching co-twin sample. A total of 626 paired samples were analyzed, including 256 samples from MZ twins. Expression values are given in terms of log2(RPKM + 1). No further normalization was applied. We used MyGene.Info API (Wu *et al* 2013, Xin *et al* 2016) via the MyGene.py Python wrapper to convert between gene symbol and Entrez gene names.

2.2. Obesity score

The obesity score S_j for sample j is calculated as a linear combination of the log2 expression of 38 genes:

$$S_j \equiv \sum_{r=1}^{n} \alpha_r X_{jr} \tag{1}$$

where α_r is the coefficient of the rank r gene in table 1 and X_{jr} is its log2 expression in sample j. The set of 38 genes and their coefficients shown in table 1 were determined in Font-Clos *et al* (2017) using a dataset unrelated to the one analyzed in this manuscript.

Table 2. KEGG pathways with PDS highly correlated with BMI. The top nine pathways that have the highest correlation of the PDS with BMI, as measured by Pearson's *R* coefficient, and as shown in figures 3 and 4. The reported number of genes includes only genes found in the TwinsUK (Buil *et al* 2015) dataset. *p*-values are corrected for multiple testing; see Methods for details.

		PDS × BMI		ΔPDS × ΔBMI	
KEGG pathway	#Genes	Pearson *R*	*p*-value	Pearson *R*	*p*-value
Tyrosine metabolism	33	0.64	7.0×10^{-71}	0.48	4.1×10^{-6}
Arginine biosynthesis	21	0.61	1.2×10^{-61}	0.53	1.8×10^{-7}
Glyoxylate and dicarboxylate metabolism	28	0.60	1.9×10^{-58}	0.47	7.3×10^{-6}
Alanine, aspartate and glutamate metabolism	35	0.59	5.6×10^{-57}	0.39	4.0×10^{-4}
Vitamin digestion and absorption	24	0.59	9.8×10^{-57}	0.53	1.8×10^{-7}
Glycine, serine and threonine metabolism	39	0.58	6.6×10^{-54}	0.46	1.1×10^{-5}
Phenylalanine metabolism	16	0.57	1.3×10^{-53}	0.43	5.4×10^{-5}
Antifolate resistance	30	0.57	5.5×10^{-53}	0.42	8.7×10^{-5}
Porphyrin and chlorophyll metabolism	41	0.55	3.1×10^{-49}	0.47	5.2×10^{-6}

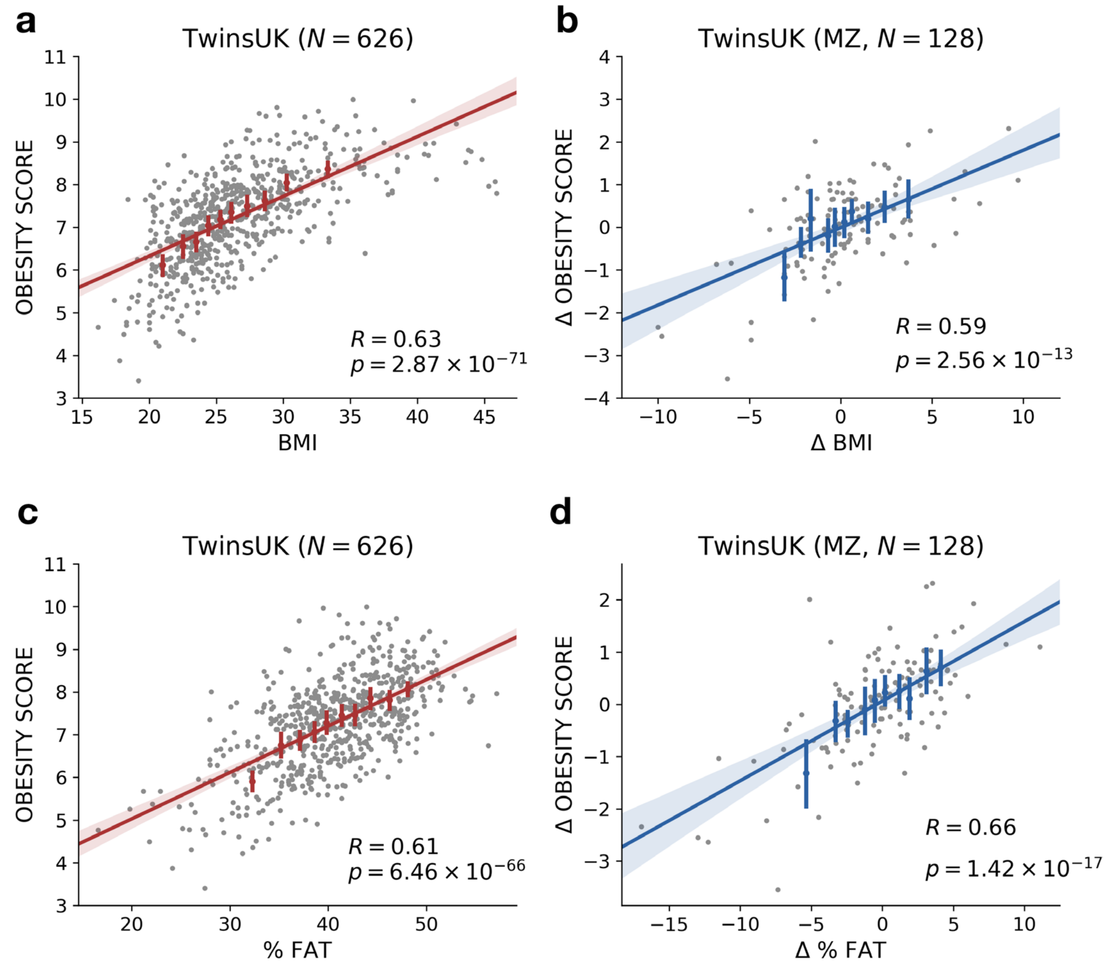


Figure 1. The obesity score correlates with BMI and fat mass. (a) and (c): Scatter plots (gray dots) and linear regressions (colored lines) between the obesity score and BMI or fat mass. Each point represents a sample from a single twin. (b) and (d): Scatter plots (gray dots) and linear regressions (colored lines) between the change in obesity score and change in BMI or change in fat mass. Each point represents a MZ twin pair. Shaded regions show 95% confidence intervals for the regression line, computed by bootstrapping. Colored dots are obtained binning the data into evenly sized bins and taking averages. The 95% confidence intervals of such averages are shown as colored vertical lines.

2.3. Statistical analysis

We used Python for all data processing and statistical analysis. Correlation coefficients and associated *p*-values were computed using the `scipy.stats.pearsonr` function. The linear regressions in figures 1–3 were computed using the `seaborn.regplot` function. *p*-values in table 2 were corrected for multiple testing, using the whole set of 626 tests performed, one for each pathway. We used a Benjamini–Yekutieli correction as the correlation structure of the pathways set is not known.

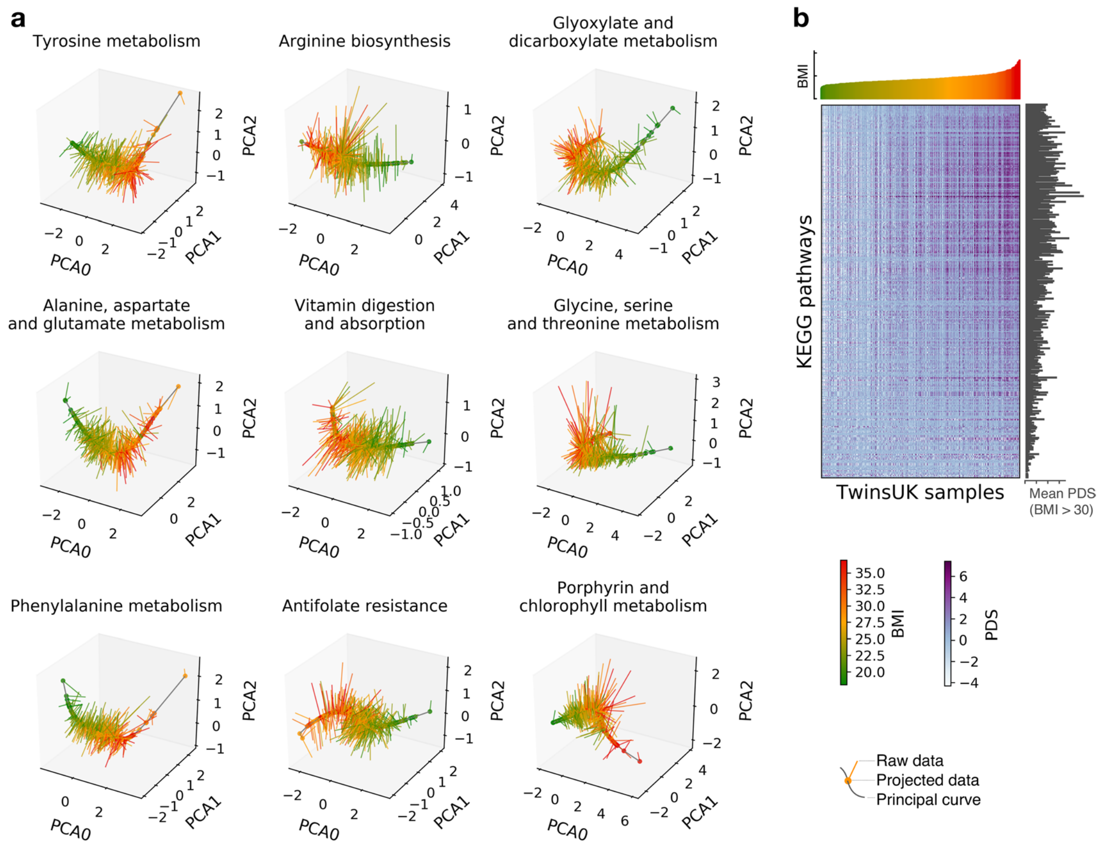


Figure 2. Pathway deregulation scores for all KEGG pathways. (a) The top nine pathways with the highest correlation of PDS with BMI. The plots show the first three PCA components of the raw data, and its projection onto the principal curve. Projection lines are colored according to BMI. (b) Heatmap of PDSs for all KEGG pathways. Rows represent pathways and are sorted by their correlation with BMI. Columns represent samples and are sorted by BMI, as illustrated by the filled colored curve on top. The heatmap coloring represents PDSs. Horizontal bars in the right-most area indicate the average PDS among obese subjects (BMI $>$ 30).

2.4. Pathway deregulation scores

Pathway deregulation scores (PDSs) were first introduced by Drier *et al* (2013) as a tool to quantify the deregulation of each pathway with respect to a reference sample. They are computed by fitting a non-parametric, non-linear one-dimensional curve through the 'middle' of the transcriptomic data, in the subspace generated by the genes of that pathway, usually through the *principal curve* algorithm (Hastie and Stuetzle 1989). We follow the algorithm presented in Drier *et al* (2013) with a small modification introduced in Font-Clos *et al* (2017): the value of 0 is placed at the mean value of the reference sample, instead of at the extremal point of the curve.

3. Results

3.1. The transcriptomic signature of obesity

In a recent publication (Font-Clos *et al* 2017) we found a robust transcriptomic signature (5σ) of obesity composed of 38 genes. Here we give a brief overview of that work, inviting the interested reader to see Font-Clos *et al* (2017) for further details. Our analysis revolved around *SVDmerge* (https://github.com/ComplexityBiosystems/SVDmerge), an algorithm to remove batch effects, and pathway deregulation scores (PDSs), a pathway-based dimensionality reduction technique. Combining these two methodologies allowed us to (i) merge several publicly available datasets, increasing the number of samples in the analysis, and (ii) transition from a gene-based to a pathway-based perspective, decreasing the number of variables from ∼20 000 genes to ∼1000 pathways. In this way we substantially improved the samples-to-variables ratio and were able to identify pathways related to adhesion molecules, inflammation, salivary secretion and digestive problems. We also proposed a simple obesity score, computed as a linear combination of the expression of the 38 genes, and showed that it correlates well with BMI in several independent validation datasets. We verified that such correlations are gender independent and tissue specific. Finally, we pointed out that some of the deregulation patterns found in obesity are also seen in breast tumor samples.

It is interesting to compare our transcriptomic signature with existing results on obesity based on GWAS (Locke *et al* 2015). These studies have revealed a set of genetic loci that are associated with BMI variations. We

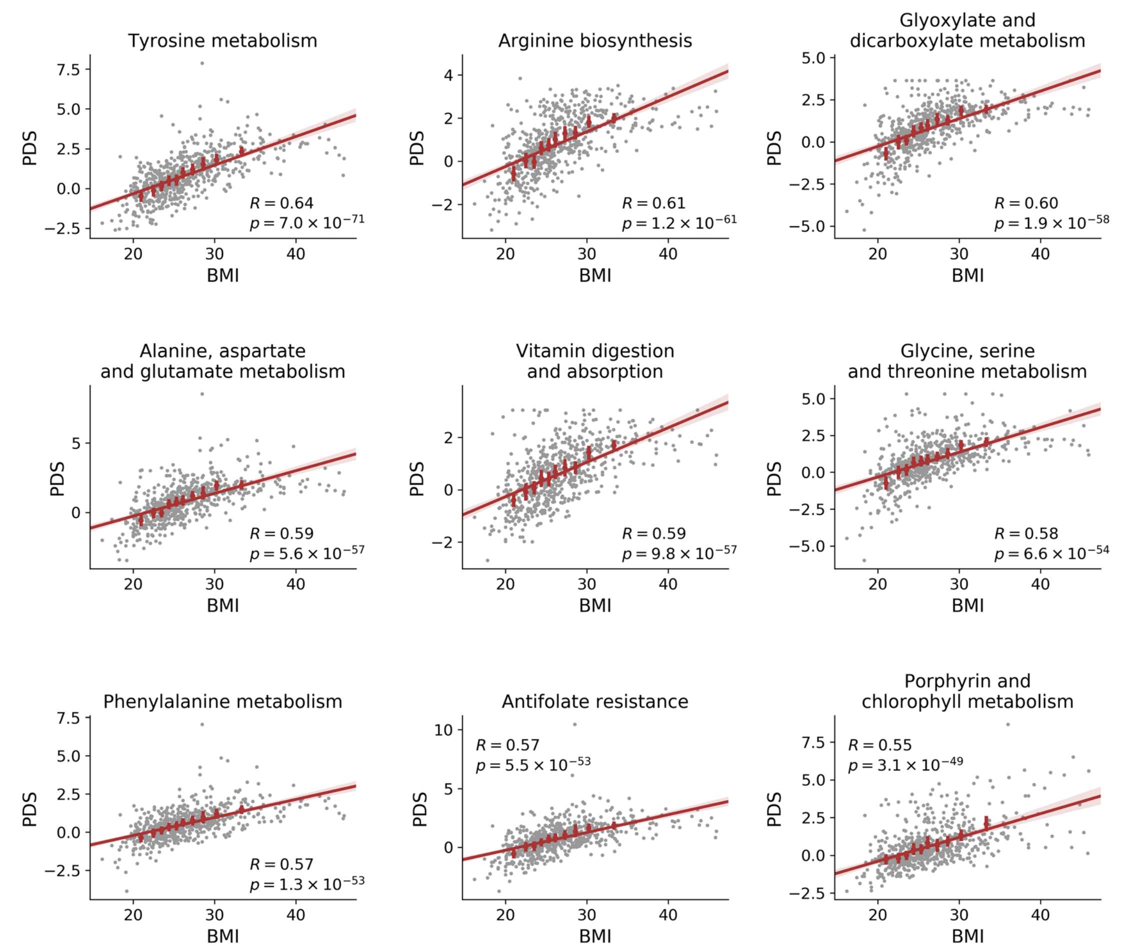


Figure 3. PDS correlate with BMI. Scatter plots (Gray dots) and linear regressions (colored lines) between PDS and BMI, for the top 9 pathways that have the highest correlation of the PDS with BMI. Each dot represents a single sample. Shaded regions show 95% confidence intervals for the regression line, computed by bootstrapping. Colored dots are obtained binning the data into evenly-sized bins and taking averages. 95% confidence intervals of such averages are shown as colored vertical lines.

have compared the list of genes in our signature with the list of genes reported in Locke *et al* (2015) as significantly associated with BMI. The two lists have no intersection. Similarly, the list of significant pathways revealed in Locke *et al* (2015) has no intersection with the list reported in Font-Clos *et al* (2017). Therefore, our approach allows to identify genes that normally are not highlighted because we are able to analyze more datasets due to the removal of the batch effect using *SVDmerge* (https://github.com/ComplexityBiosystems/SVDmerge) (Font-Clos *et al* 2017). The power of big-data analysis is actually to uncover things that are not easy to see, in this case genes and pathways at the roots of the problem.

4. Transcriptomic signature correlates with obesity

We applied the strategy described in the previous section to study transcriptomic data from a large cohort of monozygotic (MZ) pair twins (256 samples) and from a set of heterozygous twins (370 samples); see Methods for details. Figure 1(a) shows that the obesity score correlates with BMI ($R = 0.63, p = 2.87 \times 10^{-71}$) considering all the 626 samples of the batch. The TwinsUK dataset is particularly interesting because it contains samples from 128 MZ twin pairs whose BMI can be discordant. Because MZ twins are genetically identical, BMI variations between a subject and its co-twin should be due exclusively to environmental factors and lifestyle. Figure 1(b) shows indeed that the variations in BMI correlate strongly with variations in score ($R = 0.59, p = 2.55 \times 10^{-13}$) when considering only pairing between co-twins. Hence, the signature in co-twins reflects merely the BMI, rather than the genetic background that should be identical in co-twins and different in randomly paired subjects. This suggests that our transcriptomic signature is associated with obesity rather than any underlying genetic differences in the subjects. To corroborate this finding, we considered also the percentage of fat mass and show that it correlates again very strongly with the obesity score considering all 626 samples in TwinsUK ($R = 0.61, p = 6.46 \times 10^{-66}$; figure 1(c)). Furthermore, changes in fat mass between siblings in MZ twin pairs correlate strongly with changes in score ($R = 0.66, p = 1.42 \times 10^{-17}$; figure 1(d)).

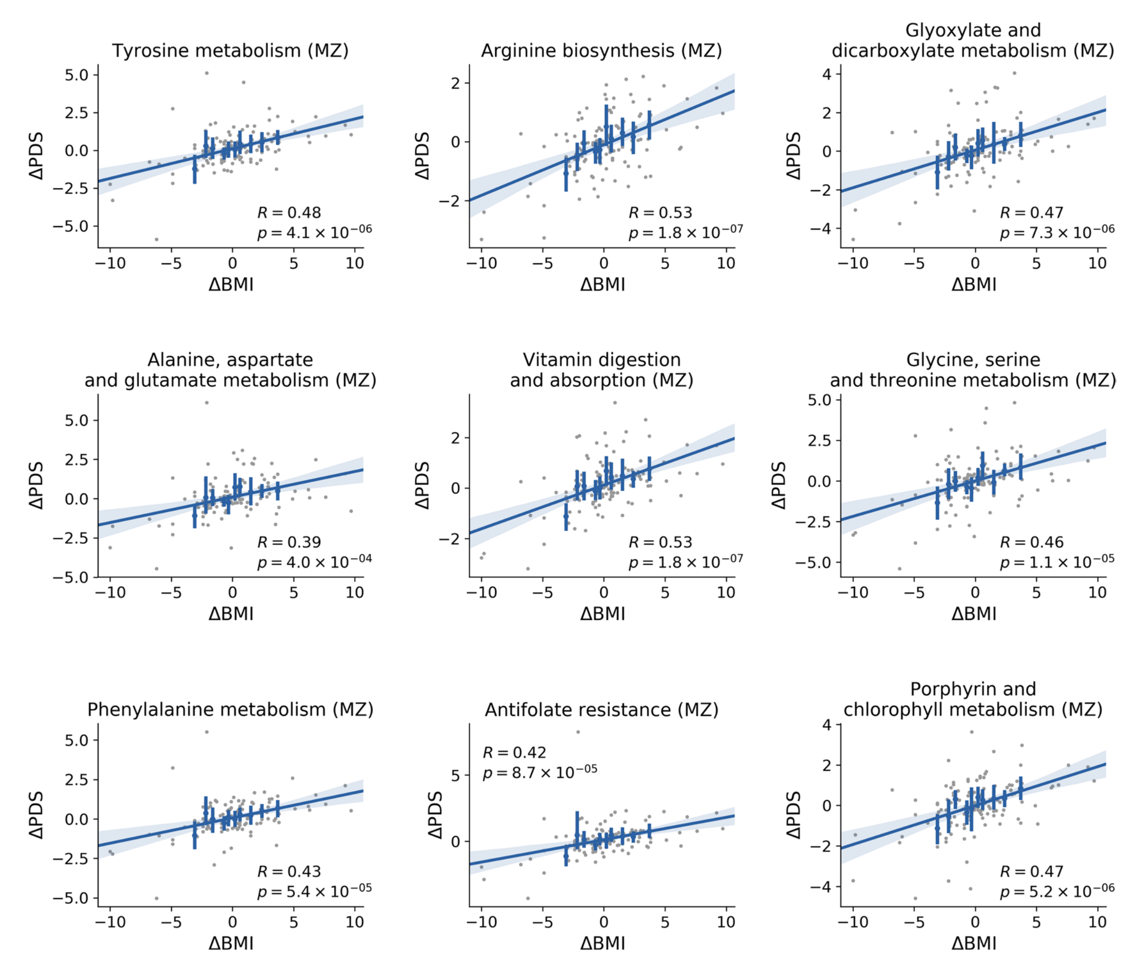


Figure 4. Changes in PDS correlate with changes in BMI in paired MZ twins samples Scatter plots (gray dots) and linear regressions (colored lines) between changes in PDS and changes in BMI, for the top nine pathways that have the highest correlation of the PDS with BMI. Each dot represents a MZ twin pair. Shaded regions show 95% confidence intervals for the regression line, computed by bootstrapping. Colored dots are obtained binning the data into evenly sized bins and taking averages. 95% confidence intervals of such averages are shown as colored vertical lines.

4.1. Pathway deregulation in obesity

To understand which pathways are most affected by obesity, we computed PDSs (see Methods (Drier *et al* 2013)) for all the samples in the TwinsUK database, among all KEGG pathways; see figure 2(b). We then computed the correlation between PDS and BMI and sorted the pathways accordingly. Figure 2(a) shows a 3-component PCA view of the raw data and its projection onto the principal curves defining the PDSs for the top nine pathways reported in table 2. The most significant pathways are all related with metabolic activities and play a clear role in metabolic misfunction. It is remarkable that samples tend to cluster by BMI, forming a colored, elongated cloud from green (lean) to orange (overweight) to red (obese). An alternative representation is given in figure 3, where we show scatter plots and linear regressions between PDS and BMI for these same pathways. *R* coefficients and *p*-values are given in table 2. Finally, we restricted the scope to paired MZ twins and inspected the relation between changes in BMI and changes in PDS. Figure 4 shows scatter plots and linear regressions for these nine pathways. Notice that each point represents an MZ twin couple, so that changes in BMI/PDS are always computed between subjects with identical genetic material.

5. Discussion

AQ3 Rare genetic mutations in the leptin gene and elsewhere in the genome can cause extreme obesity (Ahima 2008), but the importance of genetics with respect to epigenetic and environmental factors in the current obesity pandemia is still debated. An approach that tries to combine and analyze all the available transcriptomes AQ4 published in the public repositories has the clear advantage of having more data, making it easier to discriminate the real signal from the noise. This is the same approach used to analyze collective data on Google or to follow the connections between people, or for fish schools or birds. The problem in biology is that the amount of data is not so big and therefore the noise could be relevant. We previously resolved the problem of batch effects due to the fact that nonhomogeneous data are put together in a recent paper, wherein we analyzed the transcriptomes of obese and lean subjects (Font-Clos *et al* 2017). Thanks to this methodological approach, we could report a robust signature of 38 genes that is able to identify all the complex features of obesity, from inflammation to cancer to mood and reproductive disorders (Font-Clos *et al* 2017).

Here, we used the same robust signature to analyze a large cohort of heterozygotic twins (370 subjects) with respect to homozygotic pair twins (256 subjects) in dependence of BMI and fat mass (TwinsUK database (Buil *et al* 2015)). The analysis of twin pairs with different BMI offers, of course, a very good opportunity to shed some light in the debate of the role of genetic background in obesity. The most important barrier to overcome in order to answer this simple question is the number of subjects analyzed for each study and therefore the possibility to find a significant signal out of the noise. For example, a recent paper reported the gene expression profile in 23 subcutaneous adipose tissue samples of BMI-discordant monozygotic twin Finnish pairs without finding any molecular or clinical changes associated with subtypes of obesity (Muniandy *et al* 2017). Since we have identified a robust signature of the obesity phenotype using a big data approach in our previous paper (Font-Clos *et al* 2017), we used this signature to analyze a much larger cohort of twin pairs (TwinsUK (Buil *et al* 2015)), including twins with the same genetic background. Our results clearly show that in these subjects obesity is correlated with a 38-gene transcriptomic signature in a BMI-dependent manner. Therefore, our results highlight the important role of the environment instead of the genetic background. A direct consequence is that since obesity is linked to issues of behavior and lifestyle, the only way to fight this disease is to return to these aspects. The other consequence is that since obesity is not due to the 'bad luck' of the subject due to the hereditary of unlucky genes, each subject can reverse his/her condition.

In light of our results, we need to study obesity in a broader context where many external and internal factors cooperate. The broad patterns of deregulated pathways observed in obese subjects provide a striking indication of the interconnected and multi-scale nature of human physiology. Ideas and tools coming from the emerging field of network physiology and network medicine, as recently outlined (Ivanov *et al* 2016), could thus contribute to build a new perspective to tackle the obesity pandemic. AQ2

Acknowledgments

FFC and SZ are supported by the ERC advanced grant SIZEFFECTS. This work was possible thanks to access to TwinsUK data under project E879: Transcriptomic signature of obesity. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guys and St Thomas NHS Foundation Trust in partnership with Kings College London.

Author contributions

FFC analyzed data. SZ and CAMLP designed the research and wrote the paper with the assistance of FFC.

Additional information

Competing financial interests The authors declare no competing financial interests.

AQ5 References

AQ6 Ahima R S 2008 *J. Clin. Invest.* **118** 2380–3
Blakemore A I F and Froguel P 2010 *Ann. New York Acad. Sci.* **1214** 180–9
Buil A *et al* 2015 *Nat. Genet.* **47** 88–91
Drier Y, Sheffer M and Domany E 2013 *Proc. Natl Acad. Sci.* **110** 6388–93
Font-Clos F, Zapperi S and La Porta C A M 2017 *NPJ Systems Biology and Applications* vol 3 (https://doi.org/10.1038/s41540-017-0018-z)
Hastie T and Stuetzle W 1989 *J. Am. Stat. Assoc.* **84** 502–16
Ivanov P C, Liu K K L and Bartsch R P 2016 *New J. Phys.* **18** 100201
Locke A E *et al* 2015 *Nature* **518** 197–206
Muniandy M, Heinonen S, Yki-Järvinen H, Hakkarainen A, Lundbom J, Lundbom N, Kaprio J, Rissanen A, Ollikainen M and
AQ7 Pietiläinen K H 2017 *Int. J. Obesity*
Swinburn B A, Sacks G, Hall K D, McPherson K, Finegood D T, Moodie M L and Gortmaker S L 2011 *Lancet* **378** 804–14
Wu M, Frieboes H B, McDougall S R, Chaplain M A, Cristini V and Lowengrub J 2013 *J. Theor. Biol.* **320** 131–51
Xin J *et al* 2016 *Genome Biol.* **17** 91