Response to the comments made by Examiner "1136"

A reasonably comprehensive literature review was provided, however of concern, over 8.5 pages of the introduction is effectively a regurgitation of the 'Hallmarks of cancer' review by Weinberg and Hanahan. This could be substantially cropped.

Given this thesis is focused on a known environmental influence on cancer (obesity), it was a surprise the section on environmental factors associated with cancer incidence (1.2.3) is underdeveloped and superficial.

I understand the Examiner's concern about the relatively brief description of the known environmental factors that affect cancer incidence. As you may know, the focus of this project was the relationship between cancer and obesity, and how obesity affects cancer (which is covered more in depth in Section 1.3.2).

Section 1.2 was written as a section introducing cancer in general, and was not specifically focused on any single cancer type, nor environmental factor. So it follows that Section 1.2.3 describes, in very general terms, how *some* (not *all*) environmental factors can influence the cancer incidence, and I think the amount of information placed in this subsection is sufficient for this purpose. Additionally, the environmental factors chosen to be included in this section was based on either (1) (very) strong evidence of their role in cancer onset/progression in the literature (in terms of both epidemiological and biological evidence/association), or (2) some sort of relationship with obesity.

Perhaps I could emphasise these points within Section 1.2.3 and guide the readers to Section 1.3.2 for specific details on the effect of obesity on cancer.

In Section 1.2.4, Provide a citation for a recent list of molecular targets in cancer therapy, and their companion approved targeted agents (or present it in a table in appendix). This first paragraph is superficial and naïve. What is the rationale for only mentioning the RAS/RAF/MEK/ERK and PI3K/AKT pathways? Is it because these are the most relevant to those impacted by obesity? It would follow that those pathways be more in alignment with inflammatory signalling pathways rather than those associated with cell proliferation (as mentioned later in section 1.3.2). The rationale for cherry *picking select cell signalling information for this section of the introduction is unclear.*

What is the rationale for selection of one TCGA cancer dataset over another? Not all cancers are known to have obesity as a risk factor (as mentioned at the beginning of this report). It would therefore follow that not all cancers would be expected to have a signature consistent with a high BMI score metagene. A more targeted approach would perhaps have been better to rationalise which datasets to examine based on known risk of association with obesity, then look within any of those relevant cancer types for datasets for which height and weight data were available. This may have aided the pathway enrichment analysis.

The only reason why the eight TCGA data sets were chosen for the analysis was because these cancer types were the only data sets that had both height and weight information of the samples. I agree that there may have been other cancer types more suitable than the eight that were chosen for this project, but those cancer types did not have the crucial clinical information (i.e. height or weight). In addition, it was of interest in this project to identify novel genes/pathways that associated with obesity across multiple cancer types, and hence including as many cancer types was important to get as many DEGs that were potentially obesity-associated in the first instance.

But, as you mentioned in your comment, perhaps for the purpose of pathway enrichment analysis, using only the cancer types that were known to be associated with obesity may have been more useful than using all eight cancer types. Since this analysis will take a considerable amount of time and addition of information to the thesis, I would like to dismiss this idea for the current project, but I will add this to the discussion.

It was not clear why it was necessary to run the simulation of the number of differentially expressed genes between 2 or more ICGC datasets when the cell of origin of these datasets were so highly variable (p83).

As mentioned above, it was of interest to find novel obesity-associated pathways or genes that were common across multiple cancer types, so DEG analysis was carried out on all available cancer types, regardless of the tissue of origin. The purpose of the simulation was to make sure, statistically, that the number of DEGs identified between multiple cancer types were significant, and not by chance. However, as you may know from the report, majority of these DEGs would have been false positives.

Again, using only the cancer types that were known to be associated with obesity may have aided the discovery of significant pathways and/or genes.

On page 110, it is suggested that as the obesity-associated genetic signatures were derived from breast cancer datasets, that it made sense that these signatures were significant in the bladder cancer data sets as 'basal-like' and 'luminal-like' subtypes had been identified in bladder cancer. As adenocarcinomas (epithelial-cell derived, and glandular in origin) all have these characteristics due to cell of origin context, I am not convinced by interpretations made in section 5.1.2, and feel the statements made are ungrounded.

Minor corrections

Acronym list are not all acronyms – some are definitions, or abbreviations.

Changed the heading to "Acronyms, Abbreviations, and Definitions" from "Acronyms".

Endothelial should be epithelial (p16)

Changed 'endothelial' to 'epithelial'.

". . . only logical to induce. . . " should be ". . . only logical to deduce. . . " (p16)

Changed 'induce' to 'deduce'.

Define NER (p20)

NER is mentioned in the previous page (p19) as "nucleotide-excision repair" and also in the Acronyms list.

Use full first and last names when referencing personal communication

Changed "Assoc. Prof. Mik Black" to "Assoc. Prof. Michael Black".

Throughout the thesis, the use of hyphens are inconsistent (e.g. obesity-associated, RMA-normalised, etc.)

These errors have been corrected.

"Hormone" status is incorrect – it should be "receptor" status (p73)

Changed "hormone" to "receptor".

"…none of the metagenes were not generalisable in other cancer datasets." does not make grammatical sense (p83)

Changed "…none of the metagenes were not generalisable in other cancer datasets" to "…none of the metagenes were generalisable in other cancer datasets"

The font used in the plots for figures 19-22 are exceedingly small (p100, 106)

I agree with the Examiner that the fonts are indeed small, but due to the nature of the plot arrangement for these figures the fonts cannot be made any larger (at least not trivially). Also, the font size was chosen so that the statistical values and letters did not obscure the points and boxes of the plots, and I would like to avoid this as much as possible, so I have left the font size unchanged.

Response to the comments made by Examiner "1137"

Individually published papers could have been critically appraised in the Introduction when it was introduced.

Potential bias and confounding in previously published studies due to the patient selection and the use of BMI as a marker of obesity are highly pertinent to thesis and could have been explored in more detail.

The effect of obesity on treatment response in cancer, and the influence of obesity on the likelihood of initial diagnosis, may also have implications for obesity-gene expression associations found in the literature.

Potential effects on gene expression in tumours by lifestyle factors associated with obesity, but not recorded in most data sets was not discussed in great detail, however this is relevant to the concepts explored later in Section 3.3.

It would be interesting to read here about the growth inhibiting mechanisms in cancer cells of anti-obesity drugs like orlistat.

The interesting literature from in vitro studies about the effect on cancer cells of factors produced by adipocytes in culture was not explored.

Potential mechanisms of obesity-altering hormone receptor expression could have been explored in more depth at the level of cancer cells. The effects of endocrine factors like leptin and the IGFs were discussed in general but not the cancer cell type-specific gene expression changes they cause.

A general in-depth exploration of the 'meaning' of the different data types used in the thesis and how they are affected by normalisation and statistical test choice would have been appropriate in the Introduction.

How does the overlap between the genes that constitute each metagene influence the statistical analysis?

Biologically important multiple splice variants of many RNAs are available in both the Affymetrix microarray and TCGA/ICGC RNAseq data. Could the use of all probe sets/isoform variants have provided useful additional information?

Although reasoning behind exclusion of some patients due to missing clinical data is discussed, I am not completely clear about this after reading the description, and the numbers in Table 2 do not concord with what I know of these data sets from my own research. I am sure that patients will only have been excluded with good reason and it seems unlikely these exclusions will have introduced bias, however I would have liked a clearer summary of what patients had been excluded and why. Perhaps I have missed this?

From Table 2, the NZBC and the Gatza *et al.* data sets were the only data set where the samples were excluded. For NZBC data set, only the samples that had BMI or obesity status information was included in the analysis (not stated in the thesis – will be included after revision). For Gatza *et al.* data set, only the samples that were genotyped with Affymetrix HGU133A platform was used in the analysis, as all of the other three microarray data sets used this platform.

These information were stated in Section 2.2.1 under the corresponding data sets, but perhaps it might have been better if it was stated separately in a section of its own.

What basic quality control was possible on each data set?

Why was the BRCA RNAseq data set in TCGA not used? Surely this could have been informative?

I completely agree with the Examiner and the BRCA data set would have aided greatly in the exploration of obesity-associated signatures from both Creighton *et al.* and Fuentes-Mattei *et al.* data sets. However, at the time of data retrieval (March 31st 2015), there were no height, weight or BMI information for the BRCA data set, and could not use the data set for the analyses.

On page 93 you say "The fact that some pathway associated metagenes were consistent across different data sets suggested that some of these genetic signatures were reliable and did not depend on the data set the transformation matrices were derived from." However, could the pathway expression signatures that were highly correlated in all of the data sets instead reflect a technical issue, perhaps an artefact affecting a dominant probe set?

This is an interesting point. All of the pathway-associated signatures were derived from the Gatza *et al.* data set and was then transformed into other data sets. For the pathway signatures to highly correlate between the data sets, all of the probe sets included in the pathway signature have to have similar expression in the other data sets as well. With this in mind, I doubt that there was a similar technical error in all of these data sets, for a particular set of gene probe sets, given that each one of these data sets was from an independent study.

The justification presented about the appropriateness of using SVD and transformation matrices is convincing but I wonder about the effects of small numbers of probe sets/RNAs with specific biological connotations dominating some metagenes.

I agree that there will be some probe sets/RNAs that would have influenced the metagene scores more than the others, and the identification of these influencial probes/RNAs would be of great interest. However, I feel like an investigation of the most informative probes/RNAs for each of the obesity- or pathway-associated signatures (or any genetic signature in question) would require an effort of a whole new project.

It would be interesting to explore whether the differences in risk ratio between cancer types was due to technical issues, or represented real biology.

Cancer patients can have vast differences in BMI at different points in their treatment. Was the clinical timing of weight measures standardised in some way to reduce the impact of this problem, e.g. using weight at initial diagnosis?

This is a good point, and none of the papers define at which stage of the disease the patients' clinical data was taken (check).

In the Introduction, add a few sentences of additional explanation about potential bias and confounding introduced into previously published studies of obesity-cancer gene expression relationships, including potential confounding by lifestyle factors associated with obesity.

In the Introduction, refer to the published in vitro studies of the effect on cancer cells of factors produced by adipocytes in culture.

In the Introduction, add a few additional sentences to more fully introduce potential mechanisms at the level of cancer cells for obesity to alter biology through changes in hormone receptor expression and transcriptional effects of endocrine factors like leptin and the IGFs.

Add a table supplementing existing tables to more clearly show what patients from each data set were included, or excluded, and why.

This will be considered after discussion with Mik and HOD.

Add additional description about what basic quality control was possible for each data set (perhaps in section 2.3), and the results of the QC. The discussion in section 5.2.2 is hard to interpret without this.

Discuss whether the biology underpinning differences between cancer types in terms of obesity risk ratios in Table 1 could be explored statistically using genomic data.

When discussing future work, suggest specifically how the limitations of using BMI as an indicative biomarker could be addressed. What additional clinical data fields would be helpful to include as factors in the models?

I have added an extra sentence in the last paragraph addressing the concerns about the use of BMI as a measure to classify obesity.