Reviewer's report

Title: Biobanking across the Phenome

Version: 1 Date: 28 September 2012

Reviewer: Naomi Allen

Reviewer's report:

Major Comments

1. Overall, this is a thoughtful and timely paper, although I think it would benefit from being more focused. It's part a general overview on the importance of biobanks for future research (away from a rather reductionist approach to a more agnostic approach with an emphasis on data handling and bioinformatics), and part a detailed review of pleiotropic genetic effects, and to a lesser extent, on cellular aging. It is not therefore clear to me what the aim of this paper is and it came across as quite disjointed. I would be inclined to focus more on the general rationale for biobanks to meet the changing needs of researchers & the different approaches used to measure ‘phenotype’, rather than a detailed view of pleiotropy or cellular senescence.

2. I was surprised to see no mention of metabolomics/proteonomics and its technological consequences for bioinformatics, given that this will become of major interest in the field of ‘phenotyping’.

3. It wasn’t entirely clear why the processes behind cellular aging were given such a strong emphasis. Presumably, it is included in this article as an example of why research into biomarkers is important (i.e. to differentiate the aging process from that of the disease process). Perhaps more should be made about the similarity in approach in trying to identify mechanisms/pathways and intermediate phenotypes in understanding both aging and disease pathology.

Minor comments

1. There is no clear definition of what a ‘phenome approach’ is. Please clarify.

2. Some clarification on the difference between cohorts and biobanks is needed (abstract; top of pg 4) and it would help to emphasize the fact that many biobanks do not have detailed information on environmental factors. Hence, why there is so much interest in the large cohorts outlined in table 1.

3. Can you include some examples (pg 3) of biomarkers of (a) mechanisms, (b) proxy for environmental exposures. Biomarkers also used as markers of functional phenotypes that are otherwise difficult to characterise (eg creatinine for kidney function; liver enzymes, etc.). An example of an exposome and etiome (Pg 5) would also be helpful.

4. Pg 5. The paragraph on etiome profile is not very clear and seemed to be a
mixture of clustering of risk factors and diseases. Perhaps this is as it should be, but the language needs to be clearer. (e.g., are ‘disorders’ synonymous with ‘disease phenotypes’?)

5. Pg 9 – last para. Informed consent. Most large cohorts (at least those that were established recently) have broad informed consent that covers investigation of future health research in all guises (i.e. consent is not specific to a certain type of research), so there is no conflict as such.

6. Table 1.
   • deCODE is in Iceland.
   • What do you mean by ‘mega’?
   • Why is size of the country relevant? No cohort studies (excoet national linkage studies) are nation-wide.
   • Million Women Study only collected blood on a sub-sample of women
   • China Kadoorie Study and Mexico City Prospective Study should be included. There is also the Canadian Partnership for Tomorrow Project.

7. English spelling/grammar is poor in some places, leading to poorly phrased sentences and confusion for the reader. Please clarify where possible.

Minor – suggestions

1. Perhaps details of UK Biobank and China Kadoorie Study could be combined; they both include 500,000 participants, national-representative from the general population; collection of lifestyle and biological samples, etc. worth also mentioning that these studies have longitudinal follow-up for health outcomes of all participants via linkages to national health-related records.

2. Pg 4: disease networks. I think it’s worth emphasizing that, up until now, the focus has been on accurate exposure assessment in epidemiological studies, with limited attention on the outcome assessment (largely due to the limitation of relying on death/cancer registries, with few studies having the technical/ethical ability to link to detailed outcome data from medical records). The authors mention asthma as an example of different clinical entities, but perhaps another example is stroke where different sub-types have different aetiologies and prognoses and where the identification of different (genetic or non-genetic) risk factors may have real clinical relevance for different sub-types.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests:

I declare that I have no competing interests