Analysis of proposed changes required for The Ark in order to facilitate management of all Raine data.

**Biospecimen management modules (**perhaps the best way to express these costs would be in terms of the personnel required (time and level) for each module)

* Currently I believe The Ark is a very good match to Raine’s data, given we are already holding some of Raine’s biospecimen data in the near-outdated, but similar, WAGER system. And have performed several migrations of WAGER data already successfully. Section 1 contains some potentially recommended necessities and/or improvements.

Eric, Travis may not have had the chance to fill you in on our meeting while you are away. But we had a good meeting around the interesting potential for us to fund some substantial work on the Ark, based around our critical need to upgrade the Software/Management Infrastructure of the Raine Study.

Travis, as we discussed, we are keen to upgrade all aspects of the Raine Study Infrastructure to ensure its future viability. Based on our meeting, it appears that the ARK could potentially meet all of the current and future needs of the Raine Study. Plus, there could be a two-way benefit as any modifications done for the purpose of the Raine Study could (1) allow you/Eric to fund continued work on the Ark and (2)  be used more widely, thereby benefitting a wider number of research groups.

It would be very helpful at this stage to get a preliminary idea of the one-off costs required to modify the Ark to meet the requirements of the Raine Study.

I have gone through my notes from our meeting and think that the following are what we would need:

**Biospecimen management modules (**perhaps the best way to express these costs would be in terms of the personnel required (time and level) for each module)

        Phenotypic Data Management (store all phenotypic data, currently approx 85,000 variables per participant. I note that there is already a tab for ‘phenotype’ in the Ark – this would need to be modified for Raine)

        Pedigree (module already in place, but modification would be necessary for Raine – participants, siblings, parents, grandparents, children)

        Study Management (very keen to integrate a management module into Ark that allows us to manage all of our forms, data access requests, manuscript forms, project requests, etc, etc)

        Data extraction and reporting (phenotypic and genetic)

        Subject access (have a portal that allows participants to securely access selected reports of their own data)

**Other costs** (estimates of one-off setup costs + any annual costs would be helpful here)

        Data storage (I guess this mainly relates to the genetic data – GWAS, EWAS, Exome, etc)

        Data backup (as per our discussion).

Section 1: LIMS/Biospecimen Module.

Currently I believe The Ark is a very good match to Raine’s data, given we are already holding some of Raine’s biospecimen data in the near-outdated, but similar, WAGER system. And have performed several migrations of WAGER data already successfully.

* There will be a need to migrate biospecimen data from many disparate systems. We would need to know all about all of the systems and the data to precisely estimate, but we can broadly guess 2-9 (1.5-4.5 \* 2FTE) months of developer time for migration and changes and 1-2 months of testing time assuming there is no large fundamental changes required to the system (this may not be needed due to our “biospecimen and biocollection custom fields functionality that already exists for non-standard fields). Migration time is very directly correlated to the quality and consistency of the data coming in. If we don’t have it
  + Please provide some rough numbers on the amount of data and systems from which we will move.
* My experience in industry would have me recommend setting up of (preferably automated, continuous) testing for all functionality. Even very well manually tested systems can have faults, which can be very costly once logic in applications becomes increased to the point a human can’t be aware of every bit of code/logic possible. For biospecimen functionality I would suggest the amount of time to develop automated test cases would be 4-18 months depending on the depth of coverage required – and the extent of automation of testing we go for.

Phenotypic Data Management (store all phenotypic data, currently approx 85,000 variables per participant. I note that there is already a tab for ‘phenotype’ in the Ark – this would need to be modified for Raine)

Again we would need to try out a rough simulation of the data to know for sure.

1. I would recommend with such a quantity of data we analyze and review performance of all functions related to phenotypic data. Having performed this kind of review and analysis in other sections we were able to have several orders of magnitude improvement in performance. Given this section was written slightly later, I believe there may be less room for improvement, but we will need all the performance we can.

2. There is also a fundamental question on if the data, as it currently exists, needs to be refactored (i.e. Can this data adequately/optimally perform the function for it’s users in it’s current form.

Even refactoring this data into an optimal format, there is still a sheer mass of data that would have me recommend the following functionality be added to The Ark.

3. Add Groupings and Subgroupings for phenotypic custom fields in order to make the data more sense. We already have the concept of custom fields being reusable and collected into “custom field groups” (also referred to as Questionnaires or clinical Datasets). This significantly reduces the amount of work required to create and maintain questionnaires/fields if obvious fields like heart rate. But adding groupings in regards to how users can find fields might be easier. Eg; enabling a study to make “cardiovascular”, “sleep”, “lifestyle”, “medication”, “medical history” sections and appropriate subsections in order to allow finding the questions to make questionnaires. These groupings could also be utilized in data extraction. Estimate: 3-4 months development. 1-2 month test development

4. Commonly requested feature for phenotypic data, which we have not enabled yet is “skip logic” for display of certain questions based on previously answered questions. This could be useful for data entry but also as a model for how to open up to subject/participants answering questionnaires more accurately, conveniently and quickly.

5. My experience in industry would have me recommend setting up of (preferably automated, continuous) testing for all functionality. Even very well manually tested systems can have faults, which can be very costly once logic in applications becomes increased to the point a human can’t be aware of every bit of code/logic possible. For phenotypic functionality I would suggest the amount of time to develop automated test cases would be 4-18 months depending on the depth of coverage required – and the extent of automation of testing we go for.

Pedigree (module already in place, but modification would be necessary for Raine – participants, siblings, parents, grandparents, children)

We have a lot of these concepts already. There are remaining things that you may want;

1. Pedigree data integrated in the Data extraction module. It is currently extracted separately and not at all integrated.
2. Pedigree module is currently utilizing an existing library for 2D image rendering. This process could do with some streamlining to ensure it is more easily maintained and setup.

Study Management (very keen to integrate a management module into Ark that allows us to manage all of our forms, data access requests, manuscript forms, project requests, etc, etc)

This concept doesn’t clearly exist in The Ark. We just keep very minimal information on a study. Estimates;

1. Create module to manage all forms at a study level – given the increase in data, we may also change the model of how we are storing data in general to maintain adequate performance. ARK-1155 (11d) & ARK 1185 (20d)

Data extraction and reporting (phenotypic and genetic)

1. Expand and add relationships between filters (or’s, brackets, etc)
2. Compulsory testing of improvements
3. Automated Testing

Subject access (have a portal that allows participants to securely access selected reports of their own data)

* set up subject portal
* subject login screen
* subject starter screen
* subject subject report startup screen
* 4 basic reports?
* (do you want to allow subject access to “questionaires”?)