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

**P**erhaps the best way to express these costs would be in terms of the personnel required (time and level) for each module);

 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 –** 20-75 days migration and data testing time. Dev time not assumed needed until I am told of new requirements for our biospecimen 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) 20-90 days migration/testing of migrated data, 45-250 days development and testing time to improve The Ark for Raine data

        Pedigree (module already in place, but modification would be necessary for Raine – participants, siblings, parents, grandparents, children) 37 days development time, not sure about migration time involvement – your staff or we can use the interface or the uploaders in the ark. Let’s say 5-20 days migration and data testing time

        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). Based on assumptions below approximately 45+ developer and testing days depends upon the extent of functionality and workflow requested

        Data extraction and reporting (phenotypic and genetic) 90-200 developer and testing days dependent upon changes requested and any custom reports

        Subject access (have a portal that allows participants to securely access selected reports of their own data) 140-180 developer and testing days

**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) – should be free but must sort out back up system (discussed below)

        Data backup (as per our discussion). (estimate $150-300 and up per TB per year to maintain data – options outlined below)

        Support analyst ~1.25 FTE over 2 years for support, training, management implementing bug fixes, keeping servers and backups up to date. We might start around 1.5 first year, and then work on training Raine staff to take over as much as possible subsequent years.

**Other work I would recommend** to enhance the system and make the data as secure and safe as possible but all subject to what you think is best for Raine and best for the grant (up to 500+ days)

**So, in summary, we are looking at a total estimate around;**

45-185 days developer time for data migration and testing of migrated data.

330 to 750 days of developer time to provide a minimum system improvement. **Plus,** potentially another 500 developer days of improvements I see/recommend as outlined below – if we refine requirements these numbers will get more precise numbers.

We can probably estimate getting 200-230 developer days per year out of a FTE developer/tester/support analyst. These assumptions are based on at low level 8, high level 7 software engineer. Around 80-90K per year over the next few years. Generally - Cheaper resources will take more time, more expensive resources will work faster.

So we can estimate 2-3.5 FTE for 1-2 years depending on functionality required (I would estimate more toward the higher end given I think a lot of required functionality will become more obvious once we start specifying requirements more carefully). Followed by ~1.5 FTE support first year, once the system is live. Obviously, the support period and development can be incrementally merged if certain functionality is less urgent. We can also look at you data storage requirements more to give a dollar amount.

Section 1: LIMS/Biospecimen Module.

Currently I believe The Ark is a very good match to Raine’s biospecimen data, given we are already holding some of Raine’s biospecimen data in the soon to be replaced, but similar, WAGER system. And The Ark Team has performed several successful migrations of WAGER data already.

* 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 developer time for migration and changes, 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.
  + Please provide some rough numbers on the amount of data and systems from which we will import biospecimen data. My recall was biospecimen data was in 9+ places? 3-12 (1.5-12 \* 2FTE) months of developer time. Plus 1-3 FTE months testing time is our guess in the interim.
  + This sort of thing always depends upon the quality of the data and consistency of the data. If it is inconsistent and doesn’t maintain the structure expected then a lot of manual intervention is needed
  + For now I will just put a very conservative estimate of
    - Migration of data from all existing systems and XL documents (3-12 developer months) ARK-1187
    - Manual Testing of migrated data (1-3 months) ARK-1188
    - Changes to system ? Do we envisage any changes needed? Xx months?
    - Setting up biospecimen custom fields?
* 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 introduced any time logic/functionality is added, 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 for all of the possible uses of The Ark LIMS module would be 3-6 months depending on the depth of coverage required – and the extent of automation of testing we go for. If Raine solely have the interest of storing LIMS data and not processing it, then it would not be necessary to push this agenda for Raine’s situation. Estimates at end of document.

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 what we are dealing with.

1. Migration time depends upon the data. Perhaps it can all be handled by our uploaders, and all we need is very significant testing time on migrated data (1-4 months).

2. 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. 1-3 months developer time optional

3. 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;

4. Add some sort of organizational 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: 1 months development. 0.3 month test development ARK-1157

5. Commonly requested feature for phenotypic data, which we have not implemented 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. Optional 1.5 months developer time 0.3 month w test time ARK-1165

6. 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 1-4 (plus initial cost of setting up to enable automated testing) months depending on the depth of coverage required – and the extent of automation of testing we go for. All automated test script estimates at the end

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. 30d ARK-xxxx
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. 7d ARK-1126

Please let us know if there is anything more you might need.

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)

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 (25d)

2. Create data access requests page and possibly link to roles for whom to email? Perhaps have some logic regarding pending requests being sent follow up emails or having reminders on screen. ARK-1186 (20d)

3. manuscript forms ? are these just being stored…if so then its just like point one (but a good reminder we may want “types” of forms/files to store. Or is that we want some kind of “specific data entry forms / questionnaires” which are related to studies? (x days)

4. Project requests. Like above, see point “3” (x days)

NOTE/ASSUMPTION: Please elaborate how this might all work.

I am not sure of the extent to which we wish to take all of this. If we forget about things like workflows, etc and just keep various files and maybe have a structure and various appropriate information related to then it’s not a very large amount of work. If you wish me to quote a workflow system I may need to sit down with someone for a while and run through some scenarios/requirements

Data extraction and reporting (phenotypic and genetic)

1. Expand and add relationships between filters in data extraction (eg; OR’s, brackets, cross module grouping/joining of filters/conditions etc). (~60 days)
2. Expand to include consent (and component consent) fields in output and filters? (optional 40 days)
3. Increase validation and smart error message on filters for less technical staff ARK-953 (6 days)
4. Manual testing of improvements (30d)
5. Automated Testing (see end of document)
6. There is already a list of requested improvements in JIRA. These would all be great if time permits. We could review these if going ahead with the grant application (optional 60+ days)

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

* The more details or requirements I have, the better estimates I could have
* set up subject portal (30d)
* subject login screen (30d)
* subject starter screen(30d)
* subject subject report startup screen (10d)
* 4 basic reports? (40d)
* (do you want to allow subject access to “questionnaires”? i.e.; people can enter their own phenotypic data) (30-40 days)
* bells and whistles such as facebook connect (to save users adding a new login – and entice them in more easily). (15d)

**Essential Costs beyond The Ark enhancements;**

1. Management of an instance of the data, user training, user documentation/videos once production system is live;
   1. Management of a separate instance of The Ark for 2 years (user support, implement bug-fixes, upgrades, facilitate use of The Ark during initial learning curve). 2 years of 1 FTE.
   2. If this is opened up to allow all subjects access their could be a period of extra support being needed to facilitate subjects (it expands the users from 5-40 scientists to potentially many more users accessing through the “subject portal”). Potentially adds 2 years at an average of approximately 0.25FTE.
   3. A suitable alternative on support may be to integrate one/some of Raines’ FTE’s to take over this role as time goes on. In which case there may be an upfront amount of training for staff and then just a smaller amount of time per week as time goes on to make sure the skills are learnt to ensure longevity.
2. The physical environment to host The Ark (not including storing sequence data)
   1. If we can get a free instance from UWA, NSP, etc then there is no cost for physical storage. I believe this should be available, even if it is a little under-resourced and underpowered. If you choose, an offsite backup of some sort for your data, it should cost less than $500 a year I would imagine.
   2. If a commercial service such as amazon is used the cost would $800-4000 p.a.
3. The storage of raw sequence data, BAM files, results related to these. (please provide estimate of quantiy of genetic data size for quotes)
   1. If iVEC, RDSI, etc provide adequate space then there is no cost for that BUT they insist upon having a backup stored somewhere. Your options include;
      1. Utilizing GOHaD’s Red Box backup techinique (reasonable reliability, slow speed backup only, stored in iVEC supercomputer room). This may cost in the realm of $150 per year per terabyte up to $300 per year per terabyte.
      2. Maintaining a solution of your own for backup, similar to GOHaD or
      3. Seek a quote from IBM, Dell, HP, etc on mass storage (they may offer a great deal as a “sponsorship” or donation…if so GOHaD would be happy to participate in such a proposal)
      4. Utilizing Amazon storage monthly bill could be in the thousands.

**Other highly recommended work.**

1. Create structure and architecture to facilitate automatic continuous testing. ARK-1189 125d
2. Automated biospecimen tests 100d
3. Automated subject and custom subject data tests 75d
4. Automated study management tests 25d (depends upon the extent of functionality)
5. Automated phenotypic data tests 50d
6. Automated Data Extraction testing 80d
7. Data de-identification and control of new roles to graceful control what is de-identified. (60+ days)
8. Expand Interface to enable storage / details of genetic data.
   1. Expand interface to keep records of genetic data – and relate it back to subjects, biospecimens, etc 30 days (do you want this)
   2. Expand data to have relevant meta-data regarding genetic data (what sort of genetic data, coverage, etc) 30 days (do you want this and how much metadata do you want to track)
   3. If you don’t require all of this, you could simply store your data on ivec separately on your own (remember you need to keep a backup) OR we could help you OR we could store it and keep the back up. Xx days (let’s discuss what you want before I consider estimates)