**HIC/OMERO Pilot & GoDARTS Scenario Notes**

**From steering meeting 28/04/2011:**

* Generate a 'best' reference by using multiple sources (GWAS, Metabochip etc) to define overlay for the same people.
* The reference genome is a moving target; want to overlay data to generate the best understanding of what is happening at anyone position - as algorithms, micro arrays etc improve then new data can be applied/imputed to further improve the reference; as the knowledge expands the dimensions expand; a kin to a version control of the reference.
* At the moment this is a manual process that has to be done each time; KZ did this previously by imputing the Metabochip data onto the GWAS data resulting in 80K common (markers?); CP would like to redo this (using the 1000 genomo data) upto 120k; this data could then be linked with the clinical data to support research questions.

**From sub group (AJ,SW,CP,KZ) meeting 10/06/2011 (notes & transcript of audio):**

* Start from 1000 human genome (imputed gwas to hapmap), around 3700 people. **Colin to get Natalie to provide us with access to the data (already stored at CLS)**;
* Apply algorithm to overlap metabochip (directly typed) to build a “common panel”. Have file of 88,000 common SNPs when did this previously. We would hope to achieve 110-120,000 SNPs covering 10,500 people.
* Does the algorithm exist on the internet – no; there are 3 different calling algorithms on the metabochip, pairwise, concurrency between algorithms and quality control. **Kaixin has defined how to merge and how to qc the results of the algorithms (provide this code to us).** The algorithm generates a file of genotypes.
* Use plink (linear regression against phenotype) to get residuals which then use against genotypes.
* Could test with some lipid questions – statin response / untreated lipids
* Phase 2 could look to also do the imputation step using hapmap
* Phase 3 could look to use probabilities
* Aim is to understand the process from (geno/pheno) raw data to the analysis.
* Want to repeat something we have done previously and possibly add-to.
* **What GWAS data do we need to use? [@1min].** The sample genetics data Kaixin has already provided is not the imputed GWAS. We need the imputed GWAS to CCC2 data to mimic what did previously, or to do something novel we should use the imputed GWAS to 1000 genome. Best idea is then to get the 1000 genome GWAS data (from Natalie, via Colin). Covers 3700 people typed on the (1 million) GWAS chip, imputed to a reference panel of 12-15 million SNPs, then basic qc applied which brings coverage down to about 7-8 million. Very large files, hosted at CLS – need access to these rather than copies.
* **What is the process of overlaying the Metabochip data? [@6min].** Apply algorithm to generate a common panel between the two datasets (metabochip different/larger population with overlap). Currently have a file with about 88,000 common SNPs between the 2 approaches. In theory this should be able to get us to about 100/110,000 common SNPS. The total number of SNPs on the metabochip is 180,000 but QC this which reduces it significantly. Kaixin developed the QC (decision process) algorithms (perl scripts) – there isn’t a public definite algorithm, there are 3 different calling algorithms (directly typed) on the metabochip and they all had a core of SNPs they all agreed on, and about 5000 SNPs each that were non-sense; need to be able to judge the SNPs where the algorithms fell over. We need this code to be able to get the baseline data, not sure if code still exists but the logic isn’t too bad – pair wise comparison of the 3 files – looking for consensus e.g. if 3 called want 2+ agree, call rate, etc. Using a single calling algorithm you will get things that are wrong and things that are missing, there is a lot of rare material which the software may not detect. This process is really to extract the highest quality and amount of data from the directly typed data. Can use the GWAS as a scaffold/template to impute all the rest of the genomic sequence on. Can’t do this within the metabochip because it has patchy coverage of the genome, don’t have enough systematic random data. The metabochip data is in 3 files for everybody. Stitching together the populations from the 2 approaches using concordancy between calling algorithms on the metabochip and basic qc on the GWAS imputation to establish a new reference genotype for those individuals. Would expect to increase the number of SNPS to 100/110,000 SNPS for the combined population size of 10,500 individuals (n.b. population should not change, number of SNPs should change).
* **How did you use the data? [@21min].** Working on plain text files for the metabochip files, gwas binary plink format. Heavy processing is just on the imputation which won’t need to worry about this first scenario. Ideal to be able to get to an updatable imputation but this shouldn’t be the focus at the moment. Already phased directly typed material, which is the best point to start from (phased haplotype file – can impute rapidly on). Task at the moment is not to do this, but to make the overlay file and do a statin response / lipid related phenotype analysis, using the HIC data (it doesn’t require the SCIDC data). It is Louise Donnelly that we would need to speak to nearer the time. There will be useful cleaning/modeling routines that we can apply to the clinical data too. The full linked analysis can be done in plink as linear regression analysis. More efficient to develop the phenotypic model of population independent of genotypic data (output residuals), which you then use against the genotypic data. The researchers tend to work on the binary plink files. Sensible to limit this to working with real genotype data, using probabilities could cause a lot of complications. This could be something we need to consider in the future.
* **Actions:**
	+ **Work with Natalie to get the gwas data.**
	+ **Work with Kaixin to get the metabochip data, (pseudo) code for merging the metabochip 3 files with the 1 gwas data file to create 1 reference file, and the specification of QC processes (concordancy, hwe?).**