RSV Data Sorting

RSV NGS File Contents

- A4634_0001-1244_QC_Summary.pdf
  - QC probe/batch level performed for samples
  - Not clear if these samples have yet been removed from the extracted data folder
- A4634_0001-1244_QC.txt
  - Individual sample QC summary statistics and Pass/Fail
- A4634_0001-1244_Genotyping_Data
  - Very large .txt file (4.4GB) is this all the data processed and combined?
- A4634_0001-1244_ProbeSet_Summary_Table.txt
  - 99MB, also very large does this have the snp IDs probably?
- Raw data > .cel files
  - Array intensity data, binary format
  - Needs to be converted to VCF > Plink
  - also comes with attached audit, jpeg, arr files
  - Analysed in batches

Conversion to VCF files

- Going from cel files to vcf would require mapping etc.
- Birdseed is a genotyping algorithm for genotype calling of SNP6.0 arrays
- Call genotypes using apt-power tools to get CHP files
- Then once data is called convert to vcf/txt > PLINK
- See github link for wrapper on this workflow

Thoughts

APT powertools has some options to process and convert data, however it's hard to know where to start…

If the .txt file is all of the processed/QC'd data, then according to the guidebook we would need annotation files (.ps), and call files? which I don't think we have. Also what version is everything. All very unclear

Or we run the genotype calling alrgorithm for the raw cel files and work from there, but that it beyond what I think we usually do! I wouldn't know how to evaluate the output of the calling algortihm/settings…would probably just have to stick with defaults

Conclusion: Quite hard to determine how to proceed without all the info

Relevant links

https://assets.thermofisher.com/TFS-Assets/LSG/manuals/axiom_genotyping_solution_analysis_guide.pdf

http://media.affymetrix.com/support/developer/powertools/changelog/index.html see apt-convert http://media.affymetrix.com/support/developer/powertools/changelog/apt-format-result.html From this we would need to generate calls file first

https://github.com/freeseek/gtc2vcf This might be the way forward if not: Uses apt-powertools snp calling algorithm > CHP files > VCF files Also has info on getting ref genome/array annotation files