temp

Øyvind Helgeland

May 3, 2018

Predefined (static) input files

- *A* - *"/media/local-disk2/helgeland/rotterdam1/inferred-pedigree/permanent_bad-sample-IIDs.txt"*
- *B* - *"/media/local-disk2/helgeland/rotterdam1/inferred-pedigree/permanent_reconstruct-fam.txt"* The input file *A* with pre-solved problematic pedigrees contained **7** resolved families with **20** index individuals. The input file *B* with pre-identified problematic samples (often accidental duplicates of other samples) contained **65** individual IDs. ### Upstream (dynamic) input files Genetic files with **17742** individuals reached this module. The original .fam file listed **5815** fathers (V3 column) and **5832** mothers (V4 column) who had genotypes (i.e., were listed in V2 column); also **66** fathers and **58** mothers without genotypes. The final .fam file will have them reset as missing (respective numbers **0** and **0**). ### Thresholds and procedures for relationship and sex inference The thresholds used to identify paren-offspring relationship were $Z1 > 0.8$; twin or dublicated samples - $PI\_HAT \geq 0.8$; full-siblings relationship - $Z1 \geq 0.35$, $Z1 \leq 0.65$, $PI\_HAT \geq 0.35$, $PI\_HAT \leq 0.65$. The Y chromosome genotype count threshold used to separate males from females was $YC > 92$. The X chromosome F threshold used to separate females from males was $F < 0.648$. Genetic sex was inferred based on both criteria. When criteria disagreed (**4** cases), samples were flagged as not suitable for analyses *(phenotypeOK=FALSE)* and genetic sex was infered from the X chromosome data. ### Modifications to .fam file Being found in the input file *B*, **20** samples in the .fam file got assigned their true family IDs, genetic parents and genetic sex. These samples are suitable for analyses and are not flagged as problematic. Being found in the input file *A*, **64** samples in the .fam file got assigned dummy family IDs (e.g. "prblm001"), got founder's status (i.e., parental IDs were set to "0") and their declared sex was set to their genetic sex. They were flagged as not suitable for future analyses *(phenotypeOK=FALSE)*. In sex inference for all these updated samples, there were **0** cases where Y-chromosome and X-chromosome data did not agree (likely Klinefelter). The declared and inferred sex did not match in **35** samples. The remaining .fam file contained **11549** declared parent-offspring relationships (**5756** paternal, **5793** maternal). Genetic inferrence of the same data detected **11549** parent-offspring relationships and **0** pairs of dublicated (twin) samples. If there is a difference between declared and inferred numbers, the auto-generated .pdf report should be manually inspected to detect new sample-identity problems. In sex inference for all these samples, there were **4** cases where Y-chromosome and X-chromosome data did not agree (likely Klinefelter), and **4** cases where declared and inferred sex did not agree. These samples were flagged as not suitable for analyses *(phenotypeOK=FALSE)*. ### Summary In total, the number of samples flagged as not suitable for analyses *(phenotypeOK=FALSE)* is **68**. We do not trust the identity of these samples. The remaining **17674** samples were flagged as OK *(phenotypeOK=TRUE)*. The updated .fam file contained **11562** declared parent-offspring relationships (**5762** paternal, **5800** maternal). The genetic (Xchr) sex was assigned to all the samples. ### Problematic genotyping arrays The samples from the input file *A* where enriched in these genotyping arrays:

chip	prob_count
201641480150	8
201680470069	8
201689730161	7
201570320148	1
201641510161	1
201641770011	1

(the table is trimmed) The samples from the input file *B* where enriched in these genotyping arrays: (the table is trimmed)

chip	prob_count
201692730188	3
201629310100	2
201629310157	2
201680470065	2
201641780148	1
201641790166	1