**Grossly Uninvolved Tissue**

Grossly uninvolved’ (GU) tissue refers to tissue that is beyond the margins of the tumor and pathologically determined to contain 0% tumor. In the absence of peripheral blood, DNA extracted from GU tissue can be used for germline (inherited) DNA. To establish adequacy of this presumption, we undertook a pilot study to compare high-density SNP genotyping (Illumina 610Q assay that we used for the SGS) from peripheral blood and formalin-fixed paraffin-embedded (FFPE) GU DNA. All FFPE GU material was reviewed by a pathologist. We selected 10 breast cancer cases with both peripheral blood DNA and GU FFPE material available. The FFPE material for these 10 women were fixed between 1975 and 2006, a range that is representative of the blocks in the pedigrees. We genotyped 20 samples: 10 paired blood DNA-GU FFPE tissue DNA. The concentration (PicoGreen) and volume of DNA from GU material was very good on average, although the range was wide (Concentration:average 47.8 ng/μl, range 9.3-152.2 ng/μl; Volume: average 1,818 ng, range 353-5,784 ng). As expected, the quality of the FFPE DNA was lower than for the blood samples, as indicated by wider scatter on the allele calling plots, however, the algorithm was able to call genotypes for the vast majority of SNPs and all samples produced results. Genotype call-rates for the GU DNA samples ranged from 67.0% to 99.2% (average 90.6%; 67.0% was an extreme outlier). Of the called alleles, the genotypes from blood and FFPE GU DNA were extremely concordant (average 99.6%, range 98.3-100.0%) indicating that when a call was made that it was accurate. On average, 498,300 accurate SNP calls were made for GU DNA samples. Similar results have been found in GU FFPE samples from prostate tumors using a similar protocol.(1)

**Calculating Standardized Incidence Risk (SIR) Ratios**

For each index family:

The standardized incidence ratio (SIR) for a cancer group D is estimated by the ratio of observed O and expected cases E. E is calculated as:

$$E= \sum\_{ij}^{}py\_{ij}\*λ\_{ij}$$

where $py\_{ij}$ are the observed person-years in the index family for age group $i$ and calendar period $j$, and $λ\_{ij}$ the corresponding incidence rates for cancer group D in the reference population cohort. In the reference population cohort, the incidence rate $λ$ for the cancer group D is calculated as:

$$λ\_{ij}=\frac{d\_{ij}}{n\_{ij}}$$

where $d\_{ij}$ is the observed cases and $n\_{ij}$ is the observed person-years in the reference population cohort for age group $i$ and calendar period $j$.

**Calculating Gower Distance**

The Gower distance is a general similarity coefficient for defining distance between families. Similarity between families *i* and *j* with values *xi* and *xj*, *i, j* = 1,….,*n*, based on *p* variables, is defined as

$$s\left(x\_{i} , x\_{j}\right)= \sum\_{k=1}^{p}s\_{k}\left(x\_{ik} , x\_{jk}\right)δ\_{k}\left(x\_{ik} , x\_{jk}\right)w\_{k} / \sum\_{k=1}^{p}δ\_{k}\left(x\_{ik} , x\_{jk}\right)w\_{k}$$

Where $δ\_{k}\left(x\_{ik} , x\_{ik}\right)$ = 1 when *i* and *j* are both non-missing and 0 otherwise. Optional weights *wk* can be included to raise the importance of certain variables at are *a priori* considered more relevant. Cancers known to be associated with BRCA1 and BRCA2 received the highest weights (Figure 1). The other cancers were weighted using a weight inversely proportional to the population incidence of each cancer category, with rare cancers receiving higher weight and common cancers receiving lower weights. We constructed a 5,045 x 52 matrix (families x 2 risk metrics per cancer type, including breast) and calculated the Gower general coefficient (ade4 R package) for distance metric to allow for the mixture of both variable types.

Sensitivity analyses without weights and rank ordered weights were also conducted. The results varied based on the weighting scheme, however the models were more stable with the weights assigned and we chose to present the weighted results here. The score $s\_{k}\left(x\_{ik} , x\_{jk}\right)$ captures the similarity between samples *i* and *j* w.r.t. variable *k*. The score is defined as:

* Qualitative variables: $s\_{k}\left(x\_{ik} , x\_{jk}\right)= \left\{\begin{matrix}1, i and j agree in k\\0, i and j differ in k\end{matrix}\right.$
* Quantitative variables: $s\_{k}\left(x\_{ik} , x\_{jk}\right)=1-\left|x\_{ik}- x\_{jk}\right|/R\_{k}$

where Rk is the observed range of variable *k*

The distances are derived from the similarity values $d\_{k}\left(x\_{ik} , x\_{jk}\right)=1-s\_{k}\left(x\_{ik} , x\_{jk}\right)$.

**Supplementary Table 1. Description of cancer classification for cancer groups 1 – 26 used in the familial multi-cancer configuration (FMC) clustering by Surveillance, Epidemiology, and End-Result (SEER) Code.**

|  |  |  |
| --- | --- | --- |
| SEERCode | Definition | Category |
| 20010 | Lip | 1 |
| 20020 | Tongue |
| 20030 | Salivary Gland |
| 20040 | Floor of Mouth |
| 20050 | Gum and Other Mouth |
| 20060 | Nasopharynx |
| 20070 | Tonsil |
| 20080 | Oropharynx |
| 20090 | Hypopharynx |
| 20100 | Other Oral Cavity and Pharynx |
| 22020 | Larynx |
| 22010 | Nose, Nasal Cavity and Middle Ear |
| 21020 | Stomach | 2 |
| 21030 | Small Intestine | 3 |
| 21041 | Cecum | 4 |
| 21043 | Ascending Colon |
| 21045 | Transverse Colon |
| 21047 | Descending Colon |
| 21048 | Sigmoid Colon |
| 21049 | Large Intestine, NOS |
| 21051 | Rectosigmoid Junction |
| 21052 | Rectum |
| 21060 | Anus, Anal Canal and Anorectum |
| 21044 | Hepatic Flexure |
| 21046 | Splenic Flexure |
| 21100 | Pancreas | 5 |
| 21010 | Esophagus | 6 |
| 22030 | Lung and Bronchus | 7 |
| 24000 | Soft Tissue including Heart | 8 |
| 25010 | Melanoma of the Skin | 9 |
| 26000 | Breast | 10 |
| 27010 | Cervix Uteri | 11 |
| 27020 | Corpus Uteri | 12 |
| 27030 | Uterus, NOS |
| 27040 | Ovary | 13 |
| 28010 | Prostate | 14 |
| 28020 | Testis | 15 |
| 29010 | Urinary Bladder | 16 |
| 29030 | Ureter |
| 29040 | Other Urinary Organs |
| 29020 | Kidney and Renal Pelvis | 17 |
| 31010 | Brain | 18 |
| 31040 | Cranial Nerves Other Nervous System | 19 |
| 32010 | Thyroid | 20 |
| 32020 | Other Endocrine including Thymus |
| 33011 | Hodgkin - Nodal | 21 |
| 33012 | Hodgkin - Extranodal |
| 33041 | NHL - Nodal | 22 |
| 33042 | NHL - Extranodal |
| 34000 | Myeloma | 23 |
| 35011 | Acute Lymphocytic Leukemia | 24 |
| 35012 | Chronic Lymphocytic Leukemia |
| 35013 | Other Lymphocytic Leukemia |
| 35021 | Acute Myeloid Leukemia | 25 |
| 35022 | Chronic Myeloid Leukemia |
| 35023 | Other Myeloid/Monocytic Leukemia |
| 21071 | Liver | 26 |

|  |
| --- |
| **Supplementary Table 2.** Summary of Utah Population Database (UPDB) Data Contributors and Data Contributed |
| **Contributor** | **Data Contributed** |
| Family History Records | 1.9 million individuals (early 1800s – mid-1970s)-  The original genealogical component of the UPDB is a vast set of Utah family histories organized into pedigrees. These genealogies are based on three-generation documents from the Genealogical Society of Utah where each one holds demographic and kinship information (births, deaths, marriages, locations) for a couple, their children, and the couple’s parents. These data depict Utah’s founding population and their descendants. These large pedigrees represent persons affiliated with the Church of Jesus Christ of Latter-day Saints (LDS, Mormons) as well as non-LDS individuals. -  Records transfer of 90 million deceased individuals from FamilySearch have been linked to the UPDB. Individuals likely related to individuals represented in present UPDB and extend largely to the US and Europe.  |
| Center for Medicare and Medicaid Services (CMS) Claims | ~ 1 million individuals (1992–2012) -  97.8% link into UPDB. These data provide vast information on individuals 65+ with respect to morbidity, mortality, and health care. |
| Utah Statewide Inpatient Discharge | ~5.5 million encounters (1996 – present)-  Records include admission date, discharge date, and diagnoses and procedure codes. |
| Utah Statewide Ambulatory Surgery | 6.3 million encounters (1996 – present)-  These data contain surgery information occurring in hospital outpatient departments, hospital-affiliated ambulatory surgery centers, and freestanding ambulatory surgery centers. Records include admission date, discharge date, and diagnoses and procedure codes.  |
| Utah All Payers Claim Data  | 105 million medical claims (2013 – 2017)114 million pharmacy claims (2013 – 2017) |
| University of Utah Healthcare (UUH) EDW | ~2.1 million individuals (1996 – present)-  This resource covers all University of Utah hospitals and outpatient clinics. Medical information and demographic records from UUH link to UPDB. |
| Intermountain Healthcare (IH) EDW | ~4.3 million individuals (1996 – present)-  IH is a not-for-profit health care system covering Utah and southeastern Idaho. Medical information and demographic records from UUH link to UPDB.  |
| Driver License Division | ~3.9 million licenses (early 1990s – present)-  All driver license data collected for Utah residents including height, weight, and address |
| Utah Cancer Registry (UCR) | 363,000 cancer diagnoses (1966 – present)-  UCR is a statewide cancer registry established in 1966 and monitors cancer incidence and mortality and is a member of the NCI Surveillance, Epidemiology, and End Results Program. UCR holds statewide data on stage of disease at diagnosis and patient characteristics (i.e., age at diagnosis, sex, race/ethnicity, place of residence at diagnosis, and cause of death). |
| Utah Department of Health (UDOH) -- State Birth Records | ~2.93 million Birth Certificates (1931 – present)-  Utah birth certificates from the UDOH are linked into sibships and families and form the basis of parental reproductive histories in UPDB. They provide basic information about the timing of all births, conceived naturally or otherwise, in addition to basic demographic and extensive clinical information about all mother-father-child triads. A new-born on a birth certificate appears later as a parent on another, a feature that UPDB exploits to create large pedigrees separate from the GSU. |
| UDOH -- State Death Records | 865,000 Death Certificates (1904 – present)10,500 Fetal Death Certificates (1978 – present)-  Causes of death are coded using ICD revisions 6-10 for all death & fetal deaths. Industry/occupation info are coded to generate socioeconomic scores. |
| Utah Voter Registration | 2.1 million records |
| Individual Level Census Records | 2.3 million records (1880, 1900 – 1940 [1890 UT Census records were destroyed]-  Federal regulations prevent access to individual Census records within 72 years of the present. Overall 2.3 million records  |

**Supplementary Figure 1.** Silhouette plot for the final familial multi-cancer configuration (FMC) model (K=5)



**Supplementary Figure 2.** Consensus Matrix Heatmap for the Final Familial Multi-Cancer Configuration (FMC) K=5

**Supplementary Figure 3 –** Heat plot by familial multi-cancer configurations (FMC) displaying the direction of risk across the 26 cancer types.

****

References

1. Cannon-Albright LA, Cooper KG, Georgelas A, Bernard PS. High quality and quantity Genome-wide germline genotypes from FFPE normal tissue. BMC research notes. 2011;4:159.