Lunch & Learn: Science, Statistics, and Getting it Right

ASIP 2018 Annual Meeting at Experimental Biology

Vignette 2

A group of 1175 healthy subjects (43% Caucasian, 33% African or African American, 24% Hispanic/Latino) were recruited from college campuses in the Boston area (from among 26 different colleges) and were asked to provide a buccal swab for DNA sequencing along with a detailed questionnaire regarding their family history and medical health as well as a tube of blood for laboratory testing. They also agreed to complete a follow up survey every 5 years for the next 25 years in order to look at new diagnoses and diseases. For each patient, an aliquot of blood as well as the buccal swab were both used to sequence each patient to 40X coverage as well as perform comparative genomic hybridization to a sequenced and assembled reference genome. All genomes were cataloged for mutations included insertions/deletions, single nucleotide polymorphisms, and gene duplication. The survey included questions about all of the following: diabetes, hypertension, malignancy (specifically of breast, lung, colon, prostate, kidney and/or brain), infections (including frequency and specifically for mononucleosis, ear infections, head colds, urinary tract infections, toenail infections, persistent/excessive acne), diet, and exercise habits. All of the subjects were counseled to use a free pedometer (provided by the study team) which was connected to the internet and report their daily activity, which was monitored by the study.

After 10 years (3 total surveys), a manuscript was published by a non-competing group in a mouse model showing that a specific mutation of pyruvate dehydrogenase kinase 4 (PDK4) caused a massive decrease in mouse activity as well as obesity in mice. You propose to look at the pedometer data of the study's subjects' activity to see if there is an association with fewer steps and mutations in PDK4. Your PI, however, thinks that such an association may be polygenetic (or even spurious in the mouse) and the entire genome should be examined in the context of all of the data.

Questions:

1. How would you go about investigating any potential associations in your data set?
2. What statistical considerations are important in thinking about this question?
3. How should the pedometer data be parsed for the analysis?