Group 5 – Break Out Session

Developing a collaborative genomic sequencing (and other -omics?) strategy

John Danesh

Hakon Hakonarson

Group 5 – Break Out Session

- What kinds of sequencing or other -omic data would be useful for individual cohorts?

- What aspects of a collaborative sequencing strategy, in addition to low cost, would facilitate obtaining and sharing these data?

- What methods/tools are optimal for data harmonization across different sites to address platform diversity/uniformity, batch effects and related issues

Group 5 – Break Out Session

Group 5 had representatives from multiple cohorts across 9 countries:

Chile

Japan

UK

Sweden

China

Canada

Iran

Brazil

USA

Group 5 – Break Out Session

What are the key questions we can only address though large-scale international collaboration:

- **Genomic Diversity** – to do these studies properly at the scale needed
- **Exposure Diversity** – gene/environment interaction studies
- **Migrant Studies** – potentially extremely valuable to understand the role of the environment

Rare Diseases

We need the numbers to be successful

Human KO studies – null variants and homozygous CNVs

PCSK9-like opportunities

Unmet clinical need/potential drug repurposing opportunities

"Drugable genome" focus – to attract pharma

Address **global problems** such as Obesity, exposures to toxic substances and alike

Group 5 – Break Out Session

We can drive sequencing and other omics costs down through processing of millions of samples

Large scale GWAS and follow-up PheWas, including across-disease cohort analysis remains and area of interest (analyses across multiple psychiatric, neurodevelopmental, autoimmune, cancer etc)

Alcohol-related diseases – effect sizes on health vary considerably between countries

Question-driven approach, milestones and goals

- Functional omics studies (KO) across different populations
- Role of microbiome across different ethnicities

Having centralized study co-ordinator(s) cataloging cohort information so study investigators can pop a question and they identify the lead informative cohorts to address

Group 5 – Break Out Session

Question 1: What kinds of sequencing or other -omic data would be useful for individual cohorts?

General consensus that most valuable data would be whole genome sequencing and that WGS data files (TBD) would be shared – realizing will probably not happen in near future

WES did not get much vote (identifying new variants in known GWAS genes)

Initially – leveraging existing and generating and sharing new SNP-array genotyping data across all samples would be of interest/value to the group (**$50-100M**)

- still a lot left to discover by GWAS/CNV analysis (Brazil, Chile, Iran and multiple other countries still untapped)
- this would need WGS data for some populations (few hundred) to establish the correct SNP-array content to capture population variability (GSA, Affy-biobanking not enough – low cost customized arrays required – include known KO variants)
- homozygous deletions (human KO) to be found at the fraction of the cost of exome (in CAG biobank alone, in an analysis of 100k, 2,700 HD (human KO) were uncovered)

Group 5 – Break Out Session

Question 2: What aspects of a collaborative sequencing strategy, in addition to low cost, would facilitate obtaining and sharing these data?

- Leverage existing efforts presented yesterday on data file harmonization (GA4GH and others)
  - Efforts aimed at minimizing batch effects
  - Availability to cloud
  - Sharing smallest files possible (vcf initially) compressed CRAM files to follow
- Low income countries needs and multiple other sites low in funding
  - how can we make them competitive so they have incentive to participate
  - **Funding** appears to be the most important incentive (academia/industry)
- Some concerns over population-specific rare variants for ID
- Group felt it would be important to have access to limited numbers of RNAseq, proteome, epigenetics data, microbiome data, metabolome data

Group 5 – Break Out Session

Question 3: What methods/tools are optimal for data harmonization across different sites to address platform diversity/uniformity, batch effects and related issues

- Exomes have to much variability (wet-lab variability on top of seq data)
- Imputed SNP-array data proven to work across multiple continents
- WGS data – data file harmonization at US sequencing centers proven effective (processed at different sites and then merged/harmonized)
  - TOPMed with >100k WGS data files across multiple centers
  - Other programs
- Coordinate across working groups
- Phenotyping data – disease diagnosis; group felt no issues with sharing
- Metabolome data – if can be generated cost-effectively ($25/sample) then potential great value in sharing

At a recent international meeting a colleague whispered to a countryman sitting next to him – I can assure you that the US leaders will get it right ...after they have tried everything else...