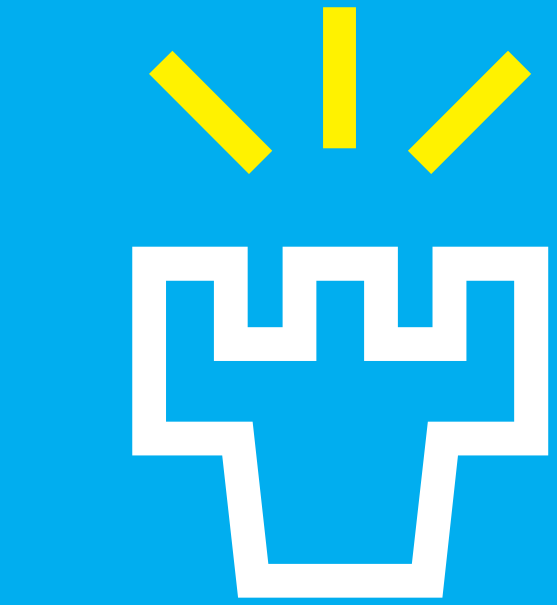
Challenges and solutions in human genetics research – case CSC

7.10.2020

Johannes Kettunen



UNIVERSITY
OF OULU

Background

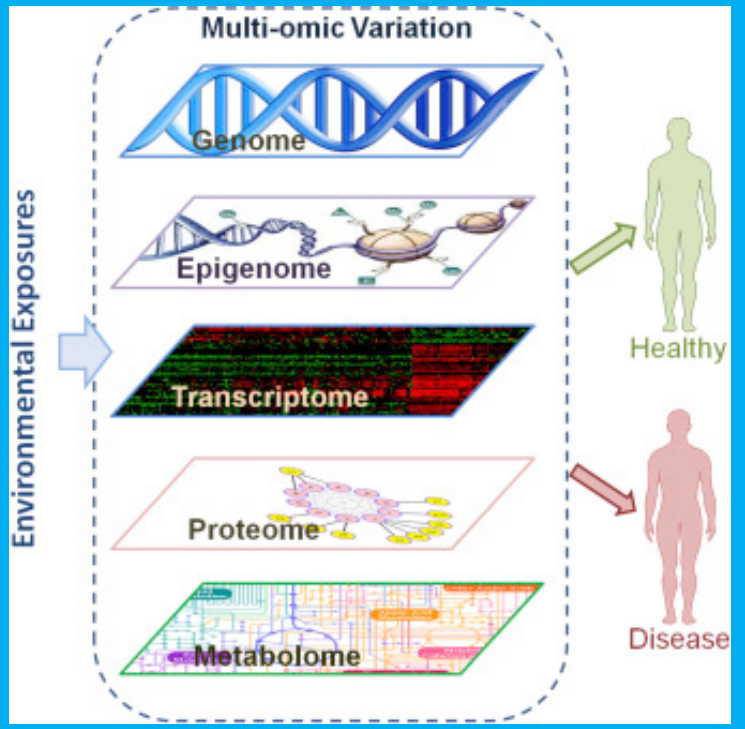
Past solutions

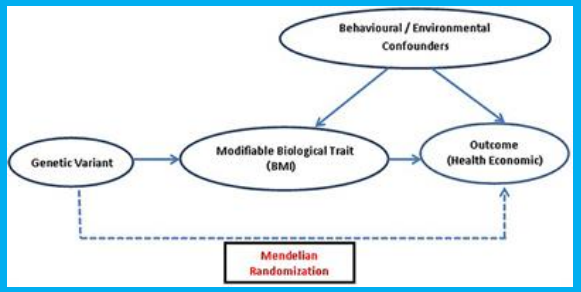
Challenges

Future prospects

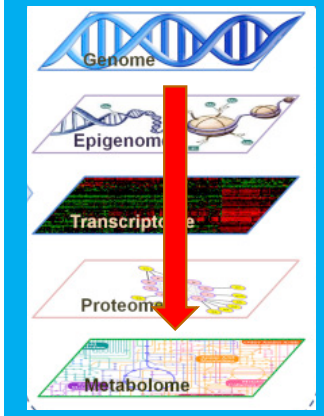
Group Kettunen, systems medicine

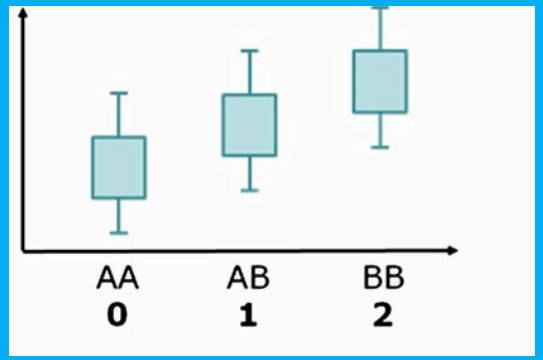
- **Epidemiology, biomarkers**
  - Risk prediction: metabolic biomarkers for all-cause mortality etc
- **Genetic epidemiology**
  - Biomarkers: Metabolomics, inflammation
  - Diseases: FINNGEN
- **Causal inference**
  - Utilizing genetics in triangulation to understand causality of biomarkers, and to proxy trial effects





Background – genome to metabolome

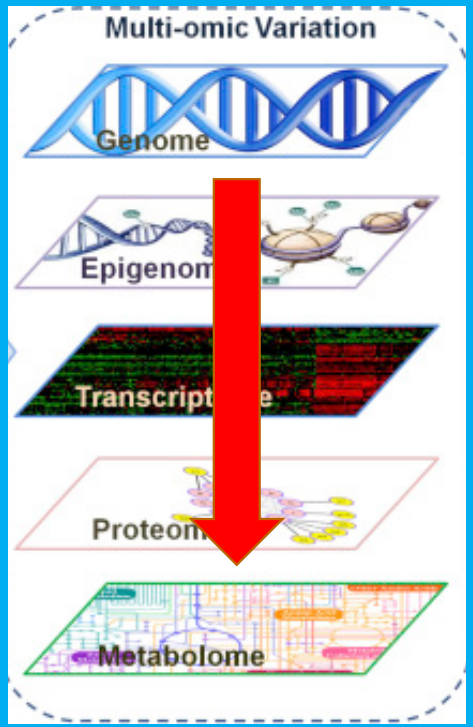




Computational burden

- **Each study cohort has genomic information for all participants for whole genome**
  - 20 million genomic variables for each participant
  - Typically each cohort has thousands of participants
  - Does not change, serves as basis
- **All other variables are correlated against genomic markers**
- **Very simple models**
- **I/O intensive**
- **Parallelization**

Past solutions and need for special settings



NB! our data does not have names or personal details attached

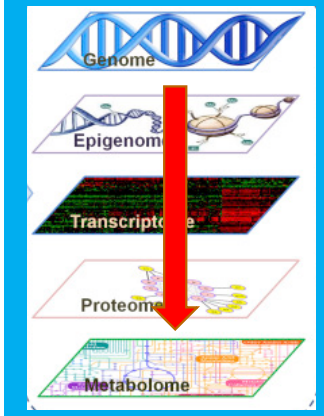
Past (before GDPR)

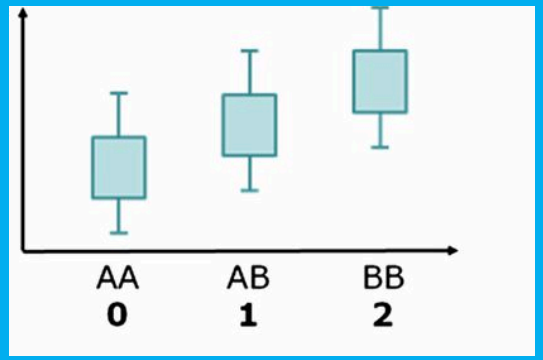
- Parallel computing
- Using similar environments to CSC puhti
- Thousands to tens of thousands of cores

GDPR

- Law did not change, interpretation did
- One measurement (LDL cholesterol concentration) is personal information (pseudonymization)
- Puhti and similar solutions not in our reach

Usual workflow



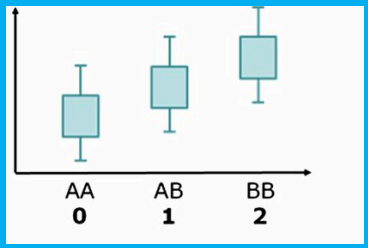


- **Genome-wide association study of serum lipids**
  - NFBC66 cohort, 3300 participants from born in Northern Finland in 1966
  - Whole genome information for all
  - 800 serum lipids measured for all participants from 46-year follow-up survey
  - Linear regression of genetic variables against all 800 serum lipid concentrations
- **Generates 800 output files with 20 million lines**
- **1 process takes approx 3 hours**
- **800*22 processes = 17600 ~ 53000 cpu hours ~ 6 cpu years**

In case of 1 cohort:
NFBC cohort, 3300 participants

Each 20 million lines

20 million markers



Lipid 1 concentrations → Output for lipid 1 concentrations

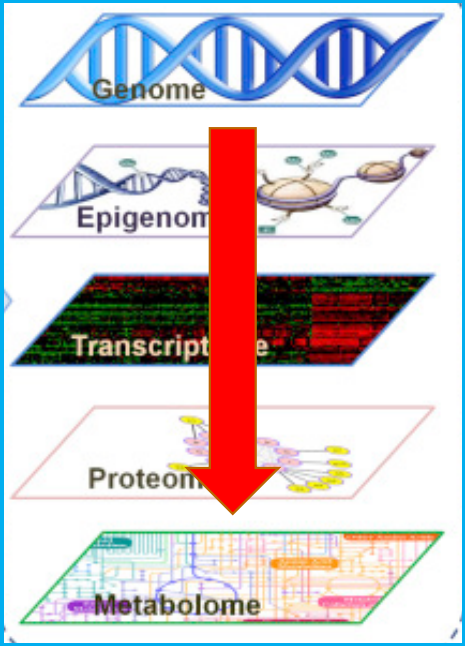
Lipid 2 concentrations → Output for lipid 2 concentrations

Lipid 3 concentrations → Output for lipid 3 concentrations

Lipid 4 concentrations → Output for lipid 4 concentrations

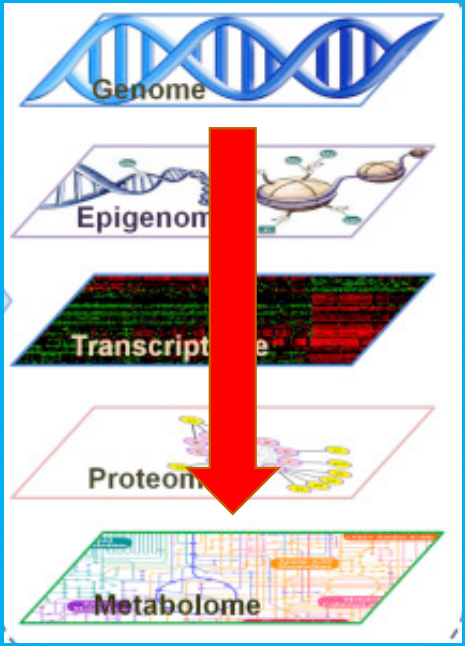
Lipid 880 concentrations → Output for lipid 880 concentrations

Usual workflow - challenges



- **6 years, one study?**
- **Usually several cohorts**
- **Other linear models?**
- **Another layer of complexity, sensitive data and ePouta**
  - Data has to be analyzed in secure environment, standard CSC puhti is not available
  - ePouta has solved our needs
  - Data in secure environment, and sensitive personal data can be analyzed
- **Note that the output files are not sensitive data, the starting files are**

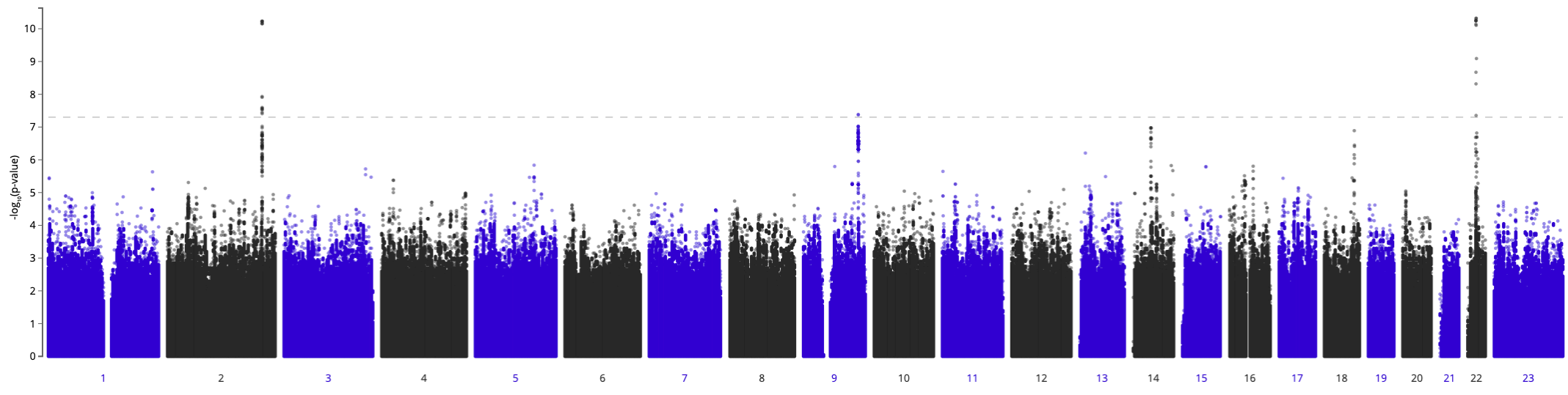
Usage, present and future



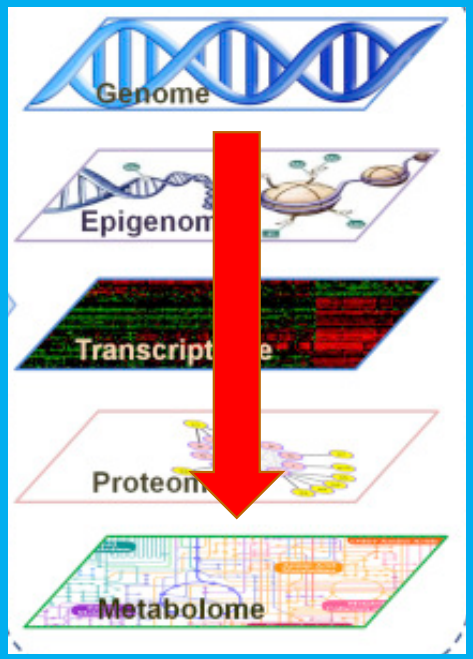
- We have used 86 CPU **years** of computing at CSC in various projects during 5 years
- These projects have generated data that requires 20 Tb active use storage
- We have new data coming and we foresee that we will use same amount of core time and storage in the coming few years
- CSC now provides a platform that fits our current needs well

Usual workflow – to final result

- **Several cohorts**
  - Quality control
  - Meta-analysis
- **Final results**



Future prospects



- **Current needs are well met by CSC**
  - Sensitive data analysis and parallelization of thousand processes in ePouta keeps us working
  - These types of projects will run for years to come
  - Note, no need for supercomputing

- **LUMI in genomics?**
  - There will very likely be need to change from very simple models to optimized models
  - Using genetic information in disease risk prediction
  - Analyzing large numbers of genetic markers simultaneusly in more complex models for optimization

Cetainly a task where LUMI will be a great asset

Thank you for your attention!


UNIVERSITY
OF OULU