Precision Medicine: Aligning Research Standards with Clinical Standards
1st July 2019

The CDx R&D Perspective

Mike Furness

Pistoia Alliance
Outline of Talk

• NGS in Discovery
• Growth of CDx
• NGS for CDx
• NGS in medical diagnostics
• NGS in the clinical testing
• Standards in NGS sequencing
• Standards in research and clinic
• Benefits of aligning both going forward
• Examples of other related initiatives
• My goals for today
Changes in NGS Technology – July 2016

NGS Tech 2018
(All data accessed 13 March 2018)

https://lnkd.in/gdTtFmm

https://emea.illumina.com/systems/sequencing-platforms/novaseq/specifications.html


https://nanoporetech.com/products#modal=comparison
http://en.mgitech.cn/product/detail/139.html

https://emea.illumina.com/systems/sequencing-platforms/novaseq.html


NGS Tech 2019

(All data accessed March 2019)
NGS Tech 2019

(accessed June 2019)

Gigs per $1000 lower and Gigs per $1000 higher

Speed of Sequencing - 2018

Time taken to sequence 1 Petabase by Wellcome Sanger Institute

Wellcome Sanger Institute celebrates reading more than 5,000,000,000,000,000,000 bases of DNA

1Pb = 8-10,000 human genomes (assuming Human whole genome = 100-120 Gb)
Large Scale Genomic Computing – Incyte 2000

1,106 Pentium
425 Alpha
64 Sparc
1,634 CPUs

135 Tb Storage
3952 ft² Data Centre

The Linux “Farm”
3,500 machines by March 2000

Integrating Genomics and Proteomics to understand Cellular Physiology and Drug Action (2000)
L.M.Furness. “Pharmacogenomics and Pharmacogenetics Euroconference”. Pasteur Institute, Paris
Large Scale Genomic Computing – Amazon 2014

Amazon has 11 cloud regions across the world, said James Hamilton, an Amazon distinguished engineer, during a presentation at re:Invent. Each region has multiple sets of data centers, and there are 28 total sets across the world. Each of those has one or more data centers, with a typical facility containing 50,000 to 80,000 servers. A conservative estimate puts Amazon over 1.5 million servers globally. Lydia Leong, an analyst at research firm Gartner, puts it at 2 million or more.

By comparison, Rackspace Hosting has a little over 100,000 servers spread across six data centers. Google has three regions with eight total sets, and Microsoft has 17 regions. Got all that? Last year, Steve Ballmer, then Microsoft’s CEO, said the company had over a million servers within its data center infrastructure and that Google had even more.

Amazon’s cloud could soon get even bigger. Hamilton told me that he saw no reason why Amazon couldn’t eventually have a data center in every U.S. state if companies adopt cloud computing as enthusiastically as people predict.

14 Nov 2014
## Use of Next-Generation Sequencing in the regulated domain of drug development

Next-generation sequencing (NGS) has moved from the realm of research into the regulated domains of drug development, diagnostic development, and clinical decision-making. This article summarizes some of the technical and regulatory challenges posed by these technologies and the efforts being made to address them.

### Algorithms

A typical NGS data processing pipeline includes the following steps:

1. The sequencing platform conducts image production and the assignment of base calls is performed.

### Results

<table>
<thead>
<tr>
<th>Project</th>
<th>(Target) Cohort Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AstraZeneca 2M Genomes Project</td>
<td>2,000,000</td>
</tr>
<tr>
<td>Ancestry.com</td>
<td>1,400,000</td>
</tr>
<tr>
<td>23andMe</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Million Veteran Program</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Precision Medicine Initiative</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Korea Biobank Project</td>
<td>618,958</td>
</tr>
<tr>
<td>European Network for Genetic and Genomic Epidemiology (ENGAGE)</td>
<td>600,000</td>
</tr>
<tr>
<td>Resilience Project</td>
<td>589,306</td>
</tr>
<tr>
<td>China Kadoorie Biobank Repository</td>
<td>512,000</td>
</tr>
<tr>
<td>deCode Genetics</td>
<td>500,000</td>
</tr>
<tr>
<td>Kaiser Permanente: Genes, Environment, and Health (RPGEH) Repository</td>
<td>500,000</td>
</tr>
<tr>
<td>UK Biobank Repository, Consortium</td>
<td>500,000</td>
</tr>
<tr>
<td>Regneron/Kaiser Permanente MyCode® Community Health Initiative Repository</td>
<td>250,000</td>
</tr>
<tr>
<td>French Genome Project</td>
<td>235,000</td>
</tr>
<tr>
<td>Vanderbilt’s BioVU Repository</td>
<td>215,000</td>
</tr>
<tr>
<td>BioBank Japan Repository Specimens</td>
<td>200,000</td>
</tr>
<tr>
<td>Leiden Open Variation Database (LOVD) Repository</td>
<td>170,000</td>
</tr>
<tr>
<td>Psychiatric Genomics Consortium (PGC)</td>
<td>170,000</td>
</tr>
<tr>
<td>100K Wellness Project</td>
<td>100,000</td>
</tr>
<tr>
<td>Actionable Cancer Genome Initiative (ACGI) Data-Sharing Project</td>
<td>100,000</td>
</tr>
<tr>
<td>East London Genes &amp; Health</td>
<td>100,000</td>
</tr>
<tr>
<td>Genome Asia 100K Consortium</td>
<td>100,000</td>
</tr>
<tr>
<td>Genomics England</td>
<td>100,000</td>
</tr>
<tr>
<td>Saudi Human Genome Program</td>
<td>100,000</td>
</tr>
<tr>
<td>Turkish Genome Project</td>
<td>100,000</td>
</tr>
<tr>
<td>Exome Aggregation Consortium (ExAC)</td>
<td>60,706</td>
</tr>
<tr>
<td>Electronic Medical Records and Genomics (eMERGE) Network Repository</td>
<td>55,028</td>
</tr>
</tbody>
</table>

---

From Biomarker to CDx

Steps from Biomarker to Diagnostic/Outcome Measure

- **Intended use?**
- **Early Development**
  - Drug Mechanism Readout (TE/PD)
- **Late Development**
  - Diagnostic / Outcome Measure
  - **Regulatory Process**
    - Approved Diagnostic
    - Accepted Outcome measure
  - **Clinical Qualification**
    - Diagnostic
    - Outcome measure (surrogate)
  - **Assay Validation**
    - Prototype assay to Assay “lock down”
    - Assay manufacturing (e.g. kit)
    - Assay standardization
  - **Clinical Application**
    - “Fit for purpose”
  - **Fit for Purpose Assay**
    - Custom developed assay
  - **Exploratory Biomarker**
    - Exploration of candidate biomarker
  - **Biomarker Identification**
    - Hypothesis driven or Un-biased
    - Modality? / Invasive or Non-invasive?

[http://www.johanluthman/151007170031/lva1/app6892](http://www.johanluthman/151007170031/lva1/app6892)
“.....in a survey by McKinsey in 2007 indicated that, on average, 30 to 50 percent of drugs in development have an associated biomarker program, and suggested this number was likely to increase.... while the most aggressive players have biomarker programs for 100 percent and companion diagnostics for 30 percent or more of their compounds, the average company has far fewer (30 to 50 percent and less than 10 percent respectively...”
What will drive growth of your NGS usage over the next two to three years? (clinical labs)

- Liquid biopsy cancer testing: 59%
- Increased breadth of testing (e.g., moving from single gene to gene-panels, gene-panels to exome, exome to whole genome): 52%
- Tissue-based cancer testing: 43%
- Companion diagnostic testing: 39%
- Germline disease testing: 33%
- Sequencing for pharma: 26%
- Direct-to-consumer testing: 20%
- Other (please specify): 8%
- We do not expect growth of NGS usage: 1%

N=90

% respondents who selected this as a factor

Source: William Blair and GenomeWeb 2017 NGS Survey
Growth of CDx usage

Figure 3: Companion diagnostic deals signed by year (A) and stage of companion diagnostic deal (B).

Pharmgenomics Pers Med. 2015; 8: 99–110
Figure 1 US Food and Drug Administration-approved companion diagnostic drugs (2012).

Notes: *Other includes respiratory, systemic hormones, dermatologicals, alimentary tract and metabolism, nervous system, and various.
# List of Cleared or Approved Companion Diagnostic Devices (In Vitro and Imaging Tools)

<table>
<thead>
<tr>
<th>Diagnostic Name</th>
<th>PMA/ 510(k)/ HDE</th>
<th>Diagnostic Manufacturer</th>
<th>Trade Name (Generic) - NDA/BLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRACAnalysis CDx</td>
<td>P140020/S016</td>
<td>Myriad Genetic Laboratories, Inc.</td>
<td>Breast Cancer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Lynparza (olaparib) - NDA 208558</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Talzenna (talazoparib) - NDA 211651</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ovarian Cancer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Lynparza (olaparib) - NDA 208558</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Rubraca (rucaparib) - NDA 209115</td>
</tr>
<tr>
<td>therascreen EGFR RGQ PCR Kit</td>
<td>P120022/S018</td>
<td>Qiagen Manchester, Ltd.</td>
<td>Non-small cell lung cancer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Iressa (gefitinib) - NDA 206995</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Gilotrif (afatinib)- NDA 201292</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Vizimpro (dacomitinib)- NDA 211288</td>
</tr>
<tr>
<td>cobas EGFR Mutation Test v2</td>
<td>P120019/S019</td>
<td>Roche Molecular Systems, Inc.</td>
<td>Non-small cell lung cancer (tissue and plasma)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Tarceva (erlotinib) - NDA 021743</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Tagrisso (osimertinib) - NDA 208065</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Iressa (gefitinib) - NDA 206995</td>
</tr>
<tr>
<td>PD-L1 IHC 22C3 pharmDx</td>
<td>P150013/S011</td>
<td>Dako North America, Inc.</td>
<td>Non-small cell lung cancer, gastric or gastroesophageal junction adenocarcinoma, cervical cancer, and urothelial carcinoma</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Keytruda (pembrolizimab) - BLA 125514</td>
</tr>
<tr>
<td>Abbott RealTime IDH1</td>
<td>P170041</td>
<td>Abbott Laboratories</td>
<td>Acute myeloid leukemia</td>
</tr>
</tbody>
</table>

AstraZeneca has collaborated with Roche to develop the cobas® EGFR Mutation Test v2 as the companion diagnostic for AZD9291 (Osimertinib).

an EGFR-TKI, a targeted cancer therapy, designed to inhibit both the activating, sensitising mutations (EGFRm), and T790M, a genetic mutation responsible for EGFR-TKI treatment resistance. Nearly two-thirds of NSCLC patients who are EGFR mutation-positive and experience disease progression after being treated with an EGFR-TKI develop the T790M resistance mutation, for which there have been limited treatment options.....
Companion Diagnostics for Cystic Fibrosis

Majority of People With CF Carry a CFTR Mutation of Known Molecular and Clinical Consequence

Triple Combination Has Potential to Treat 90% of People With CF

Precision Medicine Growth Areas

Driven by Multi-omics diagnostics and CDx biomarker-based targeted therapeutics is Precision Medicine expected to reach $134.24 Bn by 2025

Core Precision Medicine Segment: Growth Potential by Targeted Therapeutic Areas and Technologies, Global, 2016–2025

<table>
<thead>
<tr>
<th>Market Growth / Potential</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short Term</td>
</tr>
</tbody>
</table>

- Targeted Therapeutics (Rx)
  - CDx Biomarkers
  - Oncology
  - non-Oncology

- By Omics Technologies and Solutions
  - MDx
  - POCT
  - Biobanks
  - Imaging Dx
  - Dx Informatics

Key: ◀ Low ▶ Medium ▲ High

Source: Frost & Sullivan
Growth of NGS in CDx

02 Jan 2019
Next-generation sequencing will shape the companion diagnostics market in the next five years

The global companion diagnostics (CDx) market will be shaped by next-generation sequencing (NGS) technologies over the next five years, according to GlobalData, a leading data and analytics company.

The company’s latest report: ‘Biomarkers and Companion Diagnostics in Oncology’ states that one of the most important drivers of the market for CDx will be the increased emphasis on cost-effective healthcare.

Fern Barkalow, PhD, Oncology and Hematology Director at GlobalData, says: “Measuring many biomarkers in one test using NGS panels will become increasingly important as targeted therapies continue to become embedded in treatment algorithms for many different types of cancer.

“This technology is likely to transform the companion diagnostics market in the next five years due to its high efficiency and gradually decreasing cost. There are currently five marketed next-generation sequencing CDx in the US, the first of which was approved in 2016.”

GlobalData’s analysis also finds that the US leads the way in terms of clarity on CDx regulation.

Barkalow continues: “The FDA’s guidance is stringent regarding analytical and clinical validity, but it does facilitate CDx market entry by providing multiple marketing authorization pathways, and the agency is pursuing tighter regulation of laboratory-developed tests. CDx regulation in Japan is less advanced than in the US in certain aspects, while CDx were only recently formally defined in Europe and remain undefined in China in terms of regulatory processes.”

https://www.globaldata.com/next-generation-sequencing-will-shape-the-companion-diagnostics-market-in-the-next-five-years/
Prospectively Stratify participants entering drug trial

Sub-population with target phenotype and/or genotype
AbbVie, Genomics Medicine Ireland and WuXi NextCODE Announce Landmark Population Genomics Alliance

Jan 09, 2017

Effort will sequence the genomes of 45,000 participants from across Ireland to identify novel targets and advance the clinical development of better treatments for a range of serious diseases

- Focus is on key AbbVie therapeutic areas including oncology, neuroscience and immunology
- Genomics Medicine Ireland will partner with Ireland’s leading clinicians and researchers and propel bioscience innovation in Ireland
- WuXi NextCODE’s integrated genomics platform will be used to organize and mine a secure, scalable database of whole genome sequence and medical data to power novel disease insights

NORTH CHICAGO, ILL.; DUBLIN; CAMBRIDGE, Mass., 9 January 2017 – AbbVie (NYSE: ABBV), a global biopharmaceutical company, life-sciences startup Genomics Medicine Ireland Limited (GMI), and WuXi NextCODE, the global contract genomics organization, today announced the launch of a long-term strategic alliance to conduct population genomics research in Ireland aimed at advancing the discovery and development of novel therapeutic approaches to a range of serious diseases. The 15-year collaboration will focus on major chronic diseases within oncology, neuroscience and immunology that affect hundreds of thousands of people in Ireland and hundreds of millions worldwide. Financial terms were not disclosed.
Stratify participants in Biobank/Cohort

Trial 1

Trial 2

stratify
Retrospectively Stratify participants after drug trial

80% response with marker

20% response without marker

30% response

FAIL

PASS with CDx
Drug Discovery & Development – NGS & Omics

Clinical Diagnostics Regs
- GLP
- CE-IVD
- ISO 9001
- ISO 27001/2
- ISO 13485
- ISO 15189

R&D Regs
- GLP
- ISO 9001
- ISO 27001/2

Clinical Trials Regs
- GCP
- CE-IVD
- ISO 9001
- ISO 27001/2
- ISO 14155
Companion Diagnostics (CDx) for Therapeutics

Discovery Regs

Clinical Regs

Parallel Development of Drug and Diagnostic

CDx Development

Clinical Diagnostic Regs

Front. Oncol., 16 May 2014  https://doi.org/10.3389/fonc.2014.00105
Precision medicine needs pioneering clinical bioinformaticians

Gonzalo Gómez-López, Joaquín Dopazo, Juan C. Cigudosa, Alfonso Valencia and Fátima Al-Shahroul

Corresponding author. Fátima Al-Shahroul, Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro, 3. ES-28029 Madrid, Spain. Tel.: +34 91 732 80 00; E-mail: falshahroul@cnio.es
Some other activities in this Area

- EMA Concept paper on development and lifecycle of personalised medicines and companion diagnostics
- July 15, 2016, FDA released the draft guidance, "Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product."
- https://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm301431.htm
- SAFE-T (Safer and Faster Evidence-based Translation) http://www.imi-safe-t.eu/
- https://hive.biochemistry.gwu.edu/htscrs/ There are several synergistic efforts which can be leveraged to the development of HTS computational standards.
  - precisionFDA
  - NIH BD2K
  - The Human Variome Project
  - Global Alliance for Genomics and Health
  - Global Biological Standards Institute
  - The Common Workflow Language (CWL)
  - BioCompute Objects
  - Genome in a Bottle
  - myExperiment.org
  - BioSharing
  - Reproducible Bioinformatics Project (RBP)
One reference genome is not enough

Construction of JRG (Japanese reference genome) with single-molecule real-time sequencing


Chromosome-scale assembly comparison of the Korean Reference Genome KOREF from PromethION and PacBio with Hi-C mapping information

Hui-Su Kim, Sungwon Jeon, Changiae Kim, Yeon Kyung Kim, Yun Sung Cho, Asta Blazyte, Andrea Manica, Semin Lee, Jong Bhak

doi: https://doi.org/10.1101/674804

3.5KJPNv2, An allele frequency panel of 3,552 Japanese Individuals


doi: https://doi.org/10.1101/529529

A personal and population-based Egyptian genome reference

Inken Wohlers, Axel Künßter, Matthias Munz, Michael Olbrich, Anke Fährnich, Caixia Ma, Misa Hirose, Shaaban El-Mosallamy, Mohamed Salama, Hauke Busch, Saleh Ibrahim

doi: https://doi.org/10.1101/681254

Insights into human genetic variation and population history from 929 diverse genomes

Anders Bergström, Shane A. McCarthy, Ruoyun Hui, Mohamed A. Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kammar, Hélène Blanché, Jean-François Deleuze, Howard Cann, Swapnan Mallick, David Reich, Manjinder S. Sandhu, Pontus Skoglund, Aylwyn Scally, Yali Xue, Richard Durbin, Chris Tyler-Smith

doi: https://doi.org/10.1101/674986
Essential guidelines for computational method benchmarking

1. Define the purpose and scope of the benchmark.
2. Include all relevant methods.
3. Select (or design) representative datasets.
4. Choose appropriate parameter values and software versions.
5. Evaluate methods according to key quantitative performance metrics.
6. Evaluate secondary measures including computational requirements, user-friendliness, installation procedures, and documentation quality.
7. Interpret results and provide recommendations from both user and method developer perspectives.
8. Publish results in an accessible format.
9. Design the benchmark to enable future extensions.
10. Follow reproducible research best practices, by making code and data publicly available.

Fig. 3 Example of an interactive website allowing users to explore the results of one of our benchmarking studies [27]. This website was created using the Shiny framework in R.
CDx / NGS & Regulation

Unmet Needs:
With the burgeoning use of genomics in drug discovery and development, and its potential to drive new technological and regulatory challenges arise.

Biomarkers are being identified which can be used to identify patients and to help stratify them into relevant populations from clinical studies or biobank populations. Equally, some of these biomarkers can help identify patient subgroups in drug development regimes, either by defining pharmacogenomic markers which identify fast or poor metabolizers, or by defining specific therapeutic agents associated with a biomarkers presence or absence (e.g., EGFR in cancer).

The biomarkers that define populations for prescribing have typically been called Companion Diagnostics (CDx) and these are being used to inform the choice of most treatments and have had clear regulatory definitions. In the USA, the FDA also has a program called Complementary Diagnostics. These have a set of regulations relating to IVDs (In Vitro Diagnostic) test processes being updated by both the FDA and the EMA.

As large genome projects are being launched in pharmaceutical companies (e.g. Astellas’ 50k cancer project and Regeneron’s 250,000 exomes project with Geisinger Health), large amounts of data are being collected. While the data is being collected in a clinical environment, it is currently being used in a discovery project.

Proposal/Value Proposition:
We would like to bring together members of the academic and biotech companies, along with regulatory agencies, to look at ways that the data collection and analyses could be aligned across domains to enable the data from these large projects to be used retrospectively as clinical evidence.

Project Type: Standards
Project Domain: Translational Science

Team
Interest Group: Andrea Haworth, Congenica Limited
Carmen Nitsche, Pistoia Alliance
John Wise, Pistoia Alliance, Inc.
Jolyon Holdstock, Oxford Gene Technology Ltd.
Keith Nangre, Pistoia Alliance
Mariangela Galante, Illumina
Michael R. Crusoe, Common Workflow Language p... 
Neil Adams, Illumina
Wendy Tindsley, Aigenpulse
Will Spooner, Genomics England
Yvonne Wilding, AstraZeneca
Manfred Remer, RC

Supporting Documents
1. Mike Furness - Pistoia Alliance Workshop Presentation 2018.pdf
4. Steven Lee - Pistoia Alliance Workshop Presentation 2018.pdf
5. Liz Harrison - Pistoia Alliance Workshop Presentation 2018.pdf
7. 2018_04_27 Pistoia Alliance CDx NGS Regulation Workshop Report.pdf
More information

Please see the problem statement on the Pistoia Alliance Interactive Project Portfolio Platform at:
https://ip3.pistoiaalliance.org/subdomain/main/end/node/1852

The Pistoia Alliance webinar on “Faster Safe Companion Diagnostics (CDx) by Aligning Discovery & Clinical Data in the Regulatory Domain” was held on Tuesday 6th March. A recording of the webinar is available at:
http://www.pistoiaalliance.org/pistoia-alliance-debates-webinar-series/

Pistoia Alliance Publications on:
- Use of Next-Generation Sequencing in the regulated domain of drug development
- The positive impacts of Real-World Data on the challenges facing the evolution of biopharma

To explore this challenge, the Pistoia Alliance has created a short questionnaire that only takes a few minutes to complete. We would be very grateful if you would complete the questionnaire. It can be found at:
https://www.surveymonkey.co.uk/r/YQ8L2H3
## Workshop – AGENDA
11th April 2018, Royal Society of Chemistry, London

<table>
<thead>
<tr>
<th>Morning Session</th>
<th>Afternoon Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00</td>
<td>Registration and Coffee</td>
</tr>
<tr>
<td>10:00</td>
<td>Overview of the day and planned outcomes (Mike Furness, Pistoia Alliance)</td>
</tr>
<tr>
<td>10:30</td>
<td>Overview of current large-scale NGS data projects (John Whittaker, VP Target Discovery, GSK)</td>
</tr>
<tr>
<td>11:00</td>
<td>Coffee</td>
</tr>
<tr>
<td>11:30</td>
<td>How good is your Next Generation Sequencing? (Simon Patton, European Molecular Genetics Quality Network (EMQN))</td>
</tr>
<tr>
<td>12:00</td>
<td>What are Companion Diagnostics and the regulatory requirements around them? (Stephen Lee, Senior Regulatory Policy Manager – IVD Devices Division, MHRA)</td>
</tr>
<tr>
<td>12:30</td>
<td>Notified Bodies – what are they and what do they do? (Liz Harrison, Technical Team Manager, BSI Group)</td>
</tr>
<tr>
<td>13:00</td>
<td>Lunch</td>
</tr>
<tr>
<td>14:00</td>
<td>Industry Regulatory – Perspectives Challenges of NGS Companion Diagnostics Development (Stewart McWilliams, VP Quality &amp; Regulatory Affairs, Almac)</td>
</tr>
<tr>
<td>14:30</td>
<td>Cross-Expertise Delegate Breakout Groups (chair: Nadia Anwar, Translational Research Solution Consultant, Oracle)</td>
</tr>
<tr>
<td>15:15</td>
<td>Coffee</td>
</tr>
<tr>
<td>15:45</td>
<td>Plenary session: Collect ideas, identify next steps and define targeted outcomes</td>
</tr>
<tr>
<td>16:30</td>
<td>Networking Reception</td>
</tr>
<tr>
<td>17:30</td>
<td>Close of Workshop</td>
</tr>
</tbody>
</table>
My Goals for Today: points to think about…

• Identify key challenges to address
  • Standards within technologies?
  • Standards across technologies?
  • Reference data sets?
  • Benchmarking?
  • New challenges - Automation/AI/ML?

• Identify possible ways to address these challenges
  • Where can we effect ‘standards’?
  • How/Who can implement these?
  • Develop recommendations or best practices across/from industry?

• Identify other people who can help us move forward
  • National and International Initiatives?
  • Regulatory and Industry Bodies?
Analyzing the heterogeneity of rule-based EHR phenotyping algorithms in CALIBER and the UK Biobank

Spiros Denaxas, Helen Parkinson, Natalie Fitzpatrick, Cathie Sudlow, Harry Hemingway

doi: https://doi.org/10.1101/685156

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract

Electronic Health Records (EHR) are data generated during routine interactions across healthcare settings and contain rich, longitudinal information on diagnoses, symptoms, medications, investigations and tests. A primary use-case for EHR is the creation of phenotyping algorithms used to identify disease status, onset and progression or extraction of information on risk factors or biomarkers. Phenotyping however is challenging since EHR are collected for different purposes, have variable data quality and often require significant harmonization. While considerable effort goes into the phenotyping process, no consistent methodology for representing algorithms exists in the UK. Creating a national
Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions

Sulev Reisberg, MSc¹,²,³, Kristi Krebs, MSc⁴,⁵, Maarja Lepamets, MSc⁴,⁵, Mart Kals, MSc⁶, Reedik Mägi, PhD⁴, Kristjan Metsalu, MSc⁴, Volker M. Lauschke, PhD⁶, Jaak Vilo, PhD¹,²,³ and Lili Milani, PhD⁴,⁷

Fig. 1 Pipeline for extracting pharmacogenetically relevant alleles from existing genotyping data. Panel (a) depicts the different data sets, their overlap (Venn diagram), and how the data were processed. Panel (b) zooms into the detection of star alleles according to specific definition tables. ES exome sequencing, GS genome sequencing, GSA Global Screening Array, OMNI HumanOmiExpress.
Establishing reference samples for detection of somatic mutations and germline variants with NGS technologies

Li Tai Fang1*, Bin Zhu2*, Yongmei Zhao3*, Wanqiu Chen4, Zhaowei Yang4,31, Liz Kerrigan2, Kurt Langenbach5, Maryellen de Mars5, Charles Lu6, Kenneth Idler6, Howard Jacob6, Ying Yu7, Luyao
Best practices for benchmarking germline small-variant calls in human genomes

Table 1 | Contingency table describing the GA4GH definitions of TP, FP, FN, FP.AL, FP.GT, and unknown (UNK)

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Ref/ref</th>
<th>Ref/var1</th>
<th>Var1/var2</th>
<th>Var1/var1</th>
<th>Outside bed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ref/ref</td>
<td>-</td>
<td>FN</td>
<td>FN</td>
<td>FN</td>
<td>-</td>
</tr>
<tr>
<td>Ref/var1</td>
<td>FP</td>
<td>TP</td>
<td>FP.GT</td>
<td>FP.GT</td>
<td>UNK</td>
</tr>
<tr>
<td>Ref/var2</td>
<td>-</td>
<td>-</td>
<td>FP.AL</td>
<td>FP.AL</td>
<td>-</td>
</tr>
<tr>
<td>Ref/var3</td>
<td>-</td>
<td>-</td>
<td>FP.AL</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Var1/var2</td>
<td>FP</td>
<td>FP.GT</td>
<td>TP</td>
<td>FP.GT</td>
<td>UNK</td>
</tr>
<tr>
<td>Var1/var3</td>
<td>-</td>
<td>-</td>
<td>FP.AL</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Var2/var3</td>
<td>-</td>
<td>-</td>
<td>FP.AL</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Var3/var4</td>
<td>-</td>
<td>-</td>
<td>FP.AL</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Var1/var1</td>
<td>FP</td>
<td>FP.GT</td>
<td>FP.GT</td>
<td>TP</td>
<td>UNK</td>
</tr>
<tr>
<td>Var2/var2</td>
<td>-</td>
<td>FP.AL</td>
<td>FP.GT</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Var3/var3</td>
<td>-</td>
<td>-</td>
<td>FP.AL</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Matches counted as FP.GT and FP.AL are additionally counted as both FP and FN, since our tool's default matching stringency requires genotypes to match. Query variants outside the truth bed file are counted as UNK. Boxes with dashes are not possible when comparing two VCFs. Matches counted as FP.GT and FP.AL (bold) are additionally counted as both FP and FN, since our tool's default matching stringency requires genotypes to match.

Table 2 | Examples of several combinations of truth and query SNV genotypes and how they are counted as TP, FP, FN, FP.GT, and FP.AL
An open resource for accurately benchmarking small variant and reference calls

Fig. 1 | Arbitration process used to form our benchmark set from multiple technologies and callsets. a. The arbitration process has two cycles. The first cycle ignores ‘filtered outliers’. Calls that are supported by at least two technologies in the first cycle are used to train a model that identifies variants from each callset with any annotation value that is an ‘outlier’ compared to these two-technology calls. In the second cycle, the outlier variants and