What kind of data analysis will be required in the PM cooperation & how do we communicate results?

Amanda Mocroft
a.mocroft@ucl.ac.uk

The statistics team at UCL

Amanda Mocroft
Anything and everything but not theoretical

Al Cozzi Lepri
Biological data and everything theoretical

Leah Shepherd
PhD Student, cancer

Matt Bickley
PhD Student, renal function

Andrew Phillips
Head of group, modelling

Anna Schultze
PhD Student, TB and virology

Andrew Phillips
Head of group, modelling
What can we offer?

- To supervise data warehouse, data structures, SOPs for data transfer, data harmonisation and cleaning
- To provide methodological strength and statistical rigour to interdisciplinary (collaborative) clinical research programmes across PERSIMUNE.
- To check research proposals for feasibility, suitability of the suggested analytical method and statistical validity
- To provide statistical training and support to existing and proposed projects
- Facilitate winning of major clinical research grants and the placing of publications in the most highly esteemed academic journals
Collaboration and communication is key

Areas of expertise

- The design and analysis of epidemiological and other observational studies
- The design and analysis of nested case control studies
- The design and analysis of randomised trials and other clinical experiments (including the evaluation of complex interventions)
- Survival and Poisson regression analysis including risk modelling and development and validation of prognostic scores
- Complex statistical modelling, applications to high-dimensional biomarker data and clustered or correlated outcomes arising from measurements repeated over time (including repeated events or interventions)
- Modern statistical approaches to causal inference
Ongoing work in PERSIMUNE

- Wide range of HIV studies
- Proof of concept – Checking statistical validity
- BK viremia and risk of kidney transplantation failure
- Renal function and various outcomes in transplanted patients
- PTLD in EBV
- Cancers in transplant recipients
- Cancers in different immune suppressed patients
- CMV infections in MATCH
- ALT flares in transplanted patients
- Neutropenia

Collaborations
- Response to novel treatment in people with psoriasis (BRC/Kings College/MISP camp: De Rinaldis, Cozzi-Lepri, Fontes)

Communication

Peer reviewed publications
Quality not quantity

Conference abstracts
International audience, clinically meaningful research

Regular meetings / TCs

SIGs

Listen to each other

Cross-Institute working
Persimune Modelling and Computation

Magnus Fontes
International Group for Data Analysis
Institut Pasteur France
Centre for Mathematical Sciences
Lund University Sweden
One way of attack: Looking at Genetic Variation and Human Disease with GWAS. The HapMap project started in 2002. First results published in Nature 2005:

**A haplotype map of the human genome**

The International HapMap Consortium*

Inherited genetic variation has a critical but as yet largely uncharacterized role in human disease. Here we report a public database of common variation in the human genome: more than one million single nucleotide polymorphisms (SNPs) for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations, including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. These data document the generality of recombination hotspots, a block-like structure of linkage disequilibrium and low haplotype diversity, leading to substantial correlations of SNPs with many of their neighbours. We show how the HapMap resource can guide the design and analysis of genetic association studies, shed light on structural variation and recombination, and identify loci that may have been subject to natural selection during human evolution.
Google trends Interest over time for Personalized Medicine. Personalized Medicine became a buzzword precisely in 2005

Partially thanks to the first successful GWAS studies connected with e.g. HapMap

NIH Public Access
Author Manuscript

Published in final edited form as:
Science. Author manuscript; available in PMC 2008 November 10.

Was the Human Genome Project Worth the Effort?

Stephen P. Daiger
The author is at the Human Genetics Center, the University of Texas Health Sciences Center, Houston, TX 77030, USA. E-mail: stephen.p.daiger@uth.tmc.edu

One of the promises of the Human Genome Project was that it would provide tools for identifying genetic factors that contribute to common, complex diseases such as cancer and diabetes. Finding these factors would, in turn, suggest possible targets for drug therapy and other forms of treatment. Three papers in this week’s issue—by Edwards et al. (1) on page 421, Haines et al. (2) on page 419, and Klein et al. (3) on page 385—deliver on this promise for a debilitating, blinding disease called age-related macular degeneration (AMD). Using several genome-derived tools applied to nonoverlapping groups of AMD patients, the three groups report that a common variant in the complement factor H gene (CFH) on human chromosome 1q41 contributes a substantial fraction of the difference between affected and unaffected individuals. AMD affects more than 10 million Americans and is the leading cause of blindness among the elderly. The complement system is the target of a number of modulatory drugs and treatments. Together with the new findings, these facts confirm the broad potential health benefits of “genomic science.”
BUT WE NEED TO BE VERY CAREFUL AND WE NEED SOUND MATHEMATICAL MODELLING

“All models are wrong but some are useful.”

p-values need to be backed up with biology

Searching for genetic determinants in the new millennium

Neil J. Rick
Department of Genetics, Stanford University School of Medicine, Stanford, California 94305 USA

Human genetics is now at a critical juncture. The molecular methods used successfully to identify the genes underlying rare mendelian disorders are failing to find the numerous genes causing more common, familial, non-mendelian disorders. With the human genome sequence nearing completion, new opportunities are being presented for unravelling the complex genetic basis of non-mendelian disorders based on large-scale genome-wide studies. Considerable debate has arisen regarding the best approach to take. In this review I discuss these issues, together with suggestions for optimal post-genome strategies.

The current challenge:
Integration of Omics data and biological modelling

- Integrative and holistic analysis-Combining different omics data
- Deconvoluting complex biological signals
- Uncovering the dynamics of biological processes-
  Time series of omics data
Pushing for reproducibility

Over the past year, Nature has published a string of articles that highlight and reinforce the need for reproducibility of research (collected and freely available in the online edition of this journal). The problems arise in laboratories, but journals such as this one compound them by subjecting the results that they publish, and when they are not subjected to adequate scrutiny, the information for other researchers to access results properly.

From next month, Nature and the Nature research journals will introduce editorial measures to address the problem by improving the consistency and quality of reporting in life-sciences articles. To ease the interpretation and improve the reliability of published results, we will more systematically ensure that key methodological details are reported, and we will give more space to methods sections. We will examine statistics more closely and encourage authors to be transparent, for example by including their raw data.

Search for Precision Medicine on Google trends

[Graph showing interest over time and regional interest map]
NCBI web resources

GATK pipeline from the Broad Institute

The Genome Analysis Toolkit or GATK is a software package for analysis of high-throughput sequencing data, developed by the Data Science and Data Engineering group at the Broad Institute. The toolkit offers a wide variety of tools, with a primary focus on variant discovery and genotyping as well as strong emphasis on data quality assurance. Its robust architecture, powerful processing engine and high-performance computing features make it capable of taking on projects of any size.
Database information (example from KEGG)

Visualizing and analyzing pathways or graphs
General questions: What kind of data? What do we measure?

Can be hard to answer in biology. Sometimes we need to settle for describing how we measure as precisely as possible.

Think about e.g.

mRNA expression (microarrays, RNA-Seq, etc.)

Problems with normalization, housekeeping genes, spike-ins?

Flow Cytometry Problems with gating, non specific binding. Output: proportions of different cell populations (Intensities?)

Microarray data example with N=36 samples and p=9124 variables (Gene-IDs) Inflammatory Bowel Disease (IBD)

Data from NCBI's Gene Expression Omnibus (GEO) on Crohn's disease and Ulcerative Colitis, GDS2642
**Challenges:** High dimensional data with (many) more variables than samples, noise, different types of artefacts and variations

**p times N data matrix**

Measurements of
- p variables
- N samples
- +annotations

With p >> N and often a need to do integrative analysis.

---

**What should it mean to be similar or different?**

It is the choice of the modeller. Example: L^r distances in \( \mathbb{R}^p \)

One natural distance on \( \mathbb{R}^p \) is the Euclidean distance. The Euclidean distance, or \( L^2(\mathbb{R}^p) \) distance, between a sample point \( x = (x_1, x_2, \ldots, x_p) \) and a sample point \( y = (y_1, y_2, \ldots, y_p) \) is then given by

\[
d(x, y) := \left( \sum_{k=1}^{p} (x_k - y_k)^2 \right)^{1/2}.
\]

The underlying rationale is that this distance reflects how similar the two samples are across all the p variables measured.

Fixing a real number \( r \geq 1 \) the same thing is true for the generalized so called \( L^r(\mathbb{R}^p) \)-distances

\[
d_r(x, y) := \left( \sum_{k=1}^{p} (x_k - y_k)^r \right)^{1/r}.
\]

The insistence of keeping \( r \geq 1 \) here is in order to ensure that the triangle inequality (see section on Metric Spaces below) holds.
General distance functions-Metrics

A metric space is a set $M$ together with a fixed distance function or metric $d : M \times M \rightarrow [0, \infty)$ such that for all $x, y$ and $z$ in $M$ we have

1. $d(x, y) \geq 0$ with equality if and only if (iff) $x = y \quad (2.1)$
2. $d(x, y) = d(y, x)$ (symmetry) \quad (2.2)
3. $d(x, y) \leq d(x, z) + d(z, y)$ (the triangle inequality) \quad (2.3)

Examples:
- **Edit distances** for e.g. sequencing data
- **L2 or Euclidean distance** for quantitative data
- **L1 or Manhattan (or Taxicab) distance**

Effect of different distances


The normal “unit sphere” in the standard L$^2$ norm

The “unit sphere” in the L$^\infty$ norm
Exploratory Data Analysis and Visualizations

John Wilder Tukey (1915-2000)
Inventor of the FFT, the Box plot and the word “bit”.

The box and whiskers plot with outliers on the IBD dataset
Top 4 discriminators performing an ANOVA on Healthy-Crohn’s-Ulcerative Colitis
PCA duality and biplots

Dual PCA plots after ANOVA filtering down to FDR=0.05, giving 359 genes in the IBD data. Variables colored according to mean expression in the Healthy group (blue).

ISOMAP-Multidimensional scaling (MDS) using graph-distances

Graphs created in the space spanned by 217 variables. Visualization with PCA.
Deconvoluting variation

Deconvoluting different sources of variation by e.g.
- Model based methods, e.g. *Gaussian Mixture models*
- Supervised PCA is a good first step to find e.g. artefacts, outliers and batch effects

PCA for artefact detection

The IBD-data GDS2642: Filtering on variance for Crohn’s patients (gold) and Healthy (blue) to achieve an optimal 3D signal, resulting in 886 genes
Looking only at Affected vs Healthy shows another outlier in the IBD data set

Time series data

Most phenomena we study are dynamical, although time scales might differ between systems.

Example
Dangers with exploratory analyses

Filtering using ANOVA on a 10000*100 dataset with significance threshold 0.05 on random data resulting in 452 "significant discoveries" + PCA visualization.

The Solution is the use of solid statistical measures

From Nature Volume 514, Issue 7524 2014 The top 100 papers

The top 100 papers
Click through to explore the Web of Science's all-time top-cited papers. (Data provided by Thomson Reuters, extracted on 7 October 2014).

Rank: 59 Citations: 15,898
Controlling the false discovery rate: a practical and powerful approach to multiple testing. Benjamini, Y. & Hochberg, Y.
The projection score - an evaluation criterion for variable subset selection in PCA visualization

Magnus Fontes and Charlotte Soneson

\[ \alpha_2(\mathbf{A}_X, S) = \sum_{k \in S} \frac{\lambda^2_k}{\sum_{k=1}^{K} \lambda_k}. \]

\[ \tau(\phi_m(X), S, \mathcal{P}_{\phi_m}(X)) = \left( \alpha_2(\mathbf{A}_{\phi_m(X)}, S) \right)^{1/2} \]


Research within IGDA

Kerstin Johnsson
Developing automatic flow cytometry sorting strategies

http://www.milieuinterieur.fr/

Jacob Antonsson
Genetic and environmental determination of the immune system

http://www.milieuinterieur.fr/

Gabriel Illanes
Developing visualization tools to understand complex immunological data from the MI project

http://www.milieuinterieur.fr/

J Boussier
Kinetics analysis aids unravelling divergent signaling following LPS stimulation

http://www.milieuinterieur.fr/

Rasmus Henningsson
Develop a mathematical framework to model the Adaptive Fitness Landscape of a RNA virus

http://www.milieuinterieur.fr/
Microbial and Human Genomic Activities

- Genotyping
  - Cohorts of patients → SNPs, Exome, WGS
  - Extreme phenotypes → example of IC patients TRIO
  - Microbial genome → eg HIV seq

- Functional tests (NOXA1 & protein trap)
  - 4000 samples, 200 samples per week
  - Detection limit 0.5%

- Identification of Infectious agents
  - Microbiome
  - Microbes in blood

- Technique still to be developed
  - DNA
  - RNA

NIH funded HIV & Influenza WGS
New opportunity!
Exome Sequencing

Reduction in cost of DNA sequencing has fueled genomic research
Miniaturization: 4,000,000 DNA sequencing reactions in one chip

Schematic overview of exome sequencing

Biesecker & Green (2014) NEJM
Diagnostic exome sequencing for rare, clinically unrecognizable, or puzzling disorders that are suspected to be monogenic in origin

Different discrete filtering steps can be applied to narrow down search for candidate disease-causing genetic variants

Key assumptions
- Assumes causal variant is coding, ignoring regulatory elements and other non-coding variants outside of exon definitions
- Assumes causal variant alters protein sequence, ignoring rare cases of functional synonymous changes
- Assumes causal variant has complete penetrance
- Assumes causal variant has complete detectance

Filtering steps
- Targeted sequencing of exons
- Remove synonymous variants
- Remove previously identified variants
- Restrict to variants fitting dominant/recessive model of inheritance

Adapted from Stitziel et al. (2011) Genome Biol
Analysis of additional exomes help to prioritize variants

- Sequenced individual

Overlap analysis

Family linkage analysis

Parent-child trio analysis

Adapted from Bamshad et al. (2011) Nat Rev Genet

(A) Homozygous rare recessive variants and (B) de novo mutations are few in numbers and prime candidates to be causative of disease

Parent-offspring trio sequencing

(A)

\[a: \text{common allele} \]
\[b: \text{rare allele (<1\% in population)} \]

(B)

Germline de novo mutations occur in egg or sperm cell before fertilization or immediately after fertilization
Mutations do not come with information about their interpretation: Various information can be used to assess potential pathogenicity.

Exome sequencing on 500 consecutive patients: Diagnostic rate of 30%
Exome sequencing is an effective diagnostic method, but ...

... should it be possible for the patient to get his raw genetic data?

... should genetic data be distributed within the healthcare system?

... should we report incidental findings to the patient?
Do genetic defects play a prominent role in explaining phenotypes with increased susceptibility to infectious complications?

Individuals with increased susceptible to infections not caused by known factors (immunosuppressive therapy, HIV, etc)

Mutations do not come with information about their interpretation: Various in silico evidence can be used to assess potential pathogenicity

Cohort mutation rate relative to background population

Gene function
- Known disease gene
- Gene in disease-associated pathway
- Gene expressed in relevant tissue

Mutation impact
- Mutation type
- Evolutionary conservation
- Protein domain prediction

Mutations in patients

Multiple patients with mutations in same gene/pathway
Strategies to prioritize variants from exome studies

- Male
- Female
- Affected
- Heterozygous carrier
- Sex-linked heterozygous carrier
- Mating
- Consanguineous mating

---

Strategies to prioritize variants from exome studies

- Linkage based strategy
- Homozygosity based strategy
- Double-hit based strategy
- Overlap based strategy
- De novo based strategy
- Candidate based strategy

Mendelian disease: Diseases in which the phenotypes are largely determined by the action, lack of action, of mutations at individual loci

Autosomal dominant
Autosomal recessive
X-linked

Clonal evolution in the TNBC inferred from single cell exome and copy number data
### Frequency of disorder

<table>
<thead>
<tr>
<th>Extremely rare disorder</th>
<th>Very rare disorder</th>
<th>Rare disorder</th>
<th>...</th>
<th>Common disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schinzel-Giedion syndrome</td>
<td>Miller syndrome</td>
<td>Complex I deficiency</td>
<td>...</td>
<td>Intellectual disability</td>
</tr>
</tbody>
</table>

#### Mutational target

- **Locus specific**: Single gene
- **Few genes**: Few genes
- **Many genes**: Many genes

---

**Identify patient**

**Test proband and other relevant**

**Evaluate genetic variants**

---

**Human Genome Project**

**International HapMap Project**

**1000 Genomes**

A Deep Catalog of Human Genetic Variation

---

Timeline:

- **2002**: 3 billion basepairs
- **2008**: A Deep Catalog of Human Genetic Variation
The search for disease-causing variants

Adapted from Nature 461, 747-753 (2009)
Human Genetic Diseases

• **Complex Disorder**
  • Polygenic, many genes.
  • Low penetrance/effect size.
  • Multifactorial, environmental, dietary.
  • Examples: heart disease, diabetes, obesity, autism, etc.

• **Mendelian Disorder**
  • Monogenic or polygenic.
  • Full or high penetrance/effect size.
  • Examples: sickle cell anemia and cystic fibrosis.

Definitions

**Locus**: *Location* on the genome

**SNP**: “Single Nucleotide Polymorphism” a mutation found in >1% of the population, that produces a single base pair change in the DNA sequence

- alleles
- alternate forms of a SNP

- genotypes
- both alleles at a locus form a genotype

- haplotypes
- the pattern of alleles on a chromosome

**Genetic Association**: Correlation between (alleles/genotype/haplotype) and a phenotype of interest.
**Why Find Disease Genes?**

1. In a “Mendelian”, single-gene trait, one gene is sufficient to cause (most of) the disease phenotype
2. In a polygenic/multifactorial, “complex” trait, no one gene is sufficient to cause the disease phenotype

**Hunting for Disease Genes**

1. In a “Mendelian”, single-gene trait, one gene is sufficient to cause (most of) the disease phenotype
2. In a polygenic/multifactorial, “complex” trait, no one gene is sufficient to cause the disease phenotype
Rare inherited diseases

The scale of rare diseases

1 in 17 people will suffer from a rare disease at some point in their lives.

In the UK alone that equates to approximately 3.5 million people.

Only a quarter of rare diseases have had their molecular basis defined, meaning many risk being undiagnosed and therefore untreated.

There are at least 6,000 rare diseases.

Many rare diseases (approximately 80%) are of genetic origin.

Seventy-five per cent of rare diseases affect children.

30% of rare disease patients die before their fifth birthday.

Frequency of disorder

<table>
<thead>
<tr>
<th>Rare</th>
<th>Low frequency</th>
<th>Common</th>
</tr>
</thead>
<tbody>
<tr>
<td>(&lt;1/10,000)</td>
<td>(1/10,000–1/100)</td>
<td>(&gt;1/100)</td>
</tr>
</tbody>
</table>

Mutational target

e.g. CHARGE syndrome (1/10,000)

e.g. Noonan syndrome (1/2,000)

e.g. intellectual disability (2/100)

<table>
<thead>
<tr>
<th>PTPN11</th>
<th>RAF1</th>
<th>SOS1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD7</td>
<td>KRAF</td>
<td>BRAF</td>
</tr>
<tr>
<td>NRAS</td>
<td>MAPK1</td>
<td></td>
</tr>
</tbody>
</table>

Single gene  2–100 genes  >100 genes
# Persimune - Microbial Genetics

Using high-throughput sequencing for identification of gut-derived microbial infections in immunocompromised patients

Rasmus Lykke Marvig
Center for Genomic Medicine
Rigshospitalet, Copenhagen University Hospital

Systemic Inflammatory Response Syndrome (SIRS): Symptoms of systemic infection, but no microbial pathogen

Patient cohorte

Definitions

SIRS  Sepsis  Infection

Immunocompromised patients with SIRS
Strategy for identification of potential pathogen:
Deep sequencing of DNA from host tissue that may contain pathogen

Caveats of deep sequencing of metagenomic DNA libraries:
Example - Contamination likely explains ‘food genes in blood’ claim

Claim:
Complete Genes May Pass from Food to Human Blood
Sándor Spisák, Norbert Solymosi, Péter Ittész, András Bodor, Dániel Kondor, Gábor Vattay, Barbara K. Barták, Ferenc Sipos, Orsolya Galamb, Zsolt Tulassay, Zoltán Szállás

Disproof:
Diverse and Widespread Contamination Evident in the Unmapped Depths of High Throughput Sequencing Data
Richard W. Lusk
Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, United States of America
Bacterial translocation:  
Is the gut microbiota the source of infection?

Sampling  
- Gut metagenome
- Isolate genome

DNA sequencing  
- Pathogen sequence reads

Analysis  
- Resistance genes
- Virulence factors
- Phylogenetic analysis

Questions for discussion

Who are the relevant patients at Rigshospitalet?
How many patients are available and how should they be prioritized?
Is it relevant to search for pathogen DNA in blood samples?
Is it possible to obtain fecal samples?
Using whole-exome sequencing for identification of disease-causing genetic variants in patients with impaired immune function

Do genetic defects play a prominent role in explaining phenotypes with increased susceptibility to infectious complications?

Individuals with increased susceptible to infections not caused by known factors (immunosuppressive therapy, HIV, etc)
**De novo mutations: Mutations that are not inherited from the parents**

Germline de novo mutations occur in egg or sperm cell before fertilization or immediately after fertilization.

**Parent-offspring trio sequencing**

**Homozygous rare recessive variants may be causative of disease**

Impaired immunity.

Parent-offspring trio sequencing

<table>
<thead>
<tr>
<th>a: common allele</th>
<th>b: rare allele (&lt;1% in population)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>ab</td>
</tr>
<tr>
<td>bb</td>
<td></td>
</tr>
</tbody>
</table>
Overview of experimental steps for whole-exome sequencing

Exome contains 85% of disease-causing mutations in Mendelian disorders (Rabbani, J Hum Genet, 2014)

Mutations do not come with information about their interpretation:
Various *in silico* evidence can be used to assess potential pathogenicity

- Cohort mutation rate relative to background population
- Gene function
  - Known disease gene
  - Gene in disease-associated pathway
  - Gene expressed in relevant tissue
- Mutation impact
  - Mutation type
  - Evolutionary conservation
  - Protein domain prediction

Mutations in patients

Multiple patients with mutations in same gene/pathway
Relationship between the size of the mutational target and the frequency of disorders caused by *de novo* mutations

<table>
<thead>
<tr>
<th>Frequency of disorder</th>
<th>Size of mutational target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely rare</td>
<td>Locus specific</td>
</tr>
<tr>
<td>Rare</td>
<td>Few genes</td>
</tr>
<tr>
<td>Common</td>
<td>Many genes</td>
</tr>
</tbody>
</table>

- Schinzel-Giedion syndrome
- Complex I deficiency
- Intellectual disability

Questions for discussion

Is this an illusion?

What is the throughput of the relevant patients at Rigshospitalet?

Is both patients and their parents available?

How similar are the patient phenotypes?

Are phenotypes likely caused by *de novo* mutations or rare homozygous recessive variant?

What is the heterogeneity of disease-causing genetic variants?

Is it possible to correlate specific genetic defects with susceptibility to infection from certain microbial pathogens?
Marc Bennedbæk MSc.Eng. Biotechnology
HIV sequencing for resistance mutations

START Project
• Genome sequencing of HIV-1 genome for detection of minor resistant virus populations

Subject
• Baseline plasma samples of ~4,000 HIV-1 positive individuals from Western and developing countries

Process:
Identification of disease-causing genetic variants in immunocompromised patients

Rasmus Lykke Marvig, PhD
Center for Genomic Medicine
Rigshospitalet

Hypothesis: Genetic variants play a prominent role in explaining phenotypes with increased susceptibility to infectious complications

Individual with increased susceptible to infections not caused by known factors (immunosuppressive therapy, HIV, etc)

Why find disease gene?
• Understand biological defect
• Diagnostics
Reduction in cost of DNA sequencing has fueled genomic research

Miniaturation: 4,000,000 DNA sequencing reactions in one chip
Different discrete filtering steps can be applied to narrow down search for candidate disease-causing genetic variants

Key assumptions
- Assumes causal variant is coding, ignoring regulatory elements and other non-coding variants outside of exon definitions
- Assumes causal variant alters protein sequence, ignoring rare cases of functional synonymous changes
- Assumes causal variant has complete penetrance
- Assumes causal variant has complete disease

Filtering steps
1. Targeted sequencing of exons
   - ~3 million SNPs/individual
2. Remove synonymous variants
   - 15,000-20,000 coding SNPs/individual
3. Remove prioritized identified variants
   - 7,000-10,000 non-synonymous coding SNPs/individual
4. Restrict to variant fitting dominant/recessive mode of inheritance
   - 200-500 novel non-synonymous coding SNPs/individual
5. One or several putative genes

Analysis of additional exomes help to prioritize variants

Overlap analysis
- Sequenced individual

Family linkage analysis

Parent-child trio analysis

Adapted from Bamshad et al. (2011) Nat Rev Genet
Homozygous rare recessive variants and *de novo* mutations are few in numbers and prime candidates to be causative of disease.

**Patient-parents trio case 1:**

- Three variants passing filtering:
  - One (●) *de novo* mutation (parents do not have mutation)
  - Two (●) rare homozygous variants (parents heterozygous)

**Annotation and classification of variants using various databases**

Patient from Terese Katzenstein

---

**Patient case 1: Neutrophils unable to kill invading bacteria**

Discovery of a gene not previously implicated in phenotype

- Protein complex generating reactive oxygen species to kill bacterium
Exome sequencing project...

... approved by ethical committee and the data protection agency

... will be performed on 60 patients and their parents (180 persons total)

... will improve diagnostics and identify novel immune mechanisms
Biobank Scientific Interest Group
Ruth Frikke-Schmidt & Henrik Ullum

PERSIMUNE - BIOBANK

Aim and focus of activities
- Automated professional biobanking
- Optimal use of in-house logistic systems

Current status
- Establishment of local Labka biobank-codes
  - PERSIMUNE level 1, feces, saliva, and 3 pediatric codes
- Establishment of customized Labka profiles including biobank codes

Vision
- For research collaborations provider of
  - Automated sample retrieval
  - Routine and specialized analyses of biobank-material
  - Customized data extracts
Imaging Scientific Interest Group
Malene Fischer & Andreas Kjær

IMAGING

Aim and focus of activities
• To bring together in a SIG expertise necessary for supporting PERSIMUNE in the use of advanced imaging within infection, inflammation and rejection as well as pursue development of new imaging tools.

Current status
• Retrospective analysis of the use of PET/CT in immunocompromised patient
• PET/MR for diagnosing rejection after SOT transplantation
• PET/MR for evaluating inflammation in liver and gut

Vision
• Evaluate available, non-FDG PET tracers in immunocompromised patients as well as more general in infection, inflammation and rejection.
• Develop new specific tracers for visualization of immune response.
• Implement molecular imaging (prospective trial) in the surveillance of immunocompromised patient based on an individual risk profile.
Microbiome and Human Genomics
Finn Cilius Nielsen & Rasmus Lykke Marvig

Aim and focus of activities

• The aim of the G-SIG is to ensure that Persimune harness advances in high throughput sequencing in order to integrate information on genomes and microbiomes in medicine.
• Advances in high-throughput DNA sequencing techniques have triggered a revolution in our ability to generate genomic information in disease research. These techniques may be used to sequence single genomes from both humans and microbial pathogens, but also allow for deep sequencing of the collective genomes of the microorganisms that reside on and within humans. Accordingly, the technological advancements open for an opportunity to use knowledge about the genetic makeup of humans, pathogens, and microbiomes to predict and understand phenotypes.
• Specific projects and goals include:
  (1) Identification of disease-causing genetic variants in immunocompromised patients.
  (2) Use of deep metagenomic sequencing for detection of microbial infections in patients with neutropenic fever.
  (3) Investigating infectious disease outbreaks using pathogen whole genome sequencing.
  (4) Association between the composition of the microbiome just prior to and shortly after iatrogenic induced immune paralysis and subsequent risk of invading infection.

Current status
• Members of G-SIG has put a major effort in supporting the establishment of feces and saliva biobanks within PERSIMUNE.

Vision
• The vision of G-SIG is to gather experts and stakeholders with an interest in personalized medicine guided by genomic information. Areas of focus include personal genomics, pathogen genomics, and the human microbiome.
Immunologic SIC
Sisse Ostrowski & Jakob Thaning Bay

**Aim and focus of activities**
- Develop and implement an immunologic characterization platform that provides differentiated analyses of immune phenotype and function for implementation in clinical routine at RH and in PM projects
- Reveal and (try to) fulfill clinical & research needs
  - Dialog, development, studies & evidence
- Develop a real-time & research tool for data output & data analysis

**Current status**
- Ongoing development of 10-color real-time flow cytometry panel
  - In external review
- Different CMV immunity assays under review
- Immune function assays under review/evaluation & discussion

**Vision**
- Provide a flexible real-time state-of-the-art immunologic characterization platform that is both clinically relevant and applicable as a high impact research tool
  - The platform must be flexible and easy to expand/modify in order to meet future clinical needs and research requirements
Aim and focus of activities
– Developed an expert tool providing personalized medicine to immunologically impaired patients implemented as a routine clinical tool.
– Building a comprehensive data warehouse structure with link to local, regional and national databases for big data research aiming to identify risk factors for development of infections in immunologically impaired patients at RH.

Current Status - immediate plans
– Secure data retrieval and storage of data from
  – Dansk CancerBiobank
  – RegionH Biobank
  – RegionHF MK / prescription database
  – LPR, CPR registry
  – Danish HIV cohort
  – Vævstypelab and Blodbanken
– Formulate SOP for data clean up documentation
– Formulate procedure for handling and communication regarding data QA feedback from researchers and statisticians
– Apply MiBa mapping
– Create rules for clean up (Biochemistry, Microbiology, Medications)

Vision
– Create a data warehouse and tools for data management, visualisation, and analysis to improve patient care by precision medicine, and to support cross functional research collaboration and idea generation within PM at RH and internationally and establish a research platform for collaborating researchers own research
Immunologic Colloquia SIG
Marie Helleberg & Alvaro Borges

SIG Immunologic Colloquia

Aim and focus of activities:
To convene monthly colloquiums on topics that are relevant to the PERSIMUNE strategic vision. The colloquia are aimed to a broad audience of basic and clinical researchers, practicing physicians, data management experts and statisticians.

Current status: 6 colloquia so far. Wide range of topics discussed: clinical challenges, advanced molecular biology technics, ethical aspects of access to data, external experts, etc.

Vision:
To foster academic debates of high level; to increase visibility of PERSIMUNE within Rigshospitalet; to create opportunities for the different SIGs to meet and discuss research collaborations and common challenges.
Epidemiology and Biostatistics SIG
Amanda Mocroft and Alessandro Cozzi-Lepri

Aim and focus of activities

• To provide methodological strength and statistical rigour to interdisciplinary (collaborative) clinical research programmes across PERSIMUNE

Current status

• Website populated with basic description of mission/vision
• Identification of members- first face to face meeting Feb 10th 2016

Vision

• To supervise the construction of the data warehouse to facilitate statistical analysis
• To closely interact with the Bioinformatics SIG to best harmonise classic epidemiological thinking with cutting edge methods for data mining/precision medicine revolution
• To be leaders in making sophisticated statistical research methods available to and usable by PERSIMUNE researchers
Collision: how to introduce bias by trying to control for confounding

Does HIV resistance have an effect on CD4 count change?

HIV drug resistance

True baseline CD4 count

Observed CD4 count

Observed change in CD4 count

Measurement error
Collision: how to introduce bias by trying to control for confounding

**Does HIV resistance have an effect on CD4 count change?**

- HIV drug resistance
- True baseline CD4 count
- Observed CD4 count
- Observed change in CD4 count
- Measurement error

**Does BK viraemia increase the risk of kidney transplantation failure?**

- After July 1, 2011 at Rigshospitalet...
  - Immunosuppressive therapy
  - BK viremia(t)
  - BK viremia(t+1)
  - Kidney transplantation failure

*For the causal effect of immunosuppressive therapy on the risk of transplantation failure, BK viremia is a time-dependent confounder affected by prior treatment.*