

# What is Integrative Genomics (IG)??

Objective: To give a presentation 60 minutes at the end of the course and write 20+ page report covering the selected aspects of IG and also discussing which research could be done.

IG will here mean analysis of cell function/dynamics using multiple levels simultaneously. In this sense, IG is also a reality, in that there has been an increasing number of studies successfully involving multiple levels. The last decade has seen the rise of a variety of high throughput technologies, whose ability continues to expand both quantitatively and qualitatively. It started with sequencing (Genomics), was followed by measuring transcription levels (Transcriptomics), but has now spread to protein and metabolite concentrations (Proteomics and Metabolomics) and much more. Analysis of a given problem would additionally include multiple sources of knowledge about the problem. Navigating optimally in such a situation is hard and can only be done by simultaneous use of a series of computational tools ideally integrated into an environment.

## *A. The main categories of data are:*

**G - Sequence Data (Genomics)** – will in coming years increase enormously so possibly thousands of genome equivalents ( $>10^{12}$  nucleotides) will be available for the individual researchers. The main challenge will be rational use of such data, not the obtain it. Sequence data will represent species, population and ontogenetic variation.

**E - Epigenomics** - comprises the features outside of DNA sequence that affect cellular processes such as structural DNA changes.

**G - Transcriptomics** - is the level of different transcripts of all genes. This leads to crucial insights into the dynamics of gene regulation and expression patterns that characterize different tissue types and disease states.

**P - Proteomics** - is the concentration of proteins and modified proteins. Is an important complement to expression levels, but is at present very noisy and incomplete.

**M - Metabonomics** – is the concentration of all metabolites. Is important in monitoring physiological changes over time and in response to different conditions.

**P - Phenomics** - is the set of phenotypic characteristics which comprehensively characterise phenotype of an individual. Is often the feature of main interest, but is statistically hard to define due to its subjective and non-molecular nature.

**Other Data Types** – this includes novel types of microscopy and data at different levels obtained by traditional non-throughput experiments. This will by nature be a very heterogeneous class that however cannot be ignored.

## *B. Existing Knowledge, Meta-analysis and Previous Investigations:*

All investigations are done within a conceptual framework – like the central dogma – and using existing data, analysis, literature and conclusions. Relating efficiently to this is important in using new data and data optimally.

## *C. Concepts and Models:*

Data are analyzed by the use of concepts and models. Key components are:

*Genotype to Phenotype Functions ( $G \rightarrow F$ )* – due to the inherent difficulty of mechanistic modelling of different levels in biology, purely phenomenological models are very important. By far the most important class of models connect the two levels furthest apart – the Genome and the Phenome – which also happens to be the levels with the best data and the level of greatest interest respectively.

*Networks* – have risen to prominence in the last decade and represent a very general class of models indeed. In interpreting biological phenomena, graphs are often used as the end goal.

*Hidden Structures and Processes* – not all structures and processes can be observed and often their dynamics and structure must be inferred through their influence on what can be observed. Key examples are genome annotation using Hidden Markov Models (HMMs) and Hidden Gaussian processes for fluctuating molecule concentrations, but this approach is part of much more wide applicable principle.

*Mathematical Models* – these are an increasing part of the analysis. Molecular dynamics and kinetic modelling are two examples of methods that will generate something very similar to high throughput data and will become increasingly reliable and ubiquitous in their use.

*Genealogical Relationships and Evolution* – comparative genomics, association mapping and the cancer genome project are three examples of projects that compare multiple sequences that are sampled from different species, different individuals and different cells respectively.

These 5 classes of concepts are extremely wide spanning and covers most to the conceptual background for the present biosciences.

### ***A + B + C = Integrated Analysis/Genomics (IG) and Functional Explanation***

Combining data, knowledge and concept around a specified problem creates integrated analysis or integrated genomics. Traditionally, only a single type was analysed. Then two data types started to be combined, but increasingly many types are used in the same analysis. Despite an explosion in both quantity and quality of data, IG has far to go to create reliable descriptions of the underlying biological phenomena. However, it is clear that both data and IG will expand enormously in coming years. A major challenge in IG has been to provide a link to known molecular processes and structures, which could be called functional explanation.

The questions and contents below are meant as motivators and need not be followed. Since we give several lectures on IG, you should probably try to give a new angle in the presentation or focus on a few new and exciting publications. I can also supply very good publications on epigenomics, transcriptomics, proteomics, metabonomics and phenomics. This project is challenging because the available literature is so large, often has very specific applications, there is no uniformly accepted definition of what IG is and I give lectures related to the topic. To avoid being confronted with excessive amounts of literature, it will be important to create a focus immediately and start with some very good applications/papers instead of trying to get an overview of the field.

The big questions are: What are the key classes of data (OMICS)? What is the inherent variation within one level? Which classes are often combined in analysis? What are the main benefits of combination? Which models are used to analyze them?

#### **Possible Contents of Presentation/report:**

What are the discussed data types?

What are their dimensions?

What do say about the underlying quantities that are of interest?

How variable are they?

3 recent case studies

*Comments:* Clearly this is a big area and deciding on a focus immediately is important. This can be done by starting from a few major papers or looking main of cases of functional explanation, a few data types, the classes of  $G \rightarrow F$  functions used [its dependency on environment??]

#### **Recommended literature:**

Davies, Rafnar, Hellenthal and Hein (2009) “Integrative Genomics and Functional Explanation” downloadable from <http://www.stats.ox.ac.uk/research/genome/publications>

**Epigenomics:** Meissner et al. (2008) “Genome-scale DNA methylation maps of pluripotent and differentiated cells” Nature 454.766-70.

**Transcriptomics:** Emilson et al. (2008) “Genetics of Gene Expression and its effect on Disease” Nature 452.423-30.

**Proteomics:** Cox and Mann, (2007) “Is Proteomics the New Genomics?” Cell 130,395-8

**Metabonomics:** Sreekumar et al. (2009) “Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression” Nature 457. 910-15

**Phenomics:** The Mouse Phenotype Database Integration Consortium Mammalian (2007) Integration of mouse phenome data resources Genome 18, 157 163