Introduction to Bioconductor

# Section 1: What we measure, why, and how

## Overview

This week, we lay the foundations for learning Bioconductor by installing software and discussing the biology and biotechnology background needed to understand and analyse genomic data.

* Learn how to install Bioconductor and related packages.
* Review basic molecular biology concepts including basic ideas of DNA replication, chromosomes and DNA variants, gene models and gene expression, and concepts of epigenetics.
* Discuss the biotechnology behind microarrays and next-generation sequencing to better understand how our experimental data is produced

### Assessment

* Try using R to explore the genes in the Mammaprint gene signature
  + MammaPrint® is **a gene profile test (a genomic test of breast cancer tumour cells)**. It examines 70 different genes to look for changes associated with a higher risk of breast cancer recurrence after treatment.
* The diagnostic signature assesses the risk that breast cancer will progress using the gene expression levels of 70 genes
* Signature is found in the genefu package
  + Provides info and functions relevant for gene expression analysis
* Info on the 70 gene signature used in the Mammaprint algorithm is in the sig.gene70 dataframe

## Molecular basis for phenotypic variation

* Phenotypic variation between organisms and cells is partly explained by their DNA.
* DNA encodes all of the information for making proteins, the building blocks of life. Messages within the DNA are transcribed into RNA, which is then translated into protein.
* Phenotypic variation depends on differences in the DNA and also on differences in which parts of the DNA are expressed as proteins.
* We use genomic technologies to measure differences in DNA and gene expression, such as changes in the sequence or amounts of expression, and relate those changes to phenotypic variation.

Assessment

* Explore the use of data frames to store phenoypes (columns) from several individuals (rows).

## DNA: chromosomes, replication, SNPs and other variants

* GWAS are a major tool of genetic epidemiologists
* In case control design, specific disease cases are identified along with similar disease-free individuals (controls)
* SNP chips or DNA sequencing is used to obtain genotypes for a large number of SNPs
* Genotype distributions for all SNPs are compared between cases and controls – SNPs exhibiting association with disease are identified
  + Investigate for relationship to gene regulation and function

Assessment

* Gwascat package includes info on a catalog of GWAS results
* Load package and check version of GWAS catalog stored in GRCh37 (hg19) coordinates

## How microarray hybridization works

* Essentially is counting molecules
* 500,000 cells on genechip array
  + Millions of DNA probes in each cell all targeting the same gene
  + If DNA from two different samples is labelled with different flourophores they can be hybridised at the same time
* Process
  + Denaturation into ssDNA
  + Label DNA with fluorescent probe
  + Hybridise ssDNA to probes on array
    - Probes are on a specific part of the array, allowing detection when complements bind
  + Probe binding is detected
* Applications of microarrays
  + Gene expression (RNA) microarray
    - Multiple probes targeting different parts of the gene, usually towards 3’ end as RNA degrades 5’ to 3’
    - Quantify gene expression
  + Genotyping (SNP) array
    - Design two probes, one for each allele
  + ChIP microarray
    - Detection of transcription factor binding sites
    - Where on the genome is a specific protein found?
      * Fragment DNA
      * Keep DNA with protein, remove protein
      * Amplify
      * Label and hybridise to genome – “tiling probes”
      * Determine which parts of genome are intensely lit up – binding sites

## Introduction to NGS

* Applications
  + Resequencing
    - Genotyping, SNP discovery etc
  + RNA-seq for differential gene expression
    - Where is coverage higher – more expression
  + ChIP-seq
    - Get DNA fragments that were bound to protein
    - Sequence
    - Map to identify which sites bound protein