Experimental Design Practical

Contents

1. **Human practical: Identification of prognostic biomarkers in human prostate cancer patients** **2**

2. **Mouse practical: Gene expression profile of wild and mutant HHEX in brain and liver development** **3**

3. **Cell culture practical: Investigation into the effect of RARα on transcription in breast cancer tissue treated with estrogen** **4**

4. **ChIP-seq practical: Investigation into the effect of RARα on transcription in breast cancer tissue treated with estrogen.** **5**

1. Human practical: Identification of prognostic biomarkers in human prostate cancer patients

A study is being run to look for prognostic biomarkers of recurrence free survival in prostate cancer in the UK.

A biomarker is a term often used to refer to measurable characteristics that reflects the severity or presence of some disease state. More generally a biomarker is anything that can be used as an indicator of a particular disease state. Biomarkers can be measured from many different things including blood, urine, saliva or tissue samples.

Study aim: To find a panel of prognostic biomarkers to predict those with poor recurrence free survival from prostate cancer to be able to offer them more aggressive treatment and spare those with a better prognosis from a treatment that is too aggressive.

1. What would the ideal study cohort be? What factors prevent this from being the study cohort? In reality what cohort would be chosen? What issues might this cause?
2. What are the pros and cons for using a retrospective or prospective cohort for this study? [A retrospective cohort study looks back over time and takes data from patients notes and/or interviews. A prospective cohort study follows the cohort through time collecting data as it goes].
3. What would the outcome measure(s) be for this study? What issues can arise in the definition of the outcome measures for this study?
4. What factors might affect the outcome? What information may you need to collect for use in the study analysis?
5. What are your groups? Do you need a control group in addition? If so what will your control group be?
6. What effect would a short follow-up time have? Which group will this affect more?
7. Which technologies could you use to screen for your panel of biomarkers?
8. What factors might influence your choice of technology for this study?
9. Once your study has a panel of biomarkers from which a test for recurrence has been formulated. What is the next step?
10. What characteristics would a clinical test formed from your biomarkers need?

2. Mouse practical: Gene expression profile of wild and mutant HHEX in brain and liver development

Hematopoietically Expressed Homeobox (HHEX) is a transcription factor which, plays an important role in the proper development of brain and liver in mouse. A mutant HHEX (where all Serine and Tyrosine residues are mutated to Alanine) is hyperactive and induces fetal death in mutant homozygous mice. We are interested in identifying gene expression changes in brain and liver and in determining key pathways involved, in response to hyperactivity of HHEX gene. Note that, samples are collected from 15 day-old fetuses that are homozygous wildtype (Wt/Wt), heterozygous mutant (Wt/Mt) and homozygous mutant (Mt/Mt), which manifest distinguishing morphological characteristics.

Experimental design related question and answers:

1) What are the scientific questions of interest in this experiment?
2) What are you measuring?
3) What controls samples should be included in this experiment? Why is control needed in the experiment?
4) How many replicates you need to include for each group? Discuss what factors might have influence in selecting the number of replicates?
5) Which experimental groups will be included?
6) How will any findings be validated?
7) What contrasts (sample group comparisons) you make with the data?
8) What are possible sources of bias and confounding variables in the experiment?
9) How can these sources of bias and confounding be controlled?

3. Cell culture practical: Investigation into the effect of RARα on transcription in breast cancer tissue treated with estrogen

Outline:

RARα is a transcription factor that appears to interact with estrogen (E2) in ER+ breast cancer. We are interested in characterising this interaction by looking at how gene expression changes in breast cancer cells treated with estrogen when RARα is not present (using a siRNA in cultured cells). We wish to identify which estrogen-induced and estrogen-repressed genes are impacted by the presence or absence of RARα, and to analyse the key pathways involved.

Design-related questions:

1. What are your objectives?
2. What are you measuring?
3. What are your primary sample groups of interest?
4. What controls will you use each type of sample group?
5. What constitutes a replicate in this experiment? Are they biological or technical? How many samples/replicates should be collected?
6. Sketch out the design as a matrix, with sample numbers
7. What sample group comparisons (contrasts) will you make with the data? Which gene set(s) will you use for pathway analysis?
8. What are possible confounding factors and sources of bias?
9. How will you confirm effective silencing?
10. What information about your experiment should be recorded to help identify any problems should there be any?
11. Will you be multiplexing samples? How will you assign barcodes? Will you use pooled libraries? How many pools? How will samples be assigned to pools?
12. What are the sequencing parameters you need to be aware of (e.g. sequencing type and depth)?
13. What other types of data might be useful to assay, and how might the sequencing parameters need to change to accommodate this?
14. Can you think of any other design related issues that could/should be addressed?

4. ChIP-seq practical: Investigation into the effect of RARα on transcription in breast cancer tissue treated with estrogen.

You are studying the evolution of mouse gene regulation by assaying transcription factor binding. You have stable colonies of three inbred mouse strains housed in the Biological Resource Unit (they are called BL6, CAST and SPRET, see the picture below for evolutionary relationships). These strains are derived from different species and subspecies in the Mus genus. You have access to other data which suggests that transcription factor binding changes between species, and you want to characterise the divergence of transcription factor binding over this short evolutionary distance. You have picked a model tissue (liver), and you plan to use Chromatin immunoprecipitation and sequencing (ChIPSeq) to assay transcription factor binding in all three strains. You have selected four transcription factors to assay plus Polymerase II as a proxy for transcription of genes (HNF1α, HNF4α, HNF6, CEBPα, PolII).

Design-related questions

1. What is the main aim of the experiment?
2. What are you measuring?
3. What is your control?
4. How will you validate your results?
5. What is your study design? How many replicates will you use? Will you use biological or technical replicates? What is the total number of samples you will prepare and sequence?
6. What are possible confounding factors and sources of bias? The reference genome for mouse was generated from the BL6 mouse strain; does this have any effect on your experiment?
7. What information should you collect about your samples to help identify any problems downstream?
8. When will you collect your samples? Which mice will you choose? How will you store the tissue samples? How and when will you perform the Chromatin immunoprecipitation? How and when will you prepare your libraries? Where will you go to find out the information necessary to make these decisions?
9. Where can you go to find out how to test the quality and quantity of your resulting libraries? Who can you ask for feedback as to whether your libraries are likely to sequence well?
10. NGS Libraries are frequently mixed together and sequenced as pools. Your kit only allows you to pool a maximum of 12 libraries. How many pools will you make for your study, and how many samples will be in each pool? Which libraries will you pool together? Where can you go to find out the information necessary to make these decisions?

Advanced Given that ChIPseq analysis is based on counting reads, and you have decided you need 10 – 20 million reads for your analysis, what type of sequencing will you use, and how many lanes do you need? Where can you go to find out the information necessary to make these decisions?

Note: Our HiSeq 2500 currently provides 150 – 250 million reads/lane

If you were using a kit which allowed you to pool 96 libraries, how many lanes would you need for this experiment? What could you do to reduce the chance of paying for unnecessary sequencing?