MEDI6215

UNIVERSITY OF Southampton

Principles of Downstream functional analysis

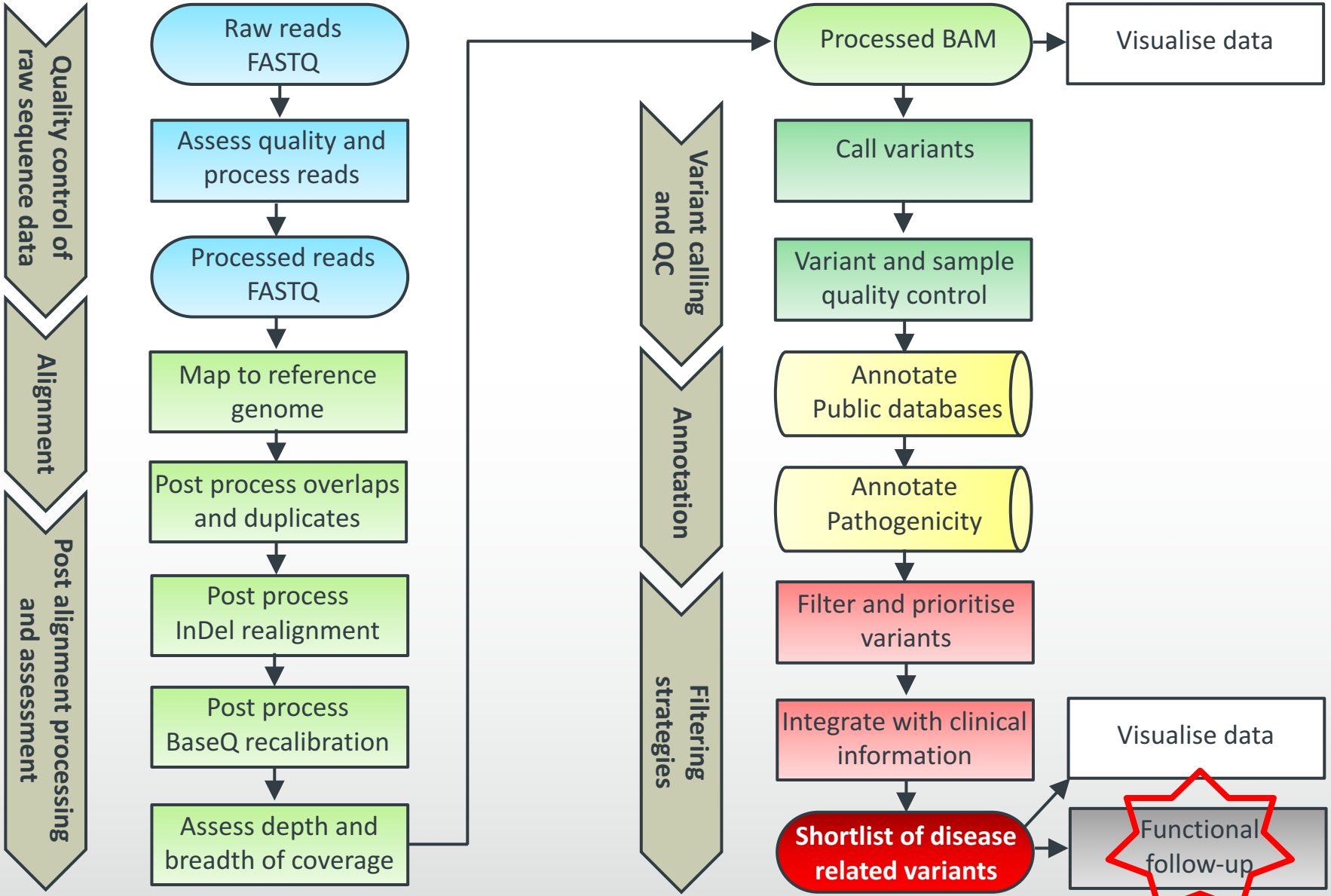
Bioinformatics, Interpretation, and Data Quality Assurance in Genome Analysis

The Greatest challenge of the "post-GWAS" era is to understand the functional consequences of these loci.
-Freedman et al. (2011) *Nature Genetics*. 43(6): 513-518

Christopher Woelk, PhD

December 16th, 2015

Analysis workflow



STRATEGY

UNIVERSITY OF Southampton

BOX 1 STRATEGIES TO PROGRESS FROM TAG SNP TO MECHANISM

1) Target resequencing efforts using linkage disequilibrium (LD) structure.

2) Use other populations to refine LD regions (for example African ancestry with shorter LD and more heterogeneity).

3) Determine expression levels of nearby genes as a function of genotype at each locus (eQTL).

4) Characterize gene regulatory regions by multiple empirical techniques bearing in mind that these are tissue and context specific.

5) Combine regulatory regions with risk loci using coordinates from multiple reference genomes to capture all variation within the shorter regulatory regions that correlates with the tag SNP at each locus.

6) Multiple experimental manipulations in model systems are needed to progressively implicate transcription units (genes) in mechanisms relevant to the associated loci:
i) Knockouts of regulatory regions in animal (difficult and may be limited by functional redundancy, but new targeting methods in rat are promising) models followed by genome-wide expression analysis.
ii) Use chromatin association methods (3C, CHIA-PET) of regulatory regions to determine the identity of target genes (compare with eQTL data).
iii) Targeted gene perturbations in somatic cell models.
iv) Explore fully genome-wide eQTL and miRNA quantitative variation correlation in relevant tissues and cells.

7) Explore epigenetic mechanisms in the context of genome-wide genetic polymorphism.

8) Employ cell models and tissue reconstructions to evaluate mechanisms using gene perturbations and polymorphic variants. The human cancer cell xenograft has re-emerged as a minimal *in vivo* validation of these models.

9) Above all, resist the temptation to equate any partial functional evidence as sufficient. Published claims of functional relevance should be fully evaluated using the steps detailed above.

Freedman et al. (2011) Nature Genetics. 43(6): 513-518

OUTLINE

UNIVERSITY OF Southampton

- Background
- Regulatory Regions
  - DNase Sensitivity Assay
  - ChIP-Seq
  - ChIA-PET
- Expression quantitative traits locus (eQTL)
- Causal Inference Test (CIT)
- Model Systems
  - siRNA knockdown
  - CRISPR/Cas9
  - Cre/Lox
  - Patient derived tumour xenografts (PDTX)
- Summary

Genomic Context

UNIVERSITY OF Southampton

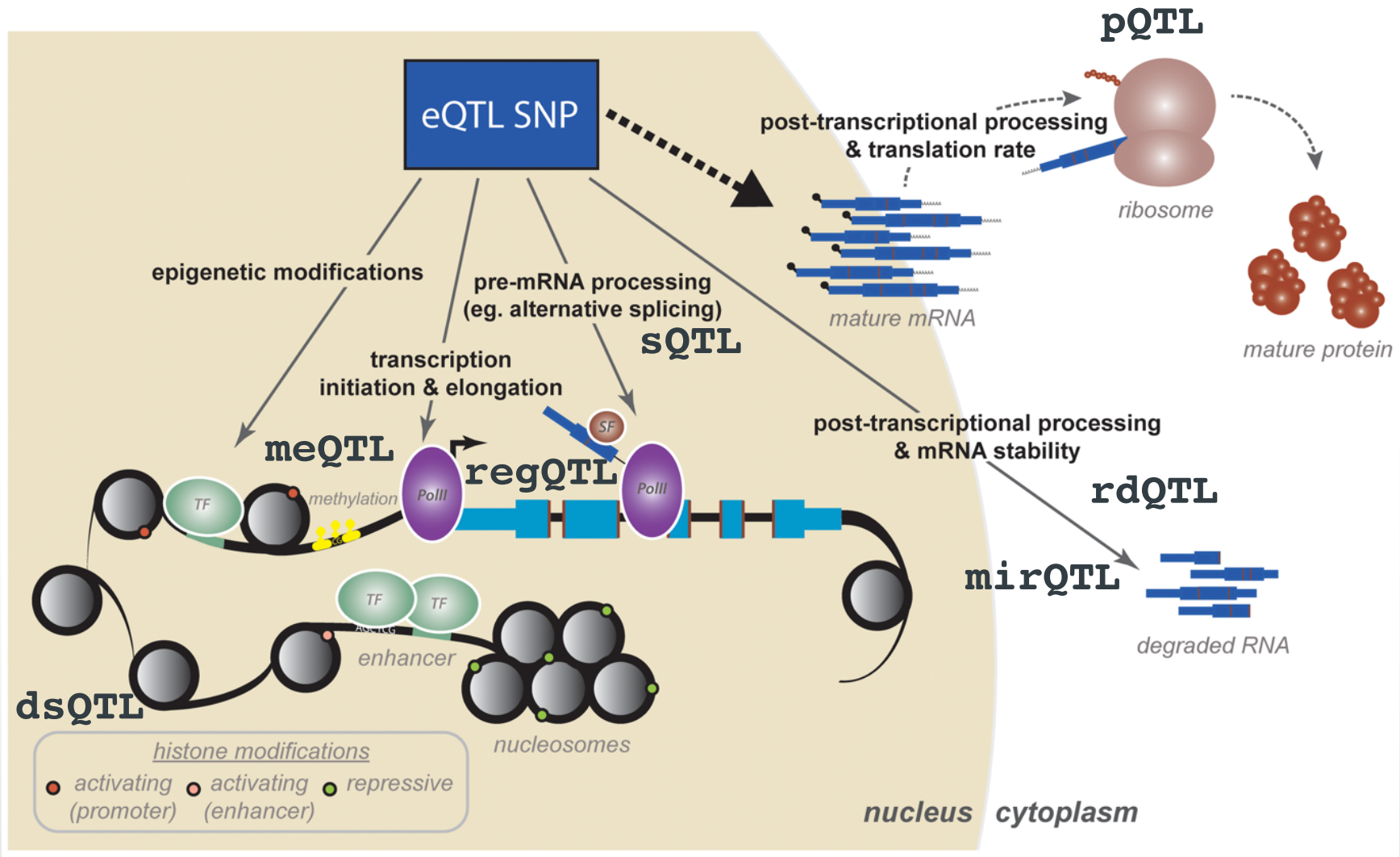
Table 1 The genomic context in which a variant is found can be used as preliminary functional analysis

Classification	Approximate percentages[a]	Approximate numbers[a]
Intronic	40	1,047
Intergenic	32	838
Within non-coding sequence of a gene	10	262
Upstream	8	210
Downstream	4	105
Non-synonymous coding	3	79
3′ untranslated region	~1	26
Synonymous coding	~1	26
5′ untranslated region		
Regulatory region		
Nonsense-mediated decay transcript		
Unknown	~1	26
Splice site		
Gained stop codon		
Frameshift in a coding sequence		

The table broadly summarizes the genomic context of disease- and trait-associated SNPs annotated in the Catalog of Genome-Wide Association Studies (http://www.genome.gov/gwastudies/) as of December 9th, 2010: 1,212 published genome-wide associations with $P < 5 \times 10^{-8}$ for 210 traits totaling 2,619 SNPs. Most of the SNPs are located in intergenic and intronic positions, but a small percentage are located upstream and downstream of genes, as well as in regulatory regions and splice sites. SNPs in these locations can be analyzed in more detail using more specific bioinformatics tools.
[a]Values are indicative and dependent on genomic boundaries used.

Freedman et al. (2011) Nature Genetics. 43(6): 513-518

SNP Functional Effects



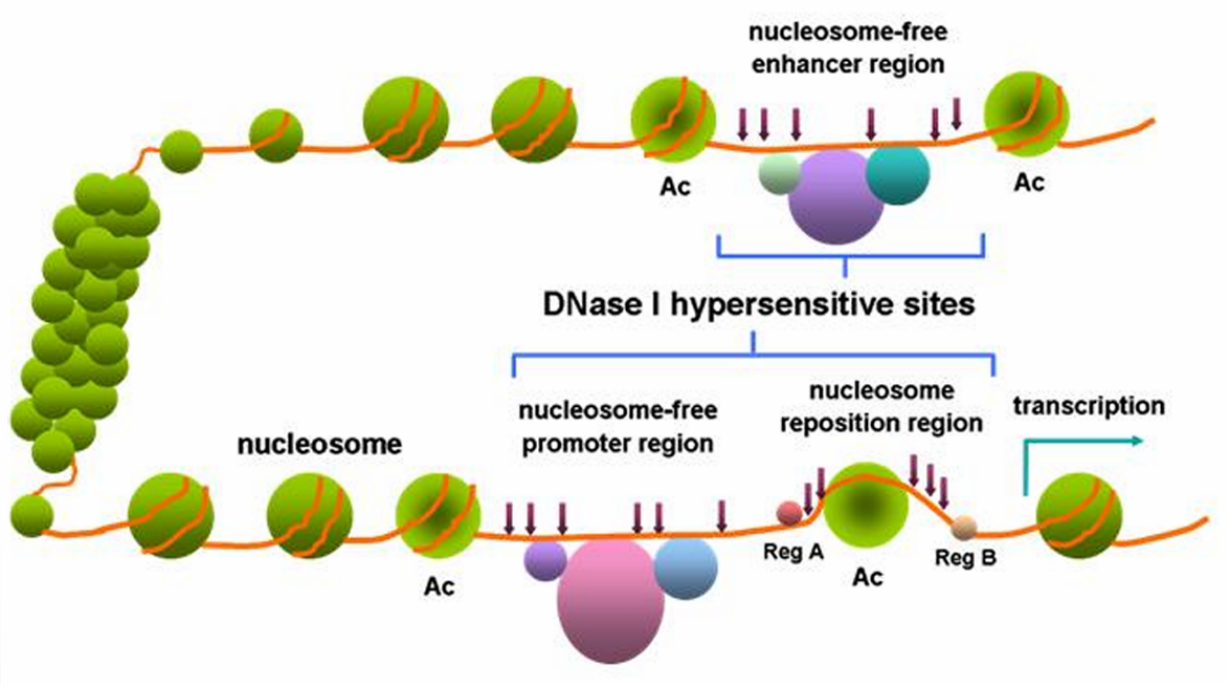
Pai et al. (2015) PLoS Genetics. 11(1):e1004857

UNIVERSITY OF Southampton

Regulatory Regions

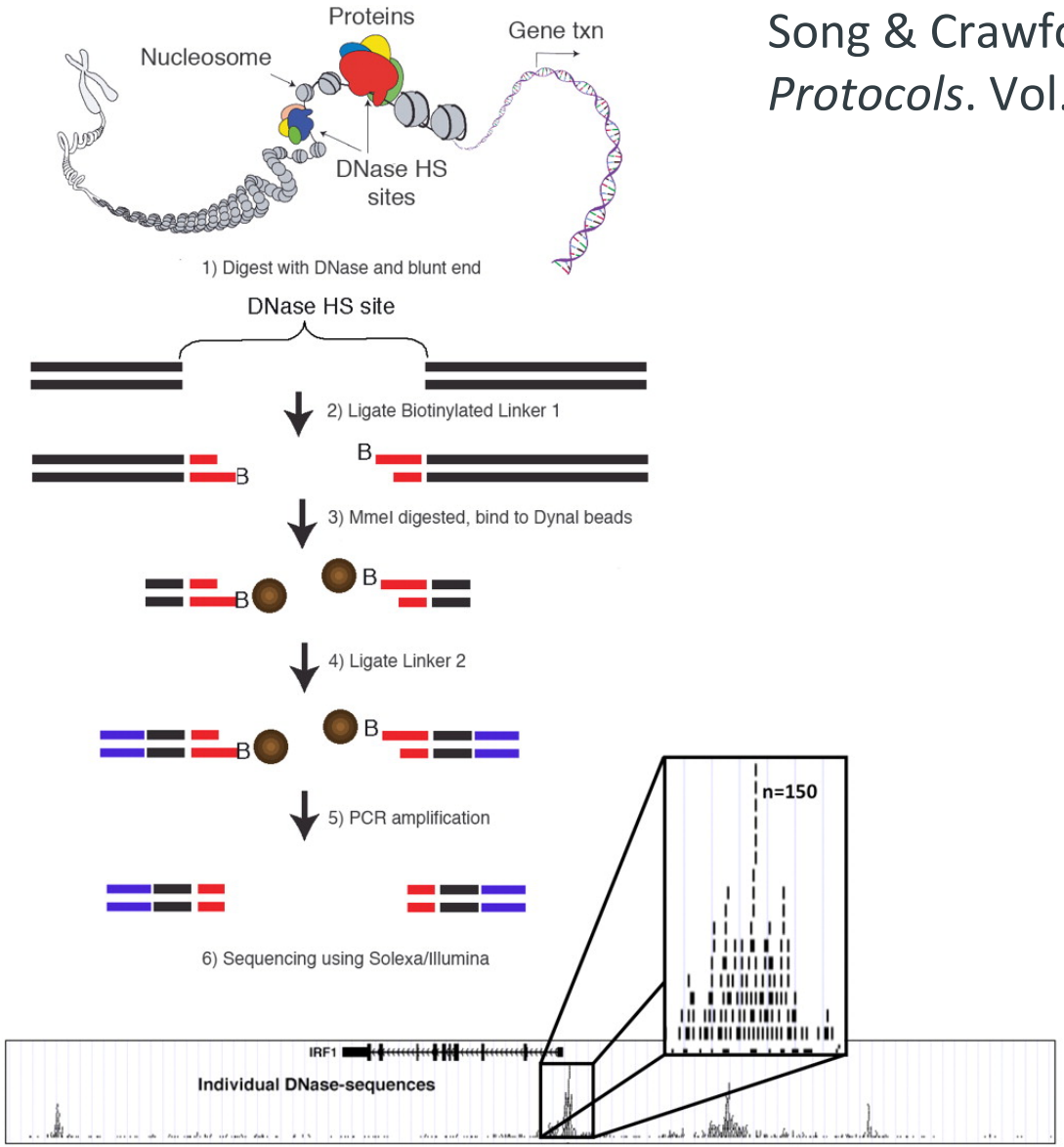
DNase I Sensitivity Assay

DNAse Sensitivity Assay



- Assay to detect active chromatin.
- DNase I hypersensitivity sites (DHSs) are regions of chromatin sensitive to cleavage by the DNase I enzyme.
- Chromatin has lots its condensed structure at a DHSs which functionally related to transcriptional activity since this remodelled stated is necessary for the binding of proteins such as transcription factors.

DNase-Seq



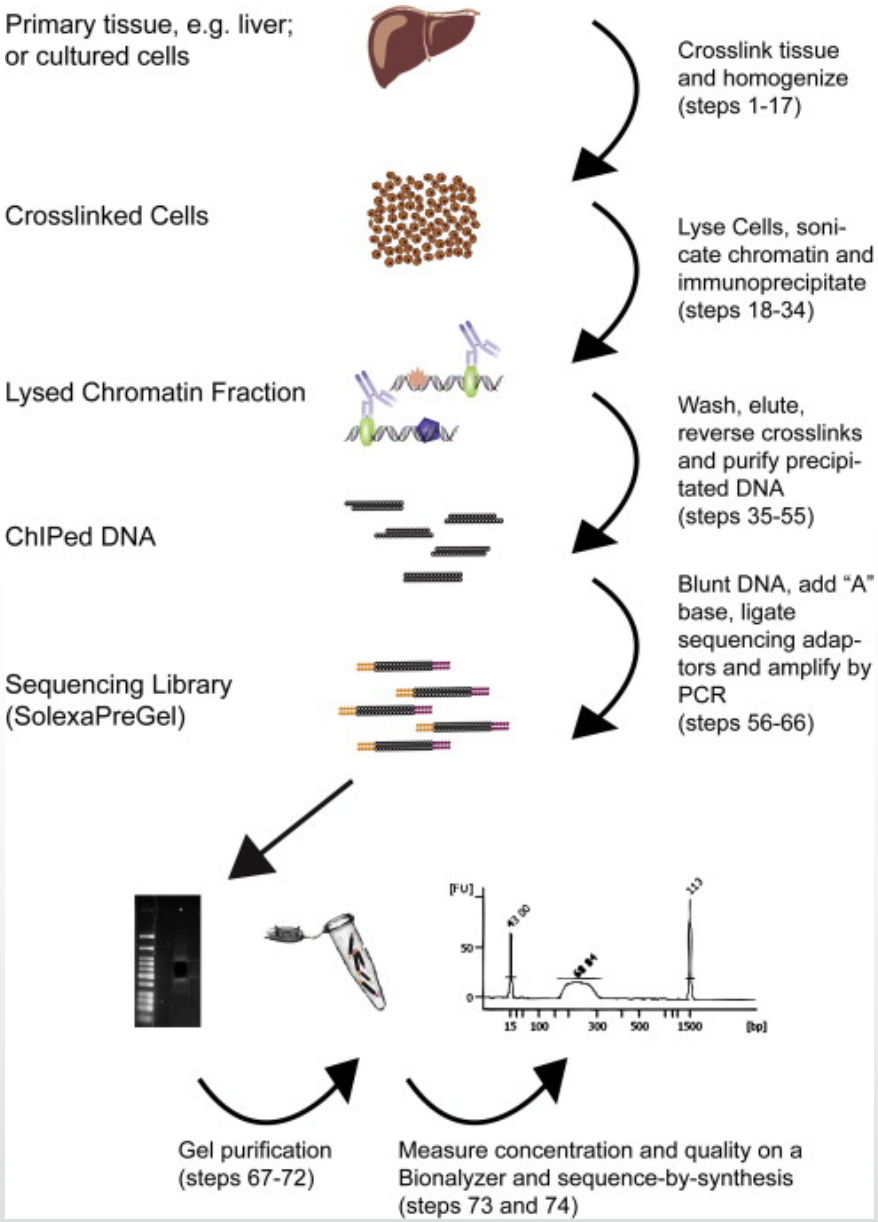
Song & Crawford (2010) *Cold Spring Harbor Protocols*. Vol. 2010 (2): pdb.prot5384

UNIVERSITY OF Southampton

Regulatory Regions

Chromatin Immunoprecipitation (ChIP)

ChIP-Seq



Schmidt et al. (2009) *Methods*. 48(3): 240-248.

- Transcription factor (TF) binding is primarily determined by the sequence context and is less driven by chromatin state.
- Many changes in TF binding do not seem to result in measurable changes in gene expression levels.
- We do not yet know how to distinguish between binding events that effect gene expression and those that do not (additional binding of TFs modifies complex?).

Strategy

Approaches for identifying targets of regulatory sequences:

- Knock out of regulatory sequences in model systems followed by genome-wide gene expression analysis to identify candidate targets (e.g., siRNA knockdown followed by RNA-Seq).
- Using regulatory sequences as baits in chromatin confirmation capture-based studies (e.g., ChIA-PET)
- Identify correlations between different genotypes of trait-associated SNPs and variations in the transcript abundance of candidate genes at these loci (e.g., eQTL and CIT analysis).

UNIVERSITY OF Southampton

Regulatory Regions

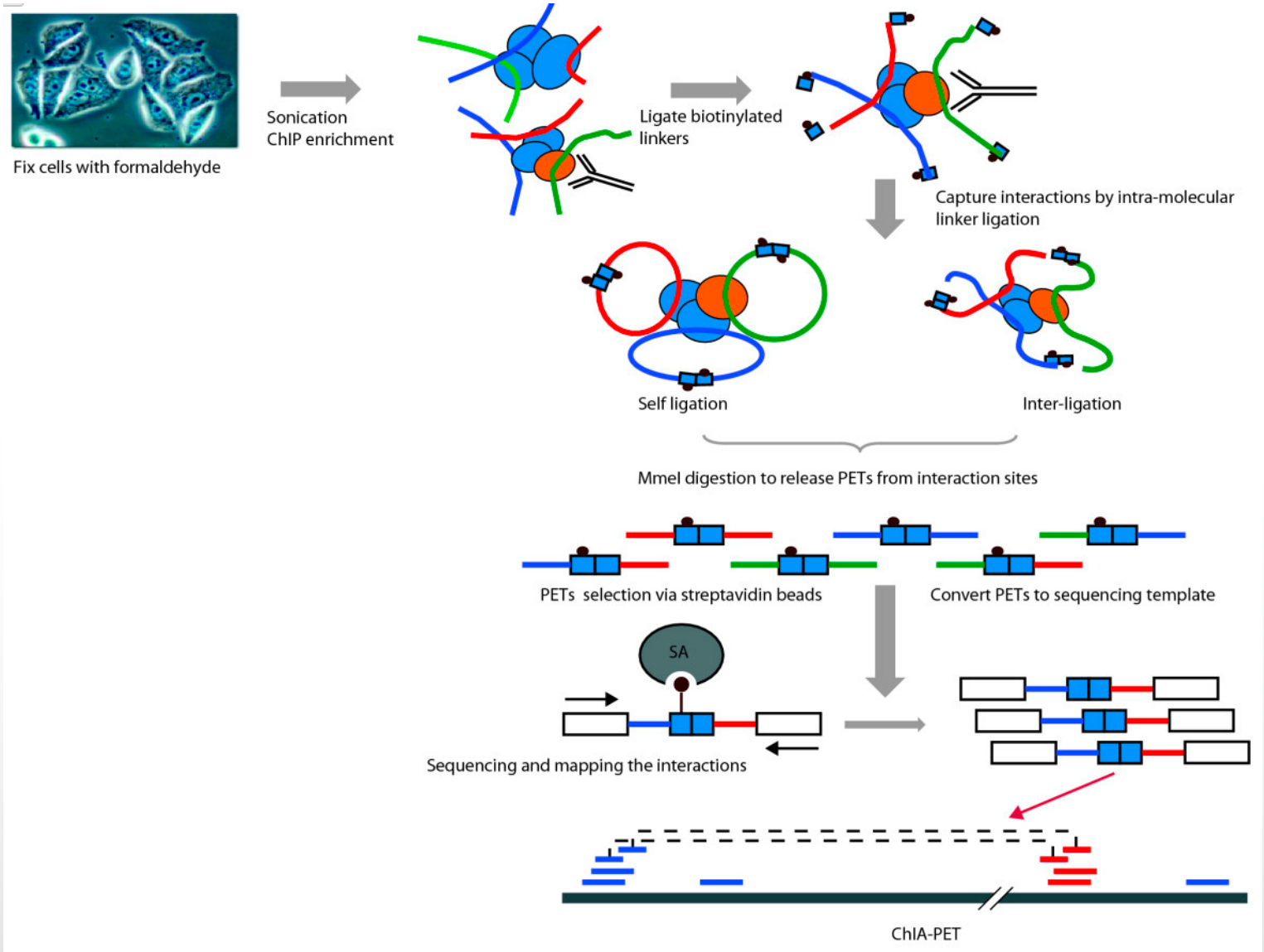
Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)

ChiA-PET

UNIVERSITY OF Southampton

- ChIP-Seq can only reveal the functional genome in a linear fashion.
- However, genes can be regulated by regions far from the promoter such as regulatory elements, insulators and boundary elements, and transcription-factor binding sites (TFBS).
- ChIA-PET is a technique that incorporates ChIP-based enrichment, chromatin proximity ligation, Paired-End Tags, and High-throughput sequencing to determine de novo long-range chromatin interactions genome-wide
- Specifically, ChIA-PET is be used to identify unique, functional chromatin interactions between distal and proximal regulatory transcription-factor binding sites and the promoters of the genes they interact with.

ChiA-PET



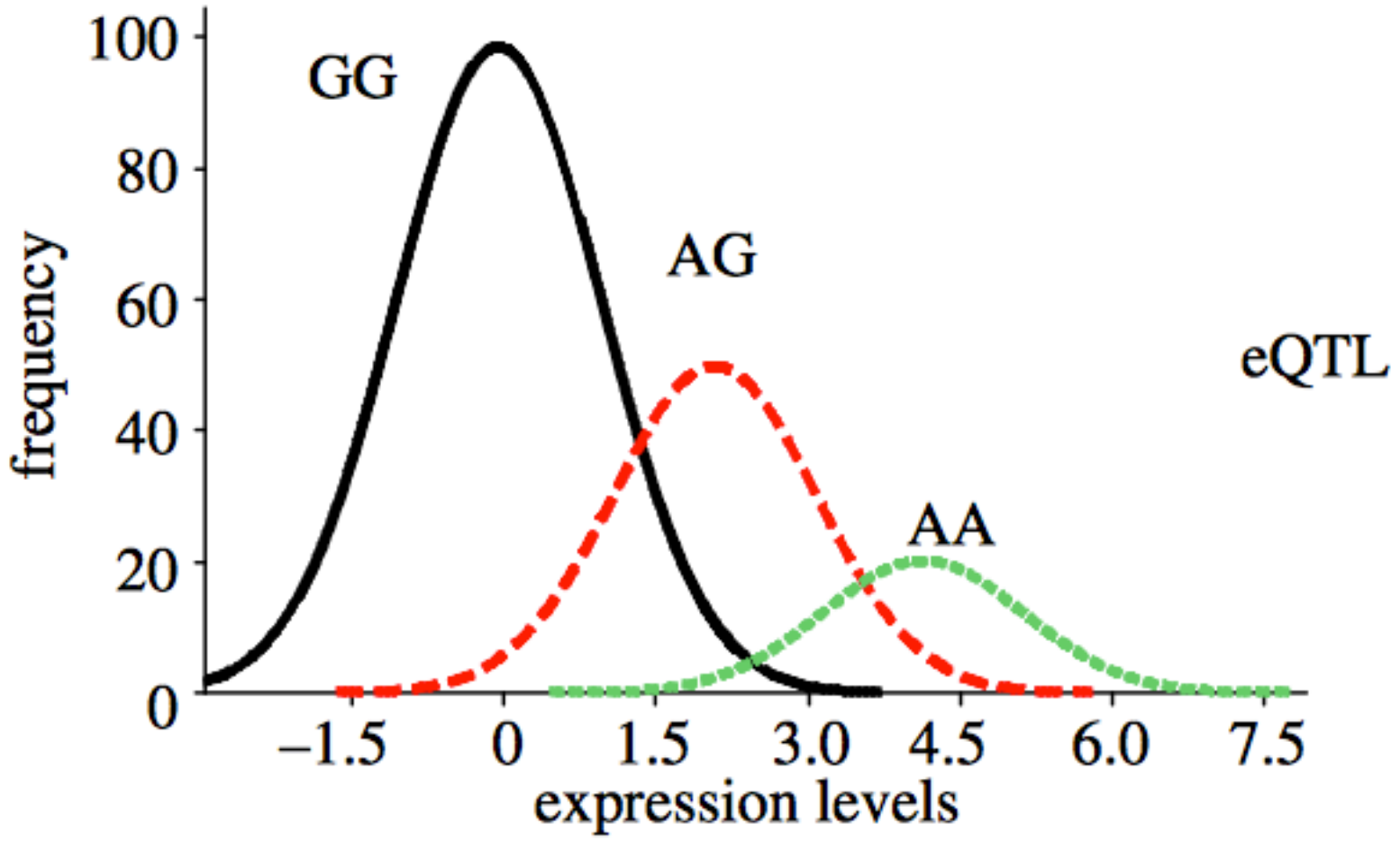
Wong & Wei (2009) *Genome Medicine*. 1:89

UNIVERSITY OF Southampton

eQTL

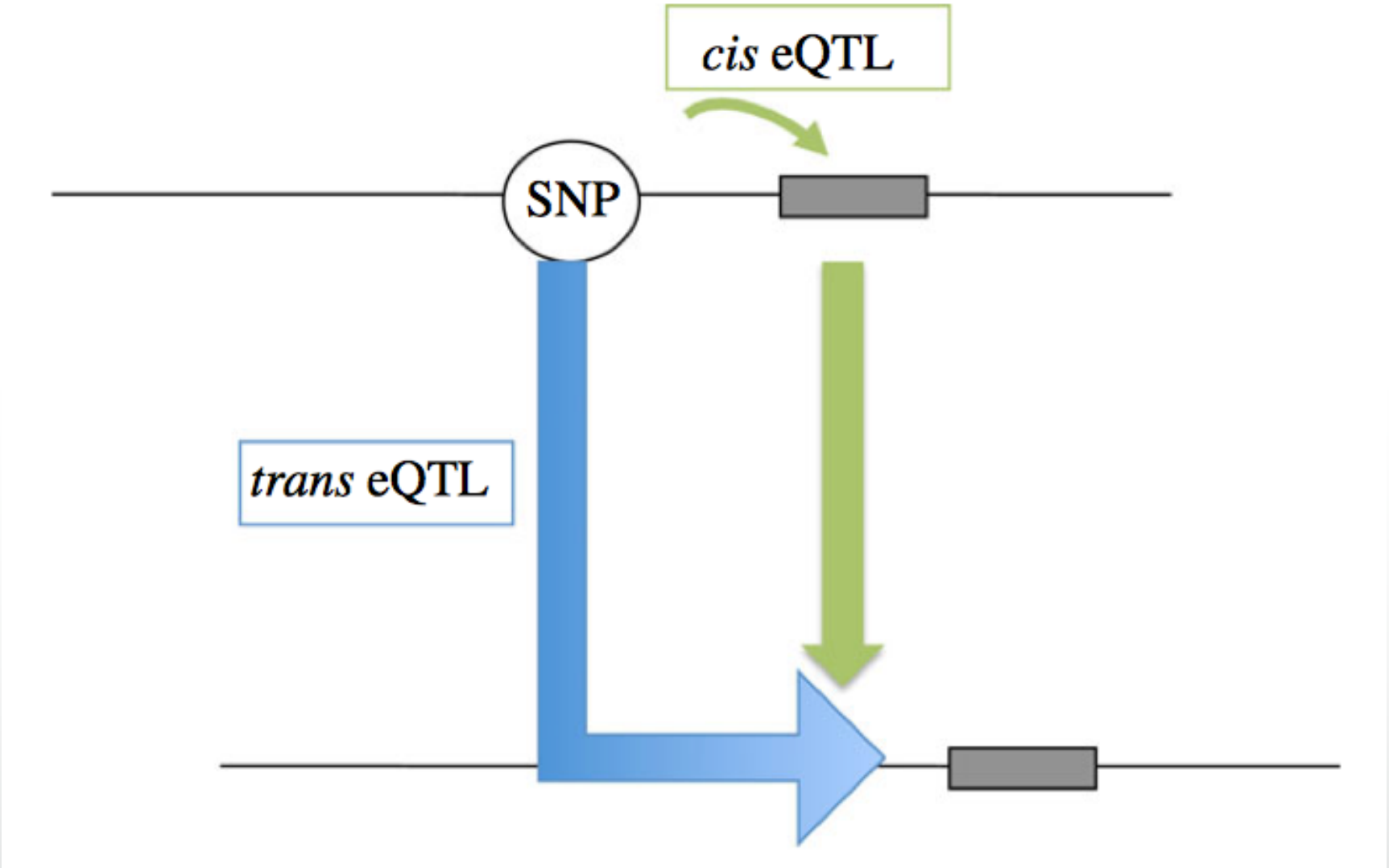
Expression Quantitative Traits Locus

Definition



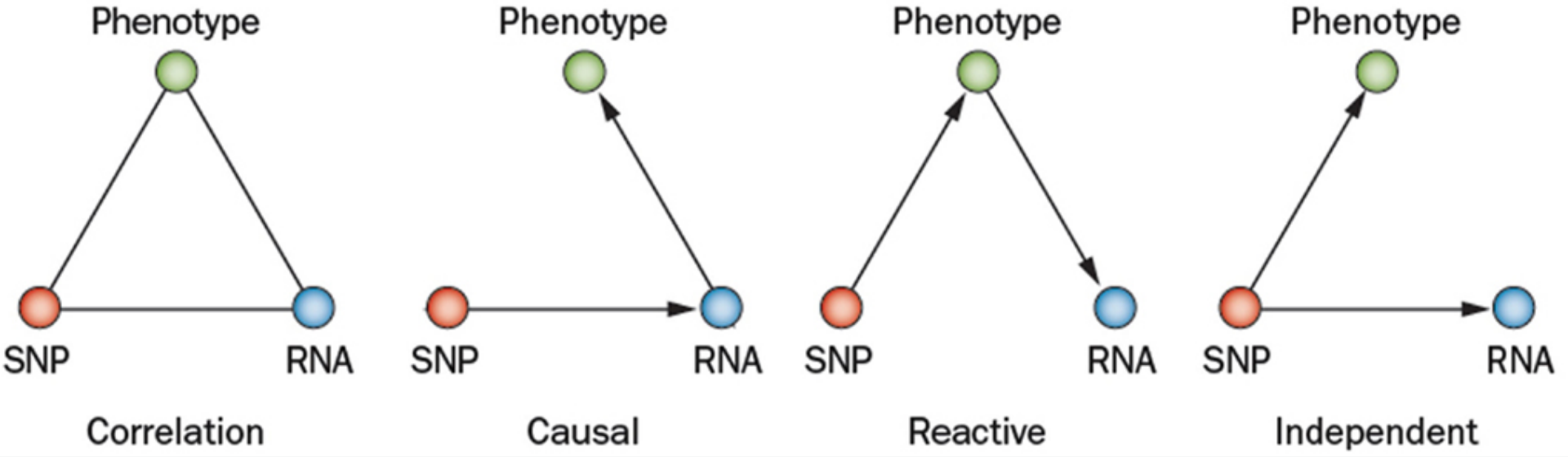
Nica et al. (2013) *Phil Trans R Soc* B 368: 20120362

Cis vs. *Trans*



Nica et al. (2013) *Phil Trans R Soc* B 368: 20120362

Relationships

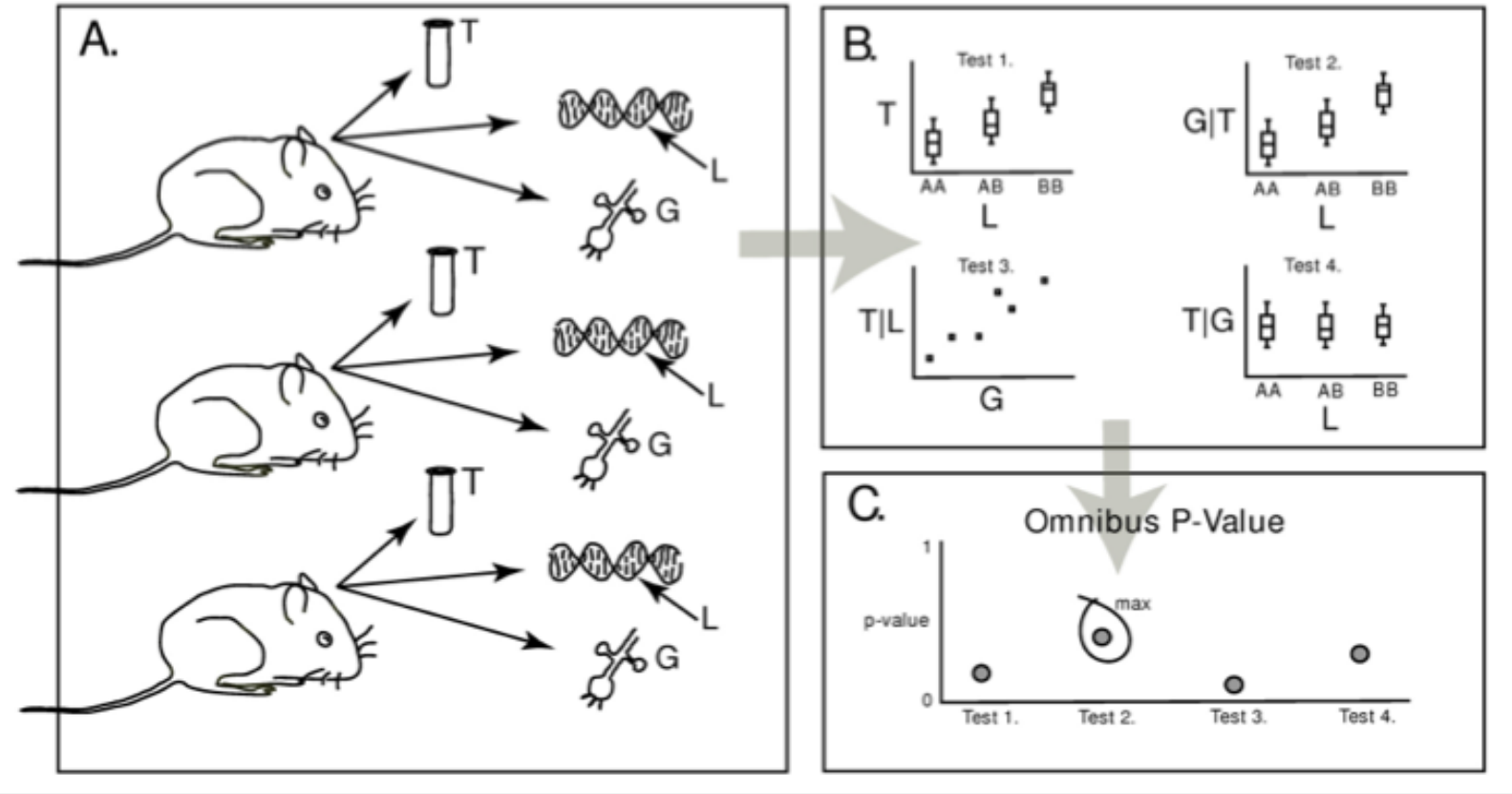


MacLellan et al. (2012) *Nature Reviews Cardiology*. 9:172-184

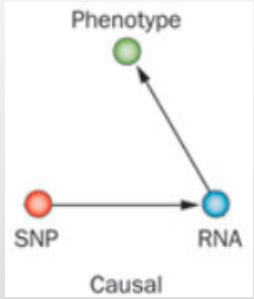
Causal Inference Test

- A likelihood-based hypothesis testing approach is implemented for assessing causal mediation.
- Causal inference is treated as a 'chain' of mathematical conditions that must be satisfied to conclude that the potential mediator is causal for the trait, where the inference is only as good as the weakest link in the chain.
- For example, it could be used to test for mediation of a known causal association between a DNA variant (SNP, aka L), and a clinical outcome or phenotype (T) by the potential mediator - gene expression (G).
- The hypothesis test generates a *p*-value or permutation-based FDR value with confidence intervals to quantify uncertainty in the causal inference.
- The outcome (T) can be represented by either a continuous or binary variable, the potential mediator is continuous (G), and the instrumental variable can be continuous or binary (SNP) and is not limited to a single variable but may be a design matrix representing multiple variables.
- *P*-values are computed for the component conditions, which include tests of linkage and conditional independence.
- The Intersection-Union Test, in which a series of statistical tests are combined to form an omnibus test, is then employed to generate the overall test result.

Causal Inference Test



1) L and T are associated,
2) L is associated with gene expression after adjusting for trait (G|T),
3) G is associated with trait after adjusting for SNP locus (T|L),
4) L is independent of trait after adjusting for gene expression (T|G).



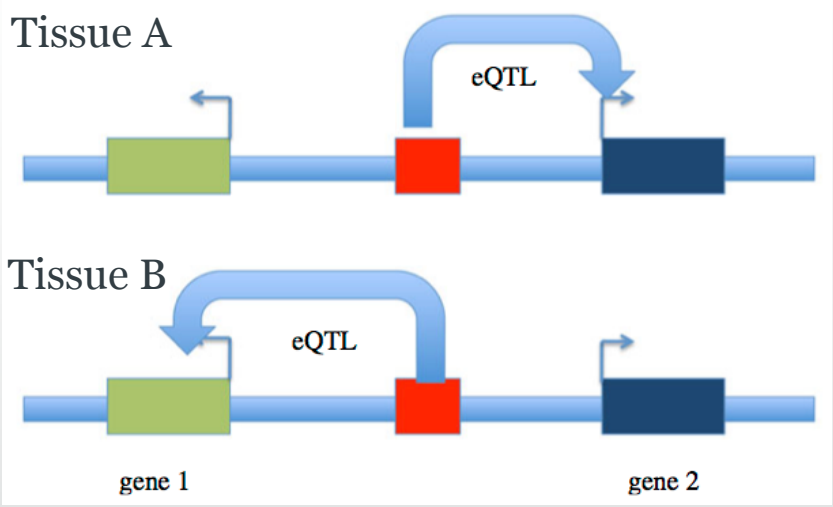
Implementation

- CRAN package for R: “cit: Casual Inference Test”
- GEPdb (http://ercsbweb.ewha.ac.kr/GEPdb/main.html): “A database for discovering the ternary association of genotype, gene expression and phenotype”.
- Sherlock (http://sherlock.ucsf.edu): “Discover disease genes in GWAS using eQTL signature matching”

Finding disease genes is elementary, my dear Watson!

Complexities

- Biobanks are needed to collect samples for additional assays (*e.g.*, RNA-Seq).
- Virtually ALL genes are likely to have at least one *cis*-acting SNP.
- Heritability studies suggest that >50% of the genetically explained variance in gene expression is due to *trans*-acting SNPs.
- Detection of *trans*-acting SNPs complicated by lower effect sizes and strict multiple test corrections.
- Common forms of disease are probably not the result of single SNPs with a single outcome but rather the outcome of perturbations of gene networks which are affected by complex genetic and environmental interactions.
- Since many traits manifest themselves only in certain tissues, such methods are only informative if expression measurements from disease-relevant cell-types are compared.
- Limited tissue interrogation will give misleading biological interpretations about the gene mediating the regulatory effect to increase disease risk.



UNIVERSITY OF Southampton

Model Systems

Cells, Tissues & Animals

Model Systems

Once there is sufficient evidence in support of a candidate susceptibility gene then more detailed functional studies are required to characterize the gene's role in the pathogenesis of the trait.

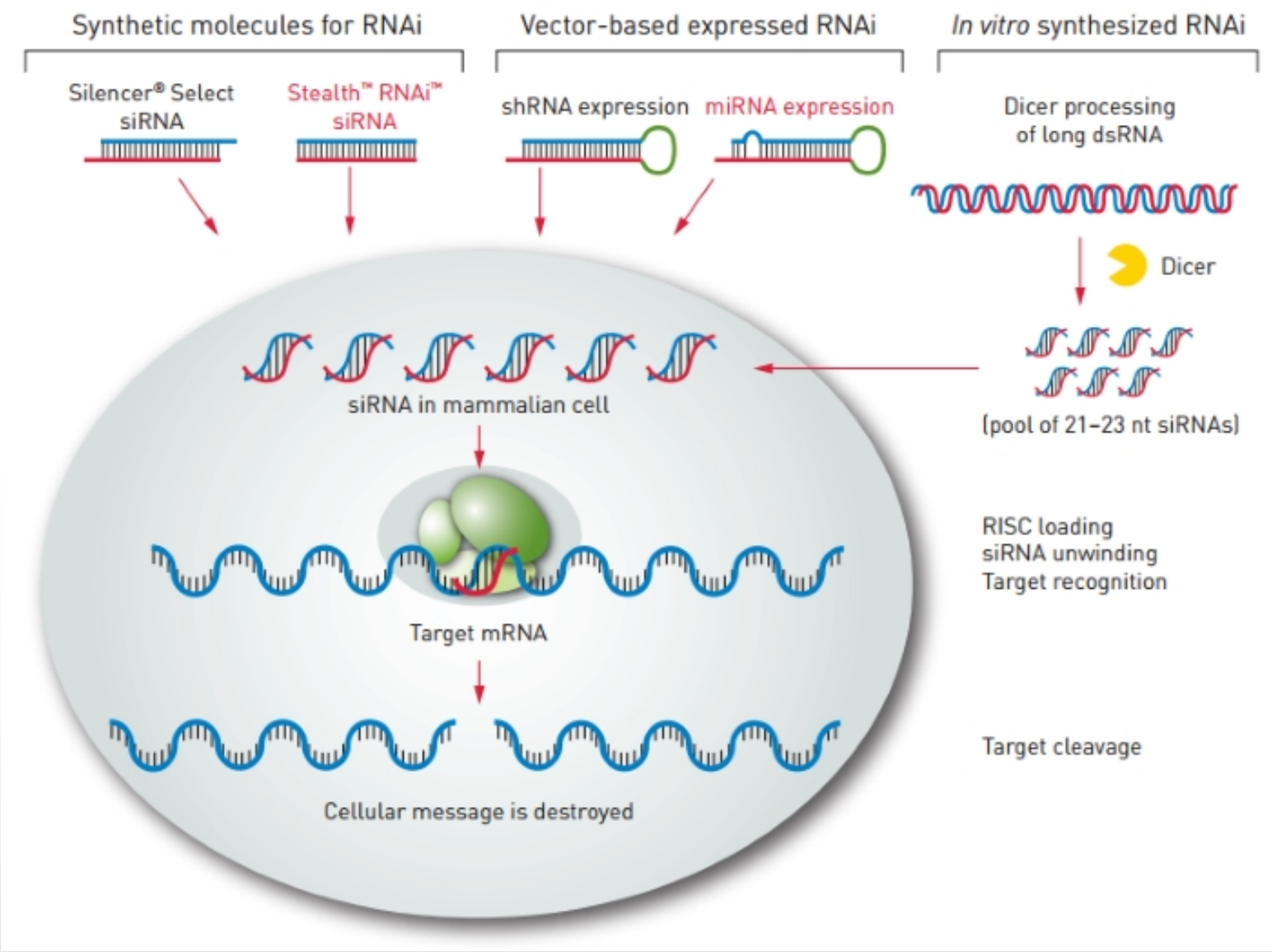
In vitro models:

- Cell lines
- Primary cells
- Tissue cultures
- 3D models (e.g. bioelectrosprayer)

In vivo models:

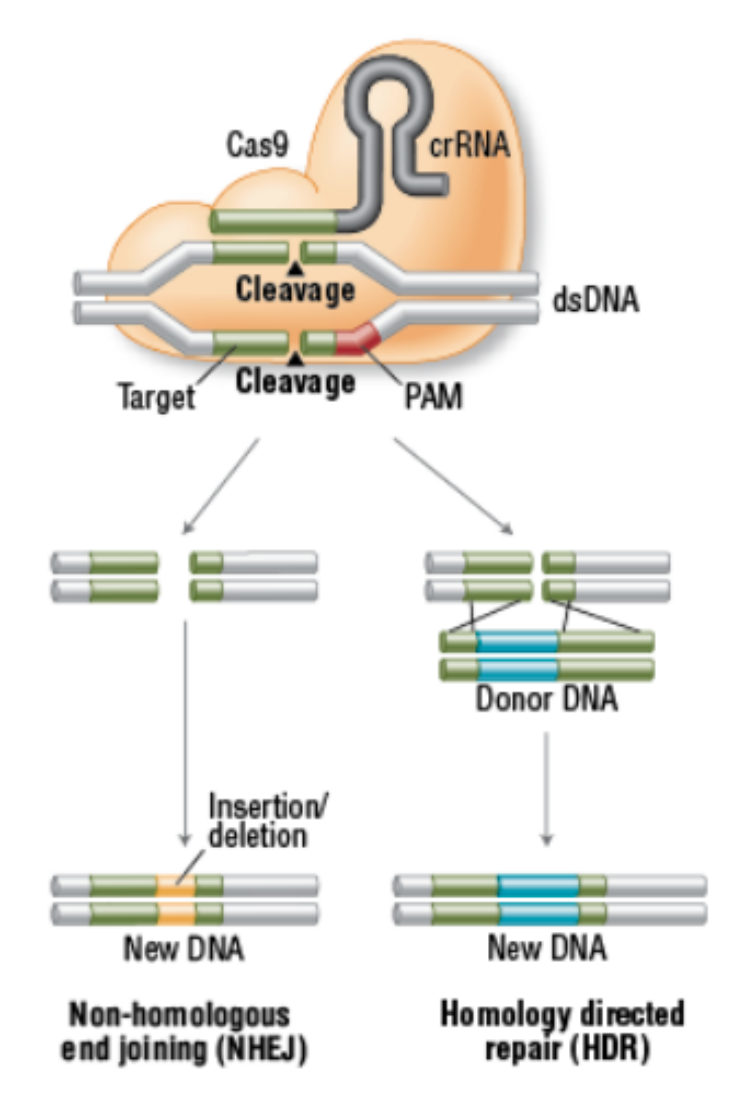
- Mouse, rat, monkey, other animal models
- Cre/Lox mouse strains
- Patient derived tumour xenografts (PDTX)

siRNA Knockdown



https://www.thermofisher.com/uk/en/home/references/gibco-cell-culture-basics/transfection-basics/rnai-non-coding-rna-research.html

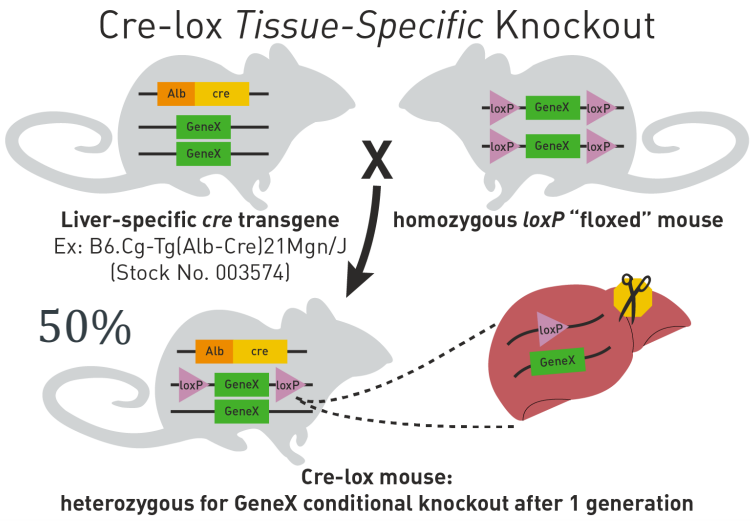
CRISPR/Cas9

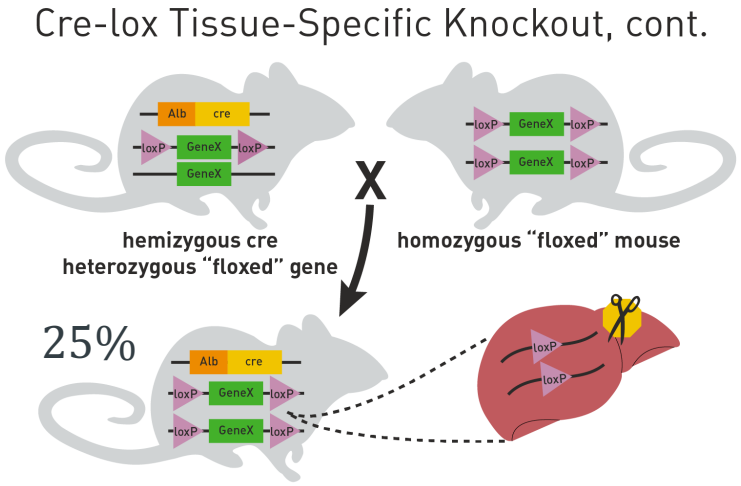


- Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated (Cas) genes are essential in adaptive immunity in select bacteria and archaea.
- They enable the organisms to respond to and eliminate invading genetic material.
- For example, *S. thermophilus* can acquire resistance against a bacteriophage by integrating a genome fragment of an infectious virus into its CRISPR locus.
- Cas9 can site-specifically cleave double-stranded DNA resulting in the activation of the doublestrand break (DSB) repair machinery.
- DSBs can be repaired by the cellular Non-Homologous End Joining (NHEJ) pathway, resulting in insertions and/or deletions (indels) which disrupt the targeted locus.
- Alternatively, if a donor template with homology to the targeted locus is supplied, the DSB may be repaired by the homology-directed repair (HDR) pathway allowing for precise replacement mutations to be made.

https://www.neb.com/tools-and-resources/feature-articles/crispr-cas9-and-targeted-genome-editing-a-new-era-in-molecular-biology

Cre-Lox

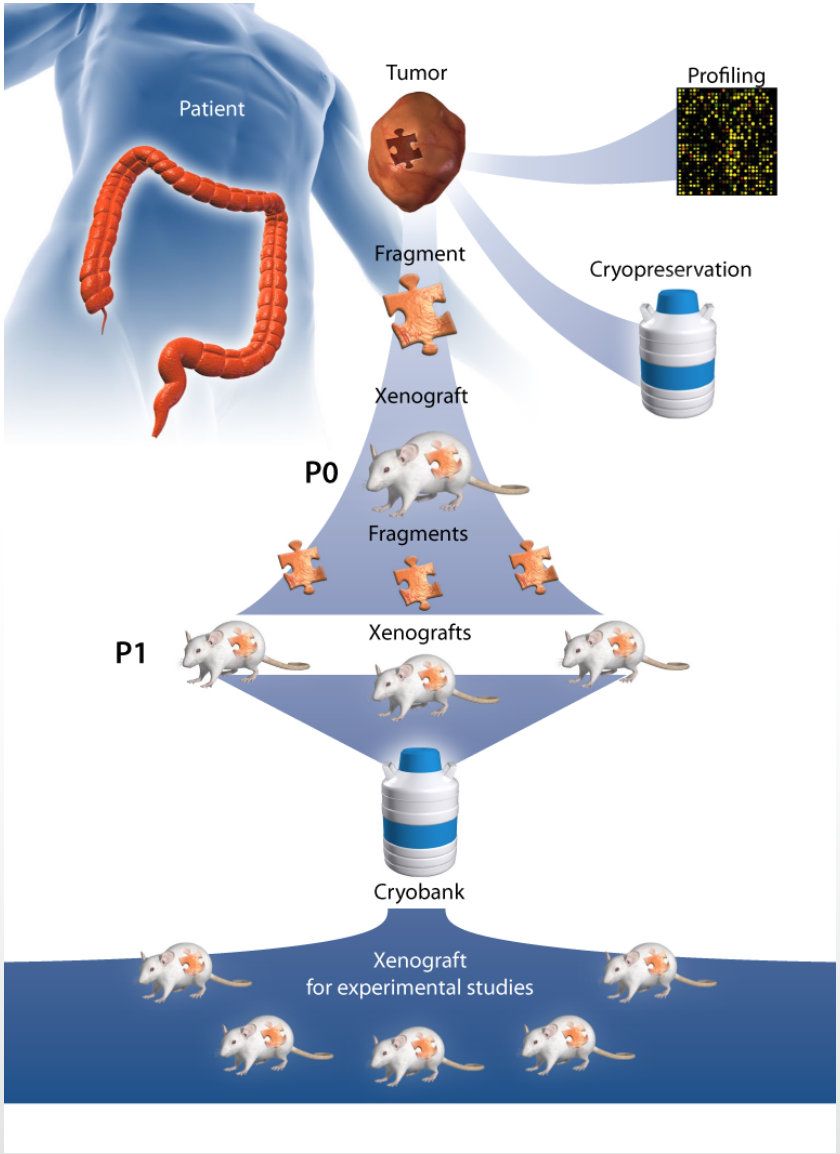




- Cre/lox is usually used to make knockout alleles, but it can also be used to activate gene expression.
- loxP sequences can be artificially inserted into animals or plants and used for the precise excision of DNA.
- Transgenic mice containing a gene surrounded by loxP sites are mated with transgenic mice that have the *cre* gene expressing only in one cell type.
- The resulting mice have the *cre* gene and the loxP-flanked gene.
- In tissues with no *cre* gene the target gene will be present and function normally.
- In a cell where *cre* is expressed it catalyses a double stranded cut of the DNA at both loxP sites, which are then ligated back together resulting in deletion of the target gene.

https://www.jax.org/news-and-insights/jax-blog/2011/september/cre-lox-breeding-for-dummies

Patient Derived Tumor Xenografts



http://www.intechopen.com/source/html/40843/media/image6.png

Complexities

- Models are generally limited to studying one variant or gene at a time and thus do not assess the complex interplay between SNPs.
- Short duration of experiments in models in contrast to diseases that develop over several decades.
- Differences in human vs. animal physiology.
- Differences in the structure and sequence of non-coding regions.
- Limited modelling of gene-environment interactions.
- Greatest functional impact of a SNP/gene interaction may be seen in healthy tissue but most cell line models are akin to cancer and thus represent an aberrant background.

SUMMARY

- Background
- Regulatory Regions
  - DNase Sensitivity Assay
  - ChIP-Seq
  - ChIA-PET
- Expression quantitative traits locus (eQTL)
- Causal Inference Test (CIT)
- Model Systems
  - siRNA knockdown
  - CRISPR/Cas9
  - Cre/Lox
  - Patient derived tumour xenografts (PDTX)