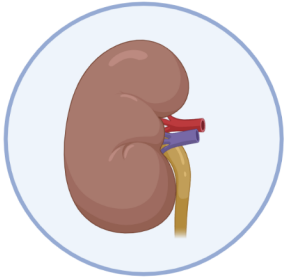
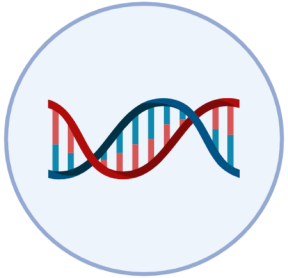
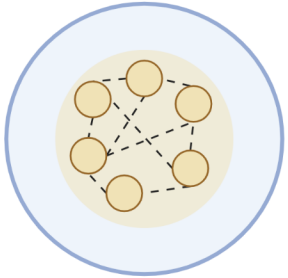
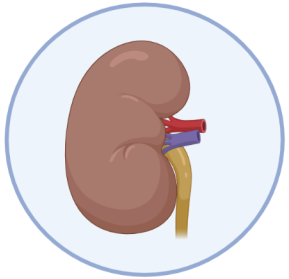
Data-driven Diagnosis: Using R to Advance Kidney Disease Research

Claire Hill
Research Fellow at Queen's University Belfast

A bit about me

2011 - 2016	Undergraduate MSci Biochemistry *Queen's University Belfast*	*No real coding experience* *Beginner bioinformatics*
2016 - 2021	PhD Interdisciplinary Bioscience *University of Oxford*	*Python* *Pandas* *Jupyter notebook* *Bioinformatics*
2021 - present	Molecular Epidemiology Research Fellow *Queen's University Belfast*	*R / R studio* Big data analysis eCarlton - R for Biologists - Free course



Why focus on kidney disease research?

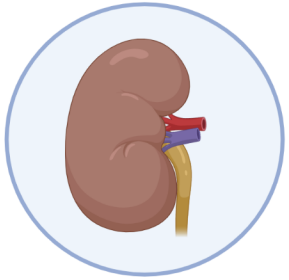
	2017	2040
- Leading cause of death:	12th	5th

- End-stage kidney disease, requiring **dialysis / kidney transplant.**

- UN Goal 3 - Decrease non-communicable disease 1/3 by 2030

SUSTAINABLE DEVELOPMENT GOALS

United Nations, 2015
Carney, 2020
Bikbov et al., 2020
Foreman et al., 2019



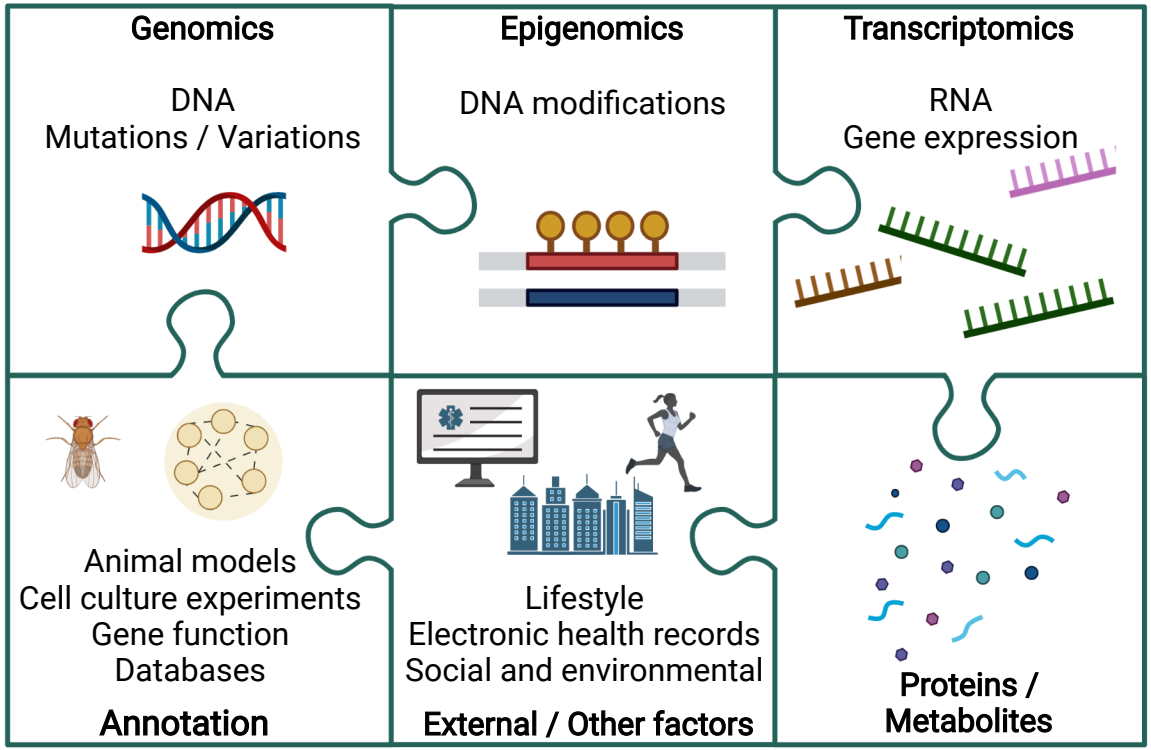
Methods to study kidney disease

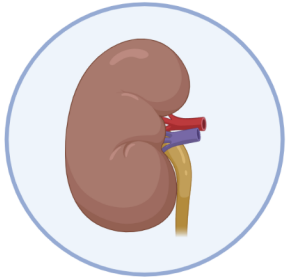
- Study **multiple factors** to understand complexity.

- Data reliability and quality control.

- **Belfast Renal Transplant** cohort:

50 years follow-up





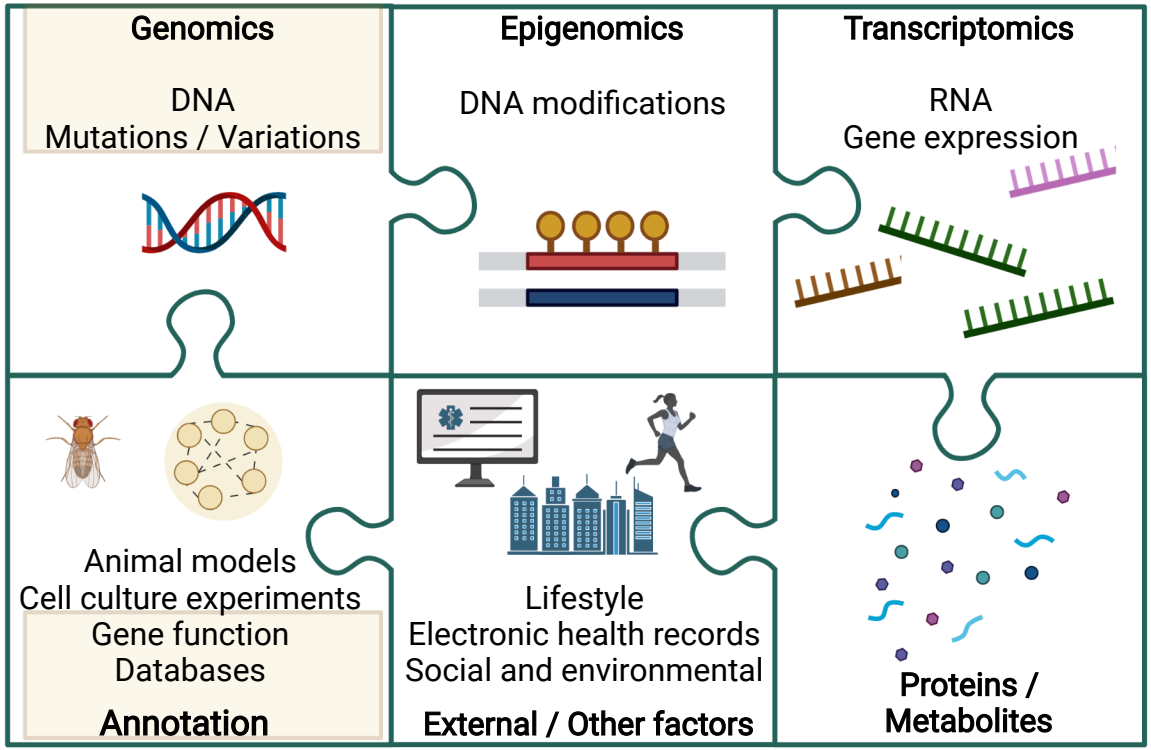
Methods to study kidney disease

- Study **multiple factors** to understand complexity.

- Data reliability and quality control.

- **Belfast Renal Transplant** cohort:

50 years follow-up

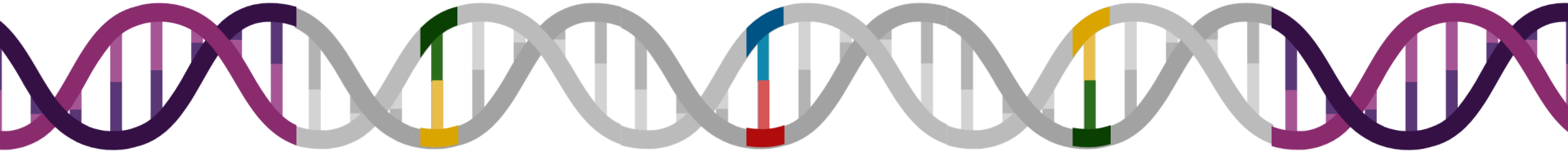


Genomic analysis of kidney disease

- Genetic differences influence **onset** and **progression**.

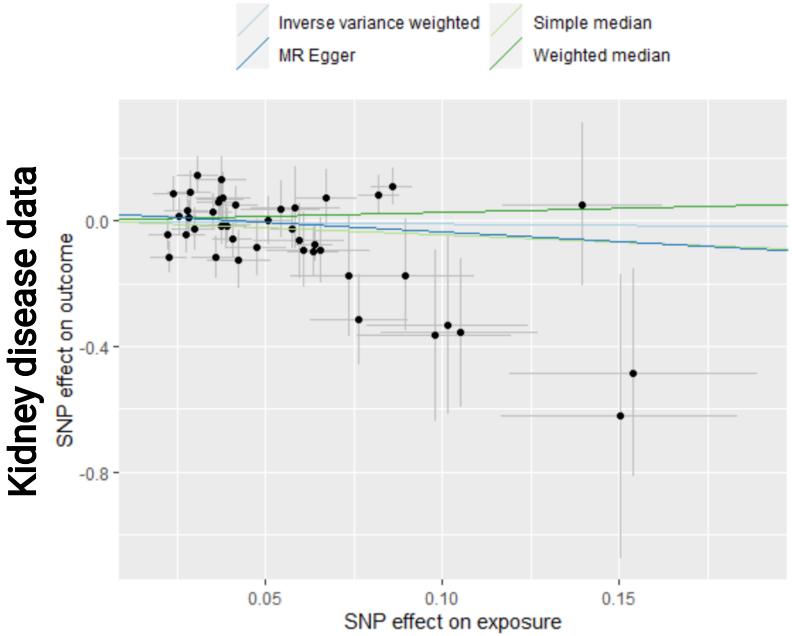
- Identifying **genetic variants** to **aid diagnosis**.

- **Overlap** with variants associated with modifiable risks to identify **causal associations**, e.g **Telomere** shortening.



Genomic analysis of kidney disease

```
library(MendelianRandomization)
library(TwoSampleMR)
#read in outcome data (e.g. kidney disease)
outcome_data <- read_outcome_data(....)
#read in exposure data (e.g. telomere associated variants)
exposure_df <- read_csv("~/MyDocuments/exp.csv")
#format your data for MR
exposure_data <- format_data(exposure_df, type="exposure")
#harmonise datasets
harmonised_data <- harmonise_data(
  exposure_dat = exposure_data ,
  outcome_dat = outcome_data)
#run multiple MR analyses at once
harmonised_res <- mr(harmonised_data,
method_list=c("mr_ivw", "mr_simple_median",
"mr_weighted_median","mr_egger_regression"))
#generate a range of plots e.g.
harmonised_p1 <- mr_scatter_plot(harmonised_res, harmonised_data)
```



Park at al., 2021

Hill and Duffy et al., in preparation

Genomic analysis of kidney disease

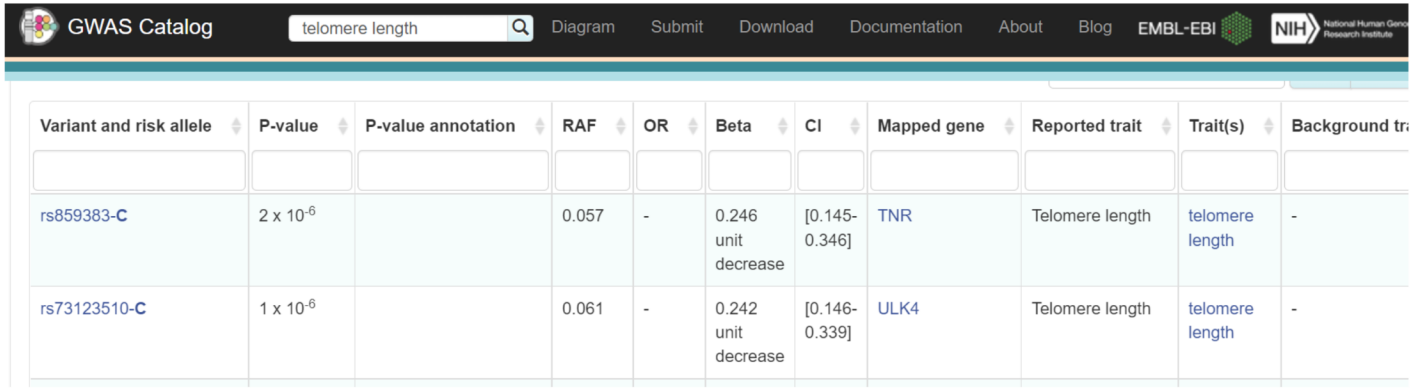
- Online version: MRBASE University of BRISTOL MRC Integrative Epidemiology Unit CANCER RESEARCH UK

- Previously published summary statisitcs:

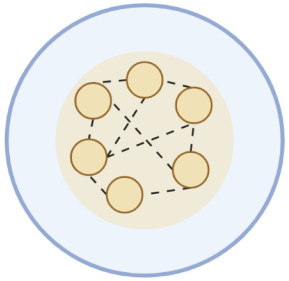
No identifiable data relased - **data protection**.

Sharing resources saves research time, effort and money.

R library **MRInstruments** (*MRC Integrative Epidemiology Unit*) pulls this data into R.



Variant and risk allele	P-value	P-value annotation	RAF	OR	Beta	CI	Mapped gene	Reported trait	Trait(s)	Background tra
rs859383-C	2 x 10^{-6}		0.057	-	0.246 unit decrease	[0.145-0.346]	TNR	Telomere length	telomere length	-
rs73123510-C	1 x 10^{-6}		0.061	-	0.242 unit decrease	[0.146-0.339]	ULK4	Telomere length	telomere length	-

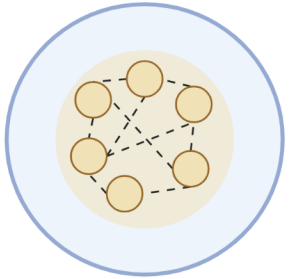


Gene ontology analysis of kidney disease

- Discover novel genetic variants associated with kidney disease, but…
  what do these genes, or the proteins they encode, do?

- Annotation via gene ontology databases:
  Molecular function,
  Biological processes,
  Cellular compartments.

- R package for analysis and **data visualisation** to aid interpretation:
  ViSEAGO (*Brionne et al., 2019*)



Gene ontology analysis of kidney disease

ViSEAGO

Background genes

Genes of interest Kidney disease genes

Enriched gene ontologies

Clustering of similar gene ontologies

Highlight hits for future study

```
##################
# load genes background
background<-scan(
    "background.txt",
    quiet=TRUE,
    what=""
)

##################
# load gene selection
selection<-scan(
    "selection.txt",
    quiet=TRUE,
    what=""
)
```

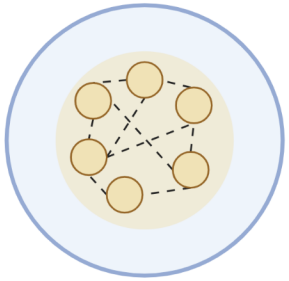
```
##################
# connect to EntrezGene
EntrezGene<-ViSEAGO::EntrezGene2GO()

 ##################
 # load GO annotations from EntrezGene
 myGENE2GO<-ViSEAGO::annotate(
     "EntrezGene_id",
     EntrezGene
 )

 ##################
 # create topGOdata for BP
 BP<-ViSEAGO::create_topGOdata(
     geneSel=selection,
     allGenes=background,
     gene2GO=myGENE2GO,
     ont="BP",
     nodeSize=5
 )
```

```
##################
# initialyse
myGOs<-ViSEAGO::build_GO_SS(
    gene2GO=myGENE2GO,
    enrich_GO_terms=BP_sResults
)
```

```
##################
# compute all available Semantic Similarity (SS) measures
myGOs<-ViSEAGO::compute_SS_distances(
    myGOs,
    distance="Wang"
)
```

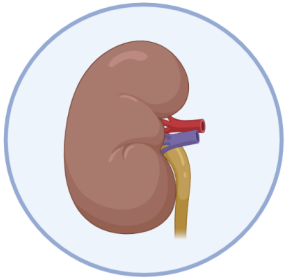
```
Wang_clusters_wardD2<-ViSEAGO::GOterms_heatmap(
    myGOs,
    showIC=TRUE,
    showGOlabels=TRUE,
    GO.tree=list(
        tree=list(
            distance="Wang",
            aggreg.method="ward.D2"
        ),
        cut=list(
            dynamic=list(
                pamStage=TRUE,
                pamRespectsDendro=TRUE,
                deepSplit=2,
                minClusterSize =2
            )
        )
    ),
    samples.tree=NULL
)
```

Brionne et al., 2019



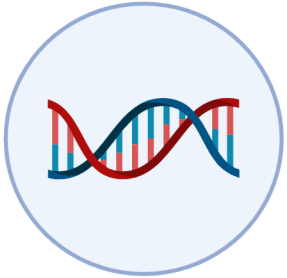
Hill and Duffy et al., in preparation

Key points



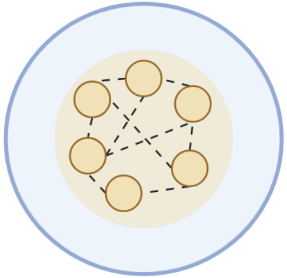
- Kidney disease is a **leading cause of death**, globally.

- **Integrating multiple approaches** advances diagnosis and treatment.



- **Genetic variants** associated with kidney disease used to aid diagnosis.

- **Mendelian Randomisation** used to identify new causal risk exposures.



- **Gene annotation** translates this knowledge into **functional insights**.

- Guiding experimental analysis for **diagnostic and therapeutic development**.

Awknowledgements

- Centre for Public Health
Molecular Epidemiology group:
Prof. AJ McKnight
Dr. Laura Smyth
Dr. Katie Kerr
Dr. Seamus Duffy
Jill Kilner
all staff and students.

- Jordan Jones

- Women Techmakers Belfast Team

- Slides created via **Biorender.com**

Fondúireacht Eolaíochta Éireann
Dá bhfuil romhainn
Science Foundation Ireland
For what's next

DfE
Department for the Economy
www.economy-ni.gov.uk

Key Tools / Links

R for Biologists (eCarlton) - Free online course

Mendelian Randomisation (Yavorska and Staley)

Two sample MR (MRC Integrative Epidemiology Unit)

MRBase (MRC Integrative Epidemiology Unit)

MRInstruments (MRC Integrative Epidemiology Unit)

ViSEAGO (Brionne et al., 2019)

References

Bikbov B, Purcell CA, Levey AS, Smith M, Abdoli A, Abebe M, et al. **Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017** . Lancet. 2020;395(10225):709–33.

Foreman KJ, Marquez N, Dolgert A, Fukutaki K, Fullman N, McGaughey M, et al. **Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories.** Lancet. 2018;392(10159):2052–90.

Jankowski J, Floege J, Fliser D, Böhm M, and Marx N, **Cardiovascular Disease in Chronic Kidney Disease: Pathophysiological Insights and Therapeutic Options.** Circulation. 2021;43(11):1157-1172

United Nations. UN General Assembly, **Transforming our world: the 2030 Agenda for Sustainable Development** (A/RES/70/1). 2015.

Carney EF. **The impact of chronic kidney disease on global health.** Nat Rev Nephrol. 2020;16(5):251.

Park S, Lee S, Kim Y, Cho S, Kim K, Kim YC, et al. **A Mendelian randomization study found causal linkage between telomere attrition and chronic kidney disease** . Kidney Int. 2021

Hemani G, Zheng J, Elsworth B, Wade KH, Baird D, Haberland V, et al, **The MR-Base Collaboration. The MR-Base platform supports systematic causal inference across the human phenome**. eLife. 2018;7:e34408.

Yavorska O, Burgess S, **MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data.** International Journal of Epidemiology. 2017;46(6):1734–1739.

Burgess S, Davey Smith G, Davies NM, Dudbridge F, Gill D, Glymour MM, et al. **Guidelines for performing Mendelian randomization investigations** . Wellcome Open Res. 2020;4:186.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. **The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019** . Nucleic Acids Research. 2019;47:D1005-D1012.

Brionne A., Juanchich A. and Hennequet-Antier C. **ViSEAGO: a Bioconductor package for clustering biological functions using Gene Ontology and semantic similarity** . BioData Mining; 2019;12(16).

Brionne A, Juanchich A, and Hennequet-Antier C. **An overview of ViSEAGO: Visualisation, Semantic similarity, Enrichment Analysis of Gene Ontology.** [Internet]. 2020 Available from: https://bioconductor.org/packages/devel/bioc/vignettes/ViSEAGO/inst/doc/ViSEAGO.html

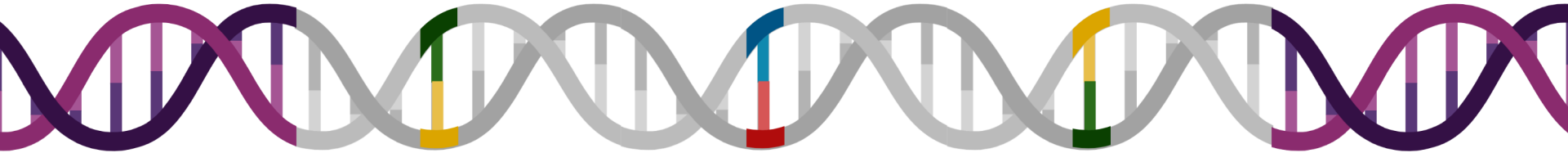
Levin A, Reznichenko A, Witasp A, Liu P, Greasley PJ, Sorrentino A, Bruchfeld A, et al. **Novel insights into the disease transcriptome of human diabetic glomeruli and tubulointerstitium** . Nephrol. Dial. Transplant. 2059–2072, 2020

Extra slides

Genomic analysis of kidney disease

- **Telomere shortening:**
  - Premature aging
  - Age related diseases
  - Diabetes mellitus
  - Cardiovascular disease
  - Hypertension

Is this associated with onset / progression of kidney disease?

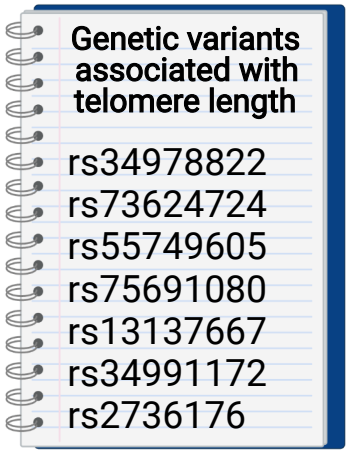


Genomic analysis of kidney disease

- Mendelian Randomisation (MR):

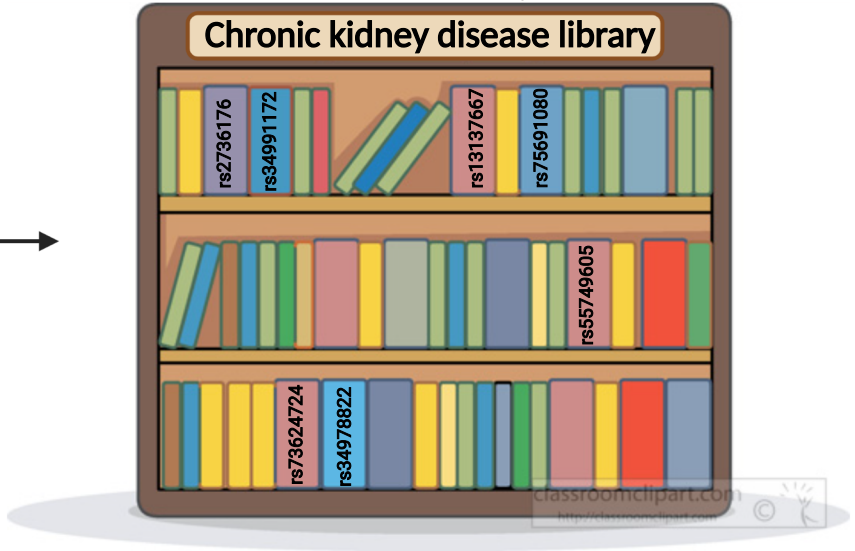
Step 1

*Genetic variants known to be associated with **proposed risk exposure***



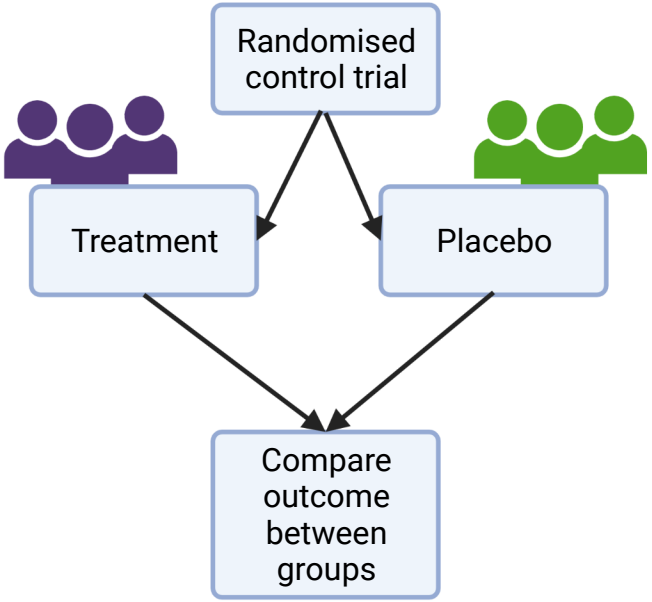
Step 2

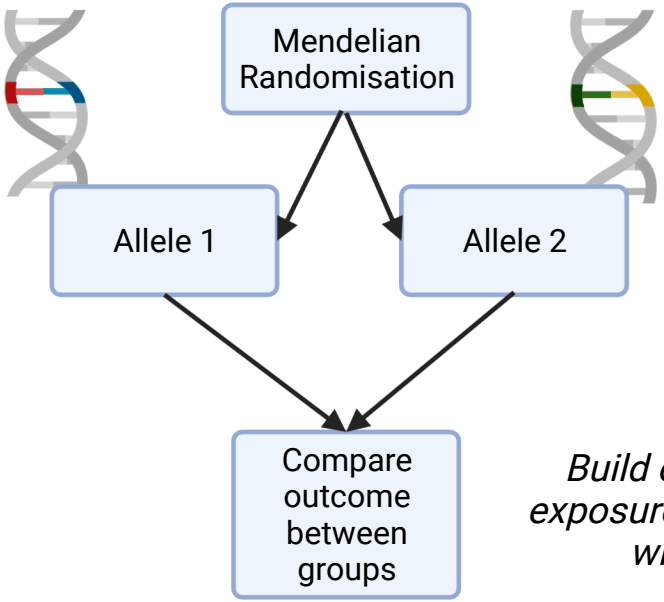
*Build evidence that this risk exposure is **causally associated** with kidney disease*



Genetic analysis of kidney disease

- Mendelian Randomisation (MR):





Step 1

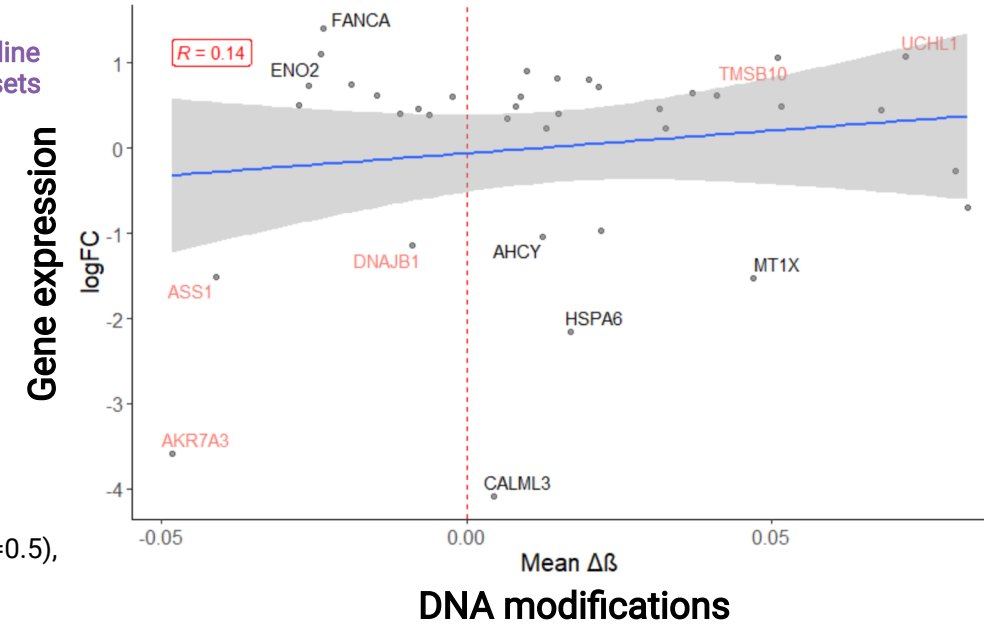
Genetic variants known to be associated with ***proposed risk exposure***

Step 2

Build evidence that this risk exposure is ***causally associated*** *with kidney disease*

Genomic analysis of kidney disease

```
ggplot(dataframe, aes(x=m, y=logFC)) +   #pull in your data
 geom_point(alpha = 0.4)+    #make points slightly transparent
 geom_vline(xintercept=0, linetype="dashed", color = "red")+  #add vertical line
 geom_text_repel(data=subset(dataframe, m < 0 & logFC < -1),   #label subsets
           aes(label=Gene, colour='red'))+
 geom_text_repel(data=subset(dataframe, m > 0 & logFC > 1),
           aes(label=Gene, colour='red'))+
 geom_text_repel(data=subset(dataframe, m > 0 & logFC < -1),
           aes(label=Gene))+
 geom_text_repel(data=subset(dataframe, m < 0 & logFC > 1),
           aes(label=Gene))+
 scale_y_continuous(breaks=seq(-4,2,1))+  #axis limits and increments
 theme(panel.grid.major = element_blank(),   #how graph looks
      panel.grid.minor = element_blank(),
      panel.background = element_blank(),
      axis.text.x = element_text(size=12),
      axis.text.y = element_text(size=12),
      legend.position = 'none',
      plot.title = element_text(color="black", size=12, face="bold.italic", hjust=0.5),
      axis.title.x = element_text(size = 15, colour = 'black'),
      axis.title.y = element_text(size = 15, colour = 'black'),
      axis.line = element_line(colour = "black") )+
 labs(x='Mean Δ β',y='logFC')+     #label axis
 geom_smooth(method='lm')+     #linear model
 stat_cor(aes(label = ..r.label..), color = "red", size=4, geom = "label")    #add correltion label
```



DNA modifications

Levin et al, 2020

Hill and Duffy et al., in preparation