Building a Human Infertility Genome Database

Presenting Author: Wieland, J
Conflicts of Interest: None.
E-poster Presentation: EP18.011

Authors: Wieland, J[1]., Buchan, S[1]., Sen Gupta, S[2]., Mantzouratou, A[1].
1 Department of Life and Environmental Sciences, Faculty of Science and Technology, Bournemouth University, Poole BH12 5BB, UK
2 Institute for Women's Health, 86-96 Chenies Mews, University College London, London WC1E 6HX, UK

1) Introduction

Infertility is a major problem affecting approximately 10-15% of the global population[1-2]. Currently, there are three (in)fertility associated databases[3-5]. One common theme amongst these databases is that none provide a patient focused approach for studying the genetics of infertility. In addition, all depend upon published research to deliver upon their aim. Herein we describe a novel database to investigate the genetic basis of infertility. The difference between this database (Fig 1) and current databases[3-5], is that a patient focussed approach is provided for studying genetics of infertility which is not dependent upon published research for core data structure.

The aim of this database (Fig 1) is to provide a novel patient focused approach for studying the genetics of infertility.

2) Methods: Database Construction

The database was constructed in 2 parts. The first part involved data processing and formatting using awk programming Fig 2(A-D). The second part required the use of MySQL to design and construct the database (Fig 1). This database contains data from 50 anonymised patients from 25 infertile couples. Data comprised annotated genomic datafiles. Additionally, these data are accompanied with embryonic data associated to each couple within the database.

(A)
```
awk 'BEGIN { FS ="\t"; OFS="\t"; }
  { print $1, $2, $3, $4, $5 }'
  file_name.txt>output_file.csv
```

(B)
```
awk 'BEGIN{ FS="\t"; OFS="\t"; }; { loc =
 $1 ":" $2 "-" $3; FS=";"; $0 = $3; }; {
   print loc } ; { FS="\t"; };'
   file_name.txt>output_file.csv
```

(C)
```
awk 'BEGIN { FS="\t"; OFS="\t"; }; $4
 {column_name = $1; column_name = $2;
 column_name = $3; FS=";"; $0 = $4; }; {
for (i=1; i<=NF; i++){ print column_name,
   column_name, column_name, $1 ; } }; {
FS="\t"; };' file_name.txt>output_file.csv
```

(D)
```
   cat file_name.txt | awk
  'BEGIN{FS="\t";OFS="\t"; re=",
[[:digit:]]+ \\([[:digit:]]\\)$";}; {
    sub(/[[:space:]]+$/,"",$3);
        if(match($5,re)) {
  $4=substr($5,RSTART+2,RLENGTH-6);
      $5=substr($5,1,RSTART-1)
substr($5,RSTART+RLENGTH-4);};print};' >
      output_file.csv
```

Figure 2: Database Construction workflow. (A) Awk programme used to extract the columns of interest. (B) Awk programme to generate the Coordinates column in the datasets. (C) Awk programme that splits each entry per row within a column at a delimiter and inserts into a new row. (D) Awk programme that is used in conjunction with (C) to generate Clinical_Associations datasets.

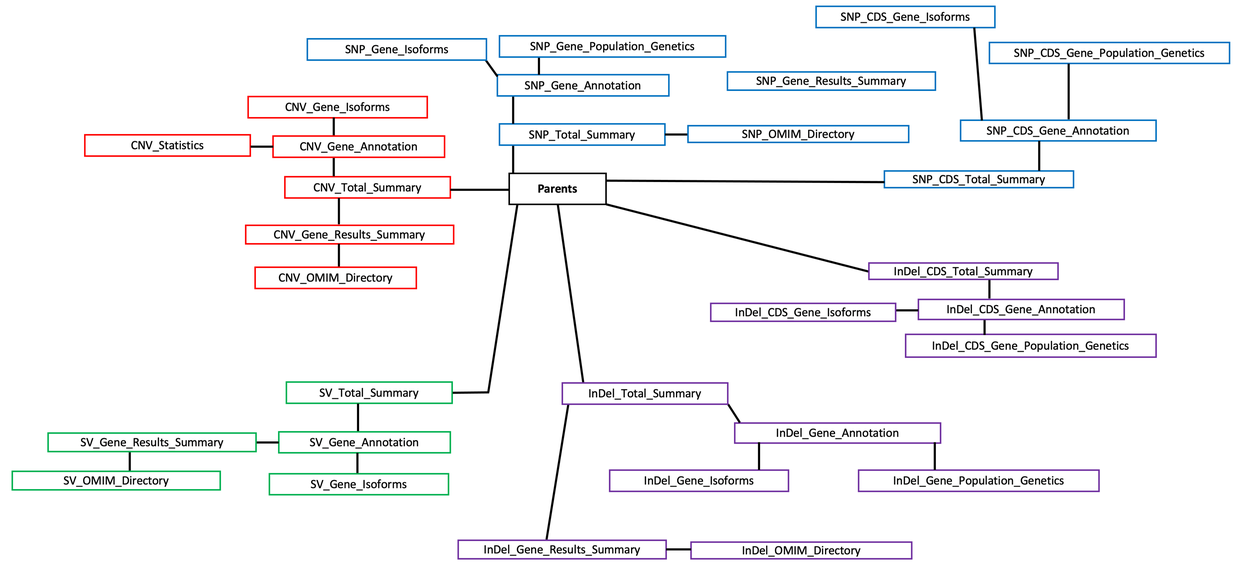

Figure 1 Database schematic model. This presents a simplified schematic Structural Query Language (SQL) database. The red outline boxes presents CN datasets. The Purple outlined boxes presents the InDels, including CDS regions. The green boxes presents the SV datasets. The blue outlined boxes presents the SNPs with CDS only regions.

3) Results

The database contains 1,611 dataset tables where each patient has their own table per dataset type; Copy Number Variation (CNV), Insertion/Deletions (InDels), including only coding regions (CDS), Structural Variant (SV) datasets, and Single Nucleotide Polymorphisms (SNPs) with CDS only regions. The patient cohort consists of patients with recurrent miscarriage (n=8), advanced maternal age (n=3), recurrent implantation failure (n=5), asthenospermia (n=1), azoospermia (n=2), poor sperm parameters (n=4), normal parameters but requiring intracytoplasmic sperm injection (ICSI) (n=9), oligoasthenoteratozoospermia (OAT) (n=1), patients from a miscarriage unit (n=6), and validation genomes (n=7).

4) Discussion

This is a novel Structured Query Language (SQL) database which aims to provide a novel patient focused approach for analysing the genetics of infertility. The next step is to analyse data within the database to identify genetic signatures associated with infertility. Additional future work will add accessory information from NCBI such as gene ontology (GO).

5) Conclusion

We have developed a novel database that aims to support future research into the genetics of infertility by providing a tailored approach from datasets of 50 anonymised infertile patients.

References

1. Sharlip, I. D., et al., 2002. Best practice policies for male infertility. *Fertility and Sterility*, 77 (5), 873-882.
2. Borght, M. V and Wyns, C., 2018. Fertility and infertility: Definition and epidemiology. *Clinical Biochemistry*, 62, 2-10.
3. Jasilioniene, A., et al., 2016. Data Resource Profile: The Human Fertility Database. *International Journal of Epidemiology* 45 (4),1077–1078.
4. Wu et al., 2021. IDDB: a comprehensive resource featuring genes, variants and characteristics associated with infertility. *Nucleic Acids Research*, 49 (1), 1218-1224.
5. Sinha, S., et al. 2021. Contraceptive and Infertility Target DataBase: a contraceptive drug development tool for targeting and analysis of human reproductive specific tissues. *Biology of Reproduction*, 105 (6), 1366-1374.