# Tecnologia de Processamento de Dados 2019/2020

# HuGS-DW

Human GWAS-SNP Cross-Reference Data Warehouse

## Course Project - Part 1


**Grupo 20**

| Nome | Student Number | Contribution (hours) |
|---|---|---|
| Elias Barreira | 40821 | 15 |
| José Matos | 49652 | 15 |
| Miguel Casanova | 24475 | 15 |
| Telmo Silva | 54013 | 15 |

**1. Describe your original data set. Identify the most important information. Identify missing or incomplete data, and identify possible strategies to use (or discard) them**

Neste projeto, propomos estabelecer um Data Warehouse (DW) em que vamos cruzar dados genómicos de duas fontes, o "1000 Genomes Project" e o "NHGRI-EBI Catalog of published genome-wide association studies", ou GWAS. Será destes projectos que vamos obter os datasets que vão formar o core do nosso DW. Os dados contidos nestas fontes são de uma enorme profundidade mas, sempre que possível, tentaremos enriquecê-los usando outras fontes.

O "1000 Genomes Project" foi um projecto de investigação internacional, com o objectivo de criar um dos catálogo mais detalhados de variação genética humana. Na fase final do projecto, 2504 indivíduos de diferentes grupos étnicos foram sequenciados para encontrar a maior parte das variantes genéticas com frequências de pelo menos 1% nas populações estudadas. O catálogo GWAS é um base de dados online onde estão compilados os dados de vários estudos de GWAS, resumindo os dados obtidos de várias fontes literárias. Assim, o GWAS constitui uma base de dados curada, consistente e pública das associações entre variantes genéticas comuns e doenças complexas.

Os dados do "1000 Genomes Project" contêm vários tipos de informação:

- Metadados acerca dos 2504 indivíduos, incluindo a sua etnia e sexo.
- Informação acerca de todos os SNPs identificados no estudo, incluindo seu ID, posição no cromossoma, o genótipo de referência e o alternativo e vários outros tipos de informações, como frequência em determinadas populações, tipo de SNP (mudança de base, deleção, inserção, "copy-number-variation" (CNV)).
- Informação, para cada um dos SNPs, dos genótipos encontrados em cada um dos indivíduos. Esta informação está organizada em ficheiros de formato Variant Call Format (VCF).

Os dados mais relevantes deste dataset, são os metadados dos vários indivíduos, nomeadamente o género e a população a que pertencem, os metadados dos SNPs (nomeadamente dados referentes à sua localização genómica) e, finalmente, os dados genómicos dos diferentes indivíduos, para cada um dos SNPs identificados.

Os dados GWAS contêm vários tipos de informação, para cada estudo de GWAS publicado. Os dados incluem:

- Dados acerca do estudo (publicação, tipo de amostra usada, tecnologias usadas).
- Dados acerca das variantes genómicas associadas com a doença complexa estudada (localização de SNPs, proximidade de genes, contexto genómico dos SNPs identificados (intergénico, intrónico, exónico, etc...).
- Vários dados estatísticos acerca da associação entre variante genética e doença.
- Links para várias ontologias relacionadas com a doença estudada.

Os dados mais relevantes deste dataset e que vamos manter no nosso DW, são dados básicos relacionados com o estudo, a doença estudada, a informação acerca dos SNPs identificados, incluindo a sua localização genómica, o contexto onde eles se encontram e as estatísticas da associação.

No que diz respeito a dados incompletos ou em falta, os datasets que obtemos são bastante sólidos e extensos. De qualquer modo, há a salientar os seguintes pontos:

- Nos dados genómicos, pode haver posições no genoma onde os resultados da sequenciação não são 100% fiáveis. Nestes casos, os dados têm a descrição que não passam um filtro de qualidade, pelo que será muito fácil descartá-los.
- Relativamente à informação acerca dos indivíduos do "1000 genomes", esta informação é anonimizada e portanto, os únicos dados que temos referem a população a que pertencem e o seu sexo.

A informação contida nos "1000 Genomes Project" é enorme, com informação para cerca de 84 milhões de SNPs para cada um de 2504 indivíduos. Por essa razão, a nossa estratégia inicial vai focar-se na análise do cromossoma 19. A razão pela qual optámos por este cromossoma, prende-se com o facto de ser o cromossoma com maior número de estudos de GWAS (normalizado por tamanho do cromossoma).

**2. Describe what other data sets (if any) are going to be used for the construction of the data warehouse. How will they complement the existing information**

Os dados genómicos que vamos utilizar, vão ser enriquecidos com datasets contendo informações de certas características do indivíduo, doenças e SNPs:

- Em relação ao indivíduo iremos usar dados sobre a sua população e o seu sexo, como esperança média de vida, aspectos socioeconómicos e culturais, entre outros.
- A população está definida como sendo oriunda duma cidade, que por sua vez está associada a uma etnia e pertence a um país, para os quais também é possível recolher dados.
- Em relação aos SNPs vamos recolher informação sobre o cromossoma em que está situado, nomeadamente, densidade génica, tamanho dos cromossomas, etc. Além disso, vamos adicionar à dimensão SNPs, informação importante acerca do contexto genómico (localização em exões, intrões, regiões reguladoras, intergénicos, etc.) e listar genes que se encontrem na proximidade.
- Em relação às doenças iremos recolher dados sobre a ontologia da mesma, recorrendo à informação da Human Disease Ontology (DOID). Esta informação vai permitir obter dados acerca do tipo de doença, que órgãos afecta e qual a taxa de mortalidade/esperança de vida associada com a doença, entre outros.

**3. Identify and characterize the facts table. Identify and characterize the grain of each element. If more than one fact table is present identify the grain for each of them.**

O nosso DW vai conter duas tabelas de factos: SNP-Person e Diseases-GWAS.

**Fact table - SNP Person, definição do grão:**

Nesta tabela cada linha corresponde à variante de um dado SNP para um dado indivíduo. Incluirá os valores para os dois alelos do indivíduo (nonadditive facts) e referências para tabelas de dimensões contendo informações sobre esse indivíduo, como população e superpopulação, e informações sobre o SNP.

**Fact table - Disease GWAS, definição do grão:**

Nesta tabela cada linha corresponde a uma referência num estudo que associa um SNP a uma doença (no mesmo estudo, podem existir várias linhas, correspondentes aos vários SNPs identificados no estudo como tendo uma associação com a doença).

**4. Identify all the possible dimensions with the data set and the accessory data found. Here we just need to identify the Dimensions, not define their structure.**
As dimensões que identificamos para o nosso dataset, são:

- **Tabela de factos SNP Person**
    - Person, Gender, Population, SNP e Chromosome

- **Tabela de factos Disease GWAS**
    - Disease, Study e SNP (partilhada com tabela de factos SNP Person)

**5. Identify how many data marts are going to be used and what types of business processes might use them. If more than one data mart is present, define the Bus matrix and identify their position in a feasibility/value matrix**
Os dados armazenados no DW que planeamos construir vão permitir construir vários data marts, de acordo com diferentes processos de negócio que podem querer usar e analisar os dados de perspectivas distintas.

Assim sendo, os data marts que podemos propor inicialmente, são:
- **Análise da variação genética das populações humanas**
- **Análise da associação entre doenças e variação genética**
- **Determinação da prevalência de doenças raras em diferentes populações humanas**

Os data marts e os "business processes" que os podem usar estão representados nas seguintes "bus matrix" e "feasibility/value matrix":

| Business Processes | Gender | Population | Person | SNP | Chromosome | Disease | Study |
|---|---|---|---|---|---|---|---|
| Variabilidade Genética | X | X | X | X | X | | |
| Variantes Genéticas em Doenças | | | | X | X | X | X |
| Doenças complexas em diferentes populações | | X | X | X | X | X | X |

**Tabela 1. Bus Matrix**

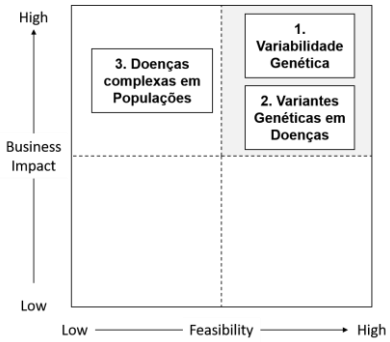


**Tabela 2. Feasibility/Value Matrix**

## 6. Define the star schema of the proposed data warehouse

Este é o “Star schema” do nosso DW:

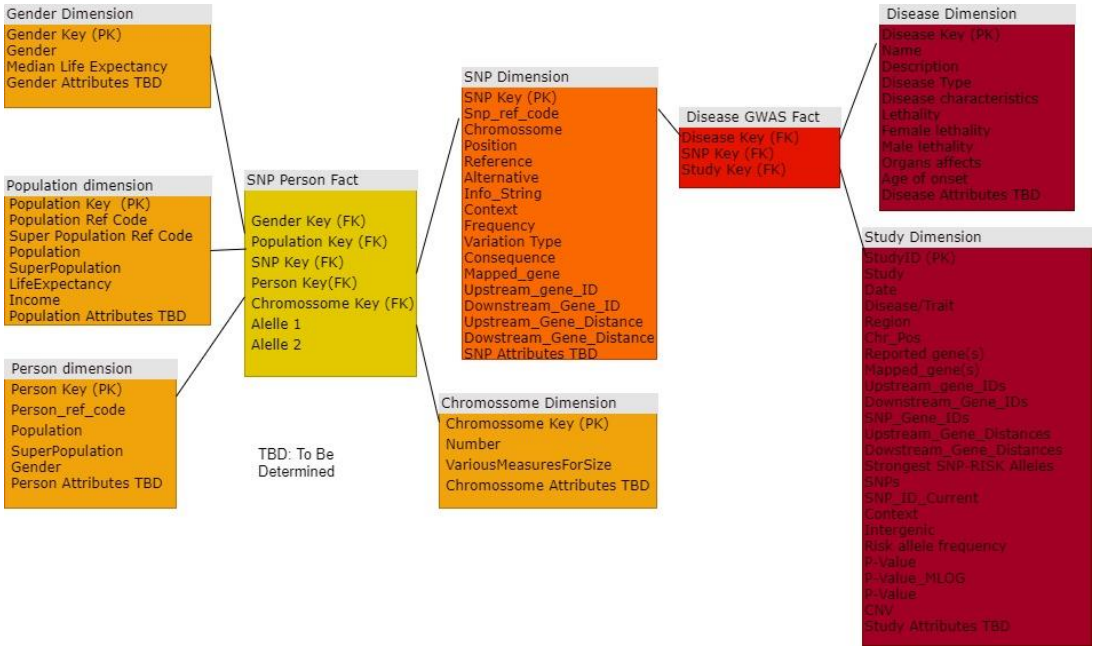


**Figura 1. Star schema do Data Warehouse**

### a. For each dimension identify and describe its columns

As dimensões do nosso DW vão ter a seguinte estrutura:

- **SNP dimension:** Esta dimensão vai ter informações acerca dos SNPs, nomeadamente, a referência, dados acerca da sua localização e contexto genómico, informações acerca da frequência em diferentes populações e consequências moleculares associadas.
- **Population dimension:** Nesta dimensão agrega-se informação acerca de populações de determinados países, cidades e etnias, incluindo aspectos socioeconómicos, como esperança média de vida, remuneração base média e outros aspectos a determinar como aspectos educacionais da população, por exemplo.
- **Person dimension:** Dimensão com dados sobre os indivíduos no “1000 Genomes Project”, como sexo e etnia.
- **Gender Dimension:** Dados sobre indivíduos de diferentes sexos como esperança média de vida por sexo e outros aspectos a determinar.
- **Chromosome Dimension:** Dimensão onde há informação sobre cromossomas como o seu número de identificação, tamanho e outras características a determinar.
- **Disease Dimension:** Dimensão onde se agrega a informação acerca de doenças humanas (da “Human Disease Ontology” - DOID). Esta informação inclui dados como letalidade, efeitos da doença e a sua categorização.
- **Study Dimension:** Esta dimensão envolve informação provenientes de estudos de GWAS. Aqui são agrupadas informações sobre o estudo, os SNPs identificados como associados à doença, assim como genes potencialmente envolvidos. Nesta dimensão, encontram-se também vários parâmetros estatísticos para a associação entre a doença e as variantes genéticas identificadas.

**b. Identify and characterize the measures of the facts table**

As medidas das nossas tabelas de factos são:

- **SNP Person Fact:** Esta tabela contém factos não-aditivos referentes aos dois alelos do SNP observados em cada indivíduo (alelo 1 e alelo 2).
- **Disease GWAS Fact:** Esta tabela não contém medidas, é uma "Factless fact table".

**c. Estimate its size and growth over time (in number of lines in tables)**

As dimensões estimadas do nosso DW vão ser:

- **Tabelas de factos:**
    - **SNP Person Fact:** Esta tabela de factos, que vai focar-se no cromossoma 19, vai ter informação acerca dos 1,832,506 SNPs localizados no cromossoma, para os 2504 indivíduos, ou seja, 4,588,595,024 linhas.
    - **Disease GWAS Fact:** Cada linha desta tabela vai conter cada um dos SNPs identificados nos cerca de 179,365 estudos. Para termos a dimensão final desta tabela, necessitamos de processar a tabela GWAS e extrair a sua informação.

- **Tabelas de dimensões:**
    - **Gender Dimension:** Contém dados relacionados com o sexo. Esta dimensão tem portanto, duas linhas.
    - **Population Dimension:** Esta tabela vai conter informação acerca de cada uma das populações definidas no "1000 Genomes Project", ou seja, 26 linhas.
    - **Person Dimension:** Contém informação acerca dos 2504 indivíduos estudados no "1000 Genomes Project". Cada linha corresponde a um indivíduo.
    - **SNP Dimension:** Cada linha corresponde a cada um dos SNPs identificados no "1000 Genomes Project", ou seja 848,018,801 SNPs. Como esta tabela de dimensões seria demasiado grande, vamos fazer um subsampling dos SNPs encontrados no cromossoma 19, ou seja, 1,832,506 SNPs.
    - **Chromosome Dimension:** Cada linha contém informação acerca de cada um dos 24 cromossomas humanos (22 autossomas e 2 cromossomas sexuais, X e Y). Apesar de só nos focarmos no cromossoma 19, iremos construir esta dimensão para assegurar que o nosso DW é "futureproof".
    - **Disease Dimension:** Esta tabela vai ter 12,292 linhas, correspondentes a cada uma das doenças descritas na DOID ontology.
    - **Study Dimension:** Nesta dimensão teremos os 179,365 estudos agregados no GWAS. Cada estudo corresponde a uma linha da tabela.

No que diz respeito ao crescimento do nosso DW, este vai-se concentrar em torno da tabela de factos Disease GWAS (uma vez que os dados dos "1000 Genomes Project" estão publicados e não vão ser actualizados). Assim sendo, antecipamos o crescimento das seguintes tabelas:

- **Disease GWAS Fact:** À medida que novos estudos de associação forem publicados e introduzidos na base de dados do GWAS, novas linhas serão adicionadas por cada um dos SNPs identificados no estudo.
- **SNP Dimension:** Sempre que forem detectados novos SNPs, estes serão adicionados a esta tabela.
- **Study Dimension:** Sempre que um novo estudo for publicado, uma nova linha será adicionada a esta tabela.

- **Disease Dimension:** Sempre que uma actualização da DOID ocorrer, esta tabela será actualizada também.

Devido à dimensão total do DW que propomos, talvez seja necessário fazer um sampling dos dados. Se isto se confirmar, iremos escolher aleatoriamente 1% dos SNPs do cromossoma 19, mantendo o número original de indivíduos.

**7. Identify the main usage of the data system and propose different analysis queries for the system. It's just necessary to write them down in plain language**

O sistema de dados que propomos construir, vai estabelecer uma ferramenta útil que une dados de variabilidade genética de diferentes populações humanas, com dados acerca da associação de doenças a variantes genéticas humanas. O uso principal deste DW, vai orbitar em torno de questões relacionadas com filogenias e variabilidade genética humana. Nomeadamente, vai permitir analisar variantes genéticas distribuídas de forma distinta entre diferentes populações humanas e a sua predisposição para certas doenças.

A maneira como vamos organizar os nossos dados e, principalmente, as camadas de informação extra que as diferentes dimensões vão ter, vai permitir elaborar questões muito interessantes. Por exemplo:

- Que populações têm uma maior frequência de variantes genéticas em genes vs regiões intergénicas e/ou reguladoras?
- Existe uma relação entre variabilidade genética e aspectos socioeconómicos e culturais das populações? (maior taxas de mutações, maior isolamento populacional?)
- Em que populações uma determinada doença complexa tem maior probabilidade de ocorrer?
- As variantes genéticas identificadas estão em que contexto genómico? Existe uma maior associação de doenças com genes ou com regiões não codificantes?
- Em populações com igual frequência de alelos de risco para uma determinada doença, qual o impacto que aspectos socioeconómicos têm sobre a incidência da doença?

Concluindo, o nosso DW vai ser útil a vários tipos de "processos de negócio", criando um sistema de dados ricos e com potencial para crescer as suas dimensões, aumentando assim o número de questões/processos ao qual o DW poderá responder.

**Recursos:**

**IGSR: The International Genome Sample Resource -** https://www.internationalgenome.org/
**GWAS Catalog -** https://www.ebi.ac.uk/gwas/home
**Disease Ontology -** https://www.ebi.ac.uk/ols/ontologies/doid
**dbSNP NCBI -** https://www.ncbi.nlm.nih.gov/snp/
**dbSNP ensembl -** https://www.ensembl.org/info/genome/variation/index.html

**Bibliografia:**

Ralph Kimball e Margy Ross, The data warehouse toolkit: The complete guide to dimensional modeling, Wiley, 3a edição, 2013 - Capítulos 1, 2, 16