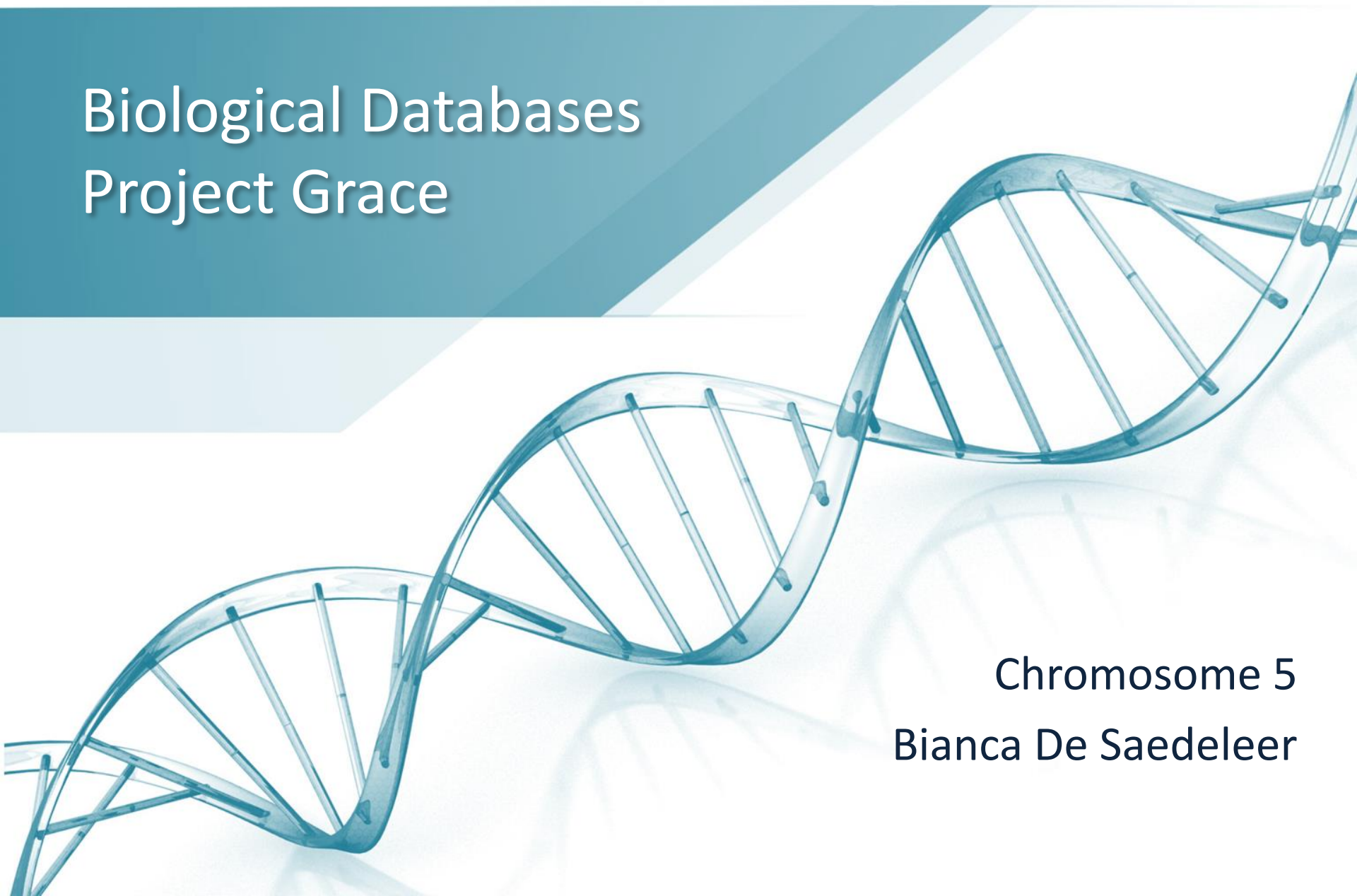
Biological Databases
Project Grace



Chromosome 5

Bianca De Saedeleer

Chromosome 5

- General overview
- Strategy, building a DB
- Allergies and diseases
- Further ideas

General overview

- 5th largest human chromosome, but low coding gene density
- Responsible for growth and development (cell division), associated with allergies, asthma and immune deficiency disorders

Length (bps)	181,538,259
Coding genes	887 (incl. 21 readthrough)
Non coding genes	1,196
Small non coding genes	219
Long non coding genes	858 (incl. 6 readthrough)
Misc non coding genes	119
Pseudogenes	734
Short Variants	41,202,467

Strategy

- Retrieved all genes and all coding genes with **pybiomart**

index	Gene_ID	Gene_Name	Gene_des	Start_pos	End_pos	Transcript_type
0	ENSG00000286094	AC026740.3	zinc finger DHHC-type containing 11B [Source:NCBI ...	716808	766919	protein_coding
1	ENSG00000050767	COL23A1	collagen type XXIII alpha 1 chain [Source:HGNC Sym...	178237618	178590393	protein_coding
2	ENSG00000113240	CLK4	CDC like kinase 4 [Source:HGNC Symbol;Acc:HGNC:136...	178602664	178630615	protein_coding
3	ENSG00000169131	ZNF354A	zinc finger protein 354A [Source:HGNC Symbol;Acc:H...	178711512	178730659	protein_coding
4	ENSG00000285891	AC113348.1	uncharacterized LOC102724657 [Source:NCBI gene;Acc...	178694605	178697695	protein_coding
5	ENSG00000178338	ZNF354B	zinc finger protein 354B [Source:HGNC Symbol;Acc:H...	178859953	178888122	protein_coding
6	ENSG00000198939	ZFP2	ZFP2 zinc finger protein [Source:HGNC Symbol;Acc:H...	178895898	178933212	protein_coding
7	ENSG00000178187	ZNF454	zinc finger protein 454 [Source:HGNC Symbol;Acc:HG...	178941191	178966433	protein_coding
8	ENSG00000113262	GRM6	glutamate metabotropic receptor 6 [Source:HGNC Sym...	178977587	178996206	protein_coding
9	ENSG00000234284	ZNF879	zinc finger protein 879 [Source:HGNC Symbol;Acc:HG...	179023752	179035064	protein_coding
10	ENSG00000177932	ZNF354C	zinc finger protein 354C [Source:HGNC Symbol;Acc:H...	179060415	179083537	protein_coding
11	ENSG00000087116	ADAMTS2	ADAM metallopeptidase with thrombospondin type 1 m...	179110853	179345461	protein_coding
12	ENSG00000176783	RUFY1	RUN and FYVE domain containing 1 [Source:HGNC Symb...	179550554	179610012	protein_coding
13	ENSG00000185261	KIAA0825	KIAA0825 [Source:HGNC Symbol;Acc:HGNC:28532]	94152966	94618597	protein_coding
14	ENSG00000173930	SLCO4C1	solute carrier organic anion transporter family me...	102233986	102296284	protein_coding
15	ENSG00000186335	SLC36A2	solute carrier family 36 member 2 [Source:HGNC Sym...	151314972	151347590	protein_coding
16	ENSG00000175745	NR2F1	nuclear receptor subfamily 2 group F member 1 [Sou...	93583222	93594611	protein_coding
17	ENSG00000186493	C5orf38	chromosome 5 open reading frame 38 [Source:HGNC Sy...	2752131	2755397	protein_coding
18	ENSG00000239389	PCDHA13	protocadherin alpha 13 [Source:HGNC Symbol;Acc:HGN...	140882124	141012347	protein_coding

Strategy

- 56 048 SNPs given
- **Myvariant**
  - **vcf**: reference and alternative allele
    → For which RS IDs has Grace the alternative allele

    26877 SNPs left

```
"SELECT chr5.RS_ID, chr5.GENOTYPE, typing.REF, typing.ALT \
   FROM chr5 INNER JOIN typing WHERE chr5.rs_id = typing.rs_id"
```

  - **clinvar**: associated traits and genes based on the RS IDs
    → lot of RS IDs not included in clinvar

    354 SNPs left

```
info = mv.getvariant(rsid, fields='clinvar')
```

Strategy

- Clinvar


```
{'accession': 'RCV000211244',
 'clinical_significance': 'drug response',
 'conditions': {'identifiers': {'medgen': 'CN236607'},
  'name': 'methotrexate response - Toxicity/ADR'},
 'last_evaluated': '2018-01-30',
 'number_submitters': 1,
 'origin': 'germline',
 'preferred_name': 'NM_002454.2(MTRR):c.66A>G (p.Ile22Met)',
 'review_status': 'reviewed by expert panel'},
```


SNP_ID	RS_ID	ACCESSION	SIGNIF	NAME
1	rs2075786	RCV000648948	Benign	Idiopathic fibrosing alveolitis, chronic form (IPF...
2	rs2075786	RCV000648948	Benign	Dyskeratosis congenita, autosomal dominant, 2 (DKC...
3	rs2736100	RCV000497563	association	Chronic osteomyelitis
4	rs2735940	RCV000498732	association	Chronic osteomyelitis
5	rs6347	RCV000180550	Benign	not specified
6	rs6347	RCV000625455	Benign	Infantile Parkinsonism-dystonia (PKDYS1)
7	rs13181449	RCV000117855	Likely benign	not specified
8	rs13181449	RCV000333710	Benign	Intellectual Disability, Recessive
9	rs13181449	RCV000715386	Benign	History of neurodevelopmental disorder
10	rs1801394	RCV000007444	risk factor	Neural tube defects, folate-sensitive, susceptibil...
11	rs1801394	RCV000007445	risk factor	Down syndrome, susceptibility to
12	rs1801394	RCV000126873	Benign	not specified
13	rs1801394	RCV000144926	Uncertain significance	Gastrointestinal stroma tumor (GIST)
14	rs1801394	RCV000211244	drug response	methotrexate response - Toxicity/ADR
15	rs1801394	RCV000264714	Benign	Disorders of Intracellular Cobalamin Metabolism

Strategy

- **PheGeni NCBI**: associated traits, genes, ethnic background

#	Trait	rs #	Context	Gene	Location	P-value	Source	Study	Population	PubMed
51	Inflammatory Bowel Diseases	rs2188962	intron	C5orf56	5: 132,435,113	1.000×10^{-52}	NHGRI		European	23128233
52	Recombination, Genetic	rs6889665	intergenic	PRDM9, CDH10	5: 23,532,534	2.000×10^{-52}	NHGRI		African African American	21775986
53	Metabolism	rs272889	intron	LOC553103	5: 132,329,685	3.000×10^{-51}	NHGRI		European	24816252
54	Metabolism	rs272889	intron	SLC22A4	5: 132,329,685	3.000×10^{-51}	NHGRI		European	24816252
55	Recombination, Genetic	rs2914276	intergenic	PRDM9, CDH10	5: 23,542,602	1.000×10^{-50}	NHGRI		European	20981099
56	Inflammatory Bowel Diseases	rs56167332	intergenic	RNU4ATAC2P, ADRA1B	5: 159,400,761	7.000×10^{-50}	NHGRI		European	26192919
57	Fibrinogen	rs2106854	intron	C5orf56	5: 132,433,482	2.000×10^{-48}	NHGRI		European	23969696
58	Eyebrows	rs16891982	missense	SLC45A2	5: 33,951,588	7.000×10^{-48}	NHGRI		European	27182965
59	Cholesterol	rs12916	UTR-3	HMGCR	5: 75,360,714	9.000×10^{-47}	NHGRI		European	20686565
60	Lupus Erythematosus, Systemic	rs10036748	intron	TNIP1	5: 151,078,585	1.000×10^{-45}	NHGRI		European	26502338
61	Cholesterol, LDL	rs3843482	intron	HMGCR	5: 75,343,434	2.000×10^{-45}	NHGRI		European	25961943
62	Cholesterol, LDL	rs12916	UTR-3	HMGCR	5: 75,360,714	5.000×10^{-45}	NHGRI		European	20686565
63	Crohn Disease	rs11741861	intron	ZNF300	5: 150,898,347	6.000×10^{-44}	NHGRI		European	26192919
64	Cholesterol	rs10038095	intron	HMGCR	5: 75,341,886	3.000×10^{-43}	NHGRI		European	25961943
65	Inflammatory Bowel Diseases	rs56399423	intron	LOC553103	5: 132,336,964	4.000×10^{-43}	NHGRI		European	26192919
66	Inflammatory Bowel Diseases	rs56399423	intron	SLC22A4	5: 132,336,964	4.000×10^{-43}	NHGRI		European	26192919
67	Inflammatory Bowel Diseases	rs17622378	intron	C5orf56	5: 132,442,760	1.000×10^{-42}	NHGRI		European	26192919
68	Inflammatory Bowel Diseases	rs6871626	intergenic	RNU4ATAC2P, ADRA1B	5: 159,399,784	1.000×10^{-42}	NHGRI		European	23128233
69	Psoriasis	rs7709212	intron	LOC285626	5: 159,337,169	2.000×10^{-42}	NHGRI		European	26626624
70	Crohn Disease	rs56167332	intergenic	RNU4ATAC2P, ADRA1B	5: 159,400,761	2.000×10^{-41}	NHGRI		European	26192919
71	Metabolism	rs11950562	intron	SLC22A4	5: 132,316,836	2.000×10^{-41}	NHGRI		European	24816252
72	Metabolism	rs11950562	intron	LOC553103	5: 132,316,836	2.000×10^{-41}	NHGRI		European	24816252
73	Cholesterol	rs6882076	nearGene-5	TIMD4	5: 156,963,286	5.000×10^{-41}	NHGRI		European	24097068
74	Alcohol Drinking	rs10515739	intron	SGCD	5: 156,590,621	7.151×10^{-41}	dbGaP	phs000342	European	
75	Fibrinogen	rs4705952	intergenic	IRF1, IL5	5: 132,503,926	9.369×10^{-41}	dbGaP	phs000930	European	26561523
76	dimethylarginine	rs37369	missense	AGXT2	5: 35,037,010	1.000×10^{-40}	NHGRI		European	24159190
77	Adenocarcinoma of lung	rs2853677	intron	TERT	5: 1,287,079	3.000×10^{-40}	NHGRI		East Asian European	22797724

Strategy

```
Frequency European : 0.8660170523751523
Frequency African American : 0.07917174177831912
Frequency Asian : 0.022330491270807957
```

- INNER JOIN the SNPs with alternative allele with information retrieved from PheGeni
- INNER JOIN outcome clinvar with PheGeni (85 entries)

rs20541	IL13	Asthma, susceptibility to	Immunoglobulin E	missense	risk factor
rs20541	IL13	Asthma, susceptibility to	Psoriasis	missense	risk factor
rs2736100	TERT	Chronic osteomyelitis	Lung Diseases. Interstitial	intron	association
rs2736100	TERT	Chronic osteomyelitis	Idiopathic Pulmonary Fibrosis	intron	association
rs2736100	TERT	Chronic osteomyelitis	Glioma	intron	association

Strategy

- **Codegen** (65 SNPS left)

```
"CREATE TABLE codegen_linked SELECT * FROM alternatives \
  INNER JOIN codegen WHERE codegen.rsid = alternatives.rs_id \
  ORDER BY score DESC "
```

SNP_ID	RS_ID	ALT	record_id	rsid	score_class	score	freq	descr
1852	rs13166360	T	5	rs13166360	info	3.02	Not given	Heterozygous at ADCY2 gene. Associated with bipola...
26178	rs11746443	A	2	rs11746443	bad	3.02	29.18%	1.19x risk of kidney stone in Japanese
25425	rs3095870	T	1	rs3095870	info	3.02	37.02%	1.7x increased risk for SLE (lupus)
23094	rs2910164	G	4	rs2910164	info	3.02	Not given	higher/earlier cancer likelihood??
21515	rs13153971	C	6	rs13153971	warning	2.93	12.36%	1.58x higher risk of Asthma.
22899	rs3212227	G	7	rs3212227	warning	2.79	46.03%	Significantly increased risk of developing cervica...
19710	rs13357391	C	10	rs13357391	info	2.5	4.44%(rare)	slightly earlier age at menarche
172	rs31489	A	11	rs31489	info	2.47	10.23%	Reduced lung cancer risk?
22912	rs7709212	C	12	rs7709212	good	2.39	47.80%	0.82x risk of Psioriasis in Chinese
7172	rs1992662	G	13	rs1992662	info	2.31	16.12%	increased risk for Crohn's disease
12999	rs2544677	C	17	rs2544677	warning	2.25	82.36%	1.7x risk of type-1 diabetes
6418	rs3194051	G	14	rs3194051	warning	2.25	34.41%	1.12x risk of type-1 diabetes
26186	rs1801020	G	16	rs1801020	good	2.25	27.83%	Normal risk of developing heart disease
160	rs2736100	A	18	rs2736100	warning	2.23	26.56%	higher risk of Interstitial lung disease, and test...
6077	rs16891982	G	25	rs16891982	info	1.89	7.56%	Generally European; Light skin; Possibly an increa...
22911	rs2546890	G	26	rs2546890	good	1.89	31.33%	Lower risk of multiple sclerosis
21214	rs1042713	A	28	rs1042713	info	1.85	49.88%	1.3x increased risk that pediatric inhaler use may...
21593	rs1000113	T	29	rs1000113	info	1.76	33.42%	1.5x risk of Crohn's disease
14026	rs2549782	T	31	rs2549782	info	1.67	49.44%	1.3x increased risk for preeclampsia in most popul...
21585	rs13361189	C	32	rs13361189	info	1.58	42.26%	1.3x increased risk for Crohn's disease
14423	rs383830	T	33	rs383830	info	1.53	60.89%	1.9x risk1.9x higher risk for coronary artery dise...
190	rs464049	G	34	rs464049	info	1.5	47.66%	increased risk of schizophrenia in limited study

Strategy

- Data imputation with **Kaviar**
- Search for corresponding RD IDs based on position (using myvariant **dbSNP**)
- Also tried querying **UCSC**, but did not succeed


```
#In Ubuntu cmd
##Make sql statements from the imputed data (chrom_positions.txt) retrieved by Gaetans guideline
$ cd /mnt/c/Users/bianc/OneDrive/Documenten/Unief/AJ2018-2019/Semester2/BiologicalDatabases/Grace/Imputation_updated
$ awk 'BEGIN {FS=":"; OFS=""} {print "SELECT rsId, valid, base1, base2 from dbSnpRSHg WHERE chrom = \"chr",$1,"\" and chromStart =
",$2,";"}' chrom_positions.txt > snps.sql

#In Windows command line
##Connect with mysql
$ cd C:\path\to\MAMP\bin\mysql\bin
$ mysql --host=localhost -u root -p

##Query UCSC (this gives a syntax error)
$ mysql --user=genome --host=genome-euro-mysql.cse.ucsc.edu -A -D hg19 -N -e "your statement";
$ mysql -h genome-mysql.cse.ucsc.edu -u genome -A -D hg19 --skip-column-names < snps.sql;
```

Strategy

- As with the given data, check whether Grace has the alternative allele in certain SNP
- Link those RS IDs with associated traits using Clinvar and PheGeni

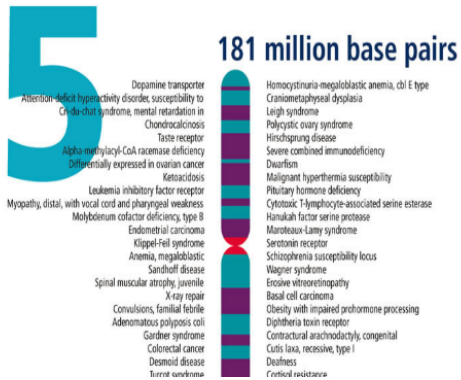
```
record = 0
total = imputation.shape[0]
for imp in imputation.iloc[:,1]:
    record += 1
    print('Record {} of {}.\n'.format(record,total))
    q = mv.query("chr5:{}-{}".format(str(imp),str(imp)))
    hits = q['hits']

    for h in range(len(hits)):
        print(h)
        pos = hits[h]['vcf']['position']

        if 'clinvar' in hits[h].keys():
            if hits[h]['clinvar']['type'] == 'single nucleotide variant':
                gene = hits[h]['clinvar']['gene']['symbol']
                rsid = hits[h]['clinvar']['rsid']
                alt = hits[h]['clinvar']['alt']
                ref = hits[h]['clinvar']['ref']

                if type(hits[h]['clinvar']['rcv']) is list: #if it's a list, you have multiple acce
                    accessions = len(hits[h]['clinvar']['rcv'])
                    for j in range(accessions):
                        acc = hits[h]['clinvar']['rcv'][j]['accession']
                        significance = hits[h]['clinvar']['rcv'][j]['clinical_significance']
                        name = hits[h]['clinvar']['rcv'][j]['conditions']['name']
                        sql = "INSERT INTO grace.imputed_clinvar \
                            (RS_ID,POSITION,REF,ALT,ACCESSION,GENE,SIGNIF,NAME) \
                            values"+ " ('"+rsid+"','"+pos+"','"+ref+"',\
                                '"+alt+"','"+acc+"','"+gene+"','"+significance+"', '"+name+"')"
                        cursor.execute(sql)
                        db.commit()

                elif type(hits[h]['clinvar']['rcv']) is dict: #only 1 accession for that rsid
                    acc = hits[h]['clinvar']['rcv']['accession']
```

Front-end

- Show results in php web interface
- SNP data of certain genes from NCBI – dbSNP

Welcome to Grace's 5th chromosome

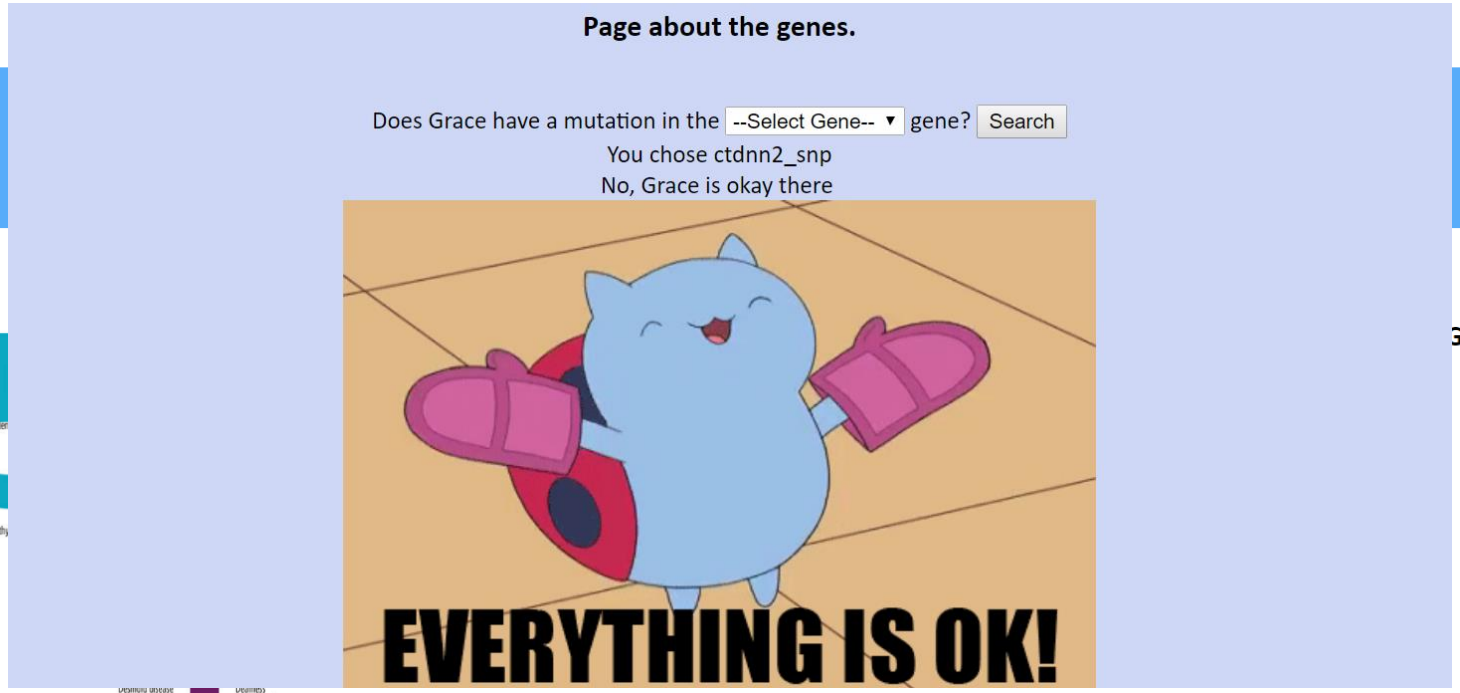
Genes Allergies Diseases



This page will give you more information about the 5th chromosome of Grace. Have fun!

Front-end

- Show results in php web interface
- SNP data of certain genes from NCBI – dbSNP



Asthma

- rs20541 – IL13
- rs13153971 – SLC6A7
- “Asthma gene”: multiple candidate genes (> 100) on chromosomes 5, 6, 11, 14, and 12 with chr 5 most ‘suspicious’

Allergic rhinitis

- rs20541 – IL13

Obesity, Diabetes

- rs6235 – PCSK1, LOC10192971
- rs3194051 – IL7R
- rs2544677 – not annotated

Mental health

rs16875288	Schizophrenia	ADAMTS16
rs16875288	Psychiatric Status Rating Scales	ADAMTS16
rs16875288	Depressive Disorder	ADAMTS16
rs16875288	Bipolar Disorder	ADAMTS16
rs2288447	Dextroamphetamine	NSUN2
rs472402	Drug Resistance	SRD5A1
rs531241	Dextroamphetamine	SRD5A1
rs566202	Dextroamphetamine	SRD5A1
rs248797	Dextroamphetamine	SRD5A1
rs17826816	Bipolar Disorder	ADCY2
rs26377	Dextroamphetamine	ZNF622
rs33692	Dextroamphetamine	FAM134B
rs34973	Personality	RASGRF2
rs6453541	Dextroamphetamine	RASGRF2
rs253959	Schizophrenia	COMMD10
rs253959	Bipolar Disorder	COMMD10
rs111294930	Schizophrenia	LINC01470
rs111294930	Schizophrenia	LINC01470
rs17113771	Emotions	LINC01470
rs4262150	Schizophrenia	LINC01470
rs4262150	Bipolar Disorder	LINC01470

Glucocorticoid Resistance syndrome

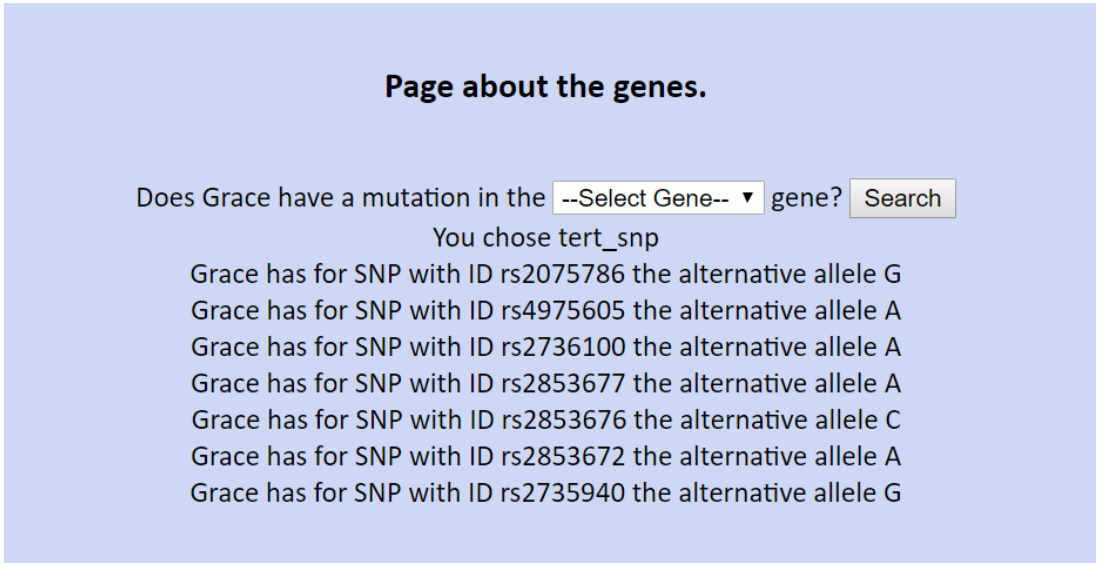
- rs6198 – NR3C1
  rs6191

- Increased levels corticotropin and cortisol
  → hypertension, hypokalaemic alkalosis (i.e. low level of potassium in blood serum), fatigue

- Elevated levels of androgens
  → Hyperandrogenism in females: hirsutism (i.e. excessive body hair), male pattern of baldness and menstrual irregularities

Chronic osteomyelitis

- rs2736100 – TERT
  rs2735940



- Chronic Bone Infection
- Associated with diabetes → found in the DB
- Males more affected → Grace might have hyperandrogenism (link?)

Further ideas

- Look into Graces drug responses

rs1801394	RCV000211244	drug response	methotrexate response - Toxicity/ADR
rs1042713	RCV000211205	drug response	salbutamol response - Efficacy
rs1042713	RCV000211334	drug response	salmeterol response - Efficacy
rs6295	RCV000417153	drug response	paroxetine response - Efficacy

- Use Gaetans word classifier to search for certain diseases in the tables of the database
- SPARQL
- Chromosome passport
- HTML – nicer interface