Synthetic datasets from RD-Connect GPAP

Last update: 11-05-2021
Contact: platform@rd-connect.eu
Website: https://platform.rd-connect.eu/

INDEX

1) Summary

This synthetic dataset includes **6 cases** (5 from rare diseases + 1 inherited cancer).

Each case corresponds to a trio of WGS (index case, mother and father) and was built using the same three public genomic datasets from HapMap samples NA12877, NA12878 and NA12882 (https://www.genome.gov/10001688/international-hapmap-project). In each of these cases, we have spiked-in real causative variants which correlate with the phenotypic data provided for each case.

For each of the 6 cases/families (6* WGS trio = 18 individuals) you will find:

1) phenotypic information = clinical description (phenopackets) and pedigree (.ped file)
2) genomic data = FASTQs, BAMs and VCFs

2) Type of cases included:

The six cases included in this synthetic dataset correspond to the following type of disorders:

a) Rare disease cases:
CASE 1- Congenital myasthenic syndrome (Autosomal Dominant -de novo)
CASE 2- Macular dystrophy (Autosomal Dominant)

CASE 3- Muscular dystrophy (Autosomal Recessive-compound heterozygous)
CASE 4- Mitochondrial disorder (Autosomal Recessive-homozygous-consanguineous)

b) Inherited cancer case:
CASE 5- Breast cancer (Autosomal Dominant)

c) Case for matchmaking purposes:
CASE 6- Same as case 1: Congenital myasthenic syndrome (Autosomal Dominant-de novo) – duplicated for matchmaking purposes. Same causative variant as case 1, phenotype and date of birth slightly different.

3) Files included per each case/family:

The following files are included for each of the individuals from each of the six cases. Note that all cases are submitted as trios = 3 individuals per family.

1. **Phenopacket:** one per individual
2. **Ped file:** one per case/family
3. **FASTQ:** 1 pair per individual = 2 files
4. **BAM:** 1 per individual + corresponding index = 2 files
5. **gVCF:** 25 per individual, one of them containing the inserted variant

Note that in each case directory you will find 3 subdirectories corresponding to each family member (see files structure):

- case1_index = index case from case 1
- case1_mother = mother from case 1
- case1_father = father from case 1

4) Technical information:

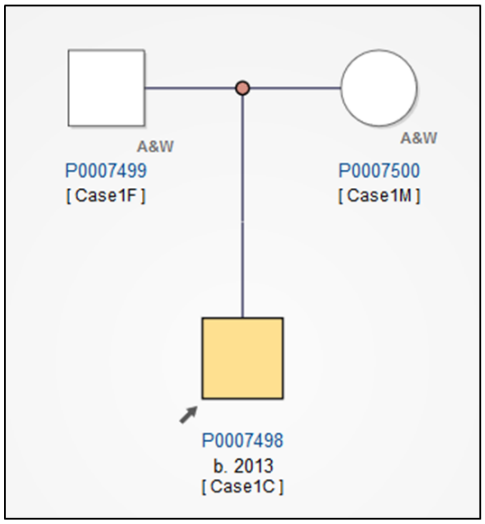
1) Genomic data
Information from original genomic data can be found here:
https://emea.illumina.com/platinumgenomes.html
Sequencing platform: HiSeq 2000
Sequencing protocol: WGS PCR-Free, 2x100bp
Coverage: 50x
Reference genome: GRCh37
Alignment: BWA-MEM
Variant calling: GATK 3.6

2) Phenotypic data
Phenopackets version: v1 (http://phenopackets.org/)

4) Detailed cases description:

CASE 1

Case1 description	
Gender	Male
Age	5 years
Referral	Congenital myasthenic syndrome
Onset	Congenital
Global pace of progression	Progressive (slow)
Main clinical features	• Neonatal hypotonia • Distal arthrogryposis • Inability to walk • Recurrent lower respiratory tract infections



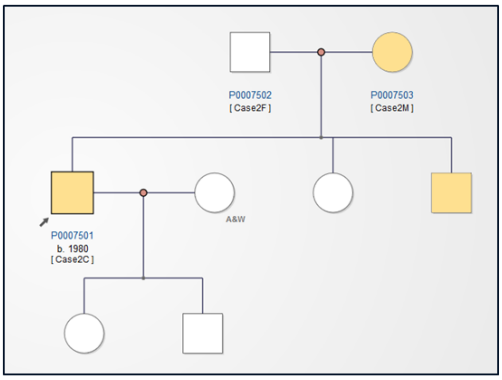
Information on the corresponding causative variant:

Disease: Central core disease
Gene: RYR1
Chr Coordinate: 19:39062815G>C
Variant: c.13888G>C p.Glu4630Gln (missense)
See this case in RD-Connect GPAP: https://playground.rd-connect.eu/genomics/?filters=56

FAM0001814	Index Case (male)	Father	Mother
Individual ID	case1C	case1F	Case1M
Phenopacket ID	P0007498	P0007499	P0007500
Genotype	0/1	0/0	0/0
Clinical status	Affected	Healthy	Healthy

CASE 2

Case 2 description	
Gender	Male
Age	38 years
Referral	Macular dystrophy
Onset	Adult
Global pace of progression	Progressive
Main clinical features	• Progressive visual loss • Scotoma • Abnormality of retinal pigmentation



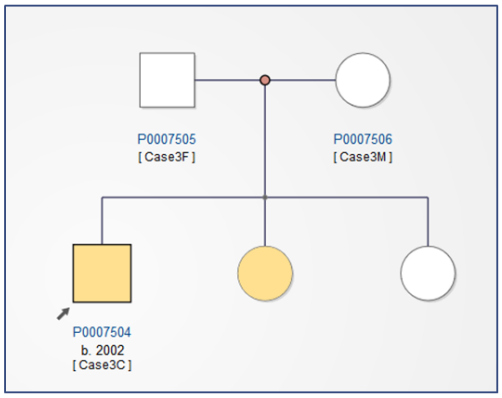
Information on the corresponding causative variant:

Disease: Macular dystrophy
Gene: PROM1
Chr Coordinate: 4:16037357C>T
Variant: c.303+1G>A (splicing)
See this case in RD-Connect GPAP https://playground.rd-connect.eu/genomics/?filters=57

FAM0001815	Index Case (male)	father	mother
Individual ID	Case2C	Case2F	Case2M
Phenopacket ID	P0007501	P0007502	P0007503
Genotype	0/1	0/0	0/1
Clinical status	Affected	Healthy	Affected

CASE 3

Case 3 description	
Gender	Male
Age	16 years
Referral	Muscular dystrophy
Onset	Juvenile
Global pace of progression	Progressive
Main clinical features	• Muscle weakness • Dystrophic muscle biopsy • Quadriceps muscle atrophy • Myalgia



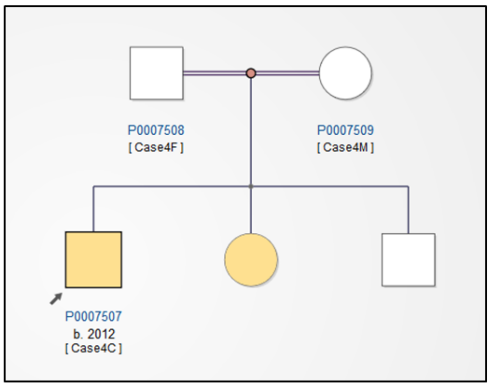
Information on the corresponding causative variants:

Disease: MUSCULAR DYSTROPHY, LIMB-GIRDLE, TYPE 2L; LGMD2L
Gene: ANO5
Chr Coordinates: 11:22257752G>T (mother allele) + 11:22242646C>CA (father allele)
Variants: c.692G>T p.Gly231Val (mother allele) + c.191dupA p.Asn64LysfsTer15 (father allele)
See this case in RD-Connect GPAP: https://playground.rd-connect.eu/genomics/?filters=59

FAM0001816	Index Case (male)	Father	Mother
Individual ID	Case3C	Case3F	Case3M
Phenopacket ID	P0007504	P0007505	P0007506
Genotype	**0/1; 0/1**	**0/0; 0/1**	**0/1; 0/0**
Clinical status	**Affected**	**Healthy**	**Healthy**

CASE 4

Case 4 description	
Gender	Male
Age	6 years
Referral	Mitochondrial disorder
Onset	Infantile
Global pace of progression	Progressive
Main clinical features	• Microcephaly • Epileptic encephalopathy • Intellectual disability • Motor deterioration



Consanguinity

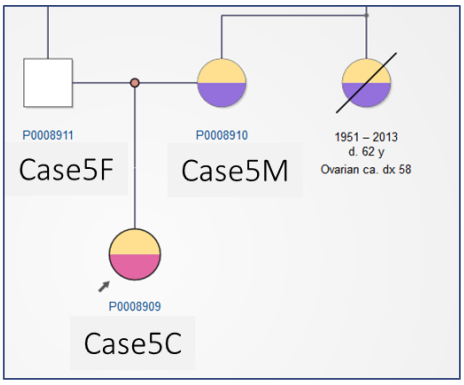
Information on the corresponding causative variant:

Disease: MITOCHONDRIAL DNA DEPLETION SYNDROME
Gene: POLG
Chr Coordinate: 15:89861960A>C
Variant: c.3294T>G p.Asn1098Lys (missense)
See this case in RD-Connect GPAP: https://playground.rd-connect.eu/genomics/?filters=60

FAM0001817	Index Case (male)	Father	Mother
Individual ID	Case4C	Case4F	Case4M
PhenoTips ID	P0007507	P0007508	P0007509
Genotype	1/1	0/1	0/1
Clinical status	Affected	Healthy	Healthy

CASE 5

Case 5 description	
Gender	Female
Age	35 years
Referral	Breast Cancer
Onset	Adult
Main clinical features	• Breast carcinoma • Neoplasm of the breast • NO Decreased fertility • NO Colon cancer



Information on the corresponding causative variant:

Disease: Breast Cancer
Gene: BRCA1
Chr Coordinate: 17:41215920G>T
Variant: c.4235C>A p.Ala1412Glu (missense)
See this case in RD-Connect GPAP: https://playground.rd-connect.eu/genomics/?filters=61

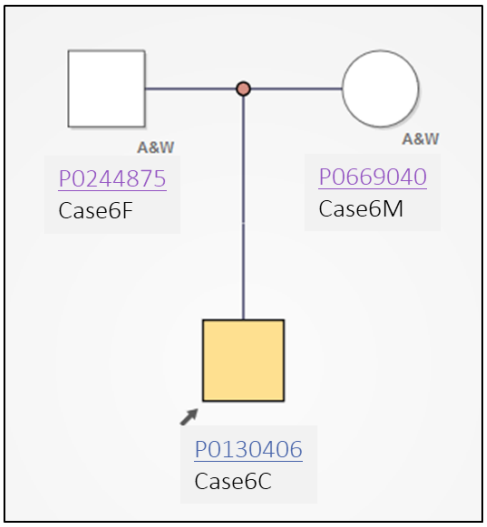
FAM0001822	Index Case (female)	Father	Mother
Individual ID	Case5C	Case5F	Case5M
Phenopacket ID	P0008909	P0008911	P0008910
Genotype	0/1	0/0	0/1
Clinical status	Affected	Healthy	Affected

CASE 6

This case was made for matchmaking purposes: similar phenotype and same causative variant as case 1.

Case 6 description

Gender	Male
Age	5 years
Referral	Congenital myasthenic syndrome
Onset	Neonatal
Global pace of progression	Progressive
Main clinical features	• Neonatal hypotonia • Distal arthrogryposis • Skeletal muscle atrophy • Muscular hypotonia of the trunk



Information on the corresponding causative variant:

Disease: Central core disease
Gene: RYR1
Chr Coordinate: 19:39062815G>C
Variant: c.13888G>C p.Glu4630Gln (missense)
See this case in RD-Connect GPAP: https://playground.rd-connect.eu/genomics/?filters=56

FAM0569511	Index Case (male)	Father	Mother
Individual ID	Case6C	Case6F	Case6M
Phenopacket ID	P0130406	P0244875	P0669040
Genotype	0/1	0/0	0/0
Clinical status	Affected	Healthy	Healthy

5) Files structure

Example of directories for case 4:

```
├── b1mg.synthetic.dataset.case4
    ├── case4_father
    │   ├── Case4_F.10.g.vcf.gz
    │   ├── Case4_F.11.g.vcf.gz
    │   ├── Case4_F.12.g.vcf.gz
    │   ├── Case4_F.13.g.vcf.gz
    │   ├── Case4_F.14.g.vcf.gz
    │   ├── Case4_F.15.g.vcf.gz
    │   ├── Case4_F.16.g.vcf.gz
    │   ├── Case4_F.17.g.vcf.gz
    │   ├── Case4_F.18.g.vcf.gz
    │   ├── Case4_F.19.g.vcf.gz
    │   ├── Case4_F.1.g.vcf.gz
    │   ├── Case4_F.20.g.vcf.gz
    │   ├── Case4_F.21.g.vcf.gz
    │   ├── Case4_F.22.g.vcf.gz
    │   ├── Case4_F.2.g.vcf.gz
    │   ├── Case4_F.3.g.vcf.gz
    │   ├── Case4_F.4.g.vcf.gz
    │   ├── Case4_F.5.g.vcf.gz
    │   ├── Case4_F.6.g.vcf.gz
    │   ├── Case4_F.7.g.vcf.gz
    │   ├── Case4_F.8.g.vcf.gz
    │   ├── Case4_F.9.g.vcf.gz
    │   ├── Case4_F.bam
    │   ├── Case4_F.bam.bai
    │   ├── Case4_F.MT.g.vcf.gz
    │   ├── case4_F_phenopacket.json
    │   ├── Case4_F.R1.fastq
    │   ├── Case4_F.R2.fastq
    │   ├── Case4_F.X.g.vcf.gz
    │   └── Case4_F.Y.g.vcf.gz
    ├── case4_index
    │   ├── Case4_IC.10.g.vcf.gz
    │   ├── Case4_IC.11.g.vcf.gz
    │   ├── Case4_IC.12.g.vcf.gz
    │   ├── Case4_IC.13.g.vcf.gz
    │   ├── Case4_IC.14.g.vcf.gz
    │   ├── Case4_IC.15.g.vcf.gz
```

```
|   ├── Case4_IC.16.g.vcf.gz
|   ├── Case4_IC.17.g.vcf.gz
|   ├── Case4_IC.18.g.vcf.gz
|   ├── Case4_IC.19.g.vcf.gz
|   ├── Case4_IC.1.g.vcf.gz
|   ├── Case4_IC.20.g.vcf.gz
|   ├── Case4_IC.21.g.vcf.gz
|   ├── Case4_IC.22.g.vcf.gz
|   ├── Case4_IC.2.g.vcf.gz
|   ├── Case4_IC.3.g.vcf.gz
|   ├── Case4_IC.4.g.vcf.gz
|   ├── Case4_IC.5.g.vcf.gz
|   ├── Case4_IC.6.g.vcf.gz
|   ├── Case4_IC.7.g.vcf.gz
|   ├── Case4_IC.8.g.vcf.gz
|   ├── Case4_IC.9.g.vcf.gz
|   ├── Case4_IC.bam
|   ├── Case4_IC.bam.bai
|   ├── Case4_IC.MT.g.vcf.gz
|   ├── case4_IC_phenopacket.json
|   ├── Case4_IC.R1.fastq
|   ├── Case4_IC.R2.fastq
|   ├── Case4_IC.X.g.vcf.gz
|   └── Case4_IC.Y.g.vcf.gz
├── case4_mother
|   ├── Case4_M.10.g.vcf.gz
|   ├── Case4_M.11.g.vcf.gz
|   ├── Case4_M.12.g.vcf.gz
|   ├── Case4_M.13.g.vcf.gz
|   ├── Case4_M.14.g.vcf.gz
|   ├── Case4_M.15.g.vcf.gz
|   ├── Case4_M.16.g.vcf.gz
|   ├── Case4_M.17.g.vcf.gz
|   ├── Case4_M.18.g.vcf.gz
|   ├── Case4_M.19.g.vcf.gz
|   ├── Case4_M.1.g.vcf.gz
|   ├── Case4_M.20.g.vcf.gz
|   ├── Case4_M.21.g.vcf.gz
|   ├── Case4_M.22.g.vcf.gz
|   ├── Case4_M.2.g.vcf.gz
|   ├── Case4_M.3.g.vcf.gz
|   ├── Case4_M.4.g.vcf.gz
|   ├── Case4_M.5.g.vcf.gz
|   ├── Case4_M.6.g.vcf.gz
```

```
|   ├── Case4_M.7.g.vcf.gz
|   ├── Case4_M.8.g.vcf.gz
|   ├── Case4_M.9.g.vcf.gz
|   ├── Case4_M.bam
|   ├── Case4_M.bam.bai
|   ├── Case4_M.MT.g.vcf.gz
|   ├── case4_M_phenopacket.json
|   ├── Case4_M.R1.fastq
|   ├── Case4_M.R2.fastq
|   ├── Case4_M.X.g.vcf.gz
|   └── Case4_M.Y.g.vcf.gz
└── case4_pedigree.ped
```