

Overview and evaluation of the Diagnostic Mutation Database

Mike Cornell
NGRL Manchester

Introduction

- ❑ DMuDB provides a route for sharing mutation data within and between UK diagnostic laboratories.
- ❑ Diagnostic molecular genetics laboratories in the UK analyse and report hundreds of mutations per year.
- ❑ Mutation data are not generally added to the many online databases or reported in journal publications.
- ❑ DMuDB presently contains more than 37,807 individual variants in 50 genes. There are 238 active users from 39 diagnostic laboratories.
- ❑ Majority of variants are BRCA or HNPCC variants but the range of data is expanding.
- ❑ Data belongs to the submitter.

Overview

- ❑ NGRL Manchester are conducting a review of DMuDB contents.
- ❑ We intend to expand data collection to other countries.
- ❑ The data generated by diagnostic labs could potentially be used to benchmark software for the diagnostic community.

Overview

- ❑ Analysed DMuDB data for BRCA1, BRCA2, APC, MLH1, MSH2 and MSH6.
- ❑ Compared with data from BIC, BRCA1 and BRCA2 publication databases and InSiGHT database (APC, MLH1, MSH2 and MSH6).
- ❑ Generated a test set of pathogenic and non-pathogenic missense variants to validate missense tools.

Types of variants submitted

- ❑ Diagnostic labs are not all submitting the same types of variants.
- ❑ Some labs are submitting exclusively pathogenic variants.
- ❑ Very few variants are classified as “probably pathogenic” or “probably non-pathogenic”.
- ❑ More than 90% of DMuDBs APC variants are classified as pathogenic.

Use of correct HGVS nomenclature

- ❑ DMuDB variants were checked using Mutalyzer 2.
- ❑ Use of correct HGVS nomenclature will allow comparison with other LSDBs.
- ❑ A few variants have been named incorrectly. These are being analysed further.

| Gene | HGVS errors |
|-------|-------------|
| BRCA1 | 9 |
| BRCA2 | 308* |
| MLH1 | 11 |
| MSH2 | 2 |
| MSH6 | 0 |
| APC | 7 |

*U43746.1:c.1114A>C,
Mutalyzer predicts a C at this
position.

Data Consistency

Have contradictory classifications (i.e. pathogenic and non-pathogenic) been recorded for variants within the same database?

DMuDB

| Gene | Non Consistent | Consistent |
|-------|----------------|------------|
| APC | 0 | 61 |
| BRCA1 | 0 | 146 |
| BRCA2 | 1 | 171 |
| MLH1 | 0 | 43 |
| MSH2 | 0 | 61 |
| MSH6 | 0 | 14 |

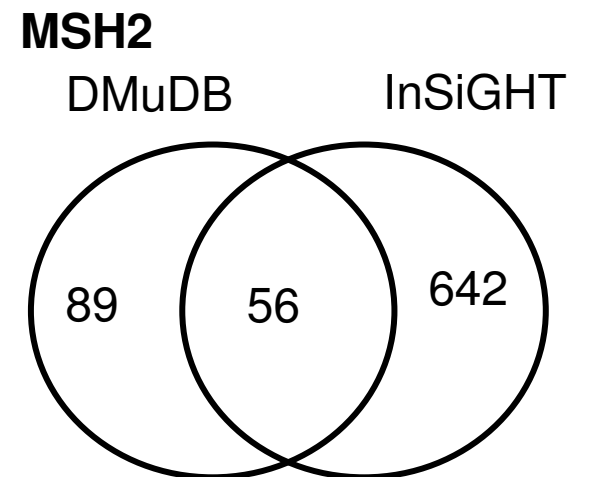
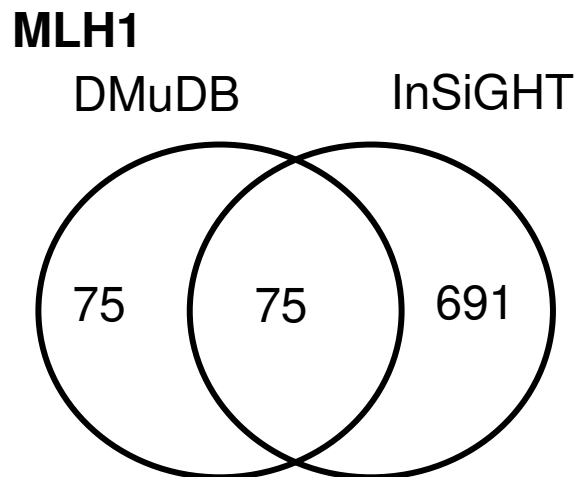
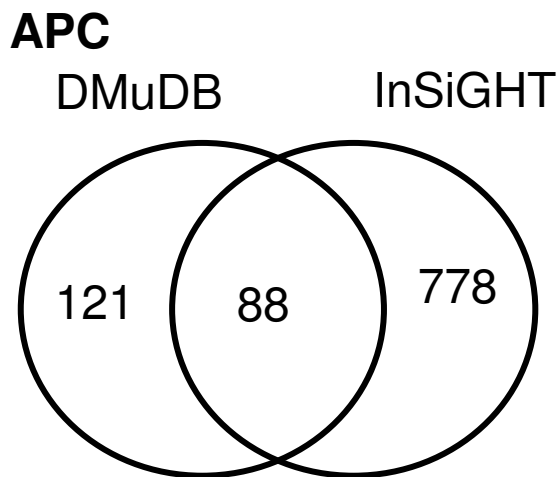
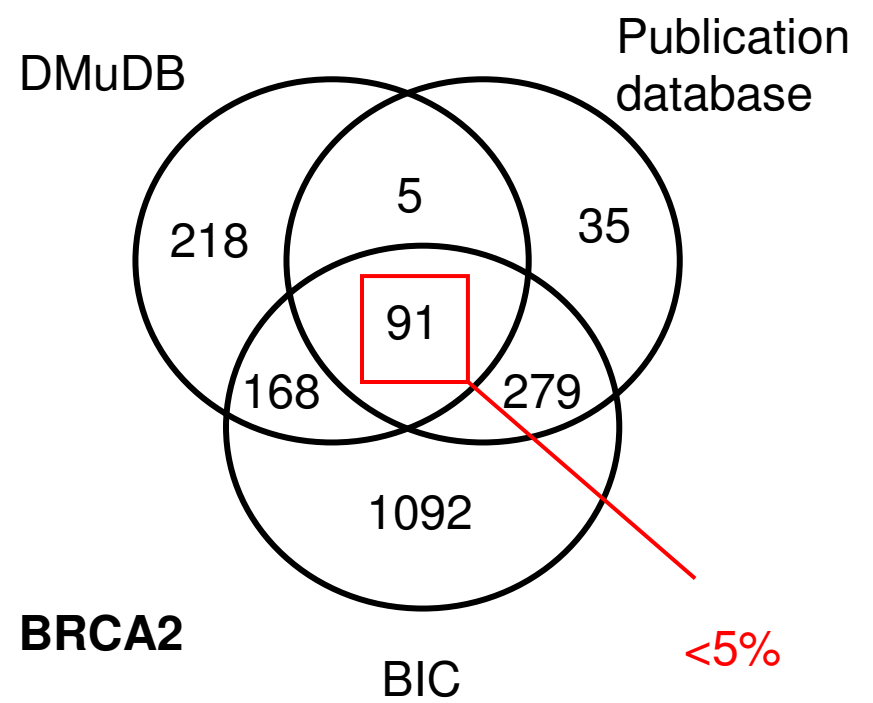
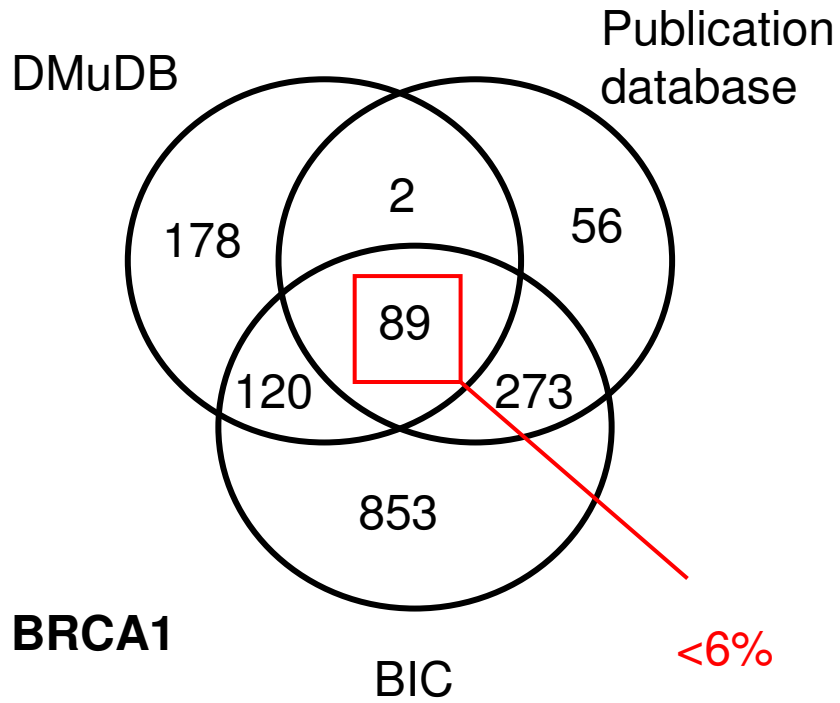
External databases

| Gene | Non Consistent | Consistent |
|-----------|----------------|------------|
| BIC BRCA1 | 0 | 365 |
| BIC BRCA2 | 0 | 377 |
| BRCA1 | 25 | 128 |
| BRCA2 | 2 | 60 |
| APC | 0 | 253 |
| MLH1 | 97 | 155 |
| MSH2 | 44 | 160 |
| MSH6 | 3 | 55 |

Classification of VUS entries

For DMuDB entries classified as “unknown significance”, have other users been able to provide a “pathogenic” or “non-pathogenic” classification?

| | Pathogenic | Non-Pathogenic |
|-------|------------|----------------|
| BRCA1 | 10 | 31 |
| BRCA2 | 46 | 8 |
| APC | 1 | 0 |
| MLH1 | 7 | 2 |
| MSH2 | 1 | 1 |
| MSH6 | 0 | 0 |



Consistency with external databases

| DMuDB | External database | Agreement | Disagreement |
|--------------|-------------------|------------|--------------|
| BRCA1 | BIC BRCA1 | 160 | 19 |
| BRCA1 | LOVD BRCA1 | 51 | 100 |
| BRCA2 | BIC BRCA2 | 151 | 6 |
| BRCA2 | LOVD BRCA2 | 20 | 41 |
| APC | InSiGHT APC | 79 | 58 |
| MLH1 | InSiGHT MLH1 | 68 | 98 |
| MSH2 | InSiGHT MSH2 | 60 | 74 |
| MSH6 | InSiGHT MSH6 | 20 | 23 |

Assessment of missense tools

- ❑ Use combined set of 2328 missense variants to evaluate tools.
- ❑ Tools assessed so far:
 - SIFT
 - AlignGVGD
 - Polyphen2
 - Panther
 - nsSNP
 - PMut
 - CanPredict

Defining pathogenic

- Different tools return different classifications
 - SIFT: Tolerant / Intolerant
 - AlignGVGD: Classes C0 / C15 / C35 / C45 / C55 / C65
 - Polyphen2: benign / possibly damaging / probably damaging
 - Panther: Pdeleterious <0.5 / ≥ 0.5
 - ns SNP: neutral / unknown / disease
 - PMut: neutral / pathological
 - CanPredict : likely not cancer / likely cancer



Tolerated



Not Tolerated



VUS

SIFT: Input proteins or alignments

- ❑ SIFT allows users to input an alignment **OR** input a protein and the software generates an alignment.
- ❑ Using own alignments reduces number of low quality predictions.
- ❑ How do predictions compare?

Input protein

| | Tolerated | Not Tolerated |
|-----------------|------------------------------|------------------|
| Input alignment | Tolerated 568 (55% agree) | 459 |
| Input alignment | 263 | 1037 (80% agree) |

Depth of alignment

- Compared AlignGVGD predictions for BRCA1 variants using “human to frog” and “human to sea urchin” alignments.
- For some variants the choice of alignment completely alters predictions.

| | sea urchin prediction | frog prediction |
|--------|-----------------------|-----------------|
| S864A | C0 | C65 |
| S186Y | C15 | C65 |
| K1208E | C0 | C55 |
| S1722F | C15 | C65 |
| E1694K | C0 | C55 |
| L1854P | C15 | C65 |
| E638K | C0 | C55 |
| E1214K | C0 | C55 |

| | sea urchin prediction | frog prediction |
|--------|-----------------------|-----------------|
| P1150S | C0 | C65 |
| S864L | C0 | C65 |
| S1651F | C0 | C65 |
| K1406N | C0 | C65 |
| I641T | C0 | C65 |
| R133C | C0 | C65 |
| M1400T | C0 | C65 |
| S1655A | C0 | C65 |
| S1655P | C0 | C65 |

Similarity of predictions

| | Align GVDG | Polyphen2 | Panther | Pmut | ns SNP | Can Predict |
|------------|------------|-----------|---------|------|----------------|----------------|
| SIFT | 66.4 | 56.7 | 68.3 | 54.7 | 11.3 (77.6) | 66.9 |
| Align GVDG | | 54.1 | 70.4 | 52.7 | 19.3 (74.5) | 57.1 |
| Polyphen 2 | | | 58.1 | 50.1 | 27.3 (64.2) | 62.1 |
| Panther | | | | 64.0 | 14.0 (78.0) | 74.0 |
| Pmut | | | | | 13.1 (76.6) | 58.8 |
| ns SNP | | | | | | 23.6 (56.6) |

Producing test sets

- ❑ Generated test sets of pathogenic and benign variants.
- ❑ Included variants from DMuDB, BIC, InSiGHT and BRCA1/2 publication databases.
- ❑ Variant needed to be reported at least twice with no contradictory reports.
- ❑ 132 pathogenic and 211 benign variants.

Quality measurements

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

Quality Statistics

| | Sensitivity | Specificity | Accuracy | MCC |
|-------------------|--------------------|--------------------|-----------------|-------------|
| SIFT | 0.85 | 0.47 | 0.38 | 0.32 |
| AlignGVGD | 0.7 | 0.85 | 0.4 | 0.55 |
| Polyphen2 | 0.87 | 0.62 | 0.39 | 0.48 |
| Panther | 0.75 | 0.71 | 0.37 | 0.45 |
| Pmut | 0.65 | 0.56 | 0.38 | 0.2 |
| ns SNP | 0.72 | 0.86 | 0.62 | 0.56 |
| CanPredict | 0.96 | 0.3 | 0.52 | 0.35 |

Summary

- ❑ Diagnostic labs generate high-quality data to SOPs.
- ❑ Many of the variants in DMuDB are unique to the database.
- ❑ DMuDB data is more consistent than some LSDBs.
- ❑ There are considerable differences in the classification of variants across LSDBs.
- ❑ Submitting data to DMuDB may aid the reclassification of VUSs.
- ❑ It will also enable us to develop test sets for testing bioinformatics tools.
- ❑ Reliable test sets would allow development of reliable guidelines for alignments.

Acknowledgements

- ❑ Glen Dobson
- ❑ Bharathi Kattamuri
- ❑ Kathryn Robertson
- ❑ Andrew Devereau
- ❑ Beth Hellen
- ❑ Support from Department of Health.
- ❑ Data submitters throughout the UK.