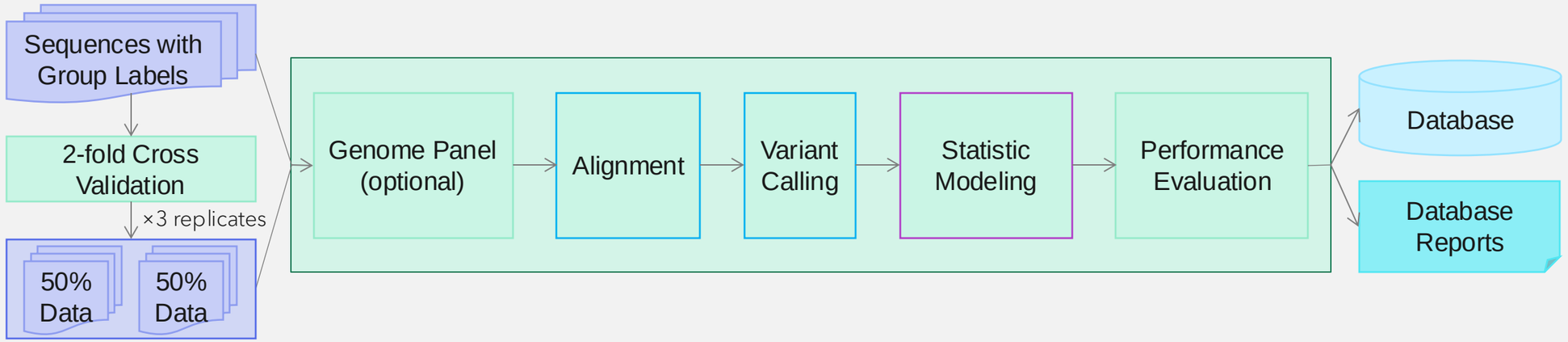
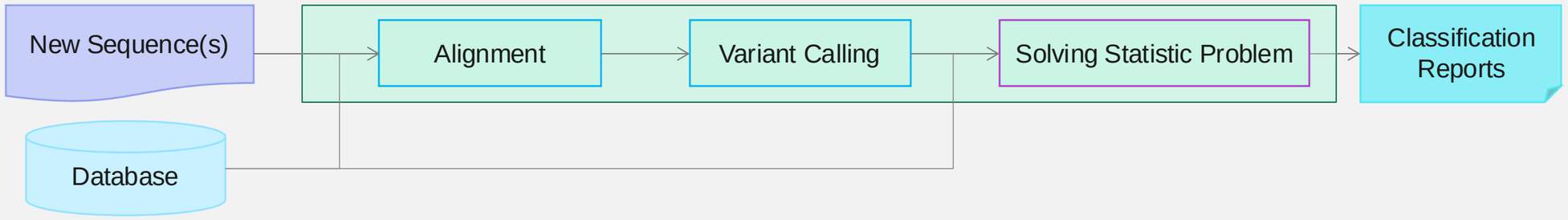
Server (Julia)

## Database Building




## Sample Classification

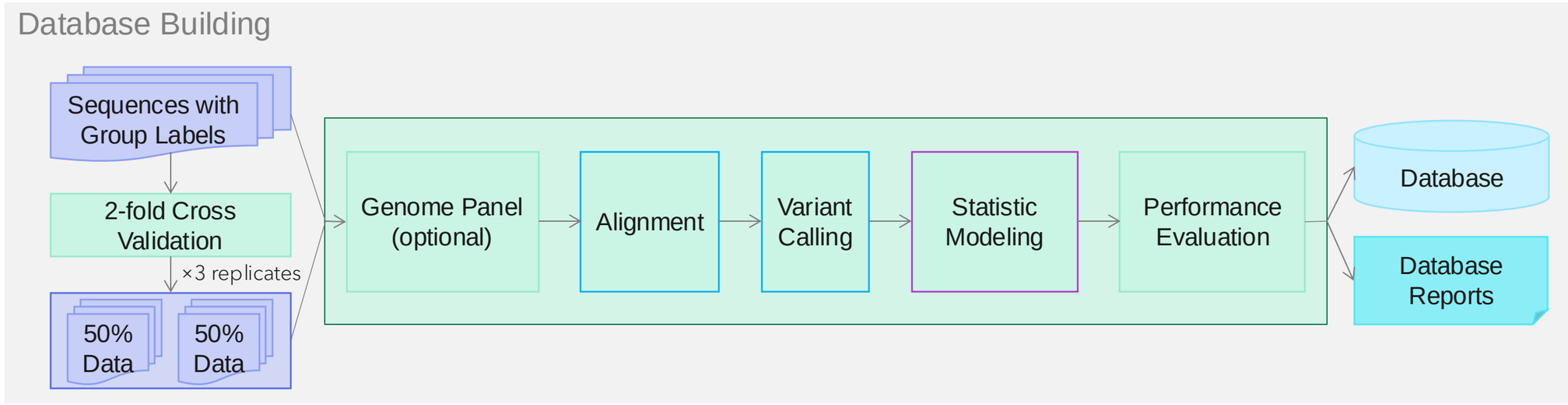


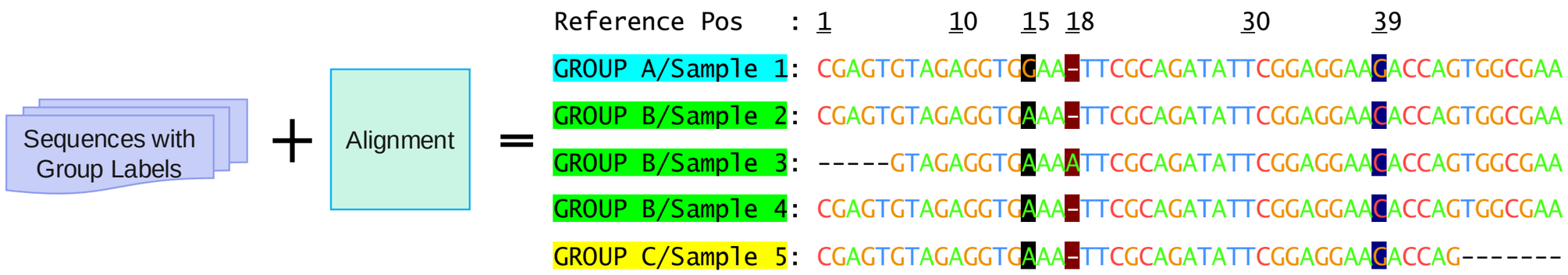

## Supportive Modules

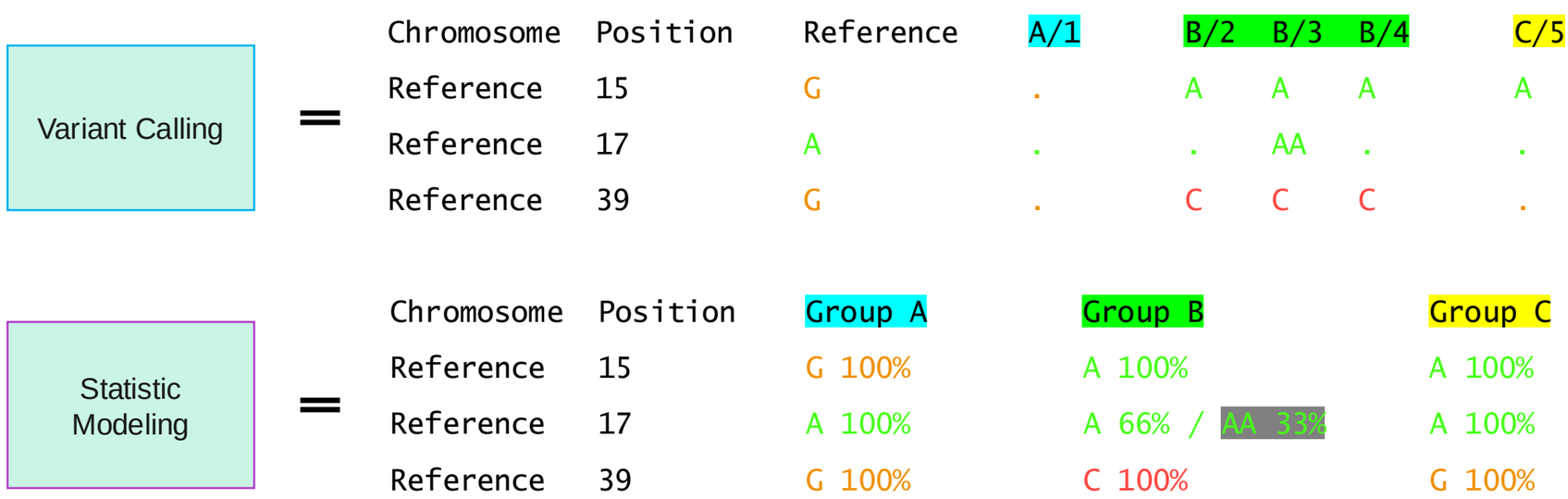
- Workload Manager
- Web API Server
- Database Cache
- Authentication
- User and File Database

## User Interface (Vue + Quasar Framework)

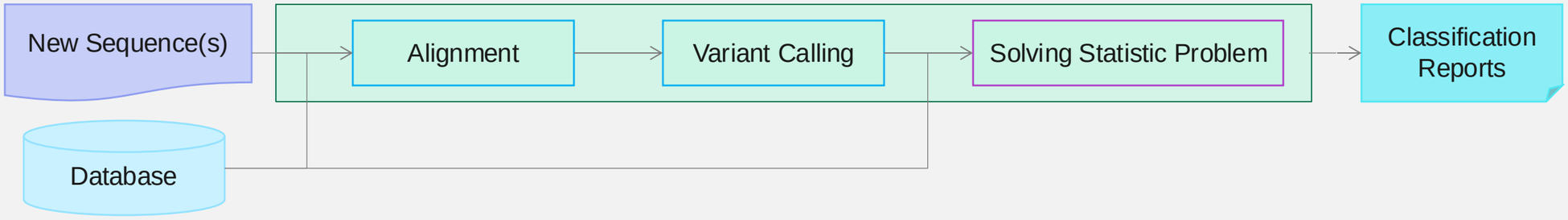
- New Analysis
- Database Performance
- Report Query
- Database Building
- User Space
- Login / Register


Database Building
Sequences with Group Labels
2-fold Cross Validation
×3 replicates
50% Data
50% Data
Genome Panel (optional)
Alignment
Variant Calling
Statistic Modeling
Performance Evaluation
Database
Database Reports


Sequences with Group Labels
+
Alignment
=
Reference Pos : 1 10 15 18 30 39
GROUP A/Sample 1: CGAGTGTAGAGGTGCAA-TTCGCAGATATTCGGAGGAACACCAGTGGCGAA
GROUP B/Sample 2: CGAGTGTAGAGGTGAAA-TTCGCAGATATTCGGAGGAACACCAGTGGCGAA
GROUP B/Sample 3: -----GTAGAGGTGAAAATTCGCAGATATTCGGAGGAACACCAGTGGCGAA
GROUP B/Sample 4: CGAGTGTAGAGGTGAAA-TTCGCAGATATTCGGAGGAACACCAGTGGCGAA
GROUP C/Sample 5: CGAGTGTAGAGGTGAAA-TTCGCAGATATTCGGAGGAACACCAG-------


Variant Calling
=
Chromosome Position Reference A/1 B/2 B/3 B/4 C/5
Reference 15 G . A A A A
Reference 17 A . . AA . .
Reference 39 G . C C C .
Statistic Modeling
=
Chromosome Position Group A Group B Group C
Reference 15 G 100% A 100% A 100%
Reference 17 A 100% A 66% / AA 33% A 100%
Reference 39 G 100% C 100% G 100%
=
Database

## Sample Classification




Database =

| Chromosome | Position | Group A | Group B | Group C |
|---|---|---|---|---|
| Reference | 15 | G 100% | A 100% | A 100% |
| Reference | 17 | A 100% | A 66% / AA 33% | A 100% |
| Reference | 39 | G 100% | C 100% | G 100% |

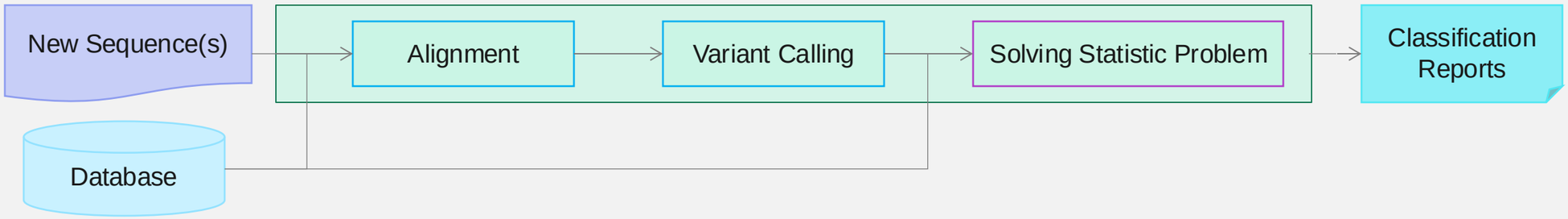
New Sequence(s) + Alignment =

```
Position        : 1         10    15 18            30        39
Reference       : CGAGTGTAGAGGTGCAA-TTCGCAGATATTCGGAGGAACACCAGTGGCGAA
New Sequence    : CGAGTGTAGAGGTGAAA-TTCGCAGATTTTCGGAGGAACACCAGTGGCGAA
```

Variant Calling =

| Chromosome | Position | Reference | New Sequence |
|---|---|---|---|
| Reference | 15 | G | A 100% |
| Reference | 17 | A | . 100% |
| Reference | 28 | A | T 100% |
| Reference | 39 | G | . 100% |

## Sample Classification




Solving Statistic Problem =

| Chromosome | Position | New Sequence | .× | Group A | Group B | Group C |
|---|---|---|---|---|---|---|
| Reference | 15 | A 100% | .× | G 100% | A 100% | A 100% |
| Reference | 17 | A 100% | .× | A 100% | A 66% / AA 33% | A 100% |
| Reference | 39 | G 100% | .× | G 100% | C 100% | G 100% |
| **maximum score** a sample can get | | | | 3.0 | 2.66 | 3.0 |

↓

| Chromosome | Position | Group A | Group B | Group C |
|---|---|---|---|---|
| Reference | 15 | A×G = 0% | A×A = 100% | A×A = 100% |
| Reference | 17 | A×A = 100% | A×A + A×AA = 66% + 0% = 66% | A×A = 100% |
| Reference | 39 | G×G = 100% | G×C = 0% | G×G = 100% |
| **actual score** the sample gets | | 2.0 | 1.66 | 3.0 |
| **maximum score** a sample can get | | 3.0 | 2.66 | 3.0 |
| **identity** = actual / maximum | | 66% | 62.5% | 100% |

We assume the variant frequencies of each location operate as a discrete Markov chain. The classification groups are hidden from observations of variants, so the problem fits a Hidden Markov Model (HMM) and can be solved using the maximum likelihood method.