```
┌─────────────────────────────────────┐        ┌──────────────────────────────────────────┐
│ Read all the files in a directory   │        │ We expect some experiments to be           │
│ split into functions.               │        │ performed as the process is refined.       │
│ Remove comments and blank lines     │        │  • Should the code be tokenized before     │
└─────────────────────────────────────┘        │    hashing to eliminate variable name      │
                  │                             │    differences                             │
                  ▼                             │  • Should we have finer parsing than       │
┌─────────────────────────────────────┐        │    just functions                          │
│ Create a subfolder for each input   │        │        • A rolling collections of lines    │
│ file and fill each with a file for  │        │          at a time                         │
│ each function found – or just "main"│        │        • Code separated with comments      │
└─────────────────────────────────────┘        │          (referring to the originals       │
                  │                             │          source)                           │
                  ▼                             └──────────────────────────────────────────┘
┌─────────────────────────────────────┐
│ Hash each function                  │
│ (finer code blocks may be selected  │
│ later)                              │
└─────────────────────────────────────┘
                  │
                  ▼
┌─────────────────────────────────────┐
│ Build an internal dictionary of     │
│ hash values and code block          │
│ identifiers                         │
└─────────────────────────────────────┘
                  │
                  ▼
┌─────────────────────────────────────┐
│ Select pairs of "close" hash        │
│ matches                             │
└─────────────────────────────────────┘
                  │
                  ▼
            For each pair
┌─────────────────────────────────────┐        ┌──────────────────────────────────────────┐
│ Further comparison---               │───────▶│ Clone Analysis –                           │
│ Tokenize code and re-hash           │        │ Classify the clone pairs                   │
│                                     │        │ (type I, type II)                          │
│                                     │        │ Try to find a suitable measure of          │
│                                     │        │ clone exactness                            │
└─────────────────────────────────────┘        └──────────────────────────────────────────┘
```