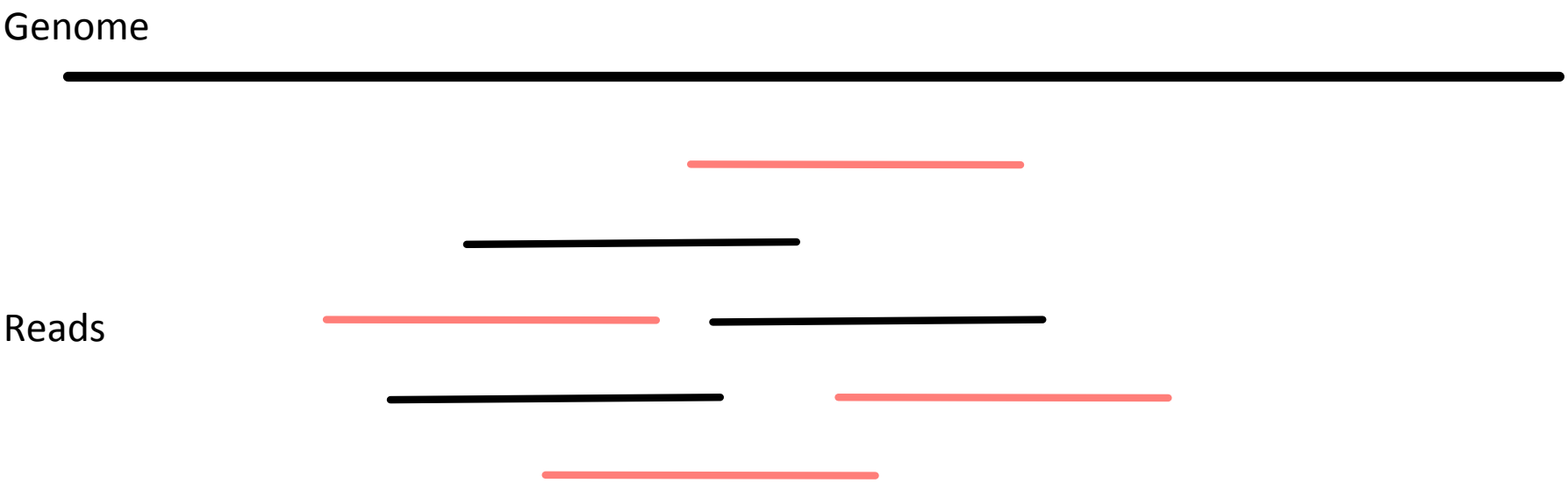
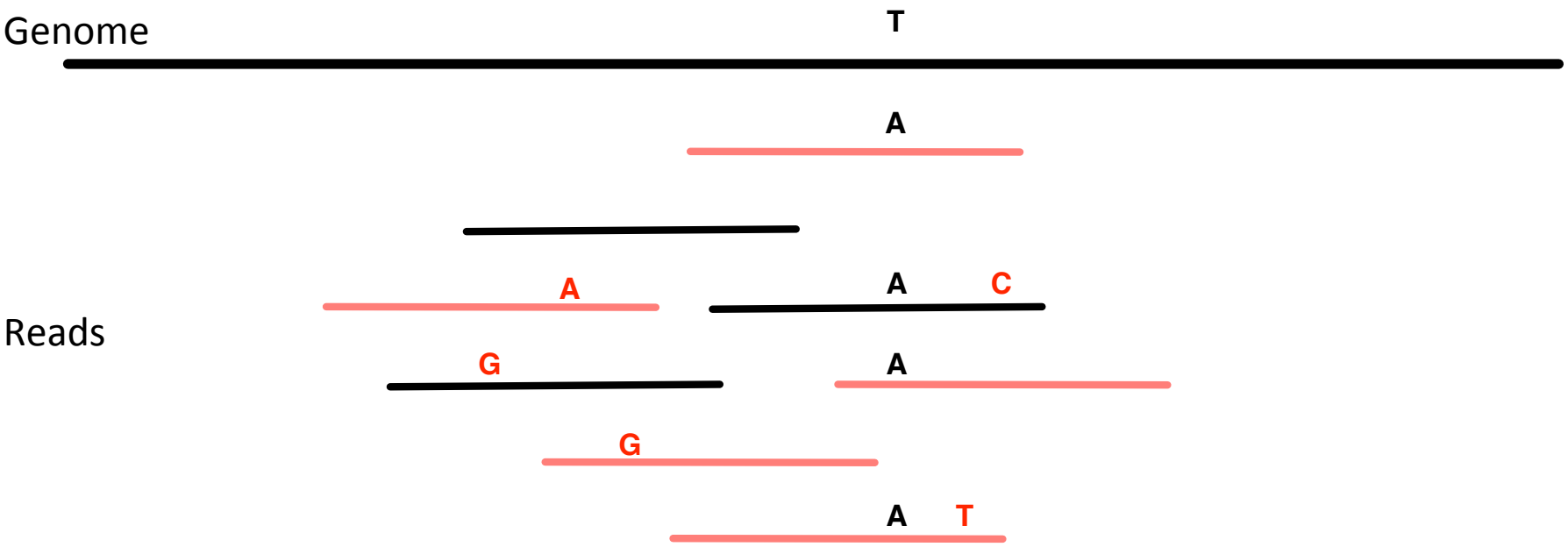
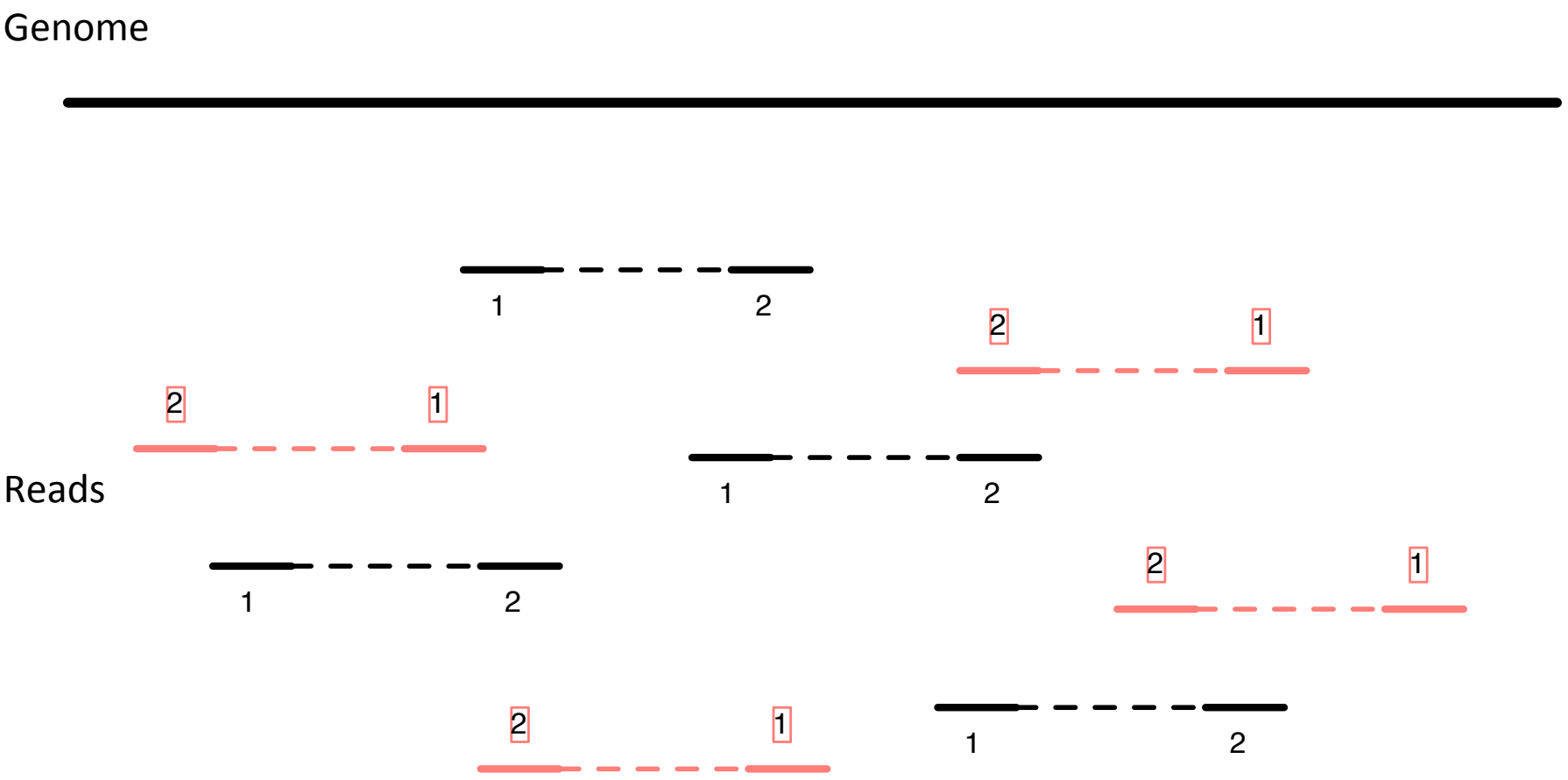
Shotgun sequencing



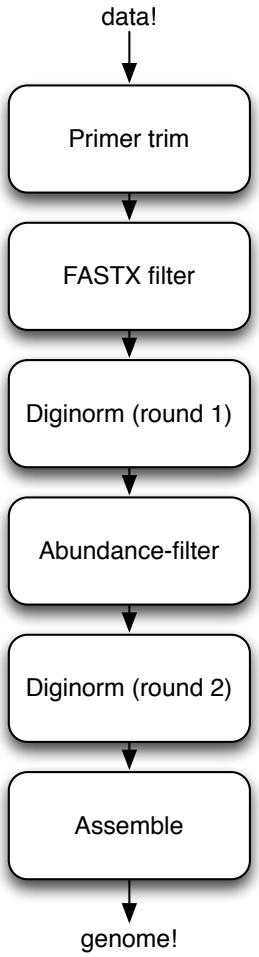
Mismatches: random vs alignable.



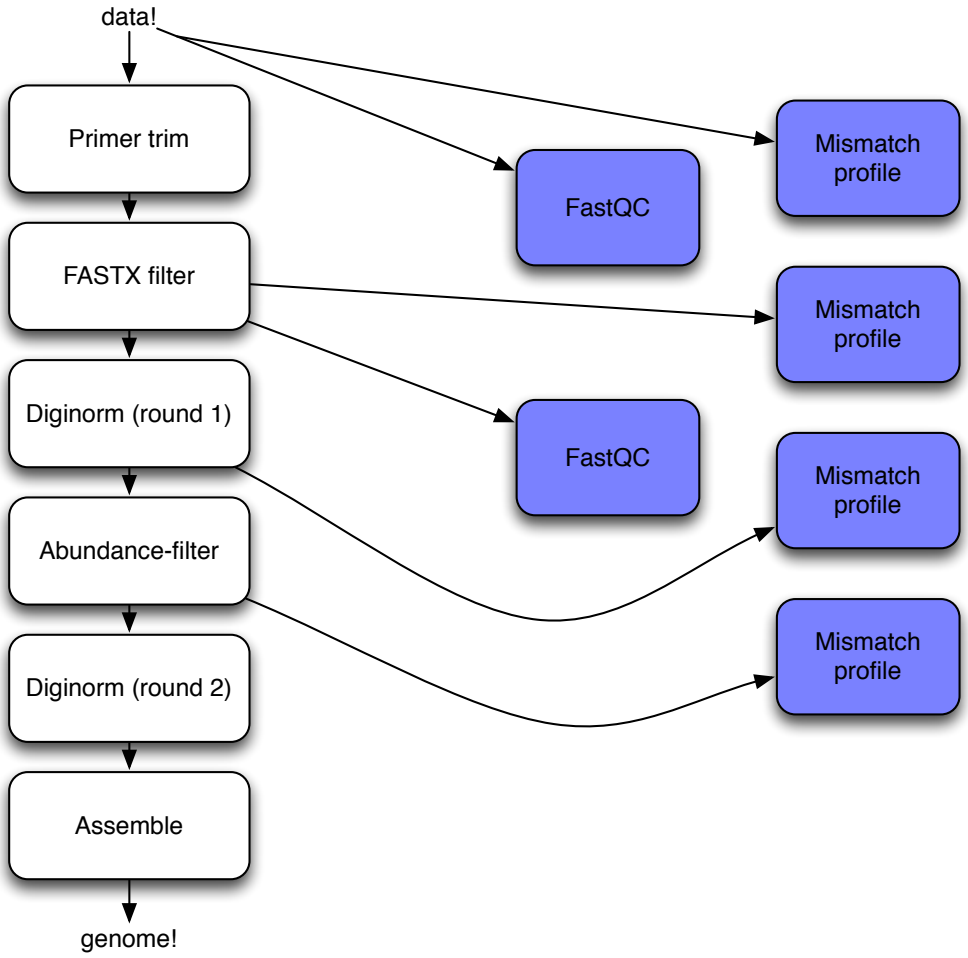
Paired end reads



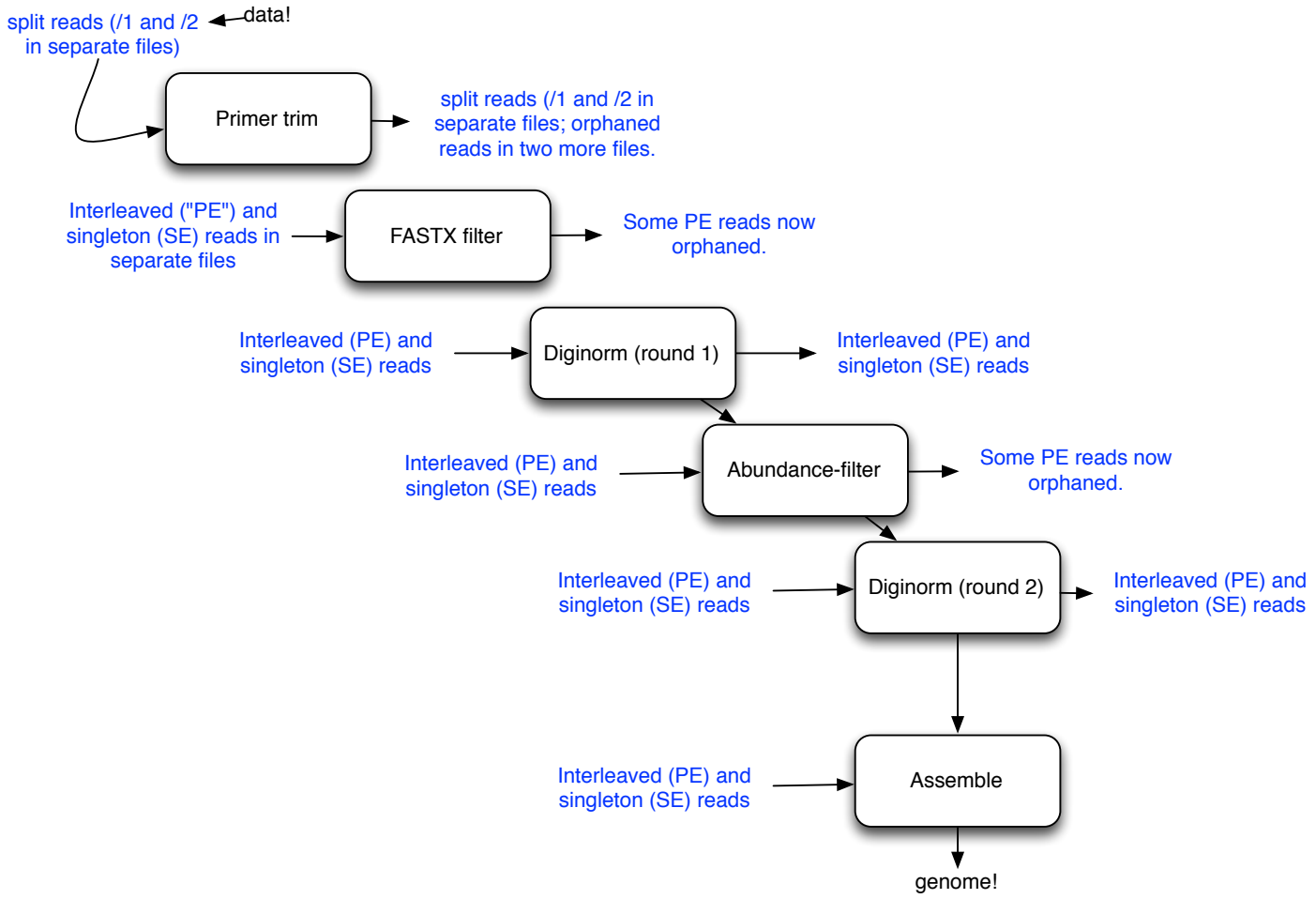
Assembly pipeline for E. coli



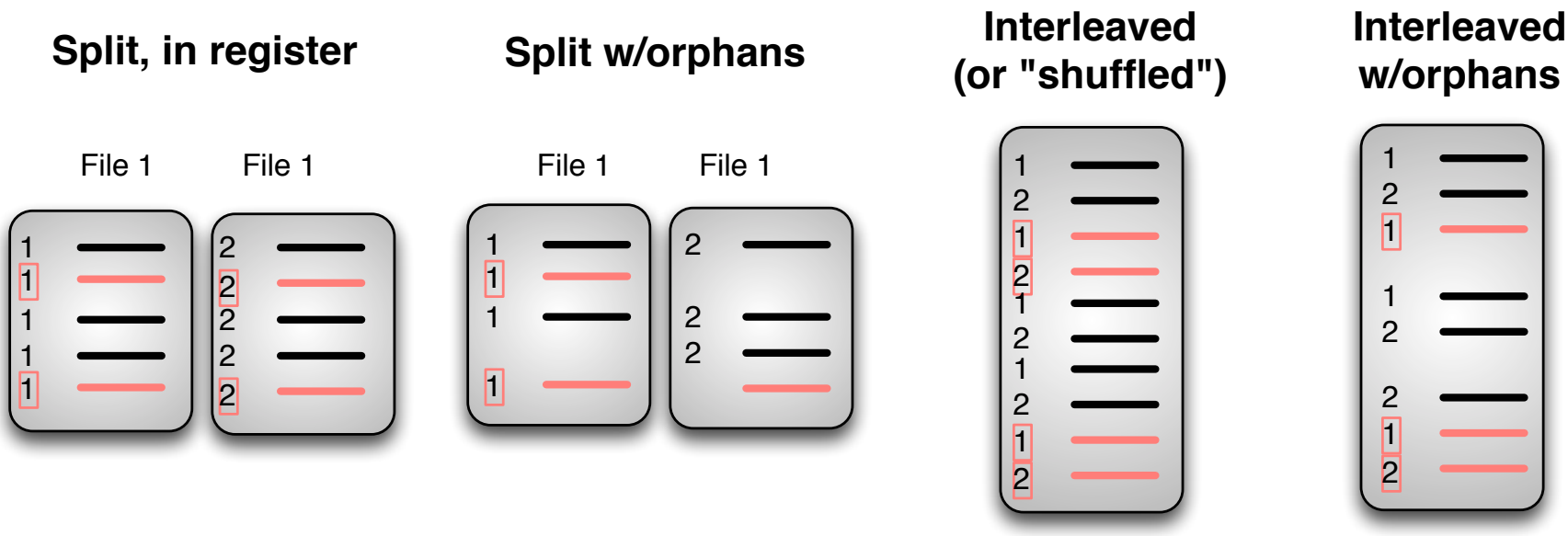
Assembly pipeline… with diagnostics



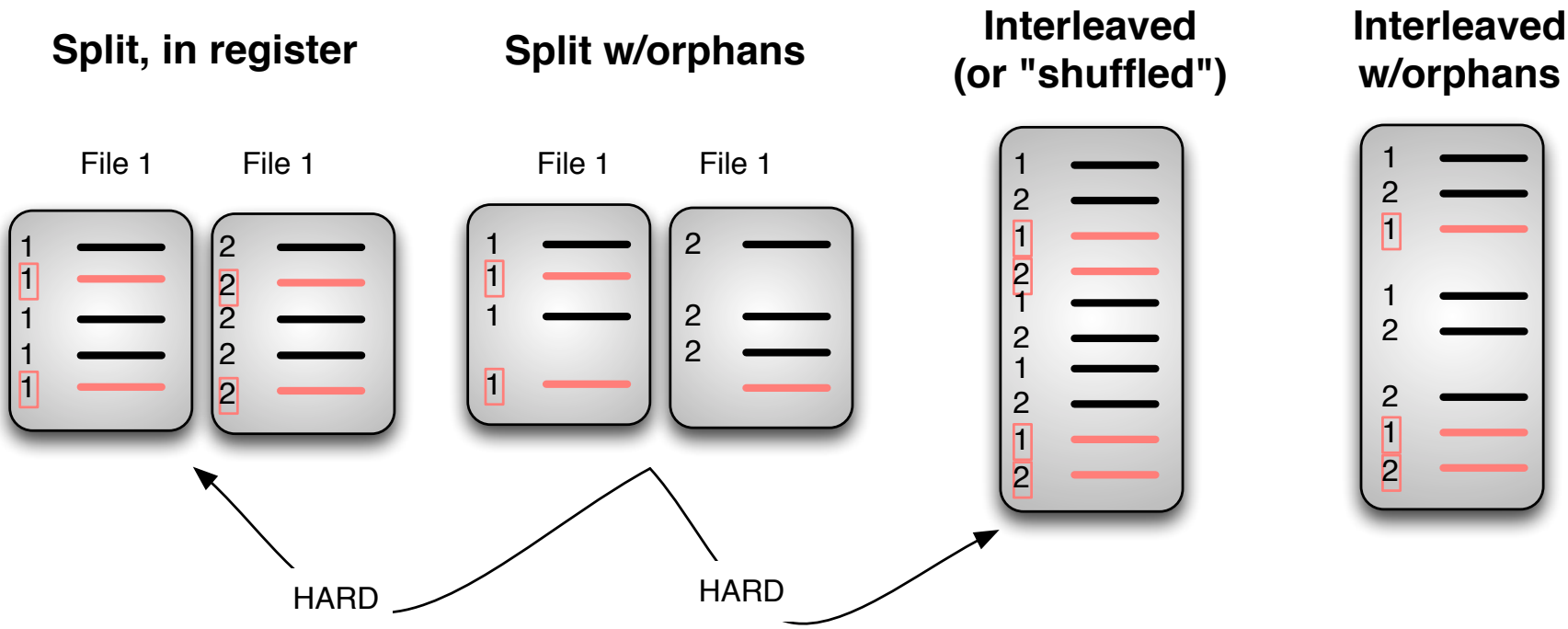
Tracking paired ends!



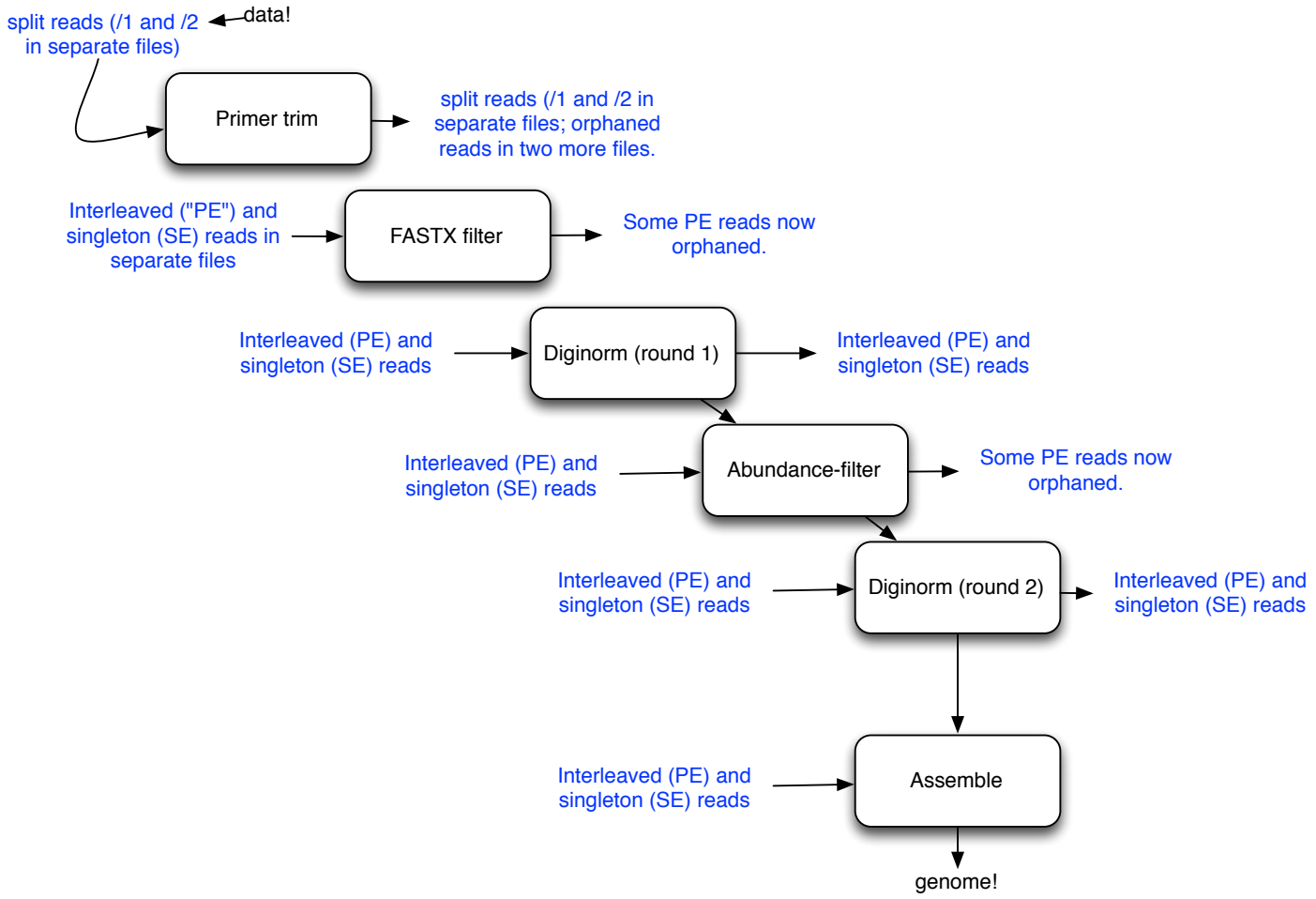
Storing reads: a taxonomy



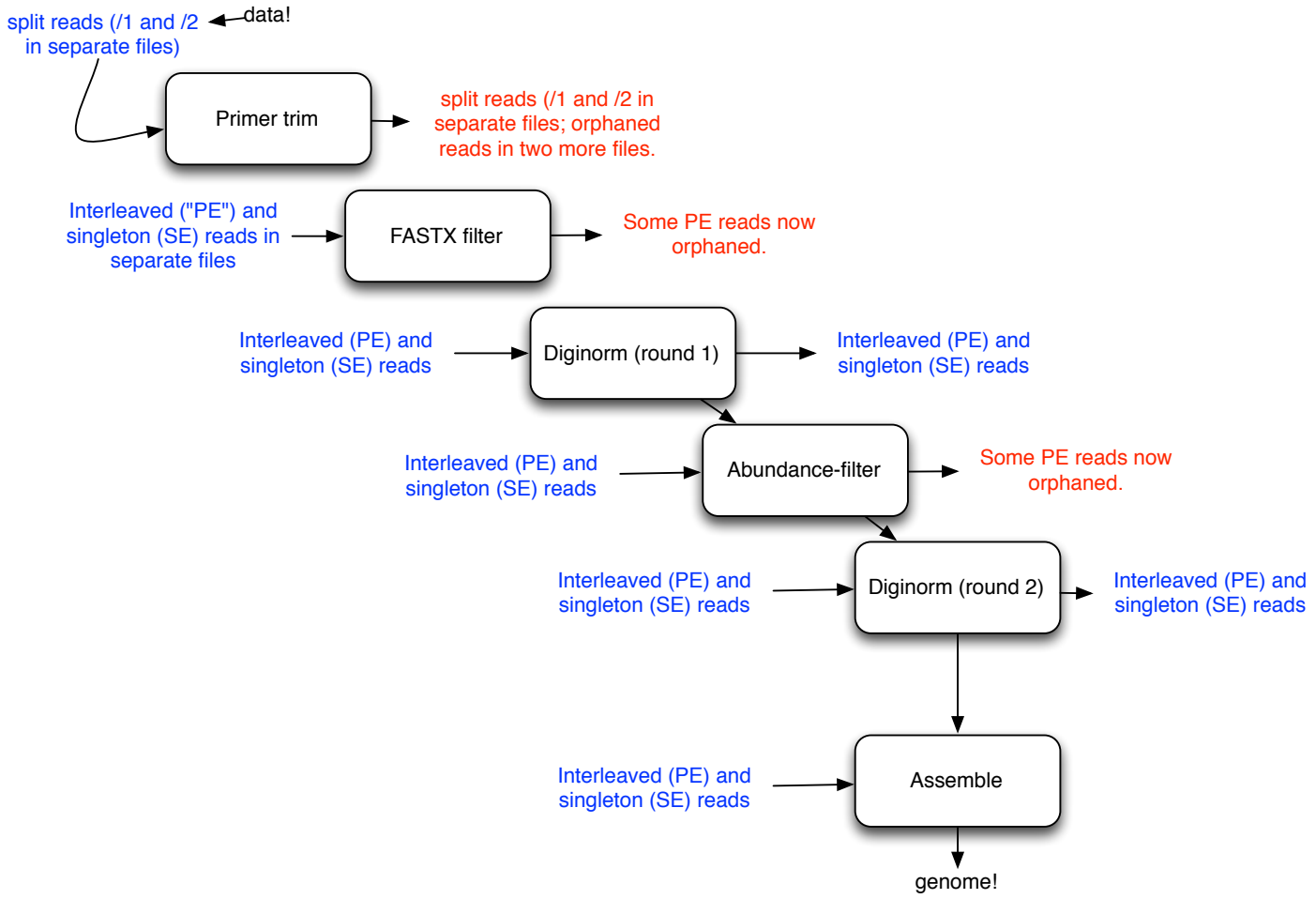
Interconverting can be hard.



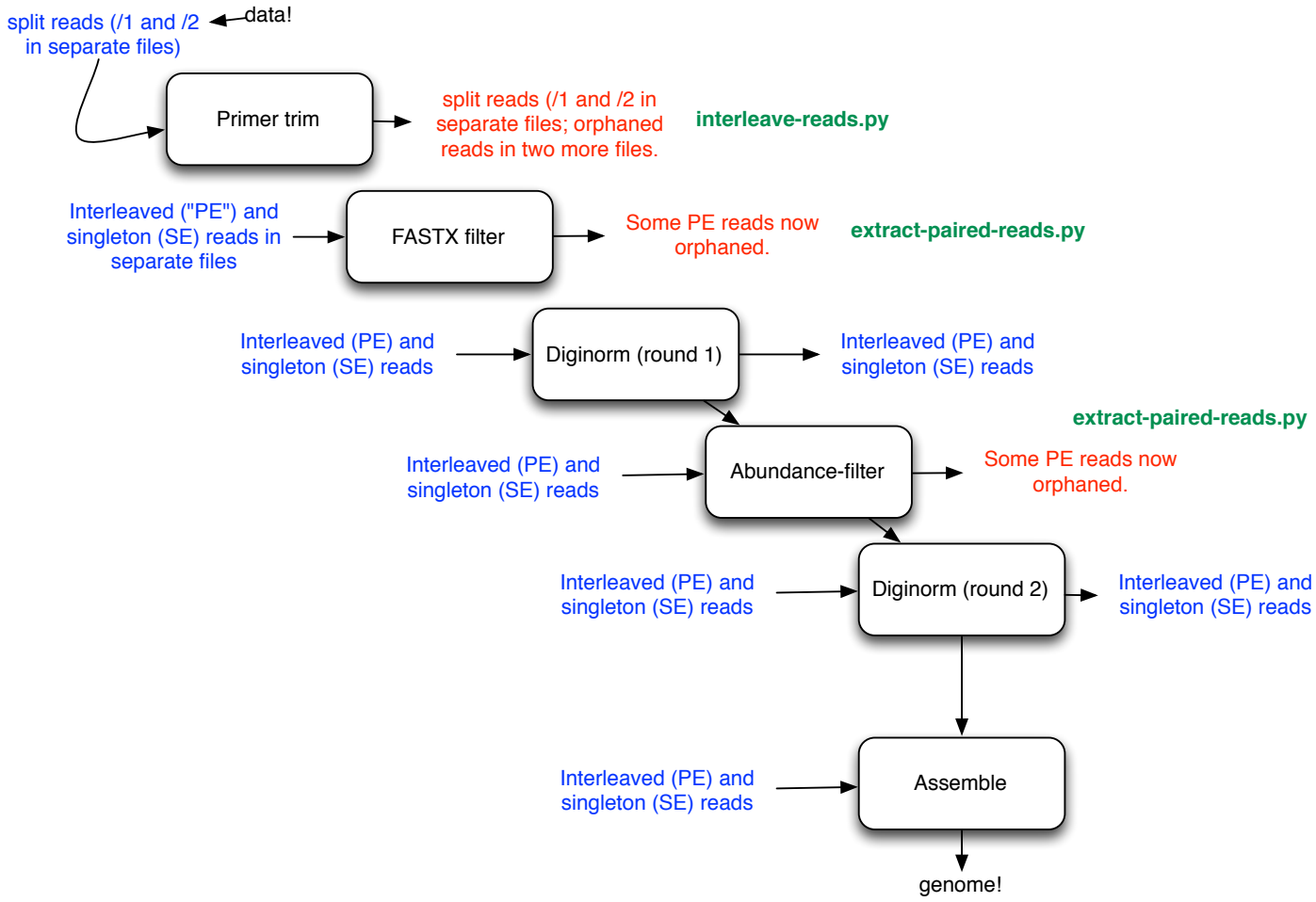
Tracking paired ends:



Must convert.



=>Explicit conversion steps in protocol



Stochastic output

- Velvet (like most assemblers) has stochastic components.
- What this means is that even for same input, same parameters, same machine => different output!?
- Two causes:
  - Random number generator;
  - Multithreading/multiprocessing (multiple "workers")

Everyone got different results from assembly; did it affect CRP gene?

```
crp             MVLGKPQTDPTLEWFLSHCHIHKYPSKSTLIHQGEKAETLYYIVKGSVAVLIKDEEGKEM
a               MVLGKPQTDPTLEWFLSHCHIHKYPSKSTLIHQGEKAETLYYIVKGSVAVLIKDEEGKEM
b               MVLGKPQTDPTLEWFLSHCHIHKYPSKSTLIHQGEKAETLYYIVKGSVAVLIKDEEGKEM
d               MVLGKPQTDPTLEWFLSHCHIHKYPSKSTLIHQGEKAETLYYIVKGSVAVLIKDEEGKEM
c               MVLGKPQTDPTLEWFLSHCHIHKYPSKSTLIHQGEKAETLYYIVKGSVAVLIKDEEGKEM
                ************************************************************

crp             ILSYLNQGDFIGELGLFEEGQERSAWVRAKTACEVAEISYKKFRQLIQVNPDILMRLSAQ
a               ILSYLNQGDFIGELGLFEEGQERSAWVRAKTACEVAEISYKKFRQLIQVNPDILMRLSAQ
b               ILSYLNQGDFIGELGLFEEGQERSAWVRAKTACEVAEISYKKFRQLIQVNPDILMRLSAQ
d               ILSYLNQGDFIGELGLFEEGQERSAWVRAKTACEVAEISYKKFRQLIQVNPDILMRLSAQ
c               ILSYLNQGDFIGELGLFEEGQERSAWVRAKTACEVAEISYKKFRQLIQVNPDILMRLSAQ
                ************************************************************

crp             MARRLQVTSEKVGNLAFLDVTGRIAQTLLNLAKQPDAMTHPDGMQIKITRQEIGQIVGCS
a               MARRLQVTSEKVGNLAFLDVTGRIAQTLLNLAKQPDAMTHPDGMQIKITRQEIGQIVGCS
b               MARRLQVTSEKVGNLAFLDVTGRIAQTLLNLAKQPDAMTHPDGMQIKITRQEIGQIVGCS
d               MARRLQVTSEKVGNLAFLDVTGRIAQTLLNLAKQPDAMTHPDGMQIKITRQEIGQIVGCS
c               MARRLQVTSEKVGNLAFLDVTGRIAQTLLNLAKQPDAMTHPDGMQIKITRQEIGQIVGCS
                ************************************************************

crp             RETVGRILKMLEDQNLISAHGKTIVVYGTR
a               RETVGRILKMLEDQNLISAHGKTIVVYGTR
b               RETVGRILKMLEDQNLISAHGKTIVVYGTR
d               RETVGRI----------------------
c               RETVGRILKMLEDQNLISAHGKTIVVYGTR
                *******
```

Program output in large-scale analyses

- You can almost always get something that makes *some* sense, i.e. isn't wrong.
- But: Insensitive? Incomplete? Biased?

Next two weeks in bioinfo

- Annotating genome
- Mapping & variant visualization
- Mapping & Quantification