Vall d'Hebron
Institut de Recerca

UNITAT
D'ESTADÍSTICA I
BIOINFORMÀTICA

A pipeline for studying minor variants in complex genetic populations using long reads from high-throughput sequencing technologies

Ortega-Serrano, I.; Quer, J.; Rodriguez-Frias F.; Sánchez-Pla, A.

ciberehd
MINISTERIO DE CIENCIA E INNOVACIÓN
Instituto de Salud Carlos III
FEDER
FONDO EUROPEO DE DESARROLLO REGIONAL
UCTS

Outline

- Motivation and experimental goals
  - HCV, HBV viruses
- Why Next Generation Sequencing (NGS) technologies?
- Which NGS technology?
- Pipeline to process reads
  - Non-specific filter to remove low-quality reads
  - Statistical modelling of erroneous variants
- Results

Motivation and experimental goals

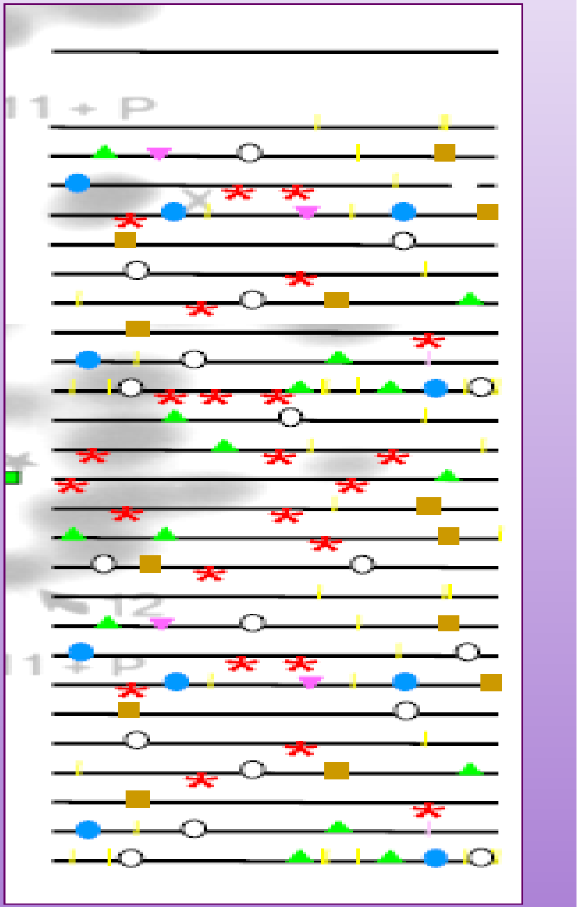
- Hepatitis C and Hepatitis B viruses circulate as quasispecies

In an infected patient:

(1) The population of viruses presents high rates of mutation and replication. It is a complex mixing of different mutants.

(2) These mutants (*variants*) are related amongst them

(3) They are subjects to competition and natural selection.



Motivation and experimental goals

- Goals of the study:
  - Detection and quantification of **mutations** or **combination of mutations** that could confer resistance to viral inhibitors in samples from chronically infected patients.

- HCV: mutations in NS3 protease
- HBV: mutations in polymerase

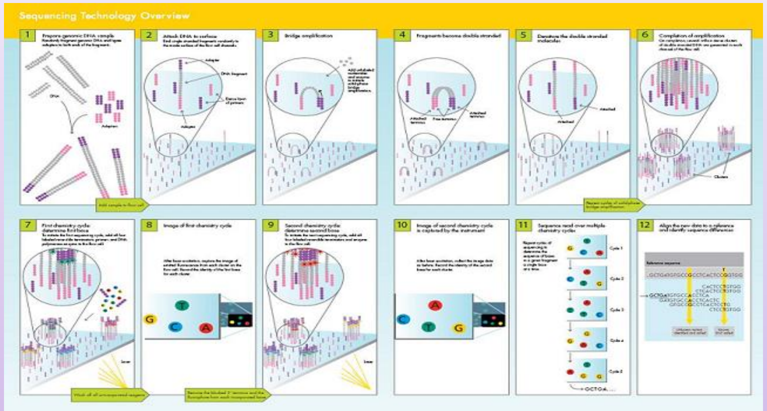
Why NGS technologies?

- Minor variants often play an important role in the development of resistance to antiviral treatments in patients, even if they are present in a very low percentage in the population.

- Minor variants may not be detected by classical sequencing methods
  - You obtain hundreds of sequences with much effort and high cost

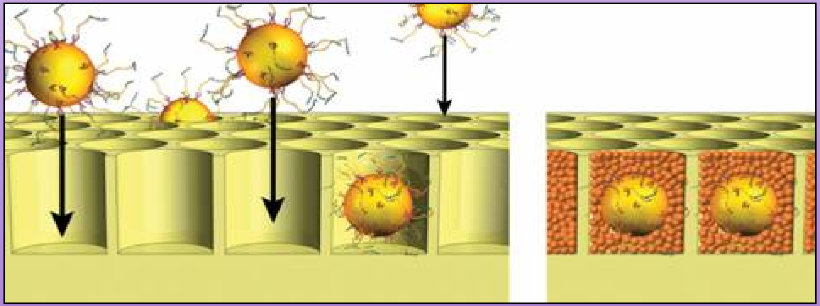
- NGS tools allow to detect minor variants efficiently
  - You obtain thousands of sequences with relatively low cost

Which NGS technology?



- Solexa-Illumina GA/AB SOLiD/....
  - 35-100 Short reads, useful for the study of SNPs or short regions
- 454-Roche GS FLX platform:
  - With longer reads (250-400) a “wider picture” is obtained
  - This is an advantage for the study of combination of mutations in a same sequence
  - Artifactual indels in homopolymers:

```
...CGGCCGGGACAAAAACCAGGTGGA....
...CGGCCGGGACAAAACCAGGTGGA.....
...CGGCCGGACAAAAACCAGGTGGA.....
...CGGCCGGGACAAAAAACCAGGTGGA...
```

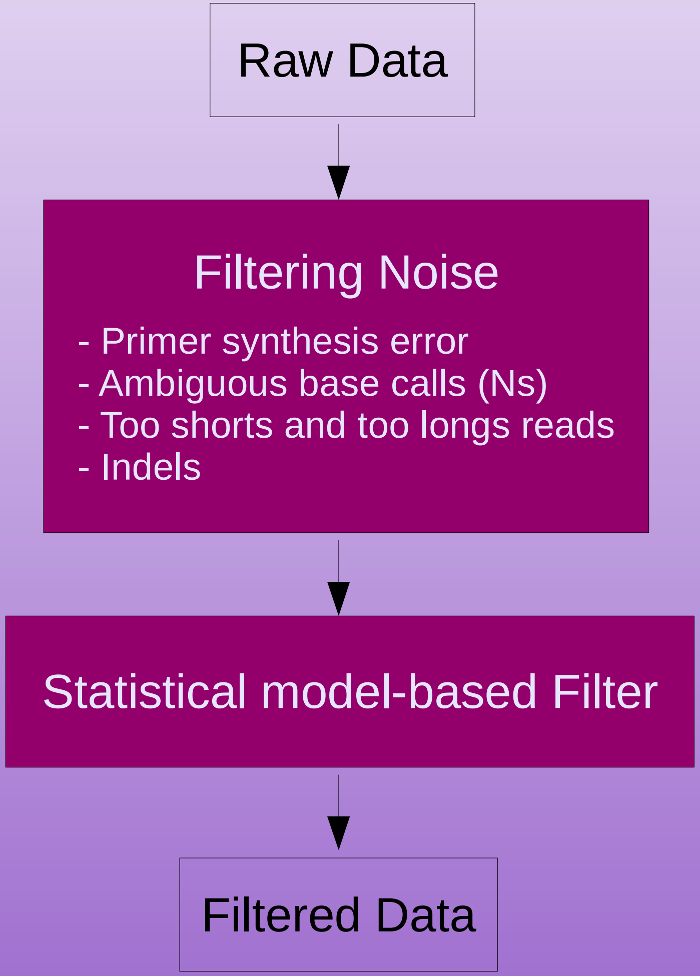


Nucleotide reference sequence of HBV polymerase:

```
5'
ATGTTTCCCTCATGTTGCTGTAC.AAA.ACC.TAC.GGA.TGG.AAA.TTG.CAC.CTG.TAT.TCC.CAT.CCC.ATC.GTC.CTG.GGC.TTT.CGC.AAA.
ATA.CCT.ATG.GGA.GTG.GGC.CTC.AGT.CCG.TTT.CTC.TTG.GCT.CAG.TTT.ACT.AGT.GCC.ATT.TGT.TCA.GTG.GTT.CGT.AGG.GCT.
TTC.CCC.CAC.TGT.TTG.GCT.TTC.AGC.TAT.ATG.GAT.GAT.GTG.GTA.TTGGGGGCCAAGTCTGTACAG 3'
```

Pipeline to process reads

- Goal: to obtain accurate estimations of proportions of the variants.



Raw Data

Filtering Noise

- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

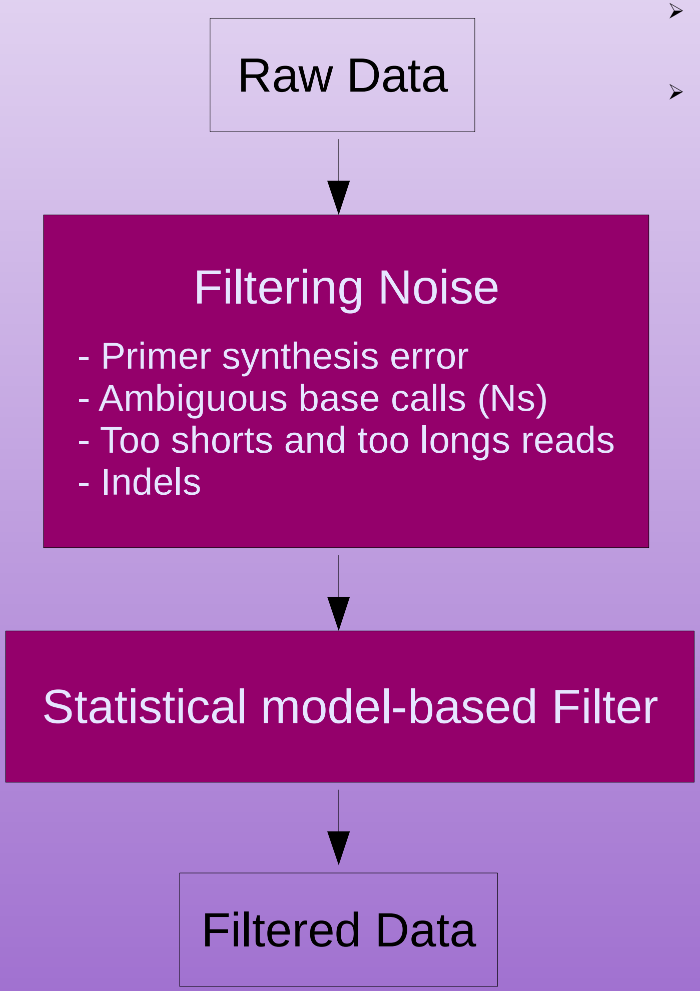
Statistical model-based Filter

Filtered Data

- Two main steps:
  - Filtering low-quality reads more accurate estimations will be obtained.
  - Statistical model for the remaining error: modelling the errors that still remain in our reads.

Assessing the pipeline performance

- To check the performance of the pipeline:
  - Pyrosequencing 3 independent PCR products from a clon
  - As we know the sequence of the clon any change in the reads is considered a process error.



Raw Data

Filtering Noise

- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

Statistical model-based Filter

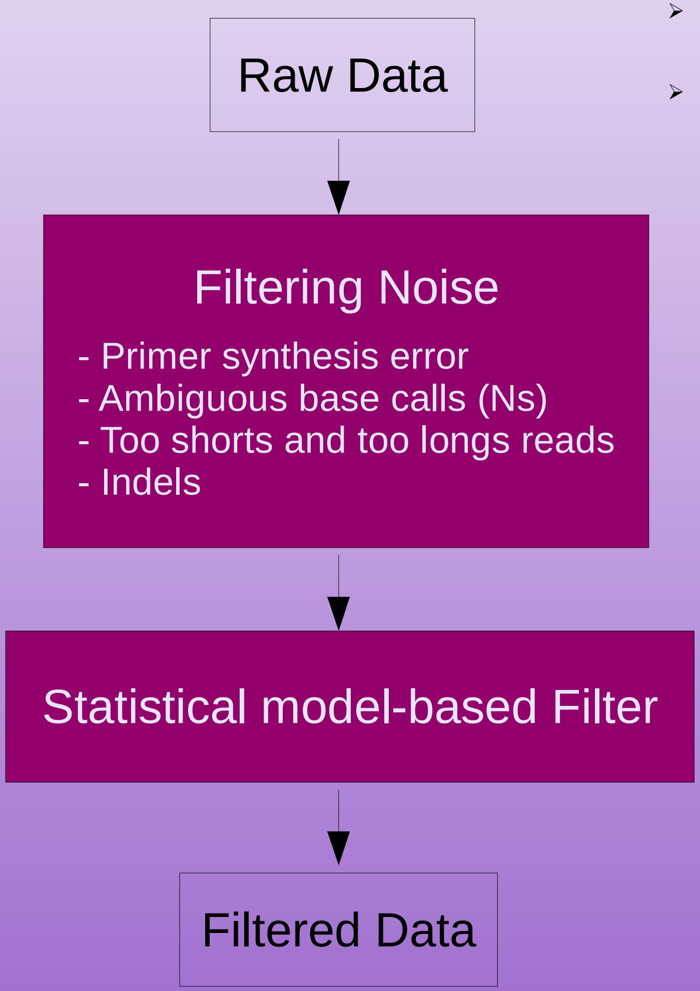
Filtered Data

Nucleotide reference sequence of the clon:

```
5'
GCTGGCCCGCTCCCCAAGGTGCCCGCTCACTGACACCCTGCACTTGCGGCTCCTCGGACCTTTACCTG
GTCACGAGGCACGCCGATGTCATTCCCGTACGCCGGCGGGGTGATGGCAGGGGCAGCCTGCTTTCGCC
CCGGCCCATCTCTTACCTGAAAGGCTCCTCGGGGGGCCCACTGCTGTGCCCCGCGGGACACGCCGTAG
GCATTTTCAGAGCCGCGGTGTGCACCCGTGGAGTGGCTAAAGCGGTGGACTTTATCCCCGTAGAGGGC
CTAGAGACAACCATGAGGTCCCCGGTGTTCTCGGACAATTCCTCC 3'
```

```
5'
GCTGGCCCGCTCCCCAAGGTGCCCGCTCTCTGACACCCTGCACTTCCGGCTCCTCGGACCTTTACCTG
GTCACGAGGCACGCCGATGTCATTCCCGTACGCCGGCGGGGTGATGGCAGGGGCAGCCTGCTTTCGCC
CCGGCCCATCTCTTACCTGAAAGGATCCTCGGGGGGCCCACTGCTGTGCCCCGCGGGACACGCCGTAG
GCATTTTCAGAGCCGCGGTGTGCACCCGTGGAGTGGCTAAAAGCGGTGGACTTTATCCCCGTAGAGGG
CCTAGAGACAACCATGAGGTCCCCGGTGTTCTCGGACAATTCCTCC 3'
```

Assessing the pipeline performance


Raw Data
Filtering Noise
- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels
Statistical model-based Filter
Filtered Data

- To check the performance of the pipeline:
  - Pyrosequencing 3 independent PCR products from a clon
  - As we know the sequence of the clon any change in the reads is considered a process error.

Nucleotide reference sequence of the clon:

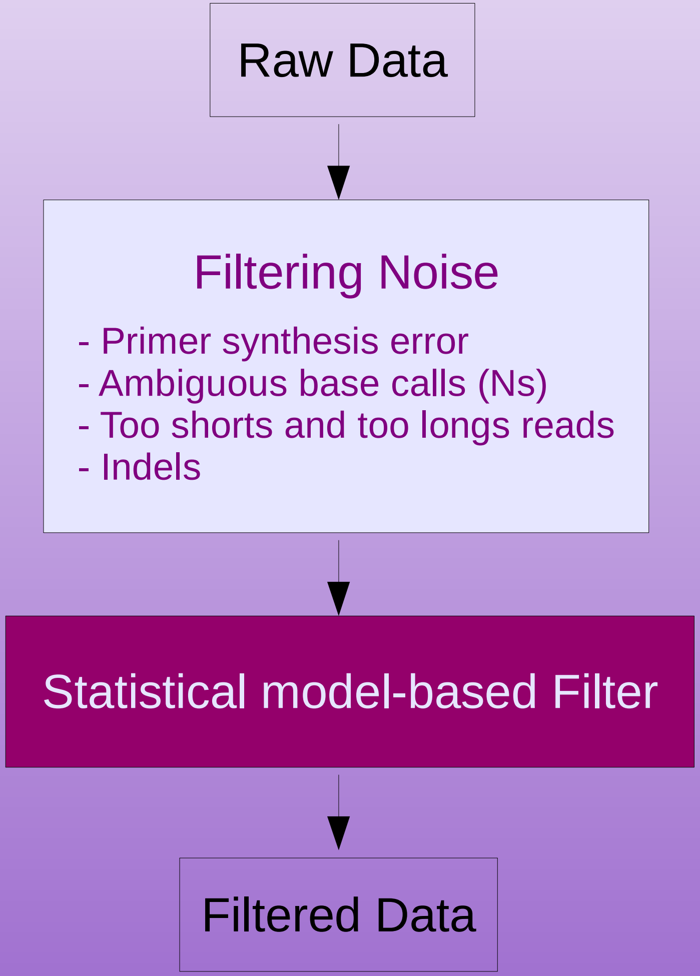
```
5'
GCTGGCCCGCTCCCCAAGGTGCCCGCTCACTGACACCCTGCACTTGCGGCTCCTCGGACCTTTACCTG
GTCACGAGGCACGCCGATGTCATTCCCGTACGCCGGCGGGGTGATGGCAGGGGCAGCCTGCTTTCGCC
CCGGCCCATCTCTTACCTGAAAGGCTCCTCGGGGGGCCCACTGCTGTGCCCCGCGGGACACGCCGTAG
GCATTTTCAGAGCCGCGGTGTGCACCCGTGGAGTGGCTAAAGCGGTGGACTTTATCCCCGTAGAGGGC
CTAGAGACAACCATGAGGTCCCCGGTGTTCTCGGACAATTCCTCC 3'
```

```
5'
GCTGGCCCGCTCCCCAAGGTGCCCGCTCTCTGAC          ACTTCCGGCTCCTCGGACCTTTACCTG
GTCACGAGGCACGCCGATGTCATTCCCGTAC           GGGTGATGGCAGGGGCAGCCTGCTTTCGCC
CCGGCCCATCTCTTACCTGAAAGGATC           GCCCACTGCTGTGCCCCGCGGACACGCCGTAG
GCATTTTCAGAGCCGCGGTGTGCA           AGTGGCTAAAAGCGGTGGACTTTATCCCCGTAGAGGG
CCTAGAGACAACCATGAGGTC           TTCTCGGACAATTCCTCC 3'
```

Process errors

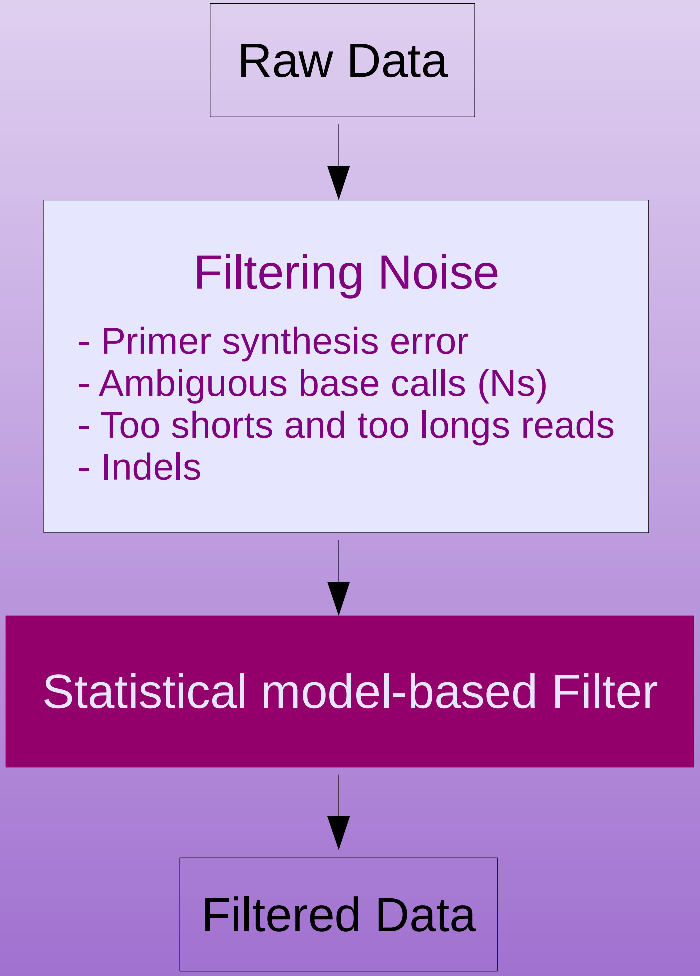
Filtering noise

- Remove reads with more than 25% of primer synthesis errors


Raw Data
Filtering Noise
- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels
Statistical model-based Filter
Filtered Data

Filtering noise

- Remove reads with more than 25% of primer synthesis errors
- Remove reads with ambiguous base calls (N)



Raw Data

Filtering Noise

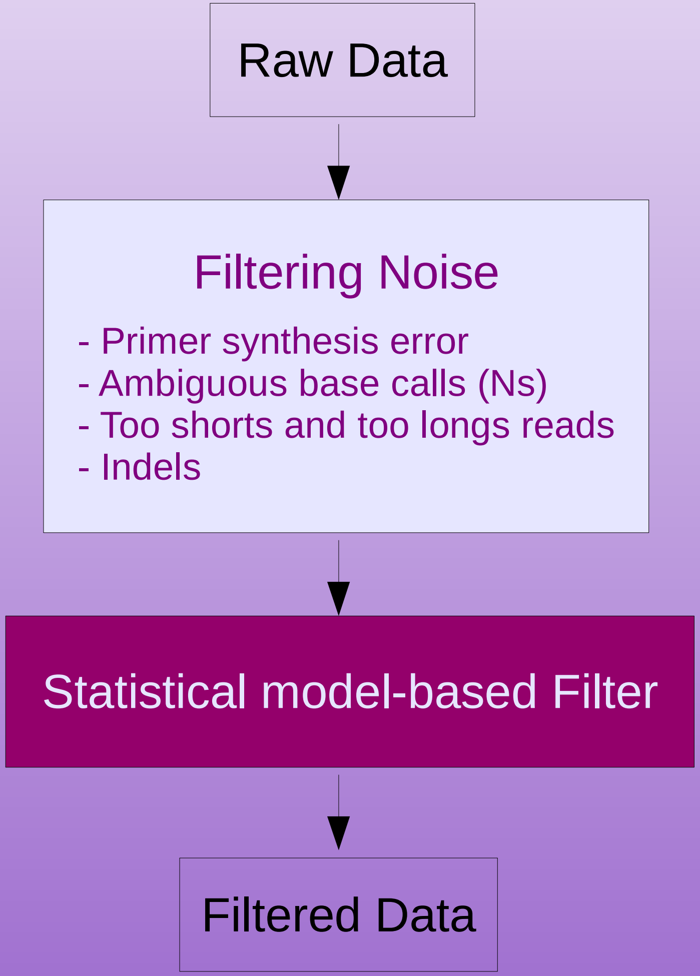
- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

Statistical model-based Filter

Filtered Data

Filtering noise

- Remove reads with more than 25% of primer synthesis errors
- Remove reads with ambiguous base calls (N)



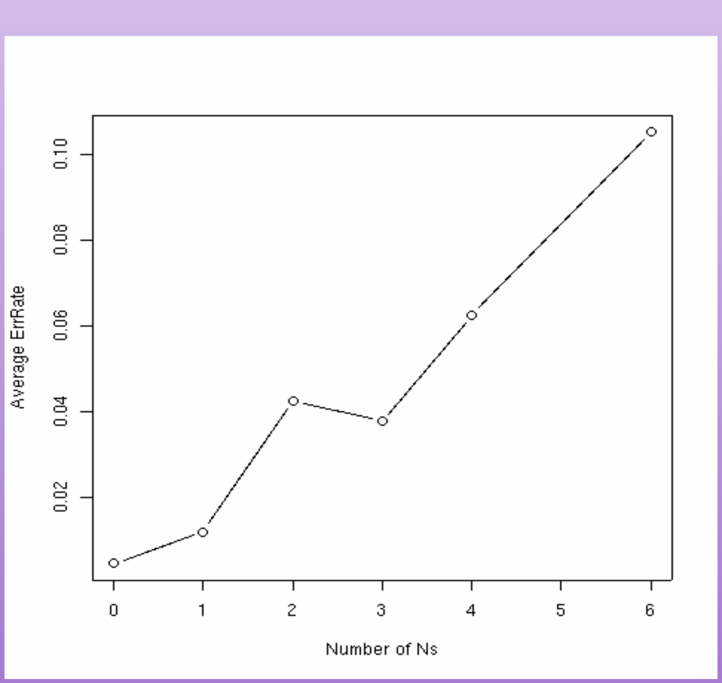
Raw Data

Filtering Noise

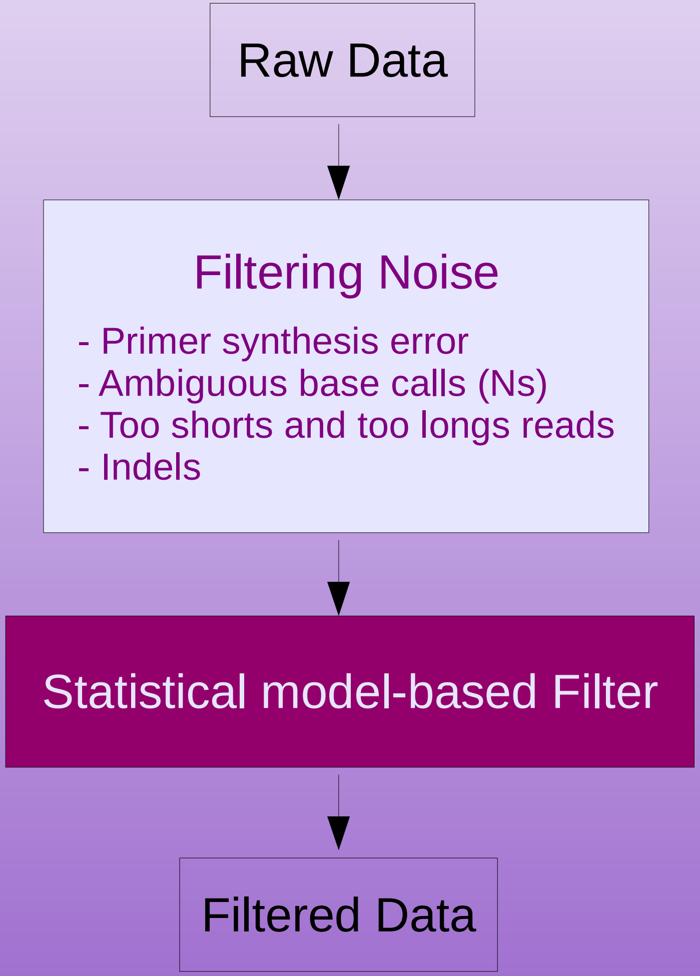
- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

Statistical model-based Filter

Filtered Data



Filtering noise



Raw Data

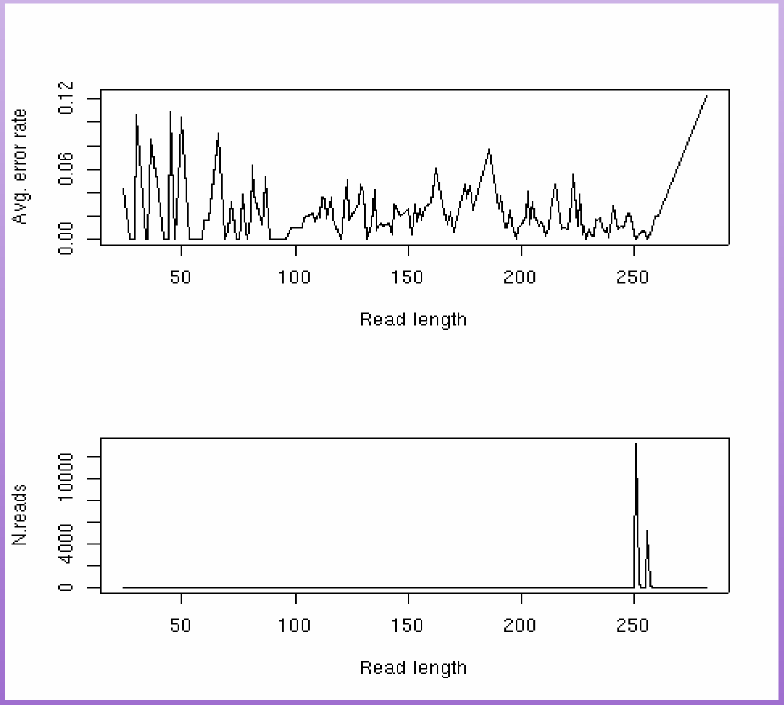
Filtering Noise

- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

Statistical model-based Filter

Filtered Data

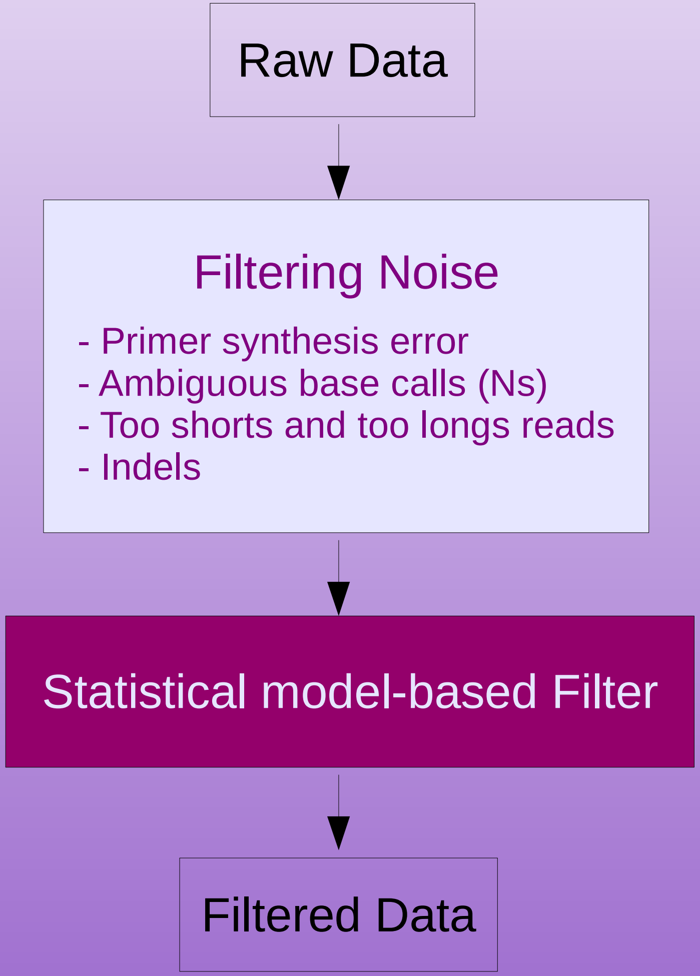
- Remove reads with more than 25% of primer synthesis errors
- Remove reads with ambiguous base calls (N)
- Remove short reads:



Vall d'Hebron
Institut de Recerca

UNITAT
D'ESTADÍSTICA I
BIOINFORMÀTICA

Filtering noise



Raw Data

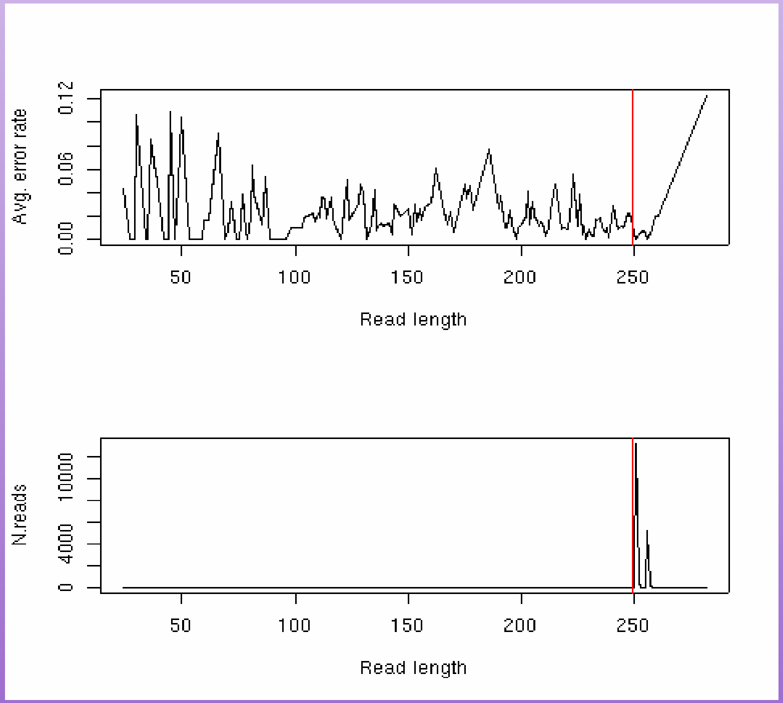
Filtering Noise

- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

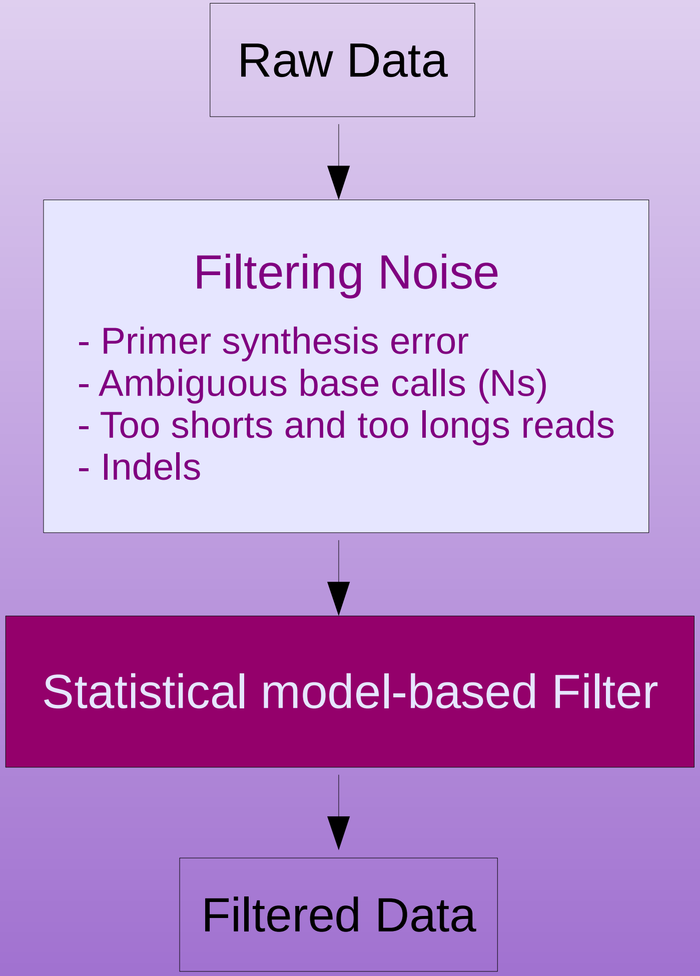
Statistical model-based Filter

Filtered Data

- Remove reads with more than 25% of primer synthesis errors
- Remove reads with ambiguous base calls (N)
- Remove short reads:



Filtering noise



Raw Data

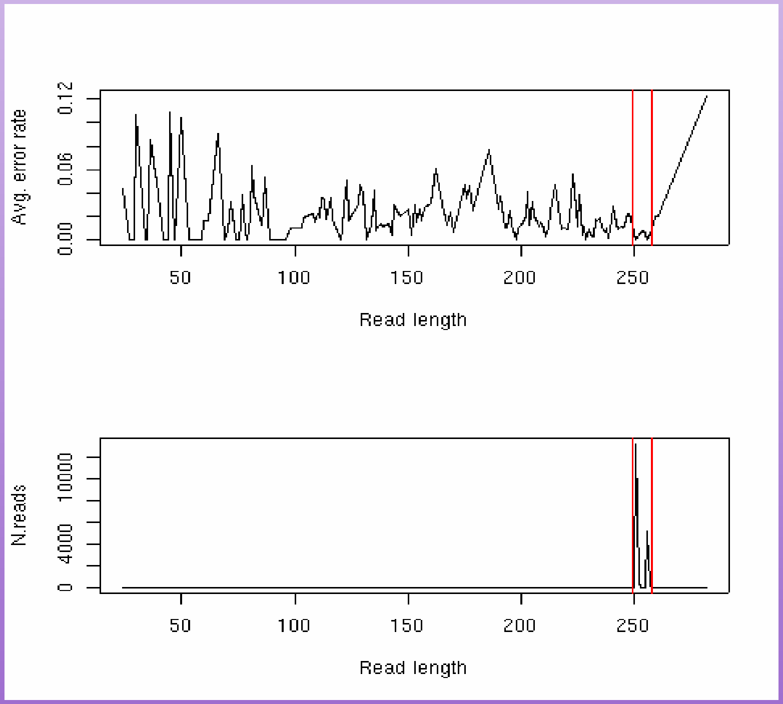
Filtering Noise

- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

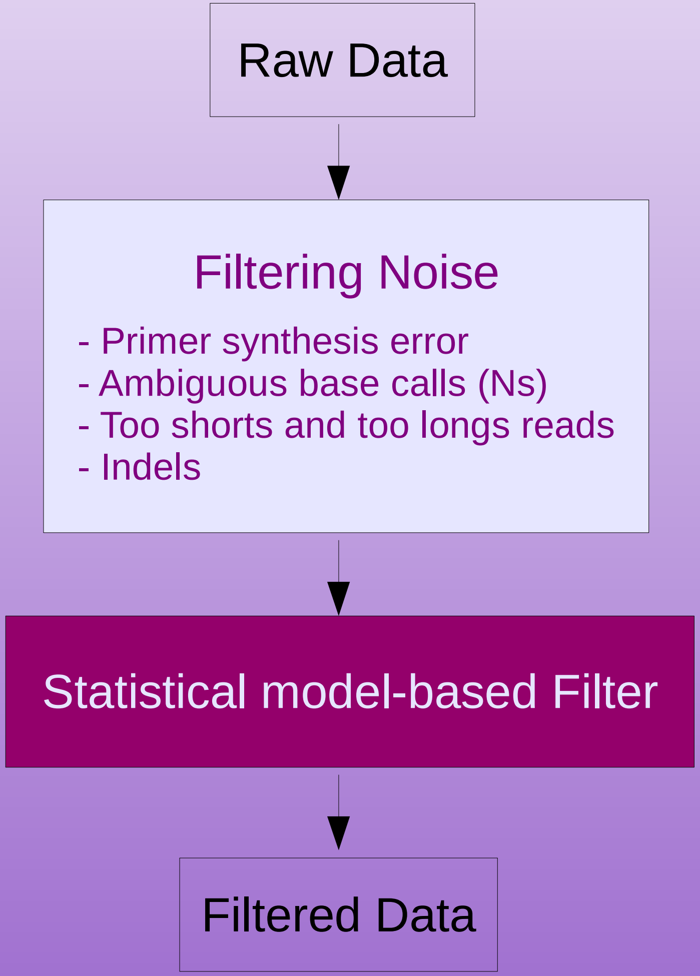
Statistical model-based Filter

Filtered Data

- Remove reads with more than 25% of primer synthesis errors
- Remove reads with ambiguous base calls (N)
- Remove short reads and trim too long ones:



Filtering noise



Raw Data

Filtering Noise

- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

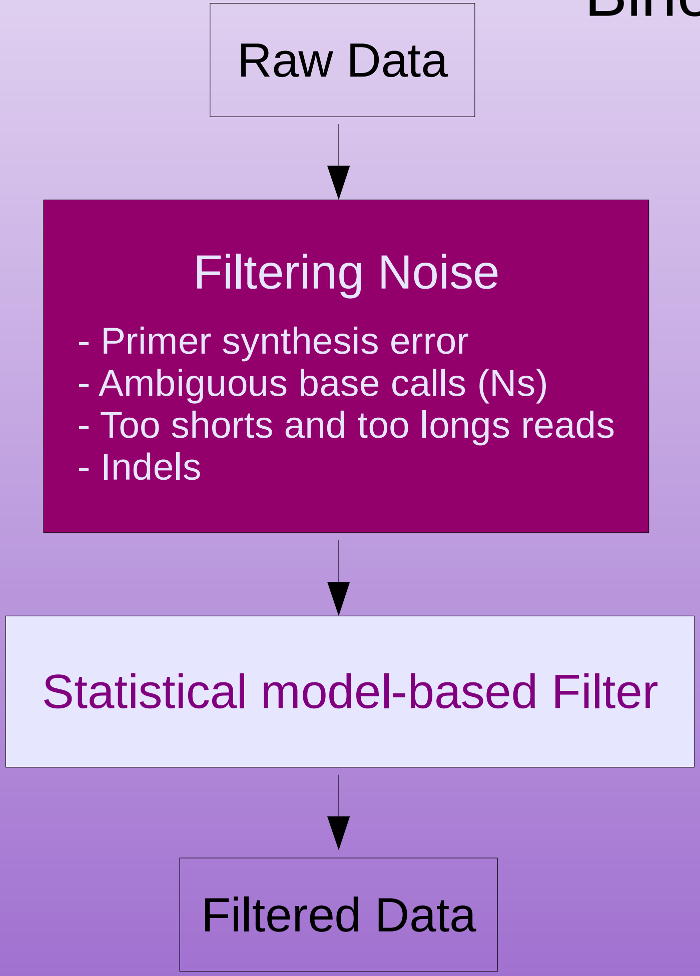
Statistical model-based Filter

Filtered Data

- Remove reads with more than 25% of primer synthesis errors
- Remove reads with ambiguous base calls (N)
- Remove short reads and trim too long ones
- Remove reads containing indels

- **(Huse et al. *Gen. Biol.* 2007)**

Modelling the remaining error



Raw Data

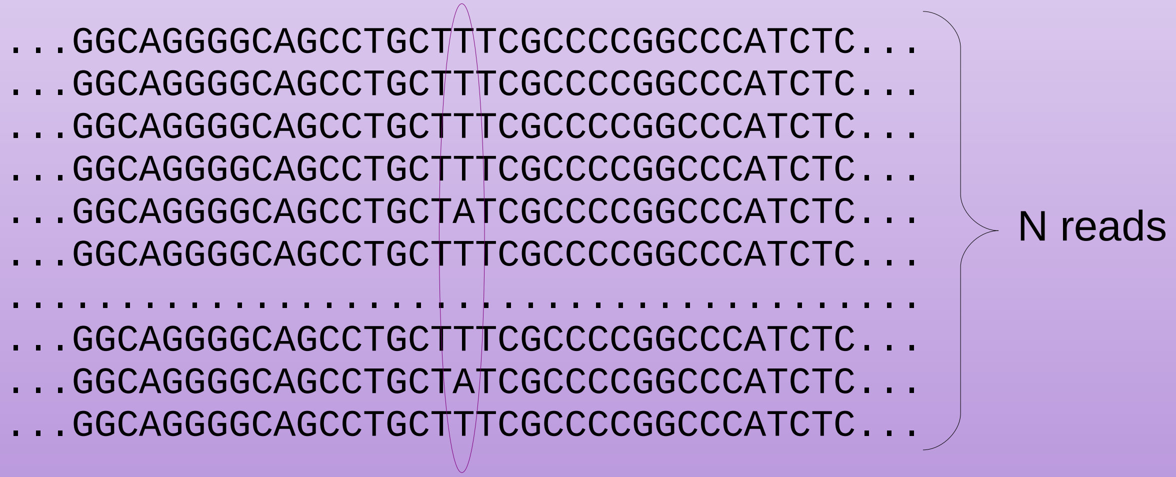
Filtering Noise

- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

Statistical model-based Filter

Filtered Data

- Mismatch errors per site can be modeled with a Binomial distribution



```
...GGCAGGGGCAGCCTGCTTTCGCCCCGGCCCATCTC...
...GGCAGGGGCAGCCTGCTTTCGCCCCGGCCCATCTC...
...GGCAGGGGCAGCCTGCTTTCGCCCCGGCCCATCTC...
...GGCAGGGGCAGCCTGCTTTCGCCCCGGCCCATCTC...
...GGCAGGGGCAGCCTGCTATCGCCCCGGCCCATCTC...
...GGCAGGGGCAGCCTGCTTTCGCCCCGGCCCATCTC...
.........................................
...GGCAGGGGCAGCCTGCTTTCGCCCCGGCCCATCTC...
...GGCAGGGGCAGCCTGCTATCGCCCCGGCCCATCTC...
...GGCAGGGGCAGCCTGCTTTCGCCCCGGCCCATCTC...
```

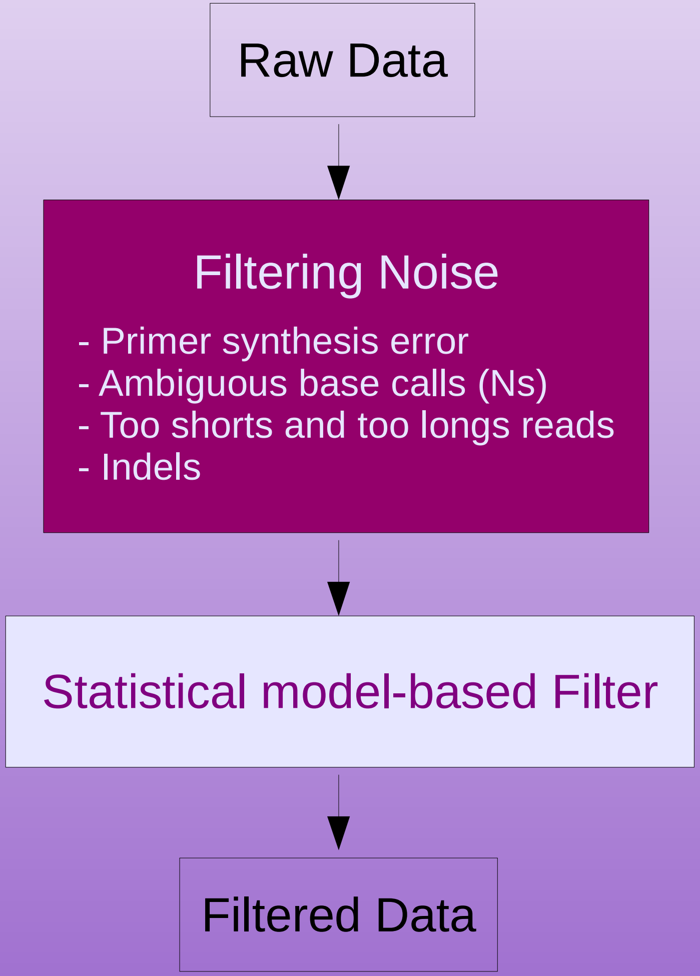
N reads

$$\mathrm{Bin}(N,p) \approx \mathrm{Poiss}(Np)\ (N \to \infty,\ p \to 0)$$

where p = prob{mismatch},

N = number of reads

Modelling the remaining error

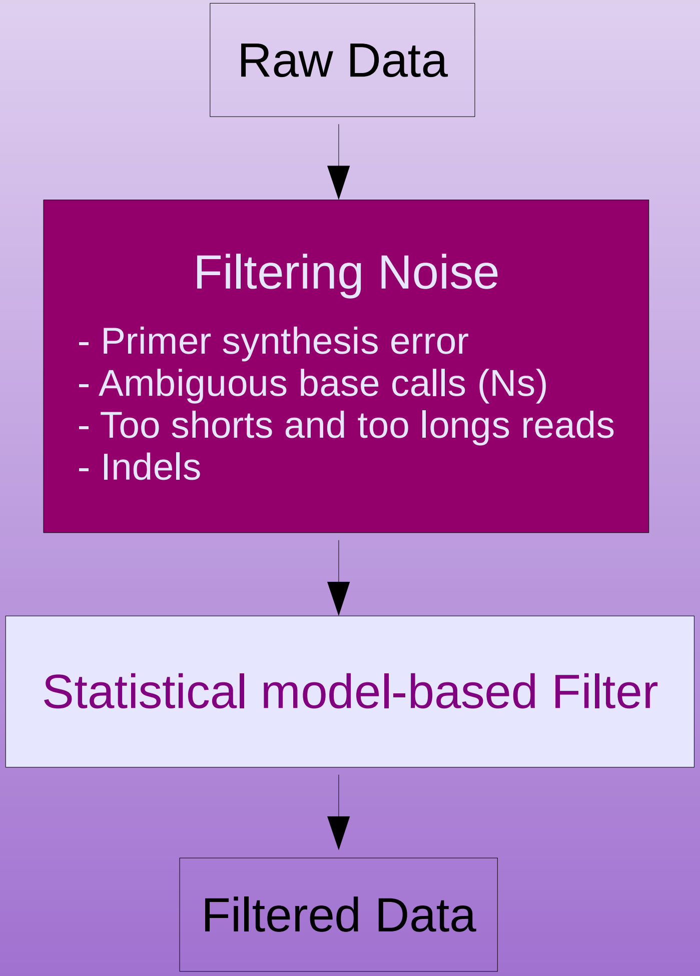

Raw Data
Filtering Noise
- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels
Statistical model-based Filter
Filtered Data

- Modelling error: **(Wang et al. *Gen. Res.* 2007)**
  - Mismatch errors ~ Poiss(λ_r), r in {“h”, “nh”}
    - For a variant observed *n* times: the probability of getting it >= *n* times if it was an error is given by:

$$P = 1 - \sum_{i=1}^{n-1} \frac{e^{-\lambda} \cdot \lambda^i}{i!}$$

$$\lambda = N \cdot \lambda^r,\ r \in \{h, nh\}$$

Modelling the remaining error



Raw Data

Filtering Noise

- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

Statistical model-based Filter

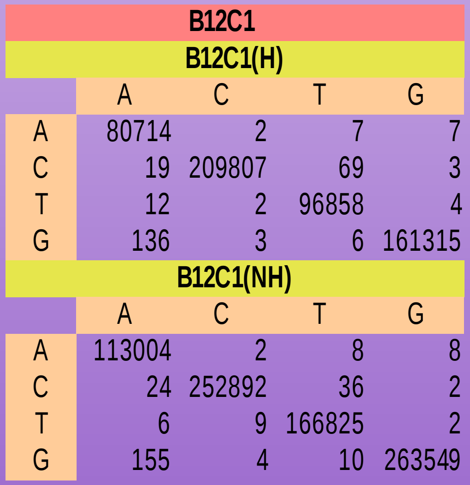
Filtered Data

- For a variant observed *n* times: the probability of getting it >= *n* times if it was an error is given by:

$$P = 1 - \sum_{i=1}^{n-1} \frac{e^{-\lambda} \cdot \lambda^i}{i!}$$

$$\lambda = N \cdot \lambda_{ij}^{r},\ r \in \{h, nh\},\ i,j \in \{A, C, T, G\}$$

- Matrices of mismatch counts and error rates:



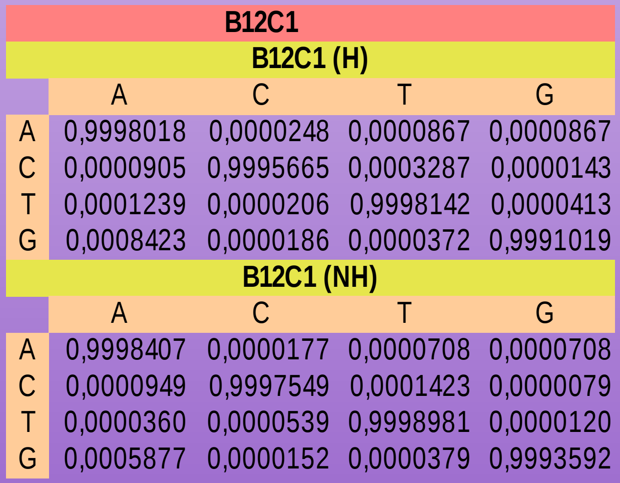
B12C1

B12C1(H)

	A	C	T	G
A	80714	2	7	7
C	19	209807	69	3
T	12	2	96858	4
G	136	3	6	161315

B12C1(NH)

	A	C	T	G
A	113004	2	8	8
C	24	252892	36	2
T	6	9	166825	2
G	155	4	10	263549



B12C1

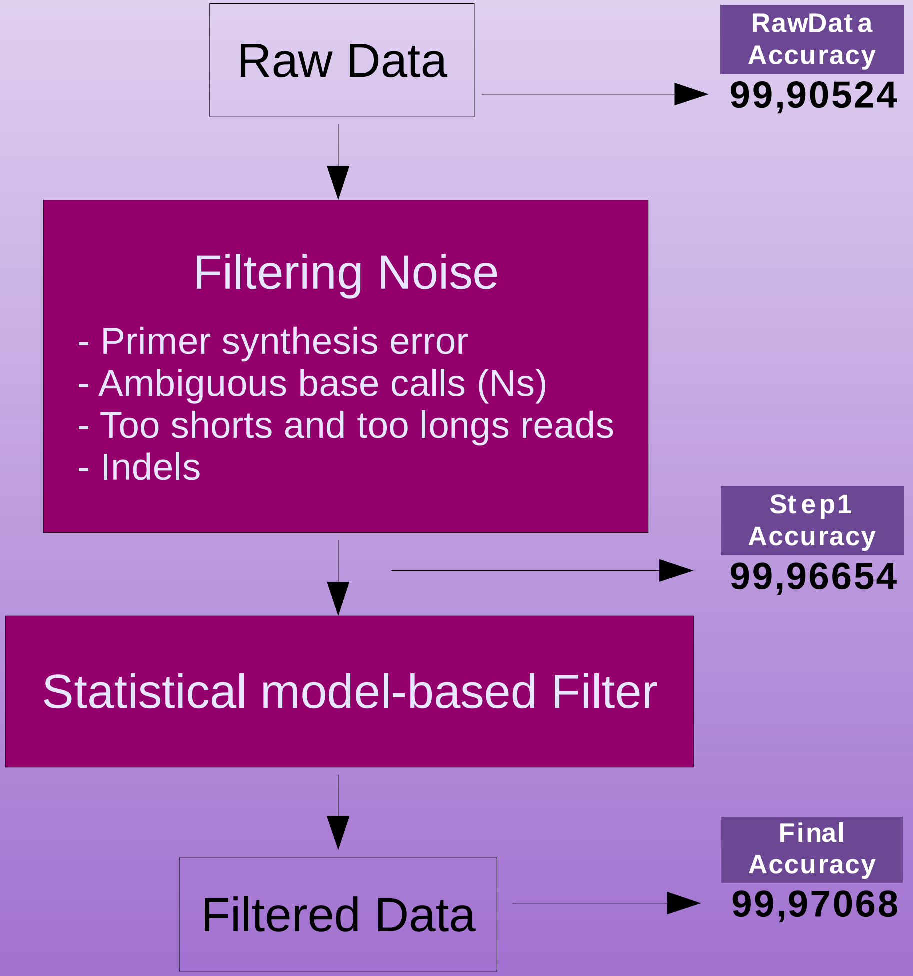
B12C1 (H)

	A	C	T	G
A	0,9998018	0,0000248	0,0000867	0,0000867
C	0,0000905	0,9995665	0,0003287	0,0000143
T	0,0001239	0,0000206	0,9998142	0,0000413
G	0,0008423	0,0000186	0,0000372	0,9991019

B12C1 (NH)

	A	C	T	G
A	0,9998407	0,0000177	0,0000708	0,0000708
C	0,0000949	0,9997549	0,0001423	0,0000079
T	0,0000360	0,0000539	0,9998981	0,0000120
G	0,0005877	0,0000152	0,0000379	0,9993592

Results and conclusions



Raw Data

RawData Accuracy
99,90524

Filtering Noise

- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

Step1 Accuracy
99,96654

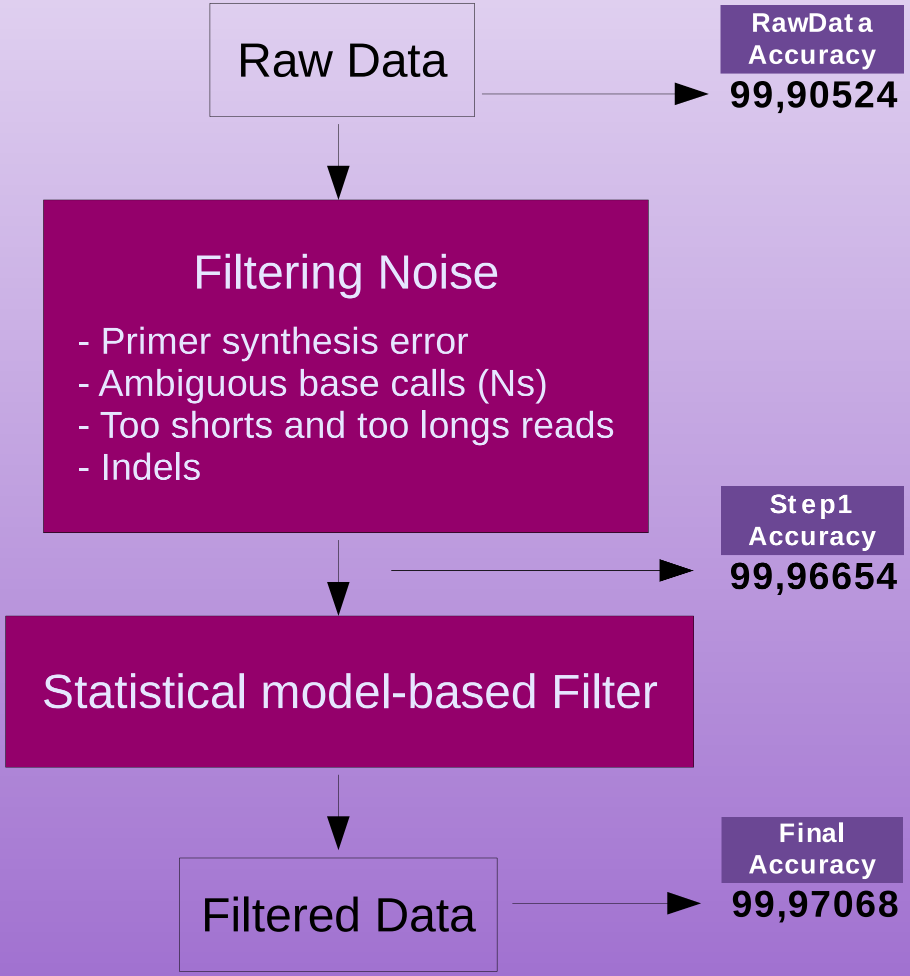
Statistical model-based Filter

Filtered Data

Final Accuracy
99,97068

- Filtering noise before computing estimations of proportions is a useful approach

Results and conclusions



Raw Data

RawData Accuracy
99,90524

Filtering Noise

- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

Step1 Accuracy
99,96654

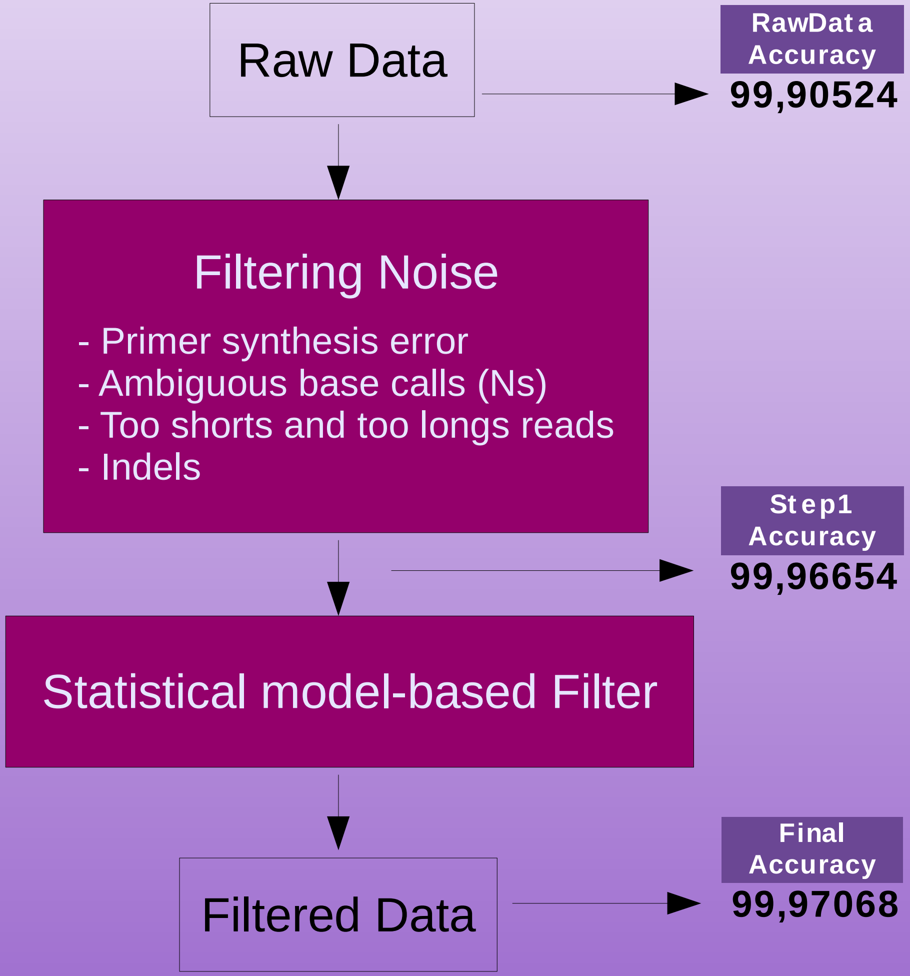
Statistical model-based Filter

Filtered Data

Final Accuracy
99,97068

- Filtering noise before computing estimations of proportions is a useful approach
- The statistical model:
  - improves slightly the accuracy
  - It provides a probabilistic score per base

Results and conclusions



Raw Data

RawData Accuracy
99,90524

Filtering Noise

- Primer synthesis error
- Ambiguous base calls (Ns)
- Too shorts and too longs reads
- Indels

Step1 Accuracy
99,96654

Statistical model-based Filter

Filtered Data

Final Accuracy
99,97068

- Filtering noise before computing estimations of proportions is a useful approach
- The statistical model:
  - improves slightly the accuracy
  - It provides a probabilistic score per base
- We still have to fine tune some points but we think that this pipeline could be useful in detecting minor variants.

Acknowledgements

- Alex Sánchez, Josep Lluis Mosquera, Alejandro Artacho

(Unitat Estadística i Bioinformàtica, Institut de Recerca HUVH)

- Fátima Nuñez, Paqui Gallego, and the rest of colleagues of the UCTS

(Unitat Científico-Tècnica de Suport, Institut de Recerca HUVH)

- Josep Quer, María Cubero, María Homs, Paco Rodríguez...

(Malalties Hepàtiques, Institut de Recerca HUVH)

Thanks for your attention!