Quasi-species Identification by Model Based Clustering

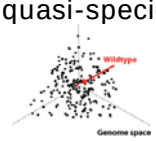
Massively parallel sequencing

Bie Verbist
Ghent University – Janssen

- Introduction
- Goal
- Massively parallel sequencing
  - Technology
  - Quality scores
- Model-based clustering
- Results

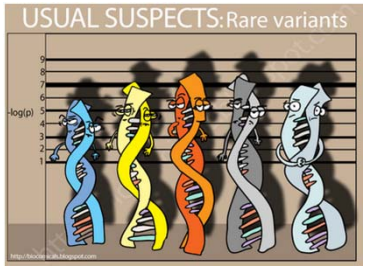
Introduction

- Identify and quantify quasi-species.



- Focus on low-frequency variants
- Deep sequencing: in-depth characterization of sequence variation.

Goal



Massively Parallel Sequencing

- illumina

Massively Parallel Sequencing

- Sequencing read:

```
@HWUSI-EAS1524:12:FC:1:1:18845:1091 1:N:0:
ATGACCCATCAAAAGACTTAATAGCAGAAATACAGAAGCAGGGGCAAGGCC
+
IIIIIIIIIIIIIIIIHIIIIIIIIIIIIIIIIIIIHIIHIHIHFGHF
```

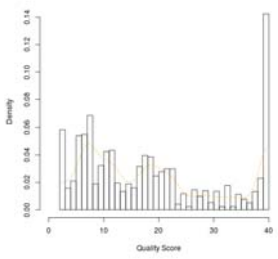
- ASCII-coded:

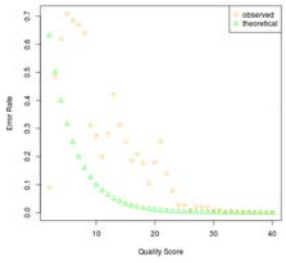
"#$%&'()*+,./0123456789:;<=>?@ABCDEFGHI $Q_{il} \in [2,40]$

- $Q_{il} = -10 \log_{10}(p_{il})$ with p_{il} error probability

Massively Parallel Sequencing

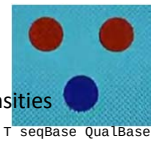
- Distribution of Q_{il} at error positions
- Q_{il} doesn't reflect true error probabilities





Massively Parallel Sequencing

- Other metrics:



– Corrected and censored raw intensities

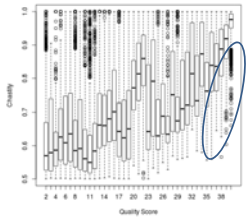
```
lane tile cycle     x     y    A   C   G   T seqBase QualBase
   1    1     1  4797 11926 2017   2   0   0      A        I
   1    1     1  2727 14009 1114   573 450 443    A        I
   1    1     1  2430 16718   42   73  0    10    C        :
```

– Second best base call

```
    First Read (R) @HWUSI-EAS1524:12:FC:1:1:4797:11926
ATCTGCTCCTGTATCTAATAGAGCTTCCTTTAGTTGCCCCCCTATCTTTAT
    Second Read (S) @HWUSI-EAS1524:12:0:1:1:17484:13206
CAACAGCTTCTGTATCTTATAGAGCCTTCACCTCCTATTTTAGCCAACCCC
```

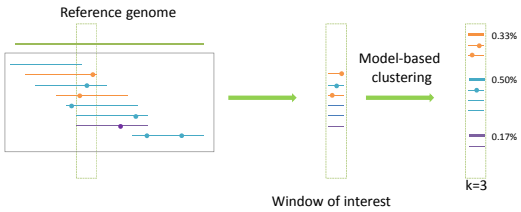
Massively Parallel Sequencing

– Chastity (C_{il}):
max. intensity / sum of the 2 highest intensities



High Quality Score
Low Chastity

Model-Based Clustering



- Clusters = haplotypes,
  – missing data problem: EM-algorithm

Model-Based Clustering

- Notation:
  – $\mathbf{r}_i$: best base calls of read *i* (*i*=1 ... n)
  – $\mathbf{s}_i$: second best base calls of read *i* (*i*=1 ... n)
  – z_{ij}: z_{ij}=1 when read i belongs to haplotype *j* (*j*=1...k)
  – τ_j: probability to belong to haplotype *j*

- Complete Data Likelihood:

$$L = \prod_{i=1}^{n}\prod_{j=1}^{k}\left(f_j(\mathbf{r}_i,\mathbf{s}_i)\tau_j\right)^{z_{ij}} \quad with\ f_j(\mathbf{r},\mathbf{s}) = Prob(\mathbf{r},\mathbf{s}|\text{haplotype } j)$$

Model-Based Clustering

$$L = \prod_{i=1}^{n}\prod_{j=1}^{k}\left(f_j(\mathbf{r}_i,\mathbf{s}_i)\tau_j\right)^{z_{ij}}$$

$$with\ f_j(\mathbf{r}_i,\mathbf{s}_i) = \prod_{l=1}^{m}\theta_{ril}^{I(r_{il}=h_{jl})}\theta_{sil}^{I(s_{il}=h_{jl})}\theta_{oil}^{\left(1-I(r_{il}=h_{jl})\right)\left(1-I(s_{il}=h_{jl})\right)}$$

$$Prob(\text{base } l \text{ of read } i = \text{base } l \text{ of haplotype } j) = \begin{cases}\theta_{ril} & base\\ \theta_{sil} & second\ base\end{cases}$$

$$\theta_{oil} = 1 - \theta_{ril} - \theta_{sil}$$

- Model is overidentified

Model-Based Clustering

- Model θ parameters:

$$log\frac{\theta_{ril}}{\theta_{oil}} = \beta_{0r} + \beta_{1r}Q_{il} + \beta_{2r}Q_{i(l-1)} + \beta_{3r}C_{il} + \beta_{4r}H_{il}$$

$$log\frac{\theta_{sil}}{\theta_{oil}} = \beta_{0s} + \beta_{1s}Q_{il} + \beta_{2s}Q_{i(l-1)} + \beta_{3s}C_{il} + \beta_{4r}H_{il}$$

  - Complexity reduced to 10 β-parameters
  - Likelihood still depends on haplotype membership

$$L = \prod_{i=1}^{n}\prod_{j=1}^{k}\left(f_j(\mathbf{r}_i, \mathbf{s}_i)\tau_j\right)^{z_{ij}}$$

Model-Based Clustering

- EM algorithm
  - E step: update posterior probability

$$\widehat{z_{ij}} = Prob(\text{read } i \text{ is of haplotype } j | \text{observed data})$$

  - M step: update β parameters by maximizing expected likelihood

till convergence

Model-Based Clustering

End result

- After conversion
  - $\widehat{z_{ij}} = Prob(\text{read } i \text{ is of haplotype } j | \text{observed data})$
  - haplotype frequency τ_j

$$\hat{\tau}_j = \frac{\sum_{i=1}^{n}\widehat{z_{ij}}}{n}$$

Model-Based Clustering

- Benchmark dataset
  - 5 HIV-1 plasmids

SDM090122	SDM090126	SDM090141	SDM090147	SDM090151
0,63	0,30	0,050	0,014	0,0046

  - 10 mutation positions

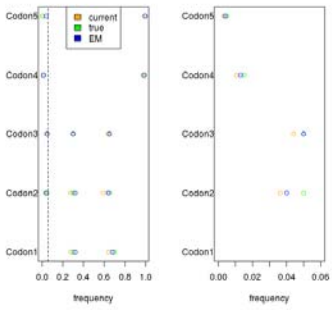
	Codon1		Codon2			Codon3		Codon4		Codon5
Pos	1	3	8	9	199	241	242	826	828	844
Ref	A	A	A	A	A	T	A	A	C	A
	0,6988	0,6988	0,6988	0,9498	0,37	0,9498	0,6486	0,986	0,986	0,9954
Variant	C	C	G	T	G	G	G	T	A	G
	0,3012	0,3012	0,3012	0,0502	0,63	0,0502	0,3514	0,014	0,014	0,0046

Results

	Codon1		Codon2		Codon3		Codon4		Codon5
	1	3	8	9	241	242	826	828	844
	A	A	A	A	T	A	A	C	A
	0,70	0,70	0,70	0,95	0,95	0,65	0,986	0,986	0,995
	C	C	G	T	G	G	T	A	G
	0,30	0,30	0,30	0,05	0,05	0,35	0,014	0,014	0,005

$\hat{\tau}_j$	Codon 1	Codon 2	Codon 3	Codon 4	Codon 5
Reference	AAA 68.4%	AAA 64.2%	TAT 64.5%	ACC 98.7%	ATA 99.6%
Haplotype 1	CAC 31.6%	AGA 31.8%	TGT 30.5%	TCA 1.3%	GTA 0.4%
Haplotype 2		AAT 4.0%	GGT 5.0%		

Results

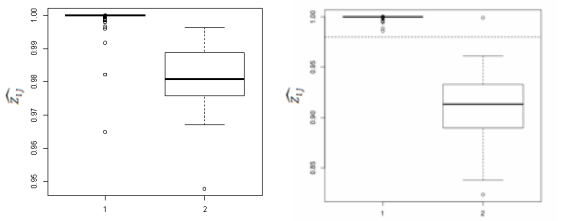


Underestimation of the low frequencies due to loss of coverage

Results

Rare variant at 0.5%

Error position of 0.5%



Distinction between error and true low frequency variant is not clear yet

Conclusion

- Model-Based Clustering approach to identify and quantify quasi-species.
- Benchmark data:
  - All haplotypes are discovered
  - Slighlty underestimation of frequency
  - Error variants at same frequency have lower posterior probabilities however distinction is not yet clear.

Conclusion

- Future research:
  - Check behaviour in error prone region of Illumina
  - Relax assumption number of haplotypes j is known
  - Allow for missing data in window
  - Model optimization at each codon position

Acknowledgement

- Promoters: *Prof.Dr. O. Thas[1] and Dr.L.Bijnens[2]*
- Translational genomics team
- Yves Wetzels, Tobias Verbeke, Joris Meys[1]

BACK-UP

Corrected raw intensities

- Corrected for
  - signal amplitude: due to inequality of fluor emission intensities
  - spectral cross talk: overlap of emission frequency correction using the calculated matrix file
  - phasing and pre-phasing values: estimates of the proportion of molecules out of sync - behind or ahead

- Split in forward/reverse data frame
  - Perform analysis twice – same results??

Start position codon		
In algorithm (Codon level)	In excel list Koen (Ncl level)	In slides
598	2850	1
604	2856	7
838	3090	241
1423	3675	826
1441	3693	844
1258 (error pos)		

FEEDBACK CONFERENCE