Problem statement for MAQ: genotyping/SNP calling

Genotyping = figuring out collection of alleles in an individual

SNP = single nucleotide polymorphism
Mapping short DNA sequencing reads and calling variants using mapping quality scores

Heng Li,¹ Jue Ruan,² and Richard Durbin¹,³

¹The Wellcome Trust Sanger Institute, Hinxton CB10 1SA, United Kingdom; ²Beijing Genomics Institute, Chinese Academy of Science, Beijing 100029, China
**Input:** Reference genome

**Input:** short reads (~50bp) w/ per-base quality scores

Mapping/alignment

Mapped reads

Calculate mapping quality

Mapped reads with per-base quality

Highest posterior probability SNP calling
Input: short reads (~50bp) w/ per-base quality scores

Calculate mapping quality

Mapped reads with per-base mapping quality

Highest posterior probability SNP calling

Input: Reference genome

Mapping/alignment

Mapped reads
"Fingerprint hashing" - match nonconsecutive letters instead of consecutive letters

ATATGTGA      read
11110000  ->  ATATnnnn
00001111  ->  nnnnGTGA
10101010  ->  AnAnGnGn
01010101  ->  nTnTnTnA

Hashes guarantee catching positions with 1 mismatch and a certain probability of catching strings with 2 mismatches
Mapping/alignment

If degenerate hits are found, randomly choose one

ATATGTGA       read
11110000   ->   ATATnnnn
00001111   ->   nnnnGTGA
10101010   ->   AnAnGnGn
01010101   ->   nTnTnTnA

Scan the reference genome against these hashes; use hash matches as seeds locations for alignment

Methods > Single end read mapping

pg 6
Input: Reference genome

Input: short reads (~50bp) w/ per-base quality scores

Mapping/alignment

Mapped reads

Calculate mapping quality

Mapped reads with per-base mapping quality

Highest posterior probability SNP calling
Mapping Quality

\[ Q_s = "phred\text{-scaled probability that the true alignment is not the one found by MAQ."} \]

- evaluate alignment quality before calling SNPs
- \( phred \)-scaled
- degeneracy in mapping of a read \( \rightarrow \) MAQ score = 0
  - avoids variant calling for repetitive sites

During SNP calling stage, infer genotype with highest posterior distribution and assign \( phred \) quality score. Call SNPs by comparing inferred genotype to reference genome.

Methods > Single end mapping qualities
pg 6, Supp. pgs 1-3
Mapping Quality

Three types of errors (error probabilities $\epsilon_{M1} - \epsilon_{M3}$):

1. Read does not come from reference
   a. source = contamination
      i. assumed to be negligible $\Rightarrow \epsilon_{M1} = 0$

2. True position is missed by alignment program
   i. heuristic algorithm does not guarantee $\epsilon_{M2} = 0$
      1. tradeoff between speed and accuracy

3. "Best" hit is not the true read locus, $\epsilon_{M3}$
   a. Incorrectly mapped because of mismatches / indels
Derivation: Board Work

• General quality score formula:
  \[ Q_B = -c \log(\epsilon_B) \]
  - \[ \epsilon_B = \frac{\text{true b}}{p(\hat{B} \neq B)} \]  
    (observed base in read ≠ true base)
  - Phred quality requires change of base
    - \[ c = \frac{10}{\log(10)} \approx 4.343 \]

• Equation 1 (for observed read \( \hat{Z} \)):
  \[ P\{ \hat{Z} \mid \text{true } Z = b_1b_2\ldots b_l \} = \prod_{i=1}^{l} p(\hat{b}_i \mid b_i) \]
  - Assumption: site independence

Supplementary Text > Single-End Mapping Errors
Supp. pgs 1-3
Input: short reads (~50bp) w/ per-base quality scores

Input: Reference genome

Mapping/alignment

Calculate mapping quality

Mapped reads with per-base mapping quality

Highest posterior probability SNP calling
Genotype calling

First, per-base qualities are redefined as $\min(\text{read quality}, \text{mapping quality})$.

Methods > Consensus Genotype Calling
pg 7, Supp. pgs 4-6
Genotype calling

? | ? | genome

? | ? | reads w/ per base score

________ A _
    __ G _____
    ____ A ___
    --- A ______

Pr(<A, A>| data)
Pr(<G, G>| data)
Pr(<A, G>| data)

choose genotype with maximum posterior probability

Methods > Consensus Genotype Calling
pg 7, Supp. pgs 4-6
Genotype calling - correlated errors

\[ \alpha_{nk} = \Pr\{\text{exactly } k \text{ errors in } n \text{ bases}\} \]

\[ \beta_{nk} = \begin{cases} 
\Pr(\text{more than } k \text{ errors | more than } k-1 \text{ errors}) & (k>0) \\
\text{or} \\
\Pr(\text{more than 0 errors in } n \text{ bases}) & (k=0) 
\end{cases} \]
Comparison
more paranoid than alternative calling strategies

Results > Figure 2
Practical considerations

Ubuntu: sudo apt-get install maq
Requires two inputs: reference fasta and fastq reads

Splits up reads into batches of ~2-5 million to limit memory usage
|   | V1       | V2   | V3   | V4   | V5   | V6   | V7   | V8   | V9   | V10  | V11  | V12  | V13  | V14  | V15  | V16   |
|---|---------|------|------|------|------|------|------|------|------|------|------|------|------|------|-------|
| 1 | ERR000436.798006 | chrX | 4401731 | -    | -136 | 18   | 99   | 70   | 70   | 0    | 0    | 1    | 0    | 30   |       |
| 2 | ERR000439.3234397 | chrX | 4401731 | +    | 72   | 18   | 99   | 43   | 43   | 1    | 5    | 0    | 1    | 30   |       |
| 3 | ERR000411.2382028 | chrX | 4401731 | +    | 104  | 18   | 99   | 74   | 74   | 0    | 0    | 1    | 0    | 35   |       |
| 4 | ERR000439.2805081 | chrX | 4401733 | +    | 174  | 18   | 93   | 79   | 14   | 0    | 0    | 1    | 0    | 32   |       |
| 5 | ERR000439.2805081 | chrX | 4401877 | -    | -174 | 18   | 93   | 14   | 14   | 0    | 0    | 1    | 2    | 30   |       |

| V1: read id |
| V2: chromosome |
| V3: position |
| V4: strand |
| V5: insert size (outer distance) |
| V6: flag |
| V7: mapping quality |

| V8: single end mapping quality |
| V9: alternative mapping quality |
| V10: number of mismatches of the best hit |
| V11: sum of qualities of mismatched bases of the best hit |
| V12: number of 0-mismatch hits of the first 24bp |
| V13: number of 1-mismatch hits of the first 24bp |

Break
Background on HMM

• Hidden Markov Models may take different forms, depending on the application
  o e.g. GENSCAN HMM for eukaryotic gene finding:
• Profile HMM is useful for specific applications
  o Domain identification
  o SNP discovery
    ▪ We will discuss this today

Background on Profile HMMs

- Common for building a model of a family of sequences, with some implicit alignment
  - Family "profile" can be aligned to query sequences
    - Pfam DB\(^1\), protein family / domain identification:

- Follow same algorithms as normal HMMs
  - Forward and Backward dynamic programming algs.

- Alignment facilitated by HMM structure
  - HMM states: Start, Insertion, Deletion, Match, End
    - "strongly linear, left-right models"\(^2\)

---

1. [http://pfam.sanger.ac.uk](http://pfam.sanger.ac.uk)
Profile HMM Structure

• Only specific transitions can be made
  ○ e.g. \( P(Q_{t+1} = I_2 \mid Q_t = I_1) = 0 \)

• Notion of "emission" must be redefined
  ○ Character emissions happen in match / insert states
    ▪ Query sequence "matches" the profile / model
    ▪ Insert state emits a character "between columns"
BAQ: Profile HMM for SNP discovery

- Existing SNP discovery methods
  - Ambiguity in read alignment: mismatch or indel?
    - "deceive SNP callers into calling false SNPs"
  - Solution: evaluate Base Alignment Quality (BAQ)
  - Claim: using bq\textsubscript{i} = \min(bq\textsubscript{i}, \text{BAQ}\textsubscript{i}) improves accuracy
BAQ Profile HMM transition matrix

• \( \alpha = P(\text{gap opening}) \)
  - site independent
  - default = 0.001

• \( \beta = P(\text{gap extension}) \)
  - default = 0.1

• \( \gamma = (2^*L)^{-1} \)
  - length "control factor"

\[
\begin{pmatrix}
M & I & D & S & E \\
M & (1 - 2\alpha)(1 - \gamma) & \alpha(1 - \gamma) & \alpha(1 - \gamma) & 0 & \gamma \\
I & (1 - \beta)(1 - \gamma) & \beta(1 - \gamma) & 0 & 0 & \gamma \\
D & 1 - \beta & 0 & \beta & 0 & 0 \\
S & (1 - \alpha)/L & \alpha/L & 0 & 0 & 0 \\
E & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
**BAQ transition matrix example**

- \( \alpha_{M1,M2} = (1-2\alpha)(1-\gamma) \)
  - 1 - 2\( \alpha \) = 1 - 2*P(gap opening)
    - 2\( \alpha \) because opening could be in read or reference
  - 1 - \( \gamma \) correction factor for read length (shorter read more likely to transition to end state, longer read more likely to continue to have non-terminal transitions)

\[
(a_{ij})_{5\times5} =
\begin{pmatrix}
M & I & D & S & E \\
M & (1 - 2\alpha)(1 - \gamma) & \alpha(1 - \gamma) & \alpha(1 - \gamma) & 0 & \gamma \\
M & (1 - \beta)(1 - \gamma) & \beta(1 - \gamma) & 0 & 0 & \gamma \\
M & 1 - \beta & 0 & \beta & 0 & 0 \\
M & (1 - \alpha)/L & \alpha/L & 0 & 0 & 0 \\
M & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Nomenclature

- Read: \( y = c_0c_1c_2...c_lc_{l+1} \) where \( c_0 = \text{'^'} \), \( c_{l+1} = \text{'$'} \)
- Reference: \( x = r_1r_2r_3...r_L \)
- Substitution probability of \( c_i = \mathcal{E}_i \), for \( i = 1..l \)

Emission Probabilities

- \( l_k = 0.25 \) (uniform over four bases)
- \( P(c_i \mid M_k) = e_{ki} = \begin{cases} 1 - \mathcal{E}_i & \text{if } r_k = c_i \text{ (if ref. matches read)} \\ \mathcal{E}_i / 3 & \text{otherwise (mismatch, } r_k \neq c_i) \end{cases} \)
Forward and Backward Algorithms

• Recall, for regular HMMs:\(^1\):

Definition 17.9. (Forward and Backward Variables) \( \forall k \in S \) and \( t \in [1, L] \) define the forward variables \( f_k(t) \) and the backward variables \( b_k(t) \) to be:

- \( f_k(t) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t, Q_t = k) \)
- \( b_k(t) = \mathbb{P}(X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}, \ldots, X_T = x_T | Q_t = k) \)

• Same idea: recursive marginalization over dynamic programming / profile HMM paths

• --> Board work

• Result: update \( i^{th} \) base quality (\( bq_i \)) as

\[
\min(bq_i, BAQ_i)
\]

1. Reference: CS 194-1 Lecture 17 scribe notes, Spring 2012
BAQ Score Function (asymptotic)
BAQ performance

- Test data:
  - 1000 genomes project dataset
  - ts / tv is > 2 in Human
    - SNP calls should reflect this ratio
    - using BAQ gives improvement

---

1. [1000genomes.org](http://1000genomes.org)
2. Explanation of base transition vs. transversion: [http://genome.sph.umich.edu/wiki/SNP_Call_Set_Properties](http://genome.sph.umich.edu/wiki/SNP_Call_Set_Properties)
BAQ software

• Part of SAMtools package
• Ubuntu: sudo apt-get install samtools
• http://samtools.sourceforge.net/mpileup.shtml