A general approach to single-nucleotide polymorphism discovery

G Matrh et al.
1999

SNPs

- a one base DNA sequence variation between two individuals of a same species
- it is the most abundant sequence variation in populations
- almost all SNPs have only two alleles (forms)
- minor allele frequency

data in 1999

- ~1M bp of finished human reference sequence in 10 regions
- data base of express sequence tags(ESTs)
- ESTs are reverse transcribed RNA sequences form different individuals
- length=~300bp

alignment

- first, repeats in the reference is covered
- ESTs are aligned to sequence according to a common anchor
- then, error, gaps, inserts are propagated in the reminding EST
- 1365 hits were found in 147 clusters
- representing 80,469 bp of sequence 38% single, 81% by 8 for fewer ESTs coverage

paralogue

- paralogue are regions highly similar DNA sequences of an individual
- they may have arisen from the same evolutionary root
- must not confuse difference in paralogue with SNPs

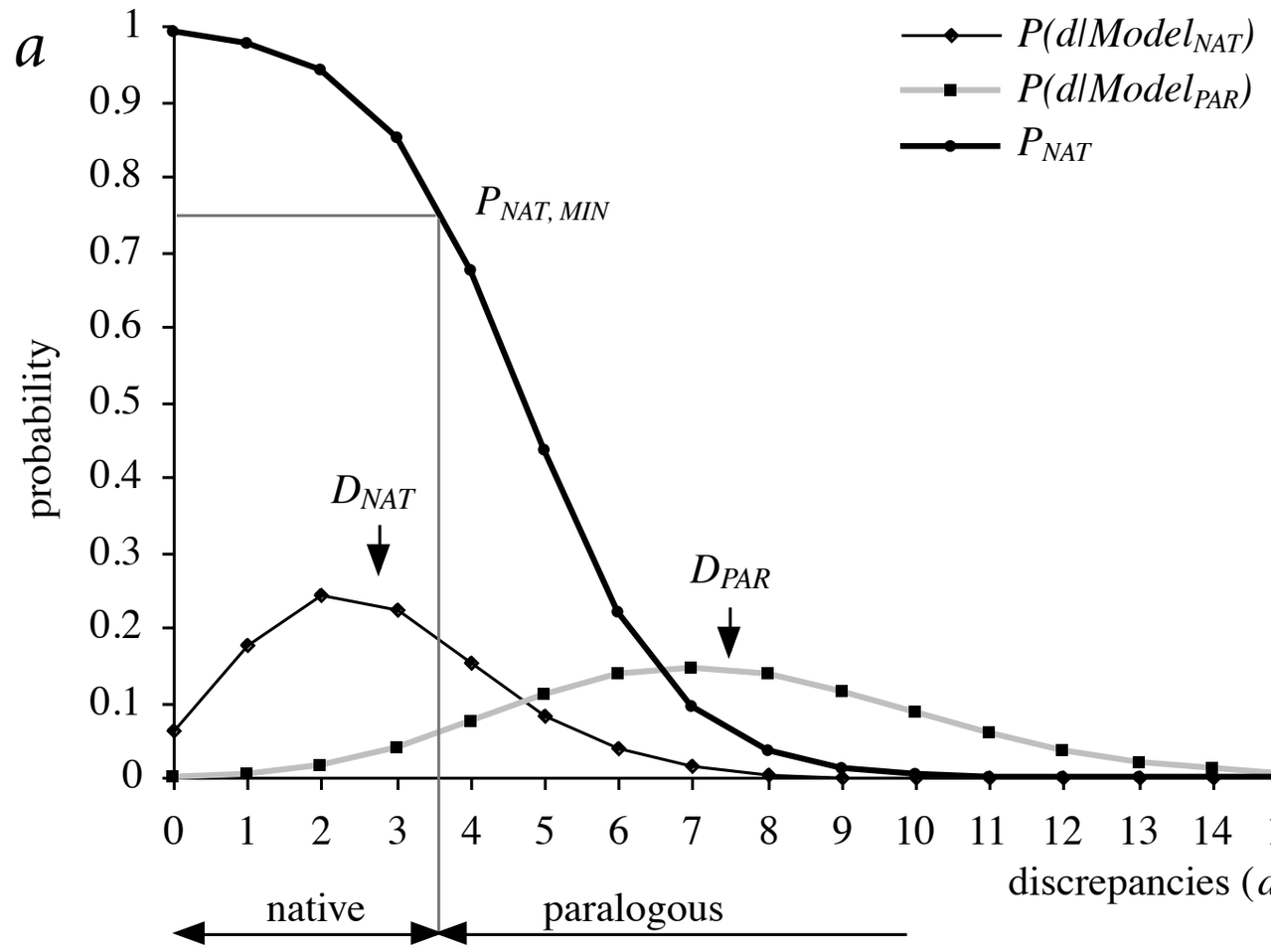
paralogue discrimination

- paralogous sequences have pairwise dissimilarity rate higher than P_{PAR}=2%
- SNP rate is P_{SNP}=0.1%
- they can be differentiated by percent different bases

model for paralogue discrimination

- in a sequence with length L
- expected number of base difference:
  form paralogue: $D_{PAR} = L \times P_{PAR} + E$
  from SNP: $D_{NAT} = L \times P_{POLY,2} + E$
- They can be approximated as a Poisson distribution
- ratio: $$P(Model_{NAT} \mid d) = \frac{1}{1 + e^{(D_{NAT} - D_{PAR})} \cdot \left(D_{PAR} / D_{PAR} \right)}$$

paralogue

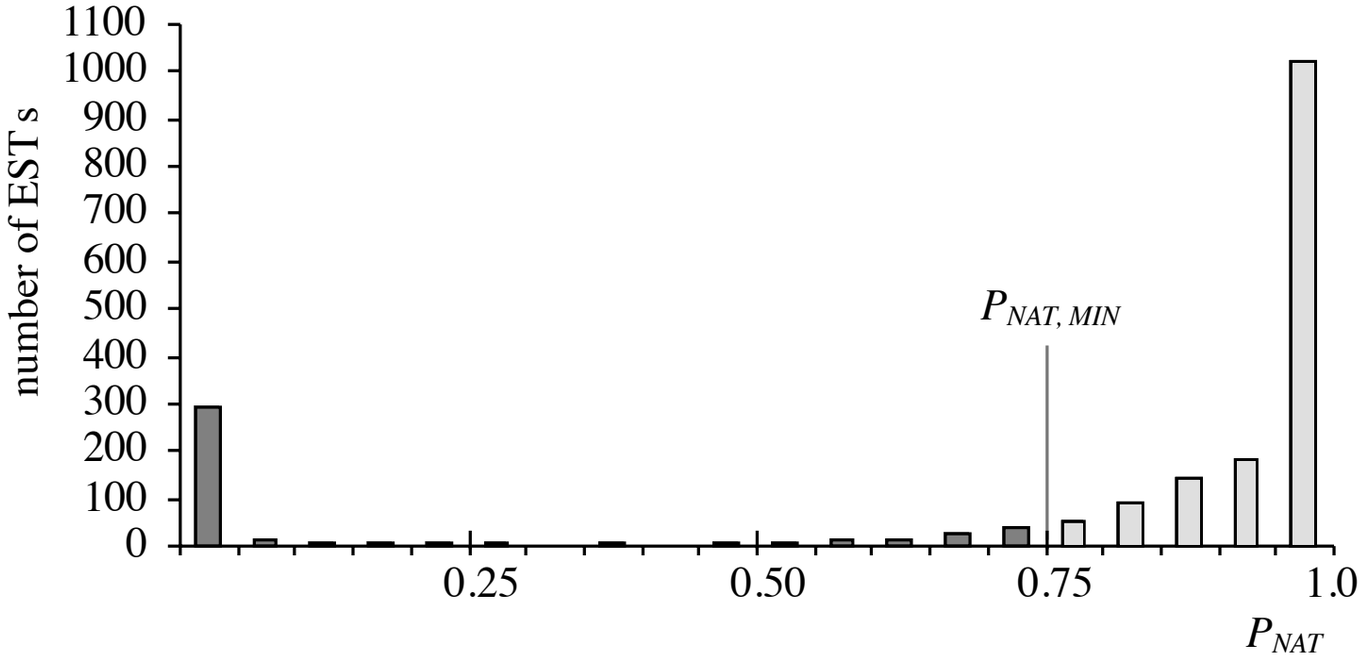


hypothetical uniform gene distribution

250 bp long EST E=2.525

no mistake in reference

paralogue



experimental results for 1954 cluster members
(anchored to 10 genomic sequences)

Bayesian SNP inference

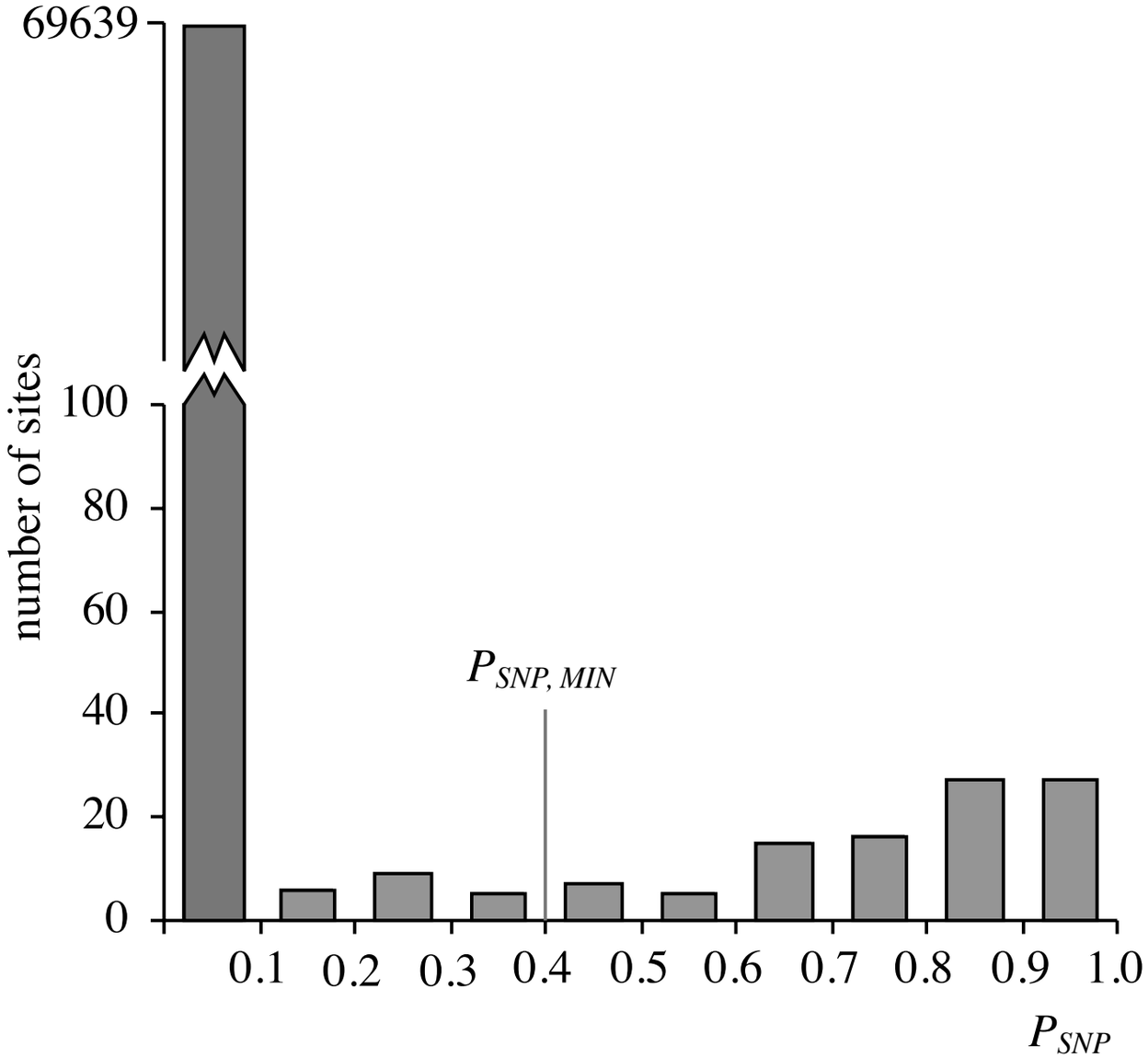
- prior probability for SNP is 0.003
- likelihood of data given sequence (probability of correct sequencing) is estimated form chromatograms
- with these two piece of information, posterior probability is calculated

Bayesian SNP inference

$$P(S_1,\ldots,S_N \mid R_1,\ldots,R_N) = \frac{\frac{P(S_1 \mid R_1)}{P_{Prior}(S_1)} \cdot \ldots \cdot \frac{P(S_N \mid R_N)}{P_{Prior}(S_N)} \cdot P_{Prior}(S_1,\ldots,S_N)}{\sum_{every\ (S_{i1},\ldots,S_{iN})} \frac{P(S_{i1} \mid R_1)}{P_{Prior}(S_{i1})} \cdot \ldots \cdot \frac{P(S_{iN} \mid R_1)}{P_{Prior}(S_{iN})} \cdot P_{Prior}(S_{i1},\ldots,S_{iN})}$$

- posterior probability for all $S_1 \ldots S_N$ are calculated
- SNP probability is the sum of all probability associated with $S_1 \ldots S_N$ that have a SNP

Results

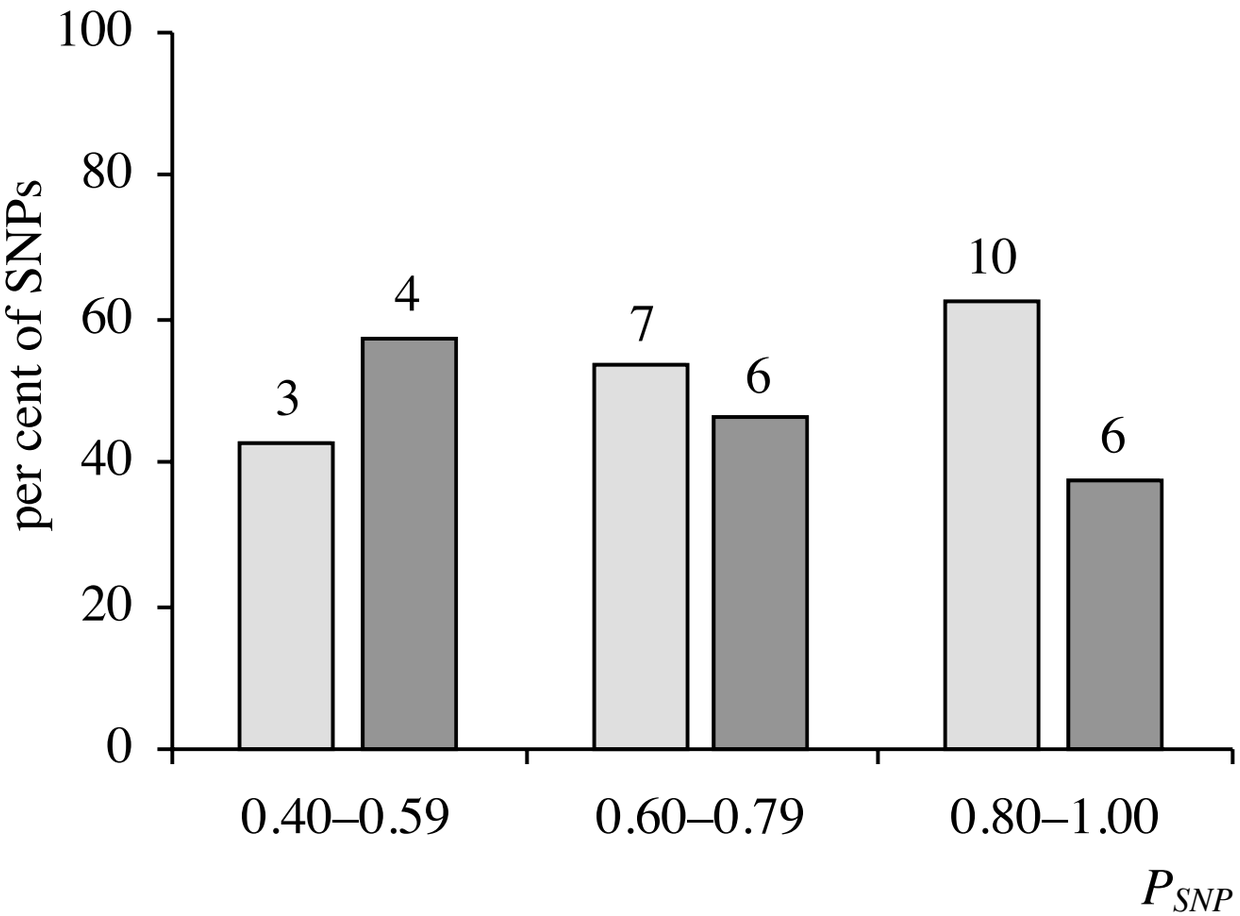


- 97 candidates were found out of 80,469 bp covered sequences

verification

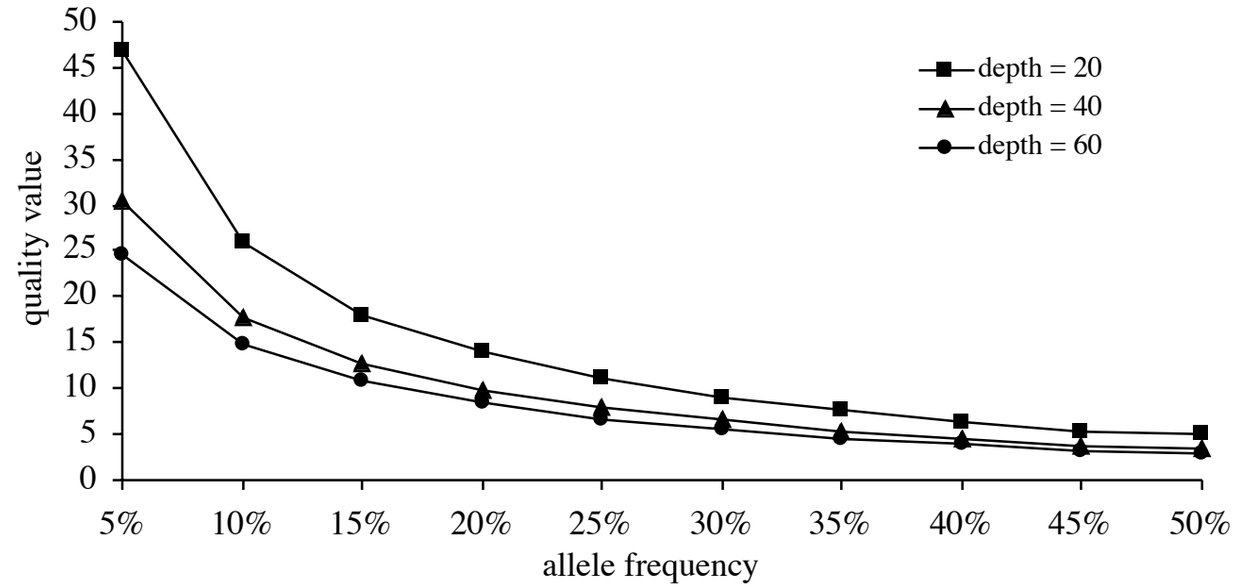
- 38 of these are near the 3' end of ESTs, and verified to have problems with cDNA liberty construction
- 18 of these can not be analyzed for various reasons
- for the remaining 36 sites, they confirmed 20 sites at least 1 in 4 individuals screened
- overall confirmation rate is reported to be 56%
- another SNP was found in 11,455bp STS

verification

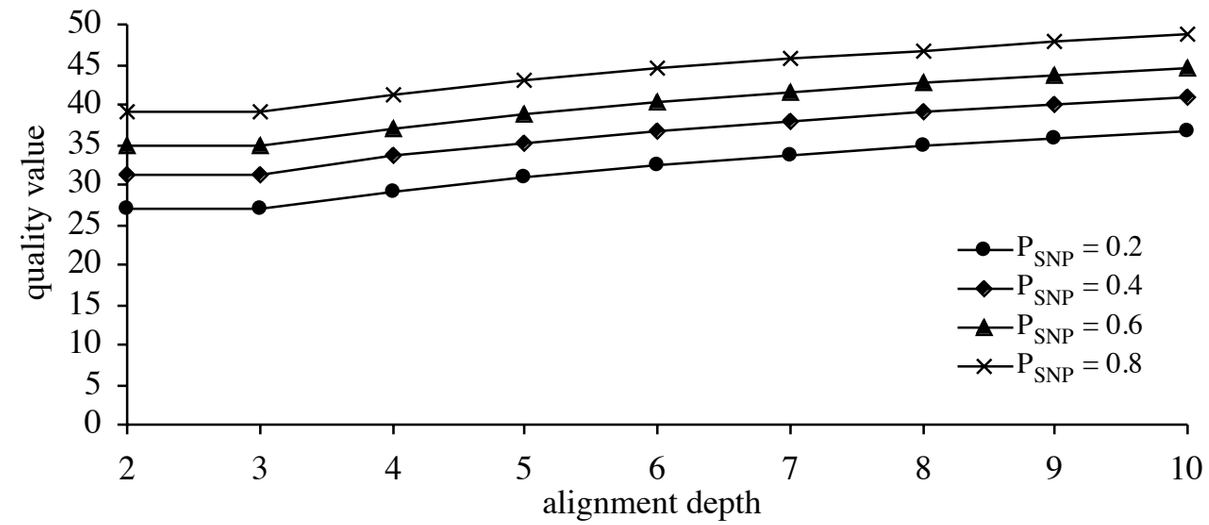


results

a


quality value
50
45
40
35
30
25
20
15
10
5
0
5%
10%
15%
20%
25%
30%
35%
40%
45%
50%
allele frequency
depth = 20
depth = 40
depth = 60

b


quality value
50
45
40
35
30
25
20
15
10
5
0
2
3
4
5
6
7
8
9
10
alignment depth
P_SNP = 0.2
P_SNP = 0.4
P_SNP = 0.6
P_SNP = 0.8

conclusion

- Poly-bayes offers a relative straight forward way of finding SNP sites
- reasonable sensitivity and accuracy
- designed for long reads