Optimal pooling strategies for SNP detection using next generation sequencing experiments

ANDREAS FUTSCHIK & DAVID RAMSEY
UNIVERSITY OF VIENNA
AND
UNIVERSITY OF LIMERICK

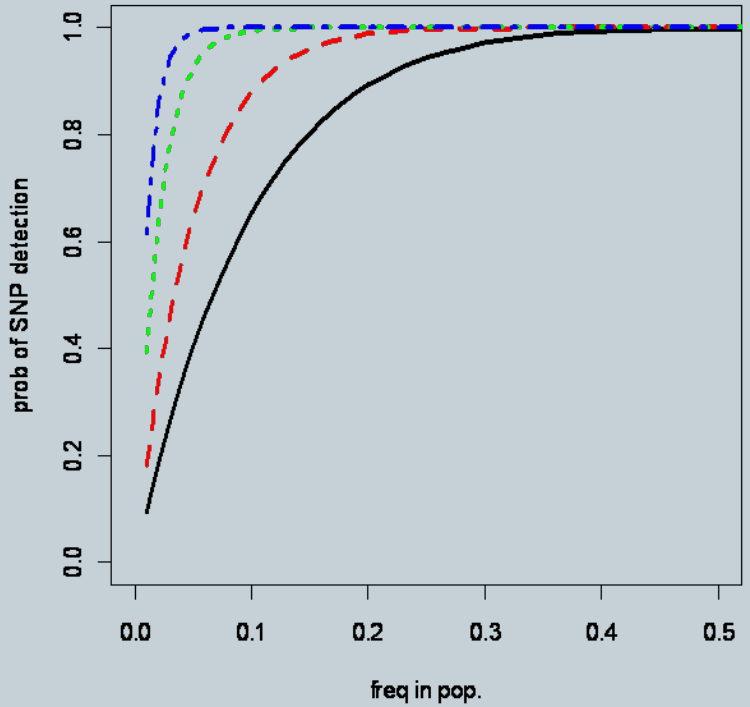
Next Generation Sequencing

- Next Generation Sequencing—more and more genomic data at lower and lower cost
- With increased sequencing capacities, hope to be able to detect also more SNPs involving rare minor alleles

SNP detection: Individual Sequencing versus Pooling

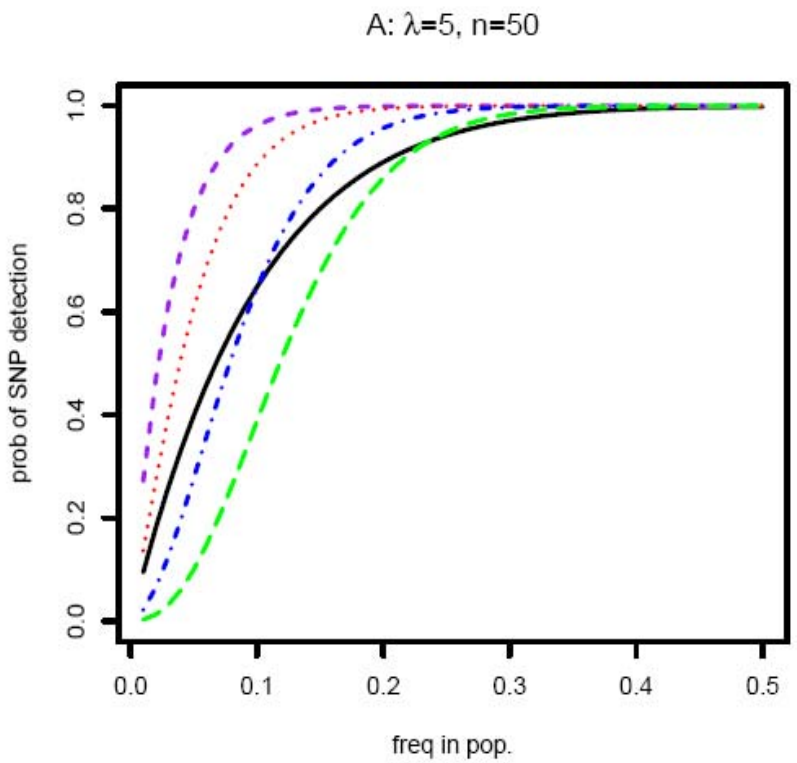
- Individual sequencing: With sufficiently high coverage, controlling for sequencing errors fairly straightforward
- Pooling—a cost effective alternative that permits to sequence also larger samples
- Larger samples should increase the chance of capturing rare alleles—so is it better to pool?
- *Futschik and Schlötterer (2010)*

Without sequencing errors it is always be better to pool



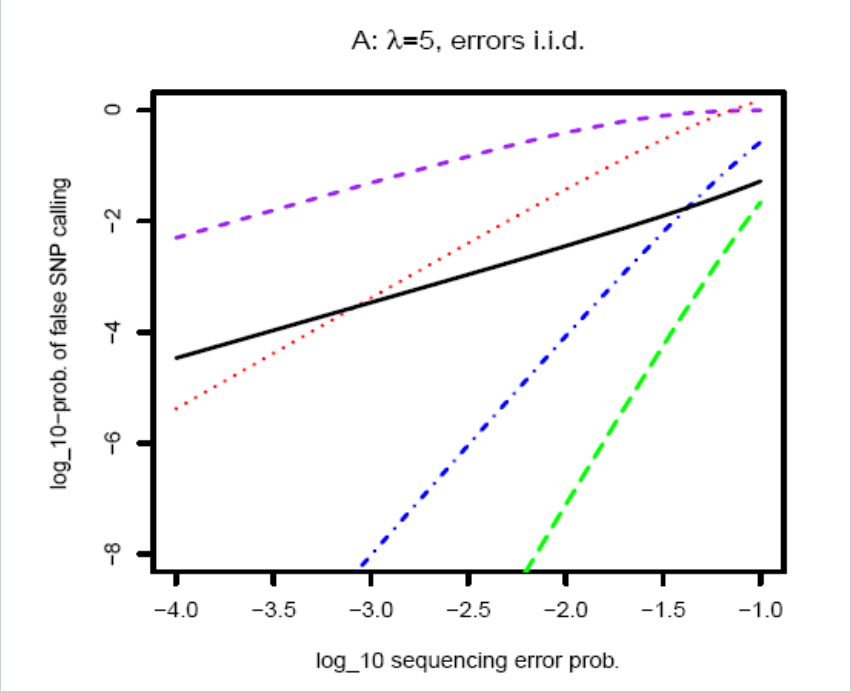
- Individual sequencing of 10 individuals, each with coverage 30 *(black line)*
- Pooling experiment with same total sequencing effort:
  - red: pool of size 20
  - green: pool of size 50
  - blue: pool of size 100

With sequencing errors situation gets more complicated



- Black individual sequencing of 10 individuals, each with coverage 5
- Pooling experiment with same sequencing total effort:
  - Purple: no error correction
  - red: m.a.f. > 1
  - blue: m.a.f. > 3
  - green: m.a.f > 5
- Futschik & Schlötterer (2010)

False positive rate



Dependent errors:

$$(1 - F_{(P)}(b-1, \lambda k\epsilon))[1 - F_{(P)}(0, \lambda k(1-\epsilon))]$$

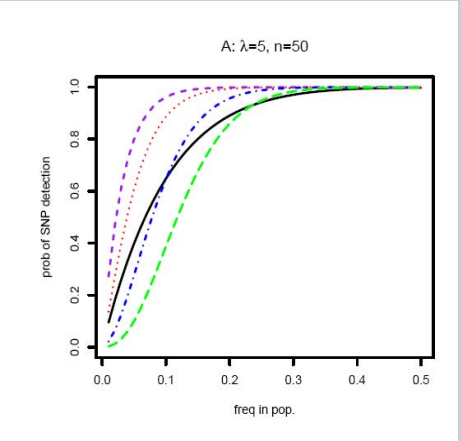
Independent errors (upper bound):

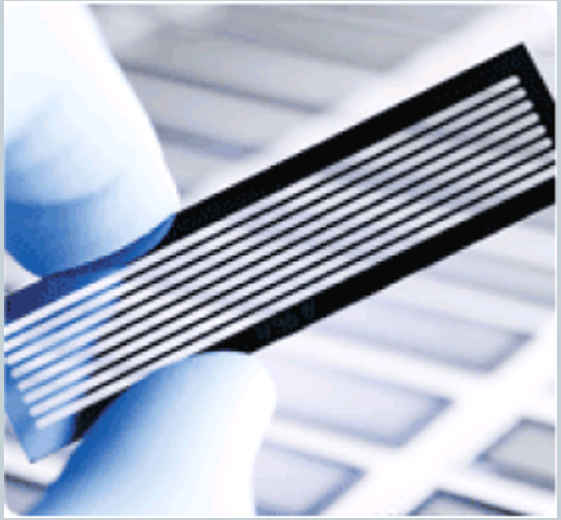
$$3\left(1 - F_{(P)}(b-1, \lambda k\epsilon/3)\right)$$

$$F_{(P)}(b,\gamma) = \sum_{i=0}^{b} \frac{\gamma^i}{i!} \exp(-\gamma)$$

How to find rare alleles?

- Low power with one lane/pool
- Suppose budget sufficient for *k* lanes:
  - How to optimally design an experiment involving k lanes?
  - How to test for rare SNP‘s in such an experiment?





Testing for SNPs

- True minor allele frequency *p* in sample.
- Data:
  - Pool of size *m* for each lane
  - Number of reads from each of k lanes: $\mathbf{R} = (R_1, R_2, \ldots, R_k)$
  - Minor allele frequencies for each lane: $\mathbf{X} = (X_1, X_2, \ldots, X_k)$
- Consider position at which polymorphism is observed.
  - H_0: $p = 0$ (polymorphism caused by sequencing errors)
  - H_1: $p > 0$ (SNP position found)
  - Protect against false positives

Likelihood Ratio Test

$$LR = \frac{\max_p \prod_{i=1}^{k} \sum_{a_i=0}^{m} \binom{m}{a_i} p^{a_i} (1-p)^{(m-a_i)} [q(a_i)]^{x_i} [1-q(a_i)]^{r_i - x_i}}{\epsilon^{\sum_{i=1}^{k} x_i} (1-\epsilon)^{\sum_{i=1}^{k} (r_i - x_i)}}.$$

$$q(a) = \frac{a(1-\epsilon)}{m} + \frac{\epsilon(m-a)}{m}.$$

p … population frequency of minor allele
ε … (maximum) probability of sequencing error
a_i … true minor allele frequency for pool *i*
m … pool size

• Chi-square approximation of *2log(LR)* does not work well, critical values via simulation.

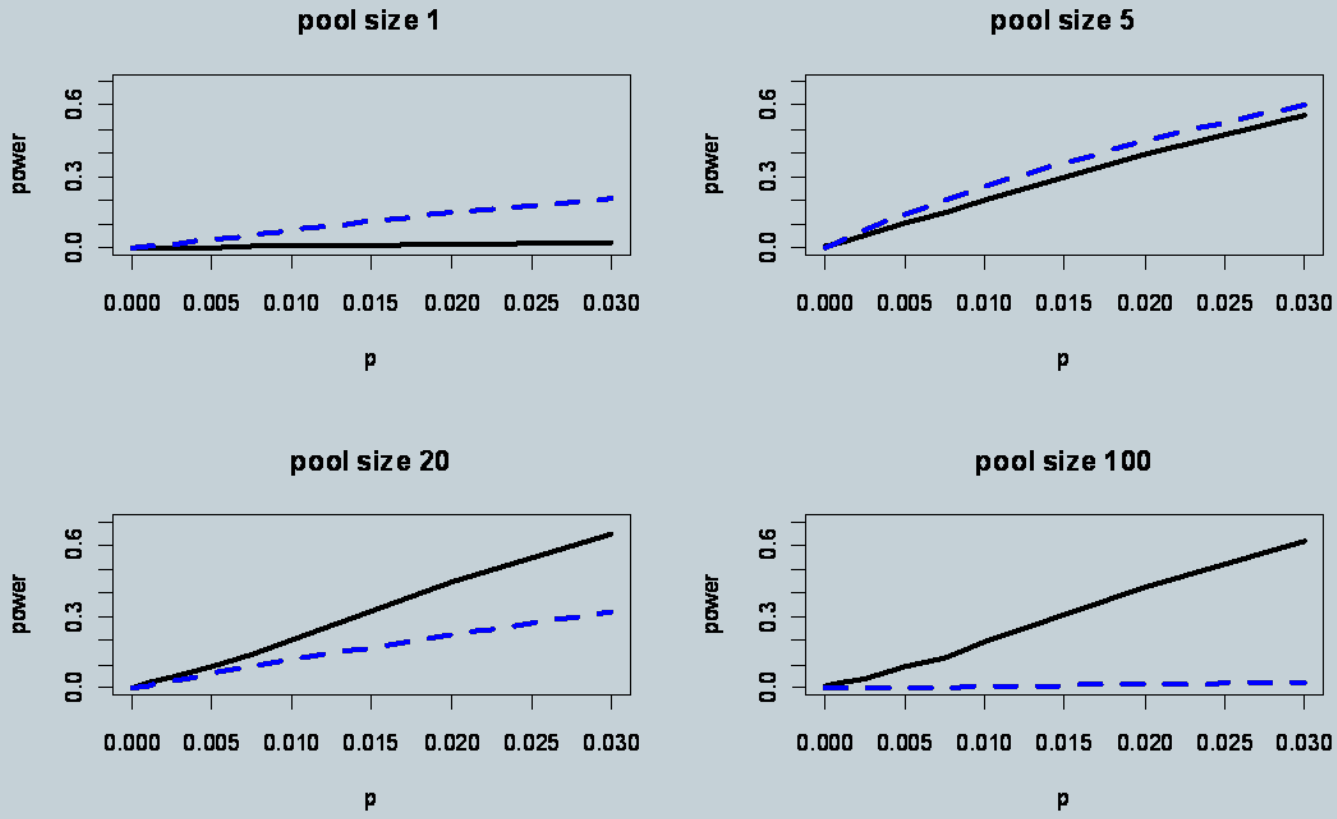
Maximum Test

$$U_k = \max_{1 \le i \le k} X_i$$

Call position a SNP, if $U_k > u_k$

- Critical value: u_k $\sqrt[k]{1-\alpha}$ quantile of $\text{Poisson}(\lambda\epsilon)$

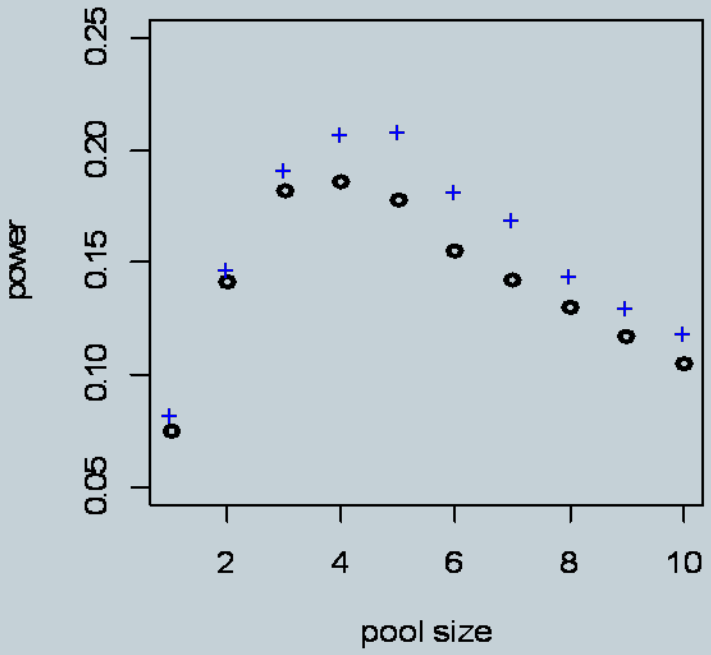
Which test is more powerful?



8 lanes, $\lambda=20$, $\varepsilon=0.01$, $\alpha=0.01$

Blue ---Maximum Test, **black** … likelihood ratio test

Optimization of pool size is important



16 lanes, p=0.005, λ=20, ε=0.01, α=0.001

- With larger pools :
  - increased chance of inclusion of rare alleles
  - smaller number of reads per individual …thus harder to distinguish rare alleles from sequencing errors
- There is an optimum pool size that maximizes chance of SNP detection!

Optimizing Pool Size for a fixed Number of Lanes I

Lemma. Suppose the pool size is fixed and ignore the possibility of errors and assume that there are b individuals with the minor allele in the sample. Then the distribution of the maximum number of reads of a minor allele across all lanes is stochastically smallest, when one individual with the minor allele appears in each of b lanes.

Lower bound on power of maximum test

$$P[D] = \sum_{b=0}^{\infty} P[D|B=b]P(B=b) \geq \sum_{b=1}^{\infty} P[D|B=b]P(B=b).$$

$$P[D|B=b] = P(U > u_c|B=b) = 1 - P(U_k \leq u_k|B=b) \geq 1 - P(V_1 \leq u_k)^b,$$

$$P[D|B=b] = 1 - r_k^b,$$

$$r_k = \sum_{j=0}^{u_k} \frac{e^{-\lambda/m}(\lambda/m)^j}{j!}.$$

Approximate Optimization Problem

$$
\begin{aligned}
P[D] &\approx \sum_{b=1}^{\infty} \frac{e^{-\mu}\mu^b[1-r_k^b]}{b!} = \sum_{b=1}^{\infty} \frac{e^{-\mu}\mu^b}{b!} - \sum_{b=1}^{\infty} \frac{e^{-\mu}(\mu r_k)^b}{b!} \\
&= (1-e^{-\mu}) - e^{-\mu(1-r_k)} \sum_{j=1}^{\infty} \frac{e^{-r_k\mu}(\mu r_k)^b}{b!} \\
&= (1-e^{-\mu}) - e^{-\mu(1-r_k)}(1-e^{-r_k\mu}) = 1 - e^{-\mu(1-r_k)}.
\end{aligned}
$$

Optimizing Pool Size for a fixed Number of Lanes

- Maximum test permits relatively simple approximate computation of optimum pool size, solve

$$0 = \sum_{j=u_k+1}^{\infty} \frac{e^{-\lambda/m}(\lambda/m)^j[1+\lambda/m-j]}{j!}.$$

pool of size m, expected coverage λ, critical value $u_k = u_k(\varepsilon, \alpha)$

Minimum minor allele frequency that can be detected with power β

- Define:

$$f_k(m;p,\alpha) = e^{-\mu(1-r_k)} = \exp\left[-mkp \sum_{j=u_k+1}^{\infty} \frac{e^{-\lambda/m}(\lambda/m)^j}{j!}\right]$$

- Need to solve:

$$1 - f_k(m_k^*;p,\alpha) = \beta$$

Minimum minor allele frequency that can be detected with power β

$$p_{min}(k;\beta,\alpha) = \frac{-\ln(1-\beta)}{m_k^* k[1-r_k]}.$$

$$r_k = \sum_{j=0}^{u_k} \frac{e^{-\lambda/m}(\lambda/m)^j}{j!}.$$

Power β, k lanes, optimum pool size m_k^*

Number of lanes needed to achieve a given power (with optimum pool size)

$$k(p;\beta,\alpha) = \min_{k} p_{min}(k;\beta,\alpha) \leq p.$$

Numerical Example

	$k = 16$	$k = 40$	$k = 80$	$k = 120$
$\epsilon = 0.01$	$u = 3, m^* = 4,$ $p_{min} = 0.0637$	$u = 4, m^* = 3,$ $p_{min} = 0.0314$	$u = 4, m^* = 3,$ $p_{min} = 0.0157$	$u = 4, m^* = 3,$ $p_{min} = 0.0105$
$\epsilon = 0.005$	$u = 3, m^* = 4,$ $p_{min} = 0.0637$	$u = 3, m^* = 4,$ $p_{min} = 0.0255$	$u = 3, m^* = 4,$ $p_{min} = 0.0127$	$u = 3, m^* = 4,$ $p_{min} = 0.00849$
$\epsilon = 0.002$	$u = 2, m^* = 6,$ $p_{min} = 0.0482$	$u = 2, m^* = 6,$ $p_{min} = 0.0193,$	$u = 2, m^* = 6,$ $p_{min} = 0.00964$	$u = 3, m^* = 4,$ $p_{min} = 0.00849$
$\epsilon = 0.001$	$u = 2, m^* = 6,$ $p_{min} = 0.0482$	$u = 2, m^* = 6,$ $p_{min} = 0.0193$	$u = 2, m^* = 6,$ $p_{min} = 0.00964,$	$u = 2, m^* = 6,$ $p_{min} = 0.00643$

α=0.01, λ = 20, power β = 0.95

Conclusions

- Chance of SNP detection considerably higher with pooled samples than for individual sequencing
- Without costing additional false positives, clever choice of pool size can increase power of SNP detection considerably
- For the maximum test, optimum pool size can be obtained without much computational effort, solution provides approximation also for LR-test.
- Further research: tagged reads, optimal pool size in population genetic inference, e.g. when testing for selection …