



Application of Multi-Objective Optimization to Pooled Experiments of Next Generation Sequencing for Detection of Rare Mutations

Julius Žilinskas^{1*}, Algirdas Lančinskas¹, Mario Rosario Guarracino²

1 Institute of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania, **2** High Performance Computing and Networking Institute, National Research Council, Naples, Italy

Abstract

In this paper we propose some mathematical models to plan a Next Generation Sequencing experiment to detect rare mutations in pools of patients. A mathematical optimization problem is formulated for optimal pooling, with respect to minimization of the experiment cost. Then, two different strategies to replicate patients in pools are proposed, which have the advantage to decrease the overall costs. Finally, a multi-objective optimization formulation is proposed, where the trade-off between the probability to detect a mutation and overall costs is taken into account. The proposed solutions are devised in pursuance of the following advantages: (i) the solution guarantees mutations are detectable in the experimental setting, and (ii) the cost of the NGS experiment and its biological validation using Sanger sequencing is minimized. Simulations show replicating pools can decrease overall experimental cost, thus making pooling an interesting option.

Citation: Žilinskas J, Lančinskas A, Guarracino MR (2014) Application of Multi-Objective Optimization to Pooled Experiments of Next Generation Sequencing for Detection of Rare Mutations. PLoS ONE 9(9): e104992. doi:10.1371/journal.pone.0104992

Editor: Maurizio D'Esposito, Institute of Genetics and Biophysics, Italy

Received: March 31, 2014; **Accepted:** July 15, 2014; **Published:** September 2, 2014

Copyright: © 2014 Žilinskas et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper.

Funding: This research was funded by a grant (No. MIP-063/2012) from the Research Council of Lithuania. This work has been partially funded by the Italian Flagship Project Interomics, The Italian PON02_00619 projects, and F.A.R. LAB-GTP initiative. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: julius.zilinskas@mii.vu.lt

Introduction

In the 1990s it became clear it was not possible to sequence all or large parts of the human genome using already available technologies. Indeed, sequencing the 3.6 billions base pairs using sequences shorter than one thousand base pairs would have been expensive and time demanding. Therefore, the problem was solved concurrently, dividing the long DNA sequence in to sub-sequences, and analyzing them in parallel.

The Next Generation Sequencing (NGS) technologies use this idea, producing a large number (in the order of ten of millions) of *short reads*, each being the result of the sequencing of a DNA fragment. These fragments are obtained firstly breaking the original DNA sequence, using ultrasounds or restriction enzymes, in pieces of a fixed size, and then cloning them in a process known as *amplification*. The result of this process is then sequenced, producing the short reads, whose length, at the moment of writing, varies between 50 and 500 base pairs, depending on the technology used by the sequencer. These nucleotide string sequences are stored in flat files, together with information on the quality of each base call. Then, each short read is aligned on the reference genome. The number of reads that aligns on a specific position is called *coverage*. The result of the alignment is a file containing the information about the alignment, such as the position on the reference genome where the short read has been mapped and its error probability, and the potential insertions or

deletions needed to obtain the alignment. The data are then analyzed to detect variations with respect to the reference genome.

An example of mutation is a Single Nucleotide Polymorphism (SNP), where the mutation consists in a change of a single nucleotide into another. This can be detected in a certain position, looking at the nucleotides in each short read that are aligned with the position of interest. If the SNP is only on one allele (heterozygous mutation), then a half of the short reads might be different from the reference. Some of these polymorphisms are due to the natural variability of organisms, and some might be unique for a specific individual. Some of those mutations in protein coding genes can alter the associated protein, whereas some other have no effect, due for example to the degeneracy of genetic code. In case of rare genetic disease studies, we are interested in detecting mutations in the DNA sequence, that are unique for an individual and causative of the disease. Due to its rarity and complexity of its treat, these mutations might appear only in one patient in the population sample under study.

The widespread availability of high-throughput sequencing technologies makes it possible to analyze large panels of patients affected by rare genetic diseases, although the time and cost of such experiments are still a limiting factor. The existing solutions rely, when possible, in the analysis of pools of patients. This pooling technique consists in analyzing a mixture of DNA from a group of individuals, and assigning the discovered causative mutations to a single patient. The inspiration for the present work comes from a study conducted on muscular diseases, whose aim is

to analyze the samples from a thousand patients affected by muscular diseases. To this extent, the sequencing experiment has been planned targeting a panel of genes taken from a previous study, where all known genes related to muscular diseases have been reviewed [1], and using Agilent HaloPlex Target Enrichment kit.

Examples of studies using pooled experiments, in various genetic and genomic applications, can be found in [2–4]. The limit of such studies is that no evaluation has been done to assess the group size of the pools, the probability to detect a mutation, the cost associated with the experiment and its biological validation.

When dealing with pooled DNA sequencing experiments, there are various aspects that need to be considered. First, the contribution of DNA from each patient in a pool is in general different. This is due to the fact that this operation is usually manually executed, and even a robot or a skilled professional biologist might produce a mixture in which the most represented individual contributes with a quantity of DNA that is double or more, with respect to the less represented one. Therefore, the number of short reads produced during sequencing and belonging to each patient might be very different. Since a certain percentage of reads is affected by base call and mapping errors, in the worst case, all mutations of a specific patient might be in reads that are unusable and therefore discarded from the analysis. For this reason, a rule of thumb used by biologists dealing with this kind of experiments is that the number of short reads covering each position of interest should be at least thirty. Such number should be multiplied by the number of patients in the pool, thus limiting the maximum number of patients m_{max} per pool, due to the limited coverage capability of sequencing methods.

Supposing that the disease has a complex genetic trait and is rare, we might expect each patient to bring a certain number of unique mutations, that will be detectable only in the short reads belonging to that specific patient. Therefore, there will be a certain number of mutations that will appear in the experiment. If we want to attribute these mutations to a specific patient, we need to test all patients for all found mutations, using for example low throughput Sanger sequencing technology. Although Sanger tests are orders of magnitude cheaper than the NGS ones, even when dealing with a low number of expected mutations in a pool of patients, this might result in a large number of experiments with a cost comparable with NGS. We will show that in some cases, it might be convenient to explore some controlling strategies that replicate patients in pools used in the high throughput experiment, in such a way that the number of Sanger sequencing experiments is decreased (or even eliminated), together with the total cost of the experiments. This will be achieved sequencing each patient twice, and allocating the patient in two different pools, which we call *main* and *replicated* pools.

In this paper we propose a technique to plan NGS pooled experiments. The solution takes into account experimental setups with or without replication of the patients. A combinatorial organization of pools has been devised with some optimality characteristics, whose advantages are: (i) the solution guarantees that the mutation is detectable in the experimental setting, and (ii) the cost of the experiment and its biological validations using Sanger sequencing is minimized. Then we use a probabilistic model to determine the number of patients for each pool, and a multi-objective optimization formulation of the problem to obtain optimal grouping of patients with a given probability to detect a mutation.

Methods

We recall that when patients are organized in pools and analyzed with NGS technology, the assignment of found mutations to patients is accomplished using Sanger sequencing technology. In the following we provide a mathematical formulation of the problem, together with an analysis of results in terms of costs associated to the optimal allocation of patients in pools.

Pooling without Replica

Suppose there are n patients each of which might bring at most N_m rare mutations. All n patients can be allocated in p pools consisting of m_1, \dots, m_p patients, such as:

$$\sum_{j=1}^p m_j = n. \quad (1)$$

The size of each pool is restricted by a maximum number m_{max} of patients, as explained in the previous section, therefore:

$$m_j \leq m_{max}, j = 1, \dots, p. \quad (2)$$

If c_1 and c_2 are the costs of NGS and Sanger sequencing, respectively, then the total cost of the experiment can be calculated by:

$$C_N = p \cdot c_1 + \sum_{j=1}^p c_2 \cdot N_m \cdot m_j^2. \quad (3)$$

The total cost C_N of the experiment depends on the cost c_1 for sequencing each of the p of pools, and the cost c_2 of the Sanger validations. Once $N_m \cdot m_j$ mutations have been found in the j -th pool, they need to be assigned to one of the m_j patients, thus requiring $N_m \cdot m_j^2$ validations.

Pooling with Replica

The validation of a biological experiment is usually obtained replicating the experiment at least twice. In this section we investigate two different strategies for pooled experiments. Both strategies are based on construction of pools so that (i) each patient is replicated two times in different pools, and (ii) if a mutation is identified in both main and replicated pool, it refers to at most 1 patient.

Transposition. The first strategy, called Transposition, is based on the concept of matrix transposition as it is illustrated in the left and center panels of Figure 1, where the left image illustrates the set of 15 patients distributed into 5 main pools, and the middle image represents the replicated pools. Naturally this approach can be applied if the maximum number of patients allowed in a single pool is greater than or equal to the number of main pools ($m_{max} \geq p_m$). If the latter requirement is not satisfied ($m_{max} < p_m$), then the number of patients in a control pool would exceed the maximum number of patients n_{max} in a pool. Therefore a larger number of control pools must be constructed using a similar strategy, as it is shown in the right panel of Figure 1, in order to correctly fit all patients in replicated pools. In general, the number of replicated pools (p_c) can be chosen by taking into account the following two constraints:

- (i) the number of replicated pools cannot be larger than the total number of patients ($p_c \leq n$), in order to have enough patients to complete all pools;
- (ii) the number of replicated pools cannot be smaller than the number of patients in the largest main pool: $p_c \geq m$, where

$$m = \max_{1 \leq j \leq p_m} m_j. \tag{4}$$

The second constraint is based on the requirement that all patients from a single main pool must be allocated in separate replicated pools, thus ensuring that identification of a SNP in any two pools would refer to at most 1 patient. Therefore, a SNP, detected in one main and one replicated pools, can be assigned to a single patient. Since no Sanger tests are needed, the price of pooling for this strategy can be expressed as

$$C_T = (p_m + p_c) \cdot c_1, \tag{5}$$

where p_m is the number of main pools, p_c is the number of replicated pools, and c_1 the price for NGS for each pool.

OptReplica. The second strategy, called OptReplica, is built so that for a given number m of patients in the largest pool the smallest number p of pools is chosen subject to:

$$p > = \frac{2n}{m}, p > = m + 1. \tag{6}$$

The first inequality is based on the requirement to fit all allocations of n patients repeating each of them twice. The second inequality is based on the requirement to discriminate the SNPs belonging to any patient: identification in any two pools of a SNP must match at most 1 patient. So if m patients are allocated to the first pool, m other pools are needed with each of these patients allocated to different pools, so that no two or more patients belonging to the same main pool are together in any other pool.

In the case of $p = m + 1$, allocation of patients to pools correspond to 2-combinations of pools. Therefore the maximum

number of patients to allocate is

$$\binom{p}{2} = \frac{p(p-1)}{2} = \frac{pm}{2}. \tag{7}$$

The algorithm to define allocations can be described as: *allocate each patient in both the first pool that is not yet completely filled, and in the first pool with the smallest number of allocated patients.* The same strategy can be used when $p > m + 1$. Examples are shown in Figure 2.

Since no Sanger tests are needed, the price of pooling for this strategy is

$$C_{OR} = p \cdot c_1, \tag{8}$$

where p stands for the number of pools, and c_1 for the price for NGS of each pool.

Grouping of Patients. Both strategies Transposition and OptReplica allow identification of an SNP belonging to a single patient without any Sanger test. Naturally, the total cost of the experiment depends on the total number of pools, used for the allocation of the patients, and the lowest price can be obtained when the number of pools is smallest (see (5) and (8)). However, the lowest number of pools is restricted by the maximum number of patients (m_{max}) allowed in a single pool.

For example, if we use Transposition strategy for allocation of $n = 128$ patients with $m_{max} = 16$, it would be optimal to use $p_m = 8$ main pools with $m = 16$ patients each. Since the number of replicated pools $p_c \geq m$, this would lead to $p_c = 16$ replicated pools, each with $m = 8$ patients, thus resulting in half loaded replicated pools. The number of these pools can be reduced by aggregating patients into groups, each of which is considered as a single unit. For example, if $n = 128$ patients are grouped into 64 groups with 2 patients per group, then an optimal number $p = 8$ of replicated pools can be used. In this case, when a variation is found in both the main and replicated pools, the grouping will lead to an indecision between the two patients belonging to the group present in both pools, that need to be resolved with a Sanger sequencing.

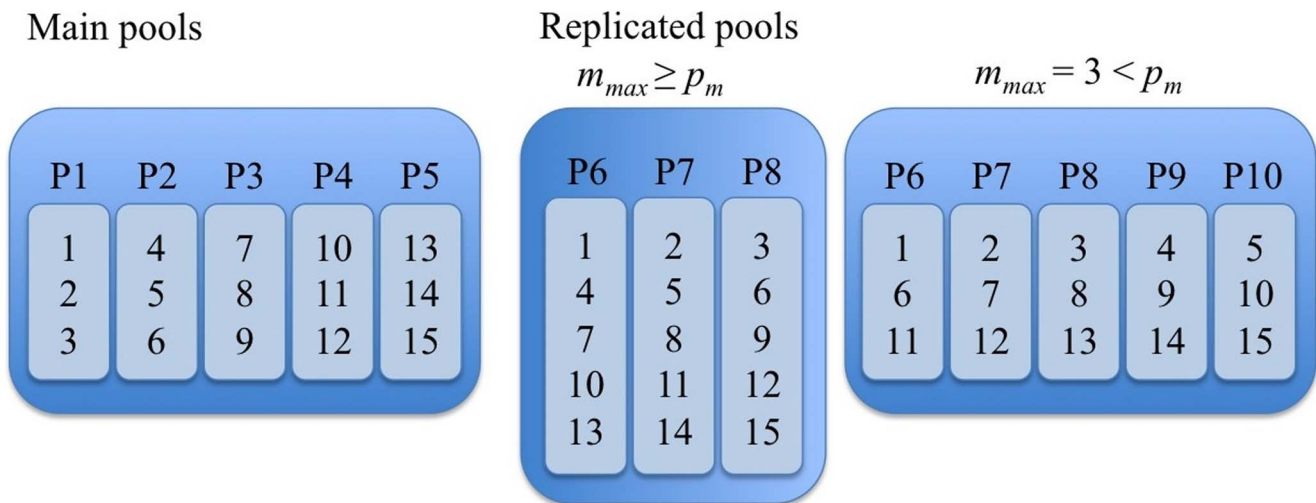


Figure 1. Illustration of allocation of patients in main pools (left), and in control pools, when $m_{max} \geq p_m$ (center) and $m_{max} < p_m$ (right), using transposition strategy.
doi:10.1371/journal.pone.0104992.g001

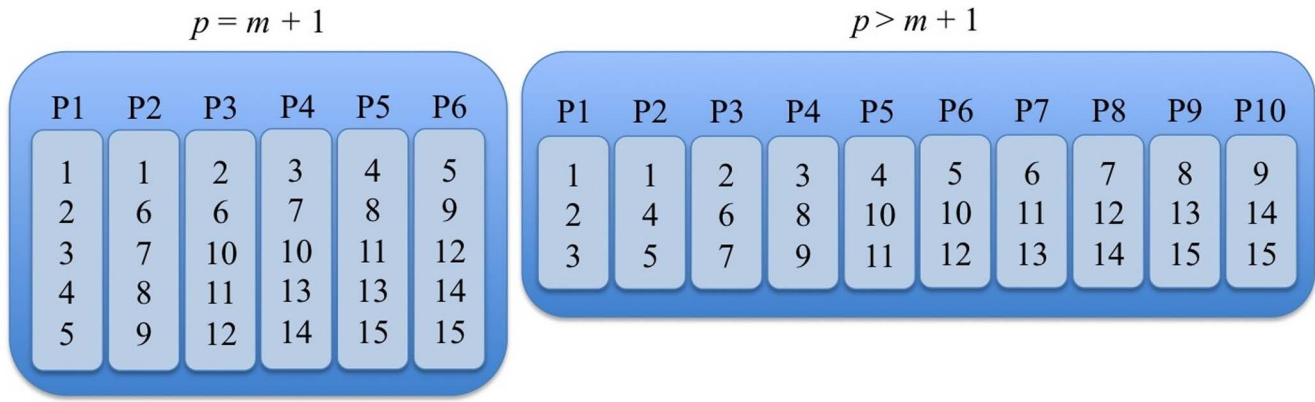


Figure 2. Illustration of allocation of patients in control pools using strategy OptReplica, when $p = m + 1$ (left) and $p > m + 1$ (right).
doi:10.1371/journal.pone.0104992.g002

The dependence of the minimum number of pools on the total number of patients using Transposition strategy is illustrated in Figure 3, where the horizontal axis corresponds to the total number of patients, the vertical axis to the number of pools, and different curves to the different sizes of groups: no grouping and with groups size equals 2.

On one hand, the grouping of patients can reduce the number of pools, on the other hand it leads to an identification of a group of patients instead of a single one. Therefore, Sanger tests must be additionally used to solve the indecision between patients in an identified group. The total cost of the experiment with grouping of patients can be expressed as:

$$C_T^g = (p_m + p_c) \cdot c_1 + c_2 \cdot N_m \cdot n_g, \tag{9}$$

where c_1 and c_2 stand for the cost of NGS and Sanger tests respectively, N_m stands for the number of rare mutations per patients, and $n_g > 1$ is the number of patients in a group.

Multi-objective Optimization

The maximum number m_{max} of patients allowed in a single pool may be determined with respect to the unity probability π to detect the mutation. If the number of patients in a pool is lower than or equal to m_{max} , then $\pi = 1$; otherwise, if the number of patients in a pool exceeds the value of m_{max} then the probability to detect the mutation becomes lower than 1, and decreases when the number of patients in a pool increases. This is due to the fact that for increasing number of patients, the number of reads representing DNA strings of a single patients decreases. However, the total cost of the sequencing can be notably reduced increasing the number of patients in a single pool. Therefore two conflicting objectives can be envisaged.

In order to analyze the influence of sequencing errors in the planning of NGS pooled experiments with respect to the total price of the experiment, a multi-objective optimization problem should be formulated by taking into account the following two objectives:

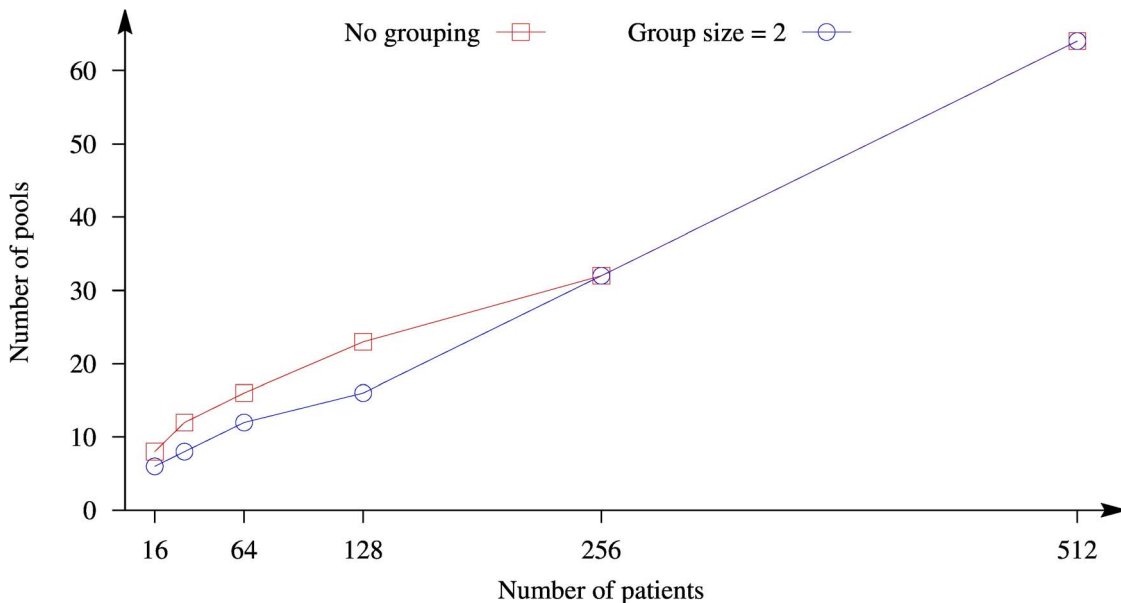


Figure 3. Dependence of the minimum number of pools on the total number of patients using Transposition strategy with $m_{max} = 16$.
doi:10.1371/journal.pone.0104992.g003

- minimize the cost C of sequencing, and
- maximize the probability π to detect the mutation.

It is assumed that the probability $\pi=1$ if the number (m) of patients in the largest pool is not greater than the maximum number allowed (m_{max}), and π is zero if m is greater than a given number \bar{m}_{max} . If the number of patients in the largest pool increases from m_{max} to \bar{m}_{max} , then the probability π decreases from 1 to 0. The value of probability π can be expressed as

$$\pi = \begin{cases} 0 & \text{if } m > \bar{m}_{max}; \\ 1 & \text{if } m \leq m_{max}; \\ 1 - \frac{m - m_{max}}{\bar{m}_{max} - m_{max}} & \text{otherwise.} \end{cases} \quad (10)$$

When objectives are conflicting, usually it is impossible to find a single solution, which would be the best according to both objectives. The solution which is the best by one objective can be worse, or even the worst, by another objective [5]. In such a case it is usual to find a set of compromising solutions, which are all optimal in some sense, thus providing to decision maker an opportunity to make a choice.

In general, two different solutions can be related to each other by the dominance relation. It is said that solution x dominates solution y (denoted by $x \succ y$) if, and only if [6]

- solution x is not worse than solution y according to all objectives, and
- solution x is strictly better than solution y according to at least one objective.

If $x \succ y$ then solution x is called *dominator* of solution y . A solution with no dominators is called *non-dominated* (or *Pareto-optimal*), and the set of such solutions is called *Pareto set* [7]. The corresponding set in the objective space (values of the objective functions) is called *Pareto front*. For more detailed analysis of the Pareto optimality, reader can refer to [8–10].

Since we deal with two objectives one of which is subject to minimize (the cost) and another – to maximize (the probability), the shape of Pareto front of the problem should be similar to one illustrated in Figure 4. One can see in the figure, the solution corresponding to the largest probability, corresponds to the largest cost of the experiment as well. On the other side the solution corresponding to the cheapest experiment, corresponds to the lowest probability as well. However it is possible to find a set of intermediate solutions which are all optimal in Pareto sense, and decision maker can choose the most appropriate one, taking into account various considerations, such as the given budget or/and the desired probability to detect the mutations, in our case.

Results

Experiments on Pooling without Replica

The first experiment is aimed at comparing the cost of NGS pooled experiments without replica. It is expected to detect $N_m = 5$ rare mutations per patient, considering the cost of NGS sequencing $c_1 = 1000$ per pool, and $c_2 = 8$ the cost of each Sanger sequencing to check whether a found mutation belongs to a patient. This ratio between costs is only indicative, and subject to vary over time. The total number of patients is chosen to be $n \in \{32, 64, 96, \dots, 1024\}$, and the maximum number of patients in a pool is assumed to be $m_{max} \in \{4, 8, 16, 24\}$. For each value of m_{max} the lowest number of pools is selected so that the maximum

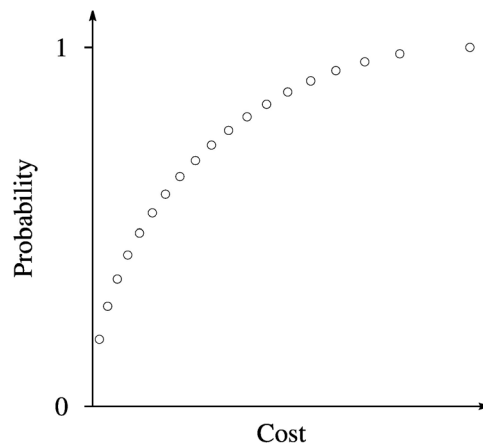


Figure 4. Illustration of Pareto front.
doi:10.1371/journal.pone.0104992.g004

number of patients in a pool would not exceed:

$$p_{min} = \lceil \frac{n}{m_{max}} \rceil. \quad (11)$$

It is assumed that pools are composed by patients so that the difference between the largest and the lowest pools would be minimal, i.e. not greater than 1; e.g. if $n = 17$ and $m_{max} = 4$, then 5 pools should be used, where 2 of them consist of 4 patients, and the remaining 3 of 3 patients. The results, obtained using p_{min} pools, are compared with the results, obtained using the optimal number of pools (p_{opt}), which is determined by the complete enumeration with respect to minimization of the cost:

$$p_{opt} = \arg \min C_N(p), \quad (12)$$

where:

$$\lceil \frac{n}{m_{max}} \rceil \leq p \leq n. \quad (13)$$

Results of the experiment are presented in Figure 5, where the horizontal axis corresponds to the number of patients and the vertical one to the cost of the pooled experiment. Different curves correspond to the different value of m_{max} , for which a minimal number p_{min} of pools is chosen, except the continuous curve (Opt. pooling), where the optimal number p_{opt} of pools is determined. The figure shows that the total cost of the sequencing increases together with the increment of the maximum number of patients in a pool, when the lowest number of pools (p_{min}) is used. However the lowest cost of the sequencing is obtained using the optimal number (p_{opt}) of pools, as it is illustrated by the black curve in Figure 5. The results, obtained using optimal values of m_{max} , are the same as those obtained with $m_{max} = 24$, except the case of $m_{max} = 4$ when the cost of the sequencing is slightly higher. This can be explained by the numerical results of the experiment, performed with $m_{max} = 24$. The optimal numbers p_{opt} of pools and the respective optimal numbers m_{opt} of patients in a pool using different numbers of patients are presented in Table 1 together with the number of Sanger tests needed and the total cost of the sequencing. The table shows that the optimal number of pools

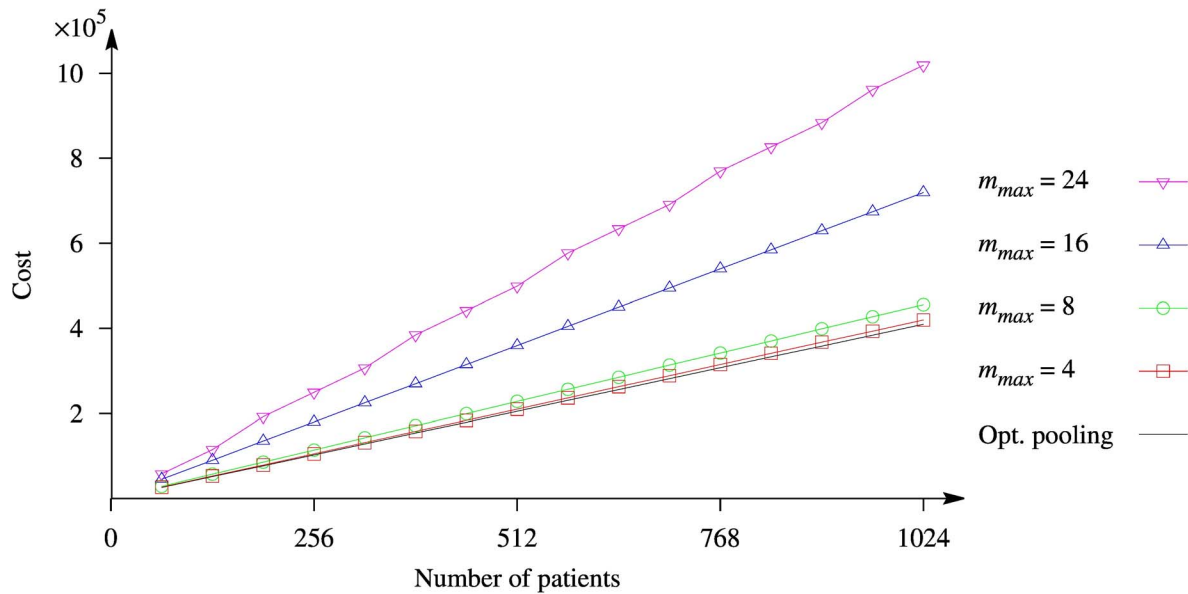


Figure 5. The cost of the sequencing versus the number of patients when allocating the maximum number m_{max} of patients per pool without any replica.
doi:10.1371/journal.pone.0104992.g005

increases when the total number of patients increases, however the optimal number of patients in a pool remains almost stable and very much less than the maximum number of patients in a pool ($m_{max} = 24$).

Experiments on Pooling with Replica

The next experiment is aimed at comparing the cost of the sequencing using two replica strategies, which lead to the sequencing without any indecision or indecision between a group of patients. The same values of m_{max} are used as they were in the previous computational experiment. The complete enumeration is used to determine the optimal numbers of main (p_m) and control (p_c) pools for Transposition strategy (see (5)), whereas the lowest number p such as $p \cdot (p - 1) > 2n$ of pools is chosen for OptReplica strategy (see (8)). The results are presented in Figure 6, where the hollow marks stand for the cost obtained using Transposition strategy and the filled marks for the cost obtained by OptReplica; different curves correspond to different value of m_{max} . The line without marks stands for the cost of the sequencing obtained with pooling without replica into pools of optimal size.

The figure illustrates that the costs of the sequencing decrease when the maximum number of patients in a pool increases, independent whether Transposition or OptReplica strategy is used. The slight difference in cost between the strategies can be distinguished when a small number of patients are pooled into small pools and the OptReplica is then slightly better. It is also clear from the figure that usage of a replication decreases the cost of the sequencing.

The impact of grouping of patients to the total cost of the sequencing using Transposition strategy is illustrated in Table 2, where the first column is the total number of patients, the second column is the cost of the sequencing without grouping of patients, and the last column is the cost of the sequencing, when patients are grouped into groups of 2. One can see in the table that it is useful to use the grouping of patients if the number of patients is small (16 and 32).

The comparison of the total cost of the pooled sequencing with replication using Transposition and OptReplica strategies is

presented in Figure 7. The figure shows the total cost, obtained using Transposition strategy is higher than the cost, obtained using OptReplica strategy, however a difference appears when the number of patients is small. Both curves have refractive points where curves start rising much faster. These points correspond to the number of patients whose optimal allocation requires to assign more patients in a pool than it is allowed by m_{max} . Therefore, it is necessary to increase the number of pools, thus increasing the number of NGS tests.

Solution of the Multi-objective Optimization Problem

The following experiment is aimed at solution of multi-objective optimization problem, where the probability to find the mutation using larger pools is simultaneously considered. The computations are performed for $n \in \{512, 1024, 2048, 4096\}$ patients, considering the extreme values $m_{max} = 24$ and $\bar{m}_{max} = 50$, which are determined using previously developed model for the evaluation of the probability to detect a mutation in pooled experiments, when the number of patients increases [11]. The fixed error rate $P_e = 0.01$ is used as it is a realistic approximation of what is found

Table 1. Cost of experiments and total number of Sanger tests needed to detect $N_m = 5$ mutations in each patient, when using optimal number p_{opt} of pools without replica, and m_{opt} patients in a pool, for a fixed total number n of patients.

n	p_{opt}	m_{opt}	Number of Sanger tests	Cost
128	26	5	3160	51280
256	51	6	6430	102440
512	102	6	12860	204880
1024	205	5	25580	409640

doi:10.1371/journal.pone.0104992.t001

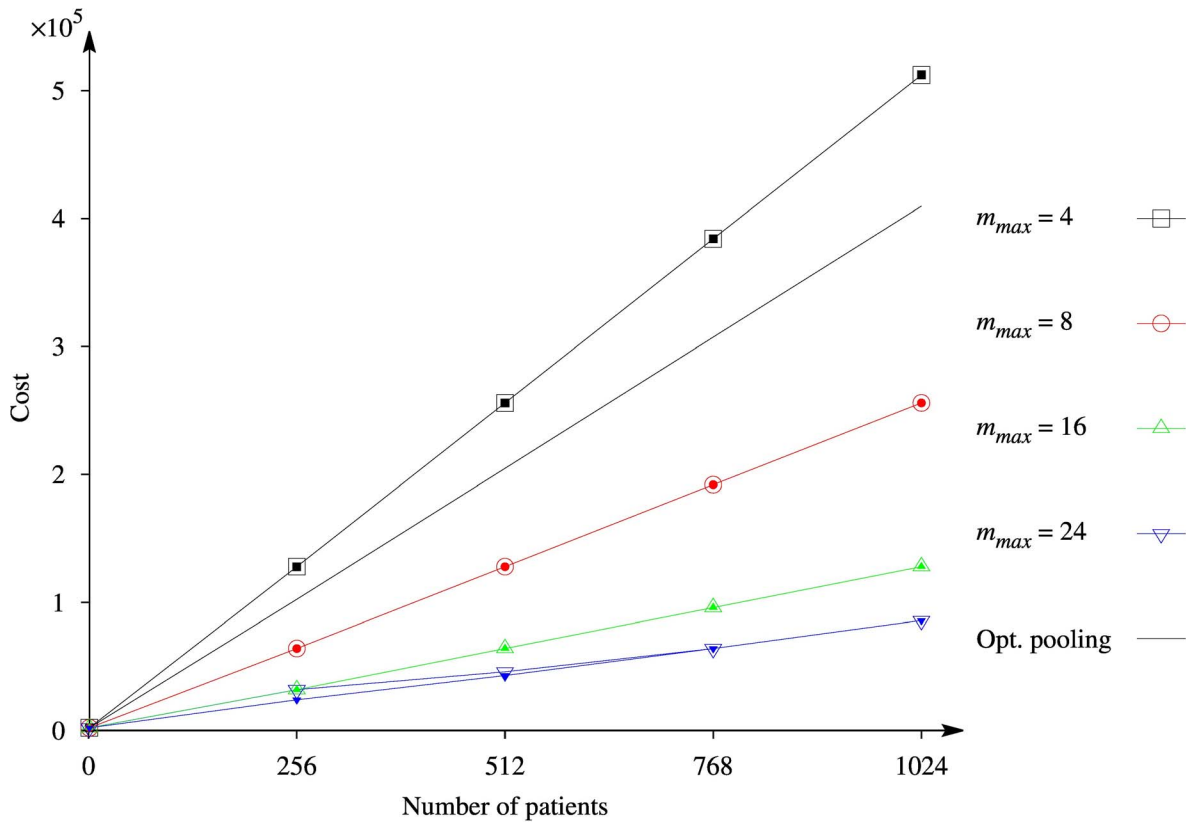


Figure 6. Comparison of the cost of NGS pooled experiment using different values of m_{max} and different strategies of pooling: Transposition (hollow marks) and OptReplica (filled marks).
doi:10.1371/journal.pone.0104992.g006

in this kind of experiments. Since the number of short reads affected by error is independent from the number of patients, in the worst case the signal produced by a SNP can be lost in the noise produced by the experiment. These values are realistic approximations, obtained using real data of pooled experiments.

Pareto fronts, obtained by the complete enumeration of values of p_m and p_c for Transposition strategy and p for OptReplica, are illustrated in Figure 8, where the horizontal axis corresponds to the cost of the sequencing, the vertical one to the probability of detecting the mutations, and different plots to different number of patients. One can see from the figure that the Pareto front consists of a single point using Transposition strategy, and several points using OptReplica strategy when $n = 512$. It is natural since all 512 patients can be allocated into such a number p_m of main pools and a number p_c of control pools so that the cost of NGS test would be minimal and without violation of the maximum pool size (e.g. $p_m = 22$ and $p_c = 24$). Meanwhile the optimal number of pools for OptReplica strategy is $p = 33$ ($33 \times 32 \geq 2 \times 512$), which corresponds to $m = 32 > m_{max}$. Therefore, a trade off is possible between increasing the number of pools thus increasing the total cost, or exceeding m_{max} , thus decreasing the probability to find the mutations.

When the number of patients is larger, the size of the Pareto front increases, although the Pareto fronts obtained using OptReplica strategy are larger in all cases. Moreover, the cost obtained using OptReplica strategy is always not greater than the cost obtained using Transposition strategy considering the fixed probability. Therefore we can conclude that usage of OptReplica strategy leads to a better solution of the formulated multi-objective optimization problem, comparing to Transposition strategy.

Discussion

Since the costs of NGS experiments change over time, we decided not to use a currency unit, such as dollar or euro, but rather a ratio of costs between low and high throughput technologies. We fixed a cost of 1000 for one pooled NGS experiment and 8 for each Sanger sequencing.

In case of no replica, small pools decrease costs

In case of experiments without replica, all found mutations need to be assigned to patients present in the pool. As shown in equation (3), the number of Sanger sequencing increases quadratically with the number of pooled individuals, and it is proportional to the number of expected mutations. Therefore, even if the cost of NGS versus Sanger sequencing is 1000:8, it is better to decrease the size of each pool, rather than increasing the number of patients in each

Table 2. Impact of grouping of the patients to the total cost of the sequencing with replication using Transposition strategy.

n	No grouping	Grouping by 2
16	8000	7280
32	12000	10560
64	16000	17120
128	23000	26240

doi:10.1371/journal.pone.0104992.t002

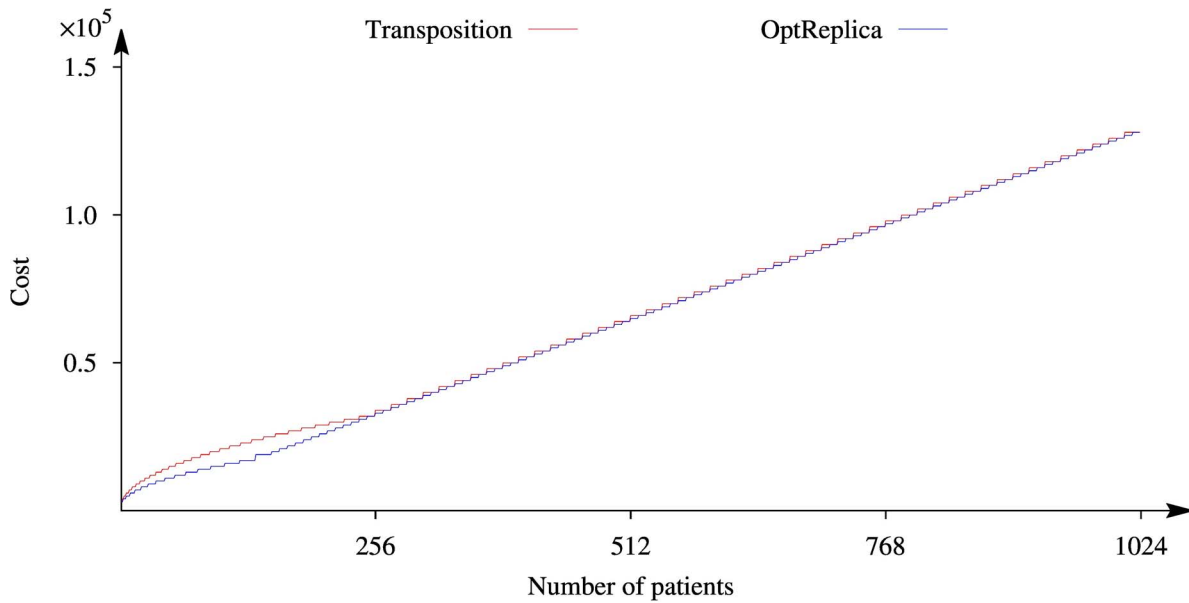


Figure 7. Comparison of cost of sequencing using Transposition and OptReplica strategies with $m_{max} = 16$.
doi:10.1371/journal.pone.0104992.g007

pool. In the present setting, the optimal number of patients per pool is 5 or 6.

Replication of pools decreases costs

When replicated pools are used, the situation is just the opposite with respect to unreplicated pooling: the experiments are less expensive when the number of patients per pool increases. In Figure 6, the only case in which pooling is more expensive that

Optimal pooling without replica, is when there are only four patients per pool. Increasing this number, the reduction of costs is significant, and makes the pooling strategy competitive with respect to unreplicated pools. In both Figures 5 and 6, we report the costs of unreplicated experiments with optimal allocation of patients per pool (Opt. pooling). When comparing the two strategies, it is evident that replicating pools is always the best option. Both Transposition and OptReplica achieve similar results, although in some cases the latter achieves lower costs.

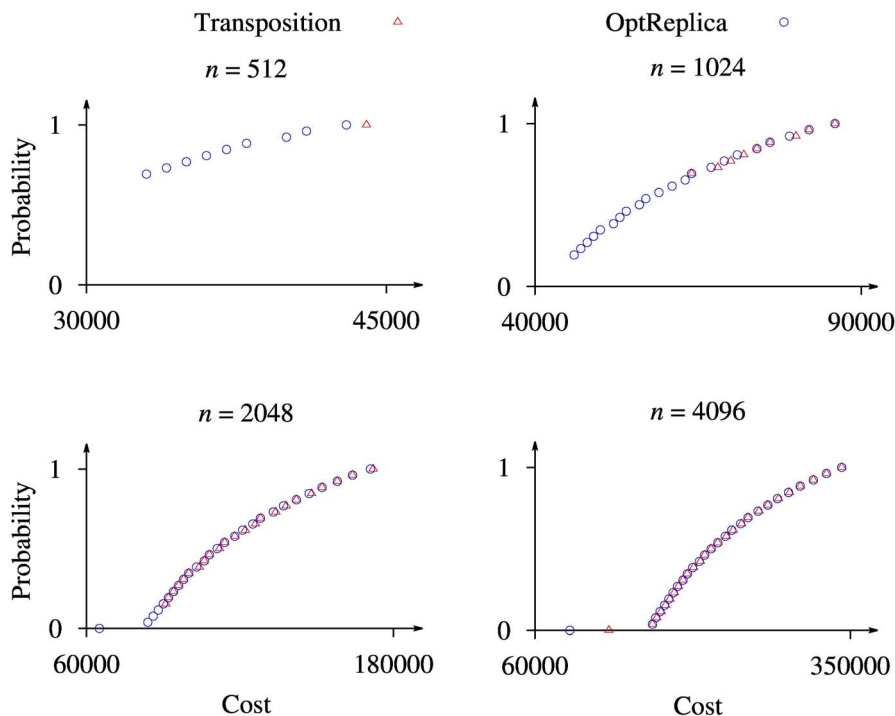


Figure 8. Illustration of Pareto fronts obtained by the complete enumeration with different values of n .
doi:10.1371/journal.pone.0104992.g008

Grouping of patients can reduce the number pools, but Sanger tests must be additionally used to solve the indecision between patients in an identified group. Therefore, grouping can reduce costs only when the total number of patients is small.

Trade-off between costs and detection probability

When the total budget for sequencing is limited, multi-objective optimization might provide useful information regarding the plan of the experiment. With the proposed model, it is possible to predict the probability to detect mutations, given the number of patients and the costs. In some cases, the difference between the available budget and the required budget might marginally affect the probability to successfully detect mutations. For example, in case of $n=4096$ patients (Figure 8), the total cost of all experiments is 342000. If the available budget is only 316000, then the probability to detect mutation is still 0.92, which might be acceptable.

References

1. Kaplan J (2012) The 2013 version of the gene table of monogenic neuromuscular disorders. *Neuromuscular Disorders* 22: 1108–1135.
2. Calvo SE, Tucker EJ, Compton AG, Kirby DM, Crawford G, et al. (2011) High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nature Genetics* 42: 851–860.
3. Futschik A, Schlotterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* 186: 207–218.
4. Wang T, Pradhan K, Ye K, Wong IJ, Rohan TE (2011) Estimating allele frequency from next-generation sequencing of pooled mitochondrial dna samples. *Frontiers in genetics* 2.
5. Deb K (2005) Multi-objective optimization. In: Burke EK, Kendall G, editors, *Search Methodologies*, Springer US. pp. 273–316.
6. Chinchuluun A, Pardalos PM (2007) A survey of recent developments in multiobjective optimization. *Annals of Operations Research* 154: 29–50.
7. Voorneveld M (2003) Characterization of Pareto dominance. *Operations Research Letters* 31: 7–11.
8. Luc DT (2008) Pareto optimality. In: Chinchuluun A, Pardalos PM, Migdalas A, Pitsoulis L, editors, *Pareto Optimality, Game Theory And Equilibria*, Springer New York, volume 17 of *Springer Optimization and Its Applications*. pp. 481–515.
9. Chinchuluun A, Pardalos PM, Migdalas A, Pitsoulis L, editors (2008) *Pareto Optimality, Game Theory And Equilibria*, volume 17 of *Springer Optimization and Its Applications*. Springer New York.
10. Zopounidis C, Pardalos PM, editors (2010) *Handbook of Multicriteria Analysis*, volume 103 of *Applied Optimization*. Springer Berlin Heidelberg.
11. Ferraro MB, Guarracino MR (2013) Prediction of single-nucleotide causative mutation for muscular diseases. In: *Computational Intelligence in Bioinformatics and Biostatistics*, Springer New York, *Lecture Notes in Bioinformatics*.

Conclusions

In this work we propose and study novel strategies for planning the NGS experiments of a large number of patients. We provide optimal configurations of the experiments, in terms of sequencing costs. We develop two different replica strategies, both providing lower overall costs. Finally, we investigate the solution to this problem when not only the cost, but also the probability to detect a mutation, is taken into account.

Author Contributions

Conceived and designed the experiments: JŽ AL MRG. Performed the experiments: AL. Analyzed the data: JŽ MRG. Contributed to the writing of the manuscript: JŽ AL MRG.