Bioinformatics, 31(8), 2015, 1199–1203
doi: 10.1093/bioinformatics/btu805
Advance Access Publication Date: 4 December 2014
Original Paper

OXFORD

Sequence analysis

Semiconductor sequencing: how many flows do you need?

Jan Budczies[1,2,3,*], Michael Bockmayr[1], Denise Treue[1], Frederick Klauschen[1] and Carsten Denkert[1,3]

[1]Institute of Pathology, Charité University Hospital, Charitéplatz 1, 10117 Berlin, Germany, [2]German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany and [3]German Consortium for Translational Cancer Research (DKTK), Berlin partner site, Charitéplatz 1, 10117 Berlin, Germany

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on September 9, 2014; revised on November 7, 2014; accepted on December 1, 2014

Abstract

Motivation: Semiconductor sequencing directly translates chemically encoded information (A, C, G or T) into voltage signals that are detected by a semiconductor device. Changes of pH value and thereby of the electric potential in the reaction well are detected during strand synthesis from nucleotides provided in cyclic repeated flows for each type of nucleotide. To minimize time requirement and costs, it is necessary to know the number of flows that are required for complete coverage of the templates.

Results: We calculate the number of required flows in a random sequence model and present exact expressions for cumulative distribution function, expected value and variance. Additionally, we provide an algorithm to calculate the number of required flows for a concrete list of amplicons using a BED file of genomic positions as input. We apply the algorithm to calculate the number of flows that are required to cover six amplicon panels that are used for targeted sequencing in cancer research. The upper bounds obtained for the number of flows allow to enhance the instrument throughput from two chips to three chips per day for four of these panels.

Availability and implementation: The algorithm for calculation of the flows was implemented in R and is available as package ionflows from the CRAN repository.

Contact: jan.budczies@charite.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

For more than 25 years, the chain-termination method developed by Sanger et al. (1977) was the predominating method for DNA sequencing. Pyrosequencing was one of the first methods for DNA sequencing that is based on sequencing by synthesis rather than chain termination (Ronaghi *et al.*, 1998). Since the late 1990s, further next-generation sequencing (NGS) methods have been developed and a technology development race started to achieve the $1000 genome. Most of the NGS technologies such as 454 pyrosequencing (Margulies *et al.*, 2005; Wheeler *et al.*, 2008) or Illumina sequencing (Bentley *et al.*, 2008) rely on the generation of light signals upon incorporation of one of the four nucleotides A, C, G or T into the new DNA strand. In contrast, semiconductor sequencing (Rothberg *et al.*, 2011) uses a different approach that is based on direct measurement of electric potential changes induced by the release of protons (H+) during strand synthesis.

In semiconductor sequencing, the reaction wells are sequentially flooded by one of the nucleotides A, C, G and T. Whenever the offered nucleotide is complementary to the template DNA strand at the actual position, it is incorporated into the new strand and induces a change in the electric potential in the reaction well. Using the scalable technology of ion-sensitive field-effect transistors (ISFET) organized in highly integrated circuits, millions of wells can be interrogated at the same time. An ISFET detects a voltage signal,

only if the current nucleotide is complementary to the template at the actual position, otherwise no signal is induced. If homopolymer repeats are present in the template, multiple nucleotides are incorporated during a single flow. As incorporation of two nucleotides into the new DNA strand doubles the voltage signal and incorporation of multiple nucleotides into the new strand multiplies the voltage signal, the length of the homopolymer stretch can be determined based on the voltage.

Ion Torrent technology (Thermo Fisher Scientific, Inc.) allows sequencing of templates of a length of up to 200 bp or 400 bp using a novel chemistry. The Ion Personal Genome Machine (PGM) system is suited for targeted sequencing of disease relevant regions of the human genome or complete sequencing of smaller genomes, e.g. of microbes. Ion PGM Sequencing 200 kit includes reagents for sequencing of amplicons up to 200 bp. Using this sequencing kit, a total number of 1000 flows can be executed, which are usually split to 2×500 and running of two sequencing chips.

Here, we present an approach to compute the number of flows required to sequence a DNA template of length n. These are the main results of our study: (i) while $k = 3n + 1$ flows are needed to sequence each template of length n to the full length, in most situations, a much lower number of flows suffices. (ii) In a random sequence model, we calculate the cumulative distribution function of the number of required flows and its expected value and variance. (iii) We provide an algorithm to calculate the number of required flows for a concrete amplicon panel from a BED file of genomic positions.

In semiconductor sequencing, the sequencing depth and the number of flows are independent quantities. We present a formula to calculate the sequencing depth from the number of amplicons and the number of samples that are sequenced in parallel.

2 Materials and Methods

2.1 Stochastic model

We construct a stochastic model to calculate the number of required flows to sequence a template of length n. The four nucleotides A, C, G and T flow into the reaction well in a defined order that is repeated until the maximal number of flows k is reached (Fig. 1A). A special situation occurs, when a homopolymer in the template is reached. In this situation, the complementary homopolymer is assembled during the one flow, leading to stronger change of pH value and voltage. In the stochastic model, this situation can be accounted for by not considering the assembly of the homopolymer in one step, but by treating the incorporation of each of the nucleotides separately. Using this trick, there are four different alternatives for assembly at each position of the template: The nucleotide is incorporated after 0 flows (if the nucleotide is a repeat of the nucleotide before) or the nucleotide is incorporated after one to three flows (if the nucleotide is not a repeat of the nucleotide before). Figure 1A illustrates the ongoing strand synthesis during four consecutive flows and each of the four alternatives. Obviously, each of the four alternatives occurs with same probability $P = \frac{1}{4}$.

2.2 Calculation of the distribution function

As main ingredient for the stochastic model, we introduce four random variables X_0, X_1, X_2 and X_3 describing the number of positions in the template, where 0, 1, 2 or 3 flows are needed for sequence assembly at this position. Completing the model, we assume that the random variables sum to up to the total number of nucleotides, $X_0 + X_1 + X_2 + X_3 = n$ and are distributed according to the multinomial distribution with equal probabilities $P_0 = P_1 = P_2 = P_3 = \frac{1}{4}$,

$$P(X_0 = n_0, X_1 = n_1, X_2 = n_2, X_3 = n_3) = \frac{n!}{n_0!n_1!n_2!n_3!}\left(\frac{1}{4}\right)^n. \quad (1)$$

After these preparations, the number of required flows to sequence the template of length n is given by $F(X_0, X_1, X_2, X_3) = 0X_0 + 1X_1 + 2X_2 + 3X_3 + 1$. The cumulative distribution function of F can be calculated as

$$P(F \le k) = \sum_{n_1=0}^{n}\sum_{n_2=0}^{n-n_1}\sum_{n_3=0}^{n-n_1-n_2} \theta(n_1 + 2n_2 + 3n_3 + 1 \le k) \times \frac{n!}{(n-n_1-n_2-n_3)!n_1!n_2!n_3!}\left(\frac{1}{4}\right)^n, \quad (2)$$

with the step function $\theta(x \le x_0)$ equaling 1 if $x \le x_0$ and 0 if $x > x_0$.

Further, the expected value of the number of required flows can be inferred from the expected value of the multinomial distribution $\mathrm{E}(X_i) = \frac{1}{4}n$ as

$$\mathrm{E}(F) = \sum_{i=0}^{3} i\mathrm{E}(X_i) + 1 = \frac{3}{2}n + 1. \quad (3)$$

Finally, inserting variance $\mathrm{Var}(X_i) = \frac{3}{16}n$ and covariance $\mathrm{Cov}(X_i, X_j) = -\frac{1}{16}n$ of the multinomial distribution, we obtain the variance of the number of required flows as

$$\mathrm{Var}(F) = \sum_{i=0}^{3} i^2\mathrm{Var}(X_i) + 2\sum_{0 \le i < j \le 3} ij\,\mathrm{Cov}(X_i, X_j) = \frac{5}{4}n. \quad (4)$$

Thus, we obtained exact expressions for expected value and variance of the number of required flows.

The cumulative distribution function was evaluated and visualized using the statistical language R (R Core Team, 2014) and the R package ggplot2 (Wickham, 2009).

2.3 Simulation of flows

An algorithm for simulation of the nucleotide flows and sequence assembly was implemented in R and is available as package *ionflows* from the CRAN repository (Bockmayr and Budczies, 2014). A BED file with the genomic positions of the amplicons is taken as input and a list with the number of required flows for each of the amplicons is delivered as output. The number of flows is calculated separately for sequencing in forward direction and sequencing in backward direction. As an example, the number of required flows was calculated for six amplicon panels for targeted sequencing in cancer research.

2.4 Sequencing depth

In semiconductor sequencing, parallel sequencing takes place in millions of wells of a sequencing chip. Each well contains a single magnetic bead with clonal DNA attached. Typically, using a single chip, multiple target sequences (amplicons) and several samples are sequenced in parallel. In an ideal situation, all samples are covered with the same depth and all amplicons are covered with the same depth. In this situation, the sequencing depth d can calculated as

$$d = \frac{e\,w}{m\,a}, \quad (5)$$

where e is the efficiency in using the chip (the percentage of wells with suitable beads), w is the number of wells on the chips, m is the number of samples (multiplexing) and a is the number of amplicons

A

	Well 1	Well 2	Well 3	Well 4
Previous flow: T	TAAAACCA AT|TTT →	GATAGCGA CT| →	CAGCGTGA GT| →	TACGTTAA AT| →
First flow: A	TAAAAGCA AT|TTT	GATAGCGA CT|A	CAGCGTGA GT|	TACGTTAA AT|
Second flow: C	TAAAAGCA AT|TTTC	GATAGCGA CT|A	CAGCGTGA GT|C	TACGTTAA AT|
Third flow: G	TAAAAGCA AT|TTTCG	GATAGCGA CT|A	CAGCGTGA GT|CG	TACGTTAA AT|G

B

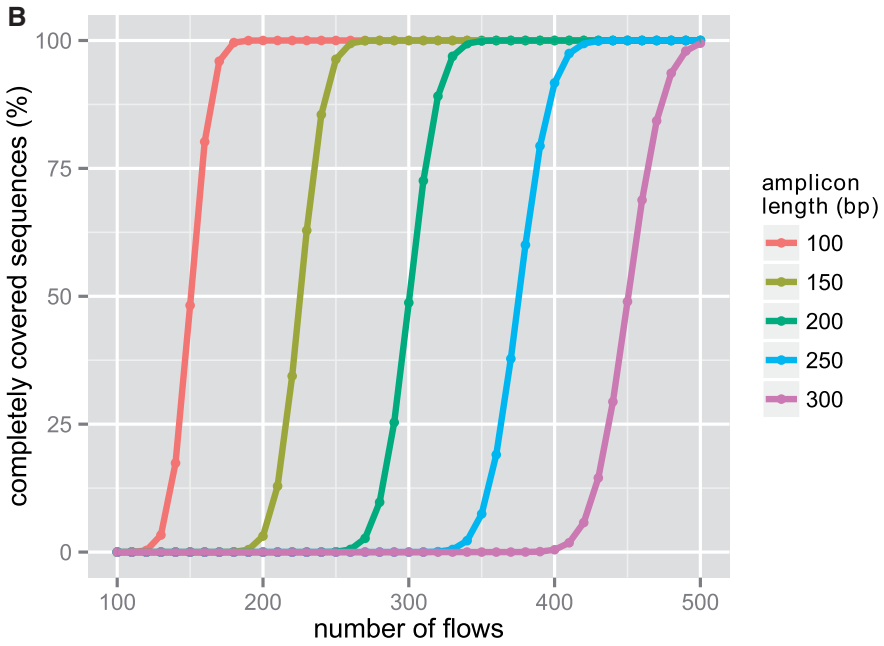

Fig. 1. (**A**) Schematic diagram of semiconductor sequencing. For construction of a stochastic model, we consider the situation where a nucleotide at a fixed position of the template (the next one right to the cursor position) is sequenced. There are four different possibilities: the nucleotide is a repeat of the nucleotide that was sequenced before (well 1), the nucleotide is sequenced during the next flow (well 2), the nucleotide is sequenced during the next but one flow (well 3) or the nucleotide is sequenced during the third flow (well 4). Accordingly, the number of additional flow to sequence the nucleotide is 0, 1, 2 or 3, and each of these possibilities occurs with the same probability $P = \frac{1}{4}$. (B) Cumulative distribution function of the number of required flows in a random sequence model. The percentage of random sequences that can be completely sequenced for an given number of flows is shown for templates of lengths of 100 bp, 150 bp, . . , 300 bp. For sequences of length n, the distribution of required flows is centered at the expected value $\frac{3}{2}n + 1$ and has variance $\frac{5}{4}n$

in the panel. For the Ion Torrent technology, the total efficiency e is a product of the four efficiency number in the Ion Sphere Particle (ISP) summary: the percentage of wells with ISPs, the percentage of sufficiently enriched ISPs, the percentage of clonal ISPs and the percentage of fragments suitable for the final library. In a typical example of targeted deep sequencing using the Ion 318 Chip (11.1 million wells, efficiency 50%), 8-fold multiplexing and a panel of 150 amplicons, a sequencing depth of 4625 can be obtained.

3 Results and discussion

A stochastic model was developed to calculate the number of flows required for semiconductor sequencing of random sequences (Fig. 1A). As it can be obtained from the formulas derived in Section 2, $151 \pm 11, 301 \pm 16$ and 451 ± 19 (expected value $\pm$ standard deviation) flows are required for sequencing of templates of length 100 bp, 200 bp and 300 bp. In the most unfavorable situation, $k = 3n + 1$ flows are required for sequencing of a template of length n. By contrast, the distribution function of the number of flows is centered at $k = \frac{3}{2}n + 1$ with a relatively small variance of $\frac{5}{4}n$. Figure 1B shows the percentage of random sequences that are completely covered in dependence of the number of applied flows for templates of the length 100 bp, 150 bp, . . . , 300 bp.

In a typical experimental setup, e.g. using the Ion Torrent technology, a total number of 1000 flows is available, which is split into two halves (500 flows per chip) by default. Alternatively, it is possible to split the 1000 flows in three-thirds (330 flows per chip). Applying our random sequence model to this situation, we obtain the following results: using the split into two halves, 99.5% of random sequences of length 300 bp can be completely sequenced, whereas $> 99.9999\%$ of sequences of length 250 bp or shorter can be completely sequenced. Using the split into three-thirds, 96.9% of sequences of length 200 bp can be completely sequenced, whereas $>$ 99.9999% of sequences of length 150 bp or shorter can be completely sequenced.

Formalin fixation and subsequent paraffin embedding represent the standard method for tissue fixation and storage in the diagnostic pathology workflow. Large collections of formalin-fixed,

Table 1. The number of required flows for six amplicon panels for targeted sequencing in cancer research. For each of the amplicons in a panel, the target sequence was retrieved and the flows were simulated for both, sequencing in forward direction and sequencing in backward direction. Values for amplicon length and for numbers of flows should be read as follows: mean value (minimum value – maximum value). The number of amplicons that require more than 300 flows for complete coverage are shown in the last column

Panel name	Panel type	DNA type	Number of primer pools	Number of amplicons	Amplicon length (bp)	Number of required flows	Amplicons > 300 flows
Comprehensive Cancer Panel	Ion Torrent	FFPE	4	15 992	109 (50–187)	167 (59–292)	0
Cancer Hotspot Panel v2	Ion Torrent	FFPE	1	207	106 (50–141)	165 (64–240)	0
Colon and Lung Cancer Panel	Community	FFPE	1	92	111 (52–138)	173 (72–232)	0
TP53 Panel	Community	FFPE	2	24	107 (72–138)	164 (101–244)	0
AML Research Panel	Community	Standard	4	264	129 (68–216)	200 (96–340)	6
BRCA1 and BRCA2 Panel	Community	Standard	3	167	145 (71–242)	211 (100–365)	8

paraffin-embedded (FFPE) tissues available at pathology laboratories are an invaluable resource for clinical research. However, DNA extracted from FFPE tissues is fragmented and chemically modified, which renders its use challenging for molecular studies (Budczies *et al.*, 2011; Hedegaard *et al.*, 2014). To cope with DNA fragmentation in sequencing studies, the length of fragments in the library is usually chosen to be restricted to a length <150 bp. In such situations, the split of the 1000 flows into three-thirds represents a convenient opportunity as we demonstrate in the above example of the Ion Torrent technology.

Additionally, we implemented a simulation algorithm to calculate the number of required flows to sequence a concrete panel of genomic DNA sequences. Using this code, we analyzed six cancer panels for targeted sequencing that are publicly available (Table 1). For each of the panels, simulations were done to exactly calculate the number of required flows for each of the amplicons (Supplementary Material S1–S6). Four of the panels (Comprehensive Cancer Panel Cancer, Cancer Hotspot Panel, Colon and Lung Cancer Panel and TP53 Panel) were designed for the analysis of DNA from FFPE tissues and included for the most parts amplicons of length <150 bp. Accordingly, <300 flows were sufficient to cover each of the amplicons of these panels. Thus, for all of these panels, the 1000 flows can be split into three-thirds, significantly saving operating time and costs. Both of the other panels (AML Research Panel and BRCA1 and BRCA2 Panel) included amplicons of length >200 bp. Accordingly, there were a few (6 and 8) amplicons that required more than 300 flows for complete sequencing.

In targeted sequencing, multiplexing helps to enhance sample throughput. Multiplexing can be done by adding barcodes to the target sequences that uniquely label the samples. For the Ion Torrent platform, 10-bp barcodes are available that can be used to label up to 16 or up to 96 samples. As the barcodes are sequenced together with the target sequences, extra flows are needed for barcode sequencing. For 10 bp barcodes, maximal 30 flows are needed for barcode sequencing. Thus, more than 300 flows are available for sequencing of the targets when a split of 1000 flows into three-thirds is used.

The 1000 available flows can be either used for two chips (standard protocol) or for three chips (modified protocol). Our simulations show that 300 flows suffice to sequence all amplicons for each of the first four panels in Table 1. Thus, the modified protocol instead of the standard protocol can be used in these cases. Actually, the first four panels were designed for FFPE tissues, whereas the last two panels were designed for frozen tissues, include amplicon lengths >150 bp and therefore require a higher number of flows. For the last two panels, the standard protocol needs to be used to ensure that all amplicons can be completely sequenced. The benefit of the modified protocol compared the standard protocol is to decrease both costs and instrument times: One third of the costs for sequencing kits can be saved and sequencer throughput can be enhanced from two chips per day to three chips per day.

In semiconductor sequencing, the sequencing depth and the number of flows are independent quantities. In Section 2, we derived a formula to calculate the sequencing depth from the number of amplicons and the number of samples. However, in practice, different amplicons in a panel are covered to a varying extend. Often, there are several amplicons that perform considerably worse than the average. There are recommendations of the chip manufacturer available, how many samples should be run in parallel to obtain 30× or 500× coverage for 95% of the amplicons in a panel (Life Technologies, 2014).

In summary, we presented an approach to determine the number of nucleotide flows that are required in semiconductor sequencing. In a model of random sequences, exact expressions were presented for the cumulative distribution function of the number of flows, the average number of flows and the corresponding variance. Furthermore, we implemented an algorithm to calculate the number of required flows for each of the items in a concrete list of genomic DNA sequences. In targeted sequencing, our methods allow to calculate the number of required flows for an amplicon panel and thus to optimize time requirements and costs.

Funding

This work was supported by the German Consortium for Translational Cancer Research (DKTK).

Conflict of Interest: none declared

References

Bentley,D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* **456**, 53–59.

Bockmayr,M. and Budczies,J. (2014) *ionflows: Calculate the Number of Required Flows for Semiconductor Sequencing*. R package version 1.1. http://cran.r-project.org/web/packages/ionflows/.

Budczies,J. *et al.* (2011) Genome-wide gene expression profiling of formalin-fixed paraffin-embedded breast cancer core biopsies using microarrays. *J. Histochem. Cytochem.*, **59**, 146–157.

Hedegaard,J. *et al.* (2014) Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. *PLoS One,* **9**, e98187.

Life Technologies (2014) *Ion AmpliSeq DNA and RNA Library Preparation User Guide, Revision B.0 (MAN0006735)*. p. 36. http://ioncommunity.lifetechnologies.com/docs/DOC-3005/.

Margulies,M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–80.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ronaghi,M. *et al.* (1998) A sequencing method based on real-time pyrophosphate. *Science*, **281**, 363, 365.

Rothberg,J.M. *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

Sanger,F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. U S A*, **74**, 5463–5467.

Wheeler,D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.

Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.