Molecular Robots Obeying Asimov's Three Laws of Robotics

Gal A. Kaminka*,**
Bar Ilan University

Rachel Spokoini-Stern†
Augmanity

Yaniv Amir†
Augmanity

Noa Agmon**
Bar Ilan University

Ido Bachelet†
Augmanity

Abstract Asimov's three laws of robotics, which were shaped in the literary work of Isaac Asimov (1920–1992) and others, define a crucial code of behavior that fictional autonomous robots must obey as a condition for their integration into human society. While, general implementation of these laws in robots is widely considered impractical, limited-scope versions have been demonstrated and have proven useful in spurring scientific debate on aspects of safety and autonomy in robots and intelligent systems. In this work, we use Asimov's laws to examine these notions in molecular robots fabricated from DNA origami. We successfully programmed these robots to obey, by means of interactions between individual robots in a large population, an appropriately scoped variant of Asimov's laws, and even emulate the key scenario from Asimov's story "Runaround," in which a fictional robot gets into trouble despite adhering to the laws. Our findings show that abstract, complex notions can be encoded and implemented at the molecular scale, when we understand robots on this scale on the basis of their interactions.

Keywords
Robotics, nanorobotics, Asimov's laws of robotics, DNA origami

1 Introduction

The prolific science fiction author and biochemist Isaac Asimov devised the three laws of robotics during the first half of the 20th century [3]. These laws became central to fictional literary works by him and others who dealt with the foreseen integration of autonomous robots into human society, technology, and culture. The three laws (L) are generally phrased as follows:

L1: A robot may not harm a human being, or allow, by inaction, a human being to come to harm.

L2: A robot must obey orders given by a human being, unless these orders are in conflict with L1.

L3: A robot must protect its own existence, unless doing so is in conflict with L1 or L2.

* Contact author.
** Department of Computer Science, Bar Ilan University, Ramat Gan, Israel. E-mail: galk@cs.biu.ac.il (G.A.K.); agmon@cs.biu.ac.il (N.A.)
† Augmanity, 8 Hamada St., Rehovot 7670308, Israel. E-mail: chaliks@gmail.com (R.S-S.); yaniv.augmanity@gmail.com (Y.A.); dogbach@gmail.com (I.B.)

Together, these laws construct a reasoning process that a robot must undergo before making a decision of any kind (Figure 1). This process filters out any move that violates L1, L2, or L3 (in that order), assuming that no decision (including doing nothing) is allowed to bypass a stage in the process.

Despite their importance for sci-fi plots, it is widely accepted among roboticists that general implementation of Asimov's laws is not practically feasible. However, limited-scope versions of the laws have been developed as part of artificial intelligence research into autonomous decision making [18]. These investigations always assumed that a single complex robot makes its decisions and applies the three-law filter process described above; the versions of the rules were developed to explore decision-making capabilities.

In the past two decades robotics has successfully entered the molecular scale as well. Of particular interest was utilizing DNA as a building block for molecular robots. Three phenomena have been central in this context: toehold-mediated strand displacement [20], which provides actuation in DNA-based devices; the immobile DNA junction [13], which enabled the programmable fabrication of nanostructures by self-assembly, including the techniques collectively known as DNA origami [11]; and aptamers, which sense diverse environmental cues [5]. Combining these components enabled the construction of programmable DNA-fueled actuators [9, 14, 17], bipedal walkers [7, 19], logic circuits and molecular computers [10, 12, 15, 16], a nanoscale factory [6], and robotic devices that link sensing of external cues to actuation, including in biological systems [1, 2, 4].

The technical challenges associated with the design and fabrication of molecular robots are being gradually tackled. However, molecular robots differ extremely from macro-scale robots, and are driven by different basic principles. It is therefore not clear whether and how general paradigms and concepts of robotics can be translated into this scale as well. To examine these questions, we developed an appropriately scoped version of Asimov's laws, implemented using molecular robots. We first focused on designing a mechanism that recognizes harm and responds to it. While harm might be an abstract notion in general, biological harm can be signaled by many molecular phenomena, for example, anomalous extracellular presence of intracellular components such as ATP, or the cell surface expression of early apoptotic markers such as phosphatidylserine. Using

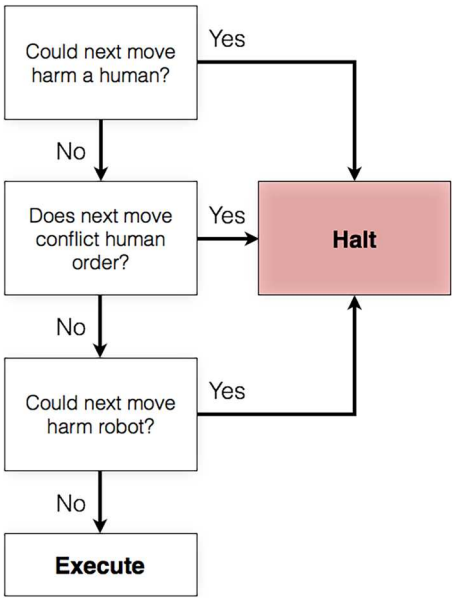

Figure 1. A reasoning process constructed by Asimov's three laws of robotics. A schematic representation of Asimov's laws as a flow diagram. Any move (including doing nothing) proceeds through this decision process without bypassing any stage, with moves that violate either one of the three laws being filtered out in real time.

aptamers or other molecular sensors, molecular robots can be readily programmed to detect these signals and respond accordingly. As a proof of principle, we chose a microRNA molecule (a human miR-16 analogue) as the damage signal for several reasons: It can be detected by a tractable, toehold-based strand displacement sensor [20], it is expressed in many copies in the cell, and its expression profile could tag specific cell types such that not only is damage detected, but also its cellular source. Obviously, miR-16 can be detected only after certain cell damage has occurred, causing its release; however, a population of molecular robots can be programmed to halt upon sensing microscopic damage before it poses a macroscopic threat. This is allowed by the infinitesimal increments of damage caused by single molecular robots.

2 Results

As robots for this study, we used slightly modified versions of the two-state DNA origami robots previously described [1]. These robots interact to carry out logic functions, which served as a starting point for the design work in this study (Supplementary Notes 1 and 2, which can be found in the Online Supplementary Materials here: http://www.mitpressjournals.org/doi/suppl/10.1162/ARTL_a_00235).

Combined, L1 and L2 make a NOT gate, with miR-16 being the input and the effector activity of the robots being the output (Figure 2a). As long as no damage is detected, the robots carry out their task (L2: input = 0, output = 1), but when miR-16 is detected at a certain concentration, they halt (L1: input = 1, output = 0). As shown previously, a NOT gate can be constructed from two robot

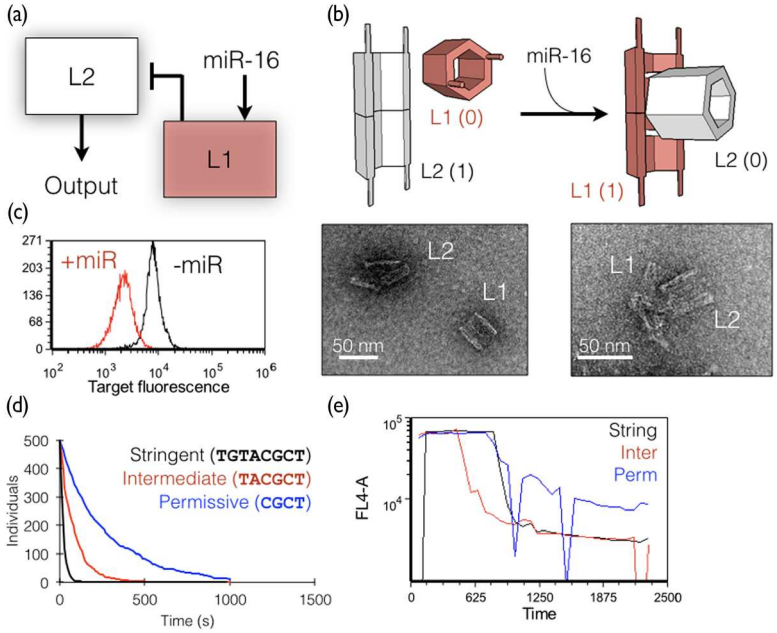

Figure 2. A logical NOT gate responding to damage by turning off. (a) A schematic of the NOT gate constructed by L1 and L2 robots. L1 robots sense miR-16 (an indicator of cellular damage), activate, and expose strands that force L2 robot closure. (b) Schematic and corresponding representative TEM images of robots at either state of the NOT gate. (c) Flow-cytometric analysis of L2 robot output in the presence (red) and absence (black) of miR-16. (d) Visual DSD simulations of the stringency of L1 closure of L2 (sequence 15* in Supplementary Note 3). Shown sequences are those of the clasp arm of L1, which hybridize with L2 gate toeholds at varying kinetics depending on length. (e) Experimental validation of the vDSD simulations (FL4-A, target fluorescence; string, stringent; inter, intermediate; perm, permissive; graph colors correspond to colors in d).

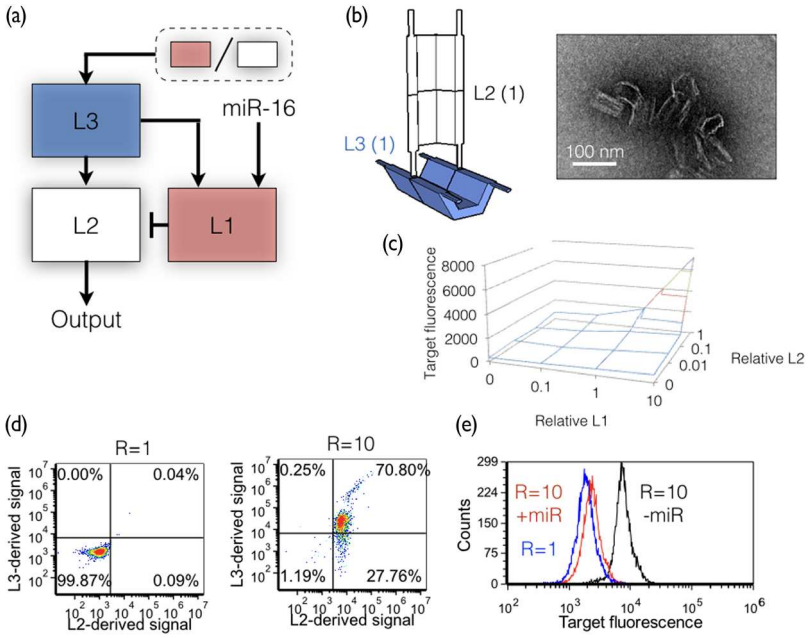

Figure 3. The L1-L2-L3 system that obeys Asimov's three laws of robotics. (a) A schematic of the L1-L2 logical NOT gate integrated with L3 that senses the L1:L2 ratio and provides a physical platform for L1 and L2 to function. (b) A schematic and the corresponding representative TEM image of L2 and L3 robots attached to each other according to the system design. (c) Experimental validation of L3 state (as measured by fluorescent signal on target microparticles) at various L1:L2 ratios, showing response at a ratio of 10:1. (d) Experimental confirmation that L2 robots are not active at a wrong L1:L2 ratio, and that they are attached to L3 robots at the proper ratio, as shown by correlated double staining of target microparticles for L2 and L3 (at FL3 and FL1 channels, respectively). R = 1 and R = 10 denote L1:L2 ratios. (e) Flow-cytometric analysis of target microparticles at three system states: wrong L1:L2 ratio (R = 1, blue), proper L1:L2 ratio without miR-16 (R = 10, –miR, black), proper L1:L2 ratio with miR-16 (R = 10, +miR, red).

types, here termed L1 and L2 robots. This design was successfully implemented with fluorescently tagged, antibody-loaded L2 robots whose output was represented by state-specific attachment to antigen-coated polystyrene beads as targets (Figure 2b,c). Interestingly, Asimov recognized that decisions on damage cannot be simply binary. Rather, robots are expected to carry out tasks given a certain chance for human injury rather than halting immediately, as is the case of the robot surgeons featured in some of Asimov's writings [3]. To implement this, we redesigned the strand displacement reaction to allow varying degrees of stringency, emulating the varying levels of acceptable risk in robot activity. This was confirmed using visual DSD [8] simulations and experiments with real robots (Figure 2d,e).

The proper function of the L1-L2 NOT gate depends on the proper molar ratio between L1 and L2 robots. Excess of L2 robots will lead to exaggerated activity, while excess of L1 robots will repress the system indefinitely. Securing the necessary L1:L2 ratio is analogous to the robotic system protecting itself (L3). To do this, we designed a third type of robot, called an L3 robot, which serves as a physical substrate or scaffold that allows L1 and L2 robots to interact only when they are at a proper ratio (Figure 3a–c, Supplementary Note 3).

Programming L3 robots to respond to a proper L1:L2 ratio is enabled by the basic single robot design, which requires displacement of both gates in order for the robot to open, with the displacement of each arm tuned to occur at a given input concentration. The inputs for the L3 robot are DNA strands attached externally to L1 and L2 robots, meaning that the inputs of L3 are invariably linked to the concentrations of L1 and L2 robots. The large number of robots present relative to the amount actually required to exert an effect allows some L1 and L2 robots in the population to serve as ratio indicators, while others physically interact on accessible L3 platforms to exert effects. We

validated our design using both visual DSD simulations and experiments with real robots in solution at three different states: incorrect L1:L2 ratio, correct ratio without damage, and correct ratio with damage (Figure 3d,e; Supplementary Note 4).

As a final proof of principle, we recreated the classical scenario from "Runaround," a short story by Asimov. First published in 1942, "Runaround" is the first to explicitly list the three laws, and is notable in that, unlike many of Asimov's other writings, the protagonist robot in the story is actually following the laws correctly. It is therefore an interesting test case for molecular robots.

In the story, a robot nicknamed Speedy is sent out in the fatal heat of the surface of the planet Mercury to bring selenium to its human operators. When Speedy fails to return, the humans notice it is moving irregularly and talking nonsense. They then note that its path is roughly circular, and deduce that Speedy's control is stuck at equilibrium between the second and third laws. It turns out that the selenium's position is leaking gas, which may destroy the robot. The command to the robot (second law) was given casually, making it a low-priority command. And as Speedy approaches the danger, the third law (which in Speedy is particularly strengthened because of the cost of the robot) kicks in at a level that exactly counterbalances the urgency of the command given to it. In other words, moving away from the selenium and danger causes the second law to become stronger, driving the robot towards the selenium. But moving closer to the selenium and danger strengthens the third law and drives the robot back. Introducing additional threats or further commands does not work, as the robot simply reaches a new equilibrium. The human operators finally manage to break Speedy from the equilibrium trajectory by putting themselves in danger, causing the first law to kick in and override both other laws.

The "Runaround" scenario has three distinct stages: It begins with L2 dominating, followed by a conflict between L2 and L3, causing an equilibrium to be reached. The equilibrium is terminated by the introduction of another conflict, between L1 and L2, in which L1 overrides L2. Our starting system was therefore composed of L1, L2, and L3 with proper L1:L2 ratio. The L2-L3 conflict was experimentally induced by forcing L3 robots to close using DNA strands complementing L3 gates, mimicking a state where the L1:L2 ratio is skewed. This led to a decrease in the number of L2 robots carrying out the starting task and a subsequent new L2 activity equilibrium. The L1-L2 conflict was induced by adding miR-16 to activate L1 robots, leading to another decrease in L2 activity, finally equilibrating at a near-baseline value (Figure 4).

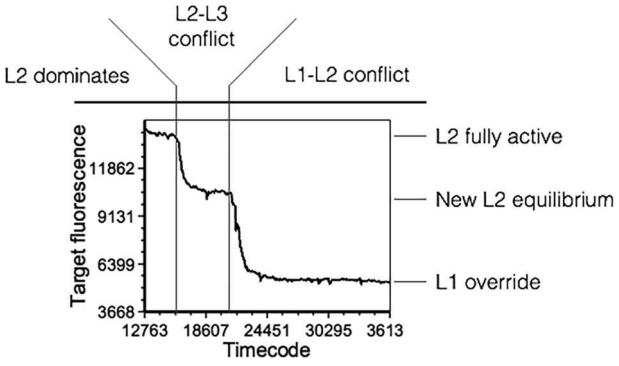

Figure 4. Experimental recreation of "Runaround" scenario using the L1-L2-L3 system. Real-time flow-cytometric analysis showing the fluorescent signal on target microparticles, that derive from active L2 robots. The graph shows the three stages of "Runaround": First, the L2 carry out their defined task; second, L3 closure occurs due to the skewed L1:L2 ratio (experimentally, adding L3 gate-clasping strands at the first marker line), leading to a lower activity equilibrium of L2; and third, damage is induced (experimentally, adding miR-16 at the second marker line), leading to a near-baseline equilibrium of L2 due to the activity of L1 robots (L1 override). In this sample the experiment duration is approximately 24 min.

3 Discussion

By letting go of the view that complex decision making is done by every single robot, we demonstrate that Asimov's laws of robotics can be implemented by molecular robots. As the behavior of each individual robot is not deterministic, this implementation is statistical, rather than classically deterministic, and is specifically enabled by the statistically reliable nature of large populations of molecular robots. Indeed, the "Runaround" experiment reliably replicates the conflicts deriving from Asimov's laws, despite the stochasticity of each individual robot and the inherent discrepancy between actual and planned ratios. The observed results reflect the net effect of approximately 100,000,000,000 robots.

The mechanism designed and described in this work only implements the first part of the first law of robotics: "A robot may not harm a human being." It takes action to stop harm from taking place. However, implementation of the second part of the law "or allow, by inaction, a human being to come to harm" is significantly more difficult to address. Consider, for instance, a different type of L2 molecular robots, whose task is to speed up healing. A blunt trauma to the body causes cells to break up, releasing miR-16. The L1 robots should not, in this case, act to stop the L2 robots. If anything, they should support and enhance their operation. The problem is the inherent ambiguity in the cause of the damage identified by miR-16 presence. However frustrating, this is actually not a problem of molecular robotics, but a general one in the first law, for robots of any size, and indeed for humans. Deciding on whether harm is caused by an action taken is difficult enough. Deciding on whether harm is caused by *not* taking an arbitrary action at any arbitrary time is clearly impractical (if not impossible). Imagine sitting next to an elderly person in the park. Can you be sure you are not causing harm, in any way, by not standing up? By not calling an ambulance? By not engaging in conversation? and so on. There are infinitely many actions not to take, at any given moment. There is no general solution of which we are aware, for molecular or macro-scale robots. But there exists a partial answer: It is to consider action, and lack of action, in the context of the robots' knowledge, the context of a specific task. For each task, we design a specific mix of L1-type robots and L2-type robots that together carry out the task while protecting from harm caused by the specific mix, or not caused by it. For surgery, the L1 robots will stop the L2 surgery robots from overacting. For healing, the (differently constructed) L1 robots will trigger L2 healing robots into operation.

The proof of principle we present is interesting on two levels. Given the inherent biocompatibility and therapeutic potential of DNA machines, this technology could lead to therapeutics that self-monitor the adverse effects that they themselves generate, and avoid them. We showed that this mechanism could be designed to respond to specific harm cues at varying degrees of stringency to allow for diverse tasks. Second, a technology that is inherently incapable of causing harm is special on more philosophical and ethical levels, and highlights unique features that could be innate in the reality of molecular entities.

Acknowledgments

The authors wish to thank S. Ittah, N. Mamet, A. Abu-Horowitz, E. Ben-ishay, B. Mizrahi, and all the members of the Bachelet lab for valuable help, comments and discussions. We additionally thank George Church for valuable comments on early drafts of this work. This study was partially supported by a European Research Council Starting Grant (335332) and a Leukemia Research Foundation (LRF) grant to I.B., and by ISF grant (1511/12) to G.K. Thanks also to K. Ushi.

G.A.K., R.S.S., Y.A., N.A., and I.B. conceived and designed the experiments, performed the experiments, and analyzed the data. G.A.K., N.A., and I.B. wrote the manuscript.

The authors declare no competing financial interests.

References

1. Amir, Y., et al. (2014). Universal computing by DNA origami robots in a living animal. *Nature Nanotechnology*, *9*, 353–357.
2. Andersen, E., et al. (2009). Self-assembly of a nanoscale DNA box with a controllable lid. *Nature*, *459*, 73–76.

3. Asimov, I. (2004). *I, robot*. Bantam Books.
4. Douglas, S., Bachelet, I., & Church, G. (2012). A logic-gated nanorobot for targeted transport of molecular payloads. *Science*, *335*, 831–834.
5. Ellington, A., & Szostak, J. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature*, *346*, 818–822.
6. Gu, H., Chao, J., Xiao, S., & Seeman, N. (2010). A proximity-based programmable DNA nanoscale assembly line. *Nature*, *465*, 202–205.
7. He, Y., et al. (2008). Hierarchical self-assembly of DNA into symmetric supramolecular polyhedra. *Nature*, *452*, 198–201.
8. Lakin, M., Youssef, S., Polo, F., Emmott, S., & Phillips, A. (2011). Visual DSD: A design and analysis tool for DNA strand displacement systems. *Bioinformatics*, *27*, 3211–3213.
9. Muscat, R., Bath, J., & Turberfield, A. (2011). A programmable molecular robot. *Nano Letters*, *11*, 982–987.
10. Qian, L., & Winfree, E. (2011). Scaling up digital circuit computation with DNA strand displacement cascades. *Science*, *332*, 1196–1201.
11. Rothemund, P. (2006). Folding DNA to create nanoscale shapes and patterns. *Nature*, *440*, 297–302.
12. Seelig, G., Soloveichik, D., Zhang, D., & Winfree, E. (2006). Enzyme-free nucleic acid logic circuits. *Science*, *314*, 1585–1588.
13. Seeman, N. (1982). Nucleic acid junctions and lattices. *Journal of Theoretical Biology*, *99*, 237–247.
14. Simmel, F., & Yurke, B. (2002). A DNA-based molecular device switchable between three distinct mechanical states. *Applied Physics Letters*, *80*, 883.
15. Stojanovic, M., Mitchell, T., & Stefanovic, D. (2002). Deoxyribozyme-based logic gates. *Journal of the American Chemical Society*, *124*, 3555–3561.
16. Stojanovic, M., et al. (2005). Deoxyribozyme-based ligase logic gates and their initial circuits. *Journal of the American Chemical Society*, *127*, 6914–6915.
17. Venkataraman, S., Dirks, R., Rothemund, P., Winfree, E., & Pierce, N. (2007). An autonomous polymerization motor powered by DNA hybridization. *Nature Nanotechnology*, *2*, 490–494.
18. Weld, D., & Etzioni, O. (1994). The first law of robotics (a call to arms). In *Proceedings of the 12th National Conference on Artificial Intelligence* (pp. 1042–1047). AAAI Press.
19. Yin, P., Yan, H., Daniell, X., Turberfield, A., & Reif, J. (2004). A unidirectional DNA walker that moves autonomously along a track. *Angewandte Chemie International Edition*, *43*, 4906–4911.
20. Zhang, D., & Seelig, G. (2011). Dynamic DNA nanotechnology using strand-displacement reactions. *Nature Chemistry*, *3*, 103–113.

Appendix: Methods

A.1 Robot Fabrication

Robots were designed and fabricated essentially as described elsewhere [1]. Briefly, M13mp18 bacteriophage DNA (20 nM) was mixed with staple oligonucleotides (final concentrations of 200 nM of each strand), in folding buffer (5 mM Tris, 10 mM $MgCl_2$, 1 mM EDTA, pH 8.0). The mixture was subjected to a thermal-annealing ramp as follows: 80 to 60°C at 2 min/°C, 60 to 20°C at 150 min/°C. Folded robots were cleaned of excess staples by sequential rounds of centrifugal filtration in standard 100-KDa cutoff filter columns. DNA concentration was measured by spectrophotometry.

A.2 Experiment Design

Experiments using microparticles as targets for L2 robots were carried out in TAE buffer supplemented with 0.1% w/v bovine serum albumin and 10 mM $MgCl_2$. The minimal amount of L2 robots was 100 fmol; other robots were included according to the appropriate molar ratios. Robots were loaded with goat anti-mouse Fab′ fragments prepared as described previously [4]. Amine-functionalized polystyrene microparticles were coated with mouse IgG (50 μg/mL) by crosslinking with ethyl-dimethylaminopropylcarbodiimide and sulfo-*N*-hydroxysuccinimide in MES-buffered saline (BupH ready-made buffer) for 1 h at room temperature, followed by buffer exchange into TAE to quench nonreacting antigen, followed by centrifugal isolation of microparticles.

A.3 Flow Cytometry

Robot assays, with biotin-loaded, fluorescently-tagged robots, were performed on antigen-coated 2-μm and 6-μm microparticles. Data were acquired on an Accuri C6 flow cytometer equipped with a 488-nm solid-state laser and 640-nm diode laser. Data were coarse-analyzed using FlowPlus software followed by analysis on FCS-Express 4.0 software (using a C6 import module).