ON THE COVER

As we marvel at the Fibonacci patterns in sunflowers, or the golden spirals in seashells, we often forget how difficult it is to extract meaningful patterns from not-yet-understood data; indeed, sometimes it is like staring into a dark and murky forest, unsure where one data-tree ends and another begins. This time, however, we've got you covered.

McGill Science Undergraduate Research Journal

THE 2012 - 2013 MSURJ COMMITTEE

Editors-in-Chief
Irene Xie
Biochemistry

Yuhao Shi (Winter 2013)
Computer Science and Biology

Natalie Gottlieb (Fall 2012)
Psychology

Managing Editor
Yuhao Shi
Computer Science and Biology

Associate Editors
Brian Krug
Biology

Jassi Pannu
Biology

Senior Editors
Janine Hsu
Immunology

Blair Jia
Quantitative Biology

Shannon Palus
Physics

Kate Sheridan
Cognitive Science

Elaine Xie
Pharmacology

Junior Editors
Ariana Aimani
Computer Science and Biology

Deborah Baremberg
Cognitive Science

Kaien Gu
Biochemistry

Kevin Liang
Neuroscience

Farah Musharbash
Physiology

Mason Sun
Chemistry and Economics

Assistant Editor:
Joanna Xu
Microbiology and Immunology

Faculty of Science Advisor:
Victor Chisholm
Office for Undergraduate Research in Science

Layout Designer
Clayton Nguyen

Graphic Designer
Irene Xie

805 Sherbrooke St. West Room 1B21
Montréal, Québec, H3A 2K6 Canada
Phone: (514) 398-6979
Fax: (514) 398-6766
Email: mcgillsurj@gmail.com
Web site: msurj.mcgill.ca

TABLE OF CONTENTS

FOREWORD

DEAR READER,

For eight years, the McGill Science Undergraduate Research Journal has worked hard to promote research to undergraduates and undergraduate research to others. Though the editors of our board have changed throughout the years, our objectives - and our love of research - never have.

Like all good scientists, we constantly question ourselves, striving to grow, to innovate, and to give back to the community that nurtures us. This year, we have chosen to do so by re-launching our blog, The Abstract (http://msurjblog.com), where current students and past graduates of the McGill community share their experiences with the turbulent, yet rewarding process leading up to the glossy research you hold in your hands. Perhaps it should have been called Materials and Methods; regardless, we hope our blog becomes a useful resource for generations of students.

Our annual publication celebrates the culmination of these efforts. As you turn the pages of this year's journal, we hope the works of your peers will inspire you to pursue your own ambitions in research and contribute to a community with infinite possibilities.

YUHAO SHI AND IRENE XIE

EDITORS-IN-CHIEF

ACKNOWLEDGEMENTS

The publication of the McGill Science Undergraduate Research Journal would not have been possible without the support and contribution of many individuals.

We thank Professor Martin Grant, Dean of the Faculty of Science, whose unwavering support throughout the past eight years has allowed this journal to flourish. We also extend our gratitude to Mr. Victor Chisholm for providing us with continued encouragement and valuable advice.

We thank all of our donors in the McGill community for their generous support:

- Faculty of Science
- Faculty of Medicine
- School of Environment
- Science Undergraduate Society
- McGill Biology Student Union
- Physiology Undergraduate League of Students
- McGill Psychology Student Association

We wish to thank the professors and post-doctoral fellows who graciously offered their time to review students' article submissions.

We would like to acknowledge the tireless efforts of the MSURJ board of editors in assembling this edition of the journal.

Lastly, we wish to recognize the student contributors whose commendable efforts have made this journal possible.

RESEARCH ARTICLE

Stress reactivity during evaluation by the opposite sex: comparison of responses induced by different psychosocial stress tests

Lydia Goff[1]*, Nida Ali[1,2,3], Dr. Jens Pruessner[1,2,3]

[1] Department of Psychology, McGill University, Montreal, QC

[2] Department of Psychiatry, McGill University, Montreal, QC

[3] Department of Neurology and Neurosurgery, McGill University, Montreal, QC

*Email Correspondence:
lydia.goff@mail.mcgill.ca

Abstract

It is becoming increasingly difficult for researchers to continue their high rates of publication when funding budgets are running tighter than ever. It is therefore in a researcher's best interest to utilize more economical tests whenever possible. This project aims to compare various stress tests in order to determine whether the new, cost-efficient Maastricht Acute Stress Test (MAST) activates a physiological and subjective stress response with the same effectiveness as pre-existing, more resource-intensive tests. This study demonstrated that the MAST produces a response similar to that of the previously predominant Trier Social Stress Test (TSST). Meanwhile, database data shows that the purely physiological Cold Pressor Task (CPT) lags behind in terms of response elicited. These findings may allow for a more cost-efficient yet highly effective stress task to become available to researchers.

Introduction

Stress is a major factor in modern daily life; we encounter it at school, at work, in our social relationships, and even at home. A stressful stimulus activates several interacting physiological stress pathways, the two main ones being the Sympathetic-Adrenal-Medullary (SAM) and the Hypothalamic-Pituitary-Adrenal (HPA) axes. Stimulation of the HPA axis following exposure to a stressor results in the release of corticotropin-releasing hormone (CRH) from the hypothalamus, which triggers the release of adrenocorticotropic hormone (ACTH) from the pituitary. ACTH then stimulates the secretion of glucocorticoids from the adrenal cortex. Conversely, in SAM axis action, ACTH also acts upon the adrenal medulla, which releases epinephrine and norepinephrine. This triggers an acute stress response from the sympathetic nervous system (such as increased blood pressure and heart rate).

While the SAM axis is predominantly responsible for an immediate fight-or-flight response (7), HPA axis activation maintains homeostasis after a challenge through a cascade of hormonal responses, which take up to 30 minutes to fully unfold (1). Thus, the two axes differ in how quickly they respond and how long they remain active. Although acute HPA activation is necessary for maintaining homeostasis, chronic activation of the HPA axis has been associated with complications such as immunosuppression, depression, and increased risk of coronary heart disease (15).

In humans, the main glucocorticoid released by stimulation of the HPA axis is cortisol, which plays an important function in modulating the stress response towards the re-establishment of homeostasis. Freely circulating levels of cortisol (i.e. cortisol not bound to proteins in the blood) serve as indicators of HPA axis activation and can be easily measured in blood and saliva (10, 11).

Furthermore, the HPA axis is especially sensitive to non-physical stressors involving a social context. Its activation is therefore considered a strong indicator of exposure to psychosocial stress. One of the most frequently used standard protocols for inducing such stress is the Trier Social Stress Test (TSST). The TSST involves a period of anticipation, then requires the participant to present a free speech in front of a panel of "experts" (personnel in lab coats), and afterwards to perform a mental arithmetic challenge (9). Other stress tests such as the Cold Pressor Task (CPT), which involves submerging one's hand in ice-cold water (20), fail to achieve similar increases in ACTH and cortisol levels as the TSST (14). Thus, it appears that a challenging task with a social-evaluative component that is outside of the individual's control and threatens their social status causes greater HPA activation than one with a physically stressful component alone (2, 18).

One limitation of the TSST is that it is time-and-resource intensive. The design of a simpler and more rapid test would be advantageous and economical. Recently, Schwabe *et al.* (2008) introduced a modified

version of the CPT that included a social evaluative component: the Socially Evaluated Cold Pressor Task (SECPT) (18). This test induces higher levels of salivary cortisol than the CPT alone, reinforcing the idea that social evaluation significantly increases HPA axis activity.

Most recently, Smeets *et al.* (2012) introduced the Maastricht Acute Stress Test (MAST) consisting of fixed SECPT intervals (CPT with social evaluation) separated by mental arithmetic challenge periods. The authors compared the responses achieved with the MAST with those from other stress paradigms, in this case CPT and SECPT, or TSST. They concluded that subjects' cortisol levels after MAST appeared higher than in both CPT and SECPT. A comparison of the MAST and TSST suggested that cortisol response curves were comparable in magnitude, and followed a similar progression (19). Carrying out all four tests within the framework of the same study allowed variables such as collection techniques, exclusions, protocol, and group size to be controlled. These results therefore suggest that the MAST may be a valid alternative to the TSST. The effectiveness of the MAST may be result of the fact that its CPT component provides a physical stressor that predominantly activates the SAM axis (indicated by elevated heart rate and blood pressure), while the social evaluation and mental arithmetic components of the TSST activate the HPA axis (indicated by higher levels of salivary free cortisol) (19).

The effect of subjects' sex is another potentially significant factor. Numerous studies have shown that the two groups who display greatest cortisol responses to a standard TSST are women in the luteal phase of their menstrual cycle, and men (10). However, a recent study by Duchesne *et al.* (2011) has shown that women in the follicular phase of their menstrual cycle and men show higher cortisol levels compared to women in the luteal phase and controls when the social evaluation during the TSST was carried out by an experimenter of the opposite sex (4). This is possibly a result of the fact that women in the follicular phase of their menstrual cycle display increased responsiveness and physiological arousal towards men and masculine stimuli (e.g. facial and vocal cues of men with high testosterone levels) (4, 6, 17).

However, it is unclear whether any such gender differences exist in the MAST. The work by Smeets *et al.* (2012), introducing the MAST, tested it on twenty men and no women. The results were subsequently compared to men-only TSST and CPT versions. As of yet, no project has directly compared all three stress tests at one time while including women. Such a study would be able to determine whether the MAST affects women in the follicular phase of the menstrual cycle differently from men when they are evaluated by experimenters of the opposite sex.

We hypothesized that the results of the current project would demonstrate that the administration of a psychosocial stress test by an experimenter of the opposite sex would lead to a comparable stress response in women in the follicular phase of their menstrual cycle and men. Further, we hypothesized that the magnitude of the stress response would differ between different psychosocial stress tests (TSST, CPT and MAST), in that the MAST and the TSST would induce a greater cortisol (HPA axis) and autonomic (SAM axis) stress response than the CPT.

Materials & Methods

Subjects

For the MAST, we recruited 9 men (ages = 20-27 years, mean = 23.33 years, sd = 2.45 years) and 8 women (ages = 18-21 years, mean = 19.50 years, sd = 1.20 years) via advertisements posted in a variety of Montréal electronic classifieds. Interested individuals filled out a screening questionnaire, and we contacted those eligible for the study in order to conduct a follow-up interview. In order to avoid any influence of hormonal changes associated with age (such as puberty and menopause) the participants recruited were all between 18 and 35 years of age. We excluded applicants who smoked more than 7 cigarettes a day, had a past history of psychiatric illness, had a body mass index (BMI, in kg/m^2) outside the range of 18-27, or were currently using steroid hormone medications and/or recreational drugs, as these criteria have been known to affect baseline hormone levels (4, 10). We scheduled female applicants who reported not using any oral contraceptives to undergo the MAST when they were in the follicular phase of their menstrual cycle (by testing them 2-13 days following the onset of menstruation; assessed via self-report and monitoring through phone calls). Women reporting themselves as pregnant were not allowed to participate. The Douglas Research Ethics Board approved this project

For the TSST, we included a total of 23 subjects (n_{men} = 12, n_{women} = 11; age mean = 23.09 years), and for the CPT, a total of 17 subjects (n_{men} = 11, n_{women} = 6, age mean = 23.35). Although different experimenters carried out each test, laboratory standards in all stress tests ensured that the same exclusion criteria were used, that participants were within the same age brackets and chosen from the same (predominantly undergraduate) student population, and that the same cortisol analysis methods and measurement instruments were used. Furthermore, only participants tested by an opposite-sex experimenter were included in our analysis so as to match our MAST sample. Only opposite-sex testing was employed due to time and resource constraints on carrying both same-sex and opposite-sex testing. Across all stress tests, the only outstanding difference that emerged during our analysis was that the CPT study involved "high" and "low" self-esteem groups (as measured by the Rosenberg Self-Esteem Scale (RSES)). Therefore, analyses were run with CPT divided into CPT-high and CPT-low groups.

General Procedure

Testing took place at the Douglas Research Institute between 1 and 5

pm to control for diurnal variations of cortisol, as cortisol levels are highest in the morning and decline throughout the day (10). Upon the participant's arrival in the laboratory, they were greeted by Experimenter #1 (the "coordinator") and taken to a waiting room for 40 minutes. This waiting period allowed hormonal levels and physiological measures to return to baseline should they have been increased due to previous spontaneous stimulation by random factors outside of the control of the experiment. During this time period, we gave the participants consent form to sign and the Visual Analogue Scale (VAS) of subjective stress to fill out. We took baseline saliva sample, heart rate (HR), and blood pressure (BP) measurements. Following the rest period, we moved the participant to the MAST testing room and introduced to Experimenter #2.

Experimenter #2 administered the MAST and was always of the opposite sex to the participant. We seated the participant with a pail of ice water (4 ± 0.5°C) placed next to their dominant hand and read them a set of instructions explaining the testing protocol. The experimenter stood in front of them next to a video camera. This camera was solely used to further induce social stress and was not operational; however, we told participants that it would be recording their facial expression for later inspection by behavioural analysts, and that they should therefore keep their gaze directed toward the camera for the duration of the procedure.
The MAST intervals were as follows: [90s-45s], [60s-60s], [60s-90s], [90s-45s] & [60s], where the first number in each bracket indicates the amount of time the hand must be submerged in ice cold water and the second number indicates the time spent doing the mental arithmetic challenge. The arithmetic challenge consisted of counting down as rapidly and accurately as possible from 2043 by steps of 17 (with their hand out of water). In order to increase the stressfulness of the situation, if a mistake in the counting was made, the experimenter stopped the participant and instructed them to begin again from 2043. If the participant was counting down with sufficient accuracy and speed, the experimenter would instruct them to speed up. After the completion of the test, the participant remained in the room for the following hour such that all samples and measurements could be obtained. Finally, we debriefed the participant about the procedure and hypothesis, and compensated them with $50 for their time.

We took saliva samples, heart rate, and blood pressure at ten-minute intervals during the entire length of the testing protocol: three times before the MAST and six times after it, for a total of ten saliva samples (to measure cortisol levels) and ten stress analogue scales (to measure the degree of subjective stress) (Fig.1). HR and BP served as indicators of SAM axis activity.

The protocol we used for the TSST data acquisition was based upon Kirschbaum's original task description (9) and the specific details follow those in Duchesne's study (4). The protocol used for the CPT was the same as that described by von Baeyer *et al.* (2005).

Physiological and Psychological Measurements

Each subject provided samples of salivary cortisol, measures of systolic and diastolic blood pressure, heart rate, and subjective stress ratings at 10-minute intervals throughout the procedure (producing a total of 10 measures per category) (Fig.1). We collected saliva using Salivettes (Sarstedt Inc., Quebec City, Quebec, Canada), which we then stored at -20°C until analysis by time-resolved fluorescence immunoassay (intra- and inter-assay variabilities of 10% and 12%, respectively) (3). We took blood pressure and heart-rate readings with a digital inflatable-wrist-cuff blood pressure monitor (Life Source, UB-512). Finally, we assessed subjective stress values using visual analogue scales (VAS), which involve asking participants to rate how stressed they feel at that time on a continuous scale from 1 to 10 (4).

Data Analysis

To determine the effectiveness of a stress test in inducing a stress response, physiological effects in response to a particular stress test (TSST, CPT or MAST) must be measured over a period of time. Thus, the dependent variables were cortisol levels (in nmol/L), heart rate (in beats per minute), systolic and diastolic blood pressure (in mmHg), and subjective stress (in centimetres, as measured on VAS),

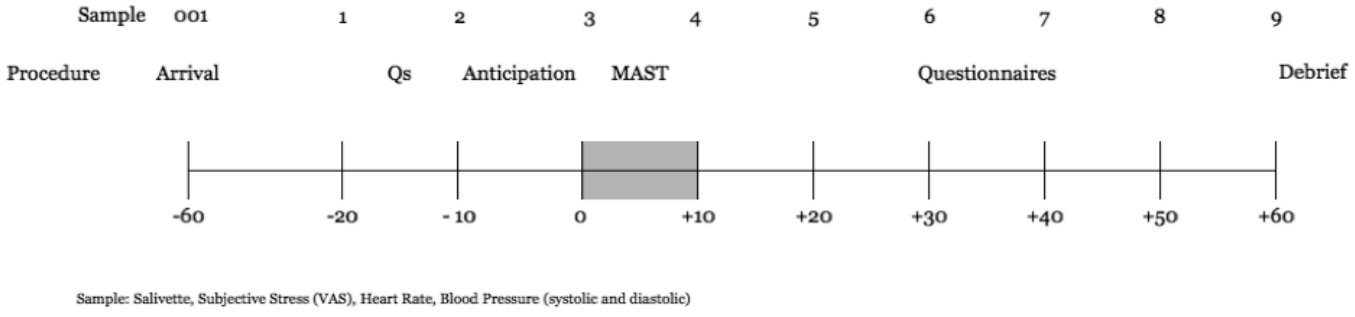


Fig. 1
Timeline of the MAST protocol. Qs= Questionnaires.

while our independent variable was time (measured in minutes).

We compared MAST with TSST and CPT data from the lab's database by using a mixed design analysis of variance (ANOVA). The ANOVA was of time by group by test (group = women in the follicular phase of their menstrual cycle or men, test = TSST, CPT or MAST) with repeated measures of cortisol, HR, BP (systolic and diastolic), and subjective stress. We determined significant main effects through pairwise comparisons corrected with Bonferroni's confidence interval adjustment. We set ninety-five-percent confidence intervals and a significance level of α=0.05 for all variables.

We conducted univariate ANOVAs on area-under-the-curve increase (AUCi), area-under-the-curve with respect to ground (AUCg) and Delta Peak (ΔPk) measures for cortisol, systolic and diastolic blood pressure, heart rate, and subjective stress. AUCi represents the increase in the area under the curve from baseline measures, AUCg represents the increase from zero, and ΔPk represents the difference between the height of peak and baseline values.

Results

Descriptive Statistics

We compared the results from the MAST (n = 17; n_{men} = 9, n_{women} = 8) with those of the TSST (n = 25; n_{men} = 13, n_{women} = 12) and the CPT (n = 18; n_{men} = 12, n_{women} = 6). Across all groups, four subjects' data had to be removed from the analysis since their scores were more than 3 SD away from the mean. We applied logarithmic transformations to non-normally distributed variables; this was the case for all measures except heart rate. Participants' age did not significantly differ between test groups or between sexes.

Effect of Sex and Stress Type on Salivary Cortisol Levels

No sex differences in cortisol levels existed in any of the stress tests. A mixed-design ANOVA with repeated measures of log-transformed cortisol levels showed that neither the main nor the interaction effects of sex were significant ($F(1,50) = 2.634$, $p = 0.111$; $F(2,50) = 1.435$, $p = 0.248$). In contrast, the same ANOVA showed that TSST and MAST produced comparable cortisol responses, which were higher than those of the CPT. All three tests produced a significant increase in cortisol levels over time within subjects (main effect of time ($F(2.818, 140.909) = 14.003$, $p = 0.000$) and time by test ($F(5.636, 140.909) = 4.104$, $p = 0.001$). Pair-wise comparisons carried out on time showed that cortisol levels at time points +10, +20, and +30 were significantly greater than cortisol levels at the -10min baseline.

The cortisol response did differ significantly between-tests (main effect of test between-subjects ($F(2,50) = 3.927$, $p = 0.026$)). Pair-wise comparisons revealed that the TSST and MAST were significantly higher than the CPT ($p<.05$; Fig.2-4).

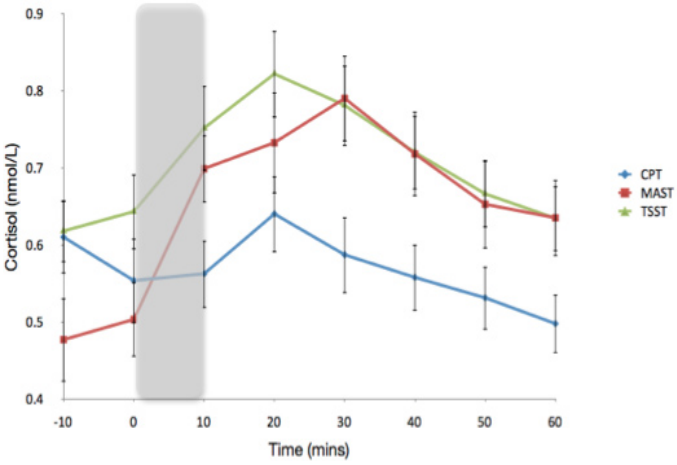


Fig. 2
Plots of estimated marginal means of log-transformed cortisol (nmol/L) levels in all three tests over time. There exists a significant effect of test between-subjects ($F(2,50)=3.927$, $p=0.026$) but only TSST and CPT are significantly different from each other ($p=0.022$). The error bars represent SEM.

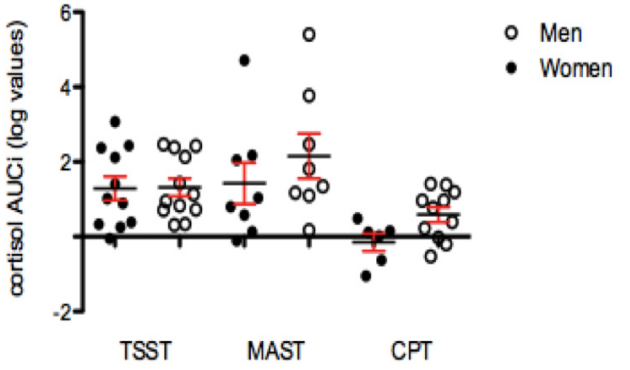


Fig. 3
Impact of three different stress tests on cortisol. Log-transformed cortisol data was used to calculate the AUCis. A significant effect of test was found in cort AUCi ($F(2,50)=8.304$, $p=0.001$). Means are displayed by full lines, SEM in red.

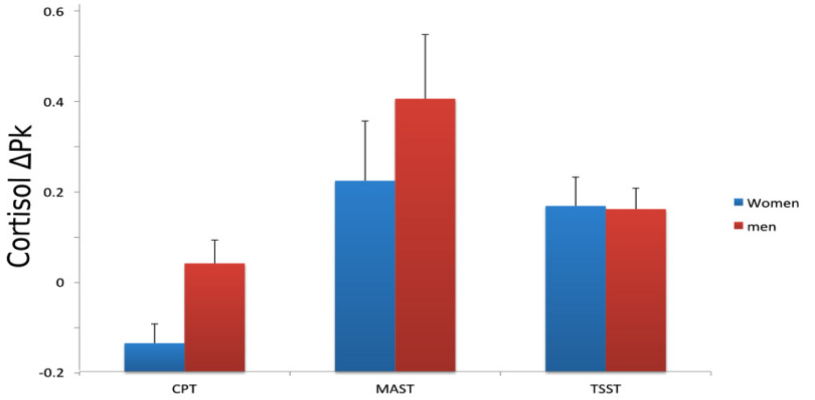


Fig. 4
Plots of estimated marginal means of log transformed cortisol ΔPk of both sexes over time in all three tests. Men's peak in cortisol is significantly greater than that of women (p=0.011). ΔPk is significantly different between TSST and CPT (p=0.024) and between MAST and CPT (p=0.006). The error bars represent SEM.

Effect of Sex and Stress Type on Blood Pressure and Heart Rate Measures

We carried out a logarithmic transformation on all blood pressure data because it was not normally distributed. HR data was normally distributed and therefore needed no transformation.

When looking at MAST data alone, the systolic BP values showed a significant between-subject effect of sex (p = 0.008) such that men's were greater.

We then ran a two-way repeated-measure ANOVA on both log systolic and log diastolic data. Both BP and HR significantly increased over time during all stress tests (lgsys F(1.823, 87.526) = 5.899, p = 0.005; lgdia F(2.852, 136.906) = 5.267, p = 0.002; HR F(4.750, 227.99) = 4.249, p = 0.001), however only HR ΔPk data showed a significant difference between tests supporting cortisol results. Responses of both TSST and MAST were comparable (p=1.000) and both were significantly greater than CPT (pTSST = 0.002; pMAST = 0.020) (Fig.5).

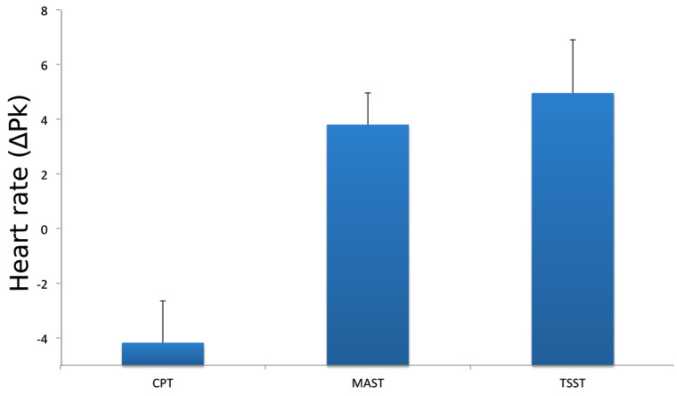


Fig. 5
Plot of estimated marginal means of heart rate ΔPk. A significant between-subjects effect of test was revealed (F(2,48)=7.021, p=0.002) where there is a significant difference between TSST and CPT (p=0.002) and between MAST and CPT (p=0.020). The error bars represent SEM.

Effect of Sex and Stress Type on Subjective Stress Ratings

Only TSST data showed a significant difference between sexes in their subjective stress responses over time (F(2.637, 55.382)=3.568, p=0.024), where women's subjective stress ratings were higher before and during the TSST.

A mixed-design repeated-measures ANOVA of subjective stress score showed that all tests produced a significant change in subjective stress ratings within-subjects over time (main effect of time (p=0.000) and time by test (p=0.031)).

Although self-esteem groups made no significant difference across all physiological measures, subjective stress did differ significantly between tests (p=0.000). Pair-wise comparisons demonstrated that subjective stress was strongly affected by whether CPT participants had high or low self-esteem; subjective stress scores of the CPT high self-esteem group were significantly lower than TSST (p=0.022), MAST (p=0.000), and CPT low self-esteem (p=0.002) (Fig.6).

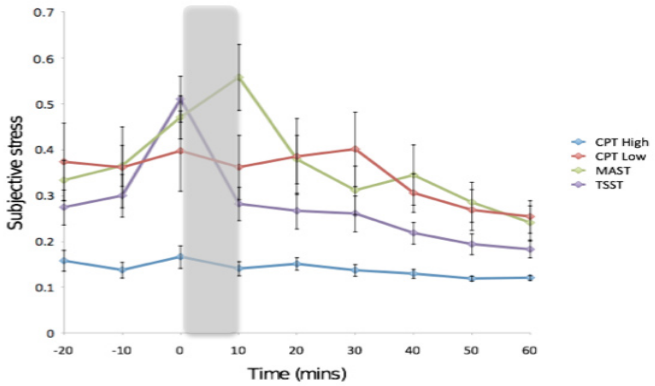


Fig. 6
Plot of estimated marginal means of log transformed subjective stress with CPT divided into high and low self-esteem groups. There exists a significant between-subjects main effect of test (F=7.420, p=0.000). CPT high-self-esteem group was significantly different from TSST (p=0.022), MAST (p=0.000) and CPT low-self-esteem (p=0.002).

Discussions

Opposite Sex Effects

Our main objective in this project was to determine whether opposite-sex testing significantly affected the stress responses of participants in a differential manner depending on their sex. Overall, opposite-sex testing did not seem to affect men and women in a differential manner. Apart from the subjective stress difference in the TSST (women rated higher in subjective stress) and the systolic BP difference in the MAST (men showed a greater systolic BP response), participants did not significantly differ in overall stress response. It is possible that the TSST produced a greater subjective stress response, being more sensitive to opposite sex evaluation due to its "job interview"-like, public speech context, period of preparation, and long speech time in front of an opposite-sex panel. Furthermore, it is likely that women in the follicular phase of their menstrual cycle are more susceptible to feeling stressed in this situation due to their heightened sensitivity to male stimuli (6).

Tests

In this project, we compared physiological and subjective responses to the TSST, the CPT, and the MAST stress tests when the tests were administered by experimenters of the opposite sex. Since the MAST is a newly developed stress task, this project was the first to both include women and directly compare its efficacy to that of the TSST and the CPT when administered by opposite-sex experimenters. In all tests, exposure produced a main effect of time in all repeated measures (cortisol, systolic and diastolic blood pressure, heart rate and subjective stress) indicating that all three stress tests were successful in producing significant responses in participants.

Our main hypothesis for this project was that the MAST administered by experimenters of the opposite sex would induce a comparable cortisol response to the TSST, and that both would generate higher responses than the CPT. Analyses of overall cortisol levels suggest that the CPT produced a significantly lower cortisol response than the TSST. The cortisol levels induced by the MAST were not significantly different from those induced by the TSST. In addition, a significant cortisol time by test effect was observed within subjects, where TSST and MAST cortisol values closely mirrored each other and CPT values were significantly lower. Subjects' cortisol responses over time were increasing following TSST, MAST and CPT. Therefore, at least in terms of HPA axis activation as indicated by salivary cortisol levels (and thus the effectiveness of a stress test to induce a stress at a psychosocial level), TSST and MAST seem to be equally effective.

Although blood pressure did not differ as a result of exposure to any of the tests, each test elicited different heart rate variations in participants over time. This suggests that both the MAST and the TSST are effective in producing an autonomic response in heart rate and that they are also both better at eliciting this response than the CPT. Therefore, it seems that stress induction tests with a psychosocial quality are not only better at activating the HPA axis, but also at producing an increased autonomic heart rate response.

When the participants of the CPT were divided into "high" and "low" self-esteem groups, there was a significant change in participants' subjective stress ratings over time. This change varied between tests. Each test thus elicited a significantly different subjective stress response within each participant throughout the testing period. Significant test effects between subjects also showed that the TSST, the MAST, and the CPT low-self-esteem groups all produced significantly greater subjective stress ratings than the CPT high-self-esteem group. This could perhaps be because individuals with very high self-esteem tend to be markedly more confident in their ability to cope with stressors and therefore perceive less stress throughout the procedure. The opposite would then be true for those with low self-esteem (i.e. less confidence) (5). In other words, self-esteem seems to buffer the subjective perception of stress during stress exposure in high-self-esteem individuals (5).

The results obtained from the MAST alone demonstrate that it causes a significant change in activation of both the HPA and SAM axes as well as in subjective stress over time in each participant. This is shown by a significant time effect within participants in cortisol levels, systolic and diastolic blood pressure, and heart rate measures as well as subjective stress ratings, in that all increase subsequently to task onset. These results further confirmed those of Smeets *et al.* (2012) as well as our hypothesis that the MAST would be an effective stress induction test that could significantly activate both the HPA and the SAM axes. Its ability to do so is possibly related to the fact that the MAST contains aspects of both psychosocial and physical stressors.

When compared to the other tests, our findings clearly demonstrate that the MAST is comparable to the effectiveness of the TSST. Furthermore, both the TSST and the MAST seem to be more effective stress induction tests than the CPT, especially when it comes to HPA axis activation and heart rate. These conclusions supported our initial statements that MAST is as effective a stress test as TSST, and that both are superior to CPT. However, our findings do not support our hypothesis since in no case did the stress response (as measured by indicators of HPA and SAM axis activation) induced by MAST exceed the one induced by TSST.

Limitations

Due to the restricted time available for this project, the major limitations were the small group sizes and the dependence on database data from previous TSST and CPT studies for analyses. Data from the TSST and the CPT were collected by different experimenters, with different protocols and different variables and covariates than the

MAST data collected specifically for this project. This was particularly true in the case of the CPT, where the added factor of self-esteem and the lack of emphasis on opposite-sex testing may have affected our results in sex differences.

Small sample sizes producing greater inter-individual variation may also have affected our results. Therefore, further studies with larger sample sizes would be required to confirm our conclusions. Finally, our main hypothesis tested for a similarity in stress responses between the sexes, thus interpreting the beta-error. As a consequence, we should have included computations for this error type in our statistical analysis, to show that if differences existed between men and women in the entire population, we possessed sufficient statistical power to demonstrate such an effect in our sample. However, because of financial and economic constraints in our testing regimen, we were unable to increase our sample size for this project, thus any power calculations would have been without consequence.

Concluding Statements

Based on the findings of this project, the MAST performed equally well in inducing a physiological stress response as the TSST in women in the follicular phase as well as men, and performed better in all cases than the CPT. Future directions of this line of research should investigate the effect of opposite versus same-sex testing, and the effect in women across all phases of the menstrual cycle or with the use of oral contraceptives.

Acknowledgements

The greatest thanks go out to Dr. Jens Pruessner, for his faith and confidence, and to Nida Ali, for her advice, assistance, and tremendous patience. Many thanks are also due to the entire, wonderful team in Dr. Pruessner's lab - especially to Annie Duchesne and Catherine D'Aguiar - for the motivation and advice they provided. Finally, a big thank you to Dr. Alan Goff and Dr. Daniel Levitin for their corrections, constructive criticism, feedback, and comments.

References

[1] Ali, N., Pruessner, J. C., *Physiological Behavior. Advance online publication.* (2011).
[2] Dickerson, S. S., & Kemeny, M. E., *Psychological Bulletin*, 130(3), **355-391** (2004).
[3] Dressendorfer et al., *The Journal of Steroid Biochemistry and Molecular Biology*, 43(7), **683-692** (1992).
[4] Duchesne, A., Tessera, E., Dedovic, K., Engert, V., & Pruessner, J. C., *Biological Psychology*. Advance online publication. (2011).
[5] Dumont, M., Provost, M.A., *Journal of Youth and Adolescence*, 28(3), **343-363** (1999).
[6] Feinberg, D.R., Jones, B.C., Law Smith, M.J., Moore, F.R., DeBruine, L.M., Cornwell, R.E., Hillier, S.G., Perrett, *D.I., Hormones and Behaviour* 49(2), **215–222** (2006).
[7] Henry, J.P., *Integrative Physiological and Behavioral Science*, 27(1), **66-83** (1992).
[8] Kirschbaum, C., & Hellhammer, D. H., *Psychoneuroendocrinology*, 19(4), **313-333** (1994).
[9] Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H., *Neuropsychobiology*, 28(1-2), **76-81** (1993).
[10] Kudielka, B. M., & Kirschbaum, C., *Biological Psychology*, 69(1), **113-132** (2005).
[11] Kirschbaum, C., Hellhammer, D., in collaboration with Allostatic Load working group, Salivary Cortisol and Challenge tests-summary. (2000).
[12] Kudielka, B. M., Hellhammer, D. H., & Wüst, S., *Psychoneuroendocrinology*, 34(1), **2-18** (2009).
[13] Kudielka, B. M., Schommer, N. C., Hellhammer, D. H., & Kirschbaum, C., *Psychoneuroendocrinology*, 29(8), **983-992** (2004).
[14] McRae, A. L., Saladin, M. E., Brady, K. T., Upadhyaya, H., Back, S. E., & Timmerman, M. A., *Human Psychopharmacology: Clinical and Experimental*, 21(6), **377-385** (2006).
[15] Miller, G. E., Chen, E., & Zhou, E. S., *Psychological Bulletin*, 133(1), **25–45** (2007).
[16] Mitchell, L. A., MacDonald, R. A. R., & Brodie, E. E., *The Journal of Pain*, 5(4), **233-237** (2004).
[17] Roney, J.R., Simmons, Z.L., *Hormones and Behavior* 53(1), **14–19** (2008).
[18] Schwabe, L., Haddad, L., & Schachinger, H., *Psychoneuroendocrinology*, 33(6), **890-895** (2008).
[19] Smeets, T., Cornelisse, S., Quaedflieg, C. W., Meyer, T., Jelicic, M., & Merckelbach, H., *Psychoneuroendocrinology*, **27**, 1998-2008 (2012).
[20] von Baeyer, C. L., Piira, T., Chambers, C. T., Trapanotto, M., & Zeltzer, L. K., *The Journal of Pain*, 6(4), **218-227** (2005).

RESEARCH ARTICLE

Studying a poxvirus gene capture model through recombination and reactivation

Katherine T. Johnson[1]*, Dr. David H. Evans[2,3]

[1] Department of Microbiology and Immunology, McGill University, Montreal, QC
[2] Department of Medical Microbiology and Immunology, University of Alberta, Edmonton, AB
[3] Li Ka Shing Institute of Virology, University of Alberta, Edmonton, AB

*Email Correspondence:
katherine.johnson@mail.mcgill.ca

Keywords
Poxvirus: *Large double-stranded DNA viruses, which infect both vertebrates and invertebrates.*
Vaccinia (VACV): *Poxvirus famously used to vaccinate humans against smallpox. Vaccinia virus is the most widely used model in poxvirus research.*
Recombination: *The breaking and rejoining of genetic material to create new combinations of genetic traits.*
Homology: *Sequence similarity between two (or more) DNA, RNA, or protein molecules.*
Titer: *Concentration of infectious virus expressed in plaque forming units per mL (PFU/mL).*

Abstract

Introduction: Vaccinia poxvirus (VACV) is a double stranded DNA virus that replicates in the cytoplasm of infected cells. Some VACV genes resemble homologs of host genes and appear to have been captured from the cell; however, since poxviruses are confined to the cytoplasm, researchers are unclear as to how these viruses acquire this homology (1). If a cellular mRNA was accidentally reverse transcribed into cDNA, which could occur during retrovirus co-infection, a poxvirus might be able to incorporate this sequence into its own genome through rare non-homologous (homology-independent) recombination.

Methods: We modeled this process using two different recombination systems and substituted a DNA encoding mycophenolic acid (MPA), a selectable marker, for the hypothetical non-homologous host cDNA. We prepared DNA constructs containing this marker along with 20 base pairs homologous to the 5' and 3' flanking regions of the VACV-encoded *Not*I restriction site. A construct without this flanking homology was also prepared. The "passive" recombination system used a helper poxvirus to reactivate VACV DNA; in contrast, VACV infected BSC40 cells were transfected with the construct in the "active" recombination system.

Results: The "passive" recombination system generated 105 PFU/mL of reactivated VACV; however, no recombinants containing the selectable marker were detected. The "active recombination" method generated 106 PFU/mL of total VACV and approximately 10 PFU/mL of recombinant virus for both homology containing and non-homology containing constructs.

Discussions: We were unable to determine the recombination frequency of the "passive system" because recombinant virus was not detected. Based on virus titers determined from plaque assays, we approximated the recombination frequency of the "active system" to be ≤ 10-5. We are currently cloning and sequencing viruses resulting from non-homologous recombination to determine where the MPA marker is located. Preliminary analysis of these types of clones (data not shown in this paper) suggests that the transfected DNAs are being incorporated into a diversity of sites, some located near the boundary of the VACV genome where the right terminal inverted repeat begins. In summary, our findings suggest that the recombination frequencies in both methods are very low and better methods of selection are needed to observe these rare events. Future studies of recombinant clones are needed to gain a better understanding of this non-homologous gene capture process.

Introduction

Poxviruses are large double stranded DNA viruses that replicate solely in the cytoplasm of infected cells. VACV was famously used to vaccinate against smallpox and is very commonly used in poxvirus research. VACV and other members of this family encode all of the enzymes required for viral replication, including DNA replication and transcription. Genetic analysis of VACV and other poxviruses has revealed sequences that appear to have cellular origins (1). Over the course of thousands of years, these viruses presumably captured

copies of cellular genes, which mutated into products that conferred advantages against host immune responses. For example, the VACV CrmE gene encodes a tumor necrosis factor (TNF) receptor, which has significant homology to the mammalian type 2 TNF receptor (1). This mutant receptor allows the virus to evade apoptosis mediated by host produced TNF-α. However, since poxviruses are confined to the cytoplasm, researchers are unclear as to how they are able to acquire these genes. Furthermore, it is also unclear how other large DNA viruses, such as Herpes virus, capture foreign genes. It is unlikely that VACV picked up genes from other sources though, but Fadly *et al.* showed that Fowlpox virus could capture the DNA form of the Reticuloendotheliosis retrovirus genome during replication (5). However, it is possible that this retrovirus DNA was simply mistaken for cellular DNA. In the current study, we investigated a potential mechanism for poxvirus "gene capture" using recombination.

Previous work done by Evans *et al.* showed that recombination occurs frequently during poxvirus replication (2). Gradually, successive recombinations lead to the evolution of new strains as DNA is exchanged between viruses. If a host gene is transcribed, processed, and then reverse transcribed back into a cDNA, a poxvirus could capture this gene through rare illegitimate (non-homologous) recombination. Presumably, the required reverse transcriptase activity would be supplied by endogenous enzymes or by a retrovirus also present within the cell (retroviral co-infection). Viral and host genomes have little sequence similarity; the cDNA fragment would be incorporated into a random location within the viral DNA, but it would likely insert into regions surrounding the telomeres, where non-essential genes are located. We also hypothesized that cDNA containing any flanking homology to a specific region of the viral genome will recombine into that locus even if the homology is only a few nucleotides. The frequency of recombination events will increase when larger sequence homologies are present (3).

This study provides preliminary data and method optimization for our future recombination experiments. Although VACV is a widely used model system in virology, little is known about its evolution or the origins of poxvirus genes. This work represents the first step in understanding this process and testing our proposed gene capture model, showing that poxviruses can integrate foreign DNA found in the cytoplasm into their genomes via non-homologous recombination.

Methods

Plasmids, Primers and PCR Conditions:

Two PCR products were constructed from the pDGloxP(del) plasmid, kindly provided by Don Gammon. One of these constructs contained 20 base pairs of flanking homology to the NotI restriction digest site in the VACV genome and the other did not have any flanking homology to the NotI site (Fig. 1). Notably, there is only one NotI digestion site in the VACV genome. Both of these constructs also contained a yellow fluorescent protein-guanine phosphoribosyltransferase mycophenolic acid selectable marker (YFP/gpt) controlled by an early poxvirus promoter. A control construct was also created, which did not contain the selectable marker, but instead contained approximately 600 base pairs of homology flanking either ends of the NotI restriction site (Fig. 1).

PCR Product A (gpt/*Not*I) was constructed using the following primers:

5'TAGACAACACACCGACGATGAGGGCGAATTGAGTGAAGGCCG3' (forward) and 5'TAGTCATATTCGTGTCTGTGGAGTCAGTGAGCGAGGAAGCGG3' (reverse) (Fig. 1).

PCR Product B (gpt) was constructed using the following primers:

5'AGGGCGAATTGAGTGAAG3' (forward) and 5'GAGTCAGTGAGCGAGGAA3' (reverse) (Fig. 1).

PCR product C was constructed from VACV DNA using the following primers, kindly provided by Li Qin:

5'ACACCCAAAAACAACCGA3' (forward) and 5'ACATACCATCGACATCCA3' (reverse) (Fig. 1).

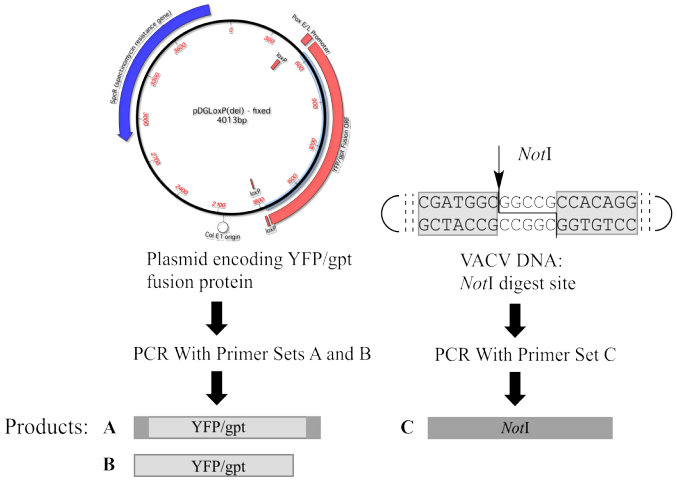


Fig. 1
Constructed PCR products containing a YFP-gpt fusion protein (MPA selectable marker). Product A is flanked by 20 base pairs of homology to the VACV NotI digestion site. Product B does not have homology to the NotI digestion site. Product C is a control for recombination experiments and encodes 600 base pair homology to the NotI digestion site.

All PCR products were verified by gel electrophoresis and purified using spin columns (Fermentas).

Plaque Assay Using Crystal Violet Staining

All viruses were titered by infecting mycoplasma-free African green monkey kidney epithelial (BSC40) cells in six well culture plates with 1mL of ten-fold dilutions (10-1 to 10-7) of virus. Titer plates were cultured for 1 day at 37°C to allow for plaque formation and then stained using crystal violet. To stain, cell medium was removed and then crystal violet was added at room temperature for 30 minutes. The stain was removed by rinsing plates with distilled water. Virus titer in plaque forming units (PFU) was calculated by selecting the well that contains approximately 10-20 plaques and then accounting for the dilution. For example, if the 10-4 dilution well had 10-20 plaques, then virus titer would be 105 PFU/ mL considering that cells were infected with 1mL of virus.

Active Recombination: Live Virus Recombination

Mycoplasma-free BSC40 cells were cultured in Minimal Essential Medium (MEM, Life Technologies), which was supplemented with 5% fetal bovine serum, 1% nonessential amino acids, 1% l-glutamine, and 1% antibiotic at 37°C. These cells were infected with VACV (Western Reserve strain) at a multiplicity of infection of one for 1 hour in phosphobuffered saline (PBS) at 37°C, which is enough time for the majority of cells to become infected. As a negative control, BSC40 cells were also mock infected with PBS for 1 hour at 37°C. VACV infected cells were then transfected with either of the PCR constructs using the following protocol. DNA and Lipofectamine 2000 transfection reagent (Invitrogen) were incubated separately in minimal serum medium (Opti-MEM, Life Technologies) for 5 minutes at room temperature and then gently added together and incubated for an additional 20 minutes at room temperature. PBS was removed from the cells and then the transfection mixture was added, along with MEM, immediately 1 hour after the VACV infection period. Cells were transfected for 1 hour at 37°C. These time frames were chosen based on work carried out by G. McFadden *et al.*, which showed that replication and recombination occur concurrently during the first phase of the VACV replication (replication persists up to 6-8 hours post-infection) (2). The transfection mixture was then removed and fresh medium was added. Cells were cultured at 37°C for 1 day to allow for viral replication and then recombinants were selected for by adding 1 X mycophenolic acid (MPA) (Fig. 2).

Recombinant virus was cloned and purified by re-plating in the presence of 1 X MPA three additional times, which ensures elimination of wild type VACV. To approximate recombination frequency, plaque assays were performed on virus samples taken from the initial round of MPA selection. Total virus (recombinant + non-recombinant) was determined by titering samples from transfected cells that were plated without drug. Approximate recombination frequency was calculated using

$$\frac{\text{Recombinant virus } (\frac{\text{PFU}}{\text{mL}})}{\text{Total virus } (\frac{\text{PFU}}{\text{mL}})}$$

Passive Recombination: VACV Reactivation Using a Helper Poxvirus

Poxvirus DNA alone is not infectious because it does not have access to either viral or cellular DNA replication machinery. Passive recombination is a reactivation strategy where new VACV is produced from viral genomic DNA. Yao *et al.* showed that VACV DNA could be reactivated into live virus if the cell is infected with another helper poxvirus (4). The helper poxvirus provides all necessary replication machinery. Furthermore, since VACV can be reactivated from a series of overlapping DNA fragments, recombination should also occur between a foreign PCR product with homology to the VACV genome and the other corresponding VACV fragments during reactivation (Fig. 3). The PCR product simply resembles another fragment of viral DNA if significant homology is present. We tested whether recombinants could be produced if the PCR product did not have any homology. This allowed us to investigate whether gene capture is a "passive" process and whether it could have occurred while the virus was inactivated (e.g. by UV light) within a host cell. We also investigated whether a double stranded break promotes recombination, by digesting VACV DNA with *Not*I before transfection (Fig. 3).

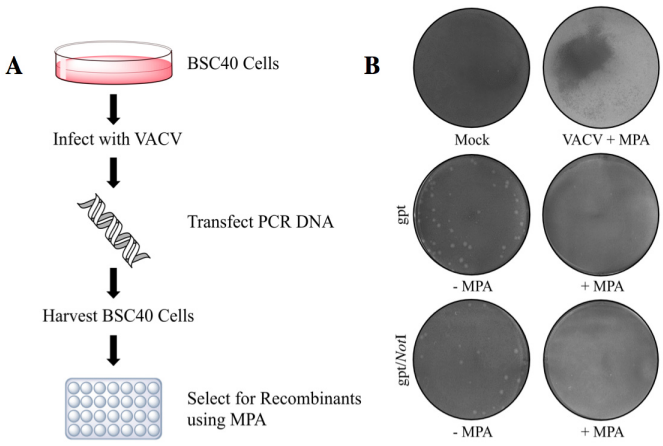


Fig. 2
(A) Active recombination system method using BSC40 cells. Only recombinant VACV containing the YFP/gpt fusion gene can survive in the presence of MPA.
(B) From top left to right: mock infected BSC40 cells; VACV (strain WR) infected BSC40 cells cultured for 3 days with MPA; VACV infected BSC40 cells transfected with PCR product B (gpt, no NotI site homology) with and without MPA selection; VACV infected BSC40 cells transfected with PCR product A (gpt/NotI, 20 base pairs homology) with and without MPA selection.

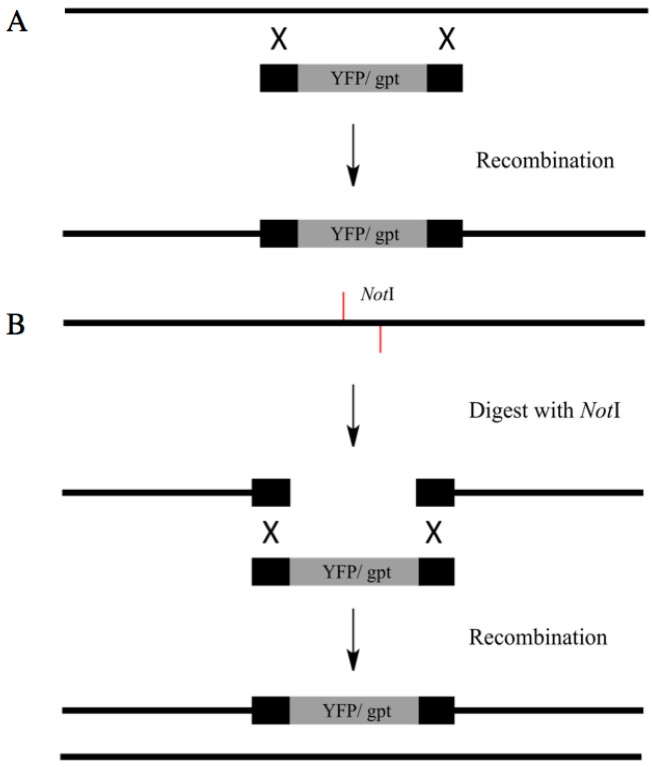


Fig. 3
(A) Homologous recombination of selectable marker and VACV DNA without digestion (double stranded break independent).
(B) Homologous recombination of selectable marker and VACV with *Not*I digestion (double stranded break dependent).

For the reactivation, mycoplasma-free Buffalo green monkey kidney (BGMK) cells were cultured in Minimal Essential Medium (MEM, Life Technologies), which was supplemented with 10% fetal bovine serum, 1% nonessential amino acids, 1% l-glutamine, and 1% antibiotic at 37°C. BGMK cells were infected with Shope Fibroma poxvirus (SFV) at a multiplicity of infection of 0.1 for 1 hour in PBS at 37°C. As a negative control, BGMK cells were also mock infected with PBS for 1 hour at 37°C. BGMK cells support the growth of both VACV and SFV (Fig. 4). SFV infected BGMK cells were then transfected with either NotI digested or undigested VACV DNA and either of the PCR constructs according to the aforementioned protocol. Cells were cultured at 37°C for 1 day to allow for VACV reactivation and replication. Reactivated virus was selected by plating on BSC40 cells, which only support VACV growth; virus was plated with 1 X MPA to select for recombinants and without 1 X MPA to determine total reactivated VACV. After culturing for 3 days at 37°C, resulting VACV was titered using plaque assay with crystal violet staining. Virus was given 3 days to replicate because very little reactivated virus is initially present.

Results

Active Recombination: Live Virus Recombination

We investigated whether replicating VACV could capture DNA containing a selectable marker. BSC40 cells were infected with VACV and then recombinants were isolated by culturing with MPA. Notably, recombinant viruses were cultured under a total of four rounds of drug selection to eliminate wild type VACV even though no plaques were visible above after the first round (Fig. 2). For cells transfected with PCR Product B (gpt, no *Not*I site homology), the viruses present in - MPA conditions are most likely wild type VACV (titer of 10^6 PFU/mL) because no recombinants were detected (Fig. 2). However, we were able to propagate recombinant virus from these + MPA samples that have very low titers (approximately 10 PFU/mL). + MPA titers increased with continued culture as the small amount of recombinant virus present replicated to detectable levels. Plaque assays of recombinant virus after four rounds of MPA selection showed high virus titers of approximately 10^7 PFU/mL (data not shown here); non-homologous recombination between the DNA and VACV was successful. VACV infected BSC40 cells transfected with PCR Product A (gpt/*Not*I, 20 base pairs homology) yielded similar results as the gpt transfected virus. The 20 base pair homology does not appear to have had a significant effect on recombination rates because comparable amounts of recombinant and total virus were recovered (10 PFU/mL and 10^6 PFU/mL). However, based on previous work done by the Evans's lab, we expect larger homologies, such as 100 base pairs, to have a more pronounced affect on recombination (3). We expect higher initial titers of recombinants with this amount of homology (> 10 PFU/ml). The overall recombination frequency for the active system was calculated to be 10^{-5} (0.00001).

Passive Recombination: VACV Reactivation Using a Helper Poxvirus

We investigated whether a segment of DNA could be incorporated into VACV DNA during reactivation by a helper poxvirus. BGMK cells were infected with SFV and then transfected with *Not*I digested or undigested VACV DNA and a PCR construct. The resulting virus was then cultured in BSC40 cells with and without drug selection.

Approximately 10^2 PFU/mL of total reactivated VACV was recovered from BGMK cells transfected with undigested VACV DNA and PCR Product B (gpt, no *Not*I site homology) (Fig. 4). However, recombinants were not detected, and we were unable to propagate the virus further as with virus from the active recombination system (Fig. 4). Interestingly, 10^5 PFU/mL of reactivated virus was recovered from cells transfected with undigested VACV DNA and PCR Product C, which has 600 base pair homology (data not shown). Less than 10^2 PFU/mL of total reactivated VACV was recovered from BGMK cells transfected with NotI digested VACV DNA and PCR Product A (gpt/*Not*I, 20 base pair

homology) (Fig. 4). Recombinants were not detected, and we were unable to propagate neither total virus nor recombinant virus further. However, the reactivation system still functions with digested DNA because approximately 10^2 PFU/mL of reactivated virus was recovered from BGMK cells transfected with digested VACV DNA and PCR Product C (data not shown).

Discussions

The active recombination system produced 10^6 PFU/mL VACV without MPA selection (total VACV) and 10 PFU/mL with MPA selection (recombinant VACV). These titers were very similar for both homology containing PCR Product A and non-homology PCR Product B recombinant viruses. The passive recombination system generated 10^2 PFU/mL of reactivated VACV for BGMK cells transfected with PCR Product B and undigested VACV DNA, but no recombinants were detected. 10^5 PFU/mL of reactivated virus was recovered from BGMK cells transfected with undigested VACV DNA and PCR Product C. These results were similar for cells transfected with PCR Product A and undigested VACV DNA. For BGMK cells transfected with NotI digested VACV DNA and PCR Product A, less than 10^2 PFU/mL reactivated VACV was recovered and no recombinants were detected. We could not recover reactivated virus for BGMK cells transfected with digested VACV DNA and PCR Product B; however, 10^2 PFU/mL reactivated VACV was recovered from BGMK cells transfected with digested VACV and PCR Product C.

Although we did not detect recombinant VACV from the reactivation system and detected few recombinants from the active recombination system, it is possible that these rare recombination events occurred more frequently than suggested. For the active system, we were able to propagate the recombinant virus even though we were not able to titer it; we did not see plaques in the 10^{-1} dilution. At low titers, it is possible to take samples of the stock that do not contain the virus. A titer using this sample would then produce a misrepresentation of the actual number of recombinants. Taking this into account, we approximated the viral titer to be 10 PFU/mL. To avoid this sampling error in future experiments, the entire stock of virus could be plated out on larger dishes, giving a more accurate reflection of the virus titer.

Yao *et al.* demonstrated that VACV reactivation is limited by the number of fragments that need to recombine to reconstruct the genome (4). With an increasing degree of fragmentation, VACV reactivation would be severely reduced or would not occur at all. To avoid this, we grew VACV and purified new DNA to ensure minimal shearing; however, reactivation frequencies did not improve. The reactivation system is also limited by transfection efficiency. Helper poxvirus infected cells must contain all appropriate overlapping DNA fragments in order to reactivate VACV. Recombination between VACV DNAs may still occur in infected cells; however, they may fail to produce live VACV if some fragments are missing. Thus, the reactivation system depends on how much DNA is taken up by helper poxvirus infected cells. With a more efficient transfection, we are more likely to find correct fragments inside a single cell. We will continue to optimize transfection efficiency by varying the concentration of the transfection reagent used and by changing the amount of DNA transfected. However, we expect that increasing the transfection efficiency will not change the recombination frequency because modern transfection protocols introduce large molar amounts of DNA into cells; there will be enough DNA present for recombination to occur optimally. The reactivation system primarily depends on having the right DNA fragments present within the cell, the likelihood of which increases if more DNA is introduced into the cell.

Interestingly, reactivated VACV was produced in samples transfected with the non-homology containing PCR product (PCR Product B). Recombination is favored when there is homology between the overlapping DNA sequences; correspondingly, the likelihood of recombination decreases significantly between non-homologous sequences (4). Yao *et al.* showed that recombination frequencies increase when the overlap in sequence between recombining DNAs is greater (3). Since recombinants were not detected in the presence of MPA, it is possible that the plaques found on drug free plates were formed by wild type VACV. This could be confirmed by DNA sequencing. Notably, SFV + VACV DNA was not under selective pressure until cell extracts were plated on BSC40 cells. Since MPA selection was not present, VACV

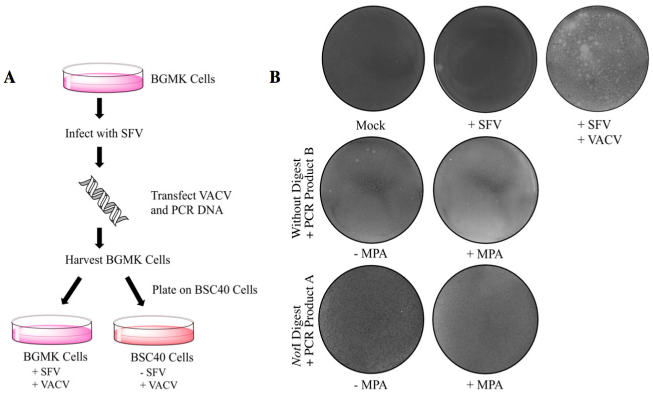


Fig. 4
(A) Passive recombination system method using BGMK and BSC40 cells. Only recombinant reactivated VACV containing the YFP/gpt fusion gene can survive in the presence of MPA.
(B) From top left to right: Mock infected BSC40 cells; SFV infected BSC40 cells; SFV and reactivated VACV infected BGMK cells; reactivated virus from BGMK cells transfected with undigested VACV DNA and PCR product B (gpt, no *Not*I site homology), with and without MPA selection; reactivated virus from BGMK cells transfected with NotI digested VACV DNA and PCR product A (gpt/*Not*I, 20 base pair homology), with and without MPA selection.

did not have to pick up the YFP/gpt marker during reactivation. If transfected BGMK cells were plated with MPA, then only virus in cells expressing the gpt marker would grow. Presumably, this is one potential method of increasing the likelihood of producing recombinant reactivated VACV.

We hypothesized that passive recombination would be favoured by the presence of a double stranded DNA break. To test this, VACV DNA was digested at a unique genome locus with the *Not*I restriction enzyme before transfection. We detected very little reactivated VACV in samples transfected with the selectable marker constructs and digested VACV DNA. Approximately 10^5 PFY/mL of total VACV was recovered for BGMK cells transfected with undigested VACV DNA and PCR product C versus 10^2 PFU/mL for cells transfected with digested VACV DNA. The digested DNA was perhaps too fragmented for reactivation to occur efficiently. Although *Not*I only cleaves the DNA at one locus, the DNA may have sheared during the purification process following digestion. Our protocol could be modified to minimize shearing, for example, by shaking gently during DNA extraction. Overall, the reactivation system must be further optimized and repeated in order to investigate whether poxviruses acquire new genes through reactivation and whether double stranded DNA breaks impact recombination and reactivation frequencies. Taking all of the above factors and titer results into account, we approximate the recombination frequency of the passive system to be less than or equal to 10^{-9}. We could not mathematically approximated the recombination frequency because no recombinants were detected.

Conclusions

Our preliminary study showed that actively replicating poxviruses are able to capture cDNAs present in the cytoplasm under selective pressure. This work suggests that poxviruses are more likely to capture genes while they are functionally replicating within infected cells. With such a low recombination frequency, it is unlikely that poxviruses acquire new genes through reactivation.

Future experiments will focus on the recombinants generated through the active method. We will continue to search for new methods that may improve efficiency and selectivity of both recombination systems. Furthermore, after purifying DNA from recombinant VACV clones and using PCR to confirm the presence of YFP/gpt fusion gene, we can subsequently sequence the recombinants to determine the regions where recombination occurred. Quantitative PCR could also be used to study these rare recombination events. Sequencing results will be confirmed via Southern blot: digestion of recombinant VACV DNA with various restriction enzymes compared with a digest of wild type VACV. Further experiments to see whether VACV can capture a stably expressed selectable marker from cells could be conducted after optimization of these recombination systems and further analysis of recombinant clones is done.

While VACV is a commonly used model in poxvirus biology, little is known about its evolution. Further research in the field of viral genetics will yield a better understanding of how poxviruses acquire mechanisms of immune evasion and about other viral genes that appear to have cellular homology.

Acknowledgements

This work was funded by NSERC Undergraduate Summer Research and Alberta Health Innovations Solutions Summer Studentship awards. Thank you to Li Qin for her invaluable mentorship and guidance in the lab, and for her primers. Thank you to Don Gammon for the YFP/gpt encoding plasmid. We also thank all members of Evans lab for their support and advice.

References

[1] Graham SC et al. J. Mol. Biol. **372**, 3 (2007).
[2] Evans DH, Stuart D, McFadden G. Journal of Virology. **62**, 2 (1988).
[3] Yao XD, Evans DH. Journal of Virology. **75**, 15 (2001).
[4] Yao XD, Evans DH. Journal of Virology. **77**, 13 (2003).
[5] García M, Narang N, Reed WM, Fadly AM. Avian Disease. **47**, 2 (2003).

RESEARCH ARTICLE

The effect of *Eurosta solidaginis* parasitism on pollinator preference in *Solidago canadensis*

Adam Meyer[1]*, Natasha Salter[1], Stephanie Shooner[1]

All authors contributed equally to this work
[1] Department of Biology, McGill University, Montreal, QC

*Email Correspondence:
gregory.meyer2@mail.mcgill.ca

Abstract

Solidago canadensis is a self-incompatible perennial species indigenous to North America that reproduces asexually via rhizomes and sexually via seeds. It is the favoured host of the gall fly, *Eurosta solidaginis*. Sexual reproduction leads to faster rates of adaptation in stressful environments and may be advantageous in the maintenance of host-parasite coevolution. The effect of infection by the gall fly on pollinator visitations at the patch and at the ramet level was assessed as a proxy for the ability to sexually reproduce. The study was conducted by analyzing pollinator preference at both the patch level and individual ramet level through successive observations of pollinator visitations. Though the variation in the number of pollinator visitations could be accounted for by time of day and median bloom stage, the percentage of infected ramets in a patch was not a significant explanatory variable. This suggests that gall formation does not affect pollinator preference and that the capacity to sexually reproduce is likely not reduced because of the host-parasite interaction. Broadly, this study served as an example of how pollinator preference may be utilized as a measure of fitness, and to further understand how selective pressures affect plant populations that reproduce both sexually and asexually.

Introduction

Solidago canadensis, the Canada Goldenrod, is a perennial species indigenous to North America that reproduces asexually via rhizomes and sexually via seeds. It colonizes abandoned fields and after one year of seedling growth it begins to reproduce vegetatively, extending out from the centre to form a circular clone. The ability to reproduce both sexually and asexually raises interesting questions as to how the plant responds to environmental stresses. For example, there is evidence that goldenrod clonal colonies (genets) can share resources among ramets (any stem belonging to the genet), but it is unclear whether a given stress to an individual ramet is indeed detrimental to the entire genet (1). One such stress is parasitization by the gall fly *Eurosta solidaginis*, which is responsible for the formation of stem ball-galls on the plant.

In the early summer, female gall flies oviposit their eggs in the developing leaves of *S. canadensis*; when the larvae hatch, they migrate to the meristem tissue and induce the formation of a gall. The stem continues to grow above the gall (2). As McCrea and Abrahamson (1985) demonstrate, there may be significant physiological and reproductive detriments to individual ramets with little to no perceived damage to the genet as a whole (1). For example, ramets infected by gallmaking parasitic insects have been found to allocate less energy resources towards both seed production and rhizome extensions (2), and devote an appreciable amount of energy to producing the gall, to the detriment of the ramet (3). Though effects to genet fitness remains apparently low, it has been suggested that if a significant number of damaged ramets were present within the genet, this may be sufficient to decrease the fitness of the whole (1). Furthermore, the fact that gall infections affect sexual reproduction in *S. canadensis* more than asexual reproduction (1), creates the possibility of selection for resistance to gallmaking insects.

Since *S. canadensis* is self-incompatible, outcrossing mediated by pollinators such as honeybees, bumblebees, soldier beetles and syrphid flies is obligatory for sexual reproduction (2). If differential resource allocation due to infection decreases the attractiveness of the clone to pollinators, thereby reducing its sexual reproductive success, then this genotype may be excluded from the population in favour of a parasite-resistant genotype. Through the sexual reproductive success of more resistant genotypes, *S. Canadensis* is both more likely to evolve resistance against parasites, and evolve resistance at faster rates. On the other hand, if the rate of cross-pollination is not affected by the presence of the parasite then population resistance may evolve more slowly compared to the rate of increased virulence of the parasite.

In order to better understand the effects of *E. solidaginis* on goldenrod evolution we focused on the plant's ability to attract pollinators.

This is an important aspect of the capacity to sexually reproduce in plants, an aspect that is absent in *Solidago* literature. There exists a rich literature describing pollinator (mainly bee) preference as being non-random with respect to floral colour, nectar concentration and other obvious indicators of fitness (4), as well as demonstrating an ability to respond to varying favourability of foraging patches (5). The importance of pollinators to sexual reproduction and to their ability to distinguish between favourable and unfavourable foraging patches demonstrates that pollinator preference should be considered among other measures of fitness already present in the literature.

Since studies have shown many instances in which the fitness of individual ramets was detrimentally affected by gall fly infection (1, 6, 7), and given the apparent ability for many pollinators, mainly bees, to make non-random decisions in choosing foraging patches, this study sought to test two hypotheses. Firstly, on the level of the genet, patches with a higher percentage of ball-gall infected ramets will receive overall less pollinator visitations and second, that on the level of individual ramets, pollinators will preferentially visit uninfected ramets over infected ramets.

Materials and Methods

Field Methods

Study Site

Our study population of *S. canadensis* grows on an east facing slope of Mont-Saint-Hilaire, an igneous montane part of the Monteregian Hills in Quebec, Canada. *S.canadensis* is one of several species that colonized an abandoned apple orchard on the Gault Nature Reserve, the private sector owned by McGill University of the Mont-Saint-Hilaire Biosphere Reserve. The orchard has been abandoned for approximately five years; *Solidago canadensis* could have been present in the field prior to the orchard being abandoned, however, its growth would have been inhibited by continual mowing. Therefore, we proceeded on the assumption that each of our study patches represented at most five years of growth. Data was collected on August 29th and 30th, 2012.

Patch Choice

Twenty-three patches were chosen. We were unable to determine if selected patches represented a single genetic individual as the use of genetic techniques was beyond the scope of this study. In order to obtain a sufficiently large sample size, the size range varied from 1.46 m^2 to 9.45 m^2 and the bloom stages varied from two to five with a median bloom stage across patches of four. Each patch was assigned a bloom stage value by randomly selecting 10 ramets per patch and identifying their bloom stage based on a predetermined number system (Table 1). In our analysis, we chose to consider the median bloom stage for each patch in order to avoid any outliers in our random ramet sample.

Patch Composition

The level of parasitization of each patch by the ball gallmaking *E. solidaginis* was determined. Each ramet was assessed for the presence of ball-galls. A total ramet count and a count of the number of infected ramets was tabulated so that the percentage of ramets parasitized by E. solidaginis could be obtained.

Flower Counts

A ramet infected with one gall and an uninfected ramet from each patch were removed at the end of the two days of fieldwork. Ramets were taken from the centre of the patch to ensure that they were of similar age and were adjacent to each other to maximize the chance of obtaining ramets from the same genetic individual. Ramets were also similar diameters just below the gall implying that they were ramets of similar size at the time of gall formation. Comparisons can only be made within a clone to ensure that differences between single gall ramets and uninfected ramets are not the consequence of genotype, but of parasitization. The number of open and closed flowers per ramet were counted to obtain the percentage of open flowers per ramet.

Pollinator Observations

To determine the number of pollinator visitations per patch and pollinator preference, visual snapshots (hereafter referred to as rounds) of each patch were taken. Assuming that bees are actively choosing patches, and ramets within patches, the probability of finding them on a given patch or ramet at any moment in time will be greater on favourable patches or ramets. Three rounds were performed consecutively with three rounds making up one trial; each patch was observed for three or four trials. Before beginning a trial, we waited 30 seconds to allow the pollinators to acclimatize to our presence. The patch was then visually assessed for the presence of pollinators. Pollinators were divided into three groups: bumblebees, honeybees, and other. Each round consisted of examining the patch for only the time necessary to document which pollinators were present and the

Patch Composition	**Bloom Stage Number**
Pre-bloom only	1
Mixed pre-bloom and full bloom	2
Full Bloom only	3
Mixed Full bloom and post-bloom	4
Post-bloom only	5

Table 1: Bloom stage classification system: Bloom stage numbers were assigned to 10 random flowers in each patch to determine the overall patch bloom stage.

type of ramet (infected versus uninfected) that they were visiting. Each ramet was only screened once per round for the presence of pollinators. Therefore, if a pollinator moved from one ramet to another, the pollinator was counted again if it landed on a ramet that had not yet been screened. To avoid the bias of time of day, the time of trials were staggered throughout the day for a given patch. The time when each trial was performed was recorded.

Statistical Methods

Patch-level Pollinator Preference

The hypothesis that if patches have higher levels of infection (percentage of ramets infected by *E.solidaginis*) then they will receive less pollinator visitations was evaluated using linear mixed-effects models with nested random factors. The process of building a model that best explained the variation in the response variable and the number of visitations by pollinators involved systematically establishing the best random structure and the significant fixed effects by comparing models using ANOVA. When models were not significantly different the most parsimonious model was chosen. The model was created using R statistical software and the package nlme (8). The percentage of infected ramets within a patch was arcsine transformed because non-normality is assumed for ratios. A graphical representation of the model was created using gamm4 in R statistical software (9).

Ramet-level Pollinator Preference

The hypothesis that pollinators would preferentially visit uninfected ramets over infected ramets was tested by taking the difference between the proportion of total visitations per patch to infected plants and the proportion of infected ramets in the patch. This difference was arcsine transformed because non-normality is assumed for ratios. A one sample t-test was then conducted to see if the mean difference across patches differed significantly from the null expectation of zero.

Flower Counts of Infected versus Uninfected Ramets

To test whether the proportion of open flowers in infected ramets differed significantly from the proportion of open flowers in uninfected ramets a two sample paired t-test was conducted. The proportion of open flowers was arcsine transformed because non-normality is assumed for ratios.

Results

Patch-level Pollinator Preference

In the 23 patches, the percentage of ramets infected by *E. solidaginis* ranged from 0% to 61.3%. The two models with the lowest Akaike information criterion (AIC) were determined after systematically isolating the linear mixed-effects model with nested random factors that best predict the response variable-number of visitations by all pollinators to a patch. The AIC is a measure of how well the model fits the data. Both models have trials nested within patch as random factors, but differ in their fixed effects. The fixed effects of model 1 are the percentage of infected ramets per patch (arcsine transformed), bloom stage median of the patch and time of day when the patch was visited (Table 2).

Random Effects	Standard Deviation		**AIC**	1327.44	
	Intercept	Residual			
Patch	4.096				
Trial	2.183				
Fixed Effects	Value	Standard Error	Degrees of Freedom	t-value	p-value
% of Infected Ramets (arcsine transformed)	-1.212	5.454	20	-0.2223	0.8263
Bloom Stage Median	-2.759	1.071	20	-2.575	0.0181
Time (min)	0.008692	0.002258	66	3.850	0.0003

Table 2:
Model 1: a linear mixed-effects model with nested random structure. This model was not the best fit because it is not statistically different from model 2 (ANOVA: Log likelihood ratio = 0.0566144 , p-value = 0.8119) which is the most parsimonious explanation for the variation in pollinator visitations between patches.

Random Effects	Standard Deviation		**AIC**	1327.44	
	Intercept	Residual			
Patch	3.993				
Trial	2.183				
Fixed Effects	Value	Standard Error	Degrees of Freedom	t-value	p-value
Bloom Stage Median	-2.755	1.047	21	-2.632	0.0156
Time (min)	0.008672	0.002258	66	3.842	0.0003

Table 3:
Model 2: a linear mixed-effects model with nested random structure. This model is the best fit for the data providing the most parsimonious explanation for the variation in pollinator visitations between patches.

The percentage of infected ramets per patch does not significantly affect the number of visitations by pollinators (P=0.8263) whereas bloom stage median of the patch and time of day do significantly affect the number of visitations (P=0.0181 and P=0.003, respectively) (Table 2). The fixed effects of model 2 are the median bloom stage of the patch and the time of day when the patch was visited (Table 3).

Bloom stage median and time of day both significantly affect the number of visitations (P=0.0156 and P=0.003, respectively) (Table 3). Model 1 and model 2 are not significantly different, thus model 2 is the best fit for the data because it is the most parsimonious predictor of pollinator visitations (ANOVA: Log likelihood ratio=0.0566144, P=0.8119). The model is a good predictor of the number of visitations by pollinators (pseudo R-squared=0.91) with little spread around the one-to-one line (Fig. 1)

Figure 2 demonstrates the relationship between the two fixed effects of model 2 and the number of visitations by pollinators. The time of day during which patches were observed for pollinator visitations ranges from 10:09 (609 minutes past midnight) to 17:35 (1055 minutes past midnight).

Number of visitations by pollinators fluctuates throughout the day, but consistently decreases as the bloom stage median increases (higher bloom stage medians are closer to the end of the bloom cycle of the plant). The number of pollinator visitations reached the highest peak near the end of the day with the most visitations occurring at this time on patches with lower bloom stage medians (Fig. 2).

Ramet-level Pollinator Preference

There was no significant difference between the number of visitations by pollinators to ramets with galls and the number of visitations that would be expected given the percentage of infected ramets within a patch (one sample t-test: t=1.18 , df=22, P=0.2497).

Flower Counts of Infected vs. Uninfected Ramets

The difference between the proportion of open flowers (mean=0.2114 (arcsine transformed)) on infected ramets and the proportion on uninfected ramets (mean=0.1936) is not statistically significant (paired t-test: t=0.2628, df=42.12, P=0.794).

Discussions

Contrary to our prediction that increased parasitization would negatively impact pollinator visitations, we found no relationship between the proportion of infected ramets within a patch and the number of pollinator visitations. However, despite an apparent apathy towards the degree of patch infection, the pollinator visitations can still be reliably predicted by the median bloom stage of the patch and the time of day (Model 2, Fig.1 and 2). The greatest number of pollinator visitations occurred in the late afternoon on patches with low median bloom stages. It has been previously demonstrated that bumblebee activity increases as temperature increases (10). It is also known that honeybees are able to remember the location of high quality plants and the time of day when the greatest reward can be obtained (5). .

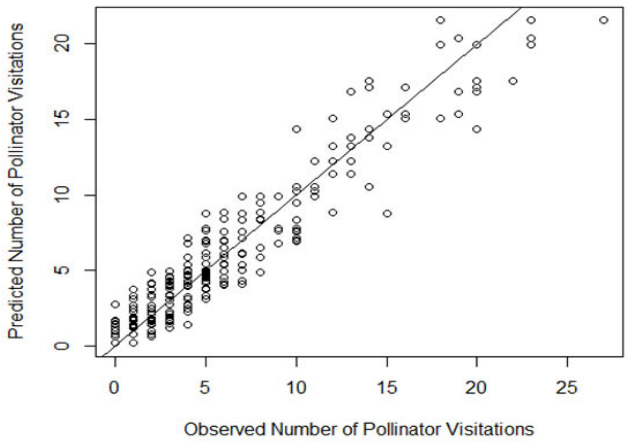


Fig. 1
The number of pollinator visitations predicted by model 2 compared to the observed. Linear regression represents a 1:1 relationship between observed and predicted number of pollinator visitations. The pseudo R-squared for model 2 = 0.91

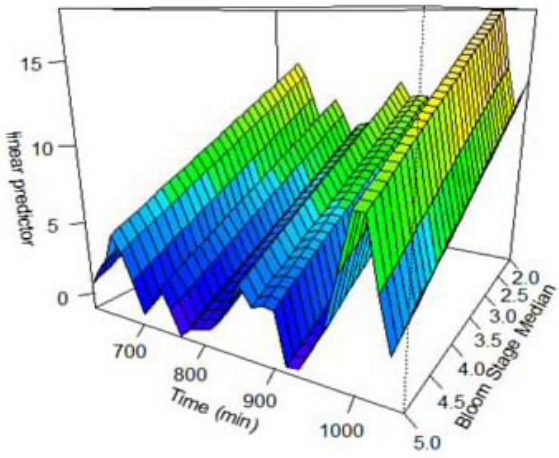


Fig. 2
Graphical representation of linear mixed-effects model with nested random structure: Model 2. Linear predictor is the number of visitations by pollinators. The warmer colours represent higher numbers of pollinator visitations.

The fluctuations throughout the day likely result from differences in peak reward time (greatest nectar production) and preferred foraging time (hottest time of day) though data is not available to confirm this prediction. Our data demonstrates that the pollinators in this study were responsive to differences in flower quality, further suggesting that galls do not affect flower attractiveness in a manner detectable by pollinators. Most importantly, this ability to make non-random foraging decisions serves to reinforce the notion that infection damage by *E. solidaginsis* at the genet level is negligible.

There may be several reasons why infection damage was apparently negligible at the genet level. Hartnett and Bazzaz (1983) find that selective pressures may be mitigated at the patch (genet) level through the physiological integration of connected ramets (11). Furthermore, they find that this may contribute to maintaining genetic diversity in the field by preventing genet death (11). Our study supports this notion by demonstrating how an environmental pressure such as a parasitic fly, which is draining to individual ramet, may be insignificant as a genet selection pressure.
This study also provides some evidence in contrast to the prediction put forth by McCrea and Abrahamson (1985), that the genet may be negatively influenced by *E. solidaginsis* infection, should the level of infection be sufficiently severe (1). Our study found that even when infection was high (roughly 50% for many patches, reaching a maximum of 61.3% galled ramets), detriment to the genet was not sufficient enough to influence pollinators. Thus, even if a patch is quite infected, an important aspect of sexual reproduction (i.e. pollinator attraction) may remain unchanged. This indicates that the damage done by *E. solidaginsis* is not significant enough to select for resistance to infection.

Additionally, the lower biomass allocated to the inflorescence in gall fly parasitized *S. canadensis* ramets (6) likely reduces the total volume of nectar available per inflorescence. Bumblebees are more sensitive to rewards based on sugar concentration, consistently choosing the reward with a higher sugar concentration even when the overall reward is equal. Therefore, if the decrease in total biomass allocated to the inflorescence resulted in flowers with nectar concentrations equivalent to those found in uninfected ramets, then it is possible that the presence of ball galls would have no effect on pollinator visitations at either the patch or the ramet-level (5).

With regards to pollinator preference within a patch, contrary to what was expected there was no non-random pollinator preference pattern with respect to infected or uninfected ramets. This could mean that individual ramets within the patches were not damaged enough by the parasite to (a) influence pollinator preference, or (b) lose out on randomly foraging pollinators. Both of these possibilities are supported by our data which demonstrates no significant difference between the proportion of open flowers on galled and ungalled ramets across patches. Additionally, while the literature puts forth evidence of many detrimental effects to galled ramets including reduced rhizome extension, seed production and biomass in general (1), it is possible that these effects were either not present in our sample size (supported by our comparison of open flower proportion) or simply not significant enough to influence pollinator behavior. Another possibility is that once pollinators have made patch-level decisions, they sample randomly within the patch.

Conclusions

The most significant findings of this study were that contrary to our prediction there was no detectable pollinator preference at the level of the genet or the ramet. While this result is negative, it remains interesting because of the larger implications it has on the understanding of selective pressures to plants as they can reproduce both sexually and asexually. For example, it is possible that in this case energy allocations between interconnected asexual ramets may provide a net benefit to the fitness of the genet (11) as it assists in keeping infected ramets healthy enough to not affect pollinator preference an important component of sexual reproduction in S. canadensis. Moreover, this study demonstrates how asexual reproduction in plants may contribute to stronger genet fitness by influencing sexual processes like pollination. The implications of this study call for further research into the effectiveness of using pollinator preference as a measure of plant fitness as well as further research towards understanding how selection pressures of varying severity act upon species, which can utilize both sexual and asexual reproduction.

Acknowledgements

We'd like to thank Felipe Perez, Simone Reader, Zophia Tanura, Andrew Hendry, Martin Lechowicz and Etienne Low-Decarie for all of the extra help and guidance.

References

[1] McCrea, K. D., Abrahamson, W. G. *Evolutionary impacts of the goldenrod ball gall maker on Solidago altissima clones.* Oecologia. 1985; 68: **20-22**.
[2] Werner PA, Gross RS, Bradbury IK. *The Biology of Canadian Weeds.* 45. Solidago canadensis L. Can. J. Plant Sci. 1980; 60: **1393-1409**.
[3] Stinner, B.R., Abrahamson, W.G., 1979. *Energetics of the Solidago canadensis–stem gall insect–parasitoid guild interaction.* Ecology 60(5), **918-926**.
[4] Nuttman, C.V., Seminda, F.M., 2006. *Visual cues and foraging choices: bee visits to floral colour phases in Alkanna orientalis (Boraginaceae).* Biological Journal of Linnean Society, (87), **427-435**.
[5] Cnaani J, Thomson JD, Papaj DR. *Flower Choice and Learning in Foraging Bumblebees: Effects of Variation in Nectar Volume and Concentration.* Ethol. 2006; 112: **278-285**.
[6] Hartnett DC, Abrahamson WG. *The Effects of Stem Gall Insects on Life History Patterns in Solidago canadensis.* Ecol. 1979; 60: **910-917**.
[7] Wise, M.J., Coffey, L.E., Abrahamson, W.G., 2008. *Nutrient Stress and Gall Flies interaction to Affect Floral-Sex ration In Gymnomonoecious Solidago Altissima.* American Journal of Botany. 95(10), **1233-1239**. 1 - 814 - 4514 - 888
[8] Pinheiro J, Bates D, DebRoy S, Sarkar D, et al. nlme: *Linear and Nonlinear Mixed Effects Models.* 2012; R package version 3.1-104.
[9] Wood S. gamm4: *Generalized additive mixed models using mgcv and lme4.* R package version 0.1-6. 2012.
[10] Comba L. *Patch use by bumblebees (Hymenoptera apidae): temperature, wind, flower density and traplining.* Ethol. Ecol. Evol. 1999; 11: **243-264**.
[11] Hartnett DC, Bazzaz FA. *Physiological Integration among Interclonal Ramets in Solidago canadensis.* Ecol. 1983; 63(4): **779-788**.

RESEARCH ARTICLE

Selecting for multicellularity in the unicellular alga *Chlamydomonas reinhardtii*

Tyler Moulton[1]*, Dr. Graham Bell[1]

[1] Department of Biology, McGill University, Montreal, QC

*Email Correspondence:
tyler.moulton@mail.mcgill.ca

Keywords:
Multicellularity: *The state defined by an organism comprised of more than one cell, often associated with division of labour among cells.*
Volvocaceae: *Family of algae representing a gradient of unicellular to multicellular species.*
Chlamydomonas reinhardtii: *A unicellular, flagellated, volvocacean alga.*
Artificial Selection: *Consistent, biased, guided selection for a phenotypic trait in a population.*

Abstract

Background: Researchers have recently begun experimentally exploring the origins of multicellularity (4-6). Their studies have found that the transition to a multicellular state may have been surprisingly simple, considering its profound implications for the history of life (3). This study experimentally selected for multicellularity in the unicellular biflagellated alga *Chlamydomonas reinhardtii*. This organism is especially interesting because it is basal to the Volvocaceae—a family of biflagellates whose evolutionary transition from unicellular organisms to complex forms have been meticulously characterised (2). The present study aimed to recreate the early stages of this transition, starting from incomplete cytokinesis after cell division.

Methods: The procedure was modeled loosely on the experiment performed by Ratcliff *et al*. (4) in which the authors successfully selected for multicellular *Saccharomyces Cerevisiae*—unicellular baker's yeast. Three experimental replicates and one control for nine strains of *C. reinhardtii* were cultured in round-bottom vials in shaking incubators. Prior to each transfer (every 3-4 days), each culture was slowly mixed, and selection lines were then gently and briefly centrifuged. This applied a selection pressure which rendered heavier (clustered) cells more fit. The very bottom ~2% of the tubes' contents was transferred, and cell cultures were examined for multicellularity.

Results: Six of nine lines of *C. reinhardtii* demonstrated an increased frequency of *C. reinhardtii* existing in a two- to four-celled state (the paired cell state accounted for 88% of these clusters)—consistent with the first step toward multicellularity as outlined by Kirk (2). A close study of cell division in the line which exhibited the strongest shift towards the multicellular phenotype suggests that true multicellularity began to evolve in this experiment. A multicellular phenotype did not become fixed in any population.

Conclusion: Our findings suggest that an artificial selection pressure is capable of inducing the evolution of multicellularity. Expanding upon this study could help us understand the mechanisms underlying the evolution of multicellularity.

Limitations: Sporadic data, possibly the result of difficulties in the procedure, prevented us from rigorously examining the effect of selection through time, limiting our ability to describe the evolutionary response. In addition, the study of individual cells, due to its time-consuming nature, was limited to one replicate of one line exhibiting a pronounced multicellular response. Thorough replication would be required before drawing a strong conclusion from this assay.

Introduction

Multicellularity is one of the most crucial innovations in the history of life, allowing for drastic increases in organismal size and complexity. Multiple phyla and clades within these phyla have independently evolved this trait (1, 2) suggesting that this innovation may be fairly easy to achieve (3, 4). Boraas (5) found that in a predator-prey chemostat, unicellular Chlorella vulgaris evolved a multicellular phenotype in response to predation from a flagellated protist. Koschwanez *et al*. (6) demonstrated that cell clumping greatly increases fitness of Saccharomyces cerevisiae in dilute sucrose solution due to metabolite sharing from secreted hydrolytic enzymes. Ratcliff *et al*. (4)* have recently reported the successful evolution of multicellular yeast simply by selecting for heavier cells via centrifugation. The yeast even

* The paper by Ratcliff *et al*. was published after the present study was carried out.

Citation: W.C. Ratcliff F. Denison, M Borrello, M. Travisano. Proc. Natl. Acad. Sci U.S.A. 109,1595-1600 (2012)

exhibited division of labor, indicating increased complexity. Such results are promising, but skeptics point out that yeast may have a multicellular ancestor within the Ascomycota, which may trivialize these results (7, 8). It is critical, therefore, that these types of studies be expanded to include other model organisms.

The family Volvocaceae presents an ideal context to study the evolution of multicellularity because the transition to multicellularity from a unicellular volvocacean ancestor occurred recently on the evolutionary timescale (between 30 and 70 mya). As a result, the family contains a group of species which very clearly demonstrates the adaptive progression from small simple colonies to the larger and comparatively complex organisms (2, 9). *Chlamydomonas reinhardtii*, a unicellular biflagellated alga, is typically used as the outgroup in Volvocacean phylogenies, a practice supported by phylogenetic studies of rRNA and ITS sequences of the Chlorphyceae (10-12). Kirk (2) outlined a putative twelve-step progression of adaptations from an ancestral volvocacean similar to C. reinhardtii all the way to the largest, most complex member of the family—*Volvox carteri*. Therefore, *C. reinhardtii* is an excellent starting point for experimentally examining the transition to multicellularity.

Several studies have already shown that *C. reinhardtii* can exist in small, somewhat ordered clumps known as palmelloids (13). This phenotype is readily induced by certain organic acids (14) and by small algal grazers (15). Both Iwasa and Murakami (14) and Lurling and Beekman (15) reported that palmelloids occur in *2n* clumps more often than expected by chance, suggesting that members originate from a common mother cell. In addition, Lurling and Beekman (15) demonstrated that palmelloid formation requires light (and presumably growth and division). But the palmelloid phenotype is a plastic response (*i.e.* certain stimuli trigger this phenotypic change) which is too rapid to be evolutionary.

In this study we subjected several lines of *C. reinhardtii* to a simple treatment—brief, gentle centrifugation at each transfer (after Ratcliff (4))—to select for increased individual size in an attempt to induce the evolution of multicellularity.

Materials and Methods

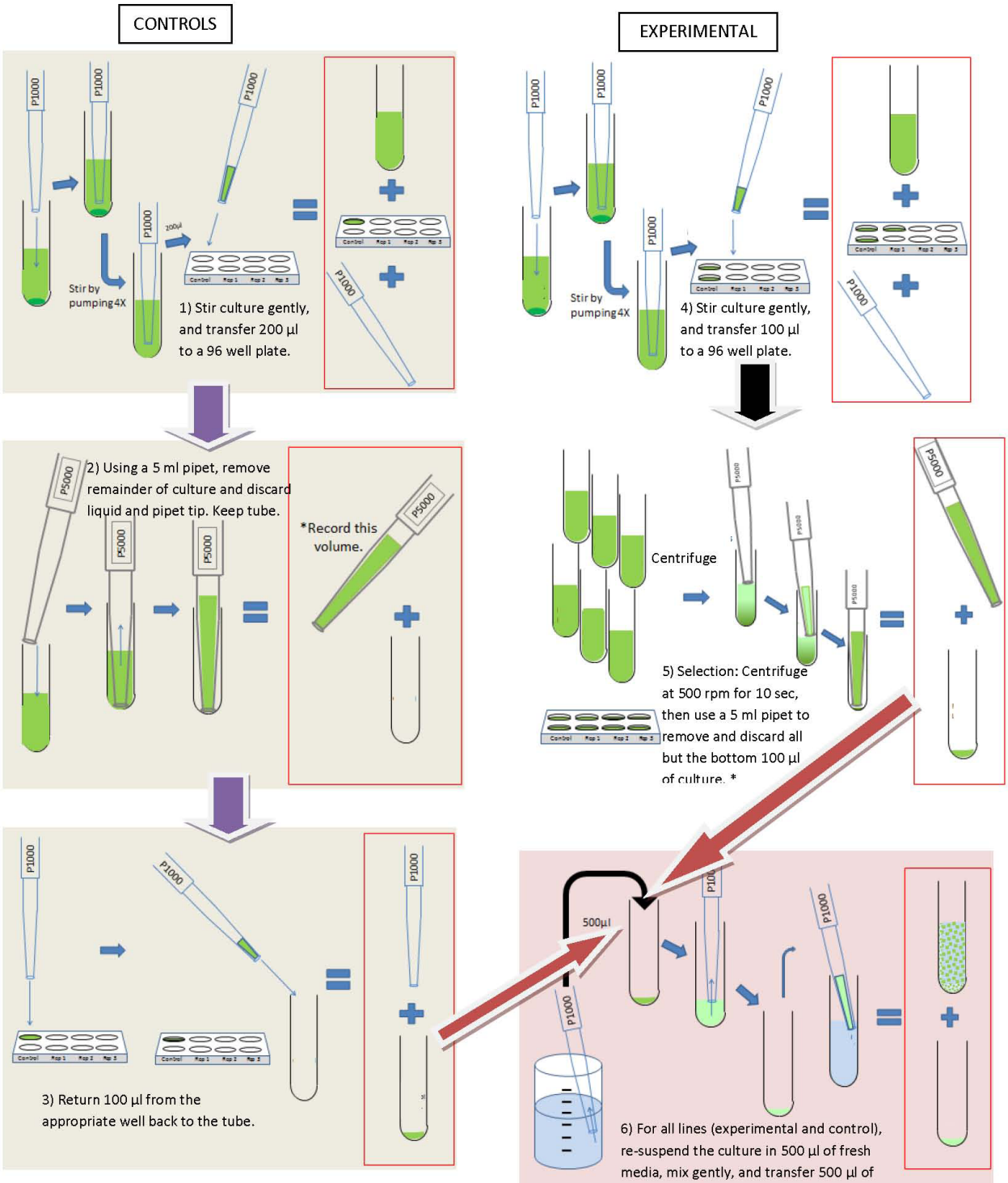
Nine *C. reinhardtii* strains were cultured in 5ml of sterile Bold's minimal media (16) in 11ml Kimax® glass culture tubes with Bacti Capall® lids and held loosely in white Nalgene® tube racks in a New Brunswick Scientific® Innova 43® incubator-shaker set to 123 rpm and 25.1°C. Available nitrogen prevented sexual reproduction, thus all cultures reproduced vegetatively (13). There were three replicates of the experimental treatment for each line, and one control (See Table 1).

Strains used in this experiment

Code	*Strain*	*Description and History*
A3	CC 2935 (*mt-*)	Wild strains isolated from Farnham, Quebec: 1993. Cultured on Bold's minimal media plates (agar 1.5%) at McGill University since 1993 (27). The population is grown on Bold's plates (1.5% agar) in the light, and is transferred every 15 days.
A4	CC 2936 (*mt+*)	
B2	A2S3 1/2	Base population for lines B2-C4. The ancestor was a cross of 3 lines performed by Clifford Zeyl in October 1992 as follows: (X•[Y•(Y•Z)]) where X is a mutant of lab strain CC-253 (*mt-)*, Y is CC-1952 wild-type (*mt-*), and Z is CC-2343 wild-type (*mt+*) (17). The population is grown on Bold's plates (1.5% agar) in the light, and is transferred every 15 days.
B3	ZIF Mass 2B	Zygotes in Flasks (ZIF): descendants of strain B2, cultured in flasks of Bold's minimal media (nitrogen-free to induce gametogenesis) with a suspended Nitex filter to serve as a substrate. Cultures are bubbled with a Pasteur pipet. After a cycle of 5 days of growth in the light followed by 5 days in the dark, the filter is removed and chloroformed, leaving only zygotes alive. A minimum of 200 zygotes is transferred to the next cycle.
B4	ZIF Mass 4	
C1	MOP Mass 1A	Mating on Plates (MOP): descendants of strain B2, grown on (1.5)% agar plates of Bold's minimal media (nitrogen-free to induce gametogenesis) in light for 10 days followed by 5 days in the dark (extra time in light because growth is slower on plates than in liquid media (13)). The plates are then chloroformed, and a minimum of 200 zygotes are transferred to the next cycle.
C2	MOP Mass 1C	
C3	Asex 1A	Descendants of strain B2, grown on nitrogen-containing Bold's minimal media plates (agar 1.5%) in light for 10 days followed by 5 days in the dark. Gametogenesis is not induced and the plates are never chloroformed, thus reproduction is strictly asexual.
C4	Unsel 2A	Descendants of strain B2, grown on Bold's minimal media plates (agar 1.5%) in light for 10 days followed by 5 days in the dark. A minimum of 200 individuals from each cycle for the subsequent inocculum. The medium is nitrogen deprived, thus allowing for gametogenesis, but the population is never chloroformed (i.e. sexual reproduction is permitted, but not selected for). This line has been exclusively asexually reproducing since the early 1990's.

Table 1: Names and descriptions of the Chlamydamonas reinhardtii strains used in the present study.

Fig. 1



* In order to remove the appropriate amount of liquid in step 5, average the volume removed from all control tubes in step 2. Add 20 µl to this number, and set the pipet to this volume.

**Note that the precise volumes of the newly inoculated vials are not known. Unequal evaporation during autoclaving causes this. This flaw was considered, but there is unequal evaporation within the incubator regardless, so this amount of error was accepted. See the 'limitations' section.

Six of the nine strains were derived from a seventh ancestral strain (B2). The ancestor was comprised of a cross of 3 wild-derived strains in 1992 (See Table 1) (17). These 7 strains were cultured under different conditions in the Bell laboratory at McGill University for 20 years prior to this study (see Table 1). All were grown in 500 µl liquid Bold's cultures for one week prior to the experiment. During the study, the cultures were kept in areas which never housed any algae other than *C. reinhardtii*, and procedures were carried out in a laminar flow hood. Thus the risk of contamination, especially from algal species, was minimal. Nine 2.8 cmX45 cm 15W Sylvania®Gro-Lux®F15T8/GRO/AQ/RP fluorescent lights (rated at 325 lumens) provided continuous illumination from the lid of the incubator, and the door window was covered with white plastic boards to create a uniformly lit environment.

Fig. 1 depicts step by step the transfer and selection procedures carried out every 3-4 days, corresponding to approximately 15 populations doublings (i.e. 5-6 cell cycles) (13, 18). The procedures were performed in a laminar flow hood to prevent contamination. Once removed from the incubator, all cultures (experimental and control treatments) were re-suspended by gentle pumping with a pipettor (Fig.1.1, Fig 1.4). At this point, 100 µl samples from each tube were stored in a 96-well plate for subsequent analysis (see below). Experimental treatment tubes were then centrifuged for 10 seconds at 500rpm (set to maximum acceleration and deceleration) using an IEC Centra GP8 ventilated centrifuge (cat. no. 216 4-place swinging bucket rotor with cat. no. 316S buckets and cat.no. 5712 bucket adaptors). All but the bottom 100 µl (2%) 'pellet' of each culture (experimental and controls) was then removed with a pipet and discarded. Centrifugation resulted in a higher proportion of heavy (clumped) cells in the experimental tubes' pellets compared to controls. This constituted a selection pressure in experimental lines which rendered the clumped phenotype more fit than a non-clumped or unicellular phenotype. The 'pellets' were re-suspended in fresh media (Fig. 1.6) and transferred to new culture tubes (~1.5% inoculum density). This procedure aimed to select solely on the basis of individual size by keeping environmental conditions the same between control and experimental cultures. This ruled out the possibility of external stimuli inducing a plastic response (15, 19).

Data Collection

The 100 µl culture samples were used for counting cells and tracking changes in individual morphology. Samples were loaded onto a haemocytometer, and viewed with a 10X objective (100X magnification). Pictures were taken with Q-Imaging© QICAM FAST-1394 in conjunction with the QCapture 2.68.6 software (©2001 Quantitative Imaging Corporation). The software settings were left at their default levels. Photographs showed the central portion of the grid, and cells were counted over an area of 0.8 mm^2 from these pictures. The photographs were taken of all populations on the day of each transfer. Counts were made for single cells, cell pairs, triplets, and quadruplets. Clumps of cells were photographed at higher magnifications to get more information about their structures.

Data Analysis

As Kirk (2) explains, the first step towards multicellularity should be the failure of daughter cells to separate following division. Therefore, we expected that putative multicellular individuals would contain 2^n cells. As the experiment continued, the frequency of cells existing in pairs increased, aligning closely with our expectations. During some samplings, 3 and 4 cell clumps were also prevalent. In such instances, the microscope hampered our ability to differentiate to differentiate between 3 and 4-celled individuals, so they were lumped together in the analysis. Larger clumps were extremely rare, and thus not quantifiable (see Notable Observations section).The evolutionary response to selection was quantified as the frequency of 2-, 3-, and 4-celled individuals relative to the total number of individuals as the response variable. In this analysis, each clump (1-cell, 2-celled, etc) was counted as an 'individual'. This fraction (log transformed to normalize the data) is plotted against time (transfer number) in Fig. 2 (generated in R with 'ggplot2', Hadley Wickham, 2012). For each strain, odds ratios compared the multicellular response of each experimental replicate to its control treatment using a logistic regression model where transfer (time) is a covariate (12 transfers, therefore DF=47, α=0.05) (Using 'glm', The R Core Team, 2007). A likelihood ratio test was then used to compute the significance of the overall effect of experimental selection treatment for each strain compared to its control ($\chi 2$ test DF=3, α=0.05) (computed in Microsoft Excel© 2010).

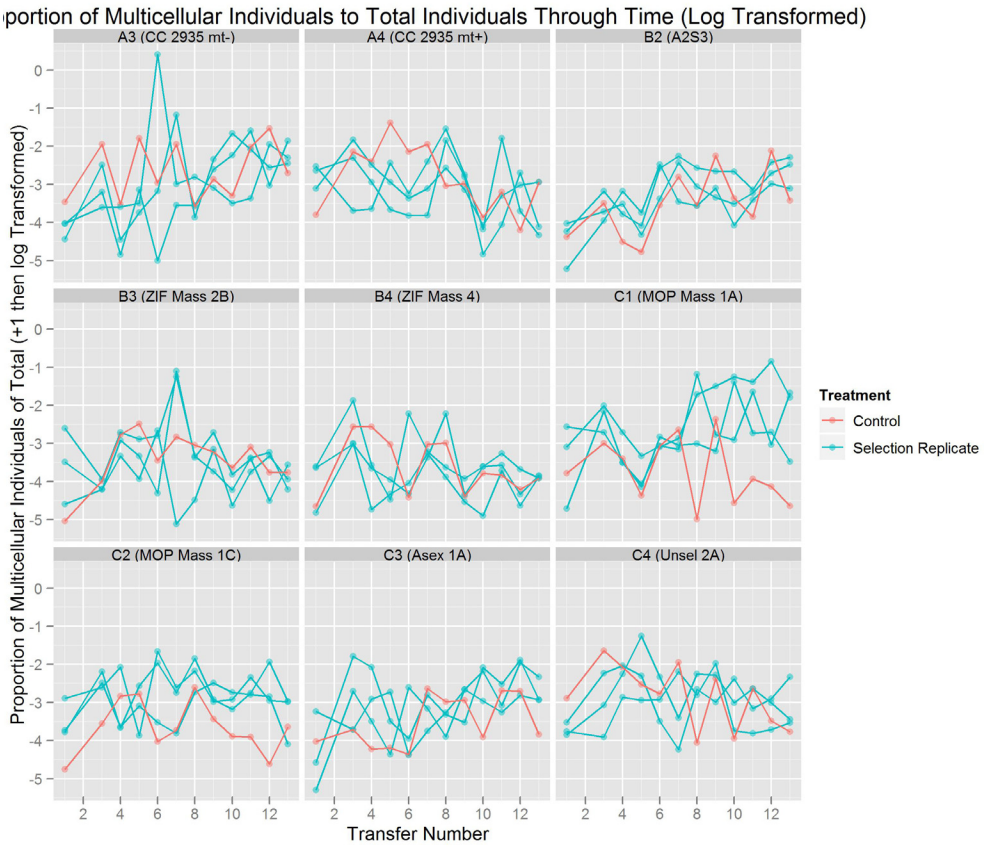


Fig. 2:
Plots showing the relative frequency of multicellular individuals through time. Refer to Table 1 for strain description.

Study of Paired-Cell Division

The object of this assay was to observe pairs of cells—putatively multicellular individuals—undergoing division. Synchronized timing of division within a pair would suggest that the individuals are daughter cells to a common mother, therefore representing a primitive multicellular state. A sterile screw-top vial containing 3.0ml of Bold's minimal medium (agarose 0.9 g/l at 45 °C) was inoculated with 50 µl of culture from line C1 line, which had by this point demonstrated the clearest trend toward purported multicellularity. The mixture was plated over solidified Bold's medium (agar 1.5 g/l), cooled, and placed open under the microscope. Fluorescent lighting in the room plus additional fluorescent and incandescent lamps around the microscope provided ample light for growth. Photographs were taken hourly when possible. When cleavage initiation was observed, pictures were taken every 20-30 min. During the initial 7 hours of illumination, photographs of 3 pairs and 2 single cells were taken from the 20X objective (200X magnification). The plate was left in dim light overnight with the lid on. Twelve hours later, illumination was restored, and a new group of cells was followed (again with the 20X objective). Once the plate was sufficiently dried out, it was possible to switch to the 40X objective (400X magnification). After another 10 hours of illumination, the room was left in absolute darkness for 24 hours, save for hourly 5 minute breaks for photography. Throughout this entire procedure, the stage of the microscope was left as undisturbed as possible so the progress of individual cells could be tracked.

Results

Count Data

The likelihood ratio tests showed significantly higher frequencies of multicellular individuals in experimental treatments compared to their respective controls in the ancestral line B2 and all of its descendants (B3-C4) except for line B4, which showed a significant result, but was skewed by a replicate which had significantly lower multicellularity frequencies than the control (See Table 2). Of the total individuals classified as multicellular (2-3-4-cell clusters) 88% were in the paired cell form, deviating minimally across strains. Selection replicates in lines C1 and C2 had the highest odds ratios (OR ranges 5.570-17.190 and 3.211-3.881 respectively) indicating a particularly high frequency of multicellular individuals compared to controls. Likelihood ratio tests also implicate treatment as a highly significant factor for these two strains, especially C1. These two strains represent historically sexually reproducing lines cultured on gel plates. Weaker responses (though still significant) were observed in C3 and C4 (historically asexual) and B3 and B4 (historically reproduced sexu-

ally in liquid media). While the selection pressure was a significant factor influencing multicellularity in the ancestor, strain B2, the response to selection was comparatively weak (OR range 0.999-2.208). The wild type isolates A3 and A4 (both single mating types) showed no significant response to selection.

Odds ratio and likelihood ratio tests for effect of selection treatment on multicellular phenotypic response (±=0.05 for all tests)				
Strain	Replicate	Odds Ratio	P|Z|(df=43)	Likelihood ratio test p|χ2|(df=3)
CC 2935 (mt-)	1	1.001	1.00E+00	3.96E-01
	2	0.989	9.60E-01	
	3	0.687	1.80E-01	
CC 2936 (mt+)	1	1.477	3.50E-01	6.82E-01
	2	1.573	3.10E-01	
	3	1.153	7.60E-01	
A2S3 1/2	1	0.999	9.90E-01	1.37E-09*
	2	1.354	1.10E-01	
	3	2.208	7.00E-06	
ZIF Mass 2B	1	3.391	5.10E-02	3.62E-02*
	2	1.762	3.91E-01	
	3	4.046	3.00E-02	
ZIF Mass 4	1	1.316	3.40E-01	2.57E-02*
	2	1.379	2.40E-01	
	3	0.695	2.40E-01	
MOP Mass 1A	1	5.570	1.80E-08	5.37E-52*
	2	6.863	1.60E-10	
	3	17.190	< 2.00E-16	
MOP Mass 1C	1	3.881	4.30E-06	8.37E-07*
	2	3.211	8.70E-05	
	3	3.764	1.30E-05	
Asex 1A	1	2.685	3.60E-05	3.32E-05*
	2	2.287	4.90E-04	
	3	1.674	3.60E-02	
Unsel 2A	1	2.041	1.10E-02	5.43E-05*
	2	0.764	4.41E-01	
	3	1.478	2.23E-01	

Table 2:
Odds Ratios compare the individual probability of being multicellular in an experimental replicate versus its control across 12 transfers. An odds ratio of 1.0 indicates equal probabilities of multicellularity in selection and control treatments. OR >1 indicates a greater probability of selection replicate individuals being multicellular. Likelihood ratio tests report overall significance of the selection treatment for each strain (α=.05).

Cell Division Study

Cell division was only examined in line C1, the line with the clearest multicellular response. Fig. 3 shows a sequence of photographs taken of two cell pairs (a-h, and i-n) over 14 hours. Note that the cell cycles of members within each pair are highly synchronized. Division is not synchronized, however, between the pairs. The pictures were not always optimally framed, but we ascertained two main points: 1) division begins after at least five hours of growth or inactive phases, and 2) division is completed within about three hours of the first appearance of the cleavage furrow. The cells were taken from line C1 Transfer15 (Nov 10th).

Discussions

Six of the seven lines derived from the common outcrossed ancestor showed an increase in the frequency of multicellular individuals. Furthermore, the ability to evolve this multicellular phenotype appears to be present in the ancestral line, B2. Lines B3and B4 (ZIF Mass—sexually reproducing in flasks) and C1 and C2 (MOP Mass—sexually reproducing on plates) have the highest diversity of allele

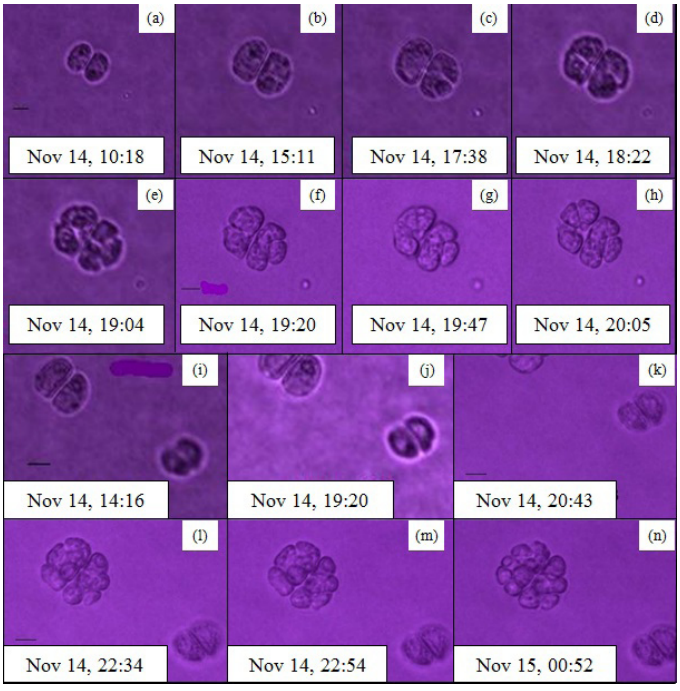


Fig. 3
Two different cell pairs (a-h, i-n) from line C1 transfer 15 undergoing division on the same agarose plate. At least five hours of growth precedes approximately three hours of synchronous multiple division of the cells within each pair.(a-f) 200X magnification; (g,h), 400X. Cleavage of second pair begins at (k). (i-j) 200X; (k-n), 400X. Scale bars = 20 μm.

combinations present in the ancestral A2S3 population because these four had been sexually reproducing since 1992 (17). Errors in crossing over could also have led to gene duplication and subsequent divergence of function, or to gene modification. That the line emerging with the strongest tendency towards multicellularity was one of these four previously sexually reproducing is consistent with the expectation that genetic diversity increases evolutionary potential (20). It is unclear, however, why line C1 exhibits such a remarkably strong and statistically significant response. Agar-base ancestral culture conditions may have somehow predisposed lines C1 and C2 to evolving multicellularity.

The multicellular response could mean one of three things: 1) cells have increased the duration of cell division or the frequency of division, 2) cells are stickier, and tend to cohere, or 3) individuals actually exist in a paired or multicellular state. To evaluate these three hypotheses, we monitored division of several cell pairs from one transfer of line C1—the line where selection appeared to be having the greatest effect (See Fig. 3). *C. reinhardtii* typically undergo a number of cleavages in rapid succession following 6-12 hours of growth (18, 21-23). The number of cleavages is determined in each cell during interphase (prior to division) and varies within a population (18, 21-23). The two pairs shown in Fig. 3 grow or retain a constant volume for at least five hours while in a paired state before rapid multiple cleavage occurs over about three hours. If the paired cells were actually just dividing cells, we would not have observed a gap between the first and subsequent cleavages. Furthermore, the cells within a pair are the same size as single cells (not shown). In addition, cultures were not synchronously dividing. (synchroniziationrequires prolonged exposure to alternating 12 hour light-dark periods (13)). There is evidence for asynchronicity in Fig. 3—one pair begins cleavage prior to 17:30, the other begins after 19:20. Within the pairs, however, cleavage is highly synchronized between cells, and each cell gives rise to the same number of daughter cells. In Fig. 3, pair 1 (i-n) undergoes one more cleavage than pair 2 (a-h). If the pairs were simply comprised of unrelated cells which had stuck together, then one would not expect this intra-pair synchronization. Therefore, these photographs suggest that both cells within a pair are derived from the same mother cell. Cytoplasmic junctions may be maintained between members of a pair of daughter cells—a crucial step in the evolution of volvocacean multicellularity (2). Or, if the cells are siblings, their cell cycles could be inherently synchronized with no need for retaining cytoplasmic junctions. An alternative possibility is that cells within a pair are unrelated, but cohered following a collision, and formed cytoplasmic junctions which allowed for cell cycle synchronization.

That a multicellular phenotype never became fixed in any line is noteworthy. It has dramatically increased in frequency, but by no means dominates the cultures. Perhaps this is due to a trade-off effect; cell clumping reduces the nutrients available for cells to consume (15). Centrifugation may have selected for heavier cells at each transfer, and thus multicellularity, but any single-celled individuals which managed to be transferred would have proliferated faster. This would keep the frequency of multicellular individuals relatively low. Another possibility is that the experiment was too short to allow for fixation of the multicellular phenotype.

Notable Observations

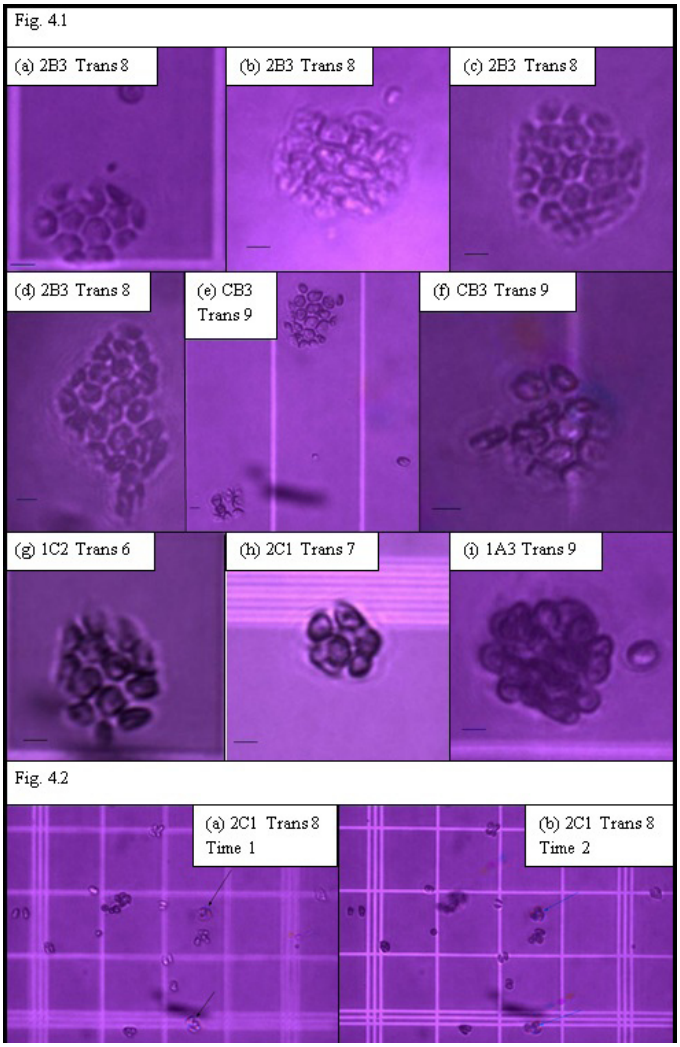


Fig. 4
Two different cell pairs (a-h, i-n) from line C1 transfer 15 undergoing division on the same agarose plate. At least five hours of growth

Flat Green Aggregates: On three separate occasions, we observed flat clumps of green cells arranged neatly in an alternating hexagon/pentagon pattern (particularly Fig. 4.1 a,c,f). These aggregates were generally circular, although several were irregular. When we attempted to wash these clumps from the haemocytometer and into a 12 well plate, they proved to be very sticky, often disintegrating rather than washing off. Given this observation and that several of

these aggregates appeared in a sample taken from a control line, we hypothesize that these clumps were a result of wall growth. Though the phenomenon was rare, such aggregates tended to co-occur, with several being found on the same slide. Perhaps this is an evolved phenotype selected by the transfer procedure. But if this is the case, it is obviously not a tremendously advantageous phenotype, because it never achieved a frequency of greater than 1% in any given culture tube. Like colonies of the genus *Gonium*, these aggregates were flat, and the interlocking cells resembled those of *Pandorina* colonies (24, 25). Unlike either of these genera—the two next highest major clades in the Volvocacea (2)—these aggregates were non-motile.

Spherical Aggregates: During the sixth transfer, we discovered an apparent sphere of 13 or more cells (Fig. 4.1 g) in line C2. It was immotile, but clearly three dimensional, and approximately 50um in diameter. Unfortunately, we failed to preserve the specimen. During the next transfer we noticed a smaller spherical group of cells (about 8) in line C1 (Fig. 4.1 h). We were unable to preserve this specimen. Both aggregates (Fig. 4.1 g,h) resembled Pandorina colonies, exhibiting the "keystone" morphology described by Angeler (25) though again, both were immotile.

On several other slides, we observed the occasional structure similar to that of palmelloids (Fig. 4.1 i) shown in Lurling and Beekman (15). Again, these occurred at very low frequencies (less than five instances observed during the whole experiment). The significance of the above formations is unknown.

Motile Aggregates: *C. reinhardtii* absorb their flagella during division, as their anchoring basal bodies are required for mitosis (26). Therefore, moving aggregates would be convincing evidence of multicellularity. Once in a control of line C4 and once in an experimental replicate of line C1, we observed loosely associated aggregates of around seven cells cartwheeling about the slide. The C1 aggregate was fairly symmetrical, but it lost a few cells over the course of 30 minutes and was determined to be nothing more than temporarily cohering cells. During transfer seven, however, we observed that nearly 25% of individuals in one replicate of line C1 (Fig. 4.2) were 3-4 cell clumps—and most were motile! They slowly rolled in apparently random paths across the surface of the slide, in a manner similar to that of *Pandorina*. We never witnessed such an event again, despite efforts to preserve this unique population.

Limitations

While this study yielded promising results, it has several limitations. One concern with the design is the imprecision of the transfer. Unequal evaporation between tubes prevented implementation of a constant and precise selection pressure. In addition, removing the supernatant proved to be a very difficult and imprecise procedure. Because there was no true 'pellet', there was inevitable mixing between the 'supernatant' and the presumptive inoculum. Undoubtedly, this lessened the selection pressure. In addition, large mats often formed at the water line on the walls of many of the tubes. These occasionally were dislodged, and ended up in the pellet. The effect of this contamination is unknown.

This imprecision may have led to the erratic data seen across transfers. In analyzing the data, we were unable to find any significant effect of time on the frequency of multicellular individuals. This is troubling, because if the response is evolutionary, we would expect to see an increase in the multicellular phenotype through time. While Fig. 2 suggests that lines B2, C1 and C2 do in fact have an increased frequency of multicellular individuals over time, we were not able to find statistical significance in this trend. This prevented us from completely ruling out the possibility that the selection procedure somehow induced a plastic palmelloid response.

In addition, low-resolution microscopy precluding detailed observations of cell morphology. As mentioned earlier, the ability to see the flagella of cells would have been very informative. Furthermore, cells in clumps may have obscured other cells within the clumps (i.e. a clump counted as 3 cells may actually have contained 4). Electron microscopy could resolve the ambiguities left by the current pictures.

Finally, this investigation lacks rigorous replication of controls. Only one control line was used per strain to compare against three experimental replicates. Given the sporadic nature of the data, a balanced design with more controls would have improved statistical power. Similarly, the study of cell division was not sufficiently replicated to extrapolate quantitative conclusions to all paired cells in this experiment. Future studies could incorporate this into their design using automated photography.

Conclusions

Selection of heavier individuals through centrifugation prior to each transfer appears sufficient to induce the beginning of a transition to true multicellularity in genetically diverse strains of *C. reinhardtii*. The increase in small clusters, particularly in the form of pairs, is consistent with the putative first steps in achieving multicellularity—incomplete daughter cell separation (2). We have gathered photographic evidence that suggests that the multicellular phenotype in this line is not merely an artifact of division or cell collision and cohesion, and that the cells within a pair are likely genetically identical. The response occurred over a fairly short period of time (around 50 generations) but the multicellular phenotype was never fixed in the population. Expanding upon this study could be invaluable in elucidating the mechanisms underlying one of the most important innovations in the history of life.

Acknowledgements

Many thanks to Etienne Low-Decarie for providing the bulk of the R scripts, and to Josianne Lachapelle for her help with various R-related issues. Both also gave plenty of much-needed advice in set-

ting up the experiment. Thanks to Larry for consultation regarding statistics. Many thanks to Dr. Bell for giving me the opportunity to undertake this experiment, and for sharing his wealth of knowledge along the way. This study was conducted September to December of 2011. Funding was provided by the National Sciences and Engineering Research Council of Canada.

References

[1] J.T. Bonner. Integr. Biol.: Issues, News, and Reviews 1, 27-36 (1998).
[2] D.L. Kirk. BioEssays 27, 299-310 (2005).
[3] R.K. Grosberg, R.R. Strathmann. Annu. Rev. Ecol. Evol. Syst. 38, 621-654 (2007).
[4] W.C. Ratcliff, paper presented to the Society for the Study of Evolution, Norman, Okl, USA. (18 June 2011).
[5] M. Boraas, D. Seale, J. Boxhorn. Evol. Ecol. 12, 153-164 (1998).
[6] J.H. Koschwanez, K.R. Foster, A.W. Murray. PLoS Biol. 9, e1001122 (2011).
[7] B. Holmes. New Sci. (2011).
[8] Y.J. Liu, B. D. Hall. Proc. Natl. Acad. Sci. U.S.A. 101, 4507-4512 (2004).
[9.] M.D. Herron, R.E. Michod. Evolution. 62, 436-451 (2008).
[10] T. Pröschold, B. Marin, U.G. Schlösser, M. Melkonian. Protist 152, 265-300 (2001).
[11] T. Muller, S. Rahmann, T. Dandekar, M. Wolf. BMC Evol. Biol. 4, 20 (2004).
[12] A. Larson, M. Kirk, D.L. Kirk. Mol. Biol. Evol. 9, 85-105 (1992).
[13] E.H. Harris, D.B. Stern, G. Witman. The Chlamydomonas Sourcebook: Introduction to Chlamydomonas and its laboratory use. (Academic Press, 2008).
[14] K. Iwasa, S. Murakami. Physiol. Plant. 21, 1224-& (1968).
[15] M. Lurling, W. Beekman. Ann. Limnol. - Int. J. Lim. 42, 65-72 (2006).
[16] H.C. Bold. Bull. Torr. Bot. Cl. 76, 101-108 (1949).
[17] G. Bell. J. evol. biol. 18, 722-34 (2005)
[18] M. Vítová, et al. Planta. 234, 599-608 (2011).
[19] D.O. Hessen, E. Vandonk. Arch. Hydrobiol. 127, 129-140 (1993).
[20] R.D.H. Barrett, D. Schluter. Trends Ecol. Evol. 23, 38-44 (2008).
[21] K. Bisova, D.M. Krylov, J.G. Umen. Plant Physiol. 137, 475-491 (2005).
[22] L. Donnan, P.C.L. John. Nat. 304, 630-633 (1983).
[23] H. Oldenhof, V. Zachleder, H. Van den Ende. Folia Microbiol. 52, 53-60 (2007).
[24] M. Hayama, T. Nakada, T. Hamaji, H. Nozaki. Phycol. 49, 221-234 (2010).24.
[25] D.G. Angeler. Hydrobiol. 369-370, 269-275 (1998).
[26] U.G.Johnson, K.R. Porter. J. Cell Biol. 38, 403-425 (1968). 27. L. Sack et al. Note. J. Phycol.. 30, 770–773(1994)

RESEARCH ARTICLE

Neuronal spiking is better than bursting at predicting motion detection in area MT

Mathura Ravishankar[1]*, Erica McKenzie[1]

[1] Faculty of Science, McGill University, Montreal, QC

*Email Correspondence:
mathura.ravishankar@mail.mcgill.ca

Keywords:
Spiking: *Spiking is the brain's mechanism to transmit stimulus information by a rate code (spikes per unit time).*
Bursting: *Bursting is a dynamic state in which a neuron repeatedly fires a number of action potentials within a time period (1).*
Middle Temporal Visual Area (MT): *The MT area of the macaque monkey is part of the extrastriate visual system, and is concerned with processing visual motion.*
aROC: *The area under the receiving operator characteristic curve is a graphical technique that determines the level of correlation between two events.*
Coherent Motion: *When a percentage of the random dot patch (RDP) moves in the same direction, the result is coherent motion. This is depicted in Fig. 1.*

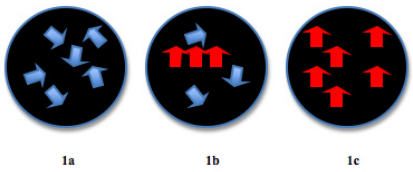


Fig. 1
This is a depiction of varying coherent motion presented to a neuron's receptive field. 1a depicts 0% coherent motion. 1b depicts 50% coherent motion. 1c depicts 100% coherent motion.

Abstract

Background: The middle temporal visual area (MT) is widely studied in visual processing and in integration of motion signals to form general perceptions. The objective of this study is to determine whether neuronal bursting in area MT of monkeys is more predictive of motion detection than neuronal spiking.

Methods: Two *Macaca mulatta* (macaque) monkeys were trained in Dr. Erik Cook's lab to detect coherent motion while connected to microelectrodes that determined their neuronal spiking activities. Using MatLab, we manipulated the collected data to determine whether spiking or bursting is more predictive of motion detection.

Results: We repeatedly found that neuronal spiking in area MT is better than bursting at predicting motion detection in macaques ($p < 0.01$).

Conclusions: Therefore, our results suggest that area MT neurons do not fire behaviourally-meaningful bursts in response to coherent motion. This finding is useful for learning about the visual processing pathway, and how information is coded in the brain.

Limitations: A key limitation of this study is that we did not exclude any experiments from analysis to control for quality of the collected data, perhaps leading to confounding factors.

Introduction

There are many ways to look at the temporal features of a neuron's spike train and determine its significance to an animal's behaviour (2). For example, studies have looked at the average spiking rate after presentation of stimulus without particular attention to the frequency of spikes or the number of inputs that a neuron integrates. Some have focused more on the firing pattern of groups of neurons rather than isolated neurons while others have looked at neural bursting activity. This paper will further look at bursting, along with average spiking.

Research in the pyramidal neurons of the hippocampus best highlights the bursting phenomenon (3). In this area, bursts of 2 – 6 spikes for a duration of 25 milliseconds or less have been tracked through extracellular recordings (4). The spiking rate of a presynaptic neuron at the central synapse is not predictive of the spiking rate at the postsynaptic neuron; this supports the notion that bursting may be a better predictor of behaviour than the average spiking rates (5). This was first studied in motoneurons and Mauthner cells, and later in different brain areas like the primary visual cortex (V1) (2). For example, both average spiking rate and bursting increase when a stimulus is presented to a V1 neuron's receptive field. However, synapses are sometimes "unreliable", which means that an action potential at the presynaptic neuron does not trigger an action potential at the postsynapstic neuron. Though this poses a problem with isolated spikes, rapid successions of presynpatic input onto the postsynaptic neuron can increase the probability of an action potential firing at the post-

synaptic neuron. Hence, bursting in a V1 neuron can encode more relevant information than isolated spikes in a V1 neuron, which is more indicative of noise (2).

The middle temporal visual area (MT), an area of extrastriate visual cortex, plays an important role in visual processing. A major input to this region is a magnocellular-dominated projection from layer 4B of V1. The MT visual area also receives projections from V2 and V3 - which are downstream of V1 - and directly from the lateral geniculate nucleus. Each MT neuron is tuned to a particular receptive field, and selective for a motion direction and speed to which it responds most vigorously. Area MT projects to downstream regions such as the ventral intraparietal area (VIP) and the medial superior temporal (MST) area (6). The MT region is crucial in motion perception, control of eye movements, and in integration of motion signals into a general perception (6). Since bursting has been shown to play an important role in V1, this paper looks at whether bursting also plays a role in area MT.

In this paper, we will explore two research ideas with respect to neuronal activity in area MT. Firstly, we will determine which bursting parameter for neurons in area MT is most predictive of detecting coherent motion. Secondly, we will compare the average spiking rate and bursting of neurons in area MT and determine which of the two is more predictive of motion detection in *Macaca mulatta* (macaque) monkeys.

Methods

Behavioural Task

Dr. Cook's lab trained two macaques to detect coherent motion while connected to tungsten microelectrodes, which recorded their neuronal spiking activities. Each experiment used one monkey, and each experiment consisted of a varying number of trials. Each trial recorded the activity of two different neurons, neuron 1 and neuron 2, from the same hemisphere of the monkey in question. Data was obtained from 19540 trials over the course of 50 experiments. The lab determined the location of each neuron's receptive field, along with its preferred speed, orientation, and direction of stimulus before the beginning of an experiment. Then, Random Dot Patches (RDP) were presented to that neuron's receptive field with increasing coherence, starting at 0% and adhering to the neuron's preferences.

There were three conditions in each trial. Condition 1 represented coherent motion in the receptive fields of both neuron 1 and neuron 2. Condition 2 presented coherent motion only in the receptive field of neuron 1. Condition 3 presented coherent motion only in the receptive field of neuron 2. Anywhere between 500 and 10,000 ms after the onset of RDP presentation, the RDP was presented with coherent motion in the receptive field(s) for 50 ms (Fig. 2). The monkeys were trained to maintain a fixation point on the screen, and release a lever for a juice reward if they correctly detected coherent motion. A trial was considered correct only if the monkey released the lever within a window of 200-800 ms after coherent motion turned off, and incorrect in every other instance. All trials were organized into two groups for analysis: correct and incorrect. We analyzed the electrophysiological data based on the 100 ms time window after coherent motion was turned off.

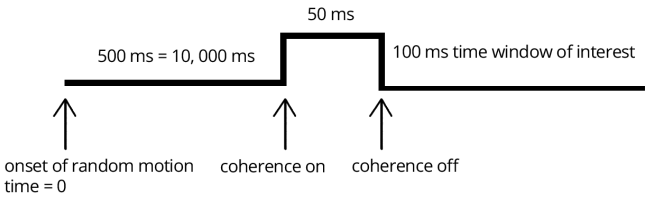


Fig. 2
This is a graph of the presented stimulus on a timeaxis. There are four times that are important for this experiment:
1) onset of random motion - this is when the RDP is presented to a neuron's receptive field at time 0
2) coherence on – this denotes when coherent motion is presented in a receptive field
3) coherence off – this denotes when the coherent motion stops
4) the 100 ms after coherence off – this was used for analysis.

Analysis of Data

We used a time period of 100 ms for our analysis because previous studies have reported that neural–behavioural covariation is greatest for this time window (7) . We used values for the standard area under the receiving operator characteristic curve (aROC) to determine the probability of motion detection that is correlated to a specific bursting parameter or to a spiking activity. If an aROC value of 0.5 is returned, this suggests that there is no correlation between the number of correct trials and the neuronal activity in question. However, aROC values greater or less than 0.5 suggest a greater predictive capacity (8) and correlation between the number of correct trials and neuronal activity. What matters is not whether the aROC value is greater or less than 0.5, but the absolute difference between the aROC value and 0.5 (ie. aROC values of 0.6 and 0.4 have an equal predictive capacity). The higher the absolute difference between 0.5 and the aROC value, the higher the predictive capacity (8). The bursting parameters we used varied from 1 – 5 spikes for 10 – 100 ms with 10 ms steps.

In order to produce aROC values for the different bursting parameters, we first calculated the distribution of the number of bursts within the 100 ms time period after coherence was turned off for both correct trials and incorrect trials averaged over both neurons in condition 1. We inputted these distributions into the MATLAB function detect probability, which outputted an aROC value. This function

was written by members of Dr. Cook's lab. These values could be compared across different bursting parameters, allowing us to determine which resulted in the best covariation with behaviour. The same metric was used to analyzed the covariation between average spiking rate and the behavioural outcome. This allowed us to determine whether bursting rate or spiking rate is more closely correlated to correct motion detection in monkeys.

Results

1) Different Bursting Parameters in Condition 1, Neuron 1 and Neuron 2 combined

We determined one aROC value for each of the different combinations of bursting parameters, which produced a 10 by 5 matrix (Fig. 3). Fig. 4 graphically depicts the aROC value for a given burst parameter. As evident by Fig. 3 and Fig. 4, the highest aROC value we obtained is 0.5742, for the bursting parameter 1 spike/10 ms. The second highest aROC value is for 1 spike/20 ms, and the third highest aROC value is for 2 spikes/20 ms. (Fig. 3 and 4). For further analysis, we discounted the aROC values for bursts with only 1 spike, because we believe a neuronal burst should have multiple spikes. Instead, we used the aROC value obtained for 2 spikes/20 ms.

With the bursting parameter 2 spikes/20 ms, we calculated the average number of bursts in a trial (Fig. 5). We found the average number of bursts for correct trials to be 1.6371, and the average number of bursts for incorrect trials to be 1.2537. After applying the t-test for the significance of the difference between the means of two independent samples, the p-value we obtained for this result is less than 0.01. All p-values in this paper were obtained through the same statistical test.

	1 spike	2 spikes	3 spikes	4 spikes	5 spikes
10 ms	0.5742	0.5593	0.5348	0.507	0.502
20 ms	0.5694	0.5666	0.556	0.5393	0.51
30 ms	0.5642	0.5636	0.5599	0.5549	0.5263
40 ms	0.562	0.5633	0.5595	0.5516	0.5431
50 ms	0.558	0.5603	0.5597	0.5559	0.5482
60 ms	0.5555	0.555	0.5584	0.559	0.5465
70 ms	0.5578	0.5577	0.5609	0.5539	0.5461
80 ms	0.5545	0.5544	0.5538	0.5537	0.5472
90 ms	0.5528	0.5528	0.5525	0.5522	0.5428
100 ms	0.5526	0.5526	0.5526	0.5534	0.5436

Fig. 3
This is the matrix obtained when we used different bursting parameters. The ones highlighted in red gave the highest aROC values.

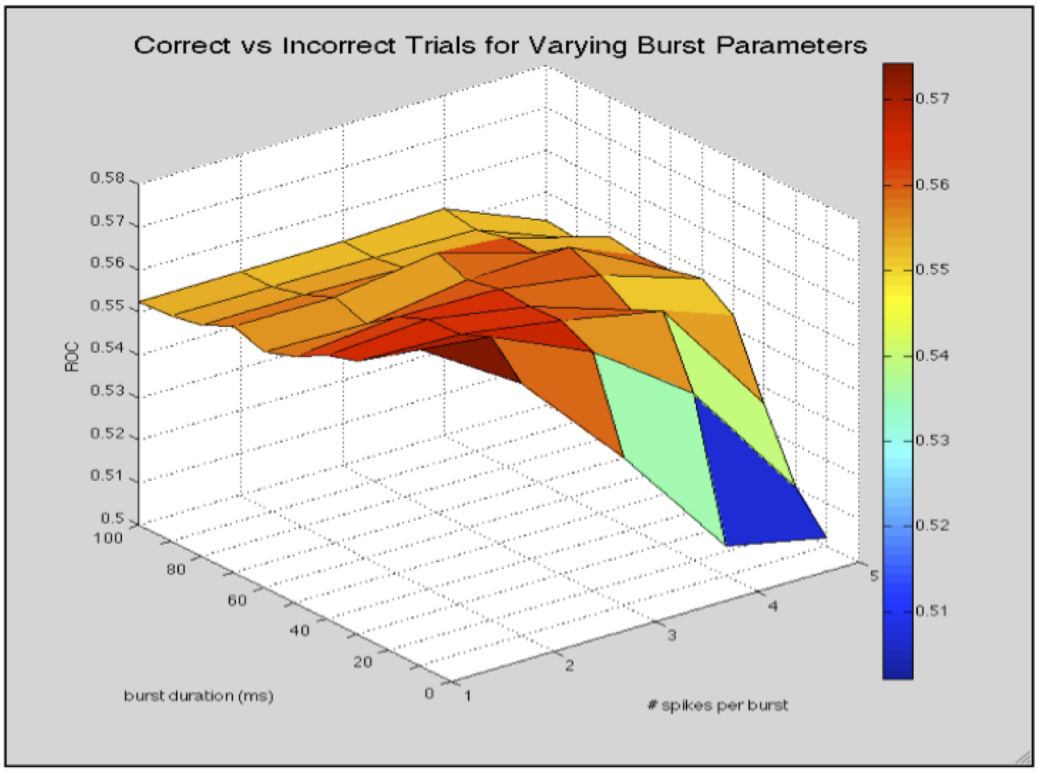


Fig. 4
This graph represents the same information as the matrix in Fig. 1.

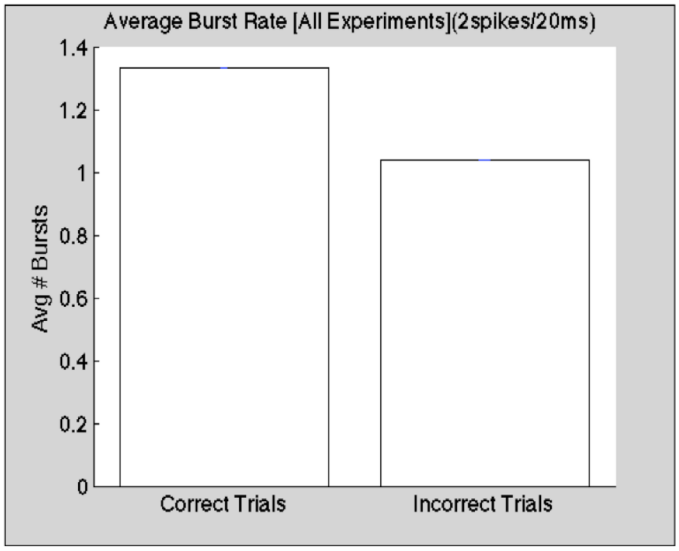


Fig. 5
This bar graph depicts the average number of bursts per correct trial and incorrect trial. As our obtained p-value was much smaller than 0.01, the plotted error bars are difficult to see.

As a post-hoc analysis, we wanted to see if the percentage of coherence in the RDP would alter our results. Of all 50 experiments, we found five experiments that represent the highest level of coherence: 80% coherent motion in both receptive field 1 and receptive field 2. When we ran our analysis again only for these five experiments, we found that the aROC values were higher in general. However, the best bursting parameters remained the same: 1 spike/10ms had an aROC value of 0.6715, 1 spike/20 ms had an aROC value of 0.6503, and 2 spikes/20 ms had an aROC value of 0.6454.

We calculated the average number of bursts only for the five experiments that had 80% coherence in the two receptive fields for the bursting parameter 2 spikes/20 ms (Fig. 6). We found the average number of bursts for correct trials to be 1.3317, and the average number of bursts for incorrect trials to be 1.0390 ($p < 0.01$).

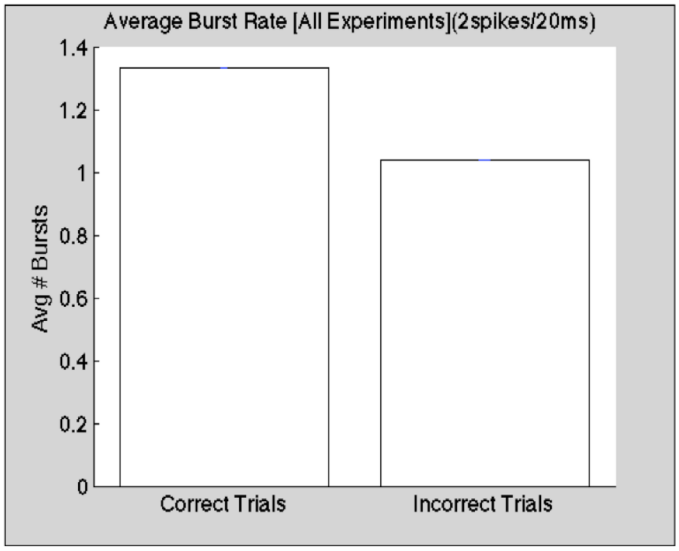


Fig. 6
This bar graph depicts the average number of bursts per correct trial and incorrect trial for the 5 experiments with 80% coherence in receptive field 1 and 2, $p < 0.01$.

2) aROC value for Isolated Spiking Activity

When the "detect probability" function was run for isolated spikes and averaged over neuron 1 and neuron 2 of all experiments, we obtained an average aROC value of 0.5961. We conducted the same post-hoc analysis (as described in section 1 of results) for the five experiments with 80% coherence in both receptive fields, and found an aROC value of 0.7163.

Discussions

After creating a matrix with different bursting parameters, we found the highest aROC values for 1 spike/10 ms, 1 spike/20 ms, and 2 spikes/20 ms. This suggests that these bursting parameters have the highest predictability of behaviour. The aROC values we obtained for 2 spikes/20 ms and for average spiking under Condition 1 are 0.5666 and 0.5691, respectively. Since both can be rounded to 0.57, the difference between spiking and bursting in Condition 1 is not significant. For the bursting parameter 2 spikes/20 ms, we found the average number of bursts for correct trials to be 1.6371, and the average number of bursts for incorrect trials to be 1.2537 ($p < 0.01$), asserting that there is a significant difference in neuronal bursting between correct trials and incorrect trials.

In order to determine if the percentage of motion coherence in the RDP played a role in our results, we analyzed the experiments that presented 80% coherence (the maximum coherence across all experiments) in both receptive fields. We found five experiments that

satisfied this condition; when we ran our analysis for the different combinations of bursting, we found the same bursting parameters (1 spike/10 ms, 1 spikes/20 ms and 2 spikes/20 ms) produced the highest aROC values. However, these aROC values were still not as high as the one obtained for spiking in these five experiments, suggesting that bursting is not better than spiking at predicting monkey behaviour - even when motion coherence is controlled for.

No experiments were excluded from analysis, perhaps contributing to potential confounding factors. There are several reasons why an experiment should be excluded: if the monkey pulled the lever before coherence was turned on or if the number of trials in an experiment was too low, for example. Furthermore, the aROC values obtained were not used to calculate a sensitivity index, which would determine the signalling reliability of the neuron (7). This type of analysis would compare the number of bursts and spikes before coherence was turned on and after coherence was turned off. It would also be worthwhile to determine if these neurons experience periods of quiescence, a characteristic of bursting neurons, and if these neurons were bursting in groups rather than in isolation.

Conclusions

Our results suggest that neuronal bursting is not behaviourally-relevant in area MT of macaques. The correlation between burst rate and behavioural outcome is weaker than the correlation between spike rate and outcome, indicating that single spikes carry more valuable information about motion to downstream processing areas. This result is congruent with Lisman *et al.*'s report on bursting in area MT, which concluded that, "if only bursts are considered, there is a marginally poorer estimate [of direction]. In this case it is clear that single spikes carry information" (2). Our finding enhances our understanding of how behaviorally-relevant information about motion is encoded by area MT neurons, and could have implications for how motion is processed in downstream areas with more complex response properties.

Acknowledgements

We would like to thank Dr. Cook and his lab at the Department of Physiology for teaching us how to use MATLAB and for guiding us through all the steps of this project.

References

[1] E. M. Izhikevich. *Scholarpedia* (2006).
[2] J.E. Lisman, *Trends Neurosci*. **20**(1) 38 (1997).
[3] S. E. Fox and J. B. Ranck, *Exp. Neurol*., **49**(1) 299-313 (1975).
[4] J. B. Ranck, *Exp. Neurol*., **41**(2) 461-531 (January 1973).
[5] J. J. Jack, S. J. Redman and K. Wong, J. Physiol., **321**(1), 65-96 (January 1981).
[6] R. T. Born and D. C. Bradley, *Annu. Rev. Neurosci.*, **28** (January 2005).
[7] J. E. T. Smith, E. P. Cook and C. A. Zhan, *J. Neurosci*., **31**(38) 13458-13468 (September 2011).
[8] N. R. Cook, *Circulation*, **115**(7) 928-35 (January 2007).

RESEARCH ARTICLE

Force-fluctuation physics of confined DNA: probing the breakdown of the Marko-Siggia law

Haider Riaz[1]*, Alex R. Klotz[1], Dr. Walter W. Reisner[1]

[1] Department of Physics, McGill University, Montreal, QC

*Email Correspondence:
haider.riaz@mail.mcgill.ca

Keywords:
Biophysics,
DNA,
Nanotechnology, Elasticity.

Abstract

Introduction: The study of polymers in nanofluidic systems such as nanopores and nanochannels is an important avenue of research in the physical and life sciences today. Complex nanofluidic devices containing varying topography are ideal for quantifying the behaviour of polymers under confinement. This study investigates the Marko-Siggia force-extension relationship under confinement. We measure the transverse fluctuations of deoxyribonucleic acid (DNA) confined between two pits in a nanofluidic slit to measure the potential breakdown of this model.

Methods: We took images of fluorescently tagged single DNA molecules in the nanopit array with video-fluorescence microscopy and analyzed the standard deviation of the peak position along the direction of the molecular extension.

Results: We were able to measure the parabolic relation between position and the strength of transverse fluctuations. By examining the peak variance as a function of slit height, our results indicate that the two dimensional version of the force-fluctuation relationship may be appropriate in the limit of strong confinement. However, we could not consistently measure the absolute length of the DNA stretched between two pits, which prevented us from fully exploring the force-extension and force-fluctuation relationships of confined DNA.

Conclusions: These experiments further demonstrate that nanofluidic confinement serves as a useful tool for testing the extreme properties of polymers, and our results suggest further investigation into the breakdown of the Marko-Siggia force law.

Introduction

The development of nanofluidic devices such as nanopores, nanochannels, and nanoslits has led to significant advancements in many areas of research (1). These devices have small volumes and provide constrained dynamics, allowing individual biological molecules to be studied with remarkable spatial and temporal resolution. Nanofluidics also have applications in next-generation genetic mapping technologies (2).The physics of deoxyribonucleic acid (DNA) molecules under confinement is fundamentally different than in bulk form (3). The unique behavior displayed by polymers under confinement is a significant area of study in nanoscience. Understanding it is key for the further development of a diverse array of fields ranging from polymer electronics (4) to polymeric membranes (5).

A polymer is a molecule made up of a number of repeating "monomer" units in a chain. There is no long-range order in orientation along the polymer, and its structure is described by a random walk (6). To maximize entropy, a polymer randomly coils with a characteristic radius. When stretched, the number of possible configurations decreases, and due to thermal fluctuations, the polymer will relax to a more condensed and higher entropy state: the polymer tends to bunch as there are more ways for it to be bunched than stretched. This phenomenon is known as entropic elasticity. Biological polymers such as DNA are often described by the worm-like chain model of entropic elasticity, which correctly predicts rigidity over a short range known as the persistence length. This model was tested by Marko and Siggia (7), who derived the eponymous force law relating the force acting on the ends of a polymer to its relative extension. More recently, Baba *et al* (8) derived a relationship between the

magnitude of transverse fluctuations of a stretched polymer and the force acting on it.

The behaviour of DNA changes when it is confined to regions below its characteristic size. While modifications to DNA's equilibrium structure have been well characterized (3), it is unknown whether the Marko-Siggia force law is modified by confinement. Here, we use a nanopit-nanoslit system to test the effects of confinement on the force-extension and force-fluctuation relationships of DNA. The nanopit-nanoslit system consists of a 100 nanometer slit, which is embedded with a lattice of square pits. This slit is small enough that the molecule behaves as if in two dimensions. The molecules have more ways to orient themselves in the pits than in the slits, so the pits can be said to act as entropic traps. Such a system has previously been used to demonstrate directed molecular organization (9) and nonlinear modifications to molecular diffusion (10). We intend to use this system to measure the transverse fluctuations and compare them to polymer theory.

The traditional worm-like behavior of the Marko and Siggia force-extension curve is described by the following equation (7):

$$\phi = \frac{f l_p}{k_B T} = \frac{l}{L} + \frac{1}{4(1-\frac{l}{L})} - \frac{1}{4} \qquad (1)$$

where Φ is the normalized force, f is the absolute force (typically piconewtons), l_p is the persistence length (typically 50 nanometers), k_B is the Boltzmann constant, T is the temperature, l is the extension of the molecules (the distance between the pits in this experiment), and L is the contour length between the two nanopits of the fluctuating polymer.

The relationship between the maximum standard deviation and the normalized force as described by Baba *et al.* (8) follows as:

$$\hat{\sigma}^2_{max} = \frac{\sigma^2_{max}}{L l_p} = \frac{\phi^{-1}}{4(1+\frac{3\phi^{-1}}{2})} \qquad (2)$$

where σ^2_{max} is the maximum variance of an individual DNA molecule and $\hat{\sigma}^2_{max}$ is the normalized variance. It should be noted that in the limit of two dimensions, the magnitude of transverse fluctuations double because they are now limited to a single spatial dimension, as discussed by Baba *et al.* The statistical behaviour of DNA in the nanopit-nanoslit system was derived by Reisner *et al* (9).

There has been speculation that the Marko-Siggia force law is modified by confinement. Binder (11) suggests that it breaks down in two-dimensional limit, while Chen *et al.* (12) claim it is modified by an increased effective persistence due to confinement, and Lin *et al.* (13) suggest that confinement acts as a bandpass filter and forbids certain fluctuation modes. We hope that the analysis of fluctuations in the wide parameter space provided by the nanopit-nanoslit system will shed light on potential modifications to the well-known force law under confinement.

Apparatus and Procedure

The experiments were performed in glass nanofluidic chips, fabricated using standard clean-room processes as described by Resiner *et al* (9). The chips contained nanofluidic slits with heights on the order of 100 nanometers. The slits were embedded with a lattice of square pits. The size, spacing, and depth of the pits as well as the height of the nanofluidic slit serve as independent variables to dictate the behaviour of the DNA (Figure 1).

Microfluidic channels in the chip served as an interface to the nanoscale system. The chip was fastened in a plastic chuck which was mounted on a Nikon Eclipse Ti-U inverted microscope. A lamp and a dichroic filter was used to stimulate the fluorescent dye molecules. DNA from a lambda phage virus (48,500 base-pairs or 16 microns in length) was stained with YOYO-1 fluorescent dye. The DNA was dissolved in 50 mM Tris buffer and pipetted into the microfluidic reservoirs of the chip. Pressure was applied to flow the DNA into the nanofluidic slits. Under equilibrium, the molecules self-assembled into the pits. We focused on geometries where the molecules occupied two pits, recording movies several minutes in length, observing the molecule fluctuating throughout its environment. After being recorded, the molecules were flown out of the field of view and new molecules were brought in.

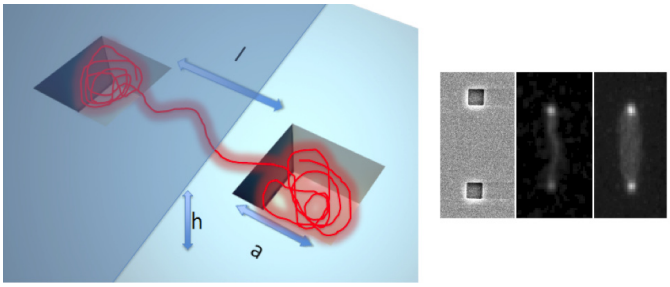


Fig. 1
Overview of the experiment.
Left: A schematic of two nanopits in a nanofluidic slit. The pits have width ***a*** , are separated by a distance ***l*** , and the roof of the system has height ***h*** above the floor. The pits have the same depth as the height of the slit.
Right: A scanning electron micrograph of two pits, an optical fluorescence micrograph of a molecule spanning two pits, and a projection of a movie of that molecule, demonstrating the variation in position. The distance between the pits is 5 microns.

Analysis

A movie of each studied molecule was opened as a three dimensional (two space and one time) matrix in MATLAB. At each frame, the brightest pixel of each row was calculated. Because the DNA is uniformly stained, maxima in recorded intensity serve as proxies for maxima in DNA concentration. For each row, an array of maximal positions over time was generated. The standard deviation of this array served as a measurement for the standard deviation of the position of the molecule at that point along its contour.

A custom built MATLAB program was used to determine the extended contour length (*L*). The first step involved subtracting the background noise in the fluorescence video image. A box was manually aligned to each of the two nanopits that contained the trapped DNA molecule. The summed pixel intensity of the two boxes was determined, as well as the pixel intensity of the entire molecule. The ratio of the intensity in the two pits to the total intensity of the molecule was used as a proxy to calculate the extended contour length between the two pits because the total contour length (L_T) of the molecule is known (19135 nm).

Results

It can be seen from the right half of Figure 2 that the standard deviation along the molecule increases as the position becomes further from the pits, peaking at the mid-point. Baba *et al*. (2012) derived this relationship as parabolic. In order to establish commonality between their optical traps and our entropic traps, we examined this relationship. In Figure 3, the variance is plotted as a function of the longitudinal coordinate and it can be seen that the parabolic relationship described by Baba *et al* (8) does indeed hold. The variance peaks at the middle of the stretched part of the molecule and behaves according to the quadratic law about this position. In our experiments, the DNA molecule was trapped at both ends, so the transverse fluctuation is symmetric about the midpoint. Once again, this result was also obtained by Baba *et al* (8) by using dual trap optical tweezers to trap the DNA molecule.

It was previously mentioned (3) that DNA behaves differently under confinement. We also postulated that the Marko-Siggia force-extension relationship (7) is not ideal to describe a DNA molecule under confinement. We measured the magnitude of transverse fluctuations of DNA in 500 nanometer pits separated by 1000 nanometers across slit heights ranging from 50 to 170 nanometers (Figure 4). We used the statistical theory of Reisner *et al* (9) to calculate the equilibrium conformation of the molecule, and coupled this result to the force-fluctuation relationship of Baba *et al* (8). to generate a theoretical prediction. As a comparison, we examined the two-dimensional limit of this relationship and the effective persistence described by Chen *et al* (12). In Fig.4, the Marko and Siggia 3D force-fluctuation curve is shown in red and the curve in blue is derived from the two dimensional limit of the force-fluctuation model. It can be seen in Fig.4, that the data points are bounded by the 3D and 2D force fluctuation curves and neither model adequately describes the behavior of the data across all regimes. The fact that the 3D force law does not ad

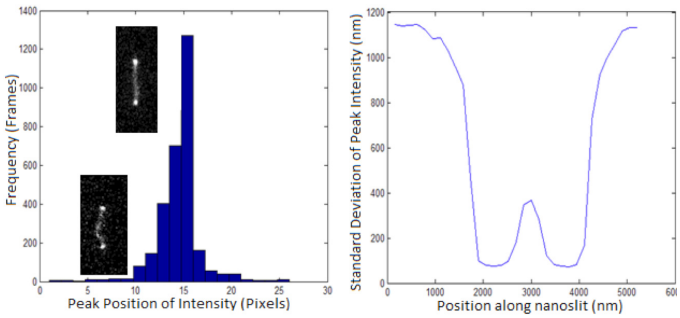


Fig. 2
Analysis to determine standard deviation. Left: A histogram of peak positions in a 3000 frame movie. The micrograph in the center of the histogram corresponds to a straight molecular configuration, while the other micrograph corresponds to a molecule fluctuation to the left. Right: The standard deviation of peak intensities, as a function of position along the nanoslit. The plateaus at the edges correspond to random noise, the two valleys correspond to the molecular contour trapped in the pits, and the peak between them represents the actual transverse fluctuations of the molecule.

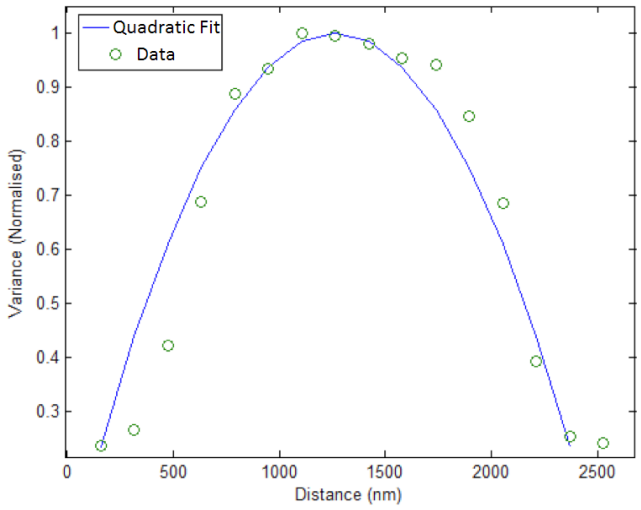


Fig. 3
Variance along molecular position for a DNA molecule. The molecule is trapped in a nanopit-nanoslit sytem with a slit height of 110 nm and with pits separated by 2000 nm. We find agreement with Baba et al.'s [6] assertion that variance behaves parabolically with distance from the end points.

equately describe the data points in the limit of strong confinement suggests that the bulk description of the Marko-Siggia force law may indeed break down under confinement. The modification of Chen *et al* (12) did not significantly affect the theoretical prediction.

Discussions

We sought to replicate the findings of Baba *et al* (8) in our solid-state nanofluidic system. We observed a parabolic relationship in transverse fluctuations with respect to the position along the molecule, which is an identical result to Baba *et al* (8). However, our observation of the force-extension relationship under confinement (Figure 4) suggests that we may have observed a breakdown of the accepted theoretical model. The behavior of the data in Figure 4 is best described as an interpolation of the 3D and 2D force fluctuation models. While these results suggest new confined polymer physics, it is regrettable that we were unable to fully explore the force-fluctuation and force-extension relationships due to difficulties in measuring the extended contour length of the molecule (the total length of the DNA stretched between the pits).

Initially, this study also expected to explore the relationship between the variance and the inverse of the normalized force, as was done by Baba *et al* (8) in the nanopit-nanoslit system. A challenge that presents itself in nanopit experiments that is not present in optical tweezer experiments is the fluctuating length of the stretched polymer, as contour moves in and out of each pit. This required a method to determine the extended contour length (L) of the DNA molecule trapped between nanopits.

In Section IV, the method used to determine the extended contour length *L* was described. This method was found to be imprecise and unstable as it was very sensitive to the positioning of the rectangular box. Even a small amount of residual background noise made a significant difference in the final computation. Overall we were unable to measure the length and instead, we had to use the statistical theory (9) as a proxy for an actual measurement of the length. We desire an objective and automatic way of calculating the extended contour length, in order to better compare our measurements to other work.

Our initial results suggest that this is potentially a fruitful avenue of polymer physics research, but to fully explore it we must develop a better image analysis method that is not as sensitive to background noise as the current approach.

Conclusions

Transverse fluctuations of single DNA molecules trapped in the nanopit-nanoslit system were studied in this paper. The findings in this paper indicate that the Marko-Siggia force-extension and force-fluctuation relationships are not the best model to describe DNA molecules under strong confinement (small slit heights). We have also demonstrated that this nanofluidic system can serve as a powerful alternative to optical trapping systems for advanced polymer physics research. Future work will further explore these relationships using improved image analysis algorithms.

Acknowledgements

Haider Riaz wishes to thank Marcel Georgin for his advice on how to construct a program for the data analysis. Alexander Klotz wishes to thank Hendrick de Haan and David Sean for their fruitful discussions on the theoretical aspects of confined polymer physics.

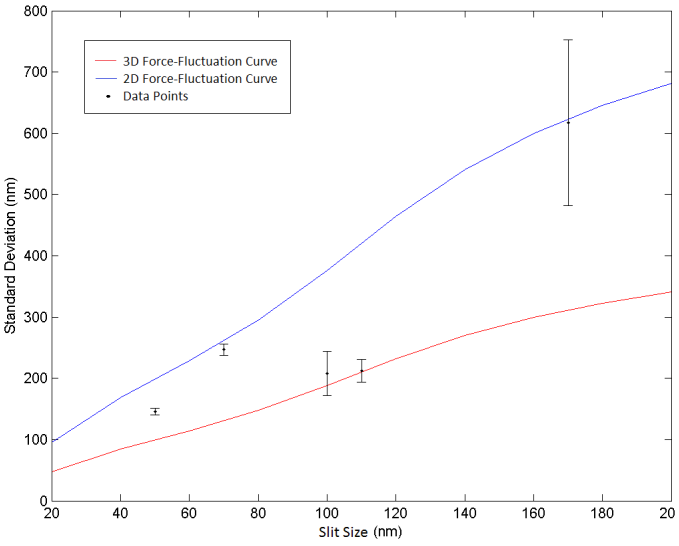


Fig. 4
Peak standard deviations of DNA in 500 nanometer pits separated by 1000 nm are measured. Theoretical curves from a combination of the statistical nanopit model and the force-fluctuation relationship are shown.

References

[1] Lam, Ernest T., Alex Hastie, Chin Lin, Dean Ehrlich, Somes K. Das, Michael D. Austin, Paru Deshpande *et al*. *"Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly*". Nature Biotechnology(2012).
[2] Welch, Robert L., Robert Sladek, Ken Dewar, and Walter W. Reisner. *"Denaturation mapping of Saccharomyces cerevisiae."* Lab on a Chip(2012).
[3] Reisner, Walter, Jonas N. Pedersen, and Robert H. Austin. *"DNA confinement in nanochannels: physics and biological applications."* Reports on Progress in Physics **75**, no.10 (2012):106601.
[4] Lee, K. S.; Smith, T. J.; Dickey, K. C.; Yoo, J. E.; Stevenson, K. J.; Loo, Y. L. *Adv. Funct. Mater.* **2006**, 16, 2409– 2414
[5] Rowe, B. W.; Freeman, B. D.; Paul, D. R. *Polymer* **2009**, 50, 5565– 5575
[6] De Gennes, Pierre-Gilles. *Scaling concepts in polymer physics*. Cornell university press, 1979.
[7] Marko, John F., and Eric D. Siggia. *"Stretching DNA."* Macromolecules **28**, no.26 (1995):8759-8770.
[8] Baba, Takeshi, Takahiro Sakaue, and Yoshihiro Murayama. *"Force–Fluctuation Relation of a Single DNA Molecule."* Macromolecules **45**, no.6 (2012):2857-2862.
[9] Reisner, Walter, Niels B. Larsen, Henrik Flyvbjerg, Jonas O. Tegenfeldt, and Anders Kristensen. *"Directed self-organization of single DNA molecules in a nanoslit via embedded nanopit arrays."* Proceedings of the National Academy of Sciences **106**, no.1 (2009):79-84.
[10] Klotz, Alexander R., Hugo B. Brandao, and Walter W. Reisner. *"Diffusion Resonance of Nanoconfined Polymers."* Macromolecules **45**, no.4 (2012):2122-2127.
[11] Hsu, Hsiao-Ping, Wolfgang Paul, and Kurt Binder. *"Breakdown of the Kratky-Porod wormlike chain model for semiflexible polymers in two dimensions."* EPL(Europhysics Letters) **95**, no.6 (2011):68004.
[12] Chen, Yeng-Long, Po-Keng Lin, and Chia-Fu Chou. *"Genearlized force-extension relation for DNA confined in sub-100nm nanoslits."* Bulletin of the American Physical Society **56** (2011).
[13] Lin, Jun, Fredrik Persson, Joachim Fritzsche, Jonas O. Tegenfeldt, and Omar A. Saleh. *"Bandpass Filtering of DNA Elastic Modes Using Confinement and Tension."* Biophysical Journal **102**, no.1 (2012):96-100.

RESEARCH ARTICLE

A study of the effect of auditory prime type on emotional facial expression recognition

Sameer Sethi[1]*, Dr. Marc D. Pell[2]

[1] Faculty of Science, McGill University, Montreal, QC
[2] School of Communication Sciences and Disorders, McGill University, Montreal, QC

*Email Correspondence:
sameer.sethi@mail.mcgill.ca

Keywords:
Vocalization: *auditory stimulus consisting of a non-verbal sound to convey an emotional meaning.*
Pseudoutterance: *auditory stimulus consisting of a verbal, speech-like (but nonsensical), sentence to convey an emotional meaning.*
Auditory Prime: *auditory stimulus that influences the interpretation of the subsequent stimuli (visual in this case).*
N400: *Negative amplitude observed in EEG data 400 msec after the onset of a stimulus; modulated by the congruency or incongruency of the prime-target pair.*

Abstract

Background: In this study, we investigated the influence of two types of emotional auditory primes - vocalizations and pseudoutterances - on the ability to judge a subsequently presented emotional facial expression in an event-related potential (ERP) study using the facial-affect decision task. We hypothesized that accuracy would be greater for congruent trials than for incongruent trials. This is due to the possibility that a congruent prime would allow the listener to implicitly identify the particular emotion of the face more effectively. We also hypothesized that the normal priming effect would be observed in the N400 for both prime types, i.e. a greater negativity for incongruent trials than for congruent trials.

Methods: Emotional primes (vocalization or pseudoutterance) were presented to participants who were then asked to make a judgment regarding whether or not a facial expression conveyed an emotion. Behavioural data on participant accuracy and experimental electroencephalogram (EEG) data were collected and subsequently analyzed for six participants.

Results: Behavioural results showed that participants were more accurate in judging faces when primed with vocalizations than pseudoutterances. ERP results revealed that a normal priming effect was observed for vocalizations in the 150 msec - 250 msec temporal window – where greater negativities were produced during incongruent trials than during congruent trials – whereas the reverse effect was observed for pseudoutterances. Few participants were tested (n = 7). Hence, this study is a pilot study preceding a further study conducted with a greater sample size (n = 25) and slight modifications in the methodology (such as the duration of auditory primes.)

Conclusions: Vocalizations showed the expected priming effect of greater negativities for incongruent trials than for congruent trials, while pseudoutterances unexpectedly showed the opposite effect. These results suggest that vocalizations may provide more prosodic information in a shorter time and thereby generate the expected congruency effect.

Introduction

On a daily basis, we are faced with the task of interpreting emotional meaning in voices by their prosody, the way they are said. A simple question such as "are you kidding?" can be asked with anger, sadness, confusion, or frustration. It can be interpreted wholly differently depending on its prosodic nature. Judging the prosody (and subsequently the meaning) of the question is up to the listener. Therefore, a popular topic of research in recent years has been to develop an understanding of the neural processes and timing involved emotional prosody integration. The investigation of how context influences semantic integration has been studied frequently through the presentation of an auditory emotional prime prior to presenting a visual target.

Priming Effects of Auditory Stimuli

Many studies have shown that the effects of prosodic cues of a voice are not limited to the interpretation of message's meaning, but also include priming effects in interpreting emotion conveyed by facial expressions (1). Information regarding the emotion of a voice is processed and integrated soon after hearing it (2), which helps explain the display of subsequent priming effects when judging a facial expression's emotions. The N400 – a characteristic negative amplitude at 400 msec post-stimulus – is particularly associated with semantic integration of emotion. Studies involving the presentation of a word after an auditory prime has established semantic context have shown that a greater negativity is generated in the N400 when the prime is emotionally unrelated to the visual stimulus (3). Paulmann & Pell

(2010) investigated the effects of the length of an emotional auditory prime on the ability of the listener to judge emotional facial expressions when they were congruent and incongruent with respect to the auditory primes. They tested both short (200 msec) and medium (400 msec) length primes in order to observe the relative amount of prosodic information required to observe such a priming effect. They found that the medium length prime generated the expected ERP pattern: a larger negativity in the N400 for incongruent trials compared to congruent trials. However, they found the reverse relationship when looking at the short length prime. This suggests that a short auditory prime may not provide the same "processing advantage" in establishing an emotional context for a facial expression. As such, long primes (longer than 400 msec) were used in the present study.

Prime Types

A factor that has not yet been studied in depth is the nature of the auditory prime used. Auditory stimuli can take the form of regular sentences with a particular prosody. However, this poses the risk of semantic context playing a role in the priming effect (in addition to the intended conveying of emotion). For this reason, this study made use of primes that did not pose the risk of establishing specific context to the listener.

The aim of the present study was to investigate the priming effect of vocalizations versus pseudoutterances in judging facial expressions by comparing the ERP pattern in the N400. *Vocalizations* refer to non-speech- like sounds that convey emotions, such as the sound "Ahhh!" to convey anger. *Pseudoutterances* resemble speech, but do not have a comprehensible meaning (an example of a pseudoutterance is "they nestered the flugs"). They are used in order to convey emotional meaning without conveying any semantic meaning. We presented both types of primes for their full duration rather than for a universal duration because the latency of the peak of emotional meaning varied from one prime to another. We created an implicit task requiring participants to answer a yes/no question regarding whether or not the presented face conveyed any of three pre-determined emotions – happiness, sadness, or anger - rather than requiring participants to determine the identity of the emotion. With respect to the behavioural data, it was hypothesized that accuracy would be greater for congruent trials than for incongruent trials due to the possibility that a congruent prime would allow the listener to implicitly identify the particular emotion of the face more effectively. With respect to the ERP data, we hypothesized that the normal priming effect would be observed in the N400 for both prime types - that is to say, a greater negativity for incongruent trials than for congruent trials. The justification for this is that both prime types convey emotional meaning without setting up any semantic context, despite the fact that they are intrinsically different in the way they convey emotion.

Methods

Participants

One left-handed and six right-handed English-speakers (6 female; mean age = 21.5 years, SD = 2.2 years) participated in the study. None of the participants reported any hearing impairments. All participants had normal or corrected-to-normal vision. Participants gave informed consent before participating in the study. Participants were compensated $10 per hour for their participation. The right-handed male participant's data was not used for the analysis, as his accuracy was less than 50%.

Stimuli

We used two types of emotional auditory cues (pseudoutterances and vocalizations) as primes and pictures of emotional or non-emotional facial expressions as targets. The primes and targets were paired for cross-modal presentation according to the 'facial affect decision task' (FADT), which involves the pairing of facial expressions with emotionally related or un-related vocal utterances. Both the auditory and visual stimuli were taken from established inventories (4-6).

Prime Stimuli. We used pseudoutterances and vocalizations of three emotion categories: anger, sadness, and happiness. The use of pseudoutterances allows for an emotion to be conveyed without being coupled with semantic information (7, 8). The pseudoutterances were presented by 10 speakers (6 female, 4 male), and were selected based on a validation study conducted by Pell *et al.* (5). Average recognition accuracy for the pseudoutterances were as follows: anger = 93%, sadness= 91%, happiness = 89%. The mean duration of the pseudoutterances was 1966.01 msec (s = 404.9 msec). The vocalizations were presented by 10 speakers (5 female, 5 male), and were constructed by Belin *et al.* (6). Average recognition accuracy for the vocalizations were as follows: anger = 75%, sadness = 77%, happiness = 81%. The mean duration of the vocalizations was 1533.13 msec (s = 799.70 msec). For both pseudoutterances and vocalizations, we presented ten items (one per speaker) for each of the three emotion categories (a total of 30 per participant). Both the pseudoutterances and vocalizations were presented in their entirety to preserve prosodic meaning.

Target Stimuli. The target stimuli consisted of 13.5 cm x 19 cm images with static facial expressions. The images were restricted to the actors' faces and were presented in colour. As with the prime stimuli, there were three distinct emotion types: anger, sadness, and happiness. Facial grimaces representing a non-emotion (a grimace) were presented in equal proportion to each emotion. The stimuli presented consisted of 10 different actors (6 female, 4 male) of various ethnicities (Caucasian, Asian, and African). We required participants to judge whether or not the target facial stimulus presented during each trial represented an emotion. Because the task was implicit, the

participant was not required to identify the specific emotion, but to simply answer "yes" or "no". Yes trials were those presented with faces in either of the three emotion categories (n=30 for each anger, sadness, happiness) and no trials were those that included the aforementioned facial grimaces (n=90). The target stimuli used in the study were retrieved from a previously established inventory (4). The accuracy of recognition for faces conveying a discrete emotion was seen to be the following: anger = 92%, sadness = 88%, happiness = 99% (4). The recognition rate for grimaces (non-emotion) was seen to be 60% in the same validation study. Fig. 1 shows an instance of each emotion as well as a facial grimace posed by one of the actors used in the study.

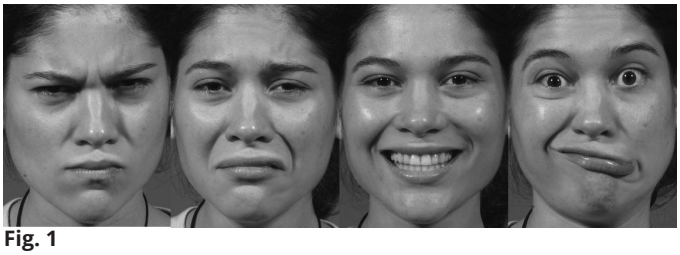


Fig. 1
Sample Facial Expression for Emotions and Grimace

Design

We required participants to make a yes/no judgment regarding whether or not the target stimulus in each trial represented a discrete emotion. The actual interpretation of the emotion in each trial was implicit since the participant had only to confirm or deny the presence of a discrete emotion in the face. The trial distribution was as follows: for each prime type (vocalizations and pseudoutterances), there were 30 trials for each combination of prime emotion (anger, sadness, happiness) and facial emotion (anger, sadness, happiness), and 90 additional trials for each combination of prime emotion and facial grimaces – a total of 1080 trials.

Procedure

We set up a 64-electrode EEG cap on each participant and seated him or her in a dimly lit room approximately 65 cm from the computer monitor. The visual stimuli on the monitor delivered the auditory stimuli through earphones at a volume set by the participant according to his or her comfort level. Participants were instructed to listen to the auditory stimulus, and carefully observe the face presented in order to subsequently make a yes/no judgment regarding whether or not it conveyed an emotion. The structure of each trial was as follows. First, we presented an auditory prime stimulus (vocalization or pseudoutterance) with a fixation cross present in the middle of the monitor screen. At the offset of the prime, we presented the visual target stimulus (emotional or grimace face) until the participant pressed *yes* or *no* on the response. After the participant's response, the instruction "Blink" was presented for 1000 msec to encourage the participant to blink at that time as opposed to during the following trial. Finally, we displayed a 1500 msec interstimulus interval with a fixation cross in the middle of the screen. The experiment was divided into groups of trials (180 trials each) – referred to as "blocks" – in order to give the participants consistent breaks during the experiment. Prior to the first experimental block, each participant completed two practice blocks consisting of 10 trials each. During the first practice block, we presented a face with an emotion or grimace and the participant was asked to make a yes/no judgment regarding whether or not the face conveyed an emotion. The purpose of this block was to help the participants distinguish faces that conveyed emotional meaning and those that did not prior to the experiment. During the second practice block, an auditory prime of either type preceded the presentation of the face, after which the participant was required to make the same yes/no judgment. There were two optional breaks within each of the six blocks (after every 60 trials), and an optional break between blocks. Each block lasted approximately 20 minutes and the entire experiment lasted approximately 3-3.5 hours (including set-up and removal of the EEG cap.)

ERP Recording and Data Analysis

We used sixty-four active Ag/AgCl electrodes (actiCAP) mounted in an elastic cap to record the EEG. The EEG was digitized at 1024 Hz during recording. We then digitized the data sets offline to a sampling rate of 250 Hz, re-referenced them to the average of the electrodes, and applied a band pass filter between 0.1 Hz and 40 Hz. Bipolar horizontal and vertical EOGs were recorded in order to help with trial rejection. We then ran an independent component analysis (ICA) and rejected trials with EOG activity greater than 50 μV. Overall, approximately 30-40% of the trials for each participant were rejected using this method. We discarded all *no* trials (i.e. trials with grimaces as the target stimulus) from this analysis, as well as incorrect *yes* trials (i.e. trials with an angry, sad, or happy target emotion where the participant incorrectly responded *no*). We averaged the remaining trials from the onset of the facial stimulus to 1000 msec after its onset.

Using ANOVA, we separately analyzed accuracy (i.e. correct yes trials) and ERP data for each condition (angry congruent, angry incongruent, etc.). We also used seven Regions of Interest (ROIs) as group factors for the analysis. The regions were divided as follows: left frontal electrodes (F3, F5, F7, FC3, FC5, FT7); right frontal electrodes (F4, F6, F8, FC4, FC6, FT8); left central electrodes (C3, C5, T7, CP3, CP5, TP7); right central electrodes (C4, C6, T8, CP4, CP6, TP8); left parietal electrodes (P3, P5, P7, PO3, PO7, O1); right parietal electrodes (P4, P6, P8, PO4, PO8, O2); and midline electrodes (FZ, Cz, CPz, Pz, POz, Oz). We used a temporal window of 440 msec – 540 msec for the mean amplitude analysis of the N400. After visual inspection, we hypothesized that the temporal window of 150 msec – 250 msec was also of interest.

Results

A two by two by three repeated measures ANOVA was carried out separately for the behavioural data (accuracy) and the ERP data. In both cases, the factors were prime type (vocalizations, pseudoutterances), prime-target congruency (congruent, incongruent), and emotion (anger, sadness, happiness), respectively.

Behavioural Results

Analysis of accuracy rates revealed a significant main effect of prime type [$F(1,5) = 19.22$, $p = 0.007$]. There were overall differences in the ability to judge whether a facial expression represented an emotion based on the prime type, with significantly greater mean accuracy for vocalizations ($\bar{x} = 88\%$, $s = 0.088\%$) than for pseudoutterances ($\bar{x} = 84\%$, $s = 0.10\%$). No significant main effects for prime-target congruency were observed. Moreover, no other significant effects with regards to emotion category or congruency status of trials were observed.

ERP Results

Temporal window of 440 msec – 540 msec. The analysis for the temporal window 440 msec – 540 msec revealed a significant main effect of prime type [$F(1,35) = 4.31$, $p = 0.045$]. Overall, we observed more negative ERP amplitudes for pseudoutterances than for vocalizations as well as for congruent trials than for incongruent trials. We noted another main effect of prime-target congruency [$F(1,35) = 9.92$, $p = 0.003$].

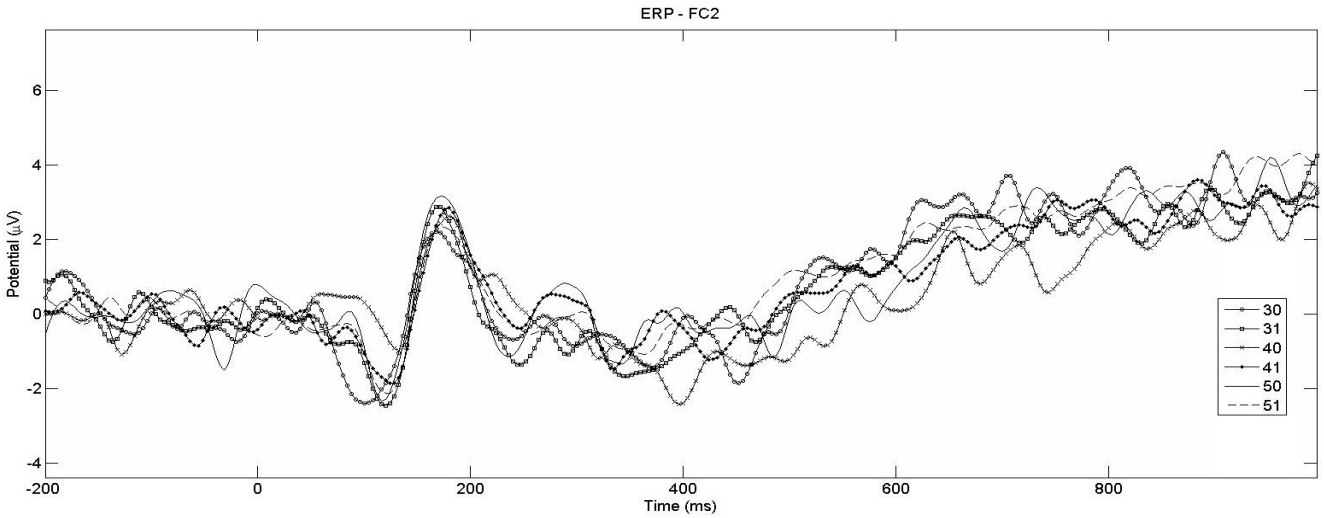


Fig. 2a Vocalizations – Averaged ERPs for FC2 electrode

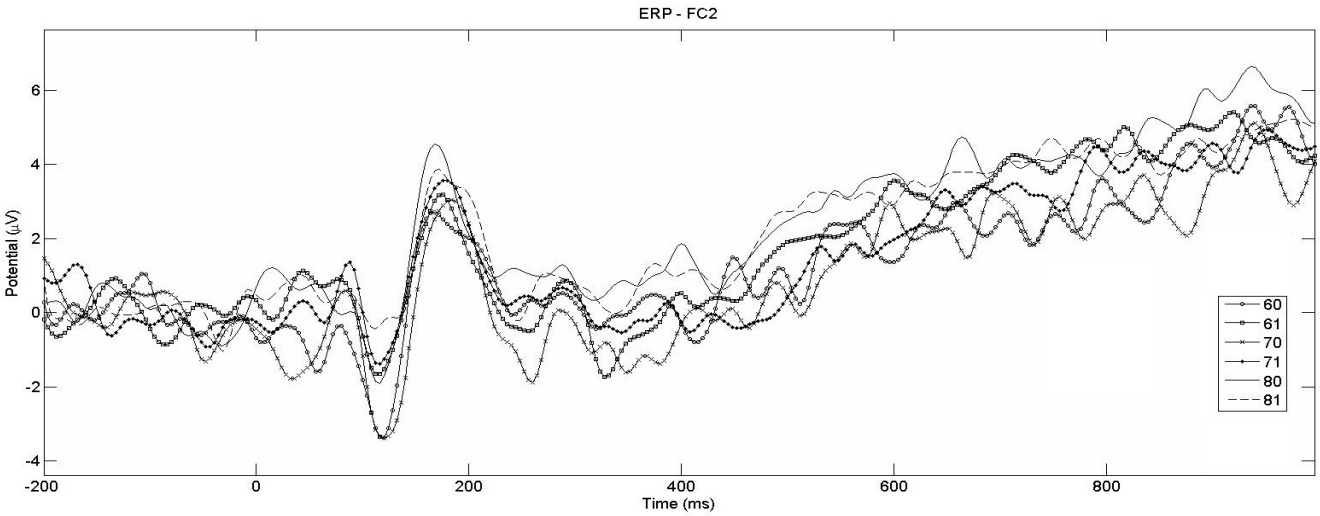


Fig. 2b Pseudoutterances – Averaged ERPs for FC2 electrode

Temporal window of 150 msec – 250 msec. The analysis for the temporal window 150 msec – 250 msec revealed a significant main effect of prime type [$F(1,35)$, $p = 0.027$]. Overall, we observed more negative ERP amplitudes for pseudoutterances than for vocalizations as well as for congruent trials than for incongruent trials. We also noted another main effect of prime-target congruency [$F(1,35) = 17.66$, $p < 0.001$]. Furthermore, a grouping factor – region of interest (ROI) – was introduced into the statistical analysis. We saw a significant interaction of prime type, prime-target congruency, and ROI [$F(6,35) = 2.60$, $p = 0.03$]. This indicated the possibility that ERP amplitudes were modulated as a function of both variables in combination in certain regions. Further analysis of specific ROIs revealed an interaction between prime type and prime-target congruency in left frontal electrodes [$F(1,5) = 15.56$, $p = 0.01$] and right frontal electrodes [$F(1,5) = 7.17$, $p = 0.04$]. In both regions, we observed significantly more negative ERP amplitudes in incongruent trials than in congruent trials for vocalizations. However, we observed significantly more negative ERP amplitudes in congruent trials than in incongruent trials for pseudoutterances. The effects in the left and right frontal regions are identical, so the FC2 electrode is used to illustrate them in Fig. 2.

Table 1 indicates the legend code to condition conversion for increased readability. A significant interaction between emotion and prime type was observed when ROI was used as a grouping factor [$F(12,70) = 1.98$, $p = 0.039$]. Further investigation showed that there was an interaction between emotion and prime type in left frontal electrodes [$F(2,10) = 4.30$, $p = 0.045$] and right frontal electrodes [$F(2,10) = 5.50$, $p = 0.024$]. In both regions, sadness was seen to elicit significantly more positive ERP amplitudes than anger and happiness for vocalizations, while it elicited significantly more negative ERP amplitudes than anger and happiness for pseudoutterances.

Code	Condition
30	Congruent Angry Vocalization
31	Incongruent Angry Vocalization
40	Congruent Sad Vocalization
41	Incongruent Sad Vocalization
50	Congruent Happy Vocalization
51	Incongruent Happy Vocalization
60	Congruent Angry Pseudoutterance
61	Incongruent Angry Pseudoutterance
70	Congruent Sad Pseudoutterance
71	Incongruent Sad Pseudoutterance
80	Congruent Happy Pseudoutterance
81	Incongruent Happy Pseudoutterance

Table 1
Code – Condition Conversion

Discussions

In the present study, we used analysis of ERP amplitudes to investigate the influence of auditory prime type on the judgment of emotional facial expressions. A previous study by Paulmann & Pell (9) investigated the effect of prime length and observed that medium length primes (400 msec or more) elicited a normal response in prime-target congruency in that incongruent trials elicited more negative ERP amplitudes than congruent trials. For this reason, this study made use of prime stimuli that were at least 400 msec in length. The main focus was to understand whether or not vocalizations and pseudoutterances convey emotional meaning to the same extent by comparing both the ERP amplitudes elicited by subsequent judgment of emotional facial expressions as well as behavioural results.

Implications of Behavioural Results

Behavioural results demonstrated that there was a significant difference in the accuracy of emotional facial expression judgment between the vocalizations and pseudoutterances as primes. As reported, vocalizations allowed for more accurate judgment overall – an average accuracy of 88% – than pseudoutterances – an average accuracy of 84%. It is speculated that a reason for this difference is the very nature of the facial expression stimuli. Fig. 1 provides an example of the typical facial stimuli used in the experiment; upon inspection, it becomes apparent that the facial expressions resemble expressions one might have in a real-life conversation after having conveyed an emotion in a fashion that closely resembles a vocalization. It is possible that the relation between the vocalization as a prime and the contextually similar facial expression allowed for more accurate responses by the participant.

Unconventional N400 Congruency Results

Studies have shown that prime-target incongruency results in more negative ERP amplitudes than congruent trials when primes of 400 msec are used (5, 10). As previously mentionedthe primes used in this study were at least 400 msec long, quite often reaching up to 2000 msec in length. We used long primes in hopes that the conventional N400 pattern of congruency would be observed. This seemed to be a reasonable deduction based on the fact that emotional speech tone has been seen to influence the early judgment of facial expressions (1, 2). However, just the opposite was observed for pseudoutterances; *congruent* trials elicited more negative ERP amplitudes than incongruent trials. This was unanticipated, and the reasons behind the phenomenon can only be speculated. It is well known that emotional information from an auditory prime is integrated early after hearing it (11). Perhaps excessive length of primes caused disintegration in the priming effect by the time the participant was required to make a judgment of the facial expression.

Another interesting effect observed was that pseudoutterances elicited greater negative ERP amplitudes than vocalizations. It is possible that vocalizations more effectively conveyed emotion, or similarly to the behavioural effect, provided emotional information that the participant found more relevant to the target faces. This would have allowed an association to be made more quickly between the prime and emotions expressed by the facial expressions when preceded by vocalizations. Absolute values for ERP amplitudes were not analyzed - however, it is likely that pseudoutterances elicited normal ERP amplitudes and that vocalizations simply elicited a relatively more positive response.

Negative ERP Amplitudes in the 150 msec – 250 msec Window

As with the 440 msec – 540 msec temporal window, the 150 msec-250 msec window showed greater negativity in trials with pseudoutterances as primes than for vocalizations. Moreover, the effect of congruency was also similar: there was greater negativity observed for congruent versus incongruent trials. This temporal window differed from the 440 msec – 540 msec window in that significant interactions were observed in the left and right frontal regions between type and congruency, and between emotion and type. The left frontal region is commonly associated with the processing of positive emotions while the right frontal region is commonly associated with the processing of negative emotions (12). This was not considered in this study; but based on the fact that the emotions of the stimuli were anger, sadness, and happiness, the observed ERPs on both sides of the brain are not surprising results. As for the particular interactions, the interaction between prime type and prime-target congruency showed greater negativity in incongruent trials than in congruent trials for vocalizations, but the opposite effect for pseudoutterances. As previously mentioned, the emotional meaning of an auditory stimulus is processed very early on. Differences in pattern between vocalizations and pseudoutterances – where vocalizations show the conventional greater negativity in incongruent trials – may be the result of the lengthier pseudoutterances favouring controlled processing. On the other hand, vocalizations tended to be less than 1000 msec in length, which may have favoured automatic processing and thus showed the expected pattern in ERP amplitudes. There was also interaction between the emotion and type of prime; greater negativity was seen for a sad prime in pseudoutterances than in vocalizations. This reiterates the idea that there may have been a weaker association between the pseudoutterance and facial expression simply due to the nature of the two prime types. Vocalizations may allow for a stronger association between the emotion of the prime and the facial expression seen.

Further Investigations

Future studies should seek to understand the ability of vocalizations to convey stronger emotional meaning than pseudoutterances as well as why full-duration pseudoutterances elicited greater negativities for congruent trials. Once these factors are understood, it would be interesting to conduct similar experiments on a greater range of emotions. This could be taken further by classifying them into positive and negative emotions in order to study further levels of interaction - for example whether positive vocalizations elicit greater or lower negativities than pseudoutterances.

Conclusions

These findings indicate that, from a behavioural standpoint, facial expressions preceded by vocalization primes are more accurately recognized than when preceded by pseudoutterance primes. This is likely due to the relations between facial expressions and vocalizations. As expected, vocalizations were shown to elicit a greater negativity for incongruent trials than for congruent trials. Interestingly, pseudoutterances caused a reversed effect. The conventional priming effect of pseudoutterances may have been absent in the averaged ERP data as a result of their long duration, as compared to previous studies. Future studies should consider capping the length of the pseudoutterances and vocalizations to see if the conventional N400 response is restored.

References

[1] G. Pourtois *et al. NeuroReport.* **11**, 1329-1333 (2000).
[2] B. De Gelder, J. Vroomen, *Cognition & Emotion.* **14**, 289-311 (2000).
[3] C. Brown, P. Hagoort, *Journal of Cognitive Neuroscience.* **5**, 34-44 (1993).
[4] M.D. Pell, *Brain & Cognition.* **48**, 499-504 (2002).
[5] M.D. Pell et al. *Journal of Phonetics.* **37**, 417-435 (2009).
[6] P. Belin, S. Fillion-Bilodeau, *Behavior Research Methods.* **40**, 531-539 (2008).
[7] D. Grandjean *et al. Nature Neuroscience.* **8**, 145-146 (2005).
[8] S. Paulmann, S.A. Kotz, *Brain & Language.* **105**, 59-69 (2008b).
[9] S. Paulmann, M.D. Pell, *Cognitive, Affective, & Behavioral Neuroscience.* **10**, 230-242 (2010).
[10] Q. Zhang, A. Lawson, C. Guo, Y. Jiang, *Brain Research Bulletin.* **71**, 316-323 (2006).
[11] S. Paulmann, S.A. Kotz, *NeuroReport.* **19**, 209-213 (2008a).
[12] R. Graham, R. Cabeza, *Cognitive Neuroscience and Neuropsychology.* **12**, 1-4 (2000).

RESEARCH ARTICLE

Proof-carrying authorization in distributed systems with Beluga: a case study

Leah Weiner[1]*

[1] School of Computer Science, McGill University, Montreal, QC

*Email Correspondence:
leah.d.weiner@gmail.com

Abstract

The rising popularity of distributed systems creates a need for a secure method of message passing. One approach to access control is accomplished through proof-carrying and proof authorization. The requesting party must provide a proof of authorization of access while the serving party must verify the validity of the proof. $PCML_5$ is a programming language which enables a programmer to encode proofs and perform proof-checking procedures more easily.

The purpose of this project is to encode $PCML_5$ with Beluga, a new functional programming language which supports dependent types and which has built in some of the more common procedures. Such an implementation will provide formal guarantees regarding the language of $PCML_5$ itself. Moreover, this case study will aid in the understanding of the development and implementation of software specically for programming with proofs. It will provide insight into the tools needed to allow any programmer to easily specify and verify complex behavioral properties of programs.

Introduction

In order to understand the research laid out in this report, one must have a solid understanding of functional programming and of security in distributed systems. This report begins with a brief description of lambda calculus, a formal calculus of functions and the basis for functional programming. Following the discussion of lambda calculus and functional programming is a section devoted to introducing Beluga, the functional programming language in which the code for this project was written.

Next is an introduction to distributed systems and security within these systems. $PCML_5$, a prototype programming language which this project implements, is introduced together with the motivations behind implementing $PCML_5$ in Beluga.

Finally, there is a detailed description of the actual steps taken towards the implementation of $PCML_5$ in Beluga.

Lambda Calculus

The lambda calculus is a formal system intended to represent and to calculate with functions. These functions can take other functions as input and/or return functions as output. Such functions are called *higher-order functions*. Lambda terms are the foundation of this system and are defined as follows:

Variables: A variable is a lambda term

Lambda Abstraction: If M is a lambda term and x is a variable, then $\lambda x.M$ is a lambda term

Application: If M and N are lambda terms, then $M\ N$ is a lambda term

The lambda term $\lambda x.M$ represents the function $f(x) = M$, where M may or may not be dependent on the variable x. The identity function for example is represented as $\lambda x.x$. Observe that variable names are unimportant. $\lambda x.x$ and $\lambda y.y$ are equivalent functions. This equivalence is known in lambda calculus as *alpha-equivalence*.

Binding

Consider the lambda term $\lambda x.M$. If x occurs in the expression M, then x is said to be *bound to M*. A variable that is not bound by a lambda abstraction is said to be *free*. The free variables of a lambda term can be deduced by the following inductive rules:

Let x be a variable and let M and N be two lambda terms. Then

$FV(x) = \{x\}$

$FV(\lambda x.M) = FV(M) \setminus \{x\}$

$FV(M\ N) = FV(M) \cup FV(N)$

where $FV(M)$ denotes the free variables of the lambda term M.

Substitution

In lambda calculus, substitution is used to mimic function application. Consider the two lambda terms $\lambda x.M$ and N, and application is represented on paper as $M\ N$. During this application, every free occurrence of the variable called "x" (the variable that is bound to M) which occurs in the body of M, is being substituted by the lambda term N. We write $[N/x]M$ to represent this substitution.

Capture-avoiding substitution

When performing substitutions as described above, certain precautions must be taken. In order to properly perform the application $M\ N$, only free variables should be substituted (bound variables should not be touched). However, in cases when variable names are not distinct, performing substitution may result in the binding of a variable which was previously free. An example of this phenomenon, called name capture, is described below. One way to avoid name capture during substitution is to rename the free variables (the variables affected by substitution) in the relevant terms uniquely. This safe substitution is known as *capture-avoiding substitution*.

To see the importance of renaming, consider the following example: In the expression $\lambda x.xz$, z is the only free variable. However, to perform the substitution $[x/z]\ \lambda x.xz$, the result, before renaming would be $\lambda x.xx$. In this way, z has been "captured", its distinction from the variable x has been lost.

The correct way to perform this substitution is to first rename the free variables. Through alpha-equivalence, $\lambda x.xz == \lambda y.yz$. Now, $[x/z]\ \lambda y.yz = \lambda y.yx$, which is the intended result (7).

Functional Programming

Functional programming is a programming paradigm based on the ideas of lambda calculus. Functional programs use function evaluation as the basis for computation. The better known imperative programming languages, such as Java or C, rely on mutating data and changing the state of the machine rather than function evaluation.

One consequence of functional programming's reliance on function evaluation is its extensive use of recursion. The more familiar imperative programs rely more heavily on iteration. Writing code in a recursive fashion generally takes less space than iterative code, allowing functional programs to have code which is cleaner and easier to read.

Functional programs also avoid side effects. That is, observable changes such as writing data to disk or changing variable names are bypassed during computation. Conversely, imperative programs do use side effects (also referred to as referential opaqueness), so the output of imperative code can change depending on the current state of the system. Functional programs, which avoid this dependence on the current state of the machine to determine the output, do not output results dependent on the state of the machine and are therefore more predictable and thus easier to reason about.

Lastly, functional programming has the property that each expression has a single value (referential transparency) (10). The use of referential transparency ensures that each expression has a unique value, eliminating confusion and giving the programmer a greater sense of control.

Recall that substitution of lambda terms must be capture-avoiding. Issues such as name capture also arise when performing function application in functional programs. Functional programs then must be able to generate fresh variable names, rename variables and perform capture-avoiding substitution appropriately. Although these procedures are extremely common in functional programming, most functional programming languages require these procedures to be manually implemented.

Beluga

Beluga is a newly developed pure functional programming language which provides a novel programming and proof environment. The motivation behind the development of Beluga was the desire to provide an environment in which to set up formal systems and proofs and in which the aforementioned lack of support for the common operations in functional programming languages (capture-avoiding substitution, renaming, fresh name generation) is remedied.(4).

While Beluga successfully achieves automatic support for these operations, it does not include support for input and output operations, references or exceptions.

Types are the foundational tools used to build programs. Built-in types common to many programming languages include integer, character or double. Beluga does not have any predefined data types but it does support dependent types, or types which depend on other variables.

Because Beluga does not have a predefined type system, a Beluga programmer must declare his own types. For example, the integer type

in Beluga can be initiated with the following piece of code:

```
nat : type.
zero : nat.
succ : nat -> nat.
```

The first line declares that natural numbers (called nat) are a type and the second line declares that zero is a natural number. The third line demonstrates how one can achieve natural numbers other than zero. This declared successor type, succ, is a dependent type. It's value is dependent on the value of the inputted nat. In the case of successor, succ N = N + 1 for any natural number N.

To get a better understanding of what code looks like in Beluga, we provide an example of the simple addition function below:

```
rec add: [. nat] -> [. nat] -> [. nat] =
fn x => fn y =>
   case x of
   | [. zero] => y
   | [. succ N] =>
       let [. R] = add [. N] y in [. succ R]
  ;
```

In line one, the keyword rec announces the start of a new (recursive) function. The name of this function is "add" and it takes in two variables of type nat and returns a variable of type nat.

Line two gives the input nat values arbitrarily chosen variable names. In this example, the first inputted natural number is called "x" and the second is called "y".

In the third line of this function, the programmer declares that there will be a case by case analysis on the value of "x", the first input to the add function. As x is a natural number, it is either equal to zero, or it is of the form succ N, where N is a natural number. It is important to notice that a natural number of the form succ N cannot equal zero.

The next line deals with the case when x equals zero. In this case, the function is told to return y. This makes sense, as 0 + y = y. This case acts as the base case in this recursive function.

The fifth line deals with the case when x is of the form succ N. It declares R to be a new variable equal to N + y. The successor of N + y is then returned as the final output to the function (after more recursive calls). To check the validity of this, notice that succ (N + y) = N + y + 1 and that x = succ N = N + 1. Thus, x + y = succ N + y = N + 1 + y = N + 1 + y, as desired.

Distributed systems

A distributed system is a set of computers which communicate through a network. Although each machine is independent, they collectively appear to an outsider (user) as one coherent system. The use of a distributed system is motivated by a system's reliability, performance and transparency.

Having a large number of machines working together clearly increases the reliability of the system as a whole. If one part fails, another is available to take over, resulting in a higher tolerance for faults. Similarly, the cooperation of a series of machines leads to the better performance of distributed systems when compared to a single centralized machine. Moreover, it is easier to share data and communicate when using a distributed system (8).

There are some disadvantages however. Writing software for a distributed system is more challenging. This difficulty may stem from the fact that each component of the distributed system may be written in a different language, that consistency across the system is hard to achieve, or that testing software in a distributed system is difficult. Also, despite a distributed system's higher fault tolerance, the greater amount of components means that more parts are prone to failure (but, as previously stated, the failure of a single piece of the system is less urgent). Additionally, data security is sometimes an issue (11).

Authentication and authorization

Certain information being passed along a network may be authorized only for the eyes of certain machines. It is therefore necessary for a distributed system to have some sort of access control. In order to grant a machine access to some set of data, it is necessary to ensure two things. Firstly, *authentication*, or certification of the identity of the requester, must be determined. Secondly, *authorization*, or the decision of allowing or disallowing the requester access to the desired information, must be figured. The set of rules by which this decision is made is called the *policy* (6).

There are two methods for access control. One possibility is having a central authority. Each request must first go through this central authority, which either approves or denies the client's request. Upon approval, the central authority tells the server to send the desired information directly to the requesting machine.

Alternatively, we can use the proof-carrying authorization method of access control wherein the middle man is cut out and the client can send a proof of authorization of access directly to the serving party. It is then the job of the server to verify this proof and, if applicable, provide the requester with the desired information. Here, the burden of proof is on the client, rather than on the server.

Systems which utilize proof-carrying authorization are both easier to implement and more efficient. In systems that don't use proof-carrying authorization, the verifier must decide whether or not to grant access to a requester. Programming such communications is difficult and time-consuming. However, in the method of proof-carry authorization, the requester can simply supply in its proof the reason why it should be granted access. Thus the method of proof-carrying authorization is faster and more straight-forward. While authorization of a proof in systems which use proof-carrying authorization takes time which is linear in the size of the proof, the verification or denial of access of a machine is much more time-consuming (6).

How to represent proofs

In order to make use of the method of proof-carrying authorization, it is imperative to be able to have a language which can represent proofs and policies. Proofs need to be represented as some sort of digital certificate. But in order for clients and servers to write and interpret proofs correctly, the language system must be unambiguous and must be easily dissected and understood by all the machines. This is difficult due to the complexity of most policies. Each distributed system has its own policy, and within a system, each set of data could have its own policy.

The problem of creating a language with which to encode the policy is discussed in the next section. Once this language is established however, implementing the language in a way that is understandable to a computer follows from converting each rule of the policy into an inference rule and representing proofs rigorously as a logical inference (6).

$PCML_5$

The growing popularity of distributed systems in the form of web-based services motivates the creation of a language in which policies can be encoded and distributed programs can be expressed (9).

Avijit, Datta and Harper introduced a programming language, $PCML_5$, in their paper entitled *Distributed Programming With Distributed Authorization* (9). $PCML_5$ is intended to simplify the programming of proofs for security in a distributed system with distributed authorization. It achieves this through its support for distributed computations, the implementation of an authorization logic, and the ability to represent both policies and proofs (9).

An authorization logic is a formal system of logic which supports the encoding of policies as theories. $PCML_5$ encodes an authorization logic by integrating the framework of the logic directly into $PCML_5$'s type system. For example, we can use an authorization logic to prove progress and preservation. The proof of progress states that if m is an expression, then either m is a value or m steps to some other expression m'. The proof of preservation states that if m is of type T and m steps to m' then m' is of type T. Together, progress and preservation ensure type safety.

Why implement $PCML_5$ in Beluga?

The successful implementation of $PCML_5$ in Beluga provides formal guarantees about $PCML_5$ itself. Specifically, we can gain insight into the theorems such as progress and preservation which together ensure type safety of the language. This case study will also aide in the understanding of both the development and implementation of software systems and programming with proofs. Implementing $PCML_5$ in Beluga raises awareness about possible issues that may arise during these sorts of implementations. Moreover, we will gain insight into the tools which are necessary to allow the average programmer to easily specify and verify complex behavioral properties.

Why Beluga?

Beluga is not the only language which can accomplish the desired task successfully. However, the use of Beluga is motivated by some of the properties inherent in Beluga programs as well as by a desire to fill some of the gaps in our knowledge of Beluga.

One advantage to using Beluga is its support for dependent types and its lack of pre-defined types. These properties allow for a simpler implementation of security procedures as defined by $PCML_5$, as we can simply define the types as prescribed by $PCML_5$. $PCML_5$ is designed to encode authorization policies and certificates easily. Beluga's support for automatic capture-avoiding substitution and renaming opens the Beluga framework to the ability to program with authorization policies and certificates, making Beluga sufficient for this project.

Moreover, this case study will aide in the understanding of some of the more technical issues regarding this task, as it is a future goal to implement security procedures in Beluga. A successful implementation of $PCML_5$ in Beluga will show that security is viable and will show simplicity of proof-carrying applications due to Beluga's built-in features. Upon successful implementation, we will have another working example of Beluga code, proof that Beluga can be learned quickly by the average programmer with no previous knowledge of Beluga, and a simpler way in which to write security procedures in Beluga. Although $PCML_5$ requires common substitutions and renaming, a Beluga programmer can essentially ignore these issues, thus encode $PCML_5$'s authorization logic more simply.

How to implement $PCML_5$ in Beluga

In order to implement $PCML_5$ with Beluga, one must first encode the primitive types. Beluga has no predefined data types, so this was all done by hand. An example of a piece of the authorization logic which $PCML_5$ encodes is expressed as:

Kinds K ::== Wld | A_1 Affirms A_2

Constructors A ::== α | A_1 x A_2 | $A_1 \rightarrow A_2$

These primitive types can be represented in Beluga as follows:

```
kind : type.                    %type kind
cnstc : type.                   %type constructor
wld : kind.
affirms : cnstc -> cnstc -> kind.
alpha : cnstc.
cross : cnstc -> cnstc -> cnstc.
arr : cnstc -> cnstc -> cnstc
```

Next, I implemented the inference rules. In order to do this, we make use of the fact that substitution and renaming is automatic in Beluga. Dependent types in $PCML_5$ (e.g. lam, exists, pi etc.) are thus easily converted into Beluga. When using these functions, a Beluga programmer need not think twice.

An example of an inference rule expressed in $PCML_5$'s logic is:

$$\frac{m{:}A_1 \times A_2 @ w}{\pi_2 m{:}A_2 @ w}$$

The statements above the horizontal line are the assumptions, while the statement below the horizontal line is the conclusion following the assumption of the statements above the line. This inference rule states that if m is of type $A_1 \times A_2$ (and thus $m = (m1, m2)$) in the world w then the second element (second projection) of m is of type A_2 in world w.

To translate this inference rule into Beluga's language, we first must have previously declared the following types:

```
cnstc : type.
term : type.
world : term.
cross : cnstc -> cnstc -> cnstc.
scnd : term -> term.
```

We then implement the "sentence structure" with which Beluga will represent $m : A @ w$:

```
is_inwld : term ->  cnstc -> cnstc -> type.
```

Finally, we can represent the inference rule itself in Beluga as:

```
isinwld_scd : is_inwld M (cross A_1 A_2) W ->
              is_inwld (scd M) A_2 W.
```

Lastly, we can make use of our previously implemented primitive types and inference rules to implement the main theorems of the paper. Proving the theorems involves a case by case analysis on the derivation of some statement in the hypothesis of the theorem.

One theorem that is proven is the progress theorem. In the $PCML_5$ paper, the statement of the theorem is written as:

If $m : A@w$ then either $mval_A$ OR $\exists m', A'$, such that $m ; A \rightarrow w\ m' ; A'$.

In plain English, this translates to: if m is of type A in world w then either m is a value under the assumptions A or there is some m' and A' such that m evaluated under A in the world w steps to some expression m' under A'.

In order to encode this statement into Beluga, we must first somehow encode this "either" in the statement. We can do this by creating a type called result, which encodes that the result is either a value or an evaluation (world shift):

```
result : term -> type.
r_val : {W:cnstc}{AA:active_prin} is_val W M AA
   -> result M.
r_ws : is_ws M AA W M' AA' -> result M.
```

With this extra tool, we can now state the theorem in Beluga by following the rules for function declaration:

```
schema apctx = active_prin ;
rec thmc4 :
   {AA:[.active_prin]} [. is_inwld M A W] ->
   [. result M] =
   mlam AA => fn d => case d of
.....
```

In the proof of this theorem, we proceed by a case by case analysis on the derivation of `is_inwld M A W.`
In the appendix of Avijit, Datta and Harper's paper, each possible derivation is followed by a systematic reasoning as to why the theorem would hold true under this derivation of the hypothesis.

For example, consider the derivation of `is_inwld M A W` , written in the appendix as:

Case: $\frac{\Delta;\cdot\vdash^{A}_{\Sigma_c;\Sigma_t} m:A_1\times A_2@w}{\Delta;\cdot\vdash^{A}_{\Sigma_c;\Sigma_t} \pi_2 m:A_2@w}$

$\Delta;\cdot\vdash^{A}_{\Sigma_c;\Sigma_t} m : A_1 \times A_2@w$ (Premise)

Either $m\ val_A$ (I.H.)

or $\exists m', A, .m; A \mapsto_w m'; A'$

Subcase: $mval_A$ (Lemma C.5)

$m =< m_1, m_2 >$

$pi_2, m_1, m_2.; A \mapsto_w m_2; A$

Subcase: $m; A \mapsto_w m'; A'$

$\pi_2 m; A \mapsto_2 \pi_2 m' : a'$

We can represent this in Beluga as:

```
| [. isinwld_scd D] => let [. R] = thmc4 [.AA] [. D] in
   (case [. R] of
    | [. r_ws T] => [. r_ws (isws_scd T)]
    | [. r_val W AA' V] =>
       (case [. V] of
        | [. isval_xV V1 V2] => [. r_ws (isws_pairscd V)]
     ))
.....
```

The Lemma C.5 that is referred to in the derivation analysis is a weakening lemma. In Beluga, it is not necessary to implement the lemma itself, and it can be implemented directly into the function (as can be seen above in the case of r_{val}). Moreover, it is not necessary to explicitly mention the impossible cases of [. V], the value. There is only one possible way the statement can be true under the assumptions of that case and so only that case needs to be expressed.

Moreover, we do not need to represent the contexts. They are either constant or can be easily included directly in the statement. This removes a lot of notation in Beluga's implementation.

Similar translations are done for all the possible derivations.

We then continue and implement a proof of preservation in Beluga. As described in PCML$_5$:

Theorem C.11 (Preservation of consistency). *Let $\Sigma_c;\Sigma_t$ be well-formed signatures, and Φ be an access control theory. Further, let all API constants declared in Σ_t be consistent in the sense of Def. C.7. If $m; A \mapsto^{\Sigma_c;\Sigma_t;\Phi}_w m'; A'$ and $\Delta;\cdot\mapsto^{A}_{\Sigma_c;\Sigma_t} m : A@w$, then $\Delta, \Phi, \cdot \mapsto^{A'}_{\Sigma_c;\Sigma_t} m' : A@w$*

Proof: Assume $m; A \mapsto^{\Sigma_c;\Sigma_t;\Phi}_w m'; A'$ and $\Delta;\cdot\mapsto^{A}_{\Sigma_c,\Phi;\Sigma_t} m : A@w$ We proceed by induction on the derivation of $m; A \mapsto^{\Sigma_c;\Sigma_t;\Phi}_w m'; A'$

This means if m is evaluated under A in world w to m' under A' and m is of type m in world w then m' is of type A in world w. This can be implemented in Beluga using a similar method as the previous proof. The statement of the theorem is written as follows:

```
rec thmc11 : [. is_ws M AA W M' AA'] ->
   [. is_inwld M A W] ->
   [. is_inwld M' A W] =
fn d => fn f => case d of
....
```

We will not review how to implement each specific case here, as it is the same methodology as the preservation cases.

Benefits of using Beluga

Following the successful implementation of PCML$_5$ in Beluga, it becomes clear that Beluga offers an advantageous environment in which to implement PCML$_5$ and encode proof-carrying code. The Beluga framework leads to a comparatively concise encoding of both the authorization logic and the theorems laid out in the PCML$_5$ paper. One factor leading to this concise representation is the elimination of the need to repeatedly represent the contexts. It was unnecessary to encode impossible or irrelevant cases, cutting down greatly on the amount of notation.

Moreover, PCML$_5$ required an implementation of separate lemmas for inversion, weakening, canonical forms, substitution or renaming. In Beluga however, these are either automatic or could be encoded directly in the function.

To get a more concrete idea of how much more concise PCML$_5$ is when implemented in Beluga, consider the exact amount of code needed. There is no way to save space when implementing the primitive types of PCML$_5$'s authorization logic in Beluga; each primitive type takes one line to encode. However, since in most cases we do not need to encode the contexts, we save a lot of symbolism when using Beluga to write out the inference rules and the theorems. The inference rules each take one line in Beluga. Due to the fact that we do not need to implement supplementary lemmas in order to successfully prove the progress and preservation theorems in Beluga,

these full proofs of these theorems take up less space on the page. Moreover, had we used another functional programming language to implement $PCML_5$'s authorization logic, we would have had to implement these lemmas since the operations are not automatic in most other programming languages.

To get a more concrete understanding of the amount of space saved, consider that the eight pages it took the appendix of the paper to fully prove the progress and preservation theorems that took Beluga a mere 250 lines of code (approximately five pages).

Conclusions

The successful implementation of $PCML_5$ in Beluga further motivates the continuing development of Beluga. The successful implementation of $PCML_5$ in Beluga demonstrates that using Beluga to code proof-carrying authorization in a distributed system can be quite simple. Beluga allows for code to be concise, and its framework is compatible with proof-carrying code. Moreover, Beluga is an accessible and useful language that can be eventually used universally. Lastly, this project offers insight into the development of proof-carrying authorization in distributed systems.

References

[1] A. Felty and B. Pientka Reasoning with Higher-Order Abstract Syntax and Context: A Comparison. In First International Conference on Interactive Theorem Proving (ITP '10), Lecture Notes in Computer Science (LNCS), vol 6172, pages 227-242, Springer, 2010.
[2] A. W. Appel and E. W. Felten. Proof-Carrying Authentication. In CSS '99: Proceedings of the 6th ACM conference on Computer and communications security, pages 52-52, Aug 1999.
[3] B. Pientka. "A Beginner's Guide to Programming in Beluga." McGill University. Sept 2010 <http://www.cs.mcgill.ca/~cs523/handouts/intro-beluga.pdf>
[4] Brigitte Pientka. Beluga: programming with dependent types, contextual data and contexts. In 10th International Symposium on Functional and Logic Programming (FLOPS '10). Lecture Notes in Computer Science (LNCS), vol 6009, pages 1-12, Springer, 2010.
[5] Deepak Garg and Frank Pfennig. Non-Interference in Constructive Authorization Logic. In Proceedings of the 19th IEEE Computer Security Foundations Workshop (CSFW 19), pages 283-296, 2006.
[6] Deepak Garg. An Introduction to Proof-Carrying Authorization. Nov 2007.
[7] Henrik Nilsson. "Untyped λ-calculus: Operational Semantics and Reduction Orders". Lecture 11. University of Nottingham. July 2012. <http://www.cs.nott.ac.uk/~nhn/G54FOP/LectureNotes/lecture11-9up.pdf>.
[8] Isup Lee. "Introduction to Distributed Systems". Univeristy of Pennsylvania. 2007. <http://www.cis.upenn.edu/~lee/07cis505/Lec/lec-ch1-DistSys-v4.pdf>
[9] Kumar Avijit, Anupam Datta, and Robert Harper. Distributed Systems With Distributed Authorization. In TLDI '10: Proceedings of the 5th ACM SIGPLAN workshop on Types in language design and implementation, pages 27-38, ACM, 2010
[10] Lujo Bauer, Michael A. Schneider, and Edward W. Felten. A Proof-Carrying Authorization System. Technical Report TR-638-01, Princeton University, April 2001.
[11] Petru Eles. "Distributed Systems". Linkoping University. <http://www.ida.liu.se/~TDDB37/lecture-notes/lect1.frm.pdf>
[12] Stephen A. Edwards. "Functional Programming and the Lambda Calculus". Columbia University. 2008. <http://www.cs.columbia.edu/~sedwards/classes/2010/w4115-spring/functional.pdf>

REVIEW ARTICLE

Operational transformation in cooperative software systems

Clarence Leung[1]*

[1] School of Computer Science, McGill University, Montreal, QC

*Email Correspondence:
clarence.leung@mail.mcgill.ca

Abstract

Modern cooperative software systems involve multiple concurrent users undertaking a common task in a real-time distributed environment, such as editing a shared text document. Maintaining data consistency, transaction causality, and replication convergence in such an environment, while providing fast client responsiveness, is a substantial challenge for classical distributed computing techniques. Operational transformation (OT) is a class of concurrency algorithms and data models that supports these functionalities, which has drawn signicant research attention in the past decade. In this review, we discuss the basic components of operational transformation models, the algorithms involved, and their actual implementations in real-world networked systems. We compare several existing OT control algorithms, the transformation functions and properties supported by each of the algorithms, and the trade-offs that are made with respect to each one. The data and operational models used in OT are well suited for high-latency environments such as the Internet, making them more frequently used in modern web services. Although many different OT control algorithms exist, choosing the most effective one often depends on the particular operations that an application must support.

Introduction

Research in computer-supported cooperative work (CSCW) in the past decade has led to the development of many new collaborative web services, such as blogs, wikis, and social networks. These services can span from only a few simultaneous clients over a local network, to millions of daily users worldwide. Handling the availability of these services becomes more difficult at a larger scale, especially in modern web applications, where notifications of user actions must be delivered in near real-time (1, 4, 7).

However, a common problem that arises when engineering real-time applications, such as collaborative editing services like Google Docs, is to find the appropriate way to handle conflict resolution due to multiple concurrent users. Traditional (pessimistic) approaches to concurrency prevent conflicts from occurring by locking a resource while a client is accessing it, but doing so leads to slower application responsiveness as multiple users are unable to modify a document simultaneously. Other approaches to concurrency, such as optimistic concurrency techniques, assume that data contention is rare, and eschew locks in favour of repairing conflicts when they occur. However, in real-time applications, conflicts are common, and thus simple optimistic concurrency techniques cannot be used without hurting performance while repairing broken states.

A new class of optimistic (non-locking) algorithms and data structures, known as operational transformation (OT), has been highly researched in the past decade in solving these conflict resolution problems. OT algorithms function by considering all user-submitted actions as units known as "operations", and handle any operations that conflict with each other by changing their properties relative to the changes that have already been applied to a particular document.

In this paper, we formally discuss the characteristics of real-time web applications, as well as the models required for us to properly verify that our documents are consistent. Using these models, we describe the operational transformation approach in detail, and survey several different approaches to OT that have been developed.

Traditional web approaches

Traditional web services handle client concurrency through the basic request/response model provided by HTTP, an asymmetric protocol. Request/response paradigms work by not having any response until a client specifically initiates a request to a web server, so no information is passed directly from the server to the client even if new updates have occurred on the server (7).

This protocol would be successful in applications that did not re-

quire real-time delivery of user actions, which most blogs, wikis, and representational state transfer (RESTful) web applications would be considered. However, for modern web applications that require real-time collaboration, there are several challenges that need to be overcome. Data must often be broadcasted from the server to the client, to keep the results of user actions up to date. This "server push" can be simulated through continuous polling of the HTTP resource, or keeping a live client-server connection, but this often results in a fragile connection, as it uses the HTTP protocol in a way that it was not intended to be used. Beyond the problems associated with response time, traditional pessimistic concurrency schemes, where a document is locked while one user is updating it, are often poor for developing real-time applications due to a lower responsiveness (2, 3).

Collaborative editing systems

One of the more well-known applications of real-time web services is those of collaborative editing systems such as Google Docs or Etherpad (14). These services provide traditional word processing and spreadsheet applications, but through a shared interface that allows simultaneous editing of a document.
There are three main characteristics that define a collaborative editing system (6):

1. **Real-time:** The local client interface should have a fast response time. Optimally, it should have a response time similar to a single-user editor.

2. **Distributed:** Each user may be geographically separated, from different machines to different communication networks, with varying latency delays.

3. **Unconstrained:** Any user can freely edit any part of the document at any time, without any constraints.

Instead of a pessimistic concurrency scheme, collaborative editing systems often use an optimistic concurrency scheme, where a non-locking transaction scheme is used. Replication is often used in order to allow the local client interface to respond quickly, by having local actions immediately update the local copy of the document. These factors provide the necessary responsiveness for real-time applications, but now lead to a new difficulty: maintaining consistency and properly resolving conflicts throughout the document, while multiple clients are concurrently editing it (8, 9).

Optimistic concurrency techniques

Optimistic concurrency techniques have been well-studied within the context of databases and distributed simulations, which make the assumption that data contention between two different transactions are rare, and conflicts rarely happen. These techniques are often used through web user interfaces, as the stateless nature of HTTP makes it difficult for locks to be used, since a user can simply leave a web page at any time without notifying the application that they have cancelled their transaction.

The simplest form of optimistic concurrency control commits a transaction only when all necessary changes have been tentatively completed, and no conflicts from other transaction have occurred. If a conflict has occurred, a "rollback" is used to revert a data object back to a previous, non-conflicted state, and the transaction which attempted to commit is annulled. This guarantees that our data is always valid and consistent, though this simple form of optimistic concurrency will not be successful in areas where many conflicting operations may occur. The lessened overhead due to locks being unnecessary, however, can lead to greater throughput if the number of conflicts is few.

However, for collaborative editing systems, conflicts occur often, making the simple optimistic concurrency technique infeasible. In later sections, we will discuss the modifications that operational transformation makes to the simple optimistic concurrency technique that make rollbacks unnecessary for resolving conflicts.

Consistency models

Consistency models define the rules that determine what data each client can see in memory, when a particular set of operations have been performed by different clients. An example of when an inconsistency can arise is when one client writes to its own local copy of data in a distributed data store, but has not yet propagated those changes over to the other clients. If those other clients wish to read or write to that data, then they may be using an older version of the data that has not taken into account the other client's modifications.

For collaborative editing systems, several consistency models have been defined to formally verify the properties that clients must support to successfully resolve conflicts in the local copies of their documents. We discuss one of the more commonly referenced models by Sun *et al.* (12), which suggests that consistency in a cooperative editing system be defined by three properties based on each of the operations (discrete transactions submitted by each client):

1. **Convergence:** When the same set of operations has been executed at each site, then the copies of the document are also identical.

2. **Causality preservation:** Given two operations O_A and O_B if $O_A \rightarrow O_B$ (O_A causally occurred before O_B), then O_A is executed before O_B at each site.

3. **Intention preservation:** For every operation *O*, the intention of *O* at the initial site where *O* is initially submitted will be identical to executing *O* at all other sites. The intention of an operation *O* is defined as the resulting document which is achieved by applying *O* on the document state from which *O* was generated.

The convergence property guarantees that the document will be correct at the end of any particular editing time period, while the causality preservation property guarantees that the document will be correct at any point during editing. The intention preservation property is similar to the convergence property, but additionally deals with the situation where operations are submitted from two different initial document states. This property is often the most difficult to achieve, and often requires a submitted operation by one client to be modified once received by another.

Operational transformation

Operational transformation (OT) is a class of optimistic concurrency algorithms and data structures that are well-suited to satisfying the three properties of the consistency model of collaborative editing systems discussed above. The OT mechanism can be divided into the models that represent the data and its changes, and the algorithms that are used to ensure the correctness of the system (1, 3, 4, 5).

Basic example

The basic operational transformation technique (inclusion transformation) can best be illustrated by a simple example. Say we are given a text document, containing the sentence:

"operational transformation is very fun"

Now, two users, A and B, each simultaneously submit an operation on this document. For the purpose of demonstration, an operation has the signature *Operation(position, string)*, where "position" represents the index of a word in the sentence, and "string" is the target string we want to operate on:

1. User A submits the operation O_A, *Insert(0, 'using')*, inserting the word "using" at the beginning of the sentence.

2. User B submits the operation O_B, *Delete(4, 'very')*, deleting the word "very" from the 4th word of the sentence.

If we perform these operations in the order above without transforming them, we get the following result:

"using operational transformation very fun"

As the second operation deletes the 4th word, which accidentally deletes "is" instead of "very". Instead, we need to transform the positional parameter of O_B by one in order to obtain the correct result, which our OT transformation algorithm can calculate and perform, giving us *Delete(5, 'very')* instead. This gives us the correct result:

"using operational transformation is fun"

Data models

Data models in OT are the representations of the data in a particular collaborative session. A data model can represent a document, a spreadsheet, a drawing, or any other object that different clients can collaboratively operate on. The most basic OT data model is that of a linear space in memory, such as that of a single string (1).

Recent developments in OT have allowed the development of more complex OT data models, such as those involving rich-text editing with markup or an advanced application state represented in JSON (JavaScript Object Notation). However, as the data models used in OT become more complex, so do the models that are necessary to represent a particular operation on the data.

Operation models

An operation is the fundamental block of the OT mechanism, and in the simplest OT data model, an operation is simply an insertion or deletion of a single character in memory. As well, more advanced operations can be created from combining insertions and deletions, such as a substitution operation (a combination of an insertion and deletion operation) (10, 11).

For advanced data models that may use more than a linear space in memory, we often find more complex operations such as moving data between memory addresses. For example, OT word processors can add operations to start and close annotation boundaries to represent markup tags, and OT JSON structures can have edit operations on particular lookup paths of a JSON object.

Past research attempted to generalize OT data and operation models, and to only allow insertion, deletion, and substitution operations on the data. However, for most modern OT implementations, application-specific data and operation models are used. These also require operation transformation functions to maintain the intention preservation property, one for each pair of operations, as each type of operation must be able to be modified by the OT control algorithm relative to each other operation type. Thus, for *N* application-specific operations, *N x N* transformation functions are required for OT (13).

Operation composition

For each client in the collaborative system, each client must perform three different activities (12):

1. **Operation generation:** an operation is specified by the user, and broadcast to all of the other clients.

2. **Operation reception:** an operation request is received from another site.

3. **Operation execution:** an operation request is executed at the site.

From these activities, we have two major sets of algorithms that handle conflict resolution for the system: the low-level operation transformation algorithms, and the high-level control (or integration) algorithms. The control algorithms handle the operation models at generation and reception, by generating a set of corresponding transformation algorithms, which then transform the target operation model before executing it at the received site (6).

We will discuss the specific actions taken during each of the activity phases when we compare the different control algorithms possible in OT.

Operation transformation

Operation transformation relies on satisfying several properties that ensure that the system is correct at each point. These are lower-level properties than the consistency model described previously, but not all properties are satisfied by each control algorithm, and where these properties are ensured may differ between each one. These transformation properties are often divided up into two different types: convergence properties and inverse properties (1, 2).

If an OT system is to support particular functionality, then it must be able to support certain transformation properties. For group editing and consistency maintenance, the system must support a transformation function known as Inclusion Transformation (IT). For group undo, where the effect of a previously executed operation is un-done at all sites, and all operations executed after it are all re-transformed, the system must support another transformation function known as Exclusion Transformation (ET).

Transformation functions

As mentioned before, the two base transformation functions that generate all other transformation functions are (11):

1. Inclusion Transformation [IT(O_A, O_B)]: transforms operation O_A against another operation O_B in such a way that the impact of O_B is effectively included.

2. Exclusion Transformation [ET(O_A, O_B)]: transforms operation O_A against another operation O_B in such a way that the impact of O_B is effectively excluded.

As previously mentioned, these transformation functions are generated for each pair of operations in the application space.

The generated transformation functions are basically generated such that the position in which the operation is applied is transformed by an offset relative to the other operation. For example, one of the most basic inclusion transformations between two insert operations can be represented as follows.

For an insert function with signature *insert(p, c, s)* where:

- p = position of the character to be inserted
- c = character to insert
- s = site priority of the client

Then, the transformation function is the algorithm:

Algorithm 1 - Inclusion transform with two inserts

```
if (p1 < p2) then
        return insert(p1, c1, s1)
else if (p1 == p2) and (s1 < s2) then
        return insert(p1, c1, s1)
else
        return insert(p1 + 1, c1, s1)
end if
```

Thus, depending on the priority and insertion location of the operation, we can transform the resulting operation as needed.

Transformation properties

Convergence properties must generally be satisfied by all control algorithms, as part of the convergence guarantee in the consistency model. There are generally two convergence properties, known as CP1 and CP2 (8, 9):

1. **CP1:** Given a state S, and two operations O_a and O_b:

   If $O'_a = IT(O_a, O_b)$ and $O'_b = IT(O_b, O_a)$
   Then: $S \circ O_a \circ O'_b = S \circ O_b \circ O_a$

2. **CP2:** Given a state S, and three operations O, O_a, and O_b:

   If $O'_a = IT(O_a, O_b)$ and $O'_b = IT(O_b, O_a)$
   Then: $IT(IT(O, O_a), O'_b) = IT(IT(O, O_b), O'_a)$

Similarly, there are three inverse properties that need to be satisfied to support the "group undo" operation. This is not a part of the consistency model above, but is a commonly implemented operation in practice (8, 9).

1. **IP1:** Given a state S, and sequence of operations $O \circ \overline{O}$:

   $S \circ O \circ \overline{O} = S$

2. **IP2:** Given any operation O, and a pair of operations O_x and $\overline{O_x}$:

   $IT(IT(O, O_x), \overline{O_x}) = IT(O, I) = 0$

3. **IP3:** Given a state S, and two operations O_a and O_b:

   If: $O'_a = IT(O, O_b)$, and
   $O'_b = IT(O_b, O_a)$, and
   $\overline{O_a}' = IT(\overline{O_a}, O'_b)$
   Then: $\overline{O_a}' = \overline{O'_a}$
   Or: $IT(\overline{O_a}, O'_b) = \overline{IT(O_a, O_b)}$

Control (integration) algorithms

The OT control algorithm is the main high-level algorithm governing the collaboration functions that are available to the system. This algorithm controls the time/space complexity of the system, handles ordering of the operations by how timestamps are applied, and processes the incoming operations into the modified transformed operations (10, 12).

Most OT systems, by design, are peer-to-peer distributed systems, but modifying the control algorithm can also determine whether the system can also be centralized, which is important when adapting an OT system over the HTTP protocol.

Comparison of control algorithms

We now compare several implementations of OT control algorithms, from ones that are historically relevant in literature to modern day OT control algorithms that are used over HTTP.

dOPT (GROVE)

dOPT, used in the GROVE groupware outline editor program, is one of the earliest concurrency control algorithms for operational transformation (1). It uses a transformation matrix to handle conflict resolution, where for m operations, there is an $m \times m$ matrix of the inclusive transformation resultant functions, for each pair of operations.

Timestamps for each client are handled by a vector timestamp, where a state vector s_i for a client C_i will have at position j, the number of operations known to have been executed by client C_j.

A request queue, Q_i, is used to queue up the requests that have been generated or received, and are waiting to be executed. Requests are handled in the form $<j, s, o, p>$ where:

- j = the site that requested the operation
- s = the vector timestamp for the requested site
- o = the operation to be performed
- p = the priority of the operation

Then, for the three activities (generation, reception, execution), dOPT performs the following:

Algorithm 2 - Generate Operations
`receive` operation O from the user
`calculate` the priority p of O
`append request` $< i, s, o, p >$ to Q_i
`multicast` $< i, s, o, p >$ to the other clients

Algorithm 3 - Receive Operations
`receive` $< j, s, o, p >$ from the network
`append` request $< j, s, o, p >$ to Q_i

Algorithm 4 - Execute Operations
`for each` request `in` Q_i `where` $s_j \leq s_i$ `do`
  `remove` request $< j, s_j, o_j, p_j >$ `from` Q_i
  `if` $(s_j < s_i)$ `then`
    $< k, s_k, o_k, p_k >$ = most recent log entry
    `where` $s_k \leq s_j$ (or $\emptyset$ otherwise)
    `while` $< k, s_k, o_k, p_k > \neq \emptyset$; `and` $o_j \neq \emptyset$; `do`
      `if` part k of s_j is $\leq$ part k of s_k `then`
        o_j = `transform`(o_j, o_k, p_j, p_k)
      `end if`
      $< k, s_k, o_k, p_k >$ = next log entry
      (or $\emptyset$ otherwise)
    `end while`
  `end if`
`end for`
`perform` operation o on i's data model
`add` request to history log
`increment` j^{th} component of s_i by 1

Although dOPT is simple and satisfies many of the correctness properties, a scenario was found where dOPT could not always ensure convergence, when remote concurrent requests with similar operations were transmitted from two different sites.

Further research later helped to solve the problem, by using different data structures for the timestamp and conflict resolution (as in the Jupiter algorithm), or by transforming the log entries themselves whenever they are used to transform an update.

Jupiter

Jupiter, a multi-user remote collaboration virtual world developed at Xerox PARC by Nichols *et al.*, addressed some of the issues found in dOPT (7). One of the major changes in Jupiter was to have a centralized coordination server that uses a change propagation algorithm to keep the clients updated and in check.

The optimistic concurrency techniques are used in the individual client-server links, where the synchronization algorithm in the individual client-server links is very similar to the dOPT algorithm.

The major changes in Jupiter, in comparison with dOPT, is that instead of a transformation matrix, Jupiter uses a function called *xform*, which takes a pair of client and server operation requests, and transforms them as a new pair of operations that lead the client and server operations to the same final state.

The *xform* technique is successful while the client and server are in the same starting state, but if the client and server diverge too much, then the operation must look back into its past history to properly calculate the converging operations. This is often done by keeping track of operation revision history, but not directly transforming saved messages, like dOPT does, which causes the incorrect scenario. Instead, Jupiter calculates past converging operations, even if it does not apply them, in order to generate the correct recent transformed operation.

Google Wave OT

Google Wave OT, the algorithm behind Google Docs and the former Google Wave product, was a modification on top of the Jupiter algorithm (14). The main changes to the Jupiter control algorithm are mainly in the client/server communication protocol, as well as optimizing the transformation functions for batch updates.

Instead of sending client operations to the server whenever new operations are requested, the Google Wave OT client must wait for a server response before sending any more operations. In the meantime, the operations are composed together into a single buffer, and then sent together once the server has processed the last batch of operations and converged.

This technique serves two major purposes. First, as the server acknowledges the client before any new operations proceed, the client is capable of predicting the operation path that the server will take, and thus always send operations to the server that are always on the path. This simplifies the server implementation significantly, as the server only needs to keep track of its own history. Instead of taking a quadratic $O(h^2)$ space for history on every possible path, it only takes up a linear $O(h^2)$ space on the server for history, where h is the number of previous operations (and possible causal operations) required to calculate the correct OT path.

Finally, the transformation algorithm handles operations in streams, rather than as single discrete operations. These operation sequence streams are guaranteed to be in order, as well as linear, so the streaming transformations can be performed in linear time.

ABT (TIPS)

A more modern protocol for OT is based on the admissibility-based transformation (ABT) framework, which has been formally verified for correctness (8,10). TIPS, an implementation of ABT, builds on top of existing HTTP protocols, and also uses a centralized server.

Clients in TIPS are able to join or leave a session at any time, and the clients also independently decide to sync with the server. As with Google Wave OT, operations are buffered and synchronized with the server only at some particular server-determined interval. This is useful, as it can be easily adapted to JavaScript-based long polling methods, which are the traditional method of maintaining an open connection over the web.

Once the server has asked all of the clients to synchronize, each client that has responded will have their operations applied to an n-way merge algorithm, which will create a sequential set of operations as output. The clients receive this sequential set of operations in another interval, which is adapted by them through another algorithm, ITSQ, which performs the inclusion transformation step that updates the document correctly.

The benefit of TIPS over existing protocols is that it is more capable of allowing clients to dynamically enter and exit a particular OT session, as well as being more robust when dealing with client or network failures.

Summary

Operational transform is still a rapidly developing technology, with many particular algorithms and system implementations that are well-suited for particular tasks. These algorithms can differ in the consistency guarantees they support, as well as functionality, and the most useful algorithm often depends on the specific application that is being built. Both the high-level design of the OT control algorithm, as well as the low-level design of the OT data and operation models is significant in determining which systems are the most practical.

Real-time collaborative applications are slowly becoming some of the more widely used applications in business environments, and with further development with Internet-accessible mobile devices, these applications will continue to flourish. Operational transformation will continue to be a significant research topic in the coming years as

networked cooperation becomes more widely used.

Acknowledgements

I would like to thank Eric Ferraiuolo of Yahoo! for assisting me in learning about the field, and for helping to provide resources and technical advice.

References

[1] C. A. Ellis and S. J. Gibbs. *Concurrency control in groupware systems.* In ACM SIGMOD89 Proceedings, **18**(2):399407, 1989.
[2] C. Sun, X. Jia, Y. Zhang, Y. Yang, and D. Chen. *Achieving convergence, causality-preservation, and intention preservation in real-time cooperative editing systems.* ACM Transactions on Computer-Human Interaction, **5**(1):63108.
[3] R. Bentley and P. Dourish. *Medium versus mechanism: Supporting collaboration through customization.* In ECSCW '95 Proceedings, 1995.
[4] A. H. Davis, C. Sun, and J. Lu. *Generalizing operational transformation to the standard general markup language.* In ACM CSCW'02, pp. 58 - 67, Nov. 2002.
[5] G. V. Cormack. *A calculus for concurrent update.* Technical Report CS-95-06, Dept. of Computer Science, University of Waterloo, Canada, 1995.
[6] D. A. Nichols, P. Curtis, M. Dixon, and J. Lamping. *High latency, low-bandwidth windowing in the Jupiter collaboration system.* In ACM UIST 95 Proceedings, Nov. 1995.
[7] B. Shao, D. Li, T. Lu, and N. Gu. *An operational transformation based synchronization protocol for Web 2.0 applications.* In CSCW 2011, pp. 563 572, Mar. 2011.
[8] D. Li and R. Li. *An approach to ensuring consistency in peer-to-peer real-time group editors.* Computer Supported Cooperative Work: The Journal of Collaborative Computing, **17**(5-6):553-611, Dec. 2008.
[9] D. Li and R. Li. *An admissibility-based operational transformation framework for collaborative editing systems.* Computer Supported Co-operative Work: The Journal of Collaborative Computing, **19**:1-43, Aug. 2009.
[10] R. Li, D. Li, and C. Sun. *A Time Interval Based Consistency Control Algorithm for Interactive Groupware Applications.* In Proc. IEEE Intl Conf. Parallel and Distributed Systems, pp. 429-436, Jul. 2004.
[11] C. Sun and C. Ellis. *Operational transformation in realtime group editors: issues, algorithms, and achievements.* In Proceedings of the ACM Conference on Computer-Supported Cooperative Work, pp. 59-68, Dec. 1998.
[12] N. Vidot, M. Cart, J. Ferrie, and M. Suleiman. *Copies Convergence in a Distributed Realtime Collaborative Environment.* In Proceedings of the ACM Conference on Computer-Supported Cooperative Work, pp. 171-180, Dec. 2000.
[13] D.Wang, A. Mah, and S. Lassen. *Google Wave operational transform.* Google Wave Federation Protocol White Papers, July 2010.

REVIEW ARTICLE

Autosomal Recessive Spastic Ataxia of Charlevoix-Saguenay (ARSACS): a once obscure neurodegenerative disease with increasing significance for neurological research

Xinlu (Crystal) Li[1]*

[1] Department of Biochemistry, McGill University, Montreal, QC

*Email Correspondence:
xinlu.li@mail.mcgill.ca

Keywords:
Autosomal Recessive Spastic Ataxia of Charlevoix-Saguenay (ARSACS): *A complex early-onset neurodegenerative disease found in high prevalence in the population of the Charlevoix-Saguenay region in Quebec due to founder effect.*
Founder Effect: *Significant reduction in genetic variability and increase in homozygosity level in a population descending from a small founder population geographically isolated from its ancestral population.*
Chaperones: *Important protein-folding quality control machineries that promote and maintain the proper folding of cellular proteins and target misfolded proteins for degradation.*
Mitochondrial Fission/Fusion Dynamics: *The delicate balance between the fission and fusion of mitochondria that is essential for its quality control, function and distribution in the cell.*

Abstract

Background: Autosomal Recessive Spastic Ataxia of Charlevoix-Saguenay (ARSACS) is a rare cerebellar ataxia occurring in the Charlevoix-Saguenay population in Quebec with high incidence as a result of founder effects. Following the discovery of the gene responsible for the disease, many other patient groups have been identified worldwide and the characterization of the gene product, sacsin, has unveiled similarities between the pathogenic mechanism of ARSACS and those of other major neurodegenerative disease.

Summary: The core symptoms of ARSACS consist of a triad of early-onset cerebellar ataxia, peripheral neuropathy and spasticity, which is accounted by degeneration of Purkinje neurons. The gene responsible for the disease is located on chromosome 13q11 and encodes for the chaperone sacsin. Drp-1, a GTPase crucial for regulating mitochondrial fission/fusion dynamics, has been identified as a potential substrate of sacsin, suggesting a link between the pathogenic mechanisms of ARSACS and prevalent neurodegenerative diseases such as Alzheimer's, Parkinson's and Huntington's diseases.

Autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS) is a complex hereditary neurodegenerative disorder characterized by a triad of early childhood-onset cerebellar ataxia (uncoordinated movements due to defects in the cerebellum (1)), peripheral neuropathy (peripheral nerve damages resulting in a variety of sensorimotor symptoms (2)) and pyramidal tract signs (traits such as spasticity, abnormal reflexes and loss of ability to perform fine motor movements due to defects in the neurons relaying signals from the cerebral cortex or midbrain to the spinal cord (3)) (4). The disease was initially named after its high incidence in the French-Canadian population of Charlevoix-Saguenay region in northeastern Quebec as a result of founder effect (5) but was later also discovered in many other regions and ethnic groups worldwide (6). *SACS*, the gene whose mutations are responsible for the ARSACS' symptoms, was identified by Engert *et al.* in 2000(7). The gene product, sacsin, constitutes one of the largest known human proteins and bioinformatics and biochemical characterizations of peptide domains in sacsin suggest that it performs its cellular function by integrating the ubiquitin-proteasome system and the chaperone pathways (6). Furthermore, a recent study of a transgenic sacsin-knockout mouse model associated the loss of sacsin's function with the disruption in neuronal mitochondrial fission/fusion dynamics (8). This finding in turn significantly elevated the research significance of ARSACS, since the disruption in neuronal mitochondrial dynamics underlies many prevalent neurodegenerative disorders such as Alzheimer's, Parkinson's and Huntington's diseases (9). Therefore, the purpose of this review is to delineate major clinical, genetic and biochemical findings on ARSACS to highlight the increasing research significance of this once obscure disease.

Clinical Features of ARSACS

The clinical features of ARSACS were first reported in 1978 by Bouchard *et al*. in a group of French-Canadian patients in Quebec whose families mostly originated from the Charlevoix-Saguenay region, where around forty families had migrated from Quebec city in the 1600's (5). The features were "remarkably homogenous" and were inherited in an autosomal recessive fashion, distinguishing the disease from most other classes of ataxia disorders except for Friedreich's Ataxia (also autosomal recessive) (5). Subsequent electrophysiological studies (i.e. electromyography and nerve conduction studies) (10), as well as brain biopsy studies using light microscopy (11), computed cerebral tomography (12) and magnetic resonance imaging (13), distinguished ARSACS from FA by pinpointing the anatomical origin of ARSACS to degeneration of the Purkinje neurons in the cerebellum, the region responsible for motor coordination and balance.

Clinical manifestation in all ARSACS patients reported by Bouchard *et al.* begins with gait unsteadiness observed in toddlers (12 to 18 months old) learning to walk (5). Spasticity (abnormal tensing of muscles) subsequently develops in the peripheries of the body and progressively increases in severity over time, accompanied by peripheral neuropathy (neuronal degeneration) and amyotrophy (muscle wasting) (5). The progression of the symptoms is much slower than in other forms of ataxia and severity of the symptoms can remain stable for a long period of time and suddenly worsen over the course of a few years (normally during early adulthood) (5). The individual is usually wheelchair-bound by the age of thirty or forty with life expectancy reduced to about fifty years of age (5).

The ARSACS Gene

A major turning point in ARSACS research occurred when *SACS*, the gene responsible for the disease, was identified by Engert *et al*. in 2000 via a series of haplotype and linkage analyses and physical gene mapping among French-Canadian ARSACS patients (7). The gene was located on the long arm of chromosome 13 (13q11) was found to contain a 12,794-base pair (bp) exon, the largest that had been identified in vertebrates (7). Furthermore, sequence analysis showed extensive conservation between mouse and human *SACS* (14). Population screening revealed a single base-pair deletion at position 6,594 (6594delT) in most ARSACS patients and a nonsense mutation at position 5254 (5254 C-T) in a small proportion of ARSACS patients (7). Both mutations were predicted to lead to the production of truncated protein products with their functions completely abolished (7). These findings, along with the migration history of the Chalevoix-Saguenay (CS) region, attributed the prevalence of the disease amongst the CS population to the increased homozygosity level in the population due to founder effect (7). In addition, though the large size of the *SACS* exon makes sequencing an overly expensive and time-consuming means for diagnosis, several fragment-based sequencing techniques were developed that could more efficiently identify patients harboring specific mutations in *SACS* (15, 16).

"ARSACS Goes Global"

The identification of *SACS* and advances in genetic linkage study tools resulted in the diagnosis of an increasing number of ARSACS patients in many other regions of the world. The trend started as early as in 2000, when several Tunisian patients previously diagnosed with another class of cerebellar ataxia were found to harbor mutations at loci linked to *SACS* (17). Since then, many other ARSACS patient groups have been identified in Italy (18-20), Japan (21-27), Britain (28), France (29-31), Spain (32-34), Netherlands (35) and Belgium (16, 36, 37). All identified ARSACS patients shared the core symptoms of ARSACS (early onset spasticity and ataxia increasing in severity over time accompanied by dysarthria/slurred speech, nystagmus/vision impairment and amyotrophy/muscle wasting) despite a few minor clinical variations, such as later onset and the lack of retinal striation or mental retardation. In fact, the trend of globalization is not unique to ARSACS. *Gomez et a*l. pointed out in an editorial that the "spreading" of ARSACS embodies a prevailing trend in the realm of neurological diseases: Several recessive neurological disorders originally considered to be restricted to a particular ethnic group or geographical location were later found to be present in many other populations around the world and Gomez *et al*. predicted that more neurological diseases were to follow suit (38).

With the identification of ARSACS patients worldwide, an increasing number of mutations in *SACS* were also revealed. To date, over 70 *SACS* mutations have been discovered in 13 countries (6). It is interesting to note that founder effect seems to play a role only in the Quebec patient population (6). Most of the currently identified *SACS* mutations are frameshift or nonsense point mutations residing within the exon (13, 16, 18, 19, 21-24, 36, 39-44) identified by *Engert et al.* However, the discovery of several point mutations and macrodeletions upstream of the exon in ARSACS patients (13, 16, 20, 25, 36, 40, 45) led to the upstream extension of the putative *SACS* gene to contain eight additional exons (6). In 2009, reverse transcriptase-polymerase chain reaction (RT-PCR) analysis of human brain mRNA, in conjunction with northern blot analysis and in-situ hybridization on a number of human tissues, confirmed that the *SACS* encodes a nine-exon transcript that is expressed in a variety of tissues including fibroblasts, skeletal muscle and the brain, particularly in the Purkinje cells of the cerebellum (46).

The Sacsin Protein

The gene product of SACS in human is a relatively large protein (~4500 amino acids) named *sacsin* (8). Based on its amino acid sequence, the secondary structure of sacsin has been predicted to consist of

two leucine zippers (a leucine-rich motif known to mediate protein dimerization (47)), three coiled coils and one hydrophilic domain in the carboxyl-terminal half of the peptide (7). To date, bioinformatics analyses have identified five distinct peptide domains in sacsin and their respective functions have been characterized to varying extents (See Fig. 1 for the location of the domains and selected ARSACS-causing mutations in sacsin) (48).

The N-terminal end of sacsin contains an ubiquitin-like domain (UbL) that was first reported by Parfitt *et al.* in 2009. Co-immunoprecipitation experiments have demonstrated the ability of N-terminal region of sacsin (up to and including the UbL domain) to directly interact with the 19S cap of the 26S proteasome, the primary protein-degradation machinery in eukaryotic cells (46). Nevertheless, this does not exclude the possibility that the observed interaction between the N-terminal region of sacsin and the proteasome could be due polyubiquitination-mediated targeting of sacsin to the proteasome instead of direct interaction between the sacsin-UbL domain and the proteaosome (46).

Towards the C-terminal direction, the next discernable structural entity consists of three SRRs (sacsin repeat regions), each of which is a supra-domain (complex of multiple independent peptide domains exerting synergistic functions) consisting of an N-terminal domain homologous to the Hsp90 (a chaperone)-ATPase domain and a hydrophobic C-terminal domain of unknown function (49). The sacsin-SRRs were first described by Engert *et al.* upon their identification of the SACS gene in 2000 and the ATPase activity of the N-terminal domain of the first SRR was verified by Anderson *et al.* to be similar to that of yeast Hsp90 (with turnover rates of 2.5/min and 1/min, respectively) (7, 49). In addition, ARSACS-causing missense mutations D168Y and R2703C in the first and third SRRs were both shown to abolish the

ATPase activity of the entire protein, thereby confirming the non-redundancy in the functions of the three SRRs in sacsin (15, 18). Bis-1-anilinoaphthalene 8-sulfonate (a molecule that becomes highly fluorescent when interacting with exposed hydrophobic surface of proteins) fluorescence measurement experiment revealed substantially increased surface hydrophobicity of the mutant SRR compared with the wild-type, indicating disruption in the normal folding of the supra-domain. However, since the study also demonstrated that the mutant protein remained soluble in the cell, the mutant SRR is most likely still able to fold to the extent of avoiding aggregation but with structural alterations significant enough to abrogate its ATPase activity (49).

The domain C-terminal to the third SRR was identified as XPCB domain by Kamionka *et al.* in 2004 (50). The sacsin XPCB-domain was found to exhibit 35% sequence identity with the hHR23 XPCB domain (50). hHR23 had been known to bind to and stabilize XPC, initiator of the global genome repair pathway. The hHR23/XPC complex is implicated in group C xeroderma pigmentosum (XP-C), a rare nucleotide excision repair deficiency disorder characterized by markedly increased photosensitivity and propensity for developing UV-induced skin cancer (50).

Interestingly, the hHR23 protein also contains an UbL domain, which has been shown to interact with the 19S regulatory subunit of proteasome *in vivo* (51). The 19S subunit of the proteasome has been found to exhibit independent chaperone-like activity, which was suggested to help maintain the proper conformation of the highly hydrophobic XPC molecule on XPCB (49). Since sacsin also contains an UbL domain, the sacsin-XPCB domain most likely function in a similar fashion as hHR23 and delivers the binding partner of sacsin to the proteasome or chaperones. However, the binding partner as well as the cellular activity of the sacsin-XPCB is yet to be identified.

The penultimate domain at the C-terminus of sacsin is the J-domain, first reported by Parfitt *et al.* in 2009 (46). J-domains are integral components of the DnaJ/Hsp40-class chaperones and mediate the interaction between Hsp40 and Hsp70 to enhance Hsp70's ATPase and substrate-binding activity (52). The Hsp40-Hsp70 complex constitutes the primary cellular machinery in assisting protein folding and quality control (52). Sequence alignment between sacsin- and Hsp40/DnaK-J-domains revealed a 60% similarity over 30 residues and the ability of the sacsin J-domain to recruit and stimulate Hsp70 and stimulate its chaperone activity was verified by in vivo complementation assays in bacteria (46). To date, the only identified ARSACS-causing mutation in the sacsin J-domain is the compound heterozygous mutation K1715X/R4331Q, with the arginine occupying a position known to be critical for the function of J-domains in most proteins (15). Note that since the R4331Q mutation alone did affect the ability of the sacsin J-domain to rescue the function of mutant bacterial DnaJ (Hsp40 class chaperone) with nonfunctional J-domain, both mutations most likely need to be present to exert a perceptible effect on the function of sacsin J-domain (46). The fact that K1715X

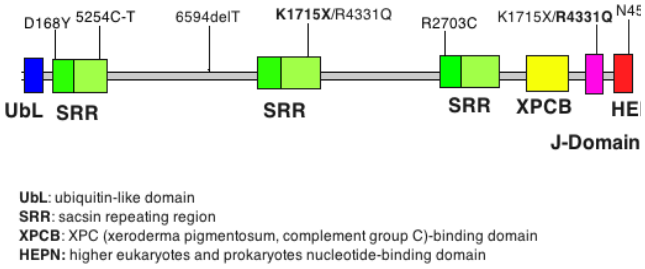


Fig. 1
A schematic diagram of sacsin showing the known structural domains and selected ARSACS-causing mutations. Note that the numbers in 6594delT and 5254C-T represent the position of the mutated nucleotide whereas the numbers in the rest of the mutations presented in the diagram indicate the position of the mutated amino acid.

resides within the second SRR suggests that the function of the sacsin J-domain may require its interaction with SRR, whose ATPase activity may contribute to J-domain's recruitment and stimulation of Hsp70.

Finally, the 110-residue region at the very C-terminus of sacsin constitutes the higher eukaryotes and prokaryotes nucleotide-binding (HEPN) domain, first reported by Crynberg *et al.* in 2003 (48). As its name suggests, the HEPN domain is known to bind nucleotides and a number of HEPN domain-containing proteins, such as the kanamycin nucleotidyl transferase family proteins, which are implicated in bacterial antibiotic resistance (48). The ability of the HEPN domain of human sacsin to bind to various nucleotides (i.e. ATP, ADP, GTP, ADP and etc.) was verified by Kozlov *et al.* in 2011 via nuclear magnetic resonance (NMR) spectroscopy and isothermal titration calorimetry (ITC) (53). Crystal structure of the human sacsin-HEPN domain revealed that it exists as a dimer, which interacts with nucleotides electrostatically via a symmetric binding pocket formed at the dimer interface (53). The ARSACS-causing mutation N4549D at the dimer interface was shown to hinder the proper folding and the dimerization of the domain, which is likely responsible for its loss of function (53). Interestingly, the sacsin-HEPN domain did not exhibit any GTPase or ATPase activity (53) and many suggested that the sacsin HEPN domain most likely functions as an "energy stockroom" to supply nucleotides (particularly GTP and ATP) for sacsin's chaperone-like activity (48, 53). However, there is a lack of direct evidence demonstrating the utilization of the HEPN-recruited nucleotides by sacsin or its binding partners and the "energy stockroom" notion therefore remains largely theoretical.

Fitting Together Pieces of the Puzzle

Experimental evidence accumulated to date strongly suggests that the sacsin protein functions as a Hsp40 family chaperone in neurons (54). The RegA region (N-terminal region of sacsin including UbL domain, the first SRR domain and the linker region preceding the second SRR domain) in human sacsin has been demonstrated to be capable of maintaining denatured firefly luciferase (FLuc, model protein-folding client in chaperone studies) in a soluble state and moderately facilitate its refolding independently from ATP, an activity characteristic of Hsp40 family chaperones (54). In addition, Reg A has been shown to be able to enhance the yield of the bacterial Hsp70 chaperone system in the absence of DnaJ (49). This observation, in addition to the presence of a J-domain in the sacsin protein, signifies that the structure of sacsin is well-suited for delivering misfolded or denatured proteins to the Hsp70 system (54). In addition, siRNA studies demonstrated that sacsin is able to reduce the number and sizes of the insoluble nuclear ataxin-1 inclusion bodies (hallmark of many cerebellar ataxia disorders) in neurons and further corroborated sacsin's potential role as a neuronal chaperone (46).

Recently, Girard *et al.* reported a transgenic sacsin-knockout (KO) mouse model that helped to elucidate several important aspects of the underlying pathogenic mechanism of ARSACS (8). The mouse model showed normal brain morphology and phenotypes (compared to control mice) at birth but displayed progressively decreasing number of Purkinje neurons from 120 days after birth, consistent with autopsy reports of significantly reduced number of Purkinje neurons in ARSACS patients (8). This observation confirmed age-dependent Purkinje neuron degeneration as one of the main pathological features of ARSACS and excluded developmental causes from the etiology of the disease. In addition, immunofluorescence experiments revealed sacsin's intracellular localization on the cytoplasmic side of mitochondria and small interference RNA (siRNA)-mediated knockdown of sacsin demonstrated remarkable changes in the morphology, function and transportation of mitochondria in sacsin-deficient neurons (8). Fluorescence recovery after photobleaching (FRAP) confirmed that mitochondria in sacsin-knockdown cells are more elongated and interconnected, indicating defective mitochondrial fission; fluorescence measurement with tetramethylrhodamine methyl ester (TMRM) showed a significantly weaker membrane potentials of these mitochondria, signifying diminished oxidative phosphorylation activity and ATP production (53). Immunofluorescence with mitochondrial marker in sacsin-knockdown neurons showed that the mitochondria in these cells fail to be delivered effectively along the dendrites but cluster within the soma and the dendritic region proximal to the soma of the neurons (8). Furthermore, the decrease in the number of dendrites as well as the thickened and disorganized morphology of the remaining dendrites in these neurons suggest that the defect in mitochondrial transport had led to degeneration of these neurons (8).

Interestingly, dynamin-related protein 1 (Drp1), a GTPase crucial for mitochondrial fission, has been found to co-immunoprecipitate and partially colocalize with sacsin in neurons (8). In light of sacsin's potential role as a molecular chaperone, Girard *et al.* suggested that sacsin most likely serves to maintain the properly-folded, functional conformation of Drp-1 either directly or indirectly by recruiting the Hsp70 family chaperones (8). Absence of sacsin from cells would therefore increase the number of misfolded, dysfunctional Drp-1 and subsequent disruption of the fission/fusion dynamics of mitochondria. The ensuing impairments in mitochondrial function and transport would in turn result in degeneration of neurons (8). The effect would be particularly detrimental to Purkinje neurons, in which the maintenance of extensively branched dendrites heavily depends the energy produced by mitochondria as well as their ability to be transported effectively along the dendrites (35). The end-result would be impairments in the cerebellum's ability to regulate and relay signals to the peripheries of the body, leading to a variety of sensory-motor symptoms such as those observed in ARSACS patients.

Disruption in the mitochondrial fission/fusion dynamics is a common theme among many neurodegenerative diseases (9). In addition, Drp1 is suggested to be implicated in many of these diseases as

well (9). Mitochondria in the neurons of Alzheimer's disease patients exhibit a spherical and fragmented morphology, which is accompanied by significant changes in the expression levels of mitochondrial fission and fusion proteins (55). This is thought to be a result of amyloid β fibril accumulation in the neurons and is hypothesized to be a Drp-1 mediated process (55). Similarly, aberrant mitochondrial fragmentation is also observed in the neurons of Parkinson's disease (PD) patients and leucine-rich repeat kinase 2 (LRRK2), the protein implicated in the most common form of PD, is hypothesized to regulate mitochondrial fission by interacting with Drp1 (56). Furthermore, a recent study attributed excessive mitochondrial fragmentation observed in neurons of Huntington's disease patients to the overstimulation of Drp-1 by huntingtin aggregates (57). Therefore, there seem to be a strong link between the pathogenic mechanisms of ARSACS and these neurodegenerative diseases, adding another level of significance to the study of ARSACS (58). In fact, it would be interesting to examine whether knocking-out or downregulating sacsin expression in neurons presented with these neurodegenerative diseases would be able to partially restore the normal mitochondrial fission/fusion dynamics.

However, disruptions in mitochondrial morphology is a common response to possible defects in a variety of mitochondrial proteins as well as the accumulation of oxidative stress (9). It is possible that sacsin affects the "health" of the mitochondria indirectly through another binding partner (or even another mechanism). Therefore, it is necessary to confirm the direct interaction between sacsin and Drp1 and its downstream cellular effects, as well as to clarify the functions of sacsin's domains and their mechanism of interaction, in order to confirm the mitochondrial model of ARSACS pathogenesis and its relation to other neurological diseases.

Conclusions

Autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS), whose primary symptoms consist of a triad of early-onset cerebellar ataxia, peripheral neuropathy and pyramidal tract signs (16), was first reported to occur in high prevalence in the Charlevoix-Saguenay population in Quebec as an evolutionary product of founder effect (5). Additional patient populations have since been identified in over thirteen countries worldwide (38). *SACS*, the gene implicated in the disease, has been mapped to the long arm of chromosome 13 (13q11) and has been found to encode the 437kDa protein sacsin, one of the largest human proteins ever identified (7). To date, over 70 ARSACS-causing missense mutations, non-sense mutations and macrodeletions have been found in patient populations identified so far and the majority of these mutations reside within the large exon 9 (6). Bioinformatic studies have identified five distinct domains within the sacsin protein, namely (from the N- to C-terminus) the UbL domain, SRR domains, XPCB domain, J-domain and HEPN domain. Furthermore, the available structural and functional information on sacsin's domains, as well as sacsin's intracellular localization pattern, suggest that sacsin most likely functions as a special class of Hsp40 chaperone that helps to regulate the dynamics of mitochondrial fission/fusion in the Purkinje neurons of cerebellum by interacting with Drp1 (8). Since disruptions in mitochondrial fission/fusion dynamics via Drp-1 mediated processeses is considered to be the underlying cause of many neurodegenerative diseases such as Alzheimer's (55), Parkinson's (56) and Huntington's diseases (57), the study of the once obscure disease ARSACS holds great importance in understanding the pathogenic mechanisms of neurodegenerative diseases and their relations to mitochondrial functions in neurons.

Acknowledgements

I would like to thank Dr. Kalle Gehring for allowing me to conduct my undergraduate research project in his laboratory and for guiding me in writing this review. Also, I would like to thank Fondation de L'Ataxie Charlevoix-Saguenay for generously providing the funding for my research project on ARSACS.

References

[1] C.M. Frederick, in *Pathophysiology of the motor systems: principles and clinical presentations,* L.K. Salain and C.M. Frederick eds. (F.A. Davis, Philadelphia, 1996).
[2] P.J. Dyck et al. *Peripheral neuropathy*. **1**, 760-870 (1984).
[3] A.G. Purves D et al, in *Neuroscience* (Sinauer Associates, Sunderland, ed. 2, 2001).
[4] J. Baets *et al. Neurology*. **13**, 1181-8 (2010).
[5] J.P. Bouchard *et al. The Canadian journal of neurological sciences.* **1**, 61-9 (1978).
[6] Y. Bouhlal et al. Hentati. *Parkinsonism Relat Disord.* **6**, 418-22 (2011).
[7] J.C. Engert *et al. Nat Genet*. **2**, 120-5 (2000).
[8] M. Girard *et al. Proc Natl Acad Sci U S A.* **5**, 1661-6 (2012).
[9] K. Itoh, K. Nakamura, M. Iijima, H. Sesaki. *Trends in cell biology.* **2**, 64-71 (2013).
[10] J.P. Bouchard et al. Can J *Neurol Sci.* 2, 185-9 (1979).
[11] J.M. Peyronnard, L. Charron, A. Barbeau. Can J *Neurol Sci.* 2, 199-203 (1979).
[12] R. Langelier, J.P. Bouchard, R. Bouchard. Can J Neurol Sci. 2, 195-8 (1979).
[13] JP Bouchard. *Hereditary neuropathies and spinocerebellar degenerations,* 451-9 (1991).
[14] A. Richter et al. *Am J Hum Genet.* 3, 768-75 (1999).
[15] S. Vermeer *et al. Neurogenetics.* 3, 207-14 (2008).
[16] J. Baets *et al. Neurology.* 13, 1181-8 (2010).
[17] N. Mrissa *et al. Neurology.* 7, 1408-14 (2000).
[18] C. Criscuolo *et al. Neurology.* 1, 100-2 (2004).
[19] G.S. Grieco *et al. Neurology.* 1, 103-6 (2004).

[20] A. Terracciano *et al. Neurogenetics. 2*, 151-5 (2009).
[21] T. Ogawa *et al. Neurology*. 1, 107-9 (2004).
[22] Y. Yamamoto *et al. J Neurol Sci.* 1, 101-4 (2005).
[23] H. Shimazaki *et al. Neurology*. 12, 2129-31 (2005).
[24] K. Hara *et al. Mov Disord*. 3, 380-2 (2005).
[25] Y. Ouyang *et al. Neurology.* 7, 1103-4 (2006).
[26] H. Shimazaki *et al. J Neurol Sci.* 1-2, 87-9 (2007).
[27] S. Miyatake *et al. Intern Med.* 16, 2221-6 (2012).
[28] A. Pyle *et al*. *Arch Neurol.* 1-4 (2012).
[29] M. Anheim *et al. Rev Neurol (Paris)*. 4, 363-8 (2008).
[30] M. Anheim *et al. Neurogenetics.* 1, 1-12 (2010).
[31] M. Anheim. *Rev Neurol (Paris)*. 5, 372-84 (2011).
[32] C. Criscuolo *et al. Mov Disord*. 10, 1358-61 (2005).
[33] A. Garcia, C. Criscuolo, G. de Michele, J. Berciano. *Muscle Nerve.* 1, 107-10 (2008).
[34] L.E. Pablo et al. *Mol Vi*s. 1871-6 (2011).
[35] W.M. *Verhoeven et al. Psychopathology*. 3, 193-9 (2012).
[36] J. Breckpot *et al. Eur J Hum Genet.* 9, 1050-4 (2008).
[37] Y. Ouyang *et al. J Neurol Sci*. 1-2, 73-6 (2008).
[38] C.M. Gomez. *Neurology*. 1, 10-1 (2004).
[39] S. Okawa *et al. J Neurol Neurosurg Psychiatry.* 2, 280-2 (2006).
[40] S. Kamada *et al. J Neurol*. 6, 803-6 (2008).
[41] G. El Euch-Fayache *et al. Arch Neurol*. 7, 982-8 (2003).
[42] Y. Bouhlal *et al. J Neurogenet*. 2, 139-48 (2008).
[43] L. Anesi, P. de Gemmis, M. Pandolfo, U. Hladnik. *J Mol Neurosci*. 3, 346-9 (2011).
[44] A.M. Richter *et al. Neurogenetics. 3*, 165-70 (2004).
[45] H.J. McMillan *et al. Muscle Nerve*. 3, 396-9 (2009).
[46] D.A. Parfitt *et al. Hum Mol Genet*. 9, 1556-65 (2009).
[47] W.H. Landschulz, P.F. Johnson, S.L. McKnight. *Science*. 4860, 1759- (1988).
[48] M. Grynberg, H. Erlandsen, *A. Godzik. Trends Biochem Sci.* 5, 224-6 (2003).
[49] J.F. Anderson, E. Siller, J.M. Barral. J *Mol Biol* 4, 665-74 (2010).
[50] M. Kamionka, J. Feigon. *Protein Sci*. 9, 2370-7 (2004).
[51] T.D. Mueller, J. Feigon. *Journal of molecular biology.* 5, 1243-55 (2002).
[52] X.B. Qiu, Y.M. Shao, S. Miao, L. Wang. *Cell Mol Life Sci*. 22, 2560-70 (2006).
[53] G. Kozlov *et al. J Biol Chem*. 23, 20407-12 (2011).
[54] J.F. Anderson, E. Siller, J.M. Barral. *J Mol Biol*. 4, 870-80 (2011).
[55] X. Wang *et al. The Journal of Neuroscience.* 28, 9090-103 (2009).
[56] X. Wang *et al. Human Molecular Genetics*. 9, 1931-44 (2012).
[57] W. Song *et al. Nature medicine*. 3, 377-82 (2011).
[58] H. Chen, D.C. Chan. *Human Molecular Genetics.* R2, R169-R76 (2009).