Summary of education and work experience

2014-2017: Post-doc in machine learning applied to epigenomics with Guillaume Bourque at McGill University, Montreal, Canada.
2013: Post-doc in machine learning with Masashi Sugiyama at Tokyo Insitute of Technology, Japan.
2012: PhD in mathematics from Ecole Normale Supérieure de Cachan, France. Studied machine learning applied to cancer genomics with Jean-Philippe Vert (Mines ParisTech, Institut Curie, INSERM) and Francis Bach (ENS, INRIA).
2009: Masters of statistics, Université Paris 6. Internship in Bayesian statistics with Mathieu Gautier and Jean-Louis Foulley, INRA Jouy-en-Josas, France.
2007-2008: Programmer and biochemist at Sangamo BioSciences, Richmond, California.
2006: Bachelor in molecular and cell biology and statistics, honors thesis with Terry Speed, UC Berkeley.

Google Scholar profile: http://scholar.google.ca/citations?user=c83d8tgAAAAJ
Email: toby.hocking@r-project.org
Web: https://github.com/tdhock
Born: 17 March 1984
Nationality: USA
Language skills: English (native speaker), French (fluent).

Reproducible research statement

I would like to emphasize that every one of my first-authored papers and posters is reproducible (whether published in a conference, journal, or pre-print archive). By “reproducible” I mean that I have also published source code and data so that it is possible for anyone to reproduce the computations (including all figures and tables). I believe it is important for the international research community to move toward using the internet to foster a more open and transparent collaboration and publishing model. I do this in practice by (1) publishing algorithms and data sets in R/Python packages, and (2) maintaining a list of links to the public source code of all of my papers on https://github.com/tdhock/reproducible-papers. In the future I will encourage any students that I mentor to do the same.

Awards

- Grant to travel to France with a “Mobilité Entrant” scholarship, one month studying optimal segmentation algorithms for ChIP-seq data with Guillem Rigaill at University of Evry (2016).
- Hocking TD. Adding direct labels to plots, won the Best Student Poster award at useR! 2011 international R conference in Warwick, England.

Invited conference tutorials

- **Hocking TD**, Killick R. Introduction to optimal changepoint detection algorithms, useR 2017, Brussels, Belgium.

- **Hocking TD**, Ekstrøm CT. Understanding and creating interactive graphics, useR 2016, Stanford, CA, USA.

Peer-reviewed publications

Note that my peer-reviewed publications occur in several distinct fields of research, including machine learning, in which results are published in highly competitive conferences such as ICML and NIPS, which accept only about 20% of submitted papers. For papers in which I am not the first author, I have written the details of my contribution.

- Shimada K, Shimada S, Sugimoto K, Nakatochi M, Suguro M, Hirakawa A, **Hocking TD**, Takeuchi I, Tokunaga T, Takagi Y, Sakamoto A, Aoki T, Naoe T, Nakamura S, Hayakawa F, Seto M, Tomita A, Kiyoi H. Development and analysis of patient-derived xenograft mouse models in intravascular large B-cell lymphoma. Leukemia 2016. **My contribution:** copy number data analysis using my supervised machine learning system (SegAnnDB).
- Chicard M, … **Hocking TD**, … Schleiermacher G. Genomic copy number profiling using circulating free tumor DNA highlights heterogeneity in neuroblastoma. Clinical Cancer Research 2016. **My contribution:** copy number data analysis using my supervised machine learning system (SegAnnDB).
- Maidstone R, **Hocking TD**, Rigaill G, Fearnhead P. On optimal multiple changepoint algorithms for large data. Statistics and Computing (2016). doi:10.1007/s11222-016-9636-3 **My contribution:** I created the code, figures, and text for the Results section (timings and accuracy on real data sets).
- **Hocking TD**, Goerner-Potvin P, Morin A, Shao X, Pastinen T, Bourque G. Optimizing ChIP-seq peak detectors using visual labels and supervised machine learning. Bioinformatics 2016 (in press). DOI:10.1093/bioinformatics/btw672
- **Hocking TD**, Rigaill G, Bourque G. PeakSeg: constrained optimal segmentation and supervised penalty learning for peak detection in count data. ICML 2015.
- Suguro M, Yoshida N, Umino A, Kato H, Tagawa H, Nakagawa M, Fukuhara N, Karnan S, Takeuchi I, **Hocking TD**, Arita K, Karube K, Tsuzuki S, Nakamura S, Kinoshita T, Seto M. Clonal heterogeneity of lymphoid malignancies correlates with poor prognosis. Cancer Sci. 2014 Jul;105(7):897-904. **My contribution:** copy number data analysis using my supervised machine learning system (SegAnnDB).
- **Hocking TD** et al. SegAnnDB: interactive Web-based genomic segmentation. Bioinformatics (2014) 30 (11): 1539-1546. DOI:10.1093/bioinformatics/btu072
- **Hocking TD**, Wutzler T, Ponting K and Grosjean P. Sustainable, extensible documentation generation using inlinedocs. Journal of Statistical Software (2013), 54(6), 1-20. DOI:10.18637/jss.v054.i06
- **Hocking TD**, Schleiermacher G, Janoueix-Lerosey I, Boeva V, Cappo J, Delattre O, Bach F, Vert J-P. Learning smoothing models of copy number profiles using breakpoint annotations. BMC Bioinfo. 2013, 14:164. DOI:10.1186/1471-2105-14-164
- **Hocking TD**,* Rigaill G,* Bach F, Vert J-P. Learning sparse penalties for change-point detection using max-margin interval regression. ICML 2013. *joint first authorship with Guillem Rigaill.
- **Hocking TD**, Joulin A, Bach F, Vert J-P. Clusterpath: an Algorithm for Clustering using Convex Fusion Penalties. ICML 2011.

- Gautier M, **Hocking TD**, Foulley JL. A Bayesian outlier criterion to detect SNPs under selection in large data sets. PloS ONE 5 (8), e11913 (2010). **My contribution:** I wrote the code for some simulations, and I helped to revise the text.
- Doyon Y, McCammon JM, Miller JC, Faraji F, Ngo C, Katibah GE, Amora R, **Hocking TD**, Zhang L, Rebar EJ, Gregory PD, Urnov FD, Amacher SL. Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. Nature biotechnology 26 (6), 702-70 (2008). **My contribution:** I did the experiments that characterized the DNA-binding specificity of the ZFNs used for the knock-outs, described in Supplementary Figure 11.

Pre-prints (not peer reviewed)

- **Hocking TD**. A breakpoint detection error function for segmentation model selection and validation. Preprint arXiv:1509.00368.
- **Hocking TD**, Bourque G. PeakSegJoint: fast supervised peak detection via joint segmentation of multiple count data samples. Preprint arXiv:1506.01286.
- **Hocking TD**, Spanurattana S, Sugiyama M. Support vector comparison machines. Preprint arXiv:1401.8008.
- **Hocking TD**, Rigaill G. SegAnnot: an R package for fast segmentation of annotated piecewise constant signals, Preprint hal-00759129.

Conference posters/talks (not peer reviewed)

- Narahara M, **Hocking TD**, Bourque G, Yamada R, Setoh K, Matsuda F, Lathrop M. Transcriptomic analysis of antibody responses to seasonal influenza vaccine reveals predictive gene signatures and potential key transcription factors. Accepted for oral presentation at the Human Genome Meeting, 5 February 2017, Barcelona, Spain (Maiko Narahara gave the talk). **My contribution**: writing code for the L1-regularized logistic regression model analysis.
- **Hocking TD** et al. Optimizing ChIP-seq peak detectors using visual labels and supervised machine learning, Canadian Epigenomics 2016 meeting at Esterel, Quebec, Canada.
- **Hocking TD**. Reproducible research works_with_R, useR 2016 conference, Stanford, CA, USA.
- **Hocking TD** and Bourque G. A supervised machine learning approach for joint peak detection in any number of samples and cell types. Poster for Epigenomics 2016, Rio Grande, Puerto Rico.
- VanderPlas S, Sievert C, and **Hocking TD**. Animint: Interactive Web-Based Animations Using Ggplot2's Grammar of Graphics. Invited session at Joint Statistical Meetings 2015, Seattle (Susan VanderPlas gave the talk). **My contribution**: initial idea, writing code.
- **Hocking TD**. Supervised, interactive genomic data analysis. Invited talk at McGill Barbados epigenomics meeting in Jan 2015.
- **Hocking TD**, Supaporn S, Sugiyama M. Support vector comparison machines. IBISML, Tokyo Tech, Nov 2013.
- **Hocking TD**. Learning to rank and compare graph layouts. MLAB Sapporo, Aug 2013.
- **Hocking TD**. A map of genomic copy number alterations in neuroblastoma based on annotation-guided breakpoint detection. Poster at Advances in Neuroblastoma Research 2012, Toronto, Canada.

- **Hocking TD**, Bach F, Vert JP. Supervised interactive breakpoint detection, MLSS Bordeaux Sept 2011.
- **Hocking TD**. Fast, named capture regular expressions in R-2.14. Lightning talk at international useR conference, Warwick, England, 2011.
- **Hocking TD**, Bach F, Vert JP. Clustering using convex fusion penalties, StatMathAppli Fréjus Aug 2010.
- **Hocking TD**. Sublogo dendrograms: visualizing correlation in biological sequence motifs. International useR conference, Rennes, France, 2009.

Free/open-source software packages

My most significant contributions (as of Dec 2016) include **directlabels** which won the Best Student Poster award at the useR 2011 conference (also several CRAN packages depend on it), **clusterpath** which has over 70 citations, **animint** which was presented at the useR 2016 conference tutorial on "understanding and creating interactive graphics," and **plotly** which has over 800 stars on GitHub. Also, I have contributed a patch https://bugs.r-project.org/bugzilla/show_bug.cgi?id=14518 that implemented named capture regular expression support in R (this code has been included with every copy of R since R-2.14 in 2011).

I am a primary maintainer of the following free/open-source software packages that have resulted from my research and collaborations. (*except the plotly package for which I designed and implemented the original ggplot interface but I am no longer a maintainer)

Package	Description	C/C++	Python	R	JavaScript
sublogo	visualize DNA sequence motifs			R	
inlinedocs	generate R package docs			R	
directlabels	text labels for multicolor plots			R	
clusterpath	convex clustering	C++		R	
quadmod	quadratic program model language			R	
bams	breakpoint detection algorithms			R	
neuroblastoma	breakpoint detection data set			R	
breakpointError	true/false positive breakpoints	C		R	
SegAnnot	optimal supervised segmentation	C	Python	R	
SegAnnDB	interactive genome segmentation	C/C++	Python		JavaScript
rankSVMcompare	SVM for ranking and comparing			R	
gganim	convert ggplots to animations			R	
animint	animated, interactive data viz			R	JavaScript
requireGitHub	reproducible research			R	
WeightedROC	ROC/AUC with weights			R	
plotly*	web data viz and sharing			R	
revector	vectors of regular expressions	C		R	
PeakError	annotation error of peak calls	C		R	
PeakSegDP	single-sample supervised peaks	C		R	
PeakSegJoint	multi-sample supervised peaks	C		R	
memtime	memory and time measurement			R	
namedCapture	regular expressions			R	
str.extractall	regular expressions in pandas		Python		
coseg	Constrained optimal segmentation	C++		R	
penaltyLearning	Learning penalty functions	C++		R	
fpop	Functional pruning optimal partitioning	C++		R	

Academic community volunteer work

- I am a reviewer for any academic journal who asks me to review a paper in my domains of expertise: machine learning, statistics, bioinformatics, and statistical software. Usually I write about 5-10 journal paper reviews per year, since completing my PhD in 2012.
- I am a reviewer for the two major machine learning conferences, ICML and NIPS. This means reviewing from 4 to 8 papers twice per year.
- Since Jan 2015 I am the lead organizer of the Montreal R User Group, which every month brings together a diverse group of people from academia and industry to discuss data analysis, visualization, and R programming.

I am a co-administrator for The R Project's participation in the Google Summer of Code, since 2012. Each year this program gives students from all over the world the chance to learn how to develop R packages, supervised by mentors who are experts in R programming. I have also mentored several students on the following projects.

- Animated interactive ggplots (animint package), Susan VanderPlas 2013, Carson Sievert 2014, Tony Tsai 2015, Kevin Ferris 2015, Faizan Khan 2016.
- Stochastic Average Gradient algorithm for computing L2-regularized linear models (bigoptim package), Ishmael Belghazi 2015.
- Coordinate descent algorithm for computing elastic net regularized interval regression linear models (iregnet package), Anuj Khare 2016.
- R interface to Google's RE2 C++ regular expression code (re2r package), Qin Wenfeng 2016.
- Performance testing R packages (Rperform package), Akash Tandon 2015-2016.