1

THE EFFECTS OF ALIGNMENT ERROR AND ALIGNMENT FILTERING ON THE SITEWISE DETECTION OF POSITIVE SELECTION

1.1 INTRODUCTION

1.1.1 *Methods for detecting sitewise positive selection*

1.1.2 *Substitution and indel processes in simulating protein-coding sequence evolution*

1.2 MODELS AND PARAMETERS FOR SIMULATING THE EVOLUTION OF MAMMALIAN GENES

1.2.1 *Distribution of selective pressures*

1.2.2 *Phylogenetic tree size and shape*

1.2.3 *Frequency and size distribution of insertions and deletions*

1.3 ANALYSIS OF THE ALIGNMENT ERROR SIMULATION RESULTS

1.4 METHODS FOR FILTERING ALIGNMENTS

1.5 ANALYSIS OF THE ALIGNMENT FILTERING SIMULATION RESULTS

2

THE EFFECTS OF ALIGNMENT ERROR AND ALIGNMENT FILTERING ON DETECTING POSITIVE SELECTION IN GENES

2.1 INTRODUCTION

2.1.1 *Existing methods for detecting gene-wide positive selection*

2.2 THE APPLICATION OF SITEWISE ESTIMATES TO THE GENEWISE DETECTION OF POSITIVE SELECTION

2.3 ANALYSIS OF THE GENEWISE DETECTION RESULTS

2.4 CONCLUSIONS AND FURTHER WORK

3

PATTERNS OF SITEWISE SELECTION IN MAMMALIAN PROTEIN-CODING GENES

3.1 INTRODUCTION

3.1.1 *The Mammalian Genome Project*

A major goal of mammalian comparative genomics has been to quantify, identify and understand the fraction of the human genome that is under evolutionary constraint. The first non-human mammalian genomes showed at least 5% of the human genome to be under purifying selection [3, 1, 2], but the small number of genomes available limited the extent to which regions of evolutionary constraint could be identified. The Mammalian Genome Project, a coordinated set of genome sequencing projects organised by the Broad Institute of MIT and Harvard, was designed with the primary purpose of increasing the accuracy and confidence with which regions of the human genome that have evolved under evolutionary constraint in mammals could be identified [**?**].

The mammalian tree of like has a star-like shape, owing to the rapid and extensive radiation of mammalian species that occurred starting *X*YZ mya.

3.1.2 *The Sitewise Likelihood Ratio test*

3.1.3 *Data quality concerns: alignment and sequencing error*

3.1.4 *Gene trees, genomic alignments, and low-coverage genomes in the Ensembl database*

3.2 METHODS TO IDENTIFY ORTHOLOGOUS SUBTREES WITHIN LARGE MAMMALIAN GENE FAMILIES

3.3 ANALYSIS OF THE GENOME-WIDE SET OF ORTHOLOGOUS MAMMALIAN TREES

3.4 ANALYSIS OF THE GLOBAL DISTRIBUTION OF MAMMALIAN SELECTIVE PRESSURES

3.5 ANALYSIS OF SITEWISE ESTIMATES FROM THREE MAMMALIAN SUB-CLADES

3.6 EVALUATION OF THE EFFECT OF GC CONTENT, RECOMBINATION RATE, AND CODON USAGE ON SITEWISE DNDS ESTIMATES AND THE DETECTION OF POSITIVE SELECTION

3.6.1 *Mammalian sitewise selective pressures are not subject to strong effects of biased gene conversion*

3.6.2 *Mammalian sitewise selective pressures suggest increased efficacy of natural selection in regions of high recombination*

3.7 CONCLUSIONS AND FUTURE WORK

4

THE USE OF SITEWISE SELECTIVE PRESSURES TO CHARACTERISE THE EVOLUTION OF GENES AND DOMAINS IN MAMMALS

4.1 COMPARISON OF SITEWISE RESULTS TO PREVIOUSLY DESCRIBED SETS OF POSITIVELY SELECTED GENES

4.2 USING SITEWISE SELECTIVE PRESSURES TO CHARACTERISE THE EVOLUTION OF GENES

4.2.1 *Identifying genes subject to positive selection*

4.2.2 *Identifying genes subject to strong or weak purifying selection*

4.3 USING SITEWISE SELECTIVE PRESSURES TO CHARACTERISE THE EVOLUTION OF PROTEIN DOMAINS

4.3.1 *Identifying protein domains subject to positive selection*

4.3.2 *Identifying protein domains subject to strong or weak purifying selection*

5

THE EVOLUTION OF PROTEIN-CODING GENES IN GORILLA AND THE AFRICAN APES

5.1 INTRODUCTION

5.1.1 *The gorilla and other primate genome projects*

5.1.2 *Incomplete lineage sorting*

5.1.3 *Effective population sizes of extant and ancestral primate populations*

5.1.4 *Measuring shifts in selective pressures using branch-specific likelihood ratio tests*

5.1.5 *Data quality concerns: sequencing, assembly and alignment error*

5.2 CONSTRUCTING CODON ALIGNMENTS OF ONE-TO-ONE ORTHOLOGOUS GENES IN SIX PRIMATE SPECIES

5.2.1 *Identification of genes with one-to-one homology*

5.2.2 *Collection of homologous DNA sequences from genome- or transcript-based multiple alignments*

5.2.3 *Filtering sequence regions with low sequence quality*

5.2.4 *Filtering sequence regions with high substitution counts*

5.2.5 *Filtering sequence regions with evidence of incomplete lineage sorting*

5.3 ANALYSIS OF PATTERNS OF DUPLICATION AND DELETION IN PRIMATE GENE FAMILIES

5.4 ANALYSIS OF THE LIKELIHOOD RATIO TEST RESULTS

5.4.1 *Genes with evidence for acceleration and deceleration in the human, chimpanzee and gorilla terminal lineages*

5.4.2 *Genes with evidence for acceleration in the African great ape ancestral branch*

5.4.3 *Genes with evidence for positive selection based on the branch-site test*

6

GORILLA PART 2

6.1 ANALYSIS OF INCOMPLETE LINEAGE SORTING IN THE AFRICAN GREAT APES WITHIN AND NEARBY PROTEIN-CODING GENES

6.2 ANALYSIS OF DN/DS LEVELS IN SIX PRIMATE GENOMES

6.2.1 *Genome-wide dN/dS in six primates and their ancestors*

6.2.2 *Genome-wide dN/dS in regions of differing sitewise constraint*

6.2.3 *Analysis of the impact of sequence and alignment filtering on primate dN/dS estimates*

6.3 CONCLUSIONS AND FUTURE WORK

BIBLIOGRAPHY

[1] RA Gibbs, GM Weinstock, ML Metzker, DM Muzny, EJ Sodergren, S Scherer, G Scott, D Steffen, KC Worley, PE Burch, G Okwuonu, S Hines, L Lewis, C DeRamo, O Delgado, S Dugan-Rocha, G Miner, M Morgan, A Hawes, R Gill, Celera, RA Holt, MD Adams, PG Amanatides, H Baden-Tillson, M Barnstead, S Chin, CA Evans, S Ferriera, C Fosler, A Glodek, Z Gu, D Jennings, CL Kraft, T Nguyen, CM Pfannkoch, C Sitter, GG Sutton, JC Venter, T Woodage, D Smith, HM Lee, E Gustafson, P Cahill, A Kana, L Doucette-Stamm, K Weinstock, K Fechtel, RB Weiss, DM Dunn, ED Green, RW Blakesley, GG Bouffard, PJ De Jong, K Osoegawa, B Zhu, M Marra, J Schein, I Bosdet, C Fjell, S Jones, M Krzywinski, C Mathewson, A Siddiqui, N Wye, J McPherson, S Zhao, CM Fraser, J Shetty, S Shatsman, K Geer, Y Chen, S Abramzon, WC Nierman, PH Havlak, R Chen, KJ Durbin, A Egan, Y Ren, XZ Song, B Li, Y Liu, X Qin, S Cawley, KC Worley, AJ Cooney, LM D'Souza, K Martin, JQ Wu, ML Gonzalez-Garay, AR Jackson, KJ Kalafus, MP McLeod, A Milosavljevic, D Virk, A Volkov, DA Wheeler, Z Zhang, JA Bailey, EE Eichler, E Tuzun, E Birney, E Mongin, A Ureta-Vidal, C Woodwark, E Zdobnov, P Bork, M Suyama, D Torrents, M Alexandersson, BJ Trask, JM Young, H Huang, H Wang, H Xing, S Daniels, D Gietzen, J Schmidt, K Stevens, U Vitt, J Wingrove, F Camara, M Mar Albà, JF Abril, R Guigo, A Smit, I Dubchak, EM Rubin, O Couronne, A Poliakov, N Hübner, D Ganten, C Goesele, O Hummel, T Kreitler, YA Lee, J Monti, H Schulz, H Zimdahl, H Himmelbauer, H Lehrach, HJ Jacob, S Bromberg, J Gullings-Handley, MI Jensen-Seaman, AE Kwitek, J Lazar, D Pasko, PJ Tonellato, S Twigger, CP Ponting, JM Duarte, S Rice, L Goodstadt, SA Beatson, RD Emes, EE Winter, C Webber, P Brandt, G Nyakatura, M Adetobi, F Chiaromonte, L Elnitski, P Eswara, RC Hardison, M Hou, D Kolbe, K Makova, W Miller, A Nekrutenko, C Riemer, S Schwartz, J Taylor, S Yang, Y Zhang, K Lindpaintner, TD Andrews, M Caccamo, M Clamp, L Clarke, V Curwen, R Durbin, E Eyras, SM Searle, GM Cooper, S Batzoglou, M Brudno, A Sidow, EA Stone, JC Venter, BA Payseur, G Bourque, C López-Otín, XS Puente, K Chakrabarti, S Chatterji, C Dewey, L Pachter, N Bray, VB Yap, A Caspi, G Tesler, PA Pevzner, D Haussler, KM Roskin, R Baertsch, H Clawson, TS Furey, AS Hinrichs, D Karolchik, WJ Kent, KR Rosenbloom, H Trumbower, M Weirauch, DN Cooper, PD Stenson, B Ma, M Brent, M Arumugam, D Shteynberg, RR Copley, MS Taylor, H Riethman, U Mudunuri, J Peterson, M Guyer, A Felsenfeld,

S Old, S Mockrin, F Collins, and Rat Genome Sequencing Project Consortium. Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428:493–521, Apr 2004. doi: 10.1038/nature02426.

[2] Kerstin Lindblad-Toh, Claire M Wade, Tarjei S. Mikkelsen, Elinor K. Karlsson, David B. Jaffe, Michael Kamal, Michele Clamp, Jean L. Chang, Edward J. Kulbokas, Michael C. Zody, Evan Mauceli, Xiaohui Xie, Matthew Breen, Robert K. Wayne, Elaine A. Ostrander, Chris P. Ponting, Francis Galibert, Douglas R. Smith, Pieter J. deJong, Ewen Kirkness, Pablo Alvarez, Tara Biagi, William Brockman, Jonathan Butler, Chee-Wye Chin, April Cook, James Cuff, Mark J. Daly, David DeCaprio, Sante Gnerre, Manfred Grabherr, Manolis Kellis, Michael Kleber, Carolyne Bardeleben, Leo Goodstadt, Andreas Heger, Christophe Hitte, Lisa Kim, Klaus-Peter Koepfli, Heidi G. Parker, John P. Pollinger, Stephen M. J. Searle, Nathan B. Sutter, Rachael Thomas, Caleb Webber, Jennifer Baldwin, Adal Abebe, Amr Abouelleil, Lynne Aftuck, Mostafa Ait-zahra, Tyler Aldredge, Nicole Allen, Peter An, Scott Anderson, Claudel Antoine, Harindra Arachchi, Ali Aslam, Laura Ayotte, Pasang Bachantsang, Andrew Barry, Tashi Bayul, Mostafa Benamara, Aaron Berlin, Daniel Bessette, Berta Blitshteyn, Toby Bloom, Jason Blye, Leonid Boguslavskiy, Claude Bonnet, Boris Boukhgalter, Adam Brown, Patrick Cahill, Nadia Calixte, Jody Camarata, Yama Cheshatsang, Jeffrey Chu, Mieke Citroen, Alville Collymore, Patrick Cooke, Tenzin Dawoe, Riza Daza, Karin Decktor, Stuart DeGray, Norbu Dhargay, Kimberly Dooley, Kathleen Dooley, Passang Dorje, Kunsang Dorjee, Lester Dorris, Noah Duffey, Alan Dupes, Osebhajajeme Egbiremolen, Richard Elong, Jill Falk, Abderrahim Farina, Susan Faro, Diallo Ferguson, Patricia Ferreira, Sheila Fisher, Mike FitzGerald, Karen Foley, Chelsea Foley, Alicia Franke, Dennis Friedrich, Diane Gage, Manuel Garber, Gary Gearin, Georgia Giannoukos, Tina Goode, Audra Goyette, Joseph Graham, Edward Grandbois, Kunsang Gyaltsen, Nabil Hafez, Daniel Hagopian, Birhane Hagos, Jennifer Hall, Claire Healy, Ryan Hegarty, Tracey Honan, Andrea Horn, Nathan Houde, Leanne Hughes, Leigh Hunnicutt, M. Husby, Benjamin Jester, Charlien Jones, Asha Kamat, Ben Kanga, Cristyn Kells, Dmitry Khazanovich, Alix Chinh Kieu, Peter Kisner, Mayank Kumar, Krista Lance, Thomas Landers, Marcia Lara, William Lee, Jean-Pierre Leger, Niall Lennon, Lisa Leuper, Sarah LeVine, Jinlei Liu, Xiaohong Liu, Yeshi Lokyitsang, Tashi Lokyitsang, Annie Lui, Jan Macdonald, John Major, Richard Marabella, Kebede Maru, Charles Matthews, Susan McDonough, Teena Mehta, James Meldrim, Alexandre Melnikov, Louis Meneus, Atanas Mihalev, Tanya Mihova, Karen Miller, Rachel Mittelman, Valentine Mlenga, Leonidas Mulrain, Glen Munson, Adam Navidi, Jerome Naylor, Tuyen Nguyen, Nga Nguyen, Cindy Nguyen, Thu Nguyen, Robert

Nicol, Nyima Norbu, Choe Norbu, Nathaniel Novod, Tenchoe Nyima, Peter Olandt, Barry O'Neill, Keith O'Neill, Sahal Osman, Lucien Oyono, Christopher Patti, Danielle Perrin, Pema Phunkhang, Fritz Pierre, Margaret Priest, Anthony Rachupka, Sujaa Raghuraman, Rayale Rameau, Verneda Ray, Christina Raymond, Filip Rege, Cecil Rise, Julie Rogers, Peter Rogov, Julie Sahalie, Sampath Settipalli, Theodore Sharpe, Terrance Shea, Mechele Sheehan, Ngawang Sherpa, Jianying Shi, Diana Shih, Jessie Sloan, Cherylyn Smith, Todd Sparrow, John Stalker, Nicole Stange-Thomann, Sharon Stavropoulos, Catherine Stone, Sabrina Stone, Sean Sykes, Pierre Tchuinga, Pema Tenzing, Senait Tesfaye, Dawa Thoulutsang, Yama Thoulutsang, Kerri Topham, Ira Topping, Tsamla Tsamla, Helen Vassiliev, Vijay Venkataraman, Andy Vo, Tsering Wangchuk, Tsering Wangdi, Michael Weiand, Jane Wilkinson, Adam Wilson, Shailendra Yadav, Shuli Yang, Xiaoping Yang, Geneva Young, Qing Yu, Joanne Zainoun, Lisa Zembek, Andrew Zimmer, and Eric S. Lander. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438:803–819, 12 2005. doi: 10.1038/nature04338.

[3] Mouse Genome Sequencing Consortium, RH Waterston, K Lindblad-Toh, E Birney, J Rogers, JF Abril, P Agarwal, R Agarwala, R Ainscough, M Alexandersson, P An, SE Antonarakis, J Attwood, R Baertsch, J Bailey, K Barlow, S Beck, E Berry, B Birren, T Bloom, P Bork, M Botcherby, N Bray, MR Brent, DG Brown, SD Brown, C Bult, J Burton, J Butler, RD Campbell, P Carninci, S Cawley, F Chiaromonte, AT Chinwalla, DM Church, M Clamp, C Clee, FS Collins, LL Cook, RR Copley, A Coulson, O Couronne, J Cuff, V Curwen, T Cutts, M Daly, R David, J Davies, KD Delehaunty, J Deri, ET Dermitzakis, C Dewey, NJ Dickens, M Diekhans, S Dodge, I Dubchak, DM Dunn, SR Eddy, L Elnitski, RD Emes, P Eswara, E Eyras, A Felsenfeld, GA Fewell, P Flicek, K Foley, WN Frankel, LA Fulton, RS Fulton, TS Furey, D Gage, RA Gibbs, G Glusman, S Gnerre, N Goldman, L Goodstadt, D Grafham, TA Graves, ED Green, S Gregory, R Guigó, M Guyer, RC Hardison, D Haussler, Y Hayashizaki, LW Hillier, A Hinrichs, W Hlavina, T Holzer, F Hsu, A Hua, T Hubbard, A Hunt, I Jackson, DB Jaffe, LS Johnson, M Jones, TA Jones, A Joy, M Kamal, EK Karlsson, D Karolchik, A Kasprzyk, J Kawai, E Keibler, C Kells, WJ Kent, A Kirby, DL Kolbe, I Korf, RS Kucherlapati, EJ Kulbokas, D Kulp, T Landers, JP Leger, S Leonard, I Letunic, R Levine, J Li, M Li, C Lloyd, S Lucas, B Ma, DR Maglott, ER Mardis, L Matthews, E Mauceli, JH Mayer, M McCarthy, WR McCombie, S McLaren, K McLay, JD McPherson, J Meldrim, B Meredith, JP Mesirov, W Miller, TL Miner, E Mongin, KT Montgomery, M Morgan, R Mott, JC Mullikin, DM Muzny, WE Nash, JO Nelson, MN Nhan, R Nicol, Z Ning, C Nusbaum, MJ O'Connor,

Y Okazaki, K Oliver, E Overton-Larty, L Pachter, G Parra, KH Pepin, J Peterson, P Pevzner, R Plumb, CS Pohl, A Poliakov, TC Ponce, CP Ponting, S Potter, M Quail, A Reymond, BA Roe, KM Roskin, EM Rubin, AG Rust, R Santos, V Sapojnikov, B Schultz, J Schultz, MS Schwartz, S Schwartz, C Scott, S Seaman, S Searle, T Sharpe, A Sheridan, R Shownkeen, S Sims, JB Singer, G Slater, A Smit, DR Smith, B Spencer, A Stabenau, N Stange-Thomann, C Sugnet, M Suyama, G Tesler, J Thompson, D Torrents, E Trevaskis, J Tromp, C Ucla, A Ureta-Vidal, JP Vinson, AC Von Niederhausern, CM Wade, M Wall, RJ Weber, RB Weiss, MC Wendl, AP West, K Wetterstrand, R Wheeler, S Whelan, J Wierzbowski, D Willey, S Williams, RK Wilson, E Winter, KC Worley, D Wyman, S Yang, SP Yang, EM Zdobnov, MC Zody, and ES Lander. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420: 520–62, Dec 2002. doi: 10.1038/nature01262.