Functional Big-Data Genomics

Ashish Agarwal, Sebastien Mondet,
Paul Scheid, Aviv Madar, Richard Bonneau, Jane Carlton, Kris Gunsalus

Center for Genomics and Systems Biology
Department of Biology
New York University

Commercial Users of Functional Programming
Copenhagen, Denmark
Sep 15, 2012


Science
THE
HUMAN
GENOME

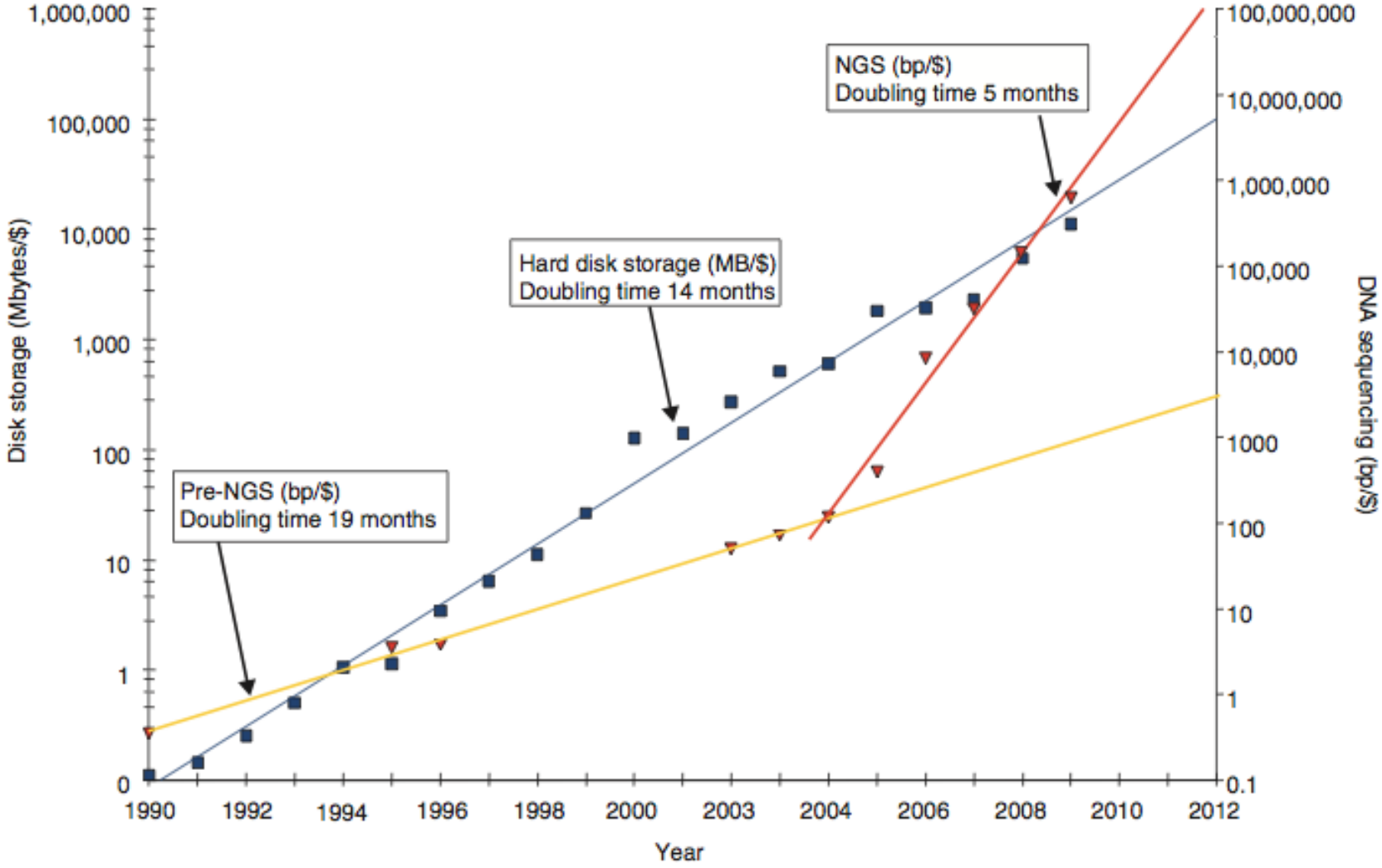

nature
the
human
genome

J. Craig Venter[1,*], Mark D. Adams[1], Eugene W. Myers[1], Peter W. Li[1], Richard J. Mural[1], Granger G. Sutton[1], Hamilton O. Smith[1], Mark Yandell[1], Cheryl A. Evans[1], Robert A. Holt[1], Jeannine D. Gocayne[1], Peter Amanatides[1], Richard M. Ballew[1], Daniel H. Huson[1], Jennifer Russo Wortman[1], Qing Zhang[1], Chinnappa D. Kodira[1], Xiangqun H. Zheng[1], Lin Chen[1], Marian Skupski[1], Gangadharan Subramanian[1], Paul D. Thomas[1], Jinghui Zhang[1], George L. Gabor Miklos[2], Catherine Nelson[3], Samuel Broder[1], Andrew G. Clark[4], Joe Nadeau[5], Victor A. McKusick[6], Norton Zinder[7], Arnold J. Levine[7], Richard J. Roberts[8], Mel Simon[9], Carolyn Slayman[10], Michael Hunkapiller[11], Randall Bolanos[1], Arthur Delcher[1], Ian Dew[1], Daniel Fasulo[1], Michael Flanigan[1], Liliana Florea[1], Aaron Halpern[1], Sridhar Hannenhalli[1], Saul Kravitz[1], Samuel Levy[1], Clark Mobarry[1], Knut Reinert[1], Karin Remington[1], Jane Abu-Threideh[1], Ellen Beasley[1], Kendra Biddick[1], Vivien Bonazzi[1], Rhonda Brandon[1], Michele Cargill[1], Ishwar Chandramouliswaran[1], Rosane Charlab[1], Kabir Chaturvedi[1], Zuoming Deng[1], Valentina Di Francesco[1], Patrick Dunn[1], Karen Eilbeck[1], Carlos Evangelista[1], Andrei E. Gabrielian[1], Weiniu Gan[1], Wangmao Ge[1], Fangcheng Gong[1], Zhiping Gu[1], Ping Guan[1], Thomas J. Heiman[1], Maureen E. Higgins[1], Rui-Ru Ji[1], Zhaoxi Ke[1], Karen A. Ketchum[1], Zhongwu Lai[1], Yiding Lei[1], Zhenya Li[1], Jiayin Li[1], Yong Liang[1], Xiaoying Lin[1], Fu Lu[1], Gennady V. Merkulov[1], Natalia Milshina[1], Helen M. Moore[1], Ashwinikumar K Naik[1], Vaibhav A. Narayan[1], Beena Neelam[1], Deborah Nusskern[1], Douglas B. Rusch[1], Steven Salzberg[12], Wei Shao[1], Bixiong Shue[1], Jingtao Sun[1], Zhen Yuan Wang[1], Aihui Wang[1], Xin Wang[1], Jian Wang[1], Ming-Hui Wei[1], Ron Wides[13], Chunlin Xiao[1], Chunhua Yan[1], Alison Yao[1], Jane Ye[1], Ming Zhan[1], Weiqing Zhang[1], Hongyu Zhang[1], Qi Zhao[1], Liansheng Zheng[1], Fei Zhong[1], Wenyan Zhong[1], Shiaoping C. Zhu[1], Shaying Zhao[12], Dennis Gilbert[1], Suzanna Baumhueter[1], Gene Spier[1], Christine Carter[1], Anibal Cravchik[1], Trevor Woodage[1], Feroze Ali[1], Huijin An[1], Aderonke Awe[1], Danita Baldwin[1], Holly Baden[1], Mary Barnstead[1], Ian Barrow[1], Karen Beeson[1], Dana Busam[1], Amy Carver[1], Angela Center[1], Ming Lai Cheng[1], Liz Curry[1], Steve Danaher[1], Lionel Davenport[1], Raymond Desilets[1], Susanne Dietz[1], Kristina Dodson[1], Lisa Doup[1], Steven Ferriera[1], Neha Garg[1], Andres Gluecksmann[1], Brit Hart[1], Jason Haynes[1], Charles Haynes[1], Cheryl Heiner[1], Suzanne Hladun[1], Damon Hostin[1], Jarrett Houck[1], Timothy Howland[1], Chinyere Ibegwam[1], Jeffery Johnson[1], Francis Kalush[1], Lesley Kline[1], Shashi Koduru[1], Amy Love[1], Felecia Mann[1], David May[1], Steven McCawley[1], Tina McIntosh[1], Ivy McMullen[1], Mee Moy[1], Linda Moy[1], Brian Murphy[1], Keith Nelson[1], Cynthia Pfannkoch[1], Eric Pratts[1], Vinita Puri[1], Hina Qureshi[1], Matthew Reardon[1], Robert Rodriguez[1], Yu-Hui Rogers[1], Deanna Romblad[1], Bob Ruhfel[1], Richard Scott[1], Cynthia Sitter[1], Michelle Smallwood[1], Erin Stewart[1], Renee Strong[1], Ellen Suh[1], Reginald Thomas[1], Ni Ni Tint[1], Sukyee Tse[1], Claire Vech[1], Gary Wang[1], Jeremy Wetter[1], Sherita Williams[1], Monica Williams[1], Sandra Windsor[1], Emily Winn-Deen[1], Keriellen Wolfe[1], Jayshree Zaveri[1], Karena Zaveri[1], Josep F. Abril[14], Roderic Guigó[14], Michael J. Campbell[1], Kimmen V. Sjolander[1], Brian Karlak[1], Anish Kejariwal[1], Huaiyu Mi[1], Betty Lazareva[1], Thomas Hatton[1], Apurva Narechania[1], Karen Diemer[1], Anushya Muruganujan[1], Nan Guo[1], Shinji Sato[1], Vineet Bafna[1], Sorin Istrail[1], Ross Lippert[1], Russell Schwartz[1], Brian Walenz[1], Shibu Yooseph[1], David Allen[1], Anand Basu[1], James Baxendale[1], Louis Blick[1], Marcelo Caminha[1], John Carnes-Stine[1], Parris Caulk[1], Yen-Hui Chiang[1], My Coyne[1], Carl Dahlke[1], Anne Deslattes Mays[1], Maria Dombroski[1], Michael Donnelly[1], Dale Ely[1], Shiva Esparham[1], Carl Fosler[1], Harold Gire[1], Stephen Glanowski[1], Kenneth Glasser[1], Anna Glodek[1], Mark Gorokhov[1], Ken Graham[1], Barry Gropman[1], Michael Harris[1], Jeremy Heil[1], Scott Henderson[1], Jeffrey Hoover[1], Donald Jennings[1], Catherine Jordan[1], James Jordan[1], John Kasha[1], Leonid Kagan[1], Cheryl Kraft[1], Alexander Levitsky[1], Mark Lewis[1], Xiangjun Liu[1], John Lopez[1], Daniel Ma[1], William Majoros[1], Joe McDaniel[1], Sean Murphy[1], Matthew Newman[1], Trung Nguyen[1], Ngoc Nguyen[1], Marc Nodell[1], Sue Pan[1], Jim Peck[1], Marshall Peterson[1], William Rowe[1], Robert Sanders[1], John Scott[1], Michael Simpson[1], Thomas Smith[1], Arlan Sprague[1], Timothy Stockwell[1], Russell Turner[1], Eli Venter[1], Mei Wang[1], Meiyuan Wen[1], David Wu[1], Mitchell Wu[1], Ashley Xia[1], Ali Zandieh[1] and Xiaohong Zhu[1]

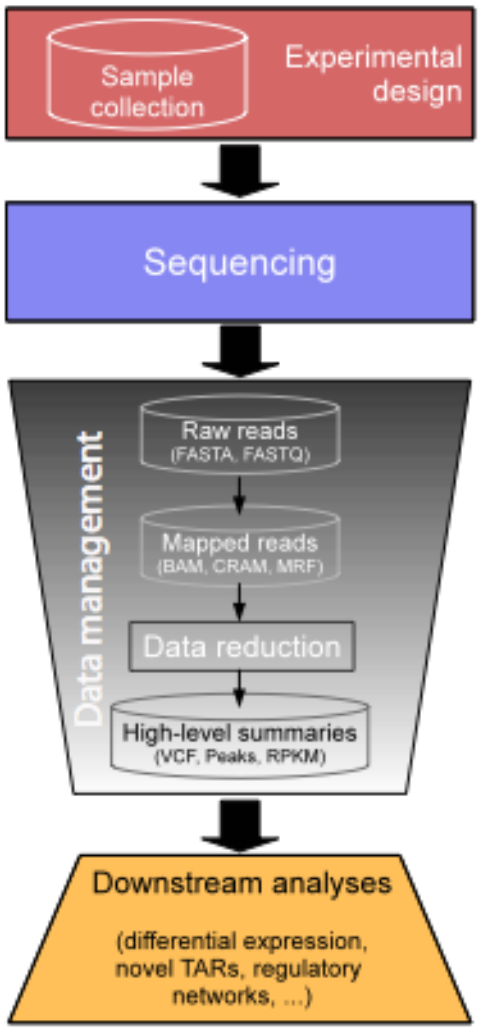
International Human Genome Sequencing Consortium Eric S. Lander[1], Lauren M. Linton[1], Bruce Birren[1], Chad Nusbaum[1], Michael C. Zody[1], Jennifer Baldwin[1], Keri Devon[1], Ken Dewar[1], Michael Doyle[1], William FitzHugh[1], Roel Funke[1], Diane Gage[1], Katrina Harris[1], Andrew Heaford[1], John Howland[1], Lisa Kann[1], Jessica Lehoczky[1], Rosie LeVine[1], Paul McEwan[1], Kevin McKernan[1], James Meldrim[1], Jill P. Mesirov[1], Cher Miranda[1], William Morris[1], Jerome Naylor[1], Christina Raymond[1], Mark Rosetti[1], Ralph Santos[1], Andrew Sheridan[1], Carrie Sougnez[1], Nicole Stange-Thomann[1], Nikola Stojanovic[1], Aravind Subramanian[1] & Dudley Wyman[1] for Whitehead Institute for Biomedical Research, Center for Genome Research:, Jane Rogers[2], John Sulston[2], Rachael Ainscough[2], Stephan Beck[2], David Bentley[2], John Burton[2], Christopher Clee[2], Nigel Carter[2], Alan Coulson[2], Rebecca Deadman[2], Panos Deloukas[2], Andrew Dunham[2], Ian Dunham[2], Richard Durbin[2], Lisa French[2], Darren Grafham[2], Simon Gregory[2], Tim Hubbard[2], Sean Humphray[2], Adrienne Hunt[2], Matthew Jones[2], Christine Lloyd[2], Amanda McMurray[2], Lucy Matthews[2], Simon Mercer[2], Sarah Milne[2], James C. Mullikin[2], Andrew Mungall[2], Robert Plumb[2], Mark Ross[2], Ratna Shownkeen[2] & Sarah Sims[2] for The Sanger Centre:, Robert H. Waterston[3], Richard K. Wilson[3], LaDeana W. Hillier[3], John D. McPherson[3], Marco A. Marra[3], Elaine R. Mardis[3], Lucinda A. Fulton[3], Asif T. Chinwalla[3], Kymberlie H. Pepin[3], Warren R. Gish[3], Stephanie L. Chissoe[3], Michael C. Wendl[3], Kim D. Delehaunty[3], Tracie L. Miner[3], Andrew Delehaunty[3], Jason B. Kramer[3], Lisa L. Cook[3], Robert S. Fulton[3], Douglas L. Johnson[3], Patrick J. Minx[3] & Sandra W. Clifton[3] for Washington University Genome Sequencing Center, Trevor Hawkins[4], Elbert Branscomb[4], Paul Predki[4], Paul Richardson[4], Sarah Wenning[4], Tom Slezak[4], Norman Doggett[4], Jan-Fang Cheng[4], Anne Olsen[4], Susan Lucas[4], Christopher Elkin[4], Edward Uberbacher[4] & Marvin Frazier[4] for US DOE Joint Genome Institute:, Richard A. Gibbs[5], Donna M. Muzny[5], Steven E. Scherer[5], John B. Bouck[5], Erica J. Sodergren[5], Kim C. Worley[5], Catherine M. Rives[5], James H. Gorrell[5], Michael L. Metzker[5], Susan L. Naylor[6], Raju S. Kucherlapati[7], David L. Nelson & George M. Weinstock[8] for Baylor College of Medicine Human Genome Sequencing Center:, Yoshiyuki Sakaki[9], Asao Fujiyama[9], Masahira Hattori[9], Tetsushi Yada[9], Atsushi Toyoda[9], Takehiko Itoh[9], Chiharu Kawagoe[9], Hidemi Watanabe[9], Yasushi Totoki[9] & Todd Taylor[9] for RIKEN Genomic Sciences Center:, Jean Weissenbach[10], Roland Heilig[10], William Saurin[10], Francois Artiguenave[10], Philippe Brottier[10], Thomas Bruls[10], Eric Pelletier[10], Catherine Robert[10] & Patrick Wincker[10] for Genoscope and CNRS UMR-8030:, André Rosenthal[11], Matthias Platzer[12], Gerald Nyakatura[11], Stefan Taudien[12] & Andreas Rump[12] for Department of Genome Analysis, Institute of Molecular Biotechnology:, Douglas R. Smith[13], Lynn Doucette-Stamm[13], Marc Rubenfield[13], Keith Weinstock[11], Hong Mei Lee[11] & JoAnn Dubois[11] for GTC Sequencing Center:, Huanming Yang[13], Jun Yu[13], Jian Wang[13], Guyang Huang[13] & Jun Gu[13] for Beijing Genomics Institute/Human Genome Center:, Leroy Hood[14], Lee Rowen[14], Anup Madan[14] & Shizen Qin[14] for Multimegabase Sequencing Center, The Institute for Systems Biology:, Ronald W. Davis[15], Nancy A. Federspiel[15], A. Pia Abola[17] & Michael J. Proctor[17] for Stanford Genome Technology Center:, Bruce A. Roe[22], Feng Chen[22] & Huaqin Pan[22] for University of Oklahoma's Advanced Center for Genome Technology:, Juliane Ramser[23], Hans Lehrach[23] & Richard Reinhardt[23] for Max Planck Institute for Molecular Genetics:, W. Richard McCombie[24], Melissa de la Bastide[24] & Nellay Dedhia[24] for Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center:, Helmut Blöcker[25], Klaus Hornischer[25] & Gabriele Nordsiek[25] for GBF—German Research Centre for Biotechnology:, Richa Agarwala[26], L. Aravind[26], Jeffrey A. Bailey[27], Alex Bateman[2], Serafim Batzoglou[1], Ewan Birney[28], Peer Bork[29,30], Daniel G. Brown[1], Christopher B. Burge[31], Lorenzo Cerutti[28], Hsiu-Chuan Chen[26], Deanna Church[26], Michele Clamp[2], Richard R. Copley[30], Tobias Doerks[29,30], Sean R. Eddy[32], Evan E. Eichler[27], Terrence S. Furey[33], James Galagan[1], James G. R. Gilbert[2], Cyrus Harmon[34], Yoshihide Hayashizaki[35], David Haussler[26], Henning Hermjakob[28], Karsten Hokamp[37], Wonhee Jang[26], L. Steven Johnson[32], Thomas A. Jones[32], Simon Kasif[38], Arek Kasprzyk[28], Scot Kennedy[39], W. James Kent[40], Paul Kitts[26], Eugene V. Koonin[26], Ian Korf[3], David Kulp[24], Doron Lancet[41], Todd M. Lowe[42], Aoife McLysaght[37], Tarjei Mikkelsen[38], John V. Moran[43], Nicola Mulder[28], Victor J. Pollara[1], Chris P. Ponting[44], Greg Schuler[26], Jörg Schultz[30], Guy Slater[28], Arian F. A. Smit[45], Elia Stupka[28], Joseph Szustakowki[28], Danielle Thierry-Mieg[26], Jean Thierry-Mieg[26], Lukas Wagner[26], John Wallis[3], Raymond Wheeler[24], Alan Williams[24], Yuri I. Wolf[26], Kenneth H. Wolfe[37], Shiaw-Pyng Yang[2] & Ru-Fang Yeh[31] for

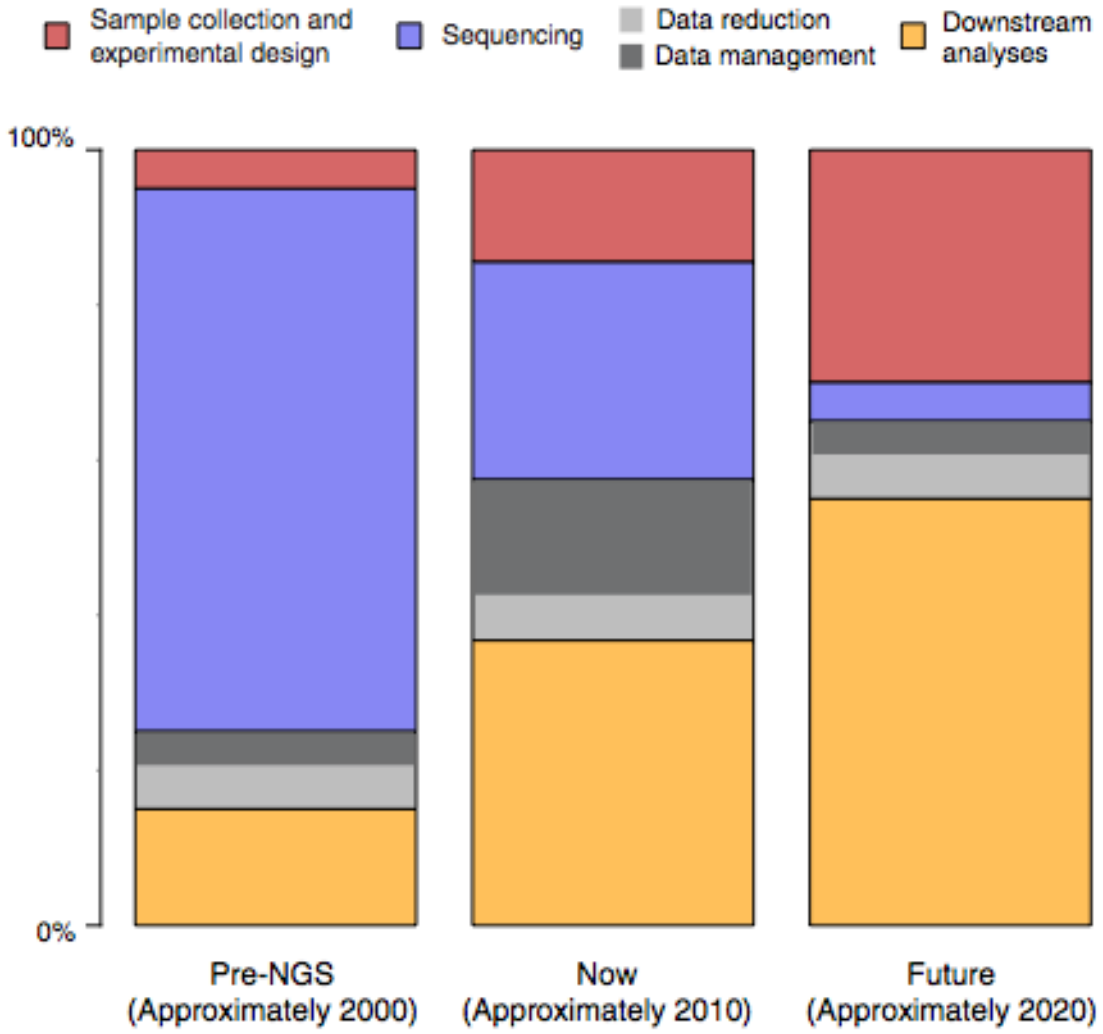
*Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings):, Francis Collins[46], Mark S. Guyer[46], Jane Peterson[46], Adam Felsenfeld[46] & Kris A. Wetterstrand[46] for Scientific management: National Human Genome Research Institute, US National Institutes of Health:, Richard M. Myers[18], Jeremy Schmutz[18], Mark Dickson[18], Jane Grimwood[18] & David R. Cox[18] for Stanford Human Genome Center:, Maynard V. Olson[19], Rajinder Kaul[19] & Christopher Raymond[19] for University of Washington Genome Center:, Nobuyoshi Shimizu[20], Kazuhiko Kawasaki[20] & Shinsei Minoshima[20] for Department of Molecular Biology, Keio University School of Medicine:, Glen A. Evans[21], Maria Athanasiou[21] & Roger Schultz[21] for University of Texas Southwestern Medical Center at Dallas:, Aristides Patrinos[47] for Office of Science, US Department of Energy: & Michael J. Morgan[48] for The Wellcome Trust:

2000:	8 YEARS	$3,000,000,000
Now:	10 days	$10,000



Stein, *Genome Biology* 2010, 11:207



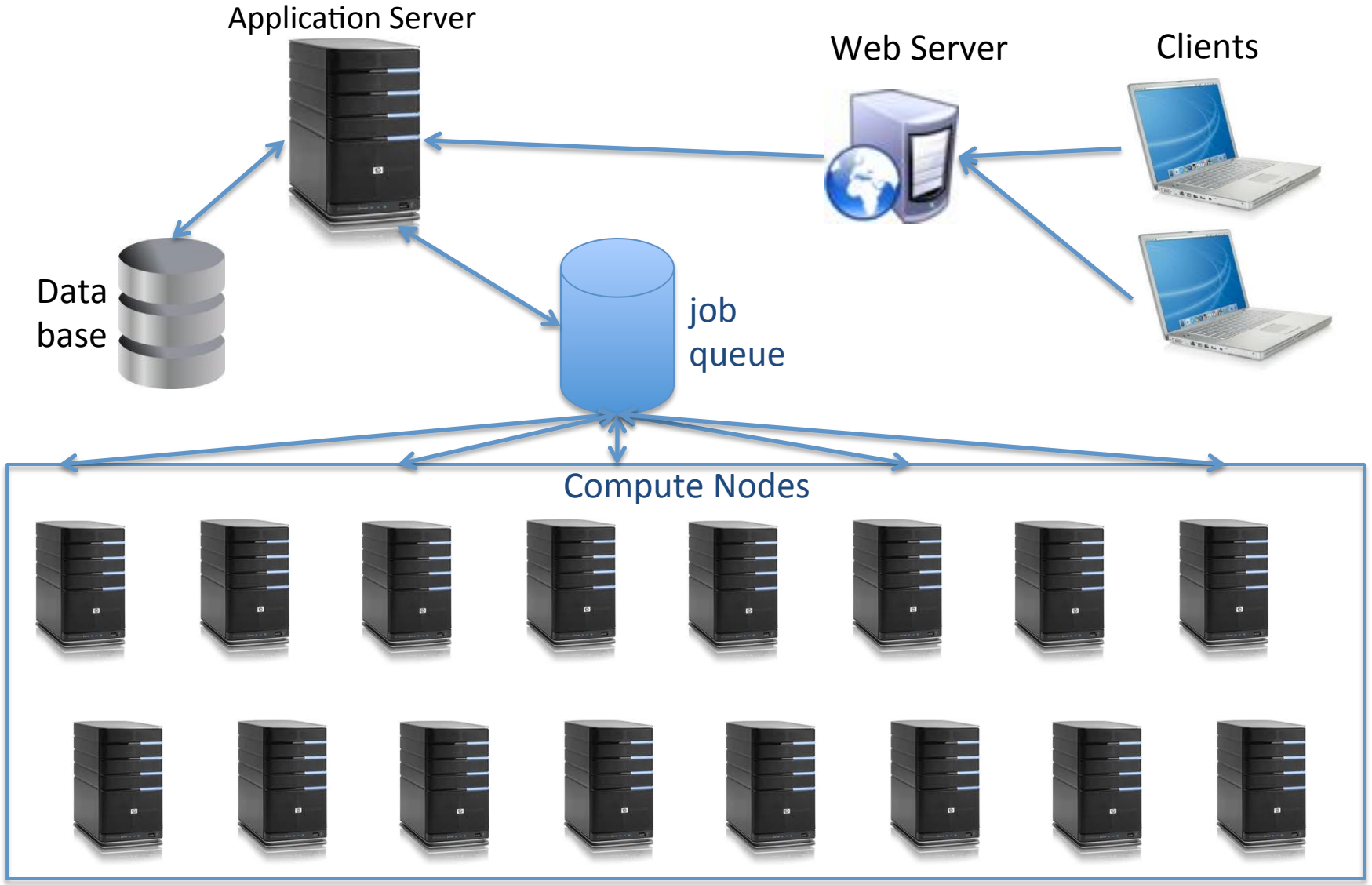


Sboner et al, *Genome Biology* 2011, 12:125

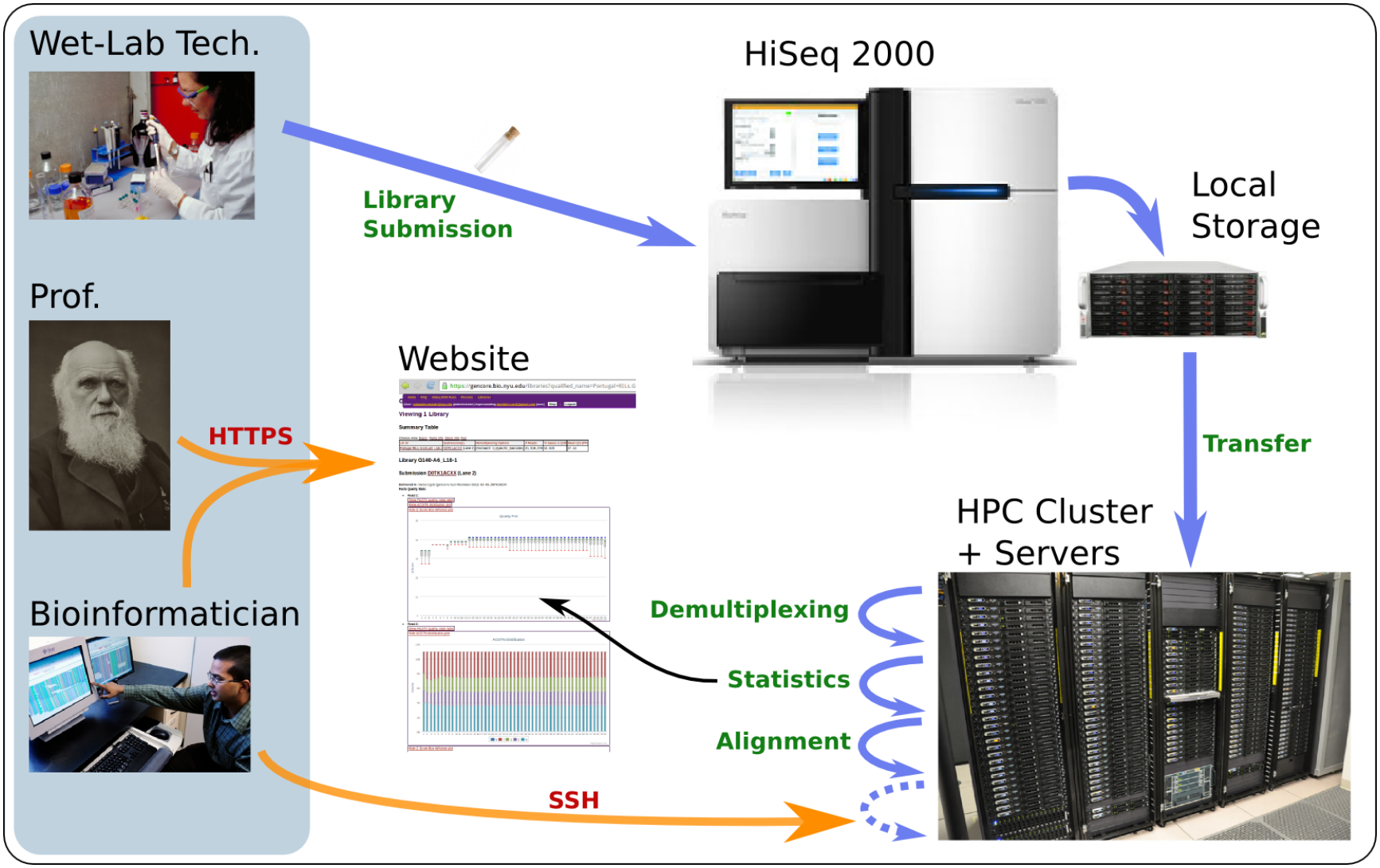
Computational Stack

- Website
  - data visualization, upload/download
- Analysis Pipeline
  - run 3rd party tools, distribute jobs, track results
- LIMS (Laboratory Information Management System)
  - metadata tracking of samples, libraries, protocols, etc
- Systems Infrastructure
  - operating systems, packages, virtual machines, network

Software Architecture



The Whole System



Big-Data is Not Just Big

- Volume
- Variety
- Velocity

The Types of our DSL

```
type =
  | Bool
  | Timestamp
  | Int
  | Real
  | String
  | Option of type
  | Array of type
  | Record of (string * type) list
  | Enumeration of string list
  | Function of type * type
  | Volume of volume
```

- More expressive than SQL
- More closely models the schemas we need
- We have full control over DSL objects

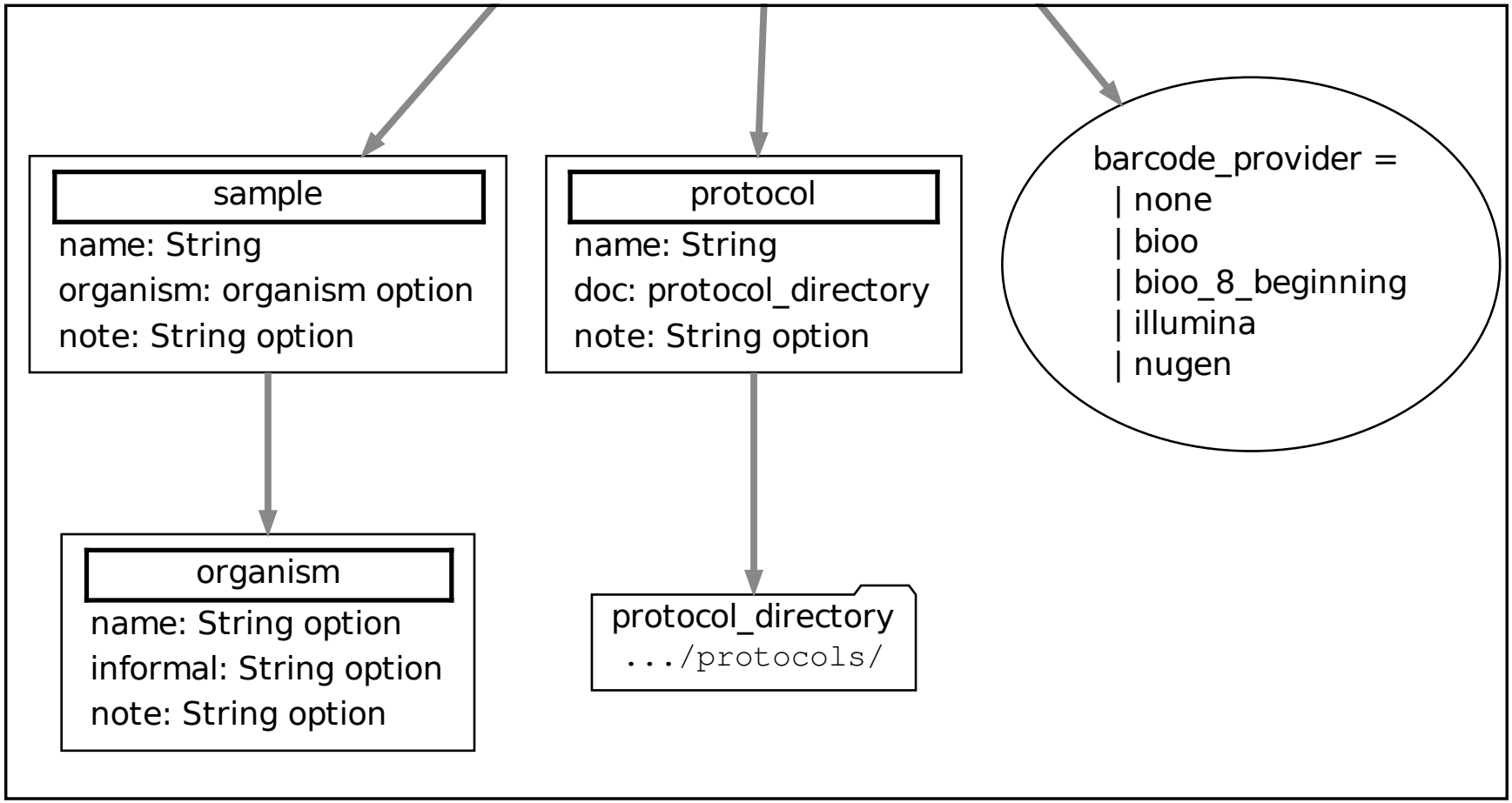
DSL Compiler

- From a given program in the DSL, we *automatically* generate
  - SQL scripts to initialize and reset the database
  - About ~10k lines (and growing) of OCaml strongly typed database reads/inserts
  - Web widgets
  - Figures in this talk

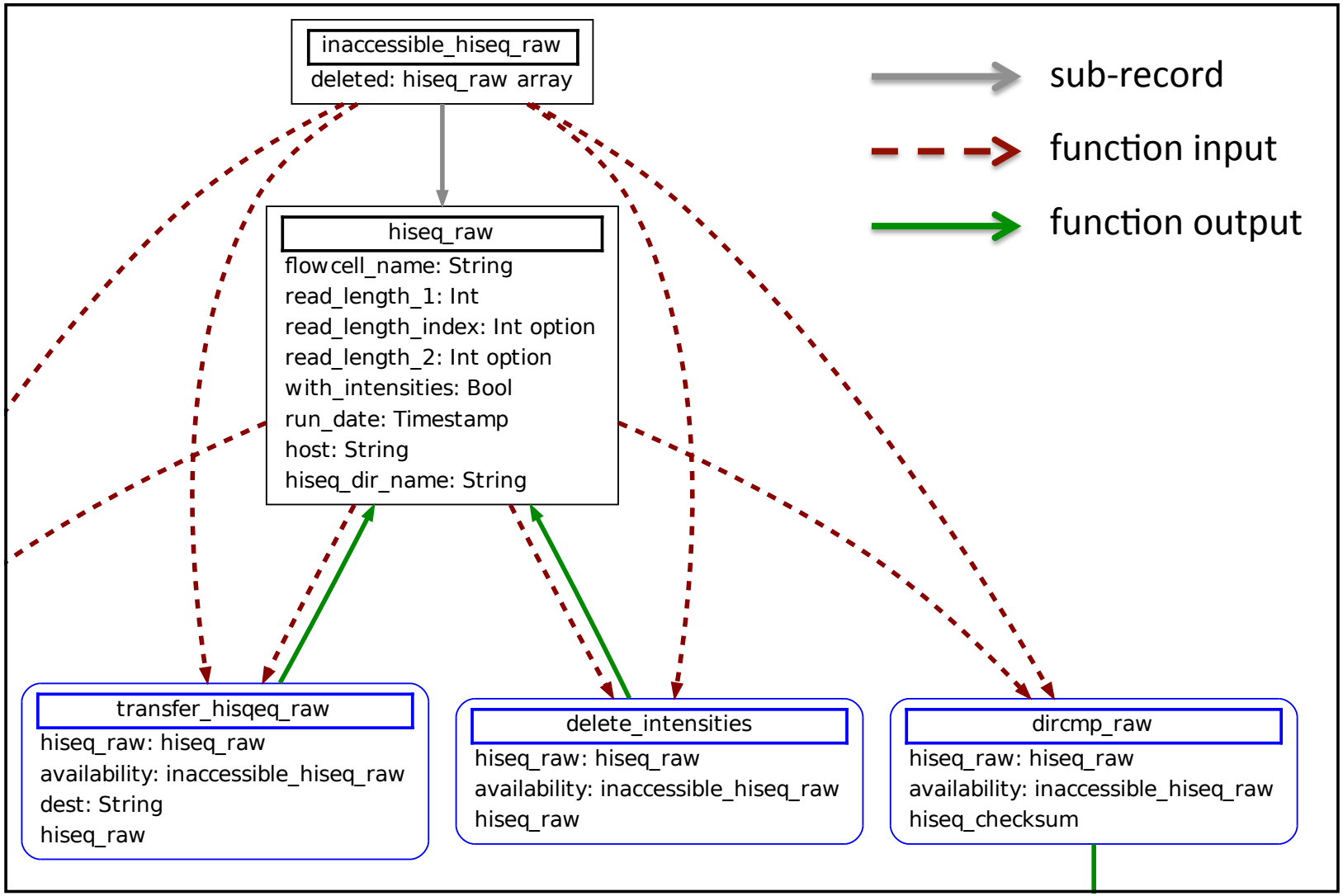
A Program in our DSL

```
(record sample
      (name       string)
      (organism   organism option)
      notable)
```

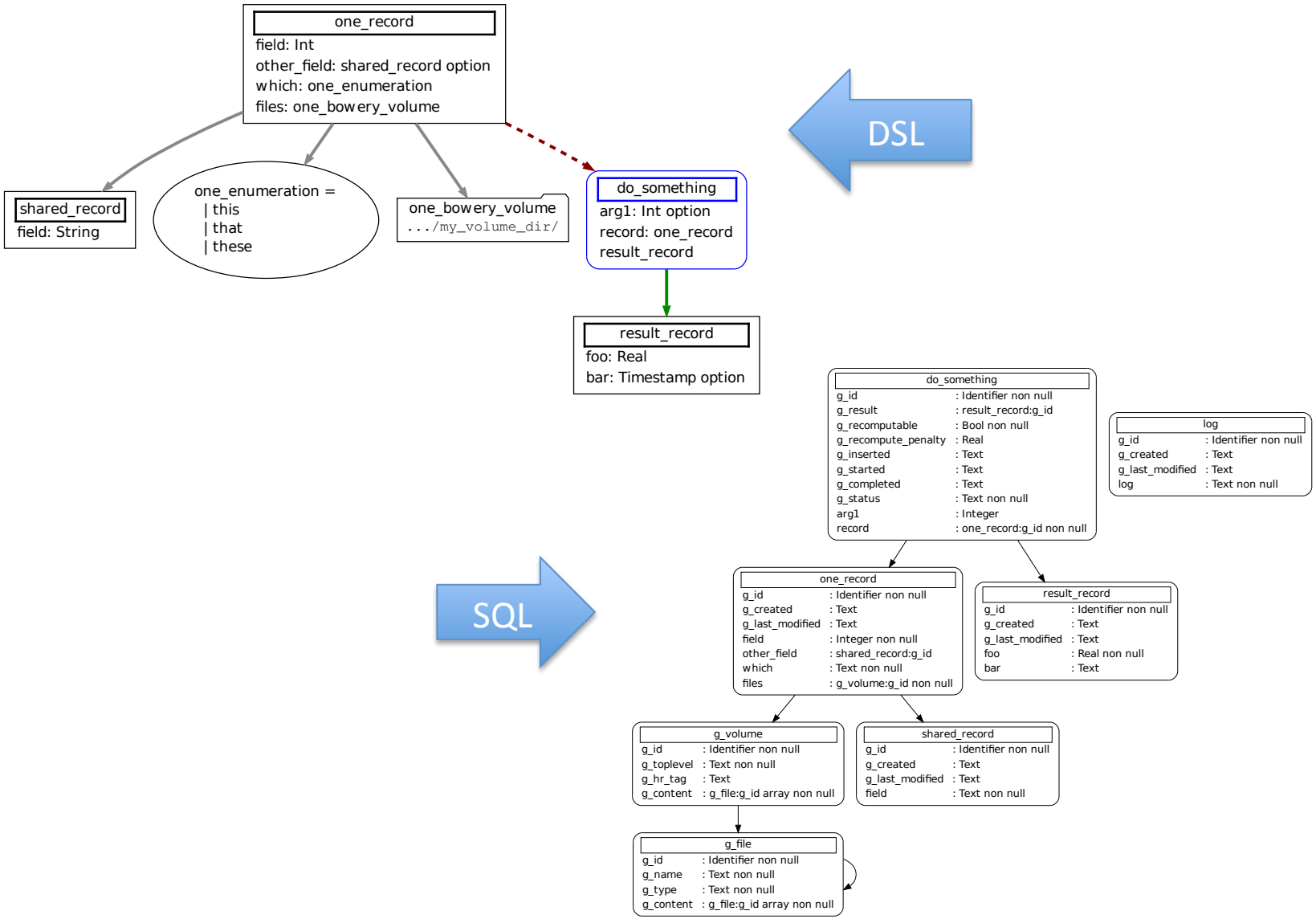
Samples, Organisms, and Protocols



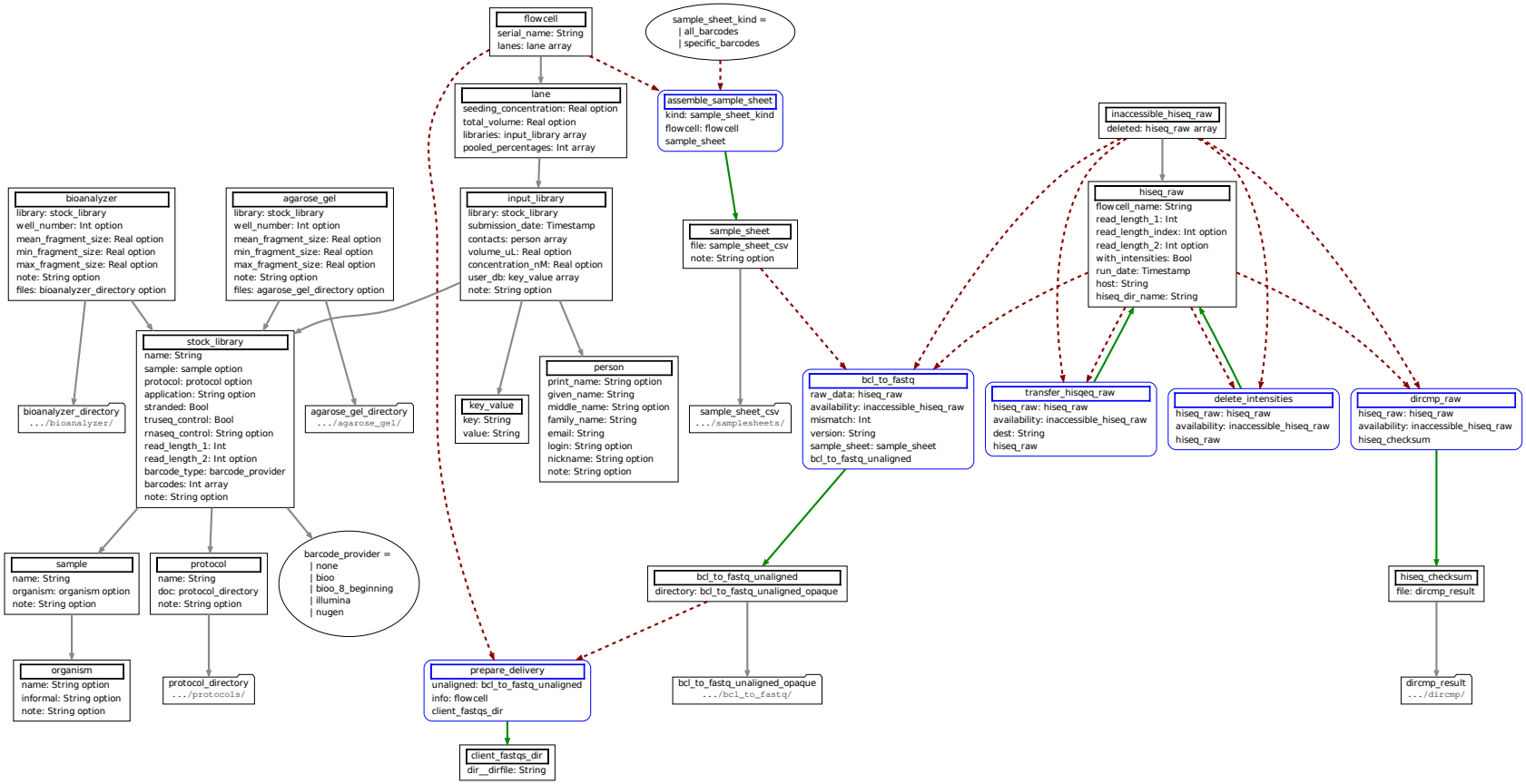
Managing HiSeq Data



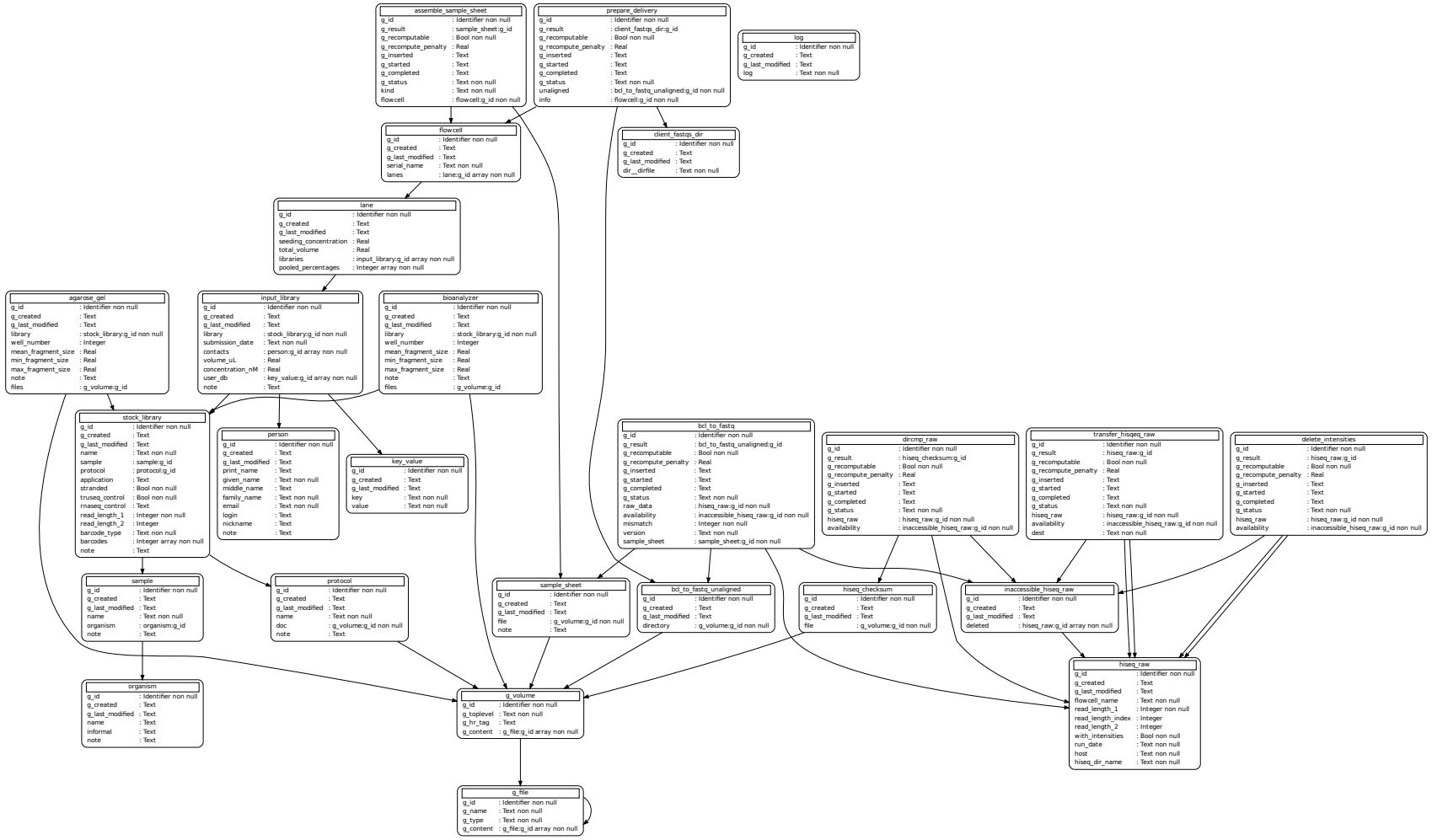
Generating SQL



Entire NYU GenCore Data Schema (in our new DSL)



Entire SQL Schema (auto-generated)



Virtual Filesystem

- Database entries have to map to physical files

- Files in DSL are often *recomputable*
- But don't have to exist on physical filesystem, or can exist in compressed form

- File metadata is recorded
  - e.g. FASTQ quality-score encoding = 33 or 64
- Easier file restructuring/renaming
- Easier migration to new clusters

Function Values

- No lambdas
- Only hardcoded function constants
- Output values are “normal” data
- Implementation must account for
  - failures
  - time lags
  - serialization

Website – powered by Ocsigen

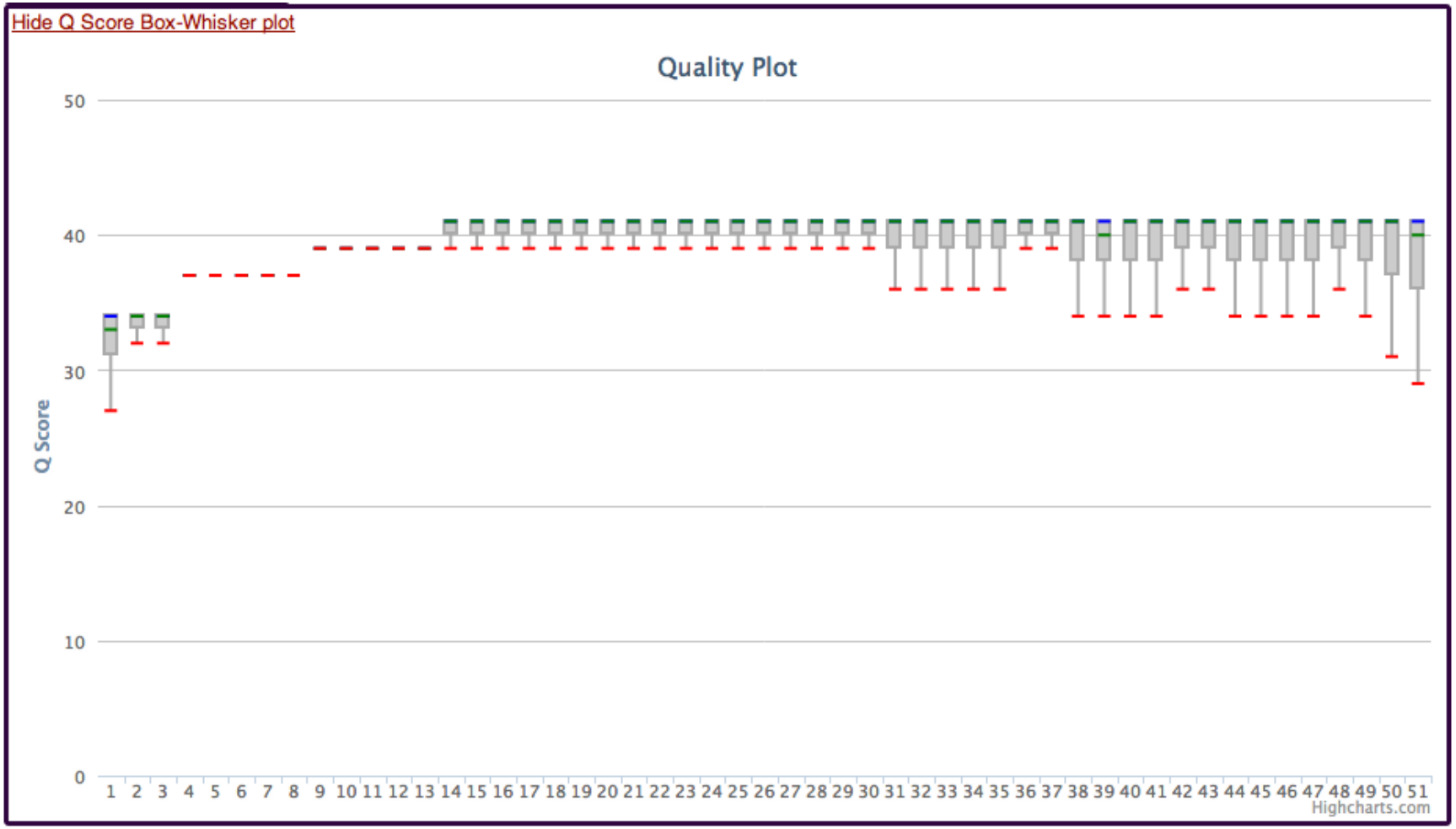
GENCORE HOME

Welcome

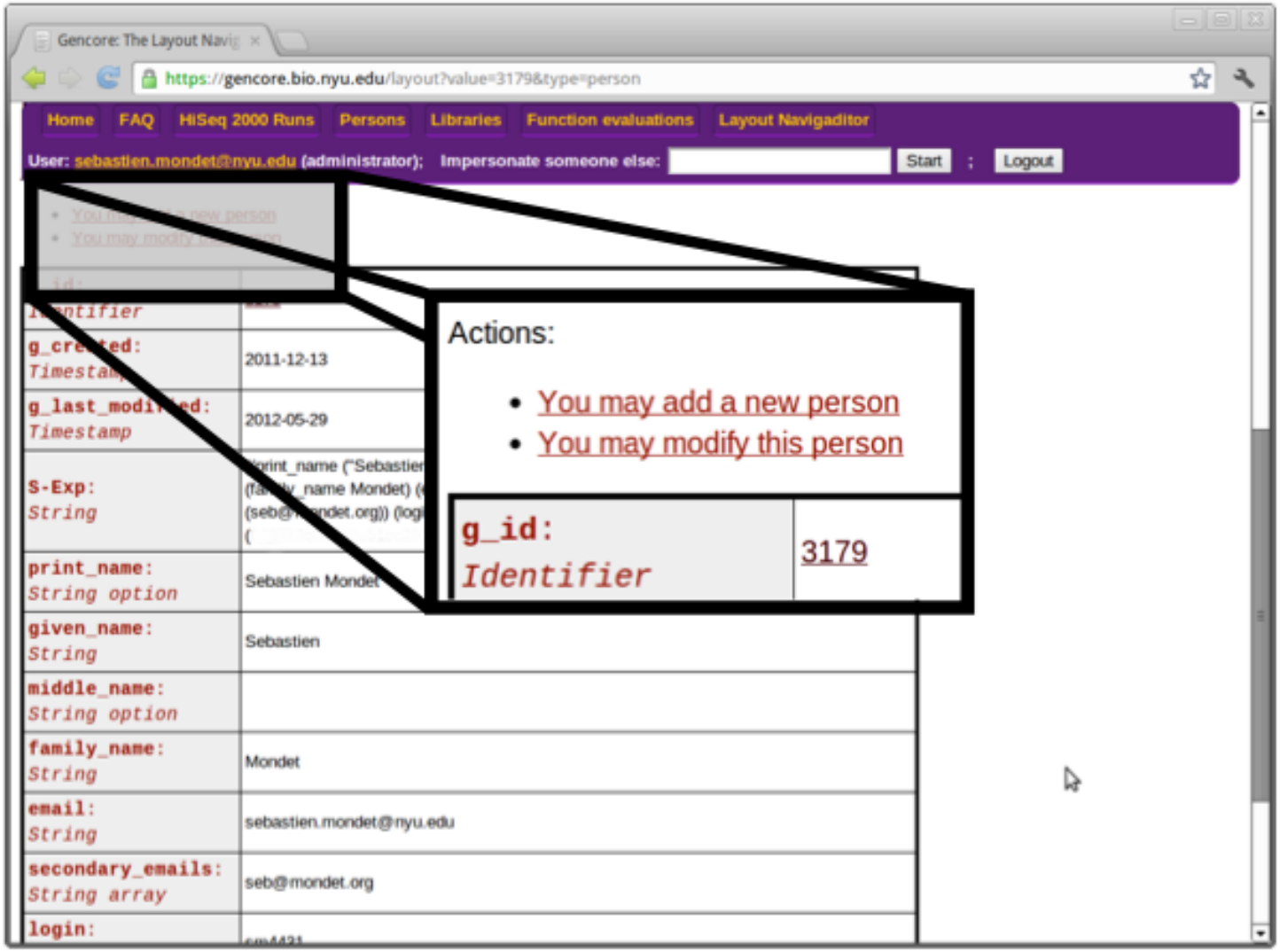
This is Gencore's website; for library submission information see the FAQ or GenCore's Google-Docs.

Menu

- HiSeq 2000 Runs
- Persons
- Libraries
- Function evaluations
- Layout Navigaditor


Hide Q Score Box-Whisker plot
Quality Plot
Q Score
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
Highcharts.com

Auto-generated Forms



Data Migrations

val migrator: S-Expression -> S-Expression

$ hitscore dump-to-file backup_v42

$ ./migrator backup_v42 backup_v43

$ hitscore wipe-out-database

$ hitscore init-database

$ hitscore load-file backup_v43

$ hitscore verify-layout

We have done over 25 data migrations!

OCaml Has Many Libraries

- Core & Batteries
- Lwt – lightweight threads
- Ocsigen – web programming framework
- Biocaml
- PG'OCaml
- Xmlm
- Ocamlnet
- … and many more

Database -> Domain Logic -> Web Interface

Experience with OCaml

- The Good
  - Good libraries
  - Industrial strength
  - Hackable
  - Option to be unsafe feels safe
  - Excellent performance
- Could Be Better
  - Public relations
  - Build system
  - Blessed libraries
  - More libraries

Functional Programming in Biology

- "Functional programming" is becoming a recognized term
- Programmers are desperately needed
- Be sure to distinguish *software engineering* from *data analysis*
- Key to success:
  - acquire domain expertise
  - build software fast
- Discuss programming scientifically

Conclusions

- The Genomics Sequencing Core at NYU CGSB runs on OCaml
- Entire system built by ~1.3 programmers
- First version: in production within 2 months
- Biology needs you!