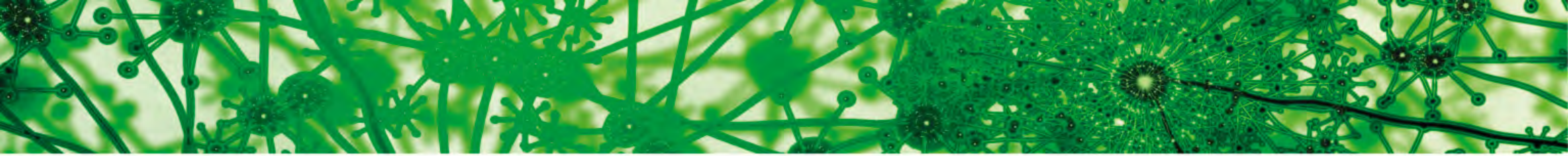


UiO : Department of Biosciences

University of Oslo

MBV4410/9410 Fall 2016

Bioinformatics for Molecular Biology

General information

Course coordinator: Jon Bråte

Email: jon.brate@ibv.uio.no

Phone: +47 922 44 582

Course administrator: Torill Rørtveit

Email: torill.rortveit@ibv.uio.no

Course web page:
https://wiki.uio.no/projects/clsi/index.php/MBV-INFX410_2016

Purpose of the course

Goals:

- Learn how to *obtain* and *analyse* different types of *biological data*
- Learn basic file handling, and how to run and use programs on the Unix command line

Background:

- Primarily molecular biology and biochemistry. No programming skills required

General information

	Level	Credits	Exam	Oblig	Extra oblig
MBV-INF4410	MSc	10	Yes	Yes	No
MBV-INF9410	PhD	10	Yes	Yes	Yes (at least 2500 words)

Home exam:

- Sent out to all participants at 15:00 Friday December 9 by email
- Must be returned at latest 15:00 Friday December 16 by email (NB! To Torill – not Jon!)

Oblig:

- Assignment (including essay) must be returned by 23:59 Friday December 2 at by email

Obligatory assignment (“oblig”)

- Exercise for oblig will be handed out at the end of course week 2.
- Will be relatively easy and similar to exercises in course weeks 1 and 2.
- Must be returned before the first lecture in course week 4 (December 5).
- PhD students (MBV-INF9410) must in addition write an essay (> 2500 words).
  - Describe how you would use 2 or more of the methods covered in the course in your own research.
- **Obligatory assignment must be approved before you can take the exam!**

Exam

The exam for this course will be a week long take-home exam. Only students who have completed and passed the obligatory assignment are allowed to take the exam.

The exam will be sent to all qualified participants at 15:00 December 9 by email.

The completed exam must be returned at latest 15:00 on December 16 by email to Torill Rørtveit (torill.rortveit@ibv.uio.no) - **NOT TO JON!**. Please put the course code and your name in the subject field (e.g. “Exam MBV-INF4410 Your Name”).

The exam must be handed in as a single PDF document. The document should be marked with the date, course code and your name.

If necessary for evaluating the exam, a small oral examination may be arranged.

MBV-INF4410: Grade scale A-F (F = fail)
MBV-INF9410: Pass/fail (Pass = B or better)

Curriculum

- All lectures
- All exercises, demos and computer labs
- Obligatory assignments
- Articles listed on the wiki

General information

Course web page

https://wiki.uio.no/projects/clsi/index.php/MBV-INFX410_2016

General information

- Everyone need to send me an email with this subject header (to jon.brate@ibv.uio.no):

“Course version” “email address” “full name”

Like this:

MBV-INF9410 jon.brate@ibv.uio.no Jon Bråte

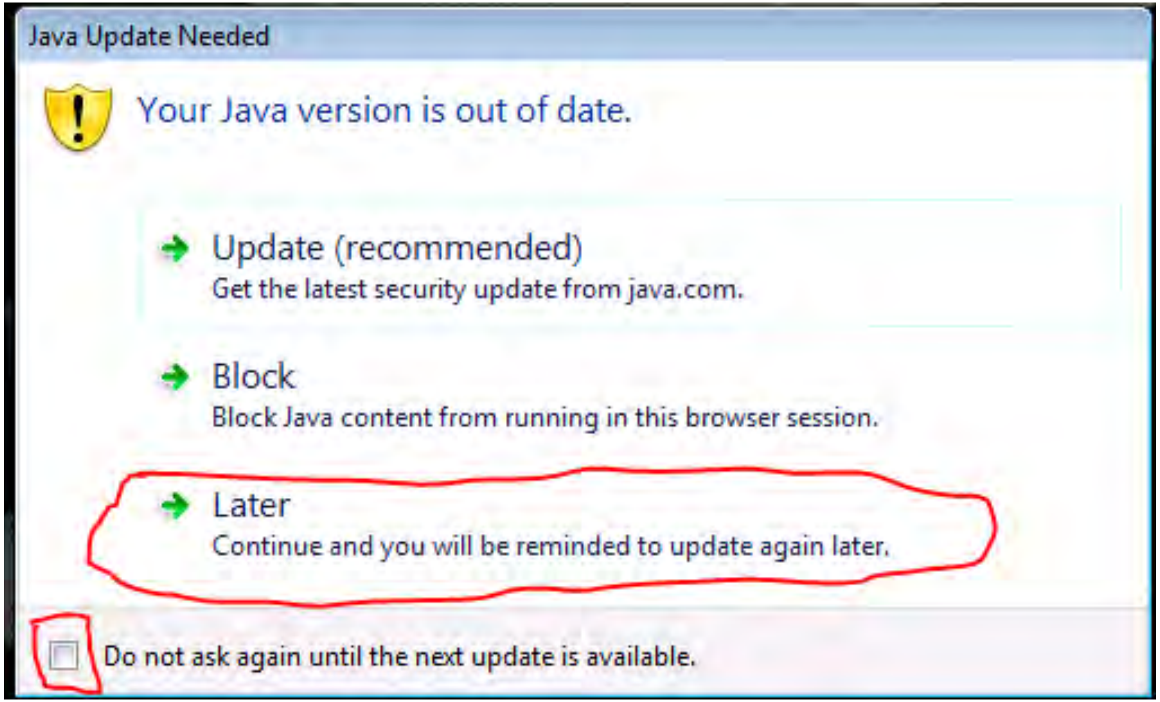
NB!
I will send you the obligs, exam and important information to this address!

Laptops

I urge you to bring your own laptops

- We don’t have access to computer rooms every day
- You need permissions to install software on it
- Please bring and **external mouse**
- You **must** have access to the UiO network (UiO username and password – see here)
- All files should be stored on your UiO home directory – not locally on your laptop!

Don’t update Java on the Desktops!!



Java Update Needed

Your Java version is out of date.

➔ Update (recommended)
Get the latest security update from java.com.

➔ Block
Block Java content from running in this browser session.

➔ Later
Continue and you will be reminded to update again later.

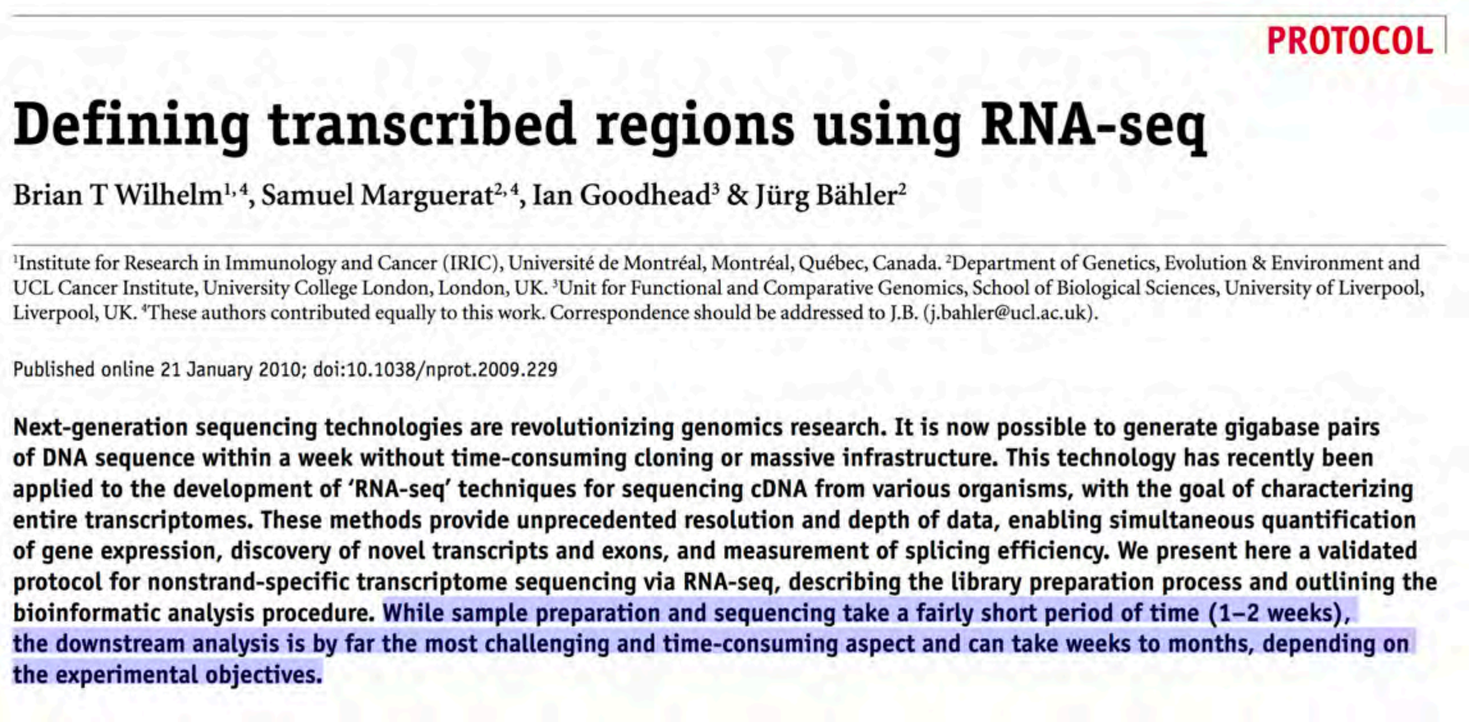
□ Do not ask again until the next update is available.

Computational resources at UiO

- Linux cluster
- Abel was (in June 2012) number 96 on the list of the most powerful computers in the world
- 258 TFLOP/s theoretical peak performance
- We will use Freebee – a ”small corner” of Abel.

Abel home page

The Abel supercomputer



PROTOCOL

Defining transcribed regions using RNA-seq

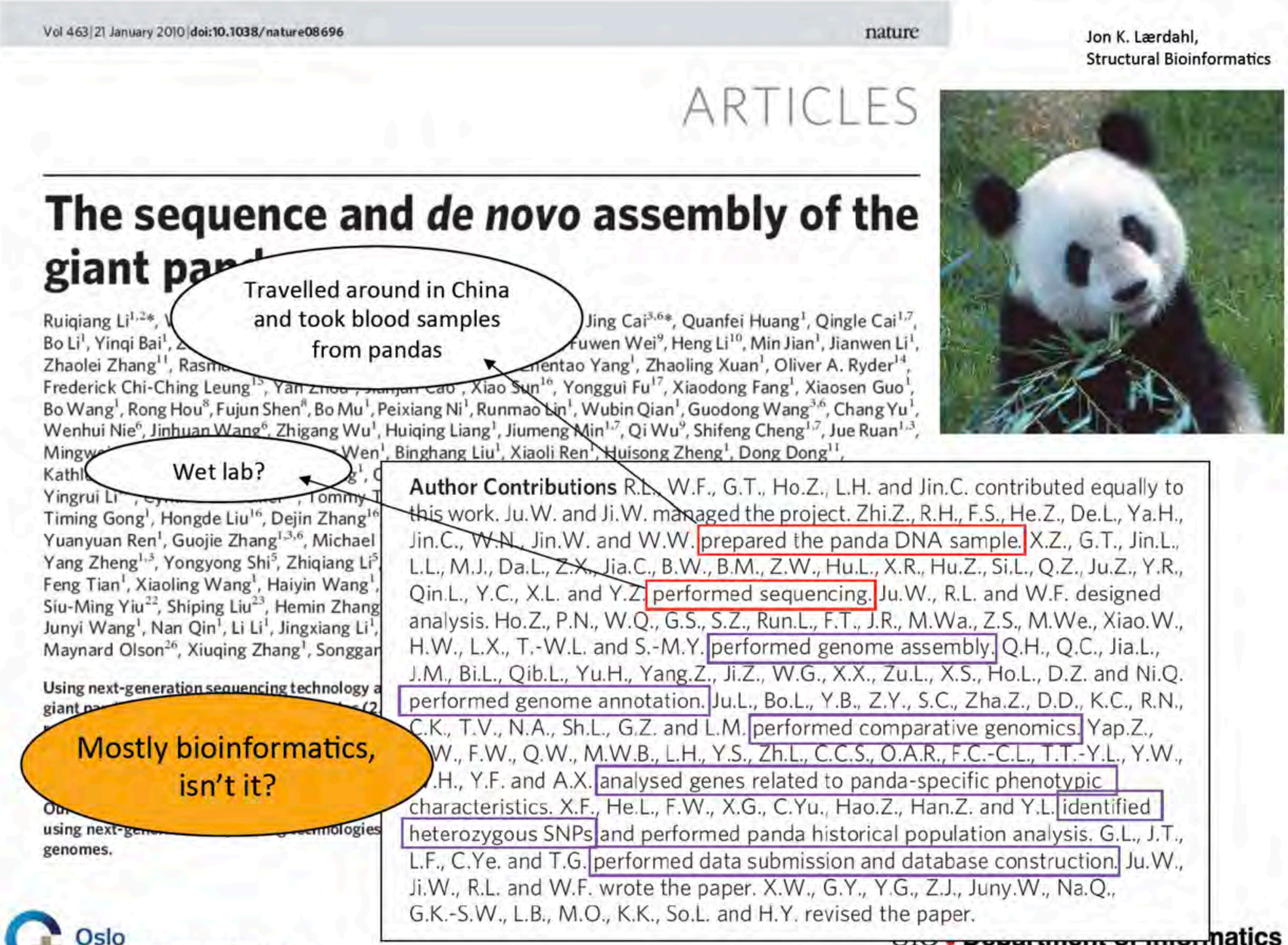
Brian T Wilhelm[1,4], Samuel Marguerat[2,4], Ian Goodhead[3] & Jürg Bähler[2]

[1]Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montréal, Québec, Canada. [2]Department of Genetics, Evolution & Environment and UCL Cancer Institute, University College London, London, UK. [3]Unit for Functional and Comparative Genomics, School of Biological Sciences, University of Liverpool, Liverpool, UK. [4]These authors contributed equally to this work. Correspondence should be addressed to J.B. (j.bahler@ucl.ac.uk).

Published online 21 January 2010; doi:10.1038/nprot.2009.229

Next-generation sequencing technologies are revolutionizing genomics research. It is now possible to generate gigabase pairs of DNA sequence within a week without time-consuming cloning or massive infrastructure. This technology has recently been applied to the development of 'RNA-seq' techniques for sequencing cDNA from various organisms, with the goal of characterizing entire transcriptomes. These methods provide unprecedented resolution and depth of data, enabling simultaneous quantification of gene expression, discovery of novel transcripts and exons, and measurement of splicing efficiency. We present here a validated protocol for nonstrand-specific transcriptome sequencing via RNA-seq, describing the library preparation process and outlining the bioinformatic analysis procedure. While sample preparation and sequencing take a fairly short period of time (1–2 weeks), the downstream analysis is by far the most challenging and time-consuming aspect and can take weeks to months, depending on the experimental objectives.

Modern biology
Wet-lab: 1 week.
Dry-lab (analysing data): months…



Vol 463 | 21 January 2010 | doi:10.1038/nature08696

nature

ARTICLES

The sequence and *de novo* assembly of the giant pa

Ruiqiang Li[1,2]*, W Jing Cai[3,6]*, Quanfei Huang[1], Qingle Cai[1,7], Bo Li[1], Yinqi Bai[1], Z Fuwen Wei[9], Heng Li[10], Min Jian[1], Jianwen Li[1], Zhaolei Zhang[11], Rasm entao Yang[1], Zhaoling Xuan[1], Oliver A. Ryder[14], Frederick Chi-Ching Leung[15], Yan Zhou, Xiao Sun[16], Yonggui Fu[17], Xiaodong Fang[1], Xiaosen Guo[1], Bo Wang[1], Rong Hou[8], Fujun Shen[8], Bo Mu[1], Peixiang Ni[1], Runmao Lin[1], Wubin Qian[1], Guodong Wang[3,6], Chang Yu[1], Wenhui Nie[6], Jinhuan Wang[6], Zhigang Wu[1], Huiqing Liang[1], Jiumeng Min[1,7], Qi Wu[9], Shifeng Cheng[1,7], Jue Ruan[1,3], Mingw Wen[1], Binghang Liu[1], Xiaoli Ren[1], Huisong Zheng[1], Dong Dong[11], Kathl g[1], C Yingrui Li , Tommy T Timing Gong[1], Hongde Liu[16], Dejin Zhang[16] Yuanyuan Ren[1], Guojie Zhang[1,3,6], Michael Yang Zheng[1,3], Yongyong Shi[5], Zhiqiang Li[5] Feng Tian[1], Xiaoling Wang[1], Haiyin Wang[1], Siu-Ming Yiu[22], Shiping Liu[23], Hemin Zhang Junyi Wang[1], Nan Qin[1], Li Li[1], Jingxiang Li[1], Maynard Olson[26], Xiuqing Zhang[1], Songgar

Using next-generation sequencing technology a giant pa (2 Ou using next-ge ologies genomes.

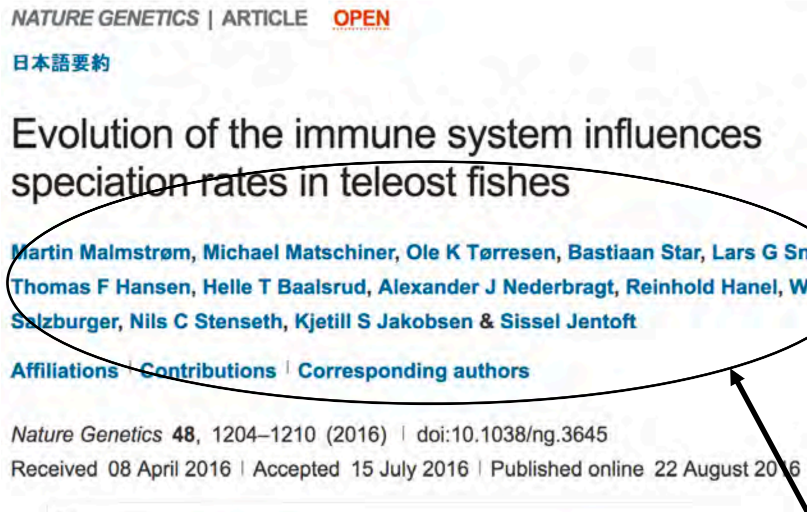
Author Contributions R.L., W.F., G.T., Ho.Z., L.H. and Jin.C. contributed equally to this work. Ju.W. and Ji.W. managed the project. Zhi.Z., R.H., F.S., He.Z., De.L., Ya.H., Jin.C., W.N., Jin.W. and W.W. prepared the panda DNA sample. X.Z., G.T., Jin.L., L.L., M.J., Da.L., Z.X., Jia.C., B.W., B.M., Z.W., Hu.L., X.R., Hu.Z., Si.L., Q.Z., Ju.Z., Y.R., Qin.L., Y.C., X.L. and Y.Z. performed sequencing. Ju.W., R.L. and W.F. designed analysis. Ho.Z., P.N., W.Q., G.S., S.Z., Run.L., F.T., J.R., M.Wa., Z.S., M.We., Xiao.W., H.W., L.X., T.-W.L. and S.-M.Y. performed genome assembly. Q.H., Q.C., Jia.L., J.M., Bi.L., Qib.L., Yu.H., Yang.Z., Ji.Z., W.G., X.X., Zu.L., X.S., Ho.L., D.Z. and Ni.Q. performed genome annotation. Ju.L., Bo.L., Y.B., Z.Y., S.C., Zha.Z., D.D., K.C., R.N., C.K., T.V., N.A., Sh.L., G.Z. and L.M. performed comparative genomics. Yap.Z., W., F.W., Q.W., M.W.B., L.H., Y.S., Zh.L., C.C.S., O.A.R., F.C.-C.L., T.T.-Y.L., Y.W., H., Y.F. and A.X. analysed genes related to panda-specific phenotypic characteristics. X.F., He.L., F.W., X.G., C.Yu., Hao.Z., Han.Z. and Y.L. identified heterozygous SNPs and performed panda historical population analysis. G.L., J.T., L.F., C.Ye. and T.G. performed data submission and database construction. Ju.W., Ji.W., R.L. and W.F. wrote the paper. X.W., G.Y., Y.G., Z.J., Juny.W., Na.Q., G.K.-S.W., L.B., M.O., K.K., So.L. and H.Y. revised the paper.

Travelled around in China and took blood samples from pandas

Wet lab?

Mostly bioinformatics, isn't it?

Oslo universitetssykehus



NATURE GENETICS | ARTICLE OPEN

日本語要約

Evolution of the immune system influences speciation rates in teleost fishes

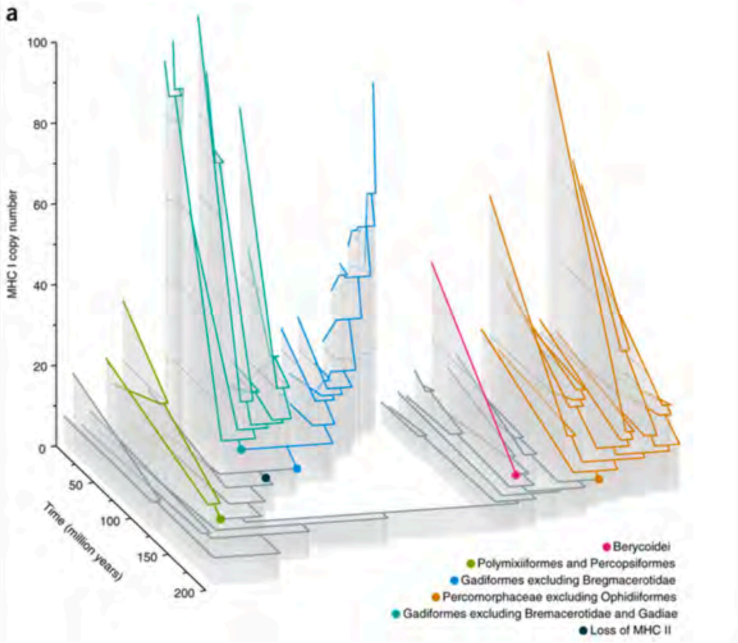
Martin Malmstrøm, Michael Matschiner, Ole K Tørresen, Bastiaan Star, Lars G Snipen, Thomas F Hansen, Helle T Baalsrud, Alexander J Nederbragt, Reinhold Hanel, Walter Salzburger, Nils C Stenseth, Kjetill S Jakobsen & Sissel Jentoft

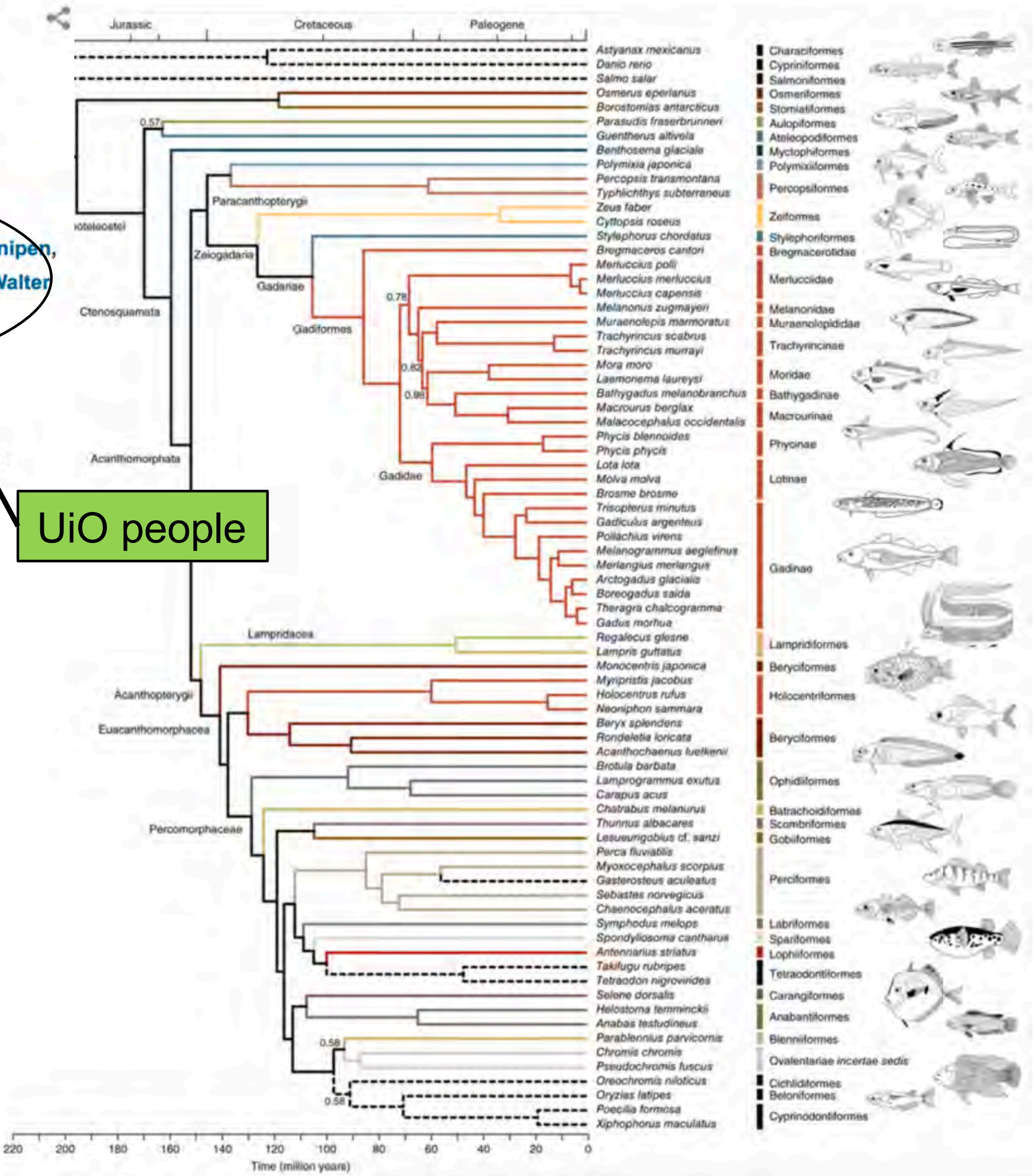
Affiliations | Contributions | Corresponding authors

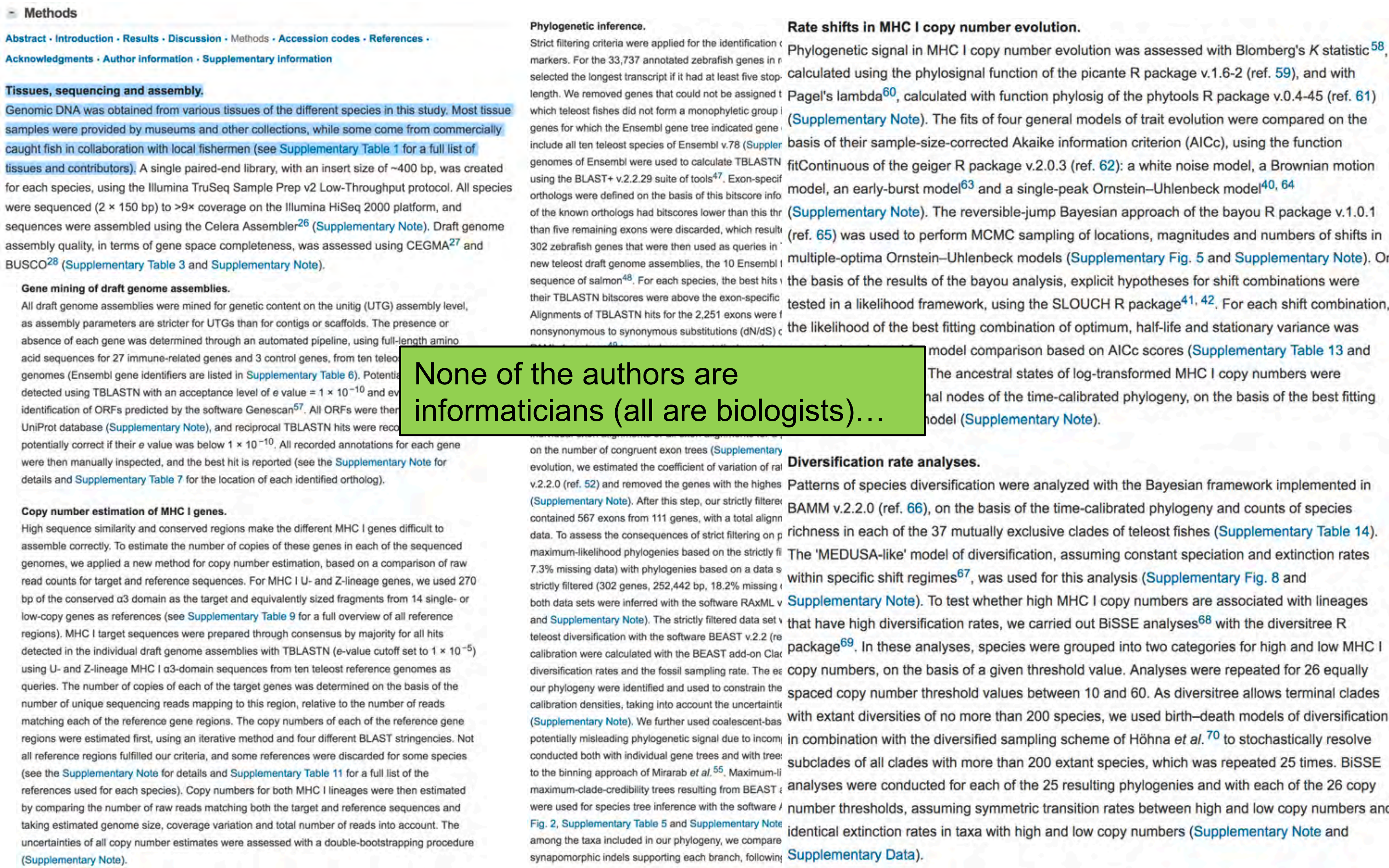
Nature Genetics **48**, 1204–1210 (2016) | doi:10.1038/ng.3645

Received 08 April 2016 | Accepted 15 July 2016 | Published online 22 August 2016

UiO people







- Methods

Abstract · Introduction · Results · Discussion · Methods · Accession codes · References · Acknowledgments · Author information · Supplementary information

Tissues, sequencing and assembly.

Genomic DNA was obtained from various tissues of the different species in this study. Most tissue samples were provided by museums and other collections, while some come from commercially caught fish in collaboration with local fishermen (see Supplementary Table 1 for a full list of tissues and contributors). A single paired-end library, with an insert size of ~400 bp, was created for each species, using the Illumina TruSeq Sample Prep v2 Low-Throughput protocol. All species were sequenced (2 × 150 bp) to >9× coverage on the Illumina HiSeq 2000 platform, and sequences were assembled using the Celera Assembler[26] (Supplementary Note). Draft genome assembly quality, in terms of gene space completeness, was assessed using CEGMA[27] and BUSCO[28] (Supplementary Table 3 and Supplementary Note).

Gene mining of draft genome assemblies.

All draft genome assemblies were mined for genetic content on the unitig (UTG) assembly level, as assembly parameters are stricter for UTGs than for contigs or scaffolds. The presence or absence of each gene was determined through an automated pipeline, using full-length amino acid sequences for 27 immune-related genes and 3 control genes, from ten teleo genomes (Ensembl gene identifiers are listed in Supplementary Table 6). Potentia detected using TBLASTN with an acceptance level of *e* value = 1 × 10^{-10} and ev identification of ORFs predicted by the software Genescan[57]. All ORFs were then UniProt database (Supplementary Note), and reciprocal TBLASTN hits were reco potentially correct if their *e* value was below 1 × 10^{-10}. All recorded annotations for each gene were then manually inspected, and the best hit is reported (see the Supplementary Note for details and Supplementary Table 7 for the location of each identified ortholog).

Copy number estimation of MHC I genes.

High sequence similarity and conserved regions make the different MHC I genes difficult to assemble correctly. To estimate the number of copies of these genes in each of the sequenced genomes, we applied a new method for copy number estimation, based on a comparison of raw read counts for target and reference sequences. For MHC I U- and Z-lineage genes, we used 270 bp of the conserved α3 domain as the target and equivalently sized fragments from 14 single- or low-copy genes as references (see Supplementary Table 9 for a full overview of all reference regions). MHC I target sequences were prepared through consensus by majority for all hits detected in the individual draft genome assemblies with TBLASTN (*e*-value cutoff set to 1 × 10^{-5}) using U- and Z-lineage MHC I α3-domain sequences from ten teleost reference genomes as queries. The number of copies of each of the target genes was determined on the basis of the number of unique sequencing reads mapping to this region, relative to the number of reads matching each of the reference gene regions. The copy numbers of each of the reference gene regions were estimated first, using an iterative method and four different BLAST stringencies. Not all reference regions fulfilled our criteria, and some references were discarded for some species (see the Supplementary Note for details and Supplementary Table 11 for a full list of the references used for each species). Copy numbers for both MHC I lineages were then estimated by comparing the number of raw reads matching both the target and reference sequences and taking estimated genome size, coverage variation and total number of reads into account. The uncertainties of all copy number estimates were assessed with a double-bootstrapping procedure (Supplementary Note).

Phylogenetic inference.

Strict filtering criteria were applied for the identification ‹
markers. For the 33,737 annotated zebrafish genes in r
selected the longest transcript if it had at least five stop
length. We removed genes that could not be assigned t
which teleost fishes did not form a monophyletic group i
genes for which the Ensembl gene tree indicated gene
include all ten teleost species of Ensembl v.78 (Suppler
genomes of Ensembl were used to calculate TBLASTN
using the BLAST+ v.2.2.29 suite of tools[47]. Exon-specif
orthologs were defined on the basis of this bitscore info
of the known orthologs had bitscores lower than this thr
than five remaining exons were discarded, which result
302 zebrafish genes that were then used as queries in
new teleost draft genome assemblies, the 10 Ensembl
sequence of salmon[48]. For each species, the best hits
their TBLASTN bitscores were above the exon-specific
Alignments of TBLASTN hits for the 2,251 exons were
nonsynonymous to synonymous substitutions (dN/dS)

on the number of congruent exon trees (Supplementary
evolution, we estimated the coefficient of variation of ra
v.2.2.0 (ref. 52) and removed the genes with the highes
(Supplementary Note). After this step, our strictly filtere
contained 567 exons from 111 genes, with a total alignn
data. To assess the consequences of strict filtering on p
maximum-likelihood phylogenies based on the strictly fi
7.3% missing data) with phylogenies based on a data s
strictly filtered (302 genes, 252,442 bp, 18.2% missing
both data sets were inferred with the software RAxML v
and Supplementary Note). The strictly filtered data set
teleost diversification with the software BEAST v.2.2 (re
calibration were calculated with the BEAST add-on Cla
diversification rates and the fossil sampling rate. The e
our phylogeny were identified and used to constrain the
calibration densities, taking into account the uncertainti
(Supplementary Note). We further used coalescent-bas
potentially misleading phylogenetic signal due to incom
conducted both with individual gene trees and with tree
to the binning approach of Mirarab *et al.*[55]. Maximum-li
maximum-clade-credibility trees resulting from BEAST
were used for species tree inference with the software
Fig. 2, Supplementary Table 5 and Supplementary Note
among the taxa included in our phylogeny, we compare
synapomorphic indels supporting each branch, followin

Rate shifts in MHC I copy number evolution.

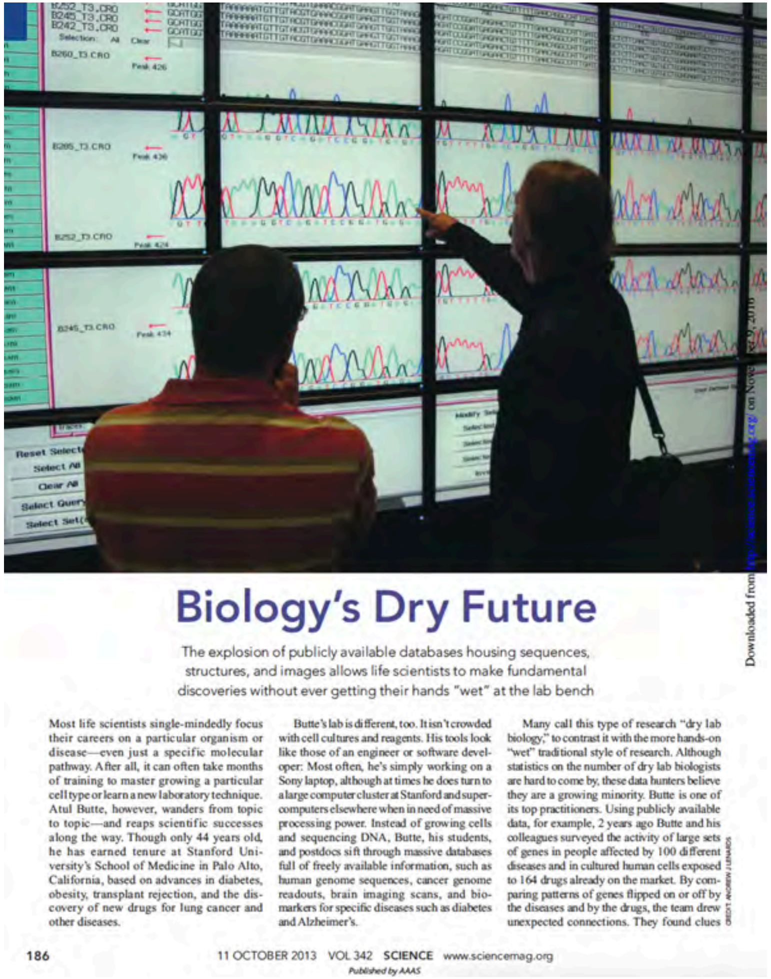
Phylogenetic signal in MHC I copy number evolution was assessed with Blomberg's *K* statistic[58], calculated using the phylosignal function of the picante R package v.1.6-2 (ref. 59), and with Pagel's lambda[60], calculated with function phylosig of the phytools R package v.0.4-45 (ref. 61) (Supplementary Note). The fits of four general models of trait evolution were compared on the basis of their sample-size-corrected Akaike information criterion (AICc), using the function fitContinuous of the geiger R package v.2.0.3 (ref. 62): a white noise model, a Brownian motion model, an early-burst model[63] and a single-peak Ornstein–Uhlenbeck model[40, 64] (Supplementary Note). The reversible-jump Bayesian approach of the bayou R package v.1.0.1 (ref. 65) was used to perform MCMC sampling of locations, magnitudes and numbers of shifts in multiple-optima Ornstein–Uhlenbeck models (Supplementary Fig. 5 and Supplementary Note). On the basis of the results of the bayou analysis, explicit hypotheses for shift combinations were tested in a likelihood framework, using the SLOUCH R package[41, 42]. For each shift combination, the likelihood of the best fitting combination of optimum, half-life and stationary variance was model comparison based on AICc scores (Supplementary Table 13 and The ancestral states of log-transformed MHC I copy numbers were nal nodes of the time-calibrated phylogeny, on the basis of the best fitting odel (Supplementary Note).

Diversification rate analyses.

Patterns of species diversification were analyzed with the Bayesian framework implemented in BAMM v.2.2.0 (ref. 66), on the basis of the time-calibrated phylogeny and counts of species richness in each of the 37 mutually exclusive clades of teleost fishes (Supplementary Table 14). The 'MEDUSA-like' model of diversification, assuming constant speciation and extinction rates within specific shift regimes[67], was used for this analysis (Supplementary Fig. 8 and Supplementary Note). To test whether high MHC I copy numbers are associated with lineages that have high diversification rates, we carried out BiSSE analyses[68] with the diversitree R package[69]. In these analyses, species were grouped into two categories for high and low MHC I copy numbers, on the basis of a given threshold value. Analyses were repeated for 26 equally spaced copy number threshold values between 10 and 60. As diversitree allows terminal clades with extant diversities of no more than 200 species, we used birth–death models of diversification in combination with the diversified sampling scheme of Höhna *et al.*[70] to stochastically resolve subclades of all clades with more than 200 extant species, which was repeated 25 times. BiSSE analyses were conducted for each of the 25 resulting phylogenies and with each of the 26 copy number thresholds, assuming symmetric transition rates between high and low copy numbers and identical extinction rates in taxa with high and low copy numbers (Supplementary Note and Supplementary Data).

None of the authors are informaticians (all are biologists)…

No more wet lab biology?



Biology's Dry Future

The explosion of publicly available databases housing sequences, structures, and images allows life scientists to make fundamental discoveries without ever getting their hands "wet" at the lab bench

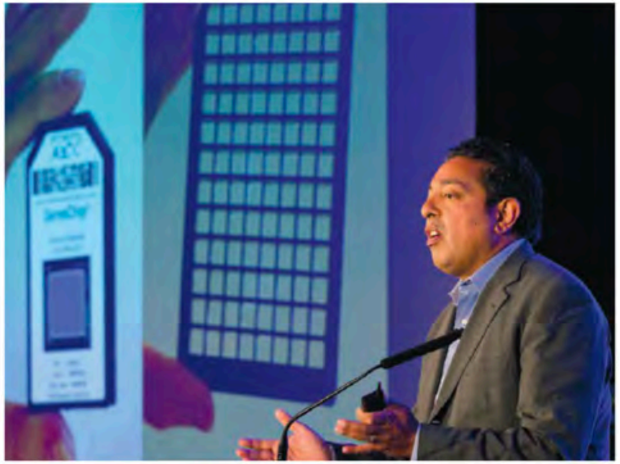
Most life scientists single-mindedly focus their careers on a particular organism or disease—even just a specific molecular pathway. After all, it can often take months of training to master growing a particular cell type or learn a new laboratory technique. Atul Butte, however, wanders from topic to topic—and reaps scientific successes along the way. Though only 44 years old, he has earned tenure at Stanford University's School of Medicine in Palo Alto, California, based on advances in diabetes, obesity, transplant rejection, and the discovery of new drugs for lung cancer and other diseases.

Butte's lab is different, too. It isn't crowded with cell cultures and reagents. His tools look like those of an engineer or software developer: Most often, he's simply working on a Sony laptop, although at times he does turn to a large computer cluster at Stanford and supercomputers elsewhere when in need of massive processing power. Instead of growing cells and sequencing DNA, Butte, his students, and postdocs sift through massive databases full of freely available information, such as human genome sequences, cancer genome readouts, brain imaging scans, and biomarkers for specific diseases such as diabetes and Alzheimer's.

Many call this type of research "dry lab biology," to contrast it with the more hands-on "wet" traditional style of research. Although statistics on the number of dry lab biologists are hard to come by, these data hunters believe they are a growing minority. Butte is one of its top practitioners. Using publicly available data, for example, 2 years ago Butte and his colleagues surveyed the activity of large sets of genes in people affected by 100 different diseases and in cultured human cells exposed to 164 drugs already on the market. By comparing patterns of genes flipped on or off by the diseases and by the drugs, the team drew unexpected connections. They found clues

CREDIT: ANDREW J LENARDS

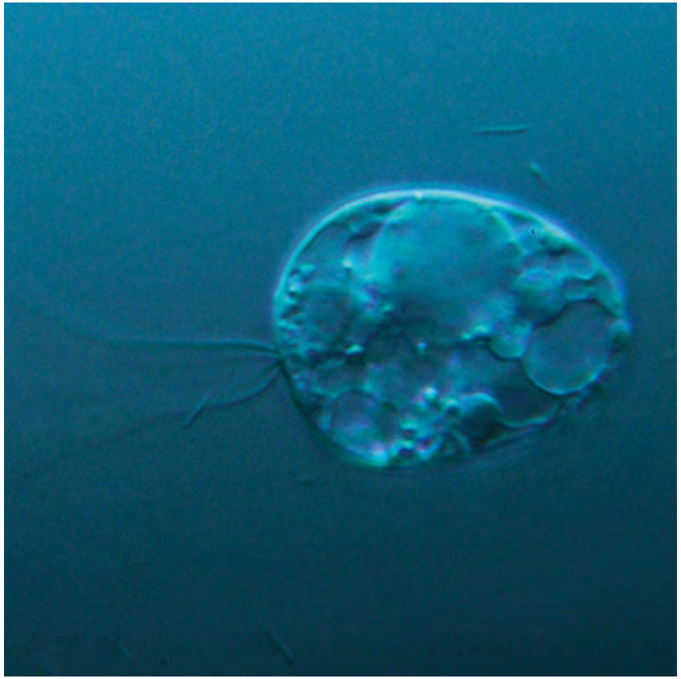
186 11 OCTOBER 2013 VOL 342 SCIENCE www.sciencemag.org
Published by AAAS

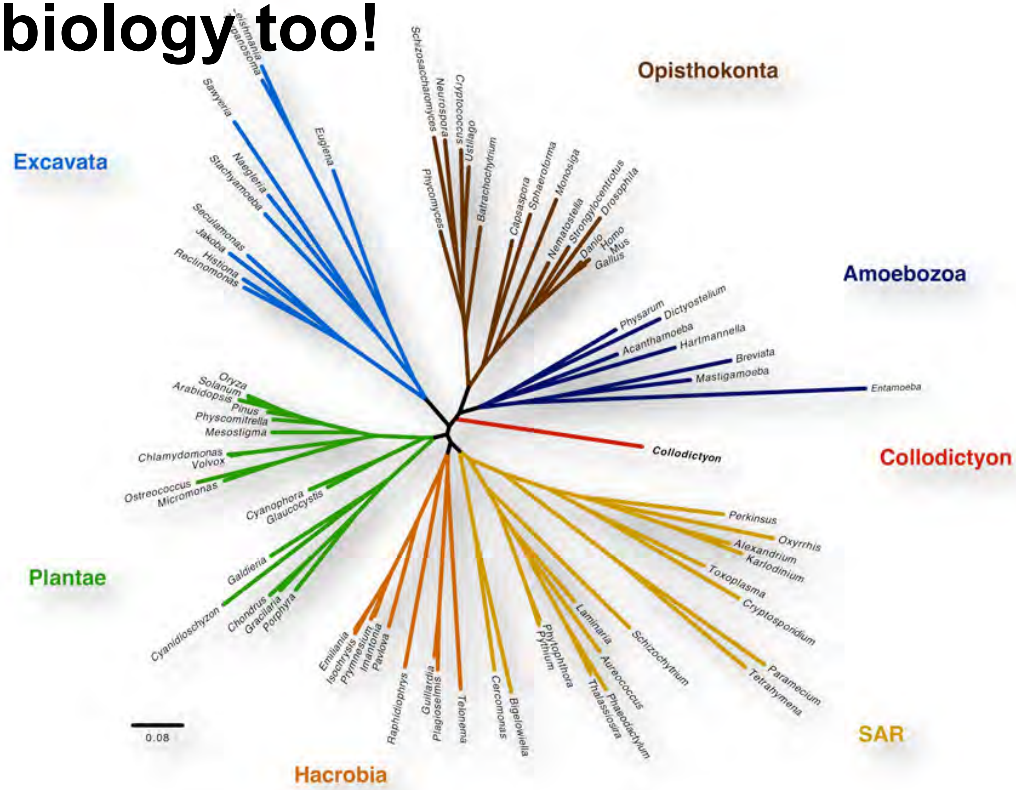


"I'm like a **kid in a candy store.** There is so much we can do."

—Atul Butte, Stanford University School of Medicine

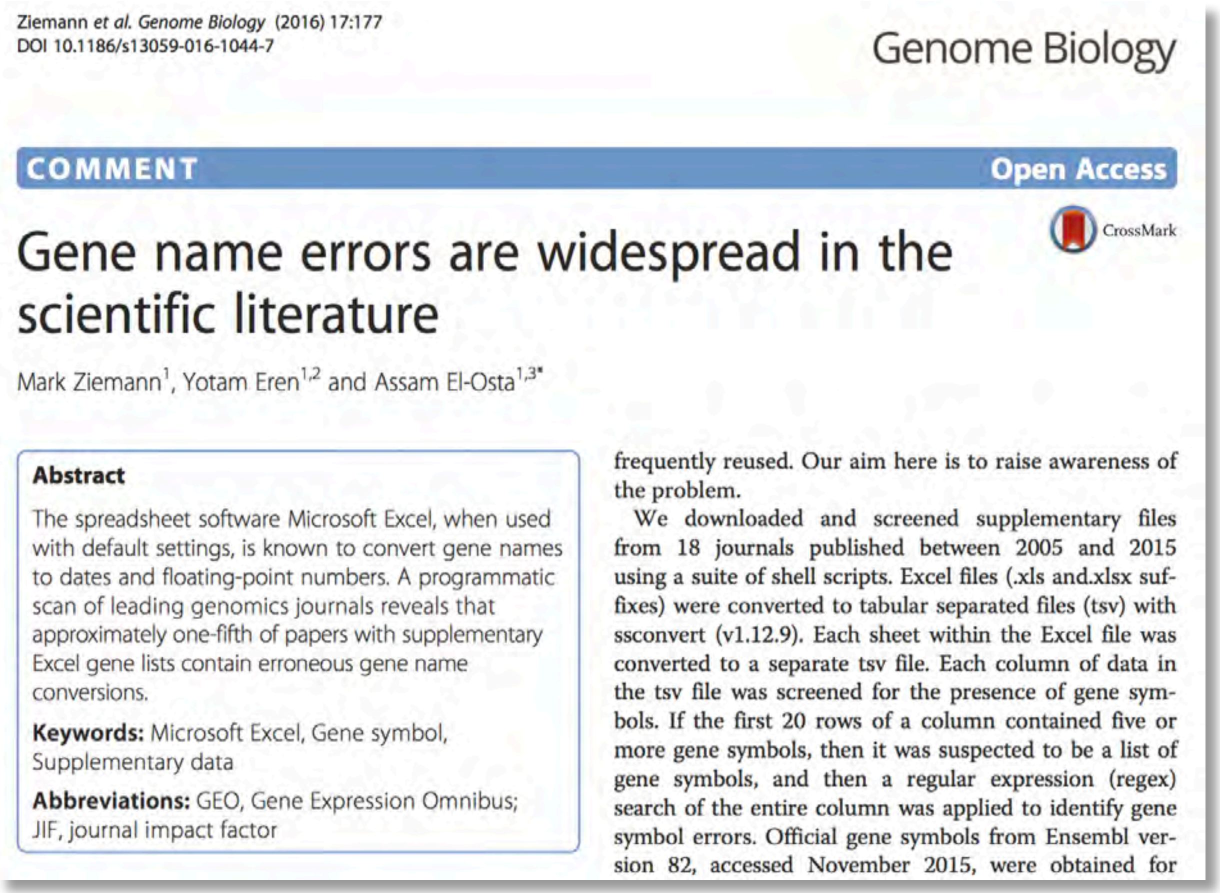
We need traditional biology too!





Collodictyon – collected in Ås in the 1980's. Placed on the Tree of Life in 2012.

Zhao et al. MBE. 2012



Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

CrossMark

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

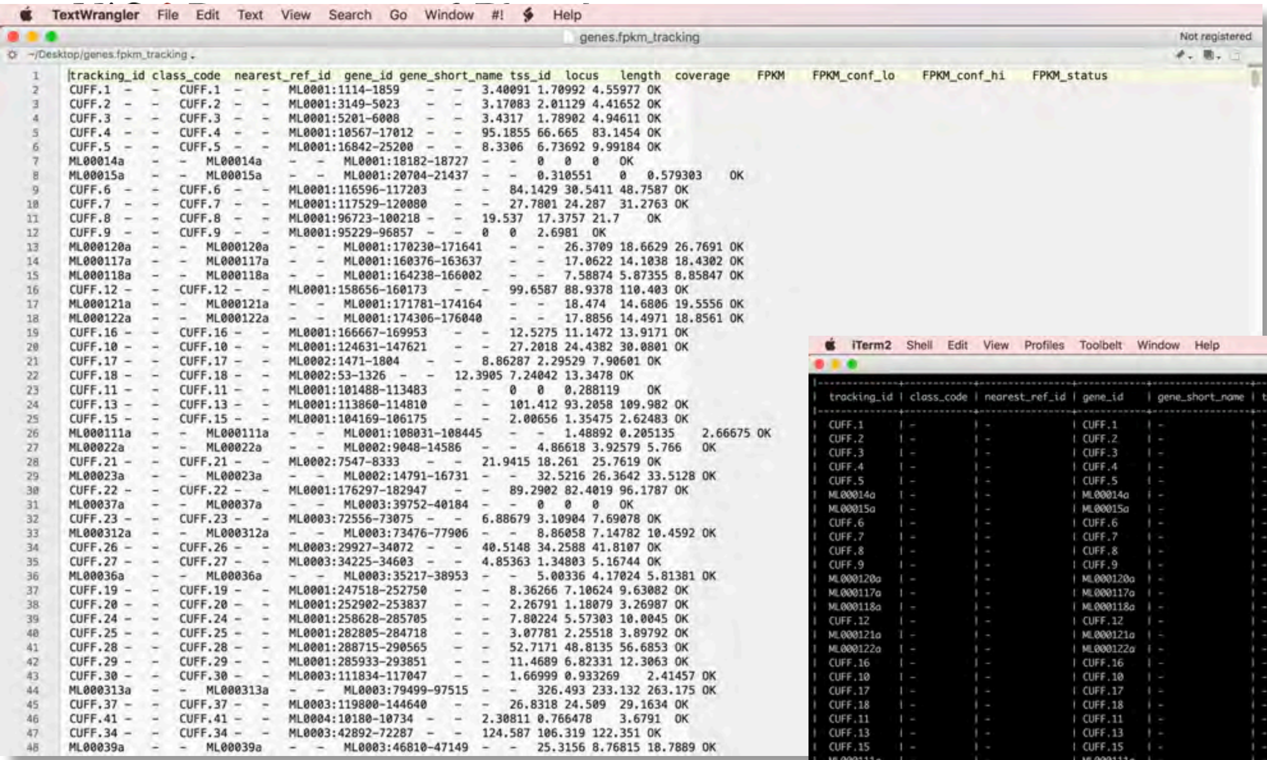
The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

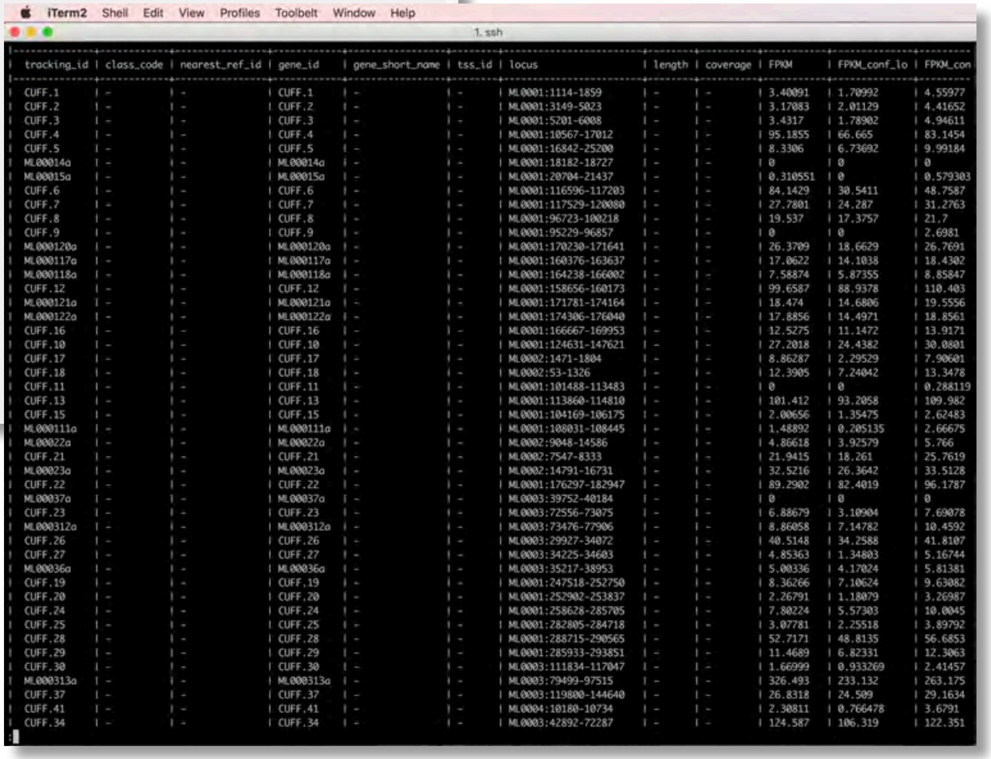
Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl version 82, accessed November 2015, were obtained for



```
tracking_id class_code  nearest_ref_id  gene_id gene_short_name tss_id  locus   length  coverage    FPKM    FPKM_conf_lo    FPKM_conf_hi    FPKM_status
CUFF.1  -   -   CUFF.1  -   -   ML0001:1114-1859    -   -   3.40091 1.70992 4.55977 OK
CUFF.2  -   -   CUFF.2  -   -   ML0001:3149-5023    -   -   3.17083 2.01129 4.41652 OK
CUFF.3  -   -   CUFF.3  -   -   ML0001:5201-6008    -   -   3.4317  1.78902 4.94611 OK
CUFF.4  -   -   CUFF.4  -   -   ML0001:10567-17012  -   -   95.1855 66.665  83.1454 OK
CUFF.5  -   -   CUFF.5  -   -   ML0001:16842-25200  -   -   8.3306  6.73692 9.99184 OK
ML00014a    -   -   ML00014a    -   -   ML0001:18182-18727  -   -   0   0   0   OK
ML00015a    -   -   ML00015a    -   -   ML0001:20704-21437  -   -   0.310551    0   0.579303    OK
CUFF.6  -   -   CUFF.6  -   -   ML0001:116596-117203    -   -   84.1429 30.5411 48.7587 OK
CUFF.7  -   -   CUFF.7  -   -   ML0001:117529-120080    -   -   27.7801 24.287  31.2763 OK
CUFF.8  -   -   CUFF.8  -   -   ML0001:96723-100218 -   -   19.537  17.3757 21.7    OK
CUFF.9  -   -   CUFF.9  -   -   ML0001:95229-96857  -   -   0   0   2.6981  OK
ML000120a   -   -   ML000120a   -   -   ML0001:170230-171641    -   -   26.3709 18.6629 26.7691 OK
ML000117a   -   -   ML000117a   -   -   ML0001:160376-163637    -   -   17.0622 14.1038 18.4302 OK
ML000118a   -   -   ML000118a   -   -   ML0001:164238-166002    -   -   7.58874 5.87355 8.85847 OK
CUFF.12 -   -   CUFF.12 -   -   ML0001:158656-160173    -   -   99.6587 88.9378 110.403 OK
ML000121a   -   -   ML000121a   -   -   ML0001:171781-174164    -   -   18.474  14.6806 19.5556 OK
ML000122a   -   -   ML000122a   -   -   ML0001:174306-176040    -   -   17.8856 14.4971 18.8561 OK
CUFF.16 -   -   CUFF.16 -   -   ML0001:166667-169953    -   -   12.5275 11.1472 13.9171 OK
CUFF.10 -   -   CUFF.10 -   -   ML0001:124631-147621    -   -   27.2018 24.4382 30.0801 OK
CUFF.17 -   -   CUFF.17 -   -   ML0002:1471-1804    -   -   8.86287 2.29529 7.90601 OK
CUFF.18 -   -   CUFF.18 -   -   ML0002:53-1326  -   -   12.3905 7.24042 13.3478 OK
CUFF.11 -   -   CUFF.11 -   -   ML0001:101488-113483    -   -   0   0   0.288119    OK
CUFF.13 -   -   CUFF.13 -   -   ML0001:113860-114810    -   -   101.412 93.2058 109.982 OK
CUFF.15 -   -   CUFF.15 -   -   ML0001:104169-106175    -   -   2.00656 1.35475 2.62483 OK
ML000111a   -   -   ML000111a   -   -   ML0001:108031-108445    -   -   1.48892 0.205135    2.66675 OK
ML00022a    -   -   ML00022a    -   -   ML0002:9048-14586   -   -   4.86618 3.92579 5.766   OK
CUFF.21 -   -   CUFF.21 -   -   ML0002:7547-8333    -   -   21.9415 18.261  25.7619 OK
ML00023a    -   -   ML00023a    -   -   ML0002:14791-16731  -   -   32.5216 26.3642 33.5128 OK
CUFF.22 -   -   CUFF.22 -   -   ML0001:176297-182947    -   -   89.2902 82.4019 96.1787 OK
ML00037a    -   -   ML00037a    -   -   ML0003:39752-40184  -   -   0   0   0   OK
CUFF.23 -   -   CUFF.23 -   -   ML0003:72556-73075  -   -   6.88679 3.10904 7.69078 OK
ML000312a   -   -   ML000312a   -   -   ML0003:73476-77906  -   -   8.86058 7.14782 10.4592 OK
CUFF.26 -   -   CUFF.26 -   -   ML0003:29927-34072  -   -   40.5148 34.2588 41.8107 OK
CUFF.27 -   -   CUFF.27 -   -   ML0003:34225-34603  -   -   4.85363 1.34803 5.16744 OK
ML00036a    -   -   ML00036a    -   -   ML0003:35217-38953  -   -   5.00336 4.17024 5.81381 OK
CUFF.19 -   -   CUFF.19 -   -   ML0001:247518-252750    -   -   8.36266 7.10624 9.63082 OK
CUFF.20 -   -   CUFF.20 -   -   ML0001:252902-253837    -   -   2.26791 1.18079 3.26987 OK
CUFF.24 -   -   CUFF.24 -   -   ML0001:258628-285705    -   -   7.80224 5.57303 10.0045 OK
CUFF.25 -   -   CUFF.25 -   -   ML0001:282805-284718    -   -   3.07781 2.25518 3.89792 OK
CUFF.28 -   -   CUFF.28 -   -   ML0001:288715-290565    -   -   52.7171 48.8135 56.6853 OK
CUFF.29 -   -   CUFF.29 -   -   ML0001:285933-293851    -   -   11.4689 6.82331 12.3063 OK
CUFF.30 -   -   CUFF.30 -   -   ML0003:111834-117047    -   -   1.66999 0.933269    2.41457 OK
ML000313a   -   -   ML000313a   -   -   ML0003:79499-97515  -   -   326.493 233.132 263.175 OK
CUFF.37 -   -   CUFF.37 -   -   ML0003:119800-144640    -   -   26.8318 24.509  29.1634 OK
CUFF.41 -   -   CUFF.41 -   -   ML0004:10180-10734  -   -   2.30811 0.766478    3.6791  OK
CUFF.34 -   -   CUFF.34 -   -   ML0003:42892-72287  -   -   124.587 106.319 122.351 OK
ML00039a    -   -   ML00039a    -   -   ML0003:46810-47149  -   -   25.3156 8.76815 18.7889 OK
```



```
|-------------+------------+----------------+-----------+-----------------+--------+----------------------+--------+----------+----------+--------------+---------
| tracking_id | class_code | nearest_ref_id | gene_id   | gene_short_name | tss_id | locus                | length | coverage | FPKM     | FPKM_conf_lo | FPKM_con
|-------------+------------+----------------+-----------+-----------------+--------+----------------------+--------+----------+----------+--------------+---------
| CUFF.1      | -          | -              | CUFF.1    | -               | -      | ML0001:1114-1859     | -      | -        | 3.40091  | 1.70992      | 4.55977
| CUFF.2      | -          | -              | CUFF.2    | -               | -      | ML0001:3149-5023     | -      | -        | 3.17083  | 2.01129      | 4.41652
| CUFF.3      | -          | -              | CUFF.3    | -               | -      | ML0001:5201-6008     | -      | -        | 3.4317   | 1.78902      | 4.94611
| CUFF.4      | -          | -              | CUFF.4    | -               | -      | ML0001:10567-17012   | -      | -        | 95.1855  | 66.665       | 83.1454
| CUFF.5      | -          | -              | CUFF.5    | -               | -      | ML0001:16842-25200   | -      | -        | 8.3306   | 6.73692      | 9.99184
| ML00014a    | -          | -              | ML00014a  | -               | -      | ML0001:18182-18727   | -      | -        | 0        | 0            | 0
| ML00015a    | -          | -              | ML00015a  | -               | -      | ML0001:20704-21437   | -      | -        | 0.310551 | 0            | 0.579303
| CUFF.6      | -          | -              | CUFF.6    | -               | -      | ML0001:116596-117203 | -      | -        | 84.1429  | 30.5411      | 48.7587
| CUFF.7      | -          | -              | CUFF.7    | -               | -      | ML0001:117529-120080 | -      | -        | 27.7801  | 24.287       | 31.2763
| CUFF.8      | -          | -              | CUFF.8    | -               | -      | ML0001:96723-100218  | -      | -        | 19.537   | 17.3757      | 21.7
| CUFF.9      | -          | -              | CUFF.9    | -               | -      | ML0001:95229-96857   | -      | -        | 0        | 0            | 2.6981
| ML000120a   | -          | -              | ML000120a | -               | -      | ML0001:170230-171641 | -      | -        | 26.3709  | 18.6629      | 26.7691
| ML000117a   | -          | -              | ML000117a | -               | -      | ML0001:160376-163637 | -      | -        | 17.0622  | 14.1038      | 18.4302
| ML000118a   | -          | -              | ML000118a | -               | -      | ML0001:164238-166002 | -      | -        | 7.58874  | 5.87355      | 8.85847
| CUFF.12     | -          | -              | CUFF.12   | -               | -      | ML0001:158656-160173 | -      | -        | 99.6587  | 88.9378      | 110.403
| ML000121a   | -          | -              | ML000121a | -               | -      | ML0001:171781-174164 | -      | -        | 18.474   | 14.6806      | 19.5556
| ML000122a   | -          | -              | ML000122a | -               | -      | ML0001:174306-176040 | -      | -        | 17.8856  | 14.4971      | 18.8561
| CUFF.16     | -          | -              | CUFF.16   | -               | -      | ML0001:166667-169953 | -      | -        | 12.5275  | 11.1472      | 13.9171
| CUFF.10     | -          | -              | CUFF.10   | -               | -      | ML0001:124631-147621 | -      | -        | 27.2018  | 24.4382      | 30.0801
| CUFF.17     | -          | -              | CUFF.17   | -               | -      | ML0002:1471-1804     | -      | -        | 8.86287  | 2.29529      | 7.90601
| CUFF.18     | -          | -              | CUFF.18   | -               | -      | ML0002:53-1326       | -      | -        | 12.3905  | 7.24042      | 13.3478
| CUFF.11     | -          | -              | CUFF.11   | -               | -      | ML0001:101488-113483 | -      | -        | 0        | 0            | 0.288119
| CUFF.13     | -          | -              | CUFF.13   | -               | -      | ML0001:113860-114810 | -      | -        | 101.412  | 93.2058      | 109.982
| CUFF.15     | -          | -              | CUFF.15   | -               | -      | ML0001:104169-106175 | -      | -        | 2.00656  | 1.35475      | 2.62483
| ML000111a   | -          | -              | ML000111a | -               | -      | ML0001:108031-108445 | -      | -        | 1.48892  | 0.205135     | 2.66675
| ML00022a    | -          | -              | ML00022a  | -               | -      | ML0002:9048-14586    | -      | -        | 4.86618  | 3.92579      | 5.766
| CUFF.21     | -          | -              | CUFF.21   | -               | -      | ML0002:7547-8333     | -      | -        | 21.9415  | 18.261       | 25.7619
| ML00023a    | -          | -              | ML00023a  | -               | -      | ML0002:14791-16731   | -      | -        | 32.5216  | 26.3642      | 33.5128
| CUFF.22     | -          | -              | CUFF.22   | -               | -      | ML0001:176297-182947 | -      | -        | 89.2902  | 82.4019      | 96.1787
| ML00037a    | -          | -              | ML00037a  | -               | -      | ML0003:39752-40184   | -      | -        | 0        | 0            | 0
| CUFF.23     | -          | -              | CUFF.23   | -               | -      | ML0003:72556-73075   | -      | -        | 6.88679  | 3.10904      | 7.69078
| ML000312a   | -          | -              | ML000312a | -               | -      | ML0003:73476-77906   | -      | -        | 8.86058  | 7.14782      | 10.4592
| CUFF.26     | -          | -              | CUFF.26   | -               | -      | ML0003:29927-34072   | -      | -        | 40.5148  | 34.2588      | 41.8107
| CUFF.27     | -          | -              | CUFF.27   | -               | -      | ML0003:34225-34603   | -      | -        | 4.85363  | 1.34803      | 5.16744
| ML00036a    | -          | -              | ML00036a  | -               | -      | ML0003:35217-38953   | -      | -        | 5.00336  | 4.17024      | 5.81381
| CUFF.19     | -          | -              | CUFF.19   | -               | -      | ML0001:247518-252750 | -      | -        | 8.36266  | 7.10624      | 9.63082
| CUFF.20     | -          | -              | CUFF.20   | -               | -      | ML0001:252902-253837 | -      | -        | 2.26791  | 1.18079      | 3.26987
| CUFF.24     | -          | -              | CUFF.24   | -               | -      | ML0001:258628-285705 | -      | -        | 7.80224  | 5.57303      | 10.0045
| CUFF.25     | -          | -              | CUFF.25   | -               | -      | ML0001:282805-284718 | -      | -        | 3.07781  | 2.25518      | 3.89792
| CUFF.28     | -          | -              | CUFF.28   | -               | -      | ML0001:288715-290565 | -      | -        | 52.7171  | 48.8135      | 56.6853
| CUFF.29     | -          | -              | CUFF.29   | -               | -      | ML0001:285933-293851 | -      | -        | 11.4689  | 6.82331      | 12.3063
| CUFF.30     | -          | -              | CUFF.30   | -               | -      | ML0003:111834-117047 | -      | -        | 1.66999  | 0.933269     | 2.41457
| ML000313a   | -          | -              | ML000313a | -               | -      | ML0003:79499-97515   | -      | -        | 326.493  | 233.132      | 263.175
| CUFF.37     | -          | -              | CUFF.37   | -               | -      | ML0003:119800-144640 | -      | -        | 26.8318  | 24.509       | 29.1634
| CUFF.41     | -          | -              | CUFF.41   | -               | -      | ML0004:10180-10734   | -      | -        | 2.30811  | 0.766478     | 3.6791
| CUFF.34     | -          | -              | CUFF.34   | -               | -      | ML0003:42892-72287   | -      | -        | 124.587  | 106.319      | 122.351
:
```

File naming

NO

```
myabstract.docx
Joe’s Filenames Use Spaces and Punctuation.xlsx
figure 1.png
fig 2.png
JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt
```

YES

```
2014-06-08_abstract-for-sla.docx
joes-filenames-are-getting-better.xlsx
fig01_scatterplot-talk-length-vs-interest.png
fig02_histogram-talk-attendance.png
1986-01-28_raw-data-from-challenger-o-rings.txt
```

https://rawgit.com/Reproducible-Science-Curriculum/rr-organization1/master/organization-01-slides.html#1

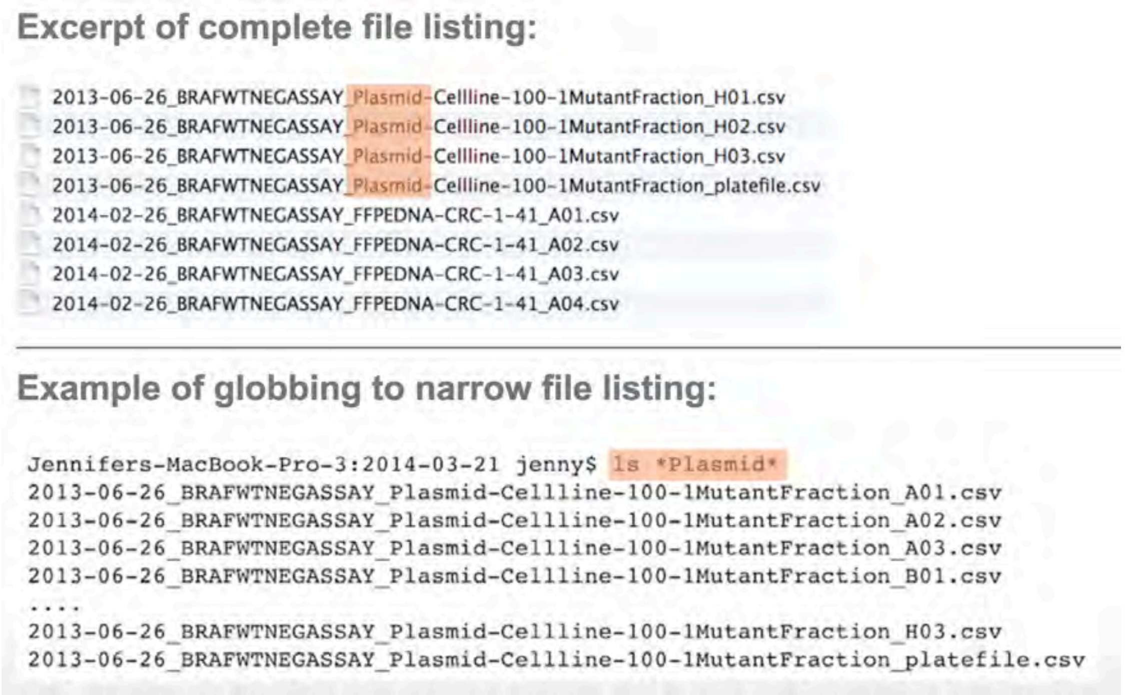
File naming

Three principles for filenames

- Machine readable (no spaces, consistency in naming, prefix, suffix, punctuation)
- Human readable (name contains info on content)
- Plays well with default ordering (numbers first, ISO 8601 standard for dates, left pad with zeroes)

https://rawgit.com/Reproducible-Science-Curriculum/rr-organization1/master/organization-01-slides.html#1

File naming



Excerpt of complete file listing:

```
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv
```

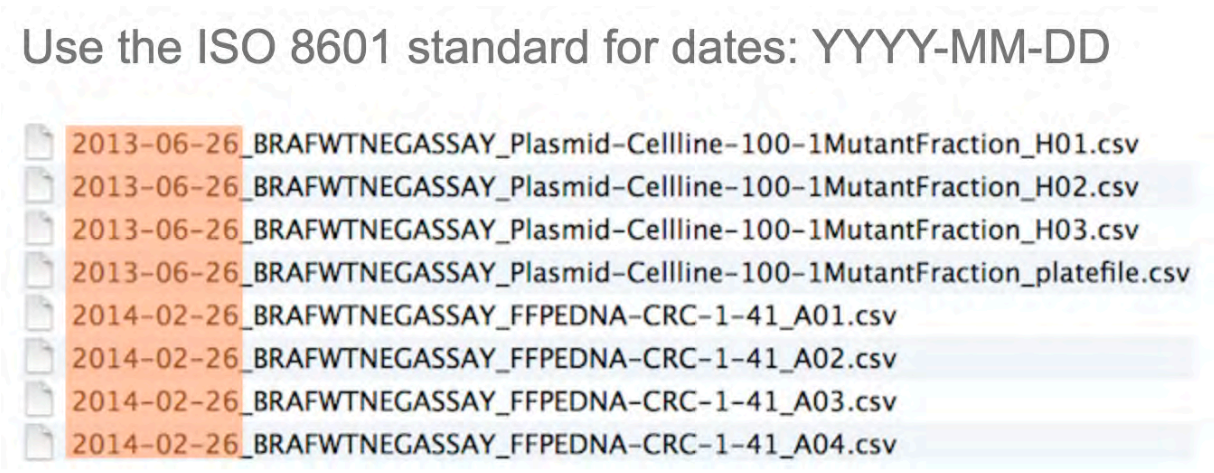
Example of globbing to narrow file listing:

```
Jennifers-MacBook-Pro-3:2014-03-21 jenny$ ls *Plasmid*
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B01.csv
....
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
```

https://rawgit.com/Reproducible-Science-Curriculum/rr-organization1/master/organization-01-slides.html#1

File naming

For sorting chronologically:



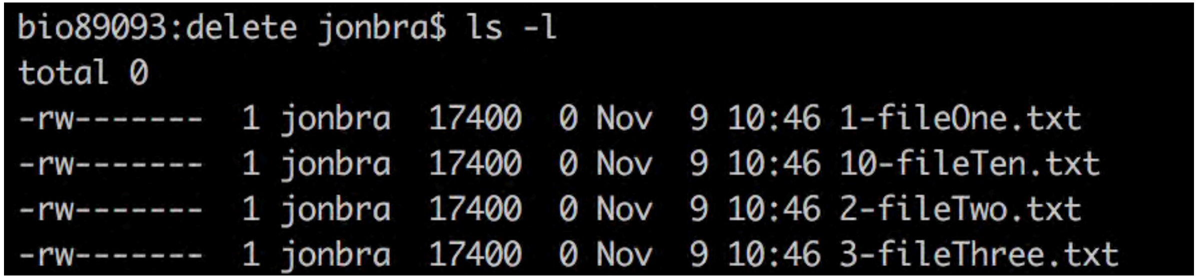
Use the ISO 8601 standard for dates: YYYY-MM-DD

```
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv
```

https://rawgit.com/Reproducible-Science-Curriculum/rr-organization1/master/organization-01-slides.html#1

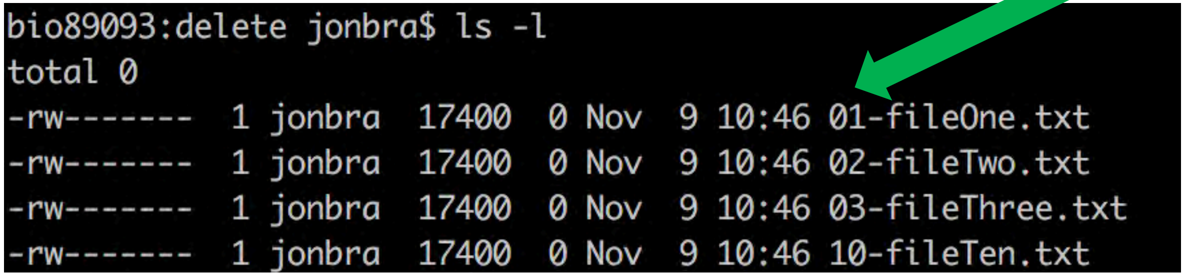
File naming

For sorting logically:



```
bio89093:delete jonbra$ ls -l
total 0
-rw-------  1 jonbra  17400  0 Nov  9 10:46 1-fileOne.txt
-rw-------  1 jonbra  17400  0 Nov  9 10:46 10-fileTen.txt
-rw-------  1 jonbra  17400  0 Nov  9 10:46 2-fileTwo.txt
-rw-------  1 jonbra  17400  0 Nov  9 10:46 3-fileThree.txt
```

Left pad with zeroes



```
bio89093:delete jonbra$ ls -l
total 0
-rw-------  1 jonbra  17400  0 Nov  9 10:46 01-fileOne.txt
-rw-------  1 jonbra  17400  0 Nov  9 10:46 02-fileTwo.txt
-rw-------  1 jonbra  17400  0 Nov  9 10:46 03-fileThree.txt
-rw-------  1 jonbra  17400  0 Nov  9 10:46 10-fileTen.txt
```

https://rawgit.com/Reproducible-Science-Curriculum/rr-organization1/master/organization-01-slides.html#1

Open science (Lex snakker ikke så mye om dette)

Passer kanskje ikke her, men vil gjerne få inn noe om dette. Kanskje senere når de har lært litt mer?

Jeg vil gjerne lære de litt ”good practice” og muligheter for open science.

- Markdown?
- Raw data as read only
- Tools and programs to use (text editors and not word etc.)
- Filnavn?
- GitHub?

Quote from Wikipedia

Budskapet her er at Bioinformatikk ikke er én ting, og at det på mange måter er alt. (mao et ubrukelig begrep).