Building Genomic Analysis Pipelines in a Hackathon Setting with Bioinformatician Teams: DNA-seq, Epigenomics, Metagenomics and RNA-seq

Ben Busby[1*], Allissa Dillman[2], Claire L. Simpson[3], Ian Fingerman[1], Sijung Yun[4], David M. Kristensen[1], Lisa Federer[5], Naisha Shah[6], Matthew C. LaFave[7], Laura Jimenez-Barron[8-9], Manjusha Pande[10], Wen Luo[11], Brendan Miller[12], Cem Meydan[13], Dhruva Chandramohan[13-14], Kipper Fletez-Brant[15-16], Paul W. Bible[17], Sergej Nowoshilow[18], Alfred Chan[19], Eric JC Galvez[20], Jeremy Chignell[21], Joseph N. Paulson[22-23], Manoj Kandpal[24], Suhyeon Yoon[25], Esther Asaki[26-27], Abhinav Nellore[15,28], Adam Stine[1], Robert Sanders[1], Jesse Becker[1], Matt Lesko[1], Mordechai Abzug[1], Eugene Yaschenko[1]

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America
[2] Surgery, Center for Prostate Disease Research, Uniformed Services University of the Health Sciences, Bethesda, Maryland, United States of America
[3] Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America
[4] Laboratory of Cell Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland, United States of America
[5] NIH Library, Division of Library Services, Office of Research Services, National Institutes of Health, Bethesda, Maryland, United States of America
[6] Systems Genomics and Bioinformatics Unit, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, United States of America
[7] Translational and Functional Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America
[8] Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America
[9] Centro de Ciencias Genomicas, Universidad Nacional Autonoma de Mexico, Cuernavaca, Morelos, Mexico
[10] Bioinformatics Core, University of Michigan, Ann Arbor, Michigan
[11] Cancer Genomics Research Laboratory, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Gaithersburg, Maryland, United States of America
[12] Department of Biology, Johns Hopkins University, Baltimore, Maryland, United States of America
[13] Institute for Computational Biomedicine, Weill Cornell Medical College, New York, New York, United States of America
[14] Tri-Institutional Training Program in Computational Biology and Medicine, New York, New York, United States of America
[15] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America
[16] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America
[17] Laboratory of Skin Biology, National Institute of Arthritis and Musculoskeletal and Skin

46   Diseases, National Institutes of Health, Bethesda, Maryland, United States of America

47   [18] Center for Regenerative Therapies, Technische Universität Dresden, Dresden, Germany

48   [19] Translational Immunology, John Wayne Cancer Institute at Saint John's Health Center, Santa

49   Monica, California, United States of America

50   [20] Microbial Immune Regulation Group, Helmholtz Centre for Infection Research,

51   Braunschweig, Germany

52   [21] Chemical and Biological Engineering, Colorado State University, Fort Collins, Colorado,

53   United States of America

54   [22] Graduate Program in Applied Mathematics & Statistics, and Scientific Computation,

55   University of Maryland, College Park, Maryland, United States of America

56   [23] Center for Bioinformatics and Computational Biology, University of Maryland, College Park,

57   Maryland, United States of America

58   [24] Division of Health and Biomedical Informatics, Department of Preventive Medicine, Feinberg

59   School of Medicine, Northwestern University, Chicago, Illinois, United States of America

60   [25] Genetics and Molecular Biology Branch, National Human Genome Research Institute,

61   National Institutes of Health, Bethesda, Maryland, United States of America

62   [26] Bioinformatics and Molecular Analysis Section, Center for Information Technology, National

63   Institutes of Health, Bethesda, Maryland, United States of America

64   [27] SRA International, Fairfax, Virginia, United States of America

65   [28] Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, United

66   States of America

67

68   *Corresponding author

69   Email: ben.busby@gmail.com (BB)

70

Abstract

72   We assembled teams of genomics professionals to assess whether we could rapidly

73   develop pipelines to answer biological questions commonly asked by biologists and others new

74   to bioinformatics by facilitating analysis of high-throughput sequencing data. In January 2015,

75   teams were assembled on the National Institutes of Health (NIH) campus to address questions in

76   the DNA-seq, epigenomics, metagenomics and RNA-seq subfields of genomics. The only two

77   rules for this hackathon were that either the data used were housed at the National Center for

78   Biotechnology Information (NCBI) or would be submitted there by a participant in the next six

79   months, and that all software going into the pipeline was open-source or open-use. Questions

80   proposed by organizers, as well as suggested tools and approaches, were distributed to

81   participants a few days before the event and were refined during the event. Pipelines were

82   published on GitHub, a web service providing publicly available, free-usage tiers for

83   collaborative software development (https://github.com/features/). The code was published at

84   https://github.com/DCGenomics/ with separate repositories for each team, starting with

85   hackathon_v001.

86

87   **Introduction**

88   Genomic analysis leverages large datasets generated by sequencing technologies in order

89   to gain better understanding of the genomes of humans and other species. Given its reliance on

90   large datasets with complex interactions and its fairly regularized metadata, genomic analysis is

91   an exemplar of "big data" science (1). Genomic analysis has shown great promise in finding

92   actionable variants for rare diseases (2), as well as directing more specific clinical action for

93   common diseases (3, 4).

94   Due to its potential for significant clinical and basic science discoveries, genomics has

95   drawn many newcomers from the biological and computational sciences, as well as investigators

96   from new graduate programs in bioinformatics. While many of these investigators can run

97   established pipelines on local, public, or combined datasets, most do not have the expertise or

98   resources to establish and validate novel pipelines. Additionally, highly experienced genomic

99   investigators often lack the resources necessary to generate and distribute pipelines with broader

100   applicability outside their specific area of research. In this study, we aimed to assess whether we

101   could close this gap by bringing genomics experts from around the world together to establish

102   public pipelines that can be both used by newcomers to genomics and refined by other seasoned

103   professionals.

104   We sought to achieve this by hosting a hackathon at the National Institutes of Health

105   (NIH) in Bethesda, Maryland. Hackathons are events in which individuals with expertise in

106   various areas of software development and science come together to collaborate intensively over

107   a period of several days, typically focusing on a specific goal or application. This form of

108   crowdsourcing facilitates innovative ideas and agile development of solutions for challenging

109   questions (5). Hackathon participants also benefit from the opportunity to learn new skills,

110   network with other professionals, and gain the personal satisfaction of using their expertise to

111   help the community.

112   Those of us who had previously attended hackathons had been motivated by learning

113   from the experience, personal curiosity, professional networking, and the satisfaction of applying

114   one's skills to help the community. These previously attended hackathons included

115   IT/programming-centric events, such as the Kaiser Permanente-sponsored hackathon for apps

116   that could help prevent and fight obesity (6); bioinformatics-oriented events, such as the

117   Illumina-sponsored hackathon to create apps in their proprietary BaseSpace cloud computing

118   environment (7); and events focused on social issues, such as exploring women's empowerment

119   and nutrition in the developing world (8).

120   Hackathons are more common in software development communities than in

121   bioinformatics and medicine, and may represent a valuable opportunity to accelerate biomedical

122   discovery and innovation. Hackathons have gained popularity in the bioinformatics community

123   in the last several years, though the environment and culture of a bioinformatics hackathon

124   differs from that of typical IT/programming-centric hackathons (9). Bioinformatic hackathons

125   more closely resemble a scientific discussion and provide an opportunity to learn and delve more

126   deeply into specific areas. Typically a successful hackathon requires participants with both

127   coding/programming skills and domain-specific knowledge (e.g. DNA-seq and RNA-seq).

128   Four teams of 4-6 participants came together for this hackathon to answer questions in

129   the fields of DNA-seq, RNA-seq, metagenomics and epigenomics. The initial topics for

130   exploration were based on questions bioinformaticians frequently ask computational cores.

131   Massive amounts of data have been generated and made publicly available in these fields using

132   high-throughput experimental technologies. Powerful computational pipelines are needed to

133   handle such large datasets, and to help analyze the data to answer biomedical questions. In

134   addition, these topics have seen a significant increase in publications in recent years, as

135   demonstrated in Figure 1.

136   **Fig. 1: Articles indexed in MEDLINE for topics related to each of the four teams.**

137   ## Background for Hackathon Team Goals

138   *DNA-seq Team*

139   The goal for the DNA-seq Team was to create an easy-to-use integrated pipeline using

140   existing tools to predict somatic variants from exome sequencing data, find shared and unique

141   variants between samples, and filter and annotate the variants. Somatic mutation calling is

142   particularly relevant to cancer genome characterization (10). Several methods exist to predict

143   somatic variants by interrogating genomic sequences of tumor-germline paired samples (11-13).

144   Issues with these variant calling algorithms are well known; each of the algorithms have

145   strengths and weaknesses, and are sensitive to the datasets used (14). One way to test the

146   reliability of the called variants and gain more confidence is to combine putative calls from

147   several algorithms and use strict filtering criteria. Thus, we wanted to create a pipeline that

148   included several existing calling algorithms. In addition, we intended to build a module within

149   the pipeline to predict eQTLs using the called somatic variants and RNA-seq data from the same

150   individuals. To achieve our goal of building an integrated pipeline, a paired exome sequence

151   dataset as well as an RNA sequence dataset were required.

152   *Epigenomics Team*

153   Computational cores contain an abundance of epigenetic data encompassing a wide

154   variety of markers over many different cell types. This abundance of information empowers labs

155   to infer how these different markers contribute to gene expression and chromatin state. However,

156   little standardization or collaboration exists among investigators seeking to elucidate

157   relationships among epigenetic modifiers, leading to inconsistencies between analyses. The

158   Epigenomics Team sought to analyze the extent to which DNA methylation and histone

159   modifications affect gene expression using regression models. In addition to modelling, we

160   wanted to create a framework for integrating experimental ChIP-seq or DNA methylation data

161   provided by users into models of gene expression that were informed by publicly available data.

162   A resource that allows users to make sense of their data in the context of the existing wealth of

163   public data on epigenetic regulation provided an attractive and worthwhile goal.

164   *Metagenomics Team*

165   The Metagenomics Team worked on the problem of identifying the presence of viral

166   sequences within human genomic or metagenomic sequences. The ability to locate these viral

167   sequences offers the prospect of using computational tools for diagnostic purposes (15). Viral

168   sequences may be embedded within human genomic sequence data as endogenous retroviruses

169   or associated with bacterial consortia of the human microbiome as either lytic phages or

170   prophages integrated into the bacterial genomes. Several studies have demonstrated the

171   association between human-specific endogenous retroviruses (HERV) and diseases such as

172   breast cancer (16, 17). In the context of the human microbiome, previous studies have used

173   metagenomics to describe differences in bacterial consortia (18), but only a few studies have

174   applied a metagenomics workflow to viral sequences present in human microbiome data (19,

175   20).

176   This team's aim was to develop and compare analytical pipelines to identify and quantify

177   viral sequences within human genomic and metagenomic datasets. The pipeline could be used in

178   the characterization of the viral community in understanding the role viruses play in various

179   environmental niches and diseases eventually associating differentially abundant viruses in

180   disease or phenotype, despite database and sparsity issues potentially using marker-gene

181   methodologies (21).

182   *RNA-seq Team*

183   RNA sequencing (RNA-seq) has become a useful technique for detecting gene

184   expression levels, alternate splicing, and gene fusions (22). RNA variant detection can be

185   important in cancer research to detect significant changes in tumor progression (23). The team's

186   initial goal was to build a variant calling pipeline using RNA-seq data from melanoma samples

187   to differentiate germline and somatic mutations from RNA editing. Some sample datasets from

188   the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA)

189   were suggested. Data could then be used to compare germline variants with known germline

190   variants from the dbGaP version of the 1000 Genomes Project and ClinVar (24, 25). Another

191   task was to determine isoform specificity based on mapping to 454 data using available software

192   for the correlation with RNA structure prediction. Data could also be compared with the NCI-60

193   cell line variants. Also, we could try to determine mosaicism with germ-line vs intratumor data.

194   Additional tasks were to detect systematic quality score variants indicating RNA editing with

195   Illumina and Pacbio reads, correlate with RNA structure prediction and to fix HTseq annotation

196   of tiny exons in UTRs.

Materials and Methods

Advertising, Preparation, and Logistics

199   Four team leads were selected by Busby to participate in the event. Team leads were

200   experts in particular areas and they suggested scientific areas, datasets, or tools that they were

201   familiar with. In order to recruit participants, we sent an announcement (Supporting Information

202   1) to contacts in Busby's network encouraging genomics professionals to apply for the

203   hackathon, as well as posting it on several NCBI social media outlets, such as Facebook, Twitter,

204   and Meetup.

205   After the application deadline, Busby reviewed the applicants for minimum necessary

206   experience, and sent the applicants' written statements to team leads, encouraging them to pick

207   five members and two alternates. Team leads reviewed the 131 qualified applicants and selected

208   a total of 27 people to fill the four teams of about 6 people each. Members were chosen solely on

209   the interest and motivational statements on their forms, and credentials and experience were not

210   further researched. One goal of the hackathon was to bring together scientists with different

211   expertise. Some of the participants were more traditional bench biologists and some were

212   bioinformaticists with more computational knowledge. This was a networking experience for

213   local and international scientists, including two who traveled from Germany and several from

214   outside of Maryland.

215   Technical professional staff from NCBI, including programmers and system

216   administrators, were also brought in and "embedded" with the team to facilitate programming

217   knowledge transfer and debugging. In addition, a librarian from the NIH Library served as an

218   editor, providing guidance on writing and organization of the manuscript.

219   The teams convened in a large meeting space that had Wi-Fi access, a nearby cafeteria,

220   plenty of tables to provide adequate room to spread out, and easels with markers for easier team

221   discussion. Space was available outside of the meeting room if participants needed a quiet space

222   apart from the larger group. To optimize time available to work on the projects, an hour was set

223   aside each day for a "working lunch" with the option to order in as a group. There was also an

224   optional group dinner each night that provided more time to network, socialize, and continue

225   work on the scientific problems.

226   The schedule was circulated the week prior to the hackathon. The agenda was semi-

227   structured with time allocated for various activities such as obtaining data, pipeline building, and

228   code checking. There were several checkpoints for each team to present their progress to the

229   larger group. A tour of the NCBI data center was conducted, as well as several brief

230   informational presentations. For example, Eugene Yaschenko of NCBI presented a group of

231   application program interfaces (APIs) relating to the SRA Toolkit collectively known as the

232   software development kit (26). The schedule was designed to give team members a common

233   starting point, but allowed for modification as needed for each team.

234   **Delegation of roles/responsibilities**

235   One of the first tasks at the hackathon was for each team to assign roles/responsibilities to

236   the team members. Most of the members were meeting each other for the first time, so it was a

237   good way to learn about each other's strengths. Roles were loosely distributed to cover areas

238   including:

239   • Systems administration: decided where data and tools went, implemented git for

240   versioning, and dealt with any technical issues;

241   • Metadata: tracked and recorded the information describing the contents of the datasets;

242   • Data Acquisition: located and downloaded appropriate publicly available datasets for

243   analysis;

244   • Data Normalization: prepared data from multiple datasets to work in a pipeline, such as

245   making read counts comparable between collection sites.

246   • Downstream Analysis: assigned annotation or function to genomic predictions and

247   looked at statistically meaningful overrepresentation

248   • Documentation: prepared code and text summaries, including drafting this manuscript.

Team Organization and Communication

250   In order to facilitate communication among team members, Google mail groups were

251   created for each team. Most team members did not know each other prior to the start of the

252   hackathon, but some teams used their Google mail group to communicate prior to the event.

253   Other collaboration tools included GitHub for program version control and Gitter, an instant

254   messaging tool, for sharing links, especially while teams were looking for datasets. Documents

255   were collaboratively edited using Google Docs.

Data Sources and Computational Tools

257   The basic workflow for the teams' activities is demonstrated in Fig. 2. Scripting and data

258   analysis took place in an Amazon Elastic Compute Cloud (EC2) where team players downloaded

259   data from NCBI repositories and used open source bioinformatic tools. System administrators

260   (SysAdmin) for each team were charged with code installation, compilation, storage and nodes

261   expansion while the teams were developing the pipelines, which were pushed to GitHub.

262   **Figure 2. Hackathon teams workflow.**

263   *DNA-seq Team*

264   The DNA-seq team acquired high-throughput sequencing data for matched tumor-normal

265   pairs from the NCBI SRA. The first three datasets found were unusable due to corrupted files,

266   mismatched samples, or missing header information (see Discussion below for more details).

267   The data used for designing the DNA-seq pipeline were found by searching the BioProject

268   database for the terms "homo sapiens NOT cell line," filtering for exome and SRA datasets, and

269   rejecting the 81 datasets that matched the term "HapMap".

270   We designed our pipeline around a publicly available exome dataset submitted to SRA,

271   BioProject PRJNA268172 (27). It consisted of exome sequences from four meningioma samples

272   and a peripheral blood DNA sample from a 61-year-old female suffering from sporadic multiple

273   meningiomas. Meningiomas are tumors originating from the membranous layers surrounding the

274   central nervous system and are generally considered benign. Malignant meningiomas, while rare,

275   are associated with a higher risk of local tumor recurrence and have a median survival time of

276   less than two years (28). We were interested in finding and comparing somatic mutations,

277   specifically SNPs, found in each of the meningioma tumor samples from the patient.

278   We downloaded the relevant files by using the SRA Toolkit, and converted the SRA files

279   to SAM and BAM files for further analysis. We found that the SAM files contained reads with

280   sequence and quality scores of different length, which would halt the BAM file conversion, so

281   we added a filter in our pipeline to remove such reads. Ideally, a module for trimming/masking

282   and re-aligning the FASTQ file would have been the most appropriate; however, due to time

283   restrictions we were unable to add such a module to our pipeline. Here, we assumed that the

284   SRA submitted files were properly aligned and followed appropriate quality control steps.

285   We used five different algorithms to call somatic variants. Four of these were in the Cake

286   pipeline (29), namely Bambino (11), CaVEMan (30), SAMtools mpileup (31), and VarScan2

287   (12). To further increase our confidence in the called variants, we added the MuTect algorithm

288   (13) to those used by Cake. Xu et al have compared somatic variation calling algorithms (14).

289   A first level of filtering is provided in the pipeline for the resulted VCF files containing

290   called somatic variants. This was accomplished using the Cake filtering module. We kept most

291   of the default parameters, and those that were changed are explained in Table 1. Once the VCF

292   files were filtered, they would be annotated using the ANNOVAR software, downloaded on

293   January 5th, 2015 (32), with five different databases (RefGene, KnownGene, ClinVar, Cosmid

294   and Cosmic), and the top 1% most deleterious CADD scores.

295   **Table 1: Cake filtering module parameters modified by the DNA-seq Team.**

Parameter	Explanation	Flag or Value Used
EXONIC_FILTER	Exome data is being used	FALSE
INDEL_FILTER	Only SNPs are being considered	FALSE
COSMIC_ANNOTATION_FLAG	Cosmic Annotations are added with ANNOVAR	FALSE
TUMOR_MIN_DEPTH	As the coverage of these samples was too low, we set the threshold to a less stringent value.	5
NORMAL_MIN_DEPTH	As the coverage of these samples was too low, we set the threshold to a less stringent value.	5
EMPTY_CONSEQ_FILTER	VEP annotations were not used	FALSE

296

297   To compare the predicted somatic variations between two matched tumor-control

298   samples, we created a module in our pipeline that used VCFtools to find shared and unique

299   variations (33). Lastly, we combined these modules into a single pipeline using Bpipe (34), as

300   well as a Unix Bash script. The full list of software used by the DNA-seq Team is included in

301   Table 2.

302   **Table 2: Tools and software used by the DNA-seq Team.**

Software/Tool	URL
SRA tool kit	http://www.ncbi.nlm.nih.gov/sra
Cake	http://sourceforge.net/projects/cakesomatic/
MuTect	https://github.com/broadinstitute/mutect/releases/tag/1.1.5
somatic sniper	http://gmt.genome.wustl.edu/packages/somatic-sniper/install.html
varscan	http://sourceforge.net/projects/varscan/
bambino	https://cgwb.nci.nih.gov/goldenPath/bamview/documentation/index.html
picard	http://broadinstitute.github.io/picard/
Annovar	http://www.openbioinformatics.org/cgi-bin/annovar_download.cgi
SAMtools	http://www.htslib.org/
bpipe	http://github.com/ssadedin/bpipe
Python	https://www.python.org/
GNU Parallel	http://ftp.gnu.org/gnu/parallel/ (35)
VCFtools	http://vcftools.sourceforge.net

303

304   *Epigenomics Team*

305   We gathered data with the intention of modeling transcription (mRNA-seq) based on

306   DNA methylation (RRBS or Bisulfite-seq) and histone states (ChIP-seq). To simplify analysis,

307   we focused on marks associated with enhancers and their regulatory status: H3K27ac,

308   H3K4me1, and H3K4me3. Ultimately, we required that included tissues have matching

309   H3K27ac, RNA-seq and DNA methylation data for preliminary modeling. Data files were drawn

310   from human cell lines and tissues in the NIH Epigenomic Roadmap that fit our criteria (outlined

311   in Table 3), from the site's FTP mirror (36) using the rsync command with the -av option. All

312   files were already aligned, and were required to have been generated using the hg19 reference

313   genome. Several files were identified that appeared to be re-aligned or uncorrected versions of

314   other downloaded files, and were removed from the analysis. Tissues acceptable for analysis

315   were identified by using the data matrix view on the Roadmap site, as well as searching for non-

316   partial datasets with the Data Grid view of the International Human Epigenome Consortium

317   (IHEC) Data Portal (37). Samples were initially downloaded from ENCODE: Encyclopedia of

318   DNA Elements (38), but were not included in initial analysis for reasons of uniformity and time.

319   **Table 3: NIH Epigenomic Roadmap Cell Lines/Tissues used by Epigenomics Team**

Cell/Tissue Type	
H1	Left_Ventricle
IMR90	Pancreas
CD34_mobilized_primary_cells	Right_Atrium
hESC-derived_CD184+_endoderm_cultured_cells	Sigmoid_Colon
HUES64_cell_line	Spleen
pancreatic_islets	Penis_Foreskin_Fibroblast_Primary_Cells
Skeletal_muscle	Penis_Foreskin_Keratinocyte_Primary_Cells
Adrenal_Gland	Penis_Foreskin_Melanocyte_Primary_Cells
Aorta	Neurosphere_Cultured_Cells_Ganglionic_Eminence_Derived
Esophagus	Brain_Hippocampus_Middle
Gastric	Ovary

320

321   *Metagenomics Team*

322   We designed six pipelines to compare different strategies for identifying and quantifying

323   viral sequences among human genomic information. The pipelines are:

324   1. Identify and quantify HERV sequences in assembled reads using blastn,

325   2. Identify and quantify HERV sequences in non-assembled reads from a human genome

326   using search tools from NCBI's SRA Toolkit,

327   3. Identify and quantify HERV sequences in non-assembled reads from a human genome

328   using a standard blastn search of reads in FASTA format.

329   Pipelines 4-6 repeat pipelines 1-3 to identify all viral sequences within a sample from a human

330   bacterial microbiome.

331   For pipeline 1, whole genome sequence raw reads of human CEU NA12878 (39) were

332   obtained from NCBI's SRA database and converted from SRA to FASTQ file format using the

333   fastq-dump command provided by the SRA Toolkit with default filter settings. The resulting

334   FASTQ file was moved to an Amazon Elastic Compute Cloud (EC2) node and assembled into

335   contigs with the ABySS assembler (40, 41). Contigs were used as queries for blastn against a

336   database consisting of all *Retroviridae* RefSeq genomes that was constructed using the

337   makeblastdb command. For pipeline 2, the SRA files containing raw reads for NA12878 were

338   used directly as a database for a SRA blast (blastn_vdb) using the *Retroviridae* RefSeq genomes

339   as query. For pipeline 3, the FASTA file of non-assembled reads from pipeline 1 was used as a

340   query for a blastn search against the database of *Retroviridae* RefSeq genomes from pipeline 1.

341   For pipelines 4 – 6, we used samples from the Human Microbiome Project (HMP) that

342   had passed preliminary quality checks (42). The plan for pipeline 4 was the same as for pipeline

343   1, except for using SOAPdenovo2 for the assembly of the microbiome FASTQ reads. For

344   pipeline 5, rather than start with SRA files, raw Illumina WGS reads in FASTQ format first were

345   converted to SRA format using the FASTQ loader tool latf-load within the SRA Toolkit.

346   Pipeline 6 follows pipeline 3, but uses the microbiome FASTA as its query and the total viral

347   RefSeq genome as its database. Full details about the data and tools we used are located in

348   Tables 4 and 5, respectively.

349   **Table 4. Data sources used by Metagenomics Team**

350

Data Source	FTP Location
refseq viral	ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral
1000 genomes contigs	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/
Human microbiome	ftp://public-ftp.hmpdacc.org/HMASM/PGAs/
Human microbiome raw reads for right_retroauricular_crease	ftp://public-ftp.hmpdacc.org/Illumina/right_retroauricular_crease/SRS015381.tar.bz2
Homo_sapiens_dna	ftp://ftp.ensembl.org/pub/current_fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.10.fa.gz

351

352   **Table 5: Tools and software used by the Metagenomics Team.**

Software/Tool	URL
SRAtoolkit	http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software
ABySS (from source, version 1.5.2)	http://www.bcgsc.ca/platform/bioinfo/software/abyss/releases/1.5.2
R statistical package	http://www.r-project.org/
BLAST+	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

353

354   *RNA-seq Team*

355   The RNA-seq Team determined that the samples initially suggested by the team lead

356   were not appropriate, so we looked for paired tumor/normal datasets from the NCBI Sequence

357   Read Archive (SRA). One dataset was a deep high-throughput transcriptome sequencing (RNA-

358   seq) performed on three pairs of matched tumor and adjacent non-tumors (NT) tissues from HCC

359   patients of Chinese origin, accession PRJNA149267 (43, 44) Another identified dataset was a

360   study to identify a prognostic signature in colorectal cancer (CRC) patients with diverse

361   progression and heterogeneity of CRCs, accession PRJNA218851 (45, 46). Thirty-six paired

362   samples from this study (18 tumor samples: GSM1228184-GSM1228201 and 18 matched

363   normal samples GSM1228202-GSM1228219) were also determined to suitable. Additional tools

364   used by the RNA-seq Team are listed in Table 6.

365   **Table 6: Tools and software used by the RNA-seq Team.**

Software/Tool	URL
HISAT	http://www.ccb.jhu.edu/software/hisat/index.shtml
Illumina iGenome	http://support.illumina.com/sequencing/sequencing_software/igenome.html
Bowtie	http://bowtie-bio.sourceforge.net/index.shtml
tophat	http://ccb.jhu.edu/software/tophat/index.shtml
Bambino	https://cgwb.nci.nih.gov/goldenPath/bamview/documentation/index.html
SAMtools	http://SAMtools.sourceforge.net/
SRA Toolkit	http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software
R	http://www.r-project.org/
python	https://www.python.org/
perl	http://www.perl.org/

366

367   # Results

368   ## Evolution of the Projects

369   *DNA-seq team*

370   Our initial goal was to build a bioinformatics pipeline to predict somatic mutations using

371   several calling algorithms and integrate the mutations with RNA-seq data to find eQTLs.

372   However, one of the main hurdles we encountered was finding an appropriate dataset that could

373   be used to test and help create our pipeline. Initially, we found a neuroblastoma dataset

374   submitted to SRA that included both the exome and RNA-seq data for matched samples.

375   However, the dataset was unusable because of corrupted files. The majority of our time at the

376   hackathon was spent on finding an appropriate dataset and debugging issues with the publicly

377   available files. This hindered our efforts to create a fully functional pipeline to achieve our goals.

378   Due to a lack of a working dataset of both DNA-seq and RNA-seq from the same individuals, we

379   were unable to write a module within our pipeline to find eQTLs. However, we were able to

380   design a pipeline that would find somatic mutations using five calling algorithms for given

381   matched samples using five different algorithms, filter and annotate mutations, and find shared

382   and unique mutations between two matched sample pairs (see Methods for more details).

383   *Epigenomics Team*

384   We initially considered different scenarios in which a lab might utilize our proposed

385   pipeline based upon their available datasets and how investigators might want to model their

386   datasets. For example, investigators might want to find a relationship between DNA methylation

387   levels and histone enrichment. Given the variation in epigenetic data available as part of publicly

388   available datasets, we recognized the need for flexibility about which data components would be

389   required to model different epigenetic relationships. Time limitations prevented us from

390   generating a workflow for every potential scenario (for example, RNA-seq and ChIP-seq, or

391   ChIP-seq and DNA methylation but no RNA-seq). Instead, we considered common questions

392   that might interest a general epigenetics laboratory. Investigators with epigenetic data often want

393   to understand how this data correlates to gene expression. Thus, we decided to focus our model

394   on elucidating relationships between a given variable collection of epigenetic data and gene

395   expression. A lab can use publicly available datasets to create a model with which to test their

396   own epigenetic data.

397   *Metagenomics Team*

398   We considered several different options for metagenomics searches before the final six

399   pipelines were settled. We spent significant time discussing the requirements for filtering or

400   other QC steps on raw reads prior to running assembly (pipelines 1 and 4) or SRA blast steps.

401   We also debated the relative merits of assembly for each task, eventually deciding to compare

402   searches with and without assembly in different pipelines, which significantly added to our

403   workload and may have contributed to the fact that only two pipelines were completed during the

404   hackathon. With regard to assembly, our group considered using the established MG-RAST

405   pipeline for a "brute force" blastn into raw reads (47), as well as using metAMOS (48), a

406   comprehensive pipeline for metagenomics analysis for comparison. Eventually we decided to

407   forego established pipelines, due to computational requirements within the timeframe of the

408   hackathon and the desire to focus on developing novel workflows of our own.

409   For pipelines 1-3 we initially planned to use human genomic information from 1000

410   Genomes (24) but found the format (base genomic sequence with list of variants) more difficult

411   to handle for our purposes than the human CEU NA12878 data that we eventually used.

412   Likewise, for the microbiome task, we initially planned to apply our pipelines to several

413   microbiome sample types, but eventually decided to focus on a skin microbiome, since skin

414   samples tend to contain abundant viruses and multiple datasets may be available due to ease of

415   sampling.

416   *RNA-seq Team*

417   Our original goals were very ambitious, especially the task of determining RNA editing

418   without DNA controls. We were interested in looking at the variants in paired cancer samples,

419   and spent a fair amount of time to find an appropriate dataset. We decided to align the samples

420   using HISAT and determine the types and counts of variants (particularly A-I transitions) that

421   may suggest RNA editing. We also wanted to determine variants in genes and then possibly

422   correlate the genes to Gene Ontology.

423   We discussed a variety of aligners, such as STAR (49) and HISAT (50), which are very

424   fast. Information about the recently released HISAT program was shared via the Google Group

425   prior to the start of the hackathon so everyone had a chance to review this program. We selected

426   HISAT because of its speed, a decision partly driven by the time constraints of the hackathon.

427   We encountered some technical difficulties processing these in HISAT, so we ran the dataset

428   with the 3 pairs of tumor/normal using tophat and bowtie so that we had some results for

429   downstream processing, while another team member continued to develop the HISAT portion of

430   the pipeline.

431   We selected bambino as our variant caller based on team members' past successes with

432   using this tool. A collection of Python and Perl scripts was written to filter out unmapped and

433   low-quality reads and, more importantly, multi-mapped reads that did not map to a unique locus

434   within the genome, since bambino would call each of these multiple alignments separately. The

435   BAMs also had to be sorted by chromosome to prepare input for bambino. Bambino was then

436   used to generate a variant call table and a Perl script was created to filter for coverage on both

437   strands and give a sparser table for downstream analysis. Another program counted the variant

438   info. We then applied the Fisher's exact test to the data using R.

439   We discussed creating a command line to get gene ontologies of the set of genes, but

440   were concerned about users keeping up-to-date versions of the GO database. Gene ontology

441   could be determined by using web sites such as PANTHER (51) or DAVID (52, 53).

442   **Technical Problems**

443   *DNA-seq Team*

444   We found the SRA website challenging to use for locating data, and the quality of

445   available datasets was inconsistent. Although many data sources validate user-submitted files, a

446   number of files that had been improperly validated and thus were not usable. Some files were

447   corrupted and could not be used, such as those in BioProject PRJNA76777 (54). Our pipeline

448   needed datasets that had a paired tumor-normal sample from the same patient. For some datasets,

449   paired samples were not available, and other datasets were marked as being paired, when in fact

450   they were not, such as BioProject PRJNA217947 (55). In addition, we observed that multiple

451   datasets in the SRA database were missing the header information required to create BAM files

452   used by downstream analysis tools, such as BioProject PRJDB1903 (56). In other cases, SRA

453   data were found to be malformed, and caused certain tools to crash. Specifically, files from

454   BioProject PRJNA268172 (27) contained reads with differing length sequence and quality scores

455   (e.g. 34 bases of sequences, 70 bases of quality information). Files with such mismatches cannot

456   be used in SAMtools to convert to BAM files, as a difference in these field lengths is

457   inconsistent with the SAM format specification (57).

458   We also encountered problems with upstream bioinformatics code quality, such as poor

459   or incorrect documentation. The tools we employed had a variety of installation methods, and

460   few were available for easy installation through a package manager. For example, core software,

461   such as R version 3, was not available as a package from the operating system vendor. Installing

462 from a third-party repository is not complex, but may be daunting to someone inexperienced in

463 systems administration.

464 *Epigenomics Team*

465   When searching for epigenetic datasets that belong to a given cell type, we found that in

466 many cases all of the necessary data were not available in one centralized location. Thus, we had

467 to search through multiple websites and databases to find enough epigenetic data for a given cell

468 type we wanted to model. In some cases, the metadata for a given file was either corrupt or

469 unavailable. In other cases, the assembly used to align reads for a given set of files was not

470 clearly indicated, so these files were discarded. When dealing with wiggle (wig) and bigWig

471 files, sometimes the format of the file was inconsistent and needed to be edited on the fly.

472 *Metagenomics Team*

473   Technical difficulties generally were resolved expediently, but still hindered timely

474 analysis within the hackathon context. For example, some Amazon EC2 nodes would suddenly

475 become completely unresponsive for unexplained reasons, requiring that we shut down and re-

476 initiate the nodes. By the end of the hackathon, results were only available from the pipelines

477 that used the SRA BLAST, in part because the SRA BLAST took about an order of magnitude

478 less computing time than the standard blast program. In both cases, many Amazon compute

479 nodes were available, but only the SRA BLAST was able to handle the large volume of human

480 genome and human microbiome read data efficiently. In contrast, a huge amount of the

481 processing power available to the standard blast program (several tens of nodes) was simply

482 wasted while the program waited for data.

483 *RNA-seq Team*

484   It is important to recognize the difficulties of variant calling, especially with RNA-seq

485   data. First, bias impacts genes expressed at lower levels. As gene expression itself varies from

486   sample to sample, depth of coverage for any particular variant may differ. For instance, a variant

487   in a sample with high gene expression would be called without difficulties, but may not be called

488   in a sample that also carries the variant but whose expression is too low to call with confidence.

489   Another source of variance lies within the heterogeneity of the tissue sample. Most tissue

490   samples harbor multiple cell types, and not all of these cells will carry a somatic mutation. This

491   problem is encountered in both DNA-seq and RNA-seq data, but results can be difficult to

492   interpret on a per-variant basis when the fluctuation in overall coverage in gene expression is

493   also considered. Thus, we decided to deal with overall global effects rather than selecting

494   particular singular changes.

495   **Project Results**

496   *DNA-seq Team*

497   The test dataset was downloaded from SRA website. The SRA Toolkit utility called

498   prefetch allows the user to download SRA data files, but we found it initially troublesome to use

499   due to configuration and storage issues; by default, prefetch stores all files in user home

500   directories, which are often limited in storage capacity. We therefore wrote a faster web-scraper

501   script to download the files from the SRA website. Given our time limitations, we had to rely on

502   the user-submitted aligned and trimmed files, but we recommend that files submitted to the SRA

503   should be validated prior to upload.

504   Our pipeline was designed to find somatic mutations using five different algorithms, filter

505   and annotate the mutations, and compare the predicted mutations between matched tumor-

506   normal samples. However, due to time constraints and initial difficulties with finding an

507   appropriate data set and software installation, we were unable to complete our analysis. A

508   diagram of the final DNA-seq Team pipeline design is presented in Fig 3.

509   **Figure 3. DNA-seq Team pipeline.**

510   *Epigenomics Team*

511   We sought to rectify previously described inconsistencies in analyses by developing a

512   more efficient, novel pipeline. Our pipeline uses RNA-seq counts, ChIP-seq peaks, and DNA

513   methylation data in order to generate a model to predict relationships between gene expression

514   and epigenetic data. These models can then be used to predict changes in gene expression with

515   respect to changes in these epigenetic signals. Publicly available datasets can be utilized to

516   generate a model, which investigators can then use to predict the state of the chromatin based on

517   their own epigenetic data. The pipeline uses a combination of Python, R, and command line-

518   based tools.

519   For each gene in a given cell type, epigenetic marks positioned locally to the gene are

520   considered, as are distal enhancer elements that may also play in a role in that gene's expression.

521   To calculate the local epigenetic effects on transcription, an arbitrary distance on the 5' and 3'

522   ends of a gene is binned into regions and the scores of epigenetic marks that reside in each of

523   these bins are collected. The distal effect of transcription on a given gene is given by peak scores

524   of enhancer elements that are at most one megabase (Mb) upstream or downstream of the gene.

525   The scores for each epigenetic mark and enhancer for a given gene are standardized and

526   stored in a data matrix, where each row corresponds to a given gene for a given sample condition

527   or cell type. Transcript gene counts generated from RNA-seq data are also stored. This pipeline

528   generates a unique model for each gene in a given cell type by considering the gene count values

529   as Y-values and each of the epigenetic scores as X-values. Corresponding coefficients are

530   calculated for each X-value. Investigators can use these coefficients to input a new set of

531   epigenetic data and receive a testable hypothesis of predicted levels of expression for each gene

532   based on the new epigenetic data. Over time, different datasets can be used to train a given

533   model to make it more reliable. A diagram of the final DNA-seq Team pipeline is presented in

534   Fig 4.

535   **Figure 4. Epigenomics Team pipeline.**

536   *Metagenomics Team*

537   Although the goals were similar across all six of our pipelines, differences in file formats

538   and analysis approaches between the pipelines required the team to split their efforts rather than

539   work together on a single pipeline. One result of this fragmentation was some lack of consistency

540   in analytical methods (for example, choice of query versus database) between the pipelines.

541   Moreover, due to time limitations of the hackathon only one assembly was completed: the

542   ABySS assembly of the NA12878 human genome. Likewise, while we completed a versatile

543   script for conversion of FASTQ files to SRA format with the latf-load tool, time allowed only for

544   its demonstration on a single human microbiome sample.

545   Initially our plan included comparison of ERV sequence abundances between NA12878

546   genomes sequenced by several different sequencing technologies. Likewise, we initially planned

547   to compare viral sequence abundances between several different microbiome sample types. Due

548   to the complexity of these tasks, we decided to demonstrate our pipelines with a single sample

549   type for each application: Illumina HiSeq 2000 reads from NA12878 and a single sample from

550   the right retroauricular crease for the HMP application. Of the six pipelines that were planned

551   and designed, we built four (pipelines 1-3 and 5).

552   We found that the most successful approach for searching a human genome for

553   endogenous retroviruses was to use reads converted to SRA format (pipeline 2) via latf-load. The

554   blastn in pipeline 2 was completed in 50-60 minutes. For pipeline 1, while an assembly of

555   NA12878 was completed using ABySS within the time constraints of the hackathon, the blastn

556   search using the assembled contigs to query the ERV database required excessive computational

557   time; after more than 4 hours using 30 cores, the search still had not finished. In contrast, the

558   blastn for pipeline 3 finished in 5 hours. Part of the increased time for the blastn search in

559   pipeline 3 may have been due to alteration of the FASTA database by merging of forward and

560   reverse paired-ends.

561   Pipeline 5 includes a set of scripts that we developed to create a versatile pipeline for

562   searching a human microbiome sample for all viruses. These scripts may be adjusted to conduct

563   BLAST searches using other types of SRA files. A shell script downloads the relevant datasets

564   for the assembled and non-assembled sequences from HMP as well as for total viral sequences

565   from RefSeq. Scripts and a wrapper, written in R, were developed to convert FASTQ data to

566   SRA format with the latf-loader tool, convert the loaded data to .kar format, run a BLAST search

567   with the blast_vdb command, and parse the data into a viral-by-sample count matrix. The

568   resulting sparse matrix may be normalized and handled in a way similar to previously published

569   methods for sparse matrices of high-throughput 16S survey data (21). Additional available code

570   executes blastn on the assembled contigs. A diagram of the final Metagenomics Team pipeline is

571   presented in Fig 5.

572   **Figure 5: Metagenomics Team pipeline.**

573   *RNA-seq Team*

574   We developed and ran a Python script that reads a user-defined manifest file to extract

575   the read sequence information from the SRA files, stores the data in FASTQ format, and

576   launches the jobs to align the sequences using HISAT. Due to technical difficulties and time

577   constraints, we decided to manually download and process a smaller set of 3 pairs of

578   tumor/normal samples, as opposed to the set of 18 pairs we had initially considered. We aligned

579   the sequences using HISAT to prepare the data for use in subsequent parts of the pipeline. The

580   aligned SAM files were filtered to remove the unmapped, low-quality or ambiguous reads, such

581   as reads that map at multiple different locations.

582   The filtered data were run through bambino to create a variant call table in which each

583   line contains a call variant at a particular location within the genome and the reference base at

584   the same location. We counted nucleotide change variants in the tumor and normal samples and

585   ran a Fisher's exact statistical test using R to identify potential RNA editing. We found no

586   significant global changes of overrepresentation, but it is important to recognize the limitations

587   of our small sample size and our focus on specific changes. RNA editing most likely only

588   comprises a small number of A to G variants, and we would not be able to identify these changes

589   by considering global total numbers as opposed to looking at each site's overall counts

590   individually. This limitation does not affect overrepresentation in a global manner, but a small

591   set of specific local changes might not be identified with this study design.

592   Before the end of the hackathon, we were able to use the initial Python script to

593   download all 36 samples and launch the alignment tool jobs, but were able to complete fewer

594   than 10 samples given the amount of time required to finish debugging. However, when

595   completed, this automation script will greatly simplify the process of accessing and launching of

596   alignment jobs for RNA-seq datasets. A diagram of the final RNA-seq Team pipeline is

597   presented in Fig 6.

598   **Figure 6: RNA-seq Team pipeline.**

Discussion and Conclusions

599

600   Feedback from hackathon participants was generally positive, and the enthusiasm that

601   participants felt was evident during the event. Participants voluntarily stayed past the planned

602   ending time each night, and many participants did not even want to take a break when lunch

603   arrived. Even more than a week after the hacakathon had ended, teams continued to

604   communicate about and work on the problems, as well as this paper.

605   Every participant in the hackathon contributed not only to the research but also to

606   drafting the paper. Each group appointed a lead writer, who worked closely with the librarian

607   editor and coordinated with the other members of their team. Because each of the team members

608   worked on different parts of the project, every individual wrote at least a portion of the sections

609   of the paper covering their work. The use of Google Docs allowed multiple authors to work on

610   the paper simultaneously and all changes to be reflected in real time. Google Docs' comment

611   functionality also facilitated communication among authors. Once the writing was considered

612   complete, the librarian editor organized and edited the draft in order to create a coherent and

613   consistent paper, then returned this final draft to all authors for their approval. Though

614   coordinating with so many authors is challenging, here we demonstrated that it is possible for a

615   large group of individuals to contribute substantively to an article.

616   Participants reported that they appreciated having structured roles within the teams. Team

617   leads were also important for the success of the team, though their presence was not necessary

618   for the entire hackathon. For example, inclement weather on the second day prevented one of the

619   team leads from attending, but the team still made progress on pipeline production. Given that

620   members of each team came from diverse backgrounds with experience working with a

621   multitude of different data types and resources, the hackathon promoted innovation through team

622   science and consensus-building. For example, it was essential that each pipeline utilize an

623   appropriate test dataset, but many teams had difficulty with data that were located across

624   multiple repositories or could not be used due to errors in metadata or formatting. Thus, teams

625   had to brainstorm other datasets to use or create new ways to process the data. Because each

626   problem encompassed technical challenges inherent in many biological fields, teams needed to

627   consolidate ideas from each member. This allowed teams to not only transcend the difficult data

628   landscape, but fostered a strong learning environment.

629   Although the ultimate goal of the hackathon was to solve biological problems,

630   participants emphasized that they appreciated this unique opportunity for career development

631   and networking. Participants with strong backgrounds in computer science effectively mentored

632   those who were less computationally savvy, and those with strong biology backgrounds were

633   able to share insights with those who lacked this expertise. Additionally, the hackathon brought

634   together individuals from different research communities who otherwise may have never met and

635   created the potential for establishing new collaborations. In particular, participants early in their

636   careers were able to meet prominent researchers in various fields and receive helpful training

637   advice from the more senior participants. We anticipate that the participants will share their

638   experiences upon returning to their respective institutes.

639   The organizers learned some valuable lessons from this event. Surprisingly, although the

640   organizers had kept the goals somewhat loosely structured, participants generally asked for more

641   structure, particularly concerning datasets. In the future, the organizers intend to prepare videos

642   for team members concerning the scientific directions of the projects prior to the event. Other

643   informational materials distributed in advance of the event could help participants learn how to

644   complete tasks that took time away from pipeline development, such as how to locate and

645   download datasets. Specific attention will be paid to using the SRA SDK to process small parts

646   of many genomes simultaneously. One team was unable to complete their pipeline, and other

647   teams were affected by time constraints, so moving some of the preparatory work of locating and

648   downloading datasets would help ensure that the teams had adequate time for more substantive

649   work on the pipelines.

650   From an institutional perspective, the hackathon was also helpful as a means to test NCBI

651   public data repositories. Over the course of this hackathon, several technical issues with respect

652   to data storage, metadata and corruptions were illuminated. These issues as well as constructive

653   feedback about how NCBI should host data were discussed directly with NCBI Director David

654   Lipman.

655   Finally, we hope that this hackathon will help to stimulate the community to continue to

656   improve these pipelines. We chose these topics and questions because they are of interest to

657   many biologists and introductory bioinformaticians. Because the data is publicly available,

658   investigators should be able to access the datasets from NCBI in order to replicate the work done

659   in creating these pipelines. We encourage members of the community to extend, expand and alter

660   these pipelines, which are licensed under a Creative Commons Attribution License (CC-BY). We

661   hope that the community will continue working with these pipelines to suit their needs and repost

662   them as they see fit.

663

664   **Acknowledgements**
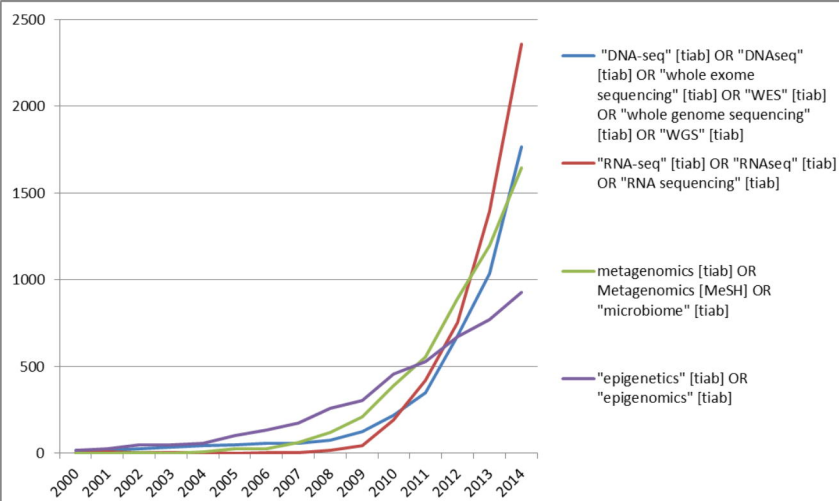
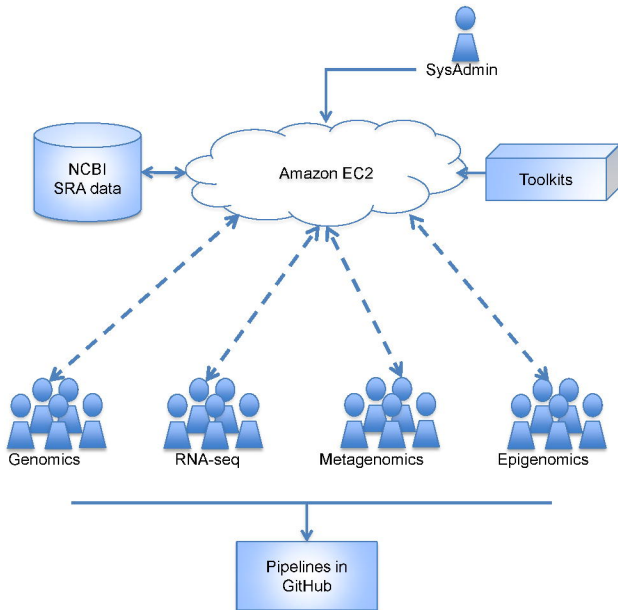669   guidance about BLAST.

670

671   **References**

672   1.   Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale
673   data management and analysis. Nature reviews Genetics. 2010;11(9):647-57. doi: 10.1038/nrg2857.
674   PubMed PMID: PMC3124937.
675   2.   Strauss KA, Puffenberger EG, Morton DH. One Community's Effort to Control Genetic Disease.
676   American Journal of Public Health. 2012;102(7):1300-6. doi: 10.2105/AJPH.2011.300569. PubMed PMID:
677   PMC3477994.
678   3.   Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, et al. Clinical interpretation
679   and implications of whole-genome sequencing. Jama. 2014;311(10):1035-45. Epub 2014/03/13. doi:
680   10.1001/jama.2014.1717. PubMed PMID: 24618965; PubMed Central PMCID: PMCPmc4119063.
681   4.   Palomaki GE, Melillo S, Neveux L, Douglas MP, Dotson WD, Janssens AC, et al. Use of genomic
682   profiling to assess risk for cardiovascular disease and identify individualized prevention strategies--a
683   targeted evidence-based review. Genetics in medicine : official journal of the American College of
684   Medical Genetics. 2010;12(12):772-84. Epub 2010/11/04. doi: 10.1097/GIM.0b013e3181f8728d.
685   PubMed PMID: 21045709.
686   5.   Celi LA, Ippolito A, Montgomery RA, Moses C, Stone DJ. Crowdsourcing Knowledge Discovery
687   and Innovations in Medicine. Journal of Medical Internet Research. 2014;16(9):e216. doi:
688   10.2196/jmir.3761. PubMed PMID: PMC4180345.
689   6.   The Health 2.0 Developer Challenge. Health 2.0⬜s Washington DC HD&IW Code-a-thon:
690   Preventing Obesity 2012 [cited 2015 January 13]. Available from:
691   http://www.health2con.com/devchallenge/washington-dcs-hdi-code-a-thon-preventing-obesity/.
692   7.   Illumina. Illumina BaseSpace Worldwide Developers Conference (WWDC) 2014 [cited 2015
693   January 13]. Available from: http://eventregistration.illumina.com/events/illumina-basespace-
694   worldwide-developers-conference-at-embl-heidelberg/event-summary-
695   3fd0d62eb6d94cc2bc017d682e9bca51.aspx.
696   8.   HelpMeViz. Hackathon: Women's Empowerment & Nutrition 2014 [cited 2015 January 13].
697   Available from: http://helpmeviz.com/2014/06/28/hackathon-womens-empowerment-nutrition/.
698   9.   DePasse JW, Carroll R, Ippolito A, Yost A, Santorino D, Chu Z, et al. Less noise, more hacking:
699   how to deploy principles from MIT's hacking medicine to accelerate health care. International journal of
700   technology assessment in health care. 2014;30(3):260-4. Epub 2014/08/07. doi:
701   10.1017/s0266462314000324. PubMed PMID: 25096225.
702   10.   Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous
703   cell lung cancers. Nature. 2012;489(7417):519-25. Epub 2012/09/11. doi: 10.1038/nature11404.
704   PubMed PMID: 22960745; PubMed Central PMCID: PMCPmc3466113.
705   11.   Edmonson MN, Zhang J, Yan C, Finney RP, Meerzaman DM, Buetow KH. Bambino: a variant
706   detector and alignment viewer for next-generation sequencing data in the SAM/BAM format.
707   Bioinformatics (Oxford, England). 2011;27(6):865-6. Epub 2011/02/01. doi:
708   10.1093/bioinformatics/btr032. PubMed PMID: 21278191; PubMed Central PMCID: PMCPmc3051333.
709   12.   Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation
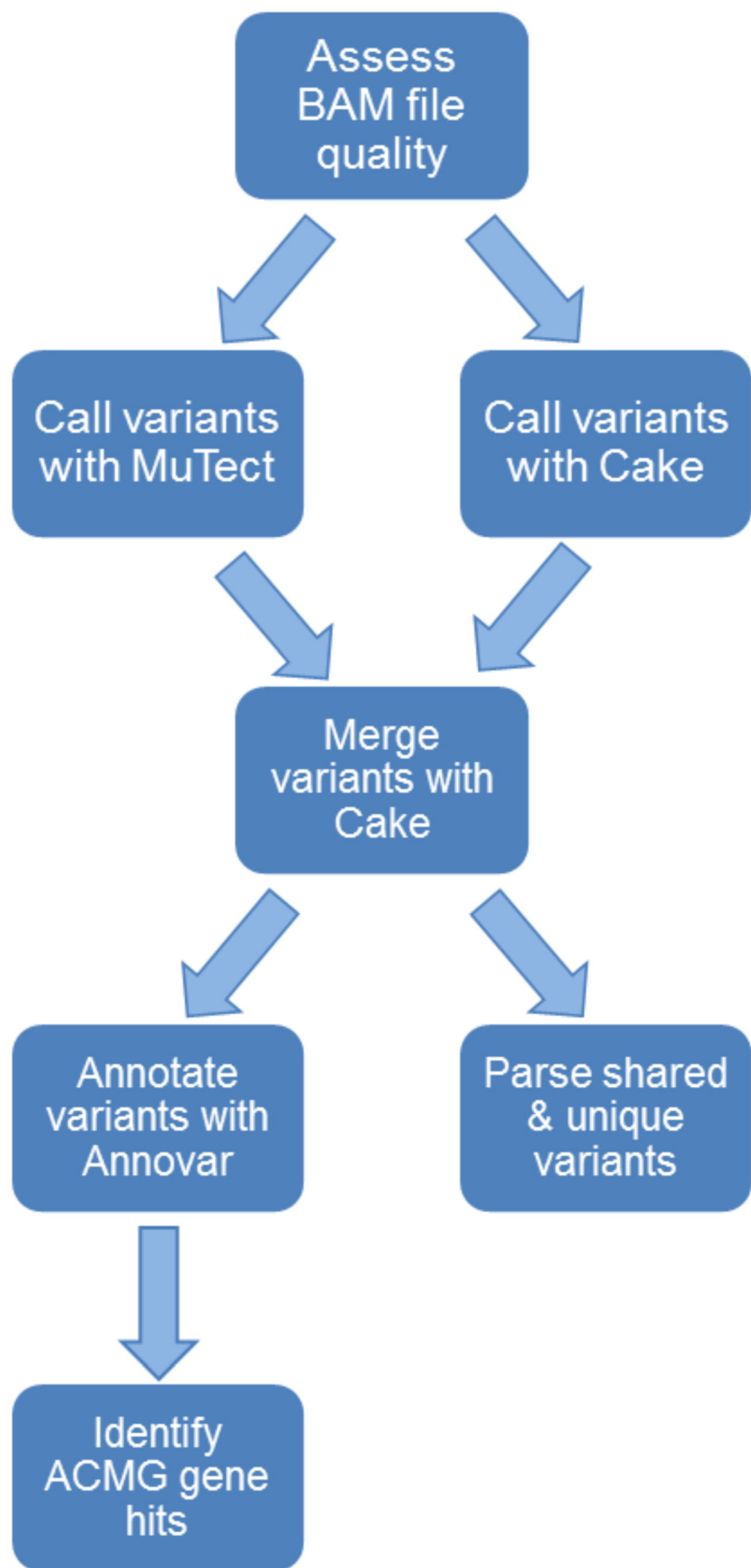710   and copy number alteration discovery in cancer by exome sequencing. Genome research.

711   2012;22(3):568-76. Epub 2012/02/04. doi: 10.1101/gr.129684.111. PubMed PMID: 22300766; PubMed
712   Central PMCID: PMCPmc3290792.
713   13.   Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection
714   of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology.
715   2013;31(3):213-9. Epub 2013/02/12. doi: 10.1038/nbt.2514. PubMed PMID: 23396013; PubMed Central
716   PMCID: PMCPmc3833702.
717   14.   Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in
718   amplicon and whole exome sequence data. BMC genomics. 2014;15:244. Epub 2014/04/01. doi:
719   10.1186/1471-2164-15-244. PubMed PMID: 24678773; PubMed Central PMCID: PMCPmc3986649.
720   15.   Pallen MJ. Diagnostic metagenomics: potential applications to bacterial, viral and parasitic
721   infections. Parasitology. 2014;141(14):1856-62. Epub 2014/03/01. doi: 10.1017/s0031182014000134.
722   PubMed PMID: 24576467; PubMed Central PMCID: PMCPmc4255322.
723   16.   Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd KK, Lenz J. Many human endogenous
724   retrovirus K (HERV-K) proviruses are unique to humans. Current biology : CB. 1999;9(16):861-8. Epub
725   1999/09/02. PubMed PMID: 10469592.
726   17.   Wang-Johanning F, Frost AR, Jian B, Epp L, Lu DW, Johanning GL. Quantitation of HERV-K env
727   gene expression and splicing in human breast cancer. Oncogene. 2003;22(10):1528-35. Epub
728   2003/03/12. doi: 10.1038/sj.onc.1206241. PubMed PMID: 12629516.
729   18.   Pop M, Walker AW, Paulson J, Lindsay B, Antonio M, Hossain MA, et al. Diarrhea in young
730   children from low-income countries leads to large-scale alterations in intestinal microbiota composition.
731   Genome biology. 2014;15(6):R76. Epub 2014/07/06. doi: 10.1186/gb-2014-15-6-r76. PubMed PMID:
732   24995464; PubMed Central PMCID: PMCPmc4072981.
733   19.   Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, et al. Viruses in the faecal
734   microbiota of monozygotic twins and their mothers. Nature. 2010;466(7304):334-8. Epub 2010/07/16.
735   doi: 10.1038/nature09199. PubMed PMID: 20631792; PubMed Central PMCID: PMCPmc2919852.
736   20.   Waller AS, Yamada T, Kristensen DM, Kultima JR, Sunagawa S, Koonin EV, et al. Classification and
737   quantification of bacteriophage taxa in human gut metagenomes. The ISME journal. 2014;8(7):1391-
738   402. Epub 2014/03/14. doi: 10.1038/ismej.2014.30. PubMed PMID: 24621522; PubMed Central PMCID:
739   PMCPmc4069399.
740   21.   Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-
741   gene surveys. Nature methods. 2013;10(12):1200-2. Epub 2013/10/01. doi: 10.1038/nmeth.2658.
742   PubMed PMID: 24076764; PubMed Central PMCID: PMCPmc4010126.
743   22.   Han L, Vickers KC, Samuels DC, Guo Y. Alternative applications for distinct RNA sequencing
744   strategies. Briefings in bioinformatics. 2014. Epub 2014/09/24. doi: 10.1093/bib/bbu032. PubMed PMID:
745   25246237.
746   23.   Antonarakis ES, Lu C, Wang H, Luber B, Nakazawa M, Roeser JC, et al. AR-V7 and resistance to
747   enzalutamide and abiraterone in prostate cancer. The New England journal of medicine.
748   2014;371(11):1028-38. Epub 2014/09/04. doi: 10.1056/NEJMoa1315815. PubMed PMID: 25184630;
749   PubMed Central PMCID: PMCPmc4201502.
750   24.   Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated
751   map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65. Epub 2012/11/07.
752   doi: 10.1038/nature11632. PubMed PMID: 23128226; PubMed Central PMCID: PMCPmc3498066.
753   25.   National Center for Biotechnology Information. ClinVar 2015 [cited 2015 January 13]. Available
754   from: http://www.ncbi.nlm.nih.gov/clinvar/.
755   26.   National Center for Biotechnology Information. SRA Toolkit 2015 [cited 2015 January 13].
756   Available from:
757   http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software.

758   27.   National Center for Biotechnology Information. Whole exome sequencing in a case of sporadic
759   multiple meningiomas 2014 [cited 2015 January 13]. Available from:
760   http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA268172.
761   28.   Perry A, Scheithauer BW, Stafford SL, Lohse CM, Wollan PC. "Malignancy" in meningiomas: a
762   clinicopathologic study of 116 patients, with grading implications. Cancer. 1999;85(9):2046-56. Epub
763   1999/05/01. PubMed PMID: 10223247.
764   29.   Rashid M, Robles-Espinoza CD, Rust AG, Adams DJ. Cake: a bioinformatics pipeline for the
765   integrated analysis of somatic variants in cancer genomes. Bioinformatics (Oxford, England).
766   2013;29(17):2208-10. doi: 10.1093/bioinformatics/btt371. PubMed PMID: PMC3740632.
767   30.   Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of
768   cancer genes and mutational processes in breast cancer. Nature. 2012;486(7403):400-4. Epub
769   2012/06/23. doi: 10.1038/nature11017. PubMed PMID: 22722201; PubMed Central PMCID:
770   PMCPmc3428862.
771   31.   Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
772   format and SAMtools. Bioinformatics (Oxford, England). 2009;25(16):2078-9. Epub 2009/06/10. doi:
773   10.1093/bioinformatics/btp352. PubMed PMID: 19505943; PubMed Central PMCID: PMCPmc2723002.
774   32.   Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-
775   throughput sequencing data. Nucleic acids research. 2010;38(16):e164. Epub 2010/07/06. doi:
776   10.1093/nar/gkq603. PubMed PMID: 20601685; PubMed Central PMCID: PMCPmc2938201.
777   33.   Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format
778   and VCFtools. Bioinformatics (Oxford, England). 2011;27(15):2156-8. Epub 2011/06/10. doi:
779   10.1093/bioinformatics/btr330. PubMed PMID: 21653522; PubMed Central PMCID: PMCPmc3137218.
780   34.   Sadedin SP, Pope B, Oshlack A. Bpipe: a tool for running and managing bioinformatics pipelines.
781   Bioinformatics (Oxford, England). 2012;28(11):1525-6. Epub 2012/04/14. doi:
782   10.1093/bioinformatics/bts167. PubMed PMID: 22500002.
783   35.   Tange O. GNU Parallel - The Command-Line Power Tool. ;login: The USENIX Magazine.
784   2011;36(1):42-7.
785   36.   National Center for Biotechnology Information. NIH Epigenomic Roadmap 2015 [cited 2015
786   January 5]. Available from: ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/roadmapepigenomics/.
787   37.   International Human Epigenome Consortium. IHEC Data Portal 2015 [cited 2015 January 13].
788   Available from: http://epigenomesportal.ca/ihec/grid.html.
789   38.   ENCODE Consortium. ENCODE: Encyclopedia of DNA Elements 2015 [cited 2015 January 13].
790   Available from: https://www.encodeproject.org/.
791   39.   National Center for Biotechnology Information. ERX069505: Whole Genome Sequencing of
792   human CEU NA12878 2015 [cited 2015 January 13]. Available from:
793   http://www.ncbi.nlm.nih.gov/sra/ERX069505%5Baccn%5D.
794   40.   BC Cancer Agency. ABySS: Assembly By Short Sequences - a de novo, parallel, paired-end
795   sequence assembler 2014 [cited 2015 January 13]. Available from:
796   http://www.bcgsc.ca/platform/bioinfo/software/abyss.
797   41.   Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for
798   short read sequence data. Genome research. 2009;19(6):1117-23. Epub 2009/03/03. doi:
799   10.1101/gr.089532.108. PubMed PMID: 19251739; PubMed Central PMCID: PMCPmc2694472.
800   42.   National Institutes of Health. NIH Human Microbiome Project 2015 [cited 2015 January 13].
801   Available from: http://hmpdacc.org/.
802   43.   Chan TH, Lin CH, Qi L, Fei J, Li Y, Yong KJ, et al. A disrupted RNA editing balance mediated by
803   ADARs (Adenosine DeAminases that act on RNA) in human hepatocellular carcinoma. Gut.
804   2014;63(5):832-43. Epub 2013/06/15. doi: 10.1136/gutjnl-2012-304037. PubMed PMID: 23766440;
805   PubMed Central PMCID: PMCPmc3995272.

806   44.   National Center for Biotechnology Information. Transcriptome sequencing of human
807   hepatocellular carcinoma (human) 2011 [cited 2015 January 13]. Available from:
808   http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA149267.
809   45.   Kim SK, Kim SY, Kim JH, Roh SA, Cho DH, Kim YS, et al. A nineteen gene-based risk score classifier
810   predicts prognosis of colorectal cancer patients. Molecular oncology. 2014;8(8):1653-66. Epub
811   2014/07/23. doi: 10.1016/j.molonc.2014.06.016. PubMed PMID: 25049118.
812   46.   National Center for Biotechnology Information. Gene expression profiling study by RNA-seq in
813   colorectal cancer (human) 2013 [cited 2015 January 13]. Available from:
814   http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA218851.
815   47.   Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST
816   server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC
817   bioinformatics. 2008;9:386. Epub 2008/09/23. doi: 10.1186/1471-2105-9-386. PubMed PMID:
818   18803844; PubMed Central PMCID: PMCPmc2563014.
819   48.   Center for Bioinformatics and Computational Biology. metAMOS: assembly and analysis toolkit
820   for metagenomics 2015 [cited 2015 January 13]. Available from:
821   http://cbcb.umd.edu/software/metAMOS.
822   49.   Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-
823   seq aligner. Bioinformatics (Oxford, England). 2013;29(1):15-21. Epub 2012/10/30. doi:
824   10.1093/bioinformatics/bts635. PubMed PMID: 23104886; PubMed Central PMCID: PMCPmc3530905.
825   50.   Kim D, Langmead B, Salzberg S. HISAT: Hierarchical Indexing for Spliced Alignment of
826   Transcripts2014 2014-01-01 00:00:00.
827   51.   Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function,
828   and other gene attributes, in the context of phylogenetic trees. Nucleic acids research.
829   2013;41(Database issue):D377-86. Epub 2012/11/30. doi: 10.1093/nar/gks1118. PubMed PMID:
830   23193289; PubMed Central PMCID: PMCPmc3531194.
831   52.   Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists
832   using DAVID bioinformatics resources. Nature protocols. 2009;4(1):44-57. Epub 2009/01/10. doi:
833   10.1038/nprot.2008.211. PubMed PMID: 19131956.
834   53.   Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the
835   comprehensive functional analysis of large gene lists. Nucleic acids research. 2009;37(1):1-13. Epub
836   2008/11/27. doi: 10.1093/nar/gkn923. PubMed PMID: 19033363; PubMed Central PMCID:
837   PMCPmc2615629.
838   54.   National Center for Biotechnology Information. Sequencing of multiple tumors from a
839   neuroblastoma patient 2011 [cited 2015 January 13]. Available from:
840   http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA76777.
841   55.   National Center for Biotechnology Information. Homo sapiens exome and transcriptome 2013
842   [cited 2015 January 13]. Available from: http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA217947.
843   56.   National Center for Biotechnology Information. Whole exome analysis of myelodysplastic
844   syndromes 2014 [cited 2015 January 13]. Available from:
845   http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJDB1903.
846   57.   The SAM/BAM Format Specication Working Group. Sequence Alignment/Map Format
847   Specication 2014 [cited 2015 January 13]. Available from: http://samtools.github.io/hts-
848   specs/SAMv1.pdf.
849

Legend:

- "DNA-seq" [tiab] OR "DNAseq" [tiab] OR "whole exome sequencing" [tiab] OR "WES" [tiab] OR "whole genome sequencing" [tiab] OR "WGS" [tiab]
- "RNA-seq" [tiab] OR "RNAseq" [tiab] OR "RNA sequencing" [tiab]
- metagenomics [tiab] OR Metagenomics [MeSH] OR "microbiome" [tiab]
- "epigenetics" [tiab] OR "epigenomics" [tiab]

SysAdmin

NCBI
SRA data

Amazon EC2

Toolkits

Genomics

RNA-seq

Metagenomics

Epigenomics

Pipelines in
GitHub

1. Input Data — GEO

	RNA-Seq	DNA methylation	ChIP-Seq	Enhancer
Cell-Type A	Counts	Score	Peak Area	Peak Area
Cell-Type B	Counts	Score	Peak Area	Peak Area
Cell-Type C	Counts	Score	Peak Area	Peak Area

2. Solve $Y = A X_1 + B X_2 + C X_3$
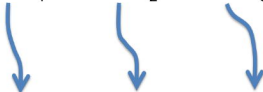
3. Model $Y = A(data) + B(data) + C(data)$
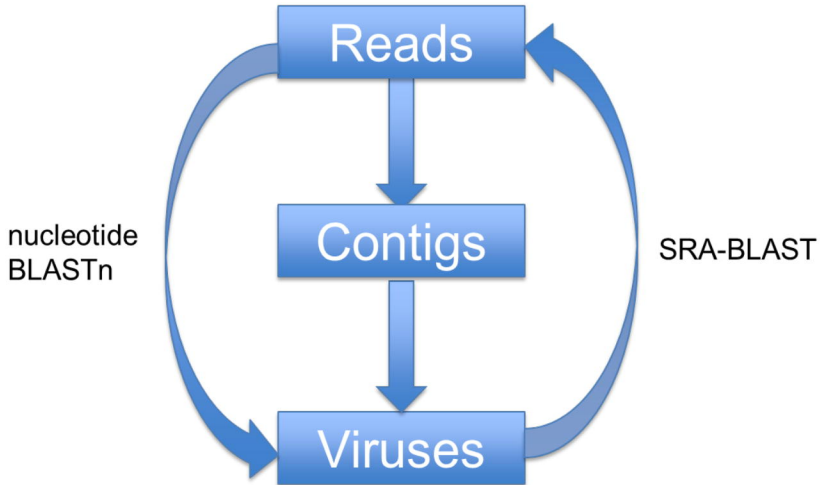
Reads

Contigs

Viruses

nucleotide
BLASTn

SRA-BLAST

SRA toolkit
(...... FASTQ)

HISAT
(Alignment)

Samtools
(filtering, sorting,
indexing)

Bambino
(variant calling)

Tabulating
counts for each
variant type

Mapping
locations to
genes

Gene ontology
analysis