BIO392
Bioinformatics of Genome Variations

Genome Variation | Function | Data Formats | Resources | Privacy

Michael Baudis **UZH SIB**
Computational Oncogenomics

University of Zurich[UZH]

BIO392: Course Schedule

- Introduction, File Formats & Genome Browsers (Michael Baudis)
  - 2018-09-18 (Tue), 13-17
  - 2018-09-19 (Wed), 09-17
  - 2018-09-20 (Thu), 09-17
  - 2018-09-21 (Fri), 09-17
- Tools & Programmatic Solutions (Izaskun Mallona)
  - 2018-09-25 (Tue), 13-17
  - 2018-09-26 (Wed), 09-17
  - 2018-09-27 (Thu), 09-17
  - 2018-09-28 (Fri), 09-17
- Genome Variants to Modified Proteins (Elif Özkirimli Olmez)
  - 2018-10-02 (Tue), 13-17
  - 2018-10-03 (Wed), 09-17
  - 2018-10-04 (Thu), 09-17
  - 2018-10-05 (Fri), 09-17
- Review & Write-Up (Michael Baudis)
  - 2018-10-09 (Tue), 13-17
  - 2018-10-10 (Wed), 09-17
    - Written exam

University of Zurich[UZH]

BIO320: Course Resources

University of Zurich[UZH]

- Course repository & website on Github
  - links to articles and information resources
  - downloads

https://baudisgroup.github.io/BIO392-Github/
https://github.com/baudisgroup/BIO392-Github

BIO392-Github

View the Project on GitHub
baudisgroup/BIO392-Github

This project is maintained by
baudisgroup

Hosted on GitHub Pages — Theme by orderedlist

BIO392

This is a repository for materials related to the BIO392 *Bioinformatics of Genome Variation* introductory course at the University of Zürich, Autumn 2018.

The home page of this repository can be found at https://baudisgroup.github.io/BIO392-Github/.

Abstract

One of the fastest growing areas of bioinformatics is in the analysis, warehousing and representation of genomic and protein sequence variants, particularly with view on the use of molecular data in personalised health and biomedical applications in general. This course will engage participants to explore common data formats, online resources and analysis techniques, with a focus on human genome variation data.

Links

- BIO392 in the UZH OLAT system
- BIO392 in the UZH directory

Literature and Resources

- articles and other reading (files)
- article links
- resource links (browsers and online repositories)

Schedule

The course schedule consists of 1 afternoon (Tuesday) and 3 “full” days (Wednesday - Friday) in the first two weeks, with 1.5 days in the last week.

The “full days” run from 09:00 - 17:00, with a break from 12:00 - 13:30.

Conceptually, mornings are predominantly related to introductions, presentations and discussion of the previous days, and are afternoons reserved for independent work on the examples and tasks.

Skill assessment

Github? R? Python? Perl? YAML?

BIO392
Bioinformatics of Genome Variations

Genomes: Core of "Personalized Health" & "Precision Medicine"

Michael Baudis **UZH SIB**
Computational Oncogenomics

University of Zurich[UZH]

Many names for one concept or many concepts in one name?

Stratified, personalised, precision, individualised, P4 medicine or personalised healthcare – all are terms in use to describe notions often referred to as the future of medicine and healthcare. But what exactly is it all about, and are we all talking about the same thing?

Source: PHG Foundation

While medicine has always been "personal" and "precise" in the given context of available knowledge and technologies, the concept of "**Personalised Medicine**" describes the use of individual genome information, concept based metadata and individually targeted therapies.

Slides freely downloadable from lectures.GersteinLab.org & “tweetable” (via @markgerstein)

Personal Genomics as a Gateway into Biology

Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.

Genome analyses at the core of Personalized Health™

- Genome analyses (including transcriptome, metagenomics) are the **core technologies** for Personalized Health™ applications
- In the context of **academic medicine**, this requires
  - standard sample acquisition procedures & central **biobanking**
  - **core sequencing facility** (large throughput, cost efficiency, uniform sample and data handling procedures)
- secure **computing/analysis** platform
- Standardized **data formats** and **sample identification** procedures
- Metadata rich, reference **variant resource(s)** & expertise
- participation in reciprocal, international **data sharing** and **biocuration** efforts

Genome analyses at the core of Personalized Health™

Susceptibility, Pharmacogenomics, Classification, Infectious Diseases, Outcome Prediction, Lifestyle ...

doi:10.1038/nature19057

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2*], Eric V. Minikel[1,2,5*], Kaitlin E. Samocha[1,2,5,6*], Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2], Daniel P. Birnbaum[2], Jack A. Kosmicki[1,2,6,13], Laramie E. Duncan[1,2,6], Karol Estrada[1,2], Fengmei Zhao[1,2], James Zou[2],

Barkur S. Shastry

SNP alleles in human disease and evolution

DISEASE MECHANISMS

Mechanisms underlying structural variant formation in genomic disorders

Claudia M. B. Carvalho[1,2] and James R. Lupski[1,3,4,5]

Abstract | With the recent burst of technological developments in genomics, and the clinical implementation of genome-wide assays, our understanding of the molecular basis of genomic disorders, specifically the contribution of structural variation to disease burden, is evolving

Rapid whole genome sequencing and precision neonatology

Joshua E. Petrikin, MD[a,], Laurel K. Willig, MD, FAAP[b], Laurie D. Smith, MD, PhD[c], and Stephen F. Kingsmore, MB, BAO, ChB, Dsc, FRCPath[d,e]*

insight progress

Cancer genetics

Bruce A. J. Ponder

Genomic Classification of Cutaneous Melanoma

The Cancer Genome Atlas Network[1,*,*,*]
[1]Cancer Genome Atlas Program Office, National Cancer Institute at NIH, 31 Center Drive, Bldg. 31, Suite 3A20, Bethesda, MD 20892, USA
*Correspondence: irwatson@mdanderson.org (I.R.W.), jgershen@mdanderson.org (J.E.G.), lchin@mdanderson.org (L.C.)
http://dx.doi.org/10.1016/j.cell.2015.05.044

Consequences of genomic diversity in *Mycobacterium tuberculosis*

Mireia Coscolla[a,b], Sebastien Gagneux[a,b,*]
[a] *Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Socinstrasse 57, 4002 Basel, Switzerland*
[b] *University of Basel, Petersplatz 1, Basel 4003, Switzerland*

doi:10.1111/pcn.12128

Psychiatry and Clinical Neurosciences

PCN Frontier Review

Copy-number variation in the pathogenesis of autism spectrum disorder

Emiko Shishido, PhD[1,2,3], Branko Aleksic, MD, PhD[3] and Norio Ozaki, MD, PhD[3]*
...he Promotion of Science, ...pan

Common gene variants, mortality and extreme longevity in humans

B.T. Heijmans[a,b], R.G.J. Westendorp[b], P.E. Slagboom[a,*]

RESEARCH ARTICLE Open Access

Integrative genome-wide expression profiling identifies three distinct molecular subgroups of renal cell carcinoma with different patient outcome

...nfred Beleut[1,5*], Philip Zimmermann[2], Michael Baudis[3], Nicole Bruni[4], Peter Bühlmann[4], Oliver Laule[2], ...-Duc Luu[1], Wilhelm Gruissem[2], Peter Schraml[1*] and Holger Moch[1]

NEURODEVELOPMENT

Genes, circuits, and precision therapies for autism and related neurodevelopmental disorders

Mustafa Sahin* and Mriganka Sur*

RESEARCH ARTICLE Open Access

Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens

Haoyang Cai[1,2], Nitin Kumar[1,2], Homayoun C Bagheri[3], Christian von Mering[1,2], Mark D Robinson[1,2*] and Michael Baudis[1,2*]

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.
N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

The landscape of somatic copy-number alteration across human cancers

Rameen Beroukhim[1,3,4,5*], Craig H. Mermel[1,3*], Dale Porter[8], Guo Wei[1], Soumya Raychaudhuri[1,4], Jerry Donovan[8], Jordi Barretina[1,3], Jesse S. Boehm[1], Jennifer Dobson[1,3], Mitsuyoshi Urashima[9], Kevin T. Mc Henry[8], Reid M. Pinchback[1], Azra H. Ligon[4], Yoon-Jae Cho[6], Leila Haery[1,3], Heidi Greulich[1,3,4,5], Michael Reich[1], Wendy Winckler[1], Michael S. Lawrence[1], Barbara A. Weir[1,3], Kumiko E. Tanaka[1,3], Derek Y. Chiang[1,3,13], Adam J. Bass[1,3,4], Alice Loo[8], Carter Hoffman[1,3], John Prensner[1,3], Ted Liefeld[1], Qing Gao[1], Derek Yecies[3], Sabina Signoretti[3,4], Elizabeth Maher[10], Frederic J. Kaye[11], Hidefumi Sasaki[12], Joel E. Tepper[13], Jonathan A. Fletcher[4], Josep Tabernero[14], José Baselga[14], Ming-Sound Tsao[15], Francesca Demichelis[16], Mark A. Rubin[16], Pasi A. Janne[3,4], Mark J. Daly[1,17], Carmelo Nucera[7], Ross L. Levine[18], Benjamin L. Ebert[1,4,5], Stacey Gabriel[1], Anil K. Rustgi[19], Cristina R. Antonescu[18], Marc Ladanyi[18], Anthony Letai[3], Levi A. Garraway[1,3], Massimo Loda[3,4], David G. Beer[20], Lawrence D. True[21], Aikou Okamoto[22], Scott L. Pomeroy[6], Samuel Singer[18], Todd R. Golub[1,3,23], Eric S. Lander[1,2,5], Gad Getz[1], William R. Sellers[8] & Matthew Meyerson[1,3,5]

The trouble with human genome variation

Conclusions from the analysis of variation in the human genome

- 1. Humans are all very similar to each other
  - Two humans will show about 99.9% sequence identity with each other. In other words, only about 1 in 1’000 bp is different between two individuals.
  - Humans show about 98% sequence identity to chimps. So two humans are still much more similar to each other than either is to the monkey.

- 2. Humans are very different from each other
  - Two typical humans will likely have over 1’000’000 independent sequence differences in their genomes.

Slide from Michael Hengartner’s Human Genome Lecture

Finding Somatic Mutations In Cancer:
Many Needles in a Large Haystack

- a typical human genome (~3 billion base pairs) has ~5 million variants
- most of them are "**rare**"; i.e. can only be identified as recurring when sequencing thousands of people
- cancer cells accumulate additional variants, only **few** of which ("**drivers**") are relevant for the disease

Graphic adapted from Mark Gerstein (GersteinLab.org; @markgerstein)

Mutations & genomic rearrangements in cancer

Lengauer et al. Genetic instabilities in human cancers. Nature (1998) vol. 396 (6712) pp. 643-9

a. small mutation (di-pyrimidine exchange at p53 in Xeroderma pigmentosum patient)

b. two-base deletion in *TGFB* in a colorectal cancer patient with mismatch repair deficiency

c. chromosomal losses (FISH; red=3, yellow=12) in CRC

d. t(1;17) in neuroblastoma, whole-chromosomal painting

e. *MYCN* gene amplification (multiple copies inserted into chromosome 1 derived marker)

Generation of copy number imbalances in cancer through imbalanced cytogenetic rearrangements - partial deletion of 11q, gain of 11pterq21 and 2 addl. copies of 17q

RL Stallings: Are chromosomal imbalances important in cancer? Volume 23, Issue 6, p278–283, 2007

Quantifying Somatic Mutations In Cancer

GENOMIC COPY NUMBER IMBALANCES PROVIDE WIDESPREAD SOMATIC VARIANTS IN CANCER

On average ~15% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on >30'000 cancer genomes from arraymap.org

CANCERS SHOW THOUSANDS OF SINGLE NUCLEOTIDE VARIANTS PER SAMPLE, MOSTLY IN NON-CODING REGIONS

Pan-Cancer Analysis of Whole Genomes (PCAWG) data show widespread mutations in non-coding regions of cancer genomes (Khurana et al., Nat. Rev. Genet. (2016)

Nobody is perfect (?)

A 10.7 Mb Interstitial Deletion of 13q21 Without Phenotypic Effect Defines a Further Non-Pathogenic Euchromatic Variant
Andreas Roos, Miriam Elbracht, Michael Baudis, Jan Senderek, Nadine Schönherr, Thomas Eggermann, and Herdit M. Schüler
American Journal of Medical Genetics Part A 146A:2417 – 2420 (2008)

Genomes Everywhere

Large Genome Data Generation, Analysis & Sharing Initiatives

Organization / Initiative: Name	Organization / Initiative: Category	Cohort
100K Wellness Project	Research Project	107 unaffected individuals (scaling up to 100,000)
23andMe	Organization	>1 million customers (>80% consented to research)
Actionable Cancer Genome Initiative (ACGI)	Data-Sharing Project	Goal: 100,000 individuals
Ancestry.com	Organization	1.4 million customer DNA samples (what % consented to research?)
BioBank Japan	Repository	Specimens from >200,000 patients and unaffected controls
Cancer Moonshot2020	Consortium	Phase 1: 20,000 cancer patients
Children's Hospital of Philadelphia Biorepository	Repository	Capacity for 8.6 million samples
China Kadoorie Biobank	Repository	>512,000 participants (general population, China). Genotyping data available for ~100,000.
CIMBA	Consortium	>15,000 BRCA1 carriers, >8,000 BRCA2 carriers
Clinical Sequencing Exploratory Research (CSER)	Consortium	~4,000 patients and healthy controls
DECIPHER	Repository	19,014 patients (international)
deCode Genetics	Organization	500,000 participants (international)
East London Genes & Health	Research Project	100,000 unaffected individuals (East London, Pakistani or Bangladeshi heritage)
Electronic Medical Records and Genomics (eMERGE) Network	Repository, Consortium, Research Project	55,028 patients
European Network for Genetic and Genomic Epidemiology (ENGAGE)	Research Project	80,000 GWAS scans, and DNA and serum/plasma from >600,000 individuals
Exome Aggregation Consortium (ExAC)	Consortium	60,706 individuals
GENIE/AACR	Data-Sharing Project	>17,000 cancer patients (international)
Genome Asia 100K	Consortium	Goal: 100,000 individuals (Asia)
Genomics England	Organization	Goal: 100,000 genomes from 70,000 individuals (rare disease & cancer patients, and their relatives)
GoT2D	Consortium, Data-Sharing Project	Multiple case-control cohorts
International Cancer Genome Consortium (ICGC)	Consortium	currently data from >16'000 samples
International Genomics of Alzheimer's Project (IGAP)	Consortium	40,000 patients with Alzheimer's disease
International Multiple Sclerosis Genetics (IMSG) Consortium	Consortium	Goal: >50,000 patients with MS
Kaiser Permanente: Genes, Environment, and Health (RPGEH)	Repository, Research Project	200,000 DNA samples (scaling up to 500,000)
Leiden Open Variation Database (LOVD)	Repository	>170,000 individuals
Million Veteran Program	Research Project	Goal: 1 million individuals; first 200,000 is complete.
MyCode® Community Health Initiative	Repository, Research Project	Goal: >250,000 patients
Precision Medicine Initiative	Research Project	Goal: >1 million participants, starting in 2016 (US)
Psychiatric Genomics Consortium (PGC)	Consortium	>170,000 subjects
Resilience Project	Research Project	589,306 individuals
Saudi Human Genome Program	Research Project	Goal: ~100,000 patients and controls (Saudi Arabia)
Scottish Genomes Partnership (SGP)	Research Project	>3,000 individuals (Scotland)
T2D-GENES	Consortium, Data-Sharing Project	10,000 patients and controls (five ethnicities); 600 individuals (Mexican American)
TBResist	Consortium	>2,600 samples
UK Biobank	Repository, Consortium, Research Project	500,000 individuals (age 40-69 years; UK)
UK10K	Research Project	10,000 participants (6,000 patients and 4,000 controls)
Vanderbilt's BioVU	Repository	>215,000 samples

Global Alliance for Genomics & Health

Genome Sequencing

whole genome sequencing (WGS)

100%

all DNA
(3.1 billion base pairs)

exome sequencing

~1%

protein-coding DNA only
(~31 million base pairs)

What does it cost to sequence a genome?

	Human Genome Project (HGP): 1991-2003	today: 2017	today: 2017
cost:	$2.7 billion	~$1,500	~$530
time:	12+ years	< 2 days	~3 days

Ryan L. Collins & Brad Wierbowski (2017)

Fig 1. Growth of DNA sequencing. The plot shows the growth of DNA sequencing both in the total number of human genomes sequenced (left axis) as well as the worldwide annual sequencing capacity (right axis: Tera-basepairs (Tbp), Peta-basepairs (Pbp), Exa-basepairs (Ebp), Zetta-basepairs (Zbps)). The values through 2015 are based on the historical publication record, with selected milestones in sequencing (first Sanger through first PacBio human genome published) as well as three exemplar projects using large-scale sequencing: the 1000 Genomes Project, aggregating hundreds of human genomes by 2012 [3]; The Cancer Genome Atlas (TCGA), aggregating over several thousand tumor/normal genome pairs [4]; and the Exome Aggregation Consortium (ExAC), aggregating over 60,000 human exomes [5]. Many of the genomes sequenced to date have been whole exome rather than whole genome, but we expect the ratio to be increasingly favored towards whole genome in the future. The values beyond 2015 represent our projection under three possible growth curves as described in the main text.

Citation: Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195

DNA Sequencing Costs: Data

Data from the NHGRI Genome Sequencing Program (GSP)

https://www.genome.gov/sequencingcostsdata/

Date	Cost per Mb	Cost per Genome
Sep-01	JS$ 5'292.39	US$ 95'263'072
Mar-02	JS$ 3'898.64	US$ 70'175'437
Sep-02	JS$ 3'413.80	US$ 61'448'422
Mar-03	JS$ 2'986.20	US$ 53'751'684
Oct-03	JS$ 2'230.98	US$ 40'157'554
Jan-04	JS$ 1'598.91	US$ 28'780'376
Apr-04	JS$ 1'135.70	US$ 20'442'576
Jul-04	JS$ 1'107.46	US$ 19'934'346
Oct-04	JS$ 1'028.85	US$ 18'519'312
Jan-05	US$ 974.16	US$ 17'534'970
Apr-05	US$ 897.76	US$ 16'159'699
Jul-05	US$ 898.90	US$ 16'180'224
Oct-05	US$ 766.73	US$ 13'801'124
Jan-06	US$ 699.20	US$ 12'585'659
Apr-06	US$ 651.81	US$ 11'732'535
Jul-06	US$ 636.41	US$ 11'455'315
Oct-06	US$ 581.92	US$ 10'474'556
Jan-07	US$ 522.71	US$ 9'408'739
Apr-07	US$ 502.61	US$ 9'047'003
Jul-07	US$ 495.96	US$ 8'927'342
Oct-07	US$ 397.09	US$ 7'147'571
Jan-08	US$ 102.13	US$ 3'063'820
Apr-08	US$ 15.03	US$ 1'352'982
Jul-08	US$ 8.36	US$ 752'080
Oct-08	US$ 3.81	US$ 342'502
Jan-09	US$ 2.59	US$ 232'735
Apr-09	US$ 1.72	US$ 154'714
Jul-09	US$ 1.20	US$ 108'065
Oct-09	US$ 0.78	US$ 70'333
Jan-10	US$ 0.52	US$ 46'774
Apr-10	US$ 0.35	US$ 31'512
Jul-10	US$ 0.35	US$ 31'125
Oct-10	US$ 0.32	US$ 29'092
Jan-11	US$ 0.23	US$ 20'963
Apr-11	US$ 0.19	US$ 16'712
Jul-11	US$ 0.12	US$ 10'497
Oct-11	US$ 0.09	US$ 7'743
Jan-12	US$ 0.09	US$ 7'666
Apr-12	US$ 0.07	US$ 5'901
Jul-12	US$ 0.07	US$ 5'985
Oct-12	US$ 0.07	US$ 6'618
Jan-13	US$ 0.06	US$ 5'671
Apr-13	US$ 0.06	US$ 5'826
Jul-13	US$ 0.06	US$ 5'550
Oct-13	US$ 0.06	US$ 5'096
Jan-14	US$ 0.04	US$ 4'008
Apr-14	US$ 0.05	US$ 4'920
Jul-14	US$ 0.05	US$ 4'905
Oct-14	US$ 0.06	US$ 5'731
Jan-15	US$ 0.04	US$ 3'970
Apr-15	US$ 0.05	US$ 4'211
Jul-15	US$ 0.015	US$ 1'363
Oct-15	US$ 0.014	US$ 1'245

- Labor, administration, management, utilities, reagents, and consumables
- Sequencing instruments and other large equipment (amortized over three years)
- Informatics activities directly related to sequence production (e.g., laboratory information management systems and initial data processing)
- Submission of data to a public database
- Indirect Costs (http://oamp.od.nih.gov/dfas/faq/indirect-costs#difference) as they relate to the above items

Slides freely downloadable from lectures.GersteinLab.org & “tweetable” (via @markgerstein)

Number genomes exploding since ‘10

Growth of Genome Data Repositories: Example EGA

The EGA contains a growing amount of data

International Cancer Genome Consortium

IRDiRC INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

UK 10K

wellcome trust sanger institute

** Files encrypted in different formats are counted only once*

European Genome-phenome Archive

Source: Jordi Rambla (CRG Barcelona) 2016

What is a PB, for human genomes? It depends.

- 2 bits per base are sufficient to encode TCGA
  - using 00, 01, 10, 11
  - [TCGA]{3’000’000’000}
  - 2 * 3 * 10^9b = 6, 000, 000, 000 b
  - perfect genome (no overhead): ~715 MB
  - 1PB => ~1400000 genomes
- according to Swiss online store (Sep 2017) ~45'000CHF (100x10TB disks
- this is less than a PhD position per year in Switzerland …
- (real costs are 2x that, + duplication, facilities, service … => ~500’000CHF)
- **However: A single 30x BAM file => 100GB**
- Still: 500’000CHF => 1PB => 10’000 genomes => 50CHF/genome (BAM format)

Bioinformatics: **File Formats**

- **text** or **binary** file formats, optimised for specific types of biological data
- examples from genomics:
  - **BAM** - compressed binary version of Sequence Alignment/ Map (SAM)
  - **BED** (Browser Extensible Data) - flexible way to define the data lines in an genome browser annotation tracks
  - **VCF** (Variant Call Format)

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

BED file example

The VCF file format
Standard for variant representation

Example

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body

```
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                FORMAT     SAMPLE1   SAMPLE2
1       1   .     ACG  A,AT    .   PASS    .                   GT:DP      1/2:13    0/0:29
1       2   rs1   C    T,CT    .   PASS    H2;AA=T             GT:GQ      0|1:100   2/2:70
1       5   .     A    G       .   PASS    .                   GT:GQ      1|0:77    1/1:95
1     100   .     T    <DEL>   .   PASS    SVTYPE=DEL;END=300  GT:GQ:DP   1/1:12:3 0/0:20
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

Deletion

SNP

Large SV

Insertion

Other event

Variant
Call
Format

- stores the results of a single or multiple interpretations of genome sequencing datasets, in comparison to a reference genome
- standard format for file-based storage of human genome variants

Global Alliance for Genomics & Health

Task: Estimate Storage Requirements for 1000 Genomes

- How much computer storage is required for 1000 Genomes
  - WES & WGS
  - Different file formats
    - SAM
    - BAM
    - VCF
    - FASTA
  - Associated costs
    - Cost afctors
    - Raw Storage costs

Reference Genome Resources...

RESOURCES FOR GENOMICS: UCSC GENOME BROWSER

- Originated from the Human Genome Project
- Most widely used general genome browser
- many default tracks
- many species
- customization with "BED" files

genome.ucsc.edu

RESOURCES FOR GENOMICS: HUMAN GENOME RESOURCES AT NCBI

- Entry point for genome reference data
- Human genome assemblies
- Human variant collections (dbVar, ClinVar, dbSNP) for download

www.ncbi.nlm.nih.gov/projects/genome/guide/human/

Download	GRCh38	GRCh37
Reference Genome Sequence	Fasta	Fasta
RefSeq Reference Genome Annotation	gff3	gff3
RefSeq Transcripts	Fasta	Fasta
RefSeq Proteins	Fasta	Fasta
ClinVar	vcf	vcf
dbSNP	vcf	vcf
dbVar	vcf	vcf

RESOURCES FOR GENOMICS: ENSEMBL

- Entry point for many genome data services and collections
- Downloads ("BioMart"), REST API

www.ensembl.org/Homo_sapiens/Info/Index

Where to find (cancer) genome data ...

Reference Resources for Human Genome Variants

- NCBI:dbSNP
  - single nucleotide polymorphisms (SNPs) and multiple small-scale variations
  - including insertions/deletions, microsatellites, non-polymorphic variants
- NCBI:dbVAR
  - genomic structural variation
  - insertions, deletions, duplications, inversions, multinucleotide substitutions, mobile element insertions, translocations, complex chromosomal rearrangements
- NCBI:ClinVar
  - aggregates information about genomic variation and its relationship to human health
- EMBL-EBI:EVA
  - open-access database of all types of genetic variation data from all species
- Ensembl
  - portal for many things genomic...

RESOURCES FOR CANCER GENOMICS

COSMIC
Catalogue of somatic mutations in cancer

Home ▼ | Resources ▼ | Curation ▼ | Tools ▼ | Data ▼ | News ▼ | Help ▼ | About ▼ | Search COSMIC... | Login ▼

COSMIC v79, released 14-NOV-16

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below, or by browsing a region of the human genome using the map to the right.

eg: Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell SEARCH

Browse the **genomic landscape** of cancer

R Resources

Key COSMIC resources

- Cell Lines Project
- COSMIC
- Whole Genomes
- Cancer Gene Census
- Drug Sensitivity
- Mutational Signatures
- GRCh37 Cancer Archive

T Tools

Additional tools to explore COSMIC

- Cancer Browser
- Genome Browser
- GA4GH Beacon
- CONAN

C Expert Curation

High quality curation by expert postdoctoral scientists

- Drug Resistance New
- Cancer Gene Census
- Curated Genes
- Gene Fusions
- Genome-Wide Screens

D Data

Further details on using COSMIC's content

- Downloads
- License
- Submission
- Genome Annotation
- Datasheets
- Help
- FAQ

Cancer Gene Census Update

7 genes have been added to the Cancer Gene Census -

EPAS1 - Endothelial PAS domain protein 1.
PTPRT - Protein tyrosine phosphatase, receptor type T.
PPM1D - Protein phosphatase, Mg2+/Mn2+ dependent 1D.
BTK - Bruton tyrosine kinase.
PREX2 - Phosphatidylinositol-3,4,5-trisphosphate dependent Rac exchange factor 2.
TP63 - Tumour protein p63.
QKI - QKI, KH domain containing RNA binding.

For full details, see the Datasheet.

RESOURCES FOR CANCER GENOMICS

National Cancer Institute — U.S. National Institutes of Health | www.cancer.gov

Cancer Genome Anatomy Project

CGAP How To | Tools | Genes | Chromosomes | Tissues | SAGE Genie | RNAi | Pathways

CGAP Info
- Educational Resources
- Slide Tour
- Team Members
- References

CGAP Data
- Download

Quick Links:
- ICG
- NCI Home
- NCICB Home
- NCBI Home
- OCG

Cancer Genome Anatomy Project (CGAP)

The NCI's Cancer Genome Anatomy Project sought to determine the gene expression profiles of normal, precancer, and cancer cells, leading eventually to improved detection, diagnosis, and treatment for the patient. Resources generated by the CGAP initiative are available to the broad cancer community. Interconnected modules provide access to all CGAP data, bioinformatic analysis tools, and biological resources allowing the user to find "in silico" answers to biological questions in a fraction of the time it once took in the laboratory.

The CGAP Website

Interconnected modules provide access to all CGAP data, bioinformatic analysis tools, and biological resources allowing the user to find "in silico" answers to biological questions in a fraction of the time it once took in the laboratory.

Genes
Gene information, clone resources, SNP500Cancer, GAI, and transcriptome analysis.

Tissues
cDNA library information, methods, and EST-based gene expression analysis.

Pathways
Diagrams of biological pathways and protein complexes, with links to genetic resources for each known protein.

RNAi
RNA-interference constructs, targeted specifically against cancer relevant genes. New addition: Validated set of shRNAs.

Chromosomes
FISH-mapped BAC clones, SNP500Cancer, and the Mitelman database of chromosome aberrations.

SAGE Genie
Analysis of gene expression using long and short SAGE tag data for both human and mouse.

Tools
Direct access to all analytic and data mining tools developed for the project.

International Cancer Genome Consortium

Enter keywords | Search

Home | Cancer Genome Projects | Committees and Working Groups | Policies and Guidelines | Media

ICGC Cancer Genome Projects

Committed projects to date: 89

Sort by: Project

Biliary Tract Cancer, Japan	Biliary Tract Cancer, Singapore	Bladder Cancer, China
Bladder Cancer, United States	Blood Cancer, China	Blood Cancer, Singapore
Blood Cancer, South Korea	Blood Cancer, United States	Blood Cancer, United States
Blood Cancer, United States	Blood Cancer, United States	Bone Cancer, France
Bone Cancer, United Kingdom	Bone Cancer, United States	Brain Cancer, Canada
Brain Cancer, China	Brain Cancer, United States	Brain Cancer, United States
Breast Cancer, China	Breast Cancer, European Union / United Kingdom	Breast Cancer, France
Breast Cancer, Mexico	Breast Cancer, South Korea	Breast Cancer, South Korea

ICGC Goal: To obtain a **comprehensive** description of **genomic, transcriptomic and epigenomic changes** in **50 different tumor types and/or subtypes** which are of clinical and societal importance across the globe.

Read more »

Launch Data Portal »

Apply for Access to Controlled Data »

Announcements

23/August/2016 - The ICGC Data Coordination Center (DCC) is pleased to announce ICGC data portal data release 22 (http://dcc.icgc.org).

ICGC data release 22 in total comprises data from more than 16,000 cancer donors spanning 70 projects and 21 tumour sites.

17/April/2016 - ICGCmed is pleased to announce the release of its white paper (http://icgcmed.org).

The International Cancer Genome Consortium for Medicine (ICGCmed) will link genomics data to clinical information, health and responses to therapies.

18/November/2015 - The International Cancer Genome Consortium (ICGC) PanCancer dataset generated by the PanCancer Analysis of Whole Genomes (PCAWG) study is now available on Amazon Web Services (AWS), giving cancer researchers access to over 2,400 consistently analyzed genomes corresponding to over 1,100 unique ICGC donors (https://icgc.org/icgc-in-the-cloud).

VARIANT RESOURCES FOR CANCER GENOMICS

Resource name	Primary institute	Constituent Knowledge base	Cancer focused	Therapeutic evidence	Predisp. evidence	Diagnostic evidence	Prognostic evidence	Variant emphasis	URL
Cancer Genome Interpreter (CGI)	Institute for Research in Biomedicine, Barcelona, Spain	x	x	x				Somatic	https://www.cancergenomeinterpreter.org/home
Clinical Interpretation of Variants in Cancer (CIViC)	Washington University School of Medicine (WashU)	x	x	x	x	x	x	All variants	www.civicdb.org
JAX Clinical Knowledgebase (CKB)	The Jackson Laboratory	x	x	x	x	x	x	Somatic	https://ckb.jax.org/
Molecular Match	Molecular Match	x	x	x			x	Somatic	https://app.molecularmatch.com/
OncoKB	Memorial Sloan Kettering Cancer Center	x	x	x				Somatic	http://oncokb.org/#/
Precision Medicine Knowledgebase (PMKB)	Weill Cornell Medical College	x	x	x	x	x	x	Somatic	https://pmkb.weill.cornell.edu/
BRCA exchange	GA4GH	x	x		x			Germline	http://brcaexchange.org/
Cancer Driver Log (CanDL)	Ohio State University (OSU) / James Cancer Hospital		x	x				Somatic	https://candl.osu.edu/
Gene Drug Knowledge Database	Synapse		x	x		x	x	Somatic	https://www.synapse.org/#!Synapse:syn2370773/wiki/62707
MatchMiner	Dana-Farber Cancer Institute		x					Somatic	http://bcb.dfci.harvard.edu/knowledge-systems/
COSMIC Drug Resistance Curation	Wellcome Trust Sanger Institute		x	x				Somatic	http://cancer.sanger.ac.uk/cosmic/drug_resistance
My Cancer Genome	Vanderbilt University		x	x		x	x	Somatic	https://www.mycancergenome.org/
NCI Clinical Trials	National Cancer Institute of the National Institutes of Health		x					Somatic	www.cancer.gov/about-cancer/treatment/clinical-trials
Personalized Cancer Therapy Database	MD Anderson Cancer Center		x	x	x	x	x	Somatic	https://pct.mdanderson.org/#/home
ClinGen Knowledge Base	ClinGen				x			Germline	https://www.clinicalgenome.org/resources-tools/
ClinVar	National Center for Biotechnology Information			x	x			All variants	http://www.ncbi.nlm.nih.gov/clinvar/
Pharmacogenomics Knowledgebase (PharmGKB)	Stanford University			x				Germline	https://www.pharmgkb.org/
The Human Gene Mutation Database (HGMD)	Institute of Medical Genetics in Cardiff				x			Germline	http://www.hgmd.cf.ac.uk

Wagner et al. A harmonized meta-knowledgebase of clinical interpretations of cancer genomic variants. Nature Genetics (under review 2018)

The arrayMap Cancer Genome Resource

arrayMap

- Search Samples
- Search Publications
- Gene CNA Frequencies
- User Data
- Array Visualization
- Progenetix

University of Zurich

- Citation
- User Guide
- Registration & Licensing
- People
- External Links ↗

FOLLOW US ON twitter

130.60.23.21

visualizing cancer genome array data @ arraymap.org

arrayMap is a curated reference database and bioinformatics resource targeting copy number profiling data in human cancer. The arrayMap database provides an entry point for meta-analysis and systems level data integration of high-resolution oncogenomic CNA data.

The current data reflects:

- 63060 genomic copy number arrays
- 763 experimental series
- 145 array platforms
- ICD-O 141 ICD-O cancer entities
- 554 publications (Pubmed entries)

Genomic copy number imbalances on chromosome 9 in a case of Glioblastoma (**GSM491153**), indcating, among others, a homozygous deletion involving CDKN2A/B.

ICD-O | Locus | ? | HG18 | HG19

For the majority of the samples, probe level visualization as well as customized data representation facilitate gene level and genome wide data review. Results from multi-case selections can be connected to downstream data analysis and visualization tools, as we provide through our Progenetix project.

arrayMap is developed by the group "Theoretical Cytogenetics and Oncogenomics" at the Institute of Molecular Life Sciences of the University of Zurich.

BRAIN TUMOURS	5653 samples ↗	[?]
BREAST CANCER	8329 samples ↗	[?]
COLORECTAL CANCER	3238 samples ↗	[?]
PROSTATE CANCER	991 samples ↗	[?]
STOMACH CANCER	1062 samples ↗	[?]

ARRAYMAP NEWS

2016-08-03: SVG graphics
2016-05-17: Transitioning to Europe PMC
More news ...

Feel free to use the data and tools for academic research projects and other applications. If more support and/or custom analysis is needed, please contact Michael Baudis regarding a collaborative project or a special license.

© 2000 - 2016 Michael Baudis, refreshed Mon, 19 Sep 2016 10:20:09 GMT in 6.87s on server 130.60.240.68. No responsibility is taken for the correctness of the data presented nor the results achieved with the Progenetix tools.

FIND CNAS BY GENE OR REGION: TP53 | [ERBB2] 17:35097862-35138441:1 [?]
REGION SIZE | MAX COVERAGE (KB): 0 kb - 5000 | 250000 kb [?]
CLINICAL DATA: no followup required
CITY: 20 km [?]
Query Database

1949 of 65042 cases matched the selection criteria.

SUBSET	PERCENT IN SUBSET
8507/3: Invasive micropapillary carcinoma (13/39)	33.3
C692: retina (14/82)	17.1
8260/3: Papillary adenocarcinoma, NOS (11/65)	16.9
8500/3: Invasive carcinoma of no special type (1201/8188)	14.7
8560/3: Adenosquamous carcinoma (3/21)	14.3
Carcinomas: breast ca. (1254/8837)	14.2
C50: breast (1254/8929)	14.0
8500/2: Ductal carcinoma in situ, NOS (25/225)	11.1
C32: larynx (3/29)	10.3
8010/2: Carcinoma in situ, NOS (2/20)	10.0
C187: sigmoid incl. rectosigmoid junction (13/140)	9.3
8480/3: Mucinous adenocarcinoma (12/132)	9.1
8522/3: Infiltrating duct and lobular carcinoma (4/44)	9.1
8460/3: Micropapillary serous carcinoma [C56.9] (32/513)	6.2
8130/1: Urothelial papilloma, NOS (11/184)	6.0
C680: other urinary organs (11/184)	6.0
C54: corpus uteri (19/330)	5.8
8441/3: Serous adenocarcinoma, NOS (31/542)	5.7
Carcinomas: esophagus ca. (32/571)	5.6
Carcinomas: gastric ca. (80/1492)	5.4

ICD Morphologies

2021 samples from arraymap have an associated "ICDMORPHOLOGYCODE" label.

9470/3: Medulloblastoma, NOS (M-94703)

Synonyms

- Medulloblastoma, NOS
- Melanotic medulloblastoma

UID	SERIESID	PMID	ICDMORPHOLOGYCODE	ICDTOPOGRAPHYCODE
GSM1000061	GSE36942	23457519	8070/3	C10
GSM1000062	GSE36942	23457519	8070/3	C10
GSM1001316	GSE40777	23571474	8070/3	C53
GSM1001317	GSE40777	23571474	8010/3	C34
GSM1001318	GSE40777	23571474	8070/3	C09
GSM1001319	GSE40777	23571474	8010/3	C34
GSM1002668	GSE40834	24047479	9823/3	C42
GSM1002669	GSE40834	24047479	9823/3	C42
GSM1002670	GSE40834	24047479	9823/3	C42
GSM1002671	GSE40834	24047479	9823/3	C42
GSM1002672	GSE40834	24047479	9823/3	C42
GSM1002673	GSE40834	24047479	9823/3	C42
GSM1002674	GSE40834	24047479	9823/3	C42
GSM1002675	GSE40834	24047479	9823/3	C42
GSM1002676	GSE40834	24047479	9823/3	C42
GSM1002677	GSE40834	24047479	9823/3	C42
GSM1002678	GSE40834	24047479	9823/3	C42
GSM1002679	GSE40834	24047479	9823/3	C42
GSM1002680	GSE40834	24047479	9823/3	C42

arrayMap

Population stratification in cancer samples based on SNP array data

- 2504 genome profiles from 1000 Genome project phase 1 as reference
- 5 superpopulations: South Asia, Europe, South America, East Asia and Africa.
- SNP positions used in 9 Affymetrix SNP arrays are extracted to train a population admixture model.

University of Zurich[UZH]

SIB

Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB | Swiss Institute of Bioinformatics
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

arraymap.org

progenetix.org

info.baudisgroup.org

sib.swiss/baudis-michael

imls.uzh.ch/en/research/baudis

Global Alliance
for Genomics & Health

elixir