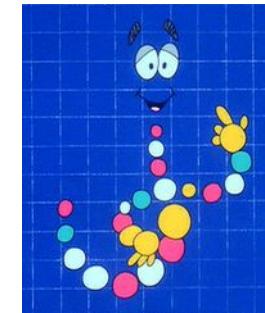


Bioinformatic tales and tools

Tristan M. Carland, PhD
for
SIO Bioinformatic Users Group (SIO-BUG)

Tales and Tools

- Career Overview
- Once I was a graduate student
 - Mega, Topali, Chimera, LibreOffice
- The tale(s) of too many projects
 - Unix, BWA, SamTools, GATK, SVS
 - What is a cluster and/or HPC?
- How does cancer work?
- What is it, you do here?
 - Scrum ftw, what is testing?

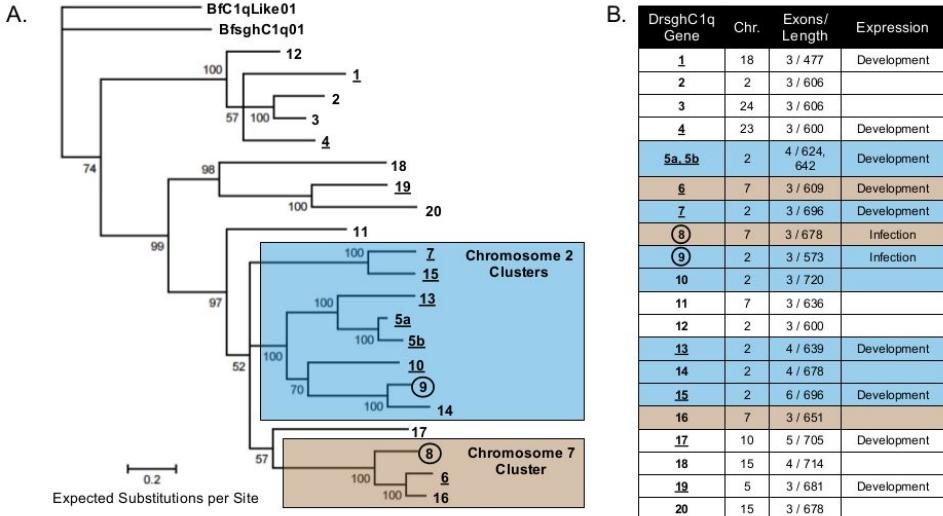
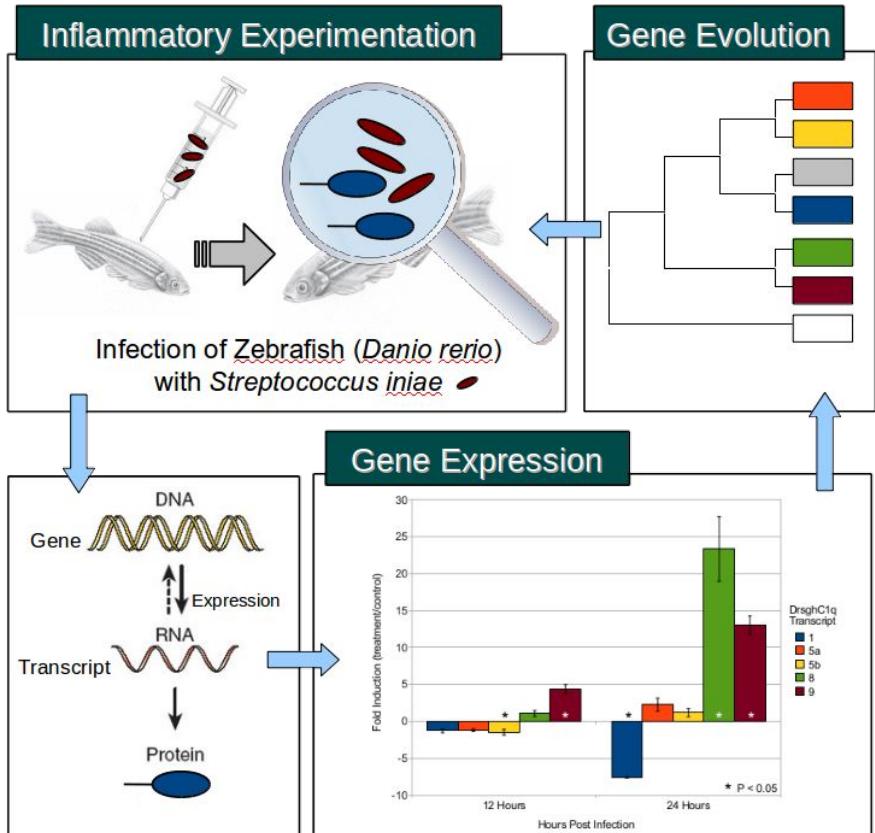


My path to wherever

1. UNCW - Majored in Computer Science
 - a. Added a major in Marine Biology (fish are cool)
 - b. Went to the Smithsonian (worms are cool)
2. Came to SIO for PhD in Marine Biology
 - a. Learned Molecular Biology (in the lab with Lena)
 - b. Dabbled in Bioinformatics
3. Postdoc in Genomic Medicine at TSRI
 - a. Moved to Human Biology at JCVI
4. Cancer Genomics Investigator with Avera
5. Freelance Consultant in Genomics
6. Software Engineer II (SDET) at Illumina



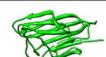
Once I was a “graduate student”



Carland, T. M., Locke, J. B., Nizet, V., & Gerwick, L. (2012). Differential expression and intrachromosomal evolution of the sghC1q genes in zebrafish (*Danio rerio*). *Developmental and Comparative Immunology*, 36(1), 31–38.

Carland, T. M., & Gerwick, L. (2010). The C1q domain containing proteins: Where do they come from and what do they do? *Developmental & Comparative Immunology*, 34(8), 785–790.

Globular Head C1q (ghC1q)
Both cghC1q and sghC1q proteins



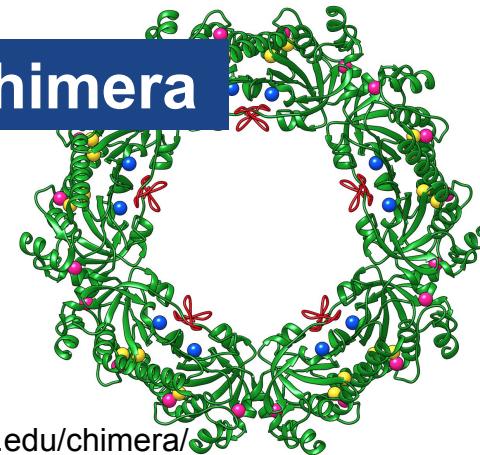
MEGA, Topali, Chimera, InkScape, LibreOffice

The screenshot shows the MEGA software interface with a title bar "MS: Alignment Explorer (zebrafish.coelacanth:P450)". The menu bar includes Data, Edit, Search, Alignment, Web, Sequencer, Display, Help. The main window displays a sequence alignment of 20 entries, each starting with a protein ID like CYP1A2, followed by a multi-colored sequence logo. The alignment is presented in a grid format with horizontal lines for gaps. Below the alignment, there is a legend for the color scheme and some descriptive text.

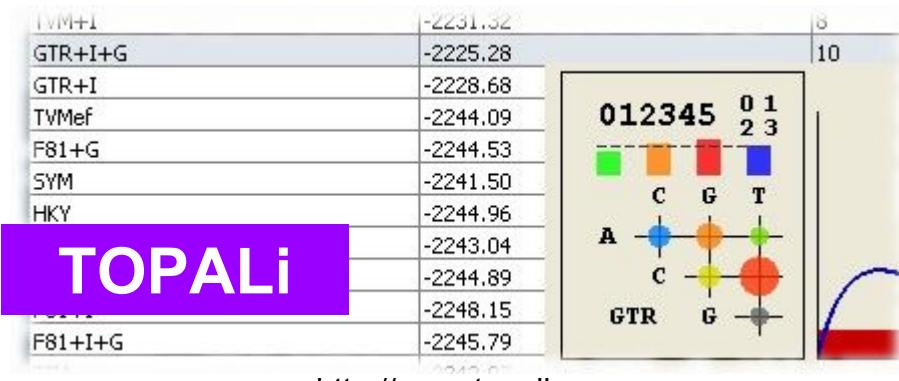
MEGA

<http://www.megasoftware.net/>

UCSF Chimera



<https://www.cgl.ucsf.edu/chimera/>



TOPALI

<http://www.topali.org>

Inkscape Project

Software Developer · inkscape.org

Inkscape is a free and open-source vector graphics editor; it can be used to create or edit vector graphics such as illustrations, diagrams, line arts, charts, logos and complex paintings.
[Wikipedia](#)

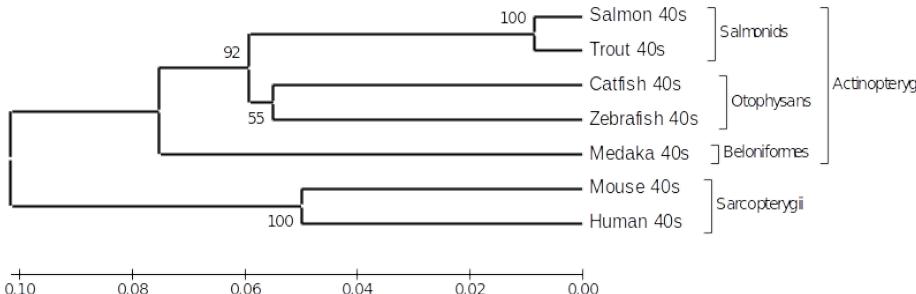
Founder: [Nathan Hurst](#)



<https://inkscape.org/en/>

The Three beats of Phylogenetics

- Sequence Choice (NCBI)
 - Organism and/or Gene (family)
- Multiple Sequence Alignment (MEGA)
 - Refine Gap Penalty Parameters
- Phylogenetic Reconstruction (TOPALi)
 - Many ways, many models



A collage of images related to phylogenetics. It features a man and a woman dancing, a large rainbow trout, a zebrafish, and a salmon. Below the collage is a screenshot of a computer interface for sequence alignment and phylogenetic analysis.

The screenshot shows the "Alignment Explorer" window from the MEGA software. The title bar reads "M5: Alignment Explorer (zebrafish.coelacanth.P450s.fasta)". The window displays a multiple sequence alignment of 20 protein sequences, each corresponding to a different CYP1 gene (CYP1IA2, CYP1B2, CYP1C1ZL, CYP1C2ZL, CYP1D1ZL, CYP2R1ZL, CYP2Y1ZL, CYP2Y3ZL, CYP2Y4ZL, CYP3A6G, CYP3C1ZL, CYP3C2L, CYP3C4L, CYP4F2, CYP4F4ZL, CYP4V7, CYP4V8, CYP5A1H). The sequences are color-coded by amino acid residue, and gaps are represented by dashes. The alignment shows high conservation of certain amino acids across the different species.

~80%

of your time doing bioinformatics will be spent trying to get things to work

just keep swimming

Personalized Genomic Medicine



The image shows the front cover of the book "The Creative Destruction of MEDICINE: How the Digital Revolution Will Create Better Health Care" by Eric Topol. The cover has a grey background with the title in large, light blue letters. Below the title, it says "HOW THE DIGITAL REVOLUTION WILL CREATE BETTER HEALTH CARE". At the bottom, there are five navigation links: "Home", "Praise", "Author", "Excerpt", and "Buy". A small red USB flash drive icon is positioned in the top right corner of the cover area.

About Eric Topol

SHARE

Eric Topol, M.D. is professor of genomics and holds the Scripps endowed chair in innovative medicine. He is the director of the Scripps Translational Science Institute in La Jolla, California. Previously, he led the Cleveland Clinic to its #1 ranking in heart care, started a new medical school, and led key discoveries in heart disease. He lives with his family in La Jolla, California.

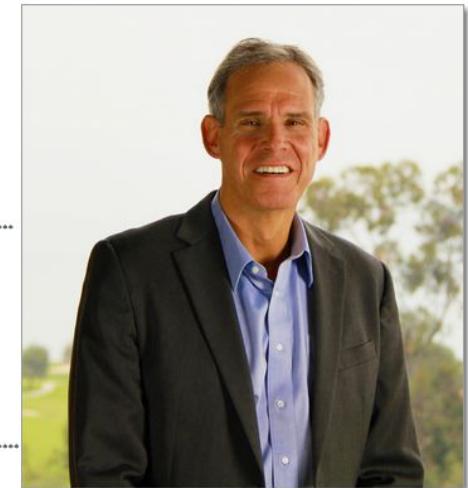
To learn more about Dr. Topol and his work, please visit his [Leigh Bureau Speakers](#) page as well as the [Scripps Translational Science Institute](#), [Scripps Health](#), [Scripps Research Institute](#), and [West Wireless Health Institute](#) websites.

Follow Dr. Topol on [Twitter](#), Facebook, and [LinkedIn](#).

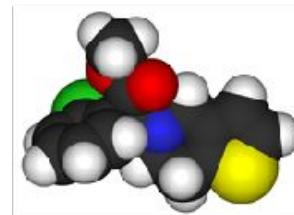
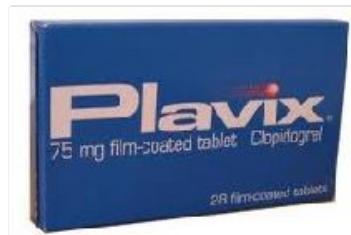
Watch his TED talk [here](#).

Dr. Topol on NPR's [Science Friday](#)

Watch *Medscape's One on One*: Dr. John Reed interviews Dr. Topol about how technology is changing medicine, making diagnosis and treatment faster, better, and more accurate.



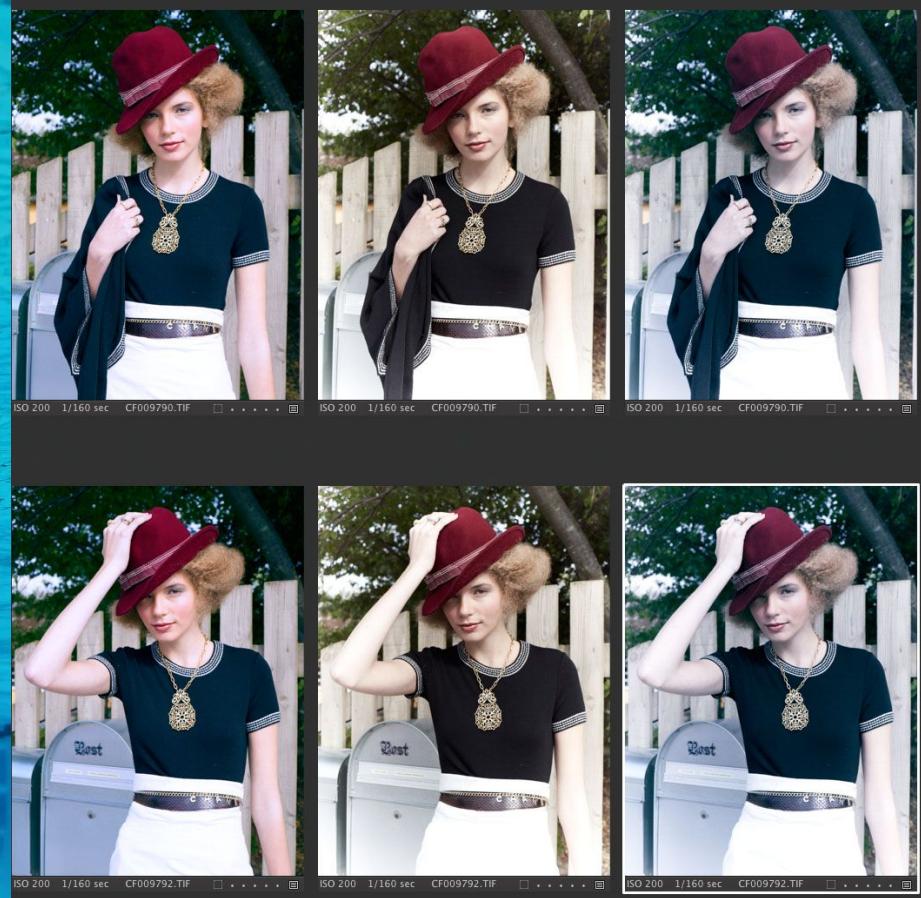
So you've had a heart attack



ATCGATCGCGCTGAGCATAGCGATCGCGCCGATCG
ATCGATCGCGCTGAGCATAGCGATCGCGCCGATCG
ATCGATCGCGCTGAGCATAGCGATCGCGCCGATCG
ATCGATCGCGCAYellowGAGCATAGCGATCGCGCCGATCG

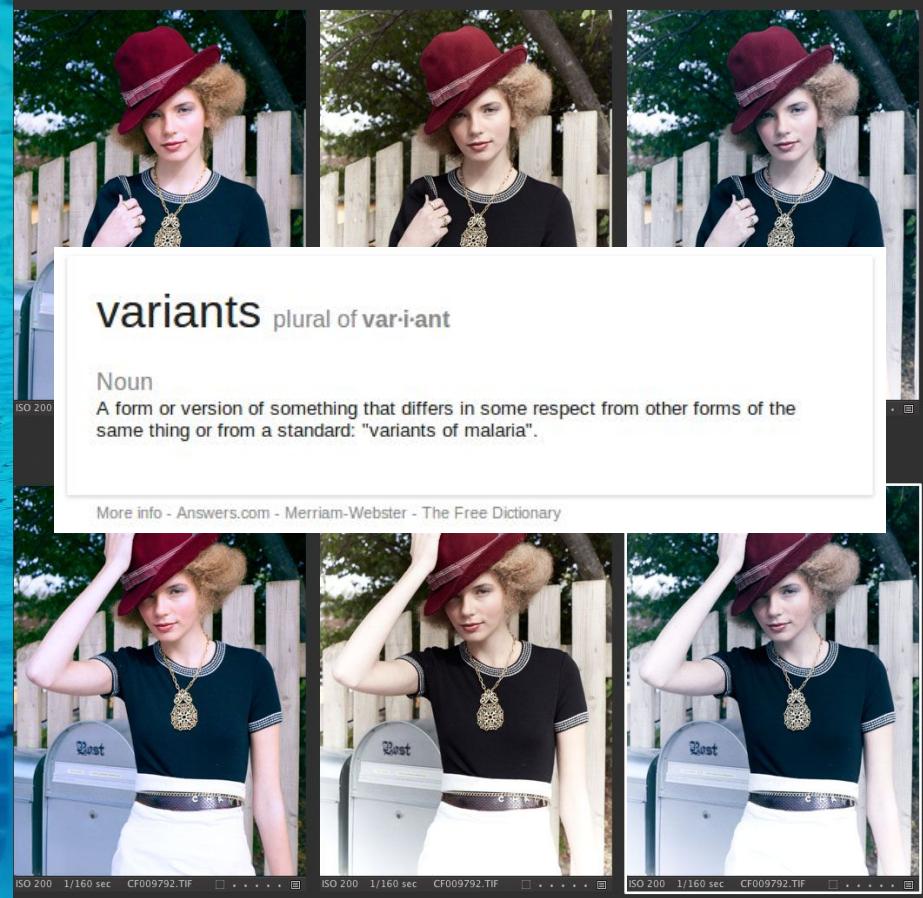
What genomic info should we save? Why?

- Every human genome should be over 99% identical
- Storing the raw data for each sample makes sense
- What should we store to run computations on?
 - We need something to compare easily



What genomic info should we save? Why?

- Every human genome should be over 99% identical
- Storing the raw data for each sample makes sense
- What should we store to run computations on?
 - We need something to compare easily
 - The differences are what define the samples



Variant Call Format

```
1 ##fileformat=VCFv4.1
2 ##FILTER=<ID=LowQual,Description="Low quality">
3 ##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
4 ##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
5 ##contig=<ID=1,length=249250621,assembly=b37>
6 ##contig=<ID=2,length=243199373,assembly=b37>
7 ##contig=<ID=GL000206.1,length=41001,assembly=b37>
8 ##contig=<ID=GL000240.1,length=41933,assembly=b37>
9 ##reference=file:///projects/stsi4/HUTs/HT020/scripts/../ref/human_g1k_v37.fasta
10 #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT HT020en HT020et
11 1 10177 . A AC 176.41 . AC=2;AF=0.500;AN=4;BaseQRankSum=-1.899;DP=38;FS=8.038;HaplotypeScore=125.8695;MLEAC=2;MLEAF=0.500;NPA=0.000;PValue=0.000;QD=4.11;RefDepth=38;S=0.000;VDB=-0.000;VQS=-0.000;WQS=-0.000;Zygosity=Heterozygous
12 1 10389 . AC A 288.40 . AC=2;AF=0.500;AN=4;BaseQRankSum=2.179;DP=71;FS=10.976;HaplotypeScore=272.6294;MLEAC=2;MLEAF=0.500;NPA=0.000;PValue=0.000;QD=4.11;RefDepth=71;S=0.000;VDB=-0.000;VQS=-0.000;WQS=-0.000;Zygosity=Heterozygous
13 1 10396 . AC A 287.40 . AC=2;AF=0.500;AN=4;BaseQRankSum=1.603;DP=68;FS=3.176;HaplotypeScore=269.5334;MLEAC=2;MLEAF=0.500;NPA=0.000;PValue=0.000;QD=4.11;RefDepth=68;S=0.000;VDB=-0.000;VQS=-0.000;WQS=-0.000;Zygosity=Heterozygous
14 1 10439 rs112766696 AC A 239.40 . AC=2;AF=0.500;AN=4;BaseQRankSum=1.599;DB;DP=50;FS=5.969;HaplotypeScore=236.8964;MLEAC=2;MLEAF=0.500;NPA=0.000;PValue=0.000;QD=4.11;RefDepth=50;S=0.000;VDB=-0.000;VQS=-0.000;WQS=-0.000;Zygosity=Heterozygous
15 1 61989 rs77573425 G C 48.17 . AC=2;AF=1.00;AN=2;DB;DP=3;Dels=0.00;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;NPA=0.000;PValue=0.000;QD=4.11;RefDepth=3;S=0.000;VDB=-0.000;VQS=-0.000;WQS=-0.000;Zygosity=Homozygous
16 1 569984 . A C 35.20 . AC=1;AF=0.500;AN=2;BaseQRankSum=1.495;DP=10;Dels=0.00;FS=3.680;HaplotypeScore=0.0000;MLEAC=1;MLEAF=0.500;NPA=0.000;PValue=0.000;QD=4.11;RefDepth=10;S=0.000;VDB=-0.000;VQS=-0.000;WQS=-0.000;Zygosity=Heterozygous
17 1 706330 . A AG 183.59 . AC=3;AF=0.750;AN=4;BaseQRankSum=-0.751;DP=20;FS=0.000;HaplotypeScore=1.8880;MLEAC=3;MLEAF=0.750;NPA=0.000;PValue=0.000;QD=4.11;RefDepth=20;S=0.000;VDB=-0.000;VQS=-0.000;WQS=-0.000;Zygosity=Heterozygous
18 1 714427 rs12028261 G A 743.20 . AC=4;AF=1.00;AN=4;DB;DP=32;Dels=0.00;FS=0.000;HaplotypeScore=0.0000;MLEAC=4;MLEAF=1.00;NPA=0.000;PValue=0.000;QD=4.11;RefDepth=32;S=0.000;VDB=-0.000;VQS=-0.000;WQS=-0.000;Zygosity=Homozygous
19 1 715348 rs3131984 T G 158.02 . AC=4;AF=1.00;AN=4;DB;DP=5;Dels=0.00;FS=0.000;HaplotypeScore=0.0000;MLEAC=4;MLEAF=1.00;NPA=0.000;PValue=0.000;QD=4.11;RefDepth=5;S=0.000;VDB=-0.000;VQS=-0.000;WQS=-0.000;Zygosity=Homozygous
20 1 723798 rs34882115 CAG C 1003.16 . AC=4;AF=1.00;AN=4;DB;DP=24;FS=0.000;HaplotypeScore=5.4571;MLEAC=4;MLEAF=1.00;MQ=0;NPA=0.000;PValue=0.000;QD=4.11;RefDepth=24;S=0.000;VDB=-0.000;VQS=-0.000;WQS=-0.000;Zygosity=Homozygous
21 1 723891 rs2977670 G C 370.27 . AC=4;AF=1.00;AN=4;DB;DP=12;Dels=0.00;FS=0.000;HaplotypeScore=0.0000;MLEAC=4;MLEAF=1.00;NPA=0.000;PValue=0.000;QD=4.11;RefDepth=12;S=0.000;VDB=-0.000;VQS=-0.000;WQS=-0.000;Zygosity=Homozygous
```

Variant Call Format

	#CHROM	POS	ID	REF	ALT	QUAL	HT020en	HT020et
1	1	10177	0	A	AC	176.41	0/1:5,4:9:24:67,0,24	0/1:6,5:11:71:147,0,71
1	1	10389	0	AC	A	288.4	0/1:7,7:14:99:140,0,154	0/1:10,9:22:99:186,0,304
2	##file	1	10396	0	AC	287.4	0/1:6,8:14:99:209,0,124	0/1:11,8:21:99:116,0,270
3	##FILT	1	10439	rs112766696	AC	239.4	0/1:10,4:14:93:93,0,241	0/1:7,7:17:99:184,0,163
4	##FORM	1	61989	rs77573425	G	48.17	1/1:1,2:3:6:74,6,0	<u>J</u>
5	##INFO	1	569984	0	A	35.2	0/1:6,2:8:62:62,0,137	<u>J</u>
6	##cont	1	706330	0	A	183.59	0/1:1,3:4:15:80,0,15	1/1:0,5:5:15:142,15,0
7	##cont	1	714427	rs12028261	G	743.2	1/1:0,13:13:33:377,33,0	1/1:1,17:19:33:393,33,0
8	##cont	1	715348	rs3131984	T	158.02	1/1:0,2:2:6:78,6,0	1/1:0,3:3:9:106,9,0
9	##refe	1	723798	rs34882115	CAG	1003.16	1/1:0,13:13:39:547,39,0	1/1:0,11:11:33:492,33,0
10	#CHROM	1	723891	rs2977670	G	370.27	1/1:0,6:6:18:193,18,0	1/1:0,6:6:18:204,18,0
11	1	725104	rs201386617	C	A	42.42	0/0:1,0:1:3:0,3,33	0/1:1,2:3:24:70,0,24
12	1	725106	rs200821612	C	G	38.42	0/0:1,0:1:3:0,3,33	0/1:1,2:3:24:66,0,24
13	1	726202	0	C	T	53.44	0/1:2,2:4:52:55,0,52	0/1:2,1:3:27:27,0,39
14	1	726481	rs3131980	T	G	205.87	1/1:0,4:4:12:153,12,0	1/1:0,2:2:6:79,6,0
15	1	729679	rs4951859	C	G	197.26	<u>J</u>	1/1:0,6:6:18:224,18,0
16	1	6198	1	734460	0	TG	T	1/1:0,3:4:9:109,9,0
17	1	5699	1	734491	rs7518433	T	C	1/1:0,2:2:3:39,3,0
18	1	7063	1	734566	rs139489325	G	A	1/1:0,1:1:3:38,3,0
19	1	7144	1	739142	rs2340527	T	A	0/1:7,9:16:99:214,0,185
20	1	7153	1	752566	rs3094315	G	A	1/1:0,8:8:24:302,24,0
21	1	7237	1	752721	rs3131972	A	G	1/1:1,67:67:99:2327,181,0
22	1	7238	1	754182	rs3131969	A	G	1/1:0,5:5:12:155,12,0
		1	754192	rs3131968	A	123.52	1/1:0,2:2:3:34,3,0	1/1:0,4:4:9:115,9,0
		1	754334	rs3131967	T	151.02	1/1:0,2:2:6:76,6,0	1/1:0,3:3:9:101,9,0

Variant Call Format

	#CHROM	POS	ID	REF	ALT	QUAL	HT020en	HT020et	
1	1	10177	0	A	AC	176.41	0/1:5,4:9:24:67,0,24	0/1:6,5:11:71:147,0,71	
1	1	10389	0	AC	A	288.4	0/1:7,7:14:99:140,0,154	0/1:10,9:22:99:186,0,304	
2	##FILT	1	10396	0	AC	A	287.4	0/1:6,8:14:99:209,0,124	0/1:11,8:21:99:116,0,270
3	##FORM	1	10439	rs112766696	AC	A	239.4	0/1:10,4:14:93:93,0,241	0/1:7,7:17:99:184,0,163
4	##INFO	1	61989	rs77573425	G	C	48.17	1/1:1,2:3:6:74,6,0	
5	##cont	1	569984	0	A	C	35.2	0/1:6,2:8:62,6,0	
6	##cont	1	714427	rs112766696	AG	T	743.2	1/1:1,3:4:15,0	
7	##cont	1	715348	rs3131984	T	G	158.02	1/1:0,2:2:6:76,6,0	
8	##cont	1	723798	rs34882115	CAG	C	1003.16	1/1:0,1:3:13:39:547,39,0	
9	##refe	1	1 : 10177	rs2977791	A	AC	370.0	0/1 1/1:0,6:6:18:193,0	
10	#CHROM	1	725104	rs201386617	C	A	42.42	0/0:1,0:1:3:0,3,33	
11	1	1017	725106	rs200821612	C	G	38.42	0/0:1,0:1:3:0,3,33	
12	1	1038	1 : 729679	0	C	T	53.44	./. 0/1:2,2:4:52:55,0,32	
13	1	1039	726481	rs3131980	T	G	205.87	1/1:0,4:4:12:153,12,0	
14	1	1043	1 : 888234	rs495199	A	G	197.2	0/0 1/1:0,6:6:18:193,0	
15	1	6198	1 : 998233	rs734460	0	TG	173.71	1/1:0,3:4:9:109,9,0	
16	1	5699	734491	rs7518433	T	C	47.42	1/1:0,2:2:3:39,3,0	
17	1	7063	1 : 998233	rs139489325	G	A	38.41	0/1:0,1:1:3:38,3,0	
18	1	7144	739142	rs2340527	F	A	359.44	0/1:7,0:16:0:214,0,185	
19	1	7153	1 : 998233	rs309455	G	A	844.20	1/1:0,8:8:24:302,1,0	
20	1	7237	752721	rs3131972	T	A	5290.2	1/1:1,6:1:67:99:2327,181,0	
21	1	7238	1	754182	rs3131969	A	314.4	1/1:0,5:5:12:155,12,0	
		1	754192	rs3131968	A	G	123.52	1/1:0,2:2:3:34,3,0	
		1	754334	rs3131967	T	C	151.02	1/1:0,2:2:6:76,6,0	
								1/1:0,3:3:9:101,9,0	

Where is it?

What is the reference?

What is an alternative?

Our “normal” sample

Our “tumor” sample

Is it somatic?

No

Can't say

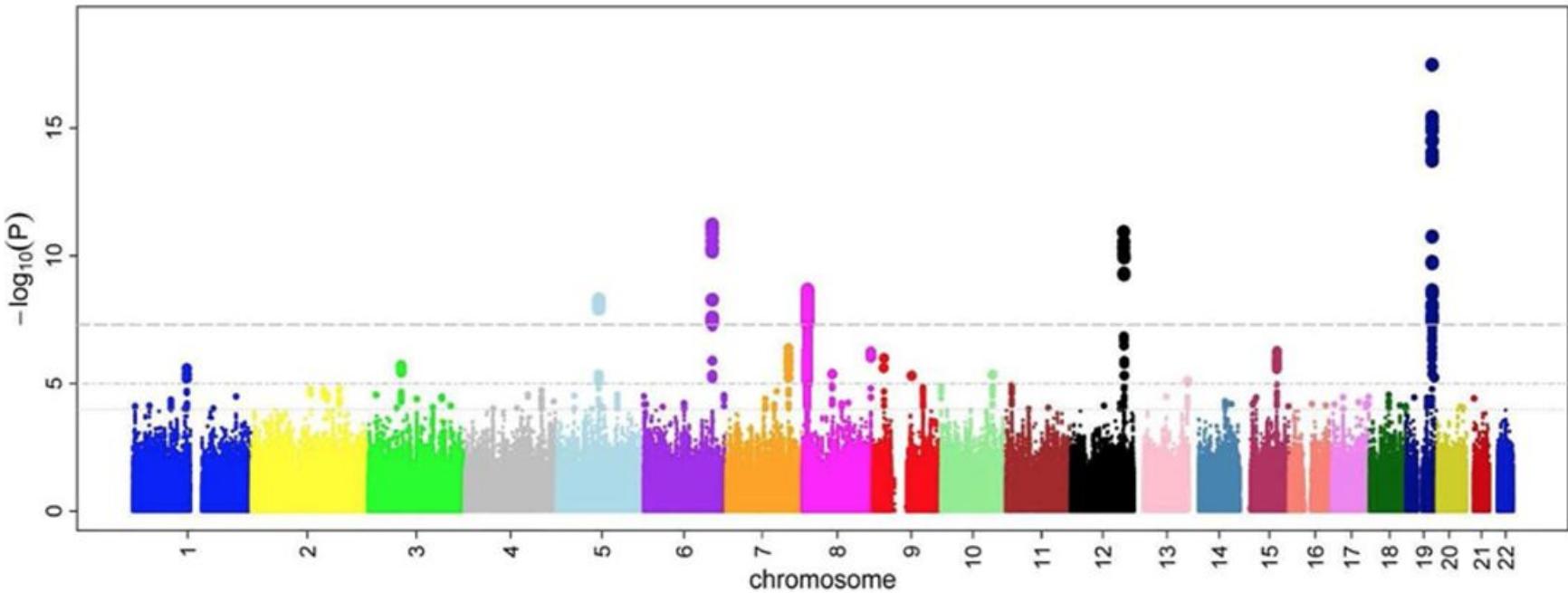
Yes

?

?

?

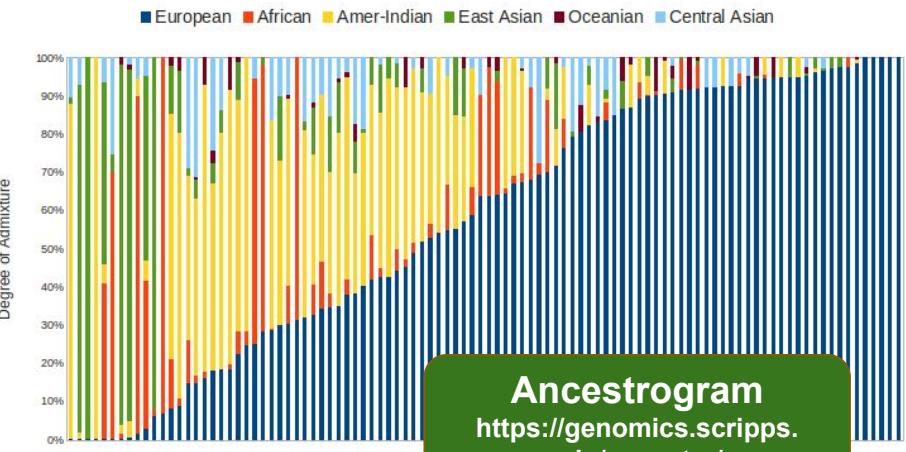
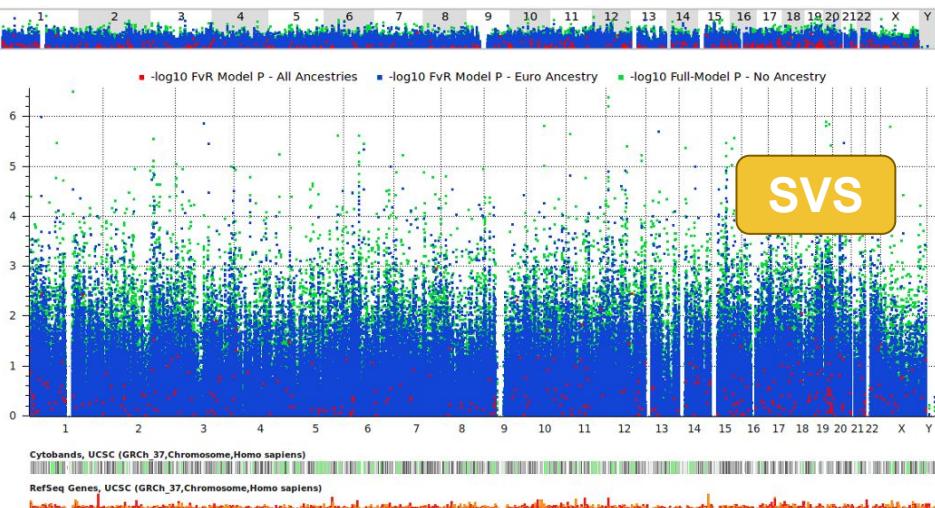
Manhattan Plots - great significance from small variants



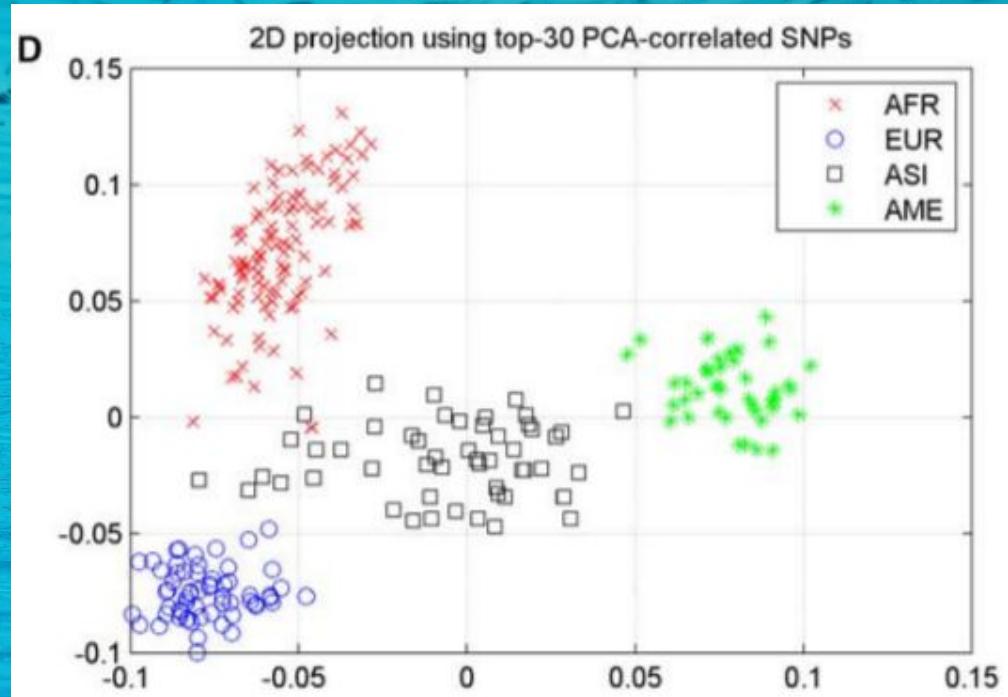
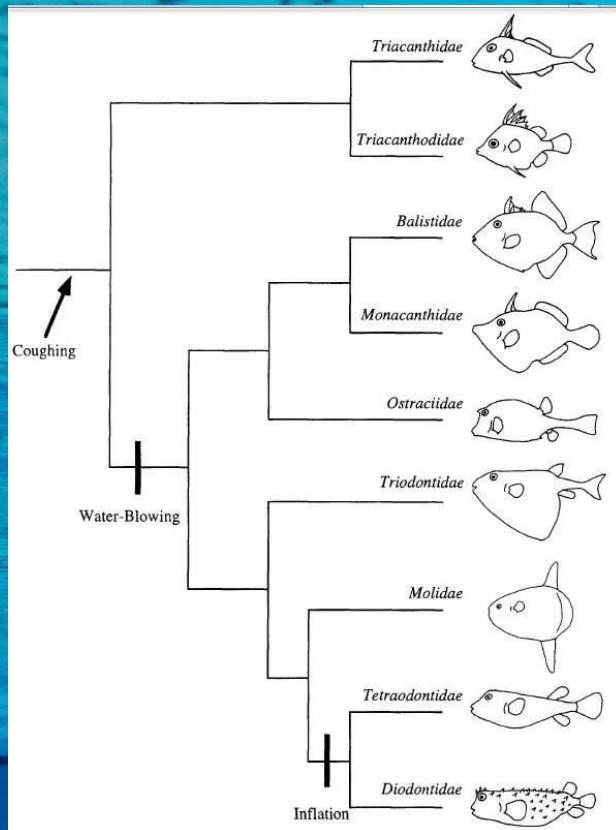
Each dot is a location within the genome, the $-\log_{10} P$ score denotes how well the variation at that position associates with the phenotype of interest.

Projects

- Personalized Cancer Medicine
- Alcoholism in Native Americans
- Group Variant Calling
- Mutated tRNA in Zebrafish
- Kidney Transplant Exome GWAS
- Cow Antibody Sequencing
- Eating Disorder Haplotype Study
- Dog Genome Assembly



Research Idea - Phylogenetic and PCA Clustering



PCA-Correlated SNPs
for Structure Identification
in Worldwide Human Populations

Peristera Paschou^{1*}, Elad Ziv^{2,3,4}, Esteban G. Burchard^{5,6}, Shweta Choudhry⁷, William Rodriguez-Cintron⁸, Michael W. Mahoney⁹, Petros Drineas¹⁰

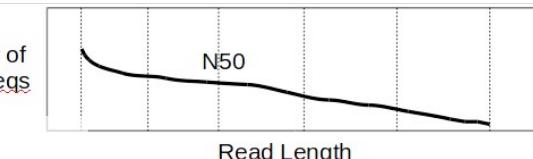
Projects

- Personalized Cancer Medicine
- Alcoholism in Native Americans
- Group Variant Calling
- Mutated tRNA in Zebrafish
- Kidney Transplant Exome GWAS
- Cow Antibody Sequencing
- Eating Disorder Haplotype Study
- Dog Genome Assembly!

Build	Sequences	Total Length	Avg Length	N0	N25	N50	N75	N90	N100
Ihm21	3531269	3479947123	985	100	3483	10420	21491	35262	158386
Ihm25	5703282	3661489441	641	100	3487	17086	38545	65633	267301
Ihm29	7799252	3771706238	483	100	1309	27886	68053	118654	524410
Iha29	7758992	3774135425	486	100	1827	30472	74333	130182	678427
Ihm31	8795319	3807721776	432	100	500	41348	106387	193997	758707
Iha31	8805458	3810569526	432	100	510	43342	112099	203252	975104
Ihm33	9858777	3829030194	388	100	108	58868	160049	298292	1846276
Iha33	9901462	3833244198	387	100	108	60492	163721	309579	1314888
Ihb33	9899263	3831539101	387	100	108	57915	156252	293855	1312357
Ihm37	2313388	2293187894	991	100	1002	2039	3641	5637	23406
Ihm41	2098570	2329859104	1110	100	1143	2347	4223	6584	27564
Ihm55	1988803	2415984266	1214	100	1394	3089	5870	9448	56111
Ihm95	459987	160020301	347	100	301	341	692	972	8080



• Lundehund

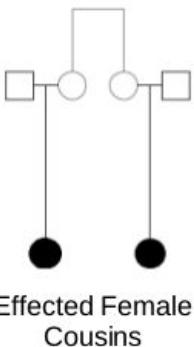


- 108bp reads, 500bp inserts
- 108bp reads, 700bp inserts
- 50Bp reads, 500bp inserts
- 36Bp reads, 3kbp inserts

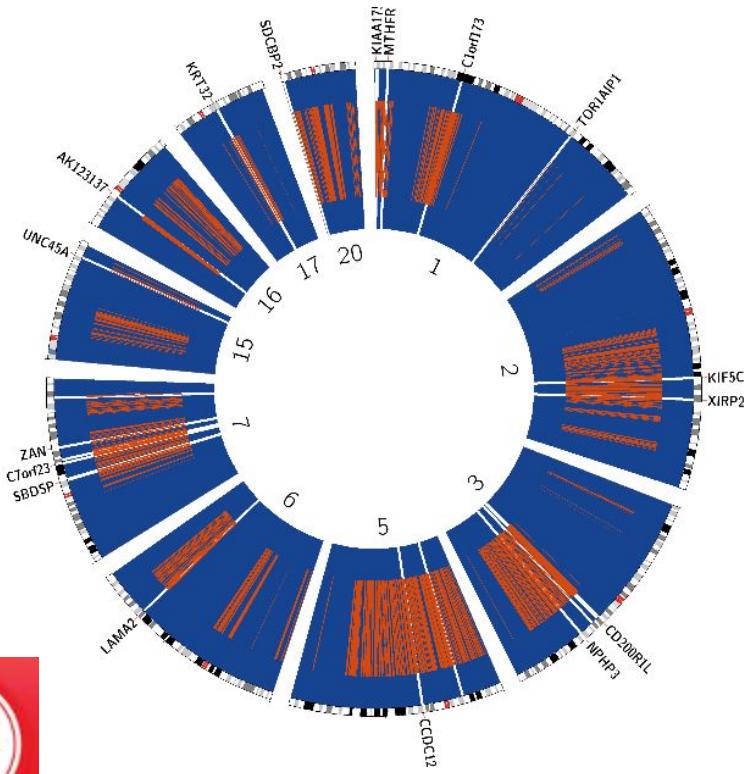
- Total Length 68,904,516,540
- 68 gbp

Projects

- Personalized Cancer Medicine
- Alcoholism in Native Americans
- Group Variant Calling
- Mutated tRNA in Zebrafish
- Kidney Transplant Exome GWAS
- Cow Antibody Sequencing
- Eating Disorder Haplotype Study
- Dog Genome Assembly



Effected Female Cousins



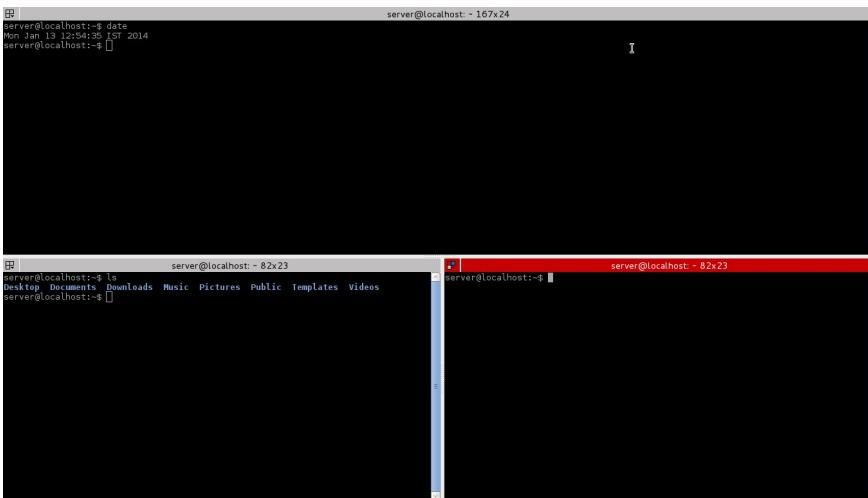
circos.ca

Shih et al. 2015 - submitted

Linux, Terminator, Sublime, awk+sed+grep



ubuntu.



Sublime Text

Sublime Text is a sophisticated text editor for code, markup and prose. You'll love the slick user interface, extraordinary features and amazing performance.

A screenshot of the Sublime Text editor window titled "Demonstration". The file is named "untitled" and contains Python code related to motion modes. The code includes imports for sublime and sublime_plugin, definitions for MOTION_MODE_NORMAL and MOTION_MODE_LINE, and a class InputState with methods for repeat digits and setting actions. The status bar at the bottom right shows "Line 1, Column 1", "Spaces: 4", and "Plain Text".

The Command Palette gives fast access to functionality. Here ⌘P is used to show the Command Palette, "sspy" (short for Set Syntax: Python) is used set the syntax of the current file to Python.

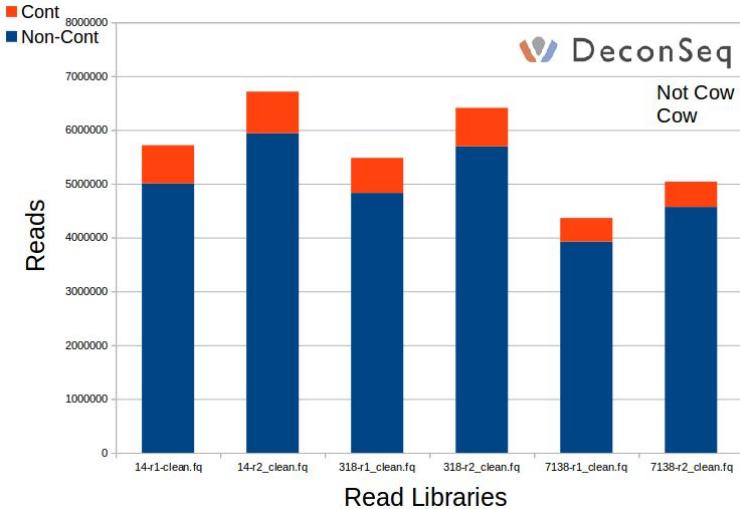
<http://www.sublimetext.com/>

Grep, AWK, SED, bash/sh/csh - learn them

- Grep - find lines in files
 - grep "thing" <file>
- awk - super versatile but I use it to parse columns and print things
 - awk -F "separator" '{print \$1" is awesome"}' <file>
- sed - replace things in files
 - sed -i s/aThing/aNewThing/g <file>
- bash/sh/csh - the shell where you run things
 - You can make scripts in here directly
 - <http://tldp.org/HOWTO/Bash-Prog-Intro-HOWTO.html>
- Perl would be good to learn (or Python)
 - <http://perldoc.perl.org/perlintro.html> - I still keep this bookmarked

Projects

- Personalized Cancer Medicine
- Alcoholism in Native Americans
- Group Variant Calling
- Mutated tRNA in Zebrafish
- Kidney Transplant Exome GWAS
- Cow Antibody Sequencing
- Eating Disorder Haplotype Study
- Dog Genome Assembly

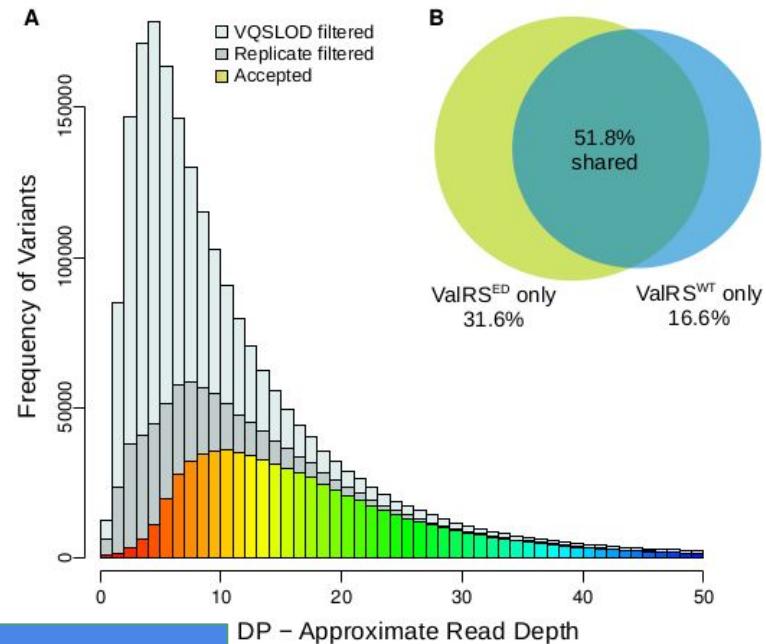


 DeconSeq

edwards.sdsu.edu/deconseq/

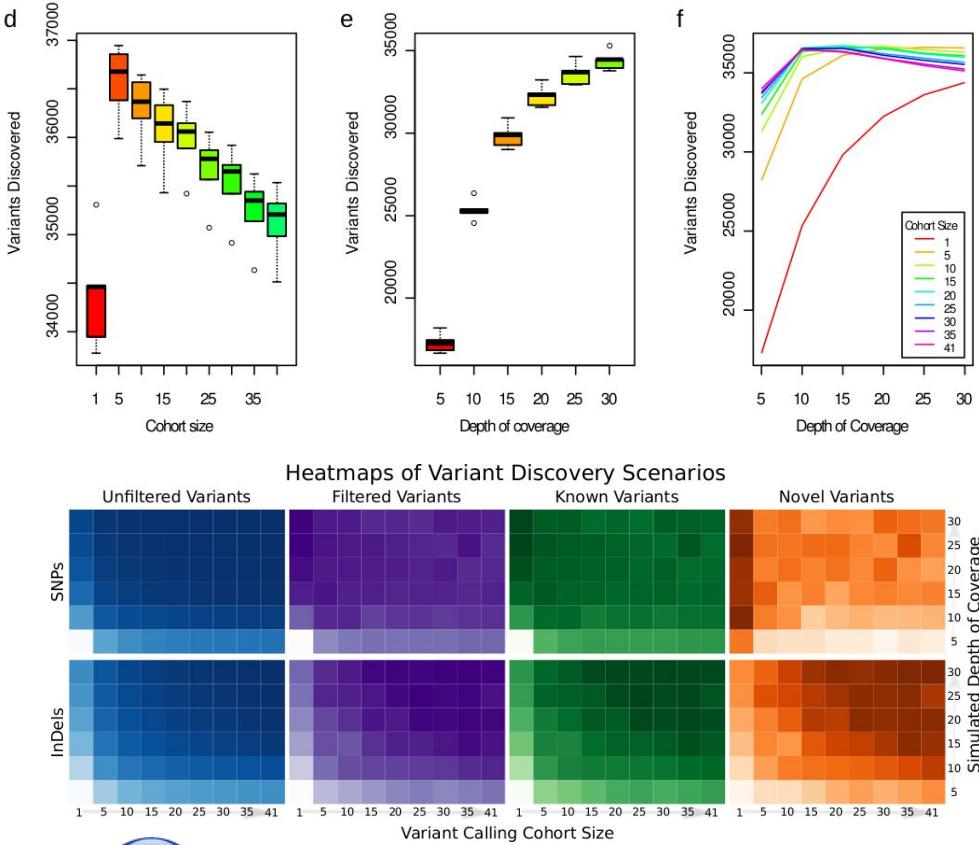
Projects

- Personalized Cancer Medicine
- Alcoholism in Native Americans
- Group Variant Calling
- Mutated tRNA in Zebrafish
- Kidney Transplant Exome GWAS
- Cow Antibody Sequencing
- Eating Disorder Haplotype Study
- Dog Genome Assembly

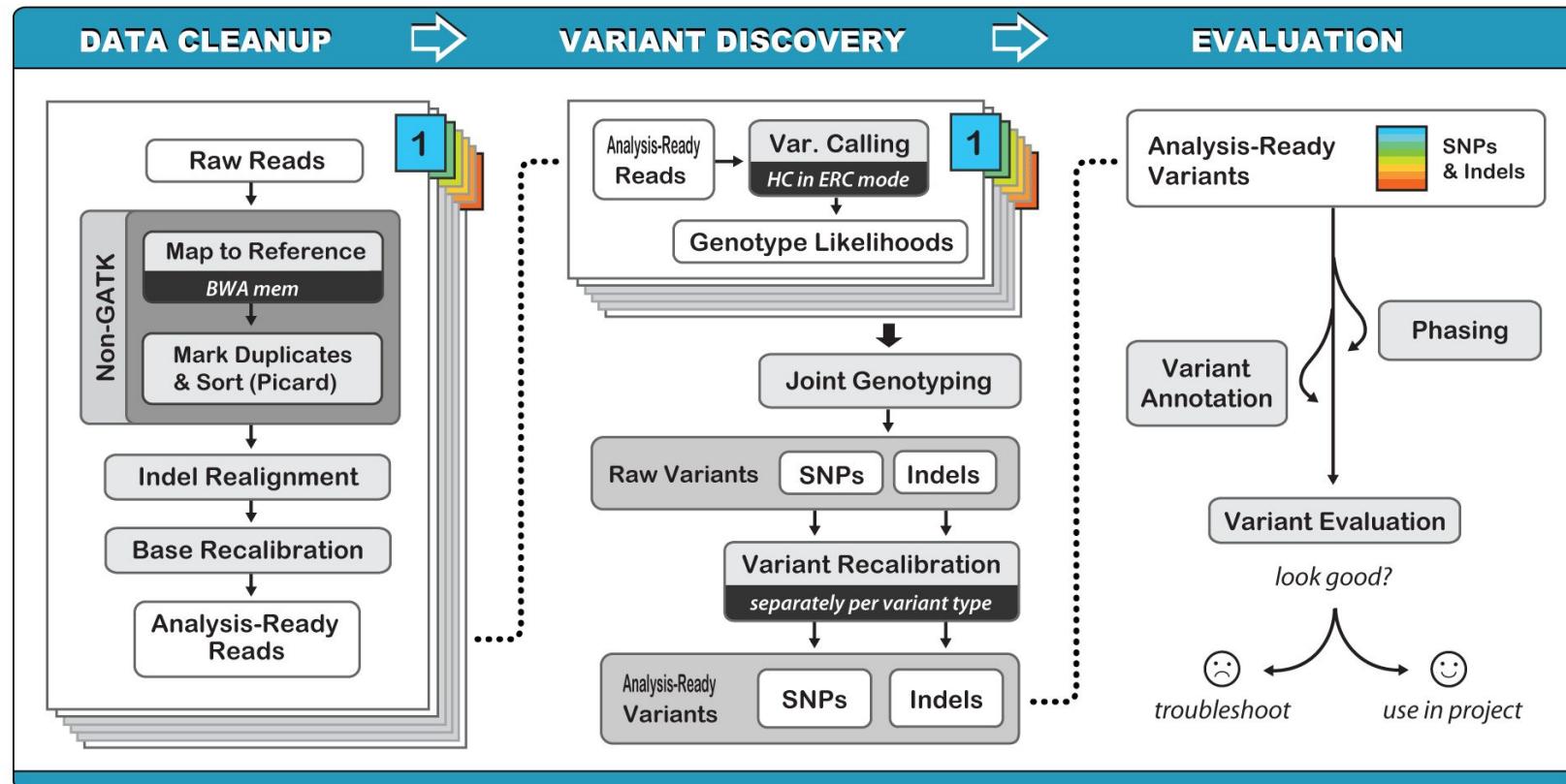


Projects

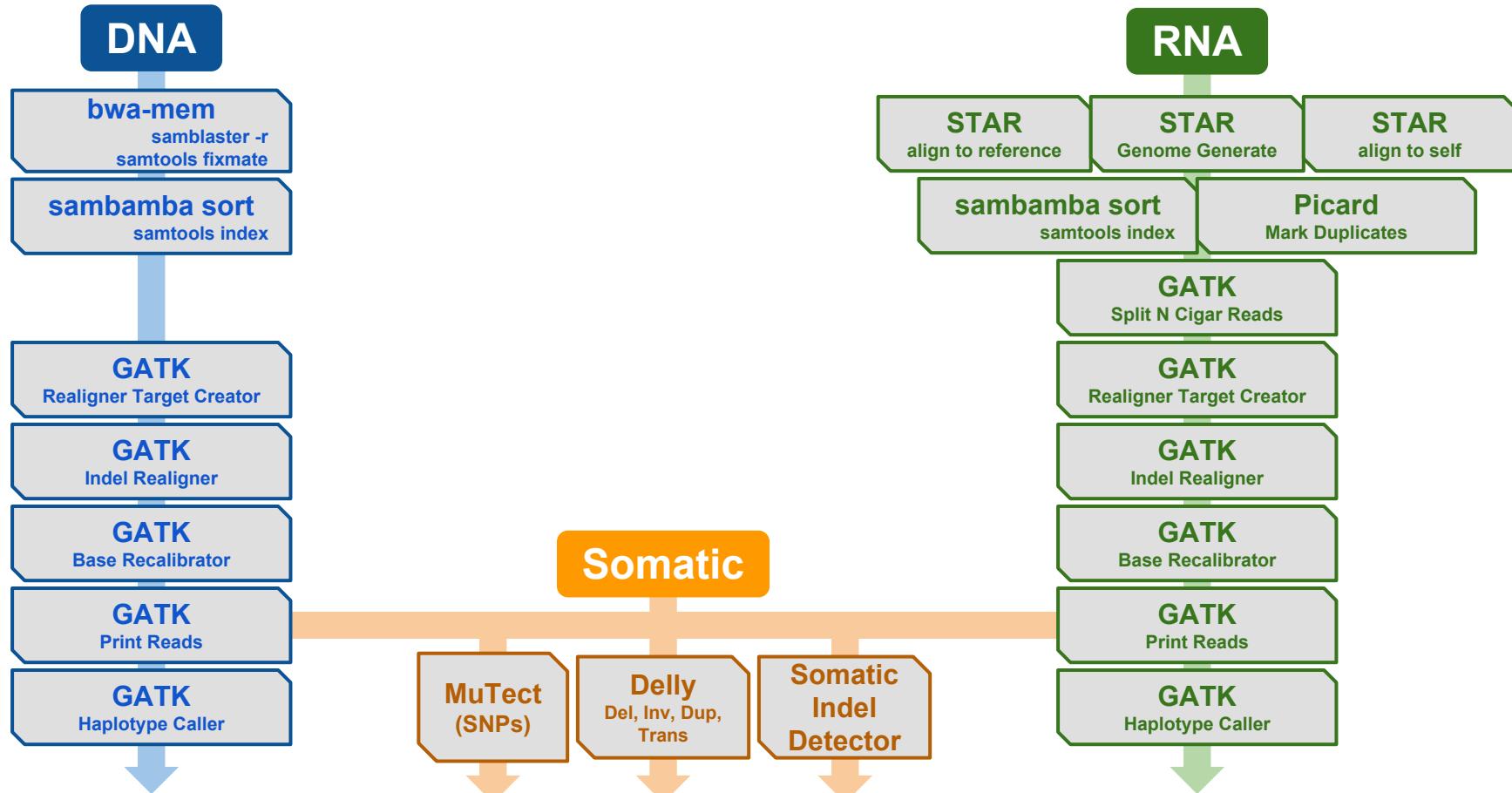
- Personalized Cancer Medicine
- Alcoholism in Native Americans
- **Group Variant Calling**
- Mutated tRNA in Zebrafish
- Kidney Transplant Exome GWAS
- Cow Antibody Sequencing
- Eating Disorder Haplotype Study
- Dog Genome Assembly



Genome Analysis ToolKit - Best Practices

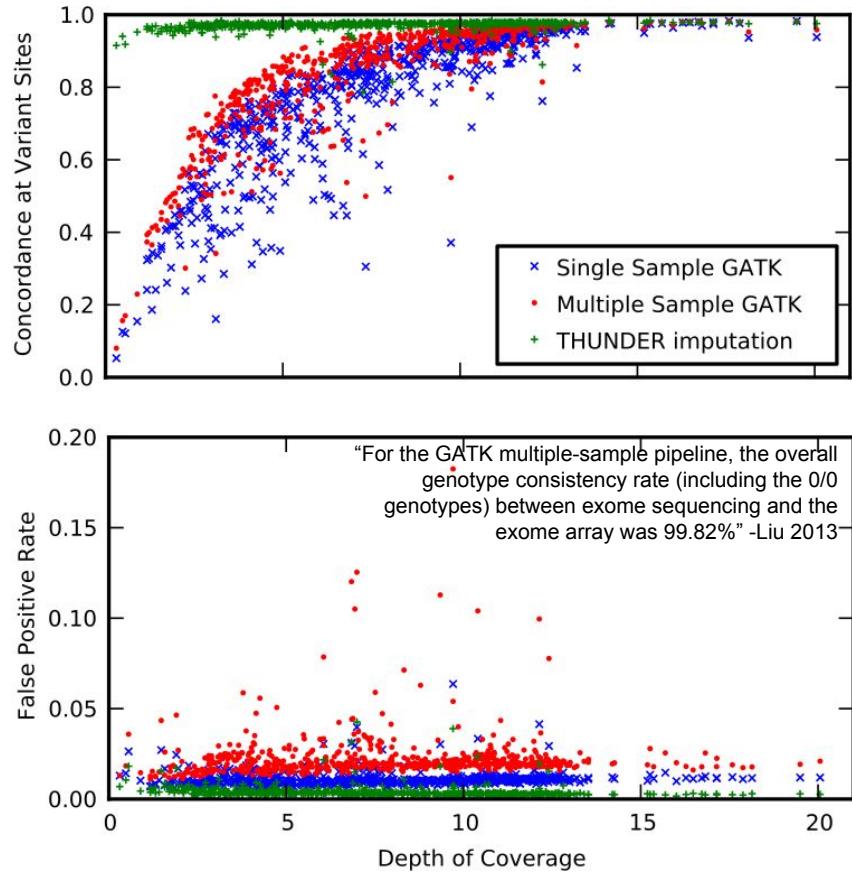


My latest GATK Best Practices pipeline



Projects

- Personalized Cancer Medicine
- Alcoholism in Native Americans
- Group Variant Calling
- Mutated tRNA in Zebrafish
- Kidney Transplant Exome GWAS
- Cow Antibody Sequencing
- Eating Disorder Haplotype Study
- Dog Genome Assembly

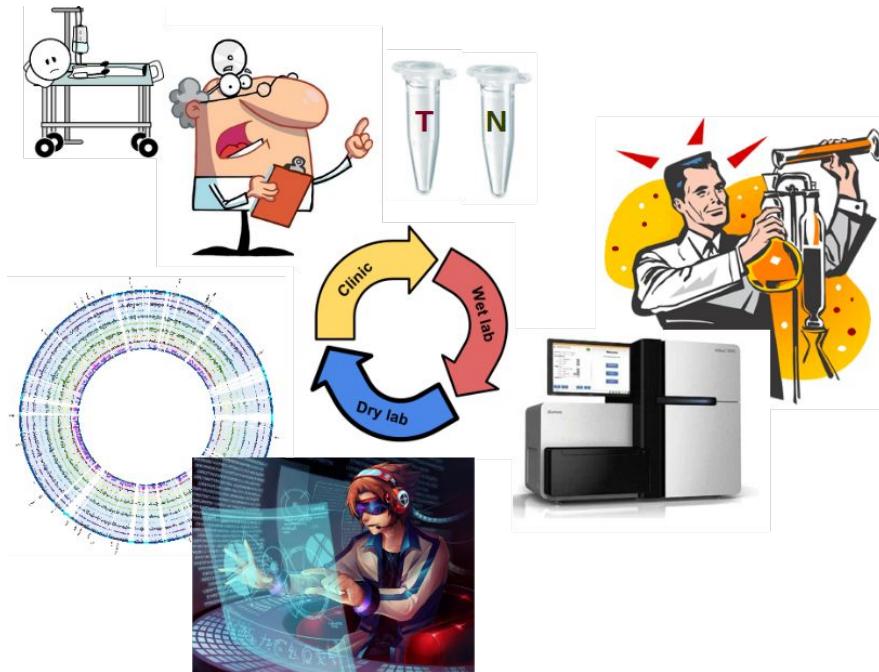


High Performance Computing

(see blackboard for details)

Projects

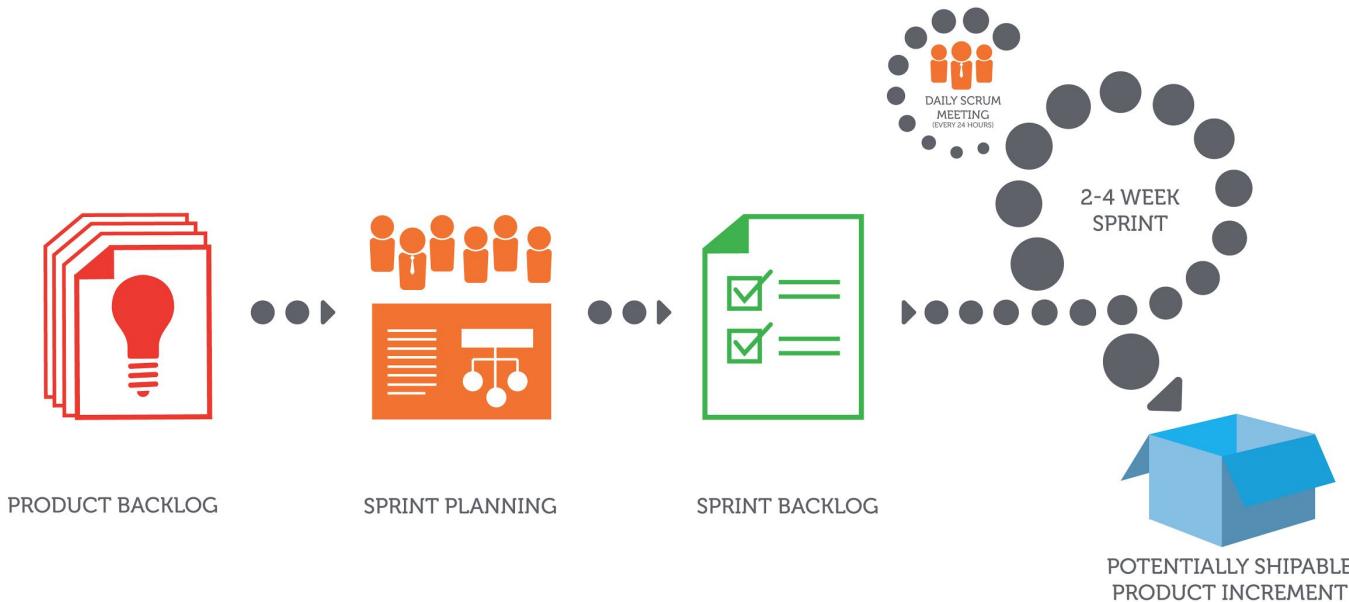
- Personalized Cancer Medicine
- Alcoholism in Native Americans
- Group Variant Calling
- Mutated tRNA in Zebrafish
- Kidney Transplant Exome GWAS
- Cow Antibody Sequencing
- Eating Disorder Haplotype Study
- Dog Genome Assembly



Cancer, how to beat it
(see blackboard for details)

What is it, you do here?

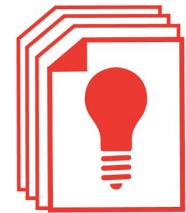
Software Engineer II, Software Development Engineer in Test



How scrum saves the world

What is it, you do here?

Software Engineer II, Software Development Engineer in Test



PRODUCT BACKLOG



SPRINT PLANNING



SPRINT BACKLOG



POTENTIALLY SHIPABLE
PRODUCT INCREMENT

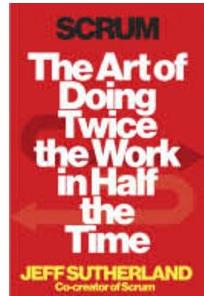
80%
of the value
comes from
20%
of the work
do that part first



Learn Scrum immediately

- Make a backlog
 - Organize it into stories
- What can you do in two weeks
 - Plan the two weeks
 - Debrief the two weeks
 - How can we make the next two weeks better?
- **One** task at a time
- Meet daily for 15 minutes
 - Same time, same place
- Work <40 hours per week

Scrum: The Art of Doing Twice the Work in Half the Time [Book]



[Shop now](#)

Sponsored

\$14.99 · [Google Play](#)

Free shipping, no tax

[View all sellers and prices](#)

[Product details](#)

Author: Jeff Sutherland

Publisher: Crown Publishing Group

Pages: 256

Format: ebook

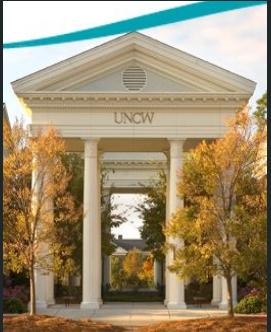
Publication Date: 2014

ISBN: 0385346468

[View more details](#)

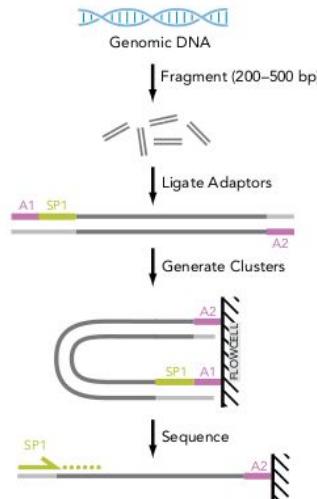


Acknowledgements



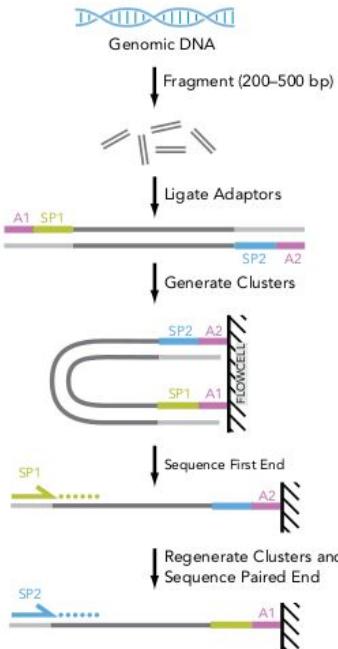
More illumina® Sequencing

- Single End
- Paired End
- Mate Pair



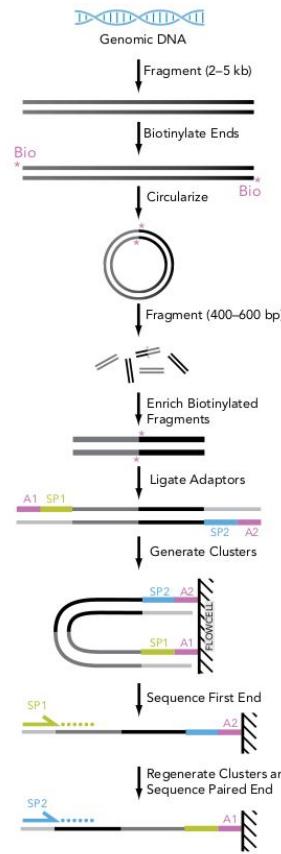
DNA to Data
 ~4 days
36 bp reads
inc. sample prep

Sample Prep
 3 hours hands-on



DNA to Data
 ~7 days
36x2 bp reads
inc. sample prep

Sample Prep
 3 hours hands-on



DNA to Data
 ~8 days
36x2 bp reads
inc. sample prep

Sample Prep
 4½ hours hands-on

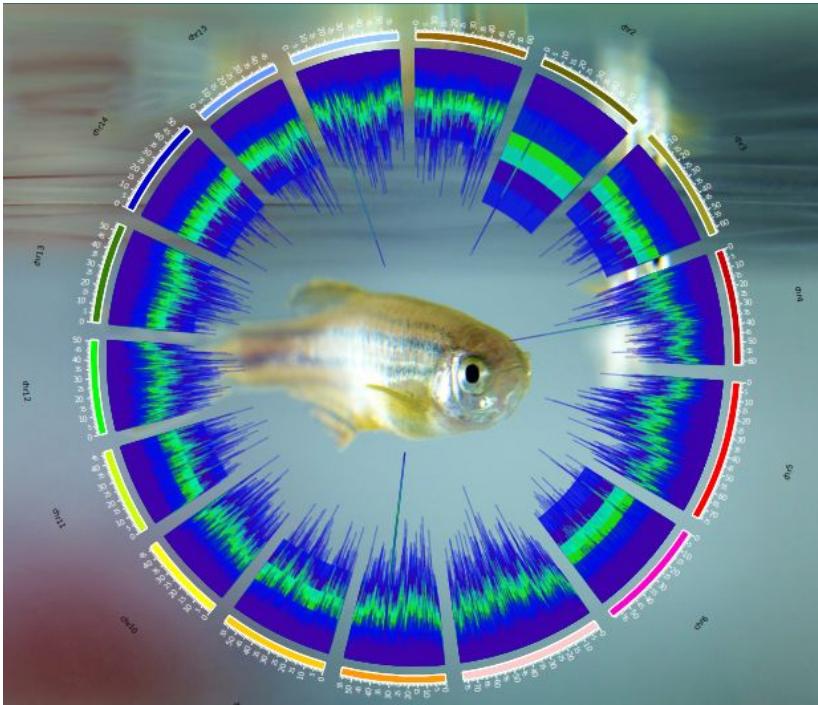
What is depth of coverage?

- Coverage - how much of the genome is covered
- Depth - how deep is a given base pair covered
- Depth of Coverage - state both..

Aligned reads

ACCGCGATTCAAGGTACCACG
GCGATTCAAGGTACCACCGC
GATTCAAGGTACCACCGCTA
TTCAGGTACCACCGTAGC
CAGGTACCACCGTAGCGC
GGTTACCACCGTAGCGCAT
TTACACCGTAGCGCATT
ACCACCGTAGCGCATTACA
CACCGTAGCGCATTACACA
CGCGTAGCGCATTACACAGA
CGTAGCGCATTACACAGATT
TAGCGCATTACACAGATTAG

Consensus contig ACGCGATTCAAGGTACCACCGTAGCGCATTACACAGATTAG



Tools to get moving - platform independent

- Illumina BaseSpace - candy store of tools, either it works or you can't do it
 - [Link](#) - feels like the app store for bioinformatics, beware older versions of tools
- Galaxy - many places including Globus
 - [Link](#) - nice for building pipelines, very appropriate for short term analyses
- Ugene - Mentioned last time, great all around resource
 - [Link](#) - platform independent, excellent youtube series ([link](#))
 - Now contains full suite for NGS analysis (not that I've tried it)
- MEGA - It was your homework for a reason
 - [Link](#) - great for phylogenetics or to just poke at sequences and alignments
- Secure Shell Chrome Extension
 - [Link](#) - Get to the command line from your browser, not great for file access, get an FTP

Tools and what they do

- Phylogenetics
 - MEGA, TOPALi, Genious et al, Ugene, Websites Galore
- Alignment tools
 - MEGA (really, go get it), Ugene, Clustal website, other websites
 - BWA, BowTie, STAR, Isaac,
- Quality Control
 - Qualimap, FastQC,
- Assembly
 - SOAP, Velvet, Abyss, Arachne
- Scripting is a language and a tool
 - R (+ggplot2), Perl, Python, Circos (is Perl or R), BedTools, etc
- Pipeline Handlers
 - Build your own in Bash/Python/Perl
 - Ptolemy or GridNexus
 - OmicsPipe (caugh)
 - BCBio
 - Galaxy or Globus
- For purchase....
 - GoldenHelix + VarSeq (\$\$ but great)

To *really* dig in..

- Upscale Text Editor
 - NotePad++ ([link](#)) excellent windows tool
 - Sublime ([link](#)) beloved Mac and Linux tool
- Ubuntu ([link](#)) or Linux Mint ([link](#)) or Bio-Linux ([link](#)) or Ubuntu Gnome ([link](#))
 - VirtualBox ([link](#)) will let you run linux inside of Windows, or really anything inside of anything
 - Parallels ([link](#)) will let you run linux (and windows) inside of MacOs
- SSH Secure Shell Client
 - [Link](#) - Might want this if you're in Windows, it's the command line and a file transfer in one
- Learn Unix and Python
 - Unix - underlying software of Linux and Mac, very powerful and dynamic ([link](#))
 - Python - popular scripting language, requires clean code to work ([link](#))
 - R - Similar to MatLab but Free and more widely used, get R-Studio ([link](#))