Using bioinformatics to advance precision medicine

High School Student Program 2016

Bioinformatics and Research Computing
Whitehead Institute

http://jura.wi.mit.edu/bio

Whitehead Institute

What is bioinformatics?

- The use of computers and software to
  - Store
  - **Analyze**
  - Integrate
  - **Interpret**

  biological information to learn about biology

What is precision medicine?

- "an innovative approach [to medicine] that takes into account individual differences in people's genes, environments, and lifestyles"

  https://www.whitehouse.gov/precision-medicine

- Many doctors are doing this already – but not typically using one's whole genome

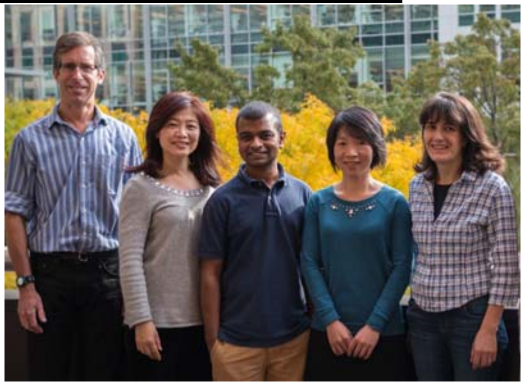
President Obama, January 30, 2015

Bioinformatics & Research Computing

Consultation and collaboration, training and education, and software in the areas of Bioinformatics and Graphics.

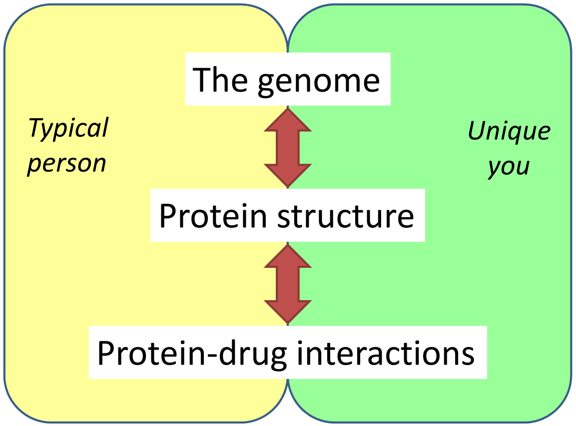
at Whitehead Institute

BaRC Bioinformatics and Research Computing

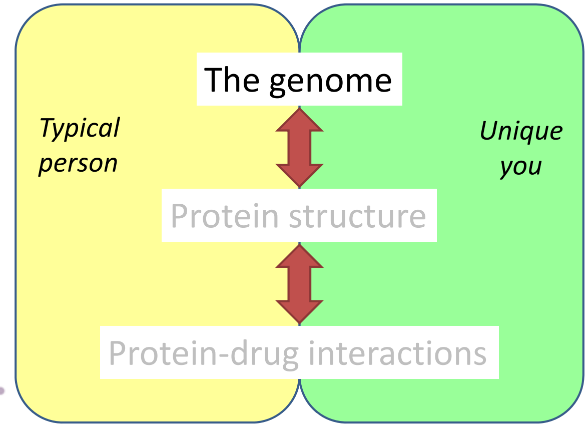


George Bell	Bingbing Yuan	Prat Thiru	Yanmei Huang	Inma Barrasa

Big challenges in precision medicine



Big challenges in precision medicine



What can we learn from one's genome?

- How does our genome differ from
  - The "reference" genome?
  - A typical genome from our "ethnic" background?
  - Our parents, siblings, and other family members?
- Are these differences due to
  - Single-letter changes ("single nucleotide variants")?
  - Insertions or deletions?
  - More or fewer copies of a repeated region?
  - [Rare] Extra or missing pieces of chromosomes?
- Is there anything "unexpected"?

Aside: the "cancer genome"

- Precision medicine can also help with cancer treatment
- Cancer is a collection of diseases, all involving different genome mutations
- To perform precision cancer medicine, one can sequence the genome of a tumor to help identify the best treatment
- This has its own set of challenges!
- We won't discuss this today.

Approaches to "genome" sequencing

- Sequence just the ~million most different locations
  - 23andMe, Ancestry.com, etc.
- Sequence just the genes (1-2% of the genome) – the "exome"
- Sequence the whole genome
  - As much of all chromosomes as possible

Sampling genome sequence

- Most (99.9%) of the genome is identical between individuals
- We want to concentrate only on the places that are the most different

Different locations in the genome

Different individuals

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
G	T	T	A	G	T	A	A	T	G	C	C	T	G	T	T	C	A	G	A
G	A	T	A	G	T	A	A	T	G	C	C	T	G	T	T	C	A	G	A
G	T	T	A	G	T	A	A	T	G	C	C	T	G	T	T	C	A	G	A
G	T	T	A	G	T	A	A	T	G	C	C	T	G	T	G	T	A	C	A
G	T	T	A	G	T	A	A	T	G	C	C	T	G	T	T	T	A	C	A
G	T	T	A	G	T	A	A	T	G	C	C	T	G	T	T	T	A	C	A
G	T	T	G	G	T	A	A	C	G	C	C	T	A	T	T	T	A	C	A
G	T	T	G	G	T	A	A	C	G	C	C	T	A	T	T	T	A	C	A
G	T	T	G	G	T	A	A	C	G	C	C	T	A	T	T	T	A	C	A
G	T	T	G	G	T	A	A	C	G	C	C	T	A	T	T	T	A	C	A
G	T	T	G	G	T	A	A	C	G	C	C	T	A	T	T	T	A	C	A

Single nucleotide polymorphisms

- SNPs (pronounced "snips") because
  - Single: were looking at just one genome position
  - Nucleotide: DNA letter differs
  - Polymorphism: variation occurring commonly in a population (in at least 1% of individuals)
- SNPs can be within a gene or between genes

4	9	14	17	19
A	T	G	C	G
A	T	G	C	G
A	T	G	C	G
A	T	G	T	C
A	T	G	T	C
A	T	G	T	C
G	C	A	T	C
G	C	A	T	C
G	C	A	T	C
G	C	A	T	C
G	C	A	T	C

But humans (like peas) are diploid

- We have 2 genomes, with 2 copies of each chromosome
- Each SNP can be
  - Homozygous (ex: CC), or
  - Heterozygous (ex: TG)

Different individuals

SNP 1	SNP 2	SNP 3	SNP 4	SNP 5
AA	TT	GG	CC	GG
AA	TT	GG	CC	GG
AA	TT	GG	CC	GG
AA	TT	GG	CT	GC
AA	TT	GG	CT	GC
AA	TT	GG	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC

Nearby SNPs are associated

- Nearby SNPs tend to stay together during meiosis
- As a result, they tend to be genetically linked
- One “tag SNP” can be used to represent a set of linked SNPs

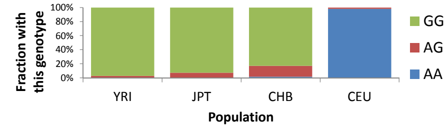
Genotype block “A” (linked SNPs) | Genotype block “B”

Different individuals

SNP 1	SNP 2	SNP 3	SNP 4	SNP 5
AA	TT	GG	CC	GG
AA	TT	GG	CC	GG
AA	TT	GG	CC	GG
AA	TT	GG	CT	GC
AA	TT	GG	CT	GC
AA	TT	GG	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC
AG	TC	GT	CT	GC

Taking ethnicity into account

- Genotypes have been collected from large-scale projects like
  - HapMap http://hapmap.ncbi.nlm.nih.gov
  - 1000 Genomes http://www.1000genomes.org
- These populations (“ethnic groups”) include
  - Yoruba in Ibadan, Nigeria (“YRI”)
  - Japanese in Tokyo, Japan (“JPT”)
  - Han Chinese in Beijing, China (“CHB”)
  - Utah residents with ancestry from northern and western Europe (“CEU”)
- Sample HapMap data for SNP rs1834640



One publicly available human genome

nature Vol 452 | 17 April 2008 | doi:10.1038/nature06884

LETTERS

The complete genome of an individual by parallel DNA sequencing

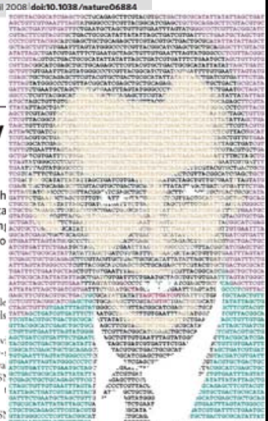
David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Ch
Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tarta
Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Son
Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinsto
& Jonathan M. Rothberg[2]†

The association of genetic variation with disease and drug response, and improvements in nucleic acid technologies, have given great optimism for the impact of 'genomic medicine'. However, the formidable size of the diploid human genome[1], approximately 6 gigabases, has prevented the routine application of sequencing methods to deciphering complete individual human genomes. To realize the full potential of genomics for human health, this limitation must be overcome. Here we report the DNA sequence of a diploid genome of a single individual, James D. Watson, sequenced to 7.4-fold redundancy in two months using massively

subject's DNA, including single nucle small insertions and deletions (indels (CNV).

The 454 base-calling software provi for each base. We developed a three- patterns of error and associated Q va software to improve the accuracy of S lion variant positions were filtered (Table 1).

Comparison of these putative S



James Watson decoded.

Precision medicine on Dr. Watson

- Concentrating on his genome sequence,
  - What can we learn about
    - Potential genetic risk for disease?
    - Expected drug response?
    - Optimal disease treatment?
- Big challenges:
  - Even though we’re only 0.1% different, with 3 billion DNA letters, it adds up to a lot
  - Which differences have something to do with our health?

Computational challenges

1. Align each piece of our genome sequence to the reference genome

   Example sequenced DNA piece:

   GACCCCGGCTGCGGCGAGGAGGAAGGAGCCAGCCTAGCAGCTTCTGCGCCTGTGGCCGCGGGTGTCCTGGAGGCCTCTCGGTGTGACGAGTGGGGGACCC
2. Repeat this process for 100 million DNA pieces
3. Identify the DNA letters that are different from the "reference genome"

Next step (the easier one)

- Compare our common DNA variants to those in lots of Genome-Wide Association Studies
- For example,
  - If rs17822931 = TT => Dry earwax; less body odor
  - If rs4988235 = GG => Lactose intolerance
  - If rs1801282 = CC => Increased diabetes 2 risk
  - If rs1799971 = AG or GG => higher odds of heroin and alcohol addiction
- (How) should one react to a finding like this?

Next step (the harder one)

- Predict the effect of our rare DNA variants
- Since these are rare, they may be less well (or not at all) studied
- An exception can be a rare variant but one that alone causes a bad disease.
  - Sickle-cell anemia
  - Tay-Sach's disease
  - Muscular dystrophy

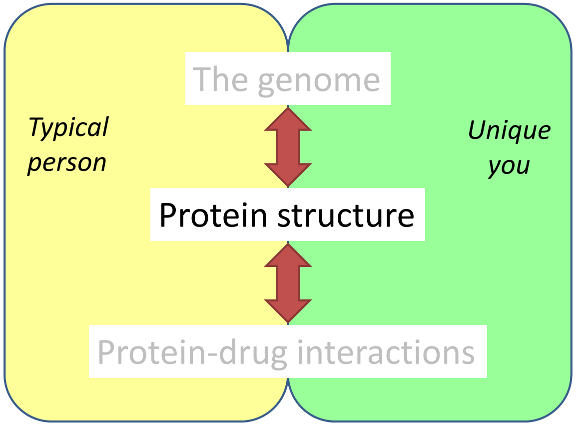
Exercise 1

Genome variant analysis

Dr. Watson's variants (summary)

- Total = 3.3 million
  - 2.7 million are common
  - 600k are rare
- 10,500 result in changes in protein sequence
  - 9000 are common
  - 1500 are rare
  - 7% of the total were predicted to be "probably damaging" to protein function

Big challenges in precision medicine

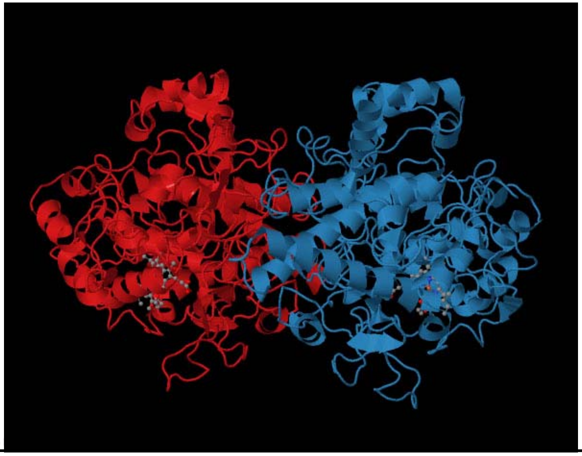


Prostaglandin synthase

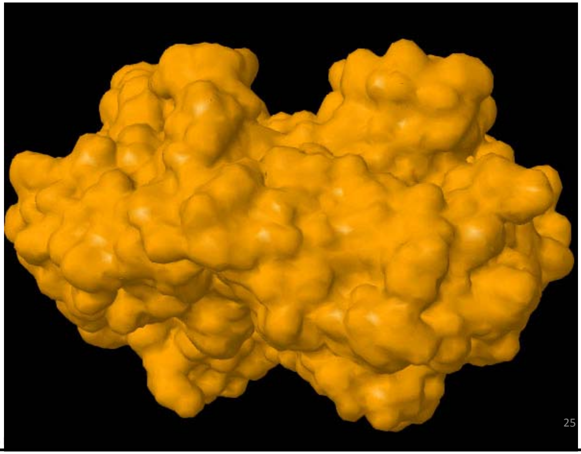
Primary sequence provides some information

```
MSRSLLLWFLLFLLLLPPLPVLLADPGAPTPVNPCCYYPCQHQGICV
RFGLDRYQCDCTRTGYSGPNCTIPGLWTWLRNSLRPSPSFTHFLLTH
GRWFWEFVNATFIREMLMRLVLTVRSNLIPSPPTYNSAHDYISWESF
SNVSYYTRILPSVPKDCPTPMGTKGKKQLPDAQLLARRFLLRRKFIP
DPQGTNLMFAFFAQHFTHQFFKTSGKMGPGFTKALGHGVDLGHIYGD
NLERQYQLRLFKDGKLKYQVLDGEMYPPSVEEAPVLMHYPRGIPPQS
QMAVGQEVFGLLPGLMLYATLWLREHNRVCDLLKAEHPTWGDEQLFQ
TTRLILIGETIKIVIEEYVQQLSGYFLQLKFDPELLFGVQFQYRNRI
AMEFNHLYHWHPLMPDSFKVGSQEYSYEQFLFNTSMLVDYGVEALVD
AFSRQIAGRIGGGRNMDHHILHVAVDVIRESREMRLQPFNEYRKRFG
MKPYTSFQELVGEKEMAAELEELYGDIDALEFYPGLLLEKCHPNSIF
GESMIEIGAPFSLKGLLGNPICSPEYWKPSTFGGEVGFNIVKTATLK
KLVCLNTKTCPYVSFRVPDASQDDGPAVERPSTEL
```

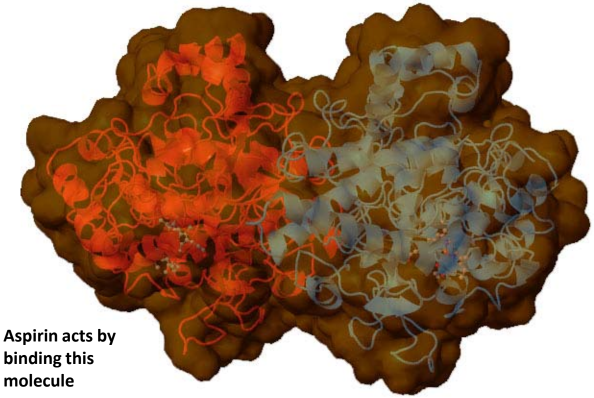

Prostaglandin synthase



Prostaglandin synthase



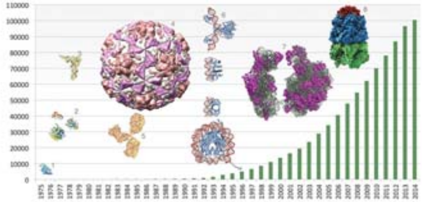
Prostaglandin synthase



Aspirin acts by binding this molecule

Generating protein structures is an art

- The Protein Data Bank holds >100k structures
- Some other structures can be predicted from sequence similarity
- Other proteins have completely unknown structure



number of structures available in the PDB

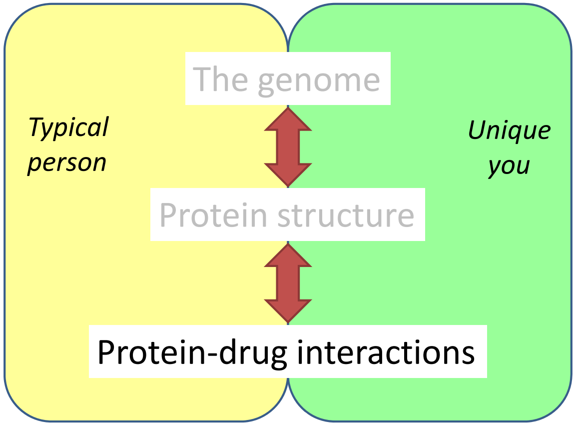
What does a mutation do?

- We can predict the effect best if we know the
  - Protein's function(s)
  - Protein's structure
  - Amino acid chemistry change
  - Active site(s) for interaction(s) with other molecules
    - Proteins
    - Metabolites
    - Drugs

Exercise 2

Protein structure analysis

Summary



Effects of variation on protein structure and function

- Changes can occur in typical protein function
- Changes may be apparent only during drug treatment
  - Typical drug may no longer bind
  - Other drug may now bind
  - Different balance between different forms of protein (inactive vs active)
- This is a major area of research

Summary

- Bioinformatics has contributed to many advances in precision medicine
- All areas of precision medicine need even more insights from computational methods
- Lots of challenges ahead for biologists, computer scientists, and doctors!