

# Computer Science and Genetics

Michael Schatz, Ph.D.

Jan 8, 2013  
CSH High School

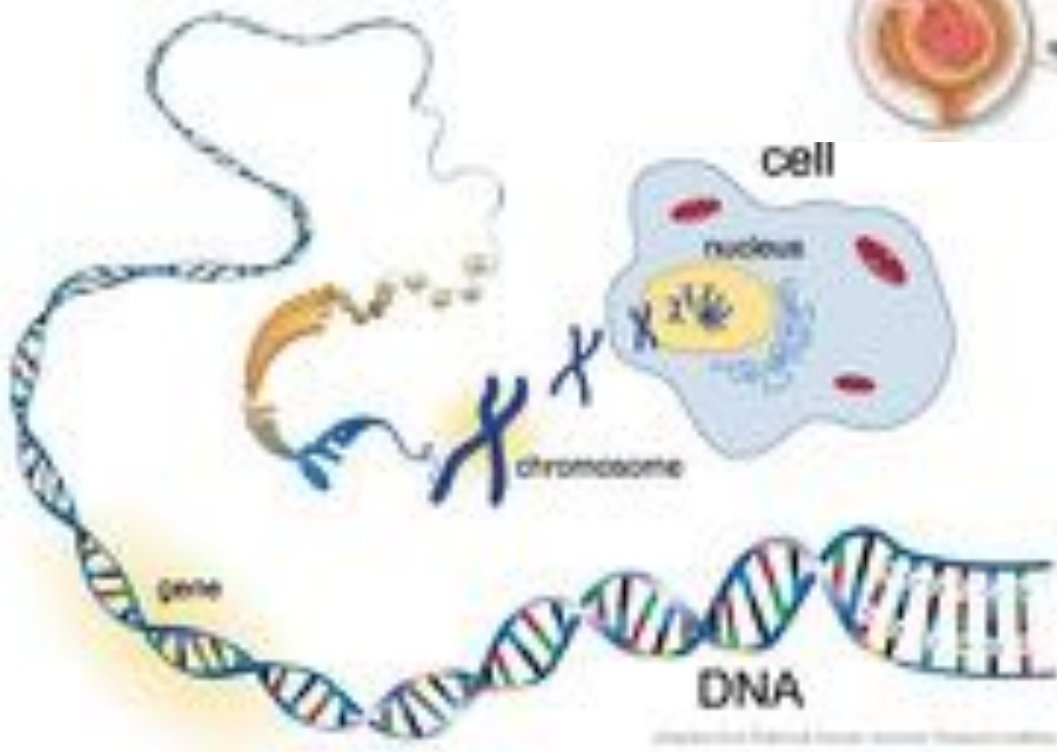
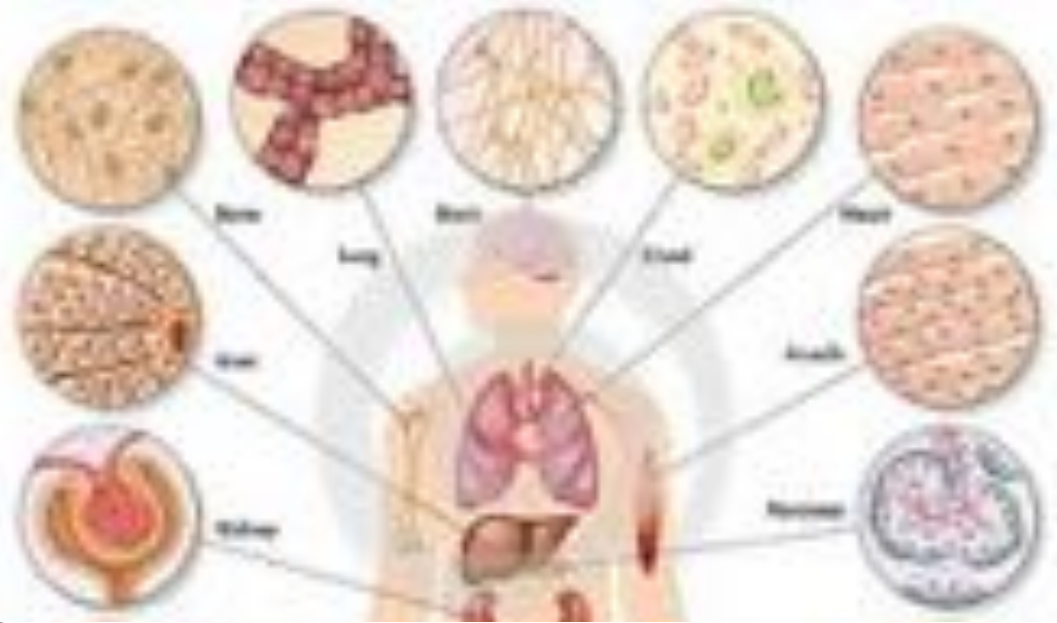




Why do you look like your parents?  
How is that information stored, transmitted, and executed?

# Cells & DNA

Your body is made  
of ~100 trillion  
different cells of  
~300 different types



As you zoom into each  
cell, you'll find each  
contains an exact copy  
of a special molecule  
called DNA

# Structure of DNA



The double helix structure makes two important properties possible:

**Base-pairing:** A always pairs with T, C always pairs with G. Therefore, a single strand of the molecule can be used as a template to make copies

**Genetic code:** Any sequence of nucleotides can be “spelled out” along the double helix. The cell can recognize those patterns as use it as a “recipe” for building cells and organizing your body.

Your genome is a 2x3B nucleotides long  
in 23 pairs of chromosomes



# Genotype to Phenotype



The particular sequence of your genome (along with your environment and experiences) shapes who you are:

- Height
- Hair, eye, skin color
- Amount of body hair
- Broad/narrow, small/large nose
- Acne prone or clear complexion
- Susceptible to disease
- Response to drug treatments

Physical traits tend to be genetic, social characteristics tend to be environmental, and everything else is a combination

# DNA Sequencing



**Illumina HiSeq 2000**

>60Gbp / day



**One human genome**

~20 DVDs / genome

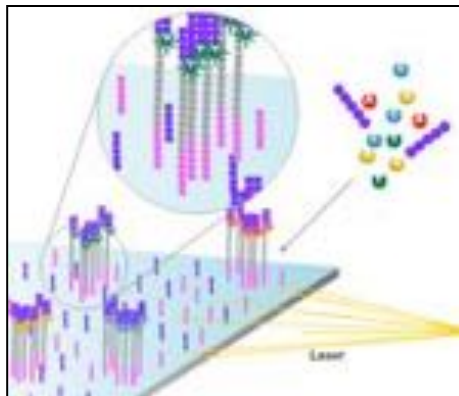
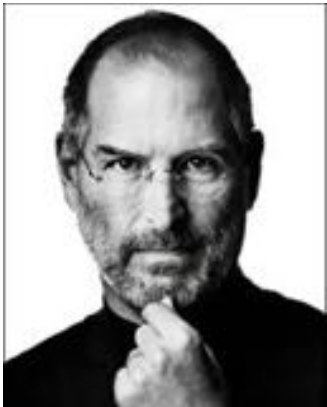
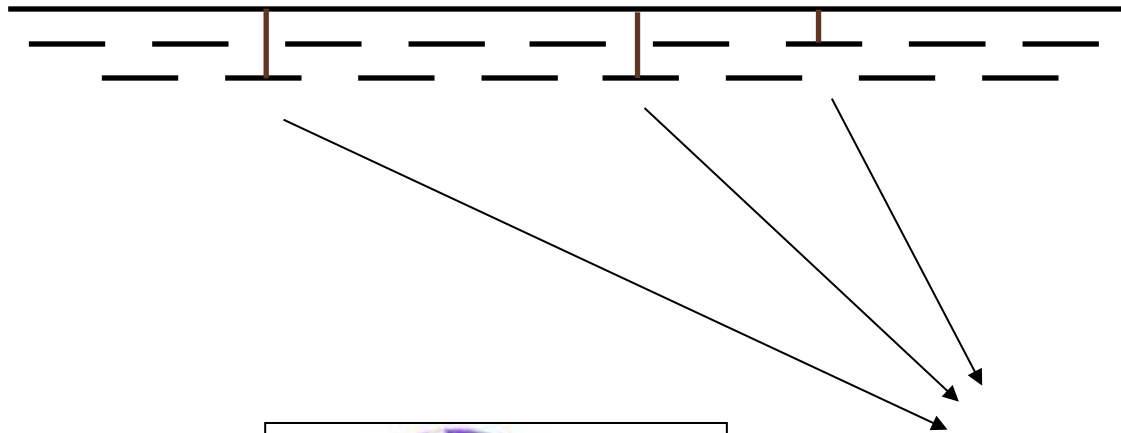


**World Wide Capacity**

>2 miles tall

# Personal Genomics

How does your genome compare to the reference?



Heart Disease  
Cancer  
Creates magical  
technology

# Searching for GATTACA

- Where is GATTACA in the human genome?
- Strategy I: Brute Force

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	...
G	A	T	T	A	C	A									

No match at offset 1



# Searching for GATTACA

- Where is GATTACA in the human genome?
- Strategy 1: Brute Force

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	...
	G	A	T	T	A	C	A								

Match at offset 2

# Searching for GATTACA

- Where is GATTACA in the human genome?
- Strategy I: Brute Force

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	...
		G	A	T	T	A	C	A	...						

No match at offset 3...

# Searching for GATTACA

- Where is GATTACA in the human genome?
- Strategy I: Brute Force

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	...
								G	A	T	T	A	C	A	

No match at offset 9 <- Checking each possible position takes time

# Brute Force Analysis



- Brute Force:
  - At every possible offset in the genome:
    - Do all of the characters of the query match?
- Analysis
  - Simple, easy to understand
  - Genome length =  $n$  [3B]
  - Query length =  $m$  [7]
  - Comparisons:  $(n-m+1) * m$  [21B]
- Overall runtime:  $O(nm)$ 
  - [How long would it take if we double the genome size, read length?]
  - [How long would it take if we double both?]

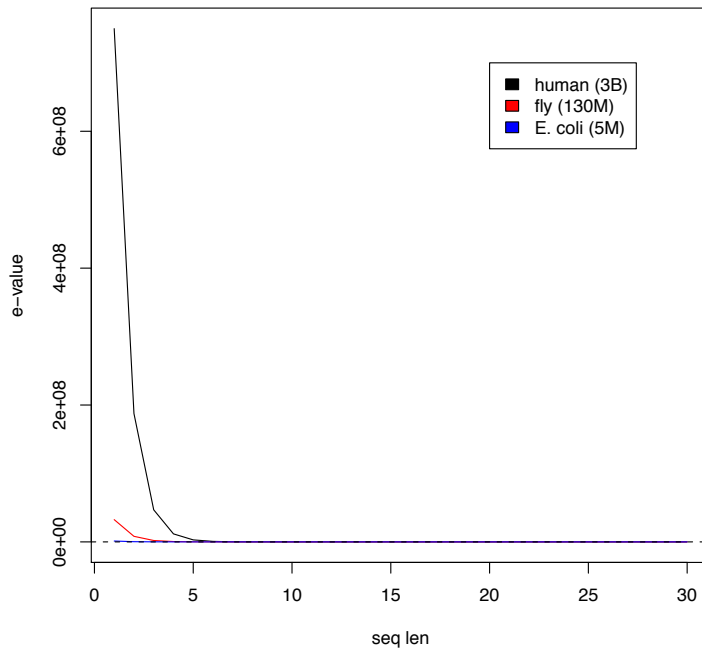
# Expected Occurrences

The expected number of occurrences (e-value) of a given sequence in a genome depends on the length of the genome and inversely on the length of the sequence

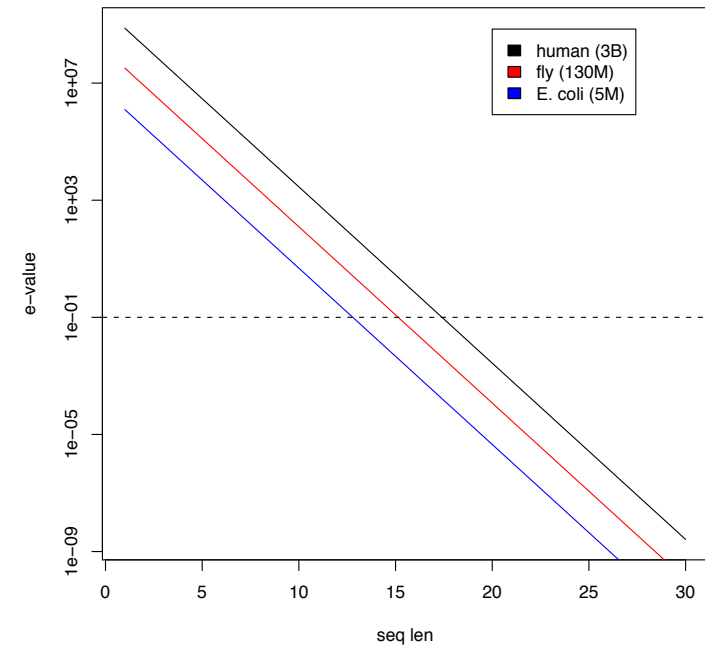
- 1 in 4 bases are G, 1 in 16 positions are GA, 1 in 64 positions are GAT, ...
- 1 in 16,384 should be GATTACA
- $E = n / (4^m)$

[183,105 expected occurrences]  
[How long do the reads need to be for a significant match?]

Value and sequence length  
cutoff 0.1



E-value and sequence length  
cutoff 0.1





# Brute Force Reflections

Why check every position?

- GATTACA can't possibly start at position 15

[WHY?]

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	...
								G	A	T	T	A	C	A	

- Improve runtime to  $O(n + m)$

[3B + 7]

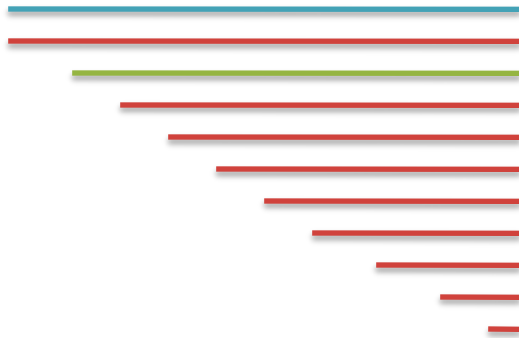
- If we double both, it just takes twice as long
- Knuth-Morris-Pratt, 1977
- Boyer-Moyer, 1977, 1991

- For one-off scans, this is the best we can do (optimal performance)

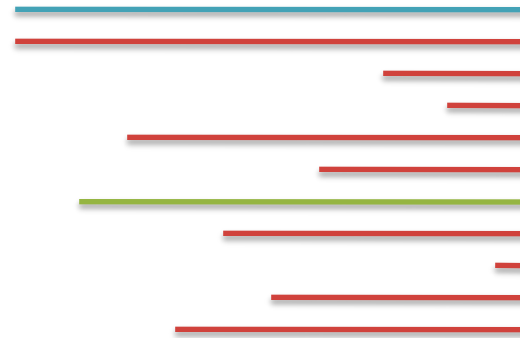
- We have to read every character of the genome, and every character of the query
- For short queries, runtime is dominated by the length of the genome

# Suffix Arrays: Searching the Dictionary

- What if we need to check many queries?
  - We don't need to check every page of the dictionary to find 'DNA'
  - Sorting alphabetically lets us immediately skip 96% (25/26) of the book *without any loss in accuracy*
- Sorting the genome: Suffix Array (Manber & Myers, 1991)
  - Sort every suffix of the genome



Split into n suffixes



Sort suffixes alphabetically

[Challenge Question: How else could we split the genome?]

# Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower
- Searching for GATTACA
  - Lo = 1; Hi = 15;

Lo  
→

#	Sequence	Pos
1	ACAGATTACC...	6
2	ACC...	13
3	AGATTACC...	8
4	ATTACAGATTACC...	3
5	ATTACC...	10
6	C...	15
7	CAGATTACC...	7
8	CC...	14
9	GATTACAGATTACC...	2
10	GATTACC...	9
11	TACAGATTACC...	5
12	TACC...	12
13	TGATTACAGATTACC...	1
14	TTACAGATTACC...	4
15	TTACC...	11

Hi  
→

# Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower
- Searching for GATTACA
  - $Lo = 1; Hi = 15; Mid = (1+15)/2 = 8$
  - $Middle = Suffix[8] = CC$

Lo  
→

#	Sequence	Pos
1	ACAGATTACC...	6
2	ACC...	13
3	AGATTACC...	8
4	ATTACAGATTACC...	3
5	ATTACC...	10
6	C...	15
7	CAGATTACC...	7
8	CC...	14
9	GATTACAGATTACC...	2
10	GATTACC...	9
11	TACAGATTACC...	5
12	TACC...	12
13	TGATTACAGATTACC...	1
14	TTACAGATTACC...	4
15	TTACC...	11

Hi  
→

# Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower
- Searching for GATTACA
  - $Lo = 1; Hi = 15; Mid = (1+15)/2 = 8$
  - Middle = Suffix[8] = CC  
=> Higher:  $Lo = Mid + 1$

Lo  
→

#	Sequence	Pos
1	ACAGATTACC...	6
2	ACC...	13
3	AGATTACC...	8
4	ATTACAGATTACC...	3
5	ATTACC...	10
6	C...	15
7	CAGATTACC...	7
8	CC...	14
9	GATTACAGATTACC...	2
10	GATTACC...	9
11	TACAGATTACC...	5
12	TACC...	12
13	TGATTACAGATTACC...	1
14	TTACAGATTACC...	4
15	TTACC...	11

Hi  
→



# Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower
- Searching for GATTACA
  - $Lo = 1; Hi = 15; Mid = (1+15)/2 = 8$
  - $Middle = Suffix[8] = CC$   
=> Higher:  $Lo = Mid + 1$
  - $Lo = 9; Hi = 15;$

#	Sequence	Pos
1	ACAGATTACC...	6
2	ACC...	13
3	AGATTACC...	8
4	ATTACAGATTACC...	3
5	ATTACC...	10
6	C...	15
7	CAGATTACC...	7
8	CC...	14
9	GATTACAGATTACC...	2
10	GATTACC...	9
11	TACAGATTACC...	5
12	TACC...	12
13	TGATTACAGATTACC...	1
14	TTACAGATTACC...	4
15	TTACC...	11

Lo  
→

Hi  
→

# Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower
- Searching for GATTACA
  - $Lo = 1; Hi = 15; Mid = (1+15)/2 = 8$
  - $Middle = Suffix[8] = CC$   
=> Higher:  $Lo = Mid + 1$
  - $Lo = 9; Hi = 15; Mid = (9+15)/2 = 12$
  - $Middle = Suffix[12] = TACC$

#	Sequence	Pos
1	ACAGATTACC...	6
2	ACC...	13
3	AGATTACC...	8
4	ATTACAGATTACC...	3
5	ATTACC...	10
6	C...	15
7	CAGATTACC...	7
8	CC...	14
9	GATTACAGATTACC...	2
10	GATTACC...	9
11	TACAGATTACC...	5
12	TACC...	12
13	TGATTACAGATTACC...	1
14	TTACAGATTACC...	4
15	TTACC...	11

Lo  
→

Hi  
→

# Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower
- Searching for GATTACA
  - $Lo = 1; Hi = 15; Mid = (1+15)/2 = 8$
  - $Middle = Suffix[8] = CC$   
=> Higher:  $Lo = Mid + 1$
  - $Lo = 9; Hi = 15; Mid = (9+15)/2 = 12$
  - $Middle = Suffix[12] = TACC$   
=> Lower:  $Hi = Mid - 1$
  - $Lo = 9; Hi = 11;$

#	Sequence	Pos
1	ACAGATTACC...	6
2	ACC...	13
3	AGATTACC...	8
4	ATTACAGATTACC...	3
5	ATTACC...	10
6	C...	15
7	CAGATTACC...	7
8	CC...	14
9	GATTACAGATTACC...	2
10	GATTACC...	9
11	TACAGATTACC...	5
12	TACC...	12
13	TGATTACAGATTACC...	1
14	TTACAGATTACC...	4
15	TTACC...	11

Lo  
→

Hi  
→

# Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower
- Searching for GATTACA
  - $Lo = 1; Hi = 15; Mid = (1+15)/2 = 8$
  - $Middle = Suffix[8] = CC$   
=> Higher:  $Lo = Mid + 1$
  - $Lo = 9; Hi = 15; Mid = (9+15)/2 = 12$
  - $Middle = Suffix[12] = TACC$   
=> Lower:  $Hi = Mid - 1$
  - $Lo = 9; Hi = 11; Mid = (9+11)/2 = 10$
  - $Middle = Suffix[10] = GATTACC$

#	Sequence	Pos
1	ACAGATTACC...	6
2	ACC...	13
3	AGATTACC...	8
4	ATTACAGATTACC...	3
5	ATTACC...	10
6	C...	15
7	CAGATTACC...	7
8	CC...	14
9	GATTACAGATTACC...	2
10	GATTACC...	9
11	TACAGATTACC...	5
12	TACC...	12
13	TGATTACAGATTACC...	1
14	TTACAGATTACC...	4
15	TTACC...	11

Lo  
→

Hi  
→

# Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower
- Searching for GATTACA
  - $Lo = 1; Hi = 15; Mid = (1+15)/2 = 8$
  - $Middle = Suffix[8] = CC$   
=> Higher:  $Lo = Mid + 1$
  - $Lo = 9; Hi = 15; Mid = (9+15)/2 = 12$
  - $Middle = Suffix[12] = TACC$   
=> Lower:  $Hi = Mid - 1$
  - $Lo = 9; Hi = 11; Mid = (9+11)/2 = 10$
  - $Middle = Suffix[10] = GATTACC$   
=> Lower:  $Hi = Mid - 1$
  - $Lo = 9; Hi = 9;$

#	Sequence	Pos
1	ACAGATTACC...	6
2	ACC...	13
3	AGATTACC...	8
4	ATTACAGATTACC...	3
5	ATTACC...	10
6	C...	15
7	CAGATTACC...	7
8	CC...	14
9	GATTACAGATTACC...	2
10	GATTACC...	9
11	TACAGATTACC...	5
12	TACC...	12
13	TGATTACAGATTACC...	1
14	TTACAGATTACC...	4
15	TTACC...	11

Lo  
Hi  
→



# Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower
- Searching for GATTACA
  - $Lo = 1; Hi = 15; Mid = (1+15)/2 = 8$
  - $Middle = Suffix[8] = CC$   
=> Higher:  $Lo = Mid + 1$
  - $Lo = 9; Hi = 15; Mid = (9+15)/2 = 12$
  - $Middle = Suffix[12] = TACC$   
=> Lower:  $Hi = Mid - 1$
  - $Lo = 9; Hi = 11; Mid = (9+11)/2 = 10$
  - $Middle = Suffix[10] = GATTACC$   
=> Lower:  $Hi = Mid - 1$
  - $Lo = 9; Hi = 9; Mid = (9+9)/2 = 9$
  - $Middle = Suffix[9] = GATTACA...$   
=> Match at position 2!

#	Sequence	Pos
1	ACAGATTACC...	6
2	ACC...	13
3	AGATTACC...	8
4	ATTACAGATTACC...	3
5	ATTACC...	10
6	C...	15
7	CAGATTACC...	7
8	CC...	14
9	GATTACAGATTACC...	2
10	GATTACC...	9
11	TACAGATTACC...	5
12	TACC...	12
13	TGATTACAGATTACC...	1
14	TTACAGATTACC...	4
15	TTACC...	11

# Binary Search Analysis

- Binary Search

Initialize search range to entire list

$\text{mid} = (\text{hi} + \text{lo}) / 2$ ;  $\text{middle} = \text{suffix}[\text{mid}]$

if query matches middle: done

else if query < middle: pick low range

else if query > middle: pick hi range

Repeat until done or empty range

[WHEN?]

- Analysis

- More complicated method

- How many times do we repeat?

- How many times can it cut the range in half?

- Find smallest  $x$  such that:  $n / (2^x) \leq 1$ ;  $x = \lg_2(n)$

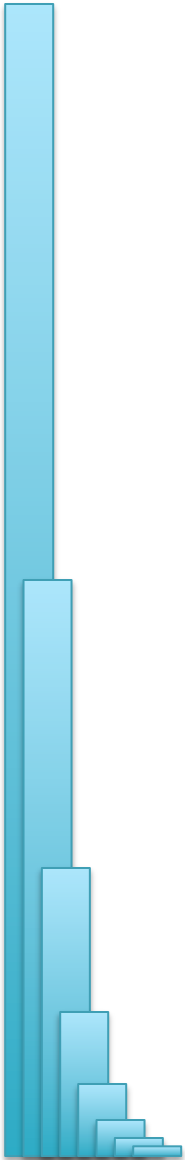
[32]

- Total Runtime:  $O(m \lg n)$

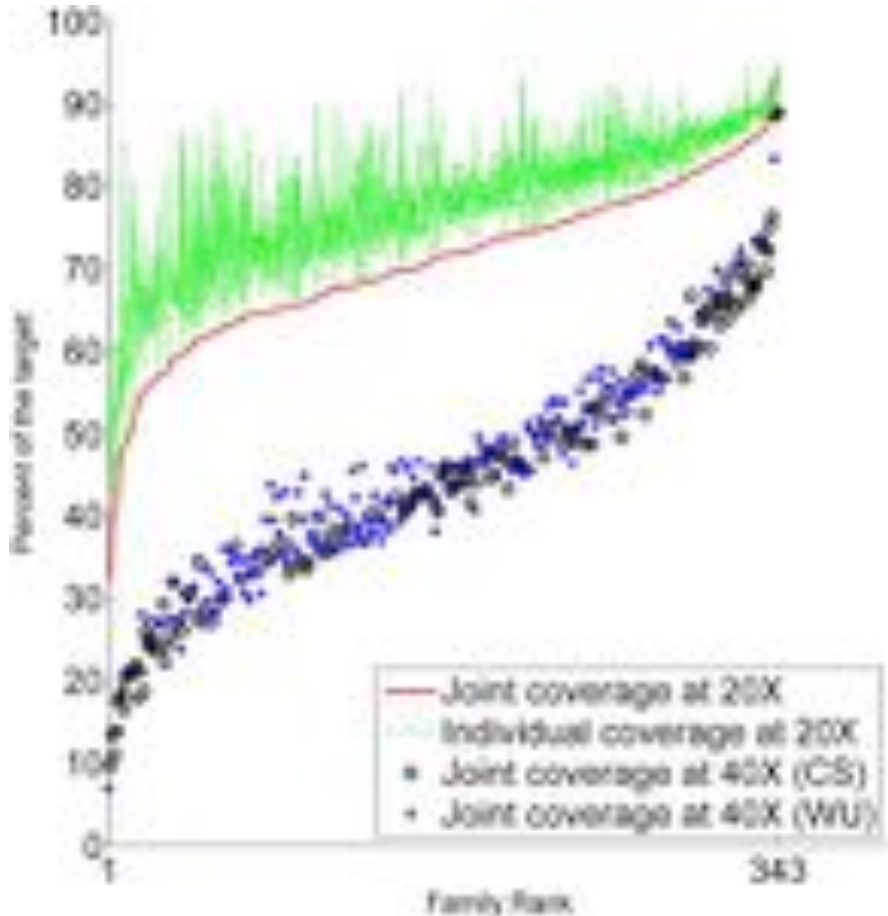
- More complicated, but **much** faster!

- Looking up a query loops 32 times instead of 3B

[How long does it take to search 6B or 24B nucleotides?]



# Genetics of Autism



Sequencing of 343 families from the Simons Simplex Collection

- Parents plus one child with autism and one non-autistic sibling
- Enriched for higher-functioning individuals

Families prepared and captured together to minimize batch effects

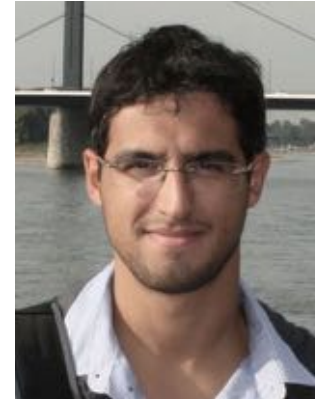
- Exome-capture performed with NimbleGen SeqCap EZ Exome v2.0 targeting 36 Mb of the genome.
- ~80% of the target at >20x coverage with ~93bp reads

**De novo gene disruptions in children on the autism spectrum**

Iossifov et al. (2012) *Neuron*. 74:2 285-299

# Scalpel: Haplotype Microassembly

G. Narzisi, D. Levy, I. Iossifov, J. Kendall, M. Wigler, M. Schatz



**Micro-assembly** pipeline for accurate detection and validation of *de novo* mutations (SNPs and indels)

```
Ref:      ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCCGGA...
Father1:  ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCCGGA...
Father2:  ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCCGGA...
Mother1:  ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCCGGA...
Mother2:  ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCCGGA...
Sib1:     ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCCGGA...
Sib2:     ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCCGGA...
Aut1:     ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCCGGA...
Aut2:     ...TCAGAACAGCTGGATGAGATCTTACC-----CCGGGAGATTGTCTTTGCCCCGGA...
```

6bp heterozygous deletion at chr13:25280526 ATP12A

# De novo mutations in Autism

- In 343 families analyzed so far, we see significant enrichment in de novo **likely gene killers** in the autistic kids
  - Overall rate basically 1:1 (432:396)
  - 2:1 enrichment in nonsense mutations
  - 2:1 enrichment in frameshift indels
  - 4:1 enrichment in splice-site mutations
  - Most de novo originate in the paternal line in an age-dependent manner (56:18 of the mutations that we could determine)
- Observe strong overlap with the 842 genes known to be associated with fragile X protein FMRP
  - Related to neuron development and synaptic plasticity
  - Suggests avenues for early interventions and possible treatments

**De novo gene disruptions in children on the autism spectrum**

Iossifov et al. (2012) *Neuron*. 74:2 285-299



# Unsolved Questions in Biology

There is tremendous interest to sequence:

- What is your genome sequence?
- How does your genome compare to my genome?
- Where are the genes and how active are they?
- How does gene activity change during development?
- How does splicing change during development?
- How does methylation change during development?
- How does chromatin change during development?
- How does is your genome folded in the cell?
- Where do proteins bind and regulate genes?
- What virus and microbes are living inside you?
- How do your mutations relate to disease?
- W
- ..

Answering these questions requires  
specialized software & quantitative analysis



# Challenges of Modern Science



**The foundations of science will continue to be *observation, experimentation, and interpretation***

- Technology will continue to push the frontier
- Measurements will be made *digitally* over large populations, at extremely high resolution, and for diverse applications

## ***Rise in Quantitative and Computational Demands***

1. *Experimental design*: selection, collection & metadata
2. *Observation*: measurement, storage, transfer, computation
3. *Integration*: multiple samples, assays, analyses
4. *Discovery*: visualizing, interpreting, modeling

***Ultimately limited by the human capacity to execute extremely complex experiments and interpret results***

# Acknowledgements

## Schatz Lab

Aspyn Palatnick  
Giuseppe Narzisi  
Shoshana Marcus  
Hayan Lee  
Rob Aboukhalil  
Mitch Bekritsky  
Charles Underwood  
James Gurtowski

Rushil Gupta  
Avijit Gupta  
Shishir Horane  
Deepak Nettem  
Varrun Ramani  
Piyush Kansal  
Alejandro Wences  
Eric Biggers

## CSHL

Hannon Lab  
Iossifov Lab  
Levy Lab  
Lippman Lab  
Lyon Lab  
Martienssen Lab  
McCombie Lab  
Ware Lab  
Wigler Lab



# Thank You!

<http://schatzlab.cshl.edu/>  
[@mike\\_schatz](#)

