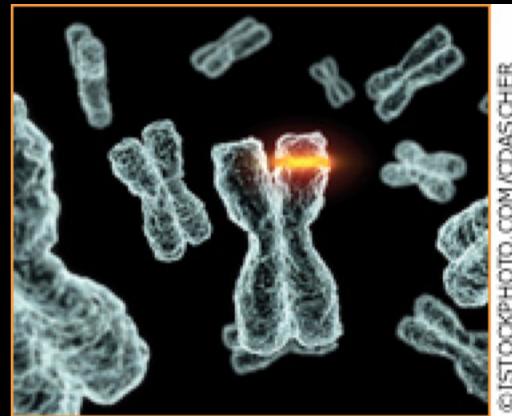


COMP90016

Computational Genomics:

Single Nucleotide Variants

SNPs



Overview

1. What are SNPs?
2. How were they discovered?
3. How are they discovered?
4. A SNP discovery project.

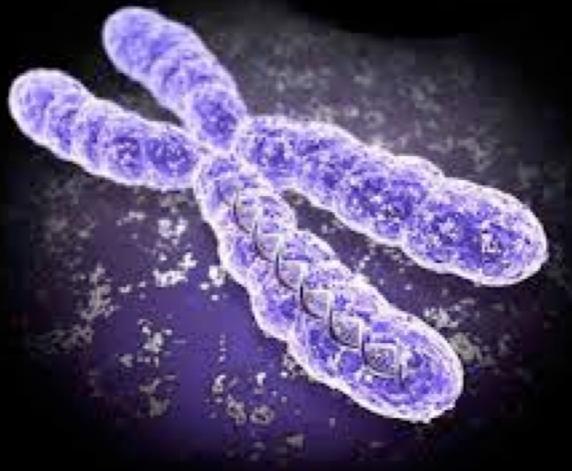


Part 1

WHAT ARE SNPs?

Alleles

- Chromosome pairs
- Two copies of the same chromosome: one from the mother, one from the father
- Similar, but not identical
- In bioinformatics we represent this as **two strings**, or **one string + a set of differences**.



X	Y	X	X	Y
X	Y	X	Y	X
X	Y	X	Y	X
X	Y	X	Y	X

SNPs: some facts and definitions

- SNP == *Single Nucleotide Polymorphism*
- SNPs are differences at single positions between two DNA molecules
 - Molecule 1: AAAAA
 - Molecule 2: AACAA
- In humans:

	heterozygous	homozygous
– Mother:	AAAAAA AAA AGAAA	↖
– Father:	AAAAACAAAAAAGAAA	↖
– Population:	AAAAAA AAA AAAAAA	
- There are **Major** and **minor** alleles in a population.

Important: The double stranded DNA is one molecule!

SNPs: more facts

- Due to random **mutation** and heredity of genomic material SNPs **propagate** through populations.
- SNPs can be **beneficial**, **detrimental**, or **inconsequential**.
- SNPs occur **everywhere** in the genome
- Of particular interest in **genes** and **regulating regions**
- Change of binding sites or amino acids can be significant to **disease**

Examples 1 and 2

- How can a SNP be **detrimental** to a phenotype?

- Consider an important gene
 - Assume the sequence

ACAGGAAGC -> TGS

Non-synonymous

ACAT**T**GAAGC -> T **STOP** S

- How can a SNP be “**inconsequential**”?

- As before:

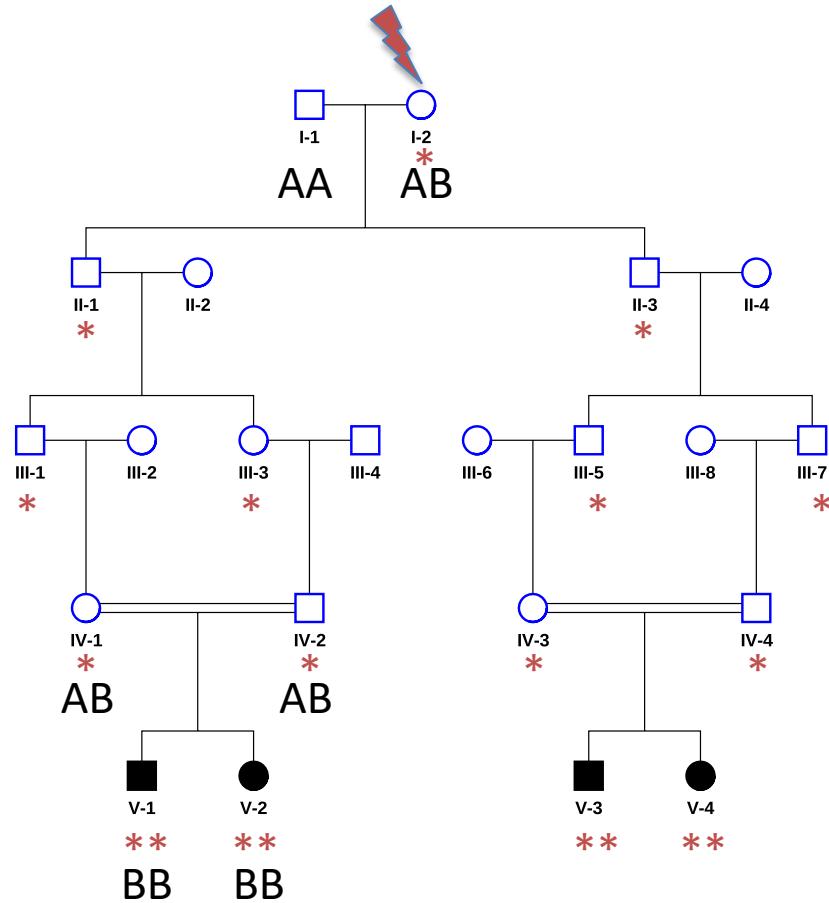
ACAGGAAGC -> TGS

Synonymous

ACT**G**GAAGC -> TGS

Example 3

How do SNPs propagate through a population?



Yet more Definitions...

- There are also **SNVs**:
 - *Single Nucleotide Variations*
- These are a superset of SNPs.
 - SNPs are only those that are actually found in a population at a certain frequency (typically $\geq 5\%$).
- In this lecture we are talking about SNPs only (but sometimes the category may be debatable).
- At a specific site we describe DNA by its **genotype** (AA, AC, AG, ..., TT in a diploid genome).
Confusingly, it is common to describe SNPs as A and B in a population setting (A for the major allele, B for the minor).
- Populations tend to develop **linkage**: 2 or more neighbouring SNPs being linked by the same **haplotype**:
 - “If there is a minor allele at position X, there is bound to be a minor allele at position Y as well”

The Computational Challenge

- There are two computational problems around SNPs:
 1. Discover new variants in a sample. What are the single-base differences between a sample and another/reference? This problem is known as variant calling.
 2. Verify the absence/existence/genotype of known variants within a sample. From prior studies we know about millions of sites in the genome that vary between individuals. Is a sample heterozygous at such a site, or homozygous for minor/major allele? This problem is known as genotyping.

Types of Variants

- There are two kinds of variants that we distinguish:
 1. Germline variants. These are the differences between our alleles and other individuals. These are the variants we are born with .
 2. Somatic variants. These variants are acquired during our lifetime in the DNA of single cells (and may proliferate from there). Such variants can turn our healthy cells into tumour cells.
- Somatic variants are investigated via variant calling. Germline variants may already be known in the population and therefore be validated with genotyping.



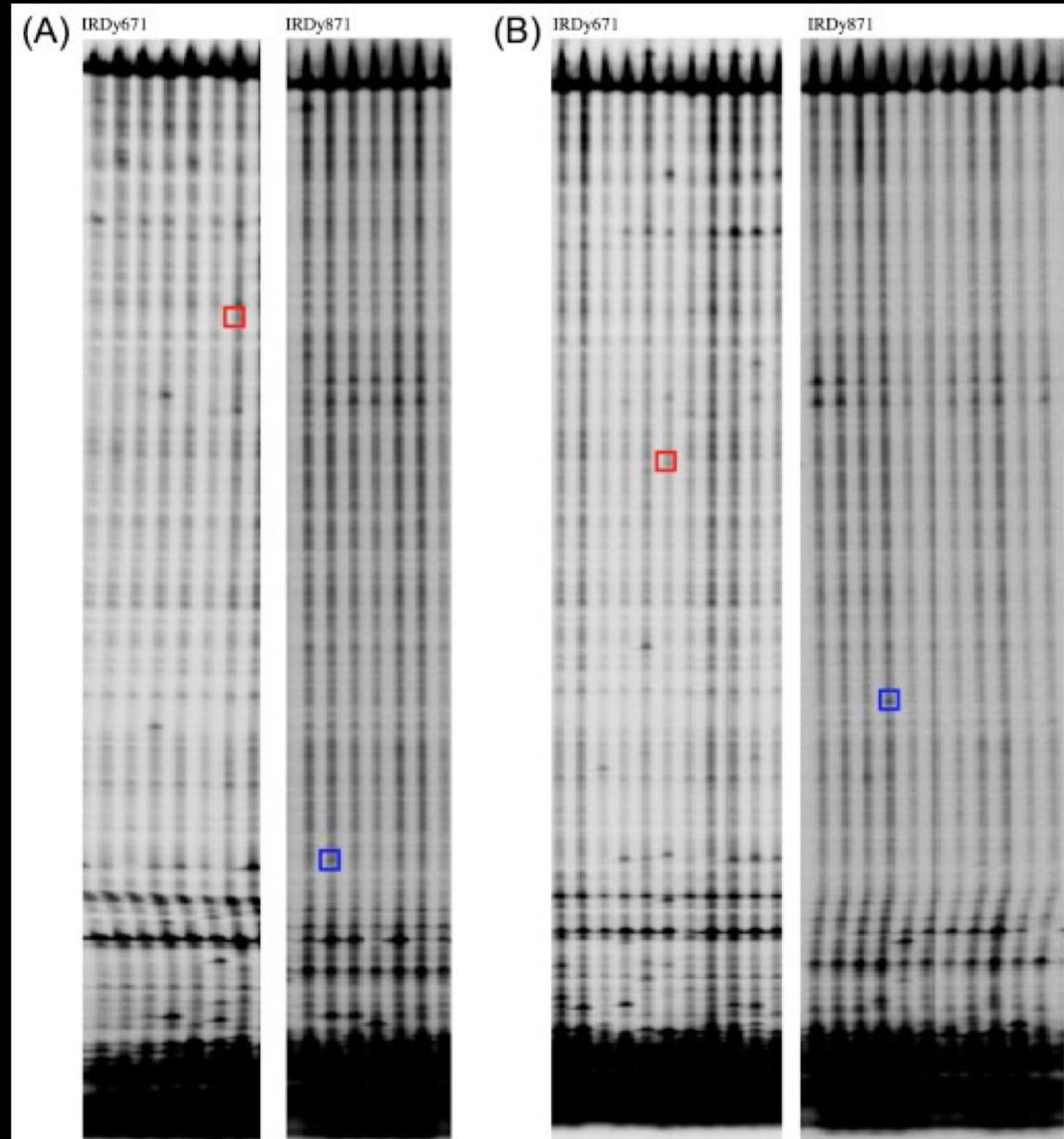
Part 2

HOW TO DISCOVER SNPs – A HISTORY

Traditional SNP Discovery (Variant Calling)

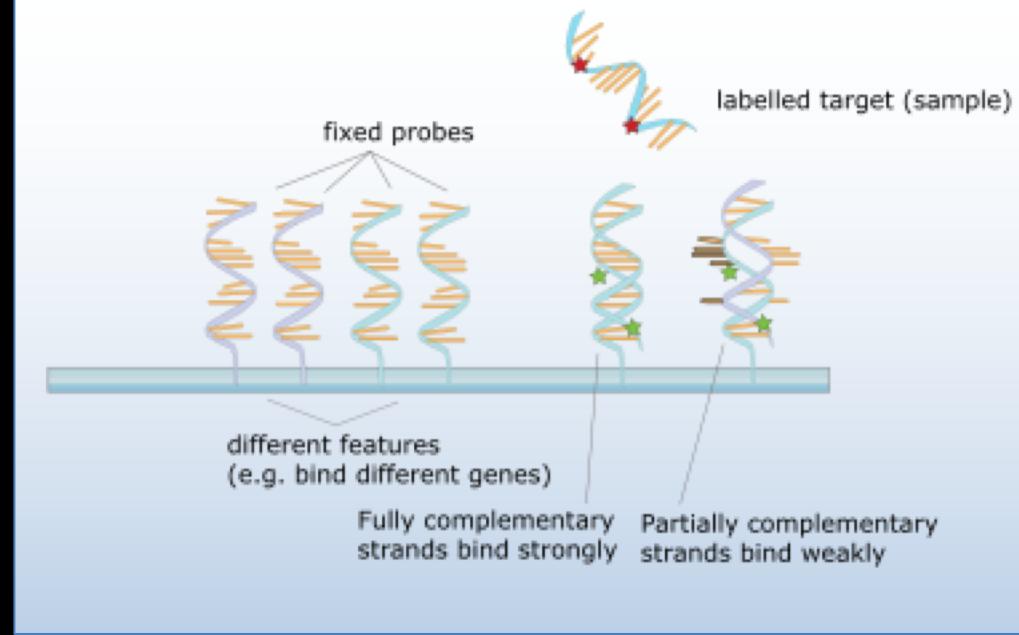
- Sanger sequencing of genes -> discovery of slight differences
- A more manual analysis of genomic differences.
- Cataloguing of SNPs led to public data bases and technologies to evaluate SNPs in individuals

SNPs on gel



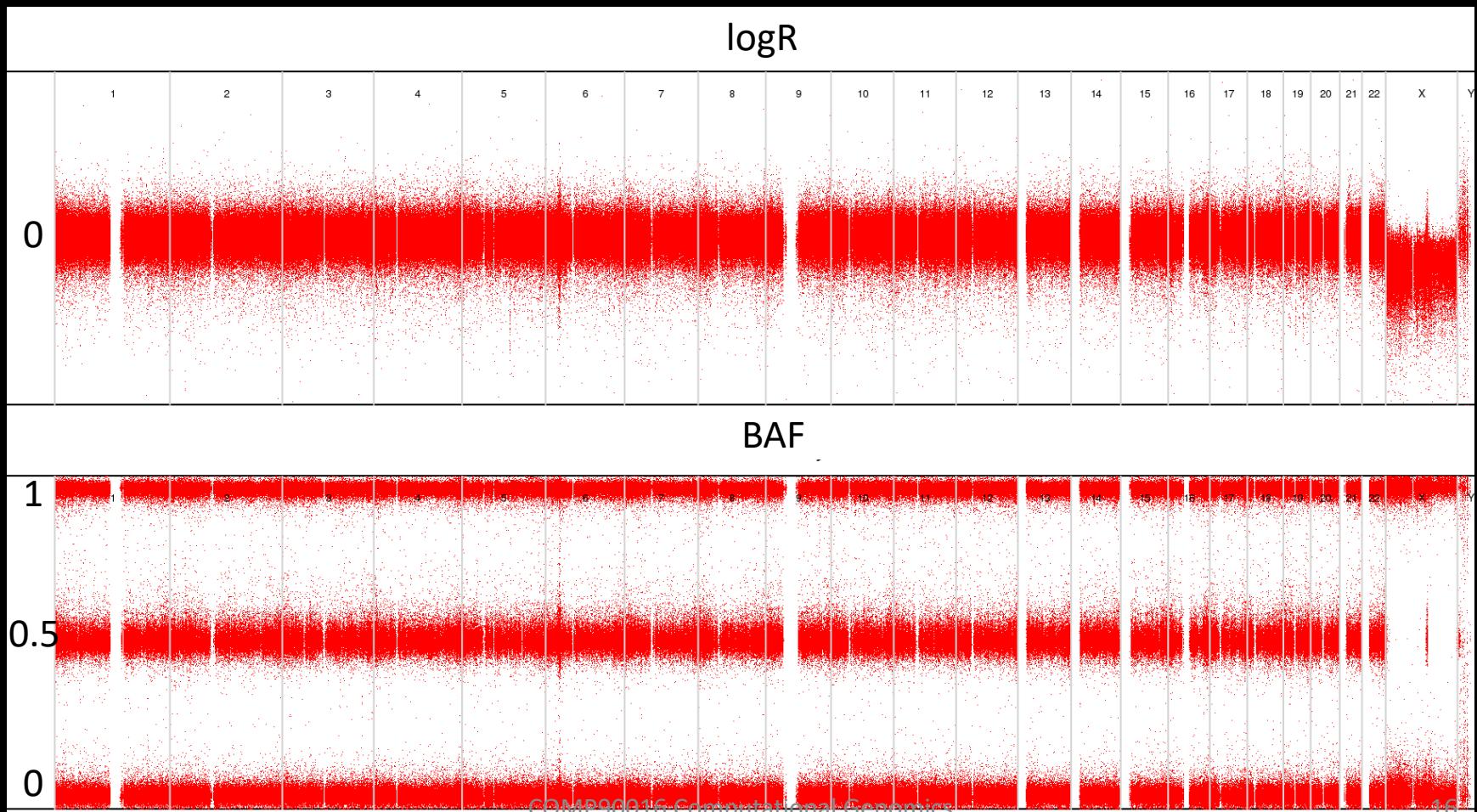
Traditional SNP Evaluation (Genotyping)

- Micro arrays
- SNP-arrays:
 - Probes == different alleles (genotypes)
 - A probe pair has the major and minor allele for a known SNP.
 - By measuring fluorescence the amount of binding to each probe pair can be established.
 - In Bioinformatics, a SNP-array delivers $2n$ intensity values for n probe pairs.



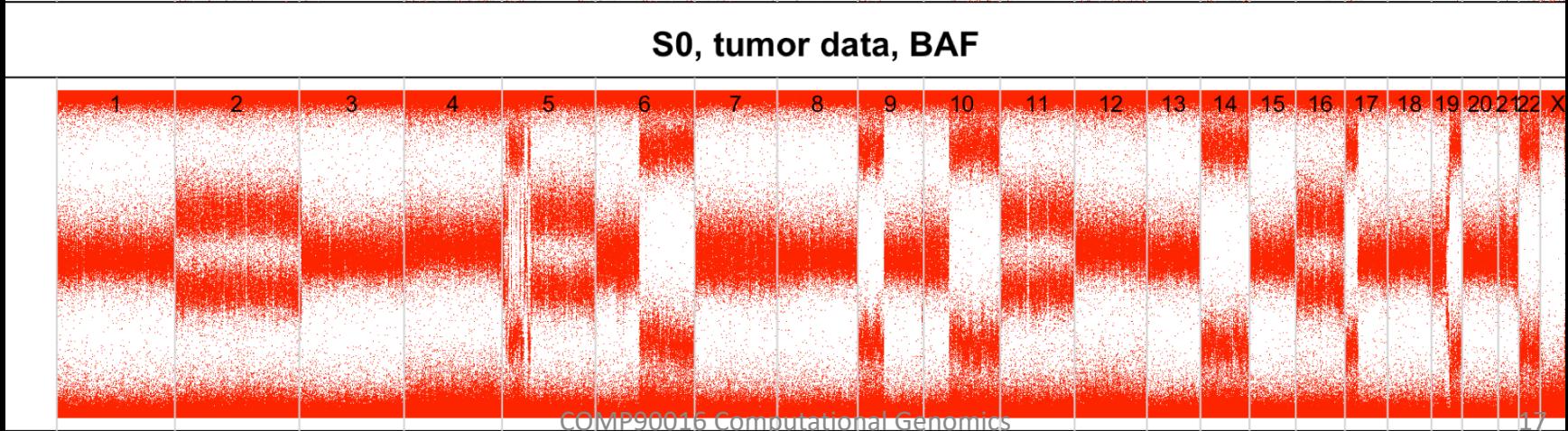
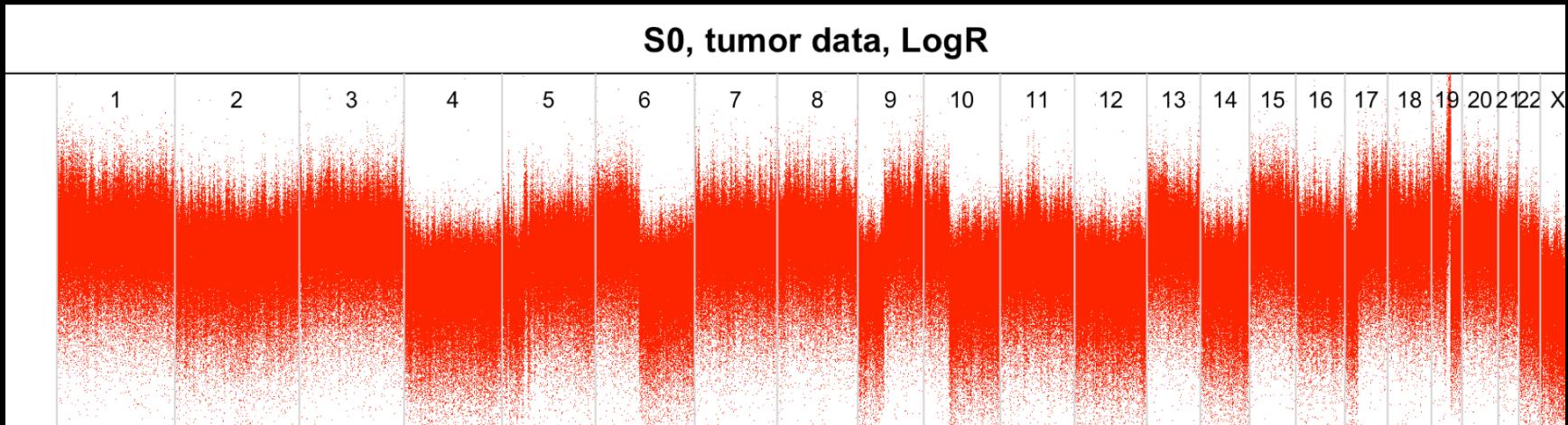
SNP Array Data

Let the array intensities be represented by I_{kj} with k an integer in $\{1..n\}$ and j in $\{1,2\}$. Define $\text{logR} := \log_2((I_{k1} + I_{k2}) / \text{mean } I)$, and $\text{BAF} := I_{k1} / (I_{k1} + I_{k2})$



SNP Array Data in Cancer

Let the array intensities be represented by I_{kj} with k an integer in $\{1..n\}$ and j in $\{1,2\}$. Define $\text{logR} := \log_2((I_{k1} + I_{k2}) / \text{mean } I)$, and $\text{BAF} := I_{k1} / (I_{k1} + I_{k2})$



Knowledge about SNPs

- dbSNP (<https://www.ncbi.nlm.nih.gov/SNP/>)
- “As of build 131 (available February 2010), dbSNP had amassed over 184 million submissions representing more than 64 million distinct variants for 55 organisms, including *Homo sapiens*, *Mus musculus*, *Oryza sativa*, and many other species” *
- “As of 23 July 2013, dbSNP listed 62,676,337 SNPs in humans.” *

* Wikipedia

Summary

- SNP: chr-a, position b, genotype c, allele frequency d
- CGAGTACGGGCTGCAGGCATACT[A/G]AAGTGAAACTGTGAGTGTGGGACC
Chromosome: 12:112241766
Gene:ALDH2
Functional Consequence: missense
Clinical significance: drug-response
Validated: by 1000G,by cluster,by frequency,by hapmap
Global MAF: A=0.0574/124
- Discovery dates back to the 1950s, but new insights have accelerated over the decades.



Part 3

CURRENT STATE OF SNP CALLING

Whole Genome Shotgun Sequencing

genome



cut many times at
random



forward-reverse paired
reads from a single
fragment



100bp

100bp

Align reads to reference
to investigate for SNPs

SNPs and Sequencing



Samtools Mpileup

TGGT1_chrIV	1053973 N	25	ttttTtttttTTTttTttTTttt	FFFJ5JIJJIGFI@GJCJEEDFAED
TGGT1_chrIV	1053974 N	25	ggggGggggGGGggGggGGgg	FFFJ5JIJJAHJ7HJCJDDEFBDD
TGGT1_chrIV	1053975 N	24	ttttttttTTTttTttTTttt	@CFJJGIIJ=HF=GHCGEEGHDED
TGGT1_chrIV	1053976 N	26	aaaaAaaaaaGAAAaAaaAAAa^Ka	@CFI5JHJJJ;IH<CICIDDIHFCCC
TGGT1_chrIV	1053977 N	26	t\$t\$ttTttttTTTttTttTTttt	@CCI;JEIJ18II=IGEJEIEIG>C@c
TGGT1_chrIV	1053978 N	24	aaAaaaaAAaaaAaAaaAAAAaa	CH5JCIIJCII<JIDJDD>HFC>D
TGGT1_chrIV	1053979 N	24	g\$gGggggGGGggGggGGgg	@G>J1JJJJGJJBJJCDDIHHCDC
TGGT1_chrIV	1053980 N	23	aAaaaaAAaaaAaTaAaAaaa	H;JJJJIEJJ=IICIEECJHEEE
TGGT1_chrIV	1053981 N	24	gGggggGGGggGggGGGggg^MG	H;J1JJJAJJ<JJCIDDCJGDDD@
TGGT1_chrIV	1053982 N	24	aAaaaaAAaaaAaAaaAAAAaa	G5IIIJJGJJ<IJ@JDDDJIDDDC
TGGT1_chrIV	1053983 N	25	gGggggGGGggGggGGGgggG^MG	F8IIIIJ@IJHIJ>IDDGJJCD@@
TGGT1_chrIV	1053984 N	25	aAaaaaAAaaaAaAaaAAAAaa	F?J1JJJEHJCJJ@JDDHJGCDCF@
TGGT1_chrIV	1053985 N	25	tTttttTTTttTttTTtttTT	F:J1JJJHGJDJJ<JDDIJIDBDDD
TGGT1_chrIV	1053986 N	25	gGggggGGGggGggGGGgggGG	DCJHJJJCJJ;JJBJDDIJBDDDF
TGGT1_chrIV	1053987 N	25	cCCCCCCCccCcCccCCCCccc	DCJEJJJEJA;JJ?DDBIJFDBDFF
TGGT1_chrIV	1053988 N	25	cCCCCCCCccCcCccCCCCccc	@DJFJIJGJG?IJ?EDDIJHC<DFF
TGGT1_chrIV	1053989 N	25	tTttttTTTttTttTTtttTT	BCJCJGJGJI3HGCHDDIJHDDDHF
TGGT1_chrIV	1053990 N	26	c\$CCCCCCCccCcCccCCCCccc^MC	@C1JJJJHJHA1ICHDDIHGDBDH@
TGGT1_chrIV	1053991 N	25	tTttttTTttTttTTtttTTTT	JGJIIHHHHF?HDDIJGDBDGH@1
TGGT1_chrIV	1053992 N	25	ggggGGggGggGGggGGGG	JHJJJHHHFJ1IEFDIJCDCDH@:
TGGT1_chrIV	1053993 N	24	tTttttTTttTgtttTTtttTT	JDIJJJEHC?JIEFEDIHBACCHHB
TGGT1_chrIV	1053994 N	25	ggggGGggGggGGggGGAggaGGGG	JIJJJJEHE8JGEFDBIIB?DDJHD=
TGGT1_chrIV	1053995 N	24	tTtttATTttTtCttTTtttTTT	JHJJJEH=3JFCFEDEIBCCCHHB
TGGT1_chrIV	1053996 N	24	ggggGGggGggGGggGGGG	H1JJJ@HBDJHFFCD@JADDIGD
TGGT1_chrIV	1053997 N	25	aaaaAAaaaAaAaaAAAAaaAAA	H1JIJBFF@JIFCFACJFCDCII?=
TGGT1_chrIV	1053998 N	25	ggggGGggGggGGggGGGG	H1JJJ3FFFJ1HDFDGIHBDCIJFA
TGGT1_chrIV	1053999 N	24	ggggGGggGggGGggGGGG	H1JJJJFFAIJGDEDGJCDDCJHJD
TGGT1_chrIV	1054000 N	25	ccccCCCCCcCcCCCCccc	HFHJI5FDGJJGDDEIJHDC>IJBA
TGGT1_chrIV	1054001 N	26	aaaaAAagAaAaaAAAAAA^MA	FDHJJ=DDGJJHDEEIJIDCDJIHA@
TGGT1_chrIV	1054002 N	27	ggggGGggGggGGggGGGGG^Mg	FGHJJADDCJJHDHEGJICDDJJHF@A
TGGT1_chrIV	1054003 N	28	ggggGGggGggGGggGGGGg^KG	FHHJJ>DD3IEIDFEBJIDDDJJGH@@J
TGGT1_chrIV	1054004 N	28	tTttttTTttTttCTtttTTCTTt	FHHJJ5ABFG@IAHC8J?DCCBCE<BAE
TGGT1_chrIV	1054005 N	28	ggggGGggGggGGggGGGGgg	FFFJJ;C@?JDICHF?J?DCDFGAED:H
TGGT1_chrIV	1054006 N	32	aaaaAAaaaTaAaaAAAAAA^MA^MA^Ka^Ja	CFFHJ@DCGJEGCJFGJ?DCDHFEFDICDD
TGGT1_chrIV	1054007 N	32	tTttttTTatTtTttTTtttTTTTt	CFFHJCDDFGEIIFIJ:DDCIHEAFCCDD
TGGT1_chrIV	1054008 N	32	g\$ggggGGGggGggGGggGGGGgg	CFFHGCDD<IHHDGFGBDDGIB:FAJCCDD
TGGT1_chrIV	1054009 N	31	ccccCTCcCcCcCCCCccc	DFHCCD<IHDDIF@JFDDDIJC?HCJFFFDD
TGGT1_chrIV	1054010 N	30	tTttTTttTttTTttTttTTt	CCHHCDD<IFGIHDJDDDDJJ?H@JFFDD
TGGT1_chrIV	1054011 N	30	ggggGGggGggGGggGGGGgg	CCFH:DFFJFJIJFCIHCDCJIC@G@JFFDD
TGGT1_chrIV	1054012 N	30	t\$t\$ttTTttTttTTttTtt	@CFH@CCFJFGJH9HBDDCHGC:H@IFFCE
TGGT1_chrIV	1054013 N	28	tTTTttTtTtTtTtTtTtTt	FCDCDFEIJFJPEBDDDJGGF9HAJFFDE

... Samtools Mpileup

TGGT1_chrIV	1053973 T	25	FFFJ5JIJJIGFI@GJCJEEDFAED
TGGT1_chrIV	1053974 G	25	FFFJ5JIJJJAHH7HJCJDDEFBDD
TGGT1_chrIV	1053975 T	24	@CFJJGIIJ=HF=GHCGEEGHDED
TGGT1_chrIV	1053976 A	26G.....^K,	@CFI5JHJJJ;IH<CICIDDIHFCCC
TGGT1_chrIV	1053977 T	26	,\$,\$.....	@CCI;JEIJ18II=IGEJEIEG>C@C
TGGT1_chrIV	1053978 A	24	CH5JCIIJCII<JIDJDD>HFC>D
TGGT1_chrIV	1053979 G	24	,\$.....	@G>J1JJJJGJJBJJCDDIHHCDC
TGGT1_chrIV	1053980 A	23T,T.....	H;JJJJIEJJ=IICIEECJHEEE
TGGT1_chrIV	1053981 G	24	H;J1JJJAJJ<JJCIDDCJGDDDD@
TGGT1_chrIV	1053982 A	24	G5IIIJJGJJ<IJ@JDDDJIDDDC
TGGT1_chrIV	1053983 G	24	F8IIIIJJ@IJHIJ>IDDGJJJCDD@
TGGT1_chrIV	1053984 A	25	F?J1JJJEHJCJJ@JDDHJGCDCF@
TGGT1_chrIV	1053985 T	25	F:J1JJJHGJDJJ<JDDIJIDBDDDD
TGGT1_chrIV	1053986 G	25	DCJHJJJCJJ;JJBJDDIJBDDDDF
TGGT1_chrIV	1053987 C	25t.....	DCJEJJJEJA;JJ?DDBIJFDBDFF
TGGT1_chrIV	1053988 C	25a.....	@DJFJIJGJG?IJ?EDDIJHC<DFF
TGGT1_chrIV	1053989 T	25	BCJCJGJGJI3HGCHDDIJHDDDHF
TGGT1_chrIV	1053990 C	26	,\$.....^M.	@CIJJJJHJHAITCHDDIHGDBDH=
TGGT1_chrIV	1053991 T	25	JGJIIHHHHF?HDDIJGDBDGH>1
TGGT1_chrIV	1053992 G	25	JHJJJHHHFJIEFEDIJGCDDHH>:
TGGT1_chrIV	1053993 C	24	tttttTTTttTgTttTTTtttTTT	JDIJJJEHC?JIEFEDIHBACCHH>
TGGT1_chrIV	1053994 G	25A,,a....	JIJJJJEHE8JGEFDIBIIB?DDJHD=
TGGT1_chrIV	1053995 T	24A.,,C,.....	JHJJJEH=3JFCFEDEIBCCCHHB
TGGT1_chrIV	1053996 G	24	HIJJJ@HBDJHFFCD@JADDIGD
TGGT1_chrIV	1053997 A	25	HIJJIJBFF@JIFCFACJFCDCII?=
TGGT1_chrIV	1053998 G	25	HIJJJ3FFFJTHDFDGIHBDCIJFA
TGGT1_chrIV	1053999 G	24	HIJJJJFFAIJGDEDGJCDDCJJHD
TGGT1_chrIV	1054000 C	25	HFHJI5FDGJJGDEIJHDC>IJBA
TGGT1_chrIV	1054001 A	26g.....^M.	FDHJJ=DDGJJHDEEIJIDCDJIA@
TGGT1_chrIV	1054002 G	27^M,	FGHJJADDCJJHDHEGJICDDJJHF@A
TGGT1_chrIV	1054003 G	27a.....	FHHJJ>DD3IEIDFEBJIDDDJJGH@€
TGGT1_chrIV	1054004 T	28C....C...	FHHJJ5ABFG@IAHC8J?DCDCBCE<BA;
TGGT1_chrIV	1054005 G	28	FFFJJ;C@?JDICHF?J?DCDFGAED:>
TGGT1_chrIV	1054006 A	32T.....^M.^M.^K,^J,	CFFHJ@DCGJEGCJFGJ?DCDHFEFDICDD
TGGT1_chrIV	1054007 T	32a.,,a.....	CFFHJCDDFGEIIFIJ:DDCIHEAFCCDD
TGGT1_chrIV	1054008 G	32	,\$.....	CFFHGCDD<IHHDGFGBDDDGIB:FAJCCDD
TGGT1_chrIV	1054009 C	31T.....\$.....	DFHHCDD<IHHDDIF@JFDDDIJC?HCJFFFDD
TGGT1_chrIV	1054010 T	30	CCHHCDD<IFGIHDJDDDDJ?@H@JFFFDD
TGGT1_chrIV	1054011 G	30	CCFH:DDFJFJFICIHCDCJIC@G@JFFFDD
TGGT1_chrIV	1054012 T	30	,\$,\$.....a.....	>>FH@CCFJFGJH9HBDDCHGC:H@IFFCE
TGGT1_chrIV	1054013 T	28,COMP90016,Computational Genomics	FCDDCCEIHFJPEBDDDGGF9H AJFFDE

Computational Strategy and Challenges

- One strategy: parse the **Mpileup** output and decide if there is a **SNP** at any given position
- Challenges with this approach:
 - Sequencing **Errors**
 - Mappability
- Computational opportunities:
 - Expectations on **frequencies**
- Necessary coping strategies:
 - Cleaning the data
 - Modeling assumptions

Dealing With Sequencing Errors: Utilization of Quality Scores

- Sequencing platforms include a quality string along with the read string, which reflects the confidence in each base.
- They make use of Phred scores*:
 - $Q = -10 \log_{10}(P)$
 - These scores are encoded in Ascii by offsetting the score by an integer (in Illumina 33 or 64):
 - A Base quality of 40 is encoded as Ascii 104: “h”

* Ewing B, Hillier L, Wendl MC, Green P. (1998): Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8(3):175–185. PMID 9521921

Phred Scores and SNP calling

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

“Typically, analyses would first involve a filtering step in which only **high-confidence** bases would be kept. The most common cutoff used would be a Phred-type quality score of **Q20** (QPhred = 20).”*

*Genotype and SNP calling from next-generation sequencing data
(Nat Rev Gen, PMC3593722)

Calling Genotypes

“... call a **heterozygous** genotype if the proportion of the non-reference allele is between 20% and 80%; otherwise, a **homozygous** genotype would be called. This is a fairly standard procedure and works well when the **sequencing depth is high** (>20x), so that the probability of a heterozygous individual falling outside the 20–80% range is small.”*

*Genotype and SNP calling from next-generation sequencing data
(Nat Rev Gen, PMC3593722, 2011)

Probabilistic Genotyping (Modelling Assumptions)

The cutoff solution to calling genotypes is **problematic**, because it does not reflect uncertainties in the calls. Neither does it allow for incorporation of prior knowledge.

$$P(E_i|D) = \frac{P(D|E_i)P(E_i)}{P(D)} = \frac{P(D|E_i)P(E_i)}{\sum_j P(D|E_j)P(E_j)}$$

Use **Bayes'** theorem to give each genotype a posterior probability.

This helps to investigate the data more holistically.

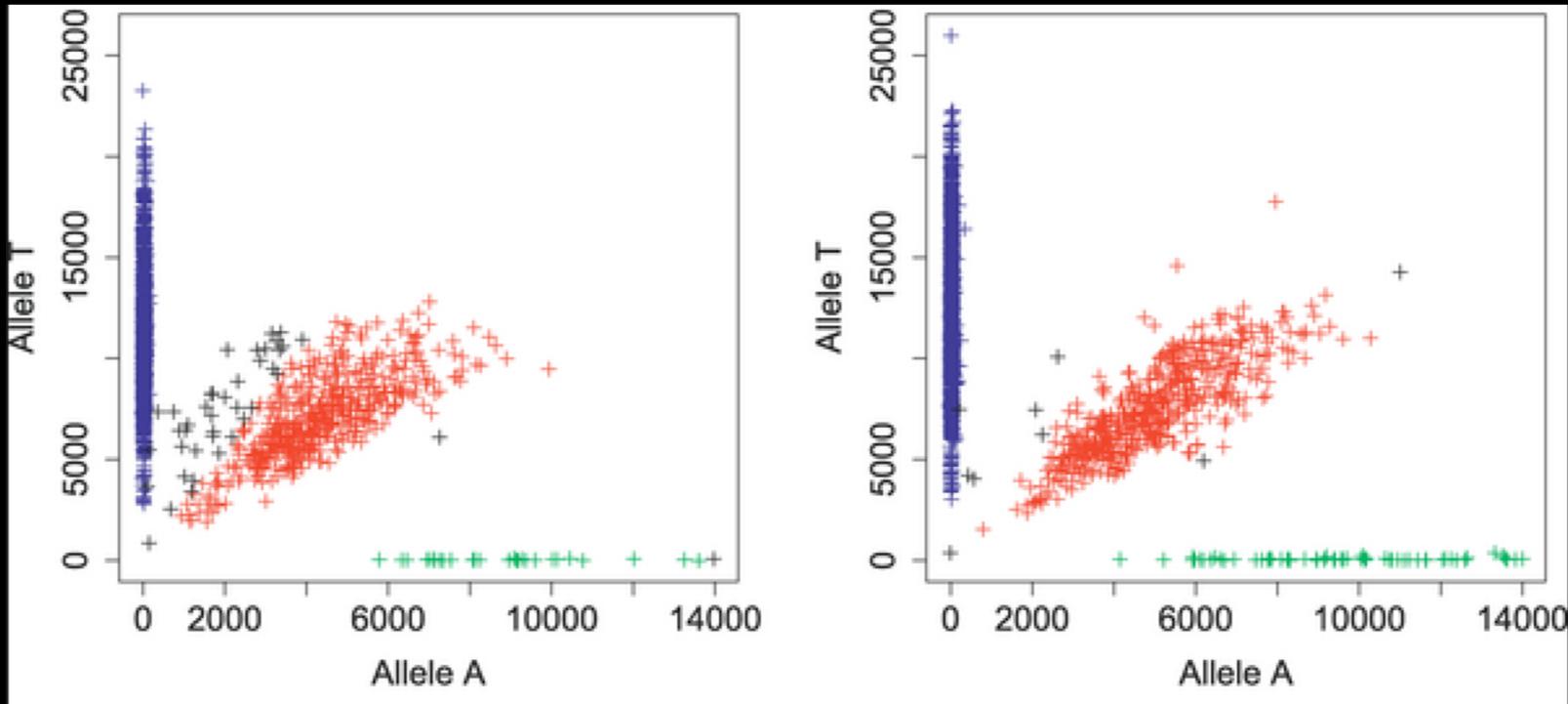
Prior knowledge finds its way into **genotyping**:

- known **population alleles** give higher prior to some alleles.
- some allele changes are more likely than others, this too can be modeled.
- framework can facilitate a range of scenarios: haploid genomes, diploid, ...

Further reading: *SNP calling using genotype model selection on high-throughput sequencing data*, Bioinformatics (2012) 28 (5): 643-650.

doi: 10.1093/bioinformatics/bts001

Genotyping of SNPs



A Method to Address Differential Bias in Genotyping in Large-Scale Association Studies

Plos Genetics 2007: <http://dx.doi.org/10.1371/journal.pgen.0030074>

SNP calling – an example algorithm: Somatic Sniper

- Somatic Sniper is designed to identify variants in a tumour/normal setting – that means variants that are present in one sample (tumour) but not the other (normal) to investigate causative changes.
- SS applies a mathematical model to establish true difference in the two samples genotypes.
- SS offers parameterised quality filtering of data:
 - Minimum mapping quality
 - Minimum base quality
- Example output:

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NORMAL TUMOR
scf_1107000998904 225 . G A . . . GT:IGT:DP:DP4:BCOUNT:GQ:JGQ:VAQ:BQ:MQ:AMQ:SS:SSC
0/0:0:0:4:1,1,0,2:2,0,2,0:1::0:33:40:36:0:. 1/1:1/1:7:0,3,2,2:4,0,3,0:65::65:37:42:44:2:30
scf_1107000998904 1758 . C T . . . GT:IGT:DP:DP4:BCOUNT:GQ:JGQ:VAQ:BQ:MQ:AMQ:SS:SSC
1/1:1/1:15:2,2,11,0:0,4,0,11:124::124:36:43:44:1:. 0/0:0:0:11:7,1,3,0:0,8,0,3:114::218:37:42:42:2:91
scf_1107000998904 1765 . T C . . . GT:IGT:DP:DP4:BCOUNT:GQ:JGQ:VAQ:BQ:MQ:AMQ:SS:SSC
1/1:1/1:18:2,2,14,0:0,14,0,4:117::117:36:43:44:1:. 0/0:0:0:13:8,2,3,0:0,3,0,10:118::226:33:42:41:2:84
```

Somatic Sniper details

- a) Probability estimation for somatic variants:

$$S = -10 \log_{10} \left(\frac{\sum_{i=0}^9 P(T|H_i)P(N|G_i)P(H_i|G_i)P(G_i)}{\sum_{j=0}^9 \sum_{k=0}^9 P(N|G_k)P(T|H_j)P(H_j|G_k)P(G_k)} \right)$$

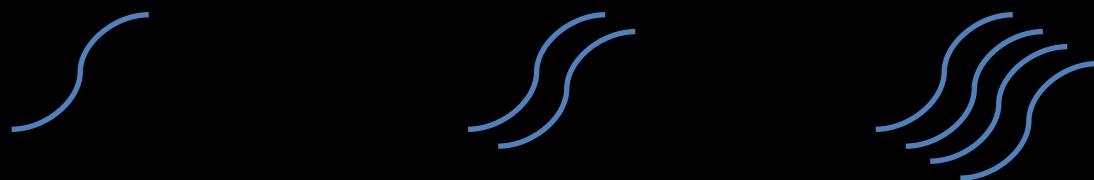
- b) Quality checks to exclude false positives:

1. Site is >10 bp from a predicted indel of quality >=50.
2. Maximum mapping quality at the site is >=40.
3. Fewer than three SNV calls in 10bp window around the site.
4. Site is covered by at least three reads.
5. Consensus quality >=20.
6. SNP quality >=20.
7. No entry in dbSNP.
8. No LOH calls

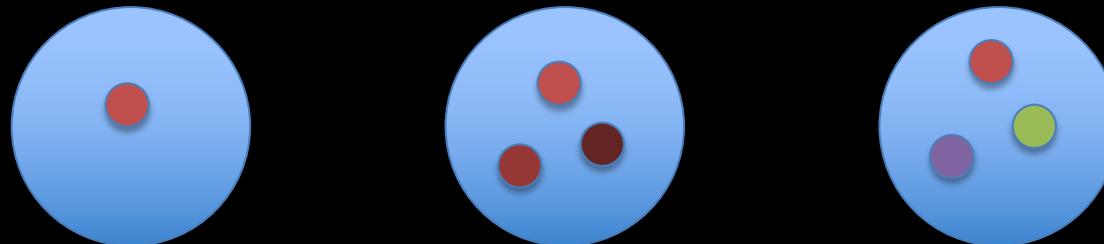
For further details check “**SomaticSniper: identification of somatic point mutations in whole genome sequencing data**” by Larson et al.

Modeling assumptions

- Samples can come from haploid, diploid, or polyploid organisms:

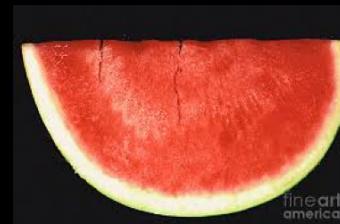


- Samples can come from single cells, clonal populations of cells, or heterogeneous populations



Ploidy Implications

- What frequencies do we expect for variants in a...
 - Haploid genome (E. Coli)?
 - Diploid genome?
 - Triploid genome (seedless watermelon (o_O)?)
 - Tetraploid genome?

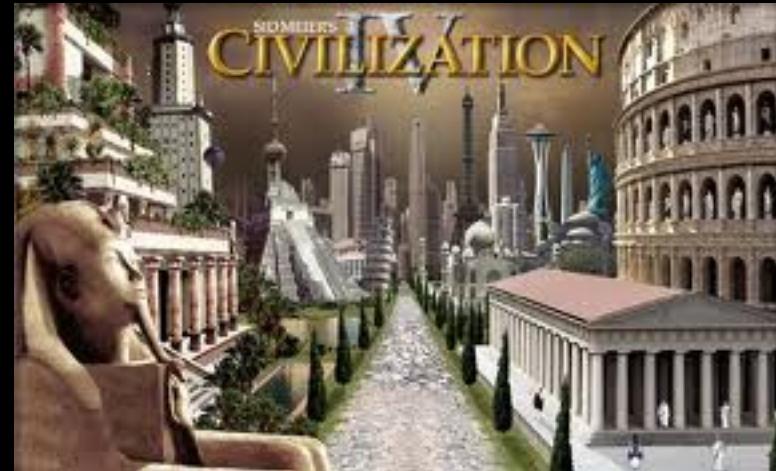


Clonality Implications

- For a diploid organism, what do we expect for variant frequencies in a...
 - Even mixture of two samples?
 - Uneven mixture of two samples?
- Clonality as well as ploidy can vary throughout the genome!
- This is a topic to come in the structural variant lectures.

Application in Bioinformatics: Additional Challenges

- Problems with **big** data:
 - Potentially **millions** of variants in output
 - Too many to eyeball for meaningful results
- Filtering of output
- Annotation of output
- Superimposing **experimental design** to discover results of interest.



Part 4

A BIOINFORMATICS EXAMPLE

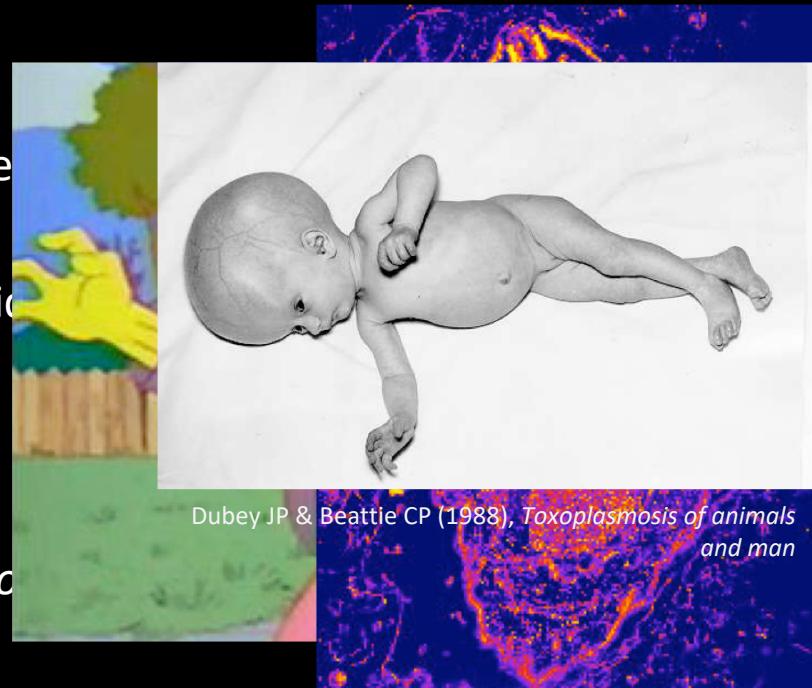
A Project: Toxoplasma Gondii

A project at WEHI with Dr. Chris Tonkin and PhD candidate James McCoy.

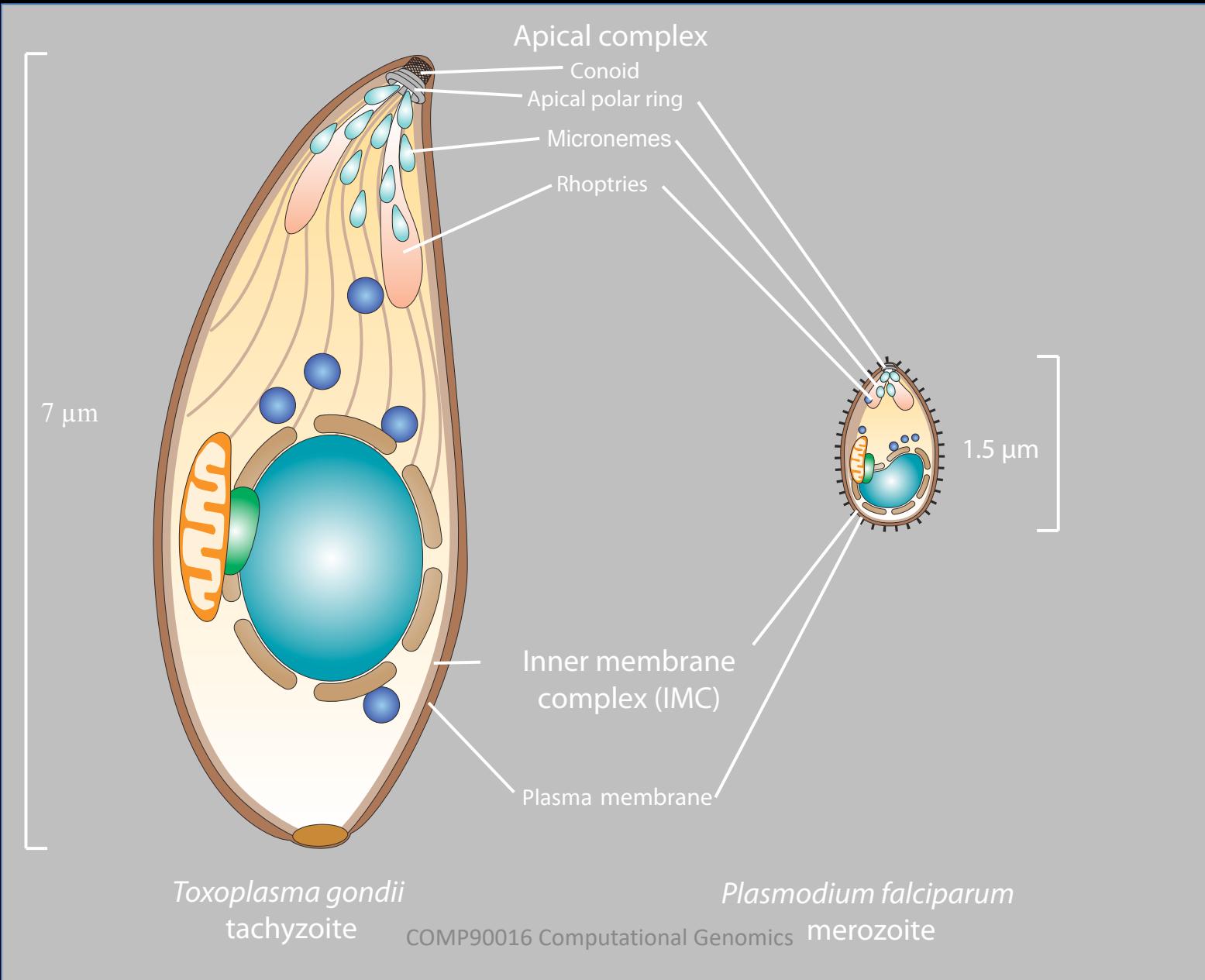
Research into signaling pathways of a parasite.

Toxoplasma gondii in disease

- “Crazy cat lady” parasite
 - Sexual life cycle in cats
 - Humans accidental secondary hosts
- 40 – 80% of individuals in a population infected
- Common cause of disease in vulnerable individuals
 - Birth defects/stillbirth
 - Neurological disease in AIDS patients
- Member of phylum Apicomplexa (with *Plasmodium*)
 - Valuable model organism

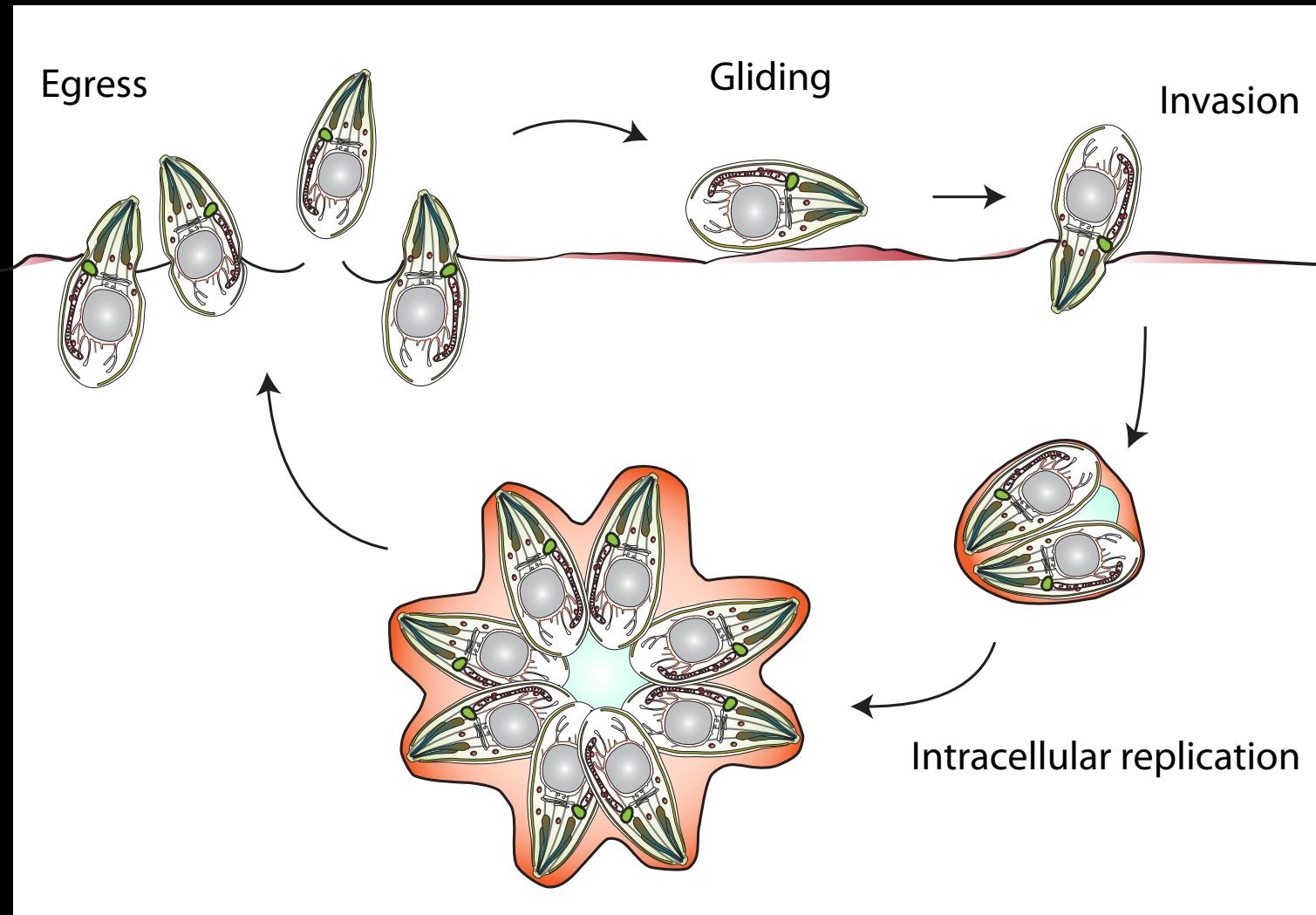


Conservation of apicomplexan invasion components



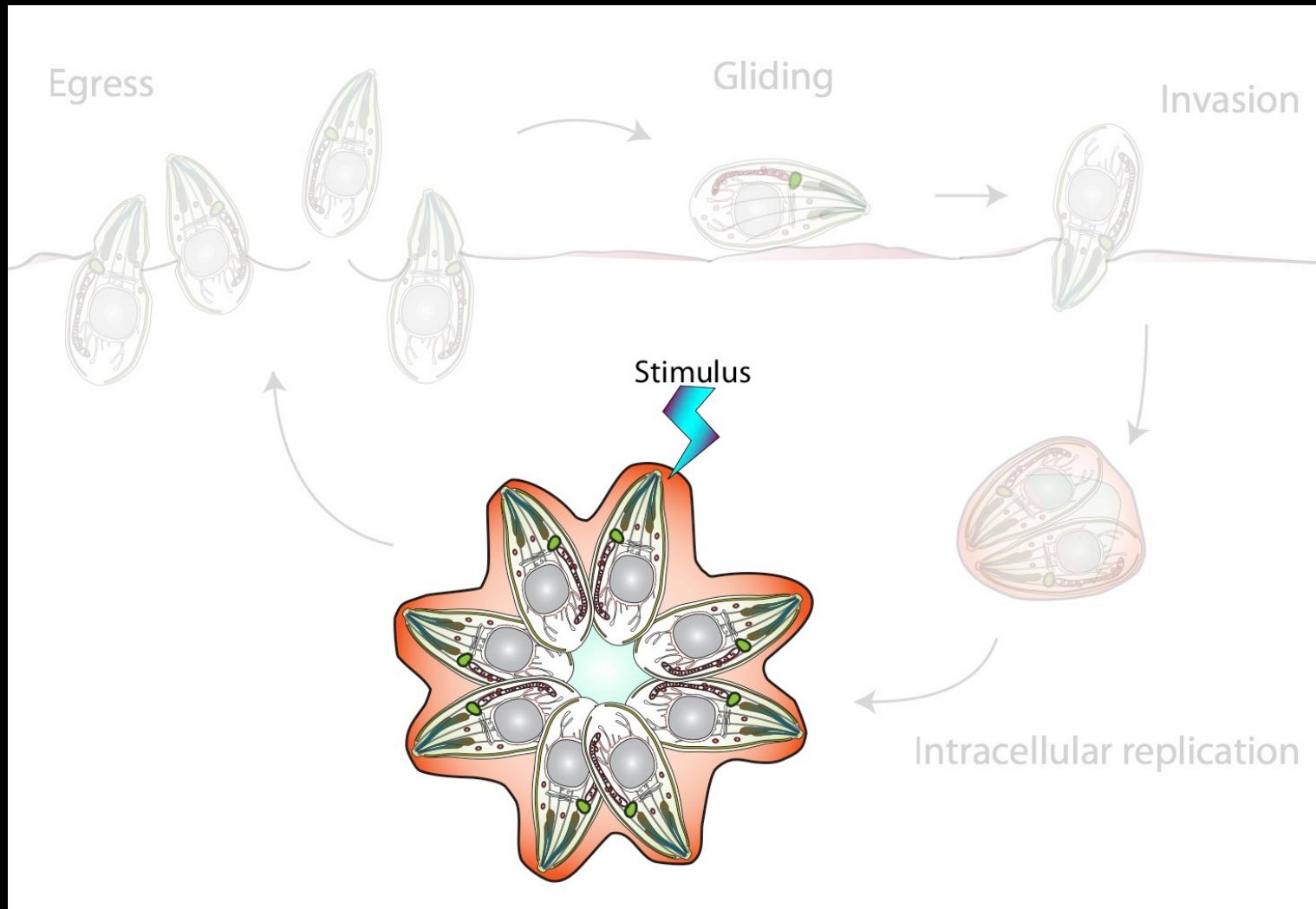
Gliding motility in Apicomplexa

- Substrate-dependent motility and invasion regulated by calcium signalling



Gliding motility in Apicomplexa

- Substrate-dependent motility and invasion regulated by calcium signalling



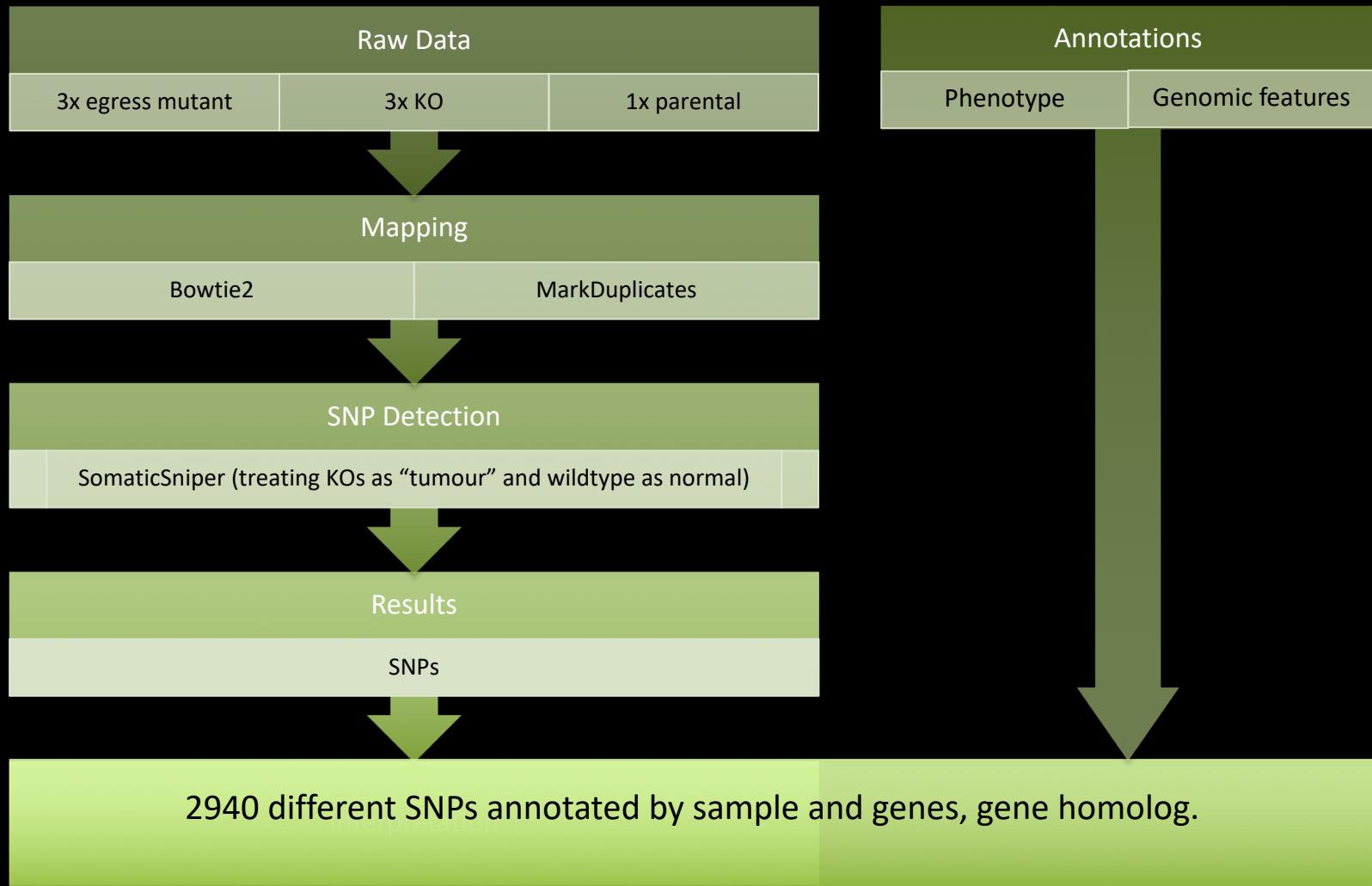
Project Overview

- Summary:
 - Knockout toxoplasma cannot egress from host cell
 - Mutagenize and isolate clones that can egress
 - Sequence mutant, KO & wildtype
 - SNP calling
 - Investigate gene candidates responsible for phenotype

Analysis overview

- Reference:
 - *Toxoplasma Gondii* GT1: ~61Mbp
- Data:
 - 3x KO mutant (no egress), 3x KO mutant (restored egress), 1x wildtype
 - Illumina HiSeq runs to 30x each
- Analysis:
 - Read alignment
 - SNP calling
 - Annotations
 - Evaluation

NGS analysis pipeline



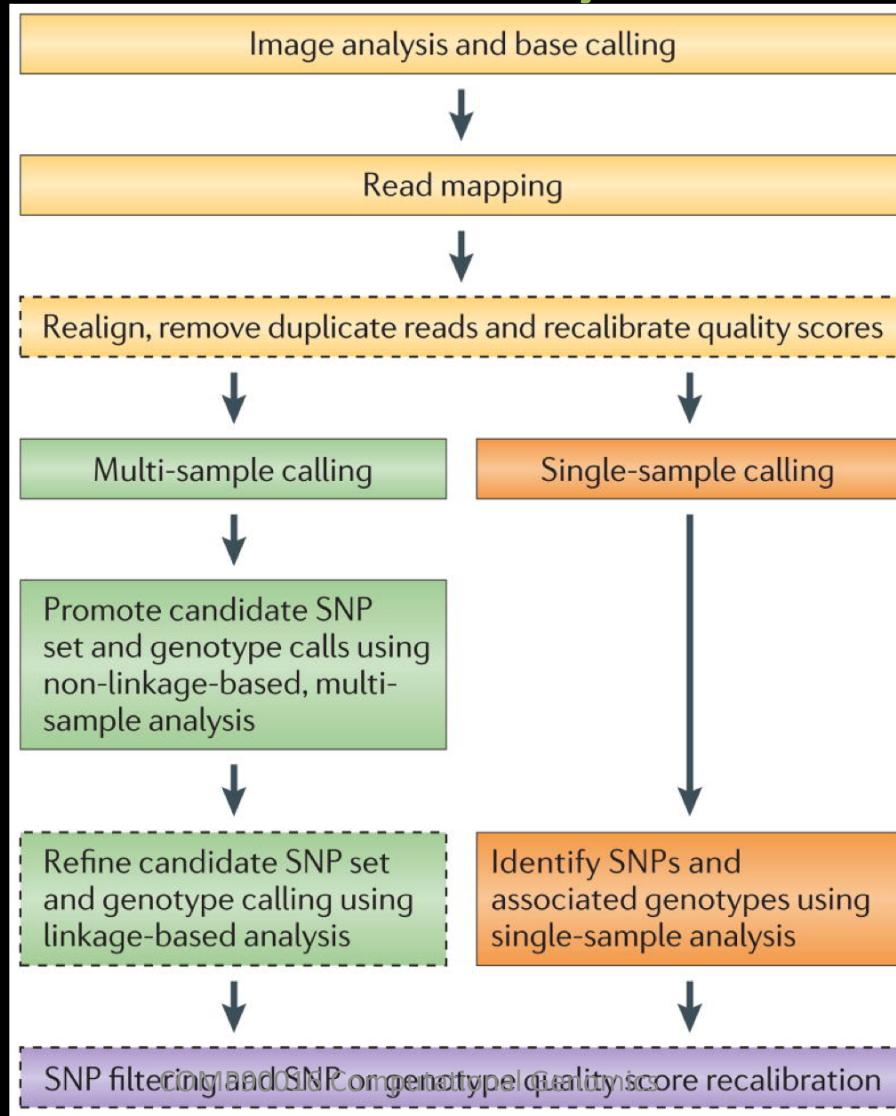
Evaluating the results

- Of the 2940 SNPs...
 - 172 lie within genes
 - 83 of those result in non-synonymous changes
 - Finally, only 3 fulfill our experimental expectation: they are present in all the egress restored samples, but none of the other KOs.
- One of the affected genes is known to be related to suppressing calcium signaling in other related species.
- The result gene candidate was then validated and investigated further in the lab.
- *A Forward-genetic Screen Identifies a Negative Regulator of Rapid Ca2+-dependent Cell Egress in the Intracellular Parasite Toxoplasma gondii.*

J Biol Chem. 2017 Mar 3. pii: jbc.M117.775114. doi: 10.1074/jbc.M117.775114.

SNP Calling from Sequencing Data Summary

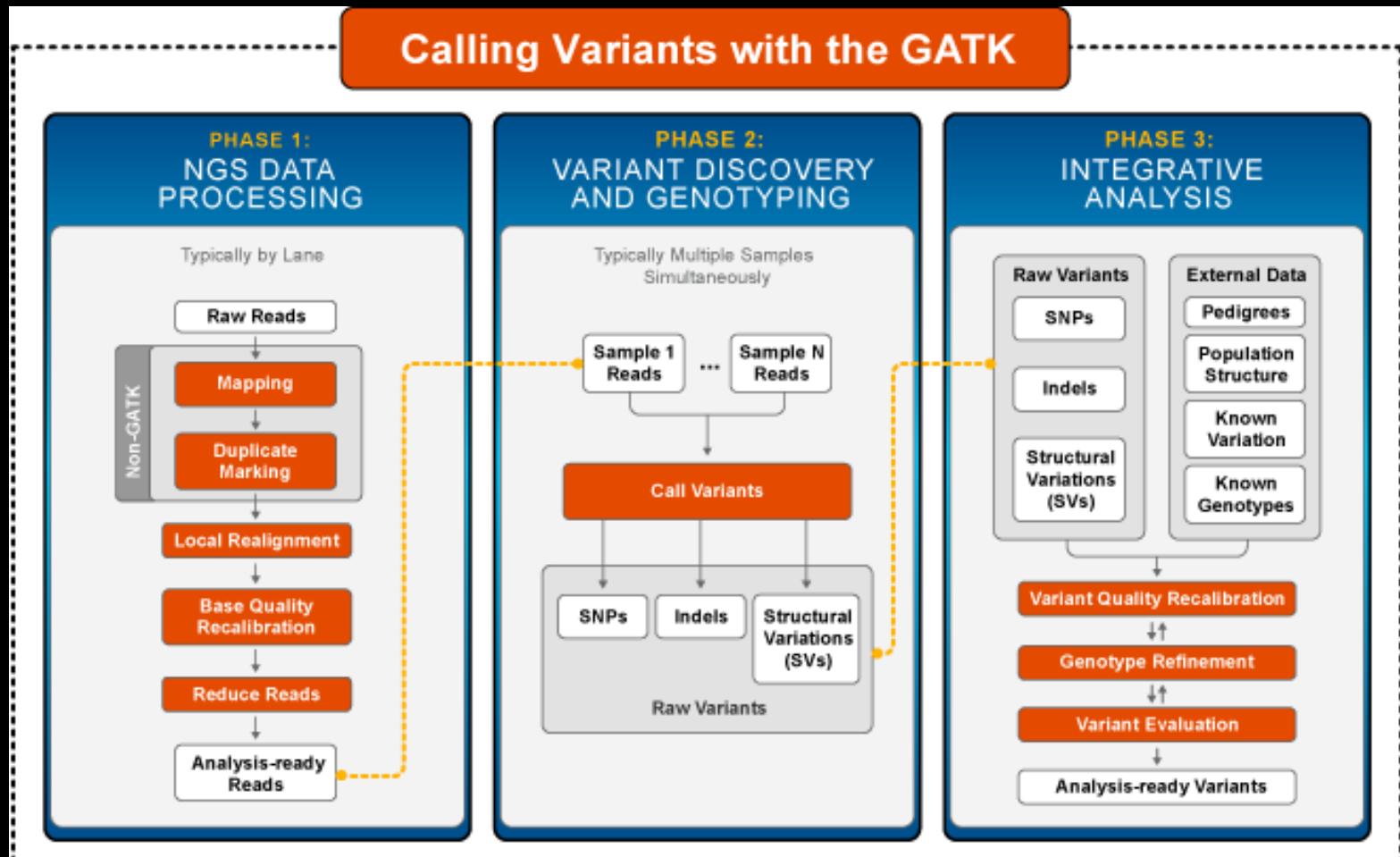
Image from *Genotype and SNP calling from next-generation sequencing data*
(Nat Rev Gen, PMC3593722)



Further reading

https://en.wikipedia.org/wiki/SNP_calling_from_NGS_data

Variant Calling and Genotyping as Done by the Broad Institute



Final Words on the state of SNP Calling

- SNP calling is challenging in general and depending on setting due to:
 - Sequencing errors
 - Mappability
 - PCR duplicates
 - Heterogeneous population of cells
 - Levels of clonality within population
- Therefore, a common practice is to use several SNP callers on the data and investigate variants that are called by more than one tool.

Variant Call Format (VCF)

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=NS,Number=1>Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1>Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##FORMAT=<ID=GT,Number=1>Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1>Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1>Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2>Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001
20 14370 . G A 29 PASS NS=3;DP=14;DB GT:GQ:DP 0|0:48:1
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP 0|0:49:3
20 1110696 . A G,T 67 PASS NS=2;DP=10;AA=T GT:GQ:DP 1|2:21:6
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP 0|0:54:7
...

```

Conclusion

- SNPs are one of the most **abundant** and **easiest** to detect variations between individuals of an organism.
 - The variations are catalogued and distinguished as major and minor alleles.
- Hopefully, you have gained **some** understanding of why they are interesting in general, and what technologies exist to analyse them.
 - Detection and confirmation methods evolved with technological advances.
- Project: **beneficial** (to the organism) SNPs are identified to further understanding of signaling in the parasite.
 - Single nucleotide changes can have a significant impact on phenotypes.