Lecture 22. Final Thoughts

Michael Schatz

April 27, 2017

JHU 600.649: Applied Comparative Genomics

Project Presentations

JHU EN.600.649: Computational Genomics: Applied Comparative Genomics

Project Presentations

Presentations will be a total of 15 minutes: 12 minutes for the presentation, followed by 3 minutes for questions. We will strictly keep to the schedule to ensure that all groups can present in class.

Schedule of Presentations

Day	Time	Team Name	Students	Title
Th 3/27	2:30-2:45	Kayarash	Kayarash Karimian	Comparing the accuracy of PacBio aligners as a function of error rate
.	.	.	.	.
Tu 5/2	1:30-1:45	STaY	Sam Kovaka, Taher Mun, Yunfan Fan	Base call free nanopore read alignment: Aligning nanopores against the reference
Tu 5/2	1:45-2:00	Rachel	Rachel Sherman	Incorporating SVs into Phased Genome Assemblies
Tu 5/2	2:00-2:15	Bang for Your Buck	Isac Lee, Suraj Kanaan, Andrew Fraser	Efficient Selection for Marks in Epigenomic Analysis
Tu 5/2	2:15-2:30	Thank God it's Genomics	Shubhi Bartaria, Saranya Akumalla	Test and apply LACHESIS on long range information from Hi-C
Tu 5/2	2:30-2:45	BS 649	Bayan Al Muhander, Shubha Tirumale	Benchmarking of non-coding mutation analysis schemes on diseased genomes
.	.	.	.	.
Th 5/4	1:30-1:45	Charlotte	Charlotte Darby	The role of mutations in computer binaries
Th 5/4	1:45-2:00	Spradling Lab	Liang-Yu Pang	Identify deletion sites in polyploid Drosophila follicle cell genome.
Th 5/4	2:00-2:15	DeepWorker	Guangyu Yang	Experiments and Extension on DeepVariant
Th 5/4	2:15-2:30	GenomeScope 2	Ravi Gaddipati, Rhyker Ranallo-Benavidez	EGSP: Extending GenomeScope Ploidy
Th 5/4	2:30-2:45	Rock & "Role"	Gherman Uritskiy, Peter DeFord, Xiuqi Chen	Investigating symbiotic roles for nutrient acquisition in a metatranscriptome analysis in rock colonies from the Atacama Desert

Your second genome?

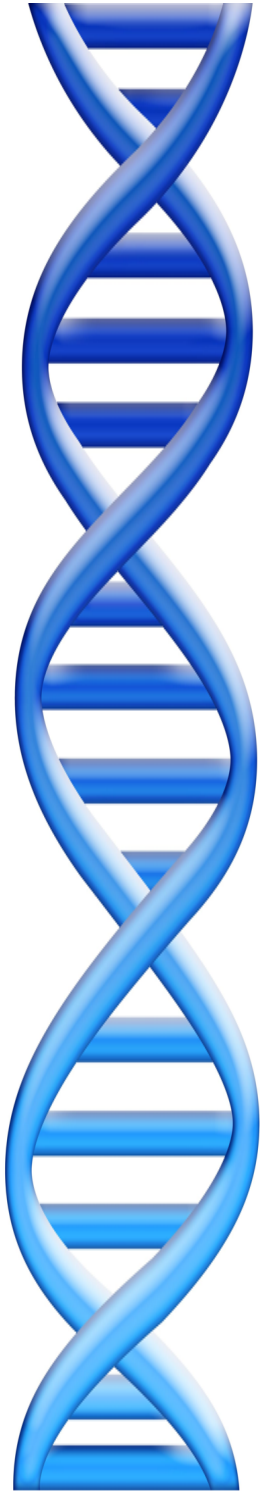
Human body:
~10 trillion cells

Human brain:
~3.3 lbs

Microbiome
~100 trillion cells

Total mass:
~3.3 lbs

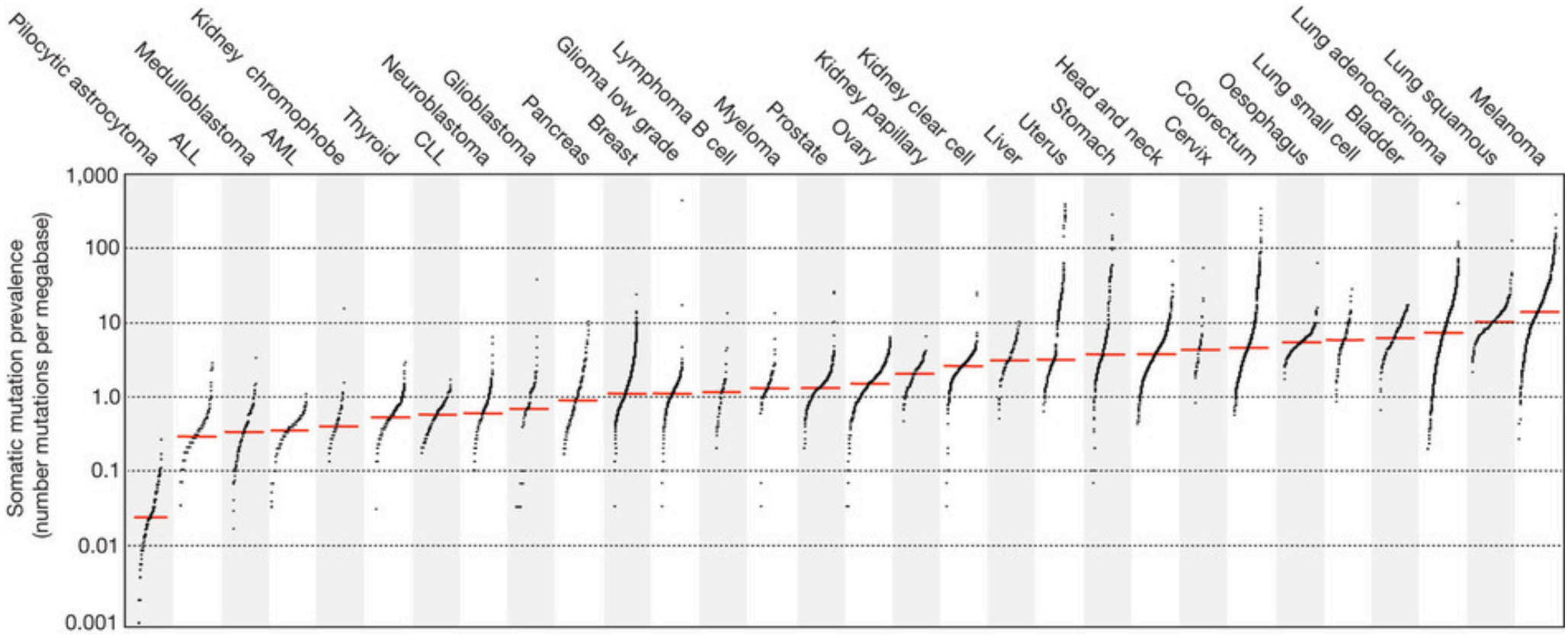
Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans
Sender et al (2016) Cell. http://doi.org/10.1016/j.cell.2016.01.013



Part 3:

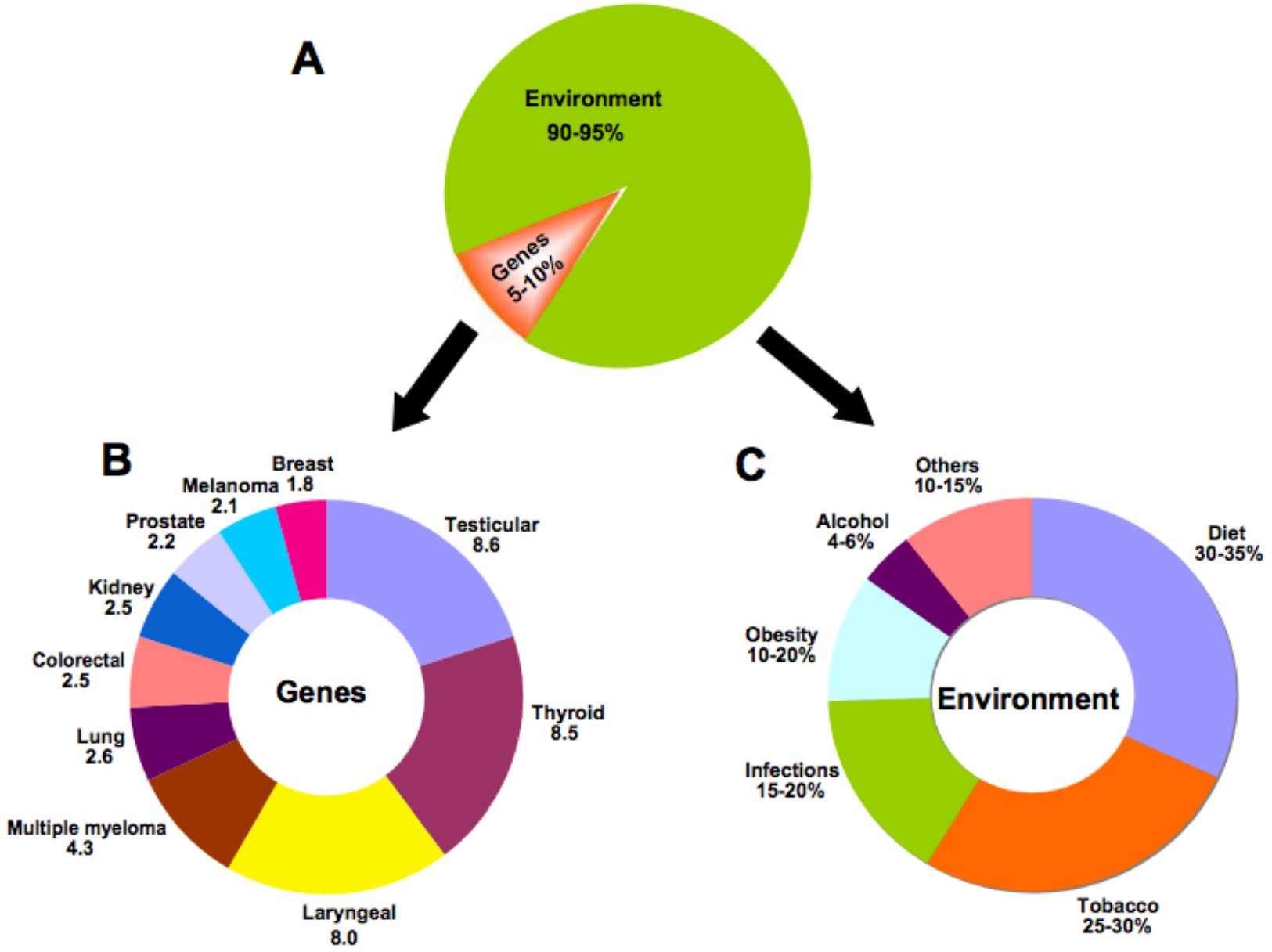
Cancer Genetics

Somatic Mutations In Cancer



Signatures of mutational processes in human cancer
Alexandrov et al (2013) *Nature*. doi:10.1038/nature12477

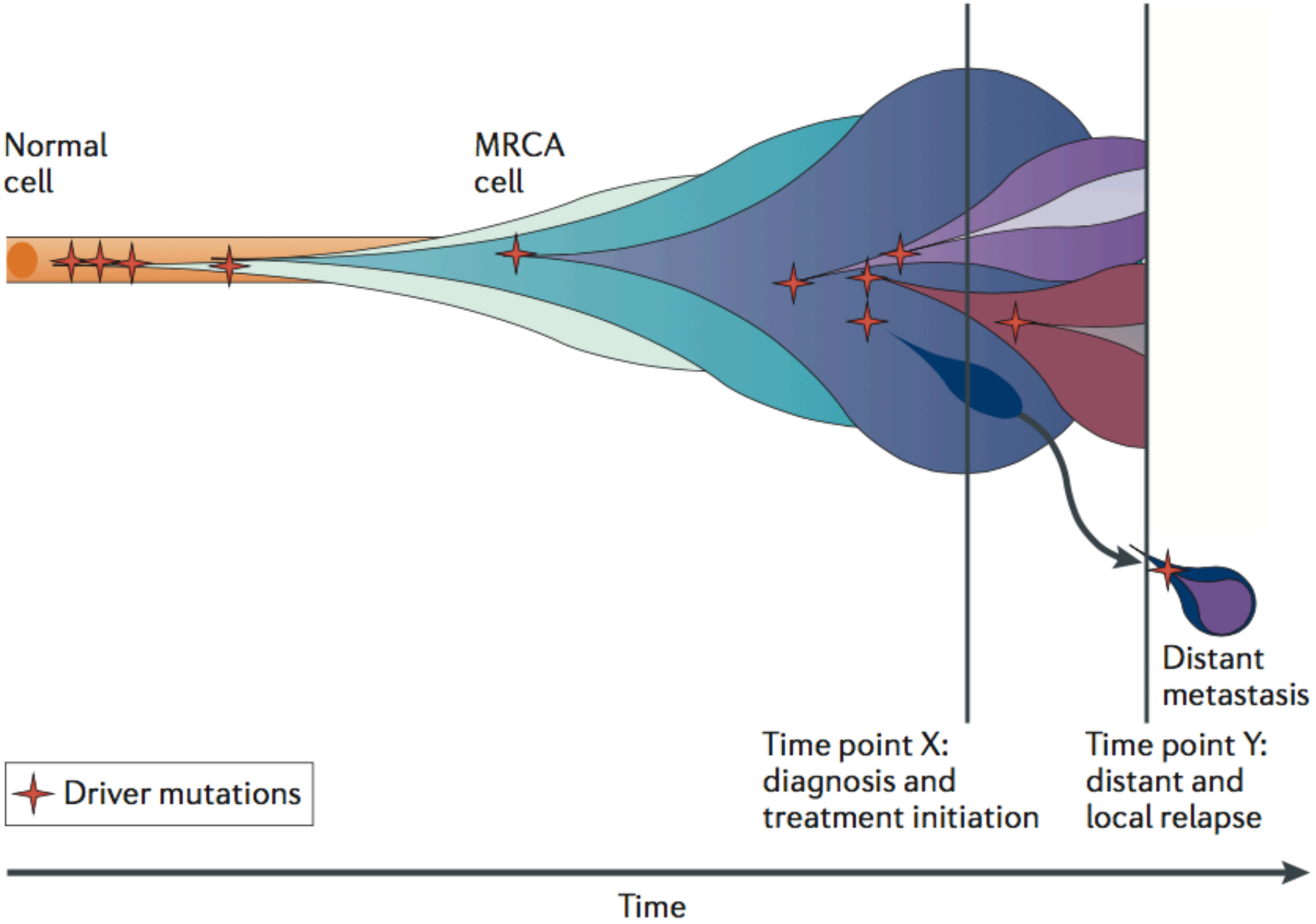
Causes of Cancer



Cancer is a Preventable Disease that Requires Major Lifestyle Changes

Anand et al (2008) Pharmaceutical Research. doi: 10.1007/s11095-008-9661-9

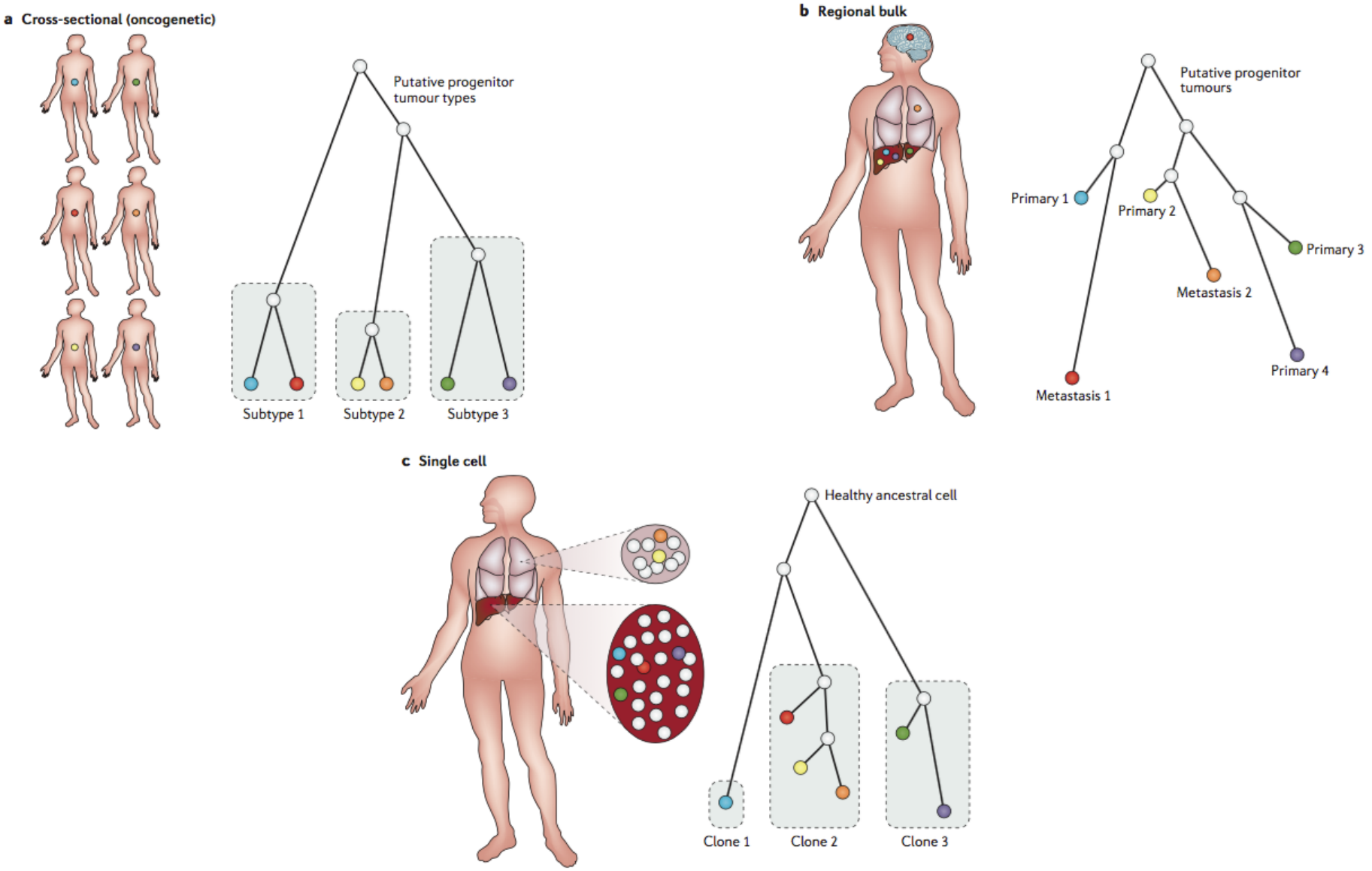
Tumor Evolution



Evolution of the cancer genome

Yates & Campbell (2012) Nature Review Genetics. doi:10.1038/nrg3317

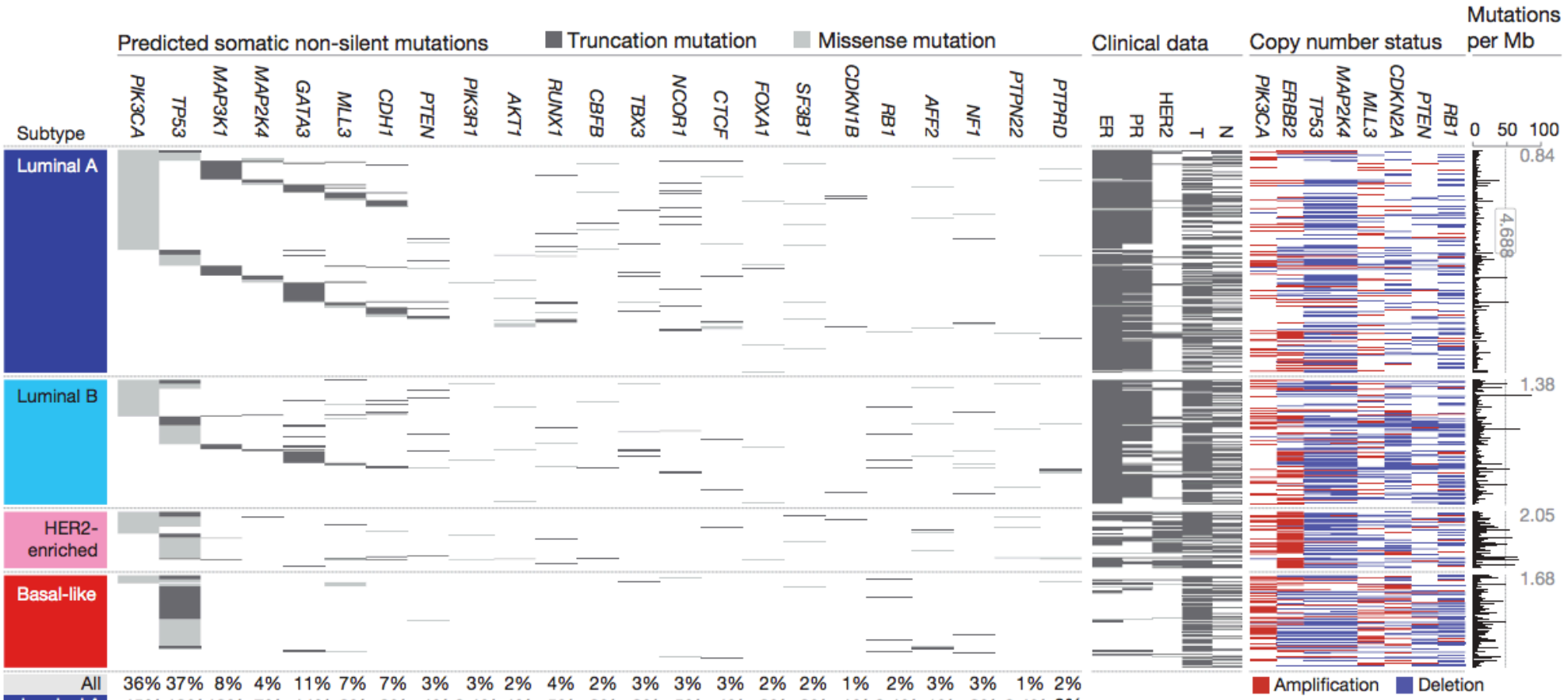
Tumor Heterogeneity



The evolution of tumour phylogenetics: principles and practice

Schwarz and Schaffer (2017) *Nature Reviews Genetics*. *doi:10.1038/nrg.2016.170*

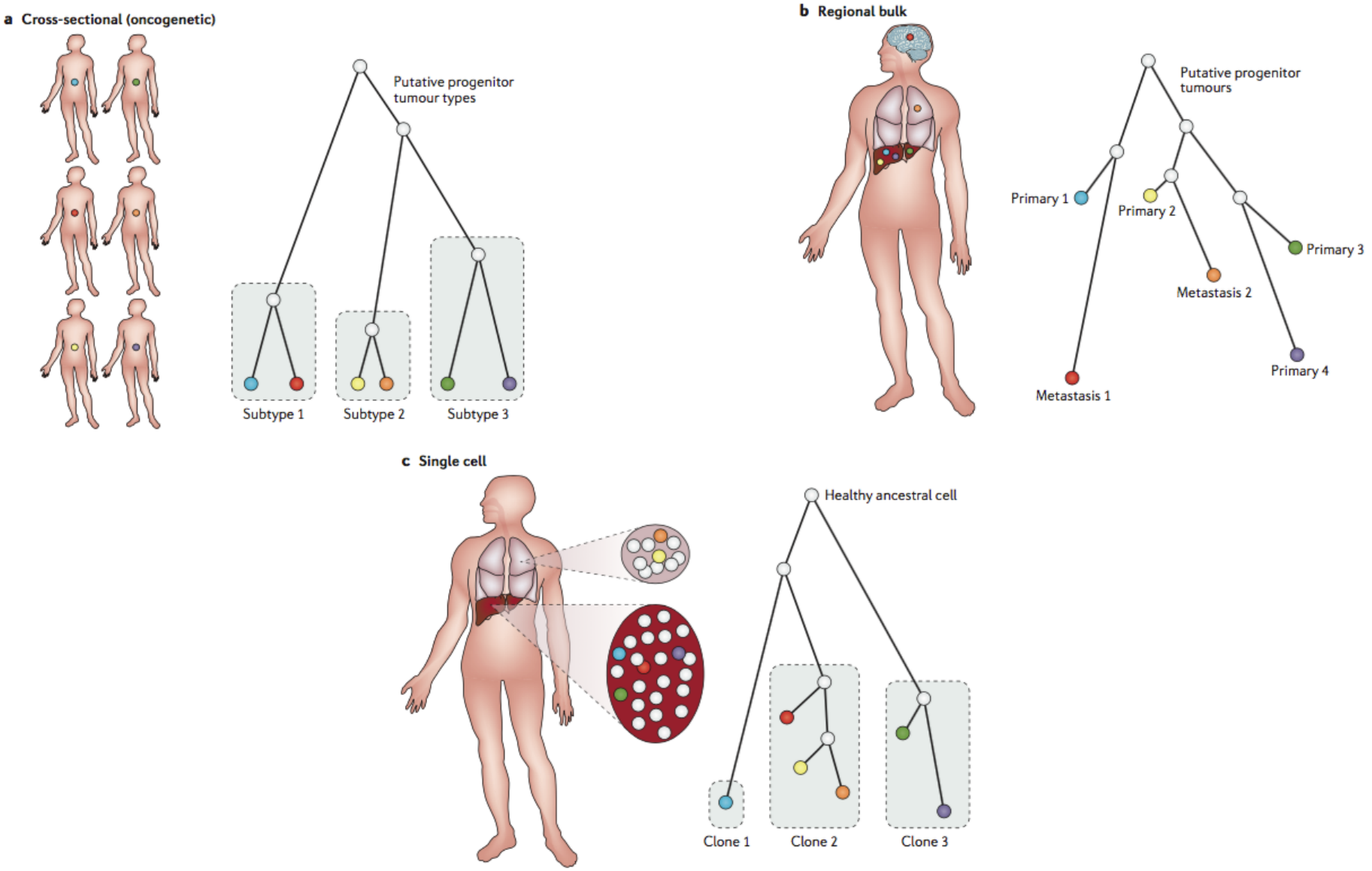
Mutations in Breast Cancer



Comprehensive molecular portraits of human breast tumours

Cancer Genome Atlas Network (2012) Nature. doi:10.1038/nature11412

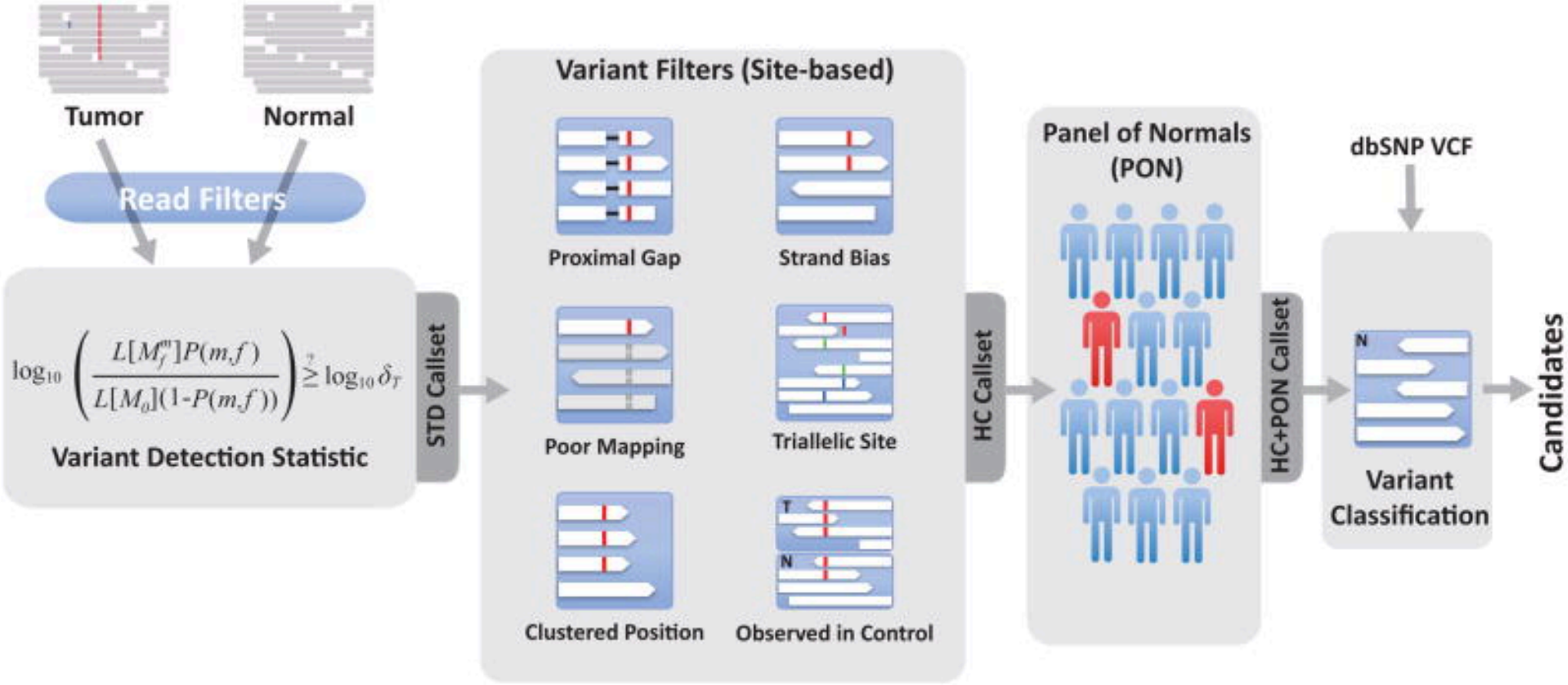
Tumor Heterogeneity



The evolution of tumour phylogenetics: principles and practice

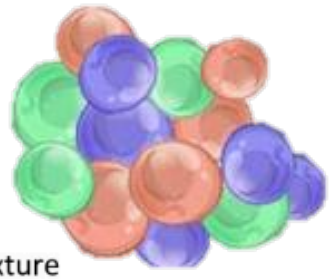
Schwarz and Schaffer (2017) *Nature Reviews Genetics*. *doi:10.1038/nrg.2016.170*

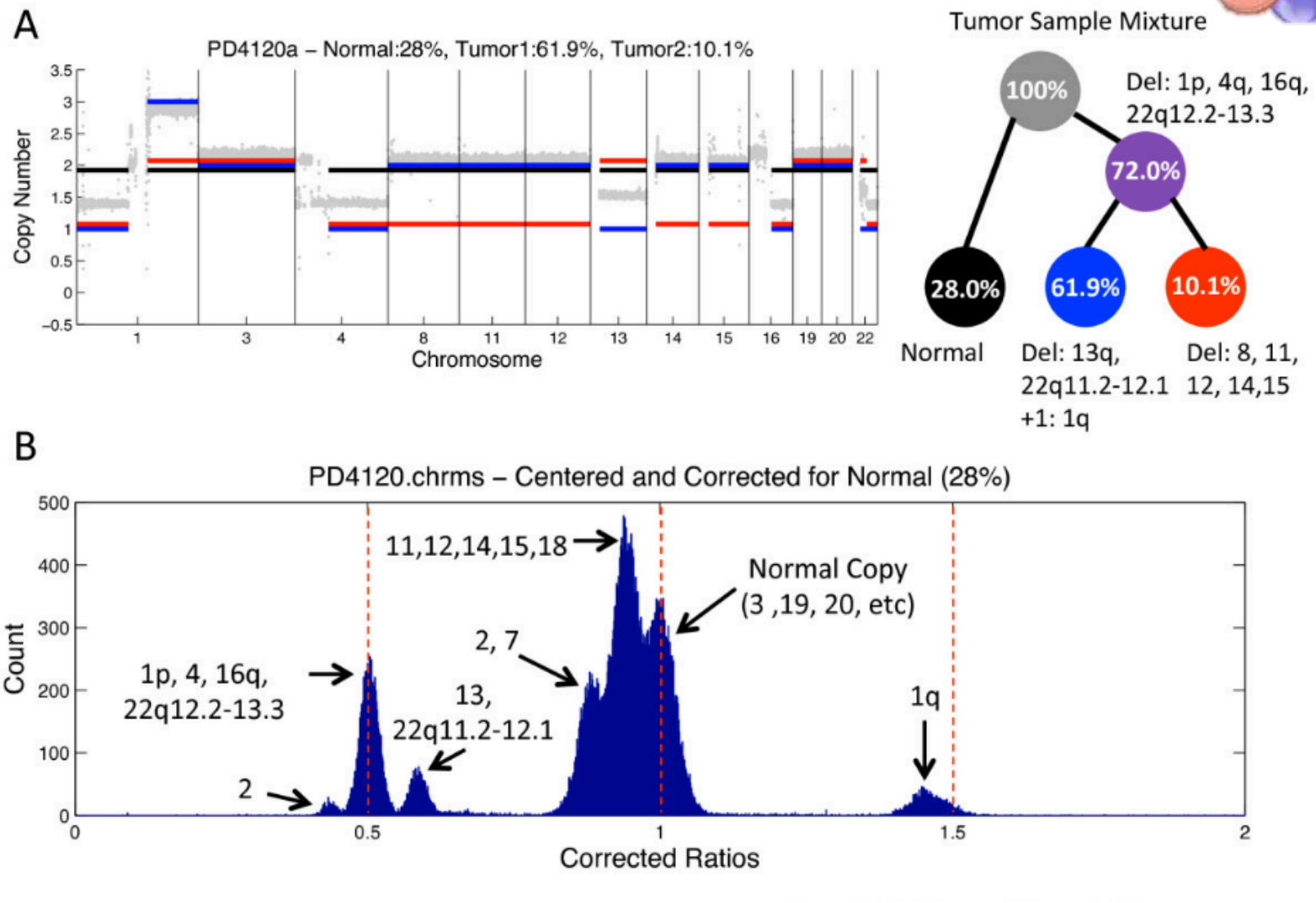
Tumor Heterogeneity



Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples
Cibulskis et al (2013) Nature Biotech. doi:10.1038/nbt.2514

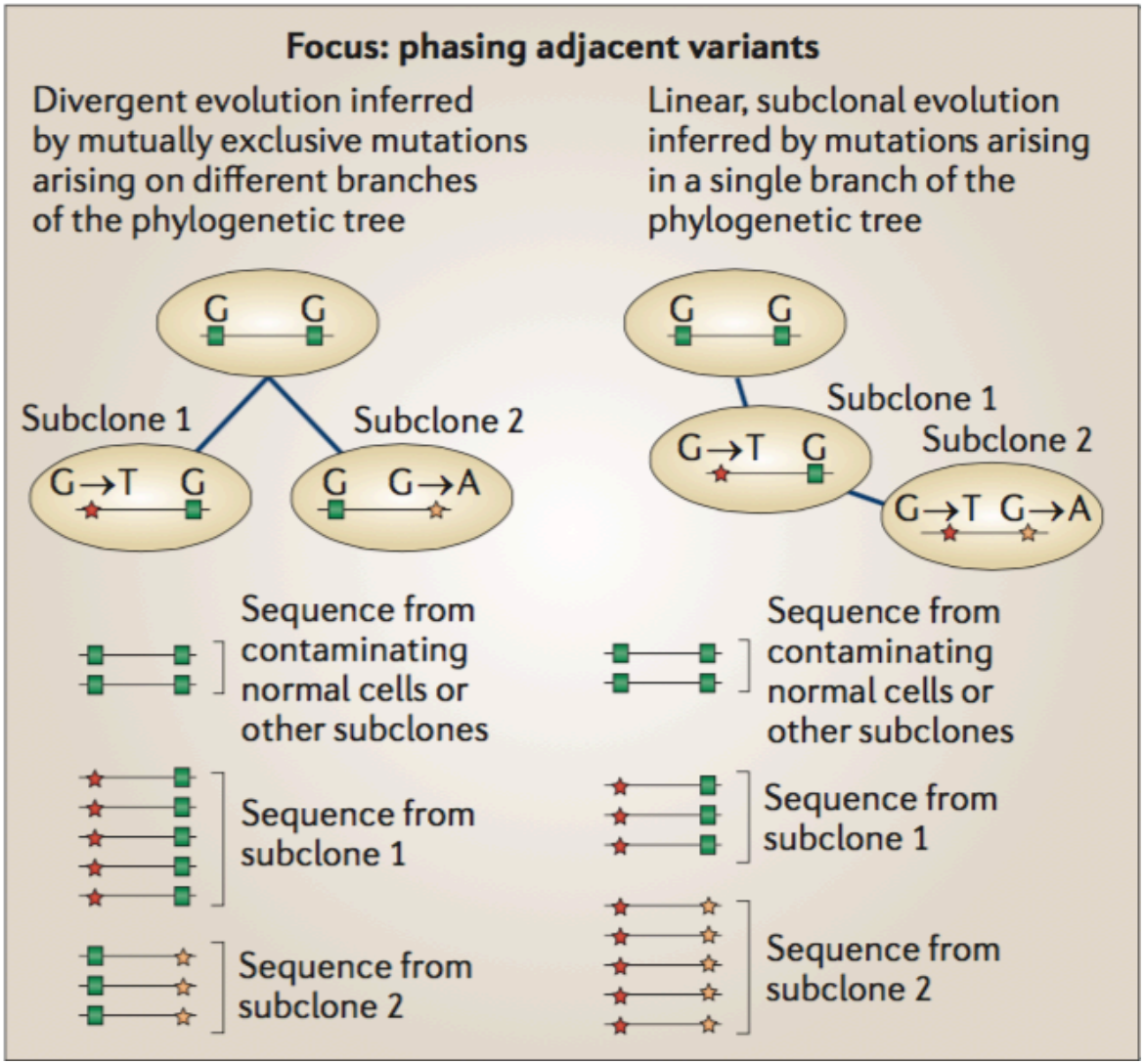
Tumor Heterogeneity





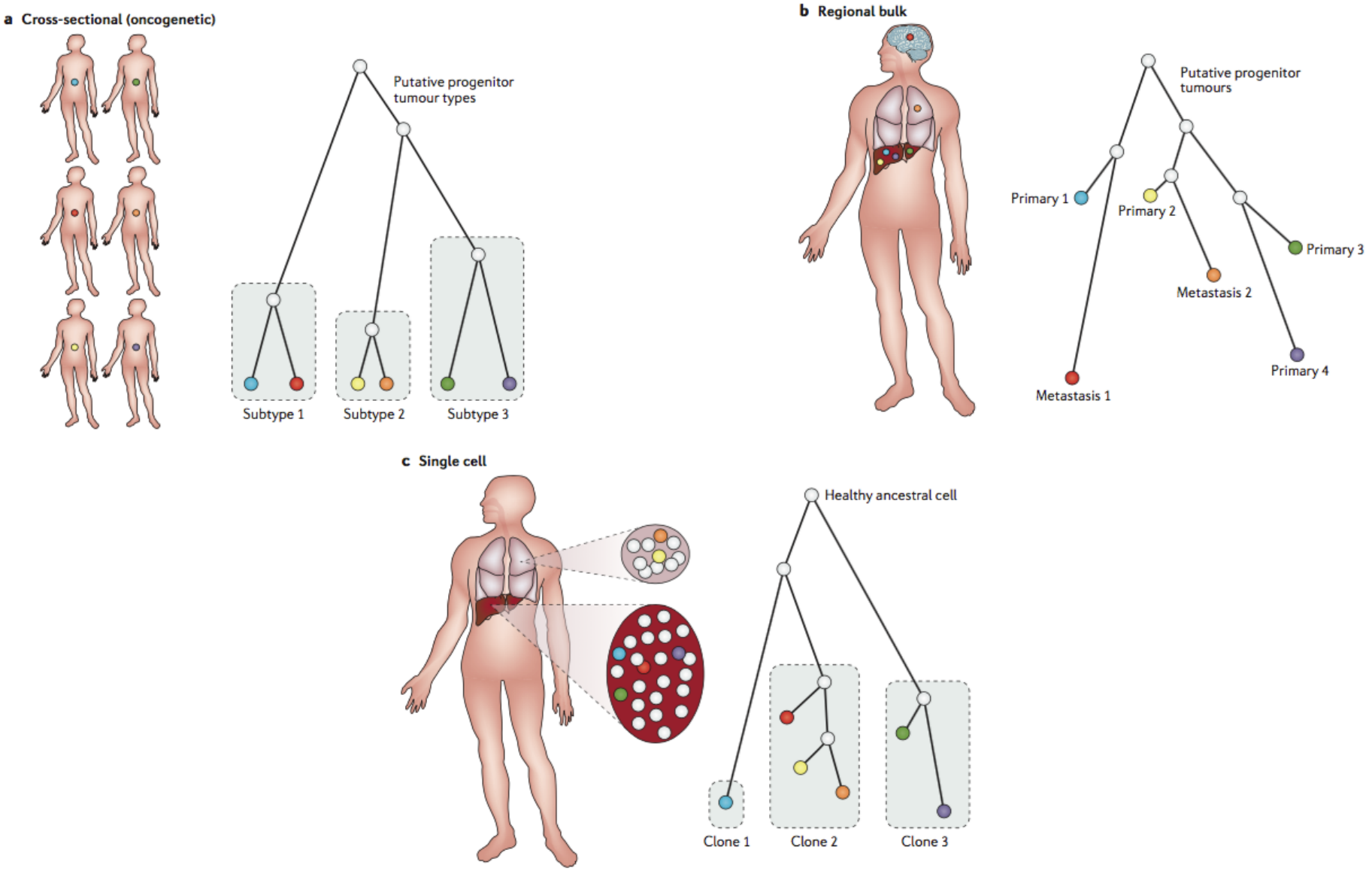
THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data
Oesperet al (2013) Genome Biology. DOI: 10.1186/gb-2013-14-7-r80

Tumor Heterogeneity



Evolution of the cancer genome
Yates & Campbell (2012) Nature Review Genetics. doi:10.1038/nrg3317

Tumor Heterogeneity



The evolution of tumour phylogenetics: principles and practice

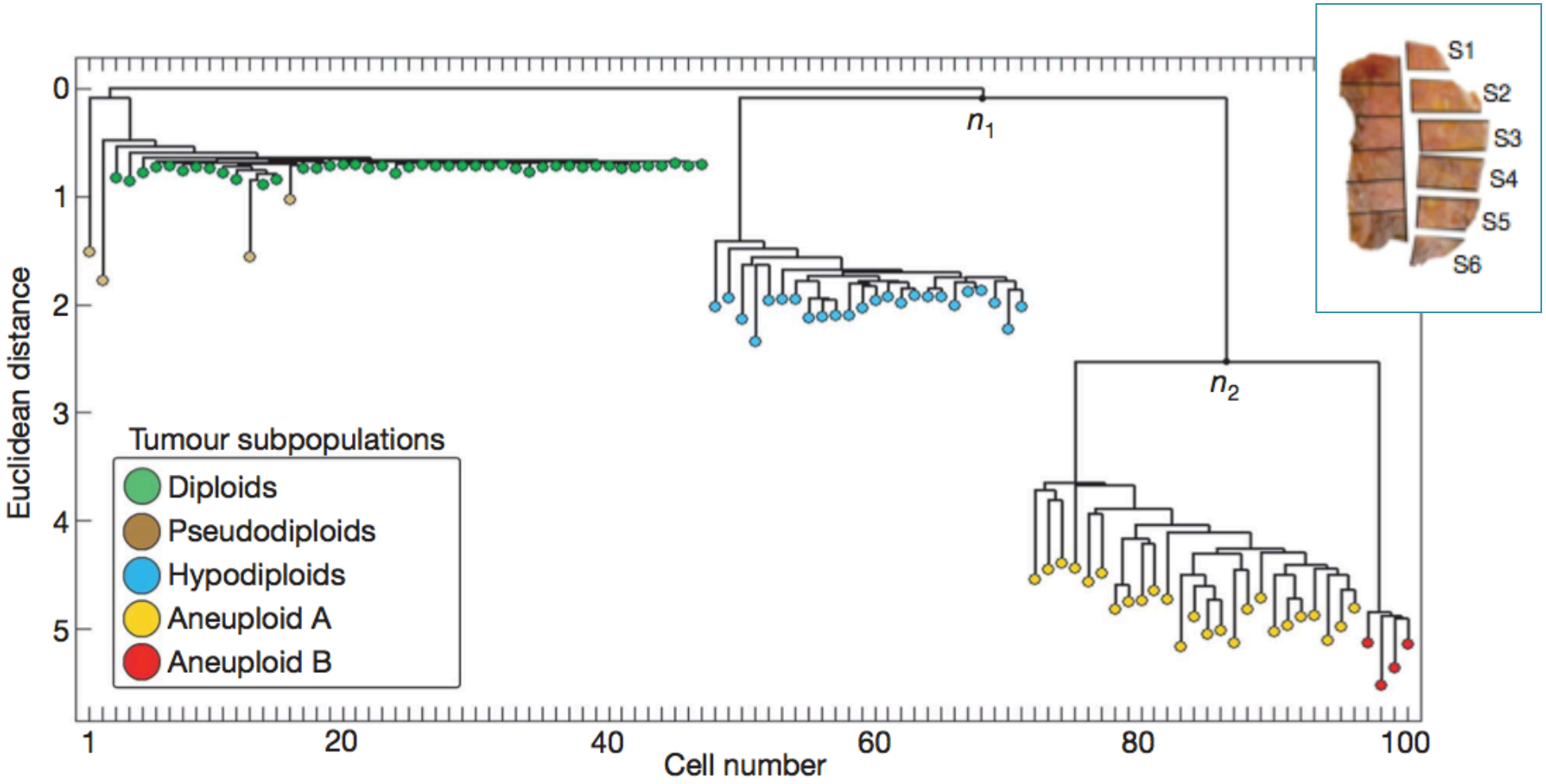
Schwarz and Schaffer (2017) *Nature Reviews Genetics*. *doi:10.1038/nrg.2016.170*

LETTER

doi:10.1038/nature09807

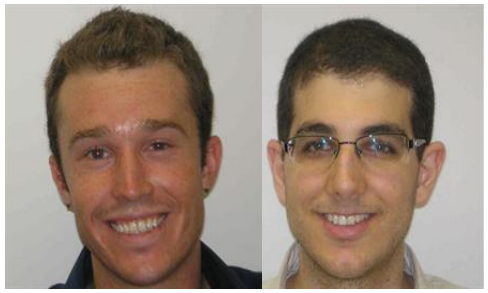
Tumour evolution inferred by single-cell sequencing

Nicholas Navin[1,2], Jude Kendall[1], Jennifer Troge[1], Peter Andrews[1], Linda Rodgers[1], Jeanne McIndoo[1], Kerry Cook[1], Asya Stepansky[1], Dan Levy[1], Diane Esposito[1], Lakshmi Muthuswamy[3], Alex Krasnitz[1], W. Richard McCombie[1], James Hicks[1] & Michael Wigler[1]



Gingko

http://qb.cshl.edu/ginkgo



Interactive Single Cell CNV analysis & clustering

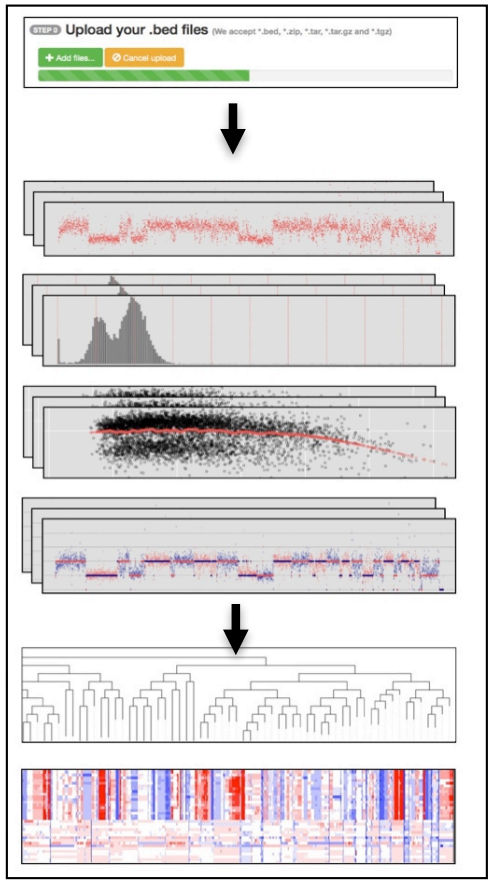
- Easy-to-use, web interface, parameterized for binning, segmentation, clustering, etc
- Per cell through project-wide analysis in any species

Compare MDA, DOP-PCR, and MALBAC

- DOP-PCR shows superior resolution and consistency

Available for collaboration

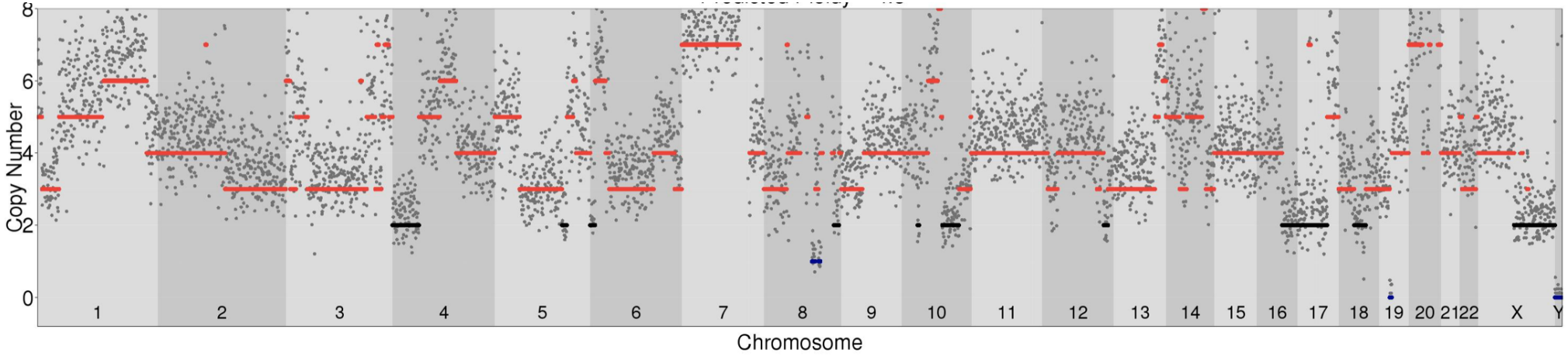
- Analyzing CNVs with respect to different clinical outcomes
- Extending clustering methods, prototyping scRNA



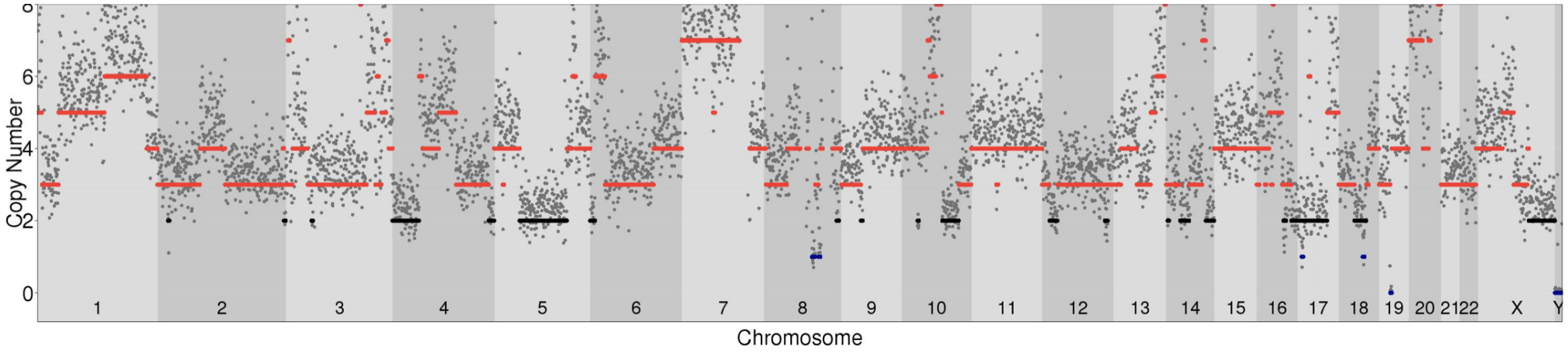
Interactive analysis and assessment of single-cell copy-number variations.
Garvin et al. (2015) Nature Methods doi:10.1038/nmeth.3578

Realtime CNV Analysis

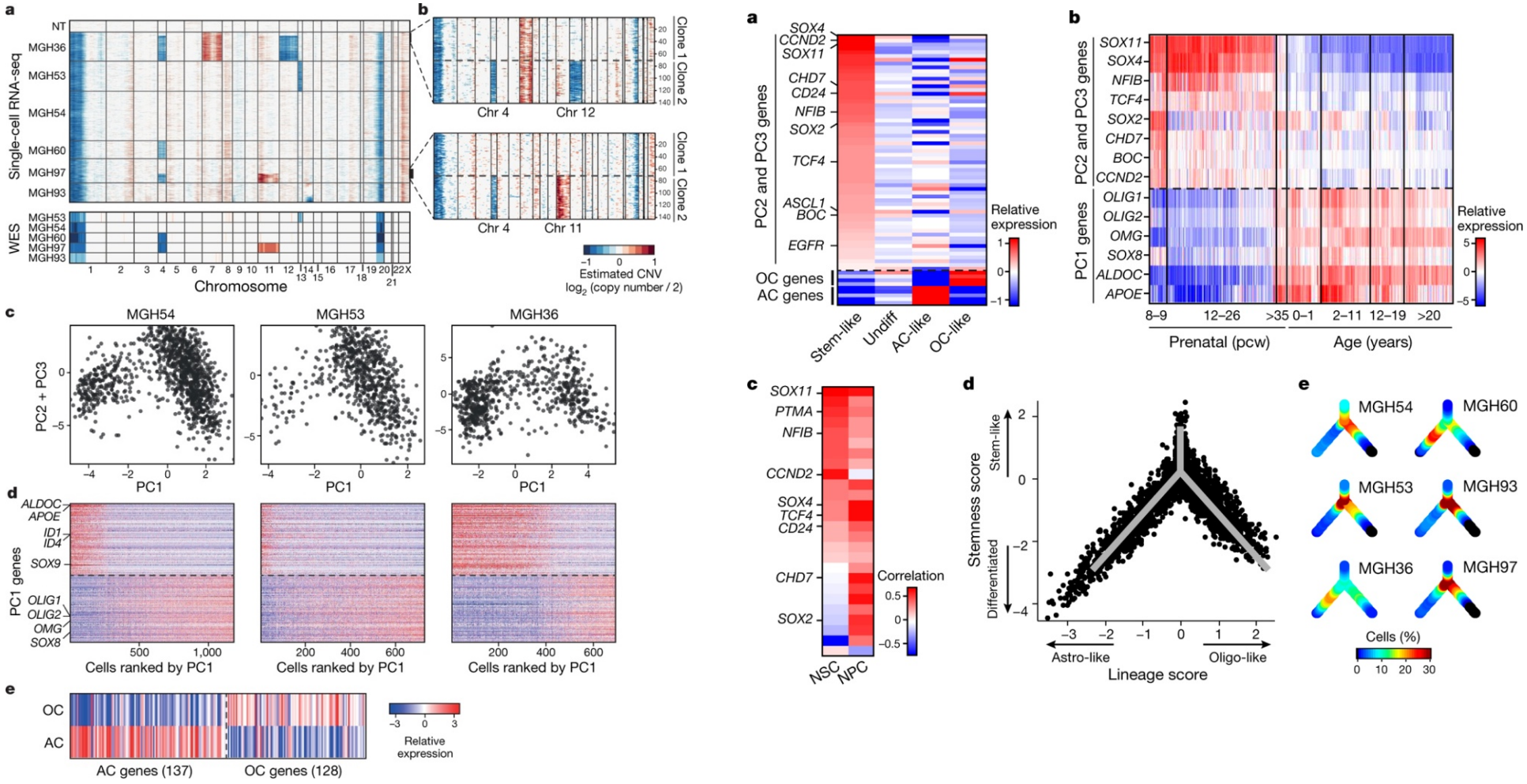
Oxford NANOPORE Technologies



illumina®

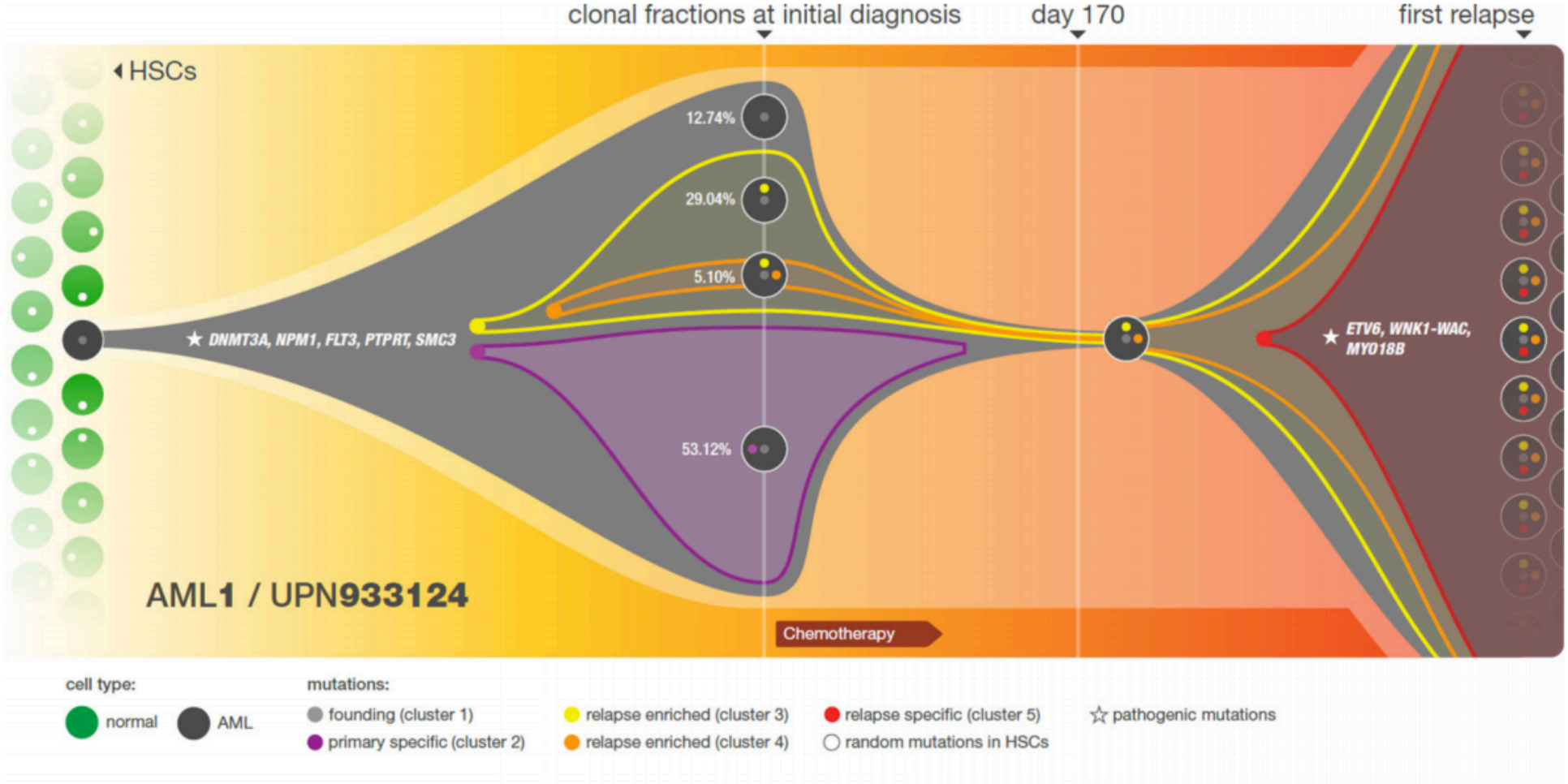


Single Cell RNA-seq of Cancer



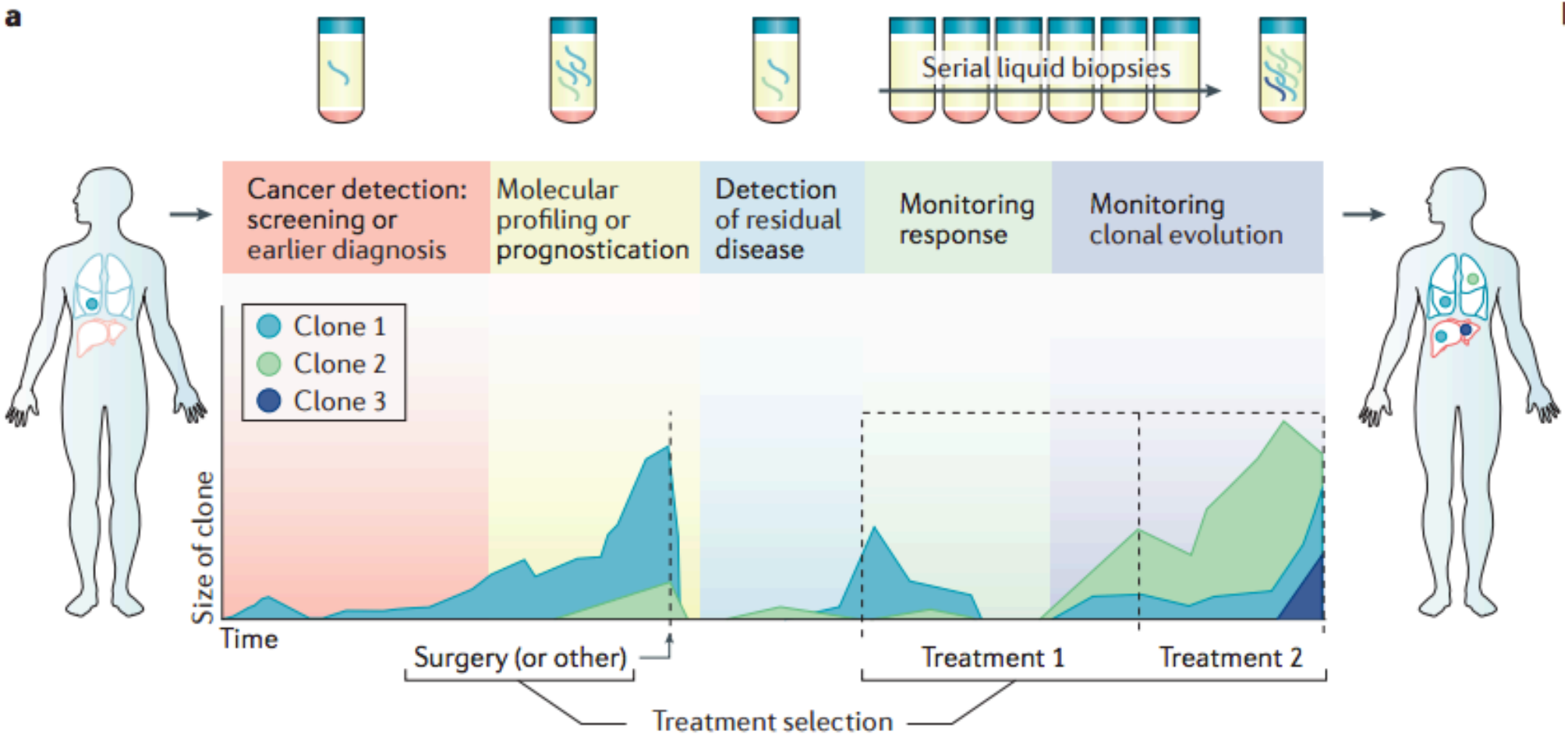
Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma
Tirosh et al (2016) Nature. doi:10.1038/nature20123

Tumor Heterogeneity and Treatment



Clonal evolution in relapsed acute myeloid leukemia revealed by whole genome sequencing
Ding et al (2012) Nature. doi:10.1038/nature10738

Liquid Biopsies



Liquid biopsies come of age: towards implementation of circulating tumour DNA
Wan et al (2017) Nature Review Cancer. doi:10.1038/nrc.2017.7

Liquid Biopsies

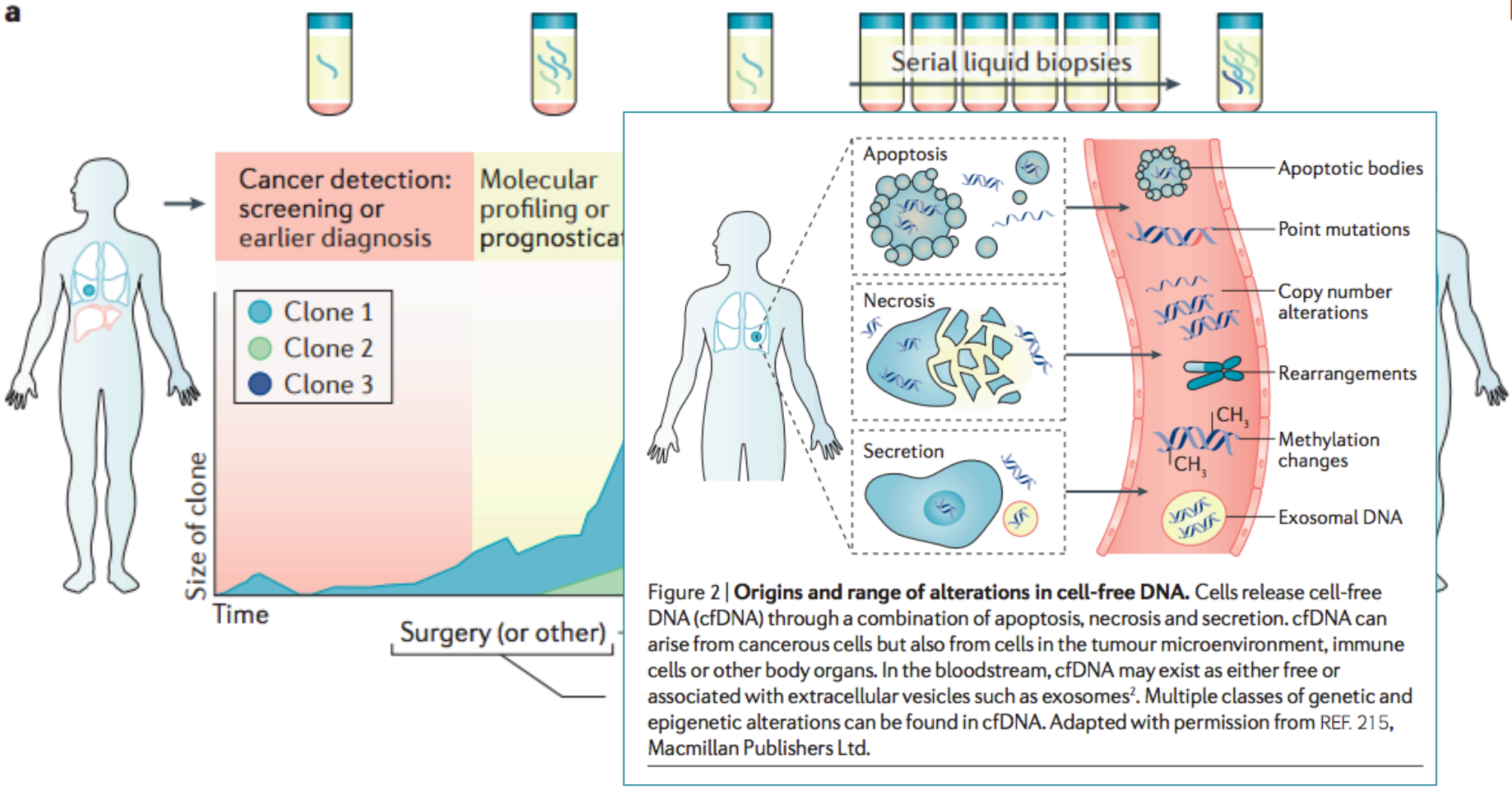
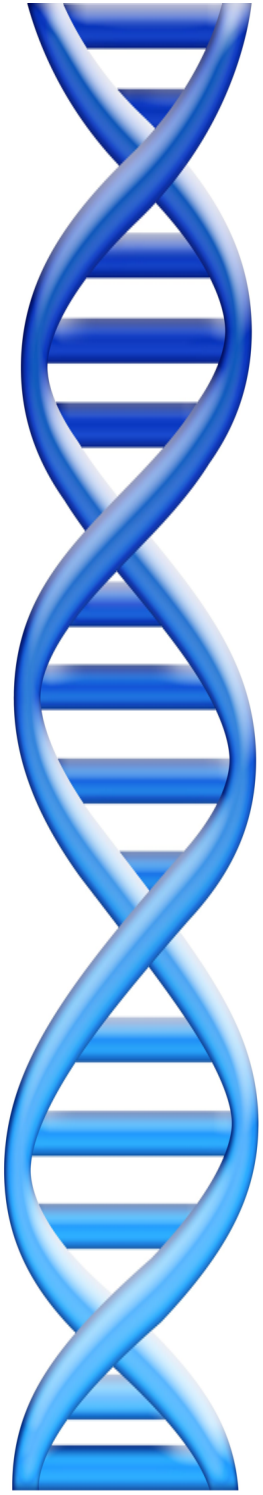


Figure 2 | **Origins and range of alterations in cell-free DNA.** Cells release cell-free DNA (cfDNA) through a combination of apoptosis, necrosis and secretion. cfDNA can arise from cancerous cells but also from cells in the tumour microenvironment, immune cells or other body organs. In the bloodstream, cfDNA may exist as either free or associated with extracellular vesicles such as exosomes[2]. Multiple classes of genetic and epigenetic alterations can be found in cfDNA. Adapted with permission from REF. 215, Macmillan Publishers Ltd.

Liquid biopsies come of age: towards implementation of circulating tumour DNA
Wan et al (2017) Nature Review Cancer. doi:10.1038/nrc.2017.7



Part 4:

Genetic Privacy

Science
AAAS

Identifying Personal Genomes by Surname Inference

Melissa Gymrek *et al.*

Science **339**, 321 (2013);
DOI: 10.1126/science.1229566

What are microsatellites

- **Tandemly repeated sequence motifs**
  - Motifs are 1 – 6 nt long
  - So far, min. 8 nt length, min. 3 tandem repeats for our analyses
- **Ubiquitous in human genome**
  - >5.7 million uninterrupted microsatellites in hg19
- **Extremely unstable**
  - Mutation rate thought to be ~10^{-3} per generation in humans
- **Unique mutation mechanism**
  - Replication slippage during mitosis and meiosis
- **May be under neutral selection**

cCTCTCTCTCTCTCTCTCTCTCTCTCa ➔ $(CT)_{13}$ tCAACAACAACAACAACAACAACAAa ➔ $(CAA)_7$

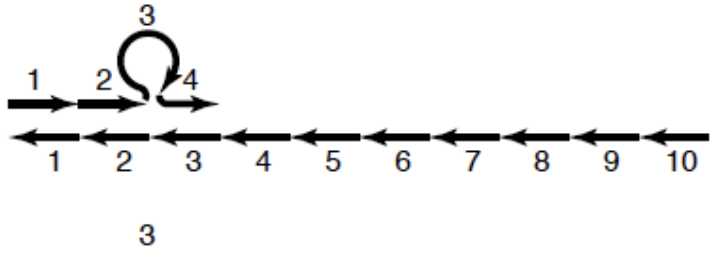
tTTGTCTTGTCTTGTCTTGTCTTGTCTTGTCc ➔ $(TTGTC)_6$ cCATTCATTCATTCATTa ➔ $(CATT)_4$

Microsatellites: Simple Sequences with Complex Evolution
Ellegren (2004) *Nature Reviews Genetics*. doi:10.1038/nrg1348

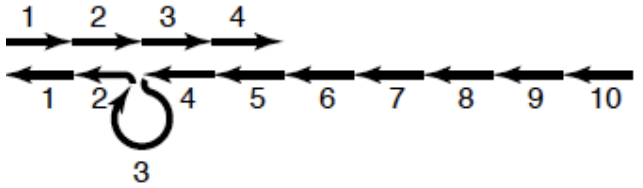
Replication slippage

- **Out-of-phase re-annealing**
  - Nascent and template strands dissociate and re-anneal out-of-phase
- **Loops repaired by mismatch repair machinery (MMR)**
  - Very efficient for small loops
  - Possible strand-specific repair
- **Stepwise process**
  - Nascent strand gains or loses full repeat units
  - Typically single unit mutations
- **Varies by motif length, motif composition, etc.**

Expansion:

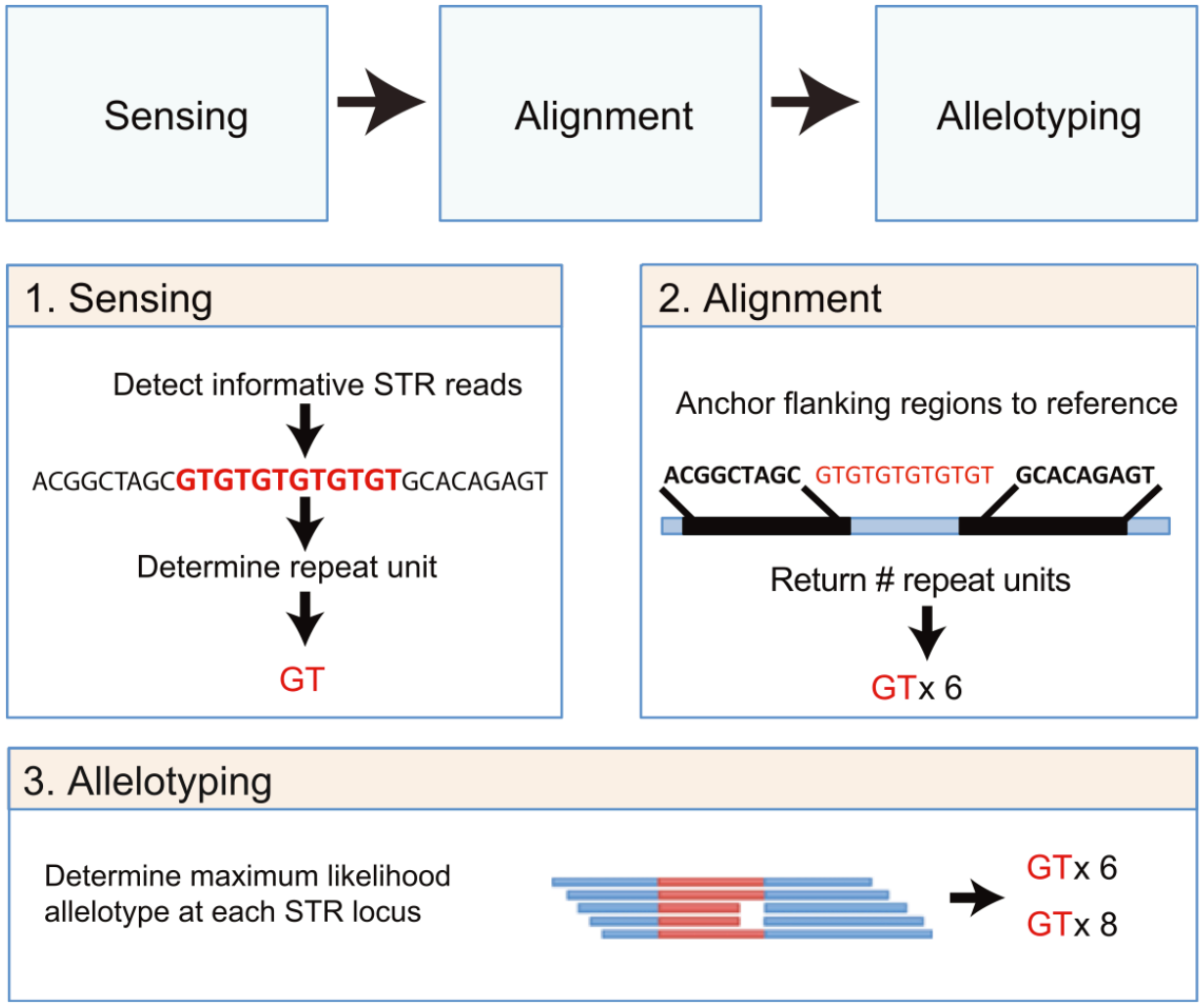


Contraction:



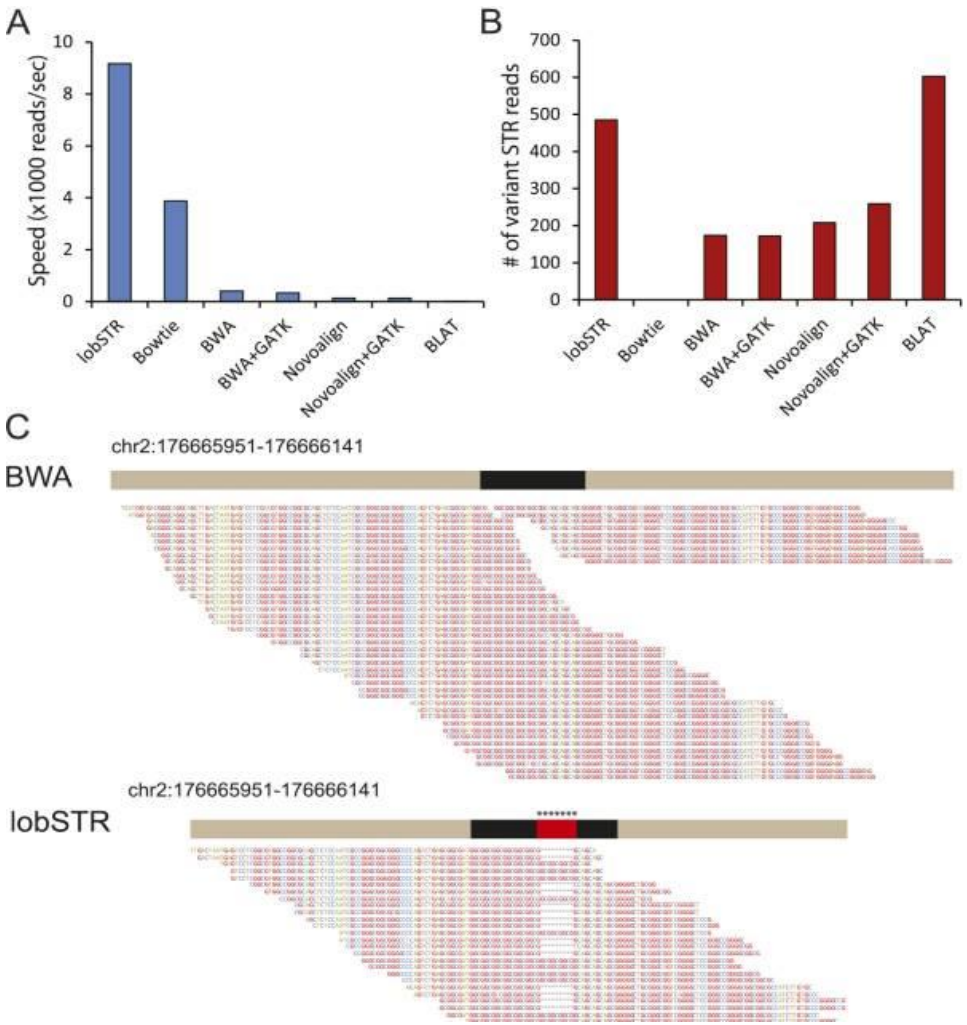
Microsatellites: Simple Sequences with Complex Evolution
Ellegren (2004) *Nature Reviews Genetics*. doi:10.1038/nrg1348

lobSTR Algorithm Overview



lobSTR: A short tandem repeat profiler for personal genomes
Gymrek et al. (2012) *Genome Research.* doi:10.1101/gr.135780.111

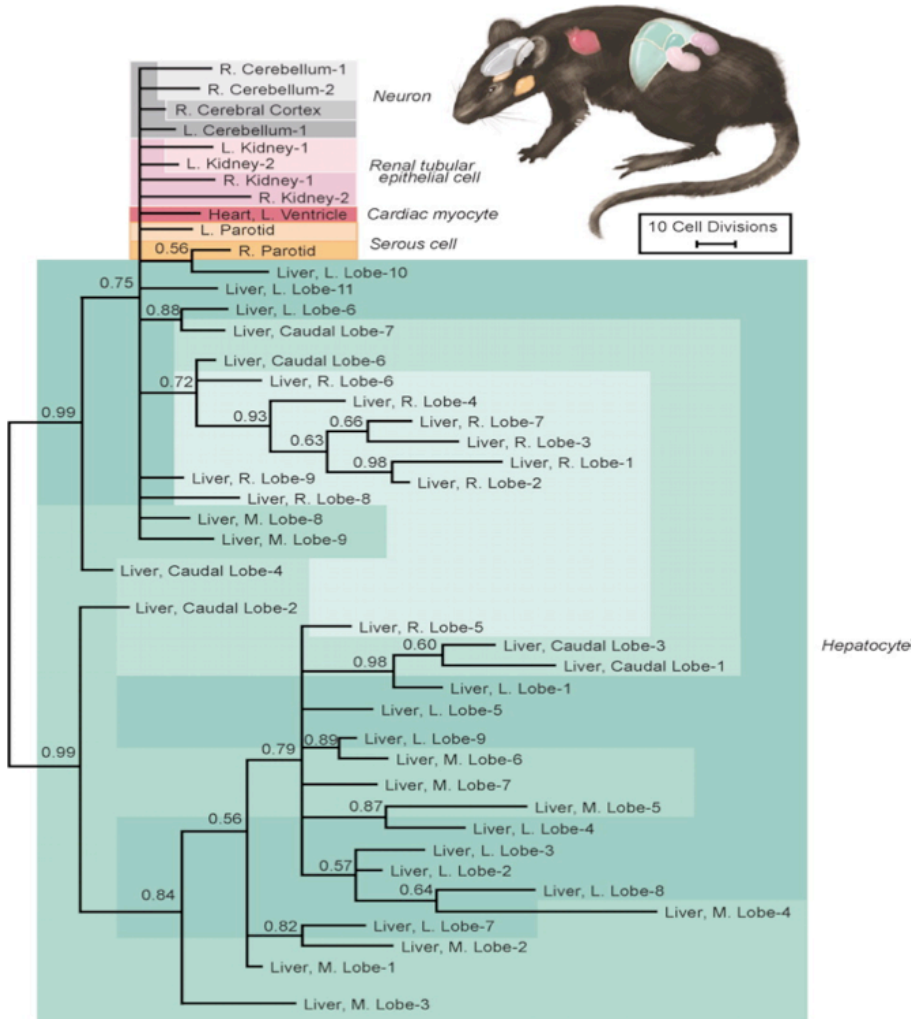
lobSTR Performance



- LobSTR processes reads between 2.5 and 1000 times faster than mainstream aligners.
- Only BLAT detected more STR variations than lobSTR.
- LobSTR accurately detects pathogenic trinucleotide expansions that are normally discarded by mainstream aligners.
  - BWA only reports normal allele.
  - LobSTR identifies both alleles present at the simulated loci.

Why should we care about microsatellites?

- Polymorphism and mutation rate variation
- Disease
  - Huntington's Disease
  - Fragile X syndrome
  - Friedrich's ataxia
- Mutations as lineage
  - Organogenesis/embryonic development
  - Tumor development



Phylogenetic fate mapping
Salipante (2006) *PNAS*. doi: 10.1073/pnas.0601265103

Combined DNA Index System (CODIS)

The Combined DNA Index System, or CODIS, blends forensic science and computer technology into a tool for linking violent crimes. It enables federal, state, and local forensic laboratories to exchange and compare DNA profiles electronically, thereby linking serial violent crimes to each other and to known offenders. Using the National DNA Index System of CODIS, the National Missing Persons DNA Database also helps identify missing and unidentified individuals.

Overview

CODIS generates investigative leads in cases where biological evidence is recovered from the crime scene. Matches made among profiles in the Forensic Index can link crime scenes together, possibly identifying serial offenders. Based upon a match, police from multiple jurisdictions can coordinate their respective investigations and share the leads they developed independently. Matches made between the Forensic and Offender Indexes provide investigators with the identity of suspected perpetrators. Since names and other personally identifiable information are not stored at NDIS, qualified DNA analysts in the laboratories sharing matching profiles contact each other to confirm the candidate match.

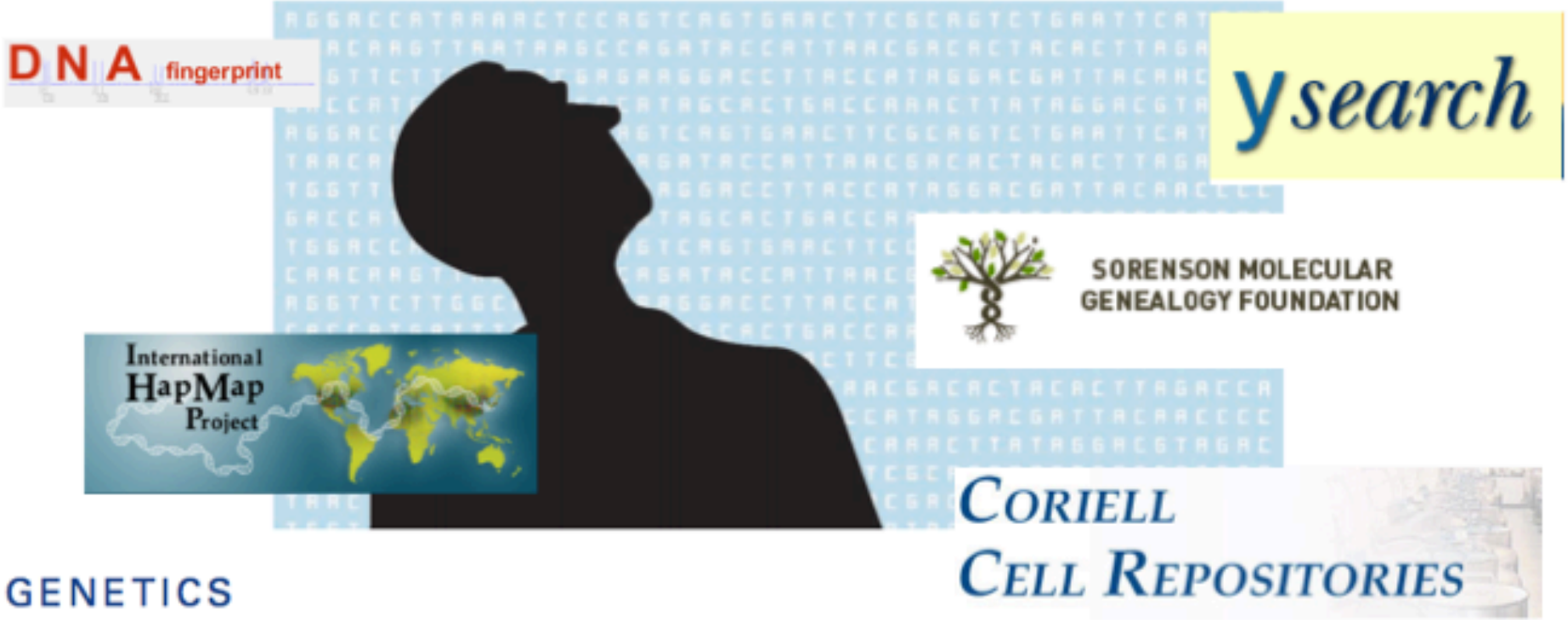
History

The FBI Laboratory's CODIS began as a pilot software project in 1990, serving 14 state and local laboratories. The DNA Identification Act of 1994 formalized the FBI's authority to establish a National DNA Index System (NDIS) for law enforcement purposes. Today, over 190 public law enforcement laboratories participate in NDIS across the United States. Internationally, more than 90 law enforcement laboratories in over 50 countries use the CODIS software for their own database initiatives.

Mission

The CODIS Unit manages CODIS and NDIS. It is responsible for developing, providing, and supporting the CODIS program to federal, state, and local crime laboratories in the United States and selected international law enforcement crime laboratories to foster the exchange and comparison of forensic DNA evidence from violent crime investigations. The CODIS Unit

Genealogy Databases



GENETICS

Genealogy Databases Enable Naming Of Anonymous DNA Donors

Surname Inference

```
>gnl|ti|1731009826 name:1095462037915 mate_pair:1731009442
AAAAAAAAACTTGTCTTAGGATAAATTCCTGGTAGTGAGATATTGGTTAAAGGAGGATGAAATTTTTATAAATTATATAGCATCTGTATGT
TACTTTCTGAATGCATCCATTAAGTTACAAGGTGCACAAGTAGCAGAAGTAGCACATGGTGCATCCCGTTCTCCACAGGACTGCCAGGAAA
GGCTTATCTTTTCAGAGAGATTTTATTAGTAATTAGACATAAAATGGTGCTTTGAACTACTGTAACTTGCATGTCTATGGTTATTAGATAG
ATACAATTTTCCCACTTTTTAAACTATTTACATTATCTCATATATGATCTTGACAATAGAGTTTTTTCCTTTTGTTTTTTACTGCTCTTGA
ATTTCCTCCTATTTTTTGAAATAACTTACTTAATATAGAGAGTTGTCTTCATGATGGTAGTATTCTTCGCAATATTTCCCAACTTCGTTGT
GTTCCTTATATAGCTCTAGTCTTCTTATGCAGGTATTTTCATTTGTATGCATTTAATTTTTTAAAGAATATCTTCTATTGCTTATAATCTG
TGAAAGTTTGTATTTCTTCATCTCCGCTTAGACCTAGTGACTCCCTCATTTTATTATCTGTGAGCTTTTCTCTCCATTAGGTTAGCACAGC
CTTCTATAATCTGACCTTATCACACCTTACAGAACCATCATTTGACACACTTCAACTCAGCACCTCAGATCTATTTGGCCCAACCCTTTTC
TTCCACACAAACGTACCATGTTCATTTATGCCTTGGGGCATTTGGCCTGTCTTAGATGCTTGCCCATTCTTATTTCCTCTACCCATTCAAA
GTCTACCTATCAGGGGACAGATACCTTGATTACACTCCTCCCCTTGATTTACTTAGTCACATGAACTCCTATACCTCATAAACTGCATAGT
CAATATTATTTATTATTATTTATT
>gnl|ti|1731009827 name:1095462037916 mate_pair:1731009443
TTTGCTTGTAACTGTGTGGCAGACTAGACATTTAGTCACTTACCTTCTGTAAGGGAAGTCACTTAATGTAGATGCTCAAGGAATCCCACGT
CCAAATACGGTCATTGCTGAATTTGATGGGAGAAAAAAAGGACTGTATTTTTGCTGGCTAATGCGGCATTTCTGGATGTTCATTAAGCTGT
TAAAAAATGTGAAATGATGACTCACTGTGAAGCCTTCCCAACCCCCTTCCCTGCTCCAATTTGCCTGCGGCTTGAAGGAATGTATACTGAC
TTGAGTTAAAGAAGAAACTTGTGGTTTCCATTATAGTCTGTAAACATCGATTGCTGAAAGCCTCAAGTGTGCTATCATCTCCTGCCACACA
CATGCTTTATATACAGTTCATCTTGGAAAACAGATTGGGAGAGCACCGTGTTTACAACTCCAGAGTAACCCAACTAATGGGGGGCTGGTCT
GCCTGGGGGTCTCGAGACCATCACTGTTAACAGTGGAGGTCCCCTCTGATGTCTTTTAAGCCAAAGTATCAAGAGAGAAGCCAAAGACAAG
CCTCATTAAAGAAAGATCTTCCTGGTACCCACAGTGGTTCTTTGCTGCTTTATCATCAGCAAAAGAGACATAAAAAAATCTTTGCTTCTCT
TTTAGGTTGTATCCTTAGGCCTTATTCTTGATGTCATTGAAAAGTTGGTTTATCTATTATTTCCAAGGCCACTGGCAATGCTATCATTTTT
GATTAATGATTAGTTTATGGCTAAGTAGTATTCCATGGTTAGGTGATGGTTATATTGAAAGCCCAGATTTCAGCACTATGTAATATAGCCA
TGTAACAAATTGCATTTGTACCCCCTAAATCTATTTTTAAAAGTTTAATTTAATTATGCAGAGTGGCAGCAGATGCAAACCATCCAGGGAT
TATTCTGGTAGTGCCTGATGGATTTAATAATACCA
>gnl|ti|1731009828 name:1095462037917 mate_pair:1731009444
CCTAAGACATTTTGAATGTATACTATACATGACTTGGAATCTGACTAAAATTTAAGATAGAAAAAATGAGCCAAAAATGCTTAAAGCTTAG
CATTACTAAAATGTGGTACAGCCATTAATTAAAACTTGTATTAGAATTTGGGGGGATGGCAAAACAAATTTTATGTTGGACATGTTGATTT
GGGCAACAGTAGAAATAAAAGTGGGTTATTGACAATTTCAGCATGATGTCAGGTCAGGGCTGGAGTGTCATAAACTTTTGAGTCCCTGAAA
ATAAAGAAATCCCTAATGGTAGAGCATATGGAAAAAGAAAGTTTGGCAAAAGTTTTCAAGCGTTAAAACTTTGATTTTTTTTTCTCCAAGC
AATTGGAGTATAACAAAGCCCCAAACGTAGAAATGTATGTGTAAGTATATGTATGTGTGTGTATATATATGTGTGTGTGTGTATGTATA
TATGCGTGTGTGTGTGTGTGTGTGTTAATGACAGAAATGTTAGAGGCATAACAGATAAAAGAAAAGTATAGTTTTAAGAGGTTTC
TATGCAGGCAAGAAAGCAAGATAAGCCAAGGACTTTCTATTCCTTAGGCAATTAACAGTGTGTTGGTATATTAAAGATAATAGTAAATATG
ACAGCAGGAATTTGGATAAAGTATATCCATCACTTCATAATTTTAAACCTTTCTCTGCTTTAAGAAACTCTCTGAACTGCCTCTTAGGGTG
CCAGCTGATATAGAAGACTTCTTTTACTGACAGTATTGAC^C
```

Whose sequence reads are these?

Identifying Personal Genomes by Surname Inference

Gymrek et al (2013) *Science*. doi: 10.1126/science.1229566

Step 1. Profile Y-STRs from the individual's genome.

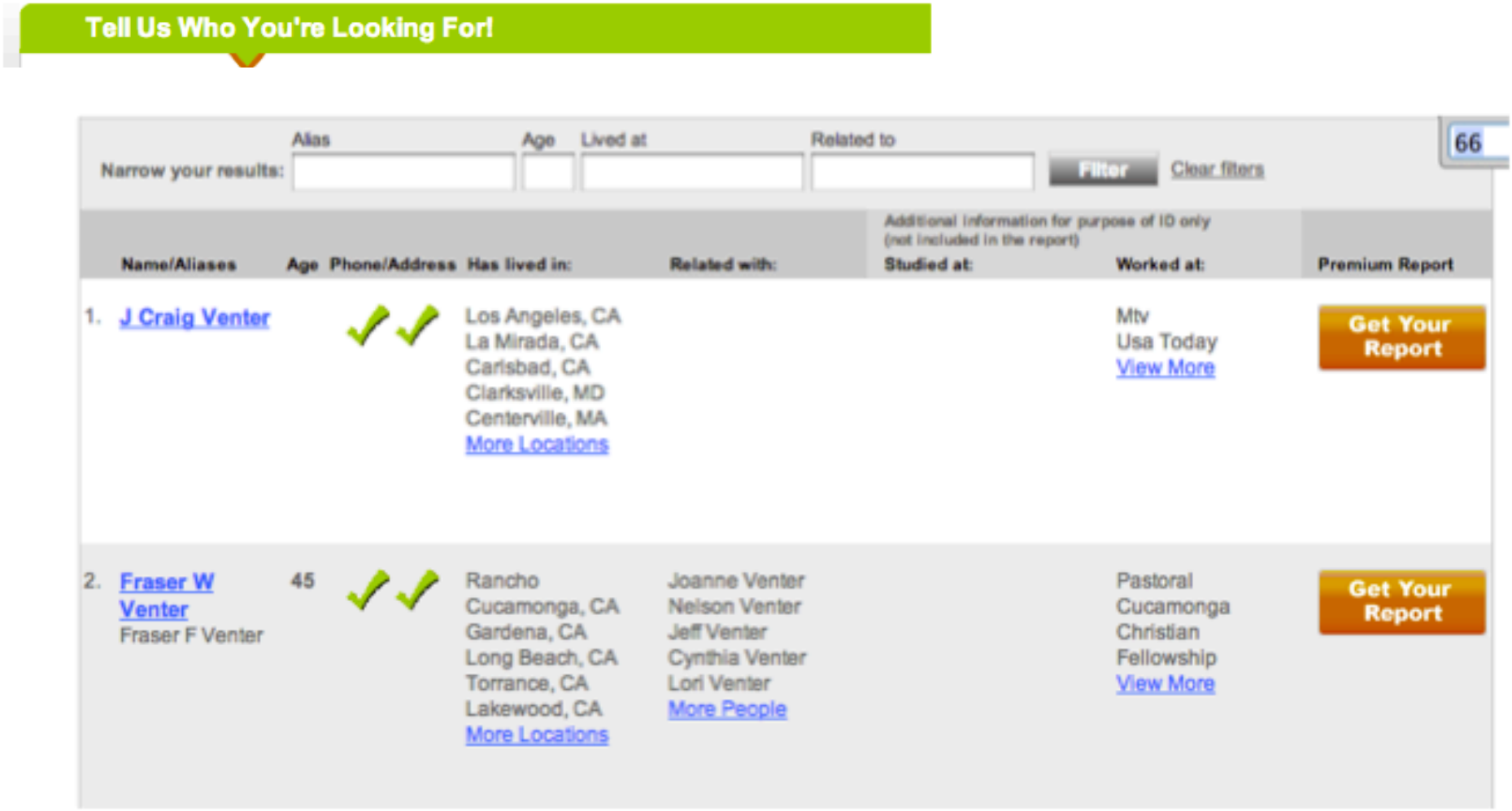
The human reference genome contains 16 copies of "TTTC". Venter has an extra copy of "TTTC", giving him a genotype of "17" at this marker. In a similar way, we can profile all other genealogical STR markers on the Y-chromosome where we know Venter's genome sequence to get the value of a whole panel of these markers.

Step 2. Search for a surname hit in online genetic genealogy databases.

http://www.ysearch.org

Step 3. Search with additional metadata to narrow down the individual.

We enter the search information: Venter, CA, and 66:

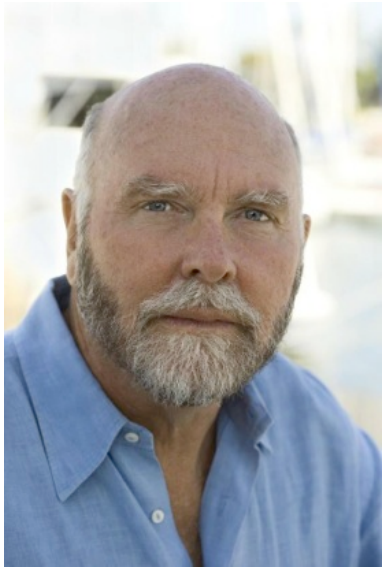


http://www.ussearch.com

Surname Inference

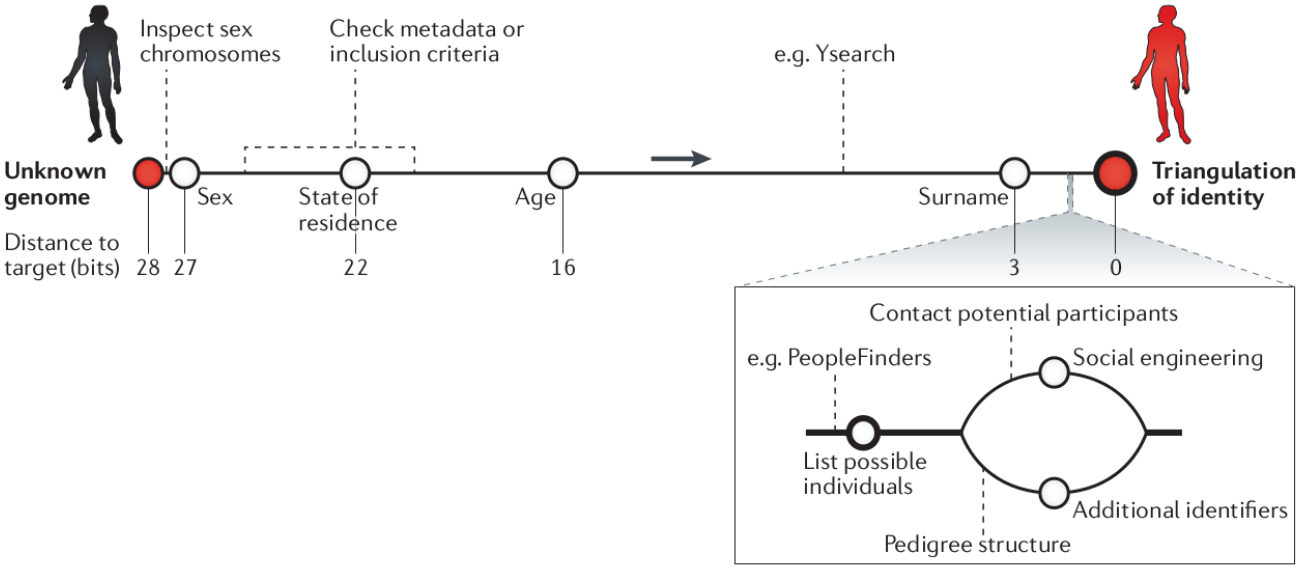
```
>gnl|ti|1731009826 name:1095462037915 mate_pair:1731009442
AAAAAAAAACTTGTCTTAGGATAAATTCCTGGTAGTGAGATATTGGTTAAAGGAGGATGAAATTTTTATAAATTATATAGCATCTGTATGT
TACTTTCTGAATGCATCCATTAAGTTACAAGGTGCACAAGTAGCAGAAGTAGCACATGGTGCATCCCGTTCTCCACAGGACTGCCAGGAAA
GGCTTATCTTTTCAGAGAGATTTTATTAGTAATTAGACATAAAATGGTGCTTTGAACTACTGTAACTTGCATGTCTATGGTTATTAGATAG
ATACAATTTTCCCACTTTTTAAACTATTTACATTATCTCATATATGATCTTGACAATAGAGTTTTTTCCTTTTGTTTTTTACTGCTCTTGA
ATTTCCTCCTATTTTTTGAAATAACTTACTTAATATAGAGAGTTGTCTTCATGATGGTAGTATTCTTCGCAATATTTCCCAACTTCGTTGT
GTTCCTTATATAGCTCTAGTCTTCTTATGCAGGTATTTTCATTTGTATGCATTTAATTTTTTAAAGAATATCTTCTATTGCTTATAATCTG
TGAAAGTTTGTATTTCTTCATCTCCGCTTAGACCTAGTGACTCCCTCATTTTATTATCTGTGAGCTTTTCTCTCCATTAGGTTAGCACAGC
CTTCTATAATCTGACCTTATCACACCTTACAGAACCATCATTTGACACACTTCAACTCAGCACCTCAGATCTATTTGGCCCAACCCTTTTC
TTCCACACAAACGTACCATGTTCATTTATGCCTTGGGGCATTTGGCCTGTCTTAGATGCTTGCCCATTCTTATTTCCTCTACCCATTCAAA
GTCTACCTATCAGGGGACAGATACCTTGATTACACTCCTCCCCTTGATTTACTTAGTCACATGAACTCCTATACCTCATAAACTGCATAGT
CAATATTATTTATTATTATTTATT
>gnl|ti|1731009827 name:1095462037916 mate_pair:1731009443
TTTGCTTGTAACTGTGTGGCAGACTAGACATTTAGTCACTTACCTTCTGTAAGGGAAGTCACTTAATGTAGATGCTCAAGGAATCCCACGT
CCAAATACGGTCATTGCTGAATTTGATGGGAGAAAAAAAGGACTGTATTTTTGCTGGCTAATGCGGCATTTCTGGATGTTCATTAAGCTGT
TAAAAAATGTGAAATGATGACTCACTGTGAAGCCTTCCCAACCCCCTTCCCTGCTCCAATTTGCCTGCGGCTTGAAGGAATGTATACTGAC
TTGAGTTAAAGAAGAAACTTGTGGTTTCCATTATAGTCTGTAAACATCGATTGCTGAAAGCCTCAAGTGTGCTATCATCTCCTGCCACACA
CATGCTTTATATACAGTTCATCTTGGAAAACAGATTGGGAGAGCACCGTGTTTACAACTCCAGAGTAACCCAACTAATGGGGGGCTGGTCT
GCCTGGGGGTCTCGAGACCATCACTGTTAACAGTGGAGGTCCCCTCTGATGTCTTTTAAGCCAAAGTATCAAGAGAGAAGCCAAAGACAAG
CCTCATTAAAGAAAGATCTTCCTGGTACCCACAGTGGTTCTTTGCTGCTTTATCATCAGCAAAAGAGACATAAAAAAATCTTTGCTTCTCT
TTTAGGTTGTATCCTTAGGCCTTATTCTTGATGTCATTGAAAAGTTGGTTTATCTATTATTTCCAAGGCCACTGGCAATGCTATCATTTTT
GATTAATGATTAGTTTATGGCTAAGTAGTATTCCATGGTTAGGTGATGGTTATATTGAAAGCCCAGATTTCAGCACTATGTAATATAGCCA
TGTAACAAATTGCATTTGTACCCCCTAAATCTATTTTTAAAAGTTTAATTTAATTATGCAGAGTGGCAGCAGATGCAAACCATCCAGGGAT
TATTCTGGTAGTGCCTGATGGATTTAATAATACCA
>gnl|ti|1731009828 name:1095462037917 mate_pair:1731009444
CCTAAGACATTTTGAATGTATACTATACATGACTTGGAATCTGACTAAAATTTAAGATAGAAAAAATGAGCCAAAAATGCTTAAAGCTTAG
CATTACTAAAATGTGGTACAGCCATTAATTAAAACTTGTATTAGAATTTGGGGGGATGGCAAAACAAATTTTATGTTGGACATGTTGATTT
GGGCAACAGTAGAAATAAAAGTGGGTTATTGACAATTTCAGCATGATGTCAGGTCAGGGCTGGAGTGTCATAAACTTTTGAGTCCCTGAAA
ATAAAGAAATCCCTAATGGTAGAGCATATGGAAAAAGAAAGTTTGGCAAAAGTTTTCAAGCGTTAAAACTTTGATTTTTTTTTCTCCAAGC
AATTGGAGTATAACAAAGCCCCAAACGTAGAAATGTATGTGTAAGTATATGTATGTGTGTGTATATATATGTGTGTGTGTGTATGTATA
TATGCGTGTGTGTGTGTGTGTGTGTGTTAATGACAGAAATGTTAGAGGCATAACAGATAAAAGAAAAGTATAGTTTTAAGAGGTTTC
TATGCAGGCAAGAAAGCAAGATAAGCCAAGGACTTTCTATTCCTTAGGCAATTAACAGTGTGTTGGTATATTAAAGATAATAGTAAATATG
ACAGCAGGAATTTGGATAAAGTATATCCATCACTTCATAATTTTAAACCTTTCTCTGCTTTAAGAAACTCTCTGAACTGCCTCTTAGGGTG
CCAGCTGATATAGAAGACTTCTTTTACTGACAGTATTGAC^C
```

It's Craig Venter!



Identifying Personal Genomes by Surname Inference
Gymrek et al (2013) *Science*. doi: 10.1126/science.1229566

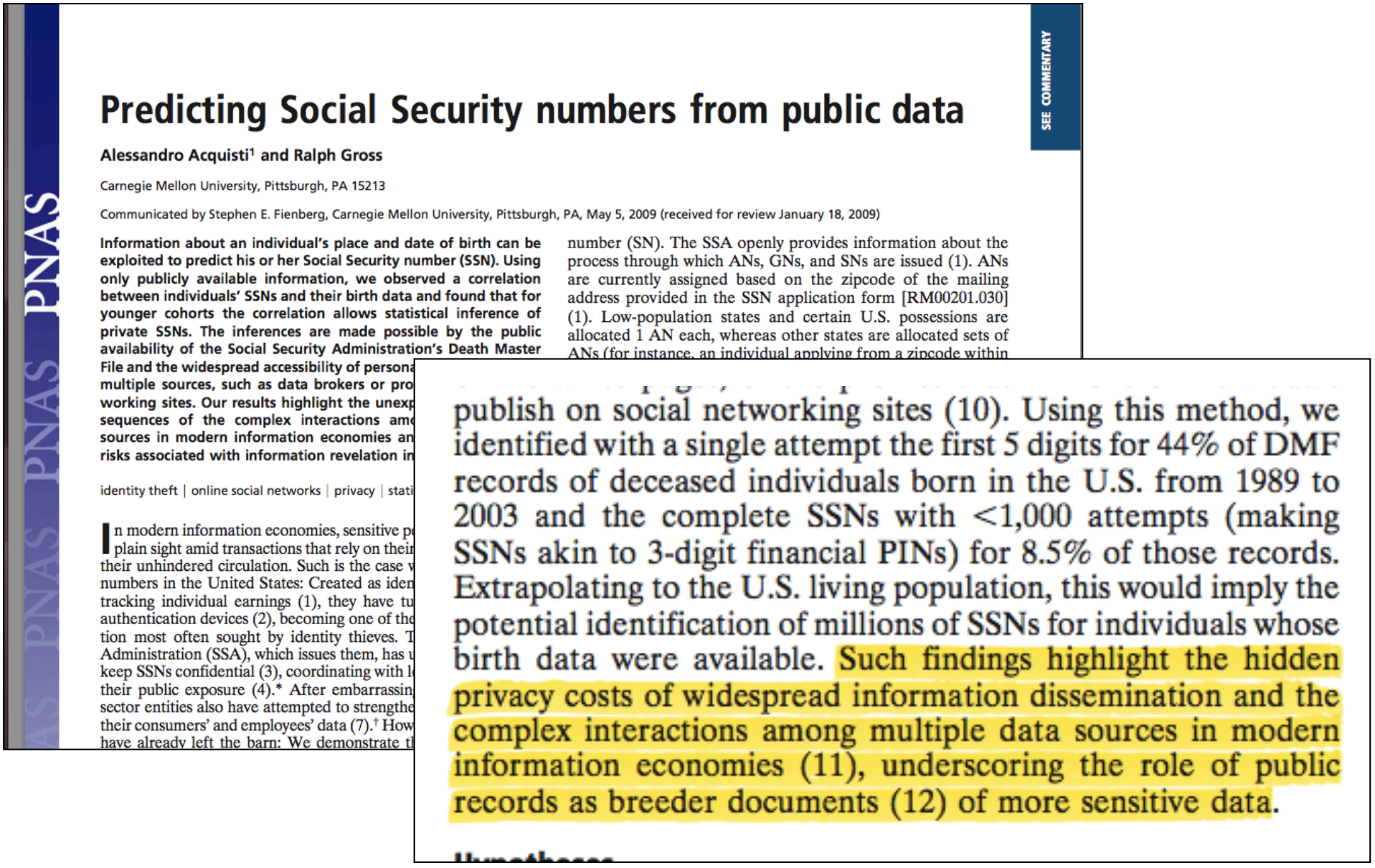
Possible route for identity tracing



- *US population: ~313.9 million individuals*
- *$\log_2$ 313,900,000 = 28.226 bits*
- *Sex ~ 1.0 information bits*
- *$\log_2$ 156,950,000 = 27.226 bits*

- Tracing attacks combine metadata and surname inference to triangulate the identity of an unknown individual.
- With no information, there are roughly 300 million matching individuals in the US, equating to 28.0 bits of entropy.
- Sex reduces entropy by 1 bit, state of residence and age reduces to 16, successful surname inference reduces to ~3 bits.

The risks of big data?



PNAS PNAS PNAS

SEE COMMENTARY

Predicting Social Security numbers from public data

Alessandro Acquisti[1] and Ralph Gross

Carnegie Mellon University, Pittsburgh, PA 15213

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, May 5, 2009 (received for review January 18, 2009)

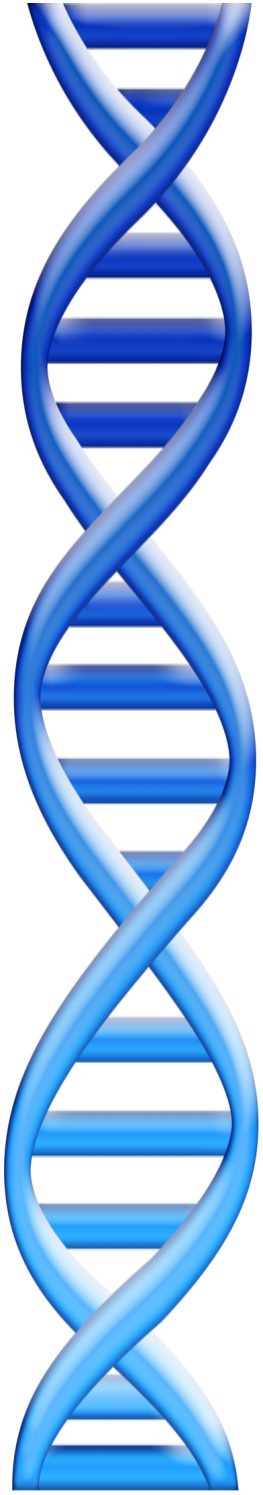
Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration's Death Master File and the widespread accessibility of persona multiple sources, such as data brokers or pro working sites. Our results highlight the unexp sequences of the complex interactions am sources in modern information economies an risks associated with information revelation in

identity theft | online social networks | privacy | stati

In modern information economies, sensitive p plain sight amid transactions that rely on their their unhindered circulation. Such is the case v numbers in the United States: Created as iden tracking individual earnings (1), they have tu authentication devices (2), becoming one of the tion most often sought by identity thieves. T Administration (SSA), which issues them, has u keep SSNs confidential (3), coordinating with l their public exposure (4).* After embarrassin sector entities also have attempted to strengthe their consumers' and employees' data (7).† How have already left the barn: We demonstrate t

number (SN). The SSA openly provides information about the process through which ANs, GNs, and SNs are issued (1). ANs are currently assigned based on the zipcode of the mailing address provided in the SSN application form [RM00201.030] (1). Low-population states and certain U.S. possessions are allocated 1 AN each, whereas other states are allocated sets of ANs (for instance, an individual applying from a zipcode within

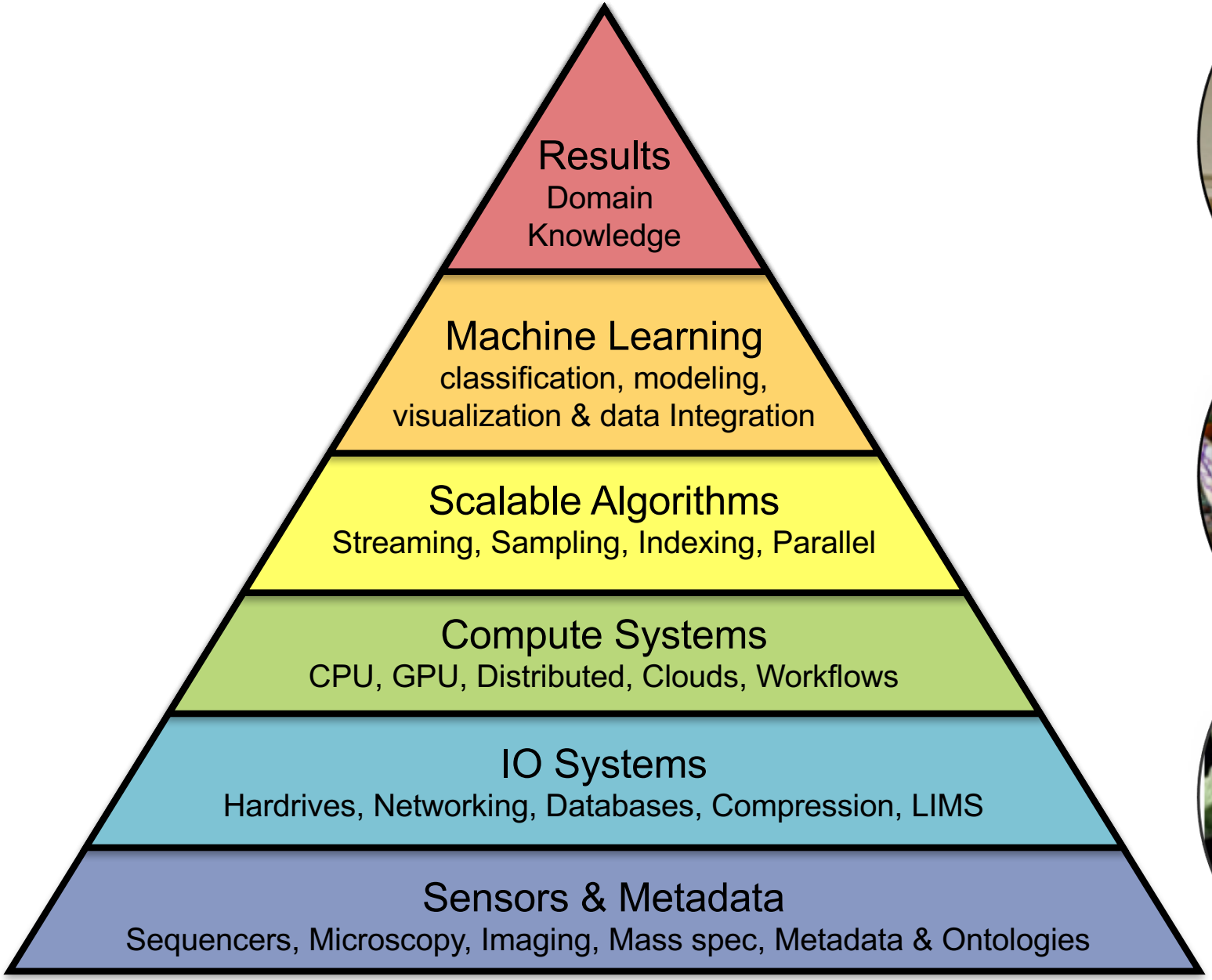
publish on social networking sites (10). Using this method, we identified with a single attempt the first 5 digits for 44% of DMF records of deceased individuals born in the U.S. from 1989 to 2003 and the complete SSNs with <1,000 attempts (making SSNs akin to 3-digit financial PINs) for 8.5% of those records. Extrapolating to the U.S. living population, this would imply the potential identification of millions of SSNs for individuals whose birth data were available. Such findings highlight the hidden privacy costs of widespread information dissemination and the complex interactions among multiple data sources in modern information economies (11), underscoring the role of public records as breeder documents (12) of more sensitive data.

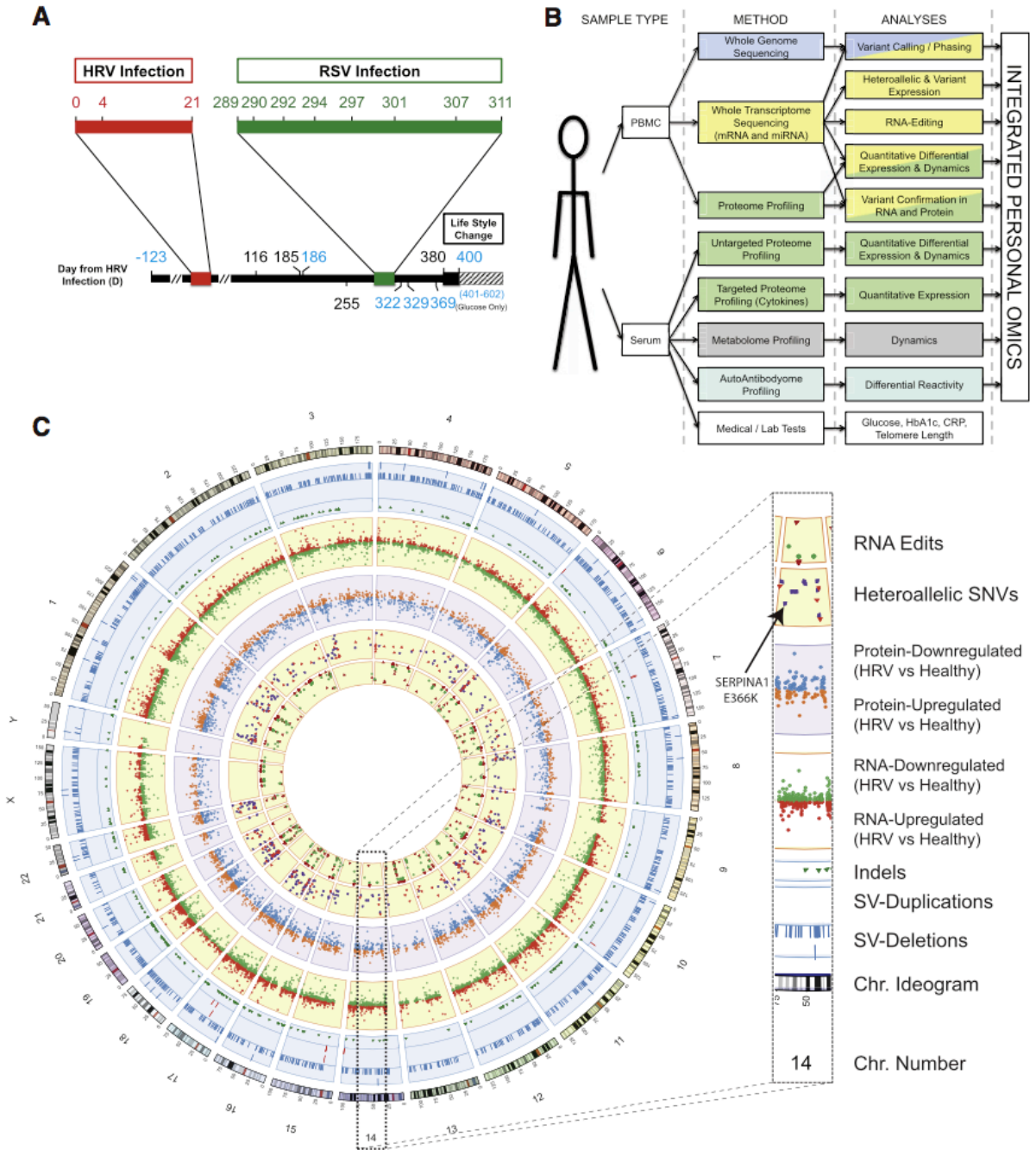


Part 5:

Final Thoughts

Comparative Genomics Technologies





Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes
Chen et al (2012) Cell. DOI 10.1016/j.cell.2012.02.009

Genomic Futures?

The rise of a digital immune system
Schatz & Phillippy (2012) GigaScience 1:4

Next steps

Tremendous power from data aggregation

- Observe the dynamics of biological systems
- Breakthroughs in medicine and biology of profound significance

Be mindful of the risks

- The potential for over-fitting grows with the complexity of the data, statistical significance is a statement about the sample size
- Reproducible workflows, APIs are a must
- Caution is prudent for personal data

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

- Technology will continue to push the frontier
- Feedback loop from the results of one project into experimental design for the next

Topics for next time?

1. Intro
2. Technologies
3. Assembly
4. Whole Genome Alignment
5. Read Mapping (Ben)
6. Read Mapping (Ben)
7. Variant Identification
8. Structural Variations
9. 3rd Gen Sequencing
10. BedTools
11. Annotation
12. RNAseq
13. MethylSeq + ChipSeq
14. ChromHMM
15. Encode
16. Midterm Review
17. MidTerm Discussion
18. Human Evolution
19. Disease Genetics
20. Cancer Genetics
21. Metagenomics
22. Final Thoughts

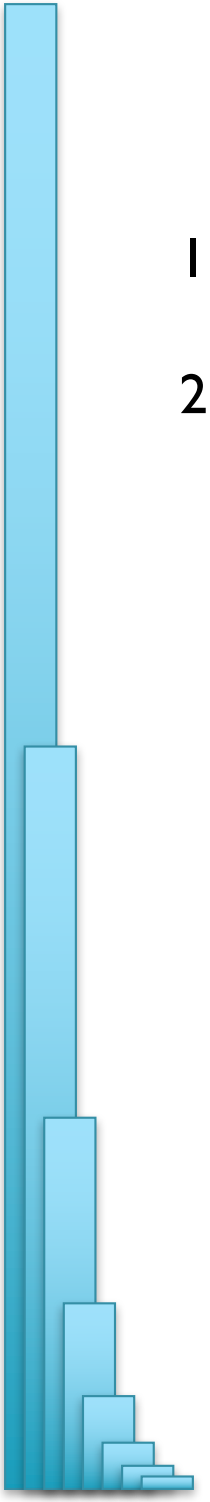
23. Presentations 1

24. Presentations 2

Next Steps

1. Questions on project?

2. Check out the course webpage




Welcome to Applied Comparative Genomics
https://github.com/schatzlab/appliedgenomics

Questions?