Weill
Cornell
Medicine

Linked-Reads and new computational techniques for analyzing metagenomics

Iman Hajirasouliha, PhD

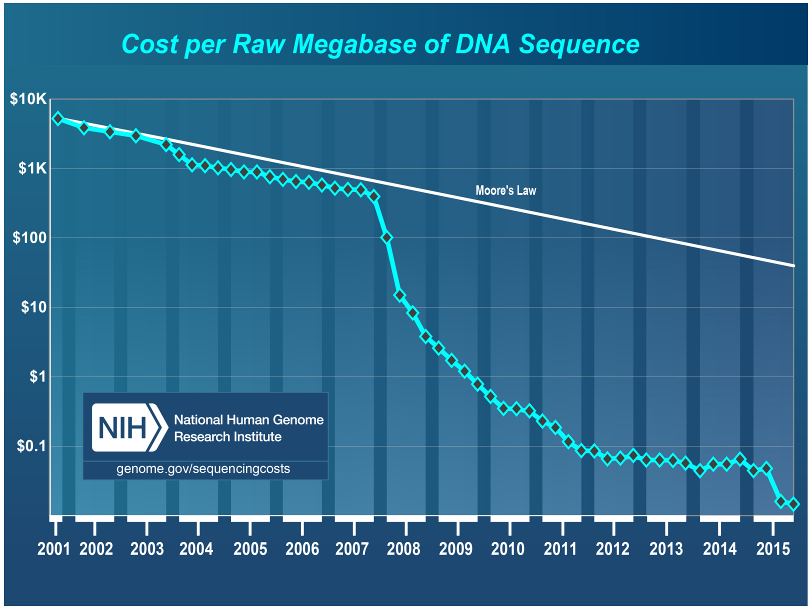
Institute for Computational Biomedicine & Institute for Precision Medicine
Weill Cornell Medical College
Cornell University

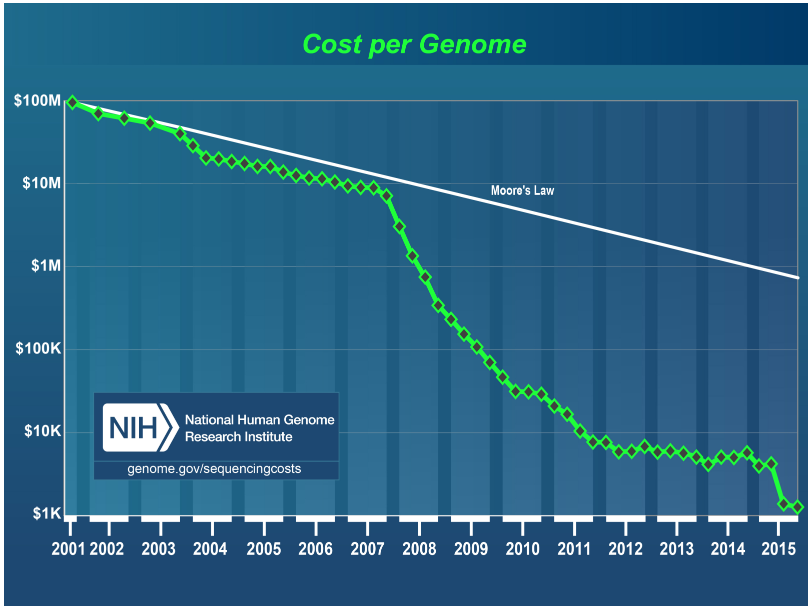
Standard Short-Read Sequencing

BIG amount of sequencing DATA

Terabyte per day for Illumina/HiSeq 2500

Fast and cheap!





1 Million genomes?

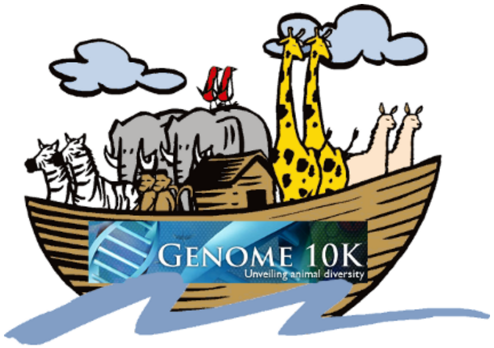
International Cancer Genome Consortium

AGRE Autism Genetic Resource Exchange A SCIENCE PROGRAM OF AUTISM SPEAKS™

Autism Genome 10K

华大基因 BGI

1000 Genomes A Deep Catalog of Human Genetic Variation

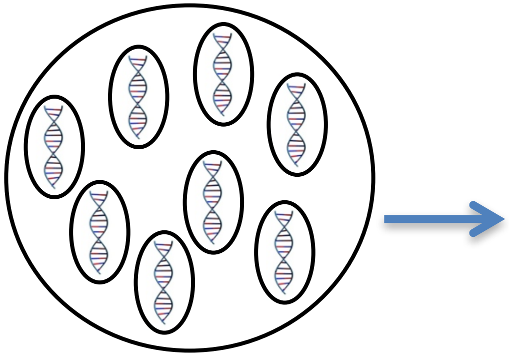


Go • NL GENOMEoftheNETHERLANDS

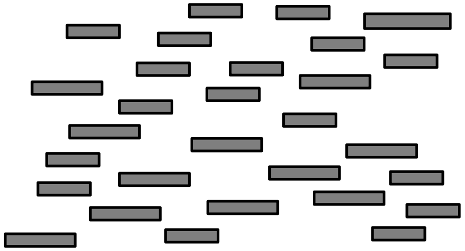
UK 10K RARE GENETIC VARIANTS IN HEALTH AND DISEASE

Genomics england

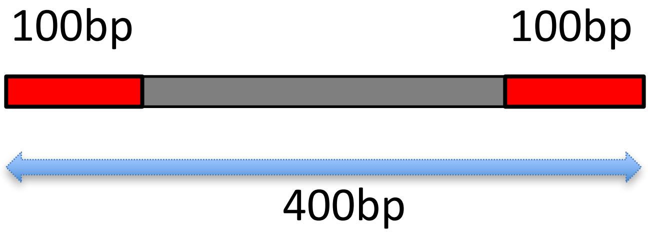
Standard short-read sequencing



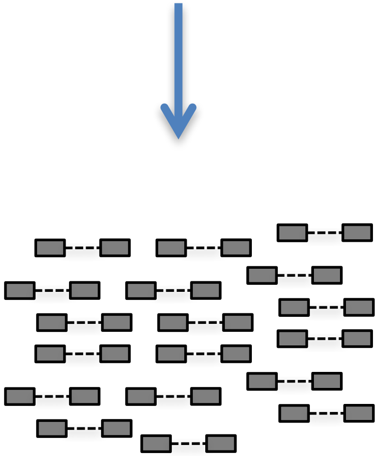
Cells



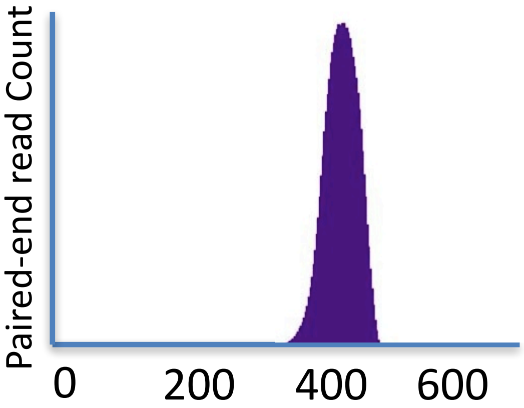
DNA is fragmented



Paired-end sequences of a fragment

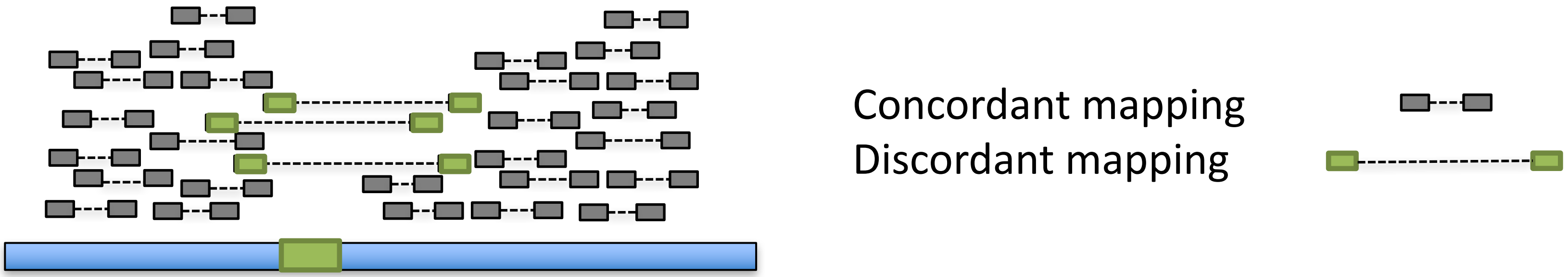


Paired-end reads



Insert size distribution

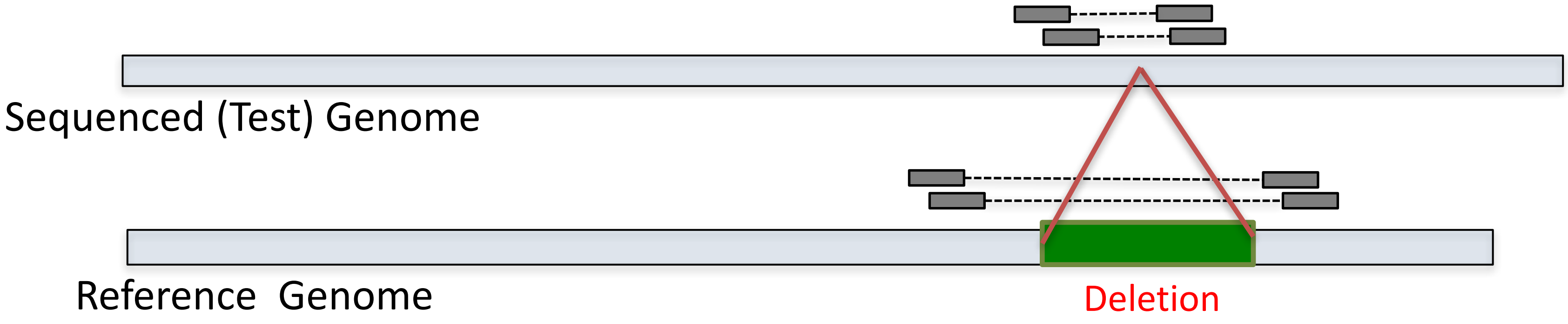
Determining Sequenced Genomes



Concordant mapping

Discordant mapping

Paired-end reads are mapped to the reference



Sequenced (Test) Genome

Reference Genome

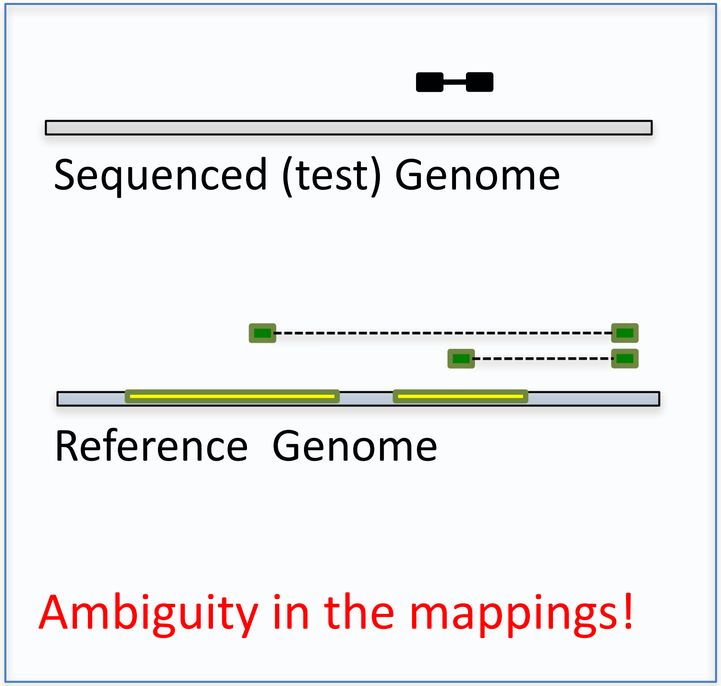
Deletion

Korbel et al. 2007, Kidd et al. 2008, Hormozdiari et al. 2009, Sindi et al. 2009

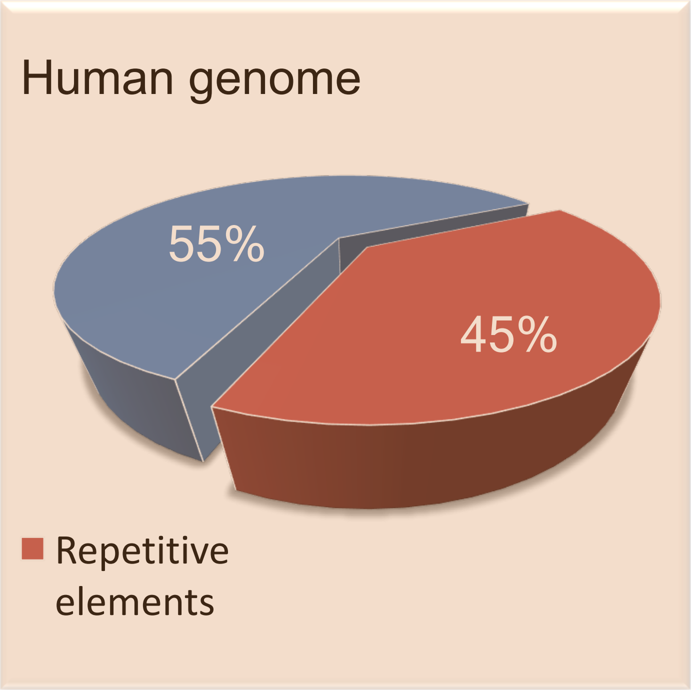
Limitations of NGS technologies

NGS produce "short reads" (e.g. 50bp to 150bp)

The human genome is repetitive!



Ambiguity in the mappings!



Challenges to determine sequenced genomes and metagenomes

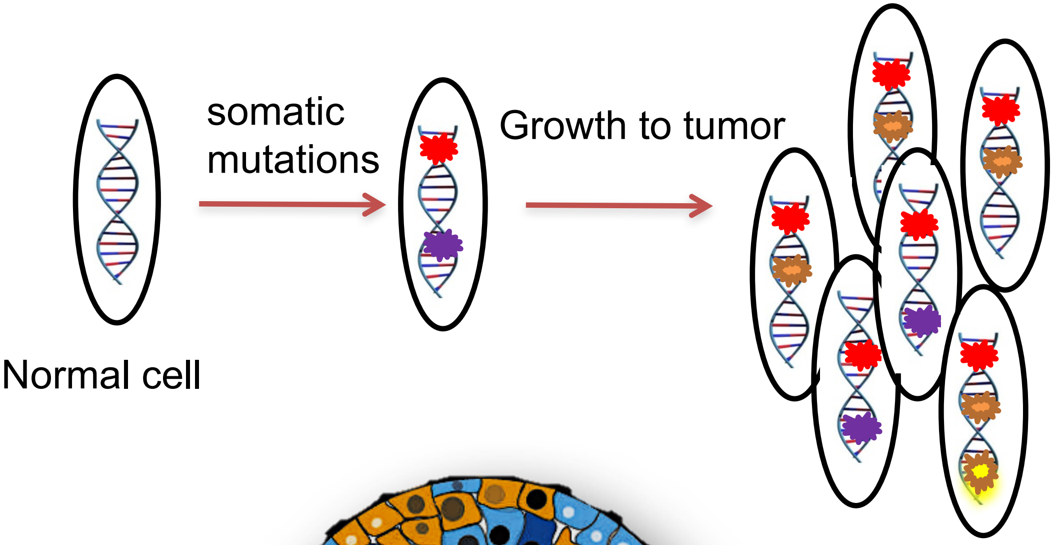
Structural Variations, including those within repetitive regions or complex events.

The reference genome is incomplete or often nonexistent for metagenomes.

In metagenomics, we need to reconstruct the entire mixture.

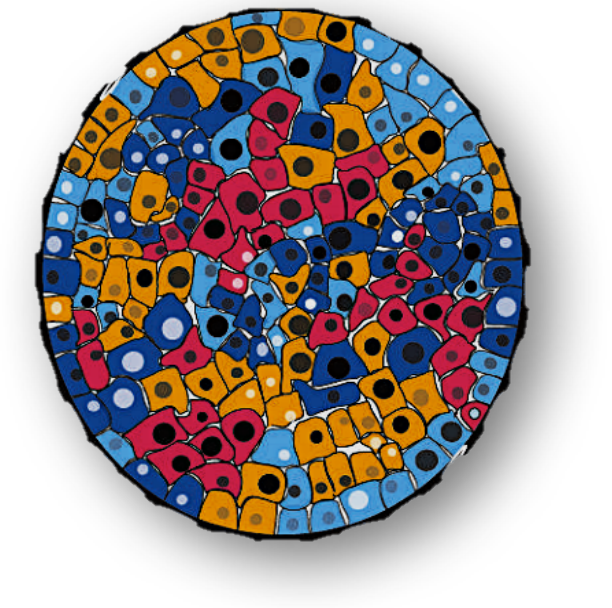
Cancer Genomes

Increased number of somatic mutations.



Mixture of tumor and normal tissue.

Cancer Heterogeneity.



Outline of two genomics projects

Project I: Using Linked-Read technologies for metagenomics

Project II: Phylogeny reconstruction using integration of bulk and single cell sequencing

Beyond short-read sequencing

Long Read:

- Pacbio
- ONT

Expensive, low throughput, high DNA input

But they are real long-reads!

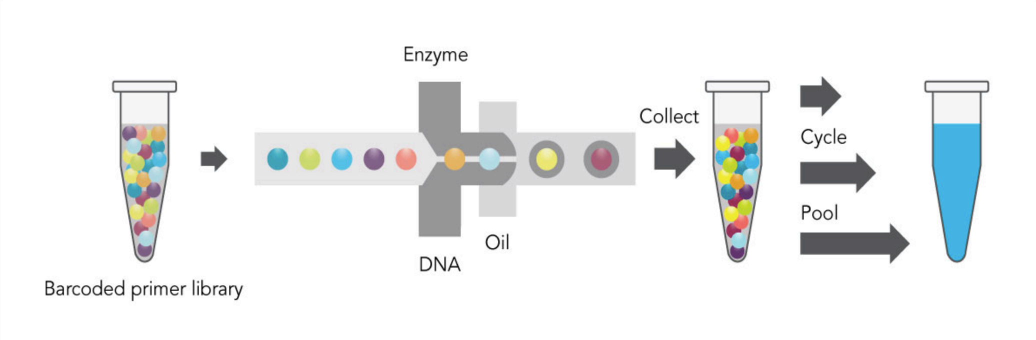
Linked Read (or read cloud technologies):

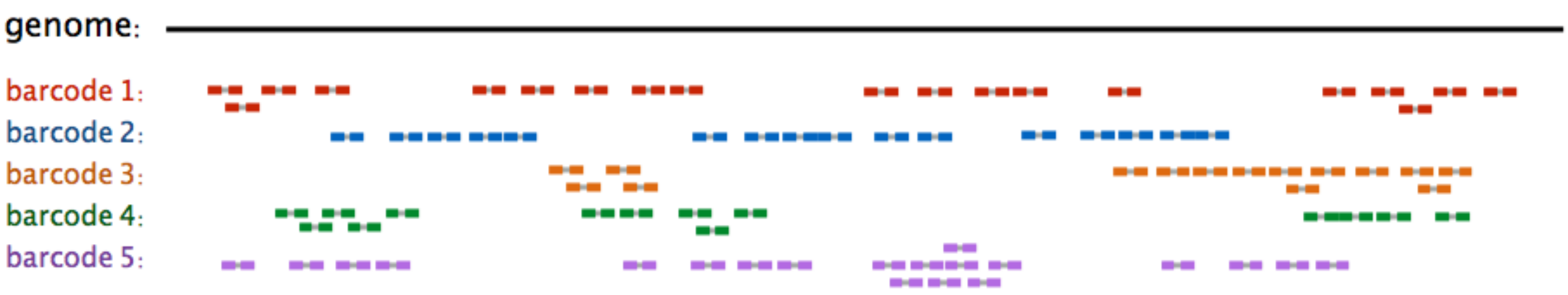
- Moleculo (Illumina Synthetic Long read)
- 10X Genomics

Cheaper, high throughput, low DNA input

But they are fake long-reads!

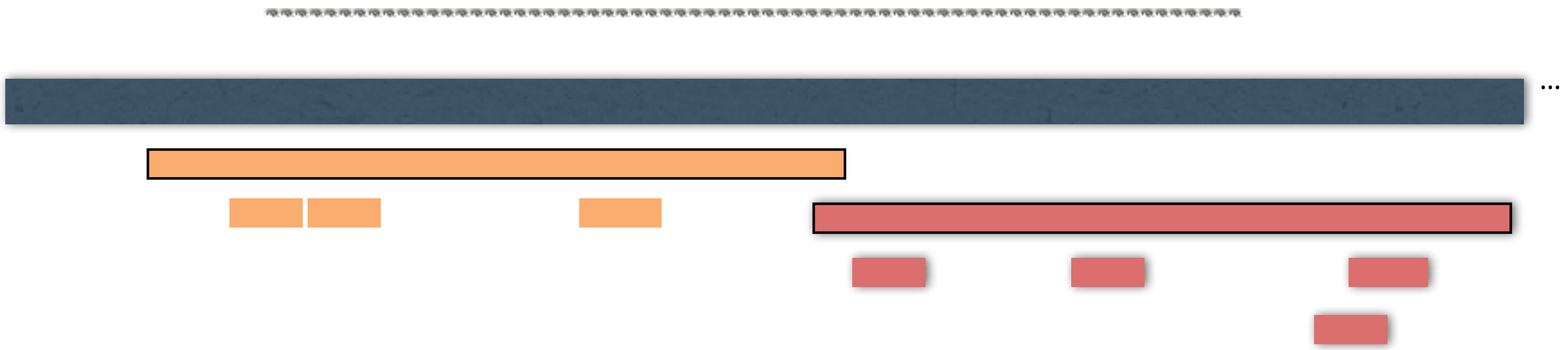
Linked-Read Technologies (e.g. 10X Genomics)





Knowing that the reads “should” form clusters, can we handle ambiguity in read mappings and SV detection better?

10X Genomics model



Long molecules / fragments:

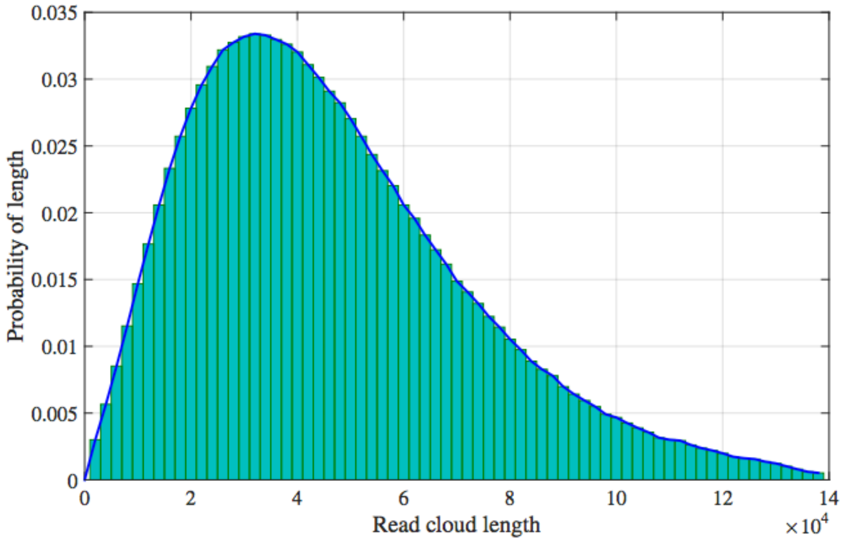
1. coverage C_F
2. mean length: ~10-100Kb

Short reads:

1. coverage C_R
2. length:150 bps

Barcodes:

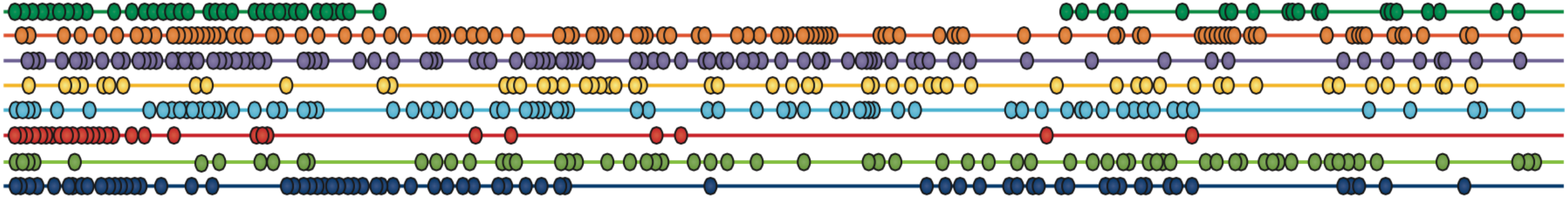
1. # useful barcode ~ 1M
2. Distribution of barcode: Poisson



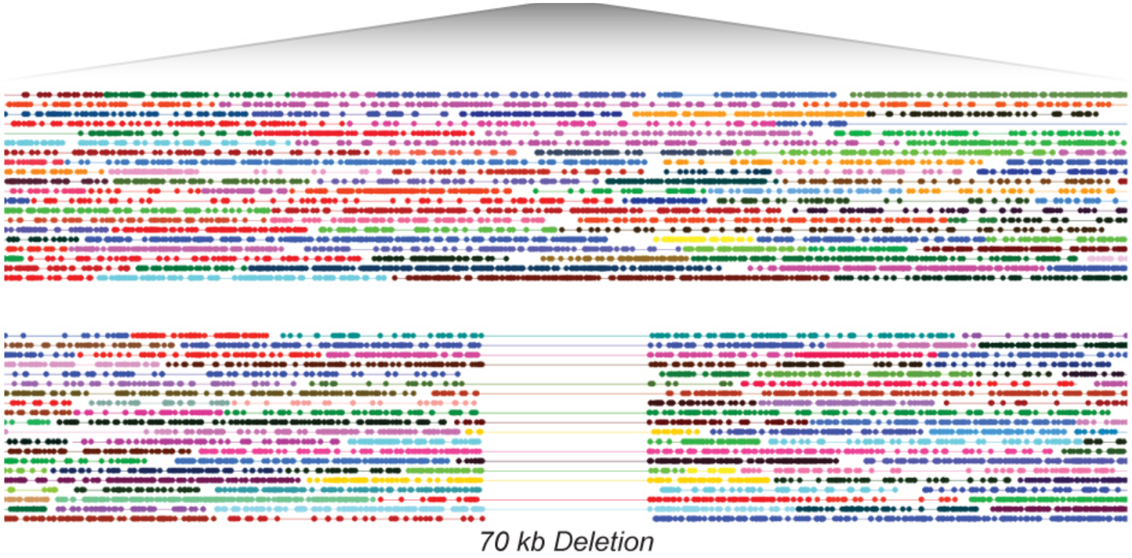
NA12878 Genome

10X Genomics application

Haplotype phasing



Large structural variation calling



a new set of algorithmic challenges

1. Each long fragment of DNA is covered only sparsely by short reads.

2. No information about the relative ordering of reads from the same fragment is preserved.

3. Typically each barcode matches reads from 2-20 long fragments of DNA.

Problem: Linked-read Deconvolution

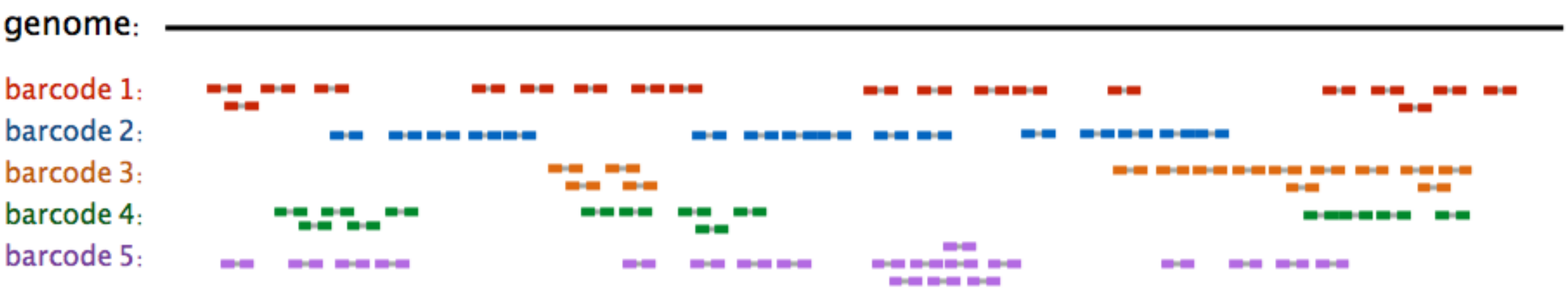
The deconvolution of reads with a single barcode into clusters that correspond to a single long fragment of DNA.

This is one particular issue common to all applications of linked-read technology!

- Any idea?!

Problem: Linked-read Deconvolution

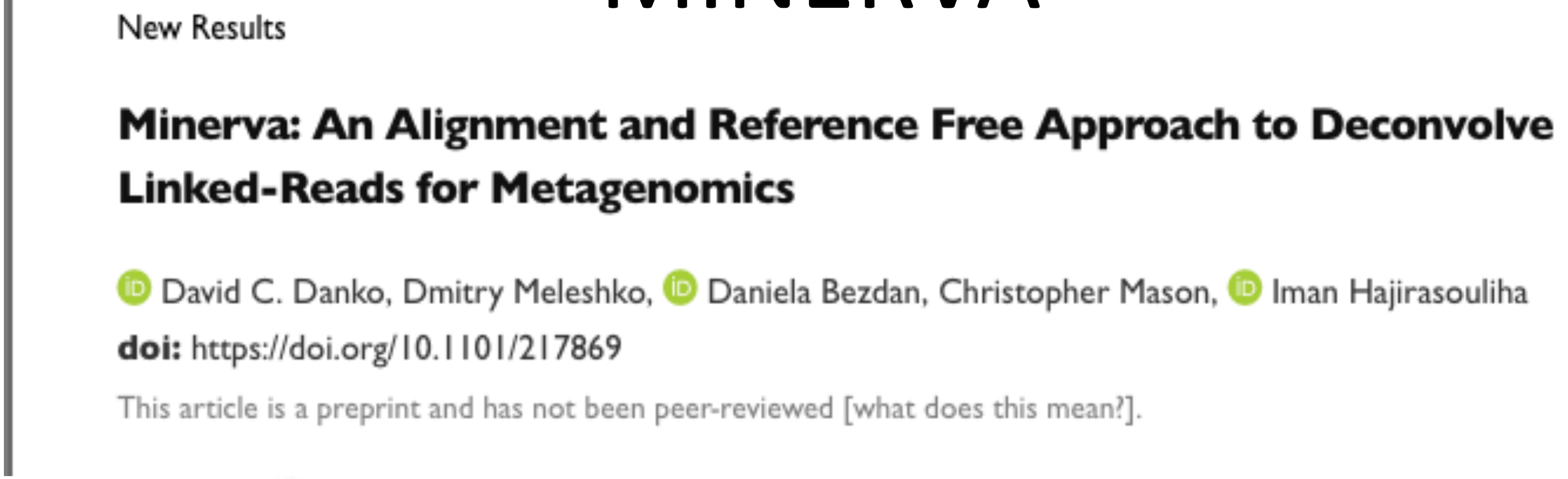
Linked-read Deconvolution when a reference is available



Linked-read Deconvolution when a reference is not available (metagenomics application?)

10X Metagenomics Consortium!

MINERVA



New Results

Minerva: An Alignment and Reference Free Approach to Deconvolve Linked-Reads for Metagenomics

David C. Danko, Dmitry Meleshko, Daniela Bezdan, Christopher Mason, Iman Hajirasouliha

doi: https://doi.org/10.1101/217869

This article is a preprint and has not been peer-reviewed [what does this mean?].

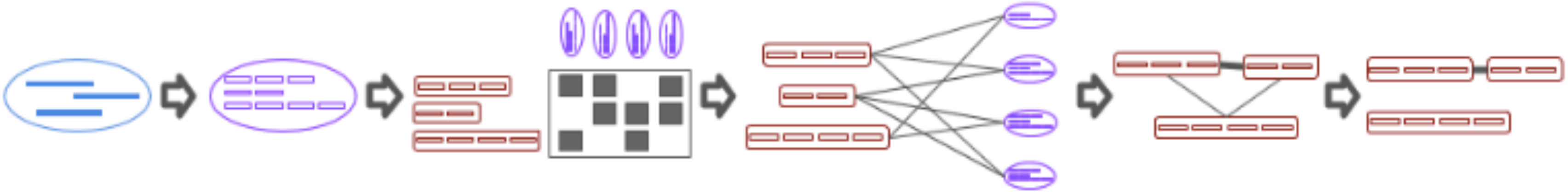
- A new graph-based algorithm for an approximate solution.
- Our approach also further uses some techniques from the field of topic modeling in Natural Language Processing (NLP).

Our graph based method

Key Observation: reads from the same fragment would tend to overlap with similar sets of reads that had different barcodes.

We justified this mathematically, while of course long repeats can be sources of errors.

Our graph based method



1) Fragments are generated
2) Fragments are sequenced and tagged
3) Reads in a given barcode are aligned to other barcodes
4) A bipartite graph between reads and barcodes is constructed
5) A graph between reads that co-occur with barcode is constructed
6) Reads are clustered into groups

primary real data sets from two microbial mock communities

- Dataset 1: 5 bacterial species: *E. coli, Enterobacter cloacae, Micrococcus luteus, Pseudomonas antarctica,* and *Staph. epidermis*.

- Dataset 2: 8 bacterial species and 2 fungi: *Bacillus subtilis, Cryptococcus neoformans, Enterococcus faecalis, E. coli, Lactobacillus fermentum, Listeria monocytogenes, Psuedomonas aeruginosa, Sachharomyces cerevisiae, Salmonella enterica,* and *Staphylococcus aureus*.

- Roughly 1ng of high molecular weight, processed using a 10X Chromium instrument, sequenced on an Illumina Hiseq with 2x150 paired-end reads.

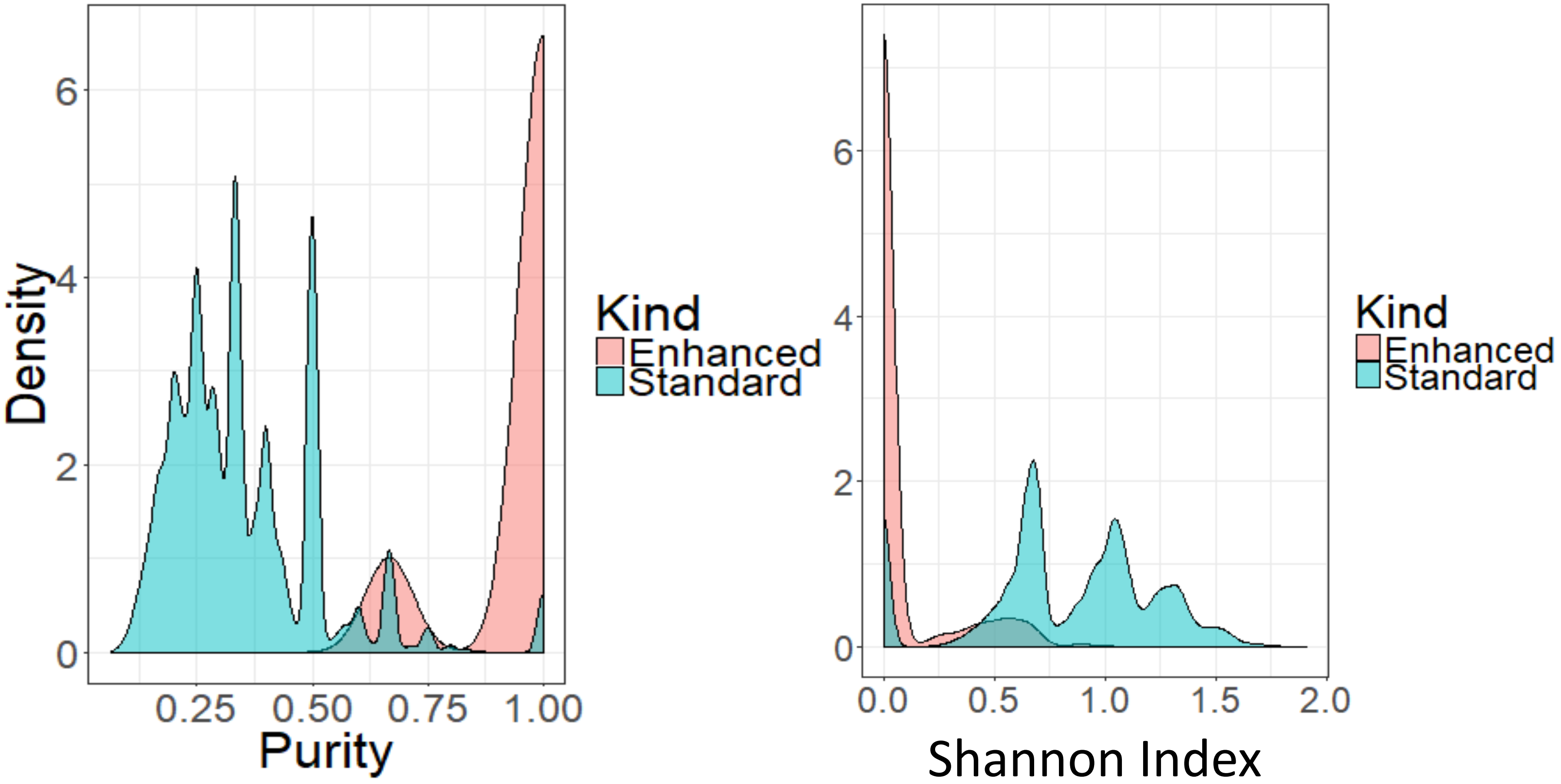
Experimental Results

- Minerva was able to identify subgroups in barcodes that largely corresponded to individual fragments of DNA. i.e. Enhanced Barcodes.

- We quantified this using two measures:
  - Shannon diversity index $H = \sum p_i \log p_i$
  - Purity $P = \max(\vec{p})$

where p_i indicates the proportion of an enhanced barcode that belongs to each fragment.

Minerva deconvolves barcodes



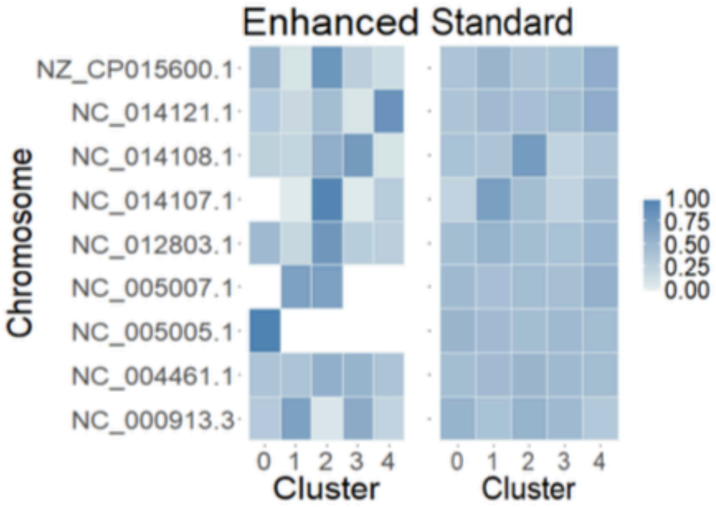
(Left) Purity for enhanced and standard barcodes

(Right) Shannon index in dataset one for enhanced and standard barcodes

Applications of Enhanced Barcodes

1. It is useful to group enhanced barcodes that likely came from the same genome.

   We used a clustering algorithm based on Latent Dirichlet Allocation (LDA), a classic model in NLP.



2. This technique can be used to improve de novo assembly algorithms. (We tested with some unpublished work from collaborators, cloudSpades!)

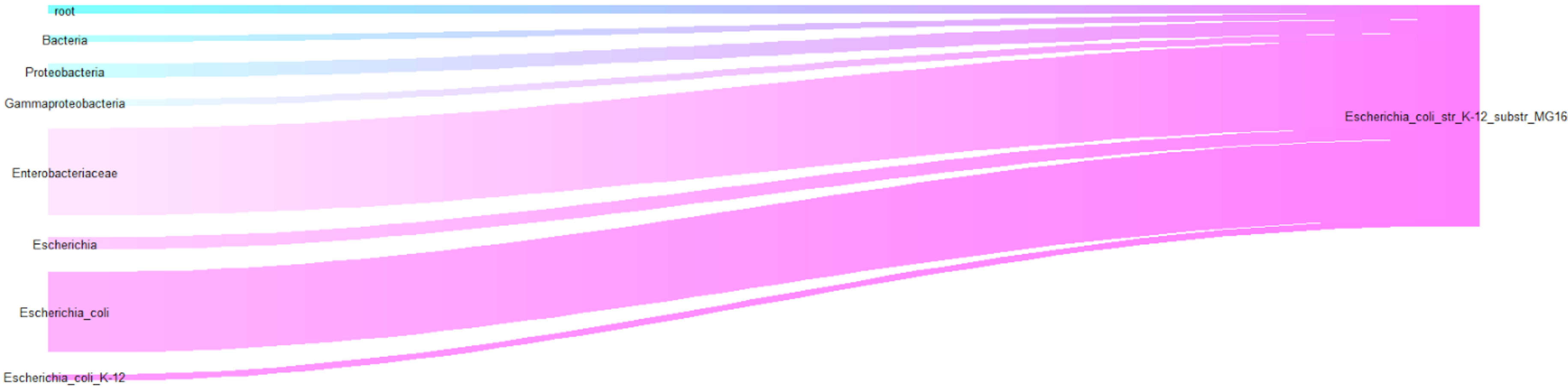
Minerva improves taxonomic assignments

- Minerva can improve the specificity of short read taxonomic assignments obtained from Kraken, a popular tool.

- All reads from the same long-fragment must have the same taxonomic rank!

- We were able to rescue a large number of reads from unspecific taxonomic assignments.

Minerva improves taxonomic assignments



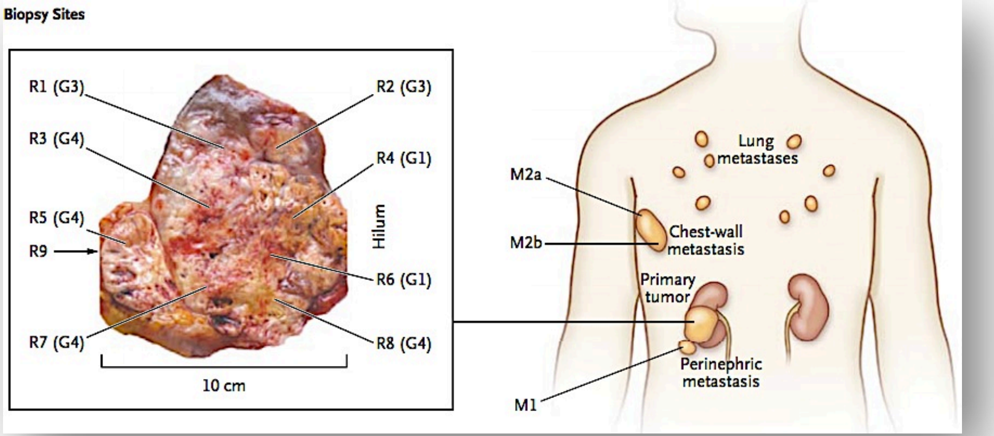
Using enhanced barcodes we can promote the taxonomic assignment of reads. Width of each frond is proportional to the number of reads promoted from a specific rank.

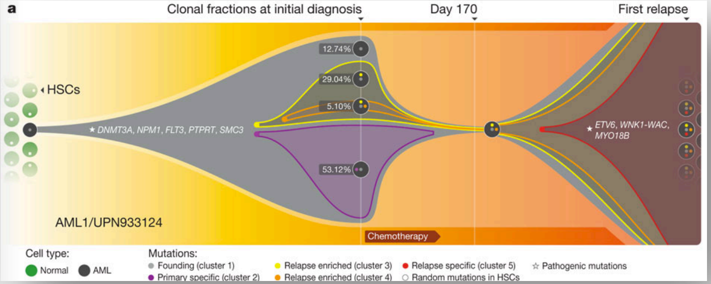
Outline of two on-going projects

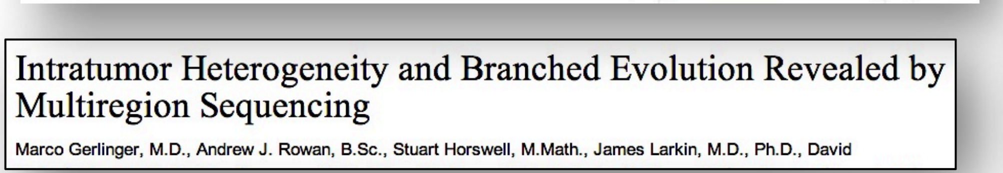
Part I: Using Linked-Read for Metagenomics

Part II: Phylogeny reconstruction using bulk and single cell sequencing

Tumor sequencing

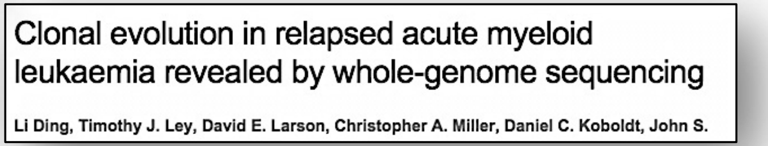






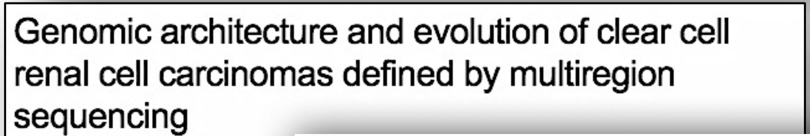
Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing

Marco Gerlinger, M.D., Andrew J. Rowan, B.Sc., Stuart Horswell, M.Math., James Larkin, M.D., Ph.D., David



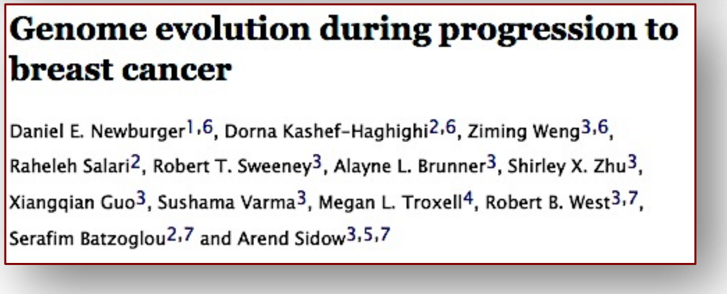
Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing

Li Ding, Timothy J. Ley, David E. Larson, Christopher A. Miller, Daniel C. Koboldt, John S.



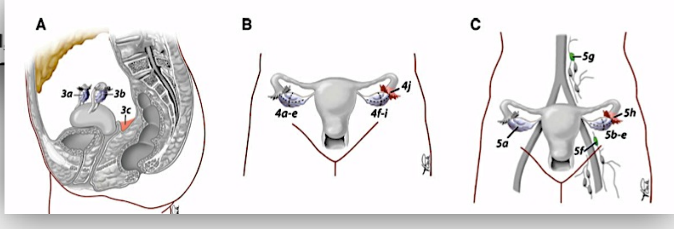
Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing

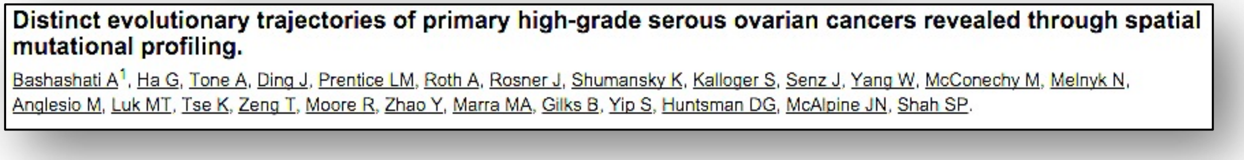
Marco Gerlinger, Stuart Horswell,



Genome evolution during progression to breast cancer

Daniel E. Newburger[1,6], Dorna Kashef-Haghighi[2,6], Ziming Weng[3,6], Raheleh Salari[2], Robert T. Sweeney[3], Alayne L. Brunner[3], Shirley X. Zhu[3], Xiangqian Guo[3], Sushama Varma[3], Megan L. Troxell[4], Robert B. West[3,7], Serafim Batzoglou[2,7] and Arend Sidow[3,5,7]

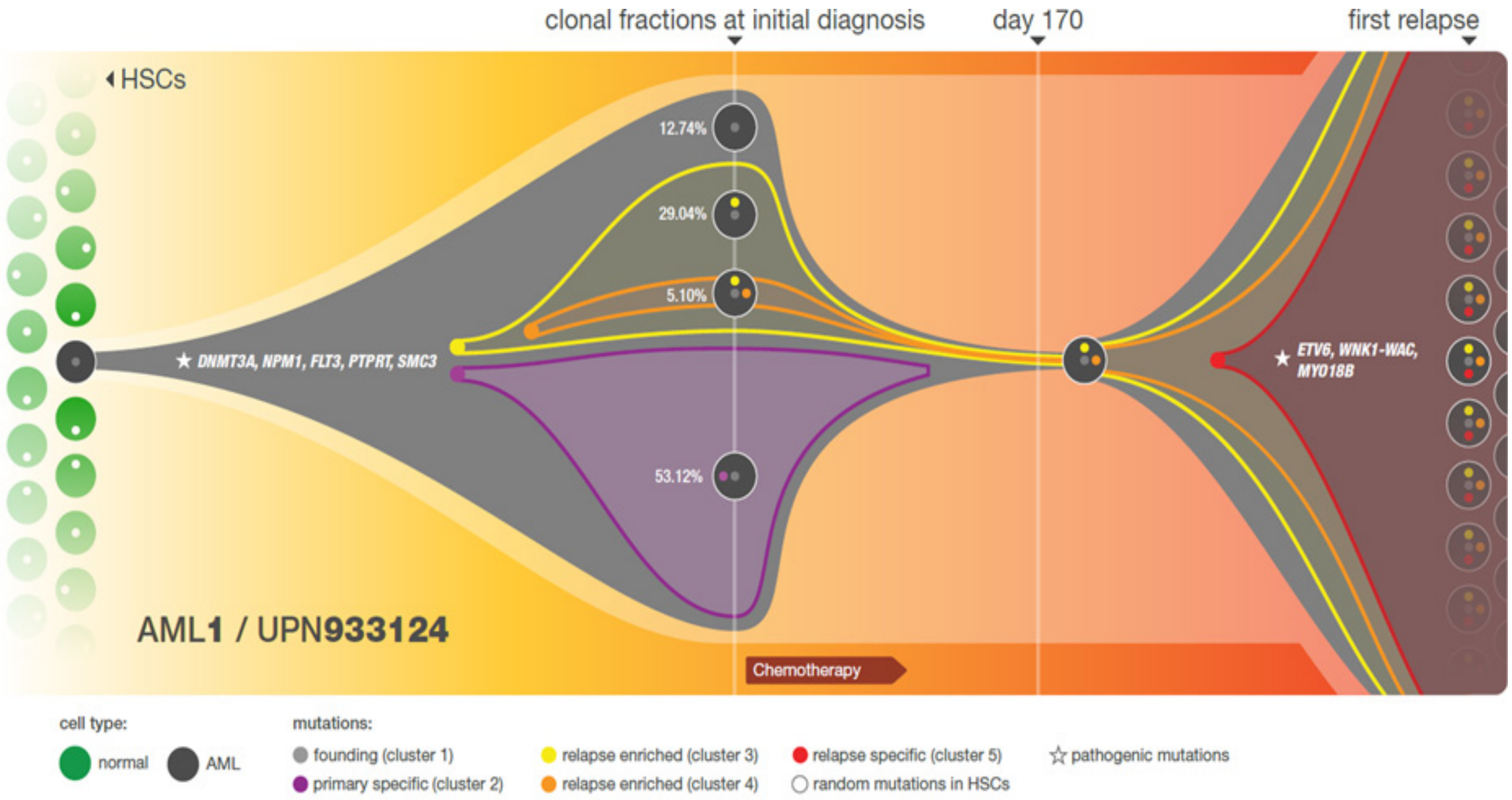




Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling.

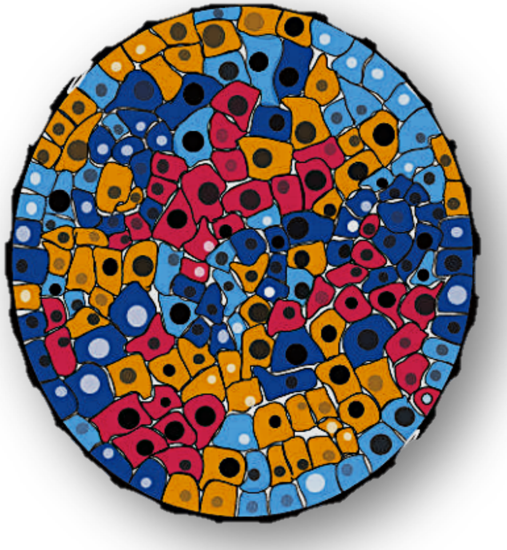
Bashashati A[1], Ha G, Tone A, Ding J, Prentice LM, Roth A, Rosner J, Shumansky K, Kalloger S, Senz J, Yang W, McConechy M, Melnyk N, Anglesio M, Luk MT, Tse K, Zeng T, Moore R, Zhao Y, Marra MA, Gilks B, Yip S, Huntsman DG, McAlpine JN, Shah SP.

Cancer Evolution

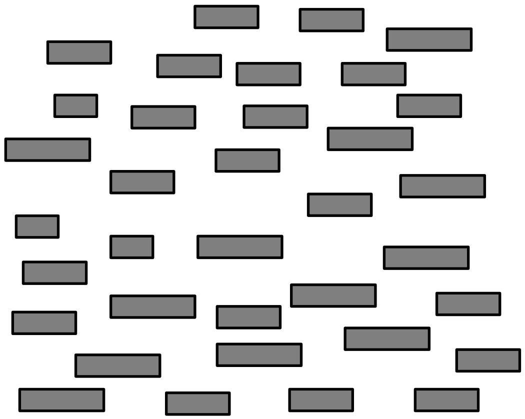


*Ding et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 2012

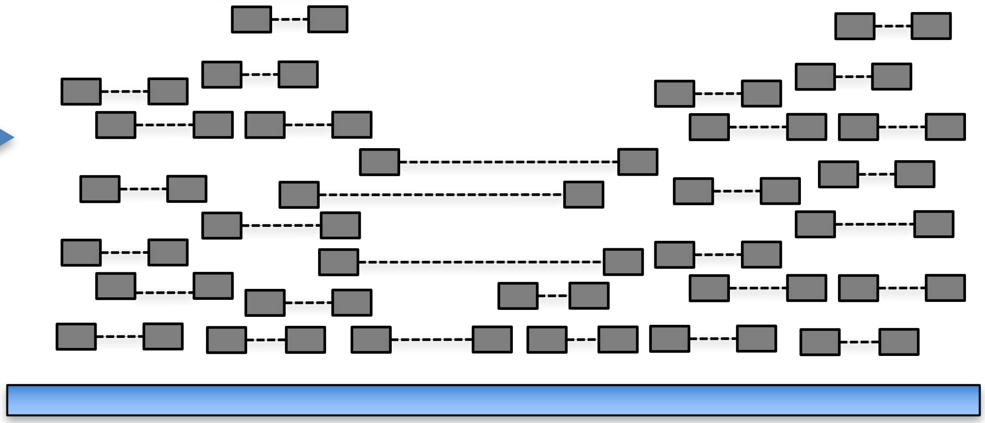
Bulk Sequencing of a tumor sample



Heterogeneous Tumor Sample



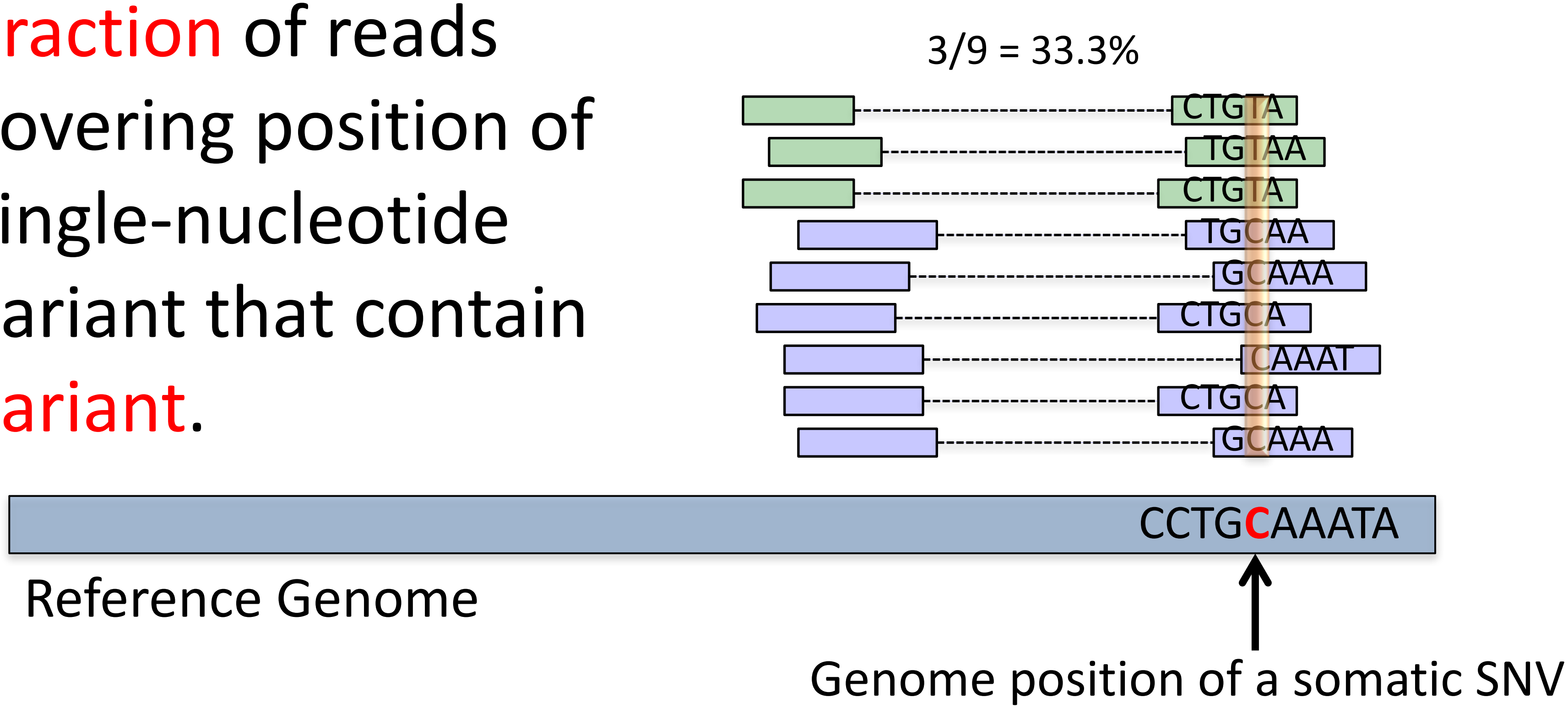
DNA is fragmented



The Reference Genome

Variant Allele Frequency (VAF)

Fraction of reads covering position of single-nucleotide variant that contain variant.

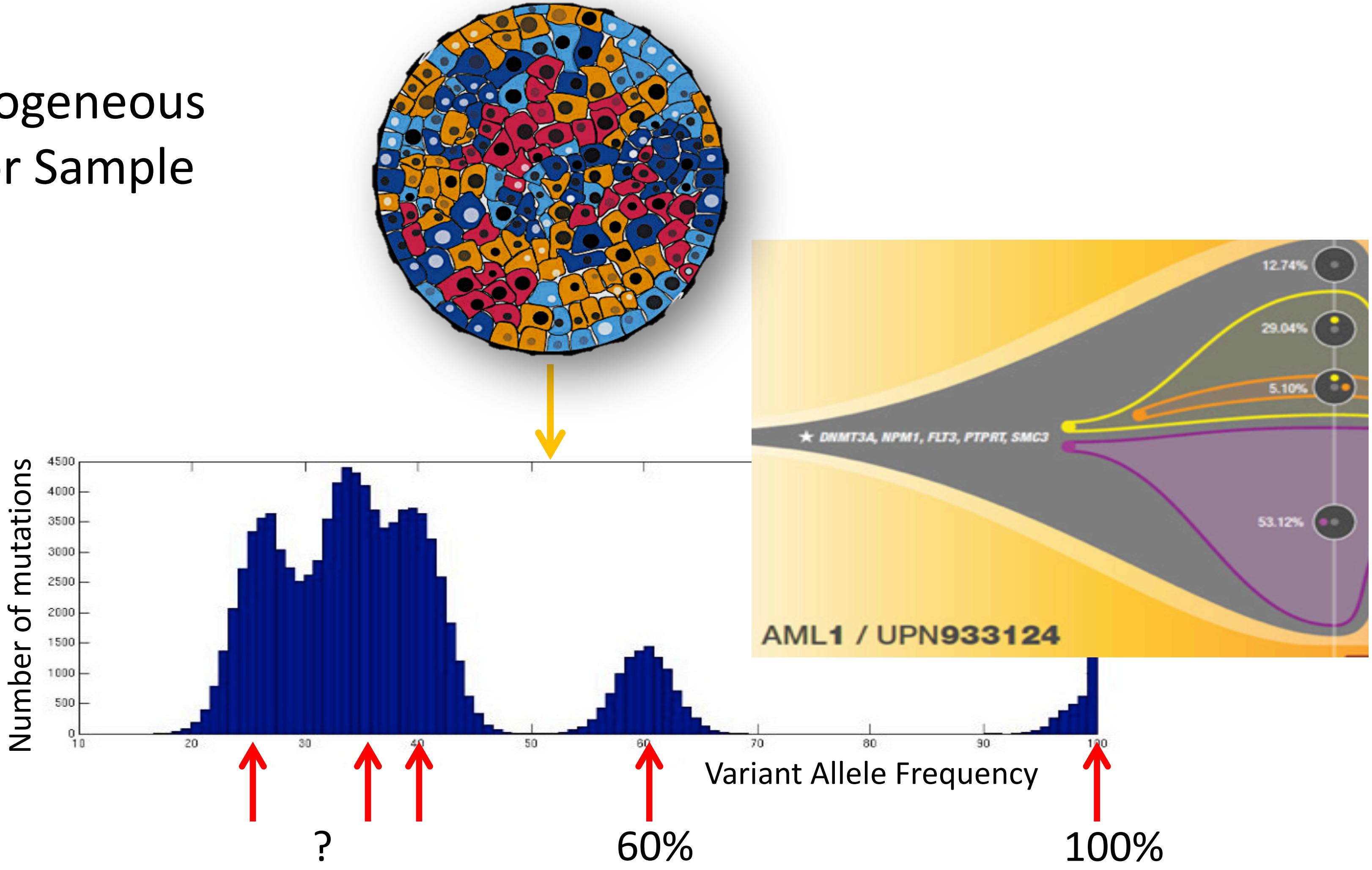


VAF ∝ fraction of tumor cells containing variant allele

**assuming no copy number aberrations*

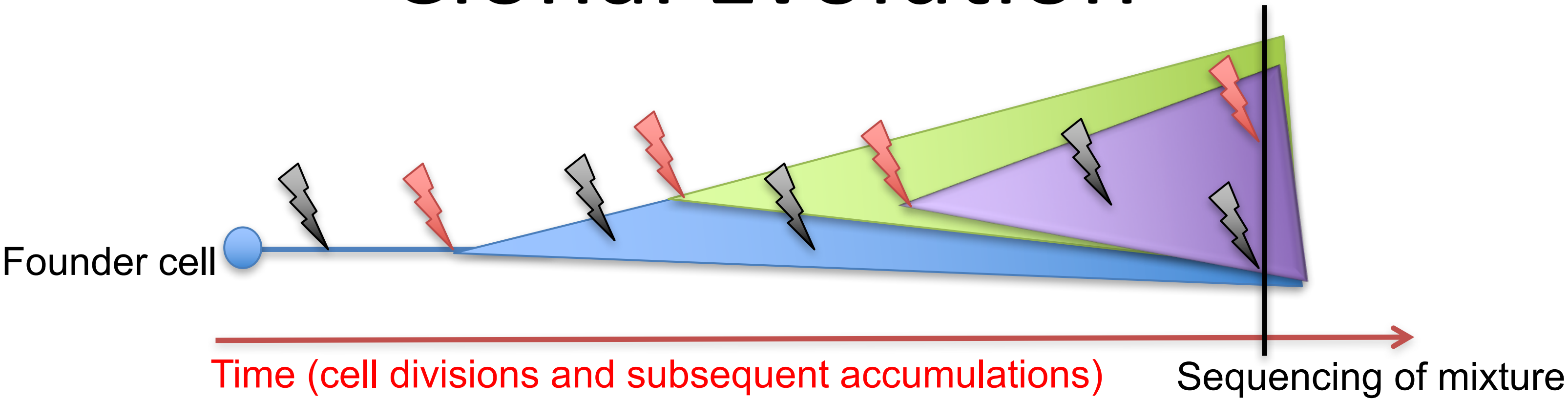
Infer Heterogeneity from VAFs

Heterogeneous
Tumor Sample



Dirichlet Process Mixture models are popular as they do not fix the number of clusters in advance.

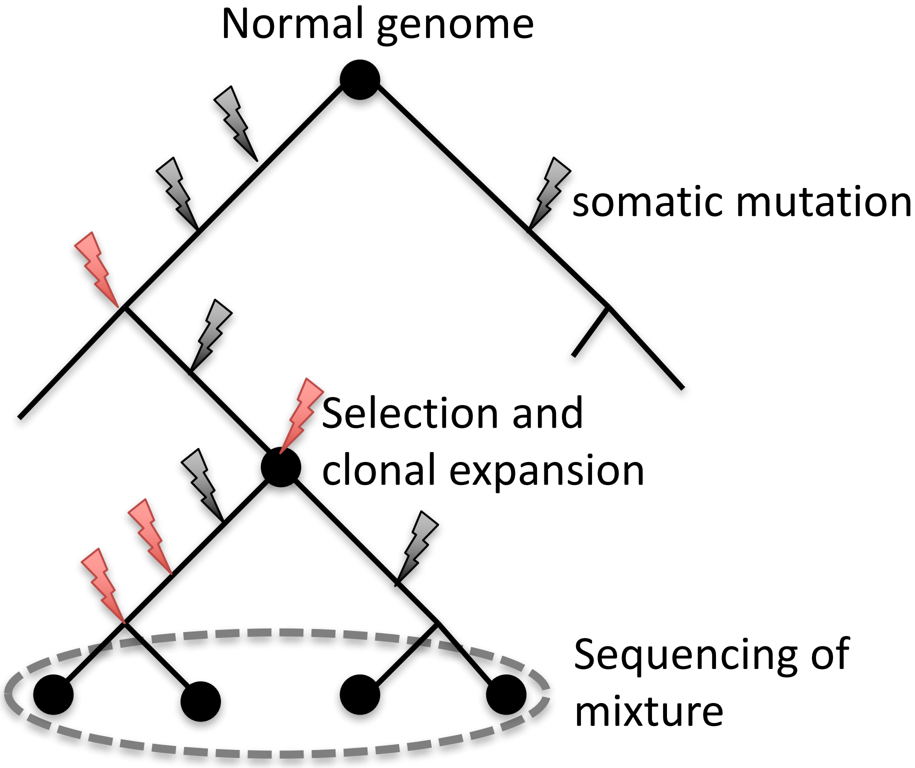
Clonal Evolution



Founder cell

Time (cell divisions and subsequent accumulations)

Sequencing of mixture



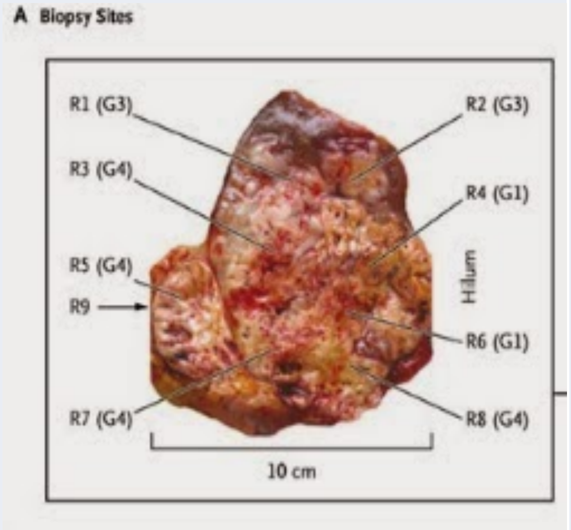
Normal genome

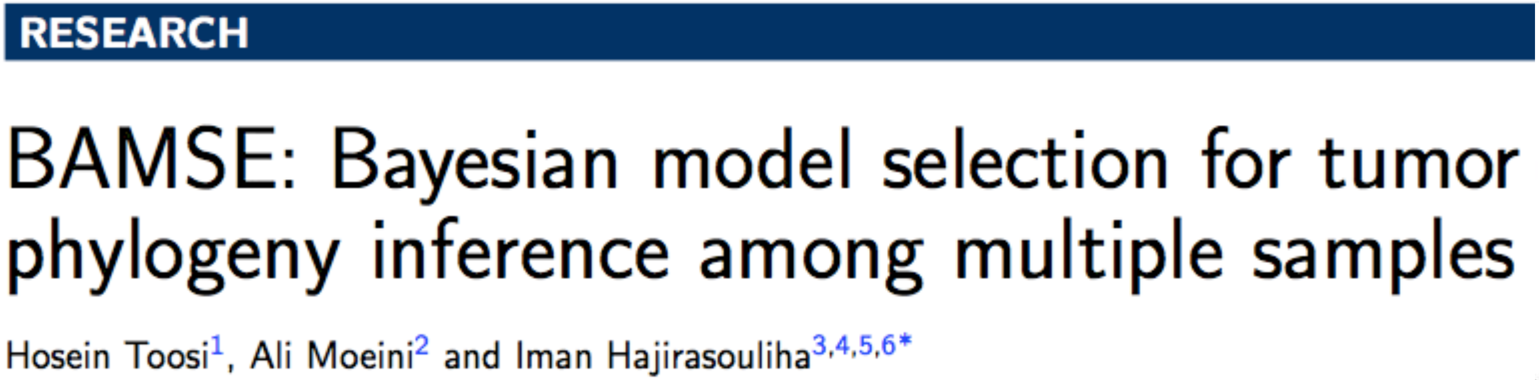
somatic mutation

Selection and clonal expansion

Sequencing of mixture

Single sample *vs.* Multiple samples

Sequencing method	Mixing	Inferring Tree
Bulk (one sample)	yes	TrAp [Strino *et al.*, 2013] **Rec-BTP [Hajirasouliha *et al.*, 2014]**
Bulk (multiple samples) 	yes	PhyloSub [Jiao *et al.*, 2014] Clomial [Zare *et al.*, 2014] **Binary *F* [Hajirasouliha *et al.*, 2014]** SubcloneSeeker [Qiao *et al.* 2014] CITUP [Malikic *et al.*, 2015] BitPhylogeny [Yuan *et al.*, 2015] **LICHeE [Popic *et al.*, 2015]** SCHISM [NikNafs *et al. 2015*] AncesTree [El-Kebir, Oesper *et al.*, 2015] **BAMSE [Toosi, Moeini, Hajirasouliha, 2017]**



RESEARCH

BAMSE: Bayesian model selection for tumor phylogeny inference among multiple samples

Hosein Toosi[1], Ali Moeini[2] and Iman Hajirasouliha[3,4,5,6]*

- BAMSE defines a Bayesian prior over all possible clustering of mutations and tree configurations

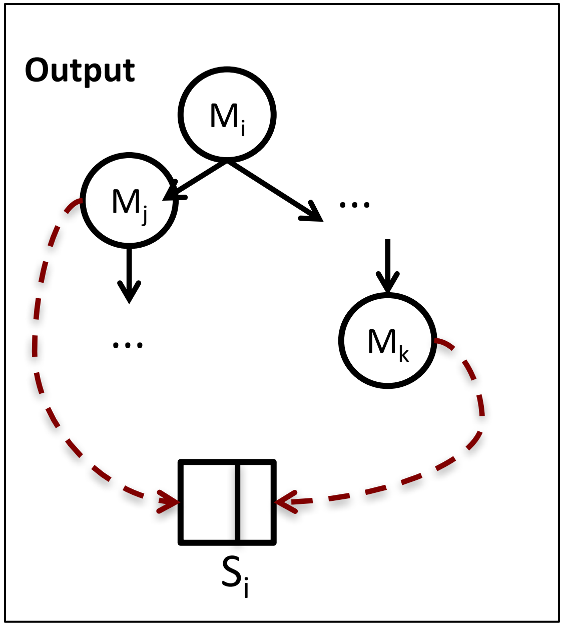
- Accurate maximum likelihood values by convex optimization

Input Data

Variant allele frequencies (VAFs) per sample

Single Nucleotide Variants (SNVs)

	#chr	position	description	Normal	S_1	S_2	S_3	…	S_M
M_1	1	184306474	A/G HMCN1	0.0	0.1	0.2	0.25		0.15
M_2	1	18534005	C/A IGSF21	0.0	0.1	0.25	0.2		0.1
M_3	1	110456920	G/A UBL4B	0.0	0.4	0.4	0.45		0.45
…									
M_N	10	26503064	C/G MYO3A	0.0	0.4	0.0	0.0		0.24



Note: In general, the method can handle any type of variant given its cell prevalence (CP) values in each sample

Perfect Phylogeny Model: Assumption

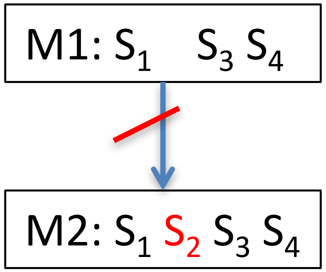
Mutations do not recur independently in different cells
⇒ cells sharing the same mutation must have inherited it
from a common ancestral cell

Perfect Phylogeny Model: Constraints

Three SNV Ordering Constraints:

1. a mutation present in a given set of samples cannot be a successor of a mutation present in a smaller subset of these samples
2. a mutation cannot have a VAF higher than that of its predecessor mutation (except due to CNVs)
3. the sum of the VAFs of mutations disjointly present in distinct subclones cannot exceed the VAF of a common predecessor mutation present in these subclones

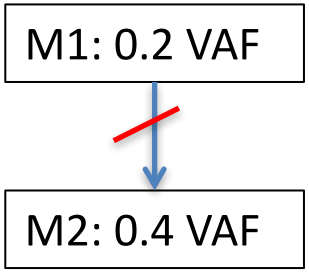
(1)



M1: S_1 S_3 S_4

M2: S_1 S_2 S_3 S_4

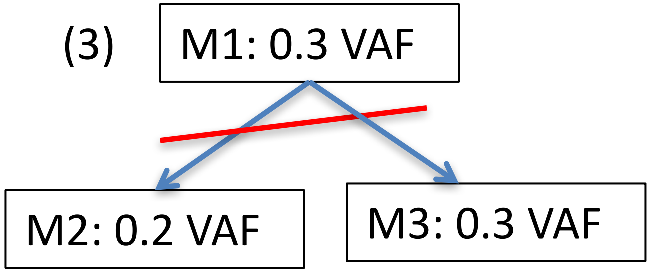
(2)



M1: 0.2 VAF

M2: 0.4 VAF

(3)



M1: 0.3 VAF

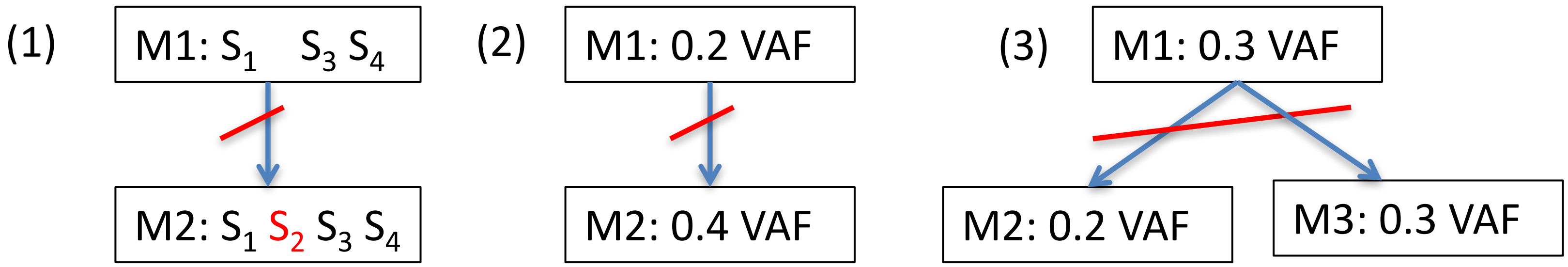
M2: 0.2 VAF

M3: 0.3 VAF

Perfect Phylogeny Model: Constraints

Three SNV Ordering Constraints:

1. a mutation present in a given set of samples cannot be a successor of a mutation present in a smaller subset of these samples
2. a mutation cannot have a VAF higher than that of its predecessor mutation (except due to CNVs)
3. the sum of the VAFs of mutations disjointly present in distinct subclones cannot exceed the VAF of a common predecessor mutation present in these subclones



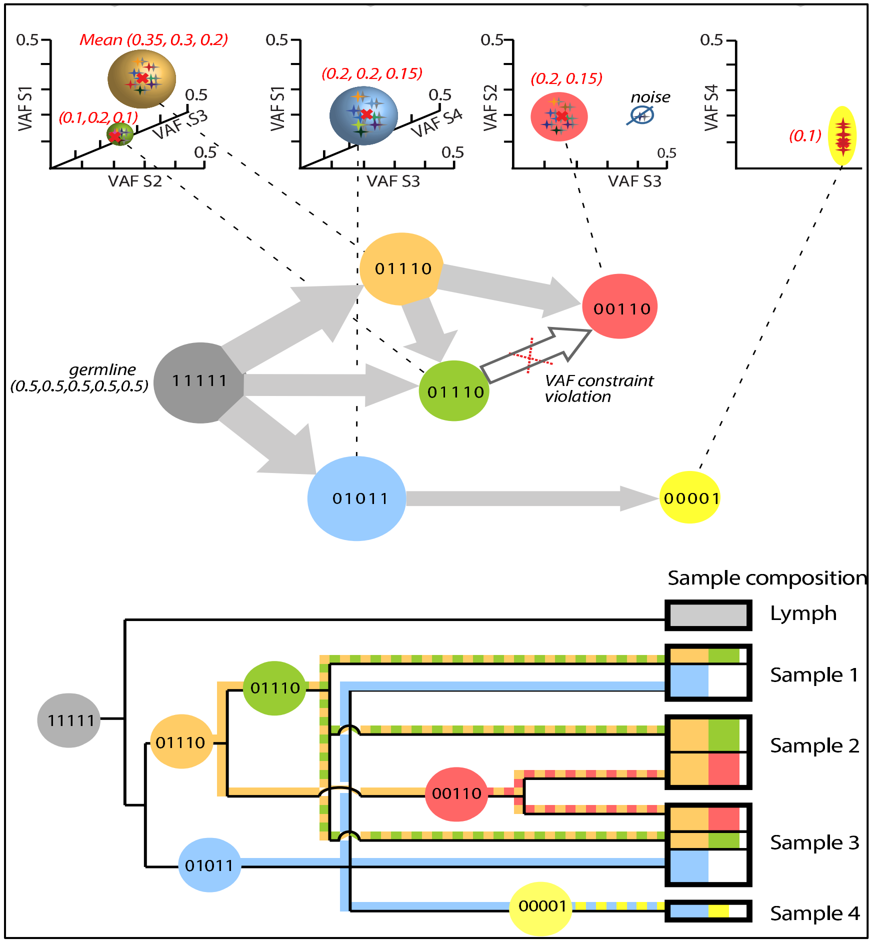
Goal: find all lineage trees that satisfy the above three constraints

Lineage tree across multiple samples

1. Group Somatic SNV.

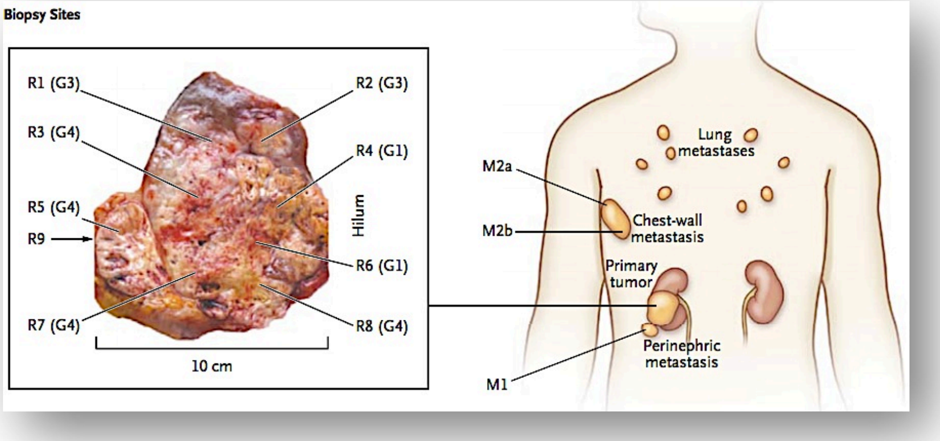
2. Construct *Evolutionary Constraint Network*.

3. Search the network for *all spanning trees*.

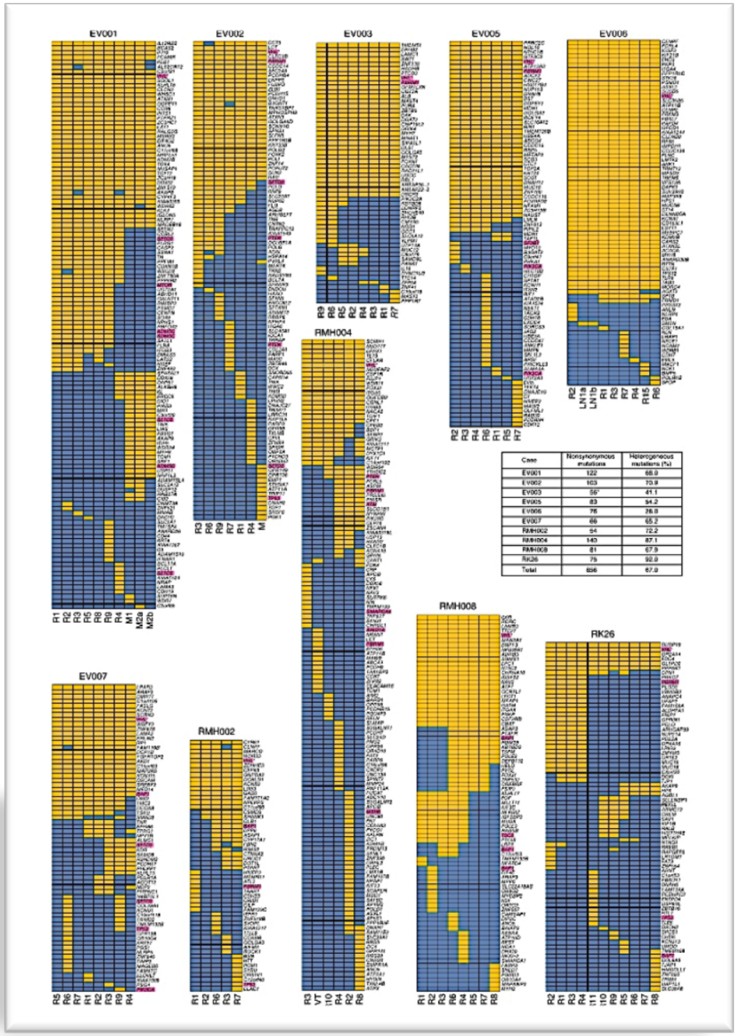


LICHeE: software package

ccRCC Study by Gerlinger et. al (2014)



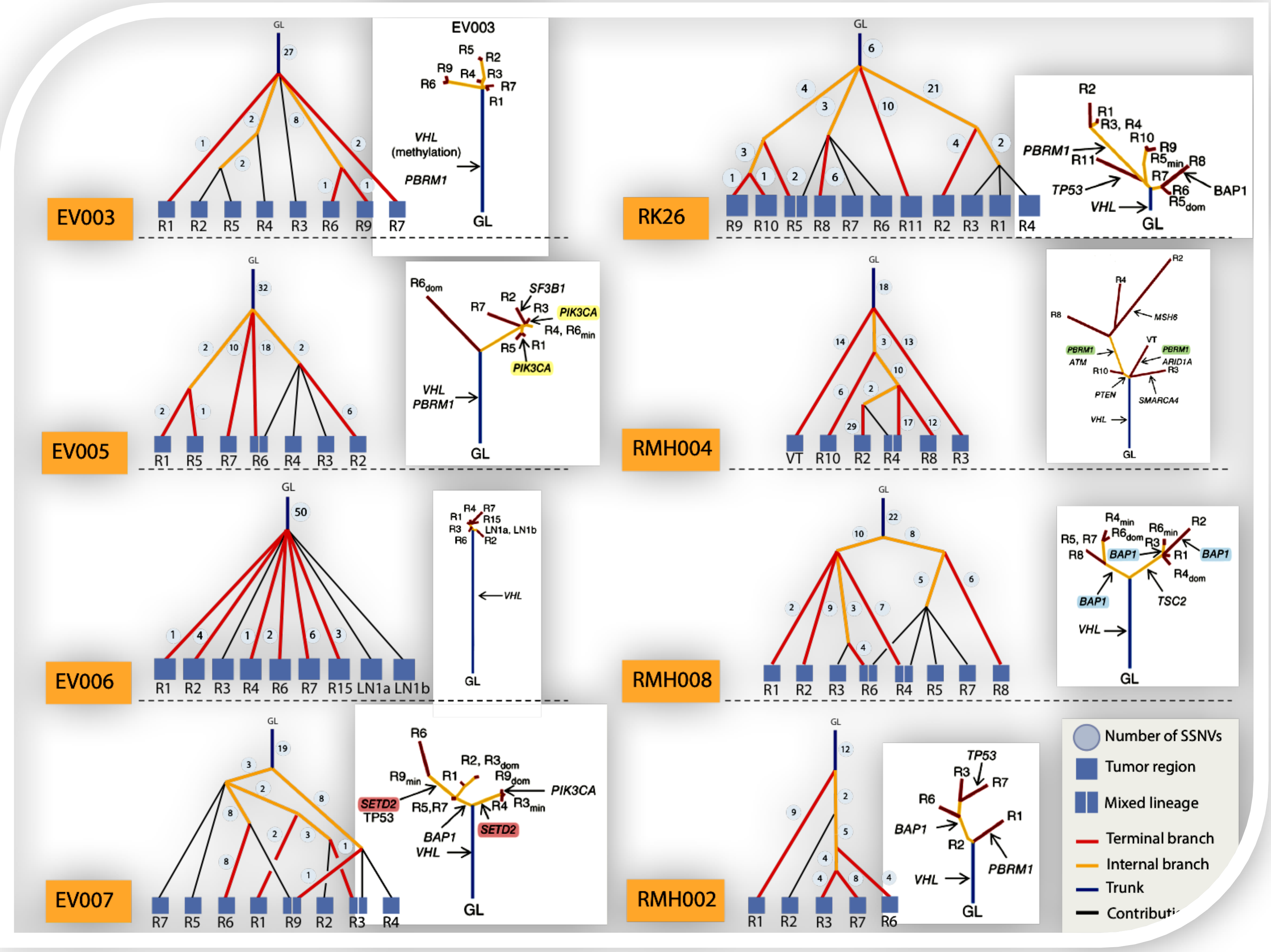
8 patients, 587 SNVs



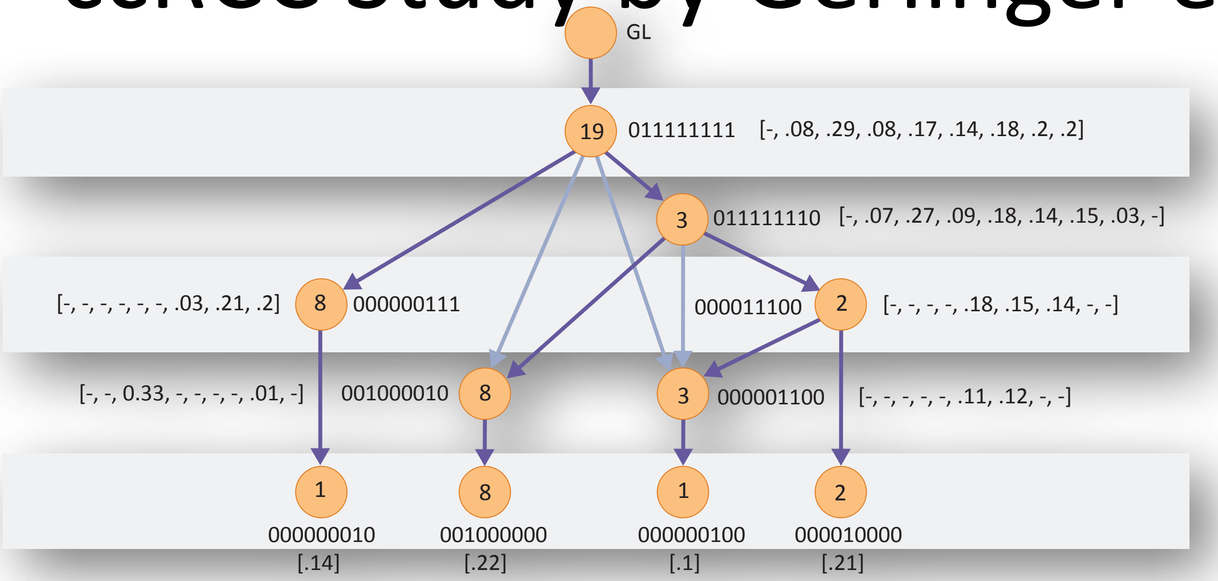
Case	Nonsynonymous mutations	Heterogeneous mutations (%)
EV001	122	68.0
EV002	103	70.9
EV003	56*	41.1
EV005	83	54.2
EV006	76	26.0
EV007	66	65.2
RMH002	54	72.2
RMH004	140	87.1
RMH008	81	67.9
RK26	75	92.0
Total	856	67.0

Gerlinger, M., et al. (2014). "Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing." Nature genetics **46**(3): 225-233.

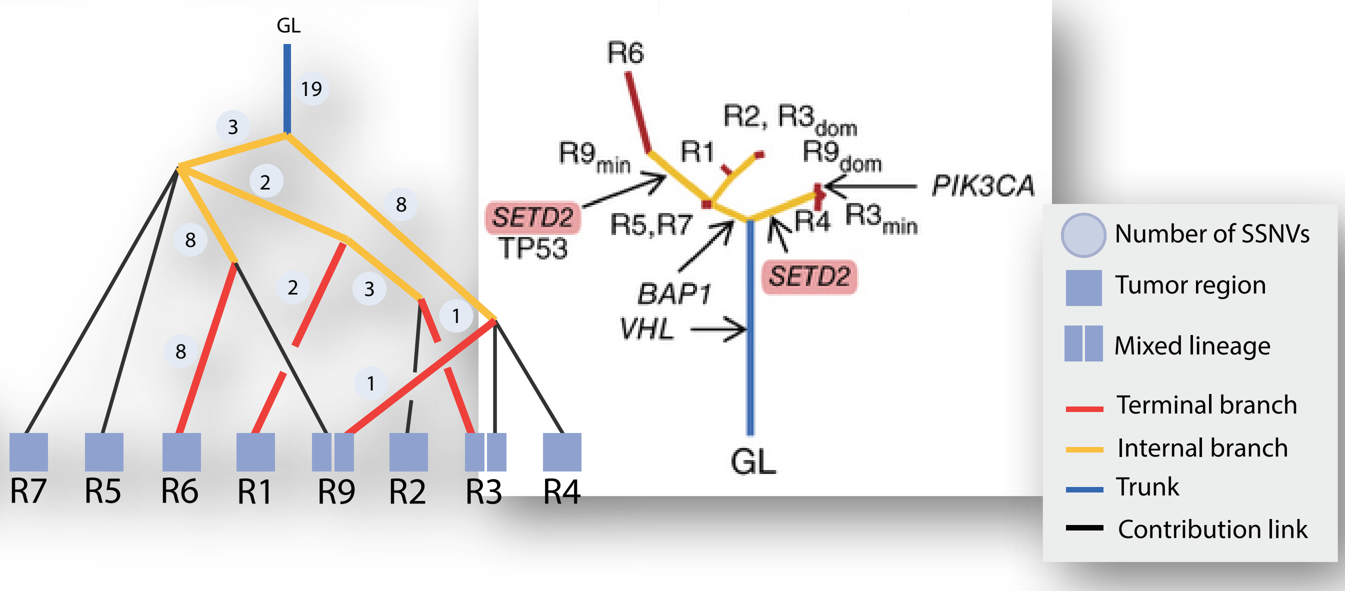
ccRCC Study by Gerlinger et. al (2014)


EV003
GL
27
R1 R2 R5 R4 R3 R6 R9 R7
R5 R2 R9 R4 R3 R6 R7 R1
VHL (methylation)
PBRM1
RK26
R9 R10 R5 R8 R7 R6 R11 R2 R3 R1 R4
R2 R1 R3, R4 R10 R9 R5min R8 R11 R7 R6 R5dom
PBRM1
TP53
VHL
BAP1
EV005
32
R1 R5 R7 R6 R4 R3 R2
R6dom R7 R2 SF3B1 R3 PIK3CA R4, R6min R5 R1 PIK3CA
VHL PBRM1
RMH004
18
VT R10 R2 R4 R8 R3
R2 R4 R8 MSH6 VT PBRM1 ATM R10 PBRM1 ARID1A R3 PTEN SMARCA4 VHL
EV006
50
R1 R2 R3 R4 R6 R7 R15 LN1a LN1b
R4 R7 R1 R15 R3 LN1a, LN1b R6 R2
VHL
RMH008
22
R1 R2 R3 R6 R4 R5 R7 R8
R4min R5, R7 R6dom R6min R2 R8 R3 BAP1 R1 BAP1 R4dom BAP1 TSC2 VHL
EV007
19
R7 R5 R6 R1 R9 R2 R3 R4
R6 R2, R3dom R9min R1 R9dom PIK3CA SETD2 TP53 R5,R7 R4 R3min BAP1 SETD2 VHL
RMH002
12
R1 R2 R3 R7 R6
TP53 R3 R7 R6 R1 BAP1 R2 PBRM1 VHL
Number of SSNVs
Tumor region
Mixed lineage
Terminal branch
Internal branch
Trunk
Contributi

ccRCC Study by Gerlinger et. al (2014)

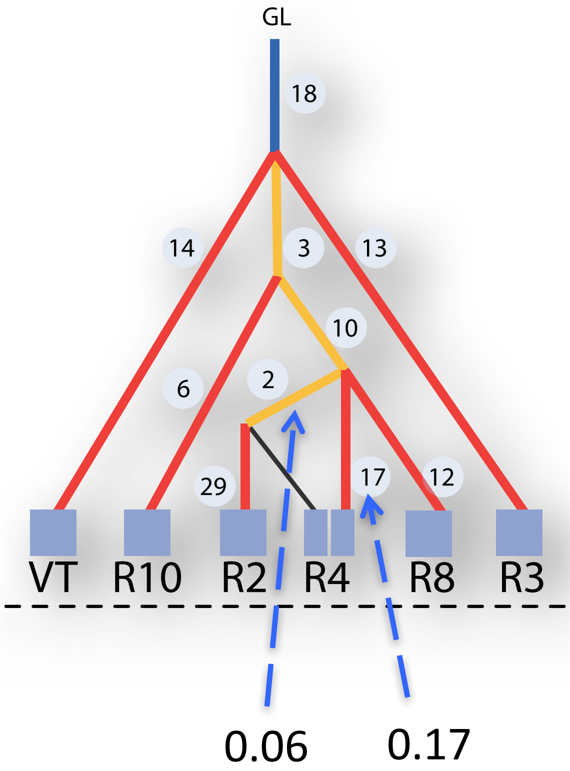


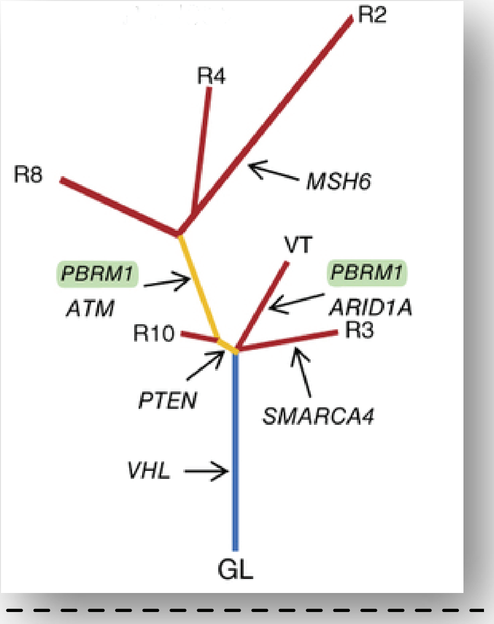
EV007



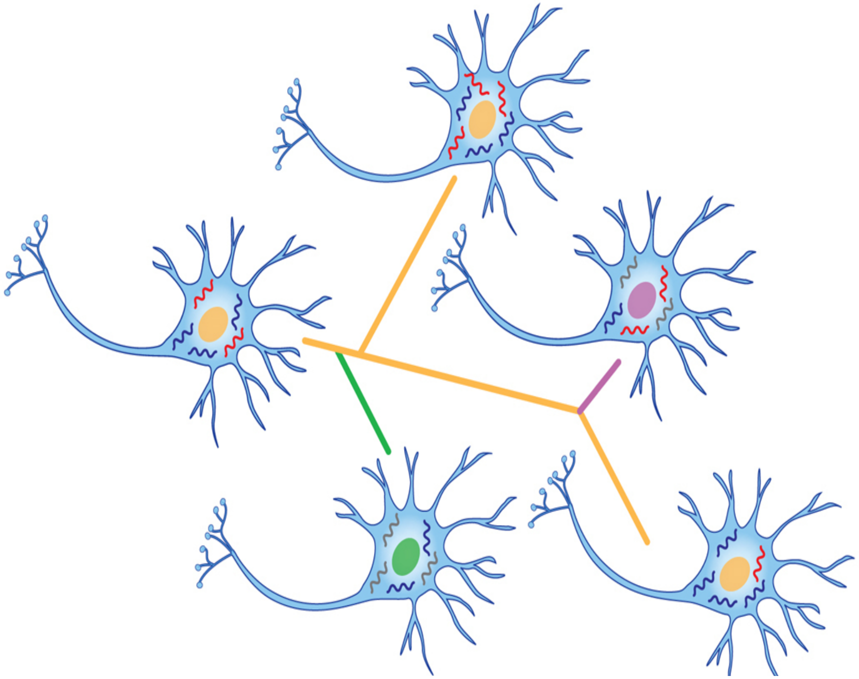
ccRCC Study by Gerlinger et. al (2014)

RMH004





Single cell genome sequencing

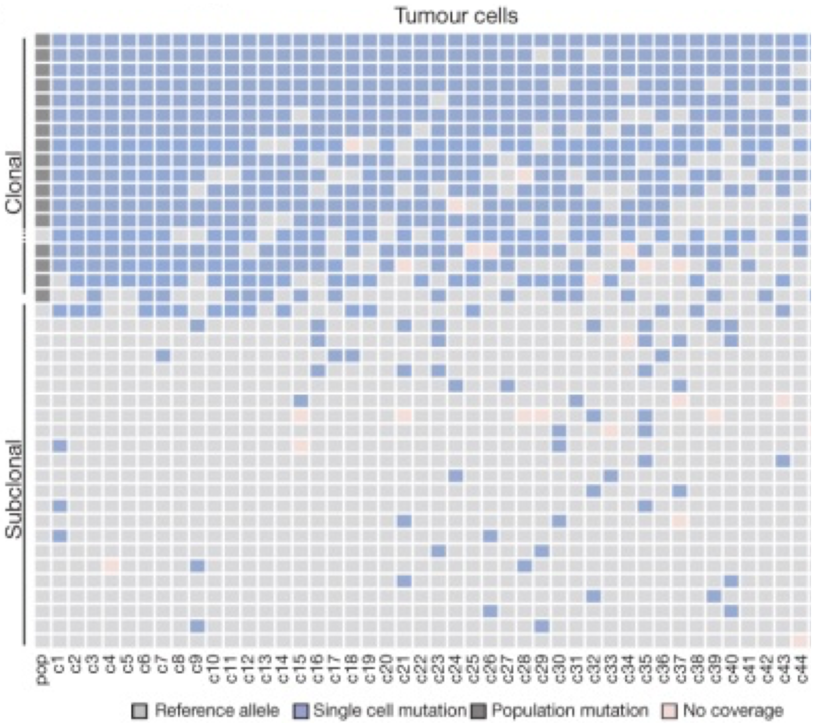


*Katie Vicari**

*Image From: Eberwine et al. ***Nature Methods*** 2014

Single cell vs. bulk sequencing

Single Cell Sequencing (SCS)



Bulk Sequencing

ID	Chromosome	Position	MutantCount	ReferenceCount	INFO
mut1	15	73021943	393	1607	geneID=BBS4
mut2	9	138702709	337	1663	geneID=CAMSAP1
mut3	3	51263127	382	1618	geneID=DOCK3
mut4	1	38226084	412	1588	geneID=EPHA10
mut5	6	133850054	201	1799	geneID=EYA4
mut6	19	40895668	654	1346	geneID=HIPK4
mut7	6	27101163	380	1620	geneID=HIST1H2AG
mut8	8	95877709	516	1484	geneID=INTS8
mut9	8	120255800	966	1034	geneID=MAL2
mut10	1	24390601	466	1534	geneID=MYOM3

$$VAF = \frac{MutantCount}{MutantCount + ReferenceCount}$$

Single cell vs. bulk sequencing

Single Cell Sequencing (SCS)

Advantages

- Better sequencing resolution
- The presence or absence of every mutation in each cell is clearly distinguishable
- New technique that can only improve as time passes
- Low rate of False Positives (read errors)

Disadvantages

- Data extracted from SCS are extremely noisy:
    - High rate of False Negatives (~15-30 % -- allelic dropout)
    - High rate of Missing Values (~10-40 %)

Bulk Sequencing

Advantages

- Better accuracy
- Cheaper than Single Cell Sequencing

Disadvantages

- Lower sequencing resolution
- More difficult interpretation of the data

Thank you!

Weill Cornell Medicine
Chris Mason
Daniela Bezdan

Salem Malikic **(SFU)**
Stephen Williams **(10X Genomics)**
Patrick Marks **(10X Genomics)**
Cenk Sahinalp **(Indiana)**
Victoria Popic **(Illumina)**

David Danko (Tri-CBM)
Simone Ciccolella (Visiting Student)
Camir Ricketts (Tri-CBM)
Dmitrii Meleshko (Tri-CBM)

10x GENOMICS®

Weill Cornell Medicine

SIMONS INSTITUTE for the Theory of Computing