DCF

Thursday, February 18, 2021 2:00 PM

genotypes (x, y, z)

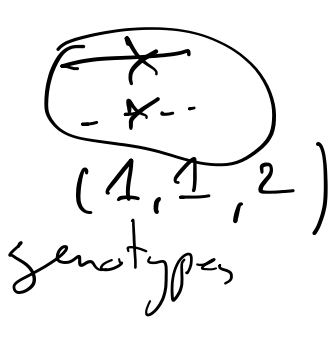
x is # maternal copies

y is # paternal "

z # mutated " "

$z \leq \max(x, y)$

$\boxed{z \leq x+y}$



$(1,1,2)$

genotype set Γ : set of all genotypes in the tumor of a given SNV locus.

$g_{(x,y,z)}$ is genotype proportion of genotype (x,y,z)

$g_{(x,y,z)} \geq 0 \quad \forall (x,y,z)$ and $\sum_{(x,y,z)} g_{(x,y,z)} = 1$

define VAF v and CCF c

Given Γ and $\bar{g}$:

$$v = \frac{\sum_{(x,y,z)\in\Gamma} z \cdot g_{(x,y,z)}}{\sum_{(x,y,z)\in\Gamma} (x+y) \cdot g_{(x,y,z)}}$$

Fractional copy number

$$\Gamma_{CCF} = \{(x,y,z) \in \Gamma : z > 0\}$$

$$c = \frac{1}{\rho} \sum_{(x,y,z)\in\Gamma_{CCF}} g_{(x,y,z)}$$

cancer cell fraction ↙ normalize by purity

~~Unnormalized~~ ↙ cellular prevalence

purity ρ is proportion of tumor cells in the sample

How to estimate CCF from data

Ⓘ VAF $\hat{v} = \frac{a}{d}$ ← #alternate/variant reads, ← #total number of reads

$\hat{v}$ ↑ after sequencing

$$\lim_{d\to\infty} \frac{a}{d} = v$$

v ↑ population

HATCHET

CHISEL

Ⓘⓘ $\bar{\mu} = [\mu_{(x,y)}]$ copy-number proportion; (x,y) ↑ copy number

$\mu_{(x,y)} \geq 0 \quad \forall (x,y)$

$\sum_{(x,y)} \mu_{(x,y)} = 1$

How to estimate CCF c given $\hat{v}$ and $\bar{\mu}$?

$$\boxed{c \approx \frac{1}{\rho}\left(\frac{\rho N_{tot} + (1-\rho)\cdot 2}{M}\right)\hat{v}}$$

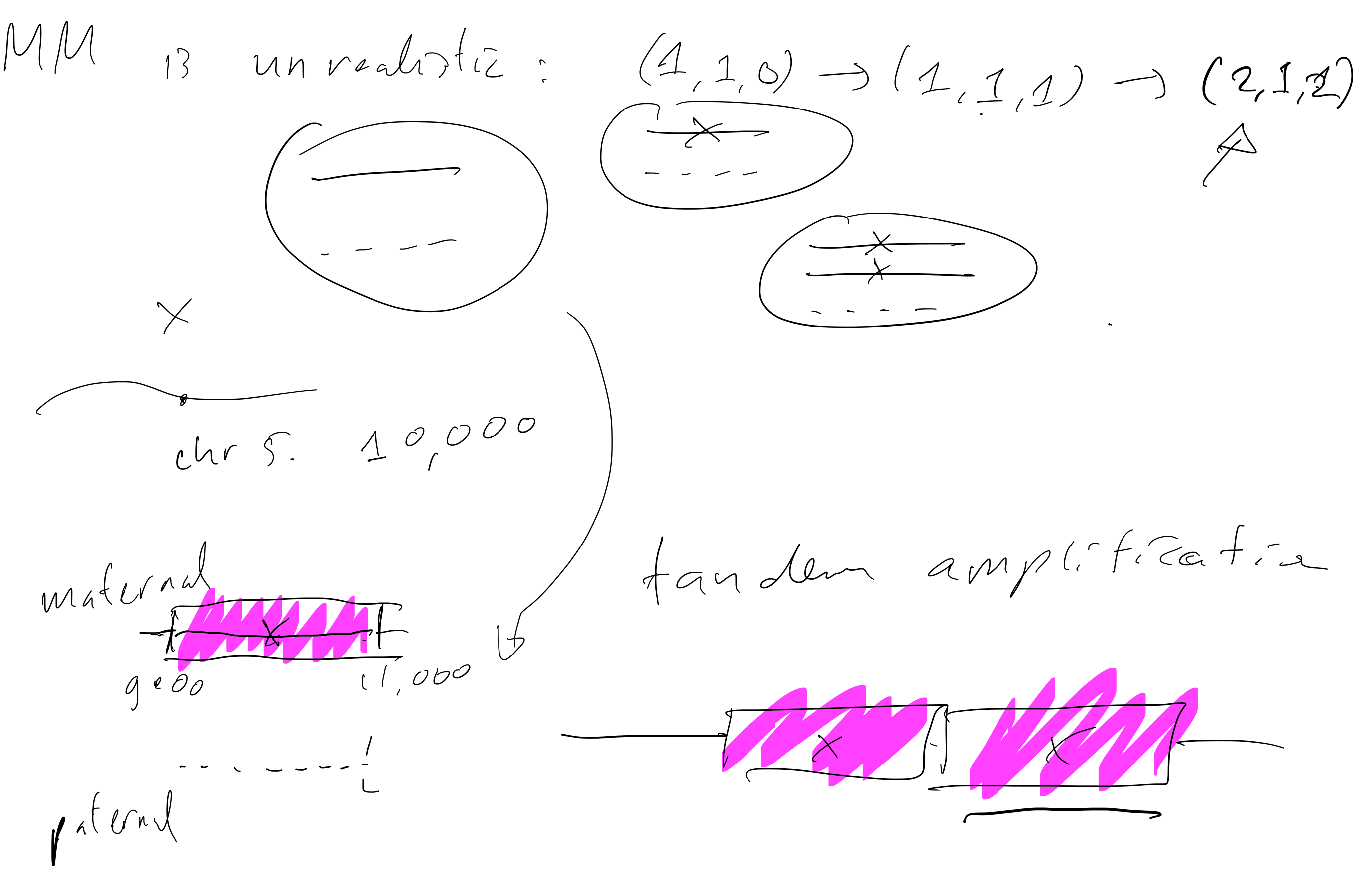
ρ : purity of sample

N_{tot} : average copy number $(x+y)$ of cancer cells.

M : mutation multiplicity

Constant Mutation Multiplicity assumption – (CMM). At every SNV locus, there exists an integer $M \geq 1$ s.t. all genotypes at the locus have the form (x, y, z) where either $z = 0$ or $z = M$.

CMM is unrealistic : $(1,1,0) \rightarrow (1,1,1) \rightarrow (2,1,2)$



genotype set Γ & genotype proportions $\bar{g}$

uniquely determine

VAF (v)

CCF (c)

- What if we're ~~given~~ v (or c) and $\bar{\mu}$ can we "say something" about uniquely determine $\bar{g}$?

- What constraint on Γ allows us to achieve ☆?

$$\mu_{(x,y)} = \sum_{z \geq 0} g_{(x,y,z)} \quad \forall (x,y)$$

Γ

Find $\bar{g}$

(x^*, y^*) → $(x^*, y^*, 0)$, → $(x^*, y^*, 1)$

$\mu_{(x^*,y^*)} = g_{x^*,y^*,0} + g_{x^*,y^*,1}$

$(x,y) \neq (x^*,y^*)$ | (x, y, z)

$\boxed{\mu_{x,y} = g_{(x,y,z)}}$

$$\boxed{\text{CCF}} \rightarrow v = \frac{\sum_{(x,y,z)\in\Gamma} z \cdot g_{(x,y,z)}}{\sum_{(x,y,z)\in\Gamma} (x+y)\, g_{(x,y,z)}}$$

constant

$$= \sum_{(x,y)} (x+y)\, \mu_{(x,y)}$$

linear system of equations

c ↓ genotype proportion ↓ VAF

Single split copy number assumption (SSCN)

Given v, $\bar{\mu}$ and SSCN genotypes Γ^* purity ρ, $\boxed{\text{CCF}}$ is uniquely determined

---

② Clustering using CCF

If CCF is fixed

$\text{CCF}_1, \ldots, \text{CCF}_k$

compute

$v_1, \ldots, v_k$

$\frac{a}{d}$

$a \sim \text{Binom}(d, v)$