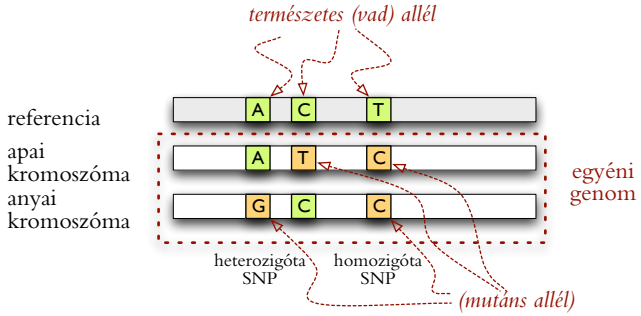
10. Egyéni variációk

EGY POPULÁCIÓBAN mindig találunk egyéni variációkat — az egyéni genomok eltérnek a referenciától A genom egy adott helyén (**lokusz**ánál) megtalálható variációt **genotípus**nak hívjük, aminek a lehetséges értékei az **allélek**. A leggyakoribb variáció a SNP (*Single Nucleotide Polymorphism*), azaz az egy nukleotidra kiterjedő pontmutáció. Diploid kromoszómák lokuszainál két allél található, az egyed lehet homozigóta (azonos allél az apai és anyai kromoszómán) vagy heterozigóta (különböző allélek).



1. ábra. Heterozigóta és homozigóta SNP-ek.

A genomhoz illesztett *read*-ek alapján megállapított genotípusokat tipikusan VCF (*Variant Call Format*) fájlban szokás leírni. Egy fájl több genom leírását is tartalmazhatja (a példában az NA00001–NA0003 mintákhoz tartozó genotipusok szerepelnek). A fájl fejlécében ## kezdetű sorok megadják az információt az adatok eredetéről, illetve az INFO, FILTER és FORMAT oszlopokban szereplő kódok jelentéséről és adatformátumokról. INFO: adatok az összes mintáról (itt NS,DP, AF,...). FILTER: lokusz szűrése az adatok minősége alapján (itt q10 és s50 lehetséges szűrőkkel). FORMAT: megadja a követő genotípus oszlopok formátumát (itt GT,GQ,DP,HQ).

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##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=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##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        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Az ALT oszlopban felsorolt allélek szerint kódoljuk a genotípust: 0=referencia, 1=első ALT, 2=második ALT, ... A **haplotípus** megadja azt is, hogy több egymás utáni lokusznál a heterozigóta allélek melyike található ugyanazon a kromoszómán, azaz a genotípusok *fázisát*.

Unphased

```
##fileformat=VCFv4.2
...
... REF ALT ... FORMAT smp
...  A   G  ... GT     0/1
...  C   T  ... GT     0/1
...  T   C  ... GT     1/1
```

Phased

```
##fileformat=VCFv4.2
...
... REF ALT ... FORMAT smp
...  A   G  ... GT     0|1
...  C   T  ... GT     1|0
...  T   C  ... GT     1|1
```

SNP calling

Az egy lokuszra illeszkedő *read*-ek alapján megállapítható a diploid genotípus. Egy helyre illeszkedő szekvenált bázisok a quality értékekkel: $\mathcal{Z} = \{(z_1, q_1), (z_2, q_2), \ldots, (z_n, q_n)\}$. Lehetséges genotípusok: 4 homozigóta, 6 heterozigóta. Ismeretlen genotípus: Y.

$$\mathbb{P}\Big\{Y = y_1y_2 \;\Big|\; \mathcal{Z}\Big\} \propto \underbrace{\mathbb{P}\Big\{\mathcal{Z} \;\Big|\; Y = y_1y_2\Big\}}_{y_1y_2 \text{ likelihood}} \times \underbrace{\mathbb{P}\{y_1y_2\}}_{y_1y_2 \text{ genotípus gyakorisága}}$$

A quality értékekkel könnyen számolható $\mathbb{P}\Big\{\mathcal{Z} \;\Big|\; Y = y_1y_2\Big\}$ akár homozigótákra ($y_1 = y_2$) akár heterozigótákra ($y_1 \neq y_2$). De mi a genotípusok a priori valószínűsége?

Hardy-Weinberg egyensúly. Egy végtelen méretű populációban, diszkrét generációk és pánmixia (nincsenek nemek, de két szülő van mindig) esetén az egyensúlyi frekvenciák a következők: $\mathsf{AA} \sim p^2$, $\mathsf{Aa} \sim 2pq$, $\mathsf{aa} \sim q^2$, ahol p a vad allél (A) és $q = 1 - p$ a mutáns allél (a) gyakorisága.

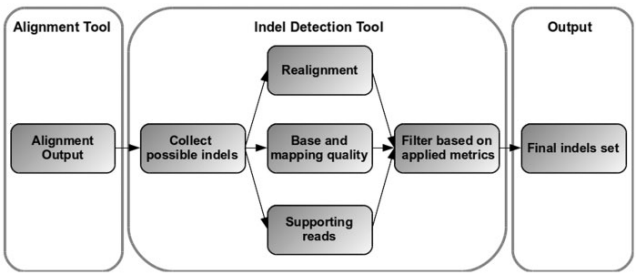
SNP frekvenciák. MAF (*Minor Allele Frequency*): mutáns allél gyakorisága a populációban. Egy SNP gyakori, ha $\mathsf{MAF} > 10\%$ és ritka, ha $\mathsf{MAF} < 5\%$. A HAPMAP (2005, 2009) és 1000 Genomes (2012) projektek az emberi populációkban felelhető egyéni variációk feltérképezését vette célba, eredetileg genotipizáló csipeket, később exom[1] és teljes genom szekvenálást használva. Így már specifikus MAF adatok is elérhetőek különböző populációkra: CEU (Észak-Amerika, európai ősökkel), YRI (Yoruba, Nigéria), JPT (Japán), CHB (Han kínai); ASW (afro-amerikai), GIH (Houston-i Gujarati), MEX (mexikóiak Los Angelesből), LWK (Luhya, Kenya), ...

[1] Az exom a genomban ismert exonok összessége. Sok egyéni szekvenáló projekt csak az exomot vizsgálja (amihez az exonokat tartalmazó DNS fragmenseket előbb ki kell válogatni), mert csak a kódoló génekbe eső mutációkat keresik.

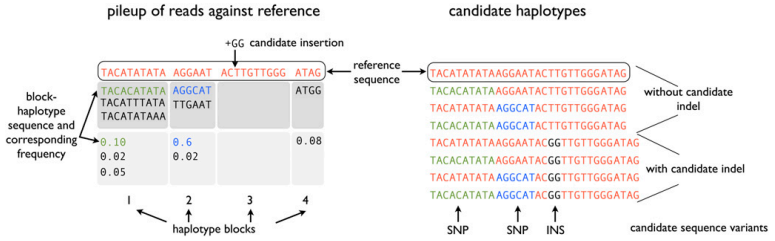
Indel variációk

A SNP-ek mellett előfordulnak törléses vagy beszúrásos polimorfizmusok is, kb. 1/8-szor ritkábban, de összességében hasonló számú bázist érintve. Egy tipikus emberi genomban lesz néhány millió SNP és néhány százezer indeles polimorfizmus.[2] Kisebb indel polimorfizmusok szekvenálásahoz érdemes újraillleszteni a *read*-eket a lokusz körül.

[2] James Watson genomjában pl. mintegy 3.3 millió SNP-et és 222 ezer indelt annotáltak.



DINDEL. A DINDEL program nagyon pontosan találja meg az indel polimorfizmusokat. Ehhez lehetséges haplotípus blokkokat definiál (egymásnak nem ellentmondó illeszkedő *read*-ek alapján), és azután statisztikus illesztést végez a read-ek és a lehetséges haplotípusok között.



2. ábra. A Dindel egyszerre legeljebb 8 lehetséges haplotípushoz illeszti az egy csúszo ablakba eső DNS olvasatokat. [Albers++]

Megbízható indelekhez 30x lefedettségre van szükség; különböző szoftvercsomagok különböző típusú adatokra lettek kihegyezve.

Tool name	Advantages	Limitations
GATK	Highly supported with good overall performance	Low sensitivity at very low coverage (<10×; can be improved by less stringent parameters)
Dindel	Best performance at low coverage	Only suitable for Illumina data analysis and has long running time
SAMtools mpileup	High PPV and simple use	Lowest sensitivity at high coverage (>50×)
VarScan	High sensitivity at intermediate/high coverage (>30×) and simple use	Low PPV at default parameter settings and low sensitivity at low coverage (<30×)

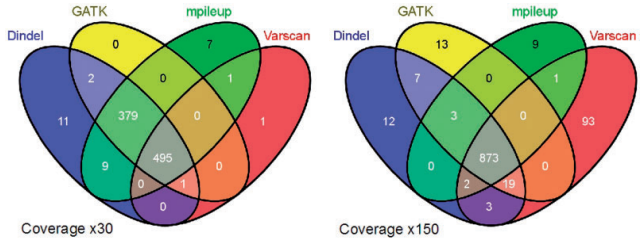


Figure 5: Venn diagram depicting the number of indels found for each software with 30× and 150× coverage and read length 72. Inclusion of indels called in any of the software results in a decrease in PPV with only a mild sensitivity improvement. Inclusion of indels supported by at least two software results in a sensitivity improvement for some of the software and a significant increase in PPV, crucial in high coverage data.

3. ábra. Genotipizáló programok és indelek. PPV (*Positive Predictive Value*) $= \frac{\mathrm{TP}}{\mathrm{TP+FP}}$, TP=true positive, FP=false positive. [Neuman++]

Struktúrális variációk

A kisebb variációk mellett találunk sok bázispárt érintő **struktúrális variációkat** is szép számban. A struktúrális variációk között gyakoriak és nehezen felfedezhetőek a kopiszám variációk (CNV-*Copy Number Variation*), amit genom-duplikációk okoznak ($\geq$ 50 bp hosszban, de akár egész kromoszómára kiterjedően). Struktúrális variációkat fel lehet fedezni több módon (ld. 5. ábra).

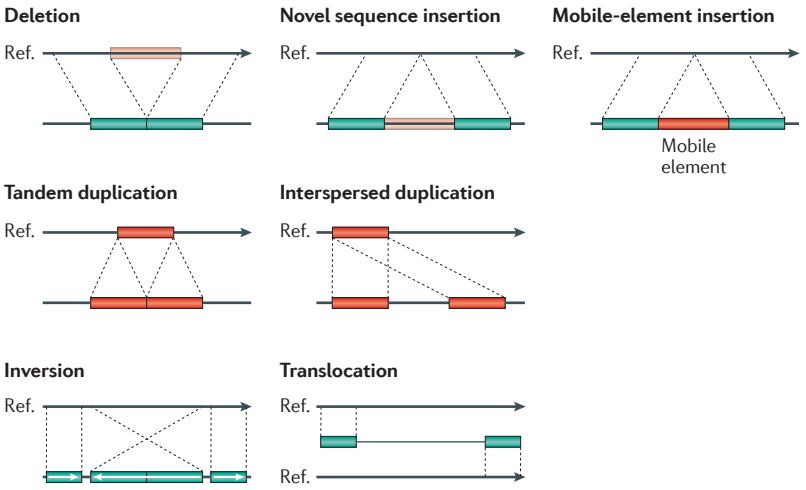
- ⋆ párosított *read*-ek illésztésénél a pár két tagja nem az elvárt módon illeszkedik a referencia genomra
- ⋆ lefedettség gyanúsan nagy vagy túl kicsi egy régióban
- ⋆ DNS fragmens egyik fele máshova illeszkedik mint a másik fele (*split read*)
- ⋆ *de novo* összerakott genom szekvencia átrendezésekkel illeszkedik a referenciához.

Hivatkozások

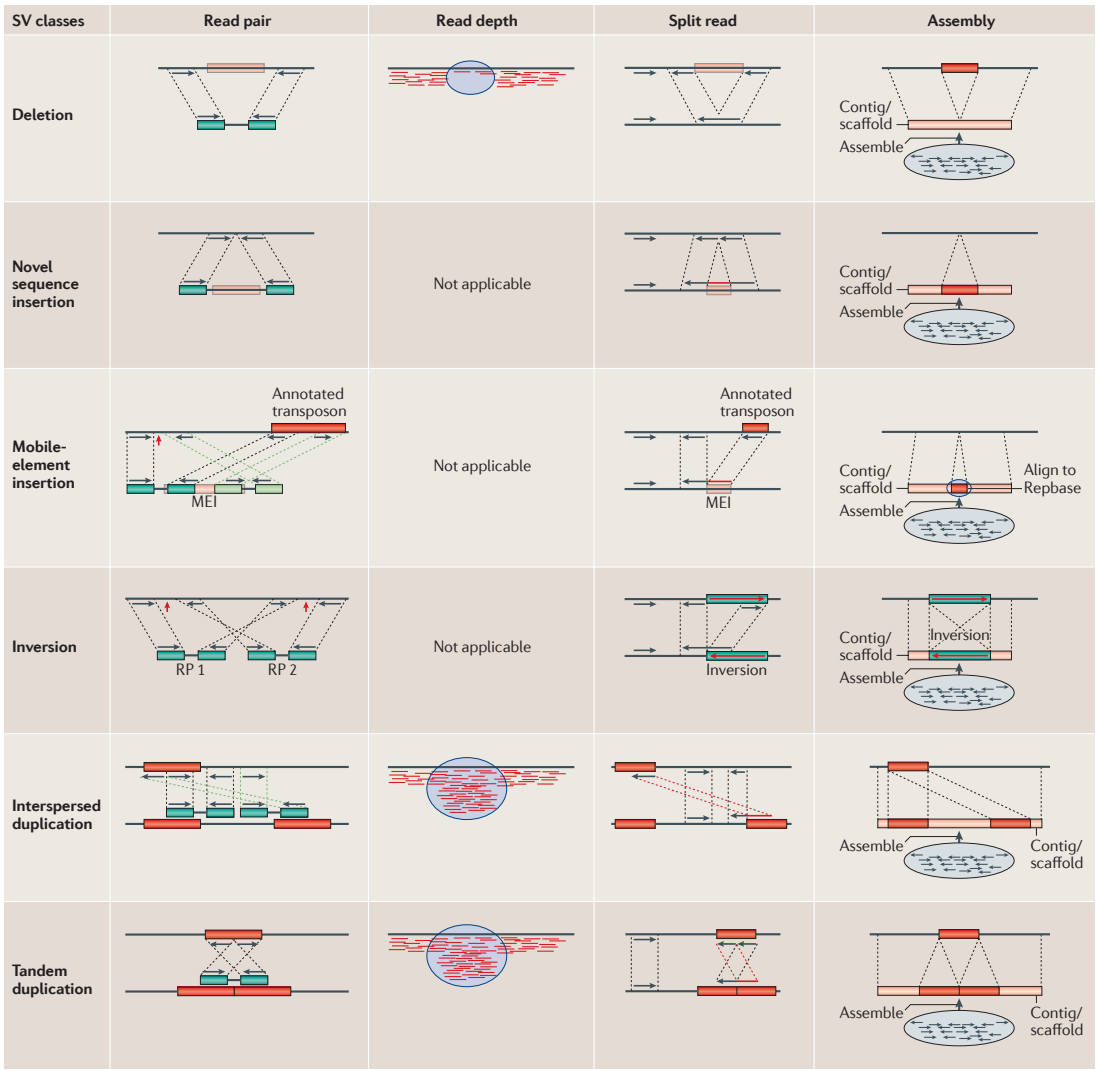
Cornelis A. Albers, Gerton Lunter, Daniel G. MacArthur, Gilean McVean, Willem H. Ouwehand, and Richard Durbin. Dindel: Accurate indel calls from short-read data. *Genome Research*, 21:961–973, 2011. DOI: 10.1101/gr.112326.110.

Can Alkan, Bradley P. Coe, and Evan E. Eichler. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12:363–376, 2012. DOI: 10.1038/nrg2958.

Joseph A. Neuman, Ofer Isakov, and Noam Shomron. Analysis of insertion–deletion from deep-sequencing data: software evaluation for optimal detection. *Briefings in Bioinformatics*, 14:46–55, 2013. DOI: 10.1093/bib/bbs013.



4. ábra. Genom struktúrális variánsok. [Alkan++]



5. ábra. Struktúrális variánsok árulkodó jegyei. [Alkan++]