**Improve INDEL discovery by integrating across platforms** (Pacbio, Illumina, ONT and 10X) (MG, KY)

The second generation sequencing technology like Illumina and 10X data have distinct error profile than the third generation PacBio and ONT. While Short-insert Illumina and haplotype-partitioned 10x reads provide reliable information for calculating short insertion and deletion, PacBio data have abundant sequence errors in homopolymer and micro-satellite regions. Thus the second generation sequence data will be combined for short indels and long single molecule third generation data will be used for detecting larger and complex indels. Nonetheless, error profiles of second and third generations reads are well described and could be simulated *in silico*. We have previous experience in parameterizing indel and complex SV detection (PMID: 21787423). Therefore, we will simulate reads from these sequencing platforms in other to maximize indel detection precision and sensitivity. A hybrid caller combining both generations of sequence data will enable us to investigate indels of full size spectrum, all forms and in diverse regions of the genome with different sequence content.

**Improving methods for lightweight genotyping using short-read data (TR, JK, MG)**

For short-read data, genotyping is mostly performed by using the local sequence information of reads piling up at a genetic locus, which we refer to as *vertical genotyping*. Long reads have, in principle, the capacity to additionally exploit phase information for genotyping. That is, the reads can be partitioned into two sets, corresponding to two haplotypes, using *multiple* variants. With about one heterozygous SNV per 1000 bp in human, long reads usually cover many heterozygous variants, which can be used jointly for haplotype partitioning and hence mutually inform each other. Such partitioning information can be leveraged for genotyping, and we refer to this process as *horizontal genotyping*.

In the framework of the HGSVC, we have taken first steps to explore this new genotyping paradigm by extending WhatsHap to perform haplotype phasing and genotyping in a joint inference process. In a first implementation, we achieved a genotype concordance of 99.3% when genotyping SNVs from HGSVC PacBio data. We argue that this approach is particularly promising for low coverage sequencing where the pileup signal alone is insufficient for genotyping due to high error rates. We will refine this methodology, extend it to indels and SVs (see “Developing advanced methods for cross-platform haplotyping” above) and chart the achievable genotyping performance from different coverages of PacBio and ONT sequencing data sets. By providing this method, we pave the way to re-genotyping SVs as long-read sequencing methods become more widespread for clinical and research purposes. This will contribute to maximizing the benefit of our SV resource for the research community.

The goal of this Aim is to develop a lightweight genotyping method to rapidly genotype base pair-resolved SV calls produced by Phased-SV in large cohorts sequenced using Illumina. We will use the method to genotype 2500 publicly available high-coverage Illumina 1KG samples, and provide the software as a resource to the community.

The base-pair resolution of SVs from Phased-SV enables genotyping in large cohorts. As part of the HGSVC, three groups performed genotyping of Phased-SV variants on separate Illumina-based cohorts individually using DELLY, SVTyper, and SMRT-Genotyper {Chaisson, 2017 #376}. Both DELLY [(Rausch et al. 2012)](https://paperpile.com/c/BDfSjR/rtAMy) and SVTyper [(Chiang et al. 2015)](https://paperpile.com/c/BDfSjR/FQdyi) were developed for genotyping deletion SVs called from Illumina reads. When applied to deletions from the Phased-SV callset, DELLY produced genotypes for XXX% of Phased-SV deletion calls in 2,600 deep-coverage human genomes (XXX/YYY) {Chaisson, 2017 #376}, and SVTyper produced genotypes for 50.4% (18,119/35,936) of the same calls in 238 high-coverage genomes. SMRT-Genotyper [(Huddleston et al. 2017)](https://paperpile.com/c/BDfSjR/b3X58), which was developed for base pair-resolved calls produced by Phased-SV, genotyped 91.6% of sites in a separate 24 high-coverage genomes across both insertions and deletions. SMRT-SV genotypes SVs by mapping reads back to both the reference genome and the sequences of breakpoints detected in Phased-SV. This mapping step requires (1) the generation of a new reference and set of breakpoint sequences, and corresponding BWA index each time the SV callset is updated, and (2) the realignment of reads to this index.

To perform lightweight genotyping, we will develop a method that will reduce the burden of generating new reference indices and global alignment of reads. In particular, a pre-filtering step will be executed where all reads will be aligned to the reference, scoring alignments as concordant if full reads have full-length alignments within the span that is predicted by insert size, and discordant otherwise.  All reads in discordant alignments are retained and mapped to a comprehensive library of SV breakpoint sequences. In order to comprehensively cover the search space of SVs we will expand the sequence of breakpoints with gaps and single nucleotide substitutions. The extended breakpoint library will be indexed for efficient read mapping. A genome will be noted as having a variant if there are a sufficient number of reads concordant with the SV breakpoint. The variant will be homozygous if there are no concordant reads supporting the reference at the SV locus, and heterozygous otherwise.

In summary, the envisioned methods will put large-scale projects using standard short-read sequencing, including NIH funded projects (such as CCDG and CMG) in a position to reliably and efficiently genotype the SVs discovered as part of this resource.